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

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 introduces a *partial flexibility* (as opposed to full flexibility under continuous reassignment) 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 arrive according to stationary Poisson processes and 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 that provides an asymptotic lower-bound for the minimum expected total holding cost. We then propose a "translation" of the solution of the fluid control problem for the original stochastic system that achieves the asymptotic lower-bound. 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 for the system under time-varying arrivals and demonstrate their effectiveness using simulation experiments. 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 reduce the expected cost of the system by 20-50% 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 stochastic fluctuations in customer arrivals and service times, the congestion level in different areas can fluctuate considerably under dedicated staffing. That is, there are likely to be times when one area could be very congested while there are servers idling in other areas. Ideally, one could assign the servers dynamically in time as soon as changes in the congestion level in different areas 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. Therefore, 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 an application to 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 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. In this work, we investigate the reassignment decision and its potential benefits. We note that

our partner hospital is not unique in having a partitioned ED. For instance, in the ED discussed in Song et al. (2015) lower acuity patients (ESI levels 4 or 5) are treated in an area separate from the main ED, namely the Rapid Care Area.

Although our work is primarily motivated by the nurse staffing application described above, our model and analysis provides 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. We refer the reader to Martonosi (2011) for a detailed discussion of an application in airport security check-lines. In addition to service systems with physically separated areas, the partially flexible design could also be beneficial for services where there are diseconomies associated with full flexibility (or full pooling) of servers. For instance, in the case of a call center, Jouini et al. (2008) illustrates how migrating from a pooled system to a dedicated system where a team of servers is assigned to each customer class could increase service rates and reduce call-back rates 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. When a server becomes idle, the next customer to start service is the one at the head of the queue to which the server is assigned. The objective is to dynamically assign servers to the server pools in order to minimize the expected waiting cost of the system, 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 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). In a parallel server system, there are 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 which differentiates our work from the literature on dynamic control of parallel server systems 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 $\times$ service rate) is known to be optimal even under more general conditions and different cost-criteria. We refer the reader to Van Mieghem (1995) and Mandelbaum and Stolyar (2004) for a summary of these results and their generalizations for systems with convex delay costs.

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 a larger action space as well as more complex transition probabilities compared to the fully flexible design. The larger action space is easily observed noting that rather than choosing the next class to serve at each epoch, the controller needs to set the number of servers assigned to each pool. Further, unlike the control problem under the flexible design 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 $c\mu$ policy under the fully flexible design does not easily "translate" into useful policies for the problem under the partially flexible design. Typically, introducing additional flexibility is expected to improve system performance. However, we find instances where "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.

In this work, we propose an asymptotic framework and analyze associated fluid control problems that allow us to obtain asymptotically optimal polices that perform well for the original stochastic system. We also use the fluid models to derive insights and highlight key tradeoffs introduced by the shift constraint. Intuitively, if the system parameters are such that the queue length processes reach their steady-states quickly during a shift, then the (long-run average) expected cost of the system is approximately close to that of using dedicated staffing for all shifts. In contrast, if the system does not reach steady-state quickly during the shifts, as we show in this work, one can reduce the expected cost of the system by exploiting the observed congestion level at the beginning of the shifts.

1.1. Summary of contributions

In this work, we consider 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. In doing so, we propose and analyze a multiserver queueing model which allows for dynamic server allocation at discrete intervals. Our main results and contributions can be summarized as follows:

- We propose an *asymptotic parameter regime* suitable for studying the dynamic server assignment problem under the partially flexible design. By analyzing an associated discrete-time (deterministic) fluid control problem that arises in the limit, we develop an *asymptotic lower-bound* for the expected cost of the system (Theorem 1). The optimal policy for this fluid control problem can be obtained by solving a convex optimization problem. We exploit a dynamic programming formulation to show structural properties of the optimal policy (Theorem 2) and prove an important structural property for a two-class system (Theorem 3) that provides *insights into the structure of "good" policies under the partially flexible design.*

- We propose an *asymptotically optimal policy* for the stochastic system based on a "translation" of the solution to the fluid control problem. The policy is asymptotically optimal in the sense that it achieves the asymptotic lower-bound obtained for the expected cost of the system (Theorem 4). Using simulation, we evaluate the performance of the policy in terms of the expected long-run average cost of the system and *quantify the benefit of partial flexibility.* We observe that as long as the system does not quickly reach its steady-state during the shifts (that is, when the shift length is not "too long" compared to the average time between service completions) this additional flexibility can substantially reduce the expected cost (and average queue length) of the system compared to dedicated staffing.

• We leverage insights from the analysis of the fluid control problem to develop simple and intuitive heuristic policies for an extension of our model to the case with time-varying arrivals and demonstrate their effectiveness using simulation experiments. Importantly, we leverage the insight that it is important to carefully balance allocating more capacity to classes which impose the highest myopic costs (i.e., those with higher priority or congestion) versus the amount of idleness incurred in these classes during the shift, while customers wait in queues of other classes (i.e., those with lower priority or lower congestion at the beginning of the shift). 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. The results suggest that the usefulness and robustness of our results and insights go beyond our stylized model.

1.2. Related Work

There is a large literature on shift scheduling (or rostering) for service systems; we refer the reader to Van den Bergh et al. (2013) for a recent review and to Burke et al. (2004) for a review devoted to the nurse scheduling problem. This body of literature focuses on determining the number of staff required during small periods of time, and constructing work shifts and assigning available staff to them while taking into account various system and staff related constraints. The typical approach is to employ an Integer Stochastic Programming formulation without explicitly considering queueing dynamics. Accounting for queueing dynamics introduces complex stochastic interdependencies across shifts. Kolesar et al. (1975) (see also Green and Kolesar 1989 and the references therein) 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.

The complexities of provider shifts in healthcare settings are many. Batt et al. (2017) 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. They further use a simulation model to investigate the benefits of "cut-off" policies for starting new work before the end of a shift. 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 2017a); see Harrison and López (1999) and Dai and Tezcan (2011) for a general formulation. 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 fundamental 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. Two common asymptotic regimes are the conventional heavy-traffic (e.g., Harrison 1998, Harrison and López 1999) and the many-server heavy-traffic (or Halfin-Whitt) regimes (e.g., Harrison and Zeevi 2004, Armony 2005, Gurvich et al. 2008, Gurvich and Whitt 2010, Tezcan and Dai 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. The general idea is to solve an associated fluid (or transient) control problem, translate the solution for the original stochastic network, and then formally establish a result on the accuracy of the approximate control. Examples include Meyn (1997), Chen and Meyn (1999), Maglaras (1999, 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 establishes its fluid-scale asymptotic optimality in the fluid asymptotic regime formalized in Dai (1995) (see also Chen and

Mandelbaum 1991). 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. The main idea in Maglaras (2000) is to find a policy for the stochastic system such that under the fluid or Functional Strong Law of Large Numbers (FSLL) type scaling, the dynamics of the stochastic system approaches that of the optimal fluid trajectories in the limit. In that sense, the policy is called asymptotically tracking and its fluid-scale asymptotic optimality is established.

The fluid asymptotic regime used in Dai (1995) and Maglaras (2000) involves considering a sequence of increasing initial conditions and employing fluid-scaling in time and space. 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, Ataman and Kumar 2005, 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 (see Section 7.2).

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 large arrivals and number of 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.

We note that dynamic assignment of servers has been studied in other contexts and for other queueing networks, again focusing on continuous-time control and without restricting control to discrete intervals. For example, Andradóttir et al. (2003) studies the dynamic allocation of flexible servers in a general network of multiclass single-server queues with random switching times, and Tsai and Argon (2008) considers server assignment policies for assembly-type single-server queues. Related to dynamic server assignment is the problem of "load balancing" in parallel single-server queues considered in Down and Lewis (2006) and Caudillo-Fuentes et al. (2010) where customers, rather than servers, are moved from one queue to another in order to balance congestion.

Finally, our study relates to the growing literature on queueing models of patient flow in hospitals. Armony et al. (2015) presents a general description of patient flow in hospitals from a queueing perspective and based on data from an Israeli hospital. Whitt and Zhang (2017) present an aggregate stochastic model of the ED using the same data as in Armony et al. (2015). More specialized queueing models have been proposed in the literature to compare different patient-flow designs (e.g., Saghafian et al. 2012, Yildiz et al. 2015) or evaluate the impact of operational decisions on patient waiting times (Shi et al. 2015, Chan et al. 2016, Dai and Shi 2017b). 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 to satisfy certain constraints on quality of service (waiting time). In contrast, we assume a fixed number of providers (nurses) and consider their dynamic assignment while taking into account the shift constraint. Other papers in the literature have studied dynamic scheduling or control of patient flows in the inpatient ward, ED, or from the ED to wards. For instance, Mandelbaum et al. (2012) studies the problem of dynamic routing from the ED to inpatient wards with the goal of balancing idleness and workload amongst wards. Huang et al. (2015) presents an asymptotically optimal scheduling policy (in the conventional heavy-traffic regime) for patient flow control in the ED through physicians. The policy determines which patient is to be seen next by a physician while taking into account the tradeoff between seeing new patients versus the ones closer to being discharged. Dai and Shi (2017a) propose an approximate dynamic programming approach for making overflow decisions for assigning patients to inpatient wards. To the best of our knowledge, dynamic assignment of providers, the main focus of our work, has not been previously considered in the literature.

1.3. Notation

We denote the nonnegative real line $[0,\infty)$ by $\mathbb{R}_+$ 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 $\|f^n(t)-f(t)\|_T \to 0$ 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 will be provided in the Appendix.

2. Model Description and Formulation of the Control Problem

To study the impact of the shift constraint, we consider a queueing system with I customer classes, I parallel queues, and I server pools. Class $i \in \mathcal{I} \equiv \{1,\ldots,I\}$ 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.

A dynamic control policy (determining the assignment of servers to pools) in our formulation takes the form of a stochastic *assignment* process $U(t) = (U_1(t),\ldots,U_I(t))$ taking values in $\mathbb{R}^I_+$ and with RCLL sample paths. Denote by $A_i(t), i \in \mathcal{I}$ and $S_i(t), i \in \mathcal{I}$ a set of $2I$ mutually independent standard (unit rate) Poisson processes and let $X(0) \in \mathbb{R}^I_+$ 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 $\mathbb{R}_+^I$ 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 in accordance with the description given at the beginning of this section. 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 $i' \in \mathcal{I}$. 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.

Before turning to admissible policies under the partially flexible design, we make a few remarks. First, note that we do not require the policies to be non-anticipating, i.e., adapted to the natural filtration generated by the stochastic primitives A, S, and $X(0)$. Second, note that 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, 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 $k+1$

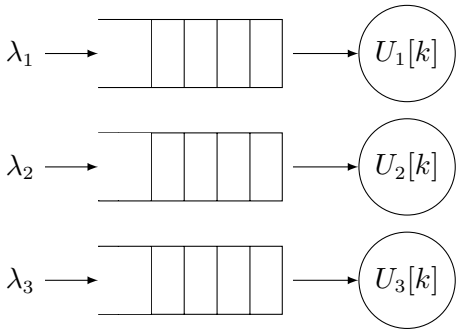

Figure 1 **A three-class system under the partially flexible design during shift** $k+1$

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. An example of the system under the partially flexible design is illustrated in Figure 1.

The requirement in (8) captures the shift constraint, distinguishing our formulation from that of the more commonly studied continuous-time control problems. We note that (1)-(2) can be viewed as a special case of the general formulation in Bassamboo et al. (2006) with a specified initial condition. In Bassamboo et al. (2006) and other similar continuous-time control formulations, e.g., Atar et al. (2010), the number of servers busy serving each class (process Z in our formulation) is taken as the control process. In our formulation, the assignment process U is introduced to model discrete-time control (or discrete-review with fixed review length) policies. In fact, 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. 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.,

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

3. Asymptotic Framework and Limiting Dynamics

In this section we consider a sequence of systems, as described above, indexed by n. The nth system has n servers, arrival rate $\lambda^n=(\lambda_1^n,\ldots,\lambda_I^n)$, service rates $\mu^n=(\mu_1^n,\ldots,\mu_I^n)$ and is initiated with

$X^n(0)$ customers in the system. Further, under the partially flexible design, the server assignment can only be controlled at the beginning of shifts of length τ^n.

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 $n \to \infty$,*

$$n^{-1}X^n(0) \to \bar{X}(0), \tag{10}$$

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 $n \to \infty$,*

$$\bar{\lambda}^n_i \equiv n^{-1}\lambda^n_i \to \bar{\lambda}_i. \tag{11}$$

We denote the stochastic processes associated with the nth system with X^n, Q^n, and Z^n. 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:

$$X^n_i(t) = X^n_i(0) + A_i(\lambda^n_i t) - S_i\left(\mu_i T^n_i(t)\right) \geq 0, \quad \text{for all } i \in \mathcal{I}, \tag{12}$$

$$T^n_i(t) = \int_0^t Z^n_i(s)ds, \quad \text{for all } i \in \mathcal{I}, \tag{13}$$

$$Q^n(t) = X^n(t) - Z^n(t) \geq 0, \tag{14}$$

$$U^n(t) - Z^n(t) \geq 0, \tag{15}$$

$$Q^n_i(t)(U^n_i(t) - Z^n_i(t)) = 0, \quad \text{for all } i \in \mathcal{I}, \tag{16}$$

$$U^n(t) \cdot \mathrm{e} \leq n, \tag{17}$$

$$U^n(t) \geq 0. \tag{18}$$

Note that the relations are the same as those in (1)–(6) except that $T^n_i(t)$ is used to express the cumulative amount of time that all n servers have spent on class i up to time t. We denote the set of all admissible policies under the fully flexible design by Π^n.

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 (12)–(18) and for almost all sample paths,

$$U^n(t) = \sum_{k=0}^{\bar{k}(t)-1} U^n[k]\mathbf{1}_{[t_k, t_{k+1})}(t), \tag{19}$$

for some sequence $\{U^n[k]; 0 \le k \le \bar{k}(t)-1\}$ 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 $\bar{k}(t)=\min\{k; t_k \ge t\}$. Note that (19) limits the set of admissible policies under the partially flexible design to those with piecewise-constant sample paths. We denote the set of discrete-review policies with shift length τ by Π^n_τ. Clearly, $\Pi^n_\tau \subset \Pi^n$.

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

$$C^n_T(U^n) = \int_0^T h\cdot Q^n(s)ds. \tag{20}$$

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

$$Y^n(t) = \int_0^t U^n(s)ds, \quad \text{and} \quad \Upsilon^n(t) = \int_0^t Q^n(s)ds, \tag{21}$$

and define the fluid-scaled processes,

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

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,*

$$\limsup_{n\to\infty} \mathbb{E}[n^{-1}C_T(U^n_*)] \le \liminf_{n\to\infty} \mathbb{E}[n^{-1}C_T(U^n)], \tag{22}$$

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

Our first result 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 (24)–(30). Equation (24) 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., Armony 2005 and Tezcan and Dai 2010, Appendix B) and holds for any sequence of admissible controls subject to the limiting relations (25)–(28). Equation (30) 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 (30) is uniform.

Proposition 1 *Consider any sequence of admissible controls* $\{U^n\}$ *and the associated processes satisfying (12)–(18). 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,*

$$\left(\bar{X}^{n_j}(t), \bar{T}^{n_j}(t), \bar{Y}^{n_j}(t), \bar{\Upsilon}^{n_j}(t)\right) \to \left(\bar{X}(t), \bar{T}(t), \bar{Y}(t), \bar{\Upsilon}(t)\right), \quad u.o.c., \tag{23}$$

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

$$\bar{X}_i(t) = \bar{X}_i(0) + \bar{\lambda}_i t - \mu_i \bar{T}_i(t) \geq 0, \quad \text{for all } i \in \mathcal{I}. \tag{24}$$

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:*

$$\bar{Q}(t) = \bar{X}(t) - \bar{Z}(t) \geq 0, \tag{25}$$

$$\bar{Z}(t) = \bar{U}(t) \wedge \bar{X}(t) \tag{26}$$

$$\bar{U}(t) \cdot e \leq 1, \tag{27}$$

$$\bar{U}(t) \geq 0. \tag{28}$$

Finally, if $\{U^n\}$ *is a sequence of discrete-review policies, i.e., it satisfies (19), then for almost all sample paths,*

$$\bar{U}^{n_j}(t) \to \bar{U}(t), \quad u.o.c., \tag{29}$$

where

$$\bar{U}(t) = \sum_{k=0}^{\bar{k}(t)-1} \bar{U}[k] \mathbf{1}_{[t_k, t_{k+1})}(t), \tag{30}$$

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

We use this proposition later in developing asymptotic lower bounds and establishing the asymptotic optimality of our proposed policy.

4. Fluid Control Problems and Asymptotic Lower Bounds

In this section we develop an asymptotic lower bound for the expected cost of the stochastic system under any family of admissible polices Π^n, and a tighter bound for the subset of discrete-review admissible policies Π^n_τ. To this end, we first introduce two deterministic fluid control problems.

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

$$J_T(x_0) \equiv \min_{u(\cdot)} \int_0^T h \cdot q(s) ds \tag{31}$$

$$\text{s.t.} \quad q(t) = (x(t) - u(t))^+ \quad \forall t \in [0, T], \tag{32}$$

$$\dot{x}(t) = \bar{\lambda} - \hat{\mu} \cdot (x(t) \wedge u(t)), \quad \forall t \in [0, T], \tag{33}$$

$$x(0) = x^0, \tag{34}$$

$$u(t) \cdot \mathrm{e} \leq 1, \quad \forall t \in [0,T], \tag{35}$$

$$u(t) \geq 0, \quad \forall t \in [0,T], \tag{36}$$

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 (33)–(36). 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 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 (31).

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 (33) with a constant allocation $u(t) = u[k]$ during stage (shift) $k+1$. Since the RHS of (33) 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

$$g^k(x,u) = \int_{\tau k}^{\tau(k+1)} h \cdot q(s) ds = \int_{\tau k}^{\tau(k+1)} h \cdot \left(f^k(x,u,s) - u\right)^+ ds. \tag{37}$$

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

$$J_{T,\tau}(x^0) \equiv \min_{\{u[k]; k \in \mathcal{K}\}} \sum_{k=0}^{N-1} g^k(x[k], u[k]) \tag{38}$$

$$\text{s.t.} \quad x[k+1] = f^k(x[k], u[k], (k+1)\tau), \quad \forall k \in \mathcal{K}, \tag{39}$$

$$x[0] = x^0, \tag{40}$$

$$u[k] \cdot \mathrm{e} \leq 1, \quad \forall k \in \mathcal{K}, \tag{41}$$

$$u[k] \geq 0, \quad \forall k \in \mathcal{K}. \tag{42}$$

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

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*

$$\liminf_{n\to\infty} \mathbb{E}[n^{-1}C_T^n(U^n)] \geq \mathbb{E}[J_T(\bar{X}(0))]. \tag{43}$$

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

$$\liminf_{n\to\infty} \mathbb{E}[n^{-1}C_T^n(U^n)] \geq \mathbb{E}[J_{T,\tau}(\bar{X}(0))], \tag{44}$$

and

$$\mathbb{E}[J_{T,\tau}(\bar{X}(0))] \geq \mathbb{E}[J_T(\bar{X}(0))]. \tag{45}$$

5. 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 $u^d \cdot \mathrm{e} \leq 1$. Such a policy serves as a simple benchmark analogous to how dedicated staffing does in the original stochastic system. In this case, given an initial condition $x^0 \in \mathbb{R}_+^I$ each class $i \in \mathcal{I}$ behaves as an independent multiserver fluid queue governed by the ODE

$$\dot{x}_i(t) = \bar{\lambda}_i - \mu_i(x_i(t) \wedge u_i^d), \quad i \in \mathcal{I},\ t \in [0,T]. \tag{46}$$

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 and easy to show 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 $t \to \infty$. In the language of dynamical systems, $\bar{\rho}_i$ is referred to as a *globally asymptotically stable stationary point*. Note that under a dedicated allocation with $\bar{\rho}_i < u_i^d$ 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.

The existence of a stable dedicated allocation has an important implication for the fluid control problems presented in Section 4. Specifically, since any dedicated allocation is a feasible solution for both control problems, given the existence of a stable dedicated allocation, the optimal cost of Problems 1 and 2 must also remain finite as $T \to \infty$. The existence of a stable dedicated solution is clearly not a necessary condition for the total cost of the infinite horizon system to remain finite under continuous or discrete-time control. In fact, for the continuous-time Problem 1, as we discuss below, the optimal cost is finite given $\sum_{i\in\mathcal{I}} \bar{\rho} < 1$. For Problem 2, the existence of a stable dedicated allocation provides a simple, yet practical, sufficient condition. In Section 5.2.2, we use the finiteness of the total cost to prescribe stationary (time-independent) policies for Problem 2. Note that since many systems utilize a dedicated allocation, it is reasonable to consider the system in such a regime. That said, our analysis for the finite horizon control problem does not require such an assumption.

5.1. 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). The optimality of the $c\mu$ rule discussed here for the multiserver setting can be formalized using the associated Hamiltonian formulation of the problem. Here we discuss the policy and present some numerical examples of the optimal solution in order to contrast its structure to that of Problem 2. We will return to the $c\mu$ policy in Section 7.1 where we use the solution of the continuous-time control problem to construct a discrete-review policy for the time-varying stochastic system.

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 $h_1\mu_1 \geq \ldots \geq h_I\mu_I$. Then, under the $c\mu$ policy, the vector of capacity allocation $\hat{u}(t)$ is given by

$$\hat{u}_i(t) = x_i(t) \wedge \left[1 - \sum_{j=1}^{i-1} x_i(t)\right]^+, \quad i \in \mathcal{I},\, t \in [0,T], \tag{47}$$

with the optimal trajectory satisfying the following system of ODEs (expressed in integral form) starting from x^0:

$$x_i(t) = x_i^0 + \lambda_i t - \mu_i \int_0^t x_i(s) \wedge \left[1 - \sum_{j=1}^{i-1} x_i(s)\right]^+ ds, \quad i \in \mathcal{I},\, t \in [0,T]. \tag{48}$$

One can show (e.g., using a similar argument to the one in the Appendix of Atar et al. 2010) that 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. That is, $\bar{\rho}$ is also a globally asymptotically stable stationary point for the system of ODEs in (48) given that $\sum_{i\in\mathcal{I}} \bar{\rho}_i < 1$.

In general, the system of ODEs can be solved numerically. Figure 2 presents the optimal 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. Further, note the trajectories converge to the stationary point $\bar{\rho}_i$ for each class as $t \to \infty$.

5.2. Discrete-Time Control (Problem 2)

5.2.1. 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),

$$\dot{x}_i(t) = \bar{\lambda}_i - \mu_i \left(x_i(t) \wedge u_i\right), \quad i \in \mathcal{I} \text{ and } t \geq k\tau, \tag{49}$$

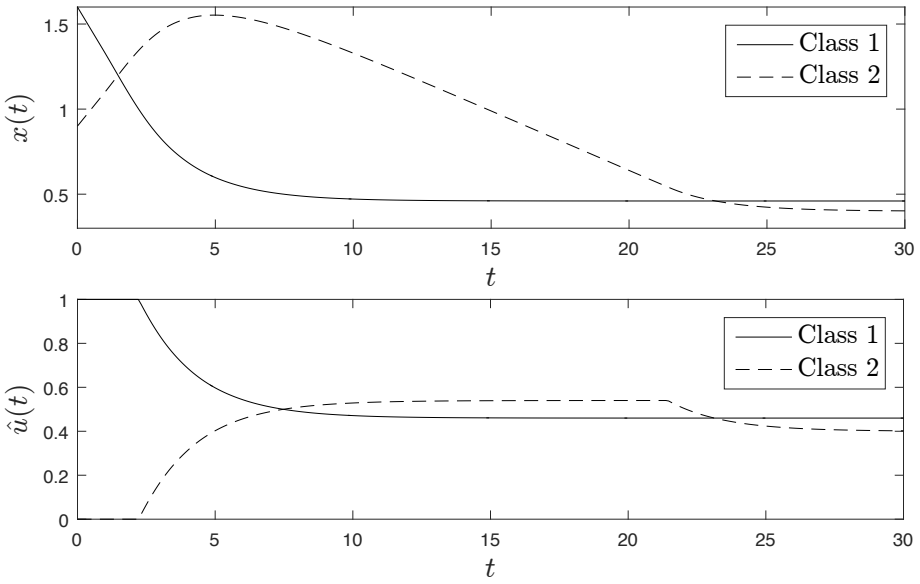

Figure 2 $c\mu$ **trajectories (top) and allocations (bottom) for a** $I=2$ **class system with** $\bar{\lambda}=(0.23,0.20)$**,** $\mu=(0.5,0.5)$**, and initial conditions** $x^0=(1.6,0.9)$**.**

given the vector of initial condition $x=x(k\tau)$ and allocation $u=u(k\tau)$. Since the problem is time-homogeneous (we discuss the time-varying case in Section 7) we can express the dynamics by doing so for the first stage and restarting time at the beginning of each stage. Denote the dynamics over the first 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 $i\in\mathcal{I}$. The solution is found by solving the ODE in (49) 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,

$$f_i(x,u,t)=\begin{cases} x_i+(\bar{\lambda}_i-\mu_i u_i)t, & u_i<x_i \text{ and } u_i\le \hat{u}_i(x,t), \quad \text{(Case 1)}\\ \bar{\rho}_i+e^{-\mu_i(t-\sigma_i)}(u_i-\bar{\rho}_i), & u_i<x_i \text{ and } u_i>\hat{u}_i(x,t), \quad \text{(Case 2)}\\ u_i+(\bar{\lambda}_i-u_i\mu_i)(t-\nu_i), & u_i\ge x_i \text{ and } u_i<\check{u}_i(x,t), \quad \text{(Case 3)}\\ \bar{\rho}_i+e^{-\mu_i t}(x_i-\bar{\rho}_i), & u_i\ge x_i \text{ and } u_i\ge \check{u}_i(x,t), \quad \text{(Case 4)}\end{cases} \tag{50}$$

where

$$\hat{u}_i(x,t)=\frac{x_i+t\bar{\lambda}_i}{1+t\mu_i}, \quad \check{u}_i(x,t)=\bar{\rho}_i+e^{-\mu_i t}(x_i-\bar{\rho}_i), \tag{51}$$

and

$$\sigma_i(x_i,u_i)=\frac{1}{\mu_i}\frac{x_i-u_i}{u_i-\bar{\rho}_i}, \quad \nu_i(x_i,u_i)=\frac{1}{\mu_i}\log\left(\frac{\bar{\rho}_i-x_i}{\bar{\rho}_i-u_i}\right). \tag{52}$$

Note that for ease of notation we have suppressed the dependence of σ_i and ν_i on x_i and u_i in (50); 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 (50) 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)

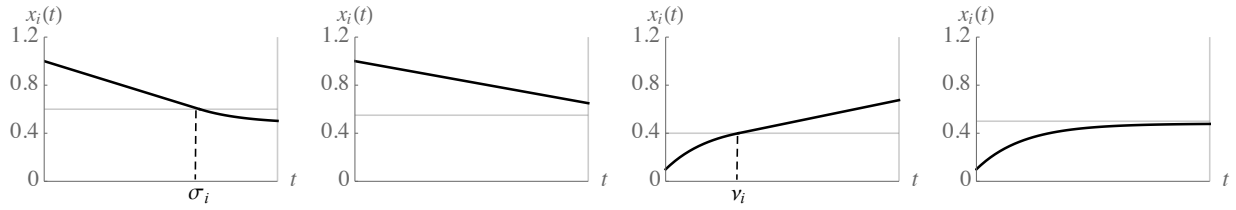

Figure 3 **Possible trajectories of $x_i(t)$ during a stage under different allocations and initial conditions. (The horizontal lines correspond to the allocations.)**

when the queue empties. Further, $\hat{u}_i(x,t)$ 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 3. 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 (37) 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

$$g(x,u) = h \cdot c(x,u), \tag{53}$$

where $c_i(x,u)$ is the class i stage cost given initial condition x and allocation u, and can be computed using (50) and (37):

$$c_i(x,u) = \begin{cases} (x_i - u_i)\tau - (\mu_i u_i - \lambda_i)\tau^2/2, & u_i < x_i \text{ and } u_i \le \hat{u}_i(x,\tau), \\ (x_i - u_i)^2/(2(\mu_i u_i - \lambda_i)), & u_i < x_i \text{ and } u_i > \hat{u}_i(x,\tau), \\ (\lambda_i - u_i\mu_i)(\tau - \nu_i)^2/2 & u_i \ge x_i \text{ and } u_i < \check{u}_i(x,\tau), \\ 0, & u_i \ge x_i \text{ and } u_i \ge \check{u}_i(x,\tau). \end{cases} \tag{54}$$

Note that the four cases in (54) are the same as those in (50) with fixed $t = \tau$. Further, it is easy to verify using (54) that c_i and hence g are continuous in (x,u).

5.2.2. 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

$$V^k(x) = \min_{u \in \mathcal{U}} \left[g(x,u) + V^{k+1}(f(x,u,\tau)) \right], \tag{55}$$

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

$$\phi^k(x) = \{u \in \mathcal{U}; V^k(x) = \tilde{V}^k(x,u)\}, \quad k \in \mathcal{K}, \tag{56}$$

where,

$$\tilde{V}^k(x,u) \equiv g(x,u) + V^{k+1}(f(x,u,\tau)). \tag{57}$$

The following result establishes the continuity of the value function $V^k(\cdot)$ and existence of an optimal (possibly non-unique) solution to Problem 2.

Proposition 2 *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 (57) is convex in (x,u) and hence evaluating the optimal policy using (55) 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 $k \in \mathcal{K}$.*

Note that, in general, the optimal policy depends on both time k and state x. However, for a sufficiently large horizon T and a total cost that remains finite we can obtain a *stationary (time-independent) policy* $\phi(\cdot)$ that is only a function of the current state. Assume that a stable dedicated allocation exists and that N (or equivalently T) can be chosen sufficiently large. As discussed in the beginning of this section, this implies that under the optimal policy and starting from any x the queues eventually empty and the future cost becomes zero. Therefore, for any x and starting at $k = 0$, there exists a finite $\bar{N}(x)$ such that the optimal future cost after $\bar{N}(x)$ is zero. Further, since the costs and transition functions are stationary, the problem at any time k is the same at 0 with $N - k$ stages to go. So with N satisfying $N \geq \bar{N}(x) + k$ the optimal allocation at any time k is identical to that at time $k = 0$. Numerically, the stationary policy $\phi(\cdot)$ can be obtained by solving Problem 2 given a starting point and choosing a sufficiently large N.

5.2.3. Structural Results and Insights In this section we discuss an important property of the optimal solutions to Problem 2. The main result provides fundamental 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 to derive a high-performing heuristic. Before presenting the main result we first present two intermediary results. The first result establishes the existence of a *non-wasteful* optimal policy that allocates all capacity in all stages. Note, 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 3 *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 $k \in \mathcal{K}$.*

The following lemma will be useful in the proof of the main result and its discussion. It 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 1 *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 $k \in \mathcal{K}$.*

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 $\phi^k(x[k])$ satisfying $\phi_1^k(x[k]) \geq \hat{u}_1(x[k], \tau) \wedge 1$.*

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 beginning of the next stage. Therefore, 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.

To facilitate the discussion of the assumptions of the theorem and to further illustrate the result, we next present an example. In Figure 4, we plot the optimal trajectories for a system with the same parameters as the experiments in Figure 2 – $\bar{\lambda} = (0.23, 0.20)$, $\mu = (0.5, 0.5)$, $h = (4, 2)$, and initial condition $x^0 = (1.6, 0.9)$ – 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. However, it does not allocate all capacity to class 1, as would be the case of continuous-time control since that would result in "too much" idleness. The optimal amount of idleness (or the time the queue empties) depends on the cost and system parameters (including the arrival rates). We see that, as the theorem predicts, the allocations are such that the queue is emptied during each stage.

We now turn to the assumptions. Note that condition (ii) is the $c\mu-$index order assumed in the continuous-time problem. Condition (iii), which is in fact necessary for the statement to hold, is slightly stronger. To understand condition (iii), note that $h_i\mu_i\tau^2/2 - h_i\tau$ is the marginal rate of decrease for class i stage-cost, when allocating more capacity to it in the absence of idleness (see

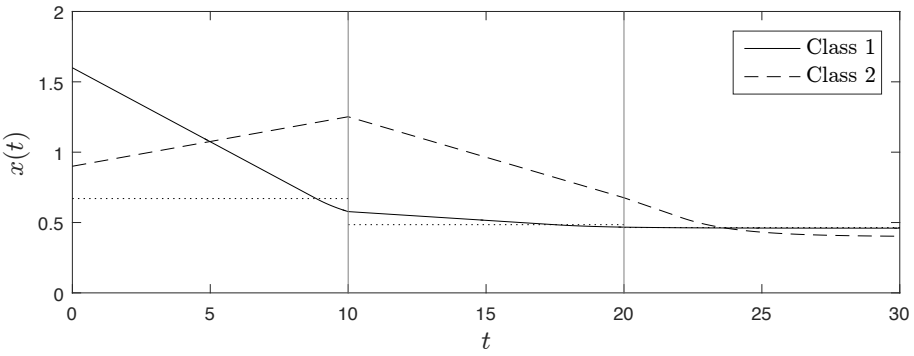

Figure 4 **Optimal class 1 (solid) and class 2 (dashed) trajectories for Problem 2 with $\bar{\lambda} = (0.23, 0.20)$, $\mu = (0.5, 0.5)$, $x^0 = (1.6, 0.9)$, $h = (4, 2)$ and $\tau = 10$. The dotted horizontal lines correspond to optimal class 1 allocations for each stage.**

(93) in the proof). Therefore, 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. More specifically, 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,

$$\phi_1^k(x[k]) \geq \hat{u}_1(x[k], \tau) \wedge 1, \quad x_1[k] > \bar{\rho}_1, \tag{58}$$

$$\phi_1^k(x[k]) \geq x_1[k] \wedge 1, \quad x_1[k] \leq \bar{\rho}_1. \tag{59}$$

Note that in each case, the lower-bound is the maximum allocation such that no idleness is incurred in class 1. Further, the lower-bound in (58) is the same as in Theorem 3 stated under the assumption $x[k] > \bar{\rho}$. (Recall that by Lemma 1, $x^0 > \bar{\rho}$ ensures $x[k] > \bar{\rho}$ for all $k \in \mathcal{K}$.)

We demonstrate the above using an additional example. In Figure 5, we plot the optimal trajectories for the same parameters as in Figure 4, but with initial condition $x^0 = (0.3, 1.5)$. Noting that the initial condition of class 1 is below the equilibrium point, i.e., $x_1 < \bar{\rho}_1$, we observe that the optimal allocation in stage 1 satisfies $x_1 \leq u_1 < \bar{\rho}_1$. 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.

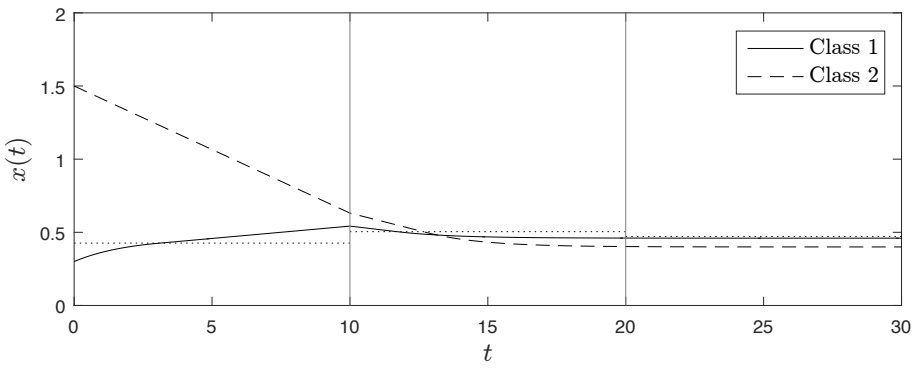

Figure 5 **Optimal class 1 (solid) and class 2 (dashed) trajectories for Problem 2 with $\bar{\lambda} = (0.23, 0.20)$, $\mu = (0.5, 0.5)$, $x^0 = (0.3, 1.5)$, $h = (4, 2)$, and $\tau = 10$. The dotted horizontal lines correspond to optimal class 1 allocations for each stage.**

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

In this section, we present an asymptotically optimal discrete-review policy as defined in Definition 1. 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 5.2.2 that the solution to Problem 2 may not be unique. Intuitively, the optimal allocation in each stage is unique unless the future cost can be kept at zero (low initial conditions) or system parameters are such that allocating capacity to different classes has the same marginal value (large initial conditions). 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 the system size (as measured by n) is not large enough and 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 could lead to poor performance especially when using a stationary solution of Problem 2 (by considering a sufficiently large number of stages N as discussed in Section 5.2.2) and evaluating the performance of the policy in terms of the long-run average cost of the system. To address this issue, we introduce a safety parameter to protect lower-priority classes in "small" systems while still maintaining asymptotic optimality. In our problem, the use of safety parameters is motivated by the discussion after Theorem 3 and 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 $\beta^n = [\beta_1^n, \beta_2^n, \ldots, \beta_I^n]$ satisfying $\beta^n = o(n)$ that is subtracted from the system state so that the server allocation decision is determined based on the modified state, $X_*^n(t) - \beta^n$. The use of safety parameters (or safety stocks)

for translation of fluid solutions have been previously used 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 8.

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 $t_{k+1} - t_k = \tau$. At each time instance t_k, 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

$$\tilde{U}^n_*[k] \equiv \hat{\phi}^k \left((\bar{X}^n_*(t_k) - \bar{\beta}^n)^+ \right). \tag{60}$$

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

$$U^n_*[k] = \left\lfloor n\tilde{U}^n_*[k] \right\rfloor, \tag{61}$$

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

By the above construction, it is easy to see that the assignment process under the DR policy is feasible and satisfies the definition of a discrete-review policy in (19), i.e., $U^n_* \in \Pi^n_\tau$ for all n. We also note that, as we exemplify in Section 8, the rounding in (61) can be carried out more carefully, e.g., using the largest reminder method, to avoid wasting capacity. 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 (44) in Theorem 1:*

$$\limsup_{n\to\infty} n^{-1}\mathbb{E}[C^n_T(U^n_*)] \le \mathbb{E}[J_{T,\tau}(\bar{X}(0))]. \tag{62}$$

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 two examples corresponding to a $I = 2$ class system with $n = 80$ servers. For both examples 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

Initial condition	Performance	Lower Bounds	
$n^{-1}X^n(0)$	$\mathbb{E}[n^{-1}C_T^n(U_*^n)]$	$J_T^\tau(x^0)$	$J_T(x^0)$
$(1.6, 0.9)$	52.36 ± 1.83	42.02	33.48
$(0.3, 1.5)$	20.14 ± 1.01	11.95	9.14

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

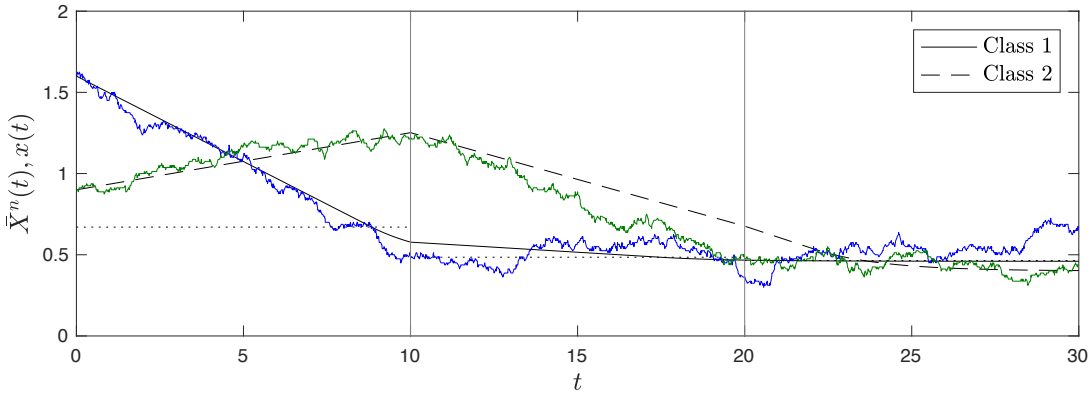

Figure 6 **A sample path of $\bar{X}^n(t)$ under the DR policy overlaid on corresponding optimal trajectories for $I = 2$ and system parameters $n = 80$, $\mu = (0.5, 0.5)$, $\lambda^n = (0.23n, 0.20n)$, $X^n(0) = (1.6n, 0.9n)$ and $\tau = 10$.**

$X_i^n(0) = nx_i^0$ for $i \in \{1, 2\}$, with $x^0 = (1.6, 0.9)$ in the first example and $x^0 = (0.3, 1.5)$ in the second. The holding cost rate is assumed to be $h = (4, 2)$. Note that the scaled parameters (i.e., $\bar{\lambda}$ and x^0) as well as the holding cost rate and shift length are the same as those in the examples of Section 5.2.3, with the corresponding optimal trajectories presented in Figures 4 and 5. In both examples, we set the safety parameter to $\beta^n = (0, 0)$.

We estimate the expected cost of the system under the DR policy over a finite horizon $T = 30$ (3 shifts) by simulating 300 sample paths. Table 1 presents the simulated costs as well as the asymptotic lower-bounds of Section 5. 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 $\mathbb{E}\,[n^{-1}C_T^n(U_*^n)]$ is due to the effect of stochastic fluctuations of $o(n)$, while the larger gap between $\mathbb{E}\,[n^{-1}C_T^n(U_*^n)]$ and $J_T(x^0)$ also includes the effect of restricting control to the beginning of shifts of length $\tau = 10$.

Figures 6 and 7 demonstrate the same (optimal) trajectories in Figures 4 and 5 with a (scaled) sample path of the stochastic $n = 80$ server system overlaid on them. Note that starting from an initial condition, the stationary DR policy solves for the optimal fluid allocations over a long horizon (here $N = 6$). 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

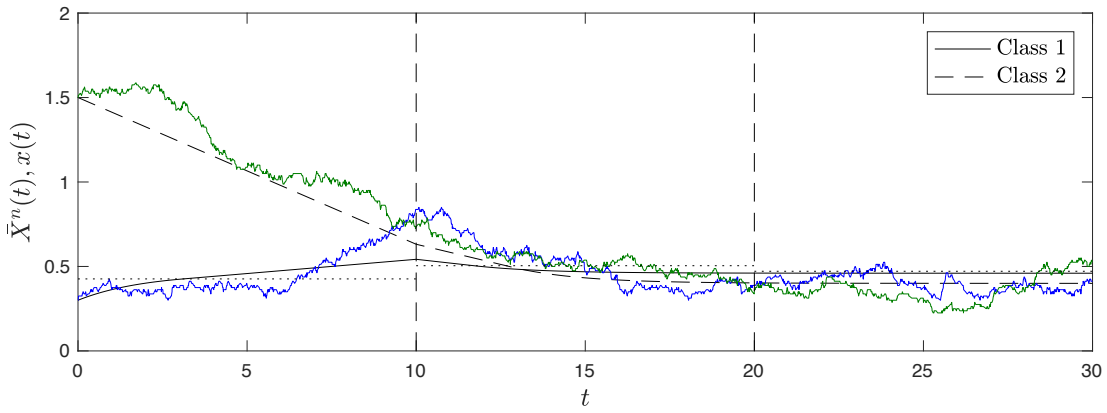

Figure 7 **A sample path of $\bar{X}^n(t)$ under the DR policy overlaid on corresponding optimal trajectories for $I=2$ and system parameters $n=80$, $\mu=(0.5,0.5)$, $\lambda^n=(0.23n,0.20n)$, $X^n(0)=(0.3n,1.5n)$ and $\tau=10$.**

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.

7. Extension to Time–Varying Arrivals and Heuristic Policies

So far we have assumed a stationary arrival process. In many application, including our motivating ED setting, the arrival process is not stationary and exhibits significant temporal variation. Therefore, we consider an extension of our results to the case where the arrivals are driven by non-homogeneous Poisson processes with periodic time-varying rates (e.g. Kim and Whitt 2014 and Armony et al. 2015). We begin with 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).

Given that Assumption 3 holds, the FSLLN for non-homogeneous Poisson processes implies that the results of Proposition 1 and Theorem 1 directly extend to the case with time-varying arrivals on compact intervals. That is, solving Problem 2 with time-dependent dynamics (to be discussed shortly), one can obtain the optimal fluid allocations for a finite horizon and use it in the DR policy

described in Section 6 to obtain an asymptotically optimal assignment for the original problem. Because the approach is similar to that of our results in Section 6, we do not provide the details here. We note that while showing convexity of Problem 2 in general seems hard, numerical examples confirm the intuition that, regardless of the input, the total cost should be convex in the allocation as formally shown in the stationary case.

Extending the result to the infinite horizon is not straightforward. Recall that in the stationary case, assuming a stable dedicated allocation was enough to keep the total cost finite as $T \to \infty$. This allowed us to obtain stationary (time-independent) policies in that setting. In contrast, with a time-varying input, the cost of the system does not necessarily remain finite as $T \to \infty$ even under appropriate stability conditions. That is, the equilibrium of the system may involve a queue that builds up and drains during the period, leading to a total cost that is increasing without bound in the length of the horizon. Therefore, with time-varying arrivals, one needs to consider the long-run average cost criterion and solve a DP where the value function takes the form of a differential sum of costs instead of the total cost-to-go considered in Section 5.2.2 (see, e.g., Bertsekas 2005). In this case, one expects the optimal policy to take the form of a periodic map from the state to a server allocation with the same period as the arrivals. We do not pursue this approach in this paper and leave a thorough investigation of the long-run average optimal control problem and its relation to the corresponding stochastic control problem for future research. Instead, we use the insights obtained from the analysis of the solution to Problem 2 in the stationary case to develop a practical and intuitive heuristic policy that performs well. In addition, we use ideas from the literature (specifically, Maglaras 2000), to develop a tracking policy based on the optimal solution of Problem 1 under time-varying arrivals.

Both heuristics construct a periodic map that takes the shift of the day and the initial state at the beginning of that shift as inputs, and returns a feasible fluid-scale allocation vector (similar to $\phi_*^k(\cdot)$ for the DR policy) for that shift. The allocations can then be translated to feasible assignments for the original n server stochastic system using a similar procedure described for the DR policy in Section 6. That is, the constructed map is applied to the normalized state of the system observed at the beginning of each shift, after possibly modifying it by a safety parameter. The result is then multiplied by n and then properly rounded to obtain an integer assignment. Note that the periodic policies depend on time, but only through the shift of the day.

7.1. The Time-Varying Heuristic (TVH) Policy

The TVH is based on the insights gained from the analysis of the stationary discrete-time control problem (Problem 2) discussed in Section 5.2.3. Recall that the main insight there was that the marginal value of allocating more capacity to the higher-priority class is the greatest as long as

no idleness is incurred in that class during the stage. However, the value starts diminishing once idleness is incurred. If all classes have the same priority level, the insight implies that the marginal value of adding capacity to any of the classes would be the same in the absence of idleness during the stage. The main idea behind the TVH is to allocate capacity to different classes at the beginning of each stage such that none of them incur *excessive idleness* during the stage.

The output of the TVH is a fluid-scale periodic allocation map, denoted by $\psi(x,l)$, where $x \in \mathbb{R}^I_+$ is the state and $l+1 \in \{1,\dots,L\}$ is the shift of the day. As discussed above, the policy can then be translated back to the stochastic n server system using a similar procedure described for the DR policy. However, unlike DR (and the CMUT policy described in the next subsection), TVH does not require a safety parameter (we discuss this point further shortly). Denote by $\{t_l; l \in \{0,\dots,L-1\}\}$ the time-instances corresponding to the hour of the day when a shift starts. Then, given a vector of initial conditions $x^0 \equiv x(t_l)$ and a piecewise-constant allocation vector $u(t) \geq 0$ satisfying $u(t)\cdot \mathrm{e} \leq 1$, the (fluid) dynamics of the system are given by the IVP,

$$\begin{aligned} \dot{x}(t) &= \bar{\lambda}(t) - \hat{\mu}\cdot(x(t) \wedge u(t)), \quad t \geq t_l, \\ x(t_l) &= x^0, \end{aligned} \tag{63}$$

which admits a unique continuous solution. 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 (63).

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 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 classes. This is achieved by setting

$$\psi(x,l) \leftarrow \tilde{\psi}(x,l) + (1 - \tilde{\psi}(x,l)\cdot \mathrm{e})/I. \tag{64}$$

- 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

$$\psi_i(x,l) = \tilde{\psi}_i(x,l)/\tilde{\psi}(x,l)\cdot \mathrm{e}. \tag{65}$$

We conclude this section with two remarks. First, note that TVH ignores the cost parameters and aims to reduce the total queue length of the system through a partial pooling of servers among different classes. Alternatively, one could take into account the priorities by allocating capacity in

the order of the $c\mu$ index, rather than the normalization done in (65). Numerical experiments show that this approach could lead to overallocation to higher priority classes, thereby making "small" systems unstable. One can address this issue by incorporating a safety parameter similar to the DR policy. However, we find that the performance in that case becomes very sensitive to the choice of safety parameters. As we show numerically in Section 8, the TVH described above 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.

7.2. The Time-Varying $c\mu$−Tracking (CMUT) Policy

The main idea behind the CMUT policy is to make allocation decisions in discrete-time with the goal of *tracking* the optimal trajectories known to be optimal for the continuous-time control problem. More formally, 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. Intuitively, one expects the policy to perform well if the system is large and the shift lengths are relatively small compared to the size of the system (or the arrival rate). We note that following a similar proof approach to Theorem 4, the above argument can be formalized by showing that the CMUT policy described below achieves the asymptotic lower bound (43) in Theorem 1 for the finite horizon problem, 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 use the idea to construct a periodic map by tracking the $c\mu$ trajectory under a time-varying (periodic) arrival rate.

Consider the system under a $c\mu$ policy described in Section 5.1 with time varying arrivals. Then, given a vector of initial conditions $x^0\equiv x(t_l)$ the dynamics of the system are governed by the IVP,

$$\begin{aligned}\dot{x}_i(t) &= \bar{\lambda}_i(t)-\mu_i\left(x_i(t)\wedge\left[1-\sum_{j=1}^{i-1}x_j(t)\right]^+\right), \quad i\in\mathcal{I} \text{ and } t\geq t_l, \\ x(t_l) &= x^0,\end{aligned} \tag{66}$$

which has a continuous solution that can be obtained numerically. The CMUT policy, similar to TVH, constructs a fluid-scale periodic allocation map $\varphi(x,l)$ which can be translated for the original system using a similar procedure for the DR policy. Let $b^l(x,u,t)\in\mathbb{R}^I_+$ denote a function that returns the state of the system at $t\in[t_l,t_{l+1})$ starting from x at time t_l (the beginning of shift $l+1$) and under the $c\mu$ allocation. The CMUT policy then sets

$$\varphi_i(x,l)=\left(x_i+\int_{t_l}^{t_{l+1}}\bar{\lambda}_i(s)ds-b_i^l(x,u,t_{l+1})\right)\Big/(\tau\mu), \quad i\in\mathcal{I}. \tag{67}$$

Observe that the allocation for each class is simply the fraction of capacity allocated to each class during a shift of length τ under the $c\mu$ policy.

We note that a naive implementation of the above policy could perform poorly when the shift length is "large". First, 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. Second, since the allocation is obtained assuming a negligible shift length, it could lead to under-allocation of capacity to lower priority classes, thereby causing instability in the long-run. To prevent this issue, we modify the normalized initial state using a safety parameter similar to that in the DR policy. As we demonstrate numerically in Section 8, the *refined* CMUT policy, i.e., after the above modifications, performs very well.

8. Numerical Study

In this section, we evaluate the performance of the proposed policies in terms of the long-run average cost of the system, i.e., $\lim_{T\to\infty} \mathbb{E}\left[C_T^n(U^n)/T\right]$, using steady-state simulation experiments. We begin by demonstrating the performance of the DR and (refined) CMUT policy under stationary arrivals in Section 8.1 where we discuss the performance of the policies in different parameter regimes and evaluate the benefits of partial flexibility by comparing the performance of the policies with that of dedicated staffing. We then turn to the case with non-stationary arrivals in Section 8.2 where we test the performance of the heuristic policies and confirm that our insights continue to hold under non-stationary arrivals. We further report the results of a case study where we evaluate the performance of the heuristics for a simulation model of the system in which we relax certain assumptions of the queueing model and calibrate the inputs using data from our partner hospital.

In order to provide a more realistic estimate of the benefits of partial flexibility, in our simulation experiments we implement a non-preemptive version of the policies. That is, when reassigning a busy server to another pool we assume that the current service must be completed first. We note that the observations made in this section continue to hold under a preemptive implementation of the policies. Further, although our theoretical results are obtained assuming preemption, one would expect the effect of preemption to be negligible in the limit under fluid-scaling, and hence the results to also hold when preemption is not allowed.

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.

8.1. Stationary Arrivals

The experiments correspond to a two-class system, i.e., $\mathcal{I}=\{1,2\}$. Denote by $R_i=\lambda_i/\mu_i$ the offered load of class $i\in\mathcal{I}$. 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* $a\in\{0,1,2,3,4,5\}$. 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.

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 and an exhaustive search.

The results for two set of experiments are summarized in Figures 8 and 9. 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 fixed at $h=(4,2)$. 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 $r=(1/4,3/4)$. The reported cost reduction is the best amongst all safety factors considered (we discuss the effect of the safety factor shortly). In both examples, we observe significant benefits compared to dedicated staffing (ranging between 25% to 53%). While this is 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%, these are substantial savings while adhering to the discrete-time control introduced by the shift constraint.

Note that the percentage of the cost reduction for each policy decreases in the size of the system, shift length and utilization. This can be understood by noting that as the rate of service completion, $n\mu$, or shift-length, τ, increases, the queues reach steady-state faster during the shift. As a result the average costs 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 shifts. Finally, note that the CMUT policy (under the best safety parameter) performs comparable and in some cases even outperforms the DR policy. As we discuss next, however, the CMUT policy is more sensitive to the choice of safety parameter.

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

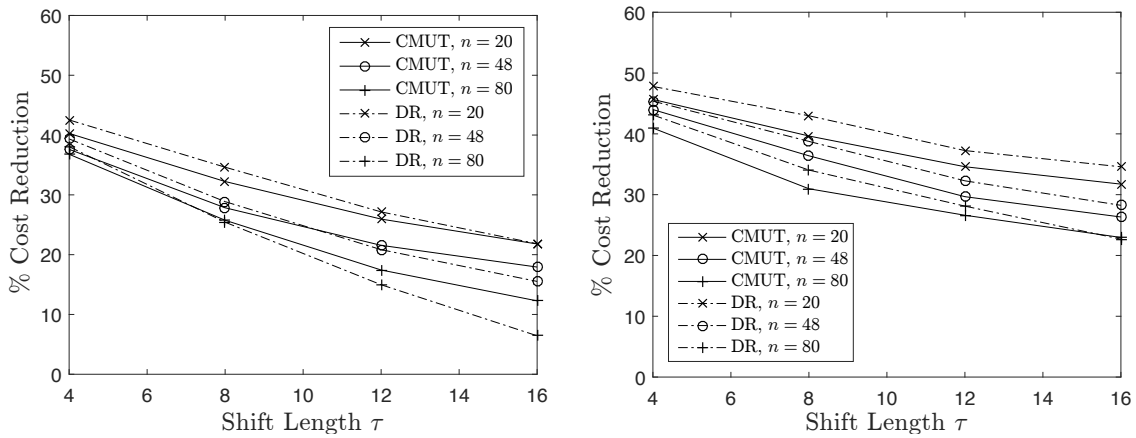

Figure 8 **Performance of the policies in terms of estimated expected long-run average cost for system parameters $\mu = (0.5, 0.5)$, $r = (1/2, 1/2)$, $\rho = 0.92$ (left) and $\rho = 0.96$ (right).**

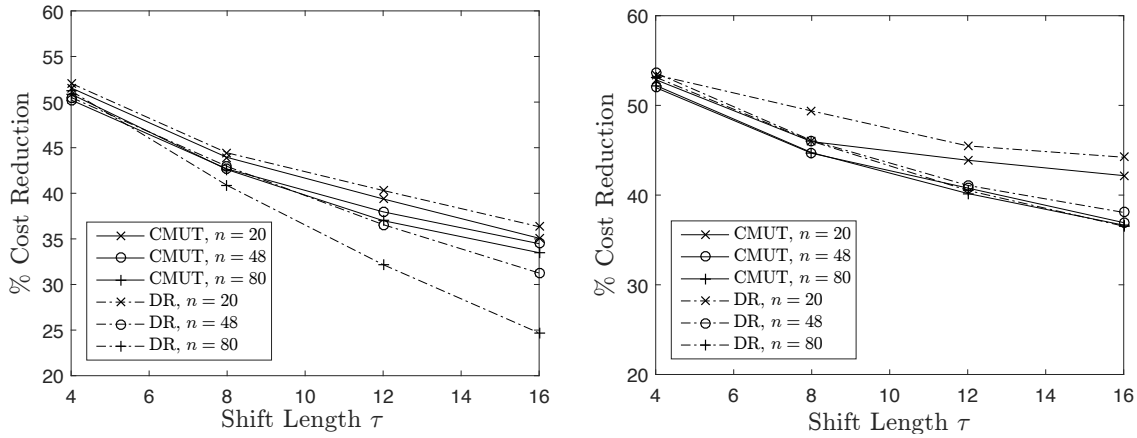

Figure 9 **Performance of the policies in terms of estimated expected long-run average cost for system parameters $\mu = (0.5, 0.5)$, $r = (1/4, 3/4)$, $\rho = 0.92$ (left) and $\rho = 0.96$ (right).**

1. 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 example with $r = (1/2, 1/2)$ and $\rho = 0.96$ in Figure 10. Using a non-zero safety factor improves the performance of both policies, but the 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 and the stochastic fluctuations of $o(n)$ not considered in the fluid control problems.

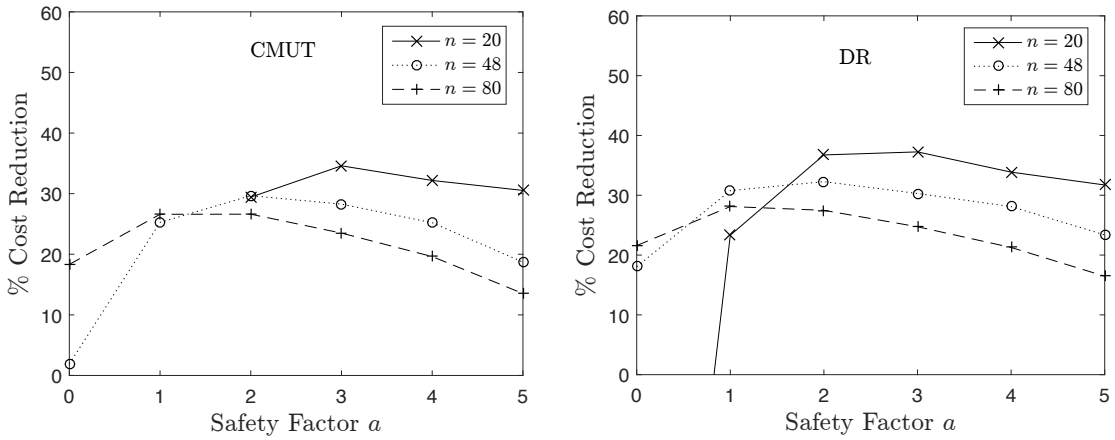

Figure 10 **Performance of the policies in terms of estimated expected long-run average cost for different safety factors and parameters $\mu = (0.5, 0.5)$, $r = (2, 2)$, $\rho = 0.96$, and $\tau = 12$.**

The sub-optimality of the CMUT allocation is due to using average allocations during a shift computed using the solution to Problem 1, which may be far from the optimal allocation when the shift length is not sufficiently small. Further, for small systems under high utilization where the stochastic fluctuations have a significant effect, the predictive power of the fluid models decreases. Since the policies favor the high-priority classes, this could lead to under-assignment of servers to the lower-priority classes over time. If the under-assignment persists long-enough (precisely if the long-run average number of servers allocated to each class is less than the offered load) the system becomes unstable. Using a safety parameter ensures that the under-assignment does not persist.

The sub-optimality of the allocations under the DR policy is due to using a finite number of stages to approximate the stationary policy and can technically be eliminated by solving Problem 2 with sufficiently large number of stages. Of course, this comes at the cost of an increased computation burden. In our experiments, using $N = 6$ is typically sufficient to obtain an stationary policy. The stochastic fluctuations may still lead to worse performance compared to dedicated staffing in very small systems with high utilization as demonstrated above.

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 10.

Expected queue length reduction. Finally, we comment on the expected queue length in steady-state (rather than the expected cost or weighted total queue length) achieved under the policies. For the cost parameters used in the above examples, i.e., $h = (4, 2)$, we observe that the cost reduction is achieved by keeping the lower-priority queue close to that under dedicated staffing while significantly reducing the high-priority queue. We illustrate this in Figure 11 where we compare the

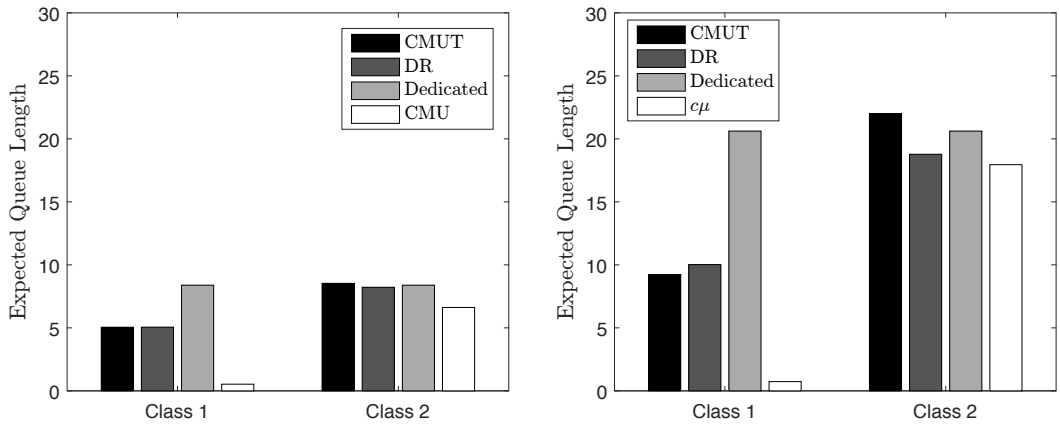

Figure 11 **Estimated (steady-state) expected queue length achieved under different policies for system parameters** $\mu=(0.5,0.5)$**,** $r=(2,2)$**,** $\tau=12$**,** $n=20$**, and** $\rho=0.92$ **(left) and** $\rho=0.96$ **(right).**

expected queue length under different policies. Indeed, by varying the cost parameters one could obtain other expected queue length values.

8.2. Time-Varying Arrivals

We now turn our attention to systems with time-varying arrival rates. Given the added complexity of analyzing such systems, we focus on developing an understanding of the performance of our heuristic policies.

8.2.1. Performance of the Heuristic Policies Focusing on the long-run average cost and queue length, we evaluate the performance of the heuristic policies proposed in Section 7 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.

Figure 12 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 $\rho=0.96$. 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 TVH policies we also plot the performance under the DR 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.

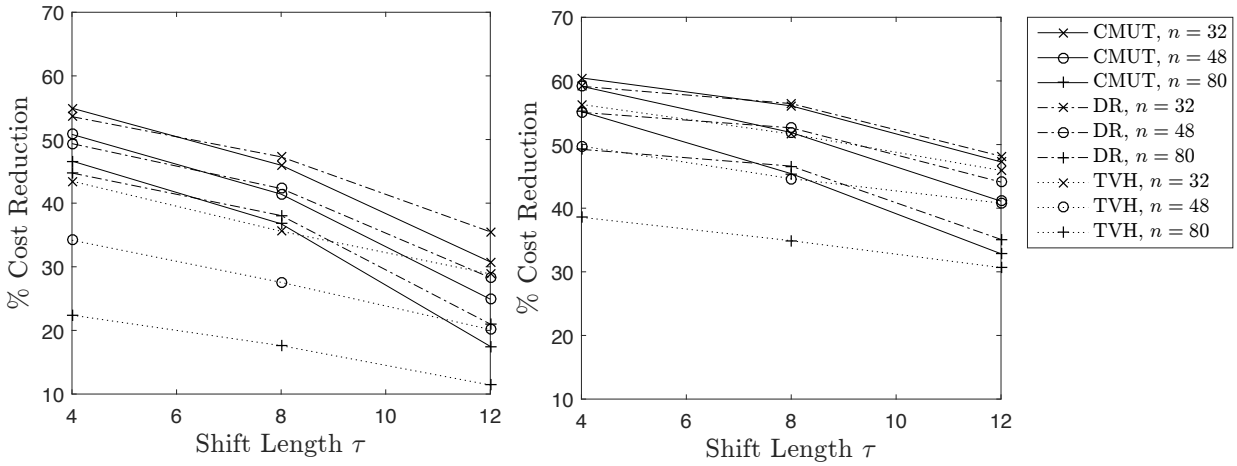

Figure 12 **Performance of the heuristic policies in terms of estimated expected long-run average cost under time-varying arrivals and system parameters $\mu_i = 0.5$ and $r_i = 1/4$ for $i \in \{1,2,3,4\}$, $\tau = 12$, $\rho = 0.92$ (left) and $\rho = 0.96$ (right).**

To further demonstrate the performance of the heuristics, in Tables 2 and 3 we report the detailed output of our simulation experiments for two examples. Table 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 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 TVH 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 TVH policy is equal for all classes. That is, cost reduction 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.

8.2.2. Case Study Finally, we evaluate the performance of the policies in Section 7 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 allow for customer abandonment. We explain the details of the simulation model and its calibration next.

	$n=32$				$n=48$			
Perf. meas.	TVH	CMUT(2)	$c\mu$	Ded.	TVH	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 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 $\pm$ correspond to %95 confidence intervals.)

	$n=32$				$n=48$			
Perf. meas.	TVH	CMUT(1)	$c\mu$	Ded.	TVH	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 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 $\pm$ correspond to %95 confidence intervals.)

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. In addition, 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 we have data for. The parameters of the Log-Normal service times are estimated using the aggregate data for length of stay (start of treatment to discharge) for discharged patients from each area. The abandonment rate for each class is set to 0.5 for all areas, so that the percentage of abandonment is approximately the same as the Left-Without-Being-Seen (LWBS) percentage observed in the data (i.e., $\sim 3\%$) in each area. The shift-length is set to $\tau = 12$ hours with the first shift of the day

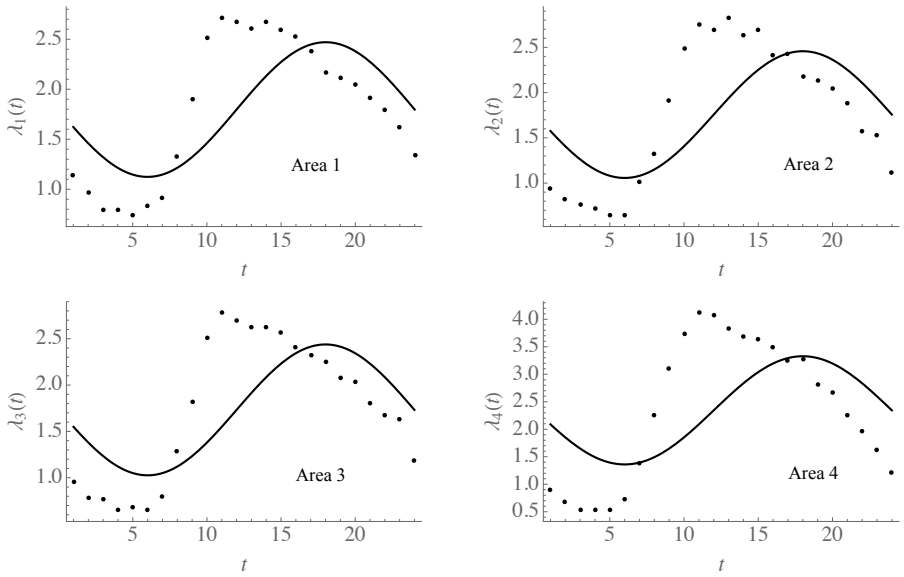

Figure 13 **Empirical hourly arrival rates for each bay and the corresponding Sinusoidal fit.**

starting at 7 am.

In the ED, nurses typically treat multiple patients at a time. We use a simple approximation and take this into account in our simulation model by linearly increasing the number of servers by a given nurse-to-patient ratio $\varrho \in \mathbb{Z}_+$. Further, when assigning servers to different classes, we limit the assignments to groups of servers (of the given ratio) by rounding the allocations obtained from our heuristic policies to multiples of ϱ. We vary $\varrho \in \{1, 2, 4\}$ in our simulation experiments. Accordingly, to choose the number of servers in our simulation model, for each class we pick the smallest multiple of 4 (largest ϱ in our experiments) that makes the class stable under dedicated staffing. That is, recalling that R_i denotes the offered load of class i, we set the number of servers for class i to $4\lceil R_i/4 \rceil$.

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 length-of-stay 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 13 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 4. Note that the system parameters correspond to a regime with a "small" number of service completions during each

		Area 1	Area 2	Area 3	Area 4
Service time	(LogMean, LogSd) (min)	$(5.90, 0.54)$	$(5.85, 0.56)$	$(5.89, 0.54)$	$(4.86, 0.75)$
	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		16	12	12	8

Table 4 Summary of the inputs for the case study.

			1 patient/nurse		4 patients/nurse	
Perf. meas.	Ded.	$c\mu$	TVH	CMUT(1)	TVH	CMUT(1)
$\mathbb{E}[Q_1]$	0.77 ± 0.02	0.12 ± 0.00	2.16 ± 0.02	1.78 ± 0.02	2.45 ± 0.02	2.05 ± 0.02
$\mathbb{E}[Q_2]$	6.08 ± 0.20	0.23 ± 0.00	2.15 ± 0.02	1.74 ± 0.01	2.53 ± 0.03	2.05 ± 0.02
$\mathbb{E}[Q_3]$	6.72 ± 0.15	1.92 ± 0.03	2.11 ± 0.02	1.75 ± 0.03	2.50 ± 0.04	2.04 ± 0.03
$\mathbb{E}[Q_4]$	2.34 ± 0.03	0.96 ± 0.02	2.33 ± 0.03	3.87 ± 0.11	2.88 ± 0.03	5.05 ± 0.12
$\sum_i \mathbb{E}[Q_i]$	15.91 ± 0.25	3.22 ± 0.04	8.75 ± 0.05	9.13 ± 0.12	10.36 ± 0.06	11.19 ± 0.12
$\sum_i h_i \mathbb{E}[Q_i]$	37.09 ± 0.74	$5.93 \pm .10$	21.64 ± 0.15	19.70 ± 0.17	25.29 ± 0.19	23.48 ± 0.25
% Aband.	0.0179 ± 0.0003	0.0036 ± 0.0001	0.0099 ± 0.0001	0.0103 ± 0.0001	0.0116 ± 0.0001	0.0125 ± 0.0002
Ave. dev. from ded.	-	-	5.12 ± 0.04	3.82 ± 0.02	5.42 ± 0.02	4.09 ± 0.04
% Queue red.	-	80%	45%	43%	35%	30%
% cost red.	-	84%	42%	47%	32%	37%

Table 5 Detailed output of the case study simulation experiments.

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. The results of the case study are presented in Table 5. Both policies achieve a significant reduction in terms of long-run average queue length and total cost over dedicated staffing. Specifically, we observe reductions of over 40% in the total queue and cost (based on cost parameters $h = (4, 3, 2, 1)$) compared to dedicated staffing, which is approximately half of that achieved under the (continuous-time, fully-flexible) $c\mu$ policy. 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 30% in the total cost and queue-length when the nurse-to-patient ratio equals 4. The CMUT policy achieves higher benefit in terms of cost reduction compared to TVH, but lower with respect to queue-length-reduction. This is because the TVH policy does not use the cost parameters and only strives to mimic a fully pooled FIFO system (rather than the continuous $c\mu$ policy). 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 equivalent to approximately 1 nurse. Thus, it may be possible to achieve these gains without requiring too many reassignments of nurses.

9. Discussion and Future Research

In this paper, we study the management and benefits of a partially flexible design for service systems that have multiple customer classes and staff who work in shifts. Such a design may be valuable in hospital Emergency Departments and/or knowledge-intensive services where switching from one customer class to another may come with loss of productivity for the servers or increase service times. 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. To investigate its benefits, we consider a queueing model with multiple customer classes and servers. We exploit a fluid asymptotic regime that allows us to obtain effective server assignment policies and further provide 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 8, 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 8 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.

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. 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), there is still work to be done in order to fully characterize 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.

Including customer abandonment. Our numerical experiments suggest that when the percentage of abandonment is small, our proposed policies can still be effective in reducing the expected waiting cost of the system. However, with a considerable abandonment rate one would expect the structure of the optimal assignment policy and the benefits of the partially flexible design to be affected. Future work should investigate the impact of abandonment on the structure of dynamic assignment policies and identify the parameter regime where despite significant abandonment partial flexibility can still be beneficial. It should be noted that in the presence of customer abandonment the optimal scheduling policy under full flexibility can be quite complex – although Atar et al. (2010) shows that for overloaded systems a simple $c\mu$-type policy is asymptotically optimal under fluid scaling.

A general parallel server system. In this work, we consider a parallel server system where all servers are cross-trained and the arrivals to different classes are exogenous. 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.

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.

Appendix. Proofs.

Proof of Proposition 1. To prove the first part we need to show that, for almost all sample paths ω, the sequence $\{\big(\bar{X}^n(\cdot,\omega),\bar{T}^n(\cdot,\omega),\bar{Y}^n(\cdot,\omega),\bar{\Upsilon}^n(\cdot,\omega)\big)\}$ 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 (24) can be shown similarly as in Theorem B.1. of Dai and Tezcan (2011). The result is a consequence of Assumption 2 implying that $\bar{X}^n(0)\to\bar{X}(0)$, the FSLLN for the arrival and service processes, and the fact that $\{\bar{T}^n(\cdot)\}$ is tight, since for all n and $i\in\mathcal{I}$,

$$\bar{T}_i^n(t_2)-\bar{T}_i^n(t_1)\le t_2-t_1,\quad 0\le t_1\le t_2. \tag{68}$$

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

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

as $j\to\infty$ and $\bar{X}(t)$ and $\bar{T}(t)$ satisfy (24). 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 $\bar{U}_i^n(t)\le 1$, we have

$$\bar{Y}_i^n(t_2)-\bar{Y}_i^n(t_1)\le t_2-t_1,\quad 0\le t_1\le t_2. \tag{69}$$

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

$$\bar{Y}^{n_j}(t)\to\bar{Y}(t),\quad u.o.c. \tag{70}$$

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 Tezcan and Dai 2010. 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 $j>j_0$, $n_j^{-1}A_i(\lambda_i^{n_j}t)\le M'\equiv 1+T\bar{\lambda}_i$. Therefore, using (12) and (14),

$$\begin{aligned}\bar{\Upsilon}_i^{n_j}(t_2)-\bar{\Upsilon}_i^{n_j}(t_1) &= \int_0^{t_2}\bar{Q}_i^{n_j}(s)ds-\int_0^{t_1}\bar{Q}_i^{n_j}(s)ds\\ &\le \int_{t_1}^{t_2} n_j^{-1}A_i(\lambda_i^{n_j}s)ds\\ &\le (t_2-t_1)M',\quad 0\le t_1\le t_2\le T,\end{aligned} \tag{71}$$

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 (68), (69) and (71) 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 realtion between $\bar{X}(\cdot)$ and $\bar{T}(\cdot)$ in (24) which was established earlier.

We next turn to establishing (25)–(28). 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 (14)–(15),

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

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

$$\bar{Q}(t) = \bar{X}(t) - \bar{Z}(t) \geq 0, \tag{72}$$

$$\bar{U}(t) - \bar{Z}(t) \geq 0, \tag{73}$$

a.e.. Next, (16) 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 $i \in \mathcal{I}$,

$$\bar{Q}_i(t)(\bar{U}_i(t) - \bar{Z}_i(t)) = 0, \tag{74}$$

holds a.e., which together with (72)–(73) establishes (26). 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 (17) we can conclude that (27) holds a.e.. A similar argument establishes (28) a.e..

Finally, assume that (19) 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 $\{(\bar{U}^{n_j}[0], \ldots, \bar{U}^{n_j}[\bar{k}(T)-1])\}$. Since for each k $\bar{U}^{n_j}[k]$ is uniformly bounded, there exists a further subsequence, say again n_j, such that as $j \to \infty$

$$(\bar{U}^{n_j}[0], \ldots, \bar{U}^{n_j}[\bar{k}(T)-1]) \to (\bar{U}[0], \ldots, \bar{U}[\bar{k}(T)-1]), \tag{75}$$

for some $(\bar{U}[0], \ldots, \bar{U}[\bar{k}(T)-1])$. 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 $i \in \mathcal{I}$,

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

$$\leq \sup_{t\in[0,T]} \sum_{k=0}^{\bar{k}(T)-1} \left|\left(\bar{U}_i^{n_j}[k]-\bar{U}_i[k]\right)\right) \mathbf{1}_{[t_k,t_{k+1})}(t)\Big|$$
$$\leq (\bar{k}(T)-1) \max_{0\leq k\leq(\bar{k}(T)-1)} \left|\left(\bar{U}_i^{n_j}[k]-\bar{U}_i[k]\right)\right|,$$

which vanishes as $j\to\infty$ by (75) 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.

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.,

$$\liminf_{n\to\infty} n^{-1}C_T^n(U^n) = \lim_{j\to\infty} n_j^{-1}C_T^{n_j}(U^{n_j}). \tag{76}$$

Applying Proposition 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 $\left\{\left(\bar{X}^{n_{j'}}(\cdot),\bar{T}^{n_{j'}}(\cdot),\bar{Y}^{n_{j'}}(\cdot),\bar{\Upsilon}^{n_{j'}}(\cdot)\right)\right\}$ converges to $\left(\bar{X}(\cdot),\bar{T}(\cdot),\bar{Y}(\cdot),\bar{\Upsilon}(\cdot)\right)$ u.o.c. with the limiting dynamics satisfying (24)–(28) a.e. as well as (29) and (30) 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 (31–36) starting from $x^0=\bar{X}(0,\omega)$. Further, if $U^n\in\Pi_\tau^n$ then $\bar{U}(t)$ is piecewise-constant and hence $(\bar{X}(t),\bar{U}(t))$ is also admissible for Problem 2 (38–42). Therefore, we have a.s.,

$$h\cdot\bar{\Upsilon}(T) = \int_0^T h\cdot\bar{Q}(s)ds \geq J_T(\bar{X}(0)), \tag{77}$$
$$h\cdot\bar{\Upsilon}(T) = \int_0^T h\cdot\bar{Q}(s)ds \geq J_{T,\tau}(\bar{X}(0)), \text{ if } U^n\in\Pi_\tau^n. \tag{78}$$

Going back to (76) and noting that $n_{j'}$ is a convergent subsequence we have

$$\lim_{j\to\infty} n_j^{-1}C_T^{n_j}(U^{n_j}) = \lim_{j'\to\infty} n_{j'}^{-1}C_T^{n_{j'}}(U^{n_{j'}}) = \lim_{j'\to\infty} h\cdot\bar{\Upsilon}^{n_{j'}}(T) = h\cdot\bar{\Upsilon}(T). \tag{79}$$

It follows from (76) and (77)–(78) that a.s.,

$$\liminf_{n\to\infty} n^{-1}C_T^n(U^n) \geq J_T(\bar{X}(0)), \tag{80}$$
$$\liminf_{n\to\infty} n^{-1}C_T^n(U^n) \geq J_{T,\tau}(\bar{X}(0)), \text{ if } U^n\in\Pi_\tau^n. \tag{81}$$

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

$$J_{T,\tau}(\bar{X}(0)) \geq J_T(\bar{X}(0)). \tag{82}$$

Taking expectation in (80) and (81) and applying Fatou's Lemma we get that

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

establishing (43) and (44). Taking expectation in (82) we have $\mathbb{E}[J_{T,\tau}(\bar{X}(0))]\geq\mathbb{E}[J_T(\bar{X}(0))]$ as claimed in (45). □

Proof of Proposition 2. To prove the claims we apply the "Maximum Theorem" (see Ok 2007, Page 306) to the minimization problem in (55), i.e.,

$$V^k(x) = \min_{u\in\mathcal{U}} \tilde{V}^k(x,u) \equiv g(x,u) + V^{k+1}(f(x,u,\tau)), \tag{83}$$

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 5.2.1 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 $k\le N-2$. It follows from (83) 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. □

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

Lemma 2 *The function $f_i(x,u,t)$ is,*

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

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

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

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,

$$\tilde{f}(x_i,u_i) = \begin{cases} x_i + (\bar{\lambda}_i - \mu_i u_i)t, & (x_i,u_i)\in\mathcal{A}_1,\\ \bar{\rho}_i + e^{-\mu_i(t-\sigma_i)}(u_i-\bar{\rho}_i), & (x_i,u_i)\in\mathcal{A}_2,\\ u_i + (\bar{\lambda}_i - u_i\mu_i)(t-\nu_i), & (x_i,u_i)\in\mathcal{A}_3,\\ \bar{\rho}_i + e^{-\mu_i t}(x_i-\bar{\rho}_i), & (x_i,u_i)\in\mathcal{A}_4, \end{cases} \tag{84}$$

with

$$\begin{aligned} \mathcal{A}_1 &\equiv \{(x_i,u_i); 0\le u_i\le x_i \text{ and } u_i\le (x_i+t\bar{\lambda}_i)/(1+t\mu_i)\},\\ \mathcal{A}_2 &\equiv \{(x_i,u_i); 0\le u_i\le x_i \text{ and } u_i\ge (x_i+t\bar{\lambda}_i)/(1+t\mu_i)\},\\ \mathcal{A}_3 &\equiv \{(x_i,u_i); 0\le x_i\le u_i\le 1 \text{ and } u_i\le \bar{\rho}_i + e^{-\mu_i t}(x_i-\bar{\rho}_i)\},\\ \mathcal{A}_4 &\equiv \{(x_i,u_i); 0\le x_i\le u_i\le 1 \text{ and } u_i\ge \bar{\rho}_i + e^{-\mu_i t}(x_i-\bar{\rho}_i)\}, \end{aligned} \tag{85}$$

obtained by fixing $t>0$ in (50) and substituting $\hat{u}$ and $\check{u}$ from (51) 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 (50)) 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 (84) 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 (84) 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:

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

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

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

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

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, $\nabla\tilde{f}_1^j\geq 0$ and $\nabla\tilde{f}_2^j\leq 0$. This together with continuity and a.e. differentiability of $\tilde{f}$ implies its monotonicity in x_i and u_i as claimed. □

We proceed with the 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 2 (i) we have that $f_i(\cdot,u,t)$ is nondecreasing. It follows directly that $f(\cdot,u,t)$ and using (37) 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 2 $f_i(\cdot,\cdot,\tau)$ 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 (37) 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

$$
\tilde{V}^k(x,u) \equiv g(x,u) + V^{k+1}(f(x,u,\tau)), \tag{86}
$$

is convex. Therefore, we can conclude from (55) and preservation of convexity under partial minimization that $V^k(\cdot)$ is convex. The claim follows by induction. □

Proof of Proposition 3. 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 $l \in \operatorname{argmin}_{i \in \mathcal{I}} u'[j]$. 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 $k \in \{0, \ldots, j-1\}$ and at $k = j$ we have $u[j] > u'[j]$ and $x[j] = x'[j]$. This implies using Lemma 2 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 2 *(ii)* and noting that by construction $u[k] = u'[k]$ for $k \in \{j+1, \ldots, N-1\}$ that $x[k] \leq x'[k]$ and $u[k] = u'[k]$ for $k \in \{j+1, \ldots, N-1\}$. Therefore, since by Lemma 2 and (37) 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. □

Proof of Lemma 1. For each $i \in \mathcal{I}'$ satisfying $x_i^0 > \bar{\rho}_i$ we show using the transition function f_i in (50) 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 (50), $x_i > \bar{\rho}_i$ implies $f_i(x,u,\tau) > \bar{\rho}_i$.

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

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

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 1 $x > \bar{\rho}$ 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 3 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 (86) we have

$$V_N^0(x) = \min_{u \ge 0} \left[g(x,u) + V_N^1(f(x,u)) + \beta(u \cdot \mathrm{e} - 1) \right], \tag{87}$$

where $\beta \ge 0$ is the Lagrangian multiplier. The necessary and sufficient optimality conditions then imply that a minimizer of (87) $\hat{\phi}^0(x)$ satisfies the following for $i = 1,2$,

$$\frac{\partial g(x,\hat{\phi}^0(x))}{\partial u_i} + \frac{\partial V_N^1(f(x,\hat{\phi}^0(x)))}{\partial x_i} \cdot \frac{\partial f_i(x,\hat{\phi}^0(x))}{\partial u_i} + \beta \ge 0, \tag{88}$$

$$\left(\frac{\partial g(x,\hat{\phi}^0(x))}{\partial u_i} + \frac{\partial V_N^1(f(x,\hat{\phi}^0(x)))}{\partial x_i} \cdot \frac{\partial f_i(x,\hat{\phi}^0(x))}{\partial u_i} + \beta \right) \hat{\phi}_i^0(x) = 0, \tag{89}$$

$$\hat{\phi}^0(x) \cdot \mathrm{e} - 1 = 0, \tag{90}$$

$$\hat{\phi}_i^0(x) \ge 0, \tag{91}$$

assuming that the derivatives in (88) and (89) exist on $\{(x,u); x > \bar{\rho}\}$. (Note that the complementary condition in (89) is introduced to handle the nonnegativity constraint $u \ge 0$.) The partials $\partial f_i/\partial u_i$ and $\partial f_i/\partial x_i$ are given by

$$(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial u_i}) f_i = \begin{cases} (1, -\mu_i t), & u_i < x_i \text{ and } u_i \le \hat{u}_i(x), \\ (e^{-\mu_i(t-\sigma_i)}, -\mu_i\sigma_i e^{-\mu_i(t-\sigma_i)}), & u_i < x_i \text{ and } u_i > \hat{u}_i(x), \\ (e^{-\mu_i\nu_i}, -\mu_i(t-\nu_i), & u_i \ge x_i \text{ and } u_i < \check{u}_i(x), \\ (e^{-\mu_i t}, 0), & u_i \ge x_i \text{ and } u_i \ge \check{u}_i(x). \end{cases} \tag{92}$$

Further, using (53) and differentiating (54) we have

$$\frac{\partial g(x,u)}{\partial u_i} = \begin{cases} -h_i(\tau + \mu_i\tau^2/2), & u_i < x_i \text{ and } u_i \le \hat{u}_i(x), \\ -h_i(\sigma_i + \mu_i\sigma_i^2/2), & u_i < x_i \text{ and } u_i > \hat{u}_i(x), \\ -h_i\big((\tau - \nu_i) + \mu_i(\tau-\nu_i)^2/2\big), & u_i \ge x_i \text{ and } u_i < \check{u}_i(x), \\ 0, & u_i \ge x_i \text{ and } u_i \ge \check{u}_i(x), \end{cases} \tag{93}$$

and

$$\frac{\partial g(x,u)}{\partial x_i} = \begin{cases} h_i\tau, & u_i < x_i \text{ and } u_i \le \hat{u}_i(x), \\ h_i\sigma_i, & u_i < x_i \text{ and } u_i > \hat{u}_i(x), \\ h_i(\tau-\nu_i)e^{-\nu_i\mu_i}, & u_i \ge x_i \text{ and } u_i < \check{u}_i(x), \\ 0, & u_i \ge x_i \text{ and } u_i \ge \check{u}_i(x). \end{cases} \tag{94}$$

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 lemma establishes the required differentiability of the value function.

Proposition 4 *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 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 $\omega(y) \le V(y)$ for all $y \in \mathcal{D}$, then V is differentiable at y.*

Proof of Proposition 4. The proof is by induction on k. For $k = N-1$ we have from (55) 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 $\omega(x) \ge V_N^{N-1}(x)$. It follows using Lemma 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 $k \le N-2$. Using (55) 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 1 $f(x, \hat{\phi}^k(x_0)) > \bar{\rho}$. 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 $\omega^k(x) \ge V_N^k(x)$ and again applying Lemma 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

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

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 (88)–(91), 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 (88) and (89) are simply

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

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

$$-\frac{\partial g(x,\hat{\phi}^0(x))}{\partial u_1} = -\frac{\partial g(x,\hat{\phi}^0(x))}{\partial u_2} = \beta, \quad \hat{\phi}_1^0(x) > 0, \hat{\phi}_2^0(x) > 0, \tag{95}$$

$$-\frac{\partial g(x,\hat{\phi}^0(x))}{\partial u_1} = \beta, -\frac{\partial g(x,\hat{\phi}^0(x))}{\partial u_2} < \beta, \quad \hat{\phi}_1^0(x) = 1, \hat{\phi}_2^0(x) = 0, \tag{96}$$

$$-\frac{\partial g(x,\hat{\phi}^0(x))}{\partial u_1} < \beta, -\frac{\partial g(x,\hat{\phi}^0(x))}{\partial u_2} = \beta, \quad \hat{\phi}_1^0(x) = 0, \hat{\phi}_2^0(x) = 1, \tag{97}$$

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 $u_1 \le \bar{\rho}_1$ then trivially $x_1 > \bar{\rho}_1 \ge u_1$, and if $u_1 > \rho_1$ then $u_1 < \hat{u}_1(x) \wedge 1 \le \hat{u}_1(x) = (x_1 + \tau\bar{\lambda}_1)/(1+\tau\mu_1)$ 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 $u_1 \le \hat{u}_1(x)$ and hence we are in Case 1 of (93). Considering the four possible cases for class 2 we observe that

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

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 (95)–(97) it follows that the allocation u can only be optimal if it satisfies (95). 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 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 (93) 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 $N \ge 2$, i.e., there exists an allocation $\hat{\phi}^0(x) \in \widehat{\mathcal{U}}(x)$ that satisfies the optimality equations (88)–(90) 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):

$$\frac{\partial V_{N-1}^0(x)}{\partial x_i} = \frac{\partial g(x,\hat{\phi}^0(x))}{\partial x_i} + \frac{\partial f_i(x,\hat{\phi}^0(x))}{\partial x_i} \cdot \frac{\partial V_{N-1}^1(f(x,\hat{\phi}^0(x))}{\partial x_i}, \quad i = 1,2, \tag{98}$$

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 (88)–(91). 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

$$-\frac{\partial g(x,u)}{\partial u_1} - \frac{\partial V_N^1(f(x,u))}{\partial x_1} \cdot \frac{\partial f_1(x,u)}{\partial u_1} \geq -\frac{\partial g(x,u)}{\partial u_2} - \frac{\partial V_N^1(f(x,u))}{\partial x_2} \cdot \frac{\partial f_2(x,u)}{\partial u_2}. \tag{99}$$

If the above inequality is strict then it follows directly that u does not satisfy the optimality conditions (88)–(90). In the case where (99) 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 (99). First, similar to the single-stage problem, $u \notin \widehat{\mathcal{U}}(x)$ implies that in the first stage we have

$$-\frac{\partial g(x,u)}{\partial u_1} = h_1(\tau + \mu_1 \tau^2/2), \quad -\frac{\partial g(x,u)}{\partial u_2} \leq h_2(\tau + \mu_2 \tau^2/2). \tag{100}$$

Further, from (92) at $t = \tau$ and with $u \notin \widehat{\mathcal{U}}(x)$ we have

$$-\frac{\partial f_1(x,u)}{\partial u_1} = \mu_1 \tau, \quad -\frac{\partial f_2(x,u)}{\partial u_2} \leq \mu_2 \tau, \tag{101}$$

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 (99). 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 1 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

$$-\frac{\partial g(\tilde{x}, \hat{\phi}^1(\tilde{x}))}{\partial u_1} - \frac{\partial V_{N-1}^0(f(\tilde{x}, \hat{\phi}^1(\tilde{x})))}{\partial x_1} \cdot \frac{\partial f_1(\tilde{x}, \hat{\phi}^1(\tilde{x}))}{\partial u_1} = \beta, \tag{102}$$

$$-\frac{\partial g(\tilde{x}, \hat{\phi}^1(\tilde{x}))}{\partial u_2} - \frac{\partial V_{N-1}^0(f(\tilde{x}, \hat{\phi}^1(\tilde{x})))}{\partial x_2} \cdot \frac{\partial f_2(\tilde{x}, \hat{\phi}^1(\tilde{x}))}{\partial u_2} \leq \beta, \tag{103}$$

where (103) 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 (92)) 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 (98) 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 $\partial f_2(\tilde{x}, \hat{\phi}^1(\tilde{x}))/\partial u_2 \neq 0$,

$$\frac{\partial V_N^1(\tilde{x})}{\partial x_1} = \frac{\partial g(\tilde{x}, \hat{\phi}^1(\tilde{x}))}{\partial x_1} + \left(-\beta - \frac{\partial g(\tilde{x}, \hat{\phi}^1(\tilde{x}))}{\partial u_1}\right) \left(\frac{\partial f_1(\tilde{x}, \hat{\phi}^1(\tilde{x}))}{\partial x_1} \Bigg/ \frac{\partial f_1(\tilde{x}, \hat{\phi}^1(\tilde{x}))}{\partial u_1}\right), \tag{104}$$

$$\frac{\partial V_N^1(\tilde{x})}{\partial x_2} \le \frac{\partial g(\tilde{x},\hat{\phi}^1(\tilde{x}))}{\partial x_2} + \left(-\beta - \frac{\partial g(\tilde{x},\hat{\phi}^1(\tilde{x}))}{\partial u_2}\right)\left(\frac{\partial f_2(\tilde{x},\hat{\phi}^1(\tilde{x}))}{\partial x_2} \Big/ \frac{\partial f_2(\tilde{x},\hat{\phi}^1(\tilde{x}))}{\partial u_2}\right), \tag{105}$$

where we note again that the inequality in (105) 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 (104) and (105) 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 (93) and (94) that

$$\frac{\partial g(\tilde{x},\hat{\phi}^1(\tilde{x}))}{\partial x_1} = h_1\theta, \quad -\frac{\partial g(\tilde{x},\hat{\phi}^1(\tilde{x}))}{\partial u_1} = h_1(\theta+\mu_1\theta^2/2), \tag{106}$$

and from (92) at $t=\tau$ that

$$\frac{\partial f_1(\tilde{x},\hat{\phi}^1(\tilde{x}))}{\partial x_1} = e^{-\mu_1(\tau-\theta)}, \quad -\frac{\partial f_1(\tilde{x},\hat{\phi}^1(\tilde{x}))}{\partial u_1} = \mu_1\theta e^{-\mu_1(\tau-\theta)}, \tag{107}$$

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 (104) as

$$\frac{\partial V_N^1(\tilde{x})}{\partial x_1} = h_1\theta + \big(\beta - h_1(\theta+\mu_1\theta^2/2)\big)\big(1/(\mu_1\theta)\big). \tag{108}$$

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 1). 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 (93), (94), and (92) (at $t=\tau$) in (105) 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 (108) at $\theta=\tau$ yields

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

which by assumption (ii) implies that (99) is satisfied with a strict inequality.

2. *$\hat{\phi}^1(\tilde{x})$ is an interior solution and $\hat{\phi}_2^1(\tilde{x}) < \tilde{x}_2$ and $\hat{\phi}_2^1(\tilde{x}) \le \hat{u}_2$:* We are in Case 1 where class 2 starts with a positive queue at stage 2 that remans positive during the stage. Substituting the corresponding terms from (93), (94), and (92) (at $t=\tau$) in (105) 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 (108) yields

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

We need to consider the two cases for θ separately.

(a) We first show that when $\theta = \sigma_1 < \tau$ (99) holds with a strict inequality. After simplifying, this is equivalent to showing

$$\beta\big((\tau/\sigma_1) - 1\big) + h_1\mu_1\tau(\tau + \sigma_1)/2 - h_2\mu_2\tau^2 > 0. \tag{109}$$

From the optimality equation (103) we know that $\beta \geq -\partial g(\tilde{x}, \hat{\phi}^1(\tilde{x}))/\partial u_2 = h_2(\tau + \mu_2\tau^2/2)$ and hence the LHS of (109) is no less than

$$\xi(\sigma_1) \equiv h_2(\tau + \mu_2\tau^2/2)\big((\tau/\sigma_1) - 1\big) + h_1\mu_1\tau(\tau + \sigma_1)/2 - h_2\mu_2\tau^2. \tag{110}$$

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 $\sigma_1 \geq \tau(h_2\mu_2)/(h_1\mu_1)$. Then it is easy to see that $\xi(\sigma_1) > h_1\mu_1\tau(\tau + \sigma_1)/2 - h_2\mu_2\tau^2 \geq 0$. 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

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

Therefore, if assumption (ii) holds with an inequality, i.e., $h_1\mu_1 > h_2\mu_2$, (99) also holds with an inequality and we are done. However, if $h_1\mu_1 = h_2\mu_2$ and also (100) and (101) are both satisfied with equality (i.e., we are in Case 1 for class 2 in the first stage) then (99) 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 $\hat{u}_2 < \hat{\phi}_2^1(\tilde{x}) < \tilde{x}_2$*: 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,

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

yielding

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

and thus

$$-\frac{\partial g(x,u)}{\partial u_1} - \frac{\partial V_N^1(f(x,u))}{\partial x_1} \cdot \frac{\partial f_1(x,u)}{\partial u_1} = h_1(\tau + \mu_1\tau^2/2) + h_1\mu_1\tau\theta, \tag{111}$$

$$-\frac{\partial g(x,u)}{\partial u_2} - \frac{\partial V_N^1(f(x,u))}{\partial x_2} \cdot \frac{\partial f_2(x,u)}{\partial u_2} \leq h_2(\tau + \mu_2\tau^2/2) + h_2\mu_2\tau\sigma_2. \tag{112}$$

It follows by assumptions (ii) and (iii) that if $\theta = \tau$ (99) 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 (102)–(103) for $\hat{\phi}^1(\tilde{x})$ reduce to

$$-h_1(\sigma_1 + \mu_1\sigma_1^2/2) = -h_2(\sigma_2 + \mu_2\sigma_2^2/2) = \beta, \tag{113}$$

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 (99) holds when $\theta = \sigma_1$, i.e., using (111) and (112),

$$h_1(\tau + \mu_1\tau^2/2) + h_1\mu_1\tau\sigma_1 \geq h_2(\tau + \mu_2\tau^2/2) + h_2\mu_2\tau\sigma_2. \tag{114}$$

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 (113) is $\sigma_1 + \mu_1\sigma_1^2/2 = z(\sigma_2 + \mu_2\sigma_2^2/2)$ which can be written as

$$\mu_1\sigma_1 = -1 + \sqrt{1 + 2\mu_1 z\sigma_2 + \mu_1\mu_2 z\sigma_2^2}. \tag{115}$$

Finally, re-writing (114), we want to show

$$\tau + \mu_1\tau^2/2 + \mu_1\tau\sigma_1 \geq z(\tau + \mu_2\tau^2/2 + \mu_2\tau\sigma_2). \tag{116}$$

First assume that $z < 1$. Then we can use (115) 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 (115) 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, (99) can only hold with equality if $h_1=h_2$ and $\mu_1=\mu_2$ and also both (101) and (100) hold with equality (i.e., we are in Case 1 for class 2 in the first stage and (112) holds with equality). This further implies using (115) 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,

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

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. □

Proof of Thereom 4. We need the following basic lemma before presenting the proof.

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

$$\limsup_{n\to\infty}\mathbb{E}\left[n^{-1}C^n_T(U^n)\right]\leq\mathbb{E}\left[\limsup_{n\to\infty}n^{-1}C^n_T(U^n)\right]. \tag{117}$$

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,

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

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 $\{n^{-1}A_i(\lambda^n_i T)\}$ is uniformly integrable for each $i\in\mathcal{I}$. It follows that $\{n^{-1}C^n_T(U^n)\}$ is also uniformly integrable. Next, we show (117) by proving that it holds for any uniformly integrable sequence $\{B^n\}$. To this end observe that,

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

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

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 4 $\{n^{-1}C_T^n(U_*^n)\}$ is uniformly integrable the lim sup is finite. By Proposition 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 (24)–(28) a.e. It follows that a.s.,

$$\begin{aligned}\limsup_{n\to\infty} n^{-1}C_T^n(U_*^n) &= \lim_{j'\to\infty} n_{j'}^{-1}C_T^{n_{j'}}(U_*^{n_{j'}})\\ &= h\cdot\bar{\Upsilon}_*(T)=\int_0^T h\cdot\bar{Q}_*(s)ds=\int_0^T h\cdot\big(\bar{X}_*(s)-\bar{U}_*(s)\big)^+ds.\end{aligned}\tag{118}$$

Next, note using (24) and (26) that $\bar{X}_*$ and $\bar{U}_*$ satisfy

$$\dot{\bar{X}}_*(t)=\bar{\lambda}-\hat{\mu}\cdot(\bar{X}_*(t)\wedge\bar{U}_*(t)).\tag{119}$$

Further, since $U_*^n\in\Pi_\tau^n$ for all n, (29) and (30) hold, that is $\bar{U}_*$ is piecewise-constant and

$$\bar{U}_*(t)=\sum_{k=0}^{\bar{k}(T)-1}\bar{U}_*[k]\mathbf{1}_{[t_k,t_{k+1})}(t),\tag{120}$$

for some $\{\bar{U}_*[k];0\le k\le\bar{k}(T)-1\}$ with $t_{k+1}-t_k=\tau$. This, together with (27) and (28), 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 (118) 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 4 to get the final result in (62).

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 (60) in (61) we can write the scaled allocation at time $k=0$ as

$$\begin{aligned}\bar{U}_*^{n_{j'}}[0]\equiv n_{j'}^{-1}U_*^{n_{j'}}[0] &= n_{j'}^{-1}\left\lfloor n_{j'}\tilde{U}_*^{n_{j'}}[0]\right\rfloor\\ &= n_{j'}^{-1}\left\lfloor n_{j'}\hat{\phi}^0\left((\bar{X}_*^{n_{j'}}(0)-\bar{\beta}^{n_{j'}})^+\right)\right\rfloor\\ &= \hat{\phi}^0\left((\bar{X}_*^{n_{j'}}(0)-\bar{\beta}^{n_{j'}})^+\right)+n_{j'}^{-1}\varepsilon,\end{aligned}\tag{121}$$

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 (121) we claim that

$$\bar{U}_*[0]=\lim_{j'\to\infty}\bar{U}_*^{n_{j'}}[0]=\hat{\phi}^0\big(\bar{X}(0,\omega)\big)\in\phi^0(\bar{X}(0,\omega)),\tag{122}$$

i.e, an optimal allocation at $k=0$ starting from $\bar{X}(0,\omega)$. Recall that $\hat{\phi}^k(x)$ denotes an arbitrary member of the correspondence $\phi^k(x)$. Since, by Proposition 2, $\phi^k(\cdot)$ is upper hemicontinuous and closed-valued, it also satisfies the closed-graph property (Ok 2007, Proposition 3). The closed-graph property of the correspondence $\phi^k(x)$ 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). Now, consider the sequence in (121) 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 $j'\to\infty$. Further, 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'. Noting that $\{n_{j'}; j' \in \mathbb{N}\}$ is a convergent subsequence, the closed-graph property of $\phi^k(x)$ establishes the claim in (122).

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

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

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 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 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. □

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.

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

Armony M (2005) Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems* 51(3):287–329.

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.

Ataman B, Kumar S (2005) Heavy traffic analysis of open processing networks with complete resource pooling: Asymptotic optimality of discrete review policies. *Ann. Appl. Probab.* 15(1A):331–391.

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, Staats BR, Patterson BW (2017) The effect of discrete work shifts on a nonterminating service system .

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

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.

Bertsekas DP (2005) *Dynamic Programming and Optimal Control* (Athena Scientific).

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.

Caudillo-Fuentes LA, Kaufman DL, Lewis ME (2010) A simple heuristic for load balancing in parallel processing networks with highly variable service time distributions. *Queueing Systems* 64(2):145–165.

Chan CW, Dong J, Green LV (2016) Queues with time-varying arrivals and inspections with applications to hospital discharge policies. *Operations Research* 65(2):469–495.

Chen H, Mandelbaum A (1991) Discrete flow networks: bottleneck analysis and fluid approximations. *Mathematics of operations research* 16(2):408–446.

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

Dai J, Shi P (2017a) Inpatient bed overflow: An approximate dynamic programming approach. Technical report, Working Paper, Cornell University, Ithaca, New York.

Dai J, Shi P (2017b) A two-time-scale approach to time-varying queues in hospital inpatient flow management. *Operations Research* 65(2):514–536.

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.

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

Down DG, Lewis ME (2006) Dynamic load balancing in parallel queueing systems: Stability and optimal control. *European Journal of Operational Research* 168(2):509–519.

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, Armony M, Mandelbaum A (2008) Service-level differentiation in call centers with fully flexible servers. *Management Science* 54(2):279–294.

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, López MJ (1999) Heavy traffic resource pooling in parallel-server systems. *Queueing systems* 33(4):339–368.

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.

Maglaras C (1999) Dynamic scheduling in multiclass queueing networks: Stability under discrete-review policies. *Queueing Systems* 31(3-4):171–206.

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, Momčilović P, Tseytlin Y (2012) On fair routing from emergency departments to hospital wards: Qed queues with heterogeneous servers. *Management Science* 58(7):1273–1291.

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.

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

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.

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

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.

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

Tezcan T, Dai J (2010) Dynamic control of n-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. *Operations Research* 58(1):94–110.

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.

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

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 mt/mt/s queues is asymptotically correct as the rates increase. *Management Science* 37(3):307–314.

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

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

Yildiz O, Kamali MF, Tezcan T (2015) Analysis of triage systems in emergency departments. Technical report, Working Paper, University of Rochester - Simon Business School, Rochester, New York.

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.