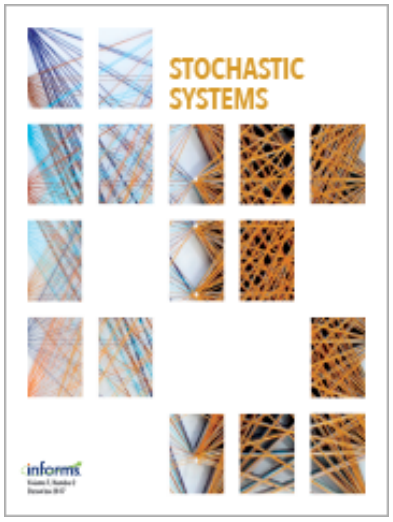
This article was downloaded by: [216.73.217.141] On: 17 November 2026, At: 12:07
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA



Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Construction of Asymptotically Optimal Control for Crisscross Network from a Free Boundary Problem

Amarjit Budhiraja, Xin Liu, Subhamay Saha

To cite this article:
Amarjit Budhiraja, Xin Liu, Subhamay Saha (2016) Construction of Asymptotically Optimal Control for Crisscross Network from a Free Boundary Problem. Stochastic Systems 6(2):459-518. https://doi.org/10.1287/15-SSY211

This work is licensed under a Creative Commons Attribution 4.0 International License. You are free to copy, distribute, transmit and adapt this work, but you must attribute this work as "*Stochastic Systems*. Copyright 2016 The Author(s). https://doi.org/10.1287/15-SSY211, used under a Creative Commons Attribution License: http://creativecommons.org/licenses/by/4.0/."

Copyright © 2016, The author(s)

Please scroll down for article—it is on subsequent pages

With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

Stochastic Systems
2016, Vol. 6, No. 2, 459–518
DOI: 10.1214/15-SSY211

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

CONSTRUCTION OF ASYMPTOTICALLY OPTIMAL CONTROL FOR CRISSCROSS NETWORK FROM A FREE BOUNDARY PROBLEM*

By Amarjit Budhiraja[†], Xin Liu[‡] and Subhamay Saha[§]

University of North Carolina at Chapel Hill[†], Clemson University[‡], and Indian Institute of Technology Guwahati[§]

An asymptotic framework for optimal control of multiclass stochastic processing networks, using formal diffusion approximations under suitable temporal and spatial scaling, by Brownian control problems (BCP) and their equivalent workload formulations (EWF), has been developed by Harrison (1988). This framework has been implemented in many works for constructing asymptotically optimal control policies for a broad range of stochastic network models. To date all asymptotic optimality results for such networks correspond to settings where the solution of the EWF is a reflected Brownian motion in $\mathbb{R}_+$ or a wedge in $\mathbb{R}^2_+$. In this work we consider a well studied stochastic network which is perhaps the simplest example of a model with more than one dimensional workload process. In the regime considered here, the singular control problem corresponding to the EWF does not have a simple form explicit solution. However, by considering an associated free boundary problem one can give a representation for an optimal controlled process as a two dimensional reflected Brownian motion in a Lipschitz domain whose boundary is determined by the solution of the free boundary problem. Using the form of the optimal solution we propose a sequence of control policies, given in terms of suitable thresholds, for the scaled stochastic network control problems and prove that this sequence of policies is asymptotically optimal. As suggested by the solution of the EWF, the policy we propose requires a server to idle under certain conditions which are specified in terms of thresholds determined from the free boundary.

Received November 2015.

*This work has been supported in part by the National Science Foundation (DMS-1016441, DMS-1305120), the Army Research Office (W911NF-10-1-0158, W911NF-14-1-0331) and DARPA(W911NF-15-2-0122).

MSC 2010 subject classifications: Primary 60K25, 68M20, 90B36; secondary 60J70.

Keywords and phrases: Stochastic networks, crisscross networks, dynamic control, heavy traffic, diffusion approximations, Brownian control problems, singular control problems, reflected Brownian motions, free boundary problems, threshold policies, large deviations.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

1. Introduction. Stochastic processing networks arise commonly in manufacturing, computer and communication systems. Optimal regulation is a key objective in the design of such networks. Construction and implementation of an optimal control can in general be challenging due to complex dynamics, lack of simple Markovian state descriptors, and high frequency and throughput characteristics. An approach pioneered by Harrison [13] is to approximate the control problems for such complex networks, when the system is in *heavy traffic*, through certain control problems for Brownian motions. These Brownian control problems (BCP) are quite non-standard in that the control processes may not even have bounded variation sample paths. A key result of Harrison and van Mieghem [16] says that in quite general settings there are *equivalent workload formulations* (EWF) of such BCP which correspond to more tractable control problems. In the EWF, controls are bounded variation processes and thus these problems fall within the classical framework of singular stochastic control, although here one has the additional feature of state constraints in non-smooth domains (typically the state space is a convex polyhedral cone). Furthermore, in many examples the EWF is of much lower dimension than the original BCP, thus providing significant model simplification. In recent years there have been many works that have developed asymptotically optimal control policies for a range of network models by analyzing the solutions of the associated BCP and EWF [1, 2, 9, 11, 14, 18]. Specifically, these works consider a sequence of control problems, indexed by a parameter n, for the underlying network such that as n becomes large the traffic intensity approaches criticality. An asymptotically optimal control is a sequence of policies $\{T_n^*\}$ such that the (scaled) cost when T_n^* is used in the n-th network is asymptotically the lowest that is achievable among all such control sequences. One simplifying feature of all the models in the above papers is that the associated EWF is a one dimensional singular control problem in $\mathbb{R}_+$ with a monotonic cost function. Such control problems have a simple solution given through a one dimensional reflected Brownian motion and this explicit form plays a key role in the proofs. Papers that develop asymptotically optimal control policies for settings with more than one dimensional EWF include [4, 24].

In [4], a well known queuing system with three buffers and two stations (see Figure 1), in heavy traffic, for which the associated EWF is a two dimensional singular control problem has been analyzed. This is perhaps the simplest non-trivial model with more than one-dimensional workload process. The model, referred to in the literature as the *crisscross network*, has been previously analyzed in [17, 21, 22, 28]. The network is of interest in its own right, but its analysis also gives insight for large networks with bottle-

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

neck sub-systems that have similar features as the crisscross network. A brief description of the crisscross network is as follows – there are 3 classes of customers (corresponding to 3 buffers) and 2 servers; for $k = 1, 2$, customers of Class k arrive according to a renewal process and receive service at Station 1; Class 1 customers leave the system once their service is completed; Class 2 customers after being processed at Station 1 proceeds to Station 2, where they are re-designated as Class 3 customers and get processed at Station 2 after which they leave the system. Precise descriptions of the control problem and the cost criterion are given in Sections 2.2 and 3, respectively. The form of an optimal control for this network depends on the underlying parameters, in particular on the strictly positive holding cost vector (c_1, c_2, c_3) and the (asymptotic) service rate vector (μ_1, μ_2, μ_3), and in general can be quite complex. Indeed the paper [22] discusses two distinct regimes where the structure of optimal control policies are expected to be quite different. These regimes are as follows. Case I: $c_1\mu_1 - c_2\mu_2 + c_3\mu_2 \leq 0$, and Case II: $c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0$. The paper [22] further differentiates Case II into four sub-cases: Case IIA: $c_2\mu_2 - c_3\mu_2 \geq 0,\ c_2\mu_2 - c_1\mu_1 \geq 0$; Case IIB: $c_2\mu_2 - c_3\mu_2 < 0,\ c_2\mu_2 - c_1\mu_1 \geq 0$; Case IIC: $c_2\mu_2 - c_3\mu_2 \geq 0,\ c_2\mu_2 - c_1\mu_1 < 0$; Case IID: $c_2\mu_2 - c_3\mu_2 < 0,\ c_2\mu_2 - c_1\mu_1 < 0$.

Case I is the simplest to analyze and it has been shown in [28] that a simple priority policy (server 1 always gives priority to Buffer 2) is (asymptotically) optimal. Asymptotically optimal control policies for Case IIA have been constructed in [4, 22]. In this case using certain monotonicity properties one can show that the optimal state process for the singular control problem is a two dimensional reflected Brownian motion in $\mathbb{R}^2_+$ with normal reflections. Furthermore, using results of [16] an explicit solution of the BCP can be given as well. The proposed policies and the proof of asymptotic optimality in [22] and [4] are quite different – the first paper uses quite technical machinery from viscosity solution analysis of Hamilton-Jacobi-Bellman(HJB) equations whereas the paper [4] proceeds by constructing a multiple threshold policy for which the associated state process closely mimics the solution of the BCP.

Cases IIB, IIC and IID, to date have remained unsolved. One of the key obstacles in their analysis has been that in these regimes the singular control problem does not admit a simple form solution. Indeed, although Harrison's framework has provided asymptotically optimal control policies for a broad range of models, all of the available results correspond to settings where the EWF has an optimal pathwise simple form solution given through a reflected Brownian motion in $\mathbb{R}_+$ or a wedge in $\mathbb{R}^2_+$. In [7] the authors studied the singular control problem that corresponds to the EWF for Cases IIB and

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

IIC. Typically, solutions of singular control problems are given in terms of an open set $\mathcal{O}$ in the state space, referred to as the *continuation region*, such that starting within $\mathcal{O}$ no control is applied until the boundary of $\mathcal{O}$ is reached; if the initial condition is in $(\bar{\mathcal{O}})^c$ (*action region*), an instantaneous control in a pre-specified direction is applied to bring the state to $\partial\mathcal{O}$; and once in $\bar{\mathcal{O}}$, the state process is constrained in the set by suitable reflection at $\partial\mathcal{O}$. In terms of the associated HJB equation, in $\mathcal{O}$ the value function satisfies a linear elliptic PDE while in $\mathcal{O}^c$ a nonlinear first order PDE is satisfied; the boundary $\partial\mathcal{O}$ separating these two regions is referred to as the *free boundary* for the system of PDE and determining this boundary is called a *free boundary problem*. In [7], it was shown that the value function J^* of the control problem is $\mathcal{C}^1$ up to the boundary and the continuation region for the optimal control is given as $G = \{x \in \mathbb{R}^2_+ : x_1 \geq \Psi(x_2)\}$ where $\Psi : \mathbb{R}_+ \to \mathbb{R}_+$ is a Lipschitz non-decreasing function given as

$$\Psi(w_2) \doteq \sup\{w_1 \geq 0 : \partial_{w_1} J^*(w_1, w_2) = 0\}\,. \tag{1.1}$$

One novel feature of this result is that here the principle of smooth fit ideas that have been used in previous works [3, 15, 25, 27] are not applicable and in fact C^2-regularity of the value functions – a crucial ingredient in these works – is not available. The paper [7] constructs an optimal controlled process as a reflected Brownian motion in G reflected at $\partial G = \partial_1 G \cup \partial_2 G$, where the direction of reflection is $e_2 = (0,1)'$ on $\partial_2 G = \{x \in \mathbb{R}^2_+ : x \cdot e_2 = 0\}$ and is $e_1 = (1,0)'$ on $\partial_1 G = \{x \in \mathbb{R}^2_+ : \Psi(x_2) = x_1\}$. The solution is not altogether explicit since in order to determine Ψ one need the value function of the control problem, however numerical methods for computing the free boundary Ψ are available [6, 19, 20].

The goal of this work is to use the solution of the free boundary problem from [7] (given by the function Ψ) to obtain an asymptotically optimal control policy for the corresponding crisscross network in regime IIB and IIC. We only study Case IIB here since treatment of the other case is expected to be quite similar (See Remark 3.6). Recall that Case IIB corresponds to $c_2\mu_2 - c_3\mu_2 < 0$, $c_2\mu_2 - c_1\mu_1 \geq 0$ (note that it implies $c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0$). In this case serving Class 1 customers reduces cost of jobs in Buffer 1 at an (asymptotic) average rate of $c_1\mu_1$. Also serving Class 2 customers reduces cost of jobs in Buffer 2 at an (asymptotic) average rate of $c_2\mu_2$ and at the same time increases cost at an (asymptotic) average rate of $c_3\mu_2$ for Buffer 3. The condition $c_1\mu_1 > c_2\mu_2 - c_3\mu_2$ implies that overall cost is reduced at a higher rate if Server 1 processes Class 1 customers instead of Class 2 customers. The second condition $c_2\mu_2 < c_3\mu_2$ says that it is cheaper to keep jobs in Buffer 2 than in Buffer 3. The third condition $c_2\mu_2 \geq c_1\mu_1$ means

the cost from the queues processed by Server 1 is reduced more rapidly if jobs in Buffer 2 are processed. The first condition suggests that a priority policy that favors Class 1 customers should be used. However, the third condition says that the minimization of immediate workload at Server 1 will be achieved by processing Class 2 customers. Also, always giving high priority to Class 1 customers may lead to an undesirable underutilization of Server 2. Thus an optimal policy needs to suitably balance these opposing considerations. Additionally, even if there are jobs in the second queue (but say no jobs in Queue 1) it may be preferable for Server 1 to idle since holding costs in Queue 3 are higher than that in Queue 2. Thus an optimal control is not expected to be a non-idling policy. In Section 3 we describe our proposed policy that suitably takes into account the various complex features of this parameter regime. Furthermore (as in [4]) the policy is designed so that the associated state process closely mimics the solution of the BCP given in Section 4. The policy we propose will require Server 1 to idle under certain circumstances which are specified in terms of a threshold determined from the free boundary Ψ. We refer the reader to Remark 3.4 and Subsection 4.1 for remarks on the form of the control policy, the role that the free boundary plays, and the key technical steps in the proof.

For asymptotic optimality we will assume that the inter-arrival and service times have finite moment generating functions in a neighborhood of 0. These conditions allow the use of certain large deviation estimates that are key in the proof of asymptotic optimality. Such large deviation techniques for obtaining asymptotically optimal control policies for stochastic networks were first introduced by Bell and Williams [1] and later also used in [2, 4, 14]. In this work, the large deviation estimates are crucially used in the proofs of Theorems 5.4 and 5.5 which together with Lemma 5.7 are perhaps the most technical parts of the paper. The statement of these two results are in the same vein as Theorems 4.8 and 4.9 in [4], however these latter results assume that the interarrival and service times are exponential and thus one can appeal to classical sample path large deviation estimates for Poisson processes. The proofs in the general distribution case treated here are substantially more technical since suitable sample path large deviation estimates for such a setting are not readily available and we instead need to rely on fixed time large deviation estimates (as in Lemma 6.1) for renewal processes. The other new challenge in this work is in arguing that the scaled workload process lies in G asymptotically as $n \to \infty$. The proof of this result given in Lemma 5.7 relies crucially on the properties of the free boundary established in [7] (see Lemma 4.4).

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Our main result is Theorem 3.7 which under Assumptions 2.1, 2.2, 2.3 and 3.1 proves the asymptotic optimality of the control policy in Definition 3.2 with a suitable choice of threshold parameters c, l_0, g_0. This result treats Case IIB but, as noted earlier, Case IIC can be treated in a similar fashion. Treatment of Case IID is a challenging open problem. For this case even the solutions of the associated EWF and the BCP are currently unavailable. We have the following conjecture(see [22] for a closely related statement) for the form of the solution to the EWF, which if resolved will be a key step forward in the construction of asymptotically optimal control policies in regime IID.

CONJECTURE 1.1. *Let B be the two dimensional Brownian motion on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ as in Definition 4.2 and consider the EWF where $\hat{h}$ is as in (4.13) and (μ_1, μ_2, μ_3), (c_1, c_2, c_3) are as in Case IID. Then there exist functions $\Psi_i : \mathbb{R}_+ \to \mathbb{R}_+$, $i = 1, 2$, that are Lipschitz, strictly increasing, and $\Psi_i(x) \to \infty$ as $x \to \infty$, such that there is a unique pair of $\{\mathcal{F}_t\}$ adapted continuous processes W_1^*, W_2^* with values in $\mathbb{R}_+$ that solve*

$$\begin{aligned} W_1^*(t) &= B_1(t) + \sup_{0\leq s\leq t} [B_1(s) - \Psi_1(W_2^*(s))]^- , \\ W_2^*(t) &= B_2(t) + \sup_{0\leq s\leq t} [B_2(s) - \Psi_2(W_1^*(s))]^- , \end{aligned} \tag{1.2}$$

and

$$I_1^*(t) = \sup_{0\leq s\leq t} [B_1(s) - \Psi_1(W_2^*(s))]^- , I_2^*(t) = \sup_{0\leq s\leq t} [B_2(s) - \Psi_2(W_1^*(s))]^-$$

is an optimal control for the EWF.

Note that Case IIB corresponds to a setting where $\Psi_1 = \Psi$, with Ψ as in (1.1), and $\Psi_2 = 0$.

The paper is organized as follows. In Section 2 we present the queuing network considered in this work. We also introduce our main assumptions on the arrival and service processes. In Section 3 we introduce our scheduling policy and state the main result which gives asymptotic optimality of the policy with suitable choices of threshold parameters. Section 4 reviews results from [7], in particular we present the solution of the Brownian control problem associated with the network from Section 2. Section 5 contains the proof of our main result: Theorem 3.7. Key steps in the proof of the theorem are contained in Theorem 5.4, Theorem 5.5 and Lemma 5.7, the proofs of which are given in Section 6. Finally the appendix summarizes

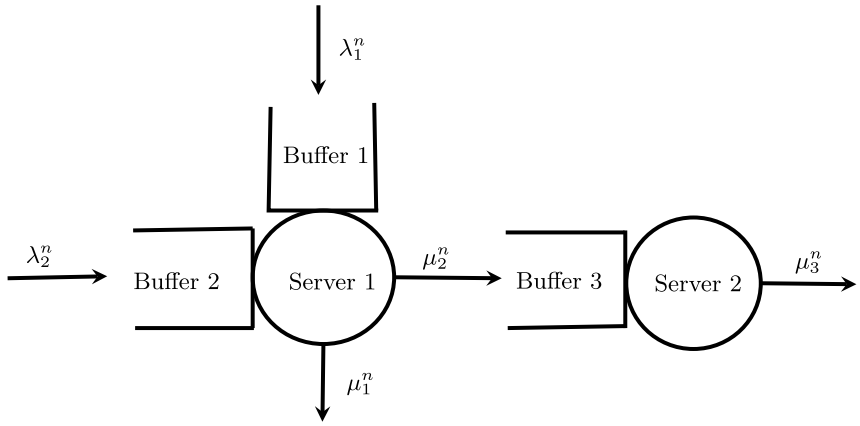


FIG 1. *The nth crisscross network.*

some elementary facts about the one dimensional Skorohod map that are appealed to in our proofs.

The following notation will be used. For a Polish space S, $D([0,\infty) : S)$ will denote the space of right continuous functions with left limits (RCLL) from $[0,\infty)$ to S equipped with the usual Skorohod topology. Define $\mathcal{D}_1 = \{f \in D([0,\infty) : \mathbb{R}) : f(0) \geq 0\}$. All stochastic processes in this work will have RCLL sample paths. A stochastic process X with values in S will be regarded as a random variable with values in $D([0,\infty) : S)$. Convergence in distribution of S valued random variables X^n to X will be denoted as $X^n \Rightarrow X$. A sequence X^n of processes with sample paths in $D([0,\infty) : S)$ is said to be C-tight if the corresponding sequence of probability laws is relatively compact (in the usual weak convergence topology) and any limit point is supported on the space of S valued continuous functions.

2. Queueing network model.

2.1. *Network structure.* Consider a sequence of networks indexed by $n \in \mathbb{N}$ of the form in Figure 1. The nth network consists of 3 classes of customers (corresponding to 3 buffers) and 2 servers. For $k = 1, 2$, customers of Class k arrive according to a renewal process with rate λ_k^n and receive service at Station 1. Class 1 customers leave the system once their service is completed. Class 2 customers after being served at Station 1 proceed to Station 2, where they are re-designated as Class 3 customers and get served at Station 2. The service rates for these 3 classes of customers are denoted by $\mu_j^n, j = 1, 2, 3$. Within each class, customers are processed using the First-Come-First-Served discipline.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

A precise mathematical description is as follows. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space; all random variables and stochastic processes described in this work are, without loss of generality, defined on this common probability space. For $k = 1, 2$, and $j = 1, 2, 3$, let $\{u_k(i) : i = 1, 2, \cdots\}$ be a sequence of i.i.d. random variables with mean 1 and standard deviation σ_k, and $\{v_j(i) : i = 1, 2, \cdots\}$ a sequence of i.i.d. random variables with mean 1 and standard deviation ς_j. In the nth network, for $k = 1, 2$ and $j = 1, 2, 3$, the inter-arrival times $\{u_k^n(i) : i = 1, 2, \cdots\}$ for Class k customers, and the service times $\{v_j^n(i) : i = 1, 2, \cdots\}$ for Class j customers are given by

$$u_k^n(i) = \frac{1}{\lambda_k^n} u_k(i), \quad v_j^n(i) = \frac{1}{\mu_j^n} v_j(i),$$

where $\lambda_k^n, \mu_j^n \in (0, \infty)$ are the arrival and service rates. We further assume that the sequences of inter-arrival times and service times are mutually independent for each $n \in \mathbb{N}$. Define

$$\xi_k^n(l) = \sum_{i=1}^{l} u_k^n(i), \ \eta_j^n(l) = \sum_{i=1}^{l} v_j^n(i) \ \text{ for } l = 1, 2, \cdots, \ k = 1, 2, \ j = 1, 2, 3.$$

Then the arrival and service processes can be described as follows: For $t \geq 0$,

$$A_k^n(t) = \sup\{l \geq 0 : \xi_k^n(l) \leq t\}, \ k = 1, 2,$$
$$S_j^n(t) = \sup\{l \geq 0 : \eta_j^n(l) \leq t\}, \ j = 1, 2, 3.$$

Thus $A_k^n(t)$ represents the numbers of customers of Class k who have arrived up to time t and $S_j^n(t)$ represents the number of customers of Class j who would have finished service up to time t if the corresponding server had continuously served Class j customers during time interval $[0, t]$. We make the following assumptions on the arrival and service rates.

ASSUMPTION 2.1. *For $k = 1, 2$ and $j = 1, 2, 3$, there exist $\lambda_k, \mu_j \in (0, \infty)$ such that $\lim_{n\to\infty} \lambda_k^n = \lambda_k$, $\lim_{n\to\infty} \mu_j^n = \mu_j$.*

The following is our main heavy traffic assumption.

ASSUMPTION 2.2. *The following relations hold for the arrival and service rate parameters:*

$$\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} = 1, \ \frac{\lambda_2}{\mu_3} = 1\,, \tag{2.1}$$

and there exist $b_i \in \mathbb{R}$, $i = 1, 2, 3$, *such that*

$$\lim_{n\to\infty} \sqrt{n}\left(\frac{\lambda_i^n}{\mu_i^n} - \frac{\lambda_i}{\mu_i}\right) = b_i, \quad i = 1, 2, \quad \lim_{n\to\infty} \sqrt{n}\left(\frac{\lambda_2^n}{\mu_3^n} - 1\right) = b_3\,. \tag{2.2}$$

Condition (2.1) says the traffic intensities at both stations converge to 1 as $n \to \infty$, while the convergence rates of the traffic intensities are characterized in (2.2).

Finally, we make the following assumption on the logarithmic moment generating functions for interarrival and service times that enables certain large deviation estimates for the renewal processes A_k^n, S_j^n (see Lemma 6.1).

ASSUMPTION 2.3. *There is a non-empty open neighborhood* $\mathcal{O}$ *of* $0 \in \mathbb{R}$ *such that for all* $l \in \mathcal{O}$,

$$\Lambda_{a,k} \doteq \log \mathbb{E}(e^{lu_k(1)/\lambda_k}) < \infty, \ \ k = 1, 2,$$
$$\Lambda_{s,j} \doteq \log \mathbb{E}(e^{lv_j(1)/\mu_j}) < \infty, \ \ j = 1, 2, 3.$$

2.2. *Scheduling control.* Scheduling control for the nth network is described by a vector-valued allocation process

$$T^n(t) = (T_1^n(t), T_2^n(t), T_3^n(t))', \ \ t \geq 0\,,$$

where for $j = 1, 2, 3$, $T_j^n(t)$ represents the cumulative amount of service time devoted to Class j customers in the time interval $[0, t]$. The idle time processes at two servers are defined as follows:

$$I_1^n(t) = t - T_1^n(t) - T_2^n(t), \quad I_2^n(t) = t - T_3^n(t), \ \ t \geq 0\,.$$

For simplicity, we assume the system is initially empty. Then the queue-length processes corresponding to the three types of customers can be described as follows. For $t \geq 0$ and $i = 1, 2$,

$$Q_i^n(t) = A_i^n(t) - S_i^n(T_i^n(t)), \quad Q_3^n(t) = S_2^n(T_2^n(t)) - S_3^n(T_3^n(t)). \tag{2.3}$$

We write $A^n = (A_1^n, A_2^n)'$, $\lambda^n = (\lambda_1^n, \lambda_2^n)'$. The quantities $S^n, I^n, Q^n, \mu^n, \lambda, \mu$ are defined similarly.

In order to precisely formulate the family of admissible control policies, we need to introduce the notion of multi-parameter filtrations and stopping times (cf. Section 2.8 of [12]). Define for $a = (a_1, a_2)' \in \mathbb{N}^2, b = (b_1, b_2, b_3)' \in \mathbb{N}^3$,

$$\mathcal{G}^n(a, b) = \sigma\left\{u_i^n(\tilde{a}_i), v_j^n(\tilde{b}_j) : \tilde{a}_i \leq a_i, \tilde{b}_j \leq b_j, \ i = 1, 2, j = 1, 2, 3\right\}.$$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Then $\{\mathcal{G}^n(a,b) : a \in \mathbb{N}^2, b \in \mathbb{N}^3\}$ is a *multiparameter filtration* with the following (partial) ordering

$$(\tilde{a},\tilde{b}) \leq (a,b) \text{ if and only if } \tilde{a}_i \leq a_i, \tilde{b}_j \leq b_j,\ i = 1,2, j = 1,2,3.$$

A $\{\mathcal{G}^n(a,b) : a \in \mathbb{N}^2, b \in \mathbb{N}^3\}$ *multiparameter stopping time* is a random variable $\mathbf{T}$ which takes values in $\bar{\mathbb{N}}^5$, where $\bar{\mathbb{N}} = \mathbb{N} \cup \{\infty\}$, such that

$$\{\mathbf{T} = (a,b)\} \in \mathcal{G}^n(a,b), \quad \text{for all } a \in \mathbb{N}^2, b \in \mathbb{N}^3.$$

The σ-field associated with such a stopping time is given by

$$\mathcal{G}^n_{\mathbf{T}} = \left\{B \in \mathcal{F} : B \cap \{\mathbf{T} = (a,b)\} \in \mathcal{G}^n(a,b), \text{ for all } a \in \mathbb{N}^2, b \in \mathbb{N}^3\right\}.$$

The scheduling control process $\{T^n(t)\}$ is required to satisfy the following conditions.

(i) For $j = 1,2,3$ and $n \in \mathbb{N}$, T^n_j is a continuous non-decreasing process with $T^n_j(0) = 0$.
(ii) For $k = 1,2$ and $n \in \mathbb{N}$, I^n_k is a continuous non-decreasing process with $I^n_k(0) = 0$.
(iii) For $j = 1,2,3$, $t \geq 0$, and $n \in \mathbb{N}$, $Q^n_j(t) \geq 0$.
(iv) Define for each $t \geq 0$ and $n \in \mathbb{N}$, a $\mathbb{N}^5$ valued random variable

$$\sigma^n_0(t) \doteq \left(A^n_i(nt) + 1, i = 1,2;\ S^n_j(T^n_j(nt)) + 1, j = 1,2,3\right).$$

Then, for each $t \geq 0$, $\sigma^n_0(t)$ is a $\{\mathcal{G}^n(a,b) : a \in \mathbb{N}^2, b \in \mathbb{N}^3\}$ stopping time. Define the filtration $\{\mathcal{G}^n_1(t) : t \geq 0\}$ as $\mathcal{G}^n_1(t) = \mathcal{G}^n_{\sigma^n_0(t)}$. Then $I^n(nt)$ is $\mathcal{G}^n_1(t)$ measurable for every $t \geq 0$.

From (i) and (ii), we see that for all $j = 1,2,3$,

(2.4) T^n_j is uniformly (in n) Lipschitz continuous with Lipschitz constant bounded by 1.

Although condition (iv) above appears somewhat technical, it is a natural non-anticipativity property and Theorem 5.4 of [5] shows that the condition is satisfied for a very broad family of control processes (see Remark 3.5).

Any process T^n satisfying the above conditions will be referred to as an admissible control policy for the nth network.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

2.3. *Scaled processes.* Now we define fluid-scaled and diffusion-scaled processes corresponding to the processes described above. For each $n \in \mathbb{N}$, define for $t \geq 0$, $\bar{T}^n(t) \doteq n^{-1}T^n(nt)$. Processes $\bar{I}^n, \bar{A}^n, \bar{S}^n, \bar{Q}^n$ are defined similarly. These will occasionally be referred to as fluid scaled processes.

We also defined diffusion-scaled processes:

$$\hat{A}^n(t) \doteq n^{-1/2}(A^n(nt) - n\lambda^n t),\ \hat{S}^n(t) \doteq n^{-1/2}(S^n(nt) - n\mu^n t),$$
$$\hat{T}^n(t) \doteq n^{-1/2}T^n(nt),\ \hat{I}^n(t) \doteq n^{-1/2}I^n(nt),\ \hat{Q}^n(t) \doteq n^{-1/2}Q^n(nt).$$

We next define the workload process $W^n = \{(W_1^n(t), W_2^n(t))' : t \geq 0\}$, which measures the amount of service needed for customers that are in the system at time t. More precisely, for $t \geq 0$, define

$$W_1^n(t) = \frac{Q_1^n(t)}{\mu_1^n} + \frac{Q_2^n(t)}{\mu_2^n},\ W_2^n(t) = \frac{Q_2^n(t)}{\mu_3^n} + \frac{Q_3^n(t)}{\mu_3^n}. \tag{2.5}$$

The fluid and diffusion scaled workload processes are defined as follows: For $t \geq 0$,

$$\bar{W}^n(t) \doteq n^{-1}W^n(nt),\ \hat{W}^n(t) \doteq n^{-1/2}W^n(nt).$$

3. Main result. We consider an expected linear infinite horizon discounted cost, associated with an admissible control T^n and the corresponding normalized queue-length process $\hat{Q}^n$, given as follows:

$$\hat{J}^n(T^n) = \mathbb{E}\left(\int_0^\infty e^{-\gamma t} c \cdot \hat{Q}^n(t)dt\right), \tag{3.1}$$

where $\gamma \in (0, \infty)$ is the discount factor and $c \equiv (c_1, c_2, c_3)'$ is a strictly positive vector of holding costs. The aim is to find a sequence $\{T^n\}$ of scheduling control policies which is asymptotically optimal, namely it satisfies

$$\lim_{n\to\infty} \hat{J}^n(T^n) = \inf \liminf_{n\to\infty} \hat{J}^n(\tilde{T}^n),$$

where the infimum is over the set of all admissible control policy sequences $\{\tilde{T}^n\}_{n\in\mathbb{N}}$. We make the following assumption on the service rates and the holding cost parameters.

ASSUMPTION 3.1. *$c_2\mu_2 - c_3\mu_2 < 0$, and $c_2\mu_2 - c_1\mu_1 \geq 0$.*

Note that the above assumption implies $c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0$. As noted in the Introduction, this parameter regime is the Case IIB considered in [22] and an optimal policy needs to suitably balance several considerations such as, overall cost is reduced at a higher rate if Server 1 processes Class 1

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

customers instead of Class 2 customers; it is cheaper to keep jobs in Buffer 2 than in Buffer 3; the cost of jobs in the queues processed by Server 1 is reduced more rapidly if jobs in Buffer 2 are processed; and Server 2 idling when there is work in Buffer 2 leads to an undesirable server underutilization.

We now describe our proposed policy which suitably takes into account the various complex features of this parameter regime. The policy is motivated by the form of the solution of the Brownian control problem associated with this control problem. This point will be explained further in Section 4 (see comments in Subsection 4.1) where the Brownian control problem is studied in detail. In particular, this section will make clear the reason the factor $\frac{\mu_1^n}{\mu_2^n}$ appears in the definition below.

Fix $c, l_0 \in (1,\infty)$ and $g_0 \in (0,\infty)$. Define $L^n \doteq \lfloor l_0 \log n \rfloor$ and $C^n \doteq \lfloor c_0 \log n \rfloor$, where $c_0 = cl_0$. Since we are only interested in asymptotic optimality, we will assume without any loss of generality, that $n \geq \bar{n}$, where $\bar{n}$ is such that for all $n \geq \bar{n}$, $C^n - L^n - 1 \geq 1$ and $\frac{\mu_1^n}{\mu_2^n}(C^n - L^n + 2) \geq 1$. The control policy will be specified in terms of the *free boundary* $\{(x_1, x_2) : x_1 = \Psi(x_2)\}$ where Ψ is as specified in (1.1) with J^* as in (4.24).

DEFINITION 3.2 (Control policy). *The policy is as follows. Server* 2 *processes jobs from Buffer 3 whenever the buffer is nonempty. The sequencing policy for Server* 1 *is as follows. At time* $s \in [0,\infty)$,

if $Q_3^n(s) - \frac{\mu_2^n}{\mu_1^n} Q_1^n(s) < L^n$,

- *serve Class* 1 *customers (provided* $Q_1^n(s) > 0$*) if either* $Q_3^n(s) \geq C^n - 1$ *or* $Q_2^n(s) = 0$;
- *serve Class* 2 *customers if* $Q_3^n(s) < C^n - 1$ *and* $Q_2^n(s) > 0$;
- *idle Server* 1 *if* $Q_1^n(s) = Q_2^n(s) = 0$;

if $Q_3^n(s) - \frac{\mu_2^n}{\mu_1^n} Q_1^n(s) \geq L^n$,

- *serve Class* 1 *customers (provided* $Q_1^n(s) > 0$*) if either* $Q_1^n(s) \geq \frac{\mu_1^n}{\mu_2^n}(C^n - L^n + 2)$ *or* $Q_2^n(s) = 0$ *or* $\{Q_1^n(s) < \frac{\mu_1^n}{\mu_2^n}(C^n - L^n + 2), Q_2^n(s) > 0, W_1^n(s) - \sqrt{n}\Psi(W_2^n(s)/\sqrt{n}) < g_0\}$;
- *serve Class* 2 *customers if* $Q_1^n(s) < \frac{\mu_1^n}{\mu_2^n}(C^n - L^n + 2)$, $Q_2^n(s) > 0$, *and* $W_1^n(s) - \sqrt{n}\Psi(W_2^n(s)/\sqrt{n}) \geq g_0$;
- *idle Server 1 if either* $Q_1^n(s) = Q_2^n(s) = 0$ *or* $\{Q_1^n(s) = 0, Q_2^n(s) > 0, W_1^n(s) - \sqrt{n}\Psi(W_2^n(s)/\sqrt{n}) < g_0\}$.

REMARK 3.3. *The free boundary* Ψ *is given in terms of the value function* J^* *of the* Workload Control Problem *in Section 4. Although this func-*

tion does not have a closed form expression, there are well developed numerical methods for solving such free boundary problems (see [6, 19, 20, 23]). The problem of developing a numerical approximation for the free boundary problem in [7] and using it for implementing the above threshold policy is currently under investigation.

REMARK 3.4. *We note that the proposed control policy addresses many of the complex features of the parameter regime. In particular the policy says that if Queue 1 is sufficiently large then it receives priority, unless there is a risk of Server 2 idling despite there being jobs in Queue 2. Furthermore, Server 1 idles when Queue 1 is empty and there is enough work in Queue 3, so as to keep jobs in the buffer with a lower holding cost. This latter property is enforced by the last line in the control policy that involves the free boundary function* Ψ. *Finally, the policy ensures that the scaled workload process* $\hat{W}^n(t) \in G = \{x \in \mathbb{R}^2_+ : x_1 \geq \Psi(x_2)\}$ *for each* t *asymptotically (see Lemma 5.7 for the precise statement). This property is motivated from the fact that the solution of the EWF,* $W^*(t) \in G$ *for each* t. *In particular, in our policy, when* $Q^n_3(s) - \frac{\mu^n_2}{\mu^n_1} Q^n_1(s) \geq L^n$, *if* $W^n_1(s) - \sqrt{n}\Psi(W^n_2(s)/\sqrt{n}) < g_0$, *the server will stop serving Class* 2 *customers to prevent* $W^n_1(s) - \sqrt{n}\Psi(W^n_2(s)/\sqrt{n})$ *from getting even smaller.*

REMARK 3.5. *From Theorem 5.4 of [5] it follows that the control policy in Definition 3.2 is an admissible control policy in the sense of Section 2.2. Roughly speaking this theorem says that a control policy that satisfies (i)–(ii) in Section 2.2 satisfies (iv) as well if it doesn't change between two consecutive events and at each event instant it is a function of previous interarrival times, service times, and routing decisions, where an event is defined to be an exogenous arrival or a service completion. Furthermore, an explicit representation for the policy in terms of a vector allocation process* T^n *can be given in a manner analogous to Remark 3.8 of [4].*

REMARK 3.6. *As noted in the Introduction, the treatment of Case IIC is expected to be similar to that for Case IIB studied here. The optimal solution of the EWF in Case IIC is given as*

$$W^*_1 = \Gamma(B_1)(t),$$
$$W^*_2 = \Gamma(B_2 - \Psi(W^*_1))(t) + \Psi(W^*_1(t)),$$

which says

$$(W^*_1(t), W^*_2(t)) \in \{x \in \mathbb{R}^2_+ : x_2 \geq \Psi(x_1)\},$$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

where Ψ is defined as in (1.1) by switching w_1 and w_2. Guided by this and our analysis for Case IIB, we propose the following scheduling policy for Case IIC, which is expected to be asymptotically optimal. Let C^n, L^n and g_0 as in Definition 3.2. Then Server 2 processes jobs as long as it is non-empty. For Server 1, at time $s \in [0,\infty)$,

if $Q_3^n(s) - \frac{\mu_2^n}{\mu_1^n} Q_1^n(s) < L^n$,

- *serve Class 1 (provided $Q_1^n(s) > 0$) if either $Q_3^n(s) \geq C^n - 1$ or $Q_2^n(s) = 0$ or $W_2^n(s) - \sqrt{n}\Psi(W_1^n(s)/\sqrt{n}) < g_0$;*
- *serve Class 2 if $Q_3^n(s) < C^n - 1, Q_2^n(s) > 0$, and $W_2^n(s) - \sqrt{n}\Psi(W_1^n(s)/\sqrt{n}) \geq g_0$;*
- *idle server 1 if $Q_1^n(s) = Q_2^n(s) = 0$;*

if $Q_3^n(s) - \frac{\mu_2^n}{\mu_1^n} Q_1^n(s) \geq L^n$,

- *serve Class 1 customers (provided $Q_1^n(s) > 0$) if either $Q_1^n(s) \geq \frac{\mu_1^n}{\mu_2^n}(C^n - L^n + 2)$ or $Q_2^n(s) = 0$;*
- *serve Class 2 customers if $Q_1^n(s) < \frac{\mu_1^n}{\mu_2^n}(C^n - L^n + 2)$ and $Q_2^n(s) > 0$;*
- *idle Server 1 if $Q_1^n(s) = Q_2^n(s) = 0$.*

See Remark 6.4 for some comments on the main technical differences in the proof of asymptotic optimality of the above policy.

The following is our main result, which gives the asymptotic optimality of the policy proposed in Definition 3.2. The limit of the cost under the proposed policy is characterized in Theorem 5.2 in terms of the solution of a Brownian control problem.

THEOREM 3.7. *Let $g_0 \in (0,\infty)$. There exist $c, \bar{l} \in (1,\infty)$ such that the sequence of scheduling controls $\{T^n\}$ defined in Definition 3.2 with threshold parameters c, l_0, g_0, with $l_0 \geq \bar{l}$, satisfies*

$$\lim_{n\to\infty} \hat{J}^n(T^n) = \inf \liminf_{n\to\infty} \hat{J}^n(\tilde{T}^n),$$

where the infimum is taken over all admissible control policy sequences $\{\tilde{T}^n\}$.

REMARK 3.8. *The parameter g_0 can be taken to be any positive real number. The parameters c and $\bar{l}$ in Theorem 3.7 can be chosen as follows.*

For $i = 1,2,3, j = 1,2$, and $l = 1,2$, define $\Theta_l^{s,i}$ and $\Theta_l^{a,j}$ as Θ_l in Lemma 6.1 with Λ^* replaced by $\Lambda_{s,i}^*$ and $\Lambda_{a,j}^*$, respectively, where $\Lambda_{s,i}^*$ and

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

$\Lambda^*_{s,j}$ *are the Legendre-Fenchel transforms of* $\Lambda_{s,i}$ *and* $\Lambda_{a,j}$, *respectively. Now let* θ_4 *be as in Theorem 5.4. From* (6.23) *one can see that,* θ_4 *can be chosen as follows:*

$$\theta_4 = \left(\inf_{n\geq 1} \frac{\mu_1^n}{2\mu_2^n(\lambda_1^n+\epsilon)} \right) \min\left\{ \lambda_1 \Theta_1^{a,1}(\lambda_1,\epsilon),\ \mu_1 \Theta_2^{s,1}(\mu_1,\epsilon),\ p_0\epsilon/(2\mu_1) \right\},$$

where $\epsilon \in (0, (\mu_1-\lambda_1)/4)$ *and* $p_0 \in \mathcal{O}$. *Let* $c = 1 + \frac{4}{\theta_4}$.

Next let γ_4 *be as in Theorem 5.5. From* (6.54) *we see that* γ_4 *can be taken to be*

$$\begin{aligned}\gamma_4 = \min \Big\{ & d\lambda_2 \Theta_1^{a,2}(\lambda_2,\epsilon_1)/K,\ d(\theta+1)(\mu_i - 2\epsilon_1)\Theta_2^{s,i}(\mu_i,\epsilon_1)/K, \\ & d\theta\mu_i \Theta_1^{s,i}(\mu_i,\epsilon_1)/K,\ dp_0(\theta+1)\epsilon_1/(2\mu_i K),\ \mu_i \Theta_1^{s,i}(\mu_i,\epsilon_1)/(4\mu_3), \\ & \lambda_2 \Theta_1^{a,2}(\lambda_2,\epsilon_1)/(4\mu_3),\ (\mu_i - 2\epsilon_1)\Theta_2^{s,i}(\mu_i,\epsilon_1)/(4\mu_3), \\ & (\lambda_2 - 2\epsilon_1)\Theta_2^{a,2}(\lambda_2,\epsilon_1)/(4\mu_3),\ p_0\epsilon_1/(8\mu_3\mu_i),\ p_0\epsilon_1/(8\mu_3\lambda_2), \\ & i = 2,3 \Big\},\end{aligned}$$

where ϵ_1, d, K, θ *are as in* (6.31). *Finally, let* $\bar{l} > \max\{1, \frac{3}{\gamma_4}\}$.

EXAMPLE 3.9. *We now give an example to illustrate how the various parameters in the policy are determined in practice. Consider the case where the interarrival and service times have Gamma distributions. Let* $\text{Gamma}(\alpha,\beta)$ *denote the gamma distribution with density function*

$$f(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad \alpha,\beta > 0 \text{ and } x \geq 0.$$

Suppose that for $k = 1,2$ *and* $j = 1,2,3$, *the random variables* $u_k(1) \sim \text{Gamma}(2,2)$ *and* $v_j(1) \sim \text{Gamma}(4,4)$. *Let arrival and service rates* $\lambda_k^n = \frac{1}{2} + \frac{1}{\sqrt{n}}$ *and* $\mu_1^n = \mu_2^n = 1, \mu_3^n = \frac{1}{2}$. *Then*

$$\lambda_k^n \to \lambda_k = \frac{1}{2}, k = 1,2, \quad \text{and} \quad \mu_1 = \mu_2 = 1, \mu_3 = \frac{1}{2}.$$

Clearly the parameters satisfy the heavy traffic condition in Assumption 2.2. Next note that

$$\begin{aligned}\Lambda_{a,k}(t) &= \log E(e^{tu_k(1)/\lambda_k}) = -2\log(1-t), \quad t<1,\ k=1,2, \\ \Lambda_{s,j}(t) &= \log E(e^{tv_j(1)/\mu_j}) = -4\log(4-t) + 4\log 4, \quad t<4,\ j=1,2, \\ \Lambda_{s,3}(t) &= \log E(e^{tv_3(1)/\mu_3}) = -4\log(2-t) + 4\log 2, \quad t<2,\end{aligned}$$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

and for $x > 0$,

$$\Lambda^*_{a,k}(x) = x - 2 + 2\log 2 - 2\log(x),\ k = 1, 2,$$
$$\Lambda^*_{s,j}(x) = 4x - 4 - 4\log(x),\ j = 1, 2,$$
$$\Lambda^*_{s,3}(x) = 2x - 4 + 4\log 4 - 4\log(2x).$$

We now follow Remark 3.8 to choose the threshold parameters for the proposed policy. Choosing $\epsilon = 1/8$,

$$\Theta^{a,1}_1(\lambda_1, \epsilon) = \Lambda^*_{a,1}(24/13) = 6.24 \times 10^{-3},$$
$$\Theta^{s,1}_2(\mu_1, \epsilon) = \Lambda^*_{s,1}(17/16) = 7.50 \times 10^{-3}.$$

Now choosing $p_0 = 1/2$, *we have* $\theta_4 = 9.6 \times 10^{-4}$ *and* $c = 4.17 \times 10^3$. *We next choose* $\epsilon_1 = 1/16, K = 72, d = 6 \times 10^5, \theta = 7.5 \times 10^{-6}$. *Then*

$$\Theta^{a,2}_1(\lambda_2, \epsilon_1) = \Lambda^*_{a,2}(48/25) = 1.64 \times 10^{-3},$$
$$\Theta^{s,2}_1(\mu_2, \epsilon_1) = \Lambda^*_{s,2}(48/49) = 8.44 \times 10^{-4},$$
$$\Theta^{s,3}_1(\mu_3, \epsilon_1) = \Lambda^*_{s,3}(48/25) = 3.29 \times 10^{-3},$$
$$\Theta^{a,2}_2(\lambda_2, \epsilon_1) = \Lambda^*_{a,2}(17/8) = 3.75 \times 10^{-3},$$
$$\Theta^{s,2}_2(\mu_2, \epsilon_1) = \Lambda^*_{s,2}(33/32) = 1.91 \times 10^{-3},$$
$$\Theta^{s,3}_2(\mu_3, \epsilon_1) = \Lambda^*_{s,3}(17/8) = 7.50 \times 10^{-3},$$

Thus $\gamma_4 = 5.28 \times 10^{-5}$ *and* $\bar{l} = 5.69 \times 10^4$. *At last, we note that the values of the parameters* c *and* $\bar{l}$ *depend on the distributions of the inter-arrival and service times and the values of* ϵ, ϵ_1, *and* p_0. *For example, if we let* $\lambda^n_k = \frac{1}{4} + \frac{1}{\sqrt{n}}, \mu^n_k = \frac{1}{2}$, *for* $k = 1, 2$, *and* $\mu^n_3 = \frac{1}{4}$ *in the above setting, we obtain that* $c = 273$ *and* $\bar{l} = 97$.

4. Brownian control problem and equivalent workload formulation. We now introduce the Brownian control problem associated with the control problem from Section 3. Roughly speaking, the BCP is obtained by taking a formal limit of the sequence of queueing control problems. Using the scaling defined in Section 2.3, we have, from (2.3), for a given sequence of admissible control policies $\{T^n\}$, and for all $t \geq 0$ and $i = 1, 2$,

$$\begin{aligned}\hat{Q}^n_i(t) &= \hat{A}^n_i(t) - \hat{S}^n_i(\bar{T}^n_i(t)) + \sqrt{n}(\lambda^n_i t - \mu^n_i \bar{T}^n_i(t)) \\ &= \hat{A}^n_i(t) - \hat{S}^n_i(\bar{T}^n_i(t)) + \sqrt{n}\mu^n_i \left(\frac{\lambda^n_i}{\mu^n_i} - \frac{\lambda_i}{\mu_i}\right) t \\ &\quad + \sqrt{n}\mu^n_i \left(\frac{\lambda_i}{\mu_i} t - \bar{T}^n_i(t)\right),\end{aligned} \tag{4.1}$$

$$\begin{aligned}\hat{Q}_3^n(t) &= \hat{S}_2^n(\bar{T}_2^n(t)) - \hat{S}_3^n(\bar{T}_3^n(t)) + \sqrt{n}(\mu_2^n \bar{T}_2^n(t) - \mu_3^n \bar{T}_3^n(t)) \\ &= \hat{S}_2^n(\bar{T}_2^n(t)) - \hat{S}_3^n(\bar{T}_3^n(t)) + \sqrt{n}\left(\mu_2^n \frac{\lambda_2}{\mu_2} - \mu_3^n\right) t \\ &\quad - \sqrt{n}\mu_2^n \left(\frac{\lambda_2}{\mu_2} t - \bar{T}_2^n(t)\right) + \sqrt{n}\mu_3^n (t - \bar{T}_3^n(t)).\end{aligned} \tag{4.2}$$

For $t \geq 0$, let

$$\begin{aligned}\hat{X}_i^n(t) &\doteq \hat{A}_i^n(t) - \hat{S}_i^n(\bar{T}_i^n(t)) + \sqrt{n}\mu_i^n \left(\frac{\lambda_i^n}{\mu_i^n} - \frac{\lambda_i}{\mu_i}\right) t, \quad i = 1, 2, \\ \hat{X}_3^n(t) &\doteq \hat{S}_2^n(\bar{T}_2^n(t)) - \hat{S}_3^n(\bar{T}_3^n(t)) + \sqrt{n}\left(\mu_2^n \frac{\lambda_2}{\mu_2} t - \mu_3^n t\right),\end{aligned} \tag{4.3}$$

and

$$\hat{Y}_i^n(t) \doteq \sqrt{n}\left(\frac{\lambda_i}{\mu_i} t - \bar{T}_i^n(t)\right), \quad i = 1, 2, \quad \hat{Y}_3^n(t) \doteq \sqrt{n}(t - \bar{T}_3^n(t)). \tag{4.4}$$

From (4.1)–(4.4), we get the following relationships:

$$\begin{aligned}\hat{Q}_i^n(t) &= \hat{X}_i^n(t) + \mu_i^n \hat{Y}_i^n(t), \quad i = 1, 2, \\ \hat{Q}_3^n(t) &= \hat{X}_3^n(t) - \mu_2^n \hat{Y}_2^n(t) + \mu_3^n \hat{Y}_3^n(t).\end{aligned} \tag{4.5}$$

Define for $t \geq 0$,

$$\bar{T}^*(t) = \left(\frac{\lambda_1}{\mu_1}, \frac{\lambda_2}{\mu_2}, 1\right)' t. \tag{4.6}$$

Write $\hat{X}^n = (\hat{X}_1^n, \hat{X}_2^n, \hat{X}_3^n)'$. It can be argued (see for example Lemma 3.3 in [5]) that, under 'reasonable' control policies, $\bar{T}^n \Rightarrow \bar{T}^*$ and consequently, using functional central limit theorem for renewal processes, under such policies

$$\hat{X}^n \Rightarrow X, \tag{4.7}$$

where X is a three-dimensional Brownian motion that starts from the origin and has drift $(\mu_1 b_1, \mu_2 b_2, \mu_3 b_3 - \mu_2 b_2)$ and covariance matrix

$$\begin{pmatrix} \sigma_1^2\lambda_1 + \varsigma_1^2\lambda_1 & 0 & 0 \\ 0 & \sigma_2^2\lambda_2 + \varsigma_1^2\lambda_2 & -\varsigma_1^2\lambda_2 \\ 0 & -\varsigma_1^2\lambda_2 & \varsigma_2^2\lambda_2 + \varsigma_3^2\mu_3 \end{pmatrix}.$$

Also,

$$\hat{I}_1^n(t) = \hat{Y}_1^n(t) + \hat{Y}_2^n(t), \text{ and } \hat{I}_2^n(t) = \hat{Y}_3^n(t), \ t \geq 0, \tag{4.8}$$

which are nondecreasing processes starting from 0.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Thus taking a formal limit as $n \to \infty$ in (4.5), we arrive at the following BCP.

DEFINITION 4.1. *Let X be a three dimensional Brownian motion as in (4.7), given on some filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$. The Brownian control problem is to find an $\mathbb{R}^3$-valued $\{\mathcal{F}_t\}$-adapted stochastic process $\tilde{Y} = (\tilde{Y}_1, \tilde{Y}_2, \tilde{Y}_3)'$, which minimizes*

$$\mathbb{E}\left(\int_0^\infty e^{-\gamma t} c \cdot \tilde{Q}(t) dt\right),$$

subject to the following conditions. For all $t \geq 0$,

$$\begin{aligned} &0 \leq \tilde{Q}_i(t) \doteq X_i(t) + \mu_i \tilde{Y}_i(t), \ i = 1, 2, \\ &0 \leq \tilde{Q}_3(t) \doteq X_3(t) + \mu_3 \tilde{Y}_3(t) - \mu_2 \tilde{Y}_2(t), \end{aligned} \tag{4.9}$$

and

$$\tilde{I}_1 \doteq \tilde{Y}_1 + \tilde{Y}_2, \ \tilde{I}_2 \doteq \tilde{Y}_3 \ \text{ are non-decreasing, and } \ \tilde{I}_i(0) = 0, \ i = 1, 2. \tag{4.10}$$

We will refer to any $\{\mathcal{F}_t\}$-adapted process $\tilde{Y} = (\tilde{Y}_1, \tilde{Y}_2, \tilde{Y}_3)'$ satisfying (4.9) and (4.10) as an admissible control for the BCP and an admissible control that achieves the minimum cost as an optimal control for the BCP.

We now introduce an equivalent workload formulation of the above BCP that makes use of a certain static deterministic linear programming (LP) problem. Recall the workload process defined in (2.5). Using (4.5) and (4.8), we have for $t \geq 0$,

$$\begin{aligned} \hat{W}_1^n(t) &= \frac{\hat{Q}_1^n(t)}{\mu_1^n} + \frac{\hat{Q}_2^n(t)}{\mu_2^n} = \frac{\hat{X}_1^n(t)}{\mu_1^n} + \frac{\hat{X}_2^n(t)}{\mu_2^n} + \hat{I}_1^n(t), \\ \hat{W}_2^n(t) &= \frac{\hat{Q}_2^n(t)}{\mu_3^n} + \frac{\hat{Q}_3^n(t)}{\mu_3^n} = \frac{\hat{X}_2^n(t)}{\mu_3^n} + \frac{\hat{X}_3^n(t)}{\mu_3^n} + \hat{I}_2^n(t). \end{aligned} \tag{4.11}$$

Fix $w_1, w_2 \in [0, \infty)$. Consider the LP problem defined as:

$$\begin{aligned} &\text{minimize}_{q_1, q_2, q_3} \ \ c_1 q_1 + c_2 q_2 + c_3 q_3 \\ &\text{subject to } \ \frac{q_1}{\mu_1} + \frac{q_2}{\mu_2} = w_1, \ \frac{q_2}{\mu_3} + \frac{q_3}{\mu_3} = w_2, \ q_1, q_2, q_3 \geq 0. \end{aligned} \tag{4.12}$$

A straightforward calculation using the fact that $c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0$ shows that the value of the LP is

$$\hat{h}(w_1, w_2) = \begin{cases} (c_1\mu_1)w_1 + \frac{\mu_3}{\mu_2}(c_2\mu_2 - c_1\mu_1)w_2, & \text{when } \ \mu_3 w_2 \leq \mu_2 w_1, \\ (c_2\mu_2 - c_3\mu_2)w_1 + (c_3\mu_3)w_2, & \text{when } \ \mu_3 w_2 \geq \mu_2 w_1, \end{cases} \tag{4.13}$$

and the optimal solution is

$$
\begin{aligned}
&q_1^* = \frac{\mu_1}{\mu_2}(\mu_2 w_1 - \mu_3 w_2),\ q_2^* = \mu_3 w_2,\ q_3^* = 0, \text{if } \mu_3 w_2 \le \mu_2 w_1,\\
&q_1^* = 0,\ q_2^* = \mu_2 w_1,\ q_3^* = \mu_3 w_2 - \mu_2 w_1,\ \text{if } \mu_3 w_2 \ge \mu_2 w_1.
\end{aligned} \tag{4.14}
$$

Using (4.13), and taking a formal limit in (4.11), we arrive at the following control problem, which is usually referred to as the equivalent workload formulation (for the BCP in Definition 4.1).

DEFINITION 4.2. *Let X and $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\ge 0}, \mathbb{P})$ be as in Definition 4.1. The equivalent workload formulation (EWF) is to find an $\mathbb{R}^2$-valued $\{\mathcal{F}_t\}$-adapted stochastic process $\tilde{I} = (\tilde{I}_1, \tilde{I}_2)'$, which minimizes*

$$
\mathbb{E}\left(\int_0^\infty e^{-\gamma t}\hat{h}(\tilde{W}(t))dt\right), \tag{4.15}
$$

subject to the following conditions. For all $t \ge 0$

$$
\begin{aligned}
&0 \le \tilde{W}_1(t) \doteq B_1(t) + \tilde{I}_1(t),\ 0 \le \tilde{W}_2(t) \doteq B_2(t) + \tilde{I}_2(t),\\
&\tilde{I}_1,\ \tilde{I}_2 \text{ are nondecreasing, and } \tilde{I}_1(0) = \tilde{I}_2(0) = 0\,,
\end{aligned} \tag{4.16}
$$

where

$$
B_1(t) = \frac{X_1(t)}{\mu_1} + \frac{X_2(t)}{\mu_2},\quad B_2(t) = \frac{X_2(t)}{\mu_3} + \frac{X_3(t)}{\mu_3}. \tag{4.17}
$$

We will refer to any $\{\mathcal{F}_t\}$-adapted process $\tilde{I} = (\tilde{I}_1, \tilde{I}_2)'$ satisfying (4.16) as an admissible control for the EWF and an admissible control that achieves the minimum cost as an optimal control for the EWF.

The BCP and EWF of the above form were first introduced by Harrison in [13], and they have been used extensively in the study of optimal scheduling for multiclass queuing networks in heavy traffic (see [1, 2, 4, 5, 8, 9, 11, 14, 16, 18]). In particular, the BCP and EWF introduced here are identical to those in [4]. The following lemma says that in order to solve the BCP it suffices to solve the associated EWF. The proof of this lemma is straightforward from (4.13) and (4.14) and we refer the reader to Section 3.1 of [4] for details.

LEMMA 4.3. *Suppose $\tilde{I}^*$ is an optimal control of the EWF. Denote by $\tilde{W}^*$ the corresponding optimal workload (which is defined by (4.16) with $\tilde{I}$*

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

replaced by $\tilde{I}^*$). Define when $\mu_3\tilde{W}_2^*(t) < \mu_2\tilde{W}_1^*(t)$,

$$\begin{aligned}(4.18)\qquad &\tilde{Y}_1^*(t) \doteq -\frac{X_3(t)}{\mu_2} + \tilde{I}_1^*(t) - \frac{\mu_3}{\mu_2}\tilde{I}_2^*(t),\ \tilde{Y}_2^*(t) \doteq \frac{X_3(t)}{\mu_2} + \frac{\mu_3}{\mu_2}\tilde{I}_2^*(t),\\ &\tilde{Y}_3^*(t) \doteq \tilde{I}_2^*(t),\end{aligned}$$

and when $\mu_3\tilde{W}_2^*(t) \geq \mu_2\tilde{W}_1^*(t)$,

$$(4.19)\qquad \tilde{Y}_1^*(t) \doteq -\frac{X_1(t)}{\mu_1},\ \tilde{Y}_2^*(t) \doteq \frac{X_1(t)}{\mu_1} + \tilde{I}_1^*(t),\ \tilde{Y}_3^*(t) \doteq \tilde{I}_2^*(t).$$

Then $\tilde{Y}^*$ is an optimal control of the BCP.

We note that if $\tilde{Y}^*$ is an optimal solution of the BCP, then the corresponding optimal queue length is

$$(4.20)\qquad \tilde{Q}_1^*(t) = \frac{\mu_1}{\mu_2}(\mu_2\tilde{W}_1^*(t) - \mu_3\tilde{W}_2^*(t)),\ \tilde{Q}_2^*(t) = \mu_3\tilde{W}_2^*(t),\ \tilde{Q}_3^*(t) = 0,$$

if $\mu_3\tilde{W}_2^*(t) < \mu_2\tilde{W}_1^*(t)$, and

$$(4.21)\qquad \tilde{Q}_1^*(t) = 0,\ \tilde{Q}_2^*(t) = \mu_2\tilde{W}_1^*(t),\ \tilde{Q}_3^*(t) = \mu_3\tilde{W}_2^*(t) - \mu_2\tilde{W}_1^*(t),$$

if $\mu_3\tilde{W}_2^*(t) \geq \mu_2\tilde{W}_1^*(t)$.

Although the BCP and EWF here are the same as those in [4], the solutions to these problems are much less straightforward than in [4]. The latter paper considers the Case IIA, where, in particular, $c_2\mu_2 - c_3\mu_2 \geq 0$ and $c_2\mu_2 - c_1\mu_1 \geq 0$. In their setting, $\hat{h}$ is a non-decreasing function of both its arguments and consequently, the solution of the EWF is trivial, in that it is given by the solution of the one-dimensional Skorohod problem (see Appendix A for the definition and properties of Skorohod problem). However for the regime considered here, $\hat{h}$ no longer has the above monotonicity property and thus a simple closed form solution of the EWF or BCP is not available. The BCP in Cases IIB and IIC has been investigated in [7] where using certain optimal stopping problems a solution of the EWF has been provided in terms of the free boundary associated with the control problem. Below we summarize some key results from [7] that will be needed here.

Consider the workload control problem in Definition 4.2 corresponding to an arbitrary initial condition. More precisely, letting B and $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ be as in Definition 4.2, and $w = (w_1, w_2)' \in \mathbb{R}_+^2$, the control problem is to find an $\mathbb{R}_+^2$-valued $\{\mathcal{F}_t\}$-adapted non-decreasing process $\tilde{I}^w = (\tilde{I}_1^w, \tilde{I}_2^w)'$ to minimize

$$(4.22)\qquad \mathbb{E}\left(\int_0^\infty e^{-\gamma t}\hat{h}(\tilde{W}^w(t))dt\right)$$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

subject to the following conditions: For $t \geq 0$,

$$\begin{aligned} 0 \leq \tilde{W}_1^w(t) &\doteq w_1 + B_1(t) + \tilde{I}_1^w(t), \\ 0 \leq \tilde{W}_2^w(t) &\doteq w_2 + B_2(t) + \tilde{I}_2^w(t). \end{aligned} \tag{4.23}$$

We refer to $\tilde{I}^w = (\tilde{I}_1^w, \tilde{I}_2^w)'$ as an admissible control for the initial condition w. Define the optimal value function as:

$$J^*(w) = \inf_{\tilde{I}^w} \mathbb{E}\left(\int_0^\infty e^{-\gamma t} \hat{h}(\tilde{W}^w(t)) dt \right), \ w \in \mathbb{R}_+^2, \tag{4.24}$$

where the infimum is taken over all admissible controls for the initial condition w.

From Theorem 3.1. of [7], J^* is a C^1 function on $\mathbb{R}_+^2$. Now define for $w_2 \in \mathbb{R}_+$, $\Psi : \mathbb{R}_+ \to \mathbb{R}_+$ as in (1.1). The following result is taken from [7].

LEMMA 4.4 (Lemma 5.1. of [7]). *The function Ψ has the following properties.*

(i) *For all $w_2 \geq 0$, $0 \leq \Psi(w_2) \leq \frac{\mu_3}{\mu_2} w_2$.*
(ii) *Ψ is non-decreasing and Lipschitz continuous with Lipschitz constant bounded by $\frac{\mu_3}{\mu_2}$.*
(iii) $\lim_{w_2 \to \infty} \Psi(w_2) = \infty$.

The following result from [7] gives an optimal solution of the workload control problem in Definition 4.2.

THEOREM 4.5 (Theorem 5.2. of [7]). *For $t \geq 0$, define*

$$\begin{aligned} W_1^*(t) &= B_1(t) + \sup_{0 \leq s \leq t} [B_1(s) - \Psi(W_2^*(s))]^- , \\ W_2^*(t) &= B_2(t) + \sup_{0 \leq s \leq t} [B_2(s)]^- , \end{aligned} \tag{4.25}$$

where for $z \in \mathbb{R}$, $z^- = -\min\{0, z\}$. Then the minimum value of (4.15) over all admissible controls is given as

$$J^*(0) = \mathbb{E}\left(\int_0^\infty e^{-\gamma t} \hat{h}(W^*(t)) dt \right). \tag{4.26}$$

Thus

$$I_1^*(t) = \sup_{0 \leq s \leq t} [B_1(s) - \Psi(W_2^*(s))]^- , \ I_2^*(t) = \sup_{0 \leq s \leq t} [B_2(s)]^- \tag{4.27}$$

is an optimal control for the EWF.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Consider the one dimensional Skorohod map $\Gamma : \mathcal{D}_1 \to D([0,\infty) : \mathbb{R}_+)$ defined as

$$\Gamma(f)(t) \doteq f(t) + \sup_{0\le s\le t} (f(s))^-, \ f \in \mathcal{D}_1, \ t \ge 0.$$

Then for $t \ge 0$,

$$W_2^*(t) \doteq \Gamma(B_2)(t), \ W_1^*(t) \doteq \Gamma(B_1 - \Psi(W_2^*))(t) + \Psi(W_2^*(t)) \,. \tag{4.28}$$

Recall $G = \{x \in \mathbb{R}_+^2 : x_1 \ge \Psi(x_2)\}$. Clearly,

$$W^*(t) \in G \ \text{ for all } t \ge 0. \tag{4.29}$$

Roughly speaking, the process W^* behaves like the Brownian motion $(B_1, B_2)'$ in the interior of G and it is reflected on the boundary of G, where the directions of reflection on $\{x \in \mathbb{R}_+^2 : x \cdot e_2 = 0\}$ and $\{x \in \mathbb{R}_+^2 : x_1 = \Psi(x_2)\}$ are e_2 and e_1, respectively.

Combining Theorem 4.5 with Lemma 4.3 we have the following corollary.

COROLLARY 4.6. *Let $W^* = (W_1^*, W_2^*)'$ and $I^* = (I_1^*, I_2^*)'$ be as in Theorem 4.5. Then Y^* defined by (4.18) - (4.19) (replacing $\tilde{I}^*$ there with I^*) is an optimal control for the BCP and Q^* defined by (4.20)–(4.21) (replacing $\tilde{W}^*$ there by W^*) is the corresponding optimally controlled state process.*

4.1. *Comments on Proof of Theorem 3.7.* The solution to the BCP given above suggests the following control policy for the n-th network. Note that the set $\{Q_3^*(t) < \frac{\mu_2}{\mu_1} Q_1^*(t)\}$ equals $\{\mu_3 W_2^*(t) < \mu_2 W_1^*(t)\}$. Since on this set $Q_3^*(t) = 0$, a good policy for the n-th network should keep $\hat{Q}_3^n(t)$ close to 0 when $Q_3^n(t) < \frac{\mu_2^n}{\mu_1^n} Q_1^n(t)$. Similarly, when $Q_3^n(t) \ge \frac{\mu_2^n}{\mu_1^n} Q_1^n(t)$ the policy should ensure that $\hat{Q}_1^n(t)$ is close to 0. This motivates the thresholds $L^n = \lfloor l_0 \log n \rfloor$ and $C^n = \lfloor c_0 \log n \rfloor$ introduced above Definition 3.2. When $Q_3^n(t) - \frac{\mu_2^n}{\mu_1^n} Q_1^n(t) < L^n$, under the policy in Definition 3.2, Server 1 processes Class 2 customers (leading to an increase in $Q_3^n(t)$) only when $Q_3^n(t) < C^n - 1$. Thus C^n can be interpreted as the level of 'safety stock' that prevents idleness of Server 2. Since in the diffusion scaling the safety stock levels approach 0 (i.e. $C^n/\sqrt{n} \to 0$ as $n \to \infty$) the policy ensures that $\hat{Q}_3^n$ is close to 0 in this regime. Similarly when $Q_3^n(t) - \frac{\mu_2^n}{\mu_1^n} Q_1^n(t) \ge L^n$, Server 1 processes Class 1 jobs as soon as $Q_1^n(t) \ge \frac{\mu_1^n}{\mu_2^n}(C^n - L^n + 2)$, ensuring that in this regime $\hat{Q}_1^n$ remains close to 0. We refer the reader to the first two convergence results in Corollary 5.6 (which are consequences of Theorem 5.4) that make these statements precise. Next (4.29) suggests

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

how a near optimal policy should depend on the free boundary Ψ, specifically under such a policy the condition $\hat{W}^n(t) \in G$ for all t should be satisfied approximately for large n. As shown in Lemma 5.7, the policy in Definition 3.2 satisfies this property. The proof of this lemma relies on Theorem 5.4, the properties of the free boundary given in Lemma 4.4, and on the key idleness property formulated in the last line of Definition 3.2. Finally, in view of (4.28), under a near optimal policy the second coordinate of the scaled workload process, i.e. $\hat{W}_2^n$, should be an approximate solution of a one dimensional Skorohod problem for $(\hat{X}_2^n + \hat{X}_3^n)/\mu_3^n$. However, $\hat{W}_2^n(t)$ is not the immediate workload of Server 2 at time t, and so the points of increase of $\hat{I}_2^n$ do not necessarily correspond to time instants when $\hat{W}_2^n$ is zero. In order to establish the desired representation as a solution of a Skorohod problem we split $\hat{I}_2^n$ as the sum of $\int_0^{\cdot} 1_{\boldsymbol{B}_d(n,s)^c} d\hat{I}_2^n(s)$ and $\int_0^{\cdot} 1_{\boldsymbol{B}_d(n,s)} d\hat{I}_2^n(s)$, where the set $\boldsymbol{B}_d(n,s)$ is defined in the statement of Theorem 5.5, and argue that the latter term converges to 0 as $n \to \infty$. This is the third convergence result of Corollary 5.6 which is a consequence of Theorem 5.5.

5. Proof of asymptotic optimality. In this section we prove Theorem 3.7 which gives the asymptotic optimality of the policy proposed in Definition 3.2 for a suitable choice of threshold parameters g_0, c and l_0. We begin with the following result which is an immediate consequence of Theorem 3.1 of [5].

THEOREM 5.1. *Let for $n \geq 1$, $\tilde{T}^n$ be an admissible control policy for the n-th network. Then with $J^*(0)$ as in (4.26), we have,* $\liminf_{n\to\infty} \hat{J}^n(\tilde{T}^n) \geq J^*(0)$.

The above theorem says that the optimal cost of the BCP is a lower bound for the asymptotic cost for any sequence of admissible control policies. Thus it suffices to show that the sequence of policies in Definition 3.2 (with a suitable choice of threshold parameters) asymptotically achieves the optimal cost of the BCP. This is done in the theorem below.

THEOREM 5.2. *There exist $c, \bar{l} \in (1,\infty)$ such that for any $g_0 \in (0,\infty)$ and $l_0 \in [\bar{l},\infty)$, the sequence of control policies $\{T^n\}$ in Definition 3.2 with threshold parameters c, l_0, g_0, satisfies the following:*

(i) $(\hat{W}^n, \hat{I}^n) \Rightarrow (W^*, I^*)$ *as* $n \to \infty$,

(ii) $\hat{J}^n(T^n) \to J^*(0)$ *as* $n \to \infty$,

where $(\hat{W}^n, \hat{I}^n)$ are defined as in Section 2 using the above sequence of control policies, and W^*, I^*, $J^*(0)$ are as in (4.25), (4.27), and (4.26) respectively.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Proof of Theorem 3.7 is immediate from Theorems 5.1 and 5.2:

PROOF OF THEOREM 3.7. From Theorem 5.1

$$\inf \liminf_{n\to\infty} \hat{J}^n(\tilde{T}^n) \geq J^*(0)$$

where the infimum is taken over all admissible control sequences $\{\tilde{T}^n\}$. Also, with $\{T^n\}$ as in Theorem 5.2,

$$\inf \liminf_{n\to\infty} \hat{J}^n(\tilde{T}^n) \leq \liminf_{n\to\infty} \hat{J}^n(T^n) = J^*(0).$$

Combining the above two inequalities, we have $\inf \liminf_{n\to\infty} \hat{J}^n(\tilde{T}^n) = J^*(0) = \lim_{n\to\infty} \hat{J}^n(T^n)$. □

Rest of this section is devoted to the proof of Theorem 5.2. The proof relies on three technical results: Theorem 5.4, Theorem 5.5 and Lemma 5.7, the proofs of which are postponed to Section 6. Throughout this section $\{T^n\}$ will denote the sequence of control policies in Definition 3.2 with some choice of threshold parameters. We begin with the following lemma from [4].

LEMMA 5.3 (Lemma 4.7 of [4]). *Let $\{f_n\}$ and $\{g_n\}$ be sequences of functions in $D([0,\infty):\mathbb{R})$, and let f and g be continuous functions from $[0,\infty)$ to $\mathbb{R}$, such that $f_n \to f$, $g_n \to g$ in $D([0,\infty):\mathbb{R})$ as $n\to\infty$. Suppose that $\int_{[0,\infty)} e^{-\gamma t} 1_{\{g(t)=0\}} dt = 0$. Let ϵ_n be a sequence of non-negative real numbers converging to 0. Then for all $T>0$, the following hold:*

$$\int_0^T e^{-\gamma t} f_n(t) 1_{\{g_n(t)\geq \epsilon_n\}} dt \to \int_0^T e^{-\gamma t} f(t) 1_{\{g(t)\geq 0\}} dt \quad \text{as } n\to\infty,$$

$$\int_0^T e^{-\gamma t} f_n(t) 1_{\{g_n(t)\leq \epsilon_n\}} dt \to \int_0^T e^{-\gamma t} f(t) 1_{\{g(t)\leq 0\}} dt \quad \text{as } n\to\infty.$$

Given $c\in(1,\infty)$, let

$$\kappa(c) \doteq \max\left\{\frac{4\mu_1}{\mu_2}, \frac{4}{c-1}, \frac{c}{c-1}, 4\right\}. \tag{5.1}$$

For $n\in\mathbb{N}$, $\kappa\geq\kappa(c)$, and $t\geq 0$, define the events:

$$\boldsymbol{A}(n,t) \doteq \left\{Q_3^n(t) - \frac{\mu_2^n}{\mu_1^n} Q_1^n(t) < L^n\right\},$$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

and

$$\mathcal{E}_\kappa(n,t) \doteq \left\{ \sup_{0\le s\le t} \hat{Q}_3^n(s) 1_{\boldsymbol{A}(n,ns)} > \frac{\kappa(C^n - L^n + 1)}{\sqrt{n}} \right\} \cup \left\{ \sup_{0\le s\le t} \hat{Q}_1^n(s) 1_{\boldsymbol{A}(n,ns)^c} > \frac{\kappa(C^n - L^n + 1)}{\sqrt{n}} \right\}. \tag{5.2}$$

Proofs of the following two results are given in Section 6.

THEOREM 5.4. *There exist $\theta_i \in (0,\infty)$, $i = 1,2,3,4$ and $n_0 \in \mathbb{N}$ such that for the sequence $\{T^n\}$ with threshold parameters $l_0 > 1, c > 1, g_0 > 0$ and with $\mathcal{E}_\kappa(n,t)$ defined as in (5.2)*

$$\mathbb{P}(\mathcal{E}_\kappa(n,t)) \le \theta_1(nt+1)^2 e^{-\theta_2 nt} + \theta_3(nt+1)^3 n^{-\theta_4(c-1)l_0}, \tag{5.3}$$

whenever $\kappa \ge \kappa(c)$, $n \ge n_0$ and $nt \ge 2$.

THEOREM 5.5. *There exist $n_1 \in \mathbb{N}$, $\epsilon \in (0,1)$, $\gamma_i = \gamma_i(c) \in (0,\infty)$, $i = 1,2,3,4$, and $d = d(c) \in (0,\infty)$, such that for the sequence $\{T^n\}$ of control policies with threshold parameters $c > 1$ and arbitrary $l_0 > 1$, $g_0 > 0$,*

$$\mathbb{P}\left[\int_{[0,t)} 1_{\boldsymbol{B}_d(n,s)} d\hat{I}_2^n(s) \ne 0 \right] \le \gamma_1(nt+1)^2 e^{-\gamma_2 nt} + \gamma_3(nt+1)^3 n^{-\gamma_4 l_0}, \tag{5.4}$$

whenever $n \ge n_1$ and $nt \ge 2/\epsilon$, where

$$\boldsymbol{B}_d(n,t) \doteq \left\{ \hat{Q}_2^n(t) \ge \frac{dl_0 \log n}{\sqrt{n}} \right\}.$$

An immediate consequence of the above two theorems is the following.

COROLLARY 5.6. *Let θ_4 be as in Theorem 5.4. Let $c = 1 + \frac{4}{\theta_4}$, and let $\gamma_i \in (0,\infty)$, $i = 1,2,3,4$ and d be as in Theorem 5.5. Choose $\bar{l} \in (1,\infty)$ to be large enough so that $\gamma_4 \bar{l} > 3$. Fix $t \ge 0$. Then for all $l_0 \ge \bar{l}$, $g_0 > 0$ and sequence $\{T^n\}$ of control policies with threshold parameters c, l_0, g_0 the probabilities in (5.3) and (5.4) tend to 0 as $n \to \infty$ for all $\kappa \ge \kappa(c)$. In particular, as $n \to \infty$,*

$$\hat{Q}_1^n(\cdot) 1_{\boldsymbol{A}(n,n\cdot)^c} \Rightarrow 0, \quad \hat{Q}_3^n(\cdot) 1_{\boldsymbol{A}(n,n\cdot)} \Rightarrow 0, \quad \int_{[0,\cdot]} 1_{\boldsymbol{B}_d(n,s)} d\hat{I}_2^n(s) \Rightarrow 0.$$

For the rest of this section we fix threshold parameters c, l_0, g_0 and constants d, κ as in Corollary 5.6. We will suppress κ and d in the notation for $\boldsymbol{B}_d(n,s), \mathcal{E}_\kappa(n,s)$. We next provide a lower bound for $W_1^n(t) - \sqrt{n}\Psi(W_2^n(t)/\sqrt{n}), t \ge 0$, which will ensure that $\hat{W}^n(t) \in G$ asymptotically. The proof is given in Section 6.

LEMMA 5.7. *Under the control policies* $\{T^n\}$ *as in Corollary 5.6, there exist* $C_1, C_2 \in (0,\infty)$ *such that for* $t \geq 0$, *we have*

$$\begin{aligned} &W_1^n(t) - \sqrt{n}\Psi(W_2^n(t)/\sqrt{n}) \\ &\geq -\frac{1}{\mu_1^n}\left|1 - \frac{\mu_3\mu_2^n}{\mu_3^n\mu_2}\right| Q_1^n(t) - \frac{1}{\mu_2^n}\left|1 - \frac{\mu_3\mu_2^n}{\mu_3^n\mu_2}\right|(2Q_2^n(t)+1) \\ &\quad - C_1(C^n - L^n + 2) + g_0 - C_2. \end{aligned} \tag{5.5}$$

The following asymptotic property of the sequence $\{T^n\}$ will play a key role. Recall $\bar{T}^*$ defined in (4.6) and the fluid scaled processes $\bar{T}^n$ introduced in Section 2.3.

LEMMA 5.8. *As* $n \to \infty$, $\bar{T}^n \Rightarrow \bar{T}^*$ *and* $\bar{Q}^n \Rightarrow 0$.

PROOF. From the second expression in (4.11), we have for $t \geq 0$,

$$\hat{W}_2^n(t) = \frac{\hat{Q}_2^n(t)}{\mu_3^n} + \frac{\hat{Q}_3^n(t)}{\mu_3^n} = \frac{\hat{X}_2^n(t)}{\mu_3^n} + \frac{\hat{X}_3^n(t)}{\mu_3^n} + \hat{I}_2^n(t),$$

and thus

$$\begin{aligned} \frac{\hat{Q}_2^n(t)}{\mu_3^n}1_{\boldsymbol{B}(n,t)} + \frac{\hat{Q}_3^n(t)}{\mu_3^n} &= \frac{\hat{X}_2^n(t)}{\mu_3^n} + \frac{\hat{X}_3^n(t)}{\mu_3^n} - \frac{\hat{Q}_2^n(t)}{\mu_3^n}1_{\boldsymbol{B}(n,t)^c} \\ &\quad + \int_0^t 1_{\boldsymbol{B}(n,s)} d\hat{I}_2^n(s) + \int_0^t 1_{\boldsymbol{B}(n,s)^c} d\hat{I}_2^n(s). \end{aligned} \tag{5.6}$$

Note that the last term on the right hand side of (5.6) is equal to 0 when $t = 0$, and is nondecreasing, and increases only when the term on the left hand side of (5.6) is 0. Therefore the left side of (5.6) can be represented in terms of the one-dimensional Skorohod map Γ (see Proposition A.2 (i)) and we have for $t \geq 0$,

$$\begin{aligned} \hat{W}_2^n(t) &= \Gamma\left(\frac{\hat{X}_2^n(\cdot)}{\mu_3^n} + \frac{\hat{X}_3^n(\cdot)}{\mu_3^n} - \frac{\hat{Q}_2^n(\cdot)}{\mu_3^n}1_{\boldsymbol{B}(n,\cdot)^c} + \int_0^{\cdot} 1_{\boldsymbol{B}(n,s)} d\hat{I}_2^n(s)\right)(t) \\ &\quad + \frac{\hat{Q}_2^n(t)}{\mu_3^n}1_{\boldsymbol{B}(n,t)^c}. \end{aligned} \tag{5.7}$$

Thus using the Lipschitz continuity property of the Skorohod map (see Proposition A.2 (ii)), we have for $t \geq 0$,

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

$$
\begin{aligned}
&\sup_{0\le s\le t} \bar{W}_2^n(s)\\
&\quad\le \frac{1}{\sqrt{n}} \sup_{0\le s\le t} \Gamma\left(\frac{\hat{X}_2^n(\cdot)}{\mu_3^n} + \frac{\hat{X}_3^n(\cdot)}{\mu_3^n} - \frac{\hat{Q}_2^n(\cdot)}{\mu_3^n} 1_{\boldsymbol{B}(n,\cdot)^c} + \int_0^{\cdot} 1_{\boldsymbol{B}(n,s)} d\hat{I}_2^n(s)\right)(s)\\
&\qquad + \frac{1}{\sqrt{n}} \sup_{0\le s\le t} \frac{\hat{Q}_2^n(s)}{\mu_3^n} 1_{\boldsymbol{B}(n,s)^c}\\
&\quad\le \frac{2}{\sqrt{n}} \sup_{0\le s\le t} \left| \frac{\hat{X}_2^n(s)}{\mu_3^n} + \frac{\hat{X}_3^n(s)}{\mu_3^n} - \frac{\hat{Q}_2^n(s)}{\mu_3^n} 1_{\boldsymbol{B}(n,s)^c} + \int_0^s 1_{\boldsymbol{B}(n,u)} d\hat{I}_2^n(u)\right|\\
&\qquad + \frac{dl_0 \log n}{\mu_3^n n}\\
&\quad\le \frac{2}{\sqrt{n}} \sup_{0\le s\le t} \left| \frac{\hat{X}_2^n(s)}{\mu_3^n} + \frac{\hat{X}_3^n(s)}{\mu_3^n}\right| + \frac{2dl_0 \log n}{\mu_3^n n} + \frac{2}{\sqrt{n}} \int_0^t 1_{\boldsymbol{B}(n,u)} d\hat{I}_2^n(u)\\
&\qquad + \frac{dl_0 \log n}{\mu_3^n n}.
\end{aligned}
$$

From functional central limit theorem for renewal processes, $\hat{A}^n$ and $\hat{S}^n$ converge weakly to Brownian motions. Combining this with the fact that $\bar{T}_i^n(t) \le t$ for $n \ge 1$ and $t \ge 0$, we have for all $t \ge 0$,

$$
\frac{2}{\sqrt{n}} \sup_{0\le s\le t} \left| \frac{\hat{X}_2^n(s)}{\mu_3^n} + \frac{\hat{X}_3^n(s)}{\mu_3^n}\right| \to 0, \text{ in probability.} \tag{5.8}
$$

Next from Corollary 5.6 we have for $t \ge 0$, $\int_0^t 1_{\boldsymbol{B}(n,u)} d\hat{I}_2^n(u) \to 0$ in probability. Finally, since $\frac{dl_0 \log n}{\mu_3^n n} \to 0$ as $n \to \infty$, we have

$$
\sup_{0\le s\le t} \bar{W}_2^n(s) \to 0, \text{ in probability.} \tag{5.9}
$$

Next using the representation for $\hat{W}_1^n$ from (4.11), we have

$$
\hat{W}_1^n(t) - \Psi(\hat{W}_2^n(t)) = \frac{\hat{X}_1^n(t)}{\mu_1^n} + \frac{\hat{X}_2^n(t)}{\mu_2^n} + \hat{I}_1^n(t) - \Psi(\hat{W}_2^n(t)).
$$

Note that although one cannot assert that $\hat{W}_1^n(t) - \Psi(\hat{W}_2^n(t))$ is always nonnegative, letting $\boldsymbol{C}(n,t) = \{\hat{W}_1^n(t) - \Psi(\hat{W}_2^n(t)) < g_0/\sqrt{n}\}$, the process $(\hat{W}_1^n(t) - \Psi(\hat{W}_2^n(t)))1_{\boldsymbol{C}(n,t)^c}$ is clearly nonnegative. This motivates the following rewriting of the above expectation.

$$
(\hat{W}_1^n(t) - \Psi(\hat{W}_2^n(t)))1_{\boldsymbol{C}(n,t)^c}
$$

$$= \frac{\hat{X}_1^n(t)}{\mu_1^n} + \frac{\hat{X}_2^n(t)}{\mu_2^n} - \Psi(\hat{W}_2^n(t)) \tag{5.10}$$
$$- (\hat{W}_1^n(t) - \Psi(\hat{W}_2^n(t)))1_{\boldsymbol{C}(n,t)} + \hat{I}_1^n(t),$$

where $\boldsymbol{C}(n,t) = \{\hat{W}_1^n(t) - \Psi(\hat{W}_2^n(t)) < g_0/\sqrt{n}\}$. Note that the scheduling policy described in Definition 3.2 is such that $\hat{I}_1^n$ is equal 0 when $t = 0$, is non-decreasing, and increases only if the left hand side of the above equation is 0. Thus using the characterizing property of the one dimensional Skorohod map we have for $t \geq 0$,

$$\hat{W}_1^n(t) = \Gamma\left(\frac{\hat{X}_1^n(\cdot)}{\mu_1^n} + \frac{\hat{X}_2^n(\cdot)}{\mu_2^n} - \Psi(\hat{W}_2^n(\cdot)) - (\hat{W}_1^n(\cdot) - \Psi(\hat{W}_2^n(\cdot)))1_{\boldsymbol{C}(n,\cdot)}\right)(t) \tag{5.11}$$
$$+ (\hat{W}_1^n(t) - \Psi(\hat{W}_2^n(t)))1_{\boldsymbol{C}(n,t)} + \Psi(\hat{W}_2^n(t)).$$

By the Lipschitz property of Γ and Ψ (see Lemma 4.4 (ii)) we now have,

$$\sup_{0\leq s\leq t} \bar{W}_1^n(s) \leq \frac{2}{\sqrt{n}} \sup_{0\leq s\leq t} \left|\frac{\hat{X}_1^n(s)}{\mu_1^n} + \frac{\hat{X}_2^n(s)}{\mu_2^n}\right| + \frac{3}{\sqrt{n}} \sup_{0\leq s\leq t} \Psi(\hat{W}_2^n(s)) \tag{5.12}$$
$$+ \frac{3}{\sqrt{n}} \sup_{0\leq s\leq t} \left|(\hat{W}_1^n(s) - \Psi(\hat{W}_2^n(s)))1_{\boldsymbol{C}(n,s)}\right|$$
$$\leq \frac{2}{\sqrt{n}} \sup_{0\leq s\leq t} \left|\frac{\hat{X}_1^n(s)}{\mu_1^n} + \frac{\hat{X}_2^n(s)}{\mu_2^n}\right| + \frac{6\mu_3}{\mu_2} \sup_{0\leq s\leq t} \bar{W}_2^n(s) + \frac{3g_0}{n}.$$

As for (5.8), we have

$$\frac{2}{\sqrt{n}} \sup_{0\leq s\leq t} \left|\frac{\hat{X}_1^n(s)}{\mu_1^n} + \frac{\hat{X}_2^n(s)}{\mu_2^n}\right| \to 0 \text{ in probability.}$$

Using this along with (5.9), we now get

$$\sup_{0\leq s\leq t} \bar{W}_1^n(s) \to 0 \text{ as } n \to \infty. \tag{5.13}$$

From (5.9),(5.13) and (4.11), we have $\bar{Q}^n \Rightarrow 0$ as $n \to \infty$. Finally, using functional central limit theorem for renewal processes again, we have

$$n^{-1/2}\hat{A}_i^n(\cdot) \Rightarrow 0, \quad n^{-1/2}\hat{S}_j^n(\bar{T}_j^n(\cdot)) \Rightarrow 0, \quad i = 1,2, \ j = 1,2,3.$$

Hence, the conclusion follows from (4.1) and the fact that $\bar{Q}^n \Rightarrow 0$. □

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

The following theorem gives certain uniform integrability properties that will be needed to prove Theorem 5.2.

THEOREM 5.9. *For $i = 1, 2$,*

$$\limsup_{n\to\infty} \int_0^\infty e^{-\gamma t}\mathbb{E}\left[\sup_{0\le s\le t} \hat{W}_i^n(s)\right]^2 dt < \infty\,, \tag{5.14}$$

and

$$\limsup_{T\to\infty} \limsup_{n\to\infty} \int_T^\infty e^{-\gamma t}\mathbb{E}\left[\sup_{0\le s\le t} \hat{W}_i^n(s)\right]^2 dt = 0\,. \tag{5.15}$$

PROOF. We only prove (5.15). The proof of (5.14) is similar. For the case $i = 2$ the proof of (5.15) is identical to that of Theorem 4.11 of [4], with one modification. Unlike [4], here we consider arrival and service processes that are general renewal processes rather than Poisson processes. Thus one cannot directly apply Doob's maximal inequality to bound $\mathbb{E}\{\sup_{0\le s\le t} |\hat{A}_i^n(s)|\}^2$ and $\mathbb{E}\{\sup_{0\le s\le t} |\hat{S}_i^n(s)|\}^2$. However, these quantities can be bounded using Lorden's inequality as in [1] (see Equation (172) therein) and we omit the details. Consider now $i = 1$. From (5.12)

$$\sup_{0\le s\le t} \hat{W}_1^n(s) \le 2 \sup_{0\le s\le t} \left|\frac{\hat{X}_1^n(s)}{\mu_1^n} + \frac{\hat{X}_2^n(s)}{\mu_2^n}\right| + \frac{6\mu_3}{\mu_2} \sup_{0\le s\le t} \hat{W}_2^n(s) + \frac{3g_0}{\sqrt{n}}. \tag{5.16}$$

The result for $i = 1$ now follows from the above estimate along with the property (5.15) for $\hat{W}_2^n$. □

We now prove the main result of this section, Theorem 5.2.

PROOF OF THEOREM 5.2. Let $c, \bar{l}$ be as in Corollary 5.6 and consider the sequence $\{T^n\}$ with threshold parameters $l_0 \ge \bar{l}$, $g_0 > 0$ and c as above. From Lemma 5.8 and functional central limit theorem for renewal processes, it follows that

$$\hat{X}^n \Rightarrow X, \tag{5.17}$$

where X is as introduced below (4.7). Define processes $\hat{Z}_i^n$, $i = 1, 2$ as

$$\hat{Z}_1^n(t) \doteq \frac{\hat{X}_1^n(t)}{\mu_1^n} + \frac{\hat{X}_2^n(t)}{\mu_2^n} - \Psi(\hat{W}_2^n(t)) - (\hat{W}_1^n(t) - \Psi(\hat{W}_2^n(t)))1_{\boldsymbol{C}(n,t)},$$

$$\hat{Z}_2^n(t) \doteq \frac{\hat{X}_2^n(t)}{\mu_3^n} + \frac{\hat{X}_3^n(t)}{\mu_3^n} - \frac{\hat{Q}_2^n(t)}{\mu_3^n}1_{\boldsymbol{B}(n,t)^c} + \int_0^t 1_{\boldsymbol{B}(n,s)} d\hat{I}_2^n(s).$$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Also let $\mathcal{I} : D([0,\infty) : \mathbb{R}) \to D([0,\infty) : \mathbb{R})$ be the identity map. From (5.7) and (5.11) we have for $t \geq 0$,

$$\hat{I}_1^n(t) = (\Gamma - \mathcal{I})(\hat{Z}_1^n)(t), \tag{5.18}$$

$$\hat{I}_2^n(t) = (\Gamma - \mathcal{I})(\hat{Z}_2^n)(t) + \int_0^t 1_{\boldsymbol{B}(n,s)} d\hat{I}_2^n(s). \tag{5.19}$$

Next, from (4.7),

$$\frac{\hat{X}_2^n}{\mu_3^n} + \frac{\hat{X}_3^n}{\mu_3^n} \Rightarrow \frac{X_2}{\mu_3} + \frac{X_3}{\mu_3} = B_2, \quad \frac{\hat{X}_1^n}{\mu_1^n} + \frac{\hat{X}_2^n}{\mu_2^n} \Rightarrow \frac{X_1}{\mu_1} + \frac{X_2}{\mu_2} = B_1. \tag{5.20}$$

Applying (5.20) and the third convergence statement in Corollary 5.6 to (5.19), and recalling (W_2^*, I_2^*) defined in (4.25) and (4.27), we see that

$$(\hat{W}_2^n, \hat{I}_2^n) \Rightarrow (W_2^*, I_2^*). \tag{5.21}$$

Also from Lemma 5.7, for some $c_1, c_2 \in (0,\infty)$,

$$\begin{aligned} &|\hat{W}_1^n(t) - \Psi(\hat{W}_2^n(t))| 1_{\boldsymbol{C}(n,t)} \\ &\leq \frac{g_0}{\sqrt{n}} + \frac{\sqrt{n}}{\mu_1^n}\left|1 - \frac{\mu_3\mu_2^n}{\mu_3^n\mu_2}\right| \bar{Q}_1^n(t) + \frac{\sqrt{n}}{\mu_2^n}\left|1 - \frac{\mu_3\mu_2^n}{\mu_3^n\mu_2}\right| (2\bar{Q}_2^n(t) + \frac{1}{n}) \\ &\quad + \frac{c_1(C^n - L^n + 2)}{\sqrt{n}} + \frac{c_2}{\sqrt{n}}. \end{aligned} \tag{5.22}$$

From Assumption 2.2 we have

$$\sqrt{n}\left|1 - \frac{\mu_3\mu_2^n}{\mu_3^n\mu_2}\right| \to \frac{|\mu_3 b_3 - \mu_2 b_2|}{\mu_3},$$

and from Lemma 5.8, $\bar{Q}^n \Rightarrow 0$. Using these observations in (5.22) we see that

$$(\hat{W}_1^n(\cdot) - \Psi(\hat{W}_2^n(\cdot))) 1_{\boldsymbol{C}(n,\cdot)} \Rightarrow 0. \tag{5.23}$$

Applying the above result and (5.20), (5.21) to (5.18), we now get $(\hat{W}_1^n, \hat{I}_1^n) \Rightarrow (W_1^*, I_1^*)$. In fact we have shown $(\hat{W}^n, \hat{I}^n) \Rightarrow (W^*, I^*)$ in $D([0,\infty) : \mathbb{R}^4)$. This proves part (i) of the theorem. For the second part of the theorem, we observe that from Theorem 5.9 and first part of this theorem,

$$\int_0^\infty e^{-\gamma t}\mathbb{E}(\hat{W}_i^n(t))dt \to \int_0^\infty e^{-\gamma t}\mathbb{E}(W_i^*(t))dt, \quad i = 1, 2. \tag{5.24}$$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

From (4.11) we see that

$$\boldsymbol{A}(n,nt) = \left\{ \hat{Q}_3^n(t) - \frac{\mu_2^n}{\mu_1^n}\hat{Q}_1^n(t) < \frac{L^n}{\sqrt{n}} \right\} = \left\{ \mu_3^n \hat{W}_2^n(t) - \mu_2^n \hat{W}_1^n(t) < \frac{L^n}{\sqrt{n}} \right\}.$$

Combining (5.24) with Lemma 5.3 and using the definition of Q_2^* we now have exactly as in the proof of Theorem 4.2 of [4],

$$\begin{aligned} &\int_0^\infty e^{-\gamma t}\mathbb{E}(\mu_2^n \hat{W}_1^n(t) 1_{\boldsymbol{A}(n,nt)^c})dt + \int_0^\infty e^{-\gamma t}\mathbb{E}(\mu_3^n \hat{W}_2^n(t) 1_{\boldsymbol{A}(n,nt)})dt \\ &\to \int_0^\infty e^{-\gamma t}\mathbb{E}(Q_2^*(t))dt. \end{aligned} \tag{5.25}$$

Finally using (5.24), (5.25), (4.11) and Corollary 5.6 it immediately follows that

$$\int_0^\infty e^{-\gamma t}\mathbb{E}(\hat{Q}_i^n(t))dt \to \int_0^\infty e^{-\gamma t}\mathbb{E}(Q_i^*(t))dt, \quad i = 1,2,3. \tag{5.26}$$

The result now follows from the definitions of $\hat{J}^n(T^n)$ and $J^*(0)$. □

6. Proofs of Theorems 5.4 and 5.5, and Lemma 5.7. Throughout this section $\{T^n\}$ will denote the sequence of control policies in Definition 3.2 with some choice of threshold parameters. We begin with the following large deviations estimate for renewal processes, which will be extensively used in our proofs. In the form stated below, the result can be found in [2] (see also [1]).

LEMMA 6.1 (Lemma 6.7 of [2]). *Let $\{\eta_i\}_{i=1}^\infty$ be a sequence of independent strictly positive random variables, where $\{\eta_i\}_{i=2}^\infty$ are identically distributed with finite mean $1/\nu$, and η_1 may have a different distribution from $\eta_i, i \geq 2$. Assume that there is a nonempty open neighborhood $\mathcal{O}$ of $0 \in \mathbb{R}$ such that*

$$\Lambda(l) \doteq \log \mathbb{E}(e^{l\eta_i}) < \infty \text{ for all } l \in \mathcal{O} \text{ and } i \geq 2. \tag{6.1}$$

For each $n \in \mathbb{N}$, let $\nu^n > 0$ be such that $\lim_{n\to\infty} \nu^n = \nu$, and for each $n \in \mathbb{N}$ and $i = 1,2,3,\ldots$, let $\eta_i^n = \frac{\nu}{\nu^n}\eta_i$. Given $\epsilon \in (0, \nu/2)$, let $n(\{\nu^n\},\epsilon) \in \mathbb{N}$ be such that when $n \geq n(\{\nu^n\},\epsilon)$,

$$|\nu^n - \nu| < \epsilon, \qquad \frac{\nu^n}{\nu}\frac{1}{\nu^n + \frac{\epsilon}{2}} \leq \frac{1}{\nu}\frac{1}{1+\frac{\epsilon}{3\nu}} < \frac{1}{\nu},$$
$$\frac{1}{\nu}\left(1 + \frac{\epsilon}{2(\nu^n - \epsilon)}\right) \geq \frac{1}{\nu}\left(1 + \frac{\epsilon}{2\nu}\right) > \frac{1}{\nu}.$$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

For each $n \in \mathbb{N}$ *and* $t \geq 0$, *let*

$$N^n(t) = \sup\left\{k \geq 0 : \sum_{i=1}^{k} \eta_i^n \leq t\right\}.$$

Let

$$\Lambda^*(x) = \sup_{l\in\mathbb{R}}(lx - \Lambda(l)),\ x \in \mathbb{R},$$

be the Legendre-Fenchel transform of Λ. *Let* $\Theta_1(\nu,\epsilon) = \Lambda^*\left(\frac{1}{\nu}\frac{1}{1+\frac{\epsilon}{3\nu}}\right)$ *and* $\Theta_2(\nu,\epsilon) = \Lambda^*\left(\frac{1}{\nu}\left(1+\frac{\epsilon}{2\nu}\right)\right)$. *Then* $\Theta_i(\nu,\epsilon) > 0$ *for* $i = 1, 2$, *and for* $n \geq n(\{\nu^n\},\epsilon)$ *and* $t \geq 2/\epsilon$,

$$\begin{aligned}\mathbb{P}(N^n(t) > (\nu^n+\epsilon)t) &\leq \exp\left(-[(\nu^n+\epsilon)t - 1]\Theta_1(\nu,\epsilon)\right)\\ &\leq \exp\left(-(\nu t - 1)\Theta_1(\nu,\epsilon)\right),\end{aligned} \tag{6.2}$$

and for $n \geq n(\{\nu^n\},\epsilon)$ *and* $t \geq 0$,

$$\begin{aligned}\mathbb{P}(N^n(t) < (\nu^n-\epsilon)t) &\leq \exp\left(-(\nu^n-\epsilon)t\Theta_2(\nu,\epsilon)\right) + \mathbb{P}\left(\eta_1^n > \frac{\epsilon}{2\nu^n}t\right)\\ &\leq \exp\left(-(\nu-2\epsilon)t\Theta_2(\nu,\epsilon)\right) + \mathbb{P}\left(\eta_1^n > \frac{\epsilon}{2\nu^n}t\right),\end{aligned} \tag{6.3}$$

Furthermore, if η_1 *has the same distribution as* $\eta_i, i \geq 2$, *then for each* $n \geq 1, t \geq 0$, *and* $0 < p_0 \in \mathcal{O}$ *and any* $m \in \mathbb{N}$,

$$\mathbb{P}\left(\max_{i=1,\ldots,m}\eta_i^n > \frac{\epsilon}{2\nu^n}t\right) \leq m\exp\left(-\frac{p_0\epsilon t}{2\nu}\right)\exp(\Lambda(p_0)). \tag{6.4}$$

For $k = 1, 2$ and $j = 1, 2, 3$, denote by $\Lambda_{a,k}^*$ and $\Lambda_{s,j}^*$ the Legendre-Fenchel transform of $\Lambda_{a,k}$ and $\Lambda_{s,j}$, respectively, where the latter functions are as introduced in Assumption 2.3. These transforms along with Lemma 6.1 applied with sequences $\{v_j(i)/\mu_j\}_{i\in\mathbb{N}}$, $\{u_k(i)/\lambda_k\}_{i\in\mathbb{N}}$, $j = 1, 2, 3$, $k = 1, 2$, will play a key role in the proofs of Theorems 5.4 and 5.5.

Let $\{\mathcal{G}_1^n(t)\}_{t\geq 0}$ be the filtration in Section 2.2 associated with the sequence $\{T^n\}$. We next introduce a family of $\{\mathcal{G}_1^n(t)\}$-stopping times as follows: For $k \in \mathbb{N}$,

$$\begin{aligned}\tau_0^n &\doteq 0,\\ \tau_{2k-1}^n &\doteq \inf\left\{t > \tau_{2k-2}^n : Q_3^n(t) - \frac{\mu_2^n}{\mu_1^n}Q_1^n(t) \geq L^n\right\},\\ \tau_{2k}^n &\doteq \inf\left\{t > \tau_{2k-1}^n : Q_3^n(t) - \frac{\mu_2^n}{\mu_1^n}Q_1^n(t) < L^n, Q_3^n(t) < C^n - 1\right\}.\end{aligned} \tag{6.5}$$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

6.1. *Proof of Theorem 5.4.* The proof is rather long and thus to guide the reader through the arguments, we divide the proof into six steps with an outline at the beginning of each step. Large deviation estimates are key in Steps 2, 4 and 5.

Step 1. *We show that for* $s \in [\tau^n_{2k-2}, \tau^n_{2k-1}), k \in \mathbb{N}$, (6.6) *and* (6.7) *hold. These equations show that, in order to estimate* $\mathbb{P}(\mathcal{E}(n,t))$*, it suffices to focus on* $[\tau^n_{2k-1}, \tau^n_{2k}), k \in \mathbb{N}$.

Note that on $[\tau^n_{2k-2}, \tau^n_{2k-1}), k \in \mathbb{N}$, $Q^n_3(\cdot)$ starts from below $C^n - 1$, and whenever $Q^n_3(\cdot)$ becomes larger than or equal to $C^n - 1$, Server 1 stops serving Buffer 2, which causes $Q^n_3(\cdot)$ to decrease. Thus we have

$$Q^n_3(t) \le C^n, \text{ for all } t \in [\tau^n_{2k-2}, \tau^n_{2k-1}), \text{ and } k \in \mathbb{N}.$$

Since $\kappa \ge \frac{c}{c-1}$, $C^n \le \kappa(C^n - L^n + 1)$ and so we have for $s \in [\tau^r_{2k-2}, \tau^n_{2k-1})$, $k \in \mathbb{N}$,

$$Q^n_3(s)1_{\boldsymbol{A}(n,s)} \le \kappa(C^n - L^n + 1). \tag{6.6}$$

We also note that for $s \in [\tau^r_{2k-2}, \tau^n_{2k-1}), k \in \mathbb{N}$, $Q^n_3(s) - \frac{\mu^n_2}{\mu^n_1} Q^n_1(s) < L^n$. Thus

$$Q^n_1(s)1_{\boldsymbol{A}(n,s)^c} = 0. \tag{6.7}$$

In view of (6.6) and (6.7), to estimate $\mathbb{P}(\mathcal{E}(n,t))$, it now suffices to focus on $[\tau^n_{2k-1}, \tau^n_{2k}), k \in \mathbb{N}$.

Step 2. *We show that with high probability there are at most* $\mathbf{k}^n$ *(which is defined in* (6.8)*) intervals* $[\tau^n_{2k-1}, \tau^n_{2k}), k = 1, \ldots, \mathbf{k}^n$*, in* $[0, nt]$*. We can then bound* $\mathbb{P}(\mathcal{E}(n,t))$ *from above by the sum of the right sides of* (6.9) *and* (6.10).

Note that each τ^n_{2k-1} corresponds to a up-crossing of $Q^n_3(\cdot) - \frac{\mu^n_2}{\mu^n_1} Q^n_1(\cdot)$ from smaller than L^n to become larger than or equal to L^n. Each up-crossing requires at least one service completion of Server 1. Let

$$\mathbf{k}^n = \lfloor nt(\lambda^n_1 + \lambda^n_2 + 2) \rfloor + 1. \tag{6.8}$$

Then we have that

$$\begin{aligned}
\mathbb{P}(\tau^n_{2\mathbf{k}^n - 1} \le nt) &\le \mathbb{P}(S^n_1(T^n_1(nt)) + S^n_2(T^n_2(nt)) \ge \mathbf{k}^n) \\
&\le \mathbb{P}(A^n_1(nt) + A^n_2(nt) \ge \mathbf{k}^n) \\
&\le \mathbb{P}(A^n_1(nt) + A^n_2(nt) \ge nt(\lambda^n_1 + \lambda^n_2 + 2)) \\
&\le \mathbb{P}(A^n_1(nt) \ge nt(\lambda^n_1 + 1)) + \mathbb{P}(A^n_2(nt) \ge nt(\lambda^n_2 + 1)).
\end{aligned}$$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

From (6.2) in Lemma 6.1, when $n \geq \max\{n(\{\lambda_1^n\},1), n(\{\lambda_2^n\},1)\}$ and $nt \geq 2$, we have

$$\begin{aligned}\mathbb{P}(\tau^n_{2\mathbf{k}^n-1} \leq nt) \leq \exp\left\{-(\lambda_1 nt - 1)\Theta_1^{a,1}(\lambda_1,1)\right\} \\ + \exp\left\{-(\lambda_2 nt - 1)\Theta_1^{a,2}(\lambda_2,1)\right\},\end{aligned} \tag{6.9}$$

where $\Theta_1^{a,i}$, $i = 1,2$, are defined as Θ_1 in Lemma 6.1 on replacing Λ^* with $\Lambda^*_{a,i}$. Also,

(6.10)

$$\begin{aligned}&\mathbb{P}(\tau^n_{2\mathbf{k}^n-1} > nt,\ \mathcal{E}(n,t)) \\ &\leq \sum_{k=1}^{\mathbf{k}^n} \mathbb{P}\left(\sup_{s\in[\tau^n_{2k-1},\tau^n_{2k}\wedge nt]} Q_3^n(s)1_{\boldsymbol{A}(n,s)} > \kappa(C^n - L^n + 1), \tau^n_{2k-1} \leq nt\right) \\ &\quad + \sum_{k=1}^{\mathbf{k}^n} \mathbb{P}\left(\sup_{s\in[\tau^n_{2k-1},\tau^n_{2k}\wedge nt]} Q_1^n(s)1_{\boldsymbol{A}(n,s)^c} > \kappa(C^n - L^n + 1), \tau^n_{2k-1} \leq nt\right) \\ &\leq \sum_{k=1}^{\mathbf{k}^n} \mathbb{P}(Q_3^n(s) > \kappa(C^n - L^n + 1) \text{ and } \boldsymbol{A}(n,s), \\ &\qquad\qquad \text{for some } s \in [\tau^n_{2k-1}, \tau^n_{2k} \wedge nt]) \\ &\quad + \sum_{k=1}^{\mathbf{k}^n} \mathbb{P}(Q_1^n(s) > \kappa(C^n - L^n + 1) \text{ and } \boldsymbol{A}(n,s)^c, \\ &\qquad\qquad \text{for some } s \in [\tau^n_{2k-1}, \tau^n_{2k} \wedge nt]).\end{aligned}$$

Step 3. *For each $k \in \mathbb{N}$, we define a sequence of stopping times $\{\eta_l^{n,k}\}_{l\geq 0}$ (see (6.11)) within $[\tau^n_{2k-1}, \tau^n_{2k})$. Statements in (6.13) and (6.14) are established from which it follows that in order to estimate $\mathbb{P}(\mathcal{E}(n,t))$ we only need to consider $[\eta^{n,k}_{2l-1}, \eta^{n,k}_{2l}]$.*

Let

$$\boldsymbol{H}(n,t) = \left\{Q_1^n(t) \geq \frac{\mu_1^n}{\mu_2^n}(C^n - L^n + 2)\right\}, \quad \boldsymbol{G}(n,t) = \{Q_3^n(t) \geq (C^n - 1)\}.$$

For $l \in \mathbb{N}$,

$$\begin{aligned}\eta_0^{n,k} &\doteq \tau^n_{2k-1}, \\ \eta^{n,k}_{2l-1} &\doteq \tau^n_{2k} \wedge \inf\{t \geq \eta^{n,k}_{2l-2} : (\boldsymbol{A}(n,t) \cap \boldsymbol{G}(n,t)) \\ &\qquad\qquad \text{or } (\boldsymbol{A}(n,t)^c \cap \boldsymbol{H}(n,t))\}, \\ \eta^{n,k}_{2l} &= \tau^n_{2k} \wedge \inf\{t \geq \eta^{n,k}_{2l-1} : \boldsymbol{A}(n,t)^c \cap \boldsymbol{H}(n,t)^c\}.\end{aligned} \tag{6.11}$$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Note that

(6.12)
$$\begin{aligned}&\{\omega : t \in [\eta_{2l-2}^{n,k}, \eta_{2l-1}^{n,k})\} \subset \\ &\{\boldsymbol{A}(n,t) \cap \boldsymbol{G}(n,t)\}^c \cap \{\boldsymbol{A}(n,t)^c \cap \boldsymbol{H}(n,t)\}^c \cap \{\boldsymbol{A}(n,t) \cap \boldsymbol{G}(n,t)^c\}^c \\ &\quad = \boldsymbol{A}(n,t)^c \cap \boldsymbol{H}(n,t)^c.\end{aligned}$$

Also, there exists $n_1 \in \mathbb{N}$ such that when $n \geq n_1$, we have $\frac{2\mu_1}{\mu_2} > \frac{\mu_1^n}{\mu_2^n}$. Since $\kappa \geq \frac{4\mu_1}{\mu_2}$, we have from (6.12) that when $n \geq n_1$,

$$\left\{ \sup_{s \in [\eta_{2l-2}^{n,k}, \eta_{2l-1}^{n,k})} Q_3^n(s) 1_{\boldsymbol{A}(n,s)} > \kappa(C^n - L^n + 1) \right\} = \emptyset, \tag{6.13}$$

$$\left\{ \sup_{s \in [\eta_{2l-2}^{n,k}, \eta_{2l-1}^{n,k})} Q_1^n(s) 1_{\boldsymbol{A}(n,s)^c} > \kappa(C^n - L^n + 1) \right\} = \emptyset. \tag{6.14}$$

Thus to estimate $\mathbb{P}(\mathcal{E}(n,t))$, it suffices to consider the intervals $[\eta_{2l-1}^{n,k}, \eta_{2l}^{n,k})$, $l, k \in \mathbb{N}$.

Step 4. *We show that with high probability there are at most* $\mathbf{k}^n$ *(which is defined in* (6.8)*) intervals* $[\eta_{2l-1}^{n,k}, \eta_{2l}^{n,k}), l = 1, \ldots, \mathbf{k}^n$ *in* $[\tau_{2k-1}^n, \tau_{2k}^n \wedge nt)$. *Additional estimates then show that,* $\mathbb{P}(\mathcal{E}(n,t))$ *can be estimated as in* (6.17).

We now estimate how many such subintervals are within $[\tau_{2k-1}^n, \tau_{2k}^n \wedge nt)$. We observe that each $\eta_{2l-1}^{n,k}$ corresponds to at least one additional arrival to Buffer 1 or one additional job completion for Buffer 3. Recall $\mathbf{k}^n$ defined in (6.8). As in the proof of (6.9) we have, for all $n \geq \max\{n(\{\lambda_1^n\}, 1), n(\{\lambda_2^n\}, 1)\}$ and $nt > 2$,

$$\begin{aligned}&\mathbb{P}(\eta_{2\mathbf{k}^n-1}^{n,k} \leq nt) \\ &\quad \leq \mathbb{P}(A_1^n(nt) + S_2^n(T_2^n(nt)) \geq \mathbf{k}^n) \\ &\quad \leq \mathbb{P}(A_1^n(nt) + A_2^n(nt) \geq \mathbf{k}^n) \\ &\quad \leq \exp\left\{-(\lambda_1 nt - 1)\Theta_1^{a,1}(\lambda_1, 1)\right\} + \exp\left\{-(\lambda_2 nt - 1)\Theta_1^{a,2}(\lambda_2, 1)\right\}.\end{aligned} \tag{6.15}$$

Thus from (6.9), (6.10) and (6.15) we have for $n \geq \max\{n(\{\lambda_1^n\}, 1), n(\{\lambda_2^n\}, 1), n_1\}$ and $nt \geq 2$,

$$\begin{aligned}\mathbb{P}(\mathcal{E}(n,t)) \leq &\sum_{k=1}^{\mathbf{k}^n} \sum_{l=1}^{\mathbf{k}^n} \mathbb{P}(Q_3^n(s) > \kappa(C^n - L^n + 1) \text{ and } \boldsymbol{A}(n,s), \\ &\qquad \text{for some } s \in [\eta_{2l-1}^{n,k}, \eta_{2l}^{n,k} \wedge nt)) \\ &+ \sum_{k=1}^{\mathbf{k}^n} \sum_{l=1}^{\mathbf{k}^n} \mathbb{P}(Q_1^n(s) > \kappa(C^n - L^n + 1) \text{ and } \boldsymbol{A}(n,s)^c,\end{aligned}$$

$$\text{for some } s \in [\eta^{n,k}_{2l-1}, \eta^{n,k}_{2l} \wedge nt))$$
$$+ (\mathbf{k}^n + 1) \exp\left\{-(\lambda_1 nt - 1)\Theta^{a,1}_1(\lambda_1, 1)\right\}$$
$$+ (\mathbf{k}^n + 1) \exp\left\{-(\lambda_2 nt - 1)\Theta^{a,2}_1(\lambda_2, 1)\right\}.$$

Next, on the set $\{Q^n_3(s) > \kappa(C^n - L^n + 1) \text{ and } \boldsymbol{A}(n,s)\}$

$$\begin{aligned} Q^n_1(s) &> \frac{\mu^n_1}{\mu^n_2}(Q^n_3(s) - L^n) > \frac{\mu^n_1}{\mu^n_2}(\kappa(C^n - L^n + 1) - L^n) \\ &= \frac{\mu^n_1}{\mu^n_2}\frac{3\kappa}{4}(C^n - L^n + 1) + \frac{\mu^n_1}{\mu^n_2}\left(\frac{\kappa}{4}(C^n - L^n + 1) - L^n\right) \\ &\geq \frac{\mu^n_1}{\mu^n_2}\frac{3\kappa}{4}(C^n - L^n + 1) + \frac{\mu^n_1}{\mu^n_2}\left(\frac{\kappa}{4}(c-1) - 1\right) l_0 \log n \\ &> \frac{\mu^n_1}{\mu^n_2}\frac{3\kappa}{4}(C^n - L^n + 1), \end{aligned}$$

where the last inequality follows on noting that $\kappa > \frac{4}{c-1}$. Let $n_2 \in \mathbb{N}$ be such that when $n \geq n_2$, we have $\frac{2\mu_1}{3\mu_2} \leq \frac{\mu^n_1}{\mu^n_2}$. Letting $\kappa' = \min\{\kappa, \frac{\kappa}{2}\frac{\mu_1}{\mu_2}\}$, we have that, when $n \geq n_2$,

$$Q^n_1(s) > \kappa'(C^n - L^n + 1).$$

Note that since $\kappa \geq \max\{\frac{2\mu_1}{\mu_2}, 4\}$, we have that

$$\kappa' \geq \frac{2\mu_1}{\mu_2}. \tag{6.16}$$

Since $\kappa \geq \kappa'$, for $n \geq \max\{n(\{\lambda^n_1\}, 1), n(\{\lambda^n_2\}, 1), n_1, n_2\}$ and $nt \geq 2$,

$$\begin{aligned} \mathbb{P}(\mathcal{E}(n,t)) \leq 2\sum_{k=1}^{k^n}\sum_{l=1}^{k^n} \mathbb{P}\Big(& Q^n_1(s) > \kappa'(C^n - L^n + 1), \\ & \text{for some } s \in [\eta^{n,k}_{2l-1}, \eta^{n,k}_{2l} \wedge nt) \Big) \\ & + (\mathbf{k}^n + 1) \exp\left\{-(\lambda_1 nt - 1)\Theta^{a,1}_1(\lambda_1, 1)\right\} \\ & + (\mathbf{k}^n + 1) \exp\left\{-(\lambda_2 nt - 1)\Theta^{a,2}_1(\lambda_2, 1)\right\}. \end{aligned} \tag{6.17}$$

Step 5. *We estimate each summand in the double summation in* (6.17) *by analyzing a* $GI/GI/1$ *queue, and show that it can be bounded by the right side of* (6.21) *(see* (6.22)*).*

Note that

$$\{\boldsymbol{A}(n,s)^c \cap \boldsymbol{H}(n,s)^c\}^c \cap \{\boldsymbol{A}(n,s) \cap \boldsymbol{G}(n,s)^c\}^c$$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

$$= \{\boldsymbol{A}(n,s) \cap \boldsymbol{G}(n,s)\} \cup \{\boldsymbol{H}(n,s) \cap \boldsymbol{G}(n,s)\} \tag{6.18}$$
$$\cup \{\boldsymbol{H}(n,s) \cap \boldsymbol{A}(n,s)^c\}.$$

From this, we see that for $s \in [\eta_{2l-1}^{n,k}, \eta_{2l}^{n,k})$, $Q_1^n(s) \geq \frac{\mu_1^n}{\mu_2^n}(C^n - L^n - 1) > 0$, and Server 1 works on Buffer 1 continuously. Also, from (6.12), we have $Q_1^n(\eta_{2l-1}^{n,k}-) < \frac{\mu_1^n}{\mu_2^n}(C^n - L^n + 2)$, and so $Q_1^n(\eta_{2l-1}^{n,k}) < \lfloor \frac{\mu_1^n}{\mu_2^n}(C^n - L^n + 2) \rfloor + 1$. Using this fact, the probability in (6.17) can be estimated by analyzing a $GI/GI/1$ queue with interarrival times $\tilde{u}_1^n, u_1^n(k), k = 2, 3, \ldots$, service times $\tilde{v}_1^n, v_1^n(k), k = 2, 3, \ldots$, and initial queue length $\lfloor \frac{\mu_1^n}{\mu_2^n}(C^n - L^n + 2) \rfloor + 1$, where $\tilde{u}_1^n$ and $\tilde{v}_1^n$ are residual interarrival and service times at time $\eta_{2l-1}^{n,k}$ in Buffer 1. Let $\{\mathcal{Q}^n(t)\}_{t \geq 0}$ be this $GI/GI/1$ queue length process and define

$$\beta^n = \inf\left\{ s > 0 : \mathcal{Q}^n(s) < \frac{\mu_1^n}{\mu_2^n}(C^n - L^n - 1) \right\}.$$

Noting that when $n \geq n_1$, $\frac{2\mu_1}{\mu_2} \geq \frac{\mu_1^n}{\mu_2^n}$ and using (6.16), we see that

$$\mathbb{P}\left(Q_1^n(s) > \kappa'(C^n - L^n + 1) \text{ for some } s \in [\eta_{2l-1}^{n,k}, \eta_{2l}^{n,k} \wedge nt] \right) \tag{6.19}$$
$$\leq \mathbb{P}\left(\mathcal{Q}^n(s) > \kappa'(C^n - L^n + 1) \text{ for some } s \in [0, \beta^n] \right).$$

Let $\epsilon \in (0, (\mu_1 - \lambda_1)/4)$, and define

$$t^n = \frac{\left\lfloor \frac{\mu_1^n}{\mu_2^n}(C^n - L^n - 1) - 2 \right\rfloor}{2(\lambda_1^n + \epsilon)}$$

Let $\tilde{A}^n$ and $\tilde{S}^n$ be the arrival and service processes of the $GI/GI/1$ queue, and consider

$$\mathcal{H}^n \doteq \left\{ \tilde{A}^n(t^n) < (\lambda_1^n + \epsilon)t^n,\ \tilde{S}^n(t^n) > (\mu_1^n - \epsilon)t^n \right\}.$$

From (6.2) and (6.3) in Lemma 6.1, for $n \geq \max\{n(\{\lambda_1^n\}, \epsilon), n(\{\mu_1^n\}, \epsilon)\}$, and $t^n \geq 2/\epsilon$, we have

$$\begin{aligned}
\mathbb{P}((\mathcal{H}^n)^c) &\leq \mathbb{P}(\tilde{A}^n(t^n) \geq (\lambda_1^n + \epsilon)t^n) + \mathbb{P}(\tilde{S}^n(t^n) \leq (\mu_1^n - \epsilon)t^n) \\
&\leq \exp\left(-(\lambda_1 t^n - 1)\Theta_1^{a,1}(\lambda_1, \epsilon) \right) + \exp\left(-(\mu_1 - 2\epsilon)t^n \Theta_2^{s,1}(\mu_1, \epsilon) \right) \\
&\quad + \mathbb{P}\left(\tilde{v}_1^n > \frac{\epsilon t^n}{2\mu_1^n} \right),
\end{aligned}$$

where $\Theta_2^{s,1}$ is defined as Θ_2 in Lemma 6.1 by replacing Λ^* with $\Lambda_{s,1}^*$.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Next using (6.2) and (6.4) in Lemma 6.1,

$$
\begin{aligned}
&\mathbb{P}\left(\tilde{v}_1^n > \frac{\epsilon t^n}{2\mu_1^n}\right) \le \mathbb{P}\left(\max_{k=1,\dots,S_1^n(nt)} v_1^n(k) > \frac{\epsilon t^n}{2\mu_1^n}\right) \\
&\le \mathbb{P}\left(S_1^n(nt) > (\mu_1^n+\epsilon)nt\right) + \mathbb{P}\left(\max_{k=1,\dots,\lfloor(\mu_1^n+\epsilon)nt\rfloor} v_1^n(k) > \frac{\epsilon t^n}{2\mu_1^n}\right) \\
&\le \exp\left(-(\mu_1 nt - 1)\Theta_1^{s,1}(\mu_1,\epsilon)\right) \\
&\quad + (\mu_1^n+\epsilon)nt \exp\left(-\frac{p_0\epsilon t^n}{2\mu_1}\right)\exp(\Lambda_{s,1}(p_0)),
\end{aligned} \tag{6.20}
$$

where $\Theta_1^{s,1}$ is defined as Θ_1 in Lemma 6.1 by replacing Λ^* with $\Lambda_{s,1}^*$ and $0 < p_0 \in \mathcal{O}$. Thus for $n \ge \max\{n(\{\lambda_1^n\},\epsilon), n(\{\mu_1^n\},\epsilon)\}$, and $t^n \ge 2/\epsilon$, we have

$$
\begin{aligned}
&\mathbb{P}((\mathcal{H}^n)^c) \\
&\le \exp\left(-(\lambda_1 t^n - 1)\Theta_1^{a,1}(\lambda_1,\epsilon)\right) + \exp\left(-(\mu_1-2\epsilon)t^n\Theta_2^{s,1}(\mu_1,\epsilon)\right) \\
&\quad + \exp\left(-(\mu_1 nt - 1)\Theta_1^{s,1}(\mu_1,\epsilon)\right) \\
&\quad + (\mu_1^n+\epsilon)nt \exp\left(-\frac{p_0\epsilon t^n}{2\mu_1}\right)\exp(\Lambda_{s,1}(p_0)).
\end{aligned} \tag{6.21}
$$

Next note that on $\mathcal{H}^n$,

$$
\begin{aligned}
\mathcal{Q}^n(t^n) &= \mathcal{Q}^n(0) + \tilde{A}^n(t^n) - \tilde{S}^n(t^n) \\
&< \frac{\mu_1^n}{\mu_2^n}(C^n - L^n + 2) + 1 + (\lambda_1^n - \mu_1^n + 2\epsilon)t^n \\
&= \frac{\mu_1^n}{\mu_2^n}(C^n - L^n + 2) + 1 - (\mu_1^n - \lambda_1^n - 2\epsilon)\frac{\left\lfloor \frac{\mu_1^n}{\mu_2^n}(C^n - L^n - 1) - 2\right\rfloor}{2(\lambda_1^n+\epsilon)}.
\end{aligned}
$$

Recall that $\epsilon \in (0, (\mu_1 - \lambda_1)/4)$. Thus there exists $n_3 \in \mathbb{N}$ such that when $n \ge n_3$, we have

$$
\begin{aligned}
&\mathcal{Q}^n(t^n) - \frac{\mu_1^n}{\mu_2^n}(C^n - L^n - 1) \\
&< \frac{3\mu_1^n}{\mu_2^n} + 1 - (\mu_1^n - \lambda_1^n - 2\epsilon)\frac{\left\lfloor \frac{\mu_1^n}{\mu_2^n}(C^n - L^n - 1) - 2\right\rfloor}{2(\lambda_1^n+\epsilon)} \\
&< 0.
\end{aligned}
$$

Since $\mathcal{Q}^n(t) \ge \frac{\mu_1^n}{\mu_2^n}(C^n - L^n - 1) - 1$ for all $t \in [0, \beta^n]$, we see that for all $n \ge n_3$, on $\mathcal{H}^n$, $t^n \ge \beta^n$. Next, on this set, for $s \in [0,\beta^n]$ and $n \ge n_3$, we

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

have

$$\begin{aligned}\mathcal{Q}^n(s) &= \mathcal{Q}^n(0) + \tilde{A}^n(s) - \tilde{S}^n(s)\\ &\le \mathcal{Q}^n(0) + \tilde{A}^n(s)\\ &< \frac{\mu_1^n}{\mu_2^n}(C^n - L^n + 2) + 1 + (\lambda_1^n + \epsilon)t^n\\ &\le \frac{\mu_1^n}{\mu_2^n}(C^n - L^n + 2) + \frac{\mu_1^n}{2\mu_2^n}(C^n - L^n - 1)\\ &= \frac{3\mu_1^n}{2\mu_2^n}(C^n - L^n + 1).\end{aligned}$$

Recall from (6.16) that $\kappa' > \frac{2\mu_1}{\mu_2}$, and so there exists $n_4 \ge n_3$ such that when $n \ge n_4$, we have $\kappa' > \frac{3\mu_1^n}{2\mu_2^n}$, and so for such n, on $\mathcal{H}^n$,

$$\mathcal{Q}^n(s) < \kappa'(C^n - L^n + 1) \quad \text{for all } s \in [0, \beta^n].$$

Let $\mathcal{I}^n = \{\mathcal{Q}^n(s) > \kappa'(C^n - L^n + 1) \text{ for some } s \in [0, \beta^n]\ \}$. Then for $n \ge n_4$, $\mathbb{P}(\mathcal{H}^n \cap \mathcal{I}^n) = 0$, and so for $n \ge \max\{n(\{\lambda_1^n\}, \epsilon), n(\{\lambda_2^n\}, \epsilon), n_1, n_4\}$,

(6.22) the probability in (6.19) is bounded by $\mathbb{P}((\mathcal{H}^n)^c)$.

Step 6. *We combine the key estimates to complete the proof.*

Combining (6.22) with (6.17) and (6.21), we finally have for sufficiently large n and $t \ge 2/n$,

$$\begin{aligned}\mathbb{P}(\mathcal{E}(n,t)) \le\ & (\mathbf{k}^n + 1)\exp\left\{-(\lambda_1 nt - 1)\Theta_1^{a,1}(\lambda_1, 1)\right\}\\ &+ (\mathbf{k}^n + 1)\exp\left\{-(\lambda_2 nt - 1)\Theta_1^{a,2}(\lambda_2, 1)\right\}\\ &+ 2(\mathbf{k}^n)^2\exp\left(-(\lambda_1 t^n - 1)\Theta_1^{a,1}(\lambda_1, \epsilon)\right)\\ &+ 2(\mathbf{k}^n)^2\exp\left(-(\mu_1 - 2\epsilon)t^n\Theta_2^{s,1}(\mu_1, \epsilon)\right)\\ &+ 2(\mathbf{k}^n)^2\exp\left(-(\mu_1 nt - 1)\Theta_1^{s,1}(\mu_1, \epsilon)\right)\\ &+ 2(\mathbf{k}^n)^2(\mu_1^n + \epsilon)nt\exp\left(-\frac{p_0\epsilon t^n}{2\mu_1}\right)\exp(\Lambda_{s,1}(p_0)).\end{aligned} \tag{6.23}$$

Thus we have shown that, there exist $\theta_1, \theta_2, \theta_3, \theta_4 \in (0, \infty)$ such that for sufficiently large n and $t \ge 2/n$, we have

$$\mathbb{P}(\mathcal{E}(n,t)) \le \theta_1(nt+1)^2\exp\{-\theta_2(nt+1)\} + \theta_3(nt+1)^3 n^{-\theta_4(c-1)l_0}. \quad \square$$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

6.2. *Proof of Theorem 5.5.* We first introduce a sequence of $\{\mathcal{G}_1^n(t)\}_{t\geq 0}$ stopping times, which will be used in the proof of Theorem 5.5. First recall the stopping times $\tau_l^n, l \in \mathbb{N}_0$ in (6.5). As in [4], for $k \in \mathbb{N}$, we define a sequence of stopping times within $[\tau_{2k-2}^n, \tau_{2k-1}^n)$. For $d, l_0 > 0$, let $D^n = dl_0 \log n$. For $m \in \mathbb{N}$,

$$
\begin{aligned}
\tilde{\eta}_0^{n,k} &\doteq \tau_{2k-2}^n, \\
\tilde{\eta}_{2m-1}^{n,k} &\doteq \min\left\{\tau_{2k-1}^n, \inf\{s > \tilde{\eta}_{2m-2}^{n,k} | Q_3^n(s) \geq C^n - 1\}\right\} \\
\tilde{\eta}_{2m}^{n,k} &\doteq \min\left\{\tau_{2k-1}^n, \inf\{s > \tilde{\eta}_{2m-1}^{n,k} | Q_3^n(s) < C^n - 1\}\right\} \\
\beta_m^{n,k} &\doteq \min\left\{\tilde{\eta}_{2m-1}^{n,k}, \inf\{s > \tilde{\eta}_{2m-2}^{n,k} | Q_2^n(s) \geq \frac{D^n}{4}\}\right\}.
\end{aligned} \tag{6.24}
$$

Recall the multi-parameter filtration $\{\mathcal{G}^n(a,b) : a \in \mathbb{N}^2, b \in \mathbb{N}^3\}$ and multi-parameter stopping times introduced in Section 2.2. Lemma 6.2 below is taken from [1].

LEMMA 6.2 (Lemma 7.6 of [1]). *Let* $\mathbf{T} = (\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3, \mathbf{T}_4, \mathbf{T}_5)'$ *be a* $\{\mathcal{G}^n(a,b) : a \in \mathbb{N}^2, b \in \mathbb{N}^3\}$ *multi-parameter stopping time. Then*

$$(u_1^n(\mathbf{T}_1), u_2^n(\mathbf{T}_2), v_1^n(\mathbf{T}_3), v_2^n(\mathbf{T}_4), v_3^n(\mathbf{T}_5))' \in \mathcal{G}_{\mathbf{T}}^n,$$

and on $\{\mathbf{T} \in \mathbb{N}^5\}$ *the conditional distribution of* $\{(u_1^n(\mathbf{T}_1 + k), u_2^n(\mathbf{T}_2 + k), v_1^n(\mathbf{T}_3 + k), v_2^n(\mathbf{T}_4 + k), v_3^n(\mathbf{T}_5 + k))' : k \in \mathbb{N}\}$ *given* $\mathcal{G}_{\mathbf{T}}^n$ *is the same as the distribution of* $\{(u_1^n(k), u_2^n(k), v_1^n(k), v_2^n(k), v_3^n(k))' : k \in \mathbb{N}\}$.

The following lemma follows along the lines of Lemma 7.5 of [1]. The proof is omitted.

LEMMA 6.3. *For* $n, k, m \in \mathbb{N}$,

$$\mathbf{T}_m^{n,k} \doteq (A_1^n(\beta_m^{n,k}), A_2^n(\beta_m^{n,k}), S_1^n(\beta_m^{n,k}), S_2^n(\beta_m^{n,k}), S_3^n(\beta_m^{n,k}))'$$

is a $\{\mathcal{G}^n(a,b) : a \in \mathbb{N}^2, b \in \mathbb{N}^3\}$ *multiparameter stopping time.*

For a time interval $[t_1, t_2)$, define

$$\mathcal{R}(t_1, t_2) = \left\{\int_{[t_1,t_2)} 1_{\{Q_2^n(s) \geq D^n\}} dI_2^n(s) \neq 0\right\}. \tag{6.25}$$

Then the LHS of (5.4) equals to $\mathbb{P}(\mathcal{R}(0, nt))$.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

PROOF OF THEOREM 5.5. Once again, since the proof is long, we divide it into five steps outlining the content at the beginning of each step. Large deviation estimates are used in Steps 2,3 and 4. Proofs of Steps 3 and 4 are the most technical parts of the argument.

Step 1. *In (6.26) we show that* $\mathcal{R}(\tau^n_{2k-1}, \tau^n_{2k}) = \emptyset$. *Furthermore, for* $[\tilde{\eta}^{n,k}_{2m-1}, \tilde{\eta}^{n,k}_{2m}) \subset [\tau^n_{2k-2}, \tau^n_{2k-1})$, $k, m \in \mathbb{N}$, *we show that* $\mathcal{R}(\tilde{\eta}^{n,k}_{2m-1}, \tilde{\eta}^{n,k}_{2m}) = \emptyset$. *From these it will follow that it suffices to consider* $\mathcal{R}(\tilde{\eta}^{n,k}_{2m-2}, \tilde{\eta}^{n,k}_{2m-1})$ *(see (6.28)).*

From the definition of stopping times in (6.5) we see that, for $s \in [\tau^n_{2k-1}, \tau^n_{2k})$, $k \in \mathbb{N}$,

$$Q^n_3(s) - \frac{\mu^n_2}{\mu^n_1} Q^n_1(s) \geq L^n, \text{ or } Q^n_3(s) \geq C^n - 1,$$

and therefore for such s, $Q^n_3(s) > 0$. Consequently, since Server 2 does not idle unless Buffer 3 is empty, we have

$$\int_{[\tau^n_{2k-1}, \tau^n_{2k})} 1_{\{Q^n_2(s) \geq D^n\}} dI^n_2(s) = 0. \tag{6.26}$$

Thus we only need to consider time intervals $[\tau^n_{2k-2}, \tau^n_{2k-1}), k \in \mathbb{N}$. We observe that for $s \in [\tilde{\eta}^{n,k}_{2m-1}, \tilde{\eta}^{n,k}_{2m})$, $Q^n_3(s) \geq C^n - 1 > 0$. Thus $Q^n_3(s) = 0$ is possible only for $s \in [\tilde{\eta}^{n,k}_{2m-2}, \tilde{\eta}^{n,k}_{2m-1})$.

Step 2. *As in Step 2 of the proof of Theorem 5.4, we show that with high probability there are at most* $\mathbf{k}^n$ *(defined in (6.8)) intervals* $[\tau^n_{2k-2}, \tau^n_{2k-1}), k = 1, \ldots, \mathbf{k}^n$ *in* $[0, nt]$, *and with high probability there are at most* $\mathbf{k}^n$ *(defined in (6.8)) intervals* $[\tilde{\eta}^{n,k}_{2m-2}, \tilde{\eta}^{n,k}_{2m-1}), m = 1, \ldots, \mathbf{k}^n$ *in* $[\tau^n_{2k-2}, \tau^n_{2k-1} \wedge nt)$. *Thus* $\mathbb{P}(\mathcal{R}(0, nt))$ *can be bounded by the sum of right sides of (6.9) and (6.28). We then write a typical summand in the double summation in (6.28) as the sum of terms in (6.29) and (6.30).*

We next estimate how many such subintervals are within $[\tau^n_{2k-2} \wedge nt, \tau^n_{2k-1} \wedge nt)$. Each $\tilde{\eta}^{n,k}_{2m-1}$ corresponds to at least one additional job completion for Buffer 3. Let $\mathbf{k}^n_1 = \lfloor (\mu^n_2 + 1)nt \rfloor + 1$. From (6.2) in Lemma 6.1, we have when $n \geq n(\{\mu^n_2\}, 1)$ and $nt > 2$,

$$\begin{aligned} \mathbb{P}(\tilde{\eta}^{n,k}_{2\mathbf{k}^n_1 - 1} \leq nt) &\leq \mathbb{P}(S^n_2(nt) \geq \mathbf{k}^n_1) \\ &\leq \exp\left(-(\mu_2 nt - 1)\Theta^{s,2}_1(\mu_2, 1)\right). \end{aligned} \tag{6.27}$$

Then from (6.27), we have when $n \geq \max\{n(\{\mu^n_2\}, 1)\}$, and $nt \geq 2$,

$$\mathbb{P}(\mathcal{R}(0, nt), \tau^n_{2\mathbf{k}^n - 1} > nt) \tag{6.28}$$

$$\leq \sum_{k=1}^{\mathbf{k}^n} \mathbb{P}\left(\int_{[\tau^n_{2k-2},\ \tau^n_{2k-1}\wedge nt)} 1_{\{Q^n_2(s)\geq D^n\}} dI^n_2(s) \neq 0 \right)$$
$$\leq \sum_{k=1}^{\mathbf{k}^n} \sum_{m=1}^{\mathbf{k}^n_1} \mathbb{P}\left(\int_{[\tilde{\eta}^{n,k}_{2m-2},\tilde{\eta}^{n,k}_{2m-1}\wedge nt)} 1_{\{Q^n_2(s)\geq D^n\}} dI^n_2(s) \neq 0 \right)$$
$$+ \mathbf{k}^n \exp\left(-(\mu_2 nt - 1)\Theta^{s,2}_1(\mu_2, 1)\right).$$

By the definition of $\beta^{n,k}_m$, for $s \in [\tilde{\eta}^{n,k}_{2m-2}, \beta^{n,k}_m)$, $Q^n_2(s) < \frac{D^n}{4}$, and so a typical summand in (6.28) is equal to

$$\mathbb{P}\left(\beta^{n,k}_m > \tilde{\eta}^{n,k}_{2m-2},\ \int_{[\beta^{n,k}_m,\ \tilde{\eta}^{n,k}_{2m-1}\wedge nt)} 1_{\{Q^n_2(s)\geq D^n\}} dI^n_2(s) \neq 0 \right) \tag{6.29}$$
$$+\ \mathbb{P}\left(\beta^{n,k}_m = \tilde{\eta}^{n,k}_{2m-2},\ \int_{[\beta^{n,k}_m,\ \tilde{\eta}^{n,k}_{2m-1}\wedge nt)} 1_{\{Q^n_2(s)\geq D^n\}} dI^n_2(s) \neq 0 \right). \tag{6.30}$$

Step 3. *We estimate the probability in* (6.29)*. Denoting by* $\mathcal{B}^{n,k}_m$ *the event in* (6.29)*, we estimate* $\mathbb{P}(\mathcal{D}^{n,k}_m \cap \mathcal{B}^{n,k}_m)$ *and* $\mathbb{P}((\mathcal{D}^{n,k}_m)^c \cap \mathcal{B}^{n,k}_m)$*, separately, with the event* $\mathcal{D}^{n,k}_m$ *defined above* (6.32)*. The estimate on* $\mathbb{P}(\mathcal{D}^{n,k}_m \cap \mathcal{B}^{n,k}_m)$ *is established in* (6.32)*, whereas* $\mathbb{P}((\mathcal{D}^{n,k}_m)^c \cap \mathcal{B}^{n,k}_m)$ *is further decomposed into* (6.36) *and* (6.37)*. The term* (6.36) *is estimated in* (6.38) *and* (6.37) *is bounded by* (6.41)*. Combining these, the estimate on* $\mathbb{P}(\mathcal{B}^{n,k}_m)$ *is given in* (6.43)*.*

In the following, we estimate the probabilities in (6.29) and (6.30) separately. We will use the following constants:

$$\begin{aligned} &\epsilon_1 \in \left(0, \min\left\{1, \frac{\mu_2-\mu_3}{8}, \frac{\mu_2}{2}, \frac{\mu_3}{2}\right\}\right), \\ &K = 32\mu_2 + \frac{4\mu_2(\mu_2-\mu_3)}{\mu_3}, \\ &d = \frac{2cK}{\mu_2-\mu_3}, \\ &\theta = \min\left\{\frac{1}{2}, \frac{\mu_2-\mu_3}{32c\mu_3}\right\}. \end{aligned} \tag{6.31}$$

(Note that from (2.1), $\mu_2 > \mu_3$.) There exists $n_1 \in \mathbb{N}$ such that when $n \geq n_1$, also noting that $\lambda_2 = \mu_3$ from Assumption 2.2, we have

$$\mu^n_2 + \epsilon_1 < 2\mu_2,\ \mu^n_3 + \epsilon_1 < 2\mu_3,\ \lambda^n_2 + \epsilon_1 < 2\lambda_2,\ \mu^n_2 - \mu^n_3 - 2\epsilon_1 > \frac{\mu_2-\mu_3}{2},$$
$$\mu^n_2 - \lambda^n_2 + 2\epsilon_1 < 2(\mu_2 - \lambda_2).$$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

For the rest of the proof we assume $n \geq n_1$. Define for $s \geq 0$,

$$\tilde{A}_2^n(s) = A_2^n(\beta_m^{n,k} + s) - A_2^n(\beta_m^{n,k}),$$
$$\tilde{S}_i^n(s) = S_i^n(T_i^n(\beta_m^{n,k}) + s) - S_i^n(T_i^n(\beta_m^{n,k})),\ i = 2, 3.$$

Now denote by $\mathcal{B}_m^{n,k}$ the event in (6.29), i.e.,

$$\mathcal{B}_m^{n,k} \doteq \left\{ \beta_m^{n,k} > \tilde{\eta}_{2m-2}^{n,k},\ \int_{[\beta_m^{n,k},\ \tilde{\eta}_{2m-1}^{n,k} \wedge nt)} 1_{\{Q_2^n(s) \geq D^n\}} dI_2^n(s) \neq 0 \right\}.$$

We claim that, with $\mathcal{D}_m^{n,k} = \{\sup_{0 \leq s \leq \frac{\theta D^n}{K}} \left| Q_2^n(\beta_m^{n,k} + s) - \frac{D^n}{4} \right| \geq \frac{1}{2}\frac{D^n}{4}\}$,

$$\begin{aligned} \mathbb{P}\left(\mathcal{D}_m^{n,k} \cap\ \mathcal{B}_m^{n,k}\right) &\leq \exp\left\{-(\frac{\mu_2 \theta D^n}{K} - 1)\Theta_1^{s,2}(\mu_2, \epsilon_1)\right\} \\ &\quad + \exp\left\{-(\frac{\lambda_2 \theta D^n}{K} - 1)\Theta_1^{a,2}(\lambda_2, \epsilon_1)\right\}. \end{aligned} \tag{6.32}$$

To see this, note that

$$\begin{aligned} \mathbb{P}\left(\mathcal{D}_m^{n,k} \cap \mathcal{B}_m^{n,k}\right) &\leq \mathbb{P}\left(Q_2^n(\beta_m^{n,k} + s) \leq \frac{1}{2}\frac{D^n}{4} \text{ for some } s \in [0, \tfrac{\theta D^n}{K}],\ \mathcal{B}_m^{n,k}\right) \\ &\quad + \mathbb{P}\left(Q_2^n(\beta_m^{n,k} + s) \geq \frac{3}{2}\frac{D^n}{4} \text{ for some } s \in [0, \tfrac{\theta D^n}{K}],\ \mathcal{B}_m^{n,k}\right). \end{aligned}$$

By the definition of $\beta_m^{n,k}$, we have $Q_2^n(\beta_m^{n,k}) \geq \frac{D^n}{4}$. For $Q_2^n(\beta_m^{n,k} + s) \leq \frac{1}{2}\frac{D^n}{4}$ for some $s \in [0, \frac{\theta D^n}{K}]$, we need at least $\frac{1}{2}\frac{D^n}{4}$ service completions for jobs in Buffer 2 in $\frac{\theta D^n}{K}$ units of time. On the other hand, noting that $\beta_m^{n,k} > \tilde{\eta}_{2m-2}^{n,k}$, we have $Q_2^n(\beta_m^{n,k}-) < \frac{D^n}{4}$ and so $Q_2^n(\beta_m^{n,k}) < \frac{D^n}{4}+1$. In order for $Q_2^n(\beta_m^{n,k}+s)$ to be greater than or equal to $\frac{3}{2}\frac{D^n}{4}$ for some $s \in [0, \frac{\theta D^n}{K}]$, we need at least $\frac{1}{2}\frac{D^n}{4} - 1$ arrivals to Buffer 2 in $\frac{\theta D^n}{K}$ time. Therefore, we have

$$\begin{aligned} \mathbb{P}\left(\mathcal{D}_m^{n,k} \cap \mathcal{B}_m^{n,k}\right) &\leq \mathbb{P}\left(\tilde{S}_2^n(\frac{\theta D^n}{K}) \geq \frac{1}{2}\frac{D^n}{4}, \beta_m^{n,k} \leq nt\right) \\ &\quad + \mathbb{P}\left(\tilde{A}_2^n(\frac{\theta D^n}{K}) \geq \frac{1}{2}\frac{D^n}{4} - 1, \beta_m^{n,k} \leq nt\right). \end{aligned}$$

Next note that

$$\begin{aligned} \frac{1}{2}\frac{D^n}{4} &> \frac{2\mu_2}{K} D^n \geq (\mu_2^n + \epsilon_1)\frac{D^n}{K} \geq (\mu_2^n + \epsilon_1)\frac{\theta D^n}{K}, \\ \frac{1}{2}\frac{D^n}{4} - 1 &> \frac{2\lambda_2}{K} D^n \geq (\lambda_2^n + \epsilon_1)\frac{D^n}{K} \geq (\lambda_2^n + \epsilon_1)\frac{\theta D^n}{K}. \end{aligned} \tag{6.33}$$

From (6.2) in Lemma 6.1, Lemmas 6.3, and 6.2, and a conditioning argu-

ment, when $nt \geq 2/\epsilon_1$ and $n \geq \max\{n(\{\mu_2^n\}, \epsilon_1), n(\{\lambda_2^n\}, \epsilon_1), n_1\}$

(6.34)
$$\mathbb{P}\left(\tilde{S}_2^n(\frac{\theta D^n}{K}) \geq \frac{1}{2}\frac{D^n}{4}, \beta_m^{n,k} \leq nt\right) \leq \mathbb{P}\left(S_2^n(\frac{\theta D^n}{K}) \geq (\mu_2^n + \epsilon_1)\frac{\theta D^n}{K}\right) \leq \exp\left\{-(\frac{\mu_2 \theta D^n}{K} - 1)\Theta_1^{s,2}(\mu_2, \epsilon_1)\right\}.$$

and

(6.35)
$$\mathbb{P}\left(\tilde{A}_2^n(\frac{\theta D^n}{K}) \geq \frac{1}{2}\frac{D^n}{4} - 1, \beta_m^{n,k} \leq nt\right) \leq \mathbb{P}\left(A_2^n(\frac{\theta D^n}{K}) \geq (\lambda_2^n + \epsilon_1)\frac{\theta D^n}{K}\right) \leq \exp\left\{-(\frac{\lambda_2 \theta D^n}{K} - 1)\Theta_1^{a,2}(\lambda_2, \epsilon_1)\right\}.$$

This proves the claim. Next,

$$\begin{aligned}
&\mathbb{P}(\mathcal{B}_m^{n,k} \cap (\mathcal{D}_m^{n,k})^c) \\
&\leq \mathbb{P}\left(\tilde{\eta}_{2m-1}^{n,k} - \beta_m^{n,k} - \frac{\theta D^n}{K} \leq \frac{D^n}{K}, (\mathcal{D}_m^{n,k})^c, \mathcal{B}_m^{n,k}\right) \\
&+ \mathbb{P}\left(\tilde{\eta}_{2m-1}^{n,k} - \beta_m^{n,k} - \frac{\theta D^n}{K} > \frac{D^n}{K}, (\mathcal{D}_m^{n,k})^c,\ \mathcal{B}_m^{n,k}\right) \\
&\leq \mathbb{P}\left(\tilde{\eta}_{2m-1}^{n,k} - \beta_m^{n,k} - \frac{\theta D^n}{K} \leq \frac{D^n}{K}, \right. \qquad (6.36) \\
&\qquad\qquad \left. \sup_{0 \leq s \leq \frac{\theta D^n}{K}} Q_2^n(\beta_m^{n,k} + s) \leq \frac{3}{2}\frac{D^n}{4},\ \mathcal{B}_m^{n,k}\right) \\
&+ \mathbb{P}\left(\tilde{\eta}_{2m-1}^{n,k} - \beta_m^{n,k} - \frac{\theta D^n}{K} > \frac{D^n}{K}, \right. \qquad (6.37) \\
&\qquad\qquad \left. \inf_{0 \leq s \leq \frac{\theta D^n}{K}} Q_2^n(\beta_m^{n,k} + s) \geq \frac{1}{2}\frac{D^n}{4},\ \mathcal{B}_m^{n,k}\right).
\end{aligned}$$

On the event in (6.36), $Q_2^n(\beta_m^{n,k}+s) \leq \frac{3}{2}\frac{D^n}{4}$ for all $s \in [0, \frac{\theta D^n}{K}]$, and $Q_2^n(\beta_m^{n,k}+s) \geq D^n$ for some $s \in [\frac{\theta D^n}{K}, \tilde{\eta}_{2m-1}^{n,k} - \beta_m^{n,k}] \subset [\frac{\theta D^n}{K}, \frac{D^n}{K} + \frac{\theta D^n}{K}]$. Thus there must be at least $\frac{1}{2}\frac{D^n}{4}$ arrivals to Buffer 2 in $\frac{D^n}{K}$ time, and so (6.36) is bounded by

$$\mathbb{P}\left(A_2^n\left(\frac{D^n}{K}\right) \geq \frac{1}{2}\frac{D^n}{4}\right).$$

From (6.33), and using similar argument as in (6.34) and (6.35), we have

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

the following upper bound for the probability in (6.36),

(6.38)
$$\mathbb{P}\left(A_2^n\left(\frac{D^n}{K}\right) \geq (\lambda_2^n+\epsilon_1)\frac{D^n}{K}\right) \leq \exp\left\{-(\frac{\lambda_2 D^n}{K}-1)\Theta_1^{a,2}(\lambda_2,\epsilon_1)\right\}.$$

Next consider the event

$$\mathcal{A}_m^{n,k} \doteq \bigcup_{i=2,3}\left\{\left|\tilde{S}_i^n\left(\frac{(\theta+1)D^n}{K}\right)-\mu_i^n\frac{(\theta+1)D^n}{K}\right| \geq \epsilon_1\frac{(\theta+1)D^n}{K}\right\}.$$

Then

(6.39)
$$\begin{aligned}\mathbb{P}(\mathcal{A}_m^{n,k}) \leq & \sum_{i=2,3}\mathbb{P}\left(\tilde{S}_i^n\left(\frac{(\theta+1)D^n}{K}\right) \geq (\mu_i^n+\epsilon_1)\frac{(\theta+1)D^n}{K}\right)\\ & + \sum_{i=2,3}\mathbb{P}\left(\tilde{S}_i^n\left(\frac{(\theta+1)D^n}{K}\right) \leq (\mu_i^n-\epsilon_1)\frac{(\theta+1)D^n}{K}\right).\end{aligned}$$

From Lemmas 6.2, 6.3 and (6.2) in Lemma 6.1, when $n \geq \max\{n(\{\mu_2^n\},\epsilon_1), n(\{\mu_3^n\},\epsilon_1)\}$ and $nt \geq 2/\epsilon_1$,

$$\begin{aligned}&\sum_{i=2,3}\mathbb{P}\left(\tilde{S}_i^n\left(\frac{(\theta+1)D^n}{K}\right) \geq (\mu_i^n+\epsilon_1)\frac{(\theta+1)D^n}{K}\right)\\ &\quad\leq \sum_{i=2,3}\exp\left\{-(\frac{\mu_i(\theta+1)D^n}{K}-1)\Theta_1^{s,i}(\mu_i,\epsilon_1)\right\}.\end{aligned}$$

Let $\tilde{v}_2^n$ and $\tilde{v}_3^n$ denote the residual service times for the jobs in service at time $\beta_m^{n,k}$ in Buffers 2 and 3, respectively. Then from (6.3) in Lemma 6.1, and once again using Lemmas 6.2, 6.3, we have

$$\begin{aligned}&\sum_{i=2,3}\mathbb{P}\left(S_i^n\left(\frac{(\theta+1)D^n}{K}\right) \leq (\mu_i^n-\epsilon_1)\frac{(\theta+1)D^n}{K}\right)\\ &\quad\leq \sum_{i=2,3}\exp\left(-(\mu_i-2\epsilon_1)\frac{(\theta+1)D^n}{K}\Theta_2^{s,i}(\mu_i,\epsilon_1)\right)\\ &\qquad+\sum_{i=2,3}\mathbb{P}\left(\tilde{v}_i^n > \frac{\epsilon_1(\theta+1)D^n}{2\mu_i^n K}\right).\end{aligned}$$

Now using (6.4), similar to (6.20), we have for $i=2,3$,

(6.40)
$$\begin{aligned}&\mathbb{P}\left(\tilde{v}_i^n > \frac{\epsilon_1(\theta+1)D^n}{2\mu_i^n K}\right)\\ &\quad\leq \exp\left(-(\mu_i nt-1)\Theta_1^{s,i}(\mu_i,\epsilon_1)\right)\\ &\qquad+(\mu_i^n+\epsilon_1)nt\exp\left(-\frac{p_0\epsilon_1(\theta+1)D^n}{2\mu_i K}\right)\exp(\Lambda_{s,i}(p_0)),\end{aligned}$$

where $0 < p_0 \in \mathcal{O}$. Combining the above estimates, we have

$$\begin{aligned}\mathbb{P}(\mathcal{A}_m^{n,k}) \le & \sum_{i=2,3} \exp\left\{-(\frac{\mu_i(\theta+1)D^n}{K} - 1)\Theta_1^{s,i}(\mu_i,\epsilon_1)\right\} \\ & + \sum_{i=2,3} \exp\left(-(\mu_i - 2\epsilon_1)\frac{(\theta+1)D^n}{K}\Theta_2^{s,i}(\mu_i,\epsilon_1)\right) \\ & + \sum_{i=2,3} \exp\left(-(\mu_i nt - 1)\Theta_1^{s,i}(\mu_i,\epsilon_1)\right) \\ & + \sum_{i=2,3} (\mu_i^n + \epsilon_1)nt \exp\left(-\frac{p_0\epsilon_1(\theta+1)D^n}{2\mu_i K}\right)\exp(\Lambda_{s,i}(p_0)).\end{aligned} \tag{6.41}$$

We next observe that on the intersection of $(\mathcal{A}_m^{n,k})^c$ and the event in (6.37), for $s \in [0, \frac{(\theta+1)D^n}{K}]$, $Q_2^n(\beta_m^{n,k}+s)$ is always nonzero, because $Q_2^n(\beta_m^{n,k}) \ge D^n/4$ and

$$\tilde{S}_2^n\left(\frac{(\theta+1)D^n}{K}\right) \le (\mu_2^n + \epsilon_1)\frac{(\theta+1)D^n}{K} \le 2\mu_2\frac{2D^n}{K} < \frac{D^n}{8}.$$

Also on this set, $\tilde{\eta}_{2m-2}^k < \frac{(\theta+1)D^n}{K} + \beta_m^{n,k} < \tilde{\eta}_{2m-1}^k$ and so

$$Q_3^n(\frac{(\theta+1)D^n}{K} + \beta_m^{n,k}) < C^n - 1. \tag{6.42}$$

Consequently, according to our policy, Server 1 works on Buffer 2 continuously over the interval $[\beta_m^{n,k}, \beta_m^{n,k} + \frac{(\theta+1)D^n}{K}]$. Thus, on this set,

$$\begin{aligned}Q_3^n\left(\beta_m^{n,k} + \frac{(\theta+1)D^n}{K}\right) &\ge \tilde{S}_2^n\left(\frac{(\theta+1)D^n}{K}\right) - \tilde{S}_3^n\left(\frac{(\theta+1)D^n}{K}\right) \\ &\ge (\mu_2^n - \mu_3^n - 2\epsilon_1)\frac{(\theta+1)D^n}{K} \\ &\ge \frac{\mu_2 - \mu_3}{2}\frac{D^n}{K} \\ &\ge C^n.\end{aligned}$$

However this contradicts (6.42) and so we must have that the intersection of $(\mathcal{A}_m^{n,k})^c$ and the event in (6.37) is empty. Thus the probability in (6.37) can be bounded by $\mathbb{P}(\mathcal{A}_m^{n,k})$, and combining this observation with (6.41) and the bound on the probability in (6.36) obtained in (6.38),

$$\begin{aligned}&\mathbb{P}(\mathcal{B}_m^{n,k}) \\ &\quad \le \exp\left\{-(\frac{\lambda_2 D^n}{K} - 1)\Theta_1^{a,2}(\lambda_2,\epsilon_1)\right\}\end{aligned} \tag{6.43}$$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

$$+\sum_{i=2,3}\exp\left\{-(\frac{\mu_i(\theta+1)D^n}{K}-1)\Theta_1^{s,i}(\mu_i,\epsilon_1)\right\}$$
$$+\sum_{i=2,3}\exp\left(-(\mu_i-2\epsilon_1)\frac{(\theta+1)D^n}{K}\Theta_2^{s,i}(\mu_i,\epsilon_1)\right)$$
$$+\sum_{i=2,3}\exp\left(-(\mu_i nt-1)\Theta_1^{s,i}(\mu_i,\epsilon_1)\right)$$
$$+\sum_{i=2,3}(\mu_i^n+\epsilon_1)nt\exp\left(-\frac{p_0\epsilon_1(\theta+1)D^n}{2\mu_i K}\right)\exp(\Lambda_{s,i}(p_0)).$$

Step 4. *Denote the event in* (6.30) *by* $\mathcal{C}_m^{n,k}$*, and consider the event* $\mathcal{H}_m^{n,k}$ *in* (6.44)*. Then* $\mathbb{P}(\mathcal{H}_m^{n,k}\cap\mathcal{C}_m^{n,k})$ *is estimated in* (6.45)*, while* $\mathbb{P}((\mathcal{H}_m^{n,k})^c\cap\mathcal{C}_m^{n,k})$ *is further written as a summation of* (6.46) *and* (6.47)*. In* (6.48)*, the term in* (6.46) *is bounded by the right side of* (6.41)*. At last, the probability* (6.47) *is analyzed below* (6.48) *and an upper bound for it is established in* (6.53)*.*

We now consider the event in (6.30), and denote it by $\mathcal{C}_m^{n,k}$, i.e.,

$$\mathcal{C}_m^{n,k}\doteq\left\{\beta_m^{n,k}=\tilde{\eta}_{2m-2}^{n,k},\ \int_{[\beta_m^{n,k},\ \tilde{\eta}_{2m-1}^{n,k}\wedge nt)}1_{\{Q_2^n(s)\geq D^n\}}dI_2^n(s)\neq 0\right\}.$$

When $k=m=1$, on $\mathcal{C}_1^{n,1}$, we have $\beta_1^{n,1}=\tilde{\eta}_0^{n,1}=0$, and so $Q_2^n(\beta_1^{n,1})=0$, which is a contradiction to the definition of $\beta_1^{n,1}$. Thus $\mathbb{P}(\mathcal{C}_1^{n,1})=0$. Consider $(k,m)\neq(1,1)$. Let

$$\mathcal{H}_m^{n,k}=$$
$$\left\{\inf_{0\leq s\leq\frac{\theta D^n}{K}}Q_3^n(\beta_m^{n,k}+s)\leq\frac{L^n}{2}\ \text{or}\ \inf_{0\leq s\leq\frac{\theta D^n}{K}}Q_2^n(\beta_m^{n,k}+s)\leq\frac{1}{2}\frac{D^n}{4}\right\}. \tag{6.44}$$

We now estimate $\mathbb{P}(\mathcal{H}_m^{n,k}\cap\mathcal{C}_m^{n,k})$. Observe that on $\{\tilde{\eta}_{2m-2}^{n,k}<\tau_{2k-1}^n\wedge nt\}$, for $k\geq 2, m=1$,

$$Q_3^n(\tilde{\eta}_{2m-2}^{n,k}-)=Q_3^n(\tau_{2k-2}^n-)\geq\min\{L^n,C^n-1\}=L^n,$$

and for $k\geq 1, m\geq 2$,

$$Q_3^n(\tilde{\eta}_{2m-2}^{n,k}-)\geq C^n-1>L^n.$$

Thus for $(k,m)\neq(1,1)$, on $\mathcal{C}_m^{n,k}$, we have $Q_3^n(\beta_m^{n,k})=Q_3^n(\tilde{\eta}_{2m-2}^{n,k})\geq L^n-1$. By the definition of $\beta_m^{n,k}$, we have $Q_2^n(\beta_m^{n,k})\geq\frac{D^n}{4}$. For $Q_2^n(\beta_m^{n,k}+s)\leq\frac{1}{2}\frac{D^n}{4}$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

for some $s \in [0, \frac{\theta D^n}{K}]$, we need at least $\frac{1}{2}\frac{D^n}{4}$ service completions for Buffer 2 in $\frac{\theta D^n}{K}$ time. Similarly, to make $Q_3^n(\beta_m^{n,k} + s) \le \frac{L^n}{2}$ for some $s \in [0, \frac{\theta D^n}{K}]$, we need at least $\frac{L^n}{4}$ service completions for Buffer 3 in $\frac{\theta D^n}{K}$ time. Thus we have

$$\mathbb{P}(\mathcal{H}_m^{n,k} \cap \mathcal{C}_m^{n,k}) \le \mathbb{P}\left(\tilde{S}_3^n(\frac{\theta D^n}{K}) \ge \frac{L^n}{4}\right) + \mathbb{P}\left(\tilde{S}_2^n(\frac{\theta D^n}{K}) \ge \frac{1}{2}\frac{D^n}{4}\right).$$

Noting that

$$\frac{L^n}{4} \ge 2\mu_3 \frac{2\theta d}{K} L^n \ge (\mu_3^n + \epsilon_1)\frac{2\theta(D^n - d)}{K} \ge (\mu_3^n + \epsilon_1)\frac{\theta D^n}{K},$$
$$\frac{1}{2}\frac{D^n}{4} \ge \frac{2\mu_2 \theta D^n}{K} \ge (\mu_2^n + \epsilon_1)\frac{\theta D^n}{K},$$

we have when $n \ge \max\{n(\{\mu_3^n\}, \epsilon_1), n(\{\mu_2^n\}, \epsilon_1), n_1\}$ and $nt \ge 2/\epsilon_1$,

$$\begin{aligned}\mathbb{P}(\mathcal{H}_m^{n,k} \cap \mathcal{C}_m^{n,k}) &\le \sum_{i=2,3} \mathbb{P}\left(\tilde{S}_i^n(\frac{\theta D^n}{K}) \ge (\mu_i^n + \epsilon_1)\frac{\theta D^n}{K}\right) \\ &\le \sum_{i=2,3} \exp\left\{-(\frac{\mu_i \theta D^n}{K} - 1)\Theta_1^{s,i}(\mu_i, \epsilon_1)\right\}.\end{aligned} \tag{6.45}$$

Next,

$$\begin{aligned}&\mathbb{P}((\mathcal{H}_m^{n,k})^c \cap \mathcal{C}_m^{n,k}) \\ &\le \mathbb{P}\left(\tilde{\eta}_{2m-1}^{n,k} - \beta_m^{n,k} - \frac{\theta D^n}{K} > \frac{D^n}{K}, (\mathcal{H}_m^{n,k})^c \cap \mathcal{C}_m^{n,k}\right) \qquad (6.46)\\ &+ \mathbb{P}\left(\tilde{\eta}_{2m-1}^{n,k} - \beta_m^{n,k} - \frac{\theta D^n}{K} \le \frac{D^n}{K}, (\mathcal{H}_m^{n,k})^c \cap \mathcal{C}_m^{n,k}\right). \qquad (6.47)\end{aligned}$$

Using a similar argument as for (6.37), we see that

$$\mathbb{P}\left(\tilde{\eta}_{2m-1}^{n,k} - \beta_m^{n,k} - \frac{\theta D^n}{K} > \frac{D^n}{K}, (\mathcal{H}_m^{n,k})^c \cap \mathcal{C}_m^{n,k}\right) \le \mathbb{P}(\mathcal{A}_m^{n,k}). \tag{6.48}$$

We now consider the probability in (6.47). Define $t_0 \doteq \frac{L^n}{4\mu_3}$, and

$$N \doteq \left\lfloor \frac{\frac{D^n}{K}}{t_0} \right\rfloor + 1 \in \left[\left\lfloor \frac{8c\mu_3}{\mu_2 - \mu_3} \right\rfloor + 1, \left\lfloor \frac{16c\mu_3}{\mu_2 - \mu_3} \right\rfloor + 1\right].$$

For $s \in [0, t_0]$, $j = 2, 3$ and $l = 0, \ldots, N-1$, define

$$S_{j,l+1}^n(s) = S_j^n\left(T_j^n\left(\beta_m^{n,k} + \frac{\theta D^n}{K} + lt_0\right) + s\right)$$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

$$- S_j^n\left(T_j^n\left(\beta_m^{n,k} + \frac{\theta D^n}{K} + lt_0\right)\right)$$

$$A_{2,l+1}^n(s) = A_2^n\left(\beta_m^{n,k} + \frac{\theta D^n}{K} + lt_0 + s\right) - A_2^n\left(\beta_m^{n,k} + \frac{\theta D^n}{K} + lt_0\right).$$

Consider the event

$$\begin{aligned}\mathcal{E}_{m,1}^{n,k} &= \left\{\left|\frac{S_{3,1}^n(t_0) - \mu_3^n t_0}{t_0}\right| > \epsilon_1 \text{ or } \left|\frac{S_{2,1}^n(t_0) - \mu_2^n t_0}{t_0}\right| > \epsilon_1\right. \\ &\quad \left. \text{or } \left|\frac{A_{2,1}^n(t_0) - \lambda_2^n t_0}{t_0}\right| > \epsilon_1\right\}.\end{aligned}$$

We first estimate $\mathbb{P}(\mathcal{E}_{m,1}^{n,k})$. Clearly,

$$\begin{aligned}\mathbb{P}(\mathcal{E}_{m,1}^{n,k}) &\le \sum_{j=2,3}\mathbb{P}\left(\left|\frac{S_{j,1}^n(t_0) - \mu_j^n t_0}{t_0}\right| > \epsilon_1\right) + \mathbb{P}\left(\left|\frac{A_{2,1}^n(t_0) - \lambda_2^n t_0}{t_0}\right| > \epsilon_1\right) \\ &= \sum_{j=2,3}\mathbb{P}\left(S_{j,1}^n(t_0) > (\mu_j^n + \epsilon_1)t_0\right) + \mathbb{P}\left(A_{2,1}^n(t_0) > (\lambda_2^n + \epsilon_1)t_0\right) \\ &\quad + \sum_{j=2,3}\mathbb{P}\left(S_j^n(t_0) < (\mu_j^n - \epsilon_1)t_0\right) + \mathbb{P}\left(A_2^n(t_0) < (\lambda_2^n - \epsilon_1)t_0\right).\end{aligned}$$

Now from Lemmas 6.3, 6.2 and (6.2) in Lemma 6.1, we have for

$$n \ge \max\{n(\{\mu_3^n\}, \epsilon_1), n(\{\mu_2^n\}, \epsilon_1), n(\{\lambda_2^n\}, \epsilon_1), n_1\} \text{ and } nt > 2/\epsilon_1,$$

$$\begin{aligned}&\sum_{j=2,3}\mathbb{P}\left(S_{j,1}^n(t_0) > (\mu_j^n + \epsilon_1)t_0\right) + \mathbb{P}\left(A_{2,1}^n(t_0) > (\lambda_2^n + \epsilon_1)t_0\right) \\ &\le \sum_{j=2,3}\exp\left\{-(\mu_j t_0 - 1)\Theta_1^{s,j}(\mu_j, \epsilon_1)\right\} + \exp\left\{-(\lambda_2 t_0 - 1)\Theta_1^{a,2}(\lambda_2, \epsilon_1)\right\}.\end{aligned}$$

Let $\check{v}_2^n, \check{v}_3^n$ denote the residual service times for jobs in service at time $\beta_m^{n,k} + \frac{\theta D^n}{K}$ in Buffers 2 and 3, respectively, and $\check{u}_2^n$ the residual arrival time at the same instant for jobs to Buffer 2. From (6.3) in Lemma 6.1, and Lemmas 6.3, 6.2 again, we have when $n \ge \max\{n(\{\mu_3^n\}, \epsilon_1), n(\{\mu_2^n\}, \epsilon_1), n(\{\lambda_2^n\}, \epsilon_1), n_1\}$,

$$\begin{aligned}&\sum_{j=2,3}\mathbb{P}\left(S_{j,1}^n(t_0) < (\mu_j^n - \epsilon_1)t_0\right) + \mathbb{P}\left(A_{2,1}^n(t_0) < (\lambda_2^n - \epsilon_1)t_0\right) \\ &\le \sum_{j=2,3}\exp\left(-(\mu_j - 2\epsilon_1)t_0\Theta_2^{s,j}(\mu_j, \epsilon_1)\right) + \sum_{j=2,3}\mathbb{P}\left(\check{v}_j^n > \frac{\epsilon t_0}{2\mu_j^n}\right) \\ &\quad + \exp\left(-(\lambda_2 - 2\epsilon_1)t_0\Theta_2^{a,2}(\lambda_2, \epsilon_1)\right) + \mathbb{P}\left(\check{u}_2^n > \frac{\epsilon t_0}{2\lambda_2^n}\right).\end{aligned}$$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Using similar arguments as in (6.20) and (6.40), we have for $j = 2, 3$,

$$\mathbb{P}\left(\check{v}_j^n > \frac{\epsilon t_0}{2\mu_j^n}\right) \le \exp\left(-(\mu_j nt - 1)\Theta_1^{s,j}(\mu_j, \epsilon_1)\right) + (\mu_j^n + \epsilon_1)nt \exp\left(-\frac{p_0\epsilon_1 t_0}{2\mu_j}\right)\exp(\Lambda_{s,j}(p_0)),$$

and

$$\mathbb{P}\left(\check{u}_2^n > \frac{\epsilon t_0}{2\lambda_2^n}\right) \le \exp\left(-(\lambda_2 nt - 1)\Theta_1^{a,2}(\lambda_2, \epsilon_1)\right) + (\lambda_2^n + \epsilon_1)nt \exp\left(-\frac{p_0\epsilon_1 t_0}{2\lambda_2}\right)\exp(\Lambda_{a,2}(p_0)),$$

where $0 < p_0 \in \mathcal{O}$. Thus

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_{m,1}^{n,k}) \le & \sum_{j=2,3} \exp\left\{-(\mu_j t_0 - 1)\Theta_1^{s,j}(\mu_j, \epsilon_1)\right\} \\
& + \exp\left\{-(\lambda_2 t_0 - 1)\Theta_1^{a,2}(\lambda_2, \epsilon_1)\right\} \\
& + \sum_{j=2,3} \exp\left(-(\mu_j - 2\epsilon_1)t_0\Theta_2^{s,j}(\mu_j, \epsilon_1)\right) \\
& + \sum_{j=2,3} \exp\left(-(\mu_j nt - 1)\Theta_1^{s,j}(\mu_j, \epsilon_1)\right) \\
& + \sum_{j=2,3} (\mu_j^n + \epsilon_1)nt \exp\left(-\frac{p_0\epsilon_1 t_0}{2\mu_j}\right)\exp(\Lambda_{s,j}(p_0)) \\
& + \exp\left(-(\lambda_2 - 2\epsilon_1)t_0\Theta_2^{a,2}(\lambda_2, \epsilon_1)\right) \\
& + \exp\left(-(\lambda_2 nt - 1)\Theta_1^{a,2}(\lambda_2, \epsilon_1)\right) \\
& + (\lambda_2^n + \epsilon_1)nt \exp\left(-\frac{p_0\epsilon_1 t_0}{2\lambda_2}\right)\exp(\Lambda_{a,2}(p_0)) \\
\doteq & \, \Gamma(t_0, nt).
\end{aligned}$$

On the intersection $(\mathcal{E}_{m,1}^{n,k})^c$ and the event in (6.47), we have $Q_2^n(\beta_m^{n,k} + \frac{\theta D^n}{K}) > \frac{1}{2}\frac{D^n}{4}$, $Q_3^n(\beta_m^{n,k} + \frac{\theta D^n}{K}) > \frac{L^n}{2}$, and for $s \in [0, t_0]$,

$$\begin{aligned}
Q_2^n\left(\beta_m^{n,k} + \frac{\theta D^n}{K} + s\right) &\ge Q_2^n\left(\beta_m^{n,k} + \frac{\theta D^n}{K}\right) + A_{2,1}^n(s) - S_{2,1}^n(s) \\
&> \frac{1}{2}\frac{D^n}{4} - (\mu_2^n + \epsilon_1)t_0
\end{aligned} \tag{6.49}$$

$$= \frac{1}{2}\frac{D^n}{4} - \frac{\mu_2^n + \epsilon_1}{2\mu_3}\frac{L^n}{2}$$
$$> \left(\frac{1}{8} - \frac{\mu_2}{2\mu_3 d}\right) D^n > 0,$$

where the last inequality follows from noting that $K \geq \frac{4\mu_2(\mu_2-\mu_3)}{\mu_3}$. We note that $[\beta_m^{n,k}+\frac{\theta D^n}{K}, \beta_m^{n,k}+\frac{\theta D^n}{K}+t_0] \subset [\tilde{\eta}_{2m-2}^n, \tilde{\eta}_{2m-1}^n)$, and for $t \in [\tilde{\eta}_{2m-2}^n, \tilde{\eta}_{2m-1}^n)$, $Q_3^n(t) < C^n - 1$ and $Q_3^n(t) - \frac{\mu_2^n}{\mu_1^n} Q_1^n(t) < L^n$. Further noting that Q_2^n is nonempty on the time interval $[\beta_m^{n,k} + \frac{\theta D^n}{K}, \beta_m^{n,k} + \frac{\theta D^n}{K} + t_0]$, we see that Server 1 will work on Buffer 2 continuously during this time interval. Consequently, we have

$$\begin{aligned} Q_3^n\left(\beta_m^{n,k} + \frac{\theta D^n}{K} + t_0\right) &\geq Q_3^n\left(\beta_m^{n,k} + \frac{\theta D^n}{K}\right) + S_{2,1}^n(t_0) - S_{3,1}^n(t_0) \\ &> \frac{L^n}{2} + (\mu_2^n - \mu_3^n - 2\epsilon_1)t_0 \\ &\geq \frac{L^n}{2}. \end{aligned} \tag{6.50}$$

Furthermore, for all $s \in [0, t_0]$,

$$\begin{aligned} Q_3^n\left(\beta_m^{n,k} + \frac{\theta D^n}{K} + s\right) &= Q_3^n\left(\beta_m^{n,k} + \frac{\theta D^n}{K}\right) + S_{2,1}^n(s) - S_{3,1}^n(s) \\ &> \frac{L^n}{2} - (\mu_3^n + \epsilon_1)t_0 \\ &= \frac{L^n}{2} - \frac{\mu_3^n + \epsilon_1}{2\mu_3}\frac{L^n}{2} \\ &> 0. \end{aligned} \tag{6.51}$$

We repeat the above analysis for the time interval $[t_0, 2t_0]$. First define the event $\mathcal{E}_{m,2}^{n,k}$ as $\mathcal{E}_{m,1}^{n,k}$ by replacing $S_{j,1}^n(s)$ and $A_{2,1}^n(s)$ with $S_{j,2}^n(s)$ and $A_{2,2}^n(s)$. We then obtain that for

$$n \geq \max\{n(\{\mu_3^n\}, \epsilon_1), n(\{\mu_2^n\}, \epsilon_1), n(\{\lambda_2^n\}, \epsilon_1), n_1\} \text{ and } nt \geq 2/\epsilon_1,$$

$\mathbb{P}(\mathcal{E}_{m,2}^{n,k})$ has the same upper bound $\Gamma(t_0, nt)$ as $\mathbb{P}(\mathcal{E}_{m,1}^{n,k})$. On the intersection $(\mathcal{E}_{m,1}^{n,k})^c \cap (\mathcal{E}_{m,2}^{n,k})^c$ and the event in (6.47), following the similar arguments to those in (6.50) and (6.51), it can be shown that

$$Q_3^n\left(\beta_m^{n,k} + \frac{\theta D^n}{K} + 2t_0\right) > \frac{L^n}{2}, Q_3^n\left(\beta_m^{n,k} + \frac{\theta D^n}{K} + t_0 + s\right) > 0, \text{ for } s \in [0, t_0],$$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

and similar to (6.49), we have for $s \in [0, t_0]$,

$$Q_2^n\left(\beta_m^{n,k} + \frac{\theta D^n}{K} + t_0 + s\right) \geq Q_2^n\left(\beta_m^{n,k} + \frac{\theta D^n}{K} + t_0\right) + A_{2,2}^n(s) - S_{2,2}^n(s)$$
$$> \left(\frac{1}{8} - \frac{\mu_2}{\mu_3 d}\right) D^n > 0.$$

Repeating this argument N times we see that on the intersection $\cap_{j=1}^N (\mathcal{E}_{m,j}^{n,k})^c$ and the event in (6.47), for $s \in [0, t_0]$, and $l = 0, \ldots, N-1$,

$$Q_3^n\left(\beta_m^{n,k} + \frac{\theta D^n}{K} + l t_0 + s\right) > 0 \tag{6.52}$$

and

$$Q_2^n\left(\beta_m^{n,k} + \frac{\theta D^n}{K} + l t_0 + s\right) \geq Q_2^n\left(\beta_m^{n,k} + \frac{\theta D^n}{K} + l t_0\right) + A_{2,l+1}^n(s) - S_{2,l+1}^n(s)$$
$$> \left(\frac{1}{8} - \frac{(l+1)\mu_2}{2\mu_3 d}\right) D^n > 0,$$

where the last inequality follows from noting that $K \geq 32\mu_2 + \frac{4\mu_2(\mu_2-\mu_3)}{\mu_3}$.

We note that, since Server 2 does not idle if there are jobs in Buffer 3, (6.52) implies that the probability of the intersection $\cap_{j=1}^N (\mathcal{E}_{m,j}^{n,k})^c$ and the event in (6.47) is 0. Thus

$$\begin{aligned} &\mathbb{P}\left(\tilde{\eta}_{2m-1}^{n,k} - \beta_m^{n,k} - \frac{\theta D^n}{K} \leq \frac{D^n}{K}, (\mathcal{H}_m^{n,k})^c \cap \mathcal{C}_m^{n,k}\right) \\ &\leq \sum_{l=1}^N \mathbb{P}(\mathcal{E}_{m,l}^{n,k}) \leq N\Gamma(t_0, nt). \end{aligned} \tag{6.53}$$

Combining the above estimate in (6.53) with (6.45), (6.48) and (6.41),

$$\begin{aligned} \mathbb{P}(\mathcal{C}_m^{n,k}) \leq &\sum_{i=2,3} \exp\left\{-(\frac{\mu_i \theta D^n}{K} - 1)\Theta_1^{s,i}(\mu_i, \epsilon_1)\right\} \\ &+ \sum_{i=2,3} \exp\left\{-(\frac{\mu_i (\theta+1) D^n}{K} - 1)\Theta_1^{s,i}(\mu_i, \epsilon_1)\right\} \\ &+ \sum_{i=2,3} \exp\left(-(\mu_i - 2\epsilon_1)\frac{(\theta+1)D^n}{K}\Theta_2^{s,i}(\mu_i, \epsilon_1)\right) \\ &+ \sum_{i=2,3} \exp\left(-(\mu_i nt - 1)\Theta_1^{s,i}(\mu_2, \epsilon_1)\right) \end{aligned}$$

$$+\sum_{i=2,3}(\mu_i^n+\epsilon_1)nt\exp\left(-\frac{p_0\epsilon_1(\theta+1)D^n}{2\mu_iK}\right)\exp(\Lambda_{s,i}(p_0))+N\Gamma(t_0,nt).$$

Step 5. *We combine the key estimates and finish the proof.*

Combining the above estimate for $\mathbb{P}(\mathcal{C}_m^{n,k})$ with (6.9), (6.28) and (6.43), we have for large enough n and $nt\geq\epsilon_1$,

$$\begin{aligned}\mathbb{P}(\mathcal{R}(n,t))\leq{}&\mathbf{k}^n\mathbf{k}_1^n\left[\exp\left\{-(\frac{\lambda_2D^n}{K}-1)\Theta_1^{a,2}(\lambda_2,\epsilon_1)\right\}\right.\\&+2\sum_{i=2,3}\exp\left\{-(\frac{\mu_i(\theta+1)D^n}{K}-1)\Theta_1^{s,i}(\mu_i,\epsilon_1)\right\}\\&+2\sum_{i=2,3}\exp\left\{-(\mu_i-2\epsilon_1)\frac{(\theta+1)D^n}{K}\Theta_2^{s,i}(\mu_i,\epsilon_1)\right\}\\&+2\sum_{i=2,3}\exp\left\{-(\mu_int-1)\Theta_1^{s,i}(\mu_2,\epsilon_1)\right\}\\&+2\sum_{i=2,3}(\mu_i^n+\epsilon_1)nt\exp\left\{-\frac{p_0\epsilon_1(\theta+1)D^n}{2\mu_iK}\right\}\exp(\Lambda_{s,i}(p_0))\\&\left.+\sum_{i=2,3}\exp\left\{-(\frac{\mu_i\theta D^n}{K}-1)\Theta_1^{s,i}(\mu_i,\epsilon_1)\right\}+N\Gamma(t_0,nt)\right]\\&+\mathbf{k}^n\exp\left\{-(\mu_2nt-1)\Theta_1^{s,2}(\mu_2,1)\right\}\\&+\exp\left\{-(\lambda_1nt-1)\Theta_1^{a,1}(\lambda_1,1)\right\}\\&+\exp\left\{-(\lambda_2nt-1)\Theta_1^{a,2}(\lambda_2,1)\right\}.\end{aligned}\tag{6.54}$$

Thus, we can find positive constants $\gamma_i, i=1,2,3,4$, such that for large enough n and $nt\geq\epsilon_1$,

$$\mathbb{P}(\mathcal{R}(0,nt))\leq\gamma_1(nt+1)^2e^{-\gamma_2nt}+\gamma_3(nt+1)^3n^{-\gamma_4l_0}.\qquad\square$$

6.3. *Proof of Lemma 5.7.* Recall the stopping times $\{\tau_k^n\}_{k\in\mathbb{N}_0}$ defined in (6.5) and note from Lemma 4.4 that when $t\in[\tau_{2k-2}^n,\tau_{2k-1}^n)$,

$$\begin{aligned}\Delta^n(t)&\doteq\frac{Q_1^n(t)}{\mu_1^n}+\frac{Q_2^n(t)}{\mu_2^n}-\sqrt{n}\Psi\left(\frac{Q_2^n(t)+Q_3^n(t)}{\sqrt{n}\mu_3^n}\right)\\&\geq\frac{Q_1^n(t)}{\mu_1^n}+\frac{Q_2^n(t)}{\mu_2^n}-\frac{\mu_3}{\mu_2}\left(\frac{Q_2^n(t)}{\mu_3^n}+\frac{Q_3^n(t)}{\mu_3^n}\right)\\&\geq\frac{Q_1^n(t)}{\mu_1^n}-\frac{\mu_3}{\mu_2\mu_3^n}\left(\frac{\mu_2^n}{\mu_1^n}Q_1^n(t)+L^n\right)+\left(\frac{1}{\mu_2^n}-\frac{\mu_3}{\mu_2\mu_3^n}\right)Q_2^n(t)\end{aligned}\tag{6.55}$$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

$$= \frac{1}{\mu_1^n}\left(1 - \frac{\mu_3\mu_2^n}{\mu_3^n\mu_2}\right) Q_1^n(t) - \frac{\mu_3}{\mu_2\mu_3^n} L^n + \left(\frac{1}{\mu_2^n} - \frac{\mu_3}{\mu_2\mu_3^n}\right) Q_2^n(t).$$

Thus for such t the inequality in (5.5) clearly holds (with a suitable choice of C_1).

Next consider intervals of the form $[\tau_{2k-1}^n, \tau_{2k}^n)$. For $k, l \in \mathbb{N}$, we define a sequence of $\{\mathcal{G}_1^n(t)\}_{t\geq 0}$ stopping times within $[\eta_{2l-2}^{n,k}, \eta_{2l-1}^{n,k}) \subset [\tau_{2k-1}^n, \tau_{2k}^n)$ as follows: For $m \in \mathbb{N}$,

$$\begin{aligned}
\zeta_0^{n,k,l} &= \eta_{2l-2}^{n,k},\\
\zeta_{2m-1}^{n,k,l} &= \tau_{2k}^n \wedge \eta_{2l-1}^{n,k} \wedge \inf\left\{t \geq \zeta_{2m-2}^{n,k,l} : W^n(t) - \sqrt{n}\Psi(W_2^n(t)/\sqrt{n}) < g_0\right\},\\
\zeta_{2m}^{n,k,l} &= \tau_{2k}^n \wedge \eta_{2l-1}^{n,k} \wedge \inf\left\{t \geq \zeta_{2m-1}^{n,k,l} : W^n(t) - \sqrt{n}\Psi(W_2^n(t)/\sqrt{n}) \geq g_0\right\}.
\end{aligned}$$

Then

$$\text{for } t \in [\zeta_{2m-2}^{n,k,l}, \zeta_{2m-1}^{n,k,l}),\ W^n(t) - \sqrt{n}\Psi(W_2^n(t)/\sqrt{n}) \geq g_0 \tag{6.56}$$

and so for such t the inequality in (5.5) holds trivially.

Also, from the definition of $\{\eta_l^{n,k}\}_{l\in\mathbb{N}_0}$ and (6.12), we see that for $t \in [\eta_{2l-2}^{n,k}, \eta_{2l-1}^{n,k})$,

$$Q_3^n(t) - \frac{\mu_2^n}{\mu_1^n} Q_1^n(t) \geq L^n,\ Q_1^n(t) < \frac{\mu_1^n}{\mu_2^n}(C^n - L^n + 2). \tag{6.57}$$

Thus from Definition 3.2 we see that over the interval $[\zeta_{2m-1}^{n,k,l}, \zeta_{2m}^{n,k,l})$ Server 1 does not work on Buffer 2. Thus letting

$$\tilde{A}_2^n(t) \doteq A_2^n(t) - A_2^n(\zeta_{2m-1}^{n,k,l}),\quad \tilde{S}_3^n(t) \doteq S_3^n(T_3^n(t)) - S_3^n(T_3^n(\zeta_{2m-1}^{n,k,l})),$$

for $t \in [\zeta_{2m-1}^{n,k,l}, \zeta_{2m}^{n,k,l})$, $Q_2^n(t) = Q_2^n(\zeta_{2m-1}^{n,k,l}) + \tilde{A}_2^n(t)$ and $Q_3^n(t) = Q_3^n(\zeta_{2m-1}^{n,k,l}) - \tilde{S}_3^n(t)$. Using the nondecreasing and Lipschitz continuity property of Ψ, we now have for $t \in [\zeta_{2m-1}^{n,k,l}, \zeta_{2m}^{n,k,l})$,

$$\begin{aligned}
\Delta^n(t) &= \frac{Q_1^n(t)}{\mu_1^n} + \frac{Q_2^n(t)}{\mu_2^n} - \sqrt{n}\Psi\left(\frac{Q_2^n(t) + Q_3^n(t)}{\sqrt{n}\mu_3^n}\right)\\
&\geq \frac{Q_2^n(\zeta_{2m-1}^{n,k,l}) + \tilde{A}_2^n(t)}{\mu_2^n}\\
&\quad - \sqrt{n}\Psi\left(\frac{Q_2^n(\zeta_{2m-1}^{n,k,l}) + \tilde{A}_2^n(t) + Q_3^n(\zeta_{2m-1}^{n,k,l}) - \tilde{S}_3^n(t)}{\sqrt{n}\mu_3^n}\right)
\end{aligned}$$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

$$
\begin{aligned}
&= \frac{\tilde{A}_2^n(t)}{\mu_2^n} - \frac{Q_1^n(\zeta_{2m-1}^{n,k,l})}{\mu_1^n} + \left[\frac{Q_1^n(\zeta_{2m-1}^{n,k,l})}{\mu_1^n} + \frac{Q_2^n(\zeta_{2m-1}^{n,k,l})}{\mu_2^n}\right.\\
&\quad \left. -\sqrt{n}\Psi\left(\frac{Q_2^n(\zeta_{2m-1}^{n,k,l}) + Q_3^n(\zeta_{2m-1}^{n,k,l})}{\sqrt{n}\mu_3^n}\right)\right]\\
&\quad - \left[\sqrt{n}\Psi\left(\frac{Q_2^n(\zeta_{2m-1}^{n,k,l}) + \tilde{A}_2^n(t) + Q_3^n(\zeta_{2m-1}^{n,k,l}) - \tilde{S}_3^n(t)}{\sqrt{n}\mu_3^n}\right)\right.\\
&\quad \left. -\sqrt{n}\Psi\left(\frac{Q_2^n(\zeta_{2m-1}^{n,k,l}) + Q_3^n(\zeta_{2m-1}^{n,k,l})}{\sqrt{n}\mu_3^n}\right)\right]\\
&\geq \frac{\tilde{A}_2^n(t)}{\mu_2^n} - \frac{Q_1^n(\zeta_{2m-1}^{n,k,l})}{\mu_1^n} + \Delta^n(\zeta_{2m-1}^{n,k,l}) - \frac{\mu_3}{\mu_3^n\mu_2}(\tilde{A}_2^n(t) - \tilde{S}_3^n(t))^+.\\
&\geq \begin{cases} -\left|\frac{1}{\mu_2^n} - \frac{\mu_3}{\mu_3^n\mu_2}\right| \tilde{A}_2^n(t) - \frac{Q_1^n(\zeta_{2m-1}^{n,k,l})}{\mu_1^n} + \Delta^n(\zeta_{2m-1}^{n,k,l}), \\ \qquad \text{when } \tilde{A}_2^n(t) \geq \tilde{S}_3^n(t), \\ \frac{\tilde{A}_2^n(t)}{\mu_2^n} - \frac{Q_1^n(\zeta_{2m-1}^{n,k,l})}{\mu_1^n} + \Delta^n(\zeta_{2m-1}^{n,k,l}), \\ \qquad \text{when } \tilde{A}_2^n(t) < \tilde{S}_3^n(t). \end{cases}
\end{aligned}
$$

Using the Lipschitz property of Ψ again we see that there exists $C_0 > 0$ (depending only on Ψ and the rate parameters) such that

$$\Delta^n(\zeta_{2m-1}^{n,k,l}) \geq \Delta^n(\zeta_{2m-1}^{n,k,l}-) - C_0 \geq g_0 - C_0,$$

where the last inequality is from (6.56). Furthermore, for all $t \in [\zeta_{2m-1}^{n,k,l}, \zeta_{2m}^{n,k,l})$, $\tilde{A}_2^n(t) \leq Q_2^n(t)$ and (from (6.57)) $Q_1^n(t) \leq \frac{\mu_1^n}{\mu_2^n}(C^n - L^n + 2)$. Thus we have for such t (in fact for all $t \in [\eta_{2l-2}^{n,k}, \eta_{2l-1}^{n,k})$),

$$\Delta^n(t) \geq -\left|\frac{1}{\mu_2^n} - \frac{\mu_3}{\mu_3^n\mu_2}\right| Q_2^n(t) - \frac{1}{\mu_2^n}(C^n - L^n + 2) + g_0 - C_0. \tag{6.58}$$

We have therefore shown that (5.5) holds for all $t \in [\eta_{2l-2}^{n,k}, \eta_{2l-1}^{n,k})$.

We finally consider the case when $t \in [\eta_{2l-1}^{n,k}, \eta_{2l}^{n,k})$(with a suitable choice of C_1, C_2). As noted below (6.18), over this interval Server 1 works on Buffer 1 continuously, and since, from the definition of $\tau_{2k-1}^n, \tau_{2k}^n$, over this interval $Q_3^n(t) \geq L^n - 1$, Server 2 works on Buffer 3 continuously. Let

$$
\begin{aligned}
\check{A}_2^n(t) &\doteq A_2^n(t) - A_2^n(\eta_{2l-1}^{n,k}),\\
\check{S}_3^n(t) &\doteq S_3^n(t - \eta_{2l-1}^{n,k} + T_3^n(\eta_{2l-1}^{n,k})) - S_3^n(T_3^n(\eta_{2l-1}^{n,k})).
\end{aligned}
$$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Thus again from the monotonicity and Lipschitz continuity property of Ψ, we have for $t \in [\eta^{n,k}_{2l-1}, \eta^{n,k}_{2l})$,

$$\begin{aligned}
\Delta^n(t) &= \frac{Q^n_1(t)}{\mu^n_1} + \frac{Q^n_2(t)}{\mu^n_2} - \sqrt{n}\Psi\left(\frac{Q^n_2(t)+Q^n_3(t)}{\sqrt{n}\mu^n_3}\right)\\
&= \frac{Q^n_1(\eta^{n,k}_{2l-1}) + [Q^n_1(t) - Q^n_1(\eta^{n,k}_{2l-1})]}{\mu^n_1} + \frac{Q^n_2(\eta^{n,k}_{2l-1}) + \check{A}^n_2(t)}{\mu^n_2}\\
&\quad - \sqrt{n}\Psi\left(\frac{Q^n_2(\eta^{n,k}_{2l-1}) + \check{A}^n_2(t) + Q^n_3(\eta^{n,k}_{2l-1}) - \check{S}^n_3(t)}{\sqrt{n}\mu^n_3}\right)\\
&= \frac{Q^n_1(t) - Q^n_1(\eta^{n,k}_{2l-1})}{\mu^n_1} + \frac{\check{A}^n_2(t)}{\mu^n_2} + \left[\frac{Q^n_1(\eta^{n,k}_{2l-1})}{\mu^n_1} + \frac{Q^n_2(\eta^{n,k}_{2l-1})}{\mu^n_2}\right.\\
&\quad \left. -\sqrt{n}\Psi\left(\frac{Q^n_2(\eta^{n,k}_{2l-1}) + Q^n_3(\eta^{n,k}_{2l-1})}{\sqrt{n}\mu^n_3}\right)\right]\\
&\quad - \left[\sqrt{n}\Psi\left(\frac{Q^n_2(\eta^{n,k}_{2l-1}) + \check{A}^n_2(t) + Q^n_3(\eta^{n,k}_{2l-1}) - \check{S}^n_3(t)}{\sqrt{n}\mu^n_3}\right)\right.\\
&\quad \left. -\sqrt{n}\Psi\left(\frac{Q^n_2(\eta^{n,k}_{2l-1}) + Q^n_3(\eta^{n,k}_{2l-1})}{\sqrt{n}\mu^n_3}\right)\right]\\
&\geq \frac{Q^n_1(t) - Q^n_1(\eta^{n,k}_{2l-1})}{\mu^n_1} + \frac{\check{A}^n_2(t)}{\mu^n_2} + \Delta^n(\eta^{n,k}_{2l-1}) - \frac{\mu_3}{\mu^n_3\mu_2}(\check{A}^n_2(t) - \check{S}^n_3(t))^+\\
&\geq \begin{cases} \frac{-Q^n_1(\eta^{n,k}_{2l-1})}{\mu^n_1} - \left|\frac{1}{\mu^n_2} - \frac{\mu_3}{\mu^n_3\mu_2}\right| \check{A}^n_2(t) + \Delta^n(\eta^{n,k}_{2l-1}), \\ \qquad \text{when } \check{A}^n_2(t) \geq \check{S}^n_3(t), \\ \frac{-Q^n_1(\eta^{n,k}_{2l-1})}{\mu^n_1} + \frac{\check{A}^n_2(t)}{\mu^n_2} + \Delta^n(\eta^{n,k}_{2l-1}), \\ \qquad \text{when } \check{A}^n_2(t) < \check{S}^n_3(t). \end{cases}
\end{aligned}$$

Also, from (6.58),

$$\begin{aligned}
\Delta^n(\eta^{n,k}_{2l-1}) &\geq \Delta^n(\eta^{n,k}_{2l-1}-) - C_0\\
&\geq -\left|\frac{1}{\mu^n_2} - \frac{\mu_3}{\mu^n_3\mu_2}\right| Q^n_2(\eta^{n,k}_{2l-1}-) - \frac{1}{\mu^n_2}(C^n - L^n + 2) + g_0 - 2C_0,
\end{aligned}$$

and using (6.12)

$$Q^n_1(\eta^{n,k}_{2l-1}) \leq Q^n_1(\eta^{n,k}_{2l-1}-) + 1 < \frac{\mu^n_1}{\mu^n_2}(C^n - L^n + 2) + 1.$$

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Furthermore, for all $t \in [\eta^{n,k}_{2l-1}, \eta^{n,k}_{2l})$,

$$\check{A}^n_2(t) \le Q^n_2(t), \text{ and } Q^n_2(\eta^{n,k}_{2l-1}-) \le Q^n_2(\eta^{n,k}_{2l-1}) + 1 \le Q^n_2(t) + 1.$$

Using the above estimates, we have for $t \in [\eta^{n,k}_{2l-1}, \eta^{n,k}_{2l})$,

$$\begin{aligned}\Delta^n(t) &\ge -\frac{1}{\mu^n_2}(C^n - L^n + 2) - 1 - \left|\frac{1}{\mu^n_2} - \frac{\mu_3}{\mu^n_3\mu_2}\right| Q^n_2(t) \\ &\quad + \Delta^n(\eta^{n,k}_{2l-1}-) - C_0 \\ &\ge -\frac{2}{\mu^n_2}(C^n - L^n + 2) - 1 - \left|\frac{1}{\mu^n_2} - \frac{\mu_3}{\mu^n_3\mu_2}\right| (2Q^n_2(t) + 1) \\ &\quad + g_0 - 2C_0.\end{aligned} \tag{6.59}$$

Thus for such t, (5.5) holds with suitable choice of C_1, C_2 as well. The result follows. □

REMARK 6.4. *In Remark 3.6 we conjectured an asymptotically optimal control policy sequence for Case IIC. The main ingredient in the proof of this result, as for the proof of Theorem 3.7, will be the convergence of the scaled workload process $\hat{W}^n$. The weak convergence analysis for $\hat{W}^n_1$ is expected to be similar to that in [4]. The main work will be in the asymptotic analysis of $\hat{W}^n_2$. For that the following representation, analogous to the one in (5.10), will be useful. For $t \ge 0$,*

$$\begin{aligned}&(\hat{W}^n_2(t) - \Psi(\hat{W}^n_1(t)))1_{\{(\hat{W}^n_2(t)-\Psi(\hat{W}^n_1(t)))\ge g_0/\sqrt{n}\}} \\ &= \frac{\hat{X}^n_2(t)}{\mu^n_3} + \frac{\hat{X}^n_3(t)}{\mu^n_3} \\ &\quad - \Psi(\hat{W}^n_1(t))) - (\hat{W}^n_2(t) - \Psi(\hat{W}^n_1(t)))1_{\{(\hat{W}^n_2(t)-\Psi(\hat{W}^n_1(t)))<g_0/\sqrt{n}\}} \\ &\quad + \int_{[0,t)} 1_{\{(\hat{W}^n_2(s)-\Psi(\hat{W}^n_1(s)))\ge g_0/\sqrt{n}\}} d\hat{I}^n_2(s) \\ &\quad + \int_{[0,t)} 1_{\{(\hat{W}^n_2(s)-\Psi(\hat{W}^n_1(s)))<g_0/\sqrt{n}\}} d\hat{I}^n_2(s).\end{aligned}$$

Clearly, the last term on the right side is equal to 0 at $t = 0$, and it is nonnegative and nondecreasing, and increases only when the left side equals 0. Using properties of the one dimensional Skorohod map, the key step in the proof of convergence of $\hat{W}^n_2$ will be arguing that $\int_{[0,\cdot]} 1_{\{(\hat{W}^n_2(s)-\Psi(\hat{W}^n_1(s)))\ge g_0/\sqrt{n}\}} d\hat{I}^n_2(s)$ converges to 0 in distribution the proof of which will require suitable modifications of the arguments in the proof of Theorem 5.5.

APPENDIX A: ONE-DIMENSIONAL SKOROHOD MAP

We recall below the definition and basic properties of the 1-dimensional Skorohod map [26]. Recall $\mathcal{D}_1 = \{x \in \mathcal{D}([0,\infty) : \mathbb{R}) : x(0) \geq 0\}$.

DEFINITION A.1 (One-dimensional Skorohod Problem (SP)). *Let $x \in \mathcal{D}_1$. A pair $(z, y) \in \mathcal{D}([0,\infty) : \mathbb{R}_+) \times \mathcal{D}([0,\infty) : \mathbb{R}_+)$ is a solution of the Skorohod problem for x if the following hold.*

(i) *For all $t \geq 0, z(t) = x(t) + y(t) \geq 0$.*
(ii) *y satisfies the following: (a) $y(0) = 0$, (b) y is nondecreasing, and (c) y increases only when $z = 0$, that is, $\int_{[0,\infty)} z(t)dy(t) = 0$.*

The following proposition summarizes some well known properties of the 1-dimensional SP (see [10] for a proof).

PROPOSITION A.2.

(i) *Let $x \in \mathcal{D}_1$. Then there exists a unique solution $(z, y) \in \mathcal{D}([0,\infty) : \mathbb{R}_+) \times \mathcal{D}([0,\infty) : \mathbb{R}_+)$ of the SP for x, which is given as follows:*

$$y(t) = -\inf_{0\leq s\leq t}(x(s) \wedge 0), \quad z(t) = x(t) - \inf_{0\leq s\leq t}(x(s) \wedge 0), \ t \geq 0.$$

We write $z = \Gamma(x)$, and refer to the map $\Gamma : \mathcal{D}_1 \to \mathcal{D}([0,\infty) : \mathbb{R}_+)$ as the Skorohod map. Let $\mathcal{I} : \mathcal{D}([0,\infty) : \mathbb{R}) \to \mathcal{D}([0,\infty) : \mathbb{R})$ be the identity functional. Then $y = (\Gamma - \mathcal{I})(x)$.

(ii) *The Skorohod map Γ is Lipschitz continuous in the following sense: For all $t \geq 0$ and $x_1, x_2 \in \mathcal{D}_1$,*

$$\sup_{0\leq s\leq t} |\Gamma(x_1)(s) - \Gamma(x_2)(s)| \leq 2 \sup_{0\leq s\leq t} |x_1(s) - x_2(s)|.$$

(iii) *Fix $x \in \mathcal{D}_1$. Let $(z, y) \in \mathcal{D}([0,\infty) : \mathbb{R}_+) \times \mathcal{D}([0,\infty) : \mathbb{R}_+)$ be such that*
   (a) *$z(t) = x(t) + y(t) \geq 0, \ t \geq 0$,*
   (b) *y is nondecreasing with $y(0) = 0$.*

   Then $z(t) \geq \Gamma(x)(t), \ t \geq 0$.

Acknowledgements. We will like to thank Marty Reiman for discussions that led to a simplification in the form of our asymptotically optimal policy.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.

REFERENCES

[1] S. L. Bell and R. J. Williams. Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: Asymptotic optimality of a threshold policy. *Ann. Appl. Probab.*, 11(3):608–649, 2001. MR1865018

[2] S. L. Bell and R. J. Williams. Dynamic scheduling of a parallel server system in heavy traffic with complete resource pooling: Asymptotic optimality of a threshold policy. *Electron. J. Probab.*, 10:1044–1115, 2005.

[3] V. E. Beneš, L. A. Shepp, and H. S. Witsenhausen. Some solvable stochastic control problems. *Stochastics*, 4(1):39–83, 1980/81. MR0587428

[4] A. Budhiraja and A. P. Ghosh. A large deviation approach to asymptotically optimal control of crisscross network in heavy traffic. *The Annals of Applied Probability*, 15(3):1887–1935, 2005.

[5] A. Budhiraja and A. P. Ghosh. Diffusion approximations for controlled stochastic networks: An asymptotic bound for the value function. *Ann. Appl Probab*, 16(4):1962–2006, 2006. MR2288710

[6] A. Budhiraja and K. Ross. Convergent numerical scheme for singular stochastic control with state constraints in a portfolio selection problem. *SIAM J. Control Optim.*, 45(6):2169–2206, 2007. MR2285720

[7] A. Budhiraja and K. Ross. Optimal stopping and free boundary characterizations for some brownian control problems. *Ann Appl. Probab.*, 18:2367–2391, 2008. MR2474540

[8] A. Budhiraja and A. P. Ghosh. Controlled stochastic networks in heavy traffic: convergence of value functions. *Ann. Appl. Probab.*, 22(2):734–791, 2012. MR2953568

[9] A. Budhiraja, A. P. Ghosh, and X. Liu, Scheduling control for Markov modulated single-server multiclass queueing systems in heavy traffic, *Queueing Systems*, 78(1), 57–97, 2014. MR3238008

[10] H. Chen and A. Mandelbeaum. Leontief systems, RBV's and RBM's. In M. H. A. Davis and R. J. Elliott, editors, *Applied Stochastic Analysis*, pages 1–43. Gordon and Breach, 1991.

[11] J. G. Dai and W. Lin. Asymptotic optimality of maximum pressure policies in stochastic processing networks. *Ann. Appl. Probab.*, 18(6):2239–2299, 2008. MR2473656

[12] S. N. Ethier and T. G. Kurtz, *Markov processes: Characterization and convergence*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986.

[13] J. M. Harrison. Brownian models of queueing networks with heterogeneous customer population. In W. Fleming and F. L. Lion, editors, *Stochastic Differential Systems, Stochastic Control Theory and Applications*, pages 147–186. Springer, New York, 1988.

[14] J. M. Harrison. Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review policies. *Ann. Appl. Probab.*, 8(3):822–848, 1998.

[15] J. M. Harrison and M. I. Taksar. Instantaneous control of Brownian motion. *Math. Oper. Res.*, 8(3):439–453, 1983. MR0716123

[16] J. M. Harrison and J. A. Van Mieghem. Dynamic control of Brownian networks: state space collapse and equivalent workload formulations. *Ann. Appl. Probab.*, 7(3):747–771, 1997. MR1459269

[17] J. M. Harrison and L. M. Wein. Scheduling networks of queues: heavy traffic analysis of a simple open network. *Queueing Systems Theory Appl.*, 5(4):265–279, 1989.

[18] Sunil Kumar. Two-server closed networks in heavy traffic: diffusion limits and asymptotic optimality. *Ann. Appl. Probab.*, 10(3):930–961, 2000.

[19] Sunil Kumar and Kumar Muthuraman. A numerical method for solving singular

stochastic control problems. *Oper. Res.*, 52(4):563–582, 2004.
[20] H. J. Kushner and L. F. Martins. Numerical methods for stochastic singular control problems. *SIAM J. Control Optim.*, 29:1443–1475, 1991.
[21] H. J. Kushner and L. F. Martins. Heavy traffic analysis of a controlled multiclass queueing network via weak convergence methods. *SIAM J. Control Optim.*, 34(5):1781–1797, 1996.
[22] L. F. Martins, S. E. Shreve, and H. M. Soner. Heavy traffic convergence of a controlled, multiclass queueing system. *SIAM J. Control Optim.*, 34:2133–2171, 1996.
[23] Kumar Muthuraman and Sunil Kumar. Solving free-boundary problems with applications in finance. *Found. Trends Stoch. Syst.*, 1(4):259–341, 2006. MR2438635
[24] V. Pesic and R. J. Williams, Dynamic scheduling for parallel server systems in heavy traffic: Graphical structure, decoupled workload matrix and some sufficient conditions for solvability of the Brownian control problem, *Preprint.*
[25] S. E. Shreve and H. M. Soner. A free boundary problem related to singular stochastic control. In *Applied stochastic analysis (London, 1989)*, volume 5 of *Stochastics Monogr.*, pages 265–301. Gordon and Breach, New York, 1991.
[26] A. V. Skorohod. Stochastic equations for diffusions in a bounded region. *Theory Probab. Appl.*, (6):264–274, 1961.
[27] H. Mete Soner and S. E. Shreve. Regularity of the value function for a two-dimensional singular stochastic control problem. *SIAM J. Control Optim.*, 27(4):876–907, 1989. MR1001925
[28] P. Yang, H. Chen, and D. Yao. Control and scheduling in a two-station queueing network. *Queueing Syst. Theory Appl.*, 18:301–332, 1994.

Amarjit Budhiraja
Department of Statistics
and Operations Research
University of North Carolina
Chapel Hill, NC 27599
USA
E-mail: budhiraj@email.unc.edu

Xin Liu
Department of Mathematical Sciences
Clemson University
Clemson, SC 29634
USA
E-mail: xliu9@clemson.edu

Subhamay Saha
Department of Mathematics
Indian Institute of Technology Guwahati
Guwahati, Assam 781039
India
E-mail: saha.subhamay@iitg.ernet.in

Downloaded from informs.org by [216.73.217.141] on 17 November 2026, at 12:07 . For personal use only, all rights reserved.