Dynamic Scheduling of a Two-Server Parallel Server System with Complete Resource Pooling and Reneging in Heavy Traffic: Asymptotic Optimality of a Two-Threshold Policy

Samim Ghamami
Department of Economics, University of California, Berkeley, Berkeley, California 94720, samim_ghamami@berkeley.edu

Amy R. Ward
Marshall School of Business, University of Southern California, Los Angeles, California 90089, amyward@usc.edu, http://www-bcf.usc.edu/~amyward/

We consider a dynamic control problem for a parallel server system commonly known as the $N$-system. An $N$-system is a two-server parallel server system with two job classes, one server that can serve both classes, and one server that can only serve one class. We assume that jobs within each class arrive according to a renewal process. The random service time of a job has a general distribution that may depend on both the job’s class and the server providing the service. Each job independently reneges, or abandons the queue without receiving service, if service does not begin within an exponentially distributed amount of time. The objective is to minimize the expected infinite horizon discounted cost of holding jobs in the system and having customers abandon, by dynamically scheduling waiting jobs to available servers.

It is not possible to solve this control problem exactly, and so, we consider an asymptotic regime in which the system satisfies both a heavy traffic and a resource pooling condition. Then, we solve the limiting Brownian control problem, and interpret its solution as a policy in the original $N$-system. We label the servers and job classes so that server 1 can only serve class 1 and server 2 can serve both classes. The policy we propose has two thresholds. There is one threshold on the total number of jobs in the system, and one threshold on the number of class 1 jobs in the system. These thresholds are used to determine which job class server 2 should serve. We show that this proposed policy is asymptotically optimal within a specified class of admissible policies in the heavy traffic limit, and has the same limiting cost as the Brownian control problem solution.

Key words: parallel server system; asymptotically optimal control; approximating diffusion control problem; threshold control; reneging; abandonment

MSC2000 subject classification: Primary: 60K25, 68M20, 90B22, 90B36; secondary: 60F17, 60J70

OR/MS subject classification: Primary: probability, stochastic model applications, queues, limit theorems; secondary: queues, diffusion models

History: Received September 2, 2012. Published online in Articles in Advance May 6, 2013.

1. Introduction. Parallel server systems arise in a variety of applications contexts, including computer systems, manufacturing, and service systems settings. One specific model that is well studied is the $N$-system; see, for example, Harrison [17], Bell and Williams [8], Osogami et al. [26], Tezcan and Dai [32], and Down and Lewis [12]. This is because the $N$-system is one of the simplest parallel server system models that retains much of the complexity inherent in more general models. The $N$-system is one of the canonical parallel server system designs according to Garnett and Mandelbaum [14].

The basic $N$-system consists of two job classes and two servers. Server 1 is dedicated and serves only one job class. Server 2 can serve both job classes. Each time server 2 finishes processing a job, and there are jobs in both buffers, there is a real-time control decision concerning which class server 2 should serve next. In the case that there are only jobs in one buffer, server 2 may either serve that class or idle. In the case that the system is empty, server 2 idles, and there is no control decision.

The basic $N$-system can be generalized by incorporating reneging, i.e., by allowing for some jobs to abandon the system without receiving service. Specifically, we study the system shown in Figure 1. Jobs arrive to buffer $k$, $k = 1, 2$, in accordance with a renewal process having rate $\lambda_k$. We call a job residing in buffer $k$ a class $k$ job. Jobs are served in a first-come, first-served manner within their class. Server 1 can serve class 1 jobs, whereas server 2 can serve both job classes. The mean time for server 1 to serve a class 1 job is $1/\mu_1$, the mean time for server 2 to serve a class 1 job is $1/\mu_2$, and the mean time for server 2 to serve a class 2 job is $1/\mu_3$. We incorporate reneging into the model by assuming that each class $k$ job independently abandons his buffer if his service has not begun within an exponentially distributed amount of time with mean $1/\delta_k$ (meaning that a job in service will not renege). The cost of holding a class $k$ job is $h_k > 0$ per unit time, and the cost of a class $k$ job reneging is $c_k > 0$. The objective is to minimize the expected infinite horizon discounted cost by dynamically deciding which job class server 2 should serve.
This problem is too difficult to solve exactly, and so we solve it asymptotically. We focus on the heavy-traffic parameter regime in Bell and Williams [8]. We assume \( \lambda_1 > \mu_1 \) so that server 2 must help server 1 in order that the number of class 1 jobs in the buffer, and the number of jobs reneging, do not become very large. This is known as a complete resource pooling condition (cf. Harrison [17], Harrison and Lopez [19], Laws [25], Kelly and Laws [23]), because in the heavy-traffic limit, under an asymptotically optimal control policy, the two servers combine to form a single “super server.” We also assume

\[
1 - \frac{\lambda_1 - \mu_1}{\mu_2} \approx \frac{\lambda_2}{\mu_3},
\]

so that the long-run fraction of time server 2 has left over after helping server 1 serve class 1 jobs is approximately the amount that the class 2 jobs require.

Our main results are as follows. First, suppose that

\[
(h_2 + c_2 \delta_2)\mu_3 < (h_1 + c_1 \delta_1)\mu_2.
\]

Then, the \( c\mu \) rule implies that server 2 should always give priority to class 1, because class 1 is the more expensive class. This implies that there are generally many class 2 customers in the system and very few class 1 customers. The rate at which jobs leave the system is then mostly determined by the service rates and the class 2 abandonment rate. When the class 2 abandonment rate \( \delta_2 \) is high compared to \( \delta_1 \), there is a fortunate alignment of cost and drift, so that having server 2 give priority to class 1 jobs simultaneously minimizes cost and also minimizes what is known as the workload in the framework of Harrison and Lopez [19]. The workload is defined as

\[
W(t) := Q_1(t) + \frac{\mu_2}{\mu_3} Q_2(t) = \mu_2 \left( \frac{Q_1(t)}{\mu_2} + \frac{Q_2(t)}{\mu_3} \right),
\]

and so \( W(t)/\mu_2 \) can be directly interpreted as the average time it would take server 2 to serve all jobs in the system. Otherwise, when \( \delta_2 \) is low compared to \( \delta_1 \), having mostly class 2 customers in the system implies a low downward drift, and so does not minimize the average workload in the system in the long run. This implies that in general a static priority policy will not be optimal.

We propose a dynamic priority policy that has two thresholds. There is one threshold \( L \) on queue 1 that serves the same purpose as in Bell and Williams [8] to prevent excessive idling of server 1 caused by server 2 helping “too much.” The other threshold \( M \) is on the workload process and is used to determine which class server 2 prioritizes. More specifically, suppose that \( Q_1(t) \) and \( Q_2(t) \) are the number of class 1 and class 2 jobs, respectively, at time \( t \). If \( Q_1(t) > L \), then server 2 should give priority to class 1 over class 2 when

\[
W(t) \leq M,
\]

and server 2 should give priority to class 2 over class 1 otherwise. If \( Q_1(t) < L \), then server 2 does not serve class 1, even if there are no class 2 jobs present. The threshold levels \( L \) and \( M \) are specified precisely in Definition 6.1 in §6.
We show in §7 that when (2) holds the number of class 1 jobs is very small, and, when (2) does not hold, the number of class 2 jobs is very small. This is desirable because when (1) holds and also \( \delta_2 \) is low compared to \( \delta_1 \) (a condition made precise in (3)), if there are many jobs in the system (more than \( M \)), it is cost effective to have those be class 1 jobs, in order to take advantage of their higher reneging rate and more quickly return the system to a state with fewer jobs. In heavy traffic, the system manager is able to rapidly obtain a completely different system configuration by switching the priority of server 2. In particular, the system manager may move from a state in which all jobs are held in one buffer to a state in which all jobs are held in the other buffer in a very short amount of time. This is because demand is large and service is fast, so that switching the priority of server 2 enables the jobs in one buffer to be served very quickly, while the new arrivals backlog in the other buffer.

The threshold parameter \( M \) is finite in the part of the parameter space in which (1) holds and the reneging parameters \( \delta_1 \) and \( \delta_2 \) are such that

\[
\frac{\delta_1 + \gamma}{\delta_2 + \gamma} > \frac{\mu_2(h_1 + c_1 \delta_1)}{\mu_3(h_2 + c_2 \delta_2)},
\]

where \( \gamma \) is the discount rate. When (1) holds but not (3), \( M \) is infinite. Then, the proposed policy exactly coincides with the policy in Definition 5.1 in Bell and Williams [8]. Otherwise, in the case that (1) does not hold, it is when (3) does not hold, so that

\[
\frac{\delta_2 + \gamma}{\delta_1 + \gamma} > \frac{\mu_1(h_1 + c_1 \delta_1)}{\mu_2(h_2 + c_2 \delta_2)}
\]

that the threshold parameter \( M \) is finite. Otherwise, when (1) does not hold, but (3) does hold, it is infinite.

The remainder of this paper is organized as follows. We first more thoroughly position our paper with respect to the current literature. We next provide the notation and terminology to be used throughout the paper. Then, in §2, we complete the description of the \( N \)-system considered here. This includes a description of the primitive stochastic processes in our model, the scheduling control policy, and a specification of the dynamic equations satisfied by the queue length processes. In §3, we describe the asymptotic regime in which we seek to analyze the performance of control policies for our system. In particular, we specify assumptions on the stochastic primitives that imply our system is in heavy traffic (cf. Harrison [18]), and satisfies the complete resource pooling condition of Harrison and Lopez [19]. Also, we describe the normalization of the queue length process via diffusive scaling, specify the associated cost function, and define an admissible policy. Next, following the general scheme proposed by Harrison [16], in §4, we state the formal Brownian control problem associated with our \( N \)-system, and, in §5, we solve the Brownian control problem. In contrast to Bell and Williams [8], this is a nontrivial control problem to solve, and is an important part of this work. In §6, we translate the solution to the Brownian control problem into a policy for the original system. We then focus on the parameter regime in which conditions (1) and (3) hold, which means that we are in the part of the parameter regime analyzed in Bell and Williams [8] (when (1) holds), but in which an asymptotically optimal policy differs from Bell and Williams [8] (when (3) also holds). In §7, we prove the weak convergence of the diffusion-scaled system processes under our proposed policy to the Brownian control problem solution. This establishes that there is a state-space collapse in that the two-dimensional queueing process can be described from a one-dimensional workload process with state-dependent drift. Then, in §8, we show that no other admissible policy can achieve a lower limiting cost in the heavy-traffic limit, which establishes that our proposed policy is asymptotically optimal. Finally, in §9, we outline how to prove weak convergence and asymptotic optimality in the other parameter regimes, both when (1) holds and (3) does not hold, and when (1) does not hold.

The main technical challenges that arise in the proofs of weak convergence and asymptotic optimality in §§7 and 8 are that (1) the “bang–bang” nature of the control seems to exclude a direct application of the framework of Bramson [10], (2) the proofs require the adaptation of the oscillation inequality in Williams [37] to include a setting where there is reneging, and (3) the fact that the diffusion control problem does not admit a so-called pathwise solution implies that the arguments to prove asymptotic optimality in Bell and Williams [8] cannot be easily extended to the \( N \)-system with reneging.

The appendix provides the following supporting material: a technical argument omitted from the proof of the proposition that shows weak convergence (Proposition 7.2), and the proofs of all the lemmas stated in the paper.

1.1. Literature review. We have already discussed the paper by Bell and Williams [8], and so we do not mention that paper here.

We begin with the paper by Ata and Rubino [2]. Their paper considers a general parallel server system (that need not be an \( N \)-system) with reneging and admission control. There are holding costs, reneging costs, and
costs for denying customers admission. They solve the limiting Brownian control problem using an average cost criterion. They then propose a policy for the general parallel server system that is motivated by the solution to the Brownian control problem. However, they prove neither weak convergence of their proposed policy nor asymptotic optimality. Instead, they verify its effectiveness only by simulation, because the weak convergence proof is very challenging. This paper provides the rigorous weak convergence proof in the case that the parallel server system is an $N$-system, and there is no admission control.

There are two further contributions in comparison to Ata and Rubino [2]. First, there is a contribution in solving the discounted formulation. In the average cost setting, the Hamilton-Jacobi-Bellman (HJB) equation is simpler to solve, because it results in a first-order equation. In contrast, the HJB equation in the discounted setting is a second-order equation. This difference means that it is harder to find a smooth solution to the HJB equation in the discounted setting. Second, we provide a fuller description of the optimal solution for the $N$-system. In particular, we identify the part of the parameter space in which the threshold on the workload level is infinite. Letting $\gamma \downarrow 0$ in (3) shows when we expect the threshold on the workload level to be infinite in the average cost setting. Because the focus of Ata and Rubino [2] is on the more general model, they do not make this condition precise; rather, the reader must determine it for himself, by specializing the Brownian control problem solution they provide in §5 to the $N$-system. (Note that the simulation study they perform in §7 for an $N$-system does satisfy (1) and (3) when $\gamma$ is set to 0, so that both papers are consistent in their recommendation of a dynamic priority threshold policy for that model.)

Our proposed policy in the case that conditions (1) and (3) both hold is similar to the policy proposed in Budhiraja and Ghosh [11], and shown to be asymptotically optimal, to control the crisscross network with no customer reneging. This is because both proposed policies have two threshold levels. However, the proposed policy in Budhiraja and Ghosh [11] has two thresholds levels that are of the same order, whereas our proposed policy has one threshold level that is of a larger order than the other.

It is also worthwhile to compare our work to recent proposed control policies in many-server models that incorporate customer abandonment. Tezcan and Dai [32] show that for the many-server $N$-system (that is, for the model in Figure 1, except that there are two large server pools, and the servers in pool 1 serve at rate $\mu_1$, and the servers in pool 2 serve class 1 at rate $\mu_2$ and class 2 at rate $\mu_3$), when there is the aforementioned fortunate alignment of cost and drift, a static priority policy is asymptotically optimal, and, in contrast to Bell and Williams [8], there is no need for a threshold on queue 1. The policy we propose is similar to the one proposed in Harrison and Zeevi [21] for a multiclass system with a single large server pool in the sense that the class priorities may change as the system state changes. Perry and Whitt [27] propose a threshold policy for a many-server X-model (a generalization of the many-server $N$-system, in which servers in both pools can serve both classes), but the purpose of the threshold is to detect overload conditions caused by an unexpected increase in one class’s arrival rate, and so is different from the purpose of the thresholds in our proposed policy.

Neither of the aforementioned papers prove asymptotic optimality. Atar [3, 4] studies parallel server systems that have a tree structure, and establishes that a policy based on the solution to the relevant diffusion control problem is asymptotically optimal in the Halfin-Whitt many-server heavy-traffic limit; however, this is done under a continuity assumption on the diffusion control problem solution that is not satisfied in our case, because of its bang–bang nature. (In the case of such a discontinuous diffusion control problem solution, the results of Atar would imply some kind of $\epsilon$-asymptotic optimality.)

Finally, the many-server model most closely related to ours is Atar et al. [6]. They establish that in a multiclass system with a single large server pool, and holding costs but no customer abandonment costs, a simple priority rule that ranks classes according to the value of the class’s

\[
\frac{\text{(holding cost per unit time)} \times (\text{service rate})}{\text{abandonment rate}}
\]

is asymptotically optimal on fluid scale in an average cost setting and under overload conditions. Translated to our setting (i.e., ignoring server 1 and focusing on server 2’s decision, and letting $c_1 = c_3 = \gamma = 0$), that rule suggests that server 2 should give static priority to class 1 when $h_2 \mu_3 / \delta_2 < h_1 \mu_2 / \delta_1$. Our proposed policy can be considered as a diffusion-scale refinement of their proposed policy in the conventional heavy-traffic setting. The adjustment to the conventional heavy-traffic setting requires placing a threshold $L$ on queue 1, so that server 2 gives static priority to class 1 whenever queue 1 exceeds the threshold, and to class 2 otherwise. Then their policy becomes our proposed policy in the case that class 2 is cheapest, so that (1) is satisfied ($h_2 \mu_3 < h_1 \mu_2$), and there is an alignment of cost and drift, so that (3) is not satisfied ($h_2 \mu_3 / \delta_2 < h_1 \mu_2 / \delta_1$), because then $M = \infty$.

Note that when (3) is satisfied in addition to (1), so that $M$ is finite, there is no inconsistency with that paper because the threshold level $M$ will disappear on fluid scale.
1.2. Notation and terminology. The m-dimensional \((m \geq 1)\) Euclidean space will be denoted by \(\mathbb{R}^m\) and \(\mathbb{R}_+\) will denote \([0, \infty)\). Let \(|\cdot|\) denote the norm on \(\mathbb{R}^m\) given by \(|x| := \sum_{i=1}^{m} |x_i|\) for \(x \in \mathbb{R}^m\). Vectors in \(\mathbb{R}^m\) should be treated as column vectors unless indicated otherwise, inequalities between vectors should be interpreted componentwise, the transpose of a vector \(a\) will be denoted by \(a'\), the diagonal matrix with the entries of a vector \(a\) on its diagonal will be denoted by \(\text{diag}(a)\), and the dot product of two vectors \(a\) and \(b\) will be denoted by \(a \cdot b\). The identity function is denoted by \(e\), so that \(e(t) = t\) for all \(t \geq 0\). For \(x \in \mathbb{R}_+\), let \([x]\) denote the integer part of \(x\).

For each positive integer \(m\), let \(D^m\) be the set of all functions \(\omega: \mathbb{R}_+ \rightarrow \mathbb{R}^m\) that are right continuous on \(\mathbb{R}_+\) and have finite left limits on \((0, \infty)\). The member of \(D^m\) that stays at the origin in \(\mathbb{R}^m\) for all time will be denoted by \(0\). For \(\omega \in D^m\) and \(t \geq 0\), let

\[\|\omega\| := \sup_{a \in [0, t]} |\omega(s)|,\]

and for \(0 \leq t_1 \leq t_2 < \infty\), let

\[\text{Osc}(\omega, [t_1, t_2]) := \sup_{t_1 \leq t \leq t_2} |\omega(t) - \omega(s)|.\]

Consider \(D^m\) to be endowed with the usual Skorokhod \(J_1\)-topology (cf. Billingsley [9] or Ethier and Kurtz [13]). Let \(\mathcal{R}^m\) denote the Borel \(\sigma\)-algebra on \(D^m\) associated with the \(J_1\)-topology. This is the same \(\sigma\)-algebra as the one generated by the coordinate maps; i.e., \(\mathcal{R}^m = \sigma(\omega(s): 0 \leq s < \infty)\). All of the continuous-time processes in this paper will be assumed to have sample paths in \(D^m\) for some \(m \geq 1\).

Suppose \(\{W^n\}_{n=1}^{\infty}\) is a sequence of processes with sample paths in \(D^m\) for some \(m \geq 1\). Then we say that \(\{W^n\}_{n=1}^{\infty}\) is tight if and only if the probability measures induced by the \(W^n\)'s on \((D^m, \mathcal{R}^m)\) form a tight sequence; i.e., they form a weakly relatively compact sequence in the space of probability measures on \((D^m, \mathcal{R}^m)\). The notation \(W^n \Rightarrow W^n\), where \(W^n\) is a process with sample paths in \(D^m\), will mean that the probability measures induced by the \(W^n\)'s on \((D^m, \mathcal{R}^m)\) converge weakly to the probability measure on \((D^m, \mathcal{R}^m)\) induced by \(W^n\). If for each \(n\), \(W^n\) and \(W\) are defined on the same probability space, we say that \(W^n\) converges to \(W\) uniformly on compact time intervals in probability if \(P(\|W^n - W\|_1 \geq \varepsilon) \rightarrow 0\) as \(n \rightarrow \infty\) for each \(\varepsilon > 0\) and all \(t \geq 0\). The abbreviations “u.o.c.,” “i.p.,” and “a.s.” refer, respectively, to uniformity on compact intervals, in probability, and almost surely.

2. The \(N\) system. We follow the notation and model assumptions in Bell and Williams [8], except that we add additional notation necessary for the model to include reneging. Subsection 2.1 sets up the stochastic primitives in our model formulation, and §2.2 details the scheduling control and system evolution equations. We delay the precise definition of the cost function until §3, because it is formulated in terms of normalized queue lengths where the normalization is on diffusion scale, commensurate with the heavy-traffic limiting regime in which we consider our model.

2.1. Stochastic primitives. All random variables and stochastic processes in our model are assumed to be defined on a complete and filtered probability space \((\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)\). The expectation operator under \(P\) will be denoted by \(E\) and \(P(A, B)\) will mean \(P(A \cap B)\).

For \(k = 1, 2\), we take as given a sequence of strictly positive i.i.d. random variables \(\{u_k(i), i = 1, 2, \ldots\}\) with mean \(\lambda_k^{-1} \in (0, \infty)\) and squared coefficient of variation \(\alpha_k^2 \in [0, \infty)\). For \(i = 1, 2, \ldots\), we interpret \(u_k(i)\) as the time between the arrival of the \((i-1)\)st and the \(i\)th arrival to class \(k\), where the “0th arrival” occurs at time zero. Setting \(\xi_k(0) = 0\) and

\[\xi_k(n) := \sum_{i=1}^{n} u_k(i), \quad \text{for } n = 1, 2, \ldots,\]

we define

\[A_k(t) := \sup\{n \geq 0: \xi_k(n) \leq t\} \quad \text{for all } t \geq 0.\]

Then \(A_k\) is a renewal process, \(A_k(t)\) counts the number of arrivals to class \(k\) that have occurred in \([0, t]\), and \(A_k\) is the long-run arrival rate to class \(k\).

For \(j = 1, 2, 3\), we take as given a sequence of strictly positive, i.i.d. random variables \(\{v_j(i), i = 1, 2, \ldots\}\) with mean \(\mu_j^{-1} \in (0, \infty)\) and squared coefficient of variation \(\beta_j^2 \in [0, \infty)\). For each \(j\), we interpret \(v_j(i), i \geq 1\), as the amount of service time required by the \(i\)th job to be processed by activity \(j\), where

- activity 1 = processing of class 1 jobs by server 1,
- activity 2 = processing of class 1 jobs by server 2,
- activity 3 = processing of class 2 jobs by server 2.
Note that $\mu_j$ is the long-run rate at which activity $j$ can process its associated class if the associated server works continuously and exclusively on this activity. For $j = 1, 2, 3$, let $\eta_j(0) := 0$,

$$\eta_j(n) := \sum_{i=1}^{n} v_j(i), \quad \text{for } n = 1, 2, \ldots,$$

and

$$S_j(t) := \sup\{n \geq 0: \eta_j(n) \leq t\} \quad \text{for all } t \geq 0.$$

Then, $S_1$, $S_2$, $S_3$ are renewal processes and $S_j(t)$ represents the number of class 1 jobs that server 1 could complete if that server worked continuously in $[0, t]$, and for $j = 2, 3$, $S_j(t)$ is the number of class $j$ jobs that server 2 could complete if that server worked continuously and exclusively on class $j$ jobs in $[0, t]$.

Also, let $N_1$ and $N_2$ be two independent, unit-rate Poisson processes. We will use $N_1$ and $N_2$ to track the cumulative number of jobs that have reneged from each buffer.

We assume that the interarrival time sequences $\{u_k(i), i = 1, 2, \ldots\}$, $k = 1, 2$, service time sequences $\{v_j(i), i = 1, 2, \ldots\}$, $j = 1, 2, 3$, and Poisson processes $N_k$, $k = 1, 2$, are all mutually independent.

### 2.2. Scheduling control.

Scheduling control of the system is exerted through a three-dimensional service time allocation process

$$T(t) := (T_1(t), T_2(t), T_3(t))^\prime, \quad t \geq 0.$$

For $j = 1, 2, 3$, $T_j(t)$ is the cumulative amount of service time devoted to activity $j$ in the time interval $[0, t]$. Then

$$I_1(t) := t - T_1(t) \quad (4)$$

is the cumulative idle time of server 1 up to time $t$,

$$I_2(t) := t - T_2(t) - T_3(t) \quad (5)$$

is the cumulative idle time of server 2 up to time $t$, $S_j(T_j(t))$ is the number of jobs served by server 1 up to time $t$, and $S_j(T_j(t))$ is the number of class $j$ jobs served by server 2, for $j = 2, 3$, up to time $t$. Let $\dot{T}_j(t)$ be the right derivative of $T_j$ at $t > 0$. Note that $\dot{T}_j(t)$ exists for all $t > 0$ and $\dot{T}_j(t) = (d/dt)T_j(t)$ whenever the derivative exists. Then, $\dot{T}_j(t)$ represents whether or not activity $j$ is being undertaken at time $t$. The number of class $k$, $k = 1, 2$, jobs at time $t$, assuming that the system is initially empty, is

$$Q_1(t) := A_1(t) - S_1(T_1(t)) - S_2(T_2(t)) - R_1(t) \quad (6)$$

$$Q_2(t) := A_2(t) - S_1(T_3(t)) - R_2(t) \quad (7)$$

where

$$R_1(t) := N_1\left(\int_0^t \delta_1(Q_1(s) - 1[\dot{T}_1(s) > 0] - 1[\dot{T}_2(s) > 0]) \, ds\right) \quad (8)$$

$$R_2(t) := N_2\left(\int_0^t \delta_2(Q_2(s) - 1[\dot{T}_1(s) > 0]) \, ds\right) \quad (9)$$

are the cumulative number of class 1 and 2 jobs that have reneged by time $t$, and $\delta_1$ and $\delta_2$ are the respective class 1 and class 2 reneging rates. Note that a job in service does not renege. Also note that the definitions of $R_1$ and $R_2$ in (8) and (9) model customers that renege independently, after an exponentially distributed amount of time, with mean $1/\delta_1$ for class 1 customers and $1/\delta_2$ for class 2 customers. This is because conditional on having $x$ class $k$ customers, $k \in \{1, 2\}$, in the system, not including any customers in service, the time until the next class $k$ customer reneges is exponential with parameter proportional to $x$.

The scheduling control $T$ must satisfy the following properties. For $j = 1, 2, 3$ and $k = 1, 2$, and $I(t) := (I_1(t), I_2(t))^\prime$ and $Q(t) := (Q_1(t), Q_2(t))^\prime$ for all $t \geq 0$ that satisfy (4)–(7),

$$T_j(\cdot) \text{ is continuous and nondecreasing with } T_j(0) = 0, \quad (10)$$

$$I_k(\cdot) \text{ is continuous and nondecreasing with } I_k(0) = 0, \quad (11)$$

$$Q_j(t) \geq 0 \quad \text{for all } t \geq 0. \quad (12)$$

Note that conditions (11)–(12) imply that for $j = 1, 2, 3$, $T_j$ is uniformly Lipschitz continuous with a Lipschitz constant less than or equal to one.
3. Heavy-traffic assumptions and the control problem. The problem of finding a control policy that minimizes costs associated with holding jobs and reneging is not solvable via an exact analysis. Therefore, we perform an asymptotic analysis. In particular, we consider the asymptotic regime associated with heavy-traffic limit theorems in which the queue length process is normalized with diffusive scaling. This corresponds to viewing the system over long intervals of time of order \( r^2 \) (where \( r \) will tend to infinity in the asymptotic limit) and regarding a single job as only having a small contribution to the overall cost, where this is quantified to be of order \( 1/r \). Formally, we consider a sequence of \( N \)-systems indexed by \( r \), where \( r \) tends to infinity through a sequence of values in \([1, \infty)\). These systems all have the same basic structure as that described in the last section; however, the arrival, service, and reneging rates, scheduling control, and cost function (defined below) may vary with \( r \). We shall indicate the dependence of relevant parameters and processes on \( r \) by appending a superscript to them. For example, for each fixed \( r \), \( T' \) is the control policy in the \( r \)th system, \( Q' \) is the associated queue lengths, and \( W' = Q'_i + (\mu'_j/\mu'_k)Q'_k \) is the associated workload. It is also useful to define the following scaled processes, for each \( t \geq 0 \),

\[
\bar{T}'(t) := r^{-2}T'(r^2t),
\]

\[
\bar{Q}'(t) := r^{-2}Q'(r^2t),
\]

\[
\bar{A}'_k(t) := r^{-1}(A'_k(r^2t) - \lambda'_k r^2 t), \quad k = 1, 2, 
\]

\[
\bar{S}'_j(t) := r^{-1}(S'(r^2t) - \mu'_j r^2 t), \quad j = 1, 2, 3, 
\]

\[
\bar{Q}'_k(t) := r^{-1}Q'_k(r^2t), \quad k = 1, 2 
\]

\[
\bar{I}'_k(t) := r^{-1}I'_k(r^2t), \quad k = 1, 2 
\]

\[
\bar{N}'_k(t) := r^{-1}(N_k(r^2t) - r^2 t).
\]

We assume that the interarrival and service times are given for each \( r, k = 1, 2, j = 1, 2, 3, i = 1, 2, \ldots \), by

\[
u'_k(i) := 1/\lambda'_k \bar{u}_k(i), \quad \nu'_j(i) := 1/\mu'_j \bar{v}_j(i),
\]

where the \( \bar{u}_k(i) \), \( \bar{v}_j(i) \) do not depend on \( r \), have mean one, and squared coefficients of variation \( \alpha^2_{\bar{u}} \), \( \beta^2_{\bar{v}} \), respectively. Also, the sequences \( \{\bar{u}_k(1), \bar{u}_k(2), \ldots \}, k \in \{1, 2\} \), and \( \{\bar{v}_j(1), \bar{v}_j(2), \ldots \}, j \in \{1, 2, 3\} \), are independent sequences of i.i.d. random variables. The Poisson processes \( N_k \) and \( N_0 \) do not vary with \( r \), but the reneging rates \( \delta'_1 \) and \( \delta'_2 \) do. The following assumption is Assumption 3.1 in Bell and Williams [8].

**Assumption 3.1.** There are strictly positive finite constants \( \lambda_1, \lambda_2, \mu_1, \mu_2, \mu_3 \) such that

(i) \( \lambda_1 > \mu_1 \),

(ii) \( 1 - (\lambda_1 - \mu_1)/\mu_2 = \lambda_2/\mu_3 \),

and, as \( r \to \infty \),

(iii) \( \lambda'_k \to \lambda_k \), \( k = 1, 2, \)

(iv) \( \mu'_j \to \mu_j \), \( j = 1, 2, 3, \)

(v) \( r \mu'_j(\lambda'_j - \mu'_j)/\mu'_j - (\lambda'_1 - \mu'_1)/\mu_2 \to 0 \),

(vi) \( r \mu'_j(\lambda'_j/\mu'_j - \lambda_2/\mu_3) \to 0 \).

Note that (i) in Assumption 3.1 is known as the complete resource pooling assumption in the literature.

Note that we assume the convergences in (v) and (vi) are to zero for convenience in solving the diffusion control problem in §5; however, we expect the same results structure to hold when the convergences in (v) and (vi) are to any real numbers.

We also require an assumption on the reneging rates. To motivate this assumption, note that in heavy traffic (i.e., under Assumption 3.1), we expect that there are of order \( r \) customers in the system. Because each customer reneges independently, in a time of order \( r^2 \), there will be of order \( r \times r^2 = r^3 \) customers that have reneged. This suggests that when the system state is scaled by \( 1/r \), the customer reneging parameters \( \delta'_1 \) and \( \delta'_2 \) should be of order \( 1/r^2 \) in order that the instantaneous system abandonment rates are finite and nonzero.

**Assumption 3.2.** There are strictly positive finite constants \( \delta'_1, \delta'_2 \) such that as \( r \to \infty \)

\[
r^2 \delta'_1 \to \delta'_1 \quad \text{and} \quad r^2 \delta'_2 \to \delta'_2.
\]
Note that Assumption 3.2 is consistent with the heavy-traffic regimes in Ward and Glynn [33] for a $M/M/1$ system with exponentially distributed reneging times, and in Reed and Ward [29] for a $GI/GI/1$ system with generally distributed reneging times.

This structure is a convenient means of allowing the sequence of systems to approach heavy traffic by simply changing arrival, service, and reneging rates while keeping the underlying sources of variability $\tilde{\mu}_k(i), \tilde{\upsilon}_j(i), \tilde{N}_1,$ and $\tilde{N}_2$ fixed. For a first reading, the reader may like to simply choose $\lambda_k^r = \lambda_k, \mu_j^r = \mu_j,$ and $\delta_k^r = \delta_k/r^2$ for $k = 1, 2$ and $j = 1, 2, 3,$ and for all $r.$

For each fixed $r,$ the cost associated with the use of control $T^r$ is

$$\tilde{J}^r := \mathbb{E}\left(\int_0^\infty e^{-\gamma v} v \cdot \tilde{Q}^r(t) \, dt\right),$$

(13)

where $\gamma > 0$ is a constant discount rate, and $v := (h_1 + c_1 \delta_1, h_2 + c_2 \delta_2)$ is the vector of holding costs that includes both the cost of physically having a job in the buffer (through $h_1$ and $h_2$) and reneging costs (through $c_1 \delta_1$ and $c_2 \delta_2.$) Note that the $\delta_k$ for $k = 1, 2$ in the above is the limit in Assumption 3.2 as $r$ becomes large, whereas the $h_k$ and $c_k$ for $k = 1, 2$ are taken to represent costs associated with the diffusion-scaled queue length.

We are interested in scheduling controls that do not anticipate the future. To define this notion precisely, let

$$\theta_k^r(t) := \inf\{u \geq t : A_k^r(u) - A_k^r(u^-) > 0\}$$

and

$$\sigma_k^r(t) := \inf\{u \geq t : S_k^r(u) - S_k^r(u^-) > 0\},$$

so that $\theta_k^r$ is the time of the first class $k$ arrival no earlier than $t$ and $\sigma_k^r(t)$ is the time of the first service completion by using activity $j$ no earlier than $t.$ Then, similar to (9) and (10) in Atar et al. [7], let

$$\mathcal{F}_t^r := \sigma\left\{A_k^r(s), S_k^r(T_j^r(s)), R_k^r(s), Q_j^r(s), T_j^r(s) : k \in \{1, 2\}, j \in \{1, 2, 3\}, s \leq t\right\}$$

$$\mathcal{B}_t^r := \sigma\left\{\int_0^t \delta_k^r(Q_j^r(s) - 1[\tilde{T}_j^r(s) > 0] - 1[\tilde{T}_j^r(s) > 0]) \, ds + u - R_k^r(t)\bigg| T_j^r(s) \right\},$$

for all $u \geq 0.$ In the above display, $\mathcal{F}_t^r$ represents the information available at time $t,$ and $\mathcal{B}_t^r$ represents future information. It is necessary to use $\theta_k^r$ and $\sigma_k^r$ in the definition of $\mathcal{B}_t^r$ because $A_k^r$ and $S_k^r$ are renewal processes, and, as such, at a given time $t > 0,$ there can be some information available about when the next event will occur.

The control problem we would like to solve is as follows.

**Definition 3.1 (The Control Problem for the $\mathcal{F}^{th}$ System):**

minimize $\tilde{J}^r(T^r)$

by choosing an admissible scheduling control

$$(T_1^r, T_2^r, T_3^r),$$

where an admissible scheduling control is one that satisfies that satisfies (10)–(12), has $\mathcal{F}_t^r$ independent of $\mathcal{B}_t^r$ for each $t > 0,$ and under which for $k \in \{1, 2\}$ and each $t > 0,$

$$N_1\left(\int_0^t \delta_k^r(Q_j^r(s) - 1[\tilde{T}_j^r(s) > 0] - 1[\tilde{T}_j^r(s) > 0]) \, ds + u - R_k^r(t)\right),$$

$$N_2\left(\int_0^t \delta_k^r(Q_j^r(s) - 1[\tilde{T}_j^r(s) > 0]) \, ds + u - R_k^r(t)\right)$$

have the same laws as $N_1(\cdot)$ and $N_2(\cdot),$ respectively (i.e., the same laws as unit-rate Poisson processes).
It appears impossible to find the control T' that minimizes \( \hat{J} \) for a fixed \( r \). Instead, we will solve the control problem in Definition 3.1 asymptotically, as \( r \) becomes large.

In addition to Assumptions 3.1 and 3.2, we make the following exponential moment assumptions that ensure that certain large deviations estimates hold for the renewal processes \( \lambda_k, S_j \) associated with the interarrival and service times. This is Assumption 3.3 in Bell and Williams [8].

**Assumption 3.3.** For \( k = 1, 2, \) \( j = 1, 2, 3, \) and all \( i \geq 1 \), let
\[
\mu_k(i) = \frac{1}{\mu_k} \tilde{u}_k(i), \quad \mu_j(i) = \frac{1}{\mu_j} \tilde{u}_j(i).
\]
Assume that there is a nonempty open neighborhood, \( \Theta \), of 0 in \( \mathbb{R} \) such that for all \( l \in \Theta \) and all \( i \geq 1 \),
\[
\Lambda_k^l(i) := \log E(e^{\lambda_k(l) i}) < \infty, \quad \text{for} \ k = 1, 2,
\]
and
\[
\Lambda_j^l(i) := \log E(e^{\lambda_j(l) i}) < \infty, \quad \text{for} \ j = 1, 2, 3.
\]

We let \( \Lambda_k^l \), \( k \in \{1, 2\} \), and \( \Lambda_j^l \), \( j \in \{1, 2, 3\} \), be the Legendre-Fenchel transforms of \( \Lambda_k^l \) and \( \Lambda_j^l \); i.e.,
\[
\Lambda_k^l := \sup_{x \in \mathbb{R}} (lx - \Lambda_k^l(l)), \quad x \in \mathbb{R}
\]
\[
\Lambda_j^l := \sup_{x \in \mathbb{R}} (lx - \Lambda_j^l(l)), \quad x \in \mathbb{R}.
\]

Our final assumption emphasizes that the evolution equations for the number of class 1 and class 2 jobs in the system in (6) and (7) has the system initially empty. This assumption is made for convenience in presentation. The emphasis is for transparency, so that it is clear to the reader that our weak convergence results in §7 hold as stated and that there is no need for a restriction to \( (0, \infty) \).

**Assumption 3.4.** The system is initially empty; that is, \( Q_1(0) = Q_2(0) = 0 \).

Note that we have not made the Assumption 3.2 in Bell and Williams [8], that \( h_1 \mu_2 \geq h_2 \mu_1 \). This is because the appropriate division of the parameter space, that was outlined in the introduction, will become apparent in §5, where we solve the (formal) Brownian control problem associated with our sequence of \( N \)-systems.

4. The Brownian control problem and equivalent workload formulation. We formally show how to derive the approximating Brownian control problem that arises in the heavy-traffic limit when passing to the limit as \( r \to \infty \) in the control problem in Definition 3.1. This requires the assumption that in fluid or law of large numbers scale, the scheduling control achieves the long-run rates required in Assumption 3.1 for a balanced system in the heavy-traffic limit; i.e., that
\[
\hat{T}' \Rightarrow \hat{T}^*, \quad \text{as} \ r \to \infty,
\]
for
\[
\hat{T}^*(t) := \left( t, \frac{\lambda_1 - \mu_1 t}{\mu_2}, \frac{\lambda_2 - \mu_2 t}{\mu_3} \right).
\]

The preceding assumption implies that the fluid scaled queue length process converges to zero in the limit. That is,
\[
(\hat{Q}_1', \hat{Q}_2') \Rightarrow (0, 0), \quad \text{as} \ r \to \infty.
\]

The first step is to note that the Equations (6)–(7) yield the following expressions for the normalized (diffusion scaled) queue length processes:
\[
\hat{Q}_1'(t) := \hat{Q}_1'(t) - \hat{S}_1'(\hat{T}_1'(t)) - \hat{S}_2'(\hat{T}_2'(t))
\]
\[
- \int_0^t (r^2 \hat{\delta}_1') \left( \hat{Q}_1'(s) - \frac{1}{r^2} \mathbf{1}[\hat{T}_1'(r^2 s) > 0] \right) ds
\]
\[
+ r(\lambda_1 t - \mu_1 \hat{T}_1'(t) - \mu_2 \hat{T}_2'(t))
\]
\[
- \int_0^t (r^2 \hat{\delta}_2') \left( \hat{Q}_2'(s) - \frac{1}{r^2} \mathbf{1}[\hat{T}_2'(r^2 s) > 0] \right) ds
\]
\[
\hat{Q}_2'(t) := \hat{Q}_2'(t) - \hat{S}_1'(\hat{T}_1'(t)) - \hat{S}_2'(\hat{T}_2'(t))
\]
\[
- \int_0^t (r^2 \hat{\delta}_2') \left( \hat{Q}_2'(s) - \frac{1}{r^2} \mathbf{1}[\hat{T}_2'(r^2 s) > 0] \right) ds
\]
\[
+ r(\lambda_2 t - \mu_2 \hat{T}_1'(t) - \mu_3 \hat{T}_3'(t))
\]
\[
+ \int_0^t (r^2 \hat{\delta}_3') \left( \hat{Q}_3'(s) - \frac{1}{r^2} \mathbf{1}[\hat{T}_3'(r^2 s) > 0] \right) ds.
\]
On combining Assumption 3.1 with the finite variance and mutual independence of the stochastic primitives \( \{ \tilde{u}_i(t), i = 1, 2, \ldots \}, \{ \tilde{v}_j(t), i = 1, 2, \ldots \}, N_k \) for \( k = 1, 2 \) and \( j = 1, 2, 3 \), we may deduce from renewal process functional central limit theorems (see, for example, Whitt [36, Corollary 7.3.1]) that
\[
(\hat{A}, \hat{S}, \hat{N}) \Rightarrow (\hat{A}, \hat{S}, \hat{N}) \quad \text{as } r \rightarrow \infty,
\]
where \( \hat{A}, \hat{S}, \hat{N} \) are mutually independent; \( \hat{A} \) is a two-dimensional driftless Brownian motion that starts from the origin and has diagonal covariance matrix \( \text{diag}(\lambda_1 \alpha_1^2, \lambda_2 \alpha_2^2) \); \( \hat{S} \) is a three-dimensional driftless Brownian motion that starts from the origin and has diagonal covariance matrix \( \text{diag}(\mu_1 \beta_1^2, \mu_2 \beta_2^2, \mu_3 \beta_3^2) \); and \( \hat{N} \) is a two-dimensional driftless Brownian motion that starts from the origin and has diagonal covariance matrix that is the identity matrix. Hence, the normalized queue length processes in (14) and (15) can be written in terms of the diffusion-scaled processes that we expect to converge to a two-dimensional Brownian motion
\[
\hat{X}_1(t) := \hat{A}_1(t) - \hat{S}_1(t) - \hat{N}_1(t) + \tilde{Y}_1(t)
\]
and using Assumptions 3.1 and 3.2, we arrive at the following Brownian control problem.

**Definition 4.1 (Brownian Control Problem (BCP)).** Let \( (\Omega, \mathcal{F}, \{ \mathcal{F}_t \}, P) \) be a complete, filtered probability space and \( \tilde{X} \) be a two-dimensional \( \mathcal{F}_t \)-Brownian motion with drift zero that starts from the origin and has diagonal covariance matrix \( \text{diag}(\lambda_1 \alpha_1^2 + \mu_1 \beta_1^2, \lambda_2 \alpha_2^2 + \mu_2 \beta_2^2, \lambda_3 \alpha_3^2 + \mu_3 \beta_3^2) \). The Brownian control problem is to minimize
\[
J(\tilde{Y}) = \mathbb{E} \left( \int_0^\infty e^{-\gamma v} \cdot \tilde{Q}(t) dt \right)
\]
using a three-dimensional control process \( \tilde{Y} = (\tilde{Y}_1, \tilde{Y}_2, \tilde{Y}_3)' \) such that
\[
\tilde{Q}_1(t) := \tilde{X}_1(t) - \int_0^t \delta_1 \tilde{Q}_1(s) ds + \mu_1 \tilde{Y}_1(t) + \mu_2 \tilde{Y}_2(t) \geq 0 \quad \text{for all } t \geq 0,
\]
\[
\tilde{Q}_2(t) := \tilde{X}_2(t) - \int_0^t \delta_2 \tilde{Q}_2(s) ds + \mu_3 \tilde{Y}_3(t) \geq 0 \quad \text{for all } t \geq 0,
\]
\[
\tilde{I}_1 := \tilde{Y}_1 \text{ is nondecreasing}, \quad \tilde{I}_1(0) = 0,
\]
\[
\tilde{I}_2 := \tilde{Y}_2 + \tilde{Y}_3 \text{ is nondecreasing}, \quad \tilde{I}_2(0) = 0,
\]
A control \( \tilde{Y} \) is admissible if it is \( \mathcal{F}_t \)-adapted, and satisfies (22)–(25).
Let $J^* = \inf J(\tilde{Y})$, where the infimum is taken over all admissible controls for the BCP. The admissible control $\tilde{Y}^*$ is optimal if $J(\tilde{Y}^*) = J^*$.

The key to solving the Brownian control problem in Definition 4.1 is to derive the equivalent workload formulation, which follows the general procedure outlined in Harrison and Mieghem [20] and is very close to the specific procedure in Ata and Rubino [2]. For $\tilde{Q}$ and $I$ as in Definition 4.1, define the workload process

$$
\tilde{W}(t) := y \cdot \tilde{Q}(t)
= \tilde{\xi}(t) - \int_0^t \left( \delta_1 \tilde{Q}_1(s) + \frac{\mu_2}{\mu_3} \delta_2 \tilde{Q}_2(s) \right) ds + \tilde{L}(t),
$$

(26)

where $y := (1, \mu_2/\mu_3)$, $\tilde{\xi} := y \cdot \tilde{X}$, and

$$
\tilde{L}(t) = \mu_1 \tilde{I}_1(t) + \mu_2 \tilde{I}_2(t) \quad \text{for all } t \geq 0.
$$

(27)

Note that the process $\tilde{\xi}$ is a Brownian motion that has zero drift and variance

$$
\sigma^2 := (\lambda_1 \alpha_1^2 + \mu_1 \beta_1^2 + \beta_1^2 (\lambda_1 - \mu_1)) + \left( \frac{\mu_2}{\mu_3} \right)^2 \lambda_2 (\alpha_2^2 + \beta_2^2).
$$

(28)

The Equation (26) specifies the workload process $\tilde{W}$ as a function of $\tilde{Q}$. In the equivalent workload formulation, we would like to think of the reverse; that is, we would like to start with $\tilde{W}$, and have $\tilde{Q}$ be a (possibly time-dependent) function of $\tilde{W}$. This is because we would prefer to solve a one-dimensional control problem involving $\tilde{W}$ than a two-dimensional control problem involving $\tilde{Q}$. Then, the objective is to state a one-dimensional control problem involving $\tilde{W}$ that is in some sense equivalent to the BCP (see Proposition 4.1 below). To do this, the first step is to set

$$
\mathcal{A}(w) := \left\{ q \in \mathbb{N}_+^2 : q_1 + \frac{\mu_2}{\mu_3} q_2 = w \right\} \quad \text{for all } w \geq 0.
$$

Then, let $q^* : \mathbb{N}_+ \to \mathbb{N}_+^2$ be a queue-length configuration function that decides how to divide the workload among the two classes at every time $t \geq 0$ and for every workload $w \geq 0$, so that we can let $\tilde{Q}(t) = q(t, \tilde{W}(t))$ for each $t \geq 0$ for some $q \in \mathcal{A}(\tilde{W}(t))$. That observation allows us to write both the objective function of the BCP and the workload process in (26) in terms of $\tilde{W}$ solely, which leads to the workload control problem defined in Definition 4.2. Its equivalence with the BCP is established in Proposition 4.1.

**Definition 4.2 (Workload Control Problem (WCP)).** Define the functions $f : \mathbb{N}_+^2 \to \mathbb{N}_+$ and $g : \mathbb{N}_+^2 \to \mathbb{N}_+$ as follows:

$$
f(q) := \delta_1 q_1 + \frac{\mu_2}{\mu_3} \delta_2 q_2
$$

and

$$
g(q) := v \cdot q = (h_1 + c_1 \delta_1) q_1 + (h_2 + c_2 \delta_2) q_2.
$$

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ be a complete filtered probability space and let $\tilde{\xi}$ be a one-dimensional $\mathcal{F}_t$-Brownian motion with drift zero and variance defined in (28). The workload control problem is to minimize

$$
\mathcal{J}_w(\tilde{\xi}, \tilde{L}) := \mathbb{E}\left( \int_0^\infty e^{-\gamma t} g(q(t, \tilde{W}(t))) \, dt \right)
$$

using a control process $\tilde{\xi}$ that satisfies

$$
\tilde{\xi}_k(t) = \int_0^t q_k(s, \tilde{W}(s)) \, ds \quad \text{for every } t \geq 0, \quad k \in \{1, 2\},
$$

(29)

$$
q(t, \tilde{W}(t)) \in \mathcal{A}(\tilde{W}(t)),
$$

(30)

$$
\tilde{W}(t) = \tilde{\xi}(t) - \int_0^t f(q(s, \tilde{W}(s))) \, ds + \tilde{L}(t) \geq 0 \quad \text{for all } t \geq 0,
$$

(31)

$$
\tilde{L}(0) = 0, \quad \text{and is nondecreasing.}
$$

(32)

A control $\tilde{\xi}, \tilde{L}$ is admissible for the workload control problem if it is $\mathcal{F}_t$-adapted, has $\tilde{W}$ and $\tilde{L}$ that are right continuous with left limits, and satisfies (29)–(32).
Let \( J_w^* = \inf J_w(\tilde{\xi}, \tilde{L}) \), where the infimum is taken over all admissible controls for WCP. The admissible control \((\tilde{\xi}^*, \tilde{L}^*)\) is optimal if \( J_w(\tilde{\xi}^*, \tilde{L}^*) = J_w^* \). We let \( \tilde{W}^* \) be the workload process under a specified optimal control \((\tilde{\xi}^*, \tilde{L}^*)\).

**Proposition 4.1.** Every admissible control \( \tilde{Y} \) for the Brownian control problem in Definition 4.1 yields an admissible control \((\tilde{\xi}, \tilde{L})\) for the workload control problem in Definition 4.2, and \( J(\tilde{Y}) = J_w(\tilde{\xi}, \tilde{L}) \). Also, given any admissible control \((\tilde{\xi}, \tilde{L})\) for the workload control problem in Definition 4.2, there exists an admissible control \( \tilde{Y} \) for the Brownian control problem in Definition 4.1, and \( J(\tilde{Y}) \leq J_w(\tilde{\xi}, \tilde{L}) \).

The proof of Proposition 4.1 is similar to Harrison and Mieghem [20, proof of Theorem 2, p. 757], and so it omitted. See also Ata and Rubino [2, proofs of Propositions 2 and 3] for a setting more similar to ours.

**Remark 4.1.** It follows from Proposition 4.1 that \( J(\tilde{Y}^*) = J_w(\tilde{\xi}^*, \tilde{L}^*) \).

5. The workload control problem solution. We show that an optimal control for the WCP has

\[
\int_0^t \tilde{W}^*(s) d\tilde{L}^*(s) = 0, \quad \tilde{L}^*(0) = 0, \quad \tilde{L}^* \text{ is continuous and nondecreasing,} \tag{33}
\]

and that there exist \( b^*, b^*_2 \in [0, \infty) \) such that \( \tilde{\xi}^*_k(t) = \int_0^t q^*_k(\tilde{W}(s)) ds, k \in \{1, 2\} \), for \( q^*: \mathfrak{H}_+ \to \mathfrak{H}_+^2 \) that satisfies

\[
q^*(w) = \begin{cases} 
(0, \frac{\mu_3}{\mu_2} w) & \text{if } w \leq b^*, \\
(w, 0) & \text{if } w > b^*, 
\end{cases} \quad \text{when } (h_2 + c_2 \delta_2) \mu_3 < (h_1 + c_1 \delta_1) \mu_2, \tag{34}
\]

and

\[
q^*(w) = \begin{cases} 
(w, 0) & \text{if } w \leq b^*_2, \\
(0, \frac{\mu_3}{\mu_2} w) & \text{if } w > b^*_2, 
\end{cases} \quad \text{when } (h_2 + c_2 \delta_2) \mu_3 > (h_1 + c_1 \delta_1) \mu_2. \tag{35}
\]

Note that \( q^* \) depends only on the value of the workload \( w \geq 0 \), and not on the time \( t \geq 0 \). Then, when \( (h_2 + c_2 \delta_2) \mu_3 < (h_1 + c_1 \delta_1) \mu_2, \tilde{W}^* \) solves the stochastic equation

\[
\tilde{W}^*(t) = \xi(t) - \int_0^t \delta_1 \tilde{W}^*(s) 1\{\tilde{W}^*(s) > b^*\} + \delta_2 \tilde{W}^*(s) 1\{\tilde{W}^*(s) \leq b^*\} ds + \tilde{L}^*(t) \geq 0,
\]

and when \( (h_2 + c_2 \delta_2) \mu_3 > (h_1 + c_1 \delta_1) \mu_2, \tilde{W}^* \) solves the stochastic equation

\[
\tilde{W}^*(t) = \xi(t) - \int_0^t \delta_1 \tilde{W}^*(s) 1\{\tilde{W}^*(s) \geq b^*\} + \delta_2 \tilde{W}^*(s) 1\{\tilde{W}^*(s) > b^*\} ds + \tilde{L}^*(t) \geq 0.
\]

Furthermore, the parameter \( b^* \) is finite when also \( (\delta_1 + \gamma)/(\delta_2 + \gamma) > \mu_2(h_1 + c_1 \delta_1)/(\mu_3(h_2 + c_2 \delta_2)) \), and is infinite otherwise. Similarly, the parameter \( b^*_2 \) is finite when also \( (\delta_1 + \gamma)/(\delta_2 + \gamma) < \mu_2(h_1 + c_1 \delta_1)/(\mu_3(h_2 + c_2 \delta_2)) \), and is infinite otherwise. To relate back to the outline of our main results given in §1, recall that the condition \( (h_2 + c_2 \delta_2) \mu_3 < (h_1 + c_1 \delta_1) \mu_2 \) is condition (1), and the condition \( (\delta_1 + \gamma)/(\delta_2 + \gamma) > \mu_2(h_1 + c_1 \delta_1)/(\mu_3(h_2 + c_2 \delta_2)) \) is condition (3).

In the case that \( (h_2 + c_2 \delta_2) \mu_3 = (h_1 + c_1 \delta_1) \mu_2 \), we show that

\[
q^*(w) = \begin{cases} 
(0, \frac{\mu_3}{\mu_2} w) & \text{if } \delta_1 < \delta_2, \\
(w, 0) & \text{if } \delta_1 > \delta_2.
\end{cases} \tag{36}
\]

If also \( \delta_1 = \delta_2 \), \( q^*(w) = (q_1, q_2) \) is optimal for any \( q \in \mathcal{A} \) (that is, for any \( q \) having \( q_1 + (\mu_2/\mu_3)q_2 = w \) for all \( w \geq 0 \)).

The first step in showing that \((\tilde{\xi}^*, \tilde{L}^*)\) is an optimal control for the WCP is to show that \( q^* \) in (34)–(35) solves the Bellman equation, and we do this in §5.1. Then, in §5.2, we show that a solution to the Bellman equation, together with the control \( \tilde{L}^* \) in (33) is an optimal control for the WCP.
5.1. The Bellman equation. The Bellman equation is

$$\min_{q \in \mathbb{A}(w)} \left\{ \frac{\sigma^2}{2} V''(w) - f(q)V'(w) - \gamma V(w) + g(q) \right\} = 0 \quad \text{for all } w \geq 0 \tag{37}$$

subject to $V$ is twice-continuously differentiable; $V'$ is bounded, nondecreasing, and has $V'(0) = 0$.

Substituting for $f$ and $g$, and performing some simple algebraic manipulations, we find that

$$\min_{q \in \mathbb{A}(w)} \left\{ \frac{\sigma^2}{2} V''(w) - f(q)V'(w) - \gamma V(w) + g(q) \right\} = \frac{\sigma^2}{2} V''(w) + (h_1 + c_1 \delta_1 - \delta_1 V'(w))w - \gamma V(w) + U_w(V'(w)),$$

for

$$U_w(x) := \min_{q_1 \in [0, \mu_2/\mu_3] \cup \mathbb{R}_+} \left( (h_2 + c_2 \delta_2) - \frac{\mu_2}{\mu_3} (h_1 + c_1 \delta_1) + \frac{\mu_3}{\mu_2} (\delta_1 - \delta_2) x \right) q_2.$$

Hence

$$\min_{q \in \mathbb{A}(w)} \left\{ \frac{\sigma^2}{2} V''(w) - f(q)V'(w) - \gamma V(w) + g(q) \right\} = \frac{\sigma^2}{2} V''(w) + (h_1 + c_1 \delta_1 - \delta_1 V'(w))w - \gamma V(w) + U_w(V'(w)) = 0. \tag{38}$$

We conclude that if there exists a solution $q^*$ to the Bellman equation (37), it will satisfy

$$q^*(w) = \begin{cases} (w, 0) & \text{if } (h_2 + c_2 \delta_2) - \frac{\mu_2}{\mu_3} (h_1 + c_1 \delta_1) + \frac{\mu_3}{\mu_2} (\delta_1 - \delta_2) V'(w) > 0, \\ (w, 0) & \text{if } (h_2 + c_2 \delta_2) - \frac{\mu_2}{\mu_3} (h_1 + c_1 \delta_1) + \frac{\mu_3}{\mu_2} (\delta_1 - \delta_2) V'(w) \leq 0. \end{cases}$$

The following two lemmas establish that $q^*$ given in (34) solves the Bellman equation when $(h_2 + c_2 \delta_2) \mu_3 < (h_1 + c_1 \delta_1) \mu_2$.

**Lemma 5.1.** Assume that

$$(h_2 + c_2 \delta_2) \mu_3 > (h_1 + c_1 \delta_1) \mu_2 \quad \text{and} \quad \frac{\delta_1 + \gamma}{\delta_2 + \gamma} > \frac{\mu_2 (h_1 + c_1 \delta_1)}{\mu_3 (h_2 + c_2 \delta_2)}.$$ 

Then, there exists a $b^* < \infty$ such that $q^*$ defined as

$$q^*(w) = \begin{cases} (0, \frac{\mu_3}{\mu_2} w) & \text{if } w \leq b^*, \\ (w, 0) & \text{if } w > b^*, \end{cases} \tag{39}$$

solves (38). In particular $(V, b^*)$ that satisfies

$$V(w) = \begin{cases} V_2(w) & \text{when } w \in [0, b^*], \\ V_1(w) & \text{when } w > b^*, \end{cases}$$

for $V_2$ and $V_1$ that solve

$$\frac{\sigma^2}{2} V_2''(w) - \delta_2 w V_2'(w) - \gamma V_2(w) + \frac{\mu_3}{\mu_2} (h_2 + c_2 \delta_2) w = 0 \quad \text{for all } w \in [0, b^*],$$

$$\frac{\sigma^2}{2} V_1''(w) - \delta_1 w V_1'(w) - \gamma V_1(w) + (h_1 + c_1 \delta_1) w = 0 \quad \text{for all } w \geq b^*,$$

subject to $V_2'$ is increasing on $[0, b^*]$; $V_2'$ is increasing on $(b^*, \infty)$ and bounded; $V_2(0) = 0$;

$$V_1(b^*) = V_2(b^*) = \frac{(h_1 + c_1 \delta_1) - (\mu_3/\mu_2)(h_2 + c_2 \delta_2)}{(\delta_1 - \delta_2)};$$

$$V_1(b^*) = V_2(b^*); \quad \text{and} \quad V_1''(b^*) = V_2''(b^*).$$

satisfies the Bellman equation (37).
Note that it follows from the assumptions in Lemma 5.1 that \((\delta_1 + \gamma)/(\delta_2 + \gamma) > 1\), which implies that \(\delta_1 > \delta_2\), so that the denominator in the expression for \(V'(b^*)\) and \(V_2'(b^*)\) in (40) is not zero.

**Lemma 5.2.** Assume that

\[
(h_1 + c_1 \delta_1) \mu_2 \geq (h_2 + c_2 \delta_2) \mu_3 \quad \text{and} \quad \frac{\delta_1 + \gamma}{\delta_2 + \gamma} \leq \frac{\mu_3(h_1 + c_1 \delta_1)}{\mu_1(h_2 + c_2 \delta_2)}.
\]

Then, \(q^*\) in (39) solves (38) when \(b^* = \infty\), so that

\[
q^*(w) = \left(0, \frac{\mu_3}{\mu_2} w\right) \quad \text{for all } w \geq 0.
\]

In particular, \(V\) that solves

\[
\frac{\partial^2}{2} V''(w) - \delta_2 w V'(w) - \gamma V(w) + \frac{\mu_3}{\mu_2} (h_2 + c_2 \delta_2) w = 0 \quad \text{for all } w \geq 0
\]

subject to \(V\) is twice continuously differentiable; \(V\) is bounded, has \(V'(0) = 0\); and \((\delta_1 - \delta_2) V'(w) \leq (h_1 + c_1 \delta_1) - \frac{\mu_3}{\mu_2} (h_2 + c_2 \delta_2) \quad \text{for all } w \geq 0.

satisfies the Bellman equation (37).

The function \(V_2\) in (40) is obtained by giving priority to class 1. Then, all jobs in the system are class 2 jobs and the linear portion of the drift is determined by the class 2 abandonment rate \(\delta_2\). Similarly, the function \(V_1\) in (40) is obtained by giving priority to class 2. Then, all jobs in the system are class 1 jobs and the linear portion of the drift is determined by the class 1 abandonment rate \(\delta_1\). These two functions must be pasted together smoothly in order to evidence a solution to the Bellman equation (37). The function \(V\) in (42) is obtained by letting \(b^* = \infty\), so that class 1 is always given priority. This intuition is consistent with the explanation provided following (1) in §1, which notes that it is only when there is an alignment between the cost and the drift that we can expect that always giving priority to class 1 is an optimal policy.

In the other half of the parameter space, when

\[
(h_2 + c_2 \delta_2) \mu_3 > (h_1 + c_1 \delta_1) \mu_2,
\]

the \(q^*\) given in (35) solves the Bellman equation. Furthermore, the \(b^*_2\) that appears in (35) is finite when \((\delta_1 + \gamma)/(\delta_2 + \gamma) < \mu_2(h_1 + c_1 \delta_1)/(\mu_3(h_2 + c_2 \delta_2))\), and is infinite otherwise. We omit the proofs for this case because the arguments are essentially identical to the proofs of Lemmas 5.1 and 5.2.

### 5.2. The workload control problem solution.

Lemmas 5.1 and 5.2 provide a candidate WCP solution. Specifically, consider \(\tilde{L}^*\) that satisfies (33) and \(q^*\) that satisfies (37). The following theorem shows that \((\tilde{\mathcal{E}}^*, \tilde{L}^*)\), where \(\tilde{\mathcal{E}}^*(t) := \int_0^t q_k(\tilde{W}(s)) \, ds\) for every \(t \geq 0, k \in \{1, 2\}\), is an optimal control for the WCP, and characterizes the associated expected infinite horizon discounted cost in terms of \(V\) that solves (37).

**Theorem 5.1.** Let \(E_x[\cdot] := E[\cdot | \tilde{W}(0) = x]\).

(i) The control \((\tilde{\mathcal{E}}^*, \tilde{L}^*)\) has expected infinite horizon discounted cost given by \(V\) that satisfies (37); that is,

\[
V(x) = E_x \left[ \int_0^\infty e^{-\gamma t} g(q^*(\tilde{W}^*(t))) \, dt \right] \quad \text{for any } x \geq 0.
\]

(ii) Under any admissible control \((\mathcal{E}, \bar{L})\) for the WCP in Definition 4.2,

\[
E_x \left[ \int_0^\infty e^{-\gamma t} g(q(t, \tilde{W}(t))) \, dt \right] \geq E_x \left[ \int_0^\infty e^{-\gamma t} g(q^*(\tilde{W}^*(t))) \, dt \right] \quad \text{for any } x \geq 0.
\]

Note that it is only in the statement and proof of Theorem 5.1 that we use the notation \(E_x[\cdot]\). This notation is not needed elsewhere in the paper because we have assumed the system starts empty.
We will use the following lemma in the proof of Theorem 5.1.

**Lemma 5.3.** Suppose $X^R$ is a reflected Brownian motion with infinitesimal drift 0, infinitesimal variance $\sigma^2$ defined in (28), and initial position $X^R(0) = W^*(0)$. Then,

$$\tilde{W}*(t) \leq_{st} X^R(t) \quad \text{for every } t \geq 0,$$

where $\leq_{st}$ represents stochastically less than.

**Proof of Theorem 5.1.** The following is helpful in proving both parts (i) and (ii). Let $\Delta \tilde{L}(t) := \tilde{L}(t) - \tilde{L}(t^-)$ and $\Delta V(\tilde{W}(t)) := V(\tilde{W}(t)) - V(\tilde{W}(t^-))$. Also let $\Delta \tilde{L}(t) := \tilde{L}(t) - \sum_{0 \leq s \leq t} \Delta \tilde{L}(s)$ denote the continuous part of $\tilde{L}$. It follows from the generalized Itô formula (see, for example, Harrison [15, §4.7]) that for any $t > 0$

$$V(\tilde{W}(t)) = V(\tilde{W}(0)) + \int_0^t \left( \frac{\sigma^2}{2} V''(\tilde{W}(s)) - f(q(s, \tilde{W}(s))) V'(\tilde{W}(s)) \right) ds + \int_0^t \sigma V'(\tilde{W}(s)) d\tilde{L}(s) + \sum_{0 \leq s \leq t} \Delta V(\tilde{W}(s)).$$

Also, integration by parts (see, for example, Harrison [15, §4.8]) shows that

$$e^{-\gamma t} V(\tilde{W}(t)) = V(\tilde{W}(0)) + \int_0^t e^{-\gamma s} \left( \frac{\sigma^2}{2} V''(\tilde{W}(s)) - f(q(s, \tilde{W}(s))) V'(\tilde{W}(s)) - \gamma V(\tilde{W}(s)) \right) ds + \int_0^t e^{-\gamma s} \sigma V'(\tilde{W}(s)) d\tilde{L}(s) + \sum_{0 \leq s \leq t} e^{-\gamma s} \Delta V(\tilde{W}(s)).$$

Since $V'$ is bounded, the stochastic integral is a martingale, and so taking expectations shows that

$$E_x[e^{-\gamma V(\tilde{W}(t))}] = E_x\left[ e^{-\gamma \int_0^t e^{-\gamma s} \left( \frac{\sigma^2}{2} V''(\tilde{W}(s)) - f(q(s, \tilde{W}(s))) V'(\tilde{W}(s)) - \gamma V(\tilde{W}(s)) \right) ds + \int_0^t e^{-\gamma s} \sigma V'(\tilde{W}(s)) d\tilde{L}(s) + \sum_{0 \leq s \leq t} e^{-\gamma s} \Delta V(\tilde{W}(s)) \right] = e^{-\gamma t} V(x).$$

**5.2.1. The argument to establish (i).** First, note that when $\tilde{L} = \tilde{L}^*, \tilde{W} = \tilde{W}^*$, and $q = q^*$, it follows from (37) and (33) that (43) becomes

$$E_x[e^{-\gamma V(\tilde{W}^*(t))}] + E_x\left[ e^{-\gamma \int_0^t e^{-\gamma s} g(q(\tilde{W}^*(s))) ds} \right] = V(x).$$

Provided we can show that

$$E_x[e^{-\gamma V(\tilde{W}^*(t))}] \to 0 \quad \text{as } t \to \infty,$$

(45)

letting $t \to \infty$ on both sides of the Equation (44) establishes part (i).

Since $V'$ is bounded, there exists $m$ such that

$$|V(x)| \leq mx + V(0) \quad \text{for all } x \geq 0.$$

Hence

$$|E_x[e^{-\gamma V(\tilde{W}^*(t))}]| \leq E_x[|e^{-\gamma t}|] \leq E_x[|e^{-\gamma t}|] \leq E_x[|e^{-\gamma t}(m\tilde{W}^*(t) + V(0))|].$$

It then follows from Lemma 5.3 that

$$|E_x[e^{-\gamma V(\tilde{W}^*(t))}]| \leq E_x[e^{-\gamma t}(mX^R(t) + V(0))].$$

Since $X^R$ is a reflected Brownian motion with infinitesimal drift 0 and infinitesimal variance $\sigma^2$, $X^R$ has the same distribution as $\tilde{W}$, where $\tilde{W}$ is a standard Brownian motion. Hence

$$E[e^{-\gamma X^R(t)}] = E[ e^{-\gamma t} |\tilde{W}(t)| ]$$

$$= \int_{-\infty}^{\infty} e^{-\gamma t} |x| \frac{1}{\sqrt{2\pi}} e^{-x^2/(2t)} dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\gamma x^2/(2t)} dx$$

$$= \frac{\sqrt{2}}{\pi} t e^{-\gamma t} \to 0, \quad \text{as } t \to \infty,$$

which implies that (45) is valid.
5.2.2. The argument to establish (ii). Suppose an arbitrary admissible control \((\tilde{c}, \tilde{L})\), where \(\tilde{c}_k(t) = \int_0^t q_k(\bar{W}(s)) \, ds\) for every \(t \geq 0, k \in \{1, 2\}\), is used until time \(t > 0\). Then, for all remaining time, the control \((\tilde{c}^*, \tilde{L}^*)\), where \(\tilde{c}_k^* = \int_0^t q_k^*(\bar{W}(s)) \, ds\) for every \(t \geq 0, k \in \{1, 2\}\), is used. This hybrid control has cost
\[
h(x, t) := \mathbb{E}_x \left[ \int_0^t e^{-\gamma s} g(q(s, \bar{W}(s))) \, ds + e^{-\gamma t} V(\bar{W}(t)) \right].
\]
To establish (ii), it is sufficient to show that
\[
h(x, t) \geq V(x) \quad \text{for all } x \geq 0 \text{ and } t > 0.
\]
To show (46), first observe that it follows from (43) that
\[
\mathbb{E}_x \left[ e^{-\gamma s} V(\bar{W}(t)) \right] = \mathbb{E}_x \left[ \int_0^t e^{-\gamma s} g(q(s, \bar{W}(s))) \, ds \right] + V(x) - \mathbb{E}_x \left[ \int_0^t e^{-\gamma s} V'(\bar{W}(s)) \, ds \right] - \mathbb{E}_x \left[ \int_0^t e^{-\gamma s} \Delta V(\bar{W}(s)) \, ds \right].
\]
Next, since \(\bar{W}(s^-) = \bar{W}(s) - \Delta \bar{L}(s)\) and \(V'(x) \geq 0\) for all \(x \geq 0\),
\[
\Delta V(\bar{W}(s)) = V(\bar{W}(s)) - V(\bar{W}(s) - \Delta \bar{L}(s)) = \int_{\bar{W}(s) - \Delta \bar{L}(s)}^{\bar{W}(s)} V'(x) \, dx \geq 0,
\]
so that
\[
\mathbb{E}_x \left[ \int_0^t e^{-\gamma s} \Delta V(\bar{W}(s)) \, ds \right] \geq 0.
\]
We conclude that (46) holds. \(\square\)

6. The proposed policy. We state our proposed policy in Definitions 6.1 and 6.2. There are two definitions because the proposed policy in the case that \((h_1 + c_1 \delta_1) \mu_1 < (h_1 + c_1 \delta_2) \mu_2\), so that class 1 is the more expensive class, differs somewhat from the case that \((h_1 + c_2 \delta_3) \mu_3 > (h_1 + c_1 \delta_1) \mu_2\), so that class 2 is the more expensive class. The fact that there is a difference is not surprising given that the model is asymmetric in the sense that server 2 can serve both classes 1 and 2 but server 1 can only serve class 1. The \(b^*\) and \(b^*_2\) that appear in Definitions 6.1 and 6.2 are those that appear in (34) and (35) in §5, and are specified by the solution to the Bellman equation (37). In the case that \((h_1 + c_2 \delta_3) \mu_3 > (h_1 + c_1 \delta_1) \mu_2\), the proposed policy is motivated by (36), and is the policy in Definition 6.1 having \(b^* = \infty\) if \(\delta_1 < \delta_2\), and is the policy in Definition 6.2 having \(b^*_2 = \infty\) if \(\delta_1 > \delta_2\). If \(\delta_1 = \delta_2\) also, there are infinitely many asymptotically optimal policies.
DEFINITION 6.1 (THE \((M', L')\) THRESHOLD POLICY WHEN \((h_2 + c_3 \delta_3) \mu_3 < (h_1 + c_4 \delta_4) \mu_2\)). For each \(r \geq 1\), let \(M' = [rb^*]\), and let \(L' = [c \log r]\) for \(c\) a positive, finite constant that is large enough (see Remark 6.3). The allocation process in the \(r\)th system, \(T'^*\), satisfies

\[
\int_0^\infty Q'_1(t) d(t - T'^*_1(t)) = 0,
\]

\[
\int_0^\infty Q'_2(t) 1\{W'(t) > M'\} d(t - T'^*_3(t)) = 0
\]

\[
\int_0^\infty (Q'_1(t) + Q'_2(t)) 1\{W'(t) > M'\} d(t - T'^*_3(t) - T'^*_3(t)) = 0
\]

\[
\int_0^\infty 1\{W'(t) < M', Q'_1(t) > L'\} d(t - T'^*_3(t)) = 0
\]

\[
\int_0^\infty Q'_2(t) 1\{W'(t) < M', Q'_1(t) \leq L'\} d(t - T'^*_3(t)) = 0
\]

\[
\int_0^\infty 1\{W'(t) < M', Q'_1(t) \leq L'\} dT'^*_3(t) = 0.
\]

In words, the \((M', L')\) threshold policy in Definition 6.1 is such that the \(r\)th system operates as follows. Server 1 works whenever possible, or, equivalently, server 1 is never idle when there are class 1 jobs waiting or at server 1. When the workload process \(W'\) exceeds the threshold level \(M'\), server 2 gives preemptive resume priority to class 2 jobs over class 1 jobs. Otherwise, when \(W'\) is at or below the level \(M'\), and \(Q'_1\) exceeds the level \(L'\), server 2 serves class 1 jobs. Finally, when \(W'\) is at or below the level \(M'\), and \(Q'_1\) is at or below the level \(L'\), server 2 does not serve class 1 jobs, and either serves a class 2 job if one is available or idles if there are no class 2 jobs.

DEFINITION 6.2 (THE \((M', L')\) THRESHOLD POLICY WHEN \((h_2 + c_3 \delta_3) \mu_3 > (h_1 + c_4 \delta_4) \mu_2\)). For each \(r \geq 1\), let \(M' = [rb^*]\), and let \(L' = [c \log r]\) for \(c\) a positive, finite constant that is large enough (see Remark 6.3). The allocation process in the \(r\)th system, \(T'^*\), satisfies

\[
\int_0^\infty Q'_1(t) d(t - T'^*_1(t)) = 0,
\]

\[
\int_0^\infty 1\{W'(t) > M', Q'_1(t) > L'\} d(t - T'^*_2(t)) = 0
\]

\[
\int_0^\infty Q'_2(t) 1\{W'(t) > M', Q'_1(t) \leq L'\} d(t - T'^*_2(t)) = 0
\]

\[
\int_0^\infty 1\{W'(t) > M', Q'_1(t) \leq L'\} dT'^*_2(t) = 0
\]

\[
\int_0^\infty Q'_2(t) 1\{W'(t) \leq M'\} d(t - T'^*_3(t)) = 0
\]

\[
\int_0^\infty 1\{W'(t) < M', Q'_1(t) > L'\} dT'^*_2(t) = 0
\]

\[
\int_0^\infty (Q'_1(t) + Q'_2(t)) 1\{W'(t) \leq M'\} d(t - T'^*_3(t) - T'^*_3(t)) = 0.
\]

In words, the \((M', L')\) threshold policy in Definition 6.2 is such that the \(r\)th system operates as follows. Server 1 works whenever possible, or, equivalently, server 1 is never idle when there are class 1 jobs waiting or at server 1. When the workload process \(W'\) is at or below the threshold level \(M'\), server 2 gives preemptive resume priority to class 2 jobs over class 1 jobs. Otherwise, when \(W'\) is at or below the level \(M'\), and \(Q'_1\) exceeds the level \(L'\), server 2 serves class 1 jobs. Finally, when \(W'\) exceeds the level \(M'\), and \(Q'_1\) is at or below the level \(L'\), server 2 does not serve class 1 jobs, and instead serves class 2 jobs.

Remark 6.1. The \((M', L')\) threshold policies in Definitions 6.1 and 6.2 ensure that, in the heavy-traffic limit, all of the jobs in the system are from the less expensive class when the workload process is below the
threshold $M'$ and from the more expensive class when the workload process is above the threshold $M'$. The purpose of placing the threshold $L'$ on queue 1 is to ensure that server 2 does not help server 1 so much that server 1 ends up with frequent idling when there are waiting class 2 jobs in the system that server 1 cannot serve.

**Remark 6.2.** The threshold $M'$ is infinite when the solution to the WCP has either $b^* = \infty$ or $b^*_2 = \infty$. In the case that $b^* = \infty$ in Definition 6.1, the proposed policy is exactly the policy proposed in Definition 5.1 in Bell and Williams [8], that is shown to be asymptotically optimal for the $N$-system with no reneging. In the case that $b^*_2 = \infty$ in Definition 6.2, the proposed policy is the static priority policy suggested by the $c\mu$-rule, in which server 2 gives priority to class 2 jobs, and that policy is known to be optimal for the $N$-system without reneging.

We prove weak convergence in §7 when the system operates under the proposed policy in Definition 6.1, and $b^*$ is finite. Then, we prove asymptotic optimality in §8 under the assumption that

$$
(h_2 + c_2 \delta_2) \mu_3 < (h_1 + c_1 \delta_1) \mu_2 \quad \text{and} \quad \frac{\delta_1 + \gamma}{\delta_2 + \gamma} > \frac{\mu_3 (h_1 + c_1 \delta_1)}{\mu_3 (h_2 + c_2 \delta_2)}.
$$

(47)

For the remaining part of the parameter space, when (47) does not hold, we outline the proofs of weak convergence and asymptotic optimality in §9. The most technically challenging part of the parameter space is when either $b^*$ or $b^*_2$ is finite; however, once we have handled the case that $b^*$ is finite, the case that $b^*_2$ is finite follows similarly.

**Remark 6.3.** There exists a constant $c_0$ such that for any $c > c_0$ in Definition 6.1, our weak convergence and asymptotic optimality proofs in §§7 and 8 are valid. The value of $c_0$ is described in the proof of Lemma 7.2 in the appendix, when establishing (B16) and (B17), and also in the last paragraph of the proof of Theorem 8.1, and also comes from the arguments in Bell and Williams [8] that we refer to in the appendix, §A; however, its description is not in terms of a concise formula, and so we do not provide it here.

**Remark 6.4.** Our weak convergence and asymptotic optimality proof methods (used to prove Theorems 7.1 and 8.1 in §§7 and 8) are valid for any sequence of thresholds $L'$ that satisfy $L' \geq c \log r$ for $c > c_0$ and $L' = o(r)$ as $r \to \infty$. Furthermore, these proofs show that as the order of the threshold size increases from log $r$ to a larger size that is still $o(r)$, the number of moments that are required to be finite decreases. In particular, looking ahead in the paper, when $L'$ is of order $\log r$, the assumption of exponential moments for the interarrival and service times in Assumption 3.3 is necessary to prove Lemma 7.2, which establishes the weak convergence of the diffusion-scaled server 1 idleness. Close inspection of this proof shows that when the threshold is of a larger order than log $r$ (for example, when $L'$ is of order $\sqrt{r}$), a less restrictive moment condition suffices.

**Remark 6.5.** When implementing the proposed policy, it may be advised to not immediately switch from having server 2 give priority to class 1 jobs to having server 2 give priority to class 2 jobs when the workload goes above the threshold level $M'$, and vice versa when the workload goes below $M'$. This is to prevent “chattering,” whereby the class that server 2 gives priority to may switch many times in a short amount of time. This chattering behavior is a fundamental reason why the weak convergence result we prove in Theorem 7.1 in the next section is not

$$
(\hat{Q}_i^r, \hat{Q}_2^r) \Rightarrow q^*(\bar{W}^r) \quad \text{as} \quad r \to \infty.
$$

In particular, we do not expect the weak convergence to hold in $D^2$.

### 7. Weak convergence

We devote §7 to proving the following weak convergence theorem.

**Theorem 7.1.** Consider the sequence of $N$-systems indexed by $r$, where the $r$th system operates under the threshold policy $T^{r,*}$ described in Definition 6.1. Then, for any $t > 0$,

$$
\int_0^t e^{-\gamma s} \hat{Q}_i^r(s) ds \Rightarrow \int_0^t e^{-\gamma s} q_i^*(\bar{W}^r(s)) ds
$$

and

$$
\int_0^t e^{-\gamma s} \hat{Q}_2^r(s) ds \Rightarrow \int_0^t e^{-\gamma s} \frac{B_2}{\mu_2} q_2^*(\bar{W}^r(s)) ds,
$$

as $r \to \infty$. Furthermore, the weak convergence holds jointly.
The proof of Theorem 7.1 is done in several steps. First, in §7.1, we prove the following state-space collapse result.

**Proposition 7.1.** Consider the sequence of $N$-systems indexed by $r$, where the $r$th system operates under the threshold policy $T^r_{\cdot \cdot \cdot}$ in Definition 6.1. Then, for any $\eta > 0$

$$\hat{Q}_r^* \{ \hat{W}^r < b^* - \eta \} \to 0 \quad \text{and} \quad \hat{Q}_r^* \{ \hat{W}^r > b^* + \eta \} \to 0 \quad \text{as} \quad r \to \infty. $$

Next, in §7.2, we prove a weak convergence result on the workload process $\hat{W}^r$. To state this result, define

$$\hat{X}^r(t) := \hat{X}_r^*(t) + \frac{\mu_r}{\mu_3} \hat{X}^r_2(t),$$

$$\hat{V}^r_1(t) := \int_0^t (r^2 \delta_1^2) \hat{Q}_1^*(s) \{ \hat{W}^r(s) \leq b^* \} \, ds,$$

$$\hat{V}^r_2(t) := \int_0^t \frac{\mu_r}{\mu_3} (r^2 \delta_2^2) \hat{Q}_2^*(s) \{ \hat{W}^r(s) > b^* \} \, ds,$$

$$\hat{V}^r(t) := \int_0^t \delta_1 \hat{W}^r(s) \{ \hat{W}^r(s) > b^* \} \, ds - \int_0^t (r^2 \delta_1^2) \hat{Q}_1^*(s) \{ \hat{W}^r(s) > b^* \} \, ds$$

$$+ \int_0^t \delta_2 \hat{W}^r(s) \{ \hat{W}^r(s) \leq b^* \} \, ds - \int_0^t \frac{\mu_r}{\mu_3} (r^2 \delta_2^2) \hat{Q}_2^*(s) \{ \hat{W}^r(s) \leq b^* \} \, ds$$

$$+ \frac{1}{r} \int_0^t (r^2 \delta_1^2) \{ \hat{T}_1^r(r^2 s) > 0 \} + \{ \hat{T}_2^r(r^2 s) > 0 \} \} + \frac{\mu_r}{\mu_3} \{ r^2 \delta_2^2 \} \hat{T}_2^r(r^2 s) > 0 \} \, ds.$$

so that

$$\hat{V}^r(t) = \hat{X}^r(t) - \hat{V}^r_1(t) - \hat{V}^r_2(t) + \hat{V}^r(t) + \mu_1 \hat{L}^r_1(t) + \mu_2 \hat{L}^r_2(t).$$

The weak convergence result we prove is the following.

**Proposition 7.2.** Consider the sequence of $N$-systems indexed by $r$, where the $r$th system operates under the threshold policy $T^r_{\cdot \cdot \cdot}$ in Definition 6.1. Then,

$$(\hat{W}^r, \hat{X}^r, \hat{V}^r_1, \hat{V}^r_2, \mu_1 \hat{L}^r_1, \mu_2 \hat{L}^r_2) \Rightarrow (\hat{W}^*, \hat{X}^*, \hat{V}^*, \mu_1 \hat{L}^*, \mu_2 \hat{L}^*) \quad \text{as} \quad r \to \infty,$$

where

$$\hat{V}(t) = \int_0^t \delta_1 \hat{W}^r(s) \{ \hat{W}^r(s) > b^* \} + \delta_2 \hat{W}^r(s) \{ \hat{W}^r(s) \leq b^* \} \, ds.$$

Finally, in §7.3, we prove Theorem 7.1.

**Remark 7.1.** The proofs of Theorem 7.1, Proposition 7.1, Proposition 7.2, and Lemmas 7.1–7.4 (that are stated in this section) hold for any finite $b > 0$, when the system operates under the $(M', L')$ threshold policy in Definition 6.1 with $b$ replacing $b^*$. For convenience in notation, in the remainder of this section, we drop the superscript $\cdot \cdot \cdot$ on $b$.

**7.1. Proof of Proposition 7.1.** We establish that

$$\hat{Q}_r^* \{ \hat{W}^r > b + \eta \} \to 0 \quad \text{as} \quad r \to \infty. \quad (48)$$

The argument to establish $\hat{Q}_r^* \{ \hat{W}^r < b - \eta \} \to 0$ as $r \to \infty$ has many similarities, and so we have put the details in the appendix.

We begin with some useful definitions:

$$\kappa_0' := 0,$$

$$\alpha_n' := \inf \{ s \geq \kappa_{n-1}' : W^r(s) \geq rb + 1 \},$$

$$\beta_n'^{k} := \inf \{ s \geq \alpha_n' : Q_k^r(s) = 0 \}, \quad k \in \{ 1, 2 \},$$

$$\gamma_n'^r := \inf \{ s \geq \alpha_n' : W^r(s) \geq r(b + \eta) \},$$

$$\kappa_n' := \inf \{ s \geq \alpha_n' : W^r(s) \leq rb \},$$

$$\mu_1 \hat{L}^r_1(t) + \mu_2 \hat{L}^r_2(t).$$
Observe that "up" excursion from below 780  

Next, we define the "good set," in which the primitive processes stay close to their mean rates. For  \( \tilde{\epsilon} > 0 \) arbitrarily small, define

\[
s' := \frac{r(\mu_3/\mu_2)b + (2(\mu_3/\mu_2) + 1)}{\mu_3' - \lambda_3' - 2\tilde{\epsilon}},
\]

and assume  \( \tilde{\epsilon} \) is small enough so that  \( \mu_3' - \lambda_3' - 2\tilde{\epsilon} > 0 \) for all large enough  \( r \). Let

\[
\eta^{*,n} := \left\{
\begin{align*}
&\|A_k^{*,n}(\cdot) - \lambda_k'(\cdot)\|_{\nu} < \tilde{\epsilon}s', \quad k \in \{1, 2\}, \\
&\|S_j^{*,n}(T_j'(\cdot)) - \mu_j'T_j'(\cdot)\|_{\nu} < \tilde{\epsilon}s', \quad j \in \{1, 2, 3\}, \\
&\|R_k^{*,n}(\cdot)\|_{\nu} \leq \tilde{\epsilon}s'.
\end{align*}
\right.
\]

Let

\[
m' := \left(\lambda_1' + \frac{\mu_3'}{\mu_3'} + 2\tilde{\epsilon}\right)r^2t.
\]

The proof of Proposition 7.1 requires the following lemma.

**Lemma 7.1.** \( P((\cap_{n=1}^{m'} \eta^{*,n})^c) \to 0, \) as  \( r \to \infty \).

We are now ready to establish (48). Fix  \( t > 0 \). Let  \( \epsilon > 0 \) be arbitrarily small. It is sufficient to show

\[
P(Q'(s)1[W'(s) > r(b + \eta)] \geq r\epsilon \quad \text{for some } s \in [0, r^2t]) \to 0, \quad \text{as } r \to \infty.
\]

We now argue that the first term on the right-hand side in the above expression converges to zero. Since every "up" excursion from below  \( rb \) to above  \( rb + 1 \) requires at least one arrival

\[
P(\alpha_{\nu'} \leq r^2t) \leq P\left(A'_1(r^2t) + \frac{\mu_3'}{\mu_3'}A'_2(r^2t) \geq m'\right).
\]

Furthermore,

\[
P\left(A'_1(r^2t) + \frac{\mu_3'}{\mu_3'}A'_2(r^2t) \geq m'\right) \leq P\left(A'_1(r^2t) \geq (\lambda_1' + \tilde{\epsilon})r^2t\right) + P\left(A'_2(r^2t) \geq (\lambda_2' + \tilde{\epsilon})r^2t\right)
\]

\[
\leq P\left(\|A'_1(\cdot) - \lambda_1'(\cdot)\|_{r^2t} > \tilde{\epsilon}r^2t\right) + P\left(\|A'_2(\cdot) - \lambda_2'(\cdot)\|_{r^2t} > \tilde{\epsilon}r^2t\right),
\]

and it follows from Lemma 9 in Ata and Kumar [1] that there exists a finite constant  \( d > 0 \), that does not depend on  \( r \), such that

\[
P\left(\|A'_1(\cdot) - \lambda_1'(\cdot)\|_{r^2t} > \tilde{\epsilon}r^2t\right) \leq \frac{d}{(r^2t)^{1+r\delta}}, \tag{50}
\]

and it follows from Lemma 9 in Ata and Kumar [1] that there exists a finite constant  \( d > 0 \), that does not depend on  \( r \), such that

\[
P\left(\|A'_1(\cdot) - \lambda_1'(\cdot)\|_{r^2t} > \tilde{\epsilon}r^2t\right) \leq \frac{d}{(r^2t)^{1+r\delta}}. \tag{50}
\]

\[\]
for any $\epsilon_1 > 0$, because Assumption 3.3 implies that the interarrival times have all moments finite. (For the reader’s convenience, we restate Ata and Kumar [1, Lemma 9], adapted to our setting, in the appendix, in the proof of Lemma 7.1.) Hence

$$P(\alpha'_{n \nu} \leq r^2 t) \to 0, \quad as \ r \to \infty.$$  

Lemma 7.1 establishes that

$$P\left(\bigcap_{n=1}^{m'} \eta^{r, \nu}_n\right)^C \to 0, \quad as \ r \to \infty.$$  

Hence to show (49), it is sufficient to establish that

$$P\left(\alpha'_{n \nu} > r^2 t, \left(\bigcap_{n=1}^{m'} \eta^{r, \nu}_n\right), Q'(s) \geq r\epsilon \text{ for some } s \in \bigcup_{n=1}^{m'} [\gamma'_n, \kappa'_n] \cap [0, r^2 t]\right) \to 0, \quad as \ r \to \infty.$$  

For this, first note that

$$P\left(\alpha'_{n \nu} > r^2 t, \left(\bigcap_{n=1}^{m'} \eta^{r, \nu}_n\right), Q'(s) \geq r\epsilon \text{ for some } s \in \bigcup_{n=1}^{m'} [\gamma'_n, \kappa'_n] \cap [0, r^2 t]\right) \leq \sum_{n=1}^{m'} P\left(\alpha'_{n \nu} > r^2 t, \eta^{r, \nu}_n, Q'(s) \geq r\epsilon \text{ for some } s \in [\gamma'_n, \kappa'_n] \cap [0, r^2 t]\right).$$  

Assume we can show that

$$P\left(\alpha'_{n \nu} > r^2 t, \eta^{r, \nu}_n, Q'(s) \geq r\epsilon \text{ for some } s \in [\gamma'_n, \kappa'_n] \cap [0, r^2 t]\right) \leq P\left(\alpha'_{n \nu} > r^2 t, \eta^{r, \nu}_n, \beta'^{-2}_n \leq \beta^{-2}_n, \beta'^{-2}_n \leq \alpha'_n + s', \beta'^{-2}_n \leq \gamma'_n\right).$$  

Then, it follows that to complete the proof it is sufficient to show that

$$\sum_{i=1}^{m'} P\left(Q'(s) \geq r\epsilon \text{ for some } s \in [\beta^{-2}_n, r^2 t \wedge \kappa'_n]\right) \to 0, \quad as \ r \to \infty.$$  

We establish (51) below. We have placed the argument to establish (52) in the appendix, because it is detailed, even though it is intuitively clear. The intuition is that after time $\beta'^{-2}_n$ the number of jobs in buffer 2 follows the evolution dynamics of an underloaded G1/G1/1 queue.

**7.1.1. The argument that (51) holds.** Assume that $\tilde{\epsilon}$ has also been chosen small enough so that

$$\left(1 + \frac{\mu_3}{\mu_2}\right)4\tilde{\epsilon}s' < \left(\frac{\delta}{2}\right)r, \quad \tilde{\epsilon}\left(3 + 2\frac{\mu_3}{\mu_3}\right)s' < \frac{\delta}{2}, \quad \text{and} \quad \lambda'_1 - \mu'_1 - (\delta'_1 r^2)\tilde{\epsilon} > 0,$$

for large enough $r$. We first show that on $\eta^{r, \nu} \cap \{\alpha'_n < \infty\}$, $\beta'^{-2}_n \leq \alpha'_n + s'$. To see this, note that

$$Q'_2(\alpha'_n + s) = Q'_2(\alpha'_n) + A'^{-2}_n(s) - S'^{-2}_n(s) - R'^{-2}_n(s) \quad \text{for } \alpha'_n + s \leq \beta'^{-2}_n.$$  

Since $W'(\alpha'_n) \leq rb + 1 + (1 + \mu_2 / \mu_3)$, it follows that $Q'_2(\alpha'_n) \leq (\mu_1 / \mu_2)(rb + 2 + \mu_2 / \mu_3)$. If $\alpha'_n + s' > \beta'^{-2}_n$, then

$$Q'_2(\alpha'_n + s') \leq \frac{\mu_3}{\mu_2} \left(rb + 2 + \frac{\mu_2}{\mu_3}\right) - s' (\mu_3' - \lambda'_3 - 2\tilde{\epsilon}).$$  

This contradicts the definition of $\beta'^{-2}_n$, because the definition of $s'$ implies that

$$\frac{\mu_3}{\mu_2} \left(rb + 2 + \frac{\mu_2}{\mu_3}\right) - s' (\mu_3' - \lambda'_3 - 2\tilde{\epsilon}) = 0.$$  

We next show that on $\eta^{r, \nu} \cap \{\alpha'_n < \infty\}$, $\beta'^{-2}_n \leq \gamma'_n$. For this, we first show that if queue 1 is small, the workload cannot become large. Define

$$\psi'_n := \inf\{u \geq 0: Q'_2(\alpha'_n + u) \geq 3\tilde{\epsilon}s'\},$$
and note that if $\psi_n' > 0$, for $0 \leq s \leq \psi_n' \wedge \delta$, 
\[
W'(\alpha_n' + s) = Q'_1(\alpha_n' + s) + \frac{\mu_2}{\mu_3} Q'_2(\alpha_n' + s) \\
\leq 3\bar{\epsilon}s' + \frac{\mu_2}{\mu_3} \left( Q'_2(\alpha_n') + A^{\epsilon,n}_2(s) - S^{\epsilon,n}_3(s) - R^{\epsilon,n}_2(s) \right) \\
\leq 3\bar{\epsilon}s' + \frac{\mu_2}{\mu_3} \left( rb + 2 + \frac{\mu_3}{\mu_3} \right) + \lambda_5's - \mu_5s + 2\bar{\epsilon}s' \\
\leq \bar{\epsilon} \left( 3 + 2\frac{\mu_2}{\mu_3} s' \right) + rb + \frac{\mu_2}{\mu_3} \\
< r \left( b + \frac{\eta}{2} \right) + \frac{1}{r} \left( 2 + \frac{\mu_2}{\mu_3} \right).
\]

Also, once queue 1 becomes large, it stays large. In particular, for $s \leq \delta$, on $\eta^{\epsilon,n}$,
\[
Q'_1(\alpha_n' + s) = Q'_1(\alpha_n' + \psi_n') + (A^{\epsilon,n}_1(s) - A^{\epsilon,n}_1(\psi_n')) - (S^{\epsilon,n}_1(T'_1(\psi_n')) - S^{\epsilon,n}_1(T'_1(s))) - (R^{\epsilon,n}_1(s) - R^{\epsilon,n}_1(\psi_n')) \\
\geq 3\bar{\epsilon}s' + \lambda_1'(s - \psi_n') - \mu_1'(s - \psi_n') - 3\bar{\epsilon}s' \\
> 0.
\]

Finally, for $\alpha_n' + s \leq \beta_n^{-1} \wedge \beta_n^{-1} \leq s \leq \delta$, since $Q'_1(s) > 0$ and $Q'_2(s) > 0$, so that $T'_1(\psi_n' + s) = T'_1(\psi_n') + s$ and $T'_2(\psi_n' + s) = T'_2(\psi_n') + s$,
\[
W'(\alpha_n' + s) = W'(\psi_n') + A^{\epsilon,n}_1(s) - A^{\epsilon,n}_1(\psi_n') \\
- (S^{\epsilon,n}_1(T'_1(\psi_n' + s)) - S^{\epsilon,n}_1(T'_1(\psi_n'))) - (R^{\epsilon,n}_1(s) - R^{\epsilon,n}_1(\psi_n')) \\
+ \frac{\mu_2}{\mu_3} \left( \begin{array}{c}
A^{\epsilon,n}_2(s) - A^{\epsilon,n}_2(\psi_n') \\
- (S^{\epsilon,n}_2(T'_2(\psi_n' + s)) - S^{\epsilon,n}_2(T'_2(\psi_n'))) \\
- (R^{\epsilon,n}_2(s) - R^{\epsilon,n}_2(\psi_n'))
\end{array} \right) \\
\leq r \left( b + \frac{\eta}{2} \right) + \frac{1}{r} \left( 2 + \frac{\mu_2}{\mu_3} \right) + \lambda_5'(s - \psi_n') - \mu_5'(s - \psi_n') \\
+ \frac{\mu_2}{\mu_3} \left( \lambda_5'(s - \psi_n') - \mu_5'(s - \psi_n') \right).
\]

Assume $r$ is large enough so that $\lambda_k' \leq \lambda_k + \bar{\epsilon}$, $k \in \{1, 2\}$, and $\mu_j' \geq \mu_j - \bar{\epsilon}$, $j \in \{1, 2, 3\}$. Then, noting that $\lambda_1 - \mu_1 + (\mu_2/\mu_3)\lambda_2 - \mu_2 = 0$, it follows that
\[
W'(\alpha_n' + s) \leq r \left( b + \frac{\eta}{2} \right) + \frac{1}{r} \left( 2 + \frac{\mu_2}{\mu_3} \right)4\bar{\epsilon}s' < r(b + \eta)
\]
for large enough $r$. This implies that on $\eta^{\epsilon,n}$, $\beta_n^{-1} \wedge \beta_n^{-1} < \gamma_n'$. Finally, note that $\beta_n^{-1} < \kappa_n$ is not an additional constraint. 

\section{Proof of Proposition 7.2.} The proof of Proposition 7.2 requires the following three lemmas, whose proofs can all be found in the appendix.

\textbf{Lemma 7.2.} Consider the sequence of $N$-systems indexed by $r$, where the $r$th system operates under the threshold policy $T^r, \epsilon$ in Definition 6.1. Then,
\[
\hat{I}_i \to 0 \quad \text{as } r \to \infty.
\]

\textbf{Lemma 7.3.} Consider the sequence of $N$-systems indexed by $r$, where the $r$th system operates under the threshold policy $T^r, \epsilon$ in Definition 6.1. Then,
\[
\left( \tilde{Q}'_1, \tilde{Q}'_2, \tilde{T}'^r, \frac{1}{r} \hat{I}_1, \frac{1}{r} \hat{I}_2 \right) \to (0, 0, \tilde{T}^*, 0, 0) \quad \text{u.o.c., a.s., as } r \to \infty.
\]
Furthermore,
\[
\hat{\tilde{X}}^r \Rightarrow \tilde{\xi}, \quad \text{as } r \to \infty,
\]
where $\tilde{\xi}$ is a Brownian motion with zero drift and variance $\sigma^2$. 

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Figure 1.}
\end{figure}
LEMMA 7.4. Consider the sequence of N-systems indexed by \( r \), where the \( r \)th system operates under the threshold policy \( T'^r \) in Definition 6.1. Then,

\[
\left\{ (\hat{W}'', \hat{X}'', \hat{e}'', \hat{V}'', \hat{I}'', \hat{I}'') \right\}
\]

is \( \mathcal{C} \)-tight (that is, it is tight in \( \mathcal{D}^X \), and any weak limit point of the sequence, obtained as a weak limit along a subsequence, has continuous paths almost surely).

It is also useful to define the function

\[
m(w) := \begin{cases} \delta_2 & w \leq b, \\ \delta_1 & w > b, \end{cases}
\]

and note that it is equivalent to write \( \hat{V}' \) as

\[
\hat{V}'(t) = \int_0^t m(\hat{W}'(s)) \hat{W}'(s) \, ds.
\]

We are now ready to prove Proposition 7.2. It follows from Lemma 7.4 that

\[
\left\{ (\hat{W}'', \hat{X}'', \hat{e}'', \hat{V}'', \hat{I}'', \hat{I}'') \right\}
\]

is \( \mathcal{C} \)-tight. Lemma 7.2 establishes that \( \hat{I}' \implies 0 \) as \( r \to \infty \), and Lemma 7.3 establishes that \( \hat{X}' \implies \hat{\xi} \) as \( r \to \infty \), where \( \hat{\xi} \) is a Brownian motion with zero drift and variance \( \sigma^2 \). Also, note our proposed policy implies that \( \mu_2 \hat{I}' \) can only increase when \( \hat{W}' > L'/r \). More formally, note that the definition of the \( (M', L') \) threshold policy in Definition 6.1 implies that for any \( 0 \leq t_1 < t_2 \),

\[
\int_{t_1}^{t_2} \mathbf{1} \{ \hat{Q}'_1(t) > \frac{L'}{r} \text{ or } \hat{Q}'_2(t) > 0 \} \, d\hat{I}'(t) = 0.
\]

Since \( \hat{W}' = \hat{Q}'_1 + (\mu_2 / \mu_1) \hat{Q}'_2 \), for any \( t > 0 \),

\[
\left\{ \hat{Q}'_1(t) \leq \frac{L'}{r} \text{ and } \hat{Q}'_2(t) = 0 \right\} \subset \left\{ \hat{W}'(t) \leq \frac{L'}{r} \right\},
\]

from which it follows that

\[
\int_{t_1}^{t_2} \mathbf{1} \{ \hat{W}'(t) > \frac{L'}{r} \} \, d\hat{I}'(t) = 0. 
\text{(53)}
\]

Let \( (\hat{W}, \hat{\xi}, \hat{E}, \hat{V}_1, \hat{V}_2, \hat{V}, \hat{0}, \hat{L}) \) be any weak limit point of the above sequence. Then,

\[
\hat{W}(t) = \hat{\xi}(t) + \hat{E}(t) - \hat{V}_1(t) - \hat{V}_2(t) - \hat{V}(t) + \hat{L}(t).
\]

In this proof, we first show that

\[
\hat{E} = \hat{V}_1 = \hat{V}_2 = 0 \quad \text{and} \quad \hat{V}(t) = \int_0^t m(\hat{W}(s)) \hat{W}(s) \, ds.
\]

Then, since \( \hat{L}' \) has \( \hat{L}'(0) = 0 \) is nondecreasing, and an argument similar to the one employed in (39)–(46) of Williams [37] shows that \( \int_0^{\infty} \hat{W}(t) \, d\hat{L}(t) = 0 \), it follows that any weak limit point satisfies

\[
\hat{W}(t) = \hat{\xi}(t) - \int_0^t m(\hat{W}(s)) \hat{W}(s) \, ds + \hat{L}(t) \quad \text{for all } t \geq 0,
\]

\[
\hat{L}(0) = 0, \quad \hat{L} \text{ is nondecreasing, and } \int_0^{\infty} \hat{W}(t) \, d\hat{L}(t) = 0. 
\text{(54)}
\]

We then use a Girsanov argument to show that there is uniqueness in law for any solution of (54).
7.2.1. The argument that any weak limit point satisfies (54). Let \( \{\mathcal{F}_t\} \) be the filtration generated by \((\xi, \mathcal{E}, \mathcal{V}_1, \mathcal{V}_2, \mathcal{V})\). Note that \( (\tilde{W}, \mu_2, \mathcal{L}) \) is a solution of the Skorokhod problem for \( \xi + \mathcal{E} - \mathcal{V}_1 - \mathcal{V}_2 - \mathcal{V} \), so that
\[
(\tilde{W}, \mu_2, \mathcal{L}) = (\Phi, \Psi)(\xi + \mathcal{E} - \mathcal{V}_1 - \mathcal{V}_2 - \mathcal{V}).
\]
Hence \( \tilde{W}(t) \) and \( \mathcal{L}(t) \) are \( \{\mathcal{F}_t\} \) adapted. Also note that \( \xi \) is a martingale relative to \( \{\mathcal{F}_t\} \), since it is a Brownian motion.

Next, \( \mathcal{E} - \mathcal{V}_1 - \mathcal{V}_2 - \mathcal{V} + \mathcal{L}(t) \) consists of terms that are continuous and monotone, and so of finite variation. Hence \( \tilde{W} \) is a continuous \( \{\mathcal{F}_t\} \)-semimartingale with quadratic variation process \( [\tilde{W}]_t = [\xi]_t = \sigma_t, t \geq 0 \). It then follows that \( \tilde{W} \) has a continuous local time process \( L^a(\cdot) \) for each \( a \in [0, \infty) \), and, by the occupation time formula (see, for example Revuz and Yor [30, Corollary 1.6])
\[
\int_0^t 1\{\tilde{W}(s) = b\} d[\tilde{W}]_s = \int_0^\infty 1\{b = a\} L^a(t) da = 0 \quad \text{for all } t \geq 0.
\]
We conclude that
\[
\int_0^\infty 1\{\tilde{W}(s) = b\} ds = 0, \quad \text{a.s.} \tag{55}
\]

We next show that \( \mathcal{V}_1 = \mathcal{V}_2 = 0 \) a.s.. Without loss of generality, by the Skorokhod representation theorem, we may replace the weak convergence along a subsequence to \((\tilde{W}, \xi, \mathcal{V}_1, \mathcal{V}_2, \mathcal{V}, 0, \mathcal{L})\) to a.s. u.o.c. convergence. Let \( \eta > 0 \) and define
\[
\hat{\mathcal{V}}^{r, \eta}_t \triangleq \int_0^t \hat{Q}^r(s)1\{\tilde{W}^r(s) > b + \eta\} ds
\]
\[
\hat{\mathcal{V}}^{r, \eta}_t \triangleq \int_0^t \hat{Q}^r(s)1\{b < \tilde{W}^r(s) \leq b + \eta\} ds.
\]

Then, for \( t \geq 0 \), a.s.,
\[
\limsup_{r \to \infty} \mathcal{V}^{r}_t \leq \limsup_{r \to \infty} \hat{\mathcal{V}}^{r, \eta}_t + \limsup_{r \to \infty} \hat{\mathcal{V}}^{r, \eta}_t.
\]
It follows from Proposition 7.1 that
\[
\limsup_{r \to \infty} \hat{\mathcal{V}}^{r, \eta}_t = 0.
\]
Also,
\[
\limsup_{r \to \infty} \hat{\mathcal{V}}^{r, \eta}_t \leq \limsup_{r \to \infty} \frac{\mu_3}{\mu_2} \int_0^t \tilde{W}^r(s)1\{b < \tilde{W}^r(s) \leq b + \eta\} ds,
\]
and also noting that \( \int_0^t \tilde{W}^r(s)1\{b < \tilde{W}^r(s) \leq b + \eta\} ds \leq (b + \eta)t \) for all \( r \), it follows from Fatou’s lemma that
\[
\limsup_{r \to \infty} \frac{\mu_3}{\mu_2} \int_0^t \tilde{W}^r(s)1\{b < \tilde{W}^r(s) \leq b + \eta\} ds \leq \limsup_{r \to \infty} \int_0^t \tilde{W}^r(s)1\{b < \tilde{W}^r(s) \leq b + \eta\} ds.
\]
Since the set of discontinuity points of \( \{\tilde{W}(t)1\{\tilde{W}(t) \in [b, b + \eta]\}, t \geq 0\} \) has Lebesgue measure 0 from (55) and \( 0 \leq \tilde{W}^r(t)1\{b < \tilde{W}^r(t) \leq b + \eta\} \leq b + \eta \) for all \( r \), it follows from Lebesgue’s dominated convergence theorem that
\[
\lim_{r \to \infty} \int_0^t \tilde{W}^r(s)1\{b < \tilde{W}^r(s) \leq b + \eta\} ds = \int_0^t \tilde{W}(s)1\{\tilde{W}(s) \in [b, b + \eta]\} ds.
\]
Hence
\[
\limsup_{r \to \infty} \mathcal{V}^{r}_t \leq \frac{\mu_3}{\mu_2} \int_0^t \tilde{W}(s)1\{\tilde{W}(s) \in [b, b + \eta]\} ds.
\]
Since the left-hand side of the above inequality does not depend on \( \eta \), we can let \( \eta \downarrow 0 \) to conclude that
\[
\limsup_{r \to \infty} \mathcal{V}^{r}_t \leq \frac{\mu_3}{\mu_2} \int_0^t \tilde{W}(s)1\{\tilde{W}(s) = b\} ds.
\]
It now follows from (55) that
\[
\limsup_{r \to \infty} \mathcal{V}^{r}_t = 0,
\]
and so
\[
\mathcal{V}_2 = 0 \quad \text{a.s.}
\]
We omit the argument that \( \mathcal{V}_1 = 0 \) a.s. because it is similar.
Hence it follows by bounded convergence that
\[ r \to 0 \quad \text{as} \quad r \to \infty, \]
implies that
\[ \int_0^t \frac{\mu_1(r^2 \delta^2)}{\mu_2} \mathbb{Q}_2 \mathbb{1} \{ \tilde{W}(t) \leq b \} \to \int_0^t \frac{\mu_1}{\mu_2} \tilde{W}(t) \mathbb{1} \{ \tilde{W}(t) \leq b \} \quad \text{a.s. u.o.c.,} \]
and
\[ \int_0^t \hat{Q}_1(t) \mathbb{1} \{ \tilde{W}(t) > b \} \to \int_0^t \tilde{W}(t) \mathbb{1} \{ \tilde{W}(t) > b \} \quad \text{a.s. u.o.c.,} \]
as \( r \to \infty \). So, by the definition of \( \hat{e}_r \), we conclude that \( \tilde{E} = 0 \) a.s.

We next show that
\[ \tilde{V}(t) = \int_0^t m(\tilde{W}(s)) \tilde{W}(s) \, ds. \]
The function \( m \) is continuous except at \( b \), and so
\[ m(\tilde{W}(\cdot, s)) \to m(\tilde{W}(\cdot, s)) \quad \text{as} \quad r \to \infty \]
for all \( s \geq 0 \) at which \( \tilde{W}(s, \omega) \neq b \). From (55), this is true for Lebesgue-almost everywhere \( s \in [0, \infty) \). Hence for such \( \omega \)
\[ m(\tilde{W}(\cdot, \omega)) \to m(\tilde{W}(\cdot, \omega)) \quad \text{almost everywhere on} \quad [0, \infty), \]
as \( r \to \infty \). We can also multiply by \( \tilde{W}(\cdot, \omega) \) to find
\[ m(\tilde{W}(\cdot, \omega)) \tilde{W}(\cdot, \omega) \to m(\tilde{W}(\cdot, \omega)) \tilde{W}(\cdot, \omega), \quad \text{as} \quad r \to \infty, \quad \text{almost everywhere on} \quad [0, \infty). \]

It follows by bounded convergence that
\[ \int_0^t m(\tilde{W}(s, \omega)) \tilde{W}(s, \omega) \, ds \to \int_0^t m(\tilde{W}(s, \omega)) \tilde{W}(s, \omega) \, ds \quad \text{u.o.c.,} \quad \text{as} \quad r \to \infty, \]
and, furthermore, the convergence is u.o.c. because the integrals are monotone and the sequence is C-tight. Hence
\[ \tilde{V}(t) = \int_0^t m(\tilde{W}(s)) \tilde{W}(s) \, ds. \]
We conclude that
\[ \tilde{W}(t) = \hat{\xi}(t) - \int_0^t m(\tilde{W}(s)) \tilde{W}(s) \, ds + \tilde{L}(t) \quad \text{for all} \quad t \geq 0. \]

7.2.2. The argument that there is uniqueness in law for any solution of (54). Assume that \( (\tilde{W}^1, \tilde{L}^1, \tilde{\xi}^1), \) \( (\tilde{\Omega}^1, \tilde{\mathbb{F}}^1, \tilde{\mathbb{P}}^1), \{ \tilde{\mathbb{F}}^1_t \} \) and \( (\tilde{W}^2, \tilde{L}^2, \tilde{\xi}^2), \) \( (\tilde{\Omega}^2, \tilde{\mathbb{F}}^2, \tilde{\mathbb{P}}^2), \{ \tilde{\mathbb{F}}^2_t \} \) are both weak solutions to (54), with \( \tilde{\xi}^1(0) = \tilde{\xi}^2(0) = 0 \). To show that there is uniqueness in law, it is enough to show that \( (\tilde{W}^1, \tilde{L}^1, \tilde{\xi}^1) \) and \( (\tilde{W}^2, \tilde{L}^2, \tilde{\xi}^2) \) have the same law under their respective probability measures. To do this, we will use the Girsanov Theorem (see, for example, Karatzas and Shreve [22, Theorem 3.5.1], and note that the process \( X \) in Karatzas and Shreve [22] is exactly \( m(\tilde{W}^k) \tilde{W} \)).

The first step is to observe that for any \( T \geq 0 \)
\[ P \left( \int_0^T (m(\tilde{W}(s)) \tilde{W}(s))^2 \, ds < \infty \right) = 1. \]
This is because
\[ m(\tilde{W}(s)) \tilde{W}(s) \leq \max(\delta_1, \delta_2) \tilde{W}(s) \quad \text{for all} \quad s \geq 0, \]
and Lemma 5.3 is valid for \( \tilde{W} \) so that, for \( X^R \) defined as in Lemma 5.3,
\[ \tilde{W}(t) \leq X^R(t) \quad \text{for every} \quad t > 0, \]
so that
\[ P \left( \int_0^T (m(\tilde{W}(s)) \tilde{W}(s))^2 \, ds < \infty \right) \leq P \left( \max(\delta_1, \delta_2)^2 \int_0^T (X^R(s))^2 \, ds < \infty \right). \]
It follows from the properties of reflected Brownian motion that
\[
P\left( \max(\delta_1, \delta_2)^2 \int_0^T (X^k(s))^2 \, ds < \infty \right) = 1.
\]

Next, define
\[
M^k(t) := \exp\left( \int_0^t m(\tilde{W}^k(s))^2 \, ds - \frac{1}{2} \int_0^t (m(\tilde{W}^k(s))^2 \, ds \right), \quad k \in \{1, 2\}.
\]

If we can show that the Novikov condition (see, for example, Corollary 3.5.13 in Karatzas and Shreve [22])
\[
E_p\left[ \exp\left( \frac{1}{2} \int_0^t (m(\tilde{W}^k(s))^2 \, ds \right) \right] < \infty, \quad 0 \leq t < \infty, \quad k \in \{1, 2\},
\]
is satisfied, then we can conclude that \(M^k\) is a martingale. This implies that we can define the probability measures \(Q^k, k \in \{1, 2\}\), according to
\[
dQ^k = M^k \, dP, \quad k \in \{1, 2\}.
\]

It follows from the Girsanov theorem that for any \(T > 0\), the processes
\[
\tilde{X}^k(t) := \tilde{\xi}^k(t) - \int_0^t m(\tilde{W}^k(s))\tilde{W}^k(s) \, ds, \quad 0 \leq t \leq T, \quad k \in \{1, 2\},
\]
are driftless Brownian motions on \((\Omega^k, \mathcal{F}^k, Q^k)\) with initial condition \(X^1(0) = X^2(0) = 0\). Also, \(L^k\) continues to satisfy the conditions
\[
L^k(0) = 0, \quad L^k \text{ is nondecreasing, and } \int_0^T \tilde{W}^k(t) \, dL^k(t) = 0
\]
on \((\Omega^k, \mathcal{F}^k, Q^k)\). Hence
\[
\tilde{W}^k(t) = \tilde{X}^k(t) + L^k(t), \quad 0 \leq t \leq T, \quad k \in \{1, 2\},
\]
is a reflected Brownian motion under \(Q^k\), and so \((\tilde{W}^1, \tilde{L}^1, \tilde{X}^1)\) and \((\tilde{W}^2, \tilde{L}^2, \tilde{X}^2)\) have the same law under their respective probability measures. This implies that \((\tilde{W}^1, \tilde{L}^1, \tilde{\xi}^1)\) and \((\tilde{W}^2, \tilde{L}^2, \tilde{\xi}^2)\) also have the same law under their respective probability measures. To see this, note that we can view \(\tilde{\xi}^k\) as defined in terms of \(\tilde{X}^k\). Then, for any \(0 = t_0 < t_1 < \cdots < t_n = T\) and \(\Gamma \in \mathcal{B}([0,T])\) a Borel set,
\[
P^1((\tilde{W}^1(t_0), \tilde{L}^1(t_0), \tilde{\xi}^1(t_0), \ldots, \tilde{W}^1(t_n), \tilde{L}^1(t_n), \tilde{\xi}^1(t_n)) \in \Gamma)
\]
\[
= \int_{\Omega^1} \frac{1}{M_1} \mathbf{1}\{ (\tilde{W}^1(t_0), \tilde{L}^1(t_0), \tilde{\xi}^1(t_0), \ldots, \tilde{W}^1(t_n), \tilde{L}^1(t_n), \tilde{\xi}^1(t_n)) \in \Gamma \} \, dQ^1
\]
\[
= \int_{\Omega^2} \frac{1}{M_2} \mathbf{1}\{ (\tilde{W}^2(t_0), \tilde{L}^2(t_0), \tilde{\xi}^2(t_0), \ldots, \tilde{W}^2(t_n), \tilde{L}^2(t_n), \tilde{\xi}^2(t_n)) \in \Gamma \} \, dQ^2
\]
\[
= P^2((\tilde{W}^2(t_0), \tilde{L}^2(t_0), \tilde{\xi}^2(t_0), \ldots, \tilde{W}^2(t_n), \tilde{L}^2(t_n), \tilde{\xi}^2(t_n)) \in \Gamma).
\]

It remains to show (56). For convenience in presentation, we drop the \(k\) superscript. Let \(\tilde{\delta} = \max(\delta_1, \delta_2)\), and note that
\[
\frac{1}{2} \int_0^T (m(\tilde{W}(s))^2 \, ds \leq \frac{1}{2} \tilde{\delta}^2 T \sup_{0 \leq t \leq T} \tilde{W}(t)^2.
\]
Since \(\tilde{W}(0) = 0\), \(\sup_{0 \leq t \leq T} \tilde{W}(t) = \text{Osc}(\tilde{W}, [0, T])\). Also note that for any \(x \in D^1\), \(\text{Osc}(x, [0, T]) \leq 2 \sup_{0 \leq t \leq T} |x(t)|\). It follows from the oscillation inequality for reflected Brownian motion (see, for example, Williams [37, Theorem 5.1]) that
\[
\text{Osc}(\tilde{W}, [0, T]) \leq \text{Osc}(\tilde{\xi}, [0, T]).
\]

Hence
\[
\sup_{0 \leq t \leq T} \tilde{W}(t) \leq 2 \sup_{0 \leq t \leq T} |\tilde{\xi}(t)|.
\]

Since \(\sup_{0 \leq t \leq T} \tilde{W}(t)^2 = (\sup_{0 \leq t \leq T} |\tilde{W}(t)|)^2\), it follows from (57) and (58) that
\[
\frac{1}{2} \int_0^T (m(\tilde{W}(s))^2 \, ds \leq \frac{1}{2} \tilde{\delta}^2 T \left( 2 \sup_{0 \leq t \leq T} |\tilde{\xi}(t)| \right)^2.
\]
Hence
\[
\exp \left( \int_0^T (m(T\tilde{W}(s))) ds \right) \leq \exp \left( 2\delta^2 T \left( \sup_{0 \leq s \leq T} (T\tilde{\xi}(t)) \right)^2 \right).
\]

Since \( \exp(2\delta^2 T(\sup_{0 \leq s \leq T} |\tilde{\xi}(t)|)^2) = (\sup_{0 \leq s \leq T} \exp(\delta^2 T\tilde{\xi}(t))^2) \), we have that
\[
\exp \left( \int_0^T (m(T\tilde{W}(s))) ds \right) \leq \left( \sup_{0 \leq s \leq T} \exp(\delta^2 T\tilde{\xi}(t))^2 \right).
\]

The process \( \tilde{\xi} \) is a martingale because it is a Brownian motion, and \( \exp(\delta^2 Tx^2) \) is a convex function in \( x \). Hence \( \exp(\delta^2 T\tilde{\xi}(t))^2 \) is a submartingale. Then, it follows from Doob’s maximal inequality (see, for example, Karatzas and Shreve [22, Theorem 1.3.8, part (iv)]) that
\[
E \left[ \left( \sup_{0 \leq s \leq T} \exp(\delta^2 T\tilde{\xi}(t))^2 \right) \right] \leq 4E(\exp(\delta^2 T\tilde{\xi}(T))^2) < \infty,
\]
which completes the proof.

We note that another potential strategy for obtaining the weak convergence of the diffusion-scaled workload process in Proposition 7.2 may be to modify Atar et al. [5, Theorem 2.11] to account for a discontinuity in the linear portion of the drift (in addition to the assumed discontinuity in the constant portion of the drift).

### 7.3. Proof of Theorem 7.1

It follows from Proposition 7.2 and the use of the Skorokhod representation theorem to replace weak convergence with a.s. u.o.c. convergence that
\[
(W^r, \tilde{V}_1^r, \tilde{V}_2^r) \rightarrow (W^*, 0, 0) \quad \text{a.s., u.o.c., as } r \to \infty.
\]

Since
\[
\tilde{W}^r = \tilde{Q}_1^r + \frac{\mu_2}{\mu_3} \tilde{Q}_2^r,
\]
and the functions \( 1\{x \geq b\}, 1\{x < b\} \) are continuous everywhere except at \( b \),
\[
\tilde{Q}_1^r(t)1\{\tilde{W}^r(t) \geq b\} + \frac{\mu_2}{\mu_3} \tilde{Q}_2^r(t)1\{\tilde{W}^r(t) \geq b\} \to \tilde{W}^*(t)1\{\tilde{W}^*(t) \geq b\}, \quad \text{as } r \to \infty,
\]
and
\[
\tilde{Q}_1^r(t)1\{\tilde{W}^r(t) < b\} + \frac{\mu_2}{\mu_3} \tilde{Q}_2^r(t)1\{\tilde{W}^r(t) < b\} \to \tilde{W}^*(t)1\{\tilde{W}^*(t) < b\}, \quad \text{as } r \to \infty,
\]
for all \( t \geq 0 \) at which \( \tilde{W}^*(t) \neq b \). On any sample path on which
\[
\int_0^\infty 1\{\tilde{W}^*(s) = b\} ds = 0,
\]
this is true Lebesgue almost everywhere \( t \in [0, \infty) \). Since for any \( t \geq 0 \)
\[
\int_0^t e^{-\gamma s} \tilde{Q}_1^r(s) 1\{\tilde{W}^r(s) < b\} ds \leq \frac{1}{r^2 \delta_1^r} \tilde{V}_1^r(t),
\]
and
\[
\int_0^t e^{-\gamma s} \tilde{Q}_2^r(s) 1\{\tilde{W}^r(s) \geq b\} ds \leq \frac{1}{r^2 \delta_2^r \mu_2} \tilde{V}_2^r(t),
\]
it follows that
\[
\int_0^t e^{-\gamma s} \tilde{Q}_1^r(s) 1\{\tilde{W}^r(s) < b\} ds \to 0 \quad \text{a.s.,}
\]
\[
\int_0^t e^{-\gamma s} \tilde{Q}_2^r(s) 1\{\tilde{W}^r(s) \geq b\} ds \to 0 \quad \text{a.s.,}
\]
as \( r \to \infty \). Then, from (59) and (60), from bounded convergence,
\[
\int_0^t e^{-\gamma s} \tilde{Q}_1^r(s) 1\{\tilde{W}^r(s) \geq b\} ds \to \int_0^t e^{-\gamma s} \tilde{W}^*(s) 1\{\tilde{W}^*(s) \geq b\} ds \quad \text{a.s.,}
\]
\[
\int_0^t e^{-\gamma s} \tilde{Q}_2^r(s) 1\{\tilde{W}^r(s) < b\} ds \to \int_0^t e^{-\gamma s} \tilde{W}^*(s) 1\{\tilde{W}^*(s) < b\} ds \quad \text{a.s.,}
\]

\[
\int_0^T (m(T\tilde{W}(s))) ds \to \int_0^T (m(T\tilde{W}^*(s))) ds \quad \text{a.s.,}
\]
as \( r \to \infty \). We conclude that
\[
\left( \int_0^r e^{-\gamma s} \hat{Q}_1(s) ds, \int_0^r e^{-\gamma s} \hat{Q}_2(s) ds \right) \\
\to \left( \int_0^r e^{-\gamma s} \tilde{W}^*(s) \mathbf{1}\{\tilde{W}^*(s) \geq b\} ds, \int_0^r e^{-\gamma s} \frac{\mu_1}{\mu_2} \tilde{W}^*(s) \mathbf{1}\{\tilde{W}^*(s) < b\} ds \right)
\]
as \( r \to \infty \). Hence
\[
\left( \int_0^r e^{-\gamma s} \hat{Q}_1(s) ds, \int_0^r e^{-\gamma s} \hat{Q}_2(s) ds \right) \\
\Rightarrow \left( \int_0^r e^{-\gamma s} \tilde{W}^*(s) \mathbf{1}\{\tilde{W}^*(s) \geq b\} ds, \int_0^r e^{-\gamma s} \frac{\mu_1}{\mu_2} \tilde{W}^*(s) \mathbf{1}\{\tilde{W}^*(s) < b\} ds \right),
\]
as \( r \to \infty \). The proof is complete because when \( b = b^* \)
\[
\left( \tilde{W}^*(s) \mathbf{1}\{\tilde{W}^*(s) \geq b\}, \frac{\mu_1}{\mu_2} \tilde{W}^*(s) \mathbf{1}\{\tilde{W}^*(s) < b\} \right) = q^*(\tilde{W}^*(s))
\]
for \( q^* \) defined in (34).

8. Asymptotic optimality. We prove the following theorem.

**Theorem 8.1.** Assume that
\[
(h_2 + c_2 \delta_2) \mu_3 < (h_1 + c_1 \delta_1) \mu_2 \quad \text{and} \quad \frac{\delta_1 + \gamma}{\delta_2 + \gamma} > \frac{\mu_2 (h_1 + c_1 \delta_1)}{\mu_3 (h_2 + c_2 \delta_2)},
\]
so that \( q^* \) given in (34) solves the WCP, for \( b^* < \infty \). When the \( r \)th system operates under the threshold policy \( T^{r,*} \) given in Definition 6.1,
\[
\lim_{r \to \infty} \hat{J}^r(T^{r,*}) = E \left[ \int_0^\infty e^{-\gamma t} g(q^*(\tilde{W}^*(t))) dt \right]. \tag{61}
\]
Furthermore, if \( \{T^r\} \) is any sequence of admissible scheduling control policies (one for each member of the sequence of \( N \)-systems), then
\[
\liminf_{r \to \infty} \hat{J}^r(T^r) \geq E \left[ \int_0^\infty e^{-\gamma t} g(q^*(\tilde{W}^*(t))) dt \right]. \tag{62}
\]
Theorem 8.1 shows that
\[
E \left[ \int_0^\infty e^{-\gamma t} g(q^*(\tilde{W}^*(t))) dt \right]
\]
is the minimum cost that can be achieved by any admissible scheduling control policy asymptotically, and that this asymptotically minimal cost is achieved by the sequence of dynamic threshold policies \( \{T^{r,*}\} \). Thus we conclude that the sequence of threshold policies \( \{T^{r,*}\} \) is asymptotically optimal.

**Proof of Theorem 8.1.** We first show (61) and then (62).

**Proof of (61).** We use the following three lemmas to establish (61).

**Lemma 8.1.** There exist constants \( d_1 > 0 \) and \( d_2 > 0 \) (that do not depend on \( r \)) such that for any \( t > 0 \)
\[
E \left[ \sup_{0 \leq s \leq t} |\hat{X}^r(s)| \right] \leq d_1 + d_2 t, \quad \text{for all } r.
\]
In particular, the family \( \{ \int_0^\infty e^{-\gamma t} \sup_{0 \leq s \leq t} |\hat{X}^r(s)| dt \} \) is uniformly integrable.

**Lemma 8.2.** There exist constants \( d_1 > 0 \) and \( d_2 > 0 \) (that do not depend on \( r \)) such that for any \( t > 0 \)
\[
E \left[ \sup_{0 \leq s \leq t} |\hat{Y}^r(s)| \right] \leq d_1 + d_2 t^2, \quad \text{for large enough } r.
\]
In particular, the family \( \{ \int_0^\infty e^{-\gamma t} \sup_{0 \leq s \leq t} |\hat{Y}^r(s)| dt \} \) is uniformly integrable.
**Lemma 8.3.** Suppose that $L \geq 0$, $0 \leq T < \infty$, and $w, x, y, z \in \mathbb{D}$ are such that $w(0) = 0$ and

1. $w(t) = x(t) - \int_0^t z(s) \, ds + y(t)$ for all $t \in [0, T]$,
2. $u(t) \geq 0$ and $z(t) \geq 0$ for all $t \in [0, T]$,
3. (a) $y(0) \geq 0$,
   (b) $y$ is nondecreasing, and
   (c) $\int_{[0,T]} 1\{w(t) > L\} \, dy(t) = 0$.

Then, there is a constant $\zeta > 0$ (not depending on $T$) such that

$\text{Osc}(w, [0, T]) \leq \zeta (\text{Osc}(x, [0, T]) + L)$.

Lemma 8.3 generalizes the one-dimensional version in Williams [37, Theorem 5.1]. The essential condition required in our proof is that the $z$ appearing in the statement of the lemma is nonnegative.

The fact that the proposed policy is admissible follows from Lemma 7.2, Lemma 8.2, and its definition. It follows from Theorem 7.1 that for any $T > 0$,

$$\int_0^T e^{-\gamma t} g(\hat{Q}(t)) \, dt \Rightarrow \int_0^T e^{-\gamma t} g(q^*(\hat{W}^*(t))) \, dt \quad \text{as } r \to \infty. \quad (63)$$

From Theorem 3.4 in Billingsley [9],

$$E\left[\int_0^T e^{-\gamma t} g(q^*(\hat{W}^*(t))) \, dt\right] \leq \liminf_{r \to \infty} E\left[\int_0^T e^{-\gamma t} g(\hat{Q}(t)) \, dt\right].$$

Also,

$$\liminf_{r \to \infty} E\left[\int_0^T e^{-\gamma t} g(\hat{Q}(t)) \, dt\right] \leq \liminf_{r \to \infty} E\left[\int_0^T e^{-\gamma t} g(q^*(\hat{W}^*(t))) \, dt\right].$$

Next,

$$\limsup_{r \to \infty} E\left[\int_0^T e^{-\gamma t} g(\hat{Q}(t)) \, dt\right] \leq \limsup_{r \to \infty} E\left[\int_0^T e^{-\gamma t} g(\hat{Q}(t)) \, dt\right] + \limsup_{r \to \infty} E\left[\int_0^T e^{-\gamma t} g(q^*(\hat{W}^*(t))) \, dt\right].$$

Since $Q'_1(t) \leq \hat{W}^*(t)$ and $Q'_2(t) \leq (\mu'_1/\mu'_2)\hat{W}^*(t)$ for all $t \geq 0$, $g(\hat{Q}(t)) \leq \left(h_1 + c_1 \hat{d}_1 + (h_2 + c_2 \hat{d}_2) \left(\frac{\mu_1}{\mu_2} + 1\right)\right)\hat{W}^*(t)$,

so that

$$E\left[\int_T^\infty e^{-\gamma t} g(\hat{Q}(t)) \, dt\right] \leq \left(h_1 + c_1 \hat{d}_1 + (h_2 + c_2 \hat{d}_2) \left(\frac{\mu_1}{\mu_2} + 1\right)\right) E\left[\int_T^\infty e^{-\gamma t} \sup_{0 \leq s \leq t} \hat{W}^*(s) \, dt\right].$$

Furthermore, by Fubini’s theorem, since $e^{-\gamma t} > 0$ and $W^*(t) \geq 0$ for all $t \geq 0$,

$$E\left[\int_T^\infty e^{-\gamma t} \sup_{0 \leq s \leq t} \hat{W}^*(s) \, dt\right] = \int_T^\infty e^{-\gamma t} E\left[\sup_{0 \leq s \leq t} \hat{W}^*(s)\right] \, dt.$$

In summary,

$$E\left[\int_0^T e^{-\gamma t} g(q^*(\hat{W}^*(t))) \, dt\right] \leq \liminf_{r \to \infty} E\left[\int_0^T e^{-\gamma t} g(\hat{Q}(t)) \, dt\right] \leq \limsup_{r \to \infty} E\left[\int_0^T e^{-\gamma t} g(\hat{Q}(t)) \, dt\right] \leq \limsup_{r \to \infty} E\left[\int_0^T e^{-\gamma t} g(q^*(\hat{W}^*(t))) \, dt\right] + \left(h_1 + c_1 \hat{d}_1 + (h_2 + c_2 \hat{d}_2) \left(\frac{\mu_1}{\mu_2} + 1\right)\right) \limsup_{r \to \infty} \int_T^\infty e^{-\gamma t} E\left[\sup_{0 \leq s \leq t} \hat{W}^*(s)\right] \, dt. \quad (64)$$
Suppose we can show that
\[
\lim_{T \to \infty} \lim_{r \to \infty} \int_0^\infty e^{-\gamma t} \mathbb{E} \left[ \sup_{0 \leq s \leq t} \hat{W}'(s) \right] dt = 0, \tag{65}
\]
and that
\[
\lim_{r \to \infty} \mathbb{E} \left[ \int_0^T e^{-\gamma t} g(\hat{Q}'(t)) dt \right] = \mathbb{E} \left[ \int_0^T e^{-\gamma t} g(q^*(\hat{W}'(t))) dt \right]. \tag{66}
\]

Since
\[
\mathbb{E} \left[ \int_0^T e^{-\gamma t} g(q^*(\hat{W}'(t))) dt \right] \to \mathbb{E} \left[ \int_0^\infty e^{-\gamma t} g(q^*(\hat{W}*\text{(t)})) dt \right], \text{ as } T \to \infty,
\]

by the monotone convergence theorem, we conclude by taking the limit as \( T \to \infty \) in (64) that
\[
\lim_{r \to \infty} \mathbb{E} \left[ \int_0^\infty e^{-\gamma t} g(q^*(\hat{W}'(t))) dt \right] = \mathbb{E} \left[ \int_0^\infty e^{-\gamma t} g(q^*(\hat{W}*\text{(t)})) dt \right].
\]

To see (65) and (66), it is useful to use the oscillation inequality in Lemma 8.3 to establish an upper bound on \( \sup_{0 \leq t \leq T} \hat{W}'(t) \). Note that
\[
\hat{W}'(t) = \hat{X}'(t) + \hat{e}'_t(t) = \int_0^t \left( (r^2 \hat{\delta}_1') \hat{Q}'(s) + \frac{\mu'_2}{\mu'_3} (r^2 \hat{\delta}_2') \hat{Q}'_2(s) \right) ds + \mu'_1 \hat{I}'_1(t) + \mu'_2 \hat{I}'_2(t),
\]

for
\[
\hat{e}'_t(t) = \frac{1}{r} \int_0^t \left( 1[\hat{I}'_1(t) > 0] + 1[\hat{I}'_2(t) > 0] \right) + \frac{\mu'_2}{\mu'_3} 1[\hat{I}'_2(t) > 0] ds.
\]

Then,
\[
w = \hat{W}',
\]
\[
x = \hat{X}' + \mu'_1 \hat{I}'_1 + \hat{e}'_t,
\]
\[
z = (r^2 \hat{\delta}_1') \hat{Q}' + \frac{\mu'_2}{\mu'_3} (r^2 \hat{\delta}_2') \hat{Q}'_2,
\]
\[
y = \mu'_2 \hat{I}'_2,
\]
satisfies the conditions of Lemma 8.3 with \( L = L'/r \). (Note that \( \hat{W}'(0) = 0 \). Also, the fact that Condition 3(c) is satisfied follows as in (53) in the proof of Proposition 7.2.) Hence there exists a constant \( \xi > 0 \) that does not depend on \( r \) such that for any \( T > 0 \),
\[
\text{Osc}(\hat{W}', [0, T]) \leq \xi \left( \text{Osc}(\hat{X}' + \hat{e}'_t + \mu'_1 \hat{I}'_1, [0, T]) + \frac{L'}{r} \right).
\]

Since \( \hat{W}'(0) = 0 \) and \( \hat{W}'(t) \geq 0 \) for all \( t \geq 0 \),
\[
\sup_{0 \leq t \leq T} \hat{W}'(t) = \text{Osc}(\hat{W}', [0, T]).
\]

Also
\[
\text{Osc}(\hat{X}' + \hat{e}'_t + \mu'_1 \hat{I}'_1, [0, T]) \leq 2 \sup_{0 \leq t \leq T} |\hat{X}'(t)| + 2 \sup_{0 \leq t \leq T} |\hat{e}'_t(t)| + 2 \mu'_1 \sup_{0 \leq t \leq T} \hat{I}'_1(t).
\]

Hence it follows that
\[
\sup_{0 \leq t \leq T} \hat{W}'(t) \leq \xi \left( 2 \sup_{0 \leq t \leq T} |\hat{X}'(t)| + 2 \sup_{0 \leq t \leq T} |\hat{e}'_t(t)| + 2 \mu'_1 \sup_{0 \leq t \leq T} \hat{I}'_1(t) + \frac{L'}{r} \right). \tag{67}
\]

The validity of (65) follows from (67), Lemmas 8.1 and 8.2, and the fact that \( \hat{e}'_t(t) \leq (t/r)(2 + \mu'_2/\mu'_3) \) by its definition.

To establish (66), recall the weak convergence in (63) and note that if the family \( \{ \int_0^T e^{-\gamma t} g(\hat{Q}'(t)) dt \} \) is uniformly integrable, then
\[
\lim_{r \to \infty} \mathbb{E} \left[ \int_0^T e^{-\gamma t} g(\hat{Q}'(t)) dt \right] = \mathbb{E} \left[ \int_0^T e^{-\gamma t} g(q^*(\hat{W}*\text{(t)})) dt \right].
\]
so that (66) holds. Since \( \tilde{Q}'(t) \leq \tilde{W}'(t) \) and \( \tilde{Q}''(t) \leq (\mu_4/\mu_3)\tilde{W}'(t) \) for all \( t \geq 0 \), to show the family \( \{ \int_0^t e^{-r\gamma}g(\tilde{Q}'(t)) \, dt \} \) is uniformly integrable, it is sufficient to show that the family \( \{ \int_0^t e^{-r\gamma}\tilde{W}'(t) \, dt \} \) is.

This follows because (67) implies that \( \{ \int_0^t e^{-r\gamma}\tilde{W}'(t) \, dt \} \) is dominated by a uniformly integrable family, since \( \{ \int_0^\infty e^{-r\gamma}S(\tilde{W}')(t) \, dt \} \) and \( \{ \int_0^\infty e^{-r\gamma}S(\tilde{W}')(t) \, dt \} \) are uniformly integrable by Lemmas 8.1 and 8.2, and \( \tilde{\epsilon}'(t) \leq (t/r)(2 + \mu_2/\mu_3) \) by its definition.

**Proof of (62).** It is useful for the argument to establish (62) to first state two lemmas, which are the equivalent in Bell and Williams [8, Lemmas 9.2 and 9.3] in our setting. To do this, in addition to the fluid-scaled processes \( \tilde{Q}' \) and \( \tilde{T}' \), also define

\[
\tilde{A}'(t) := \frac{1}{r^2} \tilde{A}'(r^2 t),
\]

\[
\tilde{S}'(t) := \frac{1}{r^2} \tilde{S}'(r^2 t),
\]

\[
\tilde{N}'(t) := \frac{1}{r^2} \tilde{N}'(r^2 t),
\]

\[
\tilde{I}_1'(t) := \frac{1}{r^2} \tilde{I}_1'(r^2 t),
\]

\[
\tilde{I}_2'(t) := \frac{1}{r^2} \tilde{I}_2'(r^2 t).
\]

Then, it follows from Equations (4)–(7) for the queue-length and idle time processes that

\[
\tilde{Q}'(t) = \tilde{A}'(t) - \tilde{S}'(t) - \tilde{N}'(t) - \tilde{I}_1'(t) - \tilde{I}_2'(t),
\]

\[
\tilde{Q}''(t) = \tilde{A}''(t) - \tilde{S}''(t) - \tilde{N}''(t) - \tilde{I}_1''(t) - \tilde{I}_2''(t),
\]

\[
\tilde{I}_1'(t) := \frac{1}{r^2} \tilde{I}_1'(r^2 t),
\]

\[
\tilde{I}_2'(t) := \frac{1}{r^2} \tilde{I}_2'(r^2 t).
\]

**Lemma 8.4.** Let \( \{ T' \} \) be any sequence of admissible scheduling control policies (one for each member of the sequence of \( N \) systems). Then,

\[
\{ (\tilde{Q}', \tilde{A}', \tilde{S}', \tilde{N}', \tilde{T}', \tilde{I}') \}
\]

is \( C \)-tight.

**Lemma 8.5.** Let \( \{ T' \} \) be a sequence of admissible scheduling controls such that

\[
\liminf_{r \to \infty} \tilde{J}'(T') < \infty.
\]

Consider a subsequence \( \{ T'' \} \) of \( \{ T' \} \) along which the liminf is achieved; i.e.,

\[
\lim_{r \to \infty} \tilde{J}'(T'') = \liminf_{r \to \infty} \tilde{J}'(T').
\]

Then, as \( r' \to \infty \),

\[
(\tilde{Q}', \tilde{A}', \tilde{S}', \tilde{N}', \tilde{T}', \tilde{I}') = (0, \lambda, \mu, e, e, 0, \tilde{T}', 0).
\]

The next step is to note that

\[
\tilde{W}'(t) = \tilde{X}'(t) - \int_0^t (r^2 \delta') \tilde{Q}'(s) + (r^2 \delta') \mu_1 \tilde{I}'(t) + (r^2 \delta') \mu_2 \tilde{I}'(t) + \tilde{\epsilon}'(t)
\]

for

\[
\tilde{\epsilon}'(t) := \int_0^t \frac{1}{r^2} \left( 1[\tilde{I}'(r^2 s) > 0] + 1[\tilde{I}'(r^2 s) > 0] + \mu_2 \mu_3 \right) \ dx.
\]

and that \( \mu_1 \tilde{I}' + \mu_2 \tilde{I}' \) is nondecreasing and has \( \mu_1 \tilde{I}'(0) + \mu_2 \tilde{I}'(0) = 0 \). Then, it follows from the minimality of the Skorokhod map (see Bell and Williams [8, Proposition B.1]) that

\[
\tilde{W}'(t) \geq \tilde{W}'(t) := \phi \left( \tilde{X}' - (r^2 \delta') \int_0^t \tilde{Q}'(s) ds - (r^2 \delta') \int_0^t \mu_2 \tilde{I}'(s) ds + \tilde{\epsilon}'(t) \right).
\]
and that
\[ \mu_1^* \hat{r}^1(t) + \mu_2^* \hat{r}^2(t) \geq \hat{L}^r(t) := \psi \left( \hat{X}^r - (r^2 \delta) \int_0^t \hat{Q}_1(s) ds - (r^2 \delta) \int_0^t \frac{\mu_3^*}{\mu_3^*} \hat{Q}_2(s) ds + \epsilon^r_t \right) (t) \]
for every \( t \geq 0 \). Because it is also true that
\[ \hat{W}^r(t) = \hat{Q}_1(t) + \frac{\mu_2^*}{\mu_3^*} \hat{Q}_2(t) \]
for every \( t \geq 0 \), there exist processes \( \hat{Q}_1(t) \) and \( \hat{Q}_2(t) \) such that
\[ 0 \leq \hat{Q}_1(t) \leq \hat{Q}_1(t) \quad \text{and} \quad 0 \leq \hat{Q}_2(t) \leq \hat{Q}_2(t), \]
and
\[ \hat{Q}_1(t) + \frac{\mu_2^*}{\mu_3^*} \hat{Q}_2(t) = \hat{W}^r(t), \quad \text{(72)} \]
for every \( t \geq 0 \). For example, define
\[ d'(t) := \hat{W}^r(t) - \hat{W}^r(t), \quad \text{for every} \ t \geq 0, \]
to be the amount the workload has dropped. Then,
\[
(\hat{Q}_1(t), \hat{Q}_2(t)) = \begin{cases} 
\left( \hat{Q}_1(t), \hat{Q}_2(t) - \frac{\mu_3^*}{\mu_2^*} d'(t) \right) & \text{if} \ \frac{\mu_3^*}{\mu_2^*} d'(t) \leq \hat{Q}_2(t), \\
(\hat{Q}_1(t) - d'(t), \hat{Q}_2(t)) & \text{if} \ \frac{\mu_3^*}{\mu_2^*} d'(t) > \hat{Q}_2(t), \ d'(t) \leq \hat{Q}_1(t), \\
(\hat{W}^r(t) - d'(t), 0) & \text{if} \ \frac{\mu_3^*}{\mu_2^*} d'(t) > \hat{Q}_2(t), \ d'(t) > \hat{Q}_1(t).
\end{cases}
\]
satisfies (73) and (72). The importance of the processes \( \hat{Q}_1 \) and \( \hat{Q}_2 \) is that for every \( r \)
\[ \hat{J}'(T^r) \geq E \left[ \int_0^\infty e^{-\gamma r} g(\hat{Q}^r(t)) dt \right]. \quad \text{(73)} \]
The following lemma establishes useful properties of \( \hat{Q}_1, \hat{Q}_2, \) and \( \hat{W}^r \).

**Lemma 8.6.** The family \( \{ \int_0^T e^{-\gamma r} g(\hat{Q}^r(t)) dt \} \) is uniformly integrable. Also,
\[ \lim_{T \to \infty} \limsup_{t \to \infty} \int_t^\infty e^{-\gamma r} E \left[ \sup_{0 \leq s \leq t} \hat{W}^r(s) \right] dt = 0 \quad \text{(74)} \]
and, on any subsequence \( r' \) on which the liminf in Lemma 8.5 is achieved, for any \( t > 0 \),
\[ \sup_{r'} E \left[ \int_0^t \hat{W}^r(s) ds \right] < \infty. \quad \text{(75)} \]
The final lemma we require in this proof establishes \( C \)-tightness.

**Lemma 8.7.** Let \( \{ T^r \} \) be a sequence of admissible scheduling controls such that
\[ \lim_{r \to \infty} \hat{J}'(T^r) < \infty. \]
Consider a subsequence \( \{ T^{r'} \} \) of \( \{ T^r \} \) along which the liminf is achieved; i.e.,
\[ \lim_{r' \to \infty} \hat{J}'(T^{r'}) = \liminf_{r \to \infty} \hat{J}'(T^r). \]
Then, \( \{ (\hat{X}^r, \int_0^T \hat{Q}_1^r(s) ds, \int_0^T \hat{Q}_2^r(s) ds) \} \) is \( C \)-tight.
If \( \lim\inf_{r \to \infty} \dot{J}(T') = \infty \), then (62) holds trivially. Hence assume \( \lim\inf_{r \to \infty} \dot{J}(T') < \infty \). It follows from Lemma 8.5 that on any subsequence \( \{T'\} \) of \( \{T'\} \) along which the \( \lim\inf \) is achieved, 
\[
(Q', \tilde{A}', \tilde{S}', \tilde{N}', \tilde{z}', \tilde{v}', \tilde{T}') \Rightarrow (0, \lambda e, \mu e, e, 0, \tilde{T}', 0),
\]
as \( r' \to \infty \). Then, it follows from Lemma 8.7 that there is a further subsequence \( r'' \) on which
\[
\left( \dot{X}', \int_0^{Q_1}(s) ds, \int_0^{Q_2}(s) ds \right) \Rightarrow (\xi', \tilde{e}_1', \tilde{e}_2'), \quad \text{as} \quad r'' \to \infty.
\]
Furthermore, on the subsequence \( r'' \), by the continuous mapping theorem and the fact that \( \epsilon_{r''} \to 0 \) as \( r'' \to \infty \),
\[
(W^{r''}, Z^{r''}) \Rightarrow (\bar{W}, \bar{L}) := \left( \phi \left( \tilde{\xi} - \delta_1 \tilde{e}_1 - \delta_2 \frac{\mu_2}{\mu_3} \tilde{e}_2 \right), \psi \left( \tilde{\xi} - \delta_1 \tilde{e}_1 - \delta_2 \frac{\mu_2}{\mu_3} \tilde{e}_2 \right) \right).
\]
Next, on the subsequence \( r'' \), we can apply in Atar et al. [7, Lemma 5], which is a special case of a result in Kurtz and Protter [24], to conclude that
\[
\int_0^{Q_2}(t) dt \Rightarrow \int_0^{d\tilde{e}_k}(t), \quad k \in \{1, 2\},
\]
and that
\[
\int_0^{e^{-\gamma}g(Q^{r''}(t)) dt} \Rightarrow \int_0^{e^{-\gamma}g(d\tilde{e}(t))},
\]
as \( r'' \to \infty \). In particular, to show that the conditions in Atar et al. [7, Lemma 5] are satisfied, first note that the process \( \int_0^{Q_2}(t) dt \), \( k \in \{1, 2\} \), has total variation over \([0, t]\) for any \( t > 0 \), \( T(\int_0^{Q_2}(t) dt) \), that is finite since
\[
T_r \left( \int_0^{Q_2}(t) dt \right) = \int_0^{Q_2}(s) ds < \infty.
\]
Furthermore,
\[
\sup_{r''} \mathbb{E} \left[ T_r \left( \int_0^{Q_2}(t) dt \right) \right] \leq \sup_{r''} \mathbb{E} \left[ \int_0^{\max \left( 1, \frac{\mu_2}{\mu_3} \right)} \bar{W}^{r''}(s) ds \right] < \infty, \quad k \in \{1, 2\},
\]
where the first inequality follows from (72) and the second follows from (75) in Lemma 8.6. Then, (77) and (78) follow when we take the process \( U^n \) in Atar et al. [7, Lemma 5] to be the identity function and \( e^{-\gamma} \), respectively, and the process \( V^n \) in that same lemma to be \( \int_0^{Q_2}(t) dt \), \( k \in \{1, 2\} \), and \( \int_0^{g(Q^{r''}(t)) dt} = \int_0^{v_1 Q_2(t) dt} + v_2 Q_2(t) dt \), respectively.

Finally, to establish (62), we first recall from (73) that
\[
\lim_{r'' \to \infty} \dot{J}(T') \geq \lim sup_{r'' \to \infty} \mathbb{E} \left[ \int_0^{\infty} e^{-\gamma} g(Q^{r''}(t)) dt \right].
\]
Next, given the weak convergence in (78), we can parallel the arguments to show (64) in the proof of part (i) to find that for any \( T > 0 \),
\[
\mathbb{E} \left[ \int_0^{T} e^{-\gamma} g(d\tilde{e}(t)) dt \right]
\leq \lim inf_{r'' \to \infty} \mathbb{E} \left[ \int_0^{\infty} e^{-\gamma} g(Q^{r''}(t)) dt \right]
\leq \lim sup_{r'' \to \infty} \mathbb{E} \left[ \int_0^{\infty} e^{-\gamma} g(Q^{r''}(t)) dt \right]
\leq \lim sup_{r'' \to \infty} \mathbb{E} \left[ \int_0^{T} e^{-\gamma} g(\tilde{Q}^{r''}(t)) dt \right]
+ \left( h_1 + c_1 \delta_1 \right) + \left( h_2 + c_2 \delta_2 \right) \left( \frac{\mu_2}{\mu_3} + 1 \right) \lim sup_{r'' \to \infty} \int_0^{T} e^{-\gamma} \mathbb{E} \left[ \sup_{\theta \in [r'' \to \infty]} \bar{W}^{r''}(s) \right] dt.
\]
From the weak convergence in (78) and the uniform integrability in Lemma 8.6,
\[
\limsup_{r'' \to \infty} \mathbb{E} \left[ \int_0^T e^{-\gamma_t} g(\dot{Q}^{r''}(t)) \, dt \right] = \mathbb{E} \left[ \int_0^T e^{-\gamma_t} g(d\tilde{\bar{Q}}(t)) \right].
\]
Since
\[
\lim_{T \to \infty} \mathbb{E} \left[ \int_0^T e^{-\gamma_t} g(d\tilde{\bar{Q}}(t)) \right] = \mathbb{E} \left[ \int_0^\infty e^{-\gamma_t} g(d\tilde{\bar{Q}}(t)) \right]
\]
by the monotone convergence theorem and
\[
\lim_{T \to \infty} \limsup_{r'' \to \infty} \int_0^T e^{-\gamma_t} \mathbb{E} \left[ \sup_{0 \leq s \leq t} \dot{W}^{r''}(s) \right] ds = 0
\]
by (74) in Lemma 8.6, we conclude from taking the limit as \( T \to \infty \) in (80) that
\[
\lim_{r'' \to \infty} \mathbb{E} \left[ \int_0^\infty e^{-\gamma_t} g(\dot{Q}^{r''}(t)) \, dt \right] = \mathbb{E} \left[ \int_0^\infty e^{-\gamma_t} g(d\tilde{\bar{Q}}(t)) \right]. \quad (81)
\]
To complete the proof, it is enough to establish that \((\tilde{\bar{Q}}, \tilde{L})\) is an admissible control for the WCP. This is because Theorem 5.1 then establishes that
\[
\mathbb{E} \left[ \int_0^\infty e^{-\gamma_t} g(d\tilde{\bar{Q}}(t)) \right] \geq \mathbb{E} \left[ \int_0^\infty e^{-\gamma_t} q^*(\dot{W}^*(t)) \, dt \right], \quad (82)
\]
and the sequence of inequalities (79), (81), and (82) establishes (62). The control \((\tilde{\bar{Q}}, \tilde{L})\) is admissible for the WCP because (29) follows from (77) when we let \( q_k(t, \dot{W}(t)) = d\tilde{\bar{Q}}_k(t) \) for \( k = 1, 2 \). To see (30), note that (72) implies
\[
\int_0^t \dot{Q}^{r''}_1(s) + \frac{\mu_3}{\mu_3} \dot{Q}^{r''}_2(s) \, ds = \int_0^t \dot{W}^{r''}(s) \, ds, \quad t \geq 0.
\]
Then, taking the limit as \( r'' \to \infty \),
\[
\int_0^t d\tilde{\bar{Q}}_1(s) + \frac{\mu_3}{\mu_3} d\tilde{\bar{Q}}_2(s) = \int_0^t \dot{W}(s) \, ds,
\]
from the weak convergence in (77), (76), and the continuous mapping theorem. Taking the derivative on both sides of the above equation then implies
\[
d\tilde{\bar{Q}}_1(t) + \frac{\mu_3}{\mu_3} d\tilde{\bar{Q}}_2(t) = \dot{W}(t).
\]

The properties of the Skorokhod map show that (31) and (32) hold, and that \( \dot{W} \) and \( \tilde{L} \) are continuous (since Lemma 8.7 implies \((\tilde{\xi}, \tilde{\xi}_1, \tilde{\xi}_2)\) are continuous).

When \( \{\mathcal{F}_t\} \) is the filtration generated by \((\tilde{\xi}, \tilde{\xi}_1, \tilde{\xi}_2)\), it follows that \((\dot{W}, \tilde{L})\) is \( \mathcal{F}_t \)-adapted. Furthermore, since any admissible scheduling control has \( \mathcal{F}_t^{r''} \)-independent of \( \mathcal{F}_t^{r''} \) for each \( r'' \) and for every \( t > 0 \) that \( \mathcal{F}_t \) is independent of \( \sigma(\xi_{t+u} - \tilde{\xi}_t; u > 0) \), and \( \tilde{\xi} \) is a \( \mathcal{F}_t \)-Brownian motion. The argument to show this is similar to the one in the last paragraph of the proof of Lemma 6 in Atar et al. [7], and so is omitted. The drift of \( \tilde{\xi} \) is 0 and its variance \( \sigma^2 \) defined in (28) for the same reasons as in the argument in the last paragraph of Lemma 7.3, since Lemma 8.5 gives the convergence \( T^{r''} \to T^* \) as \( r'' \to \infty \).

9. Asymptotic optimality of the proposed policies in the remaining parameter regimes. In §9.1, we outline the changes to the proof of Theorem 8.1 in the case that class 1 is the most expensive class and \( b^* = \infty \) in the solution to the Brownian control problem in §5; that is, in the case that
\[
(h_2 + c_2 \delta_2) \mu_3 < (h_1 + c_1 \delta_1) \mu_2 \quad \text{and} \quad \frac{\delta_1 + \gamma}{\delta_2 + \gamma} < \frac{\mu_3 (h_1 + c_1 \delta_1)}{\mu_3 (h_2 + c_2 \delta_2)}. \quad (83)
\]
In §9.2, we state the proposed policy and outline the proof changes in the case that class 2 is the most expensive class and \( b^* = \infty \); that is, in the case that
\[
(h_2 + c_2 \delta_2) \mu_3 > (h_1 + c_1 \delta_1) \mu_2 \quad \text{and} \quad \frac{\delta_1 + \gamma}{\delta_2 + \gamma} < \frac{\mu_3 (h_1 + c_1 \delta_1)}{\mu_3 (h_2 + c_2 \delta_2)}. \quad (84)
\]
Finally, in §9.3, we state the proposed policy and outline the proof changes in the case that class 2 is the most expensive class and \( b_2^* < \infty \); that is, in the case that

\[
(h_2 + c_2 \delta_2)\mu_1 > (h_1 + c_1 \delta_1)\mu_2 \quad \text{and} \quad \frac{\delta_1 + \gamma}{\delta_2 + \gamma} > \frac{\mu_2 (h_1 + c_1 \delta_1)}{\mu_1 (h_2 + c_2 \delta_2)}. \tag{85}
\]

The proof of Theorem 8.1 requires establishing (61) and (62). In all the cases (83)–(85), the argument that (62) holds can be adapted to complete the proof. In particular, Lemmas 8.1 and 8.3 remain valid, as well as the arguments following the Lemma statements.

9.1. The case that class 1 is the more expensive, and \( b^* = \infty \). In this subsection, we assume that condition (83) holds. Then, the proposed policy is the \((M', L')\) threshold policy that has \( M' = \infty \) for each \( r \geq 1 \).

Recall that the solution to the workload control problem has

\[
q^*(w) = \left( 0, \frac{\mu_2}{\mu_1} w \right) \quad \text{for all } w \geq 0.
\]

The key weak convergence result to establish is the following.

**Theorem 9.1.** Consider the sequence of \( N \)-systems indexed by \( r \), where the \( r \)th system operates under the \((\infty, L')\) threshold policy given in Definition 6.1. Then,

\[
(\hat{Q}_r^*, \hat{Q}_r^*, \mu_1 \hat{L}_r^* + \mu_2 \hat{L}_2^*) \Rightarrow (q^*(\hat{W}^*), \hat{L}^*), \quad \text{as } r \to \infty.
\]

Note that the weak convergence result in Theorem 9.1 is stronger than the weak convergence result in Theorem 7.1. In particular, it follows from the continuous mapping theorem that for any \( t > 0 \),

\[
\int_0^t e^{-\gamma s} \hat{Q}_r^*(s) \, ds \Rightarrow 0 = \int_0^t e^{-\gamma s} q_1^*(\hat{W}^*(s)) \, ds
\]

and

\[
\int_0^t e^{-\gamma s} \hat{Q}_r^*(s) \, ds \Rightarrow \int_0^t e^{-\gamma s} \hat{W}^*(s) \, ds = \int_0^t e^{-\gamma s} q_2^*(\hat{W}^*(s)) \, ds,
\]

as \( r \to \infty \).

**Outline of the Proof of Theorem 9.1.** The first step is to show that

\[
(\hat{Q}_r^*, \hat{L}_r^*) \Rightarrow (0, 0) \quad \text{as } r \to \infty.
\]

For this, first note that an argument very similar to the proof of (49) in Bell and Williams [8, Theorem 7.1], which establishes that the diffusion-scaled queue 1 queue-length process in an \( N \)-system in which there is no reneging \( (\delta_1 = \delta_2 = 0) \) weakly converges to zero, shows that \( \hat{Q}_r^* \Rightarrow 0 \) as \( r \to \infty \). Intuitively, this is because we expect that there are more jobs in the system without reneging. The next step is to observe that the same arguments used to establish Lemma 7.2 can also be used to establish that \( \hat{L}_r^* \Rightarrow 0 \) as \( r \to \infty \).

The second step is to show that \( \hat{X}^r \Rightarrow \hat{\xi}, \quad \text{as } r \to \infty \).

This follows as in the proof of Lemma 7.3, by first showing that

\[
\hat{T}^r \to \hat{T}^* \quad \text{u.o.c., a.s., as } r \to \infty.
\]

The final step is to show that

\[
(\hat{Q}_r^*, \mu_1 \hat{L}_r^* + \mu_2 \hat{L}_2^*) \Rightarrow \left( \frac{\mu_2}{\mu_1}, \hat{W}^*, \hat{L}^* \right) \quad \text{as } r \to \infty. \tag{86}
\]

To establish (86), note that

\[
\hat{W}^*(t) = \hat{X}^r(t) - \int_0^t \left( (r^2 \delta_1) \hat{Q}_r^*(s) - \frac{1}{r} \frac{1}{1} \hat{I}_r^{x^*}(r^2 s > 0) - \frac{1}{r} \frac{1}{1} \hat{I}_r^{y^*}(r^2 s > 0) \right) \, ds
\]

and

\[
\frac{\mu_2}{\mu_1} \int_0^t \left( (r^2 \delta_2) \hat{Q}_r^*(s) - \frac{1}{r} \frac{1}{1} \hat{I}_r^{x^*}(r^2 s > 0) \right) \, ds + \mu_1 \hat{L}_r^*(t) + \mu_2 \hat{L}_2^*(t).
\]
which, since \( \hat{W}^r = \hat{Q}_t^r + (\mu_2^r / \mu_1^r) \hat{Q}_t^2 \), implies that

\[
\hat{Q}_t^r(t) = \frac{\mu_2^r}{\mu_1^r} \hat{X}^r(t) - \frac{\mu_2^r}{\mu_1^r} \int_0^t \left( (r^2 \delta_1^r) \hat{Q}_s^r(s) - \frac{1}{r} \mathbf{1}[T_{i_1}^r(s) > 0] - \frac{1}{r} \mathbf{1}[T_{i_2}^r(s) > 0] \right) ds \\
- \int_0^t \left( (r^2 \delta_2^r) \hat{Q}_s^r(s) - \frac{1}{r} \mathbf{1}[T_{i_1}^r(s) > 0] - \frac{1}{r} \mathbf{1}[T_{i_2}^r(s) > 0] \right) ds + \frac{\mu_2^r}{\mu_1^r} \hat{Q}_t^1(t) - \frac{\mu_2^r}{\mu_1^r} \hat{Q}_t^2(t).
\]

Since \( \hat{Q}_t^r \) is continuous, nondecreasing, and can only increase when \( \hat{Q}_t^r \) is zero, it follows that \( (\hat{Q}_t^r, \mu_1^r \hat{I}_t^1 + \mu_2^r \hat{I}_t^2) \Rightarrow (q^*(\hat{W}^r), \hat{L}^*) \) as \( r \to \infty \).

Note that the proof of Theorem 9.2 does not require showing that \( \hat{I}_t^r \Rightarrow 0 \) as \( r \to \infty \), nor establishing any results that are helpful in proving a result similar to Lemma 8.2.

**Outline of the Proof of Theorem 9.2.** The first step is to show that

\[ \hat{Q}_t^r \Rightarrow 0 \quad \text{as} \quad r \to \infty. \]

For this, it is sufficient to show that for any \( t > 0 \)

\[ P(\hat{Q}_t^r(s) \geq 2N^r - 1 \quad \text{for some} \quad s \in [0, r^2 t]) \to 0, \]

where \( N^r = o(r) \). This can be shown using arguments similar to those used to establish (A2), except that it is not necessary to define up excursion intervals for the workload.

The next step is to establish that

\[ \tilde{T}^{r*,} \Rightarrow \tilde{T}^* \quad \text{as} \quad r \to \infty, \]

from which it follows that

\[ \hat{X}^r \Rightarrow \hat{X}^* \quad \text{as} \quad r \to \infty. \]

Note that

\[
r^{-1} \hat{Q}_t^1(t) = r^{-1} \hat{X}^r(t) - r^{-1} \frac{\mu_2^r}{\mu_1^r} \hat{Q}_t^2(t) - r^{-1} \frac{\mu_2^r}{\mu_1^r} \int_0^t \left( (r^2 \delta_1^r) \hat{Q}_s^r(s) - \frac{1}{r} \mathbf{1}[T_{i_1}^r(s) > 0] - \frac{1}{r} \mathbf{1}[T_{i_2}^r(s) > 0] \right) ds \\
- r^{-1} \int_0^t \left( (r^2 \delta_2^r) \hat{Q}_s^r(s) - \frac{1}{r} \mathbf{1}[T_{i_1}^r(s) > 0] - \frac{1}{r} \mathbf{1}[T_{i_2}^r(s) > 0] \right) ds + r^{-1} (\mu_1^r \hat{I}_t^1(t) + \mu_2^r \hat{I}_t^2(t)).
\]
It is necessary to generalize the invariance principle for semimartingale reflecting Brownian motion in Williams [37, Theorem 4.1] in one-dimension to accommodate a linear drift term so that it can be used in a setting with reneging. Then, under the proposed policy, \( \mu_1 \hat{T}_1 + \mu_2 \hat{T}_2 \) can only increase when \( \hat{Q}_x \leq 1/r \), it follows that \( (r^{-1} \hat{Q}_x, r^{-1}(\mu_1 \hat{T}_1 + \mu_2 \hat{T}_2)) \) is a solution to a linear generalization of the perturbed Skorokhod problem for

\[
r^{-1} \hat{X}_x(t) = r^{-1} \frac{\mu_2}{\mu_3} \hat{Q}_x(t) - r^{-1} \frac{\mu_2}{\mu_3} \int_0^t \left( (r^2 \delta_t') \hat{Q}_x(s) - \frac{1}{r} 1(\hat{T}_1^{r^{*}}(r^2 s) > 0) \right) ds
\]

\[
- r^{-1} \int_0^t \left( (r^2 \delta_t') \hat{Q}_x(s) - \frac{1}{r} 1(\hat{T}_1^{r^{*}}(r^2 s) > 0) - \frac{1}{r} 1(\hat{T}_2^{r^{*}}(r^2 s) > 0) \right) ds.
\]

It is straightforward to show that \( r^{-1} \hat{X}_x \to 0 \) as \( r \to \infty \). Since also \( r^{-1} \hat{Q}_x \to 0 \) as \( r \to \infty \), it follows that

\[
(r^{-1} \hat{Q}_x, r^{-1}(\mu_1 \hat{T}_1 + \mu_2 \hat{T}_2)) \to (0, 0) \quad \text{as} \quad r \to \infty.
\]

Since \( \hat{T}_1^{r^{*}} \to 0 \) as \( r \to \infty \), it follows that

\[
\tilde{T}_1^r \to \tilde{T}^r \quad \text{as} \quad r \to \infty.
\]

Then, to conclude that (87) is valid, it is sufficient to establish that

\[
\frac{1}{r} \tilde{Y}_3^r \to 0 \quad \text{as} \quad r \to \infty.
\]

(89)

This is because

\[
\frac{1}{r} \tilde{Y}_3^r(t) = \left( \frac{\lambda_2}{\mu_3} t - \tilde{T}_3^r(t) \right) \quad \text{as} \quad r \to \infty,
\]

and so it follows from (89) that

\[
\tilde{T}_1^r \to \tilde{T}^r \quad \text{as} \quad r \to \infty.
\]

Then, since

\[
\frac{1}{r} \tilde{T}_2^r = t - \tilde{T}_2^r - \tilde{T}_3^r,
\]

and \( (1/r) \tilde{T}_2^r \to 0 \) as \( r \to \infty \), it follows that

\[
\tilde{T}_2^r \to t \left( 1 - \frac{\lambda_2}{\mu_3} \right) = t \left( \frac{\lambda_1}{\mu_2} - \frac{\lambda_2}{\mu_3} \right) = \tilde{T}^r_2(t).
\]

To see that (89) holds, note that

\[
r^{-1} \hat{Q}_x(t) = r^{-1} \hat{X}_x(t) - r^{-1} \int_0^t \left( (r^2 \delta_t') \hat{Q}_x(s) - r^{-1} \frac{1}{r} 1(\hat{T}_1^{r^{*}}(r^2 s) > 0) \right) ds + r^{-1} \mu_2 \hat{Y}_3^r(t),
\]

and that \( \hat{Y}_3^r \) can only increase when \( \hat{Q}_x \) is zero. Hence \( (r^{-1} \hat{Q}_x, \mu_2 r^{-1} \hat{Y}_3^r) \) solves the one-dimensional linearly generalized Skorokhod problem for

\[
r^{-1} \hat{X}_x(t) - r^{-1} \int_0^t \left( (r^2 \delta_t') \hat{Q}_x(s) - r^{-1} \frac{1}{r} 1(\hat{T}_1^{r^{*}}(r^2 s) > 0) \right) ds.
\]

Since \( r^{-1} \hat{X}_x \to 0 \) and \( r^{-1} \hat{Q}_x \to 0 \) as \( r \to \infty \), it follows from the continuous mapping theorem and the continuity of the mapping associated with the one-dimensional linearly generalized Skorokhod problem that \( r^{-1} Y_3^r \to 0 \) as \( r \to \infty \).

The final step is to multiply (88) by \( r \) to recover the diffusion-scaled class 1 queue-length process, and to use the aforementioned generalization of the invariance principle for semimartingale reflecting Brownian motions in one-dimension to conclude

\[
(\hat{Q}_x, \mu_2 \hat{Y}_3 + \mu_2 \hat{L}_2) \Rightarrow (\hat{W}^r, \hat{L}^r) \quad \text{as} \quad r \to \infty.
\]
9.3. The case that class 2 is the more expensive and \( b_2^* < \infty \). In this subsection, we assume that condition (85) holds. Then, the proposed policy is the \((M', L')\) threshold policy given in Definition 6.2, and has \( b_2^* \) finite. Recall that the solution to the workload control problem has

\[
q^*(w) = \begin{cases} 
(w, 0) & \text{for all } w \leq b_2^*, \\
0, \left(\frac{\mu_1}{\mu_2}w\right) & \text{for all } w > b_2^*.
\end{cases}
\]

The key weak convergence result to establish is the following.

**Theorem 9.3.** Consider the sequence of \( N \)-systems indexed by \( r \), where the \( r \)th system operates under the threshold policy \( T^r, * \) described in Definition 6.2. Then, for any \( \eta > 0 \),

\[
\int_0^t e^{-\gamma s} \bar{Q}_1^r(s) ds \Rightarrow \int_0^t e^{-\gamma s} \bar{W}^r(s) 1[\bar{W}^r(s) \leq b_2^*] ds = \int_0^t e^{-\gamma s} q_1^r(\bar{W}^r(s)) ds
\]

and

\[
\int_0^t e^{-\gamma s} \bar{Q}_2^r(s) ds \Rightarrow \int_0^t e^{-\gamma s} \frac{\mu_1}{\mu_2} \bar{W}^r(s) 1[\bar{W}^r(s) > b_2^*] ds = \int_0^t e^{-\gamma s} q_2^r(\bar{W}^r(s)) ds,
\]

as \( r \to \infty \).

Note that since the proposed policy is a threshold policy, as explained in Remark 6.5, it appears difficult to have a “strong” weak convergence result like that in §§9.1 and 9.2, and instead we have a “weak” weak convergence result like that in Theorem 7.1.

**Outline of the Proof of Theorem 9.3.** The proof outline is identical to the proof of Theorem 7.1. That is, the first step is to prove the following proposition, that replaces Proposition 7.1.

**Proposition 9.1.** Consider the sequence of \( N \)-systems indexed by \( r \), where the \( r \)th system operates under the threshold policy \( T^r, * \) described in Definition 6.2. Then, for any \( \eta > 0 \),

\[
\bar{Q}_1^r \{ \bar{W}^r > b_2^* + \eta \} \Rightarrow 0 \quad \text{and} \quad \bar{Q}_1^r \{ \bar{W}^r \leq b_2^* - \eta \} \Rightarrow 0 \quad \text{as } r \to \infty.
\]

The second step is to prove the following proposition, that replaces Proposition 7.2.

**Proposition 9.2.** Consider the sequence of \( N \)-systems indexed by \( r \), where the \( r \)th system operates under the threshold policy \( T^r, * \) in Definition 6.2. Then,

\[
(\bar{W}^r, \bar{X}^r, \bar{V}_1^r, \bar{V}_2^r, \hat{V}_1^r, \hat{I}_1^r, \hat{I}_2^r) \Rightarrow (\bar{W}^*, \hat{\xi}, 0, 0, 0, \hat{V}, 0, \bar{L}^*) \quad \text{as } r \to \infty,
\]

where

\[
\hat{V}(i) = \int_0^t \delta_1 \bar{W}^r(s) 1[\bar{W}^r(s) \leq b_2^*] + \delta_2 \bar{W}^r(s) 1[\bar{W}^r(s) > b_2^*] ds.
\]

The proof of Proposition 9.1 is similar to the proof of Proposition 7.1. Lemma 7.1 can be used in the proof unchanged. The proof of Proposition 9.2 is similar to the proof of Proposition 7.2. The statements of Lemmas 7.2, 7.3, and 7.4 are unchanged, and their proofs are similar. \(\square\)

**Acknowledgments.** The authors thank Ruth Williams for outlining the weak convergence argument. The authors also thank Itai Gurvirch for further help with the weak convergence argument, Kavita Ramanan for helpful discussion concerning one-dimensional regulator mappings having a discontinuity, and Ananda Weerasinghe for helping show the existence and uniqueness of a solution to the Hamilton-Jacobi-Bellman equations in §5. Finally, the authors thank the anonymous referees for their extraordinary effort, which very much helped to improve the paper.

**Appendix.** The appendix is divided into two sections. Section A provides the details omitted from the proof of Proposition 7.1 that establish \( \bar{Q}_1^r \{ \bar{W}^r < b - \eta \} \Rightarrow 0 \) as \( r \to \infty \). The proofs of all lemmas in the manuscript can be found in §B.
Finally, let
\[ \tilde{Q}'_1(\tilde{W}' > b + \eta) \Rightarrow 0 \] as \( r \to \infty, \]
it remains to establish (52). Then, to complete the proof of Proposition 7.1, it remains to establish that
\[ \tilde{Q}'_1(\tilde{W}' < b - \eta) \Rightarrow 0 \] as \( r \to \infty. \]

We first show (52), and then show (A1).

The argument that (52) holds. Let \( N' \) be \( o(r) \), so that \( \lim_{r \to \infty} N'/r = 0 \), as \( r \to \infty \). For \( r \) large enough, \( r \infty > N' \), and so it is sufficient to show that
\[ \sum_{n=1}^{m'} P\left( Q'_2(s) \geq \left( 1 + \frac{\lambda'_2 + 1}{\lambda'_2} \right) N' \right) \to 0, \quad \text{as } r \to \infty. \] (A2)

In words, it is enough to show that in the \( n \)th up excursion interval for the workload process, once \( Q'_2 \) reaches 0, it stays small (that is, it stays below a level of order \( N' \)) for a high probability. For this, it is helpful to construct up excursion intervals from the level \( N' \), and to show that the number of class 2 customers in the system does not become too large during any of those up excursion intervals.

The following definitions are useful. For each \( n \in \{1, 2, \ldots, m'\} \), let
\[ \tau'_{n, 0} := \beta'_n, \]
\[ \tau'_{n, 2m-1} := \inf\{ s \geq \tau'_{n, 2m-2} : Q'_2(s) \geq N' \}, \]
\[ \tau'_{n, 2m} := \inf\{ s \geq \tau'_{n, 2m-1} : Q'_2(s) < N' \}, \]
for \( m \in \{1, 2, \ldots\} \). Then, \( \{\tau'_{n, 2m-1}, \tau'_{n, 2m}\} \) is the \( m \)th up excursion interval for \( Q'_2 \) within the \( n \)th up excursion interval for the workload process, starting from time \( \beta'_n \). We make use of the shifted processes
\[ A^{r', n, m}_2(s) := A'_2(\tau'_{n, 2m-1} + s) - A'_2(\tau'_{n, 2m-1}), \]
\[ S^{r', n, m}_3(s) := S'_3(\tau'_{n, 2m-1} + s) - S'_3(\tau'_{n, 2m-1}). \]

Next, we define a good set. First let
\[ s'_2 := \frac{N'}{\mu'_2 - \lambda'_2 - 2\tilde{e}_2}, \]
and assume \( 2\tilde{e}_2 < \frac{1}{2}(\mu'_2 - \lambda'_2), \tilde{e}_2 < 1, \) and \( r \) large enough so that \( \mu'_2 - \lambda'_2 - 2\tilde{e}_2 > \frac{1}{2}(\mu'_2 - \lambda'_2) \), \( \lambda'_2 + \tilde{e}_2 < \lambda'_2 + 1 \).

Define
\[ \eta^{r', n, m}_2 := \left\{ A^{r', n, m}_2(s'_2) < (\lambda'_2 + \tilde{e})s'_2, S^{r', n, m}_3(s'_2) > (\mu'_2 - \tilde{e})s'_2 \right\}. \]

Finally, let
\[ l' := (\lambda'_2 + \tilde{e})r^2t + 1. \]

The first step in establishing (A2) is to note that
\[
P\left( Q'_2(s) \geq \left( 1 + \frac{\lambda'_2 + 1}{\lambda'_2} \right) N' \right) \text{ for some } s \in [\beta'_n, r^2t \wedge \kappa'_n] \]
\[ \leq P(\tau'_{n, 2m-1} \leq r^2t) + P\left( \left( \bigcap_{m=1}^{r'} \eta^{r', n, m}_2 \right)^c \right) \]
\[ + P(\tau'_{n, 2r'-1} > r^2t, \left( \bigcap_{m=1}^{r'} \eta^{r', n, m}_2 \right), Q'_2(s) \geq \left( 1 + \frac{\lambda'_2 + 1}{\lambda'_2} \right) N' \text{ for some } s \in [\tau'_{n, 2m-1}, \tau'_{n, 2m}] \]
\[ \leq P(\tau'_{n, 2r'-1} \leq r^2t) + P\left( \left( \bigcap_{m=1}^{r'} \eta^{r', n, m}_2 \right)^c \right), \]
\[ \sum_{m=1}^{r'} P(\tau'_{n, 2m-1} > r^2t, \eta^{r', n, m}_2, Q'_2(s) \geq \left( 1 + \frac{\lambda'_2 + 1}{\lambda'_2} \right) N' \text{ for some } s \in [\tau'_{n, 2m-1}, \tau'_{n, 2m}] \). \]
Hence to establish (A2), we must show that

(i) \( m' P(\tau_{n,2r-1}^r \leq r^2 t) \to 0 \) as \( r \to \infty \);

(ii) \( m' P \left( \left( \bigcap_{m=1}^{l'} \eta_{2,m}^r \right) \right) \to 0 \) as \( r \to \infty \);

(iii) \( m' \sum_{m=1}^{l'} P \left( \tau_{n,2r-1}^r > r^2 t, \eta_{2,m}^r \right) \geq \left( 1 + \frac{\lambda_2 + 1}{2(\mu_3 - \lambda_2)} \right) N_r^r \) for some \( s \in [\tau_{n,2m-1}^r, \tau_{n,2m}^r] \to 0 \), as \( r \to \infty \).

The convergence in (i) follows because

\[ P(\tau_{n,2r-1}^r \leq r^2 t) \leq P(\Lambda_2^r(r^2 t) > (\lambda_2^r + \bar{\epsilon})r^2 t), \]

because every up excursion requires at least one class 2 arrival, and there exists a finite constant \( d > 0 \) such that

\[ P(\Lambda_2^r(r^2 t) > (\lambda_2^r + \bar{\epsilon})r^2 t) \leq \frac{d}{(r^2 t)^{1+\epsilon_1}}, \]

for \( \epsilon_1 > 0 \), by (50). The convergence in (ii) can be established using an argument similar to the one used to prove Lemma 7.1. (The key change is that Ata and Kumar [1, Lemma 9] must be used to establish a tighter bound than that in (B12)–(B14), which is straightforward because Assumption 3.3 implies that the \( \epsilon_1 \) that appears in (B15) can be as large as needed, provided it is finite.) We conclude that to establish (A2), it remains only to show (iii).

To show (iii), we first argue that on \( \eta_{2,m}^r \cap [\tau_{n,2m-1}^r < \infty], \tau_{n,2m-1}^r + s_2^m \geq \tau_{n,2m}^r \). Suppose not. Then, for any \( s \in [0, s_2^m] \), \( \tau_{n,2m-1}^r + s = \tau_{n,2m-1}^r + s \) for \( \tau_{n,2m-1}^r + s < \tau_{n,2m}^r \), and so it follows that

\[ Q_2^r(\tau_{n,2m-1}^r + s_2^m) \leq Q_2^r(\tau_{n,2m-1}^r) + A_2^r \left( \eta_{2,m}^r \right) - S_2^r \eta_{2,m}^r \]

\[ \leq N_r^r - (\mu_2^r - \lambda_2^r - 2\bar{\epsilon}_2)s_2^m \]

\[ = 0. \]

This contradicts the definition of \( \tau_{n,2m}^r \). Hence \( \tau_{n,2m-1}^r + s_2^m \geq \tau_{n,2m}^r \). Then, to see that (iii) holds, it is enough to note that for \( s < s_2^m \)

\[ Q_2^r(\tau_{n,2m-1}^r + s) \leq N_r^r + A_2^r \left( \eta_{2,m}^r \right) \leq N_r^r + (\lambda_2^r + \bar{\epsilon}_2)s_2^m, \]

and that \( \bar{\epsilon}_2 \) was chosen so that \( \lambda_2^r + \bar{\epsilon}_2 < \lambda_2^r \) and \( \mu_2 - \lambda_2^r > \frac{1}{2}(\mu_3 - \lambda_2) \), so that

\[ N_r^r + (\lambda_2^r + \bar{\epsilon}_2)s_2^m = N_r^r + (\lambda_2^r + \bar{\epsilon}_2) \cdot \frac{N_r^r}{\mu_2 - \lambda_2^r - 2\bar{\epsilon}_2} < \left( 1 + \frac{\lambda_2^r + 1}{2(\mu_3 - \lambda_2)} \right) N_r^r. \]

We conclude that for each \( n \in \{1, 2, \ldots, m'\} \) and each \( m \in \{1, 2, \ldots, l'\} \),

\[ P \left( \tau_{n,2r-1}^r > r^2 t, \eta_{2,m}^r \right) \geq \left( 1 + \frac{\lambda_2^r + 1}{2(\mu_3 - \lambda_2)} \right) N_r^r \] for some \( s \in [\tau_{n,2m-1}^r, \tau_{n,2m}^r] = 0, \)

which is sufficient to establish (iii).

\textbf{The argument to establish (A1).} The general outline for this argument is the same as the one that shows \( \hat{Q}_1^r \{ W_r^r > b + \eta \} \to 0 \) as \( r \to \infty \); however, there are changes in the details. Specifically, we define “down” excursion intervals for the workload process (rather than up excursion intervals), and then define shifted processes based on the times at which down excursions begin. Let

\[ \kappa_n^r := 0, \]

\[ \alpha_n^r := \inf \{ s \geq \kappa_n^r : W_r^r(s) \leq rb - 1 \}, \]

\[ \beta_n^r := \inf \{ s \geq \alpha_n^r : Q_1^r(s) \leq L' \}, \]

\[ \gamma_n^r := \inf \{ s \geq \alpha_n^r : W^r(s) \leq r(b - \eta) \}, \]

\[ \kappa_n^r := \inf \{ s \geq \alpha_n^r : W^r(s) \geq rb \}, \]

\[ s' := \frac{rb - 1}{\mu_2^r - \lambda_1^r - 3\bar{\epsilon}}. \]
Intuitively, (A5) follows because after time $t > 801$ where $Ghamami and Ward: Dynamic Scheduling of a Two-Server Parallel Server System$
801

where $\bar{\epsilon} > 0$ is small enough so that $\mu'_{1} + \mu'_{2} - \lambda'_{i} - 3\bar{\epsilon} > 0$ for all large enough $r$. The shifted processes $A_{n-k}^{r}, S_{k}^{r},$ and $R_{k}^{n}, k \in \{1, 2\}, j \in \{1, 2, 3\},$ and $n \in \{1, 2, 3, \ldots\},$ and the good set $\eta^{r,n}, n \in \{1, 2, 3, \ldots\},$ are defined as in the second paragraph of §7.1 (that proves Proposition 7.1). Note that in order to emphasize the similarity in proof concept, we have abused notation, and redefined $k_{n}', \alpha_{n}', \beta_{n}', \gamma_{n}', s', A_{n-k}^{r}, S_{k}^{r},$ and $R_{k}^{n}, k \in \{1, 2\}, j \in \{1, 2, 3\},$ and $n \in \{1, 2, 3, \ldots\}.$

Fix $t > 0.$ Let $\epsilon > 0$ be arbitrarily small. To show (A1), it is sufficient to show that

$$P(Q'_{i}(s)1|W'(s) < r(b - \eta)) \geq \epsilon r$$

for some $s \in [0, r^{2}t] \rightarrow 0$, as $r \rightarrow \infty.$ (A3)

For $m'$ defined as in the second paragraph of §7.1, it follows that

$$P(Q'_{i}(s)1|W'(s) < r(b - \eta)) \geq \epsilon r$$

for some $s \in [0, r^{2}t]$

$$\leq P(\alpha_{m'} > r^{2}t) + P\left(\left(\bigcap_{n=1}^{m'} \eta^{r,n}\right)^{c}\right)$$

$$+ P\left(\alpha_{m'} > r^{2}t, \left(\bigcap_{n=1}^{m'} \eta^{r,n}\right), Q'_{i}(s) \geq \epsilon r$$

for some $s \in \bigcup_{n=1}^{m'} [\gamma_{n}', \kappa_{n}'] \cap [0, r^{2}t].$

The argument that the first term on the right-hand side in the above expression converges to zero is exactly as in the fourth paragraph of §7.1. Also, Lemma 7.1 establishes that

$$P\left(\left(\bigcap_{n=1}^{m'} \eta^{r,n}\right)^{c}\right) \rightarrow 0,$$

as $r \rightarrow \infty.$

Hence to show (A3), it is sufficient to establish that

$$P\left(\alpha_{m'} > r^{2}t, \left(\bigcap_{n=1}^{m'} \eta^{r,n}\right), Q'_{i}(s) \geq \epsilon r$$

for some $s \in \bigcup_{n=1}^{m'} [\gamma_{n}', \kappa_{n}'] \cap [0, r^{2}t] \rightarrow 0,$ as $r \rightarrow \infty.$

For this, first note that

$$P(\alpha_{m'} > r^{2}t, \left(\bigcap_{n=1}^{m'} \eta^{r,n}\right), Q'_{i}(s) \geq \epsilon r$$

for some $s \in \bigcup_{n=1}^{m'} [\gamma_{n}', \kappa_{n}'] \cap [0, r^{2}t]$$

$$\leq \sum_{n=1}^{m'} P(\alpha_{m'} > r^{2}t, \eta^{r,n}, Q'_{i}(s) \geq \epsilon r$$

for some $s \in [\gamma_{n}', \kappa_{n}'] \cap [0, r^{2}t].$

Assume we can show that

$$P(\alpha_{m'} > r^{2}t, \eta^{r,n}, Q'_{i}(s) \geq \epsilon r$$

for some $s \in [\gamma_{n}', \kappa_{n}'] \cap [0, r^{2}t]$$

$$\leq P\left(\alpha_{m'} > r^{2}t, \eta^{r,n}, \beta_{n}^{r-1} < \kappa_{n}', s', A_{n-k}^{r}, S_{k}^{r}, \beta_{n}^{r-1} \leq \gamma_{n}'\right).$$

(A4)

Then, it follows that to complete the proof it is sufficient to show that

$$\sum_{n=1}^{m'} P\left(\alpha_{m'} > r^{2}t, \eta^{r,n}, \beta_{n}^{r-1} < \kappa_{n}', s', A_{n-k}^{r}, S_{k}^{r}, \beta_{n}^{r-1} \leq \gamma_{n}'\right) \rightarrow 0$$

as $r \rightarrow \infty.$ (A5)

Intuitively, (A5) follows because after time $\beta_{n}^{r-1},$ the number of jobs in buffer 1 follows the evolution dynamics of an underloaded $GI/GI/1$ queue. In particular, (A5) can be established using arguments similar to those used to show (52) in §7.1. We omit these details.
The argument to establish (A4). First, we argue by contradiction that on \( \eta_n^\ast \cap \{ \alpha_n < \infty \} \), \( \beta_n^\ast > \alpha_n + s' \). Suppose not; that is, suppose that on \( \eta_n^\ast \), \( \beta_n^\ast > \alpha_n + s' \). Then, since server 2 is processing jobs from buffer 1 during the interval \( [\alpha_n', \alpha_n + s'] \), for any \( s \in [\alpha_n', \alpha_n + s'] \),

\[
Q'_1(\alpha_n' + s') \leq Q'_1(\alpha_n') + A'_1(n)(s) - S'_1(n)(s) - S'_2(n)(s) < rb - 1 + (\lambda'_1 - \mu'_1 - \mu'_2 + 3\bar{\varepsilon})s' = 0,
\]

where the last equality follows directly from the definition of \( s' \). This is a contradiction. Hence we conclude that on \( \eta_n^\ast \), \( \beta_n^\ast \leq \alpha_n + s' \).

Next, we show that on \( \eta_n^\ast \cap \{ \alpha_n < \infty \} \), \( \beta_n^\ast \leq \gamma_n \). For this, it suffices to show that on \( \eta_n^\ast \), for any \( s \in [0, s'] \) such that \( \alpha_n + s \leq \gamma_n \), it is also true that \( W(\alpha_n' + s') > rb - r\eta \). In words, in the \( n \)th workload down-excursion interval and on the good set \( \eta_n^\ast \), when queue 1 reaches the buffer threshold, \( L' \), the workload is still above the level \( rb - r\eta \). For any \( s \in [0, s'] \), on \( \eta_n^\ast \),

\[
W'(\alpha_n' + s) \geq rb - 2 - \frac{\mu'_1}{\mu_1} + (A'_1(n)(s) - S'_1(n)(s) - S'_2(n)(s) - R'_1(n)(s)) + \frac{\mu'_1}{\mu_2} (A'_2(n)(s) - R'_2(n)(s)) \geq rb - 2 - \frac{\mu'_1}{\mu_3} + (\lambda'_1 + \lambda'_2 - \mu'_1 - \mu'_2) s' - 6\bar{\varepsilon}s'.
\]

Since

\[
\lambda'_1 - \mu'_1 - \mu'_2 + \frac{\mu'_1}{\mu_1} \lambda'_2 = \mu'_2 \left( \frac{\lambda'_1 - \mu'_1}{\mu_2} - \frac{\lambda_1 - \mu_1}{\mu_2} + \frac{\lambda'_2}{\mu_3} - \frac{\lambda_2}{\mu_3} \right),
\]

it follows from Assumption 3.1 that for large enough \( r \)

\[
s'\mu'_2 \left( \frac{\lambda'_1 - \mu'_1}{\mu_2} - \frac{\lambda_1 - \mu_1}{\mu_2} + \frac{\lambda'_2}{\mu_3} - \frac{\lambda_2}{\mu_3} \right) < 1.
\]

Hence for all \( s \in [0, s'] \), provided \( \bar{\varepsilon} > 0 \) is also chosen small enough so that \( 6\bar{\varepsilon}s' < r\eta \),

\[
W'(\alpha_n' + s) > rb - r\eta,
\]

for large enough \( r \). Hence (A4) holds. \( \Box \)

**B. Proofs of lemmas.**

**Proof of Lemma 5.1.** It is sufficient to solve (40), because solving (40) provides a solution to the Bellman equation (37). To do this, we first consider an arbitrary \( b > 0 \) and solve

\[
\frac{\sigma^2}{2} V''_2(w) - \delta_2 w V'_2(w) - \gamma V'_2(w) + \frac{\mu_3}{\mu_2} (h_2 + c_2 \delta_2) w = 0 \quad \text{for all } w \in [0, b] \quad (B1)
\]

subject to \( V'_2(0) = 0 \) and \( V'_2(b) = \frac{(h_1 + c_1 \delta_1) - (\mu_1/\mu_2)(h_2 + c_2 \delta_2)}{\delta_1 - \delta_2} \),

and

\[
\frac{\sigma^2}{2} V''_1(w) - \delta_1 w V'_1(w) - \gamma V'_1(w) + (h_1 + c_1 \delta_1) w = 0 \quad \text{for all } w \geq b \quad (B2)
\]

subject to \( V'_1(b) = \frac{(h_1 + c_1 \delta_1) - (\mu_2/\mu_1)(h_2 + c_2 \delta_2)}{\delta_1 - \delta_2} \)

\( V'_1 \) is bounded and increasing on \( [b, \infty) \).

We then show there exists \( b^* \) such that

\[
V'_1(b^*) = V_2(b^*) \quad \text{and} \quad V''_1(b^*) = V''_2(b^*) \quad \text{and} \quad V'_2 \text{ is increasing on } [0, b^*].
\]

Together (B1)–(B3) provide a solution to (40).
The solutions to (B1) and (B2) are expressed in terms of Kummer’s function

\[ M(a, b; z) := 1 + \frac{az}{b} + \frac{(a)_2 z^2}{(b)_2 2!} + \cdots + \frac{(a)_n z^n}{(b)_n n!} + \cdots , \]

where \( (a)_0 = 1 \) and

\[ (a)_n = a(a + 1)(a + 2) \cdots (a + n - 1), \quad n = 1, 2, \ldots . \]

Specifically,

\[ V_2(w) := d_1(b)M\left(\frac{\gamma}{2\delta_2}, \frac{1}{2}; \frac{\delta_2}{\sigma^2}w^2\right) + d_2M\left(\frac{\gamma}{2\delta_2}, \frac{1}{2}; \frac{\delta_2}{\sigma^2}w^2\right) \int_0^\infty \frac{\exp\left(\frac{\delta_2}{\sigma^2}y^2\right)}{M\left(\frac{\gamma}{2\delta_2}, \frac{1}{2}; \frac{\delta_2}{\sigma^2}y^2\right)} dy + \frac{\mu_1}{\mu_2} \left( h_2 + c_2 \delta_2 \right) \frac{w}{\delta_2 + \gamma} , \]

where

\[ d_1(b) := \frac{(h_1 + c_1 \delta_1) - (h_1 + c_1 \delta_1) (h_2 + c_2 \delta_2)}{\delta_1 - \delta_2} - \frac{(h_1 + c_1 \delta_1) (h_2 + c_2 \delta_2)}{\delta_1 + \gamma} \]

\[ + \frac{\sigma^2}{\gamma} \frac{1 - \frac{\mu_1}{\mu_2} (h_2 + c_2 \delta_2)}{\delta_2 + \gamma} \frac{b M\left(\frac{\gamma}{2\delta_2}, \frac{1}{2}; \frac{\delta_2}{\sigma^2}b^2\right)}{\exp\left(\frac{\delta_2}{\sigma^2}b^2\right)} \]

\[ + \frac{\mu_1}{\mu_2} \left( h_2 + c_2 \delta_2 \right) \int_0^b \frac{\exp\left(\frac{\delta_2}{\sigma^2}y^2\right)}{M\left(\frac{\gamma}{2\delta_2}, \frac{1}{2}; \frac{\delta_2}{\sigma^2}y^2\right)} dy \]

\[ d_2 := -\frac{\mu_1}{\mu_2} \left( h_2 + c_2 \delta_2 \right) \frac{w}{\delta_2 + \gamma} \]

solves (B1), and

\[ V_1(w) := \frac{(h_1 + c_1 \delta_1) - (h_1 + c_1 \delta_1) (h_2 + c_2 \delta_2)}{\delta_1 - \delta_2} - \frac{h_1 + c_1 \delta_1}{\delta_1 + \gamma} M\left(\frac{\gamma}{2\delta_1}, \frac{1}{2}; \frac{\delta_1}{\sigma^2}w^2\right) \int_0^\infty \frac{\exp\left(\frac{\delta_1}{\sigma^2}y^2\right)}{M\left(\gamma/2\delta_1, 1/2; (\delta_1/\sigma^2)y^2\right)} dy \]

\[ + \frac{(h_1 + c_1 \delta_1)}{\delta_1 + \gamma} w \]

solves (B2).

To see this, as in Ward and Kumar [34], use Polyanin and Zaitsev [28, Equation 103, p. 143] to find one solution of the homogeneous equation having the form \((\sigma^2/2)V'(w) - \delta wV(w) - \gamma V(w) = 0\). Then, apply formula (2) of Polyanin and Zaitsev [28, p. 129] to obtain the general solution to the homogeneous equation. Next, use the method of undetermined coefficients to find the general solution to the nonhomogeneous equations in (B1) and (B2). Then, use Slater [31, relation 13.4.9]

\[ M'(a, b; z) = \frac{a}{b} M(a + 1, b + 1; z) \]  
(B4)

to find

\[ V'_2(w) := d_1(b) \frac{2 \gamma}{\sigma^2} w M\left(\frac{\gamma}{2\delta_2}, \frac{1}{2}; \frac{\delta_2}{\sigma^2}w^2\right) \]

\[ + d_2 \left( 2 \frac{2 \gamma}{\sigma^2} w M\left(\frac{\gamma}{2\delta_2}, \frac{1}{2}; \frac{\delta_2}{\sigma^2}w^2\right) \int_0^w \frac{\exp\left(\frac{\delta_2}{\sigma^2}y^2\right)}{M\left(\frac{\gamma}{2\delta_2}, \frac{1}{2}; \frac{\delta_2}{\sigma^2}y^2\right)} dy + \frac{\exp\left(\frac{\delta_2}{\sigma^2}w^2\right)}{M\left(\frac{\gamma}{2\delta_2}, \frac{1}{2}; \frac{\delta_2}{\sigma^2}w^2\right)} \right) \]

\[ + \frac{\mu_1}{\mu_2} \left( h_2 + c_2 \delta_2 \right) \frac{w}{\delta_2 + \gamma} . \]
and

\[ V'_1(w) = \left( \frac{(h_1+c_2\delta_2)}{\delta_1-\delta_2} \right) - \frac{(h_1+c_1\delta_1)}{\delta_1+\gamma} \left( 2 \frac{\gamma}{\sqrt{\pi}} w M(\frac{\gamma}{2\delta_1}+1, \frac{3}{2}, \frac{\delta_1}{\delta_2}) \right) \]

that shows that the conditions \( V'_2(0) = 0 \) and

\[ V'_1(b) = \frac{(h_1+c_2\delta_2) - (\mu_2/\mu_1)(h_2+c_2\delta_2)}{(\delta_1-\delta_2)} \]

are satisfied. The relation in Slater [31, §13.1.4]

\[ M(a, b; z) = \frac{\Gamma(b)}{\Gamma(a)} e^{z-b} (1 + o(|z|^{-1})), \quad (B5) \]

as \( |z| \to \infty \) can be used to show that \( V'_1 \) is bounded. Finally, the following claim, which capitalizes in Weerasinghe and Mandelbaum [35, Lemma 3.2], shows that \( V'_1 \) is increasing.

**Claim B.1.** For all \( w \geq b \), \( V''_1(w) > 0 \).

In summary, we have provided \( V_1 \) and \( V_2 \) that solve (B1)–(B2). To establish that this \( V_1 \) and \( V_2 \) satisfy (40), and so solve the Bellman equation (38), it remains to show that (B3) holds. This is done in the following two claims.

**Claim B.2.** For any \( b \) under which \( V'_1(b) = V''_1(b) \), we have \( V''_1(w) > 0 \) for all \( w \in [0, b] \).

**Claim B.3.** There exists \( b^* \) under which \( V'_1(b^*) = V'_2(b^*) \) and \( V''_1(b^*) = V''_2(b^*) \). Furthermore, for such a \( b^* \), \( V'_1 \) is increasing on \([0, b^*]\).

The proofs of Claims B.1–B.3 follow immediately. \( \square \)

**Proof of Claim B.1.** We first show how to write the solution to (B2) in terms of \( Q_\infty \) that satisfies

\[ \frac{\sigma^2}{2} Q'_\infty(w) - \delta_1 w Q'_\infty(w) - (\delta_1+\gamma) Q_\infty(w) = 0 \quad \text{for all } w < 0 \]

\[ Q_\infty(0) = 1 \quad \text{and} \quad \lim_{w \to -\infty} Q_\infty(w) = 0. \]

Let \( U_1(w) = V'_1(w) \) for all \( w \geq 0 \). Then, to solve (B2), it is enough to find a \( V_1 \) that satisfies

\[ \frac{\sigma^2}{2} U_1''(w) - \delta_1 w U_1'(w) - h_1 + c_1 \delta_1 = 0 \quad \text{for all } w \geq b \]

subject to \( U_1(b) = \frac{(h_1+c_1\delta_1) - (\mu_2/\mu_1)(h_2+c_2\delta_2)}{\delta_1-\delta_2} \)

\( U_1 \) is bounded and nondecreasing.

Such a \( U_1 \) can be written as

\[ U_1(w) = \frac{h_1+c_1\delta_1}{\delta_1+\gamma} + Q_0(w), \]

where

\[ Q_0(w) = Q_\infty(-w) \left[ \frac{(h_1+c_1\delta_1) - (\mu_2/\mu_1)(h_2+c_2\delta_2)}{\delta_1-\delta_2} - \frac{h_1+c_1\delta_1}{\delta_1+\gamma} \right] \]

solves the homogeneous ode

\[ \frac{\sigma^2}{2} Q'_0(w) - \delta_1 w Q'_0(w) - (\delta_1+\gamma) Q_0(w) = 0 \]

subject to \( Q_0(b) = \frac{(h_1+c_1\delta_1) - (\mu_2/\mu_1)(h_2+c_2\delta_2)}{\delta_1-\delta_2} - \frac{h_1+c_1\delta_1}{\delta_1+\gamma} \).
Since $U_i(w) = V'_i(w)$ for all $w \geq 0$, $U_i(w) = (h_1 + c_1 \delta_1)/(\delta_1 + \gamma) + Q_0(w)$, and $Q_0$ is expressed in terms of $Q_\infty$, the solution to (B2) can be written in terms of $Q_\infty$.

Then, to show $V''_i(w) > 0$ for all $w \geq b$, it is sufficient to show $Q'_0(w) > 0$ for all $w \geq b$. To see this, note that $Q'_\infty(-w) > 0$ for all $w \geq b$ by Weeraseinghe and Mandelbaum [35, Lemma 3.2] and

\[
\frac{(h_1 + c_1 \delta_1) - (\mu_3/\mu_2)(h_2 + c_2 \delta_2)}{\delta_1 - \delta_2} = \frac{h_1 + c_1 \delta_1}{\delta_1 + \gamma}
\]

because $(\delta_1 + \gamma)/((\delta_2 + \gamma) > (h_1 + c_1 \delta_1)/(\mu_3/\mu_2)(h_2 + c_2 \delta_2))$. □

**Proof of Claim B.2.** Suppose we can establish that $d_1(b) > 0$. Then, since $V_2$ satisfies (25),

\[
V_2''(0) = \gamma \frac{2}{\sigma^2} V_2(0) = \gamma \frac{2}{\sigma^2} d_1(b) > 0.
\]

Also, by Lemma 1, $V''_2(b) > 0$. Now, let $F(w) = V''_2(w)$ and note that $F$ must satisfy

\[
\frac{\sigma^2}{2} F''(w) - \delta_2 w F'(w) - (\gamma + 2 \delta_2) F(w) = 0 \quad \text{for all } w \in [0, b].
\]

(B6)

The argument that $F(w) > 0$ for all $w \in [0, b]$ is by contradiction. Let $F(c) := \min_{w \in [0, b]} F(w)$, and suppose that $F(c) \leq 0$. Since $F(0) > 0$ and $F(b) > 0$, it must be that $c \in (0, b)$. Then, $F'(c) = 0$ and $F''(c) > 0$ because $c$ is a minimum. However (B6) implies that

\[
\frac{\sigma^2}{2} F''(c) = (\gamma + 2 \delta_2) F(c) \leq 0,
\]

which is a contradiction. Hence, $F(c) \geq 0$, and we conclude $V''_2(w) \geq 0$ for all $w \in [0, b]$.

To complete the proof, it remains to show $d_1(b) > 0$. For this, set

\[
h(b) := \frac{\exp(\frac{\delta_1}{2} b^2)}{M(\frac{\gamma}{2 \delta_2}, \frac{1}{2}, \frac{3}{2}, \frac{\delta_2}{\sigma^2}, b^2)} + \frac{\gamma}{\sigma^2} b M\left(\frac{\gamma}{2 \delta_2} + 1, \frac{3}{2}, \frac{\delta_2}{\sigma^2}, b^2\right) \int_0^b \frac{\exp(\frac{\delta_1}{2} y^2)}{M(\frac{\gamma}{2 \delta_2}, \frac{1}{2}, \frac{3}{2}, \frac{\delta_2}{\sigma^2}, y^2)} dy.
\]

Then,

\[
d_1(b) := \frac{(h_1 + c_1 \delta_1) - (\mu_3/\mu_2)(h_2 + c_2 \delta_2)}{\delta_1 - \delta_2} - \frac{(\mu_3/\mu_2)(h_2 + c_2 \delta_2)}{\delta_2 + \gamma} + \frac{(\mu_3/\mu_2)(h_2 + c_2 \delta_2)}{\delta_2 + \gamma} h(0)
\]

\[
= \frac{h(0)}{\delta_1 - \delta_2} \left[1 + \frac{2}{3} \frac{\delta_1}{\sigma^2} b^2 \right] > 0.
\]

Furthermore, from (B4)

\[
h'(b) = \frac{2 \delta_2}{\sigma^2} b \exp\left(\frac{\delta_1}{2} b^2 \right) \frac{\exp(\frac{\delta_1}{2} b^2)}{M(\frac{\gamma}{2 \delta_2}, \frac{1}{2}, \frac{3}{2}, \frac{\delta_2}{\sigma^2}, b^2)} + \frac{\gamma}{\sigma^2} \left(1 + \frac{2}{3} \frac{\delta_1}{\sigma^2} b^2 \right) M\left(\frac{\gamma}{2 \delta_2} + 1, \frac{3}{2}, \frac{\delta_2}{\sigma^2}, b^2\right) \int_0^b \frac{\exp(\frac{\delta_1}{2} y^2)}{M(\frac{\gamma}{2 \delta_2}, \frac{1}{2}, \frac{3}{2}, \frac{\delta_2}{\sigma^2}, y^2)} dy > 0,
\]

which implies $d_1(b) > 0$ for all $b > 0$. □

**Proof of Claim B.3.** First note that by Claim B.2, it is sufficient to show there exists $b^*$ under which $V_1(b^*) = V_2(b^*)$ and $V''_1(b^*) = V''_2(b^*)$. 

\[\text{[Proof continues...]}\]
We next argue that it is sufficient to show there exists \( b^* \) under which \( V_1(b^*) = V_2(b^*) \). For this, note that since \( V_1 \) and \( V_2 \) solve (B1) and (B2), for any \( b \) under which \( V_1(b) = V_2(b) \),

\[
\frac{\sigma^2}{2} V_2''(b) = b(\delta_2 V_1(b) - \frac{\mu_1}{\mu_2} (h_2 + c_1 \delta_2)) - \gamma V_2(b)
\]

\[
= b \left( \frac{\delta_1 (h_1 + c_1 \delta_1) - (\mu_3/\mu_2) (h_2 + c_2 \delta_2)}{\delta_1 - \delta_2} - \frac{\mu_1}{\mu_2} (h_2 + c_2 \delta_2) \right) - \gamma V_2(b)
\]

\[
= b \left( \frac{\delta_1 (h_1 + c_1 \delta_1) - (\mu_3/\mu_2) (h_2 + c_2 \delta_2)}{\delta_1 - \delta_2} - (h_1 + c_1 \delta_1) \right) - \gamma V_1(b)
\]

\[
= \frac{\sigma^2}{2} V_1''(b).
\]

Finally, to complete the proof, we show that there exists \( b \) under which \( V_1(b) = V_2(b) \). To do this, we show that for small enough \( b \), \( V_1(b) > V_2(b) \), and for large enough \( b \), \( V_1(b) < V_2(b) \). It follows by evaluating the explicit expressions for \( V_1 \) and \( V_2 \) at \( b \) that

\[
V_2(b) = \frac{\sigma^2}{2} \frac{1}{\gamma b} \frac{M \left( \frac{\delta_1^2}{2 \delta_1^2}, \frac{1}{\delta_1^2} \right)}{M \left( \frac{\delta_1^2}{2 \delta_1^2}, \frac{1}{\delta_1^2} b^2 \right)} \times \left[ \frac{\delta_1 \mu_1 (h_2 + c_2 \delta_2)}{\delta_1 + \beta} \exp \left( \frac{\delta_1^2}{2 \delta_1^2} b^2 \right) - \frac{\delta_1 \mu_1 (h_2 + c_2 \delta_2)}{\delta_1 + \beta} \exp \left( \frac{\delta_1^2}{2 \delta_1^2} \right) \right]
\]

\[
V_1(b) = \frac{\sigma^2}{2} \frac{1}{\gamma b} \frac{M \left( \frac{\delta_1^2}{2 \delta_1^2}, \frac{1}{\delta_1^2} b^2 \right)}{M \left( \frac{\delta_1^2}{2 \delta_1^2}, \frac{1}{\delta_1^2} \right)} \exp \left( \frac{\delta_1^2}{2 \delta_1^2} \right) \left[ \frac{\delta_1 \mu_1 (h_2 + c_2 \delta_2)}{\delta_1 + \beta} \exp \left( \frac{\delta_1^2}{2 \delta_1^2} b^2 \right) - \frac{\delta_1 \mu_1 (h_2 + c_2 \delta_2)}{\delta_1 + \beta} \exp \left( \frac{\delta_1^2}{2 \delta_1^2} \right) \right]
\]

Since

\[
\frac{(\mu_3/\mu_2)(h_2 + c_2 \delta_2)}{\delta_1 + \beta} - \frac{\delta_1 (\mu_3/\mu_2)(h_2 + c_2 \delta_2) - \delta_2 (h_1 + c_1 \delta_1) + \gamma \left[ (\mu_3/\mu_2)(h_2 + c_2 \delta_2) - (h_1 + c_1 \delta_1) \right]}{\delta_1 - \delta_2 (\delta_1 + \beta)}
\]

\[
= \frac{(h_1 + c_1 \delta_1) - (\mu_3/\mu_2)(h_2 + c_2 \delta_2)}{\delta_1 - \delta_2 (\delta_1 + \beta)}
\]

\[
< 0,
\]

it follows that \( V_2(b) \to -\infty \) as \( b \downarrow 0 \). Since also \( \lim_{b \uparrow 0} V_1(b) \) is finite, it follows that for small \( b \),

\[
V_1(b) > V_2(b).
\]

Next, it follows from (B5) that

\[
V_1(b) - V_2(b) \to \frac{h_1 + c_1 \delta_1}{\delta_1 + \beta} \quad \text{and} \quad V_1(b) - V_2(b) \to \frac{(\mu_3/\mu_2)(h_2 + c_2 \delta_2)}{\delta_1 + \beta}, \quad \text{as} \quad b \to \infty.
\]

It follows from the assumption \((\delta_1 + \beta)/(\delta_2 + \gamma) = (\mu_3/\mu_2)(h_2 + c_2 \delta_2)\) that

\[
\frac{\delta_1 \mu_1 (h_2 + c_2 \delta_2)}{\mu_2} - \frac{\delta_2 (h_1 + c_1 \delta_1) + \gamma \left[ (\mu_3/\mu_2)(h_2 + c_2 \delta_2) - (h_1 + c_1 \delta_1) \right]}{\delta_1 - \delta_2} > 0,
\]

and it follows from the assumption \((h_1 + c_1 \delta_1)\mu_2 \geq (h_2 + c_2 \delta_2)\mu_3\) that

\[
\frac{\mu_2 (h_1 + c_1 \delta_1)}{\mu_3 (h_2 + c_2 \delta_2)} \geq 1, \quad \text{so that} \quad \delta_1 > \delta_2.
\]
Hence
\[
\frac{(\mu_3/\mu_2)(h_2 + c_2 \delta_2)}{\delta_2 + \gamma} - \frac{h_1 + c_1 \delta_1}{\delta_1 + \gamma} = \frac{\delta_1(\mu_3/\mu_2)(h_2 + c_2 \delta_2) - \delta_2(h_1 + c_1 \delta_1) + \gamma[(\mu_3/\mu_2)(h_2 + c_2 \delta_2) - (h_1 + c_1 \delta_1)]}{(\delta_1 - \delta_2)(\delta_2 + \gamma)} > 0,
\]
and so for large \( b \),
\[
V_i(b) < V_2(b).
\]
Therefore, since \( V_i \) and \( V_2 \) are continuous, there must exists \( b^* \) such that \( V_i(b^*) = V_2(b^*) \).

**Proof of Lemma 5.2.** It is sufficient to show that there exists a unique solution to (42). We construct a solution to (42) in terms of \( Q_\infty \) that satisfies
\[
\frac{\sigma^2}{2} Q''_\infty(w) - \delta_2 w Q'_\infty(w) - (\delta_2 + \gamma) Q_\infty(w) = 0 \quad \text{for all } w < 0,
\]
\[
Q_\infty(0) = 1 \quad \text{and} \quad \lim_{w \to -\infty} Q_\infty(w) = 0.
\]
Lemma 3.2 in Weerasinghe and Mandelbaum [35] guarantees that \( Q_\infty \) exists and is unique. We can also conclude from Lemma 3.2 in Weerasinghe and Mandelbaum [35] that \( 0 \leq Q_\infty(w) \leq 1 \) and \( Q_\infty(w) > 0 \) for all \( w < 0 \).

Let \( U(w) := V'(w) \) for all \( w \geq 0 \). Then, to solve (42) it is enough to find a \( U \) that satisfies
\[
\frac{\sigma^2}{2} U''(w) - \delta_2 w U'(w) - (\delta_2 + \gamma) U(w) + \frac{\mu_3}{\mu_2} (h_2 + c_2 \delta_2) = 0 \quad \text{for all } w \geq 0
\]
subject to \( U \) is twice continuously differentiable;
\( U \) is bounded, has \( U(0) = 0 \),
and \( (\delta_1 - \delta_2) U(0) = (h_1 + c_1 \delta_1) - \frac{\mu_3}{\mu_2} (h_2 + c_2 \delta_2) \) for all \( w \geq 0 \).

Such a \( U \) can be written as
\[
U(w) = \frac{(\mu_3/\mu_2)(h_2 + c_2 \delta_2)}{\delta_2 + \gamma} + Q_0(w),
\]
where \( Q_0 \) is a solution to the homogeneous ode,
\[
\frac{\sigma^2}{2} Q'_0(w) - \delta_2 w Q'_0(w) - (\delta_2 + \gamma) Q_0(w) = 0,
\]
and has
\[
Q_0(0) = -\frac{(\mu_3/\mu_2)(h_2 + c_2 \delta_2)}{\delta_2 + \gamma}.
\]
Note that
\[
Q_0(w) = -Q_\infty(-w) \frac{(\mu_3/\mu_2)(h_2 + c_2 \delta_2)}{\delta_2 + \gamma}
\]
is a bounded solution to (B7)–(B8). Then, \( U \) is bounded and has \( U(0) = 0 \). Hence, to complete the proof, it remains only to show that
\[
(\delta_1 - \delta_2) U(w) \leq (h_1 + c_1 \delta_1) - \frac{\mu_3}{\mu_2} (h_2 + c_2 \delta_2) \quad \text{for all } w \geq 0.
\]

To show (B9), we require the following additional properties of \( U \). First, since
\[
Q'_0(w) = Q_\infty(-w) \frac{(\mu_3/\mu_2)(h_2 + c_2 \delta_2)}{\delta_2 + \gamma},
\]
and \( Q_\infty(-w) > 0 \) for all \( w \geq 0 \), it follows that \( U(w) \) is increasing in \( w \). Then, also, \( U(w) \geq 0 \) for all \( w \geq 0 \).

We are now ready to establish (B9). That the assumption
\[
(h_1 + c_1 \delta_1) \mu_2 \geq (h_2 + c_2 \delta_2) \mu_3 \quad \text{and} \quad \frac{\delta_1 + \gamma}{\delta_2 + \gamma} \leq \frac{\mu_3(h_1 + c_1 \delta_1)}{\mu_3(h_2 + c_2 \delta_2)}
\]
is satisfied, it is either the case that
\[
\delta_1 \leq \delta_2, \quad \text{so that } \frac{\delta_1 + \gamma}{\delta_2 + \gamma} \leq 1 \leq \frac{h_1 + c_1 \delta_1}{(\mu_3/\mu_2)(h_2 + c_2 \delta_2)}. \tag{B10}
\]
or that
\[
\delta_1 > \delta_2 \quad \text{so that } 1 < \frac{\delta_1 + \gamma}{\delta_2 + \gamma} \leq \frac{h_1 + c_1 \delta_1}{(\mu_3/\mu_2)(h_2 + c_2 \delta_2)}. \tag{B11}
\]
In the case that (B10) holds, it is trivial that (B9) holds, because \( U(w) \geq 0 \) for all \( w \geq 0 \), so that \((\delta_1 - \delta_2)U(w) \leq 0 \). Otherwise, in the case that (B11) holds, it is enough to show
\[
\lim_{w \to \infty} U(w) < \frac{(h_1 + c_1 \delta_1) - (\mu_3/\mu_2)(h_2 + c_2 \delta_2)}{\delta_1 - \delta_2}.
\]
This is because \( U \) is increasing and \( \lim_{w \to \infty} Q_0(w) = 0 \), which implies
\[
\lim_{w \to \infty} U(w) = \frac{(\mu_3/\mu_2)(h_2 + c_2 \delta_2)}{\delta_2 + \gamma},
\]
and
\[
\frac{(h_1 + c_1 \delta_1) - (\mu_3/\mu_2)(h_2 + c_2 \delta_2)}{\delta_1 - \delta_2} - \frac{(\mu_3/\mu_2)(h_2 + c_2 \delta_2)}{\delta_2 + \gamma}
= \frac{\delta_2(h_1 + c_1 \delta_1) - \delta_1(\mu_3/\mu_2)(h_2 + c_2 \delta_2) + \gamma[(h_1 + c_1 \delta_1) - (\mu_3/\mu_2)(h_2 + c_2 \delta_2)]}{(\delta_1 - \delta_2)(\delta_2 + \gamma)}
\geq 0,
\]
where the last inequality follows because the fact that \((\delta_1 + \gamma)/(\delta_2 + \gamma) \leq (h_1 + c_1 \delta_1)/((\mu_3/\mu_2)(h_2 + c_2 \delta_2))\) implies that
\[
\delta_2(h_1 + c_1 \delta_1) - \delta_1 \frac{\mu_3}{\mu_2}(h_2 + c_2 \delta_2) + \gamma \left[ (h_1 + c_1 \delta_1) - \frac{\mu_3}{\mu_2}(h_2 + c_2 \delta_2) \right] \geq 0. \tag*{\Box}
\]

**Proof of Lemma 5.3.** The process \( X^R \) satisfies the stochastic equation
\[
X^R(t) = X^R(0) + \tilde{\xi}(t) + L^R(t),
\]
where \( \tilde{\xi}(t) = y \cdot \tilde{\xi}(t) \) is the Brownian motion defined directly before (27), and \( L^R(0) = 0 \), \( L^R \) is nondecreasing, and \( \int_0^\infty X^R(t) dL^R(t) = 0 \). Note that we have coupled \( X^R \) and \( \tilde{W} \) by using the same Brownian motion \( \tilde{\xi} \). We establish that for these particular versions of the two processes, \( \tilde{W}^*(t) \leq X^R(t) \) for every \( t \geq 0 \) with probability 1.

The proof is by contradiction. Suppose there exists \( t > 0 \) for which \( \tilde{W}^*(t) > X^R(t) \). Since \( \tilde{W}^*(t) - X^R(t) \) has continuous sample paths, there exists \( s \in [0, t] \) such that \( \tilde{W}^*(s) = X^R(s) \) with \( \tilde{W}^*(v) > X^R(v) \) for all \( v \in (s, t] \). Because \( L^R \) is a nondecreasing process,
\[
\tilde{W}^*(t) - X^R(t) = \tilde{W}^*(s) - X^R(s) - \int_s^t f(q(\tilde{W}^*(v))) \, dv + (\tilde{L}^*(t) - \tilde{L}^*(s)) - (L^R(t) - L^R(s))
\leq \tilde{L}^*(t) - \tilde{L}^*(s).
\]
Since \( \tilde{W}^*(v) > X^R(v) \geq 0 \) for all \( v \in (s, t] \), \( \tilde{L}^*(t) - \tilde{L}^*(s) = 0 \). This implies that \( \tilde{W}^*(t) \leq X^R(t) \), which is a contradiction. \(\Box\)

**Proof of Lemma 7.1.** It follows from DeMorgan’s laws that
\[
P\left( \bigcap_{\eta=1}^{\eta'} \eta \right) = P\left( \bigcup_{\eta=1}^{\eta'} ((\eta')^C) \right).
\]
Next, note that
\[
P\left( \bigcup_{\eta=1}^{\eta'} ((\eta')^C) \right) \leq \sum_{\eta=1}^{\eta'} P((\eta')^C),
\]
and that
\[
P((\eta^r, n)^c) \leq \sum_{k=1}^{2} P(\|A_k^r, n(\cdot) - \lambda_k^r(\cdot)\|_\nu > \tilde{e} s') \\
+ \sum_{j=1}^{3} P(\|S_j^r, n(T_j^r, n(\cdot)) - \mu_j^r T_j^r, n(\cdot)\|_\nu > \tilde{e} s') \\
+ \sum_{k=1}^{2} P(\|R_k^r, n(\cdot)\|_\nu > \tilde{e} s').
\]

Since \( m' = O(r^2) \), it is enough to show that there exists a finite constant \( d > 0 \) such that for each \( n \in \{1, 2, \ldots\} \), \( k \in \{1, 2\} \), and \( j \in \{1, 2, 3\} \),
\[
P(\|A_k^r, n(\cdot) - \lambda_k^r(\cdot)\|_\nu > \tilde{e} s') \leq \frac{d}{r^7}, \quad \tag{B12}
\]
\[
P(\|S_j^r, n(T_j^r, n(\cdot)) - \mu_j^r T_j^r, n(\cdot)\|_\nu > \tilde{e} s') \leq \frac{d}{r^5}, \quad \tag{B13}
\]
\[
P(\|R_k^r, n(\cdot)\|_\nu > \tilde{e} s') \leq \frac{2d}{r^3}. \quad \tag{B14}
\]

We first establish (B12). For this, we require Ata and Kumar [1, Lemma 9, (103)]. We restate their result in our context for the reader’s convenience. Recall from the second paragraph in §3 that \( \{\tilde{u}_k(i), \tilde{u}_k(2), \ldots\} \), \( k \in \{1, 2\} \), are independent, i.i.d. sequences of mean 1 random variables, and \( \tilde{u}_k(i) = \tilde{u}_k(i)/\lambda_k^r \) for \( i \in \{1, 2, \ldots\} \).

**Ata and Kumar [1, (103)].** Let \( \epsilon_i > 0 \) be such that
\[
E[\tilde{u}_k(1)]^{2+2\epsilon_1} < \infty. \quad \tag{1}
\]
Fix \( r \). Given \( \epsilon > 0 \) and \( t > 2/\epsilon \),
\[
P\left( \sup_{0 \leq x \leq \tilde{s}, k \in \{1, 2\}} \max_{s \in [1, 2]} |A_k^r(s) - \lambda_k^r(s)| \geq \epsilon t \right) \leq \frac{C_2^r(\epsilon)}{t^{1+\epsilon_1}}, \quad \tag{B15}
\]
where
\[
C_2^r(\epsilon) = \frac{2 + 2\epsilon_1}{2\epsilon + 1} \left[ \frac{18(2 + 2\epsilon_1)^{3/2}}{(1 + 2\epsilon_1)^{1/2}} \right]^{2+2\epsilon_1} \\
\times \sum_{k=1}^{2} E \left[ \tilde{u}_k^r(1) - \frac{1}{\lambda_k^r} \right]^{2+2\epsilon_1} \left[ \left( \frac{4(\lambda_k^r)^2(\lambda_k^r + \epsilon_1)}{\epsilon^2} \right)^{1+\epsilon_1} + \left( \frac{4(\lambda_k^r)^3}{\epsilon^2} \right)^{1+\epsilon_1} \right].
\]

It follows from the properties of renewal processes that
\[
P(\|A_k^r, n(\cdot) - \lambda_k^r(\cdot)\|_\nu > \tilde{e} s') \leq P(1 + \|A_k^r(\cdot) - \lambda_k^r(\cdot)\|_\nu > \tilde{e} s').
\]

Also, letting \( \epsilon_i = 2 \) in (B15) shows that
\[
P(1 + \|A_k^r, n(\cdot) - \lambda_k^r(\cdot)\|_\nu > \tilde{e} s') = P\left( \|A_k^r, n(\cdot) - \lambda_k^r(\cdot)\|_\nu > \left( \tilde{e} - \frac{1}{s'} \right)s' \right) \leq \frac{C_2^r(\tilde{e} - 1/s')}{(s')^3}.
\]

Since \( s' \) is of order \( r \), and it follows from the expression for \( C_2^r(\epsilon) \) that \( C_2^r(\tilde{e} - 1/s') \) is of order 1, (B12) holds, assuming \( r \) is large enough so that \( \tilde{e} - 1/s' > 0 \).

We next establish (B13). Since for any \( s > 0 \), \( 0 \leq T_k^r, n(s) \leq s \), it is enough to show that
\[
P(\|S_j^r, n(T_j^r, n(\cdot)) - \mu_j^r(\cdot)\|_\nu > \tilde{e} s') \to 0, \quad \text{as } r \to \infty.
\]
This follows exactly as in the preceding paragraph.

---

1 It follows from Assumption 3.3 that (B15) holds for any finite \( \epsilon > 0 \), because all moments of the interarrival times are finite.
Finally, we show (B14), which completes the proof. First observe that for \( N_k \) a unit-rate Poisson process, since the system is empty at time 0 by Assumption 3.4,
\[
P(\| R^\prime_k (\cdot) \|_{\mathcal{V}} > \bar{\epsilon} s') \leq P(N_k(s' \delta_k^a A_k^{r_a}(s')) > \bar{\epsilon} s').
\]
Next note that
\[
P(N_k(s' \delta_k^a A_k^{r_a}(s')) > \bar{\epsilon} s')
\leq P(N_k(s' \delta_k^a A_k^{r_a}(s')) > \bar{\epsilon} s' | A_k^{r_a} \leq (\lambda_k + \bar{\epsilon}) s') + P(A_k^{r_a}(s') > (\lambda_k + \bar{\epsilon}) s').
\]
It follows from (B12) that
\[
P(A_k^{r_a}(s') > (\lambda_k + \bar{\epsilon}) s') \leq \frac{d}{r^5}.
\]
Since
\[
P(N_k(s' \delta_k^a A_k^{r_a}(s')) > \bar{\epsilon} s' | A_k^{r_a} \leq (\lambda_k + \bar{\epsilon}) s') \leq P(N_k((\lambda_k + \bar{\epsilon}) \delta_k^a(s')^2) > \bar{\epsilon} s'),
\]
it is sufficient to show that
\[
P(N_k((\lambda_k + \bar{\epsilon}) \delta_k^a(s')^2) > \bar{\epsilon} s') < \frac{d}{r^5}.
\]
For this, first note that
\[
P(N_k((\lambda_k + \bar{\epsilon}) \delta_k^a(s')^2) > \bar{\epsilon} s')
= P\left( N_k((\lambda_k + \bar{\epsilon}) \delta_k^a(s')^2) - (\lambda_k + \bar{\epsilon}) \delta_k^a(s')^2 > s' \left( \bar{\epsilon} - \frac{(\lambda_k + \bar{\epsilon}) \delta_k^a(s')^2}{s'} \right) \right)
\leq P\left( \| N_k(\cdot) - \|_\mathcal{V} > s' \left( \bar{\epsilon} - \frac{(\lambda_k + \bar{\epsilon}) \delta_k^a(s')^2}{s'} \right) \right),
\]
where the last inequality follows because \((\lambda_k + \bar{\epsilon}) \delta_k^a(s')^2\) converges to a constant as \( r \to \infty \). Then, observe that very similar argument to that used to prove (B12) can be used to show that
\[
P\left( \| N_k(\cdot) - \|_\mathcal{V} > s' \left( \bar{\epsilon} - \frac{(\lambda_k + \bar{\epsilon}) \delta_k^a(s')^2}{s'} \right) \right) < \frac{d}{r^5},
\]
because \(((\lambda_k + \bar{\epsilon}) \delta_k^a(s')^2)/s' \to 0 \) as \( r \to \infty \). \( \Box \)

**Proof of Lemma 7.2.** Fix \( t > 0 \). It is sufficient to show that for any \( \epsilon > 0 \),
\[
P(\| I_1^t \|_{\mathcal{V}^1} > r \epsilon) \to 0 \quad \text{as} \quad r \to \infty.
\]
Let
\[
\tau_0' := \inf\{ s \geq 0 : Q_1'(s) \geq L' \}.
\]
As in the last paragraph of the proof of Theorem 7.1 in Bell and Williams [8], the following inequality holds
\[
P(\| I_1^t \|_{\mathcal{V}^1} > r \epsilon) \leq P(I_1'(\tau_0') \geq r \epsilon) + P(I_1'(r^2 t) - I_1'(\tau_0') > 0).
\]
We establish the stronger result that there exist finite positive constants \( C_i, i \in \{1, 2, \ldots , 9\} \), such that for large enough \( r \)
\[
P(I_1'(\tau_0') \geq r \epsilon) \leq C_1 \exp(-C_2 L') + \left( C_3 \frac{L'}{r} \right)^{C_4 L'}, \tag{B16}
\]
and
\[
P(I_1'(r^2 t) - I_1'(\tau_0') > 0)
\leq (C_4 + C_5 r^2 t)^2(\exp(-C_6 L') + \exp(-C_7 r^2 t)) + C_8 r^2 t \left( C_9 \frac{L'}{r} \right)^{C_6 L'}, \tag{B17}
\]
because this is needed in the proof of Theorem 8.1.
The argument to establish (B16). Let $\tilde{\epsilon} > 0$ be such that $\tilde{\epsilon} < \min((\lambda_1 - \mu_1)/12, \lambda_i/2)$, and let $r_{\tilde{\epsilon}} \geq 1$ be such that for all $r \geq r_{\tilde{\epsilon}}$, $\lambda'_i - \mu'_i > (\lambda_1 - \mu_1)/2$ and $|\lambda_1 - \lambda'_i| < \tilde{\epsilon}$. Define

$$t' := \frac{8L'}{\lambda_1 - \mu_1},$$

$$\eta'_i := \{A'_i(t') \geq (\lambda'_i - \tilde{\epsilon})t', S'_i(t') \leq (\mu'_i + \tilde{\epsilon})t', N_i(\delta'_i L't') \leq \tilde{\epsilon}t'\}.$$

Then, for $r$ large enough so that $r \epsilon > t'$, since $I'_i(t) \leq t$ for all $t > 0$,

$$P(I'_i(t') \geq r \epsilon) \leq P(I'_i(t') \geq t') \leq P(\tau'_0 > t').$$

Also, since $Q'_i(t) = A'_i(t) - S'_i(t) - R'_i(t)$ and $R'_i(t) \leq N_i(\delta'_i L't')$ for all $t \in [0, \tau'_0)$, it follows that

$$P(\tau'_0 > t') \leq P(A'_i(t') - S'_i(t') - N_i(\delta'_i L't') \leq L').$$

On the set $\eta'_i$, since $\lambda'_i - \mu'_i > (\lambda_1 - \mu_1)/2$ and $\tilde{\epsilon} < (\lambda_1 - \mu_1)/12$,

$$A'_i(t') - S'_i(t') - N_i(\delta'_i L't') \geq (\lambda'_i - \mu'_i - 3\tilde{\epsilon})t' \geq \left(\frac{\lambda_1 - \mu_1}{4}\right)t' = 2L',$$

which implies

$$P(A'_i(t') - S'_i(t') - R'_i(t') \leq L'; \eta'_i) = 0.$$

Hence

$$P(A'_i(t') - S'_i(t') - N_i(\delta'_i L't') \leq L') \leq P((\eta'_i)^C).$$

Since

$$P((\eta'_i)^C) \leq P(A'_i(t') \leq (\lambda'_i - \tilde{\epsilon})t') + P(S'_i(t') > (\mu'_i + \tilde{\epsilon})t') + P(N_i(\delta'_i L't') > \tilde{\epsilon}t'),$$

it is sufficient to establish that there exists $C_1$, $C_2$, and $C_3$ such that for large enough $r$

$$P(A'_i(t') \leq (\lambda'_i - \tilde{\epsilon})t') \leq \frac{C_1}{2} \exp(-C_3L'), \quad (B18)$$

$$P(S'_i(t') > (\mu'_i + \tilde{\epsilon})t') \leq \frac{C_2}{2} \exp(-C_3L'), \quad (B19)$$

$$P(N_i(\delta'_i L't') \leq \tilde{\epsilon}t') \leq (C_3L'/r)^{C_4 t'} \quad (B20).$$

We first show that (B18) and (B19) hold. It follows from Assumption 3.3 and Appendix A in Bell and Williams [8] (in particular, see how Appendix A is applied to conclude (62) and (63) in Bell and Williams [8]) that for all large enough $r$

$$P(A'_i(t') \leq (\lambda'_i - \tilde{\epsilon})t') \leq \exp\left(-((\lambda'_i - \tilde{\epsilon})t' + 1)\Lambda_{L,t'}^{\lambda'_i - \tilde{\epsilon}}\left(\frac{1}{\lambda_1}\right)\right),$$

$$P(S'_i(t') > (\mu'_i + \tilde{\epsilon})t') \leq \exp\left(-((\mu'_i + \tilde{\epsilon})t' - 1)\Lambda_{L,t'}^{\mu'_i + \tilde{\epsilon}}\left(\frac{1}{\mu_1}\right)\right). \quad (B21)$$

Since $\tilde{\epsilon} < \lambda_1/2$, when also $r \geq r_{\tilde{\epsilon}}$, so that $|\lambda_1 - \lambda'_i| < \tilde{\epsilon}$, then the term $(\lambda'_i - \tilde{\epsilon})$ multiplying $t'$ in (B21) is positive. The inequalities in (B18) and (B19) follow from the definition of $t'$. To see the inequality in (B20), first note that since $\delta'_i L' = (r^2 \delta'_i L'/r)^{1/2}$ and $(r^2 \delta'_i L'/r) \rightarrow 0$ as $r \rightarrow \infty$, for large enough $r$, $\delta'_i L' < \tilde{\epsilon}/r$, so that

$$P(N_i(\delta'_i L't') > \tilde{\epsilon}t') \leq P\left(N_i\left(\frac{\tilde{\epsilon}}{r}t'\right) > \tilde{\epsilon}t'\right).$$

The probability on the right-hand side of the above expression can be calculated exactly, and we find that

$$P\left(N_i\left(\frac{\tilde{\epsilon}}{r}t'\right) > \tilde{\epsilon}t'\right) \leq \sum_{k=\lceil\tilde{\epsilon}t'\rceil-1}^{\infty} \exp\left(-\frac{\tilde{\epsilon}}{r}t'\right)\left(\frac{\tilde{\epsilon}t'}{r}\right)^k \frac{1}{k!} \leq \sum_{k=\lceil\tilde{\epsilon}t'\rceil-1}^{\infty} \left(\frac{\tilde{\epsilon}t'}{r}\right)^k \frac{1}{k!}.$$
It follows from the Lagrange form of the remainder term in Taylor’s theorem that there exists \( z \in [0, (\bar{e}/r)t] \) such that
\[
\sum_{k=\lceil r t \rceil - 1}^{\infty} \frac{(\bar{e}/r)t)^k}{k!} = \frac{\exp(z)}{\bar{e}t^r} \left( \frac{\bar{e}}{r} \right)^{\lceil r t \rceil}.
\]
Since \( \exp(z) \) is an increasing function, the previous two displayed equations imply that
\[
P\left( N_t \left( \frac{\bar{e}}{r} t \right) < \bar{e}t^r \right) \leq \frac{1}{\bar{e}t^r} \left( \frac{\bar{e}}{r} \right)^{\lceil r t \rceil} \exp \left( \frac{\bar{e}}{r} t^r \right).
\]
The inequality in (B20) follows by noting that for large enough \( r \)
\[
\exp \left( \frac{\bar{e}}{r} t^r \right) \leq 1,
\]
and that, if we let \( C_3 = 8\bar{e}/(\lambda_1 - \mu_1) \), then \( \bar{e}t^r = C_3 L' \).

**The argument to establish (B17).** We first note that if server 1 incurs some idle time in \([\tau_0', r^2 t]\), so that \( I_1'(r^2 t) - I_1'(\tau_0') > 0 \), then it must be that \( Q_1'(s) \leq 1 \) for some \( s \in [\tau_0, r^2 t]\). Hence
\[
P(I_1'(r^2 t) - I_1'(\tau_0') > 0) = P(Q_1'(s) \leq 1 \text{ for some } s \in [\tau_0', r^2 t]).
\]
We next make some useful definitions. Let
\[
\tau_1' := \inf \{ t > \tau_0': Q_1'(s) < L' \},
\]
\[
\tau_2' := \inf \{ t > \tau_1': Q_1'(s) \geq L' \},
\]
and, in general,
\[
\tau_{2n-1}' := \inf \{ t > \tau_{2n-2}': Q_1'(s) < L' \},
\]
\[
\tau_{2n}' := \inf \{ t > \tau_{2n-1}': Q_1'(s) \geq L' \}.
\]
Define the shifted processes
\[
A_{1i}^{\tau, n}(s) := A_{1i}^{\tau_2 - 1, s} - A_{1i}^{\tau_2 - 1},
\]
\[
S_{1i}^{\tau, n}(s) := S_{1i}^{\tau_1 - 1, s} - S_{1i}^{\tau_1 - 1},
\]
\[
R_{1i}^{\tau, n}(s) := R_{1i}^{\tau_2 - 1, s} - R_{1i}^{\tau_2 - 1}.
\]
Let \( \bar{c} < (\lambda_1 - \mu_1)/5, a < 1/(\mu_1 + 3\bar{c}) \), \( s' := aL' \), and
\[
\eta_{1i}^{n} := \{ A_{1i}^{\tau, n}(s') \geq (\lambda_1 - \bar{c})s', S_{1i}^{\tau, n}(s') \leq (\mu_1 + \bar{c})s', R_{1i}^{\tau, n}(s') \leq \bar{c}s' \}.
\]
It follows that for \( m' := [(\lambda_1 + \bar{c})r^2 t] \),
\[
P(Q_1'(s) \leq 1 \text{ for some } s \in [\tau_0', r^2 t])
\]
\[
\leq P(\tau_{2m' - 1} \leq r^2 t) + P \left( \left( \bigcap_{n=1}^{m'} \eta_{1i}^{n} \right) \right)
\]
\[
+ P \left( Q_1'(s) \leq 1 \text{ for some } s \in [\tau_0', r^2 t], \tau_{2m' - 1} > r^2 t, \bigcap_{n=1}^{m'} \eta_{1i}^{n} \right)
\]
\[
\leq P(\tau_{2m' - 1} \leq r^2 t) + P \left( \left( \bigcap_{n=1}^{m'} \eta_{1i}^{n} \right) \right),
\]
\[
+ \sum_{n=1}^{m'} P( Q_1'(s) \leq 1 \text{ for some } s \in [\tau_{2n-1}', \tau_{2n}'], \tau_{2m' - 1} > r^2 t, \eta_{1i}^{n} ).
\]
We consider each term in the above expression separately.
There must be at least one arrival every time the number of class 1 jobs in the system exceeds \( L' \), and so

\[
P(\tau_{2m'-1} \leq r^2 t) \leq P(A'_i(r^2 t) > m').
\]

Then, similar to (B21),

\[
P(A'_i(r^2 t) > m') \leq \exp \left( -((\lambda' + \bar{e})s') - 1)\Lambda^{u+}_i \left( \frac{1}{\lambda_i} \left( 1 + \bar{e}/(3\lambda_i) \right) \right) \right).
\]

(B23)

Next, observe that

\[
P\left( \bigcap_{n=1}^{m'} \eta_i'^{r,n} \right) = P\left( \bigcup_{n=1}^{m'} (\eta_i'^{r,n})^c \right)
\]

\[
\leq \sum_{n=1}^{m'} \left( P(A'_i(r^2 t) < (\lambda' - \bar{e})s') + P(S_i'(T_i'(\tau_{2n-1}')) + s) - S_i'(T_i'(\tau_{2n-1}')) > (\mu'_i + \bar{e})s') \right).
\]

Then, since

\[
P(S_i'(T_i'(\tau_{2n-1}')) > (\mu'_i + \bar{e})s') \leq P(S_i'(T_i'(\tau_{2n-1}')) + s) - S_i'(T_i'(\tau_{2n-1}')) > (\mu'_i + \bar{e})s'),
\]

similar to (84) in Bell and Williams [8],

\[
P(S_i'(T_i'(\tau_{2n-1}')) > (\mu'_i + \bar{e})s') \leq \exp \left( -((\mu'_i + \bar{e})s' - 1)\Lambda^{u+}_i \left( \frac{1}{\mu_i} \left( 1 + \bar{e}/(3\mu_i) \right) \right) \right).
\]

It follows similar to the bounds in (86) and (87) in Bell and Williams [8] that for any \( l_0 \in \mathbb{R} \) such that \( l_0 > 0 \)

\[
P(A'_i(r^2 t) < (\lambda'_i - \bar{e})s') \leq \exp \left( -((\lambda'_i - \bar{e})s')\Lambda^{u+}_i \left( \frac{1}{\lambda_i} \left( 1 + \bar{e}/(2\lambda_i) \right) \right) \right)
\]

\[+ \exp \left( - (\lambda_i r^2 t - 1)\Lambda^{u+}_i \left( \frac{1}{\lambda_i} \left( 1 + \bar{e}/(3\lambda_i) \right) \right) \right)
\]

\[+ \exp \left( 1 + \log((\lambda'_i + \bar{e})t) + 2 \log r - \frac{l_0\bar{e}s'}{2\lambda_i} + \Lambda^{u+}_i(l_0) \right).\]

Finally, for \( N_R \) a unit-rate Poisson process, since the number of class 1 jobs in the system does not exceed \( L' \) in \( [\tau_{2n-1}, \tau_{2n}] \), assuming that \( s' < \tau_{2n} - \tau_{2n-1} \) (which is shown below),

\[
P(R'_i(r^2 t) > \bar{e}s') \leq P(N_R(\delta'_i L' s') > \bar{e}s'),
\]

and, as in the argument to establish (B16),

\[
P(N_R(\delta'_i L' s') > \bar{e}s') \leq \left( \frac{\bar{e}}{r} \right)^{[\bar{e}s']}.
\]

Hence

\[
P\left( \bigcap_{n=1}^{m'} \eta_i'^{r,n} \right)^c \leq m' \left[ \exp \left( -((\mu'_i + \bar{e})s' - 1)\Lambda^{u+}_i \left( \frac{1}{\mu_i} \left( 1 + \bar{e}/(3\mu_i) \right) \right) \right) \right]
\]

\[+ \exp \left( - (\lambda_i r^2 t - 1)\Lambda^{u+}_i \left( \frac{1}{\lambda_i} \left( 1 + \bar{e}/(2\lambda_i) \right) \right) \right)
\]

\[+ \exp \left( 1 + \log((\lambda'_i + \bar{e})t) + 2 \log r - \frac{l_0\bar{e}s'}{2\lambda_i} + \Lambda^{u+}_i(l_0) \right) + \left( \frac{\bar{e}}{r} \right)^{[\bar{e}s']} \].

(B24)

We observe that if we can show that

\[
\sum_{n=1}^{m'} P(Q'_i(s) \leq 1 \text{ for some } s \in [\tau_{2n-1}', \tau_{2n}'], \tau_{2m'-1} > r^2 t, \eta_i'^{r,n}) = 0,
\]

(B25)
then, recalling that $s' = aL' = a[c \log r]$, it follows from (B22), (B23), and (B24), that there exists a finite constant $c_0$ such that for any $c > c_0$, there are finite positive constants $C_4 - C_9$ so that

$$P(I'_1(r^2t) - I'_0(\tau'_0) > 0) \leq (C_4 + C_9 r^2 t) \exp(-C_9 L') + \exp(-C_9 r^2 t) + C_9 r^2 t \left( C_9 \frac{L'}{r} \right)^{C_4 L'}.$$  

In other words, we will have established (B17). To show (B25), first note that

$$P(Q'_1(s) \leq 1 \quad \text{for some } s \in [\tau'_{2n-1}, \tau'_{2n}], \tau'_{2m'-1} > r^2 t, \eta'_4) \leq P(Q'_1(\tau'_{2n-1}) + A'_1 - \tau'_n(s) - S'_1 - \tau'_n(s) - R'_1 - \tau'_n(s) \leq 1 \quad \text{for some } s \in [0, \tau'_{2n} - \tau'_{2n-1}], \eta'_4).$$  

Next, we claim that on the set $V_r^n$,

$$\tau'_{2n} - \tau'_{2n-1} \leq s'.$$  

(B26)

To see this, observe that

$$\tau'_{2n} - \tau'_{2n-1} = \inf \{ t \geq 0 : Q'_1(\tau'_{2n-1}) + A'_1 - \tau'_n(t) - S'_1 - \tau'_n(t) - N'_r - \tau'_n(t) \geq L' \},$$

and that since $Q'_1(\tau'_{2n-1}) = L' - 1$, to show (B26), it is enough to show that on the set $\eta'_4$,

$$A'_1 - \tau'_n(s') - S'_1 - \tau'_n(s') - N'_r - \tau'_n(s') \geq 1.$$  

This follows because on the set $\eta'_4$,  

$$A'_1 - \tau'_n(s') - S'_1 - \tau'_n(s') - N'_r - \tau'_n(s') \geq (\lambda' - \bar{e})s' - (\mu'_1 - \bar{e})s' - \bar{e}s' > (\lambda_1 - \mu_1 - 5\bar{e})s',$$

and $(\lambda - \mu_1 - 5\bar{e}) > 0$, so that $(\lambda_1 - \mu_1 - 5\bar{e})s' \geq 1$ for large enough $r$. Since (B26) holds, it is sufficient to show that

$$\sum_{n=1}^{m'} P(Q'_1(\tau'_{2n-1}) + A'_1 - \tau'_n(s) - S'_1 - \tau'_n(s) - R'_1 - \tau'_n(s) \leq 1 \quad \text{for some } s \in [0, s'], \eta'_4) \to 0 \quad \text{as } r \to \infty.$$  

On the set $\eta'_4$,

$$Q'_1(\tau'_{2n-1}) + A'_1 - \tau'_n(s) - S'_1 - \tau'_n(s) - R'_1 - \tau'_n(s) \geq L' - 1 - (\mu'_1 + \bar{e})s' - \bar{e}s' \geq (1 - (\mu_1 + 3\bar{e})a)L' - 1.$$  

Since $a < 1/(\mu_1 + 3\bar{e})$, $1 - (\mu_1 + 3\bar{e})a > 0$, so that $(1 - (\mu_1 + 3\bar{e})a)L' - 1 > 1$ for large enough $r$. Hence

$$P(Q'_1(\tau'_{2n-1}) + A'_1 - \tau'_n(s) - S'_1 - \tau'_n(s) - R'_1 - \tau'_n(s) \leq 1 \quad \text{for some } s \in [0, s'], \eta'_4) = 0$$

for each $n \in \{1, \ldots, m'\}$. The proof is complete. \(\square\)

**Proof of Lemma 7.3.** It follows directly from Lemma 7.2 and the fact that for all $t \geq 0$, $\hat{T}'_1(t) + (1/r)\hat{I}'_1(t) = t$, that

$$\hat{T}'_1 \Rightarrow e \quad \text{and} \quad \frac{1}{r} \hat{I}'_1 \Rightarrow 0 \quad \text{as } r \to \infty.$$  

We next show that

$$\frac{1}{r} \hat{W}' \Rightarrow 0 \quad \text{as } r \to \infty.$$  

(B27)

From (17)--(21),

$$\frac{1}{r} \hat{W}'(t) = \frac{1}{r} \hat{X}'_1(t) + \frac{\mu'_1}{\mu'_2} \frac{1}{r} \hat{X}'_2(t) + \mu'_1 \frac{1}{r} \hat{I}'_1(t)$$

$$- \int_0^t (r^2 \delta'_1) \hat{Q}'_1(s) - \frac{1}{r^2} 1[\hat{T}'_1(r^2 s) > 0] - \frac{1}{r^2} 1[\hat{T}'_2(r^2 s) > 0]) ds$$

$$- \int_0^t \frac{\mu'_2}{\mu'_3} (r^2 \delta'_2) \hat{Q}'_2(s) - \frac{1}{r^2} 1[\hat{T}'_1(r^2 s) > 0]) ds$$

$$+ \mu'_3 \frac{1}{r} \hat{I}'_2(t), \quad \text{for all } t \geq 0.$$  

(B28)
Hence it remains to show that 1

Also, since $\bar{\delta}_t$, for all $t \geq 0$,

For this, note that by the functional strong law of large numbers and the fact that Condition 3(c) is satisfied follows as (53) in the proof of Proposition 7.2. Hence there exists a constant $\zeta > 0$ that does not depend on $r$ such that for any $T > 0$

Since $\hat{W}'(0) = 0$ and $\hat{W}'(t) \geq 0$ for all $t \geq 0$,

and so to show

it is enough to show that the right-hand side of (B29) weakly converges to 0 as $r \to \infty$. We know from Lemma 7.2 that $\mu_j (1/r) \hat{T}_j \to 0$ as $r \to \infty$, and it is obvious that

Hence it remains to show that

For this, note that by the functional strong law of large numbers and the fact that $\hat{T}_j(t) \leq t$, $j \in \{1, 2, 3\}$, for all $t \geq 0$,

Also, since $\hat{Q}_k(t) \leq (1/r^2) A_k(r^2 t)$, $k \in \{1, 2\}$, for all $t \geq 0$, and

it follows that for any $t > 0$

$$P \left( \int_0^1 (r^2 \delta_t) \left( \hat{Q}_k(s) - \frac{1}{r^2} 1[\hat{T}_j(r^2 s) > 0] - \frac{1}{r^2} 1[\hat{T}_j(r^2 s) > 0] \right) ds > \delta_1(\lambda_k t + 1)t \right) \to 0$$

$$P \left( \int_0^1 (r^2 \delta_t) \left( \hat{Q}_k(s) - \frac{1}{r^2} 1[\hat{T}_j(r^2 s) > 0] \right) ds > \delta_2(\lambda_k t + 1)t \right) \to 0.$$
as \( r \to \infty \). Hence we can also use the functional strong law of large numbers to conclude that

\[
\frac{1}{r} \mathcal{N}^r_t \left( \int_0^t \left( r^2 \delta^*_i \right) \left( \tilde{Q}_i^r(s) - \frac{1}{r^2} \mathbf{1}_i^r \left( \tilde{T}_i^r \left( r^2 s \right) > 0 \right) - \frac{1}{r^2} \mathbf{1}_i^r \left( \tilde{T}_i^r \left( r^2 s \right) > 0 \right) \right) \, ds \right) \to 0
\]

\[
\frac{1}{r} \mathcal{N}^r_t \left( \int_0^t \left( r^2 \delta^*_i \right) \left( \tilde{Q}_i^r(s) - \frac{1}{r^2} \mathbf{1}_i^r \left( \tilde{T}_i^r \left( r^2 s \right) > 0 \right) \right) \, ds \right) \to 0,
\]

as \( r \to \infty \). Also using Assumption 3.1, we conclude that

\[
\frac{1}{r} \tilde{X}_i^r(t) + \frac{\mu^e_2}{\mu_3} \tilde{Y}_i^r \to 0 \quad \text{as} \quad r \to \infty.
\]

Hence we have established (B27).

It is a direct consequence of (B27) that

\[
(\tilde{Q}_i^r, \tilde{Q}_i^r) \to 0 \quad \text{as} \quad r \to \infty.
\]

Hence we can also conclude from the representation (B28) that \((1/r)\tilde{I}_i^r \to 0\) as \( r \to \infty \). Therefore, in order to complete the argument that

\[
(\tilde{Q}_i^r, \tilde{Q}_i^r, \tilde{T}_i^r, \tilde{T}_i^r, \tilde{X}_i^r, \tilde{X}_i^r) \to (0, 0, 0, 0) \quad \text{as} \quad r \to \infty,
\]

it remains to show

\[
\tilde{T}_i^r \to \left( \frac{\lambda_1 - \mu_1}{\mu_2} \right) e \quad \text{and} \quad \tilde{T}_i^r \to \left( \frac{\lambda_2}{\mu_3} \right) e \quad \text{as} \quad r \to \infty.
\]

This follows because from (14)

\[
\tilde{Q}_i^r(t) + \int_0^t \left( r^2 \delta^*_i \right) \left( \tilde{Q}_i^r(s) - \frac{1}{r^2} \mathbf{1}_i^r \left( \tilde{T}_i^r \left( r^2 s \right) > 0 \right) - \frac{1}{r^2} \mathbf{1}_i^r \left( \tilde{T}_i^r \left( r^2 s \right) > 0 \right) \right) \, ds
\]

\[
= \frac{1}{r} \tilde{X}_i^r(t) - \frac{1}{r} \tilde{S}_i^r (\tilde{T}_i^r (t)) - \frac{1}{r} \tilde{S}_i^r (\tilde{T}_i^r (t)) + (\lambda_1 t - \mu_1 \tilde{T}_i^r (t) - \mu_2 \tilde{T}_i^r (t))
\]

\[
- \frac{1}{r} \mathcal{N}^r_t \left( \int_0^t \left( r^2 \delta^*_i \right) \left( \tilde{Q}_i^r(s) - \frac{1}{r^2} \mathbf{1}_i^r \left( \tilde{T}_i^r \left( r^2 s \right) > 0 \right) - \frac{1}{r^2} \mathbf{1}_i^r \left( \tilde{T}_i^r \left( r^2 s \right) > 0 \right) \right) \, ds \right).
\]

Since \( \tilde{Q}_i^r \to 0 \) as \( r \to \infty \), it follows that

\[
\int_0^t \left( r^2 \delta^*_i \right) \left( \tilde{Q}_i^r(s) - \frac{1}{r^2} \mathbf{1}_i^r \left( \tilde{T}_i^r \left( r^2 s \right) > 0 \right) - \frac{1}{r^2} \mathbf{1}_i^r \left( \tilde{T}_i^r \left( r^2 s \right) > 0 \right) \right) \, ds \Rightarrow 0 \quad \text{as} \quad r \to \infty,
\]

and

\[
\frac{1}{r} \mathcal{N}^r_t \left( \int_0^t \left( r^2 \delta^*_i \right) \left( \tilde{Q}_i^r(s) - \frac{1}{r^2} \mathbf{1}_i^r \left( \tilde{T}_i^r \left( r^2 s \right) > 0 \right) - \frac{1}{r^2} \mathbf{1}_i^r \left( \tilde{T}_i^r \left( r^2 s \right) > 0 \right) \right) \, ds \right) \Rightarrow 0 \quad \text{as} \quad r \to \infty.
\]

Hence

\[
\lambda_1 e - \mu_1 \tilde{T}_i^r (\cdot) - \mu_2 \tilde{T}_i^r (\cdot) \to 0.
\]

Since \( \tilde{T}_i^r \to e \) as \( r \to \infty \), we conclude that

\[
\tilde{T}_i^r \Rightarrow \frac{\lambda_1 - \mu_1}{\mu_2} e.
\]

Since \( \tilde{T}_i^r + \tilde{T}_i^r + (1/r)\tilde{I}_i^r = e \), and we have already shown \((1/r)\tilde{I}_i^r \to 0\) as \( r \to \infty \), it also follows that

\[
\tilde{T}_i^r \Rightarrow \left( 1 - \frac{\lambda_1 - \mu_1}{\mu_2} \right) e \quad \text{as} \quad r \to \infty.
\]

Since \( 1 - (\lambda_1 - \mu_1)/\mu_2 = \lambda_2/\mu_3 \) from Assumption 3.1, this completes the argument to establish (B30).

Finally, to complete the proof, it remains to show \( \tilde{X}_i^r \to \tilde{\xi} \) as \( r \to \infty \). For this, first note that the weak convergence in (16), the fact that \( \tilde{T} \Rightarrow T^* \) as \( r \to \infty \), and the random time change theorem establish that

\[
\tilde{A}_i^r + \frac{\mu^e_2}{\mu_3} \tilde{S}_i^r \circ \tilde{T}_i^r - \frac{\mu^e_2}{\mu_3} \tilde{S}_i^r \circ \tilde{T}_i^r - \frac{\mu^e_2}{\mu_3} \tilde{S}_i^r \circ \tilde{T}_i^r \Rightarrow \tilde{\xi}, \quad \text{as} \quad r \to \infty.
\]
Since 
\[(\hat{Q}_1^r, \hat{Q}_2^r) \Rightarrow (0, 0) \quad \text{as} \quad r \to \infty,\]
the functional central limit theorem and random time change theorem establish that

\[\hat{N}_1^r\left(\int_0^r t^2 \delta_1^r(\hat{Q}_1^r(s)) \, ds\right) + \hat{N}_2^r\left(\int_0^r t^2 \delta_2^r(\hat{Q}_2^r(s)) \, ds\right) \Rightarrow 0 \quad \text{as} \quad r \to \infty.\]

Assumption 3.1 completes the proof. \(\square\)

**Proof of Lemma 7.4.** Let \(\epsilon > 0\). We first observe that for any \(\eta > 0\), there exists \(M > 0\) such that for all \(r\) large enough

\[P\left(\sup_{0 \leq s \leq t} \hat{W}^r(s) \geq M\right) < \eta.\]  \(\text{(B31)}\)

To see (B31), note that it follows from the fact that \(\hat{W}^r = \hat{Q}_1^r + (\mu_2^r/\mu_3^r) \hat{Q}_2^r\) and the representations for \(\hat{Q}_1^r\) and \(\hat{Q}_2^r\) in (20) and (21) that for any \(s \geq 0\),

\[\hat{W}^r(s) = X_1^r(s) + \frac{\mu_2^r}{\mu_3^r} \hat{X}_2^r(s) + \mu_3^r \hat{I}_3^r(s) + \epsilon^r_t(s) - \int_0^r (r^2 \delta_1^r) \hat{Q}_1^r(u) + \frac{L^r}{\mu_3^r} (r^2 \delta_2^r) \hat{Q}_2^r(u) \, du + \frac{\mu_3^r}{\mu_2^r} \hat{I}_2^r(s),\]

where \(\epsilon^r_t\) is as defined in the proof of Theorem 8.1. Then,

\[w = \hat{W}^r,\]
\[x = \hat{X}_1^r + \frac{\mu_2^r}{\mu_3^r} \hat{X}_2^r + \mu_3^r \hat{I}_3^r + \epsilon^r_t,\]
\[z = (r^2 \delta_1^r) \hat{Q}_1^r + \frac{\mu_3^r}{\mu_3^r} (r^2 \delta_2^r) \hat{Q}_2^r,\]
\[y = \mu_2^r \hat{I}_2^r,\]

satisfies the conditions of Lemma 8.3 with \(L = L'/r\). (Note that \(\hat{W}^r(0) = 0\). Also, the fact that Condition 3(c) is satisfied follows as in (53) in the proof of Proposition 7.2.) Hence there exists a constant \(c > 0\) that does not depend on \(r\) such that

\[\text{Osc}(\hat{W}^r, [0, T]) \leq c\left(\text{Osc}\left(\hat{X}_1^r + \frac{\mu_2^r}{\mu_3^r} \hat{X}_2^r + \epsilon^r_t + \mu_3^r \hat{I}_3^r, [0, t]\right) + \frac{L^r}{r}\right).\]

The process \(\hat{X}_1^r + (\mu_2^r/\mu_3^r) \hat{X}_2^r + \mu_3^r \hat{I}_3^r\) is \(C\)-tight because Lemmas 7.2 and 7.3 establish weak convergence, and weak convergence establishes \(C\)-tightness. It is obvious that \(\epsilon^r_t \Rightarrow 0\) as \(r \to \infty\). Hence there exists \(M > 0\) such that for all \(r\) large enough

\[P\left(c\left(\text{Osc}\left(\hat{X}_1^r + \frac{\mu_2^r}{\mu_3^r} \hat{X}_2^r + \epsilon^r_t + \mu_3^r \hat{I}_3^r, [0, t]\right) + \frac{L^r}{r}\right) > M\right) < \eta.\]

Since \(\hat{W}^r(0) = 0\),

\[\sup_{0 \leq s \leq t} \hat{W}^r(s) = \text{Osc}(\hat{W}^r, [0, t]),\]

and so it follows that

\[P\left(\sup_{0 \leq s \leq t} \hat{W}^r(s) \geq M\right) = P(\text{Osc}(\hat{W}^r, [0, T]) \geq M) < \eta,\]

which establishes (B31).

We next argue that the processes \(\epsilon^r, \hat{V}_1^r, \hat{V}_2^r,\) and \(\hat{W}^r\) are all \(C\)-tight. For this, it follows from Theorem 15.1 in Billingsley [9] that it is sufficient to verify the following two conditions.

(i) For each \(\eta > 0\), there exists \(M\) such that

\[P\left(\sup_{0 \leq s \leq t} \max(\epsilon^r(s), \hat{V}_1^r(s), \hat{V}_2^r(s), \hat{W}^r(s)) > M\right) \leq \eta,\]

for all large enough \(r\).
(ii) For each $\epsilon > 0$ and $\eta > 0$, there exists $\xi$ such that
\[
P\left( \sup_{0 \leq t \leq 1-\xi} \max\left( w(\hat{e}', [t, t + \xi]), w(\hat{V}'_1, [t, t + \xi]), w(\hat{V}'_2, [t, t + \xi]), w(\hat{V}', [t, t + \xi]) \right) \geq \epsilon \right) \leq \eta
\]
for all large enough $r$, where for any set $S \subset [0, 1]$ and $x \in D$,
\[
w(x, S) := \sup_{u \in S} |x(u) - x(v)|.
\]
Condition (i) follows in a straightforward manner from (B31). For condition (ii), note that for any $t \in [0, 1 - \delta]$, since for large enough $r$, $r^2 \delta_1 + 1$ and $(\mu_2/\mu_3)(r^2 \delta_2) < (\mu_2/\mu_3)\delta_2 + 1,$
\[
\max( w(\hat{e}', [t, t + \xi]), w(\hat{V}'_1, [t, t + \xi]), w(\hat{V}'_2, [t, t + \xi]), w(\hat{V}', [t, t + \xi]) )
\leq 2\epsilon \left( \delta_1 + \frac{\mu_2}{\mu_3} \delta_2 + 2 \right) \sup_{0 \leq s \leq t} |\hat{W}'(s)|.
\]
Therefore, given $\epsilon > 0$, for any $M > 0$, if we choose $\xi > 0$ small enough so that
\[
2\epsilon \left( \delta_1 + \frac{\mu_2}{\mu_3} \delta_2 + 2 \right) M < \epsilon,
\]
then
\[
P\left( \sup_{0 \leq t \leq 1-\xi} \max( w(\hat{e}', [t, t + \xi]), w(\hat{V}'_1, [t, t + \xi]), w(\hat{V}'_2, [t, t + \xi]), w(\hat{V}', [t, t + \xi]) ) \geq \epsilon \right)
\cap \left\{ \sup_{0 \leq s \leq t} |\hat{W}'(s)| \leq M \right\} = 0.
\]
Hence for $M$ large enough so that (B31) holds,
\[
P\left( \sup_{0 \leq t \leq 1-\xi} \max( w(\hat{e}', [t, t + \xi]), w(\hat{V}'_1, [t, t + \xi]), w(\hat{V}'_2, [t, t + \xi]), w(\hat{V}', [t, t + \xi]) ) \geq \epsilon \right)
\leq 0 + P\left( \sup_{0 \leq s \leq t} |\hat{W}'(s)| \leq M \right)
< \eta,
\]
and we can conclude that condition (ii) holds.

The process $\hat{X}'$ is $C$-tight because it follows from Lemma 7.3 that $\hat{X}'$ weakly converges to Brownian motion as $r \to \infty$. The process $\mu_1^i \hat{I}'_1$ is $C$-tight because it follows from Lemma 7.2 that $\hat{I}'_1 \Rightarrow 0$ as $r \to \infty$.

Finally, to complete the proof, it is sufficient to show that sequence $\{\hat{W}', \mu_2^i \hat{I}'_1 \}$ is $C$-tight. For this, recall from the proof of Proposition 7.2 that
\[
(\hat{W}', \mu_2^i \hat{I}'_1, \hat{X}' - \hat{V}'_1 - \hat{V}'_2 + \hat{e}' - \hat{V}' + \mu_1^i \hat{I}'_1)
\]
is a solution to the perturbed Skorokhod problem defined in Williams [37, Theorem 5.1], with the $\delta$ in that theorem equalling $L'/r$. Therefore, in Williams [37, Theorem 5.1], for any $0 \leq t_1 < t_2$, there exists a constant $\xi$ that does not depend on $r$ such that
\[
w(\hat{W}', [t_1, t_2]) \leq \xi \left( w(\hat{X}' - \hat{V}'_1 - \hat{V}'_2 + \hat{e}' - \hat{V}' + \mu_1^i \hat{I}'_1, [t_1, t_2]) + \frac{L'}{r} \right),
\]
\[
w(\mu_2^i \hat{I}'_1, [t_1, t_2]) \leq \xi \left( w(\hat{X}' - \hat{V}'_1 - \hat{V}'_2 + \hat{e}' - \hat{V}' + \mu_1^i \hat{I}'_1, [t_1, t_2]) + \frac{L'}{r} \right).
\]
Since the sum of $C$-tight sequences is again $C$-tight, it follows from the preceding paragraphs that the sequence $\{\hat{X}' - \hat{V}'_1 - \hat{V}'_2 + \hat{e}' - \hat{V}' + \mu_1^i \hat{I}'_1 \}$ is $C$-tight. Then, noting that $L'/r \to 0$ as $r \to \infty$, it is straightforward to show that the conditions in Billingsley [9, Theorem 15.1] are satisfied to establish that the sequence $\{\hat{W}', \mu_2^i \hat{I}'_1 \}$ is $C$-tight. □
PROOF OF LEMMA 8.1. It follows from (20) and (21) that it is enough to show that the families
\[
\int_0^\infty e^{-\gamma t} \sup_{0 \leq s \leq t} \hat{A}_k(s) \, dt, \quad k = 1, 2,
\]
\[
\int_0^\infty e^{-\gamma t} \sup_{0 \leq s \leq t} \hat{S}_j(s) \, dt,
\]
(since \(\hat{T}_j(t) \leq t\) for all \(t \geq 0\), \(j = 1, 2, 3\), and \(\int_0^\infty e^{-\gamma t} \sup_{0 \leq s \leq t} \hat{N}_k(\hat{T}_k(s))\), \(k = 1, 2\), are all uniformly integrable families, and that there exist positive constants \(d_i - d_i\) such that
\[
E \left[ \sup_{0 \leq s \leq t} |\hat{A}_k(s)| \right] \leq d_2 + d_3 t, \quad k = 1, 2, \tag{B32}
\]
\[
E \left[ \sup_{0 \leq s \leq t} |\hat{S}_j(s)| \right] \leq d_4 + d_3 t, \quad j = 1, 2, 3, \tag{B33}
\]
\[
E \left[ \sup_{0 \leq s \leq t} |\hat{N}_k(\hat{T}_k(s))| \right] \leq d_6 + d_7 t, \quad k = 1, 2, \tag{B34}
\]
since it follows from Assumption 3.1 that there exists a constant \(d\) such that
\[
r \mu_2 \left( \frac{\lambda_1 - \mu_1}{\mu_2} - \frac{\lambda_1 - \mu_1}{\mu_3} \right) t + r \mu_3 \left( \frac{\lambda_2 - \mu_2}{\mu_3} - \frac{\lambda_2 - \mu_2}{\mu_3} \right) t \leq dt.
\]
Since for any \(x > 0\), \(x < 1 + x^2\), for any process \(\hat{P}\),
\[
E \left[ \sup_{0 \leq s \leq t} |\hat{P}(s)| \right] \leq E \left[ 1 + \left( \sup_{0 \leq s \leq t} |\hat{P}(s)| \right)^2 \right] = 1 + E \left[ \sup_{0 \leq s \leq t} |\hat{P}(s)|^2 \right],
\]
to establish (B32)–(B34), it is enough to show that there exist positive constants \(d_i - d_i\) such that
\[
E \left[ \sup_{0 \leq s \leq t} |\hat{A}_k(s)|^2 \right] \leq d_2^2 + d_3^2 t, \quad k = 1, 2, \tag{B35}
\]
\[
E \left[ \sup_{0 \leq s \leq t} |\hat{S}_j(s)|^2 \right] \leq d_4^2 + d_3^2 t, \quad j = 1, 2, 3, \tag{B36}
\]
\[
E \left[ \sup_{0 \leq s \leq t} |\hat{N}_k(\hat{T}_k(s))|^2 \right] \leq d_6^2 + d_7^2 t, \quad k = 1, 2. \tag{B37}
\]
Also, since for any sequence of processes \(\{\hat{P}_v\}\),
\[
E \left[ \left( \int_0^\infty e^{-\gamma t} \sup_{0 \leq s \leq t} \hat{P}_v(s) \, dt \right)^2 \right] \leq E \left[ \int_0^\infty e^{-2\gamma t} \sup_{0 \leq s \leq t} |\hat{P}_v(s)|^2 \, dt \right],
\]
the inequalities (B35)–(B37) establish the stated uniform integrability.

The inequalities (B35) and (B36) follow since the arguments that establish (172) in Bell and Williams [8] also apply here. To see (B37), first note that \(N_v\) are independent, unit-rate Poisson processes, and so \(N_v(\cdot) - (\cdot)\) is a martingale with respect to the filtration \(\sigma(N(s): 0 \leq s \leq t)\), and \(|N_v(\cdot) - (\cdot)|\) is a submartingale. Then, it follows from Doob’s maximal inequality (see, for example, in Karatzas and Shreve [22, Theorem 1.3.8, part (iii)]) that for any \(v > 0\)
\[
E \left[ \sup_{0 \leq s \leq v} (N_v(r^2 s) - r^2 s)^2 \right] = E \left[ \left( \sup_{0 \leq s \leq v} |N_v(r^2 s) - r^2 s| \right)^2 \right] \leq 4E[(N_v(r^2 v) - r^2 v)^2] = 4r^2 v. \tag{B38}
\]
Also note that it follows from the definition of \(\hat{\tau}_k\) and Lorden’s inequality for the mean value of a renewal process at time \(r^2 t\) that
\[
E[\hat{\tau}_k(t)] \leq (r^2 \delta_k)^t \frac{E[A_k^t]}{t} \leq (r^2 \delta_k)^t \frac{\lambda^k r^2 t + \alpha^2 + 1}{r^2}. \tag{B39}
\]
Then, we can use (B38) and (B39) to establish (B37) as follows

\[
E\left(\sup_{0 \leq s \leq t} (\tilde{N}_k^{(r)}(s))^2\right) \\
= \frac{1}{r^2} E\left[\sup_{0 \leq s \leq \tilde{t}_k^{(r)}} (N_k(r^2 s) - r^2 s)^2\right] \\
= \frac{1}{r^2} \int_0^{\infty} E\left[\sup_{0 \leq s \leq \tilde{t}_k^{(r)}} (N_k(r^2 s) - r^2 s)^2 | \tilde{t}_k^{(r)}(t) \in dv]\right] P(\tilde{t}_k^{(r)}(t) \in dv) \\
= \frac{1}{r^2} \int_0^{\infty} E\left[\sup_{0 \leq s \leq v} (N_k(r^2 s) - r^2 s)^2\right] P(\tilde{t}_k^{(r)}(t) \in dv) \\
\leq 4 \int_0^{\infty} v P(\tilde{t}_k^{(r)}(t) \in dv) \\
= 4E[\tilde{t}_k^{(r)}(t)] \\
= 4(r^2 \delta_k^\prime \frac{\lambda_k r^2 t + \alpha_k^2 + 1}{r^2})
\]

because

\[
(r^2 \delta_k^\prime \frac{\lambda_k r^2 t + \alpha_k^2 + 1}{r^2}) \leq (\delta_k + 1)(\lambda_k + 1) t + 1,
\]

for large enough \( r \). \( \square \)

**Proof of Lemma 8.2.** It is enough to show that there exist constants \( d_1 \) and \( d_2 \) such that

\[
E\tilde{t}_k^{(r)}(t)^2 \leq d_1 + d_2 t^4,
\]

since \( \sup_{0 \leq s \leq t} \tilde{t}_k^{(r)}(s) = \tilde{t}_k^{(r)}(t) \), for any \( x > 0 \), \( x < 1 + x^2 \), and

\[
E\left[\left(\int_0^{\infty} e^{-\gamma t} \sup_{0 \leq s \leq t} \tilde{t}_k^{(r)}(s) dt\right)^2\right] \leq E\left[\int_0^{\infty} e^{-2\gamma t} \left(\sup_{0 \leq s \leq t} \tilde{t}_k^{(r)}(s)\right)^2 dt\right] = \int_0^{\infty} e^{-2\gamma t} E[\tilde{t}_k^{(r)}(t)^2] dt,
\]

so that (B40) implies the stated uniform integrability. The argument to show (B40) is very similar to the last paragraph of the proof of Theorem 5.3 in Bell and Williams [8], except that we must also account for reneging. It follows that

\[
E\tilde{t}_k^{(r)}(t)^2 = \int_0^{r^2 t^2} P(I'_i(r^2 t) > r \sqrt{s}) ds.
\]

Next, for \( \tau_0^\prime := \inf\{s \geq 0 : Q'_i(s) \geq L'\}, \)

\[
P(I'_i(r^2 t) > r \sqrt{s}) = P(I'_i(r^2 t) > r \sqrt{s} | I'_i(\tau_0^\prime) > r \sqrt{s}) \times P(I'_i(\tau_0^\prime) > r \sqrt{s}) \\
+ P(I'_i(r^2 t) > r \sqrt{s} | I'_i(\tau_0^\prime) \leq r \sqrt{s}) \times P(I'_i(\tau_0^\prime) \leq r \sqrt{s}) \\
\leq P(I'_i(\tau_0^\prime) > r \sqrt{s}) + P(I'_i(r^2 t) - I'_i(\tau_0^\prime) > 0).
\]

Hence for \( r^\prime := 8L'/(\lambda_1 - \mu_1), \)

\[
E[\tilde{t}_k^{(r)}(t)^2] \leq \left(\frac{r^\prime}{r}\right)^2 + \int_0^{r^2 t^2} P(I'_i(\tau_0^\prime) > r^\prime) ds + r^2 t^2 P(I'_i(r^2 t) - I'_i(\tau_0^\prime) > 0).
\]

It follows from (B16) and (B17) in the proof of Lemma 7.2 in the appendix that for large enough \( r \), there exist positive, finite constants \( C_1 - C_9 \) (provided \( c_9 \) is chosen large enough), that do not depend on \( r \) or \( t \), such that

\[
P(I'_i(\tau_0^\prime) > r^\prime) \leq C_1 \exp(-C_2 L') + \left(C_3 \frac{L'}{r}\right)^{C_4 L'}
\]

and

\[
P(I'_i(r^2 t) - I'_i(\tau_0^\prime) > 0) \leq (C_4 + C_5 r^2 t)^3 \exp(-C_6 L') + \exp(-C_7 r^2 t) + C_8 r^2 t \left(C_9 \frac{L'}{r}\right)^{C_0 L'}
\]
Hence
\[
E[\hat{N}_1(t)^2] \leq \left(\frac{t'}{r}\right)^2 + r^2 \left( C_1 \exp(-C_2 L') + \left( C_1 \frac{L'}{r} \right)^{C_2 L'} \right)
+ r^2 C_4 + C_5 r^2 t (\exp(-C_6 L') + \exp(-C_7 r^2 t)) + r^4 C_8 \left( C_5 \frac{L'}{r} \right)^{C_2 L'}.
\]

Since, recalling that \( L' = [c \log r] \) for \( c > c_0 \),
- \((r'/r)^2 \to 0, r^2 (C_5 L'/r)^{C_2 L'} \to 0 \), and \( r^4 (C_8 L'/r)^{C_2 L'} \to 0 \), as \( r \to \infty \), and \( r^2 \exp(-C_2 L') \to 0 \) for \( c \) sufficiently large, and
- there exists a constant \( d_2 \) so that for large enough \( r \)
\[
r^2 r^2 t (C_4 + C_5 r^2 t) (\exp(-C_6 L') + \exp(-C_7 r^2 t)) < d_2 t^4,
\]
because, for any \( d > 0 \), \( x^i e^{-dx} \) is bounded on \([0, \infty)\) for any \( i = 1, 2, 3 \) (with \( r^2 t \) in place of \( x \)), it follows that there exists positive constants \( d_1 \) and \( d_2 \) such that \( (B40) \) holds.

**Proof of Lemma 8.3.** Let \( w, y \in \mathbf{D} \) be such that \( \underline{w}(0) = w(0) \) and
1. \( w(t) = x(t) + y(t) \) for all \( t \in [0, T] \),
2. \( w(t) \geq 0 \) for all \( t \in [0, T] \),
3. (a) \( y(0) \geq 0 \),
   (b) \( y \) is nondecreasing,
   (c) \( \int_{[0,T]} 1[w(t) > 0] d y(t) = 0 \).
Since \( w \) and \( \underline{w} \) are nonnegative processes, and \( w(0) = \underline{w}(0) = 0 \),
\[
\text{Osc}(w, [0, T]) = \sup_{0 \leq t \leq T} w(t),
\]
\[
\text{Osc}(\underline{w}, [0, T]) = \sup_{0 \leq t \leq T} \underline{w}(t).
\]
Furthermore, Theorem 5.1 in Williams [37] establishes that there exists \( c' > 0 \) such that
\[
\text{Osc}(w, [0, T]) \leq c' \text{Osc}(x, [0, T]).
\]
Therefore, if we can show that
\[
w(t) \leq \underline{w}(t) + L \quad \text{for all } t \geq 0,
\]
so that
\[
\sup_{0 \leq t \leq T} w(t) = \text{Osc}(w, [0, T]) \leq \sup_{0 \leq t \leq T} \underline{w}(t) + L = \text{Osc}(\underline{w}, [0, T]) + L,
\]
then it follows that
\[
\text{Osc}(w, [0, T]) \leq c' \text{Osc}(x, [0, T]) + L
\leq \max(c', 1)(\text{Osc}(x, [0, T]) + L).
\]
This completes the proof, by letting the constant \( c = \max(c', 1) \).

The proof that \( (B41) \) holds is by contradiction. Suppose that there exists \( t > 0 \) for which \( w(t) > \underline{w}(t) + L \).
Since \( w, \underline{w} \in \mathbf{D} \), there exists \( s \in [0, t) \) such that \( w(s) \leq \underline{w}(s) + L \) and \( w(u) > \underline{w}(u) + L \) for \( s < u \leq t \). Then,
\[
w(t) - \underline{w}(t) - L = w(s) - \underline{w}(s) - \left( \int_s^t z(v) dv \right)
+ (y(t) - y(s)) - (y(t) - y(s)) - L
\leq y(t) - y(s),
\]
since \( w(s) - \underline{w}(s) \leq L \), \( z \) is nonnegative, and \( y \) is nondecreasing. But \( w(u) > \underline{w}(u) + L \geq L \) for \( s < u \leq t \) implies \( y(t) - y(s) = 0 \). We conclude that \( w(t) - \underline{w}(t) - L \leq 0 \), which is a contradiction. □
Proof of Lemma 8.4. It follows from the functional strong law of large numbers that

$$\langle \bar{A}', \tilde{S}', \tilde{N}' \rangle \Rightarrow (\lambda e, \mu e, e) \quad \text{as } r \to \infty.$$  

Also, since they correspond to cumulative allocations of time, each of the three components of \(T'\) is uniformly Lipschitz continuous with a Lipschitz constant less than or equal to one, and this property is preserved by the fluid-scaled processes \(\bar{T}'\). It now follows that \(\{\bar{A}', \tilde{S}', \tilde{N}', \bar{T}', \tilde{T}'\}\) is \(C\)-tight.

We now argue that \(\{\bar{A}', \tilde{S}', \tilde{N}', \bar{T}', \tilde{T}'\}\) is \(C\)-tight, for which it is enough to establish that \(\{\tilde{T}'\}\) is \(C\)-tight. For this, first observe that for any \(0 < \bar{S} \leq \tilde{T} < \bar{T}\), \(\tilde{T}'(t) - \tilde{T}'(s) \leq (r^2 \delta'_0)(t - s)\bar{A}'(1)\).

The above inequality and the functional strong law of large numbers then implies that for any \(\epsilon > 0\) and \(\eta_k = \epsilon(2\lambda_0 \delta_k)^{-1}\)

$$\limsup_{r \to \infty} P\left( \sup_{0 \leq t \leq 1} |\tilde{T}'_r(t)| \geq 2\lambda_0 \delta_k \right) \leq \limsup_{r \to \infty} P\left( (r^2 \delta'_0)\bar{A}'(1) \geq 2\lambda_0 \delta_k \right) = 0,$$

$$\limsup_{r \to \infty} P\left( \sup_{|t - s| \leq \eta_k} |\tilde{T}'_r(t) - \tilde{T}'_r(s)| \geq \epsilon \right) \leq \limsup_{r \to \infty} P\left( (r^2 \delta'_0)\eta_k \bar{A}'(1) \geq \epsilon \right) = 0.$$

We conclude in Billingsley [9, Theorem 15.1] that \(\{\tau'\}\) is \(C\)-tight.

Finally, to complete the proof, note that Equations (68)–(71) combined with the \(C\)-tightness established above and a random time change theorem shows that

$$\{(\bar{A}', \tilde{S}', \tilde{N}', \tilde{T}', \bar{T}')\}$$

is \(C\)-tight. \(\square\)

Proof of Lemma 8.5. The proof is very similar to the proof of Lemma 9.3 in Bell and Williams [8]. From Lemma 8.4, it follows that

$$\{(\bar{Q}', \bar{A}', \tilde{S}', \tilde{N}', \tilde{T}', \bar{T}')\}$$

(B42)

is \(C\)-tight. Thus it suffices to show that all weak limit points of this sequence are given by \((0, \lambda e, \mu e, e, 0, \tilde{T}'^*, 0)\).

For this, suppose that \((\bar{Q}, \bar{A}, \tilde{S}, \tilde{N}, \tilde{T}, \bar{T})\) is obtained as a weak limit of (B42) along a subsequence indexed by \(r'\). Without loss of generality, by appealing to the Skorokhod representation theorem [see, for example, Ethier and Kurtz [13, Theorem 3.1.8]], we may choose an equivalent distributional representation (for which we use the same symbols) such that all of the random processes in (B42) indexed by \(r'\), as well as the limit, are defined on the same probability space and the convergence in distribution is replaced by almost sure convergence on compact time intervals, so that

$$\{(\bar{Q}', \bar{A}', \tilde{S}', \tilde{N}', \tilde{T}', \bar{T}')\} \Rightarrow \{(\bar{Q}, \bar{A}, \tilde{S}, \tilde{N}, \tilde{T}, \bar{T})\}, \quad \text{u.o.c., a.s., as } r' \to \infty.$$

From the first sentence in the proof of Lemma 8.4, a.s., \(\bar{A} = \lambda e, \tilde{S} = \mu e, \) and \(\tilde{N} = e\). It follows exactly as in (132) in Bell and Williams [8] that \(\bar{Q} = 0\) a.s.. Hence also \(\tilde{r} = 0\) a.s.. Then, by letting \(r'' \to \infty\) in (68)–(71), we have a.s., for each \(t \geq 0\),

$$0 = \lambda_1 t - \mu_1 \bar{T}_1(t) - \mu_2 \bar{T}_2(t),$$

$$0 = \lambda_2 t - \mu_1 \bar{T}_3(t),$$

$$\bar{T}_1(t) = t - \bar{T}(t),$$

$$\bar{T}_2(t) = t - \bar{T}_3(t) - \bar{T}_4(t).$$

This is exactly (133)–(136) in Bell and Williams [8], and so the conclusion that \(\bar{T} = \bar{T}^*\) and \(\bar{L} = 0\) follows exactly as in the last three sentences of the proof of Lemma 9.3 in their paper, on page 25. \(\square\)

Proof of Lemma 8.6. It follows from the oscillation inequality in Lemma 8.3 (applied with \(L = 0\)) that for any \(T > 0\), there exists a constant \(\zeta > 0\) such that

$$\text{Osc}([\bar{W}', [0, T]) \leq \frac{\zeta}{2} \text{Osc}(\tilde{X}' + \tilde{e}', [0, T]).$$
Since $\hat{W}'(0) = 0$ and $\hat{W}'(t) \geq 0$ for all $t \geq 0$,
\[
\text{Osc}(\hat{W}', [0, T]) = \sup_{0 \leq t \leq T} \hat{W}'(t).
\]
Also,
\[
\text{Osc}(\hat{X}' + \hat{\epsilon}', [0, T]) \leq 2\left( \sup_{0 \leq t \leq T} |\hat{X}'(t)| + \sup_{0 \leq t \leq T} |\hat{\epsilon}'(t)| \right).
\]
Hence it follows that
\[
\sup_{0 \leq t \leq T} \hat{W}'(t) \leq \varepsilon\left( \sup_{0 \leq t \leq T} |\hat{X}'(t)| + \sup_{0 \leq t \leq T} |\hat{\epsilon}'(t)| \right). \quad (B43)
\]
To show the family $\left\{ \int_0^T e^{-\gamma t} (\hat{Q}'(t)) dt \right\}$ is uniformly integrable (UI), it is sufficient to show the family $\left\{ \int_0^T e^{-\gamma t} \hat{W}'(t) dt \right\}$ is UI. This is because (73) and (72) imply $0 \leq \hat{Q}'(t) \leq \hat{W}'(t)$ and $0 \leq \hat{Q}'(t) \leq (\mu'_1/\mu'_2) \hat{W}'(t)$ for every $t \geq 0$, and so the families $\left\{ \int_0^T e^{-\gamma t} \hat{Q}'(t) dt \right\}, k \in \{1, 2\}$, are dominated by a UI family. The family $\left\{ \int_0^T e^{-\gamma t} \hat{W}'(t) dt \right\}$ is UI by (B43), the fact that the family $\left\{ \int_0^T e^{-\gamma t} \hat{X}'(t) dt \right\}$ is UI by Lemma 8.1, and since $\hat{\epsilon}'(t) \leq \bar{I}$ for all $0 \leq t \leq T$ for large enough $r$ from its definition.

The limit in (74) follows from (B43), the fact that there exist constants $d_1 > 0$ and $d_2 > 0$ such that
\[
E\left[ \sup_{0 \leq s \leq t} |\hat{X}'(s)| \right] \leq d_1 + d_2 t, \quad \text{for all } r,
\]
by Lemma 8.1, and because $\hat{\epsilon}'(t) \leq 1$ for large enough $r$ by its definition.

Finally, to establish (75), first note that from the Assumption (1)
\[
(h_1 + c_1 \delta_1) \hat{Q}'(t) + (h_2 + c_2 \delta_2) \hat{Q}'(t) \geq (h_2 + c_2 \delta_2) \frac{\mu'_1}{\mu'_2} \hat{W}'(t)
\]
for every $t \geq 0$. Hence for a fixed $t > 0$
\[
\hat{J}'(T') \geq (h_2 + c_2 \delta_2) \frac{\mu'_1}{\mu'_2} e^{-\gamma t} E\left[ \int_0^t \hat{W}'(s) ds \right].
\]
Since on the subsequence $r'$, $\hat{J}'(T') < \infty$, it follows that
\[
\sup_{r'} E\left[ \int_0^t \hat{W}'(s) ds \right] < \infty. \quad \square
\]

**Proof of Lemma 8.7.** On the subsequence $r'$, it follows from Lemma 8.5 that $\hat{T}' \Rightarrow \hat{T}$ as $r' \to \infty$. Therefore, just as in the last paragraph of the proof of Lemma 7.3,
\[
\hat{X}' \Rightarrow \bar{X} \quad \text{as } r' \to \infty,
\]
recalling that $\bar{X}$ is a Brownian motion. The weak convergence of $\hat{X}'$ implies $\{\hat{X}'\}$ is C-tight. Next, it follows from (B43) in the proof of Lemma 8.6, the fact that $\{\hat{X}'\}$ is C-tight, and the fact that $\hat{\epsilon}'(t) \leq 1$ for all large enough $r$, that for any $\eta > 0$, there exists $M > 0$ such that
\[
P\left( \sup_{0 \leq s \leq t} \hat{W}(s) \geq M \right) < \eta.
\]
The argument that $\left\{ \int_0^T \hat{Q}'(s) ds \right\}, k \in \{1, 2\}$, are C-tight is exactly the argument in the second paragraph of the proof of Lemma 7.4, except that $\hat{Q}'_1$ and $\hat{Q}'_2$ must replace $\hat{V}'_1$ and $\hat{V}'_2$. \quad \square

**References**


