Dynamic Scheduling of a System with Two Parallel Servers in Heavy Traffic with Resource Pooling: Asymptotic Optimality of a Threshold Policy [1]

S. L. Bell and R. J. Williams[2]

Abstract

This paper concerns a dynamic scheduling problem for a queueing system that has two streams of arrivals to infinite capacity buffers and two (non-identical) servers working in parallel. One server can only process jobs from one buffer, whereas the other server can process jobs from either buffer. The service time distribution may depend on the buffer being served and the server providing the service. The system manager dynamically schedules waiting jobs onto available servers. We consider a parameter regime in which the system satisfies both a heavy traffic condition and a resource pooling condition. Our cost function is a mean cumulative discounted cost of holding jobs in the system, where the (undiscounted) cost per unit time is a linear function of normalized (with heavy traffic scaling) queue length. We first review the analytic solution of the Brownian control problem (formal heavy traffic approximation) for this system. We "interpret" this solution by proposing a threshold control policy for use in the original parallel server system. We show that this policy is asymptotically optimal in the heavy traffic limit and the limiting cost is the same as the optimal cost in the Brownian control problem. The techniques developed here are expected to be useful for analyzing the performance of threshold-type policies in more complex multiserver systems.

Short title: Dynamic scheduling of two parallel servers.

Key words and phrases: Queueing networks, dynamic control, resource pooling, heavy traffic, diffusion approximations, Brownian control problem, large deviations.

1991 AMS Subject Classification. Primary: 60K25, 68M20, 90B22, 90B35, Secondary: 60J70.

---

[1] The research for this paper was supported in part by NSF grant DMS 9703891.

[2] Address for both authors: Department of Mathematics, University of California, San Diego, La Jolla CA 92093-0112, U.S.A.

1 Introduction

Queueing networks (otherwise known as stochastic processing networks [11]) are used as stochastic models for modern telecommunication, manufacturing and computer systems. Some of these networks allow for flexible scheduling of jobs (see e.g., [17]) through dynamic (state-dependent) alternate routing and sequencing. It is a challenging problem to design dynamic control policies for such networks that are simple to implement and yet are at least approximately optimal in an appropriate sense. As one approach to this problem, some authors (see for example [4, 14, 15, 18, 24, 25, 34]) have followed the scheme first suggested by Harrison [8] where analysis of Brownian control problems (formal heavy traffic approximations to queueing network control problems) is combined with clever interpretation of their optimal (analytic) solutions to suggest "good" policies for some queueing network control problems. These analytically derived policies (as opposed to ones derived computationally by discretization of the Brownian control problem, see e.g., [9, 10, 20, 22, 23]), have frequently involved threshold-type control. Although these policies have usually performed well when simulated, there are few proofs [19, 26] of the asymptotic optimality (in heavy traffic) of such policies.

As a step towards providing a rigorous basis for this approach to dynamic scheduling of queueing networks, here we consider a queueing system (see Figure 1) consisting of two buffers and two parallel servers with dynamic scheduling capabilities. This "parallel server system" may be viewed as a simple model for a parallel computing system where processors have overlapping capabilities, or for a manufacturing test facility where test machines have differing primary functions and some overlapping secondary functions. More importantly, we believe the approach and techniques developed here provide templates for the treatment of more complex multiserver systems, i.e., systems with two or more parallel servers.

A detailed description of our parallel server system is given in Section 2. In this Introduction, we outline the structure to facilitate a description of the main results of the paper. A schematic for our parallel server system is shown in Figure 1. The circles represent single servers and the open ended rectangles represent infinite capacity buffers for holding jobs awaiting service. Arrivals to the two infinite capacity buffers are given by independent renewal processes with a long run arrival rate of λ_k for buffer k, $k = 1, 2$. Arrivals to buffer k are called class k jobs. Within each buffer, jobs are ordered according to their arrival times, with the earliest arrival being at the head of the line. Each job requires a single service at one of the servers, subject to the restriction that server 1 can only process jobs of class 1, whereas server 2 can process class 1 and class 2 jobs. Control of the system occurs through allocations of server time to *processing activities* defined as follows:

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.

For concreteness we make the following specific assumptions concerning service protocol. Once a job has commenced service at a server, it remains at that server until its service is complete, even if its service is interrupted for some time (e.g., by preemption by a job of the other class, if allowed). A server may not start on a new job of class k until it has finished serving any class k job that it is working on or that is in suspension. A server cannot work unless it has a job to work on (in particular, if there are no class 1 jobs in buffer 1 or at server 1, and server 2 has a class 1 job in process or in suspension, then server 1 cannot work on the class 1 job assigned to server 2 and must remain idle until a new class 1 job arrives to the system). For each activity there is an associated sequence of i.i.d. random variables, specifying the amounts of service time required by the successive jobs processed by that activity. The sequences for different activities are mutually independent and are independent of the arrival processes. The mean of the service times for activity j is $1/\mu_j$, $j = 1, 2, 3$.

In this paper, we focus on the parameter regime in which the system is nominally heavily loaded, i.e., $\frac{\lambda_1 - \mu_1}{\mu_2}$ is close to $1 - \frac{\lambda_2}{\mu_3}$, which has the interpretation that the long run fraction of server 2's time needed to help process class 1 jobs is close to the long run fraction of time left over after server 2 processes class 2 jobs. In addition, we assume that a (complete) resource pooling condition (cf. [10, 12, 18, 24]) is satisfied, i.e., in addition, $\lambda_1 > \mu_1$. In particular, in the optimal solution of the Brownian control problem, the two servers combine to form a single pooled resource or "super-server". We consider a cost function which is a mean cumulative discounted cost of holding jobs in the system, where the (undiscounted) cost per unit time, per unit of normalized (in heavy traffic scale) queue length, is a constant $h_k > 0$ for class k, $k = 1, 2$. (Other cost functions could be considered — we chose this one for concreteness and since it is commonly used in stochastic control.)

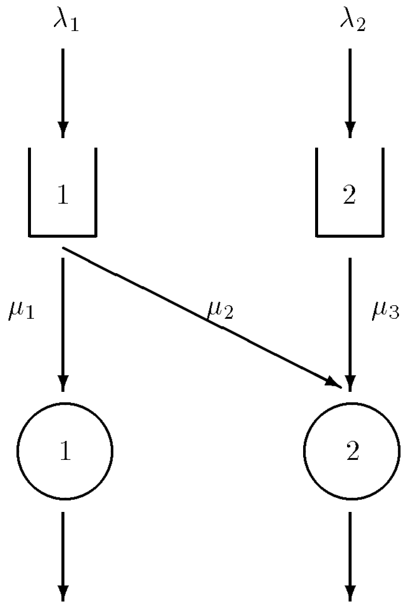

Figure 1: The Parallel Server System

The main results of the paper are the following. After reviewing the optimal solution of the Brownian control problem for our system, we "interpret" this solution by proposing a dynamic threshold control policy for use in the original parallel server system (see Definition 5.1). We show that this threshold policy is asymptotically optimal in the heavy traffic limit and that the limiting cost is the same as the optimal cost for the Brownian control problem (see Theorem 5.3). Formally (as in prior work on heavy traffic limit theorems), this involves considering a sequence of parallel server systems (see Section 3), each member of which has the same basic structure as the system described above, but in which combinations of the first order parameters approach limiting values at uniform rates (see Assumption 3.1), and in particular, $\frac{\lambda_1-\mu_1}{\mu_2} = 1 - \frac{\lambda_2}{\mu_3}$ and $\lambda_1 > \mu_1$ in the limit. In addition, here we only consider the case where $h_1\mu_2 \geq h_2\mu_3$ in the limit (see Assumption 3.2), which corresponds to class 2 being the "cheapest" (or equally cheap) class in which to hold jobs in the heavy traffic limit. For this parameter combination, one might be tempted to use a *static* priority policy suggested by an extrapolation of the classical $c\mu$ rule (see e.g., [27]), where $c = h$ here. Such a policy would require that server 1 works whenever possible and server 2 gives priority to class 1 jobs over those in class 2. As illustrated by Harrison [10] this *greedy scheduling policy* is "disastrously inefficient" and one is led to seek more efficient *dynamic* policies. In this paper, we propose such a dynamic policy and prove it is asymptotically optimal. Although not considered here, the complementary parameter regime $h_1\mu_2 < h_2\mu_3$ can be treated in a similar manner, though a little more simply, for in this case the static priority policy suggested by the $c\mu$-rule (i.e., server 2 gives priority to class 2 over class 1) is optimal.

The parallel server system treated here was considered previously by Harrison [10] under more restrictive assumptions. In particular, Harrison assumed the arrival processes were Poisson and the service times were deterministic with specific numeric values (up to a heavy traffic scale factor) for the arrival and service rates, whereas we allow i.i.d. interarrival and service times with arbitrary distributions subject to a finite exponential moment condition and a relationship between arrival and service rates that ensures the system is heavily loaded and allows (complete) resource pooling. In addition, Harrison's asymptotically optimal control policy only reviews the system status at fixed intervals of time, i.e., it is a so-called "discrete review" policy. On the other hand, we exhibit an asymptotically optimal "continuous review" policy which allows changes in service allocations to be made at random times, according to the state of the system. Our policy is easily described in terms of a threshold or safety stock level that is used to prevent unnecessary idleness of servers when there is still work in the system. A final difference is that our notion of asymptotic optimality involves a mean cumulative discounted cost whereas Harrison used a pathwise criterion.

Heavy traffic analysis of multiserver systems (parallel server systems with two or more servers) has also been considered by Kushner and Chen [20] and Harrison and López [12]. The work of Kushner and Chen considers a different parameter regime than that considered here. In a sense it is at the opposite end of the spectrum since our regime allows complete pooling and that of [20] does not allow for any resource pooling. In addition, solutions of the Brownian control problem in the regime of [20] are to be found by numerical means [21], whereas ours are derived by analytic means. The recent work of Harrison and López [12] identifies a condition for complete resource pooling in heavy traffic for the multiserver problem and proposes using the BIGSTEP discretization method of Harrison [9] to find candidates for "good" discrete review policies for this problem. However, no proof of asymptotic optimality of such policies is given in [12]. Assuming the heavy traffic resource pooling condition of Harrison and López [12], a candidate for an asymptotically optimal threshold policy is proposed in Williams [37] for the multiserver problem. The analysis of the two server problem considered in this paper is expected to play a key role in an iterative proof that this threshold policy is asymptotically optimal for the multiserver problem under the Harrison-López heavy traffic complete resource pooling condition.

The remainder of this paper is organized as follows. In Section 2 we complete the description of the parallel server system considered here. This includes a description of the primitive stochastic processes in our model, allowed scheduling control policies, and a specification of dynamic equations satisfied by the queue length process. In Section 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 asymptotically in heavy traffic (cf. [11]) and satisfies the complete resource pooling condition of Harrison and López [12]. We describe the normalization of the queue length process via diffusive scaling and we specify the associated cost function. In Section 4, following the general scheme proposed by Harrison [8], we state the formal Brownian control problem associated with our parallel server system and describe an optimal solution for this problem. (A similar description and analysis can be found in Harrison and López [12] for the multiserver system.) In Section 5, we propose a threshold control policy for the parallel server system. We then state the main results which show that this policy is asymptotically optimal in the heavy traffic limit and that the limiting cost is the same as the optimal cost in the Brownian control problem. An outline of our method of proof is given in Section 6. The details of the proofs are contained in Sections 7–9. Here a critical role is played by our analysis in Section 7 of what we call the residual process, which measures deviations of the class 1 queue length from the threshold level when our threshold policy is used. This allows us to establish a form of "state space collapse" (see Theorem 5.2) under this policy.

1.1 Notation and Terminology

The set of non-negative integers will be denoted by $\mathbb{N}$ and the value $+\infty$ will simply be denoted by ∞. For any real number x, $[x]$ will denote the integer part of x, i.e., the greatest integer that is less than or equal to x. 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 $\mathrm{diag}(a)$, and the dot product of two vectors a and b will be denoted by $a \cdot b$.

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

$$\|\omega\|_t = \sup_{s\in[0,t]} |\omega(s)|. \tag{1}$$

Consider $\mathbf{D}^m$ to be endowed with the usual Skorokhod $\mathbf{J_1}$-topology (see [6]). Let $\mathcal{M}^m$ denote the Borel σ-algebra on $\mathbf{D}^m$ associated with the $\mathbf{J_1}$-topology. This is the same σ-algebra as the one generated by the coordinate maps, i.e., $\mathcal{M}^m = \sigma\{\omega(s) : 0 \leq s < \infty\}$. All of the continuous-time processes in this paper will be assumed to have sample paths in $\mathbf{D}^m$ for some $m \geq 1$.

Suppose $\{W^n\}_{n=1}^\infty$ is a sequence of processes with sample paths in $\mathbf{D}^m$ 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 $(\mathbf{D}^m, \mathcal{M}^m)$ form a tight sequence, i.e., they form a weakly relatively compact sequence in the space of probability measures on $(\mathbf{D}^m, \mathcal{M}^m)$. The notation "$W^n \Longrightarrow W$", where W is a process with sample paths in $\mathbf{D}^m$, will mean that the probability measures induced by the W^n's on $(\mathbf{D}^m, \mathcal{M}^m)$ converge weakly to the probability measure on $(\mathbf{D}^m, \mathcal{M}^m)$ induced by W. 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 (u.o.c. in prob.) if $\mathbf{P}(\|W^n - W\|_t \geq \varepsilon) \to 0$ as $n \to \infty$ for each $\varepsilon > 0$ and all $t \geq 0$.

2 The Parallel Server System

The physicial structure of our parallel server system was described in the Introduction. This structure is the same as in the model considered by Harrison in [10]. However, our assumptions on the stochastic primitives as specified below are more general than those of [10] in that we allow non-Poisson renewal arrivals, i.i.d. *random* service times and more general rates.

2.1 Stochastic Primitives

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

We assume that the system is initially empty.

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 (variance divided by the mean squared) $\alpha_k^2 \in [0, \infty)$. 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 "0^{th} arrival" occurs at time zero. Setting $\xi_k(0) = 0$ and

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

we define

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

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 λ_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 the amount of service time required by the i^{th} job to be processed by activity j. Note that μ_j is the long run rate at which activity j can process its associated class of jobs if the associated server works continuously and exclusively on this class. For $j = 1, 2, 3$, let $\eta_j(0) = 0$,

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

and

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

Then S_1, S_2, S_3 are renewal processes and $S_1(t)$ 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 $j - 1$ jobs that server 2 could complete if that server worked continuously and exclusively on class $j - 1$ jobs in $[0, t]$.

We assume that the interarrival time sequences $\{u_k(i), i = 1, 2, \ldots\}$, $k = 1, 2$, and service time sequences $\{v_j(i), i = 1, 2, \ldots\}$, $j = 1, 2, 3$, are all mutually independent. Without loss of generality (by removing an exceptional $\mathbf{P}$-null set from Ω if necessary), we may and do assume that $A_k(t), S_j(t)$, $k = 1, 2$, $j = 1, 2, 3$, are finite-valued for all $t \geq 0$, *everywhere* on Ω. (We note that this also extends to the situation in the next section where we consider a sequence of parallel server systems.)

2.2 Scheduling Control

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

$$T(t) = (T_1(t), T_2(t), T_3(t))', \quad t \geq 0. \tag{6}$$

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

$$I_1(t) \equiv t - T_1(t) \tag{7}$$

is the cumulative idletime of server 1 up to time t,

$$I_2(t) \equiv t - T_2(t) - T_3(t) \tag{8}$$

is the cumulative idletime of server 2 up to time t, $S_1(T_1(t))$ is the number of jobs completed by server 1 up to time t, $S_j(T_j(t))$ is the number of class $j-1$ jobs completed by server 2, for $j = 2, 3$, up to time t, and for

$$Q_1(t) \equiv A_1(t) - S_1(T_1(t)) - S_2(T_2(t)), \tag{9}$$

$$Q_2(t) \equiv A_2(t) - S_3(T_3(t)), \tag{10}$$

$Q_k(t)$ is the number of class k jobs that are either in queue or "in progress" (i.e., being served or in suspension) at time t.

Now, T must satisfy certain properties that go along with its interpretation. Indeed, one could give a discrete-event type description of the properties that T must have, including any application specific constraints such as no preemption of service. Here we allow very general T's including those that may anticipate the future. For our analysis, we shall only need the following properties of the three-dimensional process $T = (T_1, T_2, T_3)'$. For $j = 1, 2, 3$, and $k = 1, 2$, and I, Q given by (7)–(10),

$$T_j(t) \in \mathcal{F} \text{ for each } t \geq 0, \tag{11}$$

$$T_j(\cdot) \text{ is continuous and non-decreasing with } T_j(0) = 0, \tag{12}$$

(13) $I_k(\cdot)$ is continuous and non-decreasing with $I_k(0) = 0$,

(14) $Q_k(t) \geq 0$ for all $t \geq 0$.

Note that conditions (12)–(13) imply that for $j = 1, 2, 3$, T_j is uniformly Lipschitz continuous with a Lipschitz constant of one.

The cost function we use for our control problem involves linear holding costs associated with the expense of holding (or storing) jobs in the system until they have completed service. We defer the precise description of this cost function to the next section, since it is formulated in terms of normalized queue lengths where the normalization is in diffusion scale, commensurate with the heavy traffic limiting regime in which we consider our model.

3 Heavy Traffic Assumptions, Scaling and the Cost Function

Even for the simple parallel server system described in the last section, the problem of finding a control policy that minimizes a cost associated with holding jobs in the system is notoriously difficult. One possible means for discriminating between policies is to look for policies that outperform others in some asymptotic regime. Here 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 of storage, where this is quantified to be of order $1/r$. Formally, we consider a sequence of parallel server 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 and service 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. 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

$$u_k^r(i) = \frac{1}{\lambda_k^r}\check{u}_k(i), \quad v_j^r(i) = \frac{1}{\mu_j^r}\check{v}_j(i), \tag{15}$$

where the $\check{u}_k(i)$, $\check{v}_j(i)$ do not depend on r, have mean one and squared coefficients of variation α_k^2, β_j^2, respectively. (The above structure is a convenient means of allowing the sequence of systems to approach heavy traffic by simply changing arrival and service rates while keeping the underlying sources of variability $\check{u}_k(i)$, $\check{v}_j(i)$ fixed. This type of setup has been used previously by others in treating heavy traffic limits, see e.g., Peterson [28]. For a first reading, the reader may like to simply choose $\lambda^r = \lambda$, $\mu^r = \mu$ for all r.)

To begin with, we make the following assumption on the first order parameters associated with our sequence of networks.

Assumption 3.1 *There are strictly positive constants, $\lambda_1, \lambda_2, \mu_1, \mu_2, \mu_3$, and real numbers θ_1, θ_2, such that*

(i) $\lambda_1 > \mu_1$,

(ii) $1 - \left(\frac{\lambda_1 - \mu_1}{\mu_2}\right) = \frac{\lambda_2}{\mu_3}$,

and as $r \to \infty$,

(iii) $\lambda_k^r \to \lambda_k, \ k = 1, 2$,

(iv) $\mu_j^r \to \mu_j, \ j = 1, 2, 3$,

(v) $r\mu_2^r \left(\frac{\lambda_1^r - \mu_1^r}{\mu_2^r} - \frac{\lambda_1 - \mu_1}{\mu_2}\right) \to \theta_1$,

(vi) $r\mu_3^r \left(\frac{\lambda_2^r}{\mu_3^r} - \frac{\lambda_2}{\mu_3}\right) \to \theta_2$.

Remark. Viewing the limiting values, λ_k, $k = 1, 2$, μ_j, $j = 1, 2, 3$, as parameters for a parallel server system of the same type as in the prelimit, we have the following interpretation of the above conditions. In the limit, the arrival rate λ_1 to class 1 exceeds the service rate μ_1 at server 1 and so the assistance of server 2 is needed to keep the class 1 queue length from growing without bound. Thus, we regard server 2 as a "helper" to server 1 in the processing of class 1. The long run fraction of server 2's time that will be required in this helper activity is $\frac{\lambda_1 - \mu_1}{\mu_2}$, since server 2 can process class 1 jobs at a long run rate of μ_2. The left member of the equation in (ii) is the long run fraction of server 2's time left over after helping process class 1 jobs and by (ii) this is exactly balanced by the long run fraction of server 2's time required to process the class 2 jobs using activity 3. Thus, we may think of the system as critically loaded in the sense that at the level of long run rates, the "capacity" of the servers is just sufficient to process the incoming load. Indeed, under the above assumption, the limiting parameters λ_k, μ_j satisfy the heavy traffic and complete resource pooling conditions of Harrison [11] and Harrison and López [12]. We have not allowed $\lambda_1 \leq \mu_1$ since $\lambda_1 = \mu_1$ would not lead to complete resource pooling, and $\lambda_1 < \mu_1$ would not satisfy the heavy traffic assumption. Conditions (v) and (vi) are the analogues for controlled networks of the usual heavy traffic conditions involving the rates at which traffic intensities approach one. Here, "nominal" long run fractions of time devoted to activities in the r^{th} system tend to limiting long run fractions, at a uniform rate across activities.

For each fixed r and control policy T^r with associated queue length Q^r and idletime I^r processes in the r^{th} system, we now define a fluid scaled process $\bar{T}^r$ and diffusion scaled processes $\hat{A}^r, \hat{S}^r, \hat{Q}^r, \hat{I}^r$. Note that A^r, S^r grow at long run average rates of λ^r, μ^r, respectively, and so they are first centered about their average rate processes before diffusion scaling is applied. For each $t \geq 0$, let

$$\bar{T}^r(t) = r^{-2} T^r(r^2 t), \tag{16}$$

$$\hat{A}^r(t) = r^{-1}\left(A^r(r^2 t) - \lambda^r r^2 t\right), \tag{17}$$

$$\hat{S}^r(t) = r^{-1}\left(S^r(r^2 t) - \mu^r r^2 t\right), \tag{18}$$

$$\hat{Q}^r(t) = r^{-1} Q^r(r^2 t), \tag{19}$$

$$\hat{I}^r(t) = r^{-1} I^r(r^2 t). \tag{20}$$

We note that the fluid scaling used here is the same as that in Williams [36], but is different from the fluid scaling used in most works concerned with stability analysis of queueing networks, where time is only accelerated by a factor of r and space is divided by a factor of r, see e.g., Bramson [2]. Though both incorporate the notion of law of large numbers scaling, we shall only need the former notion of fluid scaling here.

Now, equations (9)–(10) yield the following expressions for the normalized (diffusion scaled) queue length processes:

$$\hat{Q}_1^r(t) \equiv \hat{A}_1^r(t) - \hat{S}_1^r(\bar{T}_1^r(t)) - \hat{S}_2^r(\bar{T}_2^r(t)) + r(\lambda_1^r t - \mu_1^r \bar{T}_1^r(t) - \mu_2^r \bar{T}_2^r(t)), \tag{21}$$

$$\hat{Q}_2^r(t) \equiv \hat{A}_2^r(t) - \hat{S}_3^r(\bar{T}_3^r(t)) + r(\lambda_2^r t - \mu_3^r \bar{T}_3^r(t)). \tag{22}$$

On combining Assumption 3.1 with the finite variance and mutual independence of the stochastic primitives $\{\check{u}_k(i), i = 1, 2, \ldots\}$, $\{\check{v}_j(i), i = 1, 2, \ldots\}$, we may deduce from renewal process functional central limit theorems (cf. [16]) that

$$(\hat{A}^r, \hat{S}^r) \Longrightarrow (\tilde{A}, \tilde{S}) \quad \text{as } r \to \infty, \tag{23}$$

where $\tilde{A}, \tilde{S}$ are mutually independent, $\tilde{A}$ is a two-dimensional driftless Brownian motion that starts from the origin and has diagonal covariance matrix $\mathrm{diag}(\lambda_1\alpha_1^2, \lambda_2\alpha_2^2)$, and $\tilde{S}$ is a three-dimensional driftless Brownian motion that starts from the origin and has diagonal covariance matrix $\mathrm{diag}(\mu_1\beta_1^2, \mu_2\beta_2^2, \mu_3\beta_3^2)$.

For the r^{th} system, we consider a mean cumulative discounted holding cost for use of a control T^r having associated normalized queue length process $\hat{Q}^r$:

$$\hat{J}^r(T^r) = \mathbf{E}\left(\int_0^\infty e^{-\gamma t}\, h \cdot \hat{Q}^r(t)\, dt\right), \tag{24}$$

where $\gamma > 0$ is a constant and $h = (h_1, h_2)'$, $h_k > 0$ for $k = 1, 2$, is a constant vector of holding costs.

Remark. We could have allowed γ and h to depend on r and then assumed some limiting positive values for these constants as $r \to \infty$. Although this more general situation can be handled by our techniques, we have chosen not to include this slight generalization here to simplify the exposition without losing much generality.

We focus here on the following parameter regime.

Assumption 3.2

$$h_1\mu_2 \geq h_2\mu_3. \tag{25}$$

We shall see that this assumption means that in the (formal) Brownian control problem associated with our sequence of parallel server systems, it is cheapest (or equally cheap with equality in Assumption 3.2) to keep the "jobs" in class 2. The opposite inequality can be treated in a similar manner, although a little more simply, for in that case there is an asymptotically optimal *static priority* control policy, namely, server 2 always gives preemptive resume priority to class 2 over class 1. We leave the details for this case to the interested reader.

In addition to the above assumptions, we make the following exponential moment assumptions which ensure that certain *large deviation estimates* hold for the renewal processes A_k^r, S_j^r associated with the interarrival and service times.

Assumption 3.3 *For $k = 1,2$, $j = 1,2,3$, and all $i \geq 1$, let (cf. (15)):*

$$u_k(i) = \frac{1}{\lambda_k}\check{u}_k(i), \quad v_j(i) = \frac{1}{\mu_j}\check{v}_j(i). \tag{26}$$

Assume that there is a non-empty open neighborhood, $\mathcal{O}$, of $0 \in \mathbb{R}$ such that for all $\ell \in \mathcal{O}$ and all $i \geq 1$,

$$\Lambda_k^a(\ell) \equiv \log \mathbf{E}\left(e^{\ell u_k(i)}\right) < \infty, \quad \textit{for } k = 1,2, \tag{27}$$

and

$$\Lambda_j^s(\ell) \equiv \log \mathbf{E}\left(e^{\ell v_j(i)}\right) < \infty, \quad \textit{for } j = 1,2,3. \tag{28}$$

Remark. Note that $\Lambda_k^a(\ell), \Lambda_j^s(\ell)$ are defined, with values in $(-\infty, \infty]$ for all values of ℓ. The above assumption guarantees that there is a neighborhood of $0 \in \mathbb{R}$ where these values are finite. Note also that since the $\{u_k(i)\}_{i=1}^\infty$ (respectively, $\{v_j(i)\}_{i=1}^\infty$) are i.i.d., the Λ_k^a, Λ_j^s do not depend on i and in fact, the above conditions hold for all i if they hold for $i = 1$.

4 Brownian Control Problem

For the convenience of the reader, in this section we summarize the formulation and analysis of the Brownian control problem associated with our parallel server problem. For more details, the reader is referred to the works of Harrison et al. [8, 11, 12, 13] and Williams [37].

Using the method proposed by Harrison et al., one arrives at the following formal Brownian control problem approximation (under diffusive scaling) to the control problem for the parallel server system. One can obtain this by formally passing to the limit in the control problem for the r^{th} parallel server system. An important assumption in this formal procedure is that in the fluid scale of (16), the allocation processes achieve the long run rates for a balanced system in the heavy traffic limit, i.e., for

$$\bar{T}^*(t) \equiv \left(t, \frac{(\lambda_1 - \mu_1)}{\mu_2}t, \frac{\lambda_2}{\mu_3}t\right), \quad t \geq 0, \tag{29}$$

we have formally as $r \to \infty$,

$$\bar{T}^r \Longrightarrow \bar{T}^*. \tag{30}$$

The two-dimensional Brownian motion $\tilde{X}$ is the formal limit in distribution (as $r \to \infty$) of the sequence of processes $\{\hat{X}^r\}$ defined by (cf. (21)–(22)):

$$\hat{X}_1^r(t) \;=\; \hat{A}_1^r(t) - \hat{S}_1^r(\bar{T}_1^r(t)) - \hat{S}_2^r(\bar{T}_2^r(t)) + r\mu_2^r\left(\frac{\lambda_1^r - \mu_1^r}{\mu_2^r} - \frac{\lambda_1 - \mu_1}{\mu_2}\right)t, \tag{31}$$

$$\hat{X}_2^r(t) \;=\; \hat{A}_2^r(t) - \hat{S}_3^r(\bar{T}_3^r(t)) + r\mu_3^r\left(\frac{\lambda_2^r}{\mu_3^r} - \frac{\lambda_2}{\mu_3}\right)t, \tag{32}$$

where the functional central limit theorem result (23), a time change theorem (together with the assumption that $\bar{T}^r \implies \bar{T}^*$), and Assumption 3.1 (v)–(vi), are used to formally derive the covariance matrix and drift for this Brownian motion. The three-dimensional control process $\tilde{Y}$ in the Brownian control problem arises as a formal limit of the normalized *deviation processes* $\hat{Y}^r \equiv r(\bar{T}^* - \bar{T}^r)$.

Definition 4.1 *(Brownian control problem)*

$$\text{minimize } \;\mathbf{E}\left(\int_0^\infty e^{-\gamma t}\, h\cdot\tilde{Q}(t)\,dt\right) \tag{33}$$

using a three-dimensional control process $\tilde{Y} = (\tilde{Y}_1, \tilde{Y}_2, \tilde{Y}_3)'$ *such that*

$$\tilde{Q}_1(t) \;\equiv\; \tilde{X}_1(t) + \mu_1\tilde{Y}_1(t) + \mu_2\tilde{Y}_2(t) \geq 0 \quad \textit{for all } t \geq 0, \tag{34}$$

$$\tilde{Q}_2(t) \;\equiv\; \tilde{X}_2(t) + \mu_3\tilde{Y}_3(t) \geq 0 \quad \textit{for all } t \geq 0, \tag{35}$$

$$\tilde{I}_1(\cdot) \;\equiv\; \tilde{Y}_1(\cdot) \quad \textit{is non-decreasing, } \tilde{I}_1(0) = 0, \tag{36}$$

$$\tilde{I}_2(\cdot) \;\equiv\; \tilde{Y}_2(\cdot) + \tilde{Y}_3(\cdot) \quad \textit{is non-decreasing, } \tilde{I}_2(0) = 0, \tag{37}$$

where $\tilde{X}$ *is a two-dimensional Brownian motion with drift* $\theta = (\theta_1, \theta_2)$*, that starts from the origin and has diagonal covariance matrix* $\mathrm{diag}(\lambda_1\alpha_1^2 + \mu_1\beta_1^2 + \beta_2^2(\lambda_1 - \mu_1), \lambda_2(\alpha_2^2 + \beta_3^2))$.

For $\tilde{Q}$ satisfying the above, let

$$\tilde{W}(t) \;=\; y\cdot\tilde{Q}(t) = y\cdot\tilde{X}(t) + \tilde{V}(t) \quad \text{for all } t \geq 0, \tag{38}$$

where

$$y \;=\; (y_1, y_2)', \qquad y_1 = 1, \quad y_2 = \frac{\mu_2}{\mu_3}, \tag{39}$$

$$\tilde{V}(t) \;\equiv\; \mu_1\tilde{I}_1(t) + \mu_2\tilde{I}_2(t) \quad \text{for all } t \geq 0. \tag{40}$$

The process $\tilde{W}$ is (up to a constant scale factor) the Brownian model analogue of workload in the original parallel server system. Following Harrison and Van Mieghem [13] (see also [10, 11, 12, 37]), one can convert the Brownian control problem to an "equivalent workload formulation" expressed in terms of $\tilde{W}$ and $\tilde{V}$. This can be solved explicitly (cf. [10, 12, 37]) and one finds that a solution of the Brownian control problem is given by setting

$$\tilde{W}^*(t) = y \cdot \tilde{X}(t) + \tilde{V}^*(t), \qquad \tilde{V}^*(t) = -\inf_{0 \le s \le t} \left(y \cdot \tilde{X}(s) \right), \tag{41}$$

i.e., $\tilde{W}^*$ is a one dimensional reflected Brownian motion (cf. [7]) driven by the one-dimensional Brownian motion $y \cdot \tilde{X}$,

$$\tilde{Q}_1^*(t) = 0, \qquad \tilde{Q}_2^*(t) = y_2^{-1} \tilde{W}^*(t), \qquad \tilde{I}_1^*(t) = 0, \qquad \tilde{I}_2^*(t) = \mu_2^{-1} \tilde{V}^*(t), \tag{42}$$

and

$$\tilde{Y}_1^*(t) = 0, \qquad \tilde{Y}_2^*(t) = -\mu_2^{-1} \tilde{X}_1(t), \qquad \tilde{Y}_3^*(t) = \mu_3^{-1} \left(\tilde{Q}_2^*(t) - \tilde{X}_2(t) \right). \tag{43}$$

The associated minimum cost is

$$J^* \equiv \mathbf{E} \left(\int_0^\infty e^{-\gamma t}\, h \cdot \tilde{Q}^*(t)\, dt \right), \tag{44}$$

with $\tilde{Q}^*$ as defined in (41)–(42). The quantity J^* is finite and can be computed explicitly as in Section 5.3 of Harrison [7].

Now, even though the Brownian control problem can be analyzed exactly, its solution does not automatically translate to a policy in the original parallel server system. In particular, since server 1 only serves class 1, there is an obvious conflict in trying to achieve zero queue length for class 1 in the heavy traffic limit ($\tilde{Q}_1^* = 0$) and zero idletime for server 1 in this limit ($\tilde{I}_1^* = 0$). Even if one can guess a reasonable policy, one would still like to be able to analyze the performance of that policy. To address the aforementioned problems, in the next section we describe a dynamic threshold policy for the original sequence of parallel server systems. We then state a theorem which shows this policy is asymptotically optimal and that under this policy the associated cost $\hat{J}^r$ for the r^{th} parallel server system converges to the minimal cost J^* for the Brownian control problem as $r \to \infty$.

5 Threshold Policy and Statement of Asymptotic Optimality

We first describe our candidate for an asymptotically optimal policy. The form of this policy is motivated by the fact that the solution of the associated Brownian control problem described in the previous section suggests that in the heavy traffic limit one should try to keep all of the work in class 2 while attempting to keep both servers busy unless there is no work in the entire system. To keep the class 1 queue length low, our policy gives priority to class 1 at server 2, except when the class 1 queue length goes below a certain threshold and then priority switches to class 2 in an attempt to prevent starvation of server 1 while there is still work in the system. Starvation of server 1 will not be totally prevented with this policy, but by allowing the threshold level to grow suitably with r, we can ensure that starvation of server 1 is a rarer and rarer event as $r \to \infty$.

Definition 5.1 *(Threshold policy) Let $c_0 > 0$ be the constant described below in the proofs of Theorem 7.2 and of uniform integrability in the proof of Theorem 5.3, and let c be any constant (independent of r) such that $c > c_0$. For each $r \geq 1$, let $L^r = [c \log r]$, the integer part of $c \log r$. In the r^{th} system, the dynamic threshold control policy is described as follows.*

(i) Server 1 operates whenever possible, or equivalently, server 1 is never idle when there are jobs in buffer 1 or at server 1.

(ii) When the number of class 1 jobs in the system exceeds the threshold value L^r, server 2 gives preemptive-resume priority to class 1 jobs over class 2 jobs. In particular, when the class 1 queue length reaches $L^r + 1$ from below, server 2 immediately suspends any work on class 2 jobs and turns to service of class 1 jobs (if it has a suspended class 1 job, it resumes work on this, otherwise the server starts work on the next job in buffer 1). Similarly, when the class 1 queue length reaches a level $\leq L^r$ from above L^r, server 2 suspends its service of class 1 jobs and turns to service of class 2 jobs. If there are no class 2 jobs for server 2 to serve, the server idles until a new class 2 job arrives or the class 1 queue length reaches $L^r + 1$ again.

We let $T^{r,*}$ denote the allocation processes associated with use of the above policy in the r^{th} system.

Remark. Note that it is possible for server 1 to be idle at time t even though $Q_1^r(t) > 0$ if there is a class 1 job "in process" at server 2. However, such times will be inconsequential in the heavy traffic limit.

Remark. For our method of proof to work, c must be sufficiently large. In the proofs of Theorem 7.2 and of uniform integrability in the proof of Theorem 5.3, a means for determining a value c_0 is described such that our method works provided $c > c_0$. This value is determined from several applications of large deviation estimates for the renewal processes associated with the interarrival and service time sequences, cf. Assumption 3.3. We have not attempted to give a concise formula for c_0 nor to optimize its value, since the relevant fact is that a threshold of size $[c \log r]$ works for c sufficiently large and the order of this threshold is the smallest for which our proof works. We did not investigate whether a threshold of smaller order could be used and asymptotic optimality still achieved, since we sought to develop a method that could be readily applied to more complex multiserver systems. The reader interested in an analysis of the effects of different threshold sizes for some dynamic scheduling problems is referred to the recent work of Teh [32] in this direction.

Remark. We wish to emphasize that our proposed policy is only one of many possible asymptotically optimal policies. We have focussed on our policy because it is intuitively appealing and easy to describe. Variations of this policy are certainly possible. For instance, our method of proof would work if the threshold level satisfied $L^r \geq c \log r$ for $c > c_0$ and $L^r = o(r)$ as $r \to \infty$, e.g., $L^r = (\log r)^{1+\varepsilon}$ for any $\varepsilon > 0$ will do. We have used $L^r = [c \log r]$ as this is the smallest order threshold for which our proof works. In addition, to reduce "chattering" back and forth across a single threshold, one could introduce a second threshold at $2L^r$ and an associated "hysteretic policy" (cf. [31]) such that the additional help of server 2 is only turned on when the class 1 queue length exceeds this second threshold and that help is turned off when the class 1 queue length returns to the level L^r or below.

Remark. Although the above policy allows for preemption, there is a corresponding threshold policy without preemption that we conjecture has the same behavior in the heavy traffic limit, since in that regime a single job (in suspension or not) should not impact the asymptotic behavior of the system.

In Sections 7–8 we show that the following form of state-space collapse holds under this sequence of threshold policies in the heavy traffic limit (as $r \to \infty$).

Theorem 5.2 *Consider the sequence of parallel server systems indexed by r, where the r^{th} system operates under the threshold policy $T^{r,*}$ described above. Then the associated normalized queue length and idletime processes satisfy*

$$(\hat{Q}_1^r, \hat{Q}_2^r, \hat{I}_1^r, \hat{I}_2^r) \Longrightarrow (\mathbf{0}, \tilde{Q}_2^*, \mathbf{0}, \tilde{I}_2^*) \quad \text{as} \quad r \to \infty,$$

where $\mathbf{0}$ denotes the one-dimensional process that is identically zero, $\tilde{Q}_2^*$ is a one-dimensional reflected Brownian motion that starts from zero, has drift $\frac{\mu_3}{\mu_2}\theta_1 + \theta_2$ (where θ_1, θ_2 are defined in

Assumption 3.1) and has variance parameter $(\frac{\mu_3}{\mu_2})^2(\lambda_1\alpha_1^2+\mu_1\beta_1^2+(\lambda_1-\mu_1)\beta_2^2)+\lambda_2(\alpha_2^2+\beta_3^2)$, *i.e.,* $\tilde{Q}_2^*$ *is described by (41)–(42), and* $\tilde{I}_2^*$ *is a specific multiple of the local time at the origin of* $\tilde{Q}_2^*$, *as defined in (41)–(42).*

Recall the definition of J^* from (44). The following is proved in Section 9 using Theorem 5.2. It shows that J^* is the best that one can achieve asymptotically and that this asymptotically minimal cost is achieved by the sequence of dynamic theshold policies $\{T^{r,*}\}$. Thus we conclude that our sequence of threshold policies $\{T^{r,*}\}$ is asymptotically optimal.

Theorem 5.3 *Suppose that* $\{T^r\}$ *is any sequence of scheduling control policies (one for each member of the sequence of parallel server systems). Then*

$$\liminf_{r\to\infty} \hat{J}^r(T^r) \geq J^* = \lim_{r\to\infty} \hat{J}^r(T^{r,*}), \tag{45}$$

and $J^* < \infty$.

Remark. The notion of asymptotic optimality used here is also used for example in Puhalskii-Reiman [29].

6 Outline of the Proof

A key element in the proof of Theorem 5.2 is to first show (cf. Theorem 7.1) that under the threshold policy,

$$(\hat{Q}_1^r, \hat{I}_1^r) \Longrightarrow (\mathbf{0},\mathbf{0}) \quad \text{as } r\to\infty. \tag{46}$$

The idea behind this is that under the threshold control $T^{r,*}$, once Q_1^r has reached the threshold level L^r, the normalized class 1 queue length process $\hat{Q}_1^r$ "keeps close" to the normalized threshold level $\hat{L}^r \equiv L^r/r$, since when it is above this level, it is driven down towards the level at an average rate of $(\mu_1^r+\mu_2^r-\lambda_1^r)r$ and when it is below the level, it is driven up towards the level at an average rate of $(\lambda_1^r-\mu_1^r)r$. Indeed, large deviation estimates for the renewal processes used in defining the model (cf. Assumption 3.3) are used to show (cf. Theorem 7.2) that the probability that, on any given compact time interval, $\hat{Q}_1^r$ deviates by at least $\hat{L}^r - r^{-1}$ from the threshold level $\hat{L}^r$, goes to zero as $r\to\infty$. The result (46) follows from this. It can then be shown using the model equations for queue length and idletime (cf. Section 2), and the fact that under $T^{r,*}$ server 2 cannot be idle if there are jobs in class 2, that the fluid scaled allocations $\bar{T}^{r,*}(t) \equiv r^{-2}T^{r,*}(r^2t)$ associated with $T^{r,*}$ satisfy

$$\bar{T}^{r,*} \Longrightarrow \bar{T}^* \quad \text{as } r \to \infty, \tag{47}$$

where $\bar{T}^*$ is defined in (29), cf. Lemma 8.1. One can then combine the above to show (cf. Section 8) that

$$(\hat{Q}_2^r, \hat{I}_2^r) \Longrightarrow (\tilde{Q}_2^*, \tilde{I}_2^*) \quad \text{as } r \to \infty, \tag{48}$$

where $\tilde{Q}_2^*, \tilde{I}_2^*$ are given by (41)–(42).

For the proof of Theorem 5.3, we first show (cf. Lemma 9.3) that for any subsequence that achieves the "lim inf" on the left side of (45) as a limit and for which the "lim inf" is finite, the fluid level asymptotic behavior described in (30) must hold along the subsequence. This, together with a pathwise lower bound for $h^r \cdot \hat{Q}^r$, where $h^r = (h_1, h_2\mu_2^r\mu_3/(\mu_2\mu_3^r))'$, allows us to establish the inequality on the left side of (45). The equality on the right side of (45) follows from Theorem 5.2, after showing that a certain uniform integrability condition holds.

7 Residual Process

The main result of this section is the following theorem (cf. (46)).

Theorem 7.1 *Consider the sequence of parallel server systems indexed by r, where the r^{th} system operates under the threshold policy described in Definition 5.1. Then*

$$\left(\hat{Q}_1^r, \hat{I}_1^r\right) \Longrightarrow (\mathbf{0}, \mathbf{0}) \quad \textit{as } r \to \infty. \tag{49}$$

Throughout this section, it is assumed that in the r^{th} parallel server system we use the allocation process $T^{r,*}$ associated with the threshold policy described in Definition 5.1. To simplify notation, here we shall simply write T^r in place of $T^{r,*}$, since no other policy is considered in this section. The associated queue length and idletime processes will be denoted by Q^r, I^r, respectively.

Key to our proof of Theorem 7.1 is the behavior of what we call the *residual process* R^r defined by

$$R^r(t) = Q_1^r(t) - L^r, \quad t \geq 0, \tag{50}$$

where L^r is the threshold described in Definition 5.1. The idea here is to move the center of one's attention to the threshold and to think of Q_1^r as reaching the threshold level L^r relatively quickly and then "chattering" back and forth across this threshold, not frequently deviating far from it, so that Q_1^r rarely again goes as low as the level one or as high as the level $2L^r - 1$. When translated into the behavior of R^r, this means that we seek to show that once R^r reaches zero, it chatters back and forth across its zero level and rarely deviates more than $\pm(L^r - 1)$ from this level. In particular, the following is the main technical result of this section.

Theorem 7.2 *Let $\tau_0^r = \inf\{t \geq 0 : Q_1^r(t) \geq L^r\}$. Then, for each $t \geq 0$ and $\varepsilon > 0$,*

$$\mathbf{P}\left(I_1^r(\tau_0^r) \geq r\varepsilon\right) \quad \to \quad 0 \qquad \text{as } r \to \infty, \tag{51}$$

$$\mathbf{P}\left(\sup_{\tau_0^r \leq s \leq r^2 t} |R^r(s)| \geq L^r - 1\right) \quad \to \quad 0 \qquad \text{as } r \to \infty. \tag{52}$$

Here we have used the convention in (51) that $I_1^r(\tau_0^r) = \lim_{t\to\infty} I_1^r(t)$ on $\{\tau_0^r = +\infty\}$, and in (52) that the supremum over an empty set is defined to equal $-\infty$.

For the proof of (52), we need to establish some preliminary results concerning the properties of arrival and service processes stopped at certain levels, so that we can apply the results of Appendix A to shifted versions of these processes. We establish these preliminary results here before turning to the proofs of Theorems 7.2 and 7.1.

Definition 7.3 *For each $r \geq 1$, let $\tau_1^r = \inf\{t \geq \tau_0^r : R^r(t) = 1\}$, $\tau_2^r = \inf\{t > \tau_1^r : R^r(t) \leq 0\}$ and define recursively $\tau_{2n-1}^r = \inf\{t > \tau_{2n-2}^r : R^r(t) = 1\}$, $\tau_{2n}^r = \inf\{t > \tau_{2n-1}^r : R^r(t) \leq 0\}$, for $n = 2, 3, \ldots$. We call $[\tau_{2n-1}^r, \tau_{2n}^r)$ the n^{th} "up" excursion interval for R^r.*

Let $\mathbb{N}_\infty = \mathbb{N} \cup \{\infty\}$. Consider $\mathbb{N}_\infty^3$ to be partially ordered by componentwise inequality, i.e., $(n_1, n_2, n_3) \leq (m_1, m_2, m_3)$ if and only if $n_1 \leq m_1$, $n_2 \leq m_2$, and $n_3 \leq m_3$. Recall the definition of the cumulative interarrival time process for class 1, ξ_1^r, and the cumulative service time processes for activities 1 and 2, η_1^r, η_2^r, in the r^{th} system. For each $(p, q, s) \in \mathbb{N}_\infty^3$, let

$$\mathcal{F}_{pqs}^r = \sigma\{\xi_1^r(\cdot \wedge (p+1)), \eta_1^r(\cdot \wedge (q+1)), \eta_2^r(\cdot \wedge (s+1))\} \vee \mathcal{N}, \tag{53}$$

where $\mathcal{N}$ denotes the collection of $\mathbf{P}$-null sets in the complete probability space $(\Omega, \mathcal{F}, P)$. Then, $\{\mathcal{F}_{pqs}^r : (p, q, s) \in \mathbb{N}_\infty^3\}$ is a multiparameter filtration (cf. Ethier and Kurtz [6], p. 85).

Definition 7.4 *A (multiparameter) stopping time relative to $\{\mathcal{F}_{pqs}^r : (p, q, s) \in \mathbb{N}_\infty^3\}$ is a random variable $\mathcal{T}$ taking values in $\mathbb{N}_\infty^3$ such that*

$$\{\mathcal{T} = (p, q, s)\} \in \mathcal{F}_{pqs}^r \quad \text{for all } (p, q, s) \in \mathbb{N}_\infty^3. \tag{54}$$

The σ-algebra associated with such a stopping time $\mathcal{T}$ is

$$\mathcal{F}^r_{\mathcal{T}} = \{B \in \mathcal{F} : B \cap \{\mathcal{T} = (p,q,s)\} \in \mathcal{F}^r_{pqs} \text{ for all } (p,q,s) \in \mathbb{N}^3_\infty\}. \tag{55}$$

Lemma 7.5 *For each $r \geq 1$ and each $n \geq 1$,*

$$\mathcal{T}^r_n \equiv (A^r_1(\tau^r_{2n-1}), S^r_1(T^r_1(\tau^r_{2n-1})), S^r_2(T^r_2(\tau^r_{2n-1}))) \tag{56}$$

is a (multiparameter) stopping time relative to the filtration $\{\mathcal{F}^r_{pqs} : (p,q,s) \in \mathbb{N}^3_\infty\}$, where we adopt the convention that each of $A^r_1(\cdot)$, $S^r_1(T^r_1(\cdot))$, $S^r_2(T^r_2(\cdot))$, takes the value ∞ when evaluated at the time ∞.

Proof. This lemma can be proved in a similar manner to Lemma 8.3 in [36]. □

Lemma 7.6 *Let $\mathcal{T} = (\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3)$ be a (multiparameter) stopping time relative to the filtration $\{\mathcal{F}^r_{pqs} : (p,q,s) \in \mathbb{N}^3_\infty\}$. In the following, for notational convenience, we make the convention that each of $u^r_1(\cdot), v^r_1(\cdot), v^r_2(\cdot)$ takes the value ∞ when its argument takes the value ∞. Then,*

$$(u^r_1(\mathcal{T}_1 + 1), v^r_1(\mathcal{T}_2 + 1), v^r_2(\mathcal{T}_3 + 1)) \in \mathcal{F}^r_{\mathcal{T}}, \tag{57}$$

and on $\{\mathcal{T} \in \mathbb{N}^3\}$, the conditional distribution of $\{(u^r_1(\mathcal{T}_1+n), v^r_1(\mathcal{T}_2+n), v^r_2(\mathcal{T}_3+n)),\ n = 2,3,\ldots\}$ given $\mathcal{F}^r_{\mathcal{T}}$ is the same as the (unconditioned) distribution of the original family of i.i.d. random variables $\{(u^r_1(n), v^r_1(n), v^r_2(n)), n = 1,2,\ldots\}$.

Proof. For the proof of (57), let $\overline{\mathbb{R}}_+$ be the extended positive half line that is the compactification of $\mathbb{R}_+$ with the addition of the point at infinity. Then for any Borel set $B \subset (\overline{\mathbb{R}}_+)^3$ and $(p,q,s) \in \mathbb{N}^3_\infty$,

$$\begin{aligned} &\{(u^r_1(\mathcal{T}_1 + 1), v^r_1(\mathcal{T}_2 + 1), v^r_2(\mathcal{T}_3 + 1)) \in B\} \cap \{\mathcal{T} = (p,q,s)\} \\ &\quad = \{(u^r_1(p+1), v^r_1(q+1), v^r_2(s+1)) \in B\} \cap \{\mathcal{T} = (p,q,s)\}, \end{aligned}$$

where the last line is in $\mathcal{F}^r_{pqs}$, by the definition of that σ-algebra and the fact that $\mathcal{T}$ is a stopping time. This is even true if some or all of p,q,s take the value ∞, since for such values, the corresponding variables from among $u^r_1(p+1), v^r_1(q+1), v^r_2(s+1)$ are deterministic. Hence, (57) holds.

Now, for Borel sets $B_i \in (\overline{\mathbb{R}}_+)^3$, $i = 1, 2, ...$, and $C \in \mathcal{F}^r_{\mathcal{T}}$, we have

$$
\begin{aligned}
& \mathbf{P}\left(\{(u_1^r(\mathcal{T}_1+n), v_1^r(\mathcal{T}_2+n), v_2^r(\mathcal{T}_3+n)) \in B_{n-1}, n = 2, 3, \ldots\} \cap C \cap \{\mathcal{T} \in \mathbb{N}^3\}\right) \\
= & \sum_{(p,q,s)\in\mathbb{N}^3} \mathbf{P}\left(\{(u_1^r(p+n), v_1^r(q+n), v_2^r(s+n)) \in B_{n-1}, n = 2, 3, \ldots\} \cap C \cap \{\mathcal{T} = (p, q, s)\}\right) \\
= & \sum_{(p,q,s)\in\mathbb{N}^3} \mathbf{P}\left((u_1^r(p+n), v_1^r(q+n), v_2^r(s+n)) \in B_{n-1}, n = 2, 3, \ldots\right) \cdot \mathbf{P}\left(C \cap \{\mathcal{T} = (p, q, s)\}\right) \\
= & \mathbf{P}\left((u_1^r(i), v_1^r(i), v_2^r(i)) \in B_i, i = 1, 2, \ldots\right) \cdot \mathbf{P}(C \cap \{\mathcal{T} \in \mathbb{N}^3\}),
\end{aligned}
$$

where the last two lines follow from the fact that $C \cap \{\mathcal{T} = (p, q, s)\} \in \mathcal{F}^r_{pqs}$ by the definition of $\mathcal{F}^r_{\mathcal{T}}$, and the fact that the following sequences are i.i.d. and independent of one another: $\{u_1^r(j), j = 1, 2, \ldots\}$, $\{v_1^r(j), j = 1, 2, \ldots\}$, $\{v_2^r(j), j = 1, 2, \ldots\}$. The claim following (57) is then immediate. □

Proof of Theorem 7.2. We first establish (51). Note that for $s \le \tau_0^r$, $T_2^r(s) = 0$, $T_1^r(s) = \int_0^s 1_{\{Q_1^r(u)>0\}} du \le s$, and so

$$
Q_1^r(s) = A_1^r(s) - S_1^r(T_1^r(s)) \ \ge \ A_1^r(s) - S_1^r(s), \tag{58}
$$

$$
I_1^r(s) \ \le \ s. \tag{59}
$$

Let $0 < \tilde{\varepsilon} < \min\left(\frac{\lambda_1-\mu_1}{8}, \frac{\mu_1}{2}, \frac{\lambda_1}{2}\right)$ and $r_{\tilde{\varepsilon}} \ge 1$ such that for all $r \ge r_{\tilde{\varepsilon}}$, $\lambda_1^r - \mu_1^r \ge \frac{\lambda_1-\mu_1}{2}$, $|\lambda_1 - \lambda_1^r| < \tilde{\varepsilon}$, and $|\mu_1 - \mu_1^r| < \tilde{\varepsilon}$. Then, for $t^r = 8L^r/(\lambda_1 - \mu_1)$ and $r \ge r_{\tilde{\varepsilon}}$,

$$
\begin{aligned}
& \mathbf{P}\left(I_1^r(\tau_0^r) > t^r\right) \\
& \le \mathbf{P}\left(\tau_0^r > t^r\right) \\
& \le \mathbf{P}\left(A_1^r(t^r) - S_1^r(t^r) \le L^r\right) \\
& \le \mathbf{P}\Big(A_1^r(t^r) \ge (\lambda_1^r - \tilde{\varepsilon})t^r, \ S_1^r(t^r) \le (\mu_1^r + \tilde{\varepsilon})t^r, \ ((\lambda_1^r - \tilde{\varepsilon}) - (\mu_1^r + \tilde{\varepsilon}))t^r \le L^r\Big) \\
& \qquad + \mathbf{P}\Big(A_1^r(t^r) < (\lambda_1^r - \tilde{\varepsilon})t^r\Big) + \mathbf{P}\Big(S_1^r(t^r) > (\mu_1^r + \tilde{\varepsilon})t^r\Big).
\end{aligned} \tag{60}
$$

Now, by the choice of t^r, for $r \ge r_{\tilde{\varepsilon}}$,

$$
(\lambda_1^r - \mu_1^r - 2\tilde{\varepsilon})t^r \ge \left(\frac{\lambda_1 - \mu_1}{4}\right) t^r > L^r \tag{61}
$$

and the first probability in the last two lines of (60) is zero. Furthermore, by large deviation estimates similar to those given in Appendix A (in particular, one uses (184) with $\zeta^r(1) = 0$, $\delta^r = 0$, and $\chi^r t^r + 1$ in place of $\chi^r t$ in the right member there, (188), (181) and (190)), provided

$t^r > 2/\tilde{\varepsilon}$ we have

$$\mathbf{P}(A_1^r(t^r) < (\lambda_1^r - \tilde{\varepsilon})t^r) \leq \exp\left(-((\lambda_1^r - \tilde{\varepsilon})t^r + 1)\,\Lambda_1^{a,*}\left(\frac{1}{\lambda_1}\left(\frac{1}{1-\frac{\tilde{\varepsilon}}{3\lambda_1}}\right)\right)\right), \tag{62}$$

$$\mathbf{P}(S_1^r(t^r) > (\mu_1^r + \tilde{\varepsilon})t^r) \leq \exp\left(-((\mu_1^r + \tilde{\varepsilon})t^r - 1)\,\Lambda_1^{s,*}\left(\frac{1}{\mu_1}\left(\frac{1}{1+\frac{\tilde{\varepsilon}}{3\mu_1}}\right)\right)\right), \tag{63}$$

where $\Lambda_1^{a,*}, \Lambda_1^{s,*}$ are the Legendre-Fenchel transforms of Λ_1^a, Λ_1^s, respectively. Since the values of $\Lambda_1^{a,*}$ and $\Lambda_1^{s,*}$ appearing above are strictly positive and $t^r \to \infty$ as $r \to \infty$, it follows that

$$\mathbf{P}(I_1^r(\tau_0^r) > t^r) \to 0 \quad \text{as } r \to \infty, \tag{64}$$

and hence (51) holds since $t^r = o(r)$ as $r \to \infty$.

Since (52) clearly holds for $t = 0$, we fix $t > 0$. For the proof of (52), we will show that

$$\mathbf{P}\left(\sup_{\tau_0^r \leq s \leq r^2 t} R^r(s) \geq L^r - 1\right) \to 0 \quad \text{as } r \to \infty. \tag{65}$$

The proof of the other half, namely that

$$\mathbf{P}\left(\inf_{\tau_0^r \leq s \leq r^2 t} R^r(s) \leq -(L^r - 1)\right) \to 0 \quad \text{as } r \to \infty \tag{66}$$

is very similar. The details of this half are left to the reader, but see the next paragraph for some comments on the steps in the proof.

In outline, for (65), the proof involves showing the steps (i)–(v) below for $L^r = [c \log r]$ and c sufficiently large.

(i) The number of complete excursions that R^r makes above zero in an interval of length $r^2 t$ is bounded by the number of arrivals to class 1 during that time, which in turn is $O(r^2 t)$ with arbitrarily high probability for r sufficiently large, cf. (76).

(ii) For $0 < \tilde{\varepsilon} < \min\left(\frac{\mu_1+\mu_2-\lambda_1}{8}, \frac{\lambda_1}{2}, \frac{\mu_1}{2}, \frac{\mu_2}{2}\right)$ and r sufficiently large, for any time interval of length $s^r = O(L^r)$, the number of new arrivals to class 1 during such an interval is bounded above by $(\lambda_1^r + \tilde{\varepsilon})s^r$ with a probability that is at least as large as $1 - C_1 \exp(-C_2 L^r)$, where C_1, C_2 are positive constants that may depend on $\tilde{\varepsilon}$, but that do not depend on r.

(iii) For $\tilde{\varepsilon}$ as in (ii) and r sufficiently large, for any time interval of length $s^r = O(L^r)$ that begins at or before $r^2 t$ and during which class 1 jobs are being continually processed by activity $j \in \{1, 2\}$, the number of departures from class 1 during the interval due to such processing is at least $(\mu_j^r - \tilde{\varepsilon})s^r$ with a probability that is at least as large as $1 - C_1 \exp(-C_2 L^r)$, where C_1, C_2 are positive constants that may depend on $\tilde{\varepsilon}$, but that do not depend on r.

(iv) Properties (ii) and (iii) imply that with a probability at least as large as $1 - C_3 \exp(-C_2 L^r)$ (where C_3 is a positive constant independent of r), any given excursion of R^r above zero that starts before time $r^2 t$ lasts for at most $s^r \equiv (L^r - 3)/(\lambda_1^r + \tilde{\varepsilon})$ units of time, and, during the excursion, R^r does not reach the level $L^r - 1$.

(v) Properties (i) and (iv) imply that provided c is sufficiently large in $L^r = [c \log r]$, with a probability that can be made arbitrarily close to 1 for r sufficiently large, R^r does not reach the level $L^r - 1$ during any of the excursions above zero that start before $r^2 t$.

The details of these steps are provided in the following paragraphs. For the proof of (66), the steps need to be modified as follows. In (ii) and (iii), $0 < \tilde{\varepsilon} < \min\left(\frac{\lambda_1 - \mu_1}{8}, \frac{\lambda_1}{2}, \frac{\mu_1}{2}, \frac{\mu_2}{2}\right)$, the estimate in (ii) is replaced by an estimate of the probability that the number of arrivals that can occur in an interval of length s^r that starts at or before $r^2 t$ is bounded below by $(\lambda_1^r - \tilde{\varepsilon})s^r$, the estimate in (iii) is replaced by an estimate of the probability that the number of departures from class 1 that can occur due to continuous processing by activity 1 during an interval of length s^r is no more than $(\mu_1^r + \tilde{\varepsilon})s^r$, in (iv) one considers excursions of R^r below zero and $s^r = (L^r - 3)/(\mu_1^r + \tilde{\varepsilon})$, and in (v) one shows that with a probability that is arbitrarily close to 1 for r sufficiently large, R^r does not reach the level $-(L^r - 1)$ during any excursion below zero that starts before $r^2 t$. Since this argument rests on showing that R^r does not reach $-(L^r - 1)$ (and hence Q^r does not reach zero) during the time interval $[\tau_0^r, r^2 t]$ with probability tending to 1 as $r \to \infty$, one does not need to consider the effect of idletime.

The constant c_0 referred to in Definition 5.1 of the threshold policy can be chosen to equal the maximum of those determined such that (65), (66) and the estimates used to establish uniform integrability in the proof of Theorem 5.3 all hold. (Below we only describe how to determine a value that works for (65), but a similar estimate holds for (66) and these estimates with a sufficiently large value of c_0 are used to establish the uniform integrability in the proof of Theorem 5.3).

Fix $r \geq 1$. Consider the n^{th} "up" excursion interval for R^r, which begins at τ_{2n-1}^r. Note that if $\tau_{2n-1}^r < \infty$, then a new arrival to class 1 occurs at τ_{2n-1}^r and over the time interval $[\tau_{2n-1}^r, \tau_{2n}^r)$, both servers will be processing class 1 jobs only. We define shifted processes as follows. For each

$n \geq 1$, on $\{\tau^r_{2n-1} < \infty\}$, define for each $s \geq 0$,

$$
\begin{aligned}
(67)\quad A_1^{r,n}(s) &= A_1^r(\tau^r_{2n-1}+s) - A_1^r(\tau^r_{2n-1}),\\
(68)\quad S_j^{r,n}(s) &= S_j^r(T_j^r(\tau^r_{2n-1})+s) - S_j^r(T_j^r(\tau^r_{2n-1})), \quad j=1,2,\\
(69)\quad \breve{A}_1^{r,n}(s) &= \sup\{m \geq 0 : \xi_1^r(A_1^r(\tau^r_{2n-1})+m) - \xi_1^r(A_1^r(\tau^r_{2n-1})) \leq s\},\\
(70)\quad \breve{S}_j^{r,n}(s) &= \sup\{m \geq 0 : \eta_j^r(S_j^r(T_j^r(\tau^r_{2n-1}))+m) - \eta_j^r(S_j^r(T_j^r(\tau^r_{2n-1}))) \leq s\}, \quad j=1,2,
\end{aligned}
$$

and for concreteness, on $\{\tau^r_{2n-1} = \infty\}$, define $A_1^{r,n}, S_1^{r,n}, S_2^{r,n}, \breve{A}_1^{r,n}, \breve{S}_1^{r,n}, \breve{S}_2^{r,n}$, to be identically zero. Then on $\{\tau^r_{2n-1} < \infty\}$, for $0 \leq s \leq \tau^r_{2n} - \tau^r_{2n-1}$ we have

$$
(71)\qquad R^r(\tau^r_{2n-1}+s) = 1 + A_1^{r,n}(s) - S_1^{r,n}(s) - S_2^{r,n}(s),
$$

and, taking account of the fact that a new arrival occurs at $\tau^r_{2n-1} < \infty$ and a job may have been partially served by activity $j \in \{1,2\}$ at $\tau^r_{2n-1} < \infty$, we also have

$$
(72)\qquad A_1^{r,n}(s) = \breve{A}_1^{r,n}(s) \quad \text{and} \quad S_j^{r,n}(s) \geq \breve{S}_j^{r,n}(s).
$$

Let $0 < \tilde{\varepsilon} < \min\left(\frac{\mu_1+\mu_2-\lambda_1}{8}, \frac{\lambda_1}{2}, \frac{\mu_1}{2}, \frac{\mu_2}{2}\right)$ and let $r_{\tilde{\varepsilon}} \geq 1$ be chosen large enough such that all of the following hold for $r \geq r_{\tilde{\varepsilon}}$:

$$
\begin{aligned}
(73)\quad & r^2 t > \frac{2}{\tilde{\varepsilon}}, \qquad L^r > 3,\\
(74)\quad & |\lambda_1 - \lambda_1^r| < \tilde{\varepsilon}, \qquad |\mu_1 - \mu_1^r| < \tilde{\varepsilon}, \qquad |\mu_2 - \mu_2^r| < \tilde{\varepsilon},\\
(75)\quad & \mu_1^r + \mu_2^r - \lambda_1^r - 3\tilde{\varepsilon} > \frac{\mu_1+\mu_2-\lambda_1}{2}.
\end{aligned}
$$

Henceforth in this proof we only consider $r \geq r_{\tilde{\varepsilon}}$. Let $n^r = [(\lambda_1^r + \tilde{\varepsilon})r^2 t] + 1$. Then since each excursion of R^r from zero to one requires an arrival to class 1, using the results of Appendix A (cf. (192)) we have the following estimate on the probability that there are at least $n^r - 1$ complete and one additional partial or complete "up" excursion in $[0, r^2 t]$:

$$
\begin{aligned}
(76)\qquad \mathbf{P}(\tau^r_{2n^r-1} \leq r^2 t) &\leq \mathbf{P}\left(A_1^r(r^2 t) \geq n^r\right)\\
&= \mathbf{P}\left(A_1^r(r^2 t) > (\lambda_1^r + \tilde{\varepsilon})r^2 t\right)\\
&\leq \exp\left(-(\lambda_1 r^2 t - 1)\Lambda_1^{a,*}\left(\frac{1}{\lambda_1}\left(\frac{1}{1+\frac{\tilde{\varepsilon}}{3\lambda_1}}\right)\right)\right).
\end{aligned}
$$

Now,

$$
\begin{aligned}
(77) \qquad & \mathbf{P}(R^r(s) \geq L^r - 1 \text{ some } s \in [0, r^2t]) \\
\leq\ & \mathbf{P}\left(\tau^r_{2n^r-1} \leq r^2t\right) + \mathbf{P}\left(\tau^r_{2n^r-1} > r^2t,\ R^r(s) \geq L^r - 1 \text{ some } s \in [0, r^2t]\right) \\
\leq\ & \mathbf{P}\left(\tau^r_{2n^r-1} \leq r^2t\right) \\
& + \sum_{n=1}^{n^r} \mathbf{P}(R^r(s) \geq L^r - 1 \text{ some } s \in (\tau^r_{2n-1}, \tau^r_{2n}),\ \tau^r_{2n-1} \leq r^2t).
\end{aligned}
$$

Let $s^r = (L^r - 3)/(\lambda^r_1 + \tilde{\varepsilon})$ and for each positive integer n, let

$$
\Upsilon^{r,n} = \left\{ A^{r,n}_1(s^r) \leq (\lambda^r_1 + \tilde{\varepsilon})s^r, S^{r,n}_1(s^r) \geq (\mu^r_1 - \tilde{\varepsilon})s^r, S^{r,n}_2(s^r) \geq (\mu^r_2 - \tilde{\varepsilon})s^r, \tau^r_{2n-1} \leq r^2t \right\}.
$$

Then,

$$
\begin{aligned}
(78) \qquad & \mathbf{P}(R^r(s) \geq L^r - 1 \text{ some } s \in (\tau^r_{2n-1}, \tau^r_{2n}),\ \tau^r_{2n-1} \leq r^2t) \\
\leq\ & \mathbf{P}((\Upsilon^{r,n})^c,\ \tau^r_{2n-1} \leq r^2t) + \mathbf{P}(R^r(s) \geq L^r - 1 \text{ some } s \in (\tau^r_{2n-1}, \tau^r_{2n}), \Upsilon^{r,n}),
\end{aligned}
$$

where B^c is the complement of B in Ω. Now, on $\Upsilon^{r,n}$

$$
\begin{aligned}
(79) \qquad 1 + A^{r,n}_1(s^r) - S^{r,n}_1(s^r) - S^{r,n}_2(s^r) &\leq 1 + (\lambda^r_1 + \tilde{\varepsilon} - \mu^r_1 + \tilde{\varepsilon} - \mu^r_2 + \tilde{\varepsilon})s^r \\
&\leq 1 - \frac{(\mu_1 + \mu_2 - \lambda_1)}{2} \cdot \frac{(L^r - 3)}{(\lambda^r_1 + \tilde{\varepsilon})} < 0
\end{aligned}
$$

for all $r \geq r'_{\tilde{\varepsilon}}$ where $r'_{\tilde{\varepsilon}} \geq r_{\tilde{\varepsilon}}$ can be chosen independent of n and we have used the facts that $L^r \to \infty$ and $\mu_1 + \mu_2 - \lambda_1 > 0$. It then follows from the representation (71) of R^r on $(\tau^r_{2n-1}, \tau^r_{2n})$ when $\tau^r_{2n-1} < \infty$ that

$$
(80) \qquad \tau^r_{2n} - \tau^r_{2n-1} < s^r \text{ on } \Upsilon^{r,n} \text{ for } r \geq r'_{\tilde{\varepsilon}}.
$$

Furthermore, on $\Upsilon^{r,n}$ for $0 \leq s < s^r$,

$$
(81) \qquad 1 + A^{r,n}(s) - S^{r,n}_1(s) - S^{r,n}_2(s) \leq 1 + (\lambda^r_1 + \tilde{\varepsilon})s^r \leq L^r - 2,
$$

and so on combining this with (71) and (80) we have that for all $r \geq r'_{\varepsilon}$, on $\Upsilon^{r,n}$,

$$
(82) \qquad R^r(s) < L^r - 1 \text{ for } s \in (\tau^r_{2n-1}, \tau^r_{2n}).
$$

Thus the last probability in (78) is zero for all such r. It remains to estimate

$$
\begin{aligned}
(83) \qquad \mathbf{P}(\tau^r_{2n-1} \leq r^2t, (\Upsilon^{r,n})^c) \leq\ & \mathbf{P}(A^{r,n}_1(s^r) > (\lambda^r_1 + \tilde{\varepsilon})s^r, \tau^r_{2n-1} \leq r^2t) \\
& + \mathbf{P}(S^{r,n}_1(s^r) < (\mu^r_1 - \tilde{\varepsilon})s^r, \tau^r_{2n-1} \leq r^2t) \\
& + \mathbf{P}(S^{r,n}_2(s^r) < (\mu^r_2 - \tilde{\varepsilon})s^r, \tau^r_{2n-1} \leq r^2t).
\end{aligned}
$$

Recall the definition of $\mathcal{T}_n^r$ from Lemma 7.5. Now, the set $\{\tau_{2n-1}^r < \infty\}$ is contained in $\{\mathcal{T}_n^r \in \mathbb{N}^3\}$ by the convention adopted at the end of Section 2.1 that the arrival and service renewal processes are finite everywhere on Ω, and so

$$\mathbf{P}\left(A_1^{r,n}(s^r) > (\lambda_1^r + \tilde{\varepsilon})\, s^r, \tau_{2n-1}^r \le r^2 t\right) \quad \le \quad \mathbf{E}\left(1_{\{\mathcal{T}_n^r \in \mathbb{N}^3\}} \mathbf{P}\left(A_1^{r,n}(s^r) > (\lambda_1^r + \tilde{\varepsilon}) s^r \,|\, \mathcal{F}_{\mathcal{T}_n^r}^r\right)\right).$$

On $\{\mathcal{T}_n^r \in \mathbb{N}^3\}$, by (69), (72), $A_1^{r,n}$ is the counting process defined from the sequence of interarrival time random variables $\{u_1^r(A_1^r(\tau_{2n-1}^r)+i), i = 1, 2, \ldots\}$ where by Lemmas 7.5 and 7.6, on $\{\mathcal{T}_n^r \in \mathbb{N}^3\}$, the conditional distribution of this sequence given $\mathcal{F}_{\mathcal{T}_n^r}^r$ is equal to that of a sequence of strictly positive independent random variables where the members indexed by $i = 2, 3, \ldots$ are identically distributed with the same distribution as $u_1^r(1)$. Hence we may apply the results of Appendix A (cf. (192)) to conclude that a.s. on $\{\mathcal{T}_n^r \in \mathbb{N}^3\}$, for $s^r > \frac{2}{\tilde{\varepsilon}}$,

$$\begin{aligned}
(84) \quad \mathbf{P}\left(A_1^{r,n}(s^r) > (\lambda_1^r + \tilde{\varepsilon})\, s^r \,|\, \mathcal{F}_{\mathcal{T}_n^r}^r\right) \quad &\le \quad \exp\left(-\Big((\lambda_1^r + \tilde{\varepsilon})\, s^r - 1\Big)\Lambda_1^{a,*}\left(\frac{1}{\lambda_1}\left(\frac{1}{1 + \frac{\tilde{\varepsilon}}{3\lambda_1}}\right)\right)\right) \\
&= \quad \exp\left(-\,(L^r - 4)\,\Lambda_1^{a,*}\left(\frac{1}{\lambda_1}\left(\frac{1}{1 + \frac{\tilde{\varepsilon}}{3\lambda_1}}\right)\right)\right).
\end{aligned}$$

Similarly, for $j = 1, 2$, using (72) we have

$$\begin{aligned}
(85) \qquad & \mathbf{P}\left(S_j^{r,n}(s^r) < (\mu_j^r - \tilde{\varepsilon}) s^r, \tau_{2n-1}^r \le r^2 t\right) \\
\le \quad & \mathbf{E}\left(1_{\{\mathcal{T}_n^r \in \mathbb{N}^3\}} \mathbf{P}\left(\breve{S}_j^{r,n}(s^r) < (\mu_j^r - \tilde{\varepsilon}) s^r, \tau_{2n-1}^r \le r^2 t \,|\, \mathcal{F}_{\mathcal{T}_n^r}^r\right)\right)
\end{aligned}$$

where on $\{\mathcal{T}_n^r \in \mathbb{N}^3\}$, $\breve{S}_j^{r,n}$ is the counting process defined from the sequence of successive service time random variables $\{v_j^r(S_j^r(T_j^r(\tau_{2n-1}^r)) + i), i = 1, 2, \ldots\}$ whose conditional distribution on $\{\mathcal{T}_n^r \in \mathbb{N}^3\}$ given $\mathcal{F}_{\mathcal{T}_n^r}^r$ is that of a sequence of strictly positive independent random variables in which the members indexed by $i = 2, 3, \ldots$ are identically distributed with the same distribution as $v_j^r(1)$. Hence by a slight adaptation of the proof of (193) in Appendix A (to include the extra event $\{\tau_{2n-1}^r \le r^2 t\}$), for $\delta_j^r = \frac{\tilde{\varepsilon}}{2\mu_j^r}$ we have a.s. on $\{\mathcal{T}_n^r \in \mathbb{N}^3\}$,

$$\begin{aligned}
(86) \qquad & \mathbf{P}\left(\breve{S}_j^{r,n}(s^r) < (\mu_j^r - \tilde{\varepsilon}) s^r, \tau_{2n-1}^r \le r^2 t \,|\, \mathcal{F}_{\mathcal{T}_n^r}^r\right) \\
\le \quad & \exp\left(-\left(\mu_j^r - \tilde{\varepsilon}\right) s^r \Lambda_j^{s,*}\left(\frac{1}{\mu_j}\left(1 + \frac{\tilde{\varepsilon}}{2\mu_j}\right)\right)\right) \\
& \quad + \mathbf{P}\left(v_j^r(S_j^r(T_j^r(\tau_{2n-1}^r)) + 1) > \delta_j^r s^r, \tau_{2n-1}^r \le r^2 t \,|\, \mathcal{F}_{\mathcal{T}_n^r}^r\right) \\
\le \quad & \exp\left(-\frac{(\mu_j - 2\tilde{\varepsilon})(L^r - 3)}{(\lambda_1 + 2\tilde{\varepsilon})}\Lambda_j^{s,*}\left(\frac{1}{\mu_j}\left(1 + \frac{\tilde{\varepsilon}}{2\mu_j}\right)\right)\right) \\
& \quad + \mathbf{P}\left(v_j^r(S_j^r(T_j^r(\tau_{2n-1}^r)) + 1) > \delta_j^r s^r, \tau_{2n-1}^r \le r^2 t \,|\, \mathcal{F}_{\mathcal{T}_n^r}\right).
\end{aligned}$$

Now, using (192) and (194), for any $\ell_0 \in \mathcal{O}$ such that $\ell_0 > 0$ we have

$$
\begin{aligned}
(87)\qquad & \mathbf{P}\left(v_j^r(S_j^r(T_j^r(\tau_{2n-1}^r))+1) > \delta_j^r s^r, \tau_{2n-1}^r \le r^2 t\right) \\
\le\ & \mathbf{P}\left(\max_{i=1}^{S_j^r(r^2t)+1} v_j^r(i) > \delta_j^r s^r\right) \\
\le\ & \mathbf{P}\left(S_j^r(r^2t) > (\mu_j^r+\tilde{\varepsilon})r^2 t\right) + \mathbf{P}\left(\max_{i=1}^{[(\mu_j^r+\tilde{\varepsilon})r^2t]+1} v_j^r(i) > \delta_j^r s^r\right) \\
\le\ & \exp\left(-(\mu_j r^2 t - 1)\Lambda_j^{s,*}\left(\frac{1}{\mu_j}\left(\frac{1}{1+\frac{\tilde{\varepsilon}}{3\mu_j}}\right)\right)\right) \\
& + \exp\left(\log((\mu_j^r+\tilde{\varepsilon})r^2 t + 1) - \frac{\ell_0 \tilde{\varepsilon} s^r}{2\mu_j} + \Lambda_j^s(\ell_0)\right) \\
\le\ & \exp\left(-(\mu_j r^2 t - 1)\Lambda_j^{s,*}\left(\frac{1}{\mu_j}\left(\frac{1}{1+\frac{\tilde{\varepsilon}}{3\mu_j}}\right)\right)\right) \\
& + \exp\left(\log((\mu_j^r+\tilde{\varepsilon})t) + 1 + 2\log r - \frac{\ell_0 \tilde{\varepsilon}(L^r-3)}{2\mu_j(\lambda_1+2\tilde{\varepsilon})} + \Lambda_j^s(\ell_0)\right).
\end{aligned}
$$

Combining all of the above from (76) onwards we have for all r sufficiently large,

$$
\begin{aligned}
(88)\qquad & \mathbf{P}\left(R^r(s) \ge L^r - 1 \text{ some } s \in [0, r^2 t]\right) \\
\le\ & \exp\left(-(\lambda_1 r^2 t - 1)\Lambda_1^{a,*}\left(\frac{1}{\lambda_1}\left(\frac{1}{1+\frac{\tilde{\varepsilon}}{3\lambda_1}}\right)\right)\right) \\
& + n^r \Bigg\{ \exp\left(-(L^r-4)\Lambda_1^{a,*}\left(\frac{1}{\lambda_1}\left(\frac{1}{1+\frac{\tilde{\varepsilon}}{3\lambda_1}}\right)\right)\right) \\
& \quad + \sum_{j=1}^{2} \Bigg\{ \exp\left(\frac{-(\mu_j - 2\tilde{\varepsilon})(L^r-3)}{\lambda_1+2\tilde{\varepsilon}}\Lambda_j^{s,*}\left(\frac{1}{\mu_j}\left(1+\frac{\tilde{\varepsilon}}{2\mu_j}\right)\right)\right) \\
& \quad\quad + \exp\left(-\left(\mu_j r^2 t - 1\right)\Lambda_j^{s,*}\left(\frac{1}{\mu_j}\left(\frac{1}{1+\frac{\tilde{\varepsilon}}{3\mu_j}}\right)\right)\right) \\
& \quad\quad + \exp\left(1 + \log((\mu_j^r+\tilde{\varepsilon})t) + 2\log r - \frac{\ell_0\tilde{\varepsilon}(L^r-3)}{2\mu_j(\lambda_1+2\tilde{\varepsilon})} + \Lambda_j^s(\ell_0)\right)\Bigg\}\Bigg\}.
\end{aligned}
$$

Since $n^r \le (\lambda_1 + 2\tilde{\varepsilon})te^{2\log r} + 1$, it follows that there is a constant c_0 such that for any $c > c_0$ and $L^r = [c \log r]$, the above tends to zero as $r \to \infty$. □

Proof of Theorem 7.1. Fix $t \ge 0$ and let $\varepsilon > 0$. Then, by Theorem 7.2 and the properties of L^r, there is an $r_\varepsilon \ge 1$ such that whenever $r \ge r_\varepsilon$, $2L^r/r < \varepsilon$ and

$$
\begin{aligned}
(89)\qquad & \mathbf{P}\left(I_1^r(\tau_0^r) \ge r\varepsilon\right) < \varepsilon \\
(90)\qquad & \mathbf{P}\left(\sup_{\tau_0^r \le s \le r^2 t} |R^r(s)| \ge L^r - 1\right) < \varepsilon.
\end{aligned}
$$

Note that under the threshold policy, server one is only idle when buffer one is empty, i.e., I_1^r can increase only if $Q_1^r \le 1$. (The bound of one occurs here because there may be a class one customer in service or in suspension at server 2 when buffer one is empty.) In particular, if server one incurs some idletime in $[\tau_0^r, r^2t]$, i.e., $I_1^r(r^2t) - I_1^r(\tau_0^r) > 0$, then $R^r(s) \le -L^r + 1$ for some $s \in [\tau_0^r, r^2t]$. Thus, for all $r \ge r_\varepsilon$,

$$
\begin{aligned}
(91) \qquad & \mathbf{P}\left(\|\hat{Q}_1^r\|_t \ge \varepsilon \ \text{ or } \ \|\hat{I}_1^r\|_t \ge \varepsilon\right) \\
= \ & \mathbf{P}\left(\|Q_1^r\|_{r^2t} \ge r\varepsilon \ \text{ or } \ \|I_1^r\|_{r^2t} \ge r\varepsilon\right) \\
\le \ & \mathbf{P}\left(\sup_{\tau_0^r \le s \le r^2t} Q_1^r(s) \ge 2L^r \ \text{ or } \ I_1^r(\tau_0^r) \ge r\varepsilon \ \text{ or } \ I_1^r(r^2t) - I_1^r(\tau_0^r) > 0\right) \\
\le \ & \mathbf{P}\left(\sup_{\tau_0^r \le s \le r^2t} |R^r(s)| \ge L^r - 1\right) + \mathbf{P}\left(I_1^r(\tau_0^r) \ge r\varepsilon\right) < 2\varepsilon.
\end{aligned}
$$

Since t and ε were arbitrary, the desired result follows. □

8 Weak Convergence under the Threshold Policy

This section is devoted to the proof of Theorem 5.2. Throughout this section, as in the previous one, we assume that the allocation processes $T^{r,*}$ associated with the threshold policy described in Definition 5.1 are used in the r^{th} parallel server system, and to simplify the notation, in proofs we simply use T^r in place of $T^{r,*}$.

In addition to the scaled processes defined in (16)–(20), we define the following fluid and diffusion scaled processes. For each $t \ge 0$, let

$$\bar{A}^r(t) = r^{-2}A^r(r^2t), \tag{92}$$

$$\bar{S}^r(t) = r^{-2}S^r(r^2t), \tag{93}$$

$$\bar{I}^r(t) = r^{-2}I^r(r^2t), \tag{94}$$

$$\bar{Q}^r(t) = r^{-2}Q^r(r^2t), \tag{95}$$

and let

$$\hat{W}^r(t) = y^r \cdot \hat{Q}^r(t), \tag{96}$$

where $y^r = (y_1^r, y_2^r)'$ and $y_1^r = 1$, $y_2^r = \mu_2^r/\mu_3^r$. Recall the definition of $\bar{T}^*$ from (29). Define $e : [0,\infty) \to [0,\infty)$ such that $e(t) = t$ for all $t \ge 0$. The proof of Theorem 5.2 depends on the following lemma which we prove first.

Lemma 8.1 *For the fluid scaled allocation processes, $\bar{T}_j^{r,*}$, $j = 1, 2, 3$, we have,*

$$\bar{T}^{r,*} \Longrightarrow \bar{T}^{*} \quad \textit{as } r \to \infty. \tag{97}$$

Proof. We first note that if $\{Z^r\}$ is a sequence of processes and Z is a continuous deterministic process (such as $\bar{T}^*$ or the identically zero process $\mathbf{0}$), then $Z^r \Longrightarrow Z$ is equivalent to $Z^r \to Z$ u.o.c. in prob. (uniformly on compact time intervals in probability). This is implicitly used several times in the proof below to combine statements involving convergence in distribution to deterministic processes.

By (7),

$$0 \le t - \bar{T}_1^r(t) = \bar{I}_1^r(t) = r^{-1}\hat{I}_1^r(t) \quad \text{for all } t \ge 0, \tag{98}$$

and so, since $\hat{I}_1^r \Longrightarrow \mathbf{0}$ as $r \to \infty$ by Theorem 7.1, it follows that

$$\bar{T}_1^r \Longrightarrow \bar{T}_1^* \text{ as } r \to \infty, \tag{99}$$

where $\bar{T}_1^*(t) = t$ for all $t \ge 0$. Now, by (21), for each $t \ge 0$,

$$\bar{Q}_1^r(t) \;=\; r^{-1}\hat{A}_1^r(t) - r^{-1}\hat{S}_1^r(\bar{T}_1^r(t)) - r^{-1}\hat{S}_2^r(\bar{T}_2^r(t)) + \lambda_1^r t - \mu_1^r \bar{T}_1^r(t) - \mu_2^r \bar{T}_2^r(t). \tag{100}$$

By using the fact that $\bar{T}_j^r(t) \le t$ for all $t \ge 0$, $j = 1, 2$, to obtain an estimate like (165), we deduce from the functional central limit result (23) and the continuous mapping theorem (cf. [1], p. 77) that

$$\left(r^{-1}\hat{A}_1^r(\cdot), r^{-1}\hat{S}_1^r(\bar{T}_1^r(\cdot)), r^{-1}\hat{S}_2^r(\bar{T}_2^r(\cdot))\right) \Longrightarrow (\mathbf{0}, \mathbf{0}, \mathbf{0}) \quad \text{as } r \to \infty. \tag{101}$$

On combining (100)–(101) with Theorem 7.1, we conclude that

$$\lambda_1^r e(\cdot) - \mu_1^r \bar{T}_1^r(\cdot) - \mu_2^r \bar{T}_2^r(\cdot) \Longrightarrow \mathbf{0} \quad \text{as } r \to \infty. \tag{102}$$

It then follows from this, (99), and Assumption 3.1, that

$$\frac{(\lambda_1^r - \mu_1^r)}{\mu_2^r} e(\cdot) - \bar{T}_2^r(\cdot) \Longrightarrow \mathbf{0} \quad \text{as } r \to \infty, \tag{103}$$

and hence by (29) and Assumption 3.1,

$$\bar{T}_2^r \Longrightarrow \bar{T}_2^* \quad \text{as } r \to \infty. \tag{104}$$

It remains to show that

$$\bar{T}_3^r \Longrightarrow \bar{T}_3^* \quad \text{as } r \to \infty. \tag{105}$$

For this, since by (8),

$$\bar{T}_3^r(t) = t - \bar{T}_2^r(t) - \bar{I}_2^r(t), \quad t \geq 0, \tag{106}$$

on taking (104), the definition of $\bar{T}_3^*$ and Assumption 3.1 into account, we see that it suffices to show that

$$\bar{I}_2^r \Longrightarrow \mathbf{0} \text{ as } r \to \infty. \tag{107}$$

Now, by (96), (7)–(8), (21)–(22), we have for each $t \geq 0$,

$$\bar{W}^r(t) \equiv r^{-1}\hat{W}^r(t) = y^r \cdot \bar{Q}^r(t) = y^r \cdot \bar{X}^r(t) + \mu_1^r \bar{I}_1^r(t) + \mu_2^r \bar{I}_2^r(t), \tag{108}$$

where for $\hat{X}^r$ defined in (31)–(32),

$$\bar{X}^r \equiv r^{-1}\hat{X}^r \Longrightarrow \mathbf{0} \quad \text{as } r \to \infty, \tag{109}$$

by similar reasoning to that for (101), using (23) and Assumption 3.1. Thus, for $t \geq 0$,

$$y_2^r \bar{Q}_2^r(t) = \bar{\zeta}^r(t) + \mu_2^r \bar{I}_2^r(t) \tag{110}$$

where, by (109) and Theorem 7.1 we have

$$\bar{\zeta}^r \equiv y^r \cdot \bar{X}^r + \mu_1^r \bar{I}_1^r - \bar{Q}_1^r \Longrightarrow \mathbf{0} \quad \text{as } r \to \infty. \tag{111}$$

Here, $y_2^r > 0$, $\mu_2^r > 0$, $\bar{Q}_2^r(t) \geq 0$, $\bar{\zeta}^r(0) = 0$, $\bar{I}_2^r(0) = 0$, and $\bar{I}_2^r$ is continuous and non-decreasing. Furthermore, from the definition of the threshold policy we note that the idletime at server 2 can increase only if there are no class 2 jobs in the system, and hence $\bar{I}_2^r$ can only increase if $\bar{Q}_2^r$ is zero. It follows from these characteristics that $(y_2^r \bar{Q}_2^r,\ \mu_2^r \bar{I}_2^r)$ is the unique solution of the one-dimensional Skorokhod problem (cf. Proposition B.1) for $\bar{\zeta}^r$ and hence

$$\bar{I}_2^r(t) = -(\mu_2^r)^{-1} \inf_{0 \leq s \leq t} \bar{\zeta}^r(s) \tag{112}$$

and (107) follows from this by the continuous mapping theorem and (111). □

Proof of Theorem 5.2. For each $t \geq 0$, we have by multiplying (108) through by r that

$$\hat{W}^r(t) = y^r \cdot \hat{Q}^r(t) = y^r \cdot \hat{X}^r(t) + \mu_1^r \hat{I}_1^r(t) + \mu_2^r \hat{I}_2^r(t), \tag{113}$$

where $\hat{X}^r$ is defined by (31)–(32). By the functional central limit theorem result (23), Lemma 8.1, and a random time change theorem (cf. [1], §17), we have as $r \to \infty$,

$$\left(\hat{A}_1^r(\cdot), \hat{A}_2^r(\cdot), \hat{S}_1^r(\bar{T}_1^r(\cdot)), \hat{S}_2^r(\bar{T}_2^r(\cdot)), \hat{S}_3^r(\bar{T}_3^r(\cdot))\right) \implies \left(\tilde{A}_1(\cdot), \tilde{A}_2(\cdot), \tilde{S}_1(\bar{T}_1^*(\cdot)), \tilde{S}_2(\bar{T}_2^*(\cdot)), \tilde{S}_3(\bar{T}_3^*(\cdot))\right).$$

It then follows from the definition of $\hat{X}^r$ and Assumption 3.1 that

$$(\hat{X}_1^r, \hat{X}_2^r) \Longrightarrow (\tilde{X}_1, \tilde{X}_2) \text{ as } r \to \infty, \tag{114}$$

where

$$\tilde{X}_1(t) = \tilde{A}_1(t) - \tilde{S}_1(\bar{T}_1^*(t)) - \tilde{S}_2(\bar{T}_2^*(t)) + \theta_1 t, \tag{115}$$

$$\tilde{X}_2(t) = \tilde{A}_3(t) - \tilde{S}_3(\bar{T}_3^*(t)) + \theta_2 t, \tag{116}$$

so that $\tilde{X}$ is a two-dimensional Brownian motion with drift as described in Definition 4.1. Rearranging (113), we have

$$y_2^r \hat{Q}_2^r(t) = \hat{\zeta}^r(t) + \mu_2^r \hat{I}_2^r(t) \tag{117}$$

where

$$\hat{\zeta}^r \equiv y^r \cdot \hat{X}^r + \mu_1^r \hat{I}_1^r - \hat{Q}_1^r \Longrightarrow y \cdot \tilde{X} \quad \text{as } r \to \infty, \tag{118}$$

by (114), Theorem 7.1 and Assumption 3.1. By the same reasoning as in the proof of Lemma 8.1, using uniqueness of the solution to the Skorokhod problem, we have

$$\hat{I}_2^r(t) = -(\mu_2^r)^{-1} \inf_{0 \le s \le t} \hat{\zeta}^r(s). \tag{119}$$

It then follows from (117), (118), (119), and the continuous mapping theorem that

$$(\hat{Q}_2^r, \hat{I}_2^r) \Longrightarrow (\tilde{Q}_2^*, \tilde{I}_2^*) \quad \text{as } r \to \infty, \tag{120}$$

where $\tilde{Q}_2^*, \tilde{I}_2^*$ are given by (41)–(42). Combining this with Theorem 7.1 yields

$$(\hat{Q}^r, \hat{I}^r, \hat{W}^r) \Longrightarrow (\tilde{Q}^*, \tilde{I}^*, \tilde{W}^*) \quad \text{as } r \to \infty, \tag{121}$$

where $\tilde{Q}^*, \tilde{I}^*, \tilde{W}^*$ are defined in (41)–(42). □

9 Asymptotic Optimality of the Threshold Policy

The purpose of this section is to prove Theorem 5.3. Before proceeding with the proof, we first establish some preliminary results concerning fluid scaled processes.

In this section, $T = \{T^r\}$ will be any sequence of scheduling control policies (one for each member of the sequence of parallel server systems). The associated queue length and idletime processes will be denoted by Q^r, I^r, and the fluid and diffusion scaled versions of these processes will be denoted by $\bar{Q}^r, \bar{I}^r$ and $\hat{Q}^r, \hat{I}^r$, respectively. We also let

$$\underline{J}(T) = \liminf_{r\to\infty} \hat{J}^r(T^r), \tag{122}$$

where $\hat{J}^r(T^r)$ is defined in (24). When our sequence of threshold policies $\{T^{r,*}\}$ is used, we append a superscript * to the queue length, idletime etc. processes e.g., $Q^{r,*}, I^{r,*}$, etc.

Definition 9.1 *($\mathbf{C}$-tightness) In the following, a sequence of processes with paths in $\mathbf{D}^m$ for some $m \geq 1$ is called $\mathbf{C}$-tight if it is tight in $\mathbf{D}^m$ and any weak limit point of the sequence (obtained as a weak limit along a subsequence) has continuous paths almost surely.*

Lemma 9.2 *Let $\{T^r\}$ be any sequence of scheduling control policies (one for each member of the sequence of parallel server systems). Then*

$$\{(\bar{Q}^r(\cdot), \bar{A}^r(\cdot), \bar{S}^r(\cdot), \bar{T}^r(\cdot), \bar{I}^r(\cdot))\} \tag{123}$$

is $\mathbf{C}$-tight.

Proof. It follows from (23) that

$$(\bar{A}^r(\cdot), \bar{S}^r(\cdot)) \Longrightarrow (\lambda(\cdot), \mu(\cdot)) \quad \text{as } r \to \infty, \tag{124}$$

where $\lambda(t) = \lambda t$ and $\mu(t) = \mu t$ for all $t \geq 0$. In addition, since they correspond to cumulative allocations of time, each of the three components of T^r is uniformly Lipschitz continuous with a Lipschitz constant less than or equal to one and this property is preserved by the fluid scaled processes $\bar{T}^r$. It follows immediately from this and (124) that $\{(\bar{A}^r(\cdot), \bar{S}^r(\cdot), \bar{T}^r(\cdot))\}$ is $\mathbf{C}$-tight, cf. Theorem 15.5 in [1]. From the equations (7)–(10) for queue length and idletime we have

$$\bar{Q}_1^r(t) = \bar{A}_1^r(t) - \bar{S}_1^r(\bar{T}_1^r(t)) - \bar{S}_2^r(\bar{T}_2^r(t)), \tag{125}$$

$$\bar{Q}_2^r(t) = \bar{A}_2^r(t) - \bar{S}_3^r(\bar{T}_3^r(t)), \tag{126}$$

$$\bar{I}_1^r(t) = t - \bar{T}_1^r(t), \tag{127}$$

$$\bar{I}_2^r(t) = t - \bar{T}_2^r(t) - \bar{T}_3^r(t). \tag{128}$$

Combining these with the $\mathbf{C}$-tightness established above and a random time change theorem (cf. [1], p. 145) yields the desired result. □

The next lemma in particular implies that, when searching for an asymptotically optimal policy, we may restrict to those policies whose associated fluid scaled allocation processes converge (along a subsequence) to those given by $\bar{T}^*$.

Lemma 9.3 *Let $T = \{T^r\}$ be a sequence of scheduling control policies such that $\underline{J}(T) < \infty$. Consider a subsequence $\{T^{r'}\}$ of $\{T^r\}$ along which the* $\liminf$ *in the definition of $\underline{J}(T)$ is achieved, i.e.,*

$$\lim_{r'\to\infty} \hat{J}^{r'}(T^{r'}) = \underline{J}(T). \tag{129}$$

Then,

$$(\bar{Q}^{r'}(\cdot), \bar{A}^{r'}(\cdot), \bar{S}^{r'}(\cdot), \bar{T}^{r'}(\cdot), \bar{I}^{r'}(\cdot)) \Longrightarrow (\mathbf{0}, \lambda(\cdot), \mu(\cdot), \bar{T}^*(\cdot), \mathbf{0}) \quad \textit{as } r' \to \infty, \tag{130}$$

where $\bar{T}^*$ is defined by (29), $\mathbf{0}$ denotes the constant process that stays at the origin in $\mathbb{R}^2$ for all time, and $\lambda(t) = \lambda t$, $\mu(t) = \mu t$ for all $t \geq 0$.

Proof. From Lemma 9.2 it follows that

$$\{(\bar{Q}^{r'}(\cdot), \bar{A}^{r'}(\cdot), \bar{S}^{r'}(\cdot), \bar{T}^{r'}(\cdot), \bar{I}^{r'}(\cdot))\} \tag{131}$$

is $\mathbf{C}$-tight. Thus, it suffices to show that all weak limit points of this sequence are given by the right member of (130). For this, suppose that

$$(\bar{Q}(\cdot), \bar{A}(\cdot), \bar{S}(\cdot), \bar{T}(\cdot), \bar{I}(\cdot)), \tag{132}$$

is obtained as a weak limit of (131) along a subsequence indexed by r''. Without loss of generality, by appealing to the Skorokhod representation theorem (cf. [6], 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 (131) indexed by r'', as well as the limit (132), are defined on the same probability space and the convergence in distribution is replaced by almost sure convergence on compact time intervals, so that a.s.,

$$(\bar{Q}^{r''}(\cdot), \bar{A}^{r''}(\cdot), \bar{S}^{r''}(\cdot), \bar{T}^{r''}(\cdot), \bar{I}^{r''}(\cdot)) \to (\bar{Q}(\cdot), \bar{A}(\cdot), \bar{S}(\cdot), \bar{T}(\cdot), \bar{I}(\cdot)) \quad \text{u.o.c. as } r'' \to \infty, \tag{133}$$

From (124) we have that a.s., $\bar{A}(\cdot) = \lambda(\cdot)$ and $\bar{S}(\cdot) = \mu(\cdot)$. We next show that a.s., $\bar{Q}(\cdot) \equiv \mathbf{0}$. Combining the fact that $\lim_{r''\to\infty} \hat{J}^{r''}(T^{r''}) = \underline{J}(T) < \infty$ with (133) and Fatou's lemma, we have

$$\begin{aligned} 0 = \lim_{r''\to\infty} \frac{1}{r''}\hat{J}^{r''}(T^{r''}) &\geq \mathbf{E}\left(\int_0^\infty e^{-\gamma t} \liminf_{r''\to\infty} \left(h \cdot \bar{Q}^{r''}(t)\right) dt\right) \\ &= \mathbf{E}\left(\int_0^\infty e^{-\gamma t}\, h \cdot \bar{Q}(t)\, dt\right). \end{aligned} \tag{134}$$

Since $h_k > 0$, $k = 1, 2$, and a.s., $\bar{Q}$ has continuous paths in $\mathbb{R}_+ \times \mathbb{R}_+$, it follows from the above that a.s., $\bar{Q}(\cdot) \equiv \mathbf{0}$. Then, by letting $r'' \to \infty$ in (125)–(128) and using (124), (133), we have a.s., for each $t \geq 0$,

$$
\begin{aligned}
&(135) \qquad & 0 &= \lambda_1 t - \mu_1 \bar{T}_1(t) - \mu_2 \bar{T}_2(t), \\
&(136) \qquad & 0 &= \lambda_2 t - \mu_3 \bar{T}_3(t), \\
&(137) \qquad & \bar{I}_1(t) &= t - \bar{T}_1(t), \\
&(138) \qquad & \bar{I}_2(t) &= t - \bar{T}_2(t) - \bar{T}_3(t).
\end{aligned}
$$

Multiplying (135) by μ_2^{-1} and (136) by μ_3^{-1} and adding the two expressions together while recalling Assumption 3.1(ii) we obtain

$$
\frac{\mu_1}{\mu_2} \bar{I}_1(t) + \bar{I}_2(t) = 0. \tag{139}
$$

Since a.s., each component of $\bar{I}(\cdot)$ inherits the property from $\bar{I}^{r''}(\cdot)$ that it is non-negative for all time, it follows from (139) that a.s., $\bar{I}_1(\cdot) = \mathbf{0} = \bar{I}_2(\cdot)$. It then follows from (137) that $\bar{T}_1(\cdot) = \bar{T}_1^*(\cdot)$, and (136) together with (138) yield $\bar{T}_3(\cdot) = \bar{T}_3^*(\cdot)$ and $\bar{T}_2(\cdot) = \bar{T}_2^*(\cdot)$. □

Proof of Theorem 5.3. We first concentrate on proving the inequality on the left side of (45). For this, let $T \equiv \{T^r\}$ be a sequence of scheduling control policies. If $\underline{J}(T) = \infty$, then the inequality holds trivially and so we assume that $\underline{J}(T) < \infty$. Recall the definitions of $y = (y_1, y_2)' = (1, \mu_2/\mu_3)'$, $y^r = (1, \mu_2^r/\mu_3^r)'$, and of $\hat{W}^r$ from (96). Let

$$
h_1^r = h_1 \quad \text{and} \quad h_2^r = \frac{h_2 y_2^r}{y_2} = \frac{h_2 \mu_2^r \mu_3}{\mu_2 \mu_3^r}. \tag{140}
$$

Note that by Assumption 3.2 we have

$$
\frac{h_1}{h_2} \geq \frac{\mu_3}{\mu_2} = \frac{y_1^r}{y_2}. \tag{141}
$$

Then, using (113) (which holds for any scheduling control policy), we have for all $t \geq 0$,

$$
h^r \cdot \hat{Q}^r(t) = h_2 \left(\frac{h_1}{h_2} \hat{Q}_1^r(t) + \frac{y_2^r}{y_2} \hat{Q}_2^r(t) \right) \geq h_2 \left(\frac{y_1^r}{y_2} \hat{Q}_1^r(t) + \frac{y_2^r}{y_2} \hat{Q}_2^r(t) \right) = \frac{h_2}{y_2} \left(y^r \cdot \hat{X}^r(t) + \hat{V}^r(t) \right),
$$

where $\hat{X}^r$ is given by (31)–(32) and

$$
\hat{V}^r(t) = \mu_1^r \hat{I}_1^r(t) + \mu_2^r \hat{I}_2^r(t). \tag{142}
$$

Now, since $h^r \cdot \hat{Q}^r(t) \geq 0$ for all $t \geq 0$ and $\hat{I}_1^r, \hat{I}_2^r$ are non-decreasing and start from zero, it follows from the well known minimality of the solution of the Skorokhod problem (cf. Proposition B.1), that

$$\hat{V}^r(t) \geq - \inf_{0 \leq s \leq t} \left(y^r \cdot \hat{X}^r(s) \right) \quad \text{for all } t \geq 0, \tag{143}$$

and hence

$$h^r \cdot \hat{Q}^r(t) \geq \frac{h_2}{y_2} \varphi \left(y^r \cdot \hat{X}^r \right)(t) \quad \text{for all } t \geq 0, \tag{144}$$

where $\varphi(x)(t) \equiv x(t) - \inf_{0 \leq s \leq t} x(s)$ for all $t \geq 0$ and $x \in \mathbf{D}$ satisfying $x(0) = 0$.

Now, let $\{T^{r'}\}$ be a subsequence of $\{T^r\}$ such that $\lim_{r' \to \infty} \hat{J}^{r'}(T^{r'}) = \underline{J}(T)$. By (130), the fact that the limit there is deterministic, and (23), we have that as $r' \to \infty$,

$$(\hat{A}^{r'}(\cdot), \hat{S}^{r'}(\cdot), \bar{T}^{r'}(\cdot)) \Longrightarrow (\tilde{A}(\cdot), \tilde{S}(\cdot), \bar{T}^*(\cdot)). \tag{145}$$

By invoking the Skorokhod representation theorem, we may assume without loss of generality that the convergence above is a.s. uniform on compact time intervals (u.o.c.) and then for $\hat{X}^r, \tilde{X}$ given by (31)–(32), (115)–(116), using Assumption 3.1 we have that a.s. as $r' \to \infty$,

$$(\hat{A}^{r'}(\cdot), \hat{S}^{r'}(\cdot), \bar{T}^{r'}(\cdot), \hat{X}^{r'}(\cdot)) \to (\tilde{A}(\cdot), \tilde{S}(\cdot), \bar{T}^*(\cdot), \tilde{X}(\cdot)) \quad \text{u.o.c.} \tag{146}$$

Then, by (21)–(22) and Fatou's lemma, we have

$$\underline{J}(T) \;=\; \lim_{r' \to \infty} \hat{J}^{r'}(T^{r'}) \geq \mathbf{E} \left(\int_0^\infty e^{-\gamma t} \liminf_{r' \to \infty} \left(h \cdot \hat{Q}^{r'}(t) \right) dt \right). \tag{147}$$

Now we claim that a.s., for all $t \geq 0$,

$$\liminf_{r' \to \infty} \left(h \cdot \hat{Q}^{r'}(t) \right) \geq h \cdot \tilde{Q}^*(t), \tag{148}$$

where $\tilde{Q}^*$ is given by (41)–(42). To see this, fix $\omega \in \Omega$ such that ω is in the set of probability one where the convergence in (146) holds u.o.c., and fix $t \geq 0$. If the left member of the inequality (148) is infinite at ω, then the inequality clearly holds. On the other hand, if the left member is finite at ω, then there is a further subsequence indexed by r'' (possibly depending on ω and t) such that

$$\lim_{r'' \to \infty} \left(h \cdot \hat{Q}^{r''}(t, \omega) \right) \;=\; \liminf_{r' \to \infty} \left(h \cdot \hat{Q}^{r'}(t, \omega) \right) < \infty. \tag{149}$$

Since $h_k > 0$ and $\hat{Q}_k^{r''}(t, \omega) \geq 0$ for $k = 1, 2$, it follows that $\hat{Q}_2^{r''}(t, \omega)$ is bounded as $r'' \to \infty$, and then using the fact that $h_2^r \to h_2 > 0$ we have

$$\lim_{r'' \to \infty} (h_2 - h_2^{r''}) \hat{Q}_2^{r''}(t, \omega) = 0. \tag{150}$$

Then, using (144), (146), the continuity of φ on $\mathbf{D}$, and (41)–(42), we have

$$
\begin{aligned}
(151)\qquad \lim_{r''\to\infty} h\cdot \hat{Q}^{r''}(t,\omega) &= \lim_{r''\to\infty}\left(h_1\hat{Q}_1^{r''}(t,\omega)+h_2^{r''}\hat{Q}_2^{r''}(t,\omega)+(h_2-h_2^{r''})\hat{Q}_2^{r''}(t,\omega)\right)\\
&= \lim_{r''\to\infty}\left(h^{r''}\cdot\hat{Q}^{r''}(t,\omega)\right)\\
&\geq \liminf_{r''\to\infty}\frac{h_2}{y_2}\varphi\left(y^{r''}\cdot\hat{X}^{r''}\right)(t,\omega)\\
&= \frac{h_2}{y_2}\varphi\left(y\cdot\tilde{X}\right)(t,\omega)=\frac{h_2}{y_2}\tilde{W}^*(t,\omega)=h\cdot\tilde{Q}^*(t,\omega).
\end{aligned}
$$

Thus, (148) holds. Now, substituting this in (147), we conclude that

$$
(152)\qquad \underline{J}(T)\geq \mathbf{E}\left(\int_0^\infty e^{-\gamma t}\,h\cdot\tilde{Q}^*(t)\,dt\right)\equiv J^*.
$$

This completes the proof of the inequality in the left side of (45).

Suppose now that the threshold policy $T^{r,*}$ is used in the r^{th} parallel server system. For the purpose of establishing the finiteness of J^* and the equality in the right side of (45), by appealing to Theorem 5.2 and the Skorokhod representation theorem, we may assume that a.s.,

$$
(153)\qquad \hat{Q}^{r,*}\to\tilde{Q}^*\ \text{ u.o.c. as } r\to\infty,
$$

where $\tilde{Q}^*, \tilde{I}^*$ are given by (41)–(42). Then for

$$
(154)\qquad \hat{H}^{r,*}\equiv h\cdot\hat{Q}^{r,*}\quad\text{and}\quad \tilde{H}^*\equiv h\cdot\tilde{Q}^*
$$

we have

$$
(155)\qquad \hat{H}^{r,*}\to H^*\quad (m\times\mathbf{P})\text{-a.e.}
$$

where $dm=\gamma e^{-\gamma t}dt$ on $(\mathbf{R}_+,\mathcal{B}_+)$ and $\mathcal{B}_+$ denotes the Borel σ-algebra on $\mathbb{R}_+$. Then, since $(\mathbb{R}_+\times\Omega,\mathcal{B}_+\times\mathcal{F},m\times\mathbf{P})$ is a probability space, to establish

$$
(156)\qquad \hat{J}^r(T^{r,*})\equiv\mathbf{E}\left(\int_0^\infty e^{-\gamma t}\hat{H}^{r,*}(t)\,dt\right)\to J^*<\infty \text{ as } r\to\infty,
$$

it suffices to show that

$$
(157)\qquad \limsup_{r\to\infty}\ \mathbf{E}\left(\int_0^\infty e^{-\gamma t}\left(\hat{H}^{r,*}(t)\right)^2 dt\right)<\infty
$$

which implies the required uniform integrability. From (113) we have

$$
(158)\qquad \hat{H}^{r,*}=h\cdot\hat{Q}^{r,*}\leq\left(\frac{h_1}{y_1^r}+\frac{h_2}{y_2^r}\right)\hat{W}^{r,*}
$$

where

$$\hat{W}^{r,*} = y^r \cdot \hat{Q}^{r,*} = y^r \cdot \hat{X}^{r,*} + \hat{V}^{r,*} \tag{159}$$

and

$$\begin{aligned}
\hat{X}_1^{r,*}(t) &= \hat{A}_1^r(t) - \hat{S}_1^r\left(\bar{T}_1^{r,*}(t)\right) - \hat{S}_2^r\left(\bar{T}_2^{r,*}(t)\right) + r\mu_2^r\left(\frac{\lambda_1^r - \mu_1^r}{\mu_2^r} - \frac{\lambda_1 - \mu_1}{\mu_2}\right)t, && (160)\\
\hat{X}_2^{r,*}(t) &= \hat{A}_2^r(t) - \hat{S}_3^r\left(\bar{T}_3^{r,*}(t)\right) + r\mu_3^r\left(\frac{\lambda_2^r}{\mu_3^r} - \frac{\lambda_2}{\mu_3}\right)t, && (161)\\
\hat{V}^{r,*}(t) &= \mu_1^r \hat{I}_1^{r,*}(t) + \mu_2^r \hat{I}_2^{r,*}(t). && (162)
\end{aligned}$$

Now, by the definition of $\bar{T}^{r,*}, \hat{I}_2^{r,*}$ can only increase if $\hat{Q}_2^{r,*}$ is zero and $\hat{Q}_1^{r,*}$ is at or below the level L^r/r. Thus, it follows from an oscillation inequality for solutions of a perturbed Skorokhod problem (cf. the proof of Theorem 5.1 in [35]) that

$$\begin{aligned}
\mu_2^r \hat{I}_2^{r,*}(t) &\le -\inf_{0\le s\le t}\left(y^r \cdot \hat{X}^{r,*}(s) + \mu_1^r \hat{I}_1^{r,*}(s)\right) + y_1^r L^r r^{-1} && (163)\\
&\le \sup_{0\le s\le t}\left|y^r \cdot \hat{X}^{r,*}(s)\right| + \mu_1^r \hat{I}_1^{r,*}(t) + y_1^r L^r r^{-1}
\end{aligned}$$

where we have used the fact that $\hat{I}_1^{r,*}$ is non-decreasing to obtain the last inequality. Combining the above, we see that to prove (157) it suffices to show that as functions of t the following are all in a bounded subset of $L^1(m) \equiv L^1(\mathbb{R}_+, \mathcal{B}_+, m)$ for r sufficiently large:

$$\mathbf{E}\left(\sup_{0\le s\le t}\left(\hat{A}_k^r(s)\right)^2\right), \quad \mathbf{E}\left(\sup_{0\le s\le t}\left(\hat{S}_j^r(\bar{T}_j^{r,*}(s))\right)^2\right), \quad \mathbf{E}\left(\left(\hat{I}_1^{r,*}(t)\right)^2\right), \tag{164}$$

$k = 1, 2,\ j = 1, 2, 3$. We establish estimates that show this for the last two expectations in (164), the estimates for the first expectation being similar to those for the middle one. For later use we note that due to the exponential decay factor in m, any polynomial in t is in $L^1(m)$.

For $t \ge 0$ and $j \in \{1, 2, 3\}$, since $\bar{T}_j^{r,*}$ is continuous and $0 \le \bar{T}_j^{r,*}(s) \le s$ for all s, we have

$$\sup_{0\le s\le t}\left|\hat{S}_j^r\left(\bar{T}_j^{r,*}(s)\right)\right| \le \sup_{0\le s\le t}\left|\hat{S}_j^r(s)\right|, \tag{165}$$

so it suffices to focus on estimating the right member above. By the definition of the sum of i.i.d service times η_j^r used to define S_j^r, we have that

$$\mathcal{M}_j^r(n) \equiv \mu_j^r\ \eta_j^r(n) - n, \quad n = 0, 1, 2, \ldots, \tag{166}$$

is an L^2-martingale relative to the filtration $\{\mathcal{G}_n^{r,j}\}_{n=0}^\infty$ where $\mathcal{G}_n^{r,j} = \sigma\{v_j^r(i), i = 1, \ldots, n\} \vee \mathcal{N}$ and the quadratic variation process of this discrete-time martingale is given by $[\mathcal{M}_j^r]_n = \beta_j^2 n, n =$

$0, 1, 2, \ldots$ where β_j^2 is the squared coefficient of variation of the service times $\{v_j^r(i)\}_{i=1}^\infty$ (recall that this does not depend upon r). Now, $S_j^r(r^2t)+1$ is a stopping time for the filtration $\{\mathcal{G}_n^{r,j}\}_{n=0}^\infty$ and in a similar manner to that for equation (196) of [36] we have by Doob's inequality, the quadratic variation of $\mathcal{M}_j^r$, and Lorden's inequality for the mean-value of a renewal process at time r^2t, that

$$\mathbf{E}\left(\sup_{0\le s\le t}\left|\mathcal{M}_j^r(S_j^r(r^2s)+1)\right|^2\right) \le 4\beta_j^2\mathbf{E}\left(S_j^r(r^2t)+1\right) \le 4\beta_j^2\left(\mu_j^r r^2 t+\beta_j^2+2\right). \tag{167}$$

Now, we estimate $\hat{S}_j^r(s)$ in terms of $\mathcal{M}_j^r$ as follows:

$$\hat{S}_j^r(s) = -r^{-1}\mathcal{M}_j^r(S_j^r(r^2s)+1)+\hat{\varepsilon}_j^r, \tag{168}$$

where

$$\hat{\varepsilon}_j^r = r^{-1}\mu_j^r\left(\eta_j^r(S_j^r(r^2s)+1)-r^2s\right)-r^{-1}, \tag{169}$$

and by bounding the residual service time at r^2s by the full service time that straddles r^2s, we have

$$\begin{aligned} |\hat{\varepsilon}_j^r| &\le r^{-1}\mu_j^r v_j^r(S_j^r(r^2s)+1)+r^{-1} \\ &= r^{-1}\left(\mathcal{M}_j^r(S_j^r(r^2s)+1)-\mathcal{M}_j^r(S_j^r(r^2s))+2\right) \\ &\le 2r^{-1}\left(\sup_{0\le s\le t}|\mathcal{M}_j^r(S_j^r(r^2s)+1)|+1\right). \end{aligned} \tag{170}$$

Thus,

$$\sup_{0\le s\le t}|\hat{S}_j^r(s)| \le 3r^{-1}\left(\sup_{0\le s\le t}|\mathcal{M}_j^r(S_j^r(r^2s)+1)|+1\right) \tag{171}$$

and hence using (167) we have since $r\ge 1$,

$$\mathbf{E}\left(\sup_{0\le s\le t}|\hat{S}^r(s)|^2\right) \le 18(1+4\beta_j^2(\mu_j^r t+\beta_j^2+2)) \tag{172}$$

where the right member is in a bounded subset of $L^1(m)$.

We now turn our attention to estimating $\mathbf{E}\left((\hat{I}_1^{r,*}(t))^2\right)$. For this we will use estimates contained in the proof of Theorem 7.2. In order for these estimates to be small enough to imply the desired $L^1(m)$-boundedness, the constant c_0 in the definition of the threshold policy $T^{r,*}$ needs to be sufficiently large (and possibly larger than what is required for the results of Theorem 7.2 alone to hold). Now, since $I_1^r(r^2t)\le r^2t$, we have using the notation of Section 7,

$$\mathbf{E}\left(\left(\hat{I}_1^{r,*}(t)\right)^2\right) = \int_0^\infty \mathbf{P}\left(\left(\hat{I}_1^{r,*}(t)\right)^2 > s\right)ds = \int_0^{r^2t^2}\mathbf{P}\left(I_1^{r,*}(r^2t) > r\sqrt{s}\right)ds \tag{173}$$

$$
\begin{aligned}
&\leq \int_0^{r^2t^2} \left\{ \mathbf{P}\left(I_1^{r,*}(\tau_0^r) > r\sqrt{s}\right) + \mathbf{P}\left(\sup_{\tau_0^r \leq u \leq r^2 t} |R^r(u)| \geq L^r - 1 \right) \right\} ds \\
&\leq \left(\frac{t^r}{r}\right)^2 + \int_{(t^r/r)^2}^{r^2t^2} \mathbf{P}(I_1^{r,*}(\tau_0^r) > t^r) ds + r^2t^2\, \mathbf{P}\left(\sup_{\tau_0^r \leq u \leq r^2 t} |R^r(u)| \geq L^r - 1 \right)
\end{aligned}
$$

where $t^r = 8L^r/(\lambda_1 - \mu_1)$ and we have used the fact that $I_1^{r,*}(r^2t) - I_1^{r,*}(\tau_0^r) = 0$ if $\sup_{\tau_0^r \leq u \leq r^2 t} |R^r(u)| < L^r - 1$. By the proof of Theorem 7.2, (including the estimate analogous to (88) needed to prove (66)), there is $r_0 \geq 1$ (not depending on t) and positive, finite constants C_1–C_6 (not depending on t or r) such that for all $r \geq r_0$,

$$
\mathbf{P}\left(I_1^r(\tau_0^r) > t^r\right) \leq C_1 \exp(-C_2 L^r) \tag{174}
$$

(cf. (60)–(63)) and

$$
\mathbf{P}\left(\sup_{\tau_0^r \leq s \leq r^2 t} |R^r(s)| \geq L^r - 1 \right) \leq (C_3 + C_4 r^2 t)^2 \left(\exp(-C_5 L^r) + \exp(-C_6 r^2 t) \right). \tag{175}
$$

On substituting these estimates in (173), we have for $r \geq r_0$,

$$
\begin{aligned}
\mathbf{E}\left(\left(\hat{I}_1^{r,*}(t) \right)^2 \right) \;\leq\; & \left(\frac{t^r}{r}\right)^2 + r^2t^2 C_1 \exp(-C_2 L^r) \\
& + r^2t^2 (C_3 + C_4 r^2 t)^2 \left(\exp(-C_5 L^r) + \exp(-C_6 r^2 t) \right).
\end{aligned} \tag{176}
$$

Using the fact that for $c > 0$, $xe^{-cx}, x^2e^{-cx}, x^3e^{-cx}$ are bounded on $\mathbb{R}_+$ (with r^2t in place of x), we see that for $L^r = c \log r$ and c sufficiently large (chosen independently of t and r), the right member above defines a bounded sequence of functions in $L^1(m)$ for $r \geq r_0$. □

Acknowledgements. The authors are grateful to Marcel López and Michael Harrison for access to a preliminary version of their paper [12] and for related discussions.

A Appendix: Large Deviation Bounds for Renewal Processes

In this section we state and prove some estimates for large deviations of renewal processes. These estimates are applied in Section 7 to the arrival and service renewal processes, A^r and S^r and to shifts of these processes. While we believe that the results of this section are at least known in folklore, we could not find results in the literature that duplicate those that we need and so we give some justification here. In this section, we have reused some of the symbols defined earlier for other purposes. For more details on the properties of the Legendre-Fenchel transform used in this section, in the context of large deviations see [5], and in general see [30].

In the following, $r \geq 1$ is an index that takes values in a sequence of real numbers tending to infinity.

For $r \geq 1$ fixed, let $\{\zeta^r(i)\}_{i=1}^{\infty}$ be a sequence of strictly positive, independent random variables such that $\{\zeta^r(i)\}_{i=2}^{\infty}$ are identically distributed. Assume that there is a non-empty open neighborhood $\mathcal{O}^r$ of the origin such that

$$\Lambda^r(\ell) \equiv \log \mathbf{E}[e^{\ell \zeta^r(i)}] < \infty \quad \text{for all } \ell \in \mathcal{O}^r, \ i = 2, 3, \ldots. \tag{177}$$

It follows that $\zeta^r(i)$ for $i \geq 2$ has finite mean $m^r = E[\zeta_2^r] \in (0, \infty)$ and we let $\nu^r = 1/m^r$. For $n = 0, 1, 2, \ldots$, define

$$X^r(n) = \sum_{i=1}^{n} \zeta^r(i) \quad \text{and} \quad \check{X}^r(n) = \sum_{i=2}^{n+1} \zeta^r(i), \tag{178}$$

where an empty sum is defined to have value zero. For each $t \geq 0$, let

$$N^r(t) = \sup\{n \geq 0 : X^r(n) \leq t\}, \tag{179}$$

the renewal process associated with X^r. Recall that $[x]$ denotes the greatest integer part of any real number x. Then for $\varepsilon > 0$, $\kappa^r \equiv \nu^r + \varepsilon$ and $t > \frac{2}{\varepsilon}$, by Markov's inequality and the i.i.d. nature of $\{\zeta^r(i)\}_{i=2}^{\infty}$, we have for each $\ell \geq 0$,

$$\begin{aligned} \mathbf{P}(N^r(t) > \kappa^r t) &= \mathbf{P}(N^r(t) \geq [\kappa^r t] + 1) \\ &= \mathbf{P}(X^r([\kappa^r t] + 1) \leq t) \\ &\leq \mathbf{P}(\check{X}^r([\kappa^r t]) < t) \\ &\leq e^{\ell t}\, \mathbf{E}\left[\exp\left(-\ell \check{X}^r([\kappa^r t])\right)\right] \\ &= \exp\left(\ell t + [\kappa^r t]\Lambda^r(-\ell)\right) \\ &\leq \exp\left(\ell t + (\kappa^r t - 1)\Lambda^r(-\ell)\right) \end{aligned} \tag{180}$$

where we have used the fact that $\Lambda^r(-\ell) \leq 0$ for $\ell \geq 0$, by the non-negativity of the $\zeta^r(i)$. By minimizing the right member above over $\ell \geq 0$ (which is the same as minimizing over all ℓ since $0 < \frac{t}{\kappa^r t - 1} < m^r = \left(\frac{d}{d\ell}\Lambda^r\right)(0)$), we obtain a form of Cramér's large deviation *inequality* (cf. [5], page 27):

$$\mathbf{P}(N^r(t) > \kappa^r t) \;\leq\; \exp\left(-(\kappa^r t - 1)\Lambda^{r,*}\left(\frac{t}{\kappa^r t - 1}\right)\right) \tag{181}$$

$$\leq \exp\left(-(\kappa^r t - 1)\Lambda^{r,*}\left(\frac{1}{\nu^r + \frac{1}{2}\varepsilon}\right)\right)$$

where

(182) $$\Lambda^{r,*}(x) \equiv \sup_{\ell \in \mathbb{R}} (\ell x - \Lambda^r(\ell)), \quad x \in \mathbb{R},$$

is the Legendre-Fenchel transform of Λ^r and we have used the facts that $t > \frac{2}{\varepsilon}$ and $\Lambda^{r,*}$ takes values in $[0, \infty]$, is convex with a minimum of zero at $m^r = \frac{1}{\nu^r}$, $\Lambda^{r,*}(x)$ is decreasing (to zero) for $x < m^r$ and increasing from zero for $x > m^r$. Indeed, $\Lambda^{r,*}$ is strictly convex at $x = m^r$ and so,

(183) $$\Lambda^{r,*}\left(\frac{1}{\nu^r + \frac{1}{2}\varepsilon}\right) > 0.$$

Similarly, for $\varepsilon > 0$, $\chi^r \equiv \nu^r - \varepsilon > 0$, $\delta^r = \frac{\varepsilon}{2\nu^r}$, $t \geq 0$ and $\ell \geq 0$, we have

$$\begin{aligned}
\mathbf{P}(N^r(t) < \chi^r t) &\leq \mathbf{P}(X^r([\chi^r t] + 1) > t) \\
&= \mathbf{P}(\check{X}^r([\chi^r t]) > t - \zeta^r(1)) \\
&\leq \mathbf{P}(\check{X}^r([\chi^r t]) > t(1 - \delta^r)) + \mathbf{P}(\zeta^r(1) > \delta^r t) \\
&\leq \exp\left(-\ell t(1 - \delta^r) + \chi^r t\, \Lambda^r(\ell)\right) + \mathbf{P}(\zeta^r(1) > \delta^r t) \\
&\leq \exp\left(-\chi^r t\, \Lambda^{r,*}\left(\frac{1 - \delta^r}{\chi^r}\right)\right) + \mathbf{P}(\zeta^r(1) > \delta^r t) \qquad (184) \\
&\leq \exp\left(-\chi^r t\, \Lambda^{r,*}\left(\frac{1}{\nu^r}\left(1 + \frac{\varepsilon}{2(\nu^r - \varepsilon)}\right)\right)\right) + \mathbf{P}(\zeta^r(1) > \delta^r t)
\end{aligned}$$

where we have used the fact that $\Lambda^r(\ell) \geq 0$ for $\ell \geq 0$ and we have minimized over $\ell \geq 0$, which is the same as minimization over all ℓ since

(185) $$\frac{1 - \delta^r}{\chi^r} = \frac{1 - \frac{\varepsilon}{2\nu^r}}{\nu^r - \varepsilon} = \frac{1}{\nu^r}\left(1 + \frac{\varepsilon}{2(\nu^r - \varepsilon)}\right) > m^r.$$

Now suppose that for each r and $i = 2, 3, \ldots$,

(186) $$\zeta^r(i) = \frac{\nu}{\nu^r}\zeta(i),$$

where $\{\zeta(i)\}_{i=2}^{\infty}$ is a sequence of i.i.d. random variables with mean $m = 1/\nu$ where it is assumed that $\nu = \lim_{r \to \infty} \nu^r$ is finite and strictly positive, and there is a non-empty open neighborhood $\mathcal{O}$ of $0 \in \mathbb{R}$ such that for $i = 2, 3, \ldots$,

(187) $$\Lambda(\ell) \equiv \log \mathbf{E}\left(e^{\ell \zeta(i)}\right) < \infty \quad \text{for all } \ell \in \mathcal{O}.$$

Then (177) holds with $\mathcal{O}^r = \frac{\nu^r}{\nu}\mathcal{O}$ and for all $\ell \in \mathbb{R}$, $x \in \mathbb{R}$,

$$\Lambda^r(\ell) = \Lambda\left(\frac{\ell\nu}{\nu^r}\right), \qquad \Lambda^{r,*}(x) = \Lambda^*\left(\frac{x\nu^r}{\nu}\right), \tag{188}$$

where

$$\Lambda^*(x) \equiv \sup_{\ell \in \mathbb{R}} \left(\ell x - \Lambda(\ell)\right), \tag{189}$$

and Λ^* is convex, takes values in $[0,\infty]$, $\Lambda^*(\frac{1}{\nu}) = 0$, and $\Lambda^*(x)$ is increasing for $x > \frac{1}{\nu}$ and decreasing for $x < \frac{1}{\nu}$. Furthermore, since Λ is differentiable at 0, $\Lambda^*(x) > 0$ for $x \neq \frac{1}{\nu}$.

For $\varepsilon < \nu/2$, since $\nu^r \to \nu$, there is $r_\varepsilon \geq 1$ such that for all $r \geq r_\varepsilon$, $|\nu^r - \nu| < \varepsilon$ and

$$\frac{\nu^r}{\nu}\left(\frac{1}{\nu^r + \frac{1}{2}\varepsilon}\right) \leq \frac{1}{\nu}\left(\frac{1}{1+\frac{\varepsilon}{3\nu}}\right) < \frac{1}{\nu}, \tag{190}$$

$$\frac{1}{\nu}\left(1 + \frac{\varepsilon}{2(\nu^r - \varepsilon)}\right) \geq \frac{1}{\nu}\left(1 + \frac{\varepsilon}{2\nu}\right) > \frac{1}{\nu}. \tag{191}$$

On combining this with (188), the inequalities above for N^r yield for $r \geq r_\varepsilon$, $\kappa^r = \nu^r + \varepsilon$, $t > \frac{2}{\varepsilon}$:

$$\begin{aligned} \mathbf{P}(N^r(t) > \kappa^r t) &\leq \exp\left(-(\kappa^r t - 1)\,\Lambda^*\left(\frac{1}{\nu}\left(\frac{1}{1+\frac{\varepsilon}{3\nu}}\right)\right)\right) \\ &\leq \exp\left(-(\nu t - 1)\,\Lambda^*\left(\frac{1}{\nu}\left(\frac{1}{1+\frac{\varepsilon}{3\nu}}\right)\right)\right) \end{aligned} \tag{192}$$

and for $r \geq r_\varepsilon$, $\chi^r = \nu^r - \varepsilon$, $\delta^r = \frac{\varepsilon}{2\nu^r}$, $t \geq 0$:

$$\mathbf{P}(N^r(t) < \chi^r t) \leq \exp\left(-\chi^r t\,\Lambda^*\left(\frac{1}{\nu}\left(1+\frac{\varepsilon}{2\nu}\right)\right)\right) + \mathbf{P}(\zeta^r(1) > \delta^r t), \tag{193}$$

where the values of the quantities involving Λ^* in the above are strictly positive.

In our application, the following will be needed to estimate the last probability in (193) above. Let $\{v^r(i)\}_{i=1}^\infty$ be a sequence of i.i.d. random variables, each with the same distribution as the $\zeta^r(i)$, $i = 2, 3, \ldots$ satisfying (186) above. Then for any $n \geq 1$, and a fixed $\ell_0 \in \mathcal{O}$ such that $\ell_0 > 0$,

$$\mathbf{P}\left(\max_{i=1}^{n} v^r(i) > \delta^r t\right) \leq n\mathbf{P}\left(v^r(1) > \delta^r t\right) \leq \exp\left(\log n - \frac{\ell_0 \varepsilon t}{2\nu} + \Lambda(\ell_0)\right), \tag{194}$$

where we have used the fact that $\delta^r = \frac{\varepsilon}{2\nu^r}$ and (188) with $\ell = \ell_0\nu^r/\nu$. In our application, both t and n will depend on r and tend to infinity as r tends to infinity, in such a way that $t^r = O(\log r)$ and $n^r = O(r^2)$. We see from the above that if $t^r/\log r$ is sufficiently large for all r sufficiently large, then we will have

$$\mathbf{P}\left(\max_{i=1}^{n^r} v^r(i) > \delta^r t^r\right) \to 0 \text{ as } r \to \infty. \tag{195}$$

Indeed, this will still hold if we multiply the left member above by a quantity of order r^2, provided $t^r/\log r$ is even a little bigger.

B Appendix: One-dimensional Skorokhod Problem

The following result is used several times in the paper. We record it here for ease of reference. For a proof, see [3] or Theorem 5.1 of [35].

Proposition B.1 *Let $x \in \mathbf{D}$ such that $x(0) = 0$. Then there is a unique pair $(w, v) \in \mathbf{D}^2$ such that*

(i) $w(t) = x(t) + v(t) \geq 0$ *for all* $t \geq 0$,

(ii) v *is non-decreasing and* $v(0) = 0$,

(iii) $\int_{[0,\infty)} 1_{(0,\infty)}(w(t))\, dv(t) = 0$.

This unique solution is given by (w^*, v^*) where for each $t \geq 0$,

$$v^*(t) = -\inf_{0 \leq s \leq t} x(s), \qquad w^*(t) = x(t) + v^*(t). \tag{196}$$

Furthermore, for any pair $(w, v) \in \mathbf{D}^2$ satisfying (i)–(ii) (but not necessarily (iii)), we have for all $t \geq 0$,

$$v(t) \geq v^*(t), \qquad w(t) \geq w^*(t). \tag{197}$$

References

[1] Billingsley, P. (1968). *Convergence of Probability Measures.* John Wiley and Sons, New York.

[2] Bramson, M. (1996). Convergence to equilibria for fluid models of FIFO queueing networks. *Queueing Systems.* **22** 5–45.

[3] Chen, H. and Mandelbaum, A. (1991). Leontief Systems, RBV's and RBM's. In *Applied Stochastic Analysis,* M. H. A. Davis and R. J. Elliott (eds.). Gordon and Breach Science Publishers, New York, pp. 1-43.

[4] Chevalier, P. B. and Wein, L. (1993). Scheduling networks of queues: heavy traffic analysis of a multistation closed network. *Operations Research.* **41** 743–758.

[5] Dembo, A. and Zeitouni, O. (1998). *Large Deviations Techniques and Applications.* 2nd edition, Springer, New York.

[6] Ethier, S. N. and Kurtz, T. G. (1986). *Markov Processes: Characterization and Convergence.* Wiley, New York.

[7] Harrison, J. M. (1985). *Brownian Motion and Stochastic Flow Systems.* John Wiley and Sons, New York.

[8] Harrison, J. M. (1988). Brownian models of queueing networks with heterogeneous customer populations. In *Stochastic Differential Systems, Stochastic Control Theory and Their Applications, IMA Volume 10*, W. Fleming and P.L. Lions (eds.). Springer Verlag, New York, pp. 147–186.

[9] Harrison, J. M. (1996). The BIGSTEP approach to flow management in stochastic processing networks. In *Stochastic Networks: Theory and Applications,* F. P. Kelly, S. Zachary and I. Ziedins (eds.). Oxford University Press, pp. 57–90.

[10] Harrison, J. M. (1998). Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review policies. *Annals of Applied Probability.* **8** 822-848.

[11] Harrison, J. M. (1998). Brownian models of open processing networks: canonical representation of workload. To appear in *Annals of Applied Probability.*

[12] Harrison, J. M. and López, M. J. (1998). Heavy traffic resource pooling in parallel-server systems. To appear in *Queueing Systems.*

[13] Harrison, J. M. and Van Mieghem, J. A. (1997). Dynamic control of Brownian networks: state space collapse and equivalent workload formulations. *Annals of Applied Probability.* **7** 747–771.

[14] Harrison, J. M. and Wein, L. (1989). Scheduling networks of queues: heavy traffic analysis of a simple open network. *Queueing Systems.* **5** 265–280.

[15] Harrison, J. M. and Wein, L. (1990). Scheduling networks of queues: heavy traffic analysis of a two-station closed network. *Operations Research.* **38** 1052–1064.

[16] Iglehart, D. L. and Whitt, W. (1971). The equivalence of functional central limit theorems for counting processes and associated partial sums. *Annals of Mathematical Statistics.* **42** 1372–1378.

[17] Jordan, W. C. and Graves, C. (1995). Principles on the benefits of manufacturing process flexibility. *Management Science.* **41** 577–594.

[18] Kelly, F. P. and Laws, C. N. (1993). Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling. *Queueing Systems.* **13** 47–86.

[19] Kumar, S. (1998). Two-server closed networks in heavy traffic: diffusion limits and asymptotic optimality. To appear in *Annals of Applied Probability.*

[20] Kushner, H. J. and Chen, Y. N. (1998). Optimal control of assignment of jobs to processors under heavy traffic. To appear in *Stochastics.*

[21] Kushner, H. J. and Dupuis, P. (1992). *Numerical Methods for Stochastic Control Problems in Continuous Time.* Springer-Verlag, New York.

[22] Kushner, H. J. and Martins, L. F. (1990). Routing and singular control for queueing networks in heavy traffic. *SIAM J. Control and Optimization.* **28** 1209–1233.

[23] Kushner, H. J. and Martins, L. F. (1996). Heavy traffic analysis of a controlled multiclass queueing network via weak convergence methods. *SIAM J. Control and Optimization.* **34** 1781–1797.

[24] Laws, C. N. (1992). Resource pooling in queueing networks with dynamic routing. *Adv. Appl. Prob.* **24** 699–726.

[25] Laws, C. N. and Louth, G. M. (1990). Dynamic scheduling of a four-station queueing network. *Prob. Eng. Inf. Sci.* **4** 131–156.

[26] Martins, L. F., Shreve, S. E. and Soner, H. M. (1996). Heavy traffic convergence of a controlled, multi-class queueing system. *SIAM J. Control and Optimization.* **34** 2133–2171.

[27] Mitrani, I. (1998). *Probabilistic Modelling.* Cambridge University Press.

[28] Peterson, W. P. (1991). A heavy traffic limit theorem for networks of queues with multiple customer types. *Mathematics of Operations Research.* **16** 90–118.

[29] Puhalskii, A. A. and Reiman, M. I. (1998). A critically loaded multirate link with trunk reservation. *Queueing Systems.* **28** 157–190.

[30] Rockafellar, R. T. (1970). *Convex Analysis.* Princeton University Press.

[31] Roughan, M. and Pearce, C. E. M. (1999). A martingale analysis of hysteretic overload control. Preprint.

[32] Teh, Y. C. (1999). *Threshold Routing Strategies for Queueing Networks.* D. Phil. thesis, University of Oxford.

[33] Van Mieghem, J. A. (1995). Dynamic scheduling with convex delay costs: the generalized $c\mu$ rule. *Annals of Applied Probability.* **5** 808–833.

[34] Wein, L. (1990). Scheduling networks of queues: heavy traffic analysis of a two-station network with controllable inputs. *Operations Research.* **38** 1065–1078.

[35] Williams, R. J. (1998). An invariance principle for semimartingale reflecting Brownian motions in an orthant. *Queueing Systems.* **30** 5–25.

[36] Williams, R. J. (1998). Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Systems.* **30** 27–88.

[37] Williams, R. J. (1999). On dynamic scheduling of a parallel server system with complete resource pooling. To appear in *Proceedings of the Fields Institute Workshop on Analysis and Simulation of Communication Networks.* American Mathematical Society, Providence, RI.