The Annals of Applied Probability
2000, Vol. 10, No. 3, 930–961

TWO-SERVER CLOSED NETWORKS IN HEAVY TRAFFIC: DIFFUSION LIMITS AND ASYMPTOTIC OPTIMALITY

BY SUNIL KUMAR

Stanford University

One of the successes of the Brownian approximation approach to dynamic control of queueing networks is the design of a control policy for closed networks with two servers by Harrison and Wein. Adopting a Brownian approximation with only heuristic justification, they interpret the optimal control policy for the Brownian model as a static priority rule and conjecture that this priority rule is asymptotically optimal as the closed networks's population becomes large.

This paper studies closed queueing networks with two servers that are balanced, that is, networks that have the same relative load factor at each server. The validity of the Brownian approximation used by Harrison and Wein is established by showing that, under the policy they propose, the diffusion-scaled workload imbalance process converges weakly in the infinite population limit to the diffusion predicted by the Brownian approximation. This is accomplished by proving that the fluid limits of the queue length processes undergo state space collapse in finite time under the proposed policy, thereby enabling the application of a powerful new technique developed by Williams and Bramson that allows one to establish convergence of processes under diffusion scaling by studying the behavior of limits under fluid scaling.

A natural notion of asymptotic optimality for closed queueing networks is defined in this paper. The proposed policy is shown to satisfy this definition of asymptotic optimality by showing that the performance under the proposed policy approximates bounds on the performance under every other policy arbitrarily well as the population increases without bound.

1. Introduction. Multiclass closed queueing networks are natural models of manufacturing systems in which the work in process inventory is held constant, either by physical constraints [21] or management policy [22]. In designing scheduling policies for such systems, the objective is to minimize idleness of the servers in the network, or equivalently, to maximize the throughput rate at each server in the network. That optimal control problem is intractable, and thus it is natural to seek "good" policies using approximate analysis.

One such approximate method is the Brownian approximation approach pioneered by Harrison [11]. In this method, suitably scaled versions of performance processes, like queue lengths, are studied as the population r in the closed network is increased without bound. A limiting control problem is derived in terms of the possible limits of the scaled processes. The solution to the

Received September 1998; revised November 1999.

AMS 1991 *subject classifications*. Primary 60K25; secondary 60F17, 90F35.

Key words and phrases. Queueing networks, functional limit theorems, diffusion limits, fluid limits, scheduling policies.

limiting control problem must then be interpreted to obtain a scheduling policy for the original closed queueing network with finite population. Needless to say, this approach is most useful when the limiting control problem is easily solved and the solution to the limiting control problem is easily interpreted.

One of the successes of the Brownian approximation approach to dynamic control of closed queueing networks is the design of a control policy for *two-server* closed networks by Harrison and Wein [14]. They recast the limiting control problem in terms of the "workload imbalance" process and thus render it trivial. They interpret the optimal control policy for the limiting problem as a static priority rule, and conjecture that this priority rule is asymptotically optimal in the infinite population limit under "balanced loading." The Harrison–Wein priority rule, which prioritizes customer classes according to a certain "workload imbalance index," is by no means obvious, and no simpler means of deriving it have been proposed to date. Simulation studies [14] and numerical bounds [17, 16] suggest that the performance of this priority rule is excellent and that it could indeed be asymptotically optimal, but evaluation of this policy remains incomplete even though it was proposed a decade ago.

This paper is devoted to the asymptotic analysis of the Harrison–Wein (HW) policy in closed queueing networks with two servers that are "balanced," and it fills several gaps in the existing literature. The restriction to balanced networks, in which the relative load factors (or traffic intensity parameters) for the two servers are exactly equal, is mainly for notational convenience. All the results can be extended to the case where the relative load factors approach a common limit as the population size approaches infinity, but the generalization does not appear to be worth the added notational complexity. The main results of this paper are the following.

1. The scaled workload imbalance process under the HW priority rule is shown to converge weakly to a two-sided regulator applied to a Brownian motion (cf. [10], Chapter 5). Since this is the diffusion obtained by Harrison and Wein [14] as the solution to the limiting control problem, we establish the validity of the Brownian approximation approach used by Harrison and Wein and of their interpretation of the solution to the limiting control problem as a priority rule.
2. We show that the policy proposed by Harrison and Wein is asymptotically optimal in the following sense. Denote the fixed population level in the network by r, and the cumulative idleness incurred by server 1 up to time t by $U_1^r(t)$. Consider $\lim_{r\to\infty} \mathbf{E}[U_1^r(r^2T)]/rT$ for a fixed, large T. In a loose sense, this quantity is a measure of the rate at which the long-run average idleness approaches zero as the population increases without bound. We show that this limit exists and is finite for the HW policy and that the value of this limit is no smaller for any other nonidling policy.

The key to demonstrating these results is a powerful new technique developed in three recent papers by Williams [23, 24] and Bramson [3]. In this technique, establishing state space collapse under fluid scaling is sufficient to establish state space collapse under diffusion scaling, and that in turn is sufficient to establish convergence under diffusion scaling. Thus, the difficult problem of proving tightness of diffusion-scaled processes is reduced to the relatively easier task of establishing state space collapse under fluid scaling. There are undoubtedly many other problems of asymptotic performance analysis that will be rendered tractable by the Williams–Bramson technique. In order to enable application of this technique, we analyze the fluid limits of the performance processes and obtain the following results about the fluid limits that are interesting in their own right.

3a. Server idleness is *never* incurred in the fluid limit under the HW policy.
3b. Queue length processes undergo *state space collapse* in the fluid limit in finite time under the HW policy. That is, after a finite time only two customer classes have nonzero amounts of fluid in them.

There are very few such proofs of asymptotic optimality for scheduling policies in network settings. One such result is provided by Martins, Shreve and Soner [20] who revisit a simple open queueing network with two servers and three customer classes that was first considered by Harrison and Wein [13]. Restricting attention to a Markovian setting, [20] establishes asymptotic optimality of a scheduling policy. Their method of proof, involving viscosity solutions, is quite different from the Williams–Bramson technique used in this paper and appears less amenable to application in more general network settings.

The rest of the paper is structured as follows. In Section 2, the model of the multiclass two-server closed queueing network is described. The notion of asymptotic optimality is laid out in Section 3 and its connection to the notion of asymptotic optimality proposed by Jin, Ou and Kumar [16] is described. The HW static priority policy is described in Section 4. In Section 5, we analyze the fluid limits of the network operated under the HW policy and we establish state space collapse. A key result proved that is necessary for this analysis is that there is *never* any idleness incurred on a fluid scale under the HW policy. Then, using the result due to [3], we establish that there is state space collapse under diffusion scaling as well. In Section 6, we prove the functional limit theorem for the diffusion-scaled workload imbalance process under the HW policy, using the state space collapse result proved in the previous section. The methodology in Section 6 closely follows [23, 24] as well as [5]. In Section 7, we establish asymptotic optimality of the HW policy. First, we establish that the limit in 2. above exists and is finite for the HW policy using uniform integrability arguments. Second, we show that this limit is minimal for the HW policy by showing that the idleness process under the HW policy achieves the same weak limit as pathwise lower bounds on the idleness processes under every other policy. Finally, in Section 8, we describe some of the possible extensions of the results obtained in this paper.

2. The model.

In this section we describe the model of the two-server closed queueing network that will be analyzed in the rest of the paper. The model described here can be considered as a closed network and a restricted version of the one considered by Williams [23]. The reader will do well to consult the excellent exposition in Section 3 of [23]. All the random variables defined below are defined over a probability space $(\Omega, \mathscr{F}, \mathbf{P})$.

2.1. *Network structure.* The queueing network we consider in this paper consists of *two* servers, $j = 1, 2$. At any given time, each customer in the network belongs to one of the K customer classes, $k = 1, 2, \dots, K$. Classes $k = 1, 2, \dots, k_1$ are served by server 1 and the remaining classes $k = k_1+1, \dots, K$ are served by server 2. Let C denote the $2 \times K$ constituency matrix given by

$$C = [C_{ij}] \quad \text{where } C_{ij} = \begin{cases} 1, & \text{if } i = 1 \text{ and } 1 \le j \le k_1, \\ 1, & \text{if } i = 2 \text{ and } k_1 + 1 \le j \le K, \\ 0, & \text{otherwise.} \end{cases}$$

2.2. *Initialization of population.* In a closed network, the population in the system remains constant for all time $t \ge 0$ and hence we need to populate the system at time $t = 0$. For simplicity and concreteness we will *always* assume the following scheme for populating the network. At time $t = 0$, r customers appear instantaneously in class $k = 1$ in some particular order, and the first customer at the head of the line *begins* service at server $j = 1$ at $t = 0$. (Although the traditional notation for the population of a closed queueing network is N, we will use r to be consistent with our principal references [23, 24] and [3].) Denote by $Z^r(0)$, the $K \times 1$ initial queue length vector. Thus, we assume that $Z^r(0) = (r, 0, 0, \dots, 0)'$.

2.3. *Service times.* For $k = 1, 2, \dots, K$, we denote by $v_k(i)$ the time taken to complete service for the ith customer arriving into class k after time zero. We denote their cumulative sums by $V_k(n) := \sum_{i=1}^n v_k(i)$, and the associated vector by $V(n) = (V_k(n),\ k = 1, \dots, K)$. Note that, because of our population initialization scheme, we do not have any customers that have been partially serviced at time zero. Hence we do not have to worry about residual service times at time zero. Denote by $S_k(t)$ the associate renewal process defined by $S_k(t) = \max\{n\colon V_k(n) \le t\}$. Denote by M the $K \times K$ diagonal matrix whose (k, k)th entry is m_k. We make the following assumptions about the service times $\{v_k(i)\}$ for $i = 1, 2, 3, \dots$, and $k = 1, 2, \dots, K$:

(1) For each k, $\{v_k(i),\ i = 1, 2, \dots\}$ are i.i.d. positive random variables.

$$\mathbf{E}[v_k(1)] = m_k < \infty \tag{2}$$

and

$$\operatorname{Var}[v_k(1)] = b_k < \infty. \tag{3}$$

We also assume if $k \ne l$ then the sequences $v_k(\cdot)$ and $v_l(\cdot)$ are independent.

2.4. *Routing.* For $k = 1, 2, \dots, K$ and for $i = 1, 2, \dots$, let $\phi^k(i)$ be a random vector taking values in the set of Euclidean unit bases in $\mathbf{R}^K$, $\{e_1, e_2, \dots, e_K\}$. On completion of the ith class k service, the customer immediately becomes a class l customer if $\phi^k(i) = e_l$. We denote by $\Phi^k(n) := \sum_{i=1}^n \phi^k(i)$ the cumulative sum.

We assume that $\{\phi^k(i), i = 1, 2, \ldots\}$ are i.i.d and denote $P_{kl} = \mathbf{P}\{\phi^k(i) = e_l\}$. Let P denote the $K \times K$ matrix whose (k, l)th entry is P_{kl}. We assume that P is stochastic; that is, $\sum_{l=1}^{K} P_{kl} = 1$ for all $k = 1, \ldots K$; and is irreducible. That is, we assume there exists a unique $1 \times K$ positive vector π, with $\sum_{k=1}^{K} \pi_k = 1$, satisfying $\pi P = \pi$. Under the assumption of stochasticity, the total number of customers in the system remains constant at the initial population level r. Although the stochasticity assumption is natural and is indeed necessary for the network to be closed, the irreduciblity assumption is not without loss of generality. We are restricting attention to the class of closed queueing networks commonly called *single chain* networks. For a discussion of the applicability of the Harrison–Wein policy to multichain networks, in which P is reducible, see [14]. However, we make no claim that the results of this paper extend to multichain networks.

We use the notation $\widetilde{P}$ to denote the transpose of P, $\widetilde{P_k}$ to denote the kth column of $\widetilde{P}$ and $(\widetilde{P})^n$ to denote the nth (matrix) power of $\widetilde{P}$. Note that

$$\mathbf{E}[\phi^k(i)] = \widetilde{P^k} \quad \text{and} \quad \mathrm{Cov}[\phi^k(i)] = \Upsilon^k, \tag{4}$$

where Υ^k is the $K \times K$ matrix defined by

$$\Upsilon^k_{lm} = \begin{cases} P_{kl}(1 - P_{kl}), & \text{if } l = m, \\ -P_{kl}P_{km}, & \text{if } l \neq m. \end{cases} \tag{5}$$

Let λ^* be the unique $K \times 1$ positive vector satisfying

$$\widetilde{P}\lambda^* = \lambda^*, \tag{6}$$

$$\max(\rho_1^*, \rho_2^*) = 1 \quad \text{where } \rho^* = CM\lambda^*. \tag{7}$$

The vector λ^* above can be interpreted as the maximum sustainable long-run average rate of departures from each of the classes, and ρ^* can be interpreted as the vector of relative load factors.

ASSUMPTION 2.1. *In this paper, we only consider* balanced *networks, in which* $\rho_1^* = \rho_2^* = 1$.

The reader may be led to believe that we are only considering a very special subclass of two-station closed queueing networks. However, in order to use Brownian approximations like those used in [14], one requires that the networks be at least asymptotically balanced. That is, one can consider a sequence of systems indexed by r. In the rth system, the mean service time matrix M_r and the routing matrix P_r depend on r and the population in this system is r. One obtains the corresponding sequences λ_r^* and ρ_r^*, and one requires that $r(\rho_{r,1}^* - \rho_{r,2}^*)$ converge to some finite limit θ as $r \to \infty$. Hence, we make the balanced network assumption and only allow the population level r to vary in our sequence of systems. This assumption is also justified because the asymptotic behavior of a system with just one bottleneck server tends to resemble that of a system with only one server, and hence is less interesting. We will have more to say about relaxing Assumption 2.1 in Section 8.

2.5. *Admissible service disciplines or scheduling policies.* In this paper, we restrict attention to service disciplines that are *nonidling*. That is, the server is busy if there is at least one customer awaiting service at that server. Also, we only consider service disciplines that serve customers within a class in FIFO fashion, and devote server effort only to the first customer in the class. A customer arriving at a class (subsequent to departure from another class) joins the end of line of customers awaiting service in that class. These disciplines are the so-called *head-of-line* service disciplines.

In order to further specify the allowed service disciplines, we consider the sequence of increasing times $\{\sigma_\ell\}_{\ell=1}^{\infty}$, where σ_ℓ is the time at which the ℓth change in the cumulative number of arrivals to any one of the classes (and consequently, the ℓth change in the cumulative number of departures from some other class) takes place. Let the $r \times 2$ matrix $\mathscr{O}^\ell$ denote the (class, age) pairs of each of the r customers in the system at time σ_ℓ, where "age" denotes the time spent in the current class since arriving in that class. We require that each of the servers devote a proportion u_k of the time interval $[\sigma_\ell, \sigma_{\ell+1})$ on the head of the line customer (the one with the largest age) in each class k served by that server and that this proportion be *determined completely by* $\mathscr{O}^\ell$. Equivalently, we specify an admissible scheduling policy by a mapping u: $\{1, \dots, K\}^r \times \mathbf{R}_+^r \to [0,1]^K$. The reader should note that our class of admissible policies is fairly large and includes commonly used scheduling policies like FIFO, Shortest Next Queue, Least Work Next Queue, as well as preemptive-resume static priority rules.

2.6. *Performance processes.* All processes in this paper have paths in $\mathbf{D}_{\mathbf{R}^d}[0,\infty)$, the space of right continuous R^d-valued functions with left limits, endowed with the usual Skorohod topology (cf. [9], Section 3.5) for some appropriate dimension d. Let $\mathscr{M}_d$ be the Borel σ-algebra on $\mathbf{D}_{\mathbf{R}^d}[0,\infty)$. All stochastic processes are measurable functions from $(\Omega, \mathscr{F}, \mathbf{P})$ into $(\mathbf{D}_{\mathbf{R}^d}[0,\infty), \mathscr{M}_d)$. From here on, let e denote a vector of all ones, whose dimension is specified by the context.

Let $A_k(t)$ denote the number of customers who have arrived at class k in $[0,t]$; $D_k(t)$, the number of customers who have departed from class k in $[0,t]$; $Z_k(t)$, the number of customers in class k at time t; $T_k(t)$, the total time devoted to serving class k in $[0,t]$ by the corresponding server and $U_j(t)$ the total time server j is idle in $[0,t]$. Note that specifying $T_k(\cdot)$ for each $k = 1, \dots, K$ for each realization $\omega \in \Omega$, is equivalent to specifying the policy. Recall that $\Phi^k(\cdot)$ is the cumulative routing vector defined in Section 2.4.

Using the convention that a quantity without a subscript denotes the vector of the corresponding subscripted quantities, we obtain the following network equations under *any* nonidling, head-of-line policy. Equations (8)–(12) follow directly from the definitions of the various quantities involved.

Equation (13) is a convenient representation of the nonidling nature of the policy:

$$A(t) = \sum_k \Phi^k(D_k(t)), \tag{8}$$

$$D(t) = S(T(t)), \tag{9}$$

$$Z(t) = Z(0) + A(t) - D(t), \tag{10}$$

$$CT(t) + U(t) = et, \tag{11}$$

$$\sum_{k=1}^{K} Z_k(t) = r \quad \text{for all } t \geq 0, \tag{12}$$

$$\int_0^\infty \left(\sum_{k=1}^{k_1} Z_k(t) \right) dU_1(t) = \int_0^\infty \left(\sum_{k=k_1+1}^{K} Z_k(t) \right) dU_2(t) = 0. \tag{13}$$

These equations do not completely specify the behavior of the network operated under a given policy u. We will need to add policy specific equations in order to completely specify the behavior of the system, and we will indeed do so later, when we consider the static priority policy proposed by Harrison and Wein. In the sequel, when necessary, we will attach a superscipt r and a subscript u to each of these quantities to make explicit the dependence of these quantities on the fixed population level r and the control policy u that is employed. For example, $U^r_{u,\,j}(t)$ denotes the cumulative idleness incurred by server j up to time t when the population level is r and the control policy employed is u. (Actually, one should index the policy u by r, since the policy may depend on the population level. To simplify notation, we will not make this dependence explicit.)

3. Asymptotic performance criteria.

The analytical quest for an "optimal" policy that either maximizes the throughput of the system or minimizes server idleness at every population level r is hopeless. One is therefore driven toward looking for policies that outperform other policies asymptotically as the population increases without bound. Hence, we are led to defining asymptotic performance criteria. In this section, we define an asymptotic performance criterion that is sufficiently discriminating, and yet, for which verifying whether a given policy optimizes this criterion is tractable. One asymptotic performance criterion for weeding out poor policies is the notion of efficiency defined below.

DEFINITION 3.1. An admissible policy u is said to be *efficient* if

$$\limsup_{r\to\infty} \limsup_{T\to\infty} \frac{\mathbf{E}[U^r_{u,\,j}(T)]}{T} = 0, \tag{14}$$

for some $j \in \{1, 2\}$.

The $j \in \{1,2\}$ that achieves the limit above identifies the bottleneck server. Note that in the case of balanced networks, one expects (14) to hold for both $j = 1$ and 2 or neither.

Efficiency requires that the time averaged idleness for at least one of the servers approach zero as the population in the closed network is increased without bound. This is the most basic requirement on the asymptotic performance of any "good" policy u. However, it is not a very discriminating performance criterion in the sense that it is very likely that a whole host of policies will be efficient for a given closed network. See [18] for the discussion of a class of closed queueing networks called reentrant lines, in which at least two policies are efficient. A more discriminating performance measure will be one that not only requires the time average of the idleness to go to zero as the population increases without bound, but also requires that is decrease to zero sufficiently fast as r increases. The natural requirement for a policy u^* to be optimal under such a performance criterion is that it satisfy (15) below, adapted from [16]. For every admissible policy u, we require that

$$(15)\qquad \limsup_{r\to\infty}\limsup_{T\to\infty}\frac{r\mathbf{E}[U^r_{u^*,i}(T)]}{T} \le \liminf_{r\to\infty}\liminf_{T\to\infty}\frac{r\mathbf{E}[U^r_{u,i}(T)]}{T}$$

for some $i \in \{1,2\}$. Implicit in this definition is the assumption that the average idleness approaches zero at a rate at least as fast as r^{-1}, without which (15) is vacuous.

This definition requires that we compare long term time-average behavior under each of the policies and is quite hard to verify. Instead, for a fixed large t, if we assume that a time horizon of r^2T time units is sufficiently long (for near steady-state behavior) in a network with population r, then we can use the weaker, but easier to verify notion of asymptotic optimality described below.

DEFINITION 3.2. An admissible policy u^* is said to be *asymptotically optimal* if for every admissible policy u, every fixed time T, and some $i \in \{1,2\}$,

$$(16)\qquad \limsup_{r\to\infty}\frac{\mathbf{E}[U^r_{u^*,i}(r^2T)]}{r} \le \liminf_{r\to\infty}\frac{\mathbf{E}[U^r_{u,i}(r^2T)]}{r}.$$

Of course, Definition 3.2 only makes sense if the left-hand side above is finite for at least one policy u^*, and demonstrating this will be one of our goals in this paper. In fact, we will show that (16) holds when u^* is the HW policy and give an expression for the left-hand side of (16) in terms of the two-sided regulator applied to a Brownian motion (cf. [10], Chapter 5).

4. The Harrison–Wein policy. In this section we will describe the policy designed by Harrison and Wein [14]. Our description is equivalent, but not identical, to their original description. The treatment in this section is based on [12].

Recall that π is the invariant row vector associated with the routing matrix P. Denoting (only in the context of this section) a $1 \times K$ vector of ones by e, we can define a $K \times K$ matrix Q and a $1 \times K$ row vector $\widehat{M}$ by

$$Q = (I - \widetilde{P} + \pi' e)^{-1} \tag{17}$$

and

$$\widehat{M} = [1 \quad -1]\, CMQ. \tag{18}$$

One useful interpretation of $\widehat{M}$ is the following, based on the workload imbalance process

$$\mathscr{W}_u^r(t) = \widehat{M} Z_u^r(t). \tag{19}$$

The choice of the symbol $\mathscr{W}$ is meant to avoid confusion with the *immediate workload* process, usually denoted by W. One can write Q as

$$Q = I + \sum_{i=1}^{\infty} \Big((\widetilde{P})^i - \pi' e \Big).$$

By Assumption 2.1, we have $[1 \quad -1]\, CM(\pi' e) = 0$ because π is some constant multiple of λ^*. Hence we obtain

$$\widehat{M} = \lim_{n \to \infty} [1 \quad -1]\, H(n) \quad \text{where}$$
$$H(n) = CM \left(I + \sum_{i=1}^{n} (\widetilde{P})^i \right).$$

The (j, k)th element of the $2 \times K$ matrix $H(n)$, $H_{jk}(n)$ can be interpreted as the expected total work required from server j in completing the first n services for a customer that starts in class k. Thus, one can interpret $\widehat{M}$ as a measure of *workload imbalance*.

Loosely speaking, customers in a class k with a large value of $\widehat{M}_k$ provide a lot more work for server 1 than for server 2. A plausible policy choice is one in which each server tries to keep the other server as busy as possible, thus minimizing idleness in the system. One way to achieve this is to give *higher* priority to classes with *smaller* values of $\widehat{M}_k$ at server 1, and to reverse this rule at server 2, that is, to give *higher* priority to classes with *larger* values of $\widehat{M}_k$. This is the Harrison–Wein policy.

DEFINITION 4.1. The Harrison–Wein (HW) policy is a static, preemptive-resume, priority rule that gives higher priority to classes with *lower* values of $\widehat{M}_k$ (for $k = 1, 2, \ldots, k_1$) at server 1 and to classes with *higher* values of $\widehat{M}_k$ (for $k = k_1 + 1, \ldots, K$) at server 2. Ties are broken in lexicographic order.

We should point out that Harrison and Wein arrived at this policy via reasoning that is far more sophisticated than the loose argument given above. We will point out details of their reasoning later in this paper, when we analyze the policy defined above in greater detail. We make the following mild assumption that is almost without loss of generality. The only real assumption is the lack of ties, that is, $\widehat{M}_k \neq \widehat{M}_l$ any two classes k and l served at the same server.

ASSUMPTION 4.1. *Assume that the classes have been relabeled, and that the network primitive C, M and P result in no ties, so that the following hold:*

$$\widehat{M}_1 > \widehat{M}_2 > \cdots > \widehat{M}_{k_1}, \tag{20}$$

$$\widehat{M}_{k_1+1} > \widehat{M}_{k_1+2} > \cdots > \widehat{M}_K. \tag{21}$$

Under this relabeling of classes the lowest priority classes are class 1 at server 1 and class K at server 2. In the discussion following equation (69) of [12], it is established that $\widehat{M}_k$ for $k = 1, \ldots, K$ differ from the workload imbalance indices computed in Harrison and Wein only by a fixed additive constant. From equations (45)–(48) of [14] and the discussion following these equations there, we have

$$\widehat{M}_1 > \widehat{M}_K. \tag{22}$$

The following lemma, due to [12] describes a very important property of the workload imbalance indices that we will repeatedly use in the rest of the paper. The proof is short enough to be reproduced for completeness here.

LEMMA 4.1 (Harrison and Van Mieghem).

$$\widehat{M}(I - \widetilde{P}) = GCM, \tag{23}$$

where $G = [1 \quad -1]$.

PROOF.

$$\begin{aligned}\widehat{M}(I - \widetilde{P}) &= GCMQ(Q^{-1} - \pi' e)\\ &= GCM - GCMQ\pi' e.\end{aligned}$$

Since π is the invariant probability measure associated with P, we have $\pi e' = 1$. Thus $Q^{-1}\pi' = (I - \widetilde{P} + \pi' e)\pi' = \pi'$. So we have

$$\widehat{M}(I - \widetilde{P}) = GCM - GCM\pi' e.$$

Now π' is a multiple of λ^* (from (6) and the irreducibility of P). Thus, from (7) and Assumption 2.1, we have $GCM\pi' = 0$, completing the proof. □

5. Fluid limits and state space collapse under the HW policy.

In this section, we analyze the limits of the performance processes (8)–(13) under the HW policy, scaled in a fashion indicative of a functional law of large numbers. We will analyze these fluid limits and establish that state space collapse takes place in finite time. Using this, and invoking Bramson's result, we establish that state space collapse occurs under diffusion scaling. This result will be used in the next section to prove convergence of diffusion-scaled processes. Results that are labeled theorems in this section represent results that are not only intermediate steps in the scheme of the paper, but also results that are interesting in their own right.

The fluid scaling results in the scaled processes below are identified by a bar over the corresponding quantities. In order to simplify notation, we will omit the subscript u since we only will consider the Harrison–Wein policy in this section. We define

$$\overline{A}^r(t) = \frac{1}{r} A^r(rt), \tag{24}$$

$$\overline{D}^r(t) = \frac{1}{r} D^r(rt), \tag{25}$$

$$\overline{S}^r(t) = \frac{1}{r} S^r(rt), \tag{26}$$

$$\overline{T}^r(t) = \frac{1}{r} T^r(rt), \tag{27}$$

$$\overline{U}^r(t) = \frac{1}{r} U^r(rt), \tag{28}$$

$$\overline{\Phi}^{r,k}(x) = \frac{1}{r} \Phi^{r,k}([xr]) \tag{29}$$

and

$$\overline{Z}^r(t) = \frac{1}{r} Z^r(rt). \tag{30}$$

Note: The notation $[x]$ denotes the greatest integer less than or equal to the real number x.

The dynamics of these fluid-scaled processes is partially captured in (31)–(36), the analogs of (8)–(13) [18, 7, 3]:

$$\overline{A}^r(t) = \sum_k \overline{\Phi}^{r,k}(\overline{D}^r_k(t)), \tag{31}$$

$$\overline{D}^r(t) = \overline{S}(\overline{T}^r(t)), \tag{32}$$

$$\overline{Z}^r(t) = \overline{Z}^r(0) + \overline{A}^r(t) - \overline{D}^r(t), \tag{33}$$

$$C\overline{T}^r(t) + \overline{U}^r(t) = et, \tag{34}$$

$$\sum_{k=1}^{K} \overline{Z}^r_k(t) = 1 \quad \text{for all } t \geq 0 \tag{35}$$

and

$$\int_0^\infty \left(\sum_{k=1}^{k_1} \overline{Z}^r_k(t) \right) d\overline{U}^r_1(t) = \int_0^\infty \left(\sum_{k=k_1+1}^{K} \overline{Z}^r_k(t) \right) d\overline{U}^r_2(t) = 0. \tag{36}$$

To these equations we can add another set of equations that capture the additional properties of the HW policy. These equations capture the fact that the servers cannot be working on a lower priority class when there are higher priority customers awaiting service. Recall that $T^r_k(t)$ is the cumulative amount of time spent on class k by the corresponding server in $[0, t]$. At server 1, the HW policy forbids working on a class $k' < k$, if there is at least one customer in classes k through k_1. That is, $t - \sum_{l=k}^{k_1} \overline{T}_l^r(t)$ cannot be increasing at a time instant when $\sum_{l=k}^{k_1} \overline{Z}_l^r(t) > 0$. Similarly, at server 2, the HW policy forbids working on a class $k' > k$, if there is at least one customer in classes $k_1 + 1$ through k. That is, $t - \sum_{l=k_1+1}^{k} \overline{T}_l^r(t)$ cannot be increasing at a time instant when $\sum_{l=k_1+1}^{k} \overline{Z}_l^r(t) > 0$. Thus we have

$$\int_0^\infty \left(\sum_{l=k}^{k_1} \overline{Z}_k^r(t) \right) d\left(t - \sum_{l=k}^{k_1} \overline{T}_l^r(t) \right) = 0 \quad \text{for } k = 1, 2, \ldots, k_1 \tag{37}$$

and

$$\int_0^\infty \left(\sum_{l=k_1+1}^{k} \overline{Z}_l^r(t) \right) d\left(t - \sum_{l=k_1+1}^{k} \overline{T}_l^r(t) \right) = 0 \quad \text{for } k = k_1 + 1, \ldots, K. \tag{38}$$

The reader should note that (36) is subsumed by (37) and (38). The following theorem follows immediately from similar theorems due to [7], [3] and [18]. The connection between (31)–(38) above and (39)–(46) below is evident.

THEOREM 5.1. *Almost surely, every sequence $r \to \infty$ contains a subsequence $\{r_n\}$ such that the process $(\overline{A}^{r_n}, \overline{D}^{r_n}, \overline{Z}^{r_n}, \overline{T}^{r_n}, \overline{U}^{r_n})$ converges uniformly on compact time sets to some limit process $(\overline{A}, \overline{D}, \overline{Z}, \overline{T}, \overline{U})$. Furthermore, each limit process satisfies the following equations:*

$$\overline{A}(t) = \widetilde{P}\, \overline{D}(t), \tag{39}$$

$$\overline{D}(t) = M^{-1} \overline{T}(t), \tag{40}$$

$$\overline{Z}(t) = \overline{Z}(0) + \overline{A}(t) - \overline{D}(t), \tag{41}$$

$$C\overline{T}(t) + \overline{U}(t) = et, \tag{42}$$

$$\overline{T}(0) = \overline{U}(0) = 0,\ \overline{T}(\cdot), \overline{U}(\cdot) \textit{ are Lipschitz, increasing functions}, \tag{43}$$

$$\sum_{k=1}^{K} \overline{Z}_k(t) = 1 \quad \textit{for all } t \geq 0, \tag{44}$$

$$\int_0^\infty \left(\sum_{l=k}^{k_1} \overline{Z}_l(t) \right) d\left(t - \sum_{l=k}^{k_1} \overline{T}_l(t) \right) = 0 \quad \textit{for } k = 1, 2, \ldots, k_1 \tag{45}$$

and

$$\int_0^\infty \left(\sum_{l=k_1+1}^{k} \overline{Z}_l(t) \right) d\left(t - \sum_{l=k_1+1}^{k} \overline{T}_l(t) \right) = 0 \quad \textit{for } k = k_1 + 1, \ldots, K. \tag{46}$$

In the proof of Theorem 5.4, we will show that the above result holds for the more general notion of cluster point, as defined by [3]. The reader should not be concerned with the lack of a proof for this result here. Only the equations satisfied by the limits (39)–(46) are of interest to us in applying Bramson's results. We will call the equations (39)–(46) the *fluid model* under the HW policy.

Since every solution to these equations is Lipschitz continuous, we can talk of time derivatives at almost every time instant. Let us use the notation

$$(\dot{D}(t), \dot{Z}(t), \dot{T}(t), \dot{U}(t)) := \left(\frac{d\overline{D}(t)}{dt}, \frac{d\overline{Z}(t)}{dt}, \frac{d\overline{T}(t)}{dt}, \frac{d\overline{U}(t)}{dt}\right)$$

at any such "regular" time t where the derivatives exists. We begin our analysis of the fluid model under the HW policy by establishing a useful property of the solutions to the fluid model.

LEMMA 5.1. *For any solution to* (39)–(46)*, at every regular time t, we have*

$$\widehat{M}\dot{Z}(t) = \dot{U}_1(t) - \dot{U}_2(t). \tag{47}$$

PROOF. From (39) and (41), we have

$$\dot{Z}(t) = -(I - \widetilde{P})\dot{D}(t).$$

Therefore, using (23),

$$\widehat{M}\dot{Z}(t) = -\widehat{M}(I - \widetilde{P})\dot{D}(t) = -[1 \quad -1]\,CM\dot{D}(t).$$

Using (40) and (42) in the equation above gives us (47). □

We now present a result that is the same as Proposition 4.2 of [8]. We choose to reproduce it here for completeness and because we will use it repeatedly in the sequel.

LEMMA 5.2 (Dai and Weiss). *At any regular time t, the following are true for any solution to* (39)–(46)*:*

(i) *For any $k \in \{1, 2, \dots, K\}$, if $\overline{Z}_k(t) = 0$, then $\dot{Z}_k(t) = 0$ and $\dot{D}_k(t) = \sum_{l=1}^{K} P_{lk}\dot{D}_l(t)$.*

(ii) *At each server j, there is at most one class k^j such that $\overline{Z}_{k^j}(t) > 0$ and $\dot{D}_{k^j}(t) > 0$. Furthermore k^j is the highest priority nonempty buffer at server j. That is, if $j = 2$, then*

$$\sum_{k=k_1+1}^{k^2-1} \overline{Z}_k(t) = 0, \tag{48}$$

$$\sum_{k=k_1+1}^{k^2} m_k\dot{D}_k(t) = 1, \tag{49}$$

$$\sum_{k=k^2+1}^{K} \dot{D}_k(t) = 0. \tag{50}$$

Equations corresponding to (48)–(50) *can be obtained for server $j = 1$ as well.*

We are now ready to prove our first major result concerning the behavior of the fluid model under the HW policy. This is an important property of the HW policy, and it shows that no other policy can have better performance than the HW policy under fluid scaling. It also establishes that $\widehat{M}\,\overline{Z}$ is an invariant for (39)–(46).

THEOREM 5.2. *For any solution to* (39)–(46),

$$\overline{U}_1(t) = \overline{U}_2(t) = 0 \quad \textit{for all } t \geq 0. \tag{51}$$

Consequently,

$$\widehat{M}\,\overline{Z}(t) = \widehat{M}\,\overline{Z}(0) \quad \textit{for all } t \geq 0. \tag{52}$$

PROOF. Equation (51) is equivalent to saying that at every regular time t,

$$\dot{U}_1(t) = \dot{U}_2(t) = 0. \tag{53}$$

By (47), (53) also establishes (52). Suppose (53) is violated at some regular time t. Let us assume that $\dot{U}_1(t) > 0$ at time t. From (45), we must have $\sum_{k=1}^{k_1} \overline{Z}_k(t) = 0$ and consequently $\sum_{k=k_1+1}^{K} \overline{Z}_k(t) = 1$ and so $\dot{U}_2(t) = 0$ from (46). Also, since t is a regular point we must have, from Lemma 5.2(i),

$$\dot{Z}_k(t) = 0 \quad \text{for } k = 1, 2, \dots, k_1.$$

Let k^2 be the highest priority nonempty class at server 2, that is, $k^2 = \min\{k_1 + 1 \leq k \leq K \mid \overline{Z}_k(t) > 0\}$. Once again, from Lemma 5.2(i), we have

$$\dot{Z}_k(t) = 0 \quad \text{for } k = k_1 + 1, \dots, k^2 - 1.$$

Thus we have

$$\widehat{M}\dot{Z}(t) = \sum_{k=k^2}^{K} \widehat{M}_k \dot{Z}_k(t) = \dot{U}_1(t) - \dot{U}_2(t) > 0. \tag{54}$$

If $k^2 = K$, we must have $\dot{Z}_{k^2}(t) = -\sum_{k=1}^{K-1} \dot{Z}_k(t) = 0$, contradicting (54). If $k^2 < K$, since $\overline{Z}_{k^2}(t) = 1 - \sum_{k=k^2+1}^{K} \overline{Z}_k(t)$, we have

$$\sum_{k=k^2}^{K} \widehat{M}_k \dot{Z}_k(t) = \sum_{k=k^2+1}^{K} (\widehat{M}_k - \widehat{M}_{k^2}) \dot{Z}_k(t).$$

But for all k such that $k > k^2$, $\dot{D}_k(t) = 0$ from (50), and hence $\dot{Z}_k(t) \geq 0$. But from (21), for all $k > k^2$, we have $(\widehat{M}_k - \widehat{M}_{k^2}) \leq 0$. Hence we have

$$\sum_{k=k^2}^{K} \widehat{M}_k \dot{Z}_k(t) \leq 0,$$

contradicting (54). Thus $\dot{U}_1(t) = 0$ at all regular t. One can similarly establish that $\dot{U}_2(t) = 0$ at all regular t, thus completing the proof. □

Theorems 5.1 and 5.2 immediately yield the following result that will be needed in the next section.

COROLLARY 5.1. *Almost surely, for every fixed $t \geq 0$, under the HW policy we have*

$$\lim_{r\to\infty} \frac{U_1^r(r^2 t)}{r^2} = 0. \tag{55}$$

PROOF. Theorems 5.1 and 5.2 yield (55) along a subsequence of $\{r\}$. Since the right-hand side of (55) does not depend on the choice of the subsequence, we obtain the result. □

The following corollary is immediate from Lemma 3.5.3 of [18].

COROLLARY 5.2. *For every solution to* (39)–(46)*, we have*

$$\liminf_{t\to\infty} \frac{\overline{D}(t)}{t} \geq \lambda^*. \tag{56}$$

The next theorem, one of the central results of this paper, establishes an upper bound on the time taken for *every* solution to the fluid model to undergo state space collapse.

THEOREM 5.3. *There exists a time T depending only on M, C and P, such that for any solution to* (39)–(46)*, we have*

$$\overline{Z}_1(t) + \overline{Z}_K(t) = 1 \quad \textit{for all } t \geq T. \tag{57}$$

Furthermore, for each $\overline{Z}(0)$, there exists a $\overline{Z}^* \in \mathbf{R}_+^K$ that depends only on $\overline{Z}(0)$, such that $\overline{Z}_1^* + \overline{Z}_K^* = 1$ and

$$\overline{Z}(t) = \overline{Z}^* \quad \textit{for all } t \geq T. \tag{58}$$

Finally, if $\overline{Z}(0)$ satisfies $\overline{Z}_1(0) + \overline{Z}_K(0) = 1$, then

$$\overline{Z}(t) = \overline{Z}(0) \quad \textit{for all } t \geq 0. \tag{59}$$

This means that the collapse is not just to the manifold, but to a point on the manifold that is completely specified by the initial state.

PROOF OF THEOREM 5.3. First, given that there exists a T such that (57) holds for any solution to (39)–(46) and given a $\overline{Z}(0)$, let us establish the existence of a unique $\overline{Z}^* \geq 0$ satisfying $\overline{Z}_1^* + \overline{Z}_K^* = 1$ and (58) above. Note that we are studying all solutions to the fluid model (39)–(46), and hence any $\overline{Z}(0)$ satisfying (44) is admissible. Consider any time t such that $\overline{Z}_1(t) + \overline{Z}_K(t) = 1$. By (52), we must have

$$\widehat{M}_1 \overline{Z}_1(t) + \widehat{M}_K \overline{Z}_K(t) = \widehat{M}\, \overline{Z}(0).$$

Since $\widehat{M}_1 > \widehat{M}_K$ from (22), for any such time t, we must have

$$\begin{bmatrix} \overline{Z}_1(t) \\ \overline{Z}_K(t) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \widehat{M}_1 & \widehat{M}_K \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ \widehat{M}\,\overline{Z}(0) \end{bmatrix} =: \begin{bmatrix} \overline{Z}_1^* \\ \overline{Z}_K^* \end{bmatrix}.$$

Since $\widehat{M}\,\overline{Z}(0) \in [\widehat{M}_K, \widehat{M}_1]$, we have $\overline{Z}_1^* \geq 0$ and $\overline{Z}_K^* \geq 0$ from Cramér's rule. Choosing $\overline{Z}_2^* = \overline{Z}_3^* = \cdots = \overline{Z}_{K-1}^* = 0$ establishes (58).

Now let us establish the existence of a T such that (57) holds for any solution to (39)–(46). The key to proving this result is showing that $\overline{Z}_1(t) + \overline{Z}_K(t)$ is *always nondecreasing*, and that whenever $\overline{Z}_1(t) + \overline{Z}_K(t) < 1$ and $\dot{D}_1(t) > 0$, $\dot{Z}_1(t)$ is sufficiently positive, that is, $\overline{Z}_1(t) + \overline{Z}_K(t)$ grows sufficiently quickly.

From (52), we have established that, at any regular time t, $\widehat{M}\dot{Z}(t) = 0$. Also, from (56), for every $\varepsilon > 0$, there exists a time T such that for all $t \geq T$,

$$\int_0^t \dot{D}_1(s)\,ds \geq (\lambda_1^* - \varepsilon)\,t.$$

Now consider a regular time t such that $\dot{D}_1(t) > 0$. It must be true from Lemma 5.2 that, at this time t, for every k satisfying $2 \leq k \leq k_1$, $\overline{Z}_k(t) = \dot{Z}_k(t) = 0$. As before let k^2 be the highest priority nonempty class at server 2, that is, $k^2 = \min\{k_1 + 1 \leq k \leq K \mid \overline{Z}_k(t) > 0\}$. Once again, from (48) of Lemma 5.2, we have

$$\dot{Z}_k(t) = 0 \quad \text{for } k = k_1 + 1, \ldots, k^2 - 1.$$

If $k^2 = K$, we have

$$\overline{Z}_1(t) + \overline{Z}_K(t) = 1 \tag{60}$$

and from (52), we have

$$(\widehat{M}_1 - \widehat{M}_K)\dot{Z}_1(t) = 0.$$

From (22) and (21), we must have

$$\dot{Z}_k(t) = 0 \quad \text{for all } k = 1, 2, \ldots K, \tag{61}$$

thus establishing that $\overline{Z}_1(t) + \overline{Z}_K(t)$ is nondecreasing in this situation. This, along with the uniqueness of $\overline{Z}^*$ above, leads us to conclude that if $\overline{Z}(0)$ satisfies $\overline{Z}_1(0) + \overline{Z}_K(0) = 1$, then (59) must hold.

Now consider the situation when $k^2 < K$. From (52) we must have

$$\widehat{M}_1\dot{Z}_1(t) + \widehat{M}_{k^2}\dot{Z}_{k^2}(t) + \sum_{k=k^2+1}^{K} \widehat{M}_k\dot{Z}_k(t) = 0.$$

Since $\overline{Z}_{k^2}(t) = 1 - \sum_{k=k^2+1}^{K} \overline{Z}_k(t) - \overline{Z}_1(t)$, we have

$$\sum_{k=k^2+1}^{K} \big(\widehat{M}_k - \widehat{M}_{k^2}\big)\dot{Z}_k(t) = -\big(\widehat{M}_1 - \widehat{M}_{k^2}\big)\dot{Z}_1(t) > -\big(\widehat{M}_1 - \widehat{M}_K\big)\dot{Z}_1(t), \tag{62}$$

where the second inequality follows from (22). Therefore, from (21) and from (50), we have

$$\dot{Z}_1(t) \geq \frac{(\widehat{M}_{k^2} - \widehat{M}_{k^2+1})}{(\widehat{M}_1 - \widehat{M}_K)} \sum_{k=k^2+1}^{K} \dot{Z}_k(t) \geq 0. \tag{63}$$

Thus, once again we have established that $\overline{Z}_1(t)+\overline{Z}_K(t)$ is nondecreasing. Now we have to make sure that this quantity increases sufficiently fast. We claim that there exists a $\delta_{k^2} > 0$ that depends only on the routing matrix P such that

$$\sum_{k=k^2+1}^{K} \dot{Z}_k(t) \geq \delta_{k^2} \dot{D}_1(t). \tag{64}$$

This along with (63) would imply that there exists a δ_1 depending only on C, M and P such that

$$\dot{Z}_1(t) \geq \delta_1 \dot{D}_1(t) \quad \text{for all regular } t \geq 0. \tag{65}$$

By similar arguments (interchanging the roles of server 1 and server 2) one would also establish that there exists a δ_K depending only on C, M and P such that

$$\dot{Z}_K(t) \geq \delta_K \dot{D}_K(t) \quad \text{for all regular } t \geq 0. \tag{66}$$

Now fix $t > 0$. Suppose $\overline{Z}_1(s)+\overline{Z}_K(s) < 1$, for all $s \in [0, t]$. Then we must have

$$\begin{aligned} 1 &> \overline{Z}_1(t) + \overline{Z}_K(t) \\ &\geq \overline{Z}_1(t) + \overline{Z}_K(t) - \big(\overline{Z}_1(0) + \overline{Z}_K(0)\big) \\ &= \int_0^t \big(\dot{Z}_1(s) + \dot{Z}_K(s)\big)\, ds \\ &\geq \delta_1 \int_0^t \dot{D}_1(s)\, ds + \delta_K \int_0^t \dot{D}_K(s)\, ds \\ &= \delta_1 \overline{D}_1(t) + \delta_K \overline{D}_K(t). \end{aligned} \tag{67}$$

Thus,

$$\overline{D}_1(t) < \frac{1}{\delta_1}.$$

Now pick any k such that $2 \leq k \leq K$. Then, since P is irreducible, there exists a sequence of indices $i_1, i_2, \ldots, i_n$, with $i_1 = k$ and $i_n = 1$ such that

$$P_{i_1, i_2} P_{i_2, i_3} \cdots P_{i_{n-1}, i_n} > 0.$$

We suppress making the dependence on k of the sequence explicit for notational convenience.

We have

$$\overline{D}_1(t) < \frac{1}{\delta_1} \quad \Rightarrow \quad \overline{D}_{i_{n-1}}(t) < \frac{1}{P_{i_{n-1}, i_n}} \left[\frac{1}{\delta_1} + 1\right].$$

Repeating this procedure for one more index yields

$$\overline{D}_{i_{n-2}}(t) < \frac{1}{P_{i_{n-2},\, i_{n-1}}}\left[1 + \frac{1}{P_{i_{n-1},\, i_n}}\left[\frac{1}{\delta_1} + 1\right]\right].$$

We continue this procedure for each term in the sequence, and for each such sequence. Thus, we have

$$\overline{D}_k(t) < \varepsilon_k\left[\frac{1}{\delta_1} + \Delta_k\right] \quad \text{for all } k = 2, 3, \dots, K,$$

where each ε_k and each Δ_k is a constant that depends only on C and P. Using this, (67), and (40) and (42) we have

$$\overline{U}_1(t) > t - \sum_{k=1}^{k_1} m_k \varepsilon_k\left[\frac{1}{\delta_1} + \Delta_k\right].$$

Since from (51), $\overline{U}_1(t) = 0$, we must have

$$t < \sum_{k=1}^{k_1} m_k \varepsilon_k\left[\frac{1}{\delta_1} + \Delta_k\right].$$

Thus $\overline{Z}_1(T) + \overline{Z}_K(T) = 1$ for some $T \geq \sum_{k=1}^{k_1} m_k \varepsilon_k[1/\delta_1 + \Delta_k]$. Also, from (67), since $\overline{Z}_1(\cdot) + \overline{Z}_K(\cdot)$ is nondecreasing, $\overline{Z}_1(t) + \overline{Z}_K(t) = 1$ for all $t \geq T$. Thus, if we prove (64), we are done.

We now provide a proof of the claim (64). Let $P_{ij}^{(q)}$ denote the (i, j) element of the qth matrix power of P. By Theorem (2.1) of [1], we know that there exists a q such that $P_{1K}^{(q)} > 0$ because P is irreducible. Since $P_{1K}^{(q)} > 0$, there are only two possibilities:

(a) A product of routing probabilities of the form

$$P_{1,\, l_1} P_{l_1,\, l_2} P_{l_2,\, l_3} \cdots P_{l_{n-1},\, l_n} > 0, \tag{68}$$

for some $n \leq q$, where $l_1 \neq 1$, $l_p \neq l_{p'}$ if $p \neq p'$, $l_s \notin \{k^2, \dots, K\}$ for each $s = 1, 2, \dots n-1$, and $l_n \in \{k^2 + 1, \dots, K\}$, or

(b) The products $P_{1,\, l_1} P_{l_1,\, l_2} \cdots P_{l_{n_1}-1,\, k^2} > 0$ and $P_{k^2,\, l_{n_1}+1} P_{l_{n_1}+1,\, l_{n_1}+2} \cdots P_{l_{n-1},\, l_n} > 0$ for some $n_1 \leq n$ and $n \leq q$, where $l_1 \neq 1$, $l_p \neq l_{p'}$ if $p \neq p'$, $l_s \notin \{k^2, \dots, K\}$ if $s \neq n_1$ or $s \neq n$, and $l_n \in \{k^2 + 1, \dots, K\}$.

If any of these products in (68) is positive, we have [from (50) and (39) and (41), and Lemma 5.2(i)]

$$\sum_{k=k^2+1}^{K} \dot{Z}_k(t) \geq P_{l_{n-1},\, l_n} \dot{D}_{l_{n-1}}(t) \geq P_{l_{n-1},\, l_n} P_{l_{n-2},\, l_{n-1}} \dot{D}_{l_{n-2}}(t),$$

and so on, yielding

$$\sum_{k=k^2+1}^{K} \dot{Z}_k(t) \geq P_{l_{n-1},\, l_n} P_{l_{n-2},\, l_{n-1}} \cdots P_{1,\, l_1} \dot{D}_1(t),$$

proving (64). If we are in case (b) above, we reason as follows. Using the fact that

$$P_{1,l_1}P_{l_1,l_2}\cdots P_{l_{n_1-1},k^2} > 0,$$

mimicking the arguments used in the previous case, and since

$$\dot{Z}_{k^2}(t) = -\dot{Z}_1(t) - \sum_{k=k^2+1}^{K} \dot{Z}_k(t) \le 0$$

[from (63)], we have

$$\dot{D}_{k^2}(t) \ge P_{1,l_1}P_{l_1,l_2}\cdots P_{l_{n_1-1},k^2}\dot{D}_1(t).$$

Using this and the fact that the product $P_{k^2,l_{n_1+1}}P_{l_{n_1+1},l_{n_1+2}}\cdots P_{l_{n-1},l_n} > 0$, arguing as following (68), we establish that

$$\sum_{k=k^2+1}^{K} \dot{Z}_k(t) \ge P_{1,l_1}P_{l_1,l_2}\cdots P_{l_{n_1-1},k^2}P_{k^2,l_{n_1+1}}P_{l_{n_1+1},l_{n_1+2}}\cdots P_{l_{n-1},l_n}\dot{D}_1(t).$$

Thus, (64) must hold, completing the proof of the theorem. □

We now use this result, and the result of Bramson to establish state space collapse under diffusion scaling. This will form the basis of the analysis in the next section.

THEOREM 5.4 (Bramson). *For each $k \in \{2, 3, \ldots, K-1\}$ (i.e., for every class other than the two lowest priority classes), and for every $\varepsilon > 0$ and every fixed $T > 0$, we have*

$$\lim_{r\to\infty} \mathbf{P}\left\{ \sup_{\{0\le t\le T\}} \left| \frac{Z_k^r(r^2t)}{r} \right| \ge \varepsilon \right\} = 0. \tag{69}$$

That is, $Z_k^r(r^2\cdot)/r$ converges to $\mathbf{0}$ uniformly on compact time sets, in probability.

PROOF. This proof outlines how the results of [3] may be adapted to establish (69). The reader is well advised to keep a copy of [3] handy while reading the rest of the argument. Bramson derives a version of state space collapse that he terms "multiplicative state space collapse." This would be equivalent to (69) if we have, in the notation of Bramson (not that of this paper), $\|\widehat{W}^r(\cdot)\|_T$ is stochastically bounded for every T. This would eliminate the denominator of equation (3.35) of Bramson, and hence we would obtain (69). But, since there are never more than r customers at a server when the population level is r, we have, in the notation of Bramson, $\|\widehat{W}^r(\cdot)\|_T$ is indeed stochastically bounded for every T. Thus, if the obvious analog of Theorem 4 of Bramson were to hold for closed networks, equation (3.35) of Bramson will hold with the Δ_z in (70) defined below, and (69) above would follow.

Thus, the only unresolved issue is whether an analog of Theorem 4 of Bramson holds for closed networks. We will not reprove Theorem 4 of Bramson, but we will provide sufficient detail to argue that the "proof" of Theorem 4 in Bramson will go through essentially unmodified for closed networks. The approach we will take to establishing this is to see if making the assumptions that (in Bramson's notation) $E(\cdot) \equiv 0$, and $\phi^k(i) \neq 0$ for any k or i (on page 96 of Bramson), and the consequent stochasticity of P, will result in any of Bramson's results being inapplicable. If we find that the results are indeed applicable in this setting, we are done, because these assumptions describe exactly the situation in closed networks. We will systematically walk through all the relevant sections of Bramson's paper and verify that the results remain applicable when modified to their obvious analogs for closed networks.

We begin by noting that equations (3.3) and (3.4) of Bramson are satisfied for $v_k^r(\cdot)$ due to the assumptions of Section 2.3 in this paper. Theorem 5.3 essentially verifies that Assumption 3.1 of [3] is satisfied. Now, because of the way in which the population is initialized in Section 2.2, equation (3.34) of Bramson is satisfied, with

$$\Delta_z = \begin{bmatrix} 1/m_1 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 1/m_k \end{bmatrix}. \tag{70}$$

Equation 3.1 of Bramson holds trivially in our setting since these quantities do not depend on r. All results in Section 4 of Bramson go through unmodified because they are not formulated in a network setting at all. Now consider the results in Section 5 of Bramson. Consider equations (5.10)–(5.15) and (5.18) of Bramson. They continue to hold if (in Bramson's notation) $E^{r,m}(\cdot) \equiv 0$ and $\phi^k(i) \neq 0$ for any k or i. Since Lemma 5.1 of Bramson holds for $v^{r,T,\max}$, and since Lemma 5.2 of Bramson holds, Proposition 5.1 of Bramson continues to hold, with Equation (5.19) being redundant now. So does Proposition 5.2 of Bramson.

We need to show that Proposition 8.1 of Bramson continues to hold with one minor modification. Note that Proposition 8.1 of Bramson is the generalization of Theorem 5.1 of this paper alluded to in the discussion following Theorem 5.1. In order that our fluid limit analysis of Theorem 5.3 constitute a verification of Assumption 3.1 of Bramson, we need to show that Proposition 8.1 of Bramson holds with equations (39)–(46) of this paper. From the proof of Proposition 8.1 on page 145 of [3], since his Propositions 5.1 and 5.2 continue to hold, we obtain the analog of his Proposition 8.1 for closed networks with all of the equations (39)–(46), *except* (44) satisfied. But (44) follows trivially for every cluster point of Proposition 8.1 of Bramson because $\sum_{k=1}^{K} Z_k^r(t) \equiv r$ a.s.

Finally, we turn our attention to the proof of Theorem 4 of Bramson. In order to establish the analog of Theorem 4 of Bramson, we will need to establish that the steps in the proof of Theorem 1 of Bramson on pages 135–141 of [3] can be adapted for (a) static priorities and (b) for closed networks. Bramson provides the adaption for (a) in Section 8 of [3]. Next, we note that Proposition 6.1 of Bramson uses neither the FIFO property nor the open network property, and hence continues to hold. A careful check of the proofs of the results on pages 135–141 of [3] reveals that the assumptions of equations (3.2) and (3.3) of Bramson are used only to facilitate the application of his Proposition 6.3. We do not need his Proposition 6.3, since we can replace it by our analog of his Assumption 3.1, namely Theorem 5.3. Thus, we are justified in mimicking the rest of the proof of Theorem 1 of Bramson in order to obtain the analog of Theorem 4 of Bramson and thus obtain (69). □

6. Diffusion limits of the workload imbalance process under the HW policy. We will now use the main output of the fluid limit analysis of the previous section, Theorem 5.4 to analyze diffusion-scaled processes. It is under this scaling that Harrison and Wein carried out their original, formal analysis of limiting processes. The goal of this section is to prove Theorem 6.1 that establishes that the solution to the limiting control problem obtained by them (cf. [14], equations (40)–(44) and Proposition 4) is indeed achieved by every weak limit of the scaled workload imbalance process. This result validates both their choice of the limiting control problem, as well as their interpretation of the solution to the limiting control problem.

In this section, we will repeatedly use the notation "$\Rightarrow$". We define this as follows. Consider a sequence of stochastic processes $\{X_n\}$ taking values in $(\mathbf{D}_{\mathbf{R}^d}[0,\infty), \mathscr{M}_d)$ (cf. Section 2.6). We write $X_n \Rightarrow X$ if the probability measures induced by X_n on $(\mathbf{D}_{\mathbf{R}^d}[0,\infty), \mathscr{M}_d)$ converge to the probability measure induced by X. Also, since we will consider only the HW policy in this section, we will drop the subscript u.

We begin the analysis by defining the following diffusion-scaled processes:

$$\widehat{Z}^r(t) = \frac{1}{r} Z^r(r^2 t), \tag{71}$$

$$\widehat{\mathscr{W}}^r(t) = \frac{1}{r}\mathscr{W}^r(r^2 t) = \widehat{M}\,\widehat{Z}^r(t), \tag{72}$$

$$\widehat{S}^r(t) = \frac{1}{r}(S^r(r^2 t) - M^{-1} r^2 t e), \tag{73}$$

$$\widehat{T}^r(t) = \frac{1}{r}(T^r(r^2 t) - r^2 t M \lambda^*), \tag{74}$$

$$\widehat{U}^r(t) = \frac{1}{r} U^r(r^2 t), \tag{75}$$

$$\widehat{\Phi}^{r,k}(x) = \frac{1}{r}(\Phi^{r,k}([x r^2]) - \widetilde{P}^k r^2 t). \tag{76}$$

The main result of this section is the following. Define

$$\widehat{X}^r(t) := \widehat{M}_1 + \widehat{M}\left(\sum_{k=1}^{K} \widehat{\Phi}^{r,k}(\overline{\overline{D}}{}^r_k(t))\right) - CGM\widehat{S}^r\left(\overline{\overline{T}}{}^r(t)\right), \tag{77}$$

where $\overline{\overline{D}}{}^r(t)$ and $\overline{\overline{T}}{}^r(t)$ are two quantities that are scaled quite differently from either the fluid-scaled quantities of (24)–(30) or the diffusion-scaled quantities in (71)–(76) above, and are given by

$$\overline{\overline{D}}^r(t) := \frac{D^r(r^2t)}{r^2} \quad \text{and} \quad \overline{\overline{T}}^r(t) := \frac{T^r(r^2t)}{r^2}. \tag{78}$$

THEOREM 6.1. *Under the HW policy, for every sequence of populations $r \to \infty$,*

$$(\widehat{\mathscr{W}}^r, \widehat{X}^r, \widehat{U}_1^r, \widehat{U}_2^r) \Rightarrow (\widehat{\mathscr{W}}^*, B^*, U_1^*, U_2^*),$$

where the weak limit $(\widehat{\mathscr{W}}^*, B^*, U_1^*, U_2^*)$ is unique in distribution and satisfies

$$\widehat{\mathscr{W}}^*(\cdot) = B^*(\cdot) + U_1^*(\cdot) - U_2^*(\cdot), \tag{79}$$

with $B^*(\cdot)$ being a one-dimensional zero-drift Brownian motion starting at $\widehat{M}_1$ with covariance Γ given by

$$\Gamma = \widehat{M}\left[\sum_{k=1}^{K} \lambda_k^* \Upsilon^k\right]\widehat{M}' + GC[\operatorname{diag}(\lambda_1^* b_1, \dots, \lambda_K^* b_K)]C'G' \tag{80}$$

and $U_1^*(\cdot), U_2^*(\cdot)$ satisfy

$$U_1^*(t) = \sup_{0 \le s \le t} [\widehat{M}_K - B^*(s) + U_2^*(s)]^+, \tag{81}$$

$$U_2^*(t) = \sup_{0 \le s \le t} [B^*(s) + U_1^*(s) - \widehat{M}_1]^+. \tag{82}$$

We will prove this theorem through the sequence of lemmas that follow. The first one is merely a convenient algebraic relationship between the diffusion-scaled quantities defined in (71)–(76).

LEMMA 6.1. *The scaled workload imbalance process $\widehat{\mathscr{W}}^r(t)$ can be written as*

$$\widehat{\mathscr{W}}^r(t) = \widehat{X}^r(t) + \widehat{U}_1^r(t) - \widehat{U}_2^r(t), \tag{83}$$

where $\widehat{X}^r(t)$ is given by (77) *above.*

PROOF. The definitions (71)–(76), the performance process dynamics (8)–(13), the property of λ^* described in (6) and simple algebraic manipulations yield

$$\widehat{Z}^r(t) = \widehat{Z}^r(0) + \sum_{k=1}^{K} \widehat{\Phi}^{r,k}(\overline{\overline{D}}_k^r(t)) - (I - \widetilde{P})\widehat{S}^r(\overline{\overline{T}}^r(t)) - (I - \widetilde{P})M^{-1}\widehat{T}^r(t).$$

Using (23) and $\widehat{M}\widehat{Z}^r(0) = \widehat{M}_1$, we obtain the result. □

The reader should note that in proving Lemma 6.1, we did not use any special property of the HW policy. Hence the lemma is applicable to all admissible policies.

LEMMA 6.2. *As $r \to \infty$,*

$$\overline{\overline{T}}^r(\cdot) \Rightarrow M\lambda^*(\cdot) \quad \textit{and} \quad \overline{\overline{D}}^r(\cdot) \Rightarrow \lambda^*(\cdot), \tag{84}$$

where $\lambda^*(t) = \lambda^* t$ for all $t \geq 0$.

PROOF. The relative compactness of $\{\overline{\overline{T}}^r(\cdot)\}$ follows from the fact that this family is uniformly Lipschitz. Let $\overline{\overline{T}}^{r_n}(\cdot) \Rightarrow \overline{\overline{T}}(\cdot)$ along a subsequence $\{r_n\}$. The renewal theorem shows that $\overline{\overline{D}}^{r_n}(\cdot) \Rightarrow M^{-1}\overline{\overline{T}}(\cdot)$. From (10), one can argue that

$$(I - \widetilde{P})M^{-1}\overline{\overline{T}}(t) = 0,$$

and hence from (6), $\overline{\overline{T}}(t) = \alpha(t)M\lambda^*$, where $0 \leq \alpha(t) \leq 1$. But Corollary 5.1 and Assumption 2.1 imply that $C\overline{\overline{T}}(t) = te$ for all $t \geq 0$. From (6) and (7), this translates to $\alpha(t) = 1$ for all $t \geq 0$, completing the proof since the result does not depend on the choice of the subsequence $\{r_n\}$. □

LEMMA 6.3.

$$\widehat{X}^r(\cdot) \Rightarrow B^*(\cdot), \tag{85}$$

where $B^*(\cdot)$ is a one-dimensional zero-drift Brownian motion starting at $\widehat{M}_1$, with covariance Γ given by (80).

PROOF. This follows from the functional central limit theorem for renewal processes (cf. [15]); Lemma 6.2 and the random time change theorem (cf. [2], (17.9)); the continuous mapping theorem (cf. [2], Theorem 5.1) and (23). The details can be filled by mimicking the proof of Theorem 7.1 of Williams [23]. □

Define

$$\widetilde{\mathscr{W}}^r(t) := \widehat{M}_1 \sum_{k=1}^{k_1} \widehat{Z}_k^r(t) + \widehat{M}_K \sum_{k=k_1+1}^{K} \widehat{Z}_k^r(t).$$

Then we have

$$\widetilde{\mathscr{W}}^r(t) = \widehat{\mathscr{W}}^r(t) - \varepsilon^r(t),$$

where

$$\varepsilon^r(t) = \sum_{k=2}^{k_1}(\widehat{M}_k - \widehat{M}_1)\widehat{Z}_k^r(t) + \sum_{k=k_1+1}^{K-1}(\widehat{M}_k - \widehat{M}_K)\widehat{Z}_k^r(t).$$

By Theorem 5.4, we know that for every $t \geq 0$,

$$\sup_{0 \leq s \leq t} |\varepsilon^r(s)| \to 0 \quad \text{in probability.} \tag{86}$$

From (13) we can write $\widetilde{\mathscr{W}}^r(t)$ as

$$\widetilde{\mathscr{W}}^r(\cdot) = (\widehat{X}^r(\cdot) - \varepsilon^r(\cdot)) + \widehat{U}_1^r(\cdot) - \widehat{U}_2^r(\cdot), \tag{87}$$

where the processes $\widehat{U}_1^r(\cdot)$ and $\widehat{U}_2^r(\cdot)$ satisfy

$$\int_0^\infty \mathbf{1}_{(\widehat{M}_K,\, \widehat{M}_1]}(\widetilde{\mathscr{W}}^r(t))\, d\widehat{U}_1^r(t) = 0, \tag{88}$$

$$\int_0^\infty \mathbf{1}_{[\widehat{M}_K,\, \widehat{M}_1)}(\widetilde{\mathscr{W}}^r(t))\, d\widehat{U}_2^r(t) = 0. \tag{89}$$

For a function $f(\cdot) \in \mathbf{D}_{\mathbf{R}}[0, \infty)$, and any fixed $T > 0$, define

$$\mathbf{Osc}(f, T) = \sup_{0 \le s,\, t,\, \le T} |f(s) - f(t)|.$$

The following result is a trivial application of a more general oscillation inequality developed by [5].

LEMMA 6.4. *There exists a constant C such that for any fixed $T > 0$,*

$$\mathbf{Osc}(\widetilde{\mathscr{W}}^r, T) \le C\, \mathbf{Osc}(\widehat{X}^r - \varepsilon^r, T), \tag{90}$$

$$\mathbf{Osc}(\widehat{U}_1^r, T) \le C\, \mathbf{Osc}(\widehat{X}^r - \varepsilon^r, T), \tag{91}$$

$$\mathbf{Osc}(\widehat{U}_2^r, T) \le C\, \mathbf{Osc}(\widehat{X}^r - \varepsilon^r, T). \tag{92}$$

PROOF. This follows from Theorem 4.2 of [5]. The conditions (S.a) and (S.b) there are trivially verified for our regulator problem defined by (87)–(89), and we use the continuity of $\widehat{U}_1^r$ and $\widehat{U}_2^r$ to complete the proof. □

PROOF OF THEOREM 6.1. We are now ready to prove Theorem 6.1. We mimic the proof of Theorem 4.1 of [24]. The tightness of $\widehat{X}^r(\cdot) - \varepsilon(\cdot)$ established in Lemma 6.3 and (86), and the oscillation inequality Lemma 6.4, establishes the tightness of $(\widetilde{\mathscr{W}}^r, \widehat{X}^r, \widehat{U}_1^r, \widehat{U}_2^r)$ and hence the tightness of $(\widehat{\mathscr{W}}^r, \widehat{X}^r, \widehat{U}_1^r, \widehat{U}_2^r)$. For details of this argument, see the proof of Theorem 4.1 of [24]. Now, we need to show that for all possible weak limits $(\widehat{\mathscr{W}}^*, B^*, U_1^*, U_2^*)$, we have

$$\widehat{\mathscr{W}}^* = B^* + U_1^* - U_2^*, \tag{93}$$

along with

$$\int_0^\infty \mathbf{1}_{(\widehat{M}_K,\, \widehat{M}_1]}(\widehat{\mathscr{W}}^*(t))\, dU_1^*(t) = 0 \tag{94}$$

and

$$\int_0^\infty \mathbf{1}_{[\widehat{M}_K,\, \widehat{M}_1)}(\widehat{\mathscr{W}}^*(t))\, dU_2^*(t) = 0. \tag{95}$$

This would complete the proof of Theorem 6.1 because of the uniqueness of the two-sided regulator defined in (79)–(82); compare [10], Proposition 2.4.6. We will use the Skorohod representation theorem (cf. [9], Theorem 3.1.8) and the almost sure continuity of the limit (established by the continuity of B^* and the oscillation inequality, Lemma 6.4) to replace the sequence if processes considered with another sequence that is term-by-term equivalent and converges uniformly on compact time sets almost surely. Then, (93) follows from (87) and (86). To show (94), it suffices to show that

$$\int_0^T f_m(\widehat{\mathscr{W}}^*(t))\, dU_1^*(t) = 0 \quad \text{a.s.,} \tag{96}$$

for each $T \ge 0$ and $m = 1, 2, \dots,$ where for each $m, f_m \colon \mathbf{R} \to [0,1]$ is a continuous function with $f_m(x) = 0$ for $x \le \widehat{M}_k + 1/m$ and $f_m(x) = 1$ for $x \ge \widehat{M}_K = 2/m$. Let $\{r_n\}$ be a subsequence along which a limit is achieved. From (88), we have

$$\int_0^T f_m(\widetilde{\mathscr{W}}^{r_n}(t))\, d\widehat{U}_1^{r_n}(t) = 0 \quad \text{a.s.,}$$

for each $T \ge 0$ and $m = 1, 2, \dots,$ for every r_k. Now since f_m is continuous and bounded, and $\widetilde{\mathscr{W}}^{r_k} \to \widehat{\mathscr{W}}^*$ and $U_1^{r_n} \to U_1^*$ uniformly on compact time sets almost surely, and since $U_1^{r_n}$ is increasing for each n, from Lemma 2.4 of [6], we have

$$\int_0^T f_m(\widetilde{\mathscr{W}}^{r_n}(t))\, d\widehat{U}_1^{r_n}(t) \to \int_0^T f_m(\widehat{\mathscr{W}}^*(t))\, dU_1^*(t)$$

uniformly on compact time sets almost surely. Thus, we obtain (96) and hence (94). Similarly we can obtain (95). The uniqueness in distribution of the two-sided regulator applied to Brownian motion allows us the ignore the choice of the subsequence, thus completing the proof of the theorem. □

We note in passing that we could have used the continuity of the map from $\widehat{X}^r(\cdot) - \varepsilon^r(\cdot)$ to $\widetilde{\mathscr{W}}^r(\cdot)$ in the uniform topology to prove this theorem, but the oscillation inequality will be used in the sequel for a different purpose, and hence we chose to use this method of proof.

7. Asymptotic optimality. In this section we establish that the Harrison and Wein policy is indeed asymptotically optimal as defined in Definition 3.2. We will do this in several steps. First, we will establish that $\{\widehat{U}_1^r(T)\}_{r=1}^{\infty}$ is uniformly integrable. Then we will establish asymptotic optimality by showing that the first moment of the diffusion-scaled idleness process under the HW policy converges to that of a lower bound on the scaled idleness process under every other admissible policy.

THEOREM 7.1. *For each fixed $T > 0$, the family $\{\widehat{U}_1^r(T)\}_{r=1}^{\infty}$ is uniformly integrable. Thus, by Theorem* 6.1, *for each fixed $T > 0$, we have*

$$\lim_{r\to\infty} \frac{\mathbf{E}[U_1^r(r^2T)]}{rT} = \frac{\mathbf{E}[U_1^*(T)]}{T}, \tag{97}$$

where $U_1^*(\cdot)$ is given by (79)–(82).

PROOF. From the oscillation inequality, Lemma 6.4, we have

$$\widehat{U}_1^r(T) \le C\,\mathbf{Osc}(\widehat{X}^r - \varepsilon^r, T) \le 2C \sup_{0\le t\le T} |\widehat{X}^r(t) - \varepsilon^r(t)|.$$

Now $|\varepsilon^r(t)| \le 2\sum_k |\widehat{M}_k|$ for all t. Therefore, in order to show $\{\widehat{U}_1^r(T)\}_{r=1}^{\infty}$is uniformly integrable, it is enough to show that the family $\{\sup_{0\le t\le T} |\widehat{X}^r(t)|\}_{r=1}^{\infty}$ is uniformly integrable. Recall that, from (77), we have

$$\widehat{X}^r(t) = \widehat{M}_1 + \widehat{M}\biggl(\sum_{k=1}^{K} \widehat{\Phi}^{r,k}([\overline{\overline{D}}{}_k^r(t)])\biggr) - GCM\widehat{S}^r(\overline{\overline{T}}{}^r(t)).$$

Let us begin by showing that

$$\Bigl\{\sup_{0\le t\le T} |M\widehat{S}^r(\overline{\overline{T}}{}^r(t))|\Bigr\}_{r=1}^{\infty} \text{ is uniformly integrable.}$$

Note that, for each r, $\sup_{0\le t\le T} |\widehat{S}^r(\overline{\overline{T}}{}^r(t))| \le \sup_{0\le t\le T} |\widehat{S}^r(t)|$, since $0 \le \overline{\overline{T}}{}^r(t) \le t$. So we only need to show that

$$\Bigl\{\sup_{0\le t\le T} |M\widehat{S}^r(t)|\Bigr\}_{r=1}^{\infty} \text{ is uniformly integrable.} \tag{98}$$

To establish (98), we mimic the proof of Lemma 8.4 of [5]. Recall that $V_k(n) = \sum_{i=1}^n v_k(i)$ denotes the partial sum of the service times. So we can write $m_k\widehat{S}_k^r(t)$ as

$$\begin{aligned} m_k\widehat{S}_k^r(t) &= \biggl\{\frac{m_k[S_k^r(r^2t)+1] - V_k(S_k^r(r^2t)+1)}{r}\biggr\} \\ &\quad + \biggl\{\frac{V_k(S_k^r(r^2t)+1) - r^2t}{r}\biggr\} - \frac{m_k}{r} \\ &=: \{\mathbf{M}_k^r(t)\} + \{\eta_k^r(t)\} - \frac{m_k}{r}. \end{aligned}$$

Note that $S_k^r(r^2t)+1$ is a stopping time with respect to the filtration $\{\mathscr{G}_n\}$ where

$$\mathscr{G}_n = \sigma\{v_k(1), v_k(2), \dots, v_k(n)\}.$$

We conclude that $\mathbf{M}_k^r(t)$ is a square integrable martingale with

$$\mathbf{E}[\mathbf{M}_k^r(t)^2] = b_k m_k^2 \frac{\mathbf{E}[S^r(r^2t)+1]}{r^2}.$$

The right-hand side above is bounded for all r by Lorden's inequality (cf. [19], (III.4.1)). Now we use Doob's maximal inequality ([9], Corollary 2.2.17) to conclude that $\mathbf{E}[\sup_{0\le t\le T} |\mathbf{M}_k^r(t)|^2]$ is bounded and hence $\{\sup_{0\le t\le T} |\mathbf{M}_k^r(t)|\}_{r=1}^{\infty}$

is uniformly integrable. Since $\eta^r_k(t)$ is the overshoot of the renewal process, we have

$$\begin{aligned}\eta^r_k(t) &\le \max_{1\le i\le S^r_k(r^2t)+1} \frac{v_k(i)}{r}\\ &\le \sup_{0\le t\le T} |\mathbf{M}^r_k(t) - \mathbf{M}^r_k(t-)| + m_k/r\\ &\le 2 \sup_{0\le t\le T} |\mathbf{M}^r_k(t)| + m_k/r,\end{aligned}$$

from which (98) follows. Now, we turn our attention to establishing, for each $k = 1, 2, \dots, K$,

$$\left\{\sup_{0\le t\le T}\left|\widehat{\Phi}^{r,k}(\overline{\overline{D}}{}^r_k(t))\right|\right\}_{r=1}^{\infty} \text{ is uniformly integrable.} \tag{99}$$

Since $T^r(r^2t) \le r^2te$ for all $t \ge 0$, we have

$$\begin{aligned}\sup_{0\le t\le T}\left|\widehat{\Phi}^{r,k}_j(\overline{\overline{D}}{}^r_k(t))\right| &\le \sup_{0\le t\le T}\left|\frac{\Phi^{r,k}_j(S^r_k(r^2t)) - P_{kj}S^r_k(r^2t)}{r}\right|\\ &= \max_{n\in\{1,2,\dots,S^r_k(r^2T)\}}\left|\frac{\Phi^{r,k}_j(n) - P_{kj}n}{r}\right|.\end{aligned}$$

Now $\mathbf{E}[\max_{n\in\{1,2,\dots,N\}} |\Phi^{r,k}_j(n) - P_{kj}n|]^2 \le KN$ for some constant K that does not depend on N, for each N (using Doob's inequality ([9], Corollary 2.2.17), for example). Using this result, conditioning on $S^r(r^2T)$, and using the independence of $S^r(r^2T)$ and $\Phi^{r,k}$, we obtain

$$\mathbf{E}\left[\sup_{0\le t\le T} |\widehat{\Phi}^{r,k}([\overline{\overline{D}}{}^r_k(t)])|\right]^2 \le K' \frac{\mathbf{E}[S^r_k(r^2T)]}{r^2},$$

for some constant K'. Using Lorden's inequality ([19], (III.4.1)) once more, we obtain (99) and complete the proof of the theorem. □

COROLLARY 7.1. *Under the HW policy, we have*

$$\lim_{T\to\infty}\lim_{r\to\infty}\frac{\mathbf{E}[U^r_1(r^2T)]}{rT} = \frac{\Gamma}{2(\widehat{M}_1 - \widehat{M}_K)},$$

where Γ *is given by* (80).

PROOF. Theorems 7.1 and 6.1 yield $(\mathbf{E}[U^r_1(r^2T)]/r) \to \mathbf{E}[U^*_1(T)]$ as $r \to \infty$ for each fixed $T > 0$. But, from [10] (5.4.13) and (5.5), we know that

$$\lim_{T\to\infty}\frac{\mathbf{E}[U^*_1(T)]}{T} = \frac{\Gamma}{2(\widehat{M}_1 - \widehat{M}_K)}. \tag{100}$$

Thus we obtain the result. □

A note on this result is in order. If we had the limits above in reverse order, that is,

$$\lim_{r\to\infty}\lim_{T\to\infty}\frac{\mathbf{E}[U_1^r(r^2T)]}{rT}=\frac{\Gamma}{2(\widehat{M}_1-\widehat{M}_K)},$$

we would have established asymptotic optimality in the stronger sense of (15) in the case of *exponentially distributed* service times. This is because the right-hand side above was shown to be a lower bound on the scaled idleness in [16]. However, we will not pursue the analysis of whether or not the limits can be interchanged. We are now ready to prove our main result on asymptotic optimality of the HW policy.

THEOREM 7.2. *Under any policy u, for any $T>0$ fixed, we have*

$$\lim_{r\to\infty}\mathbf{E}[\widehat{U}^r_{HW,1}(T)]\le\liminf_{r\to\infty}\mathbf{E}[\widehat{U}^r_{u,1}(T)]. \tag{101}$$

Thus, the Harrison–Wein policy is asymptotically optimal as defined by (16).

PROOF. The proof rests on constructing two processes R^r_u and L^r_u that pathwise lower bound $\widehat{U}^r_{u,1}$ and $\widehat{U}^r_{u,2}$, respectively, under a given policy u at each r. We then show that R^r_u and L^r_u, and $\widehat{U}^r_{HW,1}$ and $\widehat{U}^r_{HW,2}$ have the same weak limit, for every policy u that can possibly be considered a candidate for violating (101), thus establishing asymptotic optimality.

First, note that the proof of Lemma 6.1 does not depend on the choice of policy, and hence we have

$$\widehat{\mathscr{W}}^r_u(\cdot)=\widehat{X}^r_u(\cdot)+\widehat{U}^r_{u,1}(\cdot)-\widehat{U}^r_{u,2}(\cdot).$$

Now fix a sample path of the process $(\widetilde{\mathscr{W}}^r_u,\widehat{X}^r_u,\widehat{U}^r_{u,1},\widehat{U}^r_{u,2})$. Since $\widehat{\mathscr{W}}^r_u(t)\in[\widehat{M}_k,\widehat{M}_1]$ for all $t\ge 0$, we can conclude that

$$\widehat{U}^r_{u,1}(t)\ge\sup_{0\le s\le t}\big[\widehat{M}_K-\widehat{X}^r_u(s)+\widehat{U}^r_{u,2}(s)\big]^+ \tag{102}$$

and

$$\widehat{U}^r_{u,2}(t)\ge\sup_{0\le s\le t}\big[\widehat{X}^r_u(s)+\widehat{U}^r_{u,1}(s)-\widehat{M}_1\big]^+. \tag{103}$$

Now construct a sequence of processes, R^r_n and L^r_n, as follows. Let $R^r_0(t)=\widehat{U}^r_{u,1}(t)$ and $L^r_0(t)=\widehat{U}^r_{u,2}(t)$, and for $n=0,1,2,\dots,$ let

$$R^r_{n+1}(t)=\sup_{0\le s\le t}\big[\widehat{M}_K-\widehat{X}^r_u(s)+L^r_n(s)\big]^+$$

and

$$L^r_{n+1}(t)=\sup_{0\le s\le t}\big[\widehat{X}^r_u(s)+R^r_n(s)-\widehat{M}_1\big]^+.$$

For each $t\in[0,T]$, using (102) and (103) and induction on n we can establish that $R^r_{n+1}(t)\le R^r_n(t)$ and $L^r_{n+1}(t)\le L^r_n(t)$ for each $n=0,1,2,\dots.$

Since these sequences are nonnegative, they converge for each t. Let $R_u^r(t) = \lim_{n\to\infty} R_n^r(t)$, and $L_u^r(t) = \lim_{n\to\infty} L_n^r(t)$. L_u^r and R_u^r are nondecreasing and not necessarily continuous processes. They satisfy

$$R_u^r(t) = \sup_{0\le s\le t} \big[\widehat{M}_K - \widehat{X}_u^r(s) + L_u^r(s)\big]^+ \tag{104}$$

and

$$L_u^r(t) = \sup_{0\le s\le t} \big[\widehat{X}_u^r(s) + R_u^r(s) - \widehat{M}_1\big]^+ \quad \text{for each } t \in [0, T]. \tag{105}$$

It might appear that R_u^r and L_u^r above may depend on $\widehat{U}_{u,1}^r$ and $\widehat{U}_{u,2}^r$, that is, the starting point of the recursions. This is not true. Given $\widehat{X}_u^r$, R_u^r and L_u^r are indeed unique. We can establish this using Lemma 4.3 of [5] and mimicking the uniqueness proof of Proposition 2.4.6 of [10]. One needs to extend the proof in Harrison from the space of continuous functions to $\mathbf{D}_{\mathbf{R}}[0,\infty)$. Lemma 4.5 of [5] provides the argument for the one-sided regulator and extending it to the two-sided regulator considered is straightforward. We will not go into the details here, as the uniqueness result is tangential to the rest of the arguments below. The reader is referred to [4] for details of generalizing such results from the space of continuous functions to the space of functions with finitely many jumps.

Note that this does not mean that R_u^r and L_u^r do not depend on the policy employed, u. Of course, $\widehat{X}_u^r$ does depend on u for each r. In the sequel, we will establish that the weak limit of $\widehat{X}_u^r$ does not depend on the policy employed, for all interesting policies u.

For convenience, let us define

$$\widehat{\mathscr{W}}_{*,u}^r(t) = \widehat{X}_u^r(t) + R_u^r(t) - L_u^r(t).$$

By construction, for every policy u, every r, and for almost every realization (i.e., for almost all $\omega \in \Omega$), we have

$$\widehat{U}_{u,1}^r(t) \ge R_u^r(t) \quad \text{for each } t \in [0, T] \tag{106}$$

and

$$\widehat{U}_{u,2}^r(t) \ge L_u^r(t) \quad \text{for each } t \in [0, T]. \tag{107}$$

In establishing (101) we need to consider only those policies u for which $\lim_{r\to\infty}(\mathbf{E}[U_u^r(r^2T)]/r^2) = 0$, since the result is trivially true for all other policies. Thus, we only consider policies for which $\overline{\overline{T}}^r(t) \to M\lambda^* t$ almost surely, uniformly on compact time sets. From Lemma 6.3, which did not use any special property of the HW policy, we can conclude that, for every such policy u, $\widehat{X}_u^r \Rightarrow B^*$, where B^* is a zero-drift Brownian motion starting at $\widehat{M}_1$, with covariance Γ given by (80). Let $C[0, T]$ denote the space of continuous functions on $[0, T]$ endowed with sup norm. Let f, g, h: $C[0, T] \to C[0, T]$ be defined by

$$f(x)(t) = x(t) + g(x)(t) - h(x)(t),$$
$$g(x)(t) = \sup_{0\le s\le t}\big[\widehat{M}_K - x(s) + h(x)(s)\big]^+$$

and

$$h(x)(t) = \sup_{0\le s\le t}\big[x(s) + g(x)(s) - \widehat{M}_1\big]^+.$$

The mappings of f, g, h are continuous functions on $C[0, T]$. Hence, using the continuous mapping theorem (cf. [2], Theorem 5.1), $(\mathscr{W}^r_{*,u}, \widehat{X}^r_u, R^r_u, L^r_u) \Rightarrow (\widehat{\mathscr{W}}^*, B^*(\cdot), U_1^*(\cdot), U_2^*(\cdot))$, where $(\widehat{\mathscr{W}}^*, B^*(\cdot), U_1^*(\cdot), U_2^*(\cdot))$ satisfies (79)–(82). Also, $(\mathscr{W}^r_{*,u}, \widehat{X}^r_u, R^r_u, L^r_u)$ satisfy the oscillation inequality, Lemma 6.4. Thus, mimiking Theorem 7.1, we have

$$\lim_{r\to\infty} \mathbf{E}[R^r_u(T)] = \mathbf{E}[U_1^*(T)].$$

Equation (101) follows from (97) and (106). □

A note on the processes R^r_u and L^r_u constructed in the proof of the previous theorem is in order. The reader should not mistake these processes for idleness processes under some "optimal" policy. They are not. For example, they need not even be continuous. One should view them as ideals that are not realizable by any policy at any population level r, but are arbitrarily well approximated by the scaled idleness processes under the HW policy as the population increases without bound.

8. Extensions. The results of this paper can be extended in several ways. The reader will find the treatment here somewhat loose and speculative: this section is meant to be an indication of possible future work. The first is the relaxation of Assumption 2.1 of a balanced network. One can consider a sequence of systems indexed by r. In the rth system, the mean service time matrix M_r and the routing matrix P_r depend on r and the population in this system is r. One obtains the corresponding sequences π_r, λ_r^* and ρ_r^* and one requires that

$$r(\rho^*_{r,1} - \rho^*_{r,2}) \text{ converges to some finite } \theta \text{ as } r \to \infty.$$

In this case, the workload imbalance indices are computed differently from (17) and (18), as

$$Q_r = (I - \widetilde{P}_r + (\pi_r)'e)^{-1}$$

and

$$\widehat{M}_r = \big[\rho^*_{r,2} \quad -\rho^*_{r,1}\big]\, CM_rQ_r.$$

The HW policy remains the same static priority rule based on the indices $\widehat{M}_r$, although now the indices need to be recomputed for each system. All the limiting arguments go through, with the Brownian motion B^* of Lemma 6.3 now having drift θ (rather than zero) and consequently $\mathbf{E}\big[U_1^*(T)\big]$ being different.

One could also try and strengthen the notion of asymptotic optimality to that involving long-run time averages of the idleness processes as in (15), rather than expected values of the scaled idleness processes evaluated at a fixed (albeit large) time. Although the interchange of limits as alluded to in the discussion following Corollary 7.1 can be carried out in some cases (cf. [18]), one needs to establish some structural properties of the underlying Markov process to do so.

Finally, one could investigate when the scaled workload imbalance process $\widehat{\mathscr{W}}_u^r$ under an admissible policy u fails to have a weak limit under diffusion scaling. Note that (87)–(89) do not depend on the policy employed. Therefore, if $\big(\widehat{X}_u^r, \varepsilon_u^r\big)$ converge weakly, $\widehat{\mathscr{W}}_u^r$ does have a weak limit. Of course, we can ignore policies for which $\widehat{X}_u^r$ does not converge to B^*, since these have poorer performance than any policy for which $\widehat{X}_u^r$ does converge to B^*. So the only way $\widehat{\mathscr{W}}_u^r$ could fail to have a weak limit even when $\widehat{X}_u^r$ does converge to B^* is if ε_u^r was badly behaved. In fact, since $\big|\varepsilon_u^r(t)\big| \le 2\sum_k\big|\widehat{M}_k\big|$, the only way $\widehat{\mathscr{W}}_u^r$ could fail to have a weak limit is if the modulus of continuity condition, equation (3.7.14) of [9] is violated by ε_u^r. Translating this condition to more easily verifiable conditions on the network primitives and the policy employed is yet another interesting possibility for future work.

Acknowledgments. I am indebted to Mike Harrison for suggesting this project, and to Ruth Williams, Maury Bramson and Jim Dai for providing early versions of their papers on which this work is based. I am also indebted to the three anonymous reviewers and the Associate Editor who found errors in, and provided extensive suggestions on, an earlier draft. They considerably improved the paper.

REFERENCES

[1] Berman, A. and Plemmons, R. (1994). *Nonnegative Matrices in the Mathematical Sciences*. SIAM, Philadelphia.

[2] Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

[3] Bramson, M. (1998). State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Sys.* **30** 89–148.

[4] Chen, H. and Mandelbaum, A. (1990). Leontief Systems, RBVs and RBMs. In *Applied Stochastic Analysis* (M. H. A. Davis and R. J. Elliott, ed.) 1–43. Gordon and Breach, Langhorne, PA.

[5] Dai, J. and Dai, W. (1998). A heavy traffic limit theorem for a class of open queueing networks with finite buffers. Preprint.

[6] Dai, J. and Williams, R. (1995). Existence and uniqueness of semimartingale reflecting Brownian motions in convex polyhedrons. *Theory Probab. Appl.* **40** 3–53.

[7] Dai, J. G. (1995). On Positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Ann. Appl. Probab.* **5** 49–77.

[8] DAI, J. G. and WEISS, G. (1996). Stability and instability of fluid models for reentrant lines. *Math. Oper. Res.* **21** 115–134.

[9] ETHIER, S. N. and KURTZ, T. G. (1986) *Markov Processes: Characterization and Convergence*. Wiley, New York.

[10] HARRISON, J. M. (1985). *Brownian Motion and Stochastic Flow Systems*. Wiley, New York.

[11] HARRISON, J. M. (1988). Brownian models of queueing networks with heterogeneous customer populations. In *Stochastic Differential Systems, Stochastic Control Theory and Applications* (W. Fleming and P. L. Lions, eds.) 147–186. Springer, New York.

[12] HARRISON, J. M. and VAN MIEGHEM, J. A. (1997). Dynamic control of Brownian networks: state space collapse and equivalent workload formulations. *Ann. Appl. Probab*. **7** 747–771.

[13] HARRISON, J. M. and WEIN, L. M. (1989). Scheduling network of queues: heavy traffic analysis of a simple open network. *Queueing Sys.* **5** 265–280.

[14] HARRISON, J. M. and WEIN, L. M. (1990). Scheduling networks of queues: heavy traffic analysis of a two-station closed network. *Oper. Res*. **38** 1052–1064.

[15] IGLEHART, D. L. and WHITT, W. (1970). Mutliple channel queues in heavy traffic I. *Ann. Appl. Probab*. **2** 150–177.

[16] JIN, H., OU, J., and KUMAR, P. R. (1997). The throughput of closed queueing networks: functional bounds, asymptotic loss, efficiency and the Harrison–Wein conjectures. *Math. Oper. Res*. **22** 886–920.

[17] KUMAR, S. and KUMAR, P. R. (1994). Performance bounds for queueing networks and scheduling policies. *IEEE Trans. Automat. Control*. **39** 1600–1611.

[18] KUMAR, S. and KUMAR, P. R. (1996). Fluid limits and the efficiency of scheduling policies for stochastic closed reentrant lines in heavy traffic. *Stochastic Networks: Stability and Rare Events. Lecture Notes in Statist*. **117** 41–64. Springer, Berlin.

[19] LINDVALL, T. (1992). *Lectures on the Coupling Method*. Wiley, New York.

[20] MARTINS, L. F., SHREVE, S. E. and SONER, H. M. (1996). Heavy traffic convergence of a controlled multiclass queueing network. *SIAM. J. Control Optim*. **34** 2133–2171.

[21] SOLBERG, J. J. (1981). Capacity planning using a stochastic workflow model. *AIIE Trans*. **13** 116–122.

[22] SPEARMAN, M. L., WOODRUFF, D. L. and HOPP, W. J. (1990). CONWIP: a pull alternative to Kanban. *Intl. J. Production Res.* **28** 879–894.

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

[24] WILLIAMS, R. (1998). An invariance principle for semimartingale reflecting Brownian motions in an orthant. *Queueing Sys.* **30** 5–25.

GRADUATE SCHOOL OF BUSINESS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-5015
E-MAIL: skumar@leland.stanford.edu