Probability in the Engineering and Informational Sciences, **19**, 2005, 141–189. Printed in the U.S.A.

OPTIMAL ROUTING IN OUTPUT-QUEUED FLEXIBLE SERVER SYSTEMS

ALEXANDER L. STOLYAR
Bell Labs
Lucent Technologies
Murray Hill, New Jersey 07974
stolyar@research.bell-labs.com

We consider a queuing system with multitype customers and nonhomogeneous flexible servers, in the heavy traffic asymptotic regime and under a *complete resource pooling* (CRP) condition. For the *input-queued* (IQ) version of such a system (with customers being queued at the system "entrance," one queue per each type), it was shown in the work of Mandelbaum and Stolyar that a simple parsimonious G$c\mu$ scheduling rule is optimal in that it asymptotically minimizes the system *customer workload* and some strictly convex queuing costs. In this article, we consider a different—*output-queued* (OQ)—version of the model, where each arriving customer must be assigned to one of the servers immediately upon arrival. (This constraint can be interpreted as immediate *routing* of each customer to one of the "output queues," one queue per each server.) Consequently, the space of controls allowed for an OQ system is a subset of that for the corresponding IQ system.

We introduce the *MinDrift* routing rule for OQ systems (which is as simple and parsimonious as G$c\mu$) and show that this rule, in conjunction with arbitrary work-conserving disciplines at the servers, has asymptotic optimality properties analogous to those G$c\mu$ rule has for IQ systems. A key element of the analysis is the notion of system *server workload*, which, in particular, majorizes customer workload. We show that (1) the MinDrift rule asymptotically minimizes server workload process among all OQ-system disciplines and (2) this minimal process matches the minimal possible customer workload process in the corresponding IQ system. As a corollary, MinDrift asymptotically minimizes customer workload among all disciplines in either the OQ or IQ system.

 0269-9648/05 $16.00

1. INTRODUCTION

1.1. The Problem

We consider a queuing system with multiple customer flows (types) $i = 1,\dots,I$ and nonhomogeneous flexible servers $j = 1,\dots,J$. This means that the mean service time μ_{ij}^{-1} of a type i customer by server j depends on both the customer type and the server. We study the "heavy traffic" asymptotic regime, when the system is close to be "critically loaded," and assume that a certain *complete resource pooling* (CRP) condition holds. Associated with the CRP condition is the notion of system *workload*, which in this article is called system *customer workload*.

The "input-queued" (IQ) version of the model (see [3,8,9,12,20]) is such that arriving customers are placed in "input" queues, one queue per each type i, where they await for service without being preassigned to any particular server until they are actually "taken for service" by one of them. It is shown in [12] that an IQ system can be asymptotically optimally controlled in heavy traffic (and under the CRP condition) by a very simple and parsimonious generalized $c\mu$ (G$c\mu$) scheduling rule, which, in particular, minimizes customer workload among virtually all service disciplines.

In this article, we consider a different—"output-queued" (OQ)—version of the model, where each arriving customer must be assigned (or *routed*) to one of the servers immediately upon arrival. (This can be viewed as immediate routing of each arriving customer to one of the "output" queues, one per each server.) Such models arise in various applications, including wireless networks, manufacturing systems, and call centers. A wireless application example is a system in which data packets ("customers") need to be delivered to multiple mobile users/destinations (which determine "customer types") via a set of transmitters ("servers"); transmission ("service") rates depend on the (different) channel qualities between different transmitters and users.

Due to the above *immediate routing* (IR) constraint, the space of controls allowed for an OQ system is a (strict) subset of that for the corresponding IQ system. (G$c\mu$ is not a valid discipline for OQ systems.) A natural question is: "Is it possible to control an OQ system in heavy traffic as efficiently as the corresponding IQ system could be controlled? For example, are there OQ-system controls that are as parsimonious as G$c\mu$ but still able to minimize the system customer workload?" The results of this article demonstrate that the answer to the latter question is "yes"—we introduce the *MinDrift* routing rule, which, in particular, minimizes the system customer workload in heavy traffic. We will describe our results shortly, after a brief literature review.

1.2. Previous Work

The IQ version of our system (with nonhomogeneous flexible servers) and the definition of the heavy traffic asymptotic regime for it (in terms of a certain linear program) were introduced in [8] (for a special two-server system) and in [9,20]

(for the general setting). For the system in heavy traffic, these articles also define the CRP condition and associated with it the notion of *customer workload*, defined as $X(t) = \sum_i \nu_i^* Q_i(t)$, $t \geq 0$, where $Q_i(t)$ is the type i queue length at time t and $\nu_i^* > 0$ is the fixed constant called *workload contribution* of a type i customer. References [8,9] propose *discrete-review* scheduling policies, which, in the heavy traffic and under CRP condition, asymptotically minimize customer workload and linear holding costs. In [3,20], *continuous-review threshold* policies are proposed that are asymptotically optimal, in the same sense and also under CRP condition. (The asymptotic optimality proofs are given for a special two-server system.) The common feature of discrete-review and continuous-review threshold policies is that they require a priori knowledge of the flows' mean arrival rates λ_i.

In [12], it is proved that a very simple G$c\mu$ scheduling rule (which, in particular, does *not* require the knowledge of arrival rates λ_i) asymptotically minimizes customer workload and strictly convex holding costs in a general IQ system under the CRP condition. Moreover, in the limit, the (appropriately rescaled) queue-length vector process $(Q_1(t), \ldots, Q_I(t))$ exhibits *state space collapse* (SSC): it "lives" on a one-dimensional manifold. (The results of [12] are closely related to earlier results in [14] for a discrete-time *generalized switch* model. Also, they generalize the earlier G$c\mu$ optimality results for a single-server system [17].) Following [14], [12] provides equivalent (geometric) characterization of the CRP condition, as follows. Let $\mathcal{M}$ be the system *service rate region*, which is roughly the set of all vectors representing feasible long-term average service rates the system is capable of *jointly* providing to different types. Then the vector $\nu^* = (\nu_1^*, \ldots, \nu_I^*)$ of workload contributions is the *unique* (up to scaling) outer normal vector to the boundary of $\mathcal{M}$ at the point $\lambda = (\lambda_1, \ldots, \lambda_I)$.

Most of the previous work on OQ systems is concentrated on load balancing schemes for systems with homogeneous servers. Much less work has been done on a heavy traffic regime in systems with nonhomogeneous flexible servers. Probably the first was [10], where a two-server system is considered, resource pooling in heavy traffic is discussed, and threshold-based policies are proposed. (See also [11] for an earlier discussion of resource pooling in systems with routing.) In a recent article [16] a two-server system (different from that in [10]) with exponential service times is considered, and the asymptotic optimality of a threshold routing policy is proved, under linear holding costs. We refer the reader to [16] for a more extensive review of the previous work on OQ systems.

1.3. Our Results

In this article, we consider a general OQ system in a heavy traffic regime and under the CRP condition. First, we give further equivalent characterization of the CRP condition, which is natural and convenient for the analysis of OQ systems; namely, we consider the *server utilization region* $\mathcal{K}$, which is the set of potential server utilization vectors that can be imposed by the input flows with mean rates λ_i or greater. Then the CRP condition, in particular, implies the uniquenes (up to

scaling) of the vector normal to $\mathcal{K}$ at the boundary point $(1,\ldots,1)$. We call components $\alpha_j^* > 0$ of the vector $\alpha^* = (\alpha_1^*,\ldots,\alpha_J^*)$, opposite of the above-mentioned normal vector, *server workload contributions* of different servers, and we call by system *server workload* the quantity ${}^uX(t) = \sum_j \alpha_j^* U_j(t)$, where $U_j(t)$ is the *unfinished work* of server j at time t. We establish the relations between customer and server workload contributions, which show that the asymptotic relation $X(t) \leq {}^uX(t)$ between customer workload and server workload exists.

We assume that a strictly convex increasing function $C_j(\cdot)$ is defined for each server j, which is interpreted as the cost rate incurred by the unfinished work on server j.

We introduce two versions of the *MinDrift* routing rule. MinDrift(U) assigns an arriving type i customer to a server

$$j \in \arg\min_j \mu_{ij}^{-1} C_j'(U_j(t)). \tag{1}$$

(This version may not be practical in many cases, because it assumes exact knowledge of the unfinished work values $U_j(t)$.) The MinDrift(Q) rule is the same as MinDrift(U), except $U_j(t)$ in (1) is replaced by the *Q-estimated unfinished work* ${}^qU_j(t) = \sum_m \mu_{mj}^{-1} Q_{mj}(t)$ of server j, where $Q_{ij}(t)$ denotes the number of type i customers in the server j queue. (MinDrift(Q) is a more practical version.)

Our main result (see Theorems 1 and 2) is that, in the OQ system in the heavy traffic asymptotic limit and under the CRP condition, the MinDrift rule (either version), in conjunction with virtually arbitrary work-conserving scheduling disciplines at the servers, minimizes (among all service disciplines) the server workload and the instantaneous and cumulative costs corresponding to the cost rate $\sum_j C_j(U_j(t))$. Moreover, in the limit, the (rescaled) unfinished work vector process $(U_1(t),\ldots,U_J(t))$ exhibits SSC such that the vector $(C_1'(U_1(t)),\ldots,C_J'(U_J(t)))$ is always proportional to α^*. (This behavior is analogous to that exhibited by IQ systems in [12,14], but with the server unfinished works replacing "input"-queue lengths and server workload contributions replacing customer workload contributions.) In addition, the minimal server workload process (attained under MinDrift) matches the minimal possible customer workload process in the corresponding IQ system. As a corollary, MinDrift(Q) (asymptotically) minimizes customer workload among all disciplines in either the OQ or IQ system (see Theorem 3). In this sense, the MinDrift rule controls an OQ system as efficiently as the corresponding IQ system (allowing a wider class of disciplines) can possibly be controlled.

Essentially as another corollary of the main results, we obtain a necessary condition (Theorem 4) for any OQ-system service discipline to have a (asymptotic) workload minimization property. Using this condition, we demonstrate (in Section 13) that even some very natural service disciplines in OQ systems, known to guarantee stability of the queues (if such is feasible at all), do *not* minimize system workload in heavy traffic.

Another contribution of the article is that, in addition to the CRP condition, we in fact identify and characterize a weaker First-Order CRP (FO-CRP) condition.

The purpose of doing this is twofold. First, FO-CRP is sufficient to establish convergence properties of *fluid sample paths*, which arise in the *fluid limit* asymptotic regime and (in addition to being an important step in proving the main heavy traffic results) are of independent interest. Second, this clarifies the role of the additional assumption, which strengthens FO-CRP to CRP, in proving the heavy traffic results.

Finally, we would like to point out that despite the fact that the technical development in this article is in many ways analogous to that in [12,14], some parts of it are quite different. In particular, the representation of the server workload process in Sections 9 and 10 (roughly, as a sum of the "driving" and "regulation" processes) is substantially different from the representation of customer workload processes in [12,14].

1.4. Outline of This Work

In Section 2, we set basic notations and conventions. The OQ-system model is formally introduced in Section 3. In Section 4, we define the (two versions of) MinDrift routing rule and discuss its basic intuition and some examples. The definition and characterization of FO-CRP and CRP conditions in terms of an IQ system are presented in Section 5. Section 6 gives an equivalent characterization of FO-CRP and CRP conditions in terms of an OQ system and establishes relations between customer and server workload contributions. The heavy traffic asymptotic regime is defined in Section 7, which also contains the definitions of and the relations between customer workload and server workload. Section 8 contains formulations of our main results (Theorems 1–4), described earlier. The analysis of fluid sample paths is the subject of Section 9. Section 10 contains the proof of Theorem 1, regarding the asymptotic optimality of the MinDrift(U) rule. The proof of Theorem 2 (regarding the MinDrift(Q) rule) can essentially be reduced to that of Theorem 1—a detailed outline of this reduction is given in Section 11. Theorem 4, a necessary condition for asymptotic workload minimization, is proved in Section 12. We conclude in Section 13 with a discussion of the relation between stability and heavy traffic optimality properties of service disciplines.

2. BASIC NOTATION AND CONVENTIONS

We use the standard notations R and R_+ for the sets of real and real nonnegative numbers, respectively; the not quite standard R_{++} is used for the set of *strictly* positive real numbers. Corresponding N-times product spaces are denoted R^N, R_+^N, and R_{++}^N. The space R^N is viewed as a standard vector-space, with elements $x \in R^N$ being row-vectors $x = (x_1, \ldots, x_N)$. We write simply 0 for the zero vector in R^N and $\mathbf{1} \doteq (1,1,\ldots,1)$ for a vector with all unit coordinates. (The dimensions of vectors 0 and $\mathbf{1}$ are either specified explicitly or are clear from the context.)

The scalar product (dot product) of $x, y \in R^N$ is

$$x \cdot y \doteq \sum_{i=1}^{N} x_i y_i$$

and the norm of x is

$$\|x\| \doteq \sqrt{x \cdot x}.$$

Vector inequalities are to be understood componentwise. As an example, for $\gamma, x \in R^N$, $\gamma < x$ means $\gamma_i < x_i$, $i = 1, \ldots, N$. Also,

$$\gamma \times x \doteq (\gamma_1 x_1, \ldots, \gamma_N x_N),$$

and if $\gamma \in R^N_{++}$, we slightly abuse notation by writing

$$1/\gamma \doteq (1/\gamma_1, \ldots, 1/\gamma_N).$$

We denote the minimum and maximum of two real numbers ξ_1 and ξ_2 by $\xi_1 \wedge \xi_2$ and $\xi_1 \vee \xi_2$, respectively.

Let $D([0,\infty), R)$ be the standard Skorohod space of right-continuous left-limit (RCLL) functions, defined on $[0,\infty)$ and taking real values. (See, for example, [7] for the definition of this space and its associated topology and σ-algebra.)

The symbol $\xrightarrow{w}$ denotes convergence in distribution of *random processes* (or other random elements) (i.e., weak convergence of their *distributions*). Typically, we consider convergence of processes in $D([0,\infty), R)$, or its N-times product space $D^N([0,\infty), R)$ equipped with product topology and σ-algebra.

The symbol $\xrightarrow{\text{u.o.c.}}$ (or the abbreviation u.o.c. after a convergence statement) stands for convergence that is *uniform on compact sets*, for *elements* of $D([0,\infty), R)$ or its N-times product $D^N([0,\infty), R)$. For functions with a bounded domain $A \subset R$, the u.o.c. convergence means uniform convergence.

We reserve the symbol $\Rightarrow$ for weak convergence of *elements* in the space $D([0,\infty), \bar{R})$; the latter is the space of RCLL functions taking values in the set $\bar{R}$ of real numbers, extended to include the two "infinite numbers" $+\infty$ and $-\infty$ (with the natural topology on $\bar{R}$). If $h, g \in D([0,\infty), \bar{R})$, then $h \Rightarrow g$ means $h(t) \to g(t)$ for every $t > 0$ where g is continuous. (Convergence at $t = 0$ is not required.) We will not need any characterization of the topology on $D([0,\infty), \bar{R})$ beyond the definition of convergence given earlier.

3. THE MODEL

We consider a queuing system with a finite number I of customer *types* and a finite number J of flexible *servers*. For notational convenience, we use the symbol I also for the set of types $\{1, \ldots, I\}$. Similarly, J also denotes the set of servers $\{1, \ldots, J\}$.

The arrival process for each type $i \in I$ is a renewal process with the time (from the initial time 0) until the first arrival being $u_i(0)$, and the rest of the interarrival times being an independent and identically distributed (i.i.d.) sequence $u_i(n), n = 1, 2, \ldots$. Let $\lambda_i = 1/E[u_i(1)] > 0$ denote the arrival rate for type i and $\alpha_i^2 = \mathrm{Var}[u_i(1)]$. The service times of type i customers by server $j \in J$ form an i.i.d. sequence $v_{ij}(n), n = 1, 2, \ldots$. Let $\mu_{ij} = 1/E[v_{ij}(1)] < \infty$ and $\beta_i^2 = \mathrm{Var}[v_{ij}(1)]$.

The convention $\mu_{ij} = 0$ is used when server j cannot serve type i. All arrival and service processes are assumed mutually independent.

A version of such a flexible (parallel) server model, which received most attention in the previous work (see [3,8,9,12,20] and references therein) is the *input-queued* model. In the input-queued model, customers of each type i that await service are waiting in a separate "input" queue i of infinite capacity. This, in particular, means that customers do not have to be assigned to the servers while waiting in the input queue; such server assignment is (irreversibly) done when the customer is "pulled" for service by one of the servers.

In this article, we concentrate on a different—*output-queued* (OQ)—model, satisfying the following (additional) *immediate routing* (IR) condition:

> *Each new customer arriving in the system must be assigned to one of the servers j immediately upon arrival, and after that, the customer can only be served by the server to which it is assigned.*

A natural way to interpret the IR condition (and this interpretation is in fact the main motivation for the OQ model) is that, upon arrival, each new customer must be *routed* to one of the servers or, more precisely, into the "output" queue associated with (or "located at") that server.

Remark 1: It should be clear that the IR condition defines the *only* difference between an OQ system and the corresponding IQ system. Therefore, in general, the class of controls (or service disciplines) for an OQ system is a *strict subset* of that for the corresponding IQ system. For example, the G$c\mu$ discipline for IQ systems, studied in [12], does *not* satisfy the IR condition and, consequently, is *not* a valid discipline for OQ systems.

A *service discipline* in an output-queued system consists of two parts: *routing (server assignment) algorithm* and *scheduling rule* employed by each server (and, generally speaking, depending on the server) to determine which customer to serve from its queue (i.e., among the customers assigned to it).

We will consider the class of service disciplines satisfying (in addition to IR) the following condition on the routing algorithm:

> (d0) *The realizations of a customer's service requirements are not known at the time when routing decision (server assignment) for this customer is made. (The distributions of the service requirements at different servers are known.)*

Sometimes, but not always, we will further restrict the class of service disciplines by imposing the following conditions on the server scheduling rules:

> (d1) *Scheduling rule of each server is nonpreemptive within each customer type; namely, a server cannot take for service a new customer of type i if it already has another type i customer "in service" (with both elapsed and residual service times being nonzero). Consequently, at any given time, a server cannot have in service more than one customer of any given type.*

(d2) *A server does not "know" the realization of a customer service time before the customer service starts.*

Note that conditions (d1) and (d2) do *allow* a server idling (even if it has customers in service) or preemption of service of one customer by another customer but of a *different type*. They also allow server-sharing by several customers but, again, each of a different type.

Remark 2: Note that the class of IQ-system service disciplines, satisfying conditions (d0)–(d2) (but not IR), is well defined, and it, first, belongs to the class of disciplines studied in [12] and, second, contains the G$c\mu$ discipline (see [12]). Moreover, this class is obviously a superset of the above-defined class of OQ-system disciplines satisfying (d0)–(d2) and IR.

4. THE MinDrift ROUTING RULE

4.1. Notation

Let $U_j(t)$ denote the *(unfinished) work* of server j at time t; namely the total amount of unfinished processing time of all customers of all types present in server j queue at time t. We denote by $Q_{ij}(t)$ the number of type i customers in queue j at time t, including those customers whose service is already started but not yet completed. The quantity

$$^qU_j \doteq \sum_i (1/\mu_{ij}) Q_{ij}(t)$$

we will call *Q-estimated (unfinished) work* of server j. Finally, by $Q_i(t)$ we will denote the total number of type i customers in the system at time t. In the OQ system, we always have

$$Q_i(t) = \sum_j Q_{ij}(t),$$

but we note that $Q_i(t)$ is well defined for both the OQ and IQ systems.

4.2. MinDrift Rule Definition

Suppose that for each server j, a cost function $C_j(\zeta)$, $\zeta \geq 0$, is given. Assume that the cost functions have the following properties:

$C_j(\cdot)$ is continuous strictly increasing convex, with $C_j(0) = 0$.

Moreover, the first derivative $C_j'(\cdot)$ is continuous strictly increasing, with $C_j'(0) = 0$.

Finally, the second derivative $C_j''(\cdot)$ is strictly positive continuous in the open interval $(0,\infty)$, with $C_j''(0) = \lim_{\zeta \downarrow 0} C_j''(\zeta) \geq 0$, where $C_j''(0)$ is either finite or is $+\infty$.

The MinDrift rule routes (assigns) customers to the servers as follows. When a new customer of type i arrives in the system, it is routed to a server j such that

$$j \in \arg\min_{j \in J} C_j'(U_j(t))/\mu_{ij}. \tag{2}$$

Ties are broken arbitrarily; for example, in favor of the smallest index j. Also, by convention, a type i customer can never be routed to a server j if $\mu_{ij} = 0$. (Throughout this article, we also adopt a related convention that any expression involving division by μ_{ij} holds under the additional assumption that $\mu_{ij} > 0$, even if we do not state this explicitly.)

Defined by (2) is the basic version of the MinDrift rule; we will refer to it as a MinDrift(U) rule.

A version of MinDrift rule, with U_j in (2) replaced by ${}^qU_j$, will be called MinDrift(Q) rule; namely the MinDrift(Q) rule routes an arriving type i customer to a server j such that

$$j \in \arg\min_{j \in J} C_j'({}^qU_j(t))/\mu_{ij}. \tag{3}$$

4.3. Nature of the MinDrift Rule

The nature of the MinDrift rule is simple—it "myopically" (or "greedily") tries to minimize the average drift of the aggregate cost function $\sum_j C_j(U_j(t))$. Indeed, $C_j'(U_j(t))/\mu_{ij}$ (see (2)) approximates the expected increment of the aggregate cost function, caused by routing one type i customer (arrived at time t) to server j; therefore, by (2), MinDrift(U) routes new arrivals in the way such that the (approximate) expected increment of $\sum_j C_j(U_j(t))$ is minimized. In other words, MinDrift(U) routing tries to *minimize the average rate of increase of* $\sum_j C_j(U_j(t))$, *due to placement of new work (or load) to the servers*. Note that in the OQ system, the "best" a service discipline can do to maximize the rate at which $\sum_j C_j(U_j(t))$ is decremented due to processing of the unfinished work, is to never idle servers when they have work to do. Thus, the MinDrift(U) routing rule (in conjunction with arbitrary work-conserving scheduling rules at the servers) strives to minimize the average drift of the aggregate cost.

The MinDrift(Q) rule is based on the same principle as MinDrift(U), except that instead of using the exact values U_j of unfinished work (which may not be available in many applications), it uses their (estimated) average values ${}^qU_j$ (which may be more readily available). As we will demonstrate, in the heavy traffic asymptotic regime we consider, the two versions MinDrift(U) and MinDrift(Q) of the rule are in a certain sense "indistinguishable," under nonrestrictive additional conditions.

The G$c\mu$ scheduling rule for IQ systems, studied in [12], is the rule that myopically tries to minimize the drift of the aggregate cost function $\sum_i C_i(Q_i(t))$ of the queue lengths Q_i, with $C_i(\cdot)$ being cost functions having the same properties as

functions $C_j(\cdot)$ defined earlier. Consequently, G$c\mu$ is such that a server j always tries to serve a queue

$$i \in \arg\max_{i \in I} C_i'(Q_i(t))\mu_{ij},$$

thus *maximizing the average rate at which* $\sum_i C_i(Q_i(t))$ *is decreased due to departures of served customers.* The G$c\mu$ does not—and *cannot*—exercise any control over the rate of increase of $\sum_i C_i(Q_i(t))$ due to new arrivals. Therefore, although both G$c\mu$ (in an IQ system) and MinDrift (in an OQ system) strive to minimize drifts of certain cost functions, they differ in that they control different system state variables: G$c\mu$ controls the rates at which queue lengths Q_i are depleted due to service, and MinDrift controls the rates at which unfinished works U_j are increased due to new arrivals.

4.4. Examples

Consider a special case when the cost functions have the form $C_j(\zeta) = \gamma_j \zeta^{\eta+1}$, with some fixed $\eta > 0$ and $\gamma_j > 0$. Then the MinDrift(U) becomes the rule routing an arriving type i customer to a server

$$j \in \arg\min_{j \in J} \gamma_j(\eta + 1)\frac{[U_j(t)]^\eta}{\mu_{ij}}, \tag{4}$$

and similarly for MinDrift(Q).

Consider a special case of the system such that, for any given pair (ij), we have either $\mu_{ij} = 0$ or $\mu_{ij} = \mu_j > 0$. In other words, the system is such that each server j has a (depending on j) subset of types i that it can serve, but the average service rates of all types within this subset are the same and equal to μ_j (and the server cannot serve at all any types i outside the subset). For this system, the MinDrift(Q) version of (4), with $\eta = 1$, becomes

$$j \in \arg\min_{j \in J} 2\gamma_j \frac{\sum_i Q_{ij}(t)}{\mu_j^2}. \tag{5}$$

Since parameters $\gamma_j > 0$ can be set arbitrarily, we see from (5) that, *for this special system*, such "popular" routing rules as "Join a server j with the shortest queue,"

$$j \in \arg\min_{j \in J} \sum_i Q_{ij}(t), \tag{6}$$

and "Join a server j with the smallest expected unfinished work,"

$$j \in \arg\min_{j \in J} \frac{\sum_i Q_{ij}(t)}{\mu_j}, \tag{7}$$

are special cases of MinDrift(Q). (We remind the reader that, in both cases, routing customers to servers where they cannot be served at all is prohibited. Also, note that if we further assume that each server employs first-in-first-out (FIFO) scheduling, then rule (7) is equivalent to the "Join the shortest expected delay" routing rule.)

It should be clear that in a general system, where service rates μ_{ij} depend on both i and j more generally than in the special system described earlier, the Join-shortest-queue and Join-smallest-expected-unfinished-work routing rules are *not* special cases of MinDrift. Consequently, the heavy traffic optimality properties (which we prove in this article for MinDrift) may not (and typically *do not*) hold for these rules.

5. COMPLETE RESOURCE POOLING CONDITION

In this section, we present the definition of the complete resource pooling (CRP) condition and related notions and results, which are "oriented toward" the analysis of the IQ model and basically follow those in [12]. However, the development in this section is more general than that in Section 5 of [12]. In particular, we consider the notion of First-Order-CRP (FO-CRP) (which is a weaker form of CRP) and prove some additional properties related to this notion. (The results of this section provide a "reference point" for the next section, which gives an equivalent characterization of FO-CRP and CRP conditions "in terms of" the OQ model.)

Consider an $I \times J$ matrix $\phi = \{\phi_{ij},\ i \in I,\ j \in J\}$, with all $\phi_{ij} \geq 0$. Each element ϕ_{ij} can be interpreted as the average rate at which server j time is allocated to the service of type i customers, in the long run. We do not call elements ϕ_{ij} "fractions" of server j time, because, for the reasons which will become clear in Section 6, it will be convenient for us *not* to assume a priori that $\sum_i \phi_{ij} \leq 1$, or even that $\phi_{ij} \leq 1$.

With a given ϕ we associate the vector $\mu(\phi) = (\mu_1(\phi), \ldots, \mu_I(\phi))$, whose coordinates are

$$\mu_i(\phi) \doteq \sum_j \phi_{ij}\, \mu_{ij}, \qquad i \in I; \tag{8}$$

this is the vector of mean long-run service rates provided to the types $i \in I$, if each server j allocates its time to serving type i at the average rate ϕ_{ij}.

Consider also a different vector-function of a matrix ϕ; namely, let the vector $\rho(\phi) = (\rho_1(\phi), \ldots, \rho_J(\phi)) \in R^J$ be defined as

$$\rho_j(\phi) \doteq \sum_i \phi_{ij}, \quad \forall j \in J. \tag{9}$$

Each component $\rho_j(\phi)$ is naturally interpreted as the total "utilization" of server j, given the average rates at which its time is allocated to service of different types i are given by ϕ_{ij}. (Again, we do not assume a priori that $\rho_j(\phi) \leq 1$.)

DEFINITION 1: *We define $\mathcal{M}$ to be the set of $\mu(\phi)$ corresponding to all possible ϕ, satisfying the condition*

$$\rho(\phi) \le \mathbf{1}. \tag{10}$$

Further, let $\mathcal{M}^*$ denote the set of all maximal *elements $\mu \in \mathcal{M}$ such that $\mu \in R^I_{++}$. ($\mu \in \mathcal{M}$ is maximal if $\mu \le \zeta \in \mathcal{M}$ implies $\zeta = \mu$.)*

Note that $\mathcal{M}$ is a bounded convex polyhedron in R^I_+. We assume that $\mathcal{M}$ is nondegenerate (i.e., has dimension I), which is equivalent to assuming that each customer type i can be served at nonzero rate μ_{ij} by at least one server j. The set $\mathcal{M}$ is in fact the closure of our system's *stability region* $\mathcal{M}^0$, which is the set of arrival rate vectors $\lambda = (\lambda_1, \ldots, \lambda_I)$ such that $\lambda < \mu(\phi)$ for some ϕ satisfying (10) (cf. [1,2,6,13–15]).

DEFINITION 2: *We say that the condition of FO-CRP holds for a vector λ if λ lies within the interior of one of the ($(I-1)$-dimensional) outer faces of $\mathcal{M}$ and $\lambda \in \mathcal{M}^*$. If, in addition, the matrix ϕ such that $\lambda = \mu(\phi)$ and (10) hold is unique, then we say that the CRP condition holds.*

Remark: The CRP condition given above is the same as in [12], and it is equivalent to that introduced earlier in [9,20] (see Assumption 3.4, Thm. 5.3 and Cor. 5.4 in [20] for a summary). The fact of equivalence will, in particular, follow from the results of this section.

When the FO-CRP condition holds, let us denote by $\nu = (\nu_1, \ldots, \nu_I)$ the (unique up to a scaling) "outer" normal vector to the polyhedron $\mathcal{M}$ at the point λ. Note that $\nu \in R^I_{++}$. (Otherwise, if some $\nu_i \le 0$, a small increase of the component λ_i would produce a vector $\lambda' \ge \lambda$, $\lambda' \ne \lambda$, and such that $\lambda' \in \mathcal{M}$—a contradiction to the maximality of λ.) For concreteness, we use the normal vector ν^*, which is the vector defined uniquely by the additional requirement that $\|\nu^*\| = 1$. The components ν^*_i of the vector ν^* are sometimes called the *workload contributions* of customers of the different types i (see [9,12,20]); in this article, we will call them *customer workload contributions*, to make a distinction from the server workload contributions introduced in Section 6.

The FO-CRP condition for λ implies, in particular, that

$$\nu^* \cdot \lambda = \max_{\mu \in \mathcal{M}} \nu^* \cdot \mu = \max_{\phi:\, \rho(\phi) \le \mathbf{1}} \nu^* \cdot \mu(\phi); \tag{11}$$

this in turn implies that, for any matrix ϕ such that (10) holds and $\lambda = \mu(\phi)$ (in fact, the equality in (10) must hold); namely, we have

$$\lambda = \mu(\phi) \quad \text{and} \quad \rho(\phi) = 1. \tag{12}$$

Given λ satisfying the FO-CRP condition, for each $j \in J$ let us denote

$$I_j = \arg\max_i \nu^*_i \mu_{ij} \doteq \left\{ i \in I \,|\, \nu^*_i \mu_{ij} = \max_l \nu^*_l \mu_{lj} \right\}.$$

Any pair (ij) such that $i \in I_j$ is called *basic activity*; therefore, I_j indicates the set of basic activities for server j. It is easy to see from (11) that, for any ϕ satisfying (12), $\phi_{ij} > 0$ implies $i \in I_j$.

LEMMA 1: *If λ satisfies the FO-CRP condition, then the corresponding graph $\mathcal{G}^*$ with nodes being type i and servers type j and arcs being basic activities is connected.*

PROOF: Suppose not. Consider any breakdown of the graph $\mathcal{G}^*$ into two components, $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$, disconnected from each other. For $m = 1,2$, denote by $I^{(m)}$ and $J^{(m)}$ the set of types and servers, respectively, within the component $\mathcal{G}^{(m)}$. By our construction, for any $m = 1,2$, $j \in J^{(m)}$ implies $I_j \subseteq I^{(m)}$.

Consider any ϕ satisfying (12). Recall that $\phi_{ij} > 0$ implies $i \in I_j$. Let us fix a small $\delta > 0$, and consider vector ν^{**} obtained from ν^* by the following modification: $\nu_i^{**} = \nu_i^*(1 + \delta)$ if $i \in I^{(1)}$, and $\nu_i^{**} = \nu_i^*$ if $i \in I^{(2)}$. Since there is no arc connecting $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$, if δ is small enough, then ϕ solves the problem

$$\max_{\phi:\rho(\phi)\le \mathbf{1}} \nu^{**}\cdot\mu(\phi)$$

as well as the problem in the right-hand side (RHS) of (11). In other words, $\lambda = \mu(\phi)$ solves the problem $\max_{\mu\in\mathcal{M}} \nu^{**}\cdot\mu$, as well as $\max_{\mu\in\mathcal{M}} \nu^{*}\cdot\mu$. This means that ν^{**} is a normal (different from ν^*) to the boundary of $\mathcal{M}$ at point λ—a contradiction to the FO-CRP condition. ■

Now, with *any* matrix ϕ let us associate graph $\mathcal{G}(\phi)$ with nodes being types i and servers j, and arcs (ij) corresponding to pairs (ij) with $\phi_{ij} > 0$.

LEMMA 2:

(i) If FO-CRP holds, then there exists ϕ satisfying (12) and such that $\phi_{ij} > 0$ if and and only if $i \in I_j$.

(ii) The FO-CRP condition for λ holds if and only if $\lambda \in \mathcal{M}^*$ and there exists ϕ such that (12) holds and the graph $\mathcal{G}(\phi)$ is connected.

(iii) If CRP condition holds, then ϕ satisfying (12) is unique, the graph $\mathcal{G}(\phi) = \mathcal{G}^*$ (i.e., $\phi_{ij} > 0$ if and only if $i \in I_j$), and this graph is a tree.

PROOF:

(i) Consider arbitrary ϕ' satisfying $\rho(\phi') = 1$ and such that $\phi'_{ij} > 0$ if and only if $i \in I_j$. Note that $\nu^*\cdot\mu(\phi') = \nu^*\cdot\lambda$, because the condition on ϕ' guarantees that ϕ' solves the problem in the RHS of (11). Let

$$\lambda'' = \frac{\lambda - \mu(\delta\phi')}{1-\delta},$$

where $0 < \delta < 1$ is fixed. We have

$$\nu^*\cdot\lambda'' = \frac{\nu^*\cdot\lambda - \delta\nu^*\cdot\mu(\phi')}{1-\delta} = \nu^*\cdot\lambda,$$

and we can always choose δ to be small enough so that λ'' lies in the interior of the same face of $\mathcal{M}$ as λ does. Then there exists ϕ'' such that $\lambda'' = \mu(\phi'')$, $\rho(\phi'') = 1$, and $\phi''_{ij} > 0$ implies $i \in I_j$. It is easy to verify directly that $\phi = (1-\delta)\phi'' + \delta\phi'$ is a matrix with the properties we seek.

(ii) Necessity follows from (i). Let us prove sufficiency. Let ϕ be a matrix such that (12) holds, $\lambda \in \mathcal{M}^*$, and the graph $\mathcal{G}(\phi)$ is connected. Since $\lambda \in \mathcal{M}^*$, there exists an outer normal vector ν^* to the boundary of $\mathcal{M}$ at point λ, and it is such that $\nu^* \neq 0$, $\nu^* \in R^I_+$. Consequently, ϕ solves the problem (in the RHS of) (11). From this, we conclude that $\nu^* \in R^I_{++}$; otherwise (if $\nu^*_i = 0$ for some i), ϕ could not be a solution to (11), since $\mathcal{G}(\phi)$ is connected. Finally, the normal ν^* must be unique (up to scaling). Indeed, consider any other vector $\nu^{**} \in R^I_{++}$, which is not a scaled version of ν^*; that is, $\max_i \nu^{**}_i/\nu^*_i > \min_i \nu^{**}_i/\nu^*_i$. Then connectedness of $\mathcal{G}(\phi)$ easily implies that since ϕ solves (11), it cannot possibly solve (11) with ν^* replaced by ν^{**}. Therefore, ν^{**} cannot be a normal to $\mathcal{M}$ at point λ.

(iii) The definition of CRP and statements (i) and (ii) of the lemma immediately apply the uniqueness of ϕ satisfying (12)—the fact that $\mathcal{G}(\phi) = \mathcal{G}^*$ and that this graph is connected. It remains to show that $\mathcal{G}(\phi)$ must be a tree. Suppose not. Let us pick any cycle on this graph. It is easy to see that we can "perturb" the (strictly positive) elements ϕ_{ij} along the arcs of the cycle so as to produce a matrix $\phi' \neq \phi$ such that $\mu(\phi') = \mu(\phi) = \lambda$ and $\rho(\phi') = \mathbf{1}$, a contradiction to the uniqueness of ϕ. ■

Remark: Just as in the case of the G$c\mu$ scheduling rule, studied for an IQ model in [12], we emphasize here that the notion of a basic activity is *not* utilized in any way (neither explicit nor implicit) by the MinDrift routing algorithm. (The algorithm need not know which activities are basic.) It is only used as a tool for the analysis of the algorithm. Similarly, the algorithm need not know the values of workload contributions.

6. EQUIVALENT CHARACTERIZATION OF THE CRP CONDITION IN TERMS OF THE OQ MODEL

In this section, we give an equivalent ("dual") characterization of the CRP condition and introduce notions and results that will be used in the analysis of our OQ-model.

First, let us give a somewhat different (although closely related) interpretation of the matrix ϕ and functions $\mu(\phi)$ and $\rho(\phi)$, defined in Section 5. Suppose a matrix ϕ is given, and assume that customers of a type i arrive (routed) to a server j at the average rate $\mu_{ij}\phi_{ij}$. Then $\mu_i(\phi)$, $i \in I$, is the total average rate at which type i customers arrive in the system, and

$$\rho_j(\phi) = \sum_i \phi_{ij} \equiv \sum_i (\mu_{ij}\phi_{ij})\mu_{ij}^{-1}, \qquad j \in J,$$

is the average rate at which the *work* (i.e., the required amount of processing time) arrives (routed) to server j. (We used the convention that $(\mu_{ij}\phi_{ij})\mu_{ij}^{-1} = 0$ if $\mu_{ij} = 0$.)

DEFINITION 3: *We define the* server utilization region $\mathcal{K} \subseteq R_+^J$ *to be the set of all possible values of* $\rho(\phi)$ *with* ϕ *satisfying the condition*

$$\mu(\phi) \geq \lambda. \tag{13}$$

Further, let $\mathcal{K}_*$ *denote the set of all* minimal *elements* $\rho \in \mathcal{K}$ *such that* $\rho \in R_{++}^J$. ($\rho \in \mathcal{K}$ *is minimal if* $\rho \geq \zeta \in K$ *implies* $\zeta = \rho$.)

Region $\mathcal{K}$ is a convex polyhedron in R_+^J, and it is nondegenerate (i.e., has dimension J) as long as $\mathcal{M}$ is nondegenerate. Note that $\mathcal{K}$ is unbounded, but it is, of course, "bounded below," say by 0, since it lies in the positive orthant.

LEMMA 3:

(i) The FO-CRP condition for a fixed vector λ *holds if and only if the following is true:*
  (a) Vector $\mathbf{1} \in R^J$ *lies within the interior of one of the* $((J-1)$*-dimensional) faces of* $\mathcal{K}$.
  (b) $\mathbf{1} \in \mathcal{K}_*$.

(ii) When the FO-CRP condition for λ *(or, equivalently, (i)(a) and (i)(b)) holds, then the unique (up to a scaling by positive constant) outer normal vector* $-\alpha^*$ *to the polyhedron* $\mathcal{K}$ *at the point* $\mathbf{1}$ *is such that* $\alpha^* = (\alpha_1^*, \ldots, \alpha_J^*) \in R_{++}^J$, *and* α^* *is related to the vector* ν^* *as follows:*

$$\alpha_j^* = \max_i \mu_{ij}\nu_i^*, \qquad j \in J, \tag{14}$$

$$\nu_i^* = \min_j \alpha_j^*/\mu_{ij}, \qquad i \in I. \tag{15}$$

In addition:
(c) $i \in I_j$ *(i.e., activity* (ij) *is basic) if and only if* $j \in J_i$, *where*

$$J_i \doteq \arg\min_j \alpha_j^*/\mu_{ij}. \tag{16}$$

(d) Any matrix ϕ *satisfying* $\rho(\phi) = \mathbf{1}$ *and (13), in fact, satisfies (12).*
(e) We have

$$\alpha^* \cdot \mathbf{1} = \nu^* \cdot \lambda. \tag{17}$$

(iii) The CRP condition for a fixed vector λ *holds if and only if (i)(a), (i)(b), and the following property hold:*
(f) the matrix ϕ *satisfying* $\rho(\phi) = \mathbf{1}$ *and (13) is unique.*
When CRP does hold, the matrix ϕ *is in fact the unique solution of (12).*

PROOF:

(i) Let us prove the necessity of (a) and (b). Consider the vector α^* defined by (14). (Note that $\alpha^* \in R^J_{++}$.) Then (15) holds. Indeed, for a fixed i, we have $\alpha^*_j = \mu_{ij}\nu^*_i$ if (ij) is basic, and $\alpha^*_j > \mu_{ij}\nu^*_i$ otherwise. (Incidentally, this means that $i \in I_j$ is equivalent to $j \in J_i$.)

Let us choose any matrix ϕ that solves (12) and such that $\mathcal{G}(\phi) = \mathcal{G}^*$. (Recall that graph $\mathcal{G}^*$ is connected.) Then since this ϕ is such that $\phi_{ij} > 0$ implies $j \in J_i$, it is easy to observe that ϕ solves the problem

$$\min_{\phi:\mu(\phi)\geq\lambda} \alpha^* \cdot \rho(\phi) \tag{18}$$

or, equivalently, **1** solves the problem

$$\min_{\rho\in\mathcal{K}} \alpha^* \cdot \rho. \tag{19}$$

(Compare problems (18) and (19) to problem (11).) This means, in particular, that **1** is a minimal element of $\mathcal{K}$, and α^* is a normal to the region $\mathcal{K}$ at point **1**. Since $\mathcal{G}(\phi)$ is connected, it is easy to see from (18) that $\rho =$ **1** could not possibly solve the problem (19) with α^* replaced by any other nonzero vector $\alpha^{**} \in R^J_+$, unless $\alpha^{**} = c\alpha^*$ for some $c > 0$. This completes the proof of necessity of (a) and (b).

The sufficiency of (a) and (b) follows simply by the symmetry between the definitions of the FO-CRP condition (for λ) and conditions (a) and (b) (expressed in terms of vector **1**); namely if for the vector $-\mathbf{1}$ and the region $-\mathcal{K}$ we define a condition analogous to FO-CRP for λ and region $\mathcal{M}$, this will be exactly conditions (a) and (b). Thus, from this condition ((a) and (b)) we can obtain a condition analogous to (a) and (b), but with **1** and $\mathcal{K}$ replaced by $-\lambda$ and $-\mathcal{M}$, respectively; this latter condition is exactly our original FO-CRP.

(ii) As part of the proof of (i), we already proved all the properties stated in (ii), except (d) and (e). If (d) would not hold, then λ could not be a maximal element of $\mathcal{M}$. Property (e) follows from the fact that any matrix ϕ, chosen as in the proof of (i), satisfies (12) and solves both problems (11) and (18).

(iii) This follows from (i), (ii), and the definition of CRP. ■

When the FO-CRP condition holds, the components α^*_j of the vector α^* we will call *server workload contributions* of different servers j.

7. HEAVY TRAFFIC REGIME

In this section, we introduce the notion of a *sequence of queuing systems in heavy traffic*. Suppose a vector λ satisfying the CRP condition is fixed. Associated with this λ are the unique matrix ϕ satisfying (12) and the corresponding normal vectors $\nu^* \in R^I$ and $\alpha^* \in R^J$. (We remark that all of the definitions and facts in this section

are valid when the weaker FO-CRP condition holds, with *arbitrary* fixed ϕ satisfying (12). However, for our main results in Section 8, the stronger CRP condition is essential.)

The quantity

$$X(t) \doteq \sum_{i=1}^{I} \nu_i^* Q_i(t), \qquad t \geq 0,$$

we will call the *customer workload* of the system. The customer workload process $X(\cdot)$ is of primary interest in the analysis of the IQ model [3,8,9,12,20].

For the OQ model, we define a different (although closely related, in the sense specified later) notion of *server workload*:

$$^{u}X(t) \doteq \sum_{j=1}^{J} \alpha_j^* U_j(t), \qquad t \geq 0.$$

In addition, we define the *Q-estimated server workload* as

$$^{q}X(t) \doteq \sum_{j=1}^{J} \alpha_j^{\,q} U_j(t), \qquad t \geq 0.$$

Informally speaking, for a service discipline satisfying constraints (d0)–(d2), $^{q}X(t)$ is a "good" (asymptotically exact) estimate of the server workload $^{u}X(t)$.

Since for any pair of $i \in I$ and $j \in J$ the inequality $\alpha_j^*/\mu_{ij} \geq \nu_i^*$ holds, we observe that the *Q*-estimated server workload cannot be less than customer workload:

$$^{q}X(t) \equiv \sum_{i=1}^{I} \sum_{j=1}^{J} (\alpha_j^*/\mu_{ij}) Q_{ij} \geq X(t). \tag{20}$$

We also have the following inequality, which we record for future reference:

$$^{q}X^{r}(t) \leq C_0 X^{r}(t), \qquad t \geq 0, \tag{21}$$

with

$$C_0 = \max_{(ij):\mu_{ij}>0} \frac{\alpha_j^*/\mu_{ij}}{\nu_i^*}. \tag{22}$$

We now consider a sequence of queuing systems, indexed by $r \in \mathcal{R} = \{r_1, r_2, \ldots\}$, where $r_n > 0$ for all n and $r_n \uparrow \infty$ as $n \to \infty$. (Hereafter in this article, $r \to \infty$ means that r goes to infinity along the sequence $\mathcal{R}$ or some subsequence of $\mathcal{R}$; the choice of the subsequence will be either explicit or clear from the context.) Each system $r \in \mathcal{R}$ has, as earlier, I customer types and J servers. The primitives and the processes corresponding to a system $r \in \mathcal{R}$ will be appended with a superscript r.

Assume that for each type i, the mean arrival rate $\lambda_i^r = 1/E[u_i^r(1)]$ is such that

$$r(\lambda_i^r - \lambda_i) \to b_i, \qquad r \to \infty, \tag{23}$$

where $b_i \in R$ is a fixed constant. Assume also convergence of the variance; that is,

$$[\alpha_i^r]^2 \to \alpha_i^2, \qquad r \to \infty. \tag{24}$$

In addition, we make the following technical assumption, needed, in particular, to apply Bramson's weak law estimates [4] (and establish (75) later): Uniformly over i and r,

$$E[(u_i^r(1))^2 1\{u_i^r(1) > x\}] \le \eta(x), \qquad x \ge 0, \tag{25}$$

where $\eta(\cdot)$ is a fixed function and $\eta(x) \to 0$ as $x \to \infty$.

For the initial interarrival times, we assume that for each i,

$$u_i^r(0)/r \to 0, \qquad r \to \infty.$$

Let $F_i^r(t)$, $t \ge 0$, denote the number of type i customers that arrived in the system by time t, excluding "initial" customers present at time 0. Assumptions (23), (24), and (25) imply a functional central limit theorem (FCLT) for these arrival processes:

$$\{r^{-1}(F_i^r(r^2 t) - \lambda_i^r r^2 t),\ t \ge 0\} \xrightarrow{w} \{\sigma_i B(t),\ t \ge 0\}, \tag{26}$$

where $\sigma_i^2 = \lambda_i^3 \alpha_i^2$, $B(\cdot)$ is a standard (zero drift, unit variance) Brownian motion, and $\xrightarrow{w}$ denotes convergence in distribution (for processes in the standard Skorohod space of RCLL functions).

The service time distributions do *not* change with the parameter r. (This in particular means that the condition analogous to (25) trivially holds for the service times $v_{i,j}^r(1)$, uniformly on (i,j) and r.) Let us denote by

$$^{\Sigma}V_{ij}^r(l) \doteq \sum_{m=1}^{l} v_{ij}^r(m), \qquad l = 0,1,2,\ldots,$$

the total amount of work (i.e., total service time) brought to server j by the first l (newly arriving) type i customers routed to it. We extend the domain of $^{\Sigma}V_{ij}^r(\cdot)$ to all real nonnegative $t \ge 0$ by adopting the convention that $^{\Sigma}V_{ij}^r(t) = {}^{\Sigma}V_{ij}^r(\lfloor t \rfloor)$. (We will use the same domain extension convention throughout the article for other functions, which are originally defined on the integers, as well.) For $^{\Sigma}V_{ij}^r$, we have the following FCLT:

$$\{r^{-1}(^{\Sigma}V_{ij}^r(r^2 t) - \mu_{ij}^{-1} r^2 t),\ t \ge 0\} \xrightarrow{w} \{\beta_{ij} B(t),\ t \ge 0\}. \tag{27}$$

For each $i \in I$, let us fix a set of integer-valued nondecreasing nonnegative functions $(^{s}N_{ij}(l),\ l = 0,1,2,\ldots)$, $j \in J$, satisfying the following conditions:

$$\sum_j {}^{s}N_{ij}(l) = l, \qquad l = 0,1,2,\ldots, \tag{28}$$

$$\max_{l \ge 0} |{}^{s}N_{ij}(l) - \frac{\mu_{ij}\phi_{ij}}{\lambda_i} l| < \infty, \qquad j \in J, \tag{29}$$

$$^{s}N_{ij}(l) \equiv 0 \quad \text{for } l = 0,1,2,\ldots, \qquad j \notin J_i. \tag{30}$$

The value of ${}^sN_{ij}(l)$ is interpreted as the number of type i customers routed to server j, out of the first l type i customers arriving in the system. Then it is clear that for each flow i, the functions ${}^sN_{ij}(\cdot)$ define a fixed ("static") pattern of routing customers to the servers in J_i, such that, for any l, the fractions of customers routed to different servers $j \in J_i$ closely track the values $\mu_{ij}\phi_{ij}/\lambda_i$; recall that $\sum_j \mu_{ij}\phi_{ij}/\lambda_i = 1$. (The MinDrift rule does not require any knowledge of this static routing pattern; it is only used as a tool for the analysis!) Such functions can be, for example, constructed recursively as follows. We set ${}^sN_{ij}(0) = 0$ for all $j \in J_i$. For each $l = 1,2,\ldots$, we set ${}^sN_{ij}(l) = {}^sN_{ij}(l-1) + 1$ for one of the $j \in J_i$ with the smallest value of ${}^sN_{ij}(l-1) - (\mu_{ij}\phi_{ij}/\lambda_i)l$, and ${}^sN_{ij}(l) = {}^sN_{ij}(l-1)$ for all other j. (The "ties" between j are broken arbitrarily, for example, in favor of the smallest one.)

For $i \in I$, let us denote by

$$A_i^r(t) \doteq \sum_{j\in J} \alpha_j^{*\Sigma} V_{ij}^r({}^sN_{ij}(F_i^r(t))) \equiv \sum_{j\in J} \alpha_j^* \sum_{m=1}^{{}^sN_{ij}(F_i^r(t))} v_{ij}^r(m), \qquad t \geq 0,$$

the total amount of server workload brought to the system by the new arrivals of flow i by time $t \geq 0$ *assuming the arrivals would be routed according to the (fixed) functions* ${}^sN_{ij}(\cdot)$. From (26)–(30) we obtain the following FCLT for the sequence of processes $A_i^r(\cdot)$:

$$\left\{ r^{-1}\left(A_i^r(r^2 t) - \lambda_i^r \frac{\sum_j \phi_{ij}\alpha_j^*}{\lambda_i} r^2 t \right), t \geq 0 \right\} \xrightarrow{w} \{{}^u\sigma_i B(t), t \geq 0\}, \tag{31}$$

where

$$\begin{aligned} {}^u\sigma_i^2 &\doteq \lambda_i \left[\sum_j \phi_{ij}\alpha_j^* \right]^2 \alpha_i^2 + \sum_j \phi_{ij}\mu_{ij}\beta_{ij}^2 [\alpha_j^*]^2 \\ &= [\nu_i^*]^2 \left[\lambda_i^3 \alpha_i^2 + \sum_j \phi_{ij}\mu_{ij}^3\beta_{ij}^2 \right]. \end{aligned} \tag{32}$$

From (31) and (17) we obtain the following FCLT for the sequence of processes $\sum_i A_i^r(\cdot)$:

$$\left\{ r^{-1}\left(\sum_i A_i^r(r^2 t) - \left[\sum_j \alpha_j^* \right] r^2 t \right), t \geq 0 \right\} \xrightarrow{w} \{at + \sigma B(t), t \geq 0\}, \tag{33}$$

where

$$\sigma^2 \doteq \sum_i {}^u\sigma_i^2, \qquad a \doteq \nu^* \cdot b. \tag{34}$$

8. MAIN RESULTS

For each value of the (scaling) parameter $r \in \mathcal{R}$, let us consider the following processes. Let $U^r(\cdot)$ and ${}^qU^r(\cdot)$ be the (vector) processes, representing (unfinished) work and Q-estimated (unfinished) work, respectively, at different servers; let ${}^uX^r(\cdot)$, ${}^qX^r(\cdot)$, and $X^r(\cdot)$ denote the scalar processes representing server workload, Q-estimated server workload, and customer workload, respectively.

Assume that in a system with index $r \in \mathcal{R}$, each server j, at any time t, incurs a *holding cost* at the (instantaneous) rate of

$$C_j^r(U_j^r(t)) = C_j(U_j^r(t)/r),$$

where $C_j(\cdot)$ are fixed convex increasing functions, with the additional properties described in Section 4.

Note that in our asymptotic regime the cost function is "rescaled" as the parameter r changes. (In other words, in a system with index r, the holding cost rate corresponding to the unfinished work $U_j^r(t)$ is $C_j(U_j^r(t)/r)$ instead of $C_j(U_j^r(t))$.) Notice, however, that in the special case (described in Section 4.4) when $C_j(\zeta) = \gamma_j \zeta^{\eta+1}$, with some fixed $\eta > 0$ and $\gamma_j > 0$, *the form of the corresponding MinDrift rule does not change with r*. Indeed, in this case, replacing $C_j'(U_j^r(t))$ in (2) with $C_j'(U_j^r(t)/r)/r$ simply does not change the routing rule.

For our main results, we need the notion of a fixed point. A vector ${}^\circ u \in R_+^J$ will be called a *fixed point* if

$$[C_1'({}^\circ u_1), \ldots, C_J'({}^\circ u_J)] = c\alpha^*, \tag{35}$$

for some constant $c \geq 0$. If we recall that each derivative $C_j'(\cdot)$ is continuous strictly increasing with $C_j'(0) = 0$, one deduces the following:

> *A fixed point* ${}^\circ u$ *corresponding to each* $c \geq 0$ *exists and is unique. Moreover,* ${}^\circ u = 0$ *for* $c = 0$, *and* ${}^\circ u \in R_{++}^I$ (*i.e., has all components strictly positive*) *for any* $c > 0$.

Thus, the set of fixed points forms a one-dimensional manifold, which can be parameterized, for example, by the corresponding server workload values $\alpha^* \cdot {}^\circ u$. In addition, it is easy to verify the following property:

> *A fixed point* ${}^\circ u$ *is the unique vector that minimizes* $\sum_j C_j(u_j)$ *among all vectors* $u \in R_+^J$ *with the same server workload (i.e., satisfying condition* $\alpha^* \cdot u = \alpha^* \cdot {}^\circ u$).

Indeed, if ${}^\circ u = 0$, the property is trivial. If ${}^\circ u \in R_{++}^J$, condition (35) implies that the (Lagrangian) function

$$\sum_j C_j(u_j) - c[\alpha^* \cdot u - \alpha^* \cdot {}^\circ u]$$

has zero gradient (with respect to u) at point ${}^\circ u$. Since this Lagrangian is strictly convex in R_+^J, it is minimized by ${}^\circ u$. Then the desired property follows from the Kuhn–Tucker theorem.

Let us define the *diffusion scaling* operator $\tilde{\Gamma}^r$, which acts on a scalar function $\Xi = (\Xi(t), t \geq 0)$ as

$$(\tilde{\Gamma}^r \Xi)(t) \doteq \frac{1}{r} \Xi(r^2 t) \tag{36}$$

and is applied to vertor-functions componentwise.

Let us consider the following *diffusion-scaled* processes: $\tilde{u}^r = \tilde{\Gamma}^r U^r$, ${}^q\tilde{u}^r = \tilde{\Gamma}^r {}^q U^r$, ${}^u\tilde{x}^r = \tilde{\Gamma}^r {}^u X^r$, ${}^q\tilde{x}^r = \tilde{\Gamma}^r {}^q X^r$, and $\tilde{x}^r = \tilde{\Gamma}^r X^r$.

8.1. Optimality of the MinDrift(U) Rule

Assume that the initial (scaled) amounts of unfinished work are deterministic and converging:

$$\tilde{u}^r(0) \to \tilde{u}(0), \tag{37}$$

where $\tilde{u}(0)$ is a fixed point, as defined earlier. Consequently, ${}^u\tilde{x}^r(0) = \alpha^* \cdot \tilde{u}^r(0) \to \alpha^* \cdot \tilde{u}(0) \doteq \tilde{w}(0)$.

Let us define the following one-dimensional reflected Brownian motion $\tilde{x} = \{\tilde{x}(t), t \geq 0\}$:

$$\tilde{x}(t) = \tilde{w}(0) + at + \sigma B(t) + \tilde{y}(t), \tag{38}$$

where $B(\cdot)$ is a standard Brownian motion,

$$\tilde{y}(t) \doteq -\left[0 \wedge \inf_{0 \leq \xi \leq t} \{\tilde{w}(0) + a\xi + \sigma B(\xi)\}\right], \tag{39}$$

and the drift a and diffusion coefficient σ are given in (34) and (32), respectively.

THEOREM 1: *Consider the sequence of queueing systems in heavy traffic, as introduced in Section 7. Assume initial condition (37). Let $\tilde{x}$ be a reflected Brownian motion defined by (38) and (39).*

(i) Suppose that the service discipline is such that the routing rule is MinDrift(U) with cost functions $C_i^r(\cdot)$, for each value of the parameter r, and each server employs an arbitrary work-conserving scheduling rule (namely the server is not allowed to idle when there is unfinished work in its queue). Then, as $r \to \infty$,

$$ {}^u\tilde{x}^r \xrightarrow{w} \tilde{x}, $$

and

$$ \tilde{u}^r \xrightarrow{w} \tilde{u}, $$

where for each $t \geq 0$, the vector $\tilde{u}(t)$ is the fixed point that is (uniquely) determined by $\alpha^* \cdot \tilde{u}(t) = \tilde{x}(t)$.

(ii) The service discipline defined in (i) is asymptotically optimal within the class of service disciplines satisfying condition (d0) in that it minimizes the server workload and the unfinished work cost rate at all times. More precisely, let $\tilde{u}^r_G$ and ${}^u\tilde{x}^r_G$ be the scaled unfinished work and server workload processes corresponding to an arbitrary service discipline G (and appropriately constructed on a common probability space with the sequence of processes defined in (i)). Then, with probability 1, *for any time* $t \geq 0$,

$$\liminf_{r\to\infty} \inf_{\xi\in[0,\,t]} [{}^u\tilde{x}^r_G(\xi) - \tilde{x}(\xi)] \geq 0 \tag{40}$$

and

$$\liminf_{r\to\infty} \sum_j C_j(\tilde{u}^r_{j,G}(t)) \geq \sum_j C_j(\tilde{u}_j(t)). \tag{41}$$

As a corollary, with probability 1, *for any* $T > 0$,

$$\begin{aligned}\liminf_{r\to\infty} \int_0^T \sum_j C_j(\tilde{u}^r_{j,G}(t))\,dt &\geq \lim_{r\to\infty} \int_0^T \sum_j C_j(\tilde{u}^r_j(t))\,dt \\ &= \int_0^T \sum_j C_j(\tilde{u}_j(t))\,dt.\end{aligned} \tag{42}$$

The proof of Theorem 1 is the subject of Sections 9 and 10.

8.2. Optimality of the MinDrift(Q) Rule

Assume that, for each r, the initial state of the system at time 0 is deterministic and it conforms to conditions (d1) and (d2) on a service discipline (which are assumed in Theorem 2 below). In particular, for each pair of i and j, server j has in its queue at most one customer of type i whose service has already started, and the realizations of service times of the customers whose service has not yet started are not known to the server. For the initial residual service times $v^r_{i,j}(0)$ (if any) of the customers whose service has already started, we assume

$$v^r_{i,j}(0)/r \to 0, \qquad r \to \infty.$$

Finally, assume that the initial (scaled) amounts of Q-estimated unfinished work are converging:

$${}^q\tilde{u}^r(0) \to \tilde{u}(0), \tag{43}$$

where $\tilde{u}(0)$ is a fixed point, as defined earlier.

It follows from the above initial conditions that

$${}^q\tilde{x}^r(0) = \alpha^* \cdot {}^q\tilde{u}^r(0) \to \alpha^* \cdot \tilde{u}(0) \doteq \tilde{w}(0) \tag{44}$$

and, in addition, with probability 1,

$$\tilde{u}^r(0) \to \tilde{u}(0) \quad \text{and} \quad {}^u\tilde{x}^r(0) \to \tilde{w}(0). \tag{45}$$

For the fixed $\tilde{w}(0)$, we consider a one-dimensional reflected Brownian motion $\tilde{x} = \{\tilde{x}(t), t \geq 0\}$, defined in (38).

THEOREM 2: *Consider the sequence of queuing systems in heavy traffic, as introduced in Section 7, and with the initial conditions described in Section 8.2.*

(i) Suppose that the service discipline is such that the routing rule is Min-Drift(Q) with cost functions $C_j^r(\cdot)$, for each value of the parameter r, and each server employs an arbitrary work-conserving scheduling rule satisfying conditions (d1) and (d2). Then, as $r \to \infty$,

$${}^q\tilde{x}^r \xrightarrow{w} \tilde{x}, \qquad {}^u\tilde{x}^r \xrightarrow{w} \tilde{x}$$

and

$${}^q\tilde{u}^r \xrightarrow{w} \tilde{u}, \qquad \tilde{u}^r \xrightarrow{w} \tilde{u},$$

where, for each $t \geq 0$, the vector $\tilde{u}(t)$ is the fixed point that is (uniquely) determined by $\alpha^* \cdot \tilde{u}(t) = \tilde{x}(t)$.

(ii) The service discipline defined in (i) is asymptotically optimal within the class of service disciplines satisfying conditions (d0)–(d2) in that it minimizes both the server workload and the Q-estimated server workload and the unfinished work cost rate at all times. More precisely, let $\tilde{u}_G^r$, ${}^q\tilde{u}_G^r$, ${}^u\tilde{x}_G^r$, and ${}^q\tilde{x}_G^r$ be the scaled unfinished work, Q-estimated unfinished work, server workload, and Q-estimated server workload processes, respectively, corresponding to an arbitrary service discipline G satisfying (d0)–(d2) (and appropriately constructed on a common probability space with the sequence of processes in (i)). Then, with probability 1, for any time $t \geq 0$,

$$\liminf_{r\to\infty} \inf_{\xi\in[0,t]} [{}^q\tilde{x}_G^r(\xi) - \tilde{x}(\xi)] = \liminf_{r\to\infty} \inf_{\xi\in[0,t]} [{}^u\tilde{x}_G^r(\xi) - \tilde{x}(\xi)] \geq 0 \tag{46}$$

and

$$\begin{aligned} \liminf_{r\to\infty} \sum_j C_j({}^q\tilde{u}_{j,G}^r(t)) &= \liminf_{r\to\infty} \sum_j C_j(\tilde{u}_{j,G}^r(t)) \\ &\geq \sum_j C_j(\tilde{u}_j(t)). \end{aligned} \tag{47}$$

As a corollary, with probability 1, for any $T > 0$,

$$\lim_{r\to\infty} \int_0^T \sum_j C_j({}^q\tilde{u}_j^r(t))\,dt = \lim_{r\to\infty} \int_0^T \sum_j C_j(\tilde{u}_j^r(t))\,dt$$
$$= \int_0^T \sum_j C_j(\tilde{u}_j(t))\,dt \qquad \textbf{(48)}$$

and

$$\liminf_{r\to\infty} \int_0^T \sum_j C_j({}^q\tilde{u}_{j,G}^r(t))\,dt = \liminf_{r\to\infty} \int_0^T \sum_j C_j(\tilde{u}_{j,G}^r(t))\,dt$$
$$\geq \int_0^T \sum_j C_j(\tilde{u}_j(t))\,dt. \qquad \textbf{(49)}$$

The proof of Theorem 2 is essentially a slightly modified (and extended) version of that of Theorem 1. It is outlined in Section 11.

8.3. Customer Workload Minimization Under the MinDrift(Q) Rule

Suppose that we are within the conditions of Theorem 2. For the initial customer workloads $\tilde{x}^r(0)$, we "automatically" (by (20) and (44)) have

$$\limsup_r \tilde{x}^r(0) \leq \lim_r {}^q\tilde{x}^r(0) = \tilde{w}(0).$$

Suppose, in addition, that in fact, $\tilde{x}^r(0)$ converges to the same limit as ${}^q\tilde{u}^r(0)$:

$$\lim_r \tilde{x}^r(0) = \lim_r {}^q\tilde{x}^r(0) = \tilde{w}(0), \qquad \textbf{(50)}$$

which is equivalent to the condition that $\lim_r \tilde{q}_{ij}^r(0) = 0$ for every nonbasic activity (ij).

Now, any service discipline in the OQ system, satisfying conditions (d0)–(d2), is within the class of disciplines for the corresponding IQ system studied in [12] (see Remark 2 in Section 3 of this article). In particular, Theorem 1 in [12] establishes that reflected Brownian motion (RBM) with exactly the same distribution as the RBM $\tilde{x}$ (defined in this article by (38)) is in fact the stochastic *lower* bound of any limit of the customer workload process $\tilde{x}^r$, under any service discipline satisfying conditions (d0)–(d2) (but not necessarily IR), which includes service disciplines for both OQ and IQ systems.

However, from (20) we have a pathwise relation

$$\tilde{x}^r(t) \leq {}^q\tilde{x}^r(t), \qquad t \geq 0, \qquad \textbf{(51)}$$

and, by Theorem 2(i), under the MinDrift(Q) rule, ${}^q\tilde{x}^r \xrightarrow{w} \tilde{x}$. Thus, $\tilde{x}$ is both the lower and upper (stochastic) bounds of $\tilde{x}^r$ and, therefore, $\tilde{x}^r \xrightarrow{w} \tilde{x}$. We have proved

the following result, which basically says that the MinDrift(Q) rule minimizes customer workload among virtually all service disciplines in either the OQ or IQ system.

THEOREM 3: *Suppose that the conditions of Theorem 2 and, in addition, condition (50) hold.*

(i) For the service discipline described in Theorem 2(i), we, in addition, have

$$\tilde{x}^r \xrightarrow{w} \tilde{x}. \tag{52}$$

(ii) The service discipline described in Theorem 2(i) asymptotically minimizes customer workload among all service disciplines, satisfying conditions (d0)–(d2), in either the OQ or IQ system; namely, the customer workload process $\tilde{x}^r_G$ under any such discipline G can be constructed on a common probability space with the RBM $\tilde{x}$, so that, with probability 1, for any time $t \geq 0$,

$$\liminf_{r\to\infty} \inf_{\xi\in[0,t]} [\tilde{x}^r_G(\xi) - \tilde{x}(\xi)] \geq 0.$$

8.4. A Necessary Condition for Workload Minimization Under Any Service Discipline: Vanishing Nonbasic Queues

Theorems 2 and 3 show that, roughly speaking, the MinDrift(Q) rule minimizes both server and customer workload in the heavy traffic limit. Theorem 4 demonstrates that (either customer or server) workload in an OQ system can be minimized by some discipline only if, under this discipline, the (scaled) queue lengths $\tilde{q}^r_{ij}$ corresponding to nonbasic activities (ij) vanish in the limit.

THEOREM 4: *Suppose that the conditions of Theorem 2 and, in addition, condition (50) hold. Suppose that under some service discipline in the OQ system, satisfying conditions (d0)–(d2), either (52) or*

$${}^u\tilde{x}^r \xrightarrow{w} \tilde{x} \tag{53}$$

or

$${}^q\tilde{x}^r \xrightarrow{w} \tilde{x} \tag{54}$$

holds. Then, for any $t \geq 0$,

$$\tilde{q}^r_{ij}(t) \xrightarrow{P} 0 \quad \textit{for any nonbasic activity } (ij). \tag{55}$$

The proof is presented in Section 12.

9. FLUID SAMPLE PATHS UNDER THE MinDrift RULE

9.1. Fluid Sample Paths: Definition and Basic Properties

In this section, we study the sequence of processes introduced in the previous section under the *fluid* (or "law of large numbers") scaling and under the MinDrift rule. More precisely, we need to consider only *sample paths* of the processes under this scaling and then their limits, which we formally define here and call *fluid sample paths* (FSPs). The key property of FSPs that must be established (in Theorem 5) is that as time increases to infinity, the queue length vector converges to a fixed point. This attraction property is used to prove (in Section 10) the state space collapse property (i.e., the property that the limit of the sequence of *diffusion scaled* processes is a process "living" on the manifold of fixed points).

Throughout Section 9 we assume the CRP condition. However, all definitions and results of this section hold under the weaker FO-CRP condition, with *arbitrary* fixed ϕ satisfying (12) and $\mathcal{G}(\phi) = \mathcal{G}^*$; in other words, the FSP definition and key properties do *not* require a solution ϕ of (12) to be unique.

First, we introduce some additional (random) functions, associated with the process for each value of the scaling parameter r. (The functions $F_i^r(t)$, $Q_i^r(t)$, and $X^r(t)$, have already been defined earlier.) Denote by $G_{ij}^r(t)$ the amount of time within $[0,t]$ that server j was serving type i customers. For each pair (i,j) we define $N_{ij}^r(n)$ as the number of type i arrivals actually routed to server j, out of the first n new type i arrivals, and denote

$$H_{ij}^r(t) \doteq N_{ij}^r(F_i^r(t)) - {}^sN_{ij}^r(F_i^r(t)).$$

Here, for any pair (i,j), ${}^sN_{ij}^r(\cdot) \equiv {}^sN_{ij}(\cdot)$ for all r, where ${}^sN_{ij}(\cdot)$ are *nonrandom* functions fixed earlier and satisfying conditions (28)–(30).

We define the *(server workload) regulation* process as follows:

$$Y^r(t) \doteq Y_{\text{idle}}^r(t) + Y_{\text{route}}^r(t), \qquad t \geq 0,$$

where

$$Y_{\text{idle}}^r(t) \doteq \sum_j \alpha_j^* \left(t - \sum_i G_{ij}^r(t) \right),$$

$$Y_{\text{route}}^r(t) \doteq \sum_i Y_{\text{route},\,i}^r(t),$$

$$Y_{\text{route},\,i}^r(t) \doteq \sum_j \alpha_j^* \mu_{ij}^{-1} H_{ij}^r(t).$$

Function Y_{idle}^r is the regulation component due to physical idleness of the servers, Y_{route}^r is the regulation component due to (possible) routing of customers to nonbasic servers, and functions $Y_{\text{route},\,i}^r$ represent the contributions into Y_{route}^r due to different flows $i \in I$.

The regulation component $Y^r_{\text{idle}}(t)$ is clearly nonnegative and nondecreasing. It is easy to verify that each regulation component $Y^r_{\text{route},\,i}(t)$ is also nonnegative and nondecreasing. (Also, so is, therefore, $Y^r_{\text{route}}(t)$.) Moreover, $Y^r_{\text{route},\,i}(t)$ is a piecewise constant function, which is constant between type i customer arrivals, and jumps by the value $\alpha^*_j/\mu_{ij} - \nu^*_i \geq 0$ when a type i arrival is routed to server j; note that the size of the jump is strictly positive if and only if server j is nonbasic for type i. Finally, note that $Y^r(t)$ does not increase over some time interval if and only if, during that interval, none of the servers idles and all new arrivals are routed to the corresponding basic servers.

We record the above facts (along with their obvious generalization) in the following lemma for future reference.

LEMMA 4: *For each value of the scaling parameter r, consider a pair of time points* $0 \leq t^r_1 < t^r_2 < \infty$ *and denote*

$$B^r_0 \doteq \frac{Y^r(t^r_2) - Y^r(t^r_1)}{t^r_2 - t^r_1},$$

$$B^r_{1,\,i} \doteq \sum_{j \notin J_i} \frac{F^r_{ij}(t^r_2) - F^r_{ij}(t^r_1)}{t^r_2 - t^r_1},$$

$$B^r_{2,\,j} \doteq \sum_{i \in I} \frac{G^r_{ij}(t^r_2) - G^r_{ij}(t^r_1)}{t^r_2 - t^r_1}.$$

Then $B^r_0 = 0$ *if and only if* $B^r_{1,\,i} = 0$ *for all* i *and* $B^r_{2,\,j} = 1$ *for all* j. *Also,* $\lim_{r\to\infty} B^r_0 = 0$ *if and only if* $\lim_{r\to\infty} B^r_{1,\,i} = 0$ *for all* i *and* $\lim_{r\to\infty} B^r_{2,\,j} = 1$ *for all* j.

Let us consider the process $Z^r = (U^r, {}^uX^r, F^r, {}^\Sigma V^r, {}^sN^r, N^r, G^r, H^r, Y^r, Y^r_{\text{idle}}, Y^r_{\text{route}})$, where

$$U^r = (U^r_j(t),\, t \geq 0,\, j \in J),$$

$${}^uX^r = ({}^uX^r(t),\, t \geq 0),$$

$$F^r = (F^r_i(t),\, t \geq 0,\, i \in I),$$

$${}^\Sigma V^r = ({}^\Sigma V^r_{ij}(l),\, l \geq 0,\, i \in I, j \in J),$$

$${}^sN^r = ({}^sN^r_{ij}(l),\, l \geq 0,\, i \in I, j \in J),$$

$$N^r = (N^r_{ij}(l),\, l \geq 0,\, i \in I, j \in J),$$

$$G^r = (G^r_{ij}(t),\, t \geq 0,\, i \in I, j \in J),$$

$$H^r = (H^r_{ij}(t),\, t \geq 0,\, i \in I, j \in J),$$

$$Y^r = (Y^r(t),\, t \geq 0),$$

$$Y^r_{\text{idle}} = (Y^r_{\text{idle}}(t),\, t \geq 0),$$

$$Y^r_{\text{route}} = (Y^r_{\text{route}}(t),\, t \geq 0).$$

For each r consider the *fluid scaled* process

$$\Gamma^r Z^r \doteq z^r = (u^r, {}^u x^r, f^r, {}^\Sigma v^r, {}^s n^r, n^r, g^r, h^r, y^r, y^r_{\text{idle}}, y^r_{\text{route}}),$$

where the fluid scaling operator Γ^r is applied componentwise and acts on a scalar function $\Xi = (\Xi(t), t \geq 0)$ as follows:

$$(\Gamma^r \Xi)(t) \doteq \frac{1}{r}\, \Xi(rt).$$

DEFINITION 4: *A fixed set of functions* $z = (u, {}^u x, f, {}^\Sigma v, {}^s n, n, g, h, y, y_{idle}, y_{route})$ *will be called a* fluid sample path *(FSP) if there exists a sequence* $\mathcal{R}_f$ *of values of* r *and a sequence of* sample paths *(of the corresponding* processes*)* $\{z^r\}$ *such that, as* $r \to \infty$ *along the sequence* $\mathcal{R}_f$,

$$z^r \to z, \quad u.o.c.,$$

and, in addition,

$$\|u(0)\| < \infty,$$

$$(f_i^r(t), t \geq 0) \to (\lambda_i t, t \geq 0), \qquad u.o.c., i \in I, \tag{56}$$

$$({}^\Sigma v_{ij}^r(t), t \geq 0) \to (\mu_{ij}^{-1} t, t \geq 0), \qquad u.o.c., i \in I, j \in J. \tag{57}$$

Remark: A sequence $\mathcal{R}_f$, the existence of which is required in Definition 4, may be completely unrelated to the sequence $\mathcal{R}$ we introduced earlier in the definition of the heavy traffic regime.

The following lemma establishes some basic properties of FPSs. We omit the simple proof, which is a direct consequence of the definitions involved.

LEMMA 5: *For any FSP* z*, all of its component functions are Lipschitz continuous and, in addition,*

$$f_i(t) = \lambda_i t, \qquad t \geq 0, i \in I,$$

$${}^\Sigma v_{ij}(t) = \mu_{ij}^{-1} t, \qquad t \geq 0, i \in I, j \in J,$$

$${}^s n_{ij}(t) = (\mu_{ij} \phi_{ij} / \lambda_i) t, \qquad t \geq 0, i \in I, j \in J,$$

$$u_j(t) = u_j(0) + \sum_i \mu_{ij}^{-1} n_{ij}(\lambda_i t) - \sum_i g_{ij}(t), \qquad t \geq 0, j \in J,$$

$${}^u x(t) = \alpha^* \cdot u(t) = {}^u x(0) + y(t), \qquad t \geq 0.$$

Furthermore, both $y(\cdot)$ *and* ${}^u x(\cdot)$ *are nondecreasing (with* $y(0) = 0$*).*

Since all component functions of an FSP are Lipschitz, they are absolutely continuous, and therefore for almost all points $t \in R_+$ (with respect to the Lebesgue measure), the following property holds:

Each component function of z *has (finite) first derivative at* t*, and each function* $n_{ij}(\cdot)$ *has (finite) first derivative at* $\lambda_i t$*.*

We refer to such time points t as *regular*. We adopt a convention that $t = 0$ is *not* a regular point (i.e., in the definition of regular points, we require that proper derivatives exist).

The dynamics of $u(t)$ satisfies the following differential equation and additional conditions at every regular point t:

$$\frac{d}{dt}u(t) = \rho^{\text{in}}(t) - \rho^{\text{out}}(t), \tag{58}$$

where the components of the J-dimensional vectors $\rho^{\text{in}}(t)$ and $\rho^{\text{out}}(t)$ are defined as

$$\rho_j^{\text{in}}(t) \doteq \sum_i \mu_{ij}^{-1}\lambda_i n'_{ij}(\lambda_i t) \in \mathcal{K}, \tag{59}$$

$$\rho_j^{\text{out}}(t) \doteq \sum_i g'_{ij}(t) \in [0,1], \tag{60}$$

and for ρ^{out}, we have

$$\rho_j^{\text{out}}(t) = 1 \quad \text{if } u_j(t) > 0. \tag{61}$$

9.2. Uniform Attraction of Fluid Sample Paths

For $u \in R_+^J$, denote

$$^*A(u) \doteq \max_j C'_j(u_j)/\alpha_j^*, \qquad {}_*A(u) \doteq \min_j C'_j(u_j)/\alpha_j^*;$$

$\Phi(u) \doteq 1 - {}_*A(u)/{}^*A(u)$ if $u \neq 0$, and $\Phi(0) \doteq 0$ by convention.

Consider the following functions associated with a fixed FSP. First, define

$$J^*(t) = \{j \in J | C'_j(u_j(t))/\alpha_j^* = {}^*A(u(t))\}$$

and, similarly, $J_*(t)$ (with $^*A$ replaced by ${}_*A$). Next, introduce

$$^*u_j(t) \doteq \{\zeta \geq 0 | C'_j(\zeta)/\alpha_j^* = {}^*A(u(t))\},$$

and note that $^*u_j(t)$ is well defined since each function $C'_j(\cdot)$ is strictly increasing continuous. Let $^*x(t) \doteq \alpha^* \cdot {}^*u(t)$, where $^*u(t) = ({}^*u_1(t), \ldots, {}^*u_J(t))$, and note that ${}^ux(t) \leq {}^*x(t)$ for all $t \geq 0$.

Finally, note that, at any time t, the following five conditions for $u(t)$ are all equivalent:

1. $u(t)$ is a fixed point.
2. $^*A(u(t)) = {}_*A(u(t))$.
3. $\Phi(u(t)) = 0$.
4. $^*x(t) = {}^ux(t)$.
5. $^*u(t) = u(t)$.

The following sequence of lemmas establishes further properties of fluid sample paths, which are less obvious than the basic properties of Lemma 5. The form of the MinDrift rule is used in the proofs in an essential way.

LEMMA 6: *Consider a fixed FSP z. Suppose $t > 0$ is a regular point and $u(t) \neq 0$. Then the following properties hold at this t:*

(i) $u_j(t) > 0$ *for all* $j \in J$.
(ii) We have

$$\sum_{j \in J^*(t)} \alpha_j^* u_j'(t) \leq 0, \qquad \sum_{j \in J_*(t)} \alpha_j^* u_j'(t) \geq 0. \tag{62}$$

(iii) Moreover, there exists a constant $\epsilon_1 > 0$, *which depends on system parameters only, such that if, in addition,* $u(t)$ *is not a fixed point (i.e.,* ${}^*A(q(t)) > {}_*A(q(t))$*), then*

$$\sum_{j \in J^*(t)} \alpha_j^* u_j'(t) \leq -\epsilon_1, \qquad \sum_{j \in J_*(t)} \alpha_j^* u_j'(t) \geq \epsilon_1. \tag{63}$$

PROOF: Let us first prove (iii). Thus, consider regular time point $t > 0$ and suppose that ${}^*A(u(t)) > {}_*A(u(t))$. The following observation is true:

$$\text{If } i \in I \text{ and } j \in J^*(t) \backslash J_i, \text{ then } n_{ij}'(\lambda_i t) = 0. \tag{64}$$

Indeed, according to the MinDrift rule and Lemma 3(ii)(c), for all sufficiently large r, the prelimit path z^r is such that in a small interval $[t, t+\epsilon]$, $\epsilon > 0$, new arriving customers of type i cannot be routed to a server $j \in J^*(t) \backslash J_i$. This easily implies that the corresponding FSP component $n_{ij}(\cdot)$ cannot increase in a small interval to the right of $\lambda_i t$, and therefore $n_{ij}'(\lambda_i t) = 0$ since t is regular. Using a similar argument, it is also easy to prove the following property:

$$\text{If } i \in I,\ J_i \backslash J^*(t) \neq \emptyset, \text{ and } j \in J^*(t), \text{ then } n_{ij}'(\lambda_i t) = 0. \tag{65}$$

Let us denote by $I^*(t)$ the (nonempty) subset of types i such that $J_i \cap J^*(t) \neq \emptyset$. Since graph $\mathcal{G}(\phi) = \mathcal{G}^*$ is connected, there exists at least one $i \in I^*(t)$ such that $J_i \backslash J^*(t) \neq \emptyset$, in which case (by (65)), we have strict inequality:

$$\sum_{j \in J^*(t)} \alpha_j^* \mu_{ij}^{-1} \lambda_i n_{ij}'(\lambda_i t) = 0 < \sum_{j \in J^*(t)} \alpha_j^* \phi_{ij}. \tag{66}$$

If $i \in I^*(t)$ and $J_i \backslash J^*(t) = \emptyset$ (i.e., $J_i \subseteq J^*(t)$), then

$$\sum_{j \in J^*(t)} \alpha_j^* \mu_{ij}^{-1} \lambda_i n_{ij}'(\lambda_i t) \leq \sum_{j \in J^*(t)} \alpha_j^* \phi_{ij}. \tag{67}$$

Indeed, using (64), the fact that $\alpha_j^* \mu_{ij}^{-1}$ is the same across all $j \in J_i$, $\sum_{j \in J} n'_{ij}(\lambda_i t) \le 1$, and $\sum_{j \in J_i} \mu_{ij} \phi_{ij} / \lambda_i = 1$, we can write

$$\begin{aligned}\sum_{j \in J^*(t)} \alpha_j^* \mu_{ij}^{-1} \lambda_i n'_{ij}(\lambda_i t) &= \sum_{j \in J_i} \alpha_j^* \mu_{ij}^{-1} \lambda_i n'_{ij}(\lambda_i t) \\ &\le \sum_{j \in J_i} \alpha_j^* \mu_{ij}^{-1} \lambda_i \frac{\mu_{ij} \phi_{ij}}{\lambda_i} = \sum_{j \in J_i} \alpha_j^* \phi_{ij} \\ &= \sum_{j \in J^*(t)} \alpha_j^* \phi_{ij}.\end{aligned}$$

As a corollary from (64), we also obtain the following property:

$$\text{If } i \notin I^*(t) \text{ and } j \in J^*(t), \text{ then } n'_{ij}(\lambda_i t) = 0. \tag{68}$$

We will now show that

$$\sum_{j \in J^*(t)} \alpha_j^* u'_j(t) \le -\epsilon, \tag{69}$$

where $\epsilon > 0$ depends only on the subset $J^*(t)$. Indeed,

$$\sum_{j \in J^*(t)} \alpha_j^* u'_j(t) = \sum_{j \in J^*(t)} \alpha_j^* \sum_{i \in I} \mu_{ij}^{-1} \lambda_i n'_{ij}(\lambda_i t) - \sum_{j \in J^*(t)} \alpha_j^*,$$

and (using (68), (66), and (67)) we have

$$\begin{aligned}\sum_{j \in J^*(t)} \alpha_j^* \sum_{i \in I} \mu_{ij}^{-1} \lambda_i n'_{ij}(\lambda_i t) &= \sum_{i \in I^*(t)} \sum_{j \in J^*(t)} \alpha_j^* \mu_{ij}^{-1} \lambda_i n'_{ij}(\lambda_i t) \\ &< \sum_{i \in I^*(t)} \sum_{j \in J^*(t)} \alpha_j^* \phi_{ij} = \sum_{j \in J^*(t)} \alpha_j^* \sum_{i \in I^*(t)} \phi_{ij} \\ &\le \sum_{j \in J^*(t)} \alpha_j^*. \end{aligned} \tag{70}$$

We have proved (69), with $\epsilon > 0$ depending only on the subset $J^*(t) \subset J$. Since there is only a finite number of subsets of J, we have proved the first inequality in (63), with some fixed $\epsilon_1 > 0$.

The second inequality in (63) is proved analogously. We denote by $I_*(t)$ the (nonempty) subset of types i such that $J_i \cap J_*(t) \neq \emptyset$. Then we use the following property (obtained using the argument analogous to that leading to (64) and (65)):

$$\text{If } i \in I_*(t) \text{ and } j \notin J_i \cap J_*(t), \text{ then } n'_{ij}(\lambda_i t) = 0.$$

We omit details.

The proof of the nonstrict inequalities in property (ii) is a straightforward extension of the proof of (iii); namely we need to consider an additional (degenerate)

case when $J^*(t) = J_*(t) = J$. In this case, for example to prove the first inequality in (62), we observe that the nonstrict inequality (67) always applies, and (70) holds with the strict inequality replaced by a nonstrict one.

Finally, (i) is proved by contradiction. Suppose, $u_j(t) = 0$ for some $j \in J$. Obviously, the set of such j is exactly $J_*(t)$. Since $u(t) \neq 0$, $u(t)$ is not a fixed point. Therefore, the second inequality in (63) should hold. However, this is impossible, because we must have $u_j'(t) = 0$ for all $j \in J_*(t)$. Indeed, the condition $u_j(t) = 0$ and the existence of $u_j'(t)$ imply that $u_j'(t) = 0$. (Otherwise, $u_j(\cdot)$ would be negative just before or right after time t.) ■

LEMMA 7: *Consider a fixed FSP. Suppose a time interval* $[t_1, t_2]$, *with* $0 \leq t_1 < t_2$, *is such that*

$$\min_{t_1 \leq t \leq t_2} \min_{j \in j} u_j(t) > 0.$$

Then, over $[t_1, t_2]$, *the functions* ${}^*A(q(t))$, ${}_*A(q(t))$, ${}^*x(t)$, *and* ${}^*u_j(t)$ *for all* $j \in I$ *are Lipschitz continuous. Moreover, for almost all* $t \in [t_1, t_2]$,

$$\frac{d}{dt}\left[{}^*A(u(t))\right] \leq 0, \qquad \frac{d}{dt}\left[{}_*A(u(t))\right] \geq 0, \qquad \frac{d}{dt}\left[{}^*x(t)\right] \leq 0, \tag{71}$$

and if, in addition, ${}^*A(u(t)) > {}_*A(u(t))$ *(i.e.,* $u(t)$ *is not a fixed point), then*

$$\frac{d}{dt}\left[{}^*x(t)\right] \leq -\epsilon_1, \tag{72}$$

where $\epsilon_1 > 0$ *is defined in Lemma 6.*

PROOF: First, the Lipschitz continuity of each function $C_j'(u_i(t))$ in $[t_1, t_2]$ follows from Lipschitz continuity of $u_j(\cdot)$ and the fact that, for the range of possible values of $u_j(t)$ in $[t_1, t_2]$, $C_j''(\cdot)$ is continuous bounded away from both infinity and zero. (This is the only place where we use the assumption that the functions $C_i(\cdot)$ are twice continuously differentiable.)

This implies that for an arbitrary fixed subset $\hat{J} \subseteq J$, the following functions are also Lipschitz continuous in $[t_1, t_2]$:

$$\max_{i \in \hat{J}} C_j'(u_j(t))/\alpha_j^*, \qquad \min_{i \in \hat{I}} C_j'(u_j(t))/\alpha_j^*.$$

In particular, ${}^*A(q(t))$ and ${}_*A(q(t))$ are Lipschitz, which (along with the fact that the second derivatives $C_j''(\cdot)$ are bounded away from zero) implies that all ${}^*u_j(t)$ and ${}^*x(t)$ are Lipschitz. We see that almost all points $t \in [t_1, t_2]$ are regular (as defined earlier) and, in addition, are such that all the max and min functions in the last display, for all (nonempty) subsets $\hat{J} \subseteq J$, have derivatives. Within the present proof, let us call such points *t strictly regular*.

Consider an arbitrary strictly regular point $t \in [t_1, t_2]$. The proof will be complete once we prove (71) and (72) for this point t. Since t is strictly regular, the derivatives $d/dt[{}^*A(u(t))]$ and $d/dt[C_j'(u_j(t))/\alpha_j^*]$ for $j \in J^*(t)$ are all equal. (In particular, this implies that ${}^*u_j'(t) = u_j'(t)$ for all $j \in J^*(t)$.) We cannot have $d/dt[{}^*A(u(t))] > 0$ because this would imply that $u_j'(t) > 0$ for all $j \in J^*(t)$, which would contradict (62). This proves the first (and, with it, the last) inequality in (71). The second inequality in (71) is proved analogously.

We can now write

$$\frac{d}{dt}[{}^*x(t)] = \sum_{j \in J} \alpha_j^* \, {}^*u_j'(t) \le \sum_{j \in J^*(t)} \alpha_j^* \, {}^*u_j'(t) = \sum_{j \in J^*(t)} \alpha_j^* u_j'(t),$$

where the inequality follows from the fact that ${}^*u_j'(t) \le 0$ for all $j \in J$ (which is implied by (71)), and the second equality is because ${}^*u_j'(t) = u_j'(t)$ for $j \in J^*(t)$. In the case ${}^*A(q(t)) > {}_*A(q(t))$, by (63), the RHS of the above display is bounded above by $-\epsilon_1$, which proves (72). ■

LEMMA 8: *Consider a fixed FSP z. Suppose $u(t_1) \neq 0$ for some $t_1 \ge 0$. Then $u(t)$ has all strictly positive components (i.e., $u(t) \in R_{++}^J$) for all $t > t_1$. Moreover, in $[t_1, \infty)$, ${}_*A(q(t))$ is nondecreasing, and both ${}^*A(q(t))$ and ${}^*x(t)$ are nonincreasing.*

PROOF: Indeed, we can always find a regular point $\xi > t_1$ arbitrarily close to t_1 so that $u(\xi) \neq 0$. By Lemma 6, $u(\xi) \in R_{++}^J$. Then, using Lemma 7, it follows that ${}_*A(u(t))$ is nondecreasing (and ${}^*A(u(t))$ and ${}^*x(t)$ are nonincreasing) starting from time ξ, and therefore $u(t) \in R_{++}^J$ for all $t \ge \xi$. Since ξ can be chosen arbitrarily close to t_1, the proof is complete. ■

LEMMA 9: *Consider a fixed FSP z. If $u(0) = 0$, then $u(t) = 0$ for all $t \ge 0$.*

PROOF: Suppose not. By continuity of ${}^*x(\cdot)$, for an arbitrarily $\epsilon > 0$, there exists time $t_1 > 0$ at which ${}^*x(t)$ reaches level ϵ for the first time. Of course, $u(t_1) \neq 0$. By Lemma 8, ${}^*x(t)$ cannot increase starting at time t_1, and therefore ${}^*x(t) \le \epsilon$ for all $t \ge 0$. Since $\epsilon > 0$ can be chosen arbitrarily small, ${}^*x(t) = 0$, and therefore $u(t) = 0$ for all $t \ge 0$. ■

The following theorem easily follows from the lemmas presented earlier in this subsection.

THEOREM 5: *For any fluid sample path, $\Phi(u(t))$ is a nonincreasing function, and the server workload ${}^u x(t)$ is a nondecreasing function. Moreover, there exist fixed constants $T_1 > 0$ and $K \ge 1$ such that, for any FSP, $u(t)$ reaches a fixed point ${}^\circ u$ within finite time ${}^u x(0) T_1$ and then stays there, and $\alpha^* \cdot {}^\circ u \le {}^u x(0) K$.*

PROOF: The fact that $x(t)$ is nondecreasing has already been established earlier.

Suppose $u(0) \neq 0$. By Lemma 8, ${}_*A(u(t))$ is nondecreasing and ${}^*A(u(t))$ is nonincreasing in $[0, \infty)$, and therefore $\Phi(u(t))$ is nonincreasing. Further, by Lemma 8,

$u(t) \in R^J_{++}$ for all $t > 0$. Then, by Lemma 7, for almost all $t > 0$, ${}^*x(t) > {}^u x(t)$ implies

$$ {}^*x'(t) \le -\epsilon_1. $$

Since ${}^*x(0) \le {}^u x(0)[\sum_j \alpha^*_j]/[\min_j \alpha^*_j]$, ${}^u x(t) \le {}^*x(t)$, and ${}^u x(t)$ is nondecreasing, we immediately see that $u(t)$ must reach a fixed point within a time proportional to ${}^u x(0)$.

Therefore, the statement of the theorem, with some fixed $T_1 > 0$ and $K \ge 1$, holds for the FSPs with $u(0) \ne 0$. By Lemma 9, it trivially holds for $u(0) = 0$ as well. ■

For future reference, we record the following property of prelimit paths.

LEMMA 10: *There exists a constant $\epsilon_2 > 0$ such that the following holds. For any prelimit (scaled) path $u^r = (u^r(t), t \ge 0)$, and $0 \le t^r_1 < t^r_2 < \infty$, the property*

$$ u^r(t) \ne 0 \quad \textit{and} \quad \Phi(u^r(t)) \le \epsilon_2, \quad \forall t \in [t^r_1, t^r_2], $$

implies that $y^r(t^r_2) - y^r(t^r_1) = 0$ or, equivalently, that in the (scaled) interval $(t^r_1, t^r_2]$, all new arriving customers are routed to their corresponding basic servers and none of the servers idles.

PROOF: First, since $u^r(t) \ne 0$ in $[t^r_1, t^r_2]$, none of the servers idles in this time interval. Second, a small value of $\Phi(u^r(t))$ implies that the vector $(C'_1(u^r_1(t)), \ldots, C'_J(u^r_1(t)))$ is "almost proportional" to vector α^*. Thus, if $\Phi(u^r(t))$ is small, it follows from the form of the MinDrift(U) rule and Lemma 3(ii)(c) that in $[t^r_1, t^r_2]$, a new arrival of any type $i \in I$ can only be routed to a server $j \in J_i$. We omit the ϵ-δ formalities. ■

10. PROOF OF THEOREM 1

For each $r \in \mathcal{R}$, consider the following process, obtained by diffusion scaling:

$$ \tilde{\Gamma}^r(U^r, {}^u X^r, F^r, {}^\Sigma V^r, {}^s N^r, N^r, G^r, H^r, Y^r, Y^r_{\text{idle}}, Y^r_{\text{route}}) $$

$$ \doteq (\tilde{u}^r, {}^u\tilde{x}^r, \tilde{f}^r, {}^\Sigma\tilde{v}^r, {}^s\tilde{n}^r, \tilde{n}^r, \tilde{g}^r, \tilde{h}^r, \tilde{y}^r, \tilde{y}^r_{\text{idle}}, \tilde{y}^r_{\text{route}}), $$

where the diffusion scaling operator $\tilde{\Gamma}^r$ is defined in (36).

To prove the properties stated in Theorem 1, it will suffice to show that for any subsequence $\mathcal{R}_1 \subseteq \mathcal{R}$ there exists another subsequence $\mathcal{R}_2 \subseteq \mathcal{R}_1$ such that these properties hold when $r \to \infty$ along $\mathcal{R}_2$. As in [14], to do this, we will choose subsequence $\mathcal{R}_2$ and construct all processes (for all $r \in \mathcal{R}_2$) on the same probability space in a way such that the desired properties hold with probability 1 (or are implied by certain probability 1 properties).

Let us fix an arbitrary subsequence $\mathcal{R}_1 \subseteq \mathcal{R}$ of indices $\{r\}$. According to Skorohod's representation theorem (see, for example, [7]), for each i, the sequences

(on r) of the processes $\{F_i^r\}$ and $\{{}^{\Sigma}V_{ij}^r\}$, $j \in J$, can be constructed on a probability space such that the convergence in (31) holds u.o.c. with probability 1 (w.p.1):

$$\left\{ r^{-1}\left(A_i^r(r^2 t) - \lambda_i^r \frac{\sum_j \phi_{ij}\alpha_j^*}{\lambda_i} r^2 t \right), t \geq 0 \right\} \xrightarrow{\text{u.o.c.}} \{{}^{u}\sigma_i B_i(t), t \geq 0\}, \qquad \textbf{(73)}$$

where B_i is a standard Brownian motion.

We can and do assume that our underlying probability space $\Omega = \{\omega\}$ is a direct product of the above I probability spaces. (Without loss of generality, we assume that this probability space is complete.) On this probability space the convergence (33) holds u.o.c. w.p.1 as well:

$$\left\{ r^{-1}\left(\sum_i A_i^r(r^2 t) - \left[\sum_j \alpha_j^* \right] r^2 t \right), t \geq 0 \right\} \xrightarrow{\text{u.o.c.}} \{at + \sigma B(t), t \geq 0\}, \qquad \textbf{(74)}$$

where B is a standard Brownian motion.

Now, from condition (25) and Bramson's weak law estimates ([4, Prop. 4.3]), we know that for any $T_3 > 0$, any $\epsilon > 0$, and any i, for all large r, we have (see the proof of property (5.19) in Proposition 5.1 of [4])

$$P\left\{ \max_{0 \leq l \leq T_3 r} \sup_{0 \leq \xi \leq 1} |f_i^r(l+\xi) - f_i^r(l) - \lambda_i \xi| \geq \epsilon \right\} < \epsilon. \qquad \textbf{(75)}$$

(The max in (75) and (76), as well as in (76)–(78), is over *integers* $l \in [0, T_3 r]$.) Also, using Proposition 4.2 of [4], it is easy to show (similarly to the derivation of property (71) in [14]) that for any $T_3 > 0$, any $\epsilon > 0$, and any pair of (i, j), for all large r, we have

$$P\left\{ \max_{0 \leq l \leq T_3 r} \sup_{0 \leq \xi \leq 1} |{}^{\Sigma}v_{ij}^r(l+\xi) - {}^{\Sigma}v_{ij}^r(l) - \mu_{ij}^{-1}\xi| \geq \epsilon \right\} < \epsilon. \qquad \textbf{(76)}$$

Estimates (75) and (76) enable us to choose a subsequence $\mathcal{R}_2 \subseteq \mathcal{R}_1$, such that as $r \to \infty$ along $\mathcal{R}_2$, with probability 1, for any $T_3 > 0$ we have

$$\max_{0 \leq l \leq T_3 r} \sup_{0 \leq \xi \leq 1} |f_i^r(l+\xi) - f_i^r(l) - \lambda_i \xi| \to 0, \qquad i \in I, \qquad \textbf{(77)}$$

and

$$\max_{0 \leq l \leq T_3 r} \sup_{0 \leq \xi \leq 1} |{}^{\Sigma}v_{ij}^r(l+\xi) - {}^{\Sigma}v_{ij}^r(l) - \mu_{ij}^{-1}\xi| \to 0, \qquad i \in I, j \in J. \qquad \textbf{(78)}$$

Properties (77) and (78), in turn, imply the following property.

With probability 1, *for any fixed* $T_4 > 0$ *and* $d > 0$, *for any* (i, j), *we have the following:*

Uniformly on any sequence of pairs (t_1^r, t_2^r), $r \in \mathcal{R}_2$, *such that* $0 \le t_1^r < t_2^r \le r^2 T_4$, $t_2^r - t_1^r \ge rd$,

$$\lim_{r\to\infty,\, r\in\mathcal{R}_2} \frac{{}^{\Sigma}V_{ij}^r({}^{s}N_{ij}^r(F_i^r(r^2 t_2))) - {}^{\Sigma}V_{ij}^r({}^{s}N_{ij}^r(F_i^r(r^2 t_1)))}{\phi_{ij}(t_2^r - t_1^r)} = 1; \tag{79}$$

Uniformly on any sequence of pairs (l_1^r, l_2^r), $r \in \mathcal{R}_2$, *such that* $0 \le l_1^r < l_2^r \le r^2 T_4$, $l_2^r - l_1^r \ge rd$,

$$\lim_{r\to\infty,\, r\in\mathcal{R}_2} \frac{{}^{\Sigma}V_{ij}^r(l_2^r) - {}^{\Sigma}V_{ij}^r(l_1^r)}{\mu_{ij}^{-1}(l_2^r - l_1^r)} = 1. \tag{80}$$

For each $j \in J$, we have

$$U_j^r(r^2 t) = U_j^r(0) + \sum_i {}^{\Sigma}V_{ij}^r(N_{ij}^r(F_i^r(r^2 t))) - \sum_i G_{ij}^r(r^2 t),$$

and therefore the expression for the scaled server workload can be written as

$$\begin{aligned}
&{}^{u}\tilde{x}^r(t) && (81)\\
&= {}^{u}\tilde{x}^r(0) + r^{-1}\left[\sum_j \alpha_j^* \sum_i {}^{\Sigma}V_{ij}^r({}^{s}N_{ij}^r(F_i^r(r^2 t))) - \sum_j \alpha_j^* r^2 t\right] && (82)\\
&\quad + r^{-1}\sum_j \alpha_j^* \left(r^2 t - \sum_i G_{ij}^r(r^2 t)\right) && (83)\\
&\quad + r^{-1}\sum_j \sum_i \alpha_j^* [{}^{\Sigma}V_{ij}^r(N_{ij}^r(F_i^r(r^2 t))) - {}^{\Sigma}V_{ij}^r({}^{s}N_{ij}^r(F_i^r(r^2 t)))] && (84)\\
&= \widetilde{w}^r(t) + \tilde{y}_{\text{idle}}^r(t) + \check{y}_{\text{route}}^r(t) = \widetilde{w}^r(t) + \check{y}^r(t), && (85)
\end{aligned}$$

where $\widetilde{w}^r(t)$ is the term (82), $\tilde{y}_{\text{idle}}^r(t)$ denotes the term (83), $\check{y}_{\text{route}}^r(t)$ denotes the term (84), and $\check{y}^r(t) \doteq \tilde{y}_{\text{idle}}^r(t) + \check{y}_{\text{route}}^r(t)$. We know from (74) that

$$(\widetilde{w}^r(t), t \ge 0) \xrightarrow{\text{u.o.c.}} (\widetilde{w}(t), t \ge 0),$$

where

$$\widetilde{w}(t) \doteq \widetilde{w}(0) + at + \sigma B(t),$$

$B(\cdot)$ is the realization of a standard Brownian motion, and the parameters a and σ are those defined in (34). (The realization $\widetilde{w}(\cdot)$ is, of course, continuous.) As seen from (85), the key step in proving Theorem 1 will be the proof of the following convergence:

$$(\check{y}^r(t), t \ge 0) \xrightarrow{\text{u.o.c.}} (\check{y}(t), t \ge 0). \tag{86}$$

In the rest of this section, we restrict ourselves to a (measurable, probability 1) subset $\Omega_2 \subseteq \Omega$ of elementary outcomes ω, such that all the specified above probability 1 properties hold, when $r \to \infty$ along $\mathcal{R}_2$.

LEMMA 11: *Consider a fixed $\omega \in \Omega_2$. As $r \to \infty$ along $\mathcal{R}_2$, the functions $\check{y}^r_{\text{route}}$ and $\tilde{y}^r_{\text{route}}$, and then $\check{y}^r$ and $\tilde{y}^r$, are "asymptotically close" in the following sense. For any fixed $T_4 > 0$ and any fixed $\delta_1 > 0$ and $\delta_2 > 0$, for all sufficiently large r, uniformly on $t \in [0, T_4]$,*

$$(1-\delta_1)\tilde{y}^r_{\text{route}}(t) - \delta_2 \le \check{y}^r_{\text{route}}(t) \le (1+\delta_1)\tilde{y}^r_{\text{route}}(t) + \delta_2 \tag{87}$$

and then

$$(1-\delta_1)\tilde{y}^r(t) - \delta_2 \le \check{y}^r(t) \le (1+\delta_1)\tilde{y}^r(t) + \delta_2. \tag{88}$$

The proof of Lemma 11 is analogous to the proof of Lemma 9 in [12]. The key observation here (which follows from (79) and (80)) is that, for fixed $T_4 > 0$ and (arbitrarily small) $d > 0$, if $t \in [0, T_4]$ and $|H^r_{ij}(F^r_i(r^2 t))| \ge rd$, then for all sufficiently large r, the ratio

$$\frac{{}^{\Sigma}V^r_{ij}(N^r_{ij}(F^r_i(r^2 t))) - {}^{\Sigma}V^r_{ij}({}^{s}N^r_{ij}(F^r_i(r^2 t)))}{\mu_{ij}^{-1} H^r_{ij}(F^r_i(r^2 t))}$$

is close to unity. We do not present the details. We note that (analogous to the situation with Lemma 9 in [12]) Lemma 11 applies to any service discipline satisfying condition (d0), and the uniqueness of ϕ (in the CRP condition) is used in the proof of Lemma 11 in an essential way.

It follows from Lemma 11 that to prove (86), it suffices to prove

$$(\tilde{y}^r(t),\, t \ge 0) \xrightarrow{\text{u.o.c.}} (\tilde{y}(t),\, t \ge 0), \tag{89}$$

because $\tilde{y}(\cdot)$ is bounded on finite intervals.

Since regulation $\tilde{y}^r$ is a nondecreasing function (for any r), for any fixed $\omega \in \Omega_2$, from any subsequence $\mathcal{R}_3(\omega) \subseteq \mathcal{R}_2$ (which may depend on ω!), it is always possible to find a further subsequence $\mathcal{R}_4(\omega) \subseteq \mathcal{R}_3(\omega)$ such that

$$\tilde{y}^r \Rightarrow \tilde{y}, \tag{90}$$

where $\tilde{y}$ is some nondecreasing RCLL function. (We will prove that this limit $\tilde{y}$ is indeed the regulation of the one-dimensional Brownian motion defined earlier.) In principle, $\tilde{y}$ may take the values $+\infty$. (In other words, $\tilde{y} \in D([0,\infty), \bar{R}$.) Recall that the notation "$\Rightarrow$" stands for convergence at every point of continuity of the limit function except maybe the point 0.) We note that (90) implies that

$${}^{u}\tilde{x}^r \Rightarrow \tilde{x} \doteq \tilde{w} + \tilde{y}, \tag{91}$$

and therefore $\tilde{x}(t) < \infty$ if and only if $\tilde{y}(t) < \infty$.

The following lemma (and its proof) is analogous to Lemma 7 in [14] and Lemma 10 in [12]; it contains key observations used in the proof of Theorem 1. The key construction of the proof, which involves "slowing down" the diffusion scaled process to consider a family of processes on the "fluid" time scale and then expoiting the uniform attraction property of fluid sample paths, is essentially the same as

that in Section 5 of [4]. This construction is central for establishing SSC in the heavy traffic asymptotic regime for multiclass queuing networks (see [4,19]).

LEMMA 12: *Suppose that the service discipline is such that the routing rule is Min-Drift(U) and scheduling rules at the servers are work-conserving. Suppose that* $\omega \in \Omega_2$ *and a subsequence* $\mathcal{R}_4(\omega) \subseteq \mathcal{R}_2$ *are fixed such that, along this subsequence, (90) holds. Suppose, a sequence* $\{\tilde{t}^r, r \in \mathcal{R}_4(\omega)\}$ *is fixed such that*

$$\tilde{t}^r \to t' \geq 0$$

and

$${}^u\tilde{x}^r(\tilde{t}^r) \to C > 0.$$

Let $\delta > 0$ *be fixed and*

$$\epsilon \doteq \sup_{\xi_1, \xi_2 \in [t'-3\delta, t'+3\delta] \cap R_+} |\widetilde{w}(\xi_1) - \widetilde{w}(\xi_2)| < C.$$

Then the following hold:

(a) $\tilde{y}$ *(and* $\tilde{x}$*) is finite in* $[0, t' + \delta]$.
(b) $\tilde{y}$ *does not increase in* $(t', t' + \delta]$ *(i.e.,* $\tilde{y}(t' + \delta) - \tilde{y}(t') = 0$*).*
(c) The following bound holds

$$C - \epsilon \leq \tilde{x}(t) \leq CK + \epsilon, \quad \forall t \in [t', t' + \delta],$$

with K defined in Theorem 5.
(d) For any $\delta' > 0$,

$$(\tilde{u}^r(t), t \in [t' + \delta', t' + \delta]) \xrightarrow{u.o.c.} (\tilde{u}(t), t \in [t' + \delta', t' + \delta]),$$

where $\tilde{u}(t)$ *is the (unique) fixed point such that* $\alpha^* \cdot \tilde{u}(t) = \tilde{x}(t)$.

If, in addition, $\tilde{t}^r = t'$ *for all r, and* $\tilde{u}^r(t') \to {}^\circ\tilde{u}$*, where* ${}^\circ\tilde{u}$ *is a fixed point (necessarily, with* $\alpha^* \cdot {}^\circ\tilde{u} = C$*), then the following hold:*

(c') $\tilde{x}(t') = C$ *and, consequently,* $\tilde{u}(t') = {}^\circ\tilde{u}$.
(d') $(\tilde{u}^r(t), t \in [t', t' + \delta]) \xrightarrow{u.o.c.} (\tilde{u}(t), t \in [t', t' + \delta])$.

PROOF: The proof essentially repeats that of Lemma 10 in [12]. For completeness, and since some adjustments are required, we present it here.

Let us consider the functions of interest on the fluid time scale; namely consider earlier defined functions ${}^ux^r(t) \equiv {}^u\tilde{x}^r(t/r)$, $y^r(t) \equiv \tilde{y}^r(t/r)$, $t \geq 0$, and similarly defined function w^r and other related ones.

Let us choose a fixed $T > 0$ as follows. Let us fix $\epsilon_3 \in (0, C - \epsilon)$, denote

$$C_3 = (C + \epsilon_3)K + \epsilon + \epsilon_3,$$

and fix arbitrary

$$T \geq C_3 T_1,$$

where K and T_1 are the constants defined in Theorem 5. As seen later in the proof, C_3 will be the upper bound of ${}^u\bar{x}^r(\cdot)$ in the interval $[\tilde{t}^r, \tilde{t}^r + \delta]$ or, equivalently, the upper bound of ${}^ux^r(\cdot)$ in the interval $[r\tilde{t}^r, r\tilde{t}^r + r\delta]$. Thus, the choice of the constant T is such that an FSP with initial server workload not exceeding C_3 will converge to a fixed point within time T.

For each integer $l \in [0, 2\delta r/T]$, consider

$${}^u\bar{x}^{r,l}(\xi) \doteq {}^ux^r(r\tilde{t}^r + Tl + \xi), \qquad \xi \geq 0,$$

and similarly defined $\bar{w}^{r,l}$, $\bar{y}^{r,l}$, and other related functions.

Let us fix arbitrary $\epsilon_4 \in (0, \epsilon_2/2)$, where ϵ_2 is defined in Lemma 10. Then the following property holds.

PROPERTY 1: *For all sufficiently large r, relation (92) below holds for all integer* $l \in [0, 2\delta r/T]$, *and relations (93)–(95) hold for all integer* $l \in [1, 2\delta r/T]$*:*

$$C - \epsilon - \epsilon_3 \leq {}^u\bar{x}^{r,l}(\xi) \leq C_3, \qquad \forall \xi \in [0, T], \tag{92}$$

$$\Phi(\bar{u}^{r,l}(\xi)) \leq \epsilon_4 \quad \text{for } \xi = 0 \text{ and } \xi = T, \tag{93}$$

$$\Phi(\bar{u}^{r,l}(\xi)) \leq 2\epsilon_4, \quad \forall \xi \in [0, T], \tag{94}$$

$$\bar{y}^{r,l}(T) - \bar{y}^{r,l}(0) = 0. \tag{95}$$

To prove Property 1, we first observe that (92) must hold for $l = 0$ for all large r, because otherwise we would be able to choose a subsequence of indices r along which the sequence of paths $\bar{z}^{r,0}$ converges to an FSP z with ${}^u\bar{x}(0) = C$ and either ${}^u\bar{x}(\xi) > CK$ or ${}^u\bar{x}(\xi) < C$ for some $\xi \in [0, T]$, which contradicts Theorem 5. Moreover, this observation shows that in fact for all large r,

$$C - \epsilon_3/2 \leq {}^u\bar{x}^{r,1}(0) \leq (C + \epsilon_3/2)K, \tag{96}$$

and, given our choice of the constant T,

$$\Phi(\bar{u}^{r,1}(0)) \leq \epsilon_4. \tag{97}$$

Next, suppose Property 1 does not hold. Then we can choose an (infinite) subsequence of r such that (along this subsequence) $l' = l'(r)$ is well defined as the smallest $l \geq 1$ such that one of the conditions (92)–(95) does not hold. (Note that, by this construction and (97), property (93) always holds for $l = l'$ and $\xi = 0$.) We will show that this construction leads to a contradiction.

Indeed, for all large r, both (92) and (93) hold for $l = l'$ and $\xi = 0$. This follows from the combination of the following facts:

1. Property (96).
2. $|\widetilde{w}(\xi_1) - \widetilde{w}(\xi_2)| \le \epsilon$ as long as $\xi_1, \xi_2 \in [t' - 3\delta, t' + 3\delta] \cap R_+$.
3. $\widetilde{w}^r \to \widetilde{w}$ uniformly in $[t' - 3\delta, t' + 3\delta] \cap R_+$.
4. Property (95) for each $1 \le l \le l' - 1$.
5. The functions $\tilde{y}^r$ and $\check{y}^r$ are asymptotically close (in the sense of Lemma 11).

Since (92) and (93), with $l = l'$ and $\xi = 0$, hold for large r, we see that (93) and (94) hold for $l = l'$ and all large r. (Otherwise, we would be able to choose a subsequence of r along which $\bar{z}^{r,l'}$ converges to an FSP z, violating Theorem 5.) Similarly, the lower bound in (92) must hold (for large r) for $l = l'$ and $\xi \in [0, T]$. This, in conjunction with (94) and Lemma 10, means that (95) holds for $l = l'$ (for large r). Finally, this and the argument we already used to prove bound (92) for $l = l'$, $\xi = 0$, shows that in fact (92) holds for $l = l'$ and all $\xi \in [0, T]$ (for large r). We have proved that (92)–(95) hold for $l = l'(r)$ for all large r. This is a contradiction with the construction of the function $l' = l'(r)$, which proves Property 1.

Property 1 (namely (92)) implies that, for all large r,

$$C - \epsilon - \epsilon_3 \le {}^u\bar{x}^{r,l}(\xi) \le C_3, \qquad \xi \in [0, T],\ 0 \le l \le 2\delta r/T.$$

Statements (a)–(c) of the lemma follow from this estimate.

To prove (d), we first notice that (a), (b), and Lemma 11 imply the following uniform convergence for the workload process:

$$({}^u\tilde{x}^r(t),\ t \in [t' + \delta', t' + \delta]) \xrightarrow{\text{u.o.c.}} (\tilde{x}(t),\ t \in [t' + \delta', t' + \delta]). \qquad \textbf{(98)}$$

Statement (d) then follows from Property 1, the fact that ϵ_4 can be chosen arbitrarily small, and convergence (98).

To prove properties (c′) and (d′), we use the same exact construction. It is easy to see that, under the additional assumptions, all conditions (92)–(95) in Property 1 hold for all integer $l \in [0, 2\delta r/T]$ (including zero). Given this, properties (c′) and (d′) are proved analogously to properties (c) and (d). We omit details. ■

The rest of the proof of Theorem 1 repeats that of Theorem 1 in [14], and that of Theorem 1 in [12], virtually verbatim. We reproduce it here, with the necessary minor adjustments, for completeness.

10.1. Proof of Theorem 1(i)

To prove this part it suffices to prove the following:

PROPERTY 2: *As $r \to \infty$ (along $\mathcal{R}_2$), for any $\omega \in \Omega_2$ (i.e., with probability 1), we have the following convergence:*

$$(\tilde{y}^r(t),\ t \ge 0) \xrightarrow{u.o.c.} (\tilde{y}(t),\ t \ge 0), \qquad \textbf{(99)}$$

where $\tilde{y}$ is defined by (39), and

$$(\tilde{u}^r(t), t \geq 0) \xrightarrow{u.o.c.} (\tilde{u}(t), t \geq 0), \tag{100}$$

where for each t, $\tilde{u}(t)$ is the fixed point such that $\alpha^* \cdot \tilde{u}(t) = \tilde{x}(t)$.

PROOF OF PROPERTY 2: Let us fix $\omega \in \Omega_2$. As explained earlier, for an arbitrary subsequence $\mathcal{R}_3(\omega) \subseteq \mathcal{R}_2$ there exists another subsequence $\mathcal{R}_4(\omega) \subseteq \mathcal{R}_3(\omega)$ such that the convergence (90) holds along this subsequence. Then, the proof of Property 2 will be complete if we can prove the following statements (for the chosen ω, with $r \to \infty$ along $\mathcal{R}_4(\omega)$). We recall that, at this point, the function $\tilde{y}$ is just *some* limit function—the fact that it is equal to the function defined by (39) is what needs to be proved in order to establish (100).

Step 1. The limit function $\tilde{y}$ is finite everywhere in $[0,\infty)$.
Step 2. The function $\tilde{y}$ is continuous, and $\tilde{y}(0) = 0$.
Step 3. If $\tilde{x}(t) > 0$, then t is not a point of increase of $\tilde{y}$.
Step 4. The function $\tilde{y}$, defined above as a limit, satisfies (39).
Step 5. Convergence (100) holds.

In this proof, we will use the convention that $\tilde{y}(0-) = 0$, $\tilde{w}(0-) = \tilde{x}(0-) = \tilde{w}(0)$. So, the case $\tilde{y}(0) > 0$ will be viewed as a discontinuity of $\tilde{y}$ (and $\tilde{x}$) at 0. Also, we will use the notation

$$\epsilon(\delta, t) \doteq \sup_{\xi_1, \xi_2 \in [t-\delta,\, t+\delta] \cap R_+} |\tilde{w}(\xi_1) - \tilde{w}(\xi_2)|.$$

Proof of Step 1. Suppose the statement does not hold. Denote $t^* = \inf\{t \geq 0 | \tilde{y}(t) = \infty\}$. The inf is attained because $\tilde{y}$ is RCLL.

We choose a small δ such that $\delta \in (0, t^*)$ if $t^* > 0$, and arbitrary $\delta > 0$ if $t^* = 0$. Let us fix $\epsilon = \epsilon(4\delta, t^*)$. Then we choose a small $\Delta t \in (0, \delta)$ and a large C such that $C > \tilde{x}(t^* - \Delta t) + \epsilon$ if $t^* > 0$, and $C > \tilde{x}(0-) + \epsilon$ if $t^* = 0$. We define

$$\tilde{t}^r = \min\{t \geq (t^* - \Delta t) \vee 0 | {}^u\tilde{x}^r(t) \geq C\}$$

and choose a further subsequence of $\{r\}$ such that

$$\tilde{t}^r \to t' \in [t^* - \Delta t, t^*].$$

(We must have $t' \leq t^*$, because the *limit* function $\tilde{y}(t)$, and therefore $\tilde{x}(t)$, is infinite for all $t \geq t^*$.) It is also easy to see (from (77)) that

$${}^u\tilde{x}^r(t) \to C.$$

The conditions of Lemma 12 are satisfied, and so $\tilde{y}$ is bounded in $[t', t' + \delta]$—a contradiction, since $t' + \delta > t^*$. Step 1 has been proved.

Proof of Step 2. Suppose that the statement does not hold. The contradiction is obtained very similarly to the way it is done in the proof of Step 1. Let t^* be a

discontinuity point (the case $t^* = 0$ is included) (i.e., $\tilde{y}(t^*-) < \tilde{y}(t^*)$). Since $\tilde{x} = \widetilde{w} + \tilde{y}$ and $\widetilde{w}$ is continuous, $\tilde{x}(t^*) - \tilde{x}(t^*-) = \tilde{y}(t^*) - \tilde{y}(t^*-)$. There are two possible cases:

(a) $\tilde{x}(t^*-) > 0$.
(b) $\tilde{x}(t^*-) = 0$.

Case (a). In this case, we must have $t^* > 0$. (Indeed, by the definition of $\widetilde{w}$ and our conventions, $\tilde{x}(0-) = \widetilde{w}(0) = \lim_r {}^u\tilde{x}^r(0)$. If $\widetilde{w}(0) > 0$, then, by Lemma 12(c′), $\tilde{x}(0) = \lim_r {}^u\tilde{x}^r(0)$, which means that $\tilde{x}$, and therefore $\tilde{y}$, has no jump at 0. If $\widetilde{w}(0) = 0$, then $\tilde{x}(0-) = 0$.) We can always fix a small $\delta > 0$ and small $\Delta t \in (0, \delta)$, such that $t' = t^* - \Delta t$ is a point of continuity of $\tilde{y}$ (and $\tilde{x}$) and $\epsilon = \epsilon(4\delta, t^*) < \tilde{x}(t') = C$. We have convergence ${}^u\tilde{x}^r(t') \to C$ (since $\tilde{x}$ is continuous at t'), and by Lemma 12, $\tilde{y}$ cannot increase in the interval $(t', t' + \delta]$ which contains t_*. So, $\tilde{x}$ cannot have a jump at t_*.

Case (b). In this case, let us fix a small $C > 0$ and then a sufficiently small $\delta > 0$ so that

$$C_1 = KC + \epsilon < \tilde{x}(t^*),$$

where $\epsilon = \epsilon(4\delta, t^*)$ and $K \geq 1$ is defined in Theorem 5 (and used in Lemma 12). Then if $t^* > 0$, we fix a small Δt such that

$$\limsup_{r\to\infty} \sup_{[t^*-\Delta t, t^*]} {}^u\tilde{x}^r(\xi) < C.$$

If $t^* = 0$, we fix an arbitrary $\Delta t > 0$. We define

$$\tilde{t}^r = \min\{t \geq (t^* - \Delta t) \vee 0 \,|\, {}^u\tilde{x}^r(t) \geq C\},$$

and choose a further subsequence of $\{r\}$ such that

$$\tilde{t}^r \to t' \in [(t^* - \Delta t) \vee 0, t^*].$$

The conditions of Lemma 12 are satisfied, and so $\tilde{x}(t) < C_1$ for all $t \in [t', t' + \delta]$, which contradicts the assumption of case (b), since t^* belongs to the latter interval. Step 2 has been proved.

Proof of Step 3. Let $t^* \geq 0$ be such that $\tilde{x}(t^*) > 0$. If $t^* = 0$, then the fact that $\tilde{y}$ does not increase in a small interval $[0, \delta]$ follows from Lemma 12(b′). If $t^* > 0$, then precisely the same construction as in the proof of Step 2(a) shows that $\tilde{y}$ does not increase in a small interval $[t', t' + \delta]$ containing t^* in its interior. Step 3 has been proved.

Proof of Step 4. Follows from the statements of Steps 2 and 3 and Proposition 1 (in the Appendix).

Proof of Step 5. It suffices to show the following:

For any $t^* \geq 0$ and any $\epsilon > 0$, there exists $\delta > 0$ such that

$$\limsup_{r\to\infty} \sup_{\xi\in[t^*-\delta,\, t^*+\delta]\cap R_+} \|\tilde{u}^r(\xi) - \tilde{u}(\xi)\| < \epsilon. \tag{101}$$

(The u.o.c. convergence will then follow from the Heine–Borel lemma.)

If $\tilde{x}(t^*) = 0$, then (101) must hold because both functions $\tilde{u}$ and $\tilde{u}^r$ (for large r) are bounded by an arbitrarily small constant in a sufficiently small neighborhood of t^*. If $\tilde{x}(t^*) > 0$ and $t^* = 0$, then (101) follows from Lemma 12(d′). If $\tilde{x}(t^*) > 0$ and $t^* > 0$, then to obtain (101) we can repeat the construction of the proof of Step 2(a) and then apply Lemma 12(d). Step 5 has been proved.

Thus, the proof of Property 2, and with it the proof of statement (i) of the theorem, is complete. ■

10.2. Proof of Theorem 1(ii)

We use the same construction of the probability space Ω, the subsequence $\mathcal{R}_2$, and the probability 1 subset Ω_2, as specified earlier. Consider an arbitrary discipline G. Sample paths for both the G$c\mu$ and G disciplines are constructed on this common probability space. For $\omega \in \Omega_2$, consider paths of ${}^u\tilde{x}^r_G$, $\tilde{y}^r_G$, and $\widetilde{w}^r_G$, corresponding to the discipline G. Since $\widetilde{w}^r_G$ is invariant with respect to the discipline, $\widetilde{w}^r_G = \widetilde{w}^r$, and therefore $\widetilde{w}^r_G \to \widetilde{w}_G = \widetilde{w}$ u.o.c.

We claim that, along the subsequence $\mathcal{R}_2$, for any $t \geq 0$,

$$\liminf_{r\to\infty} \inf_{\xi\in[0,\,t]} [{}^u\tilde{x}^r_G(\xi) - \tilde{x}(\xi)] \geq 0, \tag{102}$$

and therefore (40) holds. To prove this, we first recall that Lemma 11 holds for any discipline G satisfying condition $(d0)$.

For any subsequence $\mathcal{R}_3(\omega) \subseteq \mathcal{R}_2(\omega)$, we can choose a further subsequence $\mathcal{R}_4(\omega) \subseteq \mathcal{R}_3(\omega)$ such that $\tilde{y}^r_G \Rightarrow \tilde{y}_G$, where $\tilde{y}_G$ is some nondecreasing nonnegative RCLL function. (The case that $\tilde{y}_G(t)$ takes value $+\infty$ starting from some finite time t_* is possible.) Therefore, for any $t \geq 0$ where $\tilde{y}_G(\cdot)$ is continuous, as $r \to \infty$ along $\mathcal{R}_4(\omega)$,

$$\lim {}^u\tilde{x}^r_G(t) = \widetilde{w}(t) + \tilde{y}_G(t).$$

Since ${}^u\tilde{x}^r_G(t)$ is nonnegative, we see that $\widetilde{w} + \tilde{y}_G$ is nonnegative at every point of continuity of $\tilde{y}_G$, and therefore it is nonnegative for all $t \geq 0$ (by right-continuity). Then, by Proposition 1(ii) (in the Appendix), $\tilde{y}_G(t) \geq \tilde{y}(t)$ for all $t \geq 0$. Since $\tilde{y}$ is continuous and nondecreasing and $\tilde{y}_G$ and all $\tilde{y}^r_G$ are nondecreasing, for any $t \geq 0$ we obtain the uniform bound

$$\liminf_{r\to\infty} \inf_{\xi\in[0,\,t]} [\tilde{y}^r_G(\xi) - \tilde{y}(\xi)] \geq 0.$$

By Lemma 11, we have an analogous bound for $\check{y}^r_G$ as well:

$$\liminf_{r\to\infty} \inf_{\xi\in[0,\,t]} [\check{y}^r_G(\xi) - \tilde{y}(\xi)] \geq 0,$$

which proves (102), with $r \to \infty$ along subsequence $\mathcal{R}_4(\omega)$, and therefore along $\mathcal{R}_2$ as well (since the subsequence $\mathcal{R}_3(\omega)$ can be arbitrary). The proof of (102) (and therefore (40)) is complete.

Since the function $\sum_j C_j(u_j)$ is continuous in the vector u, and the fixed point $\tilde{u}(t)$ in (41) minimizes the value of $\sum_j C_j(u_j)$ over vectors u with server workload ${}^u\tilde{x}(t)$, property (41) also holds. Finally, the equality in (42) follows from the fact that $\tilde{u}^r \to \tilde{u}$ u.o.c., and the inequality follows from (41) and Fatou's lemma.

The proof of Theorem 1 is now complete.

11. PROOF OF THEOREM 2

The proof of Theorem 2 is a relatively straightforward extension of that of Theorem 1. The extension is based on the fact that, given the assumptions of Theorem 2(ii), the processes ${}^q\tilde{u}^r$ and $\tilde{u}^r$ are in fact "asymptotically close" (see (103)). As a result, the behavior of the system (in the diffusion limit) under MinDrift(Q) is the "same" as that under MinDrift(U). In this section, we provide a detailed sketch of such a proof extension. We believe the details can be easily filled in by a reader.

Construction of the probability space and subsequences $\mathcal{R}_1$ and $\mathcal{R}_2$. For the proof of Theorem 2, we assume that, for each r, the service times of the "initial customers" of type i at server j, whose service has not yet started at initial time 0, are given by an i.i.d. sequence $\bar{v}^r_{ij}(n)$, $n = 1,2,\ldots$. Thus, the sequence $\{\bar{v}^r_{ij}(n)\}$ is separate from the sequence $\{v^r_{ij}(n)\}$, defining service times of customers arriving after time 0, but, of course, $\bar{v}^r_{ij}(1)$ has the same distribution as $v^r_{ij}(1)$. We denote by

$${}^{\Sigma}\bar{V}^r_{ij}(n) \doteq \sum_{m=1}^{n} \bar{v}^r_{ij}(m), \qquad n = 0,1,2,\ldots,$$

the total amount of unfinished work "contained" in the the first n (in the order of them being taken for service) initial type i customers at the server j. As with other functions, we extend the domain of ${}^{\Sigma}\bar{V}^r_{ij}(\cdot)$ to all real nonnegative $t \geq 0$ and denote its fluid-scaled version by ${}^{\Sigma}\bar{v}^r_{ij} = \Gamma^r\, {}^{\Sigma}\bar{V}^r_{ij}$.

The underlying probability space is the same as in the proof of Theorem 1, except it is augmented by taking a direct product with the space on which the sequences (on r) of the processes $\{{}^{\Sigma}V^r_{ij}\}$ are defined. The subsequence $\mathcal{R}_1$ is defined exactly the same way. The property analogous to (78) holds for the processes ${}^{\Sigma}\bar{v}^r_{ij}$, as well as ${}^{\Sigma}v^r_{ij}$. Then, the subsequence $\mathcal{R}_2$ can be chosen in a way such that, additionally, the properties analogous to (78) and (80) hold for the processes ${}^{\Sigma}\bar{v}^r_{ij}$ and ${}^{\Sigma}\bar{V}^r_{ij}$, respectively.

"Asymptotic closeness" of ${}^{q}\tilde{u}^{r}$ and $\tilde{u}^{r}$. Using (78) and (80), and their analogs for ${}^{\Sigma}\bar{v}_{ij}^{r}$ and ${}^{\Sigma}\bar{V}_{ij}^{r}$, it is easy to demonstrate the following property, which holds for any service discipline within the class specified in Theorem 2(ii):

As $r \to \infty$ along $\mathcal{R}_2$, with probability 1, for any $T_3 > 0$ and any $\epsilon > 0$, we have

$$\sup_{0 \le t \le T_3} \frac{\|{}^{q}\tilde{u}^{r}(t) - \tilde{u}^{r}(t)\|}{\max(\tilde{u}^{r}(t), \epsilon)} \to 0. \tag{103}$$

This property is the key in showing that, in the heavy traffic limit, MinDrift(Q) induces the same system behavior as the MinDrift(U).

Definition and properties of the FSPs under MinDrift(Q). The process Z^r is augmented by the following components:

$$\begin{aligned} Q^r &= (Q_{ij}^r(t),\, t \ge 0,\, i \in I, j \in J), \\ {}^{\Sigma}\bar{V}^r &= ({}^{\Sigma}\bar{V}_{ij}^r(l),\, l \ge 0,\, i \in I, j \in J), \\ {}^{q}U^r &= ({}^{q}U_j^r(t),\, t \ge 0, j \in J), \\ {}^{q}X^r &= ({}^{q}X^r(t),\, t \ge 0). \end{aligned}$$

The fluid-scaled process z^r and an FSP z are augmented by the corresponding components q^r, ${}^{\Sigma}\bar{v}^r$, ${}^{q}u^r$, ${}^{q}x^r$, and q, ${}^{\Sigma}\bar{v}$, ${}^{q}u$, ${}^{q}x$, respectively. The definition of the FSP is the same, except it includes the additional conditions

$${}^{q}u(0) = u(0)$$

and an analog of (57) for ${}^{\Sigma}v_{ij}^r$. This augmented definition of an FSP easily yields the following additional FSP property (which can be added into Lemma 5):

$${}^{q}u(t) = u(t), \quad \forall t \ge 0, \tag{104}$$

which, of course, also implies ${}^{q}x(t) = {}^{u}x(t)$, $t \ge 0$. Using (104), it can be easily shown that *all* of the FSP properties established for MinDrift(U) hold for MinDrift(Q) as well.

Proof of Theorem 2. Given property (103) and the fact that FSPs under MinDrift(Q) satisfy all of the properties of FSPs under MinDrift(U) (plus (104)), the rest of the proof is the same as that of Theorem 1.

12. PROOF OF THEOREM 4

Before proceeding with the proof, note that, in addition to (51), we have another pathwise relation (see (21) and (22)):

$${}^{q}\tilde{x}^{r}(t) \le C_0 \tilde{x}^{r}(t), \qquad t \ge 0. \tag{105}$$

PROOF: The argument leading to Theorem 3 shows that, given the conditions of Theorem 4, either of the convergences (53) or (54) implies (52). So, it will suffice to prove that (52) implies (55).

Assume (52). Suppose, for some $t_1 \geq 0$, property (55), with $t = t_1$, does not hold. This implies that, in addition to the pathwise inequality $\tilde{x}^r(t_1) \leq {}^q\tilde{x}^r(t_1)$ (see (51)), we have, for some fixed constant $c > 0$,

$$\liminf_{r\to\infty} P\{{}^q\tilde{x}^r(t_1) - \tilde{x}^r(t_1) > c\} > c. \tag{106}$$

Consider a subsequence of indices r along which the distributions of both $\tilde{x}^r(t_1)$ and ${}^q\tilde{x}^r(t_1)$ (weakly) converge to some distributions, which we denote η and ${}^q\eta$, respectively. Distribution η is necessarily equal to the distribution of $\tilde{x}(t_1)$, and ${}^q\eta$ (stochastically) dominates η and *is not equal to* η, which follows from (106).

Consider the processes $\tilde{x}^r$ and ${}^q\tilde{x}^r$ restarted at time t_1. Let us fix arbitrary t_2, $t_1 < t_2 < \infty$. Then we have

$$\liminf_r P\left\{\inf_{[t_1,t_2]} \tilde{x}^r(t) < \epsilon\right\} \geq p_1, \quad \forall \epsilon > 0, \tag{107}$$

where

$$p_1 = P\{\tilde{x}(t) \text{ hits } 0 \text{ within } [t_1, t_2]\}.$$

Since pathwise inequality (105) holds, we see that (107) holds for the process ${}^q\tilde{x}^r$ as well:

$$\liminf_r P\left\{\inf_{[t_1,t_2]} {}^q\tilde{x}^r(t) < \epsilon\right\} \geq p_1, \quad \forall \epsilon > 0. \tag{108}$$

Consider now an RBM ${}^q\tilde{x}$ with the drift a and diffusion coefficient σ (same as for the RBM $\tilde{x}$), defined within interval $[t_1, t_2]$, with the initial distribution ${}^q\eta$ at time t_1. Since ${}^q\eta$ strictly dominates η,

$$P\{{}^q\tilde{x}(t) \text{ hits } 0 \text{ within } [t_1, t_2]\} = p_2 < p_1. \tag{109}$$

Using Theorem 2(ii), it is easy to see that the RBM ${}^q\tilde{x}$ is an asymptotic (stochastic) lower bound of the sequence of processes ${}^q\tilde{x}^r$ (in the sense specified in Theorem 2(ii)). From this fact we see that for any $\delta > 0$, we can choose a sufficiently small $\epsilon > 0$, so that

$$\liminf_r P\left\{\inf_{[t_1,t_2]} {}^q\tilde{x}^r(t) > \epsilon\right\} \geq (1 - p_2) - \delta. \tag{110}$$

If we fix $\delta \in (0, p_1 - p_2)$ and a corresponding $\epsilon > 0$ as above, we obtain the following from the estimates (110) and (108):

$$\liminf_r \left[P\left\{\inf_{[t_1,t_2]} {}^q\tilde{x}^r(t) > \epsilon\right\} + P\left\{\inf_{[t_1,t_2]} {}^q\tilde{x}^r(t) < \epsilon\right\}\right] \geq p_1 + (1 - p_2) - \delta > 1,$$

a contradiction, which completes the proof. ■

13. STABILITY VERSUS HEAVY TRAFFIC WORKLOAD MINIMIZATION

Consider a special case of the MinDrift routing rule, with cost functions $C_j(\zeta) = (\frac{1}{2})\zeta^2$ (see Section 4.4). The corresponding MinDrift(Q) rule is as follows: *route an arriving type i customer to a server j such that*

$$j \in \arg\min_{j \in J} {}^{q}U_j(t)/\mu_{ij}. \tag{111}$$

Our heavy traffic results apply to this rule. (The form of the rule does not change with r, as explained in Section 8.)

Using the approach of [1,2,6,14,15], it is not hard to show that under this routing rule (plus arbitrary scheduling rules satisfying (d1) and (d2)), both the queue length process $((Q_{ij}(t), i \in I, j \in J), t \geq 0)$ and the unfinished work process $((U_j(t), j \in J), t \geq 0)$ are *stable*, as long as the vector of mean rates λ is within the system stability region $\mathcal{M}^0$, defined in Section 5. In fact, as explained in [14], the analysis of the FSPs (Section 9), required to establish the heavy traffic results, is essentially a "superset" of the analysis needed to prove stability. (We do not provide details of the stability proof, as it is not the focus of this article.)

Thus, the above rule is able to both keep queues stable (as long as $\lambda \in \mathcal{M}^0$) and minimize system workload in the heavy traffic limit. The G$c\mu$ scheduling rule for the IQ system possesses the same property (see [12]) and so does the MaxWeight scheduling rule for a different, but closely related, "generalized switch" model (see [14]). All of these results may suggest the intuition that a dynamic service discipline that keeps queues stable (as long as $\lambda \in \mathcal{M}^0$), "typically" will also minimize system workload in heavy traffic. Such a "conjecture" cannot be formally correct, because it is not hard to devise some *contrived* disciplines, for which it does not hold. We note, however, that this conjecture does not hold even for *very natural* service disciplines, as the following example demonstrates.

Consider a service discipline for our OQ system, which strives to minimize the drift of the cost (Lyapunov) function

$$\sum_{ij} \frac{1}{2} Q_{ij}^2(t).$$

Then the discipline has the following form. (It is close to the class of network scheduling disciplines introduced in [15].)

Routing rule ("Join the shortest queue of your type"): Route an arriving type i customer to a server j such that

$$j \in \arg\min_{j \in J: \mu_{ij} > 0} Q_{ij}(t). \tag{112}$$

Scheduling rule ("G$c\mu$ within each server"): Server j picks a customer of type i such that

$$i \in \arg\max_{i \in I} Q_{ij}(t)\mu_{ij}. \tag{113}$$

This discipline ensures stability of the queues, when $\lambda \in \mathcal{M}^0$. (Again, the approach and techniques of [1,2,6,14,15] can be applied.)

However, it is not hard to see that, under this discipline and under the conditions of Theorem 4, condition (55) cannot possibly hold, as long as $\mu_{ij} > 0$ for at least one nonbasic activity (ij). We do not provide a formal proof here. The key part of a proof is showing the following very intuitive fact, which is implied by the nature of the routing rule: *FSPs under this discipline are such that if the initial workload is nonzero, then after some finite time, all nonbasic queue lengths are bounded away from zero*. We also exploit the fact that since the limiting workload process is lower bounded by an RBM, at any time $t > 0$ the limiting workload is nonzero with nonzero probability. Thus, by Theorem 4, none of the workload minimization properties (52)–(54), can hold.

Acknowledgment

The author is very grateful to Avi Mandelbaum for bringing output-queued flexible server systems to his attention and encouraging this work.

References

1. Andrews, M., Kumaran, K., Ramanan, K., Stolyar, A.L., Vijayakumar, R., & Whiting, P. (2004). Scheduling in a queueing system with asynchronously varying service rates. *Probability in the Engineering and Informational Sciences* 18: 191–217.
2. Armony, M. & Bambos, N. (1999). Queueing networks with interacting service resources. In *Proceedings of the 37th Annual Allerton Conference*, pp. 42–52.
3. Bell, S.L. & Williams, R.J. (2001). Dynamic scheduling of a system with two parallel servers in heavy traffic with complete resource pooling: Asymptotic optimality of a continuous review threshold policy. *Annals of Applied Probability* 11: 608–649.
4. Bramson, M. (1998). State space collapse with applications to heavy traffic limits for multiclass queueing networks. *Queueing Systems* 30: 89–148.
5. Chen, H. & Mandelbaum, A. (1991). Leontief Systems, RBV's and RBM's. In M.H.A. Davis & R.J. Elliott (eds.), *Applied stochastic analysis*. Gordon and Breach Science, pp. 1–43.
6. Dai, J.G. & Prabhakar, B. (2000). The throughput of data switches with and without speedup. In *Proceedings of the INFOCOM'2000*, pp. 556–564.
7. Ethier, S.N. & Kurtz, T.G. (1986). *Markov process: Characterization and convergence*. New York: John Wiley & Sons.
8. 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.
9. Harrison, J.M. & Lopez, M.J. (1999). Heavy traffic resource pooling in parallel-server systems. *Queueing Systems* 33: 339–368.
10. Kelly, F.P. & Laws, C.N. (1993). Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling. *Queueing Systems* 13: 47–86.
11. Laws, C.N. (1992). Resource pooling in queueing networks with dynamic routing. *Advances in Applied Probability* 24: 699–726.
12. Mandelbaum, A. & Stolyar, A.L. (2004). Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Operations Research* 52(6): 836–855.
13. McKeown, N., Anantharam, V., & Walrand, J. (1996). Achieving 100% throughput in an input-queued switch. In *Proceedings of the INFOCOM'96*, pp. 296–302.
14. Stolyar, A.L. (2004). MaxWeight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *Annals of Applied Probability* 14(1): 1–53.

15. Tassiulas, L. & Ephremides, A. (1992). Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio network. *IEEE Transactions on Automatic Control* 37: 1936–1948.
16. Teh, Y. & Ward, A. (2002). Critical thresholds for dynamic routing in queueing networks. *Queueing Systems* 42: 297–316.
17. Van Mieghem, J.A. (1995). Dynamic scheduling with convex delay costs: The generalized cμ rule. *Annals of Applied Probability* 5: 809–833.
18. Williams, R.J. (1998). An invariance principle for semimartingale reflecting Brownian motions in an orthant. *Queueing Systems* 30: 5–25.
19. Williams, R.J. (1998). Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse. *Queueing Systems* 30: 27–88.
20. Williams, R.J. (2000). On dynamic scheduling of a parallel server system with complete resource pooling. *Fields Institute Communications* 28: 49–71.

APPENDIX
The One-Dimensional Skorohod Problem

The following proposition describes standard properties of solutions to the one-dimensional Skorohod problem. (See, for example, [5] for the proof. The proof is also contained in the proof of Theorem 5.1 of [18]).

PROPOSITION 1: *Let* $w = (w(t),\ t \geq 0)$ *be a continuous function in* $D([0,\infty), R)$ *such that* $w(0) \geq 0$. *Then the following hold:*

(i) There exists a unique pair (x, y) *of functions in* $D([0,\infty), \bar{R})$, *such that the following hold:*
  (a) $x(t) = w(t) + y(t) \geq 0,\ t \geq 0$.
  (b) y *is nondecreasing and nonnegative.*
  (c) $y(0) = 0$.
  (d) For any $t \geq 0$, *if* $x(t) > 0$, *then* t *is not a point of increase of* y; *that is, there exists* $\delta > 0$ *such that* $y(\xi)$ *is constant in* $[t-\delta, t+\delta] \cap R_+$. *This unique pair is* (x°, y°), *where*

$$y^\circ(t) \doteq -\Big[0 \wedge \inf_{0 \leq u \leq t} w(u)\Big], \qquad x^\circ(t) = w(t) + y^\circ(t), \qquad t \geq 0.$$

(ii) For any pair (x, y) *of functions in* $D([0,\infty), \bar{R})$ *satisfying (a) and (b), we have*

$$y(t) \geq y^\circ(t), \qquad x(t) \geq x^\circ(t), \qquad t \geq 0.$$