OPERATIONS RESEARCH
Vol. 52, No. 6, November–December 2004, pp. 836–855
ISSN 0030-364X | EISSN 1526-5463 | 04 | 5206 | 0836

DOI 10.1287/opre.1040.0152

Scheduling Flexible Servers with Convex Delay Costs: Heavy-Traffic Optimality of the Generalized $c\mu$-Rule

Avishai Mandelbaum
Industrial Engineering and Management, Technion, Haifa 32000, Israel, avim@tx.technion.ac.il

Alexander L. Stolyar
Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, stolyar@research.bell-labs.com

We consider a queueing system with multitype customers and flexible (multiskilled) servers that work in parallel. If Q_i is the queue length of type i customers, this queue incurs cost at the rate of $C_i(Q_i)$, where $C_i(\cdot)$ is increasing and convex. We analyze the system in heavy traffic (Harrison and Lopez 1999) and show that a very simple generalized $c\mu$-rule (Van Mieghem 1995) minimizes both instantaneous and cumulative queueing costs, asymptotically, over essentially all scheduling disciplines, preemptive or non-preemptive. This rule aims at myopically maximizing the rate of decrease of the instantaneous cost at all times, which translates into the following: when becoming free, server j chooses for service a type i customer such that $i \in \arg\max_i C_i'(Q_i)\mu_{ij}$, where μ_{ij} is the average service rate of type i customers by server j.

An analogous version of the generalized $c\mu$-rule asymptotically minimizes delay costs. To this end, let the cost incurred by a type i customer be an increasing convex function $C_i(D)$ of its sojourn time D. Then, server j always chooses for service a customer for which the value of $C_i'(D)\mu_{ij}$ is maximal, where D and i are the customer's sojourn time and type, respectively.

Subject classifications: queues: diffusion models, networks, optimization; production/scheduling: sequencing/stochastic; networks: stochastic.
Area of review: Stochastic Models.
History: Received January 2002; revisions received February 2003, June 2003; accepted July 2003.

1. Introduction

In this paper, we analyze the scheduling problem for flexible servers with overlapping capabilities. Our setup is a queueing system with multitype customers, multiskilled servers, and delay costs that are convex increasing in queue lengths or sojourn times. For such a system in heavy traffic, we show that a simple generalized $c\mu$-rule ($Gc\mu$) is in fact asymptotically optimal. This is a far-reaching generalization of Van Mieghem's (1995) striking result for homogeneous servers. It constitutes a natural progression of Stolyar (2004), which is here adapted to the (parallel server) model of Harrison and Lopez (1999). (Williams 1998c is recommended as an introduction to the subject.)

To describe our $Gc\mu$-rule, let μ_{ij} denote the service rate of type i customers by server j. (μ_{ij} is the reciprocal of an average service time; $\mu_{ij} = 0$ indicates that server j cannot serve type i.) Assume first that the queue of type i incurs a queueing cost at rate $C_i(Q_i)$, which is an increasing convex function $C_i(\cdot)$ of the queue length Q_i. (Further properties of the C_i's are listed in §4, of which one should mention here only that $C_i(0) = C_i'(0+) = 0$.) Then, applying the $Gc\mu$-rule when becoming free at time t, server j takes for service a type i customer such that

$$i \in \arg\max_i C_i'(Q_i(t))\mu_{ij}.$$

As discussed later in this section, roughly speaking, the $Gc\mu$-rule is a scheduling discipline that myopically tries to maximize the rate of decrease of the instantaneous holding cost $\sum_i C_i(Q_i(t))$.

An alternative cost structure is when each type i customer incurs, up to a time t, a delay cost $C_i(D)$ which is a function of its sojourn time D (up to its service completion or up to time t, whichever comes first). Then, server j takes for service the longest-waiting (head of the line) type i customer, where

$$i \in \arg\max_i C_i'(D_i(t))\mu_{ij},$$

in which $D_i(t)$ is the longest sojourn time of a customer waiting in queue i at time t. (Heavy traffic renders irrelevant the decisions about customers who encounter idle servers upon arrival.) Our main result is Theorem 1 of §7. We show there that the above Q-version of $Gc\mu$ is optimal in heavy traffic, in that it asymptotically minimizes queueing costs at all times. An analogous result for the D-version holds with respect to sojourn time costs (Theorem 2 in §7.1).

The $Gc\mu$ scheduling rule is adaptive and robust. Indeed, its form depends on no system parameters other than service rates and cost functions; its scheduling decisions depend only on the current system state (queue lengths or sojourn times). Thus, the rule adapts automatically to environmental changes; for example, there is no need to modify it with changes in arrival rates. Additional properties of $Gc\mu$ are described in our concluding §10, notably its accommodation of linear costs (which are formally excluded in view of the assumption $C_i'(0+) = 0$).

The optimality of $Gc\mu$ is relative to essentially all scheduling disciplines, preemptive or non-preemptive, as long as each server cannot have two or more customers of the same type in service, and servers do not know the realizations of service times before customers are taken for service. Our asymptotics is for a sequence of systems that approach heavy traffic, in a way that is precisely defined in §6 and which we now describe.

A given set of service rates $\{\mu_{ij}\}$ determines the stability set M for our queueing system: this is the closure of the set of arrival rates $\lambda \in R_+^I$, with which our system is stable for at least one scheduling strategy (§5). The northeast boundary of the stability set (its maximal elements) constitute those arrival rates λ for which our system is critically loaded. The system is in heavy traffic when its vector of arrival rates is "close" to a maximal $\lambda \in M$ (§6). We further say that our system exhibits *complete resource pooling* (CRP) if the outer normal ν^* to the set M at that λ is unique up to scaling (plus some additional nondegeneracy conditions) and, in addition, all the coordinates of ν^* are strictly positive. (See §5, at the end of which we also comment on how CRP can be relaxed to merely *resource pooling*, which allows zero coordinates of ν^*.)

The quantity $X(t) = \sum_i \nu_i^* Q_i(t)$ is called the *workload* at time $t \geqslant 0$. Our main result is that, under CRP and in the heavy traffic limit, the $Gc\mu$-rule minimizes the workload $X(t)$ at any time t and, moreover, given the value of $X(t)$, the queue length vector $Q(t)$ is the one that minimizes the cost rate $\sum_i C_i(Q_i(t))$. These two properties imply minimization of the cumulative queueing costs over any finite interval. We discuss the intuition behind this unexpected result momentarily, after a brief literature survey that places the present research in historical perspective.

For single server queues, our scheduling problem has had a long history. It started at least as early as Cox and Smith (1961), who proved optimality of the $c\mu$-rule for the $M/G/1$ queue with linear waiting costs (optimality among priority rules, to be precise); and it culminated in Van Mieghem's (1995) analysis of the $G/G/1$ queue with convex costs, where asymptotic optimality of the above generalized $c\mu$-rule was established in heavy traffic. (Van Mieghem 1995 is also recommended for further motivation and references.)

The history is far leaner for multiservers, yet several streams of research had to mature into our present paper. Rather than repeat what has already been published, we mention here only the origins of these streams (to the best of our knowledge), and then refer the reader to survey papers for later developments.

Starting with stability, Tassiulas and Ephremides (1992) seem to be the first who characterized the stability set M, proving also that *MaxWeight*-type rules ensure stability (if such is at all feasible). A MaxWeight-rule is essentially $Gc\mu$ with quadratic costs: $C_i(Q_i) = \gamma_i Q_i^2$, for arbitrary constants $\gamma_i > 0$. Thus, for the model in this paper, the MaxWeight-rule is in fact a weighted "$Q\mu$"-rule: server j chooses for service the longest-waiting type i customer such that $i \in \arg\max_i \gamma_i Q_i(t)\mu_{ij}$. For more recent results on stability, readers are referred to McKeown et al. (1996), Dai and Prabhakar (2000), Armony and Bambos (1999), Andrews et al. (2004), and references therein.

As far as heavy traffic optimization is concerned, Harrison (1998) was first to go beyond single servers, with a two-type two-server model in which one server is dedicated and the other flexible. With linear costs, Harrison (1998) proved asymptotic optimality in heavy traffic of a discrete-review policy. Bell and Williams (2001) established, for that same model, asymptotic optimality of an alternative continuous-review threshold strategy. Harrison and Lopez (1999) extended Harrison (1998) to a general system, such as ours above, about which they heuristically derived conditions for heavy traffic, for complete resource pooling and for discrete-review optimality. Similarly, Williams (1998c) generalized Bell and Williams (2001), conjecturing optimality in heavy traffic for a carefully devised threshold strategy; she also provides an overview of and additional insights for Harrison (1998), Bell and Williams (2001), and Harrison and Lopez (1999), especially the conditions for complete resource pooling. The general notion of complete resource pooling in heavy traffic is closely related to the equivalent workload formulations of Brownian control problems, which were formalized by Harrison and Van Mieghem (1997). In the recent work Stolyar (2004), a (discrete-time) *generalized switch model* (which is more general than ours) was analyzed, and the MaxWeight rule ($Gc\mu$ with quadratic costs) was shown to cause a one-dimensional state space collapse and minimize the workload, in the heavy traffic limit. The proof of state space collapse in Stolyar (2004) follows the general approach pioneered by Bramson (1998) and Williams (1998b).

Our analysis (and intuition behind the main result) is analogous to that in Stolyar (2004). First we prove that, under the CRP condition and $Gc\mu$-rule, sample paths of the fluid process corresponding to a critically loaded system (input rates equal to λ) are such that the queue length vector $Q(t)$ is attracted to a fixed point $^\circ Q$, namely a point such that the vector $(C_1'(^\circ Q_1), \ldots, C_I'(^\circ Q_I))$ is proportional to the vector ν^*. (A fixed point $^\circ Q$ is exactly the point that minimizes $\sum C_i(Q_i)$, given the workload value $\sum \nu_i^* {}^\circ Q_i$.) The attraction implies that, in the heavy traffic (diffusion) limit,

the queue length process exhibits *state space collapse*—$Q(t)$ is a process "living" on the one-dimensional manifold of fixed points. (Again, the proof of such an "implication" follows the approach developed in Bramson 1998 and Williams 1998b.) By virtue of the $Gc\mu$-rule, it maximizes (roughly speaking) the value of

$$\sum_i C_i'(Q_i(t))\mu_i(t),$$

where $\mu_i(t)$ is the aggregate "instantaneous" service rate of flow i by all servers. But, as long as total queue length is nonzero, the vector $(C_1'(Q_1(t)), \ldots, C_I'(Q_I(t)))$ is approximately proportional to ν^*, and therefore the $Gc\mu$-rule "reduces to" the rule that maximizes the value of

$$\sum_i \nu_i^* \mu_i(t).$$

In other words, the server pool operates as a single "superserver," which serves the workload at the maximum possible rate. This implies the property of workload minimization. Consequently, cost rates are minimized at all times $t \geqslant 0$, because, as noted above, a fixed point minimizes the cost rate for the corresponding value of the workload.

The distinguishing feature of our analysis, compared to Stolyar (2004), is that here we deal with continuous time. In the discrete time generalized switch model of Stolyar (2004), the number of customers of each type that can be served in a "time slot" is a function of only the "scheduling decision" chosen in this slot. Therefore, the maximum possible amount of workload service is a well-defined quantity for each time slot. This makes the definition of the "superserver idleness" process (which is the cumulative amount of workload service, "wasted" by the superserver) very direct and natural. In our continuous-time case, on the other hand, we must resort to a less direct definition of the idleness process—this process becomes a "true" superserver idleness only in the limit. Another difference is our more general convex cost structure—the MaxWeight discipline in Stolyar (2004) can be viewed as a special case of $Gc\mu$ (with quadratic cost functions). With general convex costs, the one-dimensional manifold of fixed points (on which the limit process "lives") can be *nonconical* (just as in Van Mieghem 1995). One should note, however, that the model of Stolyar (2004) is more general in that it allows for arbitrary dependence between servers, and for the random variations of server "states" (i.e., their service rates μ_{ij}).

We extend the scope of $Gc\mu$ optimality from Van Mieghem's (1995) single server model to multiple nonhomogeneous servers. The above discussion brings out the reasons such a generalization is nontrivial, perhaps even nonintuitive. Indeed, in a system with a single server $j = 1$, the CRP condition holds trivially, and the components of the vector ν^* are simply the mean service times: $\nu_i^* = 1/\mu_{i1}$. Thus, the workload is simply $X(t) = \sum(1/\mu_{i1})Q_i(t)$, and, consequently, *any* work-conserving non-preemptive discipline trivially minimizes it in heavy traffic. Therefore, a scheduling rule that seeks to minimize costs need only be "concerned" with distributing workload among the queues in such a way that the instantaneous cost rate is minimized; and the $Gc\mu$-rule is a natural way to do that. In our multiserver case, on the other hand, at a first glance it may appear unlikely that a rule as parsimonious as $Gc\mu$, which does not utilize the notion of workload in any way, would be able to simultaneously and "automatically" minimize both the workload and the cost rate (given the workload). And yet, $Gc\mu$ causes our system to "self-organize" (via state space collapse) so that those properties indeed hold.

The outline of this paper is as follows. In §2, we set basic notations and conventions. Our model of a queueing system with multitype customers and flexible servers is formally introduced in §3. The generalized $c\mu$-rule ($Gc\mu$) is described in §4. In §5, we formulate the conditions for complete resource pooling (CRP), followed by the definition of heavy traffic in §6. Theorem 1, in §7, establishes the asymptotic optimality of $Gc\mu$, with respect to queueing costs; Theorem 2 is the analogous result for sojourn time costs. The rest of the paper contains steps of proving the main result, Theorem 1. Section 8 deals with state space collapse: first, fluid sample paths of a critically loaded system are introduced, and then their uniform convergence to a fixed point is proved in Theorem 3. Finally, Theorem 1 is proved in §9. (Adaptation of this proof to the proof of Theorem 2, for sojourn time costs, is outlined in the appendix.) We conclude in §10 with commentary on the robustness and useful features of $Gc\mu$ in applications, and on some possible extensions.

2. Notation and Conventions

We use the standard notations R and R_+ for the sets of real and real nonnegative numbers, respectively; and the not quite standard R_{++} 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 just 0 for the zero vector in R^N.

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, Ethier and Kurtz 1986 for the definition of this space and its associated topology and σ-algebra.)

The symbol $\stackrel{w}{\Rightarrow}$ 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 $\stackrel{\text{u.o.c.}}{\longrightarrow}$ (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 shall not need any characterization of the topology on $D([0,\infty), \bar{R})$, beyond the definition of convergence given above.

3. The Model

We consider a queueing 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 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 $v_{ij}(n), n = 1, 2, \ldots$; $v_{ij}(0)$ is the residual service time, at time 0, of type i customer at server j (if there is any). Let $\mu_{ij} = 1/E[v_{ij}(1)] < \infty$ and $\beta_{ij}^2 = \mathrm{Var}[v_{ij}(1)]$. The convention $\mu_{ij} = 0$ is used when server j cannot serve type i. All arrival and service processes are assumed mutually independent.

We allow a wide class of scheduling disciplines, which adhere to the following conditions:

Condition (d1). Once a customer is taken for service by a server, this customer cannot be served by any other server. Also, a server cannot take for service a new customer of type i if it already has another type i customer "in service" (with nonzero residual service time). Consequently, at any given time, a server cannot have in service more than one customer of any given type.

Condition (d2). Servers do not "know" the realizations of customer service times before customers are taken for service.

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.

Customers of type i that await service are waiting in queue i of infinite capacity. Denote by $Q_i(t)$ the queue length of type i customers at time t; by convention, this number includes those customers whose service already started but not yet completed. Let $D_i(t)$ be the *sojourn time* ("age") of the longest-in-system ("oldest") customer of type i, among those who have not been taken yet for service by time t.

Let $F_i(t)$ be the number of type i customers arrived to the system by time t (excluding initial customers). Denote by $\hat{F}_{ij}(t)$ the number of type i customers that were served by server j, and whose service was completed by time t. Let $U_i(t)$ be the residual interarrival time for type i at time t, and $V_{ij}(t)$ the residual service time, at time t, of the type i customer being served by server j, if there is such customer; otherwise let $V_{ij}(t) = 0$ by convention. Note that F_i and U_i are given primitives while $\hat{F}_{ij}$ and $V_{ij}(t)$ are scheduling dependent.

4. The $Gc\mu$-Rule

Suppose that for each type i, a cost function $C_i(\zeta)$, $\zeta \geqslant 0$, is given. Assume that the cost functions have the following properties: $C_i(\cdot)$ *is continuous strictly increasing convex, with* $C_i(0) = 0$; *moreover, the first derivative* $C_i'(\cdot)$ *is continuous strictly increasing, with* $C_i'(0) = 0$; *finally, the second derivative* $C_i''(\cdot)$ *is strictly positive continuous in the open interval* $(0,\infty)$, *with* $C_i''(0) = \lim_{\zeta \downarrow 0} C_i''(\zeta) \geqslant 0$, *where* $C_i''(0)$ *is either finite or is* $+\infty$.

The $Gc\mu$-rule schedules customers for service as follows. When server j becomes free, it chooses for service a customer from a queue i such that

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

and serves this customer to completion, without preemptions. Ties are broken arbitrarily, for example, in favor of the largest index i. Similarly, assignments of customers to idle servers, if such exist upon arrivals, is arbitrary: for example, in favor of the smallest index j.

REMARK 1. Due to Condition (d2), it does not matter which customer is taken for service from a queue, when the queue to serve is chosen. (The queue length process is unaffected.) So, to be concrete, we can assume that the head-of-the-line (longest waiting) customer is always taken, i.e., customers of each type are taken from the queue in FCFS order.

REMARK 2. The above version of the $Gc\mu$-rule accommodates queueing costs. An alternative, for costs of sojourn times, will be introduced in §7.1. For the sojourn time version, it is important that customers within each type are taken from the queue in FCFS order.

5. Complete Resource Pooling

Consider a "column-substochastic" matrix $\phi = \{\phi_{ij}, i \in I, j \in J\}$, namely all $\phi_{ij} \geqslant 0$ and

$$\sum_i \phi_{ij} \leqslant 1 \quad \forall\, j \in J.$$

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

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

this is the vector of mean service rates of the queues $i \in I$, if each server j allocates a fraction ϕ_{ij} of its time to queue i, in the long run.

DEFINITION. We define M to be the set of $\mu(\phi)$ corresponding to all possible ϕ as above. Further, let M^* denote the set of all maximal elements $\mu \in M$ such that $\mu \in R^I_{++}$. ($\mu \in M$ is maximal if $\mu \leqslant \zeta \in M$ implies $\zeta = \mu$.)

Note that M is a polyhedron in R^I_+. We assume that M is nondegenerate (i.e., has dimension I), which is equivalent to assuming that each queue i can be served at nonzero rate μ_{ij} by at least one server j. The set M is in fact the closure of our system's stability region M^0, which is the set of arrival rate vectors $\lambda = (\lambda_1, \ldots, \lambda_I)$ such that $\lambda < \mu(\phi)$ for some ϕ (cf. Tassiulas and Ephremides 1992, McKeown et al. 1996, Dai and Prabhakar 2000, Armony and Bambos 1999, Andrews et al. 2004, Stolyar 2004).

DEFINITION. We say that the condition of *Complete Resource Pooling* CRP holds for a vector λ if λ lies within the interior of one of the $((I-1)$-dimensional) outer faces of M, $\lambda \in M^*$, and the matrix ϕ such that $\lambda = \mu(\phi)$ is unique.

If CRP holds for λ, then the corresponding matrix ϕ is such that $\sum_i \phi_{ij} = 1$ for each j (ϕ must be column stochastic), because otherwise λ would not be maximal.

Our CRP condition is equivalent to that introduced for parallel server systems in Harrison and Lopez (1999) and Williams (1998c). (See Assumption 3.4, Theorem 5.3, and Corollary 5.4 in Williams 1998c for a summary.) In those papers, pairs (i, j) such that $\phi_{ij} > 0$ are called *basic activities*, and it is shown that the CRP condition implies the following: The graph with nodes being queues i and servers j, and arcs being basic activities, is a connected tree (with exactly $I + J - 1$ arcs).

We wish to emphasize here that the notion of a basic activity is not utilized in any way (neither explicit nor implicit) by the $Gc\mu$ scheduling algorithm. (The algorithm need not know which activities are basic.) It is only used as a tool for the analysis of the algorithm.

When the 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 M at the point λ. Note that $\nu \in R^I_{++}$. (Otherwise, if some $\nu_i \leqslant 0$, a small increase of the component λ_i would produce a vector $\lambda' \geqslant \lambda$, $\lambda' \neq \lambda$, and such that $\lambda' \in M$—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 of ν^* are sometimes called the *workload contributions* of customers of the different flows (see Harrison and Lopez 1999, Williams 1998c).

As is the case with basic activities, the $Gc\mu$-rule does not require any knowledge of the values of workload contributions. It is interesting to note, however, that ν^* can be computed by solving a certain linear program and its dual. (See Harrison and Lopez 1999, Williams 1998c, Stolyar 2004; the latter is compatible with the notations of the present paper.)

The CRP condition for λ implies, in particular, that

$$\nu^* \cdot \lambda = \max_{\mu \in M} \nu^* \cdot \mu. \tag{1}$$

For each $j \in J$, let us denote by

$$I_j = \{i \in I \,|\, \phi_{ij} > 0\}$$

the set of i such that (i, j) is a basic activity for server j. It is easy to verify that $i \in I_j$ implies $i \in \arg\max_i \nu^*_i \mu_{ij}$. (Otherwise, $\lambda = \mu(\phi)$ would not maximize the right-hand side of (1).) The converse, under the CRP condition for λ, holds as well. (Otherwise, using the fact that basic activities form a connected tree, we could "reallocate" the fractions ϕ_{ij} to produce a matrix $\phi' \neq \phi$, but such that $\mu(\phi') = \mu(\phi) = \lambda$, a contradiction to the uniqueness of ϕ.) We thus obtain

LEMMA 1. *Under the CRP condition, for any server j, $I_j = \arg\max_i \nu^*_i \mu_{ij}$.*

REMARK 1. Without the maximality requirement $\lambda \in M^*$, CRP is relaxed to a *resource pooling* (RP) condition. When the RP (but not necessarily CRP) condition holds, workload contributions ν^*_i of some flows i could be 0. Then, in analogy to Chen and Mandelbaum (1991a) and Stolyar (2004), the types i with strictly positive workload contributions (and the servers which can serve those types at a nonzero rate) form a *bottleneck subsystem*. The results of this paper

can be generalized to the case of the RP condition as follows. Suppose that for the bottleneck subsystem, in isolation, the condition of uniqueness of the matrix ϕ does hold, which implies that CRP holds for this subsystem. Then, in the heavy traffic limit, as defined in the next section, the nonbottleneck queues vanish (stay at 0), and the behavior of the bottleneck queues is the same as it would be for the bottleneck subsystem in isolation (as described by the results of this paper). The generalization can be carried out along the lines of Stolyar (2004), where the RP condition is employed (for a different, but closely related model).

REMARK 2. Suppose that the set of servers J can be partitioned into server groups of statistically identical servers: service time distributions (which may be different across types) are the same within each group. For such a system, the assumption of the uniqueness of $\{\phi\}$ in the CRP condition can be relaxed: It is sufficient to require that, for any ϕ such that $\lambda = \mu(\phi)$, the sum

$$\sum_{j\in\hat{J}} \phi_{ij}$$

is unique only for every server group $\hat{J} \subseteq J$. Under this relaxed assumption, all our results carry over as is. The generalization of our proofs for this case (which is important for applications) is straightforward; we do not pursue it here to simplify the exposition.

REMARK 3. In general, however, the requirement in CRP that matrix ϕ is unique could turn out essential in that, without it, even with ν^* unique, the $Gc\mu$-rule need not be asymptotically optimal. Roughly speaking, in the latter case, the diffusion coefficient σ (defined later in the paper) of the limiting diffusion process is not well defined.

6. The Heavy Traffic Regime

In this section, we introduce the notion of a sequence of queueing systems in heavy traffic. First, fix a vector λ satisfying the CRP condition. With λ there is an associated (unique) matrix ϕ such that $\lambda = \mu(\phi)$, and for which

$$\sum_i \phi_{ij} = 1 \quad \forall\, j \in J,$$

must hold. There is also a corresponding (unique) normal vector ν^*, in terms of which we define

$$X(t) \doteq \sum_{i=1}^{N} \nu_i^* Q_i(t) = \nu^* \cdot Q(t), \quad t \geqslant 0.$$

The process $X(\cdot)$ will be referred to as the workload of the system.

We now consider a sequence of queueing 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 paper, $r \to \infty$ means that r goes to infinity along values from 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 before, 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

$$r(\lambda_i^r - \lambda_i) \to b_i, \quad r \to \infty, \tag{2}$$

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

$$[\alpha_i^r]^2 \to \alpha_i^2, \quad r \to \infty. \tag{3}$$

In addition, we make the following Lindeberg type assumption, which is needed to apply the functional central limit theorem (FCLT), and also to apply Bramson's weak law estimates (Bramson 1998) (and establish (30) later on):

uniformly over i and r,

$$E[(u_i^r(1))^2 1\{u_i^r(1) > x\}] \leqslant \eta(x), \quad x \geqslant 0, \tag{4}$$

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

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

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

Assumptions (2)–(4) imply the FCLT for the arrival processes

$$\{r^{-1}(F_i^r(r^2 t) - \lambda_i^r r^2 t), t \geqslant 0\} \xrightarrow{w} \{\sigma_i B(t), t \geqslant 0\}, \tag{5}$$

where $F_i^r(t)$ is the number of type i customers arrived by time t, excluding customers present at time 0; $\sigma_i^2 = \lambda_i^3 \alpha_i^2$, $B(\cdot)$ 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 parameter r. (This, in particular, means that the condition analogous to (4) trivially holds for the service times $v_{i,j}^r(1)$, uniformly on (i, j) and r.) For the initial residual service times (if any) we assume for all i and j, that

$$v_{i,j}^r(0)/r \to 0, \quad r \to \infty.$$

Let us denote by $S_{ij}^r(t), t \geqslant 0$, the number of type i customers that would be served by server j if it processes type i customers continuously up to time t. Then, an FCLT applies for the processes $S_{ij}^r(\cdot)$:

$$\{r^{-1}(S_{ij}^r(r^2 t) - \mu_{ij} r^2 t), t \geqslant 0\} \xrightarrow{w} \{\sigma_{ij} B(t), t \geqslant 0\}, \tag{6}$$

where $\sigma_{ij}^2 = \mu_{ij}^3 \beta_{ij}^2$.

REMARK. Despite the fact that the service time distribution (for each (i, j)) does not vary with r, we use a superscript r in the notation $S_{ij}^r(\cdot)$ for two reasons. First, it will be important for our proofs to view the processes $S_{ij}^r(\cdot)$ with different r as different processes, not necessarily constructed on a (common) probability space of i.i.d. sequences of service times. Furthermore, because the residual service times $v_{i,j}^r(0)$ may depend on r, the processes $S_{ij}^r(\cdot)$ (with different r) do have, strictly speaking, different distributions.

7. Main Results

For each value of the (scaling) parameter $r \in \mathcal{R}$, let $Q^r(\cdot)$ and $X^r(\cdot) = \nu^* \cdot Q^r(\cdot)$ be the corresponding (vector) queue length and workload processes.

Assume that each queue i, at any time t, incurs a holding cost at the (instantaneous) rate of

$$C_i^r(Q_i^r(t)) = C_i(Q_i^r(t)/r);$$

here $C_i(\cdot)$ are convex increasing functions, with the additional properties described in §4. (An alternative cost structure, where cost is a function of customers' sojourn time, will be discussed in the next subsection.)

We note that our asymptotic regime, in which the cost function is "rescaled" as the parameter r changes, is quite standard in heavy traffic analysis (cf. Van Mieghem 1995, where the same scaling is used, for motivation and further elaboration). Also, note that if the cost functions have the form $C_i(\zeta_i) = \gamma_i \zeta_i^\alpha$, with some fixed $\alpha > 1$ and $\gamma_i > 0$, the cost functions then need not be rescaled with r (see Remark 2 in this section below).

For our main results, we need the notion of a fixed point. A vector $^\circ q \in R_+^I$ will be called a fixed point if

$$[C_1'(^\circ q_1), \ldots, C_I'(^\circ q_I)] = c\nu^* \tag{7}$$

for some constant $c \geqslant 0$. If we recall that each derivative $C_i'(\cdot)$ is continuous strictly increasing with $C_i'(0) = 0$, one deduces the following: *A fixed point $^\circ q$ corresponding to each $c \geqslant 0$ exists and is unique. Moreover, $^\circ q = 0$ for $c = 0$, and $^\circ q \in R_{++}^I$ (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 workload values $\nu^* \cdot {}^\circ q$. In addition, it is easy to verify the following property: *A fixed point $^\circ q$ is the unique vector that minimizes $\sum_i C_i(q_i)$ among all vectors $q \in R_+^I$ with the same workload, i.e., satisfying the condition $\nu^* \cdot q = \nu^* \cdot {}^\circ q$.*

Indeed, if $^\circ q = 0$, the property is trivial. If $^\circ q \in R_{++}^I$, condition (7) implies that the (Lagrangian) function

$$\sum_i C_i(q_i) - c[\nu^* \cdot q - \nu^* \cdot {}^\circ q]$$

has zero gradient (with respect to q) at point $^\circ q$. Because this Lagrangian is strictly convex in R_+^I, it is minimized by $^\circ q$. Then, the desired property follows from the Kuhn-Tucker theorem.

Applying diffusion scaling to $Q^r(\cdot)$ and $X^r(\cdot)$ gives rise to the following scaled processes:

$$\tilde{q}^r(t) \doteq r^{-1} Q^r(r^2 t), \quad t \geqslant 0,$$
$$\tilde{x}^r(t) \doteq r^{-1} X^r(r^2 t), \quad t \geqslant 0.$$

We assume that the initial queue lengths of the scaled processes are deterministic and converging:

$$\tilde{q}^r(0) \to \tilde{q}(0), \tag{8}$$

where $\tilde{q}(0)$ is a fixed point, as defined above. (We comment on this assumption after Theorem 1.) As a consequence, $\tilde{x}^r(0) = \nu^* \cdot \tilde{q}^r(0) \to \nu^* \cdot \tilde{q}(0) \doteq \widetilde{w}(0)$.

Finally, introduce the following one-dimensional reflected Brownian motion $\tilde{x} = \{\tilde{x}(t), t \geqslant 0\}$:

$$\tilde{x}(t) = \widetilde{w}(0) + at + \sigma B(t) + \tilde{y}(t), \tag{9}$$

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

$$\tilde{y}(t) \doteq -\Big[0 \wedge \inf_{0 \leqslant u \leqslant t} \{\widetilde{w}(0) + au + \sigma B(u)\}\Big], \tag{10}$$

and the drift a and diffusion coefficient σ are given by

$$a \doteq \nu^* \cdot b, \quad \sigma^2 \doteq \sum_i (\nu_i^*)^2 \Big[\sigma_i^2 + \sum_j \phi_{ij} \sigma_{ij}^2\Big]. \tag{11}$$

THEOREM 1. *Consider the sequence of queueing systems in heavy traffic, as introduced in §6.*

(1) *Suppose that the scheduling rule is $Gc\mu$ with cost functions $C_i^r(\cdot)$, for each value of the parameter r. Then, as $r \to \infty$,*

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

and

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

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

(2) *The $Gc\mu$-rule is asymptotically optimal in that it minimizes the workload and the holding cost rate at all times. More precisely, let $\tilde{q}_G^r$ and $\tilde{x}_G^r$ be the scaled queue length and workload processes corresponding to an arbitrary scheduling discipline G (and appropriately constructed on a common probability space with our sequence in heavy traffic). Then, with probability 1, for any time $t \geqslant 0$,*

$$\liminf_{r \to \infty} \tilde{x}_G^r(t) \geqslant \tilde{x}(t) \tag{12}$$

and

$$\liminf_{r \to \infty} \sum_i C_i(\tilde{q}_{i,G}^r(t)) \geqslant \sum_i C_i(\tilde{q}_i(t)). \tag{13}$$

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

$$\liminf_{r \to \infty} \int_0^T \sum_i C_i(\tilde{q}_{i,G}^r(t))\, dt \geqslant \lim_{r \to \infty} \int_0^T \sum_i C_i(\tilde{q}_i^r(t))\, dt = \int_0^T \sum_i C_i(\tilde{q}_i(t))\, dt. \tag{14}$$

REMARK 1. Suppose that assumption (8), requiring that $q(0)$ is a fixed point, does not hold. Then, the limiting one-dimensional diffusion process $\tilde{x}$ is the same as in the statement of Theorem 1, except that it starts from some fixed point $^\circ\tilde{q}(0)$ such that its workload $\nu^* \cdot {}^\circ\tilde{q}(0) \in [\nu^* \cdot \tilde{q}(0), K\nu^* \cdot \tilde{q}(0)]$, where $K \geqslant 1$ is a fixed constant spec-

ified later in Theorem 3. In addition, the weak convergence on the interval $[0, \infty)$ in Theorem 1 would be replaced by weak convergence over the open interval $(0, \infty)$. This exact phenomenon arose in Chen and Mandelbaum (1991a) for closed queueing networks and in a context close to ours in Bramson (1998), in his Theorem 3. The basic intuition is that on a fluid scale, the process trajectory reaches a fixed point within a positive finite time $K\nu^* \cdot \tilde{q}(0)$, which is negligible on a diffusion scale.

This means that if $\tilde{q}(0)$ is not a fixed point, the $Gc\mu$-rule allows the initial workload to "jump up" at time 0, i.e., $\nu^* \cdot {}^\circ\tilde{q}(0) > \nu^* \cdot \tilde{q}(0) = \tilde{w}(0)$ can hold. It is possible, of course, that a different scheduling rule, which uses a priori knowledge of system parameters, could avoid such a jump of the initial workload (and hence give rise to lower cumulative costs). However, if the drift $a < 0$, the diffusion process $\tilde{x}$ under $Gc\mu$ reaches 0 within a finite time, with probability 1, and, after that time, the $Gc\mu$-rule does minimize both the workload and cumulative costs.

REMARK 2. Consider the special case of quadratic costs: $C_i(\zeta) = \gamma_i\zeta^2/2$, where $\gamma_i > 0, i \in I$, are given constants. Then, the $Gc\mu$-rule becomes a "$Q\mu$-rule," namely each server j chooses for service a queue i such that

$$i \in \arg\max_{i \in I} \gamma_i Q_i^r \mu_{ij} \tag{15}$$

(which can be considered as a special case of the MaxWeight rule in Stolyar 2004). An important feature of this rule is that *its form does not depend on the scaling parameter r*. The above theorem then says that, in heavy traffic and under CRP, the $Q\mu$-rule minimizes workload and it thrives to keep the vector $[\gamma_1 Q_1, \ldots, \gamma_N Q_N]$ proportional to ν^* at all times: a result analogous to Stolyar (2004).

Theorem 1 deals with transient behavior in heavy traffic. It naturally gives rise to a corresponding, very plausible steady-state result, which we present as Conjecture 1 below, and which basically claims that "the limit of stationary distributions is equal to the stationary distribution of the limit." To formulate this conjecture, consider our sequence of queueing systems in heavy traffic. Suppose that the drift a in (11) is negative and $C_i(\zeta) = \gamma_i\zeta^2/2$, where $\gamma_i > 0, i \in I$, are given constants. Then, under the $Gc\mu$-rule (or, $Q\mu$ in this case) and for all r sufficiently large, the systems are stable. Indeed, the condition $a < 0$ and the CRP condition for λ guarantee that, for large r, the input rate vector λ^r is within the system stability region M^0. Then, stability under the $Q\mu$-rule is established analogously to the way it is done in Tassiulas and Ephremides (1992), McKeown et al. (1996), Dai and Prabhakar (2000), Armony and Bambos (1999), and Andrews et al. (2004) for other MaxWeight-type rules.

CONJECTURE 1. *Suppose that $a < 0$ and $C_i(\zeta) = \gamma_i\zeta^2/2$, where $\gamma_i > 0, i \in I$. Let $\tilde{q}^r(\infty)$ and $\tilde{x}(\infty)$ denote random vector and random variable with distributions equal to the stationary distributions of the processes $\tilde{q}^r$ and $\tilde{x}$, respectively. Then, as $r \to \infty$,*

$$\tilde{q}^r(\infty) \xrightarrow{w} \tilde{x}(\infty)\nu^\circ,$$

where $\tilde{x}(\infty)$ is exponentially distributed with mean $(-2a/\sigma^2)$, and

$$\nu^\circ \doteq \left[\sum_i (\nu_i^*)^2/\gamma_i\right]^{-1}\left(\frac{1}{\gamma} \times \nu^*\right).$$

REMARK 3. Conjecture 1 directly implies that, in the stationary regime, the $Q\mu$-rule stochastically minimizes the quadratic holding cost rate among all disciplines (within the class specified in §3).

7.1. Sojourn Time Costs

Suppose that, as in Van Mieghem (1995), each customer incurs a "one-time" cost that depends on its type and sojourn time in the system. More precisely, as before, consider the sequence of systems indexed by $r \in \mathcal{R}$. Suppose that at time 0 the system is "empty" for each r, i.e., $Q_i^r(0) = 0$ for all i and r. (This condition can be relaxed; we employ it to simplify the exposition.) Let $D_i^r(t, k)$ denote the sojourn time (up to time t) of the kth type-i customer to have arrived to the system by time t. Suppose that, for a fixed $T > 0$, the objective is to (asymptotically) minimize the cumulative waiting cost

$$\mathscr{C}^r(T) \doteq \frac{1}{r^2}\sum_i \sum_{k=1}^{F_i^r(r^2T)} C_i(D_i^r(r^2T, k)/r),$$

where $C_i(\cdot)$ is a cost function with the properties described in §4, and $F_i^r(r^2T)$ is the number of type i arrivals into the system by the time r^2T (as previously defined).

We define the following form of the $Gc\mu$-rule, which we call *D-$Gc\mu$*: *Customers are served without preemption. When becoming free, each server j takes for service the longest-waiting customer from a queue i such that*

$$i \in \arg\max_{i \in I} C_i'(D_i^r(t)/r)\mu_{ij},$$

where $D_i^r(t)$ is the sojourn time at time t ("age") of the longest-waiting ("oldest") type i customer (who, necessarily, has not yet been taken for service by any other server).

Then, the following result, analogously to Theorem 1, holds.

THEOREM 2. *Consider the sequence of queueing systems in heavy traffic, as introduced in §6.*

(1) *Suppose that the scheduling rule is D-$Gc\mu$, as defined above, for each value of the parameter r. Then, as $r \to \infty$,*

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

and

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

where the processes $\tilde{x}^r$, $\tilde{x}$, *and* $\tilde{q}^r$ *are defined as in previous sections, and the process* $\tilde{q}$ *is defined as follows. For each* $t \geqslant 0$, *the vector* $\tilde{q}(t)$ *is the* (*unique*) *fixed point corresponding to cost functions* $\bar{C}_i(\cdot) \doteq \lambda_i C_i(\cdot/\lambda_i)$, *with* $\nu^* \cdot \tilde{q}(t) = \tilde{x}(t)$.

(2) *The D-Gcμ-rule is asymptotically optimal in the following sense. Let* $\tilde{q}^r_G$, $\tilde{x}^r_G$, *and* $\mathscr{C}^r_G$ *be the processes corresponding to an arbitrary scheduling discipline G. Then, these random processes and the corresponding processes* $\tilde{q}^r$, $\tilde{x}^r$, $\mathscr{C}^r$, *under the D-Gcμ discipline, for all different values of r, can be constructed on a common probability space, so that the following properties hold.*
With probability 1, *for any time* $t \geqslant 0$,

$$\liminf_{r\to\infty} \tilde{x}^r_G(t) \geqslant \tilde{x}(t) \tag{16}$$

and

$$\liminf_{r\to\infty} \sum_i \bar{C}_i(\tilde{q}^r_{i,G}(t)) \geqslant \sum_i \bar{C}_i(\tilde{q}_i(t)). \tag{17}$$

Finally, with probability 1, *for any* $T > 0$,

$$\liminf_{r\to\infty} \mathscr{C}^r_G(T) \geqslant \lim_{r\to\infty} \mathscr{C}^r(T) = \int_0^T \sum_i \bar{C}_i(\tilde{q}_i(t))\, dt. \tag{18}$$

The proof of Theorem 2 is essentially an (extended) version of that of Theorem 1. It is outlined in the appendix.

8. Fluid Paths and State Space Collapse under *Gcμ*

8.1. Fluid Sample Paths for the *Gcμ*-Rule

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 $Gc\mu$-rule. More precisely, we need to consider only sample paths of the processes under this scaling, and then their limits, which we formally define below and call *fluid sample paths* (FSPs). The key property of FSPs that must be established (in Theorem 3 below) is that, as time increases to infinity, the queue length vector converges to a fixed point. Using this attraction property and the general approach developed by Bramson (1998) and Williams (1998b), we then prove (in the next section) 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.

First, we introduce some additional (random) functions, associated with the process for each value of the scaling parameter r. (The functions $F^r_i(t)$, $\hat{F}^r_{ij}(t)$, $S^r_{ij}(t)$, $Q^r_i(t)$, and $X^r(t)$, were defined earlier.)

Denote by $G^r_{ij}(t)$ the amount of time within $[0, t]$ that server j was serving type i customers. Clearly, for all $t \geqslant 0$,

$$\hat{F}^r_{ij}(t) = S^r_{ij}(G^r_{ij}(t))$$

and

$$Q^r_i(t) \equiv Q^r_i(0) + F^r_i(t) - \sum_j \hat{F}^r_{ij}(t), \quad t \geqslant 0,\ i \in I. \tag{19}$$

For each pair (i, j) we define

$$H^r_{ij}(t) = \phi_{ij} t - G^r_{ij}(t).$$

As we will clarify later, the function $H^r_{ij}(t)$ has the interpretation of server j cumulative "idleness" (up to time t) relative to the "nominal amount of service" $\phi_{ij} t$ that it could have provided to queue i had it spent exactly a fraction ϕ_{ij} of its time serving that queue. (We remind the reader that ϕ is the unique matrix such that $\mu(\phi) = \lambda$.) Note, however, that unlike "physical" idleness, this function need not be nondecreasing and may even take negative values.

We define the *total cumulative idleness* (or *regulation*) process as follows:

$$Y^r(t) = \sum_{i,j} \nu^*_i H^r_{ij}(t) \mu_{ij}, \quad t \geqslant 0.$$

It is easy to verify that, under the CRP condition, *the regulation* $Y^r(t)$ *is a nonnegative, nondecreasing function*, with $Y^r(0) = 0$. (So, it does have some properties of a "conventional" regulation process.) Indeed, for any $0 \leqslant t_1 < t_2 < \infty$,

$$\frac{Y^r(t_2) - Y^r(t_1)}{t_2 - t_1} = \sum_i \nu^*_i \sum_j [\phi_{ij} - \xi_{ij}]\mu_{ij}$$
$$= \nu^* \cdot \lambda - \nu^* \cdot \mu(\xi) \geqslant 0,$$

where $\xi_{ij} = [G^r_{ij}(t_2) - G^r_{ij}(t_1)]/[t_2 - t_1] \geqslant 0$, $\sum_i \xi_{ij} \leqslant 1$, $\xi = \{\xi_{ij}\}$, and the inequality in the last display follows from (1).

The above calculation also implies the following fact which we record for future reference. It means, roughly, that the regulation process does not increase over some time interval if and only if each server performs only basic activities during that interval.

LEMMA 2. *For each value of the scaling parameter r, consider a pair of time points* $0 \leqslant t^r_1 < t^r_2 < \infty$, *and denote*

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

$$B^r_{1,j} \doteq \sum_{i \in I_j} \frac{G^r_{ij}(t^r_2) - G^r_{ij}(t^r_1)}{t^r_2 - t^r_1}.$$

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

Let us consider the process $Z^r = (Q^r, X^r, F^r, \hat{F}^r, S^r, G^r, H^r, Y^r)$, where

$$Q^r = (Q_i^r(t), t \geqslant 0, i \in I),$$
$$X^r = (X^r(t), t \geqslant 0),$$
$$F^r = (F_i^r(t), t \geqslant 0, i \in I),$$
$$\hat{F}^r = (\hat{F}_{ij}^r(t), t \geqslant 0, i \in I, j \in J),$$
$$S^r = (S_{ij}^r(t), t \geqslant 0, i \in I, j \in J),$$
$$G^r = (G_{ij}^r(t), t \geqslant 0, i \in I, j \in J),$$
$$H^r = (H_{ij}^r(t), t \geqslant 0, i \in I, j \in J),$$
$$Y^r = (Y^r(t), t \geqslant 0).$$

For each r, consider the fluid scaled process

$$\Gamma^r Z^r \doteq z^r = (q^r, x^r, f^r, \hat{f}^r, s^r, g^r, h^r, y^r),$$

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

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

From (19), we get

$$q_i^r(t) \equiv q_i^r(0) + f_i^r(t) - \sum_j \hat{f}_{ij}^r(t), \quad t \geqslant 0, \ i \in I. \tag{20}$$

Definition. A fixed set of functions $z = (q, x, f, \hat{f}, s, g, h, y)$ 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 $\mathcal{R}_f$,

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

and, in addition,

$$\|q(0)\| < \infty,$$
$$(f_i^r(t), t \geqslant 0) \to (\lambda_i t, t \geqslant 0) \quad \text{u.o.c.}, \ i \in I, \tag{21}$$
$$(s_{ij}^r(t), t \geqslant 0) \to (\mu_{ij} t, t \geqslant 0) \quad \text{u.o.c.}, \ i \in I, \ j \in J. \tag{22}$$

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

Lemma 3. *For any FSP z, all its component functions are Lipschitz continuous and, in addition,*

$$f_i(t) = \lambda_i t, \quad t \geqslant 0, \ i \in I,$$
$$s_{ij}(t) = \mu_{ij} t, \quad t \geqslant 0, \ i \in I, \ j \in J,$$
$$q_i(t) = q_i(0) + f_i(t) - \sum_j \hat{f}_{ij}(t), \quad t \geqslant 0, \ i \in I,$$
$$\hat{f}_{ij}(t) = \mu_{ij} g_{ij}(t), \quad t \geqslant 0, \ i \in I, \ j \in J,$$
$$x(t) = \nu^* \cdot q(t) = x(0) + y(t), \quad t \geqslant 0.$$

Furthermore, both $y(\cdot)$ and $x(\cdot)$ are nondecreasing (with $y(0) = 0$).

Because all component functions of an FSP are Lipschitz, they are absolutely continuous, and therefore at almost all points $t \in R_+$ (with respect to the Lebesgue measure) the following property holds: *Each component function of z has a (finite) first derivative.* 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 vector $q(t)$ corresponding to an FSP will sometimes be called its state at time t. The dynamics of the state q is governed by the differential (vector) equation

$$\frac{d}{dt} q(t) = \lambda - v(t), \tag{23}$$

which holds at every regular point t, and where $v(t) = [v_1(t), \ldots, v_I(t)]$, $v_i(t) \doteq \sum_j \hat{f}_{ij}'(t)$.

Remark. Our notion of an FSP (as well as that in Stolyar 2004) is such that we first formally define an FSP as a limit of a sequence of (scaled) sample paths of the original process, and then derive a certain set of its properties, some of which are straightforward and "easy to guess" (as those of Lemma 3), and some may be "harder to guess" (as those in the next subsection). For our proofs of state space collapse, which use Skorohod representation (and therefore involve limits of sample-path sequences), the notion of an FSP is more natural to use than the notion of a *fluid model solution* (FMS) employed in the proofs of state space collapse in Bramson (1998) and Williams (1998b). Using the FMS notion requires that a certain set of equations (defining an FMS) is "postulated," including in our case some "less obvious" properties like those described in Lemma 4(ii) and (iii) below; then one must verify that sample-path limits of the original process in fact satisfy this set of equations; and then the rest of the required properties of the limits is derived from the set of equations. In our case, we would need (in essence) to define FSPs anyway, and then derive all their properties proved in this and the next subsection. That is why it is natural to use the notion of FSP directly from the outset.

8.2. Uniform Attraction of Fluid Sample Paths

For $q \in R_+^I$, denote

$${}^*A(q) \doteq \max_i C_i'(q_i)/\nu_i^*, \quad {}_*A(q) \doteq \min_i C_i'(q_i)/\nu_i^*,$$

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

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

$$I^*(t) = \{i \in I | C_i'(q_i(t))/\nu_i^* = {}^*A(q(t))\}$$

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

$${}^*q_i(t) \doteq \{\zeta \geqslant 0 \,|\, C_i'(\zeta)/\nu_i^* = {}^*A(q(t))\},$$

and note that ${}^*q_i(t)$ is well defined because each function $C_i'(\cdot)$ is strictly increasing continuous. Let ${}^*x(t) \doteq \nu^* \cdot {}^*q(t)$, where ${}^*q(t) = ({}^*q_1(t), \ldots, {}^*q_I(t))$, and note that $x(t) \leqslant {}^*x(t)$ for all $t \geqslant 0$.

Finally, note that, at any time t, the following five conditions for $q(t)$ are all equivalent: $q(t)$ is a fixed point, ${}^*A(q(t)) = {}_*A(q(t))$, $\Phi(q(t)) = 0$, ${}^*x(t) = x(t)$, and ${}^*q(t) = q(t)$.

The following sequence of lemmas establishes further properties of FSPs, which are less obvious than the basic properties of Lemma 3. The form of the $Gc\mu$-rule is used in the proofs in an essential way.

LEMMA 4. *Consider a fixed FSP $q(\cdot)$. Suppose that $t > 0$ is a regular point and $q(t) \neq 0$. Then, the following properties hold at this t:*

(i) *$q_i(t) > 0$ for all $i \in I$.*

(ii) *We have*

$$\sum_{i \in I^*(t)} \nu_i^* q_i'(t) \leqslant 0, \quad \sum_{i \in I_*(t)} \nu_i^* q_i'(t) \geqslant 0. \tag{24}$$

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

$$\sum_{i \in I^*(t)} \nu_i^* q_i'(t) \leqslant -\epsilon_1, \quad \sum_{i \in I_*(t)} \nu_i^* q_i'(t) \geqslant \epsilon_1. \tag{25}$$

PROOF. We start by proving (iii). To this end, consider the case ${}^*A(q(t)) > {}_*A(q(t))$.

It follows from the CRP condition (including Lemma 1) and the $Gc\mu$-rule that

$$\sum_{i \in I^*(t)} \nu_i^* q_i'(t) \leqslant -\epsilon, \tag{26}$$

where $\epsilon > 0$ depends only on the subset $I^*(t)$. Indeed, let us denote by $J^*(t)$ the (nonempty) subset of servers j such that $I_j \cap I^*(t)$ is nonempty, and for each $j \in J^*(t)$, pick a representative element $i^* = i^*(j) \in I_j \cap I^*(t)$. (Below in this proof, i^* always means $i^*(j)$, i.e., it depends on $j \in J^*(t)$.) Note that for any $j \in J^*(t)$ and any $i \in I_j \cap I^*(t)$, we have $\nu_i^* \mu_{ij} = \nu_{i^*}^* \mu_{i^*,j}$. Then, we can write

$$\sum_{i \in I^*(t)} \nu_i^* q_i'(t) = \sum_{i \in I^*(t)} \nu_i^* \lambda_i - \sum_{i \in I^*(t)} \nu_i^* \sum_{j \in J} \mu_{ij} g_{ij}'(t)$$

and

$$\begin{aligned}
\sum_{i \in I^*(t)} \nu_i^* \sum_{j \in J} \mu_{ij} g_{ij}'(t) &\geqslant \sum_{j \in J^*(t)} \sum_{i \in I^*(t)} \nu_i^* \mu_{ij} g_{ij}'(t) \\
&= \sum_{j \in J^*(t)} \nu_{i^*}^* \mu_{i^*,j} \left[\sum_{i \in I^*(t)} g_{ij}'(t) \right] \\
&> \sum_{j \in J^*(t)} \nu_{i^*} \mu_{i^*,j} \left[\sum_{i \in I^*(t)} \phi_{ij} \right] \\
&= \sum_{j \in J^*(t)} \sum_{i \in I^*(t)} \nu_i^* \mu_{i,j} \phi_{ij} \\
&= \sum_{j \in J} \sum_{i \in I^*(t)} \nu_i^* \mu_{i,j} \phi_{ij} \\
&= \sum_{i \in I^*(t)} \nu_i^* \sum_{j \in J} \mu_{i,j} \phi_{ij} = \sum_{i \in I^*(t)} \nu_i^* \lambda_i.
\end{aligned}$$

The second (strict) inequality above is crucial. It follows from the fact that, according to the $Gc\mu$-rule and Lemma 1, for all sufficiently large r, in a small interval $[t, t + \epsilon]$, the prelimit path z^r is such that any server $j \in J^*(t)$ will only serve customers from the subset $I^*(t) \cap I_j$, and, therefore,

$$\sum_{i \in I^*(t)} g_{ij}'(t) = 1 \geqslant \sum_{i \in I^*(t)} \phi_{ij}.$$

Moreover, for at least one server $j \in J^*(t)$, a strict inequality must hold. Indeed, according to the CRP condition (namely, the fact that basic activities form a tree), at least one $j \in J^*(t)$ is such that $I_j \backslash I^*(t)$ is nonempty, and, therefore,

$$\sum_{i \in I^*(t)} \phi_{ij} < \sum_{i \in I_j} \phi_{ij} = 1.$$

We have proved (26), with $\epsilon > 0$ depending only on the subset $I^*(t) \subset I$. Because there is only a finite number of subsets of I, we have proved the first inequality in (25), with $\epsilon_1 > 0$ being the minimum of all possible ϵ. The second inequality in (25) is proved analogously.

The proof of the nonstrict inequalities in property (ii) is completely analogous to the proof of (iii), except the strict inequality in the long display above is replaced by the nonstrict one.

Finally, (i) is proved by contradiction. Suppose that $q_i(t) = 0$ for some $i \in I$. Obviously, the set of such i is exactly $I_*(t)$. Because $q(t) \neq 0$, $q(t)$ is not a fixed point. Therefore, the second inequality in (25) should hold. However, this is impossible, because we must have $q_i'(t) = 0$ for all $i \in I_*$. Indeed, the condition $q_i(t) = 0$ and the existence of $q_i'(t)$ imply that $q_i'(t) = 0$. (Otherwise $q_i(\cdot)$ would be negative just before or right after time t.) □

LEMMA 5. *Consider a fixed FSP $q(\cdot)$. Suppose that a time interval $[t_1, t_2]$, with $0 \leqslant t_1 < t_2$, is such that*

$$\min_{t_1 \leqslant t \leqslant t_2} \min_{i \in I} q_i(t) > 0.$$

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

$$\frac{d}{dt}[{}^*A(q(t))] \leqslant 0, \quad \frac{d}{dt}[{}_*A(q(t))] \geqslant 0,$$

$$\frac{d}{dt}[{}^*x(t)] \leqslant 0, \tag{27}$$

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

$$\frac{d}{dt}[{}^*x(t)] \leqslant -\epsilon_1, \tag{28}$$

where $\epsilon_1 > 0$ is defined in Lemma 4.

PROOF. First, the Lipschitz continuity of each function $C_i'(q_i(t))$ in $[t_1, t_2]$ follows from Lipschitz continuity of $q_i(\cdot)$ and the fact that $C_i''(\cdot)$ is continuous bounded away from both infinity and 0, for the range of possible values of $q_i(t)$ in $[t_1, t_2]$. (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 $\hat{I} \subseteq I$, the following functions are also Lipschitz continuous in $[t_1, t_2]$:

$$\max_{i\in\hat{I}} C_i'(q_i(t))/\nu_i^*, \quad \min_{i\in\hat{I}} C_i'(q_i(t))/\nu_i^*.$$

In particular, ${}^*A(q(t))$ and ${}_*A(q(t))$ are Lipschitz, which (along with the fact that the second derivatives $C_i''(\cdot)$ are bounded away from 0) implies that all ${}^*q_i(t)$ 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 $\hat{I} \subseteq I$, 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 (27) and (28) for this point t. Because t is strictly regular, the derivatives $(d/dt)[{}^*A(q(t))]$ and $(d/dt)[C_i'(q_i(t))/\nu_i^*]$ for $i \in I^*(t)$ are all equal. (In particular, this implies that ${}^*q_i'(t) = q_i'(t)$ for all $i \in I^*(t)$.) We cannot have $(d/dt)[{}^*A(q(t))] > 0$ because this would imply that $q_i'(t) > 0$ for all $i \in I^*(t)$, which would contradict (24). This proves the first (and with it the last) inequality in (27). The second inequality in (27) is proved analogously.

We can now write

$$\frac{d}{dt}[{}^*x(t)] = \sum_{i\in I} \nu_i^*\, {}^*q_i'(t) \leqslant \sum_{i\in I^*(t)} \nu_i^{**} q_i'(t) = \sum_{i\in I^*(t)} \nu_i^* q_i'(t),$$

where the inequality follows from the fact that ${}^*q_i'(t) \leqslant 0$ for all $i \in I$ (which is implied by (27)), and the equality is because ${}^*q_i'(t) = q_i'(t)$ for $i \in I^*(t)$. In the case ${}^*A(q(t)) > {}_*A(q(t))$, by (25), the right-hand side of the above display is bounded above by $-\epsilon_1$, which proves (28). □

LEMMA 6. *Consider a fixed FSP $q(\cdot)$. Suppose that $q(t_1) \neq 0$ for some $t_1 \geqslant 0$. Then, $q(t)$ has all strictly positive components (i.e., $q(t) \in R^I_{++}$) 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 $q(\xi) \neq 0$. By Lemma 4, $q(\xi) \in R^I_{++}$. Then, using Lemma 5, it follows that ${}_*A(q(t))$ is nondecreasing (and ${}^*A(q(t))$ and ${}^*x(t)$ are nonincreasing) starting from time ξ, and therefore $q(t) \in R^I_{++}$ for all $t \geqslant \xi$. Because ξ can be chosen arbitrarily close to t_1, the proof is complete. □

LEMMA 7. *Consider a fixed FSP $q(\cdot)$. If $q(0) = 0$, then $q(t) = 0$ for all $t \geqslant 0$.*

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, $q(t_1) \neq 0$. By Lemma 6, ${}^*x(t)$ cannot increase starting at time t_1, and therefore ${}^*x(t) \leqslant \epsilon$ for all $t \geqslant 0$. Because $\epsilon > 0$ can be chosen arbitrarily small, ${}^*x(t) = 0$, and therefore $q(t) = 0$ for all $t \geqslant 0$. □

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

THEOREM 3. *For any FSP, $\Phi(q(t))$ is a nonincreasing function, and the workload $x(t)$ is a nondecreasing function. Moreover, there exist fixed constants $T_1 > 0$ and $K \geqslant 1$ such that, for any FSP, $q(t)$ reaches a fixed point ${}^\circ q$ within finite time $x(0)T_1$ and then stays there, and $\nu^* \cdot {}^\circ q \leqslant x(0)K$.*

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

Suppose that $q(0) \neq 0$. By Lemma 6, ${}_*A(q(t))$ is nondecreasing and ${}^*A(q(t))$ is nonincreasing in $[0, \infty)$, and therefore $\Phi(q(t))$ is nonincreasing. Further, by Lemma 6, $q(t) \in R^I_{++}$ for all $t > 0$. Then, by Lemma 5, for almost all $t > 0$, ${}^*x(t) > x(t)$ implies

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

Because ${}^*x(0) \leqslant x(0)[\sum_i \nu_i^*]/[\min_n \nu_n^*]$, $x(t) \leqslant {}^*x(t)$, and $x(t)$ is nondecreasing, we immediately see that $q(t)$ must reach a fixed point within a time proportional to $x(0)$.

Therefore, the statement of the theorem, with some fixed $T_1 > 0$ and $K \geqslant 1$, holds for the FSPs with $q(0) \neq 0$. By Lemma 7, it trivially holds for $q(0) = 0$ as well. □

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

LEMMA 8. *There exists a constant $\epsilon_2 > 0$, such that the following holds. For any prelimit (scaled) path $q^r = (q^r(t), t \geqslant 0)$ and $0 \leqslant t_1^r < t_2^r < \infty$, the property*

$$q^r(t) \neq 0 \quad \text{and} \quad \Phi(q^r(t)) \leqslant \epsilon_2 \quad \forall\, t \in [t_1^r, t_2^r]$$

implies that in the (scaled) interval $[t_1^r, t_2^r]$, each server j can take for new service only customers of types $i \in I_j$.

PROOF. The small value of $\Phi(q)$ implies that the vector $(C_1'(q_1), \ldots, C_I'(q_I))$ is "almost proportional" to ν^*. So, if $\Phi(q)$ is small, it follows directly from the form of the $Gc\mu$-rule and Lemma 1, that each server j can only start service of customers $i \in I_j$. We omit the ϵ-δ formalities. □

9. Proof of Theorem 1

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

$$\tilde{\Gamma}^r(Q^r, X^r, F^r, S^r, G^r, H^r, Y^r) \doteq (\tilde{q}^r, \tilde{x}^r, \tilde{f}^r, \tilde{s}^r, \tilde{g}^r, \tilde{h}^r, \tilde{y}^r),$$

where the diffusion scaling opertator $\tilde{\Gamma}^r$ is applied componentwise, and acts on a scalar function $\Xi = (\Xi(t), t \geqslant 0)$ as follows:

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

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 Stolyar (2004), 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, Ethier and Kurtz 1986), for each i the sequence of the input processes $\{F_i^r\}$ can be constructed on a probability space such that the convergence in (5) holds u.o.c. with probability 1 (w.p.1). Similarly, the sequence of service processes $\{S^r\}$ can be constructed on a probability space such that w.p.1., u.o.c. the FCLT in (6) holds for each pair (i, j):

$$(\tilde{s}_{ij}^r(t) - \mu_{ij} r t, t \geqslant 0) \xrightarrow{\text{u.o.c.}} (\sigma_{ij}^2 B(t), t \geqslant 0). \tag{29}$$

We can and do assume that our underlying probability space $\Omega = \{\omega\}$ is a direct product of the above probability spaces.

Now, from condition (4) and Bramson's weak law estimates (Bramson 1998, Proposition 4.3), we know that for any $T_3 > 0$, any $\epsilon > 0$, and any (i, j), for all large r, we have (see the proof of property (5.19) in Proposition 5.1 of Bramson 1998)

$$P\left\{\max_{0 \leqslant l \leqslant T_3 r} \sup_{0 \leqslant \xi \leqslant 1} |f_i^r(l+\xi) - f_i^r(l) - \lambda_i \xi| \geqslant \epsilon\right\} < \epsilon \tag{30}$$

and

$$P\left\{\max_{0 \leqslant l \leqslant T_3 r} \sup_{0 \leqslant \xi \leqslant 1} |s_{ij}^r(l+\xi) - s_{ij}^r(l) - \mu_{ij} \xi| \geqslant \epsilon\right\} < \epsilon. \tag{31}$$

(The max in (30) and (31), as well as below in (32) and (33), is over integers $l \in [0, T_3 r]$.)

These estimates enable one 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

$$\max_{0 \leqslant l \leqslant T_3 r} \sup_{0 \leqslant \xi \leqslant 1} |f_i^r(l+\xi) - f_i^r(l) - \lambda_i \xi| \to 0 \quad \forall\, i \tag{32}$$

and

$$\max_{0 \leqslant l \leqslant T_3 r} \sup_{0 \leqslant \xi \leqslant 1} |s_{ij}^r(l+\xi) - s_{ij}^r(l) - \mu_{ij} \xi| \to 0 \quad \forall\, (i, j). \tag{33}$$

Property (33) in turn implies the following property. *With probability* 1, *for any fixed* $T_4 > 0$ *and* $d > 0$, *uniformly on any sequence of pairs* (t_1^r, t_2^r), $r \in \mathcal{R}_2$, *such that* $0 \leqslant t_1^r < t_2^r \leqslant r^2 T_4$, $t_2^r - t_1^r \geqslant rd$,

$$\lim_{r \to \infty,\, r \in \mathcal{R}_2} \frac{S_{ij}^r(t_2^r) - S_{ij}^r(t_1^r)}{\mu_{ij}(t_2^r - t_1^r)} = 1. \tag{34}$$

For each i, we have

$$Q_i^r(r^2 t) = Q_i^r(0) + F_i^r(r^2 t) - \sum_j S_{ij}^r(G_{ij}^r(r^2 t)),$$

and, therefore, the expression for the scaled workload can be written as follows:

$$\tilde{x}^r(t) = \tilde{x}^r(0) + r^{-1}\sum_i \nu_i^*(F_i^r(r^2 t) - \lambda_i^r r^2 t) \tag{35}$$

$$+ r^{-1}\sum_i \nu_i^*\left(\lambda_i^r r^2 t - \sum_j \phi_{ij}\mu_{ij} r^2 t\right) \tag{36}$$

$$+ r^{-1}\sum_i \nu_i^* \sum_j (\phi_{ij}\mu_{ij} r^2 t - S_{ij}^r(\phi_{ij} r^2 t)) \tag{37}$$

$$+ r^{-1}\sum_i \nu_i^* \sum_j \big(S_{ij}^r(\phi_{ij} r^2 t) - S_{ij}^r(\phi_{ij} r^2 t - H_{ij}^r(r^2 t))\big) \tag{38}$$

$$= \widetilde{w}^r(t) + \breve{y}^r(t), \tag{39}$$

where $\widetilde{w}^r(t)$ denotes the sum of the first four terms, and $\breve{y}^r(t)$ denotes the last term (38). We know that

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

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 (11). (The realization $\widetilde{w}(\cdot)$ is, of course, continuous.) As seen from (39), the key step in proving Theorem 1 will be the proof of the following convergence:

$$(\breve{y}^r(t), t \geqslant 0) \xrightarrow{\text{u.o.c.}} (\tilde{y}(t), t \geqslant 0). \tag{40}$$

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 9. *Consider a fixed* $\omega \in \Omega_2$. *As* $r \to \infty$ *along* $\mathcal{R}_2$, *the functions* $\breve{y}^r$ *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* $t \in [0, T_4]$,

$$(1 - \delta_1)\tilde{y}^r(t) - \delta_2 \leqslant \breve{y}^r(t) \leqslant (1 + \delta_1)\tilde{y}^r(t) + \delta_2. \tag{41}$$

As can be seen from its proof, Lemma 9 applies to any scheduling discipline satisfying Conditions (d1) and (d2). Also, we note that in the proof of Lemma 9, the uniqueness of ϕ (in the CRP condition) is used in an essential way.

PROOF. Let us fix $T_4 > 0$. Suppose that we have a subsequence of r such that the following conditions hold: time $t \in [0, T_4]$ for all r (with t generally speaking depending on r); for each (ij), $H^r_{ij}(r^2t)$ converges to either a finite number or $+\infty$ or $-\infty$; $\max_{ij} |H^r_{ij}(r^2t)|$ converges to either a finite (nonnegative) number or $+\infty$; if $\max_{ij} |H^r_{ij}(r^2t)| \to +\infty$, then, for each (ij),

$$\frac{H^r_{ij}(r^2t)}{\max_{ij} |H^r_{ij}(r^2t)|} \to \eta_{ij}, \tag{42}$$

where $\max_{ij} |\eta_{ij}| = 1$. The proof will be complete if we can prove that (41) holds for all large r along such a subsequence.

Consider a fixed pair (i, j) in (38). If $\lim |H^r_{ij}(r^2t)|/r > 0$, then, by (34),

$$\frac{S^r_{ij}(\phi_{ij} r^2 t) - S^r_{ij}(\phi_{ij} r^2 t - H^r_{ij}(r^2t))}{\mu_{ij} H^r_{ij}(r^2t)} \to 1. \tag{43}$$

If $\lim |H^r_{ij}(r^2t)|/r = 0$, then, again by (34),

$$[S^r_{ij}(\phi_{ij} r^2 t) - S^r_{ij}(\phi_{ij} r^2 t - H^r_{ij}(r^2t))]/r \to 0 \quad \text{and}$$
$$\mu_{ij} H^r_{ij}(r^2t)/r \to 0; \tag{44}$$

in this case, the contribution of the pair (i, j) to both $\breve{y}^r(t)$ and $\tilde{y}^r(t)$ vanishes. These two observations easily imply (41) for all large r, in the case when $\max_{ij} |H^r_{ij}(r^2t)|/r$ converges to a finite number.

Now, consider the case $\max_{ij} |H^r_{ij}(r^2t)|/r \to +\infty$. Let us define τ^r (which depends on r) as follows:

$$\tau^r = \left[\max_{ij} |H^r_{ij}(r^2t)|\right]/\kappa, \tag{45}$$

where $\kappa = \min_{ij} \kappa_{ij}$ and $\kappa_{ij} = \phi_{ij} \wedge (1 - \phi_{ij})$ if $0 < \phi_{ij} < 1$, and $\kappa_{ij} = 1$ otherwise. (Note that $\kappa > 0$.) Obviously,

$$\tau^r/r \to +\infty. \tag{46}$$

For each (ij), let us define ξ^r_{ij} by the following equation:

$$\phi_{ij} - \xi^r_{ij} = H^r_{ij}(r^2t)/\tau^r. \tag{47}$$

We see from (42) and the definitions of τ^r and ξ^r_{ij}, that

$$\xi^r_{ij} \to \xi_{ij} = \phi_{ij} - \eta_{ij}\kappa \quad \forall (ij).$$

From the definition of κ and the fact that $\max_{ij} |\eta_{ij}| = 1$ (and also the facts that $\phi_{ij} = 0$ implies $H^r_{ij}(r^2t) \leqslant 0$, and $\phi_{ij} = 1$ implies $H^r_{ij}(r^2t) \geqslant 0$), we see that $\xi_{ij} \in [0, 1]$ for all (ij). For any j, we have $\sum_i \xi_{ij} \leqslant 1$ because $\sum_i H^r_{ij}(r^2t) \geqslant 0$. Thus, matrix ξ is "column-substochastic," like matrix ϕ, and $\xi \neq \phi$.

We can write

$$\begin{aligned}(r/\tau^r)\tilde{y}^r(t) &= (1/\tau^r) \sum_i \sum_j \nu^*_i H^r_{ij}(r^2t)\mu_{ij} \qquad (48)\\ &= \sum_i \sum_j \phi_{ij}\nu^*_i\mu_{ij} \\ &\quad - \sum_i \sum_j (\phi_{ij} - H^r_{ij}(r^2t)/\tau^r)\nu^*_i\mu_{ij} \\ &\to \sum_i \nu^*_i(\mu_i(\phi) - \mu_i(\xi)) \\ &= \nu^* \cdot (\mu(\phi) - \mu(\xi)) \geqslant 0. \qquad (49)\end{aligned}$$

The inequality in (49) follows from the CRP condition, which also implies that $\mu(\phi) \neq \mu(\xi)$ (because $\xi \neq \phi$).

Let us show that the case $\nu^* \cdot (\mu(\phi) - \mu(\xi)) = 0$ is impossible. Suppose that this equality does hold. Then, $\mu_i(\phi) < \mu_i(\xi)$ for at least one i. For such i, we then have

$$(1/\tau^r) \sum_j H^r_{ij}(r^2t)\mu_{ij} \to \mu_i(\phi) - \mu_i(\xi) < 0, \tag{50}$$

which, using (43) and (44), implies

$$\begin{aligned}&(1/\tau^r) \sum_j (S^r_{ij}(\phi_{ij} r^2 t) - S^r_{ij}(\phi_{ij} r^2 t - H^r_{ij}(r^2t))) \\ &\quad \to \mu_i(\phi) - \mu_i(\xi) < 0. \qquad (51)\end{aligned}$$

This in turn, easily implies that $\tilde{q}^r_i(t) \to -\infty$, which is, of course, impossible. Thus, we must have $\nu^* \cdot (\mu(\phi) - \mu(\xi)) > 0$, that is, the expression in (48) converges to a strictly positive finite constant. This, again using (43) and (44), implies (41) for all large r. □

It follows from Lemma 9 that, to prove (40), it suffices to prove

$$(\tilde{y}^r(t), t \geqslant 0) \xrightarrow{\text{u.o.c.}} (\tilde{y}(t), t \geqslant 0) \tag{52}$$

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

Because 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

$$\tilde{y}^r \Rightarrow \tilde{y}, \tag{53}$$

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, $\tilde{y} \in D([0, \infty), \bar{R})$. The notation "$\Rightarrow$" stands for convergence at every point of continuity of the limit function except maybe the point 0.) We note that (53) implies that

$$\tilde{x}^r \Rightarrow \tilde{x} \doteq \tilde{w} + \tilde{y}, \tag{54}$$

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

The following lemma and its proof are analogous to Lemma 7 in Stolyar (2004); it contains key observations that are used in the proof of Theorem 1.

LEMMA 10. *Suppose that the scheduling rule in the system is $Gc\mu$. Suppose that $\omega \in \Omega_2$ and a subsequence $\mathcal{R}_4(\omega) \subseteq \mathcal{R}_2$ are fixed such that, along this subsequence,* (53) *holds. Suppose that a sequence $\{\tilde{t}^r, r \in \mathcal{R}_4(\omega)\}$ is fixed such that*

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

and

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

Let $\delta > 0$ be fixed, and

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

Then,

(a) $\tilde{y}$ *(and $\tilde{x}$) is finite in $[0, t' + \delta]$. This in particular, means that for any (i, j), the functions $\tilde{h}^r_{ij}(\cdot)$ remain uniformly bounded in the interval $[0, t' + \delta]$ as $r \to \infty$.*

(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*:

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

with K defined in Theorem 3.

(d) *For any $\delta' > 0$,*

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

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

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

(c′) $\tilde{x}(t') = C$ *and, consequently, $\tilde{q}(t') = {}^\circ q$.*

(d′) *The following holds*:

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

PROOF. As in Stolyar (2004), the key construction in this proof, namely the construction of a set of processes $\bar{x}^{r,l}(\cdot)$ (see below) on a slower, fluid time scale, essentially follows Bramson's construction in §5 of Bramson (1998).

Let us consider the functions of interest on the fluid time scale. Namely, consider the earlier defined function $x^r(t) \equiv \tilde{x}^r(t/r)$, $t \geqslant 0$, and similarly defined functions y^r, 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 an arbitrary

$$T \geqslant C_3 T_1,$$

where K and T_1 are the constants defined in Theorem 3. As seen below in the proof, C_3 will be the upper bound of $\tilde{x}^r(\cdot)$ in the interval $[\tilde{t}^r, \tilde{t}^r + \delta]$ or, equivalently, the upper bound of $x^r(\cdot)$ 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 its initial 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

$$\bar{x}^{r,l}(u) \doteq x^r(r\tilde{t}^r + Tl + u), \quad u \geqslant 0,$$

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 8. Then, the following property holds.

PROPERTY 1. For all sufficiently large r, for all integer $l \in [1, 2\delta r/T]$, we have

$$\Phi(\bar{q}^{r,l}(0)) \leqslant \epsilon_4 \tag{55}$$

and

$$\Phi(\bar{q}^{r,l}(u)) \leqslant 2\epsilon_4 \quad \forall\, u \in [0, T]. \tag{56}$$

Indeed, suppose that Property 1 does not hold. For each r, define $l' = l'(r)$ as follows: $l' = 0$ if condition (55) is violated for $l = 1$; otherwise, l' is the smallest $l \geqslant 1$ such that condition (55) holds, but (56) does not hold for at least one $u \in (0, T]$.

Note that if $l' \geqslant 2$, then, by our construction, (56) holds for each $l = 1, \ldots, l' - 1$. This, by Lemma 8, implies that for each $l = 1, \ldots, l' - 1$, in the (fluid scaled) interval $[r\tilde{t}^r + Tl, r\tilde{t}^r + T(l+1)]$ (corresponds to the interval $[r^2\tilde{t}^r + rTl, r\tilde{t}^r + rT(l+1)]$ in the unscaled time), each new service started by a server j must be given to a customer of a flow $i \in I_j$. This does not immediately imply that

$$\bar{y}^{r,l}(T) - \bar{y}^{r,l}(0) = 0 \tag{57}$$

for $l = 1, \ldots, l' - 1$, because if some "anomalous" customers (with types $i \notin I_j$) were in service at time $r\tilde{t}^r + Tl$, then their service must continue until completion. However, using property (33) of the service processes, it is easy to see that for all large r, the service of all "anomalous" customers will be completed in the first such interval $[r\tilde{t}^r + T, r\tilde{t}^r + 2T]$; moreover $[\bar{g}^{r,1}_{ij}(T) - \bar{g}^{r,1}_{ij}(0)]/T \to 0$ as $r \to \infty$ for any (i, j) such that $i \notin I_j$. Thus, for all large r, (57) holds for each $l = 2, \ldots, l' - 1$, and (for $l = 1$)

$$[\bar{y}^{r,1}(T) - \bar{y}^{r,1}(0)]/T \to 0. \tag{58}$$

Our choice of the constant T and Theorem 3 imply that for all sufficiently large r (and l' being a function of r as defined above),

$$C - \epsilon - \epsilon_3 \leqslant \bar{x}^{r,l'}(0) \leqslant C_3. \tag{59}$$

To prove (59), we first observe that for $l = 1$, we have

$$C - \epsilon_3 \leqslant \bar{x}^{r,\,l}(0) \leqslant (C + \epsilon_3)K. \tag{60}$$

Indeed, if the upper bound in (60) does not hold, we would be able to find a subsequence $\{r\}$ along which the sequence of paths $\bar{z}^{r,\,0}$ converges to an FSP z with $x(0) = C$ and $x(T) > CK$, which contradicts Theorem 3. The lower bound in (60) holds for a similar reason—otherwise we could construct an FSP with $x(0) = C$ and $x(T) < C$. Then, (59) follows from combining the following four facts:

$|\widetilde{w}(\xi_1) - \widetilde{w}(\xi_2)| \leqslant \epsilon$ as long as $\xi_1, \xi_2 \in [t' - 3\delta, t' + 3\delta] \cap R_+$;

$\widetilde{w}^r \to \widetilde{w}$ uniformly in $[t' - 3\delta, t' + 3\delta] \cap R_+$;

condition (57) for each $l = 2, \ldots, l' - 1$, and condition (58);

the functions $\tilde{y}^r$ and $\check{y}^r$ are asymptotically close (in the sense of (41)).

Given the bound (59) on $\bar{x}^{r,\,l'}(0)$, we immediately obtain a contradiction to Theorem 3: we can find a subsequence $\{r\}$ along which the sequence of paths $\bar{z}^{r,\,l'}$ converges to an FSP z with $\Phi(0) \leqslant \epsilon_4$ and $\Phi(\xi) \geqslant 2\epsilon_4$ for some $\xi \in [0, T]$. We have thus proved Property 1.

Property 1, and an argument completely analogous to the one we have used in its proof (in particular, in the proof of (59)), imply that, for all large r,

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

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

To prove (d), we first note that (a), (b), and (41) imply the following uniform convergence for the workload process:

$$(\tilde{x}^r(t), t \in [t' + \delta', t' + \delta]) \xrightarrow{\text{u.o.c.}} (\tilde{x}(t), t \in [t' + \delta', t' + \delta]). \tag{61}$$

Statement (d) then follows from Property 1, the fact that ϵ_4 can be chosen arbitrarily small, and convergence (61).

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

9.1. Proof of Theorem 1: Part 1

This proof repeats that of Theorem 1(i) in Stolyar (2004) virtually verbatim. We reproduce it here for completeness.

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:

$$(\tilde{y}^r(t), t \geqslant 0) \xrightarrow{\text{u.o.c.}} (\tilde{y}(t), t \geqslant 0), \tag{62}$$

where $\tilde{y}$ is defined by (10), and

$$(\tilde{q}^r(t), t \geqslant 0) \xrightarrow{\text{u.o.c.}} (\tilde{q}(t), t \geqslant 0), \tag{63}$$

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

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 (53) 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 (10) is what needs to be proved to establish (63).

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 Equation (10).

Step 5. Convergence (63) holds.

In this proof, we will use the convention that $\tilde{y}(0-) = 0$ and $\widetilde{w}(0-) = \tilde{x}(0-) = \widetilde{w}(0)$. 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

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

PROOF OF STEP 1. Suppose that the statement does not hold. Denote $t^* = \inf\{t \geqslant 0 \mid \tilde{y}(t) = \infty\}$. 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

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

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

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

(We must have $t' \leqslant t^*$, because the limit function $\tilde{y}(t)$, and therefore $\tilde{x}(t)$, is infinite for all $t \geqslant t^*$.) It is also easy to see (from (32)) that

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

The conditions of Lemma 10 are satisfied, and so $\tilde{y}$ is bounded in $[t', t' + \delta]$—a contradiction, because $t' + \delta > t^*$. Step 1 has been proved. □

PROOF OF STEP 2. Suppose that the statement does not hold. The contradiction is obtained 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., $\tilde{y}(t^*-) < \tilde{y}(t^*)$. Because $\tilde{x} = \widetilde{w} + \tilde{y}$ and $\widetilde{w}$ is continuous, $\tilde{x}(t^*) - \tilde{x}(t^*-) = \tilde{y}(t^*) - \tilde{y}(t^*-)$. There are two possible cases:

(a) $\tilde{x}(t^*-) > 0$, and

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

Case (a). In this case, we must have $t^* > 0$. (Indeed, by the definition of $\tilde{w}$ and our conventions, $\tilde{x}(0-) = \tilde{w}(0) = \lim_r \tilde{x}^r(0)$. If $\tilde{w}(0) > 0$, then, by Lemma 10(c′), $\tilde{x}(0) = \lim_r \tilde{x}^r(0)$, which means that $\tilde{x}$, and therefore $\tilde{y}$, has no jump at 0. If $\tilde{w}(0) = 0$ 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 $\tilde{x}^r(t') \to C$ (because $\tilde{x}$ is continuous at t'), and by Lemma 10 $\tilde{y}$ 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 $K \geqslant 1$ is defined in Theorem 3 (and used in Lemma 10). Then, if $t^* > 0$, we fix a small Δt such that

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

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

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

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

Proof of Step 3. Let $t^* \geqslant 0$ 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 10(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. The proof follows from the statements of Steps 2 and 3 and Proposition 1 (in the appendix). □

Proof of Step 5. It suffices to show that for any $t^* \geqslant 0$ and any $\epsilon > 0$, there exists $\delta > 0$ such that

$$\limsup_{r\to\infty} \sup_{\xi\in[t^*-\delta,\, t^*+\delta]\cap R_+} \|\tilde{q}^r(\xi) - \tilde{q}(\xi)\| < \epsilon. \tag{64}$$

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

If $\tilde{x}(t^*) = 0$, then (64) must hold because both functions $\tilde{q}$ and $\tilde{q}^r$ (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 (64) follows from Lemma 10(d′). If $\tilde{x}(t^*) > 0$ and $t^* > 0$, then to obtain (64) we can repeat the construction of the proof of Step 2(a) and then apply Lemma 10(d). Step 5 has been proved. □

Thus, the proof of Property 2, and with it the proof of Part 1 of the theorem, is complete. □

9.2. Proof of Theorem 1: Part 2

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

We claim that, along the subsequence $\mathcal{R}_2$,

$$\liminf_{r\to\infty} \tilde{x}^r_G(t) \geqslant \tilde{x}(t), \quad t \geqslant 0, \tag{65}$$

and therefore (12) holds. To prove this we first observe that property (41) holds for any discipline G. 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.) Property (41) implies that we have $\check{y}^r_G \Rightarrow \tilde{y}_G$ as well. Therefore, for any $t \geqslant 0$ where $\tilde{y}_G(\cdot)$ is continuous, as $r \to \infty$ along $\mathcal{R}_4(\omega)$,

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

Because $\tilde{x}^r_G(t)$ is nonnegative, we see that $\tilde{w} + \tilde{y}_G$ is nonnegative at every point of continuity of $\tilde{y}_G$, and therefore it is nonnegative for all $t \geqslant 0$ (by right continuity). Then, by Proposition 1(ii) (in the appendix), $\tilde{y}_G(t) \geqslant \tilde{y}(t)$ for all $t \geqslant 0$. This (and property (41)) implies that $\liminf \check{y}^r_G(t) \geqslant \tilde{y}(t)$ for any $t \geqslant 0$. Thus, (65) holds along the subsequence $\mathcal{R}_4(\omega)$, and therefore along $\mathcal{R}_2$ (because the subsequence $\mathcal{R}_3(\omega)$ can be arbitrary). The proof of (65) (and therefore (12)) is complete.

Because the function $\sum_i C_i(q_i)$ is continuous in the vector q, and the fixed point $\tilde{q}(t)$ in (13) minimizes the value of $\sum_i C_i(q_i)$ over vectors q with workload $\tilde{x}(t)$, property (13) also holds. Finally, the equality in (14) follows from the fact that $\tilde{q}^r \to \tilde{q}$ u.o.c., and the inequality follows from (13) and Fatou's lemma.

The proof of Theorem 1 is now complete. □

10. Further Research and Applications

There are many directions in which the present work can be applied or extended. In this concluding section, we outline some of them.

Extending the modeling scope—feedback. Many features can be added to our model and still hopefully, leave it tractable. For example, adding Markovian feedback (as in Klimov 1974, 1978; Glazebrook and Nino-Mora 2001): upon service completion, a type i customer could immediately return for service, turning into type k with probability P_{ik}. (To simplify the discussion, suppose that P_{ik}'s do not depend on the server that performs the service.) Here, the substochastic matrix $P = [P_{ik}]$ is assumed to

have spectral radius less than unity, which guarantees that all customers eventually leave. Just as the applicability of MaxWeight-type scheduling rules extends to networks (see the maximum throughput policy in Tassiulas and Ephremides 1992), the $Gc\mu$-rule should extend to our network with feedback along the same lines. Specifically, when becoming free at time t, server j takes for service a type i customer such that

$$i \in \arg\max_i \left[C_i'(Q_i(t)) - \sum_k P_{ik} C_k'(Q_k(t)) \right] \mu_{ij},$$

unless the maximum (in the arg max above) in nonpositive, in which case the server remains idle as long as this nonpositivity condition holds. (Note that such a $Gc\mu$-rule does need to know the matrix P.) We believe that this version of a $Gc\mu$-rule retains asymptotic optimality in the (appropriately defined) heavy traffic regime, and that the basic approach developed in Stolyar (2004) and the present paper can be applied to demonstrate it. We leave this as a subject for future research.

Extending the modeling scope—large number of servers. Under what conditions do our results still apply as the number of servers grows indefinitely? Consider, as before, a sequence of queueing systems indexed by r, $r \to \infty$. All systems still have I customer types and J server skills, but now the number of servers grows linearly with r. For simplicity, let r itself be the total number of servers in the rth system, which are divided so that there are $\alpha_j r$ servers of skill j; $\alpha_j > 0$, $\sum_j \alpha_j = 1$. Suppose that the service time of a type i customer by a skill j server is exponential with parameter μ_{ij} for all r. Assume also renewal arrivals, with

$$\Lambda_i^r \doteq r\lambda_i^r = r\lambda_i + b_i$$

being the arrival rate of type i to the rth system. Then, if we "slow down" time by a factor r, we essentially obtain a system with input rates $\lambda_i^r = \lambda_i + b_i/r$, and service rates by server group j being $\alpha_j \mu_{ij}$. With a minimal additional argument, it can be shown that our results apply as is to this modified system under the diffusion scaling $(1/r)Z^r(r^2 t)$. But, this means that our results apply to the sequence of original systems with the scaling $(1/r)Z^r(rt)$. In the case of general service times, the reduction to our setting is not straightforward and requires further thought.

Linear waiting costs. In practice, waiting costs are acknowledged as being typically nonlinear (see Van Mieghem 1995, Zohar et al. 2002 and references there). Yet, the traditional scheduling literature ($c\mu$ etc.), as well as that which is approximation based (e.g., Harrison 1998, Harrison and Lopez 1999, Williams 1998c, Bell and Williams 2001), have both focused on linear delay costs. (A conceivable reason is that nonlinear costs have been viewed as not being amenable to analysis.) While linear costs violate our base assumption that $C_i'(0+) = 0$, they can still be accommodated within our framework, as will now be described.

(a) *Theory.* Let $C_i(\zeta) = c_i\zeta, , \zeta \geqslant 0$, where $c_i > 0$ are given constant cost rates. Assume that ν^* is known (either precomputed or estimated). Then, we conjecture that the following adjustment of the $Q\mu$-rule in (15) is asymptotically optimal. First, determine the "cheapest" queue i, via $i \in \arg\min_i (c_i/\nu_i^*)$; then set all γ_k, $k \neq i$ to positive constants that do not depend on r (say, all are set to 1); finally, $\gamma_i = \gamma_i(r)$ is a "sufficiently slowly" decreasing function of r (for example, $\gamma_i(r) = (\log r)^{-c}$ with any $c > 1$, or any other reciprocal of the threshold that applies in Bell and Williams 2001).

(b) *Application.* Because we do not know ν^* in advance and, moreover, it changes with circumstances (for example, as arrival rates change), we can approximate linear costs by the costs $C_i(\zeta) = c_i\zeta^{1+\epsilon}$ for some small $\epsilon > 0$. Another option is the following adaptive procedure: use the $Q\mu$-rule in (15), periodically estimate ν^* via measurements (as described below), and shift more workload into the currently "cheapest" queue $i \in \arg\min_i (c_i/\nu_i^*)$, by resetting its γ_i to a smaller value.

Estimating waiting time in real time. Suppose that it is desired to estimate the waiting time of a type i customer upon arrival in real time. This has obvious practical significance (e.g., Whitt 1999). Our results suggest the following very simple procedure, applicable under $Gc\mu$, that is based on the snapshot principle (Reiman 1984, 1988): use, as an estimate, the waiting time (age) of the longest-waiting type i customer.

On-line estimation of ν^*. The $Gc\mu$-rule does not require any knowledge of the workload contributions ν^*. Nevertheless, their (relative) values can be estimated by observing the ratios of $C_i'(Q_i)/C_k'(Q_k)$, as the system approaches heavy traffic. Under $Gc\mu$, these ratios approximate those of ν_i^*'s.

Appendix

Sketch of the Proof of Theorem 2

As in the formulation of the theorem (and the formulation and proof of Theorem 1), the variables pertaining to a specific discipline G are appended with the additional subscript G. When such a subscript is omitted, the corresponding variable pertains specifically to our D-$Gc\mu$ discipline.

Recall that we consider the class of disciplines satisfying Conditions (d1) and (d2). Therefore, without loss of generality, we can adopt the convention that, for each queue-server pair (i, j), the i.i.d. sequence $v_{ij}(n), n = 1, 2, \ldots,$ defines service times of type i customers by server j, *in the order in which they are taken from the queue for service* (and not in the order of their arrival to the system).

Consider any discipline G satisfying Conditions (d1) and (d2). Let us modify it so that the following Condition (d3) holds as well,

CONDITION (D3). When queue i is chosen for service (by any server j), the longest-waiting customer from that queue is taken for service.

More specifically, the modification of discipline G is such that whenever server j picks queue i for service (according to whatever condition the discipline G uses), it always takes the longest-waiting customer from the queue. (As we noted before, such a modification does not affect the queue length process.) Then, with probability 1 (in fact, for every realization of the process), this modification can only decrease the value of $\mathscr{C}^r_G(T)$ for any $T \geqslant 0$ (and any r). This easily follows from convexity of the cost functions $C_i(\cdot)$, using an argument completely analogous to the one used in Van Mieghem (1995) for the single server system. By definition, the D-$Gc\mu$ discipline satisfies Conditions (d1)–(d3). Therefore, we conclude that to prove the theorem, we only need to consider disciplines satisfying all three Conditions (d1)–(d3).

We use the same construction of the common probability space for all the processes with different r, as in the proof of Theorem 1. The additional process we consider (and include as a component of Z^r) is

$$D^r = ((D^r_i(t), t \geqslant 0), i \in I),$$

denoting its fluid- and diffusion-scaled versions by d^r and $\tilde{d}^r$, respectively.

Property (32) implies the following key property, which can be called the "instantaneous Little's law," and which establishes the connection between D-$Gc\mu$ and $Gc\mu$ rules: *For any discipline G (satisfying Conditions* (d1)–(d3)*), as $r \to \infty$ along $\mathscr{R}_2$, with probability* 1*, for any $T_3 > 0$ and any $\epsilon > 0$, we have*

$$\max_{i \in I} \sup_{0 \leqslant t \leqslant T_3} \frac{|\tilde{q}^r_{i,G}(t) - \lambda_i \tilde{d}^r_{i,G}(t)|}{\max(\tilde{q}^r_{i,G}(t), \epsilon)} \to 0. \tag{66}$$

We define FSPs (for D-$Gc\mu$) the same way as for $Gc\mu$, except that it will have the additional component d (as a limit of d^r), and the following additional conditions must hold:

$$d^r_i(0) \to d_i(0) = q_i(0)/\lambda_i \quad \forall\, i$$

and

$$(k^r_i(\xi), \xi \geqslant 0) \xrightarrow{\text{u.o.c.}} (\min\{\lambda_i \xi, q_i(0)\}, \xi \geqslant 0) \quad \forall\, i,$$

where $k^r_i(\xi)$ is the (fluid-scaled) number of type i customers with the (fluid-scaled) sojourn times (at initial time $t = 0$) not exceeding ξ.

With such a definition of an FSP, it is easy to prove the following additional basic property:

$$d_i(t) = q_i(t)/\lambda_i \quad \forall\, t \geqslant 0,\ \forall\, i.$$

Using this property, all the FSP properties established for the $Gc\mu$-rule are proved the same way for the D-$Gc\mu$-rule, as long as the cost functions $C_i(\cdot)$ are replaced by $\bar{C}_i(\cdot)$. Using the fact that FSPs under D-$Gc\mu$ satisfy all the properties of FSPs under the $Gc\mu$-rule, and using the key property (66), Part 1 of the theorem is proved the same way as Part 1 of Theorem 1. Properties (16) and (17) of Part 2 are proved the same way as the corresponding properties in Part 2 of Theorem 1. To prove the equality in (18) (for D-$Gc\mu$), we prove that, uniformly on $t \in [0, T]$,

$$\lim_{r \to \infty} \left| r(\mathscr{C}^r(t + 1/r) - \mathscr{C}^r(t)) - \sum_i \bar{C}_i(\tilde{q}_i(t)) \right| = 0. \tag{67}$$

The proof of (67) uses (probability 1) properties (33), (32), (66), properties of the FSPs, and continuity of the realizations of both $\tilde{x}$ and $\tilde{q}$. Then, the equality in (18) easily follows.

Finally, for an arbitrary discipline G, similarly to the proof of (67), we prove that, uniformly on $t \in [0, T]$,

$$\liminf_{r \to \infty} r(\mathscr{C}^r_G(t + 1/r) - \mathscr{C}^r_G(t)) - \sum_i \bar{C}_i(\tilde{q}_i(t)) = 0. \tag{68}$$

(The proof of (68) uses, in addition, also the continuity of $\tilde{x}$, and properties (16) and (17).) Then, (68) implies the inequality in (18).

The One-Dimensional Skorohod Problem

The following proposition describes standard properties of solutions to the one-dimensional Skorohod problem. (See for example Chen and Mandelbaum 1991b for the proof. The proof is also contained in the proof of Theorem 5.1 of Williams 1998a.)

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

(i) *There exists a unique pair (x, y) of functions in $D([0, \infty), \bar{R})$, such that*

(a) $x(t) = w(t) + y(t) \geqslant 0, t \geqslant 0$,

(b) *y is nondecreasing and nonnegative,*

(c) $y(0) = 0$,

(d) *for any $t \geqslant 0$, if $x(t) > 0$, then t is not a point of increase of y, i.e., there exists $\delta > 0$ such that $y(\xi)$ is constant in $[t - \delta, t + \delta] \cap R_+$.*

This unique pair is (x°, y°), where

$$y^\circ(t) \doteq -\Big[0 \wedge \inf_{0 \leqslant u \leqslant t} w(u)\Big], \quad x^\circ(t) = W(t) + y^\circ(t), \quad t \geqslant 0.$$

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

$$y(t) \geqslant y^\circ(t), \quad x(t) \geqslant x^\circ(t), \quad t \geqslant 0.$$

Acknowledgments

The authors are grateful to the associate editor and two referees for their constructive reviews. The research of the first author was supported by ISF (Israeli Science Foundation)

grants 388/99 and 126/02, by the Niderzaksen Fund, and by the Technion funds for the promotion of research and sponsored research. His research was partially carried out at the Mathematics Center of Lucent's Bell Labs—the hospitality of the center and its members is greatly appreciated.

References

Andrews, M., K. Kumaran, K. Ramanan, A. L. Stolyar, R. Vijayakumar, P. Whiting. 2004. Scheduling in a queueing system with asynchronously varying service rates. *Probab. Engrg. Inform. Sci.* **18** 191–217.

Armony, M., N. Bambos. 1999. Queueing networks with interacting service resources. *Proc. 37th Annual Allerton Conference on Communication, Control and Computing*, Monticello, IL, 42–51.

Bell, S. L., R. J. Williams. 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. *Ann. Probab.* **11** 608–649.

Bramson, M. 1998. State space collapse with applications to heavy traffic limits for multiclass queueing networks. *Queueing Systems* **30** 89–148.

Chen, H., A. Mandelbaum. 1991a. Stochastic discrete flow networks: Diffusion approximations and bottlenecks. *Ann. Probab.* **19** 1463–1519.

Chen, H., A. Mandelbaum. 1991b. Leontief systems, RBV's and RBM's. M. H. A. Davis, R. J. Elliott, eds. *Applied Stochastic Analysis.* Gordon and Breach Science Publishers, New York, 1–43.

Cox, D. R., W. L. Smith. 1961. *Queues.* Methuen, London, U.K., and Wiley, New York.

Dai, J. G., B. Prabhakar. 2000. The throughput of data switches with and without speedup. *Proc. INFOCOM'2000*, 556–564.

Ethier, S. N., T. G. Kurtz. 1986. *Markov Process: Characterization and Convergence.* John Wiley and Sons, New York.

Glazebrook, K. D., J. Nino-Mora. 2001. Parallel scheduling of multiclass $M/M/m$ queues: Approximate and heavy-traffic optimization of achievable performance. *Oper. Res.* **49** 609–623.

Harrison, J. M. 1998. Heavy traffic analysis of a system with parallel servers: Asymptotic optimality of discrete review policies. *Ann. Appl. Probab.* **8** 822–848.

Harrison, J. M., M. J. Lopez. 1999. Heavy traffic resource pooling in parallel-server systems. *Queueing Systems* **33** 339–368.

Harrison, J. M., J. A. Van Mieghem. 1997. Dynamic control of Brownian networks: State space collapse and equivalent workload formulations. *Ann. Appl. Probab.* **7** 747–771.

Klimov, G. P. 1974. Time sharing systems I. *Theory Probab. Appl.* **19** 532–551.

Klimov, G. P. 1978. Time sharing systems II. *Theory Probab. Appl.* **23** 314–321.

McKeown, N., V. Anantharam, J. Walrand. 1996. Achieving 100% throughput in an input-queued switch. *Proc. INFOCOM'96*, 296–302.

Reiman, M. I. 1984. Some diffusion approximations with state space collapse. *Proc. Internat. Seminar Modeling Performance Evaluation Methodology*. Lecture Notes in Control and Information Sciences. Springer, New York, 209–240.

Reiman, M. I. 1988. A multiclass feedback queue in heavy traffic. *Adv. Appl. Probab.* **20** 179–207.

Stolyar, A. L. 2004. MaxWeight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *Ann. Appl. Probab.* **14** 1–53.

Tassiulas, L., A. Ephremides. 1992. Stability properties of constrainted queueing systems and scheduling policies for maximum throughput in multishop radio network. *IEEE Trans. Automatic Control* **37** 1936–1948.

Van Mieghem, J. A. 1995. Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *Ann. Appl. Probab.* **5** 809–833.

Whitt, W. 1999. Improving service by informing customers about anticipated delays. *Management Sci.* **45** 192–207.

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

Williams, R. J. 1998b. Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse. *Queueing Systems* **30** 27–88.

Williams, R. J. 2000. On dynamic scheduling of a parallel server system with complete resource pooling. *Fields Institute Communications* **28** 49–71.

Zohar, E., A. Mandelbaum, N. Shimkin. 2002. Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Sci.* **48** 566–583.