# Gc$\mu$ Scheduling of Flexible Servers: Asymptotic Optimality in Heavy Traffic

A. Mandelbaum
Technion Institute
Haifa, 32000
ISRAEL
avim@tx.technion.ac.il

A. L. Stolyar
Bell Labs, Lucent Technologies
Murray Hill, NJ 07974
U.S.A.
stolyar@research.bell-labs.com

**Abstract**

We consider a queueing system with multi-type customers, non-homogeneous multi-skilled 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 ($Gc\mu$), known to be asymptotically optimal in a single server system [14], is in fact asymptotically optimal in our much more general setting.

## 1 Introduction

We analyze the scheduling problem for flexible servers with overlapping capabilities. Our setup is a queueing system with multi-type customers, multi-skilled 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 [14] striking result for homogeneous servers. It constitutes a natural progression of [12], which is here adapted to the (parallel server) model of Harrison and Lopez [8]. (Williams [16] is recommended as an introduction to the subject.) In this paper we present only result formulations. The detailed analysis can be found in [10].

To describe our $Gc\mu$-rule, let $\mu_{ij}$ denote the service rate of type $i$ customers by server $j$. ($\mu_{ij}$ is the reciprocal of an average service time; $\mu_{ij} = 0$ indicates that server $j$ can not 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 Section 3.) Then, applying the $Gc\mu$-rule when becoming idle at time $t$, server $j$ takes for service the longest-waiting type $i$ customer such that

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

An alternative cost structure is when each type $i$ customer incurs, upon service completion, a waiting cost $C_i(W_i)$ which is a function of its sojourn time $W_i$. Then, the type $i$ to be served is one for which

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

where $W_i(t)$ is the head-of-the-line waiting time 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 Section 6. 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 $W$-version holds with respect to sojourn times (Subsection 6.1).

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 waiting 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 [10].)

The optimality of $Gc\mu$ is relative to all scheduling disciplines, preemptive or non-preemptive, as long as each server $j$ serves customers of each type $i$ in order of their arrivals. Our asymptotics is for a sequence of systems that approach heavy traffic, in a way that is precisely defined in Section 5 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 (Section 4). The north-east boundary of the stability set (its maximal elements) constitute those arrival rates $\lambda$ for which our system is critically loaded. The system is in heavy traffic when its vector of arrival rates is 'close' to a maximal $\lambda \in M$ (Section 5). We further say that our system exhibits *complete resource pooling* (CRP) if the outer normal $\nu^*$ to the set $M$ at that $\lambda$ is unique up to scaling (plus some additional non-degeneracy conditions) and, in addition, all the coordinates of $\nu^*$ are strictly positive (Section 4). The quantity $X(t) = \sum_i \nu_i^* Q_i(t)$ is then called the *equivalent workload* at time $t \geq 0$.

Our main result is that, under CRP and in the heavy traffic limit, the $Gc\mu$-rule minimizes the equivalent 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 $\sum_i \nu_i^* C_i(Q_i(t))$. These two properties imply minimization of the cumulative queueing costs over any finite interval.

The intuition behind this unexpected result and the analysis (in [10]) are analogous to those in [12], and can be roughly described as follows. Under the CRP condition and $Gc\mu$-rule, sample paths of the *fluid* process corresponding to a critically loaded system (input rates equal to $\lambda$) 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 $\nu^*$. (A fixed point $^\circ Q$ is exactly the point that minimizes $\sum C_i(Q_i)$, given the equivalent workload value $\sum \nu_i^* {}^\circ Q_i$.) This 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. In turn, this means that, as long as total queue length is non-zero, the Gc$\mu$-rule "reduces to" the rule that maximizes the value of $\sum_i \nu_i^* \mu_i(t)$, where $\mu_i(t)$ is the "instantaneous" service rate of flow $i$. In other words, the server pool operates as a single "super-server," which serves the equivalent workload at the maximum possible rate. This implies the property of equivalent workload minimization.

The distinguishing feature of our analysis, compared to [12], is that here we deal with continuous time process and more general convex cost structure. (Although in some other respects the model in [12] is more general).

Due to space limitation, we do not present a detailed literature review (see [10] and references therein). We just want to briefly mention that our scheduling problem for *single server* queues has been studied extensively, culminating in Van Mieghem's work

[14]. For multiserver systems, the *stability* of *MaxWeight*-type rules (roughly, Gc$\mu$-rules with quadratic costs) has been demonstrated starting with the work of Tassiulas and Ephremides [13]. (See also [11, 2, 1].) The heavy traffic optimization for multiservers starts from Harrison [7], followed by works of Bell and Williams [3], Harrison and Lopez [8] (where our general model was introduced), and Williams [16]. The *equivalent workload formulation* was introduced by Harrison and Van Mieghem [7]. The general approach to proving state space collapse in heavy traffic was developed by Bramson [4] and Williams [15].

The outline of our paper is as follows. The formal model is introduced in Section 2. The Generalized c$\mu$-Rule (Gc$\mu$) is described in Section 3. In Section 4 we formulate the conditions for Complete Resource Pooling (CRP), followed by the definition of heavy-traffic in Section 5. Theorem 1, in Section 6, establishes the asymptotic optimality of Gc$\mu$, with respect to queueing costs; we then outline its adaptation to waiting costs. Section 7 contains final remarks.

*Notations:* Throughout the paper, we use the notations $R$, $R_+$, and $R_{++}$, for the sets of real, real non-negative, and real positive numbers, respectively. 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 $x = (x_1, \ldots, x_N)$. The dot-product (scalar product) of $x, y \in R^N$, is

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

# 2 The Model

We consider a queueing system with $I$ customer *types* and $J$ flexible *servers*. ($I < \infty, J < \infty$.) Types are indexed by $i = 1, \ldots, I$, which we abbreviate to $i \in I$. Similarly, servers are indexed by $j = 1, \ldots, J$, or $j \in J$.

The arrival process for each type $i$ 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 $\alpha_i^2 = Var[u_i(1)]$.

The service times of type $i$ customers by server $j$ 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 the type $i$ customer at server $j$ (if there is any). Let $\mu_{ij} = 1/E[v_{ij}(1)] < \infty$ and $\beta_i^2 = Var[v_{ij}(1)]$. The convention $\mu_{ij} = 0$ is used when server $j$ can not serve type $i$.

All arrival and service processes are assumed mutually independent.

We allow a wide class of scheduling disciplines (which in particular may be preemptive or non-preemptive). The only constraints are that each server $j$ takes customers from each queue $i$ in FIFO order, and a service of one type $i$ customer can *not* be preempted by service of another type $i$ customer. (This condition excludes disciplines which are allowed to use information on the individual customers' service requirements and thus "pick out" those with the shortest service times.)

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 $W_i(t)$ be the "age" of the longest waiting customer of type $i$, among those whose service has not yet started.

# 3 The $Gc\mu$-Rule

Suppose that, for each type $i$, a cost function $C_i(\zeta), \zeta \geq 0$, is given. Assume that each $C_i(\cdot)$ is a convex strictly increasing function with $C_i(0) = 0$; moreover, $C_i(\cdot)$ is twice continuously differentiable, and its derivative $C_i'(\cdot)$ is a strictly increasing with $C_i'(0) = 0$.

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

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

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.** The above version of the $Gc\mu$-rule accommodates queueing costs. An alternative, for waiting costs, will be introduced in Subsection 6.1.

# 4 Complete Resource Pooling

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

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

With a given $\phi$ 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 $\phi_{ij}$ of its time to queue $i$, in the long run.

We define $M$ to be the set of $\mu(\phi)$ corresponding to all possible $\phi$ as above. Further let $M^*$ denote the set of all *maximal* elements $\mu \in M$ such that $\mu \in R^I_{++}$.

Note that $M$ is a polyhedron in $R^I_+$. We assume that $M$ is non-degenerate (i.e., has dimension $I$), which is equivalent to assuming that each queue $i$ can be served at non-zero rate $\mu_{ij}$ by at least one server $j$. The set $M$ is in fact the closure of our system's *stability region*, that is the closure of the set of arrival rate vectors $\lambda = (\lambda_1, \ldots, \lambda_I)$ such that $\lambda < \mu(\phi)$ for some $\phi$. (Cf. [13, 11, 2, 1, 12].)

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

It is easy to verify that our CRP condition is equivalent to that introduced for parallel server systems in [8, 16] (see Assumption 3.4, Theorem 5.3 and Corollary 5.4 in [16] for a summary).

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 $\lambda$. Note, that $\nu^* \in R^I_{++}$. For concreteness we use $\nu^*$, which is the vector defined uniquely by the additional requirement that $\|\nu^*\| = 1$. The components of $\nu^*$ are sometimes called the *workload contributions* of customers of the different flows (see [8, 16]).

# 5 Heavy Traffic

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

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

must hold. ($\phi$ is "column-stochastic".) There is also a corresponding (unique) normal vector $\nu^*$, in terms of which we define

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

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

We now consider a sequence of queueing systems, indexed by $r \in \mathcal{R} = \{r_1, r_2, \dots\}$, 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 $r \in \mathcal{R}$ is as before, with $I$ customer types and $J$ servers. The other primitives may depend on $r$, a fact that will be acknowledged by appending $r$ as a superscript.

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

$$r(\lambda_i^r - \lambda_i) \to b_i \ , \quad r \to \infty \ , \tag{1}$$

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

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

In addition, we make the following technical assumption, needed to apply Bramson's weak law estimates [4]:
uniformly over $i$ and $r$,

$$E[(u_i^r(1))^2 1\{u_i^r(1) > x\}] \leq \eta(x) \ , \ x \geq 0 \ , \tag{3}$$

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 (1) and (2) imply a functional central limit theorem (FCLT) for the arrival processes:

$$\{r^{-1}(F_i^r(r^2 t) - \lambda_i^r r^2 t), \ t \geq 0\} \stackrel{w}{\to} \{\sigma_i B(t), \ t \geq 0\} \ , \tag{4}$$

where $F_i^r(t)$ is the number of type $i$ customers arrived by time $t$, excluding customers present at time 0; $\sigma_i^2 = \lambda_i \alpha_i^2$, $B(\cdot)$ is a standard (zero drift, unit variance) Brownian motion, and $\stackrel{w}{\to}$ denotes convergence in distribution (for processes in the standard Skorohod space of RCLL functions).

The service time distributions do *not* change with the parameter $r$. 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 $S_{ij}(t),\ t \geq 0$, the number of type $i$ customers which would be served by server $j$ if it processes type $i$ customers continuously up to time $t$. Then, a FCLT applies for each process $S_{ij}(\cdot)$:

$$\{r^{-1}(S_{ij}(r^2t) - \mu_{ij}r^2t),\ t \geq 0\} \stackrel{w}{\to} \{\sigma_{ij}B(t),\ t \geq 0\}\ , \tag{5}$$

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

# 6 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 equivalent 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 Section 3. (An alternative cost structure, where cost is a function of customers' sojourn time, will be discussed in the next subsection.)

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

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

for some constant $c \geq 0$. It is easy to see that
*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 equivalent workload, i.e. satisfying the condition* $\nu^* \cdot q = \nu^* \cdot {}^\circ q$.
The set of fixed points forms a one-dimensional manifold, which can be parameterized for example by values of the equivalent workload ($\nu^* \cdot {}^\circ q$ above).

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

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

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

$$\tilde{q}^r(0) \to \tilde{q}(0)\ , \tag{6}$$

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

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

$$\tilde{x}(t) = \tilde{x}(0) + at + \sigma B(t) + \tilde{y}(t)\ , \tag{7}$$

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

$$\tilde{y}(t) \doteq -[0 \wedge \inf_{0 \le u \le t}\{\tilde{x}(0) + au + \sigma B(u)\}] \ , \tag{8}$$

and the drift $a$ and diffusion coefficient $\sigma$ are given by

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

**Theorem 1** *Consider the sequence of queueing systems in heavy traffic, as introduced in Section 5.*

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 $r \to \infty$,*

$$\tilde{x}^r \stackrel{w}{\to} \tilde{x} \ ,$$

*and*

$$\tilde{q}^r \stackrel{w}{\to} \tilde{q} \ ,$$

*where, for each $t \ge 0$, 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 equivalent 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 equivalent 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 $t \ge 0$,*

$$\liminf_{r \to \infty} \tilde{x}_G^r(t) \ge \tilde{x}(t) \tag{9}$$

*and*

$$\liminf_{r \to \infty} \sum_i C_i(\tilde{q}_{i,G}^r(t)) \ge \sum_i C_i(\tilde{q}_i(t)) \ . \tag{10}$$

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

$$\liminf_{r \to \infty} \int_0^T \sum_i C_i(\tilde{q}_{i,G}^r(t))dt \ge \lim_{r \to \infty} \int_0^T \sum_i C_i(\tilde{q}_i^r(t))dt = \int_0^T \sum_i C_i(\tilde{q}_i(t))dt \ . \tag{11}$$

**Remark.** Suppose that assumption (6), 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 it's equivalent workload $\nu^* \cdot^\circ \tilde{q}(0) \in [\nu^* \cdot \tilde{q}(0), K\nu^* \cdot \tilde{q}(0)]$, where $K \ge 1$ is a fixed constant. 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 [5] for closed queueing networks and in a context close to ours in Bramson [4], 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 equivalent

workload to "jump up" at time 0, i.e. $\nu^* \cdot^\circ \tilde{q}(0) > \nu^* \cdot \tilde{q}(0)$ 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 equivalent 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 equivalent workload and cumulative costs.

**Remark.** 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

$$i \in \arg\max_{i \in I} \gamma_i Q_i^r \mu_{ij} \tag{12}$$

(which can be considered as a special case of the MaxWeight rule in [12]). 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 equivalent workload and it thrives to keep the vector $[\gamma_1 Q_1, \ldots, \gamma_N Q_N]$ proportional to $\nu^*$ at all times: a result analogous to [12].

Theorem 1 deals with *transient* behavior in heavy traffic. Consider our sequence of queueing systems in heavy traffic. Suppose that $a < 0$ 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. The following steady-state conjecture is very natural.

**Theorem 2** *Consider the sequence of processes, under the* $Q\mu$*-rule, as described just above. 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* $r \to \infty$*,*

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

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

$$\nu^\circ \doteq [\sum_i (\nu_i^*)^2/\gamma_i]^{-1}(\nu_1^*/\gamma_1, \ldots, \nu_I^*/\gamma_I) \ .$$

**Remark.** Theorem 2 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 Section 2).

## 6.1 Sojourn Time Costs

Suppose that, as in [14], each customer incurs a "one-time" cost that depends on its type and sojourn time in the system. More precisely, let $\hat{F}_i(T)$ denote the number of type $i$ departures from the system by time $T \geq 0$, and suppose that the objective is to minimize the cumulative waiting cost

$$\frac{1}{r^2} \sum_i \sum_{k=1}^{\hat{F}_i(r^2 T)} C_i(W_i^r(k)/r) \ ,$$

where $r$ is the scaling parameter (as before), $C_i(\cdot)$ is a cost function with the properties described in Section 3, and $W_i^r(k)$ is the sojourn time in the system of the $k$-th type-$i$ customer leaving the system.

Then it can be shown (analogously to Theorem 1), that the following form of the $Gc\mu$-rule minimizes the cumulative waiting cost:
*Each server $j$ chooses for service the longest-waiting customer from queue*

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

*where $W_i^r(t)$ is the waiting time of the (longest-waiting) type $i$ customer.*

# 7 Conclusions

We presented results showing that, surprisingly, a system as complex as our flexible server system, can be optimally controlled by a scheduling rule as parsimonious as Gc$\mu$. Further comments on the robustness of Gc$\mu$, on its accommodation of linear delay costs, and on its extensions, notably to alternative performance measures, homogeneous server groups and other relaxations of the CRP condition, large number of servers, and systems with feedback, can be found in [10].

**Acknowledgements.** The research of AM was carried out at Lucent's Bell Labs and at the Technion. The hospitality of the Mathematics Center at Bell Labs is greatly appreciated. At Technion, the research was supported by the ISF (Israeli Science Foundation) grant 388/9902 and by the Technion funds for the promotion of research and sponsored research.

# References

[1] M. Andrews, K. Kumaran, K. Ramanan, A. L. Stolyar, R. Vijayakumar, P. Whiting. Scheduling in a Queueing System with Asynchronously Varying Service Rates. 2000. (Submitted.)

[2] M.Armony, N.Bambos. Queueing Networks with Interacting Service Resources. 2000. (Submitted.)

[3] S.L.Bell, R.J.Williams. Dynamic Scheduling of a System with Two Parallel Servers in Heavy Traffic with Complete Resource Pooling: Asymptotic Optimality of a Continuous Review Threshold Policy. *Annals of Probability*, Vol. 11, (2001), pp. 608-649.

[4] M.Bramson. State Space Collapse with Applications to Heavy Traffic Limits for Multiclass Queueing Networks. *Queueing Systems*, Vol. 30, (1998), pp. 89-148.

[5] H. Chen, A. Mandelbaum. Stochastic Discrete Flow Networks: Diffusion Approximations and Bottlenecks. *Annals of Probability*, Vol. 19, (1991), pp. 1463-1519.

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

[7] J. M. Harrison. Heavy Traffic Analysis of a System with Parallel Servers: Asymptotic Optimality of Discrete Review Policies. *Annals of Applied Probability*, Vol. 8, (1998), pp. 822-848.

[8] J.M.Harrison, M.J.Lopez. Heavy Traffic Resource Pooling in Parallel-Server Systems. *Queueing Systems*, To appear.

[9] J. M. Harrison, J.A.Van Mieghem. Dynamic Control of Brownian Networks: State Space Collapse and Equivalent Workload Formulations. *Annals of Applied Probability*, Vol. 7, (1997), pp. 747-771.

[10] A.Mandelbaum, A.L.Stolyar. Scheduling Flexible Servers with Convex Delay Costs: Heavy Traffic Optimality of the Generalized c$\mu$-Rule. 2002. (Submitted.)

[11] N.McKeown, V.Anantharam, and J.Walrand. Achieving 100% Throughput in an Input-Queued Switch. *Proceedings of the INFOCOM'96*, (1996), pp. 296-302.

[12] A. L. Stolyar. MaxWeight Scheduling in a Generalized Switch: State Space Collapse and Equivalent Workload Minimization Under Complete Resource Pooling. 2001. (Submitted.)

[13] L. Tassiulas, A. Ephremides. Stability Properties of Constrainted Queueing Systems and Scheduling Policies for Maximum Throughput in Multishop Radio Network. *IEEE Transactions on Automatic Control*, Vol. 37, (1992), pp. 1936-1948.

[14] J.A.Van Mieghem. Dynamic Scheduling with Convex Delay Costs: the Generalize c$\mu$ Rule. *Annals of Applied Probability*, Vol. 5, (1995), pp. 809-833.

[15] R.J.Williams. Diffusion Approximations for Open Multiclass Queueing Networks: Sufficient Conditions Involving State Space Collapse. *Queueing Systems*, Vol. 30, (1998), pp. 27-88.

[16] R.J.Williams. On Dynamic Scheduling of a Parallel Server System with Complete Resource Pooling. *Fields Institute Communications*, (1998).