M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 11, No. 2, Spring 2009, pp. 237–253
ISSN 1523-4614 | EISSN 1526-5498 | 09 | 1102 | 0237

informs®
DOI 10.1287/msom.1070.0211
© 2009 INFORMS

Scheduling Flexible Servers with Convex Delay Costs in Many-Server Service Systems

Itay Gurvich
Columbia Business School, New York, New York 10027, ig2126@columbia.edu

Ward Whitt
IEOR Department, Columbia University, New York, New York 10027, ww2040@columbia.edu

In a recent paper we introduced the queue-and-idleness ratio (QIR) family of routing rules for many-server service systems with multiple customer classes and server pools. A newly available server serves the customer from the head of the queue of the class (from among those the server is eligible to serve) whose queue length most exceeds a specified proportion of the total queue length. Under fairly general conditions, QIR produces an important state-space collapse as the total arrival rate and the numbers of servers increase in a coordinated way. That state-space collapse was previously used to delicately balance service levels for the different customer classes. In this sequel, we show that a special version of QIR stochastically minimizes convex holding costs in a finite-horizon setting when the service rates are restricted to be pool dependent. Under additional regularity conditions, the special version of QIR reduces to a simple policy: linear costs produce a priority-type rule, in which the least-cost customers are given low priority. Strictly convex costs (plus other regularity conditions) produce a many-server analogue of the generalized-$c\mu$ ($Gc\mu$) rule, under which a newly available server selects a customer from the class experiencing the greatest marginal cost at that time.

Key words: queues; many-server queues; heavy-traffic limits for queues; service systems; cost minimization in many-server queues; skill-based routing; generalized-$c\mu$ rule; queue-and-idleness-ratio control
History: Received: June 29, 2007; accepted: November 22, 2007. Published online in *Articles in Advance* April 25, 2008.

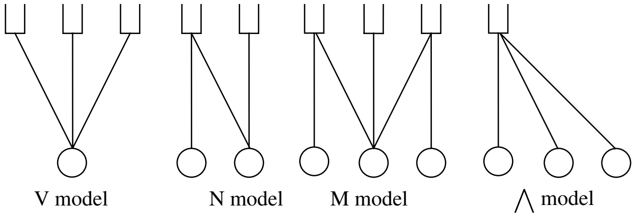
1. Introduction

The optimal control of queueing systems to minimize holding costs or maximize revenues has been the subject of extensive literature (e.g., Stidham 1993). Many have exploited Markov decision processes, but the scope of the Markov decision processes approach is necessarily limited to models that have an underlying Markovian structure. The price of trying to find the exact optimal solution is that simple solutions can be found only for relatively simple models. The curse of dimensionality tends to make large problems intractable.

An alternative, more tractable approach for complex models is to exploit *heavy-traffic asymptotics*. Approximations for description and control are generated by considering limits of a sequence of appropriately scaled queueing processes. The goal then is to generate good policies from asymptotically optimal policies. In contrast to the *conventional heavy-traffic regime*, which has increasing demand volumes with a fixed number of servers, we will be considering the *many-server heavy-traffic regime*, where the number of servers increases together with the demand volume. Thus, we will be generating approximate controls for many-server service systems, such as complex call centers.

A key paper in the conventional heavy-traffic literature is by Mandelbaum and Stolyar (2004), which extended the seminal multiclass single-server work of Van Mieghem (1995), building on Harrison (1988), to a setting with multiple nonidentical servers working in parallel. In these papers, as will also be the case in the current paper, a queue is formed for each customer class and customers leave the system after a single service completion. This stands in contrast to queueing networks, in which customers receive service in several stations before leaving the system. Examples of some simple parallel server systems are depicted in Figure 1. Each box at the top of the figure represents

Figure 1 V, N, M, and $\wedge$ Models



a queue for arriving customers of a given class; each circle at the bottom of the figure represents a pool of servers with common skills.

Mandelbaum and Stolyar (2004) showed that the generalized $c\mu$ ($Gc\mu$) rule asymptotically minimizes convex holding costs. Let $\mu_{i,j}$ be the service rate of class i customers by server j, let $Q_i(t)$ be the class i queue length at time t, and assume that the class i queue incurs a cost at rate $C_i(Q_i(t))$, where C_i is a twice-continuously differentiable, strictly increasing, strictly convex function with $C_i(0) = C_i'(0) = 0$. The $Gc\mu$ rule dictates that, when becoming free at time t, server j next serves a customer from the class i that maximizes $\mu_{i,j}C_i'(Q_i(t))$, where C_i' is the first derivative of C_i; i.e., the class to be served next by server j is

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

The classic $c\mu$ rule is obtained when C_i is linear, but strict convexity is required for this result.

The $Gc\mu$ rule has appealing simplicity, allowing the decisions to be made myopically in a decentralized manner. To make the decision for pool j, it suffices to know the queue lengths of the classes those agents can serve. The asymptotic optimality result in Mandelbaum and Stolyar (2004) builds on a useful invariance principle that implies, under a complete-resource-pooling condition, that a certain reflected Brownian motion serves as a lower bound for the aggregate workload process. As a consequence, the minimal cost can be achieved asymptotically by making sure that (a) the aggregate workload asymptotically achieves the reflected Brownian motion lower bound, and (b) given a workload level, its distribution among the different customer classes is performed optimally with respect to holding costs $C_i(\cdot)$. The $Gc\mu$ rule achieves both objectives simultaneously via a state-space-collapse result, building on the general framework introduced by Bramson (1998).

The restriction of Mandelbaum and Stolyar (2004) to strictly convex cost functions is not made for technical convenience. The simple $c\mu$ rule obtained from the $Gc\mu$ rule when the holding costs are linear indeed *fails* to asymptotically minimize the holding costs for parallel-server systems in the conventional heavy-traffic limit. In fact, applying the $c\mu$ rule to these systems may be disastrous, leading to system explosion, as illustrated by Harrison (1998) and further discussed in Dai and Tezcan (2008a).

The purpose of this paper is to extend, as much as possible, the result of Mandelbaum and Stolyar (2004) to the many-server heavy-traffic limiting regime introduced in Halfin and Whitt (1981). Unlike Mandelbaum and Stolyar (2004), our results *do* cover linear holding costs, thus underscoring the significant differences between the two heavy-traffic regimes. Section 1.1 of Dai and Tezcan (2008a) highlights these differences. Unfortunately, however, the useful invariance phenomenon that exists in the conventional heavy-traffic regime does not carry over to the many-server regime. To establish a related invariance principle in the many-server setting, we restrict attention to multiserver systems with pool-dependent service rates, i.e., to settings in which $\mu_{i,j} = \mu_j$ for every class i that can be served by servers of type j. (That effectively eliminates the μ component of the $Gc\mu$ rule.)

However, that is not all: the restriction to pool-dependent service rates is not sufficient by itself to guarantee that the aggregate workload in the system is asymptotically minimized. In contrast to the conventional heavy-traffic regime, care is needed in assigning customers to servers. Not any work-conserving policy will achieve the desired performance. Consequently, our proposed solution contains two components: *a routing component*, specifying what to do on customer arrival, and *a scheduling component*, specifying what to do on service completion.

Our solution builds on the queue-and-idleness ratio (QIR) family of controls introduced in Gurvich and Whitt (2007b). Moreover, our proofs here draw heavily on Gurvich and Whitt (2007b). Here we show that a special version of QIR (with appropriately chosen state-dependent ratio functions) asymptotically minimizes convex holding costs, including linear costs. Moreover, when restricting the attention to strictly convex holding-cost functions (with additional regularity conditions), the scheduling component of QIR reduces to the $Gc\mu$ rule (where the μ component is trivial, as indicated above).

We hasten to admit that we are by no means the first to analyze holding-cost minimization in the Halfin-Whitt regime. A multiclass but single-pool (single-server-type) model (the V model) with linear holding and abandonment costs was considered by Harrison and Zeevi (2004). A simplified setting of the V model with a common service rate for all classes and more complex, but still linear, cost structure was analyzed in Gurvich et al. (2008), where a threshold policy was proposed. Armony (2005) stochastically minimized the queue length in a multiserver-pool setting with a single customer class, which she names the inverted V (or $\bigwedge$) model. Our analysis exploits her results.

Much greater generality was achieved by Atar (2005). In his far-reaching paper, Atar covers asymptotic minimization of holding costs in general multiclass, multipool systems in the Halfin-Whitt regime. His analysis focuses on the Hamilton-Jacoby-Bellman (HJB) equations governing the limit Brownian control problem and on obtaining asymptotically optimal controls. His results are very general, but because of this generality, they provide little insight about specific cases, and the proposed controls are not as elegant as the $Gc\mu$ rule.

Our results here and in Gurvich and Whitt (2007a, b) are closely related to concurrent and independent results by Dai and Tezcan (2008a, b). Their first paper (2008b) took the important path of constructing explicit solutions for specific cases, assuming linear holding costs. That paper considers the N model, having two customer classes and two agent pools—one of which is dedicated and the other flexible. Their second paper (Dai and Tezcan 2008b) is a generalization of the first, considering general skill-based routing (SBR) systems with pool-dependent rates, just as in this paper, but still focusing on linear holding costs. Our current paper extends their results to more general convex holding costs. Thus, our current paper includes their results as a special case. Moreover, we show that the policy that Dai and Tezcan (2008b) show to be optimal for linear cost functions is also optimal much more generally; linearity is sufficient, but a weaker condition on the derivatives of the cost functions is sufficient as well.

As should be expected, the analysis here is similar to the analysis in Dai and Tezcan (2008b), but there are significant differences. Dai and Tezcan (2008b) build on their previous important work (Dai and Tezcan 2009), extending Bramson's (1998) state-space collapse framework to the Halfin-Whitt many-server regime. In contrast, we apply our own previous paper Gurvich and Whitt (2007b) and that of Armony (2005).

The remainder of the paper is organized as follows. We introduce the model and some notation in §2. Then we state the main results in §3. Included there are analogous results for a delay-cost formulation involving a special version of the waiting-and-idleness ratio (WIR)-routing rule. We provide the proofs, building on Gurvich and Whitt (2007b), in §4. Finally, we make concluding remarks in §5.

2. Model

We consider a system with a fixed set of customer classes $\mathcal{I} := \{1, \ldots, I\}$ and a fixed set of server pools, $\mathcal{J} := \{1, \ldots, J\}$. There is a queue for each customer class. If customers cannot enter service immediately on arrival, they go to the end of their queue. The number of servers in pool j is given by N_j. To define the arrival processes for the different customer classes, we let $\{A_i, i \in \mathcal{I}\}$ be a family of independent renewal counting processes $A_i := \{A_i(t), t \geq 0\}$ with interarrival-time distribution having mean 1 and squared coefficient of variation (variance divided by the square of the mean) $c^2_{a,i}$. Given a vector $\vec{\lambda} = (\lambda_1, \ldots, \lambda_I)$ of arrival rates, we let the arrival process for class i be the time-scaled renewal process $\{A_i(\lambda_i t), t \geq 0\}$. We let $\lambda := \sum_{i \in \mathcal{I}} \lambda_i$ be the total arrival rate.

The set of possible assignments of customers to servers in this system has a natural representation as a bipartite graph with vertices $V = \mathcal{J} \cup \mathcal{I}$; i.e., V is the union of the set of customer classes and the set of agent pools. We let E be the set of edges in the graph. The set E is allowed to be a strict subset of the set of all possible edges $\mathcal{E} := \{(i, j) \in \mathcal{I} \times \mathcal{J}\}$. An edge $(i, j) \in E$ corresponds to allowing pool j servers to serve class i customers. To achieve needed resource pooling (see also Assumption 2.4), we make the following assumption about E.

Assumption 2.1 (Connected Routing Graph). *The graph $G=(V,E)$ is a connected graph.*

Given the routing graph $G:=(V,E)$, which we characterize via E, let $I(j)$ be the set of classes that a pool j server can serve; i.e., $I(j):=\{i\in\mathcal{I}\colon (i,j)\in E\}$; $I(j)$ is referred to as the *skill set* of pool j servers. Similarly, let $J(i)$ be the set of all server pools that can serve class i; i.e., $J(i):=\{j\in\mathcal{J}\colon (i,j)\in E\}$. Motivated by the application to call centers, we call these skill-based-routing (SBR) systems. In that setting, servers are usually called agents; hereafter, we use these terms interchangeably.

In general, the service time of a customer can depend on both the customer's class and the pool of the agent providing the service, but otherwise (conditional on that information), we assume that the service times are mutually independent exponential random variables, independent of the arrival processes. With that assumption, the dependence is formally introduced by assuming general service rates $\mu_{i,j}$. In this paper, however, we restrict the attention to systems in which the service rates are pool dependent.

Assumption 2.2 (Pool-Dependent Service Rates). *There exist J constants $\mu_1,\ldots,\mu_J$ so that*

$$\mu_{i,j}=\mu_j \quad \textit{for all } j \textit{ and } i\in I(j).$$

Without loss of generality, we assume that the agent (server) pools are ordered in decreasing order of their processing rates, so that

$$\mu_1\geq\mu_2\cdots\geq\mu_J. \tag{1}$$

Assumption 2.2 is crucial to our asymptotic optimality results. The pool dependence allows us to asymptotically minimize the aggregate queue length in the system independently of the way this aggregate queue length is distributed among the different classes. This is analogous to the asymptotic minimality of the workload in the conventional heavy-traffic regime studied by Mandelbaum and Stolyar (2004). We emphasize that not every work-conserving policy will asymptotically achieve the minimal aggregate queue length; see Remark 3.1.

In the absence of Assumption 2.2, simple controls are not likely to emerge as asymptotically optimal solutions in the Halfin-Whitt regime. Harrison and Zeevi (2004) provide a counterexample to the asymptotic optimality of the $c\mu$ rule with linear holding costs, thus highlighting the difference between the conventional heavy-traffic regime and the many-server heavy-traffic regime, which we define next.

2.1. Many-Server Heavy-Traffic Scaling

We consider a family of systems indexed by the aggregate arrival rate λ and let $\lambda\to\infty$. The service rates μ_j and $j\in\mathcal{J}$, the routing graph G, the basic rate-1 renewal arrival processes A_i, and the ratios $a_i:=\lambda_i/\lambda$ are all held fixed. We set $A_i^\lambda(t):=A_i(\lambda_i t)$ and $A^\lambda(t):=\sum_{i\in\mathcal{I}}A_i^\lambda(t)$. The associated family of staffing vectors is $N^\lambda:=(N_1^\lambda,\ldots,N_J^\lambda)$, with N_j^λ being the number of agents in pool $j\in\mathcal{J}$. The staffing levels are assumed to satisfy the following many-server heavy-traffic condition.

Assumption 2.3 (Many-Server Heavy-Traffic Regime). *Assume that*

$$\lim_{\lambda\to\infty}\frac{N_j^\lambda}{\lambda}=\nu_j,\quad j\in\mathcal{J},$$

for some strictly positive vector $\nu=(\nu_1,\ldots,\nu_J)$ that satisfies $\sum_{j\in\mathcal{J}}\mu_j\nu_j=1$ and

$$\lim_{\lambda\to\infty}\frac{N_j^\lambda-\nu_j\lambda}{\sqrt{\lambda}}=\gamma_j,\quad j\in\mathcal{J}, \tag{2}$$

for $\gamma=(\gamma_1,\ldots,\gamma_J)$ with $-\infty<\gamma_j<\infty$ for all $j\in\mathcal{J}$.

Assumption 2.3 guarantees that the aggregate system capacity as given by $\sum_{j\in\mathcal{J}}\mu_jN_j$ is, in first order, the minimal capacity that is needed to serve an arrival stream with rate λ. This, however, is not enough, and we require also a resource-pooling condition.

Assumption 2.4 (Resource-Pooling Condition). *There exists a vector $x\in\mathbb{R}_+^{I\times J}$ that satisfies*

$$\sum_{j\in J(i)}\mu_jx_{ij}\nu_j=a_i,\quad i\in\mathcal{I},\quad \textit{and}\quad \sum_{i\in I(j)}x_{ij}=1,\quad j\in\mathcal{J}, \tag{3}$$

such that the graph $\mathcal{E}(x):=\{(i,j)\in\mathcal{I}\times\mathcal{J}\colon x_{ij}>0\}$ is a connected graph.

Assumption 2.4 guarantees that each customer class has access to more than the minimal capacity that it requires, that is, that $\sum_{j\in J(i)}\mu_j\nu_j>a_i$ (with strict inequality). Indeed, as the graph $\mathcal{E}(x)$ is connected, at least one $k\neq i$ exists such that $x_{kl}>0$ for some $l\in J(i)$ and, in particular, $x_{il}<1$ so that $\sum_{j\in J(i)}\mu_j\nu_j>\sum_{j\in\mathcal{J}(i)}\mu_jx_{ij}\nu_j=a_i$. This local excess capacity condition guarantees that if all the capacity in the set of pools $J(i)$ is directed to serve the class i queue, the queue can be drained extremely fast, and practically instantaneously as the system size grows. The capability to instantaneously decrease the number of customers in a given queue lies at the heart of state-space collapse results in both heavy-traffic regimes, the conventional and the many-server ones. It should be noted, however, that in contrast to much of the heavy-traffic literature, we do not assume that the graph $\mathscr{E}(x)$ is a tree. Our less-restrictive condition is a consequence of the assumption on pool-dependent service rates and the corresponding state-space collapse results in Gurvich and Whitt (2007b). Finally, we point out that this assumption is consistent with the heavy-traffic assumptions in Gurvich and Whitt (2007b)—see Assumptions 2.1–2.3 therein.

Assumptions 2.1–2.4 will be assumed throughout the rest of the paper. Assumption 2.3 implies that

$$\sum_{j\in\mathcal{J}} \mu_j N_j^\lambda = \lambda + \beta\sqrt{\lambda} + o(\sqrt{\lambda}) \quad \text{as } \lambda \to \infty, \tag{4}$$

where $\beta = \sum_{j\in\mathcal{J}} \mu_j \gamma_j$. Letting the traffic intensity in system λ be $\rho^\lambda := \lambda / \sum_{j\in\mathcal{J}} \mu_j N_j^\lambda$, we see that

$$\lim_{\lambda\to\infty} \sqrt{\lambda}(1-\rho^\lambda) = \beta, \tag{5}$$

which generalizes the Halfin-Whitt many-server heavy-traffic condition for the single-class single-pool $M/M/N$ queue; see Equation (2.2) in Halfin and Whitt (1981).

For the λth system, let $Q_i^\lambda(t)$ be the number of class i customers in the queue at time t and let $I_j^\lambda(t)$ be the number of idle agents in pool j at time t. Let $Q_\Sigma^\lambda(t) := \sum_{i\in\mathcal{I}} Q_i^\lambda(t)$ and $I_\Sigma^\lambda(t) := \sum_{j\in\mathcal{J}} I_j^\lambda(t)$ be the corresponding aggregate quantities. Let $N_\Sigma^\lambda = \sum_{j\in\mathcal{J}} N_j^\lambda$ be the aggregate number of agents. Finally, let the overall number of customers in the system at time t be

$$X_\Sigma^\lambda(t) := Q_\Sigma^\lambda(t) + \sum_{j\in\mathcal{J}} (N_j^\lambda - I_j^\lambda(t)).$$

Let the corresponding *scaled processes* be

$$\widehat{Q}_i^\lambda(t) := \frac{Q_i^\lambda(t)}{\sqrt{\lambda}}, \quad i\in\mathcal{I}, \qquad \hat{I}_j^\lambda(t) := \frac{I_j^\lambda(t)}{\sqrt{\lambda}}, \quad j\in\mathcal{J},$$

and

$$\widehat{Q}_\Sigma^\lambda(t) := \frac{Q_\Sigma^\lambda(t)}{\sqrt{\lambda}}, \quad \hat{I}_\Sigma^\lambda(t) := \frac{I_\Sigma^\lambda(t)}{\sqrt{\lambda}} \quad \text{and}$$
$$\widehat{X}_\Sigma^\lambda(t) := \frac{X_\Sigma^\lambda(t) - N_\Sigma^\lambda}{\sqrt{\lambda}}.$$

We consider two different objective functions: the first measures cost in terms of the queue length, and the second measures cost in terms of customer delay. Specifically, let $W_{i,k}^{\lambda,\pi}$ be the waiting time of the kth class i customer to arrive to the λth system after time 0 and under a control π. Then let

$$J_1^\lambda(\pi, T) := \int_0^T \sum_{i=1}^I C_i(\widehat{Q}_i^{\pi,\lambda}(t))\,dt \quad \text{and}$$

$$J_2^\lambda(\pi, T) := \frac{1}{A_i^\lambda(t)} \sum_{i=1}^I \sum_{k=1}^{A_i^\lambda(T)} C_i(\sqrt{\lambda} W_{i,k}^\lambda(\pi))$$

for appropriate cost functions C_i. Both the queue length and the waiting times in these objective functions are scaled so that they have proper limits as $\lambda \to \infty$ with the many-server heavy-traffic scaling. We refer the reader to §3 of Van Mieghem (1995) and §7 of Mandelbaum and Stolyar (2004) for a discussion and justification of using these scalings in the definition of the objective functions. We make the following assumption about the cost functions $\{C_i,\ i\in\mathcal{I}\}$.

Assumption 2.5 (Admissible Cost Functions). *For each $i\in\mathcal{I}$, the cost function C_i is assumed to be nondecreasing and convex with $C_i(0) = 0$.*

The essential requirement in Assumption 2.5 is that the cost functions are convex and nondecreasing. The assumption that $C_i(0) = 0$ is made without loss of generality, as we can add an arbitrary constant to each cost function without affecting the solution of the nonlinear program (11) below.

Remark 2.1. Our assumptions on the cost functions are substantially weaker than those imposed in Mandelbaum and Stolyar (2004). In contrast to Mandelbaum and Stolyar (2004), the many-server regime does allow us to consider non-strictly convex cost functions, such as linear functions and piecewise-linear functions, but the optimal policy is not $Gc\mu$ in those cases. The assumptions to get $Gc\mu$ will be similar.

Every control π needs to consist of two components: the *routing component*—specifying what to do when a customer arrives to the system—and the *scheduling component*—specifying what to do when an agent completes service and becomes available. We make additional restrictions on the family of controls; toward that end, let Π_k be the set of admissible policies for $J_k^\lambda(\pi, T)$, $k=1,2$. The sets of admissible policies for both criteria will consist of nonanticipating policies; see §4 for a formal definition. The set Π_2 will be restricted to policies that serve customers on a first-come first-served (FCFS) basis within each customer class, so that $\Pi_2 \subset \Pi_1$. Our two optimization problems are then given by

$$\inf_{\pi\in\Pi_1} J_1^\lambda(\pi, T) \quad \text{and} \quad \inf_{\pi\in\Pi_2} J_2^\lambda(\pi, T) \tag{6}$$

and are respectively referred to as the *holding-cost formulation* and the *delay-cost formulation*.

We say that a family of admissible policies $\{\pi^\lambda\}$ is *asymptotically optimal* (as $\lambda \to \infty$) for the objective function $k \in \{1,2\}$ if for any $T > 0$ and given any other sequence of admissible policies $\{\tilde{\pi}^\lambda\}$,

$$\limsup_{\lambda\to\infty} J_k^\lambda(\pi^\lambda, T) \le_{\text{st}} \liminf_{\lambda\to\infty} J_k^\lambda(\tilde{\pi}^\lambda, T), \tag{7}$$

where $\le_{\text{st}}$ denotes (conventional) *stochastic ordering*. Stochastic ordering is a strong comparison that is not available in many problems, but it turns out to be provable in our setting. Note that asymptotic optimality of a sequence π^λ does not imply uniqueness. Indeed, there might be multiple asymptotically optimal controls. Our aim is to identify one such asymptotic solution.

3. Main Results

In this section we establish the asymptotic optimality of special QIR and analogous WIR rules for the holding-cost and delay-cost formulations, respectively.

3.1. Holding-Cost Formulation

We start by formally defining QIR; see Gurvich and Whitt (2007b) for background. Toward that end, we say that an $\mathbb{R}^m$-valued function f on a subset S of $\mathbb{R}^k$ is *locally Hölder continuous* with exponent $\alpha > 0$ if, for every compact subset K of S, there exists a constant C_K such that

$$\|f(x) - f(y)\| \le C_K \|x - y\|^\alpha \quad \text{for all } x, y \in K, \tag{8}$$

where $\|\cdot\|$ is a chosen norm inducing the usual Euclidean topology, which we take to be the $\mathbb{L}^1$ norm $\|x\| := \sum_i |x_i|$. With that definition, we are ready to define the class of admissible state-dependent ratio functions.

Definition 3.1 (An Admissible State-Dependent Ratio Function). For an integer $d > 0$, a vector-valued function r: $\mathbb{R}_+ \mapsto \mathbb{R}_+^d$ is an admissible state-dependent ratio function if $\sum_{k=1}^d r_k(x) = 1$ for all $x \in \mathbb{R}_+$ and if every component r_k: $\mathbb{R}_+ \mapsto \mathbb{R}_+$ is locally Hölder continuous on the open interval $(0, \infty)$ with some exponent $\alpha_k > 0$.

Definition 3.2 (QIR for Admissible State-Dependent Ratio Functions). Given two admissible state-dependent ratio functions v and p, QIR is defined as follows:

On arrival of a class i customer at time t, the customer will be routed to an available agent in pool j^*, where

$$j^* := j^*(t) \in \operatorname*{arg\,max}_{j\in J(i),\, \hat{I}_j^\lambda(t)>0} \{\hat{I}_j^\lambda(t) - [\widehat{X}_\Sigma^\lambda(t)]^- v_j([\widehat{X}_\Sigma^\lambda(t)]^-)\};$$

i.e., the customer will be routed to an agent pool with the greatest idleness imbalance. If there are no idle agents in any of the pools in $J(i)$, the customer waits in queue i, to be served in order of arrival.

On service completion by a pool j agent at time t, the agent will admit to service the customer from the head of queue i^*, where

$$i^* := i^*(t) \in \operatorname*{arg\,max}_{i\in I(j),\, \widehat{Q}_i^\lambda(t)>0} \{\widehat{Q}_i^\lambda(t) - [\widehat{X}_\Sigma^\lambda(t)]^+ p_i([\widehat{X}_\Sigma^\lambda(t)]^+)\};$$

i.e., the agent will admit a customer from the queue with the greatest imbalance. If there are customers waiting in any of the queues in $I(j)$, the agent will remain idle.

Ties are broken in an arbitrary but consistent manner. Formally, let $\tilde{A}_i^\lambda(t)$ be the time that elapsed since the last arrival of a class i customer before time t. Then ties are broken so that the vector-valued stochastic process

$$(\widehat{Q}^\lambda, \widehat{Z}^\lambda, \tilde{A}) := (\widehat{Q}_i^\lambda(t), \widehat{Z}_{i,j}^\lambda(t), \tilde{A}_i^\lambda(t);\, i \in \mathcal{I}, j \in \mathcal{J}), \tag{9}$$

is a Markov process with stationary transition probabilities.

For given ratio functions v and p, we denote the resulting QIR control by QIR(p, v). We will be interested in special ratio functions p and v that are appropriate to achieve for our optimization objective. The routing component is relatively simple, so we start with it. Let v^* be the nonstate-dependent ratio function given by

$$v^*(\cdot) := (0, 0, \ldots, 1). \tag{10}$$

By (1) and (10), a customer of class $i \in I(J)$ will be routed to agent pool J only when all the agents in any other pool $j \in J(i)$ are busy. This component of the control essentially maximizes the throughput of the system, because it makes sure that all the idleness is concentrated in the slowest server pool.

Treating the scheduling component is much more complicated in general but will become correspondingly simple in special cases. To properly treat the scheduling component of our control, we define a deterministic convex optimization problem: Let $q_i^*(x)$, $i \in \mathcal{I}$ be an optimal solution to the nonlinear program (NLP)

$$\begin{aligned} \text{minimize} \quad & \sum_{i\in\mathcal{I}} C_i(q_i(x)) \\ \text{s.t.} \quad & \sum_{i\in\mathcal{I}} q_i(x) = x, \\ & q_i(x) \geq 0, \quad i \in \mathcal{I}, \end{aligned} \tag{11}$$

where C_i is the specified cost functions satisfying Assumption 2.5.

For each x, this NLP is a classical *separable continuous nonlinear resource allocation problem*, as in Ibaraki and Katoh (1988), Patriksson (2008), and Zipkin (1980). A solution always exists and efficient algorithms are available. We are interested in the parametric version, in which we consider the solution as a function of the resource level x. Fortunately, the special structure implies that the solutions at different resource levels are simply related: Having found a solution $q^*(x) := \{q_i^*(x), i \in \mathcal{I}\}$ at resource level x, that determined solution can be kept at resource level $x + \epsilon$; it only remains to optimally allocate the incremental ϵ resource. It suffices to perform marginal analysis: the infinitesimal incremental resource at any time should be allocated to the class(es) with the smallest (right) derivative at the current allocation. As a consequence, we have the following existence result.

Lemma 3.1 (Desired Admissible State-Dependent Ratio Function). *Under Assumption* 2.5, *there exists a parametric optimal solution $q_i^*(x)$ for $i \in \mathcal{I}$ and $x > 0$ to the resource allocation problem* (11) *such that*

$$q_i^*(x) \leq q_i^*(x+\epsilon) \leq q_i^*(x) + \epsilon \quad \textit{for all } i, x > 0 \textit{ and } \epsilon > 0,$$

so that

$$p^*(x) := q^*(x)/x, \quad x > 0 \tag{12}$$

is an admissible state-dependent ratio function, satisfying (8) *with* $\alpha = 1$.

The division by x in (12) causes difficulties in neighborhoods of 0, but the restriction to compact subsets of the open interval $(0, \infty)$ in Definition 3.1 prevents that division by x in (12) from hurting us. Let p^* be the ratio function defined in (12), with $p_i^*(x) := q_i^*(x)/x$ for all $x > 0$ and i, with p^* chosen to be an admissible state-dependent ratio function. The choice of the value at 0, $p_i^*(0)$, is of no real importance because that corresponds to epochs that the queues are empty. Hence, we may choose any value that satisfies $p_i^*(0) \geq 0$ for all $i \in \mathcal{I}$ and $\sum_{i\in\mathcal{I}} p_i^*(0) = 1$.

Finally, let $\pi_1^* := \text{QIR}(p^*, v^*)$ for v^* and p^* defined above. We are now ready to state our main result for the holding-cost criterion. Let $\Rightarrow$ denote convergence in distribution.

Theorem 3.1 (Asymptotic Optimality of QIR for Holding-Cost Criterion). *If*

$$(\widehat{Q}_i^\lambda(0), \hat{I}_j^\lambda(0);\ i \in \mathcal{I},\ j \in \mathcal{J}) \Rightarrow (\widehat{Q}_i(0), \hat{I}_j(0);\ i \in \mathcal{I},\ j \in \mathcal{J}) \quad \textit{as } \lambda \to \infty,$$

then

$$\limsup_{\lambda\to\infty} J_1^\lambda(\pi_1^*, T) \leq_{\text{st}} \liminf_{\lambda\to\infty} J_1^\lambda(\pi^\lambda, T) \tag{13}$$

for any $T > 0$ and any sequence $\{\pi^\lambda\}$ of admissible policies. Consequently, π_1^* is asymptotically optimal for the holding-cost criterion.

Remark 3.1. There is a simple explanation for the validity of Theorem 3.1. By choosing $v := v^*$, the control essentially tries to keep all servers, except the slowest ones, constantly busy. In doing so, the control asymptotically minimizes the aggregate queue length in the system. Because the ratio function p^* defined by (11) and (12) is designed to optimally distribute the aggregate queue length between the different customer classes, the asymptotic optimality follows.

In general, the scheduling component of this version of QIR can be relatively complicated, but it simplifies in many cases. The rest of this section is primarily devoted to such simplifications. Particularly appealing are the simplifications to (i) a priority-type rule and (ii) the $Gc\mu$ rule.

Remark 3.2 (A Priority-Type Rule). If all cost functions are linear, i.e., if $C_i(x) = c_i x$, $x \geq 0$, with $c_{i_0} \leq c_i$ for all $i \neq i_0$, then an optimal solution to (11) is obtained by setting $q^*_{i_0}(x) = x$ and $q^*_i(x) = 0$ for all $i \neq i_0$. Consequently, the selected ratio function p^* in (12) is the nonstate-dependent ratio function given by $p_{i_0}(\cdot) := 1$ and $p_i(\cdot) := 0$ for all $i \neq i_0$.

We now discuss the implications of this structure in QIR. First, it is easily verified that for the N model setting with linear holding costs analyzed in Dai and Tezcan (2008a), QIR reduces to the static priority $c\mu$ rule. Indeed, QIR is equivalent to a static priority rule for all settings in which the set

$$\mathop{\arg\max}_{i \in I(j),\, \widehat{Q}^\lambda_i(t) > 0} \left\{ \widehat{Q}^\lambda_i(t) - [\widehat{X}^\lambda_\Sigma(t)]^+ p_i([\widehat{X}^\lambda_\Sigma(t)]^+) \right\}$$

is identical to the set

$$\mathop{\arg\max}_{i \in I(j),\, \widehat{Q}^\lambda_i(t) > 0} \widehat{Q}^\lambda_i(t),$$

and one can verify that the N model of Dai and Tezcan (2008a) indeed satisfies this restriction. As observed in Dai and Tezcan (2008b), this restriction does not hold for most networks. Instead, a more general priority-type rule is proposed in §2.1 of that paper. This more general rule is, however, identical to the QIR rule we obtain for linear holding costs. Consequently, in the absence of abandonments, Theorem 3.1 covers the results of Dai and Tezcan (2008a, b) as special cases.

It is significant that the same simple priority-type rule, as proposed in Dai and Tezcan (2008b) for linear holding cost, is asymptotically optimal in much greater generally; our analysis shows that linearity is not the critical feature. For the priority-type rule to be asymptotically optimal, it suffices for the marginal cost of one class to be always less than the marginal costs of all other classes. Because C_i is convex, the derivative C'_i exists at all but countably many points and is nondecreasing. Hence, the *sufficient condition for the priority-type rule to be asymptotically optimal* within our convex-cost framework is for there to be a class i_0 such that

$$C'_{i_0}(\infty) := \lim_{x \to \infty} C'_{i_0}(x) \leq C'_i(0+) \quad \text{for all } i \neq i_0. \tag{14}$$

Under condition (14), the NLP has the same priority-type optimal policy. In other words, it suffices to have one low-cost class and then essentially give that class low priority. We then do not need more specific assumptions about the cost functions of the other classes.

While the priority-type rule has a desirable simplicity, it may not be asymptotically optimal when (14) is violated. One can construct simple instances of cost functions in which it is straightforward to identify the asymptotically optimal policy but in which this policy does not yield a simple priority-type rule. One such example is given below.

Remark 3.3. Condition (14) is clearly a common case for applications, making the priority-type rule a common solution. A candidate for the next simplest policy is to have a single switching point x^*, with one class having low priority if the total queue length is less than x^*, and another class has low priority for the excess above x^* if the total queue length is above x^*. That occurs if and only if there exists x^* such that

$$C'_{i_0}(x^*) \leq C'_i(0+) \quad \text{for all } i \neq i_0 \tag{15}$$

and

$$C'_{i_1}(\infty) \leq \min\{C'_{i_0}(x^*+), C'_i(0+)\} \quad \text{for all } i \notin \{i_0, i_1\}. \tag{16}$$

In this case, $q^*_{i_0}(x) = x \wedge x^*$, $q^*_{i_1}(x) = (x - x^*)^+$, and $q_i(x) = 0$ for all other i. A simple example has one linear cost function and one piecewise-linear cost function: $C_1(x) = c_1 x$, $x \geq 0$, $C_2(x) = b_2 x$, $0 \leq x < x^*$, and $C_2(x) = d_2(x - x^*) + b_2 x^*$, $x \geq x^*$, where $b_2 < c_1 < d_2$.

In this single-switching-point setting, there is a threshold, x^*. If the scaled aggregate queue length $\widehat{X}^\lambda_\Sigma(t)$ is less than this threshold, class i_0 has low priority, just as in the priority-type case above, and other classes are served in order of their queue lengths. However, when this threshold is exceeded, then this version of QIR becomes more complicated, because a term is subtracted from the scaled queue length for both classes i_0 and i_1, with the proportion subtracted for class i_1 increasing as $\widehat{X}^\lambda_\Sigma(t)$ increases, so we have a level-dependent weighted-priority scheme. Eventually, as $\widehat{X}^{\lambda}_{\Sigma}(t)$ increases high enough, class i_1 would be the low-priority class, but scheduling for the remaining classes would not simply be by the longest queue, because class i_0 would still have a large term subtracted.

Remark 3.4. Unfortunately, in general, our QIR solution in Theorem 3.1 is not preserved if we allow customer abandonment. Specifically, assume that with each customer there is an associated exponentially distributed patience random variable. A customer whose waiting time exceeds his patience abandons the system. The patience rate for a class i customer is θ_i, and the patience of different customers are independent.

First, it seems reasonable that in the presence of customer abandonment one should consider cost structures that will take into account the cost of abandonment in addition to holding or delay associated costs. However, even if one restricts attention to the same cost structure considered in the current paper, the problem is far from trivial.

When all holding costs are linear, Dai and Tezcan (2008a) showed that, provided that the class with the lowest holding cost coefficient c_i is also the one with the least patience, static priority—which is a specific case of QIR for the N model they consider (see Remark 3.2)—is asymptotically optimal. However, their result does not hold without this special ordering of costs and abandonment rates.

Because our results allow for even more general cost structures, optimality will fail in all but the most trivial case in which all patience rates are equal; i.e., $\theta_i = \theta$ for all $i \in \mathcal{I}$. However, even for this case, it seems that additional conditions need to be imposed for our results for the holding-cost criterion to hold without any change; see, e.g., the discussion in §5 of Atar et al. (2004). The reason for the failure in the more general case is very simple: our approach builds on the ability to asymptotically minimize the aggregate queue length in the system independently of the way it is distributed between the different classes. Thus, after optimizing the aggregate queue length, we may distribute it between the different classes without damaging the aggregate queue. In the presence of class-dependent abandonment rates, however, the aggregate queue length is extremely sensitive to the way it is distributed between the different classes.

3.2. $Gc\mu$ Rule

We have observed in Remark 3.2 that Theorem 3.1 provides a simple priority-type optimal policy in the case of linear holding costs and under the more general condition (14). Otherwise, the optimal QIR control can be somewhat complicated. We now show that our QIR policy reduces to the $Gc\mu$ rule when the costs are strictly convex and satisfy additional regularity conditions. Costs in practice often are regarded as convex instead of linear, so we regard this as an important conclusion, just as in Van Mieghem (1995) and Mandelbaum and Stolyar (2004).

The many-server $Gc\mu$ rule is defined as follows.

Definition 3.3 (Many-Server $Gc\mu$ Rule). The $Gc\mu$ rule for the SBR model with pool-dependent service rates is defined by changing the scheduling component p^* of QIR(p^*, v^*) to—

On service completion by a pool j agent at time t, the agent will admit to service the customer from the head of queue i^* where

$$i^* := i^*(t) \in \underset{i\in I(j),\, \widehat{Q}^{\lambda}_i(t)>0}{\arg\max}\ \mu_j C_i'(\widehat{Q}^{\lambda}_i(t)); \tag{17}$$

if there are no customers waiting in any of the queues in $J(i)$, the agent will remain idle.

Note that the μ_j in (17) is redundant. This redundancy is a result of the pool-dependence assumption. However, we choose to explicitly display μ_j to emphasize the analogy with Mandelbaum and Stolyar's $Gc\mu$ rule. We let $\pi^*_2 := Gc\mu$.

Theorem 3.2 (Asymptotic Optimality of $Gc\mu$ Rule for Holding-Cost Criterion). *If, in addition to the assumptions of Theorem* 3.1, *the cost function C_i is continuously differentiable and strictly convex with $C_i'(0) = C_i(0) = 0$ for all i, then the $Gc\mu$ rule is asymptotically optimal.*

We now outline the proof of Theorem 3.2, assuming that Theorem 3.1 has been established. To do so, we apply the Karush-Kuhn-Tucker (KKT) conditions, which exploit the assumption that the cost functions be continuously differentiable as well as convex, i.e., differentiable with a continuous derivative. (That rules out piecewise-linear cost functions.) The convexity implies that the KKT conditions are necessary and sufficient for an optimal solution. The KKT conditions say that a solution $\{q^*_i(x)\}$ is optimal for the NLP (11) if and only if there exist functions $y(x)$ and $\{\eta_i(x)\}$ satisfying the equations

$$\begin{gathered} C_i'(q_i^*(x)) - \eta_i(x) = y(x), \quad i \in \mathcal{I}, \\ \sum_{i=1}^{I} q_i^*(x) = x, \\ q_i^*(x) \geq 0, \quad \eta_i(x) \geq 0 \quad i \in \mathcal{I}, \\ \eta_i(x) q_i(x) = 0, \quad i \in \mathcal{I}. \end{gathered} \tag{18}$$

The function $y(x)$ is the Lagrange multiplier of the constraint $\sum_{i \in \mathcal{I}} q_i(x) = x$.

Moreover, there is a unique solution to these equations if the cost functions are strictly convex. If we assume in addition that $C_i'(0) = C_i(0) = 0$, then all activities receive positive resources for any $x > 0$, making $\eta_i(x) = 0$ for all i. We combine these observations in the following lemma.

Lemma 3.2 (Simple Inverse Solution). *If C_i is continuously differentiable and strictly convex with $C_i'(0) = C_i(0) = 0$ for all i, then there is a unique solution to the KKT Equations* (18) *and the NLP* (11), *satisfying*

$$q_i^* = C_i'^{-1}(y(x)), \quad x > 0 \quad \textit{for all } i, \tag{19}$$

where $C_i'^{-1}$ is the inverse of C_i', with $\eta_i(x) = 0$ for all i and $y(x) < \min_{i \in \mathcal{I}} C_i(\infty)$.

Lemma 3.2 will be used to show that QIR and $Gc\mu$ are in some sense equivalent. Theorem 3.2 will then follow from the asymptotic optimality of QIR as stated in Theorem 3.1. The details appear in §4.

3.3. Delay-Cost Formulation

We next present the results for the delay-cost criterion. We define two delay-based rules. The first is a special version of WIR, which in turn is a simple modification of QIR that replaces the individual queue length by the waiting time of the customer at the head of the queue. The second rule is a modification of the many-server $Gc\mu$ rule, which we denote as the *D-$Gc\mu$* rule, following Mandelbaum and Stolyar (2004). Toward that end, we let $W_{h,i}(t)$ be the accumulated waiting time of the customer at the head of class i queue at time t. We define the scaled version $\widehat{W}_{h,i}^{\lambda}(t) := \sqrt{\lambda} W_{h,i}^{\lambda}(t)$. Also, we let $A_{\Sigma}^{\lambda}(t) := \sum_{i \in \mathcal{I}} A_i^{\lambda}(t)$.

Definition 3.4 (WIR Rule). For the SBR model with pool-dependent service rates, WIR(p, v) is defined from QIR by replacing the scheduling component p with—

On service completion by a pool j agent at time t, the agent will next serve the customer from the head of queue i^*, where

$$i^* := i^*(t) \in \underset{i \in I(j),\, \widehat{Q}_i^{\lambda}(t) > 0}{\arg\max} \left\{ \frac{A_i^{\lambda}(t)}{A_{\Sigma}^{\lambda}(t)} \widehat{W}_{h,i}^{\lambda}(t) - [\widehat{X}_{\Sigma}^{\lambda}(t)]^+ p_i([\widehat{X}_{\Sigma}^{\lambda}(t)]^+) \right\}.$$

If there are no customers waiting in any of the queues in $J(i)$, the agent will remain idle.

For given ratio functions p and v, we denote the resulting control by WIR(p, v). For $x > 0$, we let p^{**} be an admissible state-dependent ratio function given by $p_i^{**}(x) := q_i^*(x)/x$, where for each $x > 0$, $q_i^*(x), i \in \mathcal{I}$, is a solution to the NLP (11) but with the functions $C_i(\cdot)$ replaced by $C_i^a(\cdot) := C_i(\cdot/a_i)$, where, as before, $a_i := \lambda_i/\lambda$. Finally, let $\pi_3^* = \text{WIR}(p^{**}, v^*)$.

Following Mandelbaum and Stolyar (2004), we restrict attention to the case in which all queues are empty at time $t = 0$, but this assumption is removable. However, the fluid equations corresponding to the hydrodynamic model are significantly more complicated to analyze in the absence of this condition, and we choose to impose it for simplicity. An alternative, somewhat weaker sufficient condition is given by

$$\left| \frac{\widehat{Q}_i^{\lambda}(0)}{a_i} - \widehat{W}_{h,i}^{\lambda}(0) \right| \Rightarrow 0 \quad \text{as } \lambda \to \infty \text{ for all } i \in \mathcal{I}.$$

Our notion of asymptotic optimality for the delay-cost criterion is weaker then the one we have used in Theorem 3.1, because we will restrict the attention to sequences of controls π^{λ} that are asymptotically efficient. (We define nonanticipating controls in the next section.)

Definition 3.5 (Asymptotically Efficient Controls). A sequence of nonanticipating controls $\{\pi^{\lambda}\}$ is said to be asymptotically efficient if

$$\sup_{0 \leq t \leq T} \widehat{Q}_{\Sigma}^{\lambda}(t) \wedge \hat{I}_{\Sigma}^{\lambda}(t) \Rightarrow 0 \quad \text{as } \lambda \to \infty \tag{20}$$

whenever $\widehat{Q}_{\Sigma}^{\lambda}(0) \wedge \hat{I}_{\Sigma}^{\lambda}(0) \Rightarrow 0$, as $\lambda \to \infty$. We let Π^e be the family of asymptotically efficient control sequences; i.e., $\Pi^e := \{\{\pi^{\lambda}\}\colon \pi^{\lambda} \in \Pi_2,$ and the limit (20) holds$\}$.

Asymptotic efficiency implies that there cannot be a significant number of customers in any queue while there are idle agents in some of the agent pools. In the terminology of Atar (2005), we require asymptotic *joint work conservation*; see §2.4 therein. The restriction to asymptotically efficient controls is imposed to guarantee that the family $\{Q^\lambda_\Sigma(t), \lambda > 0\}$ is C-tight. The need for C-tightness should not be surprising, given the analysis in both Van Mieghem (1995) and Mandelbaum and Stolyar (2004). Indeed, in both these papers, C-tightness of the sequence of aggregate workloads plays a crucial role in establishing a lower bound in the delay-cost case. It will also play this role here; see the proof of Proposition A.1 in the online companion to this paper. However, in the conventional heavy-traffic setting the C-tightness need not be imposed, because it is obtained as a consequence of the complete resource-pooling condition, which trivially holds in the setting of Van Mieghem (1995). The complete resource-pooling condition guarantees that any work-conserving policy will asymptotically achieve the same aggregate workload process. It remains to determine if a similar result holds in the Halfin-Whitt regime. Hence, we restrict attention to asymptotically efficient controls. This seems reasonable for practical purposes, because asymptotically efficient controls are usually desirable. (We probably would not want to consider other alternatives.) As in Van Mieghem (1995), this assumption will play a key role in establishing a lower bound for the delay cost. It should be noted that the asymptotic efficiency of QIR and WIR follows from the corresponding state-space collapse results in Theorem 4.3 below and Theorem A.1 in the online companion. Consequently, $\pi_3^* \in \Pi^e$.

Our asymptotic optimality result in the delay-cost context follows.

Theorem 3.3 (Asymptotic Optimality of WIR for Delay-Cost Criterion). *If* $\widehat{Q}^\lambda_\Sigma(0) = 0$ *for all* λ *and*

$$\hat{I}^\lambda_j(0) \;\Rightarrow\; \hat{I}_j(0) \text{ as } \lambda \to \infty \quad \text{for all } j \in \mathcal{J},$$

then

$$\limsup_{\lambda\to\infty} J_2^\lambda(\pi_3^*, T) \le_{st} \liminf_{\lambda\to\infty} J_2^\lambda(\pi^\lambda, T) \tag{21}$$

for any $T > 0$ *and any sequence* $\pi^\lambda \in \Pi^e$. *Consequently,* π_3^* *is asymptotically optimal for the delay-cost criterion within the family* Π^e.

Remark 3.5. An analog of condition (14) holds to characterize when WIR reduces to the priority-type rule, using the new scaled cost functions $C_i^a(x) := C_i(x/a_i)$ instead of the original cost functions C_i. Note that the conditions to reduce to the priority-type rule are *not* equivalent for QIR and WIR, because the scaling by the ratios a_i changes the derivatives: $C_i^{a'}(x) = C_i'(x/a_i)/a_i$. Moreover, when a priority-type rule is optimal for both, as with linear costs, the low-priority class could be different for WIR and QIR, because the slopes change from c_i to c_i/a_i.

3.4. *D-Gc*μ Rule

Definition 3.6 (Many-Server *D-Gc*μ). For the SBR model with pool-dependent service rates, the *D-Gc*μ rule is defined from the many-server *Gc*μ rule by replacing the scheduling component with the following:

On service completion by a pool j *agent at time* t, the agent will next serve the customer from the head of queue i^*, where

$$i^* := i^*(t) \in \mathop{\arg\max}_{i \in I(j),\, \widehat{Q}^\lambda_i(t) > 0} C_i'(\widehat{W}^\lambda_{h,i}(t)).$$

If there are no customers waiting in any of the queues in $J(i)$, the agent will remain idle.

We let $\pi_4^* = D\text{-}Gc\mu$

Remark 3.6 (Comparing WIR and *D-Gc*μ). The *D-Gc*μ rule has a clear advantage over any form of WIR, because WIR requires knowledge of the ratios $A_i^\lambda(t)/A_\Sigma^\lambda(t)$, but *D-Gc*$\mu$ does not.

Theorem 3.4 (Asymptotic Optimality of *D-Gc*μ for Delay-Cost Criterion). *If, in addition to the assumptions of Theorem* 3.3, *the cost function* C_i *is continuous differentiable and strictly convex with* $C_i'(0) = C_i(0) = 0$ *for all* i, *then the D-Gc*μ *rule is asymptotically optimal for the delay-cost criterion within the family* Π^e.

4. Proofs

The line of reasoning we use to establish the asymptotic optimality results can be informally summarized as follows: first, we show that, with v^* as defined in (10), QIR(p, v^*) asymptotically minimizes the aggregate queue length in the system for *any* admissible ratio function p. Once this is established, it only remains to show that we can choose p^* to optimally distribute this aggregate queue among the different classes to minimize the convex holding costs.

To show that QIR with v^* asymptotically minimizes the aggregate queue length, we show that the aggregate queue length is bounded from below by a model with multiple agent pools but a single customer class, known as the inverted-V (or $\bigwedge$) model. For the $\bigwedge$ model, Armony (2005) showed that the faster-server-first (FSF) policy is asymptotically optimal; see Definition 4.2. We will show that with v^* the SBR model with QIR is asymptotically equivalent to its lower bound and hence asymptotically optimal with respect to the aggregate queue length. For the second step, we use the state-space collapse results for QIR established in Gurvich and Whitt (2007b).

Before proceeding to the actual proofs, we make some definitions and provide notational conventions to be used throughout. All the processes in consideration are constructed on a common probability space $(\Omega, \mathbb{F}, P)$. For an integer $d > 0$, we let $D^d := D^d[0,\infty)$ be the space of all right-continuous with left limit functions with values in the d-dimensional Euclidean space $\mathbb{R}^d$, equipped with the Skorohod J_1 metric; e.g., see chapter 3 of Whitt (2002). We will write $Y^\lambda(t) \Rightarrow Y$ in D^d to emphasize that we are considering processes in D^d instead of stationary distributions on $\mathbb{R}$.

Because the limit processes we consider are either the deterministic zero function or diffusion processes, the limit processes have continuous sample paths, so the notion of convergence on the underlying function space D^d coincides with uniform convergence on closed bounded intervals. To express that, for a vector-valued process $B(t)$ in D^d, let $\|B\|^*_{s,T};= \sup_{s\le t\le T}\|B(t)\|$, where $\|B(t)\| = \sum_{k=1}^{J}|B_k(t)|$. These are defined similarly for a process $B(t)$ in $D^{d\times m}[0,\infty)$, where now $\|B(t)\| = \sum_{k=1}^{d}\sum_{l=1}^{m}|B_{k,l}(t)|$.

We will also consider a weaker notion of convergence, using the space $D^d_- := D^d(0,\infty)$, where the domain is treated as open at the left instead of closed. We again let convergence (to continuous limits) be characterized by uniform convergence over bounded intervals. The restriction to the domain $(0,\infty)$ means that we exclude uniform convergence for intervals of the form $[0,b]$. We have $Y^\lambda(t) \Rightarrow 0$ in $D^d(0,\infty)$ if and only if, for each $0 < s < T < \infty$, $\|Y^\lambda\|^*_{s,T} \Rightarrow 0$. This weaker notion of convergence is required when discussing state-space collapse, as it might not hold at $t = 0$. It will, however, hold on compact subintervals of $(0,\infty)$.

We now formally define the set of admissible policies Π_1. Although the notion of nonanticipation is intuitively clear, it requires a careful definition that takes care of measurability issues and allows, for example, the application of martingale methods. In defining this concept, we follow Definition 2 in Atar (2005). Toward that end, let $Z^\lambda_{i,j}(t)$ be the number of pool j agents giving service to class i customers at time t. The number of service completions by pool j agents of class i customers in the time interval $[0,t]$ equals $S_{i,j}(\mu_j\int_0^t Z^\lambda_{i,j}(s)\,ds)$, where $S_{i,j}(\cdot)$, $i\in\mathcal{I}$, $j\in\mathcal{J}$ is a family of independent rate-1 Poisson processes.

For $i\in\mathcal{I}$ and $j\in J(i)$, let $A_{i,j}(t)$ be the number of class i customers to be routed *on arrival* to pool j. Similarly, for $i\in\mathcal{I}$ and $j\in\mathcal{J}(i)$, let $B_{i,j}(t)$ be the number of class i customers scheduled to receive service from a pool j agent after having waited in a queue, some of which will be served although others remain in service. We set $A_{i,j}(t) = B_{i,j}(t) := 0$ whenever $j\notin J(i)$. The system dynamics then satisfy the following equations:

$$Z^\lambda_{i,j}(t) = Z^\lambda_{i,j}(0) + A^\lambda_{i,j}(t) + B^\lambda_{i,j}(t) - S_{i,j}\left(\mu_j\int_0^t Z^\lambda_{i,j}(s)\,ds\right), \tag{22}$$

$$Q^\lambda_i(t) = Q^\lambda_i(0) + A^\lambda_i(t) - \sum_{j\in\mathcal{J}} A^\lambda_{i,j}(t) - \sum_{j\in\mathcal{J}} B^\lambda_{i,j}(t). \tag{23}$$

See §4 of Gurvich and Whitt (2007b) for a more detailed discussion of this construction.

We now define two families of σ fields. The system history up to time t is given by

$$\mathcal{F}_t := \sigma\big\{A^\lambda_i(s), Q^\lambda_i(s), Z^\lambda_{i,j}(s), S_{i,j}(T^\lambda_{i,j}(t)), A^\lambda_{i,j}(s), B^\lambda_{i,j}(s);\ i\in\mathcal{I},\ j\in\mathcal{J},\ s\le t\big\}.$$

We next define future events, starting after the interarrival times in progress at time t are complete. For that purpose, let $\tau^\lambda_i(t)$ be the time of the first class i arrival after time t; i.e., $\tau_i(t) = \inf\{u\ge t\colon\ A^\lambda_i(u) - A^\lambda_i(u-) > 0\}$. Let $T^\lambda_{i,j}(t) := \mu_j\int_0^t Z^\lambda_{i,j}(s)\,ds$. Then let

$$\mathcal{G}_t := \sigma\big\{A^\lambda_i(\tau^\lambda_i(t)+u) - A^\lambda_i(\tau^\lambda_i(t)), S_{i,j}(T^\lambda_{i,j}(t+u)) - S_{i,j}(T^\lambda_{i,j}(t));\ i\in\mathcal{I}, j\in\mathcal{J}, u\ge 0\big\}.$$

We let $D^d_\uparrow$ be the subspace of D^d that consists of nondecreasing functions, where for $\mathbb{R}^d$ we use the partial ordering induced by componentwise comparison.

Definition 4.1 (Nonanticipating Policies). A policy π is a mapping $\Omega \mapsto D_{\uparrow}^{2I\times 2J}$ taking ω into

$$\{(A_{i,j}(t), B_{i,j}(t);\ i\in\mathscr{I},\ j\in\mathscr{J}),\ t\geq 0\}(\omega)$$

such that $(Z^{\lambda}_{i,j}(t), Q^{\lambda}_{i}(t);\ i\in\mathscr{I}, j\in\mathscr{J}, t\geq 0)$ satisfies Equations (22) and (23). A policy π is said to be nonanticipating if

(i) for each $t\geq 0$, $\mathscr{F}_t$ is independent of $\mathscr{G}_t$, and

(ii) for each $i\in\mathscr{I}$, $j\in\mathscr{J}$ and $t\geq 0$, the process $S_{i,j}(T^{\lambda}_{i,j}(t)+\cdot) - S_{i,j}(T^{\lambda}_{i,j}(t))$ is equal in law to $S_{i,j}(\cdot)$. We let Π_1 be the set of nonanticipative policies.

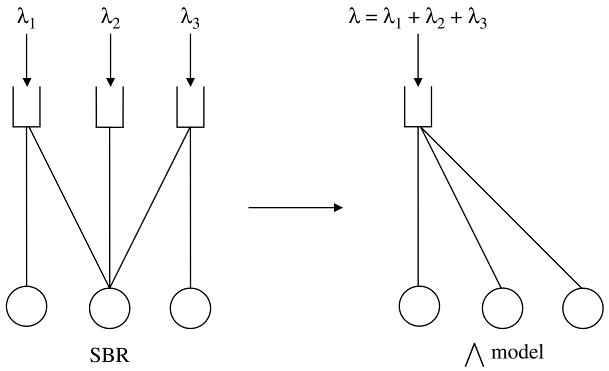
Recall that Π_2 is obtained from Π_1 by requiring that customers within each class are served on a FCFS basis. Observe that the value of $W_{h,i}(t)$ is included in the information provided by $\mathscr{F}_t$ and consequently both WIR and D-$Gc\mu$ are in Π_2.

4.1. $\bigwedge$ Model and FSF Policy

Some of the asymptotic optimality results in our previous paper (Gurvich and Whitt 2007a) relied heavily on the fact that the $M/M/N$ model served, in some sense, as a lower bound for the SBR system, with a common service rate μ. The model with a single customer class and multiple agent pools, known also as the inverted-V model (or $\bigwedge$) model, will serve as a corresponding lower bound for the SBR system here with pool-dependent service rates. More precisely, the optimally controlled $\bigwedge$ model will serve as our lower bound.

To present our stochastic ordering results, we let $\mathrm{SBR}(\mathscr{I}, \overrightarrow{\lambda}, \mathscr{J}, E, N, \mu)$ denote an SBR system with a set $\mathscr{I}$ of customer classes, arrival-rate vector $\overrightarrow{\lambda} = (\lambda_1,\dots,\lambda_I)$, a set $\mathscr{J}$ of agent pools, a routing graph E, staffing vector N, and pool-dependent service rates given by the vector $\mu=(\mu_1,\dots,\mu_J)$. The corresponding $\bigwedge$ model is denoted by $\bigwedge(\lambda,\mathscr{J},N,\mu)$ and stands for an inverted-V model with arrival rate $\lambda=\sum_{i\in\mathscr{I}}\lambda_i$ and a set $\mathscr{J}$ of agent pools with staffing vector N and service-rate vector μ. The set of admissible policies for the $\bigwedge$ model is the set of nonanticipating policies. Because the $\bigwedge$ model is a special case of the SBR model, we may use Definition 4.1 to define this set of policies. An example of an SBR system and its corresponding $\bigwedge$ model is given in Figure 2.

Figure 2 **An SBR Model and Its Corresponding $\bigwedge$ Model**



We will abbreviate and use SBR and $\bigwedge$ when the data $(\mathscr{I},\lambda,\mathscr{J},E,N,\mu)$ are clear from the context. The following result holds for each λ, so the superscript is omitted from the notation. Given admissible controls π_1 and π_2 for the SBR and $\bigwedge$ model, respectively, we let $Z^{\pi_1}_{j,\mathrm{SBR}}(t)$ and $Z^{\pi_2}_{j,\bigwedge}(t)$ be the corresponding number of busy agents in agent pool j in each of the systems under their respective controls. Similarly, we let $Q^{\pi_1}_{\Sigma,\mathrm{SBR}}(t)$ and $Q^{\pi_2}_{\Sigma,\bigwedge}(t)$ be the corresponding aggregate queue length processes. Here, the subscript Σ is used also for the $\bigwedge$ model only for purposes of notational consistency and the reader is reminded that the $\bigwedge$ model has only a single queue. We add a superscript to explicitly express the dependence on the control.

Lemma 4.1 ($\bigwedge$ Model as a Lower Bound). *Fix the data $(\mathscr{I},\mathscr{J},E,N,\overrightarrow{\lambda},\mu)$. Assume that*

$$(Z_{j,\mathrm{SBR}}(0), Q_{\Sigma,\mathrm{SBR}}(0);\ j\in\mathscr{J}) = (Z_{j,\bigwedge}(0), Q_{\Sigma,\bigwedge}(0)s;\ j\in\mathscr{J}).$$

Then, given any admissible policy π_1 for the SBR system, there exists an admissible policy π_2 for the $\bigwedge$ and a construction of the sample paths such that almost surely

$$\{Q^{\pi_1}_{\Sigma,\mathrm{SBR}}(t),\ t\geq 0\} = \{Q^{\pi_2}_{\Sigma,\bigwedge}(t),\ t\geq 0\}.$$

Consequently,

$$\{Q^{\pi_1}_{\Sigma,\mathrm{SBR}}(t),\ t\geq 0\} \overset{d}{=} \{Q^{\pi_2}_{\Sigma,\bigwedge}(t),\ t\geq 0\}.$$

The proof of Lemma 4.1 follows a very simple coupling argument based on the observation that any customer assignment that can be made in the SBR system also can be made in the corresponding $\bigwedge$ system. The complete formal argument is omitted. We now define the FSF policy proposed in Armony (2005).

Definition 4.2 (FSF Control for $\bigwedge$ Model). The FSF control is defined as follows:

On customer arrival. A customer that arrives at time $t \geq 0$ will be routed to the fastest available server, i.e, to agent pool j with

$$j \in \arg\max_{k \in \mathcal{J}:\, I_k^\lambda(t) > 0} \mu_k.$$

If all agents are busy, the customer will remain in queue.

On service completion. An agent that completes service will serve next a customer from the queue. If the queue is empty, the server will idle.

We now let $X^\lambda_{\Sigma,\bigwedge}(t)$ be the aggregate number of customers in the $\bigwedge$ model and let the scaled process be $\widehat{X}^\lambda_{\Sigma,\bigwedge}(t) := (X^\lambda_{\Sigma,\bigwedge}(t) - N^\lambda_\Sigma)/\sqrt{\lambda}$. Finally, let $\pi^*_\bigwedge := \text{FSF}$. The set of admissible policies for the $\bigwedge$ model is the set of nonanticipating policies, as defined in Definition 4.1. That is appropriate because the $\bigwedge$ model is a special case of an SBR model.

Theorem 4.1 (Asymptotic Optimality of FSF for $\bigwedge$ Model). *Fix any family of admissible policies $\{\pi^\lambda,\ \lambda > 0\}$ for the $\bigwedge$ model. Suppose that*

$$\widehat{X}^\lambda_{\Sigma,\bigwedge}(0) \ \Rightarrow\ \widehat{X}_{\Sigma,\bigwedge}(0) \text{ in } \mathbb{R} \quad \text{as } \lambda \to \infty.$$

Then, for each $T > 0$ and continuous nondecreasing function g,

$$\liminf_{\lambda\to\infty} \int_0^T g\big(\widehat{Q}^{\lambda,\pi^\lambda}_{\Sigma,\bigwedge}(t)\big)\,dt \geq_{\rm st} \limsup_{\lambda\to\infty} \int_0^T g\big(\widehat{Q}^{\lambda,\pi^*_\bigwedge}_{\Sigma,\bigwedge}(t)\big)\,dt.$$

Before proceeding with the proof of Theorem 4.1, we state a corollary that follows directly from Theorem 4.1 and Lemma 4.1.

Corollary 4.2 (A Lower Bound for SBR System). *For any family of admissible policies $\{\pi^\lambda,\ \lambda > 0\}$ for the SBR system and continuous nondecreasing function g,*

$$\liminf_{\lambda\to\infty} \int_0^T g\big(\widehat{Q}^{\lambda,\pi^\lambda}_{\Sigma,\mathrm{SBR}}(t)\big)\,dt \geq_{st} \limsup_{\lambda\to\infty} \int_0^T g\big(\widehat{Q}^{\lambda,\pi^*_\bigwedge}_{\Sigma,\bigwedge}(t)\big)\,dt,$$

for each $T \geq 0$.

Proof of Theorem 4.1. Theorem 4.1 follows from Armony (2005), but the result is not stated as such in her paper. Thus, we sketch how the different results in the Armony (2005) paper can be combined. We assume familiarity with the paper.

First, fix λ. Lemma 3.1 of Armony (2005) then establishes that there exists a sample-path construction such that a preemptive version of FSF (FSF_P) minimizes pathwise the aggregate number of customers in system. Because FSF_P is a work-conserving policy, minimizing the aggregate number of customers also minimizes the queue length. Returning to the sequence of scaled processes, we have then that the sequence of $\bigwedge$ models operated under FSF_P constitutes an asymptotic lower bound in terms of the queue length in the system.

By Proposition 4.2 and Remark 4.7 in Armony (2005), the sequence of $\bigwedge$ models operated under FSF has the same diffusion limit as the sequence operated under FSF_P. Consequently, the lower bound is asymptotically achieved and the result holds.

To complete the proof, we observe that, although Armony (2005) assumes Poisson arrivals, the results that we used above are easily extended to renewal arrivals. Indeed, Lemma 3.1 of Armony (2005) is a sample-path argument that does not use the Poisson structure of the arrivals, so it holds for any arrival process. Proposition 4.2 in Armony (2005) relies on two results. The first is the state-space collapse result (Proposition 4.1 there) that can be easily extended to renewal arrivals; see Remark 4.1 below. The second step is an functional central limit theorem (FCLT) result, Proposition 4.2, that is proved through a martingale decomposition approach. Using the FCLT for renewal processes, it is readily verified that Proposition 4.2 and Remark 4.7 remain valid (with appropriate change of the infinitesimal variance term) for renewal processes; see Theorem 7.6 of Pang et al. (2007) for an illustrative example. □

We now show that QIR with appropriately chosen parameters achieves asymptotically the same aggregate number of customers as the optimally controlled $\bigwedge$ model. Recall that v^* is the nonstate-dependent ratio function given by $v^* \equiv (0, 0, \ldots, 1)$. For the following, recall that $\pi^*_\bigwedge$ is the FSF policy for the $\bigwedge$ model.

Proposition 4.1 (Asymptotic Equivalence with $\bigwedge$ Model). *Fix any ratio function p and let $\tilde{\pi} :=$ QIR(p, v^*). If $\widehat{X}^\lambda_\Sigma(0) = \widehat{X}^\lambda_{\Sigma,\bigwedge}(0)$ for all λ and*

$$\widehat{X}^\lambda_\Sigma(0) \ \Rightarrow\ \widehat{X}_\Sigma(0) \text{ in } \mathbb{R} \quad \text{as } \lambda \to \infty,$$

then both

$$\widehat{X}^{\lambda,\tilde{\pi}}_{\Sigma,\mathrm{SBR}}(t) \ \Rightarrow\ \widehat{X}_\Sigma(t) \text{ in } D \quad \text{as } \lambda \to \infty \tag{24}$$

and

$$\widehat{X}^{\lambda,\,\pi^*_\wedge}_{\Sigma,\,\wedge}(t) \;\Rightarrow\; \widehat{X}_\Sigma(t) \textit{ in } D \quad \textit{as } \lambda\to\infty, \tag{25}$$

where $\widehat{X}_\Sigma(t)$ *is the unique solution to the following one-dimensional* SDE:

$$\widehat{X}_\Sigma(t) = \widehat{X}_\Sigma(0) - \beta t + \mu_J \int_0^t [\widehat{X}_\Sigma(s)]^- \, ds + \sqrt{1+c_a^2}B(t), \tag{26}$$

with $\{B(t),\, t\ge 0\}$ *being a standard Brownian motion and* $c_a^2 = \sum_{i\in\mathcal{I}} c_{a,\,i}^2$.

Proof. Equation (24) follows from Theorem 5.1 in Gurvich and Whitt (2007b) by setting $\theta_i = 0$ for all $i\in\mathcal{I}$ and replacing the Poisson arrivals with the renewal arrivals; see Remark 4.1 below. It is easy to see that Equation (128) in Gurvich and Whitt (2007b) remains valid provided that $\theta_i = 0$ for all $i\in\mathcal{I}$. Equation (25) follows from Proposition 4.2 in Armony (2005), with the appropriate replacement of scaling (by $\sqrt{N^\lambda}$ rather than $\sqrt{\lambda}$) and after replacing the Poisson arrival process with the renewal process. □

The implication of Proposition 4.1 is that QIR asymptotically achieves the lower bound given by the $\wedge$ model provided that the ratio function v^* is used for the routing component. Consequently, QIR minimizes the aggregate number of customers in the system as well as the aggregate queue length. This aggregate-queue-length optimality replaces the invariance phenomenon in Mandelbaum and Stolyar (2004).

In the next two subsections we will focus mainly on the general SBR model (rather then the $\wedge$ model). Hence, we omit the subscript SBR from all notation. We will explicitly use the subscript $\wedge$ when referring to the inverted-V model.

4.2. Asymptotic Optimality for Holding-Cost Formulation

The following is a direct consequence of Theorem 3.1 in Gurvich and Whitt (2007b).

Theorem 4.3 (State-Space Collapse Under QIR with Pool-Dependent Rates). *If* $(\widehat{X}^\lambda(0), \widehat{Z}^\lambda(0)) \Rightarrow (\widehat{X}(0), \widehat{Z}^\lambda(0))$ *in* $\mathbb{R}^{I+I\cdot J}$, *then we have state-space collapse*:

$$\widehat{Q}_i^\lambda(t) - \widehat{Q}_\Sigma^\lambda(t)p_i(\widehat{Q}_\Sigma^\lambda(t)) \;\Rightarrow\; 0 \textit{ in } D_- \qquad \textit{as } \lambda\to\infty,\; i\in\mathcal{I}, \tag{27}$$

and

$$\hat{I}_j^\lambda(t) - \hat{I}_\Sigma^\lambda(t)v_j(\hat{I}_\Sigma^\lambda(t)) \;\Rightarrow\; 0 \textit{ in } D_- \qquad \textit{as } \lambda\to\infty,\; j\in\mathcal{J}. \tag{28}$$

The convergence in (27) *and* (28) *is strengthened to convergence in D if we assume that*

$$\begin{aligned} &\widehat{Q}_i^\lambda(0) - \widehat{Q}_\Sigma^\lambda(0)p_i(\widehat{Q}_\Sigma^\lambda(0)) \;\Rightarrow\; 0, \quad i\in\mathcal{I}, \quad \textit{and} \\ &\hat{I}_j^\lambda(0) - \hat{I}_\Sigma^\lambda(0)v_j(\hat{I}_\Sigma^\lambda(0)) \;\Rightarrow\; 0, \quad j\in\mathcal{J}. \end{aligned} \tag{29}$$

Remark 4.1 (Relaxing Poisson-Arrival Assumption in Gurvich and Whitt 2007b). Theorem 3.1 in Gurvich and Whitt (2007b) was proved under the assumption of Poisson arrivals, but the proof is easily changed to allow for renewal arrival processes. For that purpose, Lemma 4.3 in Gurvich and Whitt (2007b) needs to be slightly changed to take care of renewal arrivals. That can be done along the lines of Proposition 6.1 in Dai and Tezcan (2009). We omit the details here.

Before we proceed, we state a lemma that will be used in the proof of Theorem 3.1. It allows us to conclude convergence of integrals starting from 0 despite the fact that state-space collapse holds only in D_-. This lemma appears as Proposition 5.2 in Gurvich and Whitt (2007b) and is proved there.

Lemma 4.2. *Fix* $T > 0$. *Let* f_i: $\mathbb{R}\mapsto\mathbb{R}$ *be continuous functions for all* $i\in\mathcal{I}$. *Then, under the conditions of Proposition* 4.1,

$$\begin{aligned} &\left(\int_0^T f_1(\widehat{Q}_i^\lambda(t))\,dt, \ldots, \int_0^T f_I(\widehat{Q}_i^\lambda(t))\,dt\right) \\ &\quad\Rightarrow \left(\int_0^T f_1([\widehat{X}_\Sigma(t)]^+p_i([\widehat{X}_\Sigma(t)]^+))\,dt, \ldots, \right. \\ &\qquad\left.\int_0^T f_I([\widehat{X}_\Sigma(t)]^+p_i([\widehat{X}_\Sigma(t)]^+))\,dt\right) \textit{ in } \mathbb{R}^I \\ &\qquad\qquad \textit{as } \lambda\to\infty, \end{aligned} \tag{30}$$

where $\widehat{X}_\Sigma(t)$ *is the diffusion process from Proposition* 4.1

Proof of Theorem 3.1. Start by fixing a family $\{\pi^\lambda,\, \lambda>0\}$ of admissible policies, i.e., $\pi^\lambda\in\Pi_1$ for all $\lambda>0$. Then, by the definition of $q_i^*(x)$, we have that

$$\sum_{i=1}^I C_i\big(\widehat{Q}_i^{\lambda,\,\pi^\lambda}(t)\big) \ge \sum_{i=1}^I C_i\big(q_i^*\big(\widehat{Q}_\Sigma^{\lambda,\,\pi^\lambda}(t)\big)\big), \tag{31}$$

for all λ and $t\ge 0$. In particular, by Lemma 4.1, there exists an admissible policy $\tilde{\pi}^\lambda$ for the $\wedge$ model so that

$$\begin{aligned} \inf_{\pi^\lambda\in\Pi_1} J_1^\lambda(\pi^\lambda) \;&\ge\; \inf_{\pi^\lambda\in\Pi_1}\int_0^T \sum_{i\in\mathcal{I}} C_i\big(q_i^*\big(\widehat{Q}_\Sigma^{\lambda,\,\pi^\lambda}(t)\big)\big)\,dt \\ &\ge_{\text{st}} \int_0^T \sum_{i\in\mathcal{I}} C_i\big(q_i^*\big(\widehat{Q}_\wedge^{\lambda,\,\tilde{\pi}^\lambda}(t)\big)\big)\,dt, \end{aligned} \tag{32}$$

where, $\widehat{Q}_{\wedge}^{\lambda,\tilde{\pi}^{\lambda}}(t)$ is the queue length at time t in the corresponding $\bigwedge$ model with initial scaled queue length $\widehat{Q}_{\Sigma}^{\lambda}(0)$, initial scaled idleness vector $\hat{I}^{\lambda}(0)$, and operated under a control $\tilde{\pi}^{\lambda} \in \widetilde{\Pi}^{\lambda}$. By Theorem 4.1,

$$\liminf_{\lambda\to\infty} \int_0^T \sum_{i\in\mathcal{I}} C_i\big(q_i^*\big(\widehat{Q}_{\wedge}^{\lambda,\tilde{\pi}^{\lambda}}(t)\big)\big)\,dt$$

$$\geq_{\text{st}} \limsup_{\lambda\to\infty} \int_0^T \sum_{i\in\mathcal{I}} C_i\big(q_i^*\big(\widehat{Q}_{\wedge}^{\lambda,\,\pi_{\wedge}^*}(t)\big)\big)\,dt. \qquad (33)$$

Applying Lemma 4.2 to the right-hand side (observe that this can be done, as the $\bigwedge$ model is just a special case of an SBR system), we have that

$$\liminf_{\lambda\to\infty} \int_0^T \sum_{i\in\mathcal{I}} C_i\big(q_i^*\big(\widehat{Q}_{\wedge}^{\lambda,\tilde{\pi}^{\lambda}}(t)\big)\big)\,dt$$

$$\geq_{\text{st}} \int_0^T \sum_{i\in\mathcal{I}} C_i(q_i^*([\widehat{X}_{\Sigma}(t)]^+))\,dt, \qquad (34)$$

where $\widehat{X}_{\Sigma}(t)$ is the limit process in Equation (26). Consequently,

$$\limsup_{\lambda\to\infty} \inf_{\pi^{\lambda}\in\Pi_1} J_1^{\lambda}(\pi^{\lambda}) \geq_{\text{st}} \int_0^T \sum_{i\in\mathcal{I}} C_i(q_i([\widehat{X}_{\Sigma}(t)]^+))\,dt. \qquad (35)$$

Having established an asymptotic lower bound, it only remains to show that this lower bound is asymptotically achieved by $\pi_1^* = \text{QIR}(p^*, v^*)$. Indeed, by Lemma 4.2

$$\int_0^T \sum_{i\in\mathcal{I}} C_i\big(\widehat{Q}_i^{\lambda,\,\pi_1^*}(t)\big)\,dt \;\Rightarrow\; \int_0^T \sum_{i\in\mathcal{I}} C_i(q_i([\widehat{X}_{\Sigma}(t)]^+))\,dt,$$

$$\text{as } \lambda\to\infty. \qquad (36)$$

Consequently, fixing any family π'^{λ} of admissible controls, we have that

$$\limsup_{\lambda\to\infty} J_1^{\lambda}(\pi^{*\lambda}) \leq_{\text{st}} \limsup_{\lambda\to\infty} J_1^{\lambda}(\pi'^{\lambda}), \qquad (37)$$

which establishes the asymptotic optimality of π_1^*. □

Proof of Theorem 3.2. By Lemma 3.2, the optimal ratio function p^* is given by $p_i^*(x) := C_i'^{-1}(y(x))/x$ for all $x > 0$. In particular, $\text{QIR}(p^*, v^*)$ chooses to serve next a class i customer with

$$i \in \operatorname*{arg\,max}_{k\in I(j):\,Q_k(t)>0} \widehat{Q}_k^{\lambda}(t) - C_k'^{-1}(y([\widehat{X}_{\Sigma}^{\lambda}(t)]^+)).$$

The result now follows from Remark 2.2 in Gurvich and Whitt (2007b). Specifically, the same asymptotic cost, as given by $\sum_{i\in\mathcal{I}} C_i(q_i([\widehat{X}_{\Sigma}(t)]^+))\,dt$, is achieved by any control that assigns to an available agent a customer from the set

$$\mathcal{U}^+ := \big\{k \in I(j)\colon\ \widehat{Q}_k^{\lambda}(t) > 0\colon\ \widehat{Q}_k^{\lambda}(t) > C_k'^{-1}(y([\widehat{X}_{\Sigma}^{\lambda}(t)]^+))\big\}.$$

As $C_i'^{-1}$ is a strictly increasing function, we also have that

$$\mathcal{U}^+ = \big\{k \in I(j)\colon\ \widehat{Q}_k^{\lambda}(t) > 0\colon\ C_k'(\widehat{Q}_k^{\lambda}(t)) > y([\widehat{X}_{\Sigma}^{\lambda}(t)]^+)\big\}.$$

Applying Remark 2.2 in Gurvich and Whitt (2007b) once again, we have that the same performance analysis is achieved by any control that assigns an available agent to a customer from class i with

$$i \in \operatorname*{arg\,max}_{k\in I(j):\,Q_k(t)>0} C_k'(\widehat{Q}_k^{\lambda}(t)) - y([\widehat{X}_{\Sigma}^{\lambda}(t)]^+).$$

This, in turn, is equivalent to choosing

$$i \in \operatorname*{arg\,max}_{k\in I(j):\,Q_k(t)>0} C_k'(\widehat{Q}_k^{\lambda}(t)),$$

which is precisely the $Gc\mu$ rule. □

Remark 4.2 (Proof of Theorems 3.3 and 3.4). The proof of Theorem 3.3 mostly follows the proof of Theorem 3.1 above. However, it requires some additional side results. These, as well as the proof of Theorem 3.4, are given in the online companion to this paper.

5. Conclusions and Directions for Future Research

In this paper we have established asymptotic optimality in the many-server heavy-traffic regime of special versions of the QIR and WIR rules for minimizing convex holding costs in many-server systems with multiple customer classes and agent pools and pool-dependent service rates. We have shown that simple elegant policies arise under extra regularity conditions. For strictly convex holding and delay costs (plus other regularity conditions), the scheduling components of our asymptotically optimal policies reduce to the $Gc\mu$ and D-$Gc\mu$ rules, respectively.

Consequently, our results extend the conventional heavy-traffic results of Mandelbaum and Stolyar (2004) to the Halfin-Whitt regime. However, this extension is only partial because we had to restrict attention to pool-dependent service rates.

The diffusion limits we obtain under QIR and $Gc\mu$ are identical to those obtained by solving the HJB equations in Atar (2005) when imposing the additional assumption of pool-dependent service rates. The solution of the diffusion control problem, however, leaves the question of finding the asymptotically optimal controls open. We make an important step forward in identifying simple asymptotically optimal controls (QIR and $Gc\mu$). The asymptotic optimality of QIR can not be deduced from Atar (2005), except for some very limited cases—we refer the reader to Gurvich and Whitt (2007b), where cases in which QIR is almost equivalent to Atar's control are identified.

It remains to be determined if simple elegant controls are asymptotically optimal with more general service rates. Limitations on what can be achieved follow from previous work: Harrison and Zeevi (2004) provide an example where a complicated bang-bang control is asymptotically optimal for the V model with linear holding costs; Atar (2005) relates the asymptotic optimality to the optimal solution of a related Brownian control problem, which in most cases results in a complex solution. Although these examples and others are discouraging, there may well exist interesting subclasses of models with elegant asymptotically optimal solutions.

It also remains to identify the class of *all* asymptotically optimal solutions. To what extent is that class large or small? It also remains to investigate how the asymptotically optimal policies perform for actual systems at typical loads.

Electronic Companion

An electronic companion to this paper is available on the *Manufacturing & Service Operations Management* website (http://msom.pubs.informs.org/ecompanion.html).

Acknowledgments

The authors thank Avi Mandelbaum, Gennady Shaikhet, and the referees for helpful comments. The second author was supported by NSF Grant DMI-0457095.

References

Armony, M. 2005. Dynamic routing in large-scale service systems with heterogenous servers. *Queueing Systems* **51**(3–4) 287–329.

Atar, R. 2005. Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Probab.* **15**(4) 2606–2650.

Atar, R., A. Mandelbaum, M. Reiman. 2004. Scheduling a multiclass queue with many iid servers: Asymptotic optimality in heavy-traffic. *Ann. Appl. Probab.* **14**(3) 1084–1134.

Bramson, M. 1998. State space collapse with applications to heavy-traffic limits for multiclass queueing networks. *Queueing Systems* **30**(1–2) 89–148.

Dai, J. G., T. Tezcan. 2008a. Dynamic control of N-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. *Oper. Res.* Forthcoming.

Dai, J. G., T. Tezcan. 2008b. Optimal control of parallel server systems with many servers in heavy traffic. *Queueing Systems.* Forthcoming.

Dai, J. G., T. Tezcan. 2009. State space collapse in many server diffusion limits of parallel server systems. *Math. Oper. Res.* Forthcoming.

Gurvich, I., W. Whitt. 2007a. Service-level differentiation in many-server service systems: A solution based on fixed-queue-ratio routing. Working paper, Columbia University, New York.

Gurvich, I., W. Whitt. 2007b. Queue-and-idleness-ratio controls in many-server service systems. Working paper, Columbia University, New York.

Gurvich, I., M. Armony, A. Mandelbaum. 2008. Service-level differentiation in call centers with fully flexible servers. *Management Sci.* **54**(2) 279–294.

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29**(3) 567–587.

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

Harrison, J. M. 1998. Heavy traffic analysis of a system with parallel servers: Asymptotic optimality of discrete-review policies. *Ann. Appl. Probab.* **8**(3) 822–848.

Harrison, J. M., A. Zeevi. 2004. Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime. *Oper. Res.* **52**(2) 243–257.

Ibaraki, T., N. Katoh. 1988. *Resource Allocation Problems: Algorithmic Approaches*. No. 4, Foundations of Computing Series, MIT Press, Cambridge, MA.

Mandelbaum, A., A. Stolyar. 2004. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Oper. Res.* **52**(6) 836–855.

Pang, G., R. Talreja, W. Whitt. 2007. Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys* **4** 193–267.

Patriksson, M. 2008. A survey on the continuous nonlinear resource allocation problem. *Eur. J. Oper. Res.* **185**(1) 1–46.

Stidham, S. 1993. A survey of Markov decision models for control of networks of queues. *Queueing Systems* **13**(1–3) 291–314.

Van Mieghem, J. A. 1995. Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *Ann. Appl. Probab.* **5**(3) 809–833.

Whitt, W. 2002. *Stochastic-Process Limits: An Introduction to Stochastic Process Limits and Their Application to Queues*. Springer-Verlag, New York.

Zipkin, P. H. 1980. Simple ranking methods for the allocation of one resource. *Management Sci.* **26** (1) 34–43.