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

Itay Gurvich* Ward Whitt †

July 28, 2007

Abstract

In a recent paper we introduced the fixed-queue-ratio (FQR) family of routing rules for many-server service systems with multiple customer classes and server pools. A newly available server next serves the customer from the head of the queue of the class (from among those he is eligible to serve) whose queue length most exceeds a specified proportion of the total queue length. Under fairly general conditions, FQR 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 FQR 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 FQR 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.

Acknowledgement

We thank Gennady Shaikhet for pointing out a missing assumption in §2.

*Columbia Business School, 4I Uris Hall, 3022 Broadway, New York, NY 10027. (ig2126@columbia.edu)

†IEOR Department, Columbia University, 304 S. W. Mudd Building, 500 West 120th Street, New York, NY 10027-6699. (ww2040@columbia.edu)

1 Introduction

The optimal control of queueing systems to minimize holding costs or maximize revenues has been the subject of extensive literature; e.g., see Stidham [18]. Many have exploited Markov decision processes (MDP's), but the scope of the MDP 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 Mandelbaum and Stolyar [14], which extended the seminal multi-class single-server work of Van Mieghem [19], building on Harrison [11], to a setting with multiple non-identical servers. In these papers a queue is formed for each customer class. Mandelbaum and Stolyar [14] 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

$$i \in \underset{i}{\operatorname{argmax}}\, \mu_{i,j} C_i'(Q_i(t)).$$

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 [14] builds on a useful invariance principle that implies, under a complete-resource-pooling condition on the network structure, that the aggregate workload in the system is asymptotically minimized by *any* work-conserving policy. As a consequence, minimizing the convex holding cost requires only appropriately distributing the aggregate workload between the different customer classes. That is achieved through a state-space-collapse result, building on the general framework introduced by Bramson [4].

The restriction of [14] 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 in parallel-server systems with single-server stations in the conventional heavy-traffic limit. In fact, applying the $c\mu$ rule to these systems may be disastrous, leading to "system explosion," as explained in §1.1 of Dai and Tezcan [6].

The purpose of the present paper is to extend, as much as possible, the result of [14] to the many-server heavy-traffic limiting regime introduced in the Halfin and Whitt [10]. Unlike [14], our results *do* cover linear holding costs, thus underscoring the significant differences between the two heavy-traffic regimes. Again, §1.1 of [6] 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. In order to establish a related invariance principle in the many-server setting, we restrict attention to multi-server 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.)

But 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 upon customer arrival, and *a scheduling component* - specifying what to do upon service completion.

Our solution builds on the fixed-queue-ratio (FQR) family of controls introduced in Gurvich and Whitt [8]. Moreover, our proofs here draw heavily upon the technical appendix to [8] in [9]. Here we show that a special version of FQR (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 FQR 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 multi-class but single-pool (single-server-type) model (the V-model) with linear holding costs was considered by Harrison and Zeevi [12]. 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. [2], where a threshold policy was proposed. Armony [1] stochastically minimized the queue-length in a multi-server-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 [3]. In his far-reaching paper, Atar covers asymptotic minimization of holding costs in general multi-class multi-pool systems in the Halfin-Whitt regime. His analysis focuses on the 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 [8, 9] are closely related to concurrent and independent results by Dai and Tezcan [6, 7]. Their first paper [6] took the important path of constructing explicit solutions for specific cases, assuming linear holding costs. Their first paper [6] considers the N model, having two customer classes and two agent pools - one of which is dedicated and the other flexible. Their second paper [7] is a generalization of the first, considering general SBR systems with pool-dependent rates, just as in this paper, but still focusing on linear holding costs. Our current paper extends their result to more general convex holding costs. Thus, our current paper includes their results as a special case. Moreover, we show that the policy that [7] shows to be optimal for linear cost functions is also optimal much more generally; linearity is sufficient, but it is not actually the critical feature.

As should be expected, the analysis here is similar to the analysis in [7], but there are significantly differences. Dai and Tezcan [7] build on their previous important work [5] extending Bramson's [4] state-space collapse framework to the Halfin-Whitt many-server regime. In contrast, we apply our own previous paper [9] and Armony [1].

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 fixed-waiting-ratio (FWR) routing rule. We provide the proofs, building on [9], in §4. Finally, we make concluding remarks in §5.

2 The Model

We consider a system with a fixed set $\mathcal{I} := \{1, \ldots, I\}$ of customer classes and a fixed set $\mathcal{J} := \{1, \ldots, J\}$ of server pools. There is a queue for each customer class. If customers cannot enter service immediately upon 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 (scv, variance divided by the square of the mean) $c^2_{a,i}$. Given a vector $\overrightarrow{\lambda} = (\lambda_1, \ldots, \lambda_I)$ 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 $V = \mathcal{J} \bigcup \mathcal{I}$; 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 serve class-i customers. In order 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., $I(j) := \{i \in \mathcal{I} : (i,j) \in E\}$; $I(j)$ 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., $J(i) := \{j \in \mathcal{J} : (i,j) \in E\}$. Motivated by the application to call centers, we call these systems **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*

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

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

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

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 [14]. However, we emphasize that, in contrast to the conventional heavy-traffic regime, not any work-conserving policy will asymptotically achieve the minimal aggregate queue length in the many-server regime; 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 [12] 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.

The 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 $\mu_j,\ j \in \mathcal{J}$, 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 (**the many-server heavy-traffic regime**) *Assume that*

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

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

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

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 $\sum_{j\in\mathcal{J}} \mu_j N_j$ 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*

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

and such that the graph $\mathcal{E}(x) := \{(i,j) \in \mathcal{I}\times\mathcal{J} : x_{ij} > 0\}$ *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). 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 both in the 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 $\mathcal{E}(x)$ 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 [8]. Finally, we point out that this assumption is consistent with the heavy-traffic assumption in [8] - see Remark 4.2 in [8] that relates the optimization setting discussed in that paper to settings with fixed staffing levels, like the one discussed here.

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

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

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 [10].

For the λ^{th} system, let $Q_i^\lambda(t)$ be the number of class-i customers in 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

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

and

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

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

$$J_1^\lambda(\pi,T) := \int_0^T \sum_{i=1}^I C_i(\hat{Q}_i^{\pi,\lambda}(t))dt \quad \text{and} \quad J_2^\lambda(\pi,T) := \frac{1}{A_i^\lambda(t)} \sum_{i=1}^I \sum_{k=1}^{A_i^\lambda(T)} C_i(\sqrt{\lambda}W_{i,k}^\lambda(\pi))$$

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 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 nonde-*

creasing and convex with $C_i(0) = 0$.

The essential part of Assumption 2.5 is the convexity. (The rest of the assumptions are without loss of generality, because we can add an arbitrary constant to each cost function and we can add a common linear function cx to all cost functions without affecting the solution of the nonlinear program (11) below.)

Remark 2.1 (**comparison to the assumptions in [14]**) Our assumptions on the cost functions are substantially weaker than those imposed in [14]. In contrast to [14], 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 non-anticipating policies; see §4 for a formal definition. The set Π_2 will be restricted to policies that serve customers first-come first-served (FCFS) 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 $\{\tilde{\pi}^\lambda\}$,

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

where $\leq_{st}$ denotes (conventional) stochastic ordering. 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 The Main Results

In this section we state our main results. We establish the asymptotic optimality of special FQR and analogous fixed-waiting-ratio (FWR) rules for the holding-cost and delay-cost formulations, respectively.

3.1 The Holding-Cost Formulation

We start by formally defining a version of FQR that allows for general state-dependent ratio functions; see [8, 9] 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

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

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 $(0, \infty)$.*

Definition 3.2 (**FQR for admissible state-dependent ratio functions**) *Given two admissible state-dependent ratio functions v and p, FQR is defined as follows:*

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

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

  i.e., the customer will be routed to an agent pool with the greatest idleness imbalance. If there are no such agents, the customer waits in queue i, to be served in order of arrival.

- **Upon 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*

$$i^* := i^*(t) \in \underset{i \in I(j), \hat{Q}_i^\lambda(t) > 0}{\operatorname{argmax}} \left\{ \hat{Q}_i^\lambda(t) - [\hat{X}_\Sigma^\lambda(t)]^+ p_i \left([\hat{X}_\Sigma^\lambda(t)]^+ \right) \right\};$$

i.e., the agent will admit a customer from the queue with the greatest queue imbalance. If there are no such customers, the agent will remain idle.

Ties are broken in an arbitrary but consistent manner, so that the vector-valued stochastic process

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

is a continuous-time Markov chain (CTMC) with stationary transition probabilities.

For given ratio functions v and p, we denote the resulting FQR control by FQR(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 non-state-dependent ratio function given by

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

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)

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

where C_i are 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 [13], Patriksson [15] and Zipkin [23]. 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 (**the 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*

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

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 $p_i^*(0) \ge 0$ for all $i \in \mathcal{I}$ and $\sum_{i\in\mathcal{I}} p_i^*(0) = 1$.

Finally, let $\pi_1^* := FQR(p^*, v^*)$ 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 FQR for the holding-cost criterion**) *If*

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

then

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

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 (**some intuition**) 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 for the slowest ones, constantly busy. By doing so, the control asymptotically minimizes the aggregate queue-length in the system. Since 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. This reasoning is very similar to the reasoning in Mandelbaum and Stolyar [14], with the exception that in conventional heavy traffic *any* work-conserving policy asymptotically achieves the minimal aggregate workload. ■

In general, the scheduling component of this version of FQR 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 non-state-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 FQR. First, it is easily verified that for the N-model setting with linear holding costs analyzed in [6], FQR reduces to the static-priority $c\mu$ rule. Indeed, FQR is equivalent to a static priority rule for all settings in which the set

$$\operatorname*{argmax}_{i\in I(j),\hat{Q}_i^\lambda(t)>0} \left\{\hat{Q}_i^\lambda(t) - [\hat{X}_\Sigma^\lambda(t)]^+ p_i\left([\hat{X}_\Sigma^\lambda(t)]^+\right)\right\}$$

is identical to the set

$$\operatorname*{argmax}_{i\in I(j),\hat{Q}_i^\lambda(t)>0} \hat{Q}_i^\lambda(t),$$

and one can verify that the N-model of [6] indeed satisfies this restriction. As observed in [7], this restriction does not hold for most networks. Instead, a more general priority-type rule is proposed in §2.1 of that same paper. This more general rule is, however, identical to the FQR rule we obtain for linear holding costs. Consequently, in the absence of abandonments, Theorem 3.2 covers the results of both [6] and [7] as special cases.

It is significant that the same simple priority-type rule, as proposed in [7] 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. Since C_i is convex, the derivative C_i' exists at all but countably many points and is nondecreasing. Hence, the **necessary and 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

$$C_{i_0}'(\infty) := \lim_{x\to\infty} C_i'(x) \le C_i'(0+) \quad \text{for all} \quad i \neq i_0\ . \tag{14}$$

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. ■

Remark 3.3 (**a single switching point**) 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^*, while 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

$$C_{i_0}'(x^*) \le C_i'(0+) \quad \text{for all} \quad i \neq i_0 \tag{15}$$

and

$$C_{i_1}'(\infty) \le \min\{C_{i_0}'(x^*+), C_i'(0+)\} \quad \text{for all} \quad i \notin \{i_0, i_1\}\ . \tag{16}$$

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: $C_1(x) = c_1 x$, $x \ge 0$, $C_2(x) = b_2 x$, $0 \le x < x^*$ and $C_2(x) = d_2(x - x^*) + b_2 x^*$, $x \ge x^*$, where $b_2 < c_1 < d_2$.

In this single-switching-point setting we have just considered, there is a threshold x^*. If the scaled aggregate queue length $\hat{X}_\Sigma^\lambda(t)$ 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 FQR 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 $\hat{X}^{\lambda}_{\Sigma}(t)$ increases, so we have a level-dependent weighted-priority scheme. Eventually, as $\hat{X}^{\lambda}_{\Sigma}(t)$ 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 (**adding abandonment**) Unfortunately, in general, our FQR 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.

When all holding costs are linear, Dai and Tezcan [6] 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 FQR 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.

Since 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 $i \in \mathcal{I}$. For this trivial case, our results for the holding-cost criterion indeed hold without any change. 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 The $Gc\mu$ 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 FQR control can be somewhat complicated. We now show that our FQR policy reduces to the $Gc\mu$ rule when the costs are strictly convex and satisfy additional regularity conditions. Since costs in practice often are regarded as convex instead of linear, we regard this as an important conclusion, just as in [19] and [14].

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

Definition 3.3 (**the 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 FQR*(p^*, v^*) *to:*

Upon 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*

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

If there are no such customers, 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 the** $Gc\mu$ **rule for the 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 optimal policy* π_1^* *can be replaced with* π_2^**; i.e., the* $Gc\mu$ *rule is asymptotically optimal.*

We now prove 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:

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

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 resource 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*

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

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

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, FQR(p^*, v^*) chooses to serve next a class-i customer with

$$i \in \underset{k \in I(j): Q_k(t) > 0}{\operatorname{argmax}} \; \hat{Q}_k^\lambda(t) - C_k'^{-1}\left(y([\hat{X}_\Sigma^\lambda(t)]^+)\right).$$

Since $C_i'^{-1}$ is a strictly increasing function where it assumes a positive finite value, this is equivalent to choosing

$$i \in \underset{k \in I(j): Q_k(t) > 0}{\operatorname{argmax}} \; C_k'(\hat{Q}_k^\lambda(t)) - y([\hat{X}_\Sigma^\lambda(t)]^+).$$

and, in turn, equivalent to choosing

$$i \in \underset{k \in I(j): Q_k(t) > 0}{\operatorname{argmax}} \; C_k'(\hat{Q}_k^\lambda(t)),$$

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

3.3 The 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 FWR, which in turn is a simple modification of FQR 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 by D-$Gc\mu$ rule, following [14]. 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 $\hat{W}^{\lambda}_{h,i}(t) = \sqrt{\lambda} W^{\lambda}_{h,i}(t)$.

Definition 3.4 (**the FWR rule**)

For the SBR model with pool-dependent service rates, FWR(p, v) is defined from FQR by replacing the scheduling component p with:

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

$$i^* := i^*(t) \in \underset{i \in I(j), \hat{Q}^{\lambda}_i(t) > 0}{\operatorname{argmax}} \left\{ \frac{A^{\lambda}_i(t)}{A^{\lambda}_{\Sigma}(t)} \hat{W}^{\lambda}_{h,i}(t) - [\hat{X}^{\lambda}_{\Sigma}(t)]^+ p_i \left([\hat{X}^{\lambda}_{\Sigma}(t)]^+ \right) \right\};$$

  If there are no such customers, the agent will remain idle.

For given ratio functions p and v, we denote the resulting control by FWR(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 $C^a_i(\cdot) := C_i(\cdot/a_i)$, where, as before, $a_i := \lambda_i/\lambda$. Finally, let π^*_3 =FWR(p^{**}, v^*).

Following [14], 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

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

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 non-anticipating controls in the next section.)

Definition 3.5 (**asymptotically efficient controls**) *A sequence of non-anticipating controls $\{\pi^{\lambda}\}$ is said to be asymptotically efficient if*

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

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

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. The restriction to asymptotically efficient controls is imposed in order 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 previous papers [19] and [14]. 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 4.2. 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 (CRP) which trivially holds in the setting of [19]). The CRP 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; that remains an open problem. 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 [19], 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 FQR and FWR follows from the corresponding state-space collapse results in Theorem 4.3 and 4.4 below. Consequently, $\pi^*_3 \in \Pi^e$.

Here is our asymptotic optimality result in the delay-cost context:

Theorem 3.3 (**asymptotic optimality of** FWR **for the delay-cost criterion**) *If* $\hat{Q}^{\lambda}_{\Sigma}(0) = 0$ *for all* λ *and*

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

then

$$\limsup_{\lambda\to\infty} J^{\lambda}_{2}(\pi^*_3, T) \leq_{st} \liminf_{\lambda\to\infty} J^{\lambda}_{2}(\pi^{\lambda}, T) \tag{21}$$

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 (**the priority-type rule**) An analog of condition (14) holds to characterize when FWR reduces to the priority-type rule, using the new scaled cost functions $C^a_i(x) := C_i(x/a_i)$ instead of the original cost functions C_i. Note that the conditions to reduce to the priority-type rule are *not* equivalent for FQR and FWR, 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 FWR and FQR, because the slopes change from c_i to c_i/a_i. ■

3.4 The D-$Gc\mu$ Rule

Definition 3.6 (**the many-server** D-$Gc\mu$)

For the SBR model with pool-dependent service rates, the D-$Gc\mu$ rule is defined from the many-server $Gc\mu$ rule by replacing the scheduling component with:

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

$$i^* := i^*(t) \in \underset{i \in I(j), \hat{Q}_i^\lambda(t) > 0}{\operatorname{argmax}} C_i'(\hat{W}_{h,i}^\lambda(t));$$

If there are no such customers, the agent will remain idle.

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

Remark 3.6 (**comparing** FWR **and** D-$Gc\mu$) The D-$Gc\mu$ rule has a clear advantage over any form of FWR, because FWR requires knowledge of the ratios $A_i^\lambda(t)/A^\lambda(t)$, while D-$Gc\mu$ does not. ■

Theorem 3.4 (**asymptotic optimality of the** D-$Gc\mu$ **for the 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 optimal policy π_3^* can be replaced with π_4^*; i.e., the D-$Gc\mu$ 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), $FQR(p, v^*)$ 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^* so as to optimally distribute this aggregate queue among the different classes to minimize the convex holding costs.

To show that FQR 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 [1] 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 FQR 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 FQR established in [9].

Before proceeding to the actual proofs, we make some definitions and 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 RCLL (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 [21]. 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 $\mathbb{R}$.

Since 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 $0 < s < T < \infty$, $\|Y^\lambda\|^*_{s,T} \Rightarrow 0$.

We now formally define the set of admissible policies Π_1. In doing so, we follow Definition 2 in [3]. Toward that end, let $Z^\lambda_{i,j}(t)$ is 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

$S_{i,j}\left(\mu_j \int_0^t Z^\lambda_{i,j}(s)ds\right)$, 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 **upon 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 queue, some of which will be served but 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 §B of [9] 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

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

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. $\tau_i(t) = \inf\{u \ge t : A^\lambda_i(u) - A^\lambda_i(u-) > 0\}$. Let $T^\lambda_{i,j}(t) := \mu_j \int_0^t Z^\lambda_{i,j}(s)ds$. Then let

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

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 (**non-anticipating policies**) *A policy π is a mapping $\Omega \mapsto D^{2I\times 2J}_\uparrow$ taking ω into*

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

such that $(Z^\lambda_{i,j}(t), Q^\lambda_i(t); i \in \mathcal{I}, j \in \mathcal{J}, t \ge 0)$ satisfies equations (22)-(23)*. A policy π is said to be non-anticipating if*

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

(ii) for each $i \in \mathcal{I}$*,* $j \in \mathcal{J}$ *and* $t \geq 0$*, the process* $S_{i,j}(T^{\lambda}_{i,j}(t) + \cdot) - S_{i,j}(T^{\lambda}_{i,j}(t) + \cdot)$ *is equal in law to* $S_{i,j}(\cdot)$.

We let Π_1 *be the set of non-anticipative policies.*

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

4.1 The $\bigwedge$ Model and the FSF Policy

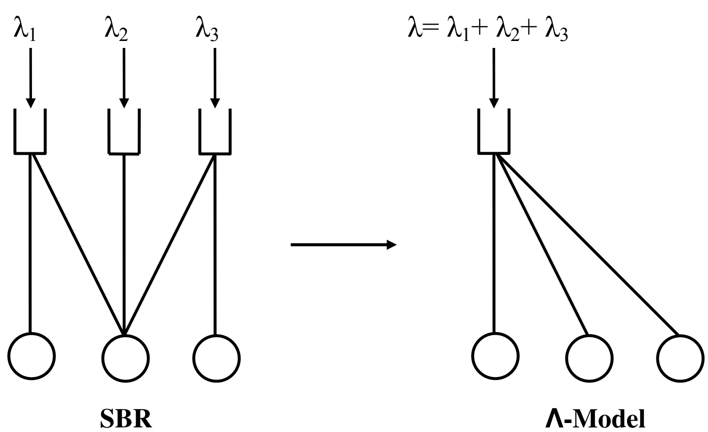

Figure 1: An SBR model and its corresponding $\bigwedge$ model

The asymptotic optimality established in Theorems 2.2 and 3.2 of our previous paper [8] 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.

In order to present our stochastic ordering results, we let $SBR(\mathcal{I}, \overrightarrow{\lambda}, \mathcal{J}, E, N, \mu)$ denote an SBR system with a set $\mathcal{I}$ of customer classes, arrival-rate vector $\overrightarrow{\lambda} = (\lambda_1, \ldots, \lambda_I)$, a set $\mathcal{J}$ of agent pools, a routing graph E, staffing vector N and pool-dependent service rates given by the vector $\mu = (\mu_1, \ldots, \mu_j)$. The corresponding $\bigwedge$ model is denoted by $\bigwedge(\lambda, \mathcal{J}, N, \mu)$ and stands for an inverted-V model with arrival rate $\lambda = \sum_{i\in\mathcal{I}} \lambda_i$, a set $\mathcal{J}$ 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 non-anticipating policies. Since 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 1.

We will abbreviate and use SBR and $\bigwedge$ when the data $(\mathcal{I}, \lambda, \mathcal{J}, E, N, \mu)$ is 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 $Z^{\pi_1}_{j,SBR}(t)$ 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 $Q^{\pi_1}_{\Sigma,SBR}(t)$ 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 (**the $\bigwedge$ model as a lower bound**) *Fix the data $(\mathcal{I}, \mathcal{J}, E, N, \overrightarrow{\lambda}, \mu)$. Assume that*

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

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

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

Consequently,

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

The proof of Lemma 4.1 follows a very simple coupling argument and it is omitted. We now define the faster-server-first (FSF) policy proposed in Armony [1].

Definition 4.2 (**the FSF control for the $\bigwedge$ model**) *The FSF control is defined as follows:*

- **Upon 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*

$$j \in \underset{k \in \mathcal{J}: I_k^\lambda(t) > 0}{\operatorname{argmax}} \ \mu_k.$$

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

- **Upon 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 $\hat{X}^\lambda_{\Sigma,\bigwedge}(t) := (X^\lambda_{\Sigma,\bigwedge}(t) - N^\lambda_\Sigma)/\sqrt{\lambda}$. Finally, let $\pi^*_{\bigwedge} := FSF$. The set of admissible policies for the $\bigwedge$ model is the set of non-anticipating policies, as defined in Definition 4.1. That is appropriate since the $\bigwedge$ model is a special case of an SBR model.

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

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

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

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

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 the SBR system**) *For any family of admissible policies $\{\pi^\lambda, \lambda > 0\}$ for the SBR system and continuous nondecreasing function g,*

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

for each $T \geq 0$.

Proof of Theorem 4.1: Theorem 4.1 follows from Armony [1], but the result is not stated as such in her paper. Thus we sketch how the different results in [1] paper can be combined. We assume familiarity with [1].

First, fix λ. Lemma 3.1 of [1] 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. Since 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 [1], 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, while [1] assumes Poisson arrivals, the results that we used above are easily extended to renewal arrivals. Indeed, Lemma 3.1 of [1] 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 [1] 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 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 [22] for an illustrative example. ■

We now show that FQR with appropriately chosen parameters achieves asymptotically the same aggregate number of customers as the optimally controlled $\bigwedge$ model. Recall that v^* is the non-state-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 the $\bigwedge$ model**) *Fix any ratio function p and let $\tilde{\pi} := FQR(p, v^*)$. If $\hat{X}^{\lambda}_{\Sigma}(0) = \hat{X}^{\lambda}_{\Sigma,\bigwedge}(0)$ for all λ and*

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

then both

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

and

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

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

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

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

Proof: Equation (24) follows from Theorem E.1 in [9] 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. However, some care is needed, because Theorem E.1 in [9] is stated only for non-state-dependent ratio functions. However, the restriction there was made only to ensure existence and uniqueness of the solution for the resulting SDE. It is easy to see that equation (A122) in [9] remains valid provided that $\theta_i = 0$ for all $i \in \mathcal{I}$. Equation (25) follows from Proposition 4.2 in Armony [1] with the appropriate replacement of scaling (by $\sqrt{N^{\lambda}}$ rather then $\sqrt{\lambda}$), and after replacing the Poisson arrival process with the renewal process. ∎

The implication of Proposition 4.1 is that FQR asymptotically achieves the lower bound given by the $\bigwedge$ model provided that the ratio function v^* is used for the routing component. Consequently, FQR 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 [14].

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

4.2 Asymptotic optimality for the Holding-Cost Formulation

The following is a direct consequence of Theorem B.1 in [9].

Theorem 4.3 (**state-space collapse under FQR with pool-dependent rates**)

If $(\hat{X}^\lambda(0), \hat{Z}^\lambda(0)) \Rightarrow (\hat{X}(0), \hat{Z}^\lambda(0))$ *in* $\mathbb{R}^{I+I\cdot J}$*, then we have state-space collapse:*

$$\hat{Q}_i^\lambda(t) - \hat{Q}_\Sigma^\lambda(t) p_i\left(\hat{Q}_\Sigma^\lambda(t)\right) \Rightarrow 0 \quad in \quad D_- \quad as \quad \lambda \to \infty, \quad i \in \mathcal{I}, \tag{27}$$

and

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

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

$$\hat{Q}_i^\lambda(0) - \hat{Q}_\Sigma^\lambda(0) p_i\left(\hat{Q}_\Sigma^\lambda(0)\right) \Rightarrow 0,\ i \in \mathcal{I},\ and\ \hat{I}_j^\lambda(0) - \hat{I}_\Sigma^\lambda(0) v_j\left(\hat{I}_\Sigma^\lambda(0)\right) \Rightarrow 0,\ j \in \mathcal{J}. \tag{29}$$

Remark 4.1 (**relaxing the Poisson-arrival assumption in [9]**) Theorem B.1 in [9] was proved under the assumption of Poisson arrivals, but the proof is easily changed to allow for renewal arrival processes. For that purpose, Lemma D.2 in [9] needs to be slightly changed to take care of renewal arrivals. That can be done along the lines of Proposition 6.1 in [5]. We omit the details here. ∎

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

$$\sum_{i=1}^{I} C_i\left(\hat{Q}_i^{\lambda,\pi^\lambda}(t)\right) \geq \sum_{i=1}^{I} C_i\left(q_i^*(\hat{Q}_\Sigma^{\lambda,\pi^\lambda}(t))\right), \tag{30}$$

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

$$\inf_{\pi^\lambda \in \Pi_1} J_1^\lambda(\pi^\lambda) \geq \inf_{\pi^\lambda \in \Pi_1} \int_0^T \sum_{i\in\mathcal{I}} C_i\left(q_i^*(\hat{Q}_\Sigma^{\lambda,\pi^\lambda}(t))\right) dt \geq_{st} \int_0^T \sum_{i\in\mathcal{I}} C_i\left(q_i^*(\hat{Q}_{\Sigma,\bigwedge}^{\lambda,\tilde{\pi}^\lambda}(t))\right) dt. \tag{31}$$

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

$$\liminf_{\lambda\to\infty} \int_0^T \sum_{i\in\mathcal{I}} C_i\left(q_i^*(\hat{Q}_{\Sigma,\bigwedge}^{\lambda,\tilde{\pi}^\lambda}(t))\right) dt \geq_{st} \limsup_{\lambda\to\infty} \int_0^T \sum_{i\in\mathcal{I}} C_i\left(q_i^*(\hat{Q}_{\Sigma,\bigwedge}^{\lambda,\pi^*}(t))\right) dt. \tag{32}$$

By equation (25), the continuity of the integral (see Theorem 11.5.1 in [21]) and the Continuous Mapping Theorem,

$$\liminf_{\lambda\to\infty}\int_0^T \sum_{i\in\mathcal{I}} C_i\left(q_i^*(\hat{Q}_{\wedge}^{\lambda,\tilde{\pi}^\lambda}(t))\right)dt \geq_{st} \int_0^T \sum_{i\in\mathcal{I}} C_i\left(q_i^*([\hat{X}_\Sigma(t)]^+)\right)dt, \tag{33}$$

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

$$\limsup_{\lambda\to\infty}\inf_{\pi^\lambda\in\Pi_1} J_1^\lambda(\pi^\lambda) \geq_{st} \int_0^T \sum_{i\in\mathcal{I}} C_i\left(q_i([\hat{X}_\Sigma(t)]^+)\right)dt. \tag{34}$$

Having established an asymptotic lower bound, it only remains to show that this lower bound is asymptotically achieved by π_1^* =FQR(p^*, v^*). Indeed, by equation (24) in Proposition 4.1, Theorem 4.3 and the Continuous Mapping Theorem,

$$\int_0^T \sum_{i\in\mathcal{I}} C_i\left(\hat{Q}_i^{\lambda,\pi_1^*}(t)\right)dt \Rightarrow \int_0^T \sum_{i\in\mathcal{I}} C_i\left(q_i([\hat{X}_\Sigma(t)]^+)\right)dt, \text{ as } \lambda\to\infty. \tag{35}$$

Consequently, fixing any family ${\pi'}^\lambda$ of admissible controls, we have that

$$\limsup_{\lambda\to\infty} J_1^\lambda({\pi^*}^\lambda) \leq_{st} \limsup_{\lambda\to\infty} J_1^\lambda({\pi'}^\lambda), \tag{36}$$

which establishes the asymptotic optimality of π_1^*. ■

4.3 Asymptotic Optimality for the Delay-Cost Formulation

The proof of Theorem 3.3 mostly follows the proof of Theorem 3.1 above. However, it requires some side results and these are proved in this section. The section is concluded with the proof of Theorem 3.3.

We start by proving that asymptotic efficiency (see Definition 3.5) implies the stochastic boundedness and C-tightness of $\hat{Q}_\Sigma^\lambda(t)$. A family $\{x^\lambda, \lambda > 0\}$ of processes in $D^d[0,T]$ is said to be *stochastically bounded* if

$$\lim_{k\to\infty}\limsup_{\lambda\to\infty} P\{\|x^r\|_T > k\} = 0.$$

It is said to be *tight* if every subsequence with $\lambda_k\to\infty$ contains a convergent subsequence and *C-tight* if the limit of each such subsequence is continuous. We refer the reader to §5 of [22] for a detailed discussion of these concepts.

Lemma 4.2 (**stochastic boundedness and C-tightness**) *For any family* $\{\pi^\lambda, \lambda > 0\} \in \Pi^e$*, the corresponding family* $\{\hat{Q}^\lambda_\Sigma(t), t \geq 0\}$ *is stochastically bounded and C-tight.*

Proof: By equation (A47) of [9], we can write

$$\hat{X}^\lambda_\Sigma(t) = \hat{X}^\lambda_\Sigma(0) - \beta t + \sum_{j\in\mathcal{J}} \mu_j \int_0^t \hat{I}^\lambda_j(s)ds + \hat{M}^\lambda_\Sigma(t) + o(1) \quad \text{as} \quad \lambda \to \infty, \tag{37}$$

where $\hat{M}^\lambda_\Sigma(t)$ is the square integrable Martingale defined prior to (A47) in [9] for each λ. Consequently,

$$|\hat{X}^\lambda_\Sigma(t)| \leq |\beta|t + \mu_1 \int_0^t \hat{I}^\lambda_\Sigma(s)ds + |\hat{M}^\lambda_\Sigma(t)|.$$

By the asymptotic efficiency assumption, $\|\hat{I}^\lambda_\Sigma - [\hat{X}^\lambda_\Sigma]^-\|_T \Rightarrow 0$ as $\lambda \to \infty$, so that

$$|\hat{X}^\lambda_\Sigma(t)| \leq |\beta|t + \mu_1 \int_0^t |\hat{X}^\lambda_\Sigma(s)|ds + |\hat{M}^\lambda_\Sigma(t)| + o(1) \quad \text{as} \quad \lambda \to \infty.$$

Since $\hat{M}^\lambda_\Sigma(t)$ is C-tight - see the proof of Lemma D.1 in [9]) - it is also stochastically bounded. Hence we can apply Gronwall's inequality to deduce that the family $\hat{X}^\lambda_\Sigma(t)$ is stochastically bounded.

To establish C-tightness, we use (37) to write

$$\hat{X}^\lambda_\Sigma(t) - \hat{X}^\lambda_\Sigma(s) = -\beta(t-s) + \sum_{j\in\mathcal{J}} \mu_j \int_s^t \hat{I}^\lambda_j(h)dh + \hat{M}^\lambda_\Sigma(t) - \hat{M}^\lambda_\Sigma(s) + o(1) \quad \text{as} \quad \lambda \to \infty$$

and, consequently,

$$|\hat{X}^\lambda_\Sigma(t) - \hat{X}^\lambda_\Sigma(s)| \leq |\beta|(t-s) + \mu_1(t-s)\|\hat{X}^\lambda_\Sigma\|_T + \mu_1 \int_0^{t-s} |\hat{X}^\lambda_\Sigma(s+h) - \hat{X}^\lambda_\Sigma(s)|dh + |\hat{M}^\lambda_\Sigma(t) - \hat{M}^\lambda_\Sigma(s)| + o(1).$$

C-tightness now follows from the stochastic boundedness of $\hat{X}^\lambda_\Sigma(t)$ and the C-tightness of $\hat{M}^\lambda_\Sigma(t)$ through an application of Gronwall's inequality, just as in the proof of Lemma D.1 in [9].

We have thus proved that the family $\hat{X}^\lambda_\Sigma(t)$ is stochastically bounded and C-tight under the asymptotic efficiency condition. To complete the proof, we apply the assumed asymptotically efficient to deduce that

$$\hat{Q}^\lambda_\Sigma(t) - \hat{Q}^\lambda_\Sigma(s) = [\hat{X}^\lambda_\Sigma(t)]^+ - [\hat{X}^\lambda_\Sigma(s)]^+ + o(1) \quad \text{as} \quad \lambda \to \infty.$$

Consequently, the C-tightness and stochastic boundedness of $\hat{X}^{\lambda}_{\Sigma}(t)$ imply these properties for $\hat{Q}^{\lambda}_{\Sigma}(t)$. ■

We turn now to the statement of the state-space collapse result for FWR.

Theorem 4.4 (**state-space collapse under FWR with pool-dependent rates**)

If $(\hat{X}^{\lambda}(0), \hat{Z}^{\lambda}(0)) \Rightarrow (\hat{X}(0), \hat{Z}^{\lambda}(0))$ *in* $\mathbb{R}^{I+I\cdot J}$ *and* $\hat{Q}^{\lambda}_{\Sigma}(0) = 0$ *for all* λ*, then we have state-space collapse:*

$$\hat{Q}^{\lambda}_{i}(t) - \hat{Q}^{\lambda}_{\Sigma}(t) p_i \left(\hat{Q}^{\lambda}_{\Sigma}(t)\right) \Rightarrow 0 \quad in \quad D \quad as \quad \lambda \to \infty, \quad i \in \mathcal{I}, \tag{38}$$

and

$$\hat{I}^{\lambda}_{j}(t) - \hat{I}^{\lambda}_{\Sigma}(t) v_j \left(\hat{I}^{\lambda}_{\Sigma}(t)\right) \Rightarrow 0 \quad in \quad D \quad as \quad \lambda \to \infty, \quad j \in \mathcal{J} . \tag{39}$$

In addition, we have

$$a_i \hat{W}^{\lambda}_{h,i}(t) - \hat{Q}^{\lambda}_{i}(t) \Rightarrow 0 \quad in \quad D \quad as \quad \lambda \to \infty, \quad i \in \mathcal{I}. \tag{40}$$

Proof: We outline the changes that should be made to the proof of Theorem 4.3 to accommodate the special features of FWR. Theorem 4.3 itself is proved in §D.2 of [9], as a special case of Theorem B.1 there. We will be making frequent reference to that section.

First, the definition of the stopping time σ^{λ} in equation (A49) of [9] should be changed to

$$\sigma^{\lambda} := \inf\{t \geq 0 | \hat{B}^{\lambda}(t) \geq 2\hat{B}^{\lambda}(0) \vee 1\} \wedge \tilde{\sigma}^{\lambda},$$

where

$$\tilde{\sigma}^{\lambda} = \inf\{t \geq 0 : \max_{i \in \mathcal{I}} |\hat{Q}^{\lambda}_{i}(t) - a_i \hat{W}^{\lambda}_{h,i}(t)| \geq \epsilon^{\lambda}\},$$

with $\epsilon^{\lambda} \to 0$ as $\lambda \to \infty$ sufficiently slow (to be precisely defined towards the end of the proof).

Then, the hydrodynamic model equations, (A71)-(A80) of [9], are augmented by the additional equation

$$\tilde{W}_{h,i}(t) = \tilde{Q}_i(t)/a_i.$$

The proof of state-space collapse now follows identically the proof of Theorem 4.3 with the exception of Lemma D.5 in [9], which should be slightly changed to take care of the new definition of the stopping time

σ^λ. Specifically, we add the following argument to the proof of Lemma D.5: First, we claim that, since state-space collapse holds on $[0, T^\lambda]$ and since $\hat{Q}^\lambda_\Sigma(0) = 0$ by assumption, the C-tightness of the sequence $\hat{Q}^\lambda(\cdot \wedge T^\lambda) = (\hat{Q}^\lambda_1(\cdot \wedge T^\lambda), \ldots, \hat{Q}^\lambda_I(\cdot \wedge T^\lambda))$ follows from that of $\hat{Q}^\lambda_\Sigma(t)$, which was proved in Lemma 4.2. Indeed, by state-space collapse $\hat{Q}^\lambda_i(t \wedge T^\lambda) \approx \hat{Q}^\lambda_\Sigma(t \wedge T^\lambda) p_i\left(\hat{Q}^\lambda_\Sigma(t \wedge T^\lambda)\right)$ with $p_i(\cdot)$ being a locally Hölder continuous function. Consequently, the tightness of $\hat{Q}^\lambda_\Sigma(t)$ implies that of $\hat{Q}^\lambda_i(t)$.

Now let $W^\lambda_i(t)$ be the virtual waiting time of class-i at time t in the λ^{th} system and $\hat{W}^\lambda_i(t) = \sqrt{\lambda} W^\lambda_i(t)$. Note that $\hat{W}^\lambda_i(t)$ is not necessarily equal to $\hat{W}^\lambda_{h,i}(t)$ as the latter refers to the cumulative waiting time of the customer at the head of the line. Having the C-tightness of $\hat{Q}^\lambda_i(\cdot \wedge T^\lambda)$, we can apply the corollary in Puhalskii [16] to establish the joint convergence

$$\left(\frac{\hat{Q}^\lambda_i(t \wedge T^\lambda)}{a_i}, \hat{W}^\lambda_i(t \wedge T^\lambda); i \in \mathcal{I}\right) \Rightarrow \left(\hat{Q}_i(t), \hat{Q}_i(t)/a_i; i \in \mathcal{I}\right) \text{ in } D^{2I} \text{ as } \lambda \to \infty,$$

where $\hat{W}_i(t) = \hat{Q}_i(t)/a_i$; see e.g. Lemma A.2 of Puhalskii and Reiman [17]. The convergence of $\hat{W}^\lambda_i(t)$ implies that the familye $\{\hat{W}^\lambda_i(t)\}$ is also stochastically bounded.

Since, by definition,

$$\hat{W}^\lambda_{h,i}(t) = \hat{W}_i(t - W^\lambda_{h,i}(t)), \tag{41}$$

we have that $\hat{W}^\lambda_{h,i}(t)$ is itself stochastically bounded and the unscaled process $W^\lambda_{h,i}(t)$ satisfies

$$W^\lambda_{h,i}(t \wedge T^\lambda) \Rightarrow 0 \text{ in } D \text{ as } \lambda \to \infty.$$

We can then apply the Random-Time-Change Theorem to equation (41) to have the joint convergence

$$\left(\frac{\hat{Q}^\lambda_i(t \wedge T^\lambda)}{a_i}, \hat{W}^\lambda_{h,i}(t \wedge T^\lambda); i \in \mathcal{I}\right) \Rightarrow \left(\frac{\hat{Q}_i(t)}{a_i}, \frac{\hat{Q}_i(t)}{a_i}\right) \text{ in } D^{2I} \text{ as } \lambda \to \infty.$$

By Theorem 11.4.8 in [21] and the continuity of the limit, we then have

$$\max_{i \in \mathcal{I}} \|\hat{Q}^\lambda_i - a_i \hat{W}^\lambda_{h,i}\|_{T^\lambda} \Rightarrow 0 \text{ in } \mathbb{R} \text{ as } \lambda \to \infty,$$

so that

$$P\left\{\max_{i \in \mathcal{I}} \|\hat{Q}^\lambda_i - a_i \hat{W}^\lambda_{h,i}\|_{T^\lambda} > \epsilon^\lambda\right\} \to 0 \text{ as } \lambda \to \infty$$

as long as $\{\epsilon^\lambda\}$ is such that $\epsilon^\lambda \to 0$ sufficiently slowly. Consequently, we have that $\tilde{\sigma}^\lambda \to \infty$. The rest of the proof follows Lemma D.5 in [9], allowing us to conclude that $\sigma^\lambda \to \infty$. Hence the adaptation of the proof of Theorem 4.3 for FWR is complete. ■

It remains to relate the proof of Theorem 3.3 to that of Theorem 3.1. That is accomplished in the following proposition, which states that both cost criteria share essentially the same lower bound. For the following, we say that a family $\{b^\lambda\}$ is $o_P^\lambda(1)$ if $b^\lambda \Rightarrow 0$ as $\lambda \to \infty$.

Proposition 4.2 *If $\{\pi^\lambda\} \in \Pi^e$ is a sequence of admissible policies, then*

$$J_2^\lambda(\pi^\lambda, T) \geq \int_0^T C_i^a \left(q_i^*(\hat{Q}_\Sigma^{\lambda,\pi^\lambda}(t)) \right) dt + o_P^\lambda(1) \quad \textit{as} \quad \lambda \to \infty,$$

where $C_i^a(\cdot) := C_i(\cdot/a_i)$ for all $i \in \mathcal{I}$.

Proof: The proof builds on the proof of Proposition 6 in Van Meighem [19]. Since there are some differences, we give a detailed proof. Since the family $\hat{Q}_\Sigma^\lambda(t)$ is C-tight by Lemma 4.2, we can choose a convergent subsequence $\{\hat{Q}_\Sigma^{\lambda^k}(t),\ k \in \mathbb{N}\}$ with $\lambda_k \to \infty$ whose limit is continuous. We will show that the result of the Proposition holds for every convergent subsequence and consequently for the whole sequence. For simplicity of presentation, we assume that $\{\lambda^k\}$ is the whole family; the reader should remember that the proof applies to the subsequence.

Denote the limit by $\hat{Q}_\Sigma(t)$. Together with the Functional Strong Law of Large Numbers (FSLLN), we have the joint convergence

$$\left(\frac{A_1^\lambda(t)}{\lambda}, \ldots, \frac{A_I^\lambda(t)}{\lambda}, \hat{Q}_\Sigma^\lambda(t) \right) \Rightarrow (a_1 t, \ldots, a_I t, \hat{Q}_\Sigma(t)) \text{ in } D^{I+1} \text{ as } \lambda \to \infty\ . \tag{42}$$

Since the space D with the J_1 topology is separable (see §11.5 of [21]), we can use the Skorohod representation Theorem, Theorem 3.2.2 in [21], to construct all the processes on an alternative probability space $(\tilde{\Omega}, \tilde{\mathbb{F}}, \tilde{P})$ such that the convergence holds for almost every $\omega \in \tilde{\Omega}$. We henceforth fix such a realization ω. We consider the sequence $\{t_k,\ k \geq 0\}$ of stopping times defined recursively as follows:

$$t_{k+1} = \min\{T, \inf\{t_k < t \leq T : |\hat{Q}_\Sigma(t) - \hat{Q}_\Sigma(t_k)| \geq \epsilon\}\},$$

where $t_0 = 0$. Fix $\omega \in \Omega$. Note that since $\hat{Q}_\Sigma(t)$ is continuous on the compact interval $[0, T]$ we have that there exists $\delta > 0$ such that

$$\inf_i (t_{k+1} - t_k) > \delta. \tag{43}$$

By Jensen's inequality,

$$\begin{aligned} J_2^\lambda(\pi^\lambda, T) &= \textstyle\sum_{i\in\mathcal{I}} \frac{1}{A_i^\lambda(t)} \sum_k \int_{t_k}^{t_{k+1}} C_i(\hat{W}_i^\lambda(s)) dA_i^\lambda(s) \\ &\geq \textstyle\sum_{i\in\mathcal{I}} \sum_k \left[\frac{1}{A_i^\lambda(t)} [A_i^\lambda(t_{k+1}) - A_i^\lambda(t_k)] \times C_i \left([A_i^\lambda(t_{k+1}) - A_i^\lambda(t_k)]^{-1} \int_{t_k}^{t_{k+1}} \hat{W}_i^\lambda(s) dA_i^\lambda(s) \right) \right] \end{aligned}$$

Since (42) holds almost surely on $\tilde{\Omega}$ and by our choice of the realization ω, we have that

$$[A_i^\lambda(t_{k+1}) - A_i^\lambda(t_k)]/\lambda = a_i(t_{k+1} - t_k) + o^\lambda(1) \quad \text{as} \quad \lambda \to \infty \,,$$

where the approximation is uniform on $[0, T]$. Thus,

$$J_2^\lambda(\pi^\lambda, T) \geq \sum_{i\in\mathcal{I}} \sum_k \left[[(t_{k+1} - t_k) + o^\lambda(1)] \times C_i \left([\lambda(t_{k+1} - t_k + o^\lambda(1))]^{-1} \int_{t_k}^{t_{k+1}} \hat{W}_i^\lambda(s) dA_i^\lambda(s) \right) \right]. \tag{44}$$

We now use Proposition 4 of [19], which - with the appropriate modification to our setting - states the following: Fix $0 \leq a < b \leq T$, then

$$\frac{1}{\lambda_i(b-a)} \int_a^b \hat{W}_i^\lambda(s) dA_i^\lambda(s) - \frac{1}{b-a} \int_a^b \hat{Q}_i^\lambda(s) ds \to 0, \text{ as } \lambda \to \infty,$$

with the convergence holding almost surely. The proof of this result is identical to the proof in [19], so it is omitted. Using (43) and recalling that $a_i := \lambda_i/\lambda$, we have

$$[\lambda(t_{k+1} - t_k + o^\lambda(1))]^{-1} \int_{t_k}^{t_{k+1}} \hat{W}_i^\lambda(s) dA_i^\lambda(s) - \frac{1}{a_i(t_{k+1} - t_k)} \int_{t_k}^{t_{k+1}} \hat{Q}_i^\lambda(s) d(s) \to 0 \text{ as } \lambda \to \infty.$$

Plugging this back into (44), we then have

$$J_2^\lambda(\pi^\lambda, T) \geq \sum_{i\in\mathcal{I}} \sum_k \left[[(t_{k+1} - t_k) + o^\lambda(1)] \times C_i \left(\frac{1}{a_i(t_{k+1} - t_k)} \int_{t_k}^{t_{k+1}} \hat{Q}_i^\lambda(s) d(s) + o^\lambda(1) \right) \right]. \tag{45}$$

Since $\hat{Q}_i^\lambda(t) \leq \hat{Q}_\Sigma^\lambda(t)$ and $\hat{Q}_\Sigma^\lambda(t)$ is bounded by its continuity on $[0, T]$, we have that $\hat{Q}_i^\lambda(t)$ is bounded. The continuity of C_i, then implies that (45) can be written as

$$J_2^\lambda(\pi^\lambda, T) \geq \sum_{i\in\mathcal{I}} \sum_k \left[(t_{k+1} - t_k) \times C_i \left(\frac{1}{a_i} \int_{t_k}^{t_{k+1}} \hat{Q}_i^\lambda(s) d(s) \right) \right] + o^\lambda(1) \quad \text{as} \quad \lambda \to \infty.$$

The C-tightness of $\hat{Q}_\Sigma^\lambda(t)$ now implies that

$$C_i \left(\frac{1}{a_i(t_{k+1} - t_k)} \int_{t_k}^{t_{k+1}} \hat{Q}_i^\lambda(s) d(s) \right) \geq C_i^a \left(q_i^*(\hat{Q}_\Sigma^\lambda(t_k)) \right) + o^\lambda(1) + O(\epsilon), \tag{46}$$

where $C_i^a := C_i(\cdot/a_i)$ and $\{q_i^*(x),\ i \in \mathcal{I}\}$ is the optimal solution for (11) with the cost functions C_i^a replacing C_i. Finally, $O(\epsilon) \to 0$ as $\epsilon \to 0$. To establish (46), note that

$$C_i \left(\frac{1}{a_i(t_{k+1} - t_k)} \int_{t_k}^{t_{k+1}} \hat{Q}_i^\lambda(s) d(s) \right) \geq C_i \left(\frac{1}{a_i(t_{k+1} - t_k)} \int_{t_k}^{t_{k+1}} \tilde{q}_i^\lambda(s) \right),$$

where $\tilde{q}_i^\lambda(s)$ is a solution to

$$\begin{array}{ll} \text{minimize} & \sum_{i\in\mathcal{I}} C_i \left(\frac{1}{a_i(t_{k+1}-t_k)} \int_{t_k}^{t_{k+1}} q_i^\lambda(s) ds \right), \\ \text{s.t.} & \sum_{i\in\mathcal{I}} q_i^\lambda(s) = \hat{Q}_\Sigma^\lambda(s),\ s \in [t_k, t_{k+1}], \\ & q_i^\lambda(s) \geq 0,\ i \in \mathcal{I},\ s \in [t_k, t_{k+1}]. \end{array}$$

However, the C-tightness of $\hat{Q}_\Sigma^\lambda(t)$ and the definition of the stopping times t_k implies that, for all λ large enough, $|\hat{Q}_\Sigma^\lambda(s) - \hat{Q}_\Sigma^\lambda(t_k)| \leq 2\epsilon$ for all $s \in [t_k, t_{k+1}]$ so that (46) follows from the continuity of C_i.

Consequently,

$$J_2^\lambda(\pi^\lambda, T) \geq \sum_{i\in\mathcal{I}} \sum_k \left[(t_{k+1} - t_k) \times C_i^a \left(q_i^*(\hat{Q}_\Sigma^\lambda(t_k)) \right) \right] + o^\lambda(1) + O(\epsilon)$$

Since ϵ was arbitrary we may invoke the definition of the Rieman integral to obtain the result for almost every $\omega \in \tilde{\Omega}$. Translating this back into the original probability space yields the claimed statement. ■

Proof of Theorem 3.4: Using Proposition 4.2 and the state-space collapse result in Theorem 4.4, the proof now follows exactly as the proof of Theorem 3.3 with the exception of the $o_P^\lambda(1)$ term - whose treatment is trivial - and the replacement of Π_1 by Π^e. ■

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 FQR and FWR 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 [14] to the Halfin-Whitt regime. However, this extension is only partial since we had to restrict attention to pool-dependent service rates. It remains to be determine if simple elegant controls are asymptotically optimal with more general service rates. Limitations on what can be achieved follow from the previous work of Harrison and Zeevi [12] and Atar [3]. Harrison and Zeevi provide an example where a complicated bang-bang control is asymptotically optimal for the V model with linear holding costs. Atar [3] relates the asymptotic optimality to the optimal solution of a related Brownian control problem, which in most cases results in a complex solution. While 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.

References

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

[2] Armony M., I. Gurvich, A. Mandelbaum. 2006. Service-level differentiation in call Centers with fully flexible servers. *Management Science*, forthcoming.

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

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

[5] Dai J.G., T. Tezcan. 2005. State space collapse in many server diffusion limits of parallel server systems. Working Paper, Georgia Institute of Technology, Atlanta, GA.

[6] Dai J.G, T. Tezcan. 2006. Dynamic control of N-systems with many servers: asymptotic optimality of a static priority policy in heavy traffic. Working Paper. Georgia Institute of Technology, Atlanta, GA.

[7] Dai J.G, T. Tezcan. 2007. Optimal control of parallel server systems with many servers in heavy traffic. Working Paper. Georgia Institute of Technology, Atlanta, GA.

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

[9] Gurvich I., W. Whitt. 2007. Service-level differentiation in many-server service systems: a solution based on fixed-queue-ratio routing, Technical Appendix. Working Paper. Columbia University, New York.

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

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

[12] 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.

[13] Ibaraki, T., N. Katoh. 1988. *Resource Allocation Problems: Algorithmic Approaches*, no. 4 in Foundations of Computing series, The MIT Press, Cambridge, MA.

[14] 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.

[15] Patriksson, M. 2006. A survey on the continuous nonlinear resource allocation problem. Dept. Mathematics, Chalmers University of Technology, Gothenburg, Sweden. Available at: http://www.cs.chalmers.se/$\sim$mipat/LATEX/survey$_0$610.pdf

[16] Puhalskii A. 1994. On the invariance principle for the first passage time, *Math. Oper. Res.* **19**(4) 946 - 954.

[17] Puhalskii A., M. Reiman. 2000. The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Advances in Applied Probablity* **32**(2) 564-595.

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

[19] Van Mieghem J.A. 1995. Dynamic scheduling with convex delay costs: the generalized $c\mu$ rule. *Ann. Appl. Probab.* **5**(3) 809-833.

[20] Van Mieghem J.A. 2003. Due date scheduling: asymptotic optimality of generalized longest queue and generalized largest delay rules. *Oper. Res.*, **51**(1) 113-122.

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

[22] Whitt W. 2006. Martingale proofs of many-server heavy-traffic limits for Markovian queues. Working Paper. Columbia University, New York.

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