MATHEMATICS OF OPERATIONS RESEARCH
Vol. 34, No. 4, November 2009, pp. 795–812
ISSN 0364-765X | EISSN 1526-5471 | 09 | 3404 | 0795

DOI 10.1287/moor.1090.0404
© 2009 INFORMS

Simplified Control Problems for Multiclass Many-Server Queueing Systems

Rami Atar
Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel,
atar@ee.technion.ac.il

Avi Mandelbaum
Department of Industrial Engineering and Management, Technion–Israel Institute of Technology, Haifa 32000, Israel,
avim@tx.technion.ac.il

Gennady Shaikhet
Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213,
shaikhet@cmu.edu

We consider scheduling and routing control problems for queueing models with I customer classes and J server pools, each consisting of many statistically identical, exponential servers. Customers require a single service that can be performed by a server from one of the pools; the service rate is $\mu_{ij} \geq 0$, which depends on the customer's class i and the server's pool j, and customers can abandon the system while waiting to be served. In the heavy traffic regime of Halfin and Whitt, these problems are formally equivalent to I-dimensional diffusion control problems. We analyze the diffusion control problems is two special cases. First, when the service rates depend only on the pool ($\mu_{ij} = \mu_j$), the diffusion control problem is shown to be similar to (but distinct from) the diffusion control problem for a *single class* model, which greatly reduces the complexity of the problem. Second, when the service rates depend only on the class ($\mu_{ij} = \mu_i$), the diffusion control problem is shown to be equivalent to a diffusion control problem for a *single pool* model, a problem that has previously been studied. In the first case, we also establish a rigorous relation between the queueing control problem and the diffusion control problem, showing that a policy for the queueing model, based on an ordinary differential equation of Hamilton-Jacobi-Bellman type, is asymptotically optimal.

Key words: many-server queueing systems; Halfin-Whitt; QED regime; heavy traffic; scheduling and routing; skills-based routing; diffusion models; stochastic control; asymptotic optimality
MSC2000 subject classification: Primary: 60K25, 68M20, 90B22, 90B36, 60F05
OR/MS subject classification: Primary: queues: limit theorems; secondary: probability: diffusion, probability: stochastic model applications
History: Received May 5, 2008; revised December 30, 2008. Published online in *Articles in Advance* October 2, 2009.

1. Introduction. A queueing system has I customer classes and J server pools. Customer arrivals for each class follow a renewal process and each pool has many statistically identical, exponential servers. These servers work independently, offering service to different classes of customers at rates μ_{ij}, which depend on the class i and the pool j. Customers may abandon while waiting to be served, according to exponential clocks with class dependent rates (see Figure 1). We consider a *routing and scheduling control* (RSC) problem in which customers are to be routed to pools in a way that performance measures, such as average queue-lengths, are minimized. As often occurs, exact analysis of the RSC problem is unavailable, and an asymptotic approach, where the queueing model is considered in a heavy traffic regime, simplifies the problem considerably. Halfin and Whitt [13] proposed a parametrization under which the arrival rates and the number of servers increase to infinity while keeping a critically loaded system. Taking formal limits under this parametrization results in an I-dimensional diffusion model and an associated diffusion control (DC) problem. The queueing model described above and the heavy traffic regime of Halfin and Whitt [13] have recently enjoyed much attention, because they capture the operational characteristics of large telephone call centers (Gans et al. [10]). The regime is now referred to in the literature as the quality- and efficiency-driven (QED) regime (Atar et al. [4]). The DC problem and its rigorous relation to the RSC problem have been analyzed by several authors (for a review see Aksin et al. [1]).

In this paper we study the DC problem alluded to above, focusing on two special cases of the model, namely where the service rates are class dependent ($\mu_{ij} = \mu_i$ for all i and j) or pool dependent ($\mu_{ij} = \mu_j$ for all i and j). These cases have been considered by several authors. In Atar [2], these conditions were shown to be sufficient for a Hamilton-Jacobi-Bellman (HJB) partial differential equation in I variables, associated with the DC problem, to be uniquely solvable, as well as for a rigorous relation with the RSC problem to be valid. Tezcan and Dai [18] first noticed that the problem degenerates to a one-dimensional problem when the rates depend only on the pools, in a setting with $I = J = 2$. This result was extended in Dai and Tezcan [8] to general I and J under the assumption that class-1 customers have the most inexpensive holding cost and the highest

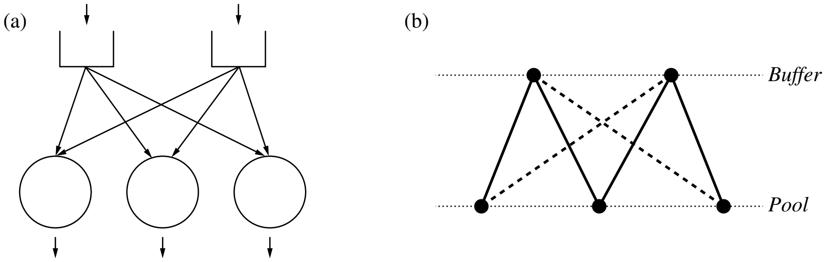

FIGURE 1. (a) A system with two buffers and three pools. (b) A corresponding graph with basic and nonbasic activities (solid and dashed lines, respectively).

abandonment rate. A static priority policy was constructed under which it was shown, using methods from Dai and Tezcan [7], that the diffusion scaled processes representing queue-lengths all converge to zero, except the one that corresponds to class-1 customers. Similarly, the processes representing idleness of servers from all pools converge to zero, except the one that corresponds to the pool that has the slowest service rate. This static policy was proved to be asymptotically optimal. Gurvich and Whitt [11] treated the pool dependent case without abandonments; using results from Gurvich and Whitt [12], asymptotically optimal policies were obtained for minimizing linear and convex holding costs. In the latter case, the routing policy is a many-server analogue of the generalized-$c\mu$ rule of Mandelbaum and Stolyar [16].

The asymptotic results of Gurvich and Whitt [11] and Dai and Tezcan [8] are related to our Examples 3.2 and 3.3 about the DC problem, where reduction from I dimensions to a single dimension is demonstrated. In both cases, the dimensionality reduction is obtained via *pathwise solutions*, in the sense that a specific control policy minimizes the diffusion limit of the weighted sum of queue-lengths, at all times, with probability 1. The existence of a pathwise minimum, however, stems from the restrictive assumptions made in Gurvich and Whitt [11] and Dai and Tezcan [8], and is not representative in general.

The first main contribution of our paper is to show that in the pool dependent case, the general DC problem is equivalent to a 1-dimensional DC problem (that need not have pathwise solutions in general). To underscore the importance of the 1-dimensional form of the problem, we show, in Example 3.4, that the DC problem is solved via a 1-dimensional HJB equation for which, unlike in the multidimensional case with possibly large values of I, efficient schemes of numerical solution are well known. Also, in the case of class dependent service rates, we show that the DC problem is equivalent to one with a single pool ($J = 1$), an I-dimensional DC problem that has been treated in Atar et al. [4].

Coming back to the pool dependent case, the dimensionality reduction provides some answers to issues raised in Atar [2] about optimality of *jointly work conserving* (JWC) policies. (While a plain *work conserving* policy allows a server to idle only when no customer that it can serve is waiting in the queue, a JWC policy allows only arrangements of customers in the system in which no server idles when a customer of *any* class is waiting in the queue.) Indeed, the reduction to one dimension is performed by first establishing a reduction to a DC problem in which the processes, that represent scaling limits of queue-length and idleness, satisfy JWC. In Atar [2], a JWC condition is assumed and shown to considerably simplify the model, but it is justified only by a heuristic discussion. The reduction of the DC problem to one in which a JWC condition holds justifies this heuristic in the case of pool dependent service rates. In the case of class dependent service rates, we also expect the question of JWC optimality to be answered affirmatively, but we were unable to prove this claim (Conjecture 4.1).

The second main contribution of this paper is the establishment of a rigorous relation between the queueing control problem and the DC problem for the pool dependent case. This result is largely based on the methods developed in Atar [3]. In a setting where the service rates may depend on both the class and the pool, Atar [3] finds such a relation between the two problems, based on an I-dimensional HJB equation. The tools developed there are applicable in the current setting as well. The rigorous relation established in the current paper is twofold. First, it is shown that under any (admissible) policy for the RSC problem, the expected discounted cost associated with weighted queue-lengths is, in the scaling limit, bounded below by the value function of a DC problem. Second, a sequence of (nonpreemptive, admissible) policies, constructed via the corresponding HJB equation, is proved to be asymptotically optimal. Adaptations of large time estimates from Atar [3], and estimates for policies that are not JWC are required to conclude the result.

We do not carry out an analogous program for class dependent services. Although the DC model obtained in this case is simpler than the general DC model derived in Atar [3], it resides in the same dimension as this general model (namely I), for which a rigorous relation has already been established (Atar [3]), and so such a result would, perhaps, be less significant.

There is potentially a third contribution, that has to do with the value of our results in supporting future research. Specifically, there are important research themes in which optimal control is only part of the overall problem. e.g., staffing (see Gans et al. [10]), and pooling design (see §4.2 in Aksin et al. [1]). In such cases, a simplified DC problem could render tractable and insightful, at least asymptotically, a problem to which a solution of its raw form seems infeasible (Gurvich and Whitt [12] is a case in point).

We introduce the DC model in §2. Sections 3 and 4 specialize this model to the pool dependent and class dependent cases, respectively. Finally, §5 is devoted to the rigorous relation between the RSC and the DC problems in the case of pool dependent service rates.

2. The diffusion model. For the queueing model described in the introduction, let $\mathcal{I} = \{1, \ldots, I\}$ and $\mathcal{J} = \{I+1, \ldots, I+J\}$ denote index sets for customer classes and server pools, respectively. Denote by $\mathcal{G}$ the graph with vertex set $\mathcal{I} \cup \mathcal{J}$, and with an edge between $i \in \mathcal{I}$ and $j \in \mathcal{J}$ if and only if the pair (i, j) forms an *activity*, in the sense that class-i can be served by pool-j servers. The edge set of $\mathcal{G}$ is denoted by $\mathcal{E}_{\text{a}}$, where "a" is mnemonic for activity. We write $i \sim j$ if (i, j) is an activity. The graph $\mathcal{G}$ is assumed to be connected. The model parameters λ_i and θ_i represent arrival and individual abandonment rates for class-i customers, and μ_{ij} represent service rate of class-i customers by pool-j servers, with the convention that $\mu_{ij} = 0$ whenever $(i, j) \in \mathcal{I} \times \mathcal{J}$ is not an activity (and, of course, $\mu_{ij} > 0$ when (i, j) is an activity). Throughout, we will assume

$$\mu_{\min} \geq \theta_{\max}, \tag{1}$$

where

$$\mu_{\min} = \min\{\mu_{ij}\colon (i, j) \in \mathcal{E}_{\text{a}}\}, \qquad \theta_{\max} = \max\{\theta_i\colon i \in \mathcal{I}\}.$$

Allowing some mean service time to exceed some mean patience would render attractive the option of "no service," in which case our approach, which requires work conservation, does not apply, and a different approach might be required (see (38) for the use of (1) in the proof).

A corresponding diffusion model representing heavy traffic limits in the regime of Halfin and Whitt has been derived in Atar [3] from the queueing model equations. This entails a number of steps that include scaling up of the arrival rates and number of servers at each pool, in such a way that an underlying static planning problem is, in an appropriate sense, critically loaded. Also, scaling and centering were applied to the processes involved in the description of the system's dynamics, in such a way that they exhibit diffusive fluctuations. While the probabilistic model is fully described in §5, we do not provide here the details of the derivation, and refer the reader to the above citation Atar [3] and to Atar et al. [5]. We reiterate that rigorous connections between the DC problem and the asymptotics of the RSC problem have been established in various settings (Atar [3], Atar et al. [4, 5], and additional references in Aksin et al. [1]).

To describe the DC problem we need to introduce additional structure. This includes the notion of *basic* activities. Roughly speaking, an activity (i, j) is basic if the proportion of the number of servers from pool j allocated to class i customers is nontrivial in the scaling limit. The precise definition is via an underlying static planning problem, and is not provided in this section (for details see §5 and Atar [3]). To formulate the diffusion model, it suffices to say that some of the activities are regarded as basic. Activities that are not basic are said to be *nonbasic*. We denote by $\mathcal{E}_{\text{ba}}$ and $\mathcal{E}_{\text{nba}}$ the subsets of $\mathcal{E}_{\text{a}}$ of basic and nonbasic activities, respectively. A principal assumption of Atar [3], Atar et al. [5], and assumed throughout this paper, is that *the graph with vertex set $\mathcal{I} \cup \mathcal{J}$ and edge set $\mathcal{E}_{\text{ba}}$ is a tree*, i.e., a graph with no cycles, connecting each pair of vertices from $\mathcal{I} \cup \mathcal{J}$ by a path of edges from $\mathcal{E}_{\text{ba}}$. This assumption corresponds to the complete resource pooling condition of Harrison and López [14] (i.e., it is equivalent to the complete resource pooling conditions, provided that the underlying fluid model is critically loaded, in an appropriate sense; see §5 and Atar [3]). Figure 2(a–b) depicts examples where our assumption is satisfied, and Figure 2(c) where it is not satisfied.

The process representing, for all $i \in \mathcal{I}$, the number of class-i customers present in the queueing system at each time, gives rise in the scaling limit to an $\mathbb{R}^I$-valued process, which we denote by $X = (X_i)_{i \in \mathcal{I}}$. Similarly, the

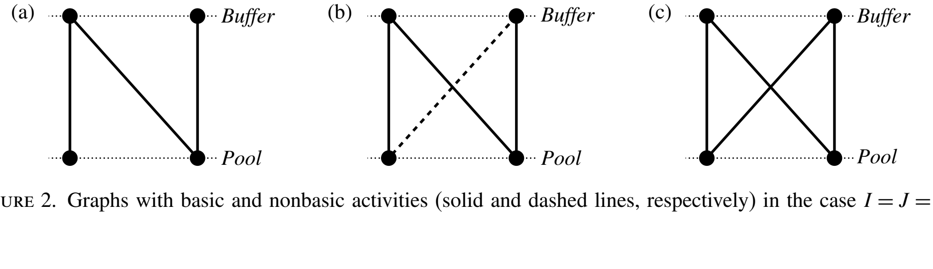

FIGURE 2. Graphs with basic and nonbasic activities (solid and dashed lines, respectively) in the case $I = J = 2$.

processes denoted by $(Y_i)_{i\in\mathcal{I}}$, and $(Z_j)_{j\in\mathcal{J}}$, correspond to queue-length of class i and the number of idle servers at pool j, respectively, and $\Psi=(\Psi_{ij})_{i\in\mathcal{I},\, j\in\mathcal{J}}$ corresponds to the number of class-i customers in service at pool j. Finally, $(W_i)_{i\in\mathcal{I}}$ is an I-dimensional Brownian motion with drift b and nondegenerate, diagonal covariance matrix Σ that represents the effect of fluctuations in arrival and service times. We refer to such a process, in the sequel, as a (b,Σ)-Brownian motion. Here b and Σ are fixed parameters that depend on μ_{ij}, θ_i, λ_i, and second moments of the interarrival times (such as the constants (ℓ, r) defined on the line following (33) of Atar [3]).

The equations satisfied by the processes (X,Y,Z,Ψ), referred to below as the *diffusion equations*, are as follows (see Atar [3] for a derivation of the equations from the queueing model):

$$X_i(t)=x_i+W_i(t)-\sum_{j\in\mathcal{J}}\mu_{ij}\int_0^t\Psi_{ij}(s)\,ds-\theta_i\int_0^t Y_i(s)\,ds,\quad i\in\mathcal{I},\tag{2}$$

$$\sum_{j\in\mathcal{J}}\Psi_{ij}=X_i-Y_i,\quad i\in\mathcal{I},\tag{3}$$

$$\sum_{i\in\mathcal{I}}\Psi_{ij}=-Z_j,\quad j\in\mathcal{J},\tag{4}$$

$$\Psi_{ij}=0,\quad i\not\sim j,\qquad \Psi_{ij}\ge 0,\quad (i,j)\in\mathcal{E}_{\mathrm{nba}},\tag{5}$$

$$Y_i\ge 0,\ Z_j\ge 0\quad i\in\mathcal{I},\ \ j\in\mathcal{J}.\tag{6}$$

We need a precise definition of a diffusion model. To this end, let a complete filtered probability space $\{\Omega,\mathcal{F},\{\mathcal{F}_t\},\mathbb{P}\}$ and an (I-dimensional) (b,Σ)-Brownian motion W be given, where $W(t)-bt$ is an $\{\mathcal{F}_t\}$-martingale. We denote by $\mathcal{M}$ the set of all processes $\mathbf{M}=(X,Y,Z,\Psi)$ on the given probability space that satisfy

- $\mathbf{M}$ is $\{\mathcal{F}_t\}$-progressively measurable,
- The diffusion Equations (2)–(6) are satisfied $\mathbb{P}$-a.s.

Any nonempty subset of $\mathcal{M}$ is said to be a *diffusion model*.

To minimize queue-length related costs, we define a subset of a given diffusion model as its *reduction* if for each member of the original model one can find, in the reduced model, a member for which the processes representing queue-length are all smaller. (See Remark 5.1 for a relation to delay costs.)

The precise definition is as follows. Throughout, we write "$\le$" for the usual partial order on $\mathbb{R}^I$ (or $\mathbb{R}^J$), and, if A and B are $\mathbb{R}^I$-valued processes, interpret $A\le B$ as $A(t)\le B(t)$ for all $t\ge 0$, $\mathbb{P}$-a.s.

Let $\mathcal{M}^1$ and $\mathcal{M}^2$ be diffusion models. Then $\mathcal{M}^2$ is said to be a reduction of $\mathcal{M}^1$, if $\mathcal{M}^2\subset\mathcal{M}^1$, and for any $(X^1,Y^1,Z^1,\Psi^1)\in\mathcal{M}^1$ there exists $(X^2,Y^2,Z^2,\Psi^2)\in\mathcal{M}^2$, such that $Y^2\le Y^1$.

If, for example, a cost $C(\mathbf{M})$ of the form

$$E\biggl[\int_0^\infty e^{-t}f(Y(t))\,dt\biggr]\tag{7}$$

is to be minimized, where $f\colon\mathbb{R}_+^I\to\mathbb{R}_+$ is nondecreasing (with respect to the usual partial order on $\mathbb{R}_+^I$), then the problem of minimizing the cost over $\mathcal{M}^1$ can be reduced to that over $\mathcal{M}^2$.

3. Simplified model in the case of pool dependent service rates.

In this section we show that when the service rates are pool dependent, the diffusion model reduces to one governed by a 1-dimensional stochastic differential equation (SDE). To this end, we will assume throughout the section that there are constants $\mu_j>0$, $j\in\mathcal{J}$, such that for every $(i,j)\in\mathcal{E}$,

$$\mu_{ij}=\mu_j.\tag{8}$$

Under (8), using (5), Equation (2) can be written as

$$X_i(t)=x_i+W_i(t)-\sum_{j\in\mathcal{J}}\mu_j\int_0^t\Psi_{ij}(s)\,ds-\theta_i\int_0^t Y_i(s)\,ds,\quad i\in\mathcal{I}.\tag{9}$$

We will write $e_i\in\mathbb{R}^I$ (or $\mathbb{R}^J$) for the ith coordinate vector, and $e\in\mathbb{R}^I$ (or $\mathbb{R}^J$) for the vector with all entries being one. Throughout, for any I- (or J-) dimensional process A, we denote by A_e the process $e\cdot A$, where "$\cdot$" denotes the usual scalar product on $\mathbb{R}^I$ (or $\mathbb{R}^J$). For $y,z\in\mathbb{R}$, we write $y\wedge z$ for $\min\{y,z\}$.

Recall that a JWC policy for the queueing system allows only arrangements of customers in the system in which no server idles when a customer of any class is waiting in the queue (a precise definition of a JWC policy for the queueing system appears in §5). This can be expressed by saying that the minimum between the total queue-length and the total number of idle servers is zero at all times. In the diffusion model, this condition corresponds to $Y_e(t) \wedge Z_e(t) = 0$, $t \geq 0$. The result below, the proof of which is deferred to the end of the section, provides a two-step reduction of the diffusion model $\mathcal{M}$. The first reduces $\mathcal{M}$ to a model $\mathcal{M}^1$ in which the JWC condition holds and nonbasic activities are not used. The second step provides a further reduction to $\mathcal{M}^2$, in which only servers in the pool with the slowest service rate may idle.

To state the following result, let a $j_0 \in \mathcal{J}$ satisfying

$$\mu_{j_0} = \mu_{\min}$$

be fixed.

Proposition 3.1.

(i) *Let $\mathcal{M}^1$ be the set of all processes $\mathbf{M} \in \mathcal{M}$, that satisfy $\mathbb{P}$-a.s.*

(a) $Y_e(t) \wedge Z_e(t) = 0$, $t \geq 0$,

(b) $\Psi_{ij}(t) = 0$ *for* $(i, j) \in \mathscr{E}_{nba}$, $t \geq 0$.

Then $\mathcal{M}^1$ is a reduction of $\mathcal{M}$.

(ii) *Let $\mathcal{M}^2$ be the set of all processes $\mathbf{M} \in \mathcal{M}^1$ that satisfy $Z(t) = Z_e(t)e_{j_0}$ for all $t \geq 0$, $\mathbb{P}$-a.s. Then $\mathcal{M}^2$ is a reduction of $\mathcal{M}^1$.*

Because of the tree structure of the basic activities it is not hard to see that JWC policies exist. Furthermore, one can ensure that all idle servers belong to a fixed pool. The above result states the intuitively clear claim that it is best to maintain all idleness at the slowest pool.

To state the main result we need some further notation. Denote

$$\mathbb{U} = \left\{ u \in \mathbb{R}^I \colon u_i \geq 0, \sum_i u_i = 1 \right\}, \tag{10}$$

and let $\mathcal{U}$ be the set of all $\mathbb{U}$-valued progressively measurable processes. We shall later also need the analogous notation

$$\mathbb{V} = \left\{ v \in \mathbb{R}^J \colon v_j \geq 0, \sum_j v_j = 1 \right\},$$

and $\mathcal{V}$ for the set of all $\mathbb{V}$-valued progressively measurable processes. For $y \in \mathbb{R}$, denote $y^+ = \max\{y, 0\}$ and $y^- = \max\{-y, 0\}$. Let μ and θ denote the vectors $(\mu_j)_{j \in \mathcal{J}}$ and $(\theta_i)_{i \in \mathcal{I}}$, respectively. For $u \in \mathcal{U}$ it is well known that there exists a unique solution $\check{X}$ to the following equation (see, for example, Theorem V.3.7 (p. 197) of Protter [17]), namely,

$$\check{X}(t) = x_e + W_e(t) + \mu_{\min} \int_0^t \check{X}^-(s)\, ds - \int_0^t \theta \cdot u(s)\, \check{X}^+(s)\, ds, \quad t \geq 0. \tag{11}$$

Owing to the tree structure of $\mathscr{E}_{\text{ba}}$, the system of equations in the unknown ϕ,

$$\begin{cases} \sum_{j \in \mathcal{J}} \phi_{ij} = a_i, & i \in \mathcal{I}, \\ \sum_{i \in \mathcal{I}} \phi_{ij} = b_j, & j \in \mathcal{J}, \\ \phi_{ij} = 0, & (i, j) \in \mathcal{I} \times \mathcal{J} \backslash \mathscr{E}_{\text{ba}}, \end{cases} \tag{12}$$

has a unique solution, whenever a and b satisfy $\sum_i a_i = \sum_j b_j$; see Proposition A.2 of Atar [2]. With

$$D_G = \left\{ (a, b) \in \mathbb{R}^I \times \mathbb{R}^J \colon \sum_{i \in \mathcal{I}} a_i = \sum_{j \in \mathcal{J}} b_j \right\},$$

denote by G: $D_G \to \mathbb{R}^{I \times J}$ the solution map, namely,

$$\phi_{ij} = G_{ij}(a, b), \quad (i, j) \in \mathcal{I} \times \mathcal{J}, \tag{13}$$

and note that this map is linear. Note also that when the relation $\Psi = G(X - Y, -Z)$ holds, Equations (3)–(5) are satisfied and, in addition, $\Psi_{ij} = 0$ for all $(i, j) \notin \mathscr{E}_{\text{ba}}$. Define

$$H_i(x, u) = -\sum_j \mu_j G_{ij}(x - x_e^+ u, -x_e^- e_{j_0}) - \theta_i x_e^+ u, \quad i \in \mathcal{I} \tag{14}$$

and $H = (H_i)_{i\in\mathcal{I}}$. Given a process $u \in \mathcal{U}$, let $\mathcal{M}^{(u)}$ denote the set of all processes (X, Y, Z, Ψ) for which

$$X(t) = x + W(t) + \int_0^t H(X(s), u(s))\, ds, \tag{15}$$

$$Y(t) = X_e^+(t)u(t), \tag{16}$$

$$Z(t) = X_e^-(t)e_{j_0}, \tag{17}$$

$$\Psi_{ij}(t) = G_{ij}(X(t) - X_e^+(t)u(t), -X_e^-(t)e_{j_0}), \quad i \in \mathcal{I},\ \ j \in \mathcal{J}, \tag{18}$$

holds for all $t \geq 0$, $\mathbb{P}$-a.s. Let

$$\mathcal{M}^0 = \bigcup_{u\in\mathcal{U}} \mathcal{M}^{(u)}.$$

Recall the notation $\mathcal{M}^2$ from Proposition 3.1.

Theorem 3.1. *One has*

(i) $\mathcal{M}^0 = \mathcal{M}^2$, *and consequently* $\mathcal{M}^0$ *is a reduction of* $\mathcal{M}$.

(ii) *For any* $u \in \mathcal{U}$, *a corresponding solution* $\check{X}$ *of* (11) *and a process* $(X, Y, Z, \Psi) \in \mathcal{M}^{(u)}$, *the equality* $X_e = \check{X}$ *holds* $\mathbb{P}$-*a.s.*

This result, proved at the end of this section, establishes a reduction of the diffusion model to a one-dimensional model. To see this, let us consider again a cost $C(\mathbf{M})$ of the form (7) where f is nondecreasing. Given a process $u \in \mathcal{U}$, let $\check{X}$ be the solution to (11) and set

$$\check{C}(u) = E\biggl[\int_0^\infty e^{-t} f(\check{Y}(t))\, dt\biggr],$$

where $\check{Y}(t) = \check{X}^+(t)u(t)$. Suppose u minimizes $\check{C}$ over $\mathcal{U}$, and let $\mathbf{M} = (X, Y, Z, \Psi)$ be defined via (15)–(18). Then by Theorem 3.1, $(X, Y, Z, \Psi) \in \mathcal{M}$ and $\check{Y} = Y$. Consequently, $C(\mathbf{M}) = \check{C}(u)$. Moreover, because by construction $\mathbf{M} \in \mathcal{M}^0$, and because by Theorem 3.1 $\mathcal{M}^0$ is a reduction of $\mathcal{M}$, we have $C(\mathbf{M}) \leq C(\widetilde{\mathbf{M}})$ for any $\widetilde{\mathbf{M}} \in \mathcal{M}$. Therefore, solving the problem of minimizing $\check{C}$ suffices for solving the original problem. In some cases, a pathwise solution to the problem exists, that is, there is $\mathbf{M} = (X, Y, Z, \Psi) \in \mathcal{M}$ such that for every $(\widetilde{X}, \widetilde{Y}, \widetilde{Z}, \widetilde{\Psi}) \in \mathcal{M}$ one has $f(Y(t)) \leq f(\widetilde{Y}(t))$, $t \geq 0$, $\mathbb{P}$-a.s. By a similar reasoning as above, to find the pathwise minimizer (if it exists), it suffices to find the u which minimizes $f(\check{Y}(t))$ for all $t \geq 0$, where, again, $\check{Y}(t) = \check{X}^+(t)u(t)$ and $\check{X}$ is the corresponding solution to (11).

As mentioned in the introduction, there is similarity between the 1-dimensional equation and the case of a single class (i.e., (11) with $\theta \cdot u$ replaced by θ_1). In both cases the term $\mu_{\min} \int_0^t \check{X}^-(s)\, ds$ arises for the same reason, namely that under the optimal policy, only servers from the slowest pool idle at any time. In fact, when $\theta_i = \theta_1$ and $c_i = c_1$ for all i, the multiclass model and the cost are identical to the single class one. However, for general $\{\theta_i\}$ and $\{c_i\}$, the multiclass model is more complicated, and how to optimally distribute workload among the classes depends on these parameters. This is demonstrated in the examples below, which show how Theorem 3.1 can be applied. The first three describe cases in which the one dimensional model $\mathcal{M}^0$ can be solved pathwise. To begin with, the case where $\theta_i = \theta_1$ for all $i \in \mathcal{I}$ is particularly simple because the dynamics (11) are not affected by the control $(\theta \cdot u(s) = \theta_1)$. This is demonstrated in the first two examples. Throughout, let $c = (c_1, \ldots, c_I)'$, $c_i > 0$, be a constant vector.

Example 3.1 (Linear Costs). Assume $\theta_i = \theta_1$ for all $i \in \mathcal{I}$. Consider the stochastic process

$$\tilde{c}(\mathbf{M})(t) = c \cdot Y(t), \quad \mathbf{M} \in \mathcal{M}. \tag{19}$$

By the discussion following Theorem 3.1, the problem of pathwise minimizing $\tilde{c}(\mathbf{M})$ (provided that a pathwise minimum exists) can be reduced to that of pathwise minimizing

$$\check{c}(u)(t) = \check{X}^+(t)c \cdot u(t), \quad u \in \mathcal{U}, \tag{20}$$

where $\check{X}$ is given by (11). Since $\check{X}$ is not affected by u, the process $\check{c}$ is clearly minimized by selecting $u(t) = e_{i_0}$, for all t, where i_0 is such that $c_{i_0} = \min c_i$. Hence, as follows from Theorem 3.1, defining $\mathbf{M}$ via (15)–(18), with the above value for u, results with pathwise minimizing the process $\tilde{c}(\mathbf{M})$ over $\mathcal{M}$. As a result, the process $\mathbf{M}$ identified above clearly minimizes a cost such as $E[\int_0^\infty e^{-t} c \cdot Y(t)\, dt]$ over $\mathcal{M}$. □

The next example is related to the generalized $c\mu$ rule of Mandelbaum and Stolyar [16].

Example 3.2 (Convex Costs). Assume $\theta_i = \theta_1$ for all $i \in \mathcal{I}$, and let

$$\tilde{c}(\mathbf{M})(t) = \sum_{i=1}^{I} C_i(Y_i), \quad \mathbf{M} \in \mathcal{M}.$$

We check whether $\tilde{c}$ can be minimized pathwise. Here the functions C_i are assumed to be strictly convex and continuously differentiable with $C_i'(0) = 0$ for all i. By Theorem 3.1, the problem can be reduced to that of pathwise minimizing the process $\check{c}$ over $\mathcal{U}$, where

$$\check{c}(u)(t) = \sum_{i=1}^{I} C_i(\check{X}^+(t) u_i(t)), \quad u \in \mathcal{U}.$$

Since $\check{X}$ is again not affected by the control u, the reduced minimization problem is solved by finding the minimum over $u(t) \in \mathbb{U}$, for each $t \geq 0$. For $y \in \mathbb{R}_+$, consider the problem

$$\min_{y_1,\ldots,y_I} \sum_i C_i(y_i), \quad \text{s.t.} \quad y_1 + \cdots + y_I = y, \quad y_i \geq 0, \ i \in \mathcal{I}.$$

Let C_i' denote the derivative of C_i. Because of the assumptions on C_i, the unique solution $\{y_1^*, \ldots, y_I^*\} = \{y_1^*(y), \ldots, y_I^*(y)\}$ satisfies

$$C_1'(y_1^*) = \cdots = C_I'(y_I^*), \qquad \text{and} \qquad \sum_i y_i^* = y.$$

This is translated into a solution to the reduced DC problem by letting $u_i(t) = y_i^*(\check{X}(t)) / \sum_k y_k^*(\check{X}(t))$, for every $t \geq 0$. In turn, the solution to the original DC problem is found by defining $\mathbf{M}$ via (15)–(18). □

Next, we consider an example with slightly more general θ_i's, where pathwise minimum can still be obtained.

Example 3.3 (Special Ordering of Parameters). Consider the problem of minimizing the linear cost (19) and assume that

$$\theta_1 \geq \theta_i, \qquad c_1 \leq c_i, \quad \text{for all } i \geq 2.$$

Then the control $u(t) = e_1$ achieves the minimum of $\check{c}$ of (20). Consequently, $\mathbf{M}$ defined via the transformation (15)–(18) minimizes pathwise $\tilde{c}(\mathbf{M})$ of (19). This claim is proved at the end of this section. □

Finally, we consider general θ_i and c_i, where a solution to the DC problem can be obtained via control theoretic tools. This example is the basis for our asymptotic analysis in §5.

Example 3.4 (Hamilton-Jacobi-Bellman Ordinary Differential (ODE)). Let

$$\widetilde{V}(x) = \inf_{\mathbf{M} \in \mathcal{M}} E\left[\int_0^\infty e^{-t} c \cdot Y(t)\, dt\right]. \tag{21}$$

By Theorem 3.1,

$$\widetilde{V}(x) = V(\xi), \tag{22}$$

where $\xi = x_e$, and

$$V(\xi) = \inf_{u \in \mathcal{U}} E\left[\int_0^\infty e^{-t} \check{X}(t)^+ c \cdot u(t)\, dt\right]. \tag{23}$$

Let us denote by ℓ and ϱ^2 the drift and, respectively, diffusion coefficient of the 1-dimensional Brownian motion W_e (cf. (11)). Let

$$L(\xi, u) = \xi^+ c \cdot u, \quad \xi \in \mathbb{R}, \ u \in \mathbb{U},$$
$$\beta(\xi, u) = \mu_{\min} \xi^- - \theta \cdot u \xi^+ + \ell, \quad \xi \in \mathbb{R}, \ u \in \mathbb{U}, \tag{24}$$

and note that (11) can be written as

$$\check{X}(t) = \xi + \varrho W_S(t) + \int_0^t \beta(\check{X}(s), u(s))\, ds, \quad t \geq 0, \tag{25}$$

where W_S is a standard Brownian motion. With

$$\overline{\mathbf{H}}(\xi, q, u) = \beta(\xi, u) q + L(\xi, u), \quad \xi \in \mathbb{R}, \ q \in \mathbb{R}, \ u \in \mathbb{U},$$
$$\mathbf{H}(\xi, q) = \inf_{u \in \mathbb{U}} \overline{\mathbf{H}}(\xi, q, u), \quad \xi \in \mathbb{R}, \ q \in \mathbb{R},$$

the HJB equation for V associated with the controlled diffusion (25) is given as

$$(1/2)\varrho^2 V''(\xi) + \mathbf{H}(\xi, V'(\xi)) - V(\xi) = 0, \quad \xi \in \mathbb{R}, \tag{26}$$

where V' and V'' denote first and second derivatives (see Fleming and Soner [9, §IV.5]). Moreover, when considered with the growth condition

$$\exists C, |V(\xi)| \leq C(1 + |\xi|^C), \quad \xi \in \mathbb{R}, \tag{27}$$

V is the only solution of Equation (26). This unique solvability statement is a standard fact, given a subexponential large time estimate on $E[\check{X}(t)]$. Such an estimate is provided by Lemma A.1(i) in the appendix. For an existence and uniqueness proof (based on large time estimate), see Atar [2]. Moreover, it is known (Atar [2]) that there exists a measurable function h: $\mathbb{R} \to \mathbb{U}$ such that

$$\overline{\mathbf{H}}(\xi, V'(\xi), h(\xi)) = \mathbf{H}(\xi, V'(\xi)), \quad \xi \in \mathbb{R}, \ \ u \in \mathbb{U}.$$

This function is the "optimal feedback" from state to control, in the sense that the unique solution $(\check{X}, u)$ to Equation (11), in conjunction with the equation $u(t) = h(\check{X}(t))$, achieves the infimum (23). □

Remark 3.1. The following two points will be important for §5.

(i) The function h can be approximated by Lipschitz functions in the following sense. For every $k \in \mathbb{N}$ there exists a globally Lipschitz function h_k: $\mathbb{R} \to \mathbb{U}$, such that

$$\overline{\mathbf{H}}(\xi, V'(\xi), h_k(\xi)) \leq \mathbf{H}(\xi, V'(\xi)) + 1/k, \quad \xi \in [-k, k], \ \ u \in \mathbb{U}. \tag{28}$$

This claim follows from the proof of Theorem 2(iv) of Atar [3].

(ii) Let

$$\tilde{\beta}(\xi, u, v) = \mu \cdot v\, \xi^- - \theta \cdot u\, \xi^+ + \ell, \tag{29}$$

$$\widetilde{\mathbf{H}}(\xi, q) = \inf_{(u,v) \in \mathbb{U} \times \mathbb{V}} [\tilde{\beta}(\xi, u, v) q + L(\xi, u)]. \tag{30}$$

Then the equation obtained from (26) by replacing $\mathbf{H}$ with $\widetilde{\mathbf{H}}$, considered with (27), has a unique solution; this solution is identical to that of Equation (26), as follows from the optimality of controls that keep $v(t) = e_{j_0}$, stated in Proposition 3.1(ii). □

We now proceed with the proofs of the results stated earlier in this section.

Proof of Proposition 3.1. (i) Let $(X, Y, Z, \Psi) \in \mathscr{M}$ be given. We will construct $(\widetilde{X}, \widetilde{Y}, \widetilde{Z}, \widetilde{\Psi}) \in \mathscr{M}^1$ such that $\widetilde{Y} \leq Y$, $\mathbb{P}$-a.s.

By summing up Equations (9) over all i's and using (4) we obtain

$$X_e(t) = x_e + W_e(t) + \sum_{j \in \mathcal{J}} \mu_j \int_0^t Z_j(s)\, ds - \sum_{i \in \mathcal{I}} \theta_i \int_0^t Y_i(s)\, ds. \tag{31}$$

Define

$$M(t) = Y_e(t) \wedge Z_e(t), \quad t \geq 0. \tag{32}$$

Observe that $M(t) \geq 0$, as follows from (6). From (3) and (4) we have $X_e = Y_e - Z_e$. Therefore, by (32), $Y_e = X_e^+ + M$ and $Z_e = X_e^- + M$. Hence there exists a process $(u, v) \in \mathscr{U} \times \mathscr{V}$ such that, for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$,

$$Y_i = (X_e^+ + M)u_i, \qquad Z_j = (X_e^- + M)v_j. \tag{33}$$

Rewrite (31) as

$$X_e(t) = x_e + W_e(t) + \int_0^t \mu \cdot v(s)\, X_e^-(s)\, ds - \int_0^t \theta \cdot u(s)\, X_e^+(s)\, ds + \int_0^t (\mu \cdot v(s) - \theta \cdot u(s)) M(s)\, ds. \tag{34}$$

We define $\widetilde{X}$ as the solution to (9) with

$$\widetilde{Y} = \widetilde{X}_e^+ u, \qquad \widetilde{Z} = \widetilde{X}_e^- v, \tag{35}$$

$$\widetilde{\Psi} = G(\widetilde{X} - \widetilde{Y}, -\widetilde{Z}). \tag{36}$$

This $\widetilde{X}$ exists and is uniquely defined, because the integrand in (9) is Lipschitz as a function of $\widetilde{X}$. Once $\widetilde{X}$ is well defined, $(\widetilde{Y}, \widetilde{Z}, \widetilde{\Psi})$ are defined via (35)–(36). By construction, $\widetilde{\mathbf{M}} = (\widetilde{X}, \widetilde{Y}, \widetilde{Z}, \widetilde{\Psi})$ is progressively measurable and satisfies the diffusion Equations (9) and (3)–(6). By (35), $\widetilde{Y}_e \wedge \widetilde{Z}_e = 0$, and by (36) and the properties of G (12), we also have $\Psi_{ij} = 0$ for $(i, j) \in \mathscr{E}_{\mathrm{nba}}$. Hence $\widetilde{\mathbf{M}} \in \mathscr{M}^1$. By construction,

$$\widetilde{X}_e(t) = x_e + W_e(t) + \int_0^t \mu \cdot v(s)\widetilde{X}_e^-(s)\,ds - \int_0^t \theta \cdot u(s)\widetilde{X}_e^+(s)\,ds. \tag{37}$$

Let $\Delta = \widetilde{X}_e - X_e$. It follows from (1) that

$$\mu \cdot v(t) \geq \theta \cdot u(t), \qquad t \geq 0. \tag{38}$$

Hence, by (34) and (37),

$$\begin{aligned}\frac{d}{dt}\Delta(t) &\leq \mu \cdot v(t)(\widetilde{X}_e^-(t) - X_e^-(t)) + \theta \cdot u(t)(X_e^+(t) - \widetilde{X}_e^+(t))\\ &\leq \mu \cdot v(t)\Delta^-(t) + \theta \cdot u(t)\Delta^-(t)\\ &\leq c\Delta^-(t),\end{aligned} \tag{39}$$

where we used the inequalities

$$a^- - b^- \leq (a-b)^-, \qquad a^+ - b^+ \leq (a-b)^+ = (b-a)^-,$$

and where $c = \max \mu_j + \max \theta_i$. The equation $(d/dt)\delta(t) = c\delta^-(t)$ with the initial condition $\delta(0) = 0$ has a unique solution $\delta = 0$. By a standard comparison theorem for ODE (Theorem 7, p. 22, Birkhoff and Rota [6]), we obtain $\widetilde{X}_e(t) \leq X_e(t)$ for all t. Thus, by (33), (35) and the positivity of M, $\widetilde{Y} \leq Y$. This proves part (i).

(ii) Let $(X, Y, Z, \Psi) \in \mathscr{M}^1$ be given. An argument similar to the one in part (i) shows that there exists a process $(u, v) \in \mathscr{U} \times \mathscr{V}$ such that

$$Y = X_e^+ u, \qquad Z = X_e^- v, \tag{40}$$

and

$$X_e(t) = x_e + W_e(t) + \int_0^t \mu \cdot v(s)\, X_e^-(s)\,ds - \int_0^t \theta \cdot u(s)\, X_e^+(s)\,ds. \tag{41}$$

Set $\tilde{u}(t) = u(t)$ and $\tilde{v}(t) = e_{j_0}$, $t \geq 0$. Let $\widetilde{X}$ be defined as the unique solution to (9) with

$$\widetilde{Y} = \widetilde{X}_e^+ \tilde{u}, \qquad \widetilde{Z} = \widetilde{X}_e^- \tilde{v}, \qquad \widetilde{\Psi} = G(\widetilde{X} - \widetilde{Y}, -\widetilde{Z}). \tag{42}$$

Let $(\widetilde{Y}, \widetilde{Z}, \widetilde{\Psi})$ be defined by the above display. By an argument as in part (i), the process $\widetilde{\mathbf{M}} = (\widetilde{X}, \widetilde{Y}, \widetilde{Z}, \widetilde{\Psi})$ thus constructed is in $\mathscr{M}^1$. By (42), $\widetilde{Z} = \widetilde{X}_e^- e_{j_0}$ hence $\widetilde{Z} = \widetilde{Z}_e e_{j_0}$ and thus $\widetilde{\mathbf{M}} \in \mathscr{M}^2$. Since

$$\mu \cdot v(t) \geq \mu \cdot \tilde{v}(t) = \mu_{\min},$$

we have, with $\Delta = \widetilde{X}_e - X_e$,

$$\begin{aligned}\frac{d}{dt}\Delta(t) &\leq \mu \cdot v(t)(\widetilde{X}_e^-(t) - X_e^-(t)) + \theta \cdot u(t)(X_e^+(t) - \widetilde{X}_e^+(t))\\ &\leq \mu \cdot v(t)\Delta^-(t) + \theta \cdot u(t)\Delta^-(t)\\ &\leq c\Delta^-(t).\end{aligned}$$

A comparison argument as in part (i) yields $\widetilde{X}_e \leq X_e$. Hence, for every $i \in \mathscr{I}$, $\widetilde{Y}_i = \widetilde{X}_e^+ u_i \leq X_e^+ u_i = Y_i$. This proves the result. □

Proof of Theorem 3.1. (i) Let $\mathbf{M} = (X, Y, Z, \Psi) \in \mathscr{M}^2$. In particular, $\mathbf{M} \in \mathscr{M}^1$, and by the proof of Proposition 3.1(ii), (40) and (41) hold for some $(u, v) \in \mathscr{U} \times \mathscr{V}$. Since Z must satisfy $Z = Z_e e_{j_0}$, we have that $v = e_{j_0}$. We show that $\mathbf{M}$ satisfies Equations (15)–(18). Equations (16) and (17) follow from (40) and the fact $v = e_{j_0}$. As a member of $\mathscr{M}^2$, $\mathbf{M}$ satisfies the diffusion equations, and in particular Ψ, X, and Y satisfy (3)–(5). Recalling the definition of G as the solution map of (12), it follows that (18) holds. Finally, the validity of (15) follows from (9) and (14). This shows $\mathscr{M}^2 \subset \mathscr{M}^0$.

Now let $\mathbf{M} = (X, Y, Z, \Psi) \in \mathscr{M}^0$. Then (15)–(18) hold. To see that $\mathbf{M} \in \mathscr{M}$, note that (16)–(18) and the definition of G imply (3)–(6). Equation (9) follows from (15) and (14). Thus $\mathbf{M} \in \mathscr{M}$. The properties $Y_e \wedge Z_e = 0$, $\Psi_{ij} = 0$ for $(i, j) \in \mathscr{E}_{\mathrm{nba}}$ and $Z = Z_e e_{j_0}$ are obvious from Equations (16)–(18) and the definition of G, and therefore $\mathbf{M} \in \mathscr{M}^2$. We conclude that $\mathscr{M}^0 = \mathscr{M}^2$. By Proposition 3.1 we therefore have that $\mathscr{M}^0$ is a reduction of $\mathscr{M}$.

(ii) Given $u \in \mathscr{U}$ let $(X, Y, Z, \Psi) \in \mathscr{M}^{(u)}$. It follows from (15)–(18) and the definition of G that X_e satisfies Equation (11). By uniqueness of solutions it follows that $X_e = \check{X}$ $\mathbb{P}$-a.s. □

Proof of Claim Made in Example 3.3. Let $\tilde{u}(t)$ be any control and denote by $\widetilde{X}$ the corresponding controlled process (11). Let $u(t) = e_1$, $t \geq 0$, and denote the corresponding controlled process by X. Using the relation $\theta_1 = \theta \cdot u \geq \theta \cdot \tilde{u}$ and repeating the comparison argument from the proof of Proposition 3.1(ii), we obtain that $X \leq \widetilde{X}$. In conjunction with the inequality $c_1 = c \cdot u \leq c \cdot \tilde{u}$ this yields $(c \cdot u)X^+ \leq (c \cdot \tilde{u})\widetilde{X}^+$. Therefore the cost (20) is pathwise minimized by u. □

4. Simplified model in the case of class dependent service rates.

In this section, we assume that there are constants $\mu_i > 0$, $i \in \mathscr{I}$, such that for every $(i, j) \in \mathscr{E}$,

$$\mu_{ij} = \mu_i. \tag{43}$$

We fix a model $\mathscr{M}$, with general I and J. We will relate it to a model with a single pool, which we will denote by $\mathscr{M}^{\mathrm{sp}}$ for which $I^{\mathrm{sp}} = I$ and $J^{\mathrm{sp}} = 1$. The diffusion equations for $\mathscr{M}^{\mathrm{sp}}$ are given by

$$X_i(t) = x_i + W_i(t) - \mu_i \int_0^t \Psi_i(s)\,ds - \theta_i \int_0^t Y_i(s)\,ds, \quad i \in \mathscr{I}, \tag{44}$$

$$\Psi_i(t) = X_i(t) - Y_i(t), \quad i \in \mathscr{I}, \tag{45}$$

$$\sum_{i \in \mathscr{I}} \Psi_i(t) = -Z(t), \tag{46}$$

$$Y_i(t) \geq 0, \quad i \in \mathscr{I}; \qquad Z(t) \geq 0. \tag{47}$$

The following result shows that the model $\mathscr{M}$ is equivalent to $\mathscr{M}^{\mathrm{sp}}$. Although $\mathscr{M}^{\mathrm{sp}}$ is an I-dimensional model, it is seen to be simpler than the (I-dimensional) model $\mathscr{M}$ (compare (2)–(6) to (44)–(47)). The model $\mathscr{M}^{\mathrm{sp}}$ was analyzed in Atar et al. [4].

Proposition 4.1. (i) *For every process $(X, Y, Z, \Psi) \in \mathscr{M}$ there exists a process $(\widetilde{X}, \widetilde{Y}, \widetilde{Z}, \widetilde{\Psi}) \in \mathscr{M}^{\mathrm{sp}}$, satisfying $\mathbb{P}$-a.s., for all $t \geq 0$,*

$$\widetilde{X}(t) = X(t), \qquad \widetilde{Y}(t) = Y(t), \qquad \widetilde{Z}(t) = Z_e(t) \tag{48}$$

and

$$\widetilde{\Psi}_i(t) = \sum_{j \in \mathscr{J}} \Psi_{ij}(t), \quad i \in \mathscr{I}.$$

(ii) *For every process $(\widetilde{X}, \widetilde{Y}, \widetilde{Z}, \widetilde{\Psi}) \in \mathscr{M}^{\mathrm{sp}}$ there exists $(X, Y, Z, \Psi) \in \mathscr{M}$, such that $\mathbb{P}$-a.s., for all $t \geq 0$, the Equations (48), $Z(t) = \widetilde{Z}(t)e_1$, and*

$$\Psi(t) = G(\widetilde{X}(t) - \widetilde{Y}(t), -\widetilde{Z}(t)e_1),$$

hold.

Proof. Part (i) follows directly from (2)–(6) and (44)–(47). Part (ii) is immediate from the same two sets of equations along with the definition of the map G. □

The question of optimality of JWC policies appears to be harder under the current setting than under that of the previous section.

Conjecture 4.1. *Let $\mathscr{M}^{\mathrm{sp,jwc}}$ denote the set of all $\mathbf{M} = (X, Y, Z, \Psi) \in \mathscr{M}^{\mathrm{sp}}$ satisfying the JWC condition*

$$Y_e(t) \wedge Z(t) = 0.$$

Then $\mathscr{M}^{\mathrm{sp,jwc}}$ is a reduction of $\mathscr{M}^{\mathrm{sp}}$.

Intuitively, given an arbitrary $\mathbf{M}$ one should be able to modify it so that the JWC condition holds, by allocating idle servers to classes where the queue length is positive, and since $\theta_{\max} \leq \mu_{\min}$, the cost would only improve. However, one must take into account how this modification affects the future of these processes, and we were unable to handle this difficulty. This open problem seems to be related to open problems of Atar et al. [4, §5.1].

We comment that if the above conjecture holds true then Proposition 4.1 can be used to lift the result to the model $\mathscr{M}^{\mathrm{jwc}}$, namely, the set of all $\mathbf{M} = (X, Y, Z, \Psi) \in \mathscr{M}$ satisfying the JWC condition

$$Y_e(t) \wedge Z_e(t) = 0$$

is a reduction of $\mathscr{M}^{\mathrm{sp}}$.

5. Pool dependent service rates: Asymptotic optimality. In this section we consider a queueing model for which the corresponding diffusion model satisfies the assumptions of §3. The notation of that section and particularly of Example 3.4 will be used throughout. As in Atar [3], we will assume throughout the section that $\mathscr{E} = \mathscr{E}_{\mathrm{ba}}$, and thus, that the graph $\mathscr{E}$ is a tree. We begin by describing the model and its parametrization, and formulate a routing-scheduling control policy based on the HJB ODE (26). We then state and prove that this policy is asymptotically optimal in the scaling limit. The proof relies on the methods developed in Atar [3], and on the reduction of the diffusion model to a JWC model (Proposition 3.1).

We describe the probabilistic queueing model. The processes and parameters will be indexed by $n \in \mathbb{N}$. For $j \in \mathscr{J}$, we denote by N_j^n the number of servers at pool j. We denote by $X_i^n(t)$ the total number of class-i customers in the system at time t, by $Y_i^n(t)$ the number of class-i customers in the queue at time t, and by $Z_j^n(t)$ the number of idle servers in pool j at time t. We also denote by $\Psi_{ij}^n(t)$ the number of class-i customers in pool j at time t. Let $X^n = (X_i^n)_{i\in\mathscr{I}}$, $Y^n = (Y_i^n)_{i\in\mathscr{I}}$, $Z^n = (Z_j^n)_{j\in\mathscr{J}}$, $\Psi^n = (\Psi_{ij}^n)_{i\in\mathscr{I},\, i\in\mathscr{J}}$.

To define arrival processes let, for each $i \in \mathscr{I}$, $\{\check{U}_i(k), k \in \mathbb{N}\}$ be a sequence of strictly positive i.i.d. random variables with $E[\check{U}_i(1)] = 1$ and $\operatorname{Var}(\check{U}_i(1)) = C_i^2 \in [0, \infty)$. Assume also that the sequences are independent. Let $U_i^n(k) = \check{U}_i(k)/\lambda_i^n$, where $\lambda_i^n > 0$. With $\sum_1^0 = 0$, define

$$A_i^n(t) = \sup\left\{l \geq 0\colon \sum_{k=1}^{l} U_i^n(k) \leq t\right\}, \quad t \geq 0. \tag{49}$$

It is assumed that the number of arrivals up to time t is $A_i^n(t)$.

To model service times as exponential independent random variables, let S_{ij}^n, $i \in \mathscr{I}$, $j \in \mathscr{J}$, be Poisson processes with rate $\mu_{ij}^n \in [0, \infty)$ (where a zero rate Poisson process is the zero process). These processes are assumed to be mutually independent, and independent of the arrival processes. Let $T_{ij}^n(t)$ denote the time up to t devoted to a class-i customer by a server, summed over all type-j servers; note that $T_{ij}^n(t) = \int_0^t \Psi_{ij}^n(s)\,ds$, $i \in \mathscr{I}$, $j \in \mathscr{J}$, $t \geq 0$. The number of service completions of class-i customers by all type-j servers up to time t is given as $S_{ij}^n(T_{ij}^n(t))$. Similarly, let $R_i^n(t)$ be Poisson processes of rate $\theta_i^n \in [0, \infty)$ and let $\mathring{T}_i^n(t)$denote the time up to t that a class-i customer spends in the queue, summed over all customers. Then $\mathring{T}_i^n(t) = \int_0^t Y_i^n(s)\,ds$, and we assume that $R_i^n(\mathring{T}_i^n(t))$ class-i customers have abandoned up to time t. With initial conditions $X_i^{0,n} := X_i^n(0)$, assumed to be deterministic, we have for $i \in \mathscr{I}$, $t \geq 0$,

$$X_i^n(t) = X_i^{0,n} + A_i^n(t) - \sum_j S_{ij}^n\left(\int_0^t \Psi_{ij}^n(s)\,ds\right) - R_i^n\left(\int_0^t Y_i^n(s)\,ds\right). \tag{50}$$

Routing and scheduling decisions are made by continuously selecting Ψ^n, subject to appropriate constraints. The policy is regarded as *preemptive* if service to a customer can be stopped and resumed at a later time, possibly in a different pool. In other words, customers can be moved instantaneously not only between a service pool and the buffer, but also between different service pools that are capable of offering service to the corresponding class. For consistency with Atar et al. [4], we abbreviate the term Preemptive Routing/Scheduling Control Policy as P-SCP. A policy is regarded *nonpreemptive* (abbreviated N-SCP) if every customer completes service with the server it is first assigned to. We collectively refer to P-SCPs and N-SCPs as *Routing and Scheduling Control Policies* (RSCPs).

Let $\zeta = (X^{0,n}; n \in \mathbb{N})$ and $p = (\Psi^n, n \in \mathbb{N})$ denote a sequence of initial conditions and SCPs, respectively. We denote by P_ζ^p the measure under which, for each n, X^n is the controlled process associated with $X^{0,n}$ and Ψ^n. Expectation under P_ζ^p is denoted by E_ζ^p.

We need a notion of SCPs that do not anticipate the future. Denote

$$\tau_i^n(t) = \inf\{u \geq t\colon A_i^n(u) - A_i^n(u-) > 0\}, \quad i \in \mathscr{I},$$

$$\mathscr{F}_t^n = \sigma\{A_i^n(s), S_{ij}^n(T_{ij}^n(s)), R_i^n(\mathring{T}_i^n(s)), \Psi_{ij}^n(s), X_i^n(s), Y_i(s), Z_j(s)\colon i \in \mathscr{I}, j \in \mathscr{J}, s \leq t\},$$

$$\mathscr{G}_t^n = \sigma\{A_i^n(\tau_i^n(t)+u) - A_i^n(\tau_i^n(t)), S_{ij}^n(T_{ij}^n(t)+u) - S_{ij}^n(T_{ij}^n(t)), R_i^n(\mathring{T}_i^n(t)+u) - R_i^n(\mathring{T}_i^n(t))\colon i\in\mathscr{I}, j\in\mathscr{J}, u\geq 0\}.$$

We say that a scheduling control policy is *admissible* if Atar et al. [4], Atar [3]

- for every t, $\mathscr{F}_t^n$ is independent of $\mathscr{G}_t^n$;
- for every i, j, and t, the process $S_{ij}^n(T_{ij}^n(t) + \cdot) - S_{ij}^n(T_{ij}^n(t))$ [resp., $R_i^n(\mathring{T}_i^n(t) + \cdot) - R_i^n(\mathring{T}_i^n(t))$] is equal in law to $S_{ij}^n(\cdot)$ [$R_i^n(\cdot)$].

Recall that we write $i \sim j$ if i and j are neighbors in $\mathcal{E}$. By assumption,

$$\Psi_{ij}^n = 0, \quad i \not\sim j. \tag{51}$$

We assume that there are constants $\lambda_i > 0$, $\nu_j > 0$, $\theta_i \geq 0$, and $\mu_j > 0$, such that, as $n \to \infty$, $n^{-1}\lambda_i^n \to \lambda_i$, $n^{-1}N_j^n \to \nu_j$, $j \in \mathcal{J}$; furthermore,

$$\theta_i^n = \theta_i, \quad i \in \mathcal{I}, \qquad \mu_{ij}^n = \mu_j, \quad i \sim j, \ \ n \in \mathbb{N}. \tag{52}$$

We set $\mu_{ij} = 0$ for $i \not\sim j$. The assumption that μ_{ij}^n depend only on j, imposed above, will enable us to use the results of §3. (We could have allowed for an $O(n^{-1/2})$, (i, j)-dependent term added to μ_{ij}^n, but this would have slightly complicated the proof, which we try to keep simple.) We also denote $\bar{\mu}_j = \nu_j \mu_j$, $j \in \mathcal{J}$. We further assume

$$\hat{\lambda}_i^n := n^{1/2}(n^{-1}\lambda_i^n - \lambda_i) \to \hat{\lambda}_i, \quad n^{1/2}(n^{-1}N_j^n - \nu_j) \to 0,$$

where $\hat{\lambda}_i \in \mathbb{R}$ are constants. To state a principal assumption on the limit parameters, that indicates that the system is critically loaded, consider the linear program of *minimizing* $\rho \in \mathbb{R}$ *subject to*

$$\sum_{j \in \mathcal{J}} \bar{\mu}_j \xi_{ij} = \lambda_i, \quad i \in \mathcal{I}; \qquad \sum_{i \in \mathcal{I}} \xi_{ij} \leq \rho, \quad j \in \mathcal{J}; \qquad \xi_{ij} \geq 0, \quad i \in \mathcal{I}, \ j \in \mathcal{J}. \tag{53}$$

The *heavy traffic* and *complete resource pooling* conditions (Harrison and López [14]) will be assumed, namely

- There exists a unique optimal solution (ξ^*, ρ^*) to (53).
- $\sum_{i \in \mathcal{I}} \xi_{ij}^* = 1$ for all $j \in \mathcal{J}$ (and, consequently, $\rho^* = 1$).
- $\xi_{ij}^* > 0$, $i \sim j$.

In what follows, ξ_{ij}^* will denote the quantities from the above condition, and $x^* = (x_i^*)$, $\psi^* = (\psi_{ij}^*)$, where $x_i^* = \sum_j \xi_{ij}^* \nu_j$, $\psi_{ij}^* = \xi_{ij}^* \nu_j$. We refer to the quantities ξ_{ij}^*, x_i^*, and ψ_{ij}^* as *the static fluid model* (Atar et al. [4], Atar [2]).

The initial conditions are assumed to satisfy the following. There are constants x_i, y_i, z_j, ψ_{ij} satisfying $y_i + \sum_j \psi_{ij} = x_i$, $z_j + \sum_i \psi_{ij} = 0$, $y_i \geq 0$, and $z_j \geq 0$, $i \in \mathcal{I}$, $j \in \mathcal{J}$, such that the initial conditions satisfy

$$\widehat{X}_i^{0,n} := n^{-1/2}(X_i^{0,n} - nx_i^*) \to x_i, \qquad \widehat{Y}_i^{0,n} := n^{-1/2} Y_i^{0,n} \to y_i, \tag{54}$$

$$\widehat{Z}_j^{0,n} := n^{-1/2} Z_j^{0,n} \to z_j, \qquad \widehat{\Psi}_{ij}^{0,n} := n^{-1/2}(\Psi_{ij}^{0,n} - \psi_{ij}^* n) \to \psi_{ij}. \tag{55}$$

The processes are centered around the static fluid model and rescaled so that they exhibit diffusive fluctuations, as follows:

$$\hat{A}_i^n(t) = n^{-1/2}(A_i^n(t) - \lambda_i^n t), \qquad \widehat{S}_{ij}^n(t) = n^{-1/2}(S_{ij}^n(nt) - n\mu_{ij}^n t), \tag{56}$$

$$\widehat{R}_i^n(t) = n^{-1/2}(R_i^n(nt) - n\theta_i^n t).$$

$$\widehat{X}_i^n(t) = n^{-1/2}(X_i^n(t) - nx_i^*), \tag{57}$$

$$\widehat{Y}_i^n(t) = n^{-1/2} Y_i^n(t), \qquad \widehat{Z}_j^n(t) = n^{-1/2} Z_j^n(t), \tag{58}$$

$$\widehat{\Psi}_{ij}^n(t) = n^{-1/2}(\Psi_{ij}^n(t) - \psi_{ij}^* n). \tag{59}$$

Our goal is to show that, given $c_i > 0$, $i \in \mathcal{I}$, a sequence of admissible N-SCPs can be constructed that, in an *asymptotic sense*, performs optimally with respect to the queue-length cost with weights c_i, in the class of all sequences of admissible SCPs, and that the optimal cost, properly scaled, converges to that of the DC problem from Example 3.4 (see also Remark 5.1 below about delay cost). The proposed sequence of N-SCP is based on Equation (26) from that example, and is defined via the functions h_k (28). The Equations (24)–(26) are considered with the constants ℓ and ϱ given by

$$\ell = \sum_i \hat{\lambda}_i, \quad \varrho^2 = \sum_i (\lambda_i C_i^2 + \lambda_i).$$

Let a subsequence $k(n)$ be given. Recall from (13) the definition of G, and let

$$\check{Y}^n(t) = \widehat{X}_e^n(t)^+ h_{k(n)}(\widehat{X}_e^n(t)), \qquad \check{Z}^n(t) = \widehat{X}_e^n(t)^- e_{j_0} \tag{60}$$

$$\check{\Psi}^n(t) = G(\widehat{X}^n(t) - \check{Y}^n(t), -\check{Z}^n(t)). \tag{61}$$

The policy attempts to keep $\widehat{\Psi}^n$ close to $\check{\Psi}^n$, and it does so by blocking activities for which $\widehat{\Psi}^n_{ij}(t) > \check{\Psi}^n_{ij}(t)$. That is, given an activity (i, j) and a time interval $[s, t)$, if $\widehat{\Psi}^n_{ij} > \check{\Psi}^n_{ij}$ holds on $[s, t)$ then no routings take place on the activity (i, j) throughout this interval. When a class-i customer is in the queue and there are pools $j \sim i$ with idle servers, and (i, j) is not blocked, the customer is instantaneously routed to the pool with the lowest index j among these pools. If there are no such pools, the customer stays in the queue. For a technical reason, that has to do with estimating the large time behavior (in Lemma 5.3 below), we modify the definition of $\check{Y}^n$ (60) for large values of t, by replacing $h_{k(n)}$ by e_{i_0}, where $i_0 \in \mathcal{I}$ is arbitrary and fixed, for all $t \ge \Theta_n$, where Θ_n is a random time (such that $\Theta_n \to \infty$ in probability) which we now define. (The processes $\check{Z}^n$ and $\check{\Psi}^n$ are still defined via (60) and (61).) For $x \in \mathbb{R}^I$, let $F^*(x) = G(x - x_e^+ h_{k(n)}(x), -x_e^- e_{j_0})$, let b_0 denote the constant $2 + \max_{i\sim j} \sup_n (\widehat{\Psi}^n_{ij}(0) - \check{\Psi}^n_{ij}(0))^+$, and define

$$\Theta_n = \inf\Bigl\{t\colon \max_{i\sim j}[\widehat{\Psi}^n_{ij}(t-) - F^*_{ij}(\widehat{X}^n(t-))] \ge b_0\Bigr\}.$$

The resulting sequence of admissible N-SCPs will be denoted by $p^*(\{k(n)\})$.

For a sequence ζ of initial conditions and a sequence p of SCPs, denote

$$\underline{V}(\zeta, p) = \liminf_{n\to\infty} E^p_\zeta\biggl[\int_0^\infty e^{-t} c \cdot \widehat{Y}^n(t)\, dt\biggr],$$

$$\overline{V}(\zeta, p) = \limsup_{n\to\infty} E^p_\zeta\biggl[\int_0^\infty e^{-t} c \cdot \widehat{Y}^n(t)\, dt\biggr].$$

Recall from Example 3.4 that V is given as the value function of the 1-dimensional DC problem (23), or equivalently as the unique solution to the ODE (26)–(27). The main result of this section relates $\underline{V}$ and $\overline{V}$ to V.

Theorem 5.1. *Let ζ be a sequence of initial conditions $\{X^{0,n}; n \in \mathbb{N}\}$ such that*

$$\widehat{X}^{0,n} = n^{-1/2}(X^{0,n} - nx^*) \to x \in \mathbb{R}^I.$$

Recall that $x_e = \sum_{i\in\mathcal{I}} x_i$. Then we have the following.

(i) *For any sequence p of admissible SCPs, one has $\underline{V}(\zeta, p) \ge V(x_e)$.*

(ii) *Provided that $E[\check{U}_i(1)^K] < \infty$, $i \in \mathcal{I}$, where K is a finite constant which depends only on the deterministic parameters introduced earlier in this section, there exists a sequence $\{k(n)\}$ such that, with $p^* = p^*(\{k(n)\})$, the sequence p^* of admissible N-SCPs satisfies $\overline{V}(\zeta, p^*) \le V(x_e)$.*

Remark 5.1 (Delay Costs). As observed in Atar et al. [4], the treatment of queue-length costs covers also delay-related costs. More precisely, for each of the customers l ever present in the system, let $\mathrm{cl}(l)$ denote the class to which l belongs, and let $\nu(l)$ denote the set of times at which customer l is in the queue. Clearly, $\widehat{Y}^n_i(t) = n^{-1/2} \sum 1_{\{t\in\nu(l)\}}$, where the sum extends over all class-i customers l. Then the cost

$$n^{-1/2} E\biggl[\sum_l c_{\mathrm{cl}(l)} \int_{\nu(l)} e^{-t}\, dt\biggr]$$

equals

$$E\biggl[\int_0^\infty e^{-t}\, c \cdot \widehat{Y}^n(t)\, dt\biggr].$$

The latter is a cost of the form treated in this paper. □

The two main ingredients of the proof are tightness of the diffusively scaled processes, and large time estimates under the policy p^*. The tightness and some related properties are stated in Lemma 5.1, and follow entirely from Atar [3]. On the other hand, the large time estimates from Atar [3] on $\widehat{X}^n(t)$, that are subexponential in t, are not valid in the current setting. Subexponential estimates on the 1-dimensional process $\widehat{X}^n_e(t)$, that are rather easy (cf. Lemma A.1(ii) in the appendix) are not sufficient. (Note that this stands in contrast to the situation dealt with in §3 where the estimate on the 1-dimensional controlled diffusion $\check{X}$, namely Lemma A.1(i), suffices for the analysis of Equation (26) and the polynomial bound (27).) The reason for this is technical, and has to do with a crucial estimate on $P(\vartheta_n < t)$, where ϑ_n is a random time defined by (71). While bounds on $\widehat{X}^n(t)$ that are uniform in n and polynomial in t imply bounds on $P(\vartheta_n < t)$ of the form $o_n(1)(1+t)^c$, that is not the case with bounds on $\widehat{X}^n_e(t)$. This issue is dealt with by an argument that combines worse bounds on $P(\vartheta_n < t)$, namely (72), and the polynomial estimates on just $\widehat{X}^n_e(t)$ (Lemma 5.3).

Finally, the results of Atar [3] imply only that admissible SCPs that are JWC (as defined later in this section) satisfy the conclusion of Theorem 5.1(i). The extension to general admissible SCPs is made possible via the results of §3, that establish a reduction of the general diffusion model to one that satisfies a JWC condition. The argument is carried out at the end of this section (see the proof of Theorem 5.1).

We fix some notation. Given a positive integer d, denote by $\mathbb{D}$ the space of functions from $\mathbb{R}_+$ to $\mathbb{R}^d$ that are right continuous on $\mathbb{R}_+$ and have finite left limits on $(0, \infty)$ (RCLL), endowed with the Skorohod J_1 topology. If X^n, $n \in \mathbb{N}$ and X are processes with sample paths in $\mathbb{D}$ (respectively, real-valued random variables) we write $X^n \Rightarrow X$ to denote weak convergence of the measures induced by X^n on $\mathbb{D}$ (respectively, on $\mathbb{R}$) to the measure induced by X, as $n \to \infty$. For $X \in \mathbb{D}$, we write $\|X\|_t^* := \sup_{0 \le s \le t} \|X(s)\|$, and, if $d = 1$, we write $\|X\|_t^*$ as $|X|_t^*$.

The fluid scale processes are defined as

$$\bar{X}_i^n(t) = n^{-1} X_i^n(t), \qquad \bar{Y}_i^n(t) = n^{-1} Y_i^n(t),$$
$$\bar{Z}_j^n(t) = n^{-1} Z_j^n(t), \qquad \bar{\Psi}_{ij}^n(t) = n^{-1} \Psi_i^n(t).$$

Denote $r_i = (\lambda_i C_i^2 + \lambda_i)^{1/2}$. Let

$$r_i \widehat{W}_i^n(t) = \hat{A}_i^n(t) - \sum_j \widehat{S}_{ij}^n\left(\int_0^t \bar{\Psi}_{ij}^n(s)\, ds\right) - \widehat{R}_i^n\left(\int_0^t \bar{Y}_i^n(s)\, ds\right), \tag{62}$$

$$\widetilde{W}^n(t) = r_i \widehat{W}_i^n(t) + \hat{\lambda}_i^n t, \tag{63}$$

$$\widehat{M}^n(t) = \widehat{Y}_e^n(t) \wedge \widehat{Z}_e^n(t) \ge 0, \tag{64}$$

and let (u^n, v^n) take values in $\mathbb{U} \times \mathbb{V}$ and be defined via the equations

$$Y^n = Y_e^n u^n, \qquad Z^n = Z_e^n v^n. \tag{65}$$

With this notation, it is shown in Atar [3, Equations (27)–(30)] that

$$\widehat{Y}_i^n + \sum_j \widehat{\Psi}_{ij}^n = \widehat{X}_i^n, \quad i \in \mathcal{I}, \qquad \widehat{Z}_j^n + \sum_i \widehat{\Psi}_{ij}^n = 0, \quad j \in \mathcal{J}, \tag{66}$$

$$\widehat{X}_i^n(t) = \widehat{X}_i^{0,n} + \widetilde{W}_i^n(t) - \sum_j \mu_{ij} \int_0^t \widehat{\Psi}_{ij}^n(s)\, ds - \theta_i \int_0^t \widehat{Y}_i^n(s)\, ds. \tag{67}$$

Recall the notation $\tilde{\beta}$ (29) from Remark 3.1(ii). It follows from equation Atar [3, (65)] along with the assumption (52), that the service rates are pool dependent, that

$$\widehat{X}_e^n(t) = \widehat{X}_e^n(0) + \widetilde{W}_e^n(t) + \int_0^t [\tilde{\beta}(\widehat{X}_e^n(s), u^n(s), v^n(s)) + \widehat{M}^n(s)(\mu \cdot v^n(s) - \theta \cdot u^n(s))]\, ds. \tag{68}$$

We have $X_e^n - Y_e^n = N_e^n - Z_e^n$ by the definition of these processes. Consequently, the diffusion scaled processes satisfy $\widehat{X}_e^n - \widehat{Y}_e^n = -\widehat{Z}_e^n$. Along with (64), this implies

$$\widehat{Y}_e^n = \widehat{M}^n + (\widehat{X}_e^n)^+. \tag{69}$$

Next, the proof of inequality Atar [3, (134)] is valid here. It states that

$$E[(\|\widehat{W}^n\|_t^*)^K] \le a_1(1+t)^{a_1}, \tag{70}$$

where K is the constant from the statement of Theorem 5.1, and a_1 is a constant that does not depend on t or n.

Define $\Lambda_{ij}^n = \widehat{\Psi}_{ij}^n - \check{\Psi}_{ij}^n$ and

$$\vartheta_n = \inf\left\{t\colon \max_{i\sim j} \Lambda_{ij}^n \ge b_1\right\}, \tag{71}$$

where $b_1 = (2b_0) \vee 13$. It is shown in the proof of Atar [3, Proposition 2, part (ii), p. 2647] that one has the estimate $\|\widehat{X}^n\|_t^* \le a_2 e^{a_2 t}(1 + \|\widehat{W}^n\|_t^* + |\widehat{M}|_t^*)$, holding for all t and n, where a_2 is a constant. Using this estimate, in place of the polynomial estimate on $\widehat{X}^n$ used to prove Atar [3, (140)], one obtains

$$P(\vartheta_n \le t) \le a_3 n^{1/4 - K/8} e^{\gamma t}, \tag{72}$$

where a_3 and γ are constants not depending on n or t.

We now introduce formally the notion of JWC policies. Fix n. Let $\mathscr{X}^n$ denote the set of all possible values of $X^n(t)$ (for fixed t) for which there is a rearrangement of customers with the property: *Either there are no customers in queue, or no server in the system is idle*, i.e.,

$$Y_e^n(t) \wedge Z_e^n(t) = 0. \tag{73}$$

A work conserving policy is said to be JWC if for every t, $X^n(t) \in \mathscr{X}^n$ implies (73). A useful property of the set $\mathscr{X}^n$ (required for the proof of Lemma 5.1 below), proved in Atar [3, Lemma 3], is that there exists a constant $\alpha_0 > 0$ such that

$$\|X^n(t) - nx^*\| \le \alpha_0 n \quad \text{implies} \quad X^n(t) \in \mathscr{X}^n. \tag{74}$$

More details on JWC policies can be found in Atar [3].

Finally, the proof of Lemma 5.2 below is based on properties of the process

$$\mathbf{K}^n(t) = \tilde{\beta}\big(\widehat{X}_e^n(t), u^n(t), v^n(t)\big)V'(\widehat{X}_e^n(t)) + c \cdot u^n(t)\widehat{X}_e^n(t) - \widetilde{\mathbf{H}}\big(\widehat{X}_e^n(t), V'(\widehat{X}_e^n(t))\big). \tag{75}$$

Note that by the definition of $\widetilde{\mathbf{H}}$ (30) we have that $\mathbf{K}^n \ge 0$.

The proof of Theorem 5.1 requires the following three lemmas.

Lemma 5.1. *Let the assumptions of Theorem* 5.1 *hold. Let*

$$J^n = \|\widehat{Y}^n - \check{Y}^n\| + \|\widehat{Z}^n - \check{Z}^n\|$$

and

$$Q_1^n = \int_0^{\cdot} \tilde{\beta}(\widehat{X}^n(s), u^n(s), v^n(s))\,ds, \qquad Q_2^n = \int_0^{\cdot} e^{-s} c \cdot \widehat{Y}^n(s)\,ds.$$

Then items (i)–(iii) *below hold under any admissible JWC SCP and under the N-SCP p^*, namely*

(i) $(\overline{X}^n, \overline{Y}^n, \overline{Z}^n, \overline{\Psi}^n) \Rightarrow (x^*, 0, 0, \psi^*)$;

(ii) $(\widehat{W}^n, \int_0^{\cdot} \widehat{M}^n, \widehat{X}^n, Q_1^n, Q_2^n)$ *are tight*;

(iii) $(\widehat{W}^n, \int_0^{\cdot} \widehat{M}^n) \Rightarrow (W, 0)$, *where W is a standard 1-dimensional Brownian motion.*

In addition,

(iv) *Under p^* one has* $\sup_{[s,t]}(J^n \vee \widehat{M}^n) \Rightarrow 0$, *for every* $0 < s < t < \infty$;

(v) *Let p be any admissible SCP* (*that is not necessarily JWC*). *Then, along any subsequence on which the cost*

$$E_\zeta^p\left[\int_0^\infty e^{-t} c \cdot \widehat{Y}^n(t)\,dt\right]$$

is bounded, we have that item (i) *above holds, $\widehat{W}^n \Rightarrow W$, and the processes* $(\widehat{W}^n, \widehat{X}^n, \int_0^{\cdot} \widehat{M}^n)$ *are tight.*

Proof. For items (i)–(iv), the proof of Proposition 1 of Atar [3] holds verbatim. To prove item (v), note that, since $c_i > 0$ for all i, and since by (64) $\widehat{M}^n \le \widehat{Y}_e^n$, we have that $\int_0^\infty e^{-t} E[\widehat{M}^n(t)]dt \le a_1$, where a_1 is a constant independent of n. Hence for any fixed T, $\int_0^T \widehat{M}^n$ are tight. Reviewing the proof of Proposition 1 of Atar [3], with $\tilde{\tau}_{n,k} := \inf\{t\colon \int_0^t \widehat{M}^n \ge k\}$ in place of τ_n, and using boundedness of u^n and v^n, one obtains that the processes $\Xi^n := (\overline{X}^n, \overline{Y}^n, \overline{Z}^n, \overline{\Psi}^n, \widehat{W}^n, \widehat{X}^n, \int_0^{\cdot} \widehat{M}^n, \int_0^{\cdot} u_i^n, \int_0^{\cdot} v_j^n)$, *when stopped at* $\tilde{\tau}_{n,k}$ are tight in n for each fixed k, on obvious modifications of the proof of that proposition. Because, as mentioned above, the random variables $\int_0^T \widehat{M}^n$ are tight in n (for fixed T), we have that the processes Ξ^n themselves are tight. The limit results (namely item (i) and $\widehat{W}^n \Rightarrow W$) follow precisely as in Atar [3]. □

Lemma 5.2. *Under the policy p^*, with an appropriate choice of the sequence* $\{k(n)\}$,

$$\limsup_{n\to\infty} E_\zeta^{p^*}\left[\int_0^{T\wedge\vartheta_n} e^{-t} c \cdot \widehat{Y}^n(t)\,dt\right] \le V(x_e). \tag{76}$$

Proof. We refer to Lemma 5.1 and denote by (W, X, Q_1, Q_2) a subsequential limit of $(\widehat{W}^n, \widehat{X}^n, Q_1^n, Q_2^n)$ under p^*.

For the policy $p^* = p^*(\{k(n)\})$, with an appropriate choice of $\{k(n)\}$, one has $\int_0^t e^{-s}\mathbf{K}^n(s)\,ds \to 0$ in probability, for every t, by the proof of Theorem 2 of Atar [3], pp. 2633–2636.

We follow the proof of Atar et al. [4, Theorem 4]. Using the smoothness of the function V (cf. Example 3.4), Ito's Lemma, Lemma 5.1, and the convergence of $\int e^{-s}\mathbf{K}^n(s)\,ds \to 0$ we obtain, along the lines of the proof of Atar et al. [4, Theorem 4, part (i)],

$$e^{-t}V(X_e(t)) = V(x_e) + \int_0^t e^{-s}V'(X_e(s))\,dW(s) - Q_2(t). \tag{77}$$

As is Atar et al. [4, Lemma 6], X is adapted to a filtration on which W is a martingale. Thus $E[Q_2(t)] \le V(x_e)$. Since $Q_2^n \Rightarrow Q_2$ and $Q_2^n(t \wedge \vartheta_n) \le Q_2^n(t)$, to prove (76) it suffices to show that, for every t, $Q_2^n(t \wedge \vartheta_n)$ are uniformly integrable. It is shown in Atar [3, the second display below (138)] that, under this policy,

$$\widehat{M}^n \le a_1 \max_{i\sim j} (\Lambda_{ij}^n)^+, \tag{78}$$

where a_1 is a constant. Fix t. It follows from the above plus (63), (68), (70), (71) and Gronwall's Lemma, that

$$E[(|\widehat{X}_e^n|_{t\wedge\vartheta_n}^*)^K] \le a,$$

where $a = a(t) < \infty$ does not depend on n. Since by (69), $c \cdot \widehat{Y}^n \le \|c\| \widehat{Y}_e^n = \|c\|((\widehat{X}_e^n)^+ + \widehat{M}^n)$ and by (71), (78) $\int_0^{t\wedge\vartheta_n} \widehat{M}^n(s)\, ds \le b_1 t$, we obtain

$$E\biggl[\biggl(\int_0^{t\wedge\vartheta_n} c \cdot \widehat{Y}^n(s)\, ds\biggr)^K\biggr] \le a_2,$$

where the constant $a_2 < \infty$ does not depend on n. Assuming without loss that $K > 1$, the uniform integrability alluded to above follows, and so does (76). □

Lemma 5.3. *Under the hypotheses of Theorem* 5.1,

$$\lim_{T\to\infty} \limsup_{n\to\infty} E_\zeta^{p^*}\biggl[\int_{T\wedge\vartheta_n}^\infty e^{-t} \|\widehat{Y}^n(t)\|\, dt\biggr] = 0. \tag{79}$$

Proof. By (71) and (78), we have, for $t \le \vartheta_n$, that $\widehat{M}^n(t) \le a_1$, where a_1 is a constant that does not depend on t or n. Hence by (69), for $t \le \vartheta_n$,

$$\|\widehat{Y}^n(t)\| = \widehat{Y}_e^n(t) \le a_1 + |\widehat{X}_e^n(t)| \le a_2(1+t) + 2|\widetilde{W}^n|_t^*,$$

for some finite constant a_2, where we used Lemma A.1(ii) in the last inequality. As a result,

$$\begin{aligned} E_\zeta^{p^*}\biggl[\int_{T\wedge\vartheta_n}^{\vartheta_n} e^{-t}\|\widehat{Y}^n(t)\| dt\biggr] &\le E_\zeta^{p^*}\biggl[\int_T^\infty e^{-t}(a_2(1+t) + 2\|\widetilde{W}^n\|_t^*)\, dt\biggr] \\ &\le \int_T^\infty e^{-t} a_3 (1+t)^{a_3}\, dt \end{aligned} \tag{80}$$

where the last inequality follows (63) and (70), and a_3 denotes a constant.

Next, for $t > \vartheta_n$ we do not have a good estimate on $\widehat{Y}^n$, and we use a crude bound. We bound $\|Y^n(t)\|$ by the initial number of customers in the system plus the total number of arrivals up to time t. This gives

$$\|Y^n(t)\| \le a_4 n + \|A^n(t)\|,$$

and thus

$$E[\|\widehat{Y}^n(t)\|^2] = n^{-1} E[\|Y^n(t)\|^2] \le a_5 n (1+t)^2. \tag{81}$$

By Cauchy-Schwartz,

$$E_\zeta^{p^*}\biggl[\int_{\vartheta_n}^\infty e^{-t}\|\widehat{Y}^n(t)\| dt\biggr] \le a_5 n^{1/2} \int_0^\infty e^{-t} P(t > \vartheta_n)^{1/2} (1+t)\, dt.$$

By (72), this is bounded above by

$$a_5 n^{1/2} \int_0^\infty e^{-t}[1 \wedge (n^{-\beta} e^{\gamma t/2})](1+t)\, dt,$$

where $\beta = K/16 - 1/8$. It is a straightforward calculation to show that the expression in the last display tends to zero as $n \to \infty$, provided that the constant β is sufficiently large. This can be ensured by selecting K to be sufficiently large. Along with (80) this establishes (79). □

Proof of Theorem 5.1. (i) For an integrable function f let $If = \int_0^{\cdot} f$. This notation is used in this proof only, and there will be no confusion with the parameter representing number of classes. We will use Lemma 5.1(v). Fix a subsequence on which the cost alluded to in that lemma is bounded (there is nothing to prove in case $\underline{V}(p, \zeta) = +\infty$). Denote by (W, X, M^I) a weak limit of $(\widehat{W}^n, \widehat{X}^n, I\widehat{M}^n)$ along a further subsequence.

By (65) and (69), $\widehat{Y}_i^n \geq 0$ and, for $0 \leq s < t \leq T$,

$$I\widehat{Y}_i^n(t) - I\widehat{Y}_i^n(s) \leq (t-s)[\|\widehat{X}^n\|_T^* + |\widehat{M}^n|_T^*].$$

The tightness of the random variables inside the square brackets above implies that the processes $I\widehat{Y}_i^n$ are tight, and that their subsequential limits are Lipschitz. A similar argument holds for $I\widehat{Z}_j^n$. Fix a further subsequence along which $I\widehat{Y}_i^n$ and $I\widehat{Z}_j^n$ converge, and denote by Y_i^I, Z_j^I their respective limits. As in the proof of Atar et al. [4, Lemma 6], the admissibility of the policy p implies that X, Y^I, and Z^I are adapted to a filtration, denoted (F_t), on which W is a martingale.

Denote by Y_i and Z_j RCLL versions of the derivatives of Y_i^I and Z_j^I. The process (Y, Z) may, and will, be chosen progressively measurable with respect to (F_t) (as follows e.g., from Jacod and Shiryaev [15, Proposition I.3.5]). Because the limits have Lipschitz sample paths, they are given as integrals of their derivatives, and so $(I\widehat{Y}_i^n, I\widehat{Z}_j^n) \Rightarrow (IY_i, IZ_j)$.

Next, by (66) and linearity of G, $I\widehat{\Psi}^n = G(I\widehat{X}^n - I\widehat{Y}^n, -I\widehat{Z}^n)$, and therefore $I\widehat{\Psi}^n \Rightarrow I\Psi$, where $\Psi := G(X - Y, -Z)$. By (67) we conclude that

$$X_i = x_i + W_i - \sum_j \mu_{ij} I\Psi_{ij} - \theta_i IY_i.$$

Hence (2) holds. Also, (3)–(5) follow from the definition of G, and (6) from the positivity of $\widehat{Y}_i^n$, $\widehat{Z}_j^n$. With the notation of §2, this shows that $(X, Y, Z, \Psi) \in \mathscr{M}$. Hence by (21) and (22),

$$E\left[\int_0^\infty e^{-t} c \cdot Y(t)\, dt\right] \geq \widetilde{V}(x) = V(x_e).$$

Given $\delta > 0$, let $T < \infty$ be such that $E[\int_0^T e^{-t} c \cdot Y(t)\, dt] \geq V(x_e) - \delta$. By Fatou's Lemma, $\liminf E_\zeta^p[\int_0^T e^{-t} c \cdot \widehat{Y}^n(t)\, dt] \geq E[\liminf \int_0^T e^{-t} c \cdot \widehat{Y}^n(t)\, dt]$. Using integration by parts, the integral on the right hand side (r.h.s.) of the above inequality can be seen to depend continuously on $\int_0^{\cdot} c \cdot \widehat{Y}^n(t)\, dt$, and therefore the lim inf on the r.h.s. equals $\int_0^T e^{-t} c \cdot Y(t)\, dt$. This shows that $\underline{V}(\zeta, p) \geq V(x_e) - \delta$. Since $\delta > 0$ is arbitrary it can be dropped, and this completes the proof of (i).

(ii) This part is immediate from Lemmas 5.2 and 5.3. □

Appendix A.

Lemma A.1. (i) *If* (11) *holds, then*

$$|\check{X}|_t^* \leq |x_e| + 2|W_e|_t^*, \quad t \geq 0. \tag{82}$$

More generally, (82) *holds provided that* $\check{X}$ *is a continuous process satisfying*

$$\check{X}(t) - \check{X}(s) \leq W_e(t) - W_e(s) + a\int_s^t \check{X}(u)^-\, ds, \tag{83}$$

$$\check{X}(t) - \check{X}(s) \geq W_e(t) - W_e(s) - a\int_s^t \check{X}(u)^+\, ds, \tag{84}$$

for all $0 < s < t$, *a.s., where* $a > 0$ *is a finite constant.*

(ii) *Let* (68) *hold. Let* $\bar{\mu} = \max_j \mu_j$. *Then*

$$|\widehat{X}_e^n(t)| \leq |\widehat{X}_e^n(0)| + 2|\widetilde{W}^n|_t^* + 2t\bar{\mu}|\widehat{M}^n|_t^*, \quad t > 0.$$

Proof. We begin with item (i). Because (11) implies (83) and (84), it suffices to prove the second part of this item. Let y be a constant such that $y > |x_e|$. If $\max_{[0,T]} \check{X} > y$ then there exist $0 \leq s < t \leq T$ for which $\check{X}(s) = |x_e|$, $\check{X}(s') \geq 0$, $s' \in (s, t)$, and $\check{X}(t) > y$. Hence by (83),

$$\check{X}(t) - \check{X}(s) \leq W_e(t) - W_e(s),$$

and therefore $y - |x_e| \leq 2|W_e|_T^*$. If $\min_{[0,T]} \check{X} < -y$ then similarly, by (84), $y - |x_e| \leq 2|W_e|_T^*$. Item (i) follows. The argument for (ii) is similar and thus omitted. □

Acknowledgments. Research for the first and second authors was supported, in part, by Technion funds for the promotion of research and sponsored research. The second author's research was also supported, in part, by two Binational Science Foundation (BSF) Grants, 2005175 and 2008480, and Israeli Science Foundation (ISF) Grant 1357/08.

References

[1] Aksin, Z., M. Armony, V. Mehrotra. 2007. The modern call center: A multi-disciplinary perspective on operations management research. *Production Oper. Management* **16**(6) 665–688.

[2] Atar, R. 2005. A diffusion model of scheduling control in queueing systems with many servers. *Ann. Appl. Probab.* **15**(1B) 820–852.

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

[5] Atar, R., A. Mandelbaum, G. Shaikhet. 2006. Queueing systems with many servers: Null controllability in heavy traffic. *Ann. Appl. Probab.* **16**(4) 1764–1804.

[6] Birkhoff, G., G.-C. Rota. 1989. *Ordinary Differential Equations*, 4th ed. John Wiley, New York.

[7] Dai, J. G., T. Tezcan. 2009. State space collapse in many server diffusion limits of parallel server systems. Working paper, Georgia Institute of Technology, Atlanta.

[8] Dai, J. G., T. Tezcan. 2008. Optimal control of parallel server systems with many servers in heavy traffic. *Queueing Systems* **59** 95–134.

[9] Fleming, W. H., H. M. Soner. 1993. *Controlled Markov Processes and Viscosity Solutions*. Springer-Verlag, New York.

[10] Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review and research prospects. *Manufacturing Service Oper. Management* **5** 79–141.

[11] Gurvich, I., W. Whitt. 2007. Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing Service Oper. Management* **11**(2) 237–256.

[12] Gurvich, I., W. Whitt. 2009. Service-level differentiation in many-server service systems via queue-ratio routing. *Oper. Res.*, ePub ahead of print October 28, http://or.journal.informs.org/cgi/content/abstract/opre.1090.0736v1.

[13] Halfin, S., W. Whitt. 1981. Heavy traffic limits for queues with many exponential servers. *Oper. Res.* **29**(3) 567–588.

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

[15] Jacod, J., A. Shiryaev. 1987. *Limit Theorems for Stochastic Processes*. Springer-Verlag, New York.

[16] 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.

[17] Protter, P. 1990. Stochastic integration and differential equations. A new approach. *Applications of Mathematics*, Vol. 21. Springer-Verlag, Berlin/Heidelberg.

[18] Tezcan, T., J. G. Dai. 2009. Dynamic control of N-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. *Oper. Res.*, ePub ahead of print July 29, http://or.journal.informs.org/cgi/content/abstract/opre.1080.0668v1.