OPERATIONS RESEARCH
Vol. 58, No. 5, September–October 2010, pp. 1427–1439
ISSN 0030-364X | EISSN 1526-5463 | 10 | 5805 | 1427

DOI 10.1287/opre.1100.0826

The $c\mu/\theta$ Rule for Many-Server Queues with Abandonment

Rami Atar, Chanit Giat, Nahum Shimkin
Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel
{atar@ee.technion.ac.il, gchanit@tx.technion.ac.il, shimkin@ee.technion.ac.il}

We consider a multiclass queueing system with multiple homogeneous servers and customer abandonment. For each customer class i, the holding cost per unit time, the service rate, and the abandonment rate are denoted by c_i, μ_i, and θ_i, respectively. We prove that under a many-server fluid scaling and overload conditions, a server-scheduling policy that assigns priority to classes according to their index $c_i\mu_i/\theta_i$ is asymptotically optimal for minimizing the overall long-run average holding cost. An additional penalty on customer abandonment is easily incorporated into this model and leads to a similar index rule.

Subject classifications: multiclass queue; customer abandonment; fluid limits.
Area of review: Stochastic Models.
History: Received April 2009; revision received September 2009; accepted December 2009. Published online in *Articles in Advance* August 17, 2010.

1. Introduction

The usefulness of the well-known $c\mu$ *rule* for service scheduling stems from its simplicity and its robustness. This scheduling policy and its generalizations have been proved to be optimal (in a precise sense Cox and Smith 1961; Baras et al. 1985—or asymptotic sense Mandelbaum and Stolyar 2004; Van Mieghem 1995, 2003) for delay and queue-length costs, in a variety of settings. Although these settings are quite general, they do not include ones where customers may abandon (or *renege*) while waiting to be served. Customer abandonment has been widely discussed in the recent queueing literature, because it is a significant modeling aspect in applications, and particularly in call centers. (For recent developments on these applications and related models, see Aksin et al. 2007 and Gans et al. 2003.) In this paper, we introduce a scheduling rule for models that include abandonment, to which we refer as the $c\mu/\theta$ *rule*. Like the $c\mu$ rule, the $c\mu/\theta$ rule is simple on one hand and performs well on the other hand. In particular, we prove that it asymptotically minimizes the long-run average holding cost under a many-server fluid regime. A preliminary version of these results was in Atar et al. (2008).

The model considered here consists of I customer classes and a server pool with n homogeneous servers. Customers of class i arrive according to a Poisson process with rate λ_i, for $i \in \mathcal{I} := \{1, \ldots, I\}$. Customers that cannot be served immediately upon arrival are kept in an infinite-capacity queue dedicated to their class. A customer that is held in queue may lose her patience and abandon the system. Customer patience is modeled by an exponentially distributed random variable with mean $1/\theta_i$, depending on the customer class i. Once admitted to service, a class-i customer is served with exponentially distributed time duration of mean $1/\mu_i$. A stochastic queueing control problem arises by considering a long-run average holding cost, in which the holding cost per unit time for a class-i customer is a given constant c_i, and the control involves dynamic assignment of servers to waiting jobs of different classes. We study this problem in a many-server fluid regime. We note that this regime is meaningful (relative to our long-term average criterion) only under overload conditions, namely when the incoming service requirement strictly exceeds the service capacity, because otherwise all queues will be empty in steady state under any nonidling service scheme. Customer abandonment allows the queue size to stabilize even in the overloaded case.

The formal scaling limit of the problem leads to a simple linear program (LP) whose solution is shown to be a lower bound on the many-server limiting cost of the stochastic queueing problem under any policy. The main result of this paper is that a simple priority policy that assigns strict priority to classes according to the order of their indices $c_i\mu_i/\theta_i$ is asymptotically optimal, in the sense that it attains the asymptotic lower bound. We first establish this result for the preemptive-service case (§5), and then extend the proof in §6 to the nonpreemptive case, where service to customers who are already in service cannot be interrupted. We note that it is not a priori obvious here that the preemptive and nonpreemptive policies should behave similarly, because service times are not accelerated under our fluid scaling.

The cost structure described above focuses on penalizing the queue size (via the holding cost parameters c_i) and does not directly account for customer abandonment. It turns

out that penalizing customer abandonments is easily incorporated into our basic model. Specifically, assume that a penalty γ_i is incurred whenever a class-i customer abandons the queue. Then (as argued in Remark 2.1 below) the modified optimization problem becomes equivalent to the original one once the cost coefficient c_i is modified to $(c_i+\theta_i\gamma_i)$. As a consequence, the modified problem admits an optimal index rule with indices $(c_i+\theta_i\gamma_i)\mu_i/\theta_i$.

It is worthwhile to note that when $\gamma_i \equiv 0$ and the abandonment rates θ_i are independent of the user class i, our modified index reduces to the index $c_i\mu_i$ of the standard $c\mu$ rule. A similar observation holds true when the abandonment parameters γ_i are proportional to the cost parameters c_i. When abandonment costs are the major concern, eliminating the holding cost parameters c_i interestingly leads to the effective index $\gamma_i\mu_i$, which is independent of θ_i. We finally observe that significant differences may exist between the abandonment rate parameters θ_i of different customer classes—for example, when these classes correspond to different modes of service such as telephone versus e-mail reply.

We proceed to survey some related literature. It appears that the $c\mu$ rule was first suggested by Smith (1956) and Cox and Smith (1961), in a deterministic and stochastic setting, respectively. The latter treated a multiclass M/G/1 system, showing optimality of the policy with respect to holding cost. Many extensions have been established since then, including Baras et al. (1985), Hirayama et al. (1989), Klimov (1974), Nain and Towsley (1994), Walrand et al. (1985). For more details, see, e.g., the discussion in Van Mieghem (1995). A generalized version of the $c\mu$ rule was introduced by Van Mieghem (1995) for the case of nonlinear, convex holding costs. The proposed rule was shown to be asymptotically optimal in diffusion scaling under heavy traffic conditions. This work was extended by Mandelbaum and Stolyar (2004) to a more general network topology.

Approximation results for many-server systems under fluid regime were obtained by Mandelbaum et al. (1998) in a variety of network settings that include time-varying parameters (this work also treated diffusion approximations). More recent development on fluid regime approximations includes Whitt (2004), which treats the so-called efficiency-driven regime, and Whitt (2006), which suggests a fluid-scale model for the G/G/N queue with abandonment. Perry and Whitt (2009) develop fluid approximations for threshold-based control policies designed to respond to unexpected overloads. Kaspi and Ramanan (2010), Reed (2010), and Kang and Ramanan (2010) obtain general fluid approximation results for G/G/N queues.

Among recent contributions to controlled queueing models in the fluid regime, we mention the works by Bassamboo et al. (2006a, b), where a two-scale parameter regime is introduced and a linear-program based approach to dynamic routing is developed. In fact, the linear program that is the basis of the present paper's development can be obtained as a special case of the one identified in the above papers. However, the approach of these references is not based on the explicit solution of the linear program, nor does it lead to a fixed-priority rule, two central ingredients of our contribution.

The rest of this paper is organized as follows. In §2 we introduce the queueing model and asymptotic framework, and in §3 we formally derive a corresponding fluid steady-state model. In §4 the solution to a linear program associated with the fluid model is shown to be a lower bound on the limiting performance for the queueing model. Section 5 shows that the lower bound is achieved by the preemptive policy. Section 6 proves an analogous result for the nonpreemptive policy, under appropriate assumptions. Some concluding remarks appear in the final §7.

Notation. We write $\mathbb{R}_+$ for $[0,\infty)$. For $x\in\mathbb{R}^I$ let $\|x\| = \sum_{i\in\mathcal{I}}|x_i|$. For $f\colon \mathbb{R}_+\to\mathbb{R}$ let $|f|_T^* = \sup_{0\leqslant t\leqslant T}|f(t)|$, and for $f\colon \mathbb{R}_+\to\mathbb{R}^I$, $\|f\|_T^* = \sup_{0\leqslant t\leqslant T}\|f(t)\|$. We use the convention that a sum $\sum_i^j$ equals zero when $j<i$. The symbol $\mathbf{1}_A$ denotes the indicator function of a given set A.

2. Model and Asymptotic Framework

2.1. Queueing Model

The model is defined on a probability space $(\Omega, F, \mathbb{P})$. Expectation with respect to $\mathbb{P}$ is denoted by $\mathbb{E}$. The queueing system consists of a pool of n servers with identical capabilities that cater to customers of I classes. We refer to this as *the nth system*, emphasizing the dependence on the number of servers. For $i\in\mathcal{I} := \{1,\dots,I\}$, denote by $X_i^n(t)$ the total headcount of class i customers in the nth system. Denote by $Q_i^n(t)$ the queue length of class-i customers, and by $Z_i^n(t)$ the number of servers that serve customers of class-i at time t. Clearly, for every $t\geqslant 0$,

$$\sum_{i\in\mathcal{I}} Z_i^n(t) \leqslant n \tag{1}$$

$$X_i^n(t) - Z_i^n(t) = Q_i^n(t) \geqslant 0, \quad i\in\mathcal{I}. \tag{2}$$

The arrival processes are denoted by A_i^n and are assumed to be Poisson processes with rates λ_i^n, $i\in\mathcal{I}$. We use $D_i^n(t)$ and $R_i^n(t)$ to denote the number of class-i service completions and number of class-i abandonments, by time t, respectively. These processes are assumed to be given by

$$D_i^n = \tilde{D}_i^n\left(\int_0^{\cdot} Z_i^n(s)\,ds\right), \quad R_i^n = \tilde{R}_i^n\left(\int_0^{\cdot} Q_i^n(s)\,ds\right), \tag{3}$$

for some Poisson processes $\tilde{D}_i^n$ and $\tilde{R}_i^n$ with rates $\mu_i^n>0$ and $\theta_i^n>0$, respectively. The $3I$ processes A_i^n, $\tilde{D}_i^n$, $\tilde{R}_i^n$, and the initial condition $X^n(0) = (X_1^n(0),\dots,X_I^n(0))$, referred to as the *stochastic primitives*, are further assumed to be mutually independent (for each n). The above processes are related via the following equation:

$$X_i^n(t) = X_i^n(0) + A_i^n(t) - D_i^n(t) - R_i^n(t), \quad i\in\mathcal{I}, t\geqslant 0. \tag{4}$$

For simplicity, it will be assumed that all customers that are initially present in the system either start their service at time zero or are queued, and determining how many customers of each class start to be served at time zero is left for the scheduling policy. Namely, given $X^n(0)$, the policy will decide the values of $Q^n(0)$ and $Z^n(0)$ (subject, of course, to (1) and (2) holding at $t = 0$).

It is well understood that if the routing decisions are made in a causal manner based on the observed histories of the processes involved, namely D^n, R^n, X^n, Q^n, Z^n, then the construction of the departure and abandonment processes via (3) assures that the customers' service and patience times are independent, exponential random variables (as a simple consequence of Brémaud 1981, Theorem 16, p. 41). A special case is that of Markovian policies, under which (probabilistic) decisions are made depending on the current state, (X^n, Q^n). However, for the treatment of this paper, it will not be necessary to require any nonanticipating property of the class of policies we consider (although the exponential structure of service and abandonment will be lost when the policy does not satisfy a nonanticipating property). It will be simpler to use an elaborate definition of the term "policy" that will only rely on the equations presented thus far and the assumptions regarding the primitive processes. More precisely, *any process*

$$\pi^n = (D^n, R^n, X^n, Q^n, Z^n) \tag{5}$$

will be referred to as a policy for the nth system, provided that Equations (1)–(4) *hold, and that the stochastic primitives satisfy our probabilistic assumptions mentioned above*. The five processes on the r.h.s. of (5) are further assumed to possess right-continuous sample paths. Given n, the collection of all policies π^n for the nth system will be denoted by Π^n. Note that policies need not satisfy any work conservation (i.e., nonidling server) condition.

For each $i \in \mathcal{I}$, let $c_i \geqslant 0$ denote the holding cost per unit time for class-i customers. Thus, the instantaneous holding cost at time t is given by

$$c \cdot Q^n(t) = \sum_{i \in \mathcal{I}} c_i Q_i^n(t).$$

For a policy π^n, consider the normalized average holding cost function

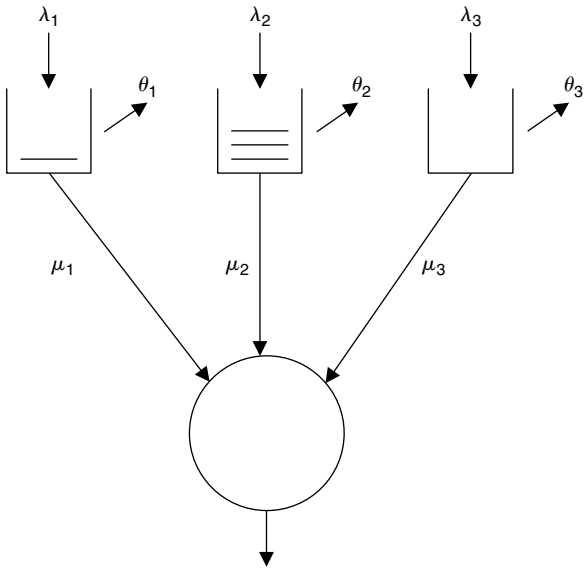
$$C_{n,T}(\pi^n) = \frac{1}{nT}\mathbb{E}\left[\int_0^T c \cdot Q^n(t)\, dt\right], \tag{6}$$

where Q^n and π^n are related via (5). Let the corresponding value be defined by

$$V_{n,T} = \inf_{\pi^n \in \Pi^n} C_{n,T}(\pi^n). \tag{7}$$

We consider a sequence of queueing systems as above, indexed by the number of servers $n \geqslant 1$. The parameters are assumed to satisfy the following assumption.

Figure 1. A model with three customer classes.



ASSUMPTION 2.1. *There exist positive constants λ_i, μ_i, and θ_i, $i \in \mathcal{I}$, such that, as $n \to \infty$,*

$$\frac{\lambda_i^n}{n} \to \lambda_i, \quad \mu_i^n \to \mu_i, \quad \theta_i^n \to \theta_i. \tag{8}$$

Note that the system may be overloaded in the sense that the workload exceeds the service capacity. However, because the abandonment rates are nonzero, stability holds automatically. The following will be assumed regarding the initial state.

ASSUMPTION 2.2. *The random variables $n^{-1}X^n(0)$ are uniformly bounded by a constant M.*

REMARK 2.1. Suppose that in addition to the holding cost $c \cdot Q^n$, we incur a penalty of size γ_i for each class-i customer that abandons the queue (i.e., reneges) before being admitted to service. To see how the cost function (6) is affected, observe that the expected reneging rate of class-i customers at time t is $\theta_i \mathbb{E}(Q_i^n(t))$, with associated penalty $\gamma_i \theta_i \mathbb{E}(Q_i^n(t))$. Therefore, we obtain the modified cost function

$$\begin{aligned} C_{n,T}(\pi^n) &= \frac{1}{nT}\mathbb{E}\left[\int_0^T \left(\sum_i c_i Q_i^n(t) + \sum_i \gamma_i \theta_i Q_i^n(t)\right) dt\right] \\ &= \frac{1}{nT}\mathbb{E}\left[\int_0^T \sum_i (c_i + \gamma_i \theta_i) Q_i^n(t)\, dt\right]. \end{aligned}$$

It is evident that this cost function reduces to the one in (6) once the cost parameters c_i are replaced by the modified values $\bar{c}_i = c_i + \gamma_i \theta_i$. Thus, all ensuing results hold for the extended cost model with reneging penalties after replacing each c_i by $\bar{c}_i$.

2.2. Scaled Processes

We introduce some notation. The fluid-scaled processes are defined as

$$\bar{X}^n = \frac{1}{n}X^n, \quad \bar{Q}^n = \frac{1}{n}Q^n, \quad \bar{Z}^n = \frac{1}{n}Z^n,$$
$$\bar{A}^n = \frac{1}{n}A^n, \quad \bar{R}^n = \frac{1}{n}R^n, \quad \bar{D}^n = \frac{1}{n}D^n.$$

Note that by Equations (1)–(2), (4), these processes satisfy

$$\sum_{i\in\mathcal{I}} \bar{Z}_i^n(t) \leqslant 1 \tag{9}$$

$$\bar{Q}_i^n(t) = \bar{X}_i^n(t) - \bar{Z}_i^n(t) \tag{10}$$

$$\bar{X}_i^n(t) = \bar{X}_i^n(0) + \bar{A}_i^n(t) - \bar{R}_i^n(t) - \bar{D}_i^n(t). \tag{11}$$

In the remainder of this subsection we provide some approximations for the scaled processes that will be useful later on in the paper. Given T and $\delta \in (0,1)$, define the event $E^n = E^n_{\delta,T}$ by

$$E^n = E_A^n \cap E_D^n \cap E_R^n, \tag{12}$$

where

$$E_A^n = \left\{ \max_i \sup_{t\in[0,T]} \left| \bar{A}_i^n(t) - \frac{\lambda_i^n}{n} t \right| < \delta \right\},$$

$$E_D^n = \left\{ \max_i \sup_{t\in[0,T]} \left| \frac{\tilde{D}_i^n(nt)}{n} - \mu_i^n t \right| < \delta \right\},$$

and

$$E_R^n = \left\{ \max_i \sup_{t\in[0,KT]} \left| \frac{\tilde{R}_i^n(nt)}{n} - \theta_i^n t \right| < \delta \right\}.$$

Above, $K = K(T) = \frac{1}{2}c_\lambda T + M + 1$, where $c_\lambda = \sup_n \max_i(\lambda_i^n/n) < \infty$, and $M < \infty$ denotes a bound on $n^{-1}\|X^n(0)\|$.

Lemma 2.1. *Let $T > 0$ and $\delta \in (0,1)$ be given. Fix a sequence of policies $\pi^n \in \Pi^n$, $n \in \mathbb{N}$. Then on the event E^n, one has, for every n,*

$$\left| \bar{D}_i^n(t) - \mu_i^n \int_0^t \bar{Z}_i^n(s)\,ds \right| \vee \left| \bar{R}_i^n(t) - \theta_i^n \int_0^t \bar{Q}_i^n(s)\,ds \right| < \delta, \quad i \in \mathcal{I},\ t \in [0,T] \tag{13}$$

and, with $K = K(T)$ as above,

$$\int_0^T \bar{Q}_i^n(s)\,ds \leqslant KT, \quad i \in \mathcal{I}. \tag{14}$$

Furthermore, $\mathbb{P}(E^n) \to 1$ as $n \to \infty$.

Proof. Fix $i \in \mathcal{I}$. By (3) and the definition of E_D^n, one has $|\bar{D}_i^n(t) - \mu_i^n \int_0^t \bar{Z}_i^n(s)\,ds| < \delta$ for every t such that $\int_0^t \bar{Z}_i^n(s)\,ds \leqslant T$. By (9), this property is valid for every $t \leqslant T$, and therefore the assertion regarding the processes $\bar{D}^n$ follows.

Similarly, by (3) and the definition of E_R^n, one has $|\bar{R}_i^n(t) - \theta_i^n \int_0^t \bar{Q}_i^n(s)\,ds| < \delta$ for every t such that $\int_0^t \bar{Q}_i^n(s)\,ds \leqslant KT$. Hence, to prove the statement regarding $\bar{R}^n$, it suffices to show that on E^n we have (14). Now, by (10), (11), and the positivity of $\bar{D}_i^n$, $\bar{R}_i^n$, and $\bar{Z}_i^n$, we have $\bar{Q}_i^n(t) \leqslant \bar{X}_i^n(0) + \bar{A}_i^n(t)$. Thus, by Assumption 2.2 and E_A^n, we have on E^n that, for all $t \leqslant T$, $\bar{Q}_i^n(t) \leqslant M + (\lambda_i^n/n)t + \delta$. Integrating from 0 to T, we obtain (14); hence, (13) follows.

Finally, the convergence $\mathbb{P}(E_A^n) \to 1$ as $n \to \infty$ is a standard consequence of the fact that, for each i, $\bar{A}_i^n = n^{-1}A_i^n$, where A_i^n is a Poisson process of rate λ_i^n and $\lim_n n^{-1}\lambda_i^n \in (0,\infty)$ (see Chen and Yao 2001, Chapter 5.6). For a similar reason, $\mathbb{P}(E_D^n \cap E_R^n) \to 1$. Consequently, $\mathbb{P}(E^n) \to 1$. □

3. The Fluid Model

In this section we consider a steady-state fluid model and corresponding linear program (LP) that are suggested by the scaled stochastic model of the previous section. We provide the simple solution of this fluid LP, and then show how it suggests (rather than implies) the $c\mu/\theta$ index rule that is the topic of this paper.

Consider quantities x, q, and z that formally represent the long-run average of the scaled processes for large values of n. These quantities are related by $x = q + z$, whereas q and z must satisfy

$$\begin{cases} \lambda_i = \mu_i z_i + \theta_i q_i \\ z_i, q_i \geqslant 0 \\ \sum_{i\in\mathcal{I}} z_i \leqslant 1. \end{cases} \tag{15}$$

This leads one to consider the problem:

$$\text{minimize } c \cdot q \text{ over all pairs } (q,z) \text{ satisfying (15).} \tag{16}$$

Denote by (q^*, z^*) a solution to this LP, let $x^* = q^* + z^*$, and let V^* denote the minimal value. To present the solution, it will be convenient to relabel the classes so that

$$\frac{c_1\mu_1}{\theta_1} \geqslant \cdots \geqslant \frac{c_I\mu_I}{\theta_I}. \tag{17}$$

It follows from the first equation in (15) that

$$\sum_i c_i q_i = \sum_i c_i \left(\frac{\lambda_i}{\theta_i} - \frac{\mu_i}{\theta_i} z_i \right).$$

Thus, the LP is equivalent to the problem:

$$\text{maximize} \sum_i \frac{c_i\mu_i}{\theta_i} z_i \quad \text{subject to} \quad 0 \leqslant z_i \leqslant \frac{\lambda_i}{\mu_i}, \quad \sum_i z_i \leqslant 1. \tag{18}$$

The solution to the latter is obviously to assign maximum values (namely λ_i/μ_i) to z_is that correspond to the larger class indices until the constraint $\sum_i z_i \leqslant 1$ is saturated. More precisely, the following is a solution:

$$z^* = \left(\frac{\lambda_1}{\mu_1}, \ldots, \frac{\lambda_{i_0-1}}{\mu_{i_0-1}}, 1 - \sum_{j=1}^{i_0-1} \frac{\lambda_j}{\mu_j}, 0, \ldots, 0\right),$$
$$q^* = \left(0, \ldots, 0, \frac{\lambda_{i_0} - \mu_{i_0} z_{i_0}}{\theta_{i_0}}, \frac{\lambda_{i_0+1}}{\theta_{i_0+1}}, \ldots, \frac{\lambda_I}{\theta_I}\right), \tag{19}$$

where

$$i_0 = \max\left\{i \in [1, \ldots, I+1]\colon \sum_{j=1}^{i-1} \frac{\lambda_j}{\mu_j} < 1\right\}. \tag{20}$$

In the case when $i_0 = I+1$, the interpretation of (19) is

$$z^* = \left(\frac{\lambda_1}{\mu_1}, \ldots, \frac{\lambda_I}{\mu_I}\right), \quad q^* = (0, \ldots, 0). \tag{21}$$

Finally, $V^* = c \cdot q^*$. Note that the solution need not be unique, as is the case, for example, when two of the indices have the same value for c_i, μ_i, θ_i, and λ_i.

The value of i_0 is related to whether or not the system is overloaded. There are three possibilities:

1. Underloaded system: $\sum_i(\lambda_i/\mu_i) < 1$. Here $i_0 = I+1$, and $q_i^* = 0$ for all i (all queues are empty at steady state).
2. Critically loaded system: $\sum_i(\lambda_i/\mu_i) = 1$. Here $i_0 = I$, but still $q_i^* = 0$ for all i.
3. Overloaded system: $\sum_i(\lambda_i/\mu_i) > 1$. Here $1 \leqslant i_0 \leqslant I$, with $z_i = \lambda_i/\mu_i$ and $q_i^* = 0$ for all $i < i_0$, whereas $z_i = 0$ and $q_i^* > 0$ for $i > i_0$.

Clearly the first two cases are degenerate, in the sense that all queues are empty in any optimal solution. Our main interest will be in the overloaded system case, where the service capacity is insufficient to handle all the service requirements, and nontrivial queues build up in fluid scale.

Let us see how the optimal fluid solution may be translated back to the original queueing system. A straightforward approach could be to assign a fixed number of servers to each service class according to the optimal fluid server share z_i^*: that is, assign $n_i = nz_i^*$ of the total n servers to class i. This server-scheduling scheme suffers from several shortcomings, among them nonsharing of servers across different service classes, and the need to precisely assess the arrival rate λ_i in order to compute z_i.

A more flexible approach, which is the one taken in this paper, relies on a priority rule that assigns strict priority to customers with smaller index i (which corresponds to larger $c_i\mu_i/\theta_i$). In the fluid limit, this would tend to indirectly realize the optimal LP solution by fully satisfying the service requirements of the higher service classes (up to i_0) while effectively starving the lower service classes beyond i_0.

We now turn to establish that this heuristic argument indeed leads to priority policies that are optimal in an appropriate asymptotic sense.

4. A Lower Bound on Performance

In this section we state and prove Proposition 4.1, which establishes a lower bound on the optimal performance. As in the previous section, we assume here and in the following that the service classes are numbered in decreasing orders of $c_i\mu_i/\theta_i$, so that (17) holds. Recall the event $E^n = E^n_{\delta,T}$ from (12).

LEMMA 4.1. *There exist constants c_0 and n_0 such that for every $\delta \in (0,1)$, $T \geqslant 1$, $n \geqslant n_0$, and $\pi^n \in \Pi^n$, the following holds on $E^n_{\delta,T}$:*

$$\|\bar{X}^n(T)\| \leqslant c_0.$$

PROOF. Let $\delta \in (0,1)$ and $T > 0$ be given. Denote $\xi^n = \sum_i \bar{X}_i^n$. Recall that $|\xi^n(0)| \leqslant M$ by Assumption 2.2. By (11),

$$\xi^n(t) = \xi^n(0) + \sum_i (\bar{A}_i^n(t) - \bar{D}_i^n(t) - \bar{R}_i^n(t)), \quad t \geqslant 0.$$

Denote throughout $\bar{\lambda}_i^n = \lambda_i^n/n$. Using Lemma 2.1, we have on $E^n = E^n_{\delta,T}$

$$\xi^n(t) = \eta^n(t) + \sum_i \int_0^t (\bar{\lambda}_i^n - \mu_i^n \bar{Z}_i^n(s) - \theta_i^n \bar{Q}_i^n(s))\, ds,$$
$$t \in [0,T],$$

where the difference $\eta^n(t)$ (defined by this equation) satisfied $|\eta^n(t)| \leqslant m_0 := M + 3I$, $t \in [0,T]$, and we used $\delta < 1$. Let $\theta_0 = \frac{1}{2}\min_{i\in\mathcal{J}} \theta_i > 0$, and let n_0 be such that $\theta_i^n \geqslant \theta_0$ for all $n \geqslant n_0$, $i \in \mathcal{J}$. Let $m = \sum_i \sup_n \bar{\lambda}_i^n < \infty$. Then for $n \geqslant n_0$,

$$\sum_i (\bar{\lambda}_i^n - \mu_i^n \bar{Z}_i^n - \theta_i^n \bar{Q}_i^n) \leqslant m - \theta_0 \sum_i \bar{Q}_i^n \leqslant m - \theta_0(\xi^n - 1),$$

using (9) and (10) in the last inequality. Hence, with $m' = m + \theta_0$, letting γ^n denote the unique solution to

$$\gamma^n(t) = \eta^n(t) + \int_0^t (m' - \theta_0 \gamma^n(s))\, ds, \quad t \in [0,T],$$

we have

$$\frac{d}{dt}(\xi^n - \gamma^n) \leqslant -\theta_0(\xi^n - \gamma^n) \quad \text{and} \quad (\xi^n - \gamma^n)(0) = 0.$$

Noting that the difference $(\xi^n - \gamma^n)$ is differentiable (as the nondifferentiable term η^n cancels out), we obtain from the above that

$$\frac{d}{dt}(\xi^n - \gamma^n) \leqslant -\theta_0(\xi^n - \gamma^n) \quad \text{and} \quad (\xi^n - \gamma^n)(0) = 0.$$

We conclude that $\xi^n \leqslant \gamma^n$ on $[0,T]$, on the event E^n. The solution γ^n is given by

$$\begin{aligned}\gamma^n(t) &= \eta^n(t) - \theta_0 \int_0^t \eta^n(s) e^{-\theta_0(t-s)}\, ds \\ &\quad + m't - \theta_0 \int_0^t m' s e^{-\theta_0(t-s)}\, ds \\ &\leqslant 2|\eta^n|_T^* + m'\theta_0 \quad \text{for all } t \in [0,T].\end{aligned}$$

Hence, $\gamma^n(t) \leqslant 2m_0 + m'\theta_0$, $t \in [0,T]$ on E^n. Consequently, $\|\bar{X}^n(t)\| = \xi^n(t)$ admits the same bound, on the same event.

The following proposition is the main result of this section.

PROPOSITION 4.1. *Denote*

$$v = \liminf_{T\to\infty} \liminf_{n\to\infty} V_{n,T},$$

then

$$v \geqslant V^*. \tag{22}$$

PROOF. Let $\epsilon > 0$ be arbitrary and fixed. Let c_0 be the constant from the above lemma. In the following, we assume, without loss of generality, that $c_0 \geqslant M \geqslant 1$. Fix $T \geqslant c_0/\epsilon$ for which

$$\liminf_n V_{n,T} \leqslant v + \epsilon.$$

Fix also a sequence of policies $\pi^n \in \Pi^n$, $n \in \mathbb{N}$, under which

$$\liminf_n C_{n,T}(\pi^n) \leqslant v + 2\epsilon.$$

In what follows we shall prove that

$$\liminf_n C_{n,T}(\pi^n) \geqslant V^* - \rho(\epsilon), \tag{23}$$

for some function ρ: $[0,\infty) \to [0,\infty)$ that is continuous and vanishing at zero. This, combined with the previous display and the fact that ϵ is arbitrary, implies $v \geqslant V^*$, hence (22).

Let X^n, D^n, R^n, etc., denote the processes corresponding to π^n via (5).

Consider the I-dimensional random vectors $q^n = (q_1^n, \ldots, q_I^n)$, $z^n = (z_1^n, \ldots, z_I^n)$ defined by

$$q^n = \frac{1}{T}\int_0^T \bar{Q}^n(s)\,ds, \quad z^n = \frac{1}{T}\int_0^T \bar{Z}^n(s)\,ds. \tag{24}$$

By (11), for every $i \in \mathcal{I}$,

$$\begin{aligned}\frac{1}{T}(\bar{X}_i^n(T) - \bar{X}_i^n(0)) &= \frac{1}{T}\bar{A}_i^n(T) - \frac{1}{T}\bar{D}_i^n(T) - \frac{1}{T}\bar{R}_i^n(T)\\ &= \bar{\lambda}_i^n - \theta_i^n q_i^n - \mu_i^n z_i^n + p_{1,i}^n + p_{2,i}^n + p_{3,i}^n,\end{aligned}$$

where

$$p_{1,i}^n = \frac{1}{T}\bar{A}_i^n(T) - \bar{\lambda}_i^n, \quad p_{2,i}^n = -\frac{1}{T}\bar{D}_i^n(T) + \mu_i^n z_i^n,$$

$$p_{3,i}^n = -\frac{1}{T}\bar{R}_i^n(T) + \theta_i^n q_i^n.$$

Fix $\delta \in (0,1)$. Using Lemma 4.1 and the fact $T \geqslant c_0/\epsilon$, we have on $E^n = E_{\delta,T}^n$

$$p_{4,i}^n(T) = \frac{1}{T}(\bar{X}_i^n(T) - \bar{X}_i^n(0)) \leqslant \frac{c_0}{T} \leqslant \epsilon,$$

and

$$p_{4,i}^n(T) \geqslant -\frac{\bar{X}_i^n(0)}{T} \geqslant -\frac{M}{T} \geqslant -\frac{c_0}{T} \geqslant -\epsilon,$$

by Assumption 2.2. By (12) and Lemma 2.1, on the event E^n, one has

$$e_i^n := \sum_{k=1}^4 |p_{k,i}^n(T)| \leqslant \frac{3\delta}{T} + \epsilon \leqslant 4\epsilon.$$

As a result, on the event E^n, the quantities (q^n, z^n) satisfy

$$\begin{cases}\tilde{\lambda}_i^n = \theta_i^n q_i^n + \mu_i^n z_i^n\\ z_i^n,\, q_i^n \geqslant 0\\ \sum_{i\in\mathcal{I}} z_i^n \leqslant 1,\end{cases} \tag{25}$$

where $|\tilde{\lambda}_i^n - \bar{\lambda}_i^n| \leqslant 4\epsilon$.

The proof of the following lemma is elementary and thus omitted.

LEMMA 4.2. *The solution V of the LP* (18) *is continuous in the parameters (λ, μ, θ) over the set $(0,\infty)^I \times (0,\infty)^I \times (0,\infty)^I$.*

As a result of (25) and the above lemma, there exists $n_0 = n_0(\epsilon)$ such that for all $n \geqslant n_0$, we have on the event E^n

$$c \cdot q^n \geqslant V^* - \rho(\epsilon),$$

for some function ρ: $[0,\infty) \to [0,\infty)$ that is continuous and vanishing at zero. Hence,

$$C_{n,T}(\pi^n) = \mathbb{E}[c \cdot q^n] \geqslant \mathbb{E}[\mathbf{1}_{E^n} c \cdot q^n],$$

and

$$\liminf_n C_{n,T}(\pi^n) \geqslant (V^* - \rho(\epsilon)) \liminf_n \mathbb{P}(E^n) = V^* - \rho(\epsilon),$$

by Lemma 2.1. This shows (23), and, as argued above, inequality (22) follows. □

5. The Preemptive $c\mu/\theta$ Policy

We next show that the preemptive $c\mu/\theta$ priority rule asymptotically achieves the lower bound V^* established in the previous section. Thus, let π_p^n denote the scheduling policy that assigns preemptive priority to service classes with higher $c_i\mu_i/c_i$. If there are two or more classes with the same index, some arbitrary but fixed order is assumed.

THEOREM 5.1. *Let Assumptions* 2.1 *and* 2.2 *hold. Then*

$$\liminf_{T\to\infty} \liminf_{n\to\infty} V_{n,T} = \limsup_{T\to\infty} \limsup_{n\to\infty} C_{n,T}(\pi_p^n) = V^*. \tag{26}$$

Throughout this section, the hypotheses of Theorem 5.1 are in force, and the processes D^n, R^n, X^n, etc., correspond to the policy π_p^n. The main tool will be the following.

LEMMA 5.1. *For every $\epsilon > 0$, there exists $T_\epsilon \in (0, \infty)$ such that for every $T \in (T_\epsilon, \infty)$,*

$$\sup_{t\in[T_\epsilon, T]} \|\bar{Z}^n(t) - z^*\| \vee \|\bar{Q}^n(t) - q^*\| < \epsilon \tag{27}$$

holds on the event $E^n_{\delta,T}$ for all $n \geqslant n_0$, where both δ and n_0 depend on ϵ and T.

PROOF. A simple analysis of the preemptive priority scheme shows the following relations between the processes $\bar{X}^n$, $\bar{Z}^n$, and $\bar{Q}^n$, namely,

$$\bar{Z}^n_i(t) = \bar{X}^n_i(t) \wedge \left[1 - \sum_{j=1}^{i-1} \bar{X}^n_j(t)\right]^+, \tag{28}$$

$$\bar{Q}^n_i(t) = \bar{X}^n_i(t) \wedge \left[\sum_{j=1}^{i} \bar{X}^n_j(t) - 1\right]^+. \tag{29}$$

Using this in (11) gives

$$\bar{X}^n_i(t) = \bar{X}^n_i(0) + \bar{\lambda}^n_i t - \mu^n_i \int_0^t \bar{X}^n_i(s) \wedge \left[1 - \sum_{j=1}^{i-1} X^n_j(s)\right]^+ ds - \theta^n_i \int_0^t \bar{X}^n_i(s) \wedge \left[\sum_{j=1}^{i} \bar{X}^n_j(s) - 1\right]^+ ds + e^n_i(t), \tag{30}$$

where

$$e^n_i(t) = (\bar{A}^n_i(t) - \bar{\lambda}^n_i t) - \left(\bar{D}^n_i(t) - \mu^n_i \int_0^t \bar{Z}^n_i(s)\, ds\right) - \left(\bar{R}^n_i(t) - \mu^n_i \int_0^t \bar{Q}^n_i(s)\, ds\right).$$

Note by Lemma 2.1 that, for given T and δ, we have on $E^n_{\delta,T}$

$$|e^n_i|^*_T \leqslant 3\delta, \quad i \in \mathcal{I}. \tag{31}$$

For $x = (x_i)_{i\in\mathcal{I}} \in [0, \infty)^I$, let $\xi = (\xi_i)_{i\in\mathcal{I}}$ denote the unique solution to the system of ordinary differential equations (ODE) (see Birkhoff and Rota 1989, Ch. V for existence and uniqueness of solutions to such systems with Lipschitz coefficients)

$$\xi_i(t) = x_i + \lambda_i t - \mu_i \int_0^t \xi_i(s) \wedge \left[1 - \sum_{j=1}^{i-1} \xi_j(s)\right]^+ ds - \theta_i \int_0^t \xi_i(s) \wedge \left[\sum_{j=1}^{i} \xi_j(s) - 1\right]^+ ds, \quad t \geqslant 0,\ i \in \mathcal{I}. \tag{32}$$

To denote the dependence on the initial condition, write $\xi(x, t)$. An argument by induction in i, provided in the appendix, shows

$$\lim_{t\to\infty} \xi(x, t) = x^*, \tag{33}$$

uniformly for $x \in [0, M]^I$. In what follows, write $\xi^n(t)$ for $\xi(\bar{X}^n(0), t)$. One has by (30) and (32),

$$\|\xi^n(t) - \bar{X}^n(t)\| \leqslant k^n(t) + m \int_0^t \|\xi^n(s) - \bar{X}^n(s)\|\, ds,$$

where $m = 2I \max_i \mu_i \vee \theta_i$, and

$$k^n(t) = (\|\lambda - \bar{\lambda}^n\| + \|\mu - \mu^n\| + K\|\theta - \theta^n\|)t + \|e^n(t)\|, \tag{34}$$

whence by Gronwall's lemma,

$$\|\xi^n - \bar{X}^n\|^*_t \leqslant |k^n|^*_t e^{mt}, \quad t \geqslant 0. \tag{35}$$

Now let $\epsilon > 0$ be given. Using (33), let T_ϵ be such that $\|\xi^n(t) - x^*\| < \epsilon/2$ for all $t \geqslant T_\epsilon$ and $n \in \mathbb{N}$. Next fix $T \in (T_\epsilon, \infty)$. We are required to show that (27) holds on $E^n_{\delta,T}$ for suitable δ and all sufficiently large n. Combining Assumption 2.1, (31), (34), and (35), it follows that δ can be chosen (depending on T) so that on $E^n_{\delta,T}$, we have $\|\xi^n - \bar{X}^n\|^*_T < \epsilon/2$, provided that n is sufficiently large. Therefore on $E^n_{\delta,T}$, for n large,

$$\|\bar{X}^n(t) - x^*\| < \epsilon, \quad t \in [T_\epsilon, T]. \tag{36}$$

Finally, due to (28) and (29), $\bar{Z}^n$ and $\bar{Q}^n$ are globally Lipschitz as functions of $\bar{X}^n$. Hence, (27) follows from (36). □

PROOF OF THEOREM 5.1. In view of Proposition 4.1, it suffices to prove

$$v_p := \limsup_{T\to\infty} \limsup_{n\to\infty} C_{n,T}(\pi^n_p) \leqslant V^*. \tag{37}$$

To this end, note first that, given t, the random variables $\|\bar{X}^n\|^*_t$, $n \in \mathbb{N}$ are uniformly integrable. Indeed, by (11), using the positivity of $\bar{X}^n$, $\bar{R}^n$, and $\bar{D}^n$, we have

$$\|\bar{X}^n\|^*_t \leqslant M + \|\bar{A}^n\|^*_t,$$

thus, the uniform integrability of $\|\bar{X}^n\|^*_t$ follows from that of $\|\bar{A}^n\|^*_t$ as a family of scaled Poisson processes.

Fix $\epsilon > 0$. Fix $T > T_\epsilon$, where T_ϵ is as in Lemma 5.1. Let δ and n_0 be the corresponding constants from Lemma 5.1. We have by (6),

$$C_{n,T}(\pi^n_p) \leqslant \frac{1}{T}\mathbb{E}\left[\mathbf{1}_{E^n_{\delta,T}} \int_0^T c \cdot \bar{Q}^n(t)\, dt\right] + \mathbb{E}[\mathbf{1}_{(E^n_{\delta,T})^c} \|c \cdot \bar{Q}^n\|^*_T].$$

By the uniform integrability of $\|\bar{X}^n\|^*_T$, and (9), the random variables $\|\bar{Q}^n\|^*_T$ are uniformly integrable. Moreover,

$$\begin{aligned}
&\frac{1}{T}\mathbb{E}\left[\mathbf{1}_{E^n_{\delta,T}} \int_0^T c \cdot \bar{Q}^n(t)\, dt\right] \\
&\quad = \frac{1}{T}\mathbb{E}\mathbf{1}_{E^n_{\delta,T}}\left[\int_0^{T_\epsilon} c \cdot \bar{Q}^n(t)\, dt + \int_{T_\epsilon}^T c \cdot \bar{Q}^n(t)\, dt\right] \\
&\quad \leqslant \frac{\|c\|K(T_\epsilon)T_\epsilon}{T} + c \cdot q^* + \|c\|\epsilon,
\end{aligned}$$

where we used Lemmas 2.1 and 5.1, and assumed $n \geqslant n_0$. Sending $n \to \infty$ and then $T \to \infty$ shows

$$v_p \leqslant c \cdot q^* + \|c\|\epsilon = V^* + \|c\|\epsilon.$$

Finally, because $\epsilon > 0$ is arbitrary, inequality (37) follows. □

6. The Nonpreemptive $c\mu/\theta$ Policy

In this section, we analyze the nonpreemptive $c\mu/\theta$ policy, denoted π^n_{np}. Nonpreemptive policies are necessary in applications where service may not be interrupted; however, their analysis is often more involved. In an asymptotic framework, sometimes one can show that the gap between optimal performance under nonpreemptive policies and under preemptive policies vanishes in the limit. This is the case, for example, in the works of Armony and Mandelbaum (2010) and Atar et al. (2004), which analyze many-server models in a diffusion regime. See also Rozenshmidt (2007) for comparison of preemptive and nonpreemptive performance under various asymptotic regimes. The main result of this section shows that under suitable assumptions, the performance of π^n_{np} asymptotically achieves the lower bound V^* from Proposition 4.1. Hence, in view of the result of §5, the optimal performance under preemptive and nonpreemptive policies is asymptotically the same.

A precise description of the policy π^n_{np} is as follows.

- $t = 0$. Given the initial condition $X^n(0)$, as many customers as possible are admitted at time zero, from the classes with highest priority.
- $t > 0$. Every time a server becomes free and some customers wait to be served, a customer from the class with the highest priority is admitted to service. Whenever a customer arrives to find a free server, it is admitted to service. Finally, service to a customer may not be stopped before it is completed.

The main result of this section is the following.

THEOREM 6.1. *Let Assumptions* 2.1 *and* 2.2 *hold. Assume, in addition, that $n^{-1}X^n(0)$ converge in distribution to x^*. Then*

$$\liminf_{T\to\infty}\liminf_{n\to\infty} V_{n,T} = \limsup_{T\to\infty}\limsup_{n\to\infty} C_{n,T}(\pi^n_{np}) = V^*. \quad (38)$$

To present the main estimate on which the proof is based, we need some notation. Fix vectors v and v' in $(0,1)^I$ with the following properties: If $i_0 > 1$, then for all $i < i_0$,

$$\text{(a)}\ \ v_i < v'_i/16, \quad \text{(b)}\ \ v_i < v'_i \inf_n \frac{\mu^n_i}{2\theta^n_i},$$
$$\text{(c)}\ \ v'_i < v'_{i_0}/i_0, \quad (39)$$

and if $i_0 \leqslant I$, then for all $i \geqslant i_0$,

$$\text{(a)}\ \ v'_{i_0} = v'_i < v_i/16, \quad \text{(b)}\ \ v'_i < v_i \inf_n \frac{\theta^n_i}{2\mu^n_i}. \quad (40)$$

For any $\epsilon > 0$, let $\epsilon_i = \epsilon v_i$ and $\epsilon'_i = \epsilon v'_i$, $i = 1,\dots,I$, and define $\mathcal{N}_\epsilon$ to be the following neighborhood of (q^*, z^*), namely,

$$\mathcal{N}_\epsilon = \prod_{i\in\mathcal{I}}(q^*_i - \epsilon_i, q^*_i + \epsilon_i) \times \prod_{i\in\mathcal{I}}(z^*_i - \epsilon'_i, z^*_i + \epsilon'_i).$$

Define the event $S^n_\epsilon \equiv S^n_{\epsilon,T}$ by

$$S^n_{\epsilon,T} := \{(\bar{Q}^n(t), \bar{Z}^n(t)) \in \mathcal{N}_\epsilon,\ \forall\, t \in [0,T]\}.$$

PROPOSITION 6.1. *Let the hypotheses of Theorem* 6.1 *hold. There exists $\epsilon_0 > 0$ such that for all $\epsilon \in (0,\epsilon_0)$ and $T > 0$, under the policy π^n_{np}, one has* $\lim_{n\to\infty}\mathbb{P}(S^n_{\epsilon,T}) = 1$.

Let us first show that Theorem 6.1 is an immediate consequence of Proposition 6.1.

PROOF OF THEOREM 6.1. In view of Proposition 4.1, we only need to show

$$\limsup_{T\to\infty}\limsup_{n\to\infty} C_{n,T}(\pi^n_{np}) \leqslant V^*.$$

Because under π^n_{np} on $S^n_{\epsilon,T}$ one has $c\cdot\bar{Q}^n(t) \leqslant c\cdot q^* + c\cdot v\epsilon$ for all $t \in [0,T]$, the proof is analogous to that of Theorem 5.1, and we omit the remaining details. □

PROOF OF PROPOSITION 6.1. Throughout, the processes D^n, R^n, X^n, etc., correspond to the policy π^n_{np}. Fix $\epsilon > 0$ and $T > 0$. (The value of ϵ will later be assumed, without loss of generality, to be sufficiently small.) The main idea of the proof is the following. For $n \in \mathbb{N}$ let

$$\tau^n = \inf\{t \geqslant 0\colon (\bar{Q}^n(t), \bar{Z}^n(t)) \notin \mathcal{N}_\epsilon\} \wedge (T+1),$$

and note that because $\mathcal{N}_\epsilon$ is an open set and $\bar{Q}^n$, $\bar{Z}^n$ have right-continuous paths, S^n_ϵ can be expressed as the event $\{\tau^n > T\}$. In order to estimate from below $\mathbb{P}(\tau^n > T)$, we prove, in four steps, the following claims.

(1) For every $i < i_0$,

$$\lim_{n\to\infty}\mathbb{P}(|\bar{Q}^n_i|^*_{\tau^n\wedge T} \geqslant \epsilon_i) = 0. \quad (41)$$

(2) For every $i < i_0$,

$$\lim_{n\to\infty}\mathbb{P}(|\bar{Z}^n_i - z^*_i|^*_{\tau^n\wedge T} \geqslant \epsilon'_i) = 0. \quad (42)$$

(3) For $i = i_0$ one has (42) and

$$\lim_{n\to\infty}\mathbb{P}(|\bar{Q}^n_i - q^*_i|^*_{\tau^n\wedge T} \geqslant \epsilon_i) = 0. \quad (43)$$

(4) For every $i > i_0$ one has (43) and (42).

The combination of the four statements shows that, with probability tending to one as $n\to\infty$, $(\bar{Q}^n(t), \bar{Z}^n(t)) \in \mathcal{N}_\epsilon$ for all $t \in [0, \tau^n \wedge T]$. By right continuity of the sample paths, on the event $\{\tau^n \leqslant T\}$ one has $(\bar{Q}^n(\tau^n), \bar{Z}^n(\tau^n)) \notin \mathcal{N}_\epsilon$. This shows that, with probability tending to one, $\tau^n > T$, thus establishing the proposition.

The four steps corresponding to the above four items appear after the following lemma.

Given $\delta > 0$, define $B^n_\delta := \{\|\bar{X}^n(0) - x^*\| < \delta\}$ and, slightly modifying the notation from §2, let $E^n_\delta := B^n_\delta \cap E^n_A \cap E^n_D \cap E^n_R$. We fix $\delta \in (0,\epsilon)$, whose precise value will be determined later.

The following is an adaptation of Lemma 2.1.

LEMMA 6.1. *There exists $\delta_0 \in (0,1)$ such that for all $\delta \in (0,\delta_0)$, $n \in \mathbb{N}$, on the event B^n_δ one has*

$$\|\bar{Q}^n(0) - q^*\| \leqslant 2I\delta, \quad \text{and} \quad \|\bar{Z}^n(0) - z^*\| \leqslant 2I\delta.$$

Consequently, the conclusions (13) *and* (14) *are valid under E^n_δ, and $\mathbb{P}(E^n_\delta) \to 1$ as $n\to\infty$.*

Proof. The last two claims are immediate consequences of Lemma 2.1 and the assumption that $\bar{X}^n(0)$ converge to x^* in distribution. It remains to prove the first claim of the lemma.

According to the policy, the classes with smallest indices receive highest priority in the initial job assignment. Let i_1 be the largest number i for which $\sum_{j=1}^{i} \bar{X}_j^n(0) \leqslant 1$. Then clearly, for $i \leqslant i_1$,

$$\bar{Z}_i^n(0) = \bar{X}_i^n(0) \in [x_i^* - \delta, x_i^* + \delta], \quad \bar{Q}_i^n(0) = 0.$$

Notice that $x^* = q^* + z^*$. By definition of i_0, $\sum_{j=1}^{i_0-1} x_j^* < 1$. Therefore, for δ small enough, $\sum_{j=1}^{i_0-1}(x_j^* + \delta) < 1$; hence, $i_1 \geqslant i_0 - 1$. Using the above display, we conclude that

$$\bar{Z}_i^n(0) \in [z_i^* - \delta, z_i^* + \delta], \quad \bar{Q}_i^n(0) = 0 \in [q_i^* - \delta, q_i^* + \delta], \quad \text{for all } i < i_0. \tag{44}$$

If $i_0 = I + 1$, we are done. Otherwise, $\bar{Z}_{i_0}^n(0) = \min\{1 - \sum_{j=1}^{i_0-1} \bar{Z}_j^n(0), \bar{X}_{i_0}^n(0)\}$, and using (44) and the fact $z_{i_0}^* = 1 - \sum_{j=1}^{i_0-1} z_j^*$, $1 - \sum_{j=1}^{i_0-1} \bar{Z}_j^n(0) \in [z_{i_0}^* - (i_0 - 1)\delta, z_{i_0}^* + (i_0 - 1)\delta]$, and $\bar{X}_{i_0}^n(0) \in [x_{i_0}^* - \delta, x_{i_0}^* + \delta]$. It follows that

$$\bar{Z}_{i_0}^n(0) \in [z_{i_0}^* - (i_0 - 1)\delta, z_{i_0}^* + (i_0 - 1)\delta],$$
$$\bar{Q}_{i_0}^n(0) = \bar{X}_{i_0}^n(0) - \bar{Z}_{i_0}^n(0) \in [q_{i_0}^* - i_0\delta, q_{i_0}^* + i_0\delta].$$

Finally, because $\sum_{j=1}^{i_0} \bar{Z}_j^n(0) \in [1 - 2(i_0 - 1)\delta, 1]$, we have for every $i \in [i_0 + 1, I]$, $\bar{Z}_i^n(0) \leqslant 2(i_0 - 1)\delta$. Because $z_i^* = 0$ for such i, the claim regarding $\bar{Z}^n(0)$ follows. Also, because $x_i^* = q_i^*$ for such i, $\bar{Q}_i^n(0) \in [\bar{X}_i^n(0) - 2(i_0 - 1)\delta, \bar{X}_i^n(0)] \subset [q_i^* - (2i_0 - 1)\delta, q_i^* + \delta]$. □

We now proceed with the proof of Proposition 6.1.

Step 1: We prove that, in case $i_0 \geqslant 2$, (41) holds for $i < i_0$. The case $i_0 = I + 1$ is very simple and treated separately at the end of this step. For now, let us assume $2 \leqslant i_0 \leqslant I$. Denote $\bar{\mathbf{Q}}^n = \sum_{i=1}^{i_0-1} \bar{Q}_i^n$ and $\epsilon_m = \min_{i=1,\ldots,i_0-1} \epsilon_i$. Let $F_i^n = \{|\bar{Q}_i^n|_{\tau^n \wedge T}^* \geqslant \epsilon_i\}$ and $F^n = \{|\bar{\mathbf{Q}}^n|_{\tau^n \wedge T}^* \geqslant \epsilon_m\}$. We shall argue that there exists $\delta > 0$ such that

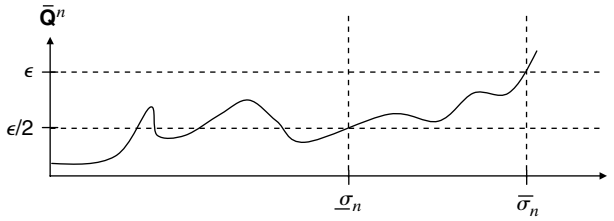
$$\lim_{n\to\infty} \mathbb{P}(E_\delta^n \cap F^n) = 0. \tag{45}$$

Because $F_i^n \subset F^n$, this will show $\lim_{n\to\infty} \mathbb{P}(E_\delta^n \cap F_i^n) = 0$, and because by Lemma 6.1 $\mathbb{P}(E_\delta^n) \to 1$, (41) will follow for $i = 1, \ldots, i_0 - 1$.

We will now analyze the event $E_\delta^n \cap F^n$, toward proving (45). The idea of the proof is that the time when the process crosses ϵ_m must be preceded by an interval on which it is strictly positive (see Figure 2). On such an interval the index policy gives priority to at least one of the classes from $\{1, \ldots, i_0 - 1\}$ over the others, and thus it is unlikely that the queue in one of these classes continues to build. Let us then denote

$$\bar{\sigma}_n = \inf\{t \geqslant 0 \colon \bar{\mathbf{Q}}^n(t) \geqslant \epsilon_m\} \wedge (T + 1), \tag{46}$$

Figure 2. The random times $\bar{\sigma}_n$ and $\underline{\sigma}_n$.



and

$$\underline{\sigma}_n = \sup\{t \leqslant \bar{\sigma}_n \colon \bar{\mathbf{Q}}^n(t) \leqslant \epsilon_m/2\}. \tag{47}$$

We assume, without loss of generality, that δ is so small that under B_δ^n one has $\bar{\mathbf{Q}}^n(0) < \epsilon_m/4$. Then it is not hard to see that on the event $E_\delta^n \cap F^n$ one has $0 < \underline{\sigma}_n \leqslant \bar{\sigma}_n \leqslant \tau^n \leqslant T$, provided that ϵ is sufficiently small. Consequently, using (46) and (47), one has on this event

$$\bar{\mathbf{Q}}^n(t) \geqslant \epsilon_m/2 \quad \text{for all } t \in [\underline{\sigma}_n, \bar{\sigma}_n],$$

and $\bar{\mathbf{Q}}^n(\bar{\sigma}_n) - \bar{\mathbf{Q}}^n(\underline{\sigma}_n -) \geqslant \epsilon_m/2$. With probability one, the size of the (upward) jump that the unnormalized queue-length process $\sum_{i=1}^{i_0-1} Q_i^n$ might have at time $\underline{\sigma}_n$ is at most one. Assuming without loss of generality that $n > 4/\epsilon_m$, the size of the jump of $\bar{\mathbf{Q}}^n$ is at most $\epsilon_m/4$, and as a result

$$\bar{\mathbf{Q}}^n(\bar{\sigma}_n) - \bar{\mathbf{Q}}^n(\underline{\sigma}_n) > \epsilon_m/4. \tag{48}$$

Recall that the $c\mu/\theta$ policy assigns to the classes $1, \ldots, i_0 - 1$ priority over the other classes. Because on the event $E_\delta^n \cap F^n$ there is a least one customer of class $i < i_0$ in the queue at every time within the interval $[\underline{\sigma}_n, \bar{\sigma}_n]$, every server that becomes free within this interval is assigned a new job from one of the classes $1, \ldots, i_0 - 1$. Consequently, the number of servers working on customers from classes $i_0, \ldots, I$ does not increase; moreover, it decreases by one each time there is service completion of a customer from one of these classes. Hence,

$$\sum_{i=i_0}^{I} (Z_i(\bar{\sigma}_n) - \bar{Z}_i(\underline{\sigma}_n)) = -\sum_{i=i_0}^{I} (D_i(\bar{\sigma}_n) - D_i(\underline{\sigma}_n)).$$

In addition, the work-conserving assumption assures that at $\underline{\sigma}_n$ and $\bar{\sigma}_n$ all servers are occupied, i.e.,

$$\sum_{i=1}^{I} Z_i^n(\bar{\sigma}_n) = \sum_{i=1}^{I} Z_i^n(\underline{\sigma}_n) = n.$$

Combining the above two displays, on $E_\delta^n \cap F^n$ one has

$$\begin{aligned}
&\sum_{i=1}^{i_0-1} (\bar{Z}_i(\bar{\sigma}_n) - \bar{Z}_i(\underline{\sigma}_n)) \\
&\quad = \sum_{i=1}^{I} (\bar{Z}_i(\bar{\sigma}_n) - \bar{Z}_i(\underline{\sigma}_n)) - \sum_{i=i_0}^{I} (\bar{Z}_i(\bar{\sigma}_n) - \bar{Z}_i(\underline{\sigma}_n)) \\
&\quad = -\sum_{i=i_0}^{I} (\bar{Z}_i(\bar{\sigma}_n) - \bar{Z}_i(\underline{\sigma}_n)) \\
&\quad = \sum_{i=i_0}^{I} (\bar{D}_i(\bar{\sigma}_n) - \bar{D}_i(\underline{\sigma}_n)).
\end{aligned} \tag{49}$$

Denote

$$\bar{\mathbf{X}}^n = \sum_{i=1}^{i_0-1} \bar{X}_i^n, \quad \bar{\mathbf{Z}}^n = \sum_{i=1}^{i_0-1} \bar{Z}_i^n, \quad \bar{\mathbf{A}}^n = \sum_{i=1}^{i_0-1} \bar{A}_i^n,$$

$$\bar{\mathbf{R}}^n = \sum_{i=1}^{i_0-1} \bar{R}_i^n, \quad \bar{\mathbf{D}}^n = \sum_{i=1}^{i_0-1} \bar{D}_i^n,$$

and

$$\bar{\mathbf{D}}_I^n = \sum_{i=1}^{I} \bar{D}_i^n.$$

On the event $E_\delta^n \cap F^n$, using (10) and (11), we obtain

$$\begin{aligned}
&\bar{\mathbf{Q}}^n(\bar{\sigma}_n) - \bar{\mathbf{Q}}^n(\underline{\sigma}_n) \\
&\quad = \bar{\mathbf{X}}^n(\bar{\sigma}_n) - \bar{\mathbf{X}}^n(\underline{\sigma}_n) - (\bar{\mathbf{Z}}^n(\bar{\sigma}_n) - \bar{\mathbf{Z}}^n(\underline{\sigma}_n)) \\
&\quad = \bar{\mathbf{A}}^n(\bar{\sigma}_n) - \bar{\mathbf{A}}^n(\underline{\sigma}_n) - (\bar{\mathbf{R}}^n(\bar{\sigma}_n) - \bar{\mathbf{R}}^n(\underline{\sigma}_n)) \\
&\qquad - (\bar{\mathbf{D}}^n(\bar{\sigma}_n) - \bar{\mathbf{D}}^n(\underline{\sigma}_n)) - (\bar{\mathbf{Z}}^n(\bar{\sigma}_n) - \bar{\mathbf{Z}}^n(\underline{\sigma}_n)) \\
&\quad = \bar{\mathbf{A}}^n(\bar{\sigma}_n) - \bar{\mathbf{A}}^n(\underline{\sigma}_n) - (\bar{\mathbf{R}}^n(\bar{\sigma}_n) - \bar{\mathbf{R}}^n(\underline{\sigma}_n)) \\
&\qquad - (\bar{\mathbf{D}}_I^n(\bar{\sigma}_n) - \bar{\mathbf{D}}_I^n(\underline{\sigma}_n)),
\end{aligned}$$

where on the last line we used (49).

In what follows, denote $\bar{\lambda}_i^n = \lambda_i^n/n$. Because the increment of $\bar{\mathbf{R}}^n$ is nonnegative, using (48) along with the definition of E_δ^n and Lemma 6.1, we have on $E_\delta^n \cap F^n$

$$\begin{aligned}
\frac{\epsilon_m}{4} &\leqslant \bar{\mathbf{A}}^n(\bar{\sigma}_n) - \bar{\mathbf{A}}^n(\underline{\sigma}_n) - \bar{\mathbf{D}}_I^n(\bar{\sigma}_n) + \bar{\mathbf{D}}_I^n(\underline{\sigma}_n) \\
&\leqslant (\bar{\sigma}_n - \underline{\sigma}_n) \sum_{i=1}^{i_0-1} \bar{\lambda}_i^n - \int_{\underline{\sigma}_n}^{\bar{\sigma}_n} \sum_{i=1}^{I} \mu_i^n \bar{Z}_i^n(s)\, ds + 4I\delta.
\end{aligned}$$

By definition of τ^n, one has $\bar{Z}_i^n(t) \geqslant z_i^* - \epsilon_i'$, for $t \in [0, \tau^n)$. Therefore, using Assumption 2.1, given $\delta' > 0$, for all sufficiently large n, the following holds on $E_\delta^n \cap F^n$

$$\begin{aligned}
\frac{\epsilon_m}{4} &\leqslant (\bar{\sigma}_n - \underline{\sigma}_n) \left[\sum_{i=1}^{i_0-1} (\lambda_i + \delta') - \sum_{i=1}^{I} (\mu_i - \delta')(z_i^* - \epsilon_i') \right] \\
&\quad + 4I\delta. \qquad (50)
\end{aligned}$$

It follows from (19) and the assumption $i_0 \leqslant I$ that

$$\sum_{i=1}^{i_0-1} \lambda_i < \sum_{i=1}^{I} \mu_i z_i^*.$$

Thus, we may assume without loss of generality that δ' is sufficiently small so that the expression in square brackets in (50) is negative. Assuming further, without loss of generality, that $4I\delta < \epsilon_m/4$, inequality (50) cannot hold. We have thus shown the following. There exists a constant $\epsilon_0 > 0$ such that for every $\epsilon \in (0, \epsilon_0)$ there exists $\delta_0 = \delta_0(\epsilon) > 0$, such that if $\delta \in (0, \delta_0)$, then $E_\delta^n \cap F^n$ is empty for all sufficiently large n. (The freedom of selecting any $\delta \in (0, \delta_0)$, rather than a particular $\delta(\epsilon)$, is not used in the present argument, but it will be important when the result of this step is used in the following steps.) This shows that (45), and consequently (41), holds for every $\epsilon \in (0, \epsilon_0)$.

Finally, the case $i_0 = I + 1$ is very easy. By (20)–(21), in this case $\sum_{i=1}^{I} z_i^* < 1$; hence $\|\bar{Z}^n - z^*\| < \epsilon$, provided ϵ is sufficiently small. This means that within $[0, \tau^n]$ there are always free servers and, by work conservation, no customers in queue. Thus, on the event $\{\tau^n \leqslant T\}$ it is impossible that $Q_i^n(\tau_n) > 0$ for any i, and (41) follows.

Step 2: We assume $i_0 \in [2, I+1]$ and prove (42), for $i < i_0$.

Fix $i < i_0$. Let $G_{i,+}^n = \{\bar{Z}_i^n(\tau^n \wedge T) - z_i^* \geqslant \epsilon_i'\}$, and $G_{i,-}^n = \{\bar{Z}_i^n(\tau^n \wedge T) - z_i^* \leqslant -\epsilon_i'\}$, $G_i^n = G_{i,+}^n \cup G_{i,-}^n$. We will establish (42) by showing that there exists δ such that

$$\lim_n \mathbb{P}(E_\delta^n \cap G_i^n) = 0. \qquad (51)$$

Consider first the event $E_\delta^n \cap G_{i,+}^n$. We use the symbols $\bar{\sigma}_n$, $\underline{\sigma}_n$ in a way similar to their use in Step 1. Namely, let $\epsilon > 0$ be given, and let

$$\begin{aligned}
\bar{\sigma}_n &= \inf\{t \geqslant 0\colon \bar{Z}_i^n(t) - z_i^* \geqslant \epsilon_i'\} \wedge (T+1), \\
\underline{\sigma}_n &= \sup\{t \leqslant \underline{\sigma}_n\colon \bar{Z}_i^n(t) - z_i^* \leqslant \epsilon_i'/2\}.
\end{aligned}$$

Without loss of generality, δ is assumed to be so small that on E_δ^n one has $|\bar{Z}_i^n(0) - z_i^*| < \epsilon_i'/4$. Then, arguing as in Step 1, on $E_\delta^n \cap G_{i,+}^n$ one has

$$\bar{Z}_i^n(t) \geqslant z_i^* + \epsilon_i'/2 \quad \text{for all } t \in [\underline{\sigma}_n, \bar{\sigma}_n], \qquad (52)$$

and, assuming n is sufficiently large,

$$\bar{Z}_i^n(\bar{\sigma}_n) - \bar{Z}_i^n(\underline{\sigma}_n) > \epsilon_i'/4. \qquad (53)$$

Sum (10) and (11) and isolate $\bar{Z}_i^n(t)$ to obtain

$$\bar{Z}_i^n(t) = \bar{X}_i^n(0) + \bar{A}_i^n(t) - \bar{D}_i^n(t) - \bar{R}_i^n(t) - \bar{Q}_i^n(t). \qquad (54)$$

We now use the results of Step 1. Note by (39a) that $\epsilon_i \leqslant \epsilon_i'/16$. As mentioned in the first step, $F_i^n \subset F^n$. Then by taking δ smaller if necessary, we have that for all sufficiently large n, $E_\delta^n \cap F^n$ is empty, thus also $E_\delta^n \cap F_i^n$. By definition of F_i^n, we have on the complement $F_i^{n,c}$ of F_i^n that $\bar{Q}_i^n(t) \leqslant \epsilon_i'/16$ for $t \in [0, \tau^n \wedge T]$. As a result, on the event $E_\delta^n \cap F_i^{n,c} \cap G_{i,+}^n$ we have, using (54) and the fact that R is nondecreasing,

$$\frac{\epsilon_i'}{8} < \bar{A}_i^n(\bar{\sigma}^n) - \bar{A}_i^n(\underline{\sigma}^n) - \bar{D}_i^n(\bar{\sigma}^n) + \bar{D}_i^n(\underline{\sigma}^n).$$

Hence, using Lemma 6.1, we have on the same event, for all sufficiently large n,

$$\frac{\epsilon_i'}{8} < (\bar{\sigma}_n - \underline{\sigma}_n)\bar{\lambda}_i^n - \int_{\underline{\sigma}_n}^{\bar{\sigma}_n} \mu_i^n \bar{Z}_i^n(s)\, ds + 4\delta.$$

Now using (52) and assuming, without loss of generality, that $4\delta < \epsilon_i'/16$, we have on $E_\delta^n \cap F_i^{n,c} \cap G_{i,+}^n$,

$$\frac{\epsilon_i'}{16} < (\bar{\sigma}_n - \underline{\sigma}_n)(\bar{\lambda}_i^n - \mu_i^n(z_i^* + \epsilon_i'/2)).$$

Because $i < i_0$, we have by (8) and (19) that $\bar{\lambda}_i^n - \mu_i^n z_i^* \to 0$ as $n \to \infty$. Because we also have that μ_i^n converge to a positive constant, we obtain that the above inequality does not hold, provided that n is sufficiently large. Thus, $\mathbb{P}(E_\delta^n \cap G_{i,+}^n) \to 0$ as $n \to \infty$. By (45), we conclude that $\delta > 0$ can be chosen so that

$$\lim_n \mathbb{P}(E_\delta^n \cap G_{i,+}^n) = 0. \tag{55}$$

The analysis of $E_\delta^n \cap G_{i,-}^n$ is very similar, and therefore we only give a sketch of the argument. Let

$$\bar{\sigma}_n = \inf\{t \geqslant 0\colon \bar{Z}_i^n(t) - z_i^* \leqslant -\epsilon_i'\} \wedge (T+1),$$
$$\underline{\sigma}_n = \sup\{t \leqslant \underline{\sigma}_n\colon \bar{Z}_i^n(t) - z_i^* \geqslant -\epsilon_i'/2\}.$$

Arguing as before, using in addition the estimate on $\bar{R}_i^n$ according to Lemma 6.1, shows that δ can be chosen so that one has on $E_\delta^n \cap F_i^{n,c} \cap G_{i,-}^n$:

$$\begin{aligned}(\bar{\sigma}_n - \underline{\sigma}_n)&(\bar{\lambda}_i^n - \mu_i^n(z_i^* - \epsilon_i'/2) - \theta_i^n \epsilon_i) \\ &\leqslant \frac{\epsilon_i'}{16} + 2\epsilon_i - \epsilon_i'/4 = -\frac{\epsilon_i'}{16}.\end{aligned} \tag{56}$$

Using once more the facts $\lim_n(\bar{\lambda}_i^n - \mu_i^n z_i^*) = 0$ and $\lim_n \mu_i^n > 0$, using (39b) by which $\epsilon_i < \epsilon_i'(\mu_i^n/(2\theta_i^n))$, it follows that the above inequality does not hold, provided that n is sufficiently large. As in the previous argument, this yields

$$\lim_n \mathbb{P}(E_\delta^n \cap G_{i,-}^n) = 0.$$

The above display and (55) establish (51), and by a further application of Lemma 6.1, it follows that $\mathbb{P}(G_i^n) \to 0$ as $n \to \infty$. To see that this implies (42), note that by definition of τ^n, $|\bar{Z}_i^n - z_i^*|_{\tau^n \wedge T}^* \geqslant \epsilon_i'$ implies $|\bar{Z}_i^n(\tau^n \wedge T) - z_i^*| \geqslant \epsilon_i'$. Thus (42) follows.

Step 3: Establishing (42) and (43) for $i = i_0$ is more subtle and requires further splitting of the events involved. Let

$$H_{i,+}^n = \{\bar{Q}_i^n(\tau^n \wedge T) - q_i^* \geqslant \epsilon_i\},$$
$$H_{i,-}^n = \{\bar{Q}_i^n(\tau^n \wedge T) - q_i^* \leqslant -\epsilon_i\},$$

and $H_i^n = H_{i,+}^n \cup H_{i,-}^n$. Then (43) and (42) are established by showing that there exists some $\delta > 0$ such that

$$\lim_n \mathbb{P}(E_\delta^n \cap H_{i_0}^n) = 0, \quad \lim_n \mathbb{P}(E_\delta^n \cap G_{i_0}^n) = 0, \tag{57}$$

respectively.

We begin by analyzing the event $E_\delta^n \cap G_{i_0,+}^n$. As follows from the previous step, the event $E_\delta^n \cap (\bigcup_{j=1}^{i_0-1} G_{j,-}^n)$ is empty for every $\epsilon \in (0, \epsilon_0)$ and sufficiently large n. Hence, it suffices to consider the event $J^n := E_\delta^n \cap (\bigcup_{j=1}^{i_0-1} G_{j,-}^n)^c \cap G_{i_0,+}^n$. Note by (39c) that $\epsilon_i' \leqslant \epsilon_{i_0}'/i_0$, for every $i = 1, \ldots, i_0 - 1$. On this event, for $t \in [0, \tau_n \wedge T]$,

$$\bar{Z}_{i_0}^n(t) \leqslant 1 - \sum_{j=1}^{i_0-1} \bar{Z}_j^n(t) \leqslant z_{i_0}^* + \epsilon_{i_0}'.$$

Thus, we have

$$\lim_n \mathbb{P}(E_\delta^n \cap G_{i_0,+}^n) = 0. \tag{58}$$

Next we analyze $E_\delta^n \cap H_{i_0,-}^n$. Clearly, if $q_{i_0}^* = 0$ this event is empty by nonnegativity of the queue-length process. We thus assume $q_{i_0}^* > 0$. Similarly to previous steps, define

$$\bar{\sigma}_n = \inf\{t \geqslant 0\colon \bar{Q}_{i_0}^n(t) - q_{i_0}^* \leqslant -\epsilon_{i_0}\} \wedge (T+1),$$
$$\underline{\sigma}_n = \sup\{t \leqslant \underline{\sigma}_n\colon \bar{Q}_{i_0}^n(t) - q_{i_0}^* \geqslant -\epsilon_{i_0}/2\}.$$

Along the lines of the previous steps, if δ is sufficiently small and n is sufficiently large, then

$$\begin{aligned}&\bar{Q}_{i_0}^n(t) \leqslant q_{i_0}^* - \epsilon_{i_0}/2 \quad \text{for all } t \in [\underline{\sigma}_n, \bar{\sigma}_n], \quad \text{and} \\ &\bar{Q}_{i_0}^n(\bar{\sigma}_n) - \bar{Q}_{i_0}^n(\underline{\sigma}_n) < -\epsilon_{i_0}/4.\end{aligned} \tag{59}$$

By (58) we may ignore $E_\delta^n \cap H_{i_0,-}^n \cap G_{i_0,+}^n$, and consider only $\tilde{J}^n := E_\delta^n \cap H_{i_0,-}^n \cap (G_{i_0,+}^n)^c$. Using Lemma 6.1, on $E_\delta^n \cap H_{i_0,-}^n$

$$\begin{aligned}-\frac{\epsilon_{i_0}}{4} > (\bar{\sigma}_n - \underline{\sigma}_n)\bar{\lambda}_{i_0}^n &- \int_{\underline{\sigma}_n}^{\bar{\sigma}_n} \mu_{i_0}^n \bar{Z}_{i_0}^n(s)\,ds \\ &- \int_{\underline{\sigma}_n}^{\bar{\sigma}_n} \theta_{i_0}^n \bar{Q}_{i_0}^n(s)\,ds - 6\delta - 2\epsilon_{i_0}'.\end{aligned} \tag{60}$$

Using (40a), without loss of generality we may assume $6\delta + 2\epsilon_{i_0}' < \epsilon_{i_0}/8$. Using (59) and the definition of $G_{i_0,+}^n$,

$$-\frac{\epsilon_{i_0}}{16} > (\bar{\sigma}_n - \underline{\sigma}_n)(\bar{\lambda}_{i_0}^n - \mu_{i_0}^n(z_{i_0}^* + \epsilon_{i_0}') - \theta_{i_0}^n(q_{i_0}^* - \epsilon_{i_0}/2)).$$

By (8) and (19), $\bar{\lambda}_{i_0}^n - \mu_{i_0}^n z_{i_0}^* - \theta_{i_0}^n q_{i_0}^* \to 0$. Because by (40b) $\epsilon_{i_0}' < \epsilon_{i_0} \inf_n \theta_{i_0}^n/2\mu_{i_0}^n$, for large-enough n the above inequality does not hold. We have thus argued that

$$\lim_n \mathbb{P}(E_\delta^n \cap H_{i_0,-}^n) = 0.$$

We analyze $G_{i_0,-}^n$. Consider first the case $q_{i_0}^* > 0$. The above argument provides us with the information that the i_0th queue is not empty until time $\tau^n \wedge T$, with high probability (namely, on the event $(H_{i_0,-}^n)^c$). Thus, on this event we necessarily have

$$\bar{Z}_{i_0}^n = 1 - \sum_{j=1}^{i_0-1} \bar{Z}_j^n$$

on $[0, \tau^n \wedge T]$. As a result, on $E^n_\delta \cap (H^n_{i_0})^c \cap (\bigcup_{j=1}^{i_0-1} G^n_{j,+})^c$,

$$\bar{Z}^n_{i_0} \geqslant 1 - \sum_{j=1}^{i_0-1} (z^*_j + \epsilon'_j) \geqslant z^*_{i_0} - \epsilon'_{i_0},$$

where we used (39c). In view of the information we already have about $H^n_{i_0}$ and $G^n_{j,+}$, $j < i_0$, this shows

$$\lim_n \mathbb{P}(E^n_\delta \cap G^n_{i_0,-}) = 0.$$

In case $q^*_{i_0} = 0$, at times when the i_0th queue is empty, it is possible that $\bar{Z}^n_{i_0} < 1 - \sum_{j=1}^{i_0-1} \bar{Z}^n_j$. During such a time period no abandonments occur, and (56) takes the form

$$(\bar{\sigma}_n - \underline{\sigma}_n)(\bar{\lambda}^n_i - \mu^n_i(z^*_i - \epsilon'_i/2)) \leqslant -\frac{\epsilon'_i}{16}.$$

Because the l.h.s. is positive for large-enough n, the remainder of the argument is similar to that provided in Step 2, and thus we omit the details.

The analysis of $E^n_\delta \cap H^n_{i_0,+}$ is very similar to the analysis of $E^n_\delta \cap H^n_{i_0,-}$. Set

$$\bar{\sigma}_n = \inf\{t \geqslant 0\colon \bar{Q}^n_i(t) - q^*_i \geqslant \epsilon_{i_0}\} \wedge (T+1),$$
$$\underline{\sigma}_n = \sup\{t \leqslant \underline{\sigma}_n\colon \bar{Q}^n_i(t) - q^*_i \leqslant \epsilon_{i_0}/2\}.$$

Intersecting $E^n_\delta \cap H^n_{i_0,+}$ with $G^n_{i_0}$ results in an empty set, whereas on $E^n_\delta \cap H^n_{i_0,+} \cap (G^n_{i_0})^c$ one has

$$\frac{\epsilon_{i_0}}{4} < (\bar{\sigma}_n - \underline{\sigma}_n)(\bar{\lambda}^n_{i_0} - \mu^n_{i_0}(z^*_{i_0} - \epsilon'_{i_0}) - \theta^n_{i_0}(q^*_{i_0} + \epsilon_{i_0}/2)). \quad (61)$$

Using the convergence $\bar{\lambda}^n_{i_0} - \mu^n_{i_0} z^*_{i_0} - \theta^n_{i_0} q^*_{i_0} \to 0$, and the choice of ϵ'_{i_0}, the above inequality does not hold, provided that n is sufficiently large. As previously, this shows

$$\lim_n \mathbb{P}(E^n_\delta \cap H^n_{i_0,+}) = 0.$$

The claim of this step follows.

Step 4: As a direct consequence of Steps 1–3,

$$\lim_n \mathbb{P}\left(\sup_{t \in [0, \tau^n \wedge T]} \sum_{i > i_0} \bar{Z}^n_i(t) > \epsilon \right) = 0. \quad (62)$$

This shows that (42) holds for $i > i_0$.

Similarly to Step 3, intersecting $H^n_{i,+}$ and $H^n_{i,-}$ with $(\bigcup_{j=1}^{i_0} G^n_j)^c$ yields (61) and (60), respectively, where now $z^*_i = 0$. Hence, for $\delta \in (0, \delta_0)$ and all large n,

$$\lim_n \mathbb{P}(E^n_\delta \cap H^n_i) = 0, \quad i > i_0.$$

This shows (43) for $i > i_0$, thus completing Step 4. This completes the proof of Proposition 6.1. □

7. Conclusion

We have identified a simple fixed-priority rule designed to minimize a combination of holding costs and abandonment penalties in a multiclass many-server queueing system with abandonment. Both the preemptive and nonpreemptive versions of this rule were analyzed and shown to minimize the long-term average cost in an asymptotic sense. The proposed priority rule is akin to the celebrated $c\mu$ rule (which was designed for systems without abandonments), but still differs in several important respects. Let us briefly elaborate on these similarities and differences.

First, the $c\mu/\theta$ rule reduces to the $c\mu$ rule when the abandonment parameters (rate and penalty) are identical for all classes. However, when these parameters differ across classes, the two indices differ accordingly.

Second, similarly to the $c\mu$ rule, the $c\mu/\theta$ rule does not depend on the arrival rates of the different customer classes. This gives the resulting policy a considerable measure of robustness and simplicity of implementation, because arrival rates are often varying and unpredictable.

Unlike the $c\mu$ rule, whose exact optimality has been demonstrated for a variety of cost functions, including finite horizon and discounted costs, our claim of optimality of the present $c\mu/\theta$ rule is restricted to the long-term average cost, under a many-server fluid scaling. In fact, an example outlined in Atar et al. (2008) demonstrates that the $c\mu/\theta$ need not be asymptotically optimal for a finite time horizon version of the cost function.

We close this paper by mentioning some directions of interest for future research. As observed before, our fluid-scale optimality criterion is meaningful only in overloaded conditions, because the queue sizes trivialize otherwise. The efficiency of the $c\mu/\theta$ rule under critically loaded or underloaded conditions is an open issue, which requires more refined tools for its analysis.

Our analysis in this paper was restricted to a model with Poisson arrivals, exponential service and patience distribution, as well as linear holding costs. It would of course be of interest to alleviate these assumptions. We conjecture that the $c\mu/\theta$ rule and its asymptotic optimality will prove insensitive to the input and service distributions, whereas both nonlinear holding costs and nonexponential patience distribution may lead to dynamic priority rules (in the style of Van Mieghem 1995). These extensions should be at the focus of our future work.

Appendix

Proof of Property (33) of Solutions to (32). Recall that ξ is a solution to the equation

$$\frac{d\xi_i}{dt} = \lambda_i - \mu_i \xi_i \wedge \left[1 - \sum_{j=1}^{i-1} \xi_j\right]^+ - \theta_i \xi_i \wedge \left[\sum_{j=1}^{i} \xi_j - 1\right]^+, \quad i \in \mathcal{I},\ \xi(0) = x \quad (63)$$

(where $ab \wedge c$ is understood as $a(b \wedge c)$). A straightforward analysis of the one-dimensional equation

$$\frac{d\zeta}{dt} = \lambda - \mu\zeta \wedge C_1 - \theta\zeta \wedge (\zeta - C_2)^+, \quad \zeta(0) = \zeta_0 \tag{64}$$

shows that the solution ζ converges to a limit as $t \to \infty$ uniformly for $\zeta_0 \in [0, M]$, irrespective of the values of the *positive* constants λ, μ, θ, C_1, and C_2. The limit value is obtained by equating the r.h.s. of (64) to zero and is easily seen to be continuous with respect to perturbations in $C_1 > 0$ and $C_2 > 0$. It is a standard fact from the theory of ODE (comparison theorem Birkhoff and Rota 1989) that a solution ζ to the one-dimensional equation

$$\frac{d\zeta}{dt} = \lambda - \mu\zeta \wedge D_1 - \theta\zeta \wedge (\zeta - D_2)^+, \quad \zeta(0) = \zeta_0, \tag{65}$$

(where D_1 and D_2 are time dependent) and a solution $\tilde{\zeta}$ to an equation of the same form, with some $(\tilde{D}_1, \tilde{D}_2)$ replacing (D_1, D_2), satisfy $\zeta \geqslant \tilde{\zeta}$ on $[0, \infty)$, provided $D_1 \leqslant \tilde{D}_1$ and $D_2 \geqslant \tilde{D}_2$ on $[0, \infty)$. As a consequence of this and the continuity property of the limit value alluded to above, given any $\epsilon > 0$ one can find $\delta > 0$ and $T \geqslant 0$ such that a solution to (65) is ϵ-close to a solution to (64) on $[T, \infty)$, provided (D_1, D_2) is δ-close to (C_1, C_2) on $[0, \infty)$.

The component ξ_1 of (63) satisfies (64) with $C_1 = C_2 = 1$, and thus converges. Hence, for large values of t, ξ_1 is nearly a constant. Consequently, on $[T, \infty)$, T being large, ξ_2 satisfies an equation of the form (64) where C_1 and C_2 are nearly constants, and by the foregoing discussion, on some $[T_1, \infty)$, it must be ϵ-close to a limit value of (64) for a suitable choice of C_1 and C_2. Because ϵ is arbitrary, this shows ξ_2 converges. It is easy to see that this argument can be iterated for $i \in \{3, \ldots, I\}$. □

Acknowledgments

Useful discussions with A. Mandelbaum and W. Whitt at the initial part of this research are gratefully acknowledged. This research is supported in part by the ISF (grant 1349/08), by the Technion's fund for promotion of research, and by grant 2006379 from the United States–Israel Binational Science Foundation (BSF).

References

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.

Armony, M., A. Mandelbaum. 2010. Design, staffing and control of large service systems: The case of a single customer class and multiple server types. *Oper. Res.* Forthcoming.

Atar, R., C. Giat, N. Shimkin. 2008. The $c\mu/\theta$ rule. *Proc. Valuetools, Athens*.

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

Baras, J. S., D.-J. Ma, A. Makowski. 1985. *K* competing queues with geometric service requirements and linear costs: The μc-rule is always optimal. *Systems Control Lett.* **6**(3) 173–180.

Bassamboo, A., J. M. Harrison, A. Zeevi. 2006a. Design and control of a large call center: Asymptotic analysis of an LP-based method. *Oper. Res.* **54**(3) 419–435.

Bassamboo, A., J. M. Harrison, A. Zeevi. 2006b. Dynamic routing and admission control in high volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Systems* **51**(3–4) 249–285.

Birkhoff, G., G. C. Rota. 1989. *Ordinary Differential Equations*, 4th ed. John Wiley & Sons, New York.

Brémaud, P. 1981. *Point Processes and Queues*. Springer, New York.

Chen, H., D. Yao. 2001. *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Springer, New York.

Cox, D. R., W. L. Smith. 1961. *Queues*. Methuen, London.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* **5**(2) 79–141.

Hirayama, T., M. Kijima, S. Nishimura. 1989. Further results for dynamic scheduling of multiclass G/G/l queues. *J. Appl. Probab.* **26** 595–603.

Kang, W., K. Ramanan. 2010. Fluid limits of many-server queues with reneging. *Ann. Appl. Probab.* Forthcoming.

Kaspi, H., K. Ramanan. 2010. Law of large numbers limits for many-server queues. *Ann. Appl. Probab.* Forthcoming.

Klimov, G. P. 1974. Time-sharing service systems. I. *Theory Probab. Appl.* **19**(3) 532–551.

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

Mandelbaum, A., W. A. Massey, M. I. Reiman. 1998. Strong approximations for Markovian service networks. *Queueing Systems* **30**(1–2) 149–201.

Nain, P., D. Towsley. 1994. Optimal scheduling in a machine with stochastic varying processing rate. *IEEE Trans. Automatic Control* **39**(9) 1853–1855.

Perry, O., W. Whitt. 2009. A fluid approximation for service systems responding to unexpected overloads. Preprint.

Reed, J. E. 2010. The G/GI/N queue in the Halfin-Whitt regime. *Ann. Appl. Probab.* Forthcoming.

Rozenshmidt, L. 2007. On priority queues with impatient customers: Stationary and time-varying analysis. Ph.D. thesis, Technion, Haifa, Israel.

Smith, W. E. 1956. Various optimizers for single-stage production. *Naval Res. Logist. Quart.* **3** 59–66.

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

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.

Walrand, J., C. Buyukkoc, P. Varaiya. 1985. The $c\mu$-rule revisited. *Adv. Appl. Probab.* **17** 237–238.

Whitt, W. 2004. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Sci.* **50**(10) 1449–1461.

Whitt, W. 2006. Fluid models for multiserver queues with abandonments. *Oper. Res.* **54**(1) 37–54.