On the asymptotic optimality of the $c\mu/\theta$ rule under ergodic cost*

Rami Atar Chanit Giat Nahum Shimkin

Department of Electrical Engineering
Technion–Israel Institute of Technology
Haifa 32000, Israel

January 12, 2010

Abstract

We consider an overloaded multi-server multi-class queueing model where customers may abandon while waiting to be served. For class i, service is provided at rate μ_i, and abandonment occurs at rate θ_i. In a many-server fluid regime, we show that prioritizing the classes in decreasing order of $c_i\mu_i/\theta_i$ asymptotically minimizes an ergodic holding cost, where c_i denotes the equivalent holding cost per unit time for class i.

1 Introduction

We consider a parallel server queueing model with I customer classes and multiple servers. Each server is capable of serving any one of the customers, and each customer has a single service requirement. Customers arrive according to renewal processes. The service time for a customer of class i is exponentially distributed with mean $1/\mu_i$. A class-i customer may abandon the system while waiting to be served, according to an exponential clock with mean $1/\theta_i$. A cost $\bar{c}_i \geq 0$ per unit time is incurred for holding a class-i customer in the queue, in addition to a penalty γ_i for each abandonment of a customer of that class. In this paper we shall be interested in minimizing the corresponding long-term average cost. Our focus will be on the overloaded system regime, where the total incoming work exceeds the service capacity. First, we argue that the cost is bounded below by the solution to a simple linear program. Then we specialize to a Markovian model (by letting arrivals be Poisson), and consider the system in a fluid limited regime where both the arrival rates and the number of servers grow without bound. Our main result shows that the lower bound alluded to above is asymptotically achieved by a static priority policy which prioritizes classes in decreasing order of $c_i\mu_i/\theta_i$, where $c_i = \bar{c}_i + \theta_i\gamma_i$. This result applies with respect to the long term *expected* average cost, as well

*Research supported in part by grants 2006379 and 2011234 from the United States–Israel Binational Science Foundation, grant 1349/08 from the Israel Science Foundation, and the Technion's fund for promotion of research.

as for the ergodic (sample-path long term average) cost. The lower bound alluded to above is also proved for a model with general service time distribution under a non-interruptible service assumption.

The policy described above, referred to as *the $c\mu/\theta$ rule*, was introduced in [1]. Both the results of [1] and those of the present paper establish optimality of this policy in the limit as the time (t) and the number of servers (n) grow without bound, where the difference lies in the order of the limits. The results of [1] state that, given $\varepsilon > 0$, one can find t such that, as n tends to infinity, the (sample path, average cost) performance of the proposed policy over the time interval $[0, t]$, is guaranteed to be optimal up to precision ε. The present paper, on the other hand, shows that for sufficiently large n, the average cost over the *infinite* time interval $[0, \infty)$ is optimal up to an arbitrary precision (depending on n). While the former approach emphasizes finite-time behavior, the latter addresses steady state.

The results of this paper require different mathematical tools from those of [1]. The lower bound (Propositions 2.1 and A.1 for exponential and general service time distribution, respund (Propositions 2.1 and A.1 for exponential and general service time distribution, respectively), is proved via a sample path analysis of the queueing process. The main tool for the upper bound (Theorem 2.2) is a Lyapunov function type argument (Lemma 3.1) that explicitly uses the form of the generator. Consequently, the extension of the upper bound beyond the Markovian setting is not straightforward.

For further references and discussion regarding the problem and suggested policy, the reader is referred to [1].

On our way to proving the main result, we analyze the Markovian model under an arbitrary priority policy, and establish the convergence of the fluid scale steady state distribution to that of the fluid model (Theorem 2.1). This result may be of interest on its own right. Fluid limits of queueing networks under priority disciplines have been considered in various works and textbooks. In [3, Section 9.3], a priority queue is considered as one of a large class of processes for which convergence to a fluid model holds. Further properties of priority queues in heavy traffic are analyzed in [4, Section 5.10]. Related results appear also in [2, Section 10]. These references are all concerned with convergence of fluid scale processes, uniformly on compact intervals of time, and therefore these results are not sufficient for the convergence of steady state distributions. One of the standard approaches to obtaining the latter is via the construction of a Lyapunov function, satisfying geometric ergodicity estimates, that are uniform both in n and t. Our proof of Theorem 2.1 is based on this approach.

The rest of this paper is organized as follows. In the next section we introduce the model with renewal arrivals, and state and prove a lower bound (Proposition 2.1). We then specialize to a Markovian setting, and state the result on fluid scale convergence of the steady state (Theorem 2.1), as well as our main result (Theorem 2.2), of asymptotic optimality of the $c\mu/\theta$ rule. We also provide a bound on the rate of convergence (Proposition 2.2) and an a.s. version of the upper and lower bounds (Proposition 2.3). Section 3 contains the proofs of Theorems 2.1 and 2.2 and Propositions 2.2 and 2.3. Finally, in the Appendix, we prove an analogue of Proposition 2.1 (lower bound) for general service time distribution and non-interruptible service.

Notation. For $x \in \mathbb{R}^I$, the ith element is denoted by x_i, and $\|x\| = \sum_{i=1}^I |x_i|$. For $x, y \in \mathbb{R}^I$, $x \cdot y = \sum x_i y_i$. For $x \in \mathbb{R}$, $x^+ = \max(x, 0)$, and $\mathbb{R}_+$ denotes the non-negative real line. For $X : \mathbb{R}_+ \to \mathbb{R}^k$, some positive integer k, we denote $\|X\|_T^* = \sup_{t \le T} \|X(t)\|$.

2 Model and results

The queueing model consists of a parallel server system with I classes of customers and n homogenous servers. It is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where expectation is denoted by $\mathbb{E}$. The arrivals are modeled as renewal processes A_i, where the inter-arrivals have finite mean $1/\lambda_i$. Service durations for class-i customers are i.i.d. exponential random variables with finite mean $1/\mu_i$. Namely, for a standard (rate 1) Poisson process $\tilde{D}_i$, the number of service completions of class-i jobs by time t is given as

$$D_i(t) = \tilde{D}_i\Big(\mu_i \int_0^t Z_i(s)ds\Big), \tag{1}$$

where $Z_i(t)$ denotes the number of class-i customers in service at time t. For a class-i customer, patience is assumed to be exponentially distributed, with mean $1/\theta_i$, where $\theta_i > 0$. This is modeled by introducing standard Poisson processes $\tilde{R}_i$, and assuming that the number of abandoning customers from buffer i by time t is given as

$$R_i(t) = \tilde{R}_i\Big(\theta_i \int_0^t Q_i(t)\Big), \tag{2}$$

where Q_i denotes the class-i queue length.

Let $X_i(t)$ denote the total number of class-i customers present in the system at time t. The initial conditions $X_1(0), X_2(0), \ldots, X_I(0)$ are assumed to be finite random variables. The $3I$ processes A_i, $\tilde{D}_i$ and $\tilde{R}_i$, and the initial condition $X(0) = (X_1(0), X_2(0), \ldots, X_I(0))$, referred to as the *stochastic primitives*, are further assumed to be mutually independent. The sample paths of A_i, $\tilde{D}_i$ and $\tilde{R}_i$ are assumed to be right-continuous.

The above processes clearly satisfy the following relations

$$X_i(t) = X_i(0) + A_i(t) - R_i(t) - D_i(t), \tag{3}$$

$$Q_i(t) = X_i(t) - Z_i(t) \ge 0, \tag{4}$$

$$Z_i(t) \ge 0, \quad \sum_{i=1}^I Z_i(t) \le n. \tag{5}$$

Service to customers may be interrupted by the system controller (so as to allow a customer of another class to be served), and resumed at a later time (provided that the customer has not abandoned in the mean time).

A control policy may be defined as a rule for allocating servers to customers, with Z understood to be the control variable. We will find it convenient to take a more abstract view, and identify any collection of processes

$$\pi = (D, R, X, Q, Z) \tag{6}$$

that comply with the description above as a *policy.* Let constants $\bar{c}_i \geq 0$ and $\gamma_i \geq 0$ be given, denoting holding cost per unit time, and abandonment cost, respectively, for class-i customers. For a policy $\pi = (D, R, X, Q, Z)$ consider the corresponding expected long term average costs

$$\underline{C}(\pi) = \liminf_{T\to\infty} \frac{1}{T}\,\mathbb{E}\Big[\int_0^T \bar{c}\cdot Q(t)dt + \gamma \cdot R(T)\Big]$$

$$\overline{C}(\pi) = \limsup_{T\to\infty} \frac{1}{T}\,\mathbb{E}\Big[\int_0^T \bar{c}\cdot Q(t)dt + \gamma \cdot R(T)\Big].$$

Using (2), it may be seen that $E(R(T)) = E(\int_0^T \theta_i Q_i(t)dt)$. We can therefore represent both cost components as holding costs with weights $c_i = \bar{c}_i + \theta_i\gamma_i$, namely,

$$\underline{C}(\pi) = \liminf_{T\to\infty} \frac{1}{T}\,\mathbb{E}\Big[\int_0^T c\cdot Q(t)dt\Big] \tag{7}$$

$$\overline{C}(\pi) = \limsup_{T\to\infty} \frac{1}{T}\,\mathbb{E}\Big[\int_0^T c\cdot Q(t)dt\Big]. \tag{8}$$

In what follows we shall always refer to the equivalent form (7)–(8) of the costs. We begin with a lower bound.

Proposition 2.1. *For any policy π,*

$$\underline{C}(\pi) \geq V_n\,,$$

where

$$V_n = \inf\Big\{c\cdot q : q \in \mathbb{R}_+^I,\ \theta_i q_i + \mu_i z_i = \lambda_i,\ i = 1,\ldots,I,\ z\in\mathbb{R}_+^I,\ \sum_i z_i \leq n\Big\}.$$

Remark 2.1. *The bound above is meaningful when the system is overloaded, in the sense that $\sum_i \lambda_i/\mu_i > 1$. When this conditions does not hold, it may be easily seen that $V_n = 0$, which is obtained for $z_i = \lambda_i/\mu_i$ and $q_i = 0$. Hence, our interest in this paper is in the overloaded regime.*

Proof. Fix a policy π. By Fatou's lemma, it suffices to prove that, with probability 1, $c{\cdot}\underline{Q} \geq V$, where

$$\underline{Q}_i = \liminf_T \frac{1}{T}\int_0^T Q_i(t)dt.$$

Note that $A_i(t)/t \to \lambda_i$ a.s., while $\tilde{D}_i(t)/t \to 1$ and $\tilde{R}_i(t)/t \to 1$ a.s. Note moreover that on the event Ω_1, where $t^{-1}\sum_i\int_0^t Q_i(s)ds$ grows without bound, there is nothing to prove. On the complement of this event, which we denote by Ω_2, the random variables $t^{-1}\int_0^t Q_i(s)ds$ and $t^{-1}\int_0^t Z_i(s)ds$ remain bounded. Consider a sequence along which these variables converge, and denote their respective limits as $\hat{Q}_i$ and $\hat{Z}_i$. It will be argued below that

$$t^{-1}X_i(t) \to 0 \qquad \text{a.s.} \tag{9}$$

Dividing by t in (3), and using (9), one has

$$\lambda_i = \lim_{t\to\infty} \Big[\frac{\tilde{R}_i(\theta_i \int_0^t Q_i(s)ds)}{\int_0^t Q_i(s)ds} \frac{\int_0^t Q_i(s)ds}{t} + \frac{\tilde{D}_i(\mu_i \int_0^t Z_i(s)ds)}{\int_0^t Z_i(s)ds} \frac{\int_0^t Z_i(s)ds}{t}\Big] = \theta_i \hat{Q}_i + \mu_i \hat{Z}_i, \quad (10)$$

a.s. on Ω_2. Note that on the event that $\int_0^t Q_i(s)ds$ remains bounded as $t \to \infty$, one cannot use the convergence $\tilde{R}_i(t)/t \to 1$ above to conclude (10). However, (10) is still valid, since in this case $\hat{Q}_i = 0$. A similar remark holds for $\hat{Z}_i$. Since the inequalities $\hat{Q}_i \geq 0$, $\hat{Z}_i \geq 0$ and $\sum \hat{Z}_i \leq n$ are clearly satisfied, it follows that $c \cdot \underline{Q} \geq V$.

It remains to prove (9). It is evident that $X_i(t)$, the number of class-i customers present at the system, can be upper-bounded by $\tilde{X}_i(t)$, the number of class-i customers that would be present at the system if no service at all would be applied to this class. Specifically, using appropriate coupling we have $X_i(t) \leq \tilde{X}_i(t)$ (a.s.). Now $\tilde{X}_i(t)$ is equivalent to the queue length process of an $G/M/\infty$ queue with service rate $\theta_i > 0$ per customer, which is well known to be stable. Consequently, $t^{-1}\tilde{X}(t) \to 0$ a.s., as $t \to \infty$. The same is implied for $X_i(t)$, for each i. This completes the proof. □

Our main result will be concerned with a specific policy, and show that it achieves the lower bound developed above, in an appropriate asymptotic sense. To present the result, we specialize to a Markovian setting. That is, we will assume throughout what follows that the arrival processes, A_i, are Poisson. We will refer to this setting as the *Markovian model.* The result will be concerned with a sequence of models, indexed by the number of servers, n. The parameters of the model, as well as the stochastic processes, will receive a superscript n, to denote their dependence on the parameter. An exception is the processes $\tilde{D}_i$ and $\tilde{R}_i$, which are still standard Poisson. Thus, for example, equation (1) defining the departure process, will now be written as

$$D_i^n(t) = \tilde{D}_i\Big(\mu_i^n \int_0^t Z_i^n(s)ds\Big),$$

and A_i^n will be a Poisson with rate λ_i^n. The parameters λ_i^n, μ_i^n and θ_i^n and initial conditions will be assumed to satisfy the following properties.

Assumption 2.1.

(i) There exist positive constants $\lambda_i, \mu_i, \theta_i$ such that, as $n \to \infty$,

$$\lambda_i^n/n \to \lambda_i, \quad \mu_i^n \to \mu_i, \quad ,\theta_i^n \to \theta_i, \qquad i = 1, 2, \ldots, I. \quad (11)$$

(ii) $\mathbb{E}[\|X^n(0)\|^2] < \infty$ for every n.

We first state a convergence result under a priority policy, that may be of interest on its own right. For the nth system, we denote by $\pi^{\mathrm{pr},n}$ the work conserving policy that gives preemptive priority to classes in increasing order of the labels. This means, in particular, that at any given time, if some server is idle then all buffers are empty. And if a customer of some class $i > 1$ is in service then no class-j customer is in the buffer, for any $j < i$. This is achieved by allowing interruption of service to customers, which are moved to the buffer until there is again an opportunity for them to be served again (however, it is possible for an interrupted customer to abandon). Denote

$$z^* = \Big(\frac{\lambda_1}{\mu_1}, \dots, \frac{\lambda_{i_0-1}}{\mu_{i_0-1}}, 1 - \sum_{j=1}^{i_0-1} \frac{\lambda_j}{\mu_j}, 0, \dots, 0\Big), \tag{12}$$

$$q^* = \Big(0, \dots, 0, \frac{\lambda_{i_0} - \mu_{i_0} z_{i_0}}{\theta_{i_0}}, \frac{\lambda_{i_0+1}}{\theta_{i_0+1}}, \dots, \frac{\lambda_I}{\theta_I}\Big), \tag{13}$$

where, with the convention $\sum_1^0 = 0$, $i_0 = \max\{i \in [1, I+1] : \sum_{j=1}^{i-1} \frac{\lambda_j}{\mu_j} < 1\}$. Let $x^* = q^* + z^*$.

Theorem 2.1. *Consider the Markovian model, and let Assumption 2.1 hold. Then, under policy* $\pi^{\mathrm{pr},n}$,

$$\lim_{n\to\infty} \limsup_{t\to\infty} \mathbb{E}[\|n^{-1}X^n(t) - x^*\|^2] = 0, \tag{14}$$

and

$$\lim_{n\to\infty} \limsup_{t\to\infty} \mathbb{E}[\|n^{-1}Q^n(t) - q^*\|^2] = 0. \tag{15}$$

For the nth system, we denote the costs of (7) by

$$\underline{C}^n(\pi) = \liminf_{T\to\infty} \frac{1}{T} \mathbb{E}\Big[\int_0^T c \cdot Q^n(t) dt\Big],$$
$$\overline{C}^n(\pi) = \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}\Big[\int_0^T c \cdot Q^n(t) dt\Big],$$

where $c_i \geq 0$ do not depend on n. It is immediate from Proposition 2.1 that, under any sequence π^n of policies,

$$\liminf_{n\to\infty} n^{-1} \underline{C}^n(\pi^n) \geq V_1, \tag{16}$$

where

$$V_1 = \inf\Big\{c \cdot q : q \in \mathbb{R}_+^I,\ \theta_i q_i + \mu_i z_i = \lambda_i,\ i = 1, \dots, I,\ z \in \mathbb{R}_+^I,\ \sum z_i \leq 1\Big\}. \tag{17}$$

The proposed policy, referred to as the *preemptive* $c\mu/\theta$ *priority rule* [1], will be denoted by $\pi^{*,n}$ for the nth system. $\pi^{*,n}$ is the work conserving policy that gives preemptive priority to classes in decreasing order of the quantities $c_i\mu_i/\theta_i$. In other words, $\pi^{*,n}$ is identical to $\pi^{\mathrm{pr},n}$ under re-labeling of the classes according to

$$\frac{c_1\mu_1}{\theta_1} \geq \frac{c_2\mu_2}{\theta_2} \geq \cdots \geq \frac{c_I\mu_I}{\theta_I}. \tag{18}$$

As a corollary of Theorem 2.1, we obtain our main result.

Theorem 2.2. *Consider the Markovian model, and let Assumption 2.1 hold. Then*

$$\limsup_{n\to\infty} n^{-1} \overline{C}^n(\pi^{*,n}) \leq V_1.$$

In view of (16), the above result expresses asymptotic optimality of the proposed policy.

One can give a bound on the rate of convergence in the above result. Let $r_n = \|\theta^n - \theta\| + \|\mu^n - \mu\| + \|n^{-1}\lambda^n - \lambda\|$.

Proposition 2.2. *Let the conditions of Theorem 2.2 hold. Then, for every n,*

$$V_1 - c_0 r_n \le n^{-1}\underline{C}^n(\pi^{*,n}) = n^{-1}\overline{C}^n(\pi^{*,n}) \le V_1 + c_0(n^{-1} + r_n)^{1/2},$$

where c_0 is a constant not depending on n.

Remark 2.2. *When the n-th system parameters are chosen nominally at $\theta^n = \theta$, $\mu^n = \mu$ and $\lambda_n = n\lambda$, we obtain $r_n = 0$. In that case the implied convergence rate of the average cost is $O(n^{-1/2})$.*

Finally, we state a sample-path version of Proposition 2.1 and Theorem 2.2, in terms of the ergodic cost function.

Proposition 2.3. *Under any sequence of policies π^n,*

$$\liminf_{n\to\infty} \liminf_{T\to\infty} \frac{1}{T}\int_0^T c \cdot Q^n(t)dt \ge V_1, \qquad a.s.$$

Moreover, let the assumptions of Theorem 2.2 hold. Then, under $\pi^{*,n}$,

$$\limsup_{n\to\infty} \limsup_{T\to\infty} \frac{1}{T}\int_0^T c \cdot Q^n(t)dt \le V_1 \qquad a.s.$$

3 Proofs

Throughout this section, the ordering (18) of the class indices is assumed, and, unless indicated otherwise, all stochastic processes are specified under $\pi^{*,n}$ (equivalently, $\pi^{\mathrm{pr},n}$). The linear program (17) can easily be seen to be solved by (12) and (13) (see [1] for more details). Moreover, since q^* and z^* attain the infimum (17), we have

$$V_1 = c \cdot q^*, \tag{19}$$

and

$$\theta_i q_i^* + \mu_i z_i^* = \lambda_i, \qquad i = 1, 2, \ldots, I. \tag{20}$$

The analysis of the policy $\pi^{*,n}$ will be based on the fact that the process X^n, under $\pi^{*,n}$, is Markovian. It is easy to see that the infinitesimal generator of X^n is given by

$$\begin{aligned}\mathcal{L}^n f(x) = &\sum_{i=1}^{I} \lambda_i^n \left(f(x + e_i) - f(x)\right) \\ &+ \sum_{i=1}^{I} \mu_i^n \mathbf{Z}_i^n(x) \left(f(x - e_i) - f(x)\right) \\ &+ \sum_{i=1}^{I} \theta_i^n \mathbf{Q}_i^n(x) \left(f(x - e_i) - f(x)\right), \qquad x \in \mathbb{Z}_+^I,\end{aligned} \tag{21}$$

where $\mathbf{Z}^n, \mathbf{Q}^n : \mathbb{R}^I_+ \to \mathbb{R}^I_+$ are defined as

$$\mathbf{Z}^n_i(x) = x_i \wedge \Big(n - \sum_{j=1}^{i-1} x_j\Big)^+, \qquad \mathbf{Q}^n_i(x) = \Big[x_i - \Big(n - \sum_{j=1}^{i-1} x_j\Big)^+\Big]^+. \tag{22}$$

We let $f^n(x) = \sum_{i=1}^I \gamma_i \, (x_i - x_i^* n)^2$, where $\gamma_i > 0$ are constants, not depending on n or x, to be determined later. Our main estimate will be the following.

Lemma 3.1. *Under the assumptions of Theorem 2.2, the constants $\gamma_i > 0$ can be chosen so that*

$$\mathcal{L}^n f^n(x) \le -a f^n(x) + a_1 \|x\| + \delta_n n^2, \qquad x \in \mathbb{Z}^I_+,\ n \ge n_0, \tag{23}$$

where $a > 0$, a_1 and n_0 are constants not depending on x or n, and δ_n is a sequence that is independent of x and converges to zero.

Proof of Lemma 3.1. Notice first that

$$\mu_i^n \mathbf{Z}_i^n(x) + \theta_i^n \mathbf{Q}_i^n(x) = \mu_i^n x_i + (\theta_i^n - \mu_i^n)\mathbf{Q}_i^n(x)\,.$$

Using this in (21), along with the identities $(a \pm 1)^2 - a^2 = \pm 2a + 1$ yields

$$\begin{aligned} \mathcal{L}^n f^n(x) = &\sum_{i=1}^I 2\gamma_i \Big(x_i - x_i^* n + \frac{1}{2}\Big)\lambda_i^n \\ &- \sum_{i=1}^I 2\gamma_i \Big(x_i - x_i^* n - \frac{1}{2}\Big)\Big[\mu_i^n x_i + (\theta_i^n - \mu_i^n)\mathbf{Q}_i^n(x)\Big]. \end{aligned} \tag{24}$$

Let $Y(x, n) = x - x^* n$. To simplify the notation, we write Y for $Y(x, n)$. Note that with this notation, $f^n(x) = \sum \gamma_i Y_i^2$. With C_1 a constant not depending on n or x (but depending on $\{\gamma_i\}$), we have

$$\begin{aligned} \mathcal{L}^n f^n(x) &\le C_1(n + \|x\|) + 2\sum_{i=1}^I \gamma_i Y_i \Big[\lambda_i^n - \mu_i^n x_i - (\theta_i^n - \mu_i^n)\mathbf{Q}_i^n(x)\Big] \\ &= C_1(n + \|x\|) - 2\sum_{i=1}^I \gamma_i \mu_i^n Y_i^2 \\ &\quad + 2\sum_{i=1}^I \gamma_i Y_i \Big[\lambda_i^n - \mu_i^n x_i^* n - (\theta_i^n - \mu_i^n)\mathbf{Q}_i^n(x)\Big]. \end{aligned} \tag{25}$$

Recall that $x^* = z^* + q^*$ and that, by (20), $\theta_i q_i^* + \mu_i z_i^* = \lambda_i$. Therefore

$$\begin{aligned} n^{-1}\lambda_i^n - \mu_i^n x_i^* &= n^{-1}\lambda_i^n - \mu_i^n (z_i^* + q_i^*) \\ &= (\theta_i^n - \mu_i^n) q_i^* + \varepsilon_n. \end{aligned}$$

Here, $\varepsilon_n \to 0$, by Assumption 2.1. Thus, the last term on the r.h.s. of (25) is given by

$$\begin{aligned}2\sum \gamma_i Y_i[n\varepsilon_n + (\theta_i^n - \mu_i^n)(nq_i^* - \mathbf{Q}_i^n(x))]\\ = 2\sum \gamma_i Y_i[n\varepsilon_n + (\theta_i^n - \mu_i^n)(\mathbf{Q}_i^n(nx^*) - \mathbf{Q}_i^n(x))],\end{aligned}$$

where we used the equality $nq_i^* = \mathbf{Q}_i^n(nx^*)$, that can be directly verified using the explicit form of q^*, z^* and x^*. By the definition of $\mathbf{Q}^n$, and the fact that, for any $a, b \in \mathbb{R}$ there exists $\rho \in [0,1]$ such that $a^+ - b^+ = \rho(a-b)$, it is not hard to see that, for any $n \in \mathbb{N}$ and $x, \tilde{x} \in \mathbb{R}_+^I$, one has

$$\mathbf{Q}_i^n(x) - \mathbf{Q}_i^n(\tilde{x}) = \rho(x_i - \tilde{x}_i) + \eta\sum_{j=1}^{i-1}(\tilde{x}_j - x_j), \tag{26}$$

where $\rho, \eta \in [0,1]$ may depend on n, x and $\tilde{x}$. Using this property, we can find functions $\rho_i, \eta_i : \mathbb{R}^I \to [0,1]$, that may depend on n, such that, with $\delta_n = |\varepsilon_n|$,

$$\begin{aligned}\mathcal{L}^n f^n(x) \le C_1(n + \|x\|) + C_2\|Y\|n\delta_n\\ - 2\sum_{i=1}^I \gamma_i\left[(1-\rho_i(x))\mu_i^n + \rho_i(x)\theta_i^n\right]Y_i^2\\ + 2\sum_{i=1}^I \gamma_i(\theta_i^n - \mu_i^n)\eta_i(x)Y_i\sum_{j=1}^{i-1}Y_j.\end{aligned}$$

Note that, for every $\rho \in [0,1]$, $(1-\rho)\mu_i^n + \rho\theta_i^n \ge \min(\theta_i^n, \mu_i^n) \ge \frac{1}{2}\min(\theta_i, \mu_i) =: m_i > 0$, provided that n is sufficiently large. Thus,

$$\begin{aligned}\mathcal{L}^n f^n(x) \le C_1(n + \|x\|) + C_2\|Y\|n\delta_n - 2\sum_{i=1}^I \gamma_i m_i Y_i^2\\ + 2\sum_{i=1}^I \gamma_i(\theta_i^n - \mu_i^n)\eta_i(x)Y_i\sum_{j=1}^{i-1}Y_j.\end{aligned}$$

Denote $A = \sup_n |\theta_i^n - \mu_i^n| < \infty$. Using the inequality $xy \le \frac{1}{2}bx^2 + \frac{1}{2}b^{-1}y^2$, which holds for $x, y \in \mathbb{R}$ and $b > 0$, we bound the last term on the above display by

$$B^n(x) := 2A\sum_{i=1}^I \gamma_i\Big[b_iY_i^2 + b_i^{-1}\Big(\sum_{j=1}^{i-1}|Y_j|\Big)^2\Big] \le 2A\sum_{i=1}^I\Big[\gamma_i b_i Y_i^2 + \gamma_i b_i^{-1}C_3\sum_{j=1}^{i-1}Y_j^2\Big],$$

where C_3 depends only on I. Now choose b_i so that $2Ab_i = m_i/2$, $i = 1, 2, \dots, I$. Next, determine γ_i inductively, as follows. Let $\gamma_1 = 1$. For $i = 2, 3, \dots, I$, let γ_i (depending on $\gamma_1, \dots, \gamma_{i-1}$) be determined by

$$2A\gamma_i b_i^{-1}C_3 = \frac{1}{2I}\min_{j \le i-1}\gamma_j m_j.$$

Then

$$B^n(x) \le \sum_{i=1}^I\Big[\frac{1}{2}\gamma_i m_i Y_i^2 + \frac{1}{2I}\sum_{j=1}^{i-1}\gamma_j m_j Y_j^2\Big] \le \sum_{i=1}^I \gamma_i m_i Y_i^2.$$

Letting $m = \min_i m_i > 0$, we obtain

$$\begin{aligned}\mathcal{L}^n f^n(x) &\le C_1(n + \|x\|) + C_2\|Y\|n\delta_n - \sum_{i=1}^{I} \gamma_i m_i Y_i^2 \\ &\le C_1(n + \|x\|) + C_2\delta_n\|Y\|^2 + C_2\delta_n n^2 - m\sum_{i=1}^{I} \gamma_i Y_i^2 \\ &\le C_1(n + \|x\|) + C_2\delta_n n^2 - \frac{m}{2}\sum_{i=1}^{I} \gamma_i Y_i^2,\end{aligned}$$

for all sufficiently large n. Thus

$$\mathcal{L}^n f^n(x) \le C_1(n + \|x\|) + C_2\delta_n n^2 - \frac{m}{2} f^n(x).$$

This completes the proof. □

Lemma 3.2. *Let the assumptions of Theorem 2.2 hold. Then, for some positive constants $\bar{C}$ and $\widetilde{C}$, not depending on n or t, $\mathbb{E}[\|X^n(t)\|] \le \mathbb{E}[\|X^n(0)\|]e^{-\widetilde{C}t} + \bar{C}n$, for all $t \ge 0$ and all sufficiently large n.*

proof. In this proof, we remove the dependence of the processes on n from the notation. By (1), (2) and (3), recalling that under the assumptions of Theorem 2.2 the setup is Markovian, we have

$$\mathbb{E}[X_i(t)] = \mathbb{E}[X_i(0)] + \lambda_i^n t - \theta_i^n \int_0^t \mathbb{E}[Q_i(s)]ds - \mu_i^n \int_0^t \mathbb{E}[Z_i(s)]ds.$$

Hence $\xi_i(t) := \mathbb{E}[X_i(t)]$ is differentiable, and, denoting $m_i = \min(\theta_i, \mu_i)/2$, we have

$$\frac{d\xi_i(t)}{dt} \le 2n\lambda_i - m_i\mathbb{E}[Q_i(t) + Z_i(t)] = 2n\lambda_i - m_i\xi_i(t),$$

provided n is sufficiently large, where we used Assumption 2.1 and then (4). Hence for $\xi(t) = \sum_i \xi_i(t)$ we have

$$\frac{d\xi(t)}{dt} \le Ln - M\xi(t), \qquad \xi(0) = \mathbb{E}[\|X(0)\|],$$

where L and M are positive constants not depending on n or t. By standard comparison of solutions to ordinary differential equations, $\xi(t) \le \nu(t)$ must hold for all $t \ge 0$, where ν solves

$$\frac{d\nu(t)}{dt} = Ln - M\nu(t), \qquad \nu(0) = \mathbb{E}[\|X(0)\|],$$

that is,

$$\xi(t) \le \mathbb{E}[\|X(0)\|]\exp\{-Mt\} + \frac{Ln}{M}(1 - \exp\{-Mt\}), \qquad t \ge 0.$$

This completes the proof. □

Proof of Theorem 2.1. Since X^n is Markovian with generator $\mathcal{L}^n$, the process

$$f(X^n(t)) - \int_0^t \mathcal{L}^n f(X^n(s))ds$$

is a martingale whenever f is a bounded function on $\mathbb{Z}_+^I$. It is easy to see by (3) that $X_i^n(t) \le X_i^n(0) + A_i^n(t)$, and since the second moment of $X^n(0)$ is assumed to be finite, and clearly $\mathbb{E}\sup_{t\le T}[\|A^n(t)\|^2] < \infty$ for every n and T, the martingale property holds also for the quadratic function f^n. Hence

$$\mathbb{E}f^n(X^n(t)) = \mathbb{E}f(X^n(0)) + \mathbb{E}\int_0^t \mathcal{L}^n f^n(X^n(s))ds. \tag{27}$$

Let us prove that

$$\limsup_{t\to\infty} \mathbb{E}[\|n^{-1}X^n(t) - x^*\|^2] \le \bar{\delta}_n, \tag{28}$$

where $\bar{\delta}_n$ is a sequence converging to zero. To this end, note by (27) that $\mathbb{E}f^n(X^n(t))$ is differentiable with respect to t. Denote $Y^n(t) := n^{-2}\mathbb{E}f^n(X^n(t))$. Then $Y^n(t) < \infty$ for every t and n. Moreover, dividing by n^2 in (27) and using Lemma 3.1, we have

$$\frac{dY^n(t)}{dt} \le -aY_t^n + \frac{a_1}{n^2}\mathbb{E}[\|X^n(t)\|] + \delta_n, \qquad t \ge 0.$$

By Lemma 3.2, for every sufficiently large n there exists $T_n < \infty$ such that $E[\|X^n(t)\|] \le 2\bar{C}n$, $t \ge T_n$. Hence, denoting $\tilde{\delta}_n = 2\bar{C}a_1n^{-1} + \delta_n$, for some n_0 and all $n \ge n_0$,

$$\frac{dY^n(t)}{dt} \le -aY_t^n + \tilde{\delta}_n, \qquad t \ge T_n.$$

By the comparison principle for solutions of differential inequalities, Y_t^n is bounded above, on $[T_n, \infty)$, by the solution y to the differential equation

$$\frac{dy}{dt} = -ay + \tilde{\delta}_n, \quad t \ge T_n, \qquad y(T_n) = Y^n(T_n).$$

Hence, for some constant C_1, for all $n \ge n_0$ and $t \ge T_n$,

$$\mathbb{E}[\|n^{-1}X^n(t) - x^*\|^2] \le C_1Y^n(t) \le C_1y(t) \le C_1Y^n(T_n)\exp\{-a(t-T_n)\} + C_1a^{-1}\tilde{\delta}_n.$$

This proves (28), hence follows (14).

To establish (15), note that we have for every n,

$$Q^n(t) = \mathbf{Q}^n(X^n(t)), \qquad t \ge 0.$$

With the notation (22), the map $\mathbf{Q}^1 : \mathbb{R}_+^I \to \mathbb{R}_+^I$, is given by

$$\mathbf{Q}_i^1(x) = \Big[x_i - \Big(1 - \sum_{j=1}^{i-1} x_j\Big)^+\Big]^+,$$

and we have $n^{-1}Q^n(t) = \mathbf{Q}^1(n^{-1}X^n(t))$, $t \ge 0$. Noting that $\mathbf{Q}^1(x^*) = q^*$, using the global Lipschitz continuity of $\mathbf{Q}^1$, we have by (28),

$$\limsup_{t\to\infty} \mathbb{E}[\|n^{-1}Q^n(t) - q^*\|^2] \le C_4\bar{\delta}_n, \tag{29}$$

for some C_4 depending only on I. This implies (15), and hence completes the proof. □

Proof of Theorem 2.2. Keeping the notation of the proof of Theorem 2.1, the inequality (29) implies that

$$\limsup_{t\to\infty} \mathbb{E}[\|n^{-1}Q^n(t) - q^*\|] \le (C_4\bar{\delta}_n)^{1/2},$$

hence

$$\frac{\overline{C}^n(\pi^{*,n})}{n} = \limsup_{T\to\infty} \frac{1}{T}\int_0^T \frac{c\cdot\mathbb{E}[Q^n(t)]}{n}dt \le c\cdot q^* + |c|(C_4\bar{\delta}_n)^{1/2}.$$

Sending $n \to \infty$, $\limsup_{n\to\infty} n^{-1}\overline{C}^n(\pi^{*,n}) \le c\cdot q^* = V_1$, where the last equality follows by (19). □

Proof of Proposition 2.2. Recall the definition of V_1 in (17), and write $V_1(\theta,\mu,\lambda)$ to denote its dependence on the parameters. It follows from Proposition 2.1 that, for any n and any policy π^n for the nth system, $n^{-1}\underline{C}^n(\pi^n) \ge V_1(\theta^n,\mu^n,\frac{\lambda^n}{n})$. It is easy to see that V_1 is Lipschitz continuous w.r.t. the three parameters. Hence follows the lower bound stated in the proposition.

For the upper bound, a review of the proofs of Theorems 2.1 and 2.2, shows that the cost $n^{-1}\overline{C}^n(\pi^{*,n})$ is bounded above by $V_1 + C_5(\tilde{\delta}_n)^{1/2} \le V_1 + C_6(n^{-1}+\delta_n)^{1/2}$, where C_5, C_6 are constants, $\tilde{\delta}_n$ is as in the proof of Theorem 2.1, and δ_n is as in the proof of Lemma 3.1. By the proof of Lemma 3.1, δ_n is bounded by a constant times r_n. This completes the proof. □

Proof of Proposition 2.3. The lower bound follows directly from the proof of Proposition 2.1, which establishes the inequality in an a.s. sense.

For the upper bound, in view of Theorem 2.2, it suffices to show that, for every n, under $\pi^{*,n}$,

$$\liminf_{T\to\infty}\frac{1}{T}\int_0^T c\cdot Q^n(t)dt = \liminf_{T\to\infty}\frac{1}{T}\mathbb{E}\Big[\int_0^T c\cdot Q^n(t)dt\Big], \qquad \text{a.s.}$$

This property follows from ergodicity of the Markov chain X^n, which can be verified by standard techniques (such as [3, Theorem 8.6]). We omit the details. □

A Appendix

We will argue that the lower bound stated in Proposition 2.1 is valid for general service time distribution, under a non-interruptible service assumption. We shall thus model service durations for class-i customers as i.i.d. positive random variables with finite mean $1/\mu_i$. To this end, assume we are given nI renewal processes $\tilde{D}_{i,k}$, $i = 1,2,\dots,I$, $k = 1,2,\dots,n$, that are mutually independent and independent of the other stochastic primitives. For each of these processes, the inter-event time has mean 1. Assume that the number of i-class jobs that server k completes by time t is given as

$$D_{i,k}(t) = \tilde{D}_{i,k}\Big(\mu_i\int_0^t Z_{i,k}(s)ds\Big), \tag{30}$$

where the process $Z_{i,k}$ takes values in $\{0,1\}$, and $Z_{i,k}(t)=1$ if a class-i customer is served by server k at time t. The relations (3)–(5) are still valid, and in addition, for each i, we have

$$\sum_{k=1}^{N} Z_{i,k} = Z_i, \qquad \sum_{k=1}^{N} D_{i,k} = D_i. \tag{31}$$

Unlike Section 2, we shall assume here that interruption of service is not possible. Thus whenever a server is assigned a new customer, it serves the customer until completion of the service requirement. The reason we do not allow interruption is that an interrupted customer may return to a different server, or even abandon the system before ever returning to service, in which cases (30) is not a valid description of the service process under an interruptible service policy (except in the exponential case), and the state description becomes more involved.

Proposition A.1. *For any policy π,*

$$\underline{C}(\pi) \ge V_n\,,$$

where V_n is as in Proposition 2.1.

Proof. The proof follows closely that of Proposition 2.1. We will only indicate where the argument differs. For each k, $\tilde{D}_{i,k}(t)/t \to 1$ a.s. Keeping the notation from the proof of Proposition 2.1, on the event Ω_2, the random variables $t^{-1}\int_0^t Q_i(s)ds$, $t^{-1}\int_0^t Z_{i,k}(s)ds$, $k=1,2,\dots,N$, and $t^{-1}\int_0^t Z_i(s)ds$, remain bounded, and, on a convergent sequence (as $t\to\infty$), we denote their respective limits as $\hat{Q}_i$, $\hat{Z}_{i,k}$ and $\hat{Z}_i$. The argument for $t^{-1}X_i(t)\to 0$ a.s. holds precisely as in the proof of Proposition 2.1. Thus dividing by t in (3), and using (31),

$$\begin{aligned}\lambda_i &= \lim_{t\to\infty}\Big[\frac{\tilde{R}_i(\theta_i\int_0^t Q_i(s)ds)}{\int_0^t Q_i(s)ds}\,\frac{\int_0^t Q_i(s)ds}{t} + \sum_{k=1}^{N}\frac{\tilde{D}_{i,k}(\mu_i\int_0^t Z_{i,k}(s)ds)}{\int_0^t Z_{i,k}(s)ds}\,\frac{\int_0^t Z_{i,k}(s)ds}{t}\Big]\\ &= \theta_i\hat{Q}_i + \sum_{k=1}^{N}\mu_i\hat{Z}_{i,k} = \theta_i\hat{Q}_i + \mu_i\hat{Z}_i,\end{aligned}$$

a.s. on Ω_2. The proof is completed as that of Proposition 2.1. □

Acknowledgement. The authors wish to thank Gennady Shaikhet for pointing out an error in an earlier version of this manuscript.

References

[1] R. Atar, C. Giat, and N. Shimkin. The $c\mu/\theta$ rule for many-server queues with abandonment. April 2009, to appear in *Operations Research.*

[2] S. Meyn. *Control Techniques for Complex Networks.* Cambridge University Press, Cambridge, 2008

[3] P. Robert. *Stochastic Networks and Queues.* Springer-Verlag, Berlin, 2003.

[4] W. Whitt. *Stochastic-Process Limits.* Springer Series in Operations Research. Springer-Verlag, New York, 2002.