Queueing Syst (2011) 67: 127–144
DOI 10.1007/s11134-010-9206-x

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

Rami Atar · Chanit Giat · Nahum Shimkin

Received: 12 January 2010 / Revised: 17 October 2010 / Published online: 29 December 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.

Keywords Queues with abandonment · Ergodic cost · Multi-class queues · Asymptotic optimality · Fluid limits

Mathematics Subject Classification (2000) 68M20 · 90B36 · 93E20

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 independent 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

Research supported in part by grants 2006379 and 2008466 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.

R. Atar (✉) · C. Giat · N. Shimkin
Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel
e-mail: atar@ee.technion.ac.il

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 (while stability is maintained due to abandonment). 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 limit 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 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, 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 possible 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, Sect. 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, Sect. 5.10]. Related results appear also in [2, Sect. 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 almost-sure 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 $X : \mathbb{R}_+ \to \mathbb{R}^I$, we denote $\|X\|_T^* = \sup_{t \le T} \|X(t)\|$.

2 Model and main results

2.1 Model

The queueing model consists of a parallel server system with I classes of customers and n homogeneous 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

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

where $Z_i(t)$ denotes the number of class-i customers in service at time t. Customers waiting to be served are said to be in the queue. While customers are in the queue they may lose patience and abandon the system. For a class-i customer, the 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

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

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

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

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

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

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 meantime). A preempted customer waiting to be re-assigned a server is considered to be in the queue, and may abandon in accordance with the model (2). One can interpret this model as if the patience time is determined once for each customer, and the clock runs only at times when the customer is in the queue. An alternative interpretation is that the patience time is drawn anew each time the customer returns to the queue; as is well known, these two interpretations are equivalent due to the memoryless property of the exponential distribution.

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

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

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

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

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

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

In what follows we shall always refer to the equivalent form (7)–(8) of the costs.

2.2 A lower bound

Proposition 2.1 *For any policy* π,

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

where

$$V_n = \inf\left\{ c \cdot q\colon\ q \in \mathbb{R}_+^I,\ \theta_i q_i + \mu_i z_i = \lambda_i,\ i = 1, \ldots, I,\ z \in \mathbb{R}_+^I,\ \sum_i z_i \le n \right\}. \quad (9)$$

Remark 2.1 The bound above is meaningful when the system is overloaded, in the sense that $\sum_i \lambda_i/\mu_i > n$. Otherwise, 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. We note that queue stability is maintained in this regime as well, due to the non-zero abandonment rates of all classes.

Proof Fix a policy π. By Fatou's lemma, it suffices to prove that, with probability 1, $U \ge V_n$, where

$$U = \liminf_{t\to\infty} \frac{1}{t} \int_0^t c \cdot Q(s)\, ds.$$

Let $\{t_k\}_{k\in\mathbb{N}}$ be a (random) sequence increasing to infinity, for which $t_k^{-1} \int_0^{t_k} c \cdot Q(s)\, ds \to U$. Note by (5) that the random variables $t^{-1} \int_0^t Z_i(s)\, ds$ remain bounded, and choose a subsequence of $\{t_k\}$ along which each of these random variables converges. Denote the respective limits as $\hat{Z}_i$ and note that they take values in $[0, n]$. By choosing a further subsequence it can be ensured, in addition, that each of the r.v.'s ${t_k}^{-1} \int_0^{t_k} Q_i(s)\, ds$ converges to a random variable taking values in $[0, \infty]$. These respective limits are denoted by $\hat{Q}_i$. It will also be argued below that, for each i,

$$t^{-1} X_i(t) \to 0 \quad \text{as } t \to \infty, \qquad \text{a.s.} \quad (10)$$

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. Dividing by t in (3), and using (10), it follows that

$$\begin{aligned} \lambda_i &= \lim_{t\to\infty} \left[\frac{\tilde{R}_i(\theta_i \int_0^t Q_i(s)\, ds)}{\int_0^t Q_i(s)\, ds} \frac{\int_0^t Q_i(s)\, ds}{t} + \frac{\tilde{D}_i(\mu_i \int_0^t Z_i(s)\, ds)}{\int_0^t Z_i(s)\, ds} \frac{\int_0^t Z_i(s)\, ds}{t} \right] \\ &= \theta_i \hat{Q}_i + \mu_i \hat{Z}_i, \end{aligned} \quad (11)$$

where the limit is taken along the subsequence. Note that on the event that $\int_0^{t_k} Q_i(s)\, ds$ remains bounded as $k \to \infty$, one cannot use the convergence $\tilde{R}_i(t)/t \to 1$ to conclude (11). However, (11) is still valid, since in this case $\hat{Q}_i = 0$. A similar remark holds for $\hat{Z}_i$. Since we assumed $\theta_i > 0$, it follows from (11) that each of the $\hat{Q}_i$ is a.s. finite. The inequalities $\hat{Q}_i \ge 0$, $\hat{Z}_i \ge 0$ and $\sum \hat{Z}_i \le n$ are clearly satisfied a.s. Thus by (9) and (11), $U = c \cdot \hat{Q} \ge V_n$ a.s.

It remains to prove (10). In the appendix, we provide a proof of this fact, based on comparison of the process X_i to the $G/M/\infty$ queue and an argument that an analogous property holds for the latter model. This completes the proof. □

2.3 Asymptotic results as $n \to \infty$

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 in what follows that the arrival processes, A_i, are Poisson. We 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, (1) defining the departure process, will now be written as

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

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 the initial conditions $X^n(0)$ 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 $n \to \infty$,

$$\lambda_i^n/n \to \lambda_i, \qquad \mu_i^n \to \mu_i, \qquad \theta_i^n \to \theta_i, \quad i = 1, 2, \ldots, I. \tag{12}$$

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

Note that the lower bound from the previous subsection can be applied to this setting as follows. From (9), it follows that

$$\frac{V_n}{n} = \inf\left\{c \cdot q\colon\ q \in \mathbb{R}_+^I,\ \theta_i^n q_i + \mu_i^n z_i = \lambda_i^n/n,\ i = 1, \ldots, I,\ z \in \mathbb{R}_+^I,\ \sum_i z_i \le 1\right\}.$$

Hence by Proposition 2.1, under any sequence π^n of policies,

$$\liminf_{n\to\infty} n^{-1} \underline{C}^n(\pi^n) \ge V_1, \tag{13}$$

where

$$V_1 = \inf\left\{c \cdot q\colon\ q \in \mathbb{R}_+^I,\ \theta_i q_i + \mu_i z_i = \lambda_i,\ i = 1, \ldots, I,\ z \in \mathbb{R}_+^I,\ \sum z_i \le 1\right\}. \tag{14}$$

To the end of showing that this lower bound is asymptotically achievable, we first present a convergence result under any priority policy. 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 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 to wait until there is an opportunity for them to be served again, or otherwise abandon.

Under the priority policy, the infinitesimal generator of X^n is given by

$$\begin{aligned}\mathcal{L}^n f(x) = &\sum_{i=1}^{I} \lambda_i^n \big(f(x+e_i) - f(x)\big) \\ &+ \sum_{i=1}^{I} \mu_i^n \mathbf{Z}_i^n(x) \big(f(x-e_i) - f(x)\big) \\ &+ \sum_{i=1}^{I} \theta_i^n \mathbf{Q}_i^n(x) \big(f(x-e_i) - f(x)\big), \quad x \in \mathbb{Z}_+^I, \end{aligned} \tag{15}$$

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

$$\mathbf{Z}_i^n(x) = x_i \wedge \left(n - \sum_{j=1}^{i-1} x_j\right)^+, \qquad \mathbf{Q}_i^n(x) = \left[x_i - \left(n - \sum_{j=1}^{i-1} x_j\right)^+\right]^+. \tag{16}$$

In the expression for the generator, the first, second and third terms correspond to transitions according to arrival, service and abandonment, respectively. The absolute priority is reflected in the expressions for $\mathbf{Z}^n$ and $\mathbf{Q}^n$, given in terms of the quantity $(n - \sum_{j=1}^{i-1} x_j)^+$. This quantity represents the number of servers available to serve class i, taking into account that $\sum_{j=1}^{i-1} x_j$ servers are occupied with higher priority customers.

Denote

$$z^* = \left(\frac{\lambda_1}{\mu_1}, \ldots, \frac{\lambda_{i_0-1}}{\mu_{i_0-1}}, 1 - \sum_{j=1}^{i_0-1} \frac{\lambda_j}{\mu_j}, 0, \ldots, 0\right), \tag{17}$$

$$q^* = \left(0, \ldots, 0, \frac{\lambda_{i_0} - \mu_{i_0} z_{i_0}}{\theta_{i_0}}, \frac{\lambda_{i_0+1}}{\theta_{i_0+1}}, \ldots, \frac{\lambda_I}{\theta_I}\right), \tag{18}$$

where $i_0 = \max\{i \in [1, I+1] : \sum_{j=1}^{i-1} \frac{\lambda_j}{\mu_j} < 1\}$, with the convention $\sum_1^0 = 0$. Let $x^* = q^* + z^*$. The vectors z^*, q^* and x^* are simply the equilibrium point of the underlying fluid model [1]

$$\dot{x}_i = \lambda_i - \mu_i z_i - \theta_i q_i, \quad z_i = \mathbf{Z}_i^1(x),\ q_i = \mathbf{Q}_i^1(x).$$

The following result relates them to the probabilistic model.

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

$$\lim_{n\to\infty} \limsup_{t\to\infty} \mathbb{E}\big[\big\|n^{-1} X^n(t) - x^*\big\|^2\big] = 0, \tag{19}$$

and

$$\lim_{n\to\infty} \limsup_{t\to\infty} \mathbb{E}\big[\big\|n^{-1}Q^n(t) - q^*\big\|^2\big] = 0. \tag{20}$$

Remark 2.2 Under policy $\pi^{\mathrm{pr},n}$, the process X^n is an ergodic Markov chain. Denote by $\mathbb{P}_s^{\mathrm{pr},n}$ its stationary distribution. Moreover, all moments of $X^n(t)$ and $Q^n(t)$ are finite under this distribution, i.e., $\mathbb{E}_s^{\mathrm{pr},n}[\|X^n(t)\|^k] < \infty$, $k \in \mathbb{N}$. These facts can be verified by standard techniques (such as [3, Theorem 8.6]) using the abandonment ingredient of the model that makes the process stable. Recalling that by Assumption 2.1(i), $\mathbb{E}[\|X^n(0)\|^2] < \infty$, similar arguments show that, for every n, $\sup_t \mathbb{E}[\|X^n(t)\|^2] < \infty$ under $\pi^{\mathrm{pr},n}$.

We note that the estimates in the last result hold also under the stationary distribution, namely $\lim_{n\to\infty} \mathbb{E}_s^{\mathrm{pr},n}[\|n^{-1}X^n(t) - x^*\|^2] = 0$, and similarly for $Q^n(t)$. Indeed, since (19) and (20) hold for any initial conditions satisfying Assumption 2.1(ii), they also hold for the invariant distribution. However, under this initial condition, the laws of $X^n(t)$ and $Q^n(t)$ do not depend on t, hence the statements in (19) and (20) imply their stationary counterparts.

For the nth system, the costs defined in (7) are denoted by

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

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

where $c \geq 0$ does not depend on n. As a consequence of Remark 2.2, we can associate with $\pi^{\mathrm{pr},n}$ the stationary cost $C_s^n(\pi^{\mathrm{pr},n}) = \mathbb{E}_s^{\mathrm{pr},n}[c\cdot Q^n(t)]$, and obtain

$$\underline{C}^n\big(\pi^{\mathrm{pr},n}\big) = C_s^n\big(\pi^{\mathrm{pr},n}\big) = \overline{C}^n\big(\pi^{\mathrm{pr},n}\big) \triangleq C^{\mathrm{pr},n}. \tag{21}$$

Thus, all three costs coincide, and are commonly denoted by $C^{\mathrm{pr},n}$ in the following.

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

$$\frac{c_1\mu_1}{\theta_1} \geq \frac{c_2\mu_2}{\theta_2} \geq \cdots \geq \frac{c_I\mu_I}{\theta_I}. \tag{22}$$

We write $C^{*,n}$ for the average cost corresponding to $\pi^{*,n}$.

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*

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

which implies, together with (13), *that*

$$\lim_{n\to\infty} n^{-1}C^{*,n} = V_1.$$

In view of (13), this result expresses the asymptotic optimality of the proposed policy.

We can give a bound on the rate of convergence of the cost 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,

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

where c_0 *is a constant not depending on* n.

Remark 2.3 When the nth 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,

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

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

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

and consequently,

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

Remark 2.4 While the results above were derived for the optimal priority policy, similar convergence properties may be seen to hold for *any* priority policy $\pi^{\mathrm{pr},n}$, after replacing the constant V_1 with a (larger) value $V(\pi^{\mathrm{pr}})$. The latter is defined by reordering the class indices according to the specified priority, computing the corresponding fluid solution (18), and letting $V(\pi^{\mathrm{pr}}) = c\cdot q$. In particular, the estimates in Theorem 2.1 may be established as before with these modified definitions, and similarly for the upper bound in Proposition 2.2. However, the lower bound of that proposition would have to be weakened, as the proof of Proposition 2.1 is not valid here. Rather, one can obtain the bound $n^{-1}C^{\mathrm{pr},n} \ge V(\pi^{\mathrm{pr}}) - c_0(n^{-1} + r_n)^{1/2}$, analogous to the upper bound in Proposition 2.2. The easiest way to see this is by replacing c with $-c$ in this result, and noting that the proof is indifferent to the sign on c.

3 Proofs

Throughout this section, the ordering (22) 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 (14) can easily be seen to be solved by (17) and (18) (see [1] for more details). Moreover, since q^* and z^* attain the infimum (14), we have

$$V_1 = c \cdot q^*, \tag{23}$$

and

$$\theta_i q_i^* + \mu_i z_i^* = \lambda_i, \quad i = 1, 2, \ldots, I. \tag{24}$$

The analysis of the policy $\pi^{*,n}$ will be based on the fact that the process X^n, under $\pi^{*,n}$, is Markovian. We let $f^n(x) = \sum_{i=1}^I \beta_i (x_i - x_i^* n)^2$, where $\beta_i > 0$ 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 $\beta_i > 0$ can be chosen so that*

$$\mathcal{L}^n f^n(x) \le -a f^n(x) + a_1 \|x\| + \delta_n n^2, \quad x \in \mathbb{Z}_+^I,\ n \ge n_0, \tag{25}$$

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

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

Using this in (15), along with the identities $(a \pm 1)^2 - a^2 = \pm 2a + 1$ yields

$$\begin{aligned}\mathcal{L}^n f^n(x) = {} & \sum_{i=1}^I 2\beta_i \left(x_i - x_i^* n + \frac{1}{2}\right)\lambda_i^n \\ & - \sum_{i=1}^I 2\beta_i \left(x_i - x_i^* n - \frac{1}{2}\right)\big[\mu_i^n x_i + \big(\theta_i^n - \mu_i^n\big)\mathbf{Q}_i^n(x)\big].\end{aligned} \tag{26}$$

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

$$\begin{aligned}\mathcal{L}^n f^n(x) = {} & \sum_{i=1}^I \beta_i \lambda_i^n + \sum_{i=1}^I \beta_i \big[\mu_i^n x_i + \big(\theta_i^n - \mu_i^n\big)\mathbf{Q}_i^n(x)\big] \\ & + 2\sum_{i=1}^I \beta_i Y_i \big[\lambda_i^n - \mu_i^n x_i - \big(\theta_i^n - \mu_i^n\big)\mathbf{Q}_i^n(x)\big]\end{aligned}$$

$$\leq C_1\big(n+\|x\|\big)+2\sum_{i=1}^{I}\beta_i Y_i\big[\lambda_i^n-\mu_i^n x_i-\big(\theta_i^n-\mu_i^n\big)\mathbf{Q}_i^n(x)\big]$$

$$= C_1\big(n+\|x\|\big)-2\sum_{i=1}^{I}\beta_i\mu_i^n Y_i^2$$

$$+2\sum_{i=1}^{I}\beta_i Y_i\big[\lambda_i^n-\mu_i^n x_i^* n-\big(\theta_i^n-\mu_i^n\big)\mathbf{Q}_i^n(x)\big]. \tag{27}$$

Recall that $x^*=z^*+q^*$ and that, by (24), $\theta_i q_i^*+\mu_i z_i^*=\lambda_i$. Therefore

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

Here, $\varepsilon_n\to 0$, by Assumption 2.1. Thus, the last term on the r.h.s. of (27) is given by

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

where we used the equality $nq_i^*=\mathbf{Q}_i^n(nx^*)$, which 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

$$\mathbf{Q}_i^n(x)-\mathbf{Q}_i^n(\tilde{x})=\rho(x_i-\tilde{x}_i)+\eta\sum_{j=1}^{i-1}(\tilde{x}_j-x_j), \tag{28}$$

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 $\delta_n=|\varepsilon_n|$,

$$\mathcal{L}^n f^n(x)\leq C_1\big(n+\|x\|\big)+C_2\|Y\|n\delta_n$$
$$-2\sum_{i=1}^{I}\beta_i\big[\big(1-\rho_i(x)\big)\mu_i^n+\rho_i(x)\theta_i^n\big]Y_i^2$$
$$+2\sum_{i=1}^{I}\beta_i\big(\theta_i^n-\mu_i^n\big)\eta_i(x)Y_i\sum_{j=1}^{i-1}Y_j.$$

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

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

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

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

where C_3 depends only on I. Now choose b_i so that $2Ab_i = m_i/2$, $i = 1, 2, \ldots, I$. Next, determine β_i inductively, as follows. Let $\beta_1 = 1$. For $i = 2, 3, \ldots, I$, let β_i (depending on $\beta_1, \ldots, \beta_{i-1}$) be determined by

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

Then

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

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

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

for all sufficiently large n. Thus

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

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

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

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

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

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

$$\frac{d\xi(t)}{dt} \le Ln - M\xi(t), \qquad \xi(0) = \mathbb{E}\big[\big\|X(0)\big\|\big], \tag{29}$$

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

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

that is,

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

(Note that this bound can alternatively be obtained by applying Gronwall's inequality, in differential form, to (29).) This completes the proof. □

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

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

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

$$\mathbb{E} f^n\big(X^n(t)\big) = \mathbb{E} f\big(X^n(0)\big) + \mathbb{E}\int_0^t \mathcal{L}^n f^n\big(X^n(s)\big)\,ds. \tag{30}$$

Let us prove that

$$\limsup_{t\to\infty} \mathbb{E}\big[\big\|n^{-1}X^n(t) - x^*\big\|^2\big] \le \bar{\delta}_n, \tag{31}$$

where $\bar{\delta}_n$ is a sequence converging to zero. To this end, note by (30) 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 (30) and using Lemma 3.1, we have

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

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

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

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 $t \ge T_n$,

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

This proves (31), and hence (19) follows.

To establish (20), note that we have for every n,

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

With the notation (16), the map $\mathbf{Q}^1 : \mathbb{R}_+^I \to \mathbb{R}_+^I$, is given by

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

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 (31),

$$\limsup_{t\to\infty} \mathbb{E}\big[\big\|n^{-1}Q^n(t) - q^*\big\|^2\big] \le C_4\bar{\delta}_n, \tag{32}$$

for some C_4 depending only on I. This shows (20). □

Proof of Theorem 2.2 The asserted equality follows from (21). Next, keeping the notation of the proof of Theorem 2.1, the inequality (32) implies

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

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 (23). □

Proof of Proposition 2.2 Recall the definition of V_1 in (14), 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 the lower bound stated in the proposition follows.

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. Using (21), 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. The upper bound follows from Theorem 2.2 and the fact that under $\pi^{*,n}$, for every n,

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

which in turn follows from the ergodicity of the Markov chain X^n under this policy (Remark 2.2). □

Acknowledgements The authors wish to thank Gennady Shaikhet for pointing out an error in an earlier version of this manuscript, and two referees for careful reading, comments and suggestions which helped to improve this paper.

Appendix

A.1 Lower bound for general service time distribution

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 $\tilde{D}_{i,k}$,

$i = 1, 2, \ldots, I$, $k = 1, 2, \ldots, n$, 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

$$D_{i,k}(t) = \tilde{D}_{i,k}\left(\mu_i \int_0^t Z_{i,k}(s)\,ds\right), \tag{33}$$

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

$$\sum_{k=1}^{N} Z_{i,k} = Z_i, \qquad \sum_{k=1}^{N} D_{i,k} = D_i. \tag{34}$$

Unlike Sect. 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 (33) 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 non-interruptible policy* π,

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

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, one can find a subsequence of $\{t_k\}$ along which the random variables $t^{-1}\int_0^t Q_i(s)\,ds$, $t^{-1}\int_0^t Z_{i,k}(s)\,ds$, $k = 1, 2, \ldots, N$, and $t^{-1}\int_0^t Z_i(s)\,ds$ converge to limits $\hat{Q}_i$, $\hat{Z}_{i,k}$ and $\hat{Z}_i$, taking values in $[0, \infty]$, $[0, \infty)$ and $[0, \infty)$, respectively. 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 (34),

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

a.s., where the limit in the first line of the above display is along the subsequence. The proof is completed as that of Proposition 2.1. □

A.2 Addendum to the proof of Proposition 2.1

The claim $t^{-1}X_i(t)$ a.s. will be proved by comparing the process $X_i(t)$ to the number-in-system process, $\tilde{Q}(t)$, for the $G/M/\infty$ model (in step 1 below), and then arguing that an analogous property holds for the latter, namely $t^{-1}\tilde{Q}(t) \to 0$ a.s. (in steps 2–4).

1. Fix i. Recall that $X_i(t) = Q_i(t) + Z_i(t)$, where $Z_i(t) \le n$. It is easy to see that one can couple (X_i, Q_i, Z_i) with a triplet $(\hat{X}, \hat{Q}, \hat{Z})$ representing number-in-system, queue-length and number-in-service for a system in which all n servers have service time $= +\infty$, while keeping $\hat{Z}(t) = Z_i(t)$ and $\hat{Q}(t) \ge Q_i(t)$ for all times.

   Next, a further coupling argument can be used to show that $\hat{X}(t)$ is stochastically dominated by the number-in-system in the infinite service time model alluded to above, operating under the policy that keeps at all times $Z = \min(n, X)$ (i.e., as many customers as possible in the 'limbo' of the n idle servers). It is not hard to see that for such a system all servers will forever be busy working on the first n customers. As a result, $\hat{X}(t) \le n + \tilde{Q}(t)$, where $\tilde{Q}$ is the number-in-system for a $G/M/\infty$ system with service rate θ (with a suitable initial condition).
2. Consider then $\tilde{Q}(t)$, the $G/M/\infty$ process with service rate θ per customer, i.i.d. inter-arrival times with finite mean λ^{-1}, and finite initial conditions $\tilde{Q}(0)$.

   It suffices to show $\lim_{t\to\infty} t^{-1}\tilde{Q}(t) = 0$ (a.s.). We were unable to deduce this from known results regarding the stability of the $G/M/\infty$ model in a direct way. The argument we provide below is based on consideration of the embedded Markov chain and a standard second moment estimate.

   Let (t_n) denote the arrival time sequence, and consider the embedded Markov chain $Q_n = \tilde{Q}(t_n)$ at these arrival times. Noting that $\tilde{Q}(t)$ is non-increasing between arrival times, and that $n^{-1}t_n \to \lambda^{-1}$ (a.s.), it suffices to show that $\lim_{n\to\infty} n^{-1}Q_n = 0$ (a.s.).
3. The second moment of (Q_n) is uniformly bounded, namely $M_2 \stackrel{\triangle}{=} \sup_n E(Q_n^2) < \infty$. To see this, note that $Q_{n+1} = 1 + Q_n - D_n$, where D_n is the number of served customers on the nth interval $[t_n, t_{n+1})$. Note that each of the Q_n customers present at t_n is served by an exponential server of rate μ. Let p be the probability that such a customer does *not* complete its service during that interval (given by $p = \int_{t=0}^{\infty} \exp(-\theta t)\, dF_T(t)$, where F_T is the inter-arrival distribution function). Note that $0 < p < 1$. Then $D_n = \sum_{i=1}^{Q_n}(1 - Z_i)$, where (Z_i) are independent Bernoulli random variables with $P(Z_i = 1) = 1 - P(Z_i = 0) = p$, and $Q_{n+1} = 1 + \sum_{i=1}^{Q_n} Z_i$. Therefore

$$\begin{aligned} E\big(Q_{n+1}^2 | Q_n\big) &= \operatorname{var}\big(Q_{n+1}^2 | Q_n\big) + E(Q_{n+1}|Q_n)^2 \\ &= Q_n \operatorname{var}(Z_i) + (1 + Q_n p)^2 \\ &= p(1-p)Q_n + (1 + Q_n p)^2. \end{aligned}$$

Since $p < 1$, the last equation implies existence of some $p < \gamma < 1$ and a finite constant M_1 so that

$$E\big(Q_{n+1}^2 | Q_n\big) \le \gamma\, Q_n^2 + M_1 \quad \text{(a.s.)}.$$

By iteration, it follows that $E(Q_n^2) \le \max\{Q_0^2, \frac{M_1}{1-\gamma}\} \stackrel{\triangle}{=} M_2 < \infty$ for all n.

4. Therefore, by Chebychef's inequality, for any $\varepsilon > 0$,

$$\sum_{n=0}^{\infty} P\left(\frac{Q_n}{n} > \varepsilon\right) \le \sum_{n=0}^{\infty} \frac{M_2}{n^2 \varepsilon^2} < \infty.$$

It now follows by the Borel–Cantelli Lemma that $n^{-1} Q_n \to 0$ (a.s.).

This completes the proof of the claim $t^{-1} X_i(t) \to 0$ a.s.

References

1. Atar, R., Giat, C., Shimkin, N.: The $c\mu/\theta$ rule for many-server queues with abandonment. Oper. Res. (2010, to appear)
2. Meyn, S.: Control Techniques for Complex Networks. Cambridge University Press, Cambridge (2008)
3. Robert, P.: Stochastic Networks and Queues. Springer, Berlin (2003)
4. Whitt, W.: Stochastic-Process Limits. Springer Series in Operations Research. Springer, New York (2002)