On the $M_t/M_t/K_t + M_t$ queue in heavy traffic

Anatolii A. Puhalskii
Institute for Problems in Information Transmission
and University of Colorado Denver

June 27, 2011

Abstract

The focus of this talk is on the asymptotics of large-time numbers of customers in time-periodic Markovian many-server queues with customer abandonment in heavy traffic. Limit theorems are obtained for the periodic number-of-customers processes under the fluid and diffusion scalings. Other results concern limits for general time-dependent queues and for time-homogeneous queues in steady state.

1 Introduction

Many-server queues with customer abandonment have been the subject of extensive research, the primary motivation coming from modelling call centres, see, e.g., Garnett, Mandelbaum, and Reiman [2], Whitt [6, 7], Zeltyn and Mandelbaum [8], and references therein. Those papers testify to the importance of the asymptotics where both the arrival rate and the number of servers tend to infinity, their ratio being maintained, whereas the service and abandonment rates are kept fixed. Most studied is the case of Poisson arrival processes and exponential service and abandonment times where the arrival, service, and abandonment rates, and the number of servers do not vary with time. Fleming, Simon, and Stolyar [1], assuming the critical load condition, obtain diffusion-scale limit theorems for the stationary number of customers. Garnett, Mandelbaum, and Reiman [2], also for the critical load, derive fluid- and diffusion-scale limits for the number-of-customers and virtual-waiting-time processes, and for the stationary distributions of these processes. Their other results are concerned with limits for the stationary fractions of abandoning customers and of customers who have to wait in the queue, as well as with computing expectations of functions of the waiting time. Similar asymptotics for the overloaded case are obtained in Whitt [6], who assumes a finite waiting room, and Talreja and Whitt [5]. In addition, Whitt [6] gains insight into the case where the number of servers is much larger than the abandonment rate. Talreja and Whitt [5] also give a proof of the virtual-waiting-time-process limit for the critically loaded queue. The Markovian assumptions are relaxed in Zeltyn and Mandelbaum [8] who study steady-state waiting times. A general framework of Markovian stochastic processing systems with time-varying rates is studied by Mandelbaum, Massey, and Reiman [4] who obtain fluid- and diffusion-scale limits for the number-of-customers processes. They do not require certain load conditions to hold. The application to many-server queues with abandonment is explored in a series of papers by Mandelbaum, Massey, Reiman, and Stolyar who consider time-varying rates, allow the possibility of retrials, and incorporate virtual-waiting-time processes, see, e.g., Mandelbaum, Massey, Reiman, Rider, and Stolyar [3] and references therein.

The purpose of this research is a study of Markovian many-server queues with customer abandonment in heavy traffic for a time-periodic case where the arrival, abandonment, and service rates, and the number of servers can be modelled as jointly periodic functions of time. Under those hypotheses, the large-time distributions of the numbers of customers are periodic. It is proved that they converge to the periodic distribution of a limiting diffusion process which arises as a particular case of the results of Mandelbaum, Massey, and Reiman [4]. The convergence of the periodic one-dimensional distributions is further extended to provide convergence of the periodic processes. The method of proof consists in establishing convergence of the number-of-customers processes and in checking the tightness of the stationary distributions of embedded discrete-time Markov chains. That makes the results of Mandelbaum, Massey, and Reiman [4] essential. Unfortunately, the proofs there contain flaws. Therefore, I first provide a separate proof of the heavy traffic convergence in distribution of the number-of-customers processes in many-server queues with time-varying rates and abandonment. The part dealing with tightness relies on bounds on the first and second moments of the numbers of customers which are uniform in time and may be of interest in their own right. Along with the application to the periodic case, I use the convergence of the processes and the moment bounds in order to establish convergence of the stationary number of customers in the time-homogeneous case for all three possible loads: supercritical, critical, and subcritical. On the one hand, this provides a unified treatment of and a different perspective on the results of Fleming, Simon, and Stolyar [1], Garnett, Mandelbaum, and Reiman [2], and Whitt [6] on the limits of the stationary number of customers. On the other hand, not only are the limits for the one-dimensional stationary distributions obtained, but also limits for the stationary versions of the corresponding processes. In addition, it is shown that allowing the abandonment and service rates to depend on the scale gives rise to extra terms in the limit distributions.

2 Convergence of the number-of-customers processes

I will consider a sequence of $M_t/M_t/K_t + M_t$ queues indexed by $n \in \mathbb{N}$. The nth queue is fed by a Poisson process of customers of rate λ_t^n at time t. The arriving customers are served by one of the K_t^n servers on a FCFS basis. If at a certain epoch, a server serving a customer becomes unavailable because K_t^n decreases, then the customer being served is relegated to the head of the queue and resumes its service from scratch the next time it enters service. The service times are i.i.d. exponential of mean 1. The service rate at time t is μ_t^n. The customers may abandon the queue after an exponentially distributed time with parameter θ_t^n. The functions λ_t^n, μ_t^n, and θ_t^n are assumed to be $\mathbb{R}_+$-valued locally integrable functions, i.e., $\int_0^t \lambda_s^n \, ds < \infty$, $\int_0^t \mu_s^n \, ds < \infty$, and $\int_0^t \theta_s^n \, ds < \infty$ for all $t \in \mathbb{R}_+$. The functions K_t^n are $\mathbb{R}_+$-valued and Lebesgue measurable. The number of customers present at time 0, the arrival process, the service times, and the abandonment times are mutually independent.

Let A_t^n denote the number of customer arrivals by time t. As mentioned, the process $A^n = (A_t^n, \, t \in \mathbb{R}_+)$ is a Poisson process with time-varying rate λ_t^n. Customer abandonment will be modelled via independent Poisson processes $R^{n,i} = (R_t^{n,i}, \, t \in \mathbb{R}_+)$, $i \in \mathbb{N}$, of rate θ_t^n and customer service will be modelled via independent Poisson processes $B^{n,i} = (B_t^{n,i}, \, t \in \mathbb{R}_+)$, $i \in \mathbb{N}$, of rate μ_t^n. Let Q_t^n represent the number of customers present at time t. The evolution of the customer population is modelled by the following equation

$$Q_t^n = Q_0^n + A_t^n - \sum_{i=1}^{\infty} \int_0^t \mathbf{1}_{\{Q_{s-}^n \geq K_{s-}^n + i\}} \, dR_s^{n,i} - \sum_{i=1}^{\infty} \int_0^t \mathbf{1}_{\{Q_{s-}^n \wedge K_{s-}^n \geq i\}} \, dB_s^{n,i} \, .$$

This equation has a unique strong solution whose trajectories belong to $\mathbb{D}(\mathbb{R}_+, \mathbb{R})$.

The next theorem establishes a fluid-scale limit. I assume as fixed $\mathbb{R}_+$-valued locally integrable functions $(\lambda_t,\, t \in \mathbb{R}_+)$, $(\mu_t,\, t \in \mathbb{R}_+)$, and $(\theta_t,\, t \in \mathbb{R}_+)$, and an $\mathbb{R}_+$-valued Lebesgue measurable function $(\kappa_t,\, t \in \mathbb{R}_+)$. Given an $\mathbb{R}_+$-valued random variable q_0, let q_t be defined by the equation

$$q_t = q_0 + \int_0^t \lambda_s \, ds - \int_0^t \theta_s (q_s - \kappa_s)^+ \, ds - \int_0^t \mu_s \, (q_s \wedge \kappa_s) \, ds.$$

Theorem 2.1. *Suppose that, as $n \to \infty$, $\int_0^t \lambda_s^n/n \, ds \to \int_0^t \lambda_s \, ds$ for all t, that $\mu_t^n \to \mu_t$ uniformly on bounded intervals, that $\theta_t^n \to \theta_t$ uniformly on bounded intervals, and that $K_t^n/n \to \kappa_t$ for all t. If the random variables Q_0^n/n converge in distribution to a random variable q_0 as $n \to \infty$, then the processes $(Q_t^n/n,\, t \in \mathbb{R}_+)$ converge in distribution in $\mathbb{D}(\mathbb{R}_+, \mathbb{R})$ to the process $(q_t,\, t \in \mathbb{R}_+)$. In particular, if q_0 is deterministic, then for all $L > 0$ and $\epsilon > 0$,*

$$\lim_{n \to \infty} \mathbf{P}\big(\sup_{t \in [0,L]} |\frac{Q_t^n}{n} - q_t| > \epsilon \big) = 0 \, .$$

Let me introduce

$$\begin{aligned}
\alpha_t^n &= \sqrt{n}\big(\frac{\lambda_t^n}{n} - \lambda_t\big), \\
\beta_t^n &= \sqrt{n}\big(\frac{\mu_t^n}{n} - \mu_t\big), \\
\gamma_t^n &= \sqrt{n}\big(\frac{\theta_t^n}{n} - \theta_t\big), \\
\delta_t^n &= \sqrt{n}\big(\frac{K_t^n}{n} - \kappa_t\big) \, .
\end{aligned}$$

The following theorem yields a diffusion-scale limit. Let processes $X^n = (X_t^n,\, t \in \mathbb{R}_+)$ be defined by

$$X_t^n = \sqrt{n}\big(\frac{Q_t^n}{n} - q_t\big) \, .$$

In the rest of the paper, $(\alpha_t,\, t \in \mathbb{R}_+)$, $(\beta_t,\, t \in \mathbb{R}_+)$, and $(\gamma_t,\, t \in \mathbb{R}_+)$ represent locally integrable functions and $(\delta_t,\, t \in \mathbb{R}_+)$ represents a locally bounded Lebesgue measurable function.

Theorem 2.2. *Let the hypotheses of Theorem 2.1 hold where $q_0 \in \mathbb{R}_+$ is deterministic. Suppose that $\int_0^t \alpha_s^n \, ds \to \int_0^t \alpha_s \, ds$, $\beta_t^n \to \beta_t$, $\gamma_t^n \to \gamma_t$, and $\delta_t^n \to \delta_t$ uniformly on bounded intervals as $n \to \infty$, and that the random variables X_0^n tend in distribution to a random variable X_0 as $n \to \infty$. Then the processes X^n converge in distribution in $\mathbb{D}(\mathbb{R}_+, \mathbb{R})$ to the process $X = (X_t \, , t \in \mathbb{R}_+)$ that is the solution of the equation*

$$X_t = X_0 + \int_0^t \big(\alpha_s - \gamma_s (q_s - \kappa_s)^+ - \beta_s (q_s \wedge \kappa_s)\big) \, ds - \int_0^t \theta_s \big(\mathbf{1}_{\{q_s > \kappa_s\}} \, (X_s - \delta_s) + \mathbf{1}_{\{q_s = \kappa_s\}} \, (X_s - \delta_s)^+ \big) \, ds$$

$$- \int_0^t \mu_s \big(\mathbf{1}_{\{q_s < \kappa_s\}} \, X_s + \mathbf{1}_{\{q_s = \kappa_s\}} \, (X_s \wedge \delta_s) + \mathbf{1}_{\{q_s > \kappa_s\}} \, \delta_s \big) \, ds + \int_0^t \sqrt{\lambda_s + \theta_s (q_s - \kappa_s)^+ + \mu_s \, (q_s \wedge \kappa_s)} \, dW_s,$$

where $W = (W_t,\, t \in \mathbb{R}_+)$ is a standard Wiener process and W and X_0 are independent.

3 Convergence of the periodic queue lengths

In this section, I will assume that the functions λ_t^n, μ_t^n, θ_t^n, K_t^n, λ_t, μ_t, θ_t, and κ_t are T-periodic, where $T > 0$. I will also assume that

$$\int_0^T (\mu_s \wedge \theta_s)\, ds > 0\,.$$

In the long term, one expects a periodic pattern to be established for the number of customers. The next lemma confirms this to be the case. Let $Q^{n,\ell} = (Q^n_{\ell T+t},\, t \in \mathbb{R}_+)$, where $\ell \in \mathbb{Z}_+$.

Lemma 3.1. *Suppose that $\int_0^T (\mu_s^n \wedge \theta_s^n)\, ds > 0$. As $\ell \to \infty$, given an arbitrarily distributed Q_0^n, the sequence of the distributions of the processes $Q^{n,\ell}$ converges in the metric of total variation in $\mathbb{D}(\mathbb{R}_+, \mathbb{R})$ to the distribution of a process $\breve{Q}^n = (\breve{Q}^n_t,\, t \in \mathbb{R}_+)$, which is a T-periodic Markov process with the same transition probability function as Q^n. The distribution of $(\breve{Q}^n_t,\, t \in \mathbb{R}_+)$ is specified uniquely and is a stationary initial distribution for the $\mathbb{D}(\mathbb{R}_+, \mathbb{R})$-valued discrete-time homogeneous Markov process $\{(Q^n_{\ell T+t},\, t \in \mathbb{R}_+),\, \ell \in \mathbb{Z}_+\}$.*

My next step is to consider periodic regimes for the deterministic approximation.

Lemma 3.2. *1. There exists a unique $q_0 \in \mathbb{R}_+$ such that the function $(q_t,\, t \in \mathbb{R}_+)$ is T-periodic. An arbitrary solution converges to this periodic solution as $t \to \infty$.*

2. If q_0 is a random variable such that the process $(q_t,\, t \in \mathbb{R}_+)$ is T-periodic, then q_0 is deterministic and has the value specified in part 1.

In what follows, $(\breve{q}_t,\, t \in \mathbb{R}_+)$ represents the T-periodic solution of Lemma 3.2.

Theorem 3.1. *Suppose that, as $n \to \infty$, $\int_0^t \lambda_s^n/n\, ds \to \int_0^t \lambda_s\, ds$ for all t, that $\mu_t^n \to \mu_t$ uniformly on bounded intervals, that $\theta_t^n \to \theta_t$ uniformly on bounded intervals, and that $K_t^n/n \to \kappa_t$ for all t. Then, for all $\epsilon > 0$ and $L > 0$,*

$$\lim_{n\to\infty} \mathbf{P}\Big(\sup_{t\in[0,L]} \Big|\frac{\breve{Q}^n_t}{n} - \breve{q}_t\Big| > \epsilon\Big) = 0\,.$$

Let

$$\breve{X}^n_t = \sqrt{n}\Big(\frac{\breve{Q}^n_t}{n} - \breve{q}_t\Big)\,.$$

The process $(\breve{X}^n_t,\, t \in \mathbb{R}_+)$ is a T-periodic Markov process.

Theorem 3.2. *Suppose that $\int_0^t \alpha_s^n\, ds \to \int_0^t \alpha_s\, ds$, $\beta_t^n \to \beta_t$, $\gamma_t^n \to \gamma_t$, and $\delta_t^n \to \delta_t$ uniformly on bounded intervals as $n \to \infty$. Then the processes $(\breve{X}^n_t,\, t \in \mathbb{R}_+)$ converge in distribution as $n \to \infty$ to process $(\breve{X}_t,\, t \in \mathbb{R}_+)$, which is a T-periodic Markov process. It is uniquely specified by the equation*

$$\breve{X}_t = \breve{X}_0 + \int_0^t \big(\alpha_s - \gamma_s(q_s - \kappa_s)^+ - \beta_s(q_s \wedge \kappa_s)\big)\, ds - \int_0^t \theta_s\big(\mathbf{1}_{\{\breve{q}_s > \kappa_s\}}\,(\breve{X}_s - \delta_s) + \mathbf{1}_{\{\breve{q}_s = \kappa_s\}}\,(\breve{X}_s - \delta_s)^+\big)\, ds$$

$$- \int_0^t \mu_s\big(\mathbf{1}_{\{\breve{q}_s < \kappa_s\}}\,\breve{X}_s + \mathbf{1}_{\{\breve{q}_s = \kappa_s\}}\,(\breve{X}_s \wedge \delta_s) + \mathbf{1}_{\{\breve{q}_s > \kappa_s\}}\,\delta_s\big)\, ds + \int_0^t \sqrt{\lambda_s + \theta_s(\breve{q}_s - \kappa_s)^+ + \mu_s\,(\breve{q}_s \wedge \kappa_s)}\, d\breve{W}_s,$$

where $(\breve{W}_t,\, t \in \mathbb{R}_+)$ is a standard Wiener process and $\breve{X}_0$ and $(\breve{W}_t,\, t \in \mathbb{R}_+)$ are independent.

4 Convergence of stationary distributions

In this section I will assume constant arrival, service, and abandonment rates, so $\lambda_t^n = \lambda^n \geq 0$, $\theta_t^n = \theta^n > 0$, $\mu_t^n = \mu^n > 0$, $\lambda_t = \lambda \geq 0$, $\theta_t = \theta > 0$, $\mu_t = \mu > 0$, $\alpha_t = \alpha$, $\beta_t = \beta$, and $\gamma_t = \gamma$. The number of servers K_t^n is also assumed to be constant which I will take as the scaling parameter n, so $\kappa_t = 1$. Accordingly, $\delta_t = 0$. The equations for the fluid- and diffusion-scale limits which appear in Theorem 2.1 and Theorem 2.2 assume the following form:

$$q_t = q_0 + \lambda t - \int_0^t \theta(q_s - 1)^+ \, ds - \int_0^t \mu \, (q_s \wedge 1) \, ds,$$

$$X_t = X_0 + \int_0^t \big(\alpha - \gamma(q_s - 1)^+ - \beta(q_s \wedge 1)\big) \, ds - \int_0^t \theta\big(\mathbf{1}_{\{q_s > 1\}} \, X_s + \, \mathbf{1}_{\{q_s = 1\}} \, X_s^+\big) \, ds$$

$$- \int_0^t \mu\big(\mathbf{1}_{\{q_s < 1\}} \, X_s - \, \mathbf{1}_{\{q_s = 1\}} \, (-X_s)^+\big) \, ds + \int_0^t \sqrt{\lambda + \theta(q_s - 1)^+ + \mu \, (q_s \wedge 1)} \, dW_s \, .$$

Lemma 4.1. *If $\lambda \geq \mu$, then $\lim_{t\to\infty} q_t = (\lambda - \mu)/\theta + 1$. If $\lambda \leq \mu$, then $\lim_{t\to\infty} q_t = \lambda/\mu$. For all t, $q_t \neq 1$ except when $q_0 = 1$ and $\lambda = \mu$ in which case $q_t = 1$ for all t .*

The Markov chain Q^n is a birth-and-death process on $\mathbb{Z}_+$ with birth rates λ^n and death rates $\mu^n(i \wedge n) + \theta^n(i - n)^+$. It admits a unique stationary distribution which is a limit in total variation of the transient distributions for any initial condition. Let $\hat{Q}^n = (\hat{Q}_t^n, \, t \in \mathbb{R}_+)$ represent the stationary version of Q^n and let $\hat{q}_0 = \lim_{t\to\infty} q_t$.

Theorem 4.1. *Suppose that $\lambda^n/n \to \lambda$, that $\mu^n \to \mu$, and that $\theta^n \to \theta$ as $n \to \infty$. Then, for all $\epsilon > 0$ and $L > 0$,*

$$\lim_{n\to\infty} \mathbf{P}\big(\sup_{t\in[0,L]} |\frac{\hat{Q}_t^n}{n} - \hat{q}_0| > \epsilon\big) = 0 \, .$$

Let process $\hat{X}^n = (\hat{X}^n(t), \, t \in \mathbb{R}_+)$ represent the stationary version of X^n ,i.e., $\hat{X}^n(t) = \sqrt{n}(\hat{Q}_t^n/n - \hat{q}_0)$.

Theorem 4.2. *Suppose that $\sqrt{n}(\lambda^n/n - \lambda) \to \alpha$, $\sqrt{n}(\mu^n/n - \mu) \to \beta$, and $\sqrt{n}(\theta^n/n - \theta) \to \gamma$ as $n \to \infty$. Then the processes $\hat{X}^n$ converge in distribution in $\mathbb{D}(\mathbb{R}_+, \mathbb{R})$ as $n \to \infty$ to a stationary continuous-path Markov process $\hat{X} = (\hat{X}_t, \, t \in \mathbb{R}_+)$.*

If $\lambda < \mu$, then the process $\hat{X}$ is Gaussian with $\mathbf{E}\hat{X}_t = \alpha/\mu - \beta \, \lambda/\mu^2$ and $\mathbf{Cov}\,(\hat{X}_u, \hat{X}_v) = (\lambda/\mu)e^{-\mu|u-v|}$. If $\lambda > \mu$, then the process $\hat{X}$ is Gaussian with $\mathbf{E}\hat{X}_t = \alpha/\theta - \gamma(\lambda - \mu)/\theta^2 - \beta/\theta$ and $\mathbf{Cov}\,(\hat{X}_u, \hat{X}_v) = (\lambda/\theta)e^{-\theta|u-v|}$. If $\lambda = \mu$, then

$$\hat{X}_t = \hat{X}_0 + (\alpha - \beta)t - \theta \int_0^t \hat{X}_s^+ \, ds + \mu \int_0^t (-\hat{X}_s)^+ \, ds + \sqrt{2\mu}\hat{W}_t \, ,$$

where the distribution of $\hat{X}_0$ has density $C \exp\big(\big((\alpha - \beta)x - (x^2/2)(\theta \, \mathbf{1}_{\{x \geq 0\}} + \mu \, \mathbf{1}_{\{x<0\}} \,)\big)/\mu\big)$, $(\hat{W}_t, \, t \in \mathbb{R}_+)$ is a standard Wiener process, and $\hat{X}_0$ and $(\hat{W}_t, \, t \in \mathbb{R}_+)$ are independent.

References

[1] P.J. Fleming, B. Simon, and A. Stolyar. Heavy traffic limit for a mobile phone system loss model. In *Proc. 2nd Internat. Conf. Telecommunication Systems, Modeling, and Anal.*, pages 158–176, Nashville, TN, 1994.

[2] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing Service Oper. Management*, 4(3):208–227, 2002.

[3] A. Mandelbaum, W. Massey, M. Reiman, B. Rider, and Stolyar A. Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems*, 21(2-4), 2002.

[4] A. Mandelbaum, W. A. Massey, and M. Reiman. Strong approximations for Markovian service networks. *Queueing Syst.*, 30:149–201, 1998.

[5] R. Talreja and W. Whitt. Heavy-traffic limits for waiting times in many-server queues with abandonment. *Ann. Appl. Probab.*, 19(6):2137–2175, 2009.

[6] W. Whitt. Efficiency-driven heavy-traffic approximations for many-server queues with abandonment. *Management Sci.*, 50(10):1449–1461, October 2004.

[7] W. Whitt. Engineering solution of a basic call center model. *Management Sci.*, 51(2):221–235, February 2005.

[8] S. Zeltyn and A. Mandelbaum. Call centers with impatient customers: many-server asymptotics of the $M/M/n + G$ queue. *Queueing Syst.*, 51(3-4):361–402, 2005.