Conjectures on symmetric queues in heavy traffic

Bert Zwart

Received: 25 January 2022 / Accepted: 28 February 2022 / Published online: 15 March 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

1 Introduction

We consider the queue-length process in the $M/\Gamma^\gamma/1$ queue with symmetric service discipline, which is defined as follows: With $n$ customers in the system, the server works on the customer in position $i$ at rate $\gamma(n, i) \geq 0$. We assume $\sum_i \gamma(n, i) = 1$ for all $n \geq 1$, to make the service discipline work conserving. The customer arrival process is Poisson with rate $\lambda$. A customer arriving to a queue of size $n$ chooses position $i$ with probability $\gamma(n + 1, i)$, moving each customer in position $k \geq i$ to position $k + 1$. Conversely, when a customer at position $i$ departs, each customer in position $k > i$ moves to position $k - 1$. We refer to this system as a symmetric queue governed by $\gamma$. Important special cases of symmetric service disciplines are last-come-first-served (LCFS), where $\gamma(n, 1) = 1$ for all $n \geq 1$, and processor sharing (PS), which is modeled by taking $\gamma(n, i) = 1/n$ for all $1 \leq i \leq n$ and all $n \geq 1$. We refer to [4] for more background.

Symmetric service disciplines have the appealing property that the stationary distribution of the queue-length process $(Q^\gamma(t), t \geq 0)$, if it exists, is insensitive to the service-time distribution, apart from its mean $m$. In particular, if the load $\rho = \lambda m < 1$, then, as $t \to \infty$, $Q^\gamma(t)$ converges weakly to a random variable $Q^\gamma(\infty)$ which satisfies

$$P(Q^\gamma(\infty) \geq k) = \rho^k, \ k \geq 0. \quad (1)$$

This note states two open problems related to the queue-length process in heavy traffic.

2 Problem statement

Assume first that the second moment of the service-time distribution is finite. Let $r$ be a scaling parameter, and let $\lambda_r$ be a sequence of arrival rates such that $\lambda_r \to 1/m$ and $r(1 - \rho_r) = r(1 - \lambda_r m)$ converges to a real-valued constant. Let $Q^\gamma_r(t), t \geq 0$ be the corresponding queue-length process in the symmetric queue governed by $\gamma$
with Poisson arrival rate $\lambda_r$. For each $r$, define the re-scaled queue-length process $\hat{Q}_r(t) = Q_r(r^2 t)/r$, $t \geq 0$.

**Problem 1** show that $\hat{Q}_r(\cdot)$ converges in distribution to a reflected Brownian motion (RBM) as $r \to \infty$.

Our second problem relates to the case where the variance of the service-time distribution is infinite. To this end, let $L$ be a slowly varying function (i.e. $L(ax)/L(x) \to 1$ as $x \to \infty$ for every $a > 0$) and suppose that the tail $\bar{F}$ of the service-time distribution satisfies $\bar{F}(x) = L(x)x^{-\alpha}$ for $\alpha \in (1, 2)$. Let $c_r$ be the solution of $c_r \bar{F}(c_r) = 1/r$. Now, for each $r$, define $\hat{Q}_r(t) = Q_r(c_r t)/r$, $t \geq 0$.

**Problem 2** identify a candidate limit process $Q^*_r(\cdot)$ and show that $\hat{Q}_r(\cdot) \to Q^*_r(\cdot)$ in distribution as $r \to \infty$.

## 3 Discussion

Problem 1 has been solved for PS in [3], [10] for general (with finite second moment) inter-arrival time distribution and a finite fourth moment assumption on the service-time distribution. The fourth moment assumption was dropped in [7] using a different method, which required Poisson arrivals. For LCFS, Problem 1 has been solved in [9] for Poisson arrivals and extended to renewal arrivals in [8].

The drift and the variance of the RBM for PS and LCFS are identical, which is no coincidence, because it is shown in [2] that the distribution of $Q^*_r(t)$ (for fixed $t$) is independent of $\gamma$, if $Q^*_r(0) = 0$. Given this, it is natural to conjecture that the scaling limit in Problem 1 should be independent of $\gamma$. It is remarkable though that this independence of $\gamma$ may break down if the Poisson arrival assumption is relaxed. For the LCFS case, this is illustrated in [8].

To solve Problem 1, one can think of several strategies. The approach in [3], [10] is to use a detailed measure-valued description of the queueing dynamics and hinges on the physical intuition that, in heavy traffic, the queue length becomes a deterministic function of the workload. This property, known as state-space collapse, should still hold for general $\gamma$. The approaches in [7] and [9] are based on connections with branching processes and excursion theory. To make such an approach work, one would first have to identify an appropriate relation between a Crump–Mode–Jagers (CMJ) branching process (see [7] and references therein for background on CMJ processes) and the service discipline $\gamma$. A more pragmatic way is to first prove Problem 1 for the special case of phase-type service-time distributions.

Problem 2 has been haunting me for the past 20 years. I view Problem 2 to be much harder than Problem 1. One reason is that the limiting queue-length process can no longer be a deterministic function of the heavy traffic limit of the workload process (for the latter, see [12]). Note that the distinction between finite and infinite variance is not needed for the invariant distribution in the case: When $\rho_r < 1$ for all $r$, $(1 - \rho_r)Q^*_r(\infty)$ converges in distribution to an exponential random variable with unit mean as $r \to \infty$.

It is possible to determine weak convergence of $\hat{Q}_r(t)$ for fixed $t$ as $r \to \infty$, assuming $Q^*_r(0) = 0$, using the result in [2]. This suggests that marginal distributions...
of a candidate limit process \( Q_{\gamma}^*(\cdot) \) should be independent of \( \gamma \). This will not be the case for the entire process, which has been determined in [1] for LCFS and in [7] for PS. In both cases, the limiting queue-length process is expressed in terms of a functional of a sequence of continuous-state branching processes, namely the height process in the LCFS case, and a Lamperti transform in the PS case; see [1] and [7] for detailed descriptions.

For PS, convergence is still an open problem, as tightness of \( \{ \hat{Q}_{PS}^r(\cdot) \}_{r} \) has not been established. In the case where \( \gamma \) is completely general, I have no conjecture on what the candidate limit process \( Q_{\gamma}^*(\cdot) \) might be.

An idea which may lead to partial results of independent interest is to try and establish a relationship between the symmetric queue governed by \( \gamma \), and some functional of a CMJ process, which should be dependent on \( \gamma \). As mentioned in the discussion of Problem 1, such a connection would be interesting in its own right. Once such a connection has been established, it may be possible to use the technique in [5], leveraging the fact that \( \hat{Q}_{r}^*(\cdot) \) is a regenerative process for each \( r \).

A sufficient condition for tightness of \( \{ \hat{Q}_{PS}^r(\cdot) \}_{r} \) has been described in [6] which is appealing, as it is stated in terms of a tail bound for the distribution of the maximum queue length during a busy cycle. Deriving such a tail bound for arbitrary symmetric service disciplines could be an interesting problem in its own right. To develop intuition, one may first approach this by assuming phase-type service times. A related tail bound for bandwidth sharing networks has been derived in [11].

References

1. Duquesne, T., Le Gall, J.-F.: Random trees, Lévy processes and spatial branching processes. *Astérisque*, (281):vi+147, (2002)
2. Fralix, B., Zwart, B.: Time-dependent properties of symmetric queues. Queueing Syst. 67(1), 33–45 (2011)
3. Gromoll, H.C.: Diffusion approximation for a processor sharing queue in heavy traffic. Ann. Appl. Probab. 14(2), 555–611 (2004)
4. Kelly, F.P.: Reversibility and Stochastic Networks. Wiley Series in Probability and Mathematical Statistics, Wiley, Chichester (1979)
5. Lambert, A., Simatos, F.: The weak convergence of regenerative processes using some excursion path decompositions. Ann. Inst. Henri Poincaré Probab. Stat. 50(2), 492–511 (2014)
6. Lambert, A., Simatos, F.: Asymptotic behavior of local times of compound Poisson processes with drift in the infinite variance case. J. Theoret. Probab. 28(1), 41–91 (2015)
7. Lambert, A., Simatos, F., Zwart, B.: Scaling limits via excursion theory: interplay between Crump-Mode-Jagers branching processes and processor-sharing queues. Ann. Appl. Probab. 23(6), 2357–2381 (2013)
8. Limic, V.: On the behavior of LIFO preemptive resume queues in heavy traffic. Electron. Commun. Probab. 5, 13–27 (2000)
9. Limic, V.: A LIFO queue in heavy traffic. Ann. Appl. Probab. 11(2), 301–331 (2001)
10. Puha, A.L., Williams, R.J.: Invariant states and rates of convergence for a critical fluid model of a processor sharing queue. Ann. Appl. Probab. 14(2), 517–554 (2004)
11. Shah, D., Tsitsiklis, J.N., Zhong, Y.: Qualitative properties of \( \alpha \)-fair policies in bandwidth-sharing networks. Ann. Appl. Probab. 24(1), 76–113 (2014)
12. Whitt, W.: Stochastic-Process Limits. Springer Series in Operations Research, Springer-Verlag, New York (2002)