Recent Papers on the Time-Varying Single-Server Queue

Ward Whitt

Department of Industrial Engineering and Operations Research, Columbia University,
New York, NY, 10027; ww2040@columbia.edu

January 20, 2020

Abstract

This is an overview of my papers in the last few years on the time-varying single-server queue.

1 Stabilizing Performance in Single-Server Queues

There is a long history of staffing (choosing a time-varying number of servers) to stabilize performance in many-server queues with time-varying arrival-rate functions, beginning with Jennings et al. [1996] and continuing with Feldman et al. [2008], Liu and Whitt [2012],Li et al. [2015] and He et al. [2016]. These draw on infinite-server approximations and the modified-offered load approximation, as reviewed in Whitt [2013].

Until recently, there were no corresponding results for single-server queues. An initial result for the single-server queue appears in Whitt [2015]. Since there is only one server, the method proposed for stabilization is a service-rate control, assuming that the service requirements come from an i.i.d. sequence, but the rate that the server works remains to be determined. Several methods are investigated. Strong positive results are obtained for the rate-matching control, which lets the service rate at time t be proportional to the arrival rate function, i.e.,

$$\mu(t) \equiv \frac{\lambda(t)}{\rho}, \tag{1.1}$$

where $\lambda(t)$ is the given arrival rate function, while ρ is the target traffic intensity; see (2.1) of Whitt [2015]. The rate-matching control is shown to stabilize the mean queue length, but not the mean waiting time. Confirming simulation experiments were conducted by Ni Ma, also reported in Ma and Whitt [2015a]. The simulation algorithm is discussed there and in Ma and Whitt [2015b].

Two other "square-root" controls are also considered. The first in (2.3) of Whitt [2015] is an analog

of the square-root-staffing formula used for many-server queues, i.e.,

$$\mu(t) \equiv \lambda(t) + \xi\sqrt{\lambda(t)}, \quad t \geq 0. \tag{1.2}$$

The second emerges from a fixed-point equation based on the pointwise-stationary approximation (PSA). Using the steady-state approximation for the $GI/GI/1$ queue, namely,

$$E[W] \approx \frac{\tau\rho V}{1-\rho}, \tag{1.3}$$

where $V \equiv (c_a^2 + c_s^2)/2$ is a variability parameter and τ is the mean service time, if we assume that it is valid locally at time t, then we obtain

$$w \equiv E[W(t)] \approx \frac{\rho(t)V}{\mu(t)(1-\rho(t))} = \frac{\lambda(t)V}{\mu(t)^2 - \mu(t)\lambda(t)} \tag{1.4}$$

where $\rho(t) \equiv \lambda(t)/\mu(t)$ is the instantaneous traffic intensity, which is a quadratic equation in $\mu(t)$. Figure 2 of Whitt [2015] shows that the first square-root rule is not too bad, but does not stabilize either the mean queue length or the mean waiting time. Figure 3 (right) shows that this second square-root formula does stabilize the mean waiting time for long cycles.

The paper Whitt [2015] has a strong result saying that it is not possible to stabilize both the mean queue length and the mean waiting time in heavy traffic; see Theorem 5.3 of Whitt [2015]. Indeed, when we stabilize the mean queue length, the mean waiting time tends to be inversely proportional to the arrival rate; see Theorem 5.2 of Whitt [2015]. It still remains to develop effective methods to stabilize the mean waiting time for short as well as long cycles.

A recent overview of time-varying queues is contained in Whitt [2018].

2 Approximating the Performance of Time-Varying $G_t/GI/1$ Queues

Just as for stationary models, diffusion approximations and heavy-traffic limits provide a basis for approximating the performance of single-server queues with time-varying arrival rates (and i.i.d. service times). Early work in this direction was done by Newell [1968a,b,c], with significant subsequent contributions by Massey [1985] and Mandelbaum and Massey [1995]. The uniform acceleration paper Massey and Whitt [1997] is also relevant. We have recently contributed in Whitt [2014, 2016]. These new papers express the heavy-traffic limits in the (now familiar) setting of Whitt [2002].

In Whitt [2014] a heavy-traffic limit is established for the periodic $G_t/GI/1$ model. As in previous work, it exploits the arrival process model in which the time-varying arrival process A is represented as the composition of a stochastic process N satisfying a FCLT and a deterministic cumulative arrival

rate function Λ, i.e., $A = N \circ \Lambda$, which embodies all the time-dependence. The limit process in the main theorem (Theorem 3.2 of Whitt [2014]) is reflected periodic Brownian motion (RPBM). Unfortunately, that RPBM limit process is not so useful directly, because not much is known about it.

Ni Ma and Wei You have done recent work to yield practical performance characterizations. An effective algorithm to simulate the general periodic $G_t/GI/1$ queue and RPBM is developed in Ma and Whitt [2016] exploiting a convenient reverse-time representation of the time-varying workload process and rare-event simulation. The algorithm draws on the rare-event simulation for the $GI/GI/1$ queue as in Asmussen [2003]. The rare-event approach is designed to efficiently calculate very small tail probabilities, but we show that it also applies to compute the time-varying mean and variance. We obtain a simulation algorithm for RPBM by considering a heavy-traffic approximation. Extensive simulation experiments show that it is effective. Additional results appear in Ma and Whitt [2019].

An effective approximation for the general periodic $G_t/GI/1$ queue and RPBM is developed in Whitt and You [2016b], exploiting a new periodic time-varying robust queueing (PRQ), which extends a first paper on robust queueing for stationary models in Whitt and You [2016a] The PRQ yields an easily computed approximation for the time-varying mean workload. It is remarkably effective in predicting the time of peak congestion within a periodic cycle. In Whitt and You [2016a,b], theoretical support is provided by new limit theorems. In important cases with appropriate parameters, the RQ bound is asymptotically correct, and so can serve as a useful approximation.

References

d

S Asmussen. *Applied Probability and Queues*. Springer, New York, second edition, 2003.

Z. Feldman, A. Mandelbaum, W. A. Massey, and W. Whitt. Staffing of time-varying queues to achieve time-stable performance. *Management Sci.*, 54(2):324–338, 2008.

B. He, Y. Liu, and W. Whitt. Stabilizing performance in nonstationary queues with non-Poisson arrivals. published online in 2016, 2016.

O. B. Jennings, A. Mandelbaum, W. A. Massey, and W. Whitt. Server staffing to meet time-varying demand. *Management Sci.*, 42:1383–1394, 1996.

A. Li, W. Whitt, and J. Zhao. Staffing to stabilize blocking in loss models with time-varying arrival rates. *Probability in the Engineering and Informational Sciences*, 30(2):185–211, 2015.

Y. Liu and W. Whitt. Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper. Res.*, 60(6):1551–1564, 2012.

N. Ma and W. Whitt. Efficient simulation of non-Poisson non-stationary point processes to study queueing approximations. *Statistics and Probability Letters*, 102:202–207, 2015a.

N. Ma and W. Whitt. Using simulation to study service-rate controls to stabilize performance in a single-server queue with time-varying arrival rate. In L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, editors, *Proceedings of the 2015 Winter Simulation Conference, Huntington Beach, CA, December 6-9, 2015*, pages 1–12. ACM, 2015b.

N. Ma and W. Whitt. A performance algorithm for periodic queues. Columbia University, Available at: http://www.columbia.edu/∼ww2040/allpapers.html, 2016.

N. Ma and W. Whitt. Minimizing the maximum expected waiting time in a periodic single-server queue with a service-rate control. *Stochastic Systems*, 9(3):261–290, 2019.

A. Mandelbaum and W. A. Massey. Strong approximations for time-dependent queues. *Mathematics of Operations Research*, 20(1):33–64, 1995.

W. A. Massey. Asymptotic analysis of the time-varying $M/M/1$ queue. *Mathematics of Operations Research*, 10 (2):305–327, 1985.

W. A. Massey and W. Whitt. Uniform acceleration expansions for Markov chains with time-varying rates. *Annals of Applied Probability*, 9(4):1130–1155, 1997.

G. F. Newell. Queues with time dependent arrival rates, I: the transition through saturation. *Journal of Applied Probability*, 5:436–451, 1968a.

G. F. Newell. Queues with time dependent arrival rates, II: the maximum queue and the return to equilibrium. *Journal of Applied Probability*, 5:579–590, 1968b.

G. F. Newell. Queues with time dependent arrival rates, III: a mild rush hour. *Journal of Applied Probability*, 5:591–606, 1968c.

W. Whitt. *Stochastic-Process Limits*. Springer, New York, 2002. Available at: http://www.columbia.edu/∼ww2040/jumps.html.

W. Whitt. Offered load analysis for staffing. *Manufacturing and Service Operations Management*, 15(2):166–169, 2013.

W. Whitt. Heavy-traffic limits for queues with periodic arrival processes. *Operations Research Letters*, 42: 458–461, 2014.

W. Whitt. Stabilizing performance in a single-server queue with time-varying arrival rate. *Queueing Systems*, 81:341–378, 2015.

W. Whitt. Heavy-traffic limits for a single-server queue leading up to a critical point. *Operations Research Letters*, 44:796–800, 2016.

W. Whitt. Time-varying queues. *Queueing Models and Service Management*, 1(2):79–164, 2018.

W. Whitt and W. You. Using robust queueing to expose the impact of dependence in single-server queues. Columbia University, Available at: http://www.columbia.edu/∼ww2040/allpapers.html, 2016a.

W. Whitt and W. You. Time-varying robust queueing. Columbia University, Available at: http://www.columbia.edu/∼ww2040/allpapers.html, 2016b.