Queueing Systems
https://doi.org/10.1007/s11134-022-09739-w

New decomposition approximations for queueing networks

Ward Whitt[1] · Wei You[2]

Received: 24 December 2021 / Accepted: 28 February 2022

1 Introduction

One of the great successes of queueing, for both theory and applications, was the development of the theory of product-form Markovian queueing networks stemming from Jackson [4]. The product-form theory motivated considering decomposition approximations for more general open queueing networks, e.g., with non-exponential service-time distributions and non-Poisson arrival processes, as developed by [5, 8]. In these early decomposition approximations, the arrival processes were partially characterized by their rate and a single variability parameter, corresponding to the variance of an interarrival time in a renewal-process approximation.

In [15] we developed a new decomposition algorithm to approximate the steady-state performance of a single-class open queueing network of single-server queues with unlimited waiting space, the first-come first-served discipline and Markovian routing. The algorithm allows non-renewal external arrival processes, general service-time distributions and customer feedback. Each flow is partially characterized by its rate and a scaled version of the variance-time curve, called the *Index of Dispersion for Counts* (IDC). Let A be an arrival counting process at a queue, i.e., $A(t)$ counts the total number of arrivals in the interval $[0, t]$. We assume that A is a stationary point process. We partially characterize A by its rate and its IDC, defined by $I_A(t) \equiv Var(A(t))/E[A(t)]$, $t \geq 0$. The required IDC functions for the external arrival processes can be calculated from the model primitives or estimated from data. Approximations for the IDC functions of the internal flows are calculated by solving a set of linear equations. The theoretical basis is provided by heavy-traffic limits for the flows established in [10, 11, 14].

✉ Ward Whitt
ww2040@columbia.edu

Wei You
weiyou@ust.hk

1 Department of Industrial Engineering and Operations Research, Columbia University, New York, USA

2 Department of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

Published online: 19 March 2022

Building on Bandi et al. [1], in [11] we developed a new *Robust Queueing* (RQ) technique to generate approximations of the mean steady-state performance at each queue from the IDC of the total arrival flow and the mean μ^{-1} and squared coefficient of variation (scv) c_s^2 of the service time at that queue. With RQ, we replace the stochastic net-input process by a deterministic instance drawn from a predetermined uncertainty set of input functions, while the RQ workload Z^* is regarded as the worst-case workload over the uncertainty set. The RQ approximation of the mean steady-state workload is

$$E[Z_\rho] \approx Z_\rho^* \equiv \sup_{x \geq 0} \left\{ -(1-\rho)x + b\sqrt{\rho x (I_A(x) + c_s^2)/\mu} \right\}, \tag{1}$$

where $b = \sqrt{2}$; see Theorem 2 of [11] and §EC.3 of its e-companion.

2. Problem statement

Improve the RQNA in [15] and extend it to more general models.

2.1 Improvements

2.1.1 Allowing multiple bottleneck queues The RQNA in [15] exploits the special case of the FCLT for the flows in Theorem 3.1 of [14] in which only a single queue in the network is a bottleneck. That leads to tractable approximations involving one-dimensional *reflected Brownian motion* (RBM) supporting the approximations in [15, 16]. With more bottleneck queues, we can exploit multidimensional RBM.

2.1.2. Statistical fitting with system data There is also great potential to exploit large system data sets together with advanced statistical techniques, e.g., machine learning, within this RQNA framework to fit the covariance functions of the internal flows, e.g., $Cov(A_{i,j}(t), A_{k,l}(t))$, that play a critical role in non-tree networks; e.g., see (28) of [15] and §§5.2–5.3 of [16].

2.2. Extensions

There are also many opportunities to extend the basic model. Such extensions are no doubt best motivated from the needs of concrete applications, as illustrated by the extensions of QNA in [8] discussed in [6]. Here are some: allow (i) multiple servers at each queue, (ii) multiple classes and/or more general routing, (iii) time-varying arrival processes.

3. Discussion

3.1 Performance comparisons with alternative algorithms

§6 of [15] compares RQNA predictions to simulation and other algorithms for difficult network examples with extensive near-immediate feedback from [2]. These examples are difficult for RQNA because the feedback induces strong dependence among the flows and the service times, as illustrated by the case of immediate feedback; see §III of [8] and §4 of [15]. Without our special techniques to eliminate near-immediate feedback, RQNA performs quite poorly, but when we incorporate these special techniques from §4 of [15], RQNA (elim) performs as well as the *sequential bottleneck decomposition* (SBD) from [2], which in turn outperforms QNA from [8] and QNET from [3].

As noted in Sects. 1.2 and 7 of [15], RQNA is effective for tree networks. Indeed, Theorem 5 of [11] shows that it is asymptotically exact in both light and heavy traffic for the $G/GI/1$. Second, Corollary 2 of [12] shows that a $GI/GI/1$ queue is fully characterized by the four tuple consisting of the rate and IDC of the arrival and service processes. Dramatic examples are provided by Tables 2 and 3 from [12], which show comparisons for queues in series exhibiting the heavy-traffic bottleneck phenomenon from [7]. The interarrival time has an H_2 distribution with scv $c_a^2 = 8.0$ but three possible values for the remaining third parameter r. Only RQNA captures the impact of r (necessarily indirectly, because r is not used). From this perspective, RQNA performs far better than the other methods.

3.2. Extensions

(i) Allowing multiple servers at each queue. As can be seen from §5.2 of [8], multiple servers at each node was allowed for QNA. An approach to robust queueing with multiple servers is in [1]. New ideas are needed to extend [10, 11] to multiple servers. *(ii) Allowing multiple classes and/or more general routing.* A provision for multiple classes, where each class had its own routing, was provided in §2.3 of [8]. The algorithm aggregated the input data to convert it into an associated approximate Markovian routing. *(iii) Allowing time-varying arrival processes.* A significant start for a single queue with time-varying arrivals is in [13], but more is needed to treat networks. For background on queues with time-varying arrivals, see [9].

References

1. Bandi, C., Bertsimas, D., Youssef, N.: Robust queueing theory. Oper. Res. **63**(3), 676–700 (2015)
2. Dai, J., Nguyen, V., Reiman, M.I.: Sequential bottleneck decomposition: an approximation method for generalized Jackson networks. Oper. Res. **42**(1), 119–136 (1994)
3. Harrison, J.M., Nguyen, V.: The QNET method for two-moment analysis of open queueing networks. Queueing Syst. **6**(1), 1–32 (1990)
4. Jackson, J.R.: Networks of waiting lines. Oper. Res. **5**(4), 518–521 (1957)
5. Kuehn, P.J.: Approximate analysis of general queuing networks by decomposition. IEEE Trans. Commun. **27**(1), 113–126 (1979)
6. Segal, M., Whitt, W.: A Queueing Network Analyzer for Manufacturing. In: Bonatti, M. (ed.) Teletraffic science for new cost-effective systems, networks and services. In: ITC 12, Proceedings of the 12th International Teletraffic Congress, pp. 1146–1152. Elsevier, North-Holland (1989)
7. Suresh, S., Whitt, W.: The heavy-traffic bottleneck phenomenon in open queueing networks. Oper. Res. Lett. **9**(6), 355–362 (1990)
8. Whitt, W.: The queueing network analyzer. Bell Lab. Tech. J. **62**(9), 2779–2815 (1983)
9. Whitt, W.: Time-varying queues. Queueing Models Serv. Manag. **1**(2), 79–164 (2018)
10. Whitt, W., You, W.: Heavy-traffic limit of the GI/GI/1 stationary departure process and its variance function. Stoch. Syst. **8**(2), 143–165 (2018)
11. Whitt, W., You, W.: Using robust queueing to expose the impact of dependence in single-server queues. Oper. Res. **66**(1), 184–199 (2018)
12. Whitt, W., You, W.: The advantage of indices of dispersion in queueing approximations. Oper. Res. Lett. **47**(2), 99–104 (2019)
13. Whitt, W., You, W.: Time-varying robust queueing. Oper. Res. **67**(6), 1766–1782 (2019)
14. Whitt, W., You, W.: Heavy-traffic limits for stationary network flows. Queueing Syst. **95**, 53–68 (2020)
15. Whitt, W., You, W.: A robust queueing network analyzer based on indices of dispersion. Nav. Res. Logist. **69**(1), 36–56 (2022)
16. Whitt, W., You, W.: Supplement to “a robust queueing network analyzer based on indices of dispersion”. http://www.columbia.edu/~ww2040/allpapers.html (2022)