# New Decomposition Approximations for Queueing Networks

Ward Whitt* ww2040@columbia.edu

Wei You† weiyou@ust.hk

January 1, 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 [7]. 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

$$I_A(t) \equiv \frac{\mathrm{Var}(A(t))}{E[A(t)]}, \quad t \geq 0. \tag{1}$$

As regularity conditions, we assume that $E[A(t)]$ and $\mathrm{Var}(A(t))$ are finite for all $t \geq 0$. 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].

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 $\mu^{-1}$ and squared coefficient of variation (scv) $c_s^2$ of the service time

*Emeritus Professor, Department of Industrial Engineering and Operations Research, Columbia University

†Department of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology

at that queue. With RQ, we replace the stochastic net-input process by a deterministic instance drawn from a pre-determined 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

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

where $b = \sqrt{2}$; see Theorem 2 of [11]. For additional discussion, see [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]. The goal is to allow more bottleneck queues and thus 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 [7] discussed in [5]. Here are some:

*2.2.1.* Allow multiple servers at each queue.

*2.2.2.* Allow multiple classes and/or more general routing.

*2.2.3.* Allow 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 [7] 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 [7] and QNET from [3].

As noted in Sections 1.2 and 7 of [15], RQNA is effective for tree networks. Indeed, RQNA is provably superior for a single queue. First, 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 [6]. These tables show the mean waiting times at each of nine exponential queues in series fed by a rate-1 renewal arrival process. The first eight queues have mean service time, and thus traffic intensity, 0.6, while the last queue has mean service time, and thus traffic intensity, 0.9, making it a bottleneck. The interarrival time has a hyperexponential ($H_2$) distribution with a squared coefficient of variation $c_a^2 = 8.0$ but three possible values for the remaining third parameter: $r = 0.1$, $r = 0.5$ (the common case of balanced means) and $r = 0.9$. (The case $r = 0.1$ makes the process like a batch Poisson process, while the case $r = 0.9$ makes it like a Poisson process; see §V of [8].)

These examples are discussed in the third paragraph of §1.2.2 of [15]. Since the first queue has a renewal arrival process, the approximations for QNA, QNET and SBD all coincide with the heavy-traffic approximation used in (44) of [7], which is independent of $r$. However, these three approximations differ at the last queue, which does not have a renewal arrival process. The main point is that only RQNA captures the impact of $r$ (necessarily indirectly, because $r$ is not used directly). From this perspective, RQNA performs far better than the other methods.

*3.2 Extensions. 3.2.1. Allowing multiple servers at each queue.* As can be seen from §5.2 of [7], multiple servers at each node was allowed for QNA. An approach to robust queueing with multiple servers was offered by [1]. New ideas seem to be needed to extend [10, 11] to multiple servers.

*3.2.2. 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 [7]. The algorithm aggregated the input data to convert it into an associated approximate Markovian routing.

*3.2.3. Allowing time-varying arrival processes.* A significant start was provided for a single queue with time-varying arrivals in [13], but much more is required to treat networks. For additional background on queues with time-varying arrivals, see [9] and references there.

## References


[1] Bandi C, Bertsimas D, Youssef N (2015) Robust queueing theory. *Operations Research* 63(3):676–700.

[2] Dai J, Nguyen V, Reiman MI (1994) Sequential bottleneck decomposition: an approximation method for generalized Jackson networks. *Operations Research* 42(1):119–136.

[3] Harrison JM, Nguyen V (1990) The QNET method for two-moment analysis of open queueing networks. *Queueing Systems* 6(1):1–32.

[4] Jackson JR (1957) Networks of waiting lines. *Operations Research* 5(4):518–521.

[5] Segal M, Whitt W (1989) A Queueing Network Analyzer for Manufacturing. Bonatti M, ed., *Teletraffic Science for New Cost-Effective Systems, Networks and Services. ITC 12, Proceedings of the 12th International Teletraffic Congress*, 1146–1152 (Elsevier, North-Holland).

[6] Suresh S, Whitt W (1990) The heavy-traffic bottleneck phenomenon in open queueing networks. *Operations Research Letters* 9(6):355–362.

[7] Whitt W (1983) The queueing network analyzer. *Bell Laboratories Technical Journal* 62(9):2779–2815.

[8] Whitt W (1984) On approximations for queues, III: Mixtures of exponential distribtuoons. *AT&T Bell Laboratories Technical Journal* 63(1):163–175.

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

[10] Whitt W, You W (2018) Heavy-traffic limit of the GI/GI/1 stationary departure process and its variance function. *Stochastic Systems* 8(2):143–165.

[11] Whitt W, You W (2018) Using robust queueing to expose the impact of dependence in single-server queues. *Operations Research* 66(1):184–199.

[12] Whitt W, You W (2019) The advantage of indices of dispersion in queueing approximations. *Operations Research Letters* 47(2):99–104.

[13] Whitt W, You W (2019) Time-varying robust queueing. *Operations Research* 67(6):1766–1782.

[14] Whitt W, You W (2020) Heavy-traffic limits for stationary network flows. *Queueing Systems* 95:53–68.

[15] Whitt W, You W (2022) A robust queueing network analyzer based on indices of dispersion. *Naval Research Logistics* 69(1):36–56.

[16] Whitt W, You W (2022) Supplement to "a robust queueing network analyzer based on indices of dispersion", http://www.columbia.edu/∼ww2040/allpapers.html.