APPENDIX
to
A Robust Queueing Network Analyzer Based on Indices of Dispersion

Ward Whitt* Wei You†

ww2040@columbia.edu weiyou@ust.hk

March 23, 2020

Abstract

(From the main paper) We develop a robust queueing network analyzer algorithm to approximate the steady-state performance of a single-class open queueing network of single-server queues with Markovian routing. The algorithm allows non-renewal external arrival processes, general service-time distributions and customer feedback. We focus on the customer flows, defined as the continuous-time processes counting customers flowing into or out of the network, or flowing from one queue to another. Each flow is partially characterized by its rate and a continuous function that measures the stochastic variability over time. This function is a scaled version of the variance-time curve, called the index of dispersion for counts (IDC). The required IDC functions for the flows can be calculated from the model primitives, estimated from data or approximated by solving a set of linear equations. A robust queueing technique is used to generate approximations of the mean steady-state performance at each queue from the IDC of the total arrival flow and the service specification at that queue. The algorithm effectiveness is supported by extensive simulation studies and heavy-traffic limits.

Keywords: *queueing networks, non-Markov queueing networks, robust queueing, index of dispersion, queueing approximations, heavy traffic*

*Department of Industrial Engineering and Operations Research, Columbia University

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

1 Overview

This appendix provides additional supporting material for the main paper. The main paper itself is a culmination of our research reported in [38, 39, 40, 41]. We presented our new approach to queueing approximations using robust queueing based on indices of dispersion in [39]. Readers should consult that source, including the e-companion, for discussion of the basic ideas. In §6 of [39] we presented a framework for a new RQNA. The present paper elaborates and refines that framework, and demonstrates that it can be remarkably effective. To develop an effective RQNA for an open queueing network (OQN), we relied on new heavy-traffic limits established in [38, 41]. We present a few more in this appendix.

We start in §2 by providing some additional literature review. In particular, we provide additonal motivation and discuss how approximations for non-Markov OQNs build on and extend the classic theory for Markov OQNs. In §3 we discuss ways to calculate the IDC from model specifications and estimate it from data, either from a real system or from a simulation model. In §4 we provide more details on the feedback elimination procedure, elaborating on §4 of the main paper. In §5 we provide additional heavy-traffic support for our IDC equations in §3.4 of the main paper. In §6 we discuss a tuning function for more flexible model tuning. Finally, in §7 we discuss additional numerical experiments.

2 Literature Review

We now supplement the main paper by providing some additional literature review.

2.1 From Markov OQNs to non-Markov OQNs

One of the most important developments in queueing theory has been the theory of Jackson networks initiated by Jackson [22]. A Jackson network is a queueing network with Poisson external arrival processes, exponential service-time distributions, FCFS service discipline and Markovian routing policy. This model is especially tractable because the queue length vector completely characterizes the system state and forms a Markov process, hence it is also called a *Markov OQN*. Jackson [22] showed that the steady-state vector for the number of customers at each queue in a Jackson network has a product-form distribution with geometric marginal distributions. Hence in steady-state the network can be viewed as if it is decomposed into mutually independent $M/M/1$ stations (in Kendall's notation), even though the queueing processes are not in fact independent.

This initial breakthrough was followed by vigorous research leading to an elaborate and useful theory, as can be seen from [10, 24, 33]. Due to its closed-form and product-form solution, Jackson networks have been widely studied, e.g. in [5, 27, 30]. Jackson networks have also been applied to many service systems. For ride-sharing economy, [7] studied the optimal platform pricing, while [32] looked at the inventory rebalancing and vehicle routing problems. [8, 26, 35, 42] analyzed resource allocation and quality-of-service in cloud computing system. For healthcare related problems, [6] studies hospital stuffing strategy to achieve optimal workflow efficiency under information security requirements; see also [20] for an overview.

However, applications in communication, manufacturing and service systems are often complicated by significant deviations from the tractable structure of a Markov OQN. For call centers and hospitals, the external arrival processes is often well approximated by Poisson processes. However, dependence in arrival processes may still be induced by over-dispersion, e.g., see [25] and references there. In most manufacturing systems, an external arrival process is often far less variable than a Poisson process by design. Even if external arrival processes can be regarded as Poisson processes, service-time distributions are often non-exponential, see [9, 18]. This is often resulted from complicated processing operations, such as those involving batching.

Non-exponential interarrival-time or service-time distributions produce complicated dependence structure in the departure processes, which will be inherited by the arrival processes at the subsequent stations. Then these processes cannot be renewal processes because (i) a departure process from any $GI/GI/1$ queue is necessarily non-renewal if the interarrival-time or service-time distribution is non-exponential and (ii) the superposition of independent renewal processes cannot be renewal unless all components are Poisson processes (in which case the superposition process is also Poisson); e.g., see [14, 15, 17].

Indeed, such dependence in departure processes is consistent with the heavy-traffic limit theorem for the stationary customer flows developed in [38, 41]. The results show that the dependence structure in stationary customer flows depends on the traffic intensity and the interarrival-time and service-time distributions in a nontrivial manner.

Furthermore, dependence among different arrival and service processes are often observed in manufacturing/communication systems. Upon service completion, jobs are directed to subsequent stations. This corresponds to splitting the departure process, which introduces dependence among the sub-flows after splitting. In hospital settings, patients may revisit a doctor after completing several tests. In manufacturing lines, products may need rework

after quality-control testings. This is referred to as *customer feedback*, which necessarily introduce dependence between the service and arrival processes.

2.2 Motivation and Main Contribution

Despite many attempts to develop more effective analyzers, early approximations such as QNA [36] or Monte Carlo simulation remain to be the most popular choices, mainly due to the ease of implementation. For example, [3] identified the major bottleneck in a health center appointment clinic, where they applied the QNA algorithm to approximate the system performance; [12] studied the effect of service interrupts and hospital resources pooling on patient flow times; [23] also applied decomposition method and two-moment approximations to analyze the impact of parallelization of care on customer sojourn time; [2] integrated simulation and optimization to find the optimal staffing allocation in an emergency department unit, where they considered a network of $M_t/G/1$ queues and the stochastic objective function is estimated by simulation; and [16] also studied the resource allocation problem in general stochastic networks by simulation optimization.

Given recent applications in modern service systems, there remains a need for fast and accurate performance analyzers for non-Markov OQNs. The main contribution of this paper is to develop a new performance analyzer that is as easy to implement as the original QNA algorithm [36], but at the same time produces much more accurate performance approximations. We develop the first approximation algorithm for non-Markov OQNs based on non-parametric traffic descriptions. Our approach specifically addresses the challenges posed by the complicated dependence in queues.

3 The Calculation and Estimation of IDC

3.1 The IDC's for Renewal Processes

For renewal processes, the variance $\mathrm{Var}(A(t))$ and thus the IDC $I_a(t)$ can either be calculated directly or can be characterized via their Laplace transforms and thus calculated by inverting those transforms and approximated by performing asymptotic analysis. Because we are interested in the steady-state behavior of the OQN, we are primarily interested in the equilibrium renewal process, as in §3.5 of [31]; see Remark 3 of the main paper.

In turns out that the variance of the equilibrium arrival renewal process $V(t) \equiv \mathrm{Var}(A(t))$ can be expressed in terms of the renewal function $m(t) \equiv E[A_0(t)]$, where A_0 is the corre-

sponding ordinary renewal process. For a function f, let $\hat{f}$ denote the Laplace transform of f, defined by

$$\hat{f}(s) \equiv \mathcal{L}(f)(s) \equiv \int_0^\infty e^{-st} f(t) dt.$$

The following formula is taken from §2 of [38]

$$\hat{V}(s) = \frac{\lambda}{s^2} + \frac{2\lambda}{s}\hat{m}(s) - \frac{2\lambda^2}{s^3} = \frac{\lambda}{s^2} + \frac{2\lambda}{s}\frac{\hat{g}(s)}{s\,(1-\hat{g}(s))} - \frac{2\lambda^2}{s^3}, \tag{1}$$

where g is the density function of the interarrival-time dsitribution. The variance function can then be obtained numerically, which is discussed in §13 of [1]. The hyperexponential (H_2) and Erlang (E_2) special cases are described in §III.G of [19].

It is also possible to carry out similar analyses for much more complicated arrival processes. [29] applies matrix-analytic methods to give explicit representations of the variance $mathrmVar(A(t))$ for the versatile Markovian point process or Neuts process; see §5.4, especially Theorem 5.4.1. Explicit formulas for the Markov modulated Poisson process (MMPP) are given on pp. 287-289.

3.2 Numerical Estimation of the IDC from Data.

Now we present an algorithm to numerically estimate the variance $V(t) = \mathrm{Var}(A(t))$ from a given realized sample path of the stationary point process $A(t)$. The main idea is based on Section 5.4 (iii) of [11].

Our goal is to estimate $V(t)$ for $0 < t < t_0$ using a realization of $A(t)$ for $0 < t < T$. The simplest way is to apply crude Monte Carlo method to estimate $V(t)$ for a fixed t and repeat over a finite grid of t's. This method divide the sample path of $A(t)$ into non-overlapping intervals of length t and count the number of arrivals in each interval. The variance is then estimated by the sample variance of the counts. This method is simple to implement but can be slow to converge.

To accelerate the crude Monte Carlo method, we apply three techniques: (i) we use overlapping intervals instead of non-overlapping ones, which introduces bias but reduces sample variance; (ii) we calculate $V(t)$ only over a finite grid equally spaced in the logarithm scale instead of the linear scale; ; and (iii) we re-use the tallied number of events for shorter intervals to calculate the total number of events for longer interval, which avoids repetitive counting. We discuss the three techniques in turn:

Remark 1 *(justifying the logarithmic scale) To justify the logarithm scale in (ii), we remark that the IDC of most stationary processes converges exponentially fast to a constant, as*

the time t increases. In particular, this holds for Markov arrival processes, which includes hyperexponential renewal process, Erlang renewal process, and Markov modulated Poission Process as special cases; e.g.. see Ch. XI of [4], [28] ot [29].

To use overlapping intervals, consider first $k = T/t$ number of non-overlapping intervals, each with length t. Now, we further divide each intervals of length t in to r intervals of the same length $\tau = t/r$. Hence we have rk number of non-overlapping intervals of length τ. Let n_i be the number of events fall in the i-th interval, consider

$$U_i \equiv A(I_i) \equiv A[i\tau, (i+r)\tau) = n_i + n_{i+1} + \cdots + n_{i+r-1}, \quad i = 0, 1, \ldots, rk - r + 1.$$

We estimate $V(t)$ with the sample variance $\bar{V}_l$ of $\{U_i\}_{i=1}^{l}$, where $l = rk-r+1$. This estimator is in general biased but can achieve lower variance compared with the one obtained with crude Monte Carlo method. In §3.3 we show that this estimator of $V(t)$ is asymptotically consistent under mild conditions that $V(t)$ is differentiable with derivative $\dot{V}(t)$ having finite positive limits as $t \to \infty$.

For the third technique, we now present a algorithm to simultaneously estimate $V(2^i\tau)$ for some $\tau > 0$ and $i = 0, 1, \ldots, l$. Let $\{I_i\}$ be the collection of non-overlapping intervals of length τ that covers $[0, T]$. Let $n_i = A(I_i)$ be the number of events on interval I_i. Then we have the following table from [11].

	time horizon t			
sample	τ	2τ	$2^2\tau$	$\cdots$
1	n_1	$n_1 + n_2$	$n_1 + n_2 + n_3 + n_4$	$\cdots$
2	n_2	$n_2 + n_3$	$n_3 + n_4 + n_5 + n_6$	$\cdots$
3	n_3	$n_3 + n_4$	$n_5 + n_6 + n_7 + n_8$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

We find the estimation of $V(2^i\tau)$ by calculating the sample variance of the corresponding column.

Now that we have a efficient algorithm to estimate $V(2^i\tau)$ for fixed τ, we have obtained the estimations of a grid equally spaced in logarithm scale. To obtain estimations for finer grids we shift the crude grid by picking several $\tau \le \tau_j \le 2\tau$ equally spaced in log scale and, for each j, simultaneously estimate $V(2^i\tau_j)$ for all i.

3.3 Consistency of the estimator

In this section we provide theoretical support for our algorithm to estimate the IDC from data in §3.2.

We show that the estimator of the variance function $V(t)$ is asymptotically consistent under mild regularity conditions that $V(t)$ is differentiable with derivative $\dot{V}(t)$ having finite positive limits as $t \to \infty$, i.e.,

$$\dot{V}(t) \to \sigma^2 \text{ as } t \to \infty,$$

for an appropriate constant σ^2. This condition is also used in §3.3 of [39].

Theorem 1 (Consistency of the estimator) *Let A be a time-stationary and ergodic point process with variance function $V(t)$ that is differentiable with derivative $\dot{V}(t)$ having finite positive limit as $t \to \infty$, i.e.,*

$$\dot{V}(t) \to \sigma^2 \text{ as } t \to \infty.$$

Then we have

$$\lim_{l\to\infty} bias(\bar{V}_l) = 0$$

for $l = rk - r + 1$, $r = t/\tau, k = T/t$ and $\bar{V}_k$ is the sample variance of $\{U_i\}_{i=1}^k$. Furthermore,

$$\lim_{l\to\infty} \bar{V}_l = V(t), \text{ w.p.1.}$$

Proof. Let $K = rk - r + 1$ be the sample size, and assume that $V(t) = I(t)t < Ct$ for some constant C. Then

$$\begin{aligned}
E\left[\bar{V}\right] &= \frac{1}{K-1}\sum_{i=1}^{K} E\left[U_i^2\right] - \frac{1}{K(K-1)} E\left[\left(\sum_{i=1}^{K} U_i\right)^2\right] \\
&= \frac{1}{K-1}\left(\sum_{i=1}^{K} E\left[U_i^2\right] - \frac{1}{K} E\left[\sum_{i=1}^{K} U_i^2 + 2\sum_{i>j} U_i U_j\right]\right) \\
&= E\left[U_1^2\right] - E\left[U_1\right]^2 - \frac{2}{K(K-1)}\sum_{i<j} \mathrm{cov}(U_i, U_j) \\
&= V(t) - \frac{2}{K(K-1)}\left(\sum_{j<i<j+r} \mathrm{cov}(U_i, U_j) + \sum_{i>j+r+1} \mathrm{cov}(U_i, U_j)\right) \\
&= V(t) - \frac{2}{K(K-1)}\left(\sum_{i=1}^{r-1}(K-i)\mathrm{cov}(U_1, U_{i+1}) + \sum_{i=r}^{K-1}(K-i)\mathrm{cov}(U_1, U_{i+1})\right) \\
&\equiv V(t) - (A + B)
\end{aligned}$$

The covariance terms can be expressed as

$$\mathrm{cov}(U_1, U_{1+i}) = \begin{cases} V(t - i\tau) + V(t + i\tau) - V(t) - V(i\tau), & i = 1, 2, \ldots, r-1 \\ V(t + i\tau) - 2V(i\tau) + V(i\tau - t), & i = r, r+1, \ldots, K-1 \end{cases} \tag{2}$$

Using the bound on $I(t)$, we have

$$\begin{aligned} A &= \frac{2}{K(K-1)} \sum_{i=1}^{r-1} (K-i)\mathrm{cov}(U_1, U_{i+1}) \\ &\leq \frac{2}{K} \sum_{i=1}^{r-1} (V(t-i\tau) + V(t+i\tau)) \\ &\leq \frac{4Ct(r-1)}{K} \leq \frac{4Ct}{k-1}, \end{aligned}$$

and

$$\begin{aligned} B &= \frac{2}{K(K-1)} \sum_{i=r}^{K-1} (K-i)\mathrm{cov}(U_1, U_{i+1}) \\ &\leq \frac{2}{K} \sum_{i=r}^{K-1} ((V(t+i\tau) - V(i\tau)) - (V(i\tau) - V(i\tau - t))) \\ &\leq \frac{2t}{K} \sum_{i=r}^{K-1} \left(\frac{V(t+i\tau) - V(i\tau)}{t} - \frac{V(i\tau) - V(i\tau - t)}{t} \right) \\ &\to 0, \text{ as } k \to \infty, \end{aligned}$$

where we used the regularity condition that $\dot{V}(t) \to \sigma^2$ as $t \to \infty$, and the fact that the average converges to 0 if the summands converge to 0.

Note that

$$\bar{V}_k \equiv \frac{1}{k-1} \sum_{i=1}^{k} U_i^2 - \frac{1}{k(k-1)} \left(\sum_{i=1}^{k} U_i \right)^2$$

By Continuous Mapping Theorem, we need only prove that both $\{U_i\}$ and $\{U_i^2\}$ follows Strong Law of Large Number (SLLN). This in turns is implied by the Strong Ergodic Theorem for stationary and ergodic sequence. The stationarity of both sequences are implied by the time-stationarity of the point process $N(t)$. The ergodicity of both sequence follows from the ergodicity of the underlying process $N(t)$. □

4 More on Feedback Elimination

The modified system is a single-server queue with a new service-time distribution and without feedback. Let N_p denote a geometric random variable with success probability $1-p$ and support $\mathbb{N}^+$, the positive natural numbers, then the new service time can be expressed as

$$S_p = \sum_{i=1}^{N_p} S_i, \tag{3}$$

where S_i's are i.i.d. copies of the original service times. This modification in service times results in a change in the service scv. By the conditional variance formula, the scv of the total service time is $\tilde{c}_s^2 = p + (1-p)c_s^2$. The new service IDC in the modified system is the IDC of the stationary renewal process associated with the new service times. To obtain the new service IDC, we need only find the Laplace Transform of the new service distribution, then apply the algorithm in §3.1.

Let g_p denote the density function of the new service time, we have

$$\begin{aligned}\hat{g}_p(s) &\equiv E\left[\exp\left(-s\sum_{i=1}^{N_p} S_i\right)\right] = E\left[E\left[\exp\left(-s\sum_{i=1}^{N_p} S_i\right)\middle| N_p\right]\right] \\ &= E\left[\prod_{i=1}^{N_p} E\left[\exp\left(-sS_i\right)\right]\right] = E\left[\hat{g}^{N_p}(s)\right] = M_p(\hat{g}(s)),\end{aligned}$$

where $\hat{g}(s)$ is the Laplace transform of the original service distribution and M_p is the probability generating function of the geometric random variable described above.

An alternative algorithm eliminates only all near-immediate feedback from the bottleneck queues, where a *bottleneck queue* is a station with a traffic intensity that equals the highest traffic intensity in the network. For each bottleneck queue in the network, by the definition of near-immediate feedback, we eliminate all feedback at this queue when we analyze the mean workload at that queue, even if the feedback flow passes through other bottleneck queues.

To help understand near-immediate feedback, consider a modified OQN with one bottleneck queue, denoted by h, while all non-bottleneck queues have service times set to 0 so that they serve as instantaneous switches. In the reduced network, we define an external arrival $\hat{A}_0$ to the bottleneck queue to be any external arrival that arrive at the bottleneck queue for the first time. Hence, an external arrival may have visited one or multiple non-bottleneck queues before its first visit to the bottleneck queue. In particular, the external arrival process can be expressed as the superposition of (i) the original external arrival process $A_{0,h}$ at station h; and (ii) the Markov splitting of the external arrival process $A_{0,i}$ at station i with probability $\hat{p}_{i,h}$, for $i \neq h$, where $\hat{p}_{i,h}$ denote the probability of a customer that enters the original system at station i ends up visiting the bottleneck station h. For the explicit formula of $\hat{p}_{i,h}$, see Remark 3.2 of [41].

In §4.2 of [41], we showed that this reduced network is asymptotically equivalent in the HT limit to the single-server queue with i.i.d. feedback that we considered in §4.1. In particular, the arrival process of the equivalent single-station system is $\hat{A}_0$ as described above, the service times remain unchanged and the feedback probability is $\hat{p}$, which is exactly the probability

of a near-immediate feedback in the original system; see (3.9) of [41] for the expression of $\hat{p}$. Hence we showed that eliminating all feedback at the bottleneck queue as described above prior to analysis is asymptotically correct in HT for OQNs with a single bottleneck queue in terms of the queue length process, the external departure process, the workload process and the waiting time process. Moreover, the different variants of the algorithm - eliminating all near immediate feedback or only the near-immediate feedback at the bottleneck queues - are asymptotically exact in the HT limit for an OQN with a single-bottleneck queue, because only the bottleneck queues have nondegenerate HT limit. In contrast, if there are multiple bottleneck queues, the HT limit requires multidimensional RBM, which is not used in our RQNA.

5 Supporting Heavy Traffic Limits

In this section, we provide detailed HT limit support for the IDC equations discussed in §3.4.

5.1 Heavy-Traffic Limits for Departure Processes

We first provide theoretical support for the approximation of the departure IDC in (17) of §3.3.1 of the main paper. That approximation is ultimately supported by the heavy-traffic limit theorem obtained in Theorem 4.1 of [41]. To use that result, we start by presenting a slight variant of it. We refer to §3.2 of [41] for the notation used here.

Lemma 1 *Under the assumption of Theorem 4.1 of* [41]*, the HT limit of the departure process of the bottleneck station h can be written as*

$$D_h^* = \tilde{Q}_h^*(0) + \tilde{A}_h^* - \tilde{Q}_h^*, \tag{4}$$

where

$$\tilde{A}_h^* = e_h'(I - P')^{-1}\left(A_0^* + (\Theta^*)'\mathbf{1}\right) \tag{5}$$

and

$$\tilde{Q}_h^* = \frac{1}{1-\hat{P}_h} Q_h^* = \psi\left(\tilde{Q}_h^*(0) + \tilde{A}_h^* - S_h^* - \lambda_h e\right). \tag{6}$$

As a result, the limiting variance function of the departure process is where

$$V_{d,h}^*(t) = w^*(\lambda_h t/c_{x,h}^2)c_{a,h}^2\lambda_h t + (1 - w^*(\lambda_h t/c_{x,h}^2))c_{s,h}^2\lambda_h t, \tag{7}$$

where $w^*(t)$ is the weight function in (19) of §3.3.1 of the main paper. The variability parameter is $c_{x,h}^2 = c_{a,h}^2 + c_{s,h}^2$ with $c_{s,h}^2$ being the service scv and $c_{a,h}^2$ being the limiting variability of the total arrival at station h, given by $c_{a,h}^2 \equiv \mathrm{Var}(\tilde{A}_h^*)/\lambda_h t$.

Proof. Start by claiming that

$$e_h' \hat{P}'_{\mathcal{H}^c,\mathcal{H}} e_h = \frac{1}{1-\hat{P}_h}, \quad \lambda_h = \frac{\hat{\lambda}_{0,h}}{1-\hat{P}_h}$$

and that

$$\frac{1}{1-\hat{P}_h}\left(e_h' + \hat{P}'_{\mathcal{H}^c,\mathcal{H}} e'_{\mathcal{H}^c}\right) = e_h'(I-P')^{-1}.$$

In fact, all three assertions can be check by writing the transition matrix in blocks according to two sets of indices $\{\mathcal{H}, \mathcal{H}^c\}$.

Now, (6) follows from dividing both sides of the limiting queue length process in Theorem 4.1 of [41] by $(1-\hat{P}_h')$ and the fact that $\psi(f/c) = \psi(f)/c$ for any function f and constant c.

The limiting variance function is derived in the exact same way as in Theorem 5.3 of [38] by noting that $\tilde{A}_h^*$ and S_h^* are two independent Brownian motions. The only change here is that we have an additional tuning function $h(\rho)$. This, however, does not change the argument, since we require that $\lim_{\rho\uparrow 1} h(\rho) = 1$. □

The approximation in (17) of §3.3.1 of the main paper is then justified by the exact same procedure as described in §6.2 of [38].

Remark 2 (The choice of the approximation in (17) of §3.3.1 of the main paper) As in any approximation based on heavy-traffic limits, it is possible to have different approximations that converge to the same limit. We propose to add a correction term to achieve exact light traffic limit while keeping the HT limit unchanged. In particular, in our approximation (18) of §3.3.1 of the main paper we have an extra ρ_i in the denominator of the term inside $w^*(\cdot)$. As $\rho_i \downarrow 0$, or equivalently, as the service time at station i become negligible in compare with the interarrival times, the departure IDC will converge to the arrival IDC. This is preserved in our approximation in our approximation (18) of §3.3.1 of the main paper by virtue of the additional correction term. □

5.2 Heavy-Traffic Limits for Splitting

We now provide additional theoretical support for the splitting approximation in §3.3.2. For that purpose, let

$$\Theta_i(n) \equiv (\Theta_{i,1}(n), \ldots, \Theta_{i,K}(n)) = \sum_{l=1}^{n} \theta_i^l$$

denote the splitting decisions up to the n-th decision at station i. Consider the diffusion-scaled processes indexed by ρ

$$D_{i,\rho}^*(t) = (1-\rho)\left[D_i((1-\rho)^{-2}t) - \lambda_i(1-\rho)^{-2}t\right],$$

$$\Theta^*_{i,\rho}(t) = (1-\rho)\left[\sum_{l=1}^{\lfloor (1-\rho)^{-2}t \rfloor} \theta^l - \mathbf{p}_i(1-\rho)^{-2}t\right] \in \mathcal{D}^K, \tag{8}$$

$$\mathbf{A}^*_{i,\rho}(t) = (1-\rho)\left[\mathbf{A}_i((1-\rho)^{-2}t) - \lambda_i \mathbf{p}_i(1-\rho)^{-2}t\right] \in \mathcal{D}^K,$$

for $t \geq 0$, where $\mathbf{p}_i \equiv E[\theta^l_i]$ is the i-th row of the routing matrix and $\mathbf{A}_{i,\rho} = (A_{i,j,\rho} : j = 1, 2, \ldots, K)$ is the vector consists of all the streams after splitting. The following result rephrases Theorem 9.5.1 in [37].

Theorem 2 *Suppose that*

$$(D^*_{i,\rho}, \Theta^*_{i,\rho}) \Rightarrow (D^*_i, \Theta^*_i) \quad as \quad \rho \uparrow 1 \quad in \quad D^{K+1} \tag{9}$$

and that almost surely D^* and $\Theta^* \circ \lambda \mathbf{e}$ have no common discontinuities of opposite sign. Then

$$\mathbf{A}^*_{i,\rho} \Rightarrow \mathbf{A}^*_i \quad in \quad D^K,$$

with

$$A^*_{i,j} \equiv p_{i,j} D^* + \Theta^*_{i,j} \circ \lambda_i e, \quad for \quad 1 \leq j \leq K, \tag{10}$$

where $e(t) = t$ is the identity mapping.

Remark 3 (*splitting the departures from a $G/GI/1$ queue*) If we split the departure process from the $GI/GI/1$ model with Markovian routing, then D^* is independent of Θ^* and Θ^* is a zero-drift K-dimensional Brownian motion with covariance matrix $\Sigma = (\sigma_{i,j}) \in \mathbb{R}^{K \times K}$, where $\sigma^2_{i,i} = p_i(1-p_i)$ and $\sigma^2_{i,j} = -p_i p_j$ for $i \neq j$. Hence, from (10) we obtain

$$\mathbf{A}^* = \mathbf{p} D^* + \Theta^* \circ \lambda e, \tag{11}$$

which is consistent with our approximation (22) of §*3.3.2* of the main paper and thus also for approximation (23) there.

Theorem 2 assumes only a joint FCLT for the flow to split and the splitting decision process, so dependence is allowed. Thus it provides support for the general splitting equation in approximation (24) of §3.3.2 of the main paper for the case where $D_{i,j}$ and $\Theta_{i,j}$ are correlated. Furthermore, define the HT-scaled correction term as

$$\alpha^*_{i,j,\rho}(t) \equiv \alpha_{i,j}((1-\rho)^{-2}t). \tag{12}$$

Finally, define the limiting correction term as

$$\alpha^*_{i,j}(t) \equiv 2\mathrm{cov}(p_{i,j} D^*_i(t), \Theta^*_{i,j}(\lambda_i t))/p_{i,j}\lambda_i t. \tag{13}$$

The following corollary follows from Theorem 2.

Corollary 1 *Under the assumptions in Theorem 2 plus the uniform integrability conditions, we have* $\alpha^*_{i,j,\rho}(t) \Rightarrow \alpha^*_{i,j}(t)$ *as* $\rho \uparrow 1$.

Proof. By the definitions of the correction term in (15) and HT-scaled processes, we write

$$\begin{aligned}
\alpha^*_{i,j,\rho}(t) &= \alpha_{i,j}((1-\rho)^{-2}t) \\
&= I_{a,i,j}((1-\rho)^{-2}t) - p_{i,j}I_{d,i}((1-\rho)^{-2}t) - (1-p_{i,j}) \\
&= \frac{\mathrm{Var}((1-\rho)A_{i,j}((1-\rho)^{-2}t))}{p_{i,j}\lambda_i t} - p_{i,j}\frac{\mathrm{Var}((1-\rho)D_i((1-\rho)^{-2}t))}{\lambda_i t} - (1-p_{i,j}) \\
&= \frac{\mathrm{Var}(A^*_{i,j,\rho}(t))}{p_{i,j}\lambda_i t} - p_{i,j}\frac{\mathrm{Var}(D^*_{i,\rho}(t))}{\lambda_i t} - (1-p_{i,j}) \\
&\Rightarrow \frac{\mathrm{Var}(A^*_{i,j}(t))}{p_{i,j}\lambda_i t} - p_{i,j}\frac{\mathrm{Var}(D^*_i(t))}{\lambda_i t} - (1-p_{i,j}) = \alpha^*_{i,j}(t).
\end{aligned}$$

□

This corollary supports the following approximation for the correction term $\alpha_{i,j}$ in

$$\alpha_{i,j}(t) \approx \alpha^*_{i,j}((1-\rho)^2 t) \tag{14}$$

with $\alpha^*_{i,j}$ defined in (13).

5.3 An Approximation Scheme for General Correction Terms

In a general open queueing network with feedback and superposition of dependent flows, the correction terms $\alpha_{i,j}$ and β_i can be non-trivial. The key idea is that, for each correction term, we select a suitable queue and assume it to be the bottleneck queue. Then we apply Theorem 4.1 of [41] to obtain HT approximation of the correction terms and utilize Corollary 5.1 of [38] to obtain explicit form of the correction term. We now discuss the two types of correction terms in turn.

5.3.1 Dependent Splitting: the Correction Term $\alpha_{i,j}$.

So that the additional correction term $\alpha_{i,j}$ is defined as

$$\alpha_{i,j}(t) \equiv I_{a,i,j}(t) - p_{i,j}I_{d,i}(t) - (1-p_{i,j}). \tag{15}$$

Unfortunately, the covariance in (13) is complicated. We do obtain a useful approximation under the extra condition that only queue i enters heavy traffic.

For any $\alpha_{i,j}$, the relevant routing flow is $A_{i,j}$ while the relevant departure flow is D_i. Naturally, we choose station i to be the HT station. So we let $\rho_i = \rho \uparrow 1$ and keep $\rho_j < 1$

for $j \neq i$. Define the HT scaled processes as in §3.2 of [41] and apply Lemma 1 with $h = i$, we have

$$D^*_{i,\rho} \Rightarrow D^*_i = \tilde{A}^*_i + \tilde{Q}^*_i(0) - \tilde{Q}^*_i. \tag{16}$$

For the routing flow $A_{i,j}$, we apply Theorem 2 so that

$$A^*_{i,j,\rho} \Rightarrow A^*_{i,j} = p_{i,j} D^*_i + \Theta_{i,j} \circ \lambda_i e \quad \text{as} \quad \rho \uparrow 1. \tag{17}$$

Define the correction term $\alpha^*_{i,j}$ as in (14), then Theorem 4.1 of [41] implies the following corollary, which leads to the correction term in approximation (25) of §3.3.2 of the main paper.

Theorem 3 *Under the assumptions in Theorem* 4.1 *of* [41] *and Theorem* 2 *plus the uniform integrability conditions, we have*

$$\begin{aligned} \alpha^*_{i,j,\rho}(t) &\Rightarrow 2\text{cov}(p_{i,j} D^*_i(t), \Theta^*_{i,j}(\lambda_i t))/(p_{i,j}\lambda_i t) \\ &= 2\xi_{i,j} p_{i,j}(1 - p_{i,j}) w^*(\lambda_i t / c^2_{x,i}), \quad as \quad \rho \uparrow 1, \end{aligned} \tag{18}$$

where $\xi_{i,j}$ is the $(i,j)^{\text{th}}$ entry of the matrix $(I - P')^{-1}$, $c^2_{x,i} = c^2_{a,i} + c^2_{s,i}$ and $c^2_{a,i}$ is the limiting variability parameter as solved from the IDC equation system in (31) *of* §3.4 *of the main paper and $c^2_{s,i}$ is the scv of the service distribution at station i.*

Proof. Apply Theorem 4.1 of [41] to obtain expression for $D^*_i(t)$, then apply Corollary 5.1 of [38] for the explicit covariance in (18). □

As a direct result of Theorem 3, we propose to define the correction term as

$$\alpha_{i,j,\rho}(t) = 2\xi_{i,j} p_{i,j}(1 - p_{i,j}) w^*((1-\rho)^{-2}\lambda_i t/(\rho c^2_{x,i})), \tag{19}$$

which is asymptotically exact as $\rho \uparrow 1$.

5.3.2 Dependent Superposition: the Correction Term β_i.

Next, we consider the correction term β_i associated with dependent superposition. From (27) of §3.3.3 of the main paper, it suffices to specify $\beta_{k,i;j,i}$ for any station i and any pair of sub-flows $(A_{j,i}, A_{k,i})$ at that station. We assume without loss of generality that (i) $\rho_j \geq \rho_k$, or (ii) $\rho_j = \rho_k$ and $\lambda_{j,i} \geq \lambda_{k,i}$. In the case (ii), we break the tie by picking the index that gives the larger rate $\lambda_{j,i}$. In both cases, we consider station j to be the HT station while keep all other stations unsaturated.

By Theorem 4.1 of [41], we have

$$\begin{aligned} A^*_\rho \Rightarrow A^* &= \tilde{A}^* + \gamma_j \left(\tilde{Q}^*_j(0) - \tilde{Q}^*_j \right) \\ D^*_{j,\rho} \Rightarrow D^*_j &= \tilde{A}^*_j + \tilde{Q}^*_j(0) - \tilde{Q}^*_j, \\ D^*_{l,\rho} \Rightarrow D^*_l &= A^*_l, \quad \text{for} \quad l \neq j, \end{aligned}$$

where

$$\tilde{A}^* = (I - P')^{-1} \left(A^*_0 + (\Theta^*)' \mathbf{1} \right),$$

$\tilde{Q}^*_j$ is defined in Lemma 1 with $h = j$ and $\gamma_j \in \mathbb{R}^K$ is defined as

$$\gamma_j = P'(I - P')^{-1} e'_j (1 - \hat{P}_j)$$

with $\hat{P}_j$ defined as in (3.9) of [41] with $\mathcal{H} = \{j\}$.

Furthermore, Theorem 2 gives

$$\begin{aligned} A^*_{j,i} &= p_{j,i} D^*_j + \Theta^*_{j,i} \circ \lambda_j e \\ &= p_{j,i} \tilde{A}^*_j + \Theta^*_{j,i} \circ \lambda_j e + p_{j,i} (\tilde{Q}^*_j(0) - \tilde{Q}^*_j) \qquad (20) \\ A^*_{k,i} &= p_{k,i} D^*_k + \Theta^*_{k,i} \circ \lambda_k e \\ &= p_{k,i} \tilde{A}^*_k + \Theta^*_{k,i} \circ \lambda_k e + p_{k,i} \gamma_{j,k} (\tilde{Q}^*_j(0) - \tilde{Q}^*_j). \qquad (21) \end{aligned}$$

We utilize the following approximations

$$A^*_{k,i} \approx p_{k,i} \tilde{A}^*_k + \Theta^*_{k,i} \circ \lambda_k e \equiv \tilde{A}^*_{k,i} \qquad (22)$$

and

$$p_{j,i} \tilde{Q}^*_j \approx \psi \left(p_{j,i} \tilde{Q}_j(0) + p_{j,i} A^*_j + \Theta^*_{j,i} \circ \lambda_j e - p_{j,i} S^*_j - p_{j,i} \lambda_j e \right) \equiv \tilde{Q}^*_{j,i}. \qquad (23)$$

By Corollary 5.1 of [38]

$$2\text{cov} \left(\tilde{A}^*_{k,i}(t), \tilde{A}^*_{j,i}(t) - \tilde{Q}^*_{j,i}(t) \right) / (\lambda_i t) = 2 \frac{\zeta_{j,i;k,i}}{\lambda_i} w^*(t / c^2_{x,j}), \qquad (24)$$

where $\tilde{A}^*_{j,i} \equiv p_{j,i} \tilde{A}^*_j + \Theta^*_{j,i} \circ \lambda_j e$ and $\zeta_{j,i;k,i}$ is the constant defined as

$$\zeta_{j,i;k,i} = \frac{1}{t} \text{cov} \left(\tilde{A}^*_{k,i}(t), \tilde{A}^*_{j,i}(t) \right). \qquad (25)$$

Note that $\zeta_{j,i;k,i}$ is a constant independent of t since $\tilde{A}^*_{k,i}(t)$ and $\tilde{A}^*_{j,i}(t)$ are Brownian motions.

Finally, we define

$$\beta_{j,i;k,i}(t) = \beta_{k,i;j,i}(t) = 2 \frac{\zeta_{j,i;k,i}}{\lambda_i} w^*((1 - \rho_j)^2 p_{j,i} \lambda_j t / (\rho c^2_{x,j,i})), \qquad (26)$$

where $c^2_{x,j,i} = p_{j,i} c^2_{a,j} + (1 - p_{j,i}) + p_{j,i} c^2_{s,j}$ and $c^2_{a,j}$ is solved from (34) of §3.4 of the main paper.

The following lemma gives explicit formula for $\zeta_{j,i;k,i}$. Let $\nu_l \equiv p_{l,i} e'_l (I - P')^{-1}$ for $l = j, k$, where e_i is the i-th unit vector.

Lemma 2

$$\zeta_{j,i;k,i} = \nu_j' \left(\mathrm{diag}(c_{a,0,i}^2 \lambda_i) + \sum_{l=1}^{K} \Sigma_l \right) \nu_k + \nu_k' \Sigma_j e_i + \nu_j' \Sigma_k e_i, \tag{27}$$

where $\mathrm{diag}(c_{a,0,i}^2 \lambda_i)$ *is the diagonal matrix with* $c_{a_0,i}^2 \lambda_i$ *as the* i*-th diagonal entry,* Σ_l *is the covaraince matrix of Brownian limit of the splitting decision process* $(\Theta_{l,i}^*)_{i=1}^K$ *at station* l *defined as* $\Sigma_l \equiv (\sigma_{i,j}^l)$ *with* $\sigma_{i,i}^l = p_{l,i}(1-p_{l,i})\lambda_l$ *and* $\sigma_{i,j}^l = -p_{l,i}p_{l,j}\lambda_l$ *for* $i \neq j$.

Proof. By the definition of $\tilde{A}^*$ and $\tilde{A}_{j,i}^*$, we have

$$\begin{aligned}
\tilde{A}_{j,i}^* \equiv p_{j,i}\tilde{A}_j^* + \Theta_{j,i}^* &= p_{j,i} e_j' (I - P')^{-1} (A_0^* + (\Theta^*)'\mathbf{1}) + \Theta_{j,i}^* \\
&= \nu_j \left(A_0^* + \sum_{l=1}^{K} \Theta_l^* \right) + e_i' \Theta_j^*, \\
\tilde{A}_{k,i}^* \equiv p_{k,i}\tilde{A}_k^* + \Theta_{k,i}^* &= p_{k,i} e_k' (I - P')^{-1} (A_0^* + (\Theta^*)'\mathbf{1}) + \Theta_{k,i}^*, \\
&= \nu_k \left(A_0^* + \sum_{l=1}^{K} \Theta_l^* \right) + e_i' \Theta_k^*,
\end{aligned}$$

where A_0^* is the Brownian limit of the external arrival processes, i.e., $A_{0,i}^* \stackrel{d}{=} c_{a_0,i} B_{a_0,i} \circ \lambda_i e$ and $\Theta^* \equiv (\Theta_1^*, \ldots, \Theta_K^*)' \in \mathbb{R}^{K \times K}$ with $\Theta_i^* = (\Theta_{i,1}^*, \ldots, \Theta_{i,K}^*)$. Recall that Θ_i^* is the the collection of the Brownian limits of the decision processes at station i, so that

$$\mathrm{cov}(\Theta_{i,j}^*, \Theta_{i,k}^*) = \begin{cases} p_{i,j}(1-p_{i,j})\lambda_i t, & j = k, \\ -p_{i,j}p_{i,k}\lambda_i t, & j \neq k. \end{cases}$$

Define

$$\Sigma_i \equiv \left(\mathrm{cov}(\Theta_{i,j}^*, \Theta_{i,k}^*)/t \right)_{j,k=1}^K \in \mathbb{R}^{K \times K}$$

so that Σ_i is a constant matrix independent of t.

Notice that $A_{0,i}^*$, Θ_j^* for $1 \leq i, j \leq K$ are mutually independent, we have

$$\begin{aligned}
\zeta_{j,i;k,i} &\equiv \frac{1}{t} \mathrm{cov} \left(\tilde{A}_{k,i}^*(t), \tilde{A}_{j,i}^*(t) \right) \\
&= \frac{1}{t} \mathrm{cov} \left(\nu_j A_0^* + \sum_{l=1}^{K} \left(\nu_j + \delta_{l,j} e_i' \right) \Theta_l^*, \nu_k A_0^* + \sum_{l=1}^{K} \left(\nu_k + \delta_{l,k} e_i' \right) \Theta_l^* \right) \\
&= \frac{1}{t} \mathrm{cov} \left(\nu_j A_0^*, \nu_k A_0^* \right) + \frac{1}{t} \sum_{l=1}^{K} \mathrm{cov} \left(\left(\nu_j + \delta_{l,j} e_i' \right) \Theta_l^*, \left(\nu_k + \delta_{l,k} e_i' \right) \Theta_l^* \right) \\
&= \nu_j' \left(\mathrm{diag}(c_{a_0,i}^2 \lambda_i) + \sum_{l=1}^{K} \Sigma_l \right) \nu_k + \nu_k' \Sigma_j e_i + \nu_j' \Sigma_k e_i.
\end{aligned}$$

6 A Tuning Function in the Departure IDC Equation

The IDC equations discussed in §3.4 of the main paper can be generalized so that the wait functions w^* include a tuning function $h(\rho)$.

We now discuss this tuning function for departure IDC as an illustration. In particular, we replace (18) of §3.3.1 of the main paper by the following

$$w_{\rho_i}(t) \equiv w^*((1-\rho_i)^2\lambda_i t/h(\rho_i)c_{x,i}^2), \quad t \geq 0.$$

The tuning function $h(\rho)$ is an increasing continuous *tuning function* of the traffic intensity ρ with $h(0) \equiv 0$ and $h(1) \equiv 1$.

The approximation in (17) of §3.3.1 of the main paper, for any tuning function $h(\rho)$, is supported by heavy-traffic limits for the stationary departure processes, where we push the queue of interest (denoted by h) to the heavy-traffic limit while keeping other stations strictly under-saturated. Such HT limits are established in Theorems 5.1-5.3 and Corollary 6.1 of [38] for the $GI/GI/1$ model and extended to cover the OQN model in Theorem 4.1 of [41]. Under regularity conditions (uniform integrability, for which it suffices to have uniformly bounded finite fourth moments of the interarrival time and service time), the approximation in in (17) of §3.3.1 of the main paper is asymptotically correct as $\rho_h \to 1$.

It remains to specify the tuning function h used in the ρ_i-dependent weight $w_{\rho_i}(t)$ in in (18) of §3.3.1 of the main paper. It is chosen to improve the quality of approximations at queues with light-to-moderate traffic intensities. In specific, we propose

$$h(\rho) \equiv \rho^2, \quad 0 \leq \rho \leq 1. \tag{28}$$

This specific choice of the tuning function is motivated from Remark 5.2 of [38], where we replace γ by γ_ρ in the pre-limit weight function and recall that the usual case of $\mu_\rho = \lambda/\rho$ corresponds to $\gamma_\rho = 1/\rho$. The tuning function in (28) also corresponds to (33) of in [39]. We remark that h_ρ can be used as a tuning parameter to improve the quality of approximations. We will illustrate in our numerical examples in §7.

7 Additional Numerical Experiments

7.1 Comparison with RQ

This example is taken from §5.2 of [39], where we consider 10 single-server queues in series. The external arrival process is a rate-1 renewal process with H_2 interarrival times, having

$c_a^2 = 5$. The first 9 queues all have Erlang service times with $c_a^2 = 0.5$ denoted by E_2, i.e., the sum of 2 i.i.d. exponential random variables. The first 8 queues have mean service time and thus traffic intensity 0.6, while the 9^{th} queue has mean service time and thus traffic intensity 0.95. The difference in variability level of the arrival and service process introduces complex variability structure underneath the first 9 queues in series. The 10^{th} queue serves as a test queue and has an exponential service-time distribution with mean and traffic intensity ρ, which is allowed to vary from 0 to 1 in order to expose the complex impact of the variability on the performance measure of the test queue.

The RQNA algorithm in this case is a simple special case of Algorithm 1 in §3.5 of the main paper. The IDC's of the external flows (I_{a_1} for external arrival at station 1 and I_{s_i} service flows) can be derived by explicitly inverse (1), see §III.G of [19]. For internal flows, we apply the departure approximation in (17) of §3.3.1 of the main paper recursively, so that for $2 \leq i \leq 9$,

$$\begin{aligned} I_{d_1}(t) &= w_1 I_{a_1}(t) + (1 - w_1) I_{s_1}(t), \quad \text{and} \\ I_{d_i}(t) &= w_i I_{d_{i-1}}(t) + (1 - w_i) I_{s_i}(t), \end{aligned} \tag{29}$$

where we used (18) of §3.3.1 of the main paper with $h(\rho) = \rho^2$ as in (28) with $\rho_i = 0.6$ for $1 \leq i \leq 8$, $\rho_9 = 0.95$, $\lambda_i = 1$. For the variability parameters, we note that $c_{x_i}^2 \equiv c_{a_i}^2 + c_{s_i}^2 = c_{a_i}^2 + 0.5$ and that, for $2 \leq i \leq 9$,

$$c_{a_i}^2 \equiv I_{a_i}(\infty) = I_{d_{i-1}}(\infty) = I_{a_{i-1}}(\infty) = \cdots = I_{a_1}(\infty) = c_{a_1}^2 = 5.$$

With $I_{a_{10}}(t) = I_{d_9}(t)$, we can now apply the RQ algorithm in (12) of §2.2 of the main paper.

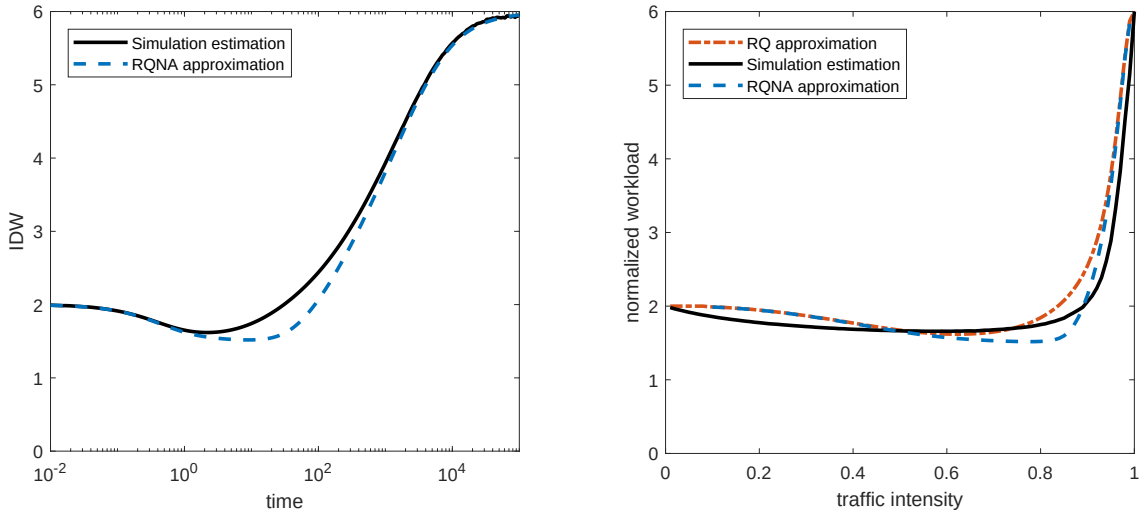
Figure 1 reports on two aspects the performance of the RQNA algorithm at the (10^{th}) test queue: (i) the approximation of the IDW, and (ii) the RQNA approximation of the steady-state mean workload. Figure 1 (left) shows that the IDC approximation in the RQNA algorithm performs very well, while Figure 1 (right) shows that both RQ (with directly estimated IDC) and RQNA are accurate.

7.2 A Single-Server Queue with i.i.d. Feedback

We start the minimal example with customer feedback, i.e., single-server queue with i.i.d. customer feedback.

In specific, we look at two settings: (1) H_2 external arrival and service distribution, both with balanced mean but $c_a^2 = 6$ and $c_s^2 = 2$, the external arrival rate is set to 1 and the feedback probability is $p = 0.5$; and (2) E_2 external arrival distribution so that $c_a^2 = 1/2$ and

Figure 1: Contrasting the RQNA approximation of the IDW at the 10-th queue and simulation estimated IDW (left) in the ten queues in series example. Simulation estimation of the steady-state mean workload, the RQ approximation in (12) of the main paper and the RQNA approximation from Algorithm 2 in §3.5 of the main paper shown in the right plot.



H_2 service distribution with balanced mean and $c_s^2 = 6$, the external arrival rate is again 1 but the feedback probability is $p = 0.75$.

To exposed the impact of the traffic intensity on the mean steady-state workload, we allow traffic in tensity to vary in the full range of $(0, 1)$.

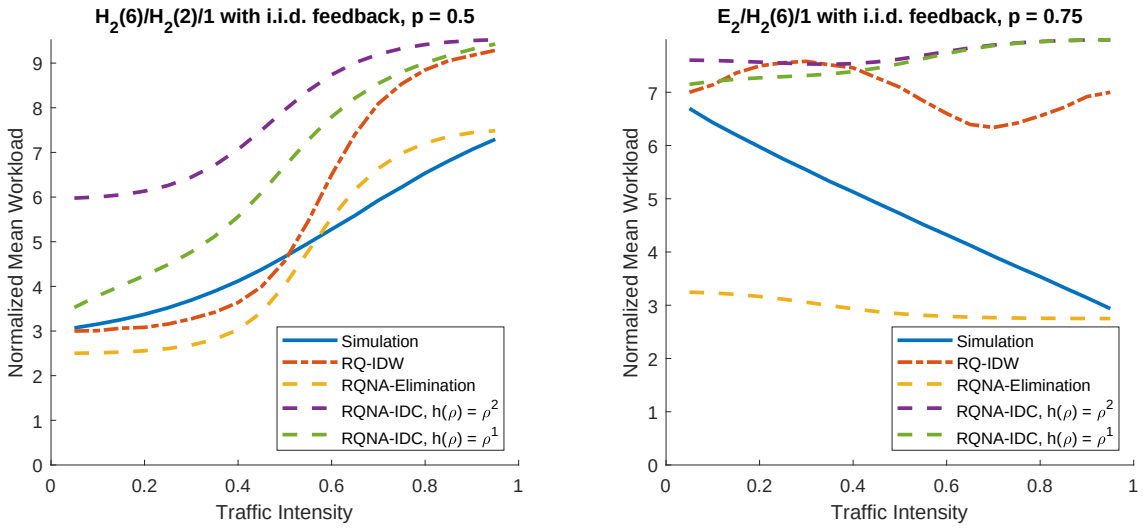
Figure 2 reports various robust queueing approximation of the two examples. We observer that feedback elimination produces exact values in the HT limit, however, it does not capture the correct LT limit. On the other hand, the RQ-IDW algorithm, as well as the RQNA-IDC algorithms with suitable tuning function gives exact LT limit, but incorrect HT limit.

7.3 Comparisons with Previous Algorithms for Queues in Series

In this section, we compare the performance of our RQNA algorithm to the performance of QNA from [36], QNET from [21], SBD from [13] and RQ from [39], for the example with 9 queues in series considered by [34]. This example was introduced by [34] to illustrate the heavy-traffic bottleneck phenomenon.

In particular, we consider an OQN with 9 stations in tandem, each with i.i.d. exponential service times. Station 1 has the only external arrival process, which is a rate-1 general renewal process. The traffic intensities at the first 8 queues are set to $\rho_i = 0.6$ for $1 \leq i \leq 8$, while

Figure 2: Contrasting the RQ algorithm in §2.2 of the main paper with simulated IDW, the RQNA algorithm with feedback elimination in §4.1 of the main paper, and the RQNA-IDC algorithm in §5 of the main paper with the simulation estimation of the mean steady-state workload, as functions of the traffic intensity ρ. For the RQNA-IDC algorithm, we display results for two different tuning functions $h(\rho)$ as specified in the legend.



the last queue has the significantly higher traffic intensity $\rho_9 = 0.9$. As in [34], two specific external renewal arrival processes are considered: (i) deterministic interarrival times with $c_{a_0}^2 = 0$; and (ii) highly variable H_2 interarrival times with $c_{a_0}^2 = 8$ (and again balanced means).

The heavy-traffic bottleneck phenomenon illustrates that the variability of the external arrival process can have only very limited impact on the performance of the following queues, especially after passing through several queues, and yet dramatically affect the performance of a later queue with a much higher traffic intensity. This phenomenon is a result of complicated long-range dependence embedded in the arrival processes, introduced by flowing through a queue (the departure processes), as revealed by the departure approximation in (17) of §3.3.1 of the main paper. This example was introduced to show the limitation of traditional decomposition methods, e.g. the QNA algorithm, and is often used as a benchmark for different approximation methods, see §3.3 of [13].

Table 1 (for low variability) and Table 2 (for high variability) compare the various approximations of the mean steady-state waiting time at each station, as well as the total waiting time in the system, to simulation estimates.

In the parentheses, we include (i) the relative half-width of the 95% confidence interval

Table 1: A comparison of four approximation methods to simulation for 9 exponential (M) queues in series fed by a deterministic arrival process with $c_a^2 = 0$.

Queue	Sim	QNA	QNET	SBD	RQ	RQNA $h(\rho) = \rho$	RQNA $h(\rho) = \rho^2$	RQNA $h(\rho) = \rho^3$
1	0.290 (2.41%)	0.45 (55%)	0.45 (55%)	0.45 (55%)	0.30 (2.3%)	0.30 (2.3%)	0.30(2.3%)	0.30 (2.3%)
2	0.491 (1.43%)	0.61 (24%)	0.66 (35%)	0.66 (35%)	0.55 (13%)	0.58 (19%)	0.53 (8.1%)	0.48 (-2.8%)
3	0.607 (1.32%)	0.72 (19%)	0.74 (22%)	0.74 (22%)	0.70 (15%)	0.72 (19%)	0.66 (9.4%)	0.60 (-1.1%)
4	0.666 (1.20%)	0.78 (17%)	0.79 (18%)	0.79 (19%)	0.77 (16%)	0.79 (19%)	0.74 (11%)	0.68 (2.1%)
5	0.706 (1.42%)	0.83 (18%)	0.82 (16%)	0.82 (16%)	0.80 (14%)	0.83 (18%)	0.79 (12%)	0.73 (3.9%)
6	0.731 (1.78%)	0.85 (16%)	0.84 (14%)	0.84 (15%)	0.83 (13%)	0.86 (18%)	0.82 (13%)	0.77 (5.7%)
7	0.748 (1.34%)	0.87 (16%)	0.85 (14%)	0.85 (14%)	0.84 (12%)	0.88 (17%)	0.85 (13%)	0.80 (7.2%)
8	0.775 (1.68%)	0.88 (14%)	0.86 (11%)	0.86 (11%)	0.85 (9.2%)	0.89 (15%)	0.86 (11%)	0.82 (6.2%)
9	5.031 (4.31%)	7.99 (59%)	6.97 (39%)	4.05 (-20%)	4.95 (-2.0%)	4.97 (-1.3%)	4.50 (-11%)	4.11 (-18%)
Total	10.05	14.0 (39%)	13.0 (29%)	10.1 (0.09%)	10.6 (5.3%)	10.8 (7.6%)	10.1 (0.13%)	9.00 (-10%)

Table 2: A comparison of four approximation methods to simulation for 9 exponential (M) queues in series fed by a highly-variable H_2 renewal arrival process with $c_a^2 = 8$.

Queue	Sim	QNA	QNET	SBD	RQ	RQNA $h(\rho) = \rho$	RQNA $h(\rho) = \rho^2$	RQNA $h(\rho) = \rho^3$
1	3.284 (3.50%)	4.05 (23%)	4.05 (23%)	4.05 (23%)	3.95 (20%)	3.95 (20%)	3.95 (20%)	3.95 (20%)
2	2.321 (4.18%)	2.92 (26%)	1.81 (22%)	1.82 (-22%)	2.61 (12%)	1.58 (-32%)	1.95 (-15%)	2.39 (3.0%)
3	1.914 (3.40%)	2.19 (14%)	1.47 (-23%)	1.49 (-22%)	2.04 (6.7%)	0.98 (-49%)	1.07 (-44%)	1.33 (-31%)
4	1.719 (4.07%)	1.73 (0.64%)	1.16 (-33%)	1.19 (-31%)	1.72 (0.31%)	0.92 (-47%)	0.94 (-41%)	0.98 (-43%)
5	1.598 (3.69%)	1.43 (-11%)	1.07 (-33%)	1.10 (-31%)	1.53 (-4.1%)	0.90 (-44%)	0.91 (-43%)	0.93 (-43%)
6	1.478 (4.13%)	1.24 (-16%)	1.03 (-31%)	1.06 (-28%)	1.41 (-4.6%)	0.90 (-39%)	0.90 (-39%)	0.91 (-39%)
7	1.423 (3.23%)	1.12 (-21%)	1.00 (-30%)	1.03 (-28%)	1.33 (-6.8%)	0.90 (-37%)	0.90 (-37%)	0.90 (-37%)
8	1.413 (4.67%)	1.04 (-26%)	0.98 (-30%)	1.01 (-29%)	1.27 (-10%)	0.90 (-36%)	0.90 (-36%)	0.90 (-36%)
9	30.12 (16.8%)	8.90 (-71%)	6.04 (-80%)	36.5 (21%)	36.9 (23%)	29.1 (-3.5%)	32.8 (9.0%)	35.3 (17%)
Total	45.27	24.6 (-46%)	18.6 (-59%)	49.8 (10%)	52.8 (17%)	40.1 (-11%)	44.4 (-2.0%)	47.6 (5.1%)

for simulation estimates (column Sim); and (ii) the relative error of the approximations compared to the simulation estimates. The first 5 columns in Table 1 and Table 2 are taken directly from Tables VIII and IX of [13], but the simulation and QNA approximations come from [34]. The last three columns are the approximations obtained from the RQNA algorithm discussed in this paper with various choice of the tuning function $h(\rho)$. The RQNA approximations of the workload are transformed into the approximations of the waiting time as in Remark 4 in §2.2 of the main paper.

To put these performance measures in perspective, note that in an $M/M/1$ queue with arrival rate 1 we would have $E[W] = \rho^2/(1-\rho)$, which would be 0.90 at the first 8 queues, but 8.1 at the last queue. For the D arrival process in Table 1, we expect that $E[W]$ will be smaller; for the the H_2 arrival process in Table 2, we expect $E[W]$ to be higher, but we see a big impact at the last queue, more than might be expected.

We make the following observations from this experiment:

1. The new RQNA algorithm does better than the QNA and QNET methods on total time spent waiting in queue, and is comparable with the SBD method, even though RQNA does not require solving an RBM.

2. The RQNA algorithm does exceptionally well at the final bottleneck queue and is competitive with all other methods for approximating the mean waiting time. The new RQNA method is based on heavy-traffic limits just as the previous methods methods, but focuses on the flows, and exploits RQ instead of analyzing an RBM.

3. The RQNA algorithm can benefit from further improvement for light-to-medium traffic intensities. As demonstrated in Table 2, the mean waiting times at queues 3-8 are pushed too much towards the $M/M/1$ values in the departure IDC approximation for light to medium traffic intensity. That remains to be a direction for future research.

References

[1] Abate J, Whitt W (1992) The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems* 10:5–88.

[2] Ahmed MA, Alkhamis TM (2009) Simulation optimization for an emergency department healthcare unit in Kuwait. *European Journal of Operational Research* 198(3):936–942.

[3] Albin SL, Barrett J, Ito D, Mueller JE (1990) A queueing network analysis of a health center. *Queueing Systems* 7(1):51–61.

[4] Asmussen S (2003) *Applied Probability and Queues* (New York: Springer), second edition.

[5] Azaron A, Ghomi SMTF (2003) Optimal control of service rates and arrivals in Jackson networks. *European Journal of Operational Research* 147(1):17–31.

[6] Bai X, Gopal R, Nunez M, Zhdanov D (2014) A decision methodology for managing operational efficiency and information disclosure risk in healthcare processes. *Decision Support Systems* 57:406–416.

[7] Banerjee S, Johari R, Riquelme C (2015) Pricing in ride-sharing platforms: A queueing-theoretic approach. *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, 639 (ACM).

[8] Bi J, Zhu Z, Tian R, Wang Q (2010) Dynamic provisioning modeling for virtualized multi-tier applications in cloud data center. *2010 IEEE 3rd International Conference on Cloud Computing*, 370–377 (IEEE).

[9] Brahimi M, Worthington D (1991) Queueing models for out-patient appointment systems-A case study. *Journal of the Operational Research Society* 42(9):733–746.

[10] Chen H, Yao DD (2001) *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization* (New York: Springer).

[11] Cox DR, Lewis PAW (1966) *The Statistical Analysis of Series of Events* (London: Methuen).

[12] Creemers S, Lambrecht M (2011) Modeling a hospital queueing network. *Queueing Networks*, chapter 18, 767–798 (Springer).

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

[14] Daley DJ (1976) Queueing output processes. *Adv. Appl. Prob.* 8(2):395–415.

[15] Daley DJ, Vere-Jones D (2008) *An Introduction to the Theory of Point Processes* (Oxford, U. K.: Springer), second edition.

[16] Dieker AB, Ghosh S, Squillante MS (2016) Optimal resource capacity management for stochastic networks. *Operations Research* 65(1):221–241.

[17] Disney RL, Konig D (1985) Queueing networks: a survey of their random processes. *SIAM Review* 27(3):335–403.

[18] Dittus RS, Klein RW, DeBrota DJ, Dame MA, Fitzgerald JF (1996) Medical resident work schedules: Design and evaluation by stimulation modeling. *Management Science* 42(6):891–906.

[19] Fendick KW, Whitt W (1989) Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue. *Proceedings of the IEEE* 71(1):171–194.

[20] Gupta D (2013) Queueing models for healthcare operations. *Handbook of Healthcare Operations Management*, 19–44 (Springer).

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

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

[23] Jiang L, Giachetti RE (2008) A queueing network model to analyze the impact of parallelization of care on patient cycle time. *Health Care Management Science* 11(3):248–261.

[24] Kelly PF (2011) *Reversibility and Stochastic Networks* (Cambridge University Press), revised edition.

[25] Kim S, Whitt W (2014) Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing and Service Oper. Management* 16(3):464–480.

[26] Lu Y, Abdelzaher T, Lu C, Sha L, Liu X (2003) Feedback control with queueing-theoretic prediction for relative delay guarantees in web servers. *The 9th IEEE Real-Time and Embedded Technology and Applications Symposium, 2003. Proceedings.*, 208–217 (IEEE).

[27] Mandelbaum A, Reiman MI (1998) On pooling in queueing networks. *Management Science* 44(7):971–981.

[28] Neuts MF (1979) A versatile Markovian point process. *Journal of Applied Probability* 16(4):764–779.

[29] Neuts MF (1989) *Structured Stochastic Matrices of $M/G/1$ Type and their Application* (New York: Marcel Dekker).

[30] Ross KW, Yao D (1989) Optimal dynamic scheduling in Jackson networks. *IEEE Transactions on Automatic Control* 34(1):47–53.

[31] Ross SM (1996) *Stochastic Processes* (New York: Wiley), second edition.

[32] Schuijbroek J, Hampshire RC, Hoeve WJV (2017) Inventory rebalancing and vehicle routing in bike sharing systems. *European Journal of Operational Research* 257(3):992–1004.

[33] Serfozo R (2012) *Introduction to Stochastic Networks*, volume 44 (Springer Science & Business Media).

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

[35] Tesauro G, Das R, Walsh WE, Kephart JO (2005) Utility-function-driven resource allocation in autonomic systems. *Second International Conference on Autonomic Computing (ICAC'05)*, 342–343 (IEEE).

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

[37] Whitt W (2002) *Stochastic-Process Limits* (New York: Springer).

[38] 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.

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

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

[41] Whitt W, You W (2020) Heavy-traffic limits for stationary network flows. *Queueing Systems* 1–16.

[42] Xiong K, Perros HG (2009) Service performance and analysis in cloud computing. *2009 Congress on Services - I*, 693–700 (IEEE).