Approximating Steady-State Performance Measures in Open Queueing Network:

An Algorithm Based on Indices of Dispersion

Wei You[1]

Columbia University

iDDA, CUHK-SZ
January 22, 2019

[1] Joint work with Ward Whitt.

Motivation

Many service systems can be modeled as open queueing networks (OQNs),

- e.g. call centers, healthcare systems, cloud computing networks and ride-sharing platforms.

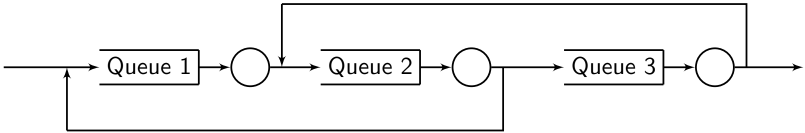


Figure: A three-station example with feedback from Dai, Nguyen and Reiman (1994)

Motivation

Performance measures

- Queue length, customer waiting time, system workload, etc.
- Important for the analysis and design of real-world systems;
- Closed-form solutions are hardly available for realistic models;

$\Rightarrow$ **resort to approximation methods.**

Background - Existing Approximation Algorithms

Decomposition approximation

- Motivated by product-form solutions of Jackson Networks.
- Treat stations as independent single-server queues.

Examples

- The Queueing Network Analyzer (QNA) by Whitt (1983),
  - approximates each station by a GI/GI/1 queue.
- Markovian Arrival Process (MAP)
  - Horváth et al. (2010), MAP/MAP/1.
  - Kim (2011a, 2011b), MMPP(2)/GI/1.

Background - Previous Approximation Algorithms

Diffusion Approximations

- Heavy-traffic limits with Reflected Brownian Motion (RBM).
  - Iglehart and Whitt (1970), Harrison (1973,1978) and Reiman (1984);
- Approximate the steady-state queue length distribution by the stationary distribution of the limiting RBM;
  - Gamarnik and Zeevi (2006), Budhiraja and Lee (2009) and Braverman, Dai and Miyazawa (2017).
- numerically calculate the steady-state mean of the RBM.

Examples

- QNET by Harrison and Nguyen (1990) for OQNs and by Dai and Harrison (1993) for CQNs;
- Sequential bottleneck decomposition (SBD) by Dai, Nguyen and Reiman (1994).

Background - Recent Developments

Recent Developments

- The first (Parametric) Robust Queueing (RQ) by Bandi et al. (2015), designed for waiting time.

All above can be classified as parametric methods.

- use a set of parameters, usually first few moments, to characterize the underlying stochastic processes.

Overview

We developed a non-parametric approximation algorithm called Robust Queueing Network Analyzer, RQNA for short.

- Designed for continuous-time workload process[2][3].
- **Main idea**: Robust optimization + Queueing theory, hence the name Robust Queueing (RQ).
  - RQ was first proposed in Bandi et al. (2015).
  - Replace probability laws by uncertainty sets, and analyze the worst case scenario.

---

[2] Use Brumelle's formula to obtain waiting time approximation.

[3] Use Little's Law to obtain queue length approximation.

Overview

- **Key component**: Index of Dispersion for Counts (IDC)

$$I_a(t) \equiv Var(A(t))/E[A(t)], \quad t \geq 0,$$

where $A(t)$ is a stationary counting process.

- Non-parametric: variability of a process is captured by continuous functions, i.e., IDCs.

- Braverman and Dai (2018), high order diffusion approximation for Erlang-C.

- **Supporting theories**: Heavy-traffic limit theorems for stationary flows and their IDCs.

Dependence in Queues

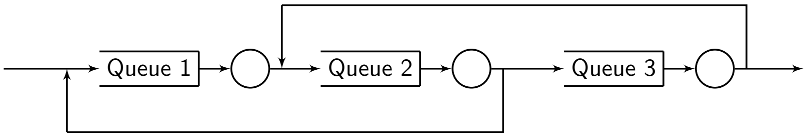


Figure: A three-station example.

Dependence rises naturally in queueing network:

- Dependence within/between the flows[4]:
  - introduced by departure, splitting, superposition and customer feedback.

[4] arrival processes, departure process, etc.

Dependence in Queues

Dependence has significant impact on performance measures

- Dependence can have complicated temporal structure.
- The **level of impact** will depend on both the temporal structure and the traffic intensity.
- Indices of dispersion can describe the temporal structure.

Indices of Dispersion for Counts (IDC)

Definition from Cox and Lewis (1966)

$$I_a(t) \equiv Var(A(t))/E[A(t)], \quad t \geq 0,$$

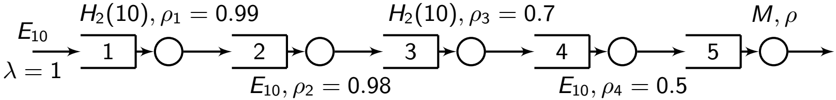
where $A(t)$ is any stationary point process.

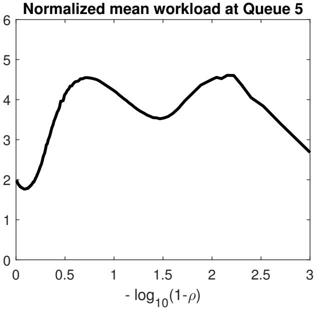
Theorem (renewal process characterization theorem)

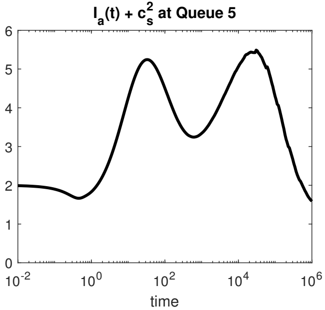
A renewal process $A(t)$ with positive rate λ is fully characterized by the IDC of its equilibrium (stationary) version $A_e(t)$.

- For $GI/GI/1$ model, the performance measure must be some function of the rates and IDCs of the arrival and service processes;
- RQNA using IDC can potentially generate more accurate and adaptive approximations.

A Five Queues in Series Example







Parametric methods (QNA, RQ by Bandi et al.) using first few moments to describe variability may fail.

Continuous-time workload process

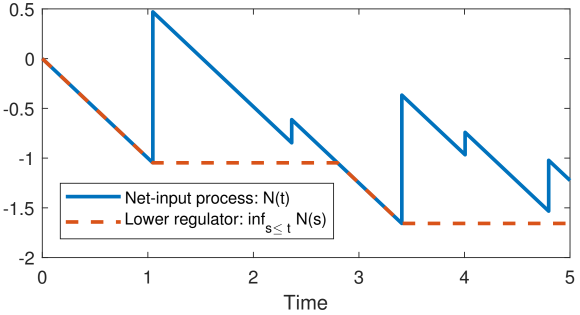
- $\{(U_i, V_i)\}$: interarrival times and service times;
- λ, μ: arrival rate and service rate;
- $A(t)$: arrival counting process associated with $\{U_k\}$;
- $Y(t)$: total input of work

$$Y(t) \equiv \sum_{k=1}^{A(t)} V_k;$$

- $N(t)$: net-input process

$$N(t) \equiv Y(t) - t.$$

Continuous-time workload process



The steady-state workload

$$Z \equiv N(0) - \inf_{-\infty \leq t \leq 0} \{N(t)\}.$$

$$= \sup_{0 \leq s \leq \infty} \{N(0) - N(-s)\} \equiv \sup_{0 \leq s \leq \infty} \{N_0(s)\}$$

- $N_0(s)$: the net-input over time $[-s, 0]$.
- With an abuse of notation, we omit the subscript in $N_0(s)$.

Stochastic versus Robust Queues

Defined in sample path sense

$$Z = \sup_{0 \le s \le \infty} \{N(s)\}.$$

- no requirement on the primitives.

Stochastic Queue

- $N(s) \equiv \sum_{k=1}^{A(s)} V_k - s$ is a stochastic process.
- Workload is a random variable.

Robust Queue

- $\tilde{N}$ is a sample path from a uncertainty set $\mathcal{U}$.
- Workload defined as the deterministic worse-case scenario

$$Z^* \equiv \sup_{\tilde{N} \in \mathcal{U}} \sup_{0 \le s \le \infty} \{\tilde{N}(s)\}.$$

Robust Queueing for continuous-time workload

Our uncertainty set is motivated from CLT

$$\mathcal{U}_b \equiv \left\{ \tilde{N} : \tilde{N}(s) \leq E[N(s)] + b\sqrt{\mathrm{Var}(N(s))},\ s \geq 0 \right\},$$

where $N(t) = \sum_{i=1}^{A(t)} V_i - t$ is the net input process associated with the stochastic queue.

- Parameter b controls the robustness.

Assume

- Arrival process is a stationary point process.
- Service times are i.i.d., independent of the arrival process.

$$E[N(t)] = \rho t - t,$$
$$\mathrm{Var}(Y(t)) = \rho t (I_a(t) + c_s^2)/\mu.$$

Robust Queueing for continuous-time workload

RQ for workload

$$Z^*(b) = \sup_{N \in \mathcal{U}_b} \sup_{0 \leq s \leq \infty} \{N(s)\},$$

where

$$\mathcal{U}_b = \left\{ \tilde{N} : \tilde{N}(s) \leq -(1-\rho)s + b\sqrt{\rho s (I_a(s) + c_s^2)/\mu},\ s \geq 0 \right\}.$$

Lemma (Dimension reduction)

The infinite-dimensional RQ problem can be reduced to

$$Z^*(b) = \sup_{0 \leq s \leq \infty} \sup_{N \in \mathcal{U}_b} \{N(s)\}$$

$$= \sup_{0 \leq s \leq \infty} \left\{ -(1-\rho)s + b\sqrt{\rho s (I_a(s) + c_s^2)/\mu} \right\}.$$

Robust Queueing for continuous-time workload

In summary, the RQ algorithm for single-server queues

$$Z^*(b) = \sup_{0 \le s \le \infty} \left\{ -(1-\rho)s + b\sqrt{\rho s (I_a(s) + c_s^2)/\mu} \right\}.$$

How to connect $Z^*(b)$ to the distribution of the steady-state workload Z?

- We propose the approximation

$$Z(p) \equiv Z(\Pi(b)) \approx Z^*(b),$$

  - $Z(p)$ denotes the p^{th} quantile of Z
  - Π: one-to-one continuous function, map b into quantile level p.

Robust Queueing for continuous-time workload

Which function Π should we use?

- For $M/M/1$ view

$$P(Z \leq z) = 1 - \rho e^{-\rho z/m}, \text{ for } m = \rho/\lambda(1-\rho)$$

Hence the p^{th} quantile is

$$Z(p) = -(m/\rho)\ln((1-p)/\rho). \qquad (*)$$

- On the other hand, for $M/M/1$ model, RQ gives

$$Z^*(b) = \frac{b^2}{2}m, \text{ for } m = \rho/\lambda(1-\rho). \qquad (**)$$

- Equating (*) to (**), we have the approximation

$$\Pi(b) \approx 1 - \rho e^{-\rho b^2/2}.$$

- **[Approximation for the mean]** From (**), we see that $b = \sqrt{2}$ corresponds to the mean.

Robust Queueing for continuous-time workload

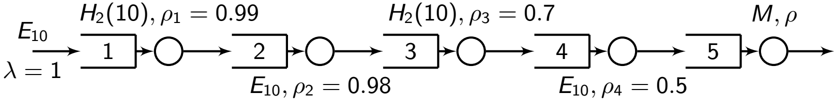
The RQ algorithm for mean steady-state workload

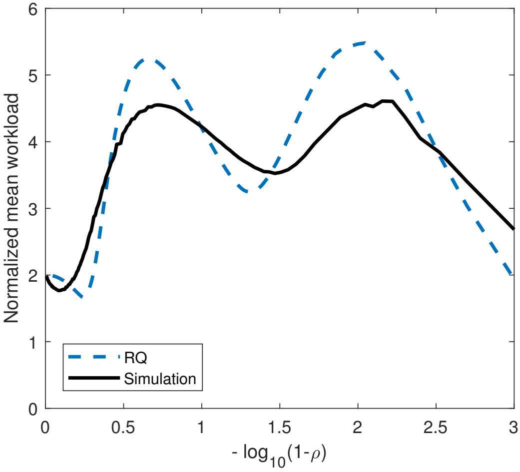
$$Z^* = \sup_{0 \le s \le \infty} \left\{ -(1-\rho)s + \sqrt{2\rho s (I_a(s) + c_s^2)/\mu} \right\}.$$

- Takes the arrival IDC $I_a(t)$ as a model input.

Theorem (RQ exact in heavy-traffic and light-traffic limits)

Under regularity assumptions, the RQ algorithm yields the exact mean steady-state workload in both light-traffic and heavy-traffic limits for $G/GI/1$ models.


E_{10}
$\lambda = 1$
$H_2(10), \rho_1 = 0.99$
1
2
$E_{10}, \rho_2 = 0.98$
$H_2(10), \rho_3 = 0.7$
3
4
$E_{10}, \rho_4 = 0.5$
M, ρ
5


Normalized mean workload
- log10(1-ρ)
RQ
Simulation
0
1
2
3
4
5
6
0
0.5
1
1.5
2
2.5
3

The Heavy-traffic Bottleneck Phenomenon

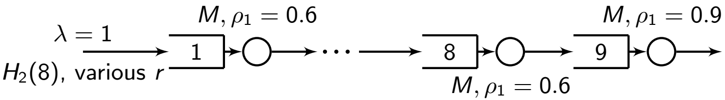


Table: Mean steady-state waiting time at each station.

r	0.5		N/A	N/A	N/A	0.9		0.1	
Queue	Sim	RQ	QNA	QNET	SBD	Sim	RQ	Sim	RQ
1	3.28	3.95	4.05	4.05	4.05	1.16	1.13	5.69	5.83
2	2.32	2.61	2.92	1.81	1.82	1.16	1.12	2.46	2.40
3	1.91	2.04	2.19	1.47	1.49	1.15	1.11	1.98	1.83
4	1.71	1.72	1.73	1.16	1.19	1.14	1.10	1.76	1.56
5	1.59	1.53	1.43	1.07	1.10	1.14	1.10	1.63	1.41
6	1.47	1.41	1.24	1.03	1.06	1.13	1.09	1.54	1.31
7	1.42	1.33	1.12	1.00	1.03	1.13	1.08	1.48	1.24
8	1.41	1.27	1.04	0.98	1.01	1.12	1.08	1.42	1.20
9	30.1	36.9	8.9	6.0	36.4	19.6	36.5	29.6	36.3
Total	45.3	52.8	24.6	18.6	49.8	28.8	45.3	47.5	53.1
Avg. abs. RE		9.7%	23%	33%	26%		13%		12%

Generalization to Queue in Series (Tandem Queues)

To generalize RQ from single-server queues to queues in series, we need the IDC of the departure process.

Literature Review - Departure Processes

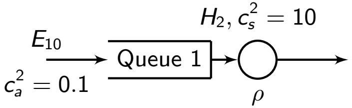
Exact characterizations

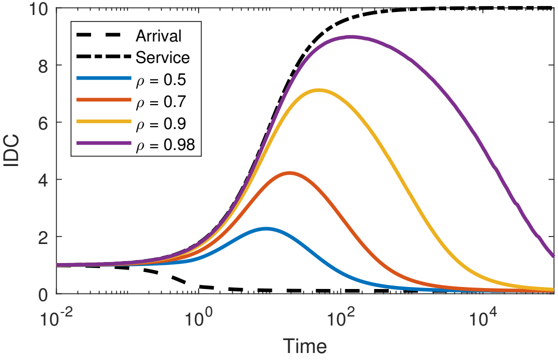
- Burke (1956): M/M/1 departure is Poisson;
- Takács (1962): the Laplace transform (LT) of the mean of the departure process under Palm distribution;
- Daley (1976): the LT of the variance function of the stationary departure from M/G/1 and GI/M/1 models;
- Green's dissertation (1999) and Zhang (2005): BMAP/MAP/1 departure is a MAP with infinite order
  - MAP with infinite order is intractable in practice, one need to resort to truncation.

Heavy-traffic limits

- Iglehart and Whitt (1970), HT limits for departure process in systems that starts empty;
- Gamarnik and Zeevi (2006) and Budhiraja and Lee (2009), HT limit for stationary queueing length process.

A numerical example





Heavy-Traffic Limit for the Departure Processes

Let $D^*_\rho(t) \equiv (1-\rho)[D_\rho((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda t]$.

Theorem (HT limit for the stationary departure process)

For GI/GI/1 queue under regularity conditions, the HT-scaled stationary departure process $D^*_\rho(t)$ converges to

$$D^*(t) = c_a B_a(\lambda t) + Q^*(0) - Q^*(t).$$

- B_a *and* B_s *are independent standard Brownian motions;*
- $Q^*(t) = \psi(Q^*(0) + c_a B_a \circ \lambda e - c_s B_s \circ \lambda e - \lambda e)$ *is the HT limit for stationary queue length process: a stationary reflective Brownian motion (RBM)* R_e *with drift* $-\lambda$, *variance* $\lambda c_x^2 \equiv \lambda c_a^2 + \lambda c_s^2$;
- $Q^*(0) \sim \exp(2/c_x^2)$ *is the exponential marginal distribution;*
- B_a, B_s *and* $Q^*(0)$ *are mutually independent.*

Heavy-Traffic Limit for the Variance Functions

Define the HT-scaled variance function of the stationary departure process

$$V^*_{d,\rho}(t) \equiv Var(D^*_\rho(t)).$$

Theorem (HT limit for the GI/GI/1 departure variance)

Under uniform integrability conditions, $V^*_{d,\rho}(t)$ converges to

$$V^*_d(t) \equiv w^*\left(\lambda t/c_x^2\right) c_a^2 \lambda t + \left(1 - w^*\left(\lambda t/c_x^2\right)\right) c_s^2 \lambda t, \quad as\ \rho \uparrow 1$$

where $c_x^2 = c_a^2 + c_s^2$,

$$w^*(t) = \frac{1}{2t}\left(\left(t^2 + 2t - 1\right)\left(2\Phi(\sqrt{t}) - 1\right) + 2\sqrt{t}\phi(\sqrt{t})(1+t) - t^2\right)$$

and ϕ, Φ *are the standard normal pdf and cdf, respectively.*

The Covariance Between BM and Stationary RBM

Corollary

Suppose $B = (B_1, B_2)$ *is a 2-d Brownian motion with zero drift and covariance matrix* $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_2^2 \end{pmatrix}$. *Let*

$$Q = \psi(B_1 + Q(0) - \lambda e)$$

be the stationary RBM associated with the drifted BM $B_1 - \lambda e$ *and* $Q(0)$ *has the stationary distribution of* Q, *which is independent of* B_1. *Then*

$$\mathrm{cov}(B_2, Q) = \left(1 - w^*(\lambda^2 t/\sigma_1^2)\right) \sigma_{1,2} t.$$

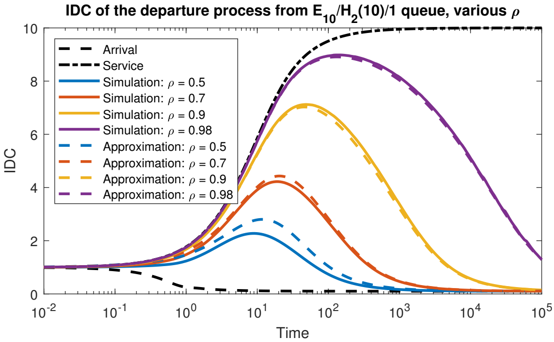
Approximation for Departure IDC

The HT theorem for variance supports the following approximation

$$I_d(t) \approx w_\rho(t) I_a(t) + (1 - w_\rho(t)) I_s(\rho t), \qquad \text{(Dep)}$$

where

$$w_\rho(t) = w^*((1-\rho)^2 \lambda t / (\rho c_x^2)),$$



Generalization to RQNA

The total arrival process at any queue:

- **superposition** of external arrival and **splitting** of **departure** processes.

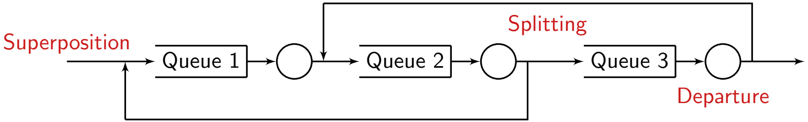


Figure: A three-station example.

Recall the departure IDC equation

$$I_{d,i}(t) = w_i(t) I_{a,i}(t) + (1 - w_i(t)) I_{s,i}(\rho t), \qquad \text{(Dep)}$$

Independent Splitting

In the case of independent splitting,

- Let $\theta^l_{i,j} = 1$ if the l-th departure from Station i is routed to Station j and 0 otherwise;
- Assume Markovian routing, so $\{\theta^l_{i,j}, l = 0, 1, \ldots\}$ are i.i.d. Bournoulli r.v. with probability $p_{i,j}$;
- Assume that D_i is independent of $\{\theta^l_{i,j}, l = 0, 1, \ldots\}$.

Independent Splitting

The customer stream $A_{i,j}(t)$ from Station i to Station j is

$$A_{i,j}(t) = \sum_{l=1}^{D_i(t)} \theta_{i,j}^l.$$

By conditional variance formula,

$$V_{a,i,j}(t) = p_{i,j}^2 V_{d,i}(t) + p_{i,j}(1 - p_{i,j})\lambda_i t,$$

or, equivalently, since $E[A_{i,j}(t)] = p_{i,j}\lambda_i t$,

$$I_{a,i,j}(t) = p_{i,j} I_{d,i}(t) + (1 - p_{i,j}).$$

Independent Splitting

$$I_{a,i,j}(t) = p_{i,j} I_{d,i}(t) + (1 - p_{i,j}). \qquad \text{(Spl')}$$

For Markovian routing, (Spl') is exact if there is no customer feedback at this station i.

However, in the presence of customer feedback, the departure process and the splitting decision are necessarily correlated.

Dependent Splitting

For the splitting with dependence, define the correction term as

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

so that

$$I_{a,i,j}(t) = p_{i,j} I_{d,i}(t) + (1 - p_{i,j}) + \alpha_{i,j}(t).$$

- In general, it is impossible to obtain exact formula for $\alpha_{i,j}(t)$.
- To approximate, we explore the joint HT limit for D_i and the splitting decision process, where only Station i is brought to heavy-traffic.

HT Limit for Splitting

Let $\theta_i^l = (\theta_{i,1}^l, \theta_{i,2}^l, \ldots, \theta_{i,K}^l)$ and define the vector of splitting decisions up to the n-th decision at station i

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

- Consider a series of system with $\rho = \rho_i \uparrow 1$ and $\rho_j < 1$ for $j \neq i$;
- Consider the usual diffusion scaling.

$$\begin{aligned}
D_{i,\rho}^*(t) &= (1-\rho)\left[D_i((1-\rho)^{-2}t) - \lambda_i(1-\rho)^{-2}t\right], \\
\Theta_{i,\rho}^*(t) &= (1-\rho)\left[\sum_{l=1}^{\lfloor (1-\rho)^{-2}t \rfloor} \theta^l - \mathbf{p}_i(1-\rho)^{-2}t\right], \\
A_{i,j,\rho}^*(t) &= (1-\rho)\left[A_{i,j}((1-\rho)^{-2}t) - \lambda_i p_{i,j}(1-\rho)^{-2}t\right], \\
Q_{i,\rho}^* &= (1-\rho)Q_i((1-\rho)^{-2}t), \\
&\ldots
\end{aligned}$$

The Correction Term α

$$A^*_{i,j,\rho} \Rightarrow A^*_{i,j} \equiv p_{i,j} D^*_i + \Theta^*_{i,j} \circ \lambda_i e, \text{ as } \rho_i \uparrow 1,$$

where

$$\begin{aligned} D^*_i &= \tilde{A}^*_i + \tilde{Q}^*_i(0) - \tilde{Q}^*_i, \\ \tilde{A}^*_i &= e_i^T (I - P^T)^{-1} \left(A^*_0 + (\Theta^*)^T \mathbf{1} \right), \\ \tilde{Q}^*_i &= \psi \left(\tilde{Q}^*_i(0) + \tilde{A}^*_i - S^*_i - \lambda_i e \right) \end{aligned}$$

and ψ is the one-dimensional reflection map.
Model primitives

- A^*_0: BM, external arrival flow;
- S^*_i: BM, service flow at station i;
- Θ^*: BM, splitting decision process.

HT Limit for Splitting

Recall that

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

Define

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

Define the limiting correction term as

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

Corollary

Under regularity conditions, we have

$$\alpha^*_{i,j,\rho}(t) \Rightarrow \alpha^*_{i,j}(t), \ \text{as } \rho \uparrow 1.$$

The Correction Term α

Recall that we obtained explicit formula for the covariance between a BM and a RBM. As a result,

$$\alpha_{i,j,\rho_i}(t) \approx 2\xi_{i,j} p_{i,j}(1 - p_{i,j}) w^*((1 - \rho_i)^{-2} \lambda_i t / (\rho_i c_{x,i}^2)),$$

$\xi_{i,j}$ is the $(i,j)^{th}$ entry of the matrix $(I - P^T)^{-1}$.

$$I_{a,i,j}(t) = p_{i,j} I_{d,i}(t) + (1 - p_{i,j}) + \alpha_{i,j}(t). \qquad \text{(Spl)}$$

HT Limit for Superposition

For dependent streams, the variance of the superposition total arrival process at queue i can be written as

$$V_{a,i}(t) \equiv \mathrm{Var}\left(\sum_{j=0}^{K} A_{j,i}(t)\right) = \sum_{j=0}^{K} \mathrm{Var}\left(A_{j,i}(t)\right) + \beta_i(t) E[A_i(t)]$$

where $A_{0,i}$ denotes the external arrival process at station i,

$$\beta_i(t) \equiv \sum_{j \neq k} \beta_{j,i;k,i}(t), \quad \text{and} \quad \beta_{j,i;k,i}(t) \equiv \frac{\mathrm{cov}\left(A_{j,i}(t), A_{k,i}(t)\right)}{E[A_i(t)]}.$$

In terms of the IDC's, we have

$$I_{a,i}(t) = \sum_{j=0}^{K} (\lambda_{j,i}/\lambda_i) I_{a_{j,i}}(t) + \beta_i(t).$$

The Correction Term β

Similar to the splitting correction term α, we explore the HT limit, where only station i is brought to heavy-traffic.

$$\beta_i(t) \equiv \sum_{j \neq k} \beta_{j,i;k,i}(t), \quad \text{and}$$

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

for some constant $\zeta_{j,i;k,i}$.

The IDC Equations

In summary, the IDC equations are

$$I_{d,i}(t) = w_i(t) I_{a,i}(t) + (1 - w_i(t)) I_{s,i}(\rho t), \tag{Dep}$$

$$I_{a,i,j}(t) = p_{i,j} I_{d,i}(t) + (1 - p_{i,j}) + \alpha_{i,j}(t), \tag{Spl}$$

$$I_{a,i}(t) = \sum_{j=0}^{K} (\lambda_{j,i}/\lambda_i) I_{a,j,i}(t) + \beta_i(t). \tag{Sup}$$

- A system of linear equations for each fixed t;
- The IDC equations have a unique solution if every customer eventually leave the system.

3 Stations with Feedback

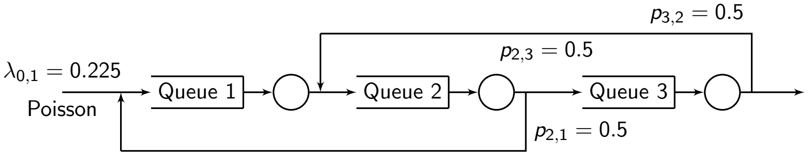


Figure: A three-station example.

Table: Traffic intensity.

Case	ρ_1	ρ_2	ρ_3
1	0.675	0.900	0.450
2	0.900	0.675	0.900
3	0.900	0.675	0.450
4	0.900	0.675	0.675

Table: Variability (squared coefficient of variation, scv) of service-time distributions.

Case	$c^2_{s,1}$	$c^2_{s,2}$	$c^2_{s,3}$
A	0.00	0.00	0.00
B	2.25	0.00	0.25
C	0.25	0.25	2.25
D	0.00	2.25	2.25
E	8.00	8.00	0.25

3 Stations with Feedback

Table: A comparison of four approximation methods to simulation for the total sojourn time in the three-station example.

Case		Simu	QNA	QNET	SBD	RQNA
A	1	40.39	20.5 (-49%)	diverging	43.0 (6.4%)	44.8 (11.0%)
	2	59.58	36.0 (-40%)	56.7 (-4.9%)	58.2 (-2.4%)	69.3 (16.4%)
	3	40.72	24.0 (-41%)	38.7 (-5.0%)	40.2 (-1.3%)	43.3 (6.3%)
	4	42.12	26.2 (-38%)	41.8 (-0.7%)	42.7 (1.3%)	41.2 (-2.2%)
B	1	52.40	42.0 (-20%)	52.6 (0.4%)	50.2 (-4.2%)	53.1 (1.4%)
	2	91.52	94.1 (2.8%)	83.7 (-8.5%)	95.3 (4.1%)	94.5 (3.2%)
	3	61.68	72.2 (17%)	61.9 (0.4%)	60.9 (-1.3%)	60.5 (-1.9%)
	4	63.34	75.8 (20%)	64.1 (1.3%)	64.7 (2.1%)	62.4 (-1.4%)
C	1	44.24	31.3 (-29%)	37.0 (-16%)	47.1 (6.4%)	42.1 (-4.8%)
	2	92.42	87.4 (-5.4%)	91.2 (-1.4%)	91.6 (-0.8%)	96.0 (3.8%)
	3	44.26	33.2 (-25%)	44.0 (-0.7%)	45.0 (1.7%)	44.0 (-0.6%)
	4	50.20	41.4 (-18%)	51.1 (1.7%)	52.2 (4.0%)	45.9 (-8.6%)
E	1	134.4	265 (97%)	155 (15%)	116 (-14%)	120 (-11%)
	2	213.1	308 (45%)	228 (7.1%)	206 (-3.3%)	173 (-19%)
	3	138.7	244 (76%)	161 (16%)	135 (-2.5%)	136 (-2.0%)
	4	155.1	252 (63%)	168 (8.2%)	147 (-5.0%)	148 (-4.8%)

3 Stations with Feedback

Table: A close look at **Case D**: $(c_{s_1}^2, c_{s_2}^2.c_{s_3}^2) = (0, 2.25, 2.25)$.

Case-Q	Simu	QNA	QNET	SBD	RQNA
D1-1	2.476	2.24 (-9.4%)	2.48 (0.3%)	2.47 (-0.1%)	2.68 (7.8%)
D1-2	10.85	14.9 (37%)	11.6 (6.5%)	11.4 (5.2%)	11.1 (2.7%)
D1-3	2.544	2.53 (-0.8%)	2.54 (-0.0%)	2.59 (1.6%)	2.53 (-0.7%)
D1-sum	55.81	71.4 (28%)	58.8 (5.3%)	58.2 (4.3%)	57.6 (3.3%)
D2-1	11.35	8.01 (-29%)	10.8 (-4.5%)	11.1 (-1.9%)	11.3 (0.1%)
D2-2	2.643	2.96 (12%)	2.75 (4.0%)	2.82 (6.7%)	3.06 (16%)
D2-3	26.87	32.9 (22%)	26.8 (-0.4%)	24.9 (-7.5%)	31.1 (16%)
D2-sum	98.36	102 (3.4%)	97.2 (-1.2%)	94.4 (-4.0%)	105 (7.1%)
D3-1	11.39	7.95 (-30%)	11.0 (-3.5%)	11.3 (-0.5%)	11.3 (-0.5%)
D3-2	2.290	2.90 (27%)	2.53 (10%)	2.26 (-1.4%)	2.10 (-8.2%)
D3-3	2.220	2.40 (7.9%)	2.38 (7.0%)	2.59 (16%)	2.43 (9.6%)
D3-sum	47.72	40.2 (-16%)	47.8 (0.2%)	48.2 (1.0%)	47.5 (0.51%)
D4-1	11.30	7.97 (-29%)	10.9 (-3.2%)	11.3 (0.3%)	11.3 (0.3%)
D4-2	2.414	2.93 (21%)	2.64 (9.5%)	2.60 (7.7%)	2.10 (-13%)
D4-3	5.886	6.83 (16%)	6.31 (7.3%)	6.17 (4.8%)	5.95 (1.1%)
D4-sum	55.24	49.3 (-11%)	56.0 (1.4%)	56.7 (2.7%)	54.3 (-1.7%)
average RE		20.24%	4.72%	4.52%	5.51%

3 Stations with Feedback

- Case E3:

$$(\rho_1, \rho_2.\rho_3) = (0.9, 0.675, 0.45)$$
$$(c^2_{s_1}, c^2_{s_2}.c^2_{s_3}) = (8, 8, 0.25)$$

Table: A comparison of six approximation methods to simulation for the sojourn time at each station of the three-station example.

Case E3, r = 0.5				
Queue	Simu	QNET	SBD	RQNA
1	31.22	35.9 (15%)	26.0 (-17%)	26.0 (-17%)
2	8.32	10.2 (23%)	11.1 (33%)	11.8 (42%)
3	2.00	1.89 (5.5%)	1.94 (3%)	0.93 (-54%)
Sum	138.7	161.3 (16%)	135.3 (-2.5%)	136.1 (-1.9%)

Case E3, r = 0.99				
Queue	Simu	QNET	SBD	RQNA
1	27.67	35.9 (30%)	26.0 (-6.0%)	26.0 (-6.0%)
2	2.67	10.2 (282%)	11.1 (316%)	6.03 (125%)
3	0.56	1.89 (236%)	1.94 (245%)	0.50 (-11%)
Sum	103.8	161.3 (55%)	135.3 (30%)	112.1 (8%)

10 Stations with Feedback

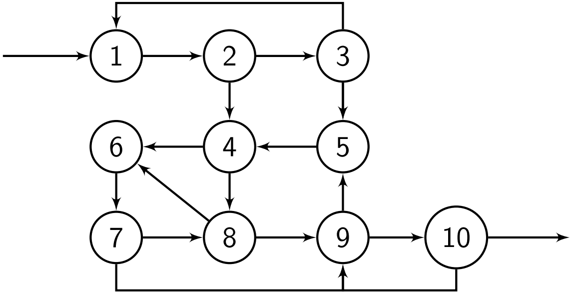


Figure: A ten-station with customer feedback example.

- The traffic intensity vector is $(0.6, 0.4, 0.6, 0.9, 0.9, 0.6, 0.4, 0.6, 0.6, 0.4)$.
- The scv's at these stations are $(0.5, 2, 2, 0.25, 0.25, 2, 1, 2, 0.5, 0.5)$

10 Stations with Feedback

Table: A comparison of five approximation methods to simulation for the mean steady-state sojourn times at each station.

Q	Simu	QNA	QNET	SBD	RQ	RQNA
1	0.99	0.97 (-2.8%)	1.00 (0.2%)	1.00 (0.4%)	0.97 (-2.0%)	1.00 (0.4%)
2	0.55	0.58 (6.0%)	0.56 (2.6%)	0.55 (0.2%)	0.55 (-0.1%)	0.56 (1.4%)
3	2.82	2.93 (4.2%)	2.90 (3.2%)	2.76 (-2.0%)	2.96 (5.0%)	2.75 (-2.5%)
4	1.79	1.34 (-25%)	1.41 (-21%)	1.76 (-1.6%)	2.34 (31%)	2.11 (18%)
5	2.92	2.49 (-15%)	2.44 (-17%)	2.81 (-3.6%)	3.77 (29%)	3.35 (15%)
6	0.58	0.64 (10%)	0.62 (7.4%)	0.59 (2.2%)	0.60 (3.8%)	0.49 (-16%)
7	0.24	0.24 (-1.7%)	0.26 (7.1%)	0.27 (11%)	0.23 (-3.0%)	0.24 (-1.3%)
8	0.58	0.64 (9.6%)	0.61 (4.6%)	0.60 (1.7%)	0.61 (3.9%)	0.59 (0.6%)
9	0.34	0.32 (-6.1%)	0.35 (2.0%)	0.43 (26%)	0.33 (-4.2%)	0.42 (21%)
10	0.29	0.30 (2.4%)	0.29 (1.4%)	0.28 (-1.7%)	0.28 (-1.5%)	0.26 (-8.7%)
Σ	22.0	20.3 (-7.9%)	20.4 (-7.3%)	22.4 (1.7%)	26.1 (18%)	24.2 (9.9%)

Thank You!

References

References on Robust Queueing:

[BBY15] C. Bandi, D. Bertsimas, and N. Youssef, Robust Queueing Theory, *Operations Research*, 2015.

[WY18a] W. Whitt, W. You, Using Robust Queueing to Expose the Impact of Dependence in Single-Server Queues, *Operations Research*, 2018.

[WY18c] W. Whitt, W. You, A Robust Queueing Network Analyzer Based on Indices of Dispersion, submitted to INFORMS Journal on Computing, 2018.

[WY18e] W. Whitt and W. You, The Advantage of Indices of Dispersion in Queueing Approximations, submitted to *Operations Research Letters*, 2018.

[WY17] W. Whitt, W. You, Time-Varying Robust Queueing, submitted to *Operations Research*, 2017.

References on queueing network approximations:

[DH93] J. G. Dai and J. M. Harrison, The QNET method for two-moment analysis of closed manufacturing systems, *Annals of Applied Probability*, 1993.

[FW89] K. W. Fendick, W. Whitt, Measurements and Approximations to Describe the Offered Traffic and Predict the Average Workload in a Single-Server Queue, *Proceedings of the IEEE*, 1989.

[HN90] J. M. Harrison, V. Nguyen, The QNET Method for Two-Moment Analysis of Open Queueing Networks, *Queueing Systems*, 1990.

[K11a] S. Kim, Modeling cross correlation in three-moment four-parameter decomposition approximation of queueing networks. Operations research, 59(2), 480-497, 2011.

[K11b] S. Kim, The two-moment three-parameter decomposition approximation of queueing networks with exponential residual renewal processes. Queueing Systems, 68(2), pp.193-216, 2011.

[SW86] K. Sriram, W. Whitt, Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data, *IEEE Journal on Selected Areas on Communications*, 1986.

[SW90] S. Suresh, W. Whitt, The Heavy-Traffic Bottleneck Phenomenon in Open Queueing Networks, *Operations Research Letters*, 1990.

[WM12] K. Wu, L. McGinnis, Interpolation Approximations for Queues in Series, *IIE Transactions*, 2012.

[WW82] W. Whitt, Approximating a Point Process by a Renewal Process: Two Basic Methods, *Operations Research*, 1982.

[WW83] W. Whitt, The Queueing Network Analyzer, *Bell System Technical Journal*, 1983.

[ZHS05] Q. Zhang, A. Heindl, E. Smirni, Characterizing the BMAP/MAP/1 Departure Process via the ETAQA Truncation, *Stochastic Models*, 2005.

References

References on HT limits:

[GZ06] D. Gamarnik, A. Zeevi, Validity of heavy traffic steady-state approximations in generalized Jackson Networks, *The Annals of Applied Probability*, 2006.

[IW70] D.L. Iglehart, W. Whitt, Multiple Channel Queues in Heavy Traffic II: Sequences, Networks and Batches. *Advanced Applied Probability*, 1970.

[Loy62] R. M. Loynes, The Stability of A Queue with Non-independent Inter-arrival and Service Times, *Mathematical Proceedings of the Cambridge Philosophical Society*, 1962.

[WY18b] W. Whitt, W. You, Heavy-traffic limit of the GI/GI/1 stationary departure process and its variance function, *Stochastic Systems*, 2018.

[WY18d] W. Whitt and W. You, Heavy Traffic Limits for the Stationary Flows in Generalized Jackson Networks, submitted to *Stochastic Systems*, 2018.

References on departure processes:

[D76] D. Daley, Queueing Output Processes, *Advances in Applied Probability*, 1976.

[B56] P. Burke, The Output of a Queuing System, *Operations Research*, 1956.

[G99] D. Green, Departure Processes from MAP/PH/1 Queues, thesis, 1999.

[T62] L. Takács, Introduction to the Theory of Queues, *Oxford University Press*, 1962.

[HKNT13] S. Hautphenne, Y. Kerner, Y. Nazarathy, P. Taylor, The Second Order Terms of the Variance Curves for Some Queueing Output Processes, `arXiv:1311.0069`, 2013.

[W84] W. Whitt, Approximations for Departure Processes and Queues in Series, *Naval Research Logistics Quarterly*, 1984.

Other Performance Measures

$$Z_\rho^* = \sup_{0 \le s \le \infty} \left\{ -(1-\rho)s + \sqrt{2\rho s I_w(s)/\mu} \right\}.$$

This RQ formulation give approximation of the mean steady-state workload. For other performance measures, we have

- Mean steady-state waiting time:

$$E[W] \approx \max\{0, Z^*/\rho - (c_s^2+1)/2\mu\}.$$

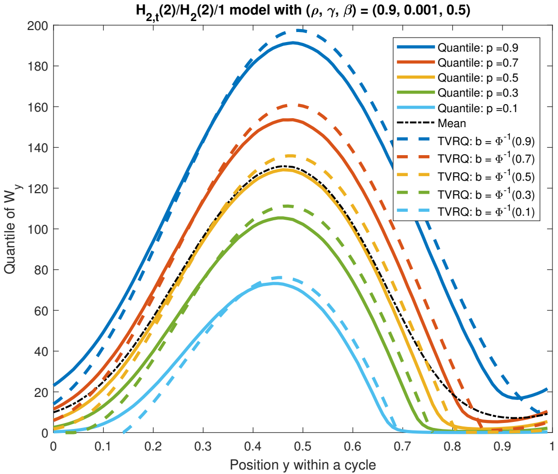
  - obtained by Brumelle's formula:

$$E[Z] = \rho E[W] + \rho \frac{E[V^2]}{2\mu} = \rho E[W] + \rho \frac{(c_s^2+1)}{2\mu}.$$

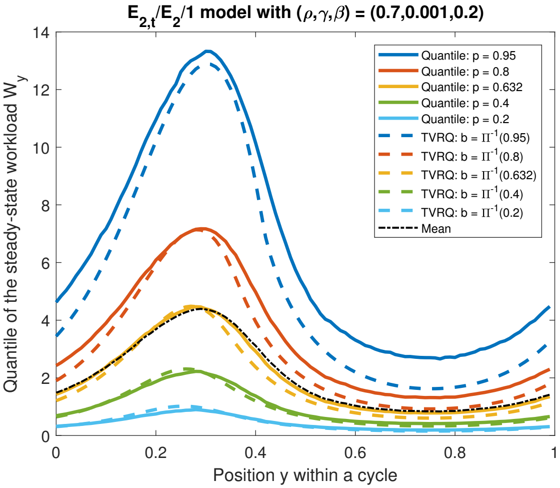
- Mean steady-state queue length, by Little's law,

$$E[Q] = \lambda E[W] = \rho E[W].$$

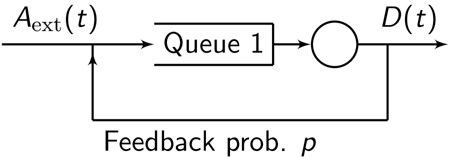
Example: Time-Varying Queue and Percentiles of the Workload



Example: Time-Varying Queue and Percentiles of the Workload



Feedback Elimination



- Normally, the immediate feedback returns the customer back to the end of the line at the same station.
- In the immediate feedback elimination procedure, the approximation step is to put the customer back at the head of the line.
  - The overall service time is then a geometric sum of the original service times.
- This does not alter the queue length process or the workload process, because the approximation step is work-conserving.

Feedback Elimination

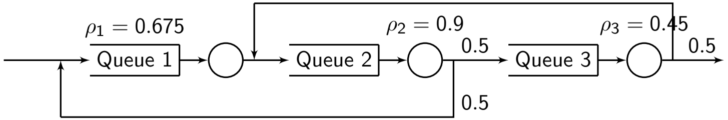


Figure: A three-station example.

For the general case,

- Near immediate feedback is defined as a feedback customer that does not go through a station with higher traffic intensity than the current station.
- For each station with feedback, we eliminate all near immediate feedback flows, the nadjust the service times just as in the single-station case.

10 Queues in Series

Table: 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
1	0.290 (2.41%)	0.45 (55%)	0.45 (55%)	0.45 (55%)	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%)
3	0.607 (1.32%)	0.72 (19%)	0.74 (22%)	0.74 (22%)	0.70 (15%)	0.72 (19%)
4	0.666 (1.20%)	0.78 (17%)	0.79 (18%)	0.79 (19%)	0.77 (16%)	0.79 (19%)
5	0.706 (1.42%)	0.83 (18%)	0.82 (16%)	0.82 (16%)	0.80 (14%)	0.83 (18%)
6	0.731 (1.78%)	0.85 (16%)	0.84 (14%)	0.84 (15%)	0.83 (13%)	0.86 (18%)
7	0.748 (1.34%)	0.87 (16%)	0.85 (14%)	0.85 (14%)	0.84 (12%)	0.88 (17%)
8	0.775 (1.68%)	0.88 (14%)	0.86 (11%)	0.86 (11%)	0.85 (9.2%)	0.89 (15%)
9	5.031 (4.31%)	7.99 (59%)	6.97 (39%)	4.05 (-20%)	4.95 (-2.0%)	4.97 (-1.3%)
Total	10.05	14.0 (39%)	13.0 (29%)	10.1 (0.09%)	10.6 (5.3%)	10.8 (7.6%)

10 Queues in Series

Table: 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
1	3.284 (3.50%)	4.05 (23%)	4.05 (23%)	4.05 (23%)	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%)
3	1.914 (3.40%)	2.19 (14%)	1.47 (-23%)	1.49 (-22%)	2.04 (6.7%)	0.98 (-49%)
4	1.719 (4.07%)	1.73 (0.64%)	1.16 (-33%)	1.19 (-31%)	1.72 (0.31%)	0.92 (-47%)
5	1.598 (3.69%)	1.43 (-11%)	1.07 (-33%)	1.10 (-31%)	1.53 (-4.1%)	0.90 (-44%)
6	1.478 (4.13%)	1.24 (-16%)	1.03 (-31%)	1.06 (-28%)	1.41 (-4.6%)	0.90 (-39%)
7	1.423 (3.23%)	1.12 (-21%)	1.00 (-30%)	1.03 (-28%)	1.33 (-6.8%)	0.90 (-37%)
8	1.413 (4.67%)	1.04 (-26%)	0.98 (-30%)	1.01 (-29%)	1.27 (-10%)	0.90 (-36%)
9	30.12 (16.8%)	8.90 (-71%)	6.04 (-80%)	36.5 (21%)	36.9 (23%)	29.1 (-3.5%)
Total	45.27	24.6 (-46%)	18.6 (-59%)	49.8 (10%)	52.8 (17%)	40.1 (-11%)

10 Queues in Series

Traffic intensity at the 10-th queue varies in $(0, 1)$.

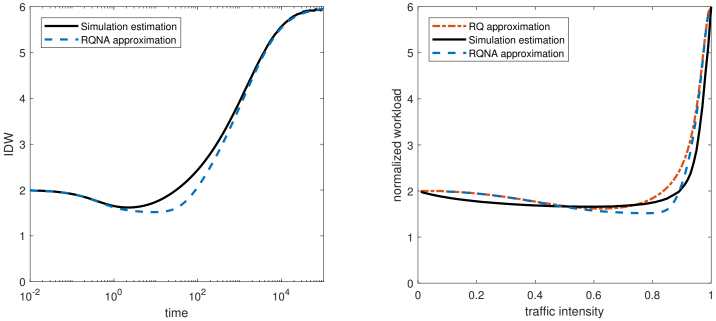


Figure: Contrasting the RQNA approximation of the IDW at the 10-th queue and simulation estimated IDW (left) in the ten queues in series example. Simulation estimation of the steady-state mean workload, the RQ approximation and the RQNA approximation shown in the right plot.

The Heavy-traffic Bottleneck Phenomenon

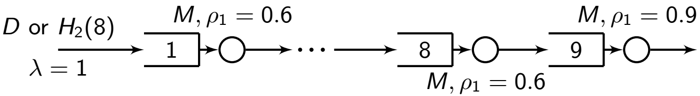
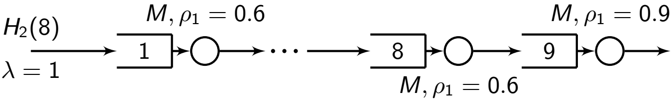


Figure: The heavy-traffic bottleneck example in Suresh and Whitt (1990).

		$H_2, c_a^2 = 8$	$D, c_a^2 = 0$
Queue 8	Simulation	1.440 ± 0.001	0.772 ± 0.000
	M/M/1	0.90 (-38%)	0.90 (17%)
	QNA	1.04 (-28%)	0.88 (14%)
	SBD	1.01 (-30%)	0.86 (11%)
Queue 9	Simulation	29.148 ± 0.049	5.268 ± 0.003
	M/M/1	8.1 (-72%)	8.1 (52%)
	QNA	8.9 (-69%)	8.0 (52%)
	SBD	36.4 (25%)	4.05 (-23%)

Table: Mean steady-state waiting times at Queue 8 and 9, compared with M/M/1 values, QNA and SBD approximations.

The Heavy-traffic Bottleneck Phenomenon



Arrival Process		$H_2, c_a^2 = 8$, $r = 0.5$	$H_2, c_a^2 = 8$, $r = 0.99$
Queue 8	Simulation	1.44	0.92
	M/M/1	0.90 (-38%)	0.90 (-2.1%)
	QNA	1.04 (-28%)	1.04 (13%)
	SBD	1.01 (-29%)	1.01 (10%)
	IR	1.20 (-17%)	1.20 (7.1%)
	RQ	1.27 (-12%)	0.92 (-0.5%)
Queue 9	Simulation	29.15	8.94
	M/M/1	8.1 (-72%)	8.1 (-9.4%)
	QNA	8.9 (-69%)	8.9 (-0.4%)
	SBD	36.5 (25%)	36.5 (308%)
	IR	21.1 (-28%)	21.1 (136%)
	RQ	37.0 (27%)	16.5 (84%)

The Heavy-traffic Bottleneck Phenomenon

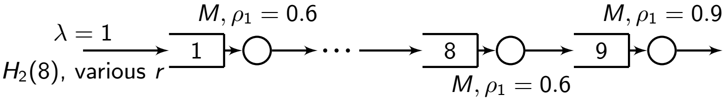
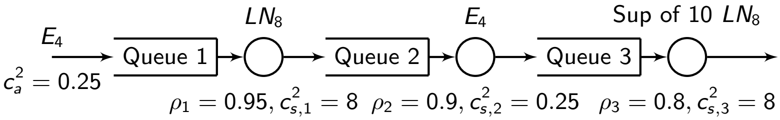
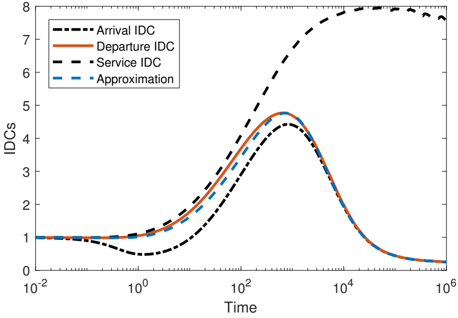


Table: Mean steady-state waiting time at each station.

r	0.9			0.5			0.1		
Queue	Sim	RQ	RQNA	Sim	RQ	RQNA	Sim	RQ	RQNA
1	1.16	1.13	1.13	3.28	3.95	3.95	5.69	5.83	5.83
2	1.16	1.12	0.95	2.32	2.61	1.58	2.46	2.40	2.71
3	1.15	1.11	0.91	1.91	2.04	0.98	1.98	1.83	1.28
4	1.14	1.10	0.90	1.71	1.72	0.92	1.76	1.56	0.97
5	1.14	1.10	0.90	1.59	1.53	0.90	1.63	1.41	0.91
6	1.13	1.09	0.90	1.47	1.41	0.90	1.54	1.31	0.90
7	1.13	1.08	0.90	1.42	1.33	0.90	1.48	1.24	0.90
8	1.12	1.08	0.90	1.41	1.27	0.90	1.42	1.20	0.90
9	19.6	36.5	27.2	30.1	36.9	29.1	29.6	36.3	29.3
Total	28.8	45.3	33.8	45.3	52.8	40.1	47.5	53.1	43.7
Avg. abs. RE		13%	20%		9.7%	34%		12%	28%

An Artificial Example





3 Stations with Feedback

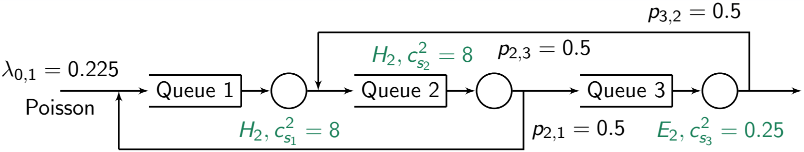


Table: The steady-state mean waiting time.

r = 0.5, (third parameter of H_2 dist., weight on one mean)				
Queue	ρ	Simu	QNET	SBD
1	0.9	31.22	35.9 (15%)	26.0 (-17%)
2	0.675	8.32	10.2 (23%)	11.1 (33%)
3	0.45	2.00	1.89 (5.5%)	1.94 (3%)
Total		138.7	161.3 (16%)	135.3 (-2.5%)
r = 0.99, (third parameter of H_2 dist., weight on one mean)				
Queue	ρ	Simu	QNET	SBD
1	0.9	27.67	35.9 (30%)	26.0 (-6.0%)
2	0.675	2.67	10.2 (282%)	11.1 (316%)
3	0.45	0.56	1.89 (236%)	1.94 (245%)
Total		103.8	161.3 (55%)	135.3 (30%)

Indices of Dispersion for Counts (IDC)

$r = 0.5$, (third parameter of H2 dist, weight on one mean)

Queue	ρ	Simu	QNET	SBD
1	0.9	31.22	35.9 (15%)	26.0 (-17%)
2	0.675	8.32	10.2 (23%)	11.1 (33%)
3	0.45	2.00	1.89 (5.5%)	1.94 (3%)
Total		138.7	161.3 (16%)	135.3 (-2.5%)

$r = 0.99$, (third parameter of H2 dist, weight on one mean)

Queue	ρ	Simu	QNET	SBD
1	0.9	27.67	35.9 (30%)	26.0 (-6.0%)
2	0.675	2.67	10.2 (282%)	11.1 (316%)
3	0.45	0.56	1.89 (236%)	1.94 (245%)
Total		103.8	161.3 (55%)	135.3 (30%)