A Robust Queueing Network Analyzer Based on Index of Dispersion

Part I

Wei You
(joint work with Ward Whitt)

Queueing Seminar, Columbia University

September 19, 2018

Motivation

- Many complex service systems can be modeled as a open network of queues.
- The estimation of performance measures in a open queueing network (OQN) is important in many OR applications.
    - Theoretical analysis are limited for queueing networks with general distributions.
    - Direct simulation estimation may be computational expensive.
- This work continues the decades-long search for an accurate and computationally light-weighted approximation algorithms.

Background - Previous Approximation Algorithms

Decomposition

approximation methods

- Motivated by the product-form solution of a Jackson Network.
- Treat each station as independent single-server queues.

Examples

- The Queueing Network Analyzer (QNA) by Whitt (1983),
  - approximates each station by a GI/GI/1 queue
  - may fail in certain cases because the dependence makes arrival process non-renewal, see Suresh and Whitt (1990).
- Kim (2011a, 2011b)
  - approximate each station by a MMPP(2)/GI/1 queue (Markov-Modulated Poisson Process);
  - dependence in the arrival process is approximated by the MMPP.

Background - Previous Approximation Algorithms

Approximations using Relfected Brownian Motion (RBM)

- Approximate the steady-state queue length distribution by the stationary distribution of the limiting RBM;
- numerically calculate the steady-state mean of the RBM.

Examples

- QNET by Harrison and Nguyen (1990) for OQNs and Dai and Harrison (1993) for CQNs;
  - computation time scales with the system.
- Sequential Bottleneck Decomposition (SBD) proposed by Dai, Nguyen and Reiman (1994),
  - decompose the network into sub-networks by traffic intensities;
  - reduced the computation burden.

Background - Recent Developments

Recent Developments

- Interpolation method (IR) by Wu and McGinnis (2014)
  - Approximate a station by the interpolation of two or more auxiliary systems.
- Robust Queueing (RQ) by Bandi et al. (2015)
  - Replace probabilistic law by uncertainty sets and utilize robust optimization.

Background - Previous Approximation Algorithms

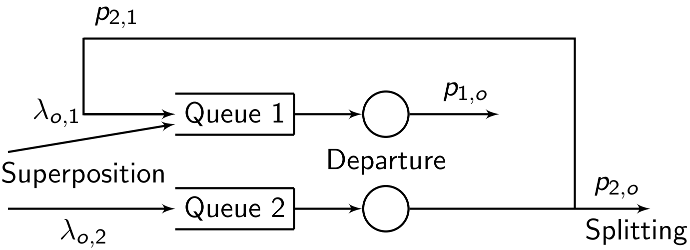
From Austin's talks

- Robust Queueing (RQ) by Bandi et al. (2015);
- a parametric RQ for customer waiting time.

We followed the RQ framework and developed

- a functional RQ for the workload process in G/G/1 models;
- bridges between RQ and RQNA
  - heavy-traffic limits of the stationary internal flows (arrival, departure processes, etc.);
- a RQNA for open queueing networks.

Dependence in Queues



Dependence rises naturally in queueing network:

- Dependence within the flows[1]:
  - introduced by departure and superposition operations.
- Dependence between the flows:
  - introduced by all three network operations.

[1] arrival processes, departure process, etc.

Dependence in Queues

Dependence has significant impact on performance measures

- Dependence may have complicated temporal structure.
- The level of impact will depend on the temporal structure and the traffic intensity.
  - As a result, parametric methods (QNA, RQ by Bandi et al.) using first two moments to describe variability may fail.
- Indices of dispersion can describe the temporal structure.
  - Fendick and Whitt (1989) first applied indices of dispersion in queueing approximation.

The Heavy-traffic Bottleneck Phenomenon

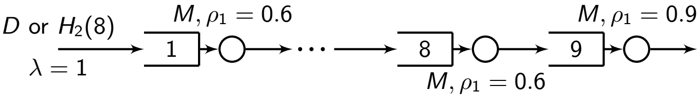
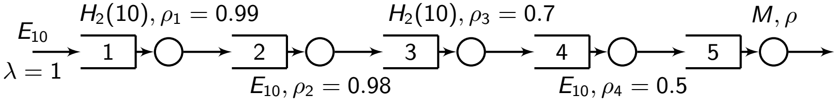


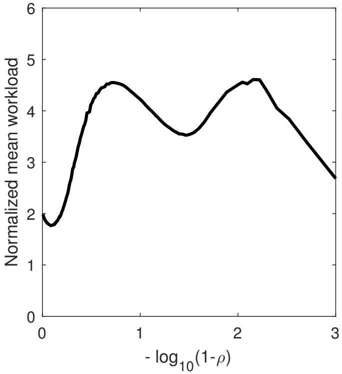
Figure: The heavy-traffic bottleneck example in Suresh and Whitt (1990).

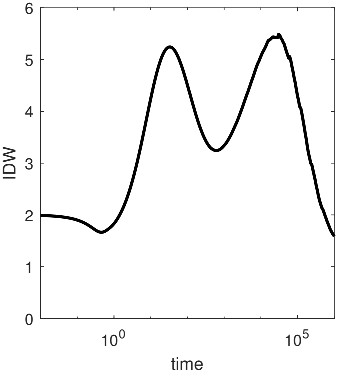
		$H_2, c_a^2 = 8$	$D, c_a^2 = 0$
Queue 8	Simulation	1.440 ± 0.001	0.772 ± 0.000
	M/M/1	0.90 (-38%)	0.90 (17%)
	QNA	1.04 (-28%)	0.88 (14%)
Queue 9	Simulation	29.148 ± 0.049	5.268 ± 0.003
	M/M/1	8.1 (-72%)	8.1 (52%)
	QNA	8.9 (-69%)	8.0 (52%)

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

One More Example





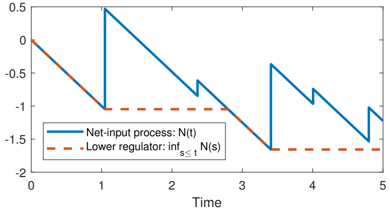


But how to match IDW to mean workload? RQ can help!

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 defined by $Y(t) \equiv \sum_{k=1}^{A(t)} V_k$;
- $N(t)$: net-input process defined by $N(t) \equiv Y(t) - t$;

Continuous-time workload process



The steady-state workload at time 0 in the queue staring empty at the remote past $-\infty$:

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

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

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

Stochastic versus Robust Queues

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

Stochastic Queue

- $N(s) \equiv \sum_{k=1}^{A(s)} V_k - s$, where $A(t)$ and $\{V_k\}$ are stationary point process and stationary sequence, respectively.

Robust Queue

- $\tilde{N}$ lies in a suitable uncertainty set $\mathcal{U}$ of total input functions to be defined later.
- There is no distribution involved, we hence focus on 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

In our specific settings, we have the following uncertainty set motivated from CLT

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

where $N_\rho(t) = Y_\rho(t) - t$ is the net input process associated with the stochastic queue with traffic intensity ρ, so

$$E[N_\rho(t)] = E[Y_\rho(t) - t] = \rho t - t,$$
$$\mathrm{Var}(N_\rho(t)) = \mathrm{Var}(Y_\rho(t) - t) = \mathrm{Var}(Y_\rho(t)).$$

- Choose $b = \sqrt{2}$ so that RQ is exact for $M/GI/1$ models.

Index of Dispersion for Work

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

The index of dispersion for work (IDW) for the net-input process $Y_\rho(t)$ is defined as

$$I_w(t) \equiv \frac{\mathrm{Var}(Y_\rho(t))}{E[Y_\rho(t)]E[V]} = \frac{\mathrm{Var}(Y_\rho(t))}{\rho t/\mu}$$

$$\Rightarrow \quad \mathrm{Var}(Y_\rho(t)) = \rho t I_w(t)/\mu$$

$$\Rightarrow \quad \mathcal{U}_\rho = \left\{ \tilde{N}_\rho : \tilde{N}_\rho(s) \leq -(1-\rho)s + \sqrt{2\rho s I_w(s)/\mu},\ s \geq 0 \right\}.$$

Robust Queueing for continuous-time workload

RQ for workload

$$Z_\rho^* = \sup_{N_\rho \in \mathcal{U}_\rho} \sup_{0 \le s \le \infty} \{N_\rho(s)\},$$

where

$$\mathcal{U}_\rho = \left\{ \tilde{N}_\rho : \tilde{N}_\rho(s) \le -(1-\rho)s + \sqrt{2\rho s I_w(s)/\mu},\ s \ge 0 \right\}.$$

Lemma (Dimension reduction)

The infinite-dimensional RQ problem can be reduced to one-dimensional

$$\begin{aligned} Z_\rho^* &= \sup_{0 \le s \le \infty} \sup_{N_\rho \in \mathcal{U}_\rho} \{N_\rho(s)\} \\ &= \sup_{0 \le s \le \infty} \left\{ -(1-\rho)s + \sqrt{2\rho s I_w(s)/\mu} \right\}. \end{aligned}$$

Furthermore, if $\rho < 1$ and $I_w(t)/t \to 0$ as $t \to \infty$, then $Z_\rho^* < \infty$.

Robust Queueing for continuous-time workload

In summary, the RQ algorithm for single-server queues

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

This formulation requires IDW I_w as model input

- I_w is defined for the stationary net-input process;
- I_w can be
  - calculated in special cases;
  - estimated by simulation or from historical data; or
  - approximated (RQNA).
- for stations without feedback, same I_w is used for all $\rho \in [0, 1)$;

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,

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

Remarks on the RQ algorithm

For $G/GI/1$ models, where

- arrival process is a general, ergodic point process;
- service times are i.i.d, independent of the arrival process.

Theorem (RQ correct in heavy-traffic and light-traffic)

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.

- For stations with customer feedback, a feedback elimination procedure is needed to obtain exact HT limit.

The Heavy-traffic Bottleneck Phenomenon

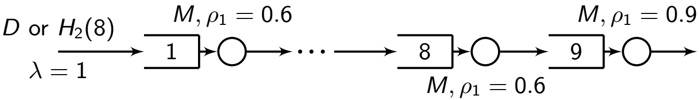
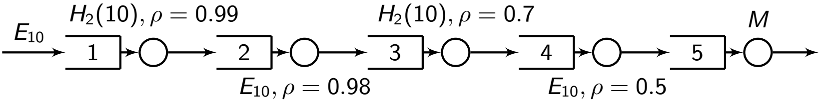
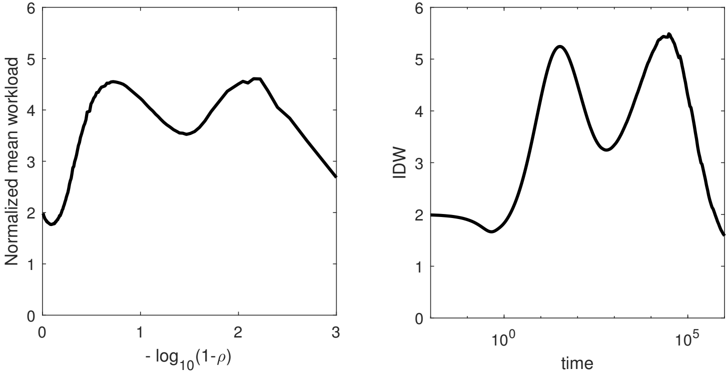


Table: The heavy-traffic bottleneck example

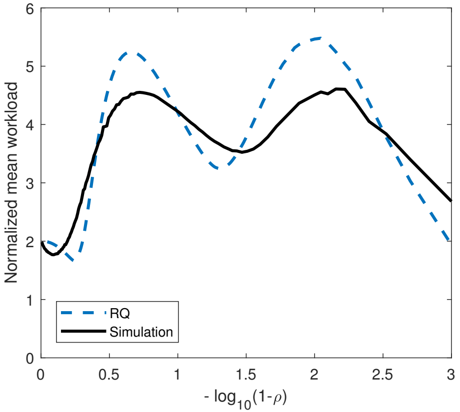
		$H_2, c_a^2 = 8$	$D, c_a^2 = 0$
Queue 8	Simulation	1.440 ± 0.001	0.772 ± 0.000
	M/M/1	0.90 (-38%)	0.90 (17%)
	QNA	1.04 (-28%)	0.88 (14%)
	SBD	1.01 (-30%)	0.86 (11%)
	IR	1.20 (-17%)	0.86 (11%)
	RQ	1.27 (-12%)	0.85 (11%)
Queue 9	Simulation	29.148 ± 0.049	5.268 ± 0.003
	M/M/1	8.1 (-72%)	8.1 (52%)
	QNA	8.9 (-69%)	8.0 (52%)
	SBD	36.4 (25%)	4.05 (-23%)
	IR	21.1 (-28%)	6.25 (19%)
	RQ	37.0 (27%)	4.95 (-6.0%)

Numerical Example: 5 queues in series





Numerical Examples - 5 Queues in series



- RQ automatically “matches” IDW to the mean workload for all traffic intensities.

More on IDW and IDC

When the service times are i.i.d., independent of the arrival process, we have

$$I_w(t) = I_a(t) + c_s^2,$$

where $I_a(t)$ is the index of dispersion for counts (IDC) associated with the arrival counting process $A(t)$

$$I_a(t) = \frac{Var(A(t))}{E[A(t)]}.$$

To calculate/estimate the IDC of a stationary point process,

- let

  $$V(t) \equiv Var(A(t))$$

  where the variance is taken under the stationary distribution.
- for stationary point process, we have $E[A(t)] = \lambda t$;

More on IDW and IDC

For stationary and ergodic point processes, taking Laplace transform on the variance function $V(t)$, we have

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

so

$$V(t) = \lambda \int_0^t (1 + 2m(u) - 2\lambda u)du.$$

- $m(t) = E^0[A(t)]$ under Palm distribution P^0, i.e., conditioning on having an arrival at time 0.
- It is the renewal function in the case of renewal processes. Let $\hat{f}(s) = \int_0^\infty e^{-st}dF(t)$, then

$$\hat{m}(s) = \frac{\hat{f}(s)}{s(1 - \hat{f}(s))}.$$

More on IDW and IDC

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

- By rearranging terms, $\hat{f}$ can be expressed by $\hat{V}(s)$;
- $\Rightarrow$ IDCs completely characterize a GI/GI/1 queue;
- By using IDW (IDC), the RQ algorithm utilizes much more information than just the first two moments, hence is potentially more accurate and adaptive.

The Heavy-traffic Bottleneck Phenomenon

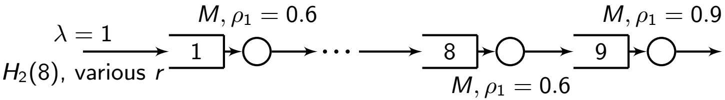


Table: Mean steady-state waiting time at each station.

r	1.0	0.9		0.5		0.1		N/A	N/A	N/A
Q	Exact	Sim	RQ	Sim	RQ	Sim	RQ	QNA	QNET	SBD
1	0.90	1.16	1.13	3.28	3.95	5.69	5.83	4.05	4.05	4.05
2	0.90	1.16	1.12	2.32	2.61	2.46	2.40	2.92	1.81	1.82
3	0.90	1.15	1.11	1.91	2.04	1.98	1.83	2.19	1.47	1.49
4	0.90	1.14	1.10	1.71	1.72	1.76	1.56	1.73	1.16	1.19
5	0.90	1.14	1.10	1.59	1.53	1.63	1.41	1.43	1.07	1.10
6	0.90	1.13	1.09	1.47	1.41	1.54	1.31	1.24	1.03	1.06
7	0.90	1.13	1.08	1.42	1.33	1.48	1.24	1.12	1.00	1.03
8	0.90	1.12	1.08	1.41	1.27	1.42	1.20	1.04	0.98	1.01
9	8.10	19.6	36.5	30.1	36.9	29.6	36.3	8.9	6.0	36.4
sum	15.3	28.8	45.3	45.3	52.8	47.5	53.1	24.6	18.6	49.8

Examples with Fixed SCV's

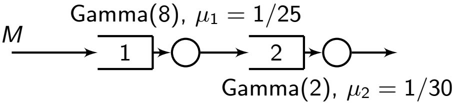


Table: Mean steady-state waiting time at station 2 of two queue in series. Example taken from Appendix C of Wu and McGinnis (2014).

ρ_2	Sim	QNA	QNET	IR	RQ
0.1	6.8	-25.7%	68.0%	97.2%	-0.8%
0.2	16.6	-27.9%	54.8%	80.2%	0.4%
0.3	30.5	-27.6%	43.4%	65.2%	0.4%
0.4	50.4	-25.1%	33.8%	52.1%	1.0%
0.5	79.5	-20.5%	25.7%	40.5%	2.4%
0.6	124.4	-14.1%	18.4%	29.3%	3.0%
0.7	198.2	-4.9%	12.7%	18.9%	2.1%
0.8	339.3	8.1%	7.9%	8.1%	-1.5%
0.9	704.3	33.0%	6.4%	-2.7%	-8.6%
0.95	1330.0	58.3%	7.4%	-9.1%	-17%

Examples with Fixed SCV's

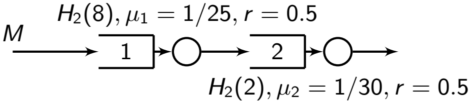


Table: Mean steady-state waiting time at station 2.

ρ_2	Sim	QNA	QNET	RQ
0.1	6.2	5.1 (-18.5%)	11.4 (84.2%)	5.2 (-15.9%)
0.2	15.1	12.0 (-20.7%)	25.7 (70.1%)	12.5 (-17.4%)
0.3	28.0	22.1 (-21.1%)	43.7 (56.2%)	23.3 (-16.9%)
0.4	47.0	37.7 (-19.6%)	67.4 (43.4%)	40.9 (-12.9%)
0.5	75.6	63.2 (-16.3%)	99.9 (32.1%)	71.8 (-5.0%)
0.6	120.7	106.9 (-11.4%)	147.3 (22.0%)	124.3 (3.0%)
0.7	197.5	188.5 (-4.5%)	223.4 (13.0%)	209.7 (6.2%)
0.8	345.7	366.8 (6.0%)	366.1 (5.9%)	354.3 (2.5%)
0.9	732.1	936.7 (27.9%)	749.4 (2.3%)	680.9 (-7.0%)
0.95	1359.8	2105.4 (54.8%)	1428.4 (5.0%)	1153.1 (-15.2%)

Examples with Fixed SCV's

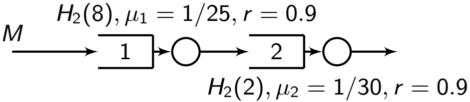


Table: Mean steady-state waiting time at station 2.

ρ_2	Sim	QNA	QNET	RQ
0.1	5.5	5.1 (-7.2%)	11.4 (107.2%)	5.0 (-8.8%)
0.2	13.1	12.0 (-8.3%)	25.7 (96.1%)	11.4 (-12.8%)
0.3	23.9	22.1 (-7.5%)	43.7 (82.8%)	19.9 (-16.7%)
0.4	39.8	37.7 (-5.2%)	67.4 (69.3%)	31.7 (-20.3%)
0.5	64.0	63.2 (-1.2%)	99.9 (56.0%)	49.8 (-22.1%)
0.6	104.3	106.9 (2.4%)	147.3 (41.2%)	82.1 (-21.2%)
0.7	178.8	188.5 (5.4%)	223.4 (24.9%)	156.9 (-12.2%)
0.8	332.8	366.8 (10.2%)	366.1 (10.0%)	355.5 (6.8%)
0.9	820.7	936.7 (14.1%)	749.4 (-8.6%)	849.0 (3.4%)
0.95	1661.8	2105.4 (26.6%)	1428.4 (-14.0%)	1472.8 (-11.3%)

The Heavy-traffic Bottleneck Phenomenon

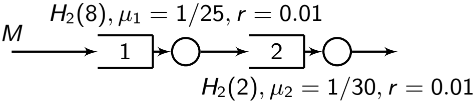


Table: Mean steady-state waiting time at station 2.

ρ_2	Sim	QNA	QNET	RQ
0.1	7.1	5.1 (-28.1%)	11.4 (60.5%)	7.5 (5.6%)
0.2	17.2	12.0 (-30.2%)	25.7 (49.4%)	18.4 (6.9%)
0.3	31.4	22.1 (-29.6%)	43.7 (39.1%)	33.7 (7.3%)
0.4	51.6	37.7 (-26.9%)	67.4 (30.6%)	55.3 (7.1%)
0.5	81.2	63.2 (-22.1%)	99.9 (23.0%)	86.0 (5.9%)
0.6	125.6	106.9 (-14.8%)	147.3 (17.2%)	131.1 (4.3%)
0.7	198.4	188.5 (-4.9%)	223.4 (12.6%)	200.9 (1.2%)
0.8	334.6	366.8 (9.6%)	366.1 (9.4%)	325.0 (-2.8%)
0.9	693.1	936.7 (35.1%)	749.4 (8.1%)	623.9 (-9.9%)
0.95	1291.4	2105.4 (63.0%)	1428.4 (10.6%)	1090.6 (-15.5%)

Generalization to RQNA

- The RQ algorithm serve as the building blocks for an Robust Queueing Network Analyzer (RQNA) algorithm;
- How do we establish connections between blocks?

Generalization to RQNA

Recall that

- RQ relies on estimating the IDW at the queue of interest;
- IDW is crucial for RQ to produce useful approximations.

A simplifying assumption

- If we assume that service times are i.i.d., independent of everything else, then

$$I_w(t) = I_a(t) + c_s^2,$$

where c_s^2 is the squared coefficient of variation (scv) of the service distribution and $I_a(t)$ is the *index of dispersion for counts* (IDC) associated with the arrival counting process $A(t)$

$$I_a(t) = \frac{Var(A(t))}{E[A(t)]}.$$

Generalization to RQNA

To extend the RQ algorithm, we need to

- (for external flows[2]) provide effective algorithm to calculate/estimate the IDC of a stationary point process;
- (for internal flows[3]) produce effective approximations internal arrival IDC at any queue within a open queueing network;

[2] Service processes, external arrival processes.

[3] Internal arrival processes, departure processes.

Generalization to RQNA: External Flows

- estimate via numerical inversion

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

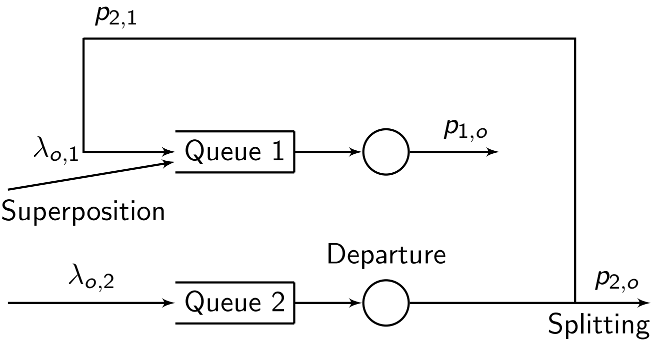
$$V(t) = \lambda \int_0^t (1 + 2m(u) - 2\lambda u)du.$$

- estimate via Monte Carlo with some variance reduction techniques.

Generalization to RQNA: Internal Flows

The total arrival process at any queue:

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



Historical Remarks on Departure Processes

- In general, departure processes are complicated, even for M/GI/1 or GI/M/1 special cases;
- Even more, the IDC we used is defined for **stationary version** of the departure process, instead of the departure from a system starting empty.
  - It is important that we use stationary version of the IDC (IDW), otherwise we do not have correct light traffic limit.

Historical Remarks on 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

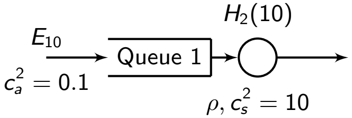
- Iglehart and Whitt (1970), HT limits for departure process in systems that starts empty;
- Gamarnik and Zeevi (2006) and Budhiraja and Lee (2009), HT limit for stationary queueing length process.

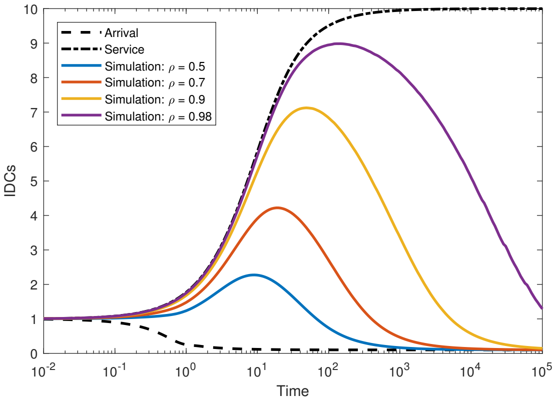
Historical Remarks on Departure Processes

Approximations

- Whitt (1982, 1983, 1984): QNA and related papers:
  - the asymptotic method: matching the long-run property of a point process
$$c_d^2 \approx c_a^2$$
  - the stationary interval method: matching the stationary interval distribution, but ignore dependence between successive departures
$$c_d^2 = c_a^2 + 2\rho^2 c_s^2 - 2\rho(1-\rho)E[W] \approx \rho^2 c_a^2 + (1-\rho^2)c_s^2$$

A numerical example





Heavy-Traffic Limit for the Departure Processes

Theorem (HT limit for the stationary departure process)

For GI/GI/1 queue under regularity conditions,

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

- B_a *and* B_s *are independent standard Brownian motions;*
- $Q^*(t) = \psi(Q^*(0) + c_a B_a - c_s B_s - e)$ *is the HT limit for stationary queue length process: a stationary reflective Brownian motion (RBM)* R_e *with drift* -1*, variance* $c_x^2 \equiv c_a^2 + c_s^2$*;*
- $Q^*(0) \sim \exp(2/c_x^2)$ *is the exponential marginal distribution;*
- B_a*,* B_s *and* $Q^*(0)$ *are mutually independent.*

Heavy-Traffic Limit for the Variance Functions

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

Under assumptions in Theorem plus uniform integrability conditions, $V^*_{d,\rho}$ converges to

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

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.*

Approximation for Departure IDC

Let $I_{d,\rho}$ be the departure IDC in the model with traffic intensity ρ. Define the weight function

$$w_\rho(t) \equiv \frac{I_{d,\rho}(t) - I_s(t)}{I_a(t) - I_s(t)} = \frac{V_{d,\rho}(t) - V_s(t)}{V_a(t) - V_s(t)},$$

where I_a and I_s are the IDC of the **base** arrival and service processes (both with rate 1). The HT-scaled weight function

$$w^*_\rho(t) = w_\rho((1-\rho)^{-2}t).$$

Approximation for Departure IDC

Corollary

Under the assumptions in the HT departure variance theorem, we have $w_{\rho}^{*}(t) \Rightarrow w^{*}(t/c_x^2)$.

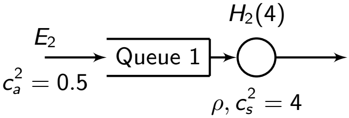
The corollary supports the following approximation

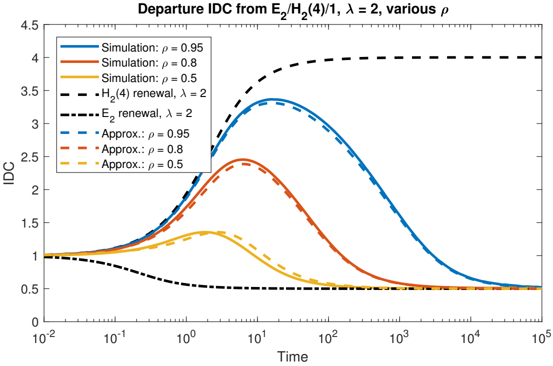
$$w_{\rho}(t) \approx w^{*}((1-\rho)^2 t/(\rho c_x^2)),$$

and

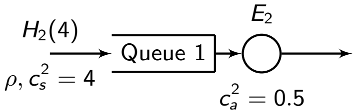
$$\begin{aligned} I_{d,\rho}(t) &= w_{\rho}(t) I_a(t) + (1 - w_{\rho}(t)) I_s(t) \\ &\approx w^{*}((1-\rho)^2 t/(\rho c_x^2)) I_a(t) + (1 - w^{*}((1-\rho)^2 t/(\rho c_x^2))) I_s(t). \end{aligned}$$

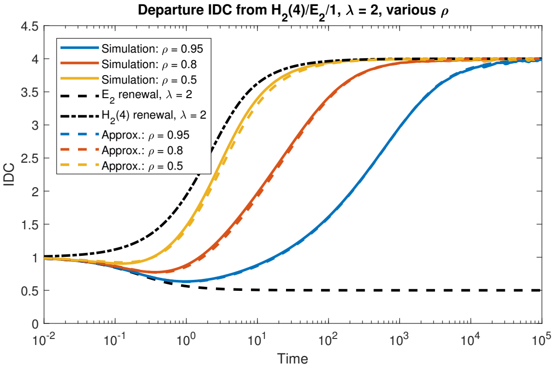
A Simple Example





A Second Example





The Heavy-traffic Bottleneck Phenomenon

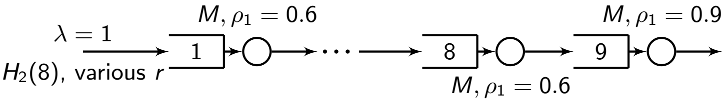


Table: Mean steady-state waiting time at each station.

r	0.9			0.5			0.1		
Q	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
sum	28.8	45.3	33.8	45.3	52.8	40.1	47.5	53.1	43.7

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.

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.

[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.

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

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.

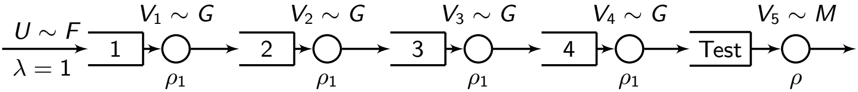
[G99] D. Green, Departure Processes from MAP/PH/1 Queues, thesis, 1999.

[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.

More Numerical Examples



Now, we look at a batch of examples:

- consider 4 identical queues in tandem:
    - same service distributions G;
    - same traffic intensity $\rho_1 = 0.7$ or 0.9;
- attach a test queue to the end of the 4 identical queues;
    - traffic intensity ρ at the test queue range from 0 to 1;
- arrival distribution F picked from: E4, LN025, LN4, H4;
- service distribution G picked from: E4, LN025, LN4, H4,M;
- a total of $2 \times 4 \times 5 = 40$ examples.

We assess the performance of RQ algorithm at the test queue.

More Numerical Examples

- $|\mathtt{RE}| = |\mathtt{RE}_\rho|$: relative error (as a function of traffic intensity) between the RQ approximation and the simulation estimation;
- `max(|RE|)`: for fixed example, the maximum relative error across different traffic intensities;
- `avg(|RE|)`: for fixed example, the simple average of the relative error across different traffic intensities;
- `Max` and `Mean` run over different example instances;

```
========================== rho = 0.7 ============================
* Max max(|RE|) for RQ = 33.01%. Mean max(|RE|) for RQ = 16.85%.
* Max avg(|RE|) for RQ = 15.47%. Mean avg(|RE|) for RQ =  7.50%.
============================= End ===============================

========================== rho = 0.9 ============================
* Max max(|RE|) for RQ = 37.36%. Mean max(|RE|) for RQ = 17.66%.
* Max avg(|RE|) for RQ = 11.69%. Mean avg(|RE|) for RQ =  6.52%.
============================= End ===============================
```

Review of Robust Queueing Theory

Bandi et al. consider a $GI/GI/1$ FCFS queue with

- $\{(U_i, V_i)\}_{i\geqslant 1}$: interarrival times and service times;
- λ, μ: arrival rate and service rate.

Lindley recursion

$$W_n = (W_{n-1} + V_{n-1} - U_{n-1})^+ = \max_{0\leqslant k\leqslant n} \{S_k^s - S_k^a\},$$

where $S_0^s \equiv 0$, $S_0^a \equiv 0$ and

$$S_k^s \equiv \sum_{i=n-k}^{n-1} V_i, \quad S_k^a := \sum_{i=n-k}^{n-1} U_i, \quad 1 \leq k \leq n.$$

- Loynes (1962) reverse-time construction;
- Lindley recursion holds for any sequence of $\{(U_i, V_i)\}$, not just i.i.d. random variables.

Review of Robust Queueing

As in usual robust optimization applications, Bandi et al. (2015) proposed to

- draw interarrival and service times from properly defined *uncertainty sets* instead of probability distributions;
- use *worst case scenario* instead of probabilistic statements (mean, distribution...) to characterize system performance.

Review of Robust Queueing

The worst case waiting time can be written as

$$W_n^* \equiv \sup_{\mathbf{U}\in\mathcal{U}^a} \sup_{\mathbf{V}\in\mathcal{U}^s} W_n(\mathbf{U},\mathbf{V}) = \sup_{\mathbf{U}\in\mathcal{U}^a} \sup_{\mathbf{V}\in\mathcal{U}^s} \max_{0\leqslant k\leqslant n} \{S_k^s - S_k^a\}$$

Motivated by CLT, Bandi et al. proposed

$$\mathcal{U}^a = \left\{(U_1,\ldots,U_n) \,\middle|\, \frac{S_k^a - k/\lambda}{k^{1/2}} \geqslant -\Gamma_a, 0 \leqslant k \leqslant n\right\},$$

$$\mathcal{U}^s = \left\{(V_1,\ldots,V_n) \,\middle|\, \frac{S_k^s - k/\mu}{k^{1/2}} \leqslant \Gamma_s, 0 \leqslant k \leqslant n\right\}.$$

- CLT suggest that $\Gamma_a = b_a\sigma_a$ and $\Gamma_s = b_s\sigma_s$.

Review of Robust Queueing

With an interchange of maximum, they reduce the problem to

$$W_n^* = \max_{0 \le k \le n} \{mk + b\sqrt{k}\}$$
$$\le \sup_{x \ge 0} \{mx + b\sqrt{x}\} = \frac{b^2}{4|m|} = \frac{\lambda b^2}{4(1-\rho)},$$

where $m = \mu^{-1} - \lambda^{-1} < 0$, $\rho = \lambda/\mu$ and $b \equiv \Gamma_a + \Gamma_s > 0$, so that $b^2 = \Gamma_a^2 + 2\Gamma_a\Gamma_s + \Gamma_s^2$.

- Closed-form solution depends only on ρ, Γ_a and Γ_s.
- The solution resembles classical heavy-traffic limit approximations or bounds, e.g., Kingman Bound

$$W_\rho^* \le \frac{\rho(\rho^{-2}c_a^2 + c_s^2)}{2\mu(1-\rho)}.$$

Review of Robust Queueing: Extension to OQN

Bandi et al. obtain an algorithm for queueing networks by assuming

- the network is **feed-forward**, i.e., no customer feedback;
- the servers are **adversary**, i.e, they pick service times such that customer waiting times are maximized.

Under assumptions above, they

- proved a (robust) **Burke's theorem**, i.e. departure falls in the same uncertainty set as the one for arrival;
- apply **linear regression** to fit Γ_a and Γ_s for external arrival processes and service processes;
- used similar **network calculus** as in QNA to determine parameters Γ_a and Γ_s;

An Artificial Example