Robust Queueing for a Series of Queues

Wei You
(with Ward Whitt)
IEOR, Columbia

INFORMS 2017
Houston, TX

October 25, 2017

Outline

Motivation

- The estimation of performance measures in a open network of queues is important in many OR applications.
- Theoretical analysis are limited for queueing network with general distributions.
- Direct simulation estimation may be computational expensive,
  - especially if doing many "what if" studies or when performing an optimization over model parameters.

Background

Traditionally, queueing systems are approximated by

- Parametric-decomposition methods using variability parameters: e.g., QNA by Whitt (1983);
- Relfected Brownian motion approximations: e.g., QNET by Dai and Harrison (1993);

More recently,

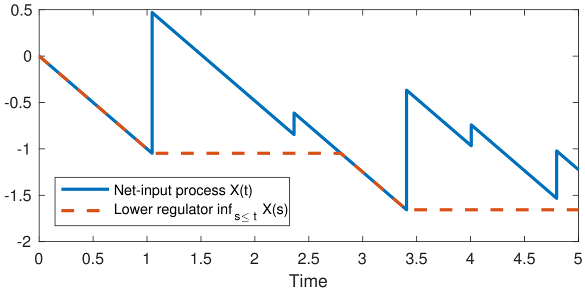
- Robust Queueing (RQ) by Bandi et al. (2015), analyzes the mean steady-state waiting time in a queueing network.
- Whitt and You (2017): RQ formulation for the workload (virtual waiting time) process in G/G/1 models.
  - Based on the Index of Dispersion for Work (IDW), see Fendick and Whitt (1989) for discussion of the IDW.

Robust Queueing for continuous-time workload

- $A_\rho(t) = A(\rho t)$: arrival counting process, $A(t)$ with rate 1;
- $\{V_i\}$: mean-1 service times;
- $Y_\rho(t) \equiv \sum_{k=1}^{A_\rho(t)} V_k \equiv Y(\rho t)$: total input of work;
- $X_\rho(t) \equiv Y_\rho(t) - t$: net-input process.

The steady-state workload at time t

$$Z_\rho \equiv X_\rho(t) - \inf_{s \le t}\{X_\rho(s)\}.$$



Robust Queueing for continuous-time workload

Under RQ framework, instead of probabilistic distribution for the net-input process, we work with the uncertainty set.

$$\mathcal{U}_\rho \equiv \left\{ \tilde{X}_\rho : \mathbb{R}^+ \to \mathbb{R} \;\middle|\; \tilde{X}_\rho(s) \leq E[X_\rho(s)] + b\sqrt{\mathrm{Var}(X_\rho(s))}, s \in \mathbb{R}^+ \right\}$$
$$= \left\{ \tilde{X}_\rho : \mathbb{R}^+ \to \mathbb{R} \;\middle|\; \tilde{X}_\rho(s) \leq -(1-\rho)s + b\sqrt{\rho s I_w(\rho s)}, s \in \mathbb{R}^+ \right\},$$

where

$$E[X_\rho(s)] = -(1-\rho)s,$$
$$\mathrm{Var}(X_\rho(s)) = \mathrm{Var}(X_\rho(s) - s) = \mathrm{Var}(Y_\rho(s)) = \mathrm{Var}(Y(\rho s))$$

and $I_w(t)$ is the *index of dispersion for work* (IDW), i.e.,

$$I_w(t) \equiv \frac{\mathrm{Var}(Y(t))}{t}.$$

Robust Queueing for continuous-time workload

The RQ algorithm

$$Z_\rho^* = \max_{X \in \mathcal{U}_\rho} Z_\rho(X) \equiv X(t) - \inf_{s \leq t}\{X(s)\}$$

where the uncertainty set is defined as

$$\mathcal{U}_\rho = \left\{ \tilde{X}_\rho : \mathbb{R}^+ \to \mathbb{R} \;\middle|\; \tilde{X}_\rho(s) \leq -(1-\rho)s + b\sqrt{\rho s I_w(\rho s)}, s \in \mathbb{R}^+ \right\},$$

Theorem (Whitt and You(2017))

The RQ solution is

$$Z^* = \sup_{s \geqslant 0} \left\{ -(1-\rho)s/\rho + \sqrt{2 s I_w(s)} \right\}.$$

Under regularity conditions, the RQ solution is asymptotically exact for G/G/1 models under light-traffic and heavy-traffic limits.

A Series of Queues

Regularity assumptions

- each queue is FCFS with a single server and unlimited waiting space;
- stationary and ergodic external arrival process
  - with finite rate and variance.
- service times have finite variance;
- traffic intensity at each queue is less than 1.

A Series of Queues

Simplifying assumption

- service times at each queue are i.i.d., independent of the external arrival process.

This implies that

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

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

$$I_a(t) \equiv \frac{Var(A(t))}{E[A(t)]};$$

RQ algorithm

$$\begin{aligned} Z^* &= \sup_{s \geqslant 0} \left\{ -(1-\rho)s/\rho + \sqrt{2sI_w(s)} \right\} \\ &= \sup_{s \geqslant 0} \left\{ -(1-\rho)s/\rho + \sqrt{2s(I_a(s) + c_s^2)} \right\} \end{aligned}$$

A Series of Queues

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

For a series of queues, the arrival process at each queue is exactly the departure from the previous queue.

Hence, extending to a series of queues simplifies to analyzing the IDC of the *departure process* of a single-server queue.

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 RQ does not yield the 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;
- BMAP/MAP/1 departure is a MAP with infinite order, see discussion in Green's dissertation (1999) and Zhang (2005).
    - 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 starting with empty system;
- Gamarnik and Zeevi (2006) and Budhiraja and Lee (2009), HT limit for **stationary** queue length process.

Historical Remarks on Departure Processes

Approximations

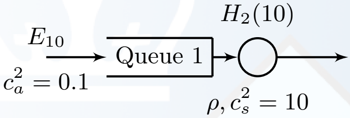
- Whitt (1982, 1983, 1984): QNA and related papers:
  - the **asymptotic method**: matching the long-run property of a point process

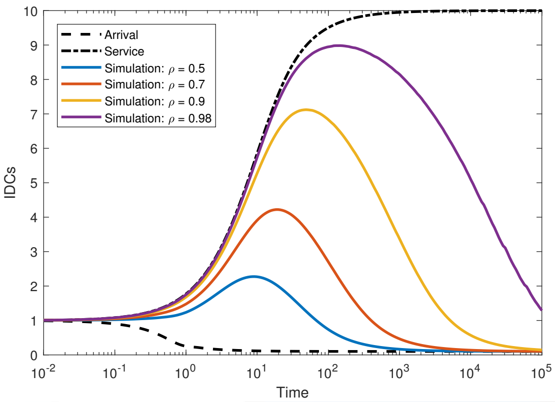
    $$c_d^2 \approx c_a^2$$

  - the **stationary interval method**: matching the stationary interval distribution, but ignore dependence between successive departures

    $$c_d^2 = c_a^2 + 2\rho^2 c_s^2 - 2\rho(1-\rho)E[W] \approx \rho^2 c_a^2 + (1-\rho^2)c_s^2$$

Departure IDC: A GI/GI/1 Example





Approximation for Departure IDC

- The numerical experiment suggests:

$$I_{d,\rho}(t) \approx w_\rho(t) I_a(t) + (1 - w_\rho(t)) I_s(t).$$

- To justify, we develop a heavy-traffic limit theorem for the weight function defined as

$$w_\rho(t) \equiv \frac{I_{d,\rho}(t) - I_s(t)}{I_a(t) - I_s(t)}.$$

- To this end, consider the HT-scaled weight function

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

  - classical HT-scaling: scale time by $(1-\rho)^{-2}$, scale space by $1-\rho$, but space scaling canceled out.

Main Theorem for Stationary Departure Processes

Theorem (HT limit for the weight function)

*For $GI/GI/1$ **stationary** departure process, under regularity conditions, we have*

$$w_\rho^*(t) \Rightarrow w^*(t/c_x^2),$$

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

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

for standard Normal cdf Φ and pdf ϕ.

- w^* is monotonically increasing and $0 \le w^* \le 1$;
- The limiting weight depend on interarrival and service distribution only through their scv's c_a^2 and c_s^2.

Approximation for Departure IDC

- Conjecture: the Theorem holds for a general class of G/G/1 models, which is supported by extensive simulation experiments.

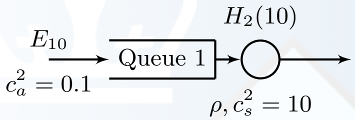
The theorem supports the following approximation

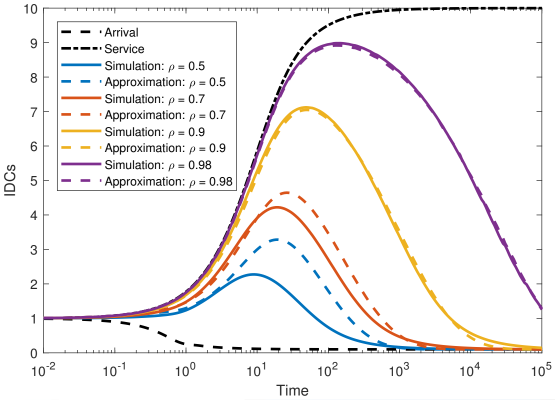
$$w_\rho(t) \approx w^*((1-\rho)^2 t/c_x^2),$$

and

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

The GI/GI/1 Example Revisited





RQ for a Series of Queues

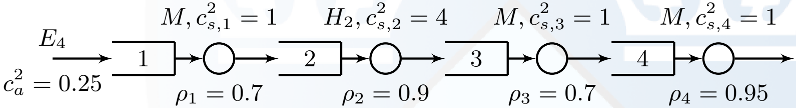
- $I_{a_1}(t)$: the IDC of the external arrival process to the first queue.
- $I_{s_i}(t)$: the IDC of the service process at queue i.
- For $i = 1, 2, \ldots, n$:
  - $c_{x,i}^2 = I_{a,i}(\infty) + I_{s_i}(\infty)$;
  - $\rho = 1/\mu_i$;
  - $w_i^*(t) = w^*((1-\rho_i)^2 t / c_{x,i}^2)$
  - $I_{a,i+1}(t) = I_{d,i}(t) = w_i^*(t) I_{a,i}(t) + (1 - w_i^*(t)) I_{s,i}(t)$
- Return $\{I_{a,i} : i = 1, 2, \ldots, n\}$

For any Queue i, apply the RQ algorithm

$$Z^* = \sup_{s \geqslant 0} \left\{ -(1-\rho)s/\rho + \sqrt{2s(I_{a,i}(s) + c_s^2)} \right\}$$

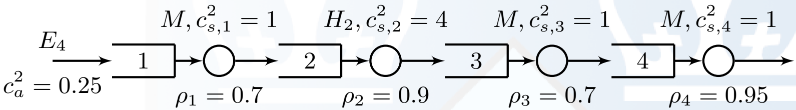
to produce approximation of the mean steady-state workload.

Numerical example: 4 Queues in Series



	Workload	RQ Approx.	Relative Error
Queue 1	1.09613	1.0583	-3.45%
Queue 2	17.6133	17.2884	-1.84%
Queue 3	2.89796	3.1702	9.39%
Queue 4	24.0131	23.5623	-1.18%
Total	45.6205	45.0792	-1.19%

Numerical example: 4 Queues in Series



By Brumelle's formula, we have

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

	Waiting Time	RQ Approx.	Relative Error
Queue 1	0.86584	0.8119	-6.23%
Queue 2	17.3204	16.9593	-2.08%
Queue 3	3.43984	3.8289	20.78%
Queue 4	24.3252	23.8524	-1.94%
Total	45.9513	45.4525	-1.09%

References

Key references:

- D. Daley, Queueing Output Processes, *Advances in Applied Probability*, 1976.
- D. Gamarnik, A. Zeevi, Validity of heavy traffic steady-state approximations in generalized Jackson Networks, *The Annals of Applied Probability*, 2006.
- W. Whitt, W. You, Heavy-traffic limit of the GI/GI/1 stationary departure process and its variance function, submitted to *Stochastic Systems*, 2017.

Queueing network approximations:

- J. G. Dai and J. M. Harrison, The QNET method for two-moment analysis of closed manufacturing systems, *Annals of Applied Probability*, 1993.
- J. G. Dai, V. Nguyen, and M. I. Reiman, Sequential bottleneck decomposition: an approximation method for generalized Jackson Networks, *Operations Research*, 1994.
- W. Whitt, The Queueing Network Analyzer, *Bell System Technical Journal*, 1983.

References

Other references:

- P. Burke, The Output of a Queuing System, *Operations Research*, 1956.
- D. Green, Departure Processes from MAP/PH/1 Queues, thesis, 1999.
- L. Takács, Introduction to the Theory of Queues, *Oxford University Press*, 1962.
- 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.
- D. Iglehart, W. Whitt, Multiple Channel Queues in Heavy Traffic II: Sequences, Networks, and Batches, *Advances in Applied Probability*, 1970.
- W. Whitt, Approximating a Point Process by a Renewal Process: Two Basic Methods, *Operations Research*, 1982.
- W. Whitt, Approximations for Departure Processes and Queues in Series, *Naval Research Logistics Quarterly*, 1984.
- Q. Zhang, A. Heindl, E. Smirni, Characterizing the BMAP/MAP/1 Departure Process via the ETAQA Truncation, *Stochastic Models*, 2005.

Thank you!

Dependent Service Times

$$
\begin{aligned}
I_w(t) &\equiv \frac{Var(Y(t))}{E[V]E[Y(t)]} \\
&= \frac{\mu^2}{\lambda t}\left(Var\left(E\left[\sum_{i=1}^{N(t)} V_i \middle| N(t)\right]\right) + E\left[Var\left(\sum_{i=1}^{N(t)} V_i \middle| N(t)\right)\right]\right) \\
&= \frac{\mu^2}{\lambda t}\left(\frac{1}{\mu^2} Var\left(N(t)\right) + E\left[\frac{1}{\mu^2} N(t) I^s_{N(t)}\right]\right) \\
&= I_a(t) + \frac{1}{\lambda t} E\left[N(t) I^s_{N(t)}\right],
\end{aligned}
$$

where

$$
I^s_k = \frac{kVar(S^s_k)}{(E[S^s_k])^2} = \frac{\mu^2}{k} Var(S^s_k)
$$

is the index of dispersion for intervals (IDI) for the service sequence and $Var(S^s_k) = \sum_{i=1}^{k} V_i$.

Corollary (Asymptotic behavior of the departure variance)

$$V_d^*(t) \sim c_a^2 \lambda t + \frac{(c_s^2 - c_a^2)c_x^2}{2\gamma^2} - \frac{8(c_s^2 - c_a^2)c_x^5}{\gamma^5} \frac{1}{\sqrt{2\pi\lambda^3 t^3}} e^{-\frac{\lambda\gamma^2 t}{2c_x^2}} \quad as\ t \to \infty.$$

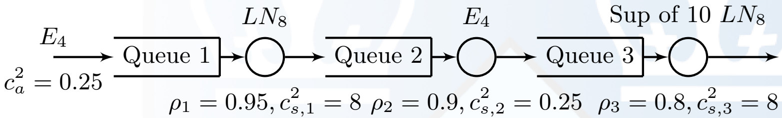
Compare to Hautphenne et al. (2013):

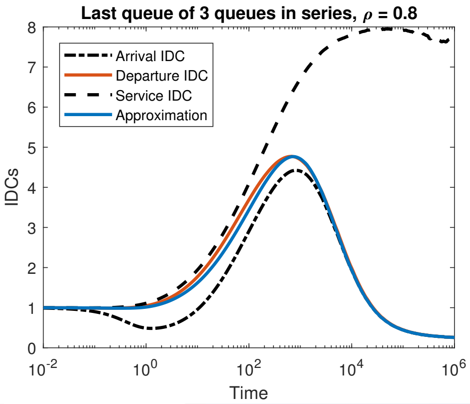
$$V_d(t) = c_a^2 t + b_\theta + o(1), \text{ as } t \to \infty.$$

- they have explicit expression for b_θ under all ρ in M/G/1;
- our have more detailed remainder for GI/GI/1 as $\rho \uparrow 1$;
- the two coincide as $\rho \uparrow 1$ in M/G/1.

Of course, our limit holds for all t, not just asymptotically.

An Artificial Example

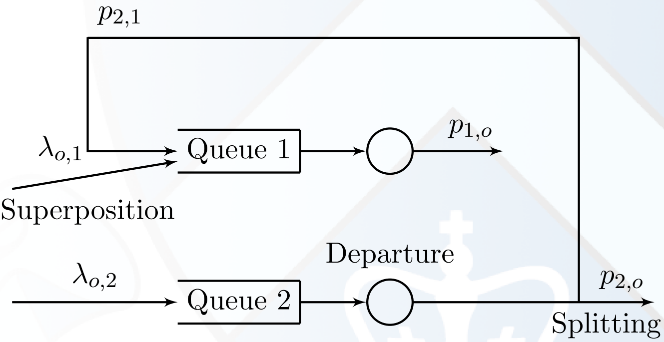




A Path to RQNA

The total arrival process at any queue:

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



Three Network Operators

In summary,

- *Splitting* under Markovian routing:

$$I_{a,j,i}(t) = p_{j,i} I_{d,j}(t) + (1 - p_{j,i}), \quad \text{for} \quad j \geq 1$$

- *Superposition* of independent streams:

$$I_{a,i}(t) = \sum_{i=0}^{k} \frac{\lambda_{j,i}}{\lambda_i} I_{a,j,i}(\lambda_{j,i} t).$$

  - adds nonlinearity
- *Departure* IDC

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

The RQNA Algorithm

- Traffic-rate equations

$$\lambda_i = \lambda_{o,i} + \sum_{j=1}^{n} \lambda_{j,i} = \lambda_{o,i} + \sum_{j=1}^{n} \lambda_j p_{j,i},$$

- Total-arrival-IDC equations

$$I_{a,i}(t) = \frac{\lambda_{o,i}}{\lambda_i} I_{a,o,i}(\lambda_{o,i} t) + \sum_{j=1}^{n} \frac{\lambda_{j,i}}{\lambda_i} \left(p_{j,i} I_{d,j}(\lambda_{j,i} t) + (1 - p_{j,i})\right)$$

The RQNA Algorithm

$$I_{a,i}(t) = \frac{\lambda_{o,i}}{\lambda_i} I_{a,o,i}(\lambda_{o,i} t) + \sum_{j=1}^{n} \frac{\lambda_{j,i}}{\lambda_i} \left(p_{j,i} I_{d,j}(\lambda_{j,i} t) + (1 - p_{j,i})\right)$$

- Departure IDC, define $\rho_i = \lambda_i / \mu_i$ and $c_{x,i}^2 = c_{a,i}^2 + c_{s,i}^2$, then

$$I_{d,i}(t) = w^*((1-\rho_i)^2 t / c_{x,i}^2) I_{a,i}(t) + (1 - w^*((1-\rho_i)^2 t / c_{x,i}^2)) I_{s,i}(t),$$

- Asymptotic-variability-parameter equations

$$c_{a,i}^2 = \frac{\lambda_{o,i}}{\lambda_i} c_{a,o,i}^2 + \sum_{j=1}^{n} \frac{\lambda_{j,i}}{\lambda_i} \left(p_{j,i} c_{a,j}^2 + (1 - p_{j,i})\right)$$

  - obtained by letting $t \to \infty$ in the total-arrival-IDC equations.
  - coincides with (24) in Whitt (1983), where we take $w_j = 1$ and $v_{ij} = 1$ there.

Solving the Total-Arrival-IDC equations

- Both the traffic-rate equations and asymptotic-variability equations are linear equations.
- Total-arrival-IDC equations
  - nonlinear due to the superposition operator;
  - simpler case: **feed-forward queueing network**, can be solved explicitly by iteration;
  - general case: forms a contraction mapping, so unique solution can be found by fixed-point-iteration method.

Extension to GI/GI/1 model

Proof sketch. From the HT limit

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

plus u.i. condition,

$$V_d^*(t) = \mathrm{Var}(c_a B_a(t)) + \mathrm{Var}(Q^*(0)) + \mathrm{Var}(Q^*(t)) + \mathrm{cov}(Q^*(0), Q^*(t)) + \mathrm{cov}(c_a B_a(t), Q^*(t)),$$

- $\mathrm{Var}(c_a B_a(t)) = c_a^2 t$;
- $\mathrm{Var}(Q^*(t)) = \mathrm{Var}(Q^*(0)) = c_x^4/4$;
- $\mathrm{cov}(Q^*(0), Q^*(t)) = \frac{c_x^4}{4} c^*(t/c_x^2)$, where c^* is the correlation function discussed in Abate and Whitt (1987,1988).
    - w^* is closely related to c^*

$$w^*(t) = 1 - \frac{1 - c^*(t)}{2t}.$$

HT limit theorem for GI/GI/1 departure variance

Proof sketch contd. The remaining term

$$\mathrm{cov}(c_a B_a(t), Q^*(t)).$$

is treated by scaling techniques. Recall that

$$Q^*(t) = \psi(Q^*(0) + c_a B_a - c_s B_s - e)$$

- Scale the original system so that we have a modified system with the same drift -1 but $\tilde{c}_a^2 = 1$.

$$\begin{aligned} &\{Q^*(0), c_a B_a(t), c_s B_s(t), -t\} \\ &\stackrel{\mathrm{d}}{=} c_a^2 \left\{ \frac{Q^*(0)}{c_a^2}, B_a(t/c_a^2), \frac{c_s}{c_a} B_s(t/c_a^2), -\frac{t}{c_a^2} \right\} \\ &\equiv c_a^2 \left\{ \frac{Q^*(0)}{c_a^2}, B_a(u), \frac{c_s}{c_a} B_s(u), -u \right\}, \end{aligned}$$

  where $u = t/c_a^2$.
- Apply results for special case $M/GI/1$ where $c_a^2 = 1$.

The Heavy-traffic Bottleneck Phenomenon

Table: The heavy-traffic bottleneck example

		High variability	Low variability
Queue 9	Simulation	29.1480 ± 0.0486	5.2683 ± 0.0025
	QNA	8.9 (-69.47%)	8.0 (51.85%)
	M/M/1	8.1 (-72.21%)	8.1 (53.75%)
	Asymp. Method	36.5 (25.22%)	4.05 (-23.13%)
	RQNA	26.88 (-7.79%)	5.44 (3.26%)
	RQ	36.98 (26.86%)	4.9509 (-6.02%)
Queue 8	Simulation	1.4403 ± 0.0005	0.7716 ± 0.0001
	QNA	1.04 (-27.79%)	0.88 (14.05%)
	M/M/1	0.9 (-37.51%)	0.9 (16.64%)
	Asymp. Method	4.05 (181.19%)	0.45 (424.88%)
	RQNA	0.9 (-37.51%)	0.895 (15.99%)
	RQ	1.267 (-12.03%)	0.853 (10.51%)