Robust Queueing for Open Queueing Networks

Wei You
(with Ward Whitt)
IEOR, Columbia University
IBM Research Center
November 1, 2017
The estimation of performance measures in an open network of queues is important in many OR applications.

Theoretical analysis are limited for queueing networks 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.
Traditionally, queueing systems are approximated by

- Parametric-decomposition methods using variability parameters: e.g., QNA by Whitt (1983);
  - QNA is widely accepted, but is known to fail in certain cases, see Suresh and Whitt (1990).
  - It relies on the approximation of the variability parameters for arrival, service and departures.
- Reflected Brownian motion approximations: e.g., QNET by Dai and Harrison (1993);
  - QNET algorithm computation time scales with the system.
  - Sequential Bottleneck Decomposition by Dai, Nguyen and Reiman (1994) proposed to relief the computation burden.
Review of Robust Queueing Theory

More recently,

- Robust Queueing (RQ) by Bandi et al. (2015) analyzed the mean steady-state waiting time in a queueing network.

We followed the RQ framework and developed

- RQ for the workload process in G/G/1 models;
- approximation of stationary departure processes, which leads to RQ for queues in series.
- RQ for $G_t/G_t/1$ models;
A Robust Queueing Theory proposed by Bandi et al. (2015)

- analyzed the mean steady-state waiting time in single server queue with general interarrival and service distributions;
- replaced probabilistic laws by uncertainty sets;
- used robust optimization and regression analysis.
- proposed an extension to feed-forward open queueing networks with adversary servers;
Review of Robust Queueing Theory

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

- $\{(U_i, V_i)\}_{i \geq 1}$: interarrival times and service times;
- $\lambda, \mu$: arrival rate and service rate.

Lindley recursion

$$W_n = (W_{n-1} + V_{n-1} - U_{n-1})^+ = \max_{0 \leq k \leq n} \{S^s_k - S^a_k\},$$

where $S^s_0 \equiv 0$, $S^a_0 \equiv 0$ and

$$S^s_k \equiv \sum_{i=n-k}^{n-1} V_i, \quad S^a_k := \sum_{i=n-k}^{n-1} U_i, \quad 1 \leq k \leq n.$$

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

$$W_n^* \equiv \sup_{U \in U^a} \sup_{V \in U^s} W_n(U, V) = \sup_{U \in U^a} \sup_{V \in U^s} \max_{0 \leq k \leq n} \{S_s^k - S_{k}^a\}$$

Motivated by CLT, Bandi et al. proposed

$$U^a = \left\{ (U_1, \ldots, U_n) \left| \frac{S_{k}^a - k/\lambda}{k^{1/2}} \geq -\Gamma_a, 0 \leq k \leq n \right. \right\},$$

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

- 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

\[ W_n^* = \max_{0 \leq k \leq n} \{ mk + b\sqrt{k} \} \]

\[ \leq \sup_{x \geq 0} \{ mx + b\sqrt{x} \} = \frac{b^2}{4|m|} = \frac{\lambda b^2}{4(1 - \rho)}, \]

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 \( \rho, \Gamma_a \) and \( \Gamma_s. \)
- The solution resembles classical heavy-traffic limit approximations or bounds, e.g., Kingman Bound

\[ W_{\rho}^* \leq \frac{\rho(\rho^{-2}c_a^2 + c_s^2)}{2\mu(1 - \rho)}. \]
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 $\Gamma_a$ and $\Gamma_s$ for external arrival processes and service processes;
- used similar network calculus as in QNA to determine parameters $\Gamma_a$ and $\Gamma_s$;
Dependence in Queues

- Dependence rises naturally in queueing network:
  - **departure** process is non-renewal beyond M/M/1 case;
  - **splitting** creates dependent flows;
  - **superposition** of different arrival streams is non-renewal unless all processes are Poisson.

- Dependence has significant impact on performance measures
  - see discussion in Section 1B of Fendick and Whitt (1989);
  - the level of impact will depend on the traffic intensity;
    - As a result, methods (QNA, RQ by Bandi et al.) using a single parameter to describe variability may fail.
An Example

Last queue of 5 queues in series (tandem queues)

$E_{10}$ $H_2(10), \rho_1 = 0.99$

$\lambda = 1$

$H_2(10), \rho_3 = 0.7$

$E_{10}, \rho_2 = 0.98$

$E_{10}, \rho_4 = 0.5$

$M, \rho$

- log$_{10}(1-\rho)$

Normalized mean workload

Traffic intensity

Normalized workload
The Heavy-traffic Bottleneck Phenomenon

\[ D \text{ or } H_2(4) \]
\[ \lambda = 1 \]
\[ M, \rho_1 = 0.6 \]
\[ 1 \rightarrow \cdots \rightarrow 8 \rightarrow 9 \]
\[ M, \rho_1 = 0.6 \]

Table: The heavy-traffic bottleneck example

<table>
<thead>
<tr>
<th>Queue 9</th>
<th>Simulation</th>
<th>( H_2, c_a^2 = 4 )</th>
<th>( D, c_a^2 = 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>M/M/1</td>
<td>29.1480 ± 0.0486</td>
<td>5.2683 ± 0.0025</td>
<td></td>
</tr>
<tr>
<td>QNA</td>
<td>8.9 (-69.47%)</td>
<td>8.0 (51.85%)</td>
<td></td>
</tr>
<tr>
<td>RQ</td>
<td>36.98 (26.86%)</td>
<td>4.9509 (-6.02%)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Queue 8</th>
<th>Simulation</th>
<th>( H_2, c_a^2 = 4 )</th>
<th>( D, c_a^2 = 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>M/M/1</td>
<td>1.4403 ± 0.0005</td>
<td>0.7716 ± 0.0001</td>
<td></td>
</tr>
<tr>
<td>QNA</td>
<td>1.04 (-27.79%)</td>
<td>0.88 (14.05%)</td>
<td></td>
</tr>
<tr>
<td>RQ</td>
<td>1.267 (-12.03%)</td>
<td>0.853 (10.51%)</td>
<td></td>
</tr>
</tbody>
</table>
Our Motivation

We want to build new RQNA algorithm

▸ with improved performance in single-server queues:
  ▸ capture dependence in the $G/G/1$ models;
  ▸ obtain correct heavy-traffic and light-traffic limits;
  ▸ provide useful approximations across all traffic intensities;

▸ to fit most open queuing networks:
  ▸ go beyond feed-forward networks;
  ▸ analyze traditional servers, as oppose to adversary servers;
  ▸ go beyond Markovian routing (work in progress);

▸ that run fast and effective.
Continuous-time workload process

- \(\{(U_i, V_i)\}\): interarrival times and service times;
- \(\lambda, \mu\): arrival rate and service rate;
- \(A(t)\): arrival counting process associated with \(\{U_k\}\);
- \(Y(t)\): total input of work defined by \(Y(t) \equiv \sum_{k=1}^{A(t)} V_k\);
- \(X(t)\): net-input process defined by \(X(t) \equiv Y(t) - t\);

The steady-state workload at time 0 in the queue starting empty at the remote past \(-\infty\):

\[
Z \equiv X(0) - \inf_{-\infty \leq t \leq 0} \{X(t)\}.
\]

\[
= \sup_{0 \leq s \leq \infty} \{X(0) - X(-s)\} \equiv \sup_{0 \leq s \leq \infty} \{X_0(s)\}
\]

- \(X_0(s)\): the net-input over time \([-s, 0]\).
- With an abuse of notation, we omit the subscript in \(X_0(s)\).
Continuous-time workload process

We now insert the traffic intensity $\rho$ into the model.

- Start with unit-rate arrival counting process $A(t)$ and mean-1 service times;
- Assume that $A_\rho(t)$ with rate $\rho$ in the $\rho$-th model satisfies:
  \[ A_\rho(t) = A(\rho t). \]
- The total input process and net-input process:
  \[ Y_\rho(t) = Y(\rho t), \text{ and } X_\rho(t) = Y(\rho t) - t. \]
- The steady-state workload:
  \[ Z_\rho = \sup_{0 \leq s \leq \infty} \{Y_\rho(s) - s\} = \sup_{0 \leq s \leq \infty} \{X_\rho(s)\}. \]
Stochastic versus Robust Queues

\[ Z_\rho = \sup_{0 \leq s \leq \infty} \{ X_\rho(s) \}. \]

**Stochastic Queue**

- \[ X_\rho(s) \equiv \sum_{k=1}^{N(rho \cdot s)} V_k - s, \]
  where \( N(t) \) and \( \{ V_k \} \) are stationary point process and stationary sequence separately.

**Robust Queue**

- \( \tilde{X}_\rho \) lies in a suitable uncertainty set \( U_\rho \) of total input functions to be defined later.
- There is no distribution involved, we hence focus on the deterministic worse-case scenario

\[ Z^*_\rho \equiv \sup_{\tilde{X}_\rho \in U_\rho} \sup_{0 \leq s \leq \infty} \{ \tilde{X}_\rho(s) \}. \]
Robust Queueing for continuous-time workload

Now, we define the uncertainty set for the net-input process.

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

where

\[ E[X_\rho(s)] = -(1 - \rho)s, \]

\[ \text{Var}(X_\rho(s)) = \text{Var}(X_\rho(s) - s) = \text{Var}(Y_\rho(s)) = \text{Var}(Y(\rho s)) \]

and \( I_w(t) \) is the index of dispersion for work (IDW) for the base net-input process \( Y(t) \), i.e.,

\[ I_w(t) \equiv \frac{\text{Var}(Y(t))}{t}. \]
Robust Queueing for continuous-time workload

RQ for workload

\[ Z^*_\rho = \sup_{X_\rho \in U_\rho} \sup_{0 \leq s \leq \infty} \{ X_\rho(s) \}, \]

where

\[ U_\rho = \left\{ X_\rho : \mathbb{R} \rightarrow \mathbb{R} \mid X_\rho(s) \leq -(1 - \rho)s + b\sqrt{\rho s I_w(\rho s)} \right\}. \]

Lemma (Dimension reduction)

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

\[ Z^*_\rho = \sup_{0 \leq s \leq \infty} \sup_{X_\rho \in U_\rho} \{ X_\rho(s) \} = \sup_{0 \leq s \leq \infty} \left\{ -(1 - \rho)s + b\sqrt{\rho s I_w(\rho s)} \right\}. \]

Furthermore, if \( \rho < 1 \) and \( I_w(t)/t \rightarrow 0 \) as \( t \rightarrow \infty \), then \( Z^*_\rho < \infty \).
In summary, the RQ algorithm for single-server queues

\[ Z^*_\rho = \sup_{0 \leq s \leq \infty} \left\{ -(1 - \rho)s + b\sqrt{\rho s I_w(\rho s)} \right\}. \]

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 approximated;
- same \( I_w \) used for all \( \rho \in [0, 1) \);
- enables convenient generalization to queueing networks.
Remarks on the RQ algorithm

\[ Z^*_\rho = \sup_{s \geq 0} \left\{ - (1 - \rho) s + b \sqrt{\rho s I_w(\rho s)} \right\}. \]

- Choose \( b = \sqrt{2} \) so that RQ is exact for \( M/GI/1 \) models.
- Slightly more general version, for \( \rho = \lambda/\mu \)

\[ Z^*(\lambda, \mu, I_w) = \sup_{s \geq 0} \left\{ - (1 - \rho) s/\rho + \sqrt{2sI_w(\mu s)/\mu} \right\} \]

**Theorem (RQ correct in Heavy-traffic and light-traffic)**

*Under regularity assumptions, the RQ algorithm with \( b = \sqrt{2} \) yields the exact mean steady-state workload in both light-traffic and heavy-traffic limits for \( G/G/1 \) models.*
Numerical Examples

Numerical Example: 5 queues in series

- $H_2(10), \rho = 0.99$
- $H_2(10), \rho = 0.7$
- $E_{10}, \rho = 0.98$
- $E_{10}, \rho = 0.5$
- $M$

Normalized mean workload

- $\log_{10}(1-\rho)$

IDW
Numerical Examples - 5 Queues in series

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

- $|\text{RE}| = |\text{RE}_\rho|$: relative error (as a function of traffic intensity) between the RQ approximation and the simulation estimation;
- $\text{max}(|\text{RE}|)$: for fixed example, the maximum relative error across different traffic intensities;
- $\text{avg}(|\text{RE}|)$: 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 $\text{max}(|\text{RE}|)$ for RQ = 33.01%. Mean $\text{max}(|\text{RE}|)$ for RQ = 16.85%.
* Max $\text{avg}(|\text{RE}|)$ for RQ = 15.47%. Mean $\text{avg}(|\text{RE}|)$ for RQ = 7.50%.
--- End ---

--- rho = 0.9 ---
* Max $\text{max}(|\text{RE}|)$ for RQ = 37.36%. Mean $\text{max}(|\text{RE}|)$ for RQ = 17.66%.
* Max $\text{avg}(|\text{RE}|)$ for RQ = 11.69%. Mean $\text{avg}(|\text{RE}|)$ for RQ = 6.52%.
--- End ---
Generalization to RQNA

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

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 \( A(t) \)

\[ I_a(t) = \frac{Var(A(t))}{E[A(t)]}. \]
To extend the RQ algorithm, we need to

- (for **external** arrival processes) provide effective algorithm to calculate/estimate the IDC of a stationary point process;
- (for **internal** arrival streams) produce effective approximations internal arrival IDC at any queue within a open queueing network;
To calculate/estimate the IDC of a stationary point process,

- let $A(t)$ be a base process with rate 1 and

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

where the variance is taken under stationary distribution.
- for stationary point process, we have $E[A(t)] = t$;
Generalization to RQNA: External Arrival Process

- estimate via numerical inversion:
  \[
  \hat{V}(s) = \frac{\lambda}{s^2} + \frac{2\lambda}{s} \hat{m}(s) - \frac{2\lambda^2}{s^3},
  \]
  \[
  V(t) = \lambda \int_0^t (1 + 2m(u) - 2\lambda u) du.
  \]

- \( m(t) = E^0[A(t)] \) under Palm distribution \( P^0 \), i.e., conditioning on having an arrival at time 0.

- renewal function in the case of renewal processes, let \( \hat{f}(s) = \int_0^\infty e^{-st} dF(t) \), then
  \[
  \hat{m}(s) = \frac{\hat{f}(s)}{s(1 - \hat{f}(s))}
  \]

- 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 **splittings** of departure processes.
Splitting and Superposition

- **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

- **Splitting** under Markovian routing:

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

- The remaining challenge is to characterize *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;
- 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** queueing length process.
Historical Remarks on Departure Processes

Approximations

  - **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
    \[ c_d^2 = c_a^2 + 2\rho^2c_s^2 - 2\rho(1 - \rho)E[W] \approx \rho^2 c_a^2 + (1 - \rho^2)c_s^2 \]
A numerical example

$E_{10}$

Queue 1

$H_2(10)$

$c_a^2 = 0.1$

$\rho, c_s^2 = 10$

Simulation: $\rho = 0.5$
Simulation: $\rho = 0.7$
Simulation: $\rho = 0.9$
Simulation: $\rho = 0.98$
Our approach

- Start with the Laplace transform for M/G/1 and GI/M/1 models in Daley (1976);
- proves HT limits for M/G/1 and GI/M/1 special cases;
- convert general GI/GI/1 to M/G/1 or GI/M/1 special cases using space-time scaling;
- obtain from the HT limit an approximation for departure IDCs in the form of convex combination.
Let $D(t)$ be the stationary departure process with finite variance, let $V_d(t) = Var(D(t))$, then

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

$$V_d(t) = \lambda \int_0^t (1 + 2m_d(u) - 2\lambda u)du.$$}

where $m_d(t) = E^0[D(t)]$ is the mean process under Palm distribution $P^0$, i.e., conditioning on having an arrival at time 0.
Laplace Transform of the Variance Function

Takacs (1962): For M/GI/1

\[ \hat{m}_d(s) \equiv \int_0^\infty e^{-st} m_d(t) dt = \frac{\hat{g}(s)}{s(1 - \hat{g}(s))} \left( 1 - \frac{s\Pi(\hat{\nu}(s))}{s + \lambda(1 - \hat{\nu}(s))} \right), \]

- \( \hat{g}(s) = E[e^{-sV}] \) is the LT of the service pdf \( g(t) \);
- \( \hat{\nu}(s) \) is the root with the smallest absolute value in \( z \) of the equation
  \[ z = \hat{g}(s + \lambda(1 - z)) \]
- \( \Pi(z) \) is the probability generating function of the distribution of the stationary queue length \( Q \)
  \[ \Pi(z) \equiv E[z^Q] = \frac{(1 - \lambda/\mu)(1 - z)\hat{g}(\lambda(1 - z))}{\hat{g}(\lambda(1 - z)) - z}. \]
Laplace Transform of the Variance Function

Daley (1976): For GI/M/1

\[ \hat{V}_d(s) = \frac{\lambda}{s^2} + \frac{2\lambda}{s^3} \left( \mu \delta - \lambda + \frac{\mu^2(1 - \delta)(1 - \hat{\xi}(s))(\mu \delta(1 - \hat{f}(s)) - s\hat{f}(s))}{(s + \mu(1 - \hat{\xi}(s)))(s - \mu(1 - \delta))(1 - \hat{f}(s))} \right), \]

- \( \lambda \) is the arrival rate,
- \( \mu \) is the service rate (with \( \lambda < \mu \));
- \( \hat{f}(s) = E \left[ e^{-sU} \right] \) is the LT of the interarrival-time pdf \( f(t) \);
- \( \hat{\xi}(s) \) is the root with the smallest absolute value in \( z \) of the equation
  \[ z = \hat{f}(s + \mu(1 - z)) \]
- \( \delta = \hat{\xi}(0) \) is the unique root in \((0, 1)\) of the equation
  \[ \delta = \hat{f}(\mu(1 - \delta)). \]
The Heavy-Traffic Scaling

Formula for both M/GI/1 and GI/M/1 are complicated

- We resort to proving a heavy traffic limit theorem.
- A family of models indexed by $\rho$
  - M/GI/1: $(\lambda, \mu) = (\rho, 1)$;
  - GI/M/1: $(\lambda, \mu) = (1, \rho^{-1})$;
- simplify by fixing the GI distribution;
- both can be easily generalized for non-unit rates.
The Heavy-Traffic Scaling

To obtain a proper heavy-traffic limit, we define

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

- classical HT-scaling from Iglehart and Whitt (1970)
  - scale time by \((1 - \rho)^{-2}\), scale space by \(1 - \rho\);
- corresponding variance function:

\[ V^*_{d,\rho}(t) \equiv (1 - \rho)^2V_{d,\rho}((1 - \rho)^{-2}t) \]

and LT

\[ \hat{V}^*_{d,\rho}(s) \equiv (1 - \rho)^4\hat{V}_{d,\rho}((1 - \rho)^2s) \]

- prove the limit for the LT and then use continuity results for the LT.
The Heavy-Traffic Limit

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

Under regularity conditions, $V^*_d,\rho$ converges to

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

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

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

and $\phi, \Phi$ are the standard normal pdf and cdf, respectively.
Extension to GI/GI/1 model

The HT limit theorem for departure variance extend naturally to the GI/GI/1 model, yielding exactly the same result.

Regularity conditions

- the interarrival-time cdf has a pdf;
- the interarrival times and service times have uniformly bounded third moments.
Extension to GI/GI/1 model

To start, we state the HT limit theorem for the departure process

Theorem (HT limit for the stationary departure process)

Under assumptions on the last slide,

\[ D^*(t) = c_a B_a(t) + Q^*(0) - Q^*(t). \]

- \( B_a \) 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.
Extension to GI/GI/1 model

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

Under assumptions in Theorem plus uniform integrability conditions, \( V_{d, \rho}^* \) converges to

\[
V_{d}^*(t) \equiv w^* \left( t/c_x^2 \right) c_a^2 \lambda t + (1 - w^* \left( t/c_x^2 \right)) c_s^2 \lambda t
\]

where \( c_x^2 = c_a^2 + c_s^2 \),

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

and \( \phi, \Phi \) are the standard normal pdf and cdf, respectively.

- Proof sketch at the end of the slides.
Let $I_{d,\rho}$ be the departure IDC in the model with traffic intensity $\rho$. 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

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

- Same HT scaling as before, but space scaling canceled out.
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

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

and

\[
I_{d,\rho}(t) = w_\rho(t)I_a(t) + (1 - w_\rho(t))I_s(t)
\approx w^*((1 - \rho)^2t/c_x^2)I_a(t) + (1 - w^*((1 - \rho)^2t/c_x^2))I_s(t).
\]
A Simple Example

\[
\begin{align*}
E_{10} & \quad \text{Queue 1} \\
\rho, c_s^2 = 10 & \quad H_2(10)
\end{align*}
\]

\[
c_a = 0.1
\]
An Artificial Example

$
\begin{align*}
E_4 & \rightarrow \text{Queue 1} \rightarrow LN_8 \rightarrow \text{Queue 2} \rightarrow E_4 \\
E_4 & \rightarrow \text{Sup of 10 LN}_8
\end{align*}$

$E_4$

$c_a^2 = 0.25$

$
\begin{align*}
\rho_1 &= 0.95, c_{s,1}^2 = 8 \\
\rho_2 &= 0.9, c_{s,2}^2 = 0.25 \\
\rho_3 &= 0.8, c_{s,3}^2 = 8
\end{align*}$

$\text{Last queue of 3 queues in series, } \rho = 0.8$
Three Network Operators

In summary,

- **Splitting** under Markovian routing:
  \[
  I_{a,j,i}(t) = p_{j,i} I_{d,j}(t) + (1 - p_{j,i}), \quad \text{for } j \geq 1
  \]

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

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

- 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-equation equations
\[ c_{a,i}^2 = \frac{\lambda_{o,i}}{\lambda_i} c_{a,o,i}^2 + \sum_{j=1}^{n} \frac{\lambda_{j,i}}{\lambda_i} (p_{j,i} c_{a,j}^2 + (1 - p_{j,i})) \]

- 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.
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 $\rho$ 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 RQNA at the test queue and compare it with RQ.
The case

* 4 identical queues in series, traffic intensity 0.70.
* Arrival distribution picked from: E4, LN025, LN4, H4.
* Service distribution picked from: E4, LN025, LN4, H4, M.
* Number of cases in total: 20.

Summary

* Max max(|RE|) for RQNA = 31.90%. Mean max(|RE|) for RQNA = 17.38%.
* Max max(|RE|) for RQ = 33.01%. Mean max(|RE|) for RQ = 16.85%.
* Max avg(|RE|) for RQNA = 21.34%. Mean avg(|RE|) for RQNA = 9.52%.
* Max avg(|RE|) for RQ = 15.47%. Mean avg(|RE|) for RQ = 7.50%.
* Min avg(|RE|) for RQNA = 0.95%. Min avg(|RE|) for RQ = 1.58%.

Compare to RQ

* Max increase of avg(|RE|) over RQ = 229.29%.
  In this case, avg(|RE|) for RQNA is 5.20%.
* Max decrease of avg(|RE|) over RQ = 72.10%.
* RQNA outperforms RQ in 8 out of 20 cases in terms of max(|RE|).
* RQNA outperforms RQ in 6 out of 20 cases in terms of avg(|RE|).
The RQNA Algorithm

Numerical Examples Revisited

The case

* 4 identical queues in series, traffic intensity 0.90.
* Arrival distribution picked from: E4, LN025, LN4, H4.
* Service distribution picked from: E4, LN025, LN4, H4, M.
* Number of cases in total: 20.

Summary

* Max max(|RE|) for RQNA = 30.00%. Mean max(|RE|) for RQNA = 12.57%.
* Max max(|RE|) for RQ = 37.36%. Mean max(|RE|) for RQ = 17.66%.
* Max avg(|RE|) for RQNA = 10.56%. Mean avg(|RE|) for RQNA = 4.40%.
* Max avg(|RE|) for RQ = 11.69%. Mean avg(|RE|) for RQ = 6.52%.
* Min avg(|RE|) for RQNA = 2.43%. Min avg(|RE|) for RQ = 1.25%.

Compare to RQ

* Max increase of avg(|RE|) over RQ = 117.58%.
  In this case, avg(|RE|) for RQNA is 2.76%.
* Max decrease of avg(|RE|) over RQ = 75.33%.
* RQNA outperforms RQ in 12 out of 20 cases in terms of max(|RE|).
* RQNA outperforms RQ in 13 out of 20 cases in terms of avg(|RE|).

End
References

- **Key references:**

- **References on queueing networks:**
Other references:


References

Key references:


Other references:

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,

\[ V_d^*(t) = \text{Var}(c_a B_a(t)) + \text{Var}(Q^*(0)) + \text{Var}(Q^*(t)) \]

\[ + \text{cov}(Q^*(0), Q^*(t)) + \text{cov}(c_a B_a(t), Q^*(t)), \]

\[ \text{Var}(c_a B_a(t)) = c_a^2 t; \]

\[ \text{Var}(Q^*(t)) = \text{Var}(Q^*(0)) = c_x^4 / 4; \]

\[ \text{cov}(Q^*(0), Q^*(t)) = \frac{c^4_x}{4} c^*(t/c_x^2), \text{where } c^* \text{ is the correlation function discussed in Abate and Whitt (1987,1988).} \]

\[ w^*(t) = 1 - \frac{1 - c^*(t)}{2t}. \]
HT limit theorem for GI/GI/1 departure variance

**Proof sketch contd.** The remaining term

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

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 $c^2_a = 1$.

$$\{Q^*(0), c_a B_a(t), c_s B_s(t), -t\} \quad \text{d} \equiv c^2_a \left\{ \frac{Q^*(0)}{c^2_a}, B_a(t/c^2_a), \frac{c_s}{c_a} B_s(t/c^2_a), -\frac{t}{c^2_a} \right\}$$

$$\equiv c^2_a \left\{ \frac{Q^*(0)}{c^2_a}, B_a(u), \frac{c_s}{c_a} B_s(u), -u \right\},$$

where $u = t/c^2_a$.

- Apply results for special case $M/GI/1$ where $c^2_a = 1$. 