Robust Queueing

Wei You (joint work with Ward Whitt)

Seminar, Chinese Academy of Sciences

Dec. 18, 2019

Dependence in Open Queueing Networks

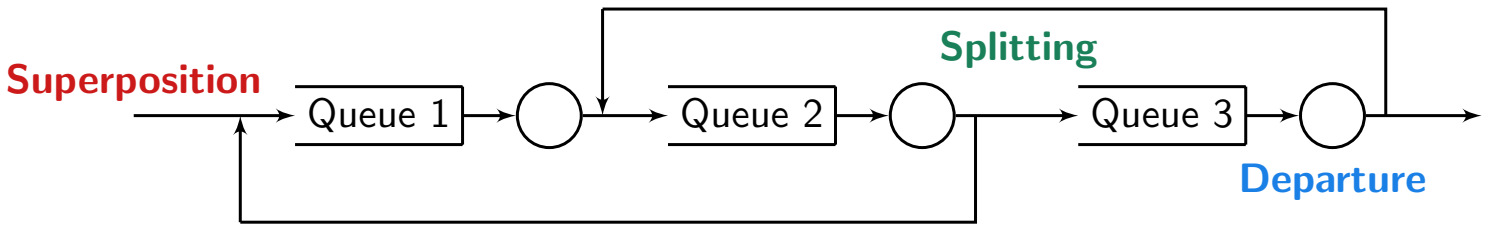


Figure: A three-station example.

Even **generalized Jackson networks** can be complicated

- Arrival process: the **superposition** of independent renewal process cannot be renewal unless all components are Poisson processes.
- **Departure** process cannot be renewal with non-Poisson arrival process or non-Exponential service-time distribution.
- Dependence among customer flows can be introduced by **splitting**.
- **Customer feedback** introduces dependence between arrival and service processes.

Dependence in Open Queueing Networks

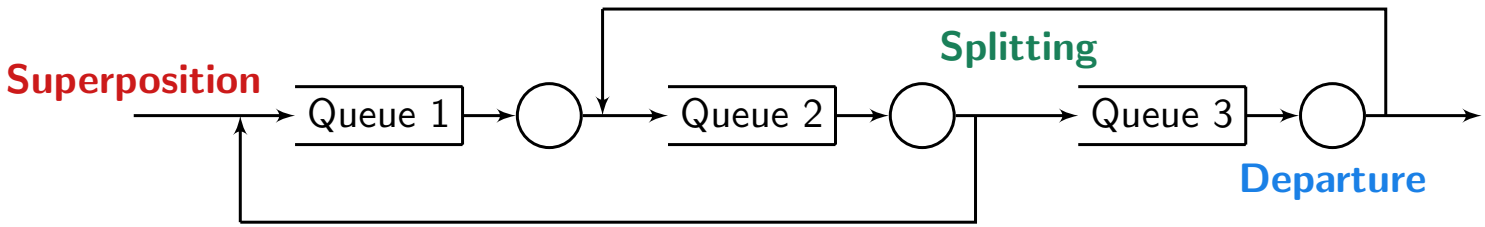


Figure: A three-station example.

Even **generalized Jackson networks** can be complicated

- Arrival process: the **superposition** of independent renewal process cannot be renewal unless all components are Poisson processes.
- **Departure** process cannot be renewal with non-Poisson arrival process or non-Exponential service-time distribution.
- Dependence among customer flows can be introduced by **splitting**.
- **Customer feedback** introduces dependence between arrival and service processes.

Dependence in Open Queueing Networks

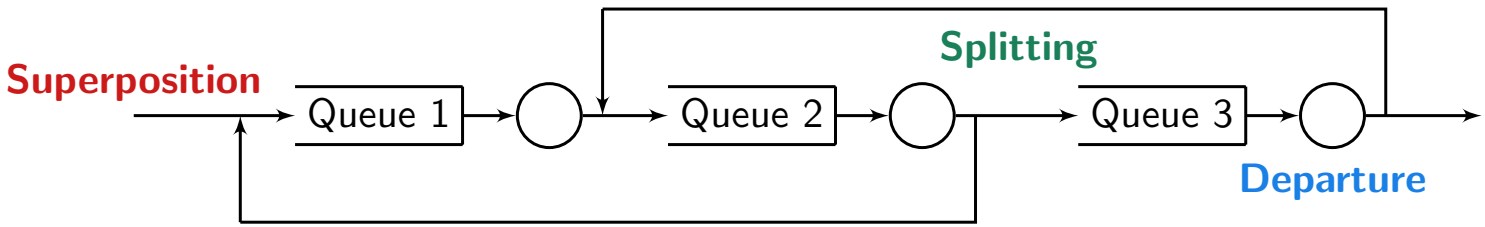


Figure: A three-station example.

Even **generalized Jackson networks** can be complicated

- Arrival process: the **superposition** of independent renewal process cannot be renewal unless all components are Poisson processes.
- **Departure** process cannot be renewal with non-Poisson arrival process or non-Exponential service-time distribution.
- Dependence among customer flows can be introduced by **splitting**.
- **Customer feedback** introduces dependence between arrival and service processes.

Dependence in Open Queueing Networks

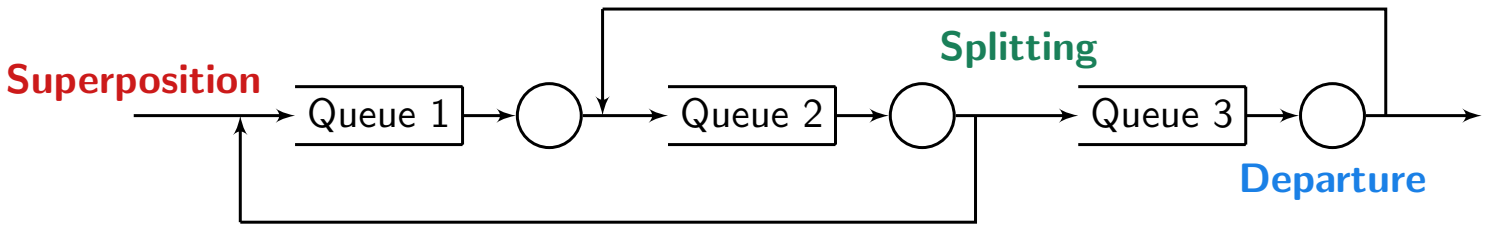


Figure: A three-station example.

Even **generalized Jackson networks** can be complicated

- Arrival process: the **superposition** of independent renewal process cannot be renewal unless all components are Poisson processes.
- **Departure** process cannot be renewal with non-Poisson arrival process or non-Exponential service-time distribution.
- Dependence among customer flows can be introduced by **splitting**.
- **Customer feedback** introduces dependence between arrival and service processes.

Dependence in Open Queueing Networks

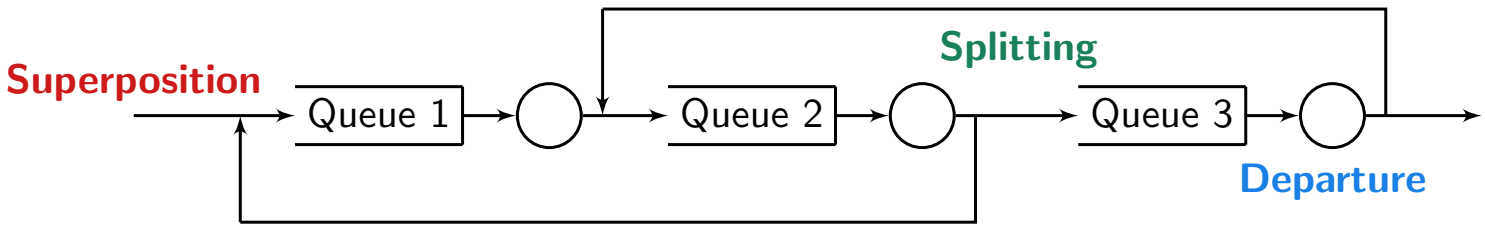


Figure: A three-station example.

Even **generalized Jackson networks** can be complicated

- Arrival process: the **superposition** of independent renewal process cannot be renewal unless all components are Poisson processes.
- **Departure** process cannot be renewal with non-Poisson arrival process or non-Exponential service-time distribution.
- Dependence among customer flows can be introduced by **splitting**.
- **Customer feedback** introduces dependence between arrival and service processes.

Dependence in Open Queueing Networks

In order to handle single nodes within a network, we inevitably faces complicated dependence structure:

- dependence in arrival process;
- dependence in service times;
- correlation between arrival process and service process.

How to approximate $G/G/1$ single-server queue under reasonably general assumptions?

Dependence in Open Queueing Networks

In order to handle single nodes within a network, we inevitably faces complicated dependence structure:

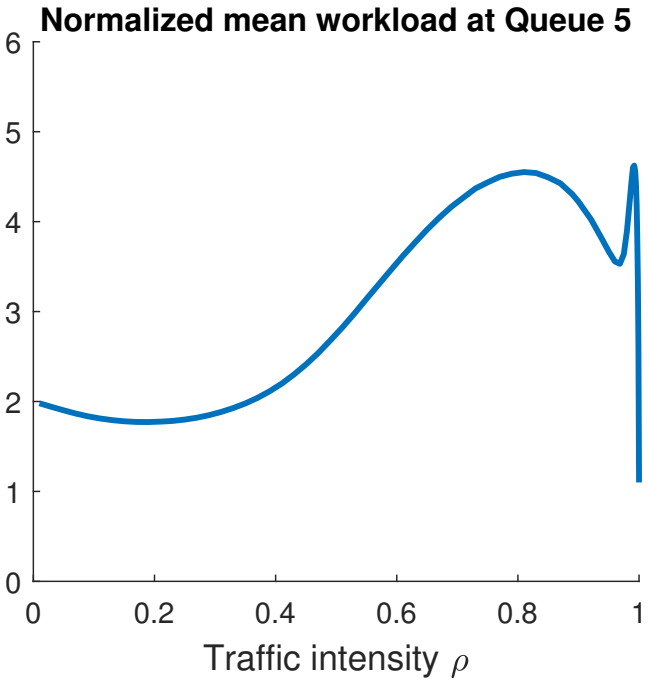
- dependence in arrival process;
- dependence in service times;
- correlation between arrival process and service process.

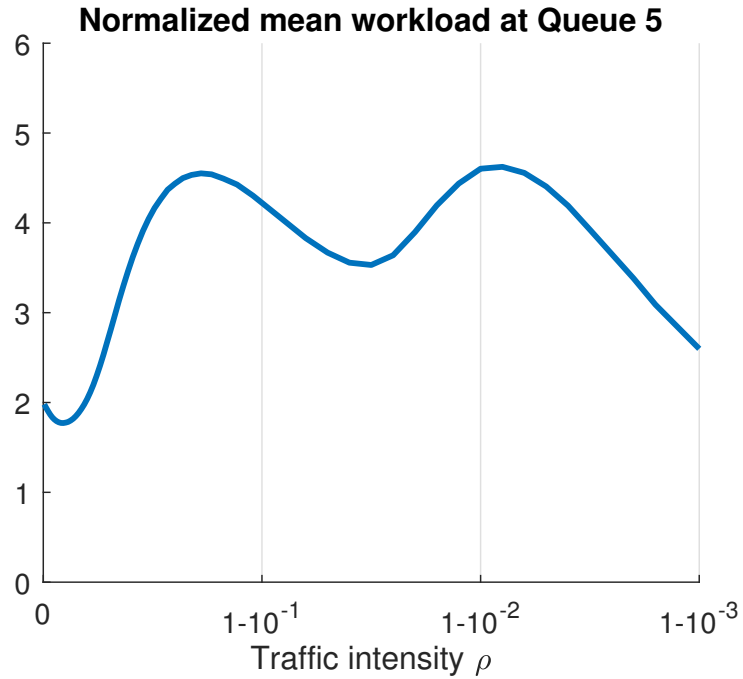
How to approximate $G/G/1$ single-server queue under reasonably general assumptions?

Dependence in Open Queueing Networks

In those non-Markov models,

dependence $\Rightarrow$ significant impact on performance measures.

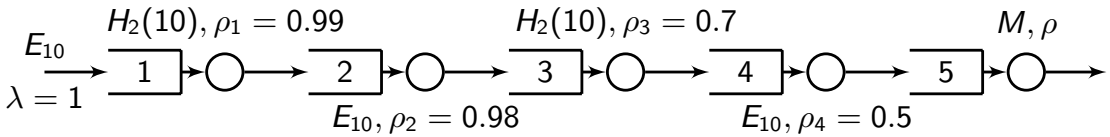


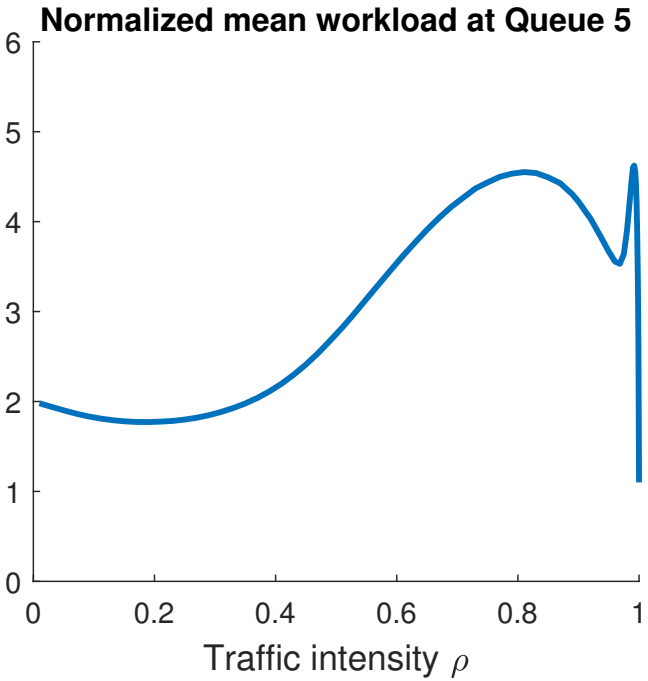


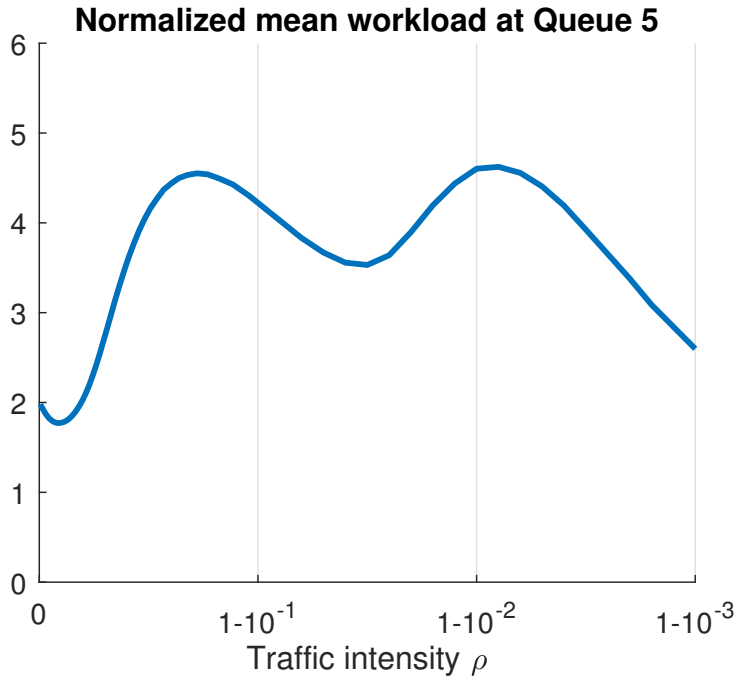
Dependence in Open Queueing Networks

In those non-Markov models,

dependence $\Rightarrow$ significant impact on performance measures.







Dependence in Open Queueing Networks

In those non-Markov models, closed-form characterization of the performance measures are rarely available

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

The purpose is to develop such an approximation algorithm to expose the impact of dependence on performance measures in non-Markov single-server open queueing networks, using non-parametric modeling.

Dependence in Open Queueing Networks

In those non-Markov models, closed-form characterization of the performance measures are rarely available

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

The purpose is to develop such an approximation algorithm to expose the impact of dependence on performance measures in non-Markov single-server open queueing networks, using non-parametric modeling.

Decomposition approximation

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

Examples

- The Queueing Network Analyzer (QNA) by Whitt (1983),
  - approximates each station by a GI/GI/1 queue.
- Markov Arrival Process (MAP)
  - Li and Hwang (1997), statistical fitting of MMPP[1].
  - Horváth et al. (2010), MAP/MAP/1.
  - Kim (2011a, 2011b), MMPP(2)/GI/1.

[1] Markov-modulated Poisson process.

Decomposition approximation

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

Examples

- The Queueing Network Analyzer (QNA) by Whitt (1983),
  - approximates each station by a GI/GI/1 queue.
- Markov Arrival Process (MAP)
  - Li and Hwang (1997), statistical fitting of MMPP[1].
  - Horváth et al. (2010), MAP/MAP/1.
  - Kim (2011a, 2011b), MMPP(2)/GI/1.

[1] Markov-modulated Poisson process.

Diffusion Approximations

- Heavy-traffic limits with Reflected Brownian Motion (RBM).
  - Iglehart and Whitt (1970), Harrison (1973), (1978) and Reiman (1984);
- Numerically calculate the steady-state mean of the RBM.
- Validity of approximation relies on exchange of limit arguments
  - Gamarnik and Zeevi (2006), Budhiraja and Lee (2009) and Braverman et al. (2017).

Examples

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

Diffusion Approximations

- Heavy-traffic limits with Reflected Brownian Motion (RBM).
  - Iglehart and Whitt (1970), Harrison (1973), (1978) and Reiman (1984);
- Numerically calculate the steady-state mean of the RBM.
- Validity of approximation relies on exchange of limit arguments
  - Gamarnik and Zeevi (2006), Budhiraja and Lee (2009) and Braverman et al. (2017).

Examples

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

More Approximations

Robust queueing approximations

- 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 to characterize the underlying stochastic processes.
  - First two moments: QNA, QNET, SBD...
  - Generator matrices for models using MAP.

Approximations based on non-parametric traffic descriptions

- Peakness function, Jagerman et al. (2004);
- Power spectrum, Li and Hwang (1992, 1993).

More Approximations

Robust queueing approximations

- 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 to characterize the underlying stochastic processes.
  - First two moments: QNA, QNET, SBD...
  - Generator matrices for models using MAP.

Approximations based on non-parametric traffic descriptions

- Peakness function, Jagerman et al. (2004);
- Power spectrum, Li and Hwang (1992, 1993).

More Approximations

Robust queueing approximations

- 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 to characterize the underlying stochastic processes.
  - First two moments: QNA, QNET, SBD...
  - Generator matrices for models using MAP.

Approximations based on non-parametric traffic descriptions

- Peakness function, Jagerman et al. (2004);
- Power spectrum, Li and Hwang (1992, 1993).

Robust Queueing Network Analyzer

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

- **Approximations for**
  - Qunatiles and mean of **workload** process[2].
  - Brumelle's formula $\Rightarrow$ mean **waiting time** approximation.
  - Little's Law $\Rightarrow$ mean **queue length** approximation.
- **Computation complexity:**
  - Solve a set of linear equations and a one dimensional optimization problem.

[2] virtual waiting time

Robust Queueing Network Analyzer

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

- **Approximations for**
  - Qunatiles and mean of **workload** process[2].
  - Brumelle's formula $\Rightarrow$ mean **waiting time** approximation.
  - Little's Law $\Rightarrow$ mean **queue length** approximation.
- **Computation complexity:**
  - Solve a set of linear equations and a one dimensional optimization problem.

[2] virtual waiting time

Robust Queueing Network Analyzer

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

- **Approximations for**
  - Qunatiles and mean of **workload** process[2].
  - Brumelle's formula $\Rightarrow$ mean **waiting time** approximation.
  - Little's Law $\Rightarrow$ mean **queue length** approximation.
- **Computation complexity:**
  - Solve a set of linear equations and a one dimensional optimization problem.

[2] virtual waiting time

Robust Queueing Network Analyzer

- **Main idea:** Robust optimization + Queueing theory, hence the name Robust Queueing (RQ).
  - Replace probability laws by uncertainty sets, and analyze the worst case scenario.
- **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.
- **Supporting theories:**
  - Heavy-traffic limit theorems for stationary flows and their IDCs.
- **Extension**: Time-varying arrival-rate and service-rate functions.

Robust Queueing Network Analyzer

- **Main idea:** Robust optimization + Queueing theory, hence the name Robust Queueing (RQ).
  - Replace probability laws by uncertainty sets, and analyze the worst case scenario.
- **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.
- **Supporting theories:**
  - Heavy-traffic limit theorems for stationary flows and their IDCs.
- **Extension**: Time-varying arrival-rate and service-rate functions.

Robust Queueing Network Analyzer

- **Main idea:** Robust optimization + Queueing theory, hence the name Robust Queueing (RQ).
  - Replace probability laws by uncertainty sets, and analyze the worst case scenario.
- **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.
- **Supporting theories:**
  - Heavy-traffic limit theorems for stationary flows and their IDCs.
- **Extension**: Time-varying arrival-rate and service-rate functions.

Robust Queueing Network Analyzer

- **Main idea:** Robust optimization + Queueing theory, hence the name Robust Queueing (RQ).
  - Replace probability laws by uncertainty sets, and analyze the worst case scenario.
- **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.
- **Supporting theories:**
  - Heavy-traffic limit theorems for stationary flows and their IDCs.
- **Extension**: Time-varying arrival-rate and service-rate functions.

Why is IDC helpful?

Characterization of Renewal Processes

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)

For a renewal process $A(t)$ with rate λ, the inter-renewal time distribution can be calculated from the IDC of its equilibrium 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;

Characterization of Renewal Processes

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)

For a renewal process $A(t)$ with rate λ, the inter-renewal time distribution can be calculated from the IDC of its equilibrium 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;

Remark

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

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

Remark

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

- 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.

Ordering of the Mean Steady-State Workload

Theorem (Ordering of the mean steady-state workload)

Consider two $GI/M/1$ queues, let I_{a_i} denote the IDC of the arrival process A_i in the i-th model $i = 1, 2$. If

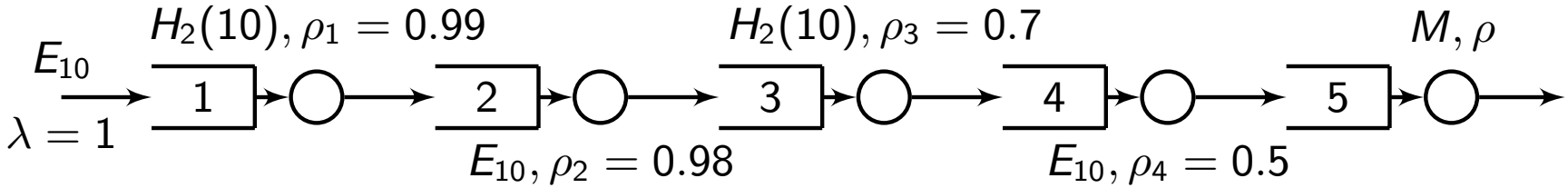
$$I_{a_1}(t) \geq I_{a_2}(t), \quad \text{for } t \geq 0,$$

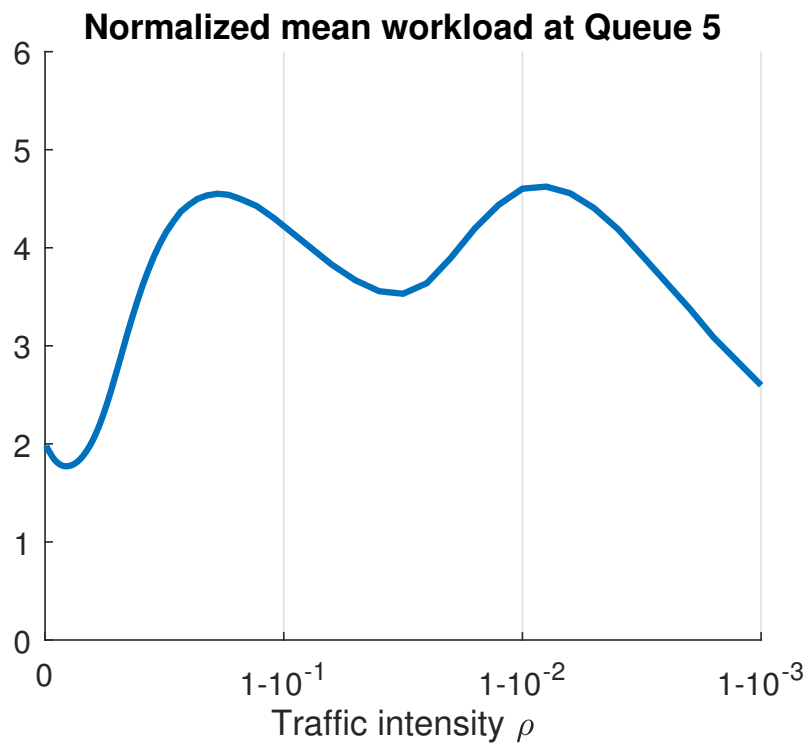
then

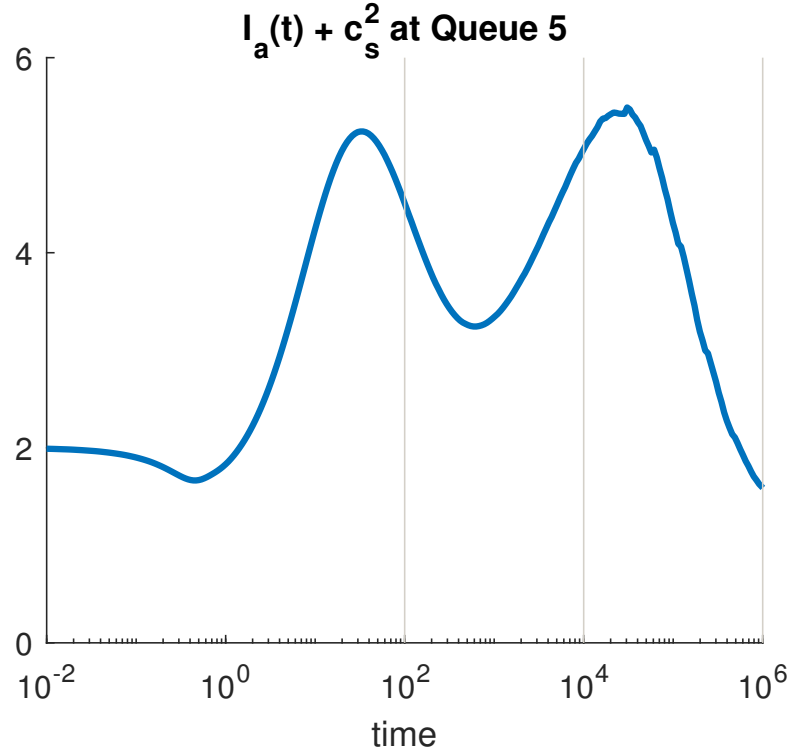
$$E[Z_{1,\rho}] \geq E[Z_{2,\rho}], \quad \text{for } \forall \rho \in (0, 1),$$

where $E[Z_{i,\rho}]$ is the mean steady-state workload in the i-th model, with traffic intensity ρ.

Revisiting the Five Queues in Series Example

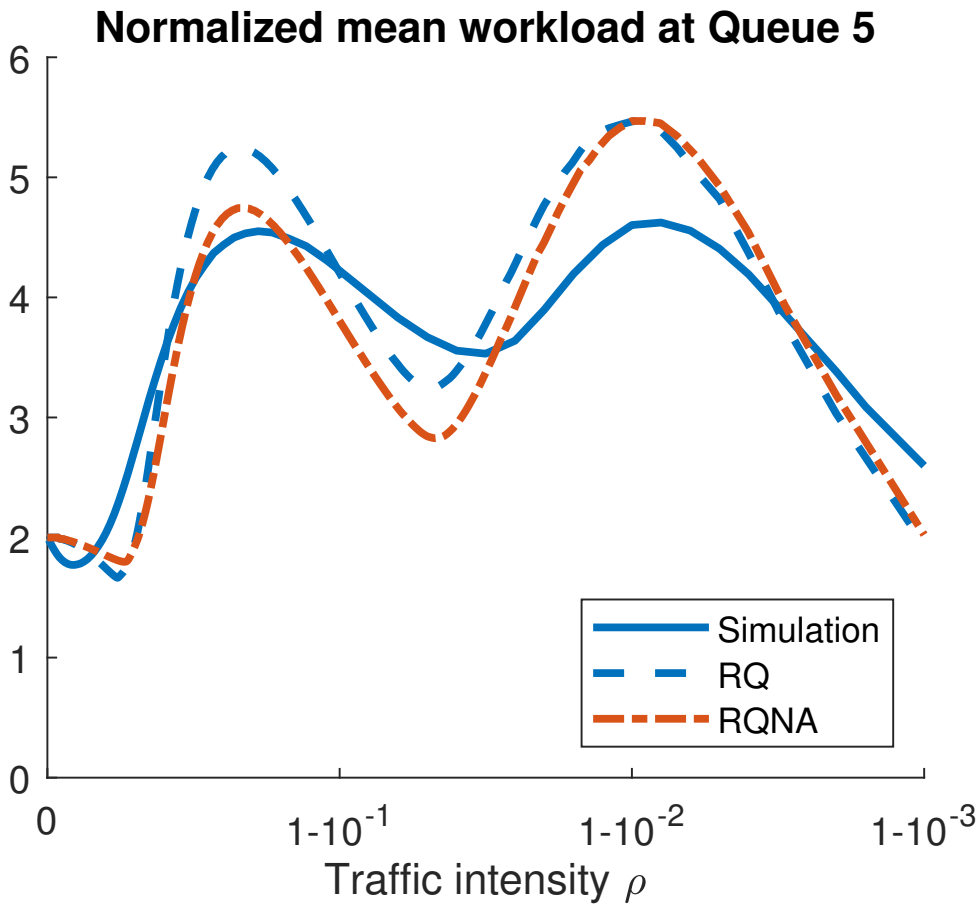






Parametric methods (QNA, RQ by Bandi et al.) using first few moments to describe variability may fail.

Revisiting the Five Queues in Series Example



Robust Queueing for Single-Server Queues

Notation

- $\{(U_i, V_i)\}$: 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

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

Notation

- $\{(U_i, V_i)\}$: 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

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

Notation

- $\{(U_i, V_i)\}$: 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

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

Notation

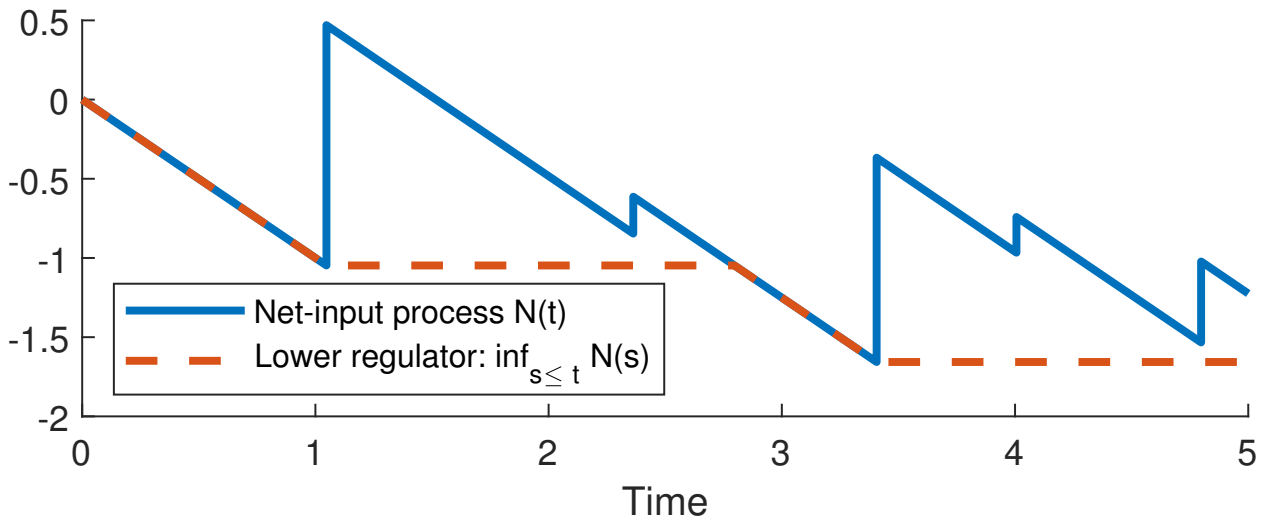
- $\{(U_i, V_i)\}$: interarrival times and service times;
- λ, μ: arrival rate and service rate;
- $A(t)$: arrival counting process associated with $\{U_k\}$;
- $Y(t)$: total input of work

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

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

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

Continuous-time workload process



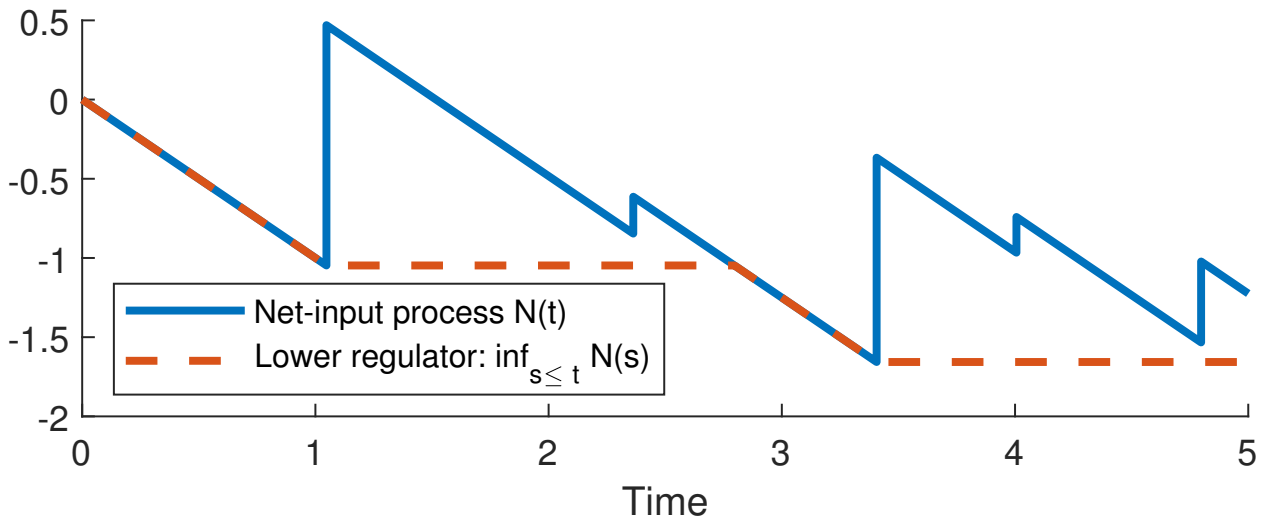
The steady-state workload

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

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

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

Continuous-time workload process



The steady-state workload

$$Z \equiv N(0) - \inf_{-\infty \le t \le 0} \{N(t)\}.$$
$$= \sup_{0 \le s \le \infty} \{N(0) - N(-s)\} \equiv \sup_{0 \le s \le \infty} \{N_0(s)\}.$$

- $N_0(s)$: 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 (deterministic) sample path from a uncertainty set $\mathcal{U}$ of functions.
- Workload defined as the (deterministic) worse-case scenario

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

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 (deterministic) sample path from a uncertainty set $\mathcal{U}$ of functions.
- Workload defined as the (deterministic) worse-case scenario

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

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 (deterministic) sample path from a uncertainty set $\mathcal{U}$ of functions.
- 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

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

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

- Parameter b allows us to approximate the quantiles.

Assume

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

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

Robust Queueing for continuous-time workload

Our uncertainty set is motivated from CLT

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

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

- Parameter b allows us to approximate the quantiles.

Assume

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

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

Robust Queueing for continuous-time workload

RQ for workload

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

where

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

Lemma (Dimension reduction)

The infinite-dimensional RQ problem can be reduced to

$$\begin{aligned} Z^*(b) &= \sup_{0 \le s \le \infty} \sup_{N \in \mathcal{U}_b} \{N(s)\} \\ &= \sup_{0 \le s \le \infty} \left\{ -(1-\rho)s + b\sqrt{\rho s (I_a(s) + c_s^2)/\mu} \right\}. \end{aligned}$$

Approximating the Quantiles

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?

- Approximate the p^{th} quantile $Z(p)$

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

  - Π: one-to-one continuous function, map b into quantile level p.

Approximating the Quantiles

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?

- Approximate the p^{th} quantile $Z(p)$

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

  - Π: one-to-one continuous function, map b into quantile level p.

Approximating the Quantiles

Which function Π should we use?

- For $M/M/1$ view

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

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 $b = \sqrt{2}$

Approximating the Quantiles

Which function Π should we use?

- For $M/M/1$ view

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

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 $b = \sqrt{2}$

Approximating the Quantiles

Which function Π should we use?

- For $M/M/1$ view

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

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 $b = \sqrt{2}$

Approximating the Quantiles

Which function Π should we use?

- For $M/M/1$ view

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

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 $b = \sqrt{2}$

Robust Queueing for the Mean Steady-State 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.

Robust Queueing for $G/G/1$ Model

The RQ algorithm for mean steady-state workload

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

where I_w is the index of dispersion for work (IDW)

$$I_w(t) \equiv \frac{Var(Y(t))}{E[V_1]E[Y(t)]}, \quad t \ge 0.$$

- Takes the arrival IDW $I_w(t)$ 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 $G/G/1$ models.

Dependent Service Sequence

If service times are i.i.d., independent of the arrival process

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

If there is dependence among service times

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

where

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

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

The Heavy-traffic Bottleneck Phenomenon

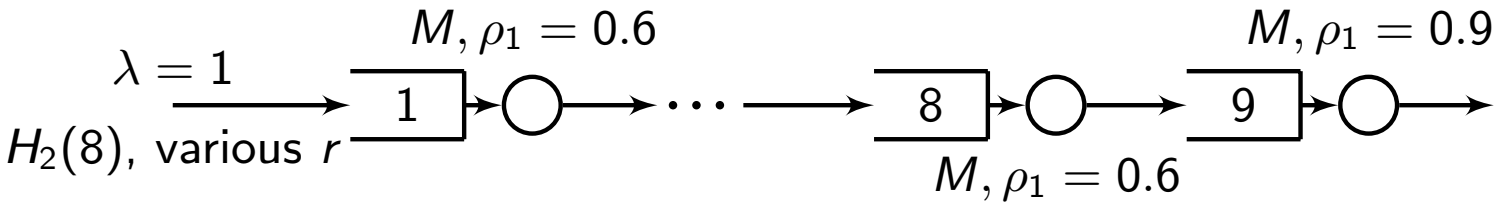


Table: Mean steady-state waiting time at each station.

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%

More on Departure Approximation

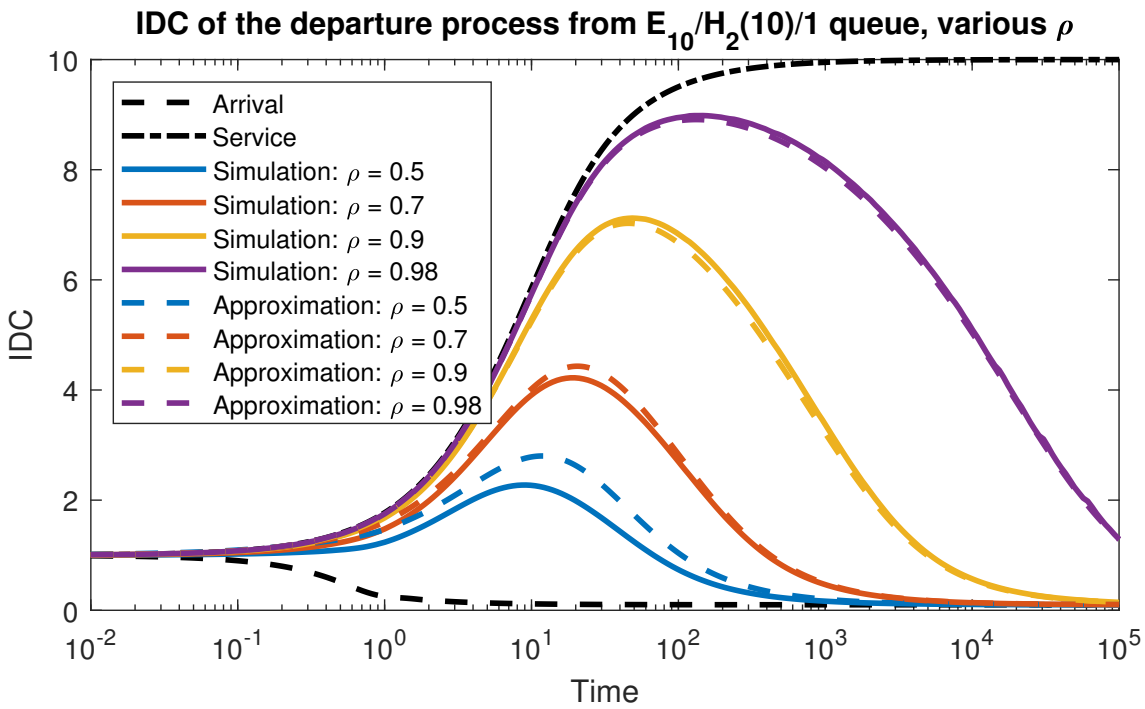
Approximation for Departure IDC

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

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



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.

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 G/G/1 to M/G/1 or GI/M/1 special cases using space-time scaling;
- produces an approximation for departure IDCs in the form of convex combination, as in original QNA paper and its refinements.

Laplace Transform of the Variance Function

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

Takàcs (1962): For M/GI/1

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

- $\hat{g}(s) = E\left[e^{-sV}\right]$ 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

$$\Pi(z) \equiv E\left[z^Q\right] = \frac{(1-\lambda/\mu)(1-z)\hat{g}(\lambda(1-z))}{\hat{g}(\lambda(1-z)) - z}.$$

Laplace Transform of the Variance Function

Daley (1976): For GI/M/1

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

- λ is the arrival rate,
- μ 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

$$\delta = \hat{f}(\mu(1-\delta)).$$

Laplace Transform of the Variance Function

- 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 ρ
  - 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

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

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

$$V^*_{d,\rho}(t) \equiv (1-\rho)^2 V_{d,\rho}\left((1-\rho)^{-2}t\right)$$

  and LT

$$\hat{V}^*_{d,\rho}(s) \equiv (1-\rho)^4 \hat{V}_{d,\rho}\left((1-\rho)^2 s\right)$$

- 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

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

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

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

and ϕ, Φ *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.

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,

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

- $\mathrm{Var}(c_a B_a(t)) = c_a^2 t$;
- $\mathrm{Var}(Q^*(t)) = \mathrm{Var}(Q^*(0)) = c_x^4/4$;
- $\mathrm{cov}(Q^*(0), Q^*(t)) = \frac{c_x^4}{4} c^*(t/c_x^2)$, 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

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

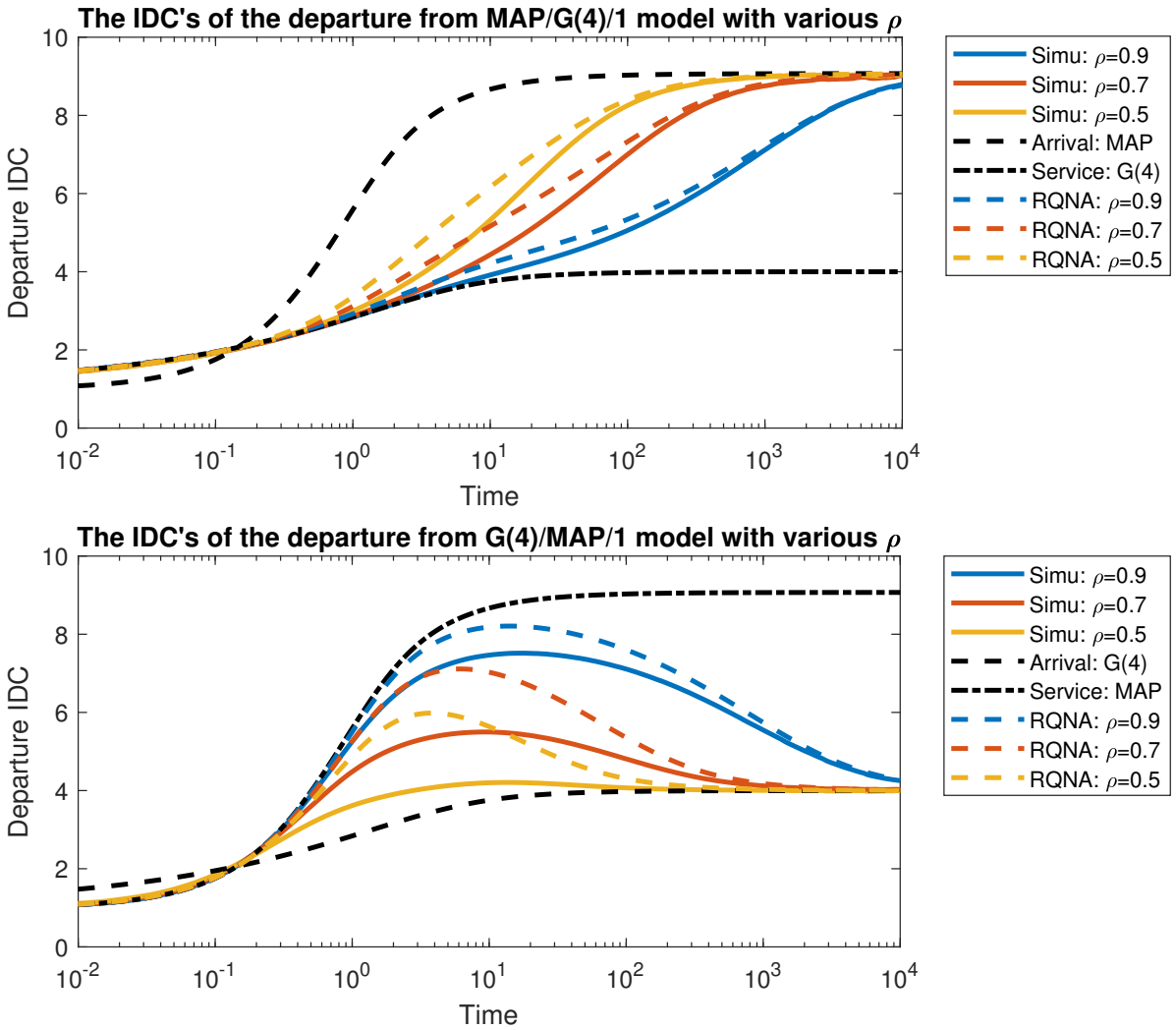
- Scale the original system so that we have a modified system with the same drift -1 but $\tilde{c}_a^2 = 1$.

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

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

Approximation for Departure IDC

Markovian arrival process (MAP) as arrival or service



A Three-Station Network with Feedback

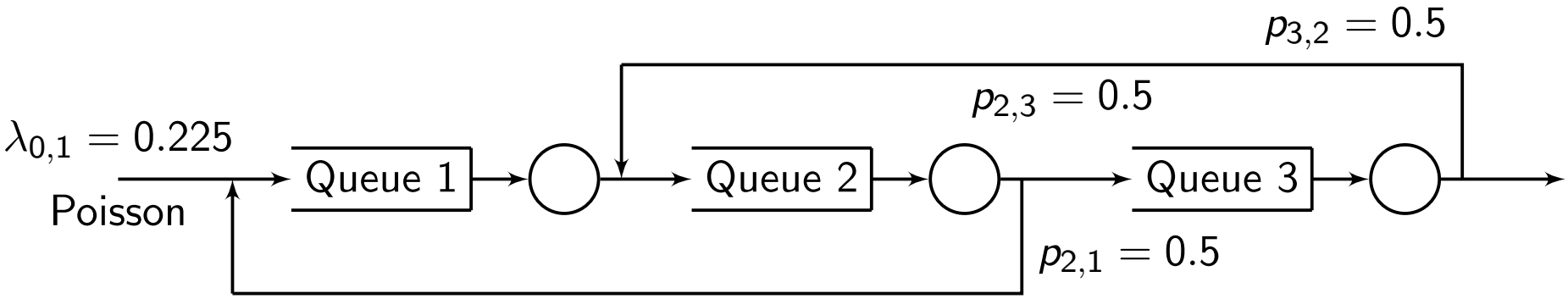


Figure: A three-station example.

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: Squared coefficient of variation of service-time distributions.

Case	$c_{s,1}^2$	$c_{s,2}^2$	$c_{s,3}^2$
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

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%)

- Case E3:

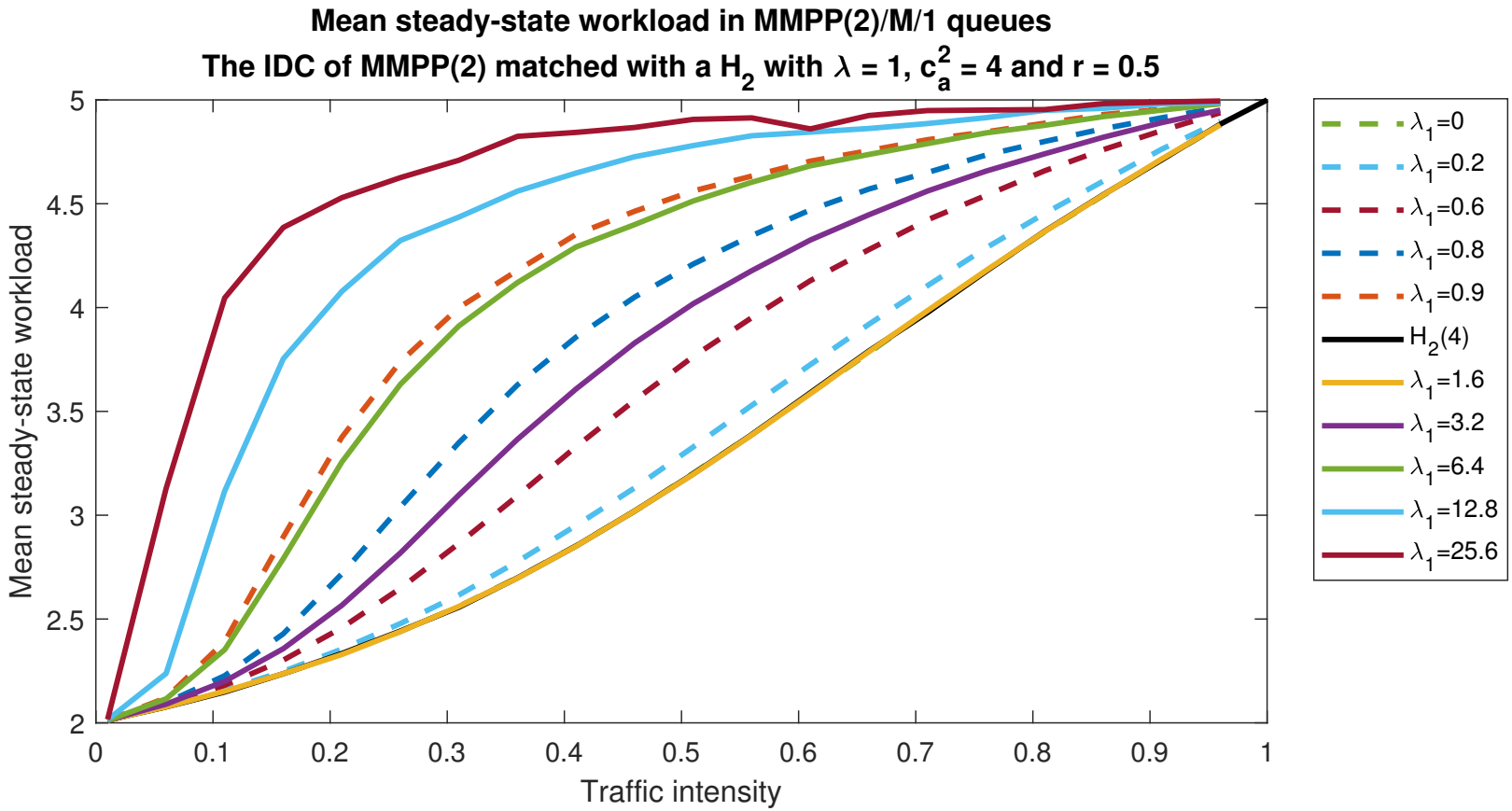
$$(\rho_1, \rho_2.\rho_3) = (0.9, 0.675, 0.45)$$

$$(c_{s_1}^2, c_{s_2}^2.c_{s_3}^2) = (8, 8, 0.25)$$

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%)

Limitations of IDC



The $G_t/G_t/1$ model

- $A(t) = N(\Lambda(t))$: the arrival process
  - $N(t)$: rate-1 *base arrival process*, a general stationary and ergodic point process.
  - $\Lambda(t)$: cumulative arrival-rate function

$$\Lambda(t) \equiv \int_0^t \lambda(s)\, ds, \quad t \geq 0.$$

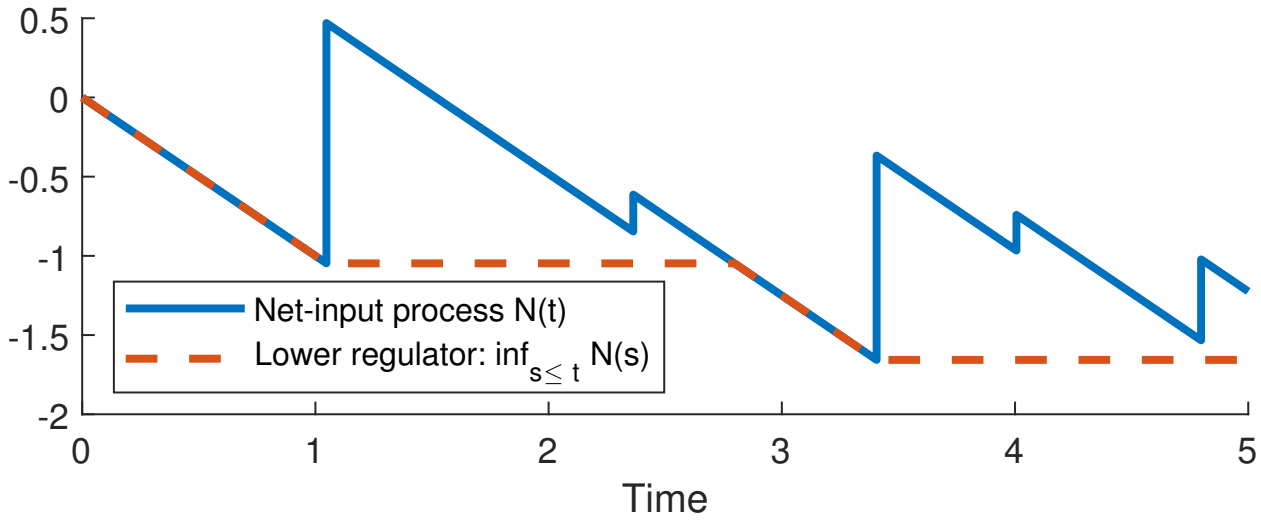
- $\{V_k\}$: stationary sequence of service times with mean 1.
- Service is offered at a variable rate of $\mu(t)$.
  - $M(t)$: cumulative service-rate function

$$M(t) \equiv \int_0^t \mu(s)\, ds, \quad t \geq 0.$$

- $X(t)$: the *net input of work*, defined by

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

Reverse-time formulation of the workload process



To obtain the workload (virtual waiting time) at time t, starting empty at time t_0, one apply the one-sided reflection mapping to $X(t)$

$$W_t(t_0) = X(t) - \inf_{t_0 \le u \le t} \{X(u)\} = \sup_{t_0 \le u \le t} \{X(t) - X(u))\}$$
$$= \sup_{0 \le s \le t - t_0} \{X_t(s)\}$$

where $X_t(s)$ is the reverse-time net input starting backwards at time t for a time period of length s.

Reverse-time formulation of the workload process

$X_t(s)$ is the reverse-time net input starting backwards at time t for a time period of length s, i.e.,

$$X_t(s) \equiv X(t) - X(t-s) \stackrel{d}{=} \sum_{k=1}^{N(\Lambda_t(s))} V_k - M_t(s)$$

with

$$\Lambda_t(s) \equiv \Lambda(t) - \Lambda(t-s), \quad s \geq 0,$$

$$M_t(s) \equiv M(t) - M(t-s), \quad s \geq 0.$$

The steady-state workload

To obtain the steady-state, we start the empty queue in a remote past, i.e., let $t_0 \to -\infty$. Hence, the steady-state workload at time t is formulated as

$$W_t \equiv W_t(-\infty) = \sup_{s \geq 0}\{X_t(s)\}$$

- For TVRQ, we aim to provide approximations for the mean and quantile of the steady-state workload $\mathbb{E}[W_t]$.

The Robust Queueing model

$$W_t \stackrel{d}{=} \sup_{s\geq 0}\left\{\sum_{k=1}^{N(\Lambda_t(s))} V_k - M_t(s)\right\} \equiv \sup_{s\geq 0}\{X_t(s)\}.$$

The idea of Robust Queueing is the replace the probability law of $X_t(s)$ by uncertainty sets and analyze the worst case scenario.

- $\tilde{X}_t \in \mathcal{U}_t$ for a suitable uncertainty set $\mathcal{U}_t$ of net input functions.
- The steady-state RQ workload is defined by

$$W_t^*(\tilde{X}_t) \equiv \sup_{s\geq 0}\{\tilde{X}_t(s)\}$$

- We use the worse-case scenario to characterized the Robust Queue:

$$W_t^* = \sup_{\tilde{X}_t\in\mathcal{U}_t} W_t^*(\tilde{X}_t).$$

TVRQ formulation using IDW

Define the *Index of Dispersion for Work* (IDW) for the underlying (time homogeneous) process

$$I_w(t) \equiv \frac{\text{Var}\left(\sum_{k=1}^{N(t)} V_k\right)}{\mathbb{E}\left[\sum_{k=1}^{N(t)} V_k\right]} = t^{-1}\text{Var}\left(\sum_{k=1}^{N(t)} V_k\right).$$

- Scaled version of the variance curve, independent of the time unit we choose.
- Captures the stochastic variability in single-server queues.
- Usually bounded in practical cases.

TVRQ formulation using IDW

Motivated from CLT, we define

$$\mathcal{U}_t \equiv \left\{ \tilde{X}_t : \tilde{X}_t(s) \leq E\left[X_t(s)\right] + b\sqrt{\mathrm{Var}\left(X_t(s)\right)} \right\}.$$

Under our stochastic settings, we have

$$E\left[X_t(s)\right] = \Lambda_t(s) - M_t(s),$$

$$\mathrm{Var}(X_t(s)) = \mathrm{Var}\left(\sum_{k=1}^{N(\Lambda_t(s))} V_k \right) \equiv \Lambda_t(s) I_w\left(\Lambda_t(s)\right),$$

The uncertainty set for TVRQ can be written as

$$\mathcal{U}_t = \left\{ X : X(s) \leq \Lambda_t(s) - M_t(s) + b\sqrt{\Lambda_t(s) I_w\left(\Lambda_t(s)\right)} \right\}.$$

The TVRQ algorithm

One can prove the following interchange of supremum

$$W_t^* = \sup_{X \in \mathcal{U}_t} \sup_{s \geq 0} \{X(s)\} = \sup_{s \geq 0} \sup_{X \in \mathcal{U}_t} \{X(s)\}$$

- The TVRQ algorithm for the time-varying steady-state workoad at time t in the general $G_t/G_t/1$ model

$$W_t^*(b) = \sup_{s \geq 0} \left\{ \Lambda_t(s) - M_t(s) + b\sqrt{\Lambda_t(s) I_w\left(\Lambda_t(s)\right)} \right\}.$$

  - Easily solvable one-dimensional optimization problem.

We now consider only the periodic case.

- Periodic Robust Queueing (PRQ).

Approximating the Quantile - Stationary Case

$$W_t^*(b) = \sup_{s \geq 0} \left\{ \Lambda_t(s) - M_t(s) + b\sqrt{\Lambda_t(s) I_w\left(\Lambda_t(s)\right)} \right\}.$$

Connecting $W_t^*(b)$ to the distribution of the steady-state workload W?

- We propose to approximate the p^{th} quantile $Z(p)$

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

  - Find appropriate Π: one-to-one continuous function, map b into quantile level p.

Approximating the Quantiles - Stationary Case

Starting with the stationary model.

- For $M/M/1$ queue

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

Hence the p^{th} quantile is

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

- On the other hand, for $M/M/1$ model, TVRQ 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.

Approximating the Quantiles - Underloaded Case

Now we consider the underloaded (UL) time-varying queues, i.e.

$$\sup_t \rho_t < 1.$$

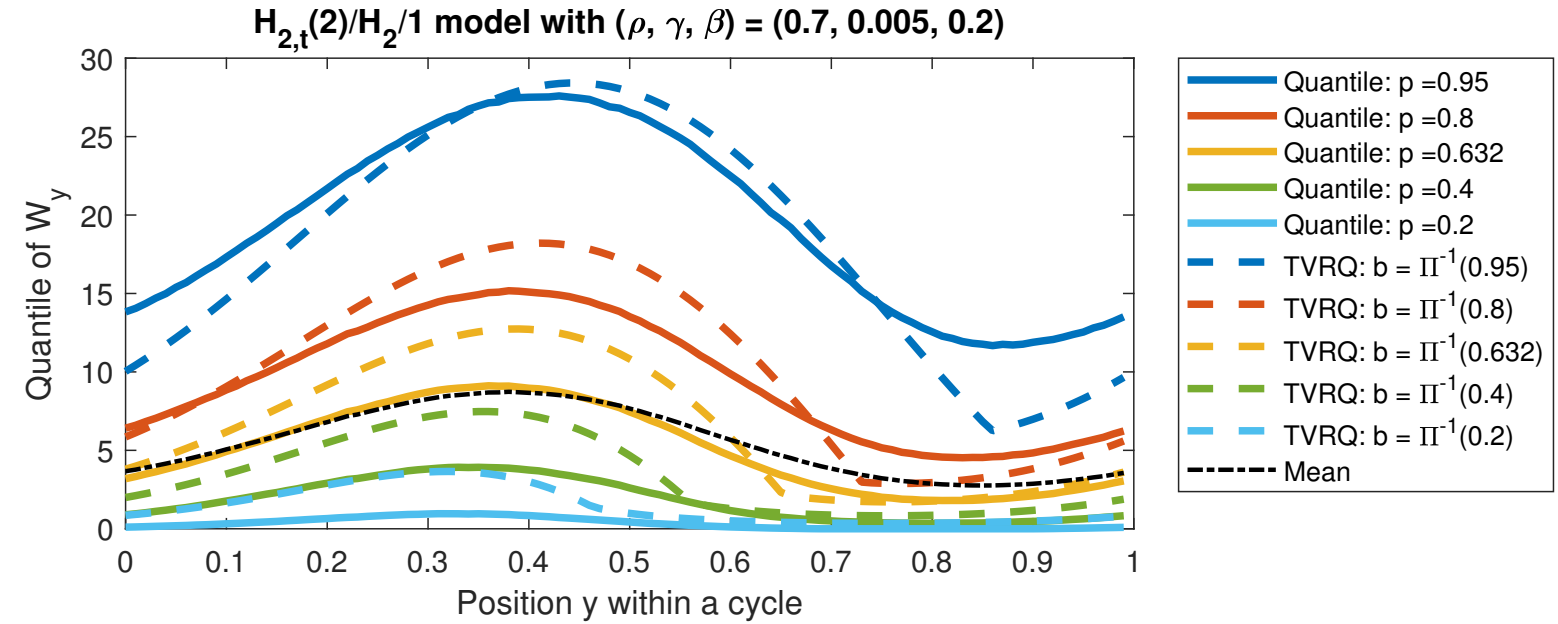
We draw insight from the pointwise stationary approximation (PSA)

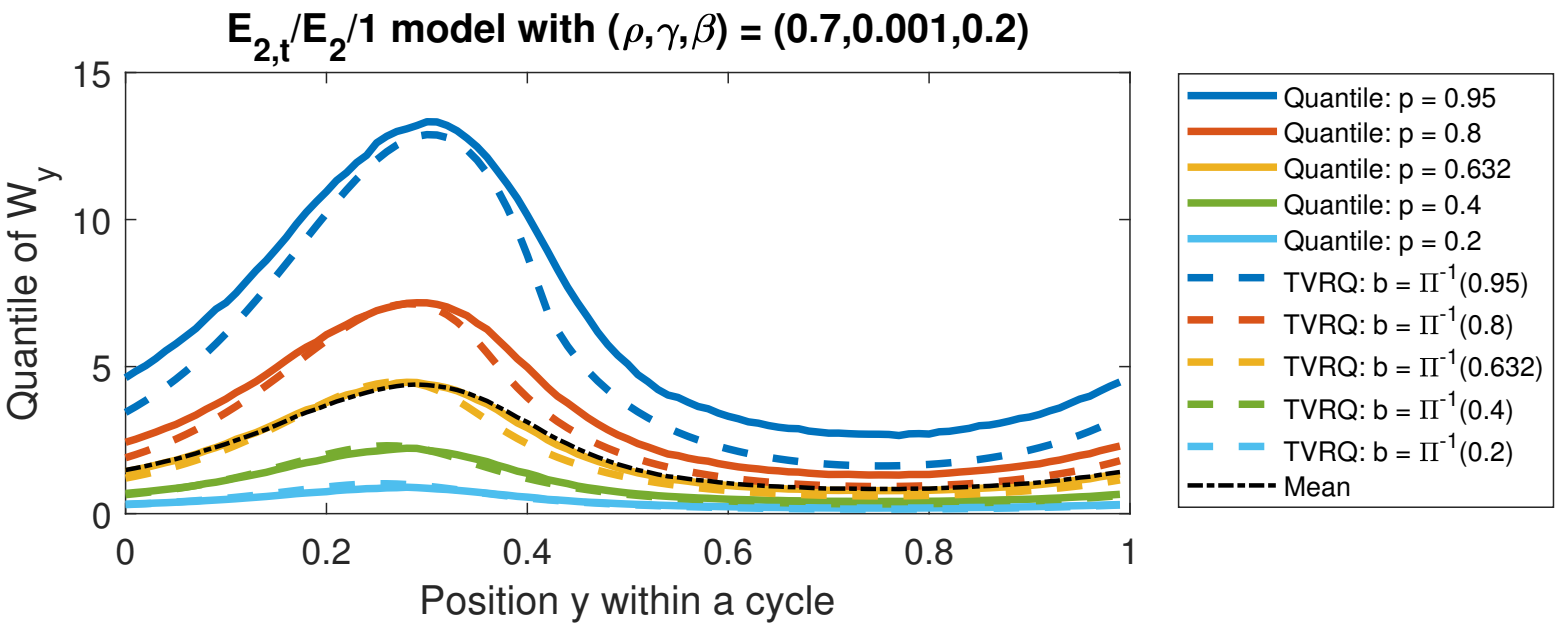
- Green and Kolesar (1991), Massey and Whitt (1998) and Whitt (1991b).
- The PSA is appropriate if the cycle length is sufficiently long that the arrival rate does not change too quickly (relative to the service times).

For the UL case, we use the same

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

$GI_t/GI/1$ model with $\lambda(t) \equiv \rho + \beta \sin(2\pi\gamma t), \quad t \geq 0.$





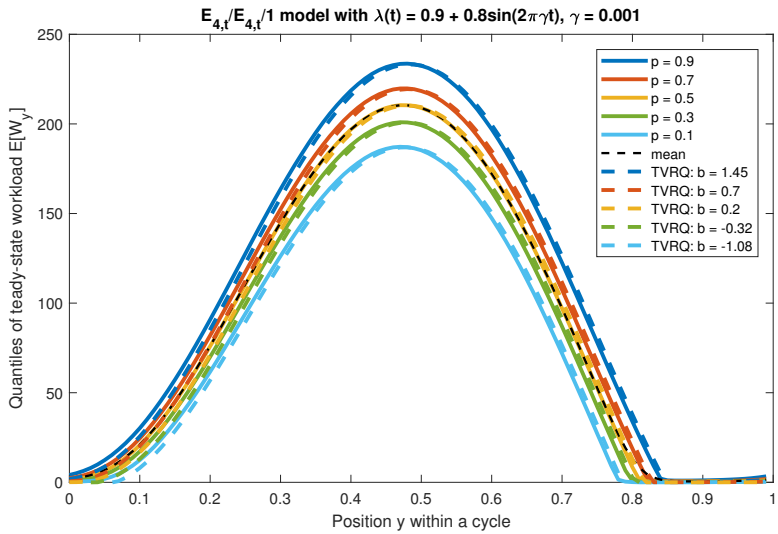
Approximating the Quantiles - Overloaded Case

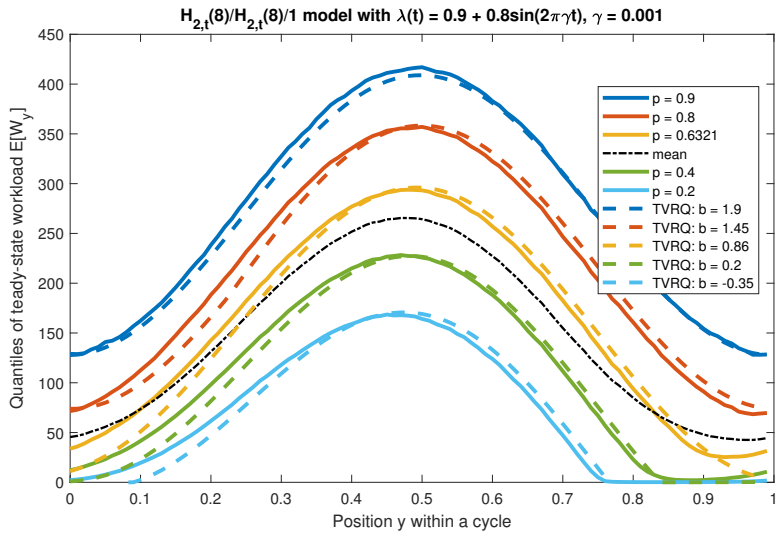
Now we consider the overloaded (OL) case,

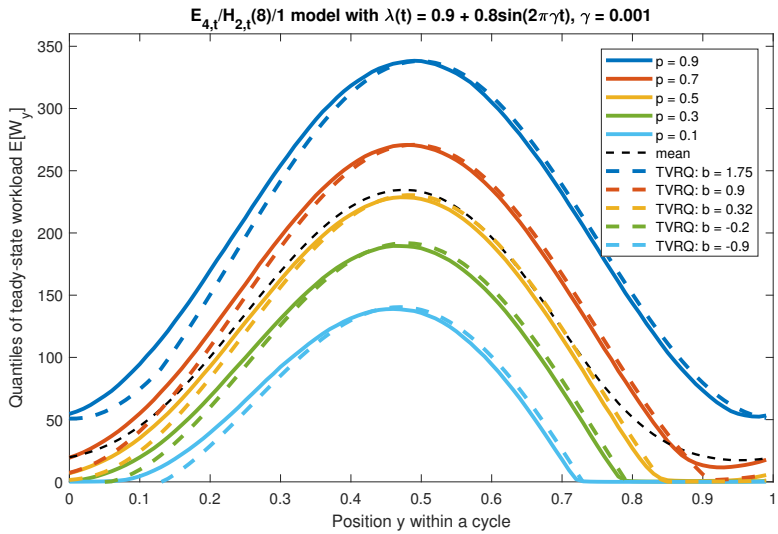
- the long run traffic intensity is below 1;
- but the instantaneous arrival rate can exceed the service rate.

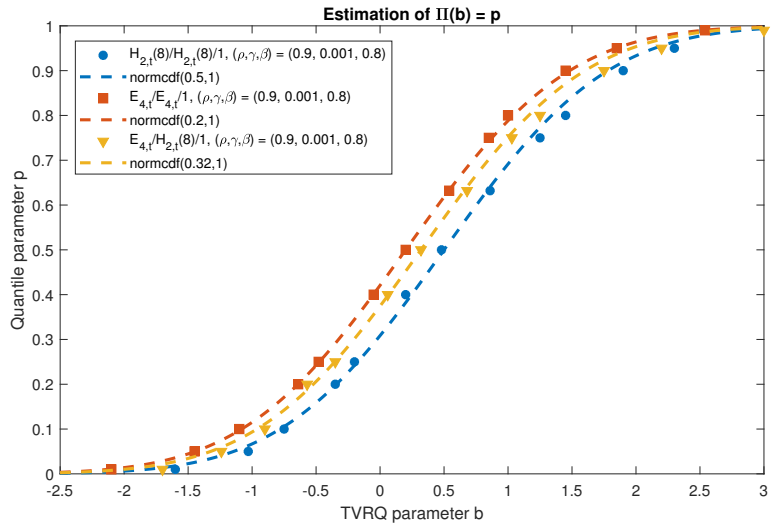
Heavy-traffic theory indicates that W_y in the OL period of the cycle should be approximately Gaussian.

$$\lambda(t) \equiv 0.9 + 0.8 \sin\left(2\pi \times 0.001 \times t\right)$$









As a simple overall approximation, we choose $\Pi(b) \approx \Phi(b; 0.5, 1.0)$, the Gaussian cdf with mean 0.5 and variance 1.

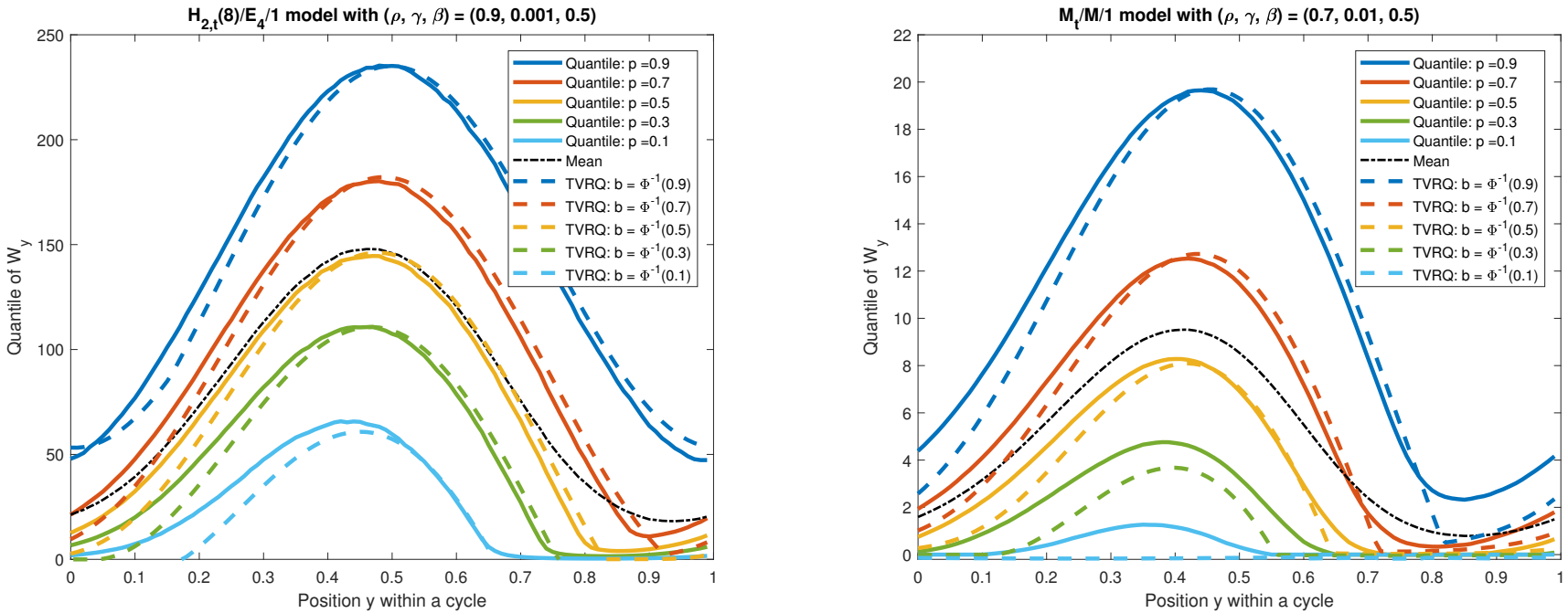


Figure: A comparison of quantiles p ranging from 0.9 to 0.1 estimated by simulation to the PRQ(b) based on Π for the $M_t/M/1$ model and the sinusoidal arrival rate function $\lambda(t) \equiv \rho + \beta \sin(2\pi\gamma t)$ with $(\rho, \beta, \gamma) = (0.9, 0.5, 0.001)$ (left) and $(0.7, 0.5, 0.01)$ (right).

Periodic queues - non-conventional heavy-traffic limits

Cumulative rate functions in the ρ-th model:

$$\Lambda_{\gamma,\rho}(t) \equiv \rho t + (1-\rho)^{-1}\Lambda_{d,\gamma}((1-\rho)^2 t), \quad t \geq 0,$$

$$M_{\gamma,\rho}(t) \equiv t + (1-\rho)^{-1} M_{d,\gamma}((1-\rho)^2 t), \quad t \geq 0,$$

$$\Lambda_{d,\gamma}(t) \equiv \int_0^t h(\gamma t)\, ds, \quad \int_0^1 h(t)\, dt = 0,$$

$$M_{d,\gamma}(t) \equiv \int_0^t r(\gamma t)\, ds, \quad \int_0^1 r(t)\, dt = 0.$$

Theorem (Heavy-traffic limits for the $G_t/GI_t/1$ from Whitt (2014))

Under regularity conditions,

$$\hat{W}_{\gamma,\rho} \Rightarrow \Psi\left(\Lambda_{d,\gamma} - e - M_{d,\gamma} + c_x B\right)$$

Periodic queues - non-conventional heavy-traffic limits

- Diffusion approximation

$$\tilde{W}_{\gamma,\rho,y} \approx \sup_{s\geq 0}\left\{\Lambda_{\gamma,\rho,y}(s) - M_{\gamma,\rho,y}(s) + c_x\tilde{B}(s)\right\}$$

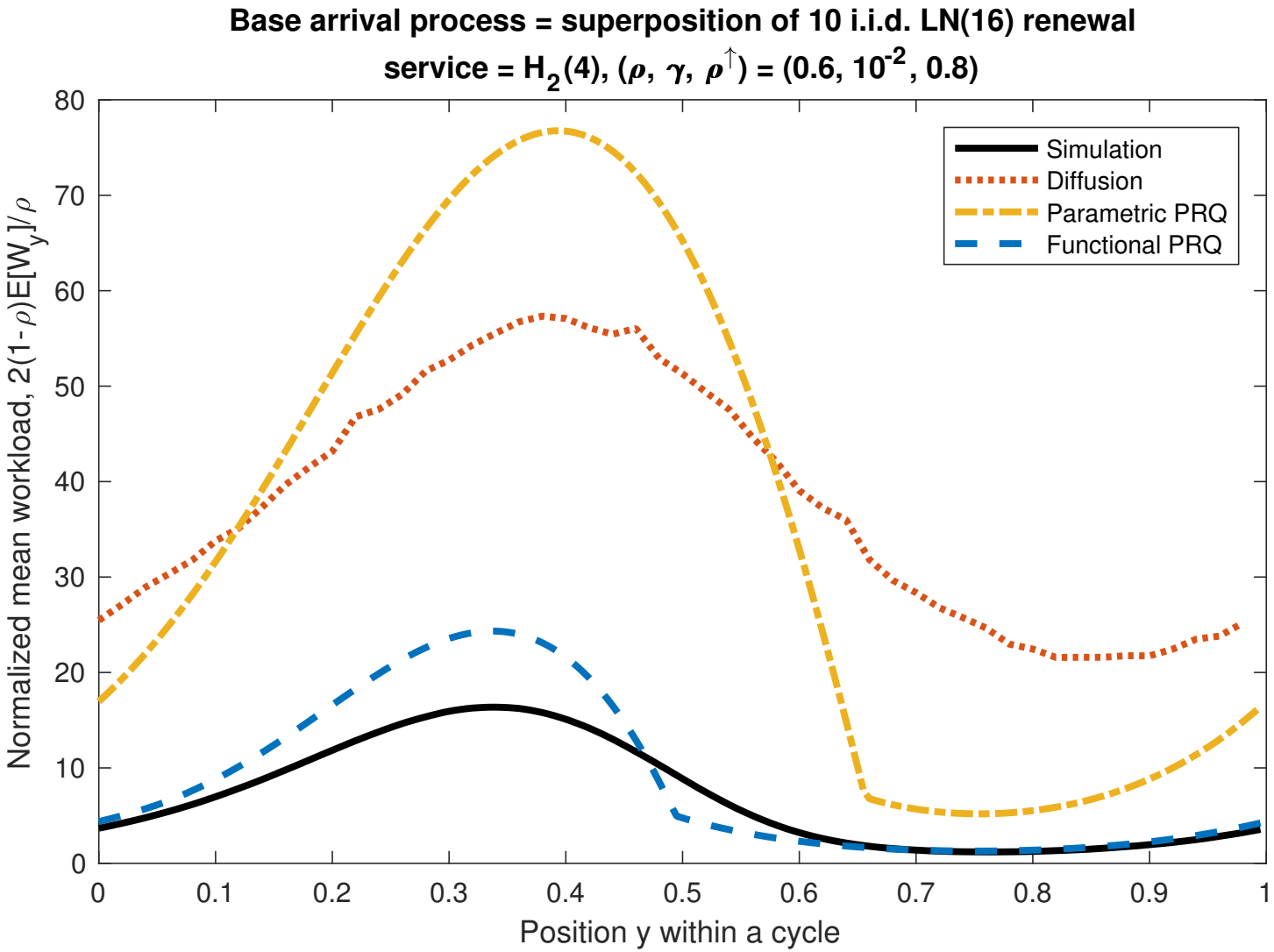
- Parametric PRQ

$$\tilde{W}^*_{\gamma,\rho,y} \equiv \sup_{s\geq 0}\left\{\Lambda_{\gamma,\rho,y}(s) - M_{\gamma,\rho,y}(s) + bc_x\sqrt{s}\right\}.$$

- Non-parametric PRQ

$$W^*_{\gamma,\rho,y} \equiv \sup_{s\geq 0}\left\{\Lambda_{\gamma,\rho,y}(s) - M_{\gamma,\rho,y}(s) + b\sqrt{\Lambda_{\gamma,\rho,y}(s) I_w\left(\Lambda_{\gamma,\rho,y}(s)\right)}\right\}.$$

Diffusion approximation versus PRQs



The heavy-traffic limit for PRQ - overloaded

Theorem (long-cycle heavy-traffic limit for PRQ in an overloaded queue)

For $G_t/G_t/1$ periodic model, the PRQ problem with the heavy-traffic scaling and $\rho^{\uparrow} > 1$ has the limit

$$(1-\rho)\lim_{\gamma \downarrow 0} \gamma \cdot W^*_{\gamma,\rho,y} = \sup_{t \geq 0} \left\{ -t + \int_{y-t}^{y} (h(s) - r(s)) ds \right\}.$$

- We need a space scaling of γ to obtain a proper limit.
- The limit depend on the traffic intensity only through a scaling of $1-\rho$.
- The limit does not depend on the stochastic structure of the associated queueing model.

The heavy-traffic limit for PRQ - underloaded

For underloaded queues, we have the Point-wise Stationary Approximation.

Theorem (long-cycle heavy-traffic limit for PRQ in an underloaded queue)

For $G_t/G_t/1$ periodic model with $\rho^{\uparrow} < 1$, PRQ is asymptotically correct as $(\gamma, \rho) \to (0, 1)$. Furthermore, we have the double limit for PRQ

$$W_y^* = \frac{b^2}{2} \cdot \frac{\rho(y) c_x^2}{2(1-\rho(y))} + o(1-\rho), \quad \text{as } (\gamma, \rho) \to (0, 1),$$

where $I_w(\infty) = c_x^2$ and $\rho(y)$ is the instantaneous traffic intensity.

- No scaling for the cycle-length parameter γ is needed.

The heavy-traffic limit for PRQ - critically-loaded

Recall that

- For underloaded case, we need a space scaling of $\gamma^0 = 1$;
- For overloaded case, we need a space scaling of γ^1;

For critically-loaded case:

- For the stochastic model, [**?**]: the additional space scaling is $\gamma^{p/(2p+1)}$.
- p is obtained from the Taylor's expansion of the arrival rate function at the critical point.

The heavy-traffic limit for PRQ - critically-loaded

Theorem (long-cycle heavy-traffic limit for PRQ in an critically loaded queue)

Assume that

$$h(t) - r(t) = 1 - ct^p + o(t^p), \tag{1}$$

for some integer $p \geq 0$. Then the long-cycle heavy-traffic limit of the PRQ solution at the critical point $y = 0$ is in the order of $O(\gamma^{-p/(2p+1)})$.

- PRQ successfully captures the correct space scaling of a critically-loaded queue in the long-cycle heavy-traffic limit.

Future Directions - Applications

Applications

- Analytic formulation and low computation complexity of RQNA $\Rightarrow$ feed into a top level optimization problem.
  - Given the service rates (resources), how to allocation?
  - Given the facilities, what is the best topological design of a service network?
  - How to balance the amount of resources used and the quality of the service?
  - Stress testing a service system? Quick robustness check.

Future Directions - Methodology

Robust Queueing

- Multi-class customer;
- Customer balking and abandonment;
- Multi-server queueing networks;
- Non-Markovian routing;
- Incorporating higher order statistics.

Indices of dispersion

- Contain rich information of stationary point processes;
- Alternative queueing approximation algorithm using indices of dispersion?
- How indices of dispersion can be applied to analyze inventory theory, supply chain management and risk management;

Miscellaneous

- Machine learning approach to identify mappings from IDC to performance measures.

Future Directions - Methodology

Robust Queueing

- Multi-class customer;
- Customer balking and abandonment;
- Multi-server queueing networks;
- Non-Markovian routing;
- Incorporating higher order statistics.

Indices of dispersion

- Contain rich information of stationary point processes;
- Alternative queueing approximation algorithm using indices of dispersion?
- How indices of dispersion can be applied to analyze inventory theory, supply chain management and risk management;

Miscellaneous

- Machine learning approach to identify mappings from IDC to performance measures.

Future Directions - Methodology

Robust Queueing

- Multi-class customer;
- Customer balking and abandonment;
- Multi-server queueing networks;
- Non-Markovian routing;
- Incorporating higher order statistics.

Indices of dispersion

- Contain rich information of stationary point processes;
- Alternative queueing approximation algorithm using indices of dispersion?
- How indices of dispersion can be applied to analyze inventory theory, supply chain management and risk management;

Miscellaneous

- Machine learning approach to identify mappings from IDC to performance measures.

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.

[DNR94] J. Dai, V. Nguyen, and M. I. Reiman. Sequential bottleneck decomposition: an approximation method for generalized Jackson networks. *Operations research*, 42(1):119-"136, 1994.

[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.

[HHT10] A. Horváth, G. Horváth, and M. Telek. A joint moments based analysis of networks of MAP/MAP/1 queues. *Performance Evaluation*, 67(9):759-778, 2010.

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

[J04] D. L. Jagerman, B. Balcioglu, T. Altiok, and B. Melamed. Mean waiting time approximations in the G/G/1 queue. *Queueing Systems*, 46(3-4):481-506, 2004.

[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.

[LH92] S. Li, C. Hwang. Queue response to input correlation functions: discrete spectral analysis. *The Conference on Computer Communications*. 1993 Dec;1(6):678-92.

[LH93] S. Li, C. Hwang. Queue response to input correlation functions: continuous spectral analysis. *IEEE/ACM transactions on networking*. 1993 Dec;1(6):678-92.

[LH97] S. Li, C. Hwang, On the convergence of traffic measurement and queueing analysis: a statistical-matching and queueing (SMAQ) tool. *IEEE/ACM transactions on networking*. 1997.

[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 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, working paper, 2019.

[WY19a] W. Whitt and W. You, The Advantage of Indices of Dispersion in Queueing Approximations, *Operations Research Letters*, 2019.

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

References on HT limits:

[BDM17] A. Braverman, J. Dai, M. Miyazawa, Heavy traffic approximation for the stationary distribution of a generalized Jackson network: The BAR approach. *Stochastic Systems*. 2017 May 5;7(1):143-96.

[BL09] A. Budhiraja, C. Lee, Stationary distribution convergence for generalized Jackson networks in heavy traffic. *Mathematics of Operations Research*. 2009 Feb;34(1):45-56.

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

[H73] J. M. Harrison. The heavy traffic approximation for single server queues in series. *Journal of Applied Probability*, 10(3):613-629, 1973.

[H78] J. M. Harrison. The diffusion approximation for tandem queues in heavy traffic. *Advances in Applied Probability*, 10(4):886-905, 1978.

[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.

[WY19b] W. Whitt, W. You, Heavy-traffic limit of the GI/GI/1 stationary departure process and its variance function, *Stochastic Systems*, 2019.

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

[H13] 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,

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

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*

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

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

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

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*

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

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.

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

$$\Theta_{i,\rho}^*(t) = (1-\rho)\left[\sum_{l=1}^{\lfloor (1-\rho)^{-2}t \rfloor} \theta^l - \mathbf{p}_i(1-\rho)^{-2}t\right],$$

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

$$Q_{i,\rho}^*(t) = (1-\rho)Q_i((1-\rho)^{-2}t),$$

$$\ldots$$

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

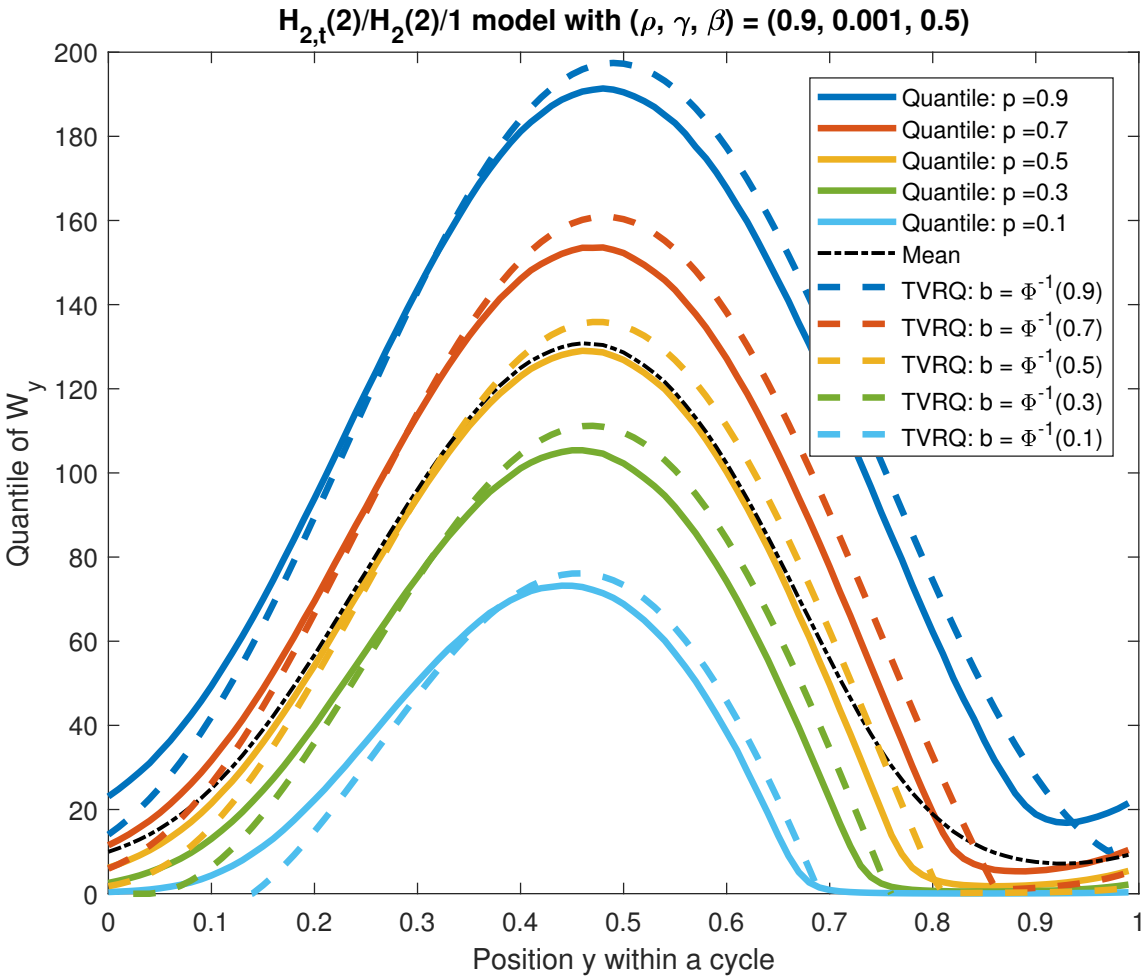
$$\alpha^*_{i,j}(t) \equiv 2\mathrm{cov}(p_{i,j} D^*_i(t), \Theta^*_{i,j}(\lambda_i t))/p_{i,j}\lambda_i t.$$

Corollary

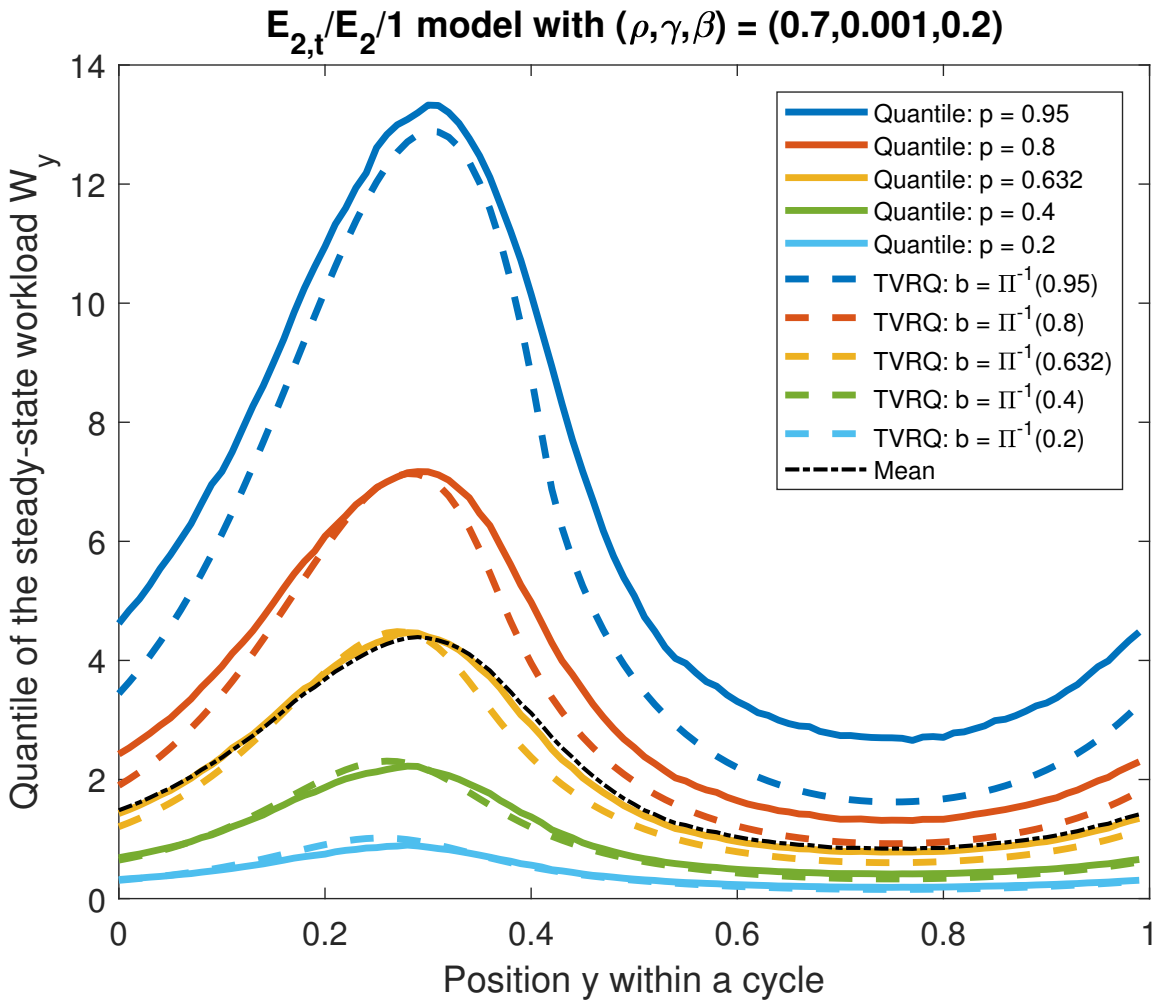
Under regularity conditions, we have

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

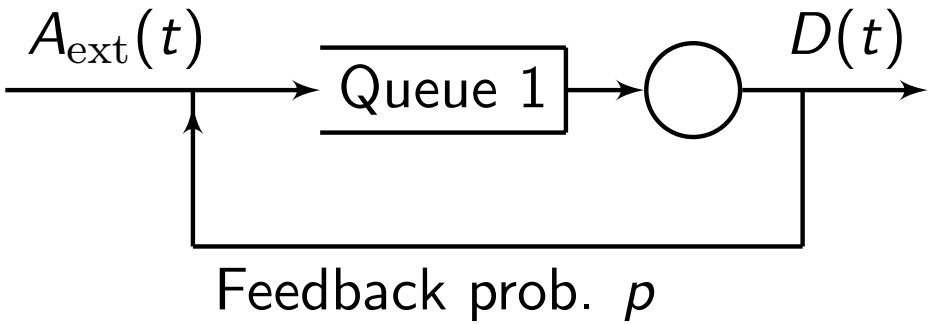
Example: Time-Varying Queue and Percentiles of the Workload



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

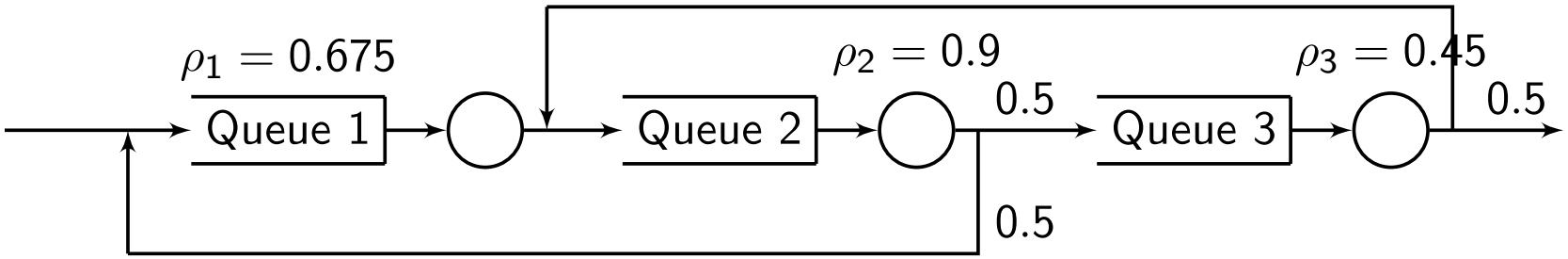


Figure: A three-station example.

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.

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%

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 $(0, 1)$.

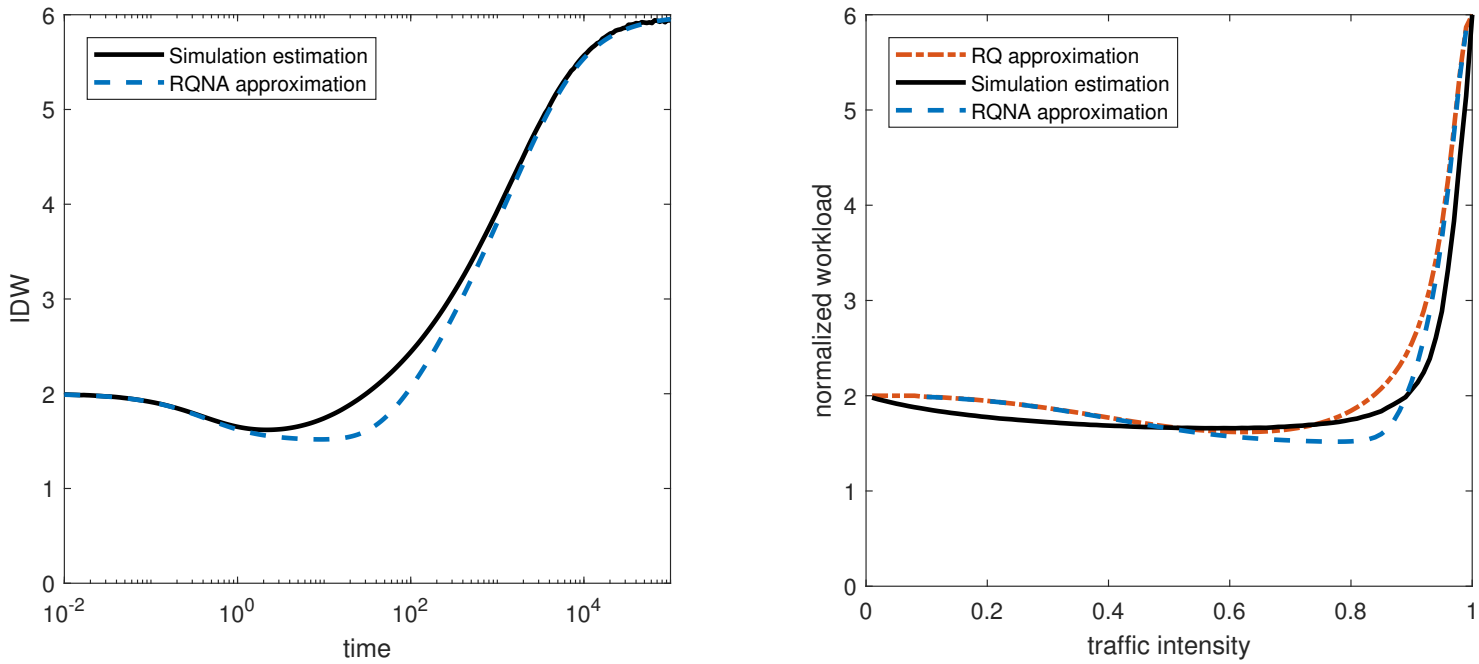


Figure: Contrasting the RQNA approximation of the IDW at the 10-th queue and simulation estimated IDW (left) in the ten queues in series example. Simulation estimation of the steady-state mean workload, the RQ approximation and the RQNA approximation shown in the right plot.

The Heavy-traffic Bottleneck Phenomenon

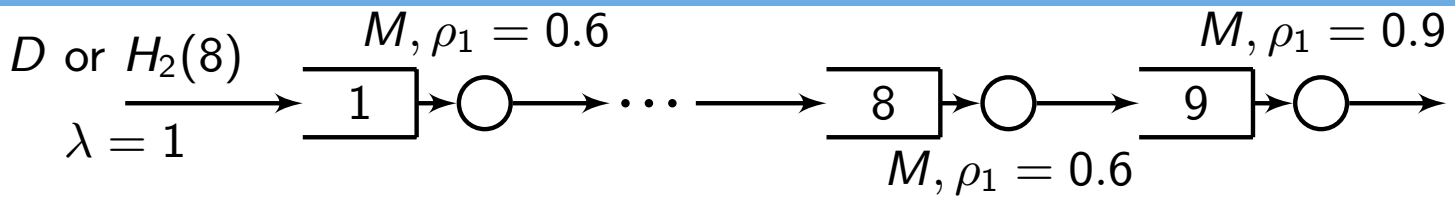
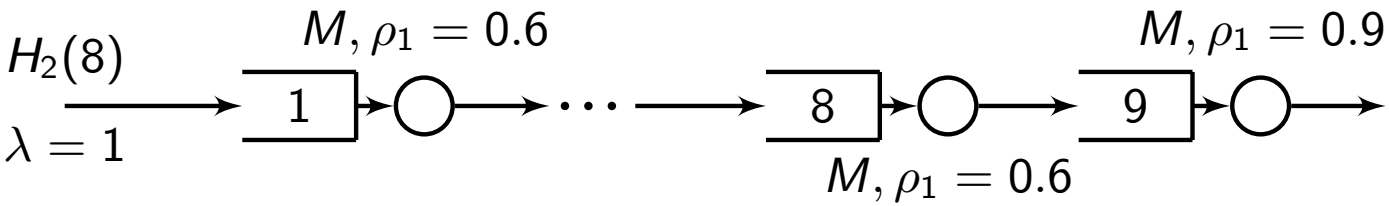


Figure: The heavy-traffic bottleneck example in Suresh and Whitt (1990).

		$H_2, c_a^2 = 8$	$D, c_a^2 = 0$
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%)

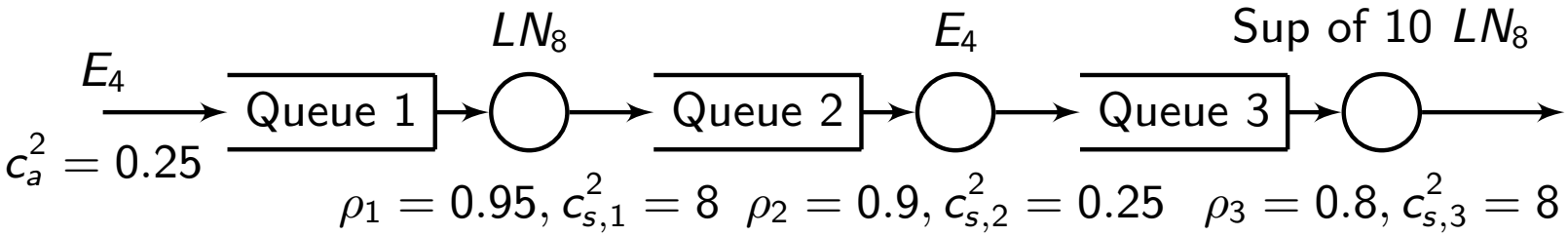
Table: Mean steady-state waiting times at Queue 8 and 9, compared with M/M/1 values, QNA and SBD approximations.

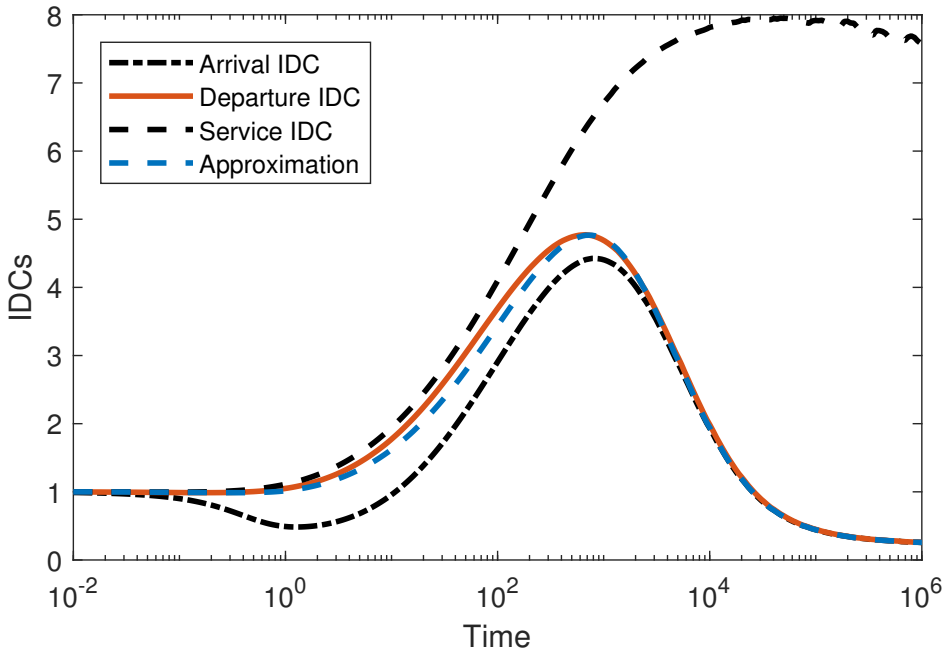
The Heavy-traffic Bottleneck Phenomenon



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%)

An Artificial Example





3 Stations with Feedback

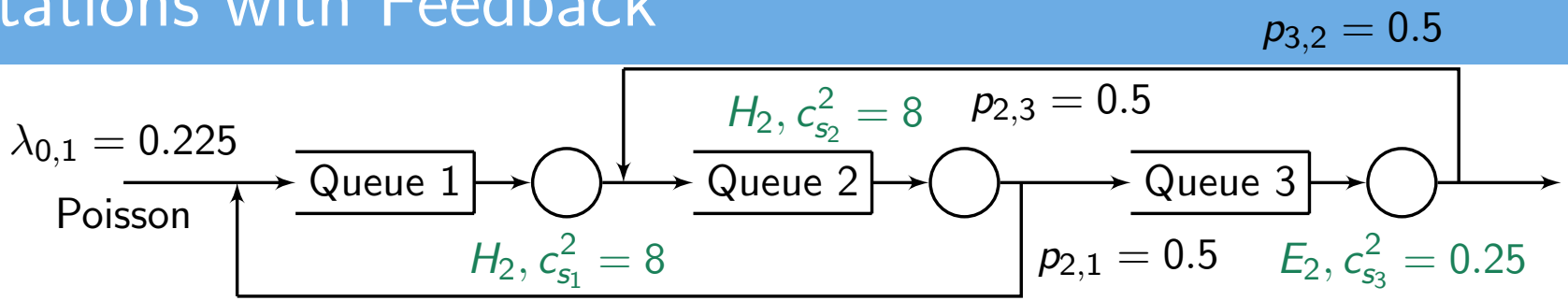
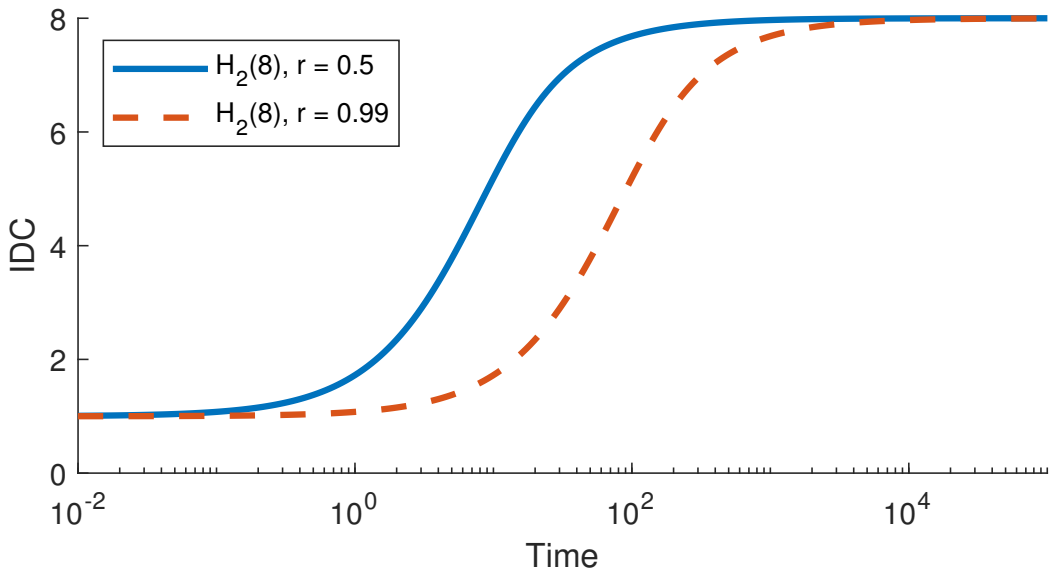


Table: The steady-state mean waiting time.

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%)



10 Stations with Feedback

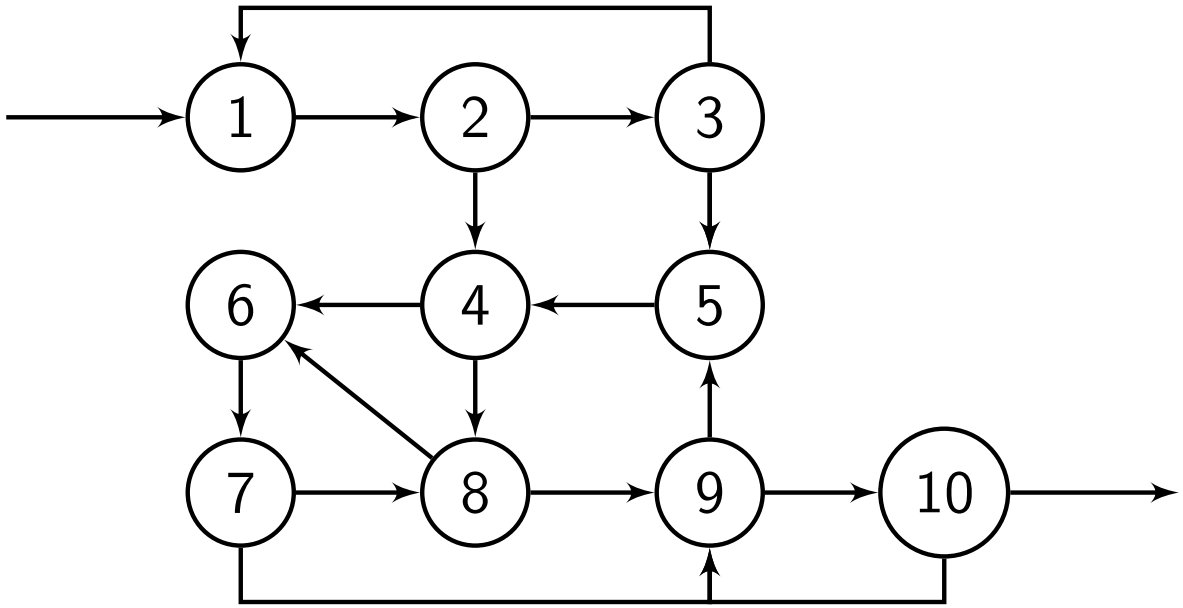


Figure: A ten-station with customer feedback example.

- 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 $(0.5, 2, 2, 0.25, 0.25, 2, 1, 2, 0.5, 0.5)$