Industrial Engineering & Operations Research Department
Columbia University

Time-Varying Robust Queueing

Wei You
joint work with Ward Whitt

IEOR, Columbia University

INFORMS APS Conference 2017
Evanston, IL

Based on "Time-Varying Robust Queueing" by Whitt and You, submitted to *Operations Research* in August 2016, revised in June 2017.

Outline

1. Backgrounds
2. Time-Varying Queueing Model
3. Time-Varying Robust Queueing
4. Periodic Robust Queueing

Backgrounds: the time-varying queues

Queueing models with time-varying arrival rates are traditionally analyzed by

- Deterministic methods: Edie (1954), Oliver and Samuel (1962);
- Numerical methods for time-varying ODEs: Koopman (1972), Kolesar et al. (1975);
  - Improved ODE approach: Rothkopf and Oren (1979), Taaffe and Ong (1987), Ong and Taaffe (1989);
- Heavy-traffic limits: Mandelbaum and Massey (1995), Whitt (2014, 2016);
- Fluid and diffusion approximations: Mandelbaum et al. (1998), Massey and Pender (2013), Pender and Massey (2017);

Backgrounds: robust optimization approaches

Robust optimization approach: replace probability laws by tractable uncertainty sets and apply deterministic optimization.

- Robust inventory theory: Bertsimas and Thiele (2006), Mamani et al. (2016);
- Robust Queueing (RQ): Bertsimas et al. (2011), Bandi et al. (2015).

Our approach

- Recently, we developed new RQ algorithm to expose the impact of dependence in the stationary $G/G/1$ model, see Whitt and You (2017).
- In this talk, we take one step forward to consider the *Time-Varying Robust Queueing* (TVRQ) for general $G_t/G_t/1$ model.
- We focus on providing useful approximations for the time-varying steady-state mean workload with structural insights.

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 $X(t)$

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

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

$$\Lambda_t(s) \equiv \Lambda(t) - \Lambda(t-s), \quad s \ge 0,$$
$$M_t(s) \equiv M(t) - M(t-s), \quad s \ge 0.$$

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 steady-state mean 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 probabilistic 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

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

- 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

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

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

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

The TVRQ algorithm

One can prove the following exchange 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

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

  - Easily solvable one-dimensional optimization problem.
- We shall focus on the Periodic Robust Queueing (PRQ) for the rest of the talk, where λ and μ are periodic functions.

Periodic queues - non-conventional heavy-traffic limits

- The heavy-traffic limits for periodic queueing models were established in Whitt (2014) and Ma and Whitt (2016).

Cumulative rate functions in the ρ-th model:

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

where

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

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

- h and r are periodic functions with period 1.
- γ is the cycle-length parameter.

Periodic queues - non-conventional heavy-traffic limits

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

Under regularity conditions,

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

- This implies that the TVRQ also generates approximation for the reflective periodic Brownian motion (RPBM).
- Diffusion approximation

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

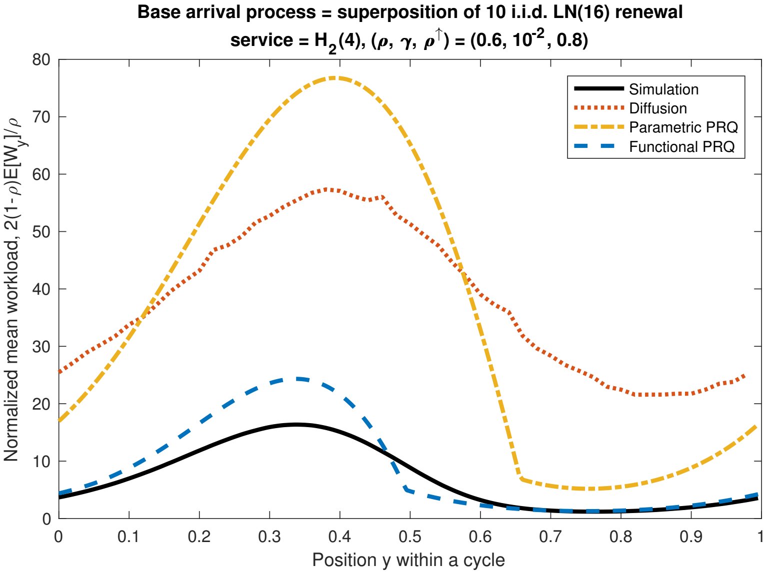
- Parametric PRQ

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

- (Functional) PRQ

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

Diffusion approximation versus PRQs



The heavy-traffic limit for PRQ

Theorem (heavy traffic limit for PRQ)

For $G_t/G_t/1$ periodic queue, if the IDW $I_w(t)$ converges to a finite $I_w(\infty) = c_x^2$, then the heavy traffic limit for PRQ is

$$\lim_{\rho \uparrow 1} \frac{2}{b^2} \cdot \frac{2(1-\rho)}{\rho c_x^2} \cdot W^*_{\gamma,\rho,y} = \sup_{s \geq 0} \left\{ f(t) + \tilde{g}_{\gamma,1,y}(t) \right\}. \tag{1}$$

- $f(t) \equiv -t + 2\sqrt{t}$ captures the corresponding $G/G/1$ model.
- $g_{\gamma,\rho,y}$ is a periodic function that captures the time-varying feature of the model

$$\tilde{g}_{\gamma,\rho,y}(t) = \frac{4}{b^2 c_x^2 \gamma \rho^2} \int_{y - \frac{b^2 c_x^2 \gamma \rho}{4} t}^{y} (h(s) - r(s)) ds.$$

The heavy-traffic limit for PRQ

- The PRQ also provides useful structural insights into the original stochastic model for different long-run traffic intensity ρ and cycle-length parameter γ.

Denote the instantaneous traffic intensity at a location y within a cycle by $\rho(y)$, let

$$\rho^{\uparrow} = \sup_{y}\{\rho(y)\}.$$

Three scenarios

1. Underloaded queues: $\rho^{\uparrow} < 1$.
2. Critically-loaded queues: $\rho^{\uparrow} = 1$.
3. Overloaded queues: $\rho^{\uparrow} > 1$.

We shall see that the space scaling needed are quite different in these cases and PRQ successfully captured this structure.

The heavy-traffic limit for PRQ - overloaded

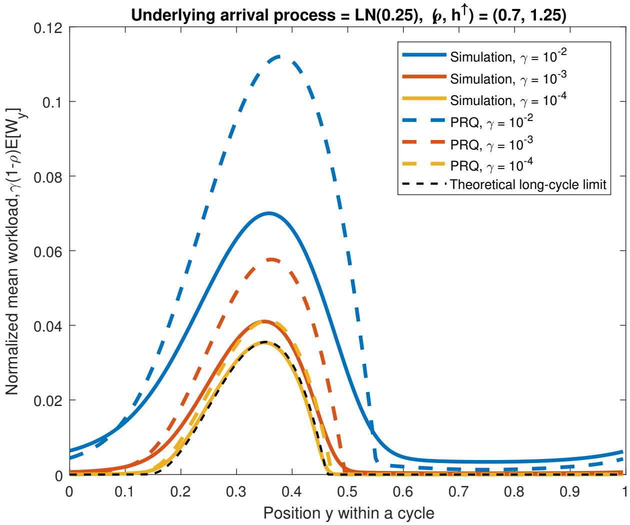
Theorem (long-cycle heavy-traffic limit for PRQ in an overloaded queue)

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

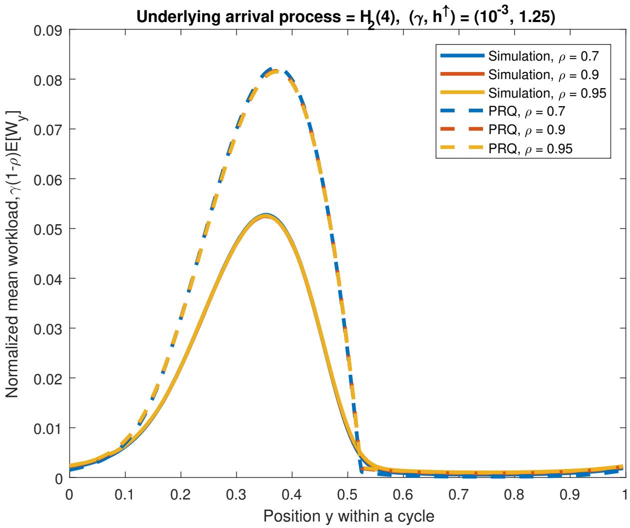
$$(1-\rho)\lim_{\gamma\downarrow 0}\gamma\cdot W^*_{\gamma,\rho,y} = \sup_{t\geq 0}\left\{-t+\int_{y-t}^{y}(h(s)-r(s))ds\right\}.$$

- 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 - overloaded



The heavy-traffic limit for PRQ - overloaded



The heavy-traffic limit for PRQ - underloaded

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

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

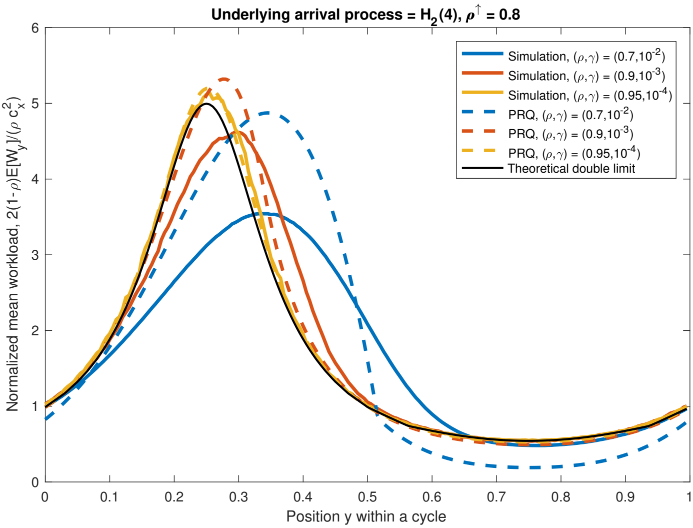
For $G_t/G_t/1$ periodic model with $\rho^{\uparrow} < 1$, PRQ is asymptotically correct as $(\gamma, \rho) \to (0, 1)$. Furthermore, we have the double limit for PRQ

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

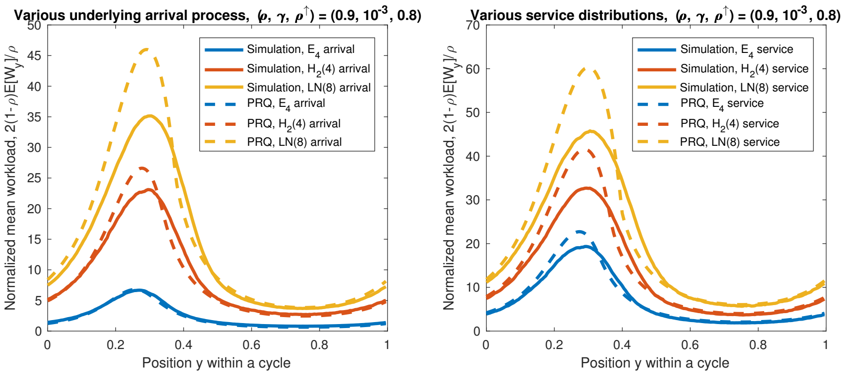
where $I_w(\infty) = c_x^2$ and $\rho(y)$ is the instantaneous traffic intensity.

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

The heavy-traffic limit for PRQ - underloaded



The heavy-traffic limit for PRQ - underloaded



- PRQ is very robust against different arrival and service distributions.

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: the space scaling depends on the detailed structure of the arrival-rate and service-rate function.

- For the original stochastic model: the scaling in the heavy-traffic FCLT is $\gamma^{p/(2p+1)}$, where p is obtained from Taylor's expansion, see Whitt (2016).
- What about PRQ? We get the same scaling!

The heavy-traffic limit for PRQ - critically-loaded

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

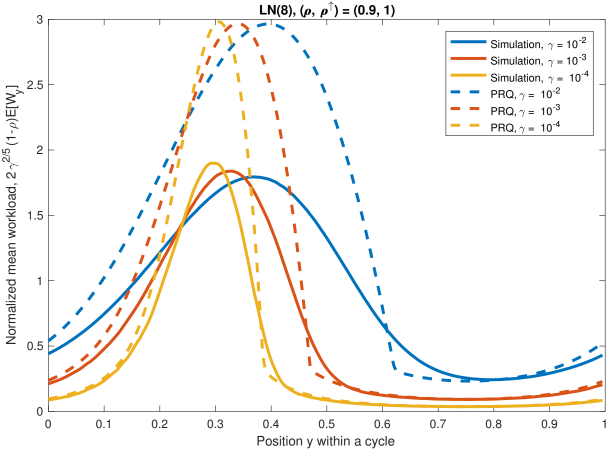
Assume that

$$h(t) - r(t) = 1 - ct^p + o(t^p), \tag{2}$$

for some real numbers $p \geq 0$. 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.

The heavy-traffic limit for PRQ - critically-loaded



- Arrival-rate function is a variant of $sin(x)$, which has power $p = 2$ for its first non-constant term in the Taylor's expansion. Thus $2/(2p + 1) = 2/5$ appears in the space scaling.

Thank you!

References

[WW14] W. Whitt, Heavy-Traffic Limits for Queues with Periodic Arrival Rates, *Operations Research Letters*, forthcoming, 2014.

[WW16] W. Whitt, Heavy-Traffic Limits for a Single-Server Queue Leading Up to a Critical Point, *Operations Research Letters*, 2016.

[WY16a] W. Whitt, and W. You, Using Robust Queueing to Expose the Impact of Dependence in Single-Server Queues, to appear in *Operations Research*, 2016. http://www.columbia.edu/ ww2040/allpapers.html.

[WY16b] W. Whitt, and W. You, Time-Varying Robust Queueing, submitted, 2016. http://www.columbia.edu/ ww2040/allpapers.html.

The long-cycle fluid limit

Furthermore, one can prove that the PRQ is asymptotically correct in the long-cycle fluid limit:

Theorem

For the periodic $G_t/GI_t/1$ model, PRQ with any b, $0 < b < \infty$, is asymptotically exact as $\gamma \downarrow 0$, i.e.,

$$\lim_{\gamma \to 0} \gamma W_{\gamma,\rho,y} \overset{w.p.1}{=} \lim_{\gamma \to 0} \gamma W^*_{\gamma,\rho,y} = \sup_{s \geq 0} \{\Lambda_{\gamma,\rho,y}(s) - M_{\gamma,\rho,y}(s)\}.$$

- A trivial limit of 0 if not overloaded.