Modeling and Analysis of Delay Announcement

Refined Approximation for Overloaded Systems

Junfei Huang

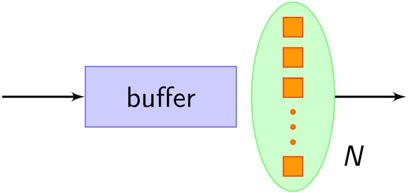
CUHK Business School

Joint work with Avi Mandelbaum, Hanqing Zhang, Jiheng Zhang

INFORMS 2013

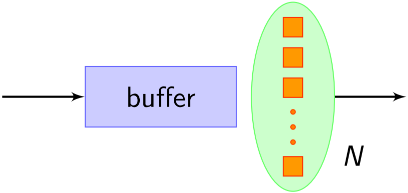
Model and Motivation

Many-server queue



Model and Motivation

Many-server queue

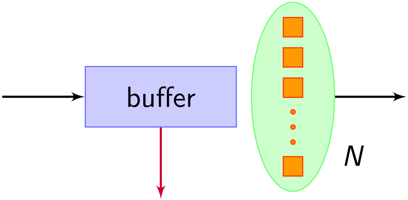


Motivation:

- Customer call centers and other services areas.

Model and Motivation

Many-server queue with abandonment



Motivation:

- Customer call centers and other services areas.

Literature Review

Many-server Queues

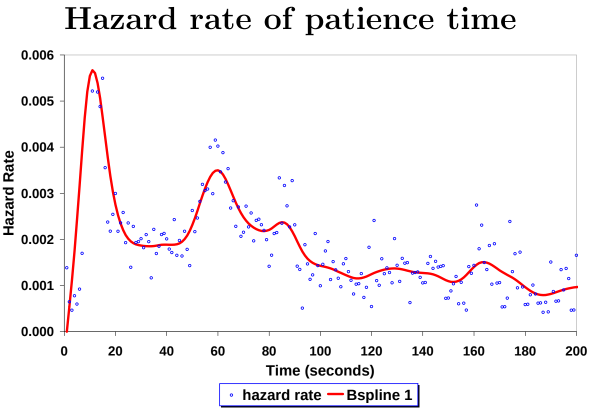
- Halfin and Whitt 1981 ($M/M/N$)
- Puhalskii and Reiman 2000 ($G/Ph/N$)
- Jelenković, Mandelbaum and Momčilović 2004 ($G/D/N$)
- Whitt 2005 ($G/H_2^*/n/m$)
- Garmarnik and Momčilović 2007 ($G/La/N$)
- Reed 2007, Puhalskii and Reed 2008 ($G/G/N$)
- Mandelbaum and Momčilović 2008 ($G/G/N$)
- Kaspi and Ramanan 2009, Kaspi 2009 ($G/G/N$)
- ......

Literature Review

Many-server Queues with Abandonment

- Whitt 2004 ($M/M/N + M$)
- Zeltyn and Mandelbaum 2005 ($G/M/N + G$)
- Whitt 2006 ($G/G/N + G$)
- Puhalskii 2008 ($M_t/M_t/N_t + M_t$)
- Kang and Ramanan 2008 ($G/G/N + G$)
- Mandelbaum and Momčilović 2009 ($G/G/N + G$)
- Dai, He and Tezcan 2009 ($G/Ph/N + G$)
- Bassamboo and Randhawa 2010 ($M/M/N + G$)
- Reed and Tezcan 2012 ($G/M/N + G$)
- Liu and Whitt 2013 ($G_t/M/N_t + G$)
- ......

Delay Announcement Story I: Announce While Waiting



- (Mandelbaum and Zeltyn: *Data-stories about (im)patient customers in tele-queues*, 2012; First observed in Brown et al. (JASA, 2005).)
- **Announcement can change customers' behavior!**
- **Should** the system make announcement? If yes, **when**?

Delay Announcement Story II: Announce Upon Arrival

- Discussed in Armony, Shimkin and Whitt (OR, 2009)
- Announcement upon arrival;
  - Announce the expected waiting time;
- An All-Exponential Model:
  - $M/M/n + GI$ **overloaded** systems ($\lambda = 140, \mu = 1, n = 100$);
  - Announce to wait $\omega = 0.224$: a proportion $e^{-\omega} = 0.2$ balk;
  - Hazard rate of the patience time distribution:

$$h(t) = \begin{cases} 0.5, & 0 \le t \le \omega; \\ 4, & t > \omega. \end{cases}$$

All-Exponential Model: A Question Left

- Fluid Approximation:

	Simulated	Fluid approximation
$\mathbb{E}[Q(\infty)]$	17.3	23.7
$\mathbb{E}[W(\infty)]$	0.157	0.225

- A well-known result:

Bassamboo and Randhawa (OR, 2010), Whitt (OR, 2006)

Under some regularity conditions, fluid approximation is very accurate for **overloaded** systems.

- **Why** is fluid approximation **not** so accurate here?
- Armony et al.: "*...remains a problem for future research*".

Theoretical Framework

- Diffusion approximation for **overloaded** systems:
  - $GI/M/n + GI$ under the ED+QED regime;
  - Approximated diffusion process: OU-type process (easy to use!);
- Why diffusion approximation for overloaded systems?
  - Why fluid approximation is not enough?
  - (Recall Bassamboo and Randhawa (OR 2010))
- Patience time distribution F^n (**Sensitivity** analysis):

$$\sqrt{n}\left(F^n(\omega + \frac{x}{\sqrt{n}}) - F^n(\omega)\right) \to f_\omega(x).$$

Asymptotic Regime

- Consider a sequence of $GI/M/s_n + G$ indexed by n.

$$\widetilde{\Lambda}^n(\cdot) \Rightarrow \widetilde{\Lambda}(\cdot) \quad \text{with} \quad \widetilde{\Lambda}^n(t) = \frac{1}{\sqrt{\lambda_n}} \left(\Lambda^n(t) - \lambda_n t \right),$$

$$\frac{\lambda_n}{s_n \mu} \to \rho > 1,$$

$$\frac{\lambda_n F_c^n(\omega) - s_n \mu}{\sqrt{\lambda_n}} \to \beta.$$

- Denote by $V^n(t)$ the virtual waiting time in the nth system, and ω its fluid limit. Define

$$\widetilde{V}^n(t) = \sqrt{\lambda_n}\Big(V^n(t) - \omega\Big).$$

Main Results

Theorem

If

$$\sqrt{\lambda_n}\Big(V^n(0) - \omega\Big) \Rightarrow \widetilde{V}_0,$$

then as $n \to \infty$*,*

$$\widetilde{V}^n \Rightarrow \widetilde{V},$$

here $\widetilde{V}$ *is the unique solution to*

$$\widetilde{V}(t) = \widetilde{V}_0 - \rho \int_0^t \Big[f_\omega(\widetilde{V}(x)) - \beta\Big]\,\mathrm{d}x + \Big[\widetilde{\Lambda}(t) - \sqrt{2\rho - 1}\mathcal{B}(t)\Big].$$

The density function of its steady state $\pi(\cdot)$ *is given by*

$$\pi(y) = C \exp\left(-\frac{2\rho}{\sigma^2}\int_0^y [f_\omega(x) - \beta]\,\mathrm{d}x\right),$$

where C *is the scaling factor.*

With this, we can easily get the convergence of queue length process.

System Performance Evaluation

For a given $GI/M/s+GI$ system with parameters $(\lambda, \theta^2, \mu, s, F)$:

$$
\begin{aligned}
\rho &:= \frac{\lambda}{s\mu}, \\
\sigma^2 &:= \theta^2 + 2\rho - 1, \\
\hat{\beta} &:= \frac{\lambda F_c(\omega) - s\mu}{\sqrt{\lambda}}, \\
\hat{f}_\omega(x) &:= \sqrt{\lambda}\Big[F(\omega + \frac{x}{\sqrt{\lambda}}) - F(\omega)\Big].
\end{aligned}
$$

Here ω is the solution to $F_c(\omega) = \dfrac{1}{\rho}$.

Approximation Formulae

- Queue Length:

$$\mathbb{E}[Q(\infty)] \approx \lambda \int_0^{\omega} F_c(x)\mathrm{d}x + \frac{1}{\rho}\sqrt{\lambda}\int_{-\infty}^{\infty} x\pi(\mathrm{d}x).$$

  Here

$$\pi(x) = C\exp\left(-\frac{2\rho}{\sigma^2}\int_0^x \left[\hat{f}_\omega(v) - \hat{\beta}\right]\mathrm{d}x\right).$$

- Probability of the waiting time $W(\infty) = V(\infty) \wedge U$, here U is the patience time:

$$\begin{aligned}\mathbb{P}(W(\infty) > y) &= F_c(y)\mathbb{P}\left(\sqrt{\lambda}(V(\infty) - \omega) > \sqrt{\lambda}(y - \omega)\right) \\ &\approx F_c(y)\int_{\sqrt{\lambda}(y-\omega)}^{\infty} C\exp\left(-\frac{2\rho}{\sigma^2}\int_0^u \left[\hat{f}_\omega(v) - \hat{\beta}\right]\mathrm{d}v\right)\mathrm{d}u.\end{aligned}$$

Example 1: Revisit Armony, Shimkin and Whitt (2009)

Customers' behavior in response to call announcement (upon arrival)

$$h(x|\tau) = \begin{cases} h_0, & x \le \tau; \\ h_1, & x > \tau. \end{cases}$$

Performances	Fluid	$h_1 = 0.5$		$h_1 = 4$	
		Simulated	Diffusion	Simulated	Diffusion
$\mathbb{E}[Q(\infty)]$	23.7	24.3	23.7	17.3	16.4
$\mathbb{E}(W(\infty); B_c)$	0.212	0.217	0.224	0.155	0.151
$\mathbb{P}(W(\infty) < \omega_e \vert S)$	1	0.512	0.512	0.754	0.756

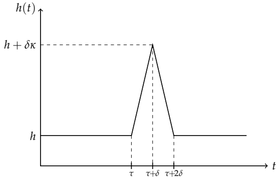
Table: Fluid v.s. Diffusion with $h_0 = 0.5$.

Example 2: When to Announce (While Waiting)?

Optimization problem:

$$\min_{s,\tau} \quad s$$
$$\text{s.t.} \quad \mathbb{P}\left(W(\infty) > w \wedge \tau\right) \le \alpha.$$

Hazard rate of the patience time distribution:



Proposition

The optimal announcement time τ^λ converges, as $\lambda \to \infty$, to the offered waiting time in the following way

$$\sqrt{\lambda}(\omega - \tau^\lambda) \to 0.$$

With this proposition, the staffing level can be reduced by $\mathcal{O}(\sqrt{\lambda})$, comparing to the situation without announcement.

Example 3: Hazard-rate Scaling

Example

Customers arrive according to Poisson process with rate λ. Service times are exponentially distributed with rate μ. The hazard rate of the patient time is as follows

$$\tilde{h}(x) = \begin{cases} h_0, & x \leq \omega; \\ h_0 + \kappa(x - \omega), & x > \omega. \end{cases}$$

Here ω is the offered waiting time.

The regular conditions in Bassamboo and Randhawa (2010) are satisfied.

Example 3: Hazard-rate Scaling

- Queue length:

Servers	Fluid	$\kappa = 20$			$\kappa = 100$		
		Simulated	Appr. G	Appr. H	Simulated	Appr. G	Appr. H
20	4	3.34 ± 0.02	3.1599	2.6690	2.70 ± 0.01	2.4354	1.9113
50	10	8.65 ± 0.04	8.7328	8.2553	7.48 ± 0.04	7.025	7.0144
100	20	18.04 ± 0.05	18.3797	17.9050	16.30 ± 0.04	1.6151	16.1387
200	40	37.56 ± 0.06	38.0092	37.5344	35.11 ± 0.06	35.5413	35.0705
400	80	77.20 ± 0.08	77.9364	77.1579	73.84 ± 0.07	74.2668	73.7983

- $\mathbb{P}(W_\infty > \omega)$: (if $F^n = F$, then the approximation is 0.4167)

Servers	$\kappa = 20$			$\kappa = 100$		
	Simulated	Appr. G	Appr. H	Simulated	Appr. G	Appr. H
20	$.35785 \pm .00143$	.3576	.3271	$.28828 \pm .00122$	.2879	.2493
50	$.36396 \pm .00210$	.3641	.3461	$.29371 \pm .00172$	.2938	.2723
100	$.37122 \pm .00188$	.3712	.3589	$.30348 \pm .00142$	.3037	.2895
200	$.37858 \pm .00175$	.3787	.3703	$.31552 \pm .00142$	.3157	.3062
400	$.38652 \pm .00124$	.3859	.3801	$.32922 \pm .00102$	.3286	.3221

- The **convergence rate** of fluid approximation may be very slow! Thus we need a refined (diffusion) approximation!

Questions?

Thank you!