Distributional Sensitivity in Many-Server Queues

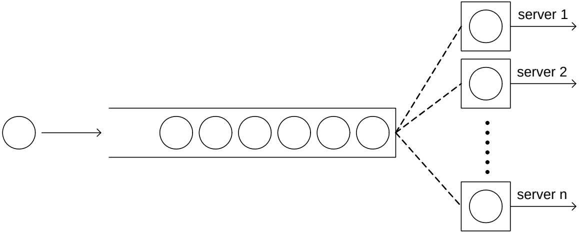
Jim Dai

Georgia Institute of Technology
The H. Milton Stewart School of Industrial and Systems Engineering

March 18, 2010

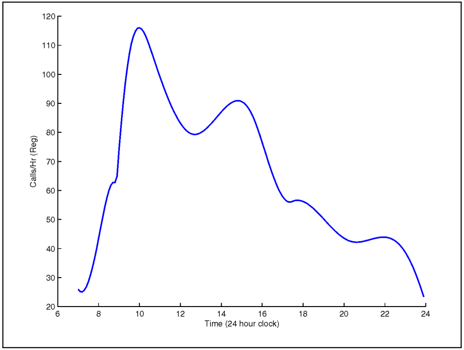
Joint work with Shuangchi He and Tolga Tezcan (UIUC → Rochester)

$G/GI/n + GI$ model

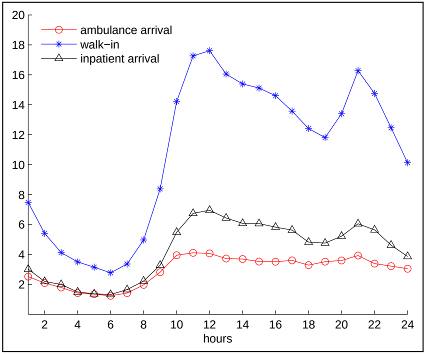


- iid service times and iid patience times
- first-in-first-out (FIFO) queue
- the number of servers n is large: call centers, web server farms, hospital beds

Time-varying arrival rates



Brown et al (05)



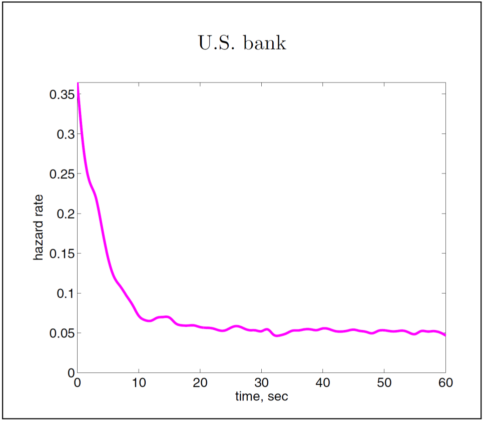
Arrivals to a hospital emergency room

Customer abandonment

Garnett, Mandelbaum & Reiman (02)

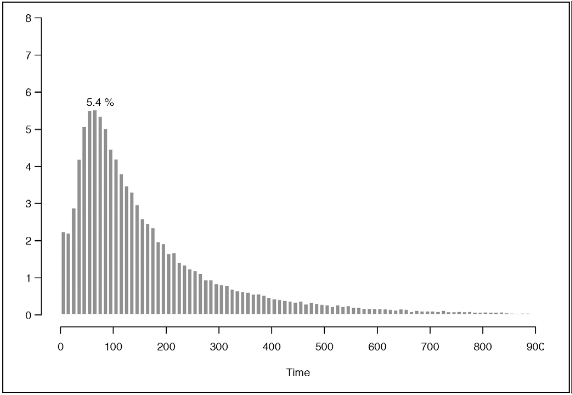
"…. There is a significant difference in the distributions of waiting time and queue length—in particular, the average waiting time and queue length are both strikingly shorter when abandonment is taken into account."

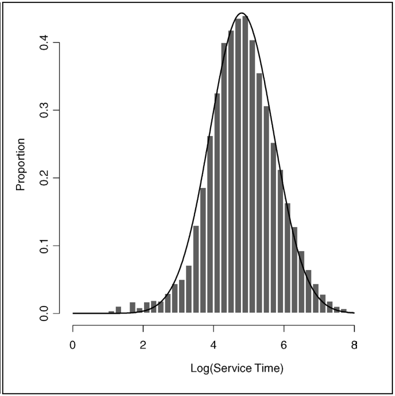
- one must model abandonment
- possibly non-exponential patience-time distribution



Mandelbaum & Zeltyn (04)

Non-exponential service-time distribution





Brown et al (2005)

Do distributions matter?

Many-server asymptotic regimes: the number of servers n is large; the call volume is high; a small to moderate fraction of customers abandon.

- critically-loaded: quality- & efficiency-driven (QED) regime, Halfin-Whitt regime
- underloaded: quality-driven (QD) regime

Distributions:

- patience-time distribution
- service-time distribution

Sensitivity on patience-time distribution F

- $M/\text{LogNormal}/100 + GI$: $\lambda = 105$, $\mu = 1$, $\sigma_s^2 = 4$, $(105 - 100)/105 = 4.76\%$
- three patience-time distributions:
  - exponential
  - uniform
  - hyperexponential
- $\alpha = F'(0+)$ is fixed;
- hyper-exponential (H_2) patience-time distribution

$$X = \begin{cases} \text{Exp}(79\alpha/30) & \text{with probability } 0.3, \\ \text{Exp}(0.3\alpha) & \text{with probability } 0.7. \end{cases}$$

Sensitivity on F with fixed $\alpha = F'(0)$

M/LogNormal/100 + GI: $\lambda = 105$, $(105 - 100)/105 = 4.76\%$

$\alpha \backslash F$	Exp	Uniform	H_2
	Abandonment probability		
$\alpha = 0.1$	0.0528	0.0530	0.0526
$\alpha = 1$	0.0701	0.0706	0.0693
$\alpha = 10$	0.0893	0.0907	0.0877
	Average queue length		
$\alpha = 0.1$	55.46	53.40	57.95
$\alpha = 1$	7.357	6.819	8.048
$\alpha = 10$	0.9373	0.7570	1.189

Sensitivity on F with mean m fixed

$$M/\text{LogNormal}/100 + GI\text{: } \lambda = 105,\ \mu = 1,\ \sigma_s^2 = 4$$

$m \backslash F$	Exp	Uniform	H_2
	Abandonment probability		
$m = 0.1$	0.0893	0.0851	0.0930
$m = 1$	0.0701	0.0645	0.0752
$m = 10$	0.0528	0.0499	0.0582
	Average queue length		
$m = 0.1$	0.9373	1.516	0.5882
$m = 1$	7.357	12.69	4.500
$m = 10$	55.46	99.77	26.54

- Mean patience time m is a wrong statistics

Replacing $G/GI/n+GI$ by $G/GI/n+M$ when $\alpha > 0$

INSIGHT

For $G/GI/n+GI$ queues in the QD/QED regime, it is generally accurate to replace the patience-time distribution F with an exponential distribution having rate $\alpha = F'(0+)$.

- Numerical algorithms such as the matrix-analytic method benefit from such a replacement; e.g., $G/Ph/n+M$ systems can be used to approximate $G/Ph/n+GI$ systems.
- Dynamic control problem can be simplified by taking advantage of the exponential patience-time distribution.
- Justifications are carried out through many-server heavy traffic limits.

Many-server asymptotic framework

- Number of servers n goes to infinity.
- Consider a sequence of $G/GI/n + GI$ queues indexed by n.
- The arrival process E^n has arrival rate λ^n that depends on n:

$$\lambda^n \approx n\lambda \quad \text{for some } \lambda > 0;$$

  $E^n(t)$ is the cumulative number of arrivals in $(0, t]$.
- The patience-time distribution F is independent of n; $F(0) = 0$ and $\alpha = F'(0)$ exists.
- The service-time distribution H is independent of n; it has finite mean $1/\mu$.
- $\rho = \lambda/\mu$; $\rho = 1$ QED or Halfin-Whitt regime; $\rho < 1$ QD .
- We assume $\rho = 1$.

Assumptions on the arrival process

- Fluid-scaling

$$\bar{E}^n(t) = \frac{1}{n} E^n(t) \quad t \geq 0.$$

- Functional weak law of large numbers (FWLLN): Assume that

$$\bar{E}^n \Rightarrow \bar{E}, \tag{1}$$

and that $\bar{E}(t) = \lambda t$ for some $\lambda > 0$. Let $\rho = \lambda/\mu$ be the traffic intensity.

- Diffusion-scaling

$$\tilde{E}^n(t) = \frac{1}{\sqrt{n}} \hat{E}^n(t) \quad \text{and} \quad \hat{E}^n(t) = E^n(t) - n\bar{E}(t) \quad \text{for } t \geq 0.$$

- Functional Central Limit Theorem (FCLT): Assume that

$$\tilde{E}^n \Rightarrow \tilde{E} \qquad \text{as } n \to \infty. \tag{2}$$

Here, we assume $\tilde{E}$ is a $(-\beta, \lambda c^2)$-Brownian motion.

Phase-type service time distributions (p, P, ν)

Definition (Neuts 1981)

A phase-type random variable is defined to be the time until absorption of a transient continuous time Markov chain.

- transient states $\mathcal{K} = \{1, \ldots, K\}$, $K + 1$ absorbing state
- initial distribution p on $\mathcal{K}$
- ν_k the rate at state (phase) $k \in \mathcal{K}$
- $P = (P_{k\ell})$ the transition probabilities on transient states $\mathcal{K}$; $I - P$ is assumed to be invertible
- Let m be the mean service time, and

$$\gamma = \frac{\text{diag}(1/\nu)\big(I + P' + (P')^2 + \ldots\big)p}{m}. \tag{3}$$

Then γ_k is interpreted as the fraction of load from phase k customers.

An example of phase-type distributions

- Two-stage hyperexponential distribution $H_2(\nu_1, \nu_2, p_1, p_2)$

$$\xi = \begin{cases} \exp(\nu_1) & \text{with probability } p_1 \\ \exp(\nu_2) & \text{with probability } p_2 \end{cases},$$

$$\mathcal{K} = \{1, 2\}, \quad p = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}, \quad \nu = \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}, \quad P = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

- Mean service time $m = p_1/\nu_1 + p_2/\nu_2$; mean service rate $\mu = 1/m$.
- Fraction of phase k load

$$\gamma_k = \frac{p_k/\nu_k}{m}, \quad \gamma_1 + \gamma_2 = 1, \qquad \gamma_k \nu_k = \mu p_k.$$

Justification: diffusion limits for $G/Ph/n + GI$ queues

Assume a phase-type service-time distribution with parameter (p, P, ν).

Let $Y_k^n(t)$, $k = 1, \dots, K$, be the number of phase k customers in system at time t and

$$\tilde{Y}_k^n(t) = (Y_k^n(t) - n\gamma)/\sqrt{n},$$

where $\gamma = \mu R^{-1} p$ and R is a $K \times K$ matrix given by $R = (I - P')\text{diag}(\nu)$.

Theorem (Dai, He & Tezcan 09)

Under some initial conditions, $\tilde{Y}^n \Rightarrow \tilde{Y}$ as $n \to \infty$. The process $\tilde{Y}$ satisfies

$$\tilde{Y}(t) = \tilde{W}(t) - R \int_0^t \tilde{Y}(s)\, ds + (R - \alpha I)p \int_0^t (e' \tilde{Y}(s))^+ \, ds,$$

where $\tilde{W}$ is a K-dimensional Brownian motion and e is the K-dimensional vector of ones.

Puhalskii & Reiman (00) for $G/Ph/n$ queues

The piecewise OU process $\tilde{Y}$

- Let $R = (I - P')\text{diag}(\nu)$. Recall that $\alpha = F'(0)$. The map $\Phi : x \in \mathbb{D}^K \to y \in \mathbb{D}^K$ is well defined via

$$y(t) = x(t) - R\int_0^t y(s)\,ds + (R - \alpha I)p\int_0^t (e'y(s))^+\,ds.$$

  Massey-Mandelbaum-Reiman (98)
- $\tilde{Y} = \Phi(B)$, where B is some K-dimensional Brownian motion.
- When $K = 1$,

$$\begin{aligned} y(t) &= x(t) - \mu\int_0^t y(s)\,ds + (\mu - \alpha)\int y(s)^+\,ds \\ &= x(t) + \mu\int_0^t y(s)^-\,ds - \alpha\int y(s)^+\,ds \end{aligned}$$

Justification: marginal limits for $G/GI/n + GI$ queues

Let $X^n(t)$ be the number of customers in system at time t and

$$\tilde{X}^n(t) = (X^n(t) - n)/\sqrt{n}.$$

Let H be the service-time distribution and $H_e(x) = \mu \int_0^x (1 - H(u))\, du$ be the equilibrium distribution of H.

Theorem (Mandelbaum & Momčilović 09)

Under some initial conditions, $\tilde{X}^n \Rightarrow \tilde{X}$ as $n \to \infty$. The process $\tilde{X}$ satisfies

$$\tilde{X}(t) = \tilde{Z}(t) + \int_0^t \tilde{X}(t-s)^+ \, dH(s) - \frac{\alpha}{\mu} \int_0^t \tilde{X}(t-s)^+ \, dH_e(s),$$

for some stochastic process $\tilde{Z}$.

Reed (09) for $G/GI/n$ queues

Measure-valued diffusion limits for $G/GI/n + GI$ queues

- Kaspi & Ramanan (09) for $G/GI/n$ queues
- A key tool: An asymptotic relationship between abandonment processes and queue length processes.

An asymptotic relationship

For the nth system in a sequence of $G/G/n+GI$ queues, let $A^n(t)$ be the number of abandonments by time t, and $Q^n(t)$ be queue length at time t.

Theorem (Dai & He (09))

Under some conditions, for each $T > 0$,

$$\frac{1}{\sqrt{n}} \sup_{0 \le t \le T} \left| A^n(t) - \alpha \int_0^t Q^n(s)\, ds \right| \to 0 \quad \text{in probability as } n \to \infty. \tag{4}$$

- A key assumption: stochastic boundedness for diffusion-scaled queue-length processes, i.e., for each $T > 0$,

$$\lim_{a \to \infty} \limsup_{n \to \infty} \mathbb{P}\left[\frac{1}{\sqrt{n}} \sup_{0 \le t \le T} Q^n(t) > a \right] = 0.$$

- The relationship holds for time-nonhomogeneous arrival processes.

A modularized approach to proving limit theorems

The asymptotic relationship suggests the following framework:

- Prove a limit theorem for queues without abandonment, using a continuous-mapping approach.
- Compare queues with abandonment and corresponding queues without abandonment to prove the stochastic boundedness of the diffusion-scaled queue-length processes.
- Apply a modified map to prove a corresponding limit theorem for queues with abandonment.

Estimating patience-time density α at zero

The asymptotic relationship suggests the following estimator: fix a $T > 0$,

$$\hat{\alpha}^n = \frac{A^n(T)}{\int_0^T Q^n(t)\,dt}.$$

- Customers who get into service have never abandoned the system and their patience times have never been observed. Thus, it is difficult to estimate the entire patience-time distribution.
- For queues in QD/QED regime, the patience-time density α at zero, rather than the entire patience-time distribution, dictates the system performance.

Consistent estimator for α

Theorem

Assume that $\lim_{n\to\infty} \mathbb{P}[\inf_{0\le t\le T} Q^n(t)/\sqrt{n} > \varepsilon] = 1$ *for some* $\varepsilon > 0$. *Then,* $\hat{\alpha}^n$ *is a consistent estimator in the sense that*

$$\hat{\alpha}^n \to \alpha \quad \text{in probability as } n \to \infty.$$

For each fixed n, $\hat{\alpha}^n$ is biased.

Consider $M(t)/GI/n(t) + GI$ queues with $\alpha = 6$ and

- time-varying arrival rate per hour

$$\lambda(t) = 1000 + 100t + 2400\sin(\pi t/12)$$

  for $0 \le t \le 12$.

- time-varying staffing level

$$n(t) = \begin{cases} 225 & 0 \le t \le 3 \\ 310 & 3 < t \le 9 \\ 275 & 9 < t \le 12 \end{cases}$$

- a lognormal service time distribution with mean 5 min and variance 10 min^2.

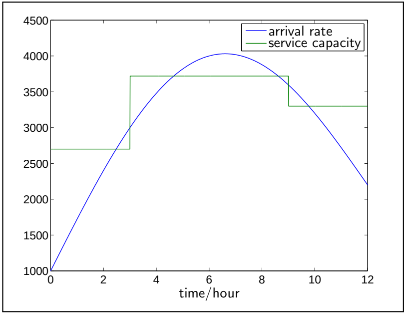


FIGURE: Arrival rate vs. service capacity.

Consistent estimator $\hat{\alpha}^n$: an example

$s \backslash F$	Exp			Uniform			H_2		
	$A^n(T)$	$\int_0^T Q^n$	$\hat{\alpha}^n$	$A^n(T)$	$\int_0^T Q^n$	$\hat{\alpha}^n$	$A^n(T)$	$\int_0^T Q^n$	$\hat{\alpha}^n$
1	1227	194.7	6.30	1187	202.2	5.87	1235	220.3	5.61
2	1128	195.1	5.78	1149	185.5	6.20	1141	194.9	5.86
3	902	150.4	6.00	926	152.5	6.07	906	156.0	5.81
4	1512	246.7	6.13	1520	241.5	6.30	1526	269.7	5.66
5	1397	234.3	5.97	1398	218.1	6.41	1395	248.3	5.62

A linear relationship in real-world call centers

The asymptotic relationship also suggests

$$\frac{A^n(T,\omega)}{T} \approx \frac{\alpha}{T}\int_0^T Q^n(s,\omega)ds \quad \text{for } T > 0. \tag{5}$$

- Mandelbaum & Zeltyn (04) proved that for $M/M/n + GI$ queues in QED regime,

$$\text{long-run abandonment rate} = \alpha \times \text{the average queue length.} \tag{6}$$

- Among a large number of data sets from call centers, there is a linear relationship between the abandonment rate and the steady-state queue length.
- It is (5), not (6), that explains this observation.

Sensitivity on service-time distributions

Two $M/H_2/100+M$ queues:

- Both have $\lambda = 110$, $\mu = 1$, $\alpha = 0.5$, $c_s^2 = 8$.
- The H_2 service distributions have $\gamma_1 = pm_1 = 0.1$ ($p = 0.8195, m_1 = 0.122, m_2 = 4.986$) and $\gamma_1 = 0.5$ ($p = 0.941, m_1 = 0.53, m_2 = 8.47$), respectively.

By the matrix-analytic method,

$$\mathbb{P}[Q_1 > 50] = 13.27\%,$$
$$\mathbb{P}[Q_2 > 50] = 7.67\%.$$

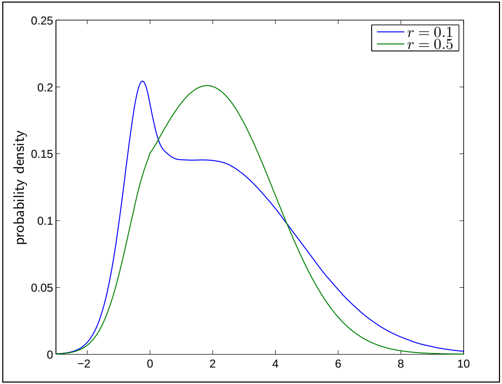


FIGURE: Steady-state distributions of diffusion limits for two $M/H_2/100+M$ queues.

Diffusion approximation via finite-element method

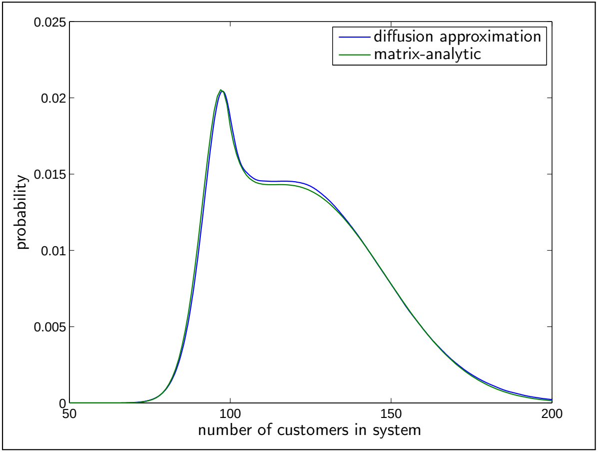


FIGURE: Steady-state distribution of an $M/H_2/100 + M$ queue with $\lambda = 110$, $\mu = 1$, $\alpha = 0.5$, $\gamma_1 = 0.1$ and $c_s^2 = 8$.

Weak invariance on service-time distributions

- Heavy-traffic limits for single-server queues or queues with a small number of servers depend only on the first two moments of the service-time distribution.
- Heavy-traffic limits for many-server queues depend on the entire service-time distribution.
- Many-server queues and single-server queues are qualitatively different.

Conjecture

Consider a sequence of $G/GI/n + GI$ queues in the QED regime. Under some initial conditions, $\tilde{X}^n(\infty) \Rightarrow \tilde{X}(\infty)$ as $n \to \infty$ and

$$\lim_{x \to \infty} \frac{1}{x^2} \log \mathbb{P}[\tilde{X}(\infty) > x] = -\frac{\alpha}{\mu(c_a^2 + c_s^2)}.$$

Weak invariance for $G/GI/n$ queues

Theorem (Gamarnik & Goldberg 2009)

For $G/GI/n$ queues in QED,

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbb{P}[\tilde{X}(\infty) > x] = -\frac{1}{\mu(c_a^2 + c_s^2)}.$$

Gamarnik & Momčilović (07) for lattice service-time distribution.

Implication of weak invariance: computing π

- Stationary density π satisfies the basic adjoint relationship (BAR)

$$\int_{\mathbb{R}^K} Gf(x)\pi(x)\,dx = 0 \quad \text{for all } f \in C_b^2(\mathbb{R}^K);$$

  see Dai and Harrison (92) for reflecting Brownian motions.
- Using a reference density $d : \mathbb{R}^K \to \mathbb{R}_+$, we compute the ratio $r(x) = \pi(x)/d(x)$ and obtain π by $\pi(x) = r(x)d(x)$.
- The algorithm is sensitive to the choice of $d(x) = d_1(x)d_2(x)$;

$$d_i(x) = \begin{cases} c_1\phi(\sqrt{2}(x+\beta)(1+c_a^2)^{-1/2}) & x < 0, \\ c_2\phi(\sqrt{2\alpha/\mu}(x+\mu\beta/\alpha)(c_s^2+c_a^2)^{-1/2}) & x \geq 0, \end{cases} \qquad (7)$$

  where c_1 and c_2 are positive constants that make d_i continuous at zero.
- Using a finite-element algorithm to compute $r(x)$

Importance of choosing a right reference density

Recall the $M/H_2/100 + M$ queue with

- $\lambda = 110$, $\mu = 1$, $\alpha = 0.5$, $\gamma_1 = 0.1$ and $c_s^2 = 8$.

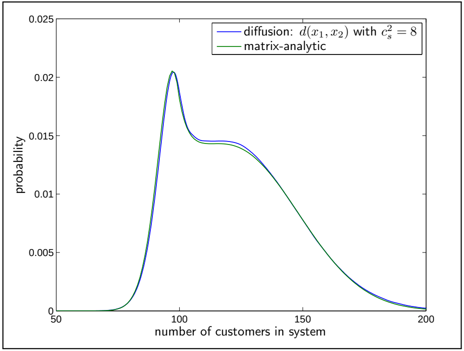


FIGURE: Gaussian reference density with $c_s^2 = 8$.

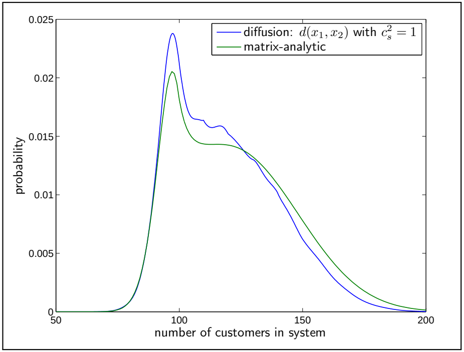


FIGURE: Gaussian reference density with $c_s^2 = 1$ (obtained from $M/M/100 + M$).

Hazard-rate scaling asymptotics

When $\alpha = 0$, if we replace $+GI$ by $+M$,

- it leads to a $G/GI/n$ queue without abandonment, possibly unstable;
- diffusion approximations based on (4) may not capture the original queue with abandonment.

We need to consider the patience-time distribution in a neighborhood of zero, rather than the origin itself. Inspired by Reed & Tezcan (09), let patience distributions depend on n with

$$F^n(x) = 1 - e^{-\int_0^x h(\sqrt{n}u)\,du}, \qquad \text{for } x \geq 0.$$

Conjecture

Under some conditions, for each $T > 0$,

$$\sup_{0 \le t \le T} \left| \frac{1}{\sqrt{n}} A^n(t) - \int_0^t \int_0^{Q^n(s)/\sqrt{n}} h(u)\,du\,ds \right| \to 0 \quad \text{in probability as } n \to \infty.$$

Hazard-rate scaling heavy-traffic limits

For $G/Ph/n + GI$ queues, recall that the K-dimensional vector $\tilde{Y}^n$ represents the number of customers in each phase in diffusion scaling.

Theorem

Under some initial conditions, $\tilde{Y}^n \Rightarrow \tilde{Y}$ as $n \to \infty$. The hazard-rate scaling diffusion process $\tilde{Y}$ satisfies

$$\tilde{Y}(t) = \tilde{W}(t) - R \int_0^t (\tilde{Y}(s) - p(e'\tilde{Y}(s))^+)\, ds - p \int_0^t \int_0^{(e'\tilde{Y}(s))^+} h(u)\, du\, ds,$$

where $\tilde{W}$ is a K-dimensional Brownian motion.

Hazard-rate scaling diffusion approximation

Consider an $M/H_2/500 + E_2$ queue with $\lambda = 522.4$ and $\mu = 1$.

- the H_2 service-time distribution is given by

$$X = \begin{cases} \text{Exp}(2.2) & \text{with probability } 0.4, \\ \text{Exp}(0.2) & \text{with probability } 0.6. \end{cases}$$

- the Erlang (E_2) patience-time distribution has $\alpha = 0$ and mean $m = 0.2$

	Abandonment probability	Average queue length	Average busy servers
Simulation	0.05025	13.90	496.1
Diffusion	0.05012	13.65	496.2

Summary

- The system performance is insensitive to the patience-time distribution as long as $\alpha = F'(0)$ is fixed and positive.
- The system performance critically depends on α; an consistent estimator of α is given.
- Many-server heavy traffic diffusion limits provide justification for replacing $+GI$ by $+M$
- Weak invariance on service-time distribution is conjectured.
- The conjectured decay rate plays a key role in choosing a right reference density for the finite-element algorithm.
- The hazard-rate diffusion limit promises a refined theory and improved performance estimates.

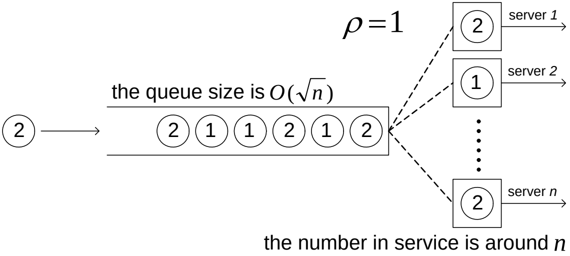
Surveys and references

1. J. G. Dai and S. He (2009), "Customer abandonment in many-server queues," *Mathematics of Operations Research*, to appear. H
2. J. G. Dai, S. He, and T. Tezcan (2009), "Many-server diffusion limits for $G/Ph/n+GI$ queues," *Annals of Applied Probability*, to appear.
3. S. Zeltyn and A. Mandelbaum (2005). Call centers with impatient customers: many-server asymptotics of the $M/M/n+G$ queue, *Queueing Systems*, **51** 361–402.
4. A. Mandelbaum and P. Momčilović (2009), "Queues with many servers and impatient customers," preprint.
5. Gans-Koole-M (03), Telephone call centers: Tutorial, review, and research prospects, *M&SOM*, **5**, 79-141.
6. Mandelbaum (06), Call centers: research bibliography with abstracts; `http://iew3.technion.ac.il/serveng/References/US7_CC_avi.pdf`

Part II: Proofs

Dai, He and Tezcan (2009), Many-Server Diffusion Limits for $G/Ph/n + GI$ Queues.

Scaling for $G/Ph/n + GI$ queues: $\rho = 1$



- $Z_k^n(t)$ the number of phase k customers in service, $X^n(t)$ in system, $Q^n(t)$ in queue, $W^n(t)$ workload; centering

$$\hat{X}^n(t) = X^n(t) - n, \quad \hat{Z}_k^n(t) = Z_k^n(t) - \gamma_k n.$$

- Diffusion-scaling

$$\tilde{X}^n(t) = \frac{1}{\sqrt{n}} \hat{X}^n(t), \quad \tilde{Z}_k^n(t) = \frac{1}{\sqrt{n}} \hat{Z}_k^n(t).$$

$$\tilde{Q}^n(t) = \frac{1}{\sqrt{n}} Q^n(t), \quad \tilde{W}^n(t) = \sqrt{n} W^n(t).$$

Critically loaded $G/Ph/n+GI$ queues: $\rho = 1$

Theorem (Dai-He-Tezcan 09)

Assume that $F(0) = 0$ and that $\alpha = F'(0)$ exists. Suppose that $(\tilde{X}^n(0), \tilde{Z}^n(0)) \Rightarrow (\xi, \eta)$. Then

$$(\tilde{Q}^n, \tilde{W}^n, \tilde{X}^n, \tilde{Z}^n) \Rightarrow (\tilde{Q}, \tilde{W}, \tilde{X}, \tilde{Z}),$$

where $(\tilde{X}, \tilde{Z})$ is a $(K+1)$-dimensional (degenerate) continuous Markov process, and

$$\tilde{Q}(t) = (\tilde{X}(t))^+ \text{ and } \tilde{W}(t) = \frac{1}{\mu}\tilde{Q}(t) \quad \text{(state space collapse).}$$

Furthermore, letting

$$\tilde{Y}(t) = p\tilde{Q}(t) + \tilde{Z}(t),$$

$\tilde{Y}$ is a K-dimensional piecewise Ornstein-Uhlenbeck (OU) process.

Puhalskii-Reiman (00) for $G/Ph/n$, Garnett-M-Reiman (02) for $M/M/n+M$

The piecewise OU process $\tilde{Y}$

- Let $R = (I - P')\text{diag}(\nu)$. Recall that $\alpha = F'(0)$. The map $\Phi : x \in \mathbb{D}^K \to y \in \mathbb{D}^K$ is well defined via

$$y(t) = x(t) - R \int_0^t y(s)\, ds + (R - \alpha I)p \int_0^t (e'y(s))^+ \, ds.$$

  Massey-Mandelbaum-Reiman (98)
- $\tilde{Y} = \Phi(B)$, where B is some K-dimensional Brownian motion.
- One can recover $(\tilde{X}, \tilde{Z})$ via

$$\tilde{X}(t) = e'\tilde{Y}(t) \quad \text{and} \quad \tilde{Z}(t) = \tilde{Y}(t) - p(\tilde{X}(t))^+, \quad t \geq 0.$$

Two-dimensional piecewise OU process

- Assume service time distribution is $H_2(\nu_1, \nu_2, p_1, p_2)$.
- For each $(x_1, x_2) \in \mathbb{D}^2$, there is a unique $(y_1, y_2) \in \mathbb{D}^2$ such that for $k = 1, 2$,

$$y_k(t) = x_k(t) - \nu_k \int_0^t y_k(s)ds + (\nu_k - \alpha)p_k \int_0^t (y_1(s) + y_2(s))^+ \, ds.$$

- The map $\Phi : x \in \mathbb{D}^2 \to y \in \mathbb{D}^2$ is well defined.
- When B is a 2-d Brownian motion with drift $-\beta p$ and covariance matrix

$$\mu \begin{bmatrix} p_1 \left(p_1 c^2 - p_1 + 2\right) & p_1 p_2 \left(c^2 - 1\right) \\ p_1 p_2 \left(c^2 - 1\right) & p_2 \left(p_2 c^2 - p_2 + 2\right) \end{bmatrix}.$$

$\tilde{Y} = \Phi(B)$ is the 2-d piecewise OU process that serves as the diffusion limit.

Diffusion approximation: $M/H_2/200 + M$

- $H_2(1/2.2, 1/.2, .4)$ service time distribution and $\alpha = F'(0) = 2/3$.
- Finite element method to solve the stationary distribution of $\tilde{Y}$; Dai-Harrison (92), Shen-Chen-Dai-Dai (02); reference density

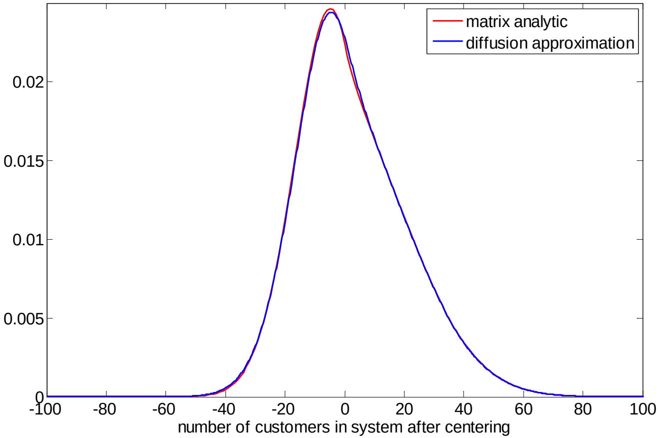
$$f(x_1, x_2) = \frac{1}{4} e^{-(x_1^2 + x_2^2)/4};$$

truncate the area $(-8, 14) \times (-8, 14)$; the grid consists of 1×1 squares.

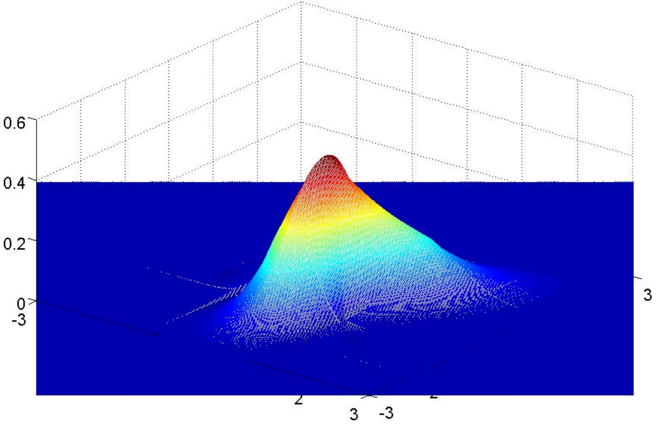
- Performance measures

	$\mathbb{E}(Q)$		$\mathbb{P}\{\text{Ab.}\}$	
λ^n	Numerical	Diffusion	Simulation	Diffusion
200	8.72	8.85	0.0290	0.0295
220	31.05	30.64	0.0940	0.0928

Steady-state density for $\hat{X}^n$ and $\sqrt{n}\tilde{X}$: $\lambda^n = 200$



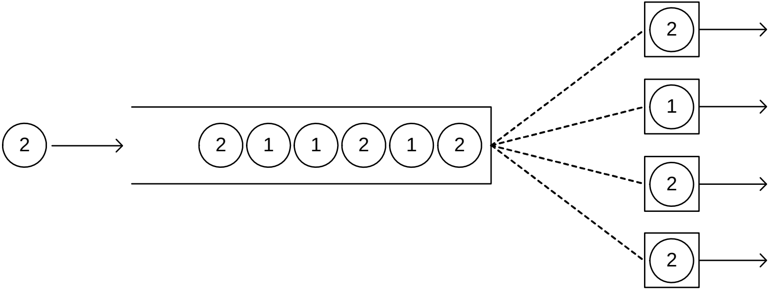
Steady-state density for $(\tilde{Y}_1, \tilde{Y}_2)$: $\lambda^n = 200$



Proof sketches: critically loaded $G/Ph/n + GI$ queues

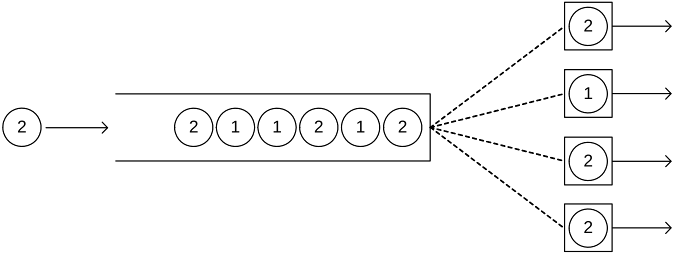
- The lemma reduces $+GI$ to $+M$
- Perturbed systems
- System representations
- Centering, scaling, applying standard tools: Donsker's theorem, continuous-mapping theorem, random-time-change theorem
- Conventional heavy traffic limits for generalized Jackson networks: Reiman (84), Johnson (83)
- Stone's theorem: Halfin-Whitt (81), Garnett-M-Reiman (02), Whitt (04), Armony-Maglaras (04)

Step 1: Perturbed systems



- Each phase has at most one customer in service, with additive service rate
- Only the leading customer in queue can abandon with additive abandonment rate

The two systems are equal in distribution



- state $(U(t), \mathcal{Q}(t), Z_1(t), Z_2(t))$, where, for example,

$$U(t) = 3.5, \quad \mathcal{Q}(t) = \{2, 1, 2, 1, 1, 2\}, \quad Z_1(t) = 1, \quad Z_2(t) = 3.$$

- Two Markov processes have the same generators.

Donsker's theorem for primitives

Primitive processes: in addition to E^n,

- service: S_k Poisson process with rate ν_k; $\hat{S}(t) = S(t) - \nu t$,
- abandonment: G Poisson process with rate α; $\hat{G}(t) = G(t) - \alpha t$,
- routing: for each $N \geq 1$ and $k = 0, 1, \ldots, K$,

$$\Phi^k(N) = \sum_{j=1}^{N} \phi^k(j); \qquad \hat{\Phi}^k(N) = \sum_{j=1}^{N} \left(\phi^k(j) - p^k \right),$$

where $p^0 = p$ and p^k is the kth column of P'.

Define diffusion-scaled processes

$$\tilde{S}^n(t) = \frac{1}{\sqrt{n}} \hat{S}(nt), \quad G^n(t) = \frac{1}{\sqrt{n}} \hat{G}(nt), \quad \tilde{\Phi}^{n,k}(t) = \frac{1}{\sqrt{n}} \hat{\Phi}^k(\lfloor nt \rfloor).$$

$$(\tilde{E}^n, \tilde{G}^n, \tilde{S}^n, \tilde{\Phi}^{0,n}, \ldots, \tilde{\Phi}^{K,n}) \Rightarrow (\tilde{E}, \tilde{G}, \tilde{S}, \tilde{\Phi}^0, \ldots, \tilde{\Phi}^K) \quad \text{as } n \to \infty.$$

System representations

$$X^n(t) = X^n(0) + E^n(t) - D^n(t) - G\left(\int_0^t Q^n(s)\,ds\right),$$

$$Z^n(t) = Z^n(0) + \Phi^0(B^n(t)) + \sum_{k=1}^{K} \Phi^k(S_k(T_k^n(t))) - S(T^n(t)),$$

$$T_k^n(t) = \int_0^t Z_k^n(s)\,ds, \quad S(T^n(t)) = (S_1(T_1^n(t)), \ldots, S_K(T_K^n(t)))'.$$

where

$$D^n(t) = -e'M^n(t) + e'R\int_0^t Z^n(s)\,ds,$$

$$e'Z^n(t) = e'Z^n(0) + B^n(t) - D^n(t),$$

$$M^n(t) = \sum_{k=1}^{K} \hat{\Phi}^k\left(S_k(T_k^n(t))\right) - (I - P')\hat{S}\left(T^n(t)\right).$$

Continuous-mapping theorem

After some centering,

$$\hat{X}^n(t) = U^n(t) - \alpha \int_0^t (\hat{X}^n(s))^+ \, ds - e' R \int_0^t \hat{Z}^n(s) \, ds,$$

$$\hat{Z}^n(t) = V^n(t) - p(\hat{X}^n(t))^- - (I - pe')R \int_0^t \hat{Z}^n(s) \, ds,$$

Thus, $(\hat{X}^n, \hat{Z}^n) = \Theta(U^n, V^n)$, where

$$U^n(t) = \hat{X}^n(0) + \hat{E}^n(t) + e' M^n(t) - \hat{G}\left(\int_0^t (\hat{X}^n(s))^+ \, ds\right),$$

$$V^n(t) = (I - pe')\hat{Z}^n(0) + \hat{\Phi}^0(B^n(t)) + (I - pe')M^n(t).$$

Because, $(\tilde{X}^n, \tilde{Z}^n) = \Theta(\tilde{U}^n, \tilde{V}^n)$, the theorem follows from

$$(\tilde{U}^n, \tilde{V}^n) \Rightarrow (\tilde{U}, \tilde{V}), \qquad \tilde{U}^n(t) = \frac{1}{\sqrt{n}} U^n(t).$$

$$\tilde{U}^n(t) = \tilde{X}^n(0) + \tilde{E}^n(t) + e'\tilde{M}^n(t) - \tilde{G}^n\left(\int_0^t (\bar{X}^n(s))^+\,ds\right),$$

$$\tilde{M}^n(t) = \frac{1}{\sqrt{n}}M^n(t) = \sum_{k=1}^{K} \tilde{\Phi}^{k,n}(\bar{S}_k^n(\bar{T}_k^n(t))) - (I - P')\tilde{S}^n(\bar{T}^n(t))$$

where, for $t \geq 0$,

$$\bar{B}^n(t) = \frac{1}{n}B^n(nt), \quad \bar{S}^n(t) = \frac{1}{n}S(nt), \quad \bar{T}^n(t) = \frac{1}{n}T^n(nt),$$
$$\bar{X}^n(t) = \frac{1}{n}\hat{X}^n(t), \quad \bar{Z}^n(t) = \frac{1}{n}\hat{Z}^n(t).$$

Because $(\bar{X}^n, \bar{Z}^n) = \Theta(\bar{U}^n, \bar{V}^n) \Rightarrow 0$, one has fluid limits

$$(\bar{S}^n, \bar{T}^n, \bar{B}^n) \Rightarrow (\bar{S}, \bar{T}, \bar{B}), \quad \text{where}$$
$$\bar{S}_k(t) = \nu_k t, \quad \bar{T}_k(t) = \gamma_k t, \quad \bar{B}(t) = \mu t.$$

More on continuous-mapping approach

- Reed (07), Kaspi-Ramanan (07), Kang-Ramanan (08) and Zhang (09) did not use continuous-mapping approach, all involving a complicated tightness argument.
- Decreusefond-Moyal (08) and Talreja-Reed (09) used continuous-mapping approach for $G/GI/\infty$ queues.
- Kaspi-Ramanan (09) measure-valued diffusion limits for $G/GI/n$ queues.