Fluid and Diffusion Limits for Many-Server Queues

Jim Dai

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

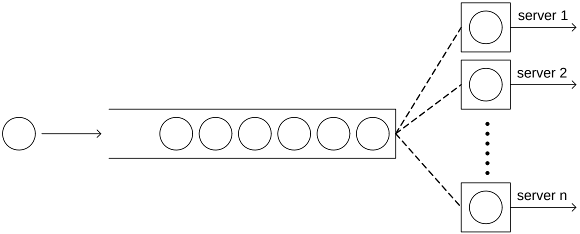
July 14, 2009

Joint work with Shuangchi He and Tolga Tezcan (UIUC)

Who is this?

Who is this?

Multi-server queues



- $G/GI/n + GI$ queues, FIFO queue, a classical model
- iid service times and iid patience times
- The number of servers n is large: call centers, web server farms, hospital beds

Performance processes

At time t,

- System size $X(t)$ — the total number of customers in system

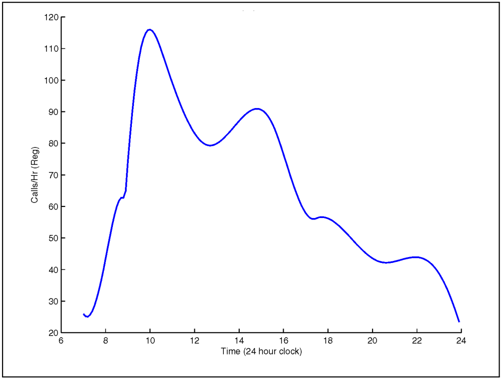
$$\hat{X}(t) = X(t) - n$$

- Queue size $Q(t) = (\hat{X}(t))^+$
- The number of idle servers $I(t) = (\hat{X}(t))^-$
- Offered waiting time at time t: $W(t)$

Examples of performance measures:

- abandonment probability $\mathbb{P}\{\text{Ab.}\}$
- average queue size $\mathbb{E}(Q)$
- average waiting time among those who are served $\mathbb{E}(W|S)$

Time-varying arrival rates: Brown et al (05)



Operating regimes

- overloaded; efficiency-driven (ED)
- critically loaded; quality- and efficiency-driven (QED); Halfin-Whitt regime
- underloaded; quality-driven (QD)

- ED: fluid model; QED: diffusion model
- Focus on QED regime

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

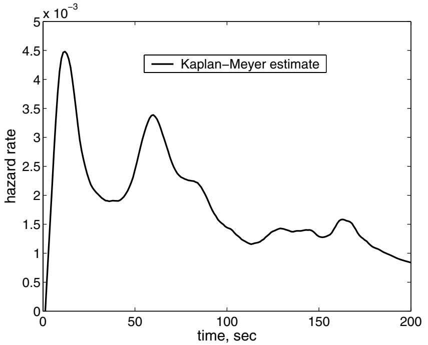
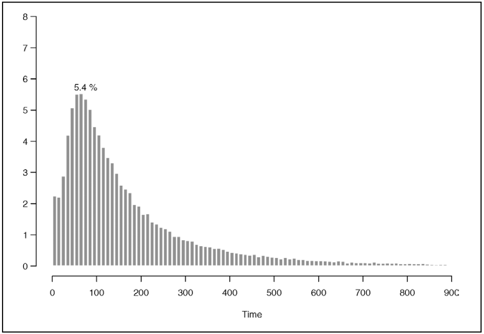
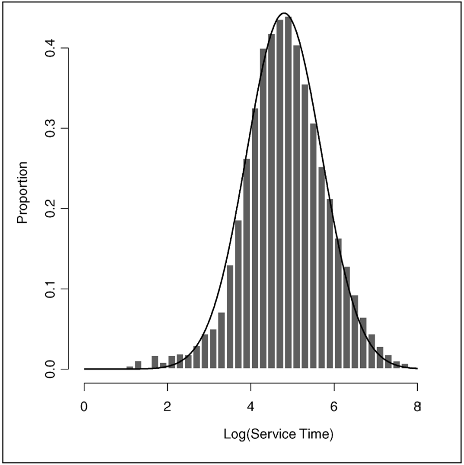


Fig. 2. Hazard rate for patience of regular customers

M-Zeltyn (04)

Non-exponential service time distribution





Brown et al (2005)

Fluid and diffusion limits for performance processes

These limits help

- understand the sensitivity of service and patience distributions on system performance
- make staffing decisions to meet certain performance targets
- predict system performance
- design near-optimal routing policies for systems with multiple server pools that serve multiple customer classes

An outline

- Asymptotic framework and phase-type distributions
- Critically loaded $G/Ph/n + GI$ queues
- Overloaded $G/Ph/n + M$ queues
- Proof sketches
- Comments on $G/GI/n + GI$ queues

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 G is independent of n; it has finite mean $1/\mu$.

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

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

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

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

An example of phase-type distributions

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

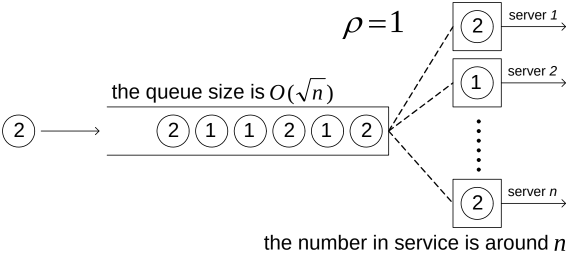
$$\xi = \begin{cases} \exp(\nu_1) & \text{with probability } p_1 \\ \exp(\nu_2) & \text{with probability } p_2 \end{cases},$$

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

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

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

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



- Let $Z_k^n(t)$ be the number of phase k customers in service at time t.
- 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 $R = (I - P')\mathrm{diag}(\nu)$. 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-M-Reiman (98)

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

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

- 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: $M/H_2/200 + M$

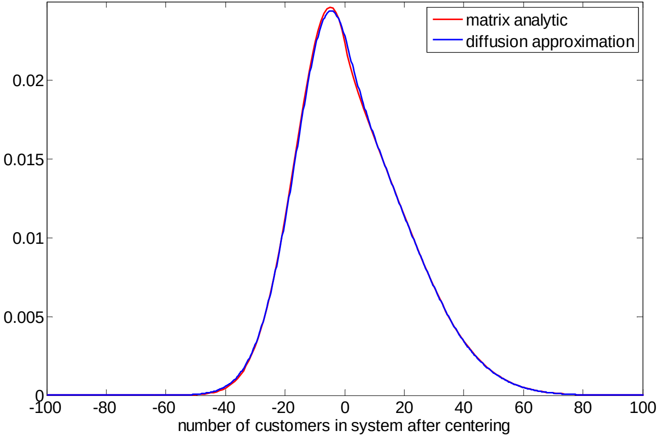
- $H_2(1/2.2, 1/.2, .4)$ service time distribution and $\alpha = F'(0) = 2/3$.
- Finite element method to solve the stationary distribution of $\tilde{Y}$; Dai-Harrison (92), Shen-Chen-Dai-Dai (02); reference density

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

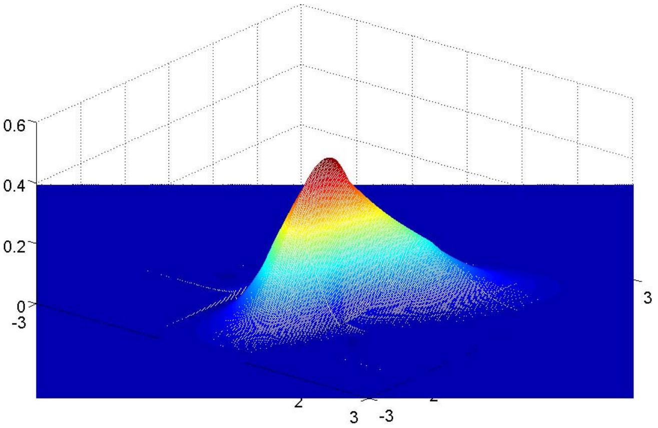
  truncate the area $(-8, 14) \times (-8, 14)$; the grid consists of 1×1 squares.
- Performance measures

	$\mathbb{E}(Q)$		$\mathbb{P}\{\text{Ab.}\}$	
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 $(\tilde{Y}_1, \tilde{Y}_2)$: $\lambda^n = 200$



Sensitivity of abandonment distribution: $\rho = 1$

Only $\alpha = F'(0)$ is used in the diffusion limits for $G/Ph/n/+GI$ queues.

Lemma (Dai-He 09, $G/GI/n+GI$ queues)

Assume that diffusion-scaled queue size is stochastically bounded: for each $T > 0$,

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

Then for any $T > 0$,

$$\frac{1}{\sqrt{n}} \sup_{0\le t\le T} \left| C^n(t) - \alpha \int_0^t Q^n(s) ds \right| \Rightarrow 0 \quad \text{as } n \to \infty,$$

where $C^n(t)$ is cumulative number of abandonments in $(0, t]$.

Sensitivity of distributions on $M(210)/GI/200 + GI$ queues

	Patience distribution			
	uniform(0, 4), $F'(0) = .25$		$H_2(1,3,0.5)$, $F'(0) = 2/3$	
Service dist.	$\mathbb{P}\{A\}$	$\mathbb{E}(Q)$	$\mathbb{P}\{A\}$	$\mathbb{E}(Q)$
H_2	.0530	42.69	.0584	18.66
	$\pm$.000	$\pm$0.46	$\pm$.001	$\pm$.146
$+M(\alpha)$	.0528	44.43	.0584	18.43
$+M(.5)$	.0569	23.87	.0569	23.87
LN	.0523	42.13	.0571	18.24
$+M(\alpha)$	.0519	43.69	.0570	17.95
$+M(.5)$	.0555	23.31	.0555	23.31

Sensitivity of distributions

- In QED regime, performance is very sensitive to patience time distribution via $F'(0)$
- Appears not sensitive to service time distribution with $\mu = 1$; Gamarnik-Momcilovic (08) for lattice service time distribution
- Mean patience time can be misleading
- The lemma suggests a linear relationship for $G/GI/n + GI$ queues in QED:

$$\lambda^n \times \mathbb{P}\{\text{Ab.}\} \approx F'(0)\mathbb{E}(Q). \tag{4}$$

  M-Zeltyn (04): empirical observations; Zeltyn-M (05) proved it for $M/M/n + GI$ queues using Baccelli-Hebuterne (81)

Overloaded $G/Ph/n+M$ queues: $\rho > 1$

$$\tilde{X}^n(t) = \frac{1}{\sqrt{n}}\Big(X^n(t) - n(1+q)\Big), \qquad q = (\lambda - \mu)/\alpha$$

Theorem (Dai-He-Tezcan 09)

Suppose that $(\tilde{X}^n(0), \tilde{Z}^n(0)) \Rightarrow (\xi, \eta)$. *Then*

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

where $(\tilde{X}, \tilde{Z}) = \Psi(\tilde{U}, \tilde{V})$ *is a* $(K+1)$*-dimensional degenerate OU process; the map* $\Psi : (u, v) \in \mathbb{D} \times \mathbb{D}^K \to (x, z) \in \mathbb{D} \times \mathbb{D}^K$ *is well defined via*

$$x(t) = u(t) - \alpha \int_0^t x(s)\, ds - e'R \int_0^t z(s)\, ds,$$

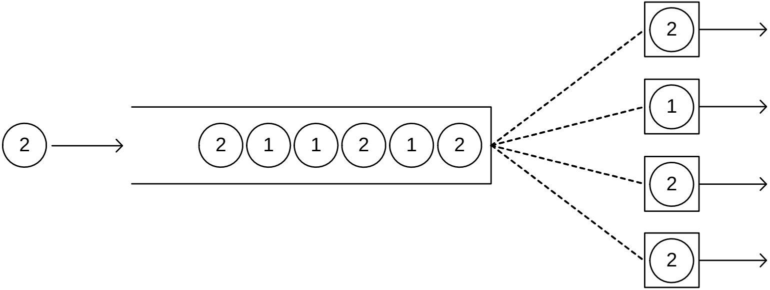
$$z(t) = v(t) - (I - pe')R \int_0^t z(s)\, ds.$$

Whitt (04) for overloaded $M/M/n+M$ queues

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

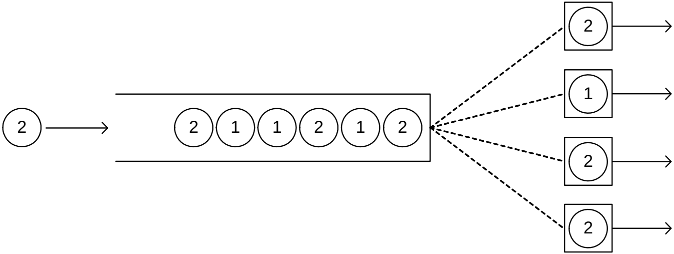
- The lemma reduces $+GI$ to $+M$
- Perturbed systems
- System representations
- Centering, scaling, applying standard tools: Donsker's theorem, continuous-mapping theorem, random-time-change theorem
- Conventional heavy traffic limits for generalized Jackson networks: Reiman (84), Johnson (83)
- Stone's 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 $k = 0, 1, \dots, K$,

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

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

Define diffusion-scaled processes

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

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

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

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

Random-time-change and fluid limits

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

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

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

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

Lemma

Let $I^n(t)$ be the number of idle servers at time t. Assume $\rho > 1$. Then for each $t > 0$,

$$\frac{1}{\sqrt{n}} \sup_{0\le s\le t} I^n(s)\, ds \Rightarrow 0 \qquad \text{as } n \to \infty.$$

Setting $\bar{U}(t) = q + (\lambda - \mu)t$, one has

$$(\tilde{X}^n, \tilde{Z}^n) = \Theta(\tilde{U}^n + \sqrt{n}\bar{U}, \tilde{V}^n),$$

For any $t > 0$, there exists $C > 0$ such that

$$\|\Theta\big(\tilde{U}^n + \sqrt{n}\bar{U}, \tilde{V}^n)\big) - \Theta(\sqrt{n}\bar{U}, 0)\|_t \le C\left(\|\tilde{U}^n\|_t + \|\tilde{V}^n\|_t\right).$$

$$\inf_{0\le s\le t} \frac{1}{\sqrt{n}}\hat{X}^n(s) > \sqrt{n}q - C\left(\frac{1}{\sqrt{n}}\|\tilde{U}^n\| + \|\tilde{V}^n\|_t\right),$$

$$\inf_{0\le s\le t} \frac{1}{\sqrt{n}}\hat{X}^n(s) \to \infty, \quad \text{which implies} \quad \sup_{0\le s\le t} \frac{1}{\sqrt{n}} I^n(s) \to 0.$$

$G/GI/n + GI$ **queues**

Whitt (06) proposed a fluid model and the following approximation when $\rho > 1$: the offered waiting time w satisfies

$$F(w) = \frac{\lambda - \mu}{\lambda}, \quad \mathbb{E}(Q^n) \approx \lambda^n \, \mathbb{E}(\eta \wedge w). \tag{5}$$

Examples: $M(120)/GI/100 + GI$

	E_2 service distribution		LN service distribution	
Patience	$\mathbb{P}\{A\}$	$\mathbb{E}(Q)$	$\mathbb{P}\{A\}$	$\mathbb{E}(Q)$
$H_2(2, 2/3, .5)$	.168	15.58	.1689	15.70
Fluid	.167	15.35	.1667	15.35
Exp(4/3)	.168	15.08	.1695	15.26
Uniform(0,2)	.167	36.34	.1665	35.97
Fluid	.168	36.67	.1667	36.67
Exp(.5)	.166	39.91	.1673	40.15

Limits for $G/GI/n + GI$ queues

- Reed (07) proved the convergence of $\tilde{X}^n$ for critically loaded $G/GI/n$ queues; uses Krichagina-Puhalskii (97) for $G/GI/\infty$ queues.
- M-Momcilovic (09) generalizes Reed (07) to $G/GI/n + GI$ queues.
- Haspi-Ramanan (07) measure-valued fluid limits for $G/GI/n$ queues; Kang-Ramanan (08) for $G/GI/n + GI$ queues.
- Zhang (09): measure-valued fluid limits for $G/GI/n + GI$ queues; "residual" processes.
- Bassamboo-Randhawa (09) justified (5) for $M/M/n + GI$ queues
- Kang-Ramanan (09) justified (5) for $G/GI/n + GI$ queues

More on continuous-mapping approach

- Reed (07), Kaspi-Ramanan (07), Kang-Ramanan (08) and Zhang (09) all involve a complicated tightness argument.
- There is a need to extend the continuous-mapping approach to the measure-valued setting; the key is to find a map on some infinitely dimensional space; diffusion limit is a piecewise-OU process.
- Decreusefond-Moyal (2008), Talreja-Reed (2009) for $G/GI/\infty$ queues.

More on continuous-mapping approach

- Reed (07), Kaspi-Ramanan (07), Kang-Ramanan (08) and Zhang (09) all involve a complicated tightness argument.
- There is a need to extend the continuous-mapping approach to the measure-valued setting; the key is to find a map on some infinitely dimensional space; diffusion limit is a piecewise-OU process.
- Decreusefond-Moyal (2008), Talreja-Reed (2009) for $G/GI/\infty$ queues.
- Kaspi-Ramanan (09) distribution-valued diffusion limits for $G/GI/n$ queues.

Fluid and diffusion limits for staffing and skill-based routing

- Fluid limits:
  Bassamboo-Harrison-Zeevi (06a,b, 08), Bassamboo-Zeevi (08), Perry-Whitt (09a,b), Bassamboo-Randhawa (09), ...
- Diffusion limits:
  Armony-Maglaras (04a, b), Atar-M-Reiman (04), Borst-M-Reiman (04), Armony (05), Atar (05), Gurvich-Armony-M (05), Tezcan (07), Gurvich-Whitt (06, 07), Dai-Tezcan (08), Tezcan-Dai (09), Armony-Ward (09), Koscaga-Ward (09), Stolyar-Tezcan (09), ...

Surveys and references

- Gans-Koole-M (03), Telephone call centers: Tutorial, review, and research prospects, *M&SOM*, **5**, 79-141.
- Mandelbaum (06), Call centers: research bibliography with abstracts; `http://iew3.technion.ac.il/serveng/References/US7_CC_avi.pdf`
- Dai, He and Tezcan, Many-server diffusion limits for $G/Ph/n + GI$ queues, December 2008.
- Dai and He, Customer abandonment in many-server queues, April 2009.