

FLUID MODELS OF MANY-SERVER QUEUES WITH ABANDONMENT

Jiheng Zhang



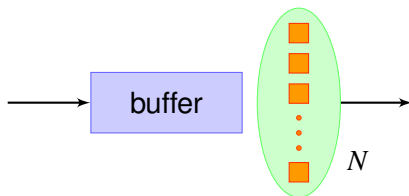
香港科技大學

THE HONG KONG UNIVERSITY OF
SCIENCE AND TECHNOLOGY

June 10, 2010

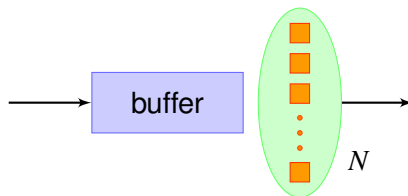
Model and Motivation

Many-server queue



Model and Motivation

Many-server queue

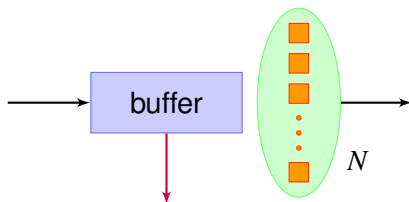


Motivation:

- Customer call centers and other services areas.

Model and Motivation

Many-server queue **with abandonment**



Motivation:

- Customer call centers and other services areas.

Many-server queue v.s. Single-server queue



Many-server queue v.s. Single-server queue



Large scale: *high demand*, need for *high capacity*

- Single server queue: increase speed
- Many-server queue: increase number of servers

A Real World Challenge

The service time is not exponentially distributed!

- Brown et. al. *Statistical analysis of a telephone call center: a queueing-science perspective*. JASA 2005

In this research

- Arrival process: **general**
- Service/patient time distribution: **general**

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

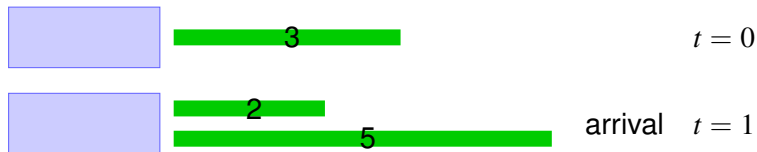
- Whitt 2004 ($M/M/N + M$)
- Zeltyn and Mandelbaum 2005 ($G/M/N + G$)
- Whitt 2006 ($G/G/N + G$)
- Puhalskii 2008 ($M_t/M_t/N_t + M_t$)
- Kang and Ramanan 2008 ($G/G/N + G$)
- Mandelbaum and Momčilović 2009 ($G/G/N + G$)
- Dai, He and Tezcan 2009 ($G/Ph/N + G$)
-

System Dynamics – example with $N = 2$

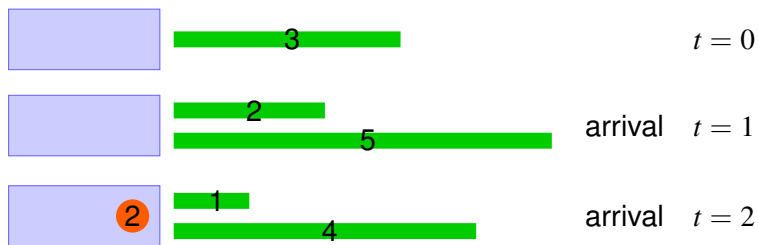


$t = 0$

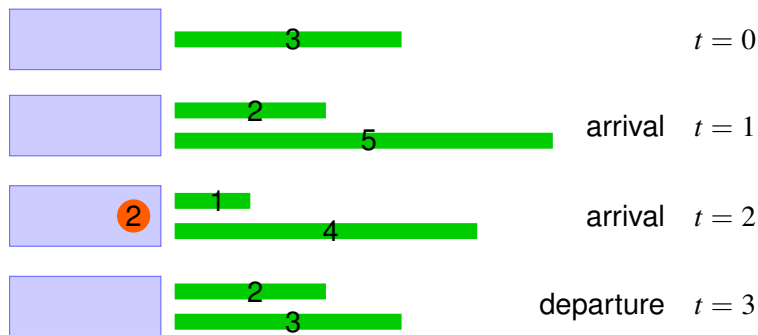
System Dynamics – example with $N = 2$



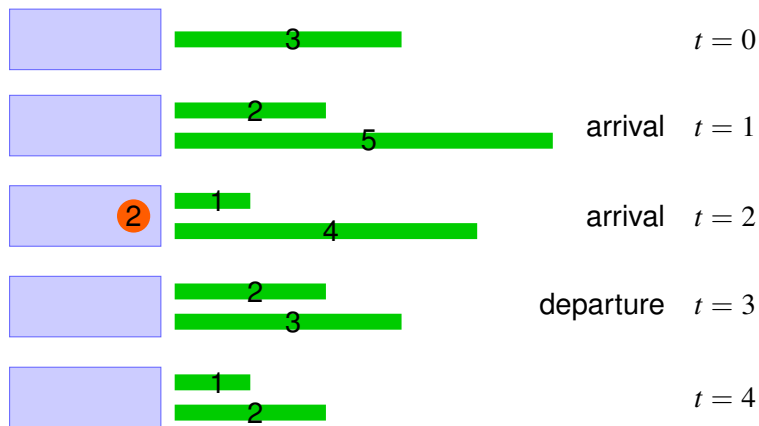
System Dynamics – example with $N = 2$



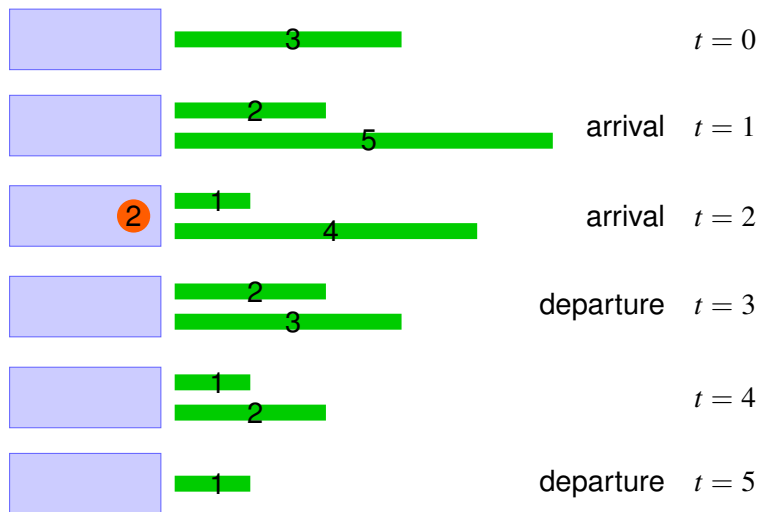
System Dynamics – example with $N = 2$



System Dynamics – example with $N = 2$



System Dynamics – example with $N = 2$



Measure-valued State Descriptor

Server pool

- $\mathcal{Z}(t)(C)$: # of customers in server with *remaining service time* in $C \subset (0, \infty)$

Measure-valued State Descriptor

Server pool

- $\mathcal{Z}(t)(C)$: # of customers in server with *remaining service time* in $C \subset (0, \infty)$

Evolution

$$\mathcal{Z}(t_0 + t)(C) = \mathcal{Z}(t_0)(C + t) + \dots (t_0, t_0 + t) \dots$$

Measure-valued State Descriptor

Server pool

- $\mathcal{Z}(t)(C)$: # of customers in server with *remaining service time* in $C \subset (0, \infty)$

Evolution

$$\mathcal{Z}(t_0 + t)(C) = \mathcal{Z}(t_0)(C + t) + \dots (t_0, t_0 + t) \dots$$

Richness

$$Z(t) = \mathcal{Z}(t)((0, \infty))$$

Measure-valued State Descriptor

Server pool

- $\mathcal{Z}(t)(C)$: # of customers in server with *remaining service time* in $C \subset (0, \infty)$

Evolution

$$\mathcal{Z}(t_0 + t)(C) = \mathcal{Z}(t_0)(C + t) + \dots (t_0, t_0 + t) \dots$$

Richness

$$Z(t) = \mathcal{Z}(t)((0, \infty))$$

Literature

Gromoll, Puha & Williams '02, Puha & Williams '02, Gromoll '06

Gromoll & Kurk '07, Gromoll, Robert & Zwart '08, ...

Zhang, Dai & Zwart '07, '08, Zhang & Zwart '08

Measure-valued State Descriptor

Virtual buffer

- $\mathcal{R}(t)(C)$: # of customers in virtual buffer with *remaining patient time* in $C \subset (-\infty, \infty)$

Measure-valued State Descriptor

Virtual buffer

- $\mathcal{R}(t)(C)$: # of customers in virtual buffer with *remaining patient time* in $C \subset (-\infty, \infty)$

Measure-valued State Descriptor

Virtual buffer

- $\mathcal{R}(t)(C)$: # of customers in virtual buffer with *remaining patient time* in $C \subset (-\infty, \infty)$

Evolution

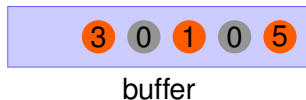


Measure-valued State Descriptor

Virtual buffer

- $\mathcal{R}(t)(C)$: # of customers in virtual buffer with *remaining patient time* in $C \subset (-\infty, \infty)$

Evolution

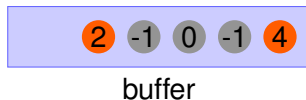


Measure-valued State Descriptor

Virtual buffer

- $\mathcal{R}(t)(C)$: # of customers in virtual buffer with *remaining patient time* in $C \subset (-\infty, \infty)$

Evolution

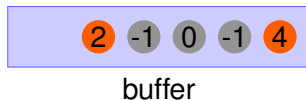


Measure-valued State Descriptor

Virtual buffer

- $\mathcal{R}(t)(C)$: # of customers in virtual buffer with *remaining patient time* in $C \subset (-\infty, \infty)$

Evolution



Richness

$$Q(t) = \mathcal{R}(t)((0, \infty))$$

$$R(t) = \mathcal{R}(t)((-\infty, \infty))$$

System Dynamics

Internal transfer process

$$B(t) = E(t) - R(t)$$

System Dynamics

Internal transfer process

$$B(t) = E(t) - R(t)$$

$1 + B(t)$: index of the next customer to be served

System Dynamics

Internal transfer process

$$B(t) = E(t) - R(t)$$

$1 + B(t)$: index of the next customer to be served

Stochastic dynamic equations

$$\mathcal{R}(t)(C) = \sum_{i=1+B(t)}^{E(t)} \delta_{u_i}(C + t - a_i), \quad C \in \mathcal{B}(\mathbb{R})$$

$$\mathcal{Z}(t)(C) = \mathcal{Z}(0)(C + t)$$

$$+ \sum_{i=1+B(0)}^{B(t)} 1_{\{u_i > \tau_i - a_i\}} \delta_{v_i}(C + t - \tau_i), \quad C \in \mathcal{B}(\mathbb{R}_+)$$

System Dynamics

Internal transfer process

$$B(t) = E(t) - R(t)$$

$1 + B(t)$: index of the next customer to be served

Stochastic dynamic equations

$$\mathcal{R}(t)(C) = \sum_{i=1+B(t)}^{E(t)} \delta_{u_i}(C + t - a_i), \quad C \in \mathcal{B}(\mathbb{R})$$

$$\mathcal{Z}(t)(C) = \mathcal{Z}(0)(C + t)$$

$$+ \sum_{i=1+B(0)}^{B(t)} 1_{\{u_i > \tau_i - a_i\}} \delta_{v_i}(C + t - \tau_i), \quad C \in \mathcal{B}(\mathbb{R}_+)$$

$\delta_{u_i - (t - a_i)} C$

System Dynamics

Internal transfer process

$$B(t) = E(t) - R(t)$$

$1 + B(t)$: index of the next customer to be served

Stochastic dynamic equations

$$\mathcal{R}(t)(C) = \sum_{i=1+B(t)}^{E(t)} \delta_{u_i}(C + t - a_i), \quad C \in \mathcal{B}(\mathbb{R})$$

$$\mathcal{Z}(t)(C) = \mathcal{Z}(0)(C + t)$$

$$+ \sum_{i=1+B(0)}^{B(t)} \mathbf{1}_{\{u_i > \tau_i - a_i\}} \delta_{v_i}(C + t - \tau_i), \quad C \in \mathcal{B}(\mathbb{R}_+)$$

$\delta_{u_i - (t - a_i)} C$

System Dynamics

Internal transfer process

$$B(t) = E(t) - R(t)$$

$1 + B(t)$: index of the next customer to be served

Stochastic dynamic equations

$$\mathcal{R}(t)(C) = \sum_{i=1+B(t)}^{E(t)} \delta_{u_i}(C + t - a_i), \quad C \in \mathcal{B}(\mathbb{R})$$

$$\mathcal{Z}(t)(C) = \mathcal{Z}(0)(C + t)$$

$$+ \sum_{i=1+B(0)}^{B(t)} \mathbf{1}_{\{u_i > \tau_i - a_i\}} \delta_{v_i}(C + t - \tau_i), \quad C \in \mathcal{B}(\mathbb{R}_+)$$

$$\delta_{u_i - (t - a_i)} C$$

$$\delta_{v_i - (t - \tau_i)} C$$

System Dynamics

Total number of customers

$$X(t) = Q(t) + Z(t)$$

Policy constraints

$$Q(t) = (X(t) - N)^+, \quad Z(t) = (X(t) \wedge N)$$

Fluid Model

- $E(\cdot): \lambda \cdot$
- $\{u_i\}: F (\vartheta_F \sim F)$
- $\{v_i\}: G (\vartheta_G \sim G)$

Fluid Model

- $E(\cdot): \lambda \cdot$
- $\{u_i\}: F (\vartheta_F \sim F)$
- $\{v_i\}: G (\vartheta_G \sim G)$

$$\bar{B}(s) = \lambda s - \bar{R}(s)$$

Fluid Model

- $E(\cdot): \lambda \cdot$
- $\{u_i\}: F (\vartheta_F \sim F)$
- $\{v_i\}: G (\vartheta_G \sim G)$

$$\bar{B}(s) = \lambda s - \bar{R}(s)$$

Fluid dynamic equations

$$\bar{R}(t)(C) = \int_{t - \frac{\bar{R}(t)}{\lambda}}^t \vartheta_F(C + t - s) d\lambda s, \quad C \in \mathcal{B}(\mathbb{R})$$

$$\begin{aligned} \bar{Z}(t)(C) &= \bar{Z}(0)(C + t) \\ &+ \int_0^t \vartheta_F\left(\frac{\bar{R}(s)}{\lambda}, \infty\right) \vartheta_G(C + t - s) d\bar{B}(s), \quad C \in \mathcal{B}(\mathbb{R}_+) \end{aligned}$$

Fluid Model

- $E(\cdot): \lambda \cdot$
- $\{u_i\}: F (\vartheta_F \sim F)$
- $\{v_i\}: G (\vartheta_G \sim G)$

$$\bar{B}(s) = \lambda s - \bar{R}(s)$$

has to be increasing!

Fluid dynamic equations

$$\bar{R}(t)(C) = \int_{t - \frac{\bar{R}(t)}{\lambda}}^t \vartheta_F(C + t - s) d\lambda s, \quad C \in \mathcal{B}(\mathbb{R})$$

$$\begin{aligned} \bar{Z}(t)(C) &= \bar{Z}(0)(C + t) \\ &+ \int_0^t \vartheta_F\left(\frac{\bar{R}(s)}{\lambda}, \infty\right) \vartheta_G(C + t - s) d\bar{B}(s), \quad C \in \mathcal{B}(\mathbb{R}_+) \end{aligned}$$

Fluid Model

- $E(\cdot): \lambda \cdot$
- $\{u_i\}: F (\vartheta_F \sim F)$
- $\{v_i\}: G (\vartheta_G \sim G)$

$$\bar{B}(s) = \lambda s - \bar{R}(s)$$

has to be increasing!

Fluid dynamic equations

$$\bar{R}(t)(C) = \int_{t - \frac{\bar{R}(t)}{\lambda}}^t \vartheta_F(C + t - s) d\lambda s, \quad C \in \mathcal{B}(\mathbb{R})$$

$$\begin{aligned} \bar{Z}(t)(C) &= \bar{Z}(0)(C + t) \\ &+ \int_0^t \vartheta_F\left(\frac{\bar{R}(s)}{\lambda}, \infty\right) \vartheta_G(C + t - s) d\bar{B}(s), \quad C \in \mathcal{B}(\mathbb{R}_+) \end{aligned}$$

Constraints

$$\bar{Q}(t) = (\bar{X}(t) - N)^+, \quad \bar{Z}(t) = (\bar{X}(t) \wedge N)$$

Existence and Uniqueness of Fluid Model Solution

Fluid model solution with initial condition $(\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0)$

- $(\bar{\mathcal{R}}(0), \bar{\mathcal{Z}}(0)) = (\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0)$
- $(\bar{\mathcal{R}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ satisfies fluid dynamic equations and constraints

Existence and Uniqueness of Fluid Model Solution

Fluid model solution with initial condition $(\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0)$

- $(\bar{\mathcal{R}}(0), \bar{\mathcal{Z}}(0)) = (\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0)$
- $(\bar{\mathcal{R}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ satisfies fluid dynamic equations and constraints

THEOREM

Assume that

G is continuous, with $0 < \mu < \infty$,

$F(\cdot)$ is Liptachitz continuous, or $\sup_{x \in [0, \infty)} h_F(x) < \infty$.

*There **exists a unique** solution to the fluid model (λ, F, G, N) for any valid initial condition $(\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0)$.*

Invariant State of Fluid Model

Invariant state $(\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty)$

- $(\bar{\mathcal{R}}(0), \bar{\mathcal{Z}}(0)) = (\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty)$ implies $(\bar{\mathcal{R}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \equiv (\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty)$

Invariant State of Fluid Model

Invariant state $(\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty)$

- $(\bar{\mathcal{R}}(0), \bar{\mathcal{Z}}(0)) = (\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty)$ implies $(\bar{\mathcal{R}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \equiv (\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty)$

THEOREM

The state $(\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty)$ is an invariant state *if and only if* it satisfies

$$\bar{\mathcal{R}}_\infty(C_x) = \lambda \int_0^w F^c(x+s) ds, \quad x \in \mathbb{R},$$

$$\bar{\mathcal{Z}}_\infty(C_x) = \min(\rho, 1) N[1 - G_e(x)], \quad x \in \mathbb{R}_+,$$

where w is a solution to the equation

$$F(w) = \max\left(\frac{\rho - 1}{\rho}, 0\right).$$

Fluid Model Analysis

Replace C by $C_x = (x, \infty)$, (note that $\vartheta_F(C_x) = F^c(x)$)

$$\bar{\mathcal{R}}(t)(C_x) = \lambda \int_{t - \frac{\bar{\mathcal{R}}(t)}{\lambda}}^t F^c(x + t - s) ds, \quad x \in \mathbb{R},$$

$$\bar{\mathcal{Z}}(t)(C_x) = \bar{\mathcal{Z}}(0)(C_x + t) + \int_0^t F^c\left(\frac{\bar{\mathcal{R}}(s)}{\lambda}\right) G^c(x + t - s) d\bar{B}(s), \quad x \in \mathbb{R}_+,$$

Fluid Model Analysis

Replace C by $C_x = (x, \infty)$, (note that $\vartheta_F(C_x) = F^c(x)$)

$$\bar{R}(t)(C_x) = \lambda \int_{t - \frac{\bar{R}(t)}{\lambda}}^t F^c(x + t - s) ds, \quad x \in \mathbb{R},$$

$$\bar{Z}(t)(C_x) = \bar{Z}(0)(C_x + t) + \int_0^t F^c\left(\frac{\bar{R}(s)}{\lambda}\right) G^c(x + t - s) d\bar{B}(s), \quad x \in \mathbb{R}_+,$$

The functional equation

$$\bar{X}(t) = \zeta_0(t) + \rho \int_0^t H((\bar{X}(t-s) - 1)^+) dG_e(s) + \int_0^t (\bar{X}(t-s) - 1)^+ dG(s)$$

where $H(x) = F^c(F_e^{-1}(\frac{\alpha}{\lambda}x))$.

The Special Case with Exponential Distribution

Now, we specialize in the case with exponential distribution, i.e.

$$F(t) = F_e(t) = 1 - e^{-\alpha t}, \quad G(t) = G_e(t) = 1 - e^{-\mu t}.$$

The Special Case with Exponential Distribution

Now, we specialize in the case with exponential distribution, i.e.

$$F(t) = F_e(t) = 1 - e^{-\alpha t}, \quad G(t) = G_e(t) = 1 - e^{-\mu t}.$$

Now the key equation becomes

$$\begin{aligned} \bar{X}(t) = & \zeta_0(t) + \rho \int_0^t \left[1 - \frac{\alpha}{\lambda} ((\bar{X}(t-s) - 1)^+) \right] \mu e^{-\mu s} ds \\ & + \int_0^t (\bar{X}(t-s) - 1)^+ \mu e^{-\mu s} ds, \end{aligned}$$

with $\zeta_0(t) = \bar{X}_0 e^{-\mu t}$.

The Special Case with Exponential Distribution

Now, we specialize in the case with exponential distribution, i.e.

$$F(t) = F_e(t) = 1 - e^{-\alpha t}, \quad G(t) = G_e(t) = 1 - e^{-\mu t}.$$

Now the key equation becomes

$$\begin{aligned} \bar{X}(t) = & \zeta_0(t) + \rho \int_0^t \left[1 - \frac{\alpha}{\lambda} ((\bar{X}(t-s) - 1)^+) \right] \mu e^{-\mu s} ds \\ & + \int_0^t (\bar{X}(t-s) - 1)^+ \mu e^{-\mu s} ds, \end{aligned}$$

with $\zeta_0(t) = \bar{X}_0 e^{-\mu t}$. After some algebra, we get

$$\bar{X}'(t) = \mu(\rho - 1) - \alpha(\bar{X}(t) - 1)^+ + \mu(\bar{X}(t) - 1)^-. \quad (\text{Whitt 04})$$

Fluid Scaling and Limiting Regimes

A sequence of systems indexed by the number of servers n .

Fluid scaling

$$\bar{\mathcal{R}}^n(t) = \frac{1}{n} \mathcal{R}^n(t), \quad \bar{\mathcal{Z}}^n(t) = \frac{1}{n} \mathcal{Z}^n(t),$$

Fluid Scaling and Limiting Regimes

A sequence of systems indexed by the number of servers n .

Fluid scaling

$$\bar{\mathcal{R}}^n(t) = \frac{1}{n} \mathcal{R}^n(t), \quad \bar{\mathcal{Z}}^n(t) = \frac{1}{n} \mathcal{Z}^n(t),$$

Arrival rate of the n th system $\lambda^n \sim n\lambda$.

$$\rho^n = \frac{\lambda^n}{n\mu_n} \rightarrow \rho \in (0, \infty)$$

Fluid Scaling and Limiting Regimes

A sequence of systems indexed by the number of servers n .

Fluid scaling

$$\bar{\mathcal{R}}^n(t) = \frac{1}{n} \mathcal{R}^n(t), \quad \bar{\mathcal{Z}}^n(t) = \frac{1}{n} \mathcal{Z}^n(t),$$

Arrival rate of the n th system $\lambda^n \sim n\lambda$.

$$\rho^n = \frac{\lambda^n}{n\mu_n} \rightarrow \rho \in (0, \infty) \left\{ \begin{array}{l} > 1, \text{ED} \\ = 1, \text{QED} \\ < 1, \text{QD} \end{array} \right.$$

Fluid Scaling and Limiting Regimes

A sequence of systems indexed by the number of servers n .

Fluid scaling

$$\bar{\mathcal{R}}^n(t) = \frac{1}{n} \mathcal{R}^n(t), \quad \bar{\mathcal{Z}}^n(t) = \frac{1}{n} \mathcal{Z}^n(t),$$

Arrival rate of the n th system $\lambda^n \sim n\lambda$.

$$\rho^n = \frac{\lambda^n}{n\mu_n} \rightarrow \rho \in (0, \infty) \begin{cases} > 1, \text{ED} \\ = 1, \text{QED} \\ < 1, \text{QD} \end{cases}$$

Constraints

$$\bar{Q}^n(t) = (\bar{X}^n(t) - \mathbf{1})^+, \quad \bar{Z}^n(t) = (\bar{X}^n(t) \wedge \mathbf{1})$$

Functional Law of Large Numbers

Assumption A:

- 1 $\bar{E}^n(\cdot) \Rightarrow \lambda \cdot$
- 2 $\vartheta_F^n \rightarrow \vartheta_F, \vartheta_G^n \rightarrow \vartheta_G$
- 3 $\mu^n \rightarrow \mu$
- 4 $(\bar{\mathcal{R}}^n(0), \bar{\mathcal{Z}}^n(0)) \Rightarrow (\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0)$
- 5 $\bar{\mathcal{R}}_0$ and $\bar{\mathcal{Z}}_0$ has no atoms

THEOREM

Under assumption A

$$(\bar{\mathcal{R}}^n(\cdot), \bar{\mathcal{Z}}^n(\cdot)) \Rightarrow (\bar{\mathcal{R}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \quad \text{as } n \rightarrow \infty,$$

where $(\bar{\mathcal{R}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ is almost surely the fluid model solution to $(\lambda, F, G, 1)$ with initial condition $(\bar{\mathcal{R}}_0, \bar{\mathcal{Z}}_0)$.

Performance Evaluation

Approximation formulas

- $\mathbb{E}(W|S) = w, \quad F(w) = \max((\rho - 1)/\rho, 0)$
- $\mathbb{E}(Q) = \frac{\lambda}{\alpha} F_e(w)$

Performance Evaluation

Approximation formulas

- $\mathbb{E}(W|S) = w, \quad F(w) = \max((\rho - 1)/\rho, 0)$
- $\mathbb{E}(Q) = \frac{\lambda}{\alpha} F_e(w)$

$M/GI/100-GI, \lambda = 120, \mu = 1, \alpha = 1$ (Whitt 2006)

Abd.	Ser.	$\mathbb{E}[Q]$	$\mathbb{E}[W S]$
E_2	E_2	40.25 ± 0.057	0.353 ± 0.00051
	LN(1, 4)	39.56 ± 0.097	0.343 ± 0.00094
Approximation		41.11	0.365
LN(1, 4)	E_2	14.51 ± 0.018	0.126 ± 0.00017
	LN(1, 4)	14.52 ± 0.043	0.125 ± 0.00027
Approximation		14.63	0.131

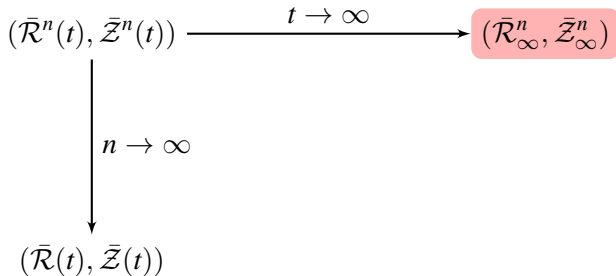
A Missing Gap

Interchange of *Steady State* and *Heavy Traffic* Limits

$$(\bar{\mathcal{R}}^n(t), \bar{\mathcal{Z}}^n(t)) \xrightarrow{t \rightarrow \infty} (\bar{\mathcal{R}}_\infty^n, \bar{\mathcal{Z}}_\infty^n)$$

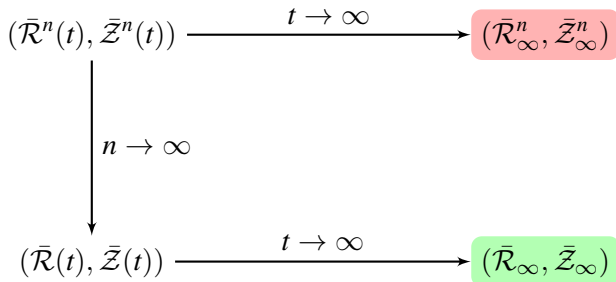
A Missing Gap

Interchange of *Steady State* and *Heavy Traffic* Limits



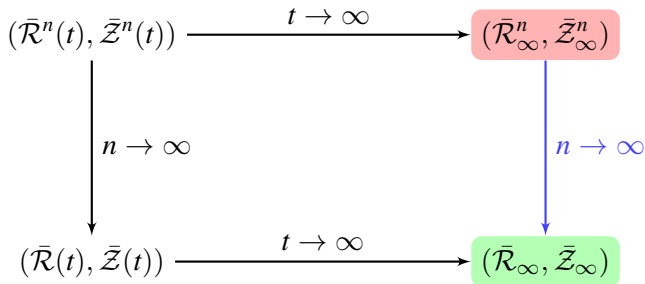
A Missing Gap

Interchange of *Steady State* and *Heavy Traffic* Limits



A Missing Gap

Interchange of *Steady State* and *Heavy Traffic* Limits



Questions?

Thank you!