Managing Time-Varying Service Systems with Flexible and Inflexible Staffing

Yunan Liu
(joint with Ward Whitt)

Department of Industrial and Systems Engineering
NC State University

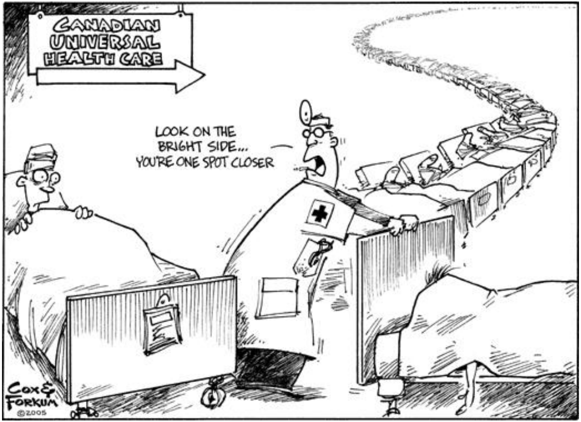
SAMSI Workshop, August 28, 2012

Motivation

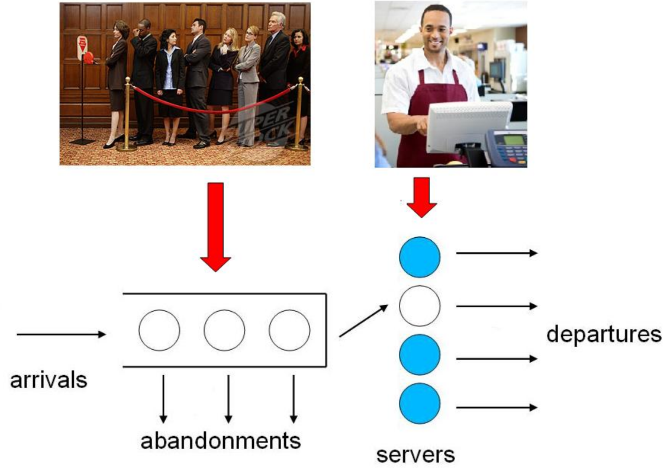
Call centers

Motivation

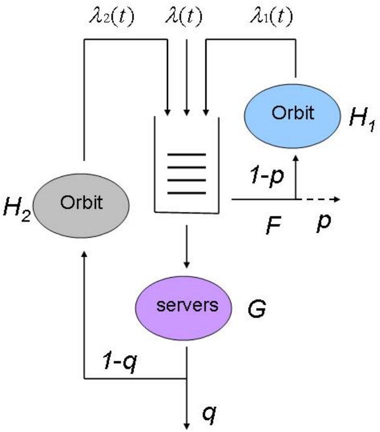
Health care



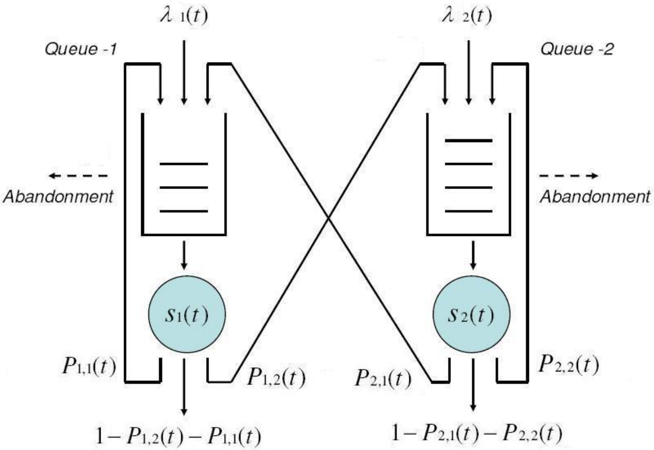
Queueing Models



Queueing Models

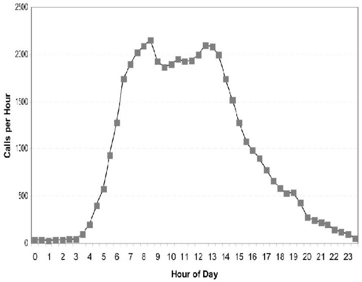


Queueing Models



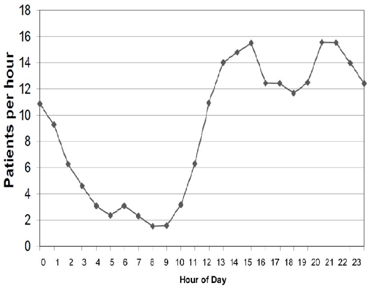
Realistic Models Features

Time-varying arrivals



call center

Green et al. (2007)

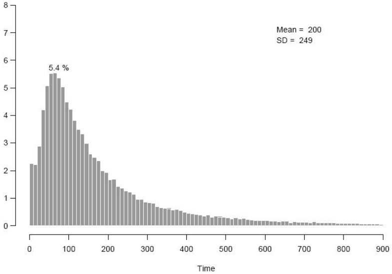


emergence room

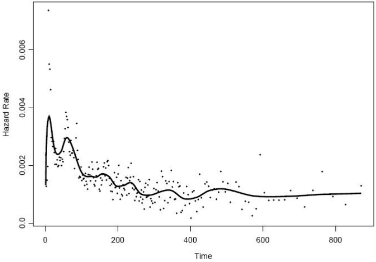
Yom-Tov and Mandelbaum (2011)

Realistic Models Features

Non-exponential service and abandonment



service



abandonment

Brown et al. (2005)

The Base Queueing Model

$\mathbf{M_t / GI / s_t + GI}$

- Poisson with a Time-varying arrival rate $\lambda(t)$ (the $\mathbf{M_t}$)
- I.I.D. service times $\sim G(x) \equiv P(S \le x)$ (the first **GI**)
- Time-varying staffing level $s(t)$ (the $\mathbf{s_t}$)
- I.I.D. abandonment times $\sim F(x) \equiv P(A \le x)$ (the $+$**GI**)
- First-Come First-Served (FCFS)
- Unlimited waiting capacity

Performance measures of interest

Manager's perspective:

- $Q(t)$: number of customers waiting in queue at t
- $B(t)$: number of customers in service at t
- $X(t) \equiv Q(t) + B(t)$: total number in system at t

Customer's perspective:

- $W(t)$: elapsed head-of-line waiting time at t
- $V(t)$: potential waiting time of a virtual customer at t
- $P_t(Delay) \equiv P(B(t) = s(t))$: probability of delay at t
- $P_t(Ab) \equiv P(W(t) > A)$: probability of abandonment at t.

Design Staffing Functions to Stabilize Performance

Objective

Service Level Agreements (SLA)

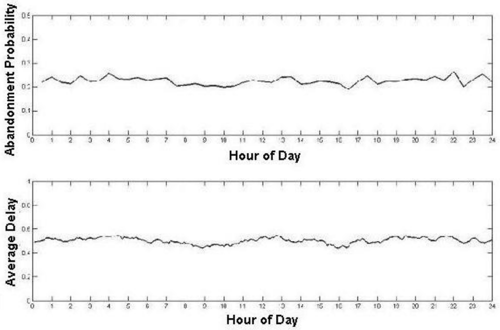
- $\mathbb{P}$(waiting $<$ 30 seconds)$>$0.8
- $\mathbb{E}$(wait)$<$30 seconds
- $\mathbb{P}$(abandonment)$<$0.02

Objective

Service Level Agreements (SLA)

- $\mathbb{P}$(waiting $<$ 30 seconds)$>$0.8
- $\mathbb{E}$(wait)$<$30 seconds
- $\mathbb{P}$(abandonment)$<$0.02

Goal: Staff to cope with arrivals and achieve SLA



Selected Literature

- **Pointwise Stationary Approximation**
  Green and Kolesar (91,97,01)
  - Short service-time, high quality-of-service

- **Modified Offered Load**
  Jagerman (75); Jennings et al.(96);
  Massey and Whitt (94,97); Feldman et al.(08)
  - Long service-time, high quality-of-service

- **Simulation-based Iterative Staffing Algorithm**
  Feldman et al.(08)
  - Stabilize probability of delay

- **Erlang-R Model**
  Yom-Tov and Mandelbaum (2011)

Specific Aim

Relating Waiting time with Delay Probability:

- Potential delay $W(t)$
- Abandonment probability $\mathbb{P}_t(Ab)$
- $\mathbb{P}_t(Ab) = \mathbb{P}(A \leq W(t)) = \mathbb{E}[F(W(t))]$

Objective:

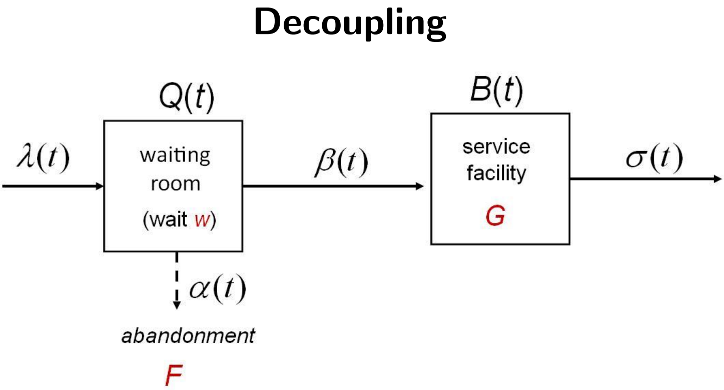
- *Input*: Given $\{\lambda(t), 0 \leq t \leq T\}, F, G$
- *Decision*: Find $\{s_t, 0 \leq t \leq T\}$
- *Aim*: For $\mathbf{M_t}/\mathbf{GI}/\mathbf{s_t} + \mathbf{GI}$ model, $\mathbb{E}[W(t)] = w$ and $\mathbb{P}_t(Ab) = \alpha$, for $0 \leq t \leq T$, for $w > 0$, $\alpha > 0$, $\alpha \approx F(w)$

An Approximating Model: Delayed Infinite-Server Model with Abandonment (DISMA)

Delayed $\mathbf{M_t/GI/\infty + GI}$ with delay w

- Poisson arrival process, time-varying rate $\lambda(t)$
- Infinitely many servers
- **Stay w in a waiting room with unlimited capacity**
- While waiting can abandon I.I.D. $A \sim F$
- If not abandoned after w, receive service I.I.D. $S \sim G$

Network of IS Queues



Two $\mathbf{M_t/GI/\infty}$ Models

Waiting Room

- Time-varying arrival rate $\lambda(t)$
- I.I.D. Service times $T = A \wedge w$, $A \sim F$
- $Q(t) \sim \text{Poisson}(E[Q(t)])$
- $E[Q(t)] = E[\lambda(t - T_e)]E[T], \qquad T = A \wedge w$

Service Facility

- Time-varying arrival rate $\beta(t) = \bar{F}(w)\lambda(t - w)$
- I.I.D. Service times $S \sim G$
- $B(t) \sim \text{Poisson}(E[B(t)])$
- $E[B(t)] = \bar{F}(w)E[\lambda(t - w - S_e)]E[S]$

Offered Load (OL) $\equiv$ **m(t)** $\equiv E[B(t)]$

Modified Offered Load Refinement

For fixed t, $s =$ some s_t, $\lambda =$ some $\lambda^{MOL}(t)$

$$(\mathbf{M_t/GI/s_t+GI}) \approx (\mathbf{M/GI/s+GI})_\mathbf{t}$$

- $\mathbf{M_t/GI/s_t+GI}$: $\lambda(t)$, s_t, $X(t)$
- $\mathbf{M/GI/s+GI}$: λ, s, X_∞

Modified Offered Load Refinement

For fixed t, $s =$ some s_t, $\lambda =$ some $\lambda^{MOL}(t)$

$$(\mathbf{M_t/GI/s_t + GI}) \approx (\mathbf{M/GI/s + GI})_{\mathbf{t}}$$

- $\mathbf{M_t/GI/s_t + GI}$: $\lambda(t)$, s_t, $X(t)$
- $\mathbf{M/GI/s + GI}$: λ, s, X_∞

Question 1: How to find such $\lambda^{MOL}(t)$?

- $\lambda^{MOL}(t) \equiv \frac{m(t)}{(1-\alpha)\mathbb{E}[S]}$ **Little's Law**

Question 2: How to find such s_t? (Aim)

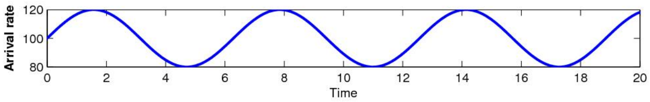
- for a given α, find s_t^{α} s.t. steady-state $P(Ab) \approx \alpha$

  in $\mathbf{M/GI/s + GI}$ with $\lambda = \lambda^{MOL}(t)$
- to compute $P(Ab)$, use approximation $\mathbf{M/GI/s + GI} \approx \mathbf{M/M/s + M(n)}$ *Whitt (2005)*

A Markovian Example

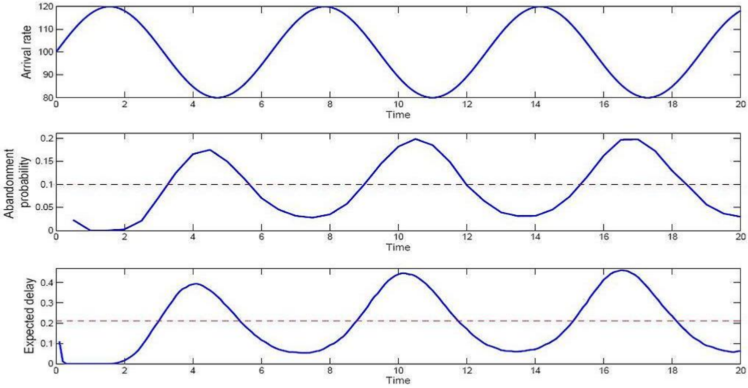
$\mathbf{M_t/M/s_t+M}$ with sinusoidal arrival rate

- $\lambda(t) = 100 + 20 \cdot \sin(t)$



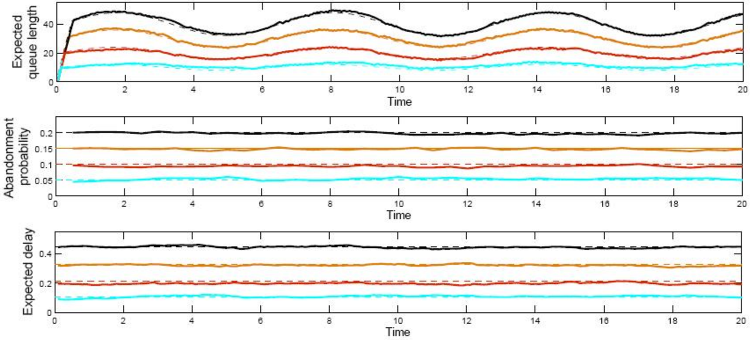
- $\bar{G}(x) = e^{-\mu x}$, $\mu = 1$
- $\bar{F}(x) = e^{-\theta x}$, $\theta = 0.5$

PSA is Bad


Arrival rate
Time
Abandonment probability
Time
Expected delay
Time

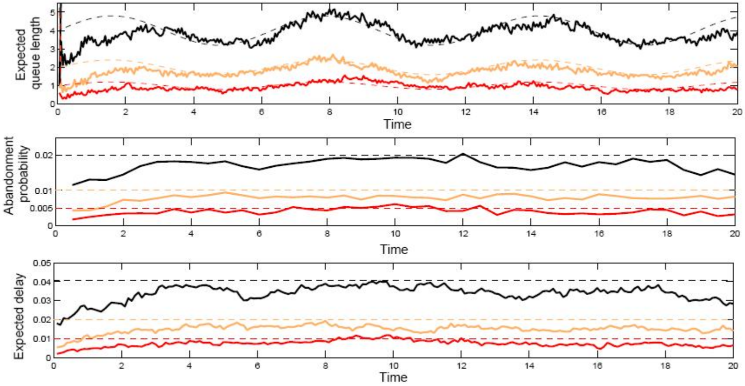
Simulation Verification

Heavy load: $5\% \leq \alpha \leq 20\%$

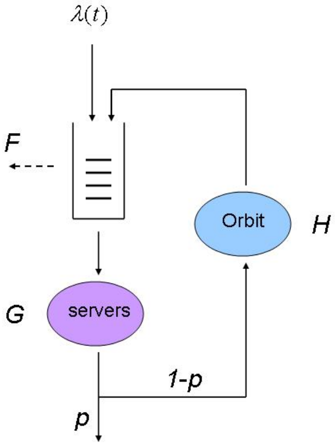


Simulation Verification

Light load: $0.5\% \leq \alpha \leq 2\%$



An Extension: Queues with Feedback



An Extension: Queues with Feedback

Step 1: Apply DISMA Approximation

- $E[Q_1(t)] = E[\lambda(t - T_e)]E[T]$
- $E[B_1(t)] = \bar{F}(w)E[\lambda(t - w - S_e)]E[S]$
- $E[O(t)] = (1 - p)E[\sigma_1(t - U_e)]E[U]$
- $E[Q_2(t)] = E[\lambda_F(t - T_e)]E[T]$
- $E[B_2(t)] = \bar{F}(w)E[\lambda_F(t - w - S_e)]E[S]$

Define the OL function $m(t) \equiv E[B_1(t)] + E[B_2(t)]$

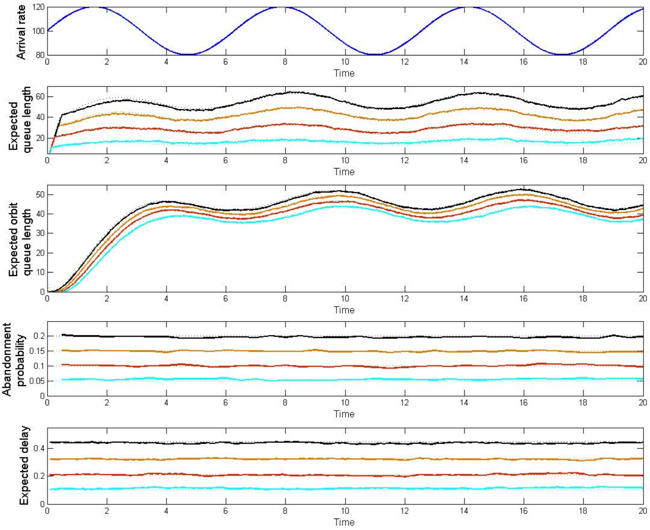
Step 2: Apply MOL refinement to $m(t)$

An Extension: Queues with Feedback

An $\mathbf{M_t/M/s_t + M}$ Example

- $\lambda(t) = 100 + 20 \cdot \sin(t)$
- $\bar{G}(x) = e^{-x}$
- $\bar{F}(x) = e^{-0.5\,x}$
- $\bar{H}(x) = e^{-x}$
- $\alpha = [0.05,\ 0.1,\ 0.15,\ 0.2]$

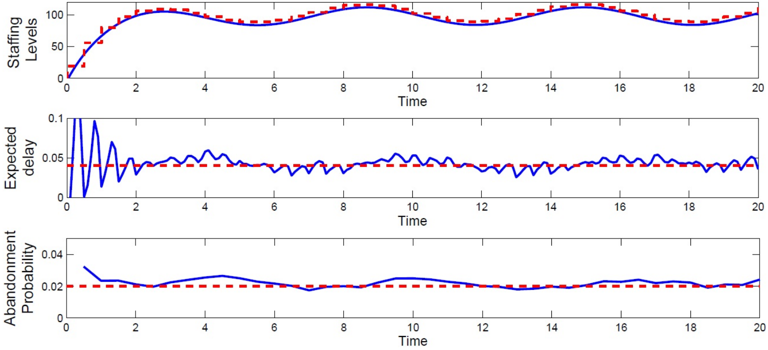
An Extension: Queues with Feedback



Approximating Performance in Systems Experiencing Periods of Underloading and Overloading

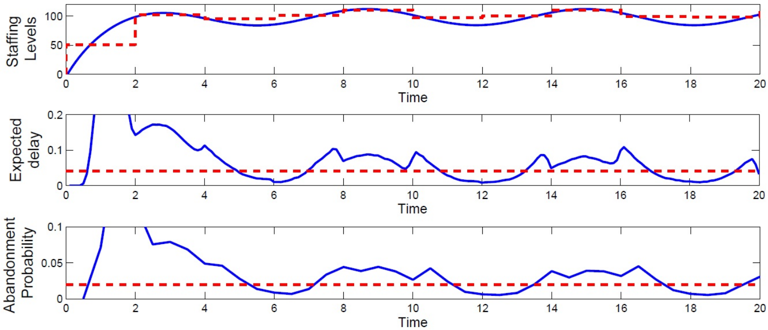
Staffing without Flexibility

$\alpha = 2\%$, staffing interval is 30 minutes



Staffing without Flexibility

$\alpha = 2\%$, staffing interval is 2 hours



Fluid Model

$$G_t \;/\; GI \;/\; s_t + GI$$

arrival	service	staffing	abandonment
$\lambda(t)$	cdf G	$S(t)$	cdf F

What Are Fluid Models

What Are Fluid Models

Bank of America

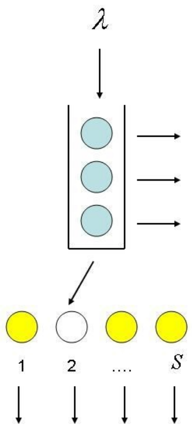
RALEIGH ROCKS

HALF MARATHON

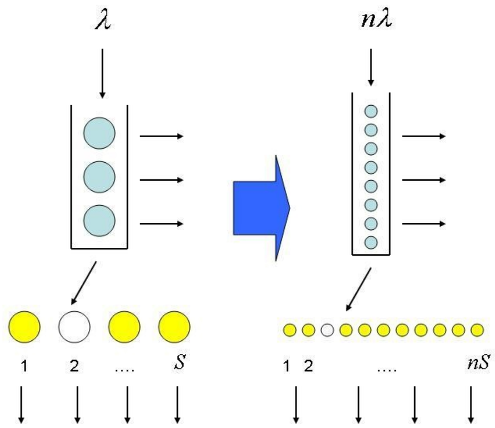
April 1, 2012

NYC Marathon (Nov.4, 2011)

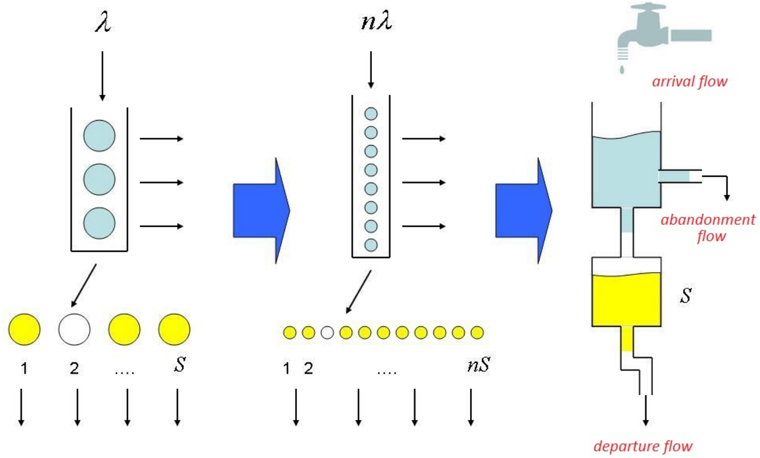
MSHT Fluid Limit



MSHT Fluid Limit



MSHT Fluid Limit



Two-Parameter Fluid Functions

Fluid content

- $Q(t,y)$: quantity of fluid in queue for up to y at t
  $\equiv \int_0^y q(t,x)dx$
- $B(t,y)$: quantity of fluid in service for up to y at t
  $\equiv \int_0^y b(t,x)dx$

Two-Parameter Fluid Functions

Fluid content

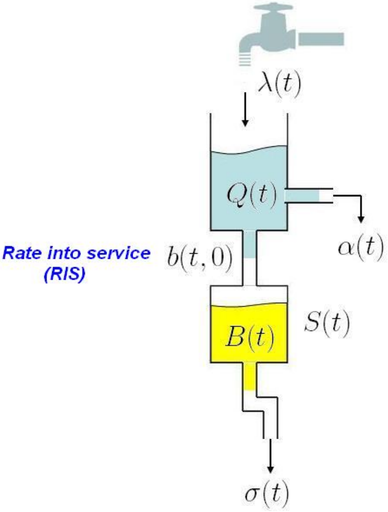
- $Q(t, y)$: quantity of fluid in queue for up to y at t $\equiv \int_0^y q(t, x) dx$
- $B(t, y)$: quantity of fluid in service for up to y at t $\equiv \int_0^y b(t, x) dx$

Rate Functions

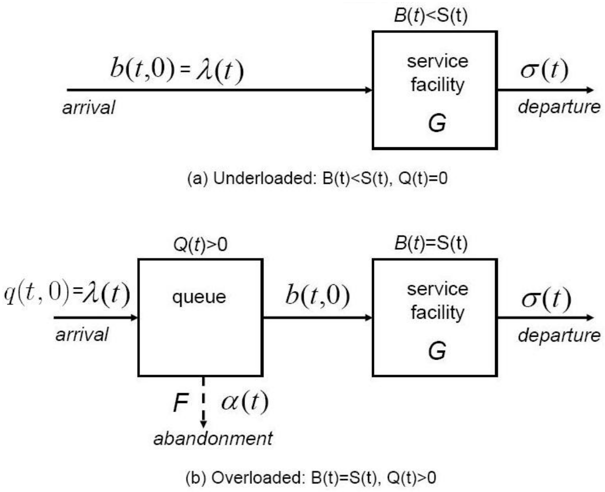
- Service completion rate: $\sigma(t) \equiv \int_0^\infty b(t, x) h_G(x) dx$
- Abandonment rate: $\alpha(t) \equiv \int_0^\infty q(t, x) h_F(x) dx$

where $h_F(x) \equiv \frac{f(x)}{\bar{F}(x)}$, $h_G(x) \equiv \frac{g(x)}{\bar{G}(x)}$

Flow Rates



A Simple Algorithm: Alternate UL and OL Regimes



- System underloaded for $t \in [0, t_1]$, overloaded for $t \in [t_1, t_2]$, ... Advance in time recursively.

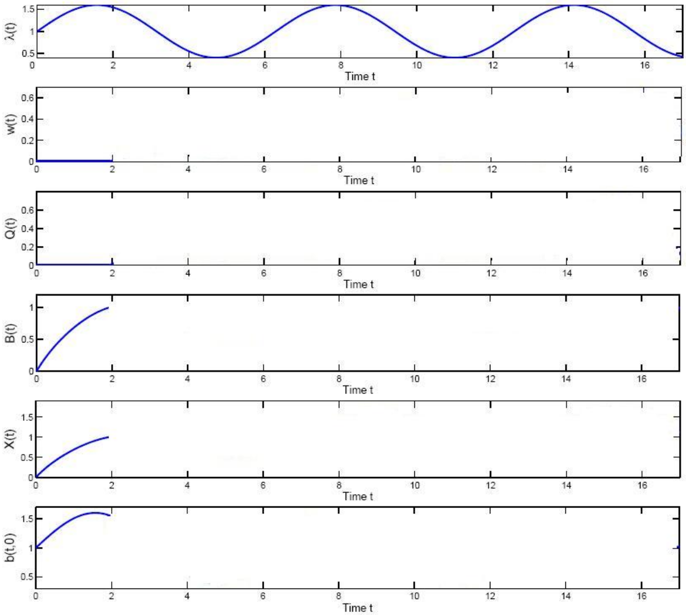
A Non-Markovian Example

$\mathbf{M_t/LN/s_t + E_2}$ fluid model

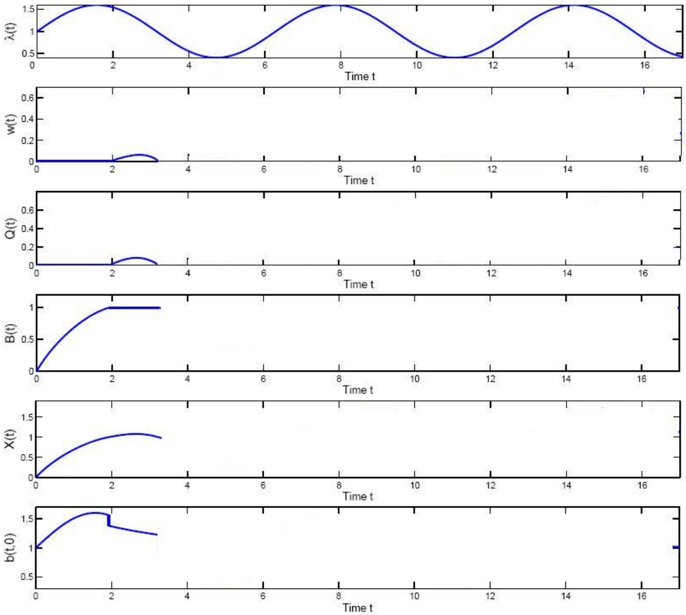
- $\lambda(t) = 1 + 0.6 \cdot \sin(t)$
- $S = 1$ (note: not a single-server queue)
- LN service: $1/\mu = 1$, $\sigma^2 = 4$ ($C_s^2 = 4$)
- E_2 abandonment: $A = X_1 + X_2$, where X_i i.i.d. $\sim \exp(1)$
- System initially empty

$\lambda(t)$ **and** S **will be scaled by** n **!**

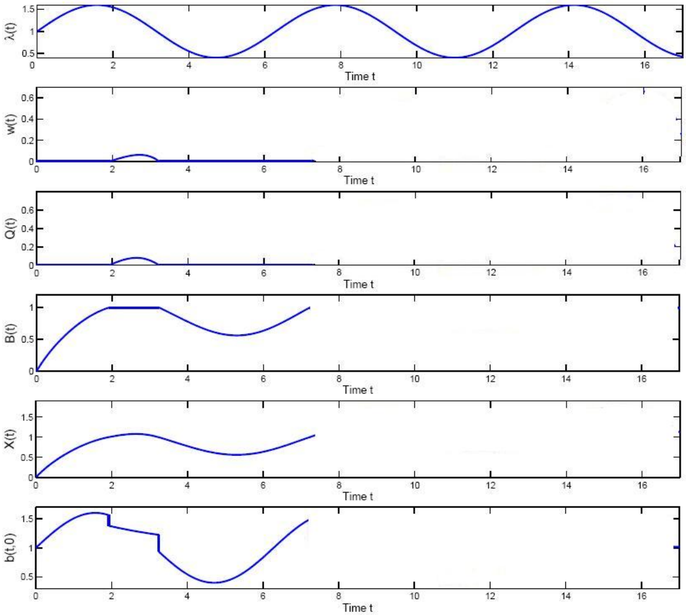
Fluid Algorithm: Alternating between OL and UL



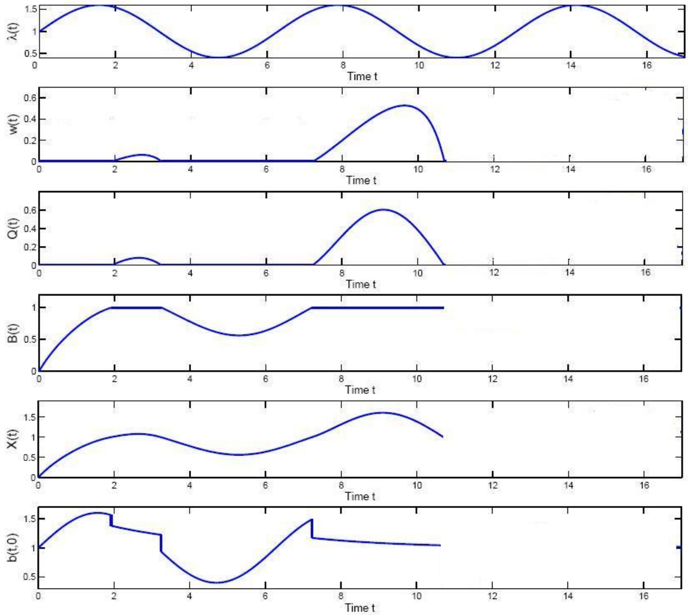
Fluid Algorithm: Alternating between OL and UL



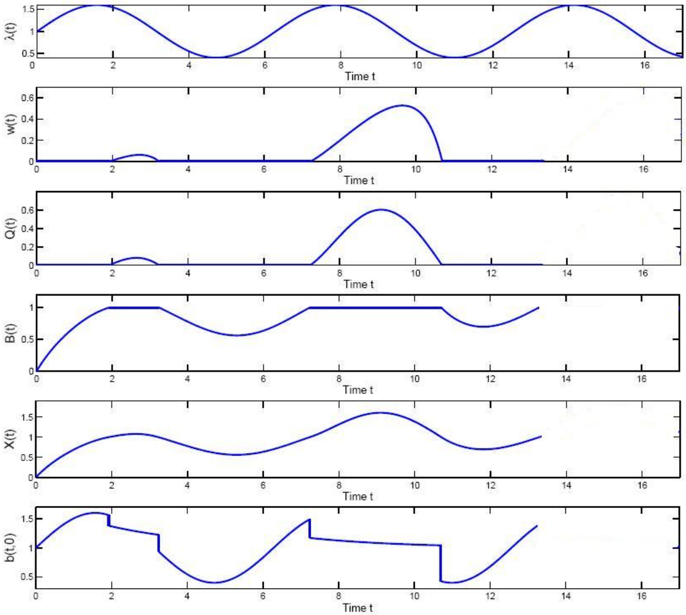
Fluid Algorithm: Alternating between OL and UL



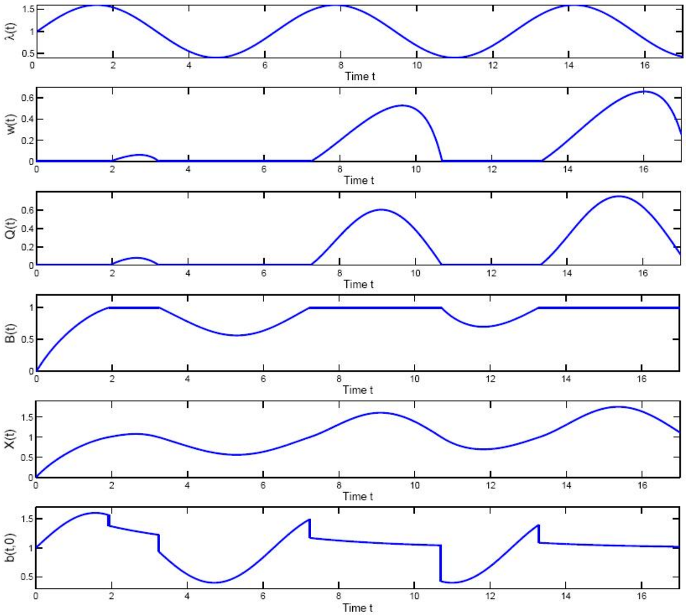
Fluid Algorithm: Alternating between OL and UL



Fluid Algorithm: Alternating between OL and UL



Fluid Algorithm: Alternating between OL and UL



Simulation Comparisons

$\mathbf{M_t/LN/s_t + E_2}$ queueing model

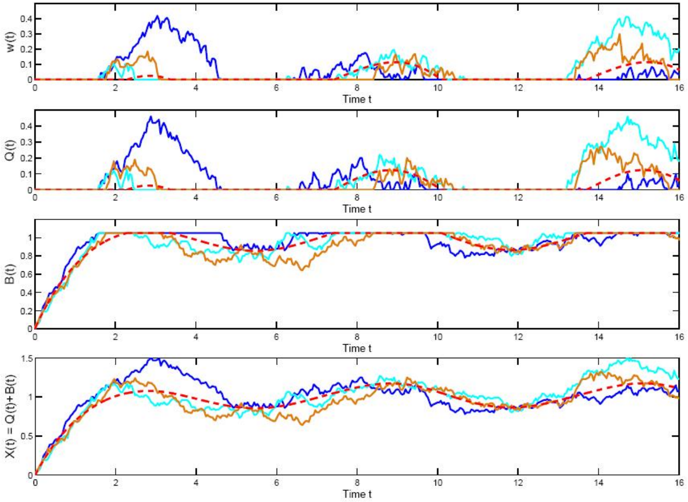
- $n = 20, 100, 2000$
- $\lambda_n(t) = n \cdot \lambda(t) = n + 0.6\, n \sin(t)$
- $S_n(t) = \lceil n\, S(t) \rceil = n$

Want to see

- When n is large:
$$\left(\frac{Q_n(t)}{n}, \frac{B_n(t)}{n}, \frac{X_n(t)}{n}, W_n(t)\right) \approx (Q(t), B(t), X(t), w(t))$$
- When n is small:
$$\left(\frac{E[Q_n(t)]}{n}, \frac{E[B_n(t)]}{n}, \frac{E[X_n(t)]}{n}, E[W_n(t)]\right) \approx (Q(t), B(t), X(t), w(t))$$

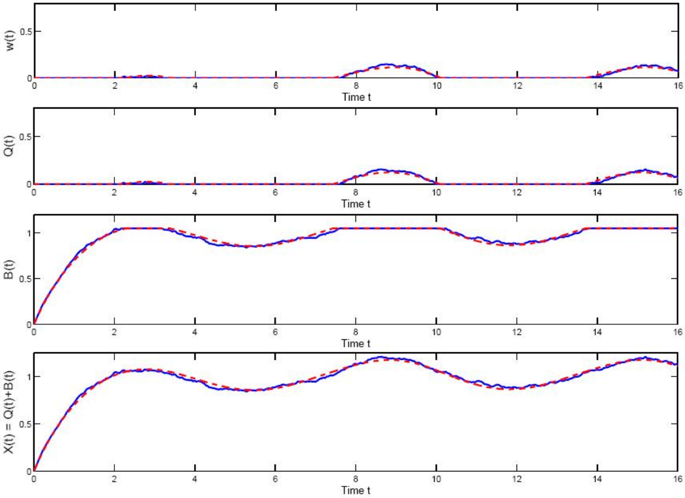
Simulation Comparisons: $\mathbf{M_t/LN/s_t + E_2}$

$n = 100$ **and 3 sample paths**



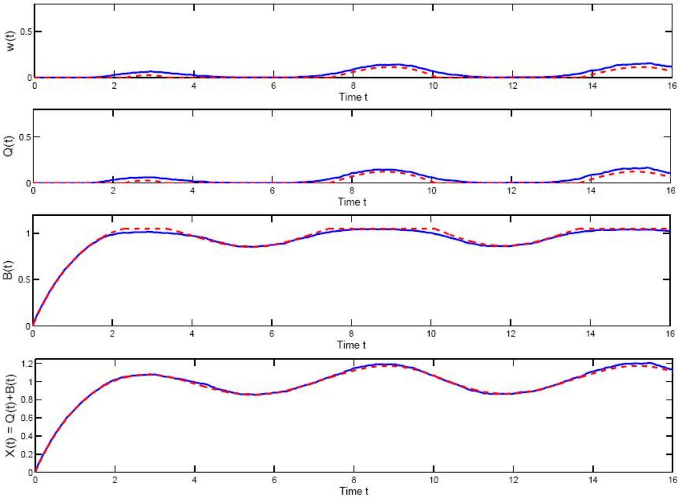
Simulation Comparisons: $\mathbf{M_t/LN/s_t + E_2}$

$n = 2000$ and a single sample path



Simulation Comparisons: $\mathbf{M_t/LN/s_t + E_2}$

$n = 100$ **and a average of 100 sample paths**



Diffusion Limits

$$G_t \,/\, M \,/\, s_t + GI$$

arrival	exponential service cdf	staffing	abandonment
$\lambda(t)$	$G(x) = 1 - e^{-\mu x}$	$S(t)$	cdf F

Separation of Variability

$$d\hat{W}(t) = H(t)\hat{W}(t)dt + J_s(t)d\mathcal{B}_s(t) + J_a(t)d\mathcal{B}_a(t) + J_\lambda(t)d\mathcal{B}_\lambda(t)$$
$$= H(t)\hat{W}(t)dt + J^*(t)d\mathcal{B}^*(t)$$

Independent Brown Motions

- $\mathcal{B}_\lambda$: arrival process
- $\mathcal{B}_s$: service times
- $\mathcal{B}_a$: abandonment times

Separation of Variability

$$d\hat{W}(t) = H(t)\hat{W}(t)dt + J_s(t)d\mathcal{B}_s(t) + J_a(t)d\mathcal{B}_a(t) + J_\lambda(t)d\mathcal{B}_\lambda(t)$$
$$= H(t)\hat{W}(t)dt + J^*(t)d\mathcal{B}^*(t)$$

Independent Brown Motions

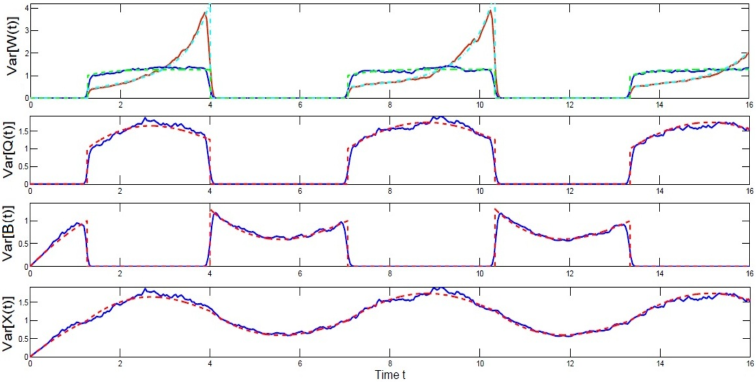
- $\mathcal{B}_\lambda$: arrival process
- $\mathcal{B}_s$: service times
- $\mathcal{B}_a$: abandonment times

Analytic coefficients

- $H(t) = -(1 - w'(t))\left(\frac{\lambda'(t-w(t))}{\lambda(t-w(t))} + h_F(w(t))\right)$
- $J_s(t) = -\frac{\sqrt{b(t,0)-s'(t)}}{\lambda(t-w(t))\bar{F}(w(t))}$
- $J_a(t) = -\frac{\sqrt{F(w(t))b(t,0)}}{\lambda(t-w(t))\bar{F}(w(t))}$
- $J_\lambda(t) = \frac{C_\lambda\sqrt{\bar{F}(w(t))b(t,0)}}{\lambda(t-w(t))\bar{F}(w(t))}$
- $J^*(t) = \frac{\sqrt{b(t,0)-s'(t)+\left(F(w(t))+C_\lambda^2\bar{F}(w(t))\right)b(t,0)}}{\lambda(t-w(t))\bar{F}(w(t))}$

Example: $\mathbf{M_t/M/s_t + H_2}$ in Both UL and OL Intervals

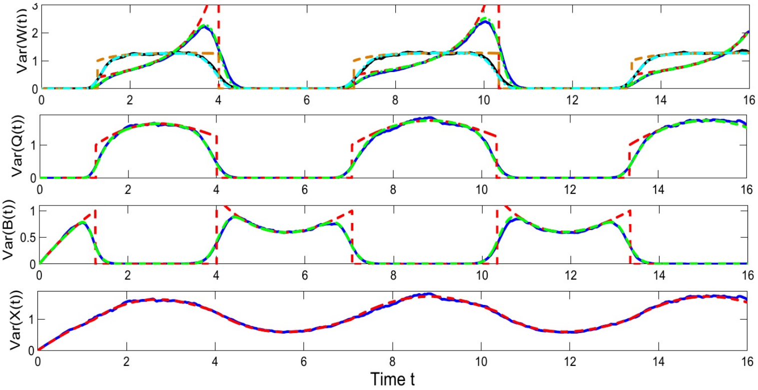
$\lambda(t) = 1 + 0.6\sin(t),\ s(t) = 1,\ \mu = 1,\ \theta = 0.5$



$n = 2000$ **and 500 sample path**

Engineering Refinement for Smaller n

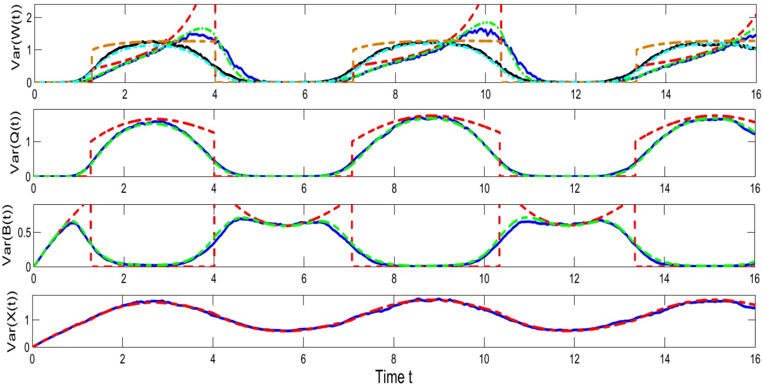
$\lambda(t) = 1 + 0.6\sin(t)$, $s(t) = 1$, $\mu = 1$, $\theta = 0.5$



$n = 100$ **and 2000 sample path**

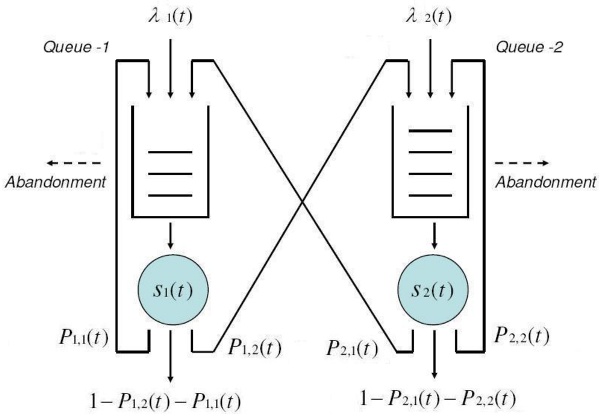
Engineering Refinement for Smaller n

$\lambda(t) = 1 + 0.6\sin(t)$, $s(t) = 1$, $\mu = 1$, $\theta = 0.5$

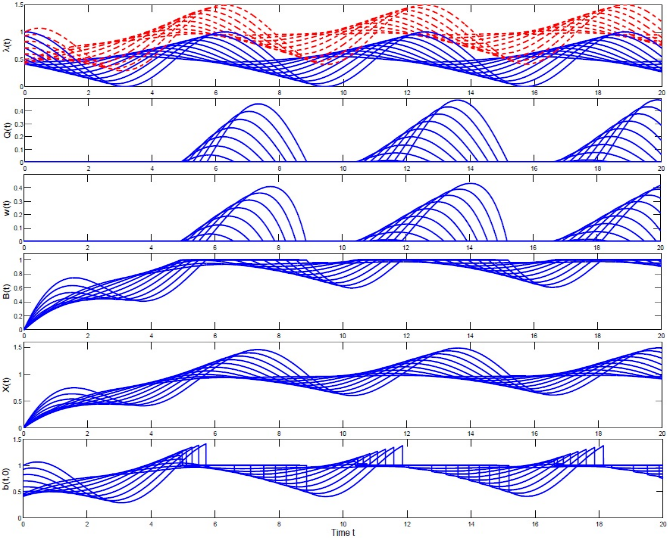


$n = 25$ **and 5000 sample path**

Extension to Networks: $(\mathbf{G_t/GI/s_t + GI})^{\mathbf{m}}/\mathbf{M_t}$



Example: Fluid Paths of $(M_t/M/s_t + M)^{10}/M_t$



Diffusion: multi-dimensional SDEs

Summary

Flexible Staffing

- Develop an approximating model: DISMA
- Provide analytic staffing formulas to stabilize performance
- Conduct simulation evaluation

Inflexible Staffing

- Develop MSHT fluid and diffusion limits
- Provide approximations for mean and variance formulas
- Conduct simulation comparisons
- Extend to network queues

THANK YOU!

References

[1] **L & W**, Stabilizing Customer Abandonment in Many-Server Queues with Time-Varying Arrivals. *Operations Research* (2012)

[2] **L & W**, The $G_t/GI/s_t + GI$ Many-Server Fluid Queue. *Queueing Systems* (2012)

[3] **L & W**, A Many-Server Fluid Limit for the $G_t/GI/s_t + GI$ Queueing Model experiencing Periods of Overloading. *Operations Research Letters* (2012)

[4] **L & W**, A Network of Time-Varying Many-Server Fluid Queues with Customer Abandonment. *Operations Research* **59** 835-846 (2011)

[5] **L & W**, Algorithms for Time-Varying Networks of Many-Server Fluid Queues. Submitted to *INFORMS Journal on Computing* (2012)

[6] **L & W**, Many-Server Heavy-Traffic Limits for Queues with Time-Varying Parameters. Submitted to *Annals of Applied Probability* (2012)

All available at *http://www.ise.ncsu.edu/liu*

Idea of the Proof of MSHT Theorems

- Recursively treat successive UL and OL intervals.
- Infinite-server (IS) MSHT limits (Pang&Whitt 2010) apply directly to treat UL intervals.
- In OL intervals first ignore flow into service; let $\tilde{\mathbf{Q}}_n(t, y)$ be the process.
- IS MSHT limits (Pang&Whitt 2010) apply to treat $\tilde{\mathbf{Q}}_n$ in OL intervals.
- To go from $\tilde{\mathbf{Q}}_n$ to $\tilde{\mathbf{Q}}_n$, focus on HOL waiting time $\mathbf{W}_n$: Equate two representations of the flow into service during OL interval:
  - new space available due to service completion and capacity change;
  - the flow into service from the queue, which occurs from the head of the line.

Fluid Constraints and Regimes

Two constraints

- Capacity constraint: $B(t) \leq S(t)$
- Non-idling constraint: $[B(t) - S(t)] \cdot Q(t) = 0$

Fluid Constraints and Regimes

Two constraints

- Capacity constraint: $B(t) \leq S(t)$
- Non-idling constraint: $[B(t) - S(t)] \cdot Q(t) = 0$

Two system regimes

- Underloaded: $Q(t) = 0$
- Overloaded: $Q(t) > 0$ (and $B(t) = S(t)$)

Flow Rates

Given $q(t,x)$ **and** $b(t,x)$

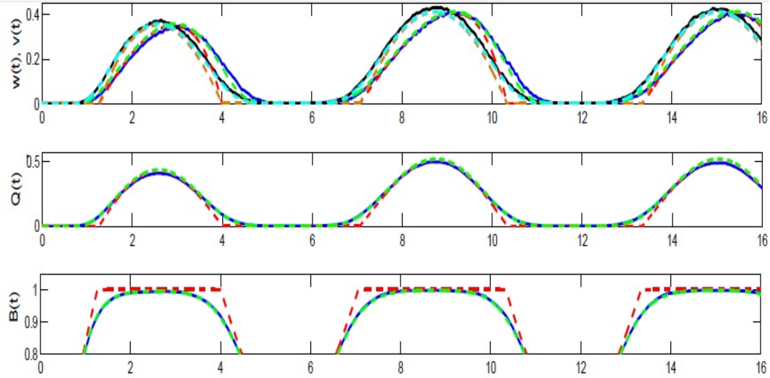
- Service completion rate: $\sigma(t) \equiv \int_0^\infty b(t,x) h_G(x) dx$
- Abandonment rate: $\alpha(t) \equiv \int_0^\infty q(t,x) h_F(x) dx$

where $h_F(x) \equiv \frac{f(x)}{\bar{F}(x)}$, $h_G(x) \equiv \frac{g(x)}{\bar{G}(x)}$

- $q(t,x)$ and $b(t,x)$ determine everything !

Engineering Refinement on Mean Values for Smaller n

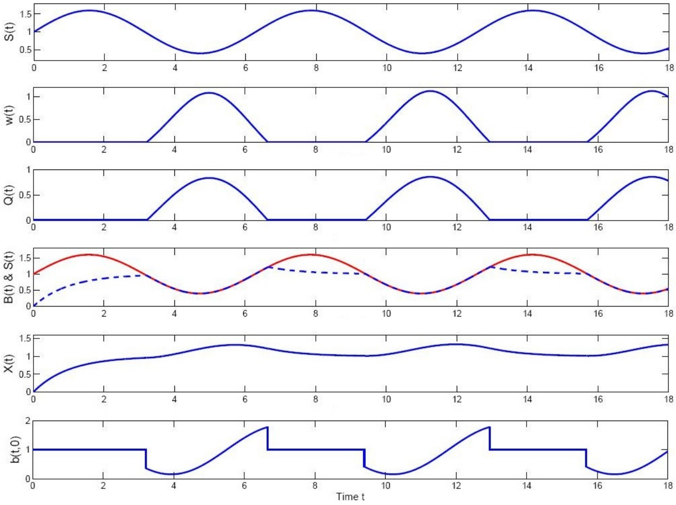
$\lambda(t) = 1 + 0.6\sin(t)$, $s(t) = 1$, $\mu = 1$, $\theta = 0.5$



$n = 25$ **and 5000 sample path**

Example: $\mathbf{M}/\mathbf{M}/\mathbf{s_t} + \mathbf{M}$ Fluid Queue

$\lambda = 1,\ s(t) = 1 + 0.6\sin(t),\ \mu = 1,\ \theta = 0.5.$



Many-Server Heavy-Traffic Limits

Fluid Limit

- LLN scaling: $\bar{Q}_n(t) \equiv \frac{Q_n(t)}{n}$, $\bar{B}_n(t) \equiv \frac{B_n(t)}{n}$, $\bar{X}_n(t) \equiv \frac{X_n(t)}{n}$
- FSLLN:
  $(\bar{Q}_n, \bar{B}_n, \bar{X}_n, W_n) \rightarrow (Q, B, X, W)$ in $\mathbb{D}^4$, as $n \rightarrow \infty$

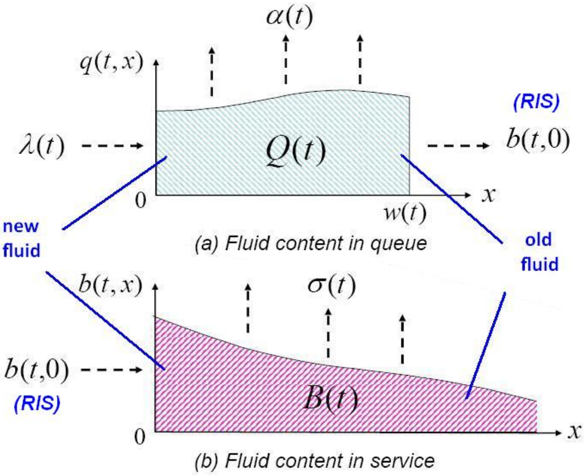
Diffusion Limit

- CLT scaling:
  $\hat{Q}_n(t) \equiv \sqrt{n}\left(\bar{Q}_n(t) - Q(t)\right) = \frac{Q_n(t) - n\,Q(t)}{\sqrt{n}}$,
  $\hat{W}_n(t) \equiv \sqrt{n}\left(W_n(t) - W(t)\right)$
- FCLT:
  $\left(\hat{Q}_n, \hat{B}_n, \hat{X}_n, \hat{W}_n\right) \Rightarrow \left(\hat{Q}, \hat{B}, \hat{X}, \hat{W}\right)$ in $\mathbb{D}^4$, as $n \rightarrow \infty$

Approximations

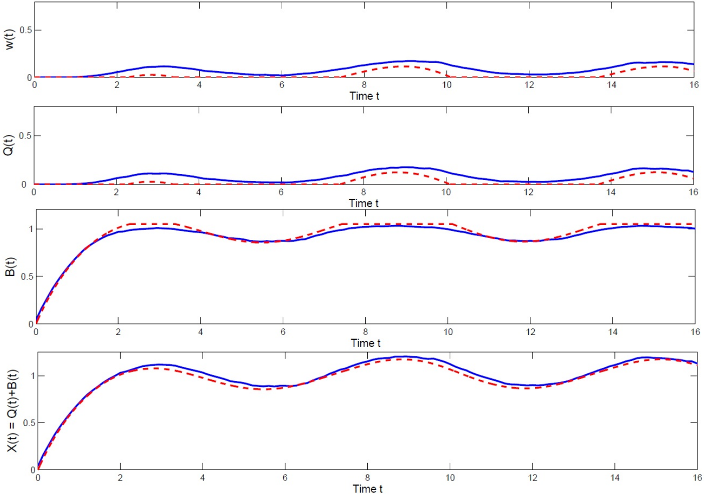
- $Q_n(t) = n\,Q(t) + \sqrt{n}\,\hat{Q}(t) + o(\sqrt{n})$
- $W_n(t) = W(t) + \frac{\hat{W}(t)}{\sqrt{n}} + o(\frac{1}{\sqrt{n}})$

Fluid Densities



Simulation Comparisons: $\mathbf{M_t/LN/s_t + E_2}$

$n = 20$ **and a average of 100 sample paths**



Separation of Variability

Diffusion for the HWT $\hat{W}$

- $\sqrt{n}(W_n - \bar{W}) \Rightarrow \hat{W}$ in $\mathbb{D}$, as $n \to \infty$
- An SDE:
$$d\hat{W}(t) = H(t)\hat{W}(t)dt + J_s(t)d\mathcal{B}_s(t) + J_a(t)d\mathcal{B}_a(t) + J_\lambda(t)d\mathcal{B}_\lambda(t)$$
$$= H(t)\hat{W}(t)dt + J^*(t)d\mathcal{B}^*(t)$$
  - $\mathcal{B}_\lambda$: arrival process
  - $\mathcal{B}_s$: service times
  - $\mathcal{B}_a$: abandonment times
  - H, J_s, J_a, J_λ and J^*: analytic functions of λ, s, F, μ, C^2_λ and fluid functions
- $\sigma^2_{\hat{W}}(t) \equiv Var(\hat{W}(t)) = \int_0^t \left(\hat{J}^2_s(t,u) + \hat{J}^2_a(t,u) + \hat{J}^2_\lambda(t,u)\right) du$

Diffusion for the PWT $\hat{V}$

- $\hat{V}(t) = \frac{\hat{W}(t+v(t))}{1-w'(t+v(t))}$

Two Waiting Times: HWT and PWT

Head-of-Line Waiting Time

- $w(t)$ = elapsed head-of-line (HOL) waiting time at t
- An ODE: $w'(t) = 1 - \frac{b(t,0)}{q(t,w(t))}$

Potential Waiting Time

- $v(t)$ = virtual waiting time of an arrival at t
- $w \to v$: $v(t - w(t)) = w(t)$ or $w(t + v(t)) = v(t)$