Queueing Models for Large-Scale Service Systems

Experiencing Periods of Overloading

Ward Whitt, Columbia University, www.columbia.edu/∼ww2040

Joint work with Yunan Liu, North Carolina State University

MIT, April 5, 2012

My Co-Author Yunan Liu on the Move

The Base Queueing Model

$G_t/GI/s_t + GI$

- Time-varying arrival rate $\lambda(t)$ (the G_t)

  (e.g., non-homogeneous Poisson)

- I.I.D. service times $\sim G(x) \equiv P(S \leq x)$ (the first **GI**)
- Time-varying staffing level $s(t)$ (the s_t)
- I.I.D. abandonment times $\sim F(x) \equiv P(A \leq x)$ (the $+$**GI**)
- First-Come First-Served (FCFS)
- Unlimited waiting capacity

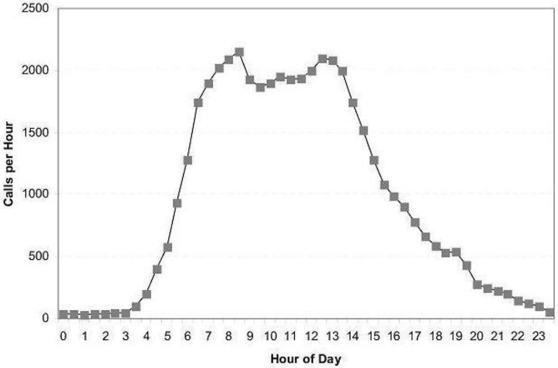
The Base Queueing Model

Performance measures of interest

- $Q(t)$: number waiting in queue at t
- $B(t)$: number in service at t
- $X(t) \equiv Q(t) + B(t)$: total number in system
- $W(t)$: elapsed head-of-line waiting time
- $V(t)$: potential waiting time

Realistic Models Features

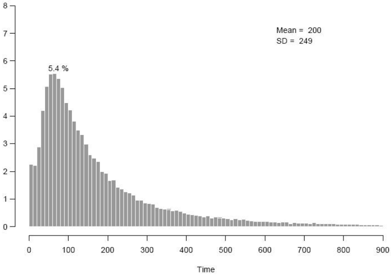
Time-varying arrivals

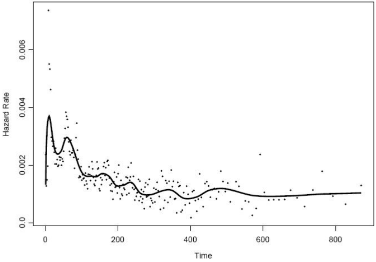


a financial services call center, from Green, Kolesar and Soares (2001)

Realistic Models Features

Non-exponential service and abandonment





service **abandonment**

Brown et al. (2005): *Statistical Analysis of a Telephone Call Center*, JASA.

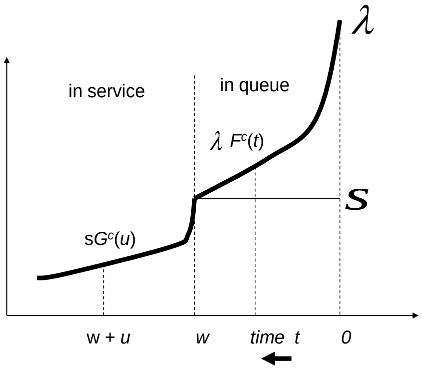
Pictures

A picture is worth n words.

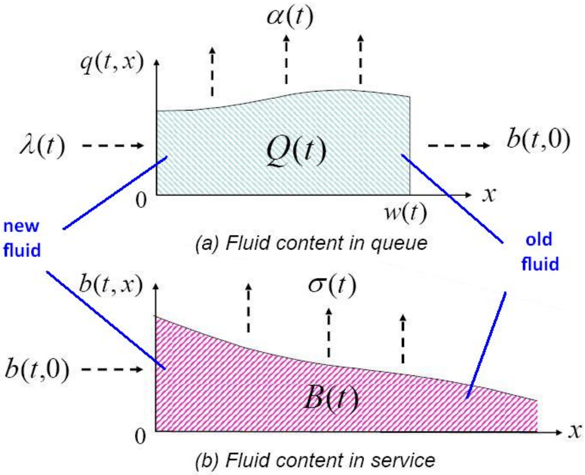
The Overloaded $G/GI/s + GI$ Fluid Queue in Steady State

The 2005 MIT talk. *Operations Research*, 2006, by W^2

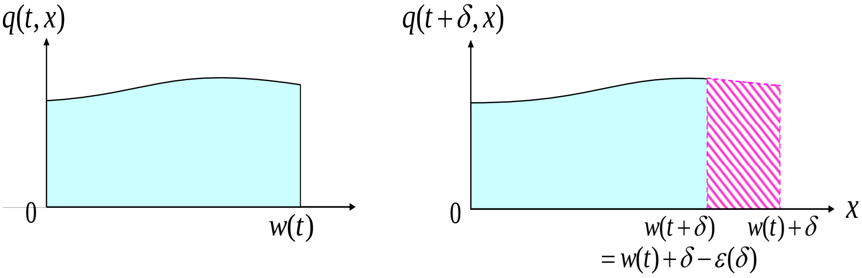
fluid density arriving time *t* in the past



The Evolution of the $G_t/GI/s_t + GI$ Fluid Queue



Deriving the ODE for the Head-of-the-Line Waiting Time



(a) Fluid content at time t

(b) Fluid content at time $t + \delta$

$$-q(t, w(t))(\Delta w(t) - \Delta t) = b(t, 0)\Delta t$$

$$\frac{\Delta w(t)}{\Delta t} = 1 - \frac{b(t,0)}{q(t, w(t))} \quad \text{so that} \quad \dot{w}(t) = 1 - \frac{b(t,0)}{q(t, w(t))}.$$

OUTLINE

1. Staffing Examples, Review: Jennings, Mand., Massey &W^2 (1996)
2. Approximations for Time-Varying Many-Server Queues
3. The $G_t/GI/s_t + GI$ Many-Server Fluid Queue with Alternating Overloaded and Underloaded Intervals (YL&W^2 2012)
4. Numerical Examples: Comparisons with Simulations
5. Extensions
    1. Asymptotic Loss of Memory, Periodic Steady State (YL&W^2 2011)
    2. Networks of Fluid Queues with Proportional Routing (YL&W^2 2011)

Business Case: H&R Block

Service Center to Help Prepare Tax Returns

- How many service representatives are needed?
- arrival rate = 100 per hour
- expected service time = 1 hour
- expected patience time $= \theta^{-1} = 2$ hours
- Erlang-A $(M/M/s+M)$ model
- Performance target: Minimum staffing such that $P(W > 0) \leq 0.20$

Business Case: H&R Block

Service Center to Help Prepare Tax Returns

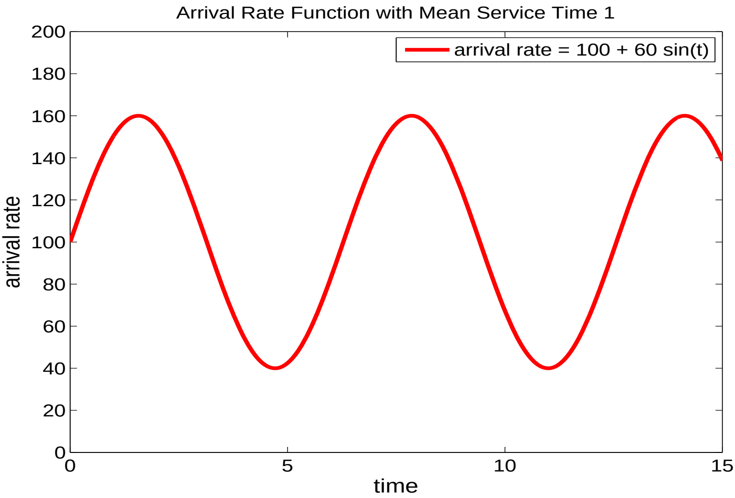
- How many service representatives are needed?
- arrival rate = 100 per hour
- expected service time = 1 hour
- expected patience time $= \theta^{-1} = 2$
- Erlang-A ($M/M/s+M$) model
- offered load = $100 \times 1 = 100$ (expected number used if unlimited)
- Square Root Staffing: $s = 100 + \sqrt{100} = 110$,
- Performance: $P(W > 0) = .19$, $P(W > .05) = .09$, $P(Ab) = .006$,

Time-Varying Arrival Rate

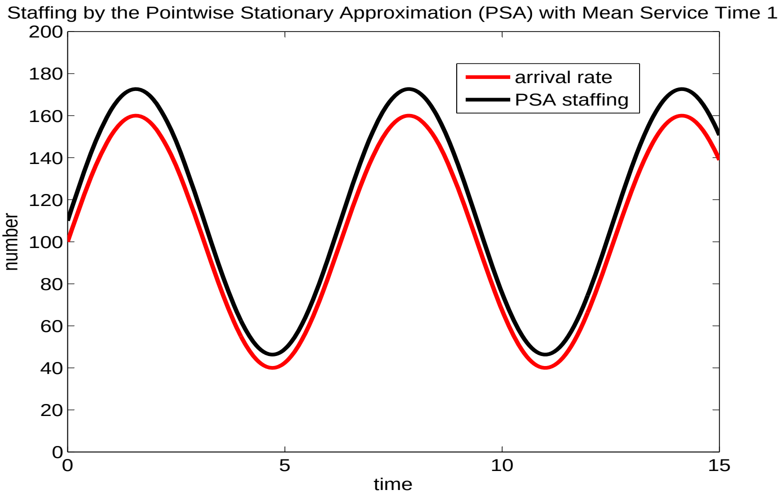
- long-run average arrival rate = 100 per hour
- expected service time = 1 hour
- expected patience time $= \theta^{-1} = 2$
- from $M/M/s+M$ model to $M_t/M/s_t+M$
- How many service representatives are needed now?
- Want to stabilize performance at similar level, e.g., $P(W>0) \approx 0.20$

Jennings et al. (1996), Feldman et al. (2008), YL&W^2 (2012)

The Arrival Rate

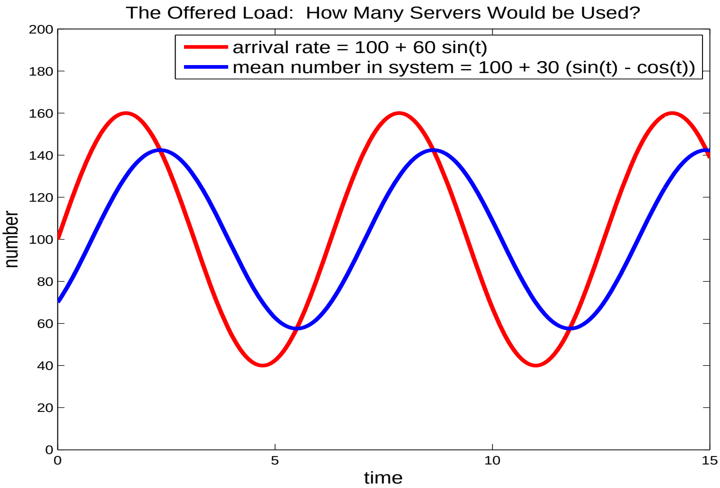


Square Root Staffing with PSA

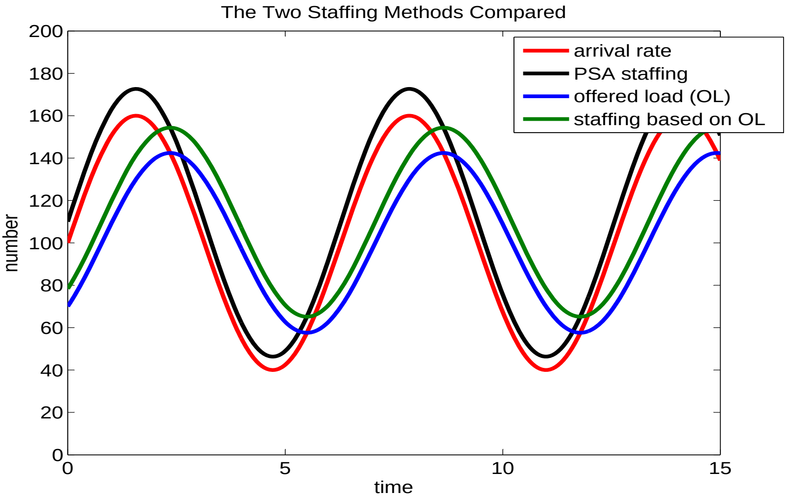


The Offered Load: Expected Number Used If Unlimited

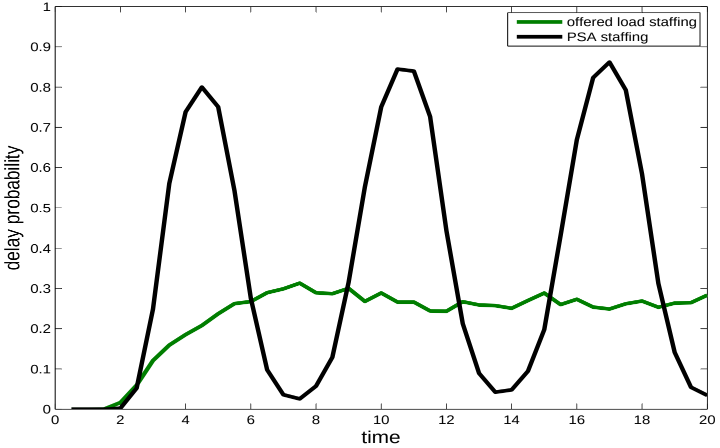
The Physics of . . . , 1993, by S. G. Eick, W. A. Massey & W^2



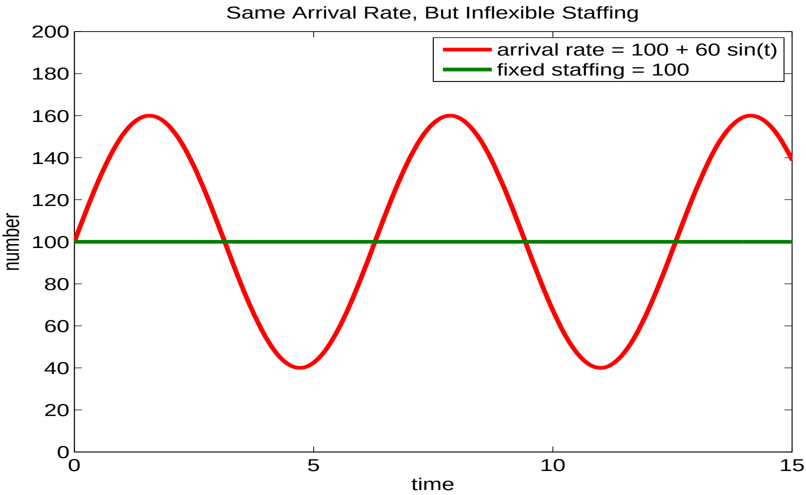
Square Root Staffing with the Offered Load



Simulation Comparison of the Two Staffing Methods



NEW TOPIC: Same Model, But Inflexible Staffing

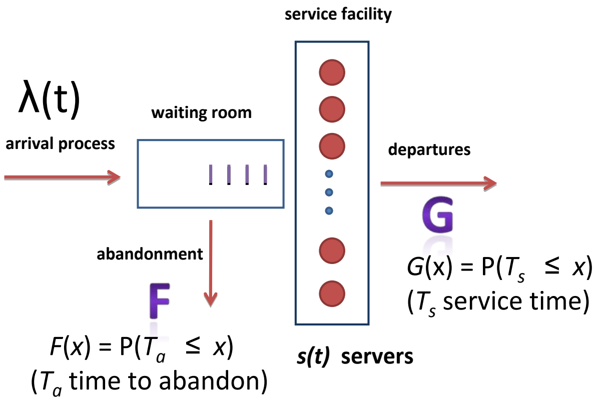


OUTLINE

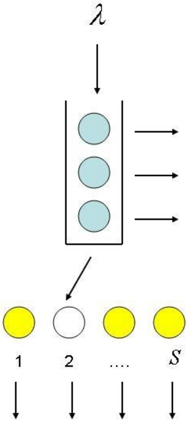
1. Staffing Examples
2. Approximations for Time-Varying Many-Server Queues
3. The $G_t/GI/s_t + GI$ Fluid Queue with Alternating Overloaded and Underloaded Intervals (Yunan Liu & W^2, *Queueing Systems*, 2012)
4. Numerical Examples: Comparisons with Simulations
5. Extensions
    1. Asymptotic Loss of Memory, Periodic Steady State
    2. Networks of Fluid Queues with Proportional Routing

2. Approximations for Many-Server Queues

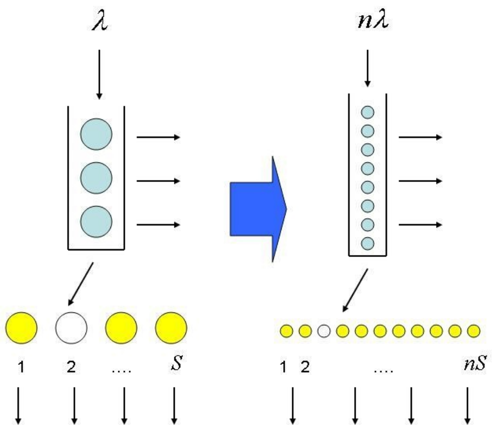
Approximation for the $\mathbf{G_t/GI/s_t + GI}$ **Stochastic Queueing Model**



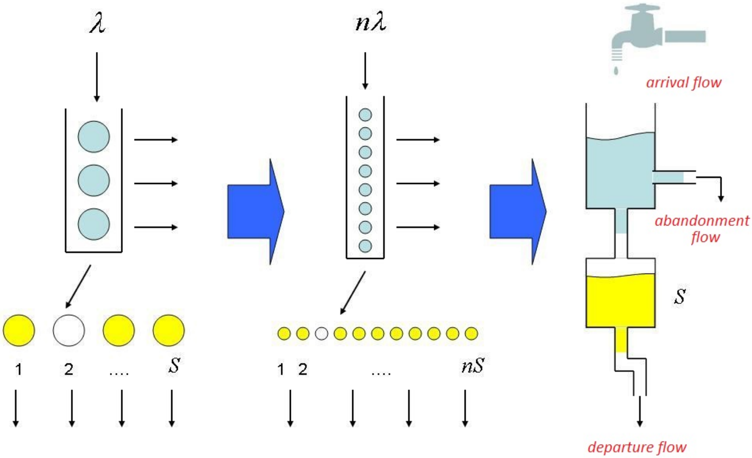
Fluid Approximation from Many-Server Heavy-Traffic limit



Fluid Approximation from MSHT limit



Fluid Approximation from MSHT limit



Many-Server Heavy-Traffic (MSHT) Limit

Increasing Scale

Increasing Scale

- a sequence of $G_t/GI/s_t + GI$ models indexed by n,
- arrival rate **grows**: $\lambda_n(t)/n \rightarrow \lambda(t)$ as $n \rightarrow \infty$,

  number of servers **grows**: $s_n(t)/n \rightarrow s(t)$ as $n \rightarrow \infty$,
- service-time cdf G and patience cdf F held **fixed** independent of n

  with mean service time 1: $\mu^{-1} \equiv \int_0^\infty x\, dG(x) \equiv 1$.

Three Levels of Approximation

- FWLLN: deterministic fluid approximation for mean values
- FCLT: stochastic approximation for full distributions
- Engineering Refinements

Focusing on alternating overloaded and underloaded intervals.

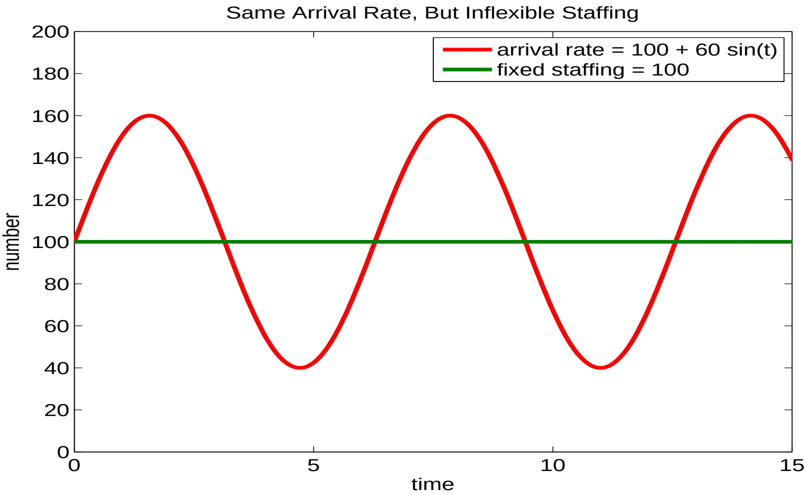
MSHT Diffusion Limit

An SDE for $\hat{W}$ (Separation of Variability)

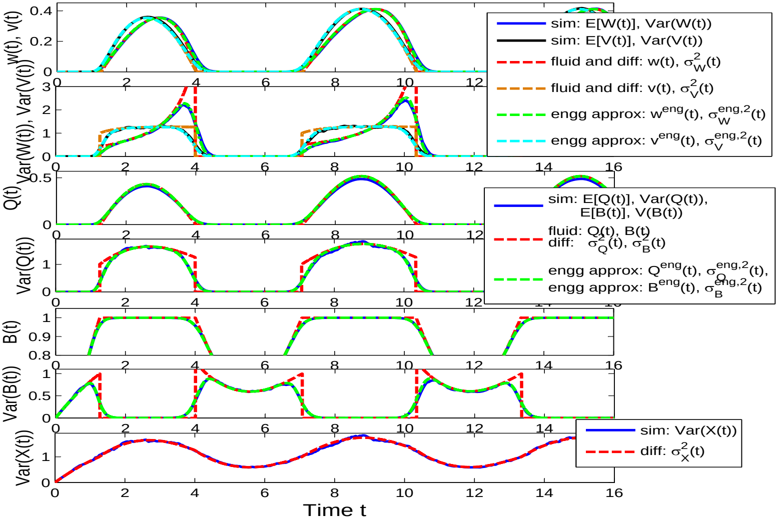
- $\sqrt{n}(W_n - \bar{W}) \Rightarrow \hat{W}$ in $\mathbb{D}$, as $n \to \infty$
- $d\hat{W}(t) = H(t)\hat{W}(t)dt$
  $$+J_s(t)d\mathcal{B}_s(t) + J_a(t)d\mathcal{B}_a(t) + J_\lambda(t)d\mathcal{B}_\lambda(t)$$
  - $\mathcal{B}_\lambda$: arrival process
  - $\mathcal{B}_s$: service times
  - $\mathcal{B}_a$: abandonment times
- $\sigma^2_{\hat{W}}(t) \equiv Var(\hat{W}(t)) = \int_0^t \left(\hat{J}^2_s(t,u) + \hat{J}^2_a(t,u) + \hat{J}^2_\lambda(t,u) \right) du$

Many-Server Heavy-Traffic Limits for ..., YL&W^2 (2011)

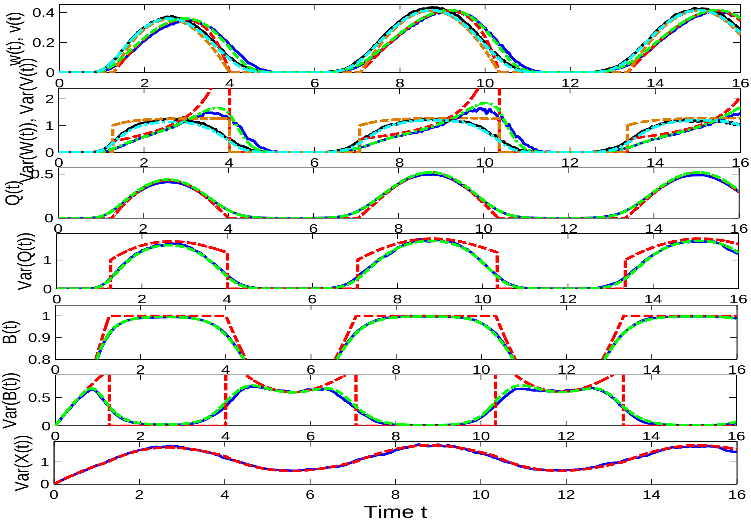
EXAMPLE: Same Model, But Fixed Staffing



Performance for $M_t/M/100 + H_2$ Model, $\theta = 0.5$



Fewer Servers: the $M_t/M/25 + H_2$ Model, $\theta = 0.5$



OUTLINE: Main Talk Begins Here.

1. Staffing Examples
2. Approximations for Time-Varying Many-Server Queues
3. The $G_t/GI/s_t + GI$ Many-Server Fluid Queue with Alternating OL and UL Intervals (Yunan Liu & W^2, *Queueing Systems*, 2012)
4. Numerical Examples: Comparisons with Simulations
5. Extensions
    1. Asymptotic Loss of Memory, Periodic Steady State
    2. Networks of Fluid Queues with Proportional Routing

What Are Fluid Models

What Are Fluid Models

Bank of America

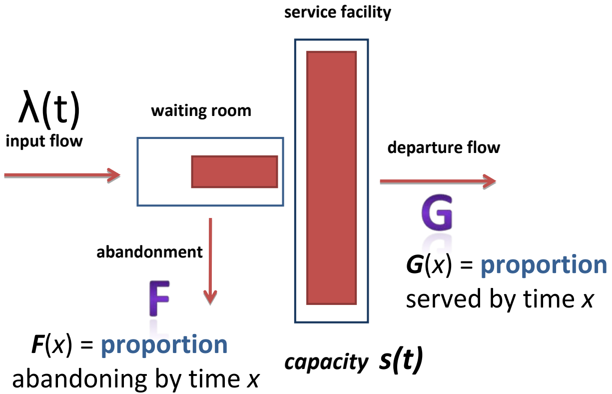
Raleigh Rocks

HALF MARATHON

April 1, 2012

3. The $G_t/GI/s_t + GI$ Fluid Model

Model data: $(\lambda(t), s(t), G(x), F(x))$ and initial conditions.



Important Issue: Feasibility of the Staffing Function

- Assume feasibility of $s(t)$ in the fluid model.
- Algorithm developed to find minimum feasible staffing function.
- Assume server switching in the stochastic model.
- Let customers be forced out in the stochastic model. (rare)

two-parameter functions or time-varying measures

Fluid content

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

Fluid densities

- $b(t,x)dx$ ($q(t,x)dx$) is the quantity of fluid in service (in queue) at time t that have been so for a length of time x.

Model Data

- $\Lambda(t) \equiv \int_0^t \lambda(u)\, du$ – input over $[0, t]$.
- $s(t) \equiv s(0) + \int_0^t s'(u)\, du$ – service capacity at time t.
- $G(x) \equiv \int_0^x g(u)\, du$ – service-time cdf.
- $F(x) \equiv \int_0^x f(u)\, du$ – patience-time cdf.
- $B(0, y) \equiv \int_0^y b(0, x)\, dx$ – initial fluid content in service for up to y.
- $Q(0, y) \equiv \int_0^y q(0, x)\, dx$ – initial fluid content in queue for up to y.

Smooth Model: Assume that $(\Lambda, s, G, F, B(0, \cdot), Q(0, \cdot))$ is differentiable with piecewise-continuous derivative $(\lambda, s', g, f, b(0, \cdot), q(0, \cdot))$.

Key Fluid Dynamics

Fundamental Evolution Equations

- $q(t+u, x+u) = q(t, x) \cdot \frac{\bar{F}(x+u)}{\bar{F}(x)}$,

  $0 \leq x \leq w(t) - u,\ u \geq 0,\ t \geq 0.$

- $b(t+u, x+u) = b(t, x) \cdot \frac{\bar{G}(x+u)}{\bar{G}(x)}$,

  $x \geq 0,\ u \geq 0,\ t \geq 0.$

Flow Rates

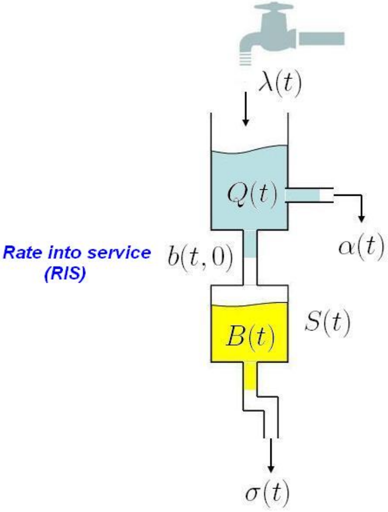
Given $q(t,x)$ **and** $b(t,x)$**,**

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

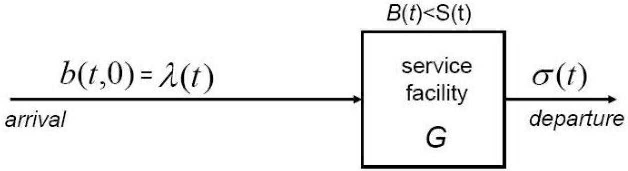
where $h_F(x) \equiv \frac{f(x)}{\bar{F}(x)}$ and $h_G(x) \equiv \frac{g(x)}{\bar{G}(x)}$

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

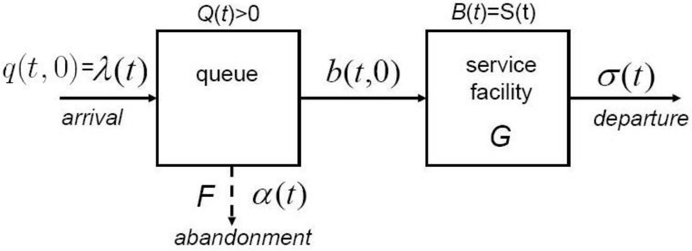
Flow Rates



Two Cases: Underloaded Intervals and Overloaded Intervals



(a) Underloaded: B(t)<S(t), Q(t)=0



(b) Overloaded: B(t)=S(t), Q(t)>0

First (Easy) Case: Underloaded Interval

$$B(t,y) \quad \text{in} \quad G_t/GI/s_t + GI \quad \text{fluid model}$$

$$\Longleftrightarrow B(t,y) \quad \text{in} \quad G_t/GI/\infty \quad \text{fluid model}$$

$$\Longleftrightarrow B(t,y) \quad \text{in} \quad M_t/GI/\infty \quad \text{fluid model}$$

$$\Longleftrightarrow E[B(t,y)] \quad \text{in} \quad M_t/GI/\infty \quad \text{stochastic model}$$

We have formulas already (Eick, Massey & W^2, 1993).

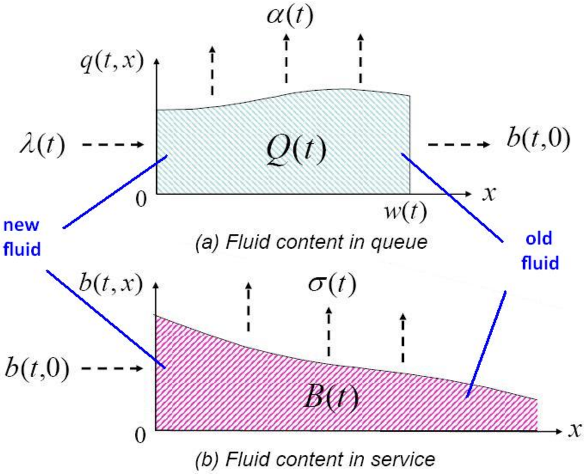
Second (Interesting) Case: Overloaded Interval

- Minimum feasible staffing function s^* exceeding s.
- b satisfies fixed-point equation.

  (Apply Banach contraction fixed point theorem.)
- w satisfies an ODE.
- PWT v obtained from BWT w via the equation:

  $v(t - w(t)) = w(t)$.

Flow enters service from left and leaves queue from right

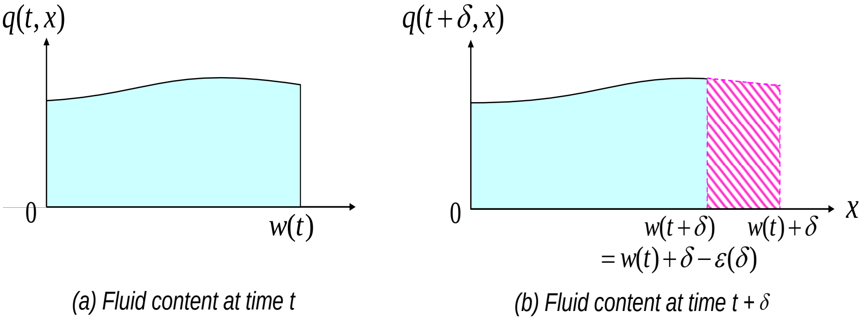


The ODE for the Boundary Waiting Time

$$w'(t) = 1 - \frac{b(t,0)}{q(t,w(t))}$$

- $q(t,w(t))$: density of fluid in queue the longest at t
- $b(t,0)$: rate into service at t
- $b(t,0) > (\leq)\, q(t,w(t)) \Rightarrow w'(t) < (\geq)\, 0$

More on the ODE for the Waiting Time



$$-q(t,w(t))(\Delta w(t) - \Delta t) = b(t,0)\Delta t$$

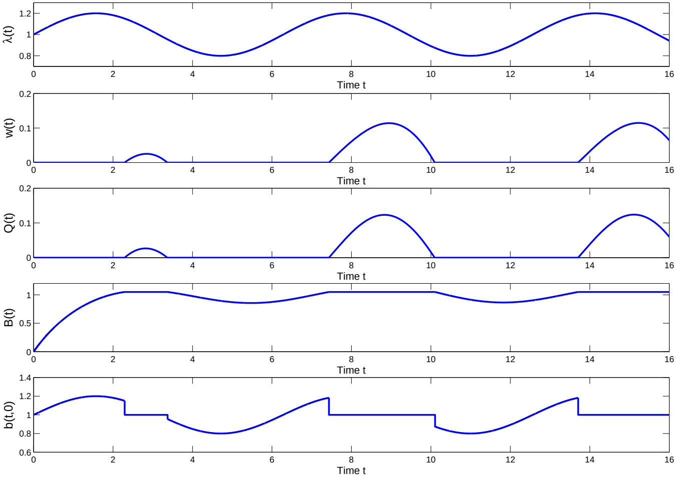
$$\frac{\Delta w(t)}{\Delta t} = 1 - \frac{b(t,0)}{q(t,w(t))} \quad \text{so that} \quad \dot{w}(t) = 1 - \frac{b(t,0)}{q(t,w(t))}.$$

OUTLINE

1. Staffing Examples
2. Approximations for Time-Varying Many-Server Queues
3. The $G_t/GI/s_t + GI$ Fluid Queue with Alternating Overloaded and Underloaded Intervals (Yunan Liu & W^2, *Queueing Systems*, 2012)
4. Numerical Examples: Comparisons with Simulations
5. Extensions
   1. Asymptotic Loss of Memory, Periodic Steady State
   2. Networks of Fluid Queues with Proportional Routing

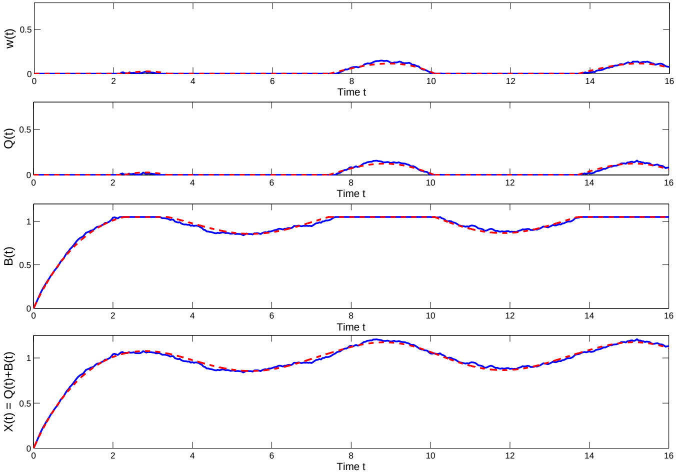
Example: $M_t/M/s+M$ Fluid Queue, $E[T_a]=2$

Arrival rate $\lambda(t) = 1 + 0.2 \cdot sin(t)$ and fixed staffing $s(t) = s = 1.05$



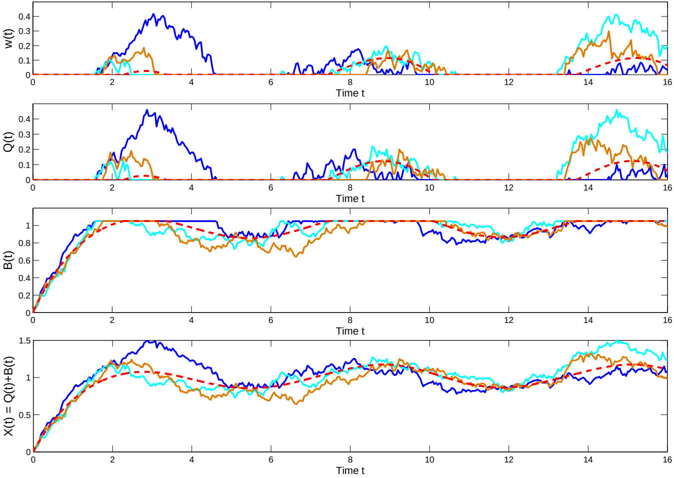
Comparison with Simulation of the $M_t/M/s+M$ Queue

$n = 2000$, single sample path ($\lambda(t) = 1000 + 200 \cdot \sin(t)$, $s = 1050$)



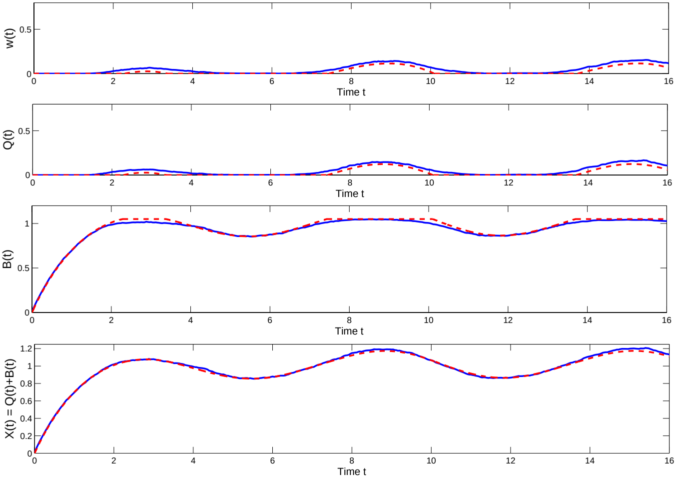
Comparison with Simulation: Smaller n

$n = 100$, 3 sample paths ($\lambda(t) = 100 + 20 \cdot sin(t)$, $s = 105$)



Comparison with Simulation: Approximate Mean Values

$n = 100$, average of 100 sample paths ($\lambda(t) = 100 + 20 \cdot sin(t)$, $s = 105$)



OUTLINE

1. Staffing Examples
2. Approximations for Time-Varying Many-Server Queues
3. The $G_t/GI/s_t + GI$ Fluid Queue with Alternating Overloaded and Underloaded Intervals (Yunan Liu & W^2, *Queueing Systems*, 2012)
4. Comparisons with Simulations
5. Extensions
    1. Asymptotic Loss of Memory, Periodic Steady State (YL&W^2 2011)
    2. Networks of Fluid Queues with Proportional Routing (YL&W^2 2011)

SUMMARY

1. Discussed approximations for time-varying many-server queues.
2. The time-varying $G_t/GI/s_t + GI$ fluid model is tractable and useful.
3. Analyzed for the case of alternating OL and UL intervals.
4. The algorithm involves: (i) a fixed-point equation for the fluid density in service, and (ii) an ODE for the boundary waiting time.
5. Extension for networks of fluid queues has been developed.
6. Asymptotic behavior as $t \to \infty$ studied (ALOM).
7. Stochastic refinements based on FCLT have been & are being developed.

THE END

References

Staffing to Stabilize Performance at Targets

- Review paper: **Coping with Time-Varying Demand when Setting Staffing Requirements for a Service System.** POMS, 2007. (with Linda V. Green and Peter J. Kolesar)
- **Server Staffing to Meet Time-Varying Demand.** Management Sci., 1996. (with Otis B. Jennings, A. Mandelbaum and W. A. Massey)
- **Staffing of Time-Varying Queues to Achieve Time-Stable Performance.** Management Sci., 2008. (with Zohar Feldman, Avishai Mandelbaum and William A. Massey)
- **Stabilizing Customer Abandonment in Many-Server Queues with Time-Varying Arrivals.** *Operations Res.*, 2012. (with Yunan Liu)

Fluid Model Papers with Yunan Liu

- **The $G_t/GI/s_t + GI$ Many-Server Fluid Queue.** QUESTA, 2012.
- **A Network of Time-Varying Many-Server Fluid Queues with Customer Abandonment.** *Operations Res.*, 2011.
- **Large-Time Asymptotics for the $G_t/M_t/s_t + GI_t$ Many-Server Fluid Queue with Abandonment.** *Queueing Systems* (QUESTA), 2011.
- **Nearly Periodic Behavior in the The Overloaded $G/D/S + GI$ Queue.** *Stochastic Systems*, 2011.
- **Algorithms for Networks of Fluid Queues.** IJoC, submitted.
- **Many-Server Heavy-Traffic Limits for Queues with Time-varying Parameters.** AnAP, submitted.

Background References on Fluid Approximations

- textbook: R. W. Hall. **Queueing Methods for Services and Manufacturing.** Prentice Hall, Englewood Cliffs, NJ, 1991.
- textbook: G. F. Newell, **Applications of Queueing Theory**, second edition, Chapman and Hall, 1982.
- tutorial: R. C. Hampshire and W. A. Massey **Dynamic Optimization with Applications to Dynamic Rate Queues.** Tutorials in OR, INFORMS 2010.
- accuracy: A. Bassamboo & R. S. Randhawa. **On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers.** Northwestern Univ., 2009. *Operations Res.*, forthcoming.

More References on Fluid Approximations

- $G/GI/s + GI$ fluid model: W^2, **Fluid models for multiserver queues with abandonments.** *Operations Res.*, 54 (2006) 37–54.
- R. C. Hampshire, O. B. Jennings and W. A. Massey **A Time Varying Call Center Design with Lagrangian Dynamics.** *PEIS*, 23 (2009) 231–259.
- application to delay announcements: M. Armony, N. Shimkin & W^2. **The Impact of Delay Announcements in Many-Server Queues with Abandonment.** *Operations Res.* 57 (2009) 66–81.
- R. Talreja & W^2, **Fluid Models for Overloaded Multi-Class Many-Server Queueing Systems with FCFS Routing.** *Man. Sci.*(2008)

Background References: MSHT Limits

- MSHT limits with time-varying arrival rates: A. Mandelbaum, W. A. Massey & M. I. Reiman. **Strong approximations for Markovian service networks.** Queueing Systems, 30 (1998) 149–201.
- survey: G. Pang, R. Talreja & W^2. **Martingale Proofs of Many-Server Heavy-Traffic Limits for Markovian Queues.** *Probability Surveys* 4 (2007) 193–267.
- MSHT limits for $G/GI/s$: H. Kaspi & K. Ramanan. **Law of large numbers limits for many-server queues.** & **SPDE limits of many-server queues**, AnAP, 2011.

MSHT Limits for Infinite-Server queues

- MSHT limits for $G/GI/\infty$: E. V. Krichagina & A. A. Puhalskii. **A heavy-traffic analysis of a closed queueing system with a GI/∞ service center.** Queueing Systems. 25 (1997) 235–280.
- MSHT limits for $G/GI/\infty$: G. Pang & W^2. **Two-parameter heavy-traffic limits for infinite-server queues.** Queueing Systems, 65 (2010) 325–364.
- MSHT limits for $G/GI/\infty$: J. Reed & R. Talreja. **Distribution-valued heavy-traffic limits for the $G/GI/\infty$ queue.** New York University, 2009.

Extra Slides

1. Many-Server Heavy-Traffic Limits

The Queueing Variables

- content processes: **two-parameter stochastic processes**
- $\mathbf{B_n(t, x)}$ number in service at time t who have been there for time $\leq x$,
- $\mathbf{Q_n(t, x)}$ number in queue at time t who have been there for time $\leq x$,
- $W_n(t)$ elapsed waiting time for customer at head of line (HOL),
- $V_n(t)$ potential waiting time for new arrival (virtual if infinitely patient),
- $A_n(t)$ number to abandon in $[0, t]$,
- $E_n(t)$ number to enter service in $[0, t]$,
- $S_n(t)$ number to complete service in $[0, t]$,
- Fluid scaling: $\bar{Y}_n \equiv n^{-1} Y_n$.

MSHT fluid limit (FWLLN)

Theorem

(FWLLN) If ..., then

$$(\bar{B}_n, \bar{Q}_n, W_n, V_n, \bar{A}_n, \bar{E}_n, \bar{S}_n) \Rightarrow (B, Q, w, v, A, E, S) \quad in \quad \mathbb{D}^2_{\mathbb{D}} \times \mathbb{D}^5,$$

as $n \to \infty$, where (B, Q, w, v, A, E, S) is deterministic, depending on the model data $(\lambda, s, G, F, B(0,\cdot), Q(0,\cdot))$, with

$$\begin{aligned} B(t,y) &\equiv \int_0^y b(t,x)\,dx, \quad Q(t,y) \equiv \int_0^y q(t,x)\,dx, \quad t \geq 0, y \geq 0, \\ A(t) &\equiv \int_0^t \alpha(u)\,du, \quad E(t) \equiv \int_0^t b(u,0)\,du, \quad S(t) \equiv \int_0^t \sigma(u)\,du. \end{aligned}$$

The Three MSHT Limiting Regimes for Stationary Models

Let the traffic intensity be $\rho_n \equiv \lambda_n / s_n \mu_n = \lambda_n / s_n$.

- Quality-and-Efficiency-Driven (QED) regime (critically loaded):

$$(1 - \rho_n)\sqrt{n} \to \beta \quad \text{as} \quad n \to \infty, \quad -\infty < \beta < \infty.$$

- Quality-Driven (QD) regime (underloaded): $(1 - \rho_n)\sqrt{n} \to \infty$.

- Efficiency-Driven (ED) regime (overloaded): $(1 - \rho_n)\sqrt{n} \to -\infty$.

In **fluid scale:** QED: $\rho = 1$, QD: $\rho < 1$ and ED: $\rho > 1$.

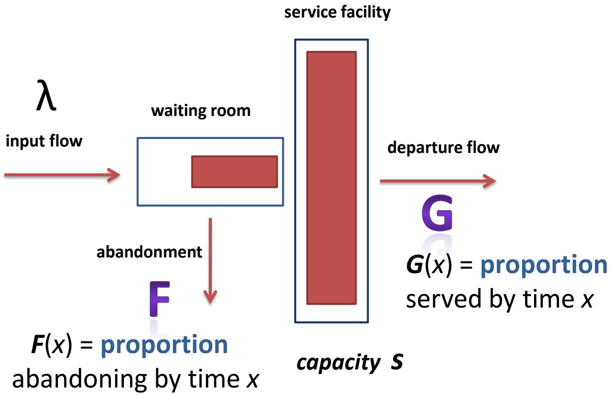
Separation of Time Scales

The MSHT limit causes a separation of time scales:

- The relevant time scale is the **mean service time**, which is fixed.
- Since the arrival rate grows, i.e., since $\lambda_n(t)/n \to \lambda(t)$ as $n \to \infty$, the arrival process matters in a long time scale, through its LLN and CLT.
- The service-time cdf G and patience cdf F matter.

2. The Stationary $G/GI/s + GI$ Fluid Model

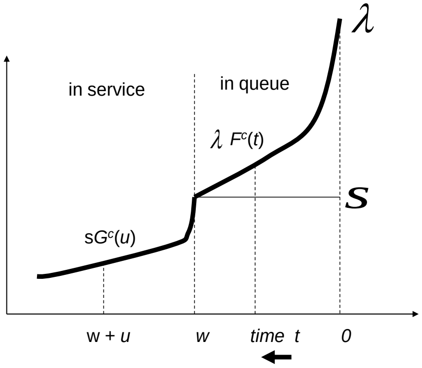
Model data: $(\lambda(t), s(t), G, F)$ and initial conditions.



The Overloaded Fluid Model in Steady State

The 2005 MIT talk.

fluid density arriving time *t* in the past



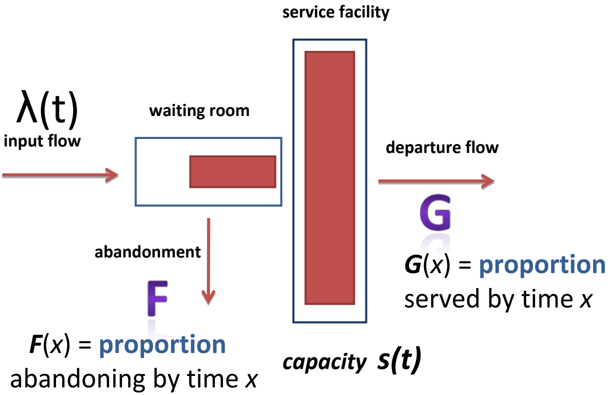
Simulations for the $M/E_2/24+GI$ Model: $\lambda = 24$

Two abandonment cdf's: Erlang E_2 and lognormal $LN(1,4)$, mean 1.

perf. meas.	E_2 sim	E_2 approx	$LN(1,4)$ sim	$LN(1,4)$ approx
$P(A)$	0.175 $\pm.0003$	0.167	0.191 $\pm.0002$	0.167
$E[Q]$	7.7 $\pm.013$	8.2	3.15 $\pm.004$	2.93
$SCV[Q]$	0.43	0.00	0.97	0.00
$E[W\|S]$	0.322 $\pm.001$	0.365	0.129 $\pm.0002$	0.131

3. The Time-Varying $G_t/GI/s_t + GI$ Fluid Model

Model data: $(\lambda(t), s(t), G(x), F(x))$ and initial conditions.



The Fluid Density in an Underloaded Interval

explicit expression:

$$\begin{aligned} b(t,x) &= \text{new content } 1_{\{x \le t\}} + \text{old content } 1_{\{x > t\}} \\ &= \bar{G}(x)\lambda(t-x)1_{\{x \le t\}} + b(0,x-t)\frac{\bar{G}(x)}{\bar{G}(x-t)}1_{\{x>t\}}. \end{aligned}$$

transport PDE:

$$b_t(t,x) + b_x(t,x) = -h_G(x)b(t,x)$$

with boundary conditions $b(t,0) = \lambda(t)$ and initial values $b(0,x)$.

The service-content density $b(t,x)$

- During an underloaded interval,

$$b(t,x) = \bar{G}(x)\lambda(t-x)1_{\{x \le t\}} + \frac{\bar{G}(x)}{\bar{G}(x-t)} b(0,x-t)1_{\{x>t\}}.$$

- During an overloaded interval,

$$b(t,x) = \mathbf{b(t-x,0)}\bar{G}(x)1_{\{x \le t\}} + + \frac{\bar{G}(x)}{\bar{G}(x-t)} b(0,x-t)1_{\{x>t\}}.$$

(i) With *M* service, $\sigma(t) = B(t) = s(t)$, $b(t,0) = s'(t) + s(t)$.

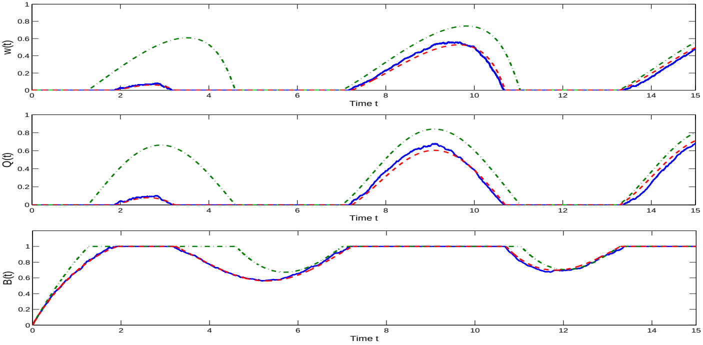
(ii) With *GI* service, $b(t,0)$ satisfies the **fixed-point equation**

$$\mathbf{b(t,0)} = a(t) + \int_0^t \mathbf{b(t-x,0)} g(x)\, dx,$$

$$\text{where} \quad a(t) \equiv s'(t) + \int_0^\infty b(0,y) g(t+y)/G(y)\, dy.$$

Non-Exponential Distributions Matter

Simulation comparison for the $M_t/GI/s+E_2$ fluid model: (i) **H_2** service (red dashed lines), (ii) **M** service (green dashed lines), (iii) sample path from **simulation** of queue with **H_2** service based on $n = 2000$ (blue solid lines).



Comparison with Simulation: Even Smaller n

$n = 20$, average of 100 sample paths ($\lambda(t) = 20 + 4 \cdot sin(t)$, $s = 21$)