Multi-Server Queues with Time-Varying Arrival Rates

Ward Whitt

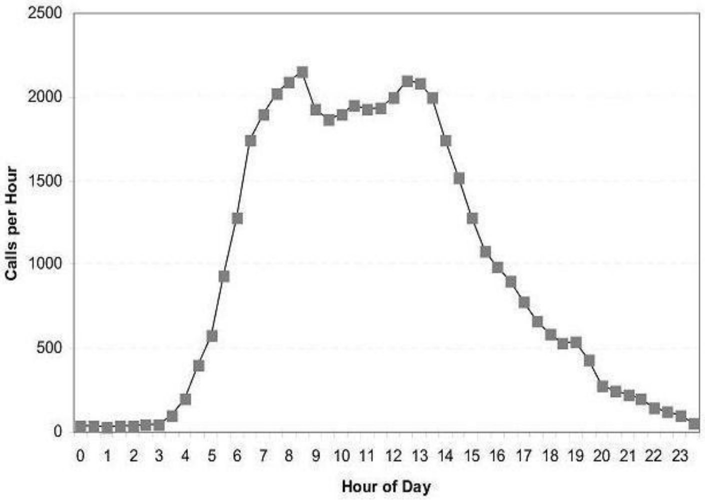
(joint work with doctoral student Yunan Liu)

Columbia University, http://www.columbia.edu/∼ww2040

INFORMS Markov Lecture, November 8, 2010

Time-Varying Arrival Rates

arrivals per hour to a medium-sized financial-services call center



The Model

$$M_t/GI/s_t + GI$$

- time-varying arrival rate $\lambda(t)$
- a large time-varying number of servers $s(t)$
- customer abandonment (the $+GI$)
- non-exponential distributions

  service-time cdf $G(x) = P(S \leq x)$, patience-time cdf $F(x) = P(A \leq x)$
- unlimited waiting room and the FCFS service discipline

(model parameters in red)

This talk has two parts.

Part 1. offered-load approximations

Recent related work has been done by discussant Galit Yom-Tov, jointly with her thesis advisor Avishai Mandelbaum.

Part 2. deterministic fluid models

Recent related work has been done by discussant Bill Massey, jointly with Robert Hampshire. See their tutorial on Tuesday.

The discussants are collaborators; e.g., see Feldman, Z., Mandelbaum, A., Massey, W. A. & WW, **Staffing of Time-Varying Queues to Achieve Time-Stable Performance**, *Management Science* 54 (2008) 324-338.

One Unifying Idea:

Exploit Associated

Infinite-Server (IS) Models

$$M_t/GI/\infty$$

Offered-Load Approximations To Set Staffing Levels

For capacity planning, specify capacity by seeing how much would be used if there were an unlimited supply, allowing for uncertainty.

The first Idea: Offered-Load (OL) Approximations

- How many servers would be used if there were an unlimited supply?
- For $M_t/GI/s_t + GI$, look at the associated $M_t/GI/\infty$ model.
- Let $X(t)$ be the number of busy servers at time t in $M_t/GI/\infty$.
- $X(t) \stackrel{\mathrm{d}}{=} \text{Poisson}(m(t)) \approx \text{Normal}(m(t), m(t))$,

where the expected number $m(t) \equiv E[X(t)]$ is called "the" offered load

and can be expressed as

$$\begin{aligned} \mathbf{m(t)} &\equiv E[X(t)] = \int_{-\infty}^{t} \lambda(u) P(S > t - u)\, du = E[\lambda(t - S_e)]E[S], \\ P(S_e \leq x) &\equiv (1/E[S]) \int_0^x P(S > u)\, du, \quad x \geq 0. \end{aligned}$$

For more on $m(t)$, see Eick, Massey & WW (1993a,b).

Implication: Square Root Staffing (SRS)

- If $X(t) \approx \mathrm{Normal}(m(t), m(t))$, then

$$\begin{aligned} P(W(t) > 0) &= P(X(t) \geq s(t)) \approx P\left(N(0,1) \geq \frac{s(t) - m(t)}{\sqrt{m(t)}}\right) \\ &= P(N(0,1) \geq \beta) \equiv 1 - \Phi(\beta). \end{aligned}$$

- Hence, the OL approximation supports the

  Square Root Staffing (SRS) formula: Given the target $\tau \equiv P(W > 0)$,

  choose β such that $1 - \Phi(\beta) = \tau$. Then let

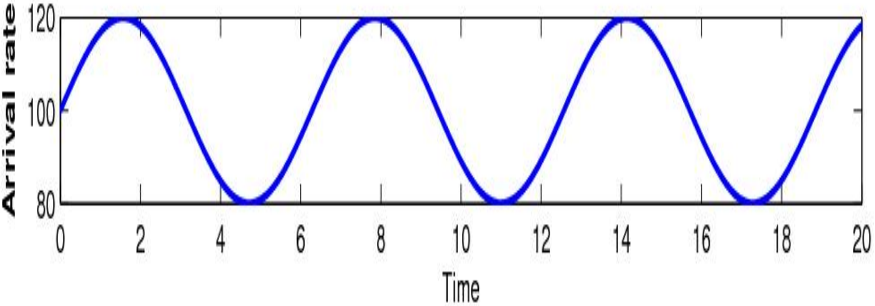
$$s(t) = m(t) + \beta\sqrt{m(t)}.$$

- The SRS is also supported by MSHT limits.

A Markovian Example (used throughout the talk)

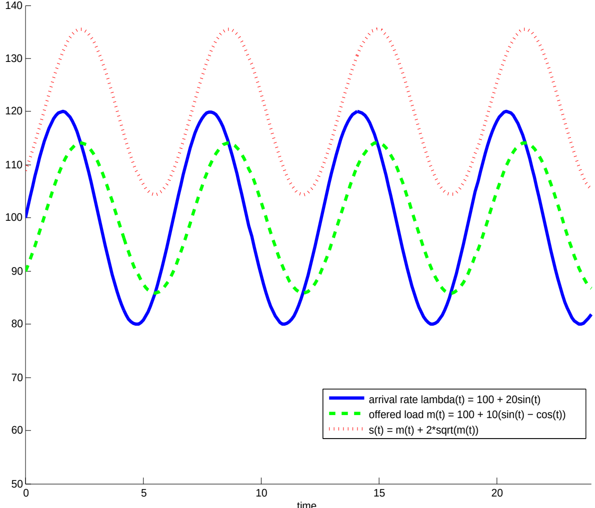
$M_t/M/s_t + M$ with sinusoidal arrival rate

- $\lambda(t) = 100 + 20 \cdot \sin(t)$
- $\bar{G}(x) = e^{-\mu x}$, $\mu = 1$; $\bar{F}(x) = e^{-\theta x}$, $\theta = 0.5$



The Example

From arrival rate to offered load to staffing



Insights

- In the stationary setting, OL is one-dimensional: $m = \lambda E[S]$.
- In the time-varying setting, OL is two-dimensional: $m(t)$.

  New methods needed when $m(t) \not\approx \lambda(t)E[S]$.
- new OL when the required service is more complicated:

  (i) OL may depend on location too, e.g. networks of queues, mobile phones; Massey & WW (1993, 1994), Leung, M & WW (1994).

  (ii) There may be time-varying service requirements; e.g, (a) disjoint intervals, as in web chat or patient contact with physicians; (b) required bandwidth for user fluctuates over time; Duffield, Massey & WW (2002).

The Second Idea: The MOL Approximation

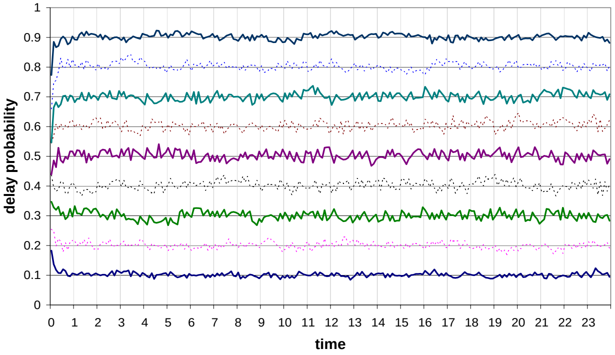
- However, the normal approximation for $P(W > 0)$ is somewhat crude, because it does not account for the actual limited number of servers.
- A better approximation can be obtained exploiting the corresponding stationary model in an appropriate time-dependent manner.

  For $M_t/GI/s_t + GI$, we look at the associated $M/GI/s + GI$ model.
- We approximate $X(t)$ in $M_t/GI/s_t + GI$ by the steady-state number $X(\infty)$ in $M/GI/s + GI$, but where $s = s(t)$ and the approximating fixed MOL arrival rate is chosen to be

$$\lambda = \lambda_{MOL}(t) \equiv \frac{m(t)}{E[S]}.$$

MOL Stabilizes Delay Probability

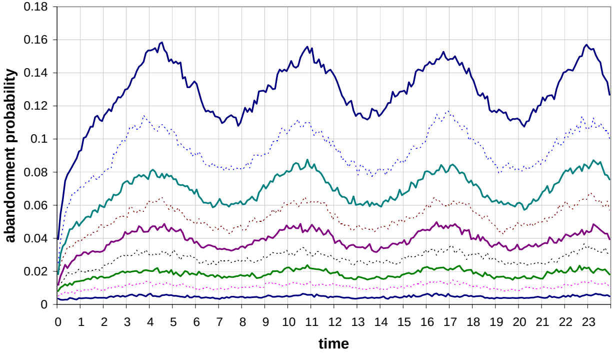
Example: $M_t/M/100+M$ model with sinusoidal arrival rate.



Plots of delay probabilities; Figure 3 from Feldman et al. (2008).

But does not always stabilize the Abandonment Probability

Same model with sinusoidal arrival rate.



Plots of abandonment probabilities; Figure 4 from Feldman et al. (2008).

A new two-step

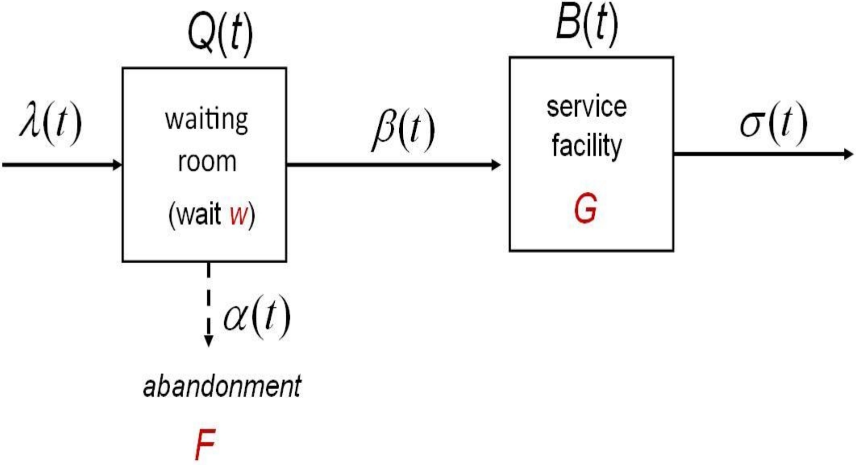
offered-load approximation:

Step 1. Approximate the $M_t/GI/s_t + GI$ system by two $M_t/GI/\infty$ queues in series: Customers wait EXACTLY w if they do not abandon.

Step 2. Create a new MOL approximation.

Two $M_t/GI/\infty$ Models in Series

Decoupling



New Modified Offered Load Approximation

- Use the new OL $m(t) \equiv E[B(t)]$ to define an MOL arrival rate

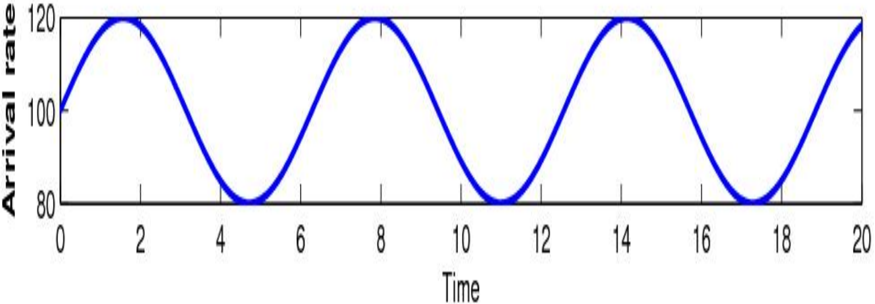
$$\lambda_{MOL}(t) \equiv \frac{m(t)}{(1-\alpha)E[S]}.$$

- Use WW (2005) to approximate the steady-state $P_\infty(Ab)$ for the $M/GI/s+GI$ model (based on approximation by $M/M/s+M(n)$).
- For any s, approximate $P_t(Ab)$ by $P_\infty(Ab)$ in the associated $M/GI/s+GI$ model using arrival rate $\lambda_{MOL}(t)$.
- Given target α, let $s_{MOL}(t) \equiv \min\{s : P_t(Ab; s) \leq \alpha\}$.

The Same Markovian Example

$M_t/M/s_t + M$ with sinusoidal arrival rate

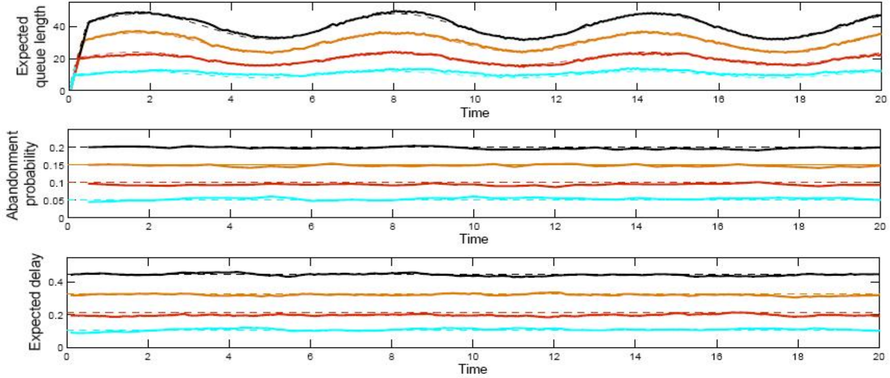
- $\lambda(t) = 100 + 20 \cdot \sin(t)$
- $\bar{G}(x) = e^{-\mu x}$, $\mu = 1$; $\bar{F}(x) = e^{-\theta x}$, $\theta = 0.5$



Validation with Simulation

Heavy load: Range of targets: $5\% \leq \alpha \leq 20\%$

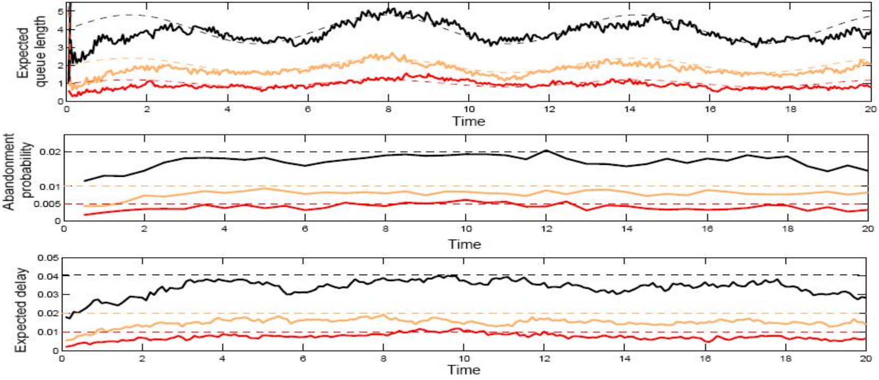
$s_{MOL}(t) \approx m(t)$: OL works without refinement.



Validation with Simulation

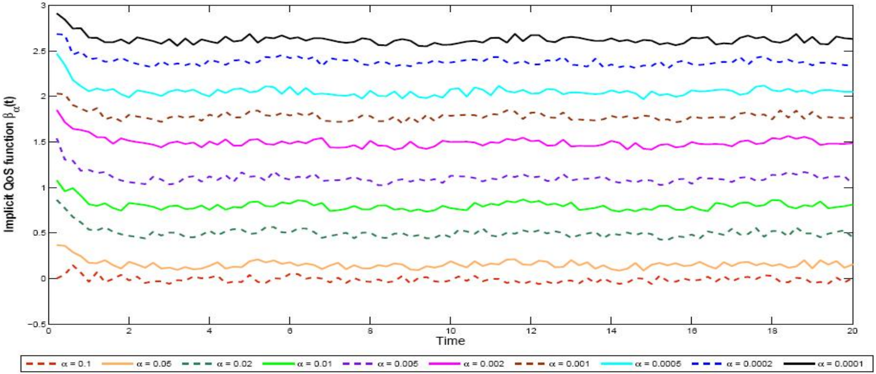
Light load: Range of targets: $0.5\% \leq \alpha \leq 2\%$

$s_{MOL}(t) > m(t)$: MOL refinement needed.



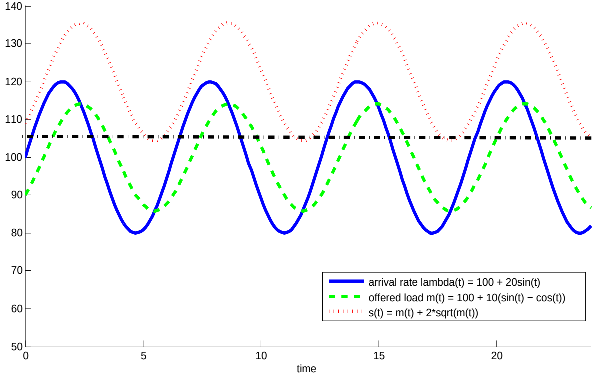
The Empirical Quality of Service

- $\beta_\alpha(t) \equiv (s^{MOL}(t) - m_\alpha(t))/\sqrt{m_\alpha(t)}$



Summary . . . and Transition to the Second Topic

From arrival rate to offered load to staffing . . . if staffing is flexible



Deterministic Fluid Approximation

for alternating

overloaded intervals and underloaded intervals

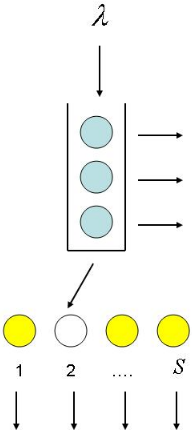
MSHT Limit

- a sequence of $G_t/GI/s_t + GI$ models indexed by n,
- arrival rate grows: $\lambda_n(t)/n \to \lambda(t) \quad \text{as} \quad n \to \infty$,

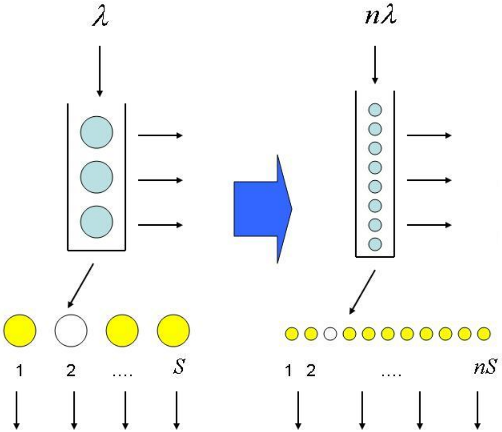
  number of servers grows: $s_n(t) \equiv \lceil ns(t) \rceil$,
- service-time cdf G and patience cdf F held fixed independent of n

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

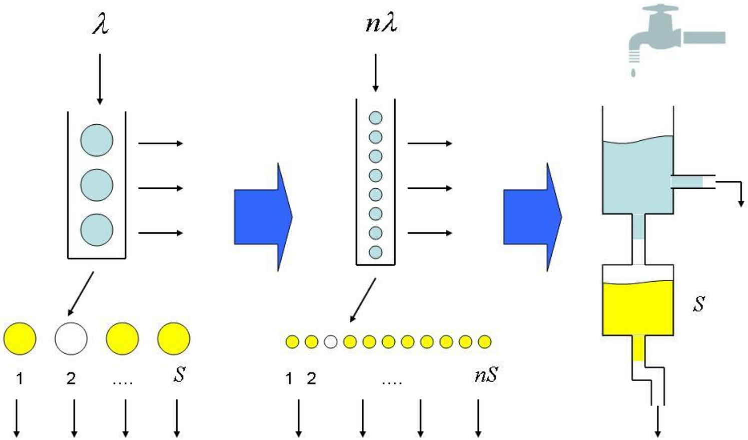
Fluid Approximation from MSHT limit



Fluid Approximation from MSHT limit



Fluid Approximation from MSHT limit



The Three MSHT Limiting Regimes for Stationary Models

Let $\lambda_n(t) = \lambda_n$ and $s_n(t) = s_n$, both constant (not time-varying).

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): $(1 - \rho_n)\sqrt{n} \to -\infty$.

A new MSHT Regime for Time-Varying Arrivals

- Instead of the QED regime,

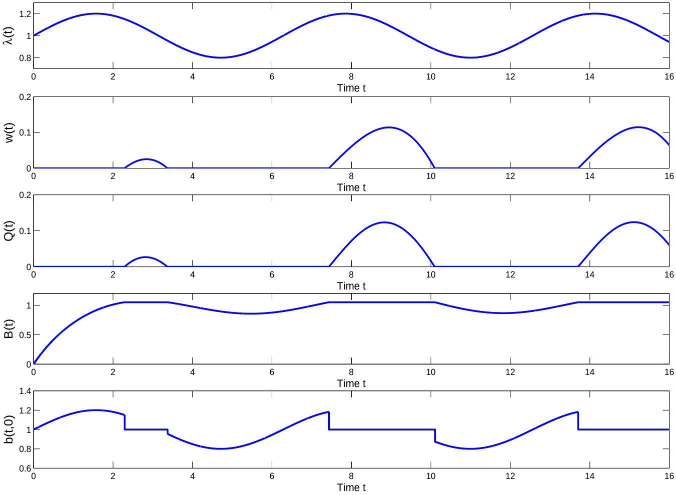
  we focus on the complement $(QED)^c = ED + QD$.

- Switching between overloaded intervals and underloaded intervals

Like the example with $\lambda(t) = 100 + 20 \cdot sin(t)$ and $s(t) = s = 105$.

Fluid Approximation for the Example

Arrival rate $\lambda(t) = 100 + 20 \cdot sin(t)$ and fixed staffing $s(t) = s = 105$



The Queueing Variables

- $B_n(t,x)$ number in service at time t who have been there for time $\leq x$,
- $Q_n(t,x)$ 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,
- $V_n(t)$ potential waiting time for new arrival (virtual w infinite patience),
- $A_n(t)$ 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 limit for alternating OL and UL intervals

Theorem

(FWLLN) If … (including regularity for the fluid model: feasible staffing, smooth model, finitely many switches between OL and UL), then

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

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

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

The Idea of the Proof

- Recursively treat successive UL and OL intervals.
- IS MSHT limits (Pang&WW10) apply directly to treat UL intervals.
- In OL intervals first ignore flow into service; let $\tilde{Q}_n(t, y)$ be the process.
- IS MSHT limits (P&WW10) apply to treat $\tilde{Q}_n$ in OL intervals.
- To go from $\tilde{Q}_n$ to Q_n, focus on HOL waiting time W_n:

  Equate two representations of the flow into service during OL interval:

  (i) new space available due to service completion and capacity change

  (ii) the flow into service from the queue, which occurs from the head of the line.

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

two-parameter functions

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 $(\lambda, s', g, f, b(0,\cdot), q(0,\cdot))$.

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 $B(t) = S(t)$)

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

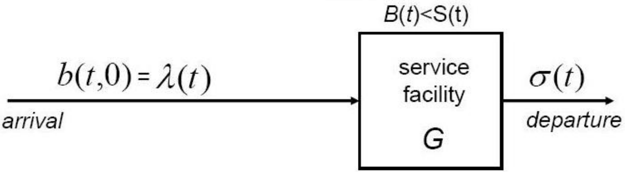
Given $q(t,x)$ and $b(t,x)$,

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

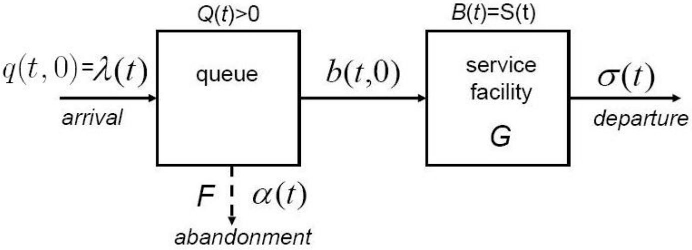
where $h_F(x) \equiv \frac{f(x)}{\bar{F}(x)}$ and $h_G(x) \equiv \frac{g(x)}{\bar{G}(x)}$

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

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

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

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

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

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

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 $b(0,x)$.

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:

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

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,

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

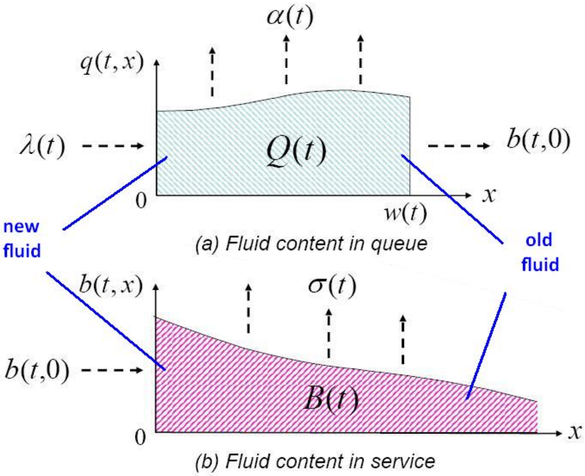
(i) With *M* service, $\sigma(t) = B(t) = s(t)$, $b(t,0) = s'(t) + s(t)$.

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

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

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

Flow enters service from left and leaves queue from right



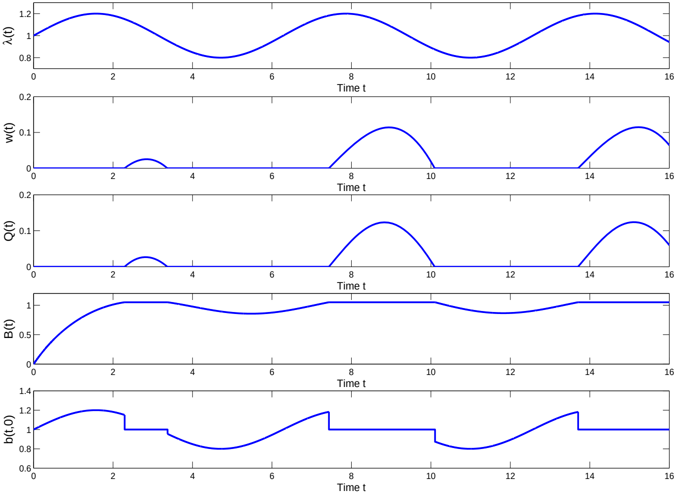
The ODE for the Boundary Waiting Time

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

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

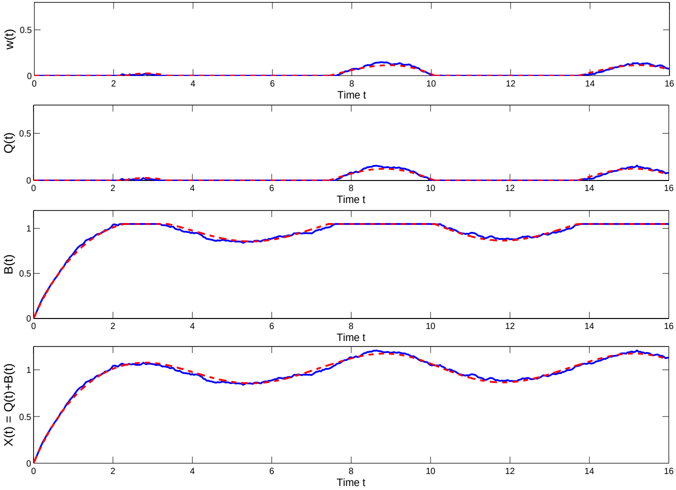
Fluid Approximation for the Example

Arrival rate $\lambda(t) = 100 + 20 \cdot sin(t)$ and fixed staffing $s(t) = s = 105$



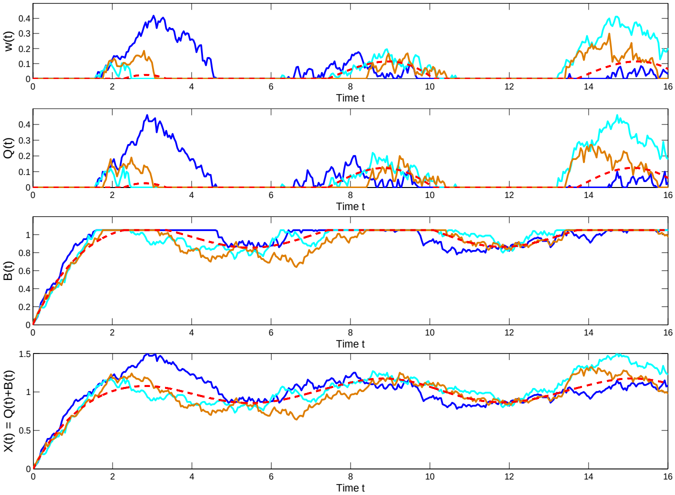
Comparison with Simulation

$n = 2000$ and a single sample path



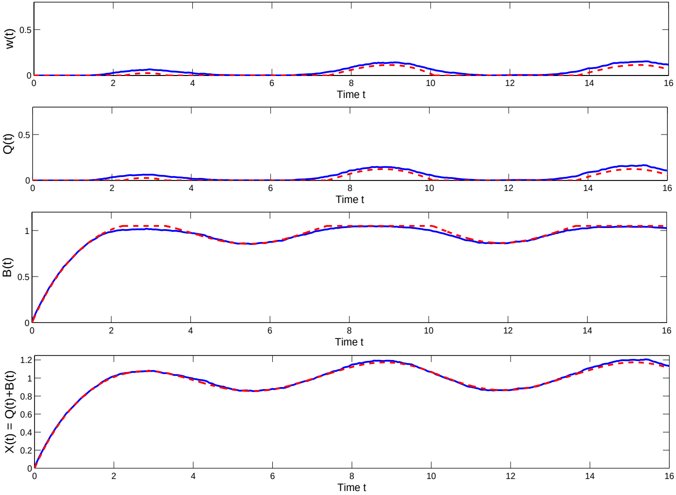
Comparison with Simulation

$n = 100$ and 3 sample paths



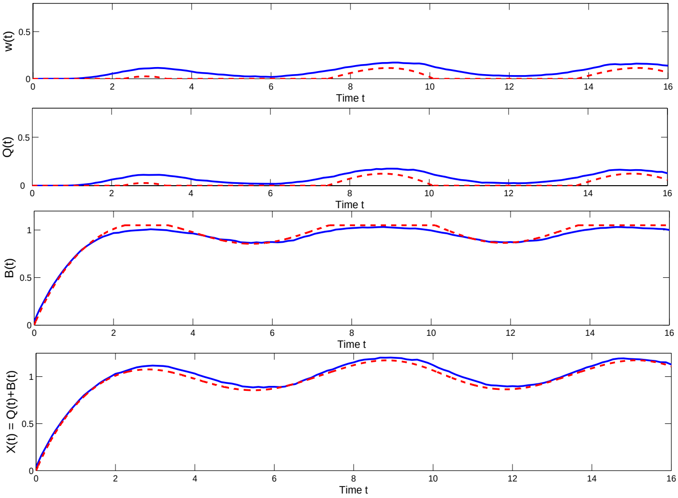
Comparison with Simulation

$n = 100$ and average of 100 sample paths



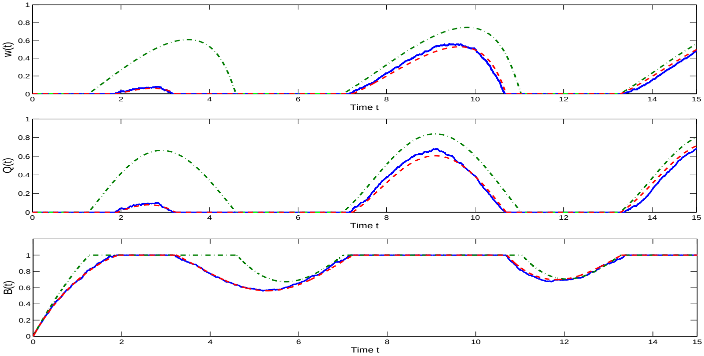
Comparison with Simulation

$n = 20$ and average of 100 sample paths



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 paths in the scaled queueing model with H_2 service based on $n = 2000$ (blue solid lines).



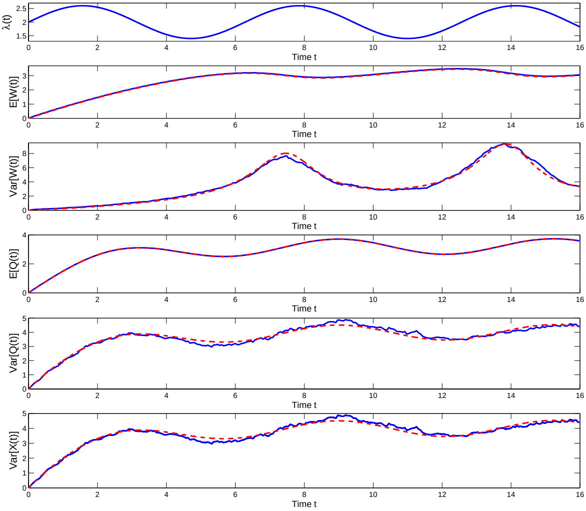
Work in Progress: FCLT and Stochastic Refinements

For smaller n, such as $n = 20$, the queueing stochastic processes experience significant fluctuations. Thus, for smaller n, we need to approximate the full distributions of the stochastic processes. That can be based on a FCLT refinement of the FWLLN plus engineering refinements. Work is underway on that.

Example: Gaussian approximation for an OL Interval

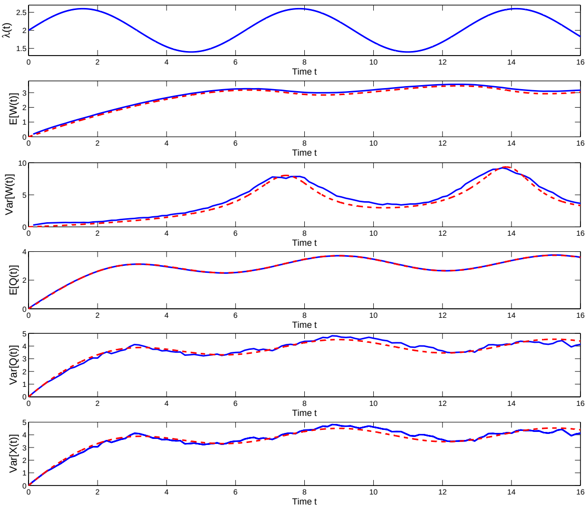
- the model: $M_t/M/s_t + M$
- $\lambda(t) = 2.0 + 6 \cdot sin(t)$, $s(t) = s = 0.4$, $\mu = 1$, $\theta = 0.5$
- initially critically loaded, $X(0) = s$
- queueing model has $n = 100$
- estimates based on 1000 replications

Comparisons with Simulation for $n = 100$



Averages of multiple (1000) sample paths

Comparisons with Simulation for $n = 25$



Averages of multiple (1000) sample paths

Summary: Two New Research Contributions

1. Developed a new **modified-offered-load** (MOL) **approximation** to **stabilize the abandonment probability**, $P_t(Ab)$, the probability that an arrival at time t eventually abandons, at any target level.

2. Developed a **deterministic fluid model** for $G_t/GI/s_t + GI$ model when the system **alternates** between **overloaded intervals** and **underloaded intervals**.

- Developed algorithm to compute all performance functions.
- Established a supporting many-server heavy-traffic (MSHT) limit.
- Developing refined stochastic approximations.

Conclusion

New effective ways to analyze and control the performance of multi-server queues with time-varying arrival rates, customer abandonment and non-exponential distributions.

Thank You!

Completed Papers by Yunan Liu and WW

- Here: **Stabilizing customer abandonment in many-server queues with time-varying arrivals**, 2009.
- Here: **The $G_t/GI/s_t + GI$ many-server fluid queue**, 2010.
- A Network of time-varying many-server fluid queues with customer abandonment, 2010. Operations Research, forthcoming.
- Large-time asymptotics for the $G_t/M_t/s_t + GI_t$ many-server fluid queue with abandonment, 2010.
- The heavily loaded many-server queue with abandonment and deterministic service times, 2010.
- All available at: www.columbia.edu/∼ww2040/allpapers.html

Background References: Offered-Load Approximations

- broad survey: L. V. Green, P. J. Kolesar & WW. **Coping with time-varying demand when setting staffing requirements for a service system.** Production and Opns. Management, 16 (2007) 13-39.
- stabilizing performance: Z. Feldman, A. Mandelbaum, W. A., Massey & WW. **Staffing of time-varying queues to achieve time-stable performance.** Management Science, 54 (2008) 324-338.
- healthcare applications: G. Yom-Tov & A. Mandelbaum. **The Erlang-*R* queue: time-varying QED queues with re-entrant customers in support of healthcare staffing.** the Technion, Israel, 2010.

Background References: Fluid Approximations

- textbook: R. W. Hall. **Queueing Methods for Services and Manufacturing.** Prentice Hall, Englewood Cliffs, NJ, 1991.
- $G/GI/s + GI$ fluid model: WW. **Fluid models for multiserver queues with abandonments.** Operations Research, 54 (2006) 37–54.
- accuracy: A. Bassamboo & R. S. Randhawa. **On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers.** Northwestern University, 2009.

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.
- MSHT for waiting times too: A. Mandelbaum, W. A. Massey, M. I. Reiman & A. Stolyar. **Waiting time asymptotics for time varying multiserver queues with abandonment and retrials.** Proceedings 37th Allerton Conference, (1999) 1095–1104.
- MSHT limits for $G/GI/s$: H. Kaspi & K. Ramanan. **Law of large numbers limits for many-server queues.** Carnegie Mellon University, 2007.

Background References: MSHT Limits for IS 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 & WW. **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.