Fluid Models for Large-Scale Service Systems

Experiencing Periods of Overloading

IEOR 8100, PhD Seminar on Queueing Theory, Professor Whitt

January 25, 2021

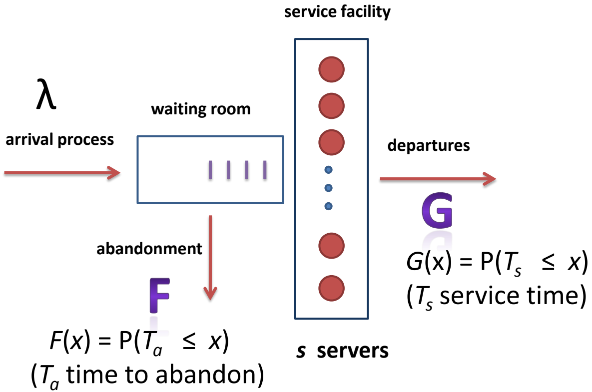
OUTLINE

1. The Overloaded $G/GI/s + GI$ Fluid Queue Model. (stationary)
   (W^2, *Operations Research*, 2006)
2. The $G_t/GI/s_t + GI$ Fluid Model with Alternating Overloaded and Underloaded Intervals. (Yunan Liu & W^2, *Queueing Systems*, 2012)
   1. Numerical Examples: Comparisons to Simulations of Stochastic Queueing Systems.
   2. An Algorithm to Compute the Performance Functions of the Fluid Model

(Relates to time-varying Little's law, offered load analysis and the infinite-server model and has extensions to networks.)

I. The Overloaded Stationary G/GI/s+GI Fluid Model

Approximation for the **G/GI/s + GI** **Stochastic Queueing Model** when overloaded



Increasing Scale

- a sequence of $G/GI/s + GI$ models indexed by n,
- arrival rate **grows**: $\lambda_n/n \to \lambda$ as $n \to \infty$,

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

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

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):

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

- 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

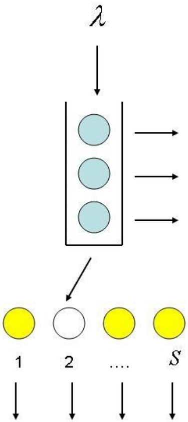
The MSHT limit causes a separation of time scales:

System View versus Customer View

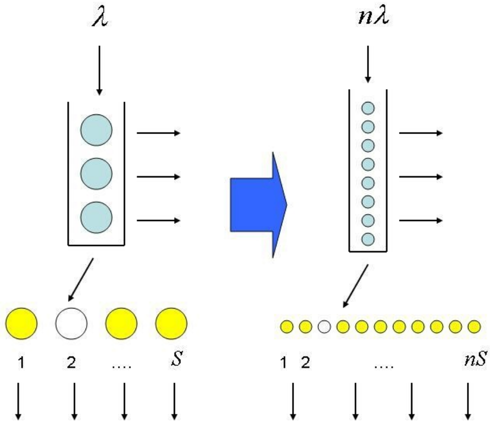
- The relevant time scale is the **mean service time**, which is fixed.
- Since the arrival rate grows, i.e., since $\lambda_n/n \to \lambda$ as $n \to \infty$,

  the arrival process matters in a long time scale, through its LLN and CLT.
- The service-time cdf G and patience cdf F matter.

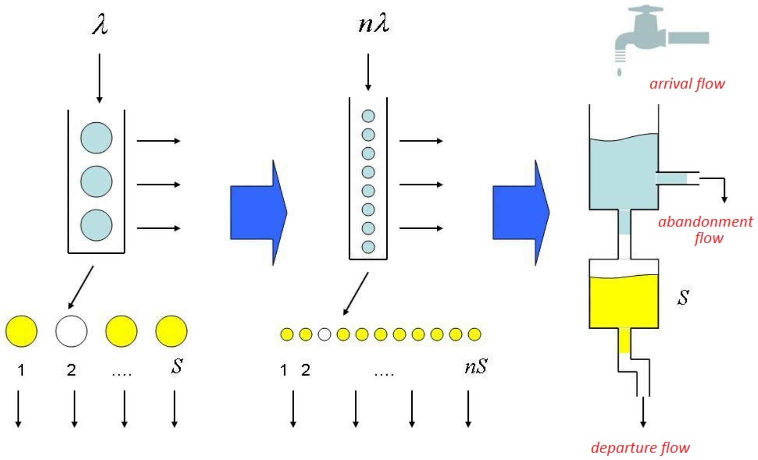
Fluid Approximation from MSHT limit



Fluid Approximation from MSHT limit



Fluid Approximation from MSHT limit



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: $\mathbf{\bar{Y}_n \equiv n^{-1} Y_n}$.

MSHT fluid limit (FWLLN)

Theorem

(FWLLN) If ..., then

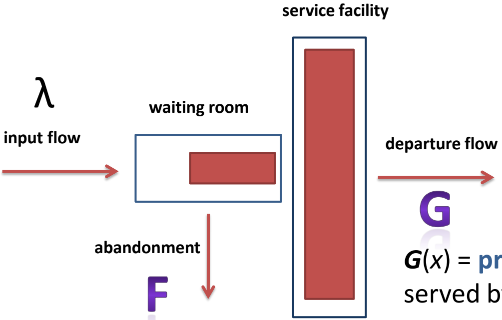
$$(\bar{B}_n, \bar{Q}_n, W_n, V_n, \bar{A}_n, \bar{E}_n, \bar{S}_n) \Rightarrow (B, Q, w, v, A, E, S) \quad \textit{in} \quad \mathbb{D}_{\mathbb{D}}^2 \times \mathbb{D}^5,$$

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 G/GI/s+GI Fluid Model

Model data: (λ, s, G, F) and initial conditions.

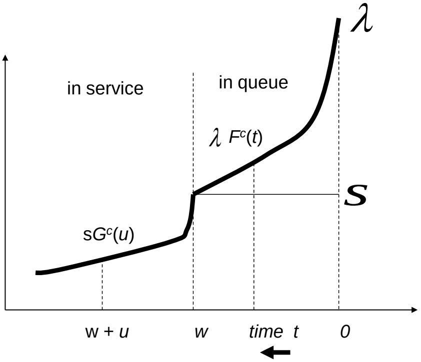


$G(x)$ = proportion served by time x

$F(x)$ = proportion abandoning by time x

The Overloaded Fluid Model in Steady State

fluid density arriving time *t* in the past



Simulations for the $M/E_2/20 + GI$ 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

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

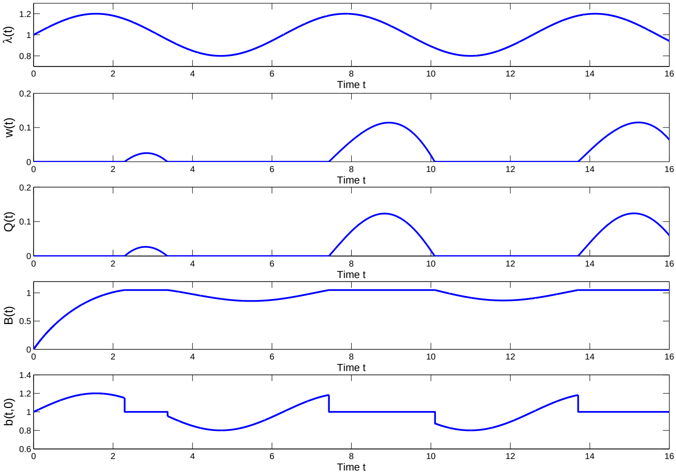
1. Numerical Examples: Comparison with Simulations of Queueing Models
2. An Algorithm to Compute All the Performance Functions of the Fluid Model

II.1. Numerical Examples

- Comparing the Algorithm for the Fluid Model to Simulations of the Stochastic Queueing Models
- Alternating Overloaded and Underloaded Intervals

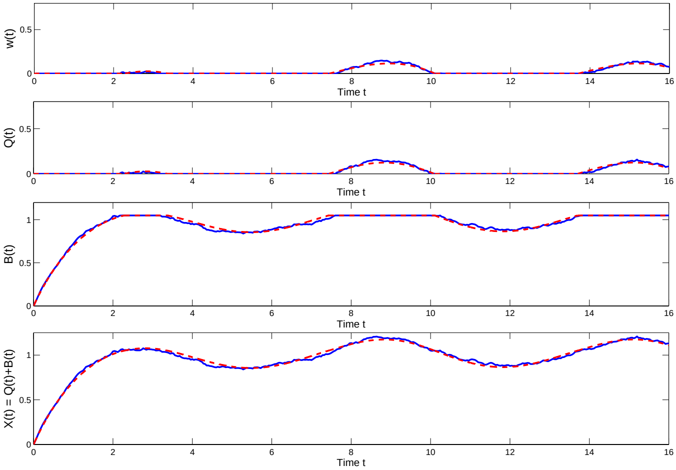
Example: $M_t/M/s + M$ Fluid Queue, $E[T_a] = 2$

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



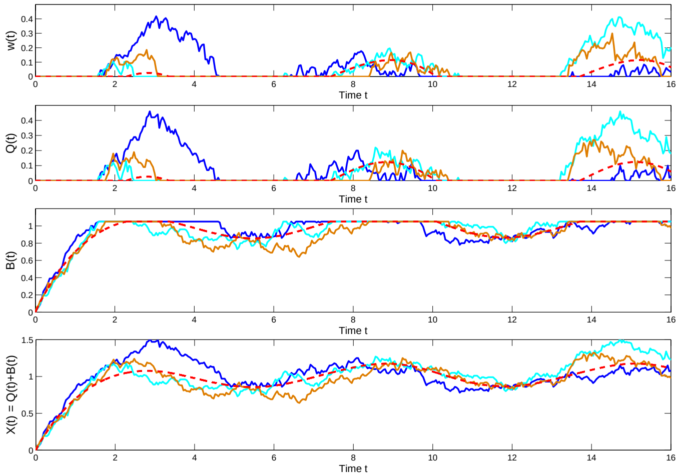
Comparison with Simulation of the $M_t/M/s+M$ Queue

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



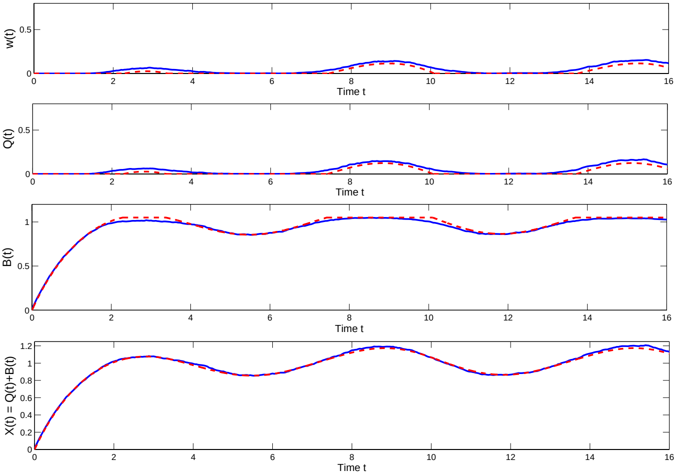
Comparison with Simulation: Smaller n

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



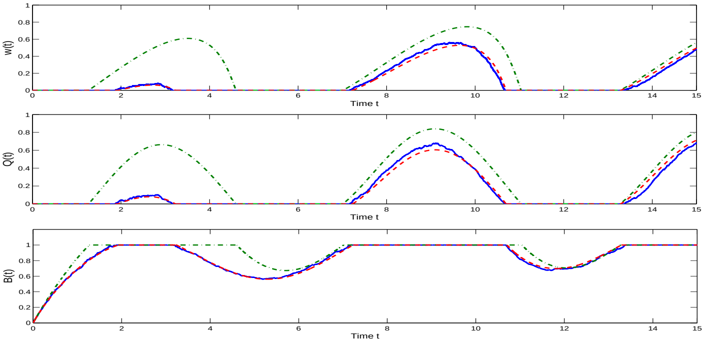
Comparison with Simulation: Approximate Mean Values

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



Non-Exponential Distributions Matter

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



II.2. The Time-Varying $G_t/GI/s_t + GI$ Fluid Model

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

- input rate $\lambda(t)$, time-varying
- service capacity $s(t)$, time-varying
- **feasible staffing** $s(t)$, (fluid not pushed out of service)
- same model data: $(\lambda(t), s(t), G, F)$ plus initial conditions
- **alternating overloaded (OL) and underloaded (UL) intervals**

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

Key Fluid Dynamics

Fundamental Evolution Equations

- $q(t+u, x+u) = q(t,x) \cdot \frac{\bar{F}(x+u)}{\bar{F}(x)}$, provided fluid does not enter service $0 \leq x \leq w(t) - u$, $u \geq 0$, $t \geq 0$.

- $b(t+u, x+u) = b(t,x) \cdot \frac{\bar{G}(x+u)}{\bar{G}(x)}$, $x \geq 0$, $u \geq 0$, $t \geq 0$.

Flow Rates

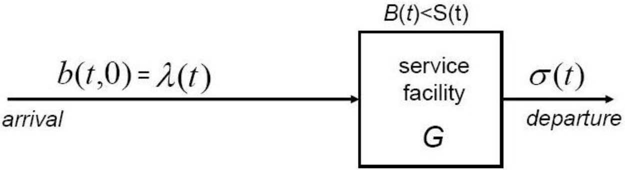
Given $q(t,x)$ **and** $b(t,x)$**,**

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

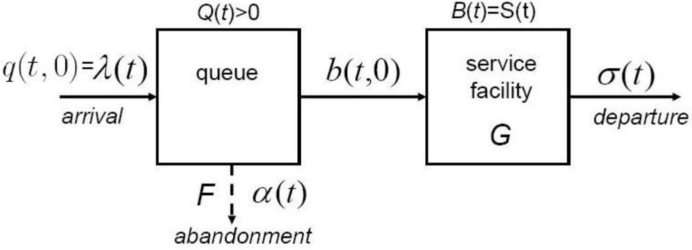
where $h_F(x) \equiv \frac{f(x)}{\bar{F}(x)}$ and $h_G(x) \equiv \frac{g(x)}{\bar{G}(x)}$

- $q(t,x)$ 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

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

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

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

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

We have formulas already (the "Physics" paper, Eick, Massey & W^2, 1993).

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 $b(0,x)$.

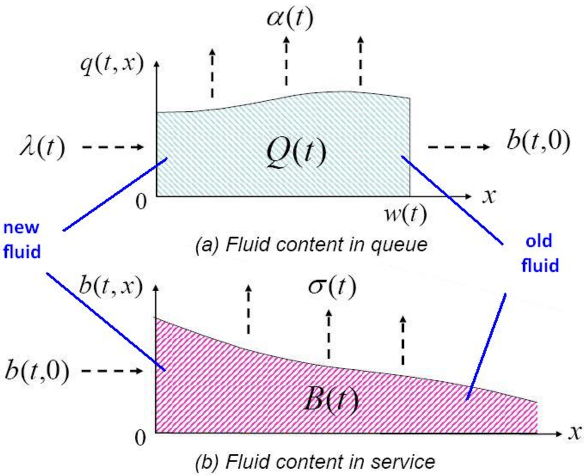
Second (Interesting) Case: Overloaded Interval

- 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:

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

Flow enters service from left and leaves queue from right



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

(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**

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

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

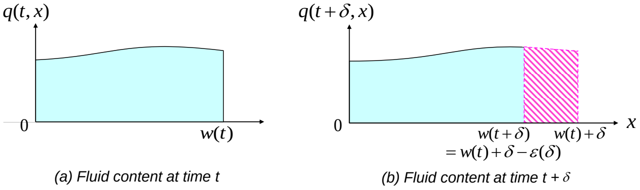
The ODE for the Boundary Waiting Time

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

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

The Amount of Fluid that Enters Service in a Small Interval

$$
\begin{aligned}
E(t+\delta) - E(t) &\approx b(t,0)\delta \approx q(t, w(t))(w(t) - [w(t+\delta) - \delta]) \\
&\equiv q(t, w(t))(-[w(t+\delta) - w(t)] + \delta), \quad \text{so} \\
\frac{b(t,0)}{q(t, w(t))} &\approx -\left(\frac{w(t+\delta) - w(t)}{\delta}\right) + 1 \quad \text{and} \\
w'(t) &\equiv \frac{dw(t)}{dt} = 1 - \frac{b(t,0)}{q(t, w(t))}.
\end{aligned}
$$



SUMMARY

1. We have discussed the fluid approximation for many-server queues.
2. It can be used to study the performance impact of delay announcements.
3. and to help make better delay predictions. [See Ibrahim&W^2 papers.]
4. The time-varying $G_t/GI/s_t + GI$ Fluid model is tractable and useful.
5. Analyzed for the case of alternating OL and UL intervals.
6. The algorithm involves: (i) a fixed-point equation for the fluid density in service, and (ii) an ODE for the boundary waiting time.
7. Extension for networks of fluid queues has been developed.
8. Stochastic refinements have been developed.

THE END

References

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.
- **Algs. for T-V Networks of Many-Server Fluid Queues.** IJoC, 2014.
- **M-S H-T Limits for Queues with T-V Parameters.** AnAP, 2014.

Background References: Fluid Approximations

- textbook: R. W. Hall. **Queueing Methods for Services and Manufacturing.** Prentice Hall, Englewood Cliffs, NJ, 1991, Ch. 6.
- $G/GI/s + GI$ fluid model: W^2. **Fluid models for multiserver queues with abandonments.** *Operations Res.*, 54 (2006) 37–54.
- 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.
- accuracy: A. Bassamboo & R. S. Randhawa. **On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers.** *Operations Res.* 58 (2010) 1398-1410.

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.

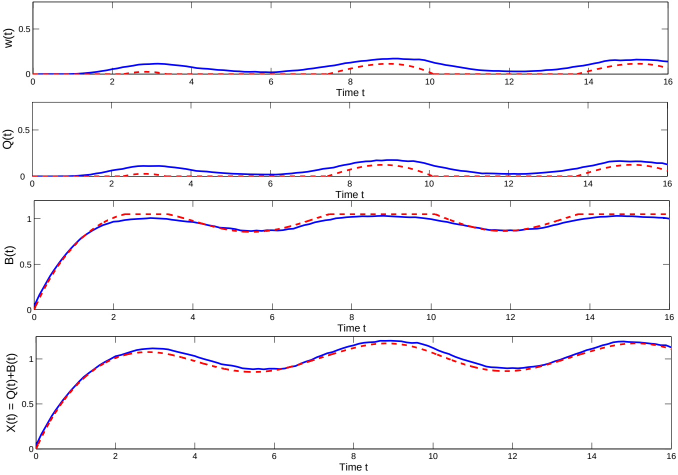
MSHT Limits for Infinite-Server queues

- 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.** AnApP. 2015 Relates to G. Kallianpur & Perez-Abreu (1988,1989).

Extra Slides

Comparison with Simulation: Even Smaller n

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



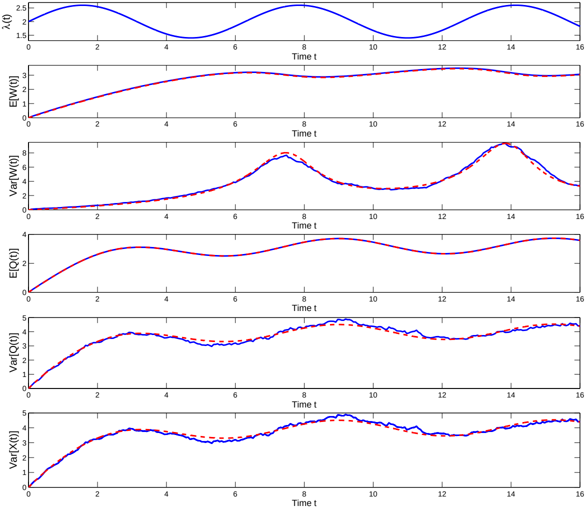
Extensions: FCLT and Stochastic Refinements

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. See: A. K. Aras, X. Chen and Y. Liu, **Many-server Gaussian limits for overloaded non-Markovian queues with customer abandonment**, Queueing Systems, 2018.

Example: Gaussian approximation for an OL Interval

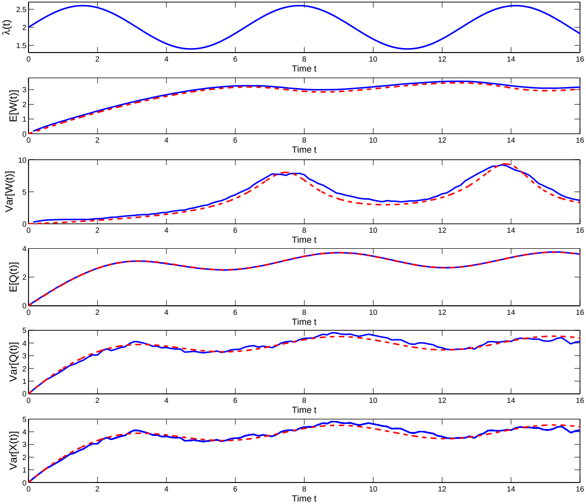
- the model: $M_t/M/s_t + M$
- $\lambda(t) = 2.0 + 6 \cdot sin(t)$, $s(t) = s = 0.4$, $\mu = 1$, $\theta = 0.5$
- initially critically loaded, $X(0) = s$
- queueing model has $n = 100$
- estimates based on 1000 replications

Comparisons with Simulation for $n = 100$



Averages of multiple (1000) sample paths

Comparisons with Simulation for $n = 25$



Averages of multiple (1000) sample paths