Staffing to Stabilize Performance at Target Levels in Many-Server Queues with Time-Varying Arrivals, Longer Service Times and Flexible Staffing

Ward Whitt, IEOR Department

Applied Probability and Risk Seminar, September 12, 2013

based on

"Stabilizing Customer Abandonment in Many-Server Queues with Time-Varying Arrivals," ***Operations Research*** **60 (2012) 1551-1564,**
by **Yunan Liu and** $\mathbf{W}^2$**.**

Who is W^2?

research areas: **applied probability, queues and asymptotic methods**

book: **Stochastic-Process Limits, Springer, 2002**

- 1969 PhD in OR, Cornell [heavy-traffic limits for queues, Iglehart]
- 1969 – 1977 Dept Administrative Sciences and Statistics, Yale
- 1977 – 2002 Research Organizations at AT&T
  - 1977 – 1986 Operations Research Center, Bell Labs, Holmdel, NJ
  - 1986 – 1996 Math. Sci. Res. Ctr., Bell Labs, Murray Hill, NJ
  - 1996 – 2002 Networking Res. Ctr., AT&T Labs, Florham Park, NJ
- 2002 – 2013 IEOR Dept., Columbia

current research focus: **many-server queues and their application to service systems such as call centers and hospitals**

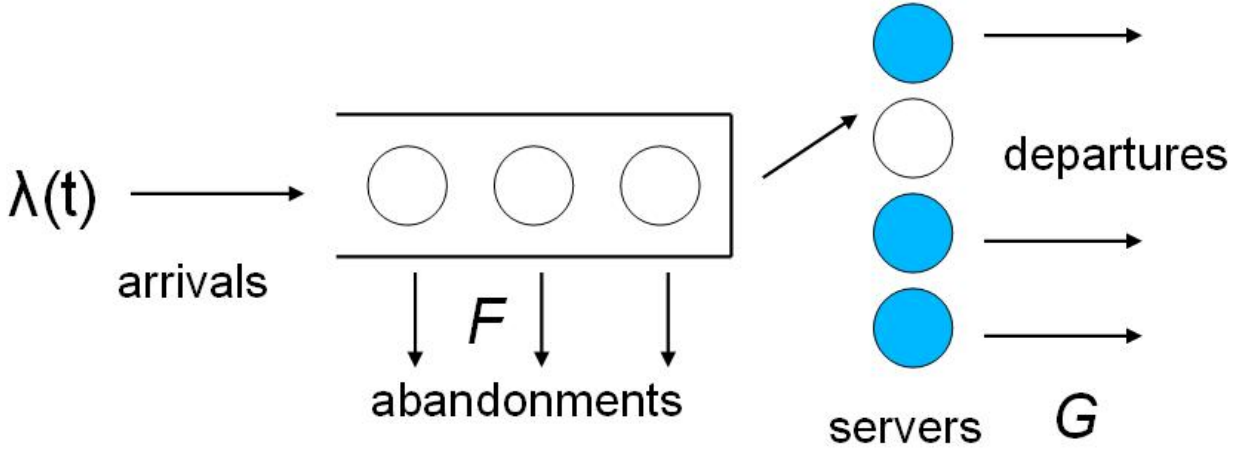
More Information

- web page: http://www.columbia.edu/∼ww2040
  - all research papers for downloading
  - IEOR 4615, Service Engineering: course in spring 2012-2014
  - link to Avishai Mandelbaum's Service Engineering web page at the Technion: http://iew3.technion.ac.il/serveng/
- email: ww2040@columbia.edu
- current research collaborators:
  - PhD student: Song-Hee (Hailey) Kim, joint advising with Carri Chan, DRO
  - postdoctoral fellow: Jamol Pender, PhD Princeton with W. A. Massey
  - former PhD students: Yunan Liu, North Carolina State University; Guodong Pang, Pennsylvania State University; Rouba Ibrahim, University College London; Itai Gurvich and Ohad Perry, Northwestern University; Rodney Wallace, IBM;

Outline

1. The $M_t/GI/s_t + GI$ Model and the relevant limits
2. Motivation
   - the full staffing problem
   - system complexities: need for M_t and GI instead of just M
3. Alternative Staffing Methods
   - PSA, OL & SRS, MOL, ISA
   - An Example: Comparisons with Simulation
4. New Method: DIS-MOL (fancy OL and MOL)
5. The Example Revisited: Simulation Verification
6. Extension to Networks
7. Summary

The Queueing Model



- Potential delay $W(t)$ ($P(W(t) > x)$ and $E[W(t)]$)
- Abandonment probability $\mathbb{P}(Ab(t))$

The Queueing Model

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

- Poisson arrival process, time-varying rate $\lambda(t)$
- I.I.D. service times $S \sim G$
- Staffing level $s(t)$
- I.I.D. abandonment times $A \sim F$
- First-Come First-Served (FCFS)
- Unlimited waiting capacity

Many-Server Heavy-Traffic (MSHT) Limits

Increasing Scale Increasing Scale

- a sequence of $G_t/GI/s_t + GI$ models indexed by n,
- arrival rate function **grows**: $\lambda_n(t)/n \to \lambda(t)$ as $n \to \infty$,
- time-varying number of servers **grows**: $s_n(t)/n \to s(t)$ as $n \to \infty$,
- service-time cdf G and patience cdf F held **fixed** independent of n with mean service time 1: $\mu^{-1} \equiv \int_0^\infty x\, dG(x) \equiv 1$.

References

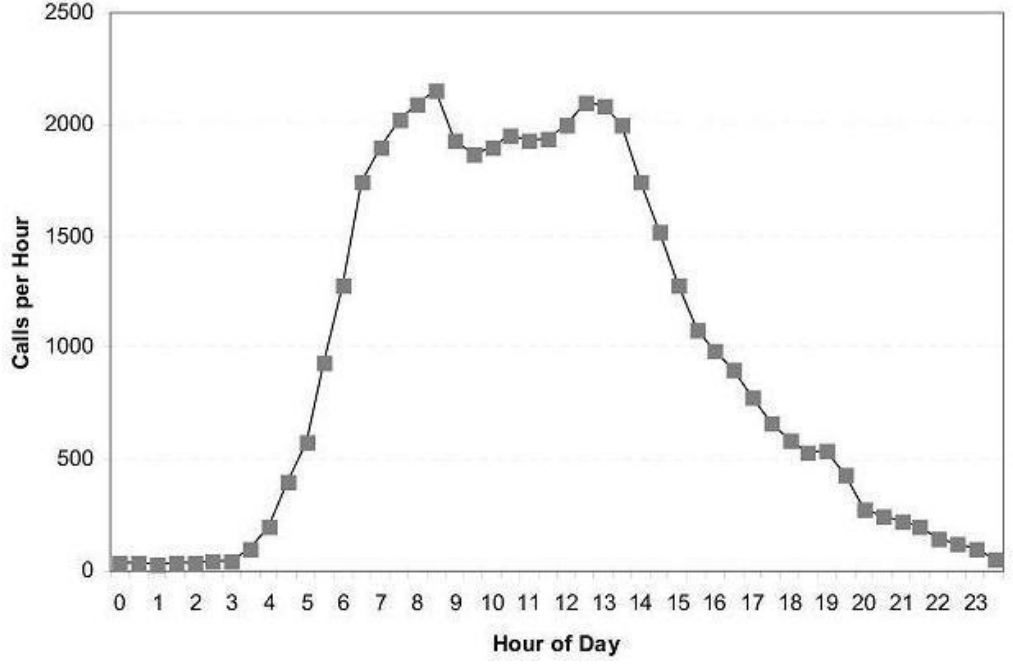
1. A. Mandelbaum, W. A. Massey, and M. I. Reiman. Strong approximations for Markovian service networks, Queueing Systems 30 (1998) 149-201.
2. Y. Liu and W^2. A Many-Server Fluid Limit for the $G_t/GI/s_t + GI$ Queueing Model Experiencing Periods of Overloading, Operations Research Letters 40 (2012) 307-312.
3. Y. Liu and W^2. Many-Server Heavy-Traffic Limits for Queues with Time-Varying Parameters, Annals of Applied Probability, published online, April 2013.

The Full Staffing Problem

1. Model Fitting ($M_t/GI/s_t + GI$)
   - Forecasting (estimating the arrival rate function from historical data)
   - Data analysis (service times and abandonment hazard rate)
2. Setting Staffing Levels (specify $s(t)$)
   - The standard method: Pointwise Stationary Approximation (PSA)
   - New methods for longer service times
     - Offered Load (OL), MOL, ISA, DIS-MOL (agree with PSA if PSA good)
3. Shift Scheduling
   - Set up shifts (meet target level subject to constraints, integer programming)
   - Assign people to shifts (personnel issues, union contracts)
4. Real-Time Control (Responding to unexpected overloads)
5. More Complex Environments: Skill-Based Routing, Distributed Call Centers

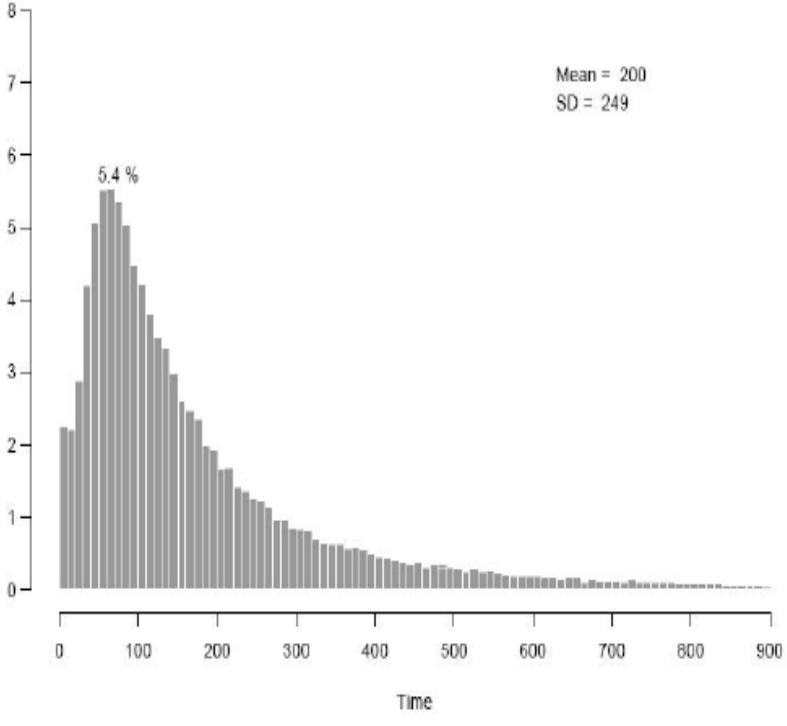
Motivation

Why hard? time-varying arrivals and longer service times

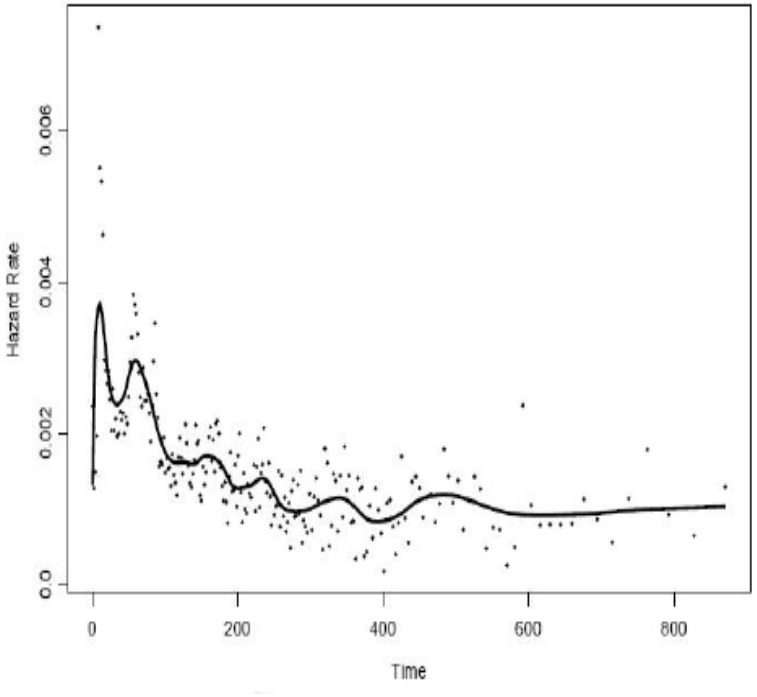


financial service call center from Green, Kolesar and Soares (2001)

Non-Exponential Service and Abandonment



service



abandonment

Brown et al. (2005)

Motivation

Staff to Meet SLA's

Service Level Agreements (SLA's)

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

Motivation

Staff to Meet SLA's

Service Level Agreements (SLA's)

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

How?

Variations of One Key Idea

Offered Load Analysis: Exploit Infinite-Server (IS) Queues.

1. **Simple Performance Formulas for an IS Queue**

   S. G. Eick, W. A. Massey and W^2. The Physics of The $M_t/G/\infty$ Queue. *Operations Research* 41 (1993) 731-742.

2. **Extends to Networks**

   W. A. Massey and W^2. Networks of Infinite-Server Queues with Nonstationary Poisson Input. *Queueing Systems* 13 (1993) 183-250.

3. Refinements using **Modified Offered Load (MOL) approximation**

   Use stationary model in nonstationary way with arrival rate depending on OL.

Alternative Staffing Methods

- **P**ointwise **S**tationary **A**pproximation (PSA)
  Standard methods, *Green and Kolesar (91,97,01), W^2 (91)*
  - Shorter service times, high quality of service

- **M**odified **O**ffered **L**oad (MOL)
  Jagerman (75); Massey&W^2 (94,97); Jennings et al. (96); Feldman et al.(08)
  - Longer service times, high quality of service

- **Simulation-Based Iterative Staffing Algorithm** (ISA)
  Feldman et al.(08)
  - Stabilize probability of delay, extends to other criteria and models, Supports MOL, But does NOT stabilize $P(\textbf{Aban}(t))$ and $E[\textbf{Wait}(t)]$

- **D**elayed-**I**nfinite-**S**erver **M**odified-**O**ffered-**L**oad (DIS-MOL)
  Liu and W^2 (12)
  - Analytical method, theoretical insights, generalizes, stabilizes $P(\textbf{Aban}(t))$ and $E[\textbf{Wait}(t)]$

Pointwise Stationary Approximation

Basic Idea (to set staffing levels $s(t)$)

- *Given* Non-stationary $\mathbf{M_t/GI/s_t + GI}$: $\lambda(t)$,
- *Use* Stationary $\mathbf{M/GI/s + GI}$: λ, s, X_∞
- For fixed t, in the stationary model
  - let $\lambda \leftarrow \lambda(t)$, choose s s.t. $\mathbb{P}(Delay) = \mathbb{P}(X_\infty \geq s) \approx \alpha$
  - Let $s(t) \leftarrow s$. Do for all t, get $\{s(t) : 0 \leq t \leq T\}$.

PSA works well when

- Short service time, high service quality

The Offered Load (OL) Approach

- The Offered Load (OL) $m(t)$ is the mean number of busy servers in the $M_t/GI/\infty$ infinite-server model (Poisson).

- We can use a normal approximation for the Poisson distribution to obtain the square root staffing (SRS) rule

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

- A refinement is to use the Modified Offered Load (MOL) method. It uses the stationary $M/GI/s+GI$ model (in a time-varying way) with arrival rate based on the offered load

$$\lambda_{MOL}(t) = \frac{m(t)}{E[S]} = m(t), \quad 0 \le t \le T\ ,$$

where $m(t)$ is the offered load. Choose $s(t)$ at time t to achieve the delay probability target for stationary model with $\lambda = \lambda_{MOL}(t)$.

Iterative staffing algorithm (ISA)

- $\mathbf{M_t/GI/s_t + GI}$
- Simulation based
- Stabilize any delay probability target $0 \leq \alpha \leq 1$
- Independent of models or parameters
- Requires very large number of simulation runs
- Supports MOL method

Iterative staffing algorithm (ISA)

Main algorithm

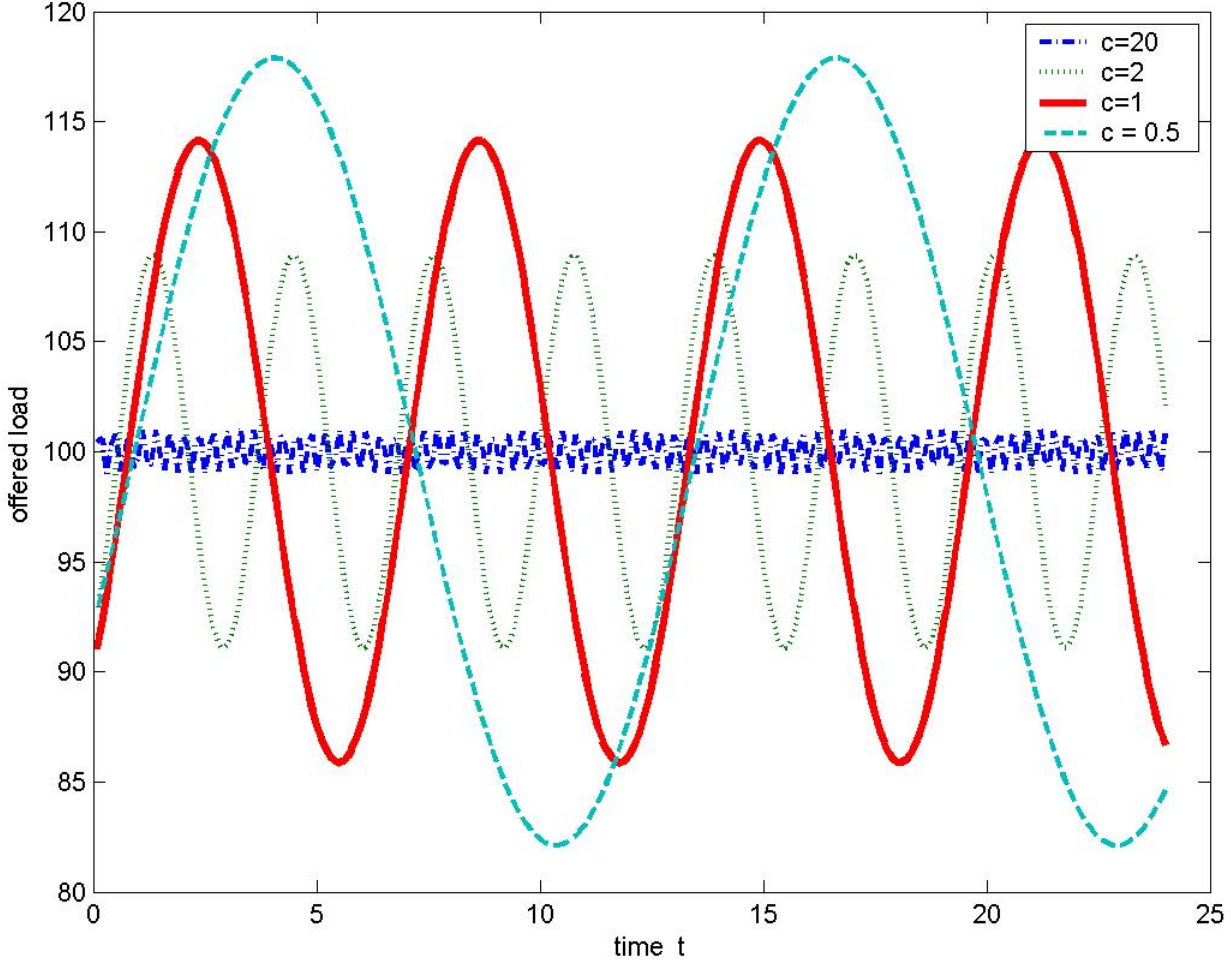
- Initialize: $s^{(1)}(t) \leftarrow s^*(t)$ for $0 \le t \le T$
  (e.g., use offered load: $s^*(t) = m(t) + \beta\sqrt{m(t)}$)
- Given $\{s^{(i)}(t) : 0 \le t \le T\}$, evaluate the distribution of $X^{(i)}(t)$ for $0 \le t \le T$ (average of many simulation runs)
- For each t, find $s^{(i+1)}(t) = s$ such that
  $\mathbb{P}(X^{(i)}(t) \ge s) \approx \alpha$
- If $\| s^{(i+1)} - s^{(i)} \| < \epsilon$, end;
  else, $i \leftarrow i + 1$, go back to line 2

An Example

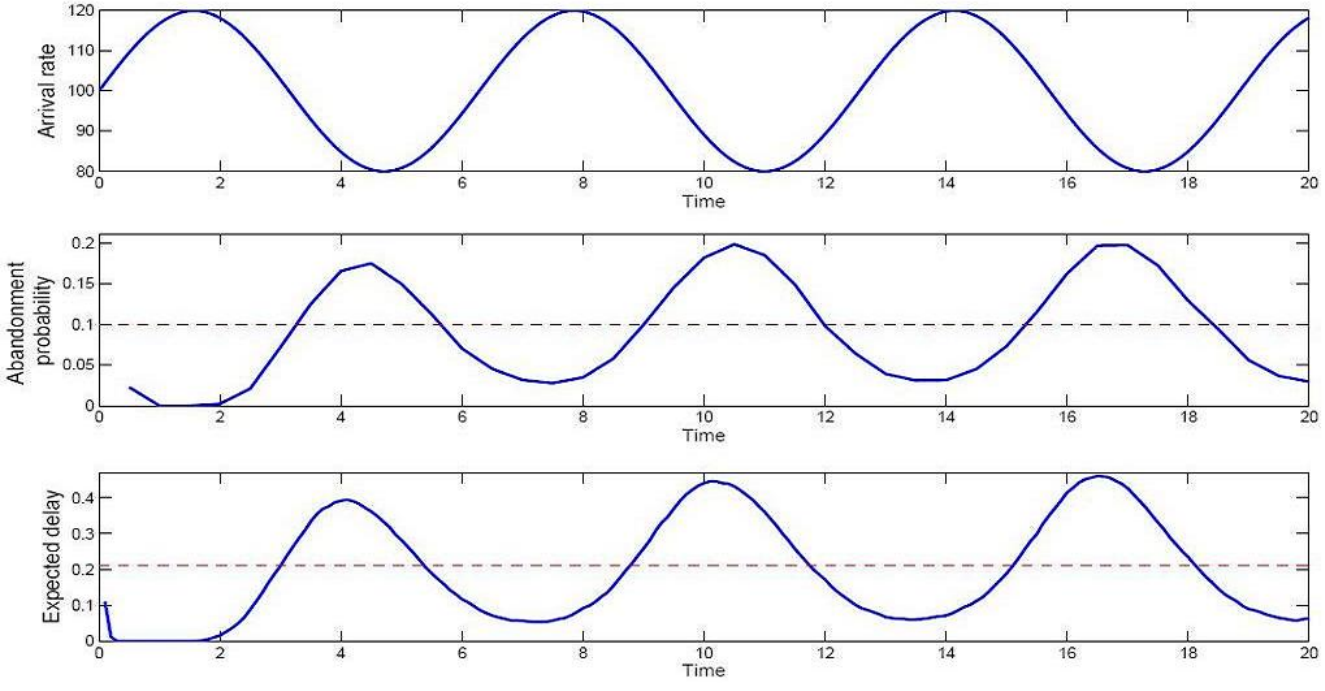
$M_t/M/s_t + M$

- Arrival Rate: $\lambda(t) = a + b\sin(ct)$
  - parameters: $a = 100$, $b = 20$, $c = 1$
- Service rate: $\mu = 1$; Abandonment Rate: $\theta = 1$
- Offered Load: mean number of busy servers in $M_t/M/\infty$ model
  - $m(t) = a + \frac{b}{1+c}(\sin(ct) - c\cos(ct))$
  - See Eick, Massey and W^2 (1993 a,b) for formulas and derivation.

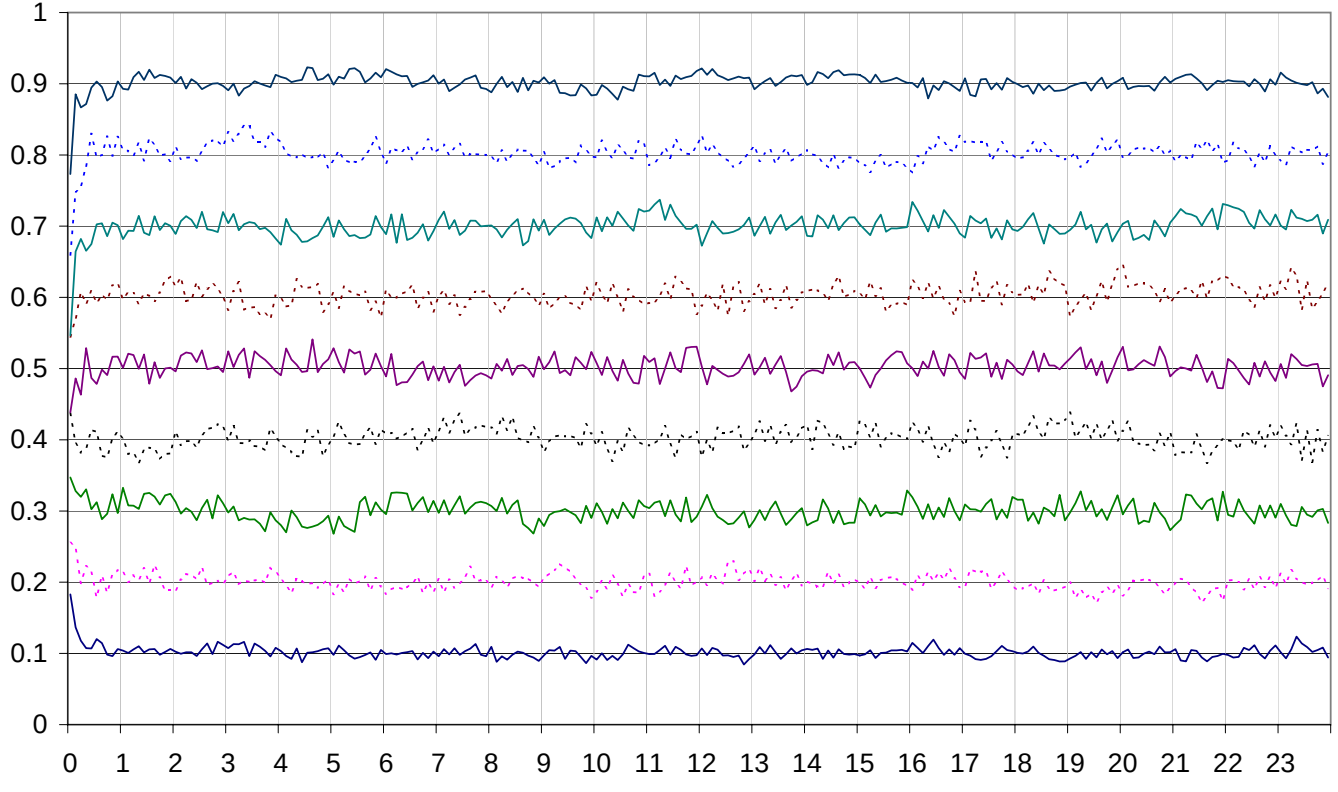
The Offered Load as a function of c in the arrival rate function $\lambda(t) = 100 + 20\sin(ct)$



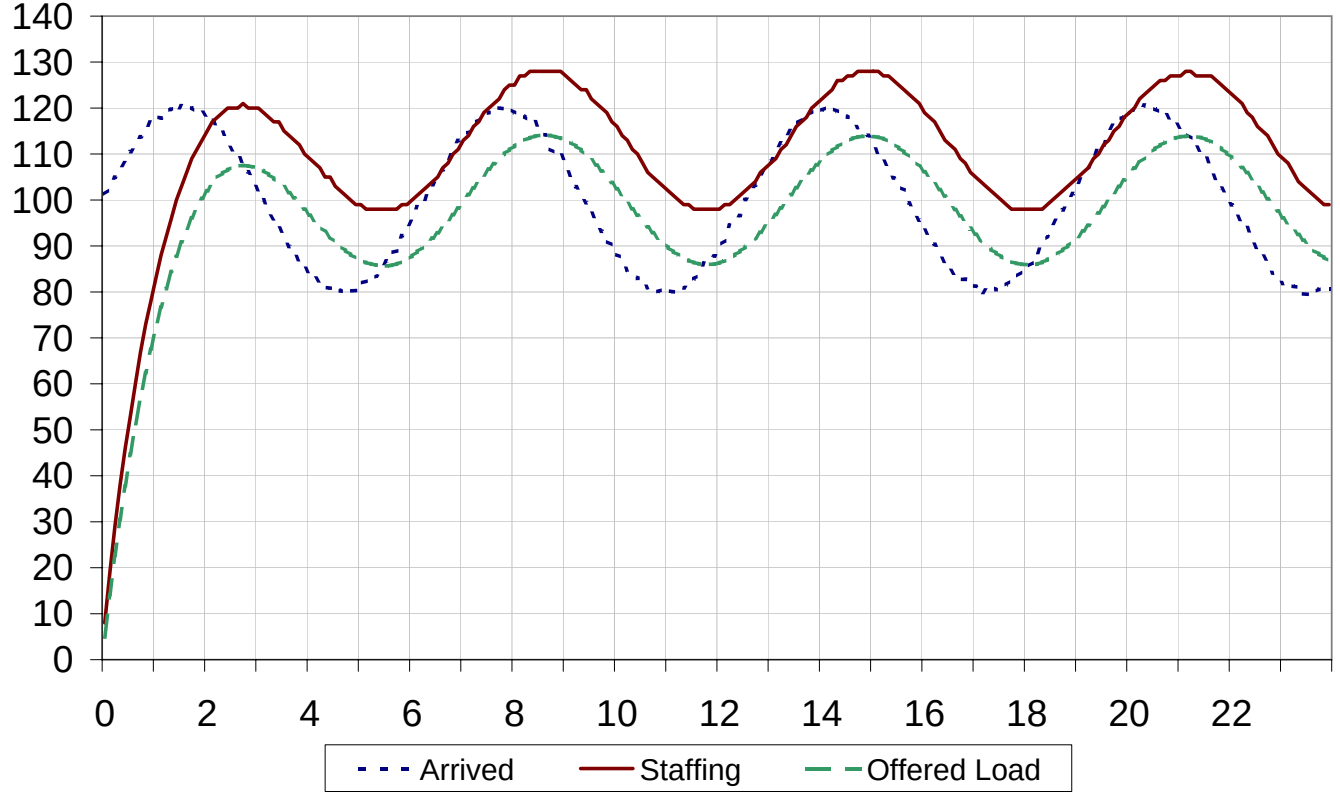
Long Service Times: PSA is Bad



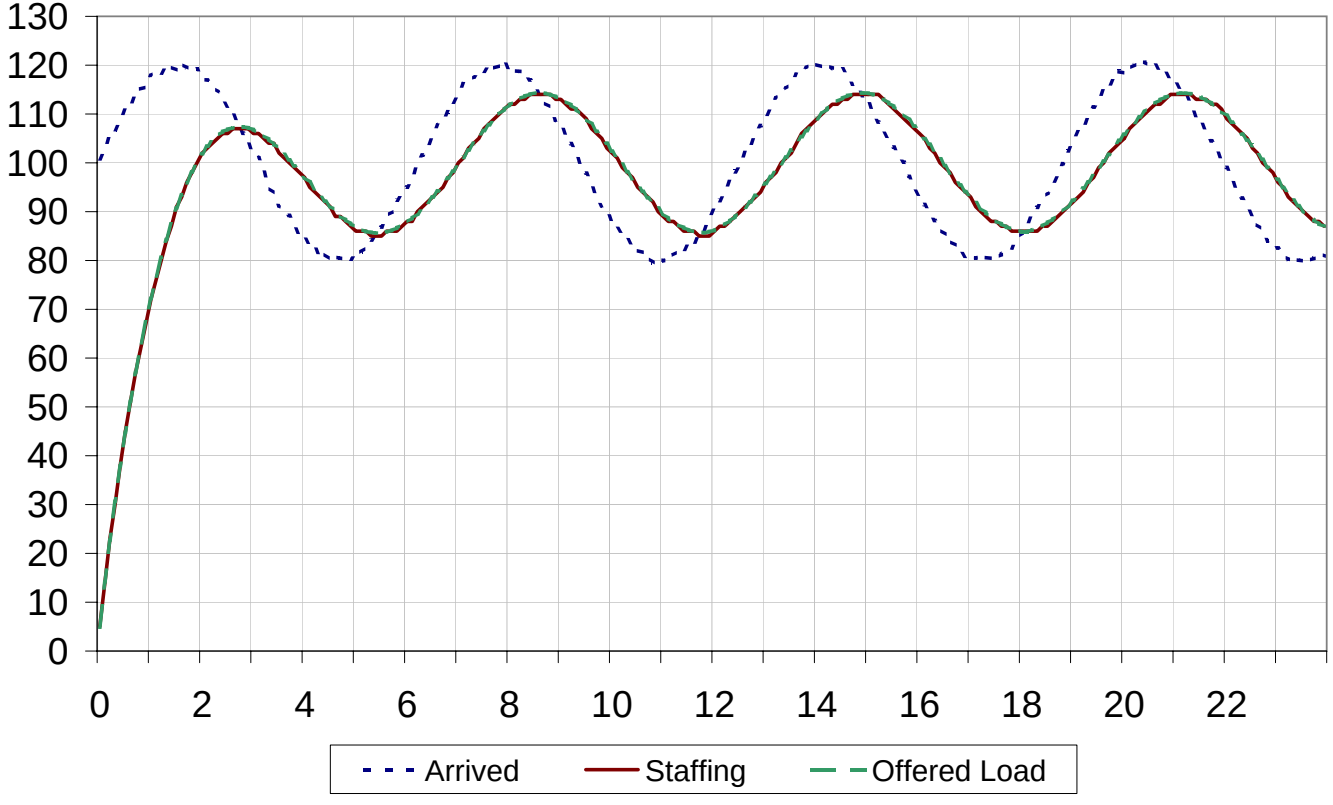
The ISA & MOL Delay Probability Functions
$P(\textbf{Delay}(t))$ for 9 Targets: $0.1, 0.2, \ldots, 0.9$



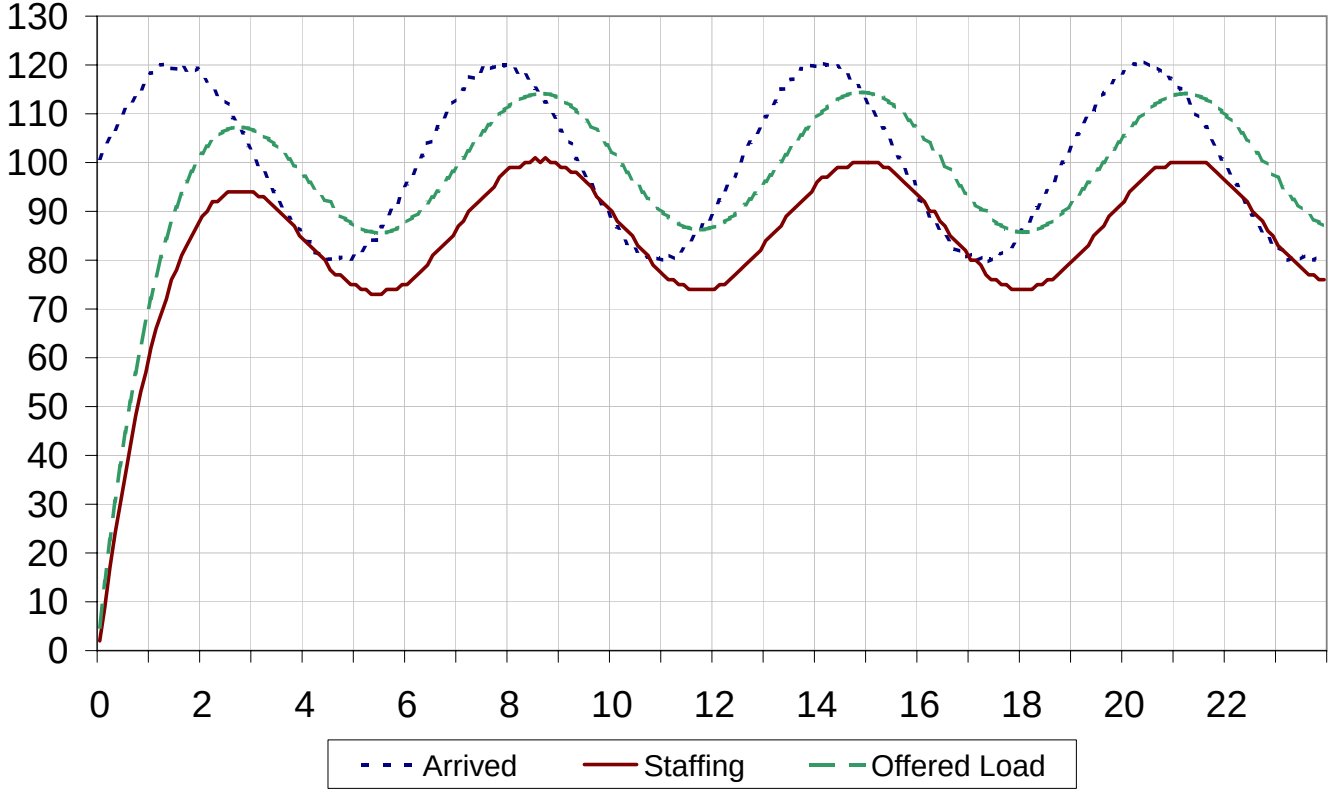
ISA & MOL Staffing for High QoS (QD): $P(\textbf{Delay}) = 0.1$



ISA & MOL Staffing for Good QoS (QED): $P(\textbf{Delay}) = 0.5$



ISA & MOL Staffing for Low QoS (ED): $P(\textbf{Delay}) = 0.9$



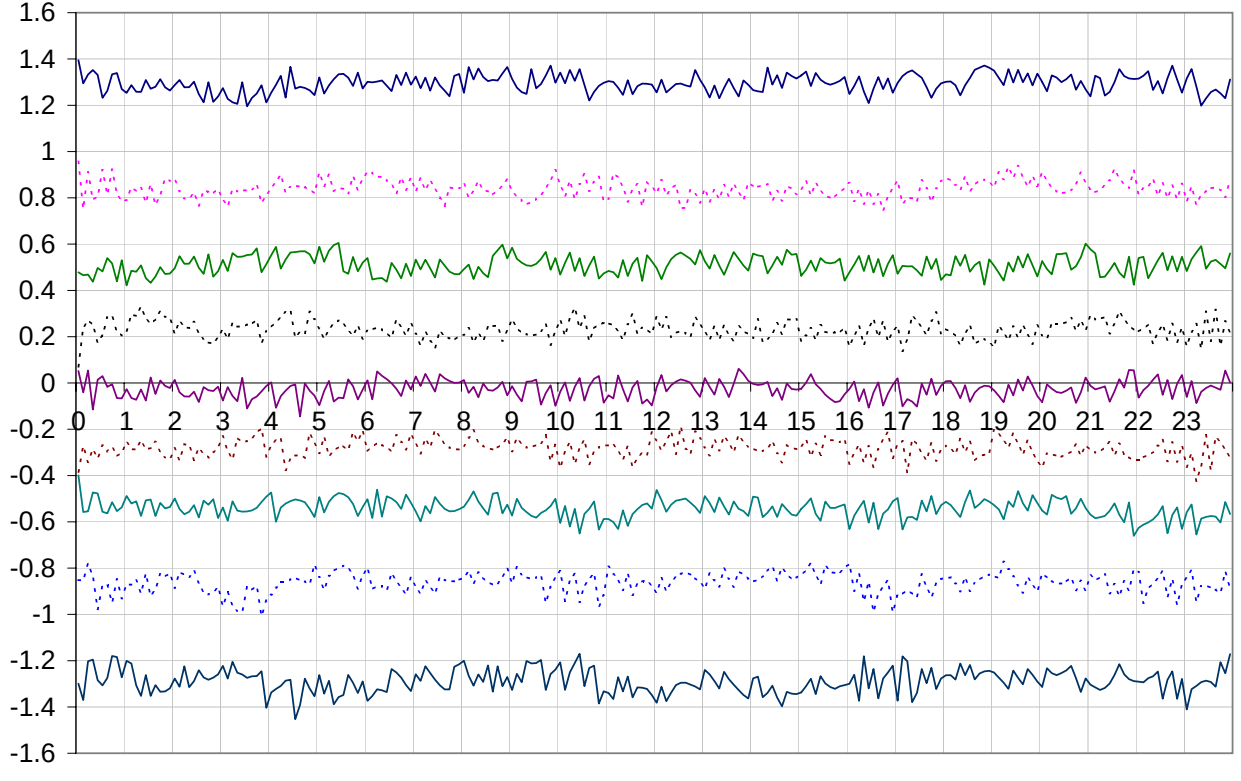
Validate the Square Root Staffing Formula

The implied empirical Quality of Service

$$\beta^{ISA}(t) \equiv \frac{s^{ISA}(t) - m(t)}{\sqrt{m(t)}}, \quad 0 \leq t \leq T \ ,$$

where $m(t)$ is the offered load.

The Implied Empirical QoS $\beta^{ISA}(t)$ for the 9 Targets: $0.1, 0.2, \ldots, 0.9$



ISA Coincides with the MOL

For the $M_t/M/s+t+M$ model, we can use MOL, i.e., the $M/M/s+M$ model with arrival rate

$$\lambda_{MOL}(t) = \frac{m(t)}{E[S]} = m(t), \quad 0 \le t \le T \ ,$$

where $m(t)$ is the offered load. Choose $s(t)$ at time t to achieve delay probability target for stationary $M/M/s+M$ model with $\lambda = \lambda_{MOL}(t)$.

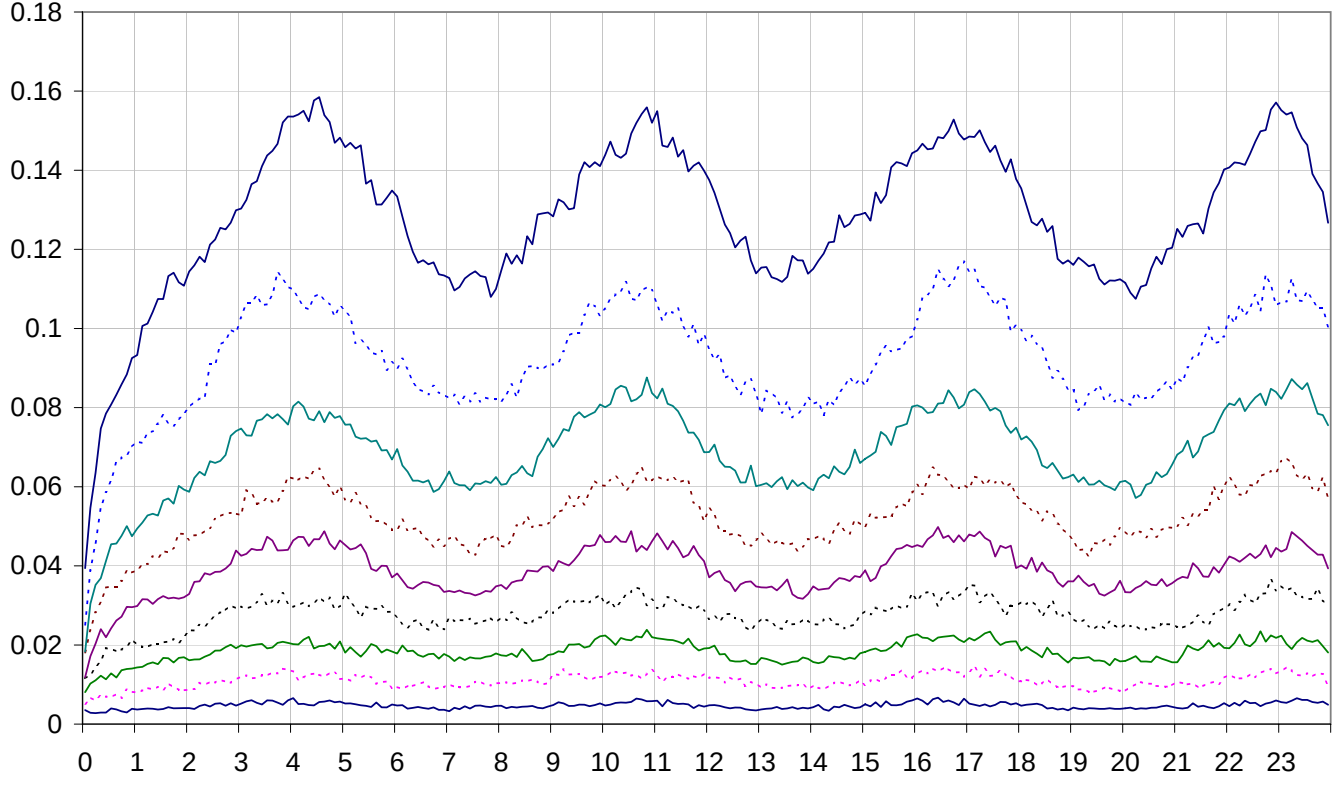
Motivation for new DIS-MOL approximation

target used so far: delay probability function $P(\textbf{Delay}(t))$

Service Level Agreements (SLA)

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

With ISA and MOL Staffing: the Abandonment Probabilities $P(\text{Aban}(t))$ for the 9 Targets: $0.1, 0.2, \ldots, 0.9$



An Approximating Model

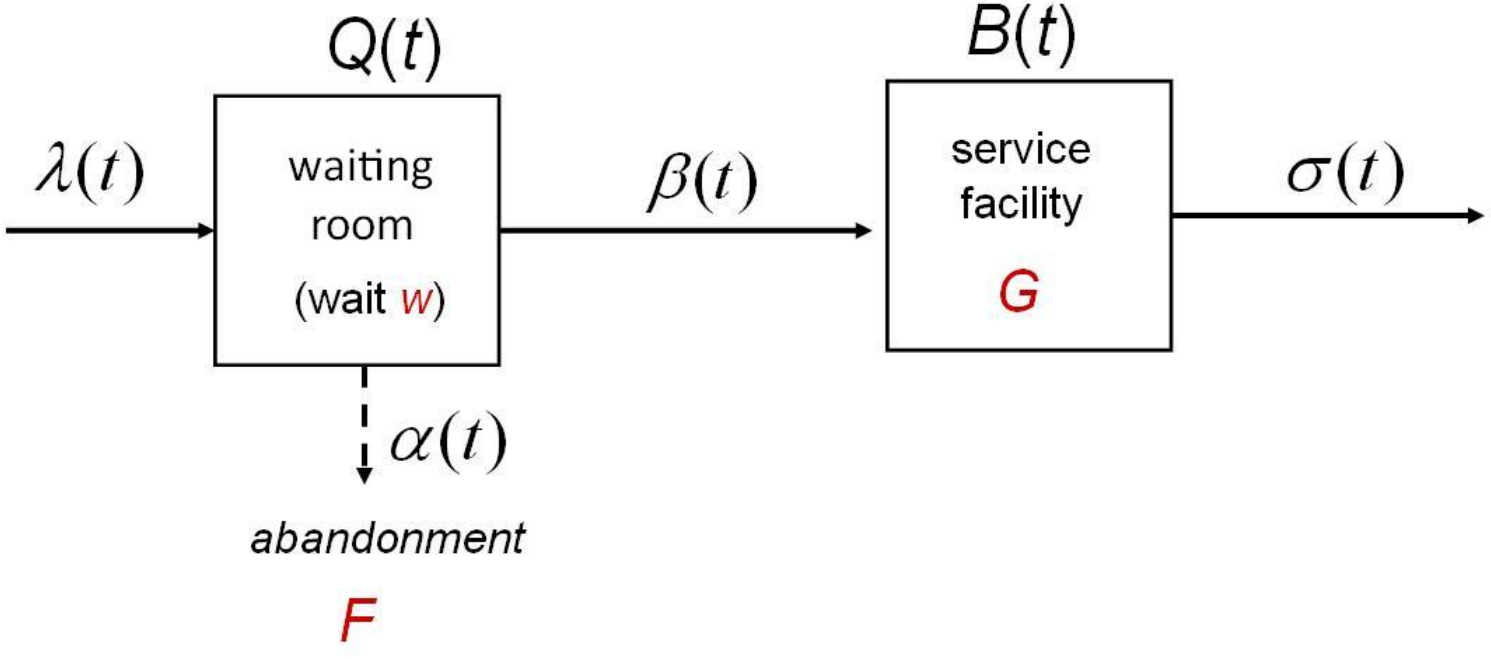
Delayed Infinite-Server (DIS) Model

Approximates overloaded $\mathbf{M_t/GI/s_t + GI}$ model, with $(\lambda(t), G, F)$.

- Given Poisson arrival process, with time-varying rate $\lambda(t)$
- Separate IS models for waiting room and service pool.
- **All customers stay exactly w in a waiting room with unlimited capacity**
- While waiting, each abandons independently with given cdf F.
- If not abandoned after w, enter service to receive with given cdf G.

Approximating DIS Model: Network of IS Queues

Decoupling



Two $M_t/GI/\infty$ Models

Fact: Recall that the departure process from an IS queue is Poisson with past at any time independent of queue length at that time, so contents of two IS queues at any one time are independent Poisson random variables.

Content of the waiting room Q(t)

- Time-varying arrival rate $\lambda(t)$
- I.I.D. Service times $T = A \wedge w$, $A \sim F$

Content of the service facility B(t)

- Time-varying arrival rate $\beta(t) = \bar{F}(w)\lambda(t-w)$
- I.I.D. Service times $S \sim G$

The Offered Load

$Q(t) \sim$ **Poisson(**$\mathbb{E}[Q(t)]$**)**

$$\mathbb{E}[Q(t)] = \mathbb{E}[\lambda(t - T_e)]\mathbb{E}[T], \quad T = A \wedge w$$

$B(t) \sim$ **Poisson(**$\mathbb{E}[B(t)]$**)**

$$\mathbb{E}[B(t)] = \bar{F}(w)\mathbb{E}[\lambda(t - w - S_e)]\mathbb{E}[S]$$

Offered Load (OL) $\equiv$ **m**$_\alpha$ **(t)** $\equiv$ **m(t)** $\equiv \mathbb{E}[B(t)]$,

where $\alpha \equiv P($**Aban**$)$ **target,** $\alpha = F(w)$ and, for any rv X, X_e has the stationary-excess distribution associated with X:

$$P(X_e \le t) \equiv \frac{1}{EX} \int_0^t P(X > u)\, du$$

Simple DIS Staffing

For high abandonment-probability (low quality-of-service) targets, use naive staffing rule:

let $s(t) = m(t)$.

The New Modified Offered Load: DIS-MOL

- **For each t, define the MOL arrival rate**

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

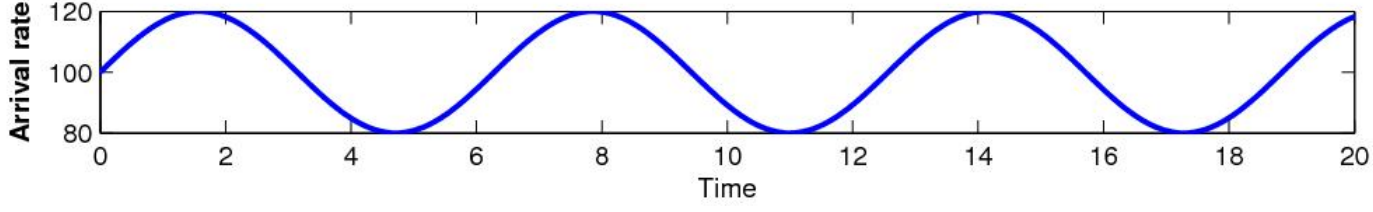
- **For each t, find $s^{\alpha}(t)$ such that steady-state $\mathbb{P}(Ab) \approx \alpha$ of $\mathrm{M/GI/s+GI}$ with $\lambda = \lambda^{MOL}(t)$, $\quad s = s^{\alpha}(t)$**

- **To carry out last step, approximate $\mathrm{M/GI/s+GI}$ model by an associated $\mathrm{M/M/s+M(n)}$ model with state-dependent abandonment rates, and apply "engineering approximation" in W^2 (2005).**

The Markovian Sinusoidal Example Revisited

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

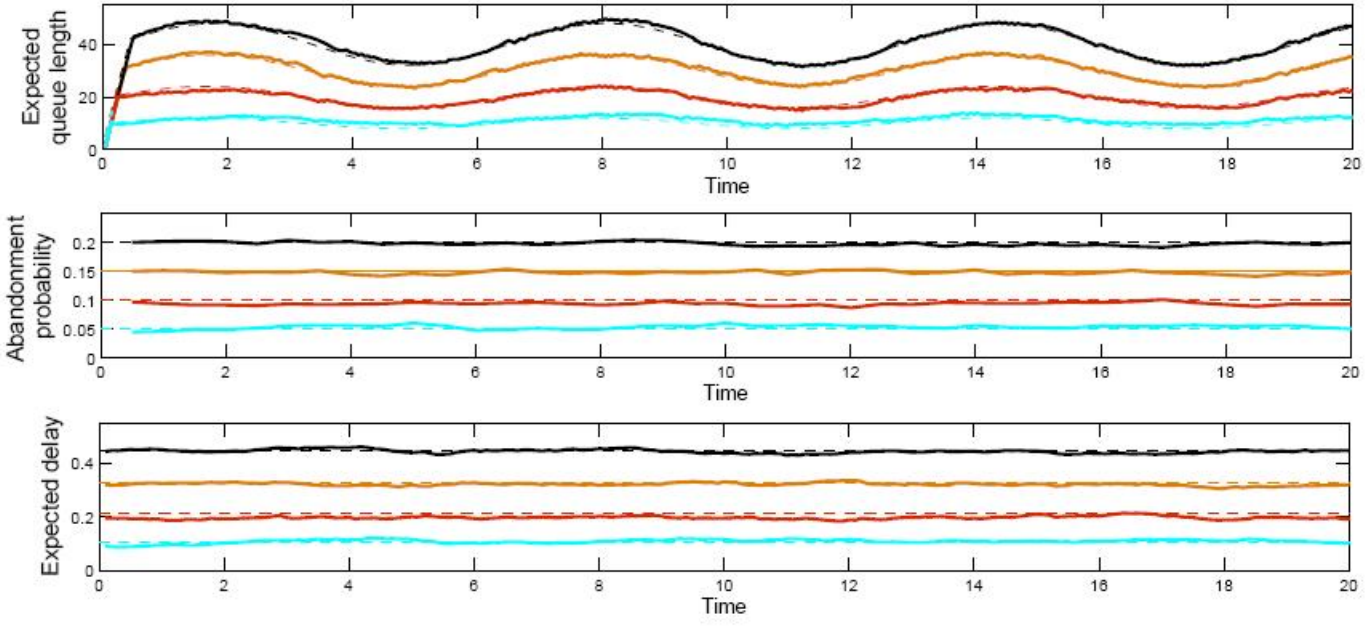
- $\lambda(t) = 100 + 20 \cdot \sin(t)$



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

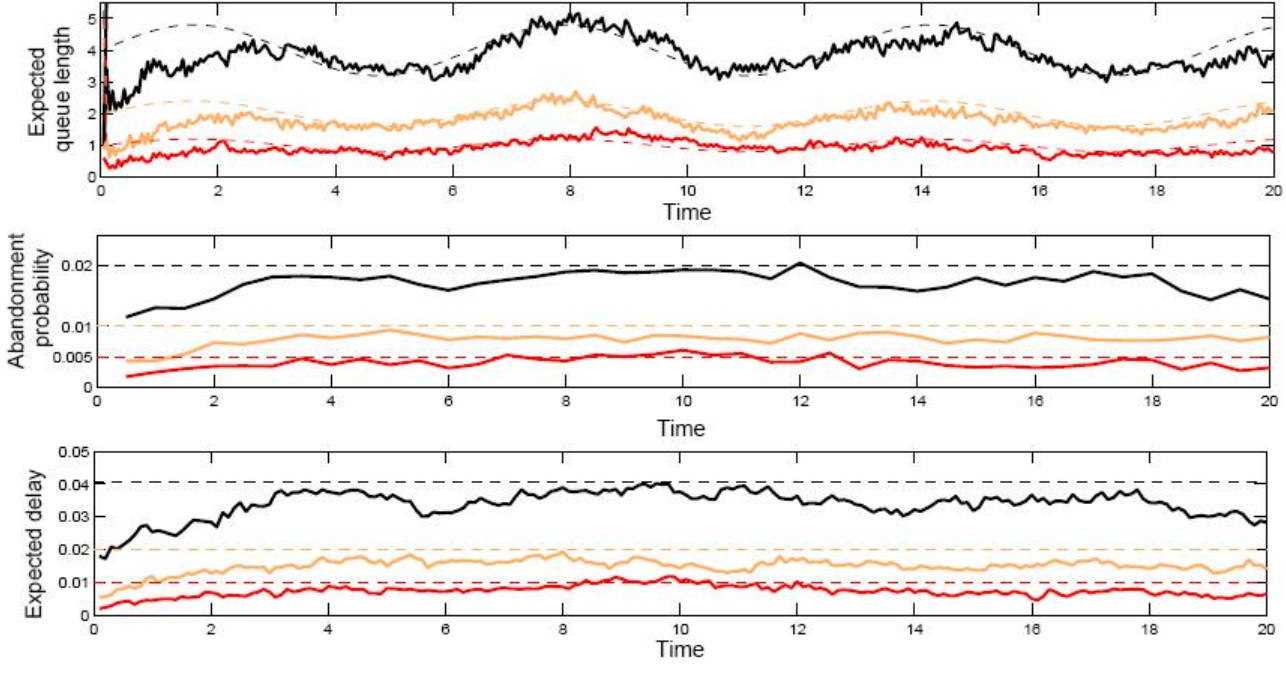
Simulation Verification

Heavy load: $5\% \leq \alpha \leq 20\%$; **DIS OL alone works.**



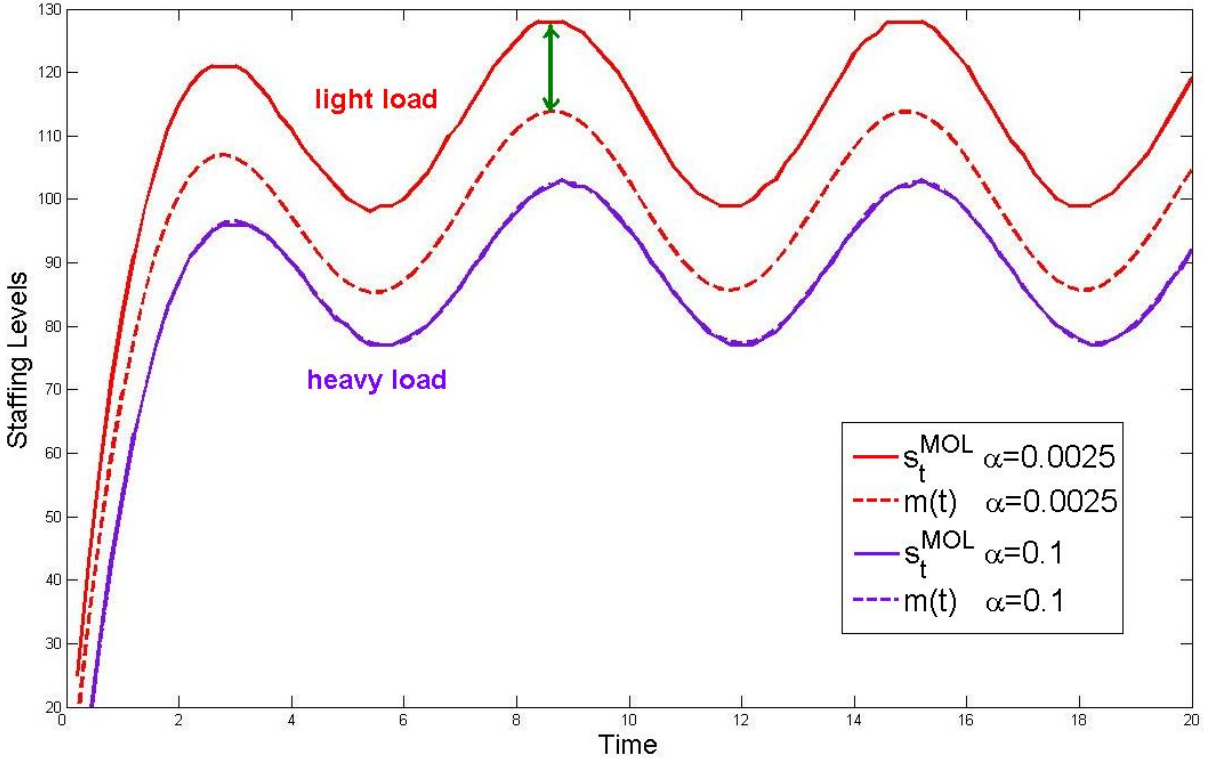
Simulation Verification

Light load: $0.5\% \leq \alpha \leq 2\%$**; need full DIS-MOL.**



A Markovian Example

$m(t)$ and s_t^{MOL}



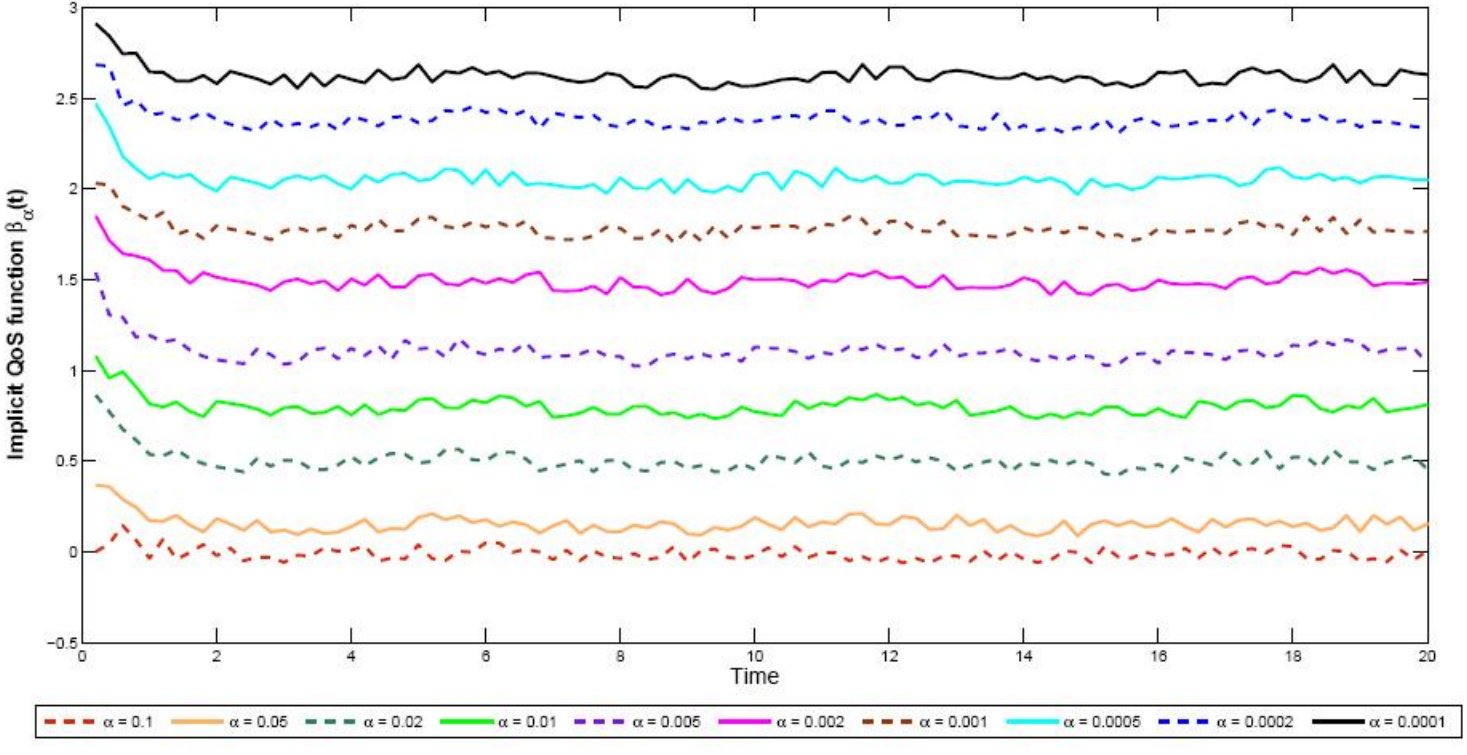
Validate the Square Root Staffing Formula

The implied empirical Quality of Service

$$\beta^{\alpha}(t) \equiv \frac{s^{MOL_{\alpha}}(t) - m_{\alpha}(t)}{\sqrt{m_{\alpha}(t)}}, \quad 0 \leq t \leq T \ ,$$

where $m_{\alpha}(t)$ is the offered load as a function of the target abandonment probability α.

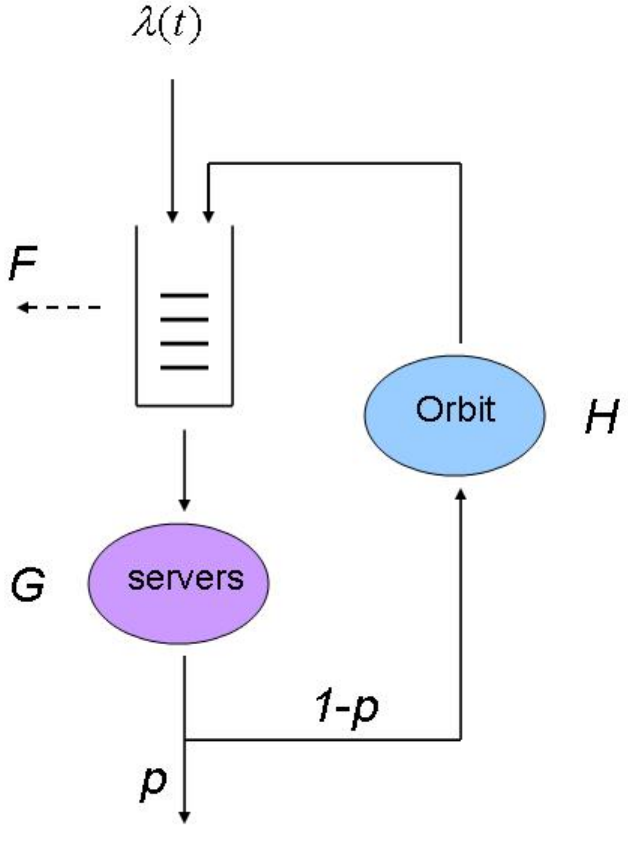
The Implied Empirical QoS $\beta_\alpha(t)$ for the 10 Targets: from 0.10 to 0.0001



Recent papers: Extension to Networks

Example: Queues with Feedback

- **G. Yom-Tov and A. Mandelbaum, Erlang-R: A Time-Varying Queue with ReEntrant Customers, in Support of Healthcare Staffing, The Technion, 2010. To appear in MSOM.**
- **Y. Liu and W^2, Stabilizing Performance in Many-Server Queues with Time-Varying Arrivals and Customer Feedback**, submitted to **Operations Research, 2013.**

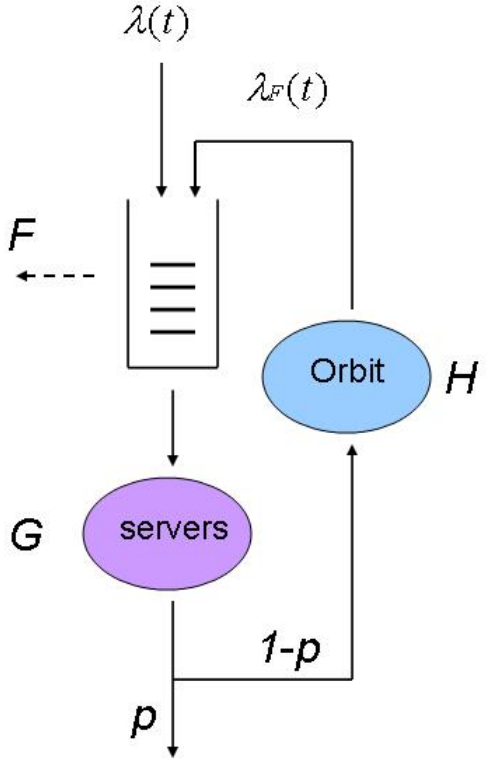

$\lambda(t)$
F
Orbit
H
G
servers
1-p
p

Extension: Queues with Feedback

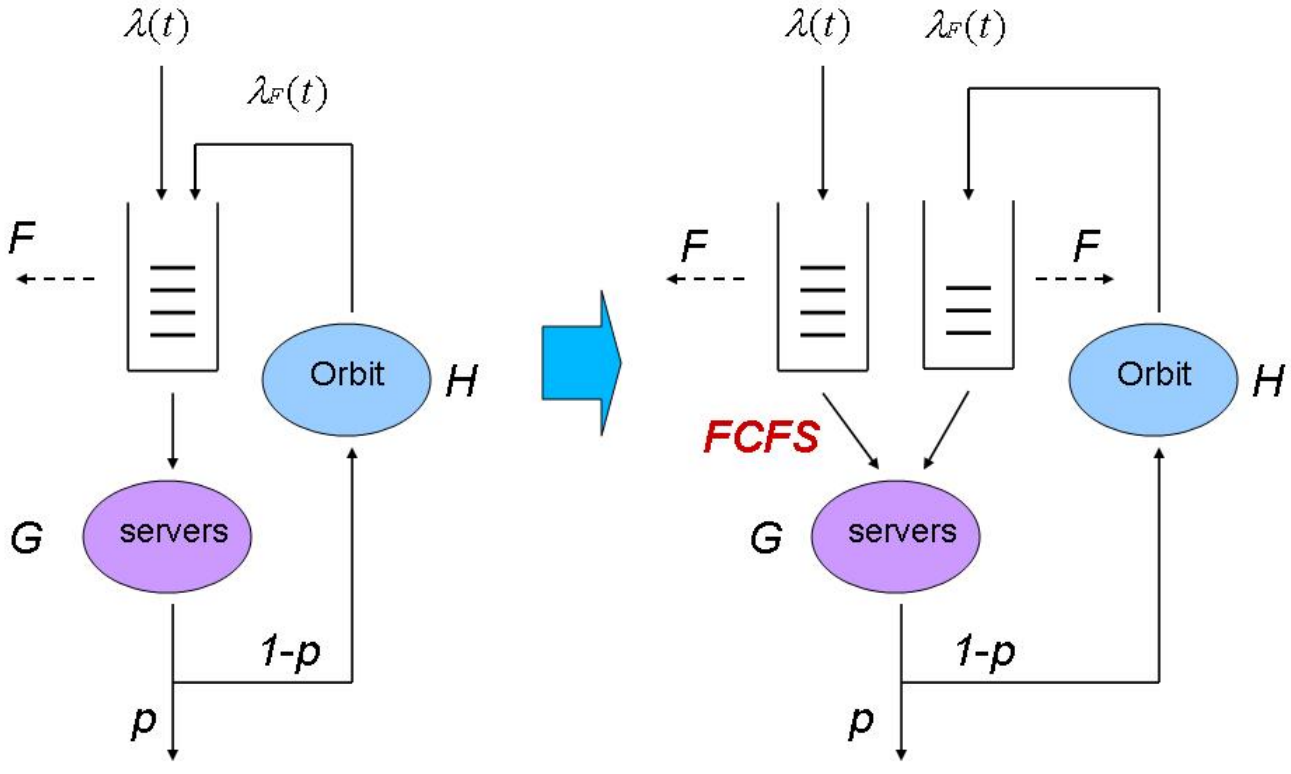
Model description

- Poisson arrival with rate $\lambda(t)$
- I.I.D. service times G
- I.I.D. patience times F
- Retrial with probability $(1-p)$
- I.I.D. orbit times H
- Customer retrial at most once

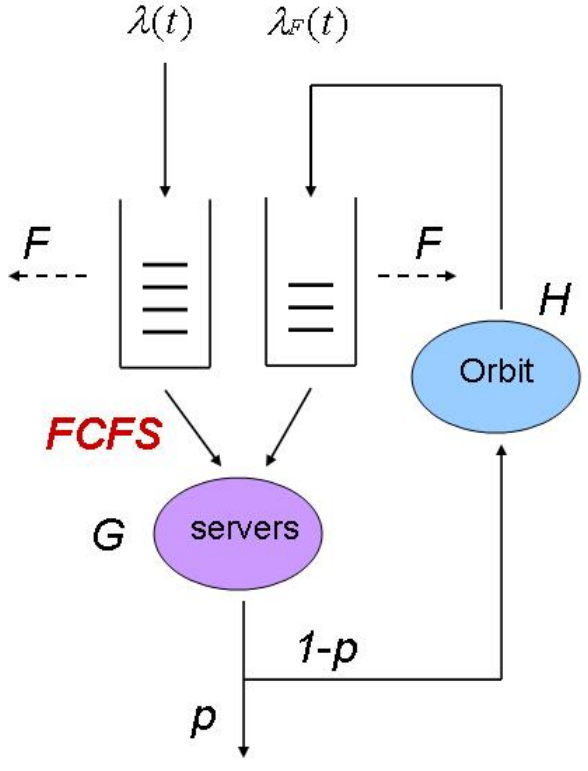
Apply variation of DIS-MOL with 5 IS queues:



Apply variation of DIS-MOL with 5 IS queues:

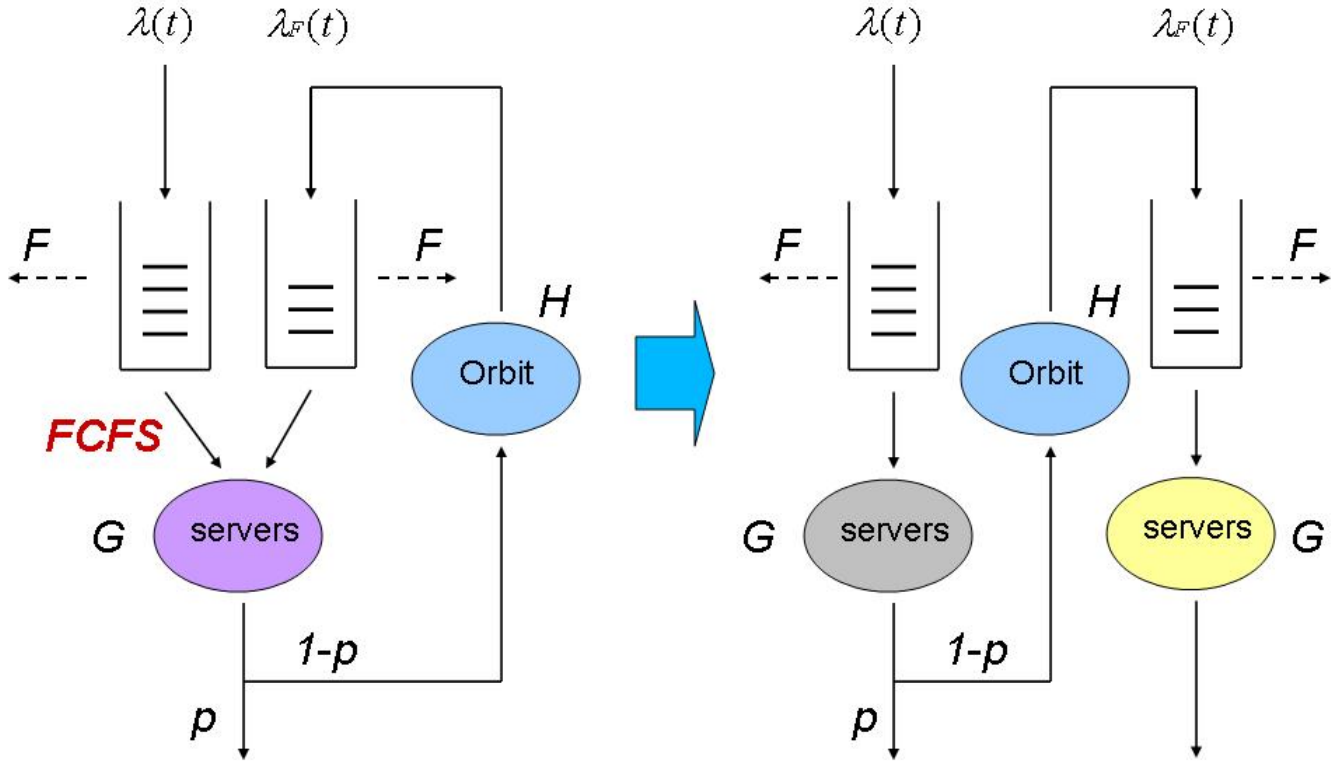


Apply variation of DIS-MOL with 5 IS queues:



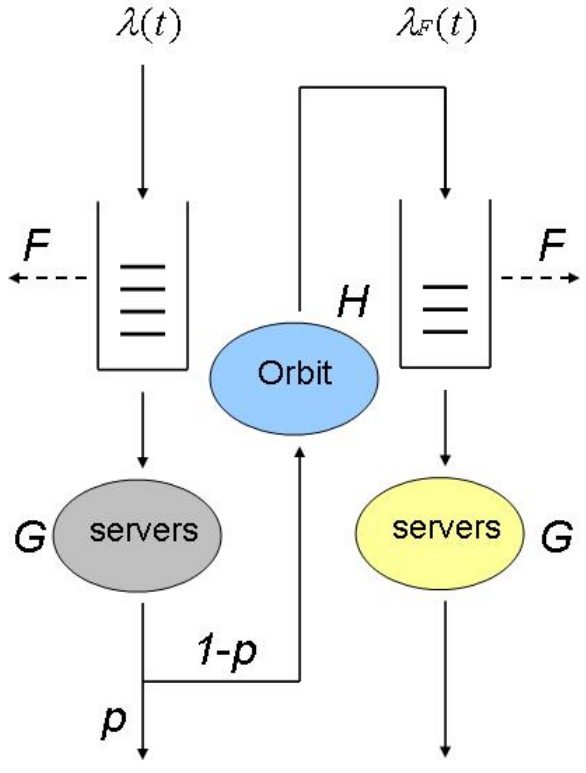
Staffing

Apply variation of DIS-MOL with 5 IS queues:



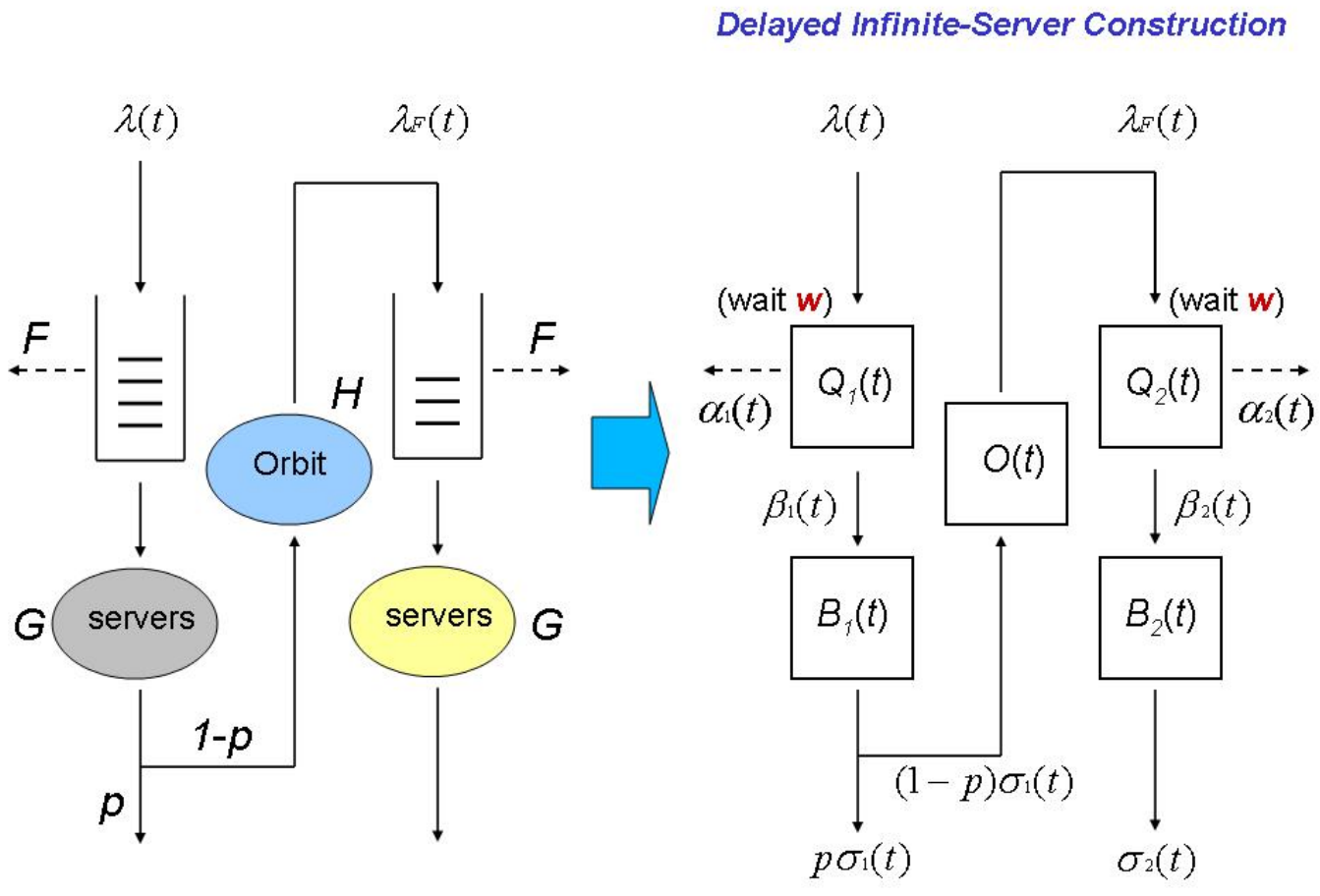
Staffing

Apply variation of DIS-MOL with 5 IS queues:



Staffing

Apply variation of DIS-MOL with 5 IS queues:



Flow Rate Formulas

- $\alpha_1(t) = E[\lambda(t-T)1_{\{T<w\}}]$
- $\alpha_2(t) = E[\lambda_F(t-T)1_{\{T<w\}}]$
- $\beta_1(t) = \lambda(t-w)\bar{F}(w)$
- $\beta_2(t) = \lambda_F(t-w)\bar{F}(w)$
- $\sigma_1(t) = \bar{F}(w)E[\lambda(t-w-S)]$
- $\sigma_2(t) = \bar{F}(w)E[\lambda_F(t-w-S)]$
- $\lambda_F(t) = (1-p)\sigma_1(t)$

The Five Expected Queue Lengths

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

The Five Expected Queue Lengths

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

$\Rightarrow$ **The Offered Load is** $m(t) \equiv E[B_1(t)] + E[B_2(t)]$

The Five Expected Queue Lengths

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

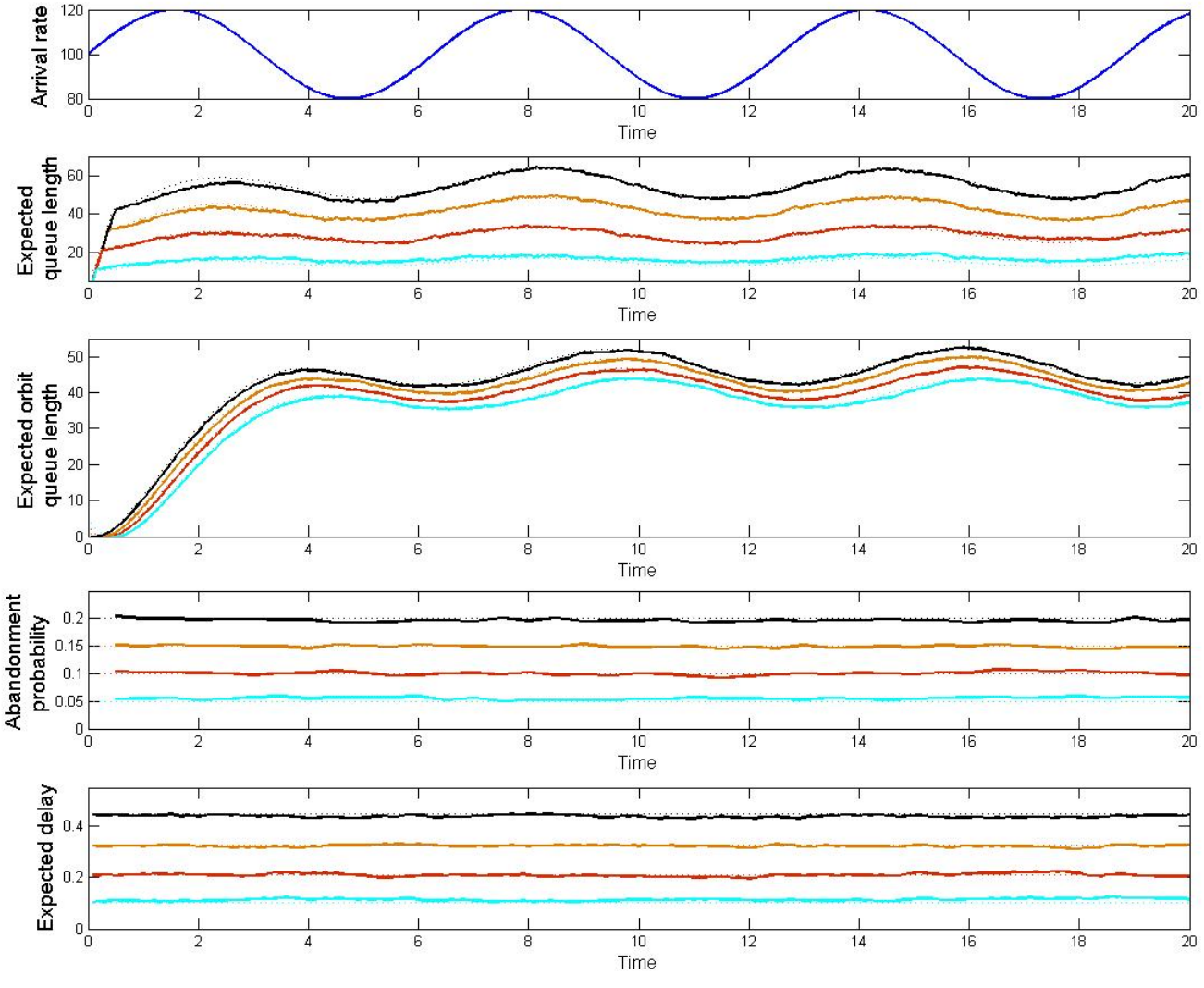
$\Rightarrow$ **The Offered Load is** $m(t) \equiv E[B_1(t)] + E[B_2(t)]$

$\Rightarrow$ **Apply MOL with OL** $m(t)$

Variation of the Previous Example

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

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


Arrival rate
120
100
80
0
2
4
6
8
10
12
14
16
18
20
Time
Expected queue length
60
40
20
Time
Expected orbit queue length
50
40
30
20
10
0
Time
Abandonment probability
0.2
0.15
0.1
0.05
0
Time
Expected delay
0.4
0.2
0
Time