Lecture 6: Offered Load Analysis

IEOR 4615: Service Engineering

Professor Whitt

February 10, 2015

What is the Problem?

- What ***capacity*** is needed in a service system?
- In order to meet uncertain exogenous demand
- Main Example: ***Staffing***, i.e., how many serivice representatives?

Offered Load Analysis

- ***Estimate the capacity needed*** to meet uncertain exogenous demand by determining the capacity that would be used if there were no limit on its availability.

NOTE: The demand could be ***defined*** directly as the offered load.

Definitions

- The ***Stochastic Offered Load*** (**SOL**): the *random* amount of capacity needed to meet uncertain exogenous demand if there were no limit on its availability.
- ***Offered Load*** (**OL**): the *expected value of the SOL.*

When is offered load analysis interesting?

1. When it is not so obvious what the offered load should be.

2. When you can afterwards apply the offered load in a non-obvious way to do a more refined analysis.

Staffing in a Service System (e.g., Call Center)

- Capacity = Number of Service Representatives
- Model for **SOL** = ***infinite-server queue***
  - Random arrivals of service requests
  - Each of random duration
- OL = expected number of busy servers

Three Stationary Concepts Coincide

- **Offered Load**
  - Expected Stochastic Offered Load
- Steady-state mean in **infinite-server model**
  - Expected Number of Busy Servers
- L in $L = \lambda W$
  - Where λ is the arrival rate
  - W is the expected required work per customer

First, a ***simple routine example***:

Business Case: H&S Schlock

Service Center to Help Prepare tax Returns

- How many service representatives are needed?
- arrival rate = 100 per hour
- expected service time = 1 hour

How many service representatives are needed?

- arrival rate = 100 per hour
- expected service time = 1 hour
- ***offered load:*** $\boldsymbol{m} = 100 \times 1 = 100$
- ***Square Root Staffing****: $\boldsymbol{s} = 100 + (100)^{1/2} = 110$

*Based on assuming an $M/GI/\infty$ model with Poisson arrivals; number is system is then Poisson, so that variance = mean; hence adding one standard deviation for slack

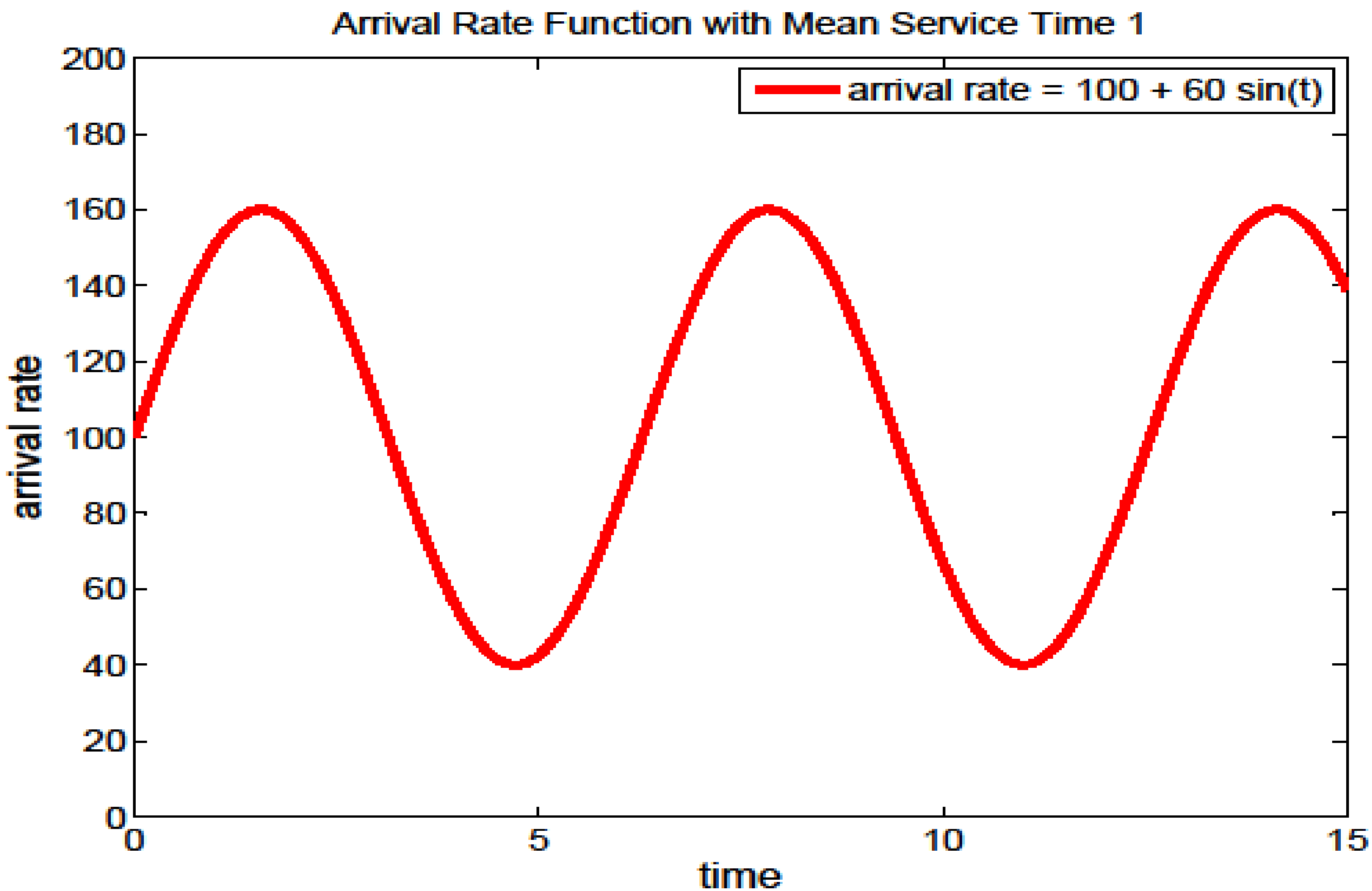
- ***Real system***: **Erlang-A (M/M/s + M) model**
- expected patience time = 2
- ***Performance***:
  - $P(Wait > 0) = .19$,
  - $P(Wait > .05) = .09$,
  - $P(Ab) = .006$,

The offered load analysis works, but we can easily do a more precise analysis with an Erlang-A model solver.

Offered Load Analysis for Staffing: Harder Case

Queueing Models with *Time-Varying Parameters*

A More Challenging Example: Time-Varying Arrival Rate



Three Time-Varying Concepts Coincide

- **Time-Varying Offered Load (TVOL)**
  - Expected Stochastic Offered Load
- Time-varying mean in $M_t/GI/\infty$
  - Expected Number of Busy Servers
  - Infinite-Server model with time-varying arrival rate
- $L(t)$ in ***Time-Varying Little's Law***
  - See second paper by Kim and Whitt (2013).

Generalize the ***simple routine example***:
Business Case: H&S Schlock
Service Center to Help Prepare tax Returns

- How many service representatives are needed?
- arrival rate = 100 per hour
- expected service time = 1 hour

Time-Varying Arrival Rate

- Long-run average arrival rate = 100 per hour
- Now $\lambda(t) = 100 + 60 \sin(t)$ (**new!!**)
- expected service time = 1 hour
- Want to stabilize performance at similar level
- $P(Wait > 0) \leq 0.20$
- How many service representatives are needed at each time now?

The Pointwise Stationary Approximation (PSA)

Approximate the Time-Varying Offered Load by

$$m_{PSA}(t) = \lambda(t)\, E[S] = \lambda(t),$$

where

$\lambda(t) = 100 + 60 \sin(t)$ time-varying arrival rate

$E[S] = 1$ expected service time

Time-Varying Offered Load (TVOL)

- Use $M_t/GI/\infty$ infinite-server model
- with nonhomogeneous Poisson arrival process
- having time-varying arrival rate $\lambda(t)$
- Use TVOL (from 1993 Physics paper)
- $m(t) = E[L(t)] = E[\lambda(t - S_e)]E[S]$, where
- $L(t)$ = number of busy servers at time t
- S = service time
- S_e = stationary excess of service time
- $P(S_e \leq x) = (1/\mathrm{E}[S]) \int_0^x P(S>u)\, du$

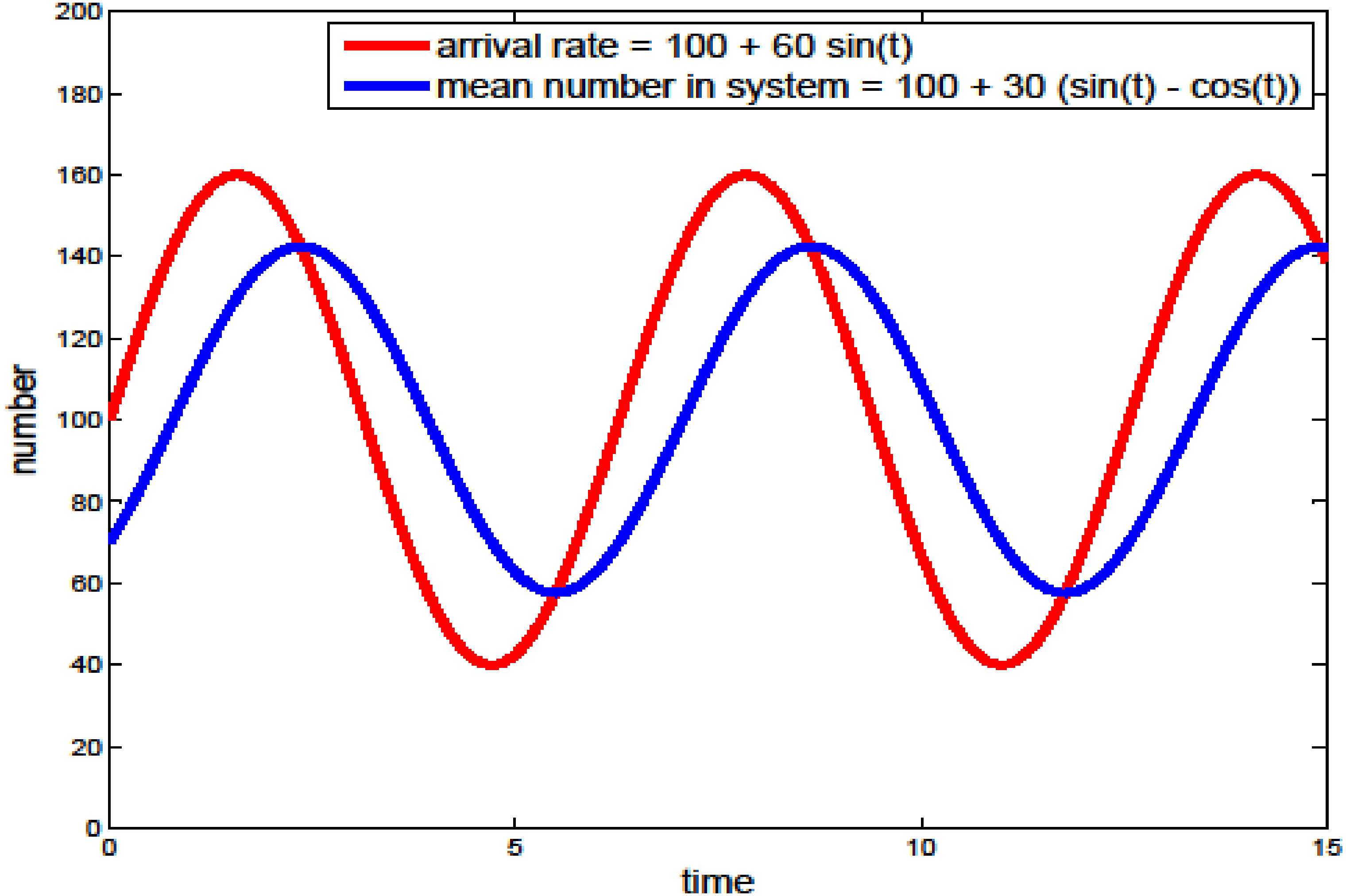
Explicit Formula for the Infinite-Server Mean

- Now $\lambda(t) = 100 + b \sin(ct)$, $b = 60$ and $c = 1$
- expected service time = 1 hour
- *Explicit formula for the mean in this case!!*
- $m(t) = 100 + (b/(1+c^2))\,(\sin(ct) - c\cos(ct))$

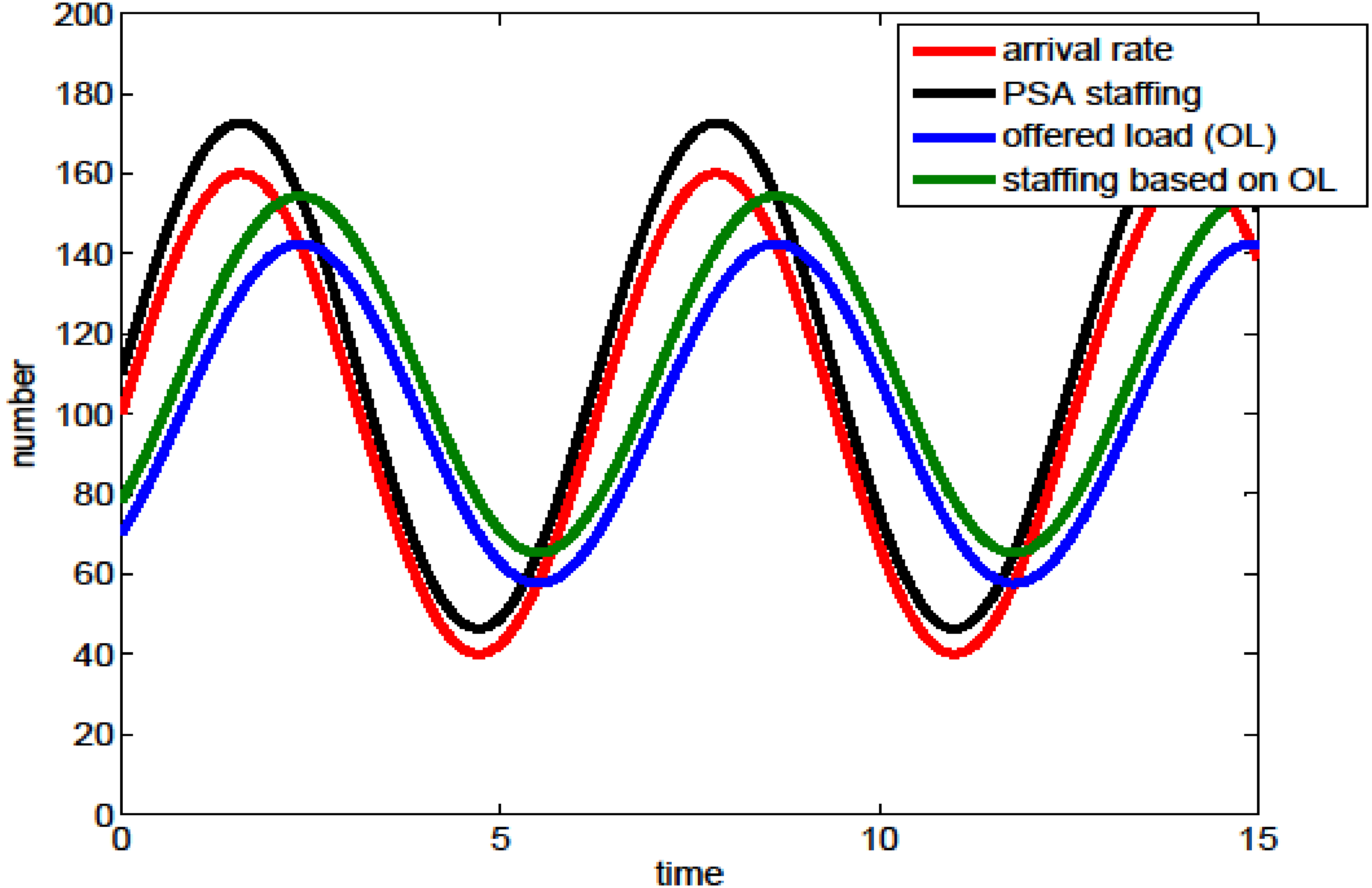
  $= 100 + 30(\sin(t) - \cos(t))$

- Eick, S. G., W. A. Massey, W. Whitt. 1993. $M_t/G/\infty$ queues with sinusoidal arrival rates. Management Sci. 39 241–252

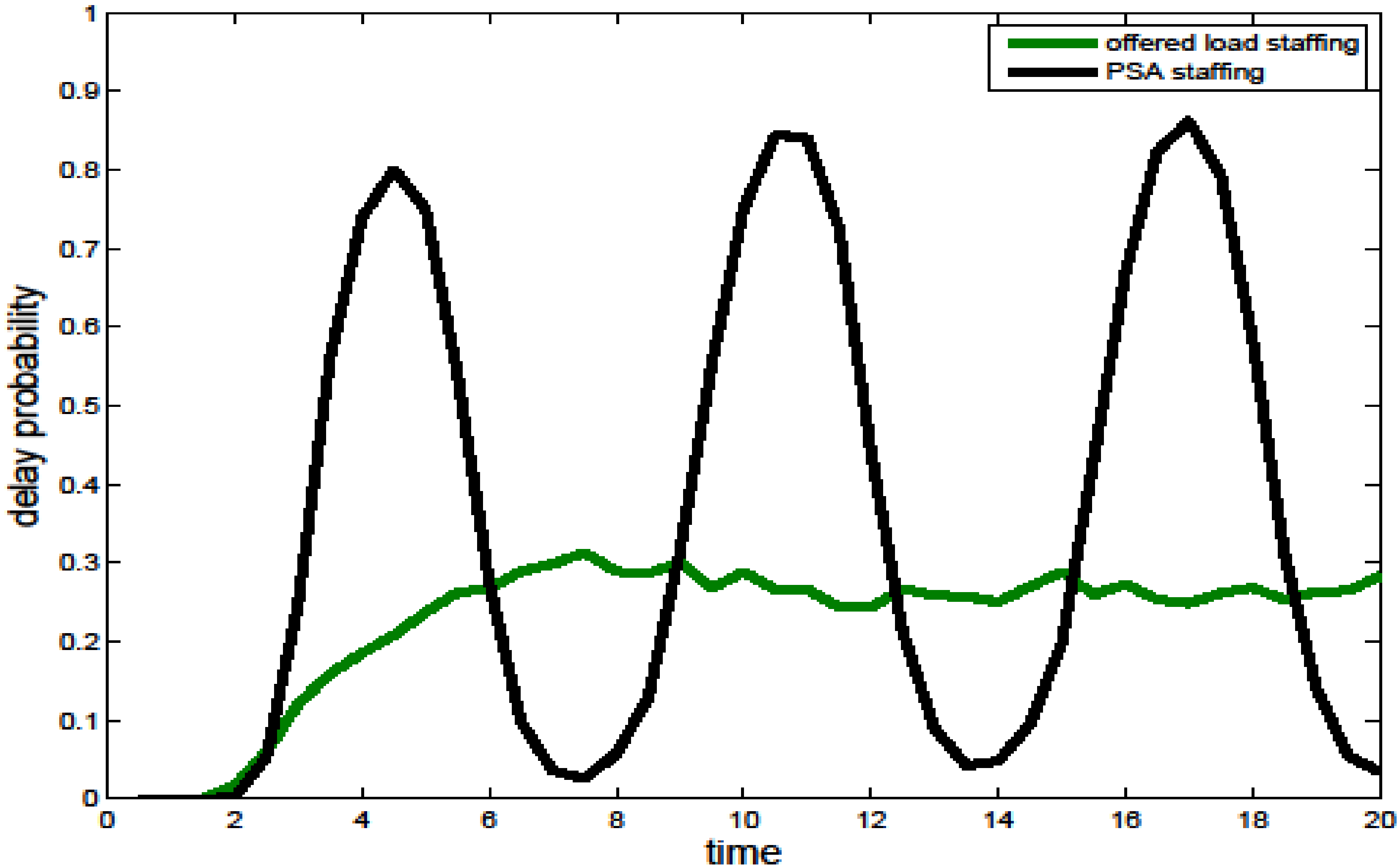
Arrival Rate and Offered Load (for $M_t/M/\infty$ model)



Staffing by PSA and TVOL



Simulation Comparison: PSA versus TVOL



Theorem from 1993 Physics paper

Theorem 1. *For each t, $Q(t)$ has a Poisson distribution with mean*

$$m(t) = E\left[\int_{t-S}^{t} \lambda(u)\,du\right] = E[\lambda(t - S_e)]E[S]. \qquad (3)$$

The departure process is a Poisson process with time-dependent rate function δ, where

$$\delta(t) = E[\lambda(t - S)]. \qquad (4)$$

For each t, $Q(t)$ is independent of the departure process in the interval $(-\infty, t]$.

Why?

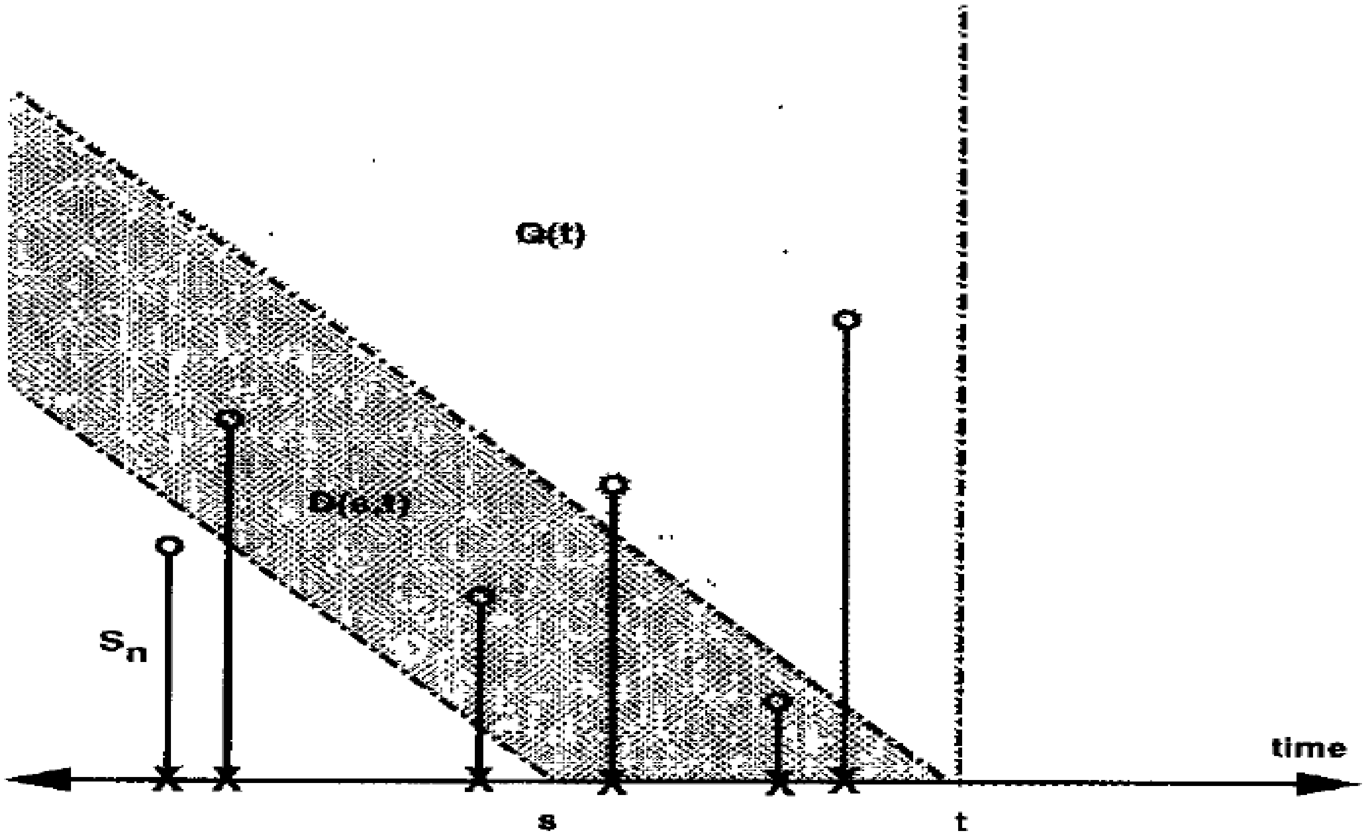


Figure 1. A possible realization of the Poisson random measure for Theorem 1; the random variables $Q(t)$ and $D(s, t)$ count the number of points in the designated subset.

Derivation of the mean number of busy servers in the $M_t/GI/\infty$ Model

$$\begin{aligned}
m(t) &\equiv E[Q(t)] = \int_{-\infty}^{t} \left(\int_{t-s}^{\infty} \lambda(s) g(z)\, dz \right) ds \\
&= \int_{-\infty}^{t} \lambda(s) G^c(t-s)\, ds \quad \text{(integrating over } z\text{)} \\
&= \int_{0}^{\infty} \lambda(t-s) G^c(s)\, ds \quad \text{(change of variables)} \\
&= \int_{0}^{\infty} \lambda(t-s) ES g_e(s)\, ds \quad (G^c(s) = E[S] g_e(s)) \\
&= E[\lambda(t-S_e)]E[S] \quad (S_e \text{ has pdf } g_e(s)).
\end{aligned}$$

(Service time S has pdf g, cdf G and $G^c(x) \equiv 1 - G(x)$.
Variable S_e has pdf $g_e(x) \equiv G^c(x)/E[S]$.)

More Complex Offered Load Models

The base model for the TVOL is the $M_t/GI/\infty$ infinite-server model, but there are other possibilities:

- Service over several disjoint time intervals, as in web chat
- Network of queues (Massey & W^2 93, McCalla & W^2 02, Yom-Tov & Mandelbaum 11)
- Service provided over space, as in mobile communications (Massey & W^2 94, Leung, Massey & W^2 94)
- The capacity used might be non-integer and time-varying, as in bandwidth usage in communication networks (Duffield, Massey & W^2 01)

More Accurate Staffing and Performance Prediction: The **Modified Offered Load** (MOL) **Approximation**

- Use steady-state performance of corresponding stationary model with capacity constraints and other details, e.g., customer abandonment, but in a nonstationary way.

- At time t, make the stationary offered load agree with $m(t)$ by letting
  - $\lambda_{MOL}(t) = m(t)/E[S]$

- where $m(t) = E[L(t)] = E[\lambda(t - S_e)]E[S]$ is TVOL, as before
  - -based on $M_t/GI/\infty$ model
  - -having time-varying arrival rate $\lambda(t)$

- **Staffing**: Let $s(t)$ = maximum s such that $P(Wait(t) > 0) \leq 0.2$, where $Wait(t)$ is steady-state wait for the stationary model at time t.

References
Offered Load Analysis for Staffing

1. Eick, S. G., W. A. Massey, W. Whitt. 1993a. **The physics of the $M_t/G/\infty$ queue**. Oper. Res. 41 731–742.
2. Eick, S. G., W. A. Massey, W. Whitt. 1993b. $M_t/G/\infty$ queues with sinusoidal arrival rates. Management Sci. 39 241–252.
3. Jennings, O. B., A. Mandelbaum, W. A. Massey, W. Whitt. 1996. **Server staffing to meet time-varying demand**. Management Sci. 42 1383–1394.
4. Massey, W. A. 2005. The analysis of queues with time-varying rates for telecommunication models. Telecommunication Systems 21:2–4, 173–204.
5. Green, L. V., P. J. Kolesar, W. Whitt. 2007. Coping with time-varying demand when setting staffing requirements for a service system. Production and Operations Management 16 13–29.
6. Whitt, W. 2013. **Offered Load Analysis for Staffing**. Manufacturing and Service Operations Management 15 166-69. (and e-companion)

References
The Pointwise Stationary Approximation (PSA)

1. Green, L. V., P. J. Kolesar. 1991. The pointwise stationary approximation. Management Sci. 37 84–97.
2. Whitt, W. 1991. The pointwise stationary approximation for Mt/Mt/s queues is asymptotically correct. Management Sci. 7 307–314.
3. Jennings, O. B., A. Mandelbaum, W. A. Massey, W. Whitt. 1996. Server staffing to meet time-varying demand. Management Sci. 42 1383–1394.
4. Massey, W. A., W. Whitt. 1998. Uniform acceleration expansions for Markov chains with time-varying rates**.** Annals of Applied Probability, 8 1130-1155.
5. Green, L. V., P. J. Kolesar, W. Whitt. 2007. Coping with time-varying demand when setting staffing requirements for a service system. Production and Operations Management 16 13–29.

References
Other Offered Load Models

1. Duffield, N. G., W. A. Massey, W. Whitt. 2001. A nonstationary offered-load model for packet networks. Telecommunication Systems 13(3-4) 271–296.
2. Leung, K. K., W. A. Massey, W. Whitt. 1994. Traffic models for wireless communication networks. IEEE Journal on Selected Areas in Communication 12(8) 1353–1364.
3. Massey, W. A. 2005. The analysis of queues with time-varying rates for telecommunication models. Telecommunication Systems 21:2–4, 173–204.
4. Massey, W. A., W. Whitt. 1993. Networks of infinite-server queues with nonstationary Poisson input. Queueing Systems 13(1) 183–250.
5. Massey, W. A., W. Whitt. 1994b. A stochastic model to capture space and time dynamics in wireless communication systems. Probability in the Engineering and Informational Science 8 541–569.
6. McCalla, C., W. Whitt. 2002. A time-dependent queueing-network model to describe life-cycle dynamics of private-line telecommunication services. Telecommunication Systems 17 9–38.
7. Yom-Tov, G., A. Mandelbaum. 2010. The Erlang-*R* queue: time-varying QED queues with re-entrant customers in support of healthcare staffing. Working paper, the Technion, Israel

References
The Modified Offered Load Approximation

1. Jagerman, D. L. 1975. Nonstationary blocking in telephone traffic. Bell System Tech. J. 54 625–661.
2. Massey, W. A., W. Whitt. 1994a. An analysis of the modified offered load approximation for the nonstationary Erlang loss model. Annals of Applied Probability 4 1145–1160.
3. Jennings, O. B., A. Mandelbaum, W. A. Massey, W. Whitt. 1996. Server staffing to meet time-varying demand. Management Sci. 42 1383–1394.
4. Massey, W. A., W. Whitt. 1997. Peak congestion in multi-server service systems with slowly varying arrival rates. Queueing Systems 25 157–172.
5. Green, L. V., P. J. Kolesar, W. Whitt. 2007. Coping with time-varying demand when setting staffing requirements for a service system. Production and Operations Management 16 13–29.
6. Feldman, Z., A. Mandelbaum, W. A. Massey, W. Whitt. 2008. Staffing of time-varying queues to achieve time-stable performance. Management Sci. 54(2) 324–338.
7. Liu, Y., W. Whitt. 2012. Stabilizing customer abandonment in many-server queues with time-varying arrivals. Operations Research 60 1551-1560.
8. Yom-Tov, G., A. Mandelbaum. 2010. The Erlang-*R* queue: time-varying QED queues with re-entrant customers in support of healthcare staffing. Working paper, the Technion, Israel