Delay Announcements

Predicting Queueing Delays for Delay Announcements

IEOR 4615, Service Engineering, Professor Whitt

Lecture 21, April 21, 2015

OUTLINE

(References at end)

The Purpose of Delay Announcements

Why might a service system manager want to **tell each customer an estimate of the delay that customer will experience?**

Improve customer satisfaction.

- **Propositions about the Psychology of Waiting**, Maister (1985), Lec. 1
  - **Uncertain Waits** feel longer than known finite waits.
  - **Anxiety** makes waits seem longer.
  - **Unexplained Waits** feel longer than explained waits.
- **Improve performance for the customers who are served.**
  - **By inducing some customers to balk or abandon earlier and then retry later** when the system is more lightly loaded.

Making Delay Announcements: Some Questions

- **What** delay to predict?
  - delay **before entering service** (assuming will not abandon)
  - response time, i.e., delay **until completing service**
- **When** to predict and announce?
  - **immediately upon arrival**
  - **throughout time in queue** (continuously or periodically)
- **What to announce/report?** Customer ability to process information?
  - **single number w, may be different for each customer**
  - **full distribution** or partial summary, e.g., mean and variance.
  - **past prediction accuracy** average error or variance.
  - **explanation** of cause of delay.

Examples of Possible Delay Announcements

- **When the delay will be short:**

  "We should be able to serve you soon. The last customer to enter service waited only one minute." (Aims to encourage customer to wait.)

- **When the delay will be long:**

  "We are currently experiencing unexpected high demand; the last customer to enter service had to wait w minutes before beginning service. We will do our best to serve you without excessive delay, but you might want to try again later." (Aims to encourage customer to balk or abandon sooner and then retry later. By doing so, aims to provide better service to the customers who are served.)

Problem for Today: Delay Prediction

- Assume the **standard multi-server queueing system** with random arrivals, service times and patience times, **given system history up to arrival time.**
- **Given that we will announce, immediately upon arrival, our best estimate of a single number w to each customer who has to wait before starting service, how should we predict w and how accurate is our prediction?**
- Two Approaches: **simplicity** versus **complexity**.

Information to Exploit

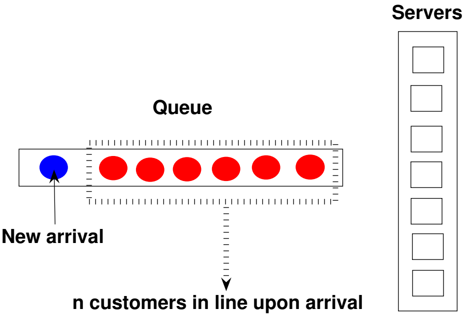
- **full model: distributions and parameters**
  - no information: steady-state $E[Wait|Wait > 0]$
- **delay history** (major focus)
- **queue length**
  - **with and without customer abandonment**
- queue length and elapsed service times
- queue length, customer classes and elapsed service times

Alternative Real-Time Delay Predictors

Assume the **standard multi-server queueing system** with random arrivals, service times and patience times, **given system history up to arrival time.**

- Mean Steady-State Delay for the model (but model error?)
- Standard Simple Queue-Length-Based Delay Predictor QL_s
- Delay Experienced by the Last to Enter Service (LES)
- Elapsed Delay of the Customer at the Head of Line (HOL)
- Refined Predictors
  - calculate $E[Wait|Q(t) = n]$ or $E[Wait|\text{entire history at } t]$
  - fluid model refinements

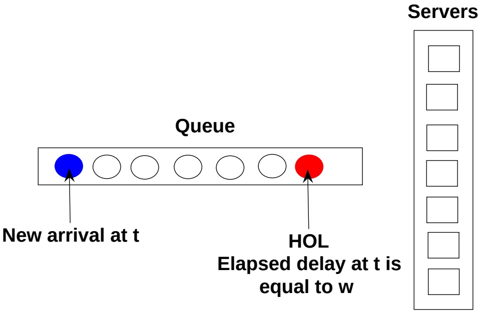
The Standard Simple Queue-Length-Based Delay Predictor



- s = number of agents, and μ^{-1} = mean service time

$$\boxed{\theta_{QL_s}(n) \equiv (n+1) \times \frac{\mu^{-1}}{s}}$$

The Head-of-Line (HOL) Delay Predictor



- w = elapsed delay of HOL customer (similar to LES delay)

$$\boxed{\theta_{HOL}(w) \equiv w}$$

Actual Random Delays After Prediction

- $W_{HOL}(w)$: distributed as the delay of a new arriving customer ***given*** that:
  - (i) there is a customer at the HOL upon arrival (non-restrictive)
  - (ii) elapsed delay of HOL customer is equal to w

- $W_Q(n)$: distributed as the delay of a new arriving customer ***given*** that:
  - (i) the customer has to wait
  - (ii) the customer finds n customers in line upon arrival

For example, with HOL, we announce $\theta_{HOL}(w) \equiv w$.

w is a single-number prediction of the random variable $W_{HOL}(w)$.

Quantifying The Accuracy of the Predictors

***Mean Squared Error* (MSE)**

$$MSE(\theta_{QL_s}(n)) = E[(W_Q(n) - \theta_{QL_s}(n))^2]$$

$$E[MSE(\theta_{QL_s}(Q^w_\infty))] = \sum_{n=0}^{\infty} MSE(\theta_{QL_s}(n))P[Q^w_\infty = n]$$

- Q^w_∞ has the conditional distribution of the steady-state QL upon arrival given that the customer must wait.

How to Evaluate Predictors with Simulation

Simulation Estimate of MSE: ***Average Squared Error (ASE)***

$$ASE \equiv \frac{1}{k}\sum_{i=1}^{k}(p_i - d_i)^2 \quad (k = \text{ sample size})$$

- p_i = predicted delay for customer i ($p_i > 0$)
- d_i = actual delay (or potential delay with abandonments)

Root Relative Average Squared Error (RASE)

$$RASE \equiv \frac{\sqrt{ASE}}{\frac{1}{k}\sum_{i=1}^{k} d_i}$$

Different Cases (Models) to Consider

Assume the **standard multi-server queueing system** with random arrivals, service times and patience times, **given system history up to arrival time.**

- $A/M/s$ (stationary model without abandonment)
- **abandonment:** $A/M/s+M$
- **non-exponential abandonment:** $A/M/s+GI$
- **time-varying arrivals:** $M_t/M/s$ and $M_t/M/s+M$

QL_s in the GI/M/s Model (or A/M/s)

$$W_Q(n) = \sum_{i=1}^{n+1} V_i$$

where V_i i.i.d. exponential with mean $(s\mu)^{-1}$

$$E[W_Q(n)] = \sum_{i=1}^{n+1} E[V_i] = \sum_{i=1}^{n+1} \frac{1}{s\mu} = \frac{n+1}{s\mu} \equiv \theta_{QL_s}(n)$$

$$MSE(\theta_{QL_s}(n)) = Var[W_Q(n)] = \sum_{i=1}^{n+1} Var[V_i] = \sum_{i=1}^{n+1} \frac{1}{s^2\mu^2} = \frac{n+1}{s^2\mu^2}$$

$\theta_{QL_s}(n)$ **is an unbiased estimator. It minimizes the MSE!**

Compare QL$_s$ to Steady-State Mean in the GI/M/s Model

In steady state, $1 + Q|W > 0$ is geometric on $\{1, 2, \ldots\}$ with mean $1/(1-\rho)$ as in $M/M/1$ with service rate $s\mu$ and $W|W > 0$ is exponential with

$$E[W|W > 0] = \frac{1}{s\mu(1-\rho)} \quad \text{and} \quad Var[W|W > 0] = \frac{1}{s^2\mu^2(1-\rho)^2}$$

$$RSE(W|W > 0) \equiv \frac{\sqrt{Var[W|W > 0]}}{E[W|W > 0]} = 1$$

In contrast,

$$E[W_Q(n)] = \frac{n+1}{s\mu}, \quad Var[W_Q(n)] = \frac{n+1}{s^2\mu^2}$$

$$RMSE(\theta_{QL_s}(n)) = \frac{\sqrt{Var[W_Q(n)]}}{E[W_Q(n)]} = \frac{1}{\sqrt{n+1}} \approx \frac{1}{\sqrt{n}}$$

Compare QL$_s$ to Steady-State Mean in the GI/M/s Model

$$\begin{aligned} E[MSE(\theta_{QL_s}(Q^w_\infty))] &= \sum_{n=0}^{\infty} MSE(\theta_{QL_s}(n))P[Q^w_\infty = n] \\ &= \sum_{n=0}^{\infty} Var(\theta_{QL_s}(n))P(Q^w_\infty = n) = \sum_{n=0}^{\infty} \frac{n+1}{s^2\mu^2} P(Q^w_\infty = n) \\ &= \frac{1}{s^2\mu^2(1-\rho)} \end{aligned}$$

Hence, it is much better to use the information:

$$\frac{Var(W|W>0)}{E[Var(W_Q(Q_\infty))]} = \frac{1/s^2\mu^2(1-\rho)^2}{1/s^2\mu^2(1-\rho)} = \frac{1}{1-\rho}.$$

HOL in the $M/M/s$ Model

$$W_{HOL}(w) = \sum_{i=1}^{A(w)+2} V_i$$

where V_i i.i.d. exponential with mean $(s\mu)^{-1}$

$$\Downarrow$$

$$E[W_{HOL}(w)] = E\left[\sum_{i=1}^{A(w)+2} V_i\right] = E[A(w)+2]E[V] \neq w \equiv \theta_{HOL}(w)$$

$MSE(\theta_{HOL}(w))$ **depends on** $Var[A(w)]$**.**

Simulations for the $GI/M/s$ Model: Poisson Arrivals

In Tables: ASE's in units of 10^{-3} (RASE in %); $\rho = \lambda/s\mu$; $c_a^2 = Var/mean^2$.

$M/M/100$

ρ	QL_s	HOL	HOL/QL_s	$(c_a^2+1)/\rho$
0.98	5.03 (14%)	10.2 (20%)	**2.03**	**2.04**
0.95	2.04 (22%)	4.27 (32%)	**2.09**	**2.11**
0.93	1.44 (26%)	3.08 (39%)	**2.14**	**2.15**
0.90	0.994 (32%)	2.19 (47%)	**2.20**	**2.22**

Similar for other renewal arrival processes: ratio $\approx (c_a^2+1)/\rho$.

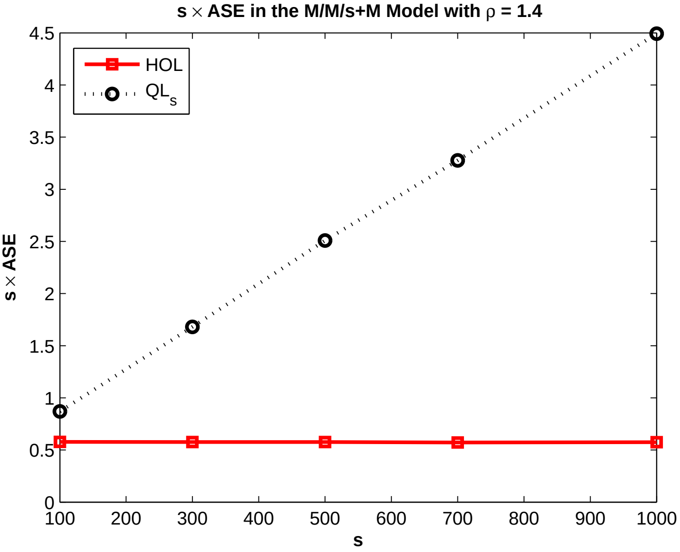
Simulations for the $GI/M/s$ Model: Deterministic Arrivals

In Tables: ASE's in units of 10^{-3} (RASE in %); $\rho = \lambda/s\mu$; $c_a^2 = Var/mean^2$.

$D/M/100$

ρ	QL_s	HOL	HOL/QL_s	$(c_a^2+1)/\rho$
0.98	2.48 (20%)	2.62 (21%)	**1.06**	**1.02**
0.95	1.01 (32%)	1.15 (34%)	**1.14**	**1.05**
0.93	0.725 (37%)	0.871 (41%)	**1.20**	**1.08**
0.90	0.519 (44%)	0.664 (50%)	**1.28**	**1.11**

Abandonments: Simulations for the $M/M/s+M$ Model



$W_Q(n)$ in the $GI/M/s + M$ Model

$W_Q(n)$: distributed as the ***potential*** delay of a new arriving customer ***given*** that:

(i) the customer has to wait

(ii) the customer finds n customers in line upon arrival

$$W_Q(n) = \sum_{i=0}^{n} X_i$$

where X_i independent exponential with mean $(s\mu + i\nu)^{-1}$

Markovian QL Predictor (QL_m)

- **The Markovian Queue-Length Predictor (QL_m)**

$$\theta_{QL_m}(n) = \sum_{i=0}^{n} \frac{1}{s\mu + i\nu}$$

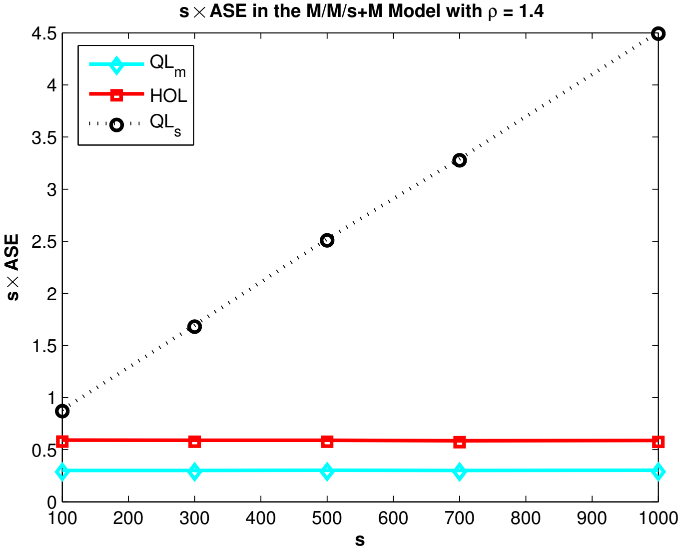
- **QL_m in the $GI/M/s+M$ Model**

$$\theta_{QL_m}(n) = \sum_{i=0}^{n} 1/(s\mu + i\nu) = E[W_Q(n)]$$

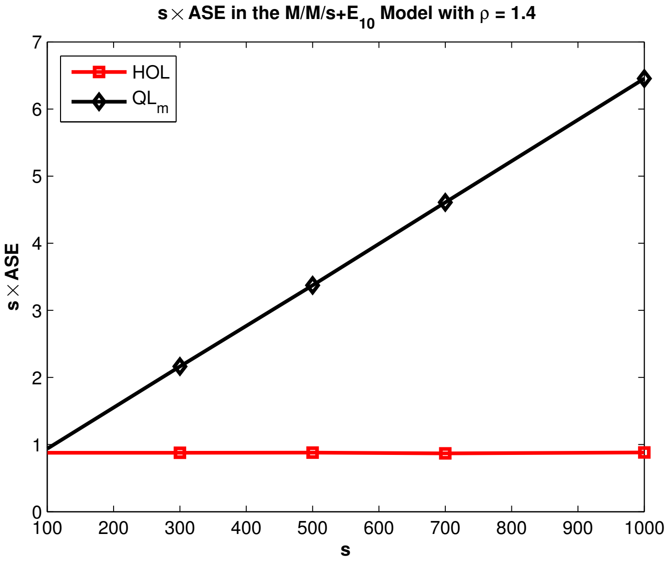
$$\Downarrow$$

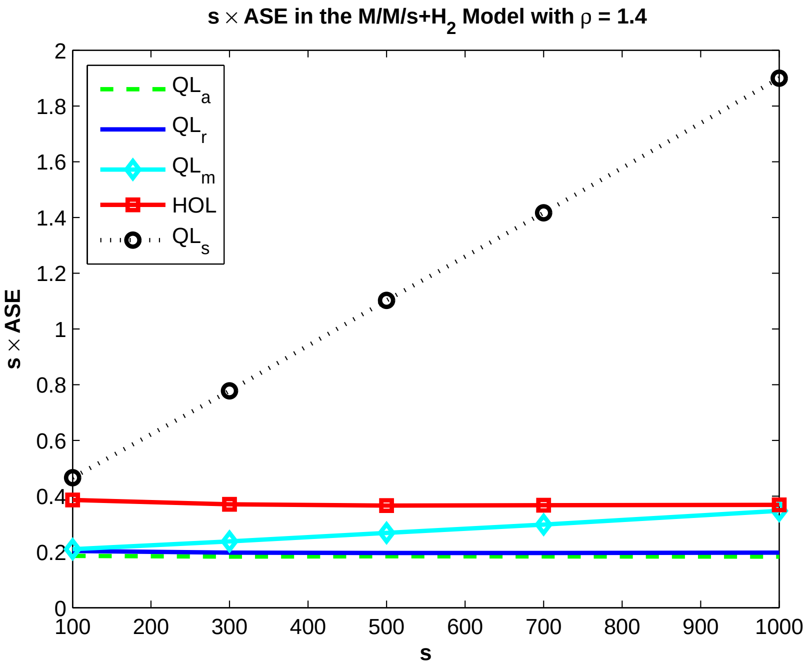
$\theta_{QL_m}(n)$ minimizes the MSE!

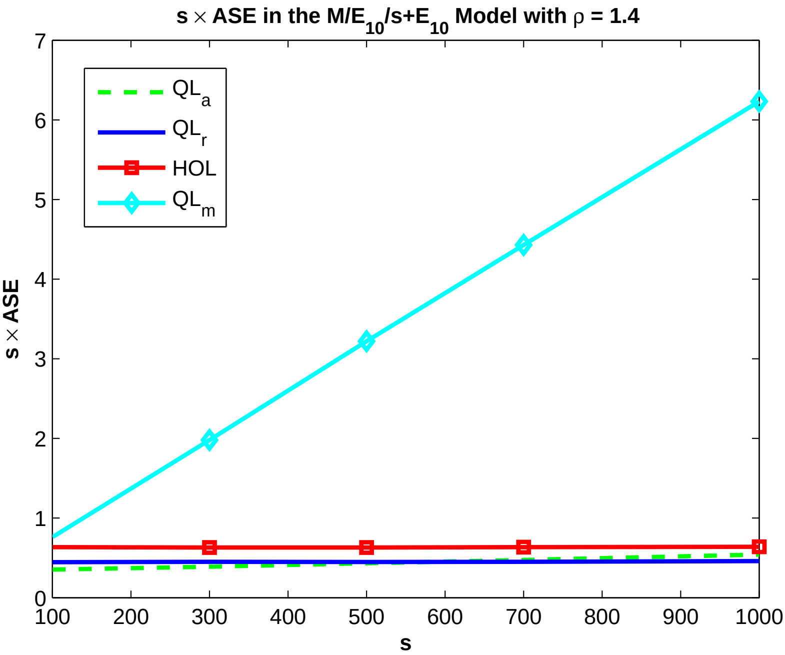
Refined QL Predictor for Same $M/M/s+M$ Model



But when the Abandonment Distribution is Not Exponential



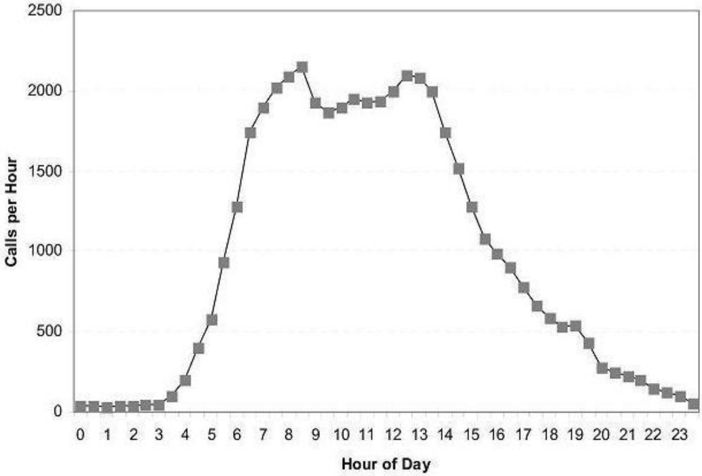

s × ASE in the M/M/s+H2 Model with ρ = 1.4
s × ASE
s
QLa
QLr
QLm
HOL
QLs
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
100
200
300
400
500
600
700
800
900
1000



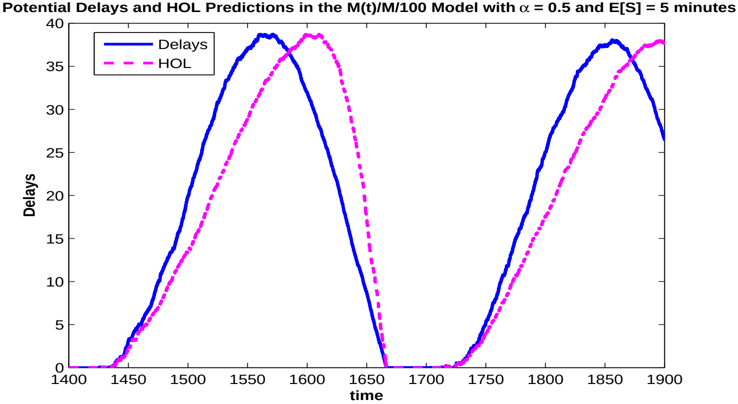
s × ASE in the M/E10/s+E10 Model with ρ = 1.4
QLa
QLr
HOL
QLm
s × ASE
s
0
1
2
3
4
5
6
7
100
200
300
400
500
600
700
800
900
1000

Time-Varying Arrival Rates

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

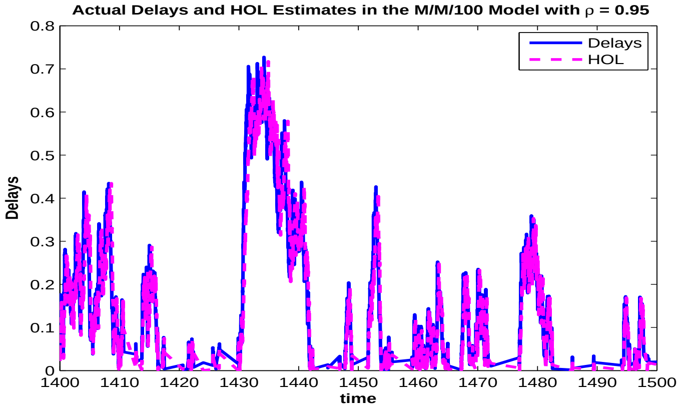


HOL Delay Prediction in the $M_t/M/100$ Model



- Arrival process: Nonhomogeneous Poisson with rate $\lambda(t)$
- **sinusoidal** $\lambda(t) = \bar{\lambda} + \alpha\bar{\lambda}\sin(\gamma t), \quad \rho = \bar{\lambda}/s\mu = 0.95$

HOL Delay Prediction With Constant Arrival Rate



- Arrival process: homogeneous Poisson with rate λ
- $\rho = \bar{\lambda}/s\mu = 0.95$

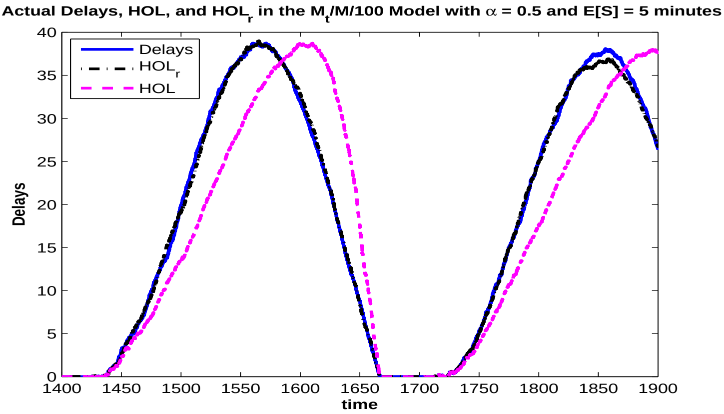
Problem: Time Lag in HOL Delay

- HOL delay was potential delay of arrival in the past.
- Use fluid model to create refined predictor.
- $v(t)$ potential delay of new arrival in fluid model at time t
- $w(t)$ HOL delay in the fluid model at time t

$$\theta_{HOL_r}(w, t) = \frac{v(t)}{w(t)} \times w,$$

where w is observed HOL delay at time t in actual system.

HOL_r Delay Prediction in the $M_t/M/100$ Model



- Arrival process: Nonhomogeneous Poisson with rate $\lambda(t)$
- $\lambda(t) = \bar{\lambda} + \alpha\bar{\lambda}\sin(\gamma t), \quad \rho = \bar{\lambda}/s\mu = 0.95$

SUMMARY

- We have shown how to predict delays to make delay announcements.
  1. The simple queue-length estimator $\theta_{QL_s}(n)$ is optimal for $G/M/s$, but HOL (and LES) is not too bad.
  2. $\theta_{QL_s}(n)$ can perform poorly with abandonments, while HOL is robust.
  3. For the $G/M/s+M$ model, a new Markovian estimator is optimal.
  4. But when the patience times are not exponential, it too can perform poorly. New refined estimators can do better. HOL remains robust.
  5. With time-varying arrivals, even HOL can perform poorly, but fluid models can be used to refine the HOL estimator.

The End

References, all but first with Rouba Ibrahim

- **Predicting Queueing Delays.** *Man. Sci.* 45 (1999) 870–888.
- **Real-Time Delay Estimation Based on Delay History.** *Manuf. Serv. Opns. Mgt.* 11 (2009) 397–415.
- **Real-Time Delay Estimation in Overloaded Multiserver Queues with Abandonment.** *Man. Sci.* 55 (2009) 1729–1742.
- **Real-Time Delay Estimation Based on Delay History in Many-Server Queues with Time-Varying Arrivals.** *Prod. Opns. Mgt.* 20 (2011) 654-667.
- **Wait-Time Predictors for Cust. Serv. Systems with Time-Varying Demand and Capacity.** *Oper. Res.* 59 (2011) 1106–1118.

More References

- M. Armony, N. Shimkin and WW. **The Impact of Delay Announcements in Many-Server Queues with Abandonment**. *Operations Res.* 57 (2009) 66–81. (See Lecture 21a in Courseworks.)
- WW. **Improving Service By Informing Customers About Anticipated Delays**. *Management Sci.* 45 (1999) 192–207.
- R. J. Batt and C, Terwiesch. **Waiting Patiently: An Empirical Study of Patient Abandonment in an Emergency Department** *Management Sci.* 61 (2015) 39–59.
- G. Allon and A. Bassamboo. **The Impact of Delaying the Delay Announcements.** *Oper. Res* 59 (2011) 1198–1210.

Even More References

- A. Senderovich1, M. Weidlich, A. Gal1 and A. Mandelbaum1 **Queue Mining Predicting Delays in Service Processes**. In *Advanced Information Systems Engineering*, eds. M. Jarke et al., Proceedings of 26th CAiSE Conference in Greece, 2014, eds. M. Jarke et al.