Making Delay Announcements

Performance Impact and Predicting Queueing Delays

Ward Whitt

With Mor Armony, Rouba Ibrahim and Nahum Shimkin

March 7, 2012

Last Class

1. The Overloaded $G/GI/s + GI$ Fluid Queue Model

   (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)

Today: Application of Fluid Models to Delay Prediction

1. The Performance Impact of Delay Announcements to New Arrivals

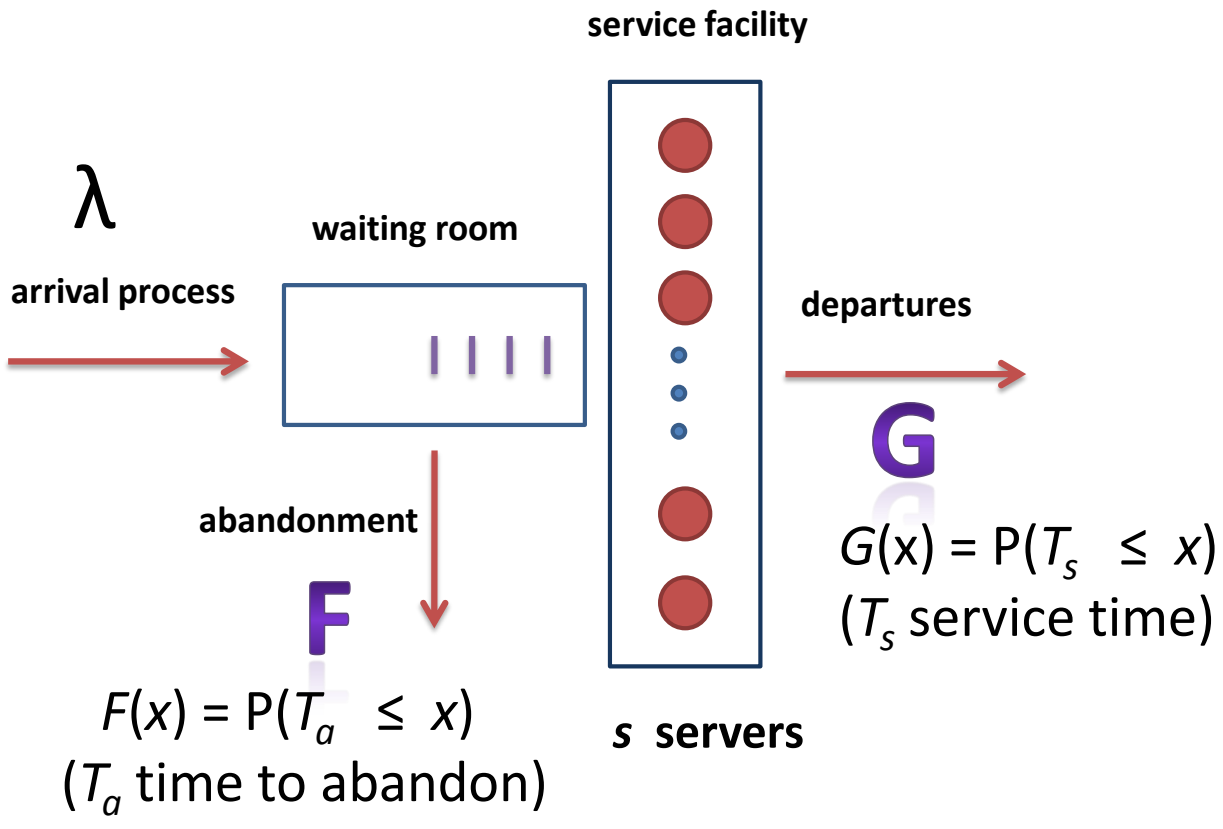
   (Mor Armony, Nahum Shimkin & W^2, *Operations Research*, 2009)

2. Real-Time Delay Predictors (including Time-Varying Arrivals)

   (Rouba Ibrahim & W^2, *Operations Research*, 2011)

1. Review: The G/GI/s+GI Fluid Model

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



Many-Server Heavy-Traffic (MSHT) Limit

Increasing Scale **Increasing Scale**

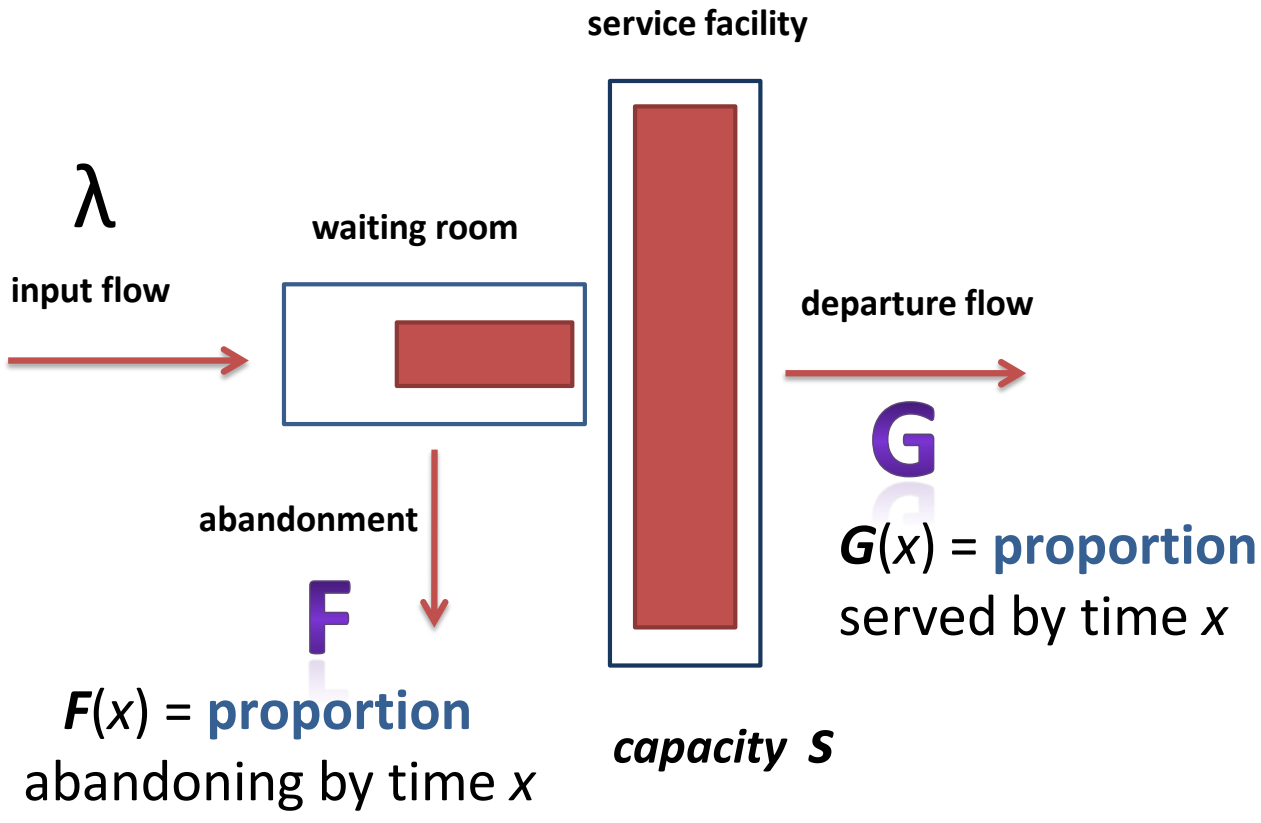
- a sequence of $G/GI/s+GI$ models indexed by n,
- arrival rate **grows**: $\lambda_n/n \to \lambda \quad \text{as} \quad n \to \infty$,

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

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

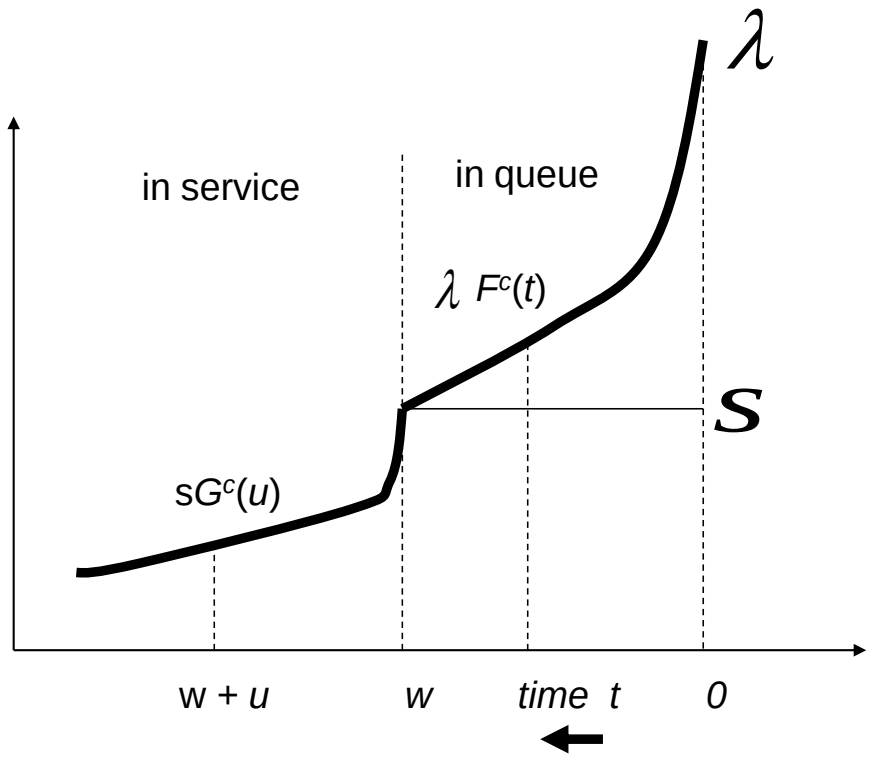
The G/GI/s+GI Fluid Model

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



The Overloaded Fluid Model in Steady State

fluid density arriving time t in the past

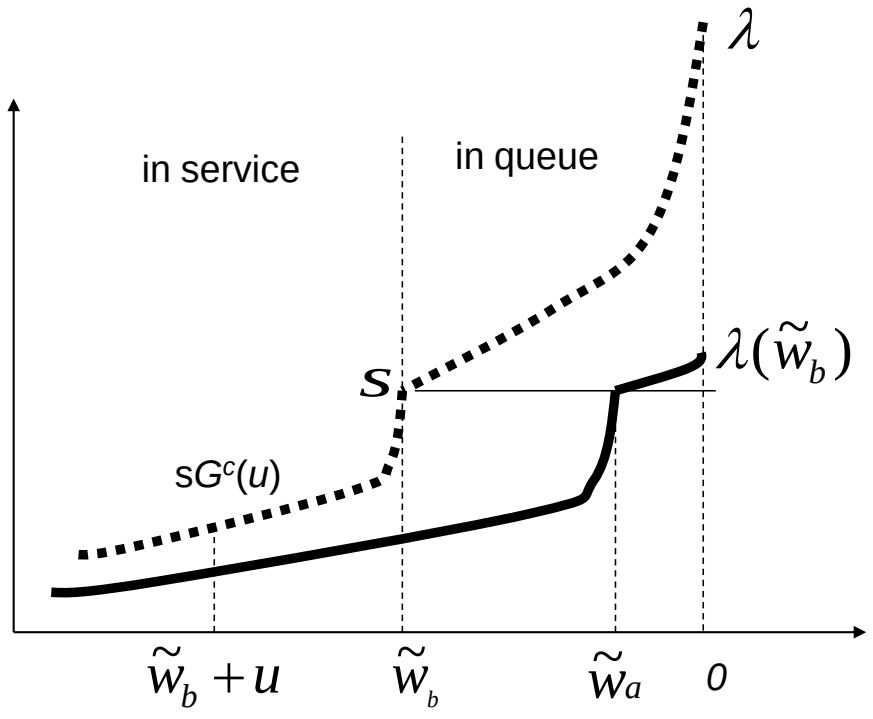


2. The Performance Impact of Delay Announcements

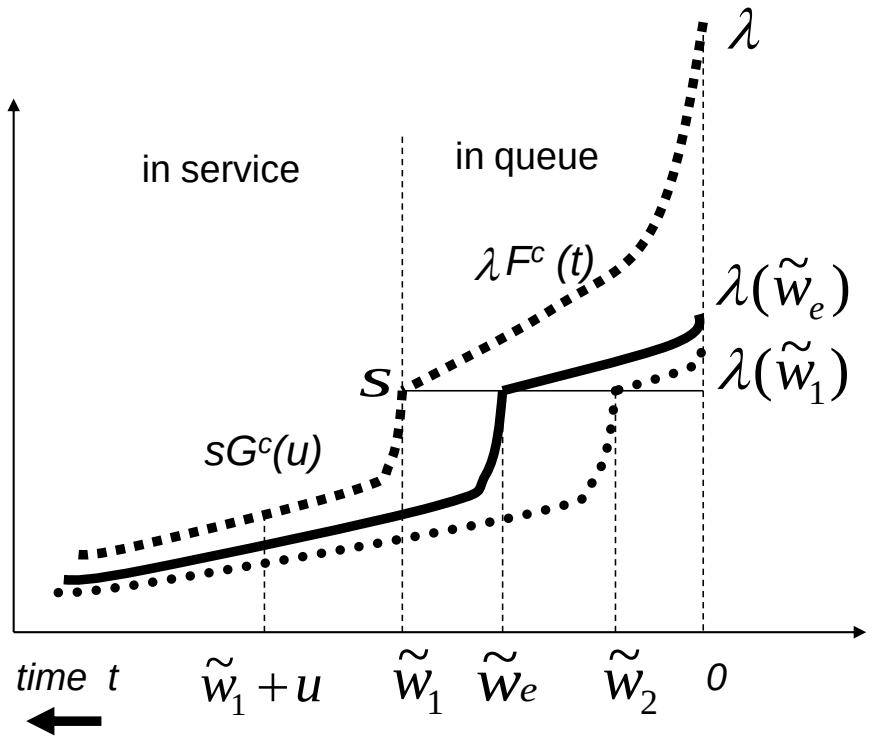
- make **delay announcement** (single-number w) to new arrivals
- helpful with invisible queues (e.g., waiting on the telephone)
- But we need to consider the **customer response**
- Assume: balk (leave immediately) with probability $B(w)$
- If join, abandon before time t with probability $F(t|w)$
- What to announce? Delay of Last to Enter Service (LES)
- What is the **performance impact** of the announcement?

Use Fluid Model to Study Performance Impact

Direct Response to Delay Announcement

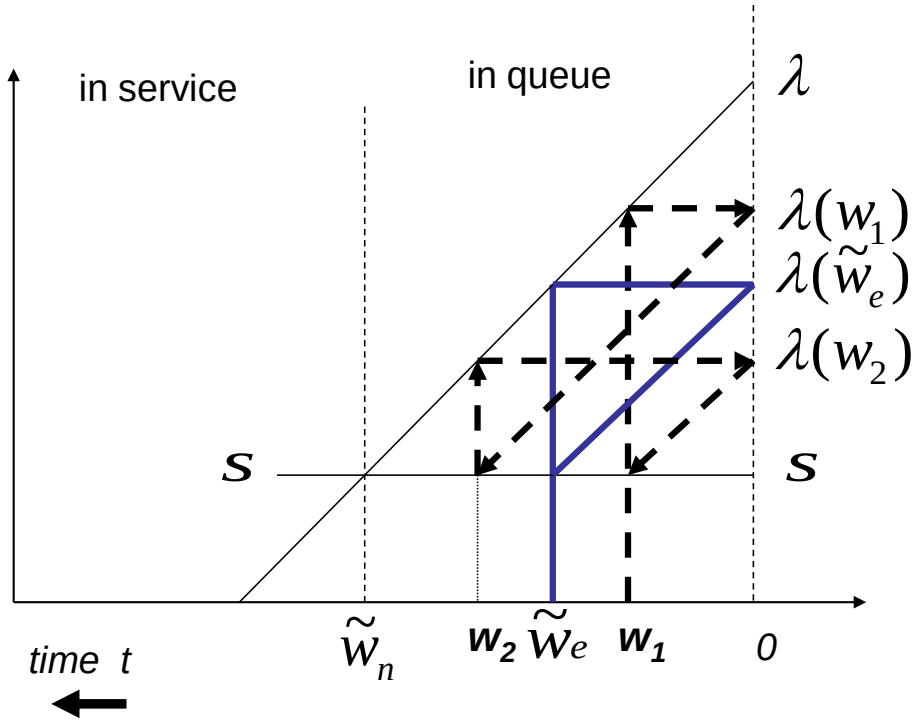


Equilibrium Delay



Use Fluid Model to Study Convergence

Cycling Around the Equilibrium Delay



Conclusions of Study, Armony et al. 2009

- Under general conditions, there exists a unique equilibrium delay in the fluid model.
- Direct iteration $w_{n+1} = d(w_n)$ as shown above can lead to oscillations.
- Damped iterations produce convergence: $w_{n+1} = pd(w_n) + (1-p)w_n$
- LES delay and fluid equilibrium delay both work well in simulations

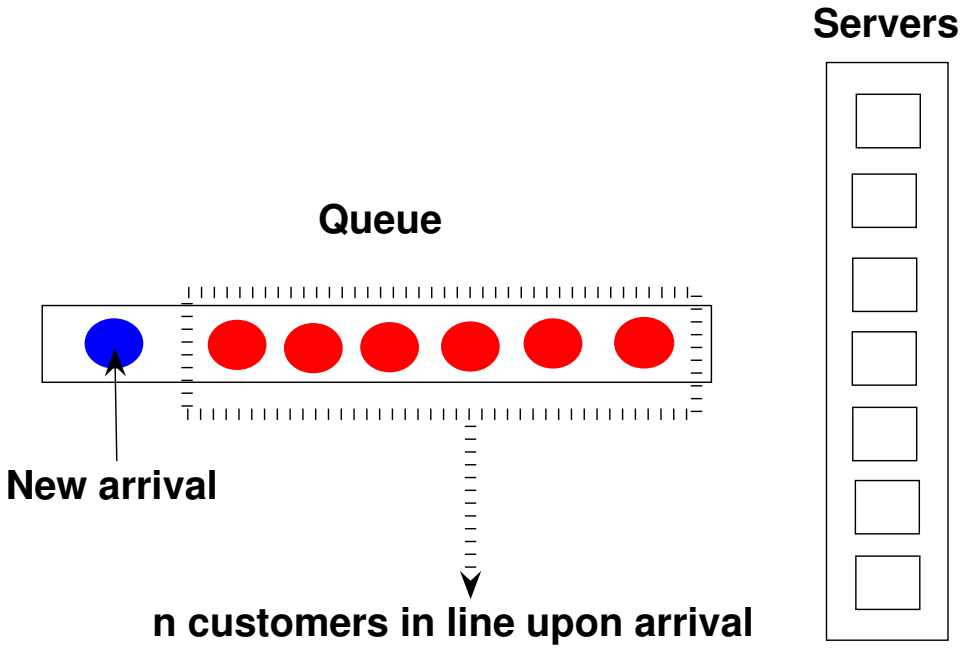
  But LES more robust, does not depend on the model.

3. Alternative Real-Time Delay Predictors

Now considered without customer response

- Delay Experienced by the Last to Enter Service (LES)
- Elapsed Delay of the Customer at the Head of Line (HOL)
- Standard Simple Queue-Length-Based Delay Predictor QL_s
- Use the **Fluid Model** to Develop Refined Predictors

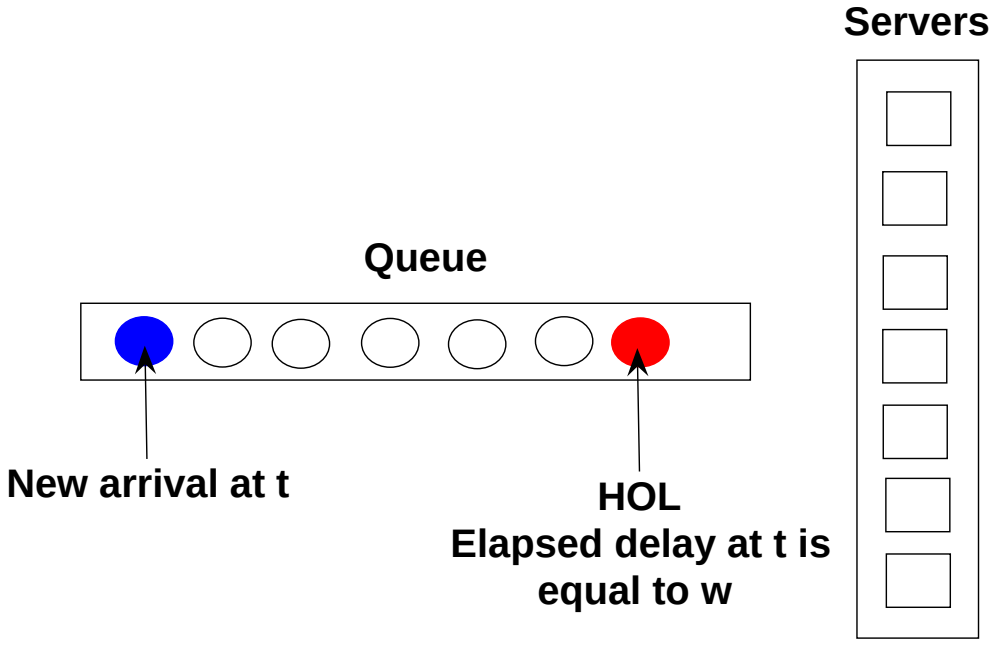
The Standard Simple Queue-Length-Based Delay Predictor



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

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

The Head-of-Line (HOL) Delay Predictor



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

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

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

$$\text{We announce } \theta_{HOL}(w) \equiv w$$

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

Quantifying The Accuracy of the Predictors

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

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

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

- 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)

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

QL_s in the GI/M/s Model

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

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

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

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

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

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 $(s\mu)^{-1}$

$$\Downarrow$$

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

$$Var[W_{HOL}(w)] = E[A(w) + 2Var[V] + Var[A(w)](E[V])^2$$

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 $\approx (c_a^2+1)/\rho$.

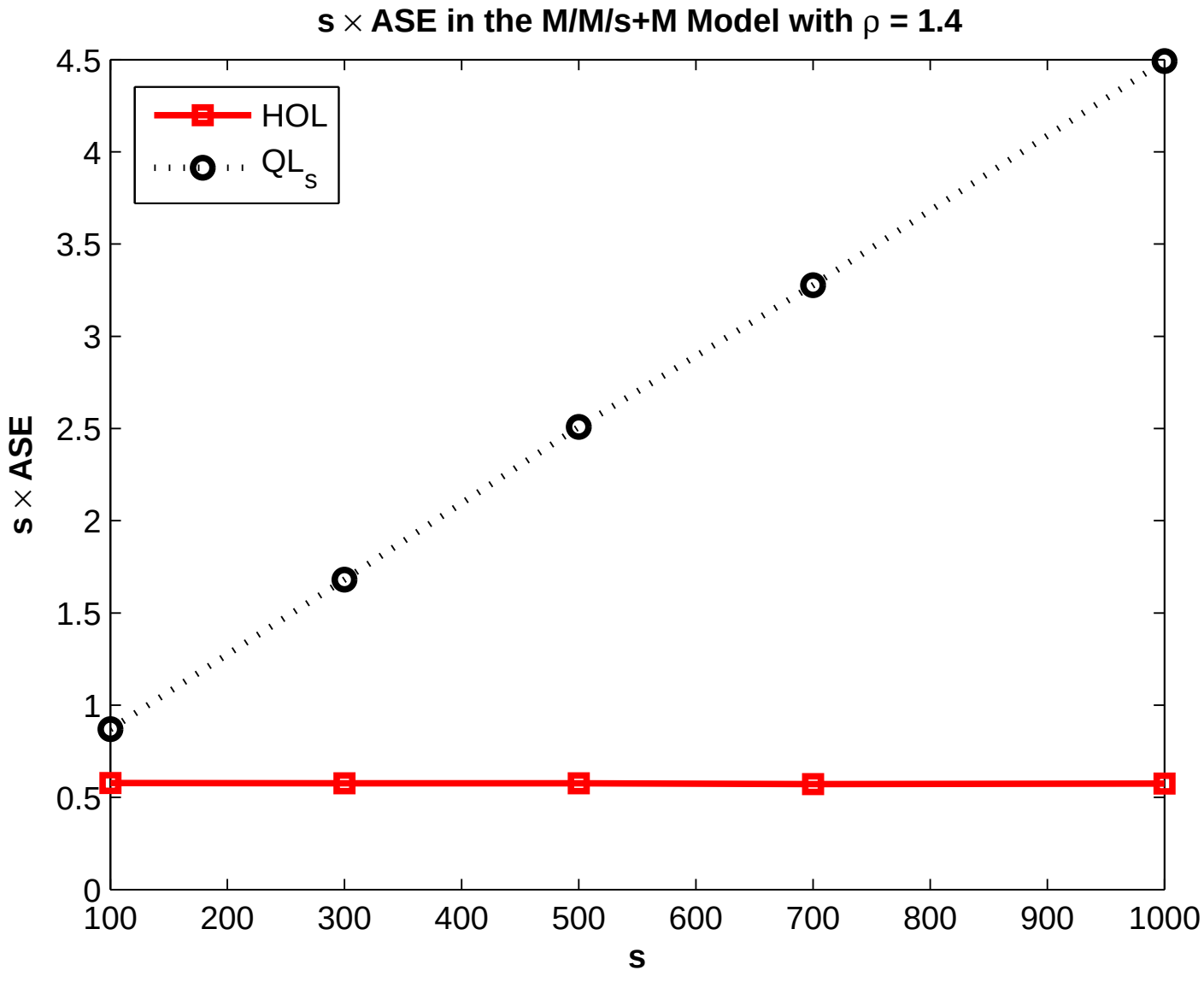
Simulations for the $GI/M/s$ Model: Deterministic Arrivals

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)**

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

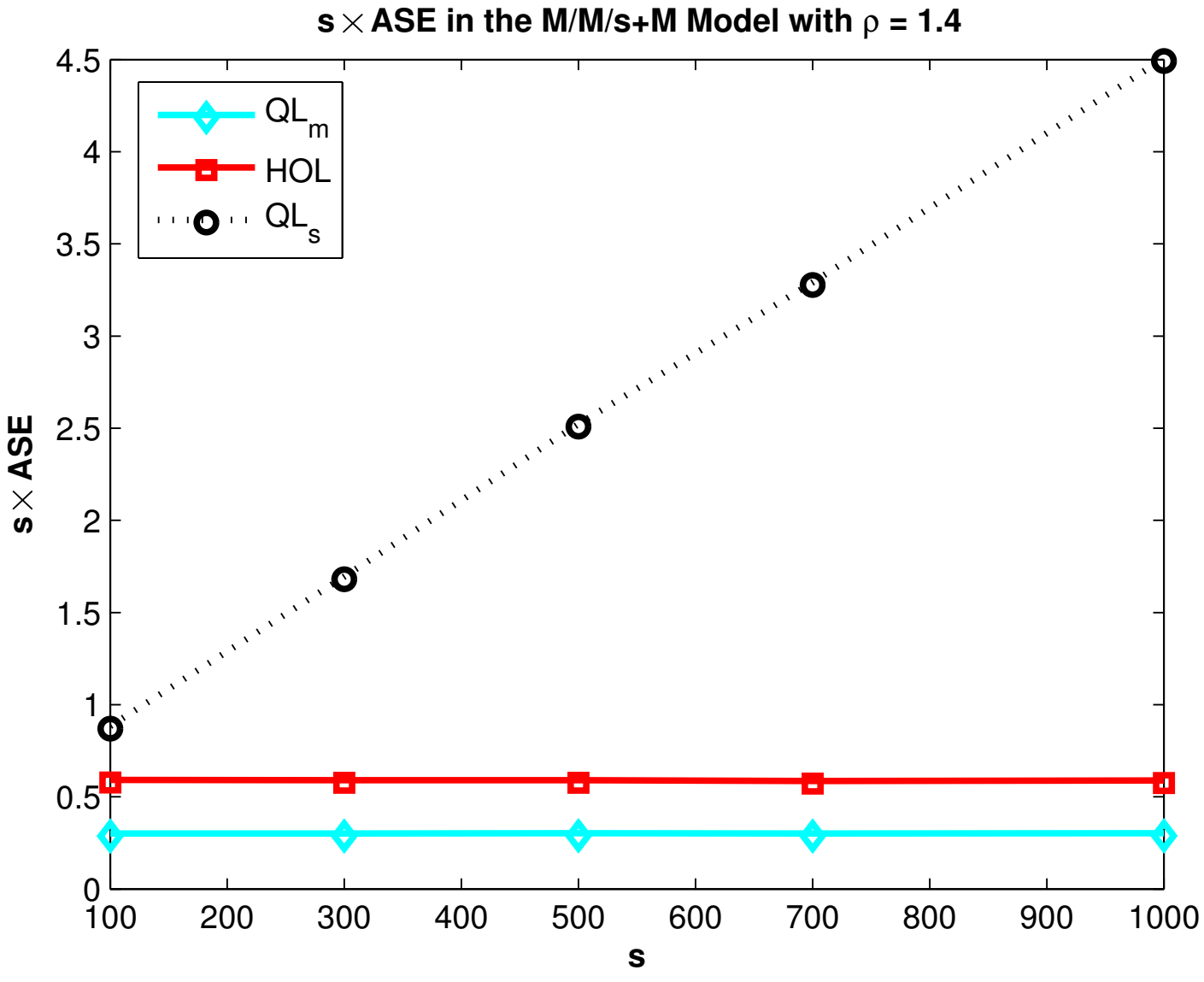
- **QL_m in the $GI/M/s+M$ Model**

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

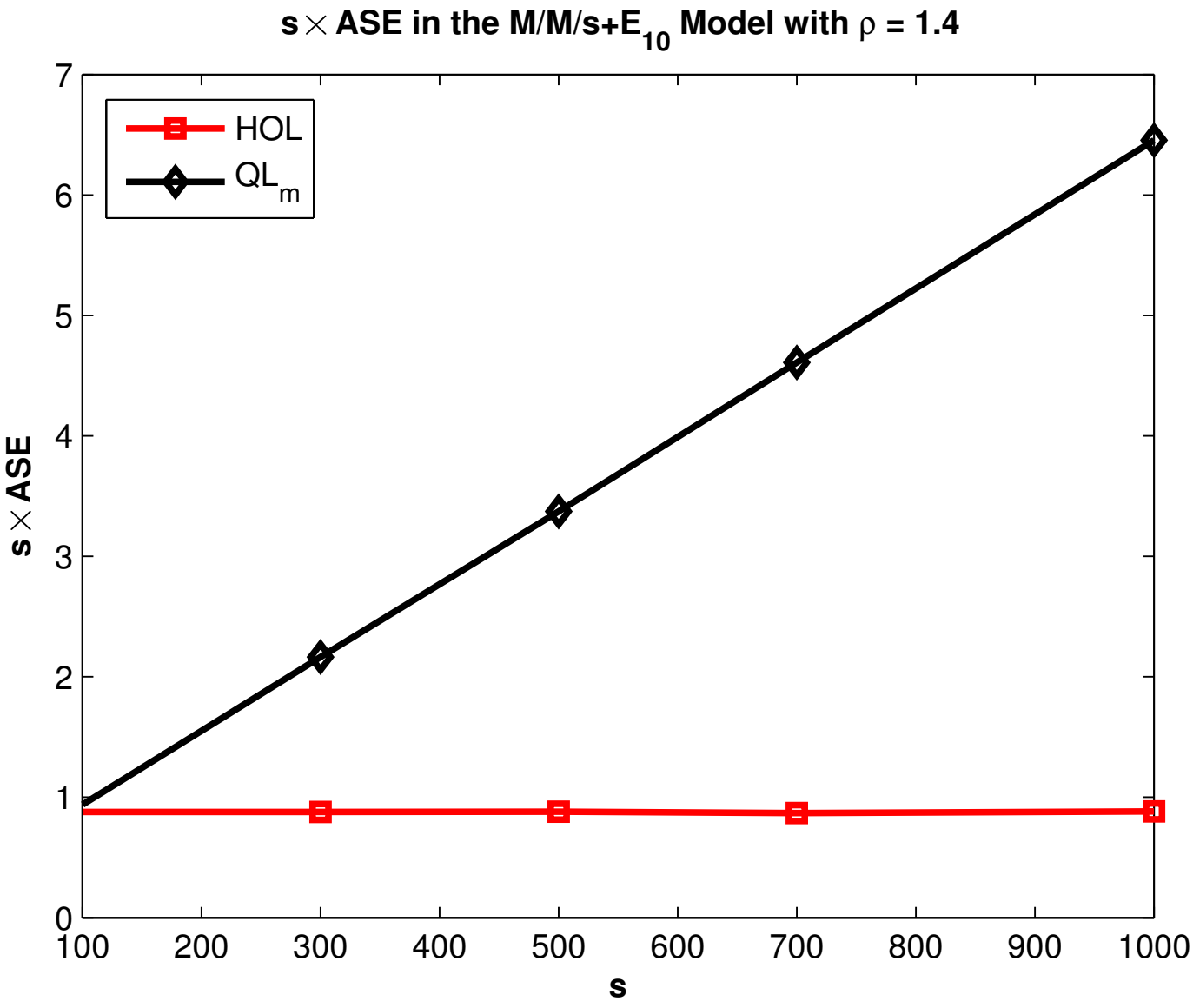
$$\Downarrow$$

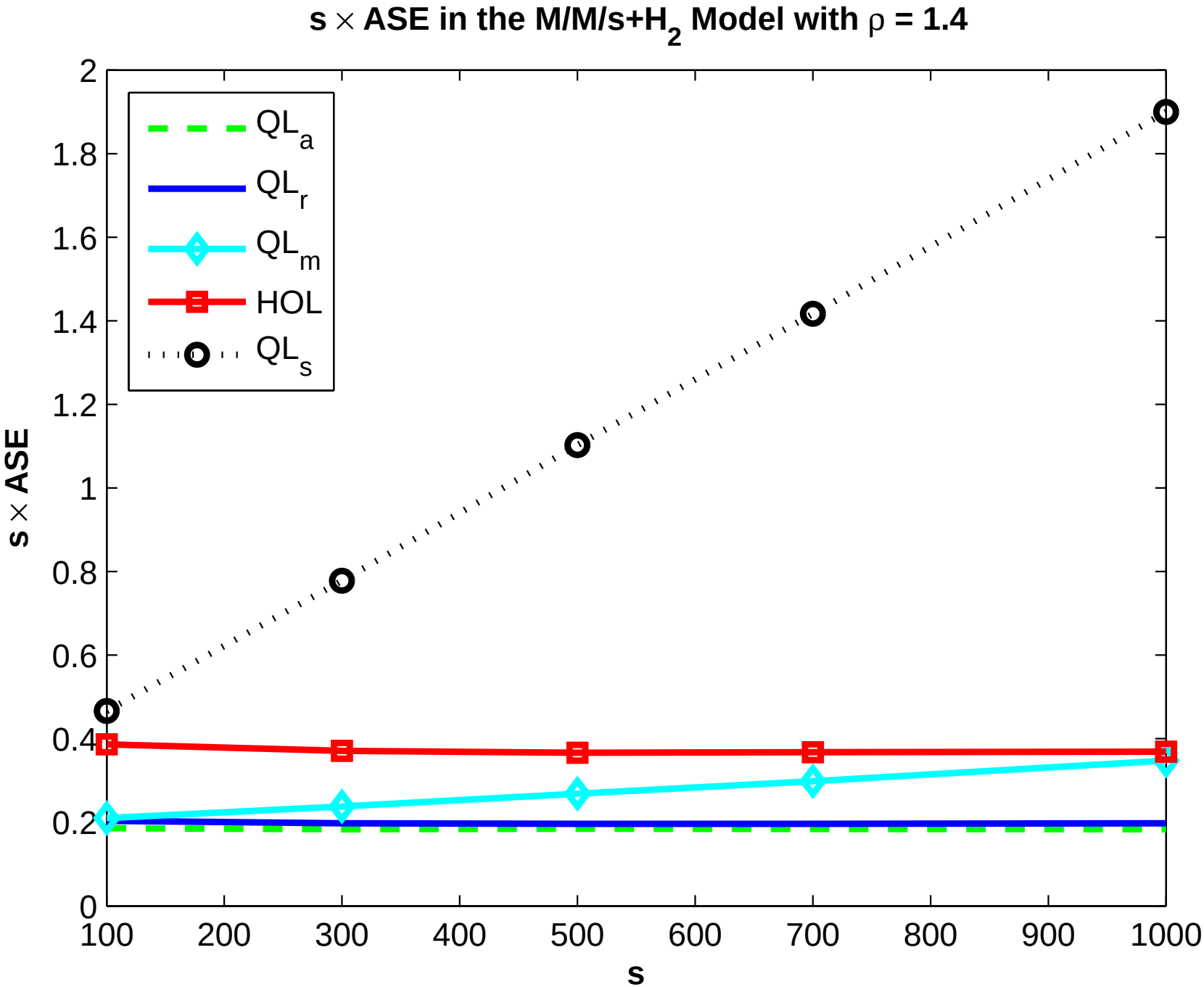
$\theta_{QL_m}(n)$ minimizes the MSE!

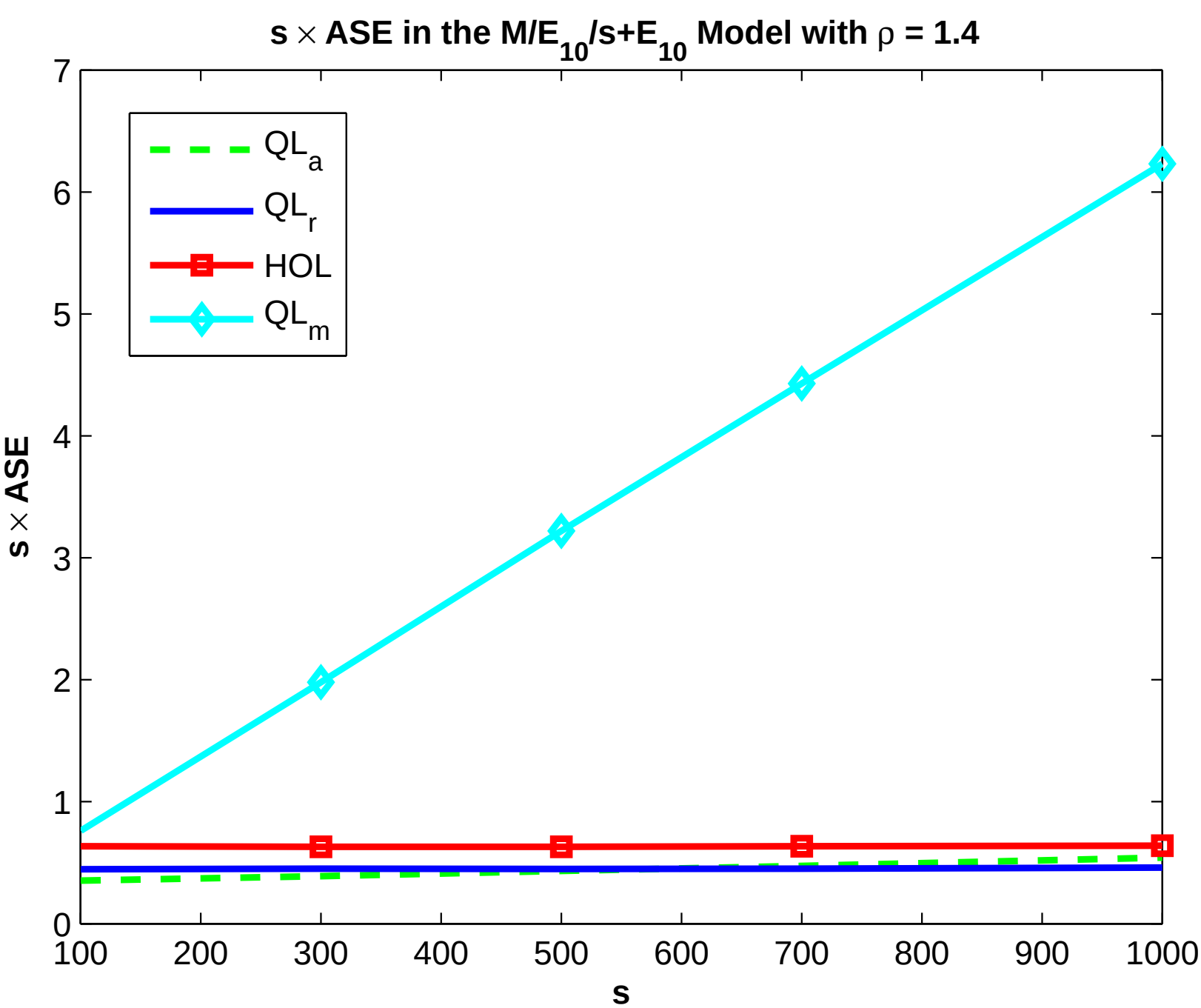
Refined QL Predictor for Same $M/M/s+M$ Model



But when the Abandonment Distribution is Not Exponential

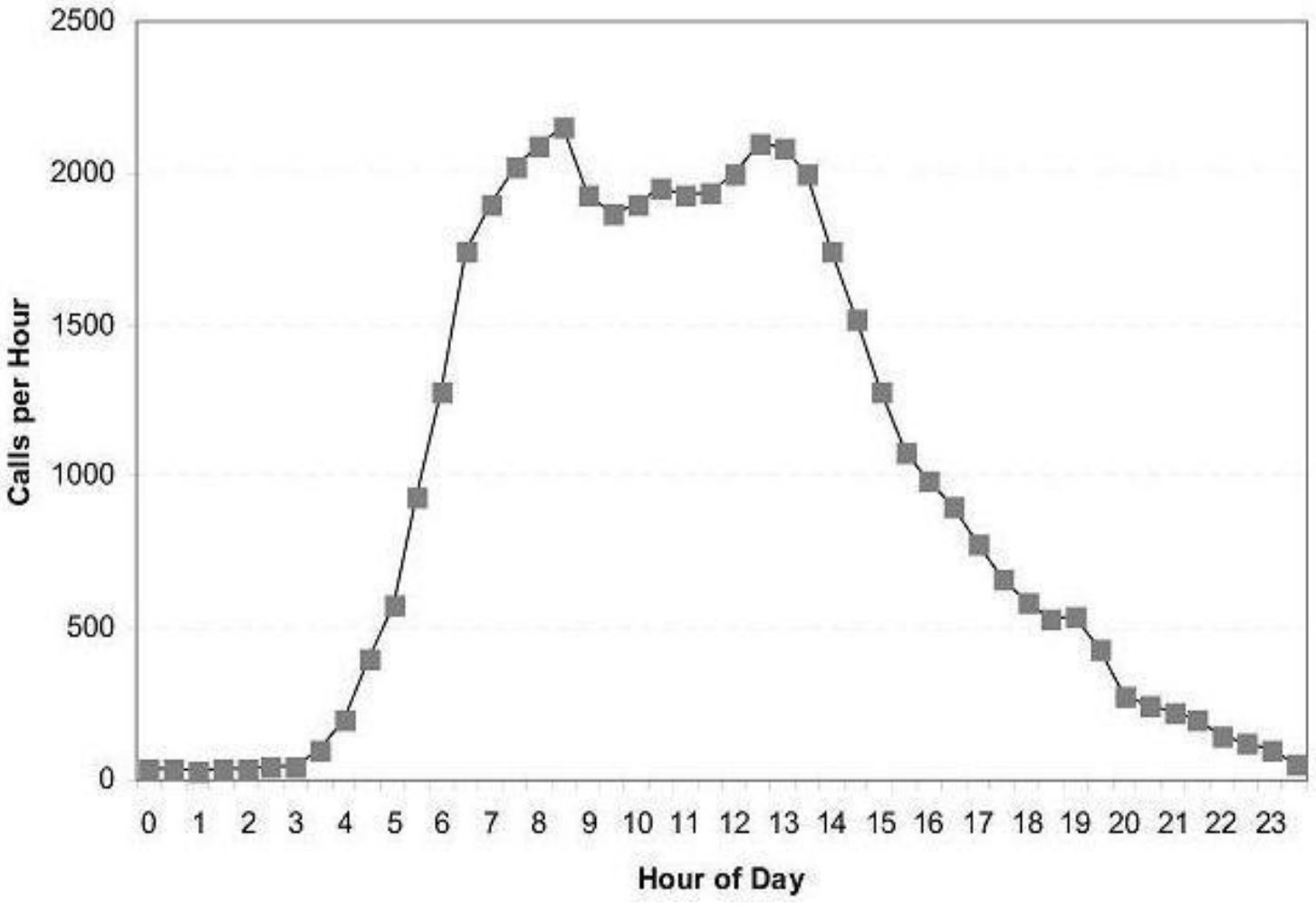



s × ASE in the M/M/s+H2 Model with ρ = 1.4
QLa
QLr
QLm
HOL
QLs
s × ASE
s
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
100
200
300
400
500
600
700
800
900
1000

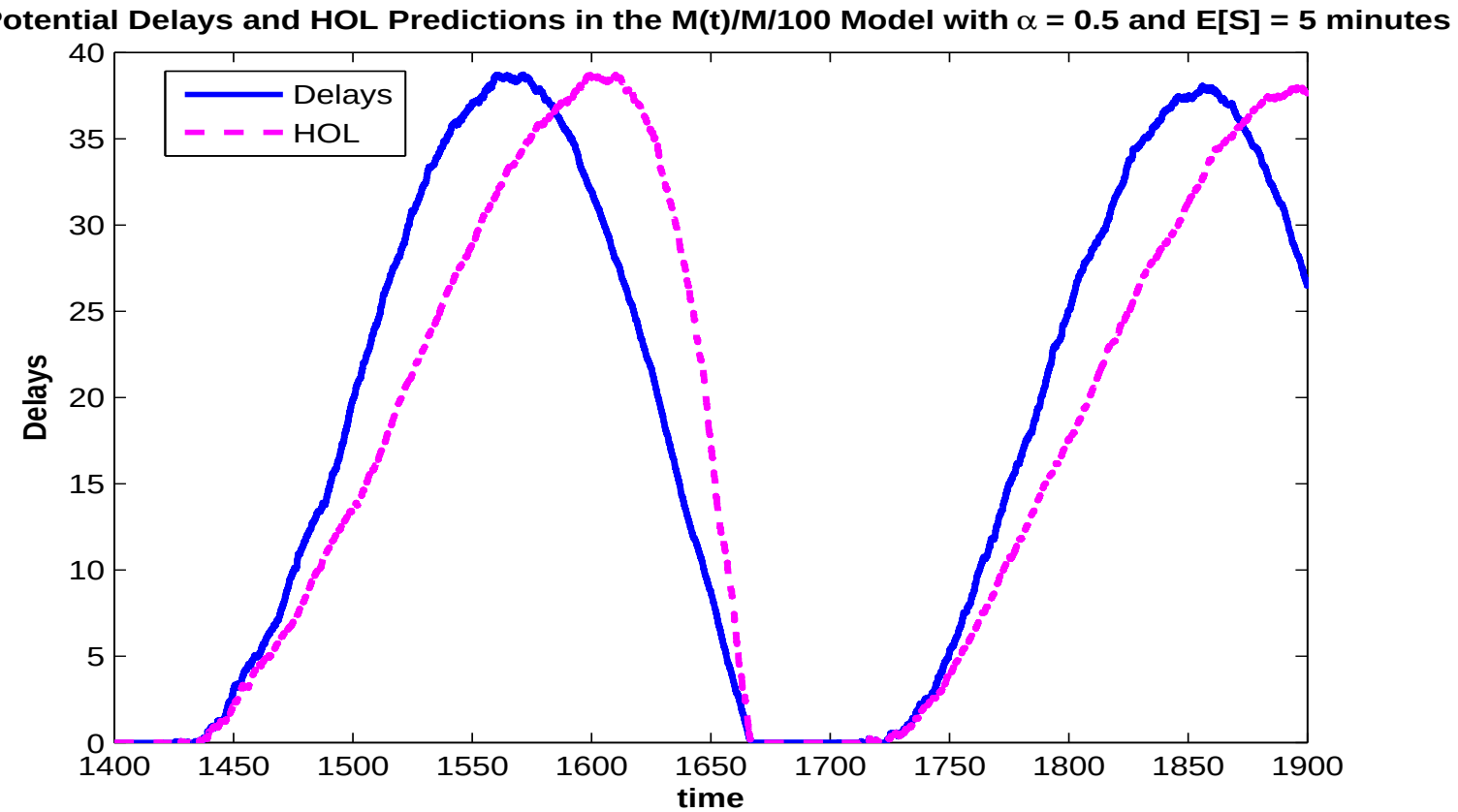

s × ASE in the M/E10/s+E10 Model with ρ = 1.4
QLa
QLr
HOL
QLm
s × ASE
s
0
1
2
3
4
5
6
7
100
200
300
400
500
600
700
800
900
1000

Time-Varying Arrival Rates

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

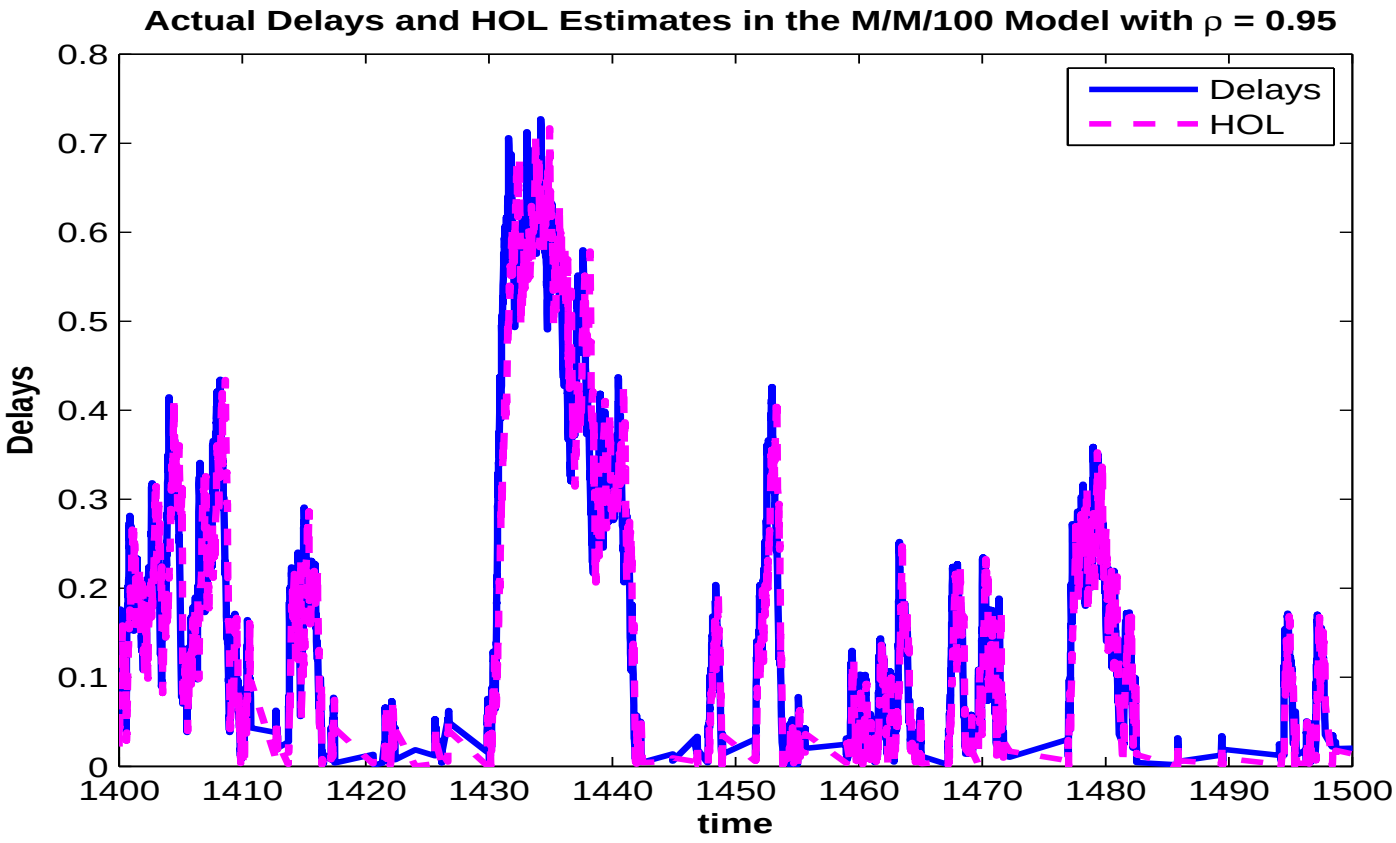


HOL Delay Prediction in the $M_t/M/100$ Model



- 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
- $\rho = \bar{\lambda}/s\mu = 0.95$

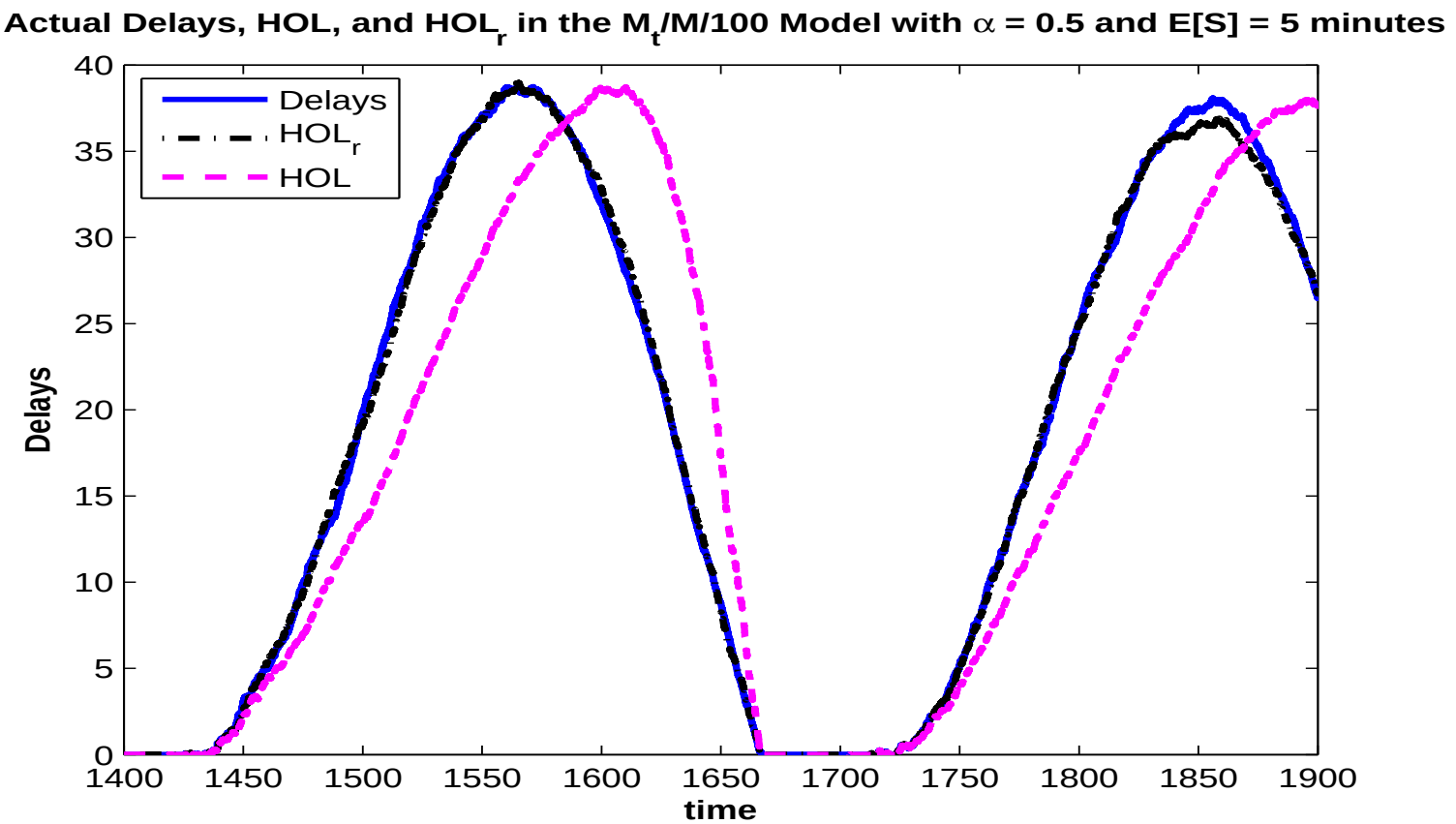
Problem: Time Lag in HOL Delay

- 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

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. It can be used to help make better delay predictions.

The End

References

Other Papers with Rouba Ibrahim

- **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.* 10 (2009) 1729–1742.
- **Real-Time Delay Estimation Based on Delay History in Many-Server Queues with Time-Varying Arrivals.** *Prod. Opns. Mgt.*.