

Real-Time Delay Estimation in Call Centers

Rouba Ibrahim and Ward Whitt

**Industrial Engineering Department
Columbia University**

Why Give Delay Announcements In Call Centers?

Customer Perspective

- Invisible Queues
- Lack of Knowledge About the System
- Solo Waits
- Long Waits (e.g., service-oriented call centers)

Managerial Perspective

- Control Congestion
- Increase Customer Satisfaction

Our Approach to the Problem

Delay Announcement Framework

- Large systems, heavily loaded
- Single delay estimate, given at the beginning of the wait
- No impact on customer behavior

Queueing Theory Approach

- $GI/M/s$ (excluding abandonment)
- $G/GI/s + GI$ (including abandonment)

Contributions

- Propose simple and efficient delay estimators
- Compare estimators based on:
 - (i) accuracy
 - (ii) amount of information required
- Theoretical results supported by simulation experiments

Two Delay Estimators

1. The Simple Queue-Length Estimator (QL_s)

- s = number of agents
- μ = individual service rate
- $Q(t) = n$ = queue length (number of customers waiting)

$$\theta_{QL_s}(n) \equiv (n + 1)/s\mu$$

2. The Head-of-Line Delay Estimator (HOL)

- w = elapsed delay of HOL customer

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

Corresponding Actual Delays

1. $W_Q(n)$ = delay of a customer *given* that:

(i) the customer has to wait

(ii) the customer finds n other customers in queue upon arrival

2. $W_{HOL}(w)$ = delay of a customer *given* that:

(i) there is a customer at the HOL (non-restrictive)

(ii) elapsed delay of HOL customer is w

Quantifying The Efficiency of the Estimators

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

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

- **Simulation Estimate: Average-Squared Error (ASE)**

$$E[MSE] \approx ASE \equiv \frac{1}{k} \sum_{i=1}^k (a_i - e_i)^2$$

a_i = actual delay experienced ($a_i > 0$)

e_i = delay estimation given

k = sample size

GI / M / S

$$\text{mean} = \lambda^{-1} \quad \text{mean} = \mu^{-1} = 1$$

$$c_a^2 = \frac{\text{Var}[U]}{(E[U])^2}$$

QL_s in the GI/M/s Model

- **Distribution of $W_Q(n)$**

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

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

- **MSE of QL_s**

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

$$\Rightarrow c_{W_Q(n)}^2 = \frac{1}{n+1} \rightarrow 0 \text{ as } n \rightarrow \infty$$

θ_{QL_s} minimizes the MSE!

HOL in the $GI/M/s$ Model

- Distribution of $W_{HOL}(w)$

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

$$\Rightarrow E[W_{HOL}(w)] = E\left[\sum_{i=1}^{A(w)+2} V_i/s\right] = \frac{E[A(w) + 2]}{s\mu}$$

- MSE of HOL

$$MSE(HOL) = E[(W_{HOL}(w) - w)^2]$$

Asymptotic approximations for MSE(HOL)

Analytical Results for the $GI/M/s$ Model

Result

$$\frac{c_{W_{HOL}(w)}^2}{c_{W_Q(n)}^2} \rightarrow \frac{c_a^2 + 1}{\rho} \quad \text{as } sw \approx n \rightarrow \infty$$

Intuition

- $W_{HOL}(w)$ becomes more variable as c_a^2 increases
- $W_Q(n)$ is *not* affected by an increase in c_a^2 (conditioning)
- $c_{W_Q(n)}^2 \downarrow 0 \Rightarrow c_{W_{HOL}(w)}^2 \downarrow 0$ as $n \rightarrow \infty$

Simulations for the $GI/M/s$ Model

ASEs are reported in units of 10^{-3} ;

M/M/100

ρ	QL_s	HOL	HOL/ QL_s	$(c_a^2 + 1)/\rho$
0.98	5.033	10.23	2.03	2.04
0.95	2.041	4.269	2.09	2.11
0.93	1.442	3.084	2.14	2.15
0.90	0.9940	2.185	2.20	2.22

D/M/100

ρ	QL_s	HOL	HOL/ QL_s	$(c_a^2 + 1)/\rho$
0.98	2.476	2.624	1.06	1.02
0.95	1.007	1.153	1.14	1.05
0.93	0.7250	0.8713	1.20	1.08
0.9	0.5189	0.6641	1.28	1.11

G / GI / s + GI

$$\text{mean} = \lambda^{-1} \quad \text{mean} = \mu^{-1} = 1$$

$$\begin{aligned} \text{mean} &= \alpha^{-1} \\ \text{cdf } &F \\ \text{hazard rate} &= \frac{f}{1-F} \end{aligned}$$

The Overloaded $G/GI/s + GI$ Model

- $\rho = \frac{\lambda}{s\mu} > 1$
- Longer Delays
- Abandonment \Rightarrow Stability

The Need to Go Beyond QL_s

ASEs are reported in units of 10^{-3} ; $\rho = 1.4$;

M/M/s + M

s	QL_s	HOL
100	8.693	5.845
500	4.942	1.136
1000	4.543	0.5699

M/LN(1, 1)/s + LN(1, 1)

s	QL_s	HOL
100	18.87	6.012
500	14.14	1.192
1000	13.43	0.6040

A Simple Refined QL Estimator (QL_{sr})

$$\theta_{QL_{sr}}(\mathbf{n}) = \beta \times \theta_{QL_s}(\mathbf{n})$$

β is a model-specific constant

Parameters Needed: $Q(t)$, s , μ , $F(x)$, λ

Derivation of $QL_{sr}(n)$

Deterministic Approximations For Large Systems

$$\rho F^c(w) = 1$$

$$q = \rho s \int_0^w F^c(x) dx$$

Approximation for QL_s

$$\theta_{QL_s}(n) = (n + 1)/s\mu \approx (q + 1)/s \approx q/s$$

Refinement of QL_s

$$\theta_{QL_{sr}}(n) = \underbrace{\left(\frac{w}{q/s}\right)}_{\beta} \times \theta_{QL_s}(n)$$

Markovian QL Estimator

- The Markovian Queue-Length Estimator (QL_m)

$$\theta_{\text{QL}_m}(n) = \sum_{i=0}^n 1/(s\mu + i\alpha)$$

Parameters Needed: $Q(t)$, s , μ , α

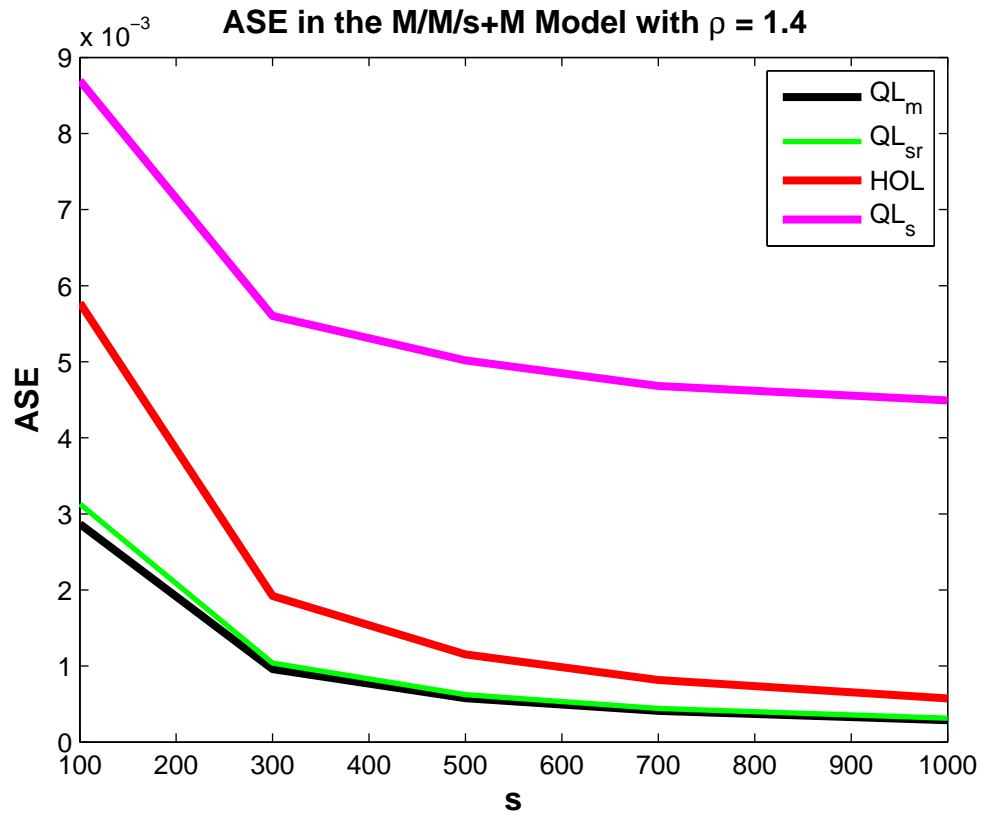
- QL_m in the M/M/s+M Model

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

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

$$E[W_Q(n)] = \sum_{i=0}^n E[X_i] = \sum_{i=0}^n 1/(s\mu + i\alpha) = \theta_{\text{QL}_m}(n)$$

QL_m minimizes the MSE!



Approximation-Based Queue-Length Estimator

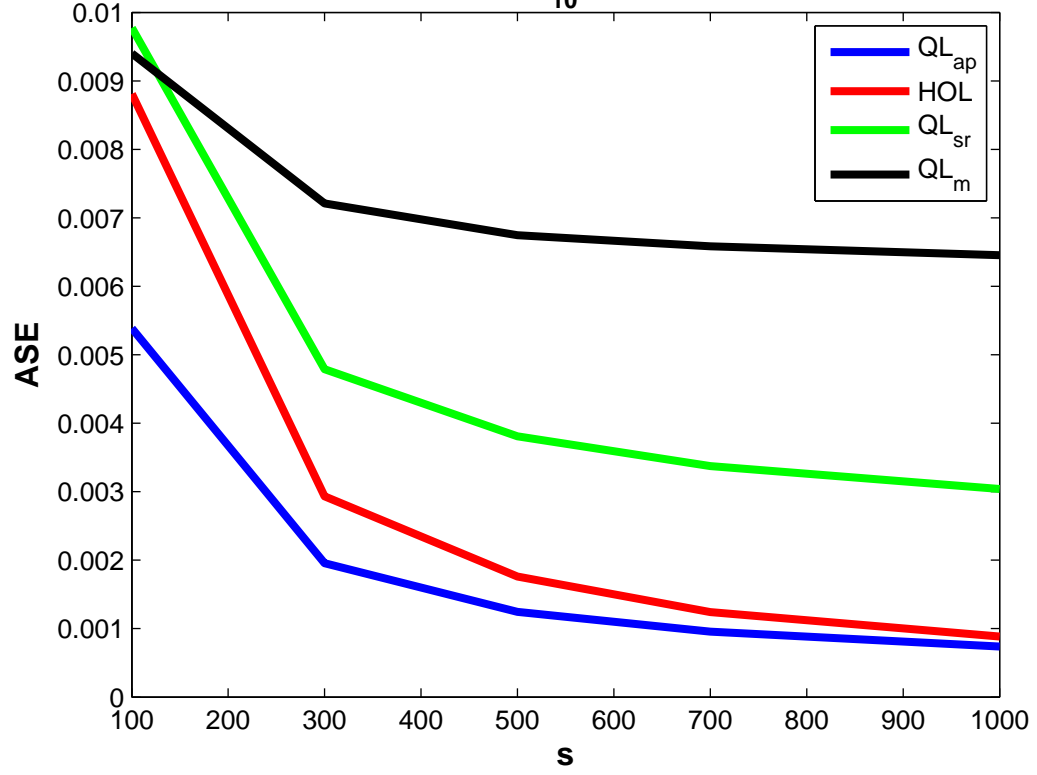
$$M/GI/s + GI \approx M/M/s + M(n) \quad (\text{Whitt 2005})$$

$$\theta_{\text{QL}_{\text{ap}}}(\mathbf{n}) = \sum_{i=0}^n \frac{1}{s\mu + \delta_i}$$

where $\delta_i = \sum_{j=n-i+1}^n h(j/\lambda)$ is the approximate total abandonment rate in state $i > 0$
and $\delta_0 = 0$

Parameters Needed: $Q(t), s, \mu, F(x), \lambda$

ASE in the M/M/s+E₁₀ Model with $\rho = 1.4$



Summary and Future Work

Summary

- New simple and efficient delay estimators
- Comparison of the estimators via theory and simulations

Future Work

- Evaluate with real call center data
- Analysis for the $G/GI/s + GI$ model
- Include customer reaction into our models

THANK YOU!