On the Accuracy of Delay-History-Based Predictors in Large Call Centers

Mor Armony, New York University

Joint work with:
Achal Bassamboo, Northwestern University
Rouba Ibrahim, Columbia University
Avi Mandelbaum, Technion

MSOM 2010

Motivating Application: Delay Announcements

Modern Call Centers

- Large
- Uncertain and time-varying demand $\Rightarrow$ inappropriate staffing
- Long waiting times (e.g., service-oriented call centers)
- Uncertainty about length of wait (invisible queues)

Delay Announcements

- Inexpensive
- Relatively easy to implement
- Improve quality of service
- Control congestion: **impact customer behavior**

Delay-History-Based Predictors

Exploit the recent history of delays in the system.

Advantages

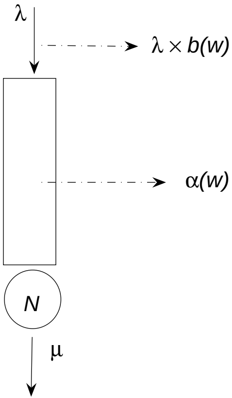
- Do not rely on system parameters
- Robust
- Easy to interpret

Last-to-Enter-Service Predictor (*LES*)

- $w(t)$ = waiting time of the customer arriving at time t
- τ_t = arrival time of the LES customer at time t

$$\boxed{\theta_{LES}(t) \equiv w(\tau_t)}$$

Single class/Single pool Model



- Arrival rate: λ
- Prob. of balking: $b(w)$ depends on the announced delay w
- Service rate: μ
- Number of servers: N
- Abandonment rate: $\alpha(w)$ depends on the announced delay w

Our Work

Question: How accurate is the *LES* delay predictor?

Part 1: Asymptotic Accuracy of *LES*

- Abandonment: independent of the announcement
- Abandonment: dependent on the announcement
- Balking: dependent on the announcement

Part 2: Empirical Study

- Real-life call center data
- Accuracy of LES
- New delay-history-based predictors

Literature Review

- Delay announcements and their effect on system dynamics: Hassin (1986), Whitt (1999a), Armony & Maglaras (2004), Guo & Zipkin (2007), Armony et al. (2009), Allon et al. (2010a,b)
- Lead time quotations in manufacturing: Duenyas & Hopp (1995), Spearman & Zhang (1999), Ata & Olsen (2007), Dobson & Pinker (2006)
- Accuracy of waiting time estimates: Nakibly (2002), Whitt (1999b), Jouini et al (2007), Ibrahim & Whitt (2009a,b,c,d)

Part 1

Asymptotic Accuracy of LES

A Sequence of Systems (QED Asymptotic Regime)

N^{th} System:

- Service rate: μ
- Number of Servers: N
- Arrival rate: $\lambda^N = N\mu + \mathcal{O}(\sqrt{N})$
- Abandonment rate: $\alpha^N(w)$
- Balking probability: $b^N(w)$

Announcement independent abandonment

- Consider the case where $\alpha^N(w) \equiv \alpha$, $b^N(w) \equiv 1$
- Waiting times are "small" (Garnett et al.)

$$w^N(t) = \mathcal{O}\left(\frac{1}{\sqrt{\lambda^N}}\right)$$

- Scaled queue length is diffusion
- Scaled queue length: almost constant between arrival and departure
- Snapshot principle: $\sqrt{N} w^N \approx \frac{1}{\mu} \frac{Q^N}{\sqrt{N}}$ (Puhalskii).

Theorem

As the system size increases,

$$\sqrt{N}|w(t^N) - w(\tau_t^N)| \Rightarrow 0, \quad \text{for all} \quad t > 0.$$

WT based on *LES* ≈ Actual WT

$$\Updownarrow$$

LES is asymptotically correct

Announcement dependent abandonment behavior

- Consider the case where $\alpha^N(w) \in [\alpha_1, \alpha_2]$, $b^N(w) \equiv 1$
- We show,

$$w^N(t) = \mathcal{O}\left(\frac{1}{\sqrt{\lambda^N}}\right)$$

- Scaled queue length may/may not be a diffusion!
- Queue length: almost constant between arrival and departure

Theorem

As the system size increases,

$$\sqrt{N}|w(t^N) - w(\tau_t^N)| \Rightarrow 0, \quad \text{for all} \quad t > 0.$$

WT based on *LES* $\approx$ Actual WT

$\Updownarrow$

LES is asymptotically correct

Sketch of the Proof

Bounding argument

- First show that

$$|\tau_t^N - t| \to 0, \quad \text{as } N \to \infty.$$

- Consider two systems initialized at time τ_t^N:
  - System I: Abandonment rate is α_1
  - System II: Abandonment rate is α_2.

- We can construct such that

$$Q_{\text{Sys II}}(t+s) \leq Q(t+s) \leq Q_{\text{Sys I}}(t+s) \quad \text{for all } s \geq 0.$$

- $Q_{\text{Sys}\,\cdot}$ converge to diffusion processes

General distribution for abandonment time

- If announcement is w, then the abandonment time of the customer has distribution F_w
- We assume that the hazard rate of F_w is uniformly bounded from above and below
- Similar bounding argument holds
- *LES* announcements are asymptotically accurate

System with Balking

- Customer balks with probability $b(w)$ if given announcement w
- If $b(w)$ is fixed and is of $\mathcal{O}(\frac{1}{\sqrt{N}})$
  - Diffusion limit holds
  - Snapshot principle implies asymptotic accuracy
- For general $b(w)$, under technical conditions *LES* is asymptotically accurate

Part 2

Statistical Analysis of Call Center Data

Description of the Data

Call Center of a US Bank

- Large call center:
  - 900-1200 agents on weekdays
  - 200-500 agents on weekends
- Multiple sites: NY, PA, RI, and MA
- Routing: skill-based, across sites
- Up to 300,000 calls/day
- Types of services: Retail, Premier, Business, Consumer Loans, Online Banking, and **Telesales**

Data Set

- Single customer class: **Telesales**
- 7769 calls registered over two weekdays:
  - 05/22/2003 (3654 calls)
  - 05/28/2003 (4115 calls)
- Working hours: 7 AM - midnight
- Around 50 agents (time-varying)

Summary Statistics

Wait = time until either entry to service or abandonment.

05/22/2003 (in secs)

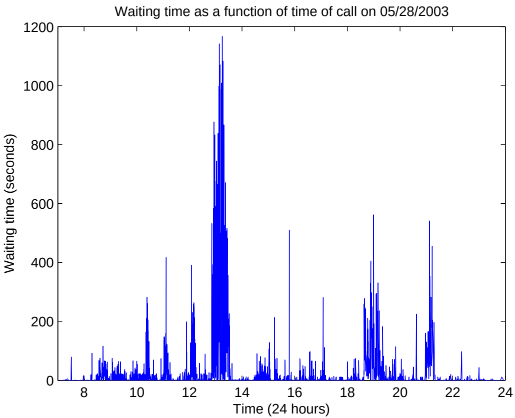
Average wait	17
Std dev. of wait	**62**
Average positive wait	43
Proportion of delayed customers	38%
Proportion of abandonment	3.0%
75^{th} percentile	5.0

05/28/2003 (in secs)

Average wait	24
Std dev. of wait	**83**
Average positive wait	54
Proportion of delayed customers	45%
Proportion of abandonment	4.0%
75^{th} percentile	14

- Large variance $\Rightarrow$ overall average is not a reliable predictor
- Need to use information about current system state

Time Variation of Waits



- Peaks correspond to a decrease in the number of available agents (e.g., lunch break around 1pm)
- Fluctuations ⇒ errors in delay-history-based predictions

Quantifying Accuracy

Sample Bias (B)

$$B \equiv \frac{1}{k}\sum_{i=1}^{k}(p_i - d_i)$$

- p_i = delay prediction for customer i
- d_i = measured delay for customer i ($d_i > 0$)
- k = sample size

Average Squared Error (ASE)

$$ASE \equiv \frac{1}{k}\sum_{i=1}^{k}(p_i - d_i)^2 \ .$$

We consider $\sqrt{ASE}$.

Biased LES Prediction

Time unit $=$ 1 second.

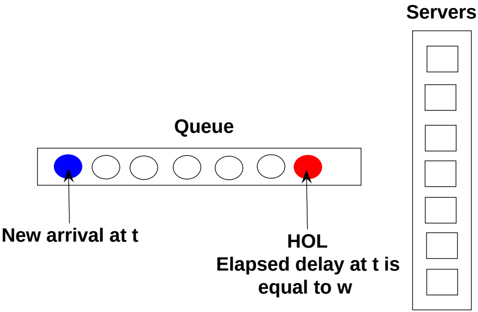
	$B(LES)$	$\sqrt{ASE(LES)}$
05/22	**-18**	99
05/28	**-7.5**	160

	$B(LES)$/Avg. Wait	$\sqrt{ASE(LES)}$/(Avg. Wait)
05/22	-0.42	2.3
05/28	-0.15	3.0

Problem: Announce $LES = 0$ to delayed customers.

- 05/22: 51% of LES announcements $= 0$
- 05/28: 58% of LES announcements $= 0$

Alternative Delay-History-Based Predictors



Head-of-Line Predictor (*HOL*)

- w_H = elapsed delay of HOL customer

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

The *HOL* announcement is positive if there is an HOL customer. Otherwise, announce *LES*.

Performance of Predictors

Time unit = 1 second.

Estimates on 05/22/2003

	HOL	LES	$(LES + HOL)/2$
B	-0.34	-18	-9.1
$\sqrt{ASE}$	120	99	98

	HOL	LES	$(LES + HOL)/2$
$B/$ Avg. Wait	-0.0078	-0.42	-0.21
$\sqrt{ASE}$ /Avg. Wait	2.7	2.3	2.2

We corrected for the bias but large variance remains.

New Predictor Based on the Traffic Intensity

M. Armony
A. Bassamboo
R. Ibrahim
A. Mandelbaum

Refined *LES* Predictor (LES_r)

- t_L = arrival time of LES customer
- t_C = arrival time of current customer
- w_H = elapsed delay of HOL customer
- w_L = delay of LES customer
- $\rho(t)$ = estimate of traffic intensity at time t

$$\boxed{\theta_{LES_r} \equiv \frac{\rho(t_C)}{\rho(t_L)} \times \frac{w_H + w_L}{2}}$$

Frame of Reference

No-Information Predictor (*NI*)

- Uses no information about current system state
- Announces average waiting time

Queue-Length-Based Predictor (*QL*)

- n = queue length upon arrival
- s = number of agents upon arrival
- m = average service time upon arrival

$$\boxed{\theta_{QL}(n) \equiv (n+1) \times \frac{m}{s}}$$

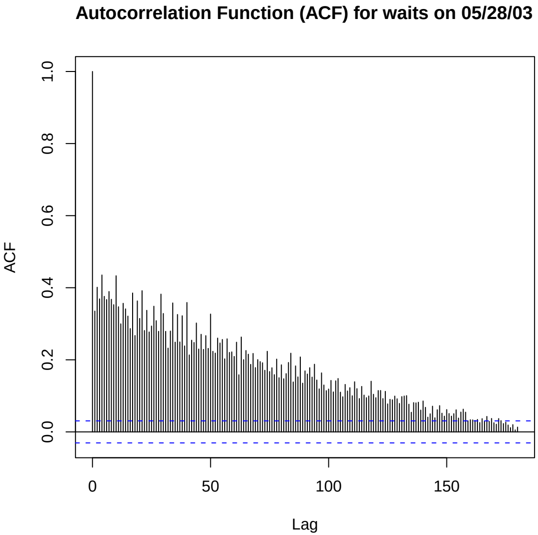
Performance Conditional on the Level of Delay

Time unit = 1 second.

		NI	QL	LES	LES_r
Delays smaller than 30 (71%)	B $\sqrt{ASE}$	0 7.7	-8.0 11	6.8 46	14 52
Delays in (30, 120) (20%)	B $\sqrt{ASE}$	0 25	-66 70	-26 87	-23 69
Delays larger than 120 (9%)	B $\sqrt{ASE}$	0 62	-274 336	-205 282	-185 270

- LES_r is more accurate than LES for long delays
- Error in LES_r prediction: large bias remains

Autocorrelation Function



Suggests averaging over several past delays.

Concluding Remarks

We studied the accuracy of *LES*

- Easy to implement
- Needs no info. regarding system parameters

Asymptotic Results

- It is asymptotically accurate in the QED regime with abandonments/balking
- The result holds even if there is no diffusion approximation!

Empirical Results

- Problem: Large variance of delays, significant time-variability
- *LES* has significant prediction error: bias + variance
- New delay-history-based predictors: significant bias remains
- Time-series analysis approach seems promising