Accuracy of the Last-To-Enter-Service Delay Predictor in Large Call Centers

Rouba Ibrahim, University of Montreal

Joint work with:
Mor Armony, New York University
Achal Bassamboo, Northwestern University
Cliff Hurvich, New York University
Avi Mandelbaum, Technion

INFORMS 2010

Motivating Application: Delay Announcements

Modern Call Centers

- Long waiting times (e.g., service-oriented call centers)
- Uncertainty about length of wait (invisible queues)

Delay Announcements

- Improve quality of service
- Control congestion: **impact customer behavior**

Need accurate ways to predict delays.

Literature Review

- **Asymptotic analysis of multiserver queues:**
  Halfin and Whitt (1981), Puhalskii and Reiman (2000), Garnett et al. (2002), Borst et al. (2004), Talreja and Whitt (2009)...

- **Delay announcements in call centers:**
  Whitt (1999), Nakibly (2002), Armony and Maglaras (2004), Guo and Zipkin (2007), Jouini et al. (2007), Armony et al. (2009), I. and Whitt (2009), Allon et al. (2010)...

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*)

- t = arrival time of new (current) customer
- w = delay of LES customer prior to t

$$\theta_{LES}(t, w) \equiv w$$

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

Our Work

Part 1: Asymptotic Accuracy of *LES*

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

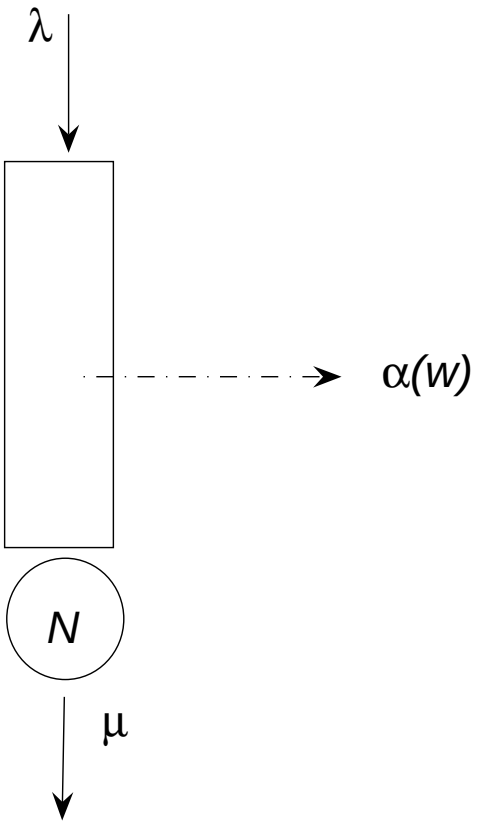
Part 2: Empirical Study

- Real-life call center data
- Accuracy of LES
- Alternative predictors

Part 1

Asymptotic Accuracy of LES

Single class/Single pool Model



- Poisson arrival process with rate λ
- Exponential service times with rate μ
- N servers
- Exponential times to abandon with rate $\alpha(w)$

A Sequence of Systems (Halfin-Whitt Asymptotic Regime)

N^{th} system:

- Service rate: μ
- Number of Servers: N
- Arrival rate: $\lambda^N = N\mu + \mathcal{O}(\sqrt{N})$
- Bounded abandonment rate: $\alpha(w) \in [\alpha_1, \alpha_2]$ for all w

Halfin and Whitt (1981), Puhalskii and Reiman (2000), Garnett et al. (2002), Talreja and Whitt (2009).

Announcement independent abandonment

Consider the case where abandonment rate α is constant

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

$\Updownarrow$

LES is asymptotically correct

Intuition: Snapshot Principle (Reiman 1982)

- Space scaling: virtual waiting times are "small" for large N

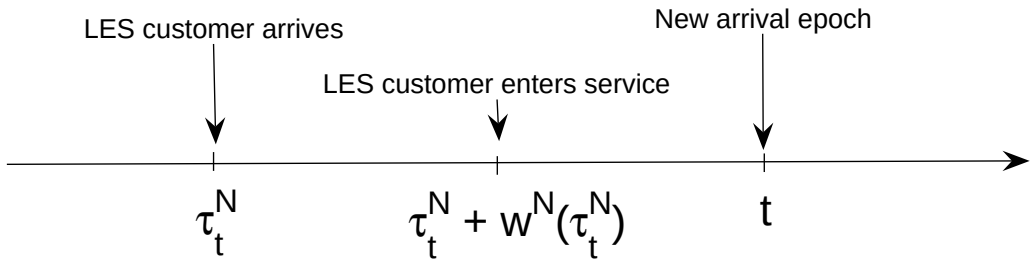
$$w^N(t) = \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$$

- No time scaling: waits change significantly over time intervals of length $\mathcal{O}(1)$

Announcement dependent abandonment behavior

N^{th} system:

- $w^N(t)$: virtual waiting time of customer arriving at time t
- τ_t^N: arrival time of last customer to enter service before t



We show that as $N \to \infty$:

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

Sketch of the Proof

Bounding argument

- First show that for all t:

$$|\tau_t^N - t| \to 0 \quad \text{as} \quad N \to \infty, \text{ w.p.1 .}$$

- 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 w.p.1:

$$Q^N_{\text{Sys II}}(\tau_t^N + s) \leq Q^N(\tau_t^N + s) \leq Q^N_{\text{Sys I}}(\tau_t^N + s) \quad \text{for all } s \geq 0.$$

- Scaled Q^N_{Sys}. converges to a diffusion processes

Simulation Experiments

Average Absolute Difference (D)

$$D \equiv \frac{1}{k} \sum_{i=1}^{k} |p_i - d_i| \ ,$$

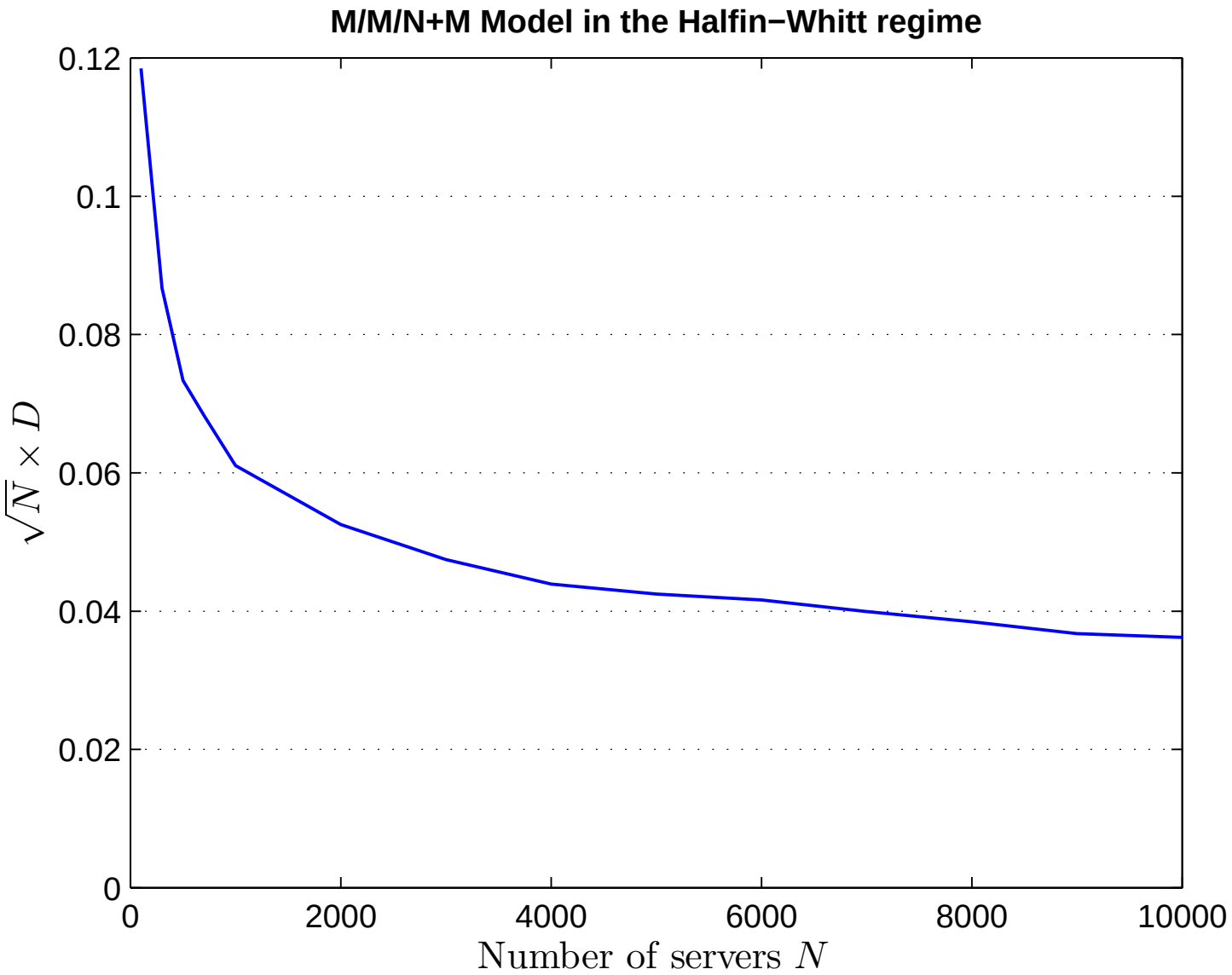
- p_i = delay prediction for customer i
- d_i = potential delay for customer i
- k = sample size

Need: $\sqrt{N} \times D \to 0$ as $N \to \infty$.

$M/M/N+M$ Models

- N ranges from 100 to 10000
- $\rho \equiv \frac{\lambda}{N\mu} = 1$
- $\alpha(w) \in \{\alpha_1, \alpha_2, \alpha_3\}$

Simulation Results



Part 2

Statistical Analysis of Call Center Data

Description of the Data

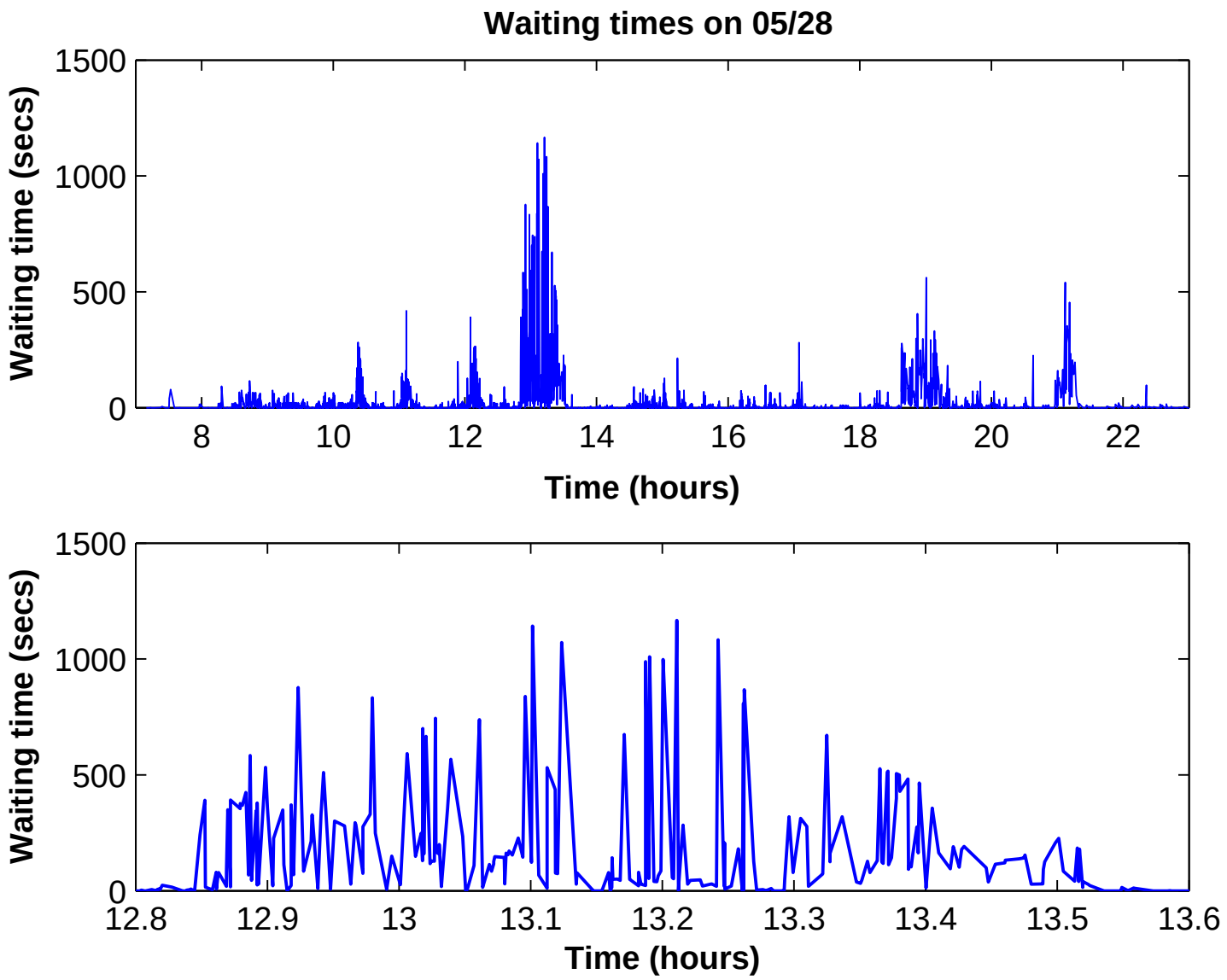
Call Center of a US Bank

- 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**
- About 83000 calls registered over May 2003
- Working hours: 7 AM - midnight

Waits on 05/28/2003



Summary Statistics of Waits on 05/28/2003

Wait = time until entry to service.

05/28/2003 (in secs)

Average wait	20
Std dev. of wait	**75**
Average positive wait	47
Proportion of delayed customers	43%
Proportion of abandonment	20%
75^{th} percentile	11
Proportion with wait $>$ 1 min	8.0%

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

(Similar behavior on 05/22/2003.)

ASE, Bias and Variance

Average Squared Error (ASE)

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

- p_i = delay prediction for customer i
- d_i = measured delay for customer i
- k = sample size

Sample Bias (B)

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

$ASE = Variance + B^2 \Rightarrow$ Best we can do is: $ASE = Variance$

Accuracy of *LES* Prediction on 05/28

Time unit = 1 second.

$B^2(LES)$	$ASE(LES)$	Var
5322	8624	3302

B/Avg. Wait	$\sqrt{ASE}$/Avg. Wait	$\sqrt{Var}$/Avg. Wait
3.6	4.6	2.9

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

The QL_s and NI Predictors

Simple Queue-Length-Based Predictor (QL_s)

- n = queue length upon arrival at t
- $s(t)$ = number of agents at t
- m = average service time

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

No-Information Predictor (NI)

Announces the average daily wait.

Performance of the Predictors

Time unit = 1 second.

Estimates on 05/28 (Avg. Wait = 20 secs)

	LES	QL_s	NI
B^2	5322	2239	0
ASE	8624	5930	5656
Var	3302	3691	5656

Concluding Remarks

Asymptotic Results

- LES is asymptotically accurate in the Halfin-Whitt regime with abandonments
- With/without customer reactions to the announcements

Empirical Results

- Problem: Large variance of delays, significant time-variability
- *LES* has significant prediction error: bias + variance
- New predictors: significant bias remains
- Investigate ways of reducing the bias

THANK YOU!

System Characteristics