

# Managing Callbacks in Call Centers

---

Mor Armony

Stern School of Business

New York University

Costis Maglaras

Graduate School of Business

Columbia University

May 2003

## Motivation

---

- **Recent trends in Customer Contact Centers:**

- Anticipated delay information.
- Offer alternative channels (e-mail, on-line chat, call-back option)

- **Research Shows:**

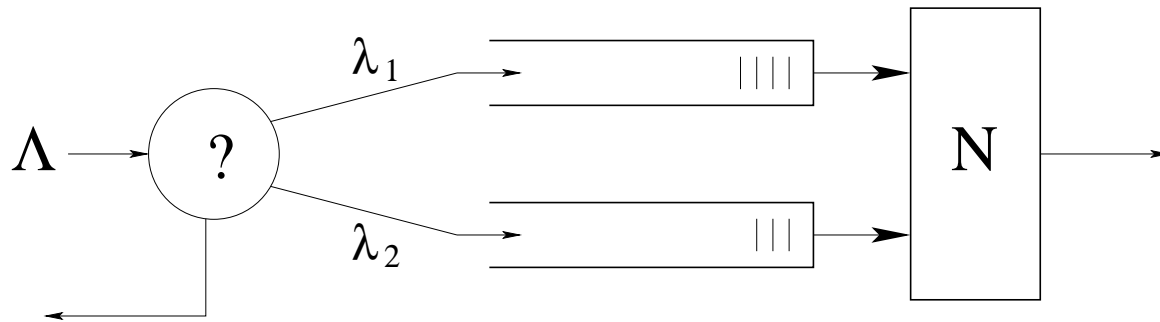
- Announcing delay information improves performance and profitability  
(Whitt (1999), Plambeck (2001))
- Offering the call back option improves performance (A. & Maglaras 2001)

- **Our Goal:** Study system that

- 1. offer a call back option, and
- 2. announce delay information,
- Performance analysis & System design (staffing rules).

## The Model

---



- Service Facility:

- $N$  statistically identical servers

- Two service modes with iid service requirements  $\sim \exp(\mu)$

1. real time service

2. postponed service (call-back option) with deadline  $D_2$ .

- System state:  $S(t) = (Q_1(s), Q_2(s), Z_1(s), Z_2(s); s \leq t)$ .

## The Model (cont.)

---

- Customer behavior:

- Aggregate arrivals  $\sim$  Poisson rate  $\Lambda$

- Available info:  $\hat{w}_1(S)$  and  $D_2$

- Customer choose among

1. real time service

2. call-back

3. balk

- The Resulting System: a two class M/M/N with state-dependent arrival rates:

$$\lambda_1(S) \triangleq \lambda_1(\hat{w}_1(S), D_2) \quad \text{and} \quad \lambda_2(S) \triangleq \lambda_2(\hat{w}_1(S), D_2).$$

### Fundamental Tradeoffs:

- Number of servers vs. congestion levels

- Call back deadline vs. extent of load balancing

## Operational Questions

---

1. How to sequence the jobs?

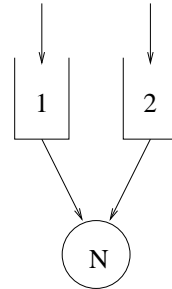
- satisfy call-back option deadline
- minimize class 1 waiting time

2. How to estimate class 1 waiting time?

- Based on current information
- What information to announce?

## Related Literature

---



### Control of the $V$ -design

- **Exact MC Solution:** Brandt & Brandt ('99), Gans & Zhou ('02), Mandelbaum & Yahalom ('03)
- **Asymptotic (many-server) Solutions:** Atar, Mandelbaum & Reiman ('02), Harrison & Zeevi ('02), Maglaras & Zeevi ('03)

### Call-Centers in the many-server asymptotic regime:

- Halfin & Whitt ('81), Jennings, Mandelbaum, Massey & Whitt ('96), Puhalskii & Reiman ('00), Garnett, Mandelbaum & Reiman ('02), Borst, Mandelbaum & Reiman ('03), Armony & Maglaras ('03), Whitt ('03).

### Lead-Time Quotation

- Duenyas & Hopp ('95), Duenyas ('95), Plambeck ('01), Dobson & Pinker ('02).

## Why is this difficult?

---

- Job Sequencing:
  - Stochastic nature of the system
  - State-dependent arrivals
  - Age based information hard to analyze
- Waiting Time Estimation:
  - Depends on the sequencing policy
  - Depends on age information
  - Multiclass and state-dependent nature
    - ⇒ Waiting times depend on future arrivals, that depend on waiting time estimation, etc.

## Our Proposal

---

- **The Threshold sequencing policy**

$$\text{Give Service Priority to class } \begin{cases} 1 & \text{if } Q_2(t) < \theta(t) \\ 2 & \text{otherwise} \end{cases}$$

$$\theta(t) = \# \text{ of arrivals into queue 2 during } (t - D_2, t]$$

$$\approx \lambda_2 D_2$$

Properties: asymptotically

a. Compliant:  $W_2 \leq D_2$

b. Optimal:  $\min \{W_1 : W_2 \leq D_2\}$

- **The Snapshot waiting time estimator**

$$\hat{w}_1(t) = \frac{Q_1(t)}{\lambda_1}$$

Asymptotically consistent ( $\approx$  approximation becomes “correct”)

## The Many-Server Heavy-Traffic Regime of Halfin-Whitt (QED regime)

---

**System Load:**  $R = \frac{\lambda_1 + \lambda_2}{\mu}$ .

**[HW81]:** As  $N \rightarrow \infty$ ,

$$P(\text{wait}) \approx \alpha, \quad 0 < \alpha < 1 \quad \text{(Customers)}$$

$\Updownarrow$

$$\rho \approx 1 - \frac{\beta}{\sqrt{N}}, \quad \beta > 0 \quad \text{(Agents)}$$

$\Updownarrow$

$$N \approx R + \beta\sqrt{R}, \quad \beta > 0 \quad \text{(Manager)}$$

**Scaling relationships:**  $\frac{Q^N - N}{\sqrt{N}} \Rightarrow X$ , a well defined limit.

- Jobs in the system =  $\mathcal{O}(N)$
- Queue Lengths =  $\mathcal{O}(\sqrt{N})$
- Waiting Times =  $\mathcal{O}(1/\sqrt{N})$  ( $\Rightarrow D_2 = \mathcal{O}(1/\sqrt{N})$ ).

## Asymptotic Analysis of proposed policy

---

**Theorem:** The threshold sequencing policy and snapshot estimator are asymptotically:

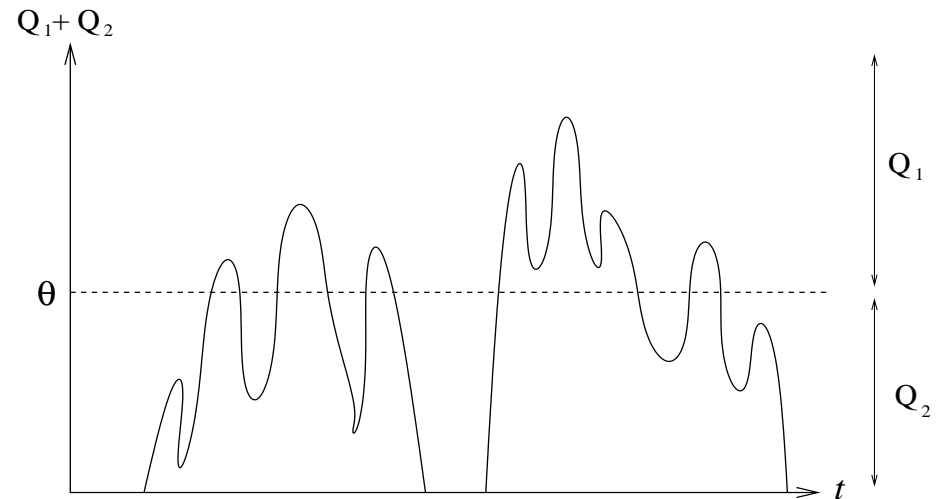
- a. Compliant:  $W_2 \leq D_2$
- b. Optimal:  $\min \{W_1 : W_2 \leq D_2\}$
- c. Consistent:  $W_1 = \hat{w}_1$

Technical details:

(1)  $\frac{Q_1^N + Q_2^N - N}{\sqrt{N}} \Rightarrow X$ , ( $X$  tractable diffusion),

(2)  $\frac{Q_1^N}{\sqrt{N}} \Rightarrow (X - \theta)^+$ ,  $\frac{Q_2^N}{\sqrt{N}} \Rightarrow X^+ \wedge \theta$ ,

(3)  $\sqrt{N}W_1^N \Rightarrow \frac{(X - \theta)^+}{\lambda_1}$ ,  $\sqrt{N}W_2^N \Rightarrow \frac{X^+ \wedge \theta}{\lambda_2}$ .



## The Diffusion Limit

---

**Proposition:** Suppose that  $\lambda_1^N(0, d_2) + \lambda_2^N(0, d_2) = N\mu - \delta\sqrt{N}\mu$ , then  $\frac{Z^N - N}{\sqrt{N}} \Rightarrow X$ , where  $X$  is a diffusion process with

$$dX(t) = [-\delta + f(X(t))] \mu dt + \sqrt{2\mu} dB(t),$$

where

$$f(x) = \begin{cases} \frac{\kappa}{\gamma} \frac{x-\theta}{\lambda_1}, & x \geq \theta \\ 0, & 0 \leq x < \theta \\ -x & x < 0 \end{cases} .$$

In addition,

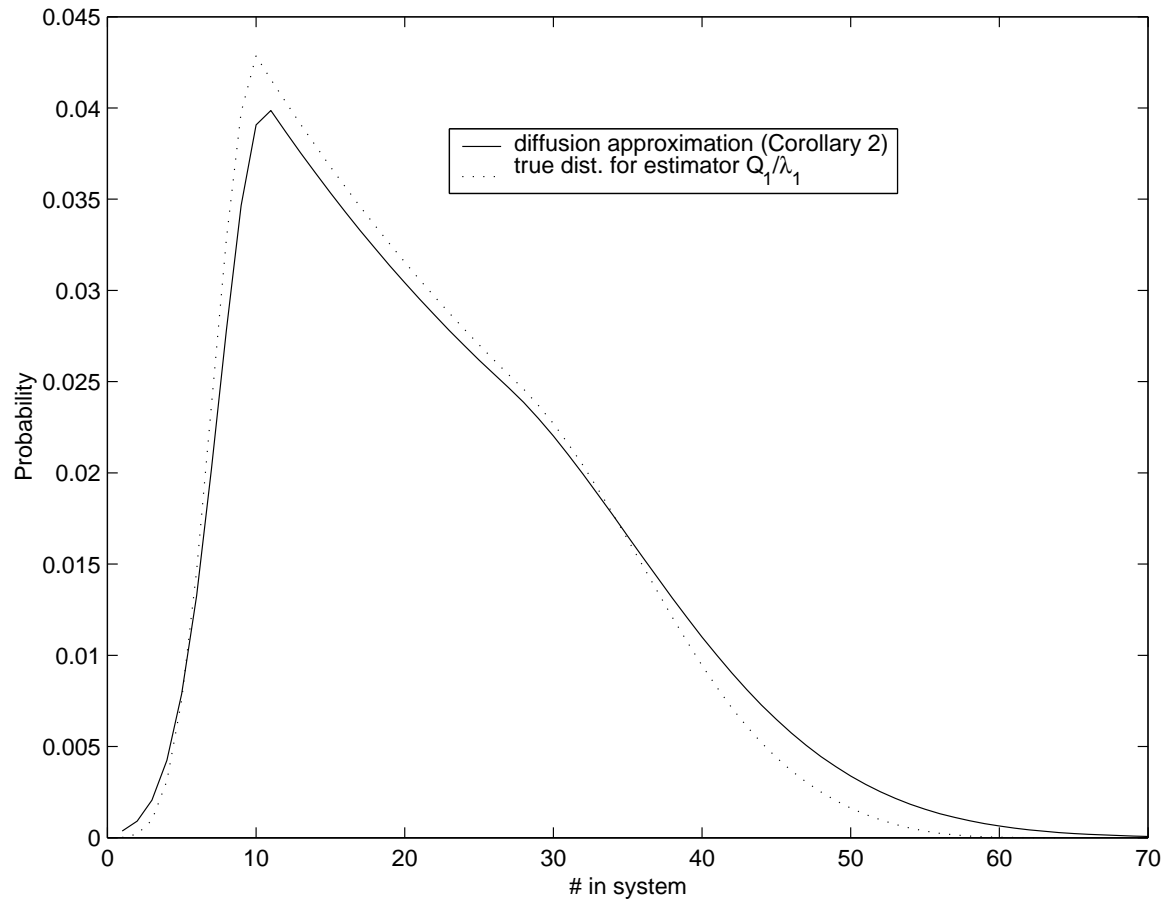
$$\sqrt{N}W_1^N \Rightarrow \frac{(X - \theta)^+}{\lambda_1}, \quad \sqrt{N}W_2^N \Rightarrow \frac{X^+ \wedge \theta}{\lambda_2}.$$

**Steady State Distribution:** has a simple form - with 3 parts:

1. Normal for  $x \geq \theta$ ,
2. Exponential for  $0 \leq x < \theta$ ,
3. Normal for  $x < 0$ .

## Performance Approximations of the Steady State Distribution

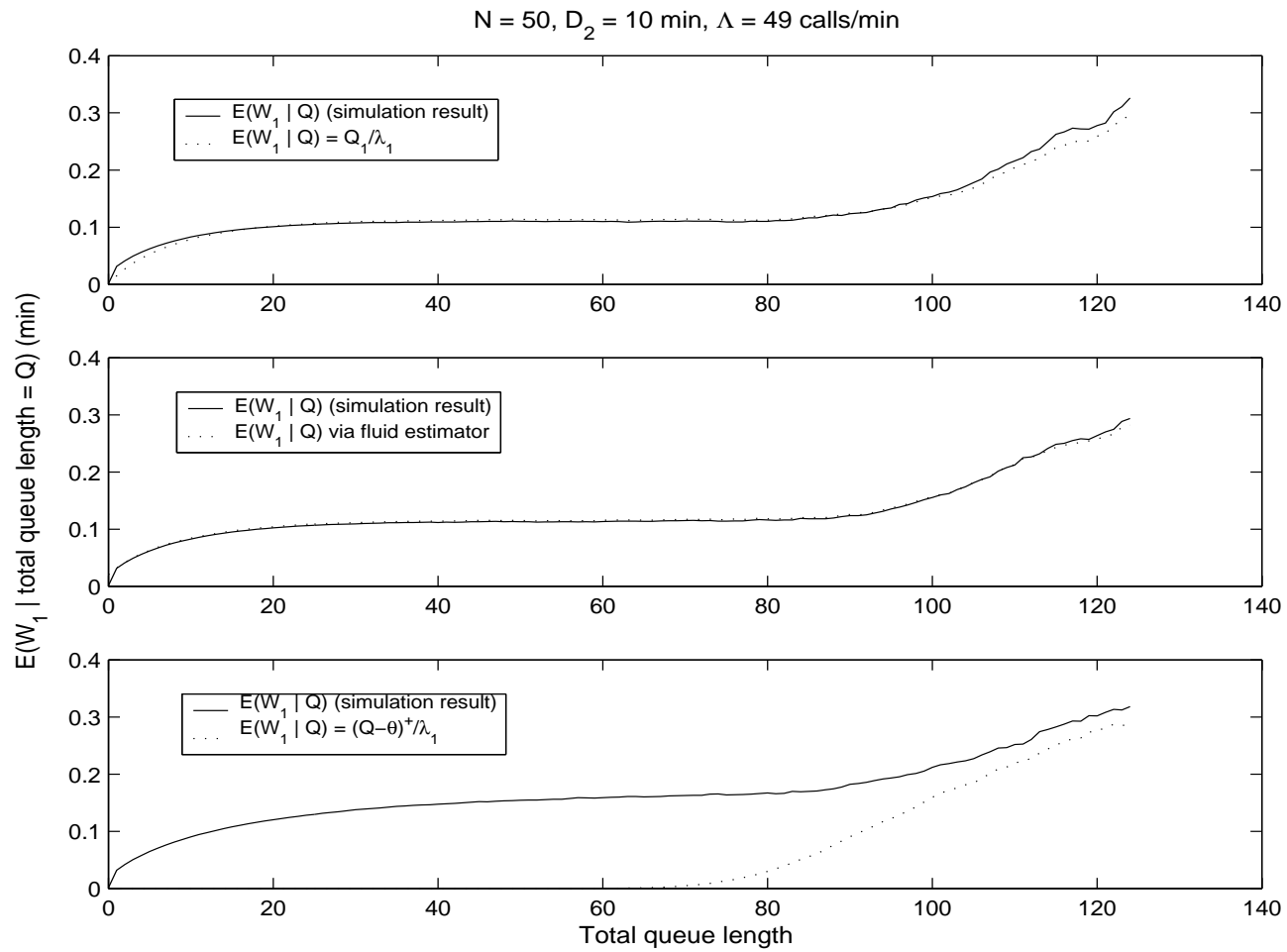
---



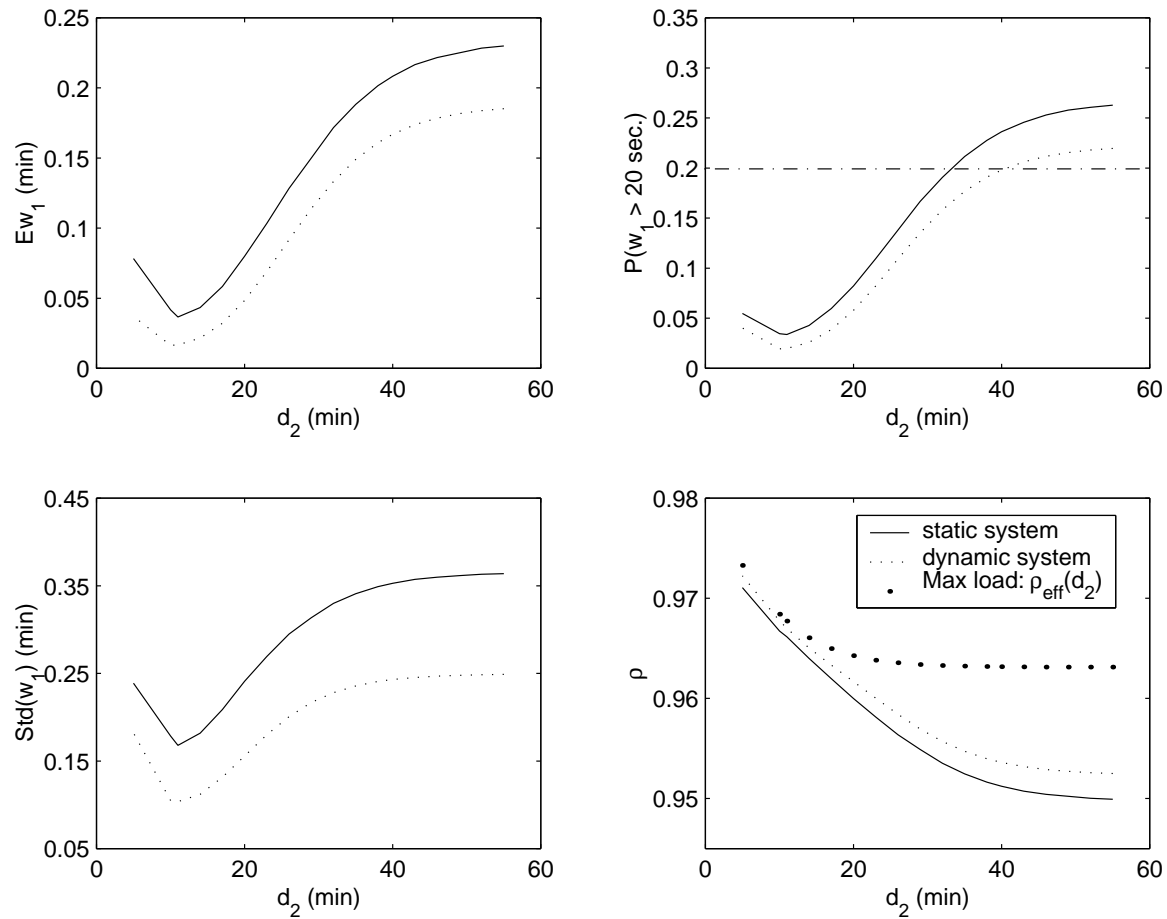
Steady state distribution for total queue length process (MNL choice model):

$$N = 10, \mu = 1, D_2 = 10, \rho_{\text{eff}}(D_2) = .97, r_1 = r_2 = 1, c_1 = .5, c_2 = .05, \nu = .3.$$

## Accuracy of waiting time estimator

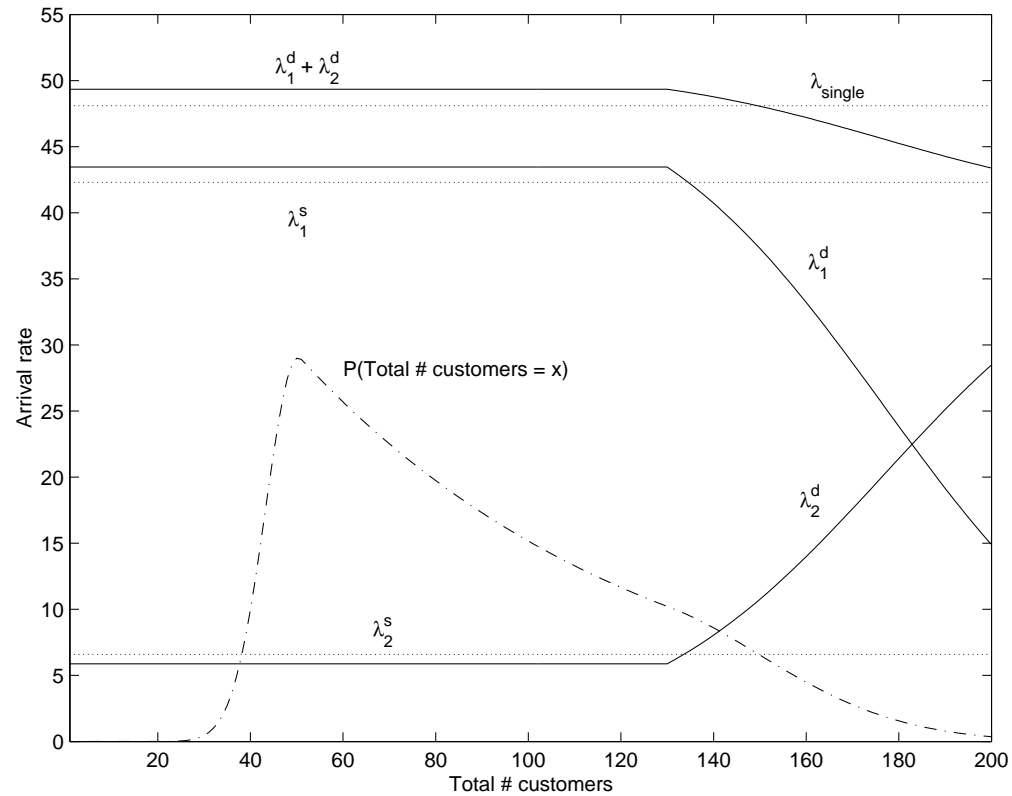


## The Value of Call-back Option and Delay Information



Steady state (static) versus state dependent (dynamic) information:  $N = 50$ ,  $\mu = 1$ ,  
MNL choice model with  $r_1 = r_2 = 1$ ,  $c_1 = .5$ ,  $c_2 = .05$ ,  $\nu = .3$  and  $\rho_{\text{eff}}(0) = .98$ .

## Load Balancing due to Call-Backs



Arrival rates vs. total number of customers for:  $N = 50$ ,  $\mu = 1$ , MNL choice model with  $r_1 = r_2 = 1$ ,  $c_1 = .5$ ,  $c_2 = .05$ ,  $\nu = .3$ ,  $\rho_{eff}(0) = 1.00$ ,  $D_2 = 12$  and  $\theta = 80$ .

## Staffing Rules

---

We would like to Minimize  $N$  subject to the following specifications:

- $EW_1 \leq \bar{w}$  (typically,  $\bar{w} = 10$  sec.),
- $P(w_1 \geq y) \leq \epsilon$  (typically,  $y = 20$  sec. and  $\epsilon = 20\%$ ),
- $P(\text{balking}) \leq \epsilon_b$  (typically,  $\epsilon_b = 1\%$ ).

**Square root staffing rules** apply of the form:

$$N(\bar{w}, y, \epsilon, \epsilon_b) = R + x^* \sqrt{R}$$

where

- $R = \Lambda/\mu$
- $x^*$  is easily computed from the diffusion limit

## Concluding Remarks

---

- Simple modelling framework for customer contact centers that
  - Offer a call-back option with a guaranteed deadline
  - Inform customers of their anticipated delay
- Threshold sequencing policy is asymptotically compliant and optimal
- Snapshot waiting time estimator is asymptotically consistent
- Diffusion approximation simplifies performance analysis
- Design considerations: square root staffing rules apply

### Extensions:

- Non-stationary arrival rates
- What information to announce to callers?