

# Design and Control of Large Call Centers: Asymptotic Analysis via Two-scale Fluid Limits

---

**Assaf Zeevi**

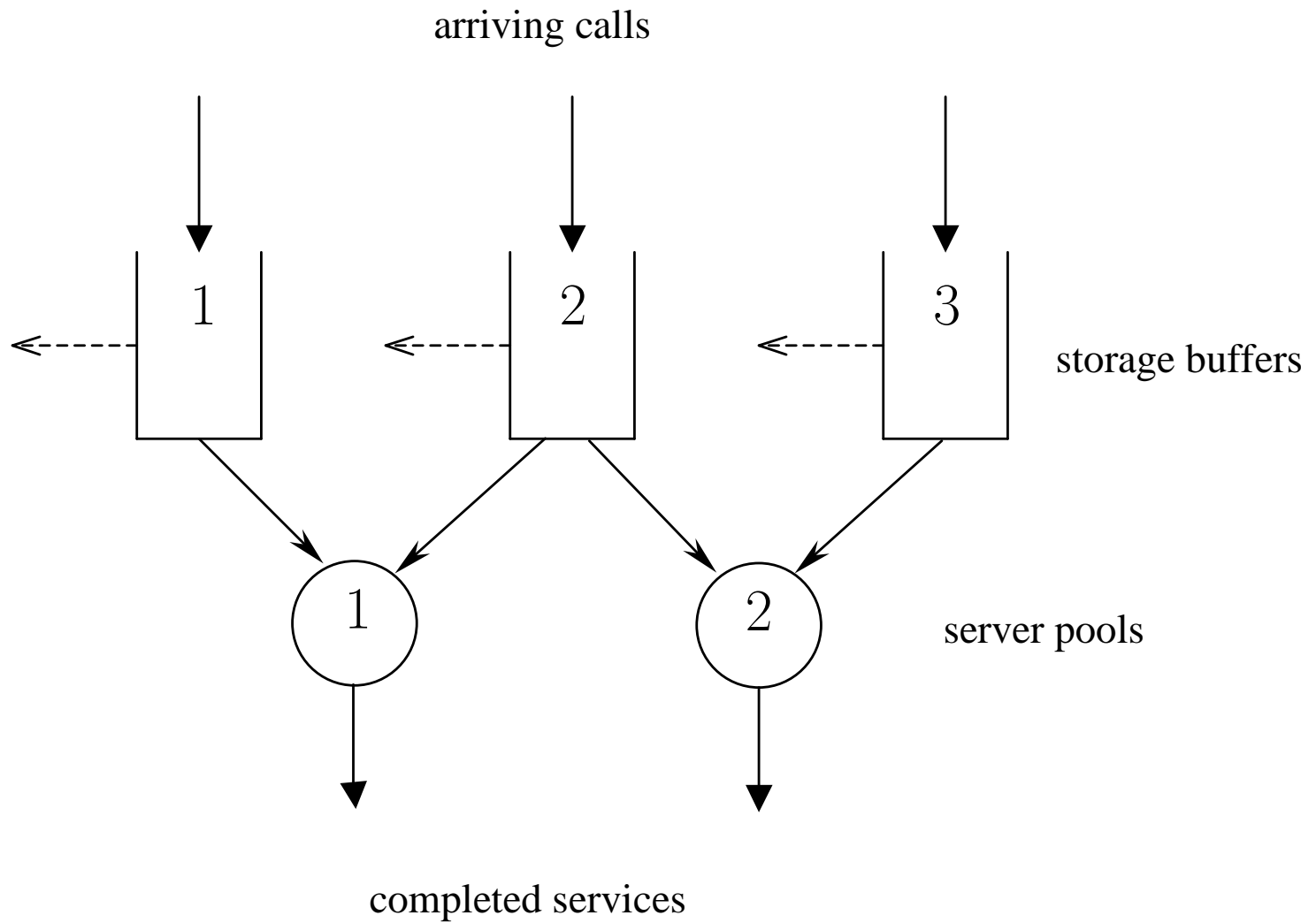
Graduate School of Business  
Columbia University

May 2004

\* based on joint work with Mike Harrison and Achal Bassamboo  
(GSB, Stanford University)

# Call center model

---



## Design objectives

---

- I. **set staffing level for each agent pools**
- II. assign agents to work schedules (workforce scheduling)
- III. **dynamically route calls to agents**

Typical goal:

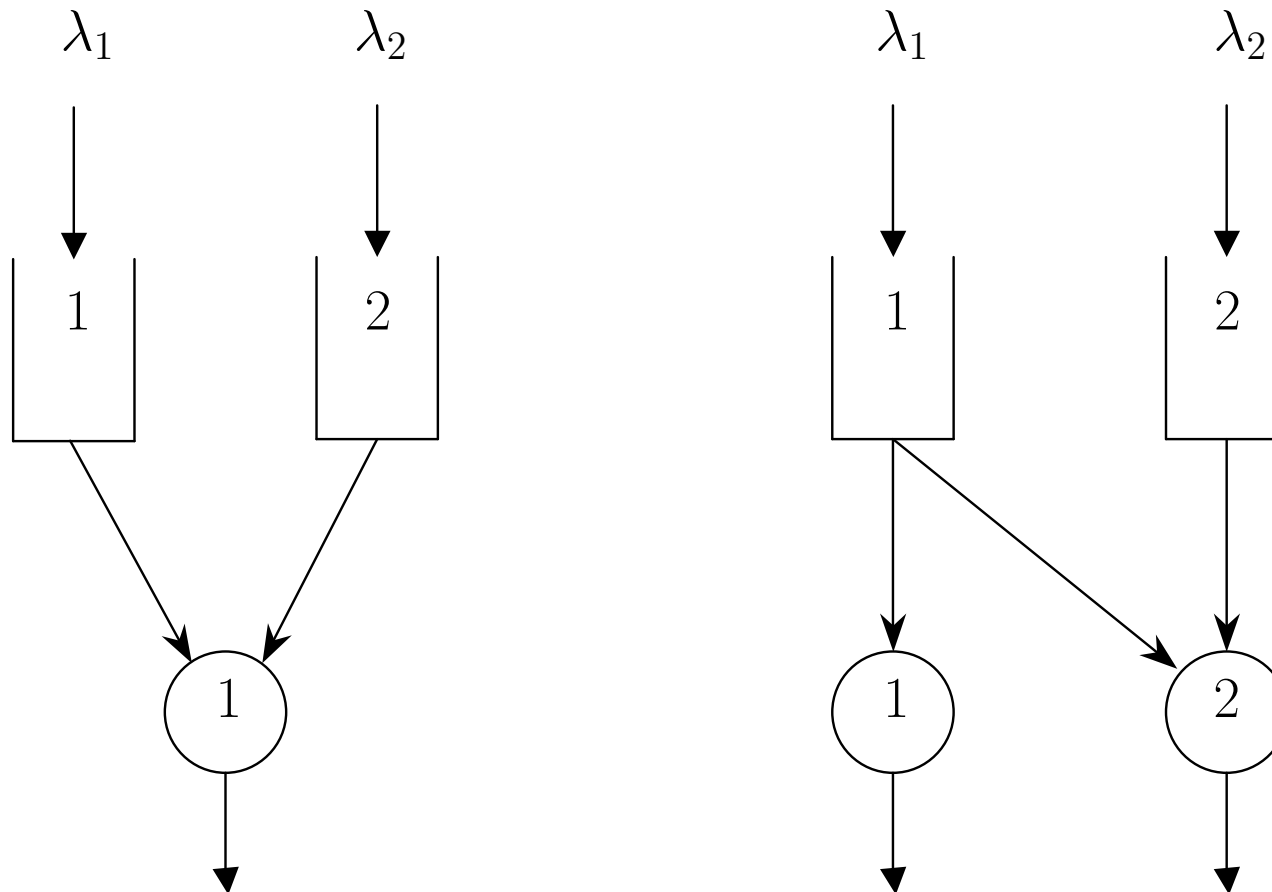
minimize **staffing cost**

subject to **quality-of-service** [waiting/abandonments etc]

solving **staffing problem** hinges on **routing policy** !

## More on the routing problem...

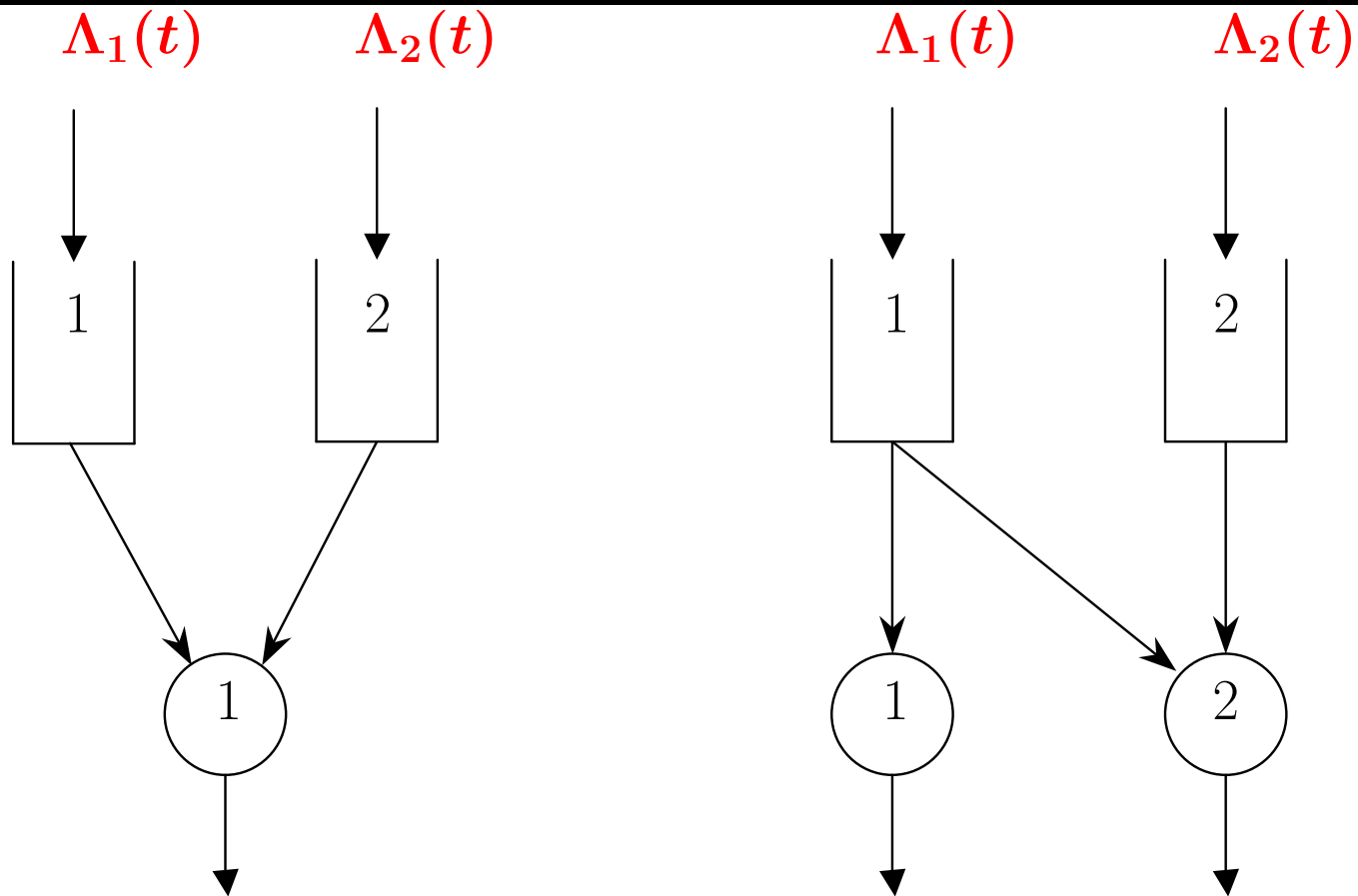
---



common tool to solve control problem: **diffusion approximations...**

## What if arrival rates not constant/deterministic?

---



performance analysis tools: **fluid limits, uniform acceleration**

## Our work

---

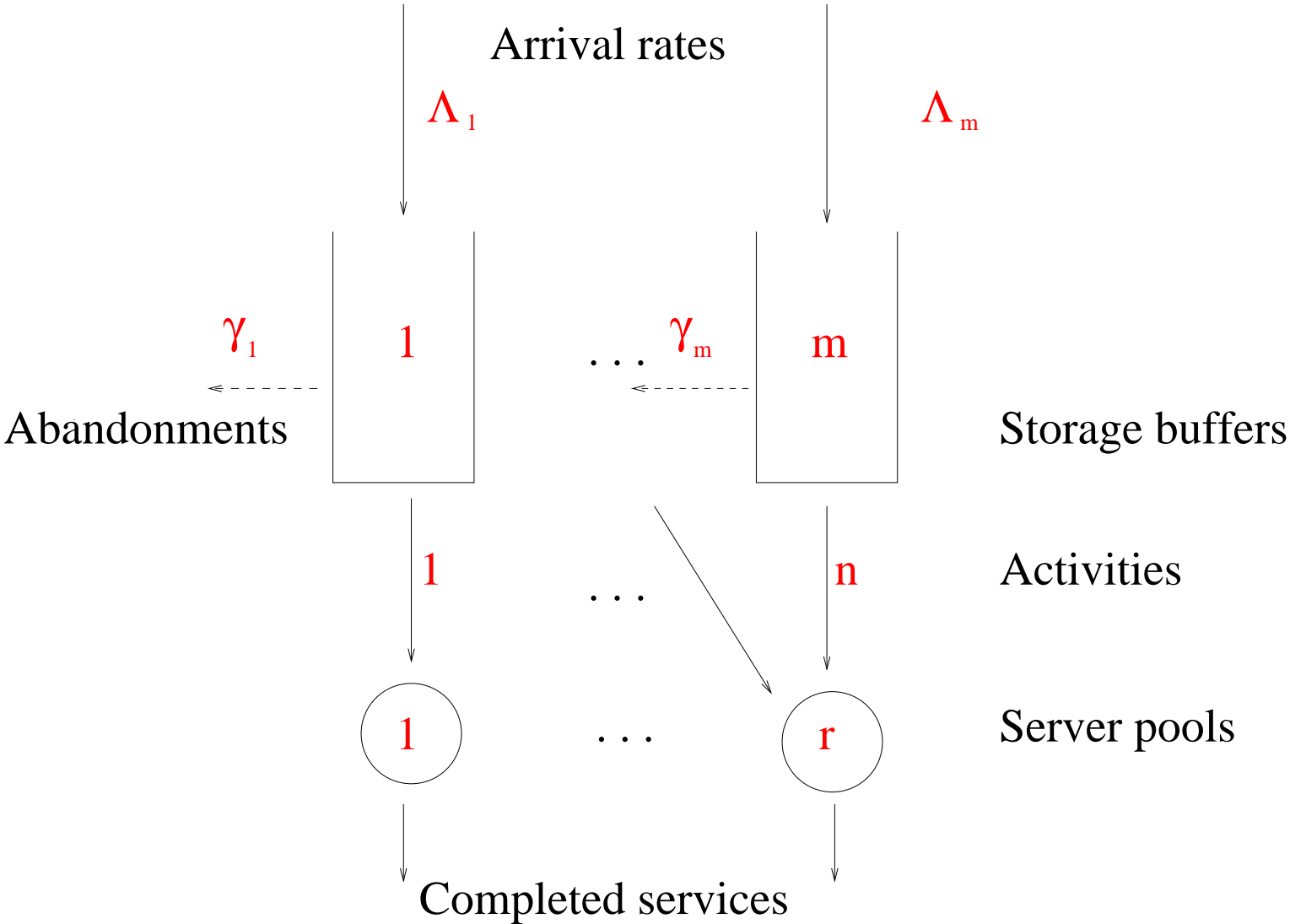
**Main theme:** how to solve **design** and **control** problems in **large service systems** when demand is **time-varying + uncertain**

### **Key ingredients:**

- ▶ formulate “reasonable” objective
- ▶ focus on temporal variations and uncertainty in arrivals
  - queueing fluctuations are negligible in comparison...
- ▶ derive approximate solutions for staffing and routing
- ▶ prove solutions are nearly optimal...

# Call center model

---



## System state dynamics

---

$$Z_i(t) = \begin{bmatrix} \text{Arrivals} \\ \text{rate: } \Lambda_i(t) \end{bmatrix} - \begin{bmatrix} \text{Completed Services} \\ \text{rate: } (RX)_i(t) \end{bmatrix} - \begin{bmatrix} \text{Abandonments} \\ \text{rate: } \gamma_i Q_i(t) \end{bmatrix}$$

- ▶  $Z$  : **headcount process**

$Z_i(t) = \#$  of class  $i$  customers present at time  $t$

- ▶  $Q$  : **queue length process**

$Q_i(t) = \#$  of class  $i$  customers not being served at time  $t$

- ▶  $X$  : **dynamic control** [ $\#$  of servers allocated to each activity]

$(RX)_i =$  rate of service in class  $i$

- ▶  $(X, Z, Q)$  satisfy

$$AX(t) \leq b, \quad Q(t) = Z(t) - BX(t) \geq 0, \quad Z(t) \geq 0, \quad X(t) \geq 0$$

- ▶  $b$  : **staffing vector**

## Objective function

---

- ▶ Minimize sum of personnel cost + expected abandonment penalties

$$\mathcal{J}(X, b) := \underbrace{c \cdot b}_{\text{staffing cost}} + \underbrace{\mathbb{E} \left[ \sum_{i=1}^m p_i \int_0^T \gamma_i Q_i(s) ds \right]}_{\text{expected "lost business"}}$$

–  $c$  : personnel cost vector

–  $p$  : penalty cost vector

- ▶ **objective:** minimize  $\mathcal{J}(X, b)$

over staffing level ( $b$ ) and admissible control ( $X$ ).

- ▶ **interpretation of  $\mathcal{J}(X, b)$ :**  $p$  dualizes service level constraints...

## Asymptotic analysis: Fluid limits

---

### Fluid limits:

- ▶ accelerate arrivals by  $\kappa$
- ▶ scale up number of servers by  $\kappa$
- ▶ normalize state and control processes by  $\kappa$

$$\kappa^{-1}(\bar{Z}^\kappa, \bar{Q}^\kappa) \rightarrow (\bar{Z}, \bar{Q}), \text{ a.s. as } \kappa \rightarrow \infty$$

where  $\bar{Z}$  solves

$$\frac{d\bar{Z}_i(t)}{dt} = \underbrace{\Lambda_i(t)}_{\text{arrival rate}} - \underbrace{(R\bar{X})_i(t)}_{\text{service rate}} - \underbrace{\gamma_i \bar{Q}_i(t)}_{\text{abandonment rate}}$$

**Fluid limit equation is not easy to solve...**

## Asymptotic analysis: Two-scale fluid limits

---

### primitives:

- ▶ Service rates **accelerated linearly** :  $R^\kappa = \kappa R$
- ▶ Abandonment rates **accelerated linearly** :  $\gamma_i^\kappa = \kappa \gamma_i$
- ▶ Arrival rates **grow faster** than service rates;

$$\Lambda^\kappa(\cdot) = g(\kappa)\kappa\Lambda(\cdot)$$

$g(x)$  grows to infinity as  $x \rightarrow \infty$

- ▶ # servers grows so that **processing capacity scales like arrival rate**

### objective fn:

- ▶ Cost per server **accelerated linearly** :  $c^\kappa = \kappa c$   
keeps cost of capacity constant...
- ▶ All other parameters (matrices  $A, B$  and penalty vector  $p$ ) remain constant

## Interpretation of scaling

---

- ▶ linear scaling of service + arrival rate  $\Rightarrow$  time acceleration

*uniform acceleration*

- ▶ “extra acceleration” of arrival rate  $\Rightarrow$

need to increase # of servers w/o bound (to match arrivals)

- ▶ acceleration of renegeing rate  $\Rightarrow$  losses are significant

**system behavior:** scale grows, equilibrates rapidly,  
“loss-like dynamics”

## Two-scale fluid limits

---

**Theorem.** For any sequence of staffing vectors  $\{b^\kappa\}$  and corresponding admissible controls  $\{X^\kappa\}$

**normalized headcount and queuelength converge:**

$$\frac{Z^\kappa}{g(\kappa)} \rightarrow Z, \quad \frac{Q^\kappa}{g(\kappa)} \rightarrow Q, \quad \text{as } \kappa \rightarrow \infty$$

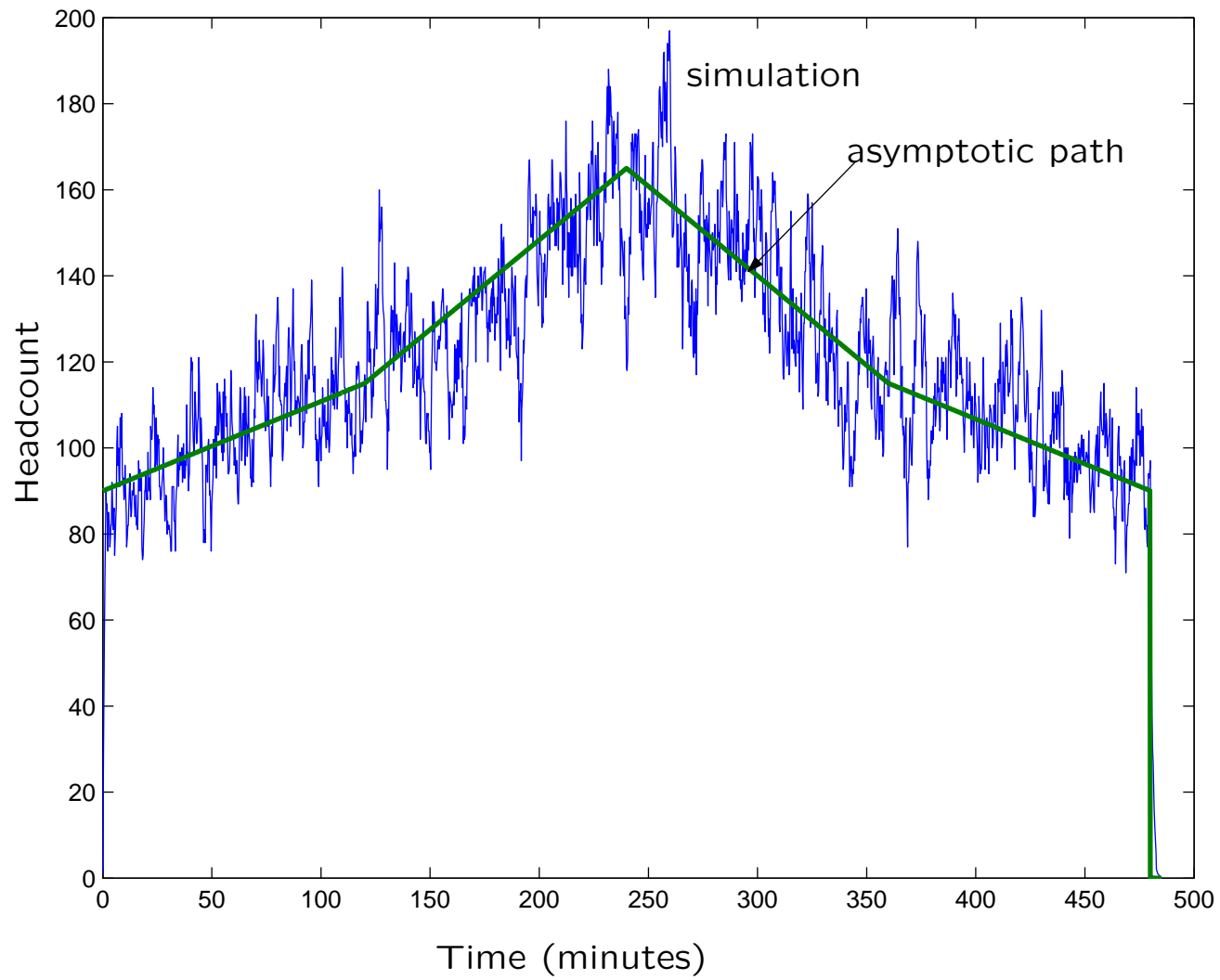
**limiting dynamics:**

$$\gamma_i Q_i(t) = \Lambda_i(t) - (RX)_i(t)$$

- ▶ abandonment loss rate = arrival loss rate...
- ▶ simple limiting dynamics, system equilibrates instantaneously!

## Picture proof...

---



## Proposed LP-based staffing method

---

- ▶ **Staffing level:** chose  $b_*$  to

$$\text{minimize} \quad c \cdot b + \mathbb{E} \left[ \int_0^T \pi(\Lambda(t), b) dt \right]$$

- ▶  $\pi(\lambda, b)$  is the optimal value of the **LP**

$$\min \quad p \cdot (\lambda - Rx)$$

$$\text{s.t.} \quad Rx \leq \lambda, \quad Ax \leq b, \quad x \geq 0$$

- ▶ **Performance estimate:**

$$\mathcal{J}_* = c \cdot b_* + \mathbb{E} \left[ \int_0^T \pi(\Lambda(t), b_*) dt \right]$$

**Q.** Why is this reasonable?

**Q.** What is the quality of  $\mathcal{J}_*$  as performance estimate?

## Lower bound on system performance

---

**Theorem.** For any sequence of staffing vectors  $\{b^\kappa\}$  and corresponding admissible dynamic controls  $\{X^\kappa\}$ ,

$$\liminf_{\kappa \rightarrow \infty} (\kappa g(\kappa))^{-1} \mathcal{J}^\kappa(X^\kappa, b^\kappa) \geq \underbrace{c \cdot b_* + \mathbb{E} \left[ \int_0^T \pi(\Lambda(t), b_*) dt \right]}_{\text{Performance Estimate } \mathcal{J}_*}$$

►  $\pi(\lambda, b)$  is the optimal value of server allocation LP

Q. Can we construct policies  $X_*^\kappa$  and  $b_*^\kappa$  such that  $\mathcal{J}^\kappa(X_*^\kappa, b_*^\kappa) \approx \kappa g(\kappa) \cdot \mathcal{J}_*$ ?

# Achieving the lower bound: Asymptotically optimal staffing/routing

---

**Prescription.** (when arrival rate is known...)

▶ **Staffing vector** : in  $\kappa^{th}$  system

$$b_*^\kappa = g(\kappa)b_*$$

▶ **Dynamic control** : allocate  $X_*^\kappa(t)$  servers at time  $t$  according to

$$\text{server allocation LP} \begin{cases} \min & p \cdot (\Lambda^\kappa(t) - Rx) \\ \text{s.t.} & R^\kappa x \leq \Lambda^\kappa(t), \quad Ax \leq b_*^\kappa, \quad x \geq 0 \end{cases}$$

**Theorem.**  $\{X_*^\kappa\}, \{b_*^\kappa\}$  are *asymptotically optimal*

**Q.** What happens when the arrival rates are not known?

## $\Lambda$ -tracking policies

---

- ▶ Estimate the arrival rate :  $\hat{\Lambda}$
- ▶ “Plug in” estimate for dynamic control :

$\hat{X}_*^\kappa(t)$  is a solution of the LP

$$\min p \cdot (\hat{\Lambda}^\kappa(t) - Rx)$$

$$\text{s.t. } R^\kappa x \leq \hat{\Lambda}^\kappa(t), Ax \leq b_*^\kappa, x \geq 0.$$

**Q.** When is a  $\Lambda$ -tracking control asymptotically optimal?

## Asymptotic optimality of $\Lambda$ -tracking policies

---

**Theorem.** If estimator is **uniformly consistent**, then sequence of  $\Lambda$ -tracking controls  $\{\hat{X}_*^\kappa\}$ , together with the staffing vectors  $\{b_*^\kappa\}$ , is asymptotically optimal.

► **Uniform consistency**

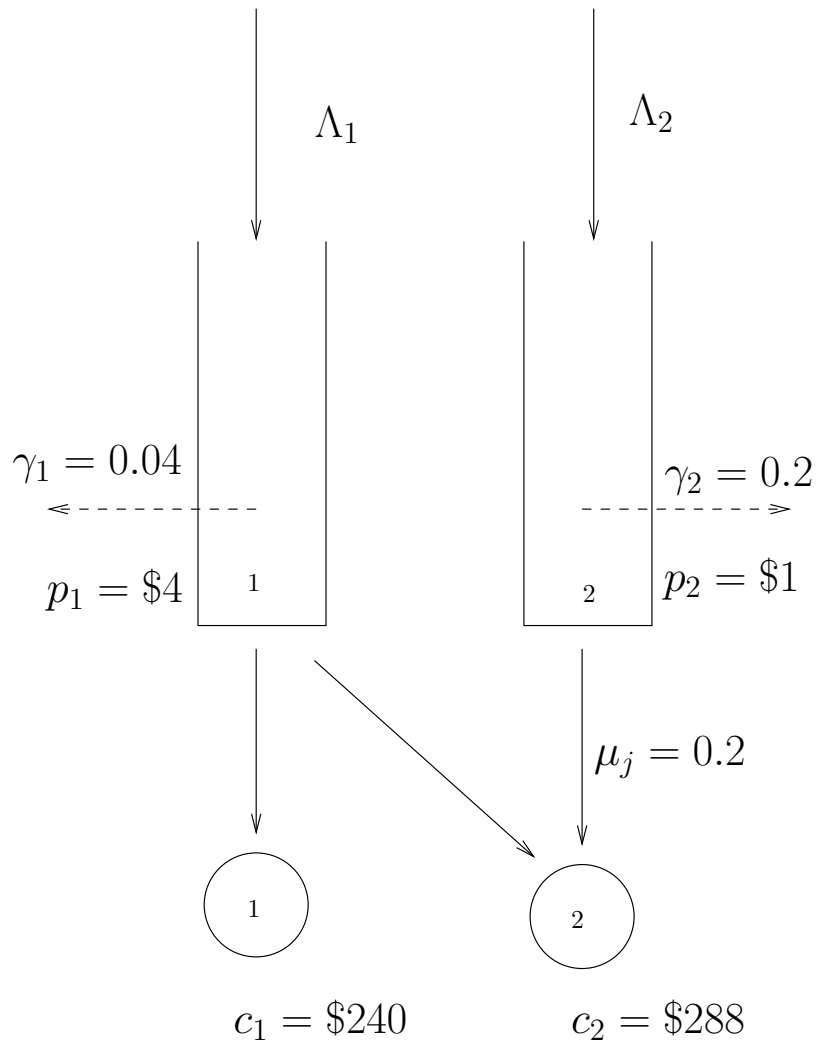
$$\frac{\hat{\Lambda}^\kappa(\cdot)}{\kappa g(\kappa)} \rightarrow \Lambda(\cdot) \quad \text{a.s., as } \kappa \rightarrow \infty, \text{ [uniformly]}$$

► Uniform consistency of estimator  $\Rightarrow$

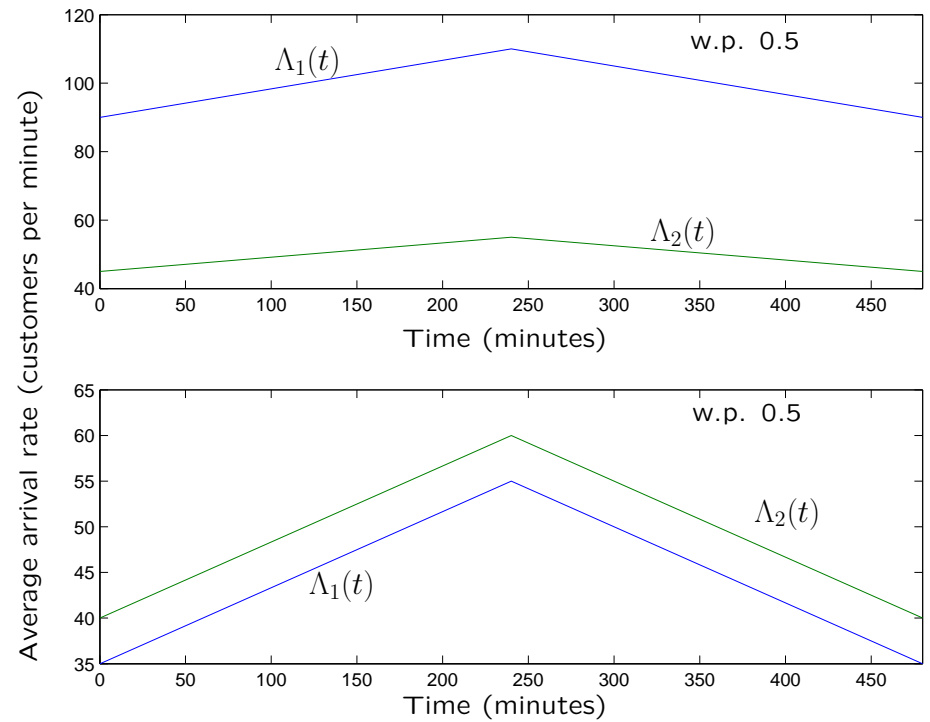
Solutions of LP w/ estimator  $\approx$  Solution of LP w/ actual arrival rates

estimated controls  $\approx$  controls when  $\Lambda$  is known

# Simulation Results: Asymptotic Optimality

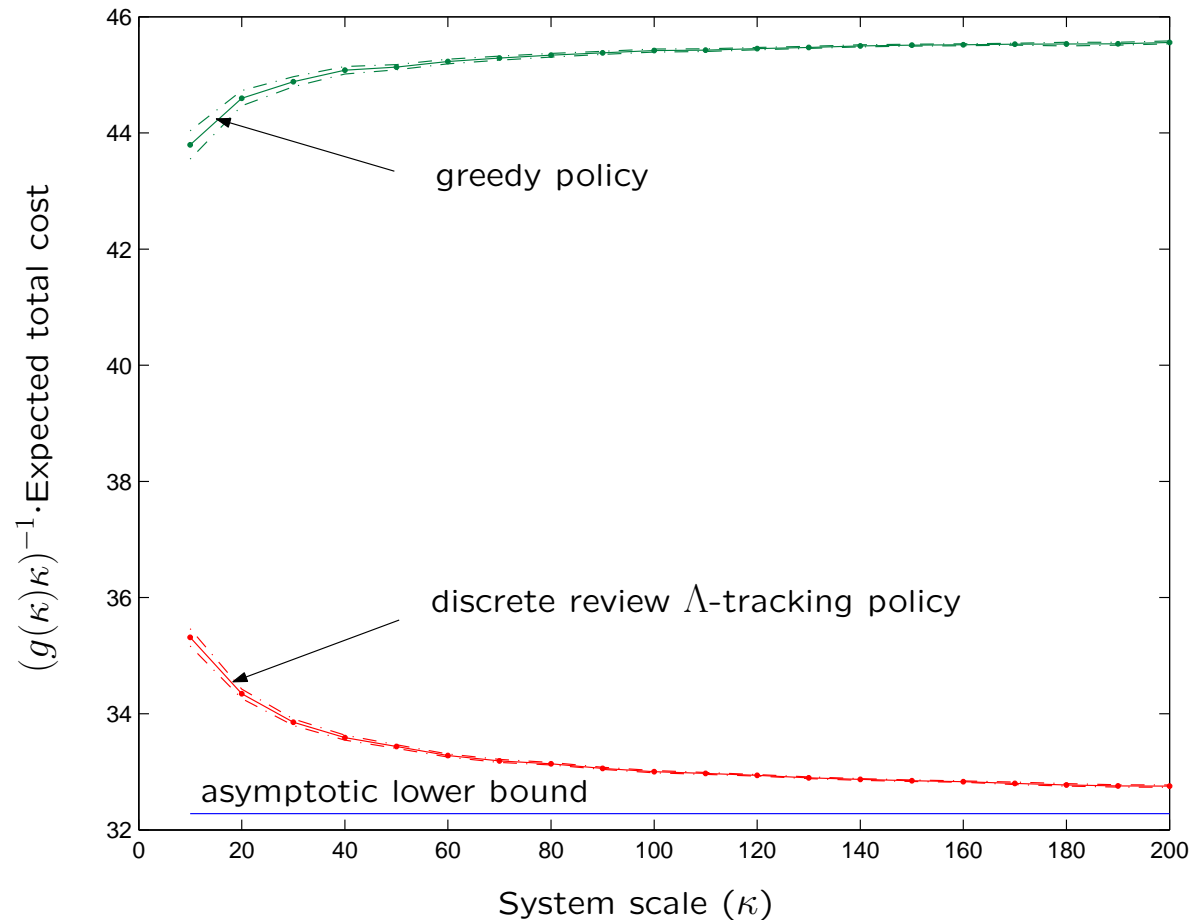


**arrival rate:**



## Simulation Results: Asymptotic Optimality (Continued)

- ▶ Sequence of systems with  $g(\kappa) = \kappa$  (= “extra acceleration”)
- ▶ Scaled expected total cost from simulations:



## Takeaway message

---

- ▶ asymptotic regime
  - **de-emphasizes queueing dynamics** [“small order” fluctuations]
- ▶ objective function
  - “dualizes” QoS constraints...
  - simple LP w/recourse [major reduction in complexity]
  - does not require analysis of routing...
- ▶ comments on method
  - fluid-scale objective fn provides performance estimate
  - can be used in **multi-class/multi-pool** systems
  - accounts for **temporal variations** and **uncertainty** in arrival rates