

# Staffing Call Centers with Differentiated Levels of Service

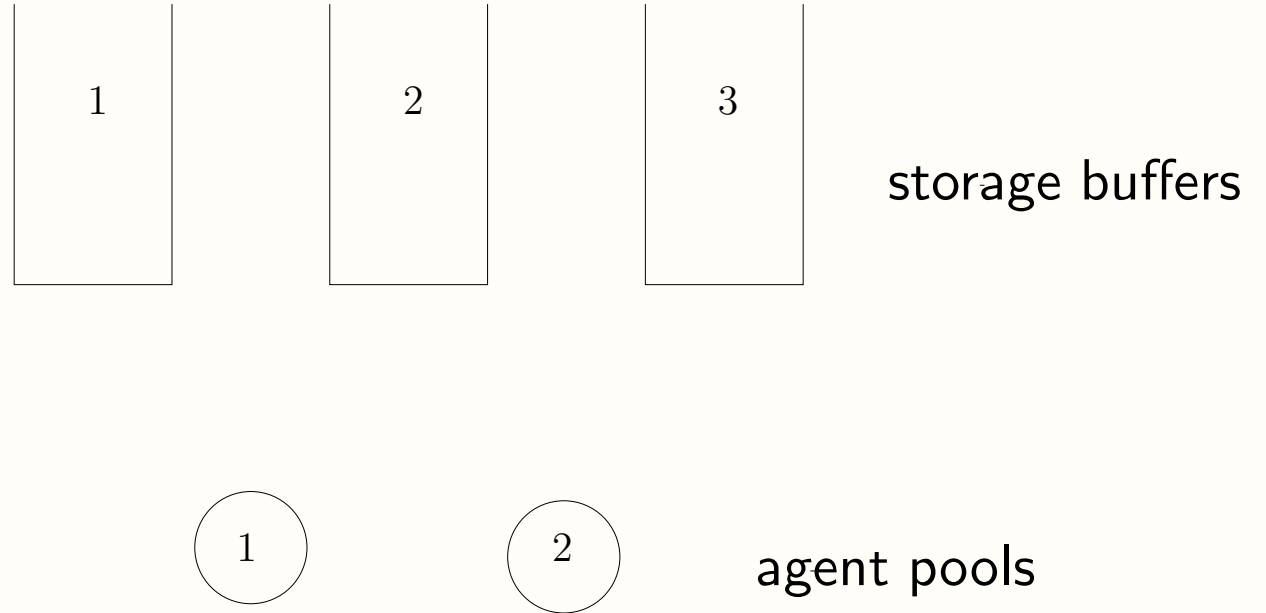
## Approximate Solutions via Constraint Dualization

Achal Bassamboo\* & Assaf Zeevi<sup>‡</sup>

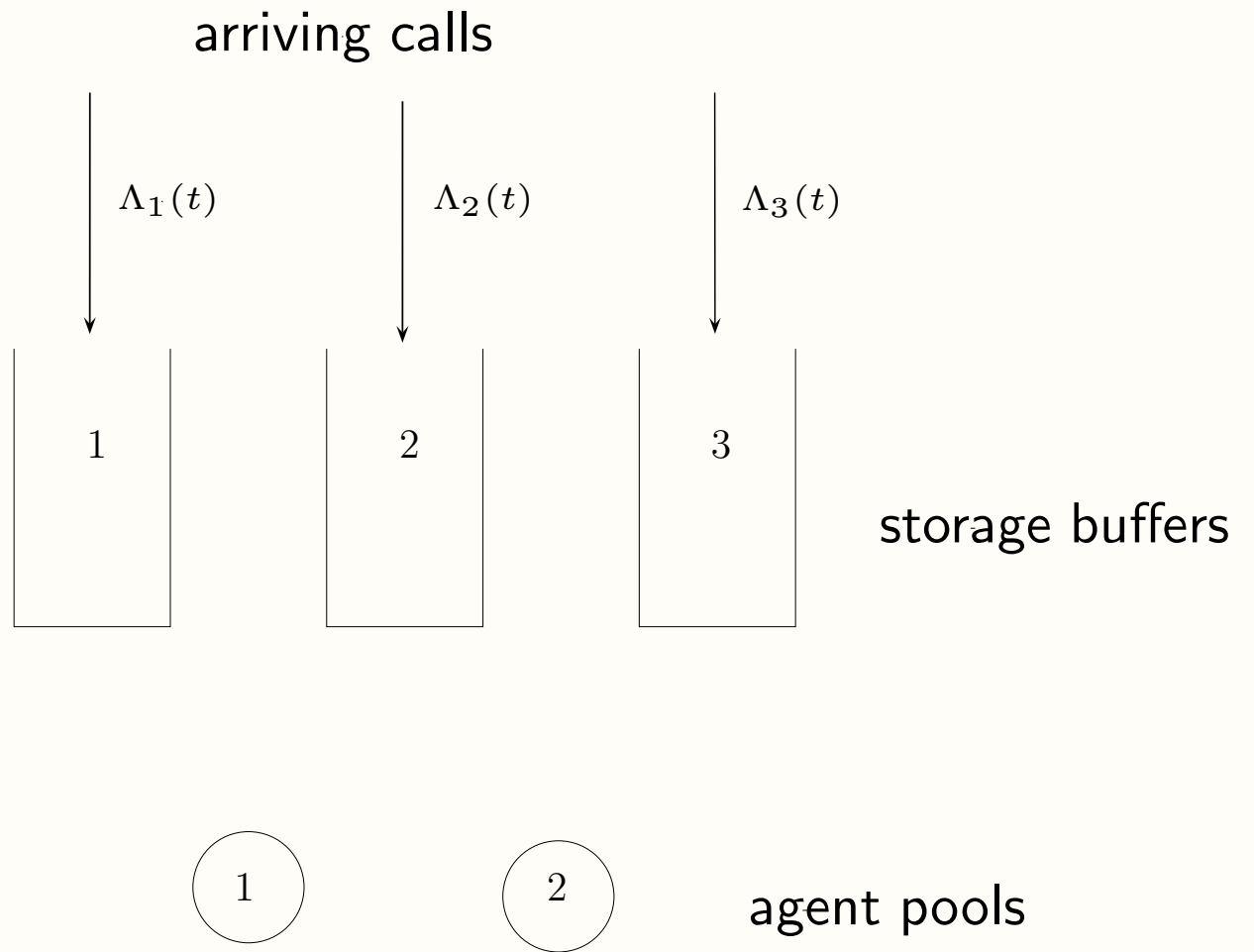
February, 2008

\* Kellogg School of Management, Northwestern University    <sup>‡</sup> GSB, Columbia University

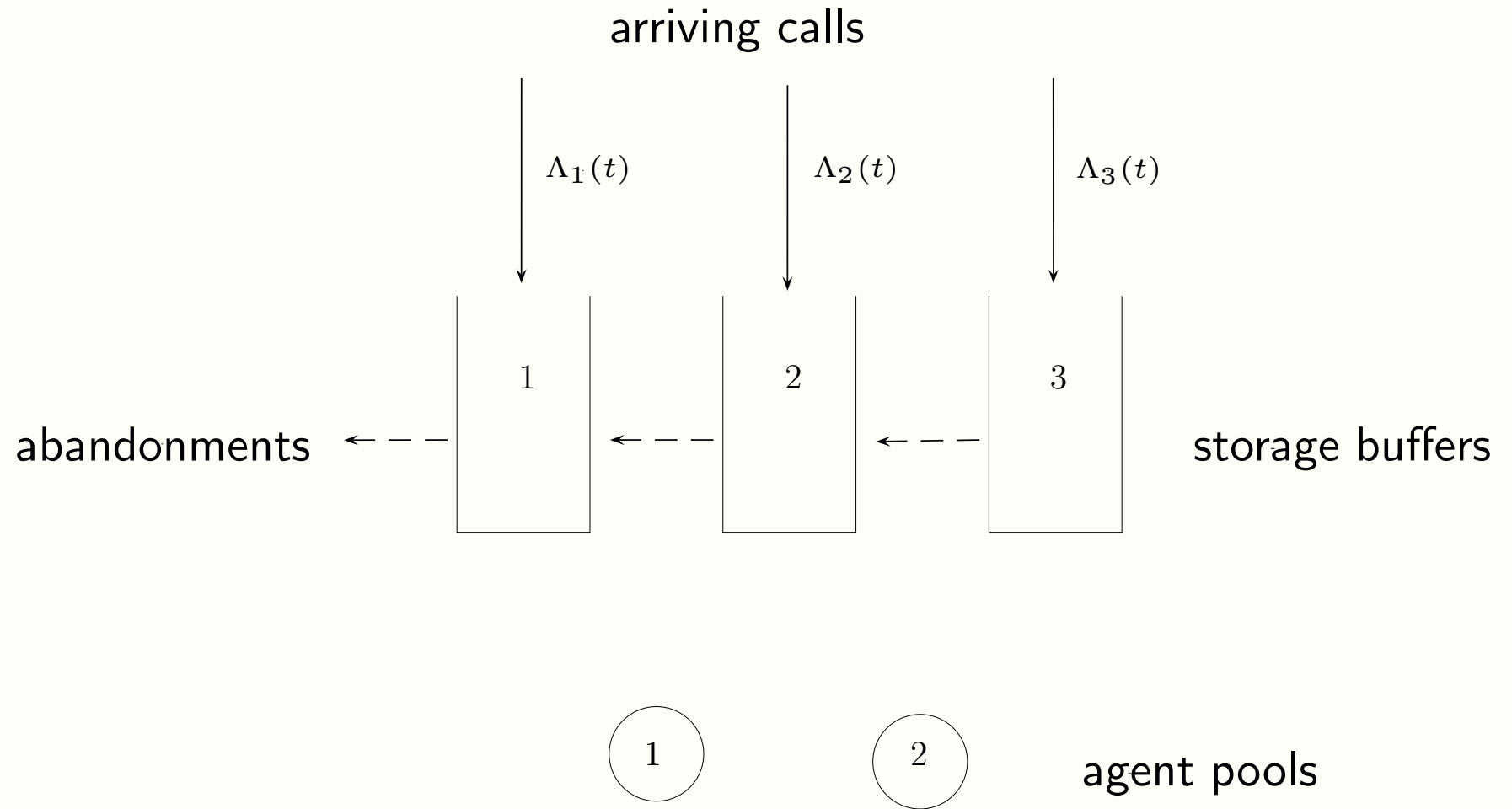
# System model



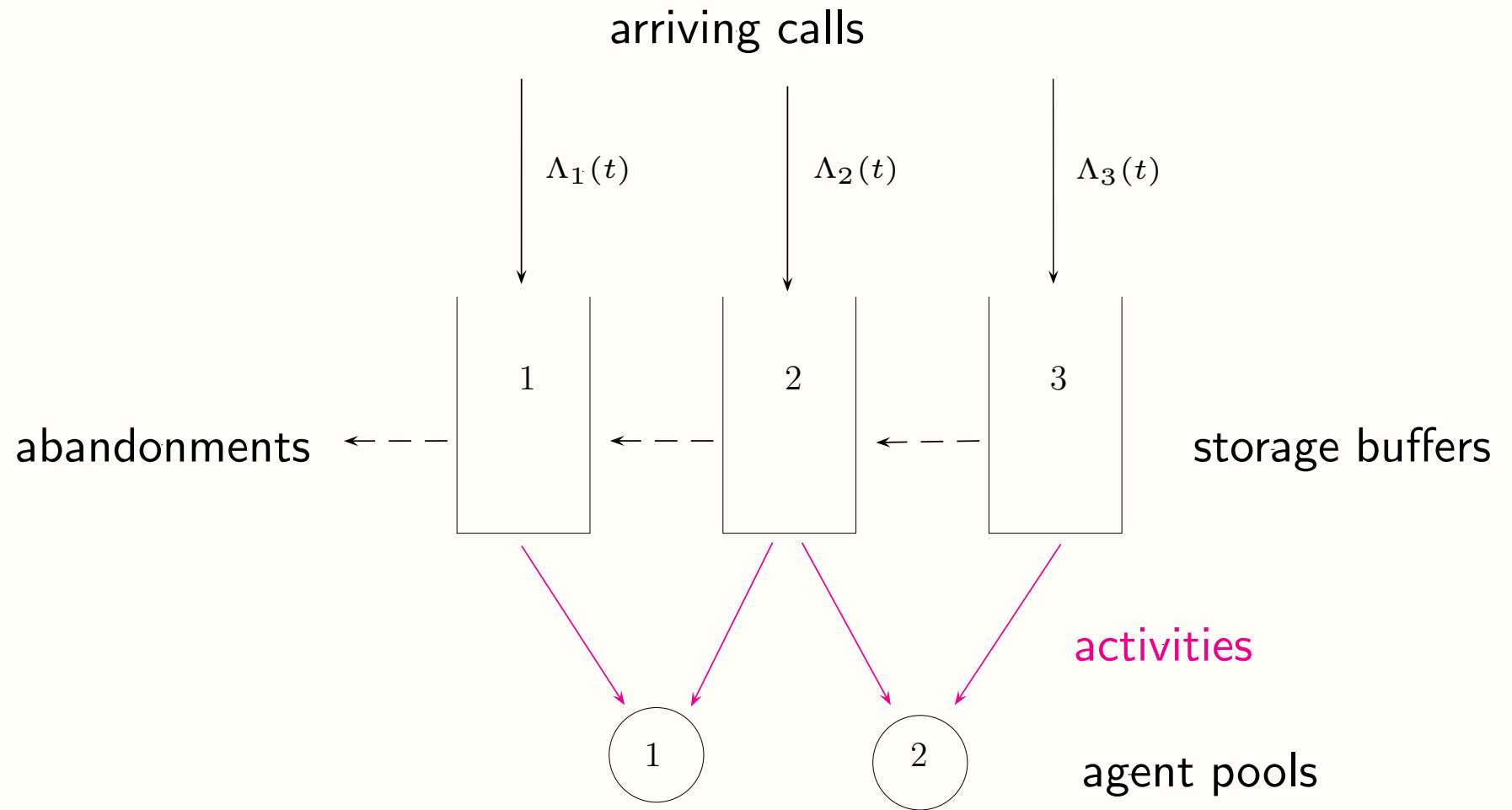
# System model



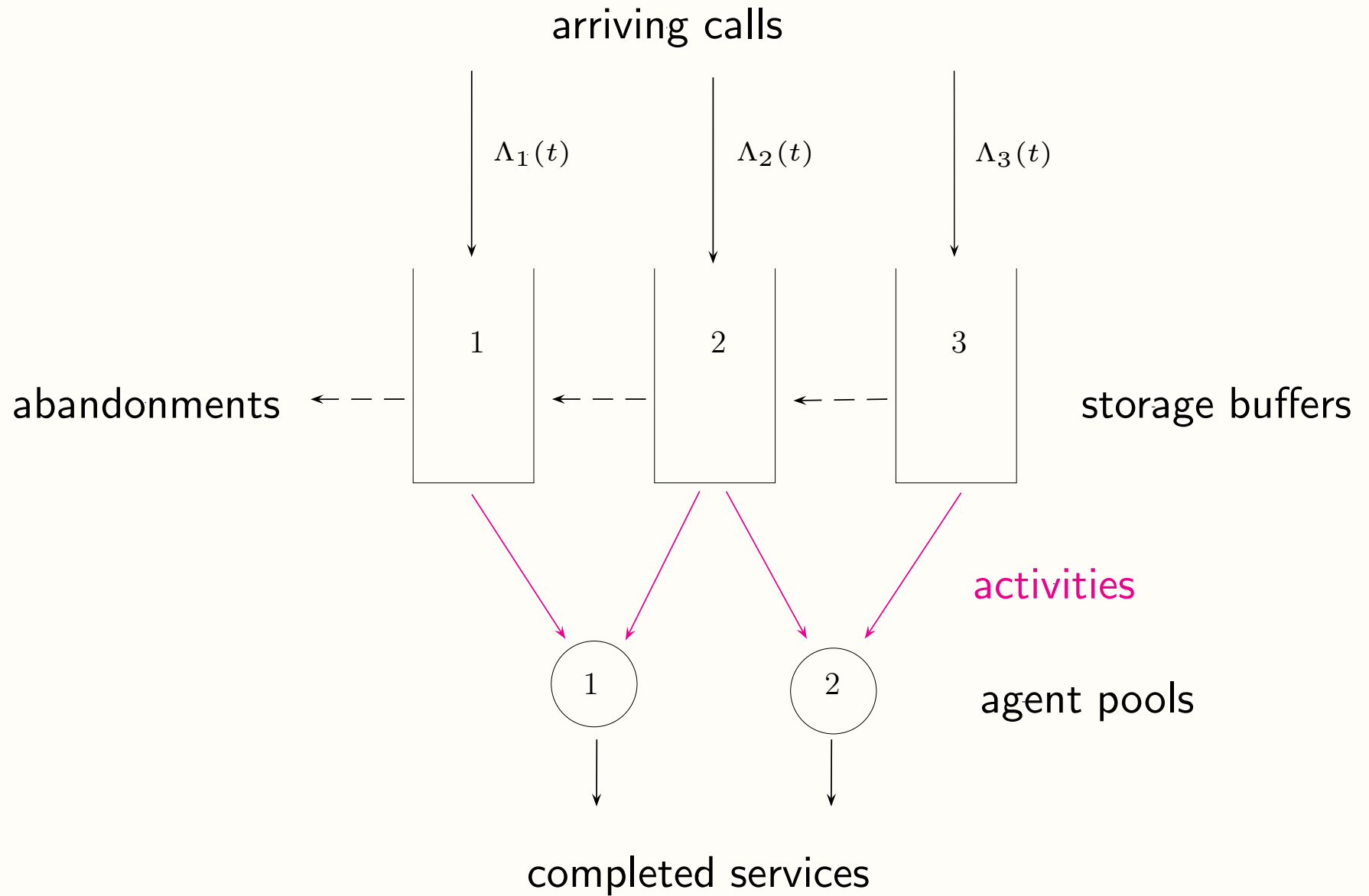
# System model



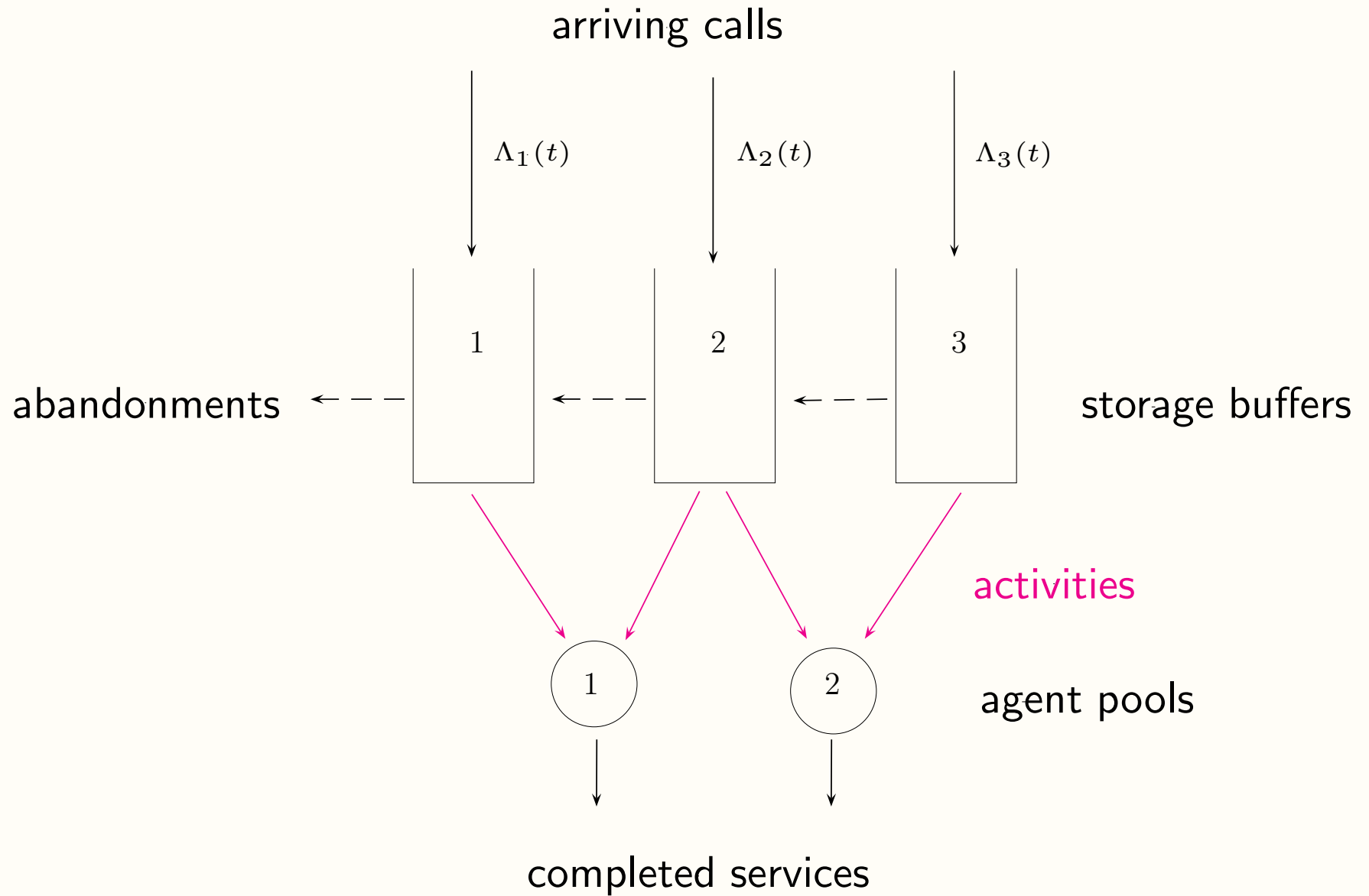
# System model



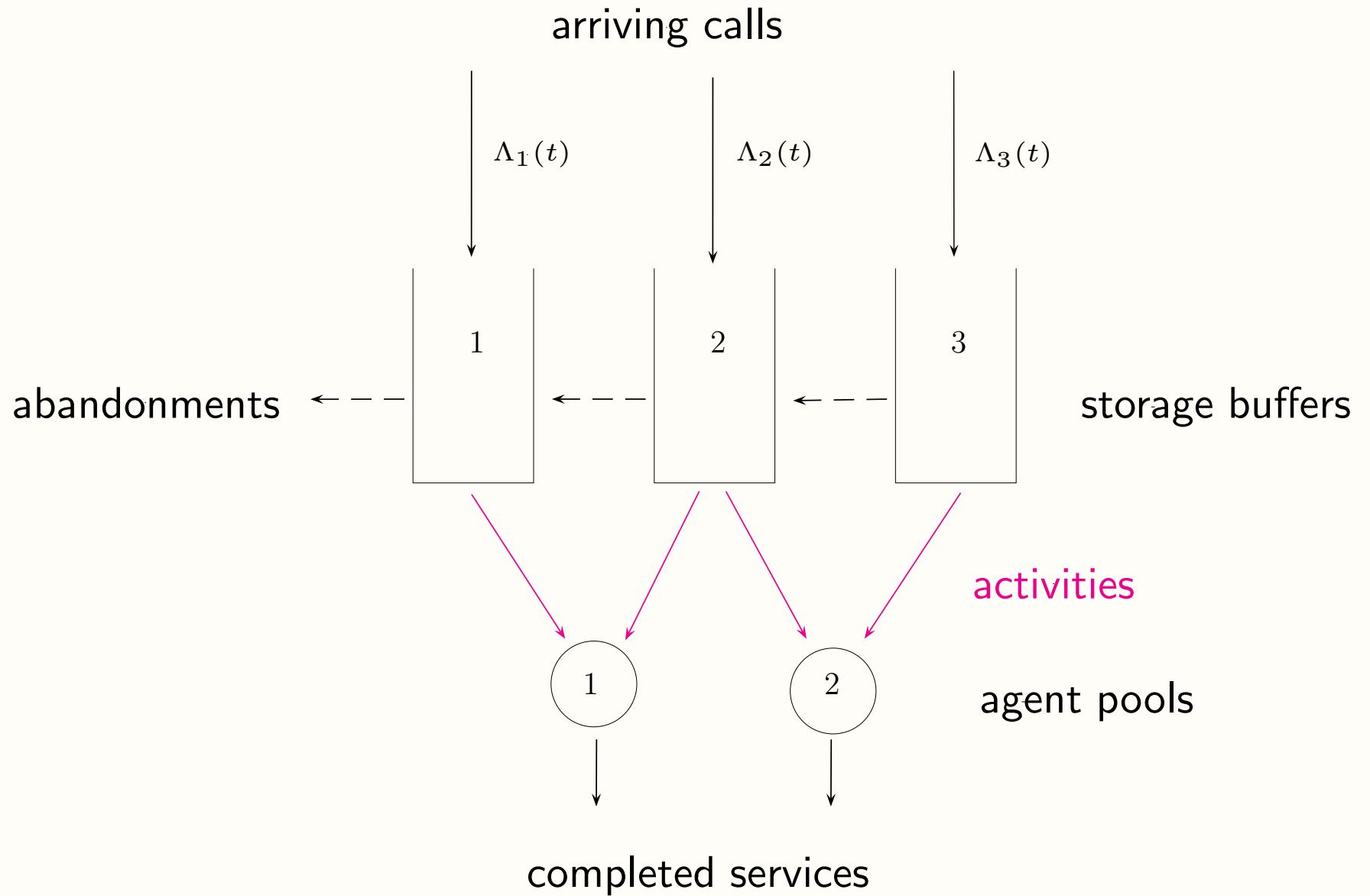
# System model



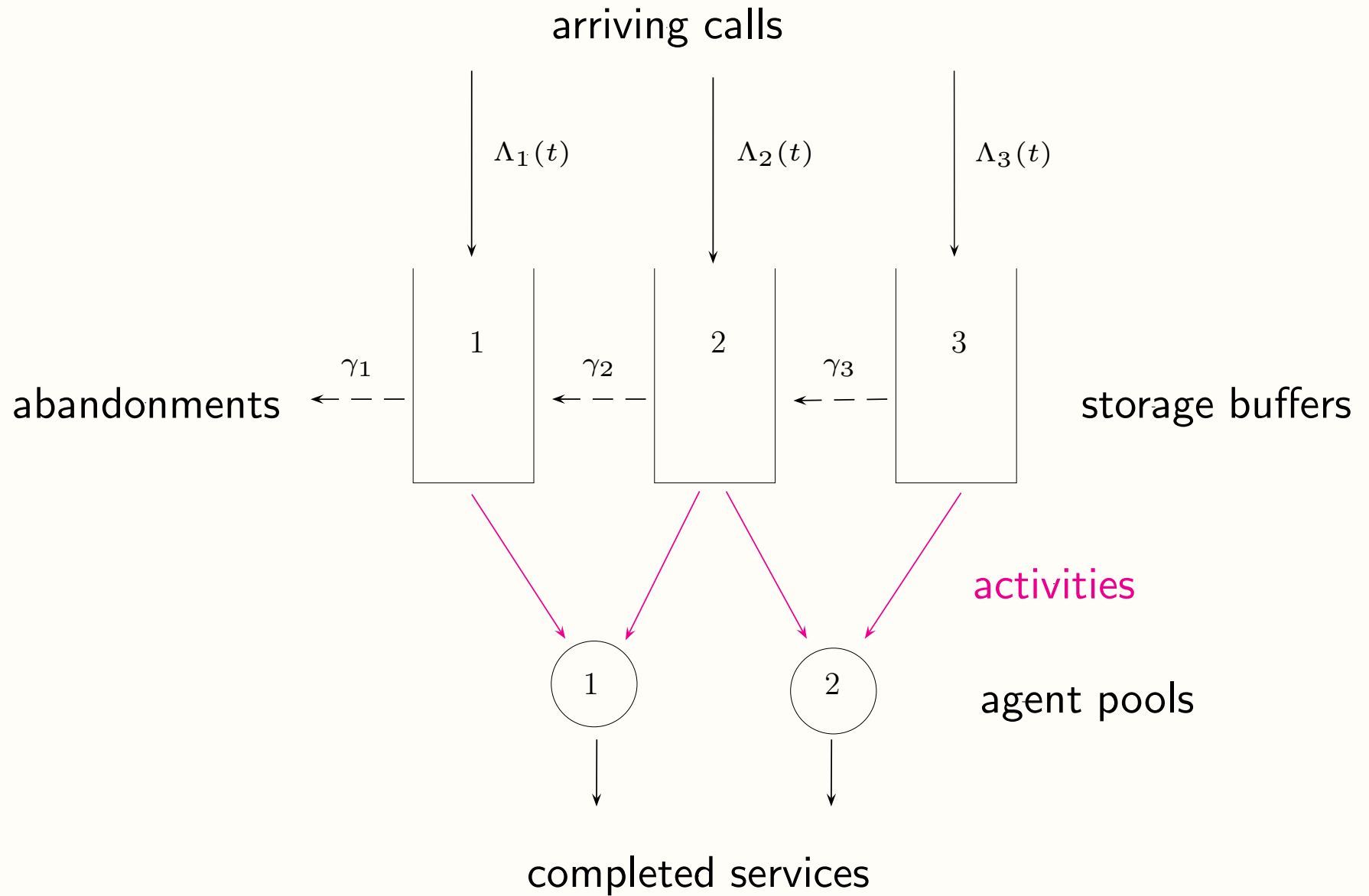
# System model



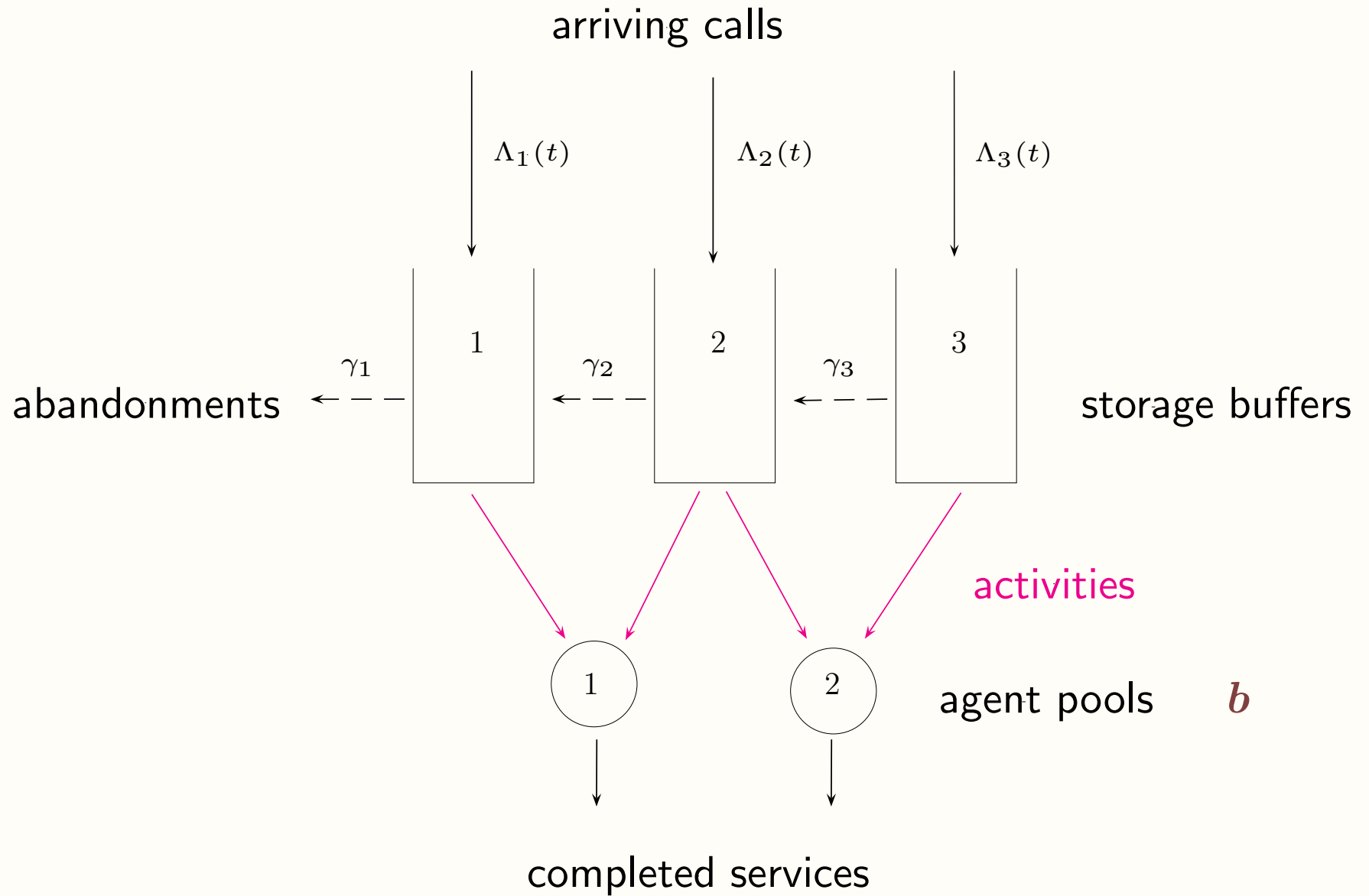
# System model



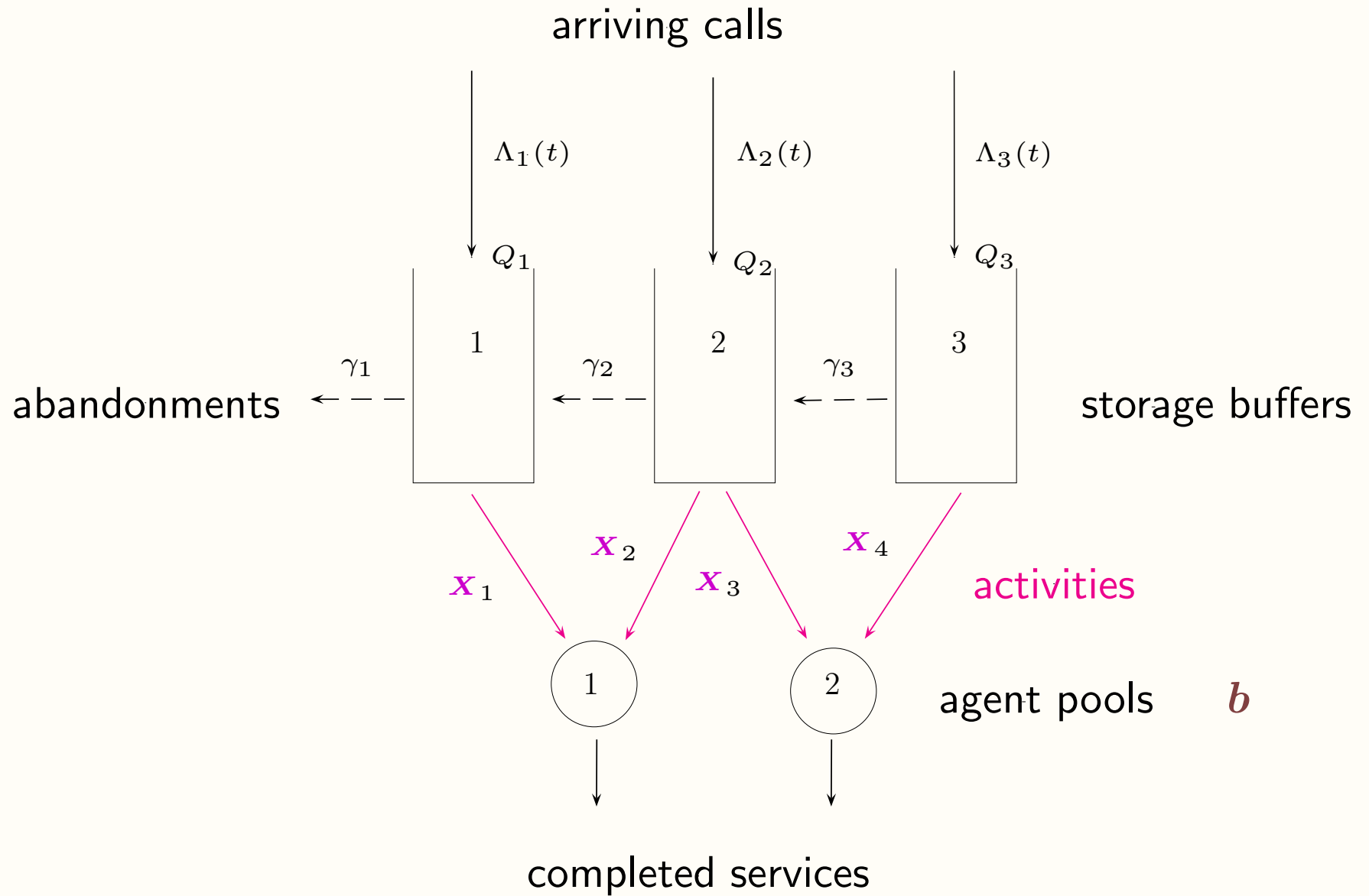
# System model



# System model



# System model



# Objective of the system manager

[P1]: Choose staffing level and a routing control to:

minimize Staffing Cost

subject to QoS Constraint

# Objective of the system manager

[P1]: Choose staffing level  $b$  and a routing control  $X$  to:

minimize  $c \cdot b$

subject to QoS Constraint

- $c$  : cost of staffing over  $[0, T]$

# Objective of the system manager

[P1]: Choose staffing level  $\mathbf{b}$  and a routing control  $\mathbf{X}$  to:

$$\begin{aligned} & \text{minimize} && c \cdot \mathbf{b} \\ & \text{subject to} && \mathbb{E} \left[ \begin{array}{c} \# \text{ of Abandonments} \\ \text{in class } i \end{array} \right] \leq \xi_i \mathbb{E} \left[ \begin{array}{c} \# \text{ of Arrivals} \\ \text{in class } i \end{array} \right] \\ & && A\mathbf{X}(t) \leq \mathbf{b}, \\ & && Q(t) \geq 0. \end{aligned}$$

- $c$  : cost of staffing over  $[0, T]$
- $\xi_i$  : maximum fraction of abandonment allowed for class  $i$  customers

# Objective of the system manager

[P1]: Choose staffing level  $\mathbf{b}$  and a routing control  $\mathbf{X}$  to:

$$\begin{aligned} & \text{minimize} && c \cdot \mathbf{b} \\ & \text{subject to} && \mathbb{E} \left[ \int_0^T \gamma_i Q_i(s) ds \right] \leq \xi_i \mathbb{E} \left[ \int_0^T \Lambda_i(s) ds \right] \text{ for all } i. \\ & && A\mathbf{X}(t) \leq \mathbf{b}, \\ & && Q(t) \geq 0. \end{aligned}$$

- $c$  : cost of staffing over  $[0, T]$
- $\xi_i$  : maximum fraction of abandonment allowed for class  $i$  customers

# Objective of the system manager

[P1]: Choose staffing level  $\mathbf{b}$  and a routing control  $\mathbf{X}$  to:

$$\begin{aligned} & \text{minimize} && c \cdot \mathbf{b} \\ & \text{subject to} && \mathbb{E} \left[ \int_0^T \gamma_i Q_i(s) ds \right] \leq \xi_i \mathbb{E} \left[ \int_0^T \Lambda_i(s) ds \right] \text{ for all } i. \\ & && A\mathbf{X}(t) \leq \mathbf{b}, \\ & && Q(t) \geq 0. \end{aligned}$$

- $c$  : cost of staffing over  $[0, T]$
- $\xi_i$  : maximum fraction of abandonment allowed for class  $i$  customers

**Note:**  $Q(t) = \#$  in system - Customers being served (depends on  $\mathbf{X}$ )

# Literature Review

- Mandelbaum and Zeltyn (2007)
- Gurvich and Whitt (2007)
- Gurvich, Armony and Mandelbaum (2005)
- Pot, Bhulai and Koole (2007)

# Literature Review

- Mandelbaum and Zeltyn (2007)
- Gurvich and Whitt (2007)
- Gurvich, Armony and Mandelbaum (2005)
- Pot, Bhulai and Koole (2007)

## Open Question

How to staff and route in a multi-class/multi-pool parallel server system to meet QoS constraint.

# Step 1: Approximating optimization problem

Problem [P1] can be approximated by the following:

[A1]: Choose the staffing level  $\mathbf{b}$  and control  $\mathbf{X}$  to

minimize  $c \cdot \mathbf{b}$

subject to  $\mathbb{E} \left[ \int_0^T \gamma_i Q_i(s) ds \right] \leq \xi_i \mathbb{E} \left[ \int_0^T \Lambda_i(s) ds \right]$  for all  $i$ .

$$Q_i(t) \geq 0$$

$$A\mathbf{X}(t) \leq \mathbf{b}.$$

# Step 1: Approximating optimization problem

Problem [P1] can be approximated by the following:

[A1]: Choose the staffing level  $\mathbf{b}$  and control  $\mathbf{X}$  to

minimize  $c \cdot \mathbf{b}$

subject to  $\mathbb{E} \left[ \int_0^T \gamma_i Q_i(s) ds \right] \leq \xi_i \mathbb{E} \left[ \int_0^T \Lambda_i(s) ds \right]$  for all  $i$ .

$$Q_i(t) \geq 0$$

$$A\mathbf{X}(t) \leq \mathbf{b}.$$

## PSFM Approximation

$$\Lambda_i(t) \approx (R\mathbf{X})_i(t) + \gamma_i Q_i(t) \quad [\text{instantaneous rate balance}]$$

# Step 1: Approximating optimization problem

Problem [P1] can be approximated by the following:

[A1]: Choose the staffing level  $\mathbf{b}$  and control  $\mathbf{X}$  to

minimize  $c \cdot \mathbf{b}$

subject to  $\mathbb{E} \left[ \int_0^T (\Lambda(s) - R\mathbf{X}(s))_i ds \right] \leq \xi_i \mathbb{E} \left[ \int_0^T \Lambda_i(s) ds \right]$  for all  $i$ .

$$R\mathbf{X}(t) \leq \Lambda(t)$$

$$A\mathbf{X}(t) \leq \mathbf{b}.$$

## PSFM Approximation

$$\gamma_i Q_i(t) \approx \Lambda_i(t) - (R\mathbf{X})_i(t) \quad [\text{instantaneous rate balance}]$$

## Step 2: Dualizing the QoS constraint

[A1]  $\Rightarrow$  Max-Min problem [A2]:

$$\max_{\mathbf{p} \geq 0} \min_{\mathbf{b} \geq 0, \mathbf{X} \in \mathcal{X}(\mathbf{b})} \left\{ c \cdot \mathbf{b} + \sum_{i=1}^m \mathbf{p}_i \mathbb{E} \left[ \int_0^T ((1 - \xi_i) \Lambda_i(s) - (R\mathbf{X})_i(s)) ds \right] \right\}$$

- $\mathcal{X}(\mathbf{b}) = \{ \mathbf{X} : R\mathbf{X}(t) \leq \Lambda(t), \quad A\mathbf{X}(t) \leq \mathbf{b} \}$
- $\mathbf{p}$  are the Lagrange multipliers

## Step 2: Dualizing the QoS constraint

[A1]  $\Rightarrow$  Max-Min problem [A2]:

$$\max_{\mathbf{p} \geq 0} \min_{\mathbf{b} \geq 0, \mathbf{X} \in \mathcal{X}(\mathbf{b})} \left\{ c \cdot \mathbf{b} + \sum_{i=1}^m \mathbf{p}_i \mathbb{E} \left[ \int_0^T ((1 - \xi_i) \Lambda_i(s) - (R\mathbf{X})_i(s) ds) \right] \right\}$$

- $\mathcal{X}(\mathbf{b}) = \{ \mathbf{X} : R\mathbf{X}(t) \leq \Lambda(t), \quad A\mathbf{X}(t) \leq \mathbf{b} \}$
- $\mathbf{p}$  are the Lagrange multipliers

### Theorem (Strong Duality)

[A1] and [A2] are equivalent.

## Step 3: Optimizing over control

[A2] reduces to the following Max-Min problem [A3]:

$$\max_{\mathbf{p} \geq 0} \min_{\mathbf{b} \geq 0} \left\{ c \cdot \mathbf{b} + \mathbb{E} \left[ \int_0^T \pi(\Lambda(s), \mathbf{b}, \mathbf{p}) ds \right] \right\}$$

## Step 3: Optimizing over control

[A2] reduces to the following Max-Min problem [A3]:

$$\max_{\mathbf{p} \geq 0} \min_{\mathbf{b} \geq 0} \left\{ c \cdot \mathbf{b} + \mathbb{E} \left[ \int_0^T \pi(\Lambda(s), \mathbf{b}, \mathbf{p}) ds \right] \right\}$$

For any vectors  $\lambda$ ,  $\mathbf{b}$  and  $\mathbf{p}$  define  $\pi(\lambda, \mathbf{b}, \mathbf{p})$  as the solution to:

$$\begin{aligned} & \text{minimize} && \sum_{i=1}^m \mathbf{p}_i ((1 - \xi_i) \lambda_i - (Rx)_i) \\ & \text{subject to} && Ax \leq \mathbf{b}, \quad Rx \leq \lambda, \quad x \geq 0. \end{aligned}$$

## Step 3: Optimizing over control

[A2] reduces to the following Max-Min problem [A3]:

$$\max_{\mathbf{p} \geq 0} \min_{\mathbf{b} \geq 0} \left\{ c \cdot \mathbf{b} + \mathbb{E} \left[ \int_0^T \pi(\Lambda(s), \mathbf{b}, \mathbf{p}) ds \right] \right\}$$

For any vectors  $\lambda$ ,  $\mathbf{b}$  and  $\mathbf{p}$  define  $\pi(\lambda, \mathbf{b}, \mathbf{p})$  as the solution to:

$$\begin{aligned} & \text{minimize} && \sum_{i=1}^m \mathbf{p}_i ((1 - \xi_i) \lambda_i - (Rx)_i) \\ & \text{subject to} && Ax \leq \mathbf{b}, \quad Rx \leq \lambda, \quad x \geq 0. \end{aligned}$$

**Note:** Decision variables in [A3] is  $\mathbf{b}$  and  $\mathbf{p}$ .

# Computing the optimal solution

## Theorem

*The optimal solution  $(\mathbf{b}^*, \mathbf{p}^*)$  of the Max-Min problem (A3) is a saddle point of the function:*

$$c \cdot \mathbf{b} + \mathbb{E} \left[ \int_0^T \pi(\Lambda(s), \mathbf{b}, \mathbf{p}) ds \right].$$

# Computing the optimal solution

## Theorem

The optimal solution  $(\mathbf{b}^*, \mathbf{p}^*)$  of the Max-Min problem (A3) is a saddle point of the function:

$$c \cdot \mathbf{b} + \mathbb{E} \left[ \int_0^T \pi(\Lambda(s), \mathbf{b}, \mathbf{p}) ds \right].$$

Apply saddle point computation approach to compute  $\mathbf{b}^*$ .

For example: Perturbation methods [Kallio and Ruszczyński(94)]

# Computing the optimal solution

## Theorem

The optimal solution  $(\mathbf{b}^*, \mathbf{p}^*)$  of the Max-Min problem (A3) is a saddle point of the function:

$$c \cdot \mathbf{b} + \mathbb{E} \left[ \int_0^T \pi(\Lambda(s), \mathbf{b}, \mathbf{p}) ds \right].$$

Apply saddle point computation approach to compute  $\mathbf{b}^*$ .

For example: Perturbation methods [Kallio and Ruszczyński(94)]

$\mathbf{p}^*$  used for constructing dynamic control

# Implementation in actual call center

- Solve the min-max problem to obtain  $(b^*, p^*)$ .

# Implementation in actual call center

- Solve the min-max problem to obtain  $(\mathbf{b}^*, \mathbf{p}^*)$ .
- Proposed staffing vector:  $\mathbf{b}^* + C\sqrt{\mathbf{b}^*}$ .

# Implementation in actual call center

- Solve the min-max problem to obtain  $(\mathbf{b}^*, \mathbf{p}^*)$ .
- Proposed staffing vector:  $\mathbf{b}^* + C\sqrt{\mathbf{b}^*}$ .
  - Will the QoS be met...

# Implementation in actual call center

- Solve the min-max problem to obtain  $(\mathbf{b}^*, \mathbf{p}^*)$ .
- Proposed staffing vector:  $\mathbf{b}^* + C\sqrt{\mathbf{b}^*}$ .
  - Will the QoS be met...Yes! if the system size is large and use an appropriate control.

# Implementation in actual call center

- Solve the min-max problem to obtain  $(\mathbf{b}^*, \mathbf{p}^*)$ .
- Proposed staffing vector:  $\mathbf{b}^* + C\sqrt{\mathbf{b}^*}$ .
  - Will the QoS be met...Yes! if the system size is large and use an appropriate control.
- Dynamic control: use the dual variables  $\mathbf{p}^*$  as the penalty vector
  - Use the control described in Bassamboo, Harrison and Zeevi (2005).

# Asymptotic Justification

# Asymptotic regime

- Time horizon grows linearly,  $[0, \kappa T]$ .
- Arrival rates grows

$$\Lambda^\kappa(\cdot) = \kappa \Lambda \left( \frac{\cdot}{\kappa} \right).$$

# Asymptotic regime

- Time horizon grows linearly,  $[0, \kappa T]$ .
- Arrival rates grows

$$\Lambda^\kappa(\cdot) = \kappa \Lambda \left( \frac{\cdot}{\kappa} \right).$$

Our operating regime

high-volume & large number of servers

# Quality of service satisfying

## Definition

A sequence of staffing levels  $\{b^\kappa\}$  are **QoS satisfying**, if for all  $\kappa$  large enough the QoS constraint is met.

# Quality of service satisfying

## Definition

A sequence of staffing levels  $\{\mathbf{b}^\kappa\}$  are **QoS satisfying**, if for all  $\kappa$  large enough the QoS constraint is met.

## Theorem

Let  $b_*^\kappa$  be the solution to Min-Max problem [A3]. Then the staffing vectors  $\hat{\mathbf{b}}^\kappa = \kappa b_*^\kappa + C \sqrt{b_*^\kappa}$  are **QoS satisfying**.

# Quality of service satisfying

## Definition

A sequence of staffing levels  $\{b^\kappa\}$  are **QoS satisfying**, if for all  $\kappa$  large enough the QoS constraint is met.

## Theorem

Let  $b_*^\kappa$  be the solution to Min-Max problem [A3]. Then the staffing vectors  $\hat{b}^\kappa = \kappa b_*^\kappa + C\sqrt{b_*^\kappa}$  are **QoS satisfying**.

# Asymptotic optimality

## Theorem

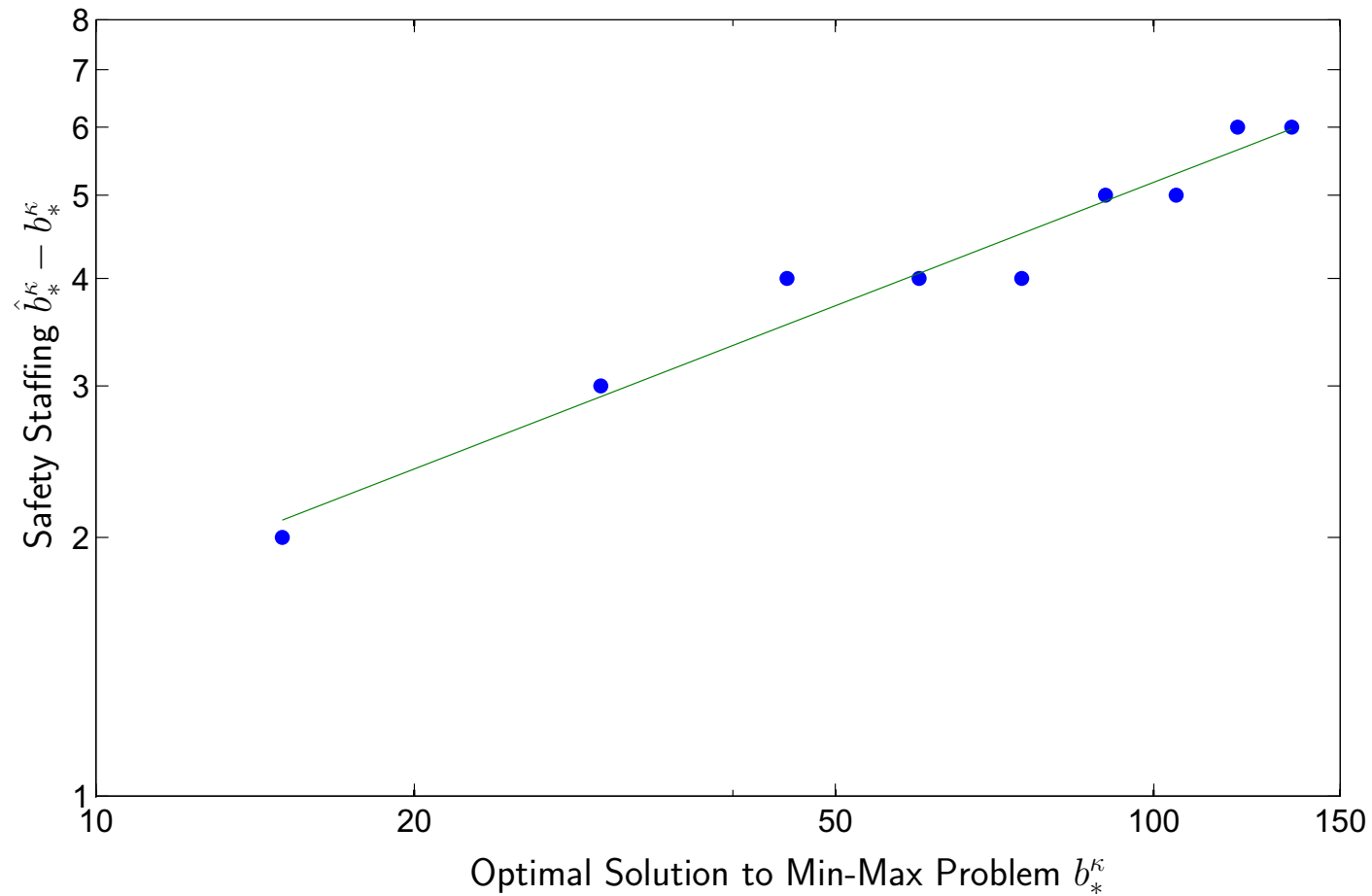
For any sequence of staffing vectors  $\{b^k\}$  that are *QoS satisfying* we have

$$\text{Cost of staffing } b^k \gtrsim \text{Cost of staffing } \hat{b}^k$$

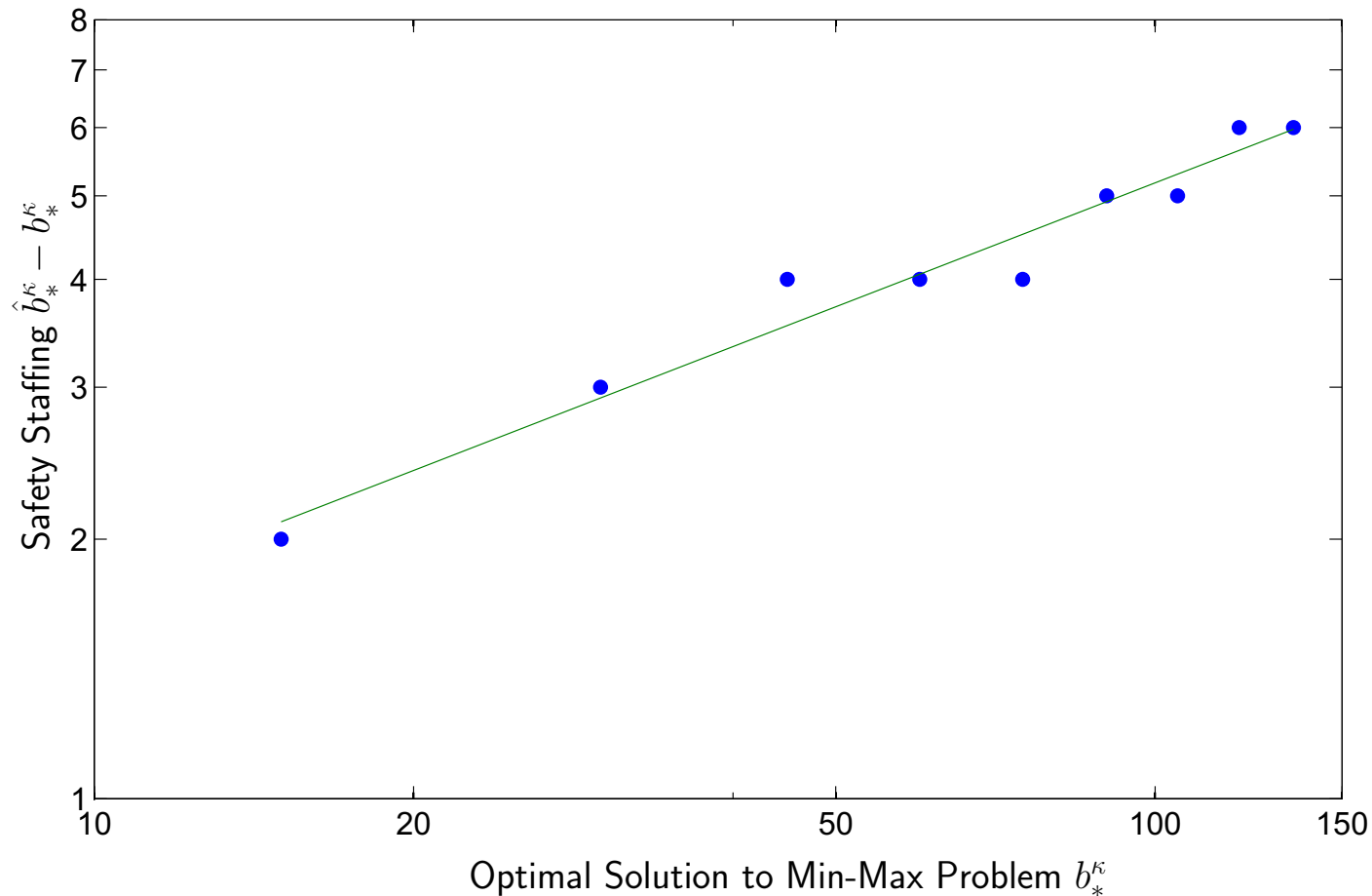
# Numerical Example: Single-class Single-pool Example

- Arrival rate is constant but random (3-point distribution)
- Service rate, Abandonment rate is 1
- QoS: fraction of abandonments is less than  $\xi = 16.6\%$ .

# Numerical Example: Single-class Single-pool Example



# Numerical Example: Single-class Single-pool Example



- Regression on log-log scale: Slope is  $0.477[\pm 0.09]$ ,  $R^2 = 0.95$
- Safety Staffing needed over  $b_*^\kappa$  is  $\mathcal{O}(\sqrt{b_*^\kappa})$ .

Thank You.