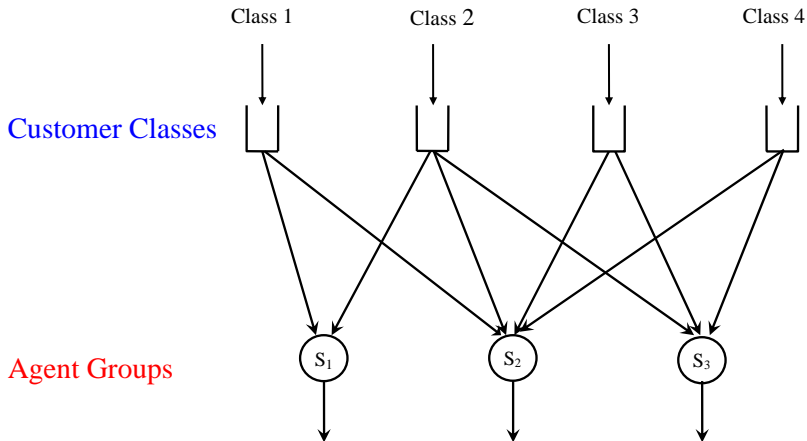


Service Level **Differentiation**  
in  
Large-Scale Service Systems

Itay Gurvich  
Columbia University

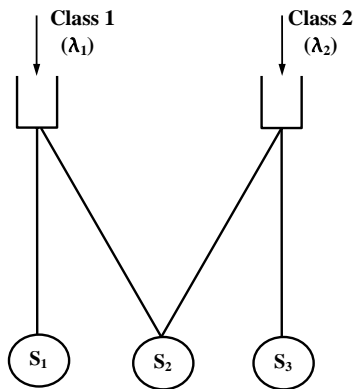
Joint work with Ward Whitt

# A network description



Different customer classes = Different service-level targets

# The objective



Service levels:

80% of class 1 wait less than 20 sec.

80% of class 2 wait less than 60 sec.

$$P\{W_i > T_i\} \leq \alpha$$

Costs:

Salary  $c_j$  for type- $j \Rightarrow \text{cost} = \sum_j c_j S_j$

Minimize labor cost through a **Staffing** and **Routing** solution

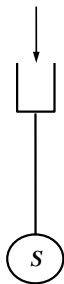
subject to **Service-Level Targets**

# Some initial observations

Single class, single type:

Exact (Calculator) *or* Simulation *or*

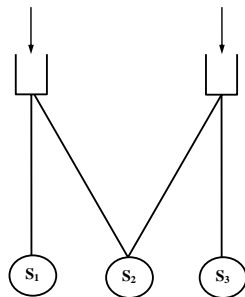
Approximations



Multi class, multi type:

Exact=MDP(?) *or* Simulation(?) *or*

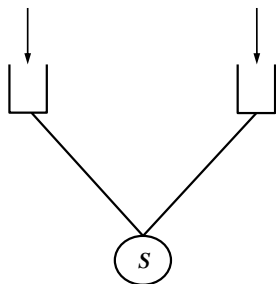
Approximations



# Our Proposed Solution - Main Idea

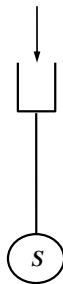
Minimize Staffing subject to

$$P\{W_i > T_i\} \leq \alpha \text{ for all } i$$



Minimize Staffing subject to

$$P\{W > T\} \leq \alpha$$



Reduction via Queue Proportions

# The Routing Rule

- ▶ Given ratios  $p_1, p_2$  ( $p_1 + p_2 = 1$ ): aim for

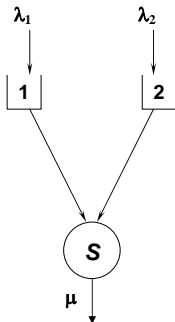
$$\frac{Q_i(t)}{Q_1(t) + Q_2(t)} \approx p_i$$

- ▶ A Simple Routing -

Fixed Queue Ratio (FQR)

Serve class  $i$  with greatest **Queue Imbalance**

$$\frac{Q_i(t)}{Q_1(t) + Q_2(t)} - p_i$$



## Doing the reduction

$$P\{W_i > T_i\} \approx P\{Q_i > \lambda_i T_i\}$$

$$[\text{Little's Law } Q_i \approx \lambda_i W_i]$$

$$\approx P\left\{p_i \left[\sum_{k=1}^2 Q_k\right] > \lambda_i T_i\right\}$$

$$\left[\frac{Q_i(t)}{Q_1(t) + Q_2(t)} \approx p_i\right]$$

$$= P\left\{\left[\sum_{k=1}^2 Q_k\right] > \sum_{k=1}^2 \lambda_k T_k\right\}$$

$$\left[p_i = \frac{\lambda_i T_i}{\lambda_1 T_1 + \lambda_2 T_2}\right]$$

$$= P\left\{Q > \sum_{k=1}^2 \lambda_k T_k\right\} = P\{W > T\} \leq \alpha$$

$$\left[T = \frac{\lambda_1 T_1 + \lambda_2 T_2}{\lambda_1 + \lambda_2}\right]$$

## Doing the reduction

$$P\{W_i > T_i\} \approx P\{Q_i > \lambda_i T_i\} \quad [\text{translating to queues}]$$

$$\approx P\left\{p_i \left[\sum_{k=1}^2 Q_k\right] > \lambda_i T_i\right\} \quad [\text{routing}]$$

$$= P\left\{\left[\sum_{k=1}^2 Q_k\right] > \sum_{k=1}^2 \lambda_k T_k\right\} \quad [\text{choosing the proportions}]$$

$$= P\left\{Q > \sum_{k=1}^2 \lambda_k T_k\right\} = P\{W > T\} \leq \alpha \quad [\text{staffing}]$$

# A Basic Setting - The V Model

- ▶ Minimize Staffing subject to

$$P\{W_i > T_i\} \leq \alpha \text{ for all } i$$

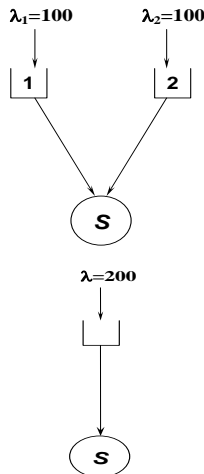
- ▶ Proposed Solution

- ▶ **Staffing:** based on single class with average target

$$T = \frac{\lambda_1}{\lambda} T_1 + \frac{\lambda_2}{\lambda} T_2$$

- ▶ **Routing:** Use Fixed Queue Ratio with ratios

$$p_i = \frac{\lambda_i T_i}{\lambda_1 T_1 + \lambda_2 T_2}$$



Staffing through aggregate - Differentiation through FQR

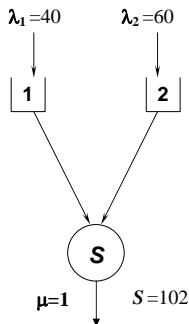
# Theoretical Support

As  $\lambda \rightarrow \infty$  (and  $S \rightarrow \infty$  accordingly)

$$\frac{Q_i(t)}{\sum_{k=1}^m Q_k(t)} \Rightarrow p_i$$

- ▶ Key: **Asymptotic proportionality**
- ▶ System behavior captured through aggregate queue length  
= **state-space collapse**.
- ▶ Extending Bramson (1998) + Dai and Tezcan (2006)

# Why does it work?



$$p_1 = 1/3 \quad p_2 = 2/3$$

Expect:

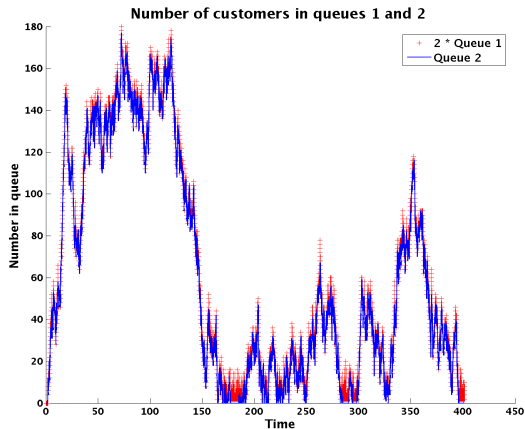
$$Q_1(t) = \frac{1}{3}(Q_1(t) + Q_2(t))$$

$$Q_2(t) = \frac{2}{3}(Q_1(t) + Q_2(t))$$

⇓

$$\frac{Q_2(t)}{Q_1(t)} = 2$$

## Why does it work? - Cont.



$$\frac{Q_2(t)}{Q_1(t)} \approx 2$$

A single sample-path!!!

## Extending FQR to Multiple Agent Types

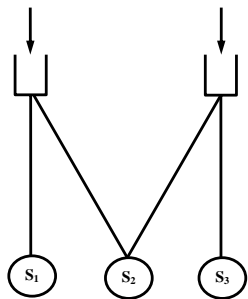
**Queue** proportions  $p = (p_1, \dots, p_I)$

**Idleness** proportions  $v = (v_1, \dots, v_J)$

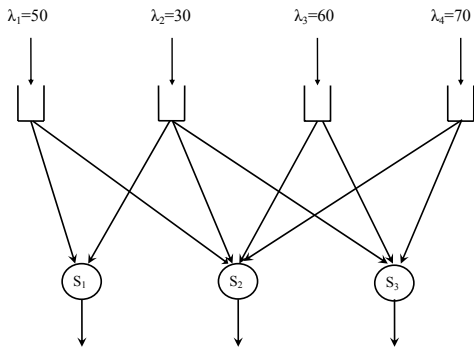
$$\frac{Q_i(t)}{\sum_{l=1}^I Q_l(t)} \approx p_i, \quad \frac{I_j(t)}{\sum_{l=1}^J I_l(t)} \approx v_j$$

- ▶ **Upon service completion:** as before
- ▶ **Upon arrival:** choose agent pool with greatest

**Idleness Imbalance**



# Example 1



$$P\{W_1 > 0.15\} \leq 0.2$$

$$P\{W_2 > 0.2\} \leq 0.2$$

$$P\{W_2 > 0.15\} \leq 0.2$$

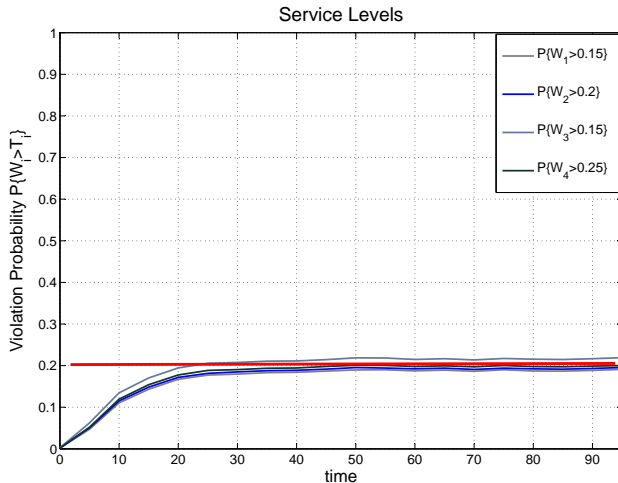
$$P\{W_2 > 0.25\} \leq 0.2$$

Class-dependent service rates  $\mu_1 = 1, \mu_2 = 0.5, \mu_3 = 1.5, \mu_4 = 2$ .

**Staffing:**  $M/M/S$  based aggregate (with avg. service time) [Distribution?]

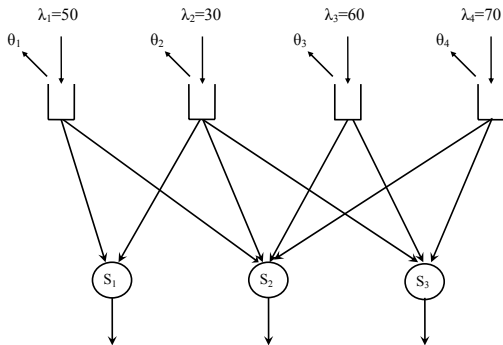
**Control:** FQR with  $p_i = \frac{\lambda_i T_i}{\sum_{k=1}^4 \lambda_k T_k}$ .

# Example 1



\* Based on simulation study with Zohar Feldman

## Example 2



$$P_1\{Ab\} \leq 3\%$$

$$P_2\{Ab\} \leq 2\%$$

$$P_3\{Ab\} \leq 5\%$$

$$P_4\{Ab\} \leq 4\%$$

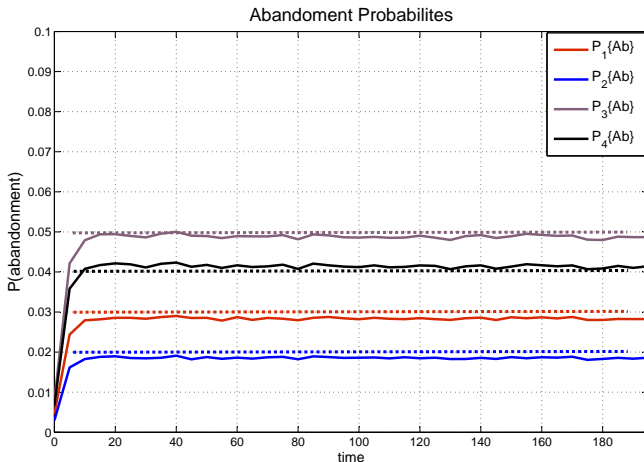
Abandonment rates:  $\theta_1 = 1, \theta_2 = 0.8, \theta_3 = 1.2, \theta_4 = 1.4$ .

**Staffing:**  $M/M/N + M$  based aggregate

[Rates?]

**Control:** FQR with  $p_i := \frac{\lambda_i \alpha_i / \theta_i}{\sum_{k=1}^4 \lambda_k \alpha_k / \theta_k}$ .

## Example 2



The End