

Staffing and Routing in Single-Class Large-Scale Service Systems with Heterogeneous Servers

Mor Armony

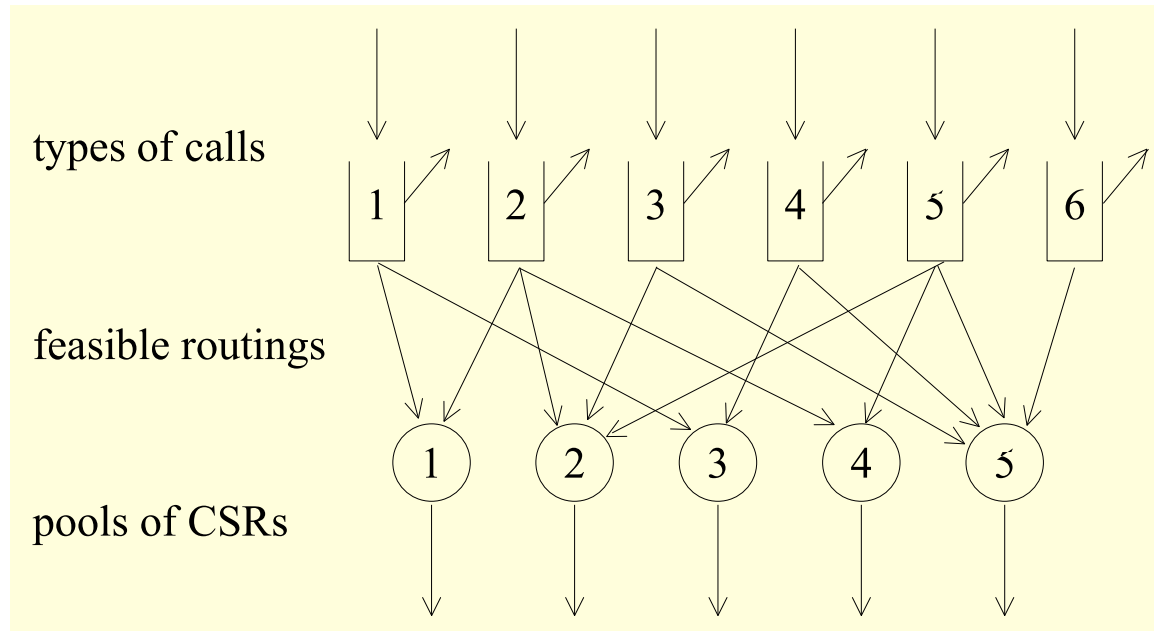
Based on joint work with

Avi Mandelbaum

Itay Gurvich

July 2004

General Multi-Skill Call-Centers

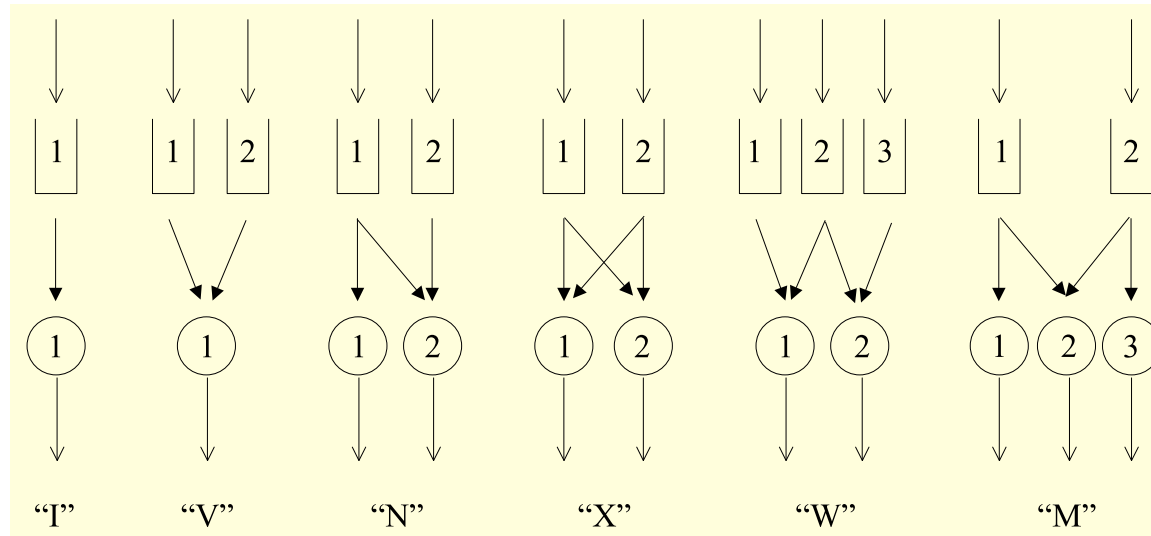


Main Operational Issues (Given a Call Volume Forecast):

- **Design** - Long Term
- **Staffing** - Short Term
- **Routing** - Real time

Very Complex: Usually treated hierarchically and unilaterally.

Design “Building-Blocks” for Multi-Skill Call Centers



Sample of Literature for Many-Server Systems:

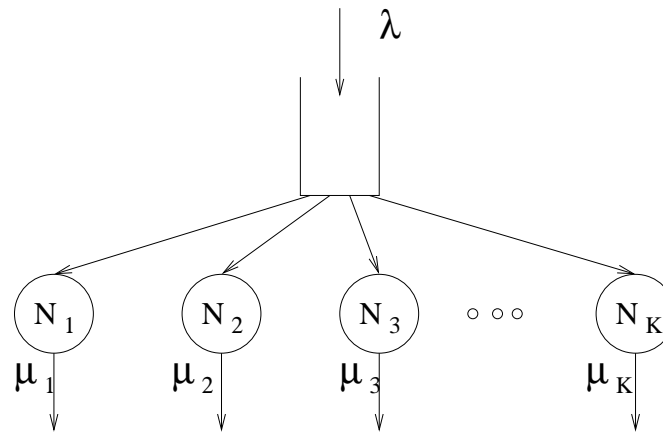
I-design: Halfin & Whitt ('81), Borst, Mandelbaum & Reiman ('03).

V-design: Brandt & Brandt ('99), Puhalskii & Reiman ('00), Koole & Bhulai ('02), Gans & Zhou ('02), Atar, Mandelbaum & Reiman ('02), A. & Maglaras ('03), , Harrison & Zeevi ('03), Maglaras & Zeevi ('03), Yahalom & Mandelbaum ('04), Gurvich ('04).

Λ-design: Rykov ('01), Luh & Viniotis ('02), de Véricourt & Zhou ('03).

General Design: Shumsky ('03), Harrison & Zeevi ('04), Basamboo, Harrison & Zeevi ('04), Gans & Zhou ('04),

The \wedge -Design



- Single customer class - Poisson arrival process with rate λ .
- K server pools (N_1, N_2, \dots, N_K servers)
- Exponential service times with rates $\mu_1 < \mu_2 < \dots < \mu_K$.

Our Focus: Staffing and Routing

- How many servers of each type are needed?
- How to route incoming or waiting customers?

Background on Staffing: Halfin-Whitt ('81)

Consider a sequence of $M/M/N$ models, $N = 1, 2, 3, \dots$

Then the following **3 points of view** are equivalent:

• Customers: $\lim_{N \rightarrow \infty} P_N\{\text{Wait} > 0\} = \alpha, \quad 0 < \alpha < 1;$

• Servers: $\lim_{N \rightarrow \infty} \sqrt{N}(1 - \rho_N) = \beta, \quad 0 < \beta < \infty;$

• Manager: $N \approx R + \beta\sqrt{R}, \quad R = \lambda/\mu \text{ large};$

Here $\alpha = \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1}$, where $\Phi(\cdot)/\phi(\cdot)$ is the standard normal distribution / density.

Extremes:

Everyone waits: $\alpha = 1 \Leftrightarrow \beta \leq 0$ **Efficiency-driven**

No one waits: $\alpha = 0 \Leftrightarrow \beta = \infty$ **Quality-driven**

Background on Staffing: Economics of $\sqrt{\cdot}$ Safety-Staffing

Borst, Mandelbaum & Reiman ('02)

Quality $D(t)$ delay cost (t =delay time).

Efficiency $C(N)$ staffing cost (N = # agents)

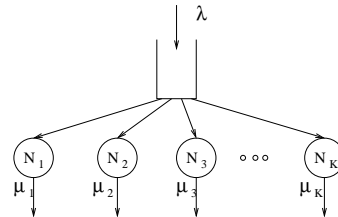
Optimization: N^* that minimizes total costs

- $C \gg D$: Efficiency-driven $N \approx R + \gamma$
- $C \ll D$: Quality-driven $N \approx R + \delta R$
- $C \approx D$: QED $N \approx R + \beta\sqrt{R}$

Framework: Asymptotic theory of $M/M/N$, $N \uparrow \infty$.

Constraint Satisfaction: Minimal N^* that adheres to delay cost constraint.

Staffing in the \wedge -design model



Challenges with the above dimensioning approach:

- R is not well defined
- Routing is not specified
- Constraint satisfaction: feasible region is multi-dimensional

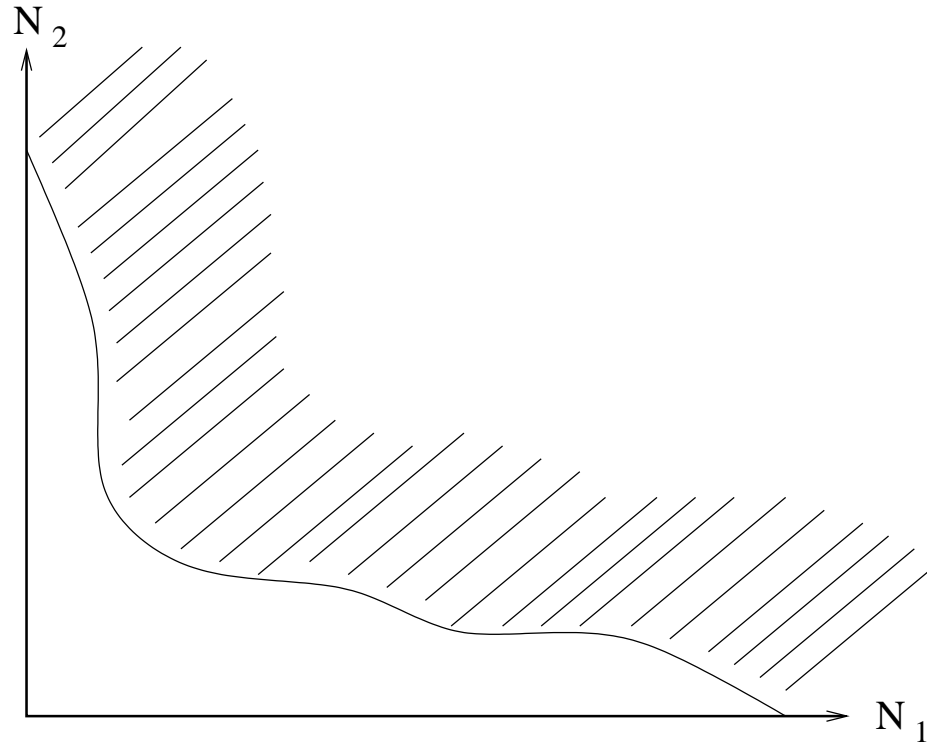
The staffing problem:

$$\text{Minimize} \quad C_1(N_1) + C_2(N_2) + \dots + C_K(N_K)$$

$$\text{Subject to} \quad P(\text{wait} > 0) \leq \alpha, \text{ for some routing policy}$$

Solution: $\mu_1 N_1 + \mu_2 N_2 + \dots + \mu_K N_K = \lambda + \text{safety-staffing}$

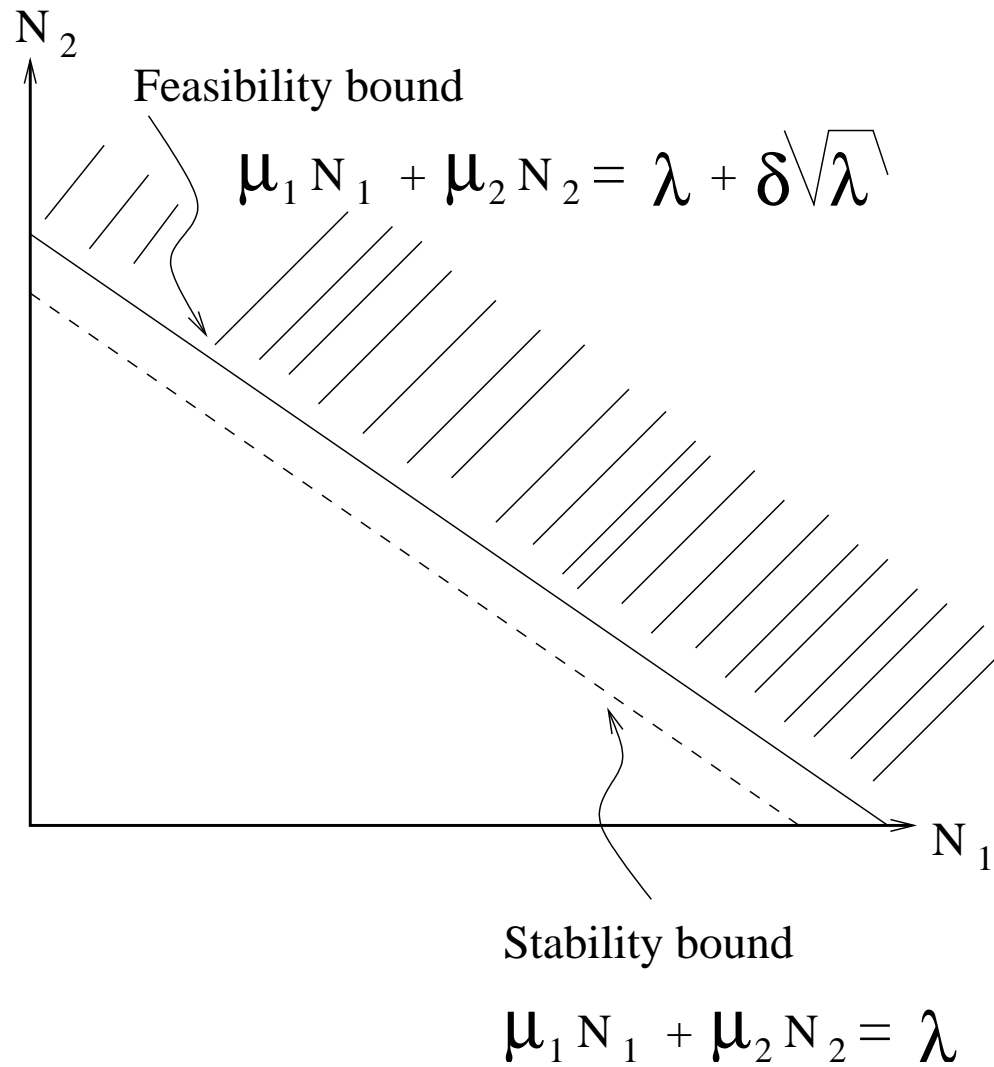
The feasible region



Problems:

- Need to find optimal routing.
 - Threshold type solutions: Luh & Viniotis ('01) and Rykov ('01).
- Difficult to find **exact** feasible region.

The asymptotic feasible region



The asymptotic feasible region

Theorem (Asymptotic Feasible Region): Consider a sequence of systems indexed by $\lambda \uparrow \infty$. Suppose that $\liminf_{\lambda \rightarrow \infty} N_1/N > 0$. Then there exists a non-preemptive policy for which

$$\limsup_{\lambda \rightarrow \infty} P_\lambda(\text{wait} > 0) \leq \alpha, \quad 0 < \alpha < 1$$

if and only if

$$\mu_1 N_1 + \mu_2 N_2 + \dots + \mu_K N_K \geq \lambda + \delta \sqrt{\lambda} + o(\sqrt{\lambda}), \quad 0 < \delta < \infty.$$

Here

$$\alpha = \left[1 + \frac{(\delta/\sqrt{\mu_1})\Phi(\delta/\sqrt{\mu_1})}{\phi(\delta/\sqrt{\mu_1})} \right]^{-1}.$$

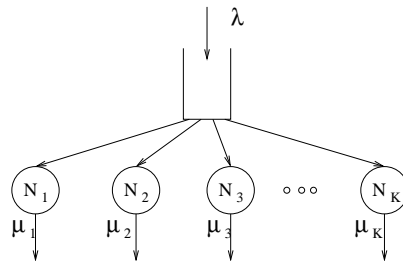
Corollary: The heterogeneous server system outperforms the homogeneous one.

Proof: Let $\mu = \sum_{k=1}^K q_k \mu_k$. Then $P(\text{wait} > 0) \leq \alpha$, iff

I-design:
$$\mu N \geq \lambda + \beta(\alpha) \sqrt{\mu} \sqrt{\lambda},$$

\wedge -design:
$$\mu_1 N_1 + \mu_2 N_2 + \dots + \mu_K N_K \geq \lambda + \beta(\alpha) \sqrt{\mu_1} \sqrt{\lambda}.$$

Background: Optimal Routing Policies for the \wedge -Design Model (Rykov ('01), Luh & Viniotis ('02))



Problem: Find a non-preemptive non-anticipative routing control to minimize the average total number of customers in the system.

Solution: The optimal solution is of **threshold** type. Assign a customer to server k if:

1. It is the fastest idle server, and
2. the number of customers in queue is m_k or more.

The thresholds have the following properties:

- $m_1 \geq m_2 \geq \dots m_K$.
- m_k may depend on the state of the other (slower) servers.

Asymptotically Optimal Routing

Proposition (Optimal Preemptive Routing): The preemptive routing policy, FSF_P , that always sends calls to the faster servers first is path-wise optimal (it minimizes total number of jobs in the system at any time $t \geq 0$).

Proof: Sample path coupling argument.

Proposition (Asymptotically Optimal Routing): The non-preemptive routing policy, FSF , that always sends incoming or waiting calls to the faster servers first is asymptotically optimal (the difference between the number of jobs under FSF and FSF_P asymptotically goes to 0, in probability, U.O.C).

Proof: State-space collapse - in the limit the fast servers are always busy.

⇒ The preemptive and non-preemptive policies are asymptotically the same.

Note (no thresholds): Thresholds are not-needed-

The Halfin-Whitt regime is different than the traditional heavy-traffic regime (Teh & Ward ('02)).

Asymptotic Feasibility

Proposition (Limiting Waiting Probability): For both Π^P and Π^{NP}

$$\lim_{\lambda \rightarrow \infty} P(\text{wait} > 0) = \alpha \quad 0 \leq \alpha \leq 1,$$

if and only if

$$\mu_1 N_1 + \mu_2 N_2 + \dots + \mu_K N_K \approx \lambda + \delta \sqrt{\lambda}, \quad 0 \leq \delta \leq \infty$$

where

$$\alpha = \left[1 + \frac{(\delta/\sqrt{\mu_1})\Phi(\delta/\sqrt{\mu_1})}{\phi(\delta/\sqrt{\mu_1})} \right]^{-1},$$

provided that $\liminf_{\lambda \rightarrow \infty} N_1/N > 0$.

Conclusion: The asymptotic feasible region.

Asymptotically Optimal Staffing: Homogeneous Cost Function

Problem:

$$\text{Minimize } C_1 N_1^p + C_2 N_2^p + \dots + C_K N_K^p, \quad p > 1$$

$$\text{Subject to } P(\text{wait} > 0) \leq \alpha, \quad \text{for some routing policy}$$

Solution: Solve

$$\text{Minimize } C_1 N_1^p + C_2 N_2^p + \dots + C_K N_K^p, \quad p > 1$$

$$\text{Subject to } \mu_1 N_1 + \mu_2 N_2 + \dots + \mu_K N_K \geq \lambda + \delta \sqrt{\lambda}$$

To get

$$\frac{C_k}{\mu_k} N_k^{p-1} = \frac{C_j}{\mu_j} N_j^{p-1} \tag{1}$$

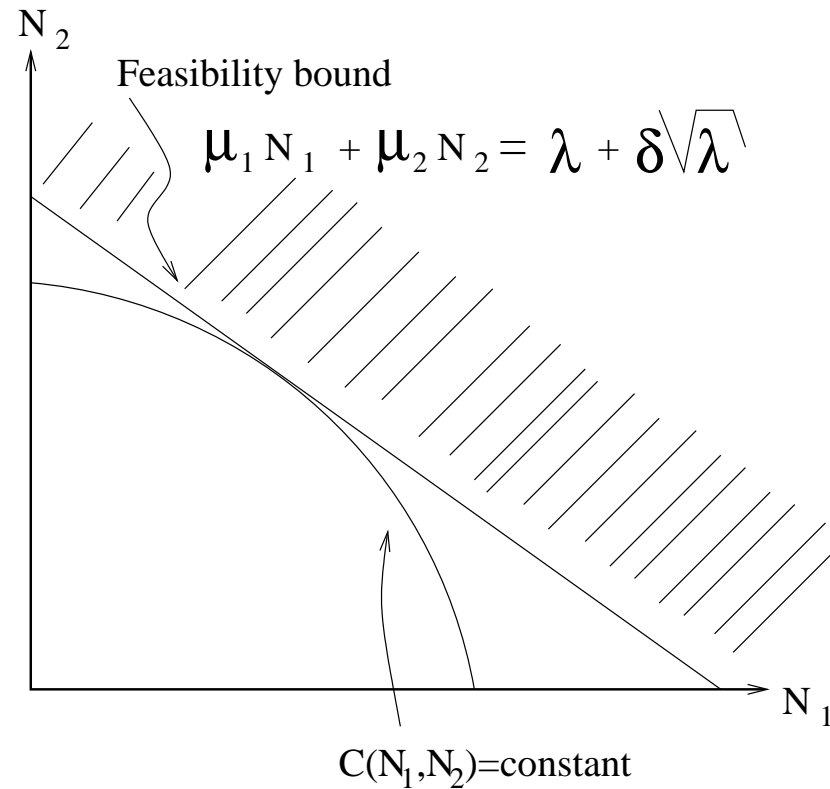
$$\mu_1 N_1 + \mu_2 N_2 + \dots + \mu_K N_K = \lambda + \delta \sqrt{\lambda} \tag{2}$$

↓

$$\frac{N_k}{N_j} = \left(\frac{\mu_k / C_k}{\mu_j / C_j} \right)^{1/(p-1)} .$$

Note: $N_1/N > 0$!!!

Asymptotically Optimal Staffing



Note: Constraints can be easily incorporated such as:

- Slow servers (training agents) need to exceed a certain portion of the total.
- Number of fast servers is limited.

Transient Analysis

$Z(t)$ = the total number of jobs in the system,

$N = N_1 + N_2 + \dots + N_K$ the total number of servers, $X^\lambda(t) = \frac{Z(t) - N}{\sqrt{N}}$.

Proposition: Suppose that

1. $\lim_{\lambda \rightarrow \infty} \frac{\mu_k N_k}{\lambda} = a_k$, $k = 1, 2, \dots, K$, $a_1 > 0$, $a_k \geq 0$, $a_1 + a_2 + \dots + a_K = 1$, and
2. $\lim_{\lambda \rightarrow \infty} \frac{\mu_k N_k - a_k \lambda}{\sqrt{\lambda}} = \delta_k$, $k = 1, 2, \dots, K$, $-\infty < \delta_k < \infty$, $\delta = \delta_1 + \delta_2 + \dots + \delta_K > 0$.

Then, if $X^\lambda(0) \Rightarrow X(0)$ then under both FSF_P and FSF, $X^\lambda \Rightarrow X$, where X is a diffusion process with infinitesimal drift and variance:

$$m(x) = \begin{cases} -\delta\sqrt{\mu} & x \geq 0, \\ -\delta\sqrt{\mu} - \mu_1 x & x < 0, \end{cases}$$

and

$$\sigma^2(x) = 2\mu.$$

Here, $\mu = (a_1/\mu_1 + a_2/\mu_2 + \dots + a_K/\mu_K)^{-1}$.

Asymptotic Feasibility

Proof:

1. Let ν_P^λ be the steady-state distribution associated with FSF_P , then ν_P^λ weakly converges as $\lambda \uparrow \infty$ and hence, $\lim_{\lambda \rightarrow \infty} P_{\nu_P^\lambda}(\text{wait} > 0) = \alpha$.

2. Coupling between FSF_P and FSF : excursions from (N_1, N_2) are upwards-identical, and downwards-shorter for FSF . Hence:
 - (a) There exists a steady-state distribution ν^λ for FSF (the state (N_1, N_2) is positive-recurrent).
 - (b) The set $\{\nu^\lambda\}_{\lambda > 0}$ is tight.
$$\Rightarrow \lim_{\lambda \rightarrow \infty} P_{\nu^\lambda}(\text{wait} > 0) = \alpha.$$

The \wedge and \vee designs

	\wedge design	\vee design
Staffing	Minimize $C(\vec{N})$ Subject to $P(wait > 0) \leq \alpha$.	Minimize N Subject to $P(wait_i > 0) \leq \alpha_i,$ $i = 1, \dots, J$
Control	Priority routing (FSF) Minimize Steady-State Delays	Priority Sequencing (LDPF) Provided that $N = R + \beta(\alpha_J)\sqrt{R}$ Satisfy service level differentiation
Thresholds	Associated with queue length	Associated with number of idle servers
State-Space Collapse	Only slow servers are idle	Only low priority customers wait

The V-design: Service Level Differentiation

A two-class example

Suppose that staffing level is $N = R + \beta\sqrt{N}$, then:

Threshold	$\sim P\{wait_1 > 0\}$	$\sim P\{wait_2 > 0\}$
a	$\alpha(\beta)\rho_1^a$	$\alpha(\beta)$
$b \ln N$	$\alpha(\beta)\rho_1^{b \ln N}$	$\alpha(\beta)$
$c\sqrt{N}$	$\alpha(\beta - c)\rho_1^{c\sqrt{N}}$	$\alpha(\beta - c)$

Conclusions and Further Research

Conclusions:

1. Square-root safety staffing is asymptotically optimal (\Rightarrow linear boundary for feasible region).
2. Asymptotic cost optimization is simple.
3. Routing to fast servers first is asymptotically optimal (no thresholds are needed).
4. \wedge -design better than I-design.

Future Research:

1. Transient (finite horizon) staffing.
2. Combine V-design and \wedge -design to study N-design.