Staffing and Control of Large Service Systems: The Case of Multiple Customer Classes and Fully Flexible Servers

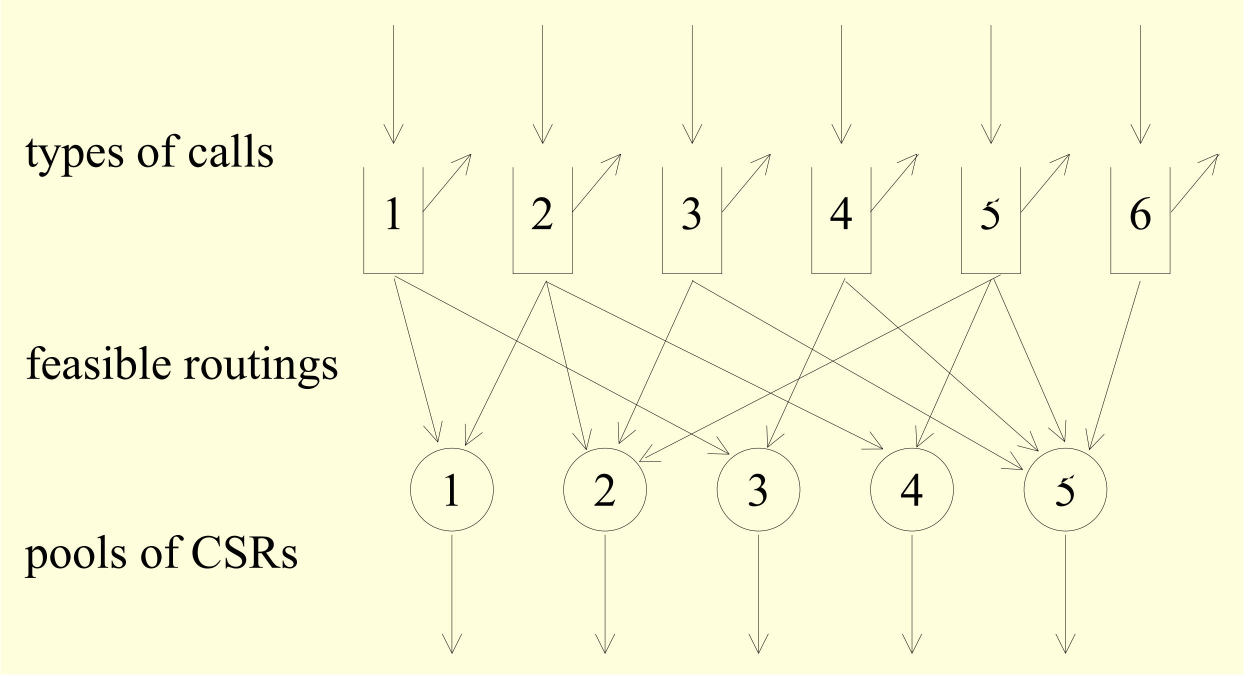
Mor Armony

Joint work with

Itay Gurvich and Avi Mandelbaum

October 2004

Multi-Skill Call-Centers

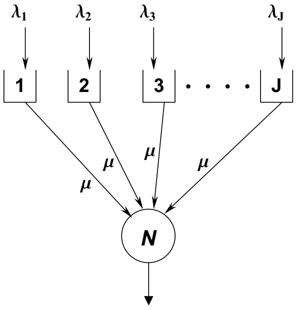


Main Operational Issues (Given a Forecast of Workload):

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

Very Complex: Hence treated hierarchically and unilaterally.

The V-Design



The Joint Staffing and Control Problem:

How many servers to use (staffing)

and how to match them with customers (control)

so as to minimize staffing + holding costs

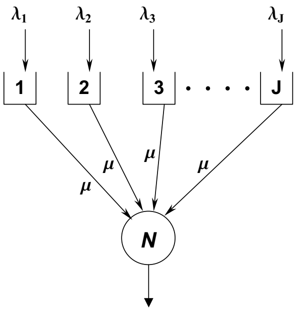
subject to service level constraints.

$$\text{minimize} \quad c_1\lambda_1 EW_1 + \ldots + c_J\lambda_J EW_J + N$$

$$\text{subject to} \quad P_\pi(W_i > 0) \le \alpha_i, \quad i = 1, \ldots, J$$

$$N \in \mathbb{Z}_+, \quad \pi \in \Pi$$

The V-Design: Motivation



Customer population heterogeneity occurs naturally:

- multilingual
- different needs (new account, billing questions, etc.),
- business vs. private consumer

or as a marketing tool:

- Willingness to pay (first vs. economy class)
- Frequency of use (frequent flyer)
- Business Potential (grandparents club)

The V-Design: Motivation

Two conflicting goals: Marketing -

- Deliver on promised quality of service,
- Provide service expertise,
- Prioritize revenue generating customers,

vs. Operations -

- High server efficiency,
- Low operating costs,
- Cross-training.

Question: Can these two goals be consolidated?

Answer: Yes! In large systems, fully flexible servers can provide high and differentiated quality of service with high server efficiencies.

Asymptotic Framework: Quality and Efficiency Driven (QED) regime.

Talk Outline

1. Background
2. Model Formulation
3. SRSS and TP policies
4. Asymptotic Optimality
5. Discussion and Conclusions

Related Literature:

Staffing and Control of Large Scale Service Systems

- I-design: Halfin & Whitt ('81), Garnett, Mandelbaum & Reiman ('02), Borst, Mandelbaum & Reiman ('03), Mandelbaum & Zeltyn ('04).

- V-design: Schaack & Larson ('86), Brandt & Brandt ('99), Koole & Bhulai ('02), Gans & Zhou ('02), Armony & Maglaras ('03), Atar, Mandelbaum & Reiman ('03), Harrison & Zeevi ('03), Yahalom & Mandelbaum ('04), Gurvich ('04).

- $\wedge$-design: Rykov ('01), Luh & Viniotis ('01), de Véricourt & Zhou ('03), Armony ('04) Armony & Mandelbaum ('04).

- N-, M- and W- designs: Shumsky ('04), Chevalier, Shumsky & Tabordon ('04), Wallace and Whitt ('04)

- General-design: Atar ('04), Harrison & Zeevi ('04), Bassamboo, Harrison & Zeevi ('04).

QED and the I-design

Theorem (Halfin-Whitt, 1981):

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

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

- Customer: $\lim_{N\to\infty} P_N\{Wait > 0\} = \alpha$, $0 < \alpha < 1$;
- Server: $\lim_{N\to\infty} \sqrt{N}\,(1 - \rho_N) = \beta$, $0 < \beta < \infty$;
- Manager: $N \approx R + \beta\sqrt{R}$, $R = \lambda/\mu$ 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**

Staffing and the I-design:

$\sqrt{\cdot}$ Safety-Staffing (SRSS)

Borst, Mandelbaum & Reiman ('02)

Quality	C (t) delay cost	($t =$ delay time).
Efficiency	S (N) staffing cost	($N = \#$ agents)

Assume $S(N) \equiv N$

Optimization: N^* that minimizes total costs

• $C << \mathbf{1}$:	Efficiency-driven	$N \approx R + \gamma$
• $C >> \mathbf{1}$:	Quality-driven	$N \approx R + \delta R$
• $C \approx \mathbf{1}$:	QED	$N \approx R + \beta\sqrt{R}$

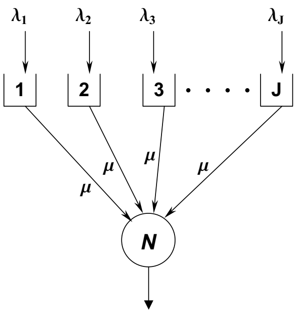
Satisfization: N^* that minimizes staffing costs s.t. delay constraints.

Here: **N* that is minimal s.t. $\mathrm{P}(\mathrm{Wait} > 0) \leq \alpha$.**

• $\alpha \approx \mathbf{1}$:	Efficiency-driven	$N \approx R + \gamma$
• $\alpha \approx \mathbf{0}$:	Quality-driven	$N \approx R + \delta R$
• $\mathbf{0} < \alpha < \mathbf{1}$:	QED	$N \approx R + \beta\sqrt{R}$

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

The V-Design



- J customer classes: arrivals Poisson(λ_j).
- N iid servers: service durations Exp(μ)
- Staffing and Control Problem:

$$\begin{array}{ll} \text{minimize} & c_1\lambda_1 EW_1 + \ldots + c_J\lambda_J EW_J + N \\ \text{subject to} & P_\pi(W_i > 0) \le \alpha_i, \quad i = 1, \ldots, J \\ & N \in \mathbb{Z}_+, \quad \pi \in \Pi \end{array}$$

Assumptions:

$\alpha_1 \le \alpha_2 \le \ldots \le \alpha_J$,

$c_1 \ge c_2 \ge \ldots \ge c_J$.

Remarks:

- A pure cost minimization problem: $\alpha_i = 1$ for all $i = 1, \ldots, J$.

- A constraint satisfaction problem: $c_i = 0$ for all $i = 1, \ldots, J$.

Staffing and Control Proposals

Staffing: Square-Root Safety-Staffing (SRSS)

$$N = R + \beta\sqrt{R}, \quad \beta > 0$$

Control: Threshold-Priorities (TP): Serve class j if

1) All higher priority queues are empty, and

2) the number of idle servers exceeds the threshold K_j.

$$\text{Priorities: } 1 > 2 > ... > J,$$

$$\text{Thresholds: } K_1 \le K_2 \le ... \le K_J.$$

Question: What are the values of β and $K_1, ..., K_J$?

Answer:

- β is a function of c_J and α_J only.
- Thresholds are $O(\log R)$ and are a function of all system parameters.

Remark: Thresholds are NOT a function of system state.

Service-Level Differentiation

Two Class Example:

Threshold K	$\sim P\{W_1^N > 0\}$	$\sim P\{W_2^N > 0\}$
a	$\alpha(\beta) \cdot \rho_1^a$	$\alpha(\beta)$
b $\ln N$	$\alpha(\beta)\rho_1^{b \ln N}$	$\alpha(\beta)$
c$\sqrt{N}, \quad c < \beta$	$\alpha(\beta - c) \cdot \rho_1^{c\sqrt{N}}$	$\alpha(\beta - c)$

In all cases $E[W_1 | W_1 > 0] = \Theta(\frac{1}{N})$ and $E[W_2 | W_2 > 0] = \Theta(\frac{1}{\sqrt{N}})$.

Without threshold ($a = 0$), both classes enjoy QED service with the same delay probability.

As the threshold increases, differentiation of service level increases as well, which is manifested through the delay probabilities (but not through average delays).

Example: Logarithmic thresholds improve dramatically the accessibility of high-priority and, at the same time, are not hurting the low-priority (who are still QED-served).

Asymptotic Framework

Consider a sequence of systems indexed by $r = R \to \infty$.
The r^{th} Staffing and Control Problem is:

$$\begin{array}{ll} \text{minimize} & c_1^r \lambda_1^r E W_1^r + ... + c_J^r \lambda_J^r E W_J^r + N^r \\ \text{subject to} & P_{\pi^r}(W_i^r > 0) \le \alpha_i^r, \ \ i = 1, ..., J \\ & N^r \in \mathbb{Z}_+, \quad \pi^r \in \Pi^r \end{array}$$

Assumptions:

- $c_i^r = c_i r^{\gamma_i}, \quad \gamma_i \ge 0, \ \ i = 1, 2, ..., J - 1,$
- $\alpha_i^r = \alpha_i r^{-\delta_i}, \quad \delta_i \ge 0, \ \ 1 = 1, 2, ..., J - 1,$
- $c_J^r = c_J$ and $\alpha_J^r = \alpha_J$.

Definitions:

- Asymptotic Feasibility: $\limsup_{r\to\infty} \frac{P_{\pi^r}(W_i^r > 0)}{\alpha_i^r} \le 1$.
- Asymptotic Optimality: Asymptotic feasibility +

$$\limsup_{r\to\infty} \frac{C^r(N^r, \pi^r) - R}{C^r(N'^r, \pi'^r) - R} \le 1,$$

for all asymptotically feasible $\{(N'^r, \pi'^r)\}$.

Asymptotic Optimality of SRSS and TP

Assumption: Class J is non-negligible

$$\liminf_{r\to\infty} \frac{\lambda_J^r}{\lambda^r} > 0.$$

Theorem: SRSS and TP are asymptotically optimal with:

$$N^{*r} = R + \beta(\alpha_J, c_J)\sqrt{R}$$

(determined by lowest priority J), and

The thresholds $K_1^r < K_2^r < \ldots < K_J^r$ are given by

$$K_j^r - K_{j-1}^r = \frac{\ln \alpha_{j-1}^{*r} - \ln \alpha_j^{*r}}{\ln \rho_{\le j-1}^r}, \; j = 2, \ldots J,$$

$$K_1^r = 1;$$

where $\alpha_j^{*r} = f(\alpha_j^r, \gamma_j)$,

and $\rho_{\le j-1}^r = \sum_{k=1}^{j-1} \lambda_k^r/(\mu N^{*r})$

Note: If $\alpha_j^r \downarrow 0$ or $c_j^r \uparrow \infty$ polynomially with r,

then $K_j^r \uparrow \infty$ as $\ln r$

Properties of the Solution

1) The joint staffing and control problem is decomposed into two separate problems.

2) The staffing is based on R and the low priority parameters: α_J and c_J only.

3) Service differentiation is obtained through a careful selection of the thresholds.

4) The QED regime is obtained as a solution rather than an assumption.

5) State-Space Collapse of class level queue lengths.

6) Performance approximations based on diffusion limits.

Implications of 2)

- Only aggregate demand forecasts are needed.
- When service is being outsourced multidimensional signalling of demand reduces to 1-dimension.

Corollary: $c\mu$ -rule

Corollary: If $\alpha_1^r = ...\alpha_J^r = 1$ for all r and c_i^r is independent of r, $i = 1, ..., J$, then the $c\mu$ rule is asymptotically optimal.

Adding Abandonment

Class i with patience parameter $0 < \theta_i < \infty$.

Assume $N = R + \beta\sqrt{R}$

Cost minimization: π that minimizes total costs

$$\sum_{i=1}^{J} c_i \lambda_i P_\pi\{Ab_i\}$$

Note: This is equivalent to Profit Maximization.

Assumptions:

1) Costs and Impatience follow the same order:

   $c_1 \geq c_2 \geq ... \geq c_J$ and $\theta_1 \geq \theta_2 \geq ... \geq \theta_J$.

2) Non-negligibility of class J: $\liminf_{\lambda\to\infty} \frac{\lambda_J}{\lambda} > 0$

Then TP is an asymptotically optimal control

- static priority $1 > 2 > \ldots > J$
- with logarithmic thresholds

Adding Abandonment (cont)

Constraint Satisfaction: Find N^* and π^* that minimizes staffing costs s.t. abandonment probability constraints.

$$\begin{array}{ll} \text{minimize} & N \\ \text{subject to} & P_\pi\{Ab_i\} \le \eta_i, \quad i = 1, ..., J \\ & N \in \mathbb{Z}_+, \ \pi \in \Pi \end{array}$$

Assumption: η_i are constant (do not change with R).

Optimal Solution:

Server pool decomposition: $N_i = \frac{\lambda_i}{\mu_i}(1 - \eta_i)$.

Allow η_i to scale with R - Solution not trivial.

Summary of Results

1. Square-Root Safety-Staffing and Threshold Priority control are asymptotically optimal and robust.

2. Staffing is based on aggregate demand and low priority delay and cost parameters only.

3. Service level differentiation obtained through threshold selection.

4. Performance approximations allow to evaluate the costs and benefits of cross-training.

5. The QED regime is obtained as a solution rather than an assumption.

Extensions (future research)

- The general abandonment model.

- Combine with the $\wedge$-design to study the N-design.