# Design and Control of the M/M/N Queue with Multi-Class Customers and Many Servers

M.Sc. Thesis

Itay Gurvich

Thesis Advisors: Prof. Avishai Mandelbaum

Prof. Mor Armony

June 2004

# Contents

# Staffing (+SBR): How Many Agents?

- Fundamental problem in service operations / call centers:

- People = 70% costs of running call centers, employing
  3% U.S. workforce; 1000's agents in a "single" Call Center.

Reality

- Workforce Management (WFM) is M/M/N-based
- Reality is complex and becoming even more so
- Solutions are urgently needed
- Technology enables smart systems
- Theory lags significantly behind needs

» Ad-hoc methods: heuristics, simulation-based

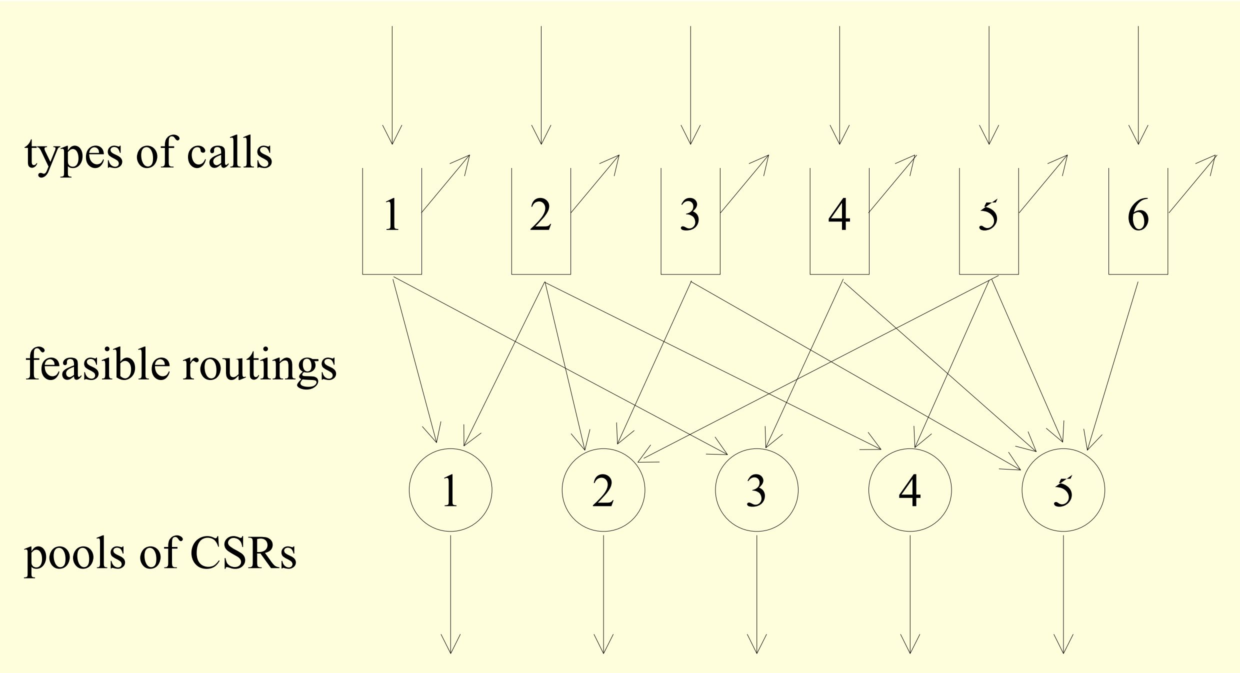
Progress is based on

- Small yet significant models for theoretical insight

the research of which gives rise to

- Principles, Guidelines, Tools: Service Engineering

# 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.

# Staffing (+SBR): How Many Agents?

- Fundamental problem in service operations / call centers:

- People = 70% costs of running call centers, employing 3% U.S. workforce; 1000's agents in a "single" Call Center.

Reality

- Workforce Management (WFM) is M/M/N-based
- Reality is complex and becoming even more so
- Solutions are urgently needed
- Technology enables smart systems
- Theory lags significantly behind needs

» Ad-hoc methods: heuristics, simulation-based

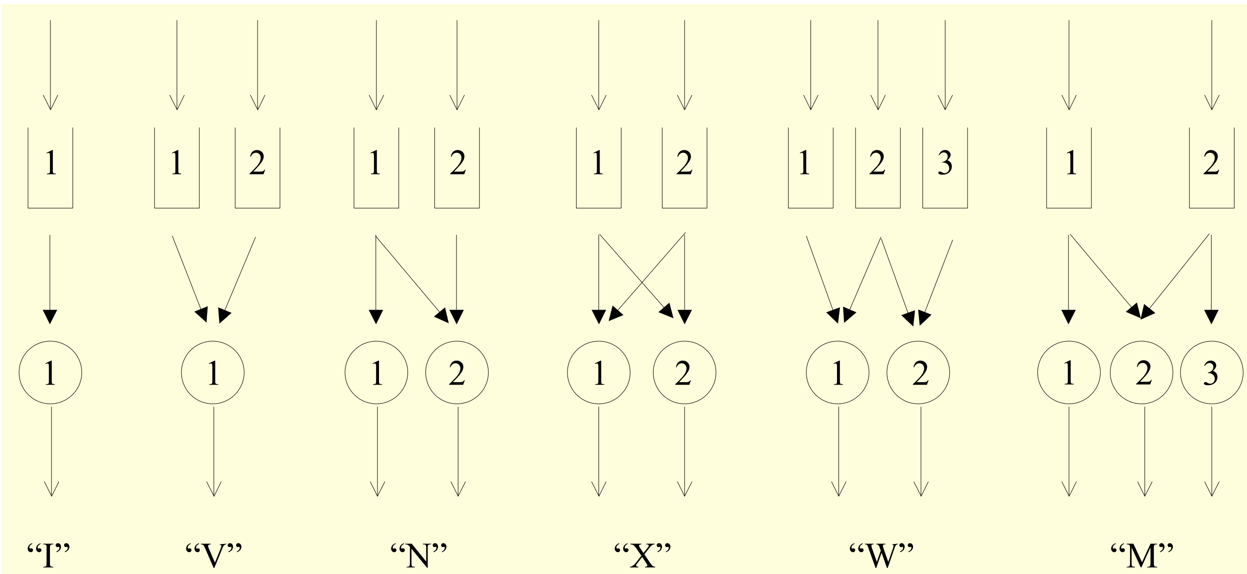
Progress is based on

- Small yet significant models for theoretical insight

the research of which gives rise to

- Principles, Guidelines, Tools: Service Engineering

# Design "Building-Blocks"




**Literature on $I$, $V$ and $\wedge$-designs:**

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

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

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

# QED 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 $\lim_{N\to\infty} P_N\{\text{Wait} > 0\} = \alpha, \qquad 0 < \alpha < 1;$
- Server $\lim_{N\to\infty} \sqrt{N}(1-\rho_N) = \beta, \qquad 0 < \beta < \infty;$
- Manager $N \approx R + \beta\sqrt{R}, \qquad R = \lambda \times \text{E(S)}$ large;

Here $$\alpha = \left[1 + \frac{\beta\phi(\beta)}{\varphi(\beta)}\right]^{-1},$$

where $\varphi(\cdot)/\phi(\cdot)$ is the standard normal density/distribution.

Extremes:

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

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

# $\sqrt{\cdot}$ Safety-Staffing: Performance

$R = \lambda \times E(S)$ Offered load (Erlangs)

$N = R + \underbrace{\beta\sqrt{R}}$ $\beta$ = "service-grade" $> 0$

$= R + \Delta$ $\sqrt{\cdot}$ safety-staffing

Expected Performance:

$$\% \text{ Delayed} \approx P(\beta) = \left[1 + \frac{\beta\phi(\beta)}{\varphi(\beta)}\right]^{-1}, \quad \beta > 0$$ Erlang-C

$$\text{Congestion index } = E\left[\frac{\text{Wait}}{E(S)} \,\middle|\, \text{Wait} > 0\right] = \frac{1}{\Delta}$$ ASA

$$\% \left\{\frac{\text{Wait}}{E(S)} > T \,\middle|\, \text{Wait} > 0\right\} = e^{-T\Delta}$$ TSF

$$\text{Servers' Utilization} = \frac{R}{N} \approx 1 - \frac{\beta}{\sqrt{N}}$$ Occupancy

# Dimensioning M/M/N: $\sqrt{\cdot}$ Safety-Staffing

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. $\mathbf{P(Wait > 0) \leq \alpha}$.**

- $\alpha \approx \mathbf{1}$ : 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 $M/M/N$, $N \uparrow \infty$.

# Economics: $\sqrt{\cdot}$ Safety-Staffing

Optimal $$\mathbf{N}^* \approx \mathbf{R} + y^*(C)\ \sqrt{\mathbf{R}}$$

where $\mathbf{C}$ = delay/waiting costs

Here $$y^*(\mathbf{C}) \approx \left(\frac{C}{1 + C\left(\sqrt{\pi/2} - 1\right)}\right)^{1/2}, \qquad 0 < C < 10$$

$$\approx \left(2 \ln \frac{C}{\sqrt{2\pi}}\right)^{1/2}, \qquad C \text{ large.}$$

Performance measures: $\Delta = y^* \sqrt{\mathrm{R}}$ safety staffing

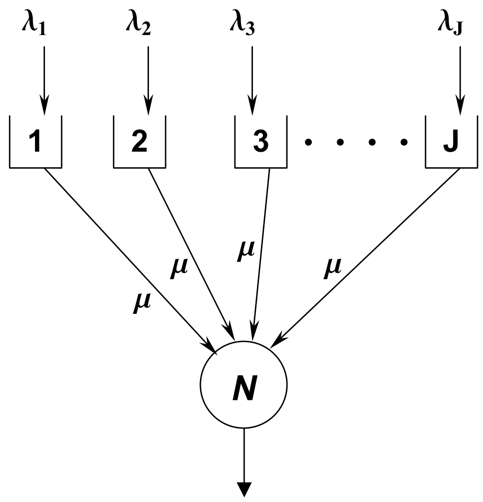
$$\mathrm{P}\{\text{Wait} > 0\} \approx \mathbf{P}(y^*) = \left[1 + \frac{y^* \phi(y^*)}{\varphi(y^*)}\right]^{-1}$$ **Erlang-C**

$$\text{TSF} = \mathrm{P}\left\{\frac{\text{Wait}}{\mathrm{E(S)}} > \mathrm{T} \,\middle|\, \text{Wait} > 0\right\} = e^{-\mathrm{T}\Delta}$$

$$\text{ASA} = \mathrm{E}\left[\frac{\text{Wait}}{\mathrm{E(S)}} \,\middle|\, \text{Wait} > 0\right] = \frac{1}{\Delta}$$

$$\text{Occupancy} = 1 - \frac{\Delta}{\mathrm{N}} \approx 1 - \frac{y^*}{\sqrt{\mathrm{N}}}$$

# The $V$-Design




- $J$ customer classes: arrivals Poisson$(\lambda_j)$.
- $N$ **iid** servers: service durations Exp$(\mu)$.

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

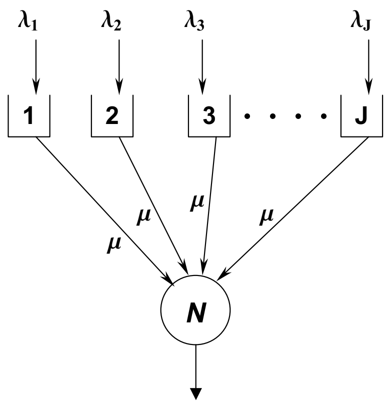
minimize $\quad N$

subject to $\quad \exists \pi \in \Pi$

$$P_\pi(W_i > 0) \le \alpha_i,\ 0 < \alpha_i < 1, \quad i = 1, ..., J$$

$$N \in \mathbb{Z}_+$$

# The $V$-Design




Add waiting costs $C_1 > C_2 > \dots$

Optimization: $N^*$ that minimizes total costs

$$\sum_{i=1}^{J} C_i \lambda_i W_i + N$$

**Optimal Control**: minimize waiting costs "$\sum_{i=1}^{J} C_i \lambda_i W_i(\cdot)$"

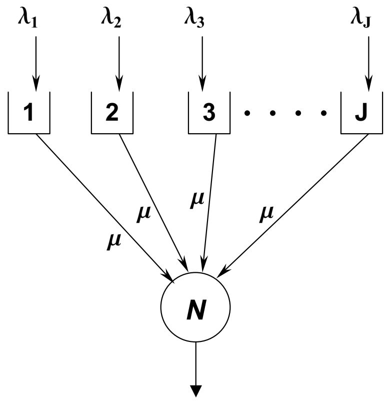
Yahalom 2003 - Blackwell optimality:

- **Static priorities $\mathbf{1 > 2 > \dots}$ with thresholds**

$$\mathbf{0 = K_1(x) \leq K_2(x) \leq \dots}$$

  i.e. a class-$j$ customer is served when the system state is $x$ if she is of the present highest-priority and the number of idle servers is more than $K_j(x)$.

# The $M/M/\{K_i\}$ Model




**Static priorities $1 > 2 > \ldots$ with thresholds**

$$\mathbf{K_1 \leq K_2 \leq \ldots K_J}$$

i.e. a class-$j$ customer is served when it is of the present highest-priority and the number of idle servers is more than $K_j$.

Let $\mathbf{K} \triangleq \mathbf{K_J}$.

Performance analysis of $M/M/\{K_i\}$ in steady-state (Schaack & Larson 1986).

# $M/M/\{K_i\}$-Performance Analysis

Two-Class case using Larson (Let $M = N - K$)

$$
\begin{aligned}
P_{M+} &= P_0 \left( \sum_{n=M}^{N-1} \left(\frac{\lambda_1+\lambda_2}{2}\right)^M \left(\frac{\lambda_1}{\mu}\right)^n \frac{1}{n!} \frac{1}{1-\lambda_2 h_1(M)} \right. \\
&+ \left. \left(\frac{\lambda_1+\lambda_2}{2}\right)^M \left(\frac{\lambda_1}{\mu}\right)^N \frac{1}{N!} \frac{1}{1-\frac{\lambda_1}{N\mu}} \frac{1}{1-\lambda_2 h_1(M)} \right) .
\end{aligned}
$$

where

$$
\begin{aligned}
P_0 &= \left[ \sum_{n=0}^{M-1} \left(\frac{\lambda_1+\lambda_2}{\mu}\right)^n \frac{1}{n!} + \sum_{n=M}^{N-1} \left(\frac{\lambda_1+\lambda_2}{2}\right)^M \left(\frac{\lambda_1}{\mu}\right)^n \frac{1}{n!} \frac{1}{1-\lambda_2 h_1(M)} \right. \\
&+ \left. \left(\frac{\lambda_1+\lambda_2}{2}\right)^M \left(\frac{\lambda_1}{\mu}\right)^N \frac{1}{N!} \frac{1}{1-\frac{\lambda_1}{N\mu}} \frac{1}{1-\lambda_2 h_1(M)} \right]^{-1} .
\end{aligned}
$$

and

$$
h_1(M) = \frac{1}{M\mu} + \sum_{k=2}^{N-M} \frac{{\lambda_1}^{k-1}}{\mu^k \prod_{j=0}^{k-1}(M+j)} + \frac{{\lambda_1}^{N-M}}{(N\mu-\lambda)\,\mu^{N-M} \prod_{l=1}^{N-M}(N-l)} .
$$

**Complicated**.

Instead:

Consider a sequence of $M/M/\{K_i\}$ systems, such that $\lambda^r$, $K^r \triangleq K^r_J$ and $N^r$ all go to $\infty$ in a certain manner.

# $M/M/\{K_i\}$-Performance Analysis : Steady State (1).

**Proposition 1** *Fix $r$ and assume $K^r > 0$. We then have the following:*

1. *The threshold system is stable if $\sum_{i=1}^{J} \lambda_i^r < (N^r - K^r)\mu$.*

2. *If $\sum_{i=1}^{J} \lambda_i > \frac{(N^r - K^r)\mu}{(1-\delta^r)} \wedge N$ where $\delta^r \leq \frac{\lambda_J^c/((N-K)\mu)}{(N-K)(1-\lambda_J^c/((N-K)\mu))}$, $\lambda_J^c = \sum_{i=1}^{J-1} \lambda_i$.*
   *The system is not stable i.e $Q^r(t) \to \infty$ as $t \to \infty$.*

*If $K^r \equiv 0$ (static priority), Condition 1 is necessary and sufficient.*

$$\boxed{R - (N - K) = O(1)}$$

i.e. if $N \approx R + \Delta$

$$\boxed{K < \Delta + O(1)}$$

# $M/M/\{K_i\}$-Performance Analysis : Steady State (2).

Assume that $\lambda_J^r/\lambda^r \to \epsilon > 0$ and define $\rho_C^r = \lambda^r/(N^r - K^r)\mu$, then:

**Proposition 2 : QED Characterization**

- *Customer:* $\lim_{r\to\infty} P\{W_J^r > 0\} = \alpha, \quad 0 < \alpha < 1;$
- *Server:* $\lim_{r\to\infty} \sqrt{N^r}\,(1 - \rho_C^r) = \beta, \quad 0 < \beta < \infty;$
- *Manager:* $N^r - K^r \approx R + \beta\sqrt{R}\,, \qquad R = \lambda/\mu$ *large.*

In that case

$$\frac{Y^r(\infty) - (N^r - K^r)}{\sqrt{N^r}} \Rightarrow X(\infty)$$

Where $Y^r$ is the total number of customers in system $r$, and $X(\infty)$ has a density:

$$f(x) = \begin{cases} \exp\{-\beta x\}\alpha(\beta) & x \geq 0 \\ \frac{\phi(\beta+x)}{\Phi(\beta)}(1 - \alpha(\beta)) & x < 0 \end{cases}$$

Also Let $Q_i^r$ be the queue of class $i$, then:

$$\frac{1}{\sqrt{N^r}} Q_i^r(\infty) \Rightarrow 0,\ i = 1, ..., J-1$$

# $M/M/\{K_i\}$-Performance Analysis : Steady State (3).

**Corollary 3**

$$\sqrt{N^r} W_J^r(\infty) \Rightarrow W \tag{1}$$

*Where*

$$W \sim \begin{cases} \exp(\epsilon\mu\beta) & w.p.\ \alpha(\beta) \\ 0 & otherwise \end{cases} \tag{2}$$

**Proposition 4** *For every* $r > 0$

$$1 \leq \frac{P\{W_i^r(\infty) > 0\}}{P\{W_J^r(\infty) > 0\} \cdot \prod_{j=i}^{J-1} (\rho_k^r)^{K_{j+1}^r - K_j^r}} \leq \left( \frac{N^r}{N^r - K^r} \right)^{K^r}, \tag{3}$$

*in particular for* $K^r = o(\sqrt{N^r})$ *and assuming* $\alpha(\beta) > 0$ *we have*

$$P\{W_i^r(\infty) > 0\} \sim \alpha(\beta) \cdot \prod_{k=i}^{J-1} (\rho_k^r)^{K_{k+1}^r - K_k^r} \tag{4}$$

# Service-Level Differentiation

Two Class Example:

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

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).

## $M/M/\{K_i\}$-Performance Analysis : Steady State (4).

**Proposition 5** *Assume that class $J$ is non-negligible. Then, for all $k = 1, ..., J-1$*

$$N^r[W_k^r|W_k^r > 0] \Rightarrow [W_k|W_k > 0] \tag{5}$$

$[W_k|W_k > 0]$ *has the Laplace transform:*

$$\frac{\mu(1-\sigma_k)(1-\tilde{\gamma}(s))}{s - \widehat{\lambda}_k + \widehat{\lambda}_k\tilde{\gamma}(s)} \tag{6}$$

*where* $\sigma_j = \lim_{r\to\infty} \sum_{i=1}^{j} \rho_i^r$, *and*

$$\tilde{\gamma}(s) = \frac{s+\mu}{2b_k\mu} + \frac{1}{2} - \sqrt{\left(\frac{s+\mu}{2b_k\mu} + \frac{1}{2}\right)^2 - \frac{1}{b_k}} \tag{7}$$

*where* $b_k = \lim\limits_{r\to\infty} \frac{\sum_{i=1}^{k-1} \lambda_i^r}{N^r}$

*Also,*

$$\begin{aligned} &N^r E[W_k^r|W_k^r > 0] \to [\mu(1-\sigma_k)(1-\sigma_{k-1})]^{-1} \\ &(N^r)^2 E[(W_k^r)^2|W_k^r > 0] \\ &\qquad \to 2(1-\sigma_k\sigma_{k-1})\left[(\mu)^2(1-\sigma_k)^2(1-\sigma_{k-1})^3\right]^{-1} \end{aligned} \tag{8}$$

$\boxed{\text{Waiting time of High Priorities is O(1/N)}}$

$\boxed{\text{Queue of High Priorities is O(1)}}$

# Asymptotic Optimality (1)

$C^r(N^r, \pi^r)$ - cost with $N^r$ servers and policy $\pi^r$

**Definition:** $\{N^r, \pi^r\}$ is asymptotically optimal with respect to $\bar{\lambda}^r$, if,

- *Asymptotic feasibility*:

$$\limsup_{r \to \infty} P_\pi\{W_i^r > 0\} \leq \alpha_i \,, \forall i = 1, ..., J$$

- *Asymptotic Optimality*: If we take any other sequence of policies $\{N_2^r, \pi_2^r\}$ that is asymptotically feasible then

$$\liminf_{r \to \infty} \frac{C^r(N_2^r, \pi_2^r) - \underline{C}^r}{C^r(N^r, \pi^r) - \underline{C}^r} \geq 1$$

# Optimal Control (1): QED Solution

$$\begin{array}{ll} \text{minimize} & N \\ \text{subject to} & P_\pi(W_i > 0) \le \alpha_i \\ & N \in \mathbb{Z}_+ \end{array}$$

Asymptotically optimal (staffing + scheduling) as follows:

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

(determined by lowest priority **J**)

$\pi^*$: static priority $1 > 2 > \ldots > J$, with thresholds $S_1 < S_2 < \ldots < S_J$, given by

$S_j = S_{j-1} + \ln \frac{\alpha_{j-1}}{\alpha_j} / \ln \rho_{j-1}^+ \, , \, j = 2, \ldots J,$
$S_1 = 1;$

i.e. a class $j$ customer served *iff* it is of the present highest priority and the number of idle agents is $S_j$ or more.

(Here $R = \sum_j \lambda_j / \mu, \quad \rho_j^+ = \sum_{k=1}^j \lambda_k / (\mu N^*)$)

Note: allowing $\alpha_j^N \downarrow 0$ polynomially with $N$,
requires $S_j^N \uparrow \infty$ as $\ln N$

# Optimal Control (2): QED Solution

Optimization: $N^*$ that minimizes total costs

$$\sum_{i=1}^{J} C_i \lambda_i W_i + N$$

Assume $C_i$'s are constants

and $\liminf_{\lambda \to \infty} \frac{\lambda_J}{\sum_j \lambda_j} = \epsilon > 0$ (non-negligible)

Then asymptotically optimal **non-preemptive** staffing and control is

- Staff with $N = R + \beta\sqrt{R}$ , $\beta = y^*(C_J)$,
- non-idling, and
- static priority $1 > 2 > \ldots > J$

**Starting point:** For any non-idling strategy, the total work in system $(\sum_j W_j)(\cdot)$ is that of an $M/M/N$, with parameters $\lambda = \sum_j \lambda_j$, $\mu$, $N$.

# Where are the Thresholds ?

Optimization: $N^*$ that minimizes total costs

$$\sum_{i=1}^{J} C_i(\lambda)\lambda_i W_i + N$$

Assume $C_J(\lambda) \equiv C_J$ is constant

and $C_i(\lambda) = d_i\lambda^{\gamma_i}$, $d_i, \gamma_i > 0$, $i \neq J$

Then, asymptotically optimal is

- Staff with $N = R + \beta\sqrt{R}$ , $\beta = y^*(C_J)$,
- Idling - $M/M/\{K_i\}$ with logarithmic thresholds.

# $M/M/\{K_i\}$-Performance Analysis : Diffusion Limits (1).

For $r = 1, 2, ....$ define the centered and scaled process

$$X^r(t) = \frac{Y^r(t) - (N^r - K^r)}{\sqrt{N^r}}$$

$$\lim_{r\to\infty} \sqrt{N^r}(1 - \rho_C^r) = \beta\,,\ 0 < \beta < \infty$$

where

$$\rho_C^r = \frac{\lambda^r}{(N^r - K^r)\mu}$$

**Proposition 6** *Assume that $X^r(0) \Rightarrow X(0)$, then*

$$X^r \Rightarrow X$$

*where $X$ is a diffusion process with infinitesimal drift given by*

$$m(x) = \begin{cases} -\beta\mu & x \geq 0 \\ -(\beta + x)\mu & x \leq 0 \end{cases}$$

*and state independent infinitesimal variance $\sigma^2 = 2\mu$.*

Remark: This is the Halffin-Whitt limit for the single class model with $N - K$ servers.

# $M/M/\{K_i\}$-Performance Analysis : Diffusion Limits (2).

**Corollary 7** ***State Space Collapse*** *Denote by* $\mathcal{E}^r(t)$ *the number of busy servers above the level of* $N^r - K^r$*, i.e.* $\mathcal{E}^r(t) = [Z^r(t) - (N^r - K^r)]^+$ $\big($*where* $[x]^+ = \max\{x, 0\}\big)$ *. Assume*

$$\lim_{r\to\infty} \frac{\lambda_k^r}{\lambda^r} = a_k,\ k = 1, ..., J; \quad a_J > 0,\ a_i \geq 0,\ i = 1, ..., J-1$$

*Then*

$$\frac{1}{\sqrt{N^r}}\mathcal{E}^r(t) \Rightarrow 0$$

$$\frac{1}{\sqrt{N^r}}Q_i^r(t) \Rightarrow 0,\ \forall i \leq J-1$$

$$\frac{1}{\sqrt{N^r}}Q_J^r(t) \Rightarrow X^+$$

**Corollary 8** *Let* $W_i^r(t)$ *be the virtual waiting time process for class* $i$*. If*

$$\exists -\infty < c < \infty\ :\ \sqrt{N}\Big(\frac{\lambda_J^r}{N^r} - a_j\mu\Big) \to c\,,$$

*then*

$$\sqrt{N^r}W_J^r \Rightarrow \frac{1}{a_J\mu}(X \vee 0)$$

$\boxed{Q_i, i < J \text{ disappears in the } \sqrt{N} \text{ scaling}}$

$\boxed{Q_J \text{ is the whole queue}}$

# Extensions - What about Abandonment ?

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

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

Optimization: $N^*$ that minimizes total costs

$$\sum_{i=1}^{J} C_i \lambda_i P_i\{Ab\} \tag{9}$$

$C_i$'s are const, s.t $C_i > C_j$ whenever $\theta_i > \theta_j$.

$\liminf_{\lambda\to\infty} \frac{\lambda_J}{\sum_j \lambda_j} = \epsilon > 0$

Then asymptotically optimal **non-preemptive** control is

- non-idling, and
- static priority $1 > 2 > \ldots > J$

Add logarithmic thresholds if $C_i$, $i \neq J$ scale polynomially.

# Extensions - What about Abandonment ?

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

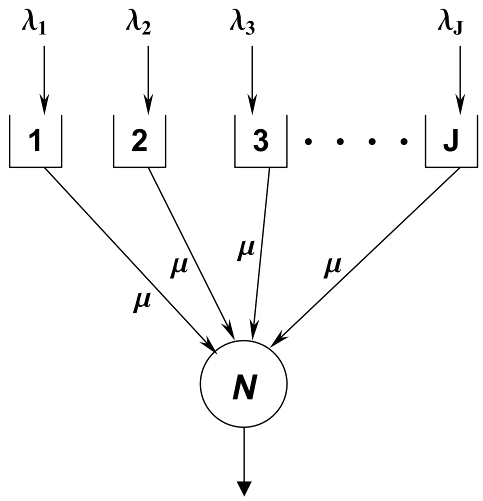
$$
\begin{array}{ll}
\text{minimize} & N \\
\text{subject to} & P_i\{Ab\} \le \alpha_i,\ 0 < \alpha_i < 1, \quad i = 1, ..., J \\
 & N \in \mathbb{Z}_+
\end{array}
$$

Optimal Solution:

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

Allow $\alpha_i$ to scale with $\lambda$ - Solution not trivial.

# Summary of Results

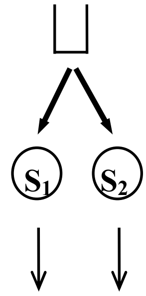



1. For both satisfization and Optimization the asymptotically optimal policy is $M/M/\{K_i\}$.

2. State space collapse allows a complete asymptotic analysis of the $M/M/\{K_i\}$ model.

# Back Up

# Reversed-V Design: Pure Routing




Homogeneous Customers

Heterogeneous Agents: **S2 = Faster**

Optimal Routing: **"Slow-Server" phenomenon** (Rykov)

- S2(=Fast) always employed, if possible;

- S1(= Slow) employed if # in queue exceeds a threshold.

**QED** regime: $\sqrt{\cdot}$ Safety-Staffing – see below (Armony)

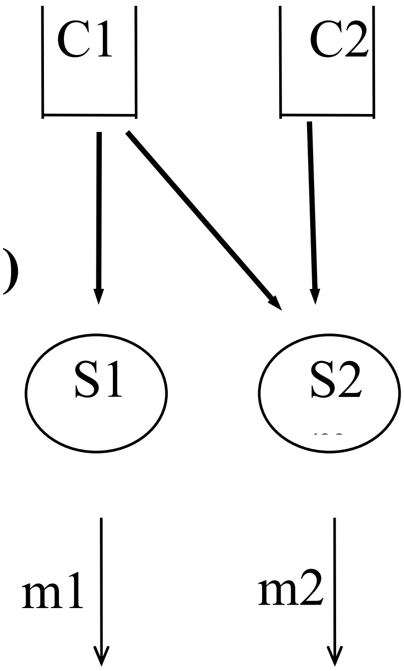
- No threshold needed: just have all servers work when possible, ensuring that the "fast" get the priority.

Asymptotically optimal staffing:

1. Given a delay probability, determine S1 + S2 via $\sqrt{\cdot}$ Safety.
2. Given staffing costs, determine S1 / S2.

Distributed call centers: in progress.

# N-Design: Routing and Scheduling




Heterogeneous Customers: **C2=VIP**

Heterogeneous Agents: **S2 = Faster (m1>m2)**

Costs: $\mathbf{D_1, D_2}$ Delay costs
$\mathbf{H_1, H_2}$ Staffing costs

Assume: $\mathbf{D_2 \gg D_1}$ **(Truly VIP)**

Assume: $\mathbf{H_1\, m_1 < H_2\, m_2}$ (Otherwise V)

**QED** regime: $\sqrt{\cdot}$ Safety Staffing – see below (Gurvich).

- $(C_1, C_2; S_2)$ operate as V-model, with "idle-thresholds"

- $(C_1; S_1, S_2)$ operate as $\Lambda$, but without "queue-thresholds"

Asymptotically optimal staffing:

1. Given a delay probability (service level), determine $\mu_1 S1 + \mu_2\, S2$ via $\sqrt{\cdot}$ Safety;
2. Given staffing costs, determine S1 / S2 via Math. Prog.

Ultimately: $\sqrt{\cdot}$ Safety-Staffing is asymptotically optimal.