QED Control and Staffing: The Cases of a Single Customer Class or a Single Server Type

with

Mor Armony

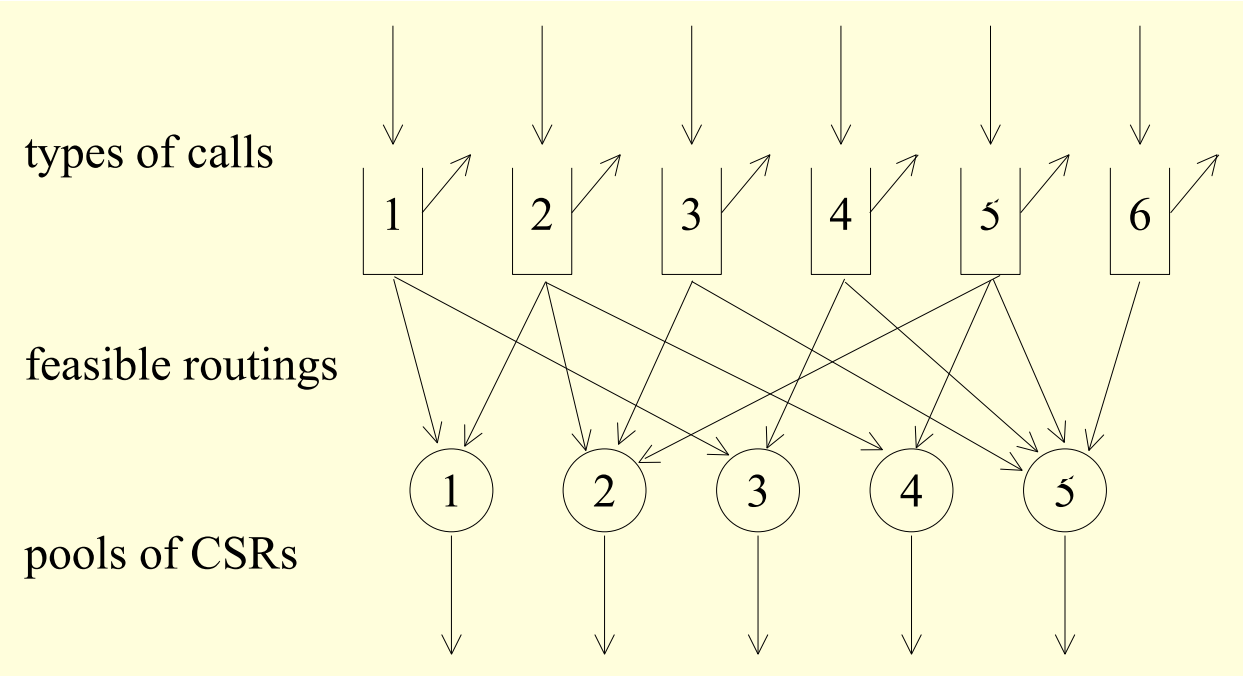
Rami Atar

Itay Gurvich

Marty Reiman

Source: SBR_FINAL_Sep3.tex

Multi-Skill Call-Centers

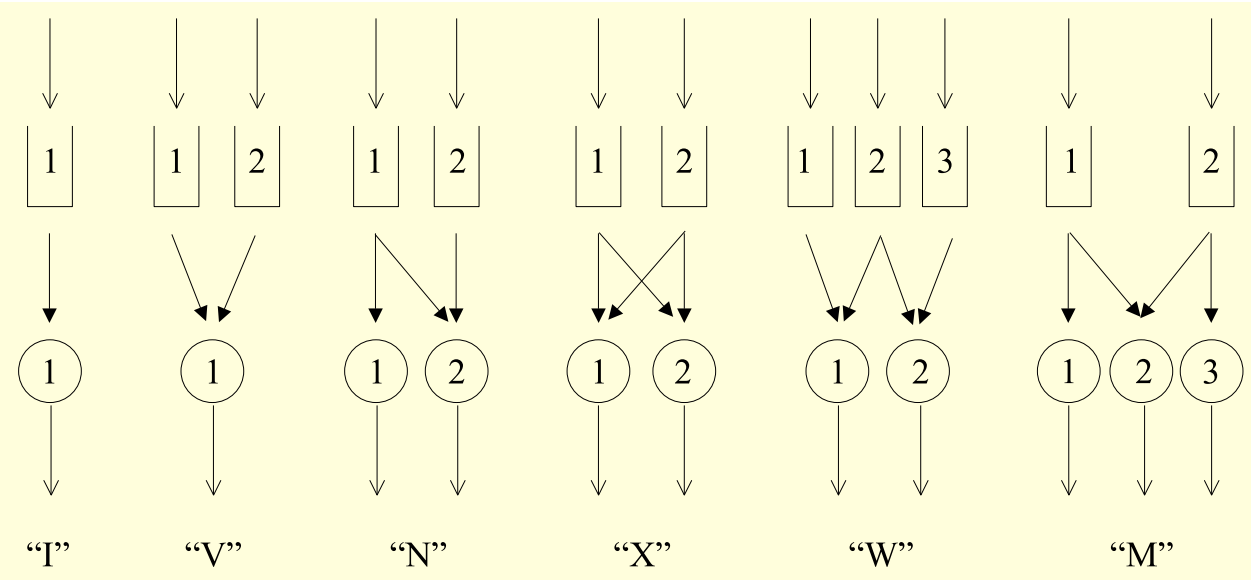


Main Operational Issues (Given a Forecast of Workload):

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

Very Complex: Hence treated hierarchically and unilaterally.

Design “Building-Blocks”



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

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

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

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

QED M/M/N in Steady State

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: $\lim_{N\to\infty} P_N\{Wait > 0\} = \alpha, \quad 0 < \alpha < 1;$
- Server: $\lim_{N\to\infty} \sqrt{N}\,(1 - \rho_N) = \beta, \quad 0 < \beta < \infty;$
- 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**

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

Borst, M. & Reiman ('02)

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

Optimization: N^* that minimizes total costs

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

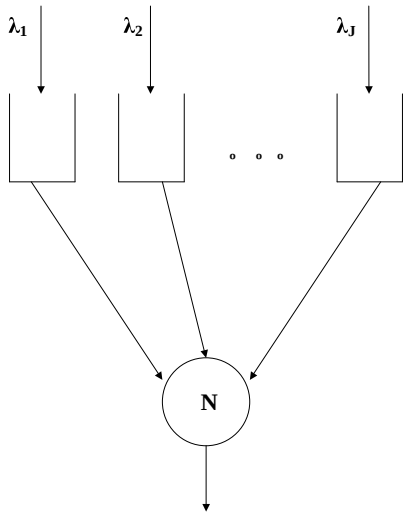
Satisfization: N^* that minimizes staffing costs s.t. delay constraints.

Here: $\mathbf{N^*}$ **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 $N \approx R + \delta R$
- $\mathbf{0 < \alpha < 1}$: 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).
- Waiting costs $C_1 > C_2 > \ldots$

Optimal Control: minimize waiting costs "$\sum_{j=1}^{J} C_j W_j(\cdot)$"

Preemptive (Coupling): non-idling with static priorities $1 > 2 > \ldots$

Non-preemptive (Yahalom 2003 - Blackwell optimality):

- **Static priorities** $1 > 2 > \ldots$ **with thresholds** $\mathbf{S_1 > S_2 > \ldots}$ i.e. a class-j customer served if it is of the present highest-priority and the number of idle servers is S_j or more.
- Performance analysis in steady-state (Schaack & Larson 1986).

Optimal Control: QED Solution

Atar, M., Reiman ('02, '03); Gurvich ('03)

Assume $N = R + \beta\sqrt{R}$ $(R = \sum_j \lambda_j/\mu)$

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

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

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

Proof: Suffices asymptotic equivalence of Preemptive and Non-Preemptive.

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 $\lambda = \sum_j \lambda_j$, μ, N.

Asymptotic Equivalence

- Total work in system $\stackrel{d}{=} M/M/N$, if non-idling

- Under static priority (preemptive or non-preemptive), the lowest priority customers (Class J) "enjoy" QED service. More precisely,

$$W_J^N \stackrel{d}{=} \Theta(\frac{1}{\sqrt{N}})$$

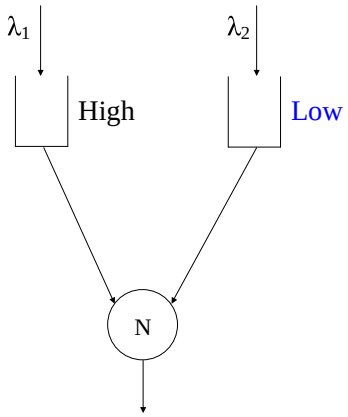
- Under static priority (preemptive or non-preemptive), the high priority customers (classes $1, \ldots, J-1$) enjoy Q-driven service (light traffic). More precisely,

$$W_j^N | W_j^N > 0 \stackrel{d}{=} \Theta(\frac{1}{N}) \quad , j = 1, \ldots J-1.$$

- Multiplying total work by $\sqrt{N}$ (preemptive **or** non-preemptive) yields asymptotic equivalence, $N \uparrow \infty$.

Asymptotic Equivalence: What's Going On?

Low Priority View



1. While **waiting**, **Low Priority** customers "see" an $\mathbf{M/G/1}$ queue: $W_2|W_2 > 0 \stackrel{d}{=} W_{M/G/1}|W > 0$.

Non-Preemptive: $G_{NP} \stackrel{d}{=} M(\lambda_1)/M(N\mu)/1$ busy period. Thus, $E(G_{NP}) = \frac{1}{N\mu(1-\rho_1)}$, where $\rho_1 = \frac{\lambda_1}{N\mu}$.

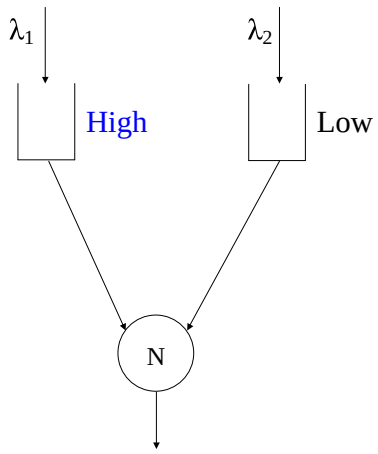
Preemptive: $G_P \stackrel{d}{=}$ Geometric number of busy periods and $Exp(N\mu)$, resulting in $E(G_P) = \cdots = \frac{1}{N\mu(1-\rho_1)}$.

2. When some servers are **idle** - same Birth & Death process for Preemptive and Non-Preemptive.

3. Rigorously: Paste excursions (as in Whitt 2003), to show $Q_1 + Q_2 \stackrel{d}{\approx} Q_2$ (queue-length)

Asymptotic Equivalence: What's Going On?

High Priority View



Preemptive: "See" $M(\lambda_1)/M(\mu)/N$ in light traffic.

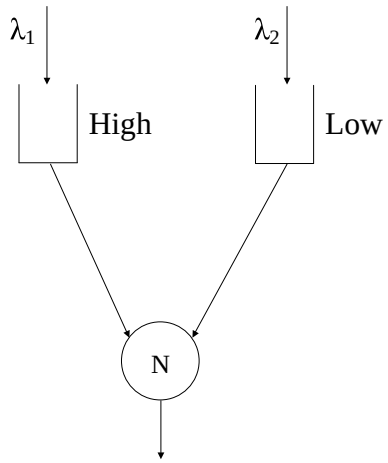
Non-Preemptive: Don't wait if less than N servers busy.

Given wait - "See" $M(\lambda_1)/M(N\mu)/1$ in light traffic.

Rigorously:

1. Prove convergence of $Q_1 + Q_2$ (QED M/M/N)
2. Prove convergence of High Priority queue Q_1 to zero: Since both Non-Preemptive and Preemptive "see" a queue in light traffic
3. Conclude $Q_1 + Q_2 \stackrel{d}{\approx} Q_2$ (queue length)

Where are the Thresholds ?



Assume $N = R + \beta\sqrt{R}$ (QED staffing)

$$\rho_1 = \limsup_{N\to\infty} \frac{\lambda_1^N}{N\mu} < 1.$$

Apply a threshold S^N: Serve Low Priority (Class 2) if the number of idle servers is S^N or more.

Stability requires $\limsup_{N\to\infty} S^N/\sqrt{N} \le \beta$. Then

$$E[W_1^N | W_1^N > 0] = \theta(\frac{1}{N}) \,,\; E[W_2^N | W_2^N > 0] = \theta(\frac{1}{\sqrt{N}}).$$

for all such thresholds. **However**,

Service-Level Differentiation

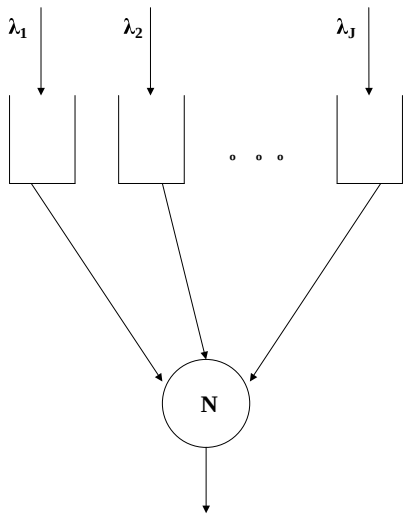
Threshold	$\sim P\{W_1^N > 0\}$	$\sim P\{W_2^N > 0\}$
a	$\alpha(\beta) \cdot \rho_1^a$	$\alpha(\beta)$
b ln N	$\alpha(\beta) \cdot N^{b \ln \rho_1}$	$\alpha(\beta)$
c$\sqrt{N}$	$\alpha(\beta - c) \cdot \rho_1^{c\sqrt{N}}$	$\alpha(\beta - c)$

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).

Dimensioning the V-Model



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

The staffing problem:

Given $0 < \alpha_1 < \alpha_2 < \ldots \alpha_J < 1$,

Min N

s.t. $P_\pi(W_j(\infty) > 0) \leq \alpha_j\,, \quad j = 1, ..., J$

for some scheduling policy π

(Could also minimize $cN + \sum_j d_j \lambda_j E W_j(\infty)$)

Dimensioning V: QED Solution

(Gurvich, 2003)

Asymptotically optimal (staffing + scheduling) as follows:

$$N^* = R + P^{-1}(\alpha_J)\sqrt{R}$$

(determined by lowest priority **J**)

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

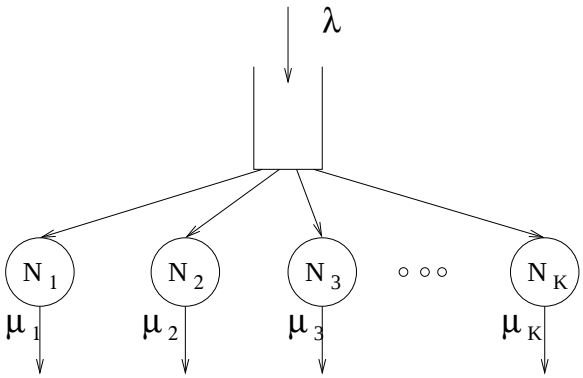
$$S_j = S_{j-1} + \ln \frac{\alpha_{j-1}}{\alpha_j} / \ln \rho_{j-1}^+ \,, \; j = 2, \ldots J,$$
$$S_1 = 1;$$

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, or exponentially requires $S_j^N \uparrow \infty$ as $\ln N$, or $\sqrt{N}$

The $\wedge$-Design (Armony & M., 2003)



- Single customer class: arrivals Poisson(λ).
- K server pools: pool k has N_k **iid** servers;
  service durations Exp with rates $\mu_1 < \mu_2 < ... < \mu_K$ (fastest).

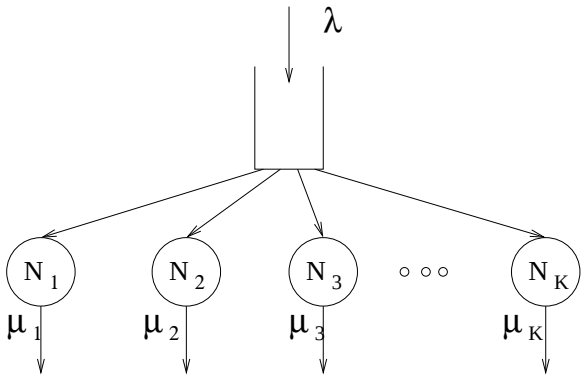
The Focus: Staffing

- How many servers of each type are needed?

Design Concerns

- What is the advantage (if any) of differentiated service rates?
- How much (de)centralization?

Staffing the $\wedge$-Model



M/M/N dimensioning requires modification:

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

WLOG - **Two** server pools (K=2).

The **Staffing Problem**

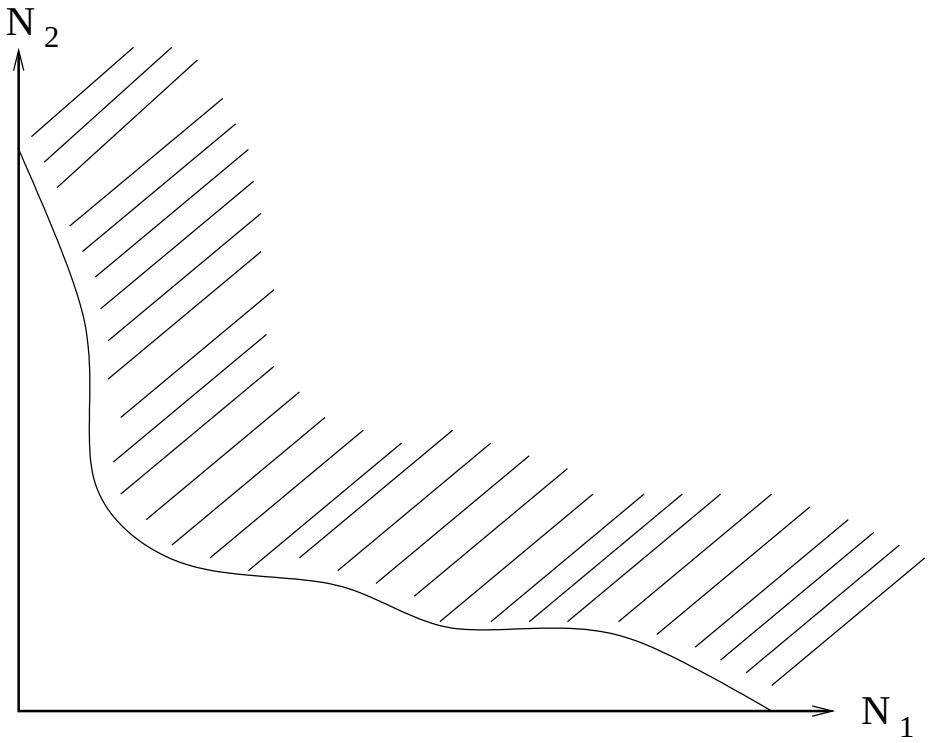
Minimize $C_1(N_1) + C_2(N_2)$

Subject to $P_\pi(\text{wait } > 0) \leq \alpha,$ for some routing policy π;

$N_1, N_2 \in \mathbb{Z}_+.$

"Solution": $\mu_1 N_1 + \mu_2 N_2 = \lambda +$ safety-staffing

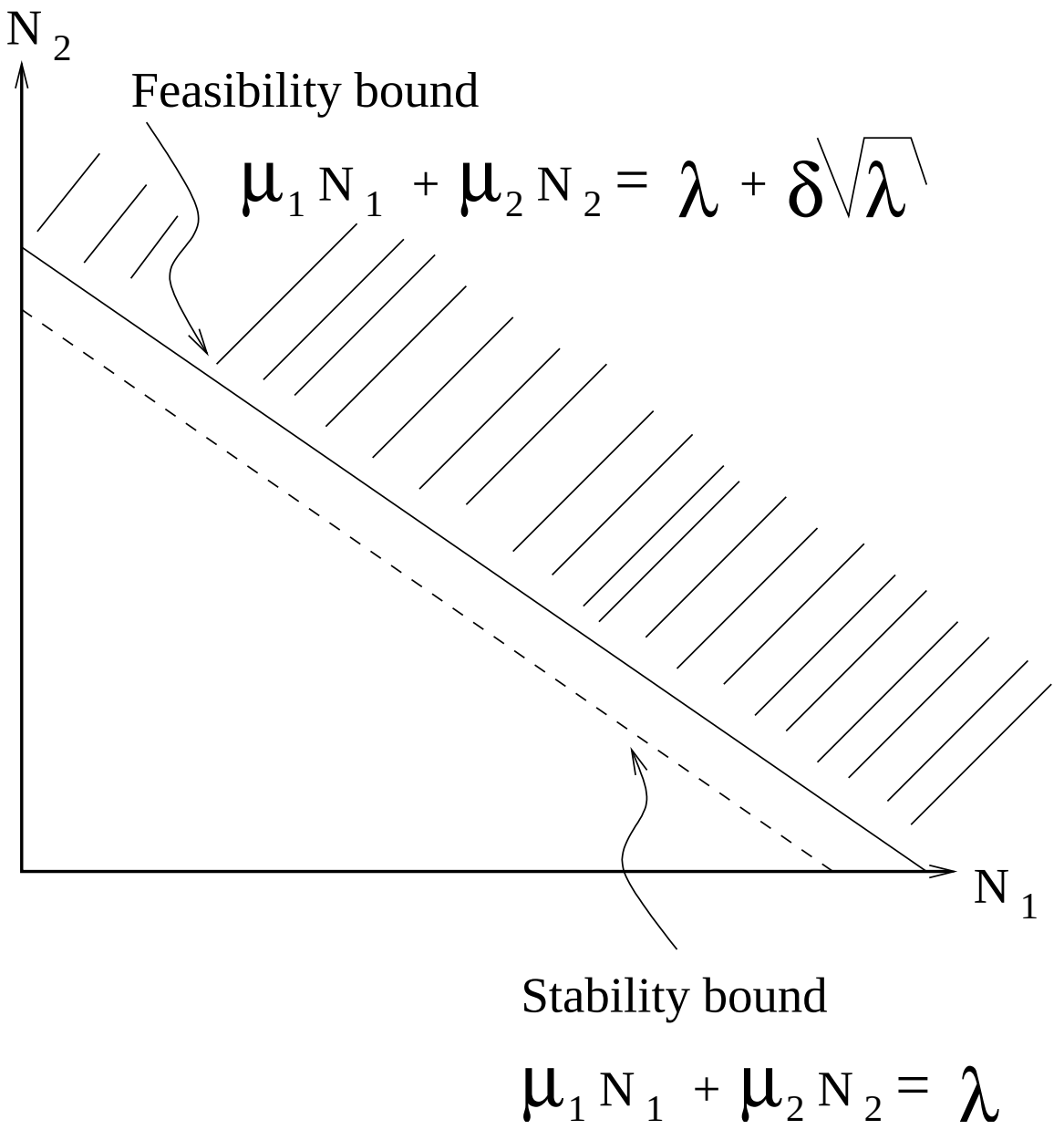
The Feasible Region



Problems:

- Must find optimal routing.
  - Threshold type solutions: Rykov (’01), Luh & Viniotis (’01) .
- Difficult to find **exact** feasible region.

The Feasible Region: QED Asymptotics



QED Feasibility: Theory

Proposition (Asymptotic Feasibility):
Consider a sequence of systems indexed by $\lambda \uparrow \infty$. Assume the number of slow servers is non-negligible: $\liminf_{\lambda\to\infty} N_1/N_2 > 0$. Then there exists a non-preemptive policy for which

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

if and only if

$$\mu_1 N_1 + \mu_2 N_2 \ge \lambda + \delta\sqrt{\lambda} + o\left(\sqrt{\lambda}\right), \quad 0 < \delta < \infty.$$

Here

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

is the Halfin-Whitt function $\alpha(\delta/\sqrt{\mu_1})$.

Corollary (Differentiated Service): The $\bigwedge-$design requires less capacity than the $I-$ design with average service rate.

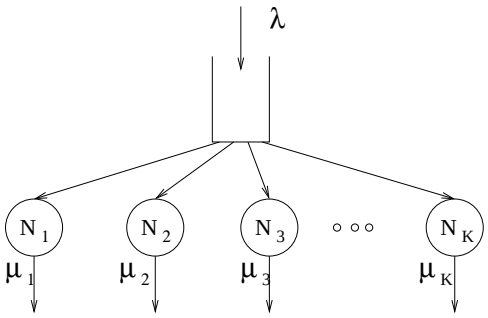
Proof: Recall $\mu_1 < \mu_2$. Let $\mu = \theta\mu_1 + (1-\theta)\mu_2$.
Then $P(\text{wait} > 0) \le \alpha$ iff

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

$\bigwedge$-design: $$\mu_1 N_1 + \mu_2 N_2 \ge \lambda + \beta(\alpha)\sqrt{\mu_1}\sqrt{\lambda} + o(\sqrt{\lambda}) .$$

The $\wedge-$Model: Exact Optimal Routing

(Rykov 2001, Luh & Viniotis 2002)



Problem: Find a non-preemptive non-anticipative routing policy that minimizes the average total number of customers in the system (or the average sojourn time).

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

1. It is the fastest idle server, and

2. the number of customers in queue is S_k or more.

Note: S_k may depend on the state of the other (slower) servers.

The $\wedge-$Model: QED Optimal Routing

Proposition (Optimal Preemptive Routing): The preemptive routing policy, Π^P, that always sends calls to the faster servers first is optimal in steady-state: it stochastically minimizes the total number of jobs in the system in steady-state.

Proof: Sample path coupling.

Note: Under Π^P, the total number of customers in the system determines how many servers of each type are working - thus, it is a one-dimensional Birth & Death process.

Corollary: Π^P stochastically minimizes the steady-state queue length and waiting time (since non-idling).

Proposition (Asymptotically Optimal Routing): The non-preemptive routing policy, Π^{NP}, that always sends incoming or waiting calls to the faster servers first is asymptotically optimal, with respect to queue length and waiting time in steady-state.

Proof: State-space collapse - in the limit, the fast servers are always busy.
$\Rightarrow$ The preemptive and non-preemptive policies are asymptotically equivalent.

Note: Thresholds are not needed above.

Asymptotic Feasibility

Proposition (Limiting Waiting Probability):

For both Π^P and Π^{NP}:

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

if and only if

$$\mu_1 N_1 + \mu_2 N_2 = \lambda + \delta\sqrt{\lambda} + o(\sqrt{\lambda}), \quad 0 \le \delta \le \infty,$$

where

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

provided that $\liminf_{\lambda\to\infty} N_1/N_2 > 0$.

Note: Choice of δ depends on α only through μ_1 - the service rate of the slowest servers.

Conclusion: The linear asymptotic feasible region.

QED Staffing: Optimality

Problem:

$$\mathbf{P}(\lambda,\alpha) \;=\; \text{Minimize} \qquad C_1 N_1^p + C_2 N_2^p, \quad p>1$$

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

$$N_1, N_2 \in \mathbb{Z}_+$$

Solution: Let $\vec{N}(\lambda,\alpha)$ be the optimal solution the auxiliary problem:

$$\mathbf{AP}(\lambda,\alpha) \;=\; \text{Minimize} \qquad C_1 N_1^p + C_2 N_2^p, \quad p>1$$

$$\text{Subject to} \qquad \mu_1 N_1 + \mu_2 N_2 \ge \lambda + \delta(\alpha)\sqrt{\lambda}$$

$$N_1, N_2 \ge 0$$

Claim: $\lceil \vec{N}(\lambda,\alpha) \rceil$ is an asymptotically optimal staffing sequence among all asymptotically feasible staffing sequences, as $\lambda \to \infty$.

Question: How to compare the costs of two staffing sequences? If

$$\vec{N} = \vec{N}(\lambda) = \lambda + o(\lambda) \;\text{ and }\; \vec{M} = \vec{M}(\lambda) = \lambda + o(\lambda),$$

then

$$\frac{C_1 N_1^p + C_2 N_2^p}{C_1 M_1^p + C_2 M_2^p} \to 1, \;\text{ as }\; \lambda \to \infty.$$

$\Rightarrow$ a finer comparison criterion is needed.

QED Optimal Staffing

Comparing Asymptotic Costs:

Let $\underline{C}(\lambda)$ be the optimal cost associated with the Stability Problem:

$$\begin{array}{lll} \underline{C}(\lambda) = & \text{Minimize} & C_1 N_1^p + C_2 N_2^p, \quad p > 1 \\ & \text{Subject to} & \mu_1 N_1 + \mu_2 N_2 \geq \lambda \\ & & N_1, N_2 \geq 0 \end{array}$$

Definition - Asymptotic Optimal Staffing: A sequence of staffing vectors $\vec{N} = \vec{N}(\lambda; \alpha)$ is said to be asymptotically optimal if:

1. It is asymptotically feasible, and
2. for every sequence $\vec{M} = \vec{M}(\lambda, \alpha)$ of staffing vectors which is also asymptotically feasible

$$\limsup_{\lambda \to \infty} \frac{C_1 N_1^p + C_2 N_2^p - \underline{C}(\lambda)}{C_1 M_1^p + C_2 M_2^p - \underline{C}(\lambda)} \leq 1.$$

Proposition (Asymptotically Optimal Staffing): Let $\vec{N}(\lambda; \alpha)$ be the optimal solution of the auxiliary problem $\mathbf{A}P(\lambda, \alpha)$. Then $\lceil \vec{N}(\lambda; \alpha) \rceil$ is an asymptotically optimal staffing, as $\lambda \to \infty$.

QED Optimal Staffing - Example

Consider the case $p = 2$, and the staffing problem:

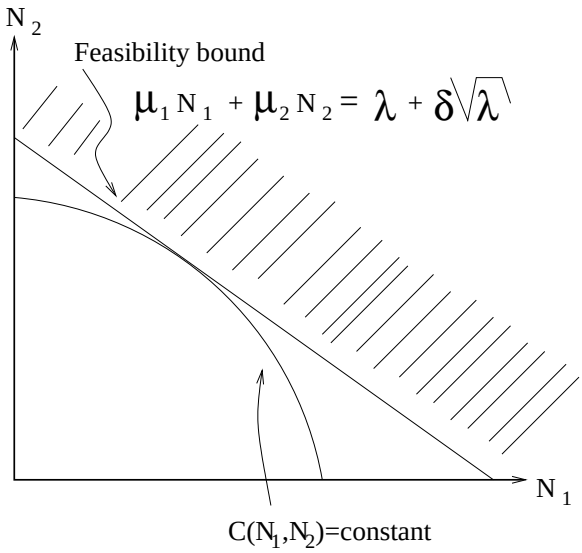
$$\begin{aligned} \mathbf{P}(\lambda, \alpha) \; = \quad & \text{Minimize} \qquad & C_1 N_1^2 + C_2 N_2^2, \\ & \text{Subject to} \qquad & P(\text{wait} > 0) \leq \alpha, \text{ for some routing policy} \\ & & N_1, N_2 \in \mathbb{Z}_+ \end{aligned}$$

Solution: **Total Capacity** (for feasibility) -

$$\mu_1 N_1 + \mu_2 N_2 = \lambda + \delta\sqrt{\lambda}, \quad \delta = \delta(\alpha, \mu_1).$$

Number of Servers in Each Pool (for optimality) -

$$\frac{N_1}{N_2} = \frac{C_2/\mu_2}{C_1/\mu_1}.$$



Transient Analysis

Goals:

- Prove equivalence between Π^P and Π^{NP} (state-space collapse).
- Characterize transient behavior of the multiple server type system in the QED regime, and compare to the $M/M/N$ system (Halfin & Whitt).

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

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

Proposition: Suppose that

1. $\lim_{\lambda\to\infty} \frac{\mu_i N_i}{\lambda} = a_i, \;\; i = 1,2, \;\; a_1 > 0, \; a_2 \geq 0, \; a_1 + a_2 = 1,$ and
2. $\lim_{\lambda\to\infty} \frac{\sum_{i=1}^2 \mu_i N_i - \lambda}{\sqrt{\lambda}} = \delta, \;\; \delta > 0.$

If $X^\lambda(0) \xrightarrow{d} X(0)$ then, under both Π^P and Π^{NP}, $X^\lambda \xrightarrow{d} X$, 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

$$\sigma^2(x) = 2\mu, \quad \mu = \left(\frac{a_1}{\mu_1} + \frac{a_2}{\mu_2}\right)^{-1}.$$

Conclusions and Further Research

Conclusions:

1. Square-root safety staffing is asymptotically optimal for both V- and $\wedge$-designs.

2. V-Model: Serving VIP customers first is asymptotically optimal (no thresholds needed for minimizing average waits, but they do arise with refined performance measures).

3. $\wedge$-Model: Routing to fast servers first is asymptotically optimal (no thresholds needed altogether, but could arise with server-related measures).

4. Asymptotic QED equivalence of non-preemptive and preemptive is fundamental (recent work by R. Atar).

Future Research:

1. Add features: Abandonment, Retrials (CRM);
   Customer-driven services: μ_j's.

2. Where are the thresholds?

3. Combine V-designs and $\wedge$-designs to study N-designs.