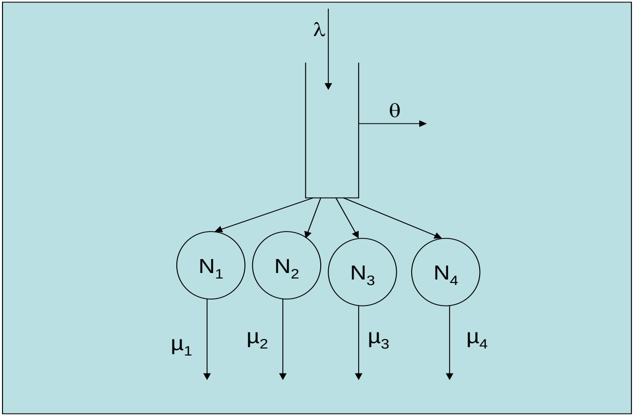
Routing and Staffing in Large-Scale Service Systems with Heterogeneous Servers and Impatient Customers

Mor Armony
New York University

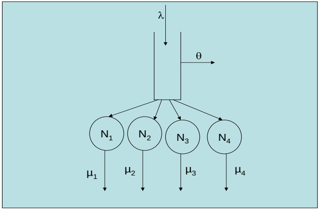
Joint Work with Avi Mandelbaum

Motivation: Call Centers

The Inverted-V Model with Abandonment



The Inverted-V Model with Abandonment: Motivation



- Heterogeneous server population
- Learning Effects
- Various server skills

The Model

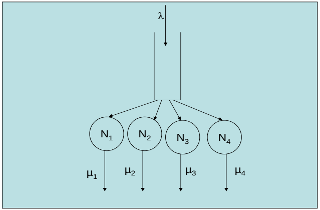
- Single customer class - Poisson arrival process with rate λ.
- K server pools ($N_1, N_2, ..., N_K$ servers)
- Exponential non-preemptive service times with rates

  $\mu_1 < \mu_2 < ... < \mu_K$.
- Exponential time to abandonment with rate θ.

Our Focus: Staffing and Routing

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

Inverted-V model Without Abandonment (Armony '05)



Minimize $\quad C_1(N_1) + C_2(N_2) + ... + C_K(N_K)$

Subject to $\quad P(W > 0) \leq \alpha,$ for some routing policy

Why Consider Abandonment?

Even little abandonment can have a significant effect on performance:

Why Consider Abandonment?

Even little abandonment can have a significant effect on performance:

- An unstable $M/M/N$ system ($\rho > 1$) becomes stable with abandonment ($M/M/N+M$).

Why Consider Abandonment?

Even little abandonment can have a significant effect on performance:

- An unstable $M/M/N$ system ($\rho > 1$) becomes stable with abandonment ($M/M/N + M$).
- Example: (Mandelbaum & Zeltyn '06) Consider $\lambda = 2000$, $\mu = 20$. Service level target: "80% of customers should be served within 30 second".

Why Consider Abandonment?

Even little abandonment can have a significant effect on performance:

- An unstable $M/M/N$ system ($\rho > 1$) becomes stable with abandonment ($M/M/N + M$).
- Example: (Mandelbaum & Zeltyn '06) Consider $\lambda = 2000$, $\mu = 20$. Service level target: "80% of customers should be served within 30 second".
  - 106 agents ($\theta = 0$).

Why Consider Abandonment?

Even little abandonment can have a significant effect on performance:

- An unstable $M/M/N$ system ($\rho > 1$) becomes stable with abandonment ($M/M/N + M$).
- Example: (Mandelbaum & Zeltyn '06) Consider $\lambda = 2000$, $\mu = 20$. Service level target: "80% of customers should be served within 30 second".
  - 106 agents ($\theta = 0$).
  - 95 agents ($\theta = 20$ (avg. patience of 3 minutes), $P(ab) = 6.9\%$)

Why Consider Abandonment?

Even little abandonment can have a significant effect on performance:

- An unstable $M/M/N$ system ($\rho > 1$) becomes stable with abandonment ($M/M/N+M$).
- Example: (Mandelbaum & Zeltyn '06) Consider $\lambda = 2000$, $\mu = 20$. Service level target: "80% of customers should be served within 30 second".
  - 106 agents ($\theta = 0$).
  - 95 agents ($\theta = 20$ (avg. patience of 3 minutes), $P(ab) = 6.9\%$)
  - 84 agents ($\theta = 60$ (avg. patience of 1 minute), $P(ab) = 16.8\%$)

Problem Formulation

$$\begin{aligned} \text{Minimize} \quad & C_1(N_1) + C_2(N_2) + \ldots + C_K(N_K) \\ \text{Subject to} \quad & P(W > T) \leq \alpha_T, \quad \text{for some routing policy} \\ & EW \leq \bar{W}, \\ & P(ab) \leq \Delta. \end{aligned}$$

Issues related to formulation:

Problem Formulation

$$
\begin{aligned}
\text{Minimize} \quad & C_1(N_1) + C_2(N_2) + \ldots + C_K(N_K) \\
\text{Subject to} \quad & P(W > T) \le \alpha_T, \quad \text{for some routing policy} \\
& EW \le \bar{W}, \\
& P(ab) \le \Delta.
\end{aligned}
$$

Issues related to formulation:

- FCFS: Natural but not necessarily optimal.

Problem Formulation

$$
\begin{array}{ll}
\text{Minimize} & C_1(N_1) + C_2(N_2) + \ldots + C_K(N_K) \\
\text{Subject to} & P(W > T) \leq \alpha_T, \quad \text{for some routing policy} \\
 & EW \leq \bar{W}, \\
 & P(ab) \leq \Delta.
\end{array}
$$

Issues related to formulation:

- FCFS: Natural but not necessarily optimal.
- Intensional Idling can improve service level.

Problem Formulation

$$\begin{aligned}
\text{Minimize} \quad & C_1(N_1) + C_2(N_2) + \ldots + C_K(N_K) \\
\text{Subject to} \quad & P(W > T) \leq \alpha_T, \quad \text{for some routing policy} \\
& EW \leq \bar{W}, \\
& P(ab) \leq \Delta.
\end{aligned}$$

Issues related to formulation:

- FCFS: Natural but not necessarily optimal.
- Intensional Idling can improve service level.
- Not all are Customer-subjective measurements.

Background: Garnett, Mandelbaum & Reiman '02

In a sequence of $M/M/N+M$ models, $N = 1, 2, 3, ...$, with $R = \lambda/\mu$, the following are equivalent:

- $N \approx R + \beta\sqrt{R}, \quad -\infty < \beta < \infty$;

Background: Garnett, Mandelbaum & Reiman '02

In a sequence of $M/M/N+M$ models, $N = 1, 2, 3, \ldots$, with $R = \lambda/\mu$, the following are equivalent:

- $N \approx R + \beta\sqrt{R}, \quad -\infty < \beta < \infty;$
- $\lim_{N\to\infty} P_N\{W > 0\} = \alpha, \quad 0 < \alpha < 1;$

Background: Garnett, Mandelbaum & Reiman '02

In a sequence of $M/M/N+M$ models, $N = 1, 2, 3, \ldots$, with $R = \lambda/\mu$, the following are equivalent:

- $N \approx R + \beta\sqrt{R}, \quad -\infty < \beta < \infty;$
- $\lim_{N\to\infty} P_N\{W > 0\} = \alpha, \quad 0 < \alpha < 1;$
- $\lim_{N\to\infty} \sqrt{N} P_N\{ab\} = \Delta, \quad 0 < \Delta < \infty;$

Here $\alpha = w(-\beta, \sqrt{\mu/\theta}), \quad \Delta = \left[\sqrt{\theta/\mu} \cdot h(\beta\sqrt{\mu/\theta}) - \beta\right]\alpha,$

$w(x,y) = \left[1 + \frac{h(-xy)}{yh(x)}\right]^{-1}, \quad h(x) = \frac{\phi(x)}{1-\Phi(x)}.$

Background: Mandelbaum & Zeltyn '06

Staffing $M/M/N+G$ systems: Minimize N subject to performance constraints. Three Operational Regimes:

Background: Mandelbaum & Zeltyn '06

Staffing $M/M/N+G$ systems: Minimize N subject to performance constraints. Three Operational Regimes:

1. QED regime: $N = R + \beta\sqrt{R}$
   - $\sqrt{\lambda}P\{ab\} \leq \Delta$, $\sqrt{\lambda}EW \leq \bar{w}$

Background: Mandelbaum & Zeltyn '06

Staffing $M/M/N+G$ systems: Minimize N subject to performance constraints. Three Operational Regimes:

1. QED regime: $N = R + \beta\sqrt{R}$
   - $\sqrt{\lambda}P\{ab\} \leq \Delta$, $\sqrt{\lambda}EW \leq \bar{w}$
2. ED regime: $N = (1-\gamma)\cdot R, \quad \gamma > 0$
   - $P\{ab\} \leq \Delta$, $EW \leq \bar{w}$

Background: Mandelbaum & Zeltyn '06

Staffing $M/M/N+G$ systems: Minimize N subject to performance constraints. Three Operational Regimes:

1. QED regime: $N = R + \beta\sqrt{R}$
   - $\sqrt{\lambda}P\{ab\} \le \Delta$, $\sqrt{\lambda}EW \le \bar{w}$
2. ED regime: $N = (1-\gamma)\cdot R, \quad \gamma > 0$
   - $P\{ab\} \le \Delta$, $EW \le \bar{w}$
3. ED + QED regime: $N = (1-\gamma)R + \delta\sqrt{R}, \quad \gamma > 0.$
   - $P\{W > T\} \le \alpha.$

Staffing of the Inverted-V model in the QED Regime

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

$$\text{Subject to} \quad \sqrt{\lambda}P(ab) \leq \Delta, \quad \text{for some routing policy.}$$

Challenges:

Proposed Staffing: Square-Root Safety-Staffing

Staffing of the Inverted-V model in the QED Regime

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

$$\text{Subject to} \quad \sqrt{\lambda} P(ab) \leq \Delta, \quad \text{for some routing policy.}$$

Challenges:

- Offered load not well-defined

Proposed Staffing: Square-Root Safety-Staffing

Staffing of the Inverted-V model in the QED Regime

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

$$\text{Subject to} \quad \sqrt{\lambda}P(ab) \leq \Delta, \quad \text{for some routing policy.}$$

Challenges:

- Offered load not well-defined
- Optimal routing unknown

Proposed Staffing: Square-Root Safety-Staffing

Staffing of the Inverted-V model in the QED Regime

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

$$\text{Subject to} \quad \sqrt{\lambda} P(ab) \leq \Delta, \quad \text{for some routing policy.}$$

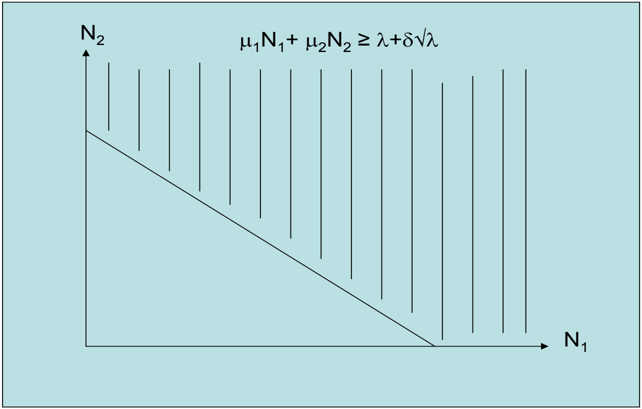
Challenges:

- Offered load not well-defined
- Optimal routing unknown

Proposed Staffing: Square-Root Safety-Staffing

$$\mu_1 N_1 + \mu_2 N_2 + \ldots + \mu_K N_K = \lambda + \delta\sqrt{\lambda}, \quad -\infty < \delta < \infty$$

Asymptotic Feasible Region



The Asymptotic Feasible Region

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

$$\limsup_{\lambda \to \infty} \sqrt{\lambda} P_\lambda(ab) \leq \Delta,$$

if and only if

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

Routing

Exact Optimal Routing: Unknown (De Vericourt and Zhou '06)

Routing

Exact Optimal Routing: Unknown (De Vericourt and Zhou '06)

Proposed Routing: Route to Faster Servers First (FSF).

Routing

Exact Optimal Routing: Unknown (De Vericourt and Zhou '06)

Proposed Routing: Route to Faster Servers First (FSF).

Potential Problem: Preemption may lead to excessive idling of fast servers.

Routing

Exact Optimal Routing: Unknown (De Vericourt and Zhou '06)

Proposed Routing: Route to Faster Servers First (FSF).

Potential Problem: Preemption may lead to excessive idling of fast servers.

Proposition (Optimal Preemptive Routing): FSF_P is optimal in the sense that it stochastically minimizes the cumulative number of abandoning customers.

Routing

Exact Optimal Routing: Unknown (De Vericourt and Zhou '06)

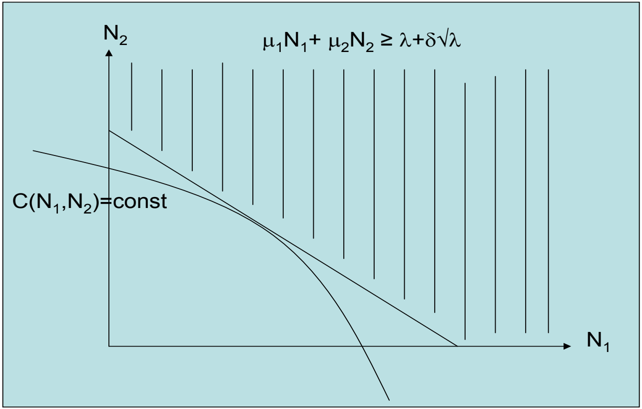
Proposed Routing: Route to Faster Servers First (FSF).

Potential Problem: Preemption may lead to excessive idling of fast servers.

Proposition (Optimal Preemptive Routing): FSF_P is optimal in the sense that it stochastically minimizes the cumulative number of abandoning customers.

Proposition (Asymptotically Optimal Routing): FSF is asymptotically optimal in the sense that in the limit FSF and FSF_P have the same performance. (**Proof:** State-space collapse - Faster servers are always busy).

Asymptotically Optimal Staffing



Asymptotically Optimal Staffing: Example

Problem:

$$\text{Minimize} \quad C_1 N_1^p + C_2 N_2^p + \ldots + C_K N_K^p, \quad p > 1$$
$$\text{Subject to} \quad \sqrt{\lambda} P\{ab\} \leq \Delta.$$

Solution:

$$\text{Minimize} \quad C_1 N_1^p + C_2 N_2^p + \ldots + C_K N_K^p, \quad p > 1$$
$$\text{Subject to} \quad \mu_1 N_1 + \mu_2 N_2 + \ldots \mu_K N_K \geq \lambda + \delta\sqrt{\lambda}.$$

To get: $\frac{N_k}{N_j} = \left(\frac{\mu_k / C_k}{\mu_j / C_j}\right)$. (Note: $N_1/N > 0$!!!)

Outperforming the Homogeneous Server System

Consider an $M/M/N+M$ system with $\mu = \sum_{k=1}^{K} q_k \mu_k$. Then $\sqrt{\lambda} P\{ab\} \leq \Delta$ if and only if:

$$\mu N \geq \lambda + \beta \sqrt{\mu} \sqrt{\lambda}.$$

Compared to:

$$\mu_1 N_1 + \mu_2 N_2 + ... + \mu_K N_K \geq \lambda + \beta \sqrt{\mu_1} \sqrt{\lambda}.$$

Conclusions

Conclusions

- Simple Solution to a Difficult Problem:

Conclusions

- Simple Solution to a Difficult Problem:
    - Separation of Staffing and Control.

Conclusions

- Simple Solution to a Difficult Problem:
  - Separation of Staffing and Control.
  - Square-root Safety Staffing is asymptotically optimal.

Conclusions

- Simple Solution to a Difficult Problem:
    - Separation of Staffing and Control.
    - Square-root Safety Staffing is asymptotically optimal.
    - Linear boundary for feasible region.

Conclusions

- Simple Solution to a Difficult Problem:
  - Separation of Staffing and Control.
  - Square-root Safety Staffing is asymptotically optimal.
  - Linear boundary for feasible region.
  - FSF is asymptotically optimal (although not exactly optimal)

Conclusions

- Simple Solution to a Difficult Problem:
  - Separation of Staffing and Control.
  - Square-root Safety Staffing is asymptotically optimal.
  - Linear boundary for feasible region.
  - FSF is asymptotically optimal (although not exactly optimal)
    - Preemptive and non-preemptive policies are asymptotically equivalent

      (General result: Atar '03)

Conclusions

- Simple Solution to a Difficult Problem:
  - Separation of Staffing and Control.
  - Square-root Safety Staffing is asymptotically optimal.
  - Linear boundary for feasible region.
  - FSF is asymptotically optimal (although not exactly optimal)
    - Preemptive and non-preemptive policies are asymptotically equivalent (General result: Atar '03)
  - Asymptotic cost minimization made simple - easy to include other constraints.

Conclusions

- Simple Solution to a Difficult Problem:
  - Separation of Staffing and Control.
  - Square-root Safety Staffing is asymptotically optimal.
  - Linear boundary for feasible region.
  - FSF is asymptotically optimal (although not exactly optimal)
    - Preemptive and non-preemptive policies are asymptotically equivalent (General result: Atar '03)
  - Asymptotic cost minimization made simple - easy to include other constraints.
- The inverted-V system outperforms its homogeneous server