Compensation and Staffing and to Trade Off Speed and Quality in Large Service Systems

Dongyuan Zhan (UCL School of Management)
Amy R. Ward (USC Marshall School of Business)

13 November 2015

UCL
SCHOOL OF
MANAGEMENT

Slowdown?

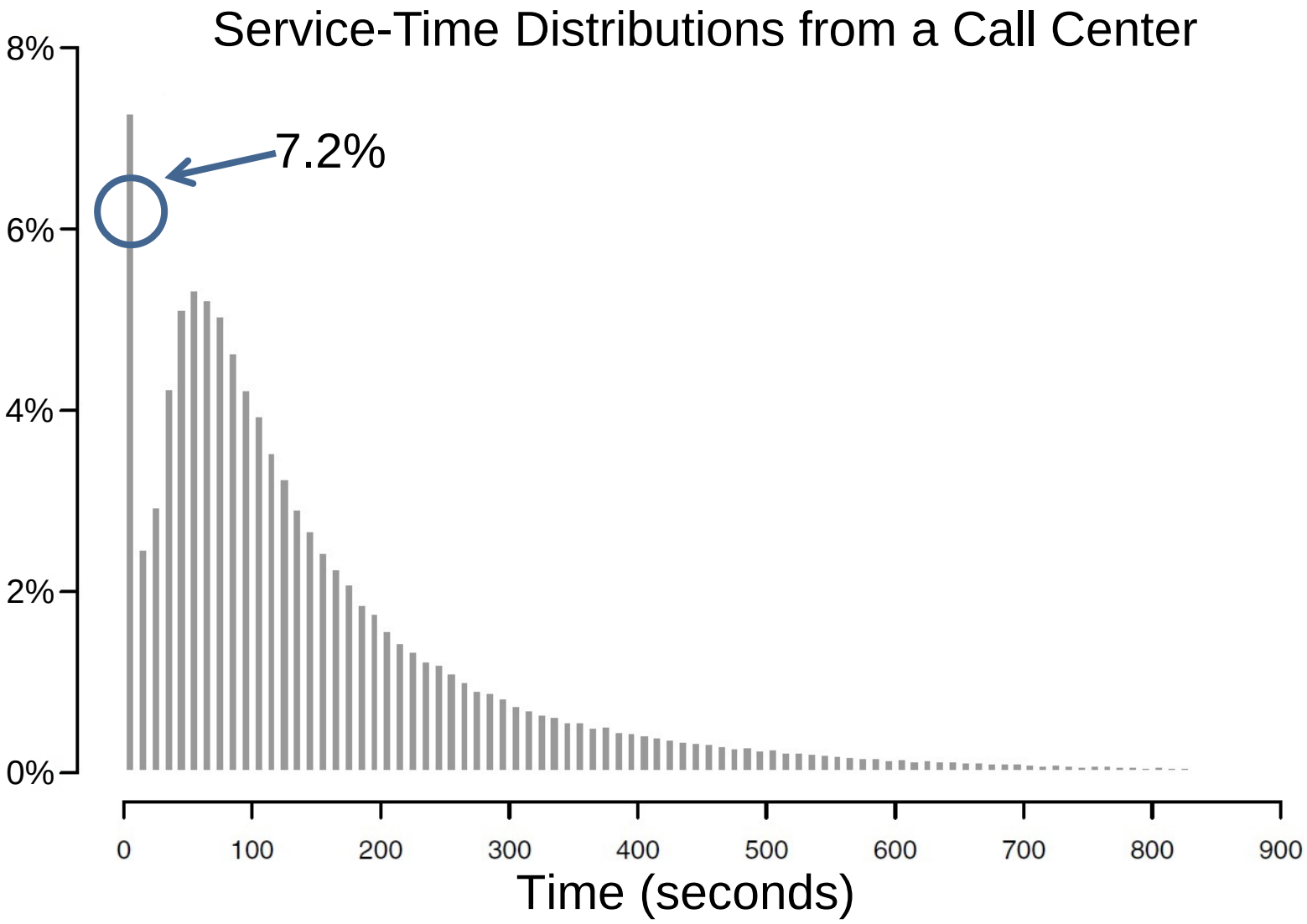
Imagine:

How do you respond?

You may respond strategically.

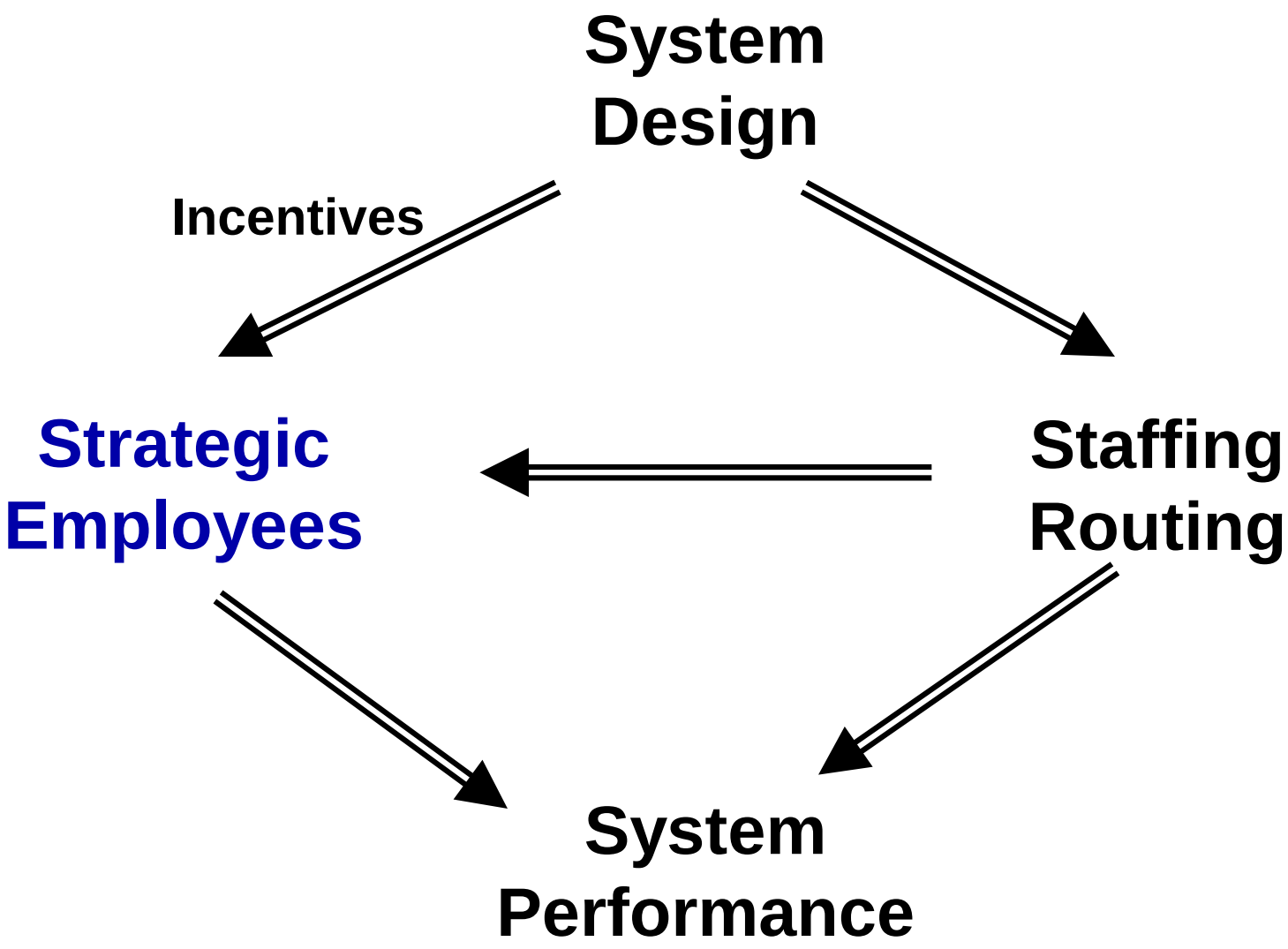
What if you were paid per review?

Employee Strategic Behavior



From Gans, Koole, Mandelbaum (2003)

The Big Picture



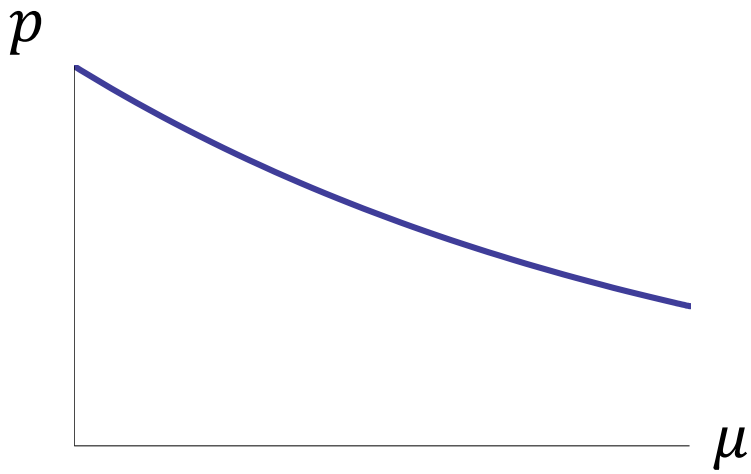
Outline

- The Employee Payment Model
- Employee Behavior: the Equilibrium Service Rate
- Optimal Design: Compensation and Staffing
- Generalizations

The Speed Quality Trade-off

Lower speed ⟹ Higher quality

$p(\mu)$: The probability of successful service.



Discretionary task completion in Hopp, Iravani, Yuen, 2007

Customer-intensive service in Anand, Pac, Veeraraghavan, 2011.

Compensation and Research Question

Each risk neutral employee

- is paid P_S per service completion.
- is penalized P_F per failed service.
- maximizes her expected payment.

The promised expected payment is c_S per unit time.

What is employees' behavior?

What is the optimal system design?

OM Literature Review

Queueing games

- Hassin and Haviv (2003)

Service rate decisions when service providers compete

- Kalai, Kamien, Rubinovitch (1992), Gilbert and Weng (1998), Cachon and Harker (2002) , Cachon and Zhang (2007)

Speed and quality trade-off

- Hopp, Iravani, Yuen(2007), Ren and Zhou (2008), Lu, Van Mieghem, Savaskan (2009), Anand, Pac, Veeraraghavan (2011), Mehrotra, Ross, Ryder, Zhou (2012), Chan, Yom-Tov, Escobar(2014), Zhan and Ward (2014)

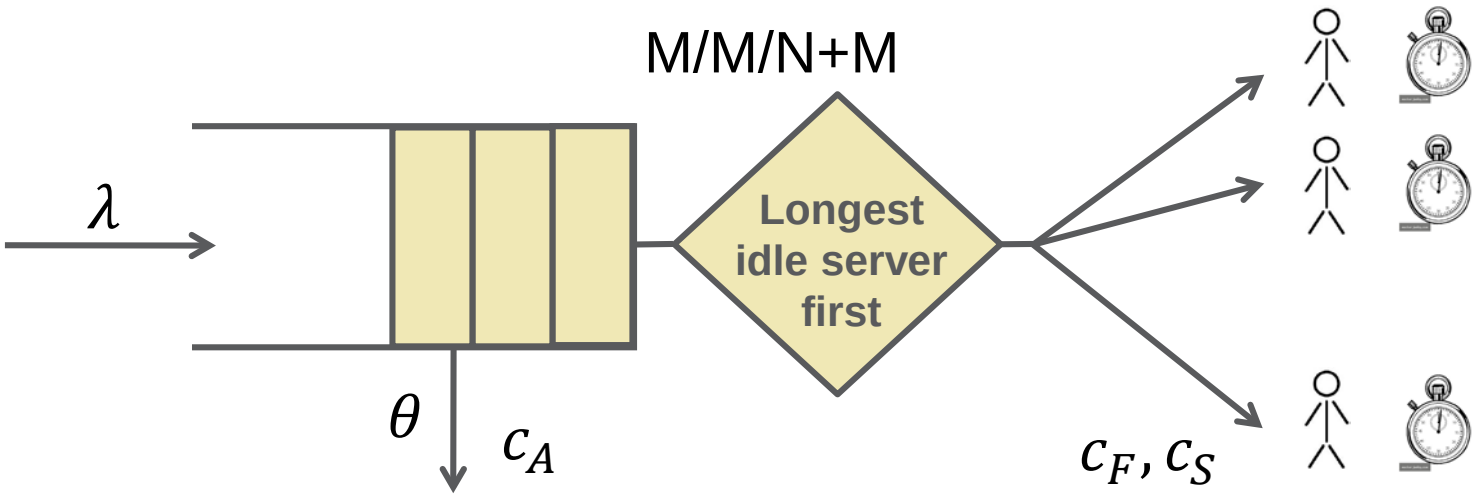
Large-scale service systems

- Haffin and Whitt (1981), Garnett, Mandelbaum, Reiman (2002), Gurvich and Whitt (2009), Tezcan and Dai (2010), Allon and Gurvich (2010), Mandelbaum and Armony (2011)

Strategic employees within an organization

- Buell, Kim, Tsay (2014), Song, Tucker, Murrell (2014), Shunko, Niederhoff, Rosokha (2014), Gopalakrishnan, Douroudi, Ward, Wierman (2014)

The Model



Given reward P_S, penalty P_F and staffing N, what service rate will the employees choose?

- Does there exist a symmetric Nash equilibrium?
- The utility is the expected payment (and later generalize).

How to choose P_S, P_F and N optimally?

- This depends on the cost structure the system manager faces.
- The costs are linear (and later generalize).

Outline

- The Employee Payment Model
- **Employee Behavior: The Equilibrium Service Rate**
- Optimal Design: Compensation and Staffing
- Generalizations

Find the Service Rate When N=1

The server chooses her service rate to maximize her expected payment (utility):

$$U(\mu) = (P_S - P_F\big(1 - p(\mu)\big)\mu\, B(\mu)$$

Reward | Penalty | Busy time portion

$$B(\mu) = 1 - \frac{1}{1 + \sum_{i=1}^{\infty} \prod_{k=0}^{i-1} \frac{\lambda}{\mu + k\theta}}$$

Proposition:

Suppose $p(\mu)\mu$ is concave in μ, given P_S and P_F, the server chooses a **unique** service rate $\boldsymbol{\mu_E}$ to maximize the expected payment.

$\boldsymbol{\mu_E}$ is continuous and decreasing in P_F/P_S.

Find the Equilibrium Service Rate

When $N > 1$, the complication is that each server i has utility that depends on the service rates chosen by the other servers.

$$U_i(\vec{\mu}) = (P_S - P_F\big(1 - p(\mu_i)\big)\mu_i B_i(\vec{\mu})$$

What is the symmetric equilibrium service rate $\boldsymbol{\mu_E}$?

Server 1 busy, no queue — Server 1 busy, queue

$$B(\mu_1, \mu) = \frac{\sum_{i=0}^{N-1}\left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!} + \left(\frac{\lambda}{\mu}\right)^i \frac{1}{(N-1)!}\sum_{i=1}^{\infty}\prod_{k=1}^{i}\frac{\lambda}{(N-1)\mu + \mu_1 + k\theta}}{\sum_{i=0}^{N-1}\left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!} + \left(\frac{\lambda}{\mu}\right)^i \frac{1}{(N-1)!}\sum_{i=1}^{\infty}\prod_{k=1}^{i}\frac{\lambda}{(N-1)\mu + \mu_1 + k\theta} + \frac{\mu_1}{\lambda}\sum_{i=0}^{N-1}\left(\frac{\lambda}{\mu}\right)^i \frac{N-i}{i!}}$$

Server 1 idle, no queue

Wishful thinking: If $B_i(\vec{\mu}) = 1$, then

$$\mu_E = \mu^* := \text{argmax}_\mu(P_S - P_F\big(1 - p(\mu)\big)\mu$$

"No competition equilibrium."

Determined by $\frac{P_F}{P_S}$.

Busy Time Approximate

Staffing: $N^{\lambda} = b\lambda + o(\lambda), b > 0$

Note if $b = \frac{1}{\mu_E}$, then the first order term is identical to the square root staffing.

Define $\hat{B}(\mu_1, \mu) = \frac{\mu}{\mu + \mu_1[b\mu - 1]^+}$

Lemma:

As $\lambda \to \infty$,

$$\sup_{\mu_1, \mu \in [\underline{\mu}, \overline{\mu}]} \left| B(\mu_1, \mu) - \hat{B}(\mu_1, \mu) \right| \to 0$$

Therefore, define $\hat{U}(\mu_1, \mu) = (P_S - P_F(1 - p(\mu_1))\mu_1 \hat{B}(\mu_1, \mu)$, we have

$$U(\mu_1, \mu) \to \hat{U}(\mu_1, \mu) \text{ as } \lambda \to \infty.$$

The Approximate Equilibrium Service Rate

For the revised problem

$$\widehat{U}(\hat{\mu}_E, \hat{\mu}_E) = \max_{\mu \in [\underline{\mu}, \overline{\mu}]} \widehat{U}(\mu, \hat{\mu}_E)$$

There exists a unique $\hat{\mu}_E \in (\frac{\lambda}{N}, \mu^*]$.

Proposition: Suppose $p(\mu)\mu$ is concave in μ, for systems with large enough λ, there exists a symmetric equilibrium service rate $\mu_E^{\lambda}(N^{\lambda}, P_S, P_F)$. As $\lambda \to \infty$,

if $b \leq \frac{1}{\mu^*}$, $\mu_E^{\lambda}(N^{\lambda}, P_S, P_F) \to \hat{\mu}_E = \mu^*$;

if $b > \frac{1}{\mu^*}$, $\mu_E^{\lambda}(N^{\lambda}, P_S, P_F) \to \hat{\mu}_E \in (\frac{1}{b}, \mu^*)$.

Outline

- The Employee Payment Model
- Employee Behavior: the Equilibrium Service Rate
- **Optimal Design: Compensation and Staffing**
- Generalizations

The System Cost

$$C(N, P_S, P_F) = \sum_{i=1}^{N} U_i(\mu_E) + c_F\big(1 - p(\mu_E)\big)\mu_E NB(\mu_E) + c_A(\lambda - \mu NB(\mu_E))$$

Controls | Staffing cost | Failed services cost | Abandonment cost

μ_E is determined by $N, \frac{P_F}{P_S}$, and is well approximated by $\hat{\mu}_E$.

Individual rationality constraint is: $U_i(\mu_E) \geqq c_S$.

$$C(N, P_S, P_F) = c_S N + c_F\big(1 - p(\mu_E)\big)\mu_E NB(\mu_E) + c_A(\lambda - \mu NB(\mu_E))$$

$$\mathrm{Min}_{N,P_S,P_F}\ C(N, P_S, P_F)$$

The First Best Control

In a **decentralized** system:

$$C(N, P_S, P_F) = c_S N + c_F\big(1 - p(\mu_E)\big)\mu_E NB(\mu_E) + c_A(\lambda - \mu NB(\mu_E))$$

In a **centralized** system:

$$C(N, \mu) = c_S N + c_F\big(1 - p(\mu)\big)\mu NB(\mu) + c_A(\lambda - \mu NB(\mu))$$

A **first best control** achieves:

$$\mathrm{Min}_{N,P_S,P_F}\, C(N, P_S, P_F) = \mathrm{Min}_{N,\mu}\, C(N, \mu)$$

First Best Compensation for M/M/1+M

No decision on staffing.

First best compensation:

$$P_S = c_A\Delta,\ P_F = c_F\Delta,$$

where Δ is set so that IR constraint binds.

Each service completion avoids the abandonment cost c_A.

Each service failure incurs the cost c_F.

Can this First Best Compensation Be Extended?

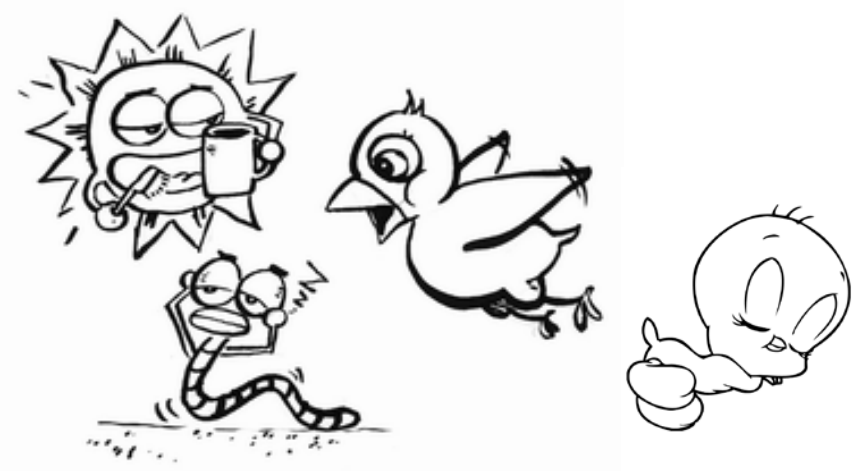
In an M/M/N+M system, is it the first best if the manager set

$$P_S = c_A\Delta,\ P_F = c_F\Delta\ ?$$

NO!

The employees have to compete for jobs.

The Early Bird Gets the Worm!



Competition ⇒ Speed up!

A **larger** $\frac{P_F}{P_S}$ is needed…

Cost of Serving One Customer

Each customer requires $\frac{1}{\mu}$ time in average, cost of salary is $\frac{c_S}{\mu}$.

The expected service failure cost is $c_F\big(1 - p(\mu)\big)$.

Define $\hat{\mu} \coloneqq \operatorname{argmin}_{\mu} \ \frac{c_S}{\mu} + c_F\big(1 - p(\mu)\big)$.

The manager wants the employees to work at $\hat{\mu}$.
The minimal cost of serving a customer is

$$\hat{c} \coloneqq \frac{c_S}{\hat{\mu}} + c_F\big(1 - p(\hat{\mu})\big).$$

Proposition (asymptotic optimal):

If $\hat{c} < c_A$, let the employees work at $\hat{\mu}$ and staff $N^{\lambda} = \left\lceil \frac{\lambda}{\hat{\mu}} \right\rceil$ in centralized systems. The policy is asymptotic optimal

$$\lim_{\lambda \to \infty} \frac{C\big(N^{\lambda}, \hat{\mu}\big)}{\lambda} = \hat{c}.$$

The Optimal Staffing and Compensation

Proposed policy: $N^{\lambda} = \left\lceil \frac{\lambda}{\hat{\mu}} \right\rceil$, $P_S = \hat{c}$, $P_F = c_F$.

Theorem (asymptotic first best)

If $\hat{c} < c_A$, the policy is asymptotic first best:

$$\lim_{\lambda \to \infty} \frac{C\left(N^{\lambda}, P_S, P_F\right)}{\lambda} = \hat{c}.$$

The policy incentivizes the servers to work at $\mu^*(P_S, P_F) = \hat{\mu}$.

The system is critically loaded: $\lambda \approx N^{\lambda}\hat{\mu}$, $B(\mu^*) \approx 1$.

Since $\hat{c} < c_A$, we have $\frac{P_F}{P_S} = \frac{c_F}{\hat{c}} > \frac{c_F}{c_A}$.

M/M/1+M: $\frac{P_F}{P_S} = \frac{c_F}{c_A}$.

A larger $\frac{P_F}{P_S}$ to counteract speed-up effect!

Outline

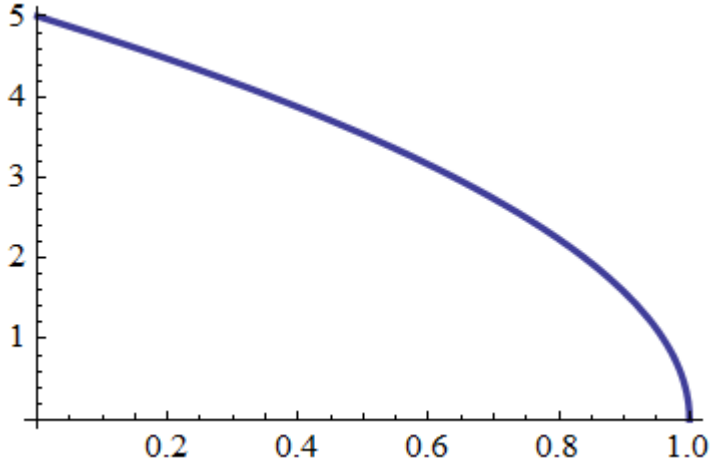
- The Employee Payment Model
- Employee Behavior: the Equilibrium Service Rate
- Optimal Design: Compensation and Staffing
- **Generalizations**

Extension on Utility and Cost Functions

$$U_i(\vec{\mu}) = (P_S - P_F\big(1 - p(\mu_i)\big)\mu_i B_i(\vec{\mu}) + U_I(B_i(\vec{\mu}))$$

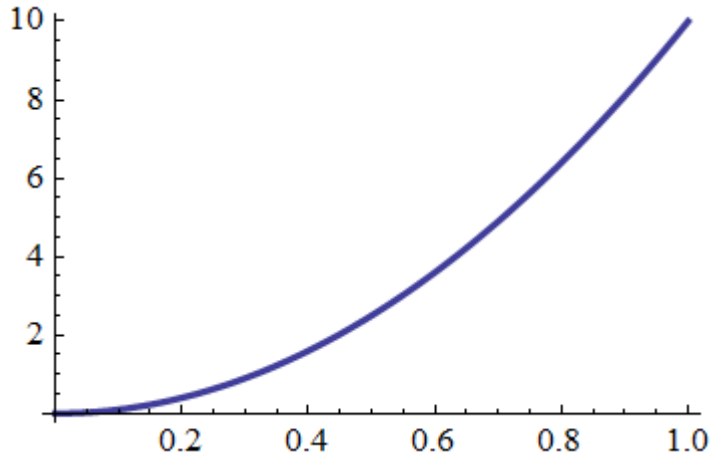
Monetary Non-monetary

Idleness value U_I



Busy time portion

Unit service failure / Abandonment cost



Service failure / Abandonment portion

An Approximate Symmetric Equilibrium

Replace B by $\hat{B}$ in the utility, so that

$$\hat{U}(\mu_1, \mu) = (P_S - P_F\big(1 - p(\mu_1)\big)\hat{B}(\mu_1, \mu) + U_I\left(\hat{B}(\mu_1, \mu)\right).$$

Find $\hat{\mu}_E$ such that

$$\hat{U}(\hat{\mu}_E, \hat{\mu}_E) = \max_{\mu \in [\underline{\mu}, \bar{\mu}]} \hat{U}(\mu, \hat{\mu}_E)$$

$\hat{\mu}_E$ is approximate symmetric equilibrium.

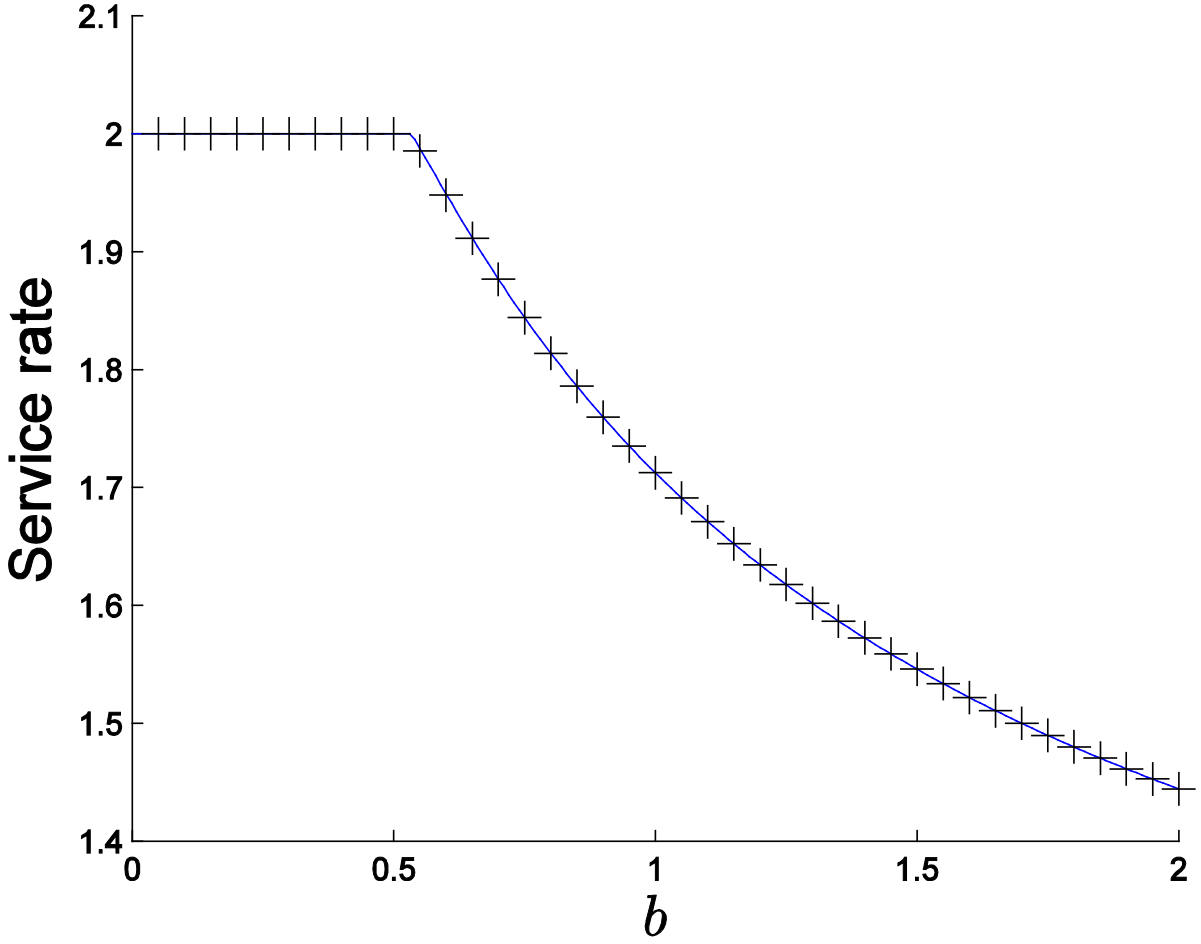
Proposition:

For any fixed $P_S > 0, P_F \geq 0$, the revised problem has a solution.

- If $b\mu^* \leq 1$, then $\hat{\mu}_E = \mu^*$ is one solution;
- Otherwise, any solution has $b\hat{\mu}_E > 1$.

Uniqueness?

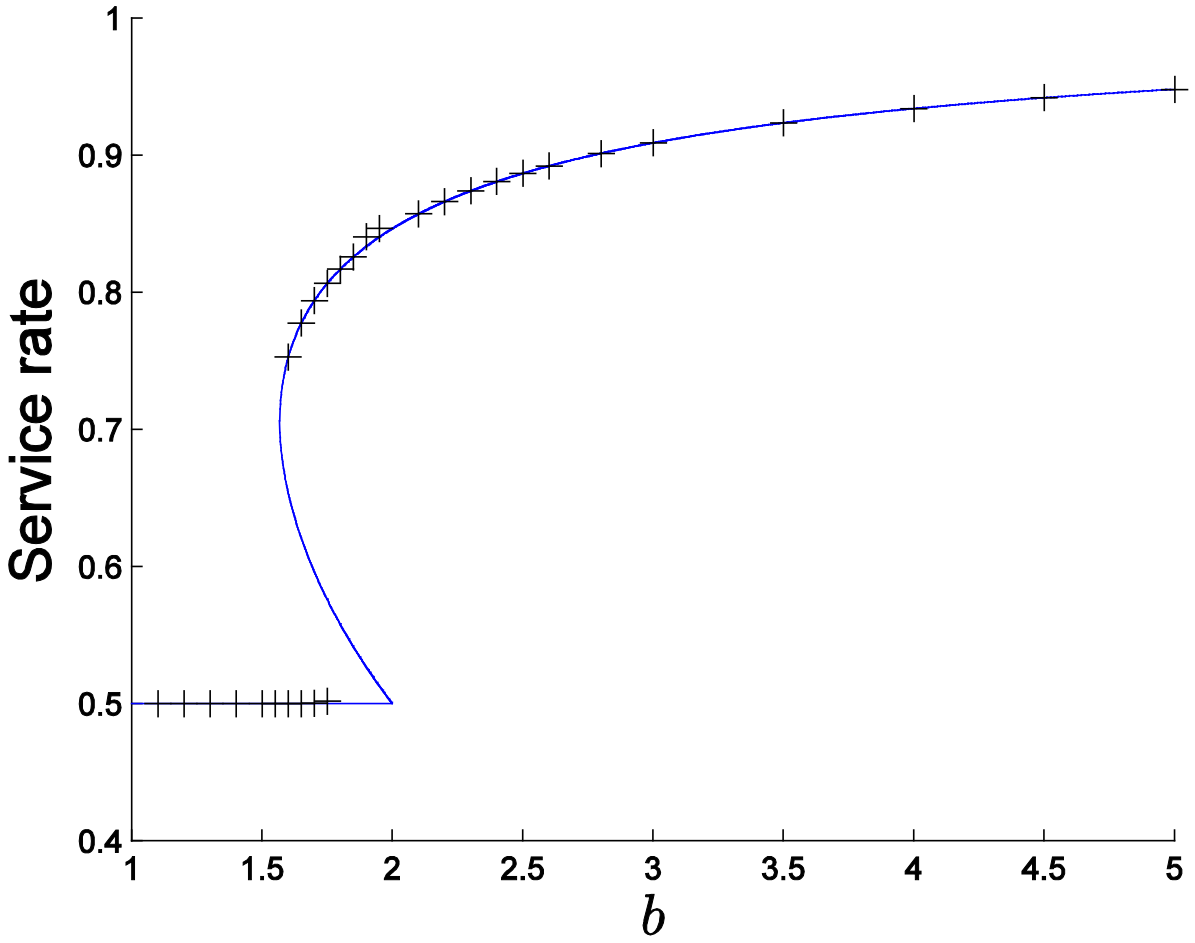
Comparison of μ_E^{λ}(cross) and $\widehat{\mu}_E$(line)



$\lambda = 1000, N^{\lambda} = b\lambda, \theta = 0.1, P_S = 20, P_F = 10;$

$p(\mu) = 1 - 0.5\mu, \mu \in [0,2]$, and $U_I(x) = 10\sqrt{1-x}$.

Comparison of μ_E^{λ}(cross) and $\widehat{\mu}_E$(line)



$\lambda = 1000, N^{\lambda} = b\lambda, \theta = 0.1, P_S = 5, P_F = 10;$

$p(\mu) = 1 - 0.5\mu, \mu \in [0,2]$, and $U_I(x) = 10\sqrt{1-x}$.

A Panorama of Optimal Limit Regimes

Utility and cost structure	Optimal Regime	Property
Linear Costs	Critically loaded	No idleness, no abandonment.
Utility = pay + concave idleness value	Underloaded	Servers enjoy idle time
Increasing abandonment cost	Overloaded	Some customers abandon
Concave idleness value + increasing abandonment cost	Underloaded + Overloaded	Customer waiting and server idling are simultaneous

Takeaways

We propose a first best compensation and staffing scheme for a large-scale service system with strategic employees.

Facing multiple strategic employees, the optimal compensation design needs to counteract the speed-up effect due to competition.

Different optimal regimes arise depending on the cost structure and employee utility function.

Thank you!

Dongyuan Zhan

Email: d.zhan@ucl.ac.uk