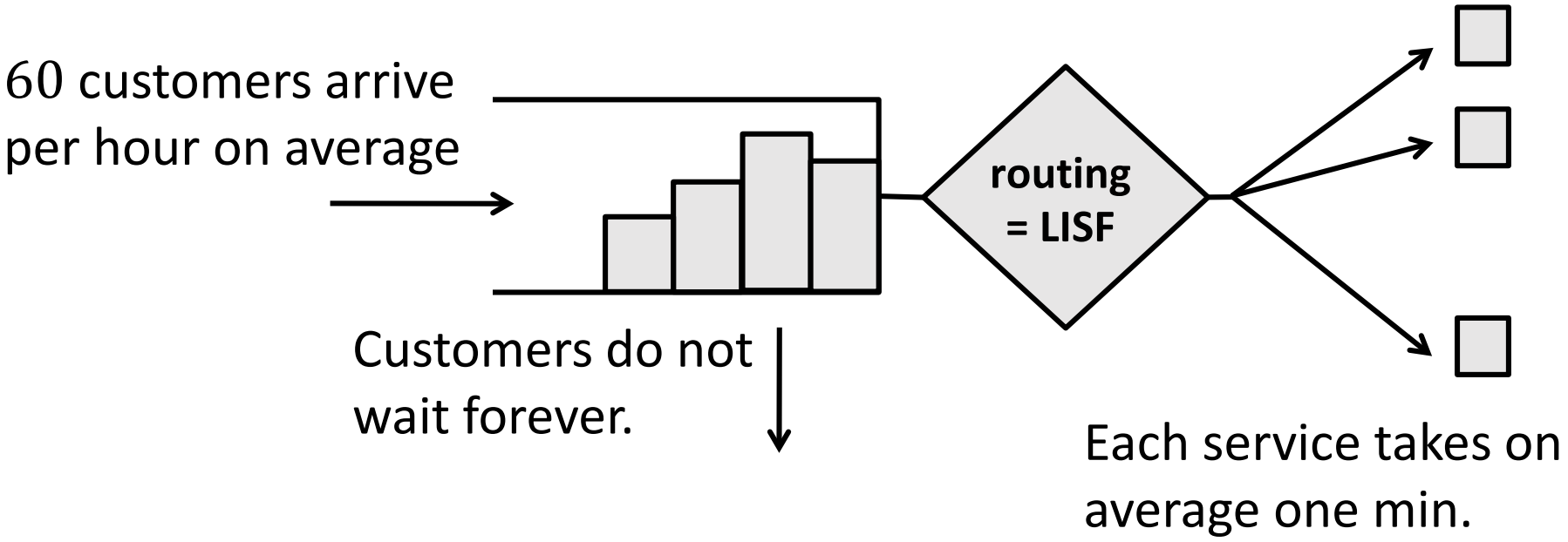
Staffing, Routing, and Payment to Trade Off Speed and Quality in Large Service Systems

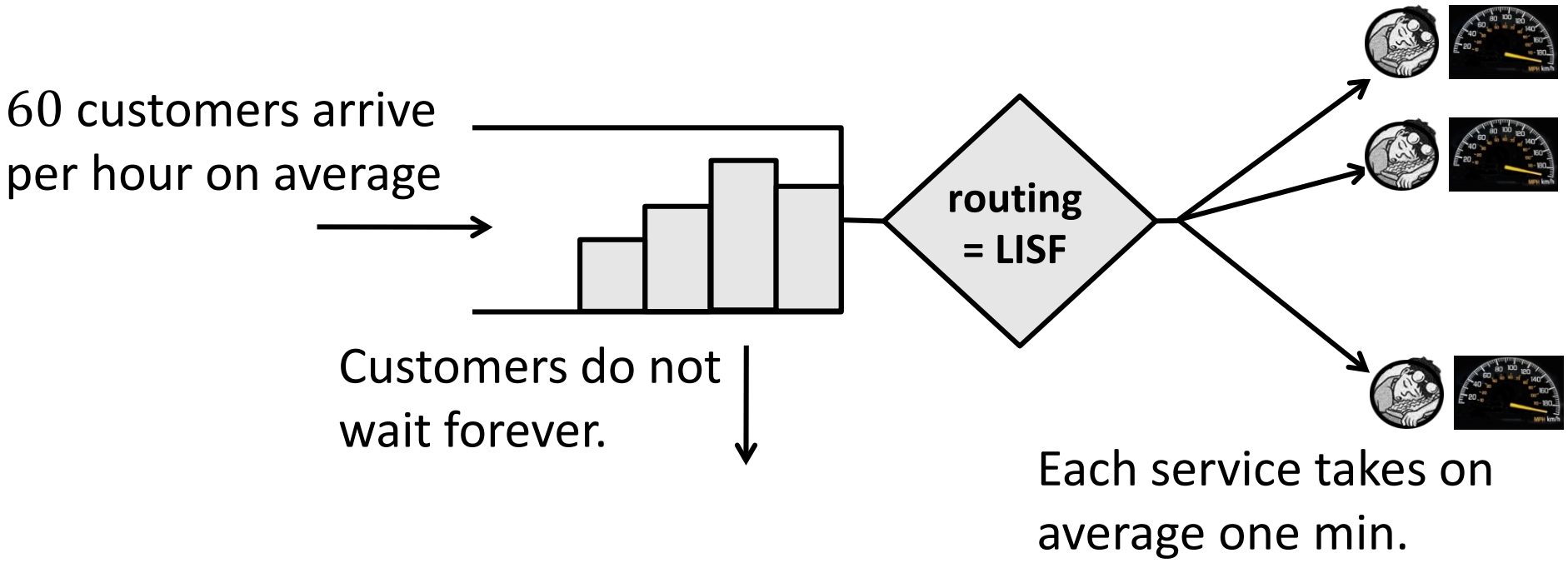
Amy Ward
The University of Chicago
Booth School of Business

*Joint work with Dongyuan Zhan, University College London

Q: How many employees to staff?



Q: Did you account for server behavior?



Example: Employees are not Machines

Imagine:

How do you respond?

You respond strategically.

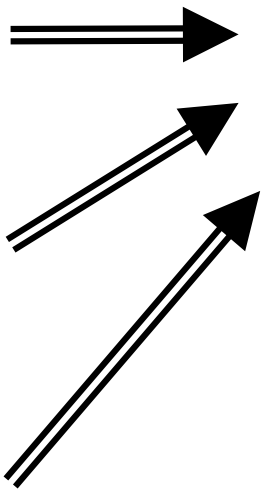
What if you were paid per review?

The takeaway: *We cannot necessarily assume a fixed average processing rate and staff accordingly.*

Research Positioning

We know

- Traditional queueing system
  - Arrival rate
  - Service rate
- Many queueing games
  - Customer utility
  - Service rate
- Our model
  - Arrival rate
  - Server utility



We want

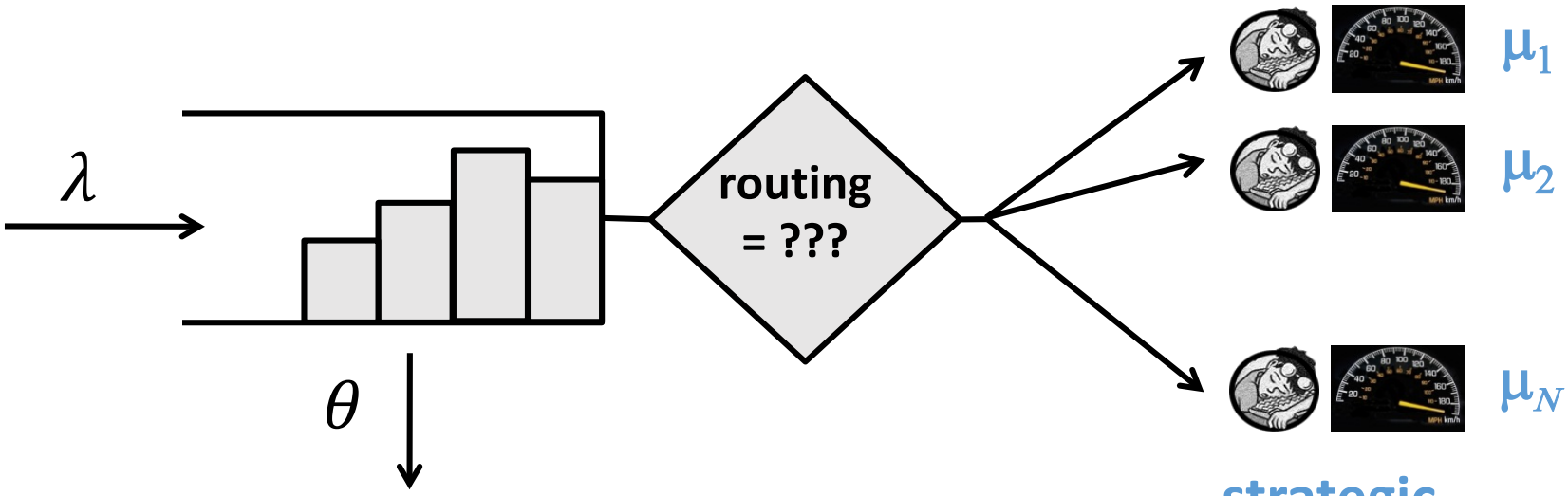
System performance
System design

- Next step
  - Customer utility
  - Server utility

Talk Outline

- The Problem Formulation
- The Centralized Control Problem
- The Decentralized Control Problem

Our Model: M/M/N$^{\text{strategic}}$ + M



- We assume a decreasing function $p(\mu)$ that represents the probability of successful service.
- Each server i selfishly chooses $\mu_i \in [\underline{\mu}, \overline{\mu}]$ to maximize payment.

risk neutral

expected

We must decide:

- How many employees to staff;
- How to route work;
- How to compensate work.

Preliminaries: The Class of Routing Policies

- An idle-time-order-based (IOB) routing policy is one that decides which server should handle an incoming arrival based only on the order in which the servers became idle (and cannot use service rate information).
- Common examples are: LISF, random.
- An IOB(T) policy delays each arriving customer for $T \in [0, \infty]$ time units (a "soft" admission control.)

Lemma 1. (Modification of Theorem 9 in Gopalakrishnan et al, 2016.) In an M/M/N+M queue, all IOB(T) routing policies have the same steady-state probabilities, and, as a consequence, result in the same expected steady-state utilization of server i,

$$B_i(\vec{\mu}, N, T) = \underbrace{\blacksquare}, i \in \{1, \ldots, N\}.$$

Closed-form expression is messy, and so not written on slides.

Some Related OM Literature

- Queueing games
  - Hassin and Haviv (2003), Hassin (2016)
- Strategic servers
  - Kalai, Kamien, Rubinovitch (1992), Gilbert and Weng (1998), Cachon and Harker (2002) , Cachon and Zhang (2007), Geng, Huh, and Nagarajan (2015)
- Speed-quality trade-off decisions
  - Hopp, Iravani, and Yuen (2007), Lu, Van Mieghem, and Savaskan (2009), Anand, Pac, and Veeraraghavan (2011)
- Large-scale “strategic” (customer or server) systems
  - Maglaras and Zeevi (2003) (2005), Armony and Maglaras (2004), Allon and Gurvich (2010), Armony and Gurvich (2010), Allon, Bassamboo and Cil (2014), Chan, Yom-Tov, and Escobar (2014), Gopalakrishnan, Doroudi, Ward, and Wierman (2016), Gurvich, Lariviere, Moreno-Garcia (2018), Ibrahim (2018)
- Empiric work
  - Hasija, Pinker, Shumsky (2010), Buell, Kim, Tsay (2016), Song, Tucker, Murrell (2015), Shunko, Niederhoff, Rosokha (2018)

A "Typical" Problem Formulation

Assume all homogeneous servers work at the same rate.

The manager decides the staffing level N and utilization β to:

$$\text{Minimize}_{\substack{N \in \{0,1,\ldots\} \\ \beta \in [0, B(\mu, N, T=0)]}} \quad c_s N + C(N, \beta).$$

Equivalently, admission delay T.

Busy time when all servers work at rate μ.

Performance-based costs;
See, for example, BMR (2004).
BMR = Borst, Mandelbaum, and Reiman.

$$C(N, \beta) =$$

$$(g_U(\beta)) \times N$$ Utilization cost.

$$+(\lambda - \mu N \beta) g_A\left(\frac{\lambda - \mu N \beta}{\lambda}\right)$$ Abandonment cost.

$$+\big(1 - p(\mu)\big)\mu N \beta\, g_F(1 - p(\mu))$$ Failed services cost.

Required Changes

Assume all homogeneous ~~servers work at the same rate~~.

The manager wants to solve the centralized control problem:

Payment from outside alternative;
The minimum the manager can pay.

Lower bound on ANY payment function.

Minimize $c_s N + C(\vec{\mu}, N, \cancel{\beta} T)$.

$\mu_i \in [\underline{\mu}, \overline{\mu}]$ $N \in \{0,1,\ldots\}$ $T \in [0,\infty]$

Let $\beta_i = B_i(\vec{\mu}, N, T)$. The cost function becomes:

Utilization cost.

Abandonment cost.

$$C(\vec{\mu}, N, T) = \sum_{i=1}^{N} g_U(\beta_i) + \left(\lambda - \sum_{i=1}^{N} \beta_i \mu_i\right) g_A\left(\frac{\lambda - \sum_{i=1}^{N} \beta_i \mu_i}{\lambda}\right)$$

$$+ \sum_{i=1}^{N} (1 - p(\mu_i)) \beta_i \mu_i g_F\left(\frac{\sum_{i=1}^{N} (1 - p(\mu_i) \beta_i \mu_i}{\sum_{i=1}^{N} \beta_i \mu_i}\right)$$

Failed services cost.

Same Service Rates are Optimal

Assume that on [0,1]

- g_U, g_F, g_A, p are all continuous;
- g_U is weakly increasing and strictly convex;
- g_F is weakly increasing and weakly convex;
- g_A is weakly increasing and $ag_A(a)$ is strictly convex;

And that on $\left[\underline{\mu}, \overline{\mu}\right]$:

- p is strictly decreasing and weakly concave.

Proposition 1:

Any solution to the centralized control problem has all servers working at the same service rate and having the same utilization.

Equivalent Centralized Control Problem

The manager decides the service rate μ, the staffing level N, and the admission delay T (which implies server utilization β) to:

$$\text{Minimize} \quad c_s N + \underbrace{C(\mu, N, T)}.$$

$\mu \in \left[\underline{\mu}, \overline{\mu}\right]$

$N \in \{0,1, \ldots\}$ Performance-based costs.

$T \in [0, \infty]$

$\beta = B(\mu, N, T)$

$$C(\mu, N, T) =$$

$$(g_U(\beta)) \times N$$

Utilization cost.

$$+(\lambda - \mu N \beta) g_A\left(\frac{\lambda - \mu N \beta}{\lambda}\right)$$

Abandonment cost.

$$+\big(1 - p(\mu)\big)\mu N \beta\, g_F(1 - p(\mu))$$

Failed services cost.

Asymptotic Analysis of the Centralized Control Problem (1/4)

We consider a sequence of systems with increasing arrival rate $\lambda \to \infty$.

Satisfies the constraints of the centralized control problem for each λ.

Definition: An admissible policy $\{(\mu^\lambda, N^\lambda, T^\lambda): \lambda \geq 0\}$ is *asymptotically optimal* if

$$\lim_{\lambda \to \infty} \frac{c_S N^\lambda + C(\mu^\lambda, N^\lambda, T^\lambda)}{c_S N_*^\lambda + C(\mu_*^\lambda, N_*^\lambda, T_*^\lambda)} = 1.$$

Optimal objective function value for given λ.

Asymptotic Analysis of the Centralized Control Problem (2/4)

Adjusted staffing cost.

For $\hat{c}_S(\beta) = \frac{c_S + g_U(\beta)}{\beta}$, solve the one-dimensional optimizations:

- $\hat{\beta}_* = \underset{\beta \in [0,1]}{\operatorname{argmin}} \{\hat{c}_S(\beta)\}$;
- $\hat{\mu}_* = \underset{\mu \in [\underline{\mu}, \overline{\mu}]}{\operatorname{argmin}} \left\{ \frac{\hat{c}_S(\beta_*)}{\mu} + \big(1 - p(\mu)\big) g_F\big(1 - p(\mu)\big) \right\}$

The cost to serve a customer, adjusted to include both utilization and service failure.

Service cost.

- $\hat{a}_* = \underset{a \in [0,1]}{\operatorname{argmin}} \left\{ (1-a)\left(\frac{\hat{c}_S(\beta_*)}{\mu_*} + \big(1 - p(\mu_*)\big) g_F\big(1 - p(\mu_*)\big) \right) + a g_A(a) \right\}$,

Abandonment cost.

and set $\hat{b}_* = \frac{1 - \hat{a}_*}{\hat{\beta}_* \hat{\mu}_*}$.

Asymptotic Analysis of the Centralized Control Problem (3/4)

Theorem 1. Under the assumption stated earlier, any policy

$\left(\hat{\mu}_*, N_{ao}^{\lambda} = \hat{b}_*\lambda + o(\lambda), \hat{T}_*\right)$ with $\hat{T}_* = \begin{cases} -\frac{\log(\hat{b}_*\hat{\mu}_*\hat{\beta}_*)}{\theta}, & \text{if } \hat{\beta}_* < 1, \hat{a}_* > 0, \\ 0, & \text{otherwise,} \end{cases}$

is asymptotically optimal. Furthermore,

$$\lim_{\lambda\to\infty} B\left(\mu_*^{\lambda}, N_*^{\lambda}, T_*^{\lambda}\right) = \underbrace{\hat{\beta}_* \text{ and } \hat{\beta}_*}_{} = \min\left(1, \frac{exp(-\theta\hat{T}_*)}{\hat{b}_*\hat{\mu}_*}\right).$$

The solution to 1-D optimization on previous slide.

Asymptotic Analysis of the Centralized Control Problem (4/4)

Define $\hat{c}_* = \frac{\hat{c}_S(\beta_*)}{\mu} + \big(1 - p(\mu)\big) g_F(1 - p(\mu))$ to be the min. cost to serve a customer.

Assume $g_U(\beta) = k\beta^r$, for $r > 1$ and $\beta \in [0,1]$.

Ab. cost	Util. cost	Optimal regime	Ab. %	Server Util.
$g_A(0) > \hat{c}_*$	$k \leq \frac{c_S}{r-1}$	Critically Loaded	$a_\star = 0$	$\beta_\star = 1$
$g_A(0) \leq \hat{c}_*$	$k \leq \frac{c_S}{r-1}$	Efficiency-Driven	$a_\star > 0$	$\beta_\star = 1$
$g_A(0) > \hat{c}_*$	$k > \frac{c_S}{r-1}$	Quality-Driven	$a_\star = 0$	$\beta_\star < 1$
$g_A(0) \leq \hat{c}_*$	$k > \frac{c_S}{r-1}$	Intentional Idling	$a_\star > 0$	$\beta_\star < 1$

Remark 1: Compare to BMR (2004), which considered a M/M/N model.
An analysis for a full spectrum of cost functions had not been done for M/M/N+M.

Asymptotic Analysis of the Centralized Control Problem (4/4)

Define $\hat{c}_* = \frac{\hat{c}_S(\beta_*)}{\mu} + \big(1 - p(\mu)\big) g_F(1 - p(\mu))$ to be the min. cost to serve a customer.

Assume $g_U(\beta) = k\beta^r$, for $r > 1$ and $\beta \in [0,1]$.

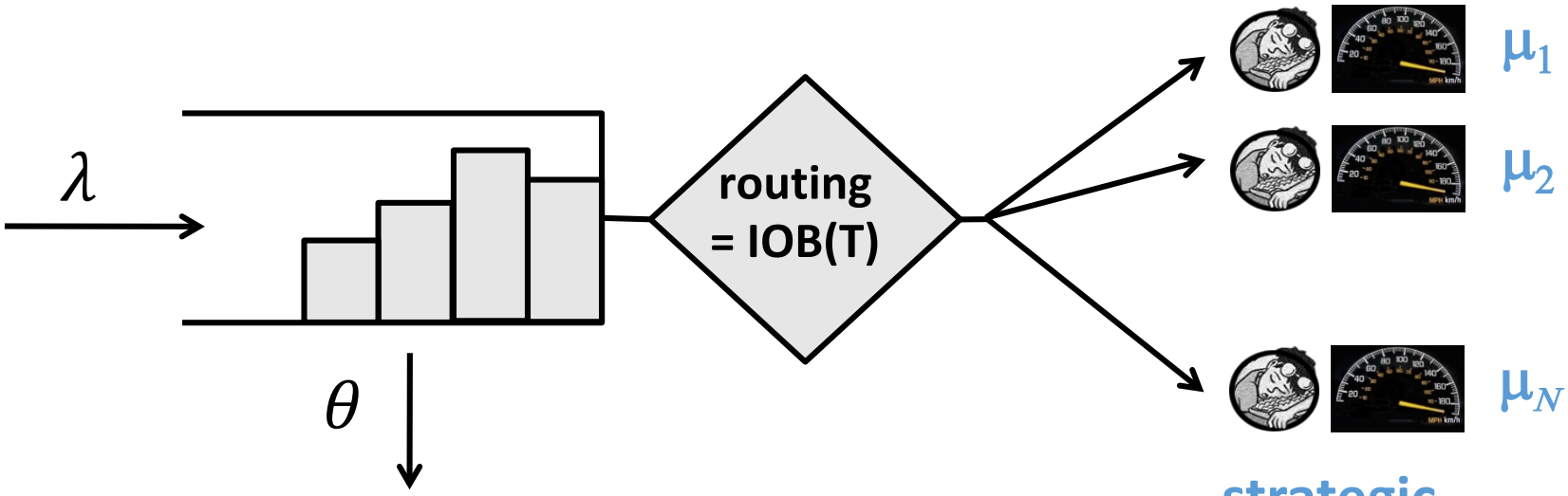
Ab. cost	Util. cost	Optimal regime	Ab. %	Server Util.
$g_A(0) > \hat{c}_*$	$k \leq \frac{c_S}{r-1}$	Critically Loaded	$a_\star = 0$	$\beta_\star = 1$
$g_A(0) \leq \hat{c}_*$	$k \leq \frac{c_S}{r-1}$	Efficiency-Driven	$a_\star > 0$	$\beta_\star = 1$
$g_A(0) > \hat{c}_*$	$k > \frac{c_S}{r-1}$	Quality-Driven	$a_\star = 0$	$\beta_\star < 1$
$g_A(0) \leq \hat{c}_*$	$k > \frac{c_S}{r-1}$	Intentional Idling	$a_\star > 0$	$\beta_\star < 1$

Remark 2: Conditions for critically loaded are consistent with Proposition 5 in Bassamboo and Randhawa, Proposition 1 in Ren and Zhou (2008), and Proposition 1 in Whitt (2006).

Talk Outline

- The Problem Formulation
- The Centralized Control Problem
- The Decentralized Control Problem

The Server Utility



- We assume a decreasing function $p(\mu)$ that represents the probability of successful service.
- Each server i selfishly chooses $\mu_i \in [\underline{\mu}, \bar{\mu}]$ to maximize payment.

risk neutral — expected

- An equilibrium is a service rate vector $\vec{\mu}$ that satisfies

$$U_i(\vec{\mu}) = \max_{v \in [\underline{\mu}, \bar{\mu}]} U_i(\mu_1, \dots, \mu_{i-1}, v, \mu_{i+1}, \dots, \mu_N), \text{ for all } i \in \{1, \dots, N\},$$

and individual rationality (IR); that is, $U_i(\vec{\mu}) \geq c_S$, for all $i \in \{1, \dots, N\}$.

c_S: Employment alternative

The De
Recall: Equivalent Centralized Control Problem

payment function

The manager decides the service rate μ, the staffing level N, and the admission delay T (which implies server utilization β) to:

$$\text{Minimize} \quad \sum_{i=1}^{N} E[P_i] + c_S N + C(\vec{\mu}_E, N, T).$$

$P \in \mathcal{P}$ ~~$\mu \in [\underline{\mu}, \bar{\mu}]$~~ *

$N \in \{0,1,\dots\}$

$T \in [0,\infty]$

$E[P_i] \geq c_S$

Performance-based costs.

$\mathcal{P}$ is the class of payment functions based on the observable or known elements:

- $\lambda, g_U, g_A, g_F, p$
- The realized number of abandonments;
- The realized number of completed and failed services in a finite time interval.

Also need to ensure an equilibrium $\vec{\mu}_E$ exists and account for potential non-uniqueness.

The Decentralized Control Problem

The manager decides the payment function $\vec{P}$,the staffing level N, and the admission delay T (which implies server utilization β) to:

Minimize $$\sup_{\vec{\mu}_E \in \mathcal{S}(\vec{P},N,T)} \sum_{i=1}^{N} \underbrace{E[P_i]}_{U_i} + C(\vec{\mu}_E, N, T)$$

The set of equilibrium service rates.

Subject to:

- $N \in \{0,1,\dots\}$
- $T \in [0,\infty]$
- $\mathcal{S}(\vec{P},N,T) \neq \emptyset$
- $\min_{\vec{\mu}_E \in \mathcal{S}(\vec{P},N,T)} E[P_i] \geq c_S$ for each $i \in \{1,\dots,N\}$.

Note that the solution to the centralized control problem is a lower bound on the decentralized control problem, because any equilibrium service rate is feasible for the centralized control problem.

Limiting First Best Payment

Let $N_{ao}^{\lambda} = \hat{b}_* \lambda + o(\lambda)$. We would like to find a sequence of contracts $\vec{P}^{\lambda} \in \mathcal{P}$ for all λ such that

$$\left\{\left(\vec{P}^{\lambda}, N_{ao}^{\lambda}, \hat{T}_*\right): \lambda \geq 0\right\}$$

satisfy the decentralized control problem constraints for all λ , and any sequence of symmetric equilibrium service rates satisfies:

$$\lim_{\lambda \to \infty} \left|\mu_E^{\lambda} - \hat{\mu}_\star\right| = 0 \text{ and } \lim_{\lambda \to \infty} E\left[P_i^{\lambda}\right] - c_S = 0, i \in \left\{1,2, \dots, N_{ao}^{\lambda}\right\}.$$

Then, the solutions to the centralized and decentralized control problem become identical as λ becomes large; that is,

$$\lim_{\lambda \to \infty} \sup_{\vec{\mu}_E \in \mathcal{S}(\vec{P},N,T)} \frac{\sum_{i=1}^{N_{ao}^{\lambda}} E\left[P_i^{\lambda}\right] + C\left(\mu_E^{\lambda}, N_{ao}^{\lambda}, \hat{T}_*\right)}{c_S N_*^{\lambda} + C\left(\mu_*^{\lambda}, N_*^{\lambda}, T_*^{\lambda}\right)} = 1,$$

under earlier stated Assumption.

Optimal centralized control problem objective function value for given λ.

Piecerate Payment

When the service rate vector is $\vec{\mu}^{\lambda}$, staffing level is N^{λ}, and routing parameter is T^{λ}, under piecerate payment, the expected payment per time unit to server i is:

$$U_i^{\lambda} = E\left[P_i^{\lambda}\right] = \left(P_S^{\lambda} - P_F^{\lambda}(1 - p(\mu_i))\right)\mu_i \times B_i\left(\vec{\mu}^{\lambda}, N^{\lambda}, T^{\lambda}\right), \mathrm{i} \in \{1, \ldots, N^{\lambda}\}.$$

Focus on tagged server 1. We would like to solve for a fixed point of

$$R^{\lambda}(\mu) = \arg\max_{\mu_1 \in [\underline{\mu}, \overline{\mu}]} \underbrace{U^{\lambda}(\mu_1, \mu)}_{\text{Expected payment to server 1 when all other servers work at rate } \mu.},$$

where

$$U^{\lambda}(\mu_1, \mu) = \left(P_S^{\lambda} - P_F^{\lambda}(1 - p(\mu_1))\right)\mu_1 \times B\left((\mu_1, \mu), N^{\lambda}, T^{\lambda}\right).$$

Such a fixed point is a symmetric equilibrium service rate when IR holds.

An Approximate Equilibrium Service Rate (1/2)

$$R^{\lambda}(\mu) = \arg\max_{\mu_1 \in [\underline{\mu}, \overline{\mu}]} U^{\lambda}(\mu_1, \mu),$$

$$U^{\lambda}(\mu_1, \mu) = \left(P_S^{\lambda} - P_S^{\lambda}(1 - p(\mu_1))\right)\mu_1 \times \hat{B}\big((\mu_1, \mu)\big), \quad \text{~~} B\big((\mu_1, \mu), N^{\lambda}, T^{\lambda}\big) \text{~~}.$$

Proposition. Fix $b \geq 0$. Under IOB(T) routing and staffing $N^{\lambda} = b\lambda + o(\lambda)$, for any $\mu_1, \mu \in [\underline{\mu}, \overline{\mu}]$,

$$\lim_{\lambda \to \infty} B\big((\mu_1, \mu), N^{\lambda}, T^{\lambda}\big) = \hat{B}\big((\mu_1, \mu)\big),$$

where

$$\hat{B}\big((\mu_1, \mu)\big) = \frac{\mu \exp(-\theta T)}{\mu \exp(-\theta T) + \mu_1 [b\mu - \exp(-\theta T)]^+}.$$

An Approximate Equilibrium Service Rate (2/2)

$$R^{\lambda}(\mu) = \arg\max_{\mu_1 \in [\underline{\mu}, \overline{\mu}]} U^{\lambda}(\mu_1, \mu),$$

$$U^{\lambda}(\mu_1, \mu) = \left(P_S^{\lambda} - P_S^{\lambda}(1 - p(\mu_1))\right)\mu_1 \times \hat{B}\left((\mu_1, \mu)\right),$$

~~$\times B\left((\mu_1, \mu), N^{\lambda}, T^{\lambda}\right)$~~.

Lemma. Given $b \geq 0, \mu \in [\underline{\mu}, \overline{\mu}]$, and $T > 0$, define

$$P_R(b, \mu, T) = \frac{1}{1 - p(\mu) - \mu p'(\mu) \max\{b\mu \exp(\theta T), 1\}}.$$

If $P_S > 0$ and $P_F \geq 0$ satisfy $\frac{P_F}{P_S} = P_R$, then μ is the unique fixed point.

Theorem 2. The piecerate payment that sets $\frac{P_F^{\lambda}}{P_S^{\lambda}}$ to equal $P_R(b^*, \mu^*, T^*)$, and P_S^{λ} to ensure IR $\left(U_i^{\lambda} \geq c_S, i \in \{1, \ldots, N_{ao}^{\lambda}\}\right)$, is limiting first best.

Some Intuition

Define $P_S^* = \lim_{\lambda\to\infty} P_S^\lambda$ and $P_F^* = \lim_{\lambda\to\infty} P_F^\lambda$.

Example 1. Suppose each service failure costs c_F ($g_F(x) = c_F$).

- Case 1: $\beta^* = 1$. Then

$$P_S^* = \frac{c_S}{c_S+g_U(1)}\underbrace{\hat{c}_*}_{\text{The min cost to serve a customer.}} \text{ and } P_F^* = \frac{c_S}{c_S+g_U(1)}c_F;$$

i.e., if there is no utilization cost, pay $\hat{c}_*$ per service completion; else, lower payment and keep ratio unchanged.

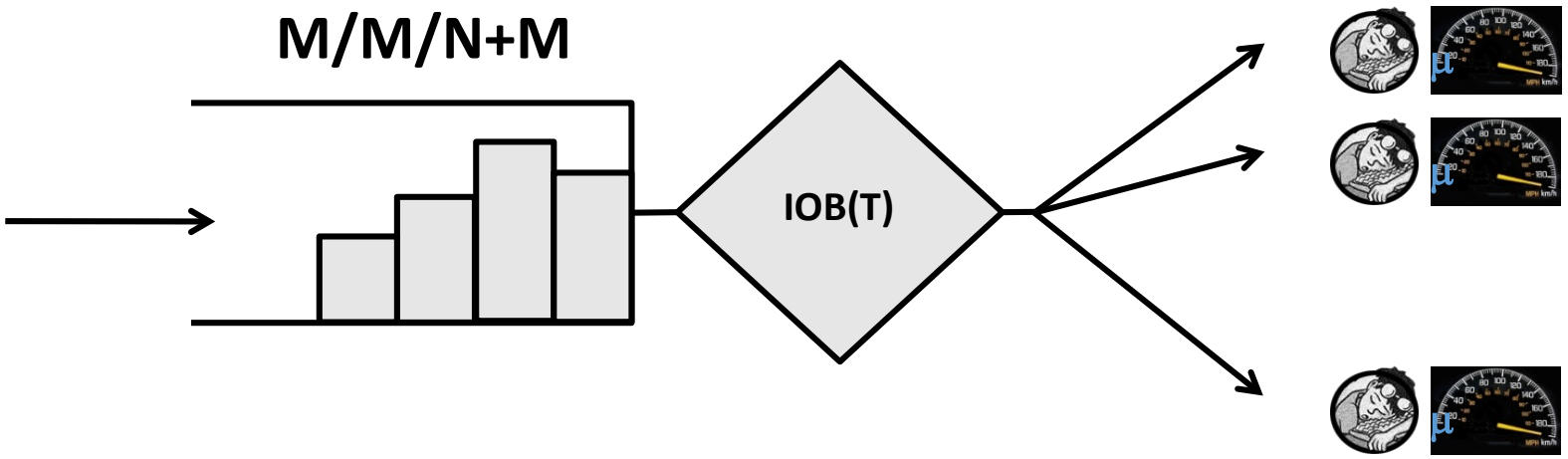
- Case 2: $\beta^* < 1$. Then

$$P_S^* = \blacksquare > \frac{c_S}{c_S+g_U(1)}\hat{c}_* \text{ and } P_F^* = \blacksquare > \frac{c_S}{c_S+g_U(1)}c_F.$$

In words, the manager transfers her costs to the servers in a way that induces the servers to work at rate $\hat{\mu}_*$.

Concluding Remarks

We need to rethink optimal system design to account for how servers respond to incentives (i.e., when servers are strategic)!



The manager should jointly optimize over the staffing, routing, and service speed, and then provide incentives to achieve her desired service speed.

Future research: Other routing policies? For example, Gopalakrishnan (2019).

*Paper forthcoming in *Operations Research*, available from my web page: http://faculty.chicagobooth.edu/Amy.Ward/publications.html