# Behavior-Aware Queueing:

## The Finite-Buffer Setting with Many Strategic Servers

Yueyang Zhong
The University of Chicago Booth School of Business

Raga Gopalakrishnan
Smith School of Business at Queen's University

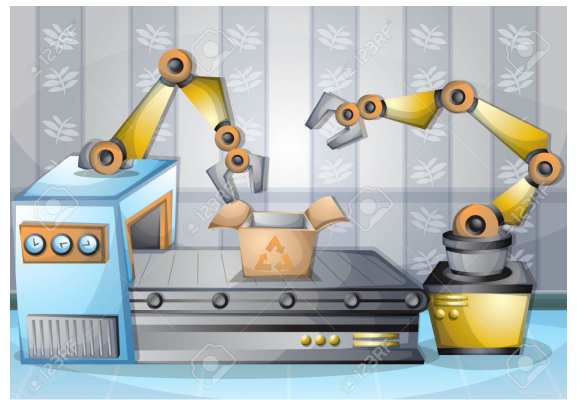
Amy Ward
The University of Chicago Booth School of Business

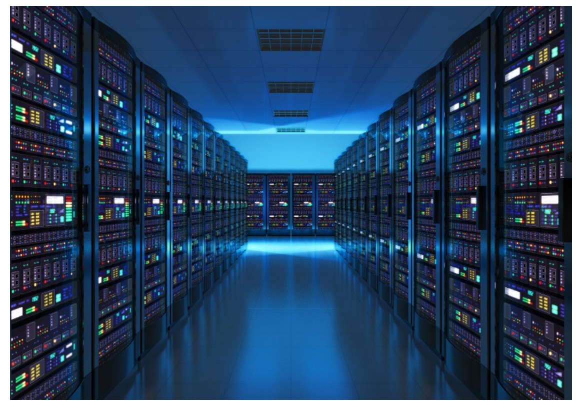
June 23, 2022

**Paper Link:** http://ssrn.com/abstract=3633435

CHICAGO BOOTH
The University of Chicago Booth School of Business

# Motivation






| | **Traditional Queueing (Manufacture/Computer Science)** | **Behavior-Aware Queueing (Service Operations)** |
|---|---|---|
| | Arrival and service processes are assumed to be exogenous | Arrival and service processes are endogenously determined by customer and worker's utility functions |
| **Demand side** | Inanimate jobs | Human customers<br>⟹ may not wait in queue forever [Naor (1969), Knudsen (1972)] |
| **Supply side** | Inanimate machines | Human workers<br>⟹ service rates are not exogeneous<br>speedup (social pressure)<br>slowdown (fatigue or social loafing) |

# Strategic Arrivals

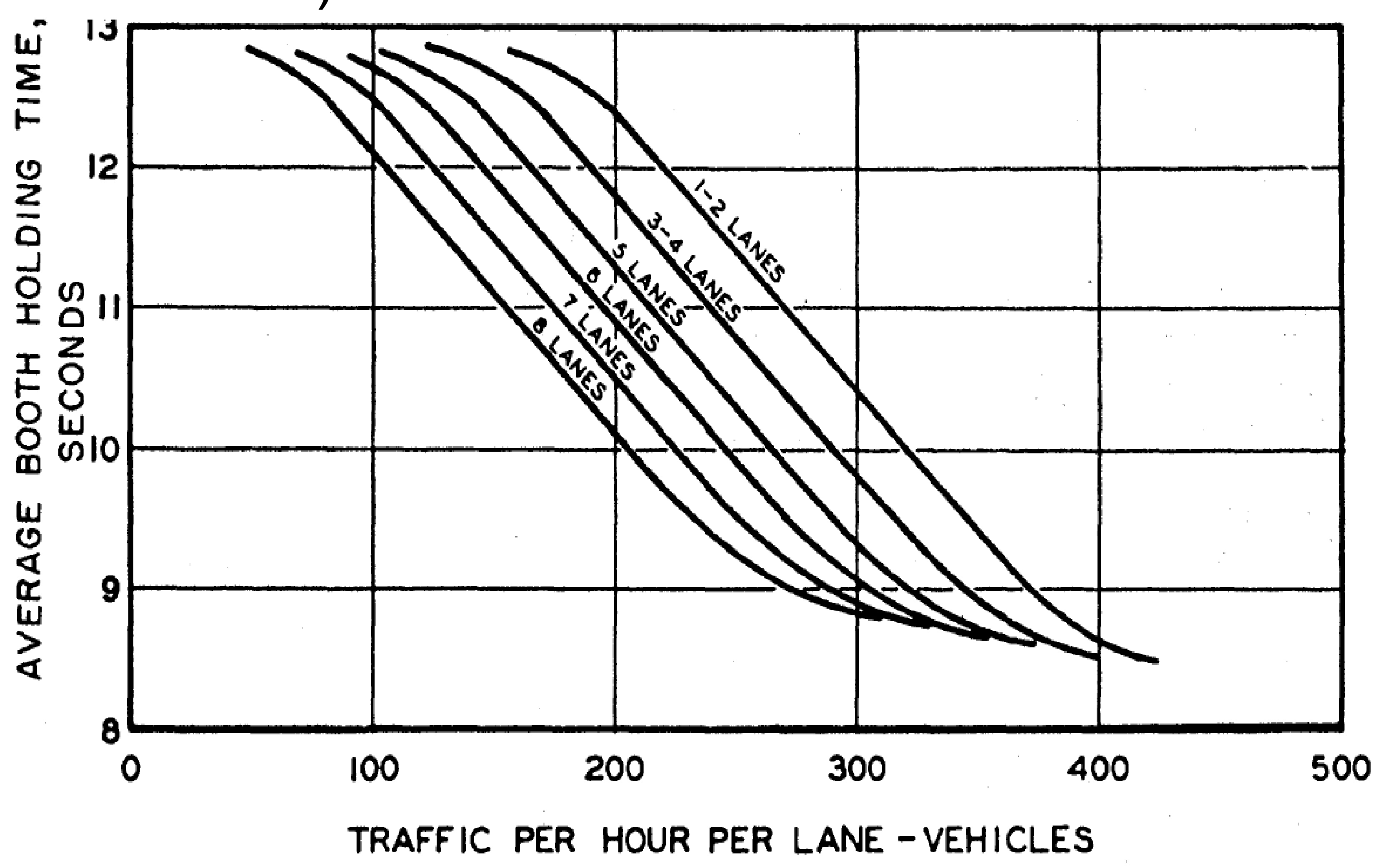
Congestion impacts joining decision.

Foundational Econometrica papers
Naor (1969), Knudsen (1972)

# Strategic Servers

Congestion impacts service time.


(Service time)




Edie (1954), Fig 7: Average booth holding time per vehicle at George Washington Bridge

Larger goal is to understand the behavior of strategic customers, strategic servers, and their interactions.

# Objective

To develop an analytical queueing model to investigate how server work speed is affected by system design decisions concerning

- How many servers to staff and how much to pay them;
- Whether and when to turn away customers.

# Literature Review

**Empirical literature**

[Delasay et al., 2019] (survey)

[Kc & Terwiesch, 2009]

[Staats & Gino, 2012]

[Mas & Morretti, 2009]

**Large system asymptotic analysis**

[Ibrahim, 2018]

[Dong and Ibrahim, 2020]

[Zhan and Ward, 2019]

[Gopalakrishnan et al., 2016]

**Queueing game**

[Hassin and Haviv, 2003] (survey book)

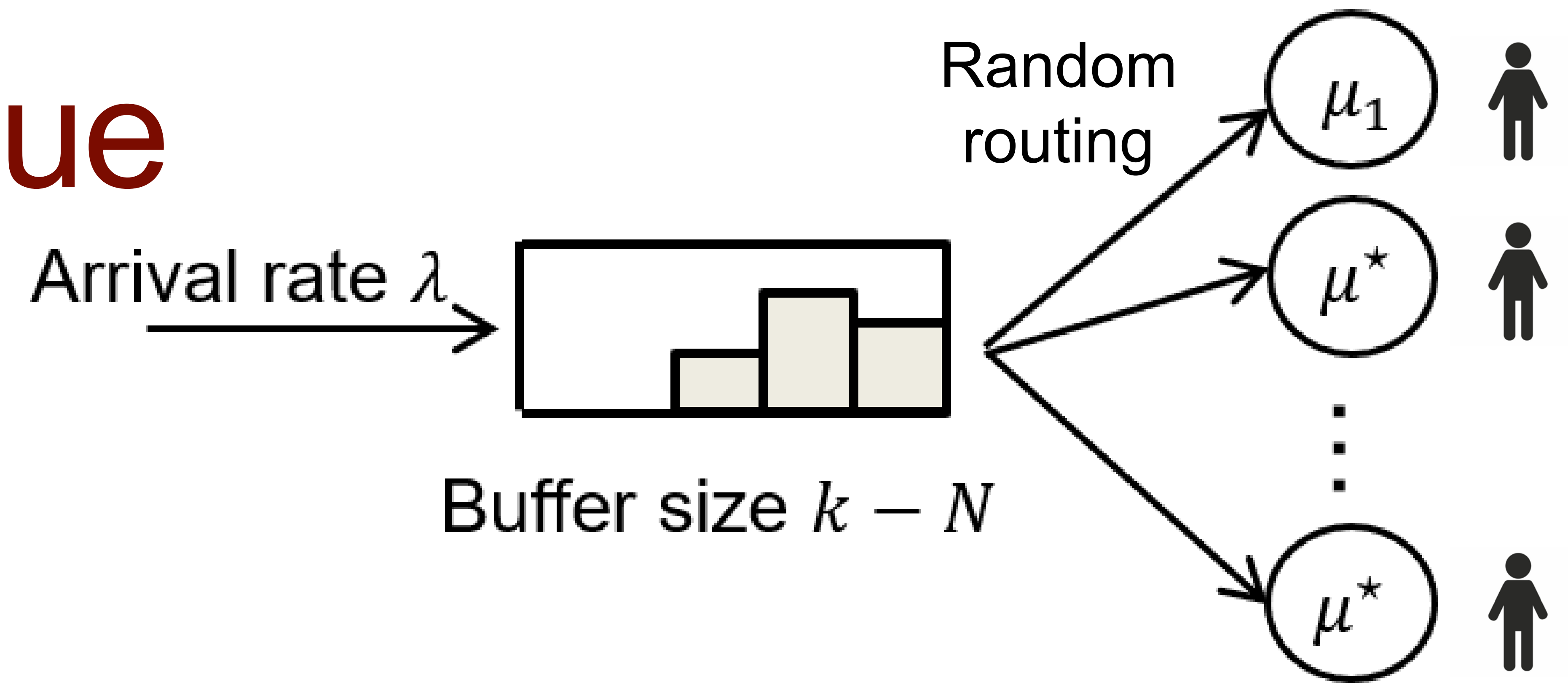
[Hassin, 2016] (survey book)

[Allon and Kremer, 2018] (survey chapter)

# Outline

- Model
- Asymptotic Analysis
- Looking Ahead

# Model: Strategic Server $M/M/N/k$ Queue




**Definition: Nash Equilibrium Service Rate**

The servers want to choose rates $\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_N)$ that satisfy

$$U_i(\vec{\mu}) = \max_{\mu_i \geq 0} \underbrace{p\mu_i B_i(\vec{\mu})}_{\text{Payment}} + \underbrace{vI_i(\vec{\mu})}_{\text{Idleness}} - \underbrace{c(\mu_i)}_{\text{Effort cost}},$$

**Definition: Symmetric Equilibrium**

$\mu^\star \in \text{argmax}_{\mu_1 \geq 0}\, U_1(\mu_1, \mu^\star)$ where

$$U_1(\mu_1, \mu) = p\mu_1 B_1(\mu_1, \mu) + vI_1(\mu_1, \mu) - c(\mu_1).$$

Notation:
$N$ = number of servers
$\lambda$ = Poisson arrival rate
$\mu_i$ = service rate of server $i$
$k$ = system size
$k - N$ = buffer, or waiting room size

**Individual Rationality:** $U(\mu^\star, \mu^\star) \geq 0$

Notation and Assumption:
$B_i(\vec{\mu})$ = Long-run average fraction of time servers are busy;
$I_i(\vec{\mu})$ = Long-run average fraction of time servers are idle, as opposed to working to serve customers;
$c(\mu)$ = Effort cost function that is continuous, differentiable, strictly increasing, strictly convex, and has $c(0) = 0$.

# Equilibrium Analysis

Equilibrium Analysis Steps: (1) Satisfy first-order condition (FOC); (2) Global maximum.

$$\left.\frac{\partial U(\mu_1,\mu)}{\partial \mu_1}\right|_{\mu_1=\mu} = 0 \quad \Leftrightarrow \quad p\big(1 - I(\mu,\mu)\big) + (v - p\mu)\left.\frac{\partial I}{\partial \mu_1}\right|_{\mu_1=\mu} = c'(\mu).$$

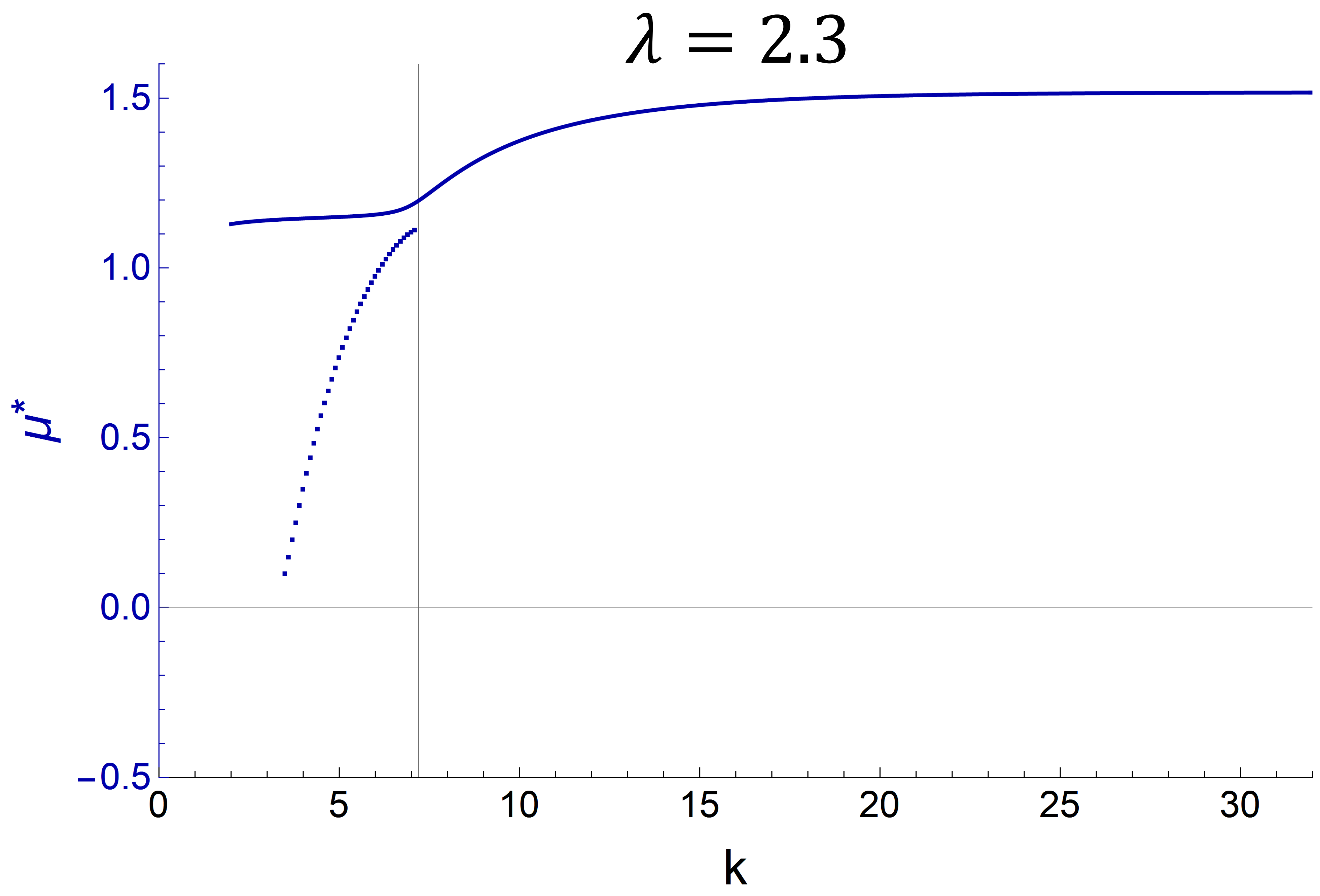
**Lemma**: In an $M/M/N/k$ system with $N-1$ servers operating at rate $\mu > 0$, and a tagged server with rate $\mu_1 > 0$.

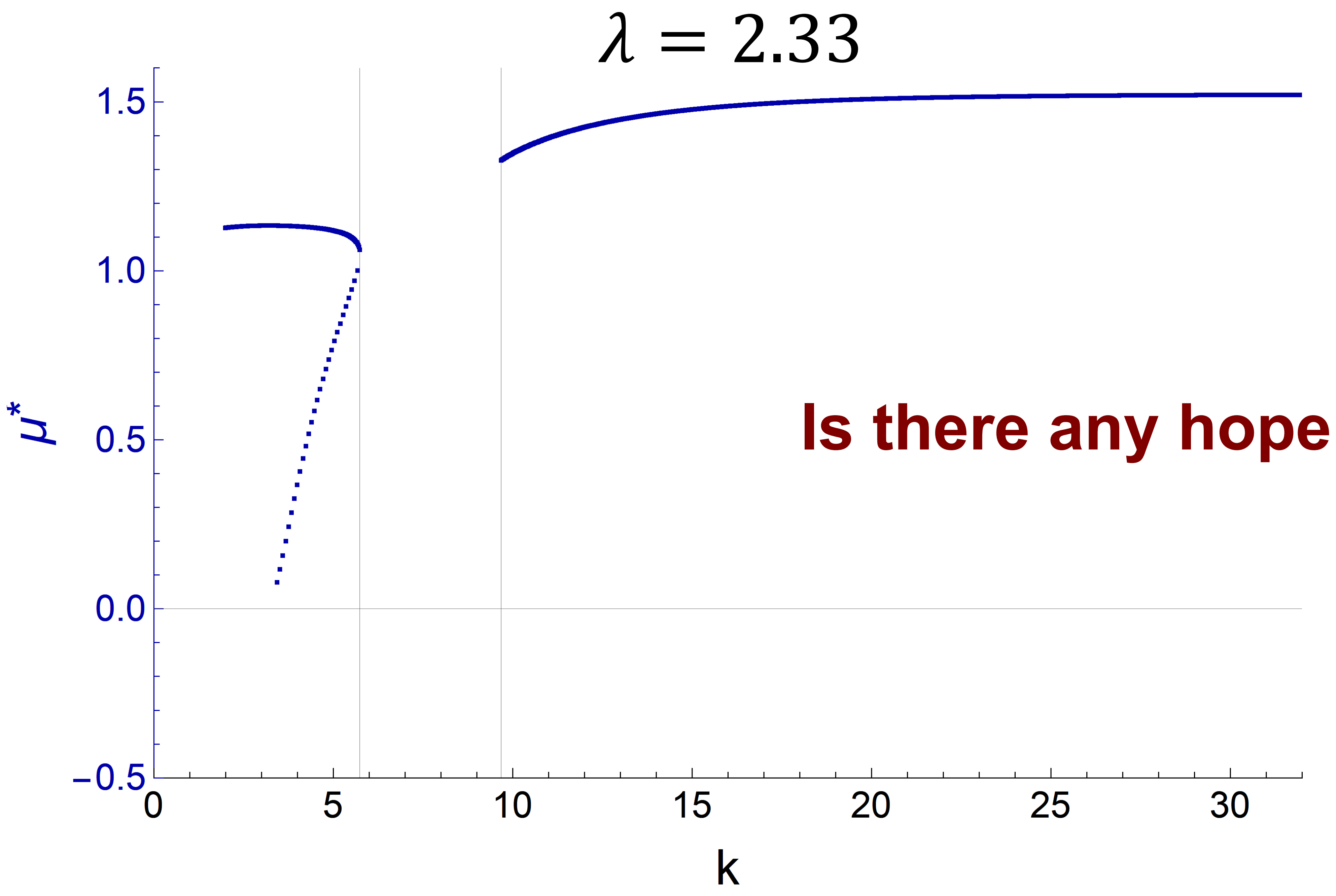
$$I(\mu_1,\mu) = \left(1 + \rho\frac{\mu}{\mu_1}\left(\frac{1-C}{N-\rho} + \left(1 - \left(\frac{\rho}{N - \left(1 - \frac{\mu_1}{\mu}\right)}\right)^{k-N}\right)\frac{C}{(N-\rho) - \left(1 - \frac{\mu_1}{\mu}\right)}\right)\right)^{-1},$$

where $\rho = \frac{\lambda}{\mu}$ and $C \coloneqq ErlC(N,\rho) = \frac{\frac{\rho^N}{N!}\cdot\frac{N}{N-\rho}}{\sum_{j=0}^{N-1}\frac{\rho^j}{j!} + \frac{\rho^N}{N!}\cdot\frac{N}{N-\rho}}$.

**How do solutions to the FOC that are equilibria behave?**

# $M/M/2/k$ System: $N = 2, p = 0, v = 1, c(\mu) = \frac{3}{32}\mu^2$




**Is there any hope for analysis?**

- Not unique
- Not continuous
- Not monotonic

# Outline

- Model
- Asymptotic Analysis
- Looking Ahead

# Asymptotic Analysis

Consider a sequence of $M/M/N^{\lambda}/k^{\lambda}$ systems, and let $\lambda$ become large:

- $N^{\lambda}$: the staffing level
- $k^{\lambda} \geq N^{\lambda}$: the system size
- $\mu^{\star,\lambda}$: prelimit equilibrium

Linear staffing with parameter $a$

**Lemma:** Fix $\mu > 0$. If $\boxed{N^{\lambda} = \frac{1}{a}\lambda + o(\lambda)}$, for $a > 0$,

$$\lim_{\lambda\to\infty} I^{\lambda}(\mu,\mu) = \left[1 - \frac{a}{\mu}\right]^{+} \text{ and } \lim_{\lambda\to\infty} \left.\frac{\partial I^{\lambda}(\mu_1,\mu)}{\partial \mu_1}\right|_{\mu_1=\mu} = \frac{a[\mu-a]^{+}}{\mu^3}.$$

Otherwise, if $N^{\lambda} = f(\lambda) + o\big(f(\lambda)\big)$ for $f(\lambda) = o(\lambda)$ or $f(\lambda) = \omega(\lambda)$, then $\lim_{\lambda\to\infty} \left.\frac{\partial I^{\lambda}}{\partial \mu_1}\right|_{\mu_1=\mu} = 0$ or $1$ (degenerate).

**Proposition:** For large enough $\lambda$, $\frac{\partial^2 U^{\lambda}(\mu_1,\mu)}{\partial \mu_1^2} < 0$ for all $\mu_1 > 0$ and $\mu > 0$.

# Asymptotic Analysis: Existence

$$N^\lambda = \frac{1}{a}\lambda + o(\lambda)$$

**FOC:** $p\left(1 - I^\lambda(\mu,\mu)\right) + (v - p\mu)\left.\frac{\partial I^\lambda(\mu_1,\mu)}{\partial \mu_1}\right|_{\mu_1=\mu} = c'(\mu).$

**Limiting FOC:** $p\left(1 - \left[1 - \frac{a}{\mu}\right]^+\right) + (v - p\mu)\frac{a[\mu - a]^+}{\mu^3} = c'(\mu)$

**Theorem:** There exists $\bar{a}(p,v)$ such that:

Underloaded Equilibria $\left(\frac{a}{\mu} < 1\right)$ → $n_u$; Overloaded Equilibria $\left(\frac{a}{\mu} > 1\right)$ → $n_o$

$p \le c'(0)$

| | $n_u$ | $n_o$ | |
|---|---|---|---|
| $a < \bar{a}(p,v)$ | 2 | 0 | ← Blue |
| $a > \bar{a}(p,v)$ | 0 | 0 | ← Purple |

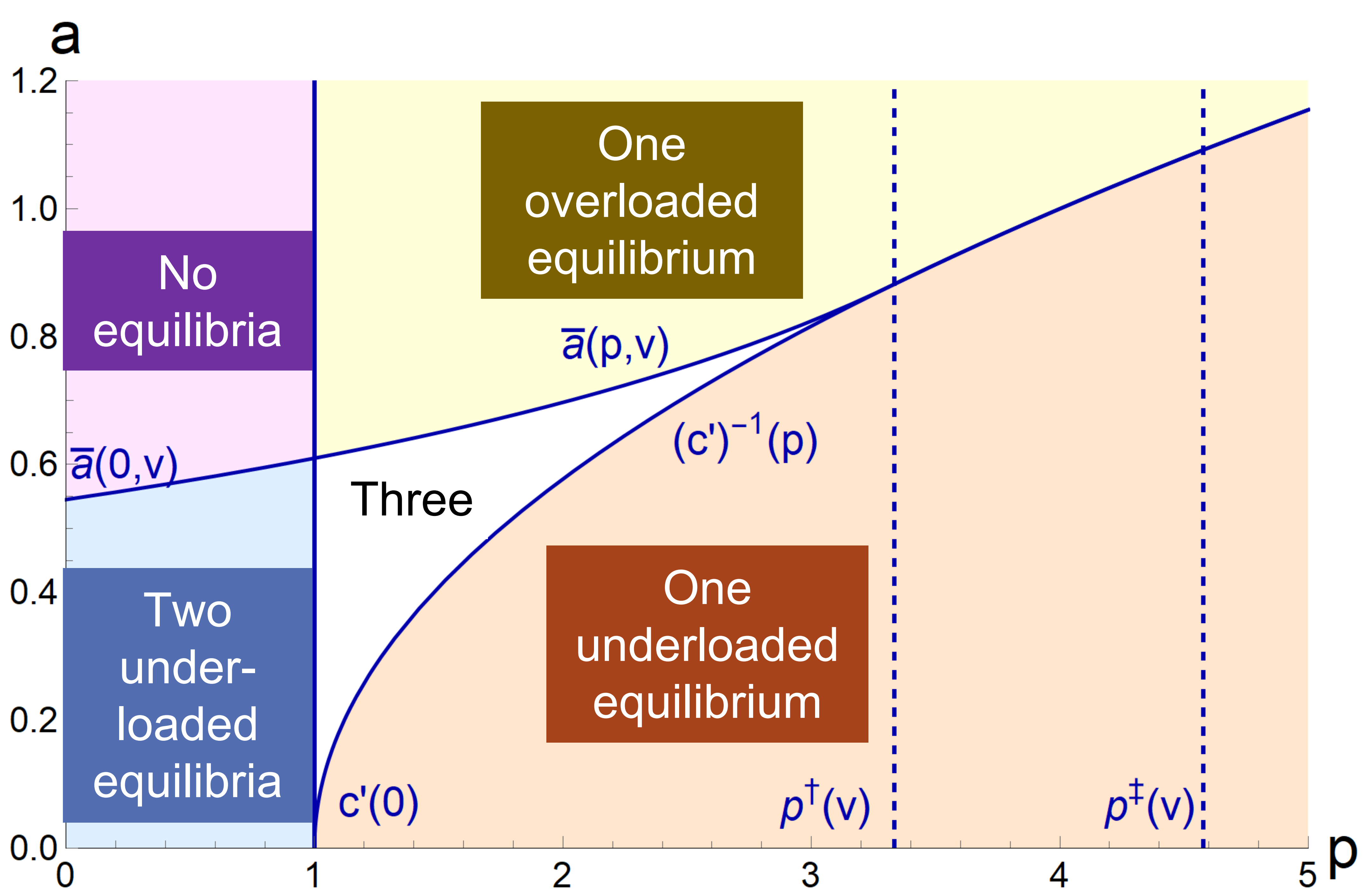
(ignoring boundary behavior; $a = \bar{a}(p,v)$ iff $n_c = 1$)

Critically loaded equilibria $\left(\frac{a}{\mu} = 1\right)$

$p > c'(0)$

| | $n_u$ | $n_o$ | |
|---|---|---|---|
| $a < (c')^{-1}(p)$ | 1 | 0 | ← Orange |
| $(c')^{-1}(p) < a < \bar{a}(p,v)$ | 2 | 1 | ← White |
| $\bar{a}(p,v) < a$ | 0 | 1 | ← Yellow |

(ignoring boundary behavior; $a = (c')^{-1}(p)$ iff $n_c = 1$)




$c(\mu) = \mu^3 + \mu$ and $v = 10$

**Takeaway**: The system manager must either staff enough servers or pay them enough to ensure equilibrium existence.

CHICAGO BOOTH
The University of Chicago Booth School of Business

# Asymptotic Analysis: Multiplicity

**Lemma (Equilibrium Selection):**
If $\mu_1^\star$, $\mu_2^\star$ are two distinct limiting equilibria with $\mu_1^\star > \mu_2^\star$, then $U(\mu_1^\star, \mu_1^\star) > U(\mu_2^\star, \mu_2^\star)$.

**Takeaway:** Servers prefer the faster equilibrium. The self-interested behavior of servers is aligned with the system manager's interest in better system performance as measured by less waiting time or larger throughput.

# Asymptotic Analysis: Monotonicity

**Proposition (Monotonicity of $a$):**

**Overloaded equilibria** (solves $p = c'(\mu)$):

$\mu_o^\star(a, p; v)$ does not depend on $a$.

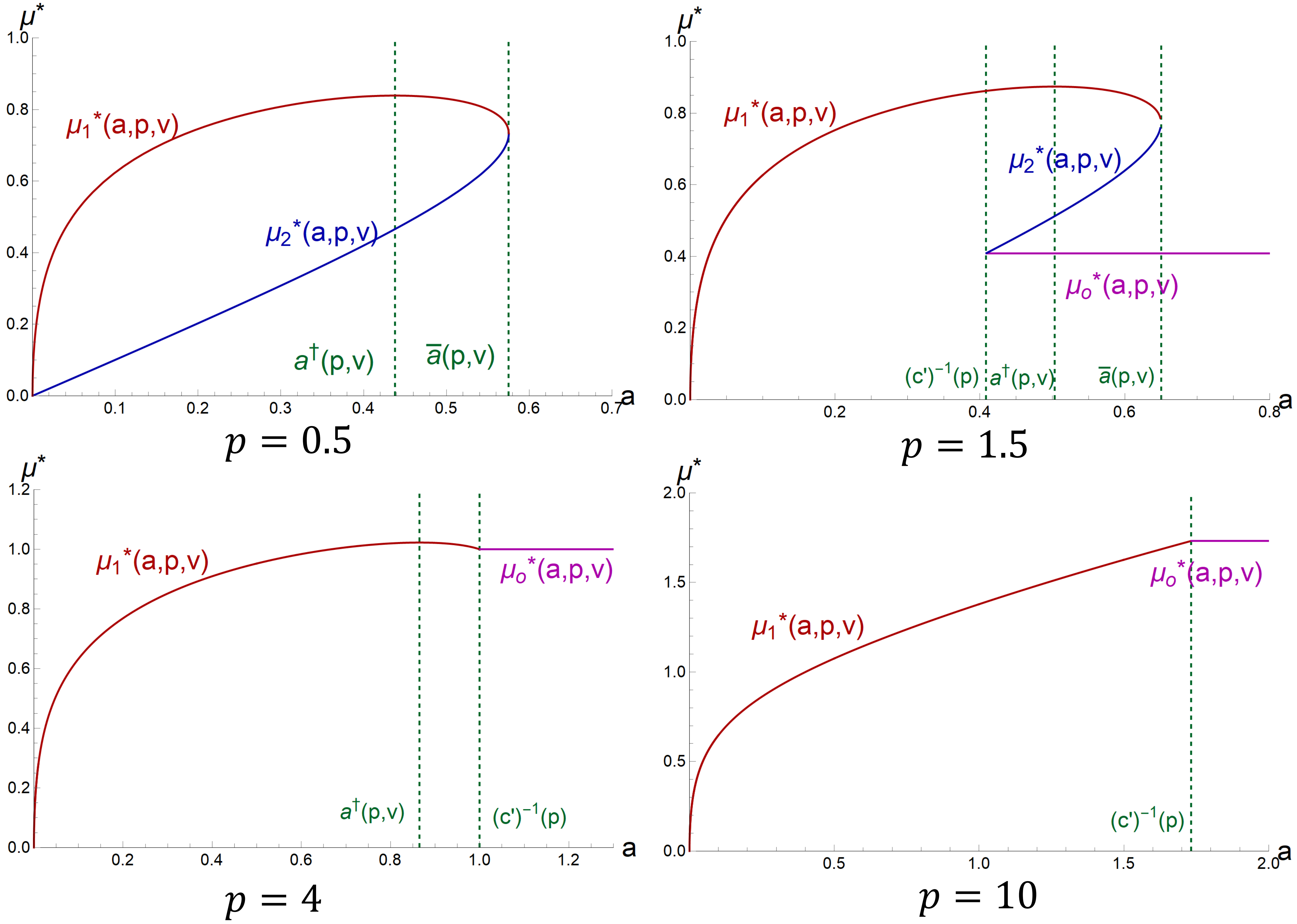
**Underloaded equilibria:**

(i) $\mu_1^\star(a, p; v)$ is strictly increasing in $a$, and then strictly decreasing for smaller $p$; otherwise, $\mu_1^\star(a, p; v)$ is strictly increasing.

(ii) $\mu_2^\star(a, p; v)$ is strictly increasing in $a$.

**Takeaway**: When servers are not paid enough, increasing workload beyond a tipping point may result in a sharp drop in system performance due to server "rebellion".

**Takeaway**: Large enough payment can (1) incentivize servers to work rather than rebel, when the workload is very high; (2) guarantee monotonic increasing behavior.




$c(\mu) = \mu^3 + \mu$

# Asymptotic Analysis: Monotonicity

**Proposition (Monotonicity of $p$):**

**Overloaded equilibria** (solves $p = c'(\mu)$)**:**

$\mu_o^\star(a, p, v)$ is strictly increasing in $p$.

**Underloaded equilibria:**

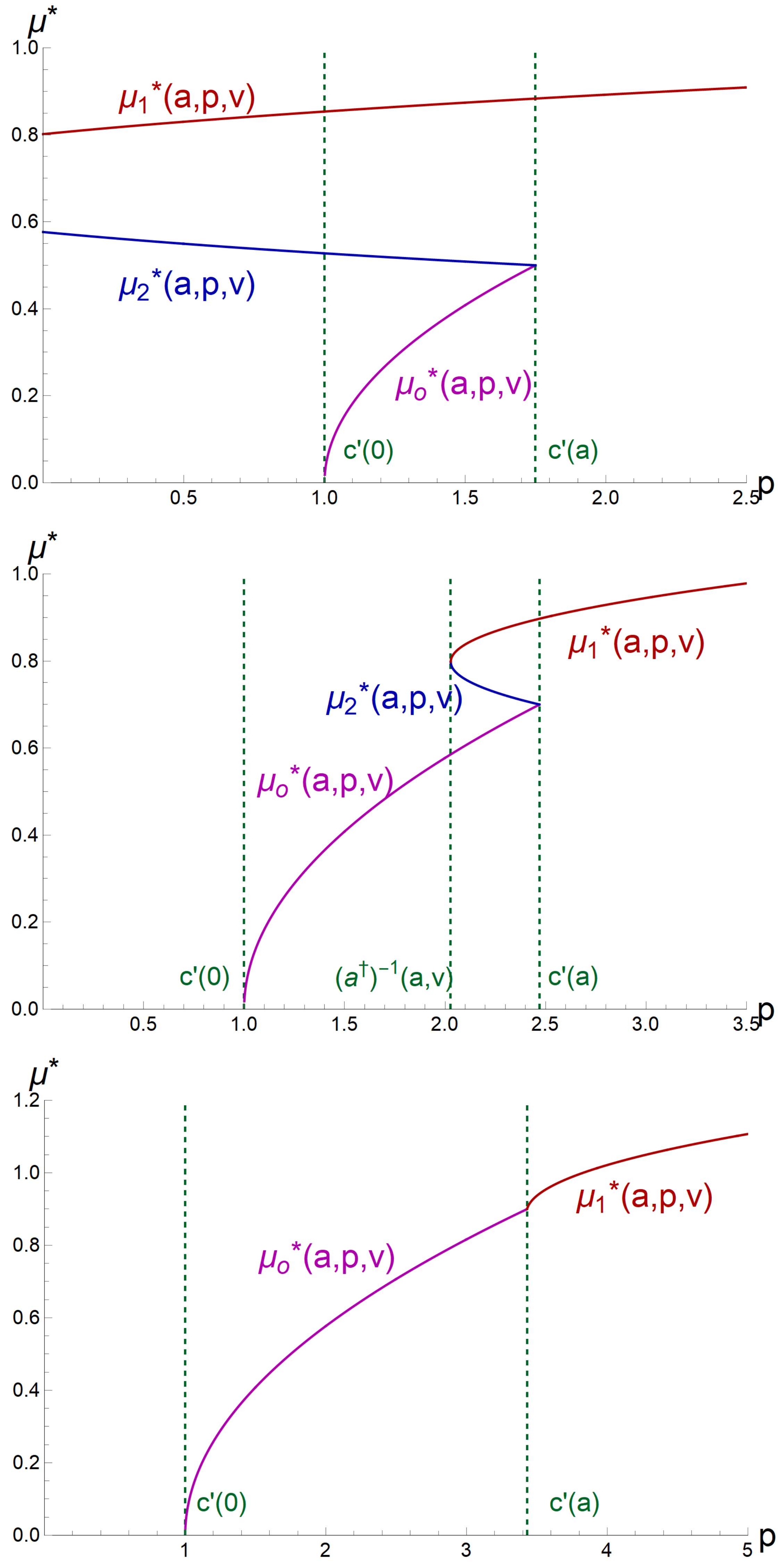
$\mu_1^\star(a, p, v)$ is strictly increasing in $p$,
and $\mu_2^\star(a, p, v)$ is strictly decreasing in $p$.

**Takeaway:** Higher wage could cause servers to speed up, and could cause servers to slow down.

$a = 0.5$

$a = 0.7$

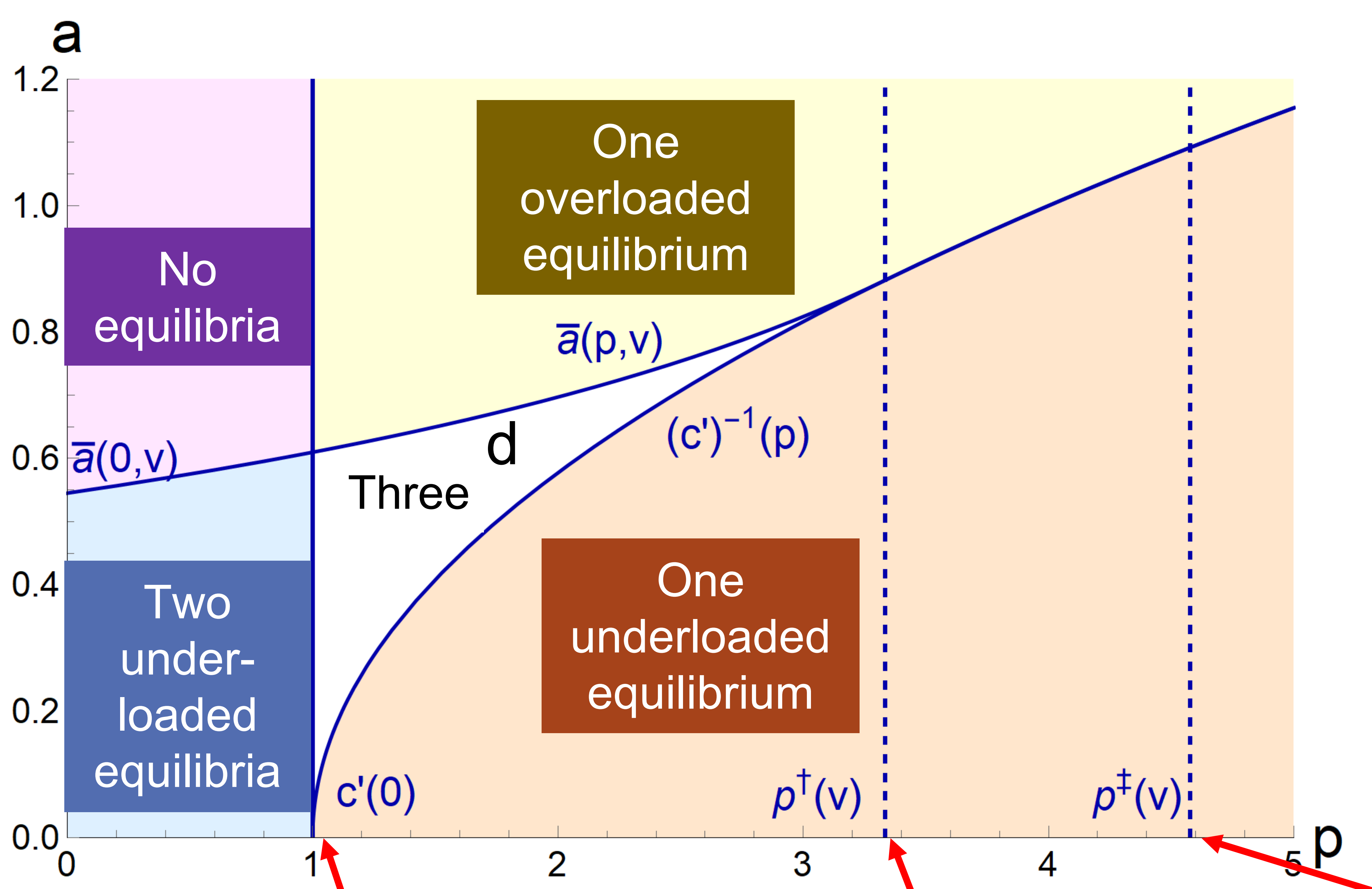
$a = 0.9$




$c(\mu) = \mu^3 + \mu$

# Asymptotic Analysis: Summary

$c(\mu) = \mu^3 + \mu$ and $v = 10$




$p > c'(0)$
guarantee existence

$p > p^{\dagger}(v)$
guarantee uniqueness

$p > p^{\ddagger}(v)$
guarantee monotonicity

**Takeaway:** More desirable equilibrium properties for all staffing levels come at increasing costs to the manager.

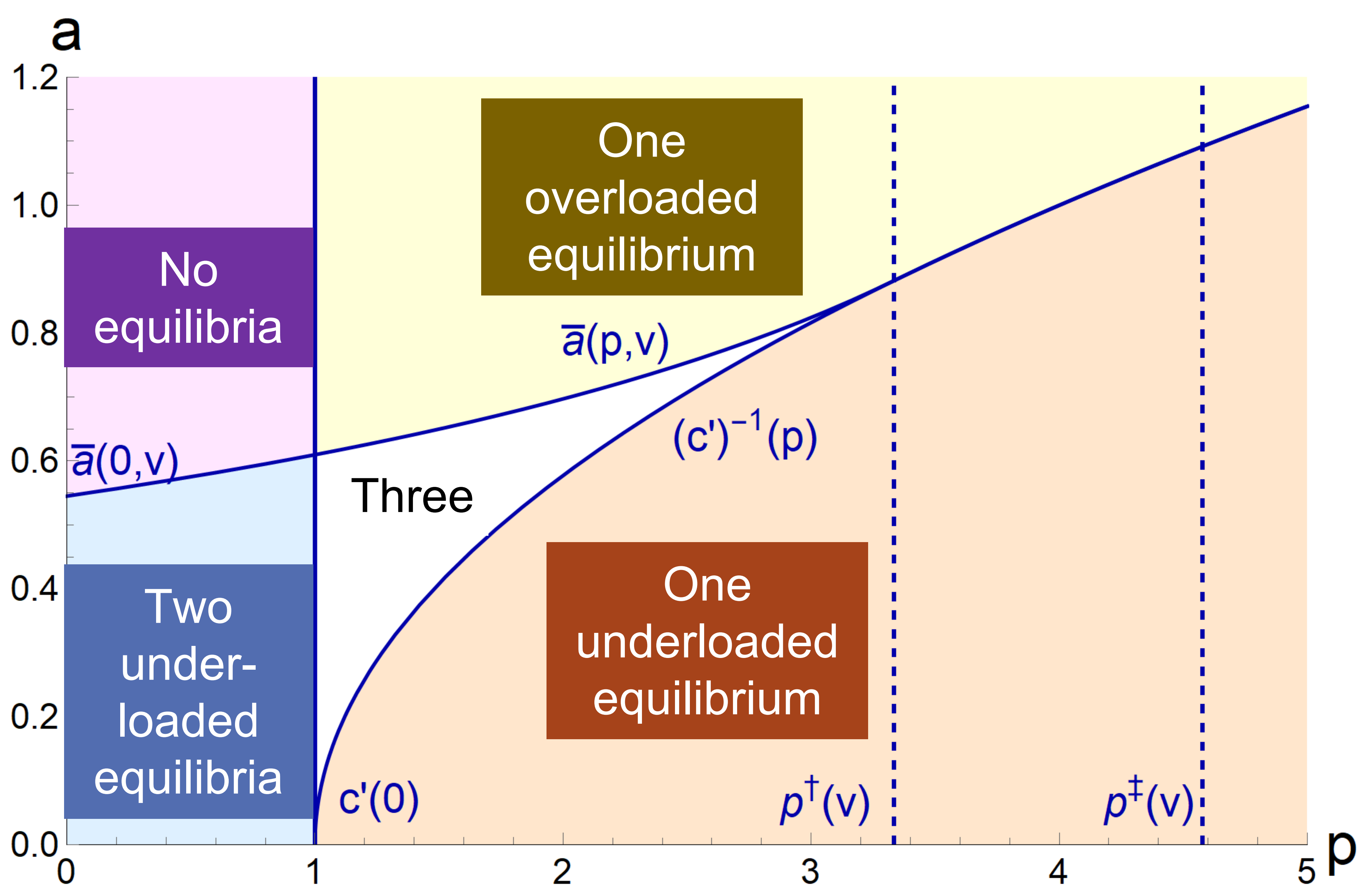
# Asymptotic Analysis: Convergence Theorem

**Theorem:** The following holds for all large enough $\lambda$:

- If $n_o = 0$, then $n_o^\lambda = 0$.
- If $n_u = 0$, then $n_u^\lambda = 0$.

- If $n_o = 1$, then $n_o^\lambda \geq 1$.
- If $n_u = 2$, then $n_u^\lambda \geq 2$.
- If $n_u = 1$, then $n_u^\lambda \geq 1$.

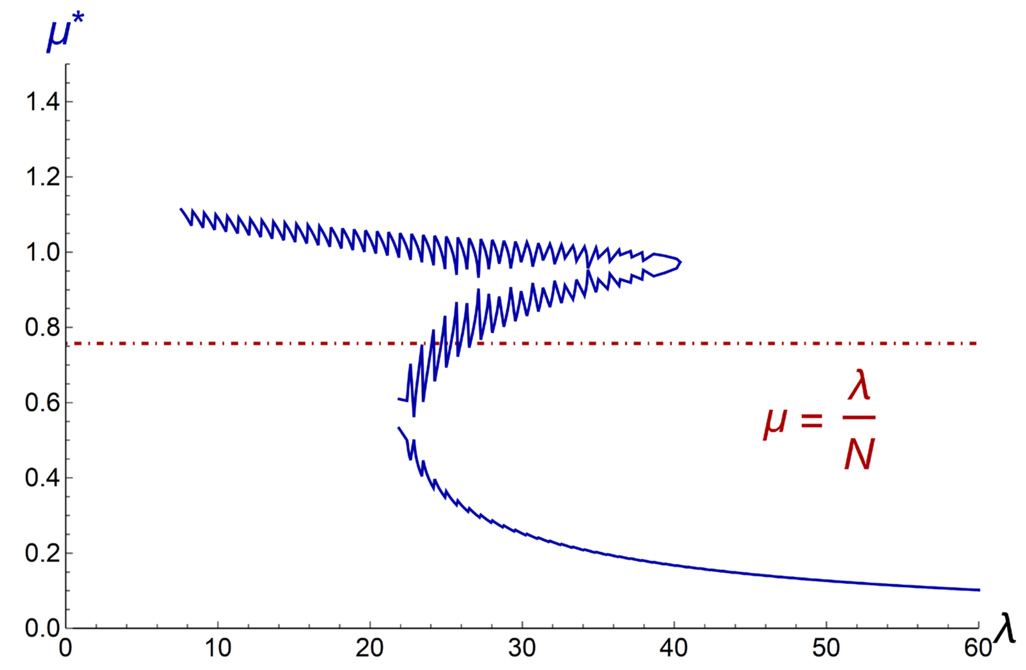
Furthermore, for any limit equilibria, there exists a prelimit equilibria that is close.



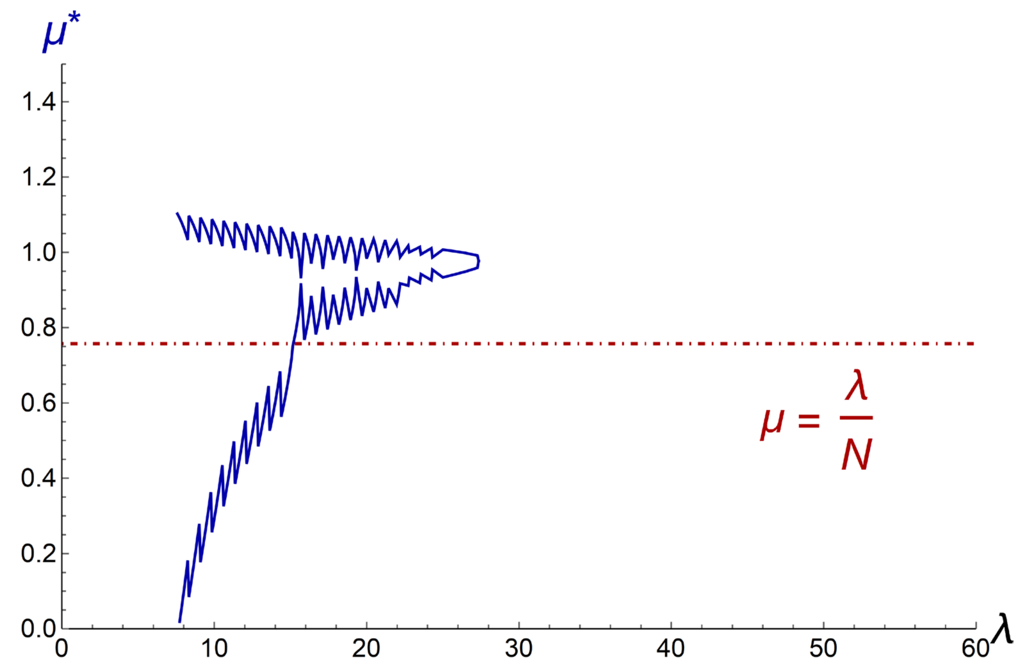
# Asymptotic Analysis: Convergence Theorem

Case: No limiting underloaded equilibrium.

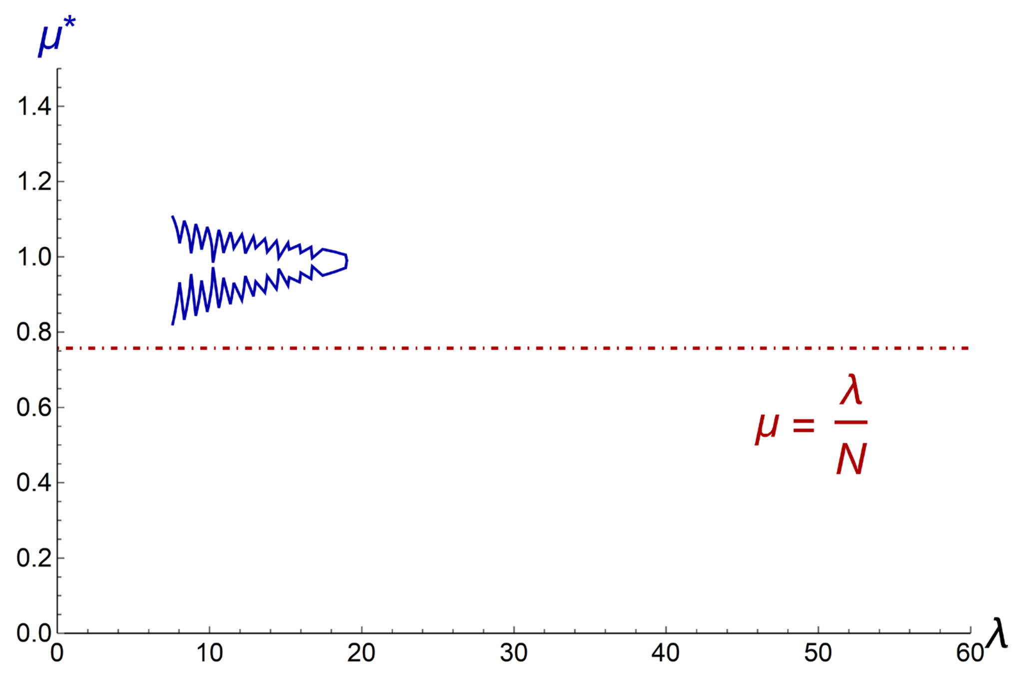
$$\boldsymbol{p = 0.5, v = 10, c(\mu) = \mu^3 + \mu}$$




$N^\lambda = \lfloor 1.4\lambda \rfloor, k^\lambda = \lfloor 1.01 N^\lambda \rfloor$

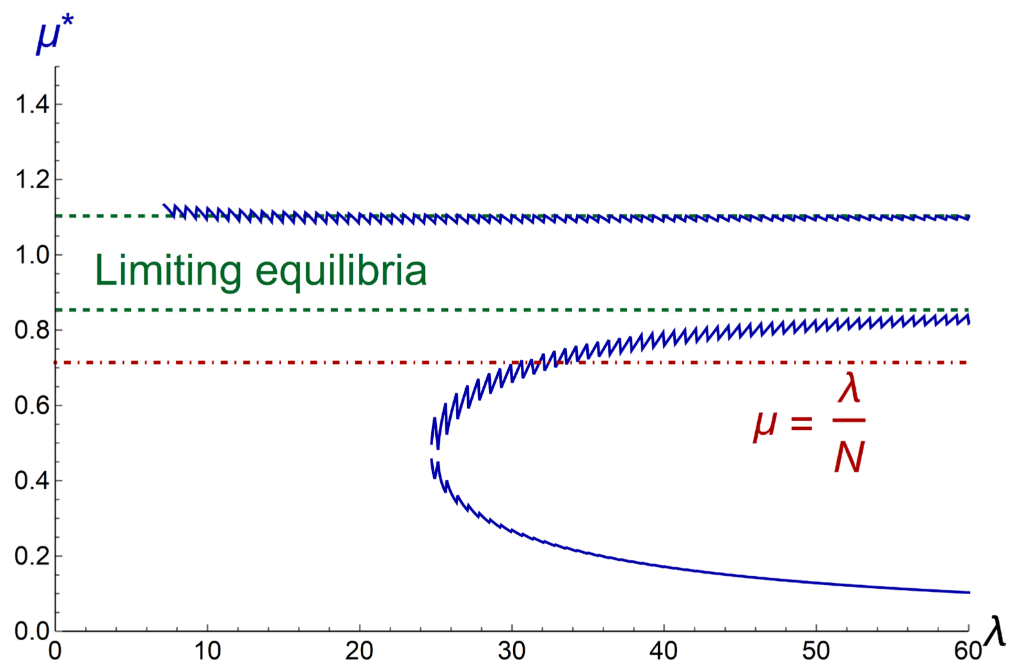



$N^\lambda = \lfloor 1.4\lambda \rfloor, k^\lambda = \lfloor 1.1 N^\lambda \rfloor$

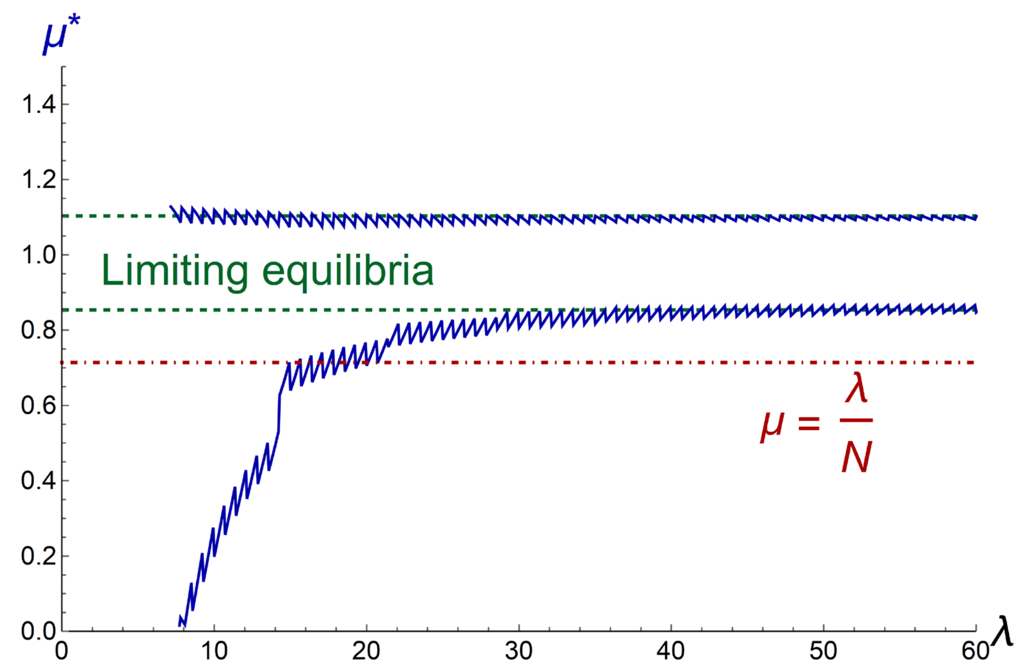



$N^\lambda = \lfloor 1.4\lambda \rfloor, k^\lambda = \lfloor 3 N^\lambda \rfloor$

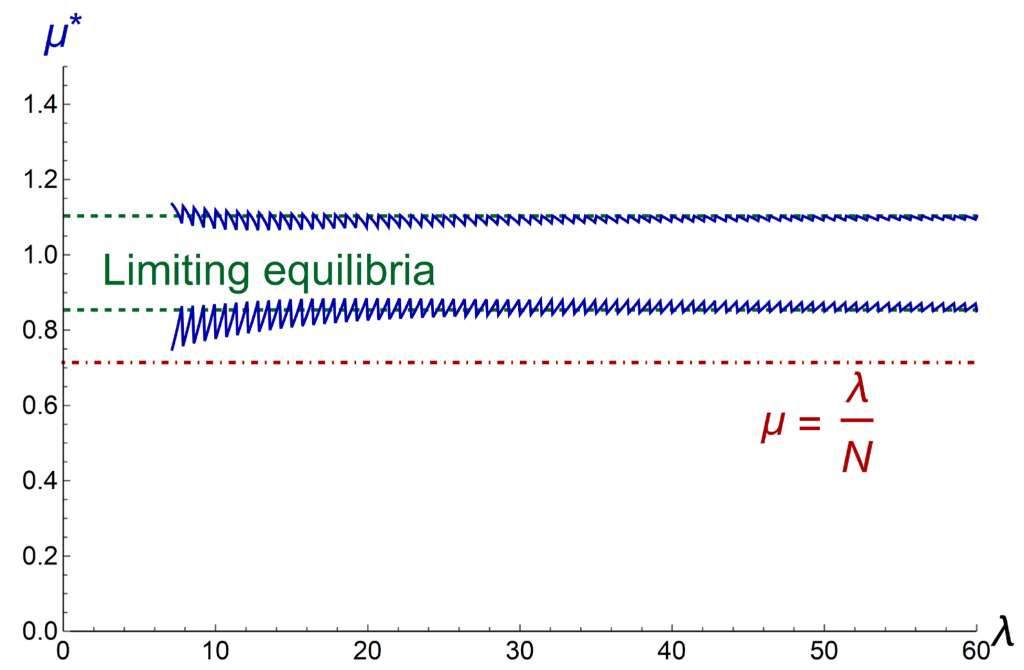
Case: Two underloaded limiting equilibria.




$N^\lambda = \lfloor 1.32\lambda \rfloor, k^\lambda = \lfloor 1.01 N^\lambda \rfloor$




$N^\lambda = \lfloor 1.32\lambda \rfloor, k^\lambda = \lfloor 1.1 N^\lambda \rfloor$




$N^\lambda = \lfloor 1.32\lambda \rfloor, k^\lambda = \lfloor 3 N^\lambda \rfloor$

**Limiting FOC:** $p\left(1 - \left[1 - \frac{a}{\mu}\right]^+\right) + (v - p\mu)\frac{a[\mu - a]^+}{\mu^3} = c'(\mu)$ $\boxed{\boldsymbol{N^\lambda = \frac{1}{a}\lambda + o(\lambda)}}$

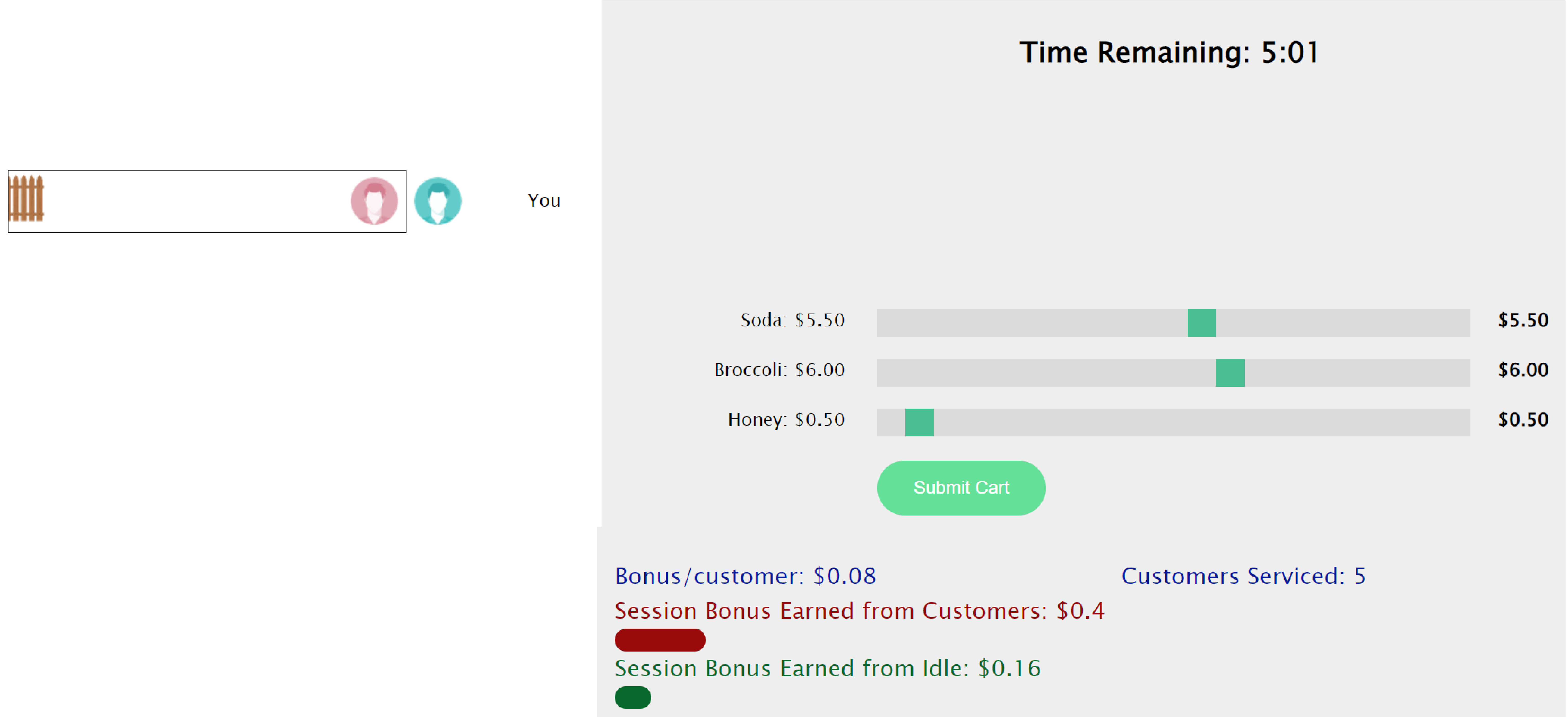
# Outline

- Model
- Asymptotic Analysis
- Looking Ahead

# Ongoing Work: Experiment to test hypotheses

**H1**: The servers work speed is increasing in payment.

**H2**: The servers work speed is first increasing and then decreasing, as the workload grows.




CHICAGO BOOTH

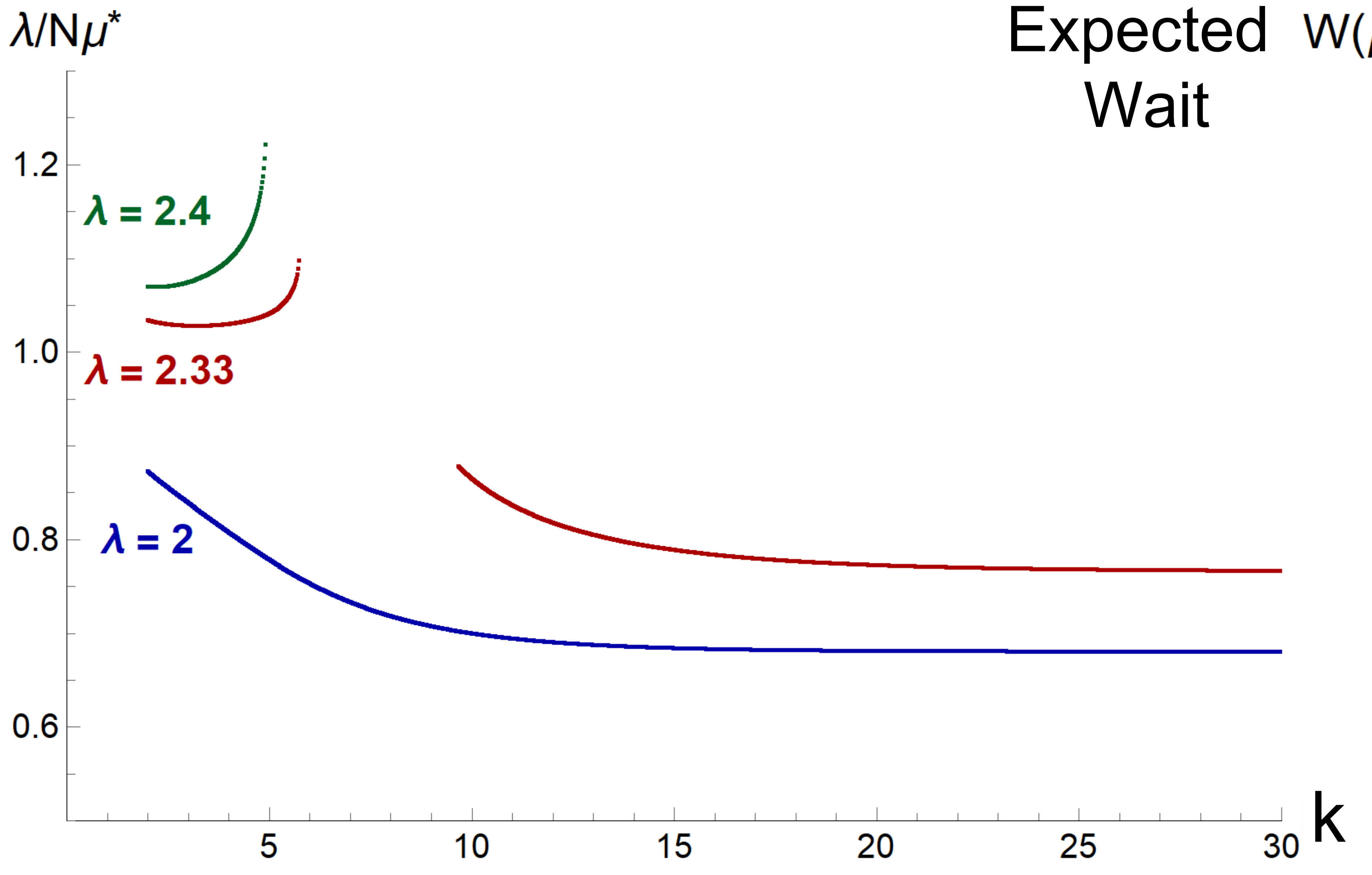
The University of Chicago Booth School of Business

# Ongoing Work: Consequences for system design

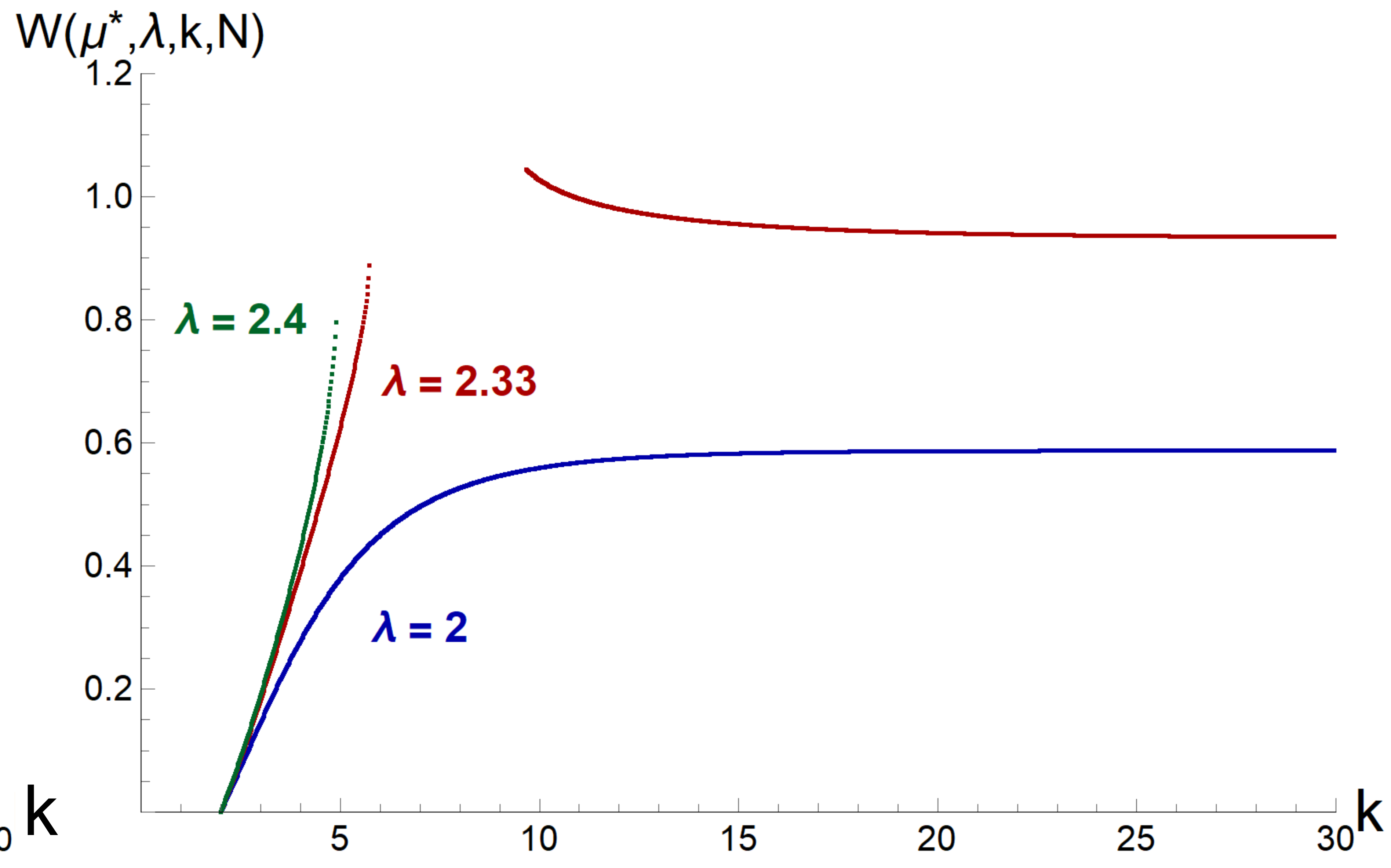
- Admission control, social welfare optimization, staffing optimization
- Queueing models that endogenize server behavior require revisiting system design “rules of thumb”.

  Ex: Does decreased load lead to decreased waiting?

System Load




Expected Wait




$$N = 2, p = 0$$
$$v = 1, c(\mu) = \frac{3}{32}\mu^2$$

# Ongoing Work: Strategic Arrival and Server Interactions

**The Customer Side:**

- A customer who arrives to find $i$ customers in the system joins if and only if $\overset{\text{Value for service}}{R} - \overset{\text{Cost of waiting}}{C} \underbrace{\left( \frac{i - N + 1}{N\mu} + \frac{1}{\mu} \right)}_{\text{Expected time in system}} \geq 0.$

- Customer equilibrium strategy is to join if and only if there are no more than $k^* - 1$ customers in the system, where $k^* = \left\lfloor \frac{RN\mu}{C} \right\rfloor$.

**Open Question: Joint Equilibrium** † **:**

$(k^*, \mu^*)$ is a Nash equilibrium if and only if $\bullet\ k^* = \left\lfloor \frac{RN\mu^*}{C} \right\rfloor \geq N$, and, $\bullet\ \mu^* \in \arg\max_{\mu_i \geq 0} U_i(\mu_i, \mu^*; \lambda, N, k^*)$.

† Only such paper is Chung, Ahn, and Righter (2020), which is restricted to $N = 1$ setting.

This paper provides foundation to study such interaction where customers' decisions endogenously induce a finite buffer.

# Summary

- Analytically studied some nontraditional server behavior documented in empirical works.
- We studied a many-strategic-server finite-buffer ($M/M/N/k$) queueing system.
- Asymptotic analysis allows us to characterize when equilibria exist, and to analyze their behavior, for large arrival rate.
- Equilibria may or may not exist, and may not be monotonic, which has consequences for system design.

# Thanks!
# Q&A

yzhong0@chicagobooth.edu