Scheduling Impatient Customers in a Multiclass Many-Server Queue

Performance 2020

38th International Symposium on Computer Performance, Modeling, Measurements and Evaluation

Amy R. Ward

*joint work with Amber Puha

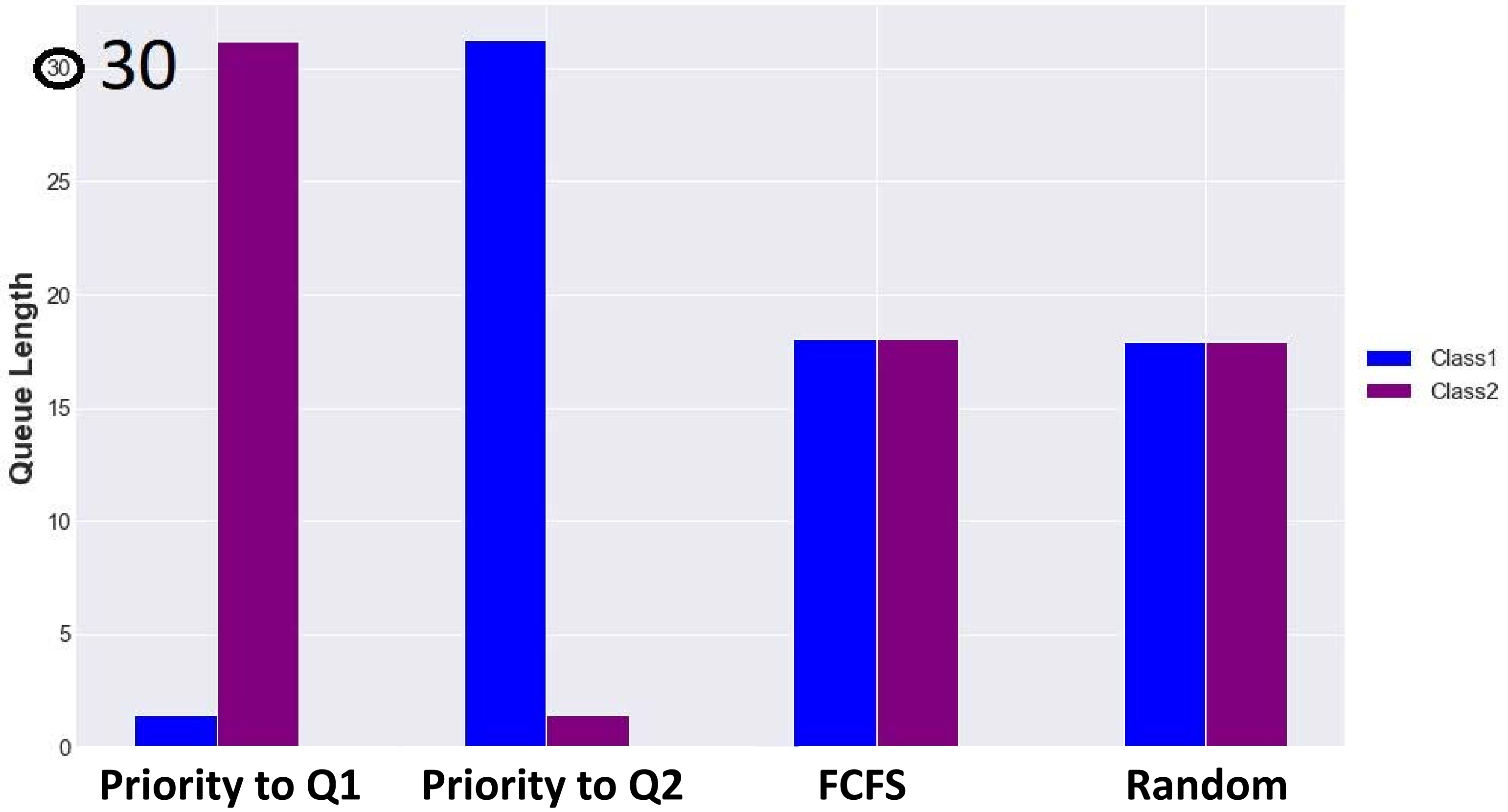
Paper link: https://faculty.chicagobooth.edu/Amy.Ward/papers/Puha-Ward-Weak-Convergence-2019.pdf

---

CHICAGO BOOTH

The University of Chicago Booth School of Business

Why Do We Care About Scheduling?



An Early (Deterministic) Scheduling Problem

Processing time V_1
Processing rate $\mu_1 = 1/V_1$
Waiting cost c_1

Processing time V_2
Processing rate $\mu_2 = 1/V_2$
Waiting cost c_2

Objective: Determine the schedule that minimizes the waiting cost.

Schedule	Cost
(1,2)	$c_2 V_1$
(2,1)	$c_1 V_2$

We prefer the schedule (1,2) iff $c_2 V_1 < c_1 V_2$, or, equivalently, $c_2\mu_2 < c_1\mu_1$.

The very appealing $c\mu$ rule:
Order classes and give priority in that order.

Reference: Smith (1956).

The $c\mu$ Rule

Optimal

- In the Stochastic Setting; Pinedo (1983)
- When there are due dates; Smith (1956) and Pinedo (1983)
- Multiclass (mc) M/G/1 with feedback; Klimov (1974)

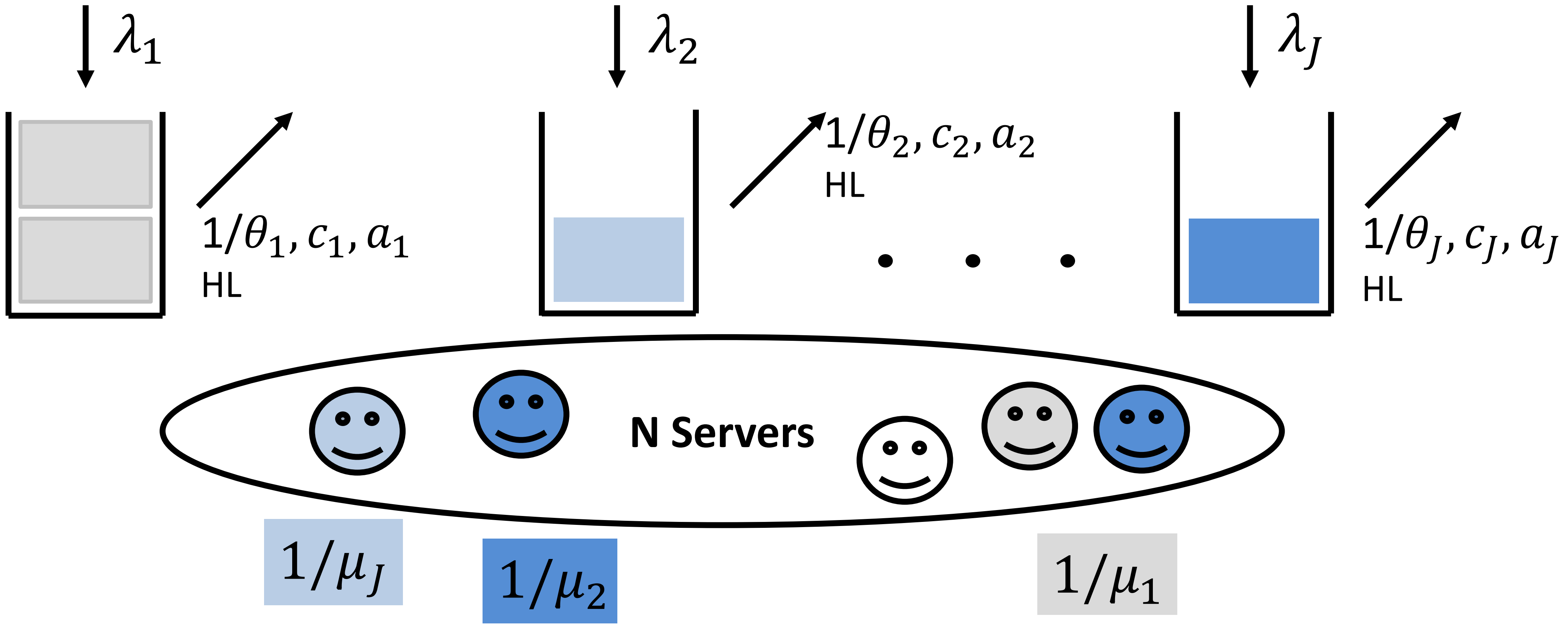
Asymptotically Optimal

- Convex delay costs in mc G/G/1; van Mieghem (1995)
- Heterogeneous servers in mc G/G/N; Mandelbaum and Stoylar (2004)
- Server pools in mc G/M/N; Gurvich and Whitt (2009)

Q: What happens when jobs will not wait forever?

We will study this question in a many-server queue with abandonment.

The $c\mu$ Rule when Jobs may Abandon

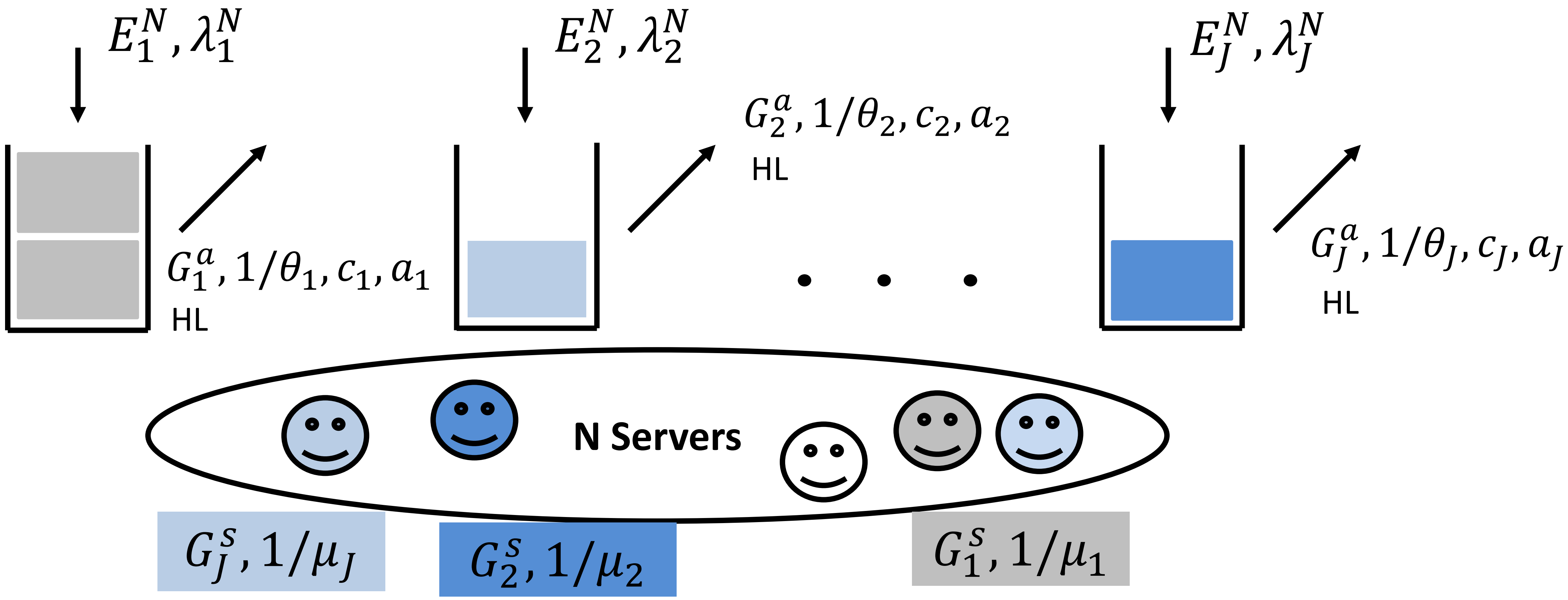


Atar, Giat, Shimkin (2010) The $\tilde{c}_j\mu_j/\theta_j$ rule asymptotically minimizes long-run average cost in a M/M/N+M queue the overloaded regime $(\tilde{c}_j = c_j + \theta_j a_j)$.

The $c\mu$ and $\tilde{c}\mu/\theta$ rules is a **static priority (SP) scheduling policy** in the sense that the decision of who to next serve does not depend on system state.

Our Research Question: What is an asymptotically optimal policy?

(Is SP still asymptotically optimal?)



Assume overloaded regime.

Admissible scheduling policy is HL, non-anticipating, does not have wild oscillations.

- Connection to learning: Should be easier if an ao policy has a simple form.

Any questions?

Our Approach

Operating under an Unspecified Admissible Scheduling Policy

Multiclass G/GI/N+GI Queue

N→ ∞

LLN Scaling

PW 2020, Theorems 1 and 2

Fluid Model

Functional Law of Large Numbers Approximation to the Stochastic System

Proposed Policy Class

PW 2020, Theorem 3

G/GI/N+GI Queue Steady-State

N→ ∞

LLN Scaling

Fluid Invariant States

Fixed Point Solutions

Fluid Control Problem

Talk Outline

1. Provide a fluid model relevant for a large class of non-preemptive HL scheduling policies.
   - Show limit points of scaled state processes are fluid model solutions (PW, Theorem 1).
   - Establish tightness (PW, Theorem 2).

2. Formulate and solve a fluid control problem for an overloaded system.
   - Characterize fluid invariant states (PW, Proposition 1).
   - Provide weak convergence theorem for appropriate scheduling policy (PW, Theorem 3).

Some Works Related to Step 1

(Provide fluid model relevant for a large class of HL scheduling policies.)

- Single Class Fluid Model for G/GI/N and G/GI/N+GI.
  - Whitt (2006) proposed a Fluid Model.
  - Liu and Whitt (2012) calculate fluid performance measures.
  - Reed (2009) and Kaspi and Ramanan (2011) proved convergence, without abandonment.
  - Kang and Ramanan (2010 and 2012) proved convergence, with abandonment.
  - Provided the framework for approaching the multiclass case.
- Multiclass Scheduling.
  - Atar, Kaspi and Shimkin (2014) analyzed SP for multiclass G/GI/N+GI, and show asymptotic optimality of SP for G/GI/N**+M.**
  - We generalize to a large class of admissible policies that include SP.

The State Space

Time elapsed since last class j arrival.

The number of class j customers in the system.

$$(\alpha^N, x^N, \nu^N, \eta^N).$$

Measure-valued processes tracking the age-in-service and the potential queue.

Primitive inputs:

- Arrival counting processes for each class;
- Sequence of i.i.d. service times for each class;
- Sequence of i.i.d. deadlines for each class.

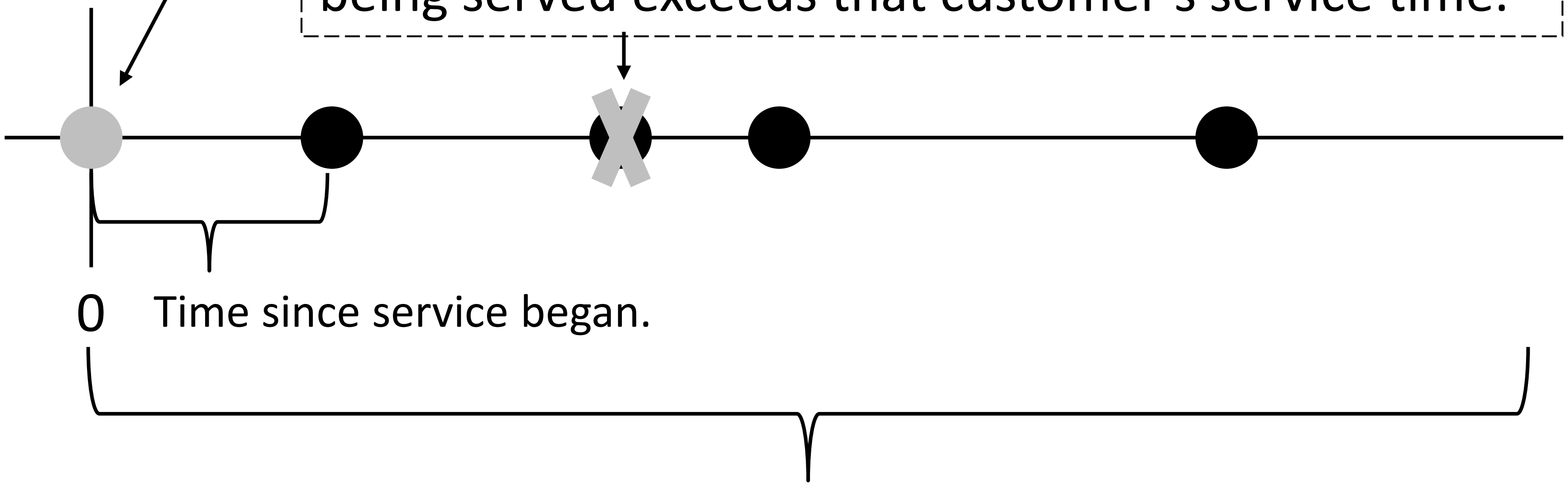
The ν Measure (for given Class j)

Note: Depends on Scheduling Control.

$$\langle 1, \nu_j(t) \rangle = \int_0^\infty \nu_j(t)(dx) = \cancel{5}\,4 \quad \text{and} \quad \sum_{j=1}^{N} \langle 1, \nu_j(t) \rangle \leq N$$

Customer entering service has age 0.

Customer is no longer tracked once the time spent being served exceeds that customer's service time.



Each dot is a unit atom whose position represents the time elapsed since a customer began service, and shifts to the right at rate 1.

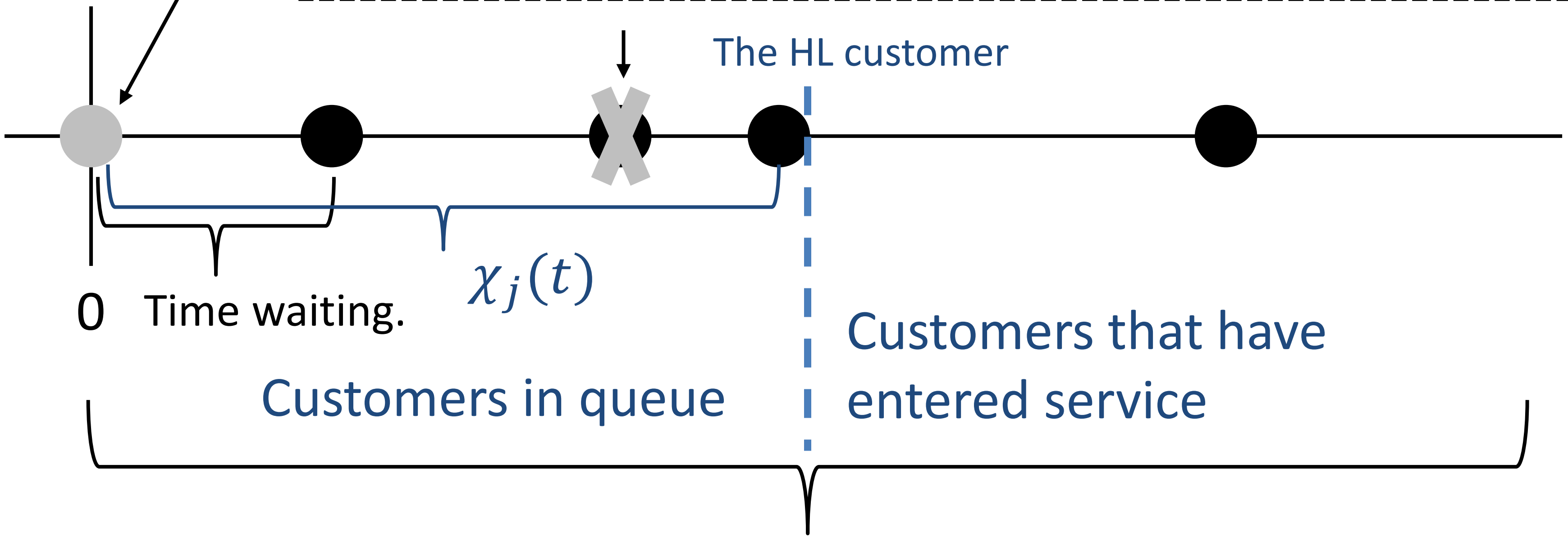
The η Measure (for given Class j)

Note: Independent of Scheduling Control.

$\langle 1, \eta_j(t) \rangle = \cancel{5}\,4 \geq Q_j(t)$ and $\left\langle 1_{[0,\chi_j(t)]}, \eta_j(t) \right\rangle = Q_j(t)$

Customer entering system has waited 0 time units.

Customer is no longer tracked once the time elapsed since arrival exceeds that customer's abandonment time.



Each dot is a unit atom whose position represents the time elapsed since a customer arrival, and shifts to the right at rate 1 until its potential abandonment time.

The Fluid Model State Space and Auxiliary Functions

Number of fluid in system

Age-in-service measure

Potential queue measure

For (X, ν, η), define for all j and $t \geq 0$,

$$B_j(t) := \langle 1, \nu_j(t) \rangle$$ (Proportion of class j fluid in service);

Service distribution hazard rate

$$\delta_j(u) := \langle h_j^s, \nu_j(u) \rangle = \int_0^\infty h_j^s(x) \nu_j(u)(dx)$$ (Instantaneous departure rate);

$$D_j(t) := \int_0^t \delta_j(u) du$$ (Cumulative departure process);

$$Q_j(t) := X_j(t) - B_j(t)$$ (Queue-length process);

$$\chi_j(t) := \inf\{x \geq 0 : \langle 1_{[0,x]}, \eta_j(t) \rangle \geq Q_j(t)\}$$ (Class *j* head-of-line wait time process);

$$R_j(t) := \int_0^t \left\langle 1_{[0,\chi_j(u)]} h_j^a, \eta_j(u) \right\rangle du$$ (Cumulative abandonment process);

Abandonment distribution hazard rate

$$K_j(t) := B_j(t) + D_j(t) - B_j(0)$$ (Cumulative entry-into-service process).

A Fluid Model Solution (Not Unique)

Non-negative, continuous, and non-decreasing J-dimensional function having domain $\Re_+$; for example, $E_j(t) = \lambda_j t$ for all j.

Let E be an arrival function. Then, (X, ν, η) is a fluid model solution for E if the following hold.

(1) For each j, K_j is non-decreasing and $\sum_{j=1}^{J} B_j(t) \in [0,1]$ for all $t \geq 0$.

(No service rule specified.)

(2) For all j and $t \geq 0$, $X_j(t) = X_j(0) + E_j(t) - R_j(t) - D_j(t)$, $0 \leq Q_j(t) \leq \int_0^\infty \eta_j(dy)$, and finiteness.

(3) For all $j, f \in C_b([0,\infty))$, and $t \geq 0$,

Service ccdf.

$$\langle f, \nu_j(t) \rangle = \left\langle f(\cdot + t) \frac{\bar{G}_j^s(\cdot + t)}{\bar{G}_j^s(\cdot)}, \nu_j(0) \right\rangle + \int_0^t f(t-u) \bar{G}_j^s(t-u) dK_j(u)$$

$$\langle f, \eta_j(t) \rangle = \left\langle f(\cdot + t) \frac{\bar{G}_j^a(\cdot + t)}{\bar{G}_j^a(\cdot)}, \eta_j(0) \right\rangle + \int_0^t f(t-u) \bar{G}_j^a(t-u) dE_j(u).$$

Abandonment ccdf.

Non-Policy Specific Convergence

Assume

- $\lim_{N\to\infty} \frac{E^N}{N} = E$ almost surely, $\lim_{N\to\infty} \mathbb{E}\left[\frac{E_j^N(t)}{N}\right] = \mathbb{E}[E_j(t)]$, for all $t \geq 0$, and E is continuous;

- Entry-into-service process oscillations can be controlled;
- Convergence of initial conditions and “goodness” of initial fluid state;
- Hazard rates of abandonment and service time distributions are either bounded or lower semi-continuous;

PW 2020 Theorem 1

Suppose that (X, ν, η) is a distributional limit point of a sequence of fluid-scaled state processes. Then, (X, ν, η) is almost surely a fluid model solution for E.

PW 2020 Theorem 2

A sequence of fluid-scaled state processes is tight.

Talk Outline

1. ~~Provide a fluid model relevant for a large class of non-preemptive HL scheduling policies.~~
   - ~~Show limit points of scaled state processes are fluid model solutions (PW, Theorem 1).~~
   - ~~Establish tightness (PW, Theorem 2).~~

$E_j(t) = \lambda_j t$ for $t \geq 0$ and all j and $\sum_{j=1}^{J} \frac{\lambda_j}{\mu_j} \geq 1.$

2. Formulate and solve a fluid control problem for an overloaded system.
   - Characterize fluid invariant states (PW, Proposition 1).
   - Provide weak convergence theorem for appropriate scheduling policy (PW, Theorem 3).

Any questions?

Fluid Model Invariant States

Definition (Feasible server effort allocation).

- $\boldsymbol{B} = \left\{ b \in \Re_+^J : b_j \leq \lambda_j / \mu_j, \sum_{j=1}^J b_j \leq 1 \right\}$

PW 2020 Proposition 1

For each $b \in \boldsymbol{B}$, there exists an invariant state such that b_j is the proportion of server effort devoted to class j, and

$$Q_j(t) = \lambda_j \frac{1}{\theta_j} f_j\left(\frac{\lambda_j - b_j \mu_j}{\lambda_j}\right) \text{ for all } t \geq 0, \text{ where } f_j(x) = G_{e,j}^a\left(\left(G_j^a\right)^{-1}(x)\right).$$

Mean patience time.

Abandonment stationary excess cdf.

Abandonment cdf.

Intuition: If exponential abandonment distribution, then

$$\frac{\lambda_j}{\theta_j} f_j\left(\frac{\lambda_j - b_j \mu_j}{\lambda_j}\right) = \frac{1}{\theta_j}\left(\lambda_j - b_j \mu_j\right) = \mathrm{q_j}.$$

Flow balance implies $\lambda_j - b_j \mu_j = \theta_j q_j$.

Fluid Model Invariant States

Definition (Feasible server effort allocation).

- $\boldsymbol{B} = \left\{ b \in \Re_+^J : b_j \leq \lambda_j / \mu_j, \sum_{j=1}^J b_j \leq 1 \right\}$

PW 2020 Proposition 1

For each $b \in \boldsymbol{B}$, there exists an invariant state such that b_j is the proportion of server effort devoted to class j, and

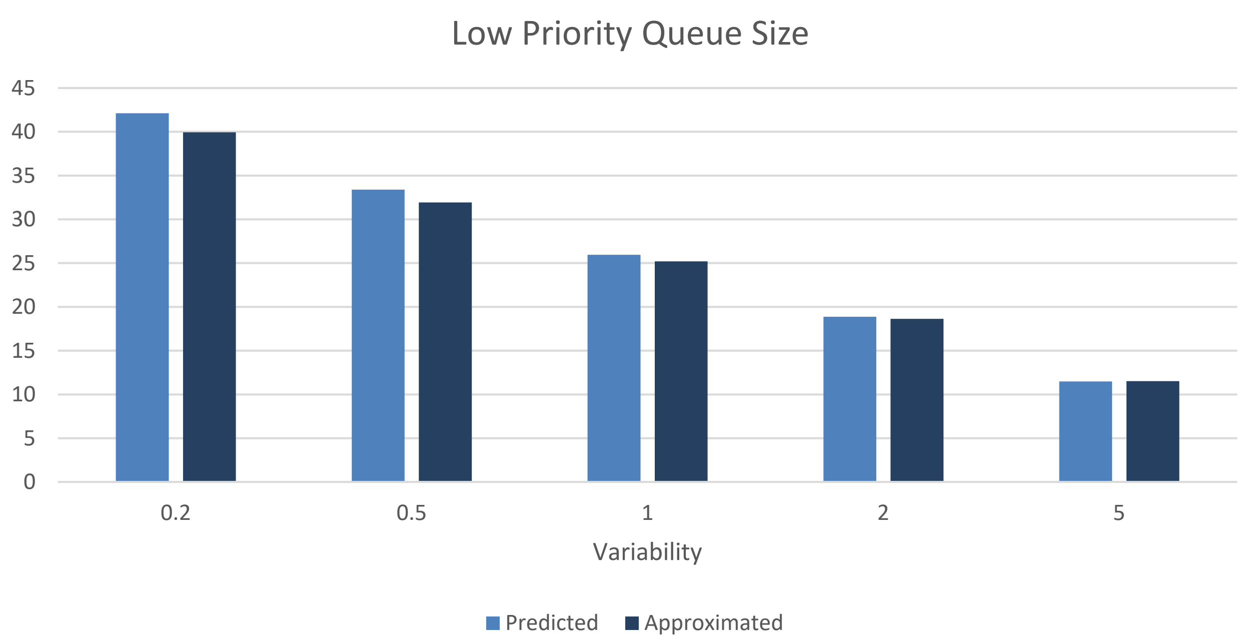
$$Q_j(t) = \lambda_j \underbrace{\frac{1}{\theta_j}}_{\text{Mean patience time.}} f_j\left(\frac{\lambda_j - b_j \mu_j}{\lambda_j}\right) \text{ for all } t \geq 0, \text{ where } f_j(x) = \underbrace{G_{e,j}^a}_{\text{Abandonment stationary excess cdf.}}\left(\left(\underbrace{G_j^a}_{\text{Abandonment cdf.}}\right)^{-1}(x)\right).$$

How good is using the function f_j to approximate the class j mean steady-state queue-length?

Performance Measure Approximation

Assume Static Priority Scheduling.

A two-class $M/LN(1{,}4)/100 + LN(1, v)$ queue, with each class having arrival rate 60 per hour.



(High priority queue has predicted size 0, and simulated size about 1.5 for all values of the variability *v*.)

Note that queue size decreases as variability increases.

A Fluid Control Problem

$$m^\star = \min_{b \in B} \sum_{j=1}^{J} \underbrace{c_j \frac{\lambda_j}{\theta_j} f_j\left(\frac{\lambda_j - b_j \mu_j}{\lambda_j}\right)}_{\text{Queue}} + \underbrace{a_j(\lambda_j - b_j \mu_j)}_{\text{Abandonments}}$$

When is the solution consistent with static priority scheduling?

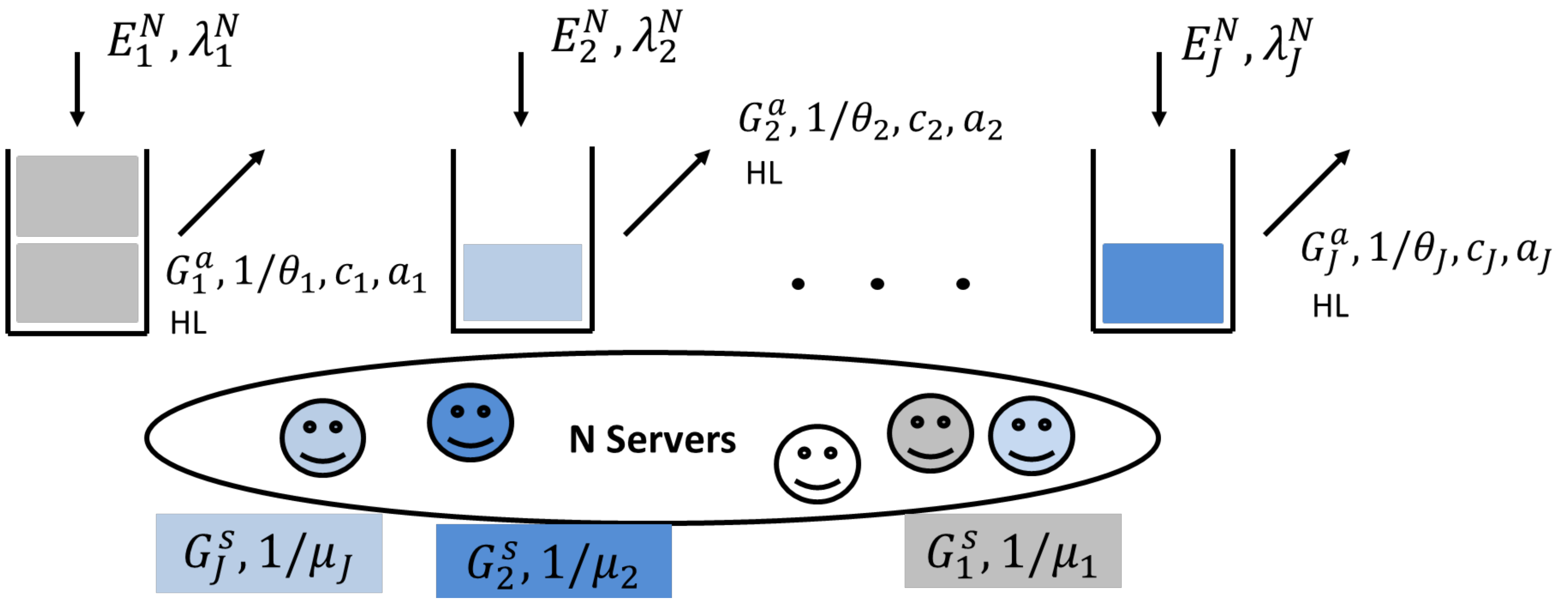
- If there is no holding cost; that is, $c_j = 0$.
  - Digression: Return to the question from earlier in the talk regarding implications for learning.
- If the abandonment distribution has increasing hazard rate (IFR), then
  - f_j is concave, and $m^\star$ is achieved by a feasible vertex.
  - I.E., the solution motivates a static priority policy.
- If the abandonment distribution has decreasing hazard rate (DFR), then
  - f_j is convex, and $m^\star$ could be attained by a non-vertex feasible point.
  - I.E., the solution motivates partially serving classes (not static priority).
  - (We have numeric examples with non-vertex feasible point solution.)

Other Examples when Static Priority Scheduling is not Optimal: Non-Overloaded Regimes

- Exact MDP Analysis
  - Down, Koole and Lewis (2011)
- Single-Server System in Heavy-Traffic Asymptotic Regime
  - Ata and Tongarlak (2013)
  - Kim and Ward (2013)
- Many-Server System in Halfin-Whitt Asymptotic Regime
  - Harrison and Zeevi (2004)
  - Atar, Mandelbaum and Reiman (2004)
  - Kim, Randhawa, and Ward (2018)
- Many-Server System in Overloaded Regime
  - Long, Shimkin, Zhang, and Zhang (2020) for GI/M/N+GI

Remaining Q: How do I schedule so as to achieve b?

Weighted Random Buffer Selection (WRBS) Scheduling



A newly available server next serves class j with probability $p_j > 0$, where $\sum_{j=1}^{J} p_j = 1$.

Policy Specific Convergence

Assume

- $\lim_{N\to\infty} \frac{E^N}{N} = E$ almost surely, $\lim_{N\to\infty} \mathbb{E}\left[\frac{E_j^N(t)}{N}\right] = \mathbb{E}[E_j(t)]$, for all $t \geq 0$, and E is continuous;
- Entry-into-service process oscillations can be controlled;
- Convergence and "goodness" of initial conditions;
- Hazard rates of service distributions are either bounded or lower semi-continuous;
- Hazard rates of abandonment distributions are bounded.

PW 2020 Theorem 5

Suppose that the queue operates under WRBS policy p. Then,

Fluid-scaled state process $\left(\frac{\boldsymbol{X}^N}{\boldsymbol{N}}, \frac{\boldsymbol{\nu}^N}{\boldsymbol{N}}, \frac{\boldsymbol{\eta}^N}{\boldsymbol{N}}\right) \Rightarrow (X, \nu, \eta)$ as $N \to \infty$,

where (X, ν, η) is almost surely a fluid model solution for E that has unique law.

PW 2020 Theorem 4

For any non-idling $b \in \boldsymbol{B}$, the WRBS policy with $p_j = \frac{\mu_j b_j}{\sum_{k=1}^J \mu_k b_k}$ has invariant state defined by b.

and many idling.

To minimize cost asymptotically, use b that solves the fluid control problem.

General Roadmap for Proving Policy Specific Convergence: Application of Theorems 1 and 2

Add policy specific equations to the multiclass G/GI/N+GI queue that uniquely characterize the dynamics.
(Note: The arrival process may have time-varying rates, as is true in many application settings.)

⬇

Add policy specific equations to the fluid model, and prove uniqueness of fluid model solutions*.

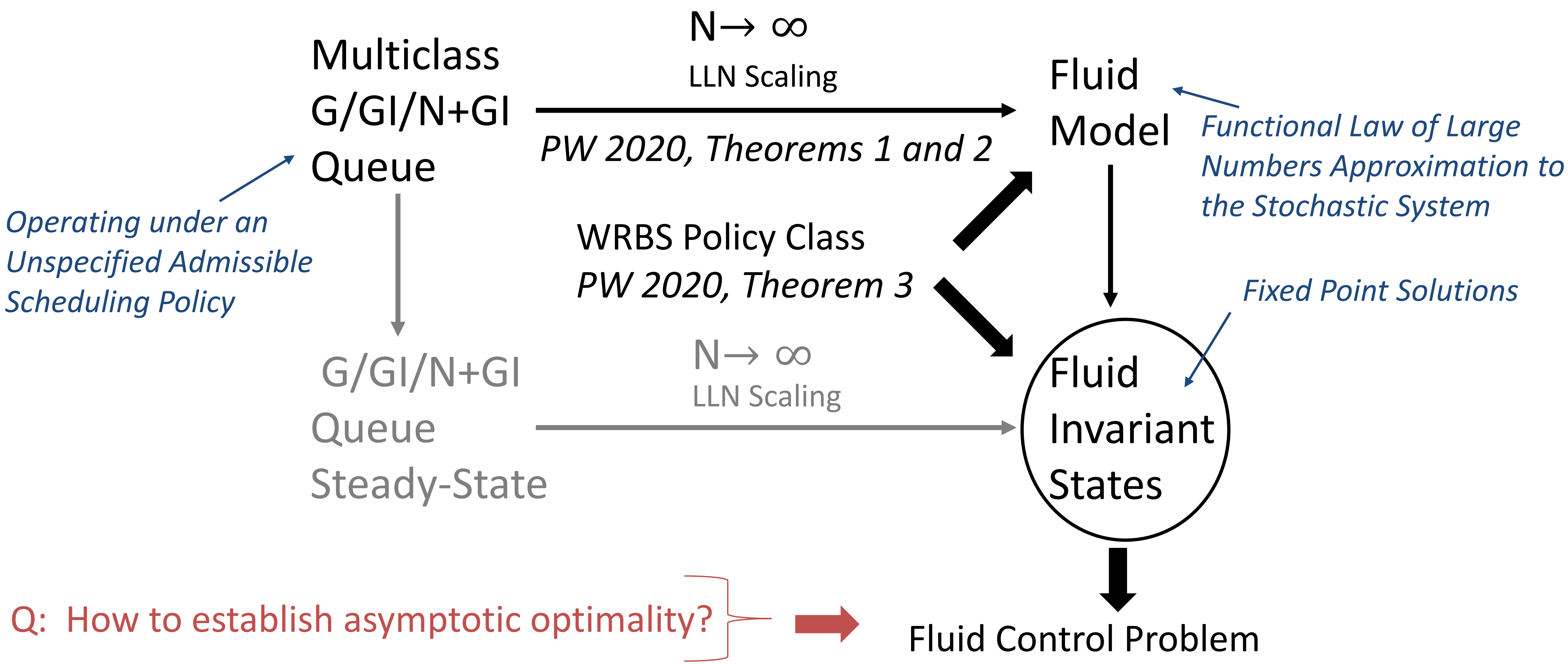
⬇

Prove that fluid limit points of fluid-scaled state processes satisfy the fluid model policy specific equations.
(Theorems 1 and 2 show that all other conditions in the definition of a fluid model solution are satisfied.)

PW 2020 Theorem 3 [The Key to Proving the Weak Convergence on the Previous Slide]
Given a fluid arrival function E, a fluid model solution for WRBS policy *p* is unique for each initial state.

*For a given arrival function and given initial condition.

Summary and Work-in-Progress



Thank you and questions.

Example with Non-Vertex Optima

$$m^\star = \min_{b \in \boldsymbol{B}_J} \sum_{j=1}^{J} \underbrace{c_j \frac{\lambda_j}{\theta_j} f_j\left(1 - \frac{b_j}{\rho_j}\right)}_{\text{Queue}} + \underbrace{a_j(\lambda_j - b_j\mu_j)}_{\text{Abandonments}}$$

Parameters: $\rho_1 = \rho_2 = \mu_1 = \mu_2 = c_1 = c_2 = 1$ and $a_1 = a_2 = 0$.

⬇

Then, $b_2 = 1 - b_1$, and we have a 1-D problem.

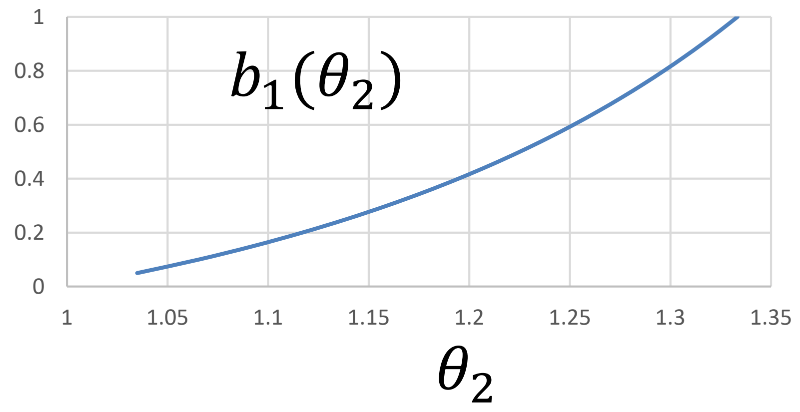
⬇

Patience densities: Class 2 is exponential(θ_2);

Class 1 has density $\frac{2e^{-x}+2e^{-2x}}{3}$ for $x > 0$, which has mean $\frac{5}{6}$.

The minimizer $b_1 \in [0,1]$ satisfies

$$\boxed{\theta_2 = \frac{2}{3b_1}\left(1 + 3b_1 - \sqrt{1 + 3b_1}\right).}$$



(This example is developed by Amber Puha's student Jacques Coulombe.)

WRBS Policy-Specific Fluid Equations

A specified WRBS fluid model solution also satisfies

$$p_j \int_s^t 1\{Q_j(u) > 0\} dD_\Sigma(u) \le K_j(t) - K_j(s) \le p_j \int_s^t dD_\Sigma(u), 1 \le j < J$$

and

Entry into service process.

$$I(t) = \left[I(t) - Q_J(t)\right]^+.$$