A Fluid Limit for an Overloaded Multi-class Many-server Queue with General Reneging Distribution

Amy R. Ward
The University of Chicago Booth School of Business

*Based on current work with Amber Puha.
A Service System Model: The Multiclass Many Server Queue

Call Centers: Garnett, Mandelbaum, Reiman (2002)
Hospital Emergency Department: Green, Soares, Giglio, and Green (2006)
A Service System Model: The Multiclass Many Server Queue

Q: Which class should the available server next serve?
Why is Scheduling Important?

Poisson arrivals, 60 per hour for both classes; 100 Servers; Exponential(1) service times; Exponential(1) patience times.

(Simulation courtesy of Huiyu Wang.)
Specialize to the M/M/N+M Queue

Atar, Giat, Shimkin (2010) The $\tilde{c}_j \mu_j / \theta_j$ rule asymptotically minimizes long-run average cost in the overloaded regime ($\tilde{c}_j = c_j + \theta_j a_j$).
The Need for Non-Static Priority Scheduling Rules

1. Static priority scheduling is not in general optimal.
   - Kim, Randhawa, and Ward (2018) for numerical experiments with non-exponential patience time distribution

2. Static priority scheduling is unfair, which can prevent its adoption.
   - Wierman (2007) for discussion in the context of computer systems
Our Research Objective
(Also serves as Talk Outline.)

*We want to understand the multiclass many server queue with abandonment, without making any distributional assumptions.*

1a. Provide a fluid model relevant for a very general class of scheduling rules.

1b. Analyze a policy class with full flexibility to partially serve classes (“as fair as desired”).

2. Use fluid model invariant states to define an approximating scheduling control problem.
Some Related Works

• Single Class Fluid Model.
  – Provided the framework for approaching the multiclass case.

• Multiclass Scheduling.
  – Atar, Kaspi and Shimkin (2014) analyzed static priority for multiclass G/G/N+G.
  – We extend to non-static priority.

• Very Recent
  – Mukherjee, Li, and Goldberg (2018)
  – Large deviations analysis in Halfin-Whitt regime (M/H₂/N+M).
An **admissible scheduling policy** cannot know the future, does not preempt service, and satisfies mild conditions to control entry-into-service oscillations.
At the moment of departure, the available server next serves class $j$ with probability $p_j$ (if possible), where $\sum_{j=1}^{J} p_j = 1$. 
The Multiclass Many-Server Queue

The State Space: \( (\alpha^N, X^N, \nu^N, \eta^N) \).

Measure-valued processes.

Time elapsed since last class j arrival.

The number of class j customers in the system.
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 $\nu$ Measure (for given Class $j$)

- Customer entering service has age 0.
- Customer is no longer tracked once the time spent being served exceeds that customer’s service time.
The $\eta$ Measure (for given Class j)

Note: Independent of Scheduling Control.

Customer entering system has waited 0 time units.

Customer is no longer tracked once the time elapsed since arrival exceeds that customer’s patience 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.
Theorem (Convergence)

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

When the queue operates under an admissible scheduling rule, under mild initial conditions, a sequence of fluid-scaled state processes operating $(\alpha^N, X^N, \nu^N, \eta^N)/N$ is tight.

Suppose that $(X, \nu, \eta)$ is a distributional limit point of $\left\{ \left( \frac{X^N}{N}, \frac{\nu^N}{N}, \frac{\eta^N}{N} \right) \right\}$.

Then, we need to characterize $(X, \nu, \eta)$. 

Scaled arrival process. 

Number of servers. 

Scaled system processes.
The Fluid Model Solution Space and Auxiliary Functions

Let $S$ be the set of r.c.l.l. functions $(X, \nu, \eta)$ such that for all $j$ and $t \geq 0$,

\[ \int_0^t \langle h_{s,j}, \nu_j(u) \rangle \, du < \infty \quad \text{and} \quad \int_0^t \langle h_{a,j}, \eta_j(u) \rangle \, du < \infty \quad \text{(Finiteness)}. \]

For $(X, \nu, \eta) \in S$, define for all $j$ and $t \geq 0$,

- $B_j(t) := \langle 1, \nu_j(t) \rangle$, $I_j(t) = 1 - \sum_{j=1}^{J} B_j(t)$ \hspace{1cm} \text{(Proportion of class j fluid in service)};
- $D_j(t) := \int_0^t \langle h_{s,j}, \nu_j(u) \rangle \, du$ \hspace{1cm} \text{(Cumulative departure process)};
- $Q_j(t) := X_j(t) - B_j(t)$ \hspace{1cm} \text{(Queue-length process)};
- $\chi_j(t) := \inf \{ x \geq 0 : \langle 1_{[0,x]}, \eta_j(t) \rangle \geq Q_j(t) \}$ \hspace{1cm} \text{(Class j head-of-line wait time process)};
- $R_j(t) := \int_0^t \left\langle 1_{[0,\chi_j(u)]}, h_{a,j}, \eta_j(u) \right\rangle \, du$ \hspace{1cm} \text{(Cumulative abandonment process)};
- $K_j(t) := B_j(t) + D_j(t) - B_j(0)$ \hspace{1cm} \text{(Cumulative entry-into-service process)}. 

Number of jobs in system. \hspace{1cm} Age-in-service measure. \hspace{1cm} Potential queue measure.

Service distribution hazard rate. \hspace{1cm} Abandonment distribution hazard rate.
A Fluid Model Solution (Not Unique)

Let \( E \) be an arrival function. Then, \((X, \nu, \eta) \in \mathcal{S}\) 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 \( t \geq 0 \), \( X_j(t) = X_j(0) + E_j(t) - R_j(t) - D_j(t) \), and \( 0 \leq Q_j(t) \leq \int_0^{H_j^r} \eta_j(dy) \).

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

   \[
   \langle f, \nu_j(t) \rangle = \left\{ f(\cdot +t) \frac{\tilde{G}_{s,j}(\cdot +t)}{G_{s,j}(\cdot)}, \nu_j(0) \right\} + \int_0^t f(t-u)\tilde{G}_{s,j}(t-u)dK_j(u)
   \]

   \[
   \langle f, \eta_j(t) \rangle = \left\{ f(\cdot +t) \frac{\tilde{G}_{a,j}(\cdot +t)}{\tilde{G}_{a,j}(\cdot)}, \nu_j(0) \right\} + \int_0^t f(t-u)\tilde{G}_{a,j}(t-u)dE_j(u)
   \]

(As in Atar, Kaspi, and Shimkin 2014, with static priority equation eliminated.)
A specified WRBS fluid model solution also satisfies
\[ p_j \int_s^t 1\{Q_j(u) > 0\} dD_\Sigma(u) \leq K_j(t) - K_j(s) \leq p_j \int_s^t dD_\Sigma(u), 1 \leq j < J \]
and
\[ I(t) = [I(t) - Q_j(t)]^+ . \]

**Lemma**: If \( E_j \) is absolutely continuous with density \( \lambda_j(\cdot) \) for each \( j \), then so are the coordinates of \( X \) and the auxiliary functions, and

\[ K_j(t) = \int_0^t (\lambda_j(u) \land p_j \delta(u)) 1\{Q_j(u) = 0\} + p_j \delta(u) 1\{Q_j(u) > 0\} du, \]

where \( \delta \) is the density of \( D_\Sigma \).
Theorem (Non-Policy Specific Convergence)

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

Under mild initial conditions, a sequence of fluid-scaled state processes \((\alpha^N, X^N, \nu^N, \eta^N)/N\) is tight.

Suppose that \((X, \nu, \eta)\) is a distributional limit point of \(\left\{ \left( \frac{X^N}{N}, \frac{\nu^N}{N}, \frac{\eta^N}{N} \right) \right\}\).

Then, under mild conditions*, \((X, \nu, \eta)\) is, almost surely, a fluid model solution for \(E\) with specified initial state.

- Conditions are similar to the single class case. Hazard rates of abandonment and service distributions are either bounded or lower semi-continuous, and \(E_j\) is continuous for all \(j\) (for example, \(E_j(t) = \lambda_j t\)).
Theorem (Weak Convergence)

Suppose \( \lim_{N \to \infty} \frac{E^N}{N} = E \) almost surely, and \( \lim_{N \to \infty} \mathbb{E} \left[ \frac{E_j^N(t)}{N} \right] = \mathbb{E}[E_j(t)] \)

for all \( t \geq 0 \).

Under the conditions of the previous theorem, and also assuming the abandonment distributions have bounded hazard rate, the sequence of fluid-scaled processes \( \left\{ \left( \frac{X^N}{N}, \frac{\nu^N}{N}, \frac{\eta^N}{N} \right) \right\} \) weakly converges to the unique WRBS(\( p \)) fluid model solution.

*Bounded hazard may seem strong, but consistent with what was assumed for SP.*
Our Research Objective
(Also serves as Talk Outline.)

We want to understand the multiclass many server queue with abandonment, without making any distributional assumptions.

1a. Provide a fluid model relevant for a—very general class of scheduling rules.

1b. Analyze a policy class with full flexibility—to partially serve classes (“as fair as desired”).

2. Use fluid model invariant states to define an approximating scheduling control problem.
Fluid Model Invariant States

Assumptions.

• **(Fluid arrival process)** For some \( \lambda \in (0, \infty) \), \( E_j(t) = \lambda_j t \) for all \( j \) and \( t \geq 0 \).

• **(Overloaded)** For each \( j \), \( \rho_1 + \rho_2 + \cdots + \rho_J > 1 \) for \( \rho_j = \frac{\lambda_j}{\mu_j} \).

• **(Mean abandonment time)** For each \( j \), \( \int_0^\infty \bar{G}_a,j(x)dx = \frac{1}{\theta_j} \).

Definition (Feasible server effort allocation).

\[
\mathcal{B} = \left\{ b \in \mathbb{R}_+^J : b_j \leq \rho_j, \sum_{j=1}^J b_j \leq 1 \right\}
\]

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

\[
Q_j(t) = \frac{\lambda_j}{\theta_j} f_j \left( 1 - \frac{b_j}{\rho_j} \right) \quad \text{for all } t \geq 0,
\]

where \( f_j(x) = G_{a,e,j} \left( \left( G_{a,j} \right)^{-1}(x) \right) \).

**Abandonment stationary excess cdf.**

**Abandonment cdf.**

Intuition: If exponential abandonment distribution, then

\[
\frac{\lambda_j}{\theta_j} f_j \left( 1 - \frac{b_j}{\rho_j} \right) = \frac{1}{\theta_j} (\lambda_j - b_j \mu_j) = q_j.
\]

Flow balance implies \( \lambda_j - b_j \mu_j = \theta_j q_j \).
Fluid Model Invariant States

Assumptions.

- **(Fluid arrival process)** For some \( \lambda \in (0, \infty)^J, E_j(t) = \lambda_j t \) for all \( j \) and \( t \geq 0 \).
- **(Overloaded)** For each \( j \), \( \rho_1 + \rho_2 + \cdots + \rho_J > 1 \) for \( \rho_j = \frac{\lambda_j}{\mu_j} \).
- **(Mean abandonment time)** For each \( j \), \( \int_0^\infty G_{a,j}(x)dx = \frac{1}{\theta_j} \).

Definition (Feasible server effort allocation).

- \( B = \{ b \in \mathbb{R}^J_+: b_j \leq \rho_j, \sum_{j=1}^J b_j \leq 1 \} \)

Theorem. For each \( b \in B \), there exists an invariant state such that \( b_j \) is the proportion of server effort devoted to class \( j \), and

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

**Q1:** For any given \( b \in B \), how should I schedule so as to achieve \( b \)?

**Q2:** What is my approximating control problem?
The Fluid Control Problem

\[ m^* = \min_{b \in B_j} \sum_{j=1}^{J} c_j \frac{\lambda_j}{\theta_j} f_j \left( 1 - \frac{b_j}{\rho_j} \right) + a_j (\lambda_j - b_j \mu_j) \]

Queue
Abandonments

Solution Properties. When is static priority (asymptotically) optimal?

If there is no holding cost; that is, \( c_j = 0 \).

If the abandonment distribution has non-decreasing hazard rate (IFR), then
- \( f_j \) is concave, and \( m^* \) is achieved by a feasible vertex.
- I.E., the solution motivates a static priority policy.
  (Consistent with earlier, but don’t know ordering).

If the abandonment distribution has non-increasing hazard rate (DFR), then
- \( f_j \) is convex, and \( m^* \) 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.)
Assume No Holding Costs and Static Priority Scheduling.

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

(Q: Why does queue size decrease as variability increases?)

(High priority queue has predicted size 0, and simulated size about 1.5 for all values of the variability $v$.)
What are the Predicted Abandonment Rates?
(Recall: Two-class $M/LN(1,4)/100 + LN(1, \nu)$ queue, with each class having arrival rate 60 per hour.)

<table>
<thead>
<tr>
<th>Class 1</th>
<th>Class 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$(\lambda_2 - b_2 \mu_2) \times N = (0.6 - 0.4) \times 100 = 20$</td>
</tr>
</tbody>
</table>

A: Even though the same number of jobs abandon, jobs that abandon do so sooner, reducing average queue-size and wait time.
The Fluid Control Problem

\[ m^* = \min_{b \in B_j} \sum_{j=1}^{J} c_j \frac{\lambda_j}{\theta_j} f_j \left( 1 - \frac{b_j}{\rho_j} \right) + a_j (\lambda_j - b_j \mu_j) \]

Queue

Abandonments

Solution Properties.  When is static priority (asymptotically) optimal?

If there is no holding cost; that is, \( c_j = 0 \).

If the abandonment distribution has non-decreasing hazard rate (IFR), then
- \( f_j \) is concave, and \( m^* \) is achieved by a feasible vertex.
- I.E., the solution motivates a static priority policy.
  —(Consistent with earlier, but don’t know ordering).

If the abandonment distribution has non-increasing hazard rate (DFR), then
- \( f_j \) is convex, and \( m^* \) 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.)
Example with Non-Vertex Optima

\[ m^* = \min_{b \in B_J} \sum_{j=1}^{J} c_j \frac{\lambda_j}{\theta_j} f_j \left( 1 - \frac{b_j}{\rho_j} \right) + a_j (\lambda_j - b_j \mu_j) \]

Queue

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(\( \theta_2 \));
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

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

(This example is developed by Amber Puha’s student Jacques Coulombe.)
Fluid Model Invariant States

Assumptions.

• **(Fluid arrival process)** For some $\lambda \in (0, \infty)^J$, $E_j(t) = \lambda_j t$ for all $j$ and $t \geq 0$.

• **(Overloaded)** For each $j$, $\rho_1 + \rho_2 + \cdots + \rho_J > 1$ for $\rho_j = \frac{\lambda_j}{\mu_j}$.

• **(Mean abandonment time)** For each $j$, $\int_0^\infty \bar{G}_{a,j}(x) dx = \frac{1}{\theta_j}$.

Definition **(Feasible server effort allocation).**

• $B = \{b \in \mathbb{R}^J_+: b_j \leq \rho_j, \sum_{j=1}^J b_j \leq 1\}$

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

$$Q_j(t) = \frac{\lambda_j}{\theta_j} f_j \left(1 - \frac{b_j}{\rho_j}\right)$$

for all $t \geq 0$, where $f_j(x) = G_{a,e,j} \left((G_{a,j})^{-1}(x)\right)$.

Abandonment stationary excess cdf.  
Abandonment cdf.

Q1: For any given $b \in B$, how should I schedule so as to achieve $b$?

Q2: What is my approximating control problem?
Conjecture: WRBS is Asymptotically Optimal

Convergence to Fluid Control Problem Solution:
If \( b \in B \) solves the fluid control problem, then the RBS policy that sets

\[
p_j = \frac{\mu_j b_j}{\sum_{k=1}^J \mu_k b_k}
\]

has cost equal to \( m^* \) on fluid scale; that is,

\[
\lim_{N \to \infty} \lim_{T \to \infty} \mathbb{E} \left[ \frac{1}{T} \sum_{j=1}^J \left( \int_0^T c_j Q_j^N(t; RBS) dt + a_j \frac{R_j^N(T; RBS)}{T} \right) \right] = m^*.
\]

*To mimic static priority, set \( b_j = \rho_j \) for high priority classes.

Lower Bound:
Under any admissible policy \( \pi \in \Pi \),

\[
\lim_{N \to \infty} \lim_{T \to \infty} \mathbb{E} \left[ \frac{1}{T} \sum_{j=1}^J \left( \int_0^T c_j Q_j^N(t; \pi) dt + a_j \frac{R_j^N(T; \pi)}{T} \right) \right] \geq m^*.
\]
Concluding Remarks

<table>
<thead>
<tr>
<th>Fluid Control Problem Assumptions</th>
<th>Scheduling</th>
</tr>
</thead>
<tbody>
<tr>
<td>No holding cost</td>
<td>Static Priority RBS</td>
</tr>
<tr>
<td>IFR</td>
<td>Static Priority RBS</td>
</tr>
<tr>
<td>DFR</td>
<td>RBS</td>
</tr>
</tbody>
</table>

**Tutorial paper (with open problems) available soon from my web page (or email me):** [http://faculty.chicagobooth.edu/Amy.Ward/publications.html](http://faculty.chicagobooth.edu/Amy.Ward/publications.html)