Asymptotic analysis of multiclass queues with random order of service

Reza Aghajani
Joint work with Ruth Williams

UC San Diego

Frontier Probability Days 2018

Enzymatic Reactions in Cells:

1. proteins produced:

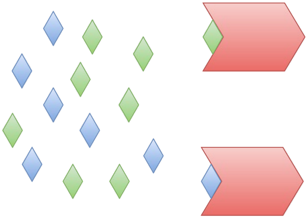
$$\emptyset \xrightarrow{\lambda_\ell} X_\ell$$

2. proteins processed by Enzyme:

$$X_\ell + E \xrightarrow{\mu_\ell} E$$

3. dilution:

$$X_\ell \xrightarrow{\gamma_\ell} \emptyset$$



- Different species of proteins are processed by a shared pool of enzymes
- The goal is to study the effect of this shared processing resources on the **correlation** between numbers of protein of different species.

Motivation

Queueing models have been used to study these molecular reactions.
jobs: proteins, servers: enzymes.

Characteristics:

- random order of service (ROS) discipline: proteins do not stand in lines!
- reneging: to models dilution.
- Multiclass: to represent different species of proteins
- many-server: there are typically more than one copy of the enzyme

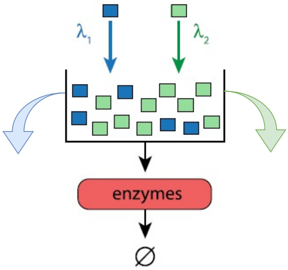
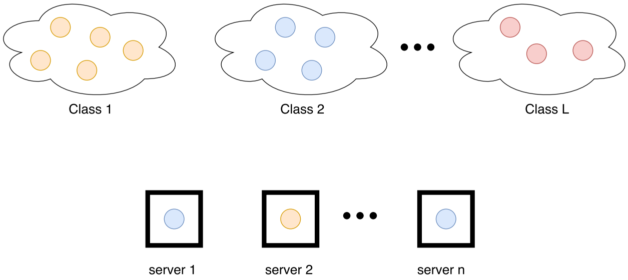


Figure taken from [Mather et al. 2010] and edited.

Model: Random Order of Service

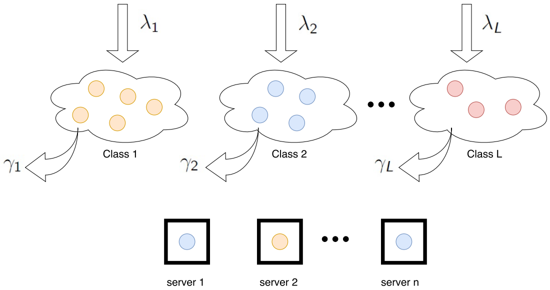
Multiclass, many-server queue with reneging under (D)ROS



- jobs are of L different classes
- jobs are processed by n homogeneous, non-idling servers
- each server can process jobs from all classes

Model: Random Order of Service

Multiclass, many-server queue with reneging under (D)ROS

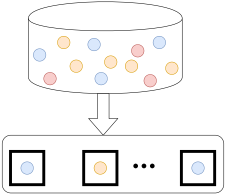


Jobs of each class ℓ:

- arrive according to a renewal process at rate λ_ℓ.
- have i.i.d. patience times with inverse mean γ_ℓ.
- have i.i.d. service requirement $\{v_{\ell,j}\}$ with inverse mean μ_ℓ
- $Q_\ell(t)$ is number of queues of class ℓ **waiting in queue** at time t.

Model: Random Order of Service

Multiclass, many-server queue with reneging under (D)ROS



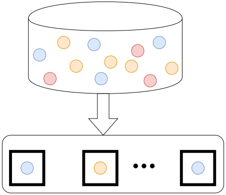
Service policy: Random Order of Service:

- upon server availability, a job is randomly selected for service entery from all jobs waiting in queue
- **ROS:** all job classes are treated equally :

$$P(\text{a given job is selected}) = \frac{1}{\sum_{\ell=1}^{L} Q_\ell(t)} = \frac{1}{Q(t)}$$

Model: Random Order of Service

Multiclass, many-server queue with reneging under (D)ROS



Service policy: Random Order of Service:

- upon server availability, a job is randomly selected for service entery from all jobs waiting in queue
- **ROS:** all job classes are treated equally :

$$P(\text{a given job is selected}) = \frac{1}{\sum_{\ell=1}^{L} Q_\ell(t)} = \frac{1}{Q(t)}$$

- **DROS**: the random selection is discriminatory:

$$P(\text{a job is selected from class } j) = \frac{p_j}{\sum_{\ell=1}^{L} p_\ell Q_\ell(t)}$$

Prior Work

Most of prior work on queues with ROS assume

- Poisson arrivals: [Burke 59], [Kingman 62], [Carter-Cooper 72], [Balmer 72], [Boxma et al. 15]
- Exponential Distribution: [Borst et al. 03], [Rogiest et al. 14]
- Exceptions are [Zwart 05], [Kim and Kim 12]

Same holds for multiclass case under DROS

- [Kim et al. 11], [Ayesta et al. 11], [Rogiest et al. 14], [Izagirre et al. 2015]

Prior Work

Most of prior work on queues with ROS assume

- Poisson arrivals: [Burke 59], [Kingman 62], [Carter-Cooper 72], [Balmer 72], [Boxma et al. 15]
- Exponential Distribution: [Borst et al. 03], [Rogiest et al. 14]
- Exceptions are [Zwart 05], [Kim and Kim 12]

Same holds for multiclass case under DROS

- [Kim et al. 11], [Ayesta et al. 11], [Rogiest et al. 14], [Izagirre et al. 2015]

However, none of the above considers reneging. In fact, ROS with reneging is only studied in

- [Barrer 57]: single class, Poisson arrivals, exponential service time, deterministic patience time
- [Kelly 1979]: multiclass*, exponential everything.
- [Mather et al. 10] multiclass, exponential everything.

Our Goal

It is known that processing times in biological systems are not always exponentially distributed, “especially when operations such as binding, folding, transcription and translation are involved”.

Our Goal: Study multiclass, many-server queues

- operating under (D)ROS
- with reneging,
- renewal arrivals
- non-exponential service requirements
- non-exponential patience times

Challenges

Challenges:

- ROS is non-head-of-the-line policy, and hard to analyze.
- For non-exponential patience times, one needs to keep track of ages (time since arrival) or residual patience times of all jobs

Any Markovian representation will be infinite-dimensional

Challenges:

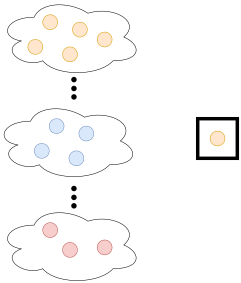
- ROS is non-head-of-the-line policy, and hard to analyze.
- For non-exponential patience times, one needs to keep track of ages (time since arrival) or residual patience times of all jobs

> Any Markovian representation will be infinite-dimensional

- As this model has not yet been studied even for single server queues, we start with that case.

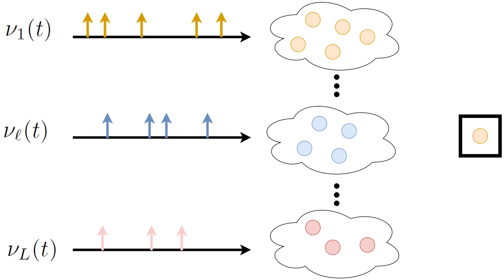
State Representation

A Measure-Valued State Representation.



State Representation

A Measure-Valued State Representation.



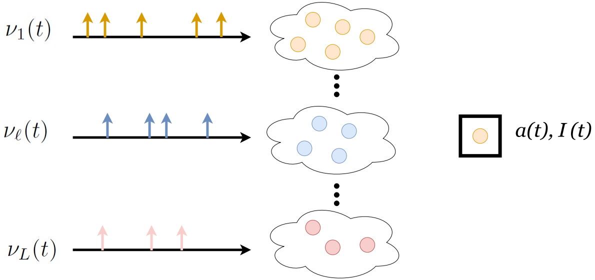
1. Ages in queue:

$$\nu_\ell(t) = \sum_{\mathcal{Q}_\ell(t)} \delta_{w_{\ell,j}(t)}$$

- $w_{\ell,j}(t)$: age in queue (time since arrival) of job j of class ℓ at time t
- $\mathcal{Q}_\ell(t)$: all jobs of class ℓ waiting in queue at time t
- Queue length of type ℓ: $Q_\ell = \langle 1, \nu_\ell \rangle$

State Representation

A Measure-Valued State Representation.

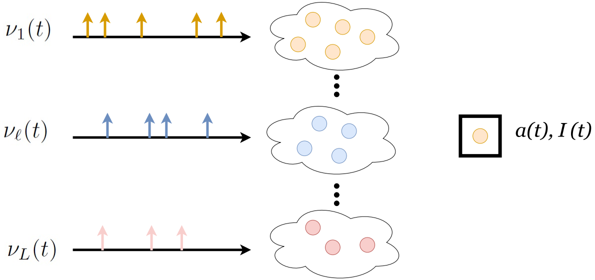


2. Job in service:
   - $a(t)$: age in service (time since service entry) of the job receiving service.
   - $I(t)$: class index of the job receiving service

3. Arrivals:
   - $R_\ell(t)$: time since last arrival of class ℓ at time t

State Representation

A Measure-Valued State Representation.



Markovian state descriptor:

$$Y(t) = \big(R_\ell(t), \nu_\ell(t); \ell = 1, ..., L,\ a(t), I(t)\big)$$

Remark. Our representation keeps track of "ages". Alternative representation may track *residual patience and service times.*

- Like many other complex stochastic network models, this model is not amenable to exact analysis.
- As the first step, we use fluid approximation to study this model.

Fluid Limit Scaling:

Consider a sequence of queueing systems, parameterized by $r \in N$:

- speed up arrivals: $E_\ell^r(t) = E_\ell(rt)$,
- speed up service rates: $v_{\ell,j}^r = \frac{1}{r} v_{\ell,j}$
- patience times unchanged.
- Queue lengths and ν_ℓs are scaled:

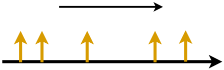
$$\bar{Q}_\ell^r(t) = \frac{Q_\ell^r(t)}{r}, \quad \bar{\nu}_\ell^r(t) = \frac{\nu_\ell^r(t)}{r}.$$

We are interested in the limit $\bar{\nu}$ of $\bar{\nu}^r = (\bar{\nu}_\ell^r)$ as $r \to \infty$.

Dynamics

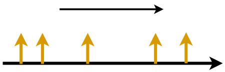
Dynamics of ν_ℓ:

1. Linear growth of ages with time: masses move to the right

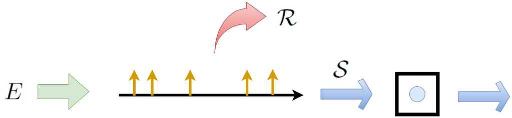


Dynamics of ν_ℓ:

1. Linear growth of ages with time: masses move to the right



2. Arrivals, renegings, and service entries.

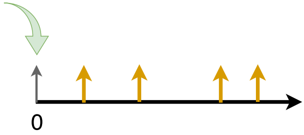


$$\langle f, \nu_\ell(t)\rangle = \langle f, \nu_\ell(0)\rangle + \langle f', \nu_\ell(t)\rangle + \mathcal{E}_\ell(t; f) - \mathcal{R}_\ell(t; f) - \mathcal{S}_\ell(t; f).$$

- dynamics of ν_ℓ for different classes is are coupled through the service entry term $\mathcal{S}$

1. Arrivals.

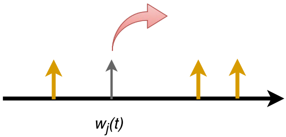
If a new jobs arrives, it only lands in queue if the server is busy, i.e., when the total number of jobs $X(t)$ in system is non-zero.



$$\mathcal{E}_\ell(t; f) = \int_0^t \mathbf{1}(X(s-) \geq 1) f(0) dE_\ell(s)$$

2. Reneging.

Each job j waiting in queue with age in queue $w_{j,\ell}(t)$ can renege:



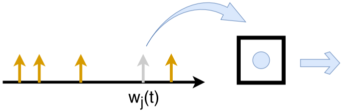
There is a martingale M_R s.t.

$$\mathcal{R}_\ell(t; f) = \int_0^t \langle f\ h_{R,\ell}, \nu_\ell(s) \rangle ds + M_R(t)$$

where $g_{R,\ell}$, $G_{R,\ell}$, and $h_{R,\ell}$ are pdf, cdf, and hazard rate of patience times for jobs of class ℓ.

3. Service Entry.

Service entries of jobs in queue happen immediately after departures.



There is a martingale M_S such that

$$\mathcal{S}_\ell(t; f) = \int_0^t h_{S,I(s-)}\,(a(s))\,\frac{p_\ell\langle f, \nu_\ell(t)\rangle}{\sum_{\ell'=1}^L p_{\ell'} Q_{\ell'}(s-)} ds + M_S(t)$$

where $g_{S,\ell}$, G_{S_ℓ}, and h_{S_ℓ} are pdf, cdf, and hazard rate of service times of jobs of class ℓ.

Fluid Limit

Theorem (Fluid Limit)

Under the assumption that $h_{R,\ell}$s are bounded, $(\bar{\nu}_1^r, ..., \bar{\nu}_L^r)$ is tight in $\mathbb{D}^L_{\mathbb{M}_F}[0,\infty)$, and each subsequential limit $(\bar{\nu}_1, ..., \bar{\nu}_L)$ satisfies

$$\langle f, \bar{\nu}_\ell(t)\rangle = \langle f, \bar{\nu}_\ell(0)\rangle + \int_0^t \langle f' - f h_{R,\ell}, \bar{\nu}_\ell(s)\rangle ds + \lambda_\ell f(0) \int_0^t 1(\bar{q}(s) > 0) ds - \int_0^t 1(\bar{q}(s) > 0) \frac{p_\ell \langle f, \bar{\nu}_\ell(s)\rangle}{\sum_{j=1}^L \frac{p_j}{\mu_j} \langle \mathbf{1}, \bar{\nu}_j(s)\rangle} ds,$$

for every $f \in C_b^1(\mathbb{R}_+)$, where $\bar{q}(t) = \sum_{\ell=1}^L \langle \mathbf{1}, \bar{\nu}_\ell(t)\rangle$.

Fluid Limit

Theorem (Fluid Limit)

Under the assumption that $h_{R,\ell}$s are bounded, $(\bar{\nu}_1^r, ..., \bar{\nu}_L^r)$ is tight in $\mathbb{D}^L_{\mathbb{M}_F}[0,\infty)$, and each subsequential limit $(\bar{\nu}_1, ..., \bar{\nu}_L)$ satisfies

$$\langle f, \bar{\nu}_\ell(t)\rangle = \langle f, \bar{\nu}_\ell(0)\rangle + \int_0^t \langle f' - f h_{R,\ell}, \bar{\nu}_\ell(s)\rangle ds + \lambda_\ell f(0) \int_0^t 1(\bar{q}(s) > 0) ds - \int_0^t 1(\bar{q}(s) > 0) \frac{p_\ell \langle f, \bar{\nu}_\ell(s)\rangle}{\sum_{j=1}^L \frac{p_j}{\mu_j} \langle \mathbf{1}, \bar{\nu}_j(s)\rangle} ds,$$

for every $f \in C_b^1(\mathbb{R}_+)$, where $\bar{q}(t) = \sum_{\ell=1}^L \langle \mathbf{1}, \bar{\nu}_\ell(t)\rangle$.

Proof steps:

1. bounds for fluctuations to get tightness.
2. Theory of point processes + martingale decomposition.
3. subsequential limits: multi-scale analysis.

About The Proof

Proof: Multi-Scale Analysis

- in the fluid scaling regime, service variables $(I^r(t), a^r(t))$ evolve on a faster time scale, compared to the slower measure-valued processes $\bar{\nu}_\ell^r$.

We need to perform a multi-scale analysis to establish an averaging principle for slow and fast components

About The Proof

Proof: Multi-Scale Analysis

- in the fluid scaling regime, service variables $(I^r(t), a^r(t))$ evolve on a faster time scale, compared to the slower measure-valued processes $\bar{\nu}^r_\ell$.

We need to perform a multi-scale analysis to establish an averaging principle for slow and fast components

- On a small interval $[s, s+\delta]$ where $\bar{\nu}^r$ is approximately constant, I^r nearly reaches to equilibrium:

$$\beta_\ell(s) \approx \frac{p_\ell \bar{Q}^r_\ell(s)}{\sum_{j=1}^L p_j \bar{Q}^r_j(s)}$$

- The limiting expected departure rate is therefore

$$\frac{1}{\sum_{\ell=1}^L \beta_\ell(s)/\mu_\ell} = \frac{\sum_{\ell=1}^L p_\ell \bar{Q}_\ell(s)}{\sum_{\ell=1}^L \frac{p_\ell}{\mu_\ell} \bar{Q}_\ell(s)}$$

Fluid Limit

Theorem (Fluid Limit)

Under the assumption that $h_{R,\ell}s$ are bounded, $(\bar{\nu}_1^r, ..., \bar{\nu}_L^r)$ is tight in $\mathbb{D}^L_{\mathbb{M}_F}[0,\infty)$, and each subsequential limit $(\bar{\nu}_1, ..., \bar{\nu}_L)$ satisfies

$$\langle f, \bar{\nu}_\ell(t)\rangle = \langle f, \bar{\nu}_\ell(0)\rangle + \int_0^t \langle f' - f h_{R,\ell}, \bar{\nu}_\ell(s)\rangle ds + \lambda_\ell f(0) \int_0^t 1(\bar{q}(s) > 0) ds - \int_0^t 1(\bar{q}(s) > 0) \frac{p_\ell \langle f, \bar{\nu}_\ell(s)\rangle}{\sum_{j=1}^L \frac{p_j}{\mu_j} \langle \mathbf{1}, \bar{\nu}_j(s)\rangle} ds,$$

for every $f \in C_b^1(\mathbb{R}_+)$, where $\bar{q}(t) = \sum_{\ell=1}^L \langle \mathbf{1}, \bar{\nu}_\ell(t)\rangle$.

The fluid limit equation is

1. a system of measure-valued equations
2. the equations are coupled through the non-linear term in the last integrand, hard yo analyze.
3. interested in uniqueness and long-time behavior.

Single-Class Case:

Consider a simplified model where

- there is a single class ($L = 1$); only one measure-valued process $\bar{\nu}$.
- overloaded cases: $\lambda > \mu$ (interesting case)
- set $\mu = 1$.

The equation reduces the single equation

$$\langle f, \bar{\nu}(t)\rangle = \langle f, \bar{\nu}(0)\rangle + \int_0^t \langle f' - f h_R, \bar{\nu}(s)\rangle ds + \lambda f(0) t - \mu \int_0^t \frac{\langle f, \bar{\nu}(s)\rangle}{\langle \mathbf{1}, \bar{\nu}(s)\rangle} ds$$

- One is often interested in limiting queue length $\bar{q}(t) = \langle \mathbf{1}, \bar{\nu}(t)\rangle$

Fluid PDE

Observation. The fluid limit equation is closed under one-parameter family of functions $\{f^x; x \geq 0\}$:

$$\left\{ f^x(u) = \frac{\overline{G}_R(u+x)}{\overline{G}_R(u)}; x \geq 0 \right\} \qquad (\overline{G}_R = 1 - G_R)$$

Note that $\mathbf{1} = f^0$. (`[A.-Xi-Ramanan 17, A.-Ramanan 15]`)

Fluid PDE

Observation. The fluid limit equation is closed under one-parameter family of functions $\{f^x; x \geq 0\}$:

$$\left\{ f^x(u) = \frac{\overline{G}_R(u+x)}{\overline{G}_R(u)}; x \geq 0 \right\} \qquad (\overline{G}_R = 1 - G_R)$$

Note that $\mathbf{1} = f^0$. (`[A.-Xi-Ramanan 17, A.-Ramanan 15]`)

- We define

$$\bar{Z}(t,x) = \langle f^x, \bar{\nu}(t) \rangle$$

Fluid PDE

Observation. The fluid limit equation is closed under one-parameter family of functions $\{f^x; x \geq 0\}$:

$$\left\{ f^x(u) = \frac{\overline{G}_R(u+x)}{\overline{G}_R(u)}; x \geq 0 \right\} \qquad (\overline{G}_R = 1 - G_R)$$

Note that $\mathbf{1} = f^0$. (`[A.-Xi-Ramanan 17, A.-Ramanan 15]`)

- We define

$$\bar{Z}(t,x) = \langle f^x, \bar{\nu}(t) \rangle$$

- plugging f^x in fluid limit equation, $\bar{Z}$ satisfies the "fluid PDE"

$$\partial_t \bar{Z}(t,x) - \partial_x \bar{Z}(t,x) = \lambda \overline{G}_R(x) - \frac{\bar{Z}(t,x)}{\bar{Z}(t,0)} \tag{1}$$

  which is a non-linear transport equation, with boundary condition $\bar{Z}(t,0) = \langle \mathbf{1}, \nu(t) \rangle = \bar{q}(t)$.

Fluid PDE

About the Fluid PDE

$$\partial_t \bar{Z}(t,x) - \partial_x \bar{Z}(t,x) = \lambda \overline{G}_R(x) - \frac{\bar{Z}(t,x)}{\bar{Z}(t,0)} \tag{2}$$

- This reduced fluid model $\bar{Z}$ is function-valued and characterized by a PDE.
- This generalized the so-called ODE method for finite-dimensional Markov Processes, we can call it the PDE method.
- PDE is non-standard: b.c. appears as external force

Fluid PDE

Conjecture (Uniqueness)

When $\rho > 1$ *and* h_R *is bounded, for every initial condition* $Z(0,\cdot) = z(\cdot) \geq 0$, *the PDE*

$$\partial_t \bar{Z}(t,x) - \partial_x \bar{Z}(t,x) = \lambda \overline{G}_R(x) - \frac{\bar{Z}(t,x)}{\bar{q}(t)}$$

has a unique solution.

Fluid PDE

Conjecture (Uniqueness)

When $\rho > 1$ and h_R is bounded, for every initial condition $Z(0,\cdot) = z(\cdot) \geq 0$, the PDE

$$\partial_t \bar{Z}(t,x) - \partial_x \bar{Z}(t,x) = \lambda \overline{G}_R(x) - \frac{\bar{Z}(t,x)}{\bar{q}(t)}$$

has a unique solution.

- proved when the initial condition satisfies $\bar{Z}(0,\cdot) > 0$.
- for zero i.c., an argument similar to `[Puha-Stolyar-Williams 06]` for Processor sharing is expected to work.

Fluid PDE

Conjecture (Uniqueness)

When $\rho > 1$ and h_R is bounded, for every initial condition $Z(0,\cdot) = z(\cdot) \geq 0$, the PDE

$$\partial_t \bar{Z}(t,x) - \partial_x \bar{Z}(t,x) = \lambda \overline{G}_R(x) - \frac{\bar{Z}(t,x)}{\bar{q}(t)}$$

has a unique solution.

- proved when the initial condition satisfies $\bar{Z}(0,\cdot) > 0$.
- for zero i.c., an argument similar to `[Puha-Stolyar-Williams 06]` for Processor sharing is expected to work.

Proof sketch.

- partially solve transport equation
- show the resulting fixed point equation for $\bar{q}(\cdot)$ has a unique solution
- a key challenge is the appearance of $\bar{q}(t)$ in denominators.

Fluid PDE

Theorem (Steady-State Solution)

When $\rho > 1$, the PDE (2) has a unique steady state solution z_* given by

$$z^*(x) = \lambda \int_x^\infty \overline{G}_R(u) e^{\frac{x-u}{q}} du \tag{3}$$

with q is the unique solution to

$$q = \lambda \hat{G}_R(\frac{1}{q}),$$

where $\hat{G}_R$ is the Laplace transform of $\overline{G}_R$.

Fluid PDE

Theorem (Steady-State Solution)

When $\rho > 1$, the PDE (2) has a unique steady state solution z_* given by

$$z^*(x) = \lambda \int_x^\infty \overline{G}_R(u) e^{\frac{x-u}{q}} du \tag{3}$$

with q is the unique solution to

$$q = \lambda \hat{G}_R(\frac{1}{q}),$$

where $\hat{G}_R$ is the Laplace transform of $\overline{G}_R$.

Proof Sketch.

1. fixed point characterization is by Laplace analysis of the PDE.
2. write the pde for $Z(t,x) - z_*(x)$, then partially solve
3. the equation gives a Gronwall-type bound for $|q(t) - q|$
4. show $q(t) \to q$, and then $Z(t,\cdot) \to z_*$

Multiclass PDE

Challenge: Analysis of multiclass fluid equations

Similar to the single-class case, we can write fluid PDEs for multiclass case:

$$\partial_t Z_k(t,x) - \partial_x Z_k(t,x) = \lambda_k \overline{G}_{R,k}(x) - \frac{p_k Z_k(t,x)}{\sum_{\ell=1}^{K} \frac{p_\ell}{\mu_\ell} Z_\ell(t,0)}. \tag{4}$$

- above is a system of coupled, non-linear PDEs.
- because of the non-linear coupling, these equations are harder to analyze
- stationary solution is identified, and shown to be unique.
- Uniqueness of the equation is ongoing.

Summary

We analyzed a multiclass queue with Random Order of Service policy and reneging, under the non-exponential service and patience time assumptions, using the framework of measure-valued processes. Our motivation is two fold:

1. better understanding of intracellular molecular reactions, using a model with more realistic assumptions, i.e., non-exponential times.

2. advance the theory of measure-valued processes and their scaling limits in the context of queueing networks.
    - use of measure-valued processes for different queueing model leads to new infinite-dimensional deterministic and stochastic evolution equations.
    - in the absence of a general theory, new challenges introduced by each model need to be addressed in a case-by-case basis.

Ongoing and Future Work

1. The PDE analysis of multiclass case is ongoing.
2. **Diffusion Approximation**
   - diffusion approximation is needed for the analysis of correlations between job classes
   - stability analysis of fluid limit is a key step
3. **Many-Server Queues**
   - many-server queue is the more relevant model for our application; there are typically more than one copy of an enzyme
   - same framework can be employed; the dynamics will be more complicated.