Hydrodynamic Limits of Randomized Load Balancing Networks

Kavita Ramanan
and
Mohammadreza Aghajani

Brown University

Stochastic Networks and Stochastic Geometry
a conference in honour of François Baccelli's 60th birthday
IHP, Paris, Jan 2015

A Plethora of Scientific Interests

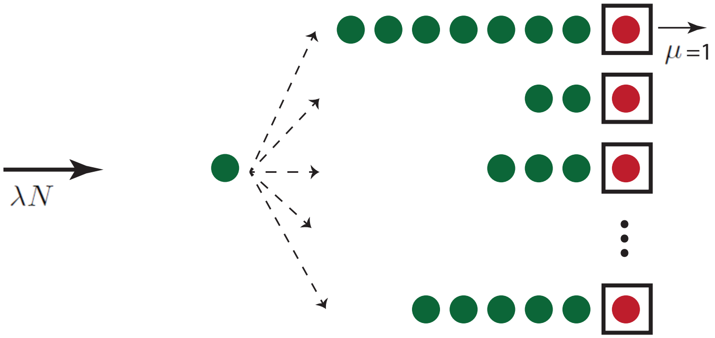
François Baccelli

- Stochastic Geometry
- Information theory
- Stochastic network calculus
- Simulation
- Performance Evaluation
- Wireless Networks
- ...
- "A Mean-Field Model for Multiple TCP Connections through a Buffer Implementing RED", François Baccelli, David R. Mcdonald, Julien Reynier, 2002.

Model of Interest

Network with

- N identical servers
- an infinite capacity queue for each server
- a common arrival process routed immediately on arrival
- FCFS service discipline within each queue

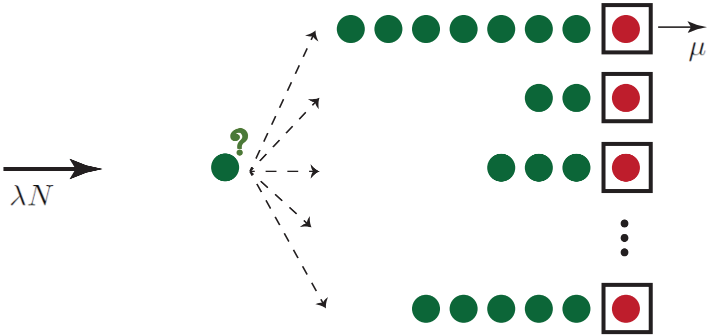


Model of Interest

Load Balancing Algorithm:

- How to assign incoming jobs to servers?
- Aim to achieve good performance with low computational cost

Goal: Analysis and comparison of different load balancing algorithms



Model of Interest

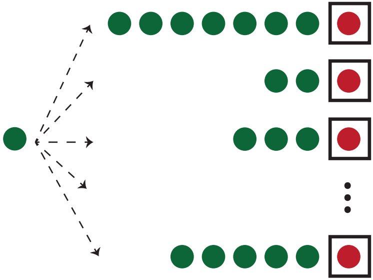
Appears in:

- Supermarkets
- Hash tables
- Distributed memory machines
- Path selection in networks
- Web Servers
- etc.

Routing Algorithm: Supermarket Model

Each arriving job

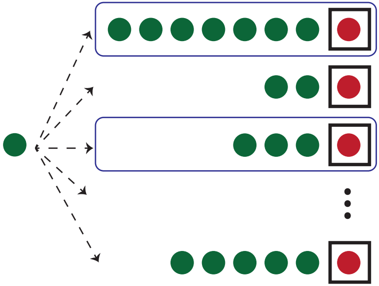
- chooses d queues out of N, uniformly at random
- joins the shortest queue among the chosen d
- ties broken uniformly at random



Routing Algorithm: Supermarket Model

Each arriving job

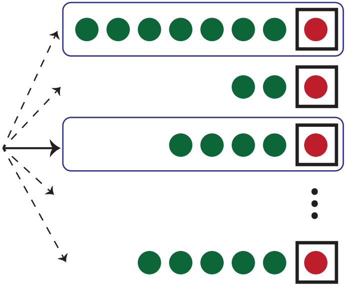
- chooses d queues out of N, uniformly at random
- joins the shortest queue among the chosen d
- ties broken uniformly at random



Routing Algorithm: Supermarket Model

Each arriving job

- chooses d queues out of N, uniformly at random
- joins the shortest queue among the chosen d
- ties broken uniformly at random



Prior Work - Exponential Service Distribution

Supermarket model for exponential service time

Fluid limit and steady state queue length decay rate is obtained as $N \to \infty$

- case $d = 2$ [Vvedenskaya-Dobrushin-Karpelevich '96]
- case $d \geq 2$ [Mitzenmacher '01]

General approach

Using Markovian state descriptor $\{S_\ell^N(t); \ell \geq 1, t \geq 0\}$

- $S_\ell^N(t)$: fraction of stations with at least ℓ jobs
- Convergence as $N \to \infty$ proved using an extension of Kurtz's theorem
- The limit process is a solution to a countable system of coupled ODEs
- Steady state queue length approximated by fixed point of the ODE sequence

Prior Work - Exponential Service Distribution

Summary of Results:

$X^{i,N}$ – length of ith queue in an N-server network

- $d = d_N = N$ (Joint the Shortest Queue - JSQ)
  - Performance: $P(X^{i,N}(\infty) > \ell) \to 0$ for $\ell \geq 1$
  - Computational Cost: N comparisons per routing (**not feasible**)

Power of two Choices: double-exponential decay for $d \geq 2$

Prior Work - Exponential Service Distribution

Summary of Results:

$X^{i,N}$ – length of ith queue in an N-server network

- $d = N$ Join the Shortest Queue (JSQ)
  - Performance: $P(X^{i,N}(\infty) > \ell) \to 0$ for $\ell \geq 1$
  - Computational Cost: N comparisons per routing (**not feasible**)

- $d = 1$ (random routing, decoupled $M/M/1$ queues):
  - Performance: $P(X^{i,N}(\infty) > \ell) \to c\lambda^{\ell}$
  - Computational cost: one random flip per routing

Power of two Choices: double-exponential decay for $d \geq 2$

Prior Work - Exponential Service Distribution

Summary of Results:

$X^{i,N}$ – length of ith queue in an N-server network

- Joint the Shortest Queue (JSQ)
  - Performance: $P(X^{i,N}(\infty) > \ell) \to 0$ for $\ell \geq 1$
  - Computational Cost: N comparison per routing (**not feasible**)

- $d \geq 2$ (supermarket model):
  - Performance: $P(X^N(\infty) > \ell) \to \lambda^{(d^\ell - 1)/(d-1)}$
  - Computational Cost: d random flips and $d-1$ comparison per routing

- $d = 1$ (random routing, decoupled $M/M/1$ queues):
  - Performance: $P(X^{i,N}(\infty) > \ell) \to c\lambda^\ell$
  - Computational cost: one random flip per routing

Power of two Choices: double-exponential decay for $d \geq 2$

Prior Work - Exponential Service Distribution

Summary of Results:

$X^{i,N}$ – length of ith queue in an N-server network

- Joint the Shortest Queue (JSQ)
  - Performance: $P(X^{i,N}(\infty) > \ell) \to 0$ for $\ell \geq 1$
  - Computational Cost: N comparison per routing (**not feasible**)
- $d \geq 2$ (supermarket model):
  - Performance: $P(X^N(\infty) > \ell) \to \lambda^{(d^\ell - 1)/(d-1)}$
  - Computational Cost: d random flips and $d-1$ comparison per routing
- $d = 1$ (random routing, decoupled $M/M/1$ queues):
  - Performance: $P(X^{i,N}(\infty) > \ell) \to c\lambda^\ell$
  - Computational cost: one random flip per routing

Power of two Choices: double-exponential decay for $d \geq 2$

Prior Work -General Service Distribution

Our Focus: General service time distributions

- almost nothing was known 5 years ago
- Mathematical challenge:
  - $\{S_\ell^N\}$ is no longer Markovian
  - need to keep track of more information
  - No common countable state space for Markovian representations of all N-server networks

Prior Work -General Service Distribution

Recent Progress:

1. When $\lambda < 1$ (proved in a more general setting)
   - Stability of N-server networks [Foss-Chernova'98]
   - Tightness of stationary distribution sequence [Bramson'10]
2. Under further restrictions – namely, service distributions with decreasing hazard rate and time-homogeneous Poisson arrivals
   - Results on decay rate of limiting stationary queue length [Bramson-Lu-Prabhakar'13]
   - Their approach (cavity method) only yields the steady-state distribution – no information on transient behavior
   - Requires showing asymptotic independence on infinite time intervals and the study of a queue in a random environment
   - According to Bramson, extending this asymptotic independence result to more general service distributions is a challenging task

Prior Work -General Service Distribution

A Phase Transition Result

Theorem (Bramson, Lu, Prabhakar '12)

Suppose the service distribution is a power law distribution with exponent $-\beta$. Then

- If $\beta > d/(d-1)$, the tail is doubly exponential
- If $\beta < d/(d-1)$, the tail has a power law
- If $\beta = d/(d-1)$ then the tail is exponentially distributed

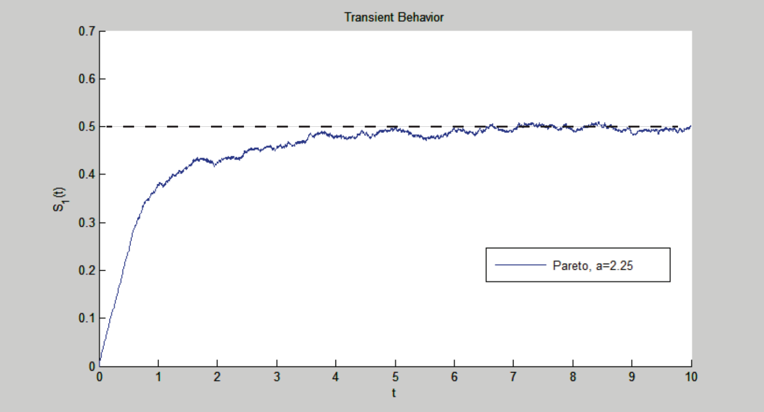
Observe: The "power of two choices" fails when $\beta \leq 2$

Motivates a better understanding of general service distributions
There is also the need to better understand transient behavior ...

Transient Behavior - Simulation (exponential service)

Simulation results for *fraction of busy servers**

- Poisson arrival with $\lambda = 0.5$
- 1000 servers
- empty initial condition

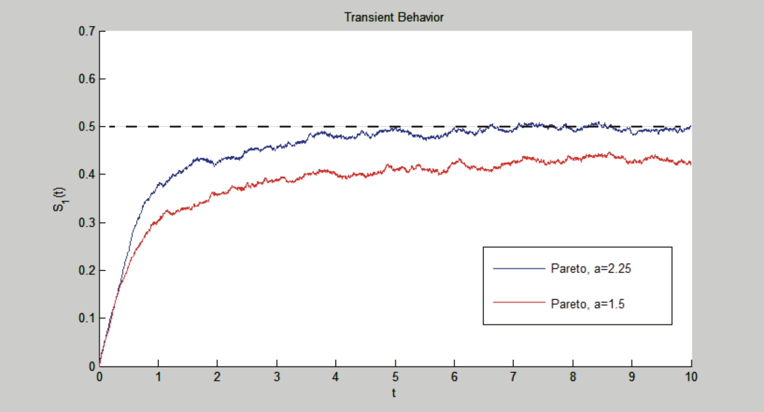


*Simulation results by Xingjie Li, Brown University

Transient Behavior - Simulation (exponential service)

Simulation results for *fraction of busy servers**

- Poisson arrival with $\lambda = 0.5$
- 1000 servers
- empty initial condition



*Simulation results by Xingjie Li, Brown University

Our Goal

Observations:

- No existing results on the time scale to reach equilibrium
- Transient behavior is also important
- No result on service distributions without decreasing hazard rate
- Existing results require Poisson arrivals

Our Goal: To develop a framework that

- Allows more general arrival and service distributions
- Sheds insight into the phase transition phenomena for general service distributions
- Captures transient behavior as well
- Can be extended to more general settings, including heterogeneous servers, thresholds, etc.

We introduce a different approach using a particle representation

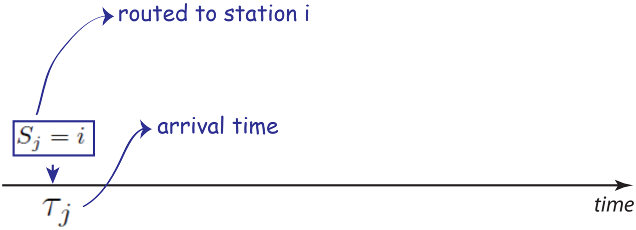
Particle Representation: The Age of a Job

The age $a_j(t)$ of job j is the time spent upto t in service

Particle Representation: The Age of a Job

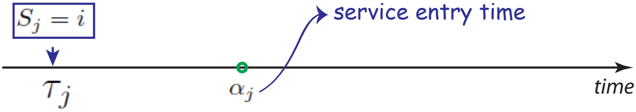
The age $a_j(t)$ of job j is the time spent upto t in service

- τ_j: arrival time of job j to network
- s_j: routing (index of chosen queue)



The age $a_j(t)$ of job j is the time spent upto t in service

- τ_j: arrival time
- S_j: routing (index of chosen queue)
- α_j: service entry time



Particle Representation: The Age of a Job

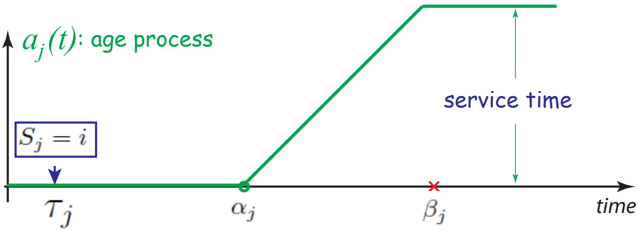
The age $a_j(t)$ of job j is the time spent upto t in service

- τ_j: arrival time
- S_j: routing (index of chosen queue)
- α_j: service entry time
- β_j: departure time

Particle Representation: The Age of a Job

The age $a_j(t)$ of job j is the time spent upto t in service

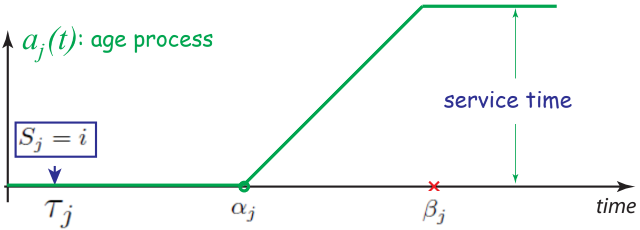
- τ_j: arrival time
- S_j: routing (index of chosen queue)
- α_j: service entry time
- β_j: departure time
- $\beta_j - \alpha_j$: service time



Particle Representation: The Age of a Job

The age $a_j(t)$ of job j is the time spent upto t in service

- τ_j: arrival time
- S_j: routing (index of chosen queue)
- α_j: service entry time
- β_j: departure time
- $\beta_j - \alpha_j$: service time

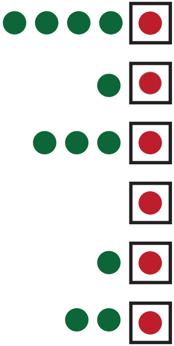


Interacting Measure-Valued Processes Representation

$\nu_\ell = \nu_\ell^N$: unit mass at ages of jobs in servers with queues of length $\geq \ell$

$$\nu_\ell^N(t) = \sum_j \delta_{a_j^N(t)},$$

where the sum is over indices of job in service at queues of length $\geq \ell$

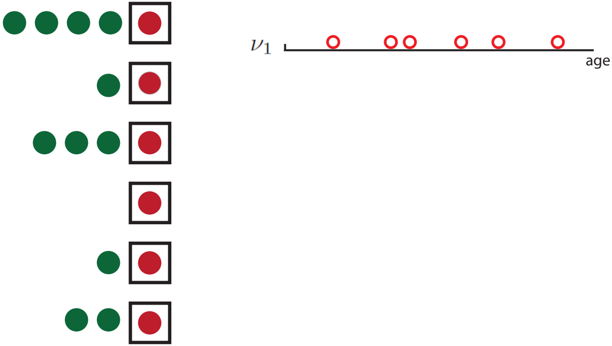


Interacting Measure-Valued Processes Representation

$\nu_\ell = \nu_\ell^N$: unit mass at ages of jobs in servers with queues of length $\geq \ell$

$$\nu_\ell^N(t) = \sum_j \delta_{a_j^N(t)},$$

where the sum is over indices of job in service at queues of length $\geq \ell$

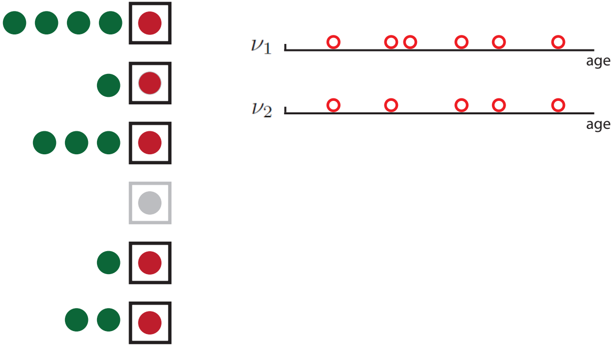


at least one job

$\nu_\ell = \nu_\ell^N$: unit mass at ages of jobs in servers with queues of length $\geq \ell$

$$\nu_\ell^N(t) = \sum_j \delta_{a_j^N(t)},$$

where the sum is over indices of job in service at queues of length $\geq \ell$



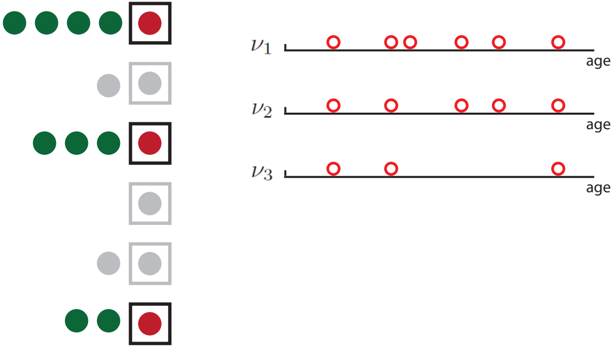
at least two jobs

Interacting Measure-Valued Processes Representation

$\nu_\ell = \nu_\ell^N$: unit mass at ages of jobs in servers with queues of length $\geq \ell$

$$\nu_\ell^N(t) = \sum_j \delta_{a_j^N(t)},$$

where the sum is over indices of job in service at queues of length $\geq \ell$

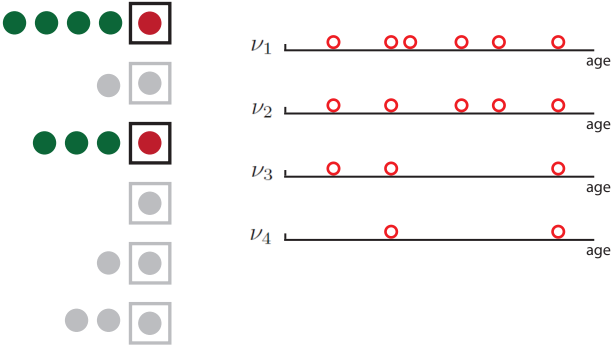


at least three jobs

$\nu_\ell = \nu_\ell^N$: unit mass at ages of jobs in servers with queues of length $\geq \ell$

$$\nu_\ell^N(t) = \sum_j \delta_{a_j^N(t)},$$

where the sum is over indices of job in service at queues of length $\geq \ell$



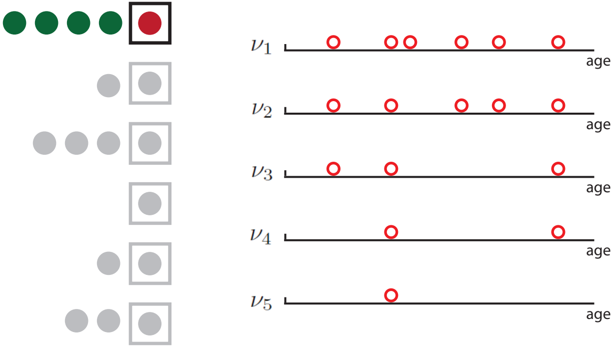
at least four jobs

Interacting Measure-Valued Processes Representation

$\nu_\ell = \nu_\ell^N$: unit mass at ages of jobs in servers with queues of length $\geq \ell$

$$\nu_\ell^N(t) = \sum_j \delta_{a_j^N(t)},$$

where the sum is over indices of job in service at queues of length $\geq \ell$

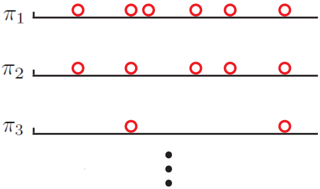


at least five jobs

Interacting Measure-Valued Processes Representation

- $\mathbb{M}_{\leq 1}[0,L)$: space of sub-probability measures on $[0,L)$ with the topology of weak convergence.
- For $\pi \in \mathbb{M}_{\leq 1}[0,L)$ and $f \in \mathbb{C}_b[0,L)$, $\langle f, \pi \rangle = \int_{[0,L)} f(x)\pi(dx)$
- $\mathcal{S}$: **space of decreasing sequences of sub-probability measures**,

$$\mathcal{S} = \{(\pi_\ell)_{\ell \geq 1} \in \mathbb{M}_{\leq 1}[0,L)^\infty | \langle f, \pi_\ell - \pi_{\ell+1} \rangle \geq 0, \forall \ell \geq 1, f \in \mathbb{C}_b[0,L)\}.$$

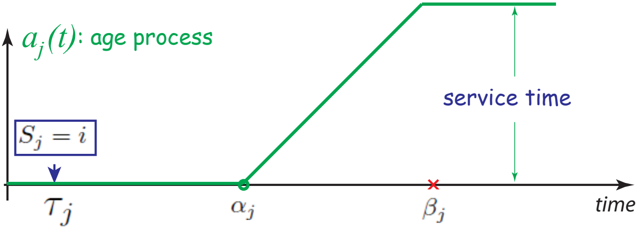


- The $\mathcal{S}$-valued process $\{\bar{\nu}^N(t) = \frac{1}{N}\left(\nu_\ell^N(t)\right)_{\ell \geq 1}; t \geq 0\}$ captures the dynamics
- $\{S_\ell^N(t) = \frac{1}{N}\langle \mathbf{1}, \nu_\ell^N(t)\rangle; \ell \geq 1, t \geq 0\}$ is Markovian in exponential case

Interacting Measure-Valued Processes Representation

Theorem 1 (Aghajani-R'14) Markovian Representation

For each $N \in \mathbb{N}$, $\{(\bar{\nu}^N_\ell(t), \ell \geq 1) : t \geq 0\}$ is a Markov process on $\mathcal{S}$ with respect to a suitable filtration $\{\mathcal{F}^N_t, t \geq 0\}$.



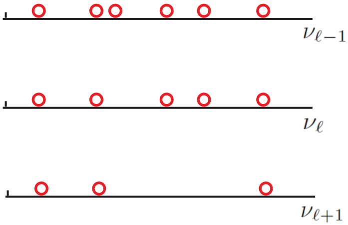
Filtration

- $\tilde{\mathcal{F}}^N_t$: information about all *events* up to time t

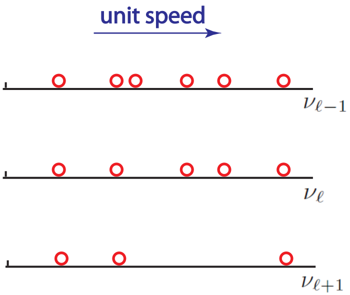
$$\tilde{\mathcal{F}}_t = \sigma\big(S_j \mathbf{1}(\tau_j \leq s), \mathbf{1}(\alpha_j \leq s), \mathbf{1}(\beta_j \leq s); j \leq 1, s \in [0,t]\big),$$

- $\{\mathcal{F}_t; t \geq 0\}$ is the associated right continuous filtration, which is completed

I. when no arrival/departure is happening, the masses move to the right with unit speed.



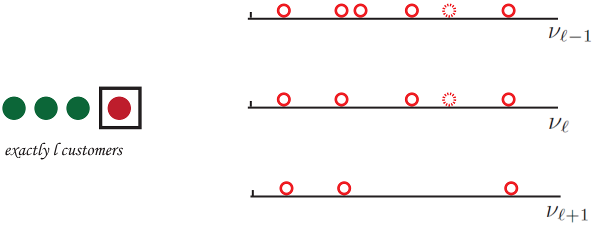
I. when no arrival/departure is happening, the masses move to the right with unit speed.



Dynamics of Measure-Valued Processes

II. Upon departure from a queue with ℓ jobs

- the corresponding mass departs from all $\nu_j, j \leq \ell$
- a new mass at zero is added to all $\nu_j, j \leq \ell - 1$ (if $\ell \geq 2$)

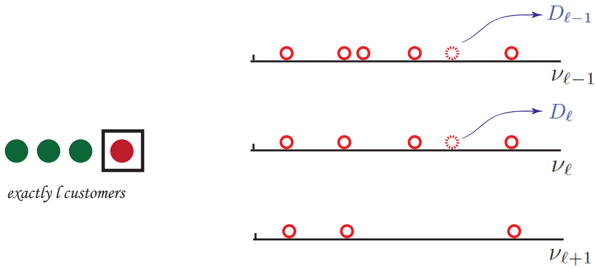


- D_ℓ: cumulative departure process from servers with at least ℓ jobs before departure.

Dynamics of Measure-Valued Processes

II. Upon departure from a queue with ℓ jobs

- the corresponding mass departs from all $\nu_j, j \le \ell$
- a new mass at zero is added to all $\nu_j, j \le \ell - 1$ (if $\ell \ge 2$)

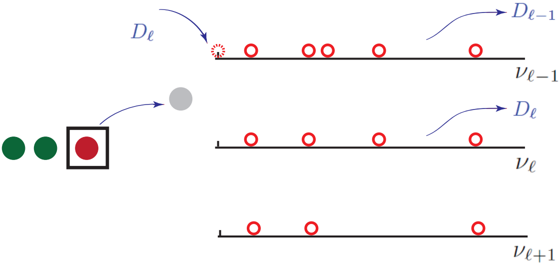


- D_ℓ: cumulative departure process from servers with at least ℓ jobs before departure.

Dynamics of Measure-Valued Processes

II. Upon departure from a queue with ℓ jobs

- the corresponding mass departs from all $\nu_j, j \leq \ell$
- a new mass at zero is added to all $\nu_j, j \leq \ell - 1$ (if $\ell \geq 2$)

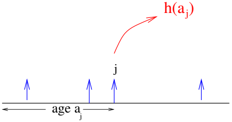


- D_ℓ: cumulative departure process from servers with at least ℓ jobs before departure.

II. Form of the cumulative departure process D_ℓ

The hazard rate function

$$h(x) = \frac{g(x)}{1 - G(x)}$$



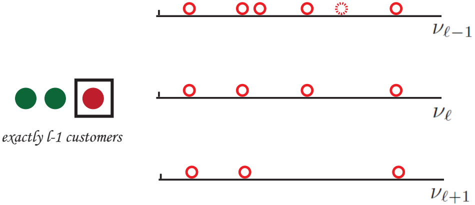
- $\langle h, \nu_\ell^{(N)}(t)\rangle = \sum_j h(a_j^N(t))$ conditional mean departure rate at time t from queues of length greater than or equal to ℓ, given ages of jobs
- the compensated departure process

$$D_\ell^N(t) - \int_0^t \langle h, \nu_\ell^N(s)\rangle \, ds$$

is a martingale (with respect to the filtration $\{\mathcal{F}_t^N\}$).

III. Upon arrival to a queue with $\ell - 1$ jobs right before arrival,

- if $\ell = 1$, a mass at zero joins ν_1
- if $\ell \geq 2$, the mass corresponding to the age of job in that particular server is added to ν_ℓ

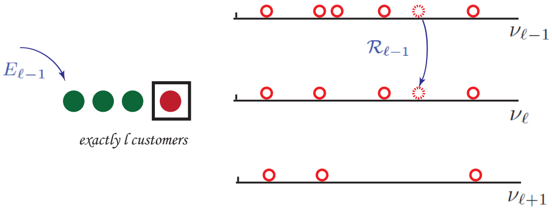


- $\mathcal{R}_\ell$: *routing measure process*

Dynamics of Measure-Valued Processes

III. Upon arrival to a queue with $\ell - 1$ jobs right before arrival,

- if $\ell = 1$, a mass at zero joins ν_1
- if $\ell \geq 2$, the mass corresponding to the age of job in that particular server is added to ν_ℓ



- $\mathcal{R}_\ell$: *routing measure process*

Routing Probabilities in the Supermarket Model

Upon arrival of j^{th} job,

- suppose queue i has ℓ jobs: $X^i = \ell$.
- ζ_j is the index of the queue to which job j is routed

$$\text{what is } \mathbb{P}\{\zeta_j = i | X^i = \ell\}?$$

1. $\mathbb{P}\{\text{queue } i \text{ has queue length} \geq \ell\} = \mathbb{P}\{\text{all picks have queue length} \geq \ell\} = S_\ell^d$.

   $$S_\ell = S_\ell^N = \frac{1}{N}\langle \mathbf{1}, \nu_\ell^N \rangle = \langle \mathbf{1}, \bar{\nu}_\ell^N \rangle : \text{fraction of queues with at least } \ell \text{ jobs}$$

2. $\mathbb{P}\{\text{queue } \zeta_j \text{ has exactly } \ell \text{ jobs}\} = S_\ell^d - S_{\ell+1}^d$.
3. Number of queues with ℓ jobs is $S_\ell - S_{\ell+1}$
4. $\mathbb{P}\{\zeta_j = i \big| X^i = \ell\} = \frac{1}{N} \frac{S_\ell^d - S_{\ell+1}^d}{S_\ell - S_{\ell+1}}$
5. When $d = 2$, $\mathbb{P}\{\zeta_j = i \big| X^i = \ell\} = \frac{1}{N}(S_\ell + S_{\ell+1})$

Hydrodynamic Limit: Assumptions

Arrival Process: Belongs to one the following two classes:

- $E^{(N)}$: (possibly time-inhomogeneous) Poisson Process with rate $\theta_N \lambda(\cdot)$ where $\theta_N / N \to 1$ as $N \to \infty$ and $\lambda(\cdot)$ is locally square integrable.
- $E^{(N)}$ is a renewal process whose interarrival distribution has a density

Service Time

has distribution G with density g and mean 1

Hydrodynamic Limit: Age Equations

Definition A process $\nu = \{\nu_\ell\}_{\ell \geq 0}$ solves the *age equations* if for all $f \in \mathbb{C}_b^1[0,\infty)$,

$$\langle f, \nu_\ell(t) \rangle = \langle f, \nu_\ell(0) \rangle$$

→ initial jobs

Hydrodynamic Limit: Age Equations

Definition A process $\nu = \{\nu_\ell\}_{\ell \geq 0}$ solves the *age equations* if for all $f \in \mathbb{C}_b^1[0,\infty)$,

$$\langle f, \nu_\ell(t)\rangle = \langle f, \nu_\ell(0)\rangle + \int_0^t \langle f', \nu_\ell(s)\rangle ds$$

linear growth of ages

Hydrodynamic Limit: Age Equations

Definition A process $\nu = \{\nu_\ell\}_{\ell \geq 0}$ solves the *age equations* if for all $f \in \mathbb{C}_b^1[0, \infty)$,

$$\langle f, \nu_\ell(t) \rangle = \langle f, \nu_\ell(0) \rangle + \int_0^t \langle f', \nu_\ell(s) \rangle ds + f(0) D_{\ell+1}(t)$$

→ service entry

Hydrodynamic Limit: Age Equations

Definition A process $\nu = \{\nu_\ell\}_{\ell \geq 0}$ solves the *age equations* if for all $f \in \mathbb{C}_b^1[0,\infty)$,

$$\langle f, \nu_\ell(t)\rangle = \langle f, \nu_\ell(0)\rangle + \int_0^t \langle f', \nu_\ell(s)\rangle ds + f(0) D_{\ell+1}(t) - \int_0^t \langle h\, f, \nu_\ell(s)\rangle ds$$

departure

Hydrodynamic Limit: Age Equations

Definition A process $\nu = \{\nu_\ell\}_{\ell \geq 0}$ solves the *age equations* if for all $f \in \mathbb{C}_b^1[0, \infty)$,

$$\langle f, \nu_\ell(t)\rangle = \langle f, \nu_\ell(0)\rangle + \int_0^t \langle f', \nu_\ell(s)\rangle ds + f(0) D_{\ell+1}(t) - \int_0^t \langle h\, f, \nu_\ell(s)\rangle ds + \int_0^t \langle f, \eta_\ell(s)\rangle ds.$$

Routing process

Hydrodynamic Limit: Age Equations

Definition A process $\nu = \{\nu_\ell\}_{\ell \geq 0}$ solves the *age equations* if for all $f \in \mathbb{C}_b^1[0,\infty)$,

$$\langle f, \nu_\ell(t)\rangle = \langle f, \nu_\ell(0)\rangle + \int_0^t \langle f', \nu_\ell(s)\rangle ds + f(0) D_{\ell+1}(t) - \int_0^t \langle h\, f, \nu_\ell(s)\rangle ds + \int_0^t \langle f, \eta_\ell(s)\rangle ds.$$

$$\langle \mathbf{1}, \nu_\ell(t)\rangle - \langle \mathbf{1}, \nu_\ell(0)\rangle = D_{\ell+1}(t) + \int_0^t \langle \mathbf{1}, \eta_\ell(s)\rangle ds - D_\ell(t),$$

→ mass balance

Hydrodynamic Limit: Age Equations

Definition A process $\nu = \{\nu_\ell\}_{\ell \geq 0}$ solves the *age equations* if for all $f \in \mathbb{C}_b^1[0,\infty)$,

$$\langle f, \nu_\ell(t)\rangle = \langle f, \nu_\ell(0)\rangle + \int_0^t \langle f', \nu_\ell(s)\rangle ds + f(0) D_{\ell+1}(t) - \int_0^t \langle h\, f, \nu_\ell(s)\rangle ds + \int_0^t \langle f, \eta_\ell(s)\rangle ds.$$

$$\langle 1, \nu_\ell(t)\rangle - \langle 1, \nu_\ell(0)\rangle = D_{\ell+1}(t) + \int_0^t \langle 1, \eta_\ell(s)\rangle ds - D_\ell(t),$$

$$D_\ell(t) = \int_0^t \langle h, \nu_\ell(s)\rangle ds,$$

departure rate

Hydrodynamic Limit: Age Equations

Definition A process $\nu = \{\nu_\ell\}_{\ell \geq 0}$ solves the *age equations* if for all $f \in \mathbb{C}_b^1[0,\infty)$,

$$\langle f, \nu_\ell(t)\rangle = \langle f, \nu_\ell(0)\rangle + \int_0^t \langle f', \nu_\ell(s)\rangle ds + f(0) D_{\ell+1}(t) - \int_0^t \langle h\, f, \nu_\ell(s)\rangle ds + \int_0^t \langle f, \eta_\ell(s)\rangle ds.$$

$$\langle \mathbf{1}, \nu_\ell(t)\rangle - \langle \mathbf{1}, \nu_\ell(0)\rangle = D_{\ell+1}(t) + \int_0^t \langle \mathbf{1}, \eta_\ell(s)\rangle ds - D_\ell(t),$$

routing measure

$$D_\ell(t) = \int_0^t \langle h, \nu_\ell(s)\rangle ds,$$

$$\eta_\ell(t) = \begin{cases} \lambda(1 - \langle \mathbf{1}, \nu_1(t)\rangle^2)\delta_0 & \text{if } \ell = 1, \\ \lambda\langle \mathbf{1}, \nu_{\ell-1}(t) + \nu_\ell(t)\rangle(\nu_{\ell-1}(t) - \nu_\ell(t)) & \text{if } \ell \geq 2. \end{cases}$$

Hydrodynamic Limit: Age Equations

Definition A process $\nu = \{\nu_\ell\}_{\ell \geq 0}$ solves the *age equations* if for all $f \in \mathbb{C}_b^1[0,\infty)$,

$$\langle f, \nu_\ell(t)\rangle = \langle f, \nu_\ell(0)\rangle + \int_0^t \langle f', \nu_\ell(s)\rangle ds + f(0) D_{\ell+1}(t) - \int_0^t \langle h\, f, \nu_\ell(s)\rangle ds + \int_0^t \langle f, \eta_\ell(s)\rangle ds.$$

$$\langle \mathbf{1}, \nu_\ell(t)\rangle - \langle \mathbf{1}, \nu_\ell(0)\rangle = D_{\ell+1}(t) + \int_0^t \langle \mathbf{1}, \eta_\ell(s)\rangle ds - D_\ell(t),$$

routing measure

$$D_\ell(t) = \int_0^t \langle h, \nu_\ell(s)\rangle ds,$$

routing probabilities

$$\eta_\ell(t) = \begin{cases} \lambda\left(1 - \langle \mathbf{1}, \nu_1(t)\rangle^2\right)\delta_0 & \text{if } \ell = 1, \\ \lambda\langle \mathbf{1}, \nu_{\ell-1}(t) + \nu_\ell(t)\rangle(\nu_{\ell-1}(t) - \nu_\ell(t)) & \text{if } \ell \geq 2. \end{cases}$$

Hydrodynamic Limit: Age Equations

Theorem 2 (Aghajani-R'14) Age Equations

Given any $\nu(0) = (\nu_\ell(0), \ell \geq 1) \in \mathbf{S}$ there exists a unique solution $\nu(\cdot) = \{(\nu_\ell(t), \ell \geq 1); t \geq 0\}$ to the age equations with initial condition $\nu(0)$.

Hydrodynamic Limit: Convergence

• Let $\{\nu^{(N)}(t) = (\nu_\ell^{(N)}(t), \ell \geq 1); t \geq 0\}$ be the measure-valued representation for the N-server system with initial condition $\nu^{(N)}(0)$.

Theorem 3 (Aghajani-R'14) Hydrodynamic Limit

If for every $\ell \geq 1$, $\nu_\ell^{(N)}(0)/N \to \nu_\ell(0)$, then

$$\frac{1}{N}\nu^{(N)}(\cdot) \Rightarrow \nu(\cdot)$$

in $\mathbf{S}$, where ν is the unique solution to the age equation corresponding to $\nu(0)$.

A Propagation of Chaos Result

Informal statement

The evolution of any subset of k queues are asymptotically independent on finite time intervals with marginal queue lengths given by the hydrodynamic equations.

Let $X^{N,i}(\cdot)$ be the process that tracks the length of the ith queue.

Theorem 4 (Aghajani-R'14) Propagation of Chaos

Suppose for each N, $\{X^{N,i}(0), i = 1, \ldots, N\}$ is exchangeable, let $\nu^N(0) \to \nu(0)$ as $N \to \infty$ and let $\nu = (\nu_\ell, \ell \geq 1)$ be the solution to the age equations associated with $\nu(0)$. Then

$$\lim_{N\to\infty} \mathbb{P}\left\{X^{N,1}(t) \geq \ell\right\} = S_\ell(t) = \langle \mathbf{1}, \nu_\ell(t) \rangle,$$

and for any $\ell_1, \ldots, \ell_k \in \mathbb{N}^k$,

$$\lim_{N\to\infty} \mathbb{P}\left\{X^{N,1}(t) \geq \ell_1, \ldots, X^{N,k}(t) \geq \ell_k\right\} = \prod_{m=1}^{k} S_{\ell_m}(t)$$

Hydrodynamics Limit: Proof of Uniqueness

Step 1:

Use (weak-sense) PDE techniques to partially solve the age equation:

Lemma (Aghajani-'R '14) Partial Solution of the Age Equations

Under suitable assumptions on D_ℓ and η_ℓ, for every $f \in \mathbb{C}_b[0,\infty)$,

$$\begin{aligned}\langle f, \nu_\ell(t)\rangle =& \langle f, \nu_\ell(0)\rangle + \int_0^t \langle f', \nu_\ell(s)\rangle ds + f(0) D_{\ell+1}(t) \\ & - \int_0^t \langle hf, \nu_\ell(s)\rangle ds + \int_0^t \langle f, \eta_\ell(s)\rangle ds \end{aligned} \qquad (1)$$

holds if and only if

$$\begin{aligned}\langle f, \nu_\ell(t)\rangle =& \langle f(\cdot + t)\frac{\bar{G}(\cdot + t)}{\bar{G}(\cdot)}, \nu_\ell(0)\rangle + \int_{[0,t]} f(t-s)\bar{G}(t-s) dD_{\ell+1}(s) \\ & + \int_0^t \langle f(\cdot + t - s)\frac{\bar{G}(\cdot + t - s)}{\bar{G}(\cdot)}, \eta_\ell(s)\rangle ds \end{aligned} \qquad (2)$$

Definition. We refer to equation (2):

$$\langle f, \nu_\ell(t)\rangle = \langle f(\cdot + t)\frac{\bar{G}(\cdot + t)}{\bar{G}(\cdot)}, \nu_\ell(0)\rangle + \int_{[0,t]} f(t-s)\bar{G}(t-s)dD_{\ell+1}(s) + \int_0^t \langle f(\cdot + t - s)\frac{\bar{G}(\cdot + t - s)}{\bar{G}(\cdot)}, \eta_\ell(s)\rangle ds$$

and the remaining age equations, (3)–(5) below, as the Hydrodynamics Equations.

$$\langle \mathbf{1}, \nu_\ell(t)\rangle - \langle \mathbf{1}, \nu_\ell(0)\rangle = D_{\ell+1}(t) + \int_0^t \langle \mathbf{1}, \eta_\ell(s) ds - D_\ell(t), \qquad (3)$$

with

$$D_\ell(t) = \int_0^t \langle h, \nu_\ell(s)\rangle ds \qquad (4)$$

and

$$\eta_\ell(t) = \begin{cases} \lambda(1 - \langle \mathbf{1}, \nu_1(t)\rangle^2)\delta_0 & \text{if } \ell = 1, \\ \lambda\langle \mathbf{1}, \nu_{\ell-1}(t) + \nu_\ell(t)\rangle(\nu_{\ell-1}(t) - \nu_\ell(t)) & \text{if } \ell \geq 2. \end{cases} \qquad (5)$$

Show that these hydrodynamic equations have a unique solution.

- Consider the special class of functions $\mathbb{F}$

$$\mathbb{F} = \left\{ \frac{\bar{G}(\cdot + r)}{\bar{G}(\cdot)} : r \geq 0 \right\}.$$

- Show that the class of functions is (in a suitable sense) invariant under the hydrodynamic equation (2)

$$\begin{aligned}\langle f, \nu_\ell(t)\rangle =& \langle f(\cdot + t)\frac{\bar{G}(\cdot + t)}{\bar{G}(\cdot)}, \nu_\ell(0)\rangle + \int_{[0,t]} f(t-s)\bar{G}(t-s)dD_{\ell+1}(s) \\ &+ \int_0^t \langle f(\cdot + t - s)\frac{\bar{G}(\cdot + t - s)}{\bar{G}(\cdot)}, \eta_\ell(s)\rangle ds\end{aligned}$$

- Show uniqueness first for this class of functions $f \in \mathbb{F}$ and then show that this implies uniqueness for all $f \in \mathcal{C}_b[0, L)$.

Hydrodynamic Limit: Proof of Convergence

Skipping details and some subtleties ...

- Identify compensators of various processes à la Baccelli-Bremaud
- Establish tightness
- Show convergence

We have obtained a general convergence result and characterized the limit.

So what ?

What can one do with this measure-valued hydrodynamic limit?
Can one use it to compute anything ?

A PDE representation

- If one is only interested in $S_\ell(t) = \langle \mathbf{1}, \nu_\ell(t) \rangle$, one can get a simpler representation.

Define

$$f^r(x) = \frac{\bar{G}(x+r)}{\bar{G}(x)} \qquad \xi_\ell(t,r) = \langle f^r, \nu_\ell(t) \rangle$$

and note that

$$\S_\ell(t) = xi_\ell(t,0) \quad \text{and} \quad \langle h, \nu_\ell(t) \rangle = -\partial_r \xi_\ell(t,0).$$

Theorem 5 (Aghajani-R '15)

Suppose, in addition, we assume time-varying Poisson arrivals and bounded hazard rate function. If ν solves the age equations associated with $\nu(0)$, then $\xi(\cdot,\cdot) = \{\xi_\ell(\cdot,\cdot), \ell \geq 1\}$ is the unique solution to a certain system of PDEs.

Details of the PDE representation

Recall

$$f^r(x) = \frac{\bar{G}(x+r)}{\bar{G}(x)} \qquad \xi_\ell(t,r) = \langle f^r, \nu_\ell(t) \rangle$$

and

$$S_\ell(t) = \xi_\ell(t,0) \quad \text{and} \quad \langle h, \nu_\ell(t) \rangle = -\partial_r \xi_\ell(t,0).$$

Then (for $d = 2$) the "PDE" takes the following form: for $t > 0$

$$\begin{aligned}\xi_\ell(t,r) =& \xi_\ell(0,t+r) - \int_0^t \bar{G}(t+r-u)\partial_r \xi_{\ell+1}(u,0)du,\\ &+\lambda \int_0^t \left(\xi_{\ell-1}(u,0) + \xi_\ell(u,0)\right)\left(\xi_{\ell-1}(u,t+r-u) - \xi_\ell(u,t+r-u)\right) du\end{aligned}$$

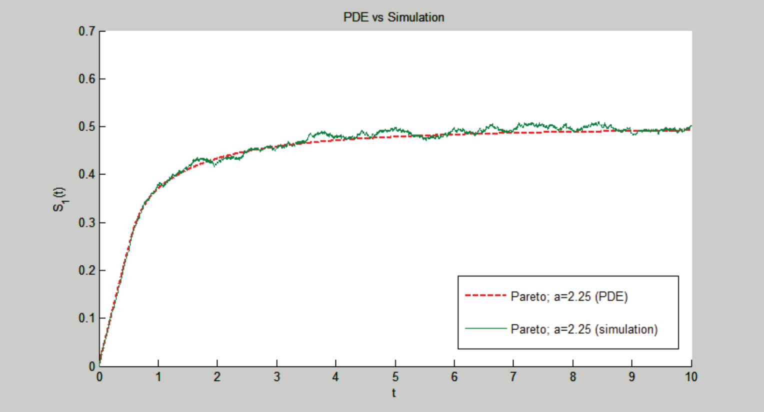
with boundary condition

$$\xi_\ell(t,0) - \xi_\ell(0,0) = \int_0^t \left(\lambda(u)\left(\xi_{\ell-1}(u,0)^2 - \xi_\ell(u,0)^2\right) - \left(\partial_r \xi_{\ell-1}(u,0) - \partial_r \xi_\ell(u,0)\right)\right) du$$

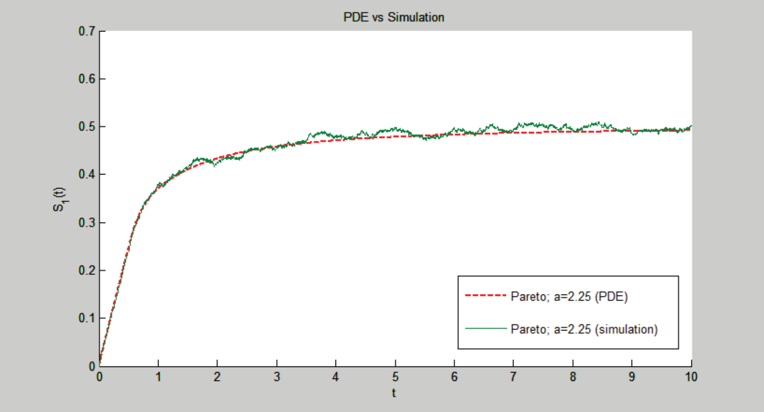
- This system of PDEs can be numerically solved to provide approximations to performance measures of the network.
- The class of functionals represented by $\{\xi_\ell(\cdot,\cdot), \ell \geq 1\}$ is rich enough to include both the queue length and the virtual waiting time

Simulation Results

We can numerically solve the PDE and compare the results to simulations

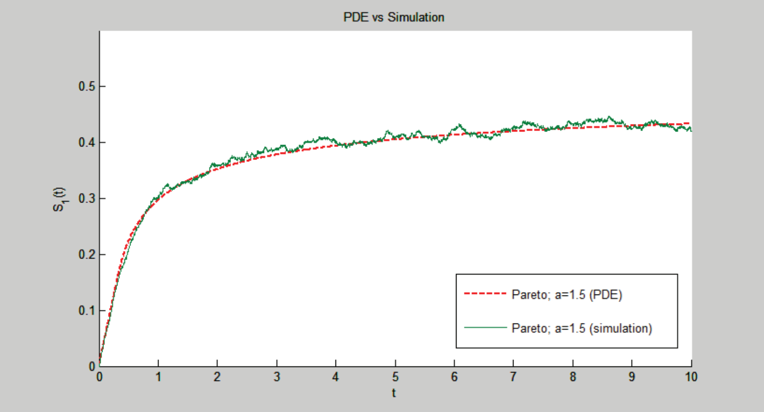


We can numerically solve the PDE and compare the results to simulations



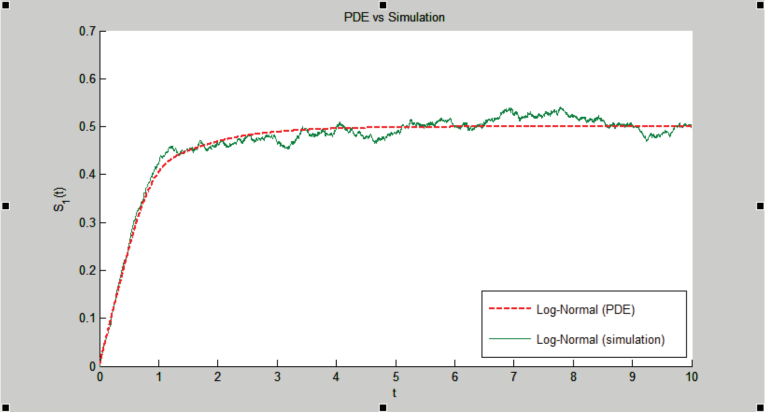
Simulation Results

We can numerically solve the PDE and compare the results to simulations



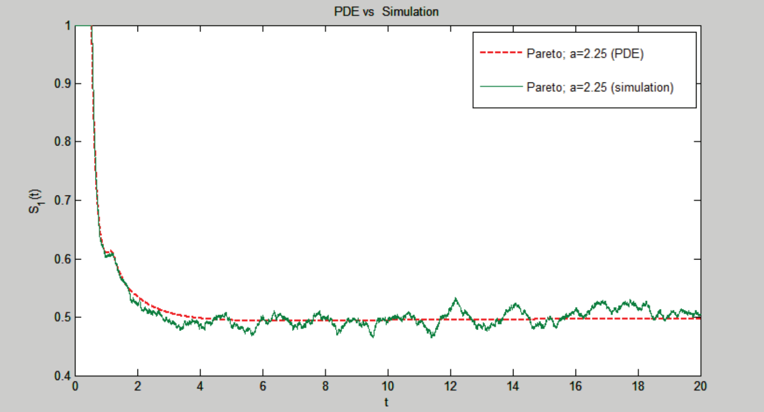
Simulation Results

We can numerically solve the PDE and compare the results to simulations



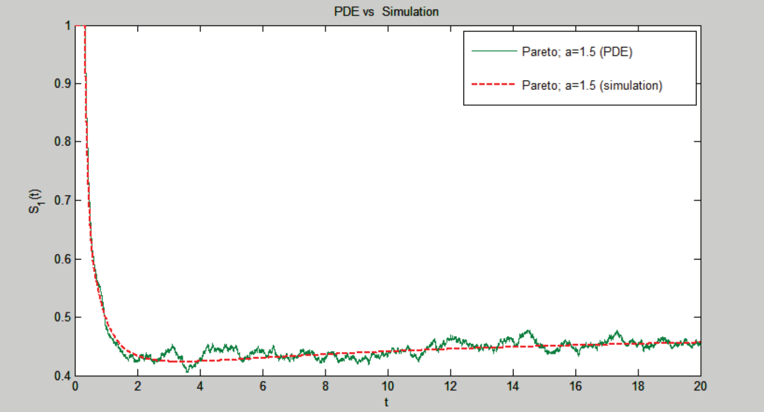
Simulation Results

We can numerically solve the PDE and compare the results to simulations



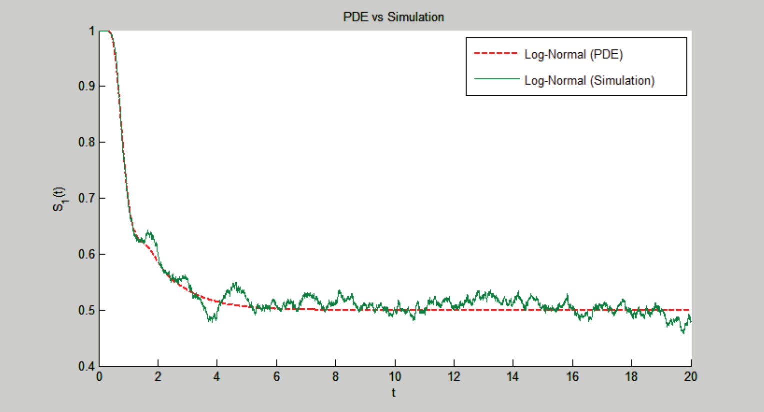
Simulation Results

We can numerically solve the PDE and compare the results to simulations



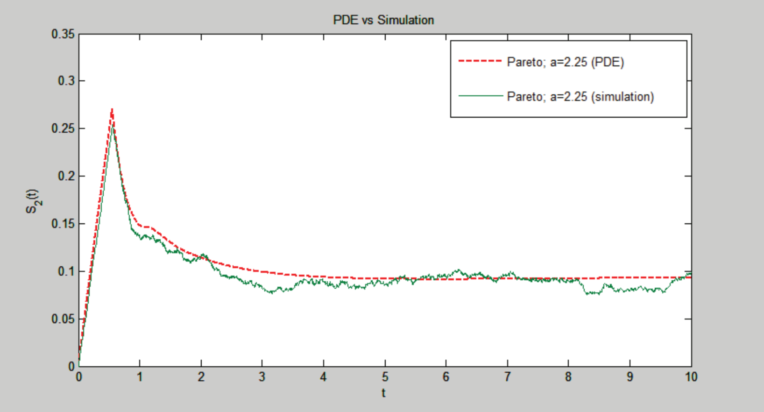
Simulation Results

We can numerically solve the PDE and compare the results to simulations

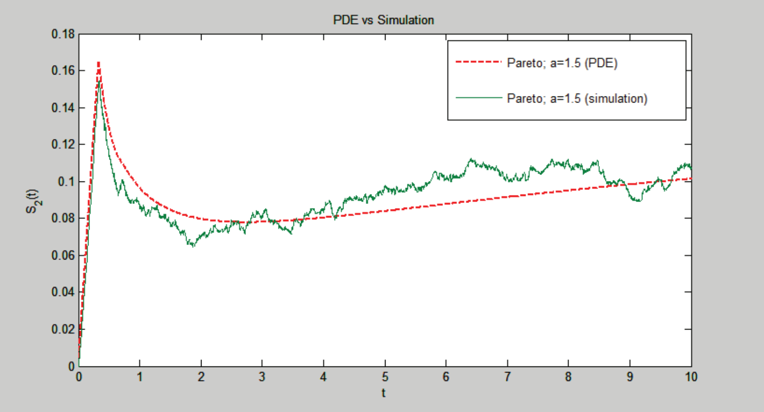


Simulation Results

We can numerically solve the PDE and compare the results to simulations

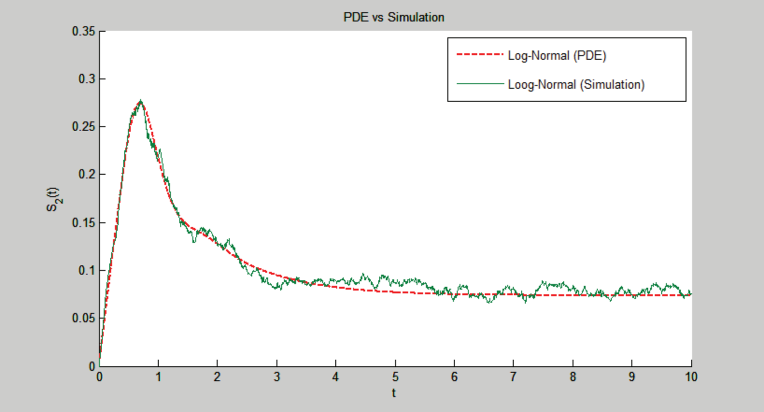


We can numerically solve the PDE and compare the results to simulations



Simulation Results

We can numerically solve the PDE and compare the results to simulations



Summary of Results

We introduced a framework for the analysis of load balancing algorithms, featuring

- Hydrodynamic limit which captures transient behavior
- Applicable to general service distributions
- Incorporates more general time varying arrival processes
- Propagation of chaos on the finite interval was established

For Exponential service distribution:

- limit process is characterized by the solution to a sequence of ODEs

For General service distribution:

- limit process is characterized by the solution to a sequence of PDEs
- Equilibrium distributions are characterized by the fixed point of the PDEs
- We can also show that uniqueness of fixed points of the PDE imply propagation of chaos on the infinite interval

Concluding Remarks

Interacting measure-valued processes framework

- Obtained a PDE that provides more efficient alternative to simulations in order to address network optimization and design questions
- Applicable for modifications of this randomized load balancing algorithm
- Can be applied to the analysis of the Serve the Longest Queue (SLQ)-type service disciplines [Ramanan, Ganguly, Robert]
- The framework can be used for other non-queueing models arising in materials science

Other Questions

- Ongoing: Analysis of fixed points of the PDE to gain insight into the stationary distribution and phase transition (ongoing)
- Implications for rate of convergence to stationary distribution
- More on Numerical solution for the PDEs