A mean-field approach to chaining networks

Alexander Aurell [1]   Göran Svensson [2,3]

[1] Division of mathematical statistics, KTH, Stockholm, Sweden

[2] Division of optimization and systems theory, KTH, Stockholm, Sweden

[3] Teleopti AB, Stockholm, Sweden

October 19, 2017

Presentation Overview

1. Chaining Networks
   - Queueing Theory
   - Chaining Systems
   - Scope and Quality of Service Measures
2. Mean-Field Approximation
   - Short Introduction
   - The Transition Rate Matrix Q
3. Example and Validation
4. Summary and Conclusions

Some Queueing Theory

The Markov chain representing a $M/M/n$ queue and a $M/M/n/n$ queue:

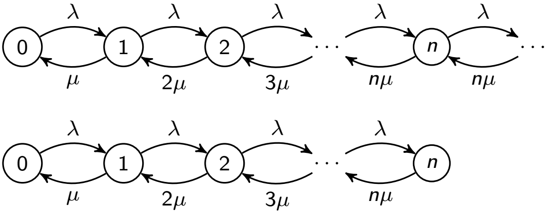


Figure 1: Markovian multi-server queues with n servers. Above with infinite buffer and below with blocking.

I. Stationary state
II. The arrivals follow a Poisson process with parameter λ
III. The service times are Exponentially distributed with parameter μ per server
IV. There are n servers

Multi-Class, Multi-Server Queues

We consider queueing networks with several job types (multi-class) and many servers (multi-server). When some servers handle multiple-classes of job types the system cannot be separated for each class and the Markov chain becomes more involved.

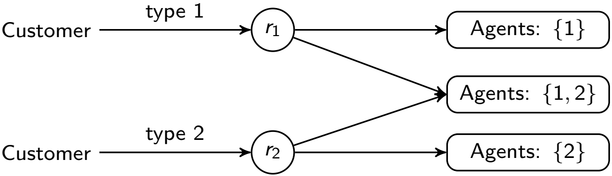


Figure 2: Queueing network with 3 agent pools, where one pool serves both types of customers.

Chaining in Short

Description and Motivation
In a chaining network each server(agent) is trained to handle two types of jobs(customers). The agents are grouped in agent pools that serve the same types of customers.

Chaining in Short

Description and Motivation
In a chaining network each server(agent) is trained to handle two types of jobs(customers). The agents are grouped in agent pools that serve the same types of customers.

All customer types can be served by exactly two agent pools.

Chaining in Short

Description and Motivation
In a chaining network each server(agent) is trained to handle two types of jobs(customers). The agents are grouped in agent pools that serve the same types of customers.

All customer types can be served by exactly two agent pools.

Such a system offers most of the benefits of the fully flexible network but to a fraction of the training cost.

Chaining in Short

Description and Motivation
In a chaining network each server(agent) is trained to handle two types of jobs(customers). The agents are grouped in agent pools that serve the same types of customers.

All customer types can be served by exactly two agent pools.

Such a system offers most of the benefits of the fully flexible network but to a fraction of the training cost.

Furthermore, the chaining structure offers a high degree of robustness, choke points can be spread out.

Chaining Networks

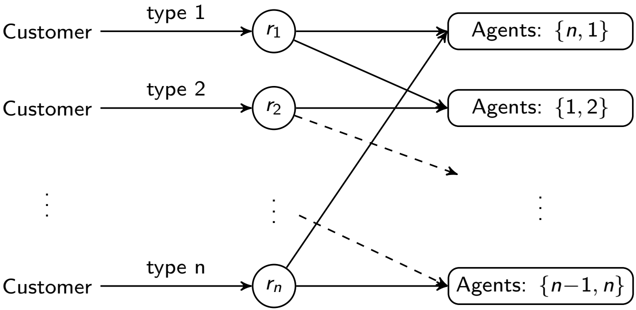


Figure 3: Queueing network of chaining type with n agent pools and customer types.

Chaining Literature

William C. Jordan and Stephen C. Graves, *Principles on the Benefits of Manufacturing Process Flexibility*, Management Science, pages 577–594, volume 41, number 4, year 1995.

Inman, R. R. and Jordan, W. C. and Blumenfeld, D. E., *Chained cross-training of assembly line workers*, International Journal of Production Research, pages 1899–1910, volume 42, number 10, year 2004

Model Assumptions and Parameters

Assumptions

I. Each agent can handle two types of jobs,
II. Agents in the same pool are trained on the same job types,
III. Agents with the same skill-set are indistinguishable,
IV. Given routing rule,
V. Piece-wise stationary demands.

Parameters

I. n different job types and agent pools,
II. Arrivals $\sim Po(\lambda_i),\ i = 1, \ldots, n$,
III. Service times $\sim Exp(\mu_{i,j}),\ i, j \in \{1, \ldots, n\}$.

Chaining Network, revisited

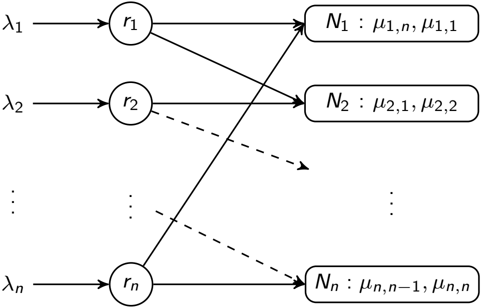

λ_1
λ_2
$\vdots$
λ_n
r_1
r_2
$\vdots$
r_n
$N_1 : \mu_{1,n}, \mu_{1,1}$
$N_2 : \mu_{2,1}, \mu_{2,2}$
$\vdots$
$N_n : \mu_{n,n-1}, \mu_{n,n}$

Figure 4: Queueing network of chaining type with n agent pools. Each pool is staffed by N_i agents with service rates given by $\mu_{i,j}$.

Scope and Measure

Scope:

1. We consider the limited case of a chaining network without a buffer[1].
2. The routing rule is assumed to be a function of the state of all servers, symmetric over each pool[2].
3. In our test case the routing rule is taken to be a coin flip between two pools conditioned on there being idle agents.

[1] Which can be compared to Erlang-B or Engset queues but in a multi-skill environment.

[2] It only matters which pool the server is located in, it is indifferent to the enumeration of the servers within each pool.

[3] A fairness of service measure can also be defined, were different customer types get served in some fair sense.

Scope and Measure

Scope:

1. We consider the limited case of a chaining network without a buffer[1].
2. The routing rule is assumed to be a function of the state of all servers, symmetric over each pool[2].
3. In our test case the routing rule is taken to be a coin flip between two pools conditioned on there being idle agents.

Quality of Service Measure:
A measure of how well the queueing network performs is needed to optimize the system.

The blocking formulation of the current problem provides a natural Quality of Service (QoS) measure, namely the blocking probability [3], i.e., what are the chances a newly arrived customer finds the system fully occupied?

[1] Which can be compared to Erlang-B or Engset queues but in a multi-skill environment.

[2] It only matters which pool the server is located in, it is indifferent to the enumeration of the servers within each pool.

[3] A fairness of service measure can also be defined, were different customer types get served in some fair sense.

Mean Field: short introduction

The heuristic behind a mean-field approximation is as follows:

> *If the network is made up of many small[4] entities, "particles", interacting in a symmetric way, each particle can be approximated by a representative, interacting with the mean-field.*

For certain routing rules, the chaining network can be decomposed into small components that interact through a symmetric function of the states.

[4] In this context, small means that each particle has minor influence, negligible in the large system limit, on any other particle.

Mean Field: short introduction

The heuristic behind a mean-field approximation is as follows:

> *If the network is made up of many small[4] entities, "particles", interacting in a symmetric way, each particle can be approximated by a representative, interacting with the mean-field.*

For certain routing rules, the chaining network can be decomposed into small components that interact through a symmetric function of the states.

From the microscopic interactions of these components, that give rise to the system dynamics, the mean-field approximation is derived.

Replacing the inter-component interaction with a mean field greatly reduces complexity of the problem, for N components with state space S, the system size reduction is

$$|S|^N \to N|S|. \tag{1}$$

[4] In this context, small means that each particle has minor influence, negligible in the large system limit, on any other particle.

Mean Field: short introduction

Queue length, mean and other distributional properties

Vvedenskaya, Nikita Dmitrievna and Dobrushin, Roland L'vovich and Karpelevich, Fridrikh Izrailevich, *Queueing system with selection of the shortest of two queues: An asymptotic approach*, Problemy Peredachi Informatsii, pages 20–34, volume 32, number 1, year 1996.

Vvedenskaya, Nikita D and Suhov, Yuri M, Fridrikh Izrailevich, *Dobrushin's mean-field approximation for a queue with dynamic routing*, INRIA, year 1997.

Dawson, Donald A and Tang, Jiashan and Zhao, Yiqiang Q, *Balancing queues by mean field interaction*, Queueing Systems, pages 335–361, volume 49, number 3-4, year 2005.

Stolyar, Alexander L *Pull-based load distribution in large-scale heterogeneous service systems* Queueing Systems, pages 341–361, volume 80, number 4, year 2015.

Mean Field: short introduction

Large system limit dynamics

Kurtz, Thomas G *Solutions of ordinary differential equations as limits of pure jump Markov processes*, Journal of applied Probability, pages 49–58, volume 7, number 1, year 1970.

Oelschlager, Karl *A martingale approach to the law of large numbers for weakly interacting stochastic processes* The Annals of Probability, pages 458–479, year 1984.

Bobbio, Andrea and Gribaudo, Marco and Telek, Miklós, *Analysis of large scale interacting systems by mean field method*, QEST'08. Fifth International Conference on Quantitative Evaluation of Systems, pages 215–224, year 2008.

Mean Field: Decomposition of the network into particles

Consider the following part of the queuing network, the "type 1 group":

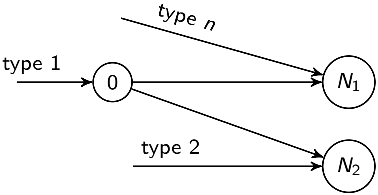


Figure 5: Arrivals to a type 1 group.

We will split this group into N_1 small components, "particles":

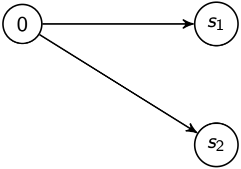


Figure 6: Typical mean-field particle, where $s_1 \in N_1$ and $s_2 \in N_2$.

Mean Field: A convergence result

Let the N_ℓ particles in the type ℓ group take values in the finite set $\mathcal{X} = \{1, \ldots, d\}$ and let $X^{k,N_\ell}(t)$ be the state of the k^{th} particle at time t. The empirical measure associated with the particles is

$$\mathbb{P}_t^{N_\ell}(\omega) := \frac{1}{N_\ell} \sum_{k=1}^{N_\ell} \delta_{X^{k,N_\ell}(t,\omega)}. \tag{2}$$

The rate at which a particle changes state from i to j at time t is assumed to be $Q_{ij}^{N_\ell}(\mathbb{P}_t^{N_\ell})$.

Theorem

Assume that there is a Lipschitz function Q_{ij} such that if $x_{N_\ell} \to x$ then $Q_{ij}^{N_\ell}(x_n) \to Q_{ij}(x)$ and assume that $\mathbb{P}_0^{N_\ell}$ converges in probability to some p as $N_\ell \to \infty$. Then $(\mathbb{P}_t^{N_\ell}; t \geq 0)$ converges in probability to $(\mathbb{P}_t; t \geq 0)$, the unique solution to

$$\dot{\mathbb{P}}_t = \mathbb{P}_t Q(\mathbb{P}_t), \quad \mathbb{P}_0 = p. \tag{3}$$

Constructing the transition rate matrix Q

Let $p_\ell(t)$ denote the fraction of agents not occupied in pool ℓ at time t, which can be written in terms of the empirical measure,

$$p_\ell(t) := \int_{\mathcal{X}} 1_{\{\text{not occupied}\}}(x)\mathbb{P}_t^{N_\ell}(dx). \tag{4}$$

It is through this quantity the particles will interact! Notice the negligible influence of the state of a single particle on p_ℓ when N_ℓ is large.

[5] Except for the last particle type, n, where a were arrivals can be routed to pool n or 1.

[6] Except for particles 1 and n.

Constructing the transition rate matrix Q

Let $p_\ell(t)$ denote the fraction of agents not occupied in pool ℓ at time t, which can be written in terms of the empirical measure,

$$p_\ell(t) := \int_{\mathcal{X}} 1_{\{\text{not occupied}\}}(x)\mathbb{P}_t^{N_\ell}(dx). \tag{4}$$

It is through this quantity the particles will interact! Notice the negligible influence of the state of a single particle on p_ℓ when N_ℓ is large.

To obtain the arrival rates we need some preliminaries. Let

$$\begin{cases} \tilde{\lambda}_\ell := \frac{\lambda_\ell}{N_\ell + N_{\ell+1}}, & \ell = 1, \dots, n-1, \\ \tilde{\lambda}_n := \frac{\lambda_n}{N_n + N_1}, \end{cases} \tag{5}$$

be the arrival rates per server of type $\ell = 1, \dots, n$.

[5] Except for the last particle type, n, where a were arrivals can be routed to pool n or 1.

[6] Except for particles 1 and n.

Under a routing rule $R(p_\ell(t))$, the arrival rates to a particle at time t are

$$\begin{array}{cccccl}
\text{pool} & \text{customer type} & \text{arrival rate} & & & \\
\ell & \ell-1 & a_{\ell,\ell-1} & = & \frac{R(p_\ell(t))\tilde{\lambda}_{\ell-1}(N_{\ell-1}+N_\ell)}{N_\ell p_\ell(t)} & \\
\ell & \ell & a_{\ell,\ell} & = & \frac{R(p_\ell(t))\tilde{\lambda}_{\ell}(N_{\ell}+N_{\ell+1})}{N_\ell p_\ell(t)} & \qquad (6) \\
\ell+1 & \ell & a_{\ell+1,\ell} & = & \frac{R(p_{\ell+1}(t))\tilde{\lambda}_{\ell}(N_{\ell}+N_{\ell+1})}{N_{\ell+1} p_{\ell+1}(t)} & \\
\ell+1 & \ell+1 & a_{\ell+1,\ell+1} & = & \frac{R(p_{\ell+1}(t))\tilde{\lambda}_{\ell+1}(N_{\ell+1}+N_{\ell+2})}{N_{\ell+1} p_{\ell+1}(t)} &
\end{array}$$

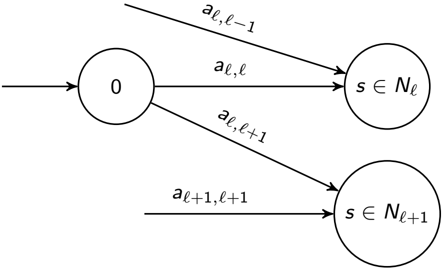


Figure 7: Arrivals to a type ℓ particle.

Constructing the transition rate matrix Q

For each group, the transition rates satisfy the convergence criterion in the theorem. For the n-group multi-particle system the transition rate matrix Q is formed by non-communicating submatrices $Q_i, i = 1, \ldots, n$, as

$$Q = \begin{bmatrix} Q_1 & 0 & \ldots & 0 \\ 0 & Q_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & Q_n \end{bmatrix}. \tag{7}$$

Constructing the transition rate matrix Q

For each group, the transition rates satisfy the convergence criterion in the theorem. For the n-group multi-particle system the transition rate matrix Q is formed by non-communicating submatrices $Q_i, i = 1, \ldots, n$, as

$$Q = \begin{bmatrix} Q_1 & 0 & \ldots & 0 \\ 0 & Q_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & Q_n \end{bmatrix}. \tag{7}$$

In the job type grouping, each server will belong to at least two particles in separate groups. For the model to be well-defined, the probability of a server to be in a certain state must be equal across the groups it belongs to! Therefore, there is a set of consistency equations that must be satisfied.

Example: 2-chain with 3 pools

Consider the chaining system with three agent pools and three customer types.

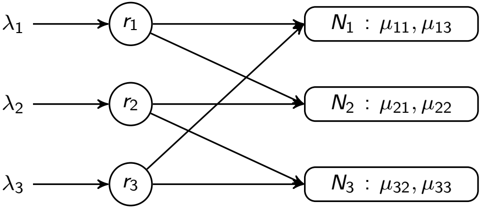

λ_1
λ_2
λ_3
r_1
r_2
r_3
$N_1 : \mu_{11}, \mu_{13}$
$N_2 : \mu_{21}, \mu_{22}$
$N_3 : \mu_{32}, \mu_{33}$

Figure 8: Example of a queueing network of chaining type with three agent pools and corresponding staffing of N_1, N_2 and N_3. The service rates are given by $(\mu_{11}, \mu_{13}), (\mu_{21}, \mu_{22})$ and (μ_{32}, μ_{33}) and the arrival rates to the system are λ_1, λ_2 and λ_3.

Job type grouping

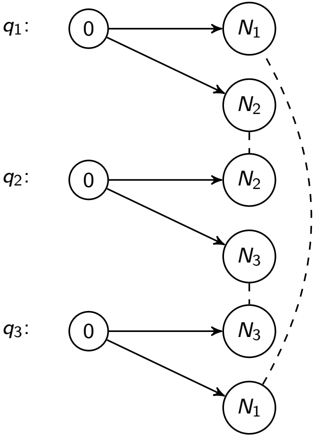


Figure 9: There will be three types of mean field-particles. The arrivals are represented by arrows and the consistency conditions are represented by dashed lines.

QoS: Blocking Probability

An incoming customer is blocked if there are no available servers with the skill to serve it. The Quality of Service measure will be the probability of being blocked. The blocking probability of a type ℓ customer at time t can be expressed as

$$\mathbb{P}\left(X_t^{i,N_\ell} \text{ fully occupied for all } i \in \{1, \dots, N_\ell\}\right). \tag{8}$$

The blocking event can not be captured by the mean-field distribution $\mathbb{P}_t$.

QoS: Blocking Probability

An incoming customer is blocked if there are no available servers with the skill to serve it. The Quality of Service measure will be the probability of being blocked. The blocking probability of a type ℓ customer at time t can be expressed as

$$\mathbb{P}\left(X_t^{i,N_\ell} \text{ fully occupied for all } i \in \{1, \dots, N_\ell\}\right). \tag{8}$$

The blocking event can not be captured by the mean-field distribution $\mathbb{P}_t$.

Assume that the number of occupied servers is $\text{Po}(m_t^\ell)$-distributed when $p_\ell(\mathbb{P}_t) \approx 1/2$, where $m_t^\ell := N_\ell(1 - p_\ell(\mathbb{P}_t))$ is the number of occupied servers. The $\text{Po}(m_t^\ell)$ distribution can be approximated by the normal distribution $\mathcal{N}\left(m_t^\ell, m_t^\ell\right)$ when N_ℓ is large.

QoS: Blocking Probability

An incoming customer is blocked if there are no available servers with the skill to serve it. The Quality of Service measure will be the probability of being blocked. The blocking probability of a type ℓ customer at time t can be expressed as

$$\mathbb{P}\left(X_t^{i,N_\ell} \text{ fully occupied for all } i \in \{1,\ldots,N_\ell\}\right). \tag{8}$$

The blocking event can not be captured by the mean-field distribution $\mathbb{P}_t$.

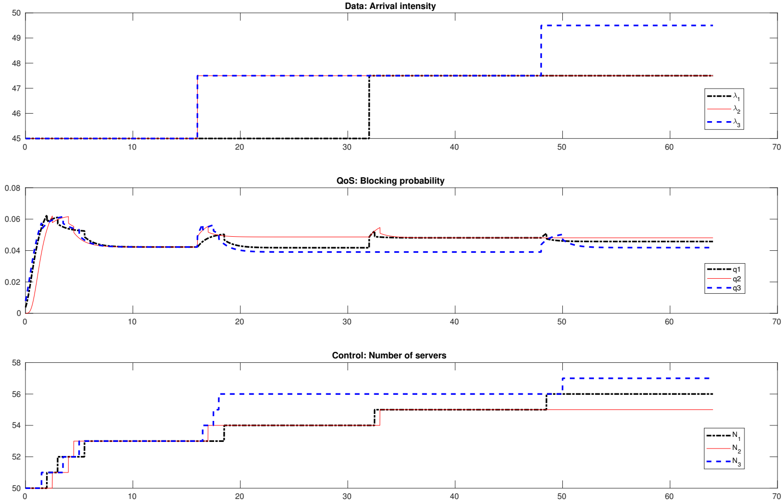
Assume that the number of occupied servers is $\text{Po}(m_t^\ell)$-distributed when $p_\ell(\mathbb{P}_t) \approx 1/2$, where $m_t^\ell := N_\ell(1-p_\ell(\mathbb{P}_t))$ is the number of occupied servers. The $\text{Po}(m_t^\ell)$ distribution can be approximated by the normal distribution $\mathcal{N}\left(m_t^\ell, m_t^\ell\right)$ when N_ℓ is large.

This is not the whole picture! The distribution is confined to the domain $[0, N_\ell]$. When $p_\ell(\mathbb{P}_t)$ is close to 0, mass gathers around the barrier N_ℓ. Intuitively, there is a "reflection" at $x = N_\ell$. Inspired by this, we assume a mixture model for the distribution of the number of occupied servers

$$\alpha\mathcal{N}(m_t^\ell, m_t^\ell) + (1-\alpha)\mathcal{N}_{\text{refl}}(m_t^\ell, m_t^\ell, N_\ell) \tag{9}$$

where α is λ_ℓ/N_ℓ-dependent.

Numerical Experiment: Control of the QoS



Numerical Experiment: Control of the QoS

A comparison of the mixture model with the empirical density constructed from 3000 simulations of the chaining network.

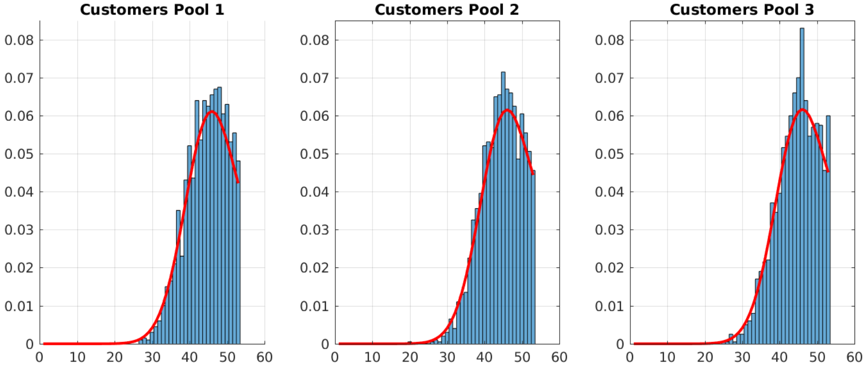


Figure 10: Right before the 1st jump when $(\lambda_1, \lambda_2, \lambda_3) = (45, 45, 45)$ and $(N_1, N_2, N_3) = (53, 53, 53)$.

Numerical Experiment: Control of the QoS

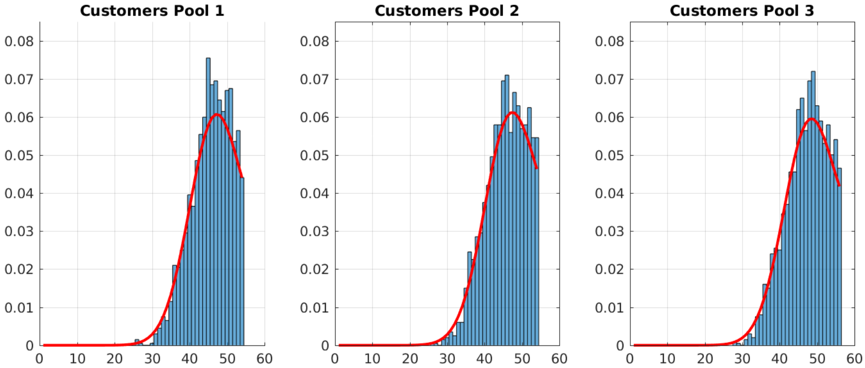


Figure 11: Right before the 2nd jump when $(\lambda_1, \lambda_2, \lambda_3) = (45, 47.5, 47.5)$ and $(N_1, N_2, N_3) = (54, 54, 56)$.

Numerical Experiment: Control of the QoS

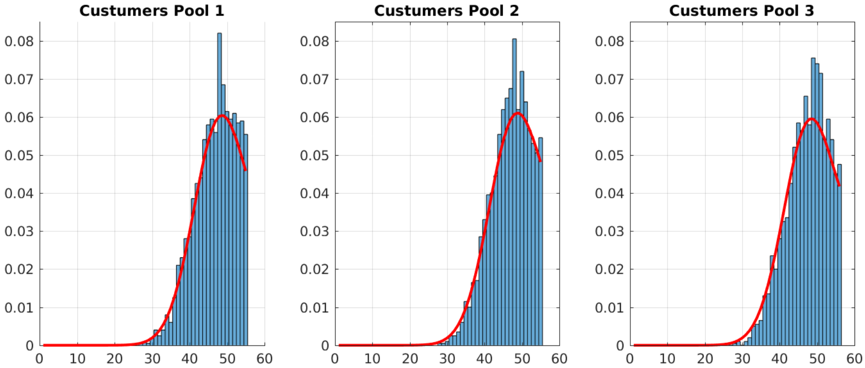


Figure 12: Right before the 3rd jump when $(\lambda_1, \lambda_2, \lambda_3) = (47.5, 47.5, 47.5)$ and $(N_1, N_2, N_3) = (55, 55, 56)$.

Numerical Experiment: Complexity

For a fixed time interval but various pool size, we solve the mean-field approximation. For comparison, we simulate the chaining network for the same settings.

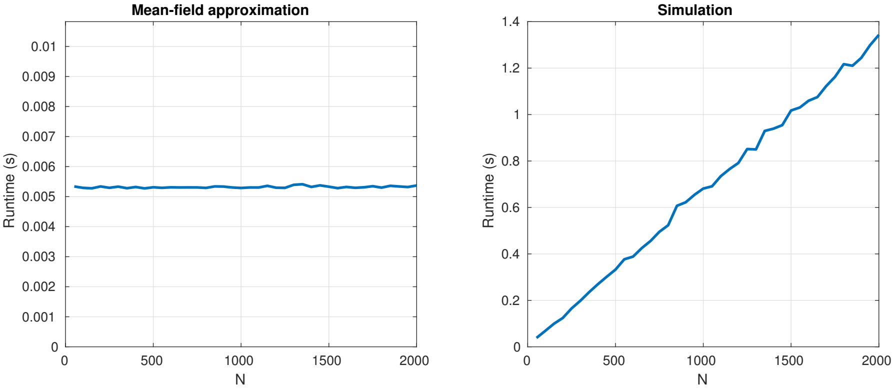


Figure 13: Runtime (evaluation of QoS excluded).

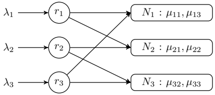
Summary and Conclusions

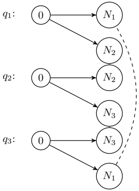
To the best of our knowledge, a novel approach to analyze chaining networks.

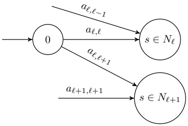
The mean-field approach could offer a fast way of controlling the QoS in a chaining network, with excellent scaling properties.

The limiting density of occupied servers is non-trivial, and this resulted in problems estimating the QoS. Similar problems are likely to arise when dealing with other QoS measures.

We have not analyzed the most general chaining network. Extensions should include waiting lines as well as other network configurations and routing rules.


λ_1
λ_2
λ_3
r_1
r_2
r_3
$N_1 : \mu_{11}, \mu_{13}$
$N_2 : \mu_{21}, \mu_{22}$
$N_3 : \mu_{32}, \mu_{33}$


q_1:
0
N_1
N_2
q_2:
0
N_2
N_3
q_3:
0
N_3
N_1


$a_{\ell,\ell-1}$
$a_{\ell,\ell}$
$a_{\ell,\ell+1}$
$a_{\ell+1,\ell+1}$
0
$s \in N_\ell$
$s \in N_{\ell+1}$