Geometric lower bounds for stochastic processing networks with limited connectivity

Diego Goldsztajn and Andres Ferragut

Universidad ORT Uruguay

SPOR Seminar – TU/Eindhoven – June 2025

Outline

Stochastic processing networks

Flexibility metrics

Overview of results

Monotone transformations of networks

Proof sketches

Final remarks

Outline

Stochastic processing networks

Flexibility metrics

Overview of results

Monotone transformations of networks

Proof sketches

Final remarks

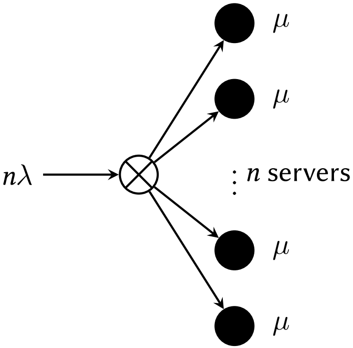
Introduction

- Load balancing plays a central role in parallel-processing systems.
- Balance incoming tasks across distributed servers.
- A lot of attention in recent decades, see e.g. [van der Boor et al., 2022].

Main goal: Efficiency, i.e. minimize idle servers.

Problem: we also need scalable policies.

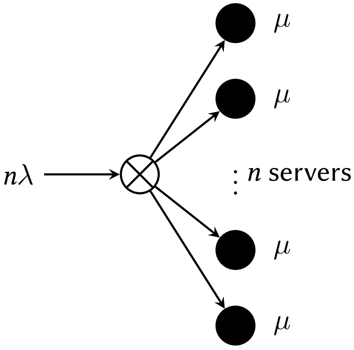
Supermarket model



$$\rho := \frac{\lambda}{\mu}$$

- Arrivals at rate λ, n parallel servers.
- Tasks are allowed to run in any server.
- Tasks are queued at the servers.
- Exponential assumptions.

Supermarket model



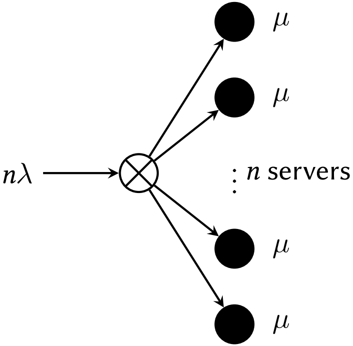
$$\rho := \frac{\lambda}{\mu}$$

- Arrivals at rate λ, n parallel servers.
- Tasks are allowed to run in any server.
- Tasks are queued at the servers.
- Exponential assumptions.

Join-the-shortest-queue (JSQ)

- Upon arrival, choose the shortest queue.
- Optimal policy. Achieves efficiency for large n.
- Large communication overhead.

Supermarket model



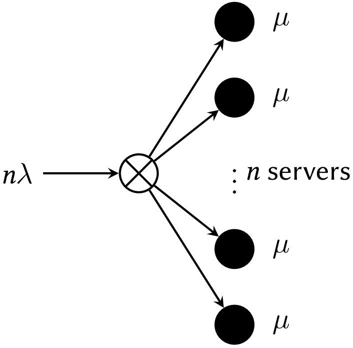
$$\rho := \frac{\lambda}{\mu}$$

- Arrivals at rate λ, n parallel servers.
- Tasks are allowed to run in any server.
- Tasks are queued at the servers.
- Exponential assumptions.

Power-of-d (PoD) [Vvedenskaya et al., 1996; Mitzenmacher, 2001]

- Upon arrival, sample d queues at random.
- Choose the shortest among them.
- Doubly exponential decay with minimal overhead.

Supermarket model



$$\rho := \frac{\lambda}{\mu}$$

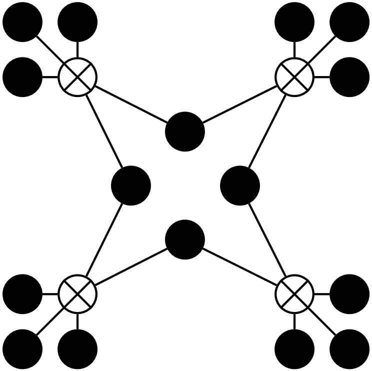
- Arrivals at rate λ, n parallel servers.
- Tasks are allowed to run in any server.
- Tasks are queued at the servers.
- Exponential assumptions.

Join-the-idle-queue (JIQ) [Lu et al., 2011]

- Upon arrival, choose an idle queue (you'll find one, trust me)
- Minimal communication overhead.
- Good if ρ away from 1.

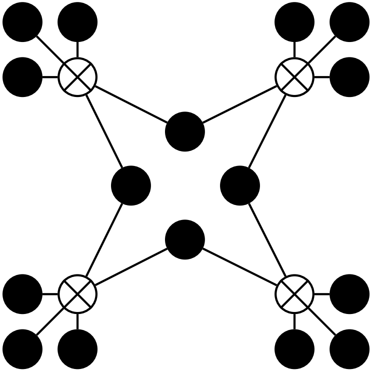
Networks are not that simple...

- Multiple entry points, types of tasks...
- Heterogeneous servers.
- Compatibility constraints: not every dispatcher is connected to every server.



Networks are not that simple...

- Multiple entry points, types of tasks...
- Heterogeneous servers.
- Compatibility constraints: not every dispatcher is connected to every server.



What's the right abstraction for this system?

Stochastic processing networks

Consider a bipartite graph $G = (D, S, E)$:

- $d \in D$ is a dispatcher, receives tasks at rate $\lambda(d)$.
- $s \in S$ is a server. Executes tasks sequentially at rate $\mu(s)$ and maintains a queue.
- $(d, s) \in E$ encodes compatibility constraints.
- We assume JSQ is used...

Definition

$\mathbf{X}(t, u)$, the number of tasks in server u at time t is the *load balancing process* associated with the bipartite graph $G = (D, S, E)$ and the rate functions $\lambda : D \to (0, \infty)$ and $\mu : S \to (0, \infty)$.

Stability conditions

[Foss and Chernova, 1998, Theorem 2.5]

For $\mathbf{X}$ to be stable (ergodic), it is sufficient that:

$$\sum_{\mathcal{N}(d) \subset U} \lambda(d) < \sum_{u \in U} \mu(u) \quad \text{for all nonempty } U \subset S$$

where:

$$\mathcal{N}(d) := \{u \in S : (d, u) \in E\}$$

denotes the set of servers that are compatible with some dispatcher d.

In what follows we always assume stability, and define $X(u)$ to be the steady-state of the load balancing process.

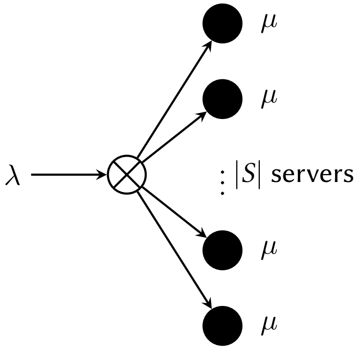
Simple processing network

Definition

Consider a network $G = (D, S, E)$ with D a singleton, $E = D \times S$, and constant λ and μ. We call the load balancing process of this network *simple* with load $\rho := \lambda/\mu$.

- I.e., the general definition contains the supermarket model under JSQ.
- Ergodic if $\rho < |S|$

Simple load balancing processes will be key to our proofs.



Performance measure

Definition (Queue occupancy measure)

Suppose that $\mathbf{X}$ is an ergodic load balancing process and let X denote its stationary distribution. The steady-state occupancy is the random sequence:

$$q(i) := \frac{1}{|S|} \sum_{u \in S} \mathbf{1}_{\{X(u) \geqslant i\}} \quad \text{for all} \quad i \in \mathbb{N}.$$

- Lower values of q imply better performance.
- $|S| \sum_i q(i) =$ total number of tasks, and thus delay by Little's law.

Queue occupancy in simple networks

Consider a sequence of simple networks of growing size $|S| = n \to \infty$, and $\lambda^{(n)} = n\lambda$, then $\rho^{(n)} = \rho$. In the mean field limit:

- $q(i) = \rho^i$ under random routing (geometric decay).
- $q(i) = \rho^{\frac{d^i - 1}{d-1}}$ under PoD (double exponential decay).
- $q(1) = \rho$ and $q(i) = 0$ for $i > 1$ under JSQ (and JIQ).

Queue occupancy in simple networks

Consider a sequence of simple networks of growing size $|S| = n \to \infty$, and $\lambda^{(n)} = n\lambda$, then $\rho^{(n)} = \rho$. In the mean field limit:

- $q(i) = \rho^i$ under random routing (geometric decay).
- $q(i) = \rho^{\frac{d^i - 1}{d-1}}$ under PoD (double exponential decay).
- $q(1) = \rho$ and $q(i) = 0$ for $i > 1$ under JSQ (and JIQ).

But the simple network is fully flexible...

Question: Can we have the same performance in the limit with less flexibility?

Asymptotic results under partial connectivity

[Rutten and Mukherjee, 2024] and others

Consider a sequence of graphs $G^{(n)}$ where $n = |S|$, $|D| = M(n)$ and $\lambda^{(n)} \equiv \frac{\lambda n}{M(n)}$ (so the total arrival rate is λn).

Introduce the regularity and diversity metrics:

$$\phi(G) = \max_u \left| \frac{|S|}{|D|} \sum_{d \in \mathcal{N}(u)} \frac{1}{\deg(d)} - 1 \right| \quad \text{and} \quad \gamma(G) = \frac{1}{|D|} \sum_d \frac{1}{\deg(d)}$$

- ϕ is a measure of (ir)regularity. How diverse are the degrees of the dispatchers.
- γ is an (in)flexibility metric. When $\gamma \to 0$, the average degree (options) grow.

Asymptotic results under partial connectivity

[Rutten and Mukherjee, 2024] and others

Theorem

If $\phi(G^{(n)}) \to 0$ *and* $\gamma(G^{(n)}) \to 0$*, then the load balancing process associated to the bipartite graph under PoD is "close" to the solution for a fully connected bipartite graph.*

Asymptotic results under partial connectivity

[Rutten and Mukherjee, 2024] and others

Theorem

If $\phi(G^{(n)}) \to 0$ *and* $\gamma(G^{(n)}) \to 0$*, then the load balancing process associated to the bipartite graph under PoD is "close" to the solution for a fully connected bipartite graph.*

- In plain terms, if the flexibility is large, and there are no dispatchers with few choices, then we recover the PoD behavior of the fully connected network.

- Similar results hold for other policies as well.

- Note $\gamma(G^{(n)}) \to 0$ implies that the average degree of dispatchers must go to ∞.

In this talk...

- We look for converse results!
- In particular, we provide geometric lower bounds for the network behavior when connectivity is limited.
- We do so by introducing a novel bottleneck measure, as well as the average degree.
- We show that, unless these metrics diverge, there will always be a geometric tail, even under JSQ.

Outline

Stochastic processing networks

Flexibility metrics

Overview of results

Monotone transformations of networks

Proof sketches

Final remarks

Bottleneck metric

Given a bipartite graph $G = (D, S, E)$, we define:

Definition (Bottleneck metric)

$$\alpha_G := \frac{1}{|S|} \sum_{u \in S} \min \{\deg(d) : d \in \mathcal{N}(u)\}$$

Interpretation:

- For a server u, $\min \{\deg(d) : d \in \mathcal{N}(u)\}$ quantifies how important is this server for some nodes.
- Now pick a server at random, what is the average “importance”.

Bottleneck metric

Further interpretation

- Assume that some dispatchers have only few options.
- Then the subset of servers that serve them are clearly a bottleneck for the network.
- Congestion will occur in these servers.
- If the size of this subset grows with $|S|$, then you're in trouble.

Average degree metric

Given a bipartite graph $G = (D, S, E)$, we define:

Definition (Average degree metric)

$$\beta_G := \frac{1}{|D|} \sum_{d \in D} \deg(d).$$

- Simpler than $\gamma(G)$, pick a dispatcher at random, what are they options on average.
- By Jensen's inequality:

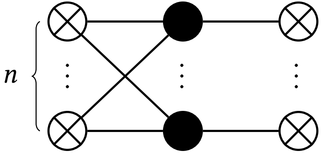
$$\gamma(G) = \frac{1}{|D|} \sum_{d} \frac{1}{\deg(d)} \geqslant \frac{1}{\frac{1}{|D|} \sum_d \deg(d)} = \frac{1}{\beta_G},$$

so again $\gamma(G) \to 0$ implies $\beta_G \to \infty$.

Examples

A network with too many (hidden) bottlenecks

Why two metrics? Let's look at the following examples:



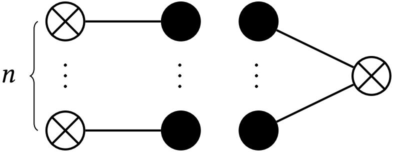
$$\alpha_{G_n^1} = 1,$$

$$\beta_{G_n^1} = \frac{n+1}{2}.$$

- In this case, α_G remains bounded, because at least half of the dispatchers crucially depend on a single server.
- However, the average degree of the network grows without bound.

Examples

A network with very flexible nodes hiding others



$$\alpha_{G_n^2} = \frac{n+1}{2},$$

$$\beta_{G_n^2} = \frac{2n}{n+1}.$$

- In this second case, half of the network is clearly disconnected, and thus has bounded $\min(d)$.
- However, the flexible dispatcher on the right makes the average $\alpha_G \to \infty$.
- Crucially in this case the average degree remains bounded.

Outline

Stochastic processing networks

Flexibility metrics

Overview of results

Monotone transformations of networks

Proof sketches

Final remarks

A useful lemma

Lemma

Let $\mathbf{X}$ be a simple and ergodic load balancing process with load ρ. Then its steady-state occupancy satisfies:

$$E\left[q(i)\right] \geqslant \frac{\left[r(\rho, |S|)\right]^i}{|S|} \quad \textit{for all} \quad i \in \mathbb{N},$$

where:

$$r(\rho, x) := \left(\frac{\rho}{x}\right)^x.$$

A useful lemma

Lemma

Let $\mathbf{X}$ be a simple and ergodic load balancing process with load ρ. Then its steady-state occupancy satisfies:

$$E\left[q(i)\right] \geqslant \frac{\left[r(\rho, |S|)\right]^i}{|S|} \quad \textit{for all} \quad i \in \mathbb{N},$$

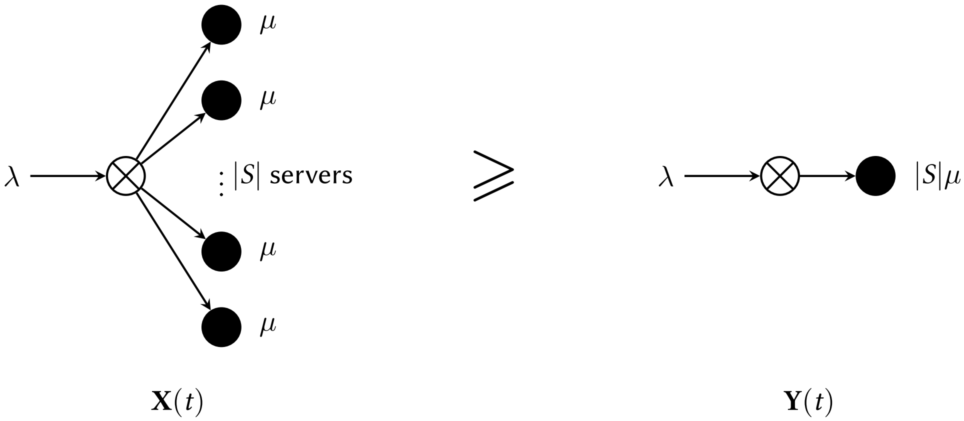
where:

$$r(\rho, x) := \left(\frac{\rho}{x}\right)^x.$$

So every simple network has a geometric tail for finite $|S|$.

Proof sketch I

Coupling with a single server queue



Proof sketch II

Coupling with a single server queue

- By construction, if both systems start empty:

$$\sum_{u \in S} \mathbf{X}(t, u) \geqslant \mathbf{Y}(t) \quad \text{for all } t$$

- $\mathbf{X}$ ergodic $\Rightarrow \lambda < \mu|S|$, therefore $\mathbf{Y}$ ergodic.

- Then in steady-state:

$$P\left(\sum_{u \in S} X(u) \geqslant i\right) \geqslant P\left(Y \geqslant i\right) \quad \text{for all} \quad i \in \mathbb{N}.$$

Proof sketch III

Coupling with a single server queue

- The above inequality implies that

$$P\left(Y \geqslant |S| i\right) \leqslant P\left(\sum_{u \in S} X(u) \geqslant |S| i\right) \leqslant P\left(\bigcup_{u \in S} \{X(u) \geqslant i\}\right) \leqslant |S| P\left(X(v) \geqslant i\right),$$

  where v is any server.

- We conclude by:

$$E\left[\mathbf{1}_{\{X(u) \geqslant i\}}\right] = P\left(X(u) \geqslant i\right) \geqslant \frac{1}{|S|} P\left(Y \geqslant |S| i\right) = \frac{1}{|S|}\left(\frac{\rho}{|S|}\right)^{|S| i} = \frac{[r(\rho, |S|)]^{i}}{|S|}$$

  for all $u \in S$, and averaging over u.

Simple network flexibility

For a simple network, note that:

$$\alpha_G = \frac{1}{|S|} \sum_{u \in S} \min \left\{ \deg(d) : d \in \mathcal{N}(u) \right\} = |S|,$$

$$\beta_G = \frac{1}{|D|} \sum_{d \in D} \deg(d) = |S|,$$

Simple network flexibility

For a simple network, note that:

$$\alpha_G = \frac{1}{|S|} \sum_{u \in S} \min \left\{ \deg(d) : d \in \mathcal{N}(u) \right\} = |S|,$$

$$\beta_G = \frac{1}{|D|} \sum_{d \in D} \deg(d) = |S|,$$

So we can deduce from the Lemma that:

$$E\left[q(i)\right] \geqslant \frac{\left[r(\rho, |S|)\right]^i}{|S|} = \frac{\left[r(\rho, \alpha_G)\right]^i}{\alpha_G} = \frac{\left[r(\rho, \beta_G)\right]^i}{\beta_G}$$

where again:

$$r(\rho, x) := \left(\frac{\rho}{x}\right)^x .$$

Bounds for general networks

The above geometric tail generalizes to any network:

Theorem (α_G bound)

Suppose that $\mathbf{X}$ *is the load balancing process of an ergodic stochastic processing network with* $\lambda(u)$ *and* $\mu(s)$.
Assume that:

$$0 < \lambda_0 \leqslant \min_{d \in D} \lambda(d) \quad \textit{and} \quad \max_{u \in S} \mu(u) \leqslant \mu_0 < \infty,$$

and let $\rho_0 := \frac{\lambda_0}{\mu_0}$. *If* q *is the steady state occupancy then:*

$$E\left[q(i)\right] \geqslant \frac{\left[r(\rho_0, \alpha_G)\right]^i}{\alpha_G} \quad \textit{for all} \quad i \geqslant \frac{1}{\rho_0},$$

Bounds for general networks

Theorem (β_G bound)

Suppose that $\mathbf{X}$ *is the load balancing process of an ergodic stochastic processing network with* $\lambda(u)$ *and* $\mu(s)$*.*
Assume that:

$$0 < \lambda_0 \leqslant \min_{d \in D} \lambda(d) \quad \textit{and} \quad \max_{u \in S} \mu(u) \leqslant \mu_0 < \infty,$$

and let $\rho_0 := \frac{\lambda_0}{\mu_0}$*. If* q *is the steady state occupancy then:*

$$E\left[q(i)\right] \geqslant C(\beta_G, \rho_0) \frac{\left[r(\rho_0, \beta_G + 1)\right]^i}{\beta_G + 1} \quad \textit{for all} \quad i \geqslant \frac{1}{\rho_0},$$

Main result

Theorem

Consider a sequence of bipartite graphs $G_n = (D_n, S_n, E_n)$ *with*

$$\alpha := \liminf_{n\to\infty} \alpha_{G_n} \quad \textit{and} \quad \beta := \liminf_{n\to\infty} \beta_{G_n}.$$

In addition, fix rate functions $\lambda_n : D_n \to (0,\infty)$ *and* $\mu_n : S_n \to (0,\infty)$ *such that*

$$\liminf_{n\to\infty} \min_{d\in D_n} \lambda_n(d) > \lambda_0 > 0 \quad \textit{and} \quad \limsup_{n\to\infty} \max_{u\in S_n} \mu_n(u) < \mu_0 < \infty.$$

If the associated $\mathbf{X}_n$ *are ergodic and* $\rho_0 := \lambda_0/\mu_0$*, then the steady-state occupancies satisfy:*

$$\liminf_{n\to\infty} E\left[q_n(i)\right] \geqslant \max\left\{\frac{\left[r(\rho_0,\alpha)\right]^i}{\alpha}, C(\beta,\rho_0)\frac{\left[r(\rho_0,\beta+1)\right]^i}{\beta+1}\right\} \quad \textit{for all} \quad i \geq \frac{1}{\rho_0}.$$

Proof sketch

- By assumption, there exists $n_0 \geq 1$ such that

$$\min_{d \in D_n} \lambda_n(d) > \lambda_0 \quad \text{and} \quad \max_{u \in S_n} \mu_n(u) < \mu_0 \quad \text{for all} \quad n \geqslant n_0.$$

- As a result, the bound Theorems imply that, for $n \geqslant n_0$,

$$E\left[q_n(i)\right] \geqslant \max\left\{\frac{\left[r(\rho_0, \alpha_{G_n})\right]^i}{\alpha_{G_n}}, C(\beta_{G_n}, \rho_0)\frac{\left[r(\rho_0, \beta_{G_n}+1)\right]^i}{\beta_{G_n}+1}\right\} \quad \text{for all} \quad i \geqslant \frac{1}{\rho_0}.$$

- Since $x \mapsto \left[r(\rho_0, x)\right]^i / x$ is continuous in $(0, \infty)$ and $\min\{\alpha_{G_n}, \beta_{G_n}\} \geqslant 1$, we get

$$\liminf_{n \to \infty} E\left[q_n(i)\right] \geqslant \max\left\{\frac{\left[r(\rho_0, \alpha)\right]^i}{\alpha}, C(\beta, \rho_0)\frac{\left[r(\rho_0, \beta+1)\right]^i}{\beta+1}\right\} \quad \text{for all} \quad i \geqslant \frac{1}{\rho_0}.$$

Unavoidable geometric tails

Remark: If either α or β are finite, then the mean steady-state occupancy cannot decay faster than geometrically in the limit,

Unavoidable geometric tails

Remark: If either α or β are finite, then the mean steady-state occupancy cannot decay faster than geometrically in the limit,

- This gives a partial converse to the results in [Mukherjee et al., 2020; Rutten and Mukherjee, 2024, 2023; Budhiraja et al., 2019; Weng et al., 2020; Zhao et al., 2022; Zhao and Mukherjee, 2023].
- They prove that, under suitable connectivity assumptions, the mean-field limit behaves as if the graph were complete, and thus decays faster that geometric.
- All of their connectivity assumptions imply $\beta = \infty$.

Unavoidable geometric tails

Remark: If either α or β are finite, then the mean steady-state occupancy cannot decay faster than geometrically in the limit,

- This gives a partial converse to the results in [Mukherjee et al., 2020; Rutten and Mukherjee, 2024, 2023; Budhiraja et al., 2019; Weng et al., 2020; Zhao et al., 2022; Zhao and Mukherjee, 2023].
- They prove that, under suitable connectivity assumptions, the mean-field limit behaves as if the graph were complete, and thus decays faster that geometric.
- All of their connectivity assumptions imply $\beta = \infty$.

The previous Theorem implies that such mean-field limits are not possible if $\beta < \infty$.

Outline

Stochastic processing networks

Flexibility metrics

Overview of results

Monotone transformations of networks

Proof sketches

Final remarks

Proof technique

We now move into the details of the proofs of the bounds.

- Key idea: make a sequence of *monotone transformations* to the original network, that preserve order (i.e. each step has better performance)

- These networks are coupled to preserve the laws of the original network.

- In the end, we end up with a series of $|D|$ (coupled) simple networks, where we can apply the previous bounds (and thus, geometric tails).

- Since these networks have better performance, the original network must also have a geometric tail.

Monotone transformations

Arrival rate decrease

We consider the following transformations $G_1 \mapsto G_2$, $\lambda_1 \mapsto \lambda_2$, $\mu_1 \mapsto \mu_2$:

Arrival rate decrease

The arrival rate of tasks is decreased for some dispatchers. Specifically:

$$\lambda_1(d) \geqslant \lambda_2(d) \quad \text{for all } d \in D_1$$

while $G_2 := G_1$ and $\mu_2 := \mu_1$.

Monotone transformations

Service rate increase

Service rate increase

The service rate of tasks is increased for some servers. Specifically:

$$\mu_1(u) \leqslant \mu_2(u) \quad \text{for all } u \in S_1$$

while $G_2 := G_1$ and $\lambda_2 := \lambda_1$.

Monotone transformations

Service rate increase

Service rate increase

The service rate of tasks is increased for some servers. Specifically:

$$\mu_1(u) \leqslant \mu_2(u) \quad \text{for all } u \in S_1$$

while $G_2 := G_1$ and $\lambda_2 := \lambda_1$.

It is clear that both transformations will produce a less congested network, i.e.:

$$P\left(X_1(u) \geqslant i\right) \geqslant P\left(X_2(u) \geqslant i\right) \quad \text{for all} \quad u \in S_1 \text{ and } i \in \mathbb{N}$$

in steady-state.

Monotone transformations

Edge simplification

- Our third transformation requires *coupling* some servers, so we need to keep track of these couplings.

- Consider a stochastic processing network $G = (D, S, E)$ with rates λ, μ.

- Associate with $\mathbf{X}$ a partition $\mathcal{S}$ of S such that all the servers in $U \in \mathcal{S}$ have the same *potential* departure process.

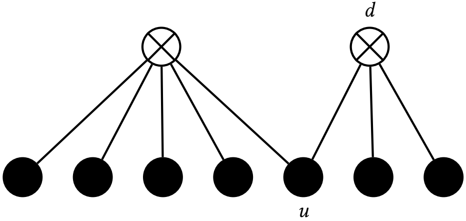
- Clearly, we must have $\mu(u) = \mu(v)$ if $u, v \in U$ and $U \in \mathcal{S}$.

- Initially, $\mathcal{S} = \{\{u\} : u \in S\}$.

Monotone Transformations

Edge simplification

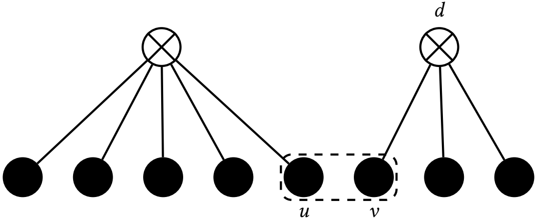
Assume that you start with the following network:



Monotone Transformations

Edge simplification

And apply edge simplification:



- Edge simplification that removes the compatibility relation (d, u) and incorporates server v and the compatibility relation (d, v).
- The servers u and v have the same potential departure process (coupling).

Monotone transformations

Edge simplification

The formal definition is:

Edge simplificatioon

A compatibility relation $(d, u) \in E_1$ is removed while a server $v \notin S_1$ and the edge (d, v) are incorporated. Specifically,

$$D_2 := D_1, \quad S_2 := S_1 \cup \{v\} \quad \text{and} \quad E_2 := (E_1 \setminus \{(d, u)\}) \cup \{(d, v)\}\,.$$

The potential departure process of v is the same as for u. Namely, suppose that U is the element of the partition $\mathcal{S}_1$ such that $u \in U$. Then we let

$$\mathcal{S}_2 := (\mathcal{S}_1 \setminus \{U\}) \cup \{U \cup \{v\}\} \quad \text{and} \quad \mu_2(v) := \mu_1(u).$$

Further, $\lambda_2(d) := \lambda_1(d)$ for all $d \in D_2$ and $\mu_2(w) := \mu_1(w)$ for all $w \in S_1$.

Edge simplification improves performance

Proposition

Suppose now that $\mathbf{X}_2$ *is obtained from* $\mathbf{X}_1$ *by edge simplification, removing* (d, u) *and incorporating server* v. *Assume* $\mathbf{X}_1$ *is ergodic, then the following inequalities hold in steady-state:*

$$P\left(X_1(u) \geqslant i\right) \geqslant P\left(X_2(v) \geqslant i\right) \quad \textit{and} \quad P\left(X_1(w) \geqslant i\right) \geqslant P\left(X_2(w) \geqslant i\right)$$

for all $w \in S_1$ *and* $i \in \mathbb{N}$.

Outline

Stochastic processing networks

Flexibility metrics

Overview of results

Monotone transformations of networks

Proof sketches

Final remarks

Proof of the α bound I

First let's recall the Theorem:

Theorem (α_G bound)

Suppose that $\mathbf{X}$ *is the load balancing process of an ergodic stochastic processing network with* $\lambda(d)$ *and* $\mu(u)$*.*
Assume that:

$$0 < \lambda_0 \leqslant \min_{d \in D} \lambda(d) \quad \textit{and} \quad \max_{u \in S} \mu(u) \leqslant \mu_0 < \infty,$$

and let $\rho_0 := \frac{\lambda_0}{\mu_0}$*. If* q *is the steady state occupancy then:*

$$E\left[q(i)\right] \geqslant \frac{\left[r(\rho_0, \alpha_G)\right]^i}{\alpha_G} \quad \textit{for all} \quad i \geqslant \frac{1}{\rho_0},$$

Proof of the α bound II

Given a graph $G = (D, S, E)$ and rates λ, μ:

- We perform an edge simplification at all the edges sequentially.
- From these transformations we get the bipartite graph $G_0 = (D_0, S_0, E_0)$ given by

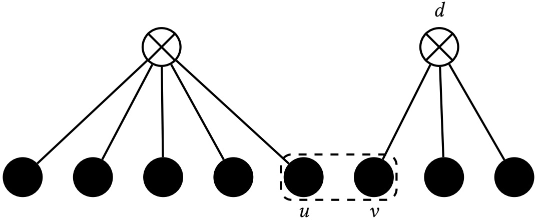
$$D_0 := D, \quad S_0 := \{u_d : (d, u) \in E\} \quad \text{and} \quad E_0 := \{(d, u_d) : (d, u) \in E\},$$

- The sets of coupled servers are given by:

$$\mathcal{S}_0 := \{\{u_d : u \in \mathcal{N}(d)\} : d \in D\}.$$

Proof of the α bound III

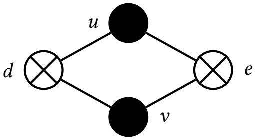
In other words we do this...

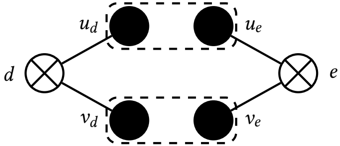


...until all dispatchers get a simple network of their own (with coupled departure processes).

Proof of the α bound IV

Moreover, after decomposition, all coupled servers end up in different neighborhoods!





Proof of the α bound V

- Now apply the arrival rate decrease and departure rate increase transformation...
- ...until all the dispatchers have the same arrival rate λ_0,
- ...and all servers have the same rate μ_0.

Call $\mathbf{X}_0$ the resulting load balancing process, it follows that $\mathbf{X}_0$ is ergodic and in steady-state:

$$P\left(X(u) \geqslant i\right) \geqslant P\left(X_0(u_d) \geqslant i\right) \quad \text{for all} \quad (d, u) \in E \quad \text{and} \quad i \in \mathbb{N},$$

Proof of the α bound VI

- Choose now weights for each edge $\theta : D \times S \to [0,1]$ such that

$$\sum_{d\in\mathcal{N}(u)} \theta(d,u) = 1 \quad \text{for all} \quad u \in S.$$

- Then we have the following for the steady-state occupancy:

$$E\left[q(i)\right] = \frac{1}{|S|}\sum_{u\in S} P\left(X(u) \geqslant i\right) \geqslant \frac{1}{|S|}\sum_{u\in S}\sum_{d\in\mathcal{N}(u)} \theta(d,u) P\left(X_0(u_d) \geqslant i\right) \text{ for all } i \in \mathbb{N}.$$

- Now for each dispatcher, we have a simple load-balancing process, with the same degree as in the original graph.

Proof of the α bound VII

- We can apply the proposition to each component to get:

$$E\left[q(i)\right] \geqslant \frac{1}{|S|}\sum_{u\in S}\sum_{d\in\mathcal{N}(u)}\theta(d,u)\frac{\left[r\left(\rho_0,\deg(d)\right)\right]^i}{\deg(d)},$$

- Observe that the function:

$$f(x) := \frac{\left[r(\rho,x)\right]^k}{x} = \frac{1}{x}\left(\frac{\rho}{x}\right)^{kx} \quad \text{for all} \quad x > 0$$

is strictly decreasing and convex in $[\rho,\infty)$.

- Observe that:

$$\sum_{u\in S}\sum_{d\in\mathcal{N}(u)}\frac{\theta(d,u)}{|S|} = 1,$$

Proof of the α bound VIII

- Therefore, using Jensen's inequality:

$$E\left[q(i)\right] \geqslant \frac{\left[r(\rho_0, \theta_G)\right]^i}{\theta_G} \quad \text{for all } i \geqslant \frac{1}{\rho_0} \quad \text{with } \theta_G := \frac{1}{|S|} \sum_{u \in S} \sum_{d \in \mathcal{N}(u)} \theta(d, u) \deg(d).$$

- Finally ecall that:

$$\alpha_G := \frac{1}{|S|} \sum_{u \in S} \min\left\{\deg(d) : d \in \mathcal{N}(u)\right\}$$

is just one possible θ (in fact is the one that achieves the sup over all θ, and thus the better bound).

Outline

Stochastic processing networks

Flexibility metrics

Overview of results

Monotone transformations of networks

Proof sketches

Final remarks

Final remarks

- Simple load-balancing networks always have geometric tails for finite number of servers.

- We defined two flexibility measures α_G and β_G that describe dispatcher-server connectivity.

- We showed that, in a sequence of growing size networks, unless $\alpha_G \to \infty$, $\beta_G \to \infty$, the geometric tails do not disappear in the limit.

- Future work: characterize the size of the bottleneck set, by defining:

$$\alpha_G(U) = \frac{1}{|U|} \sum_{u \in U} \min \{\deg(d) : d \in \mathcal{N}(u)\} \quad U \subset S$$

  and study its properties.

Bedankt!

Andres Ferragut

ferragut@ort.edu.uy

aferragu.github.io

Arxiv version

References I

A. Budhiraja, D. Mukherjee, and R. Wu. Supermarket model on graphs. *The Annals of Applied Probability*, 29(3):1740–1777, 2019.

S. G. Foss and N. I. Chernova. On the stability of a partially accessible multi-station queue with state-dependent routing. *Queueing Systems*, 29:55–73, 1998.

Y. Lu, Q. Xie, G. Kliot, A. Geller, J. R. Larus, and A. Greenberg. Join-Idle-Queue: A novel load balancing algorithm for dynamically scalable web services. *Performance Evaluation*, 68(11): 1056–1071, 2011.

M. Mitzenmacher. The power of two choices in randomized load balancing. *IEEE Transactions on Parallel and Distributed Systems*, 12(10):1094–1104, 2001.

D. Mukherjee, S. C. Borst, J. S. H. van Leeuwaarden, and P. A. Whiting. Asymptotic optimality of power-of-d load balancing in large-scale systems. *Mathematics of Operations Research*, 45(4): 1535–1571, 2020.

D. Rutten and D. Mukherjee. Load balancing under strict compatibility constraints. *Mathematics of Operations Research*, 48(1):227–256, 2023.

References II

D. Rutten and D. Mukherjee. Mean-field analysis for load balancing on spatial graphs. *The Annals of Applied Probability*, 34(6):5228–5257, 2024.

M. van der Boor, S. C. Borst, J. S. H. van Leeuwaarden, and D. Mukherjee. Scalable load balancing in networked systems: A survey of recent advances. *SIAM Review*, 64(3):554–622, 2022.

N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich. Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii*, 32(1):20–34, 1996.

W. Weng, X. Zhou, and R. Srikant. Optimal load balancing with locality constraints. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 4(3):1–37, 2020.

Z. Zhao and D. Mukherjee. Optimal rate-matrix pruning for large-scale heterogeneous systems. *arXiv preprint arXiv:2306.00274*, 2023.

Z. Zhao, D. Mukherjee, and R. Wu. Exploiting data locality to improve performance of heterogeneous server clusters. *arXiv preprint arXiv:2211.16416*, 2022.