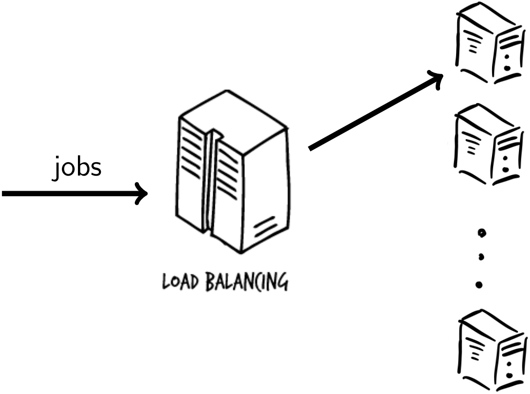
Approximations to Study the Impact of the Service Discipline in Systems with Redundancy

Nicolas Gast
Inria & Univ. Grenoble Alpes

Benny Van Houdt
University of Antwerp

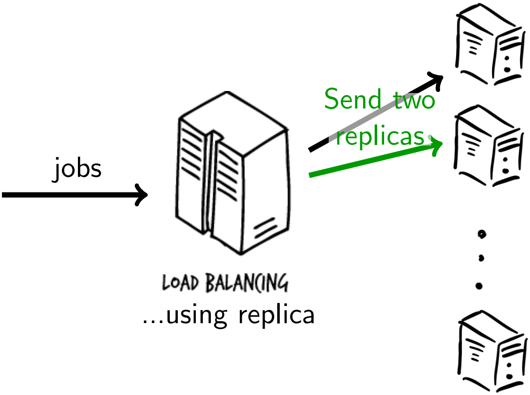
ACM SIGMETRICS 2024, Venezia

Redundancy can be used as a “load balancing” strategy



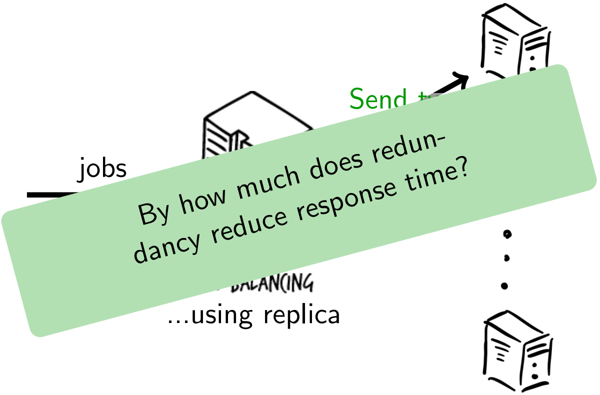
Policies: Random, JIQ, JSQ(d),...

Redundancy can be used as a “load balancing” strategy



Effective Straggler Mitigation: Attack of the Clones – Ananthanarayanan et al. NSDI 2013
The Tail at Scale – Dean and Barroso. Commun. ACM 2013

Redundancy can be used as a “load balancing” strategy



Effective Straggler Mitigation: Attack of the Clones – Ananthanarayanan et al. NSDI 2013
The Tail at Scale – Dean and Barroso. Commun. ACM 2013

There are lots of work, depending on the model considered.

- Are replica sizes: Equal? *i.i.d.*? Correlated (S&X)[1]
- Do we cancel replicas: on start? on completion?

Different metric considered:

- Stability[2]? Exact analysis[3] or Asymptotic regime[4].

---

[1] A better model for job redundancy: Decoupling server slowdown and job size. Gardner et al. 2017

[2] A Survey of Stability Results for Redundancy Systems. Anton et al 2021.

[3] Redundancy-d: The power of d choices for redundancy. Gadner et al. 2017

[4] Shneer and Stolyar. Large-scale parallel server system with multi-component jobs. QUESTA 21.

There are lots of work, depending on the model considered.

- Are replica sizes: Equal? *i.i.d.*? Correl
- Do we cancel replicas: on

Different meth

- Stability[2] act analysis[3] or Asymptotic regime[4].

All of those work (except stability) assume **FCFS**.

[1] A better model for job redundancy: Decoupling server slowdown and job size. Gardner et al. 2017

[2] A Survey of Stability Results for Redundancy Systems. Anton et al 2021.

[3] Redundancy-d: The power of d choices for redundancy. Gadner et al. 2017

[4] Shneer and Stolyar. Large-scale parallel server system with multi-component jobs. QUESTA 21.

There are lots of work, depending on the model considered.

- Are replica sizes: Equal? *i.i.d.*? Correl
- Do we cancel replicas: on

Different met

- Stability[2] analysis[3] or Asymptotic regime[4].

All of those work (except stability) assume **FCFS**.

why?

- It makes sense.
- You can use order-independent queues or asymptotic independence.

[1] A better model for job redundancy: Decoupling server slowdown and job size. Gardner et al. 2017

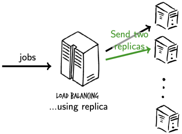
[2] A Survey of Stability Results for Redundancy Systems. Anton et al 2021.

[3] Redundancy-d: The power of d choices for redundancy. Gadner et al. 2017

[4] Shneer and Stolyar. Large-scale parallel server system with multi-component jobs. QUESTA 21.

Our Work: Impact of the Service Discipline in Redundancy

We focus on a (simple) queueing model:



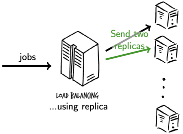
- N identical servers.
- Poisson arrival rate: $N\lambda$.
- Cancel on complete.

For each job, we send two[5] replicas, exponentially distributed, and *i.i.d.*.

[5] For $d > 2$ replicas: see paper

Our Work: Impact of the Service Discipline in Redundancy

We focus on a (simple) queueing model:



- N identical servers.
- Poisson arrival rate: $N\lambda$.
- Cancel on complete.

For each job, we send two[5] replicas, exponentially distributed, and *i.i.d.*.

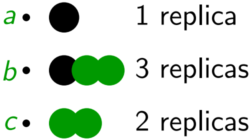
Our results

1. Service discipline does matter (even for *i.i.d* exponential replicas).
2. PS is connected to a dynamic random graph model.
3. We can build pair approximation (and triplet approximations) that accurate but not asymptocally exact.

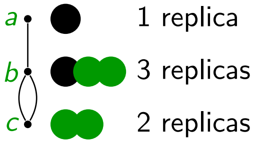
[5] For $d > 2$ replicas: see paper

Outline

Markovian representation: dynamic graph model



Markovian representation: dynamic graph model



We model the N servers by a graph with N nodes.

- For each job shared by i and j, we add an edge (i, j)

Markovian representation: dynamic graph model

1 replica

3 replicas

2 replicas

We model the N servers by a graph with N nodes.

- For each job shared by i and j, we add an edge (i, j)

Markovian representation: dynamic graph model

1 replica

3 replicas

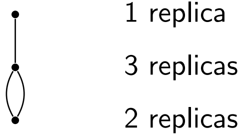
2 replicas

We model the N servers by a graph with N nodes.

- For each job shared by i and j, we add an edge (i, j)
- Each edge is created at rate $2\lambda/N$.
- Each node deletes one of its edge at rate 1.

We want to study the degree distribution (=queue length)

Markovian representation: dynamic graph model



We model the N servers by a graph with N nodes.

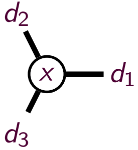
- For each job shared by i and j, we add an edge (i, j)
- Each edge is created at rate $2\lambda/N$. ⟶ Similar to Erdos-Renyi
- Each node deletes one of its edge at rate 1. ⟶ Creates dependencies

We want to study the degree distribution (=queue length)

Outline

Construction of a mean field approximation

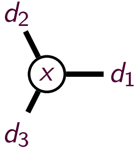
We zoom on a node that has degree x:



- $x \mapsto x+1$ at rate 2λ
- $x \mapsto x-1$ at rate $1 + \sum_{i=1}^{x} \frac{1}{d_i}$, where d_i is the degree of the ith neighboor.

Construction of a mean field approximation

We zoom on a node that has degree x:

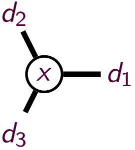


- $x \mapsto x+1$ at rate 2λ
- $x \mapsto x-1$ at rate $1+\sum_{i=1}^{x}\frac{1}{d_i}$, where d_i is the degree of the ith neighboor.

$$\mathbb{E}\left[\frac{1}{d_i}\right] = ?$$

Construction of a mean field approximation

We zoom on a node that has degree x:

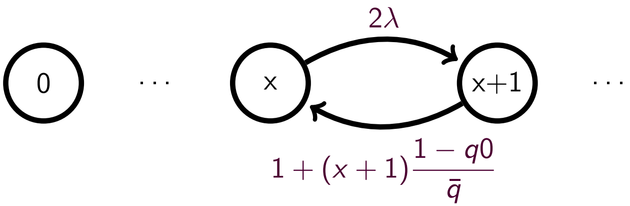


- $x \mapsto x+1$ at rate 2λ
- $x \mapsto x-1$ at rate $1 + \sum_{i=1}^{x} \frac{1}{d_i}$, where d_i is the degree of the ith neighboor.

$$\mathbb{E}\left[\frac{1}{d_i}\right] = \sum_{q\geq 1} \underbrace{\mathbf{P}\left[d_i = q\right]}_{\approx \frac{q\mathbf{P}[\text{degree}=q]}{\bar{q}} \text{ (mean field approximation)}} \frac{1}{q} = \frac{1-q_0}{\bar{q}},$$

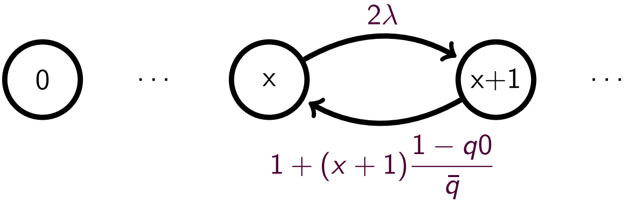
where $\bar{q} = \sum_{q} q\mathbf{P}\left[\text{degree} = q\right]$ is the average queue length.

When zooming on the node, we have a density dependent birth-death process



+ ODE easy to integrate numerically.
+ Almost closed-form fixed-point (see paper)

When zooming on the node, we have a density dependent birth-death process



+ ODE easy to integrate numerically.
+ Almost closed-form fixed-point (see paper)
- **But:** This assumes that neighboring nodes are independent.

This approximation is accurate

For $\lambda = 0.9$ and $n = 10^6$:

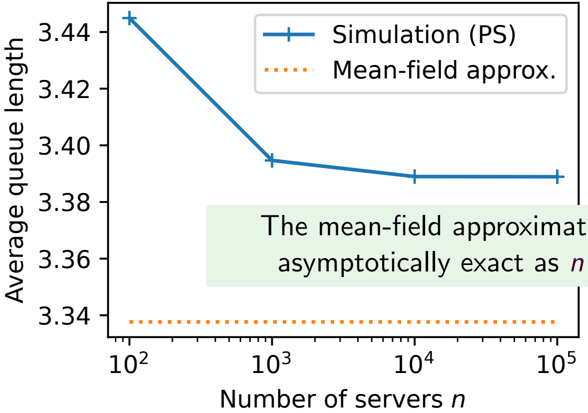
PS (simu)	PS (mean-field)	FCFS (simu)	FCFS (theory[6])
3.3889	3.3376	3.1168	3.1169

[6] Redundancy-d: The power of d choices for redundancy. Gardner et al. OR 2017

This approximation is accurate...but not asymptotically exact.

For $\lambda = 0.9$ and $n = 10^6$:

PS (simu)	PS (mean-field)	FCFS (simu)	FCFS (theory[6])
3.3889	3.3376	3.1168	3.1169

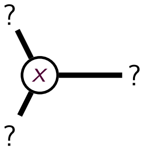


The mean-field approximation is not asymptotically exact as $n \to \infty$.

[6] Redundancy-d: The power of d choices for redundancy. Gardner et al. OR 2017

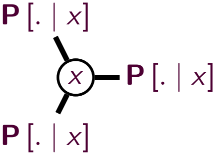
We can build a more accurate approximation: The pair-approximation

The mean-field approximation assumes that the degree of neighboring nodes are independent. They are not.



We can build a more accurate approximation: The pair-approximation

The mean-field approximation assumes that the degree of neighboring nodes are independent. They are not.

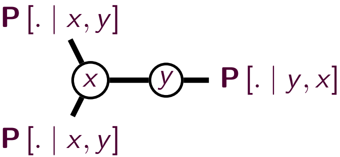


We track:

$$\pi(x, y) = \frac{1}{N} \#\{\text{connected pairs } (x, y)\}.$$

We can build a more accurate approximation: The pair-approximation

The mean-field approximation assumes that the degree of neighboring nodes are independent. They are not.



We track:

$$\pi(x, y) = \frac{1}{N} \#\{\text{connected pairs } (x, y)\}.$$

The pair-approximation is

$$\mathbf{P}[z|x, y] \approx \mathbf{P}[z|x] = \frac{\pi(x, z)}{\sum_{z'} \pi(x, z')}$$

We can construct an ODE approximation for π

The events affecting π are:

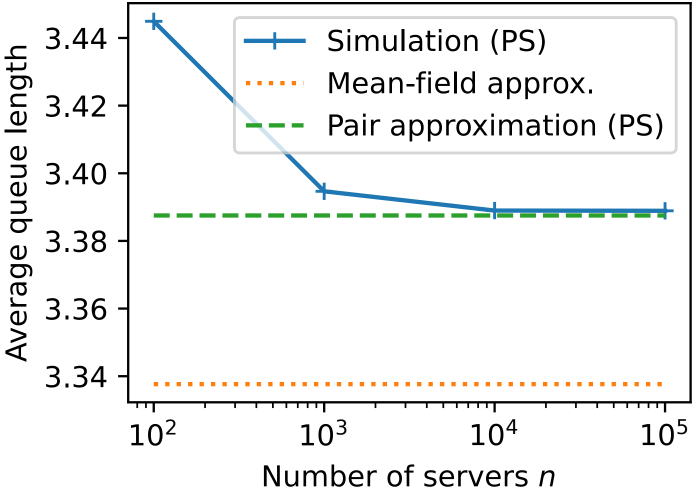
- Creation or destruction of pairs
- $(x, y) \mapsto (x+1, y)$: creation of a new neighbor of x
- $(x, y) \mapsto (x-1, y)$: departure of one of the $x-1$ neighbors of x.

$$\begin{aligned}\frac{d\pi_t(x,y)}{dt} &= \lambda q_t(x-1)q_t(y-1) + 2\lambda\left[\pi_t(x-1,y) + \pi_t(x,y-1) - 2\pi_t(x,y)\right] \\ &\quad + \pi_t(x+1,y)\left[h_t(x+1) + \frac{x}{x+1}\right] + \pi_t(x,y+1)\left[h_t(y+1) + \frac{y}{y+1}\right] \\ &\quad - \pi_t(x,y)\left[2 + h_t(x) + h_t(y)\right], \end{aligned} \tag{11}$$

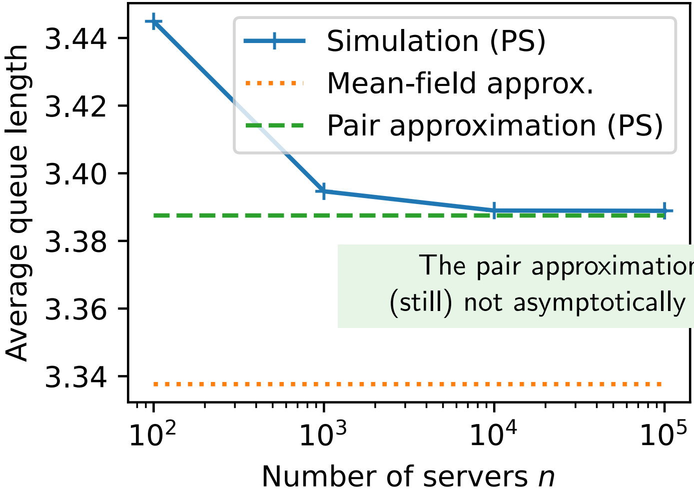
\+ Easy to integrate numerically.

– Is this asymptotically exact?

The pair approximation is more accurate than the m-f.

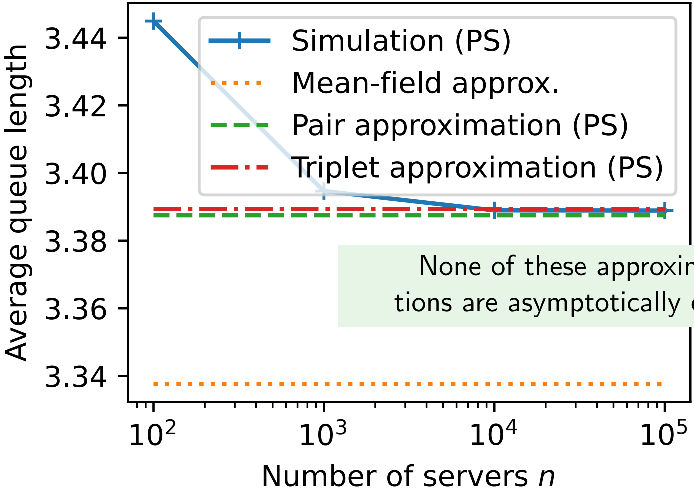


The pair approximation is more accurate than the m-f.



The pair approximation is
(still) not asymptotically exact

The pair approximation is more accurate than the m-f.



None of these approximations are asymptotically exact

Can we do triplet (but complexity is large (construction+computation)).

Outline

In the paper, we build approx. for FCFS, LCFS and LPS(K)

More complex than for PS because we need to track the replicas' positions

$$\pi(x, y, \mathrm{pos}_x, \mathrm{pos}_y)$$

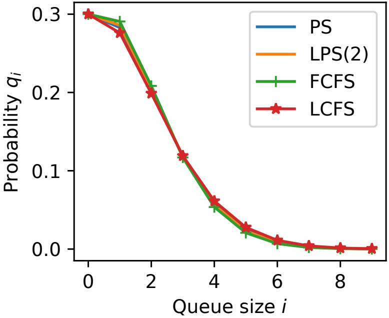
They allow to study the queue length distribution and correlations.

In the paper, we build approx. for FCFS, LCFS and LPS(K)

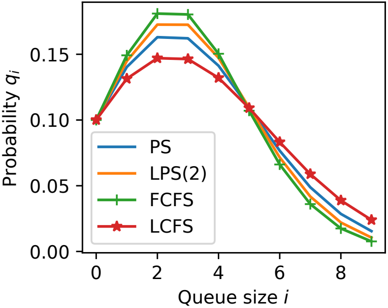
More complex than for PS because we need to track the replicas' positions

$$\pi(x, y, \mathrm{pos}_x, \mathrm{pos}_y)$$

They allow to study the queue length distribution and correlations.



$\rho = 0.7$



$\rho = 0.9$

FCFS is the best, due to correlations between replicas (see paper).

Outline

Conclusion

Service disciplines affect queue length in system with redundancy

- Even when replicas are *i.i.d.* and have exponential sizes.

We provide numerical scheme (ODE) based or mean-field or pair approximation.

- They are not asymptotically exact but very accurate.
- They confirm that FCFS performs best (correlated replicas).

Open questions and references

Future work:

- Link with JIQ + redundancy.
- More general model: non *i.i.d.*, heterogeneous, non-exponential.

Slides and references: http://polaris.imag.fr/nicolas.gast

- Approximations to Study the Impact of the Service Discipline in Systems with Redundancy. Nicolas Gast and Benny Van Houdt. ACM SIGMETRICS 2024. https://arxiv.org/abs/2401.07713