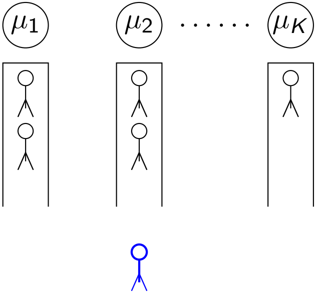
On the stability of redundancy models

Elene Anton
Based on joint work with:
U. Ayesta, M. Jonckheere and I.M. Verloop

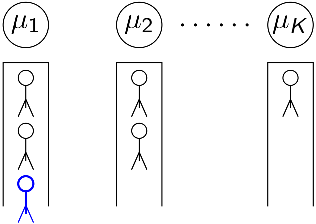
IRIT-CNRS and ENSEEIHT

December 3, 2018

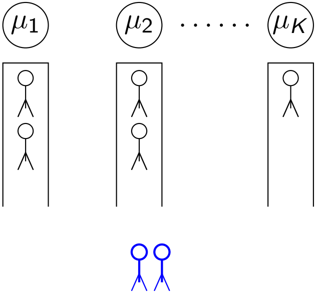
Load-balancing strategies:



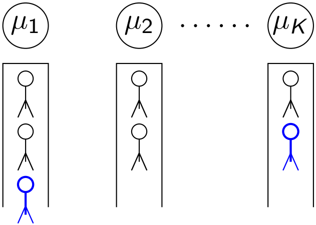
Load-balancing strategies:



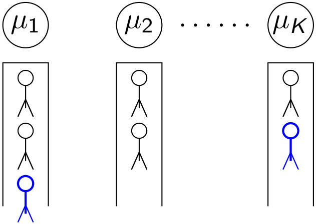
Redundancy-d: A job is dispatched into several servers.



Redundancy-d: A job is dispatched into several servers.



Redundancy-d: A job is dispatched into several servers.



Exploit variability in the workload in different queues !

Trade off of redundancy

- Positive aspect: Exploits variability in the workload.
- Negative aspect: There is additional workload added to the system.

Previous work

Theorem

Assume FCFS service policy and all the copies of a job are i.i.d. The system is stable $\iff \lambda < \mu K$.

[Gardener et al.] [1]

[1] *Kristen Gardner, Samuel Zbarsky, Sherwin Doroudi, Mor Harchol-Balter, Esa Hyytiä, and Alan Scheller-Wolf. 2016. Queueing with redundant requests: exact analysis. Queueing Systems 83, 3-4 (2016), 227–259*

Main objectives

Determine how the stability condition is impacted by:

- The scheduling policy implemented in the K servers.

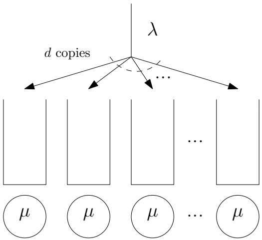
Main objectives

Determine how the stability condition is impacted by:

- The scheduling policy implemented in the K servers.
- The possible correlation between the d copies of the same job.

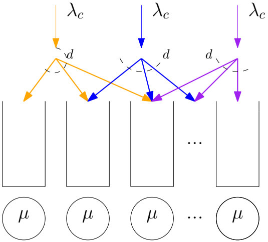
Outline

Model description



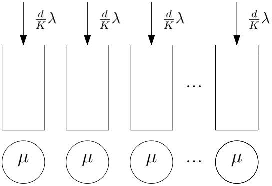
- K servers with capacity 1.
- Poisson arrivals with rate λ.
- Exponential service times with parameter μ.

Model description



- Each arrival chooses d servers at random, $s_1, \ldots, s_d$.
- This job is said to be of type $c = \{s_1, \ldots, s_d\}$.
- The set of types:
  $\mathcal{C} := \{c = \{s_1, \ldots, s_d\} \subset S \ : \ s_i \neq s_j \ \forall i \neq j\}$ and $|\mathcal{C}| = \binom{K}{d}$.
- Arrivals of type c at rate $\lambda_c = \frac{\lambda}{\binom{K}{d}}$.

Model description



- Arrival rate to a server s is $\frac{d}{K}\lambda$.
- Departure in server s due to:
  - Local copy has completed service.
  - A copy of a job in the local queue has completed service in an other server.

Model description

- The number of type-c jobs at time t is given by $N_c(t)$ and

$$\vec{N}(t) = (N_1(t), \ldots, N_{|\mathcal{C}|}(t)) \in \mathbb{Z}_+^K$$

Model description

- The number of type-c jobs at time t is given by $N_c(t)$ and

$$\vec{N}(t) = (N_1(t), \ldots, N_{|\mathcal{C}|}(t)) \in \mathbb{Z}_+^K$$

- The number of copies in server s at time t is given by $M_s(t) = \sum_{c \in \mathcal{C}(s)} N_c(t)$ and

$$\vec{M}(t) = (M_1(t), \ldots, M_K(t)) \in \mathbb{Z}_+^K$$

Model description

Service policies we consider:

- PS (Processor Sharing): service is equally shared among the copies in a server.
- FCFS: copies are served in order of arrival.
- ROS (Random Order of Service): An empty server picks a copy to serve at random.
- Priority policy: In each server, a priority law is fixed among the types it can serve.

Model description

We consider copies of a job to be:

1. **i.i.d copies.**
2. **identical copies:** All d copies of a job are identical replicas and have the same service time.

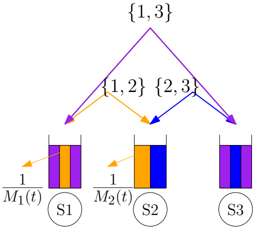
Main results

Table: Summary of stability conditions

	PS	FCFS	ROS	Priority policy
i.i.d	$\lambda < \mu K$	$\lambda < \mu K$	$\lambda < \mu K$	$\lambda << \mu K$
i.c.	$\lambda < \mu \frac{K}{d}$	$\lambda < \tilde{\mu}$	$\lambda < \mu K$	–
		$(\tilde{\mu} < \mu(K - (d-1)))$		

PS service policy with iid

Example: $K = 3$ and $d = 2$ copies, $\mathcal{C} = \{\{1,2\},\{1,3\},\{2,3\}\}$



- I.i.d copies $\Longrightarrow$ the departure rate of a type$-c$ job is

$$\sum_{s \in c} \frac{\mu}{M_s(t)}$$

.

PS service policy with iid

Theorem

Assume PS service policy and copies of a job are i.i.d.
The system is stable $\iff \lambda < \mu K$.

Proof:

- Show that fluid limit satisfies

$$\frac{\mathrm{d}m_{max}(t)}{\mathrm{d}t} = \lambda\frac{d}{K} - \mu\left(\sum_{c\in\mathcal{C}(s)}\sum_{l\in S(c)}\frac{n_c}{m_l}\right) \leq \lambda\frac{d}{K} - \mu d$$

ROS service policy with iid

Theorem

Assume ROS service policy and copies of a job are i.i.d.
The system is stable $\iff \lambda < \mu K$.

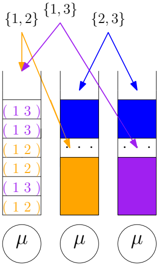
Proof:

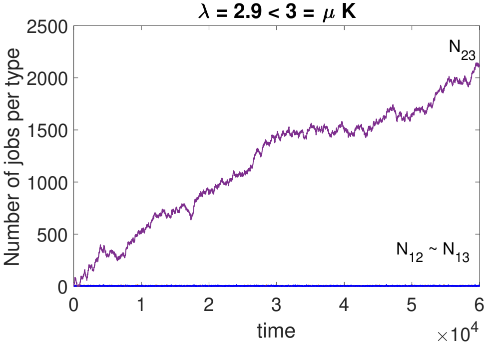
- Show that fluid limit satisfies $\frac{\mathrm{d}m_{max}(t)}{\mathrm{d}t} \leq \lambda \frac{d}{K} - \mu d$

Priority policy with K=3 servers and d=2 copies

$\mathcal{C} = \{\{1,2\},\{1,3\},\{2,3\}\}$.

Server 1: FCFS, Server 2: $\{1,2\} \preceq \{2,3\}$, Server 3: $\{1,3\} \preceq \{2,3\}$.



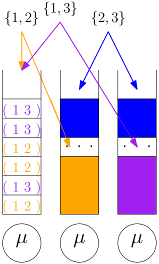


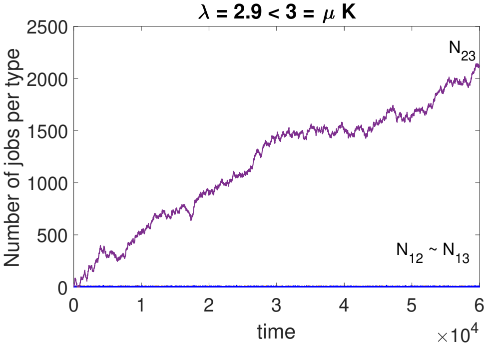
$$\frac{\mathrm{d}|\vec{n}(t)|}{\mathrm{d}t} = \lambda - (3\mu - \mu P(\text{ server 1 is empty })).$$

Priority policy with K=3 servers and d=2 copies

$\mathcal{C} = \{\{1,2\},\{1,3\},\{2,3\}\}$.

Server 1: FCFS, Server 2: $\{1,2\} \preceq \{2,3\}$, Server 3: $\{1,3\} \preceq \{2,3\}$.





The system can be unstable when $\lambda < \mu K$.

Outline

Identical copies assumption

- IID copies: $\lambda < \mu K$.

Identical copies assumption

- IID copies: $\lambda < \mu K$.
  - $d = 1 \implies K$ homogeneous servers with rate μ.
  - $d = K \implies$ single server with rate μK.

Identical copies assumption

- IID copies: $\lambda < \mu K$.
  - $d = 1 \implies K$ homogeneous servers with rate μ.
  - $d = K \implies$ single server with rate μK.
- Identical copies: All copies of a job are exact replicas with the same service time.

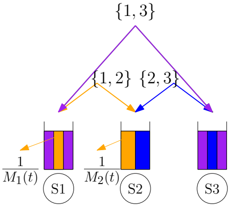
Identical copies assumption

- IID copies: $\lambda < \mu K$.
  - $d = 1 \implies K$ homogeneous servers with rate μ.
  - $d = K \implies$ single server with rate μK.
- Identical copies: All copies of a job are exact replicas with the same service time.
  - For $d = 1 \implies K$ homogeneous servers with rate μ.
  - For $d = K \implies$ single server with rate μ.

> The performance decreases in d: no longer maximum stable

PS service policy with Identical copies

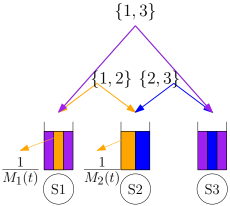
Example: $K = 3$ and $d = 2$ copies, $\mathcal{C} = \{\{1,2\},\{1,3\},\{2,3\}\}$



- $a_{cis}(t)$ attained service of the i-th type$-c$ job.

PS service policy with Identical copies

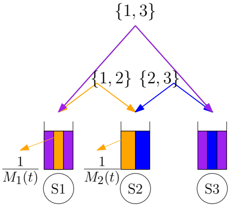
Example: $K = 3$ and $d = 2$ copies, $\mathcal{C} = \{\{1,2\},\{1,3\},\{2,3\}\}$



- $a_{cis}(t)$ attained service of the i-th type$-c$ job.

PS service policy with Identical copies

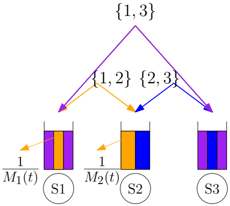
Example: $K = 3$ and $d = 2$ copies, $\mathcal{C} = \{\{1,2\},\{1,3\},\{2,3\}\}$



- $a_{cis}(t)$ attained service of the i-th type$-c$ job.
- $\frac{\mathrm{d}a_{cis}(t)}{\mathrm{d}t} = \frac{1}{M_s(t)}$.

PS service policy with Identical copies

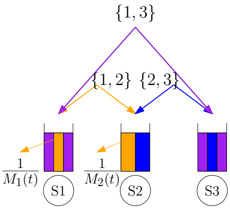
Example: $K = 3$ and $d = 2$ copies, $\mathcal{C} = \{\{1,2\},\{1,3\},\{2,3\}\}$



- $a_{cis}(t)$ attained service of the i-th type$-c$ job.
- $\frac{\mathrm{d}a_{cis}(t)}{\mathrm{d}t} = \frac{1}{M_s(t)}$.
- A job leaves the system due to a departure in server $s^*_{ci}(t) = \arg\max_{s \in c}\{a_{cis}(t)\}$.

PS service policy with Identical copies

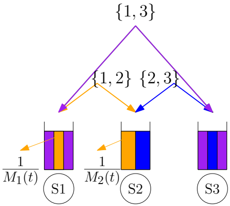
Example: $K = 3$ and $d = 2$ copies, $\mathcal{C} = \{\{1,2\},\{1,3\},\{2,3\}\}$

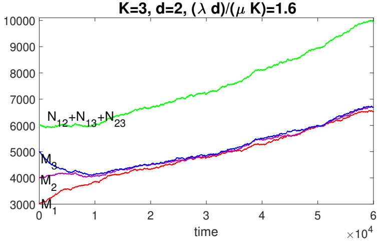


- $a_{cis}(t)$ attained service of the i-th type$-c$ job.
- $\frac{\mathrm{d}a_{cis}(t)}{\mathrm{d}t} = \frac{1}{M_s(t)}$.
- A job leaves the system due to a departure in server $s^*_{ci}(t) = \arg\max_{s \in c}\{a_{cis}(t)\}$.
- Departure rate of the i-th type-c job: $\frac{\mu}{M_{s^*_{ci}(t)}(t)}$.

PS service policy with Identical copies

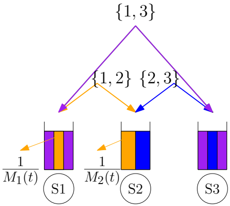
Example: $K = 3$ and $d = 2$ copies, $\mathcal{C} = \{\{1,2\},\{1,3\},\{2,3\}\}$

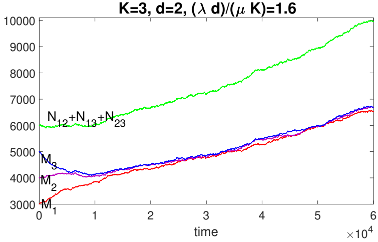




- The drift of server s: $\frac{\mathrm{d}m_s}{\mathrm{d}t} = \lambda \frac{d}{K} - \sum_{c \in \mathcal{C}(s)} \sum_{i=1}^{N_c(t)} \frac{\mu}{M_{s^*_{ci}(t)}(t)}.$

Example: $K = 3$ and $d = 2$ copies, $\mathcal{C} = \{\{1,2\},\{1,3\},\{2,3\}\}$





- The drift of server s: $\frac{\mathrm{d}m_s}{\mathrm{d}t} = \lambda \frac{d}{K} - \sum_{c \in \mathcal{C}(s)} \sum_{i=1}^{N_c(t)} \frac{\mu}{M_{s^*_{ci}(t)}(t)}$.
- When symmetric state $(M_1 = M_2 = M_3)$: $\frac{\mathrm{d}m_s}{\mathrm{d}t} = \lambda \frac{d}{K} - \mu$ which can be strictly positive when $\lambda < \mu K$.

PS system with Identical copies

Theorem

Assume PS service policy and copies of a job to be identical copies. The system is stable $\Longleftrightarrow \lambda < \mu \frac{K}{d}$.

Proof:

$\Longleftarrow)$

- Upper Bound $\vec{N}^{UP}(t)$: the system where all copies need to be served.
- $\vec{N}^{PS}(t) \leq_{st.} \vec{N}^{UP}(t)$
- $\vec{N}^{UP}(t)$ is stable iff $\lambda d < \mu K$

PS system with Identical copies

Theorem

Assume PS service policy and copies of a job to be identical copies. The system is stable $\iff \lambda < \mu \frac{K}{d}$.

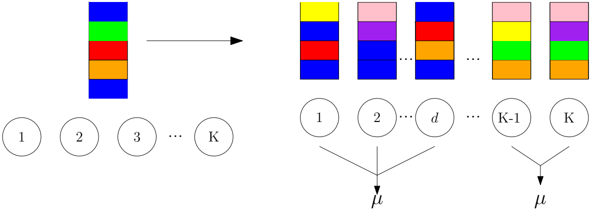
Proof:

$\Longrightarrow$)

- Lower Bound $\vec{N}^{LB}(t)$: the departure rate of a job is determined by the capacity it gets at the server with the least number of copies: $\frac{\mu}{M_{s_c^*}(t)}$ where $s_c^* = \arg\min_{s \in \mathcal{S}(c)}\{M_s(t)\}$.
- $\vec{N}^{PS}(t) \geq_{st.} \vec{N}^{LB}(t)$, since $\frac{\mu}{M_{s_{ci}^*(t)}(t)} \leq \frac{\mu}{M_{s_c^*}(t)}$.
- The fluid limit of $\vec{N}^{LB}(t)$ satisfies $\frac{\mathrm{d}m_{min}(t)}{\mathrm{d}t} = \lambda \frac{d}{K} - \mu > 0$

FCFS system with Identical copies

Stability condition reduces **at least** to $\lambda < \mu(K - d + 1)$.



FCFS system with Identical copies

Theorem

Under FCFS service policy and identical copies the system is stable $\iff$

$$\lambda < \tilde{\mu} = \sum_{i \in \tilde{S}} \tilde{\Pi}_i i \mu$$

where $\tilde{\Pi}_i$ *is the fraction of time one sees departure rate* $i\mu$ *when the system is congested.*

FCFS system with Identical copies

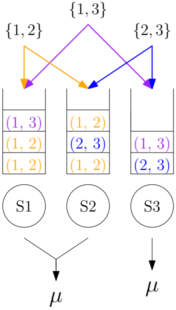
The solution of the congested system:

- K and $d = K - 1$, **Stability condition:** $\lambda < 2\mu$.

FCFS system with Identical copies

The solution of the congested system:

- K and $d = K - 1$, **Stability condition:** $\lambda < 2\mu$.
  Example: $K = 3$ and $d = 2$



FCFS system with Identical copies

The solution of the congested system:

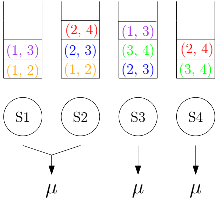
- K and $d = K - 1$, **Stability condition:** $\lambda < 2\mu$.
- For general K and d is hard to characterize.

FCFS system with Identical copies

The solution of the congested system:

- K and $d = K - 1$, **Stability condition:** $\lambda < 2\mu$.
- For general K and d is hard to characterize.

Example: $K = 4$ and $d = 2$.

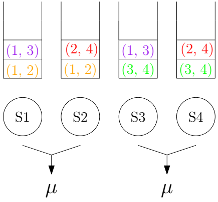


FCFS system with Identical copies

The solution of the congested system:

- K and $d = K - 1$, **Stability condition:** $\lambda < 2\mu$.
- For general K and d is hard to characterize.

Example: $K = 4$ and $d = 2$.



FCFS system with Identical copies

The solution of the congested system:

- K and $d = K-1$, **Stability condition:** $\lambda < 2\mu$.
- For general K and d is hard to characterize.
  Example: $K=4$ and $d=2$. The steady-state equations are:

$$2\mu\pi(O_2,n,O_1) = \mu\pi(O_2,n+1,O_1) + \mu\sum_{j=0}^{n}(\tfrac{1}{6})^{j+1}\pi(O_2,n-j,O_1)$$
$$+\mu\sum_{s=1}^{4}(\tfrac{1}{3})^{n}\pi(O_2,n,O_1,0,O_s) + \mu(\tfrac{1}{6})^{n+1}\pi(O_1,0,O_2)$$
$$+\mu\sum_{s=1}^{4}\sum_{j=0}^{n}\mu(\tfrac{1}{3})^{j}\pi(O_2,j,O_s,n-j,O_1)$$

$$3\mu\pi(O_3,m,O_2,n,O_1) = \mu\pi(O_3,m,O_2,n+1,O_1)$$
$$+\mu\sum_{s=1,2}\sum_{j=0}^{n}(\tfrac{1}{3})^{j+1}\pi(O_3,m+j+1,O_s,n-j,O_1)$$
$$+\mu\sum_{s=1}^{3}\sum_{j=0}^{m}(\tfrac{3}{6})^{j}\tfrac{1}{6}(O_s,m-j,O_2,n,O_1)$$
$$+\mu(\tfrac{1}{3})^{n}(\tfrac{3}{6})^{m}\tfrac{1}{6}\sum_{s=1,2}\pi(O_2,n,O_1,0,O_s)$$
$$+\mu(\tfrac{1}{3})^{n+1}\sum_{s=1,2}\pi(O_3,m+n+1,O_1,0,O_s)$$
$$+\mu\sum_{s=1,2}\sum_{j=0}^{n}(\tfrac{1}{3})^{j}(\tfrac{3}{6})^{m}\tfrac{1}{6}\pi(O_2,j,O_s,n-j,O_1)$$
$$+\mu\sum_{j=0}^{n}(\tfrac{1}{6})^{j+2}(\tfrac{3}{6})^{m}(\tilde{O},n-j,O_1)$$
$$+\mu(\tfrac{1}{6})^{n+2}(\tfrac{3}{6})^{m}\pi(O_1,0,\tilde{O}),$$

ROS system with Identical copies

At a fluid scale,
$P($ a job is served simultaneously in more than one server$) \to 0$.

Theorem

Under ROS service policy and identical copies assumption, the system is stable $\iff \lambda < \mu K$

Proof:

- Show that fluid limit satisfies $\frac{\mathrm{d}m_{max}(t)}{\mathrm{d}t} \leq \lambda \frac{d}{K} - \mu d$

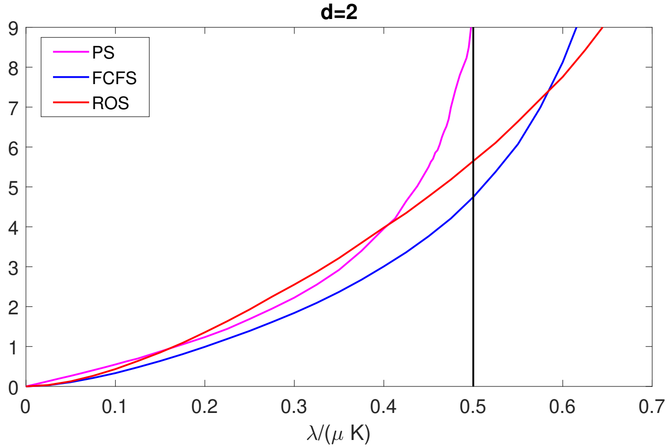
Conclusion

Table: Summary of stability conditions

	PS	FCFS	ROS	Priority policy
i.i.d	$\lambda < \mu K$	$\lambda < \mu K$	$\lambda < \mu K$	$\lambda << \mu K$
i.c.	$\lambda < \mu \frac{K}{d}$	$\lambda < \tilde{\mu}$ $(\tilde{\mu} < \mu(K-(d-1)))$	$\lambda < \mu K$	–

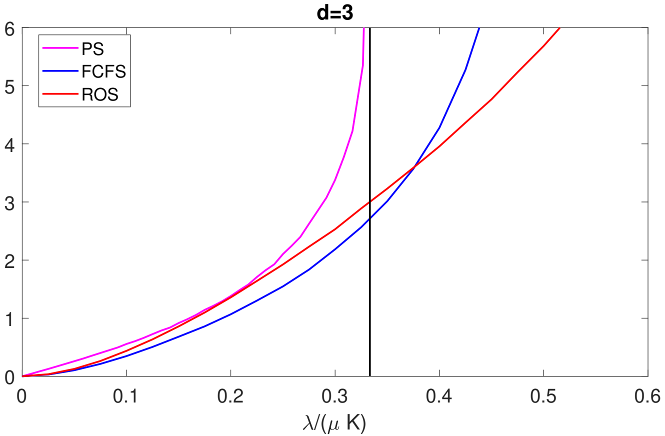
Simulations for the mean number of jobs

Mean number of jobs with identical copies and $K = 5$.



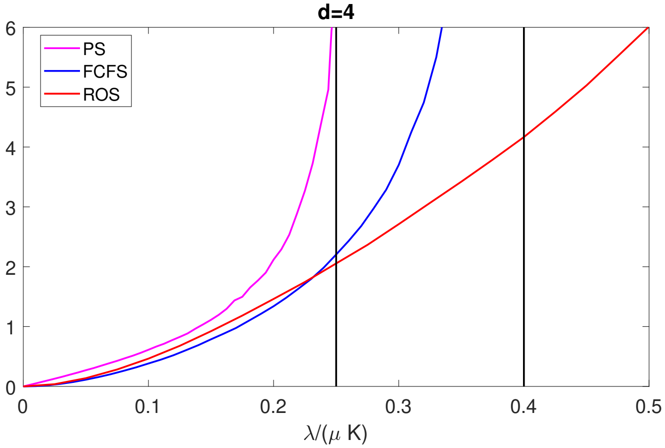
Simulations for the mean number of jobs

Mean number of jobs with identical copies and $K = 5$.



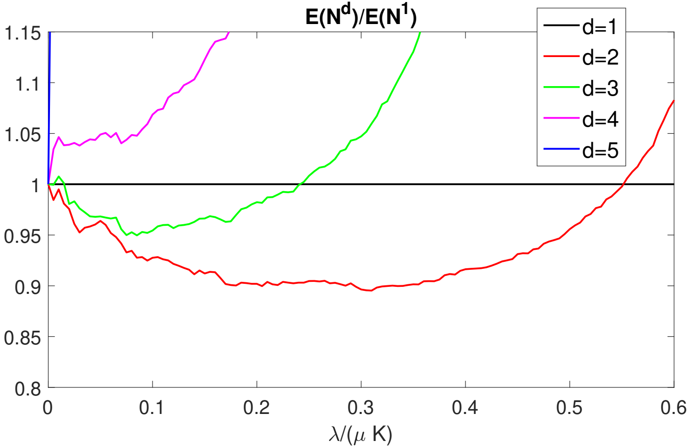
Simulations for the mean number of jobs

Mean number of jobs with identical copies and $K = 5$.



LT approximation for FCFS with identical copies

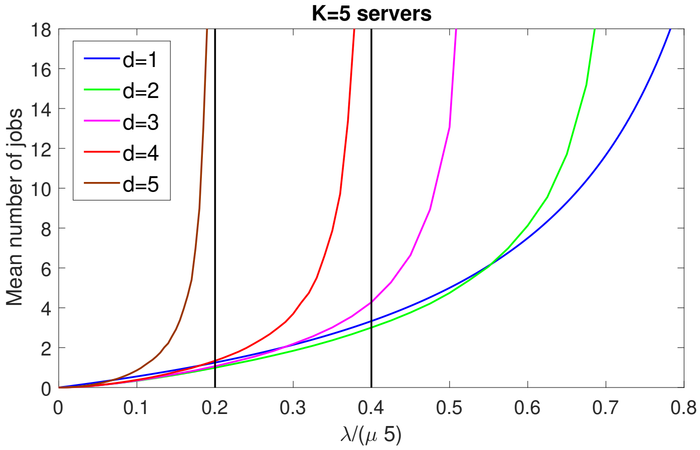
Relative mean response time under low load of λ.



$\mathbb{E}(D^{LT,FCFS}) = \frac{1}{\mu} + \frac{3\lambda}{2\mu^2}\frac{1}{\binom{K}{d}}$,

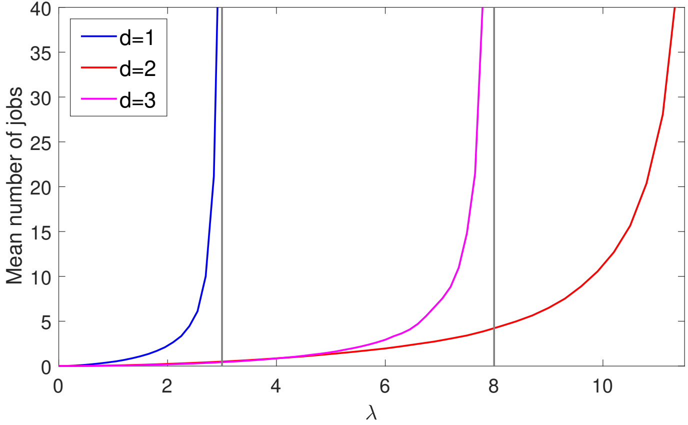
$\min \mathbb{E}(D^{LT,FCFS})$ when $d^* = \arg\max_d\{\binom{K}{d}\} = 2$.

Homogeneous servers for FCFS with identical copies



Heterogeneous speed servers for FCFS with identical copies

$K = 3$ and $\mu = (1, 4, 8)$



Conclusions

- Redundancy systems under iid assumption:
  - FCFS, PS and ROS are maximum stable.
  - Priority queues lose stability.
- Redundancy system under identical copies assumption: Stability condition strongly depends on the scheduling policy.
- Heterogeneous servers can improve stability.

Future work

- Redundancy systems under iid assumption:
  Analyse sufficient conditions for which the system is maximum stable.

Future work

- Redundancy systems under iid assumption:
  Analyse sufficient conditions for which the system is maximum stable.
- Redundancy system under identical copies assumption:
  Characterize the stability condition when variable servers: heterogeneous speed servers, S&X model,...

Thank you!