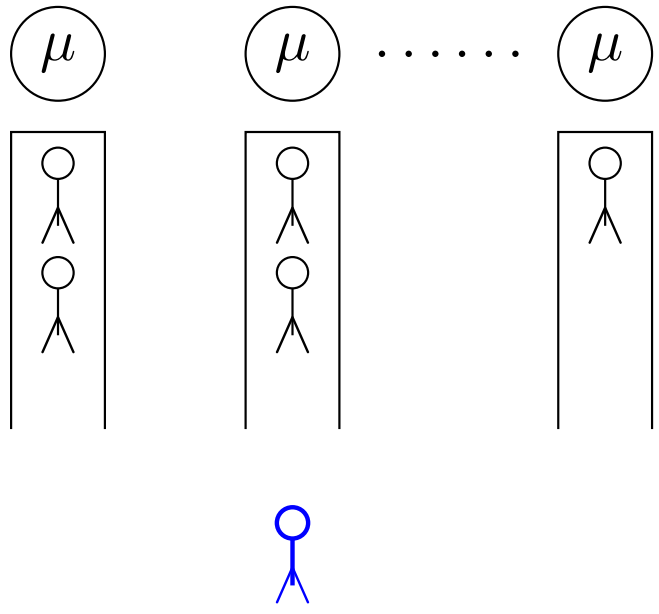
On the stability of redundancy models

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

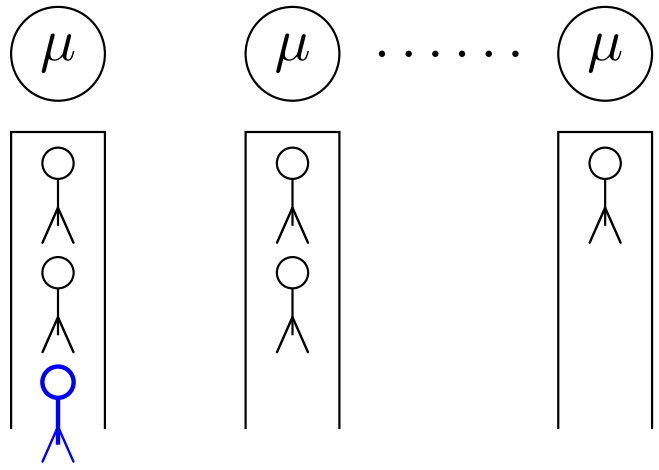
IRIT-CNRS and ENSEEIHT

November 12, 2019

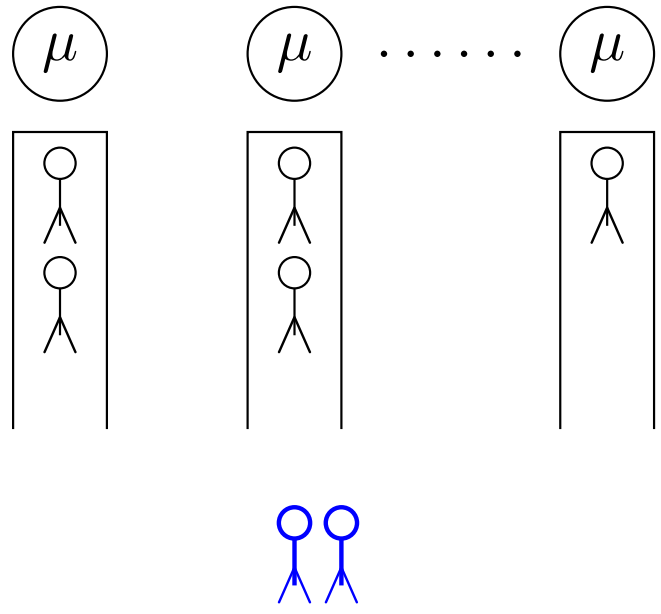
Load-balancing strategies:



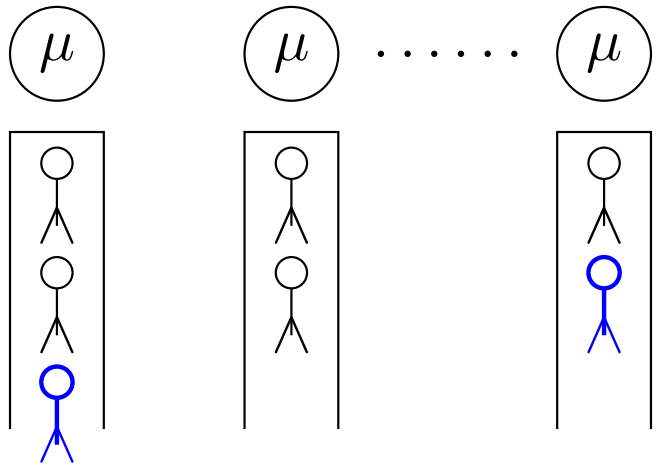
Load-balancing strategies:



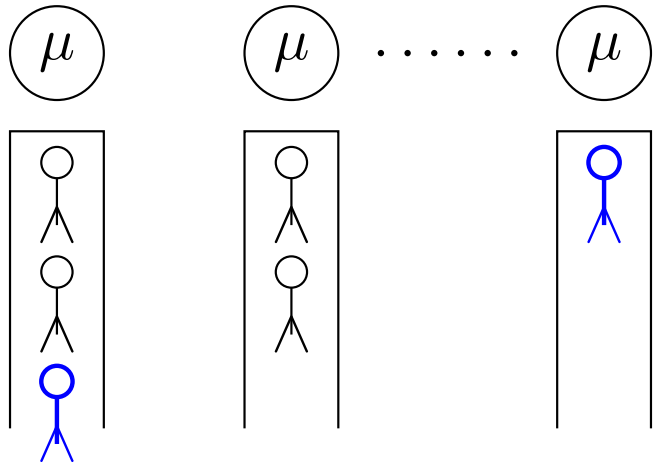
Redundancy-d: A job is dispatched into several servers.



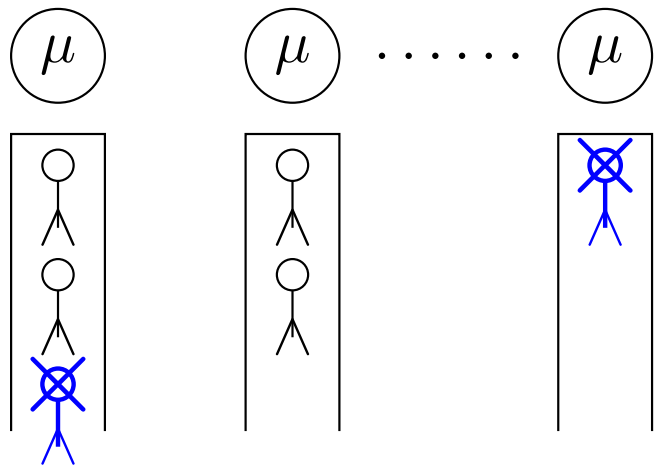
Redundancy-d: A job is dispatched into several servers.



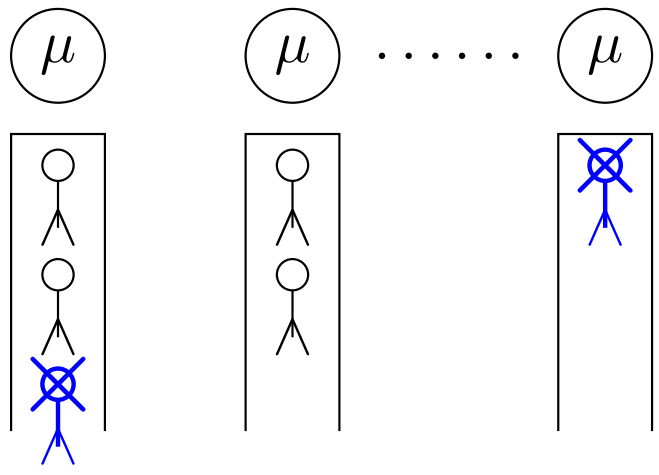
Redundancy-d: A job is dispatched into several servers.



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



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



Exploit variability in the queue lengths 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 $\iff \lambda < \mu K$.

[Gardner 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

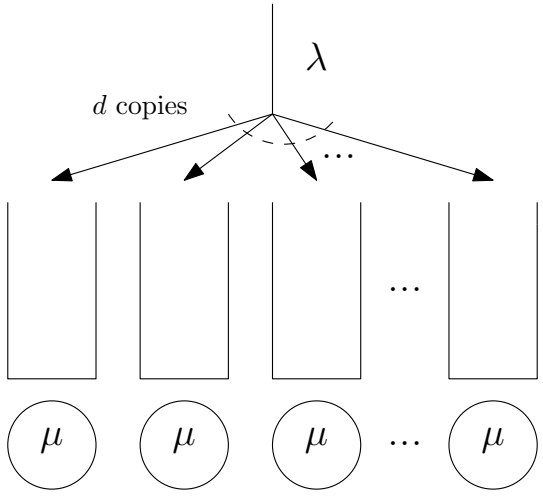
Determine how the stability condition is impacted by:

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

Outline

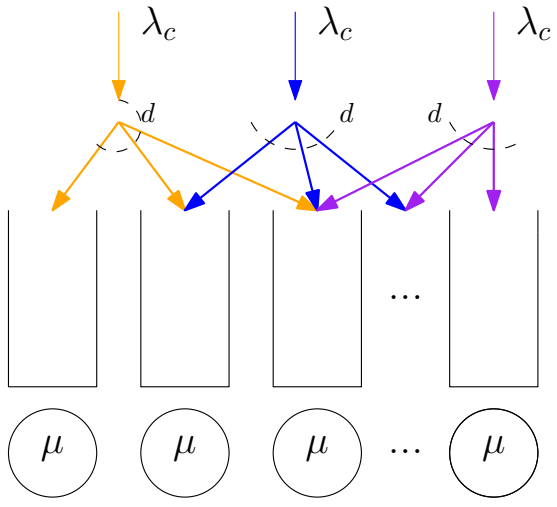
1. Model description
2. I.i.d copies
   - PS service policy
   - ROS service policy
   - Priority policy
3. Identical copies
   - PS service policy
   - FCFS service policy
   - ROS service policy
4. Numerical results
5. Conclusions

Model description



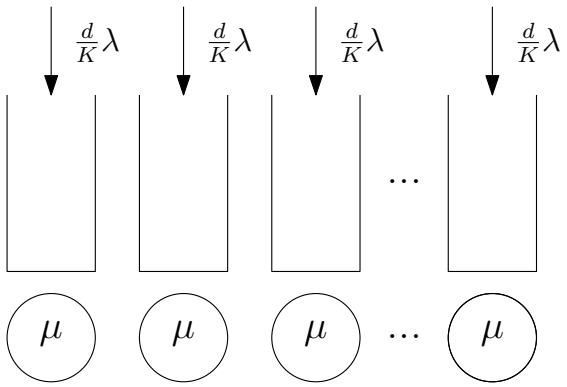
- K servers with capacity 1.
- Poisson arrivals with rate λ.
- Exponential service times with parameter μ.
- Total load in the system: $\rho = \frac{\lambda}{K\mu}$.

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 jobs at rate $\lambda_c = \frac{\lambda}{\binom{K}{d}}$.

Model description



- Arrival rate of copies to a server is $\frac{\binom{K-1}{d-1}}{\binom{K}{d}}\lambda = \frac{d}{K}\lambda$.
- 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

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

- 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

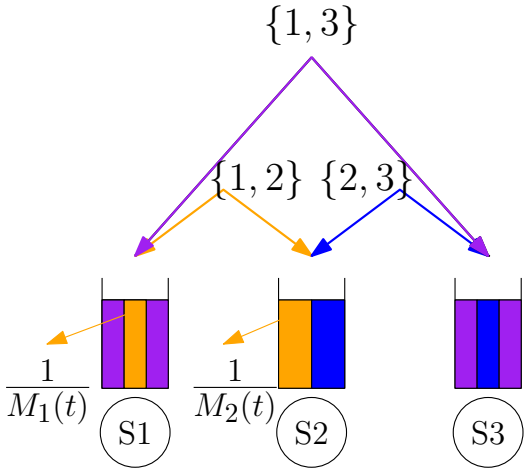
	PS	FCFS	ROS	Priority policy
i.i.d	$\lambda < \mu K$	$\lambda < \mu K$	$\lambda < \mu K$	$\lambda << \mu K$
i.c.	$\lambda < \mu \frac{K}{d}$	$\lambda < \bar{\ell}\mu$ $(\bar{\ell} < (K - (d-1)))$	$\lambda < \mu K$	–

PS service policy with iid

Theorem

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

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:

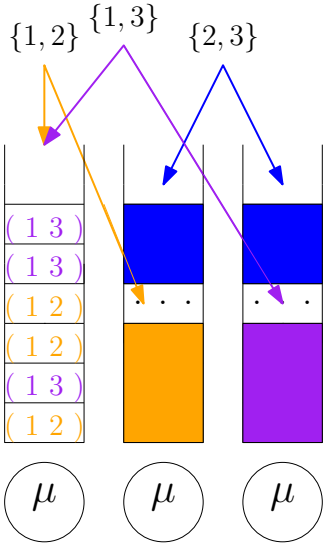
- Show that fluid limit satisfies

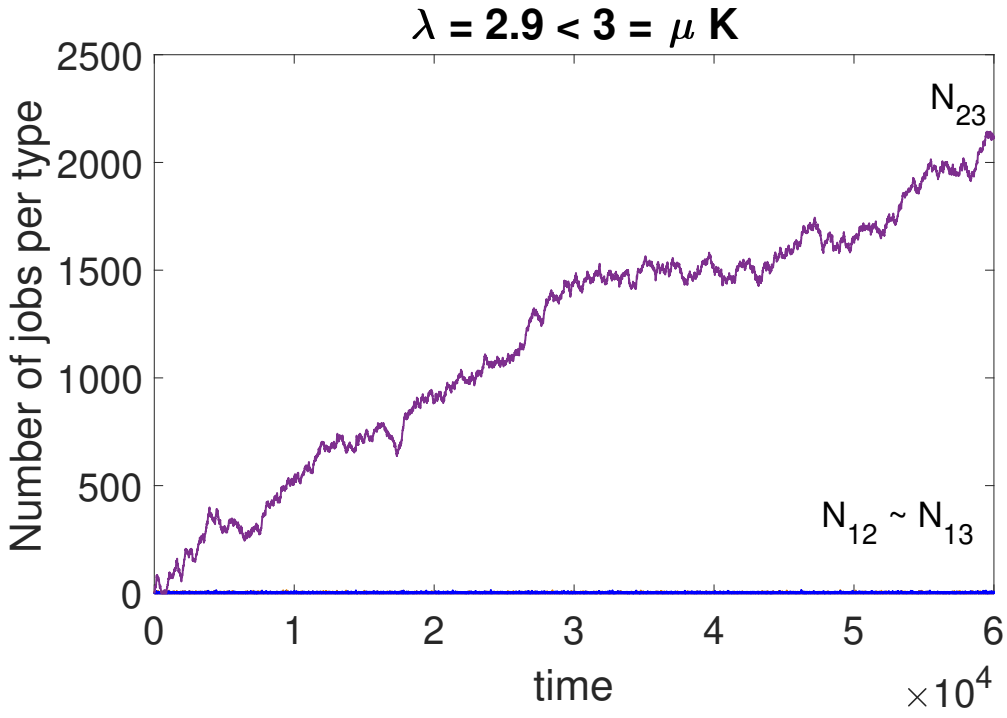
$$\frac{\mathrm{d}m_{max}(t)}{\mathrm{d}t} = \lambda\frac{d}{K} - \mu\left(\sum_{c\in\mathcal{C}(s)}\sum_{l\in S(c)}\frac{n_c}{m_l}\right) \leq \lambda\frac{d}{K} - \mu d$$

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





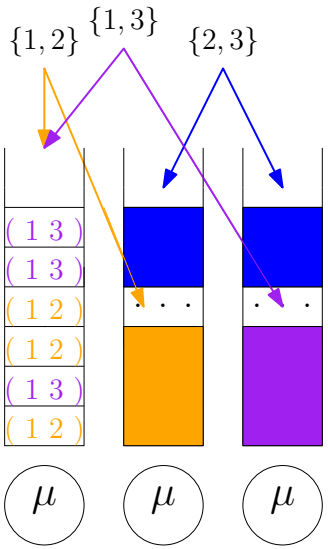
$$\frac{\mathrm{d}|\vec{n}(t)|}{\mathrm{d}t} = \lambda - (3\mu - \mu P(\text{ server 1 is empty })) =$$

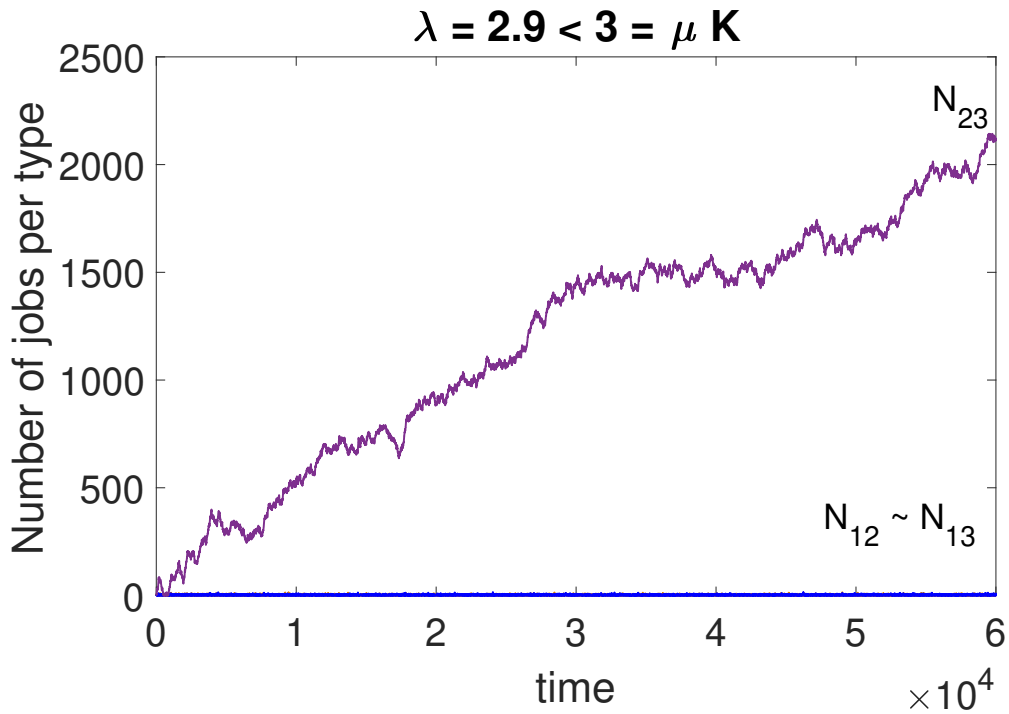
$$= \lambda - (3\mu - \mu \frac{(2\mu - \frac{\lambda}{3})^2 (3\mu - \frac{2\lambda}{3})}{4\mu^2(3\mu - \frac{2\lambda}{3}) + (\frac{\lambda}{3})^2 \mu}) = 2.9 - (3 - 0.23) = 0.13.$$

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

1. Model description
2. I.i.d copies
   - PS service policy
   - ROS service policy
   - Priority policy
3. Identical copies
   - PS service policy
   - FCFS service policy
   - ROS service policy
4. Numerical results
5. Conclusions

Identical copies assumption

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

Identical copies assumption

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

Identical copies assumption

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

Identical copies assumption

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

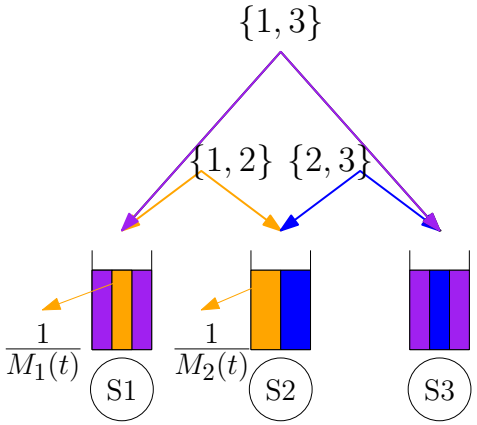
The stability region decreases in d

PS service policy with Identical copies

Theorem

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

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



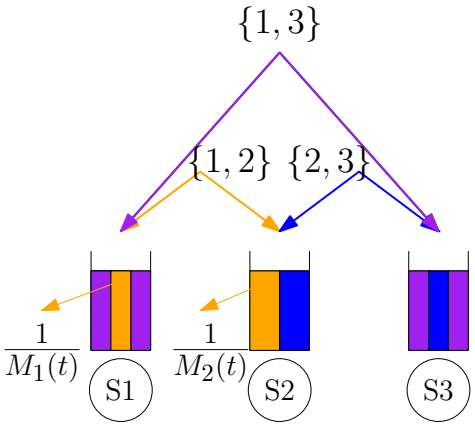
- $a_{cis}(t)$ attained service of the i-th type$-c$ job.
- $\frac{\mathrm{d}a_{cis}(t)}{\mathrm{d}t} = \frac{1}{M_s(t)}$.
- A job leaves the system due to a departure in server $s^*_{ci}(t) = \arg\max_{s \in c}\{a_{cis}(t)\}$.
- Departure rate of server s: $\sum_{c \in \mathcal{C}(s)} \sum_{i=1}^{N_c(t)}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

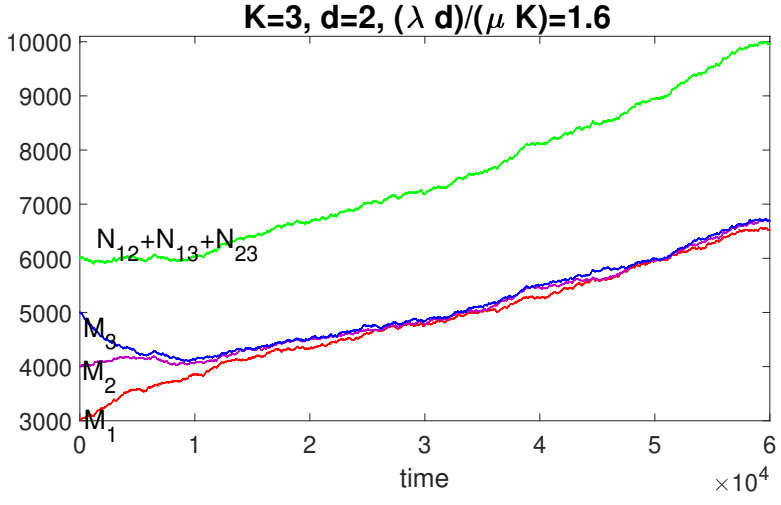
PS service policy with Identical copies

Theorem

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

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





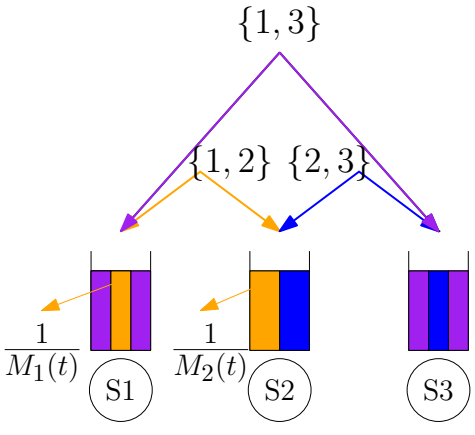
- The drift of server s: $\frac{\mathrm{d}m_s}{\mathrm{d}t} = \lambda \frac{d}{K} - \sum_{c \in \mathcal{C}(s)} \sum_{i=1}^{N_c(t)} \frac{\mu}{M_{s^*_{ci}(t)}(t)}$.

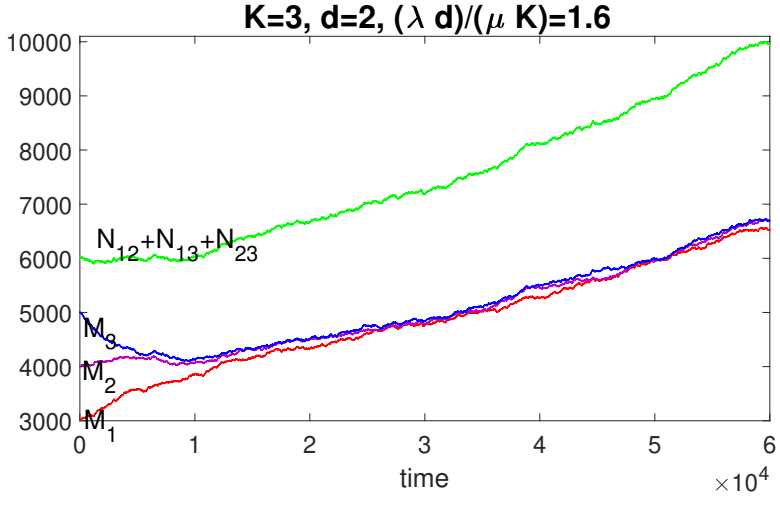
PS service policy with Identical copies

Theorem

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

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 $\iff \lambda < \mu \frac{K}{d}$.

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:

$\Longrightarrow$)

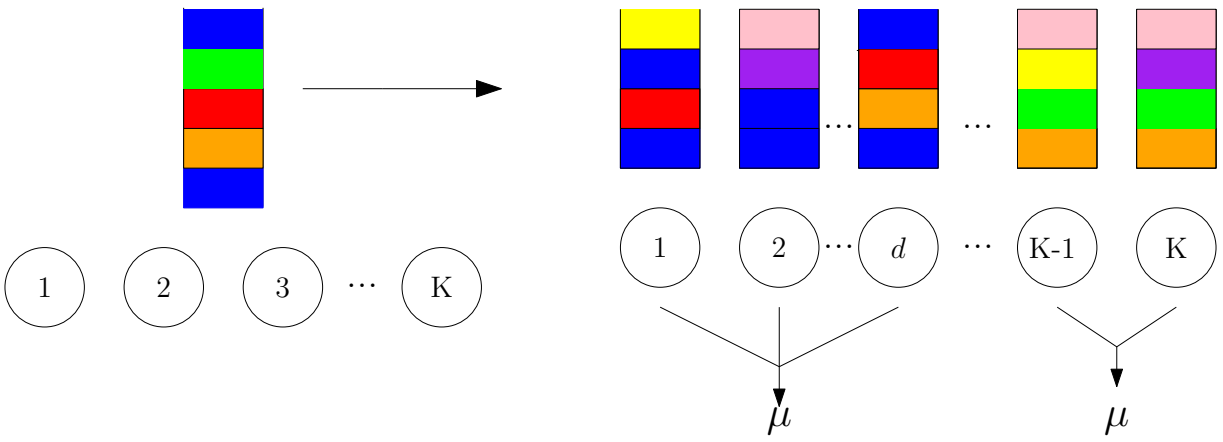
- Lower Bound $\vec{N}^{LB}(t)$: the departure rate of a job is determined by the capacity it gets at the server with the least number of copies: $\frac{\mu}{M_{s_c^*}(t)}$ where $s_c^* = \arg\min_{s \in \mathcal{S}(c)}\{M_s(t)\}$.
- $\vec{N}^{PS}(t) \geq_{st.} \vec{N}^{LB}(t)$, since $\frac{\mu}{M_{s_{ci}^*(t)}(t)}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

FCFS system with Identical copies

Theorem

Under FCFS service policy and identical copies the system is stable $\iff \lambda < \bar{\ell}\mu$.

Example:



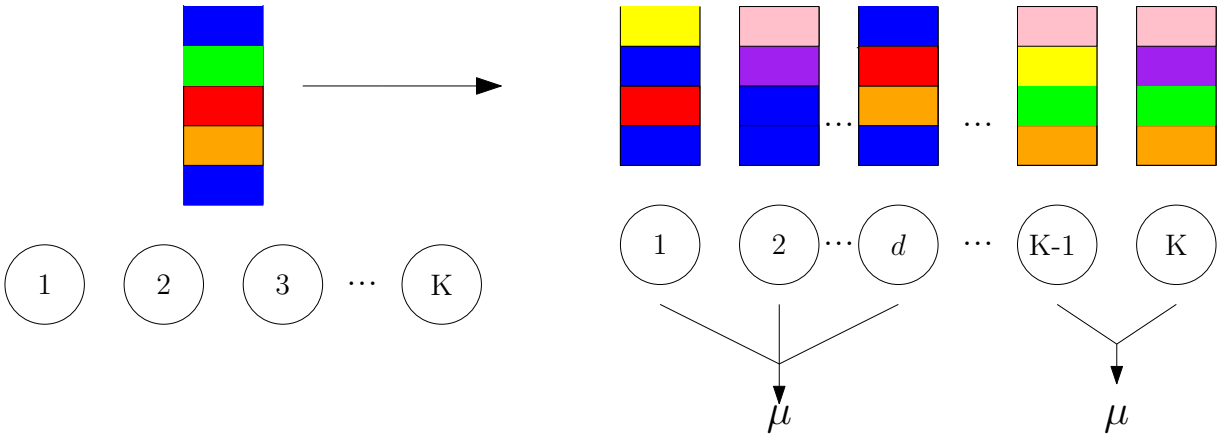
Stability condition is **at least** $\lambda < \mu(K - (d-1))$.

FCFS system with Identical copies

Theorem

Under FCFS service policy and identical copies the system is stable $\iff \lambda < \bar{\ell}\mu$.

Example:



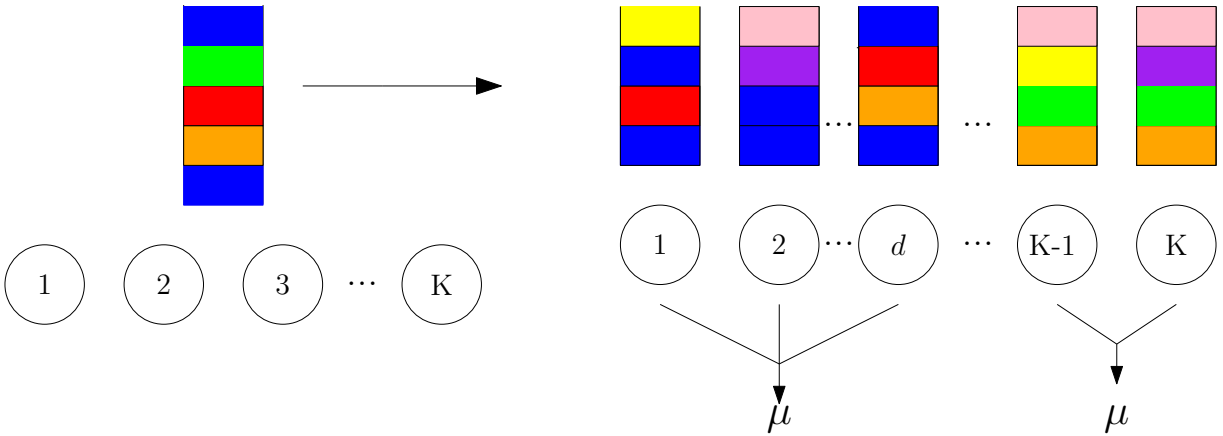
Saturated system: An infinite backlog of jobs waiting in the system, sampled uniformly over types.

FCFS system with Identical copies

Theorem

Under FCFS service policy and identical copies the system is stable $\iff \lambda < \bar{\ell}\mu$.

Example:



Saturated system: An infinite backlog of jobs waiting in the system, sampled uniformly over types.

$\bar{\ell}$ is the mean number of jobs on service on the saturated system.

FCFS system with Identical copies

Table: Value of $\bar{\ell}/K$.

$\bar{\ell}/K$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$
$d=1$	1	1	1	1	1	1
$d=2$	0.5	0.66	0.71	0.74	0.76	0.77
$d=3$		0.33	0.5	0.54	0.57	0.58
$d=4$			0.25	0.4	0.43	0.46
$d=5$				0.2	0.33	0.36
$d=6$					0.16	0.28
$d=7$						0.14

FCFS system with Identical copies

Table: Value of $\bar{\ell}/K$.

$\bar{\ell}/K$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$
$d=1$	1	1	1	1	1	1
$d=2$	0.5	0.66	0.71	0.74	0.76	0.77
$d=3$		0.33	0.5	0.54	0.57	0.58
$d=4$			0.25	0.4	0.43	0.46
$d=5$				0.2	0.33	0.36
$d=6$					0.16	0.28
$d=7$						0.14

$d=1$
$\bar{\ell}=K$

FCFS system with Identical copies

Table: Value of $\bar{\ell}/K$.

$\bar{\ell}/K$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$
$d=1$	1	1	1	1	1	1
$d=2$	0.5	0.66	0.71	0.74	0.76	0.77
$d=3$		0.33	0.5	0.54	0.57	0.58
$d=4$			0.25	0.4	0.43	0.46
$d=5$				0.2	0.33	0.36
$d=6$					0.16	0.28
$d=7$						0.14

$d = K$

$\bar{\ell} = 1$

FCFS system with Identical copies

Table: Value of $\bar{\ell}/K$.

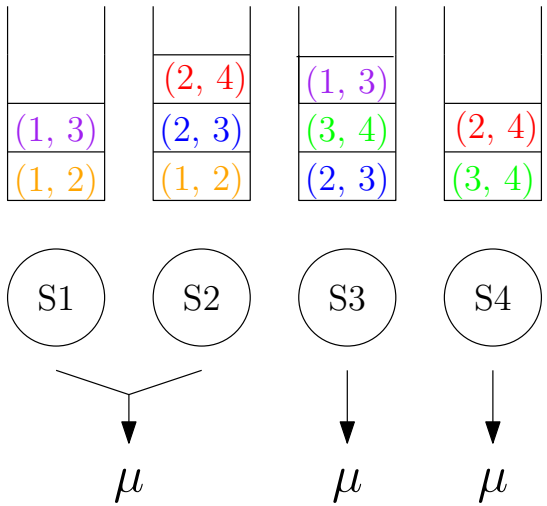
$\bar{\ell}/K$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$
$d=1$	1	1	1	1	1	1
$d=2$	0.5	0.66	0.71	0.74	0.76	0.77
$d=3$		0.33	0.5	0.54	0.57	0.58
$d=4$			0.25	0.4	0.43	0.46
$d=5$				0.2	0.33	0.36
$d=6$					0.16	0.28
$d=7$						0.14

$d = K - 1$

$\bar{\ell} = 2$

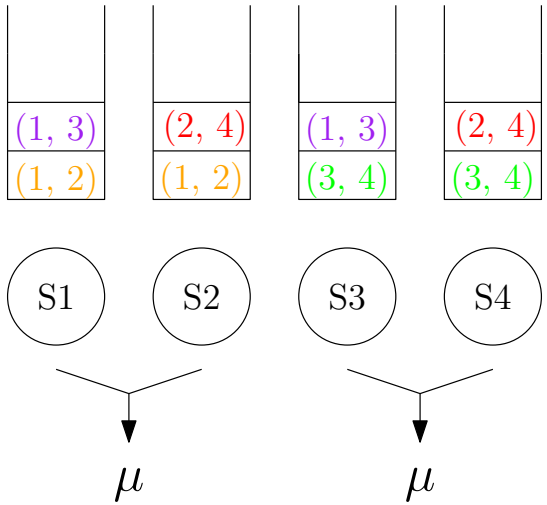
FCFS system with Identical copies

The solution of the congested system for $K = 4$ and $d = 2$.



FCFS system with Identical copies

The solution of the congested system for $K = 4$ and $d = 2$.



FCFS system with Identical copies

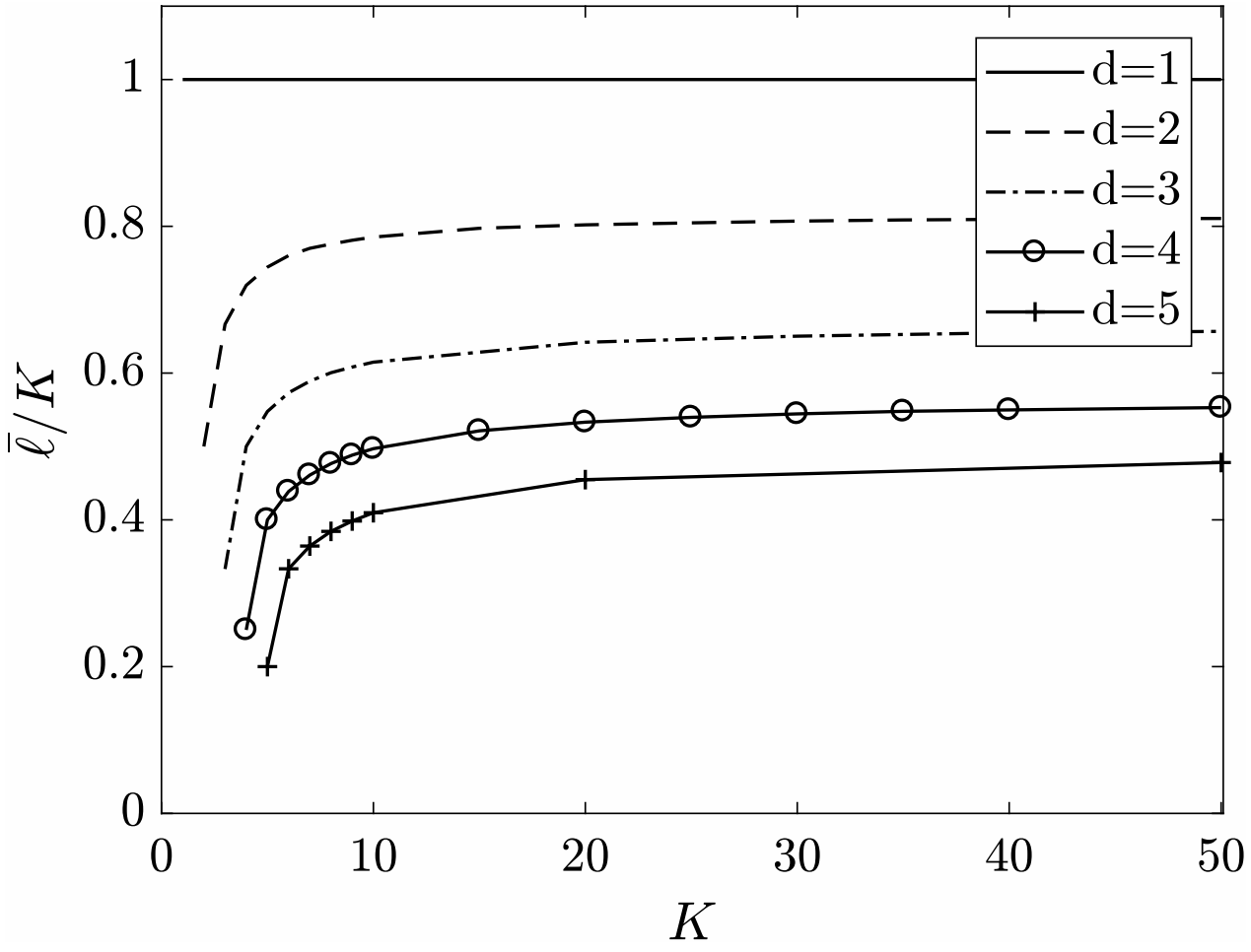
Table: Value of $\bar{\ell}/K$.

$\bar{\ell}/K$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$
$d=1$	1	1	1	1	1	1
$d=2$	0.5	0.66	0.71	0.74	0.76	0.77
$d=3$		0.33	0.5	0.54	0.57	0.58
$d=4$			0.25	0.4	0.43	0.46
$d=5$				0.2	0.33	0.36
$d=6$					0.16	0.28
$d=7$						0.14

- $\bar{\ell}/K$ is increasing on the number of servers K.

FCFS system with Identical copies

Value of $\bar{\ell}/K$.



ROS system with Identical copies

Theorem

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

Proof:

- At a fluid scale,
  $P($ a job is served simultaneously in more than one server$) \to 0$.
- Show that fluid limit satisfies $\frac{\mathrm{d}m_{max}(t)}{\mathrm{d}t} \leq \lambda \frac{d}{K} - \mu d$

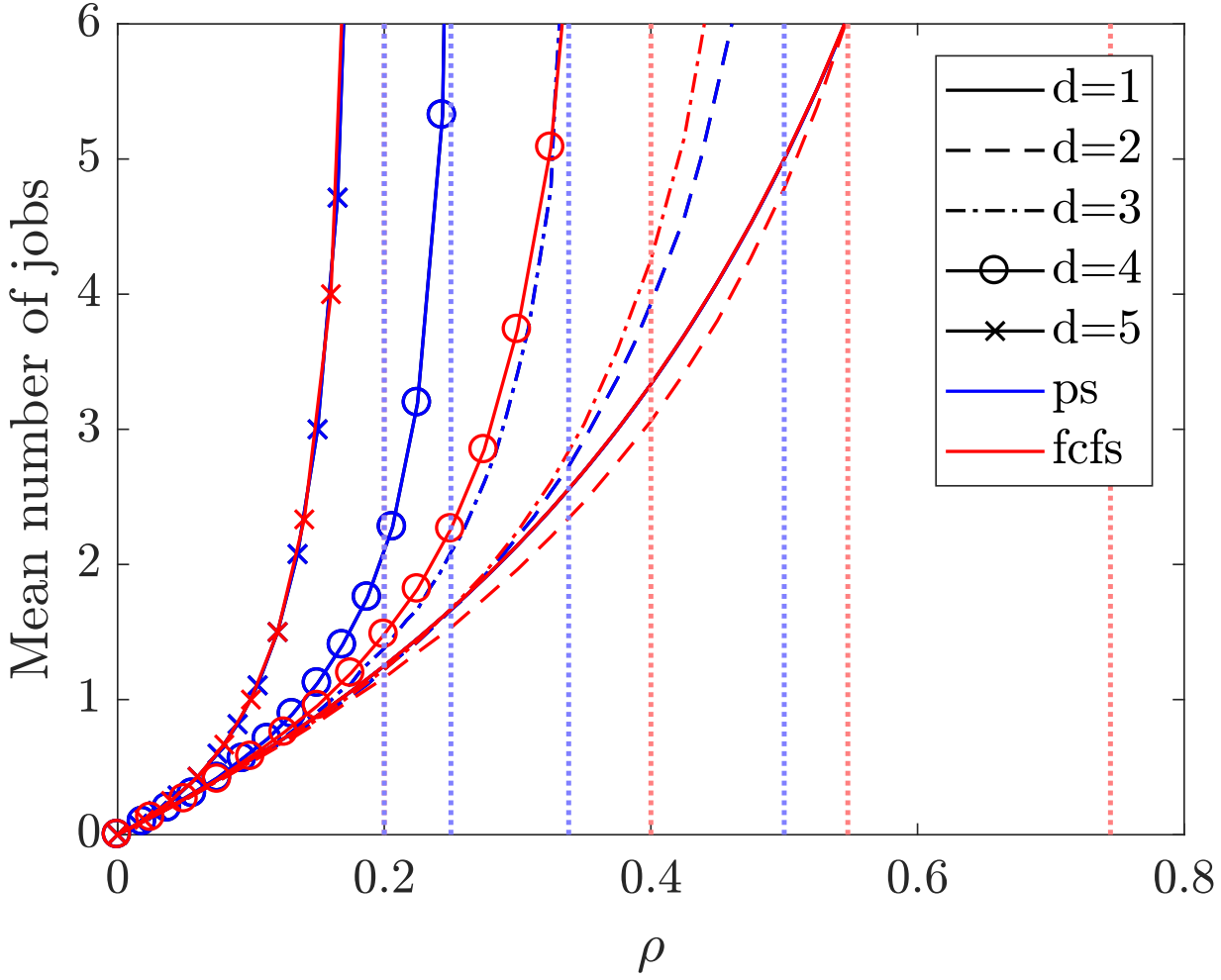
Conclusion

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.	$\lambda < \mu \frac{K}{d}$	$\lambda < \bar{\ell}\mu$	$\lambda < \mu K$	–

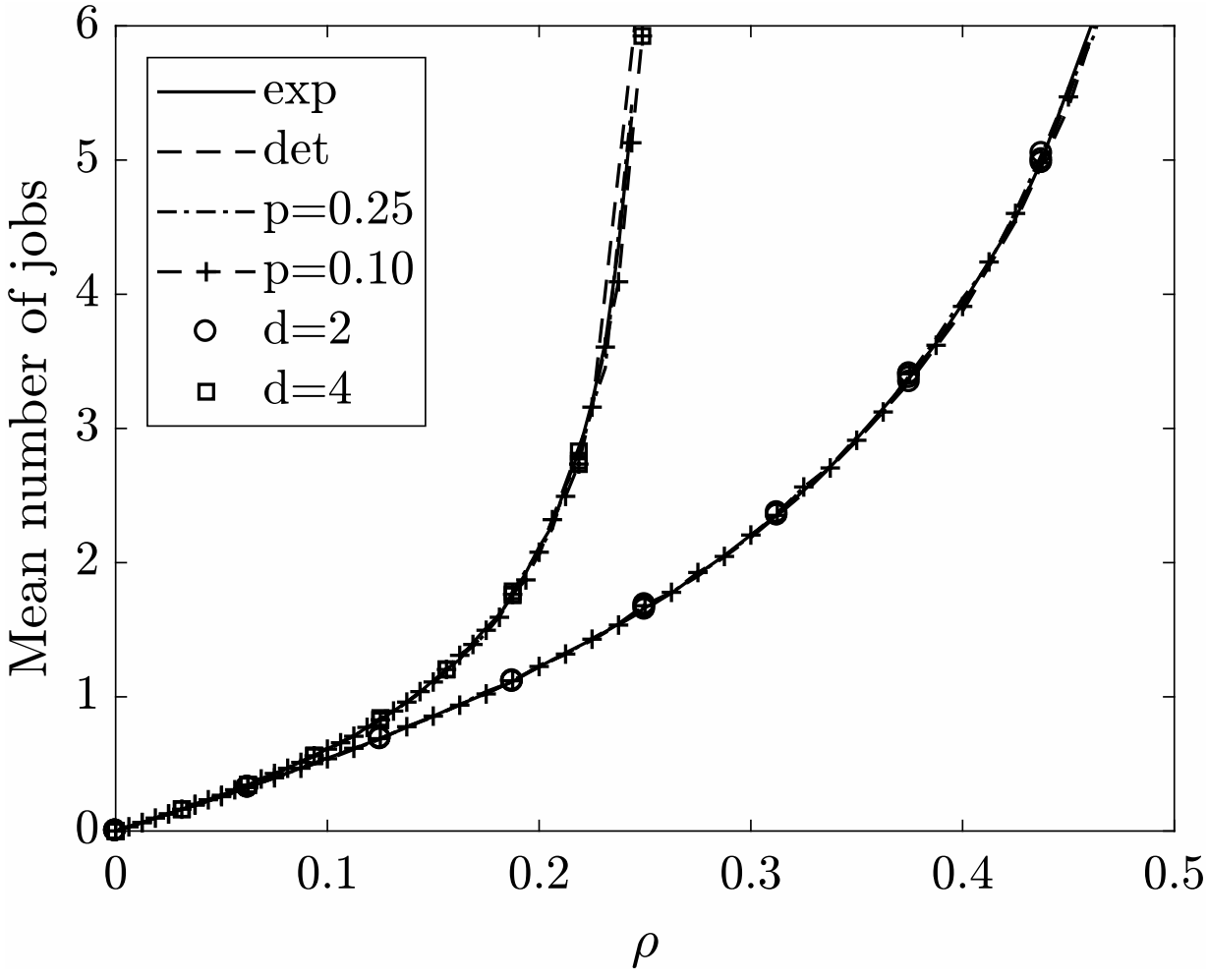
Simulations for the mean number of jobs

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



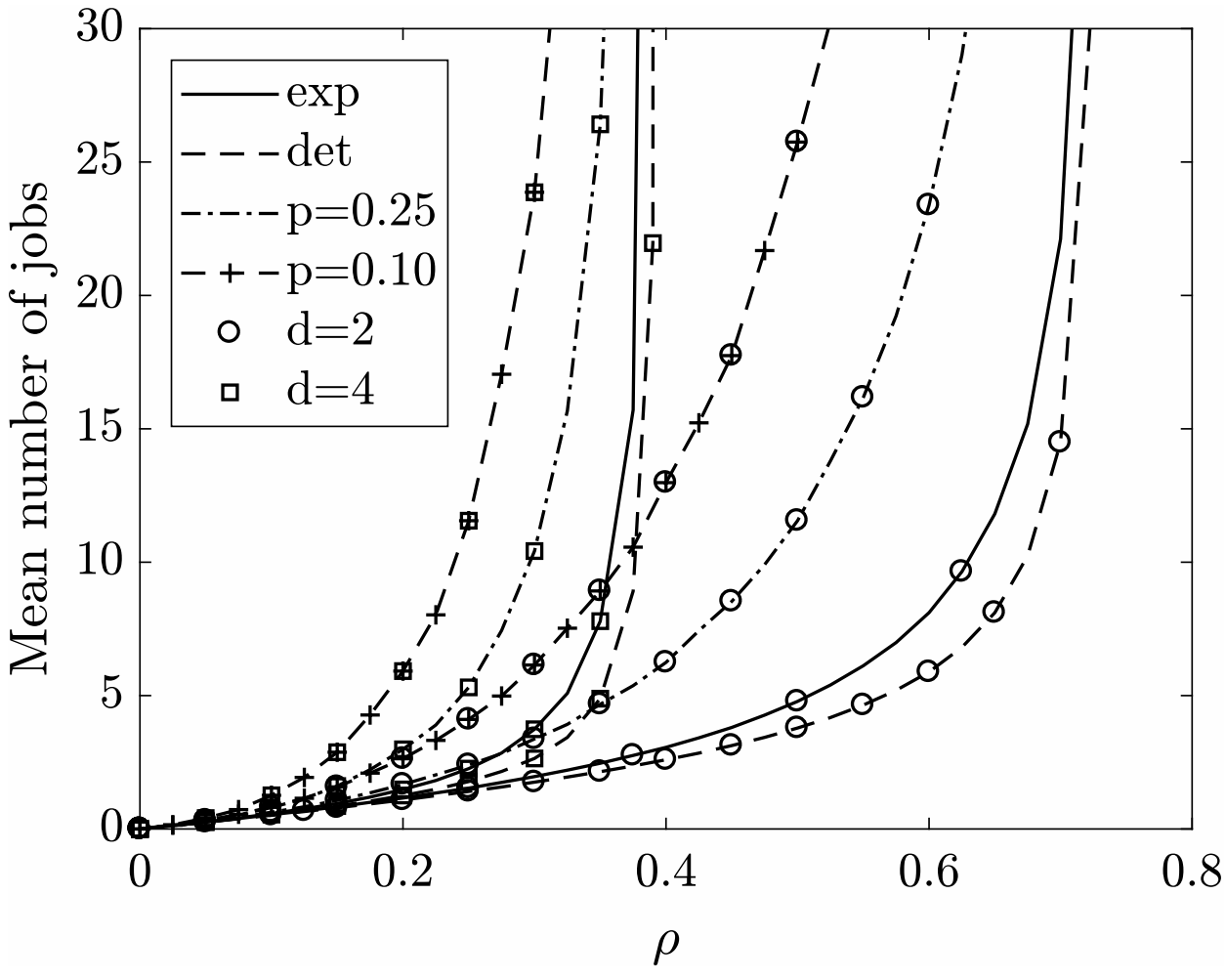
Non-exponential service requirements: PS

Mean number of jobs under PS service policy with identical copies and exponential, deterministic and degenerate hyperexponential service requirements.



Non-exponential service requirements: FCFS

Mean number of jobs under FCFS service policy with identical copies and exponential, deterministic and degenerate hyperexponential service requirements.



Conclusions

- Redundancy systems under iid assumption:
  - FCFS, PS and ROS are as stable as if there was no redundancy.
  - Priority queues lose stability.

Conclusions

- Redundancy systems under iid assumption:
  - FCFS, PS and ROS are as stable as if there was no redundancy.
  - Priority queues lose stability.

Future work: Obtain sufficient conditions for which the stability condition is not impacted by redundancy.

Conclusions

- Redundancy systems under iid assumption:
  - FCFS, PS and ROS are as stable as if there was no redundancy.
  - Priority queues lose stability.

  Future work: Obtain sufficient conditions for which the stability condition is not impacted by redundancy.
- Redundancy system under identical copies assumption:
  - Stability condition strongly depends on the scheduling policy.

Conclusions

- Redundancy systems under iid assumption:
  - FCFS, PS and ROS are as stable as if there was no redundancy.
  - Priority queues lose stability.

  Future work: Obtain sufficient conditions for which the stability condition is not impacted by redundancy.
- Redundancy system under identical copies assumption:
  - Stability condition strongly depends on the scheduling policy.

  Future work: Characterize the stability condition when variable servers: heterogeneous speed servers.

References

- ⋆ E. Anton, U. Ayesta, M. Jonckheere, I. M.Verloop. *On the stability of redundancy models* arXiv:1903.04414, 2019