Redundancy scheduling with scaled Bernoulli service requirements

YEQT 2018

Youri Raaijmakers
joint work with Onno Boxma & Sem Borst

Redundancy models

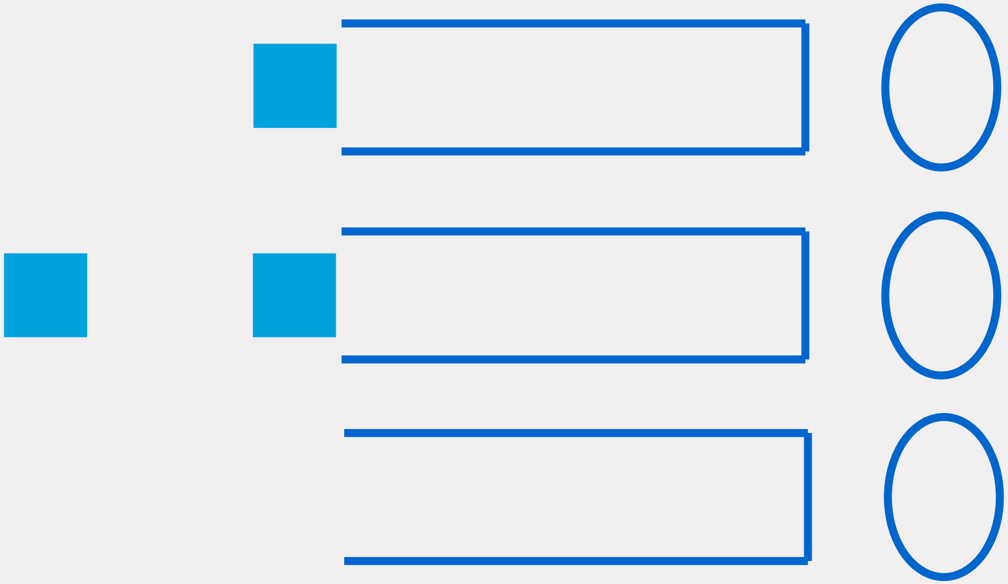
Notation:

- N servers
- d replicas

Assumptions:

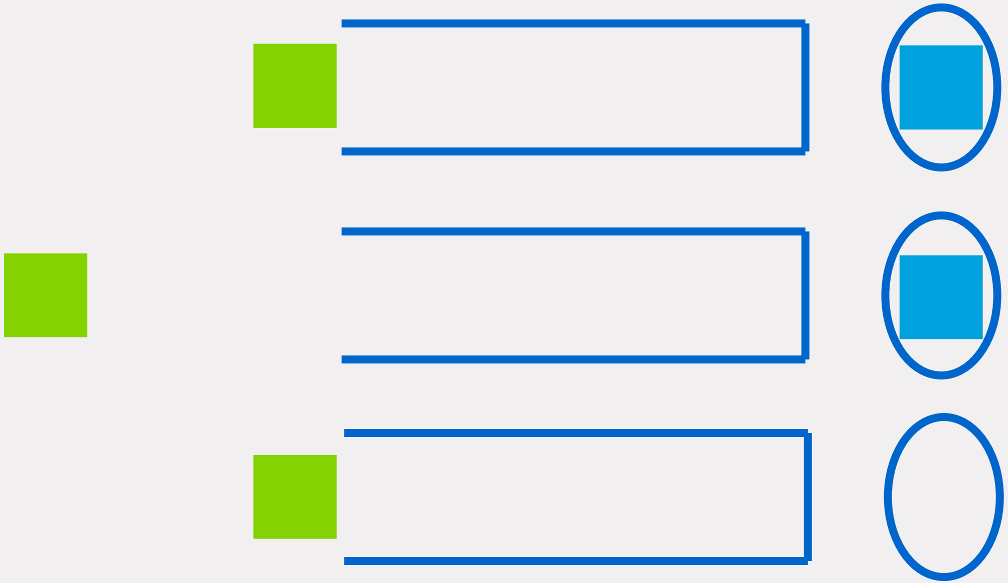
- Independent and identically distributed (i.i.d.) replicas
- Cancel on completion (c.o.c.)

Redundancy models



Scenario: $N = 3$ and $d = 2$

Redundancy models

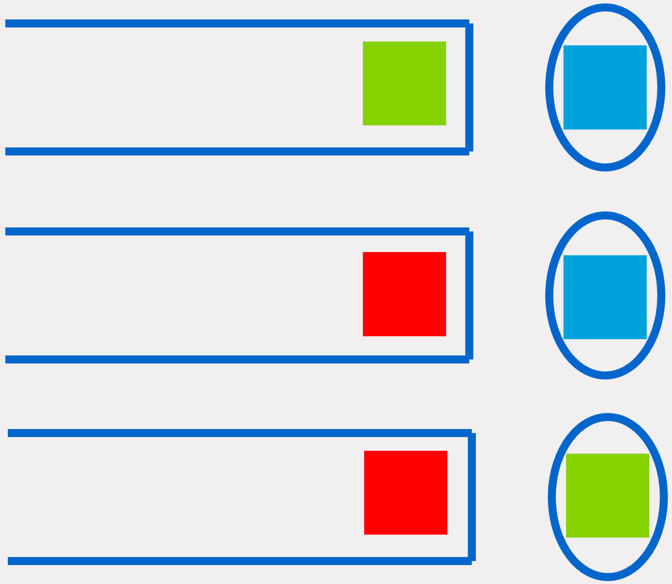


Scenario: $N = 3$ and $d = 2$

Redundancy models

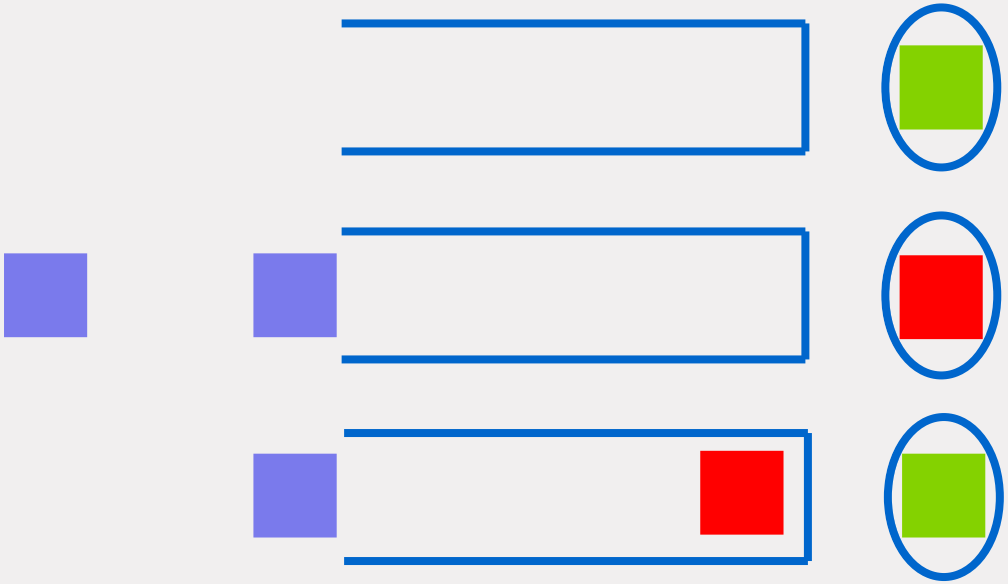
Scenario: $N = 3$ and $d = 2$

Redundancy models



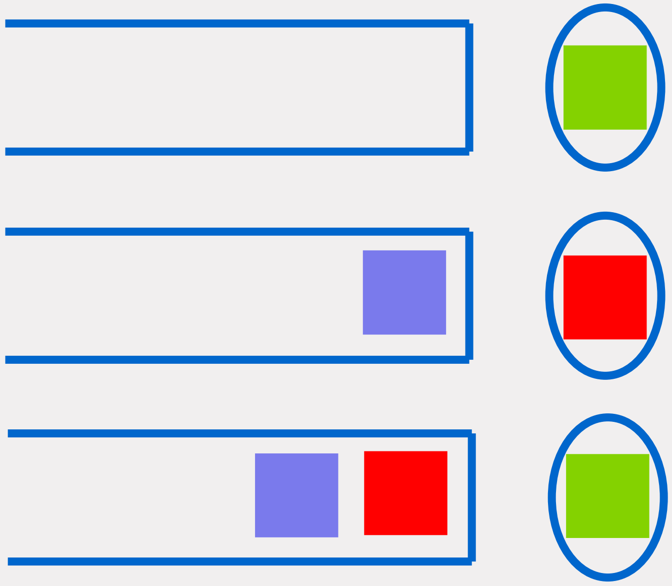
Scenario: $N = 3$ and $d = 2$

Redundancy models



Scenario: $N = 3$ and $d = 2$

Redundancy models



Scenario: $N = 3$ and $d = 2$

Outline

- What is already known regarding the stability condition?
- Sufficient stability condition
- Asymptotically necessary stability condition for scaled Bernoulli service requirements

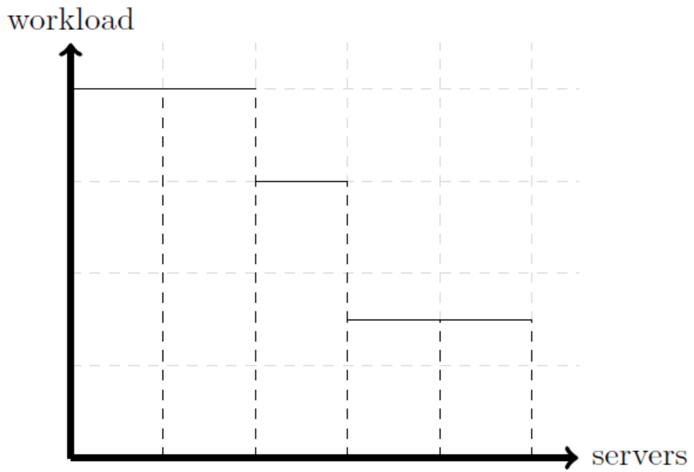
What is already known?

- Exponential service requirements studied by Gardner *et al.* [1]
  - Stability condition: $\rho = \frac{\lambda}{N\mu} < 1$
  - Observation: stability condition does NOT depend on d

[1] K. Gardner, M. Harchol-Balter, A. Scheller-Wolf, M. Velednitsky, S. Zbarsky (2017). Redundancy-d: The power of d choices for redundancy. Operation Research **65 (4)** 1078-1094.

Sufficient stability condition

- Real workloads $\boldsymbol{\omega}$
  - Example: Let $\boldsymbol{\omega} = (4.1, 4.1, 3.8, 2.5)$ and consider an arrival on servers 2 and 4 with $b = (1.5, 1.1)$ then $\boldsymbol{\omega}_{new} = (4.1, 4.1, 3.8, 3.6)$

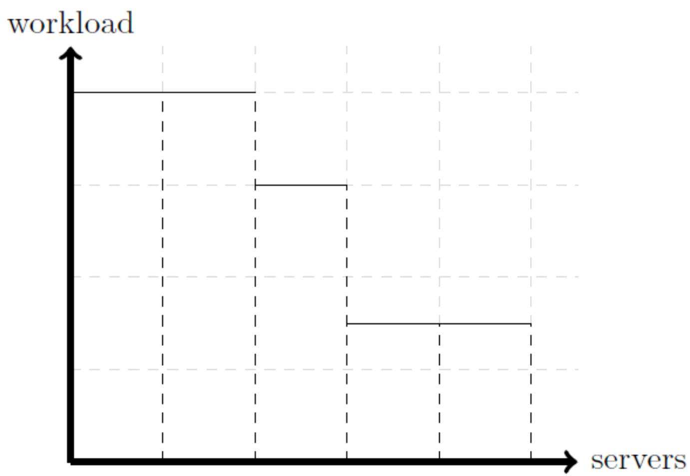


Sufficient stability condition

- **Lemma 1:** Maximum workload is upper bounded by the workload in a corresponding *M/G/1* queue with $\lambda_{MG1} = \lambda$ and $B_{MG1} = \min\{B_1, \dots, B_d\}$

- Our service requirements
  - $B = \begin{cases} X \cdot K, & w.p.\ 1 - p = \frac{1}{K} \\ 0, & w.p.\ p = 1 - \frac{1}{K} \end{cases}$

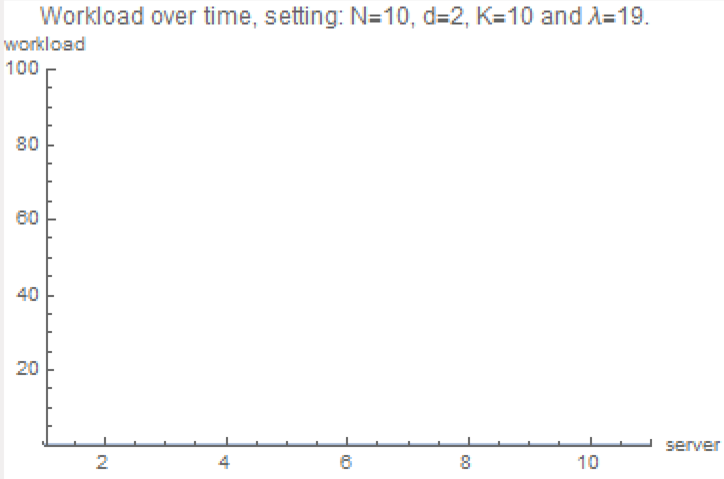
  where K is a fixed positive real number, and X is a general str
  $\mathbb{E}[X] = 1$ and we assume $\mathbb{E}[B] = 1$

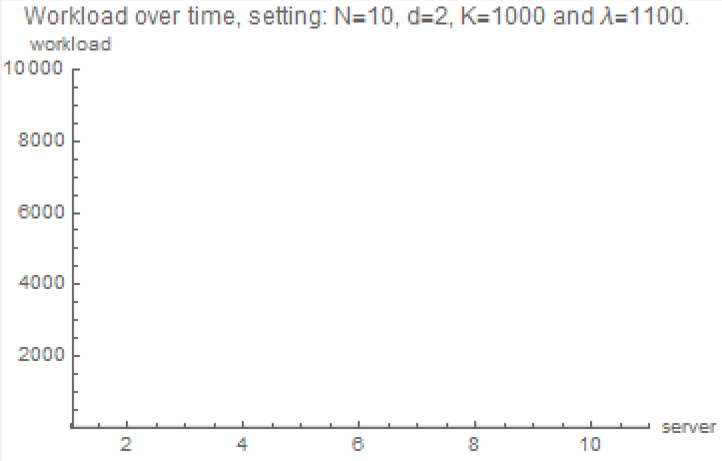


- **Theorem 1:** A sufficient stability condition is given by

$$\lambda \cdot \mathbb{E}\left[\min\{B_1, \dots, B_d\}\right] = \frac{\lambda \cdot \mathbb{E}\left[\min\{X_1 K, \dots, X_d K\}\right]}{K^d} < 1$$

Asymptotically necessary condition





$K = 10$ vs. $K = 1000$

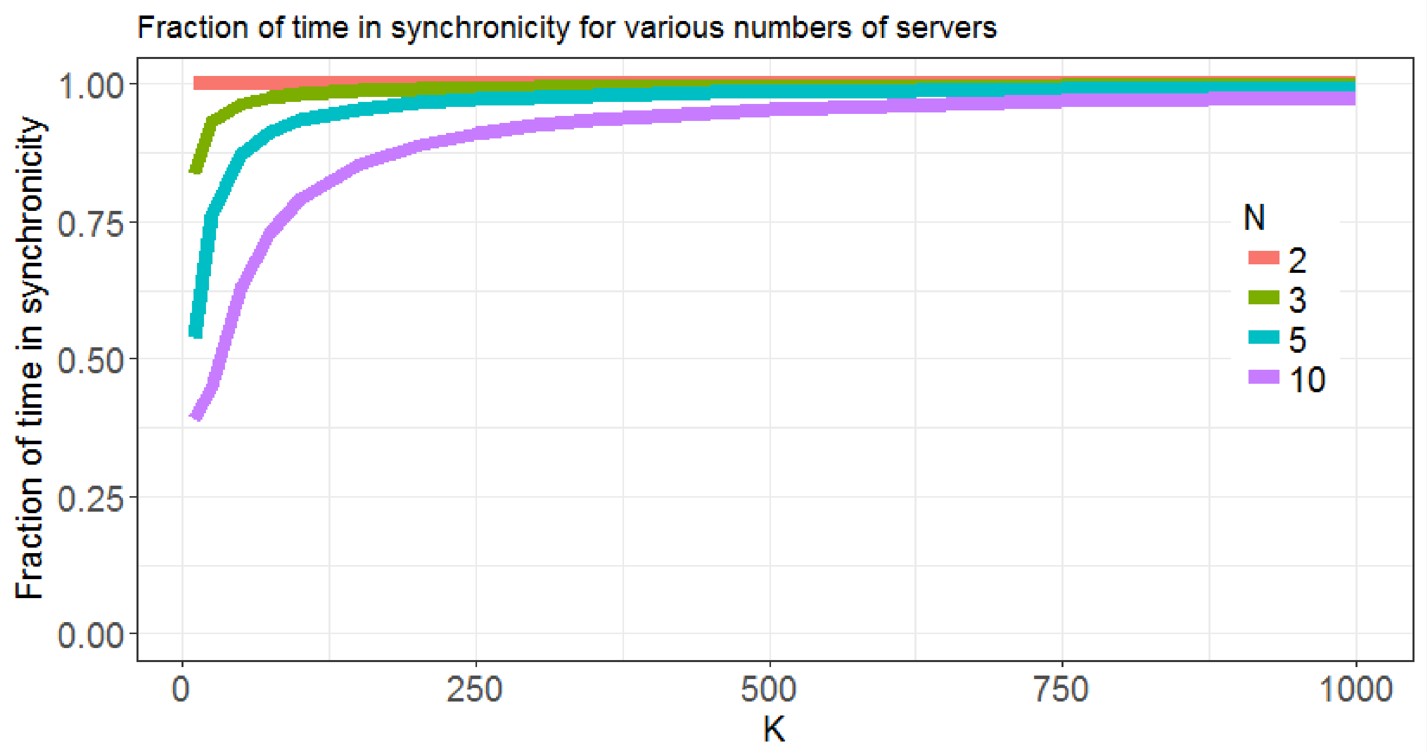
Asymptotically necessary condition

- **Lemma 2:** For every $\epsilon > 0$ there exists a $K_\epsilon(d, N)$ such that for all $K > K_\epsilon(d, N)$ the system is at least a fraction $(1 - \epsilon)$ of the time in so-called synchronicity in the long term

- **Proof:**
  1) Expected time in synchronicity: $\frac{1}{(1-p)^d \lambda} = \frac{K^d}{\lambda}$
  2) Expected time not in synchronicity: $\frac{1}{\lambda} O(K)$
     - Only jobs for which all replicas have service requirements $X \cdot K$ increase the maximum workload
     - Expected time not in synchronicity conditioned on number increases in maximum workload

$$B = \begin{cases} X \cdot K, & w.p. \quad 1 - p = \frac{1}{K} \\ 0, & w.p. \quad p = 1 - \frac{1}{K} \end{cases}$$

Asymptotically necessary condition



Asymptotically necessary condition

- **Lemma 2:** For every $\epsilon > 0$ there exists a $K_\epsilon(d, N)$ such that for all $K > K_\epsilon(d, N)$ the system is at least a fraction $1 - \epsilon$ of the time in so-called synchronicity in the long term

- **Theorem 2:** For every $\epsilon > 0$ there exists a $K_\epsilon(d, N)$ such that for all $K > K_\epsilon(d, N)$ the system with scaled Bernoulli service requirements is not stable when

$$(1-\epsilon)\frac{\lambda \cdot \mathbb{E}\left[\min\{X_1 K, \ldots, X_d K\}\right]}{K^d} > 1$$

$$B = \begin{cases} X \cdot K, & w.p. \quad 1 - p = \frac{1}{K} \\ 0, & w.p. \quad p = 1 - \frac{1}{K} \end{cases}$$

Asymptotically stability condition

- Combining Theorems 1 and 2 gives that for every $\epsilon > 0$ there exists a $K_\epsilon(d, N)$ such that for all $K > K_\epsilon(d, N)$ the stability condition for scaled Bernoulli service requirements is given by

$$\frac{\lambda \cdot \mathbb{E}\left[\min\{X_1 K, \ldots, X_d K\}\right]}{K^d} < 1$$

  - Observation: stability condition does NOT depend on N

Conclusions and further research

- Stability condition for scaled-Bernoulli service requirements depends on d, but not on N
- Extension to general arrival processes
- Extension to other service requirement distributions

Thank you!