Heavy traffic analysis of redundancy routing in queueing networks

YEQT 2016

Gal Mendelson

Technion- Israel Institute of Technology

Electrical Engineering

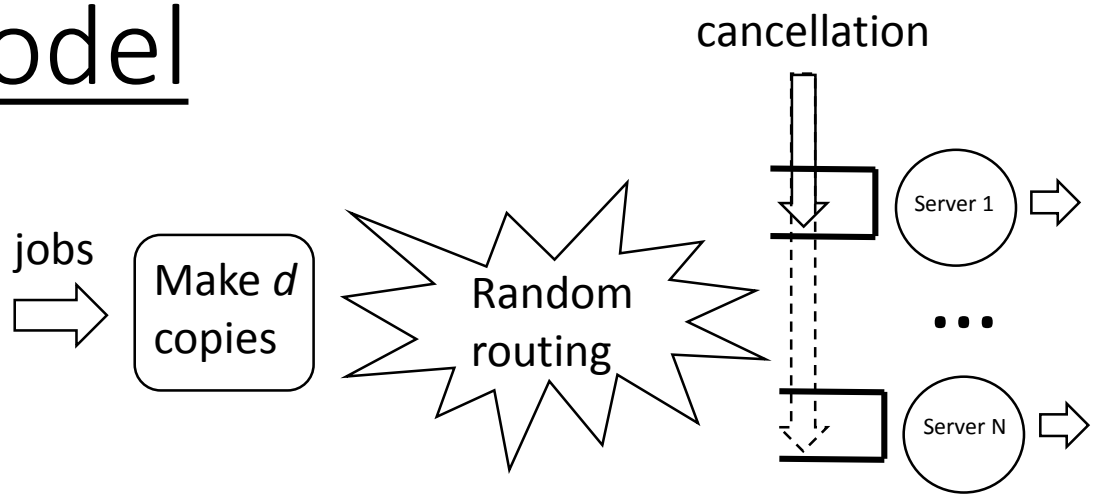
With: Rami Atar and Isaac Keslassy

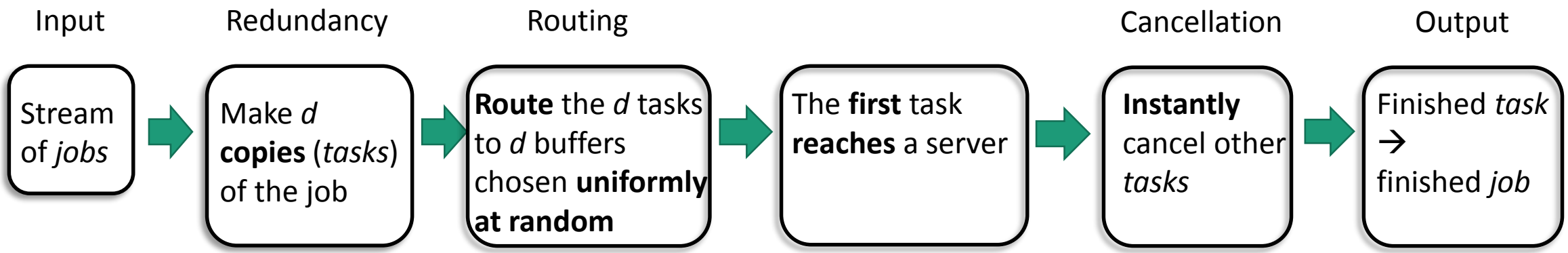
Talk overview

- **Redundancy routing** – model for data center architecture
- **Related work**
- **Goal**: heavy traffic analysis with fixed number of servers
- **Key observation** – equivalence to ‘Join the least workload’
- **Main results** – state space collapse and diffusion limit
- **Application** – optimality in ‘map-reduce’-like schemes

Redundancy routing model

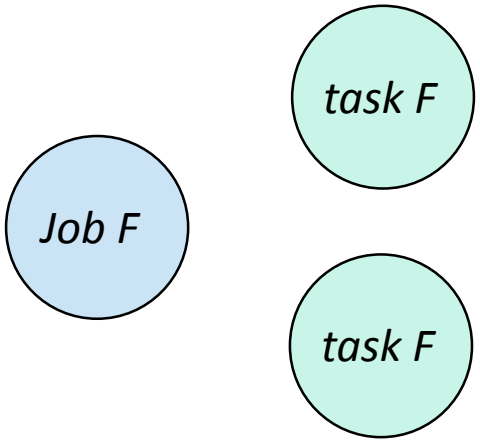
- *N* (fixed) work conserving servers
- Infinite capacity buffers
- First come first serve

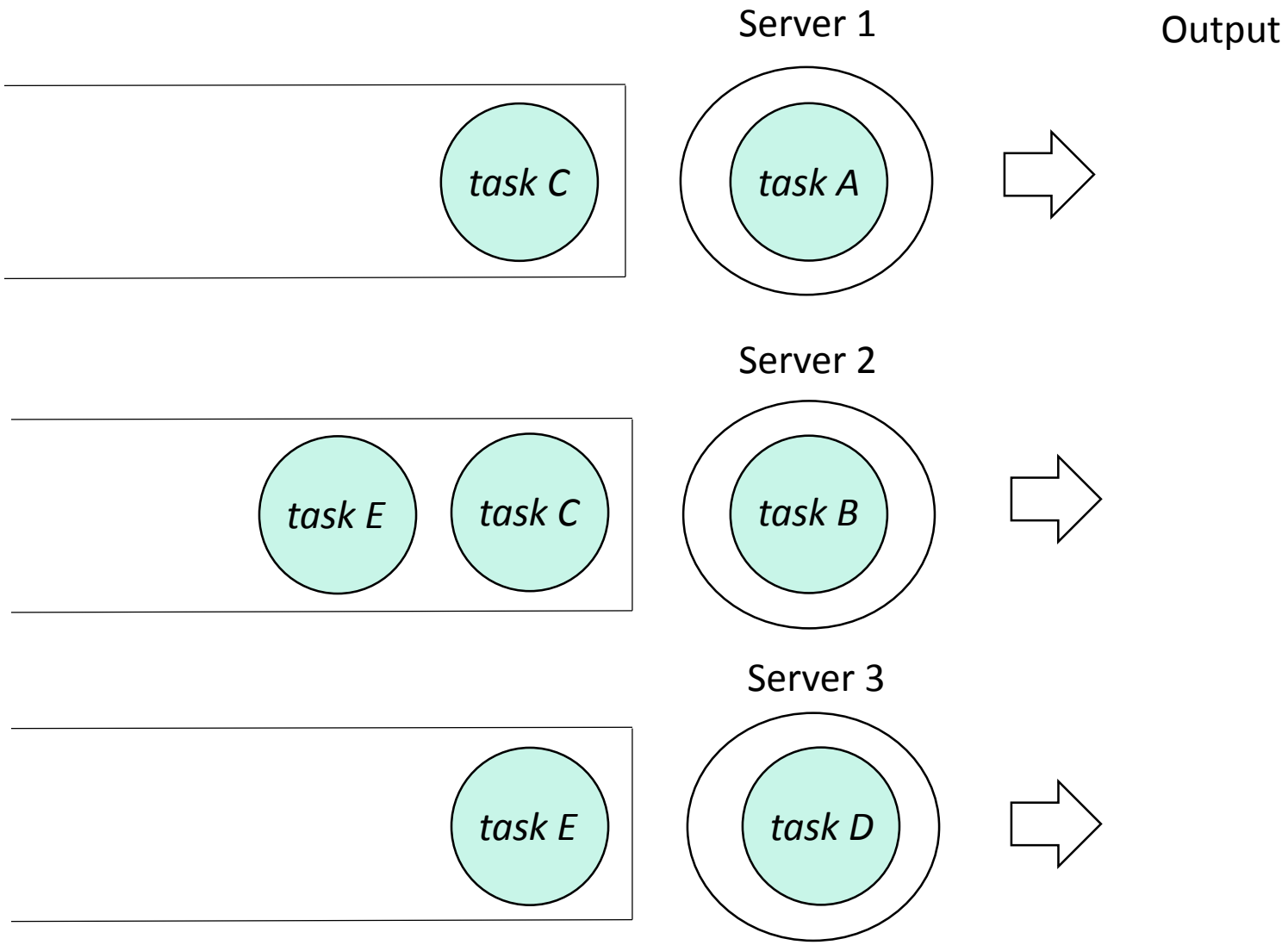




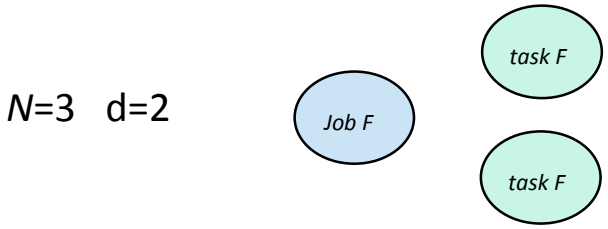
Example

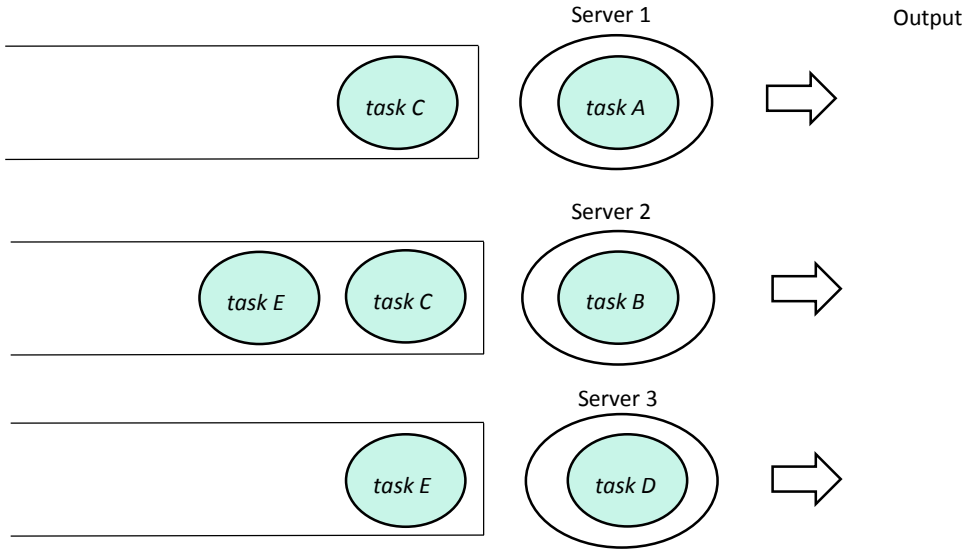
N=3 d=2





Key observation

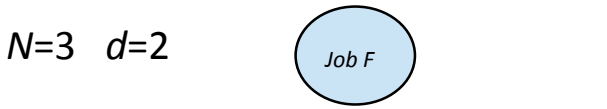


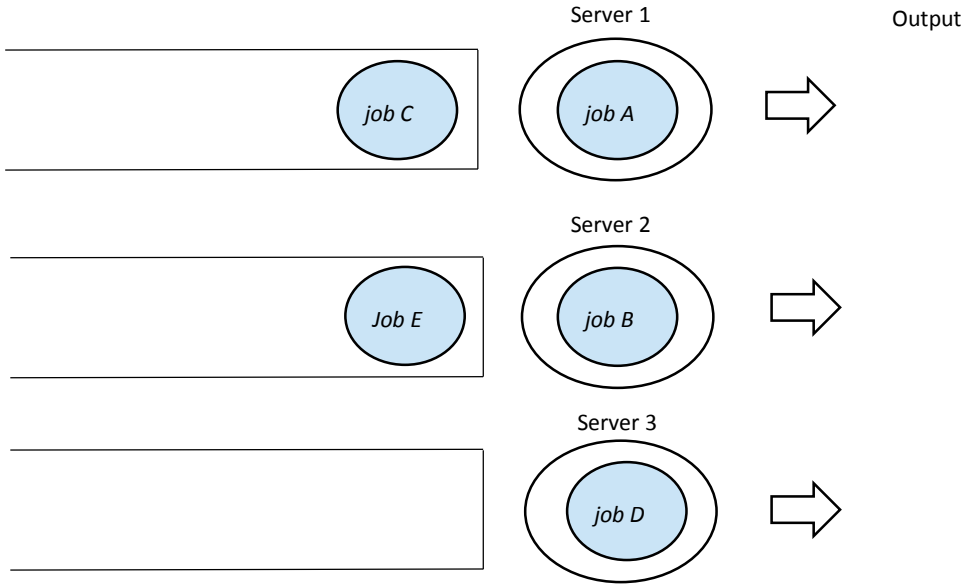


Redundancy and cancellations

Obtain results for both

Mathematically equivalent to:





'Join the least workload'

We analyze this model

JSQ

Algorithmic / simulation

Foschini, G., and J. A. C. K. Salz. **A basic dynamic routing problem and diffusion.** *IEEE Transactions on Communications* 26.3 (1978): 320-327.

G. Ananthanarayanan, A. Ghodsi, S. Shenker and I. Stoica. **Effective straggler mitigation: Attack of the clones.** In Presented as part of the 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 13), Lombard, IL, 2013.

J. Dean and L. A. Barroso. **The Tail at Scale**. CACM 56, 2 (2013), 7480

N. D. Vvedenskaya, R. L. Dobrushin and F. I. Karpelevich. **A queueing system with a choice of the shorter of two queues An asymptotic approach**. Probl. Inf. Transm. 32 1529, 1996.

Analytic

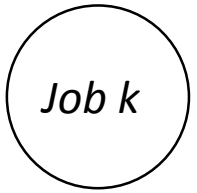
Bramson, Maury, Yi Lu, and Balaji Prabhakar. **"Randomized load balancing with general service time distributions."** *ACM SIGMETRICS Performance Evaluation Review*. Vol. 38. No. 1. ACM, 2010.

Bramson, Maury, Yi Lu, and Balaji Prabhakar. **"Decay of tails at equilibrium for FIFO join the shortest queue networks."** *The Annals of Applied Probability* 23.5 (2013): 1841-1878.

Kristen Gardner, Samuel Zbarsky, Sherwin Doroudi, Mor Harchol-Balter, Esa Hyytia, Alan Scheller-Wolf. **"Queueing with redundant requests: exact analysis."** *Queueing Systems.* 2016.

Goal: heavy traffic analysis in diffusion scale

Model - continue



$T_k^n = \left(T_{k,1}^n,\ T_{k,2}^n, \dots, T_{k,N}^n\right)$

Avg: $\left((\mu_1^n)^{-1}, (\mu_2^n)^{-1}, \dots, (\mu_N^n)^{-1}\right)$

Vectors T_k^n $i.i.d$ **across jobs.**

$\left\{T_{k,1}^n,\ T_{k,2}^n, \dots, T_{k,N}^n\right\}$ **can be dependent**.

- Sequence of systems indexed by *n*
- Input (jobs): Poisson process $\boldsymbol{A^n(t)}$, with rate $\boldsymbol{\lambda^n}$
- $\boldsymbol{T_{k,i}^n}$**:** duration of service for job *k* if it is processed by server *i*
- $\boldsymbol{\mu_i^n}$: the reciprocal mean service time for jobs processed in server *i*
- We assume: $\exists \gamma \in \mathbb{R}: \quad T_{k,i}^n < \gamma$

Heavy traffic setting

- We assume:

$$\exists\, \lambda, \hat{\lambda}: \qquad \lim_{n\to\infty} n^{-1/2}(\lambda^n - n\lambda) = \hat{\lambda}$$

$$\exists\, \mu_i, \hat{\mu}_i: \qquad \lim_{n\to\infty} n^{-1/2}(\mu_i^n - n\mu_i) = \hat{\mu}_i$$

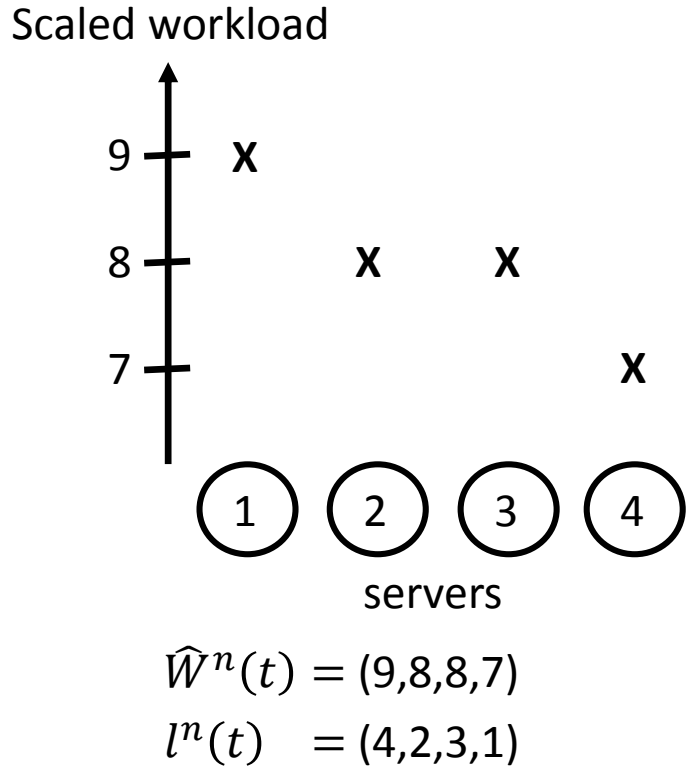
- Critical load: assume $\lambda = \sum \mu_i$

Workload

- $W^n(t) = \left(W_1^n(t), \ldots, W_N^n(t)\right)$ - workload process
- $\widehat{W}^n(t) \coloneqq n^{1/2} W^n(t)$ - scaled workload process
- $l^n(t)$ - order statistics of $\widehat{W}^n(t)$:

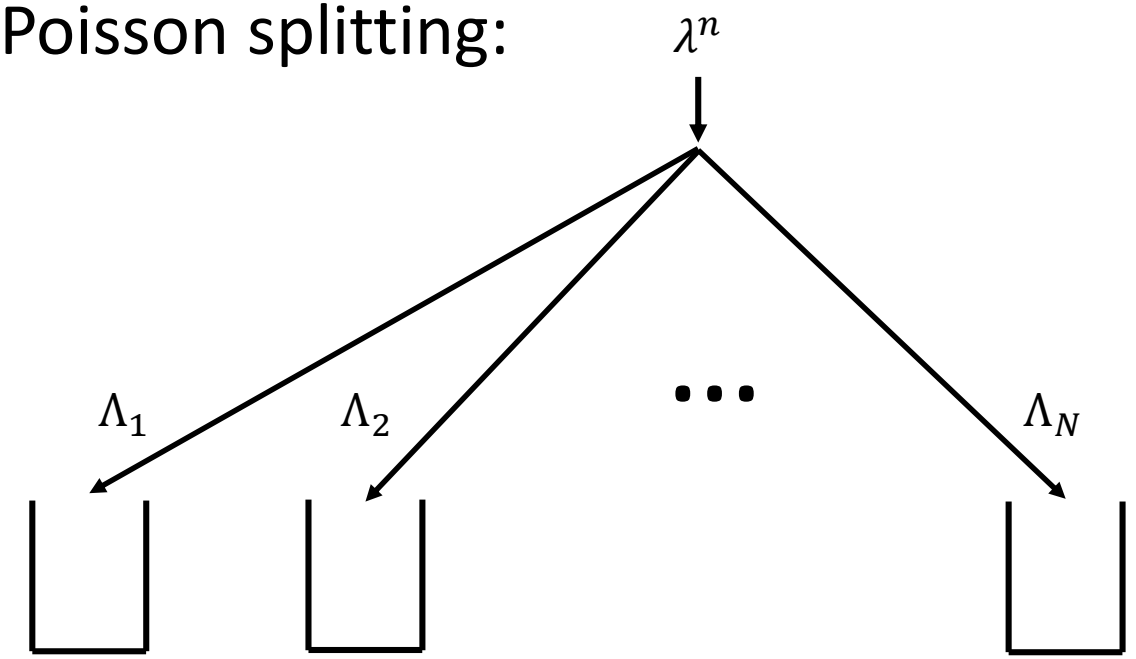
$$\widehat{W}^n_{l_1^n(t)}(t) \leq \widehat{W}^n_{l_2^n(t)}(t) \leq \cdots$$

- Ties are broken by index



Input process

- Poisson splitting:



For $d = 2$:

$$\Lambda_i = \lambda^n \frac{2}{N} \frac{N-i}{N-1} \quad , \qquad \Lambda_N = 0$$

General d:

$$\Lambda_i = \lambda^n \frac{\binom{N-i}{d-1}}{\binom{N}{d}} \quad , \qquad i \leq N - d + 1$$

$$\Lambda_N = \cdots = \Lambda_{N-d+2} = 0$$

Warning: servers change their 'identity'. Λ_i is the input rate to the server with the *i* 'th smallest workload. Not necessarily server *i*.

- $\boldsymbol{N_i^n(t)}$ – input process to server i – a compound Poisson process
- Recall that jobs are indexed by k, and each job comes with service time requirement $\left(T_{k,i}^n\right)_i$
- The workload in server i = **sum of service requirement** of jobs processed (or to be processed) in server i, *minus* its **busy time**

Denote $\boldsymbol{k_i^*(j)}$ - index of the j'th job processed (or to be processed) in server i.

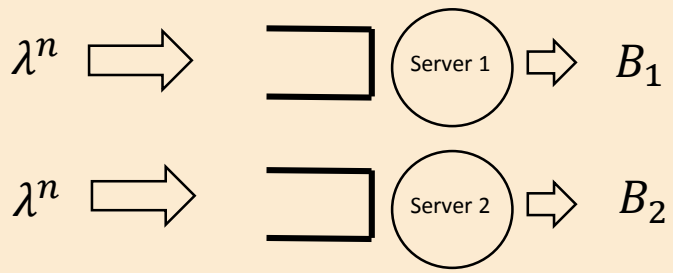
Scaled workload:

$$\widehat{W}_i^n(t) = n^{1/2} \sum_{j=1}^{N_i^n(t)} T_{k_i^*(j),i}^n - n^{1/2} \int_0^t 1_{\{\widehat{W}_i^n(s)>0\}} ds$$

Main results

Fix $T > 0$. Assume: $\frac{\mu_{max}}{\mu_{min}} < \frac{N-1}{N-d}$

Comparison: two independent M/M/1 queues:



1. State space collapse

As $n \to \infty$:

$$\max_{1 \le i,j \le N} \left\| \widehat{W}_i^n - \widehat{W}_j^n \right\|_T \to 0 \text{ , } in\ probability.$$

2. Diffusion limit

As $n \to \infty$:

$$\left(\widehat{W}_1^n, \dots, \widehat{W}_N^n \right) \Rightarrow (B, \dots, B)$$

where B is a (m, σ^2) Reflected Brownian Motion on the half line. The parameters m, σ^2 can be calculated using the model parameters. They do not depend on d!

Discussion- Redundancy routing

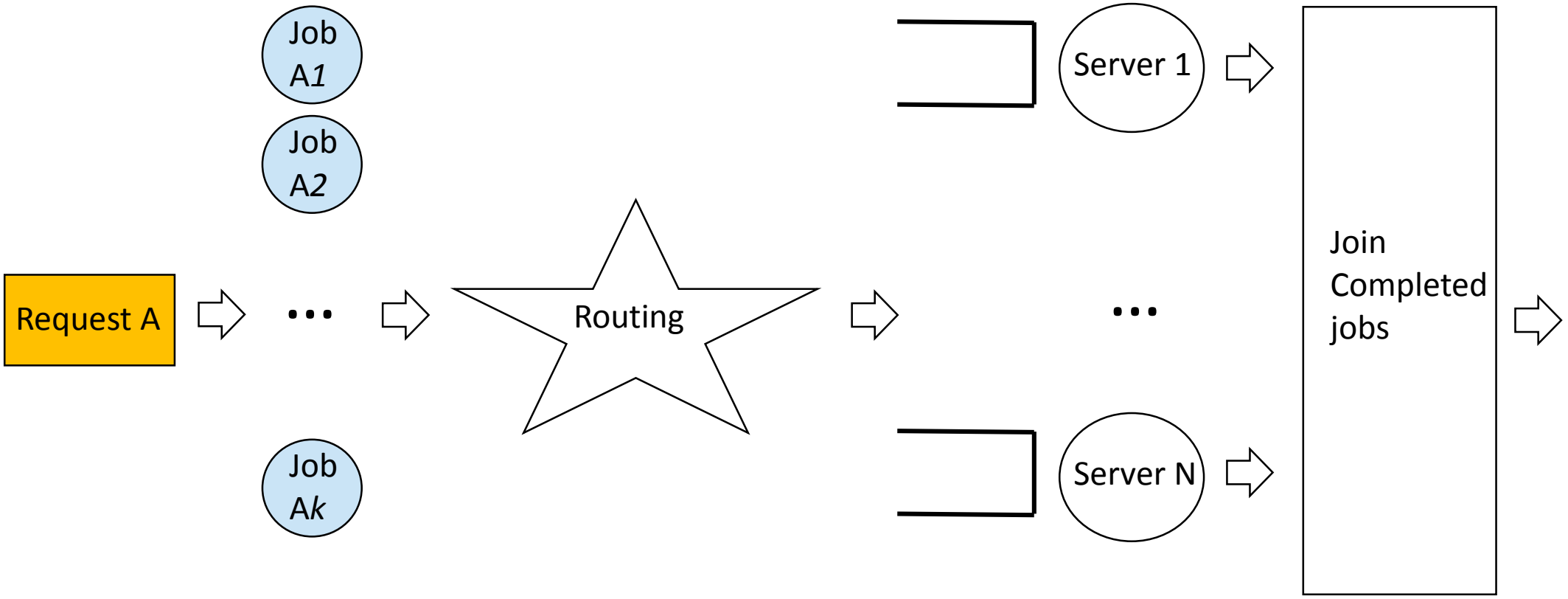
- Redundancy routing **equalizes** the delay in heavy traffic for $d \geq 2$

  (in the limit, provided the condition holds)

- **Join the least workload** can be implemented with no information on workloads, queue lengths or service rates

- **Example**: MapReduce-like applications (next slide)

Example – MapReduce-like applications



Analogy to JSQ(d) routing

We also have state space collapse results for:

- JSQ(d) routing
- A combination of JSQ(N) and redundancy routing

Thank you