Stability and Optimization of Speculative Queueing Networks

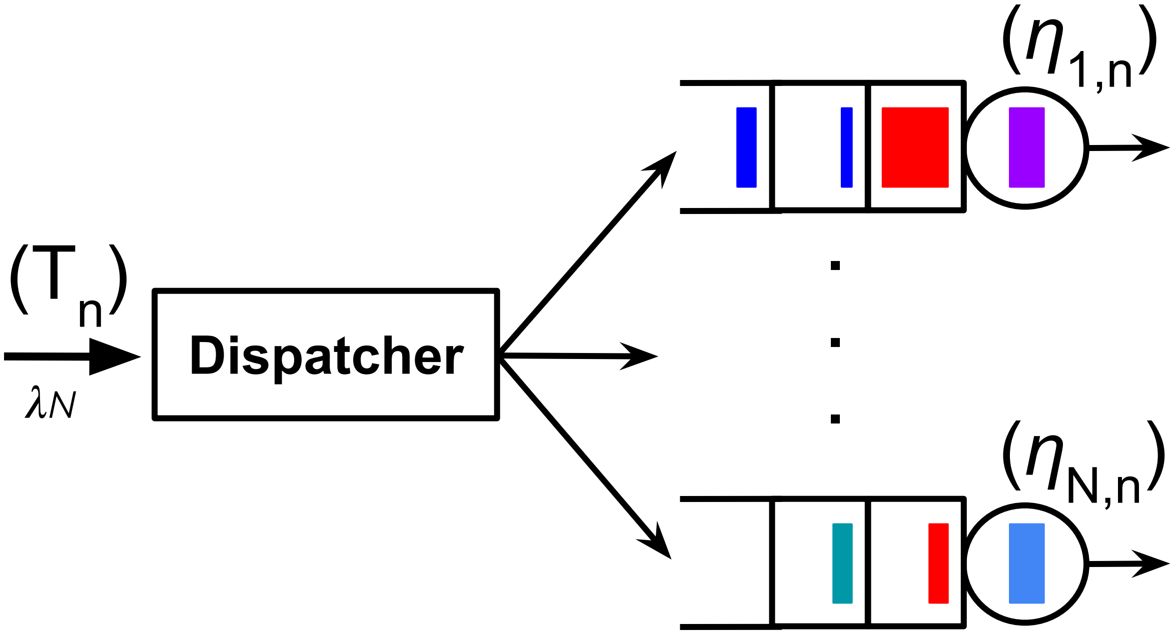
Jonatha Anselmi,

Inria

Joint work with Neil Walton, Durham University

“Standard” Load Balancing

Each incoming job is dispatched to a (unique) queue



Objective

Minimize response time

Huge Literature

Random

Round-Robin

Join-the-shortest-queue, JSQ

Power-of-*d*

Join-the-idle-queue

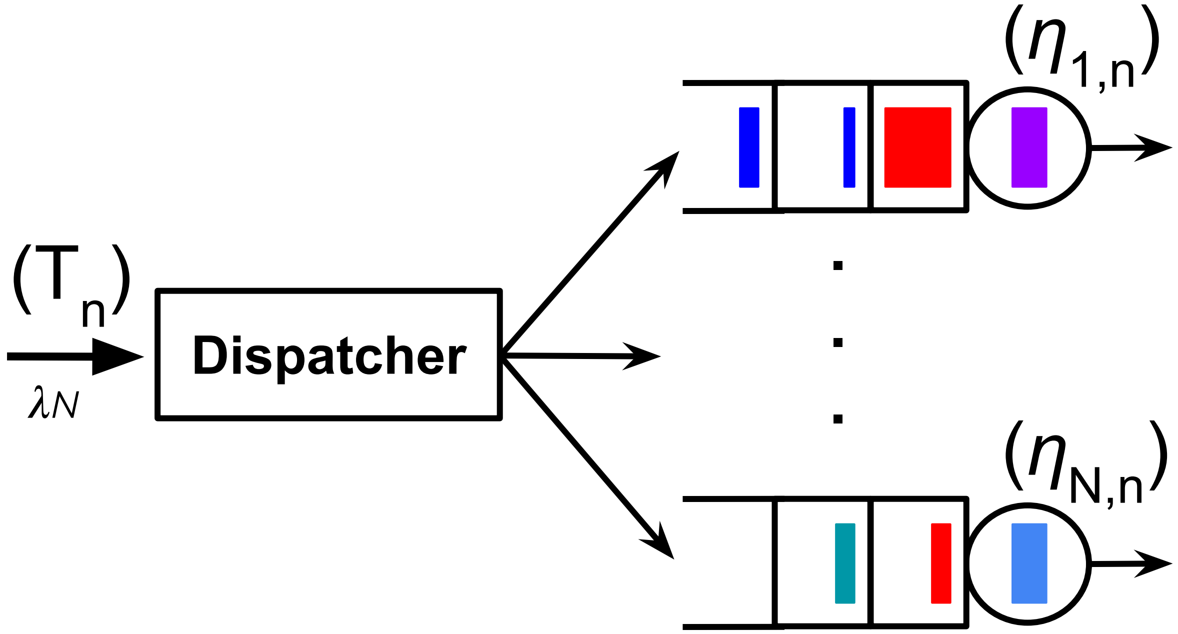
Least Left Workload, aka JSW

Size Interval Task Allocation

… and a lot more

“Standard” Load Balancing

Each incoming job is dispatched to a (unique) queue



Objective
Minimize response time

Huge Literature
Random
Round-Robin
Join-the-shortest-queue, JSQ
Power-of-*d*
Join-the-idle-queue
Least Left Workload, aka JSW
Size Interval Task Allocation

… and a lot more

Remark. All these load balancing algorithms are *stable* if and only if $\lambda\mathbb{E}[\eta] < 1$, where $\eta_{i,n} \overset{\mathrm{d}}{=} \eta$ (homogeneous case)

Recent Approach: Replicate

Motivation: to mitigate the effect of *stragglers*

Two underlying principles

Either replicate:

1) *"replicate a job upon its arrival and use the results from whichever replica responds first"*

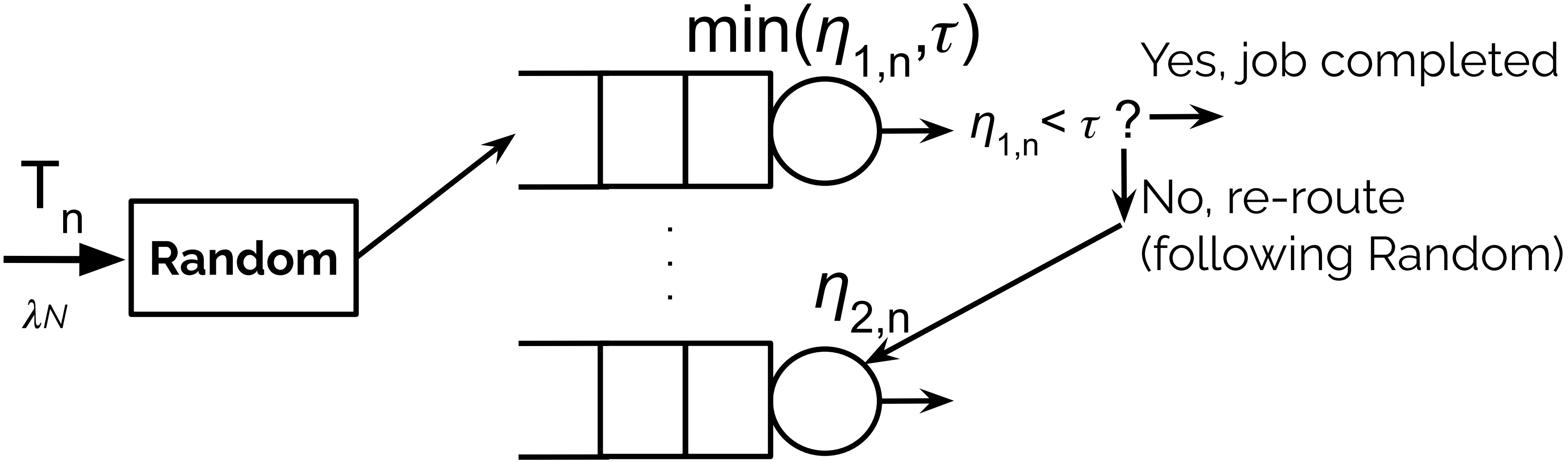
or speculate:

2) *"replicate a job as soon as the system detects it as a straggler"*

Our contribution

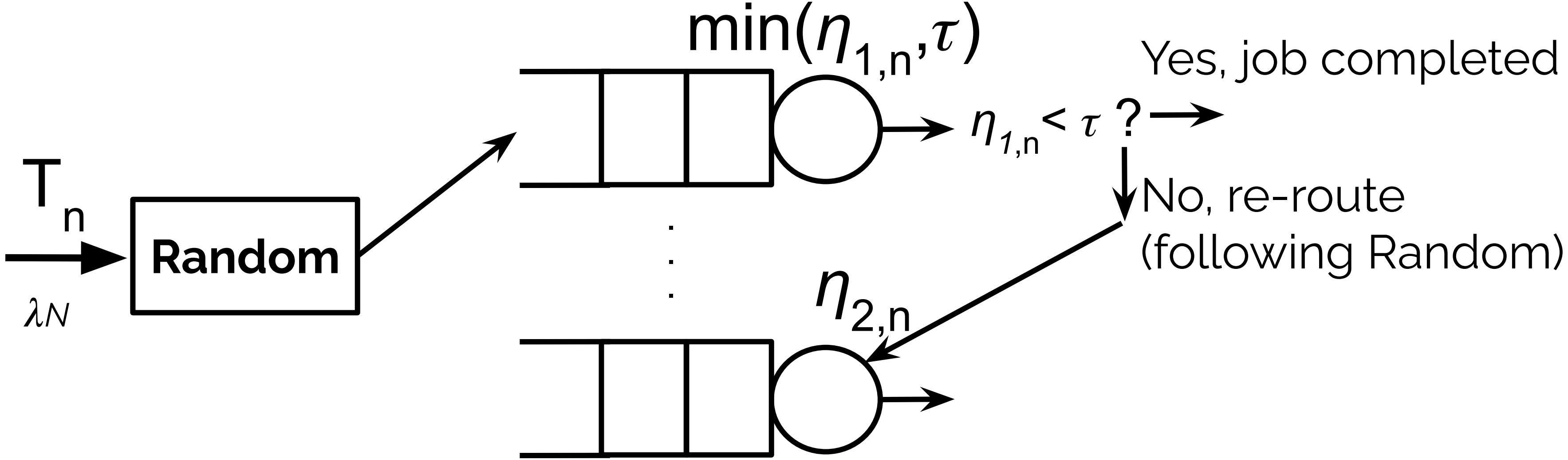
Compare Standard Load Balancing, Replication (1st principle) and Speculation (2nd principle)
⇒ Build a Markov model for speculation
⇒ Stability theorem
⇒ Optimal stopping
⇒ Comparison of the stability regions

Speculative Load Balancing



The mean processing time of any job is $\mathbb{E}[\min(\eta_1, \tau)] + \mathbb{P}(\eta_1 > \tau)\mathbb{E}[\eta_2 \mid \eta_1 > \tau]$

Speculative Load Balancing



The mean processing time of any job is $\mathbb{E}[\min(\eta_1,\tau)] + \mathbb{P}(\eta_1 > \tau)\mathbb{E}[\eta_2 \mid \eta_1 > \tau]$

Main assumptions

- Service times have the general distribution of η (heterogeneous case in [A Walton 2021])
- Fixed visit i, $\eta_{i,n}$ are IID and equal in distribution to η
- Fixed job n, $\eta_{i,n}$ have arbitrary dependency
- Head-of-the-Line scheduling disciplines (FCFS, priorities, etc.)

Stability Result

Let $\rho(\tau) := \lambda\Big(\mathbb{E}[\min(\eta_1, \tau)] + \mathbb{P}(\eta_1 > \tau)\mathbb{E}[\eta_2 \mid \eta_1 > \tau]\Big)$

Let *X(t)* denote the Markov process associated to speculative load balancing

Theorem. If $\rho(\tau)<1$, then *X* is positive Harris recurrent.

(the system is stable under the natural stability condition)

⇒ General version (general routing probabilities, heterogeneous servers) in [A Walton 2021]

Proof (outline).

- Multiclass representation
- Use the fluid framework of Dai and Bramson.
- Lyapunov function.

Stability Result

Let $\rho(\tau) := \lambda\Big(\mathbb{E}[\min(\eta_1, \tau)] + \mathbb{P}(\eta_1 > \tau)\mathbb{E}[\eta_2 \mid \eta_1 > \tau]\Big)$

Let *X(t)* denote the Markov process associated to speculative load balancing

Theorem. If $\rho(\tau)<1$, then *X* is positive Harris recurrent.

(the system is stable under the natural stability condition)

⇒ General version (general routing probabilities, heterogeneous servers) in [A Walton 2021]

Proof (outline).

- Multiclass representation
- Use the fluid framework of Dai and Bramson.
- Lyapunov function.

Remark. The stability regions of speculative load balancing, $\rho(\tau)<1$, and standard load balancing, $\lambda\mathbb{E}[\eta] < 1$, are different!

Speculative vs Standard Load Balancing

Theorem. $\rho(\tau) < \lambda\mathbb{E}[\eta]$ if and only if $\underbrace{\mathbb{E}[\eta_2 \mid \eta_1 > \tau]}_{\text{E. service time after rerouting}} < \underbrace{\mathbb{E}[\eta_1 - \tau \mid \eta_1 > \tau]}_{\text{E. remaining service time}}$

Speculative vs Standard Load Balancing

Theorem. $\rho(\tau) < \lambda\mathbb{E}[\eta]$ if and only if $\underbrace{\mathbb{E}[\eta_2 \mid \eta_1 > \tau]}_{\substack{\text{E. service time}\\ \text{after rerouting}}} < \underbrace{\mathbb{E}[\eta_1 - \tau \mid \eta_1 > \tau]}_{\substack{\text{E. remaining}\\ \text{service time}}}$

***S&X* Model.** $\eta_i = S_i X$. Slowdowns S_1 and S_2 are IID and independent of the intrinsic size X.

Theorem. Within the *S&X* model, $\rho(\tau) < \lambda\mathbb{E}[\eta]$ if there exists z such that

$$\mathbb{E}\left[Sx \wedge z\right] < \mathbb{P}\left(Sx \leq z\right) \mathbb{E}[S]\, x, \quad \forall x \in \text{support}(X)$$

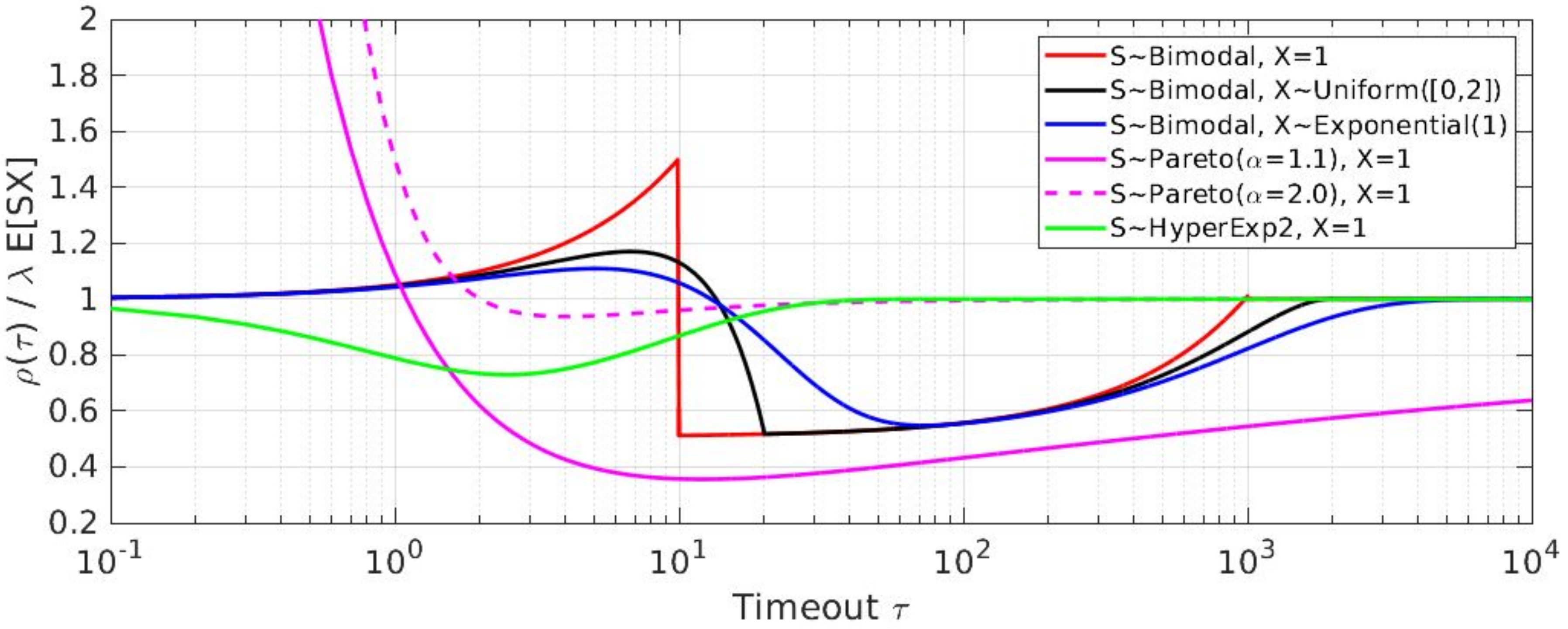
In addition, if X is deterministic, this is necessary.

⇒ Large sets of such z's exist within common service time distributions (Pareto, HyperExp, etc.)

⇒ **Increased stability region!**

Speculative vs Standard Load Balancing

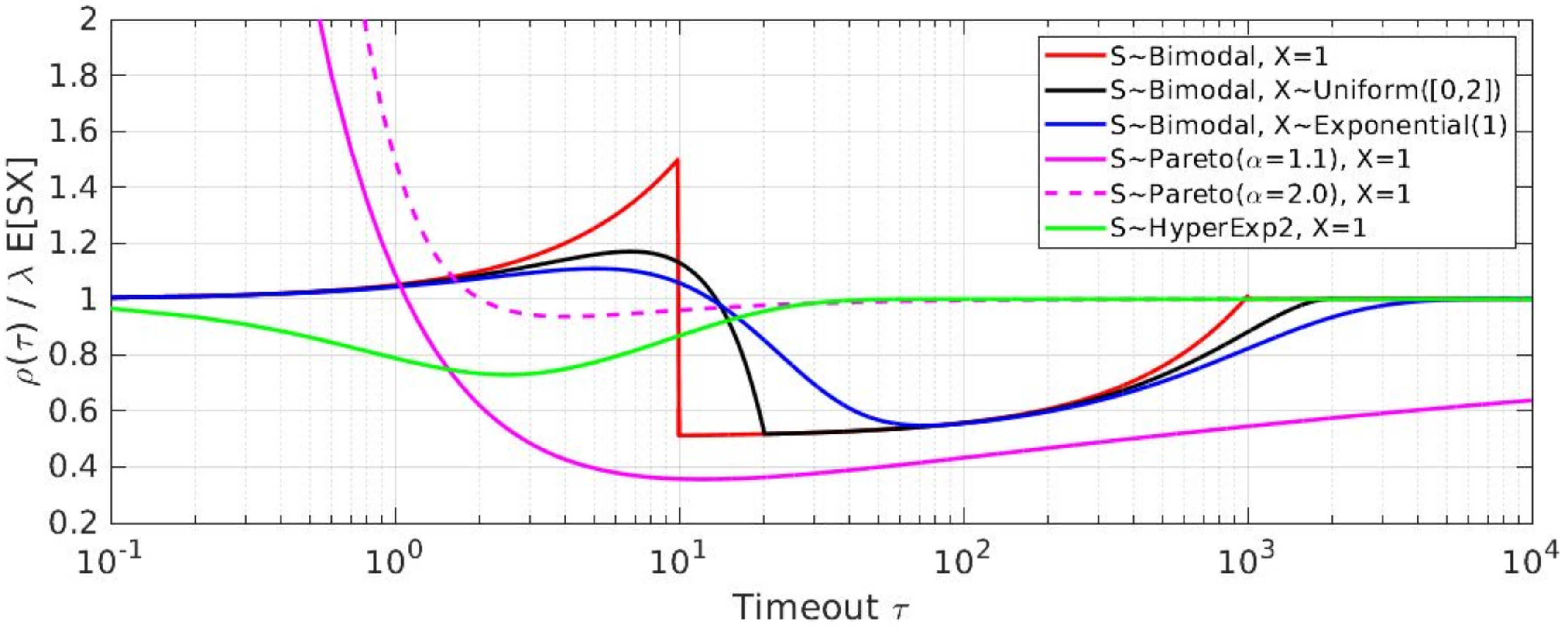
$\eta_i = S_i X$ — the S_i's are equal in distribution to S



$$\text{Bimodal:} \quad S = 10 \text{ w.p. } 0.99, S = 10^3 \text{ w.p. } 0.01.$$

Speculative vs Standard Load Balancing

Service times at server i: $\eta_i = S_i X$ — where S_i ="server slowdown" and X="job intrinsic size"



$$\text{Bimodal:} \quad S = 10 \text{ w.p. } 0.99, S = 10^3 \text{ w.p. } 0.01.$$

Optimal Timeout Design

Assumption.
The service time η has a decreasing hazard function and $t \mapsto \frac{1 + \frac{d}{dt}\mathbb{E}[\eta_2 | \eta_1 > t]}{\mathbb{E}[\eta_2 | \eta_1 > t]}$ is nondecreasing

Definition.
Let τ^* be the smallest τ such that $\frac{f(\tau)}{\int_\tau^\infty f(s)ds} \leq \frac{1 + \frac{d}{dt}\mathbb{E}[\eta_2 | \eta_1 > \tau]}{\mathbb{E}[\eta_2 | \eta_1 > \tau]}$

Optimal Timeout Design

Assumption.

The service time η has a decreasing hazard function and $t \mapsto \frac{1 + \frac{d}{dt}\mathbb{E}[\eta_2|\eta_1 > t]}{\mathbb{E}[\eta_2|\eta_1 > t]}$ is nondecreasing

Definition.

Let τ^* be the smallest τ such that $\frac{f(\tau)}{\int_\tau^\infty f(s)ds} \leq \frac{1 + \frac{d}{dt}\mathbb{E}[\eta_2|\eta_1 > \tau]}{\mathbb{E}[\eta_2|\eta_1 > \tau]}$

Theorem.

There exists a finite τ^* and minimizes the load $\rho(\tau)$.

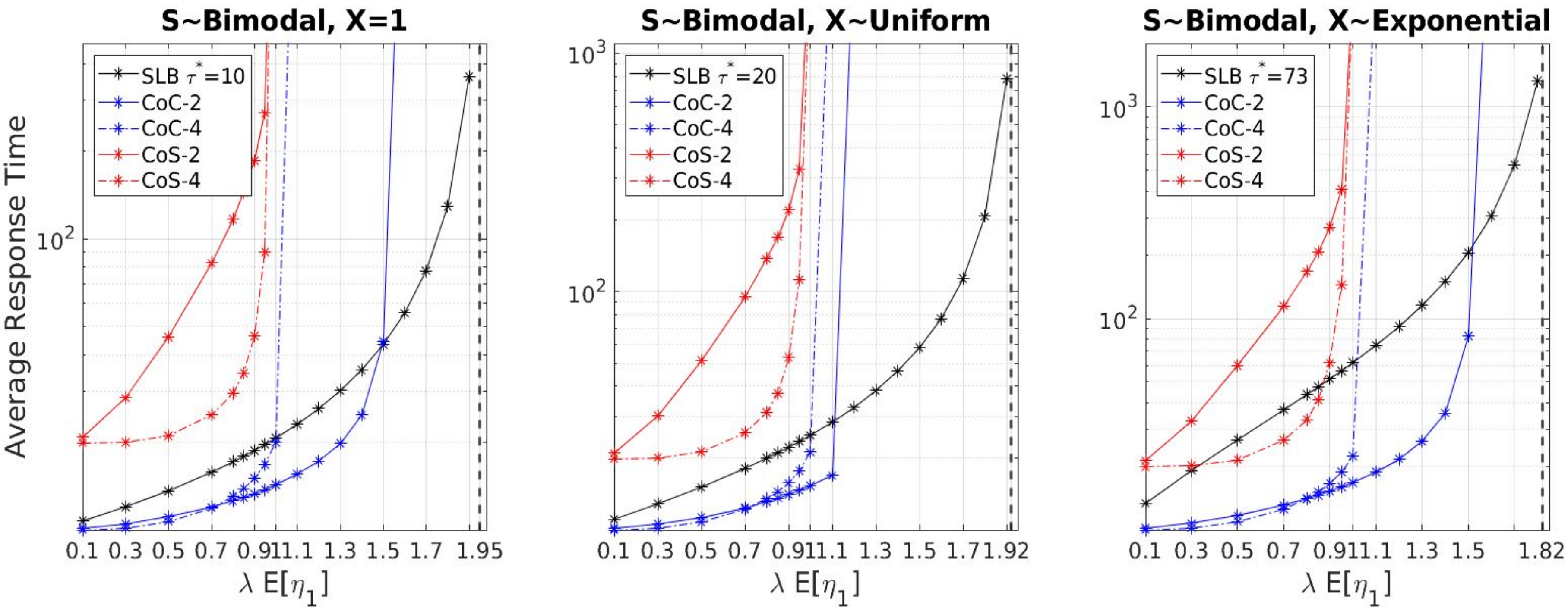
If η_1 *and* η_2 are independent, any value of τ satisfying $\frac{f(\tau)}{\int_\tau^\infty f(s)ds} = \frac{1}{\mathbb{E}[\eta_2]}$ minimizes the load.

Proof.

- Optimal stopping problem
- Markov decision process formulation
- Application of the one-step-lookahead principle.

Speculation vs Replication

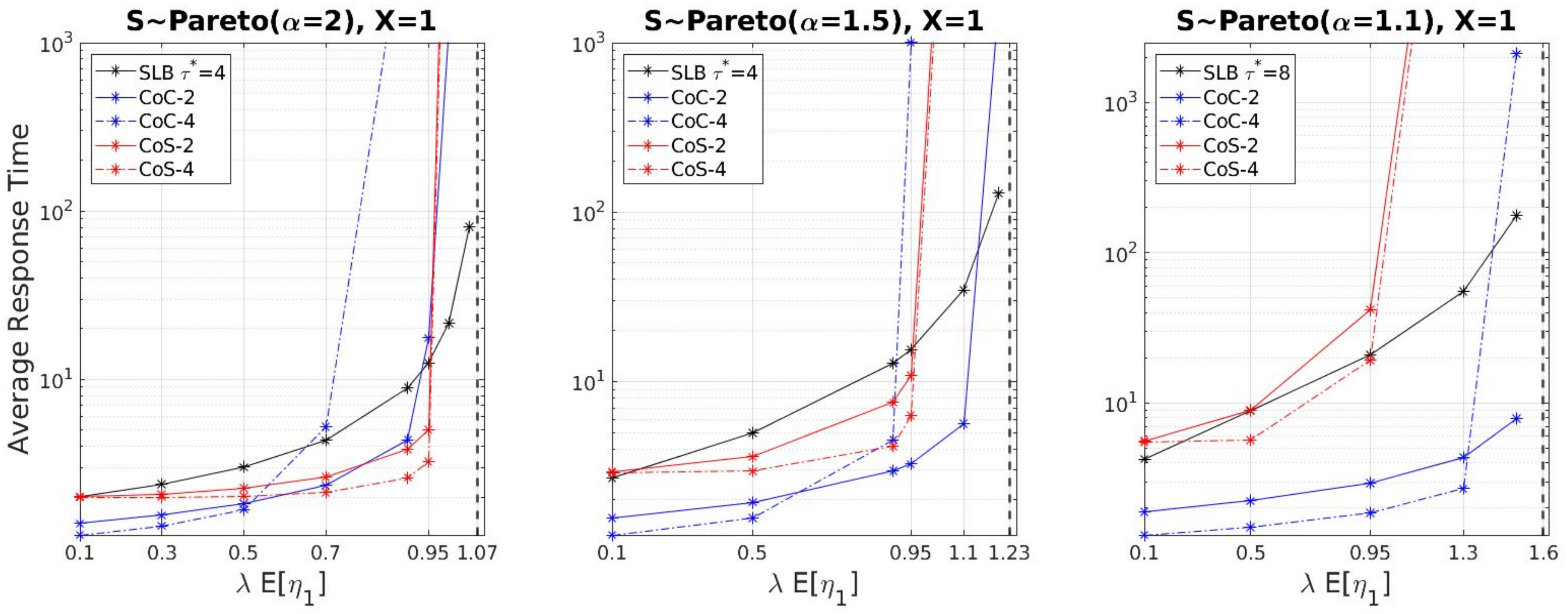
Replication strategies: Cancel-on-Complete-*d* (CoC-*d*) and Cancel-on-Start-*d* (CoS-*d*)



⇒ **Speculative Load Balancing (SLB) provides a larger stability region!**

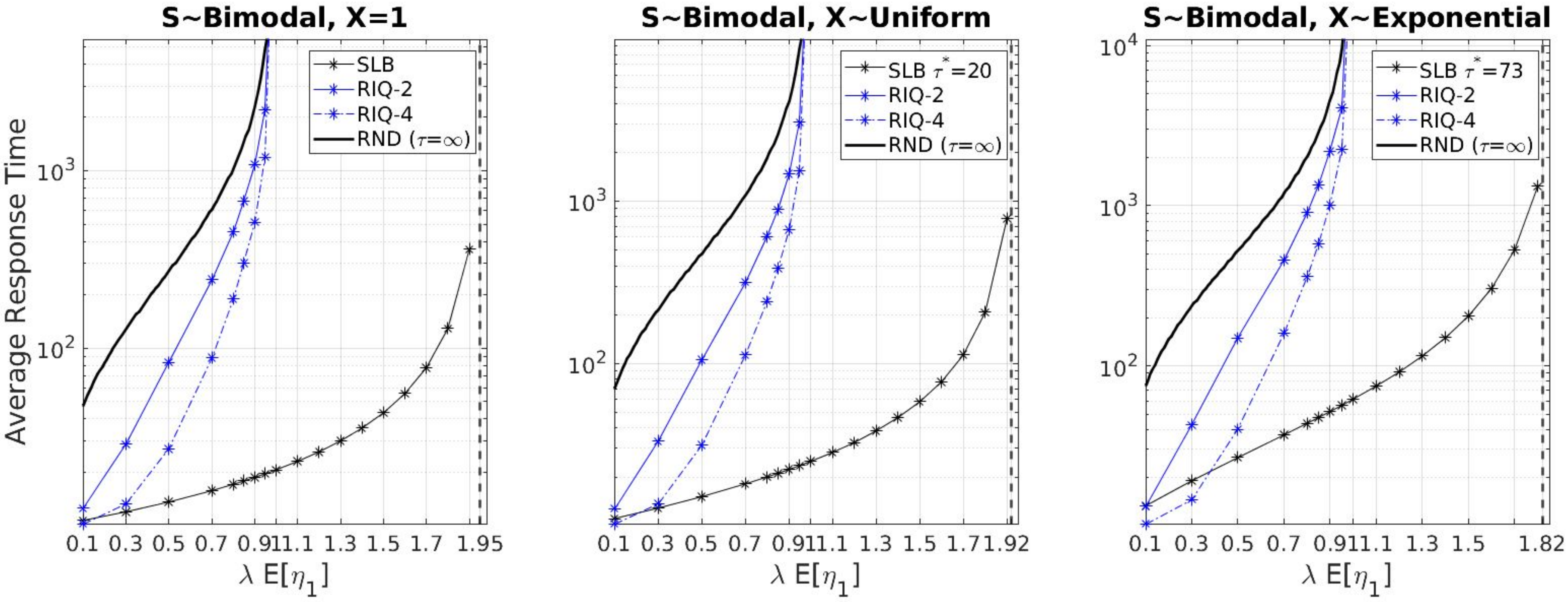
Speculation vs Replication

Replication strategies: Cancel-on-Complete-*d* (CoC-*d*) and Cancel-on-Start-*d* (CoS-*d*)



Speculation vs Replication

Replication strategy: Redundant-to-Idle-Queue-*d* (RIQ-*d*)



Response Time for Large Systems

N FCFS queues, arrival rate λN

Let $R_N(\tau)$ be the long-run average response time.

Conjecture.

Provided that $\rho(\tau)<1$, $\lim_{N\to\infty} R_N(\tau) = (1+\mathbb{P}(\eta_1 > \tau))W + \frac{\rho(\tau)}{\lambda}$

where

$$W := \frac{\lambda}{2}(1+\mathbb{P}(\eta_1 \geq \tau))\frac{M}{1-\rho(\tau)}$$

$$M := \frac{\mathbb{E}[(\eta_1 \wedge \tau)^2] + \mathbb{E}[\hat{\eta}_2^2]\,\mathbb{P}(\eta_1 > \tau)}{1+\mathbb{P}(\eta_1 > \tau)}$$

Response Time for Large Systems

N FCFS queues, arrival rate λN

Let $R_N(\tau)$ be the long-run average response time.

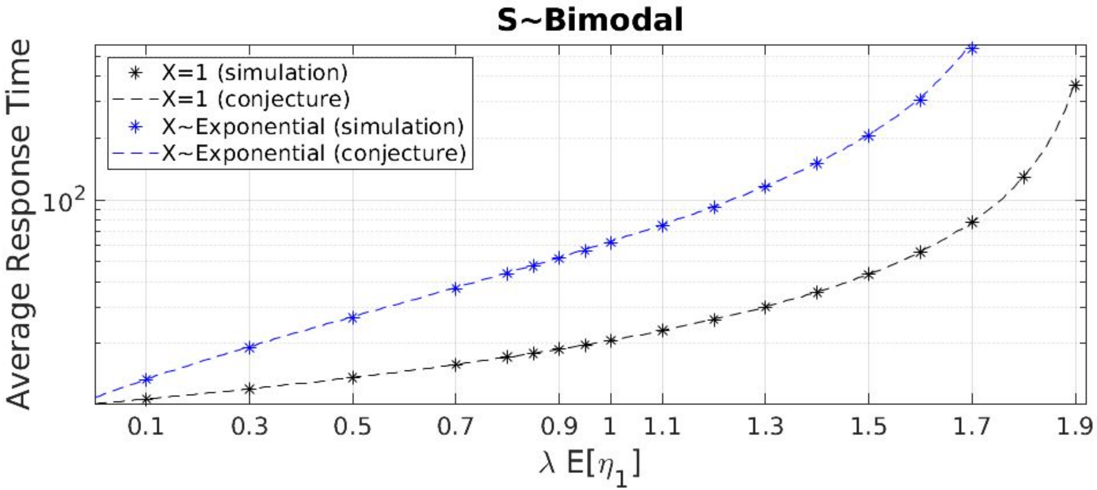
Conjecture.

Provided that $\rho(\tau)<1$, $\lim_{N\to\infty} R_N(\tau) = (1+\mathbb{P}(\eta_1>\tau))W + \frac{\rho(\tau)}{\lambda}$

where

$$W := \frac{\lambda}{2}(1+\mathbb{P}(\eta_1 \geq \tau))\frac{M}{1-\rho(\tau)}$$

$$M := \frac{\mathbb{E}[(\eta_1 \wedge \tau)^2] + \mathbb{E}[\hat{\eta}_2^2]\,\mathbb{P}(\eta_1>\tau)}{1+\mathbb{P}(\eta_1>\tau)}$$



Conclusions

In this talk

- Comparison between Speculative Load Balancing (SLB), standard load balancing and replication schemes.

Take away messages

- SLB is convenient when service times are decreasing failure rate
- Cancel-on-Complete-*d* provides better response times in light/moderate load conditions
- SLB provides better response times in heavy load conditions (larger stability region)

Future research

- Combine SLB and replication
- Multiple levels of speculation
- Re-route with other load balancing algorithms (e.g., to idle queues)

Conclusions

Take away messages

- SLB is convenient when service times are decreasing failure rate
- Cancel-on-Complete-*d* provides better response times in light/moderate load conditions
- SLB provides better response times in heavy load conditions (larger stability region)

Thank you

J. Anselmi and N. Walton, "Stability and Optimization of Speculative Queueing Networks," in IEEE/ACM Transactions on Networking, vol. 30, no. 2, pp. 911-922, April 2022, doi: 10.1109/TNET.2021.3128778.