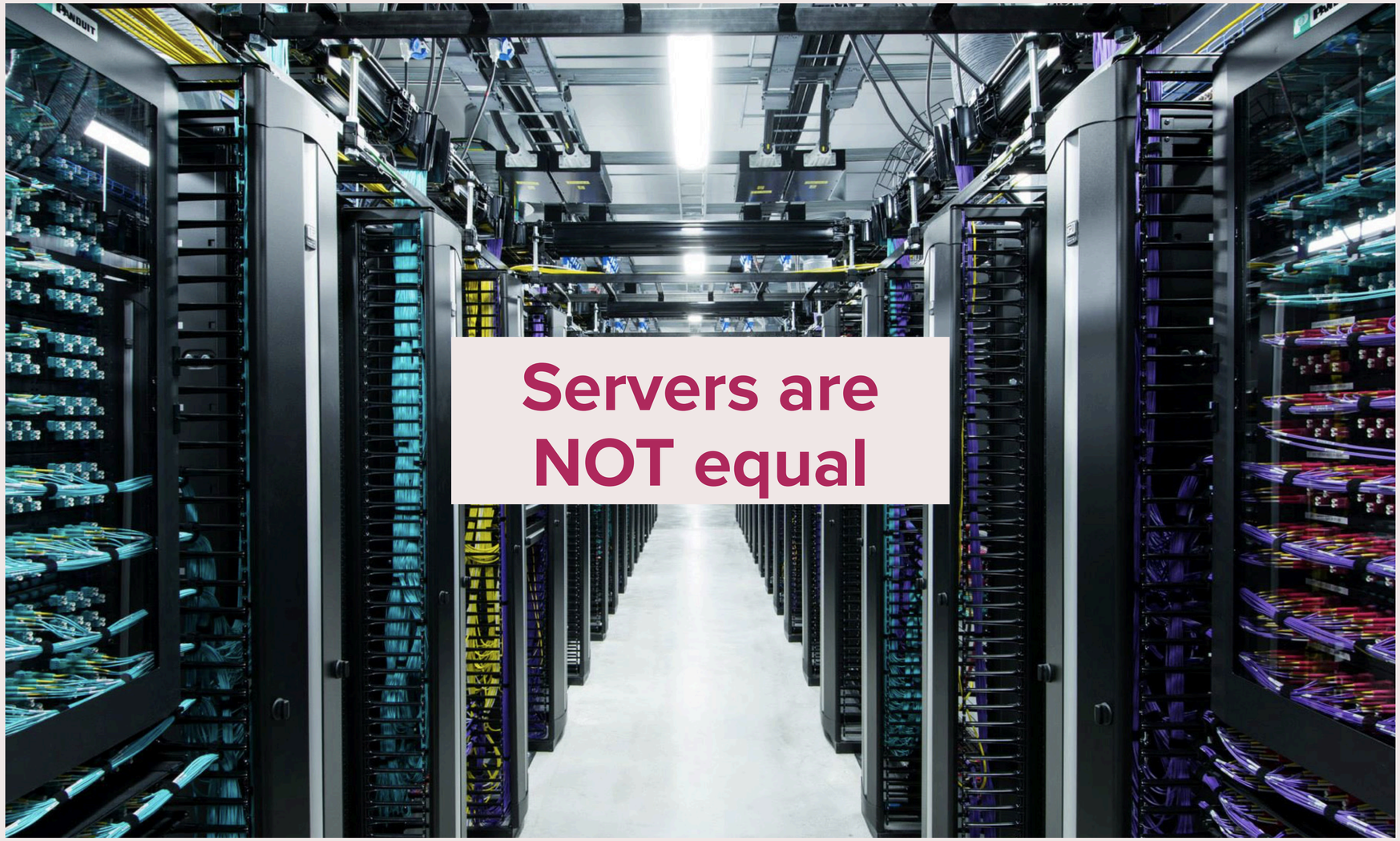
Queue Length Behavior in Load Balancing Systems Under Power-of-d Choices: Many-Server Heavy-Traffic Regime

Daniela Hurtado-Lange

INFORMS 2021

Motivation

Load Balancing in Data Centers



Outline

Model
- Load balancing system in discrete time
- Power-of-d choices

Throughput Optimality Under Heterogeneous Servers
- Necessary and sufficient conditions
- Interpretation of the conditions
- Heavy-traffic behavior

Many-Server Heavy-Traffic Regime
- Definition of the regime
- Queue length behavior

Main Takeaways

Load Balancing System

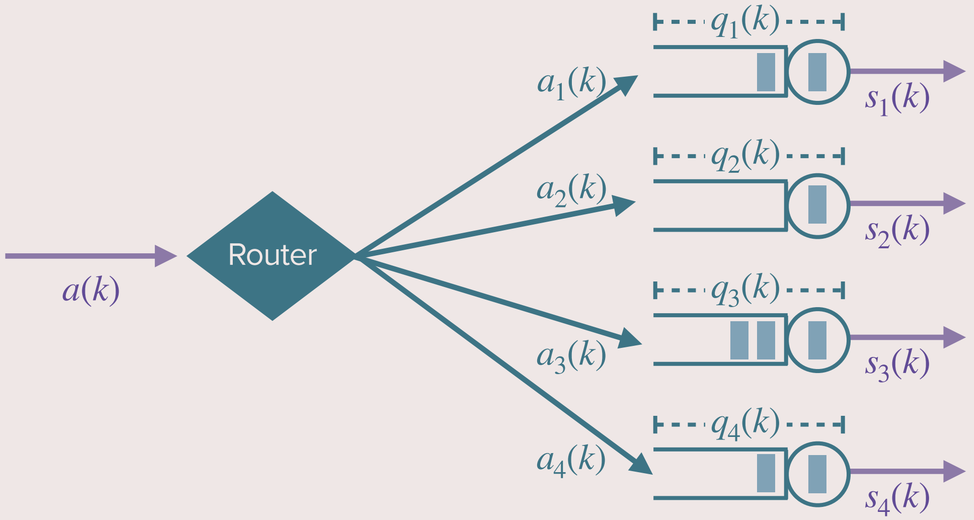
- Discrete time model
- N servers with an infinite buffer
- $q_i(k)$ = # jobs in queue i in time slot k

Arrivals:

- $a(k)$ = arrivals in time slot k
- $\{a(k) : k \in \mathbb{Z}_+\}$ is a sequence of i.i.d. random variables
- Upon arrival, jobs are routed to the queues

Service:

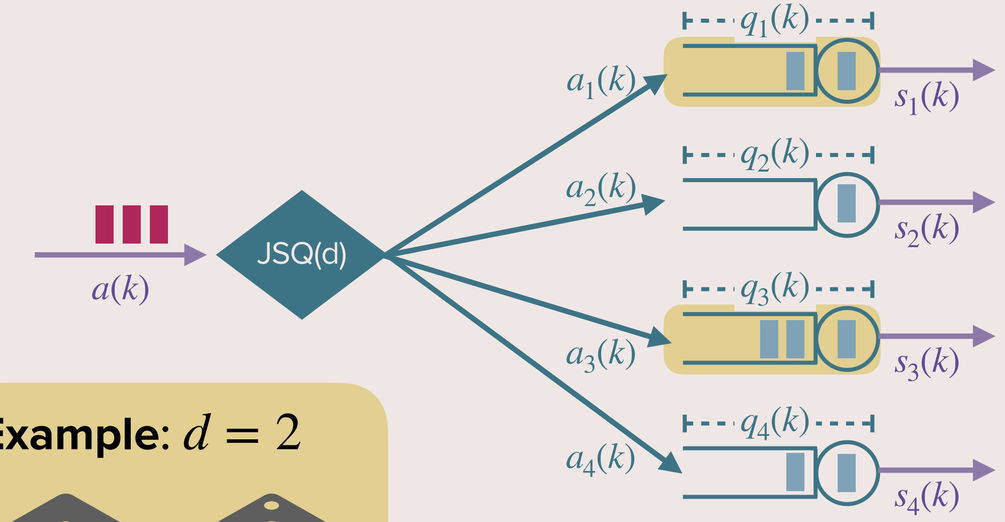
- $s_i(k)$ = potential service in queue i in time slot k
- $\{s_i(k) : k \in \mathbb{Z}_+\}$ is a sequence of i.i.d. random variables
- Service process to different queues are independent of each other



Power-of-d Choices

Algorithm: Given an integer $d \in [1, N]$, in each time slot:

1. Select d servers uniformly at random
2. Route arrivals to the shortest queue among those d



Example: $d = 2$

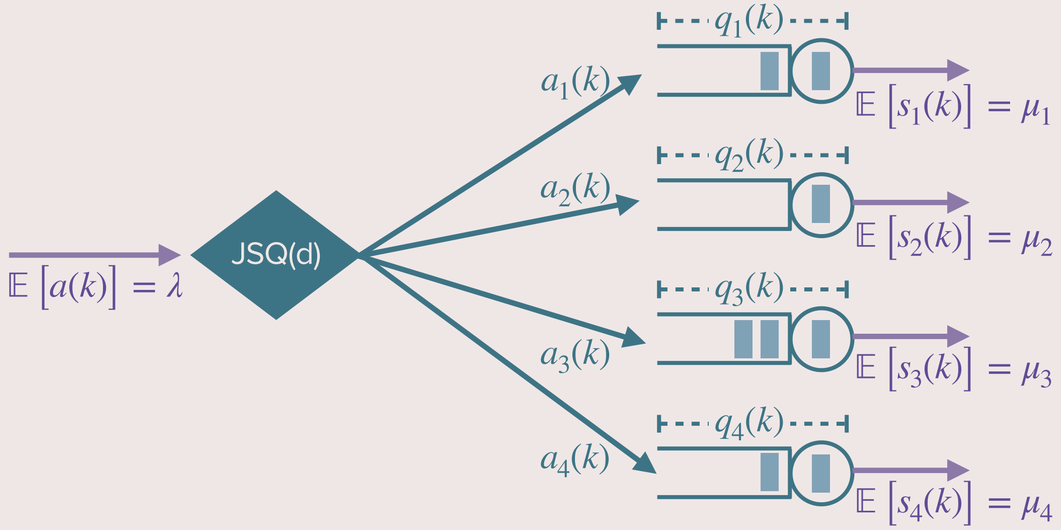
Observations:

- Also known as JSQ(d)
- If $d = N$, it is exactly JSQ
- If $d = 1$, it is random routing

Throughput Optimality

Capacity Region of the Load Balancing System:

$$\mathscr{C} = \left\{ \lambda \in \mathbb{R}_+ : \lambda \leq \sum_i \mu_i \right\}$$



Definition: Throughput optimal

A routing algorithm $\mathscr{A}$ is throughput optimal, if the load balancing system operating under $\mathscr{A}$ is stable for all $\lambda \in Int(\mathscr{C})$

Example:

- JSQ is throughput optimal
- Power-of-d is throughput optimal if all servers are equal

What happens if the servers are different?

Power-of-d When Servers are Different

Theorem: [HL, Maguluri 2020]

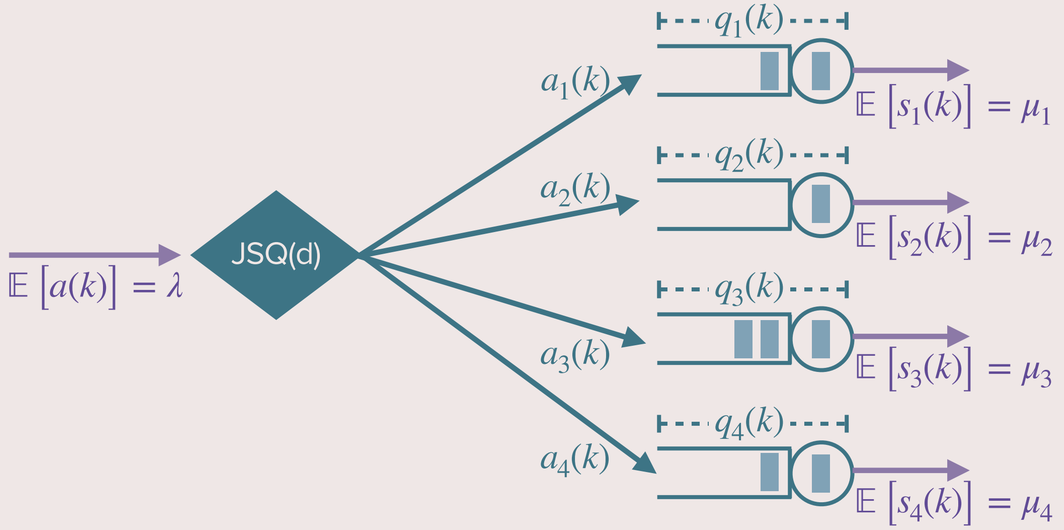
Let $1 \le d \le N-1$. Define

$$x_i = \frac{\mu_i}{\mu_\Sigma}, \text{ and } y_i = \frac{\binom{i-1}{d-1}}{\binom{N}{d}}.$$

Power-of-d is throughput optimal if and only if $\boldsymbol{x} \preccurlyeq \boldsymbol{y}$

Majorization: Let $x_{(i)}$ be the i^{th} smallest component of $\boldsymbol{x} \in \mathbb{R}^N$. Then, the notation $\boldsymbol{x} \preccurlyeq \boldsymbol{y}$ means that

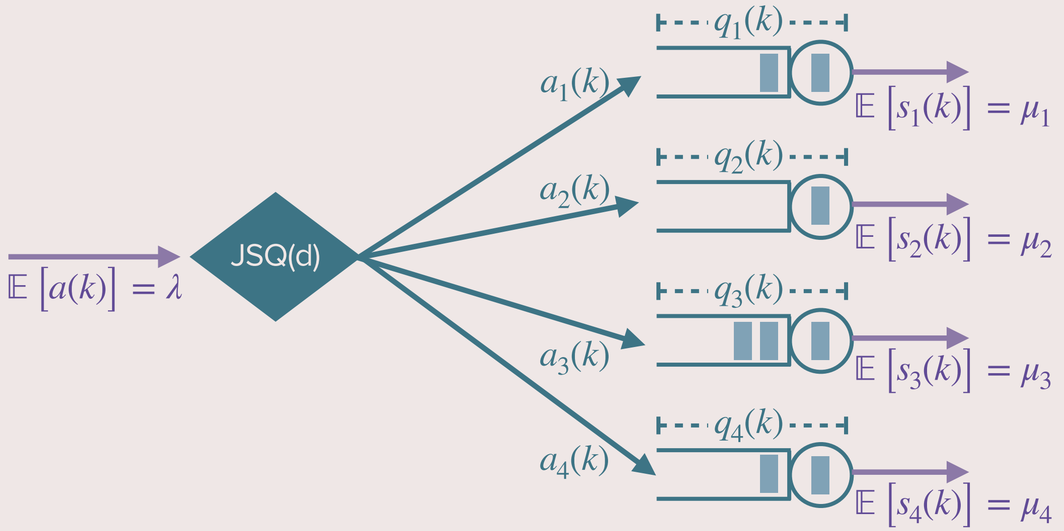
$$\sum_{i=j}^{N} x_{(i)} \le \sum_{i=j}^{N} y_{(i)} \quad \forall j \in [N]$$



Then, the necessary and sufficient condition is that

$$\frac{\sum_{i=1}^{j} \mu_{(i)}}{\mu_\Sigma} \ge \frac{\binom{j}{d}}{\binom{N}{d}} \qquad \forall d \le j \le N-1$$

Power-of-d When Servers are Different



Interpretation:

- Faster servers should be sampled sufficiently often
- Power-of-d samples **uniformly** at random, so we characterize the **amount of imbalance** that power-of-d can tolerate
- Fix N. As d gets larger, we can tolerate more imbalance.
- If $d = 1$ (random routing), then $RHS = \frac{j}{N}$
  In this case, $\mu_i = \mu_j \; \forall i \neq j$ is the only way to satisfy the inequalities
- If $d = N$ (JSQ), then $RHS = 0$
  In this case, all vectors $\boldsymbol{\mu}$ satisfy the inequalities

Theorem: Power-of-d is throughput optimal if and only if

$$\frac{\sum_{i=1}^{j} \mu_{(i)}}{\mu_\Sigma} \geq \frac{\binom{j}{d}}{\binom{N}{d}} \qquad \forall d \leq j \leq N-1$$

Heavy-Traffic Behavior

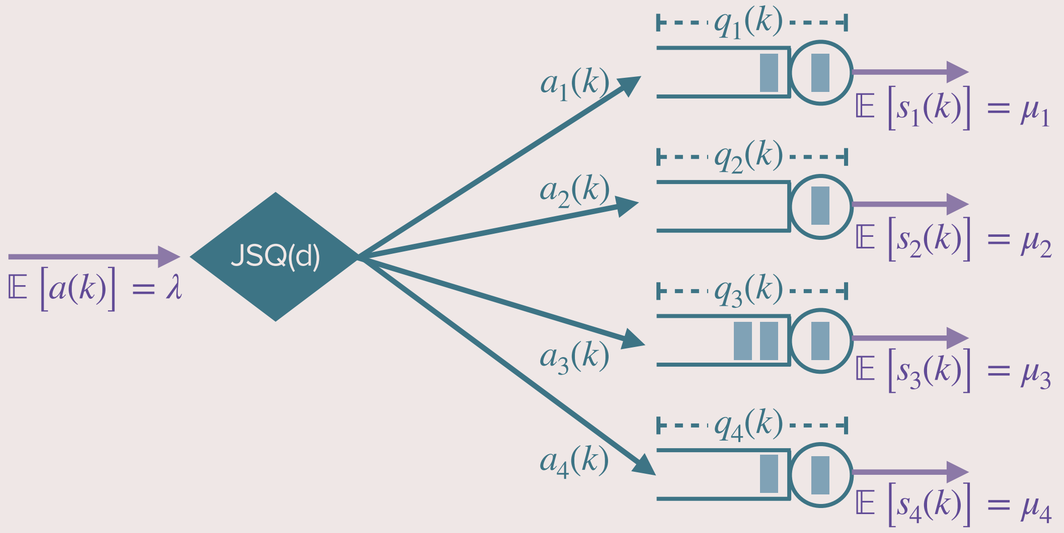
Theorem: [HL, Maguluri 2020]

If all the inequalities are satisfied **strictly**, then the system experiences SSC and, defining

$$\epsilon \triangleq \mu_\Sigma - \lambda, \epsilon \boldsymbol{q} \Rightarrow \mathbf{1}\, Expo\left(\frac{2}{\sigma_a^2 + \sum_i \sigma_{s_i}^2}\right) \text{ as } \epsilon \downarrow 0$$

Convergence in distribution in **classical** heavy traffic

What happens in **many-server heavy-traffic**?



Theorem: Power-of-d is throughput optimal if and only if

$$\frac{\sum_{i=1}^{j} \mu_{(i)}}{\mu_\Sigma} \geq \frac{\binom{j}{d}}{\binom{N}{d}} \qquad \forall d \leq j \leq N-1$$

Many-Server Heavy-Traffic Regime

Service process:

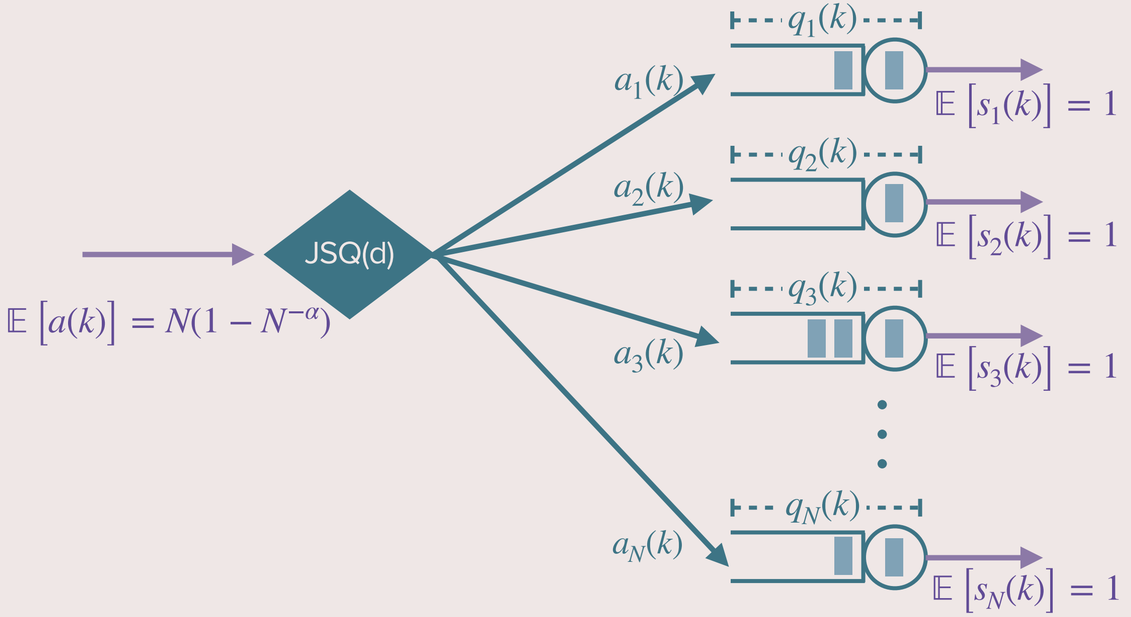
- All servers are equal and independent of each other
- $\mathbb{E}\left[s_i(1)\right] = 1$ and Var $\left[s_i(1)\right] = \sigma_s^2$

Arrival process:

- Consider $\alpha > 0$
- $\mathbb{E}\left[a(1)\right] = N(1 - N^{-\alpha})$ and Var $\left[a(1)\right] = N\sigma_a^2$
- *Arrivals "per server" have mean* $1 - N^{-\alpha}$ *and variance* σ_a^2

Literature:

- Plenty of literature for $\alpha \leq 1$
  - $\alpha = 1/2$: Halfin-Whitt regime
  - $\alpha = 1$: Nondegenerate slowdown (NDS)



We focus on $\alpha > 1$, i.e., when the load grows fast

Many-Server Heavy-Traffic Regime

Routing algorithm

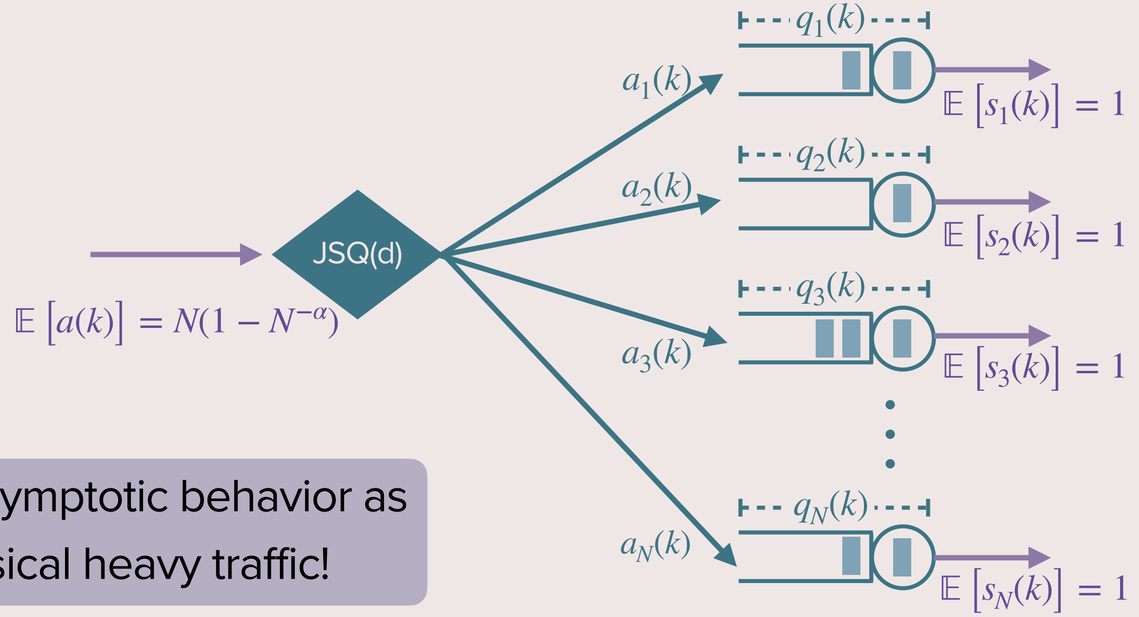
- Power-of-d choices
- $d = cN^{\beta}$, where $c, \beta \geq 0$ are constants

Theorem: [HL, Maguluri 2020]

In steady-state, and if $\alpha + \beta > 11/2$, we have

$$N^{-\alpha} \sum_{i=1}^{N} q_i \Rightarrow Expo\left(\frac{2}{\sigma_a^2 + \sigma_s^2}\right) \text{ as } N \to \infty$$

Same asymptotic behavior as classical heavy traffic!



Corollary: [HL, Maguluri 2020]

Load balancing in continuous time, operating under **JSQ**. Then, if $\alpha > 9/2$ we have

$$N^{-\alpha} \sum_{i=1}^{N} q_i \Rightarrow Expo\left(\frac{2}{\sigma_a^2 + \sigma_s^2}\right) \text{ as } N \to \infty$$

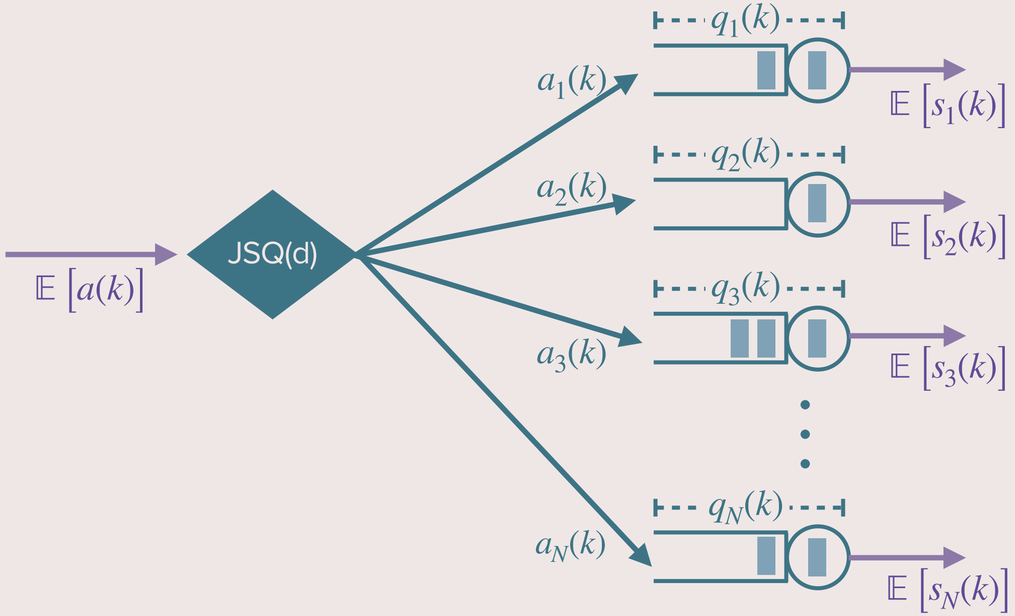
Proof sketch:

1. Multiplicative state space collapse
2. Convergence in distribution
   - Option 1: MGF method
   - Option 2: Stein's method

Main Takeaways

Power-of-d Choices with Heterogeneous Servers:

- How much imbalance can JSQ(d) tolerate?
- **Result:**
  - Sample fast servers frequently
  - Service rate vector majorized by probability of sampling slow servers
  - Similar conditions to obtain heavy-traffic optimality



Many-Server Heavy-Traffic Regime:

- Conditions on α to observe classical heavy-traffic behavior?
- **Result:** If $d = cN^{\beta}$ and $\alpha + \beta > 11/2$, then we observe heavy-traffic behavior
  - JSQ: $\beta = 1 \implies \alpha > 9/2$

Thanks!

Queue Length Behavior in Load Balancing Systems Under Power-of-d Choices: Many-Server Heavy-Traffic Regime

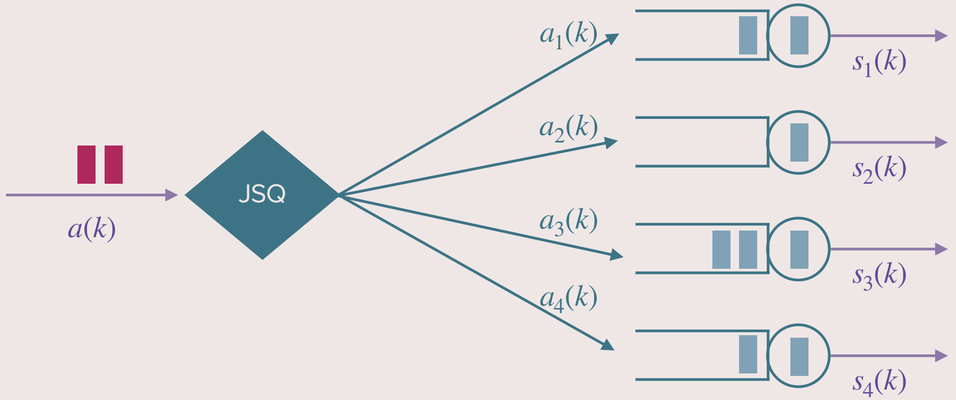
Daniela Hurtado-Lange

INFORMS 2021

Load Balancing System — JSQ Routing

JSQ:

- Join the shortest queue
- **All** arrivals are routed to the shortest queue



Classical Heavy-Traffic Regime:

- Let $\mu_i = \mathbb{E}\left[s_i(1)\right]$, and $\mu_\Sigma = \sum \mu_i$
- Arrival rate: $\mathbb{E}\left[a(1)\right] = \mu_\Sigma - \epsilon$, where $\epsilon \in (0, \mu_\Sigma)$
- Take the limit as $\epsilon \downarrow 0$

Load Balancing System — JSQ Routing (cont.)

Theorem: [Eryilmaz and Srikant 2012]

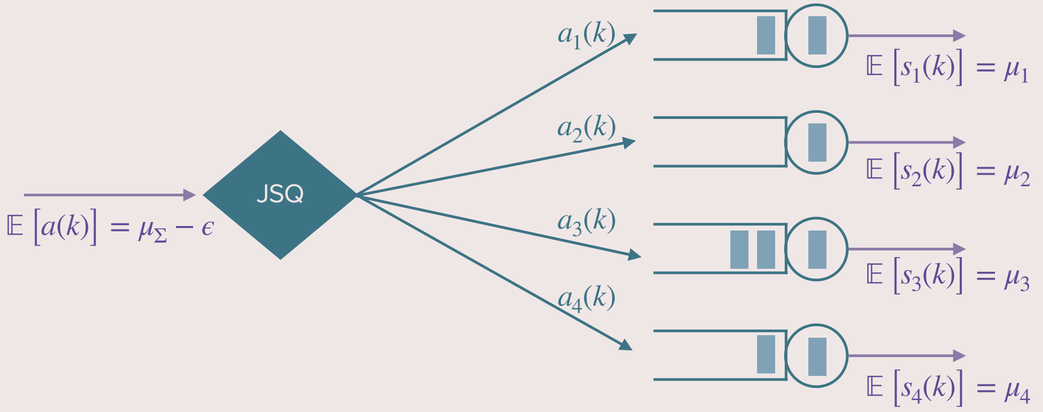
In steady-state,

$$\left| \mathbb{E}\left[\sum_{i=1}^{N} q_i\right] - \frac{\sigma_a^2 + \sum \sigma_{si}^2}{2\epsilon} \right| \text{ is } o\left(\frac{1}{\epsilon}\right),$$

where $\sigma_a^2 = \text{Var}\left[a(1)\right]$ and $\sigma_{si}^2 = \text{Var}\left[s_i(1)\right]$

$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\sum q_i\right] = \frac{\sigma_a^2 + \sum \sigma_{si}^2}{2}$$

But... What is $o\left(\frac{1}{\epsilon}\right)$?



Theorem: [HL, Varma, Maguluri 2020]

In steady-state,

$$\left| \mathbb{E}\left[\sum_{i=1}^{N} q_i\right] - \frac{\sigma_a^2 + \sum \sigma_{si}^2}{2\epsilon} \right| \leq \beta \log\left(\frac{1}{\epsilon}\right),$$

where $\sigma_a^2 = \text{Var}\left[a(1)\right]$ and $\sigma_{si}^2 = \text{Var}\left[s_i(1)\right]$, and β is a constant.

Rate of convergence to heavy-traffic

Proof Outline

Theorem: [HL, Varma, Maguluri 2020]

In steady-state,

$$\left| \mathbb{E}\left[\sum_{i=1}^{N} q_i \right] - \frac{\sigma_a^2 + \sum \sigma_{si}^2}{2\epsilon} \right| \leq \beta \log\left(\frac{1}{\epsilon}\right),$$

where $\sigma_a^2 = \text{Var}\left[a(1)\right]$ and $\sigma_{si}^2 = \text{Var}\left[s_i(1)\right]$, and β is a constant.

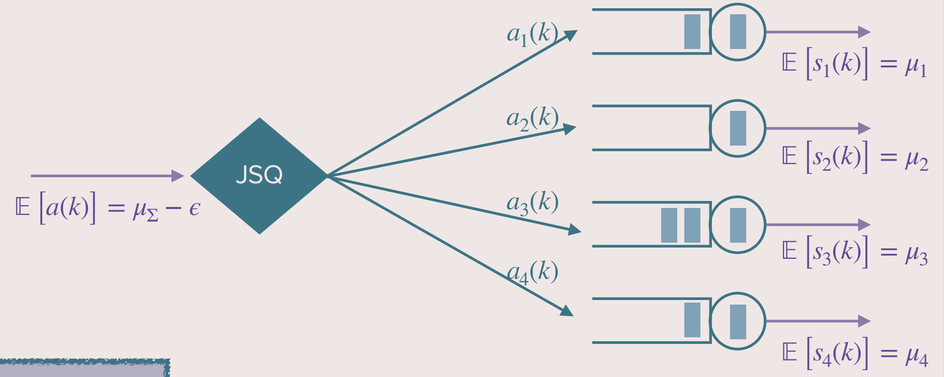
[Eryilmaz and Srikant 2012]

Set to zero the drift of the following test function:

$$V(\boldsymbol{q}) = \left(\sum q_i \right)^2$$

$$\implies \quad \mathbb{E}\left[V(\boldsymbol{q}(k))\right] = \mathbb{E}\left[V(\boldsymbol{q}(k+1))\right]$$

$$\implies \quad 2\epsilon \mathbb{E}\left[\sum q_i(k) \right] = \underbrace{\mathbb{E}\left[\left(a(1) - \sum s_i(1) \right)^2 \right]}_{= \sigma_a^2 + \sum \sigma_{si}^2 + \epsilon^2} - \underbrace{\mathbb{E}\left[\left(\sum u_i(k) \right)^2 \right]}_{\leq \beta_1 \epsilon} + \underbrace{\mathbb{E}\left[\left(\sum q_i(k+1) \right) \left(\sum u_i(k) \right) \right]}_{\to 0}$$



$$q_i(k+1) = q_i(k) + a_i(k) - s_i(k) + u_i(k)$$
$$q_i(k+1)u_i(k) = 0$$

Proof Outline (cont.)

Theorem: [HL, Varma, Maguluri 2020]

In steady-state,

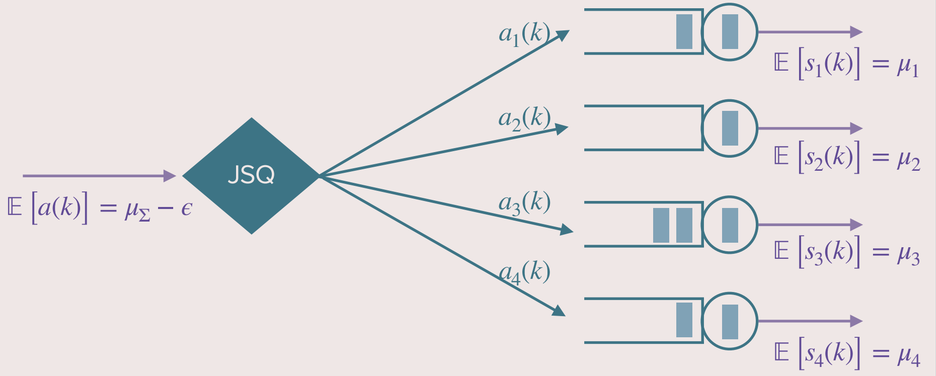
$$\left| \mathbb{E}\left[\sum_{i=1}^{N} q_i\right] - \frac{\sigma_a^2 + \sum \sigma_{si}^2}{2\epsilon} \right| \leq \beta \log\left(\tfrac{1}{\epsilon}\right),$$

where $\sigma_a^2 = \text{Var}\left[a(1)\right]$ and $\sigma_{si}^2 = \text{Var}\left[s_i(1)\right]$, and β is a constant.

$$\mathbb{E}\left[\left(\sum q_i(k+1)\right)\left(\sum u_i(k)\right)\right] = ??$$

State Space Collapse: All queues are equal in heavy traffic

$$\boldsymbol{q}_{\parallel} = \mathbf{1}\frac{\sum q_i}{N}\,,\ \boldsymbol{q}_{\perp} = \boldsymbol{q} - \boldsymbol{q}_{\parallel}$$



Eryilmaz and Srikant 2012: $\mathbb{E}\left[\|\boldsymbol{q}_{\perp}\|^r\right] \leq C(r)$

We prove that $\mathbb{E}\left[\|\boldsymbol{q}_{\perp}\|^r\right] \leq \overline{C}^r r!$

$$q_i(k+1) = q_i(k) + a_i(k) - s_i(k) + u_i(k)$$
$$q_i(k+1)u_i(k) = 0$$

$$\implies \left|\mathbb{E}\left[\left(\sum q_i(k+1)\right)\left(\sum u_i(k)\right)\right]\right| = N\left|\mathbb{E}\left[\langle \boldsymbol{q}_{\perp}(k+1), \boldsymbol{u}(k)\rangle\right]\right|$$

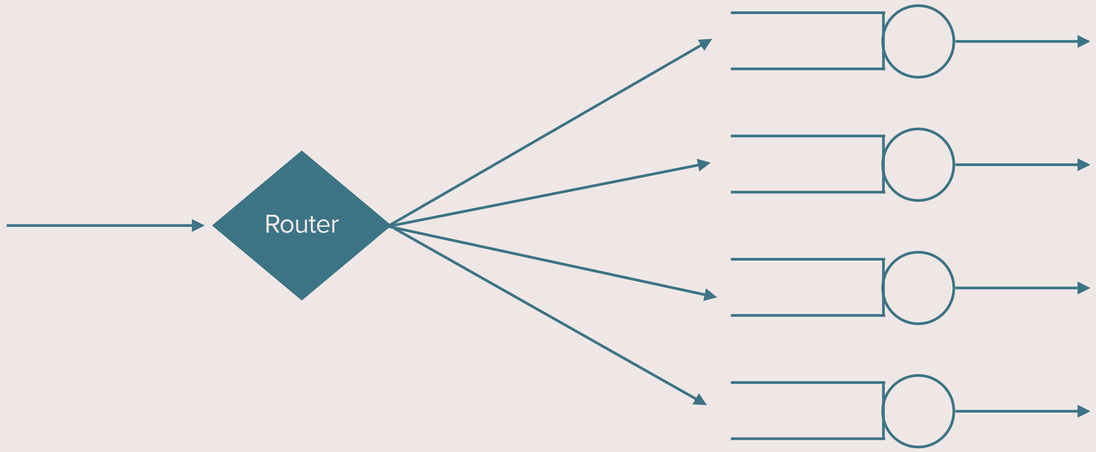
$$\leq N\mathbb{E}\left[\|\boldsymbol{q}_{\perp}\|^r\right]^{\frac{1}{r}}\mathbb{E}\left[\|\boldsymbol{u}\|^{\frac{r}{r-1}}\right]^{1-\frac{1}{r}}$$

SSC & Stirling's inequality

$$\leq \widetilde{C}\, e^{\frac{1}{r}} r^{\frac{1}{2r}+1} \epsilon^{1-\frac{1}{r}}$$

Last step: $r = \log\left(\frac{1}{\epsilon}\right)$ minimizes the upper bound.

Outline



Part I: Heavy-Traffic Bounds

- Discrete time model
- Join the Shortest Queue (JSQ) routing
- Eryilmaz and Srikant (2012): Asymptotically tight bounds
- HL, Varma, Maguluri (2020): Rate of convergence
- Essential steps in the proof

Part II: Many-Server Heavy-Traffic Regime

- What is this regime?
- Literature review
- Power-of-d choices routing
- HL, Maguluri (2020):
  - When queue lengths behave as in classical heavy-traffic?
  - Continuous vs. discrete time
- Ongoing: Water-filling

Part III: Throughput Optimality of Power-of-d choices

- What is throughput optimality?
- Maguluri and Srikant (2014): Throughput and delay optimality when all servers are equal
- HL, Maguluri (2020): Necessary and sufficient conditions for throughput optimality

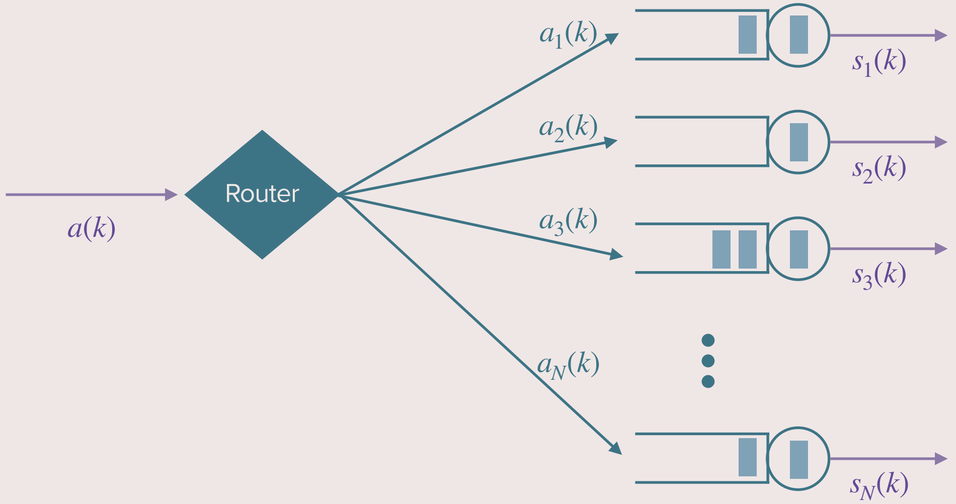
JSQ in Many-Server Heavy-Traffic Regime

Service process:

- All servers are equal
- $\mathbb{E}\left[s_i(1)\right] = 1$ and Var $\left[s_i(1)\right] = \sigma_s^2$

Arrival process:

- $\mathbb{E}\left[a(1)\right] = N(1 - N^{-\alpha})$ and Var $\left[a(1)\right] = N\sigma_a^2$
- *Arrivals "per server" have mean* $1 - N^{-\alpha}$ *and variance* σ_a^2



Theorem: [HL, Maguluri 2020]

In steady-state, and if $\alpha > 4$, we have

$$N^{-\alpha} \sum_{i=1}^{N} q_i \Rightarrow Expo\left(\frac{2}{\sigma_a^2 + \sigma_s^2}\right)$$

as $N \to \infty$

Can we get a result for $\alpha > 1$?

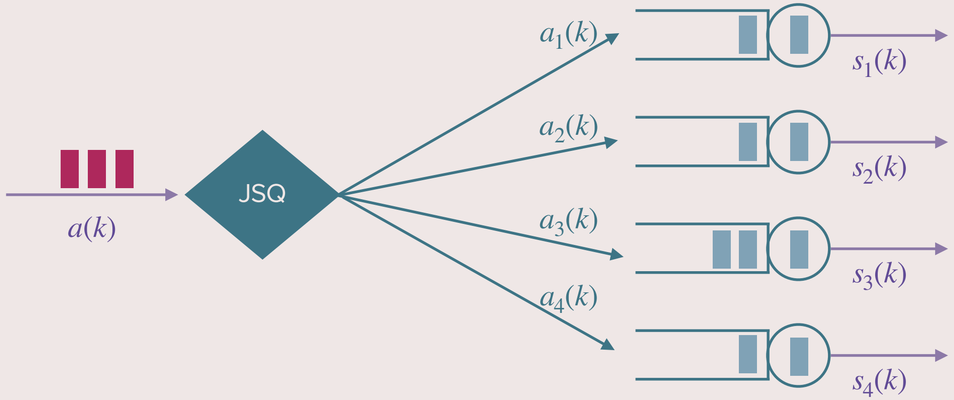
Conjecture: We can get it for $\alpha > 2$ if we change the routing policy

Same asymptotic behavior as classical heavy traffic!

What is wrong with JSQ?

- SSC: All queues are equal
  $\Longrightarrow$ Behavior is similar to a single server queue

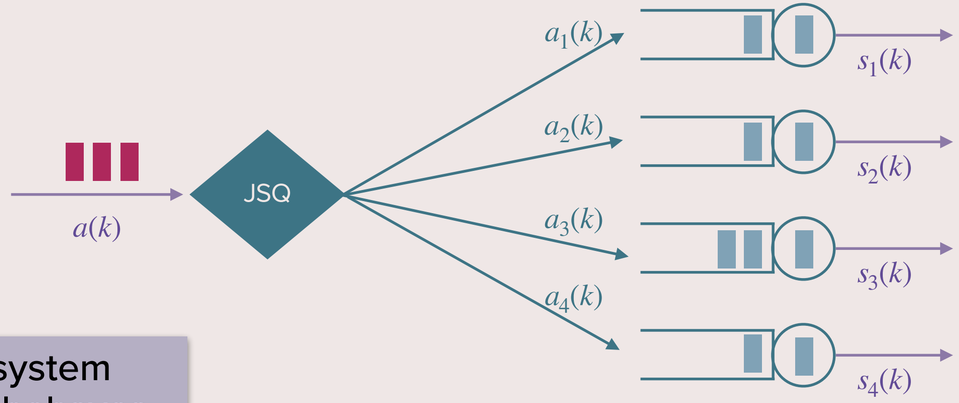
- JSQ routes **all** arrivals to the same queue
  $\Longrightarrow$ Causes “imbalance” in each time slot



Alternative: Water-filling

- Send **each of the arrivals** to the shortest queue

- “Helps” to keep all queues equal in each time slot

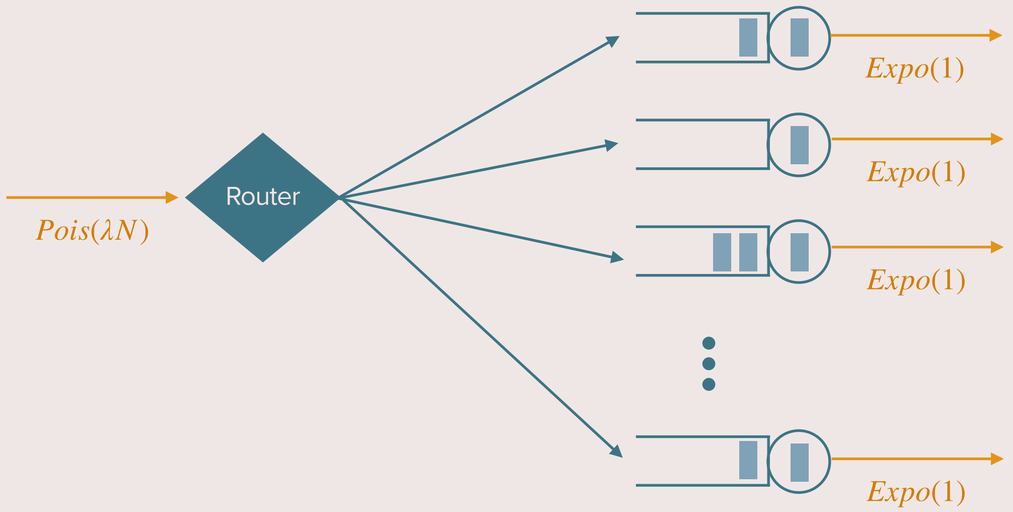


Conjecture: Load balancing system operating under water-filling behaves as in classical heavy-traffic for $\alpha > 2$

Load Balancing System in Continuous Time

Same as before

- N identical servers with an infinite buffer
- Single stream of arrivals
- $q_i(t)$: #jobs in queue i at time t



What changes?

- No more time slots! Time is continuous now!
- Service: Each job takes an $Expo(1)$ time to be processed
- Arrivals: Poisson process with rate λN

Power-of-d Choices in Many-Server Heavy-Traffic

Theorem: [HL, Maguluri 2020]

Let $d = cN^{\beta}$, where $c > 0$ and $\beta \geq 0$, and suppose $\lambda = 1 - N^{-\alpha}$. If $\alpha + \beta > 3$, then

$$N^{-\alpha} \sum_{i=1}^{N} q_i \Rightarrow Expo\,(1) \text{ as } N \to \infty$$

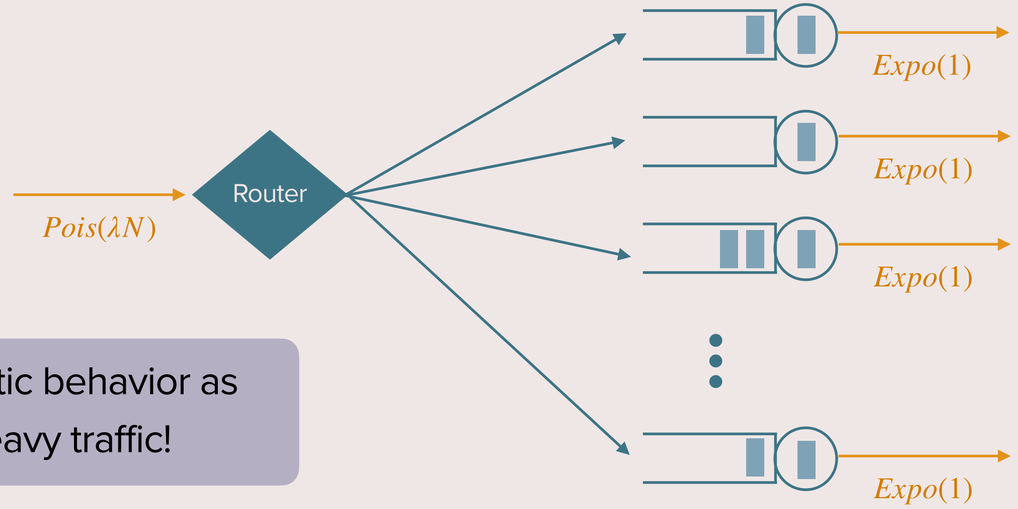
Corollary: [HL, Maguluri 2020]

Load balancing in continuous time, operating under JSQ. Then, if $\alpha > 2$ we have

$$N^{-\alpha} \sum_{i=1}^{N} q_i \Rightarrow Expo\,(1) \text{ as } N \to \infty$$

Proof: $c = 1$ and $\beta = 1$ in the theorem.

Same asymptotic behavior as classical heavy traffic!



Proof of the theorem:

Drift method in continuous time

1. Notion of State Space Collapse
2. Set to zero the drift of a function

Proof Outline: Computing $N^{-\alpha}\mathbb{E}\left[\sum q_i\right]$

Step 1: State Space Collapse

Recall $\boldsymbol{q}_{\parallel} = \mathbf{1}\dfrac{\sum q_i}{N}$ and $\boldsymbol{q}_{\perp} = \boldsymbol{q} - \boldsymbol{q}_{\parallel}$

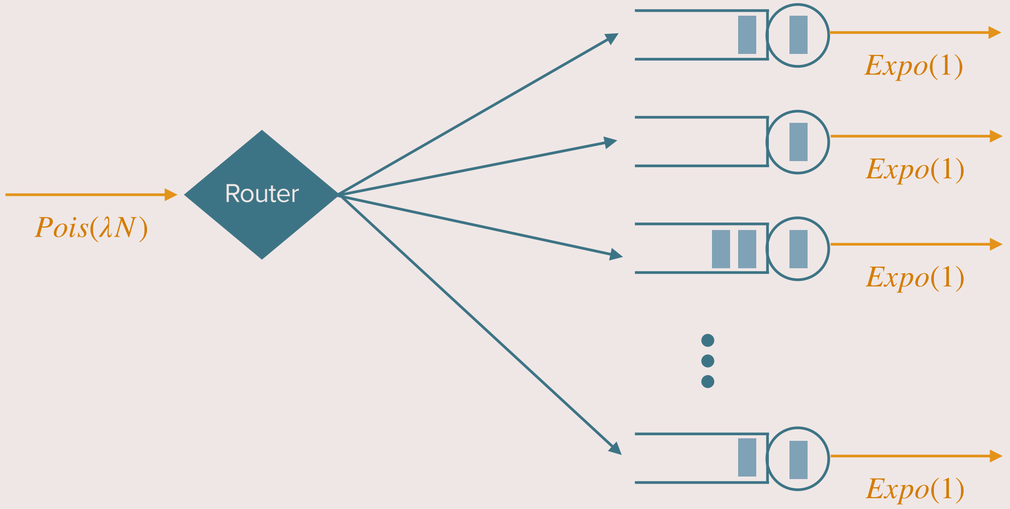
Proposition: [HL, Maguluri 2020]

Load balancing system in continuous time, operating under power-of-d with $d > 1$. Then, for any N large enough and any $r \in \mathbb{N}$ we have

$$\mathbb{E}\left[\|\boldsymbol{q}_{\perp}\|^r\right]^{\frac{1}{r}} \leq \tilde{C}\frac{rN^2}{d-1},$$

where $\tilde{C}$ is independent of d, r and N.

This bound is sufficient to show that the terms associated to $\boldsymbol{q}_{\perp}$ are negligible.



Theorem: Let $d = cN^{\beta}$, where $c > 0$ and $\beta \geq 0$, and suppose $\lambda = 1 - N^{-\alpha}$. If $\alpha + \beta > 3$, then

$$N^{-\alpha}\sum_{i=1}^{N} q_i \Rightarrow Expo\,(1) \text{ as } N \to \infty$$

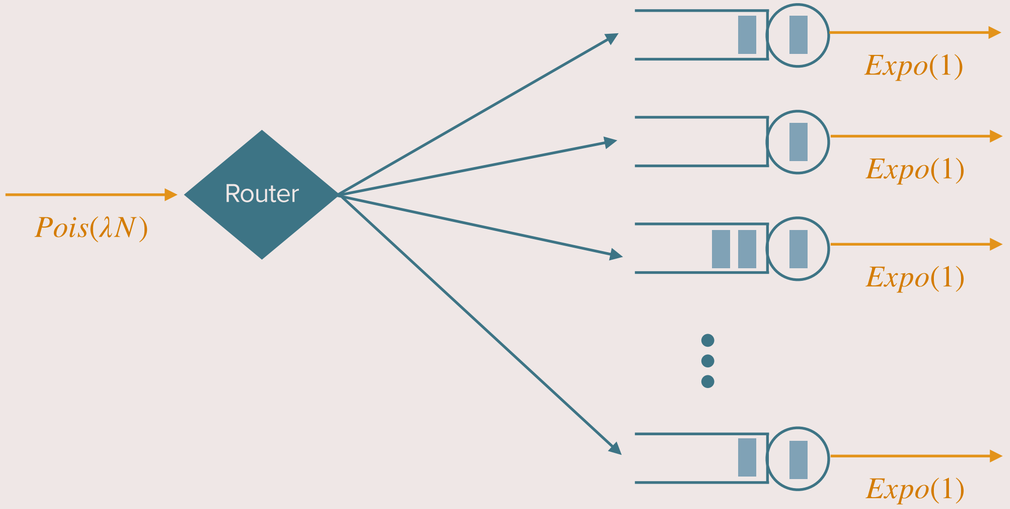
Proof Outline: Computing $N^{-\alpha}\mathbb{E}\left[\sum q_i\right]$ (cont.)

Step 2: Set to zero the drift of $V(\boldsymbol{q}) = \left(\sum q_i\right)^2$

Note: Continuous time analysis

$\Longrightarrow$ Drift defined in terms of the generator of the CTMC

$$\Delta V(\boldsymbol{q}) = \lambda N\left(\left(\sum q_i + 1\right)^2 - \left(\sum q_i\right)^2\right) + \sum_{i=1}^{N} 1\{q_i > 0\}\left(\left(\sum_{i'} q_{i'} - 1\right)^2 - \left(\sum_{i'} q_{i'}\right)^2\right)$$



$$\mathbb{E}\left[\Delta V(\boldsymbol{q})\right] = 0$$

$$\Longrightarrow N^{-\alpha}\mathbb{E}\left[\sum q_i\right] = 1 - N^{-\alpha} + \frac{1}{N}\mathbb{E}\left[\left(\sum_i q_i\right)\left(\sum_i 1\{q_i = 0\}\right)\right]$$

Need to show: $\to 0$

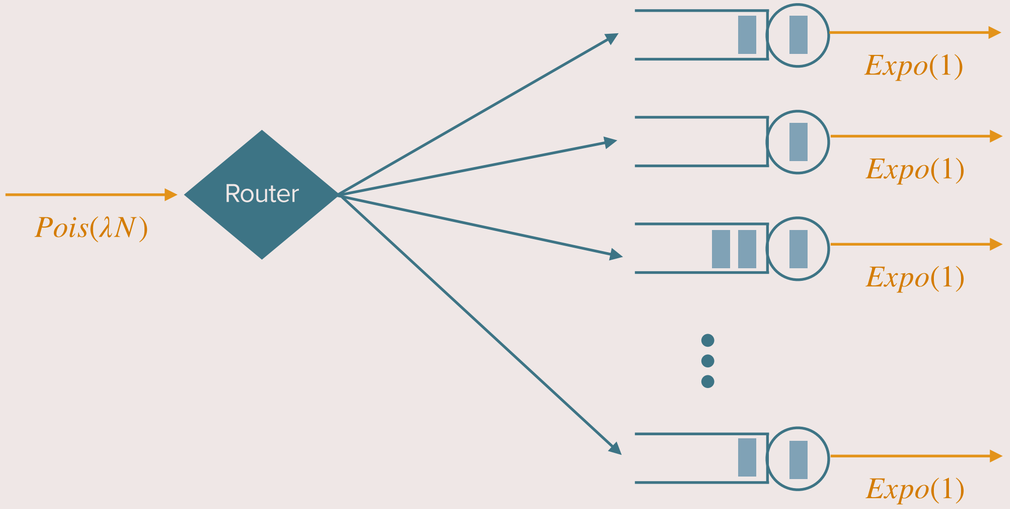
Theorem: Let $d = cN^{\beta}$, where $c > 0$ and $\beta \geq 0$, and suppose $\lambda = 1 - N^{-\alpha}$. If $\alpha + \beta > 3$, then

$$N^{-\alpha}\sum_{i=1}^{N} q_i \Rightarrow Expo\,(1) \text{ as } N \to \infty$$

Proof Outline: Computing $N^{-\alpha}\mathbb{E}\left[\sum q_i\right]$

$$\frac{1}{N}\mathbb{E}\left[\left(\sum_i q_i\right)\left(\sum_i 1\{q_i=0\}\right)\right] = ??$$

$\approx$ unused service in discrete time



Observe: $\left(\sum_i \frac{q_i}{N}\right)\left(\sum_i 1\{q_i=0\}\right) = \sum_i 1\{q_i=0\}q_{\perp i}$

$$\Longrightarrow \left|\frac{1}{N}\mathbb{E}\left[\left(\sum_i q_i\right)\left(\sum_i 1\{q_i=0\}\right)\right]\right| \leq \mathbb{E}\left[\|\boldsymbol{q}_\perp\|^r\right]^{\frac{1}{r}}\mathbb{E}\left[\sum_i 1\{q_i=0\}\right]^{1-\frac{1}{r}}$$

SSC

Set to zero the drift of $V'(\boldsymbol{q}) = \sum q_i$

$$\leq \tilde{C}\frac{rN^2}{d-1}N^{(1-\alpha)\left(1-\frac{1}{r}\right)}$$

If $d = cN^\beta$

$$\approx N^{3-\alpha-\beta}$$

Theorem: Let $d = cN^\beta$, where $c > 0$ and $\beta \geq 0$, and suppose $\lambda = 1 - N^{-\alpha}$. If $\alpha + \beta > 3$, then

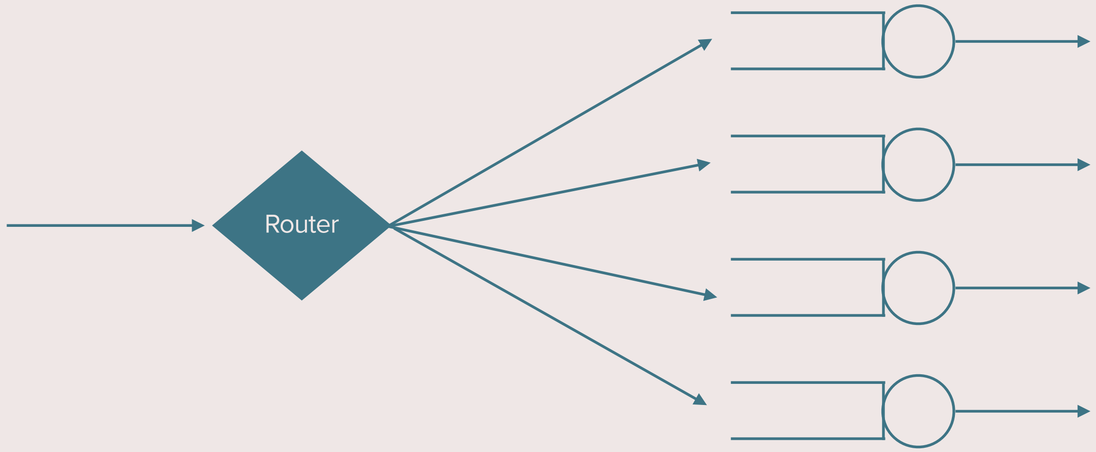
$$N^{-\alpha}\sum_{i=1}^{N} q_i \Rightarrow Expo\,(1) \text{ as } N \to \infty$$

Proposition: Load balancing system in continuous time, operating under power-of-d with $d > 1$. Then, for any N large enough and any $r \in \mathbb{N}$ we have

$$\mathbb{E}\left[\|\boldsymbol{q}_\perp\|^r\right]^{\frac{1}{r}} \leq \tilde{C}\frac{rN^2}{d-1},$$

where $\tilde{C}$ is independent of d, r and N.

Outline



Part I: Heavy-Traffic Bounds

- Discrete time model
- Join the Shortest Queue (JSQ) routing
- Eryilmaz and Srikant (2012): Asymptotically tight bounds
- HL, Varma, Maguluri (2020): Rate of convergence
- Essential steps in the proof

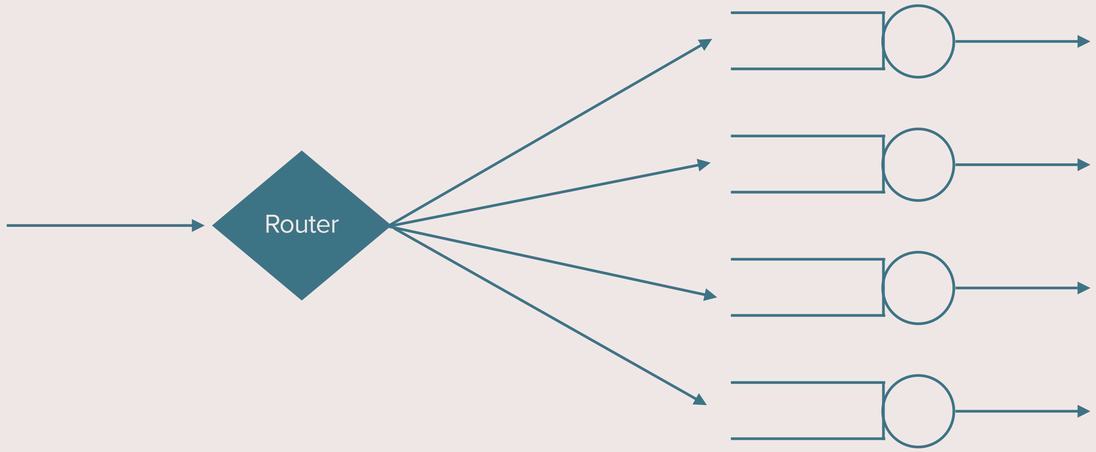
Part II: Many-Server Heavy-Traffic Regime

- What is this regime?
- Literature review
- Power-of-d choices routing
- HL, Maguluri (2020):
  - When queue lengths behave as in classical heavy-traffic?
  - Continuous vs. discrete time
- Ongoing: Water-filling

Part III: Throughput Optimality of Power-of-d choices

- What is throughput optimality?
- Maguluri and Srikant (2014): Throughput and delay optimality when all servers are equal
- HL, Maguluri (2020): Necessary and sufficient conditions for throughput optimality

Outline



Part I: Heavy-Traffic Bounds

- Discrete time model
- Join the Shortest Queue (JSQ) routing
- Eryilmaz and Srikant (2012): Asymptotically tight bounds
- HL, Varma, Maguluri (2020): Rate of convergence
- Essential steps in the proof

Part II: Many-Server Heavy-Traffic Regime

- What is this regime?
- Literature review
- Power-of-d choices routing
- HL, Maguluri (2020):
  - When queue lengths behave as in classical heavy-traffic?
  - Continuous vs. discrete time
- Ongoing: Water-filling

Part III: Throughput Optimality of Power-of-d choices

- What is throughput optimality?
- Maguluri and Srikant (2014): Throughput and delay optimality when all servers are equal
- HL, Maguluri (2020): Necessary and sufficient conditions for throughput optimality