Degree of Queue Imbalance: Overcoming the Limitation of Heavy-traffic Delay Optimality in Load Balancing Systems

Xingyu Zhou

The Ohio State University

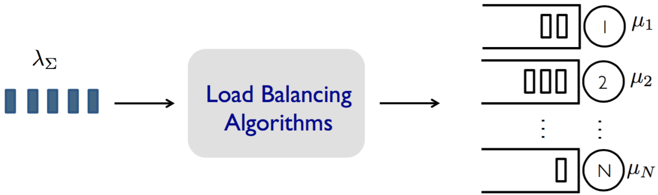
Joint work with...

Fei Wu*, OSU (co-primal)

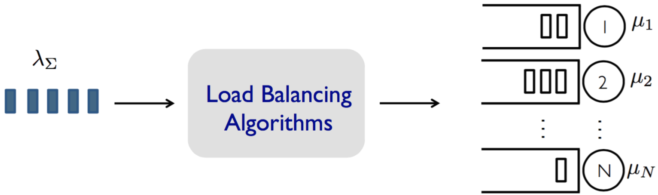
Jian Tan, OSU

Kannan Srinivasan, OSU

Ness Shroff, OSU

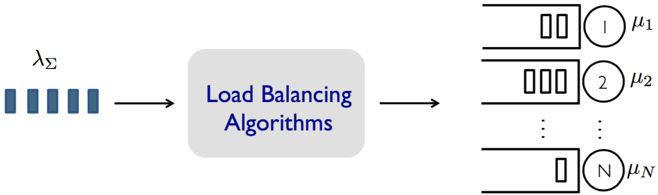


λ_Σ
Load Balancing Algorithms
1
μ_1
2
μ_2
N
μ_N



The goal of load balancing:

choose the *right* server(s) for each request.



The goal of load balancing:

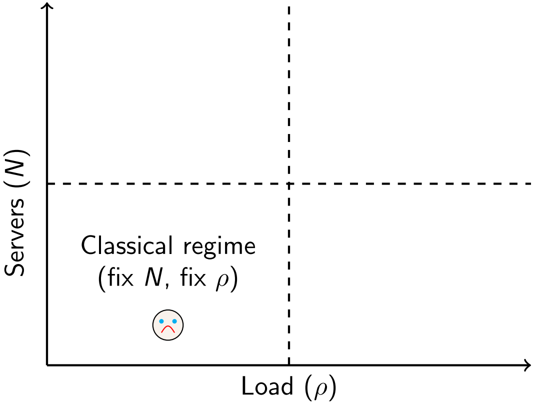
choose the *right* server(s) for each request.

What does *right* mean?

Low delay

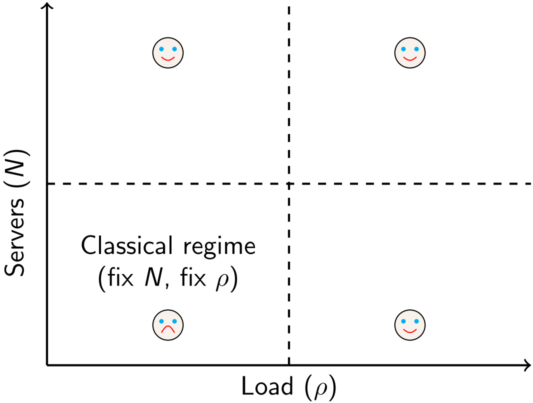
Low delay

- Classical regime is very difficult.



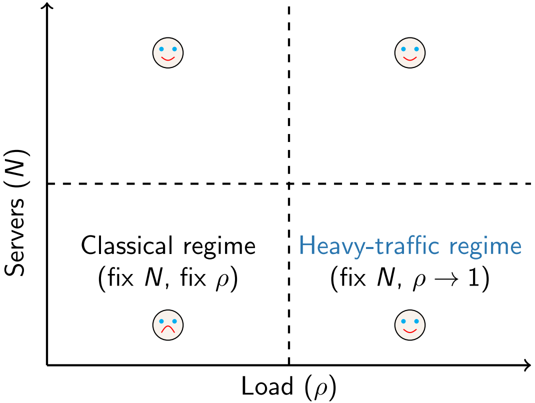
Low delay

- Classical regime is very difficult.
- Turn to asymptotic regimes for insights.



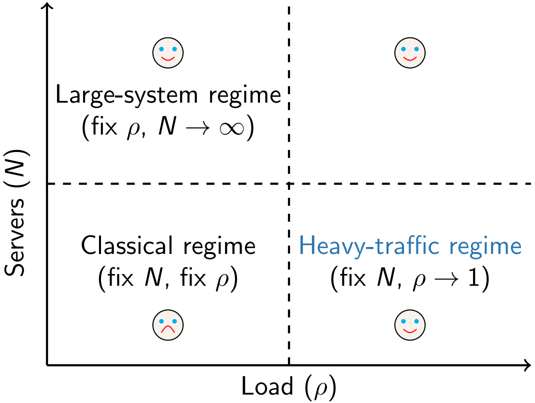
Low delay

- Classical regime is very difficult.
- Turn to asymptotic regimes for insights.



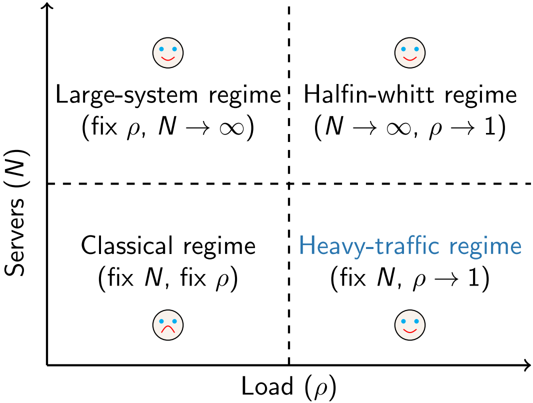
Low delay

- Classical regime is very difficult.
- Turn to asymptotic regimes for insights.



Low delay

- Classical regime is very difficult.
- Turn to asymptotic regimes for insights.



In this talk, we focus heavy-traffic regime, ask two questions below:

1. Question: How large can the difference be in the empirical delay for different 'optimal' schemes?

In this talk, we focus heavy-traffic regime, ask two questions below:

1. Question: How large can the difference be in the empirical delay for different 'optimal' schemes?
   - we know 'optimality' exists in heavy-traffic limit.

In this talk, we focus heavy-traffic regime, ask two questions below:

1. Question: How large can the difference be in the empirical delay for different ‘optimal’ schemes?
   - we know ‘optimality’ exists in heavy-traffic limit.
   - but, how much does it tell about moderate load?

In this talk, we focus heavy-traffic regime, ask two questions below:

1. Question: How large can the difference be in the empirical delay for different 'optimal' schemes?
   - we know 'optimality' exists in heavy-traffic limit.
   - but, how much does it tell about moderate load?
   - how far away from just random routing in empirical performance?

In this talk, we focus heavy-traffic regime, ask two questions below:

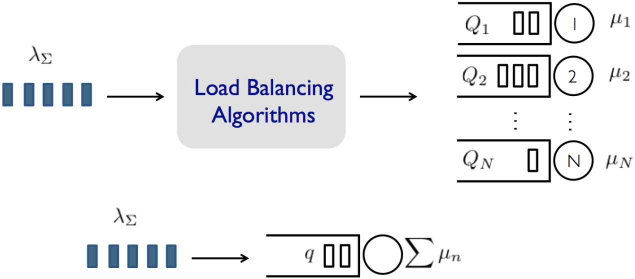
1. Question: How large can the difference be in the empirical delay for different 'optimal' schemes?
   - we know 'optimality' exists in heavy-traffic limit.
   - but, how much does it tell about moderate load?
   - how far away from just random routing in empirical performance?

2. Question: Can we characterize the difference and differentiate the policies that are 'optimal'?

Before we start...

Definition (Heavy-traffic Delay Optimal)

It can achieve the lower bound on delay when $\epsilon \to 0$, that is, $\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\sum Q_n\right] = \lim_{\epsilon \downarrow 0} \mathbb{E}\left[q\right]$



Fact: $\mathbb{E}\left[\sum Q_n\right] \geq \mathbb{E}\left[q\right]$, since packet remains in the queue until finished.

Quiz time....

Consider the following policy: at each time-slot t, it adopts JSQ w.p. p, otherwise just uses Random.

Quiz time....

Consider the following policy: at each time-slot t, it adopts JSQ w.p. p, otherwise just uses Random.

Question: Which of the following p value guarantees 'optimality'?

(A). $p = 1$ (B). $p = 0.5$ (C). $p = 0.1$ (D). $p = 0.00001$

Quiz time....

Consider the following policy: at each time-slot t, it adopts JSQ w.p. p, otherwise just uses Random.

Question: Which of the following p value guarantees 'optimality'?

(A). $p = 1$ (B). $p = 0.5$ (C). $p = 0.1$ (D). $p = 0.00001$

All the choices are correct!

Part I: Limitation of heavy-traffic optimality in load balancing

Main Result

Question: How large can the difference be in the empirical delay for different 'optimal' schemes?

Main Result

Question: How large can the difference be in the empirical delay for different 'optimal' schemes?

Answer: The empirical delay of 'optimal' policies can range from JSQ to arbitrarily close to Random ($p = 0.00001$)

Main Result

Question: How large can the difference be in the empirical delay for different 'optimal' schemes?

Answer: The empirical delay of 'optimal' policies can range from JSQ to arbitrarily close to Random $(p = 0.00001)$

- A very weak condition is enough: in the long-term, the dispatcher favors (even slightly) shorter queues.

Main Result

Question: How large can the difference be in the empirical delay for different 'optimal' schemes?

Answer: The empirical delay of 'optimal' policies can range from JSQ to arbitrarily close to Random ($p = 0.00001$)

- A very weak condition is enough: in the long-term, the dispatcher favors (even slightly) shorter queues.
- This condition is called LDPC: Long-term Dispatching Preference Condition.

Dispatching distribution and preference

Let us focus on homogeneous servers first.

The *n*th component of dispatching distribution $\mathbf{P}(t)$ is the *probability* of dispatching arrival to the *n*th *shortest* queue.

Dispatching distribution and preference

Let us focus on homogeneous servers first.

The nth component of dispatching distribution $\mathbf{P}(t)$ is the *probability* of dispatching arrival to the nth *shortest* queue.

We also define dispatching preference

$$\Delta(t) \triangleq \mathbf{P}(t) - \mathbf{P}_{\text{rand}}(t)$$

where $\mathbf{P}_{\text{rand}}(t)$ is the dispatching distribution under random routing.

Dispatching distribution and preference

Let us focus on homogeneous servers first.

The *n*th component of dispatching distribution $\mathbf{P}(t)$ is the *probability* of dispatching arrival to the *n*th *shortest* queue.

We also define dispatching preference

$$\Delta(t) \triangleq \mathbf{P}(t) - \mathbf{P}_{\text{rand}}(t)$$

where $\mathbf{P}_{\text{rand}}(t)$ is the dispatching distribution under random routing.

Let random vector $\overline{\Delta}$ denote the dispatching preference in steady-state.

$$\widetilde{\Delta} = \mathbb{E}\left[\overline{\Delta}\right]$$

Example

Let consider a homogeneous case with 3 servers.

Example

Let consider a homogeneous case with 3 servers.

- Random: randomly joins one
  - $\mathbf{P}_{\text{rand}}(t) = (1/3, 1/3, 1/3)$
  - $\Delta(t) = \overline{\Delta} = \widetilde{\Delta} = (0, 0, 0)$

Example

Let consider a homogeneous case with 3 servers.

- Random: randomly joins one
  - $\mathbf{P}_{\text{rand}}(t) = (1/3, 1/3, 1/3)$
  - $\Delta(t) = \overline{\Delta} = \widetilde{\Delta} = (0, 0, 0)$

- JSQ: always join the shortest one
  - $\mathbf{P}_{\text{JSQ}}(t) = (1, 0, 0)$
  - $\Delta_{JSQ}(t) = \overline{\Delta} = \widetilde{\Delta} = (2/3, -1/3, -1/3)$

Example

Let consider a homogeneous case with 3 servers.

- Random: randomly joins one
  - $\mathbf{P}_{\text{rand}}(t) = (1/3, 1/3, 1/3)$
  - $\Delta(t) = \overline{\Delta} = \widetilde{\Delta} = (0, 0, 0)$
- JSQ: always join the shortest one
  - $\mathbf{P}_{\text{JSQ}}(t) = (1, 0, 0)$
  - $\Delta_{JSQ}(t) = \overline{\Delta} = \widetilde{\Delta} = (2/3, -1/3, -1/3)$
- Power of 2: randomly picks two and joins the shorter one
  - $\mathbf{P}_{\text{Po2}}(t) = (2/3, 1/3, 0)$
  - $\Delta_{Po2}(t) = \overline{\Delta} = \widetilde{\Delta} = (1/3, 0, -1/3)$

Example

Let consider a homogeneous case with 3 servers.

- Random: randomly joins one
  - $\mathbf{P}_{\text{rand}}(t) = (1/3, 1/3, 1/3)$
  - $\Delta(t) = \overline{\Delta} = \widetilde{\Delta} = (0, 0, 0)$
- JSQ: always join the shortest one
  - $\mathbf{P}_{\text{JSQ}}(t) = (1, 0, 0)$
  - $\Delta_{JSQ}(t) = \overline{\Delta} = \widetilde{\Delta} = (2/3, -1/3, -1/3)$
- Power of 2: randomly picks two and joins the shorter one
  - $\mathbf{P}_{\text{Po2}}(t) = (2/3, 1/3, 0)$
  - $\Delta_{Po2}(t) = \overline{\Delta} = \widetilde{\Delta} = (1/3, 0, -1/3)$
- p-JSQ: JSQ w.p. p + Random w.p. $1 - p$, e.g., $p = 0.5$
  - $\mathbf{P}_{0.5\text{-JSQ}}(t) = (1, 0, 0)$ or $\mathbf{P}_{0.5\text{-JSQ}}(t) = (1/3, 1/3, 1/3)$
  - $\overline{\Delta} = (2/3, -1/3, -1/3)$ or $\overline{\Delta} = (1/3, 0, -1/3)$, *with equal prob.*
  - $\widetilde{\Delta} = (1/2, -1/6, -1/3)$.

Long-term Dispatching Preference Condition

Definition (LDPC)

A load balancing scheme is said to satisfy the LDPC if

$$\widetilde{\Delta}_1 \geq \widetilde{\Delta}_2 \geq \ldots \geq \widetilde{\Delta}_N \quad \text{and} \quad \widetilde{\Delta}_1 \neq \widetilde{\Delta}_N.$$

Key message: 'slightly prefer shorter queues in the long-term'

Long-term Dispatching Preference Condition

Definition (LDPC)

A load balancing scheme is said to satisfy the LDPC if

$$\widetilde{\Delta}_1 \geq \widetilde{\Delta}_2 \geq \ldots \geq \widetilde{\Delta}_N \quad \text{and} \quad \widetilde{\Delta}_1 \neq \widetilde{\Delta}_N.$$

Key message: ‘slightly prefer shorter queues in the long-term’

Theorem (LDPC $\implies$ optimality)

Any load balancing scheme satisfying LDPC is heavy-traffic delay optimal.

Long-term Dispatching Preference Condition

Definition (LDPC)

A load balancing scheme is said to satisfy the LDPC if

$$\widetilde{\Delta}_1 \geq \widetilde{\Delta}_2 \geq \ldots \geq \widetilde{\Delta}_N \quad \text{and} \quad \widetilde{\Delta}_1 \neq \widetilde{\Delta}_N.$$

Key message: 'slightly prefer shorter queues in the long-term'

Theorem (LDPC $\implies$ optimality)

Any load balancing scheme satisfying LDPC is heavy-traffic delay optimal.

Every coin has two sides:

Long-term Dispatching Preference Condition

Definition (LDPC)

A load balancing scheme is said to satisfy the LDPC if

$$\widetilde{\Delta}_1 \geq \widetilde{\Delta}_2 \geq \ldots \geq \widetilde{\Delta}_N \quad \text{and} \quad \widetilde{\Delta}_1 \neq \widetilde{\Delta}_N.$$

Key message: ‘slightly prefer shorter queues in the long-term’

Theorem (LDPC $\implies$ optimality)

Any load balancing scheme satisfying LDPC is heavy-traffic delay optimal.

Every coin has two sides:

-  we have an even larger class of optimal policies. (JSQ, Power-of-d, and more flexible ones...)

Long-term Dispatching Preference Condition

Definition (LDPC)

A load balancing scheme is said to satisfy the LDPC if

$$\widetilde{\Delta}_1 \geq \widetilde{\Delta}_2 \geq \ldots \geq \widetilde{\Delta}_N \quad \text{and} \quad \widetilde{\Delta}_1 \neq \widetilde{\Delta}_N.$$

Key message: 'slightly prefer shorter queues in the long-term'

Theorem (LDPC $\implies$ optimality)

Any load balancing scheme satisfying LDPC is heavy-traffic delay optimal.

Every coin has two sides:

- ☺ we have an even larger class of optimal policies. (JSQ, Power-of-d, and more flexible ones...)
- ☹ we have many poor polices even though they are optimal.

Long-term Dispatching Preference Condition

Definition (LDPC)

A load balancing scheme is said to satisfy the LDPC if

$$\widetilde{\Delta}_1 \geq \widetilde{\Delta}_2 \geq \ldots \geq \widetilde{\Delta}_N \quad \text{and} \quad \widetilde{\Delta}_1 \neq \widetilde{\Delta}_N.$$

Key message: 'slightly prefer shorter queues in the long-term'

Theorem (LDPC $\implies$ optimality)

Any load balancing scheme satisfying LDPC is heavy-traffic delay optimal.

Every coin has two sides:

- we have an even larger class of optimal policies. (JSQ, Power-of-d, and more flexible ones...)
- we have many poor polices even though they are optimal.
  - p-JSQ (w.p. p JSQ, otherwise Random) satisfies LDPC for any $p > 0$

Long-term Dispatching Preference Condition

Definition (LDPC)

A load balancing scheme is said to satisfy the LDPC if

$$\widetilde{\Delta}_1 \geq \widetilde{\Delta}_2 \geq \ldots \geq \widetilde{\Delta}_N \quad \text{and} \quad \widetilde{\Delta}_1 \neq \widetilde{\Delta}_N.$$

Key message: 'slightly prefer shorter queues in the long-term'

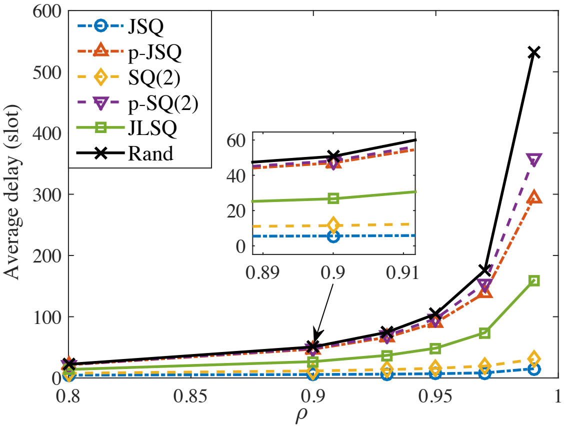
Theorem (LDPC $\Longrightarrow$ optimality)

Any load balancing scheme satisfying LDPC is heavy-traffic delay optimal.

Every coin has two sides:

- we have an even larger class of optimal policies. (JSQ, Power-of-d, and more flexible ones...)
- we have many poor polices even though they are optimal.
  - p-JSQ (w.p. p JSQ, otherwise Random) satisfies LDPC for any $p > 0$
  - Join longer or shorter queue (JLSQ) satisfies LDPC for any $p < \frac{N_1}{N}$
    - join one of the N_1 longest queue w.p. p
    - otherwise, join one of the $N - N_1$ queues.

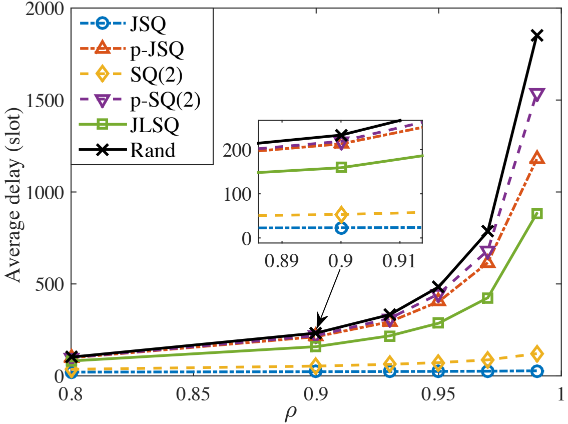
Simulations



- number of servers: $N = 10$
- p-JSQ and p-SQ(2): $p = 0.01$
- JLSQ: $N_1 = N/2$, $p = 0.49$

In this setting, delay of p-SQ(2) is 20x larger than JSQ even at $\rho = 0.99$

Simulations (Cont'd)



- number of servers: $N = 50$
- p-JSQ and p-SQ(2): $p = 0.01$
- JLSQ: $N_1 = N/2$, $p = 0.49$

In this setting, delay of p-SQ(2) is 50x larger than JSQ even at $\rho = 0.99$

What we have shown...

- For load balancing, heavy-traffic optimality may be a coarse metric.
- The practical performance of theoretically optimal scheme has huge difference:

What we have shown...

- For load balancing, heavy-traffic optimality may be a coarse metric.
- The practical performance of theoretically optimal scheme has huge difference:
  - it can range from that of JSQ to that of (arbitrarily close) Random.

What we have shown...

- For load balancing, heavy-traffic optimality may be a coarse metric.
- The practical performance of theoretically optimal scheme has huge difference:
  - it can range from that of JSQ to that of (arbitrarily close) Random.
  - since 'optimality' only requires a long-term preference on shorter queues, i.e, LDPC.

What we have shown...

- For load balancing, heavy-traffic optimality may be a coarse metric.
- The practical performance of theoretically optimal scheme has huge difference:
  - it can range from that of JSQ to that of (arbitrarily close) Random.
  - since 'optimality' only requires a long-term preference on shorter queues, i.e, LDPC.

Question: Can we characterize the difference and differentiate them?

Part II: A Refined Metric

Quiz time....

Consider the following policy: at each time-slot t, it adopts JSQ w.p. p, otherwise just uses Random.

Quiz time....

Consider the following policy: at each time-slot t, it adopts JSQ w.p. p, otherwise just uses Random.

Question: Give the order of 'goodness' of the following choices of p?

(A). $p = 1$ (B). $p = 0.5$ (C). $p = 0.1$ (D). $p = 0.00001$

Quiz time....

Consider the following policy: at each time-slot t, it adopts JSQ w.p. p, otherwise just uses Random.

Question: Give the order of 'goodness' of the following choices of p?

(A). $p = 1$ (B). $p = 0.5$ (C). $p = 0.1$ (D). $p = 0.00001$

$$A > B > C > D$$

How close to Random...

How close to Random...

Definition

The degree of dispatching preference for a given load balancing scheme is given by the L_1 norm of the long-term dispatching preference, i.e., $\left\|\widetilde{\Delta}\right\|_1$.

'degree of dispatching preference $= \left\|\widetilde{\Delta}\right\|_1$'

How close to Random...

Definition

The degree of dispatching preference for a given load balancing scheme is given by the L_1 norm of the long-term dispatching preference, i.e., $\left\|\widetilde{\Delta}\right\|_1$.

$$\text{'degree of dispatching preference} = \left\|\widetilde{\Delta}\right\|_1\text{'}$$

Note:

- it is actually the *total variation distance* from Random.
  - $\left\|\widetilde{\Delta}\right\|_1 = \left\|\widetilde{\mathbf{P}} - \mathbf{P}_{\text{rand}}\right\|_1 = 2\left\|\widetilde{\mathbf{P}} - \mathbf{P}_{\text{rand}}\right\|_{tv}$

How close to Random...

Definition

The degree of dispatching preference for a given load balancing scheme is given by the L_1 norm of the long-term dispatching preference, i.e., $\left\|\widetilde{\Delta}\right\|_1$.

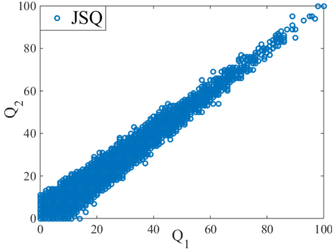
$$\text{'degree of dispatching preference} = \left\|\widetilde{\Delta}\right\|_1\text{'}$$

Note:

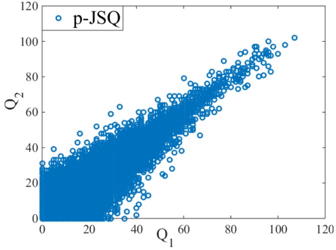
- it is actually the *total variation distance* from Random.
  - $\left\|\widetilde{\Delta}\right\|_1 = \left\|\widetilde{\mathbf{P}} - \mathbf{P}_{\text{rand}}\right\|_1 = 2\left\|\widetilde{\mathbf{P}} - \mathbf{P}_{\text{rand}}\right\|_{tv}$
- minimum attained at Random, maximum at JSQ.
- for p-JSQ, $\left\|\widetilde{\Delta}\right\|_1 \to 0$ as $p \to 0$.

What's the result of different degree of dispatching preference?

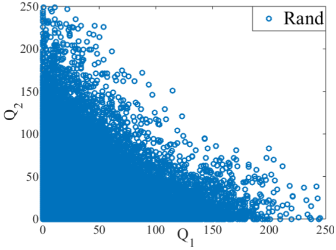
Intuition...



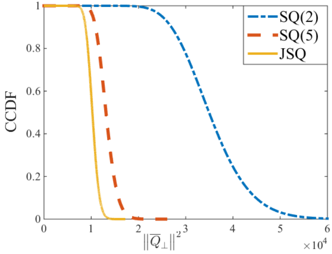
(a) $N = 2$, JSQ



(b) $N = 2$, p-JSQ ($p = 0.5$)



(c) $N = 2$, Random

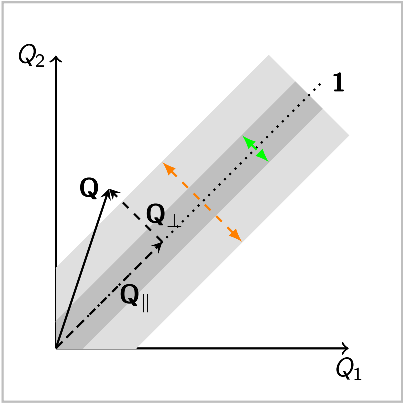


(d) $N = 10$, CCDF

A Refined Metric

Definition

The degree of queue imbalance in a load balancing system with a steady-state queue length vector $\overline{\mathbf{Q}}$ is given by $\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}\right\|^2\right]$, where $\mathbf{Q}_{\perp} \triangleq \mathbf{Q}(t) - \mathbf{Q}_{\|}(t) = \langle \mathbf{Q}, \mathbf{1} \rangle \mathbf{1}$.



The closer, The worse...

Theorem

Under any load balancing scheme satisfying LDPC, the degree of queue imbalance is on the order of

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|^2\right] = \Theta\left(\frac{1}{\left\|\widetilde{\Delta}\right\|_1^2}\right).$$

$$\text{Degree of Queue Imbalance} \approx \frac{1}{(\text{Degree of Dispatching Preference})^2}$$

The closer, The worse...

Theorem

Under any load balancing scheme satisfying LDPC, the degree of queue imbalance is on the order of

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|^2\right] = \Theta\left(\frac{1}{\left\|\widetilde{\Delta}\right\|_1^2}\right).$$

$$\text{Degree of Queue Imbalance} \approx \frac{1}{(\text{Degree of Dispatching Preference})^2}$$

Take our favorite p-JSQ and p-power-of-d for example:

The closer, The worse...

Theorem

Under any load balancing scheme satisfying LDPC, the degree of queue imbalance is on the order of

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\| \overline{\mathbf{Q}}_{\perp}^{(\epsilon)} \right\|^2 \right] = \Theta\left(\frac{1}{\left\| \widetilde{\Delta} \right\|_1^2} \right).$$

$$\text{Degree of Queue Imbalance} \approx \frac{1}{(\text{Degree of Dispatching Preference})^2}$$

Take our favorite p-JSQ and p-power-of-d for example:

- Part I shows that for any $p > 0$, they remain 'optimal'.

The closer, The worse...

Theorem

Under any load balancing scheme satisfying LDPC, the degree of queue imbalance is on the order of

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|^2\right] = \Theta\left(\frac{1}{\left\|\widetilde{\Delta}\right\|_1^2}\right).$$

$$\text{Degree of Queue Imbalance} \approx \frac{1}{(\text{Degree of Dispatching Preference})^2}$$

Take our favorite p-JSQ and p-power-of-d for example:

- Part I shows that for any $p > 0$, they remain 'optimal'.
- But, the empirical delay gets worse as $p \to 0$.

The closer, The worse...

Theorem

Under any load balancing scheme satisfying LDPC, the degree of queue imbalance is on the order of

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|^2\right] = \Theta\left(\frac{1}{\left\|\widetilde{\Delta}\right\|_1^2}\right).$$

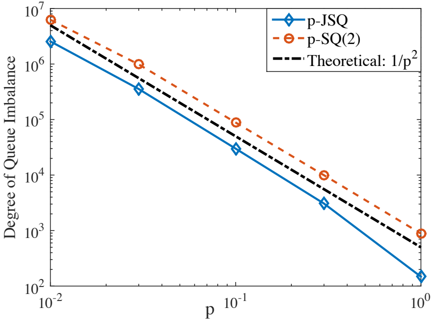
$$\text{Degree of Queue Imbalance} \approx \frac{1}{(\text{Degree of Dispatching Preference})^2}$$

Take our favorite p-JSQ and p-power-of-d for example:

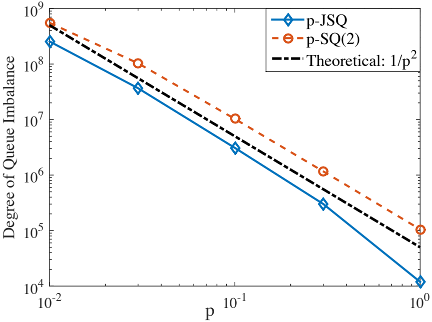
- Part I shows that for any $p > 0$, they remain 'optimal'.
- But, the empirical delay gets worse as $p \to 0$.
- The above theorem tells us the degree of queue imbalance $\to \infty$ on the order $\Theta\left(\frac{1}{p^2}\right)$ as $p \to 0$.

Degree of Queue Imbalance vs. p

$$\text{Degree of Queue Imbalance} \approx \frac{1}{p^2}$$



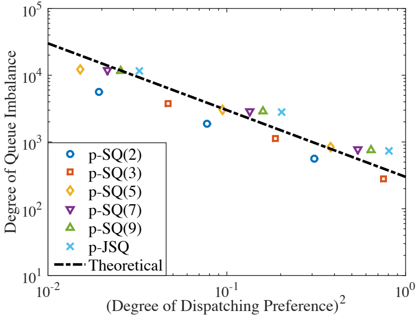
(a) $N = 10$, $\epsilon = 0.001$



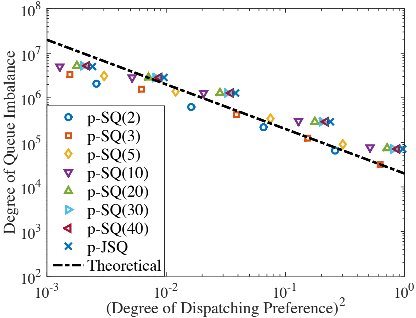
(b) $N = 50$, $\epsilon = 0.001$

Degree of Queue Imbalance vs. $\|\widetilde{\Delta}\|_1$

$$\text{Degree of Queue Imbalance} \approx \frac{1}{(\text{Degree of Dispatching Preference})^2}$$



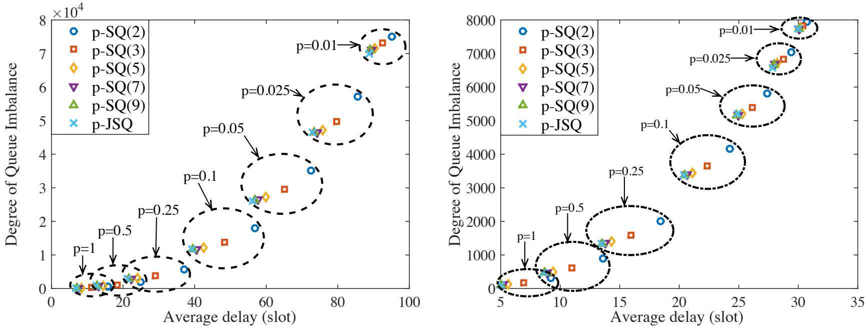
(a) $N = 10,\ \rho = 0.95$



(b) $N = 50,\ \rho = 0.95$

Degree of queue imbalance VS. Delay ($N = 10$)

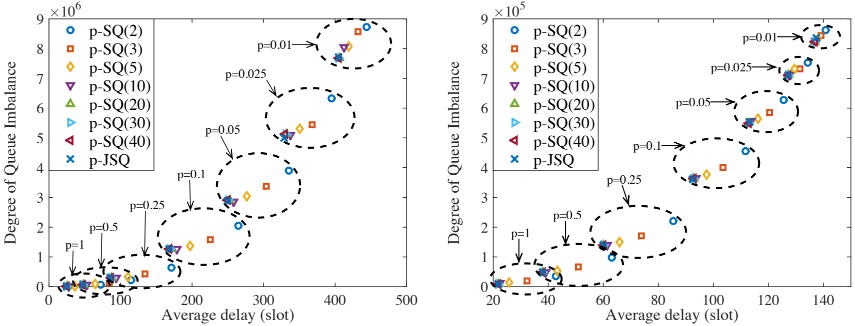
$$D_{\text{avg}}^{(\epsilon)} \leq \frac{\zeta^{(\epsilon)}}{2\lambda_{\Sigma}^{(\epsilon)}} \cdot \frac{1}{\epsilon} + \frac{M}{\lambda_{\Sigma}^{(\epsilon)}} \cdot \sqrt{\frac{\text{Degree of Queue Imbalance}}{\epsilon}},$$



(a) $N = 10, \ \rho = 0.95$ (b) $N = 10, \ \rho = 0.85$

Degree of queue imbalance VS. Delay ($N = 50$)

$$D_{\text{avg}}^{(\epsilon)} \leq \frac{\zeta^{(\epsilon)}}{2\lambda_{\Sigma}^{(\epsilon)}} \cdot \frac{1}{\epsilon} + \frac{M}{\lambda_{\Sigma}^{(\epsilon)}} \cdot \sqrt{\frac{\text{Degree of Queue Imbalance}}{\epsilon}},$$



(a) $N = 50,\ \rho = 0.95$ (b) $N = 50,\ \rho = 0.85$

What we have shown...

Question: Can we characterize the difference and differentiate 'optimal' policies?

Answer: Yes!

What we have shown...

Question: Can we characterize the difference and differentiate ‘optimal’ policies?

Answer: Yes!

- The solution is degree of queue imbalance.
  - instead of looking at the *sum queue lengths*.
  - it turns to look at the *expected queue-length difference*.
  - it can reflect the degree of dispatching preference.

Well...I want to learn some techniques!

Upper bound

Proposition

Under any load balancing scheme satisfying LDPC, the degree of queue imbalance is upper bounded by

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|^2\right] \leq \frac{1}{\left\|\widetilde{\Delta}\right\|_1^2} M_1,$$

where M_1 is some constant.

Upper bound

Proposition

Under any load balancing scheme satisfying LDPC, the degree of queue imbalance is upper bounded by

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|^2\right] \leq \frac{1}{\left\|\widetilde{\Delta}\right\|_1^2} M_1,$$

where M_1 is some constant.

Note:

Upper bound

Proposition

Under any load balancing scheme satisfying LDPC, the degree of queue imbalance is upper bounded by

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|^2\right] \leq \frac{1}{\left\|\widetilde{\Delta}\right\|_1^2} M_1,$$

where M_1 is some constant.

Note:

- key idea is still Hajek's Lemma: moments bound from drift.

Upper bound

Proposition

Under any load balancing scheme satisfying LDPC, the degree of queue imbalance is upper bounded by

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\| \overline{\mathbf{Q}}_{\perp}^{(\epsilon)} \right\|^2 \right] \leq \frac{1}{\left\| \widetilde{\Delta} \right\|_1^2} M_1,$$

where M_1 is some constant.

Note:

- key idea is still Hajek's Lemma: moments bound from drift.
- Some tricks need to extract the term $\left\| \widetilde{\Delta} \right\|_1$.

Upper bound

Proposition

Under any load balancing scheme satisfying LDPC, the degree of queue imbalance is upper bounded by

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|^2\right] \leq \frac{1}{\left\|\widetilde{\Delta}\right\|_1^2} M_1,$$

where M_1 is some constant.

Note:

- key idea is still Hajek's Lemma: moments bound from drift.
- Some tricks need to extract the term $\left\|\widetilde{\Delta}\right\|_1$.
- Hence it directly characterizes the impact of different schemes on the upper bound.

Lower bound

Proposition

Under any load balancing scheme satisfying LDPC, the degree of queue imbalance is lower bounded by

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|^2\right] \geq \frac{1}{\left\|\widetilde{\Delta}\right\|_1^2} M_2,$$

where M_2 is some constant.

Lower bound

Proposition

Under any load balancing scheme satisfying LDPC, the degree of queue imbalance is lower bounded by

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|^2\right] \geq \frac{1}{\left\|\widetilde{\Delta}\right\|_1^2} M_2,$$

where M_2 is some constant.

Note:

Lower bound

Proposition

Under any load balancing scheme satisfying LDPC, the degree of queue imbalance is lower bounded by

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|^2\right] \geq \frac{1}{\left\|\widetilde{\Delta}\right\|_1^2} M_2,$$

where M_2 is some constant.

Note:

- The same result holds for general ‘optimal’ schemes as well.

Lower bound

Proposition

Under any load balancing scheme satisfying LDPC, the degree of queue imbalance is lower bounded by

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|^2\right] \geq \frac{1}{\left\|\widetilde{\Delta}\right\|_1^2} M_2,$$

where M_2 is some constant.

Note:

- The same result holds for general 'optimal' schemes as well.
- The 'moment bounds from drift' method fails.

Lower bound

Proposition

Under any load balancing scheme satisfying LDPC, the degree of queue imbalance is lower bounded by

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}\right\|^2\right] \geq \frac{1}{\left\|\widetilde{\Delta}\right\|_1^2} M_2,$$

where M_2 is some constant.

Note:

- The same result holds for general 'optimal' schemes as well.
- The 'moment bounds from drift' method fails.
- We solve this with a novel Lyapunov function.

Universal equality

Universal equality

1. Consider the Lyapunov function: $V(\mathbf{Q}) \triangleq \sum_{i=1}^{N} \sum_{j>i}^{N} (Q_i - Q_j)^2$.

Universal equality

1. Consider the Lyapunov function: $V(\mathbf{Q}) \triangleq \sum_{i=1}^{N} \sum_{j>i}^{N} \left(Q_i - Q_j\right)^2$.
2. Setting mean drift to zero at steady-state:

$$\mathcal{B}^{(\epsilon)} := 2\mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1 \left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] = \mathcal{T}_1^{(\epsilon)} - \mathcal{T}_2^{(\epsilon)} + \mathcal{T}_3^{(\epsilon)},$$

where

Universal equality

1. Consider the Lyapunov function: $V(\mathbf{Q}) \triangleq \sum_{i=1}^{N} \sum_{j>i}^{N} \left(Q_i - Q_j\right)^2$.
2. Setting mean drift to zero at steady-state:

$$\mathcal{B}^{(\epsilon)} := 2\mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1 \left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] = \mathcal{T}_1^{(\epsilon)} - \mathcal{T}_2^{(\epsilon)} + \mathcal{T}_3^{(\epsilon)},$$

where

$$\mathcal{B}^{(\epsilon)} \rightarrow 0 \quad \text{(optimality)}$$

Universal equality

1. Consider the Lyapunov function: $V(\mathbf{Q}) \triangleq \sum_{i=1}^{N} \sum_{j>i}^{N} (Q_i - Q_j)^2$.
2. Setting mean drift to zero at steady-state:

$$\mathcal{B}^{(\epsilon)} := 2\mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1 \left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] = \mathcal{T}_1^{(\epsilon)} - \mathcal{T}_2^{(\epsilon)} + \mathcal{T}_3^{(\epsilon)},$$

where

$$\mathcal{B}^{(\epsilon)} \rightarrow 0 \quad \text{(optimality)}$$

$$\mathcal{T}_2^{(\epsilon)} := \sum_{i=1}^{N} \sum_{j>i}^{N} \mathbb{E}\left[\left(\overline{U}_i^{(\epsilon)} - \overline{U}_j^{(\epsilon)}\right)^2\right] \rightarrow 0$$

Universal equality

1. Consider the Lyapunov function: $V(\mathbf{Q}) \triangleq \sum_{i=1}^{N} \sum_{j>i}^{N} (Q_i - Q_j)^2$.
2. Setting mean drift to zero at steady-state:

$$\mathcal{B}^{(\epsilon)} := 2\mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1 \left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] = \mathcal{T}_1^{(\epsilon)} - \mathcal{T}_2^{(\epsilon)} + \mathcal{T}_3^{(\epsilon)},$$

where

$$\mathcal{B}^{(\epsilon)} \rightarrow 0 \quad \text{(optimality)}$$

$$\mathcal{T}_2^{(\epsilon)} := \sum_{i=1}^{N} \sum_{j>i}^{N} \mathbb{E}\left[\left(\overline{U}_i^{(\epsilon)} - \overline{U}_j^{(\epsilon)}\right)^2\right] \rightarrow 0$$

$$\mathcal{T}_3^{(\epsilon)} := \sum_{i=1}^{N} \sum_{j>i}^{N} \mathbb{E}\left[\left(\overline{A}_i^{(\epsilon)} - \overline{A}_j^{(\epsilon)} - \overline{S}_i^{(\epsilon)} + \overline{S}_j^{(\epsilon)}\right)^2\right] \rightarrow K$$

Universal equality

1. Consider the Lyapunov function: $V(\mathbf{Q}) \triangleq \sum_{i=1}^{N} \sum_{j>i}^{N} (Q_i - Q_j)^2$.
2. Setting mean drift to zero at steady-state:

$$\mathcal{B}^{(\epsilon)} := 2\mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1 \left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] = \mathcal{T}_1^{(\epsilon)} - \mathcal{T}_2^{(\epsilon)} + \mathcal{T}_3^{(\epsilon)},$$

where

$$\mathcal{B}^{(\epsilon)} \rightarrow 0 \quad \text{(optimality)}$$

$$\mathcal{T}_2^{(\epsilon)} := \sum_{i=1}^{N} \sum_{j>i}^{N} \mathbb{E}\left[\left(\overline{U}_i^{(\epsilon)} - \overline{U}_j^{(\epsilon)}\right)^2\right] \rightarrow 0$$

$$\mathcal{T}_3^{(\epsilon)} := \sum_{i=1}^{N} \sum_{j>i}^{N} \mathbb{E}\left[\left(\overline{A}_i^{(\epsilon)} - \overline{A}_j^{(\epsilon)} - \overline{S}_i^{(\epsilon)} + \overline{S}_j^{(\epsilon)}\right)^2\right] \rightarrow K$$

$$\mathcal{T}_1^{(\epsilon)} := 2\lambda_\Sigma^{(\epsilon)} N \mathbb{E}\left[\langle \overline{\mathbf{Q}}_\perp^{(\epsilon)}, \widetilde{\Delta} \rangle\right]$$

Universal equality

1. Consider the Lyapunov function: $V(\mathbf{Q}) \triangleq \sum_{i=1}^{N} \sum_{j>i}^{N} (Q_i - Q_j)^2$.
2. Setting mean drift to zero at steady-state:

$$\mathcal{B}^{(\epsilon)} := 2\mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1 \left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] = \mathcal{T}_1^{(\epsilon)} - \mathcal{T}_2^{(\epsilon)} + \mathcal{T}_3^{(\epsilon)},$$

where

$$\mathcal{B}^{(\epsilon)} \rightarrow 0 \quad \text{(optimality)}$$

$$\mathcal{T}_2^{(\epsilon)} := \sum_{i=1}^{N} \sum_{j>i}^{N} \mathbb{E}\left[\left(\overline{U}_i^{(\epsilon)} - \overline{U}_j^{(\epsilon)}\right)^2\right] \rightarrow 0$$

$$\mathcal{T}_3^{(\epsilon)} := \sum_{i=1}^{N} \sum_{j>i}^{N} \mathbb{E}\left[\left(\overline{A}_i^{(\epsilon)} - \overline{A}_j^{(\epsilon)} - \overline{S}_i^{(\epsilon)} + \overline{S}_j^{(\epsilon)}\right)^2\right] \rightarrow K$$

$$\mathcal{T}_1^{(\epsilon)} := 2\lambda_\Sigma^{(\epsilon)} N \mathbb{E}\left[\langle \overline{\mathbf{Q}}_\perp^{(\epsilon)}, \widetilde{\Delta} \rangle\right]$$

3. Thus, we have $\lim_{\epsilon \downarrow 0} 2\mu_\Sigma N \mathbb{E}\left[\langle \overline{\mathbf{Q}}_\perp^{(\epsilon)}, \widetilde{\Delta} \rangle\right] = -K$.

In summary, we try to go beyond heavy-traffic optimality:

1. we show that HT-optimality is coarse: it contains policies that can be arbitrarily close to Random.
   - A weak condition such as LDPC is enough.
   - As a result, slight preference in the long-term implies optimality.

In summary, we try to go beyond heavy-traffic optimality:

1. we show that HT-optimality is coarse: it contains policies that can be arbitrarily close to Random.
   - A weak condition such as LDPC is enough.
   - As a result, slight preference in the long-term implies optimality.

2. we propose a new metric Degree of Queue Imbalance, which can differentiate between good and poor policies.
   - Look at the queue-length difference among servers.
   - The closer...The worse...

Wait... one more question, how about general case?

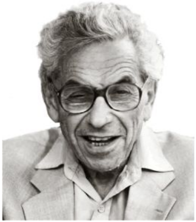
Wait... one more question, how about general case?

Have you heard *Erdős Number*?

‘Perfect Death’

“....I finish up an important theorem... Then someone in the audience shouts out, ‘*What about the general case?*’

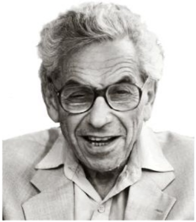
— Paul Erdős



‘Perfect Death’

“....I finish up an important theorem... Then someone in the audience shouts out, ‘*What about the general case?*’ I’ll turn to the audience and smile, say ‘I’ll leave that to the next generation,’ and then I’ll *keel over*. ”

— Paul Erdős



For heterogeneous servers:

Main results established before still hold in a weaker sense under some mild additional conditions.

Thank you!

Backup

- Can we generalize this method to other scenarios?
- Is the LDPC condition necessary for optimality?
- Sometimes, a perfect balance of queue lengths may not be good.

Here is the intuition...

Here is the intuition...

1. A sufficient and necessary condition:

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1 \left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] = 0.$$

   - $U(t)$ is the unused service due to empty queue.

Here is the intuition...

1. A sufficient and necessary condition:

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1 \left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] = 0.$$

   - $U(t)$ is the unused service due to empty queue.

2. The condition can be upper bounded by

$$\mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1 \left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] \leq N\sqrt{C\epsilon\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}(t)\right\|^2\right]}.$$

Here is the intuition...

1. A sufficient and necessary condition:

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1 \left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] = 0.$$

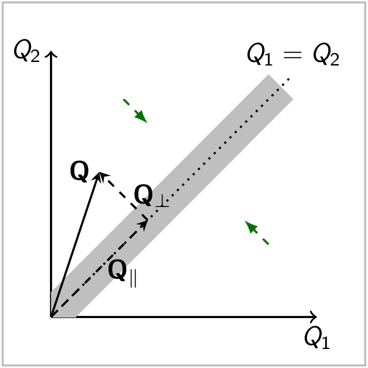
   - $U(t)$ is the unused service due to empty queue.

2. The condition can be upper bounded by

$$\mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}(t+1)\right\|_1 \left\|\overline{\mathbf{U}}^{(\epsilon)}(t)\right\|_1\right] \leq N\sqrt{C\epsilon\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp}^{(\epsilon)}(t)\right\|^2\right]}.$$

3. The moment term is upper bounded by a constant under LDPC
   - Lyapunov drift
   - T-step Hajek's Lemma

More on moments bound



- The drift  is indicated by

$$\mathbb{E}\left[\langle \mathbf{Q}_\perp, \mathbf{A} - \mathbf{S}\rangle \mid \mathbf{Q}\right].$$

- for each t, it can be either positive or negative.
- but, under LDPC, there exists finite T

$$\sum_{t=t_0}^{t_0+T-1} \mathbb{E}\left[\langle \mathbf{Q}_\perp, \mathbf{A} - \mathbf{S}\rangle \mid \mathbf{Q}(t_0)\right] \approx -\delta \left\|\mathbf{Q}_\perp\right\|$$

- that is, long term drift  is positive.