Load Balancing in Heavy-traffic Regime: Theory to Algorithms

Xingyu Zhou, Fei Wu, Jian Tan, Yin Sun, Kannan Srinivasan and Ness Shroff

Research Problems

We are interested in the following three problems regarding load balancing in heavy-traffic regime.

Can we go beyond...?

1. the previous 'optimal' policies.
2. the single dimensional state-space collapse.
3. the heavy-traffic delay optimality.

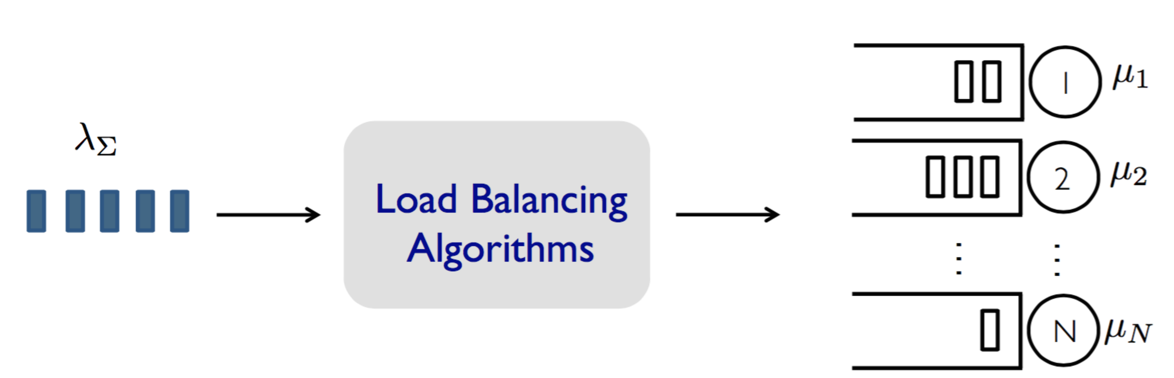
Our contributions: we provide the answers to all the three questions above.

1. ***Beyond previous 'optimal' policies?*** Yes!
   - we identify a class of 'optimal' policies.
   - we prove that JIQ is not 'optimal'.
   - we design a new pull-based policy JBT-d, which is 'optimal' while enjoying all the nice features of JIQ.
2. ***Beyond single dimensional state-space collapse?*** Yes!
   - we prove that 'optimality' holds even under multi-dimensional state-space collapse.
   - it allows us to design new flexible optimal policies.
3. ***Beyond heavy-traffic delay optimality?*** Yes!
   - we show that HT-optimality is coarse: it contains policies that can be arbitrarily close to random routing.
   - we propose a new metric Degree of Queue Imbalance, which can differentiate between good and poor policies.

Related Works

- Eryilmaz, et al. [1]: the Lyapunov-drift based framework and the HT-optimality of JSQ and MaxWeight.
- Maguluri, et al. [2] and [3]: HT-optimality of power-of-d and optimal queue-length scaling of MaxWeight in switch systems under multi-dimensional collapse.
- Wang, et al. [4]: HT-optimality of JSQ-MaxWeight in MapReduce.

Optimality Definition



Throughput Optimal: It can stabilize the system for any arrival rate in capacity region with all the moments bounded, i.e, for any $\epsilon > 0$ where $\epsilon = \sum \mu_n - \lambda_\Sigma$.

Heavy-traffic Delay Optimal: It can achieve the lower bound on delay when $\epsilon \to 0$, that is, $\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\sum Q_n\right] = \lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[q\right]$, where q is the queue length of the single-server resource pooled system.

Important Notions

Dispatching distribution $\mathbf{P}(t)$: The nth component of $\mathbf{P}(t)$ is the probability of dispatching arrival to the nth shortest queue at time-slot t.

Dispatching Preference $\Delta(t)$: $\Delta(t) = \mathbf{P}(t) - \mathbf{P}_{\text{rand}}(t)$, where $\mathbf{P}_{\text{rand}}(t)$ is the $\mathbf{P}(t)$ under (proportional) random routing.

Tilted distribution: A $\mathbf{P}(t)$ is tilted if, for some $2 \leq k \leq N$

- $\Delta_n(t) \geq 0$ for all $n < k$.
- $\Delta_n(t) \leq 0$ for all $n \geq k$

δ-tilted distribution: A $\mathbf{P}(t)$ is δ-tilted if

- $\Delta_n(t)$ is tilted.
- $\Delta_1(t) \geq \delta$, $\Delta_N(t) \leq -\delta$

A class of policies Π : A load balancing algorithm is in Π if

- $\mathbf{P}(t)$ is tilted for any t.
- every T time-slots, there exits a slot t' such that $\mathbf{P}(t')$ is δ-tilted.

Long-term Dispatching Preference Condition (LDPC): $\widetilde{\Delta}_1 \geq \widetilde{\Delta}_2 \geq \ldots \geq \widetilde{\Delta}_N$ and $\widetilde{\Delta}_1 \neq \widetilde{\Delta}_N$, where $\widetilde{\Delta} \triangleq \mathbb{E}\left[\overline{\Delta}\right]$ and $\overline{\Delta}$ is a random vector which is equal in distribution to $\Delta(t)$ in steady state.

Main Result: Beyond Previous Optimal Policies [5]

Question: Can a policy enjoys optimality, low message overhead and zero dispatching delay at the same time?

The solution is the new JBT-d algorithm:

1. every T time-slots, randomly sample d servers and take the minimum queue length as threshold.
2. each server report its ID when its queue length is not larger than the threshold for the first time.
3. if possible, randomly picks a ID and join the server.
4. otherwise, randomly picks a queue to join.

Note: If servers are heterogeneous, report μ and pick ID with proportional probability in steps 3 and 4.

Note: JIQ can be viewed as a static version of JBT-d with $T = \infty$ and $th = 0$. JIQ has low message overhead but it is not HT-optimal even for homogeneous servers, but our JBT-d is.

Theorem: JIQ is not heavy-traffic delay optimal even in homogeneous servers.

In contrast...
Theorem: For any finite T and $d \geq 1$, JBT-d is throughput and heavy-traffic delay optimal.

Actually, JBT-d belongs to the optimal class Π:
Theorem: Any policy in the class Π is throughput and heavy-traffic delay optimal.

Note: JSQ and Power-of-d are both in the class Π.

Main Result: Beyond Single Dimensional Collapse [6]

Question: Can a policy be optimal under multi-dimensional state-space collapse, and if so, how can we achieve it?

A Polyhedral Cone $\mathcal{K}_\alpha$:

$$\mathcal{K}_\alpha = \{\mathbf{x} \in \mathbb{R}^N : \mathbf{x} = \sum w_n \mathbf{b}^{(n)}, w_n \geq 0 \text{ for all } n \in \mathcal{N}\}$$

where the nth component of $\mathbf{b}^{(n)}$ is 1 and α everywhere else for some $\alpha \in [0, 1]$.

HT-optimality still holds under multi-dimensional collapse:
Theorem: Given a throughput optimal policy, if there exists an $\alpha \in (0, 1]$ such that the state-space collapses to the cone $\mathcal{K}_\alpha$, then this policy is Heavy-traffic optimal.

We can achieve it in the following flexible way:
Theorem: Given a throughput optimal policy, if there exists $\mathcal{K}_{\alpha^{(\epsilon)}}$ such that for all $\mathbf{Q}(t) \notin \mathcal{K}_{\alpha^{(\epsilon)}}$, $\mathbf{P}(t)$ is δ-tilted with parameter $\delta^{(\epsilon)}$. And $\alpha^{(\epsilon)}\delta^{(\epsilon)} = \Omega(\epsilon^\beta)$ for some $\beta \in [0, 1)$, then this policy is HT-optimal.

Main Result: Beyond Heavy-traffic Optimality [7]

Question: How large is the difference of empirical delay among HT-optimal policies and how can we differentiate it?

Huge difference: the empirical delay performance of HT-optimal can range from very good (JSQ) to very bad (arbitrarily close to random routing)
Theorem: Any policy satisfying the LDPC is HT-optimal.

Degree of Queue Imbalance: The degree of queue imbalance of a 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 $\mathbf{Q}_\perp \triangleq \mathbf{Q}(t) - \langle \mathbf{Q}, \mathbf{1}_N \rangle \mathbf{1}_N$.

We can differentiate it with new metric *Degree of Queue Imbalance:*
Theorem: For any policy satisfying LDPC, the degree of queue imbalance is on the order of

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

References

[1] Atilla Eryilmaz and R Srikant. Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems*, 72(3-4):311–359, 2012.

[2] Siva Theja Maguluri, R Srikant, and Lei Ying. Heavy traffic optimal resource allocation algorithms for cloud computing clusters. *Performance Evaluation*, 81:20–39, 2014.

[3] Siva Theja Maguluri, Sai Kiran Burle, and R Srikant. Optimal heavy-traffic queue length scaling in an incompletely saturated switch. *Queueing Systems*, 88(3-4):279–309, 2018.

[4] Weina Wang, Kai Zhu, Lei Ying, Jian Tan, and Li Zhang. Maptask scheduling in mapreduce with data locality: Throughput and heavy-traffic optimality. *IEEE/ACM Transactions on Networking (TON)*, 24(1):190–203, 2016.

[5] Xingyu Zhou, Fei Wu, Jian Tan, Yin Sun, and Ness Shroff. Designing low-complexity heavy-traffic delay-optimal load balancing schemes: Theory to algorithms. *Proc. ACM Meas. Anal. Comput. Syst.*, 1(2):39:1–39:30, December 2017.

[6] Xingyu Zhou, Jian Tan, and Ness Shroff. Flexible load balancing with multi-dimensional state-space collapse: Throughput and heavy-traffic delay optimality. *submitted.*

[7] Xingyu Zhou, Fei Wu, Jian Tan, Kannan Srinivasan, and Ness Shroff. Degree of queue imbalance: Overcoming the limitation of heavy-traffic delay optimality in load balancing systems. *Proc. ACM Meas. Anal. Comput. Syst.*, 2(1):21, 2018.