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 \( \sum n_{\alpha} - \lambda_{\alpha} \).

Heavy-traffic Delay Optimal: It can achieve the lower bound on delay when \( \epsilon \to 0 \), that is, \( \lim_{\epsilon \to 0} E \|Q\| \leq \lim_{\epsilon \to 0} E \|Q\|_{L_2} \) where \( Q \) is the queue length of the single-server resource pooled system.

Important Notions

\begin{itemize}
  \item \textbf{Dispatching distribution} \( P(t) \): The \( n \)-th component of \( P(t) \) is the probability of dispatching arrival to the \( n \)-th shortest queue at time-slot \( t \).
  \item \textbf{Dispatching Preference} \( \lambda(t) : \Delta(t) = P(t) - P_{\text{rand}}(t) \), where \( P_{\text{rand}}(t) \) is the \( P(t) \) under (proportional) random routing.
  \item \textbf{Tilted distribution}: A \( P(t) \) is \( \delta \)-tiled if 
    \( \Delta(t) > 0 \) for all \( n \), \( \Delta(t) \leq 0 \) for all \( n > k \).
  \item \textbf{\( \delta \)-tiled distribution}: A \( P(t) \) is \( \delta \)-tiled if 
    \( \Delta(t) \geq \delta, \Delta(t) \leq -\delta \).
\end{itemize}

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 \( \mu \) 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 \( d = 0 \). JIQ has low message overhead but it is not HT-optimal even for homogeneous servers, but our JBT-d is.

In contrast:

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

Main Result: Beyond Single Dimensional Collapse [6]

A Polyhedral Cone \( K_{\alpha} \):
\( K_{\alpha} = \{ x \in \mathbb{R}^N : x = \sum w_i b_i \alpha_i, w_i \geq 0 \text{ for all } n \in \mathbb{N} \} \)
where the \( n \)-th component of \( b_i \alpha_i \) is \( 1 + \alpha_i \) everywhere else for some \( \alpha_i \in [0, 1] \).

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 \( K_{\alpha} \), 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 \( K_{\alpha} \), such that for all \( \Delta(t) \in K_{\alpha} \), \( P(t) \) is \( \delta \)-tiled with parameter \( \delta \).

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 \( \mathbf{Q} \) is given by \( E \| \mathbf{Q} \|_{L_2} \), where \( Q_i = E \| Q_i \| - (Q_i)_{\mathbb{E}}/\lambda_i \).

We can differentiate it with new metric Degree of Queue Imbalance:

For any policy satisfying LDP, the degree of queue imbalance is on the order of \( \lim_{\epsilon \to 0} E \| Q_i \|_{L_2} = O(1/\Delta_0) \).

References