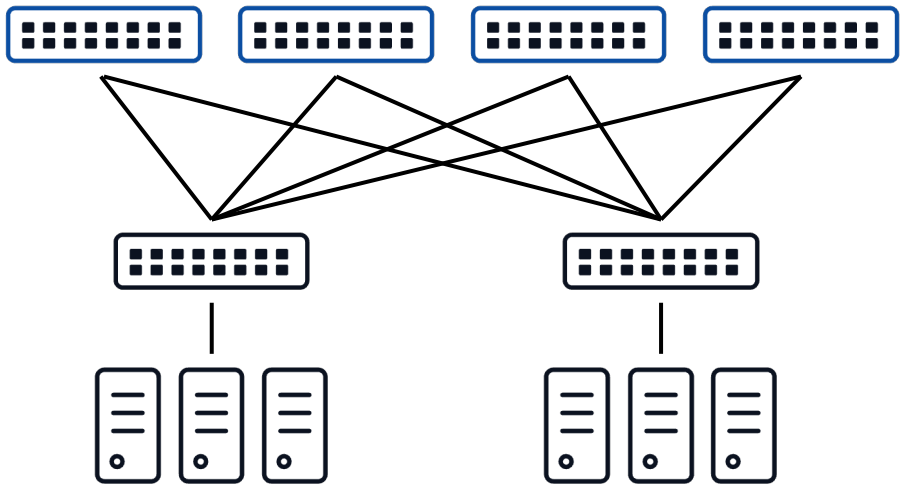
PIN: Less is More for RDMA Load Balancing

Jichun Wu[1], Ran Shu[2], Yongqiang Xiong[2], Andrew W. Moore[1]

[1] University of Cambridge, [2] Microsoft Research

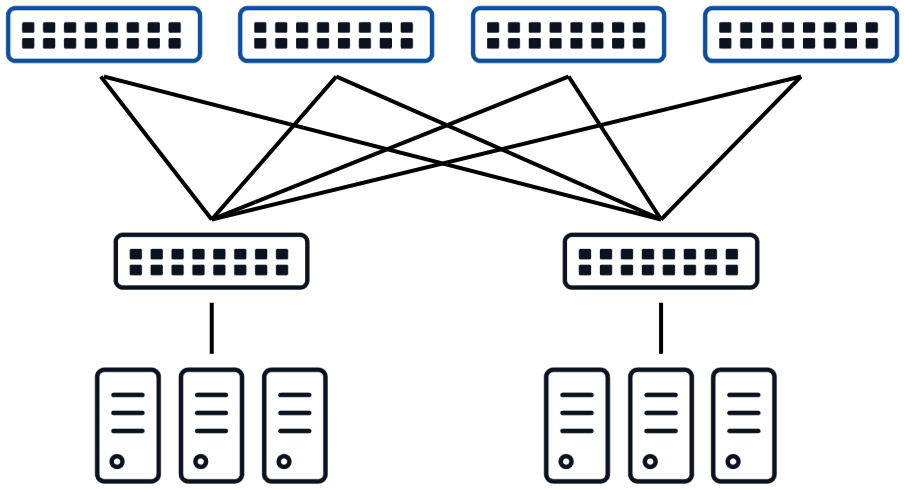
Multi-pathing

- Clos network
  - High bandwidth provided by multiple paths
- Equal Cost Multi-Path (ECMP)
  - Each flow / connection takes one path based on hash
  - Hash collisions cause contention

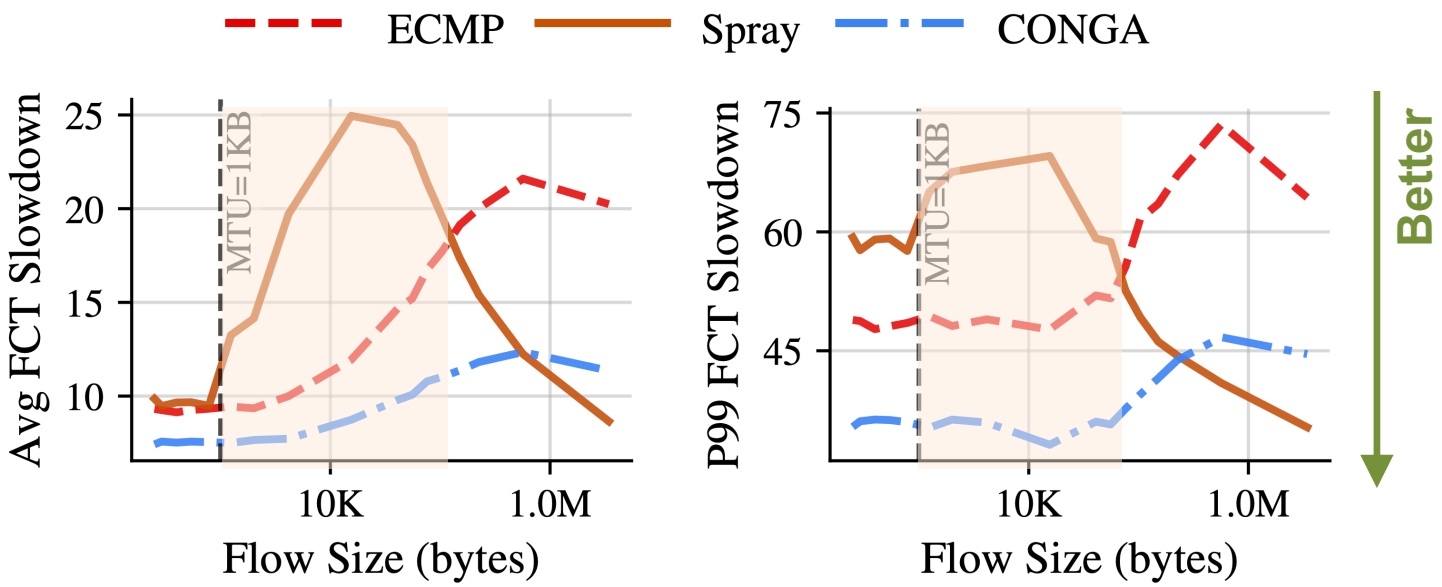


Spraying is attractive for RDMA

- Why spraying?
  - Avoid ECMP hash collisions
  - Exploit path diversity for large flows
- Why now?
  - Newer RDMA tolerate reordering

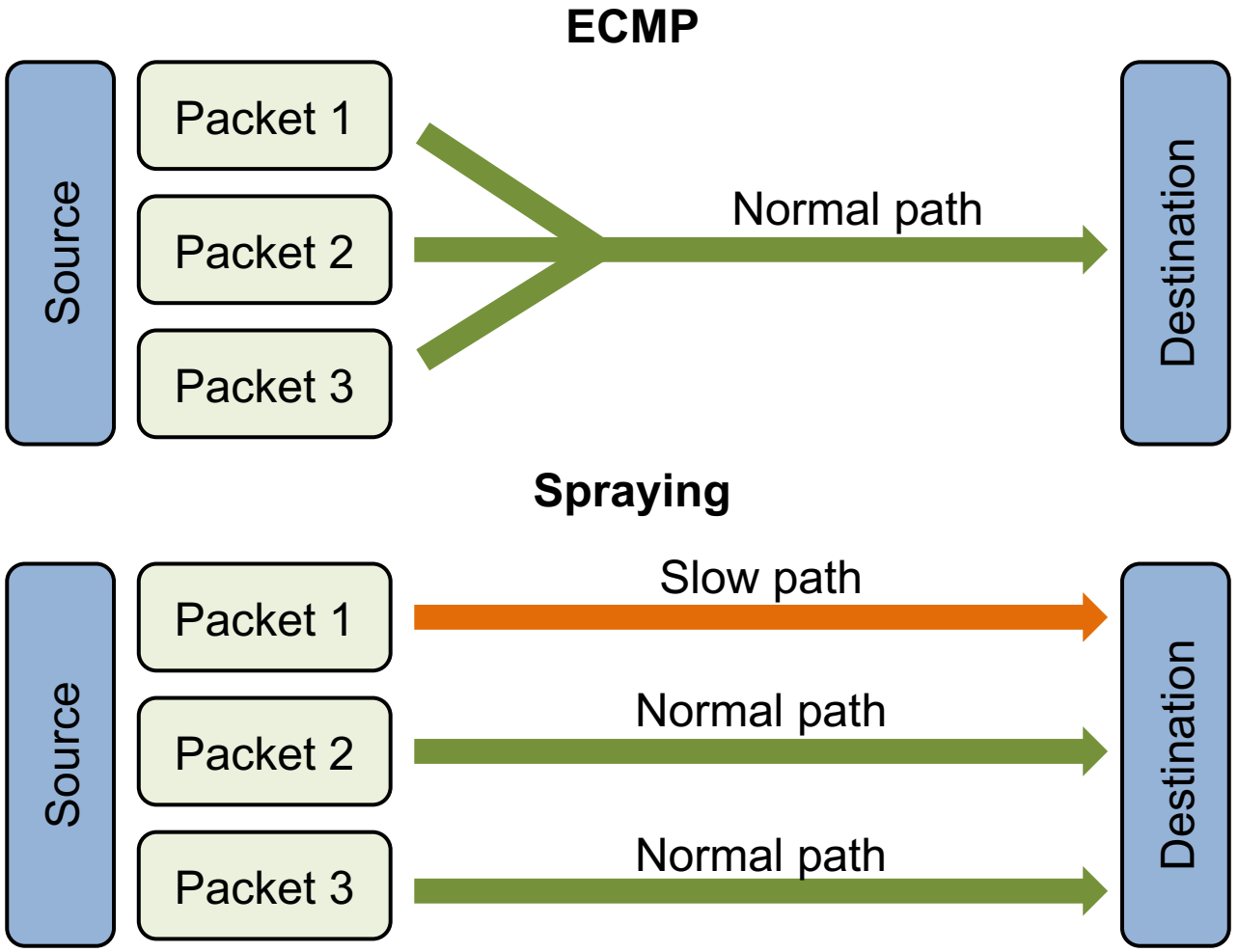


Observation



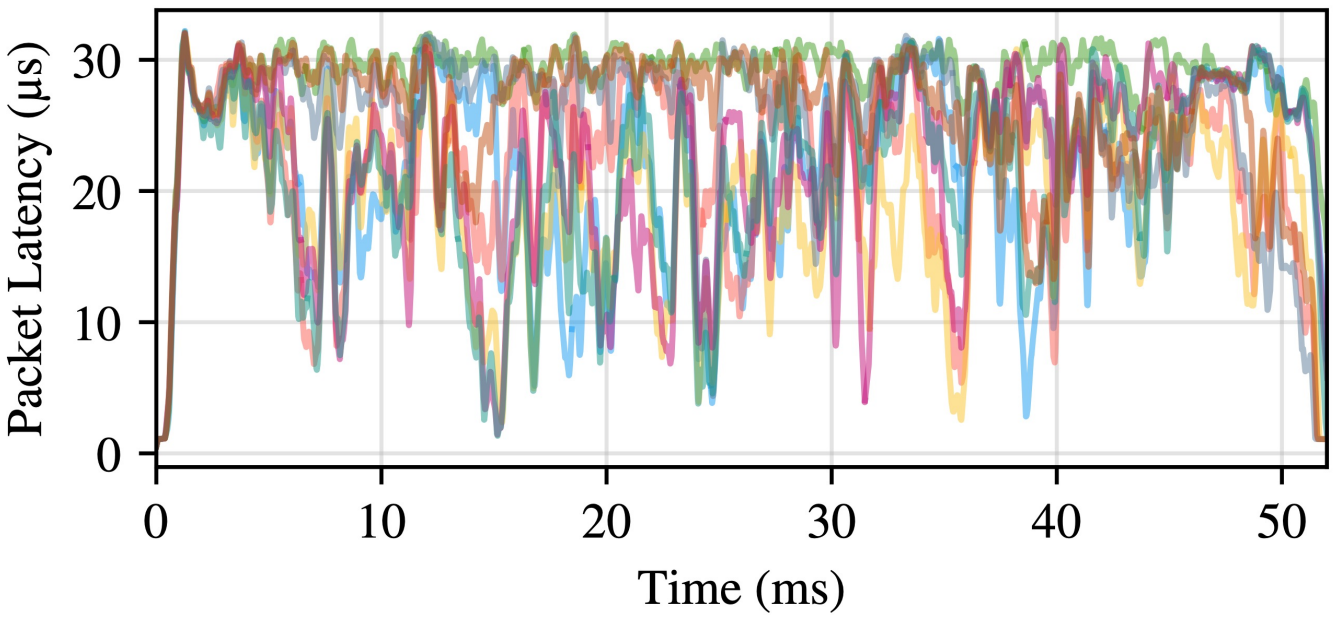
- Spraying does not benefit every flow
- Large flows benefit the most
- Short flows even worse than ECMP

Our guess



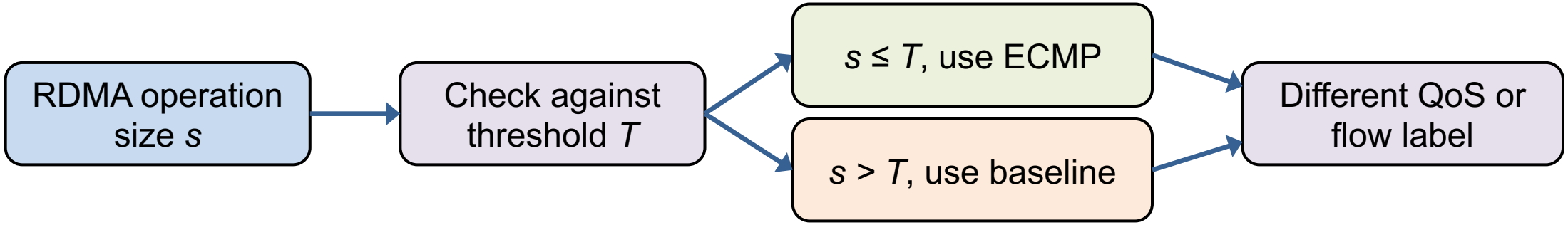
- Spraying has a larger chance of having one of the packet traverse a slower path
- For 3 packets, if half of the paths are slow:
  ECMP: $\frac{1}{2}$
  Spraying: $1 - (\frac{1}{2})^3 = \frac{7}{8}$

Path quality varies



- Spraying with 128 hosts connected by leaf-spine network
- Even healthy network experience link quality variation

PIN



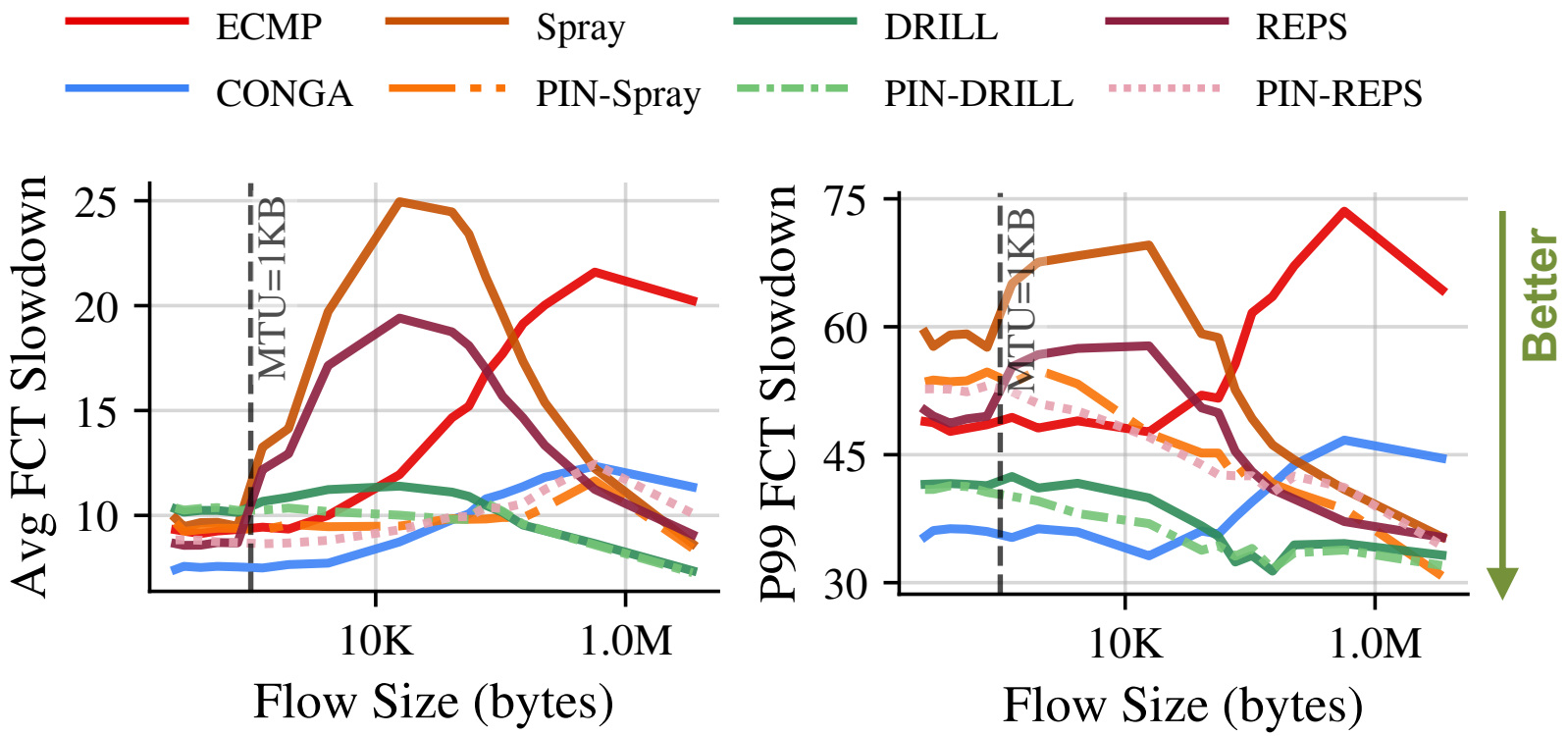
- Baseline: Spray, DRILL, REPS
- Large flows keep path diversity
- Small flows get stability

Evaluation

- Simulator: ns-3
- Workload: Alibaba Storage, Meta Hadoop
- Topology: 128-host leaf-spine, 100Gbps

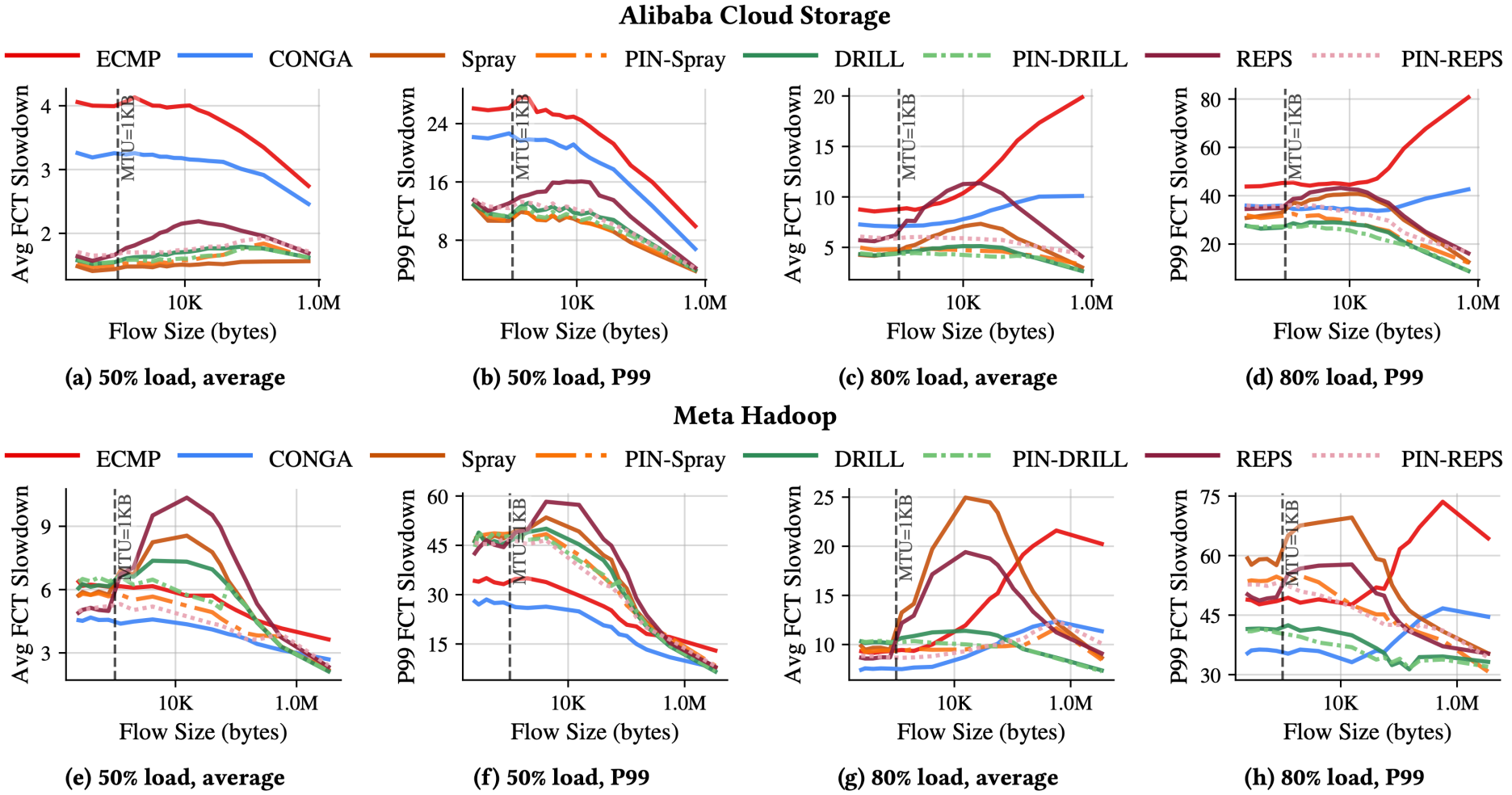
- Modified to allow reordering

Results (80% load, Meta Hadoop)



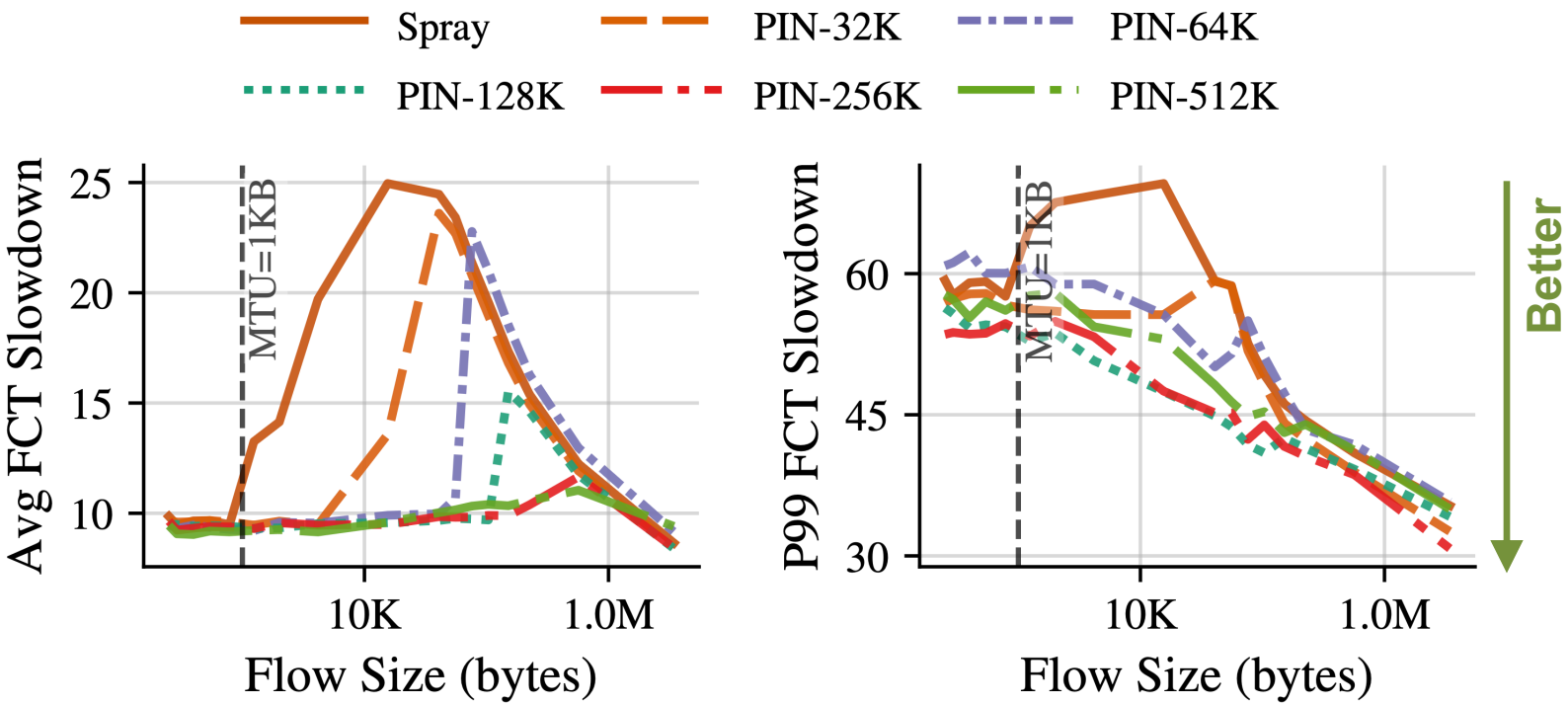
- PIN-Spray improves both avg and tail
- Other spraying LBs have similar benefits

Results



Threshold: 64KB for DRILL, 256KB for Spray and REPS

Threshold analysis



- Too small: leave part of the gap unsolved
- Too large: hurt global load balance like ECMP
- Choose a middle value, e.g., 256KB

Q&A

Takeaway:

- Fine-grained RDMA load balancing (e.g., spraying) is not uniformly beneficial
- Short flows need path stability
- PIN is a lightweight wrapper: pin the path for short flows, preserve baseline behavior for large flows