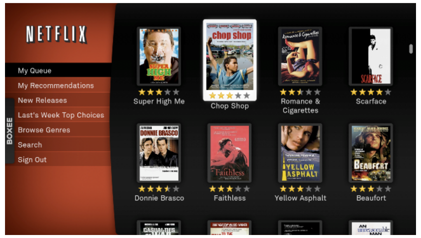
Scheduling Mix-flows in Commodity Datacenters with Karuna

Li Chen, Kai Chen, Wei Bai, Mohammad Alizadeh (MIT)
SING Group, CSE Department
Hong Kong University of Science and Technology

Datacenter Transport

- Deadline flows
  - Meeting deadlines
  - D3, D2TCP, …
- General (non-deadline) flows
  - Reduce flow completion time (FCT)
  - pFabric, PDQ, PASE, PIAS, …

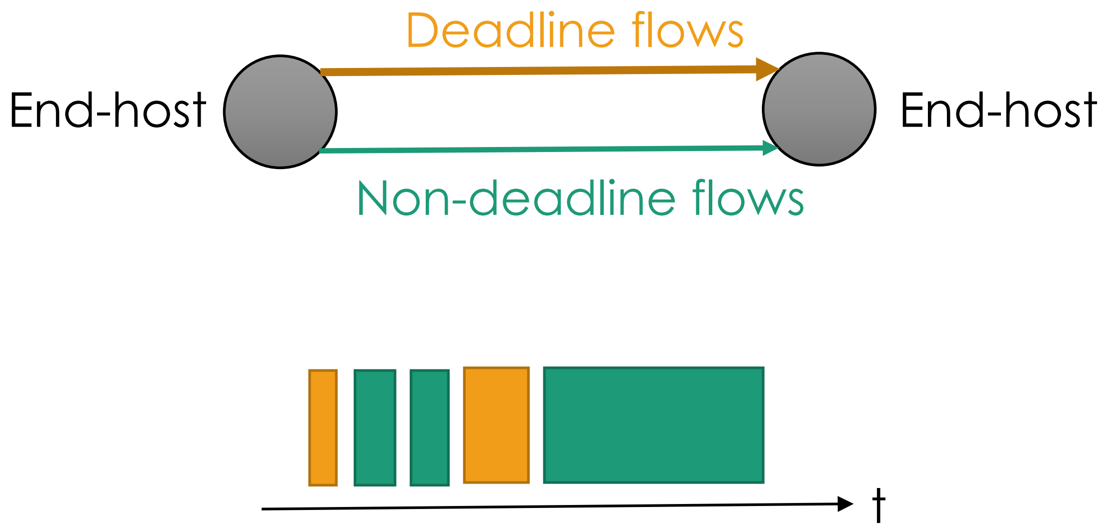


We investigate a practical, yet neglected, problem:

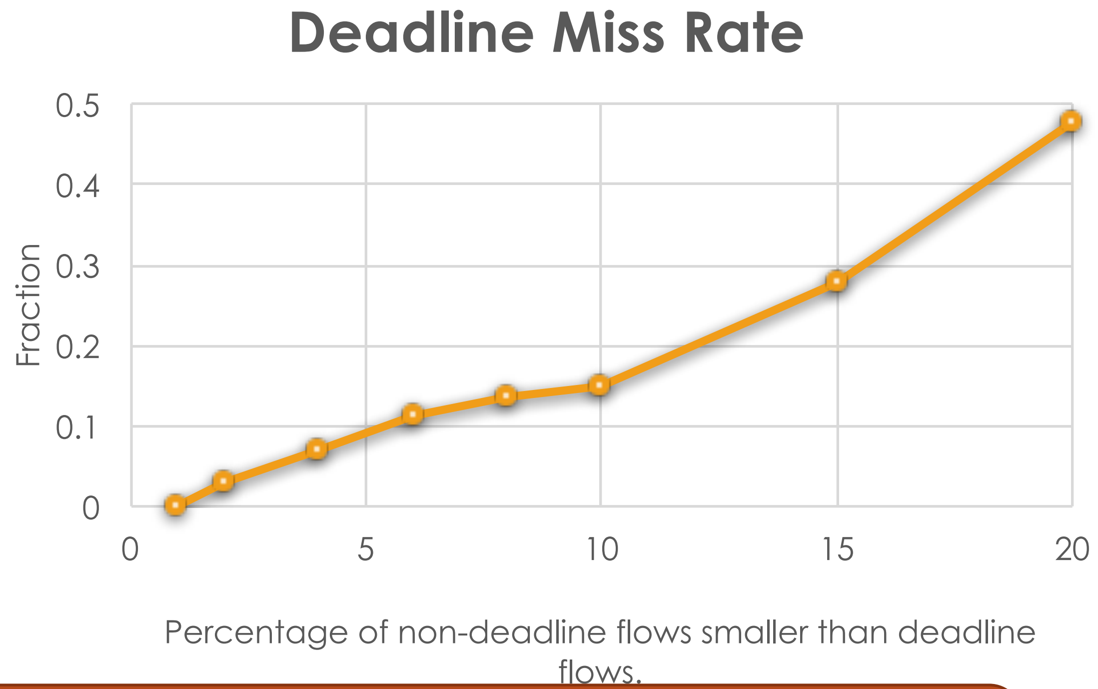
Mix-flow Scheduling

Prior solutions do not work for mix-flows

Shortest Job First (SJF) Scheduling – pFabric, PASE, PIAS, PDQ



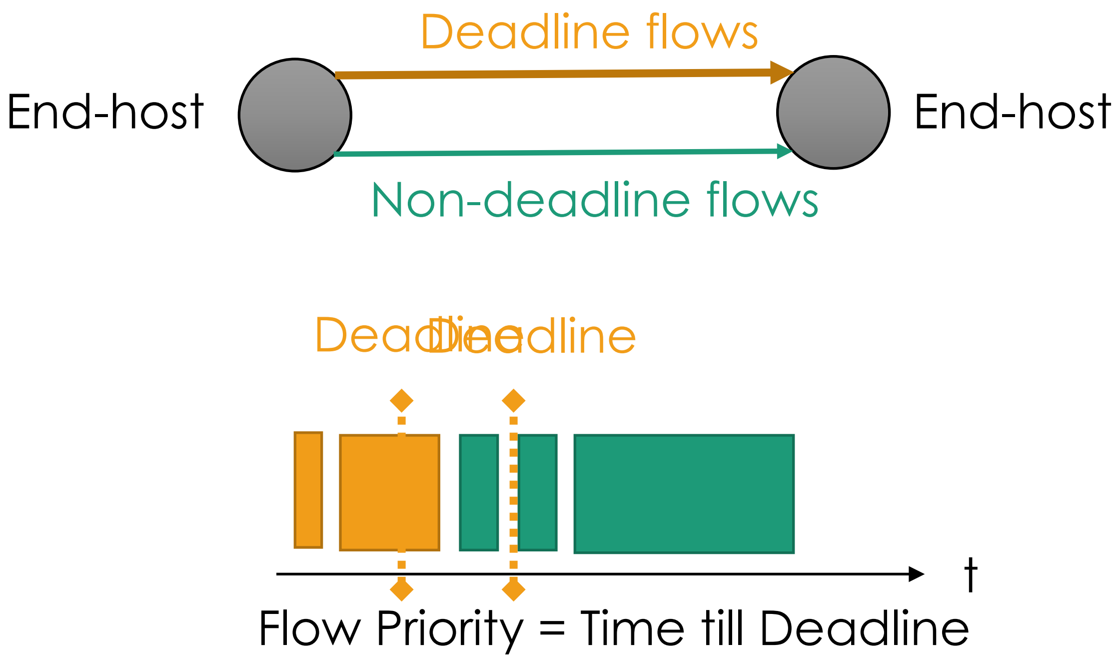
Flow Priority = Remaining size

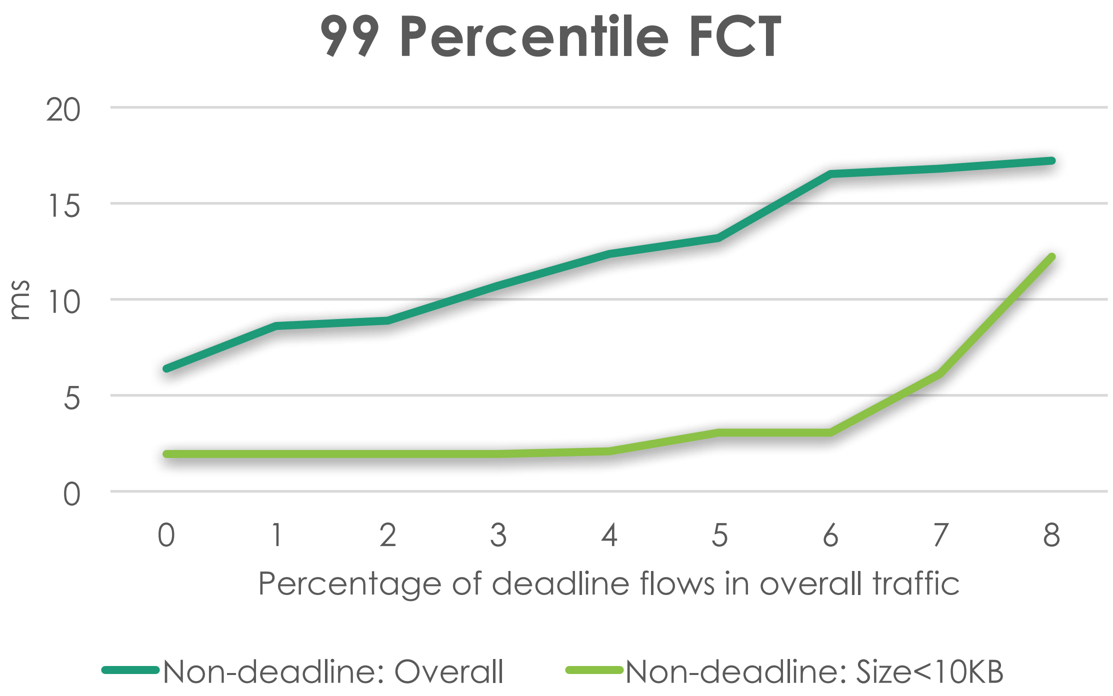


Scheduling only with sizes hurts deadline flows
Problem: unawareness of deadlines.

Prior solutions do not work for mix-flows

Earliest Deadline First Scheduling – pFabric, PASE, PIAS, PDQ





Prioritizing deadline flows hurts non-deadline flows, especially short ones.
Problem: Existing transports for deadline flows unnecessarily takes all bandwidth.

How to schedule mix-flows?

Deadline Flows

- Meet deadlines
- Flow deadline → Priority

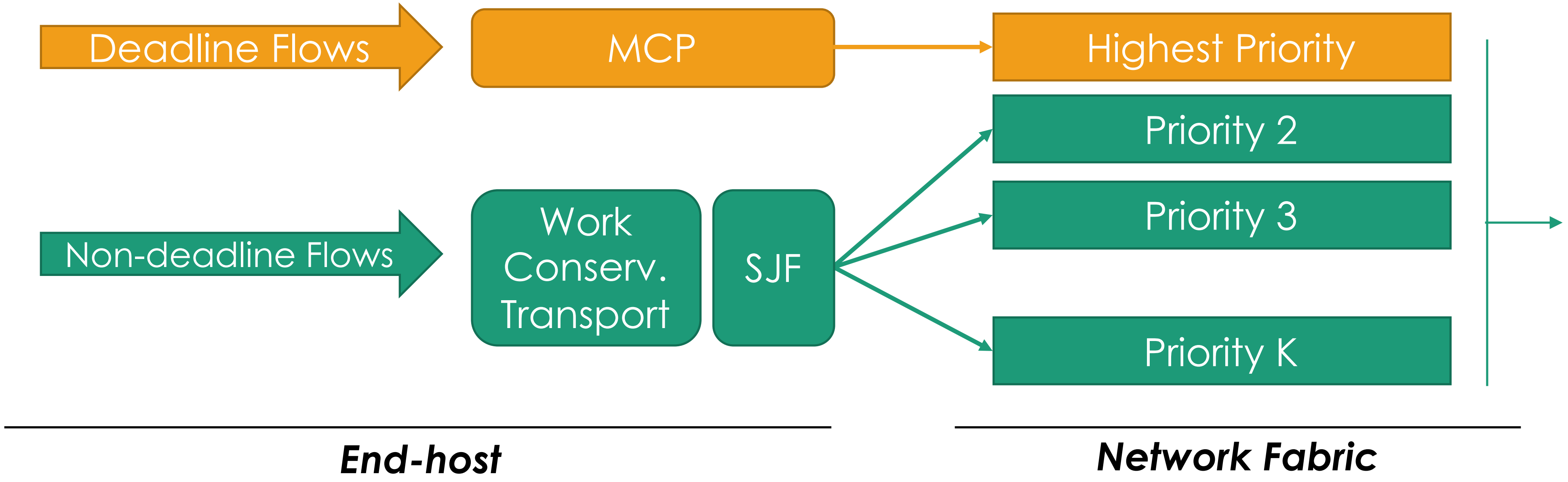
Non-deadline Flows

- Reduce FCT
- Flow Size → Priority

Karuna

Key Insight: Deadline flows should minimally impact non-deadline flows.

- Deadline flows
  - **High priority** with **minimal bandwidth** to complete just before deadlines.
- Non-deadline flows
  - **Low priority** but take **all available bandwidth** to reduce FCT.



MCP for deadline flows:

Completing deadlines with minimal bandwidth

Minimal-impact **C**ongestion control **P**rotocol

MCP: Formulation and solution

- Objective → Minimal impact
  - Per-packet latency
- Constraints:
  - Meet deadlines
  - Network capacity

$$P(\mathbf{y}(t)) = \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}\sum_s \{\sum_{l\in L(s)} d_l(y_l(t))\}$$

$$\min_{\mathbf{x}(t)} \sum_s \{V \sum_{l\in L(s)} d_l(y_l(t)) + \frac{Z_s(t)\gamma_s(t)}{x_s(t)} + \sum_{l\in L(s)} (Q_l(t)+\mu)x_s(t)\}$$

$Z_s($

$$\text{subject to } y_l(t) = \sum_{s\in S(l)} x_s(t), \forall l$$

$$W_s(t+\tau_s(t)) \leftarrow W_s(t) + \tau_s(t)(\Theta(\gamma_s(t), \frac{W_s(t)}{\tau_s(t)}) - \sum_{l\in L(s)} (Q_l(t)+\lambda_l(t)))$$

Stochastic Optimization → Lyapunov Optimization Framework [1] → Convex Optimization → Primal Solution → Per-flow congestion window update function

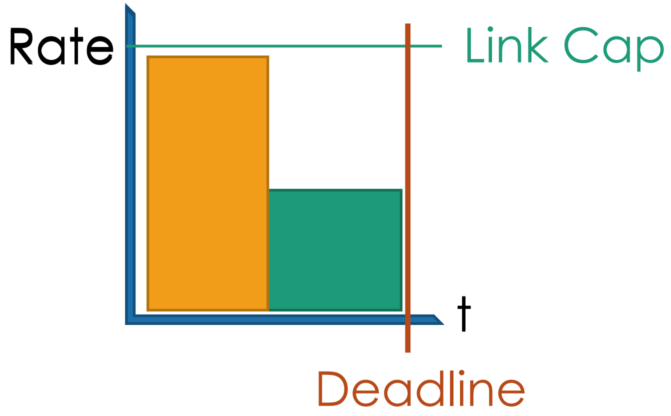
[1] M. J. Neely. *Stochastic Network Optimization with Application to Communication and Queueing Systems*, Morgan & Claypool, 2010.

MCP: Formulation and solution

- Objective → Minimal impact
  - Per-packet latency
- Constraints
  - Meet deadlines
  - Network capacity

- Solution

→ **Near-deadline completion**



Stochastic Optimization → Lyapunov Optimization Framework [1] → Convex Optimization → Primal Solution → Per-flow congestion window update function

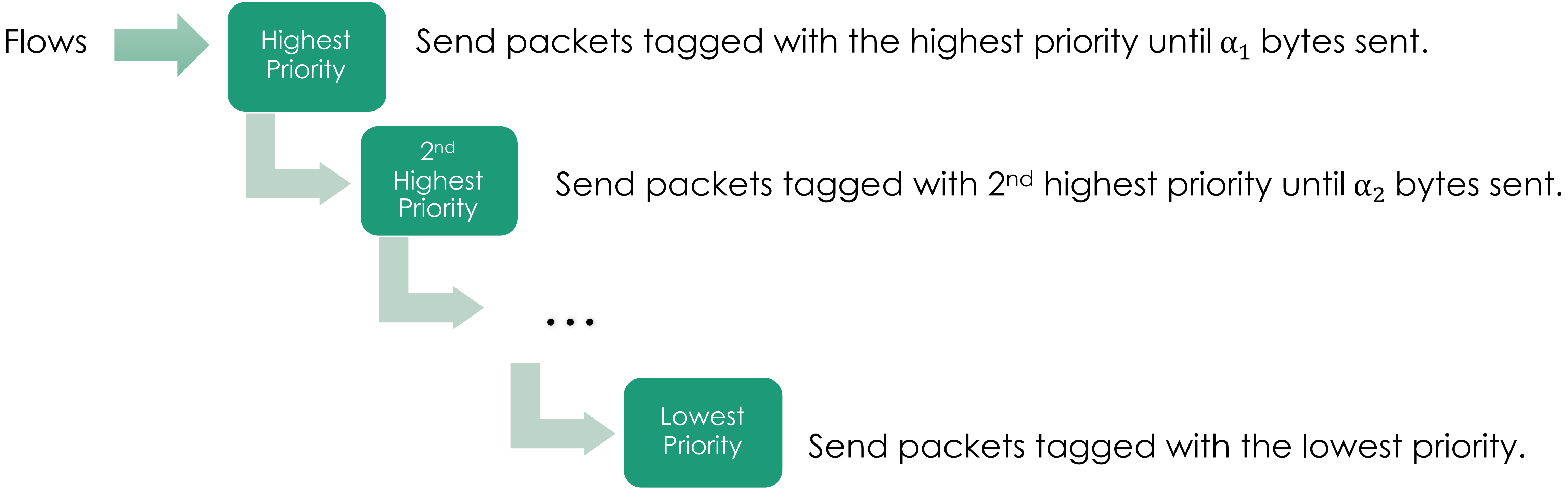
Reducing FCT for non-deadline flows

Mimicking SJF

Non-deadline flows with/out known sizes

Non-deadline flows with unknown size

- PIAS [2] is best known scheme.



[2] Wei Bai, et. al., *Information-Agnostic Flow Scheduling for Commodity Data Centers*, USENIX NSDI 2015

Karuna for non-deadline flows

- Non-deadline flows with unknown size ← PIAS
- Non-deadline flows with known size
  - Karuna extends PIAS to schedule flows with/out known sizes.

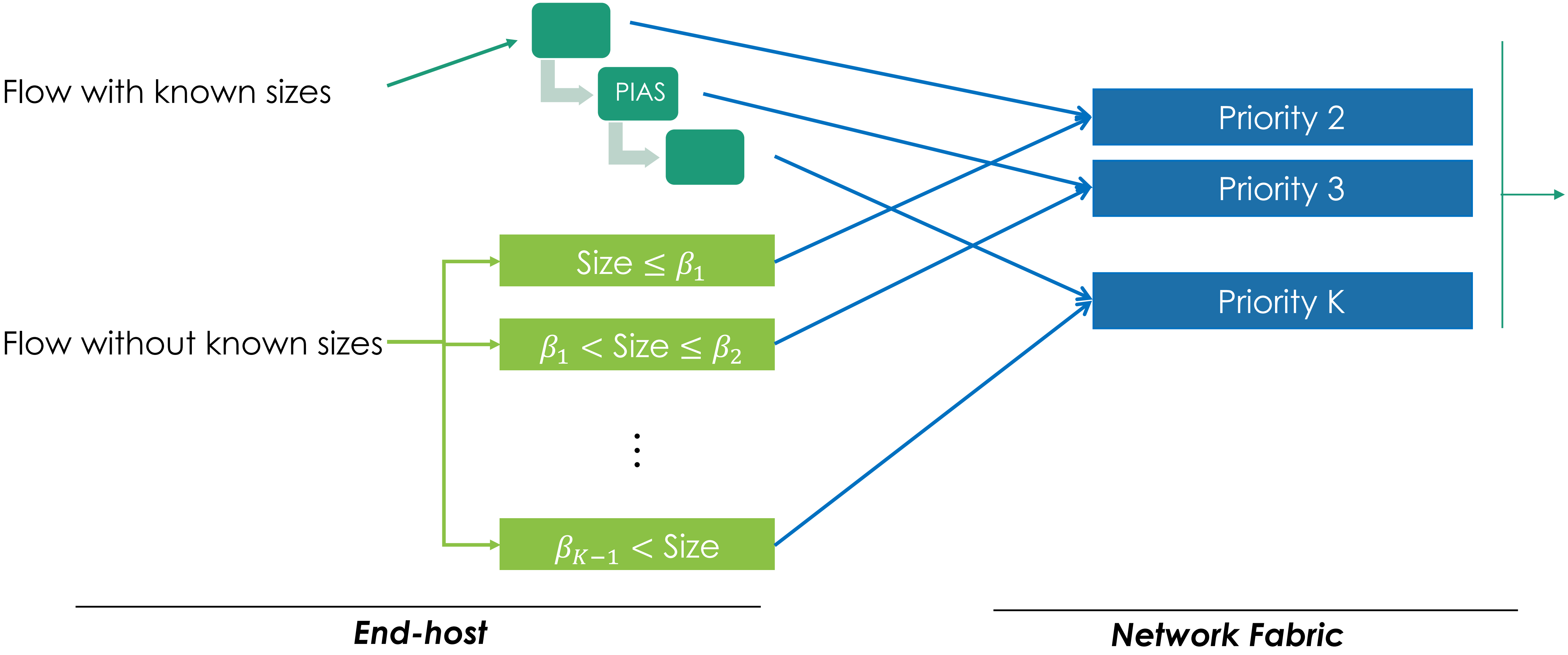
Sum of Linear Ratios Problem (PIAS) → Reformulation to include flows with known sizes → Quadratic Sum of Ratios Problem (Karuna)

Demotion Thresholds: $\{\alpha_i\}$

Demotion Thresholds: $\{\alpha_i\}$

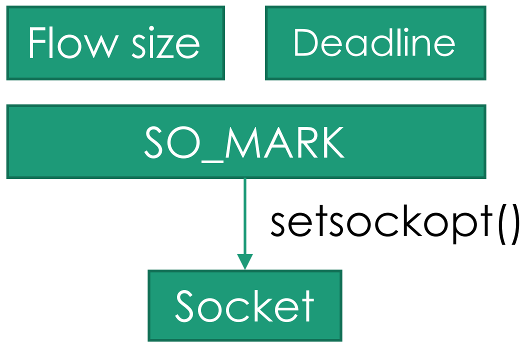
Splitting Thresholds: $\{\beta_i\}$

Karuna for non-deadline flows: mimicking SJF



Implementation

Implementation



End-host

Information passing

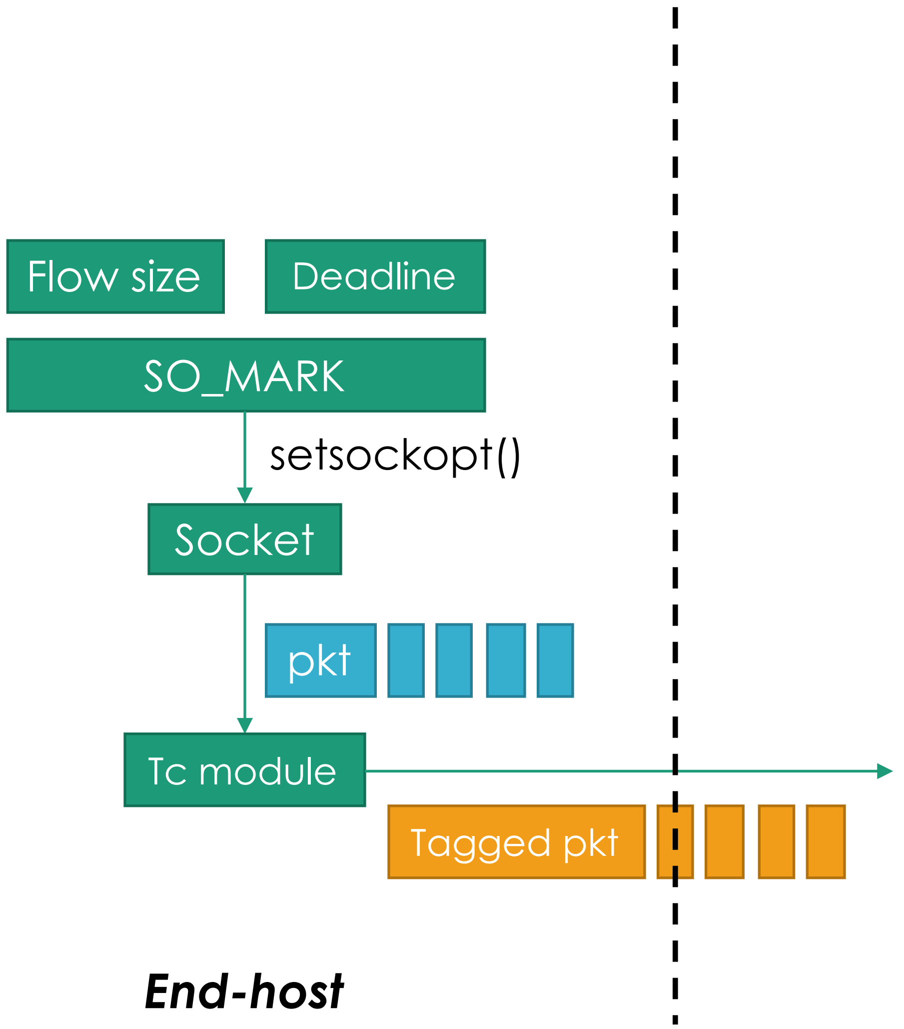
Pass flow information (deadline, size) to the kernel using SO_MARK

Network Fabric

Implementation

Information Passing
Packet tagging

TC module at the sender-side.
Tag DSCP fields in packet headers based on thresholds.

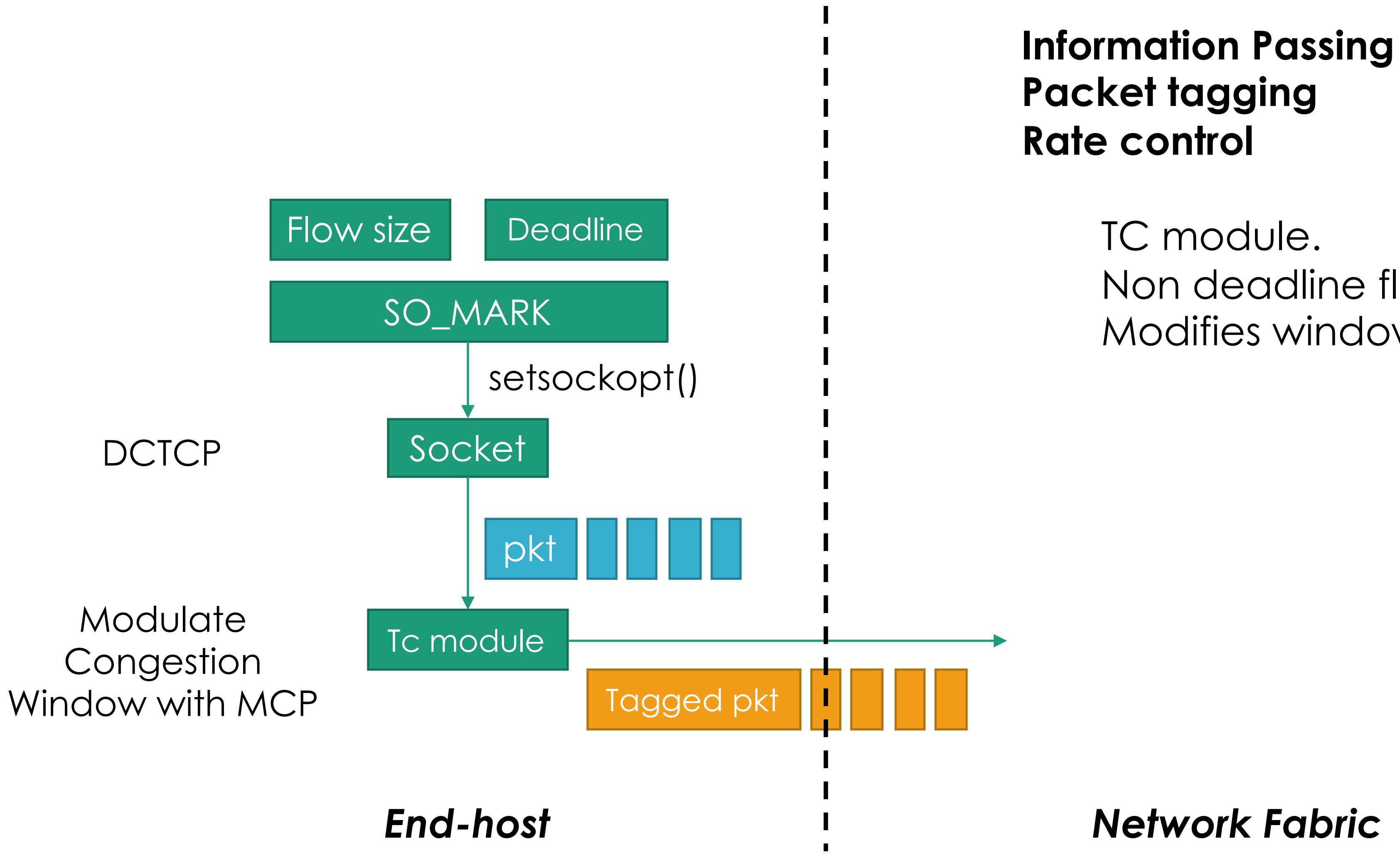


Network Fabric

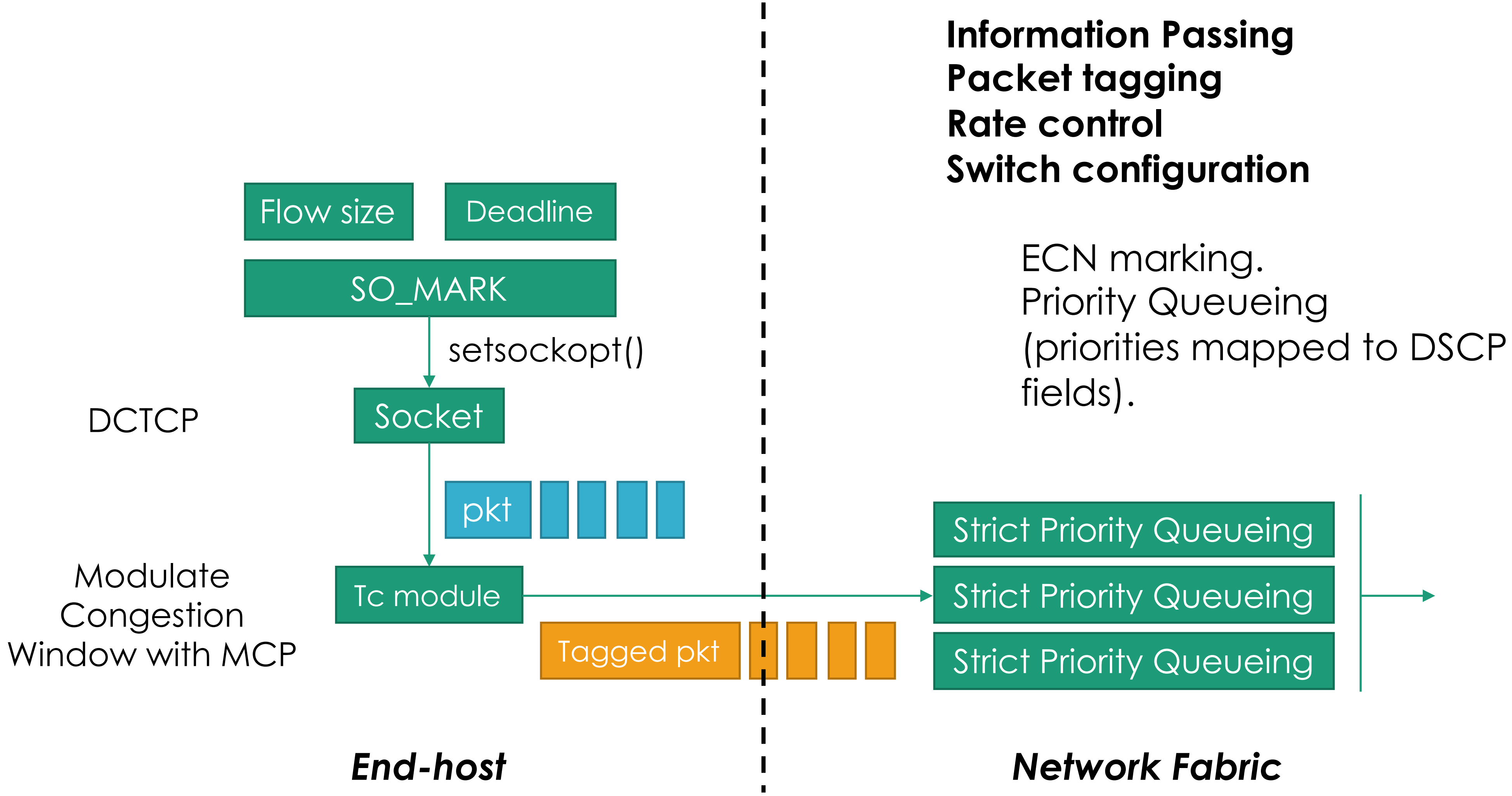
Implementation

Information Passing
Packet tagging
Rate control

TC module.
Non deadline flows use DCTCP
Modifies window size using MCP.



Implementation



Evaluation

Testbed Experiments

Simulations

Evaluation: Testbed Experiments

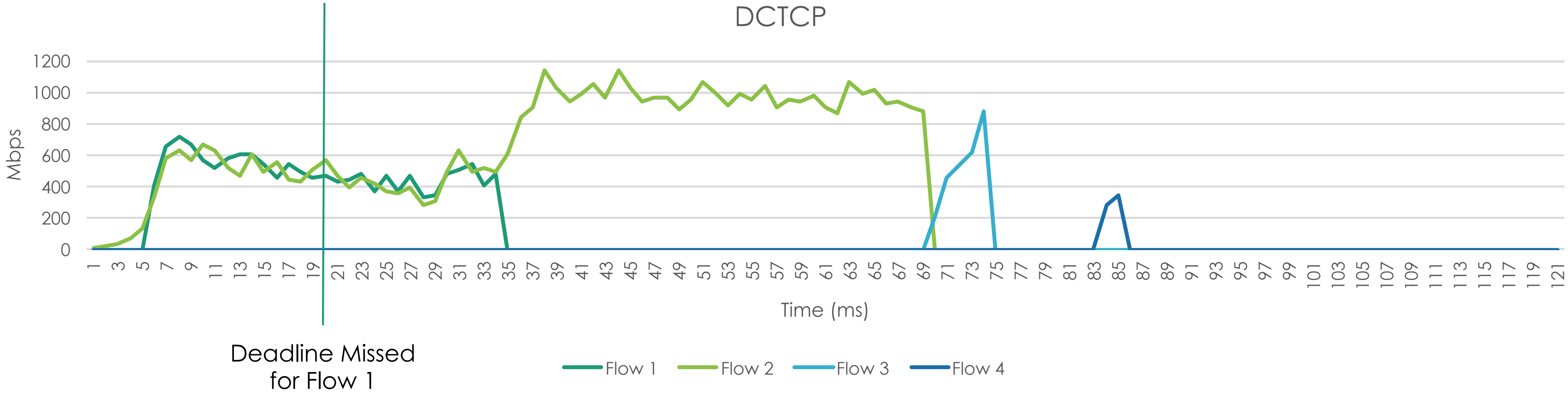
- Setup
  - 16 servers
  - A Gigabit Pronto-3295 switch
  - 8 Priority queues mapped to DSCP
  - RTT ~100us
  - Karuna kernel module
- Traffic trace
  - Web search (DCTCP [3])
  - Data mining (VL2 [4])

[3] Alizadeh, Mohammad, et al. "Data center tcp (dctcp)." *ACM SIGCOMM computer communication review*. Vol. 40. No. 4. ACM, 2010.

[4] Greenberg, Albert, et al. "VL2: a scalable and flexible data center network."*ACM SIGCOMM computer communication review*. Vol. 39. No. 4. ACM, 2009.

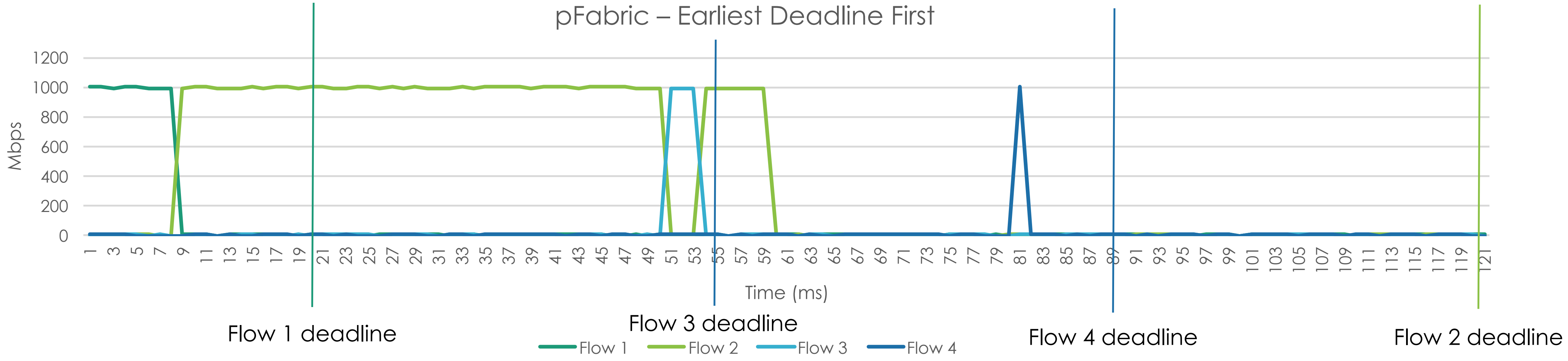
Testbed Experiments: Deadline Flows

Flow	Size	Deadline	Start Time
1	14.4MB	20ms	0ms
2	48MB	120ms	0ms
3	3MB	5ms	50ms
4	0.5MB	10ms	80ms



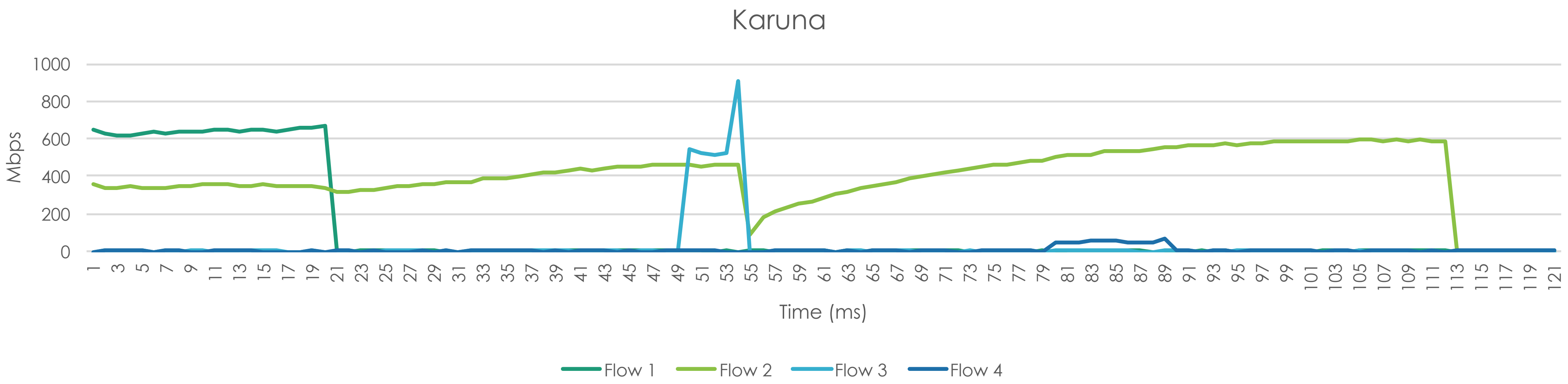
Testbed Experiments: Deadline Flows

Flow	Size	Deadline	Start Time
1	14.4MB	20ms	0ms
2	48MB	120ms	0ms
3	3MB	5ms	50ms
4	0.5MB	10ms	80ms



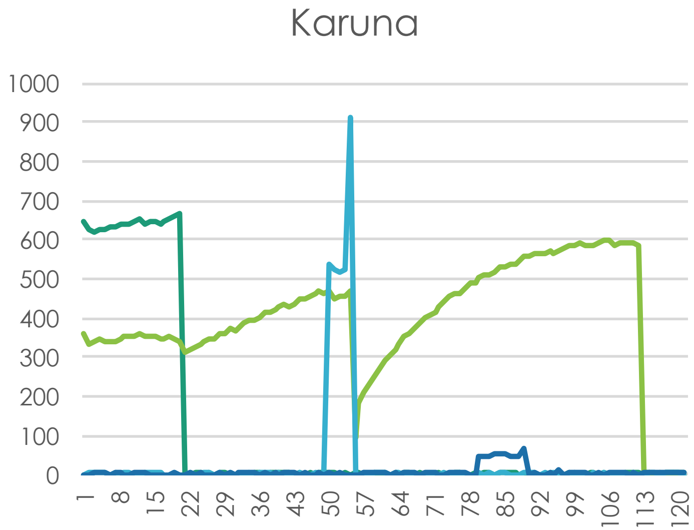
Testbed Experiments: Deadline Flows

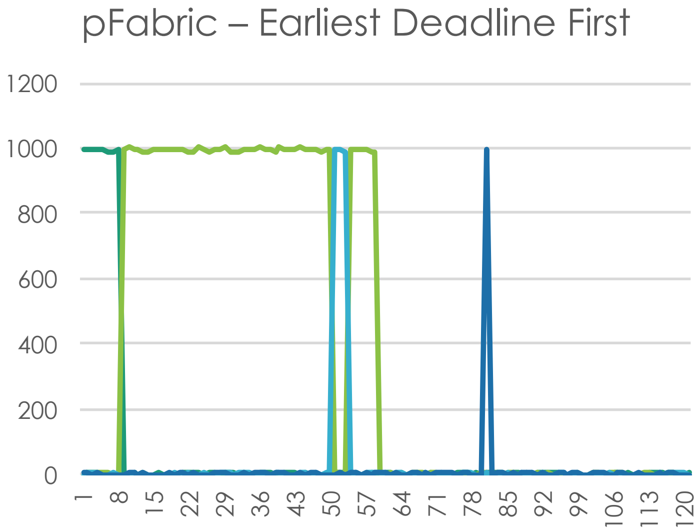
Flow	Size	Deadline	Start Time
1	14.4MB	20ms	0ms
2	48MB	120ms	0ms
3	3MB	5ms	50ms
4	0.5MB	10ms	80ms

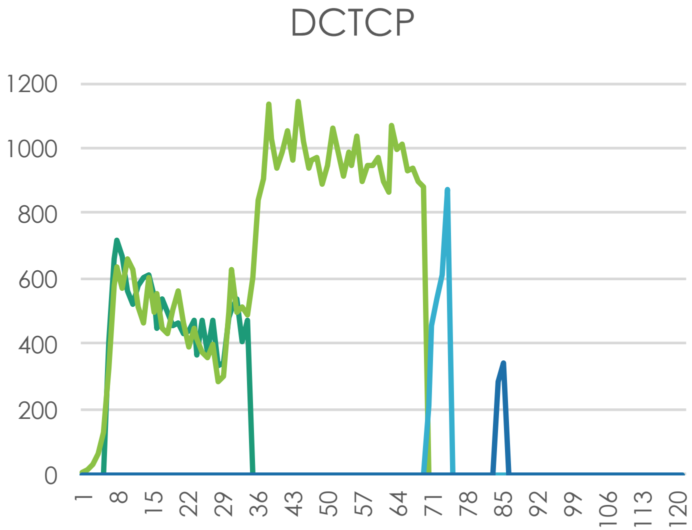


Testbed Experiments: Deadline Flows

Flow	Size	Deadline	Start Time
1	14.4MB	20ms	0ms
2	48MB	120ms	0ms
3	3MB	5ms	50ms
4	0.5MB	10ms	80ms

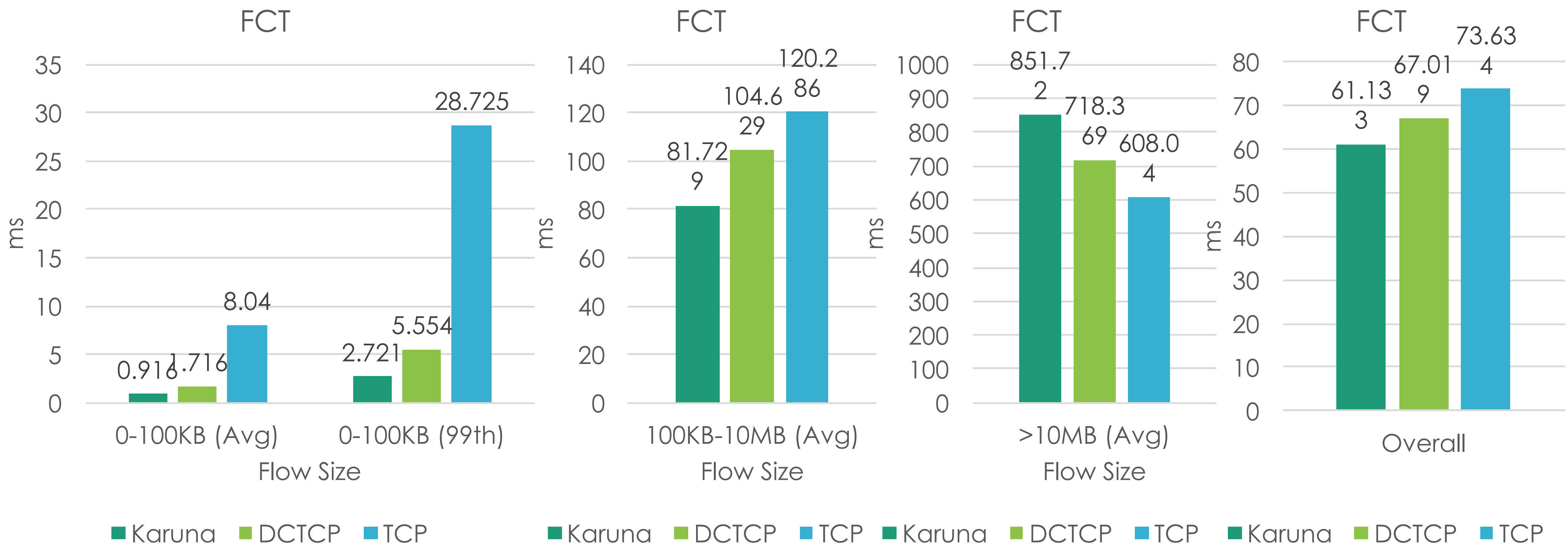






Karuna completes deadline flow just before deadline, leaving bandwidth for non-deadline flows.

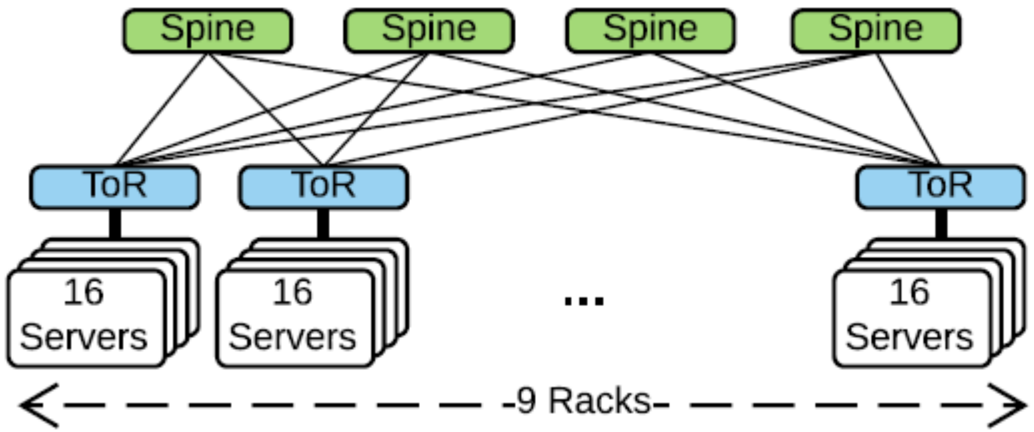
Testbed Experiments: Non-deadline Flows



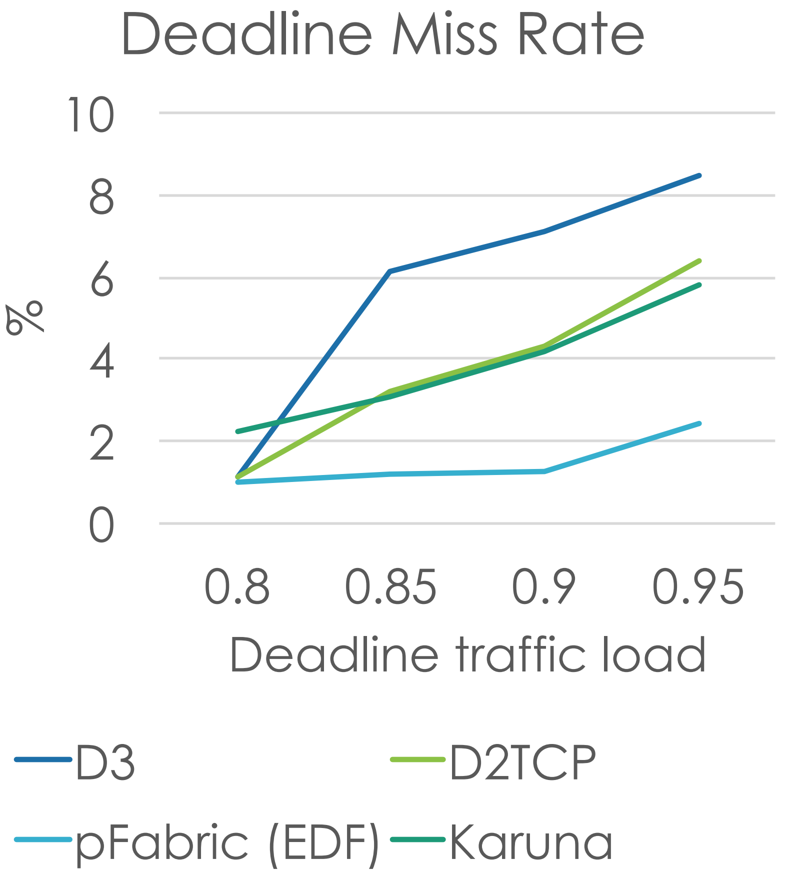
Mimics shortest job first scheduling for non-deadline flows.

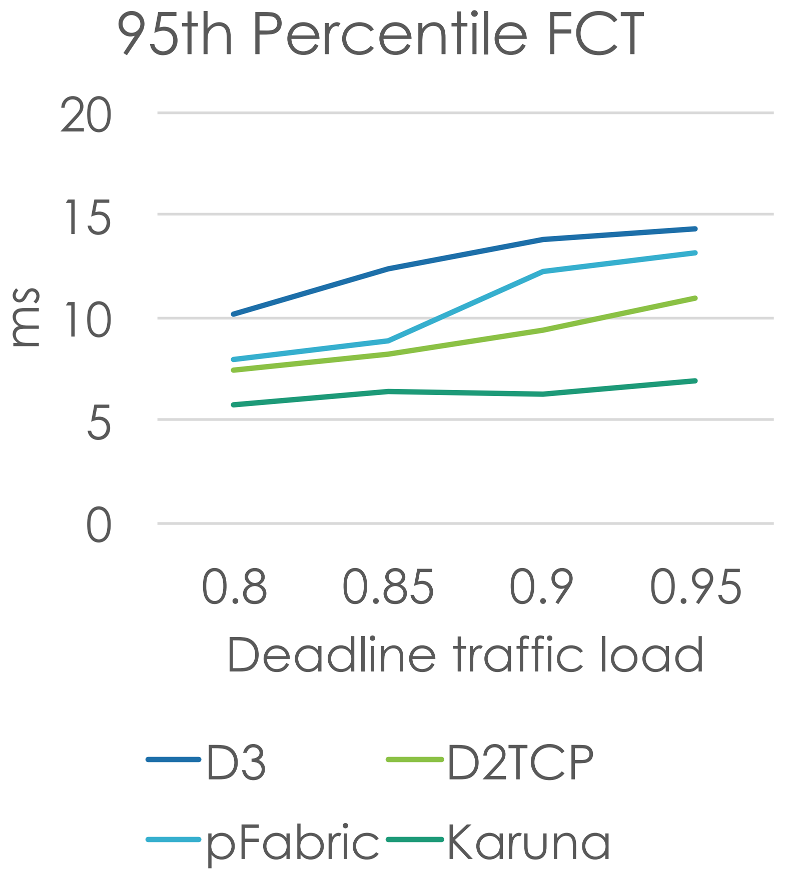
Evaluation: Simulations

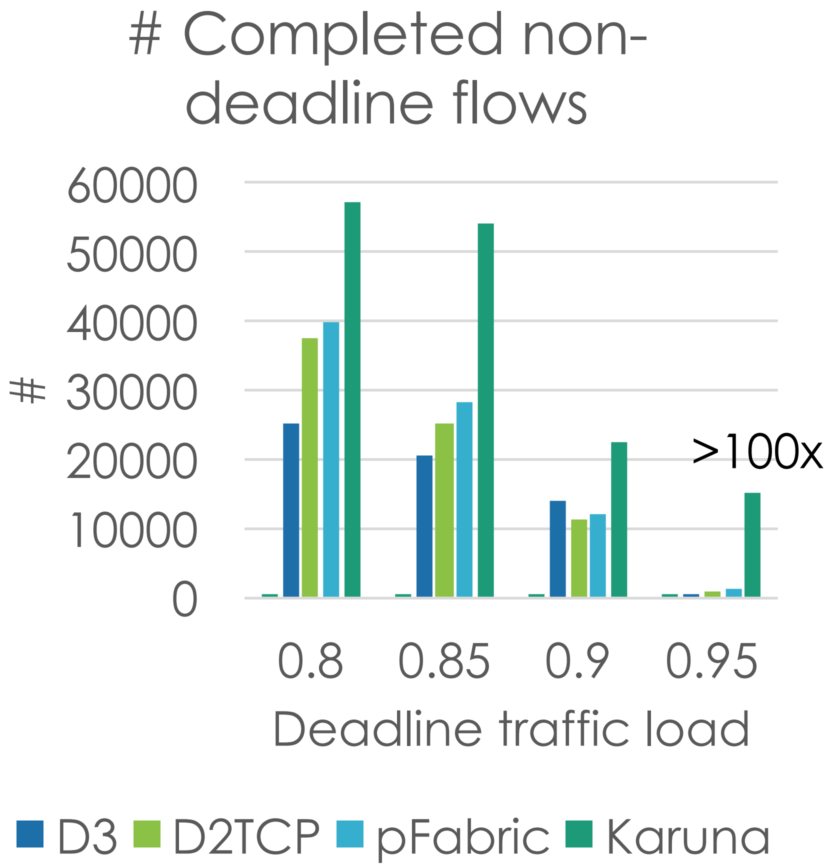
- Simulation Setup
  - Spine-leaf with 144 servers
  - 10G Server-ToR links
  - 40G ToR-Spine links
- Compare with:
  - D3
  - D2TCP
  - pFabric - EDF



Large-scale Simulations: Key Benefit of Karuna







Reducing completion times of non-deadline flows while completing deadline flows.

Concluding remarks

- Filling a gap in datacenter flow scheduling
- Karuna
  - Prioritizes deadline flows but control their rates.
  - Uses the remaining bandwidth to schedule non-deadline flows based on size.

- ***Thank you! Q & A!***