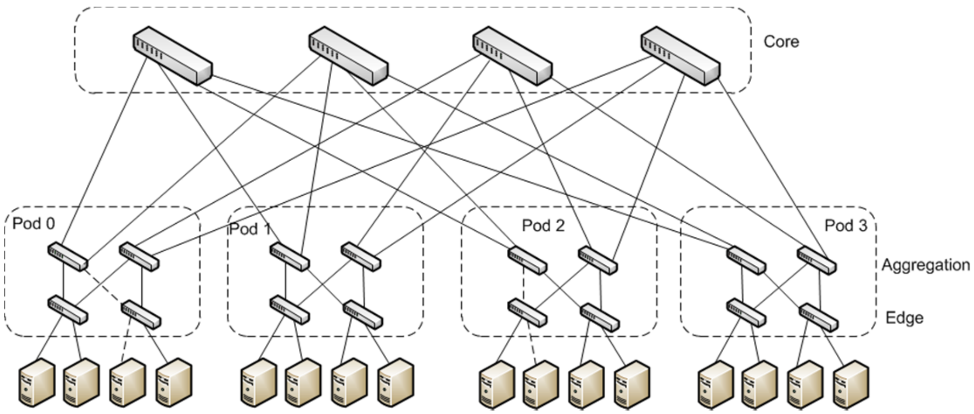
NJIT
New Jersey's Science & Technology University

Expediting the transport of Data Center Flows (DAQ: Deadline-Aware Queue)

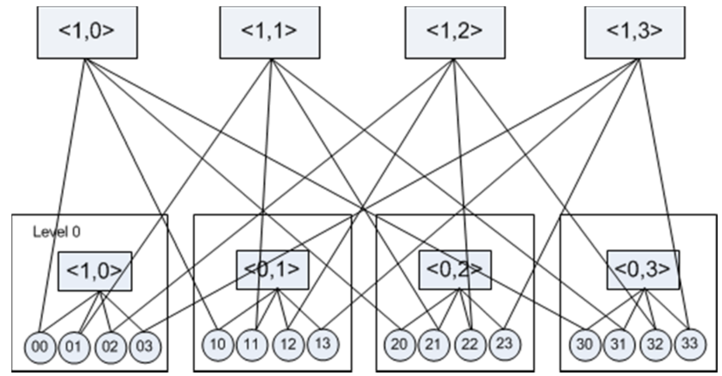
Roberto Rojas-Cessa
Networking Research Laboratory
ECE Dept.
New Jersey Institute of Technology
Newark, NJ 07102
rojas@njit.edu

NRL

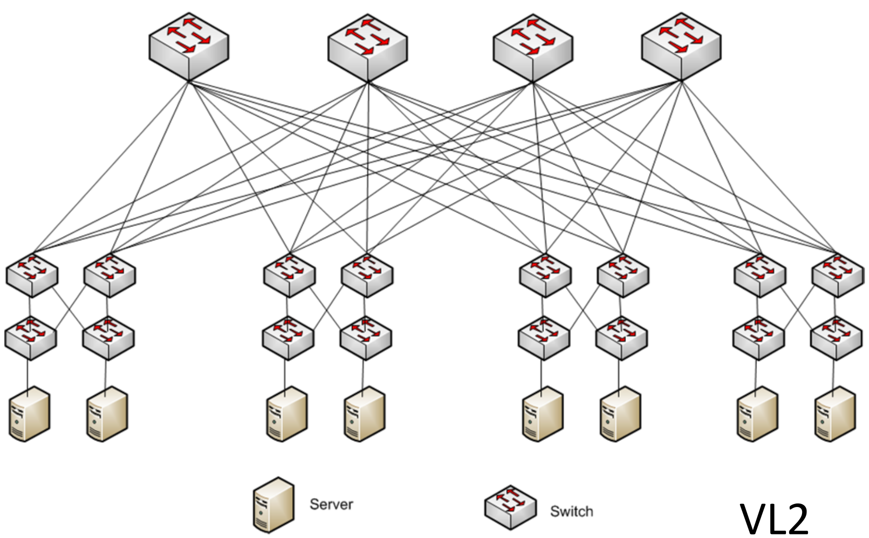
Examples of DC topologies



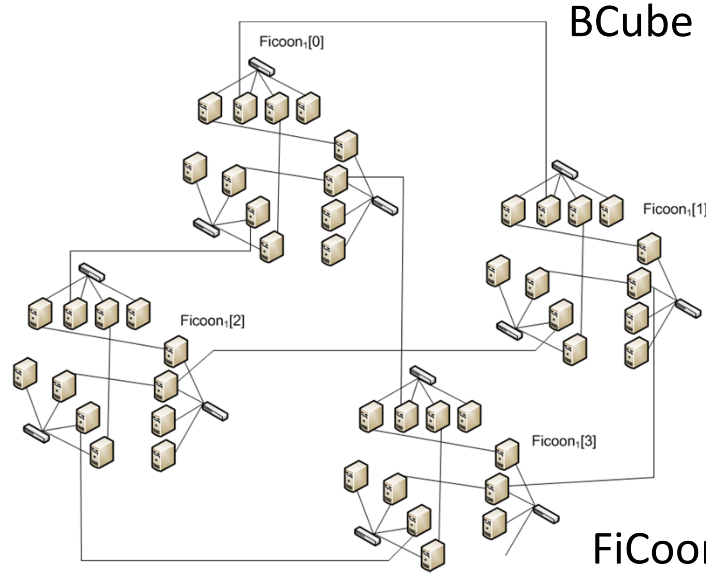
Fat-Tree



BCube

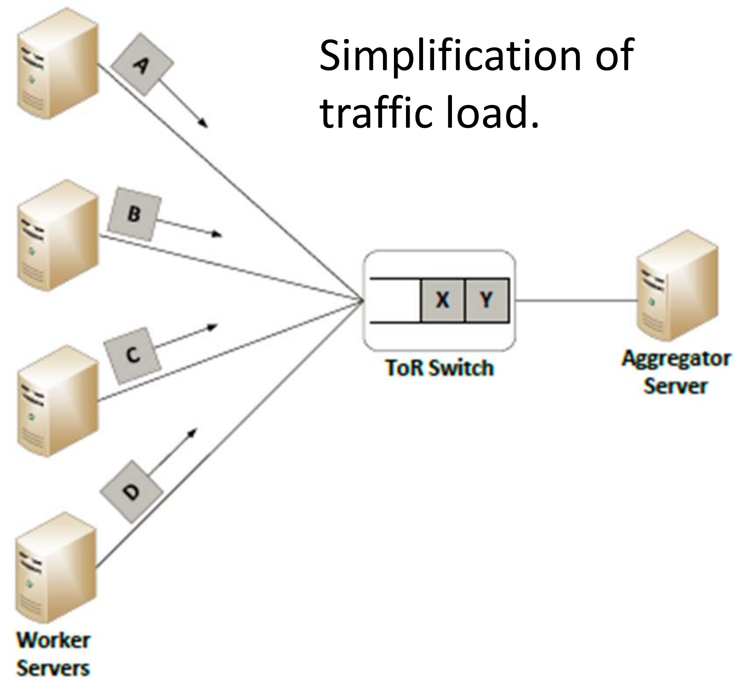


VL2



FiCoon

What is unique in Data Center Traffic? Partition-Aggregate Model



Simplification of traffic load.

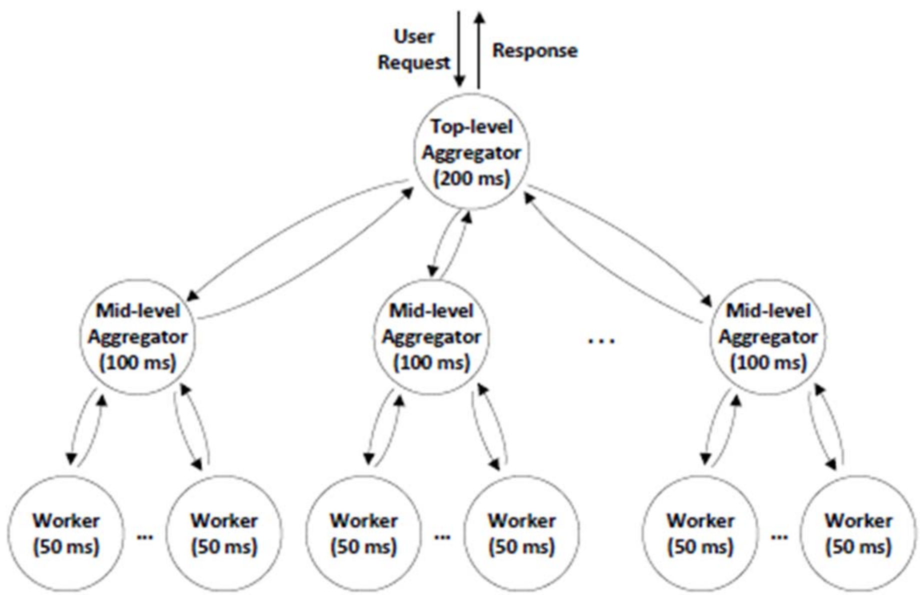
Data Aggregation

Flows may be associated with response deadlines

Deadlines are inherited by partial processes

For all flows, short **Flow Completion times (FCTs)** are desirable

For deadline-sensitive flows, short **Application Throughput** is desirable.



Data aggregation → Connection-Oriented Transport → Transmission Control Protocol (TCP)

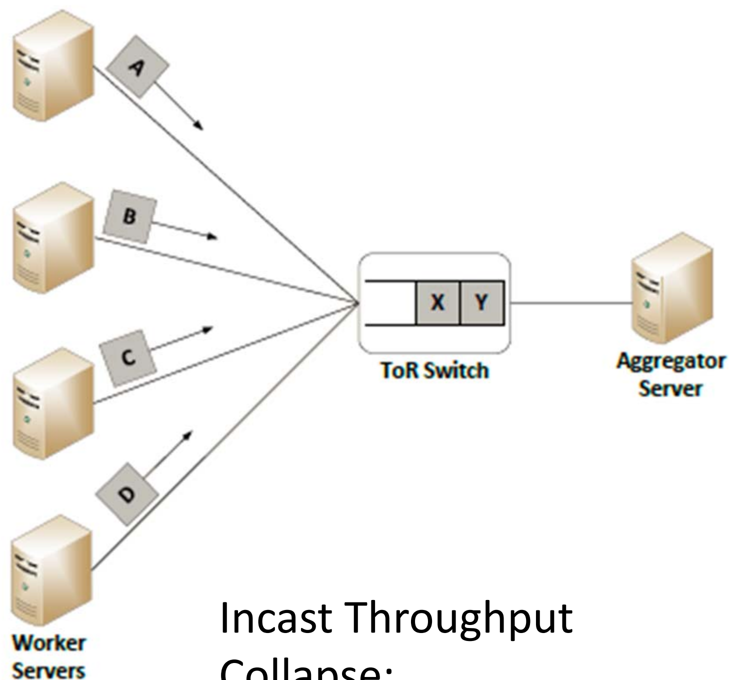
Expected requirements of a Data Center (DC) Transport Protocol

- Maximize the number of flows completing transmission before deadlines
- Guarantee a high throughput for long flows.
- Allow high, if not 100%, link utilization.
- Achieve lossless transmissions.
- Minimize the amount of state information at switches

Why TCP is not good enough?

- Data Center Flows: Long + Short Flows
- Congestion
- Multiple flows concur at aggregation switches
- Lack of a centralized scheduler

Flow control mechanisms are not transmission speed aware→ Long FCTs!



Incast Throughput Collapse: Retransmission Time Offs + Retransmission → choke bandwidth

Existing Solutions

- Earlier Congestion Notification (ECN): DCTCP
- Rate Control: D2TCP, D3, PDQ (deadline aware)
- Congestion Control: RCP
- Pacing Schemes: HULL
- Load Balancing Schemes: DeTail, CONGA, RepFlow
- Switch Modification: DAQ

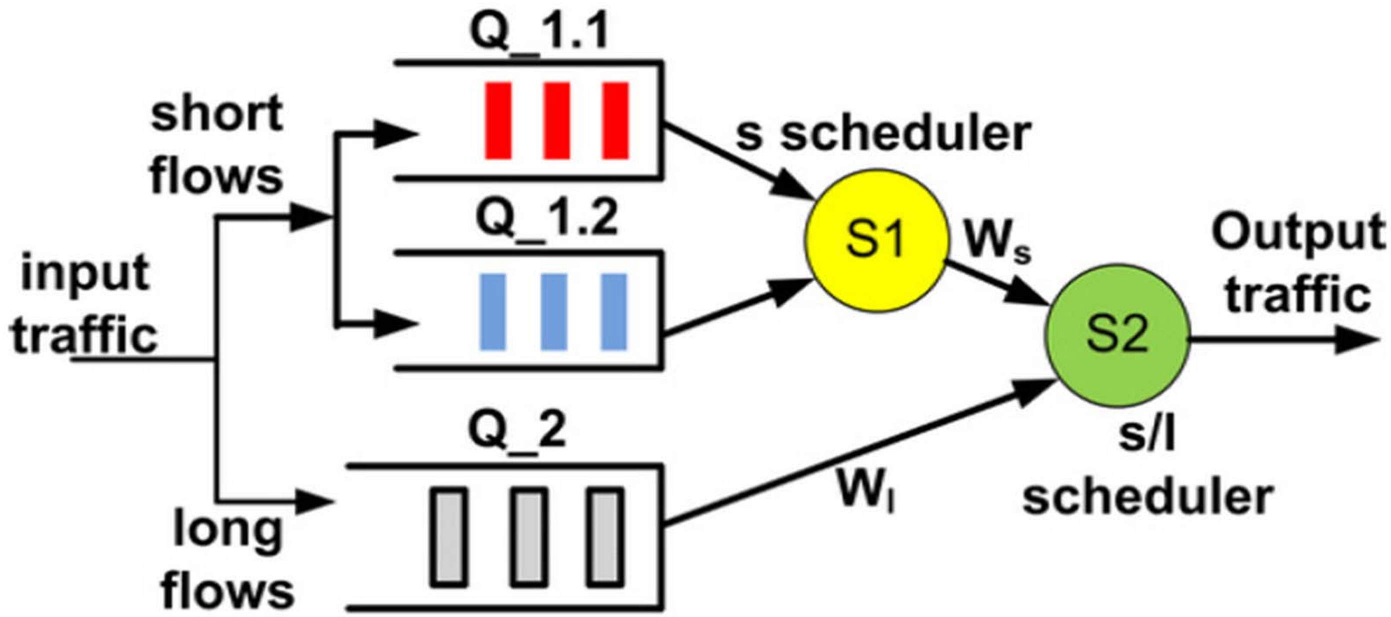
Deadline-Applicable Schemes

- **RCP** [Dukkipati05] assigns rate according to available bandwidth. Parameters must be tuned.
- **DCTCP** [Alizadeh10]: ECN + congestion window modification. Agnostic to deadlines.
- **D^3** [Wilson11] reserves transmission rates FCFS.
- **PDQ** [Hong12]: selects flows→ earliest deadline first (EDF) and the shortest job first (SJF). High complexity.

Proposed Scheme: Deadline Aware Queue (DAQ) at DC Switches

- Objectives:
  - Maximize application throughput
  - Ensure minimum bandwidth for long flows
  - Minimize flow-state information at switches
  - Minimize modification to layered protocols

Switch Architecture



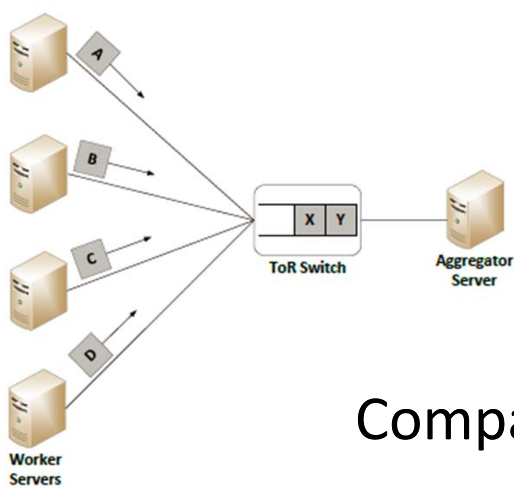
Use Three Queues: Urgent, Non-urgent, Long

Short flows: Urgent or Non-urgent

Long flows: long-flow queue + service weighted scheduling

Test setup

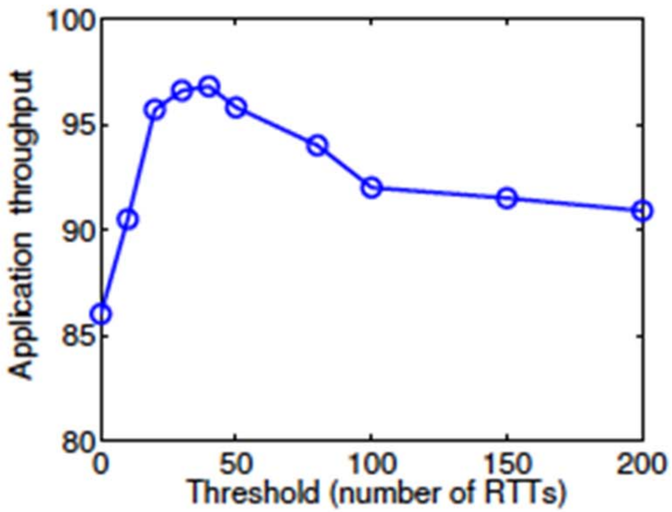
- Loss-less flow control between
  - Senders and switch
  - Switch and receiver (aggregator)
- Large congestion window size instead of slow start

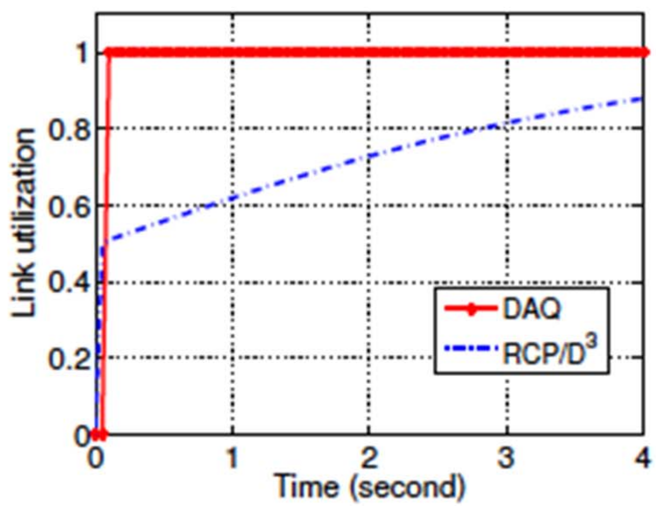


Comparison: RCP and D^3

Impact of Urgent Threshold Value

Application throughput: No. on-time flows/All arrived flows



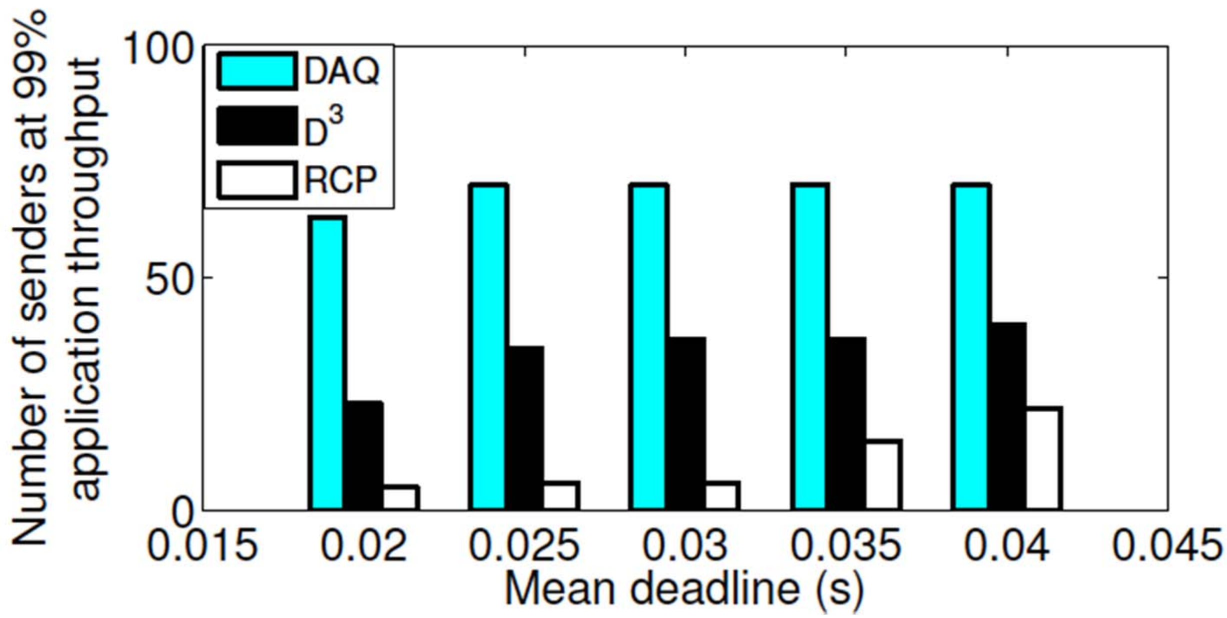


Flow size: 30KB, rate: 3600 flows/s

Number of long flows: 5

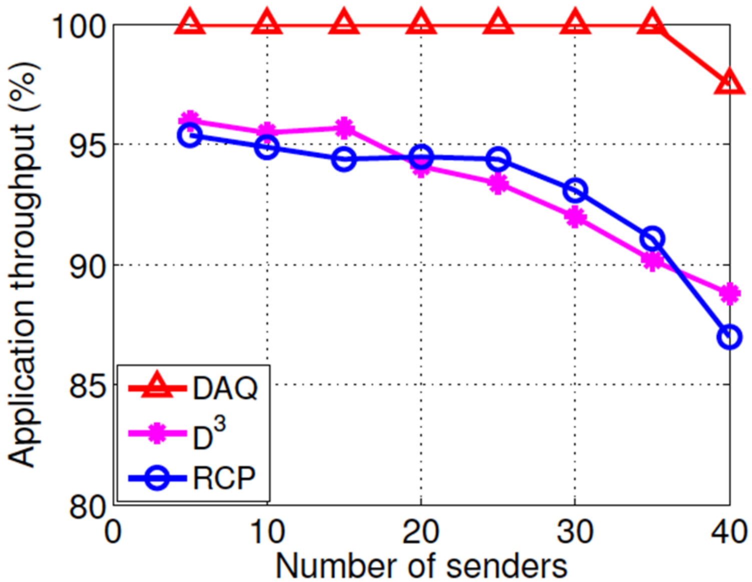
NRL

Supported number of senders

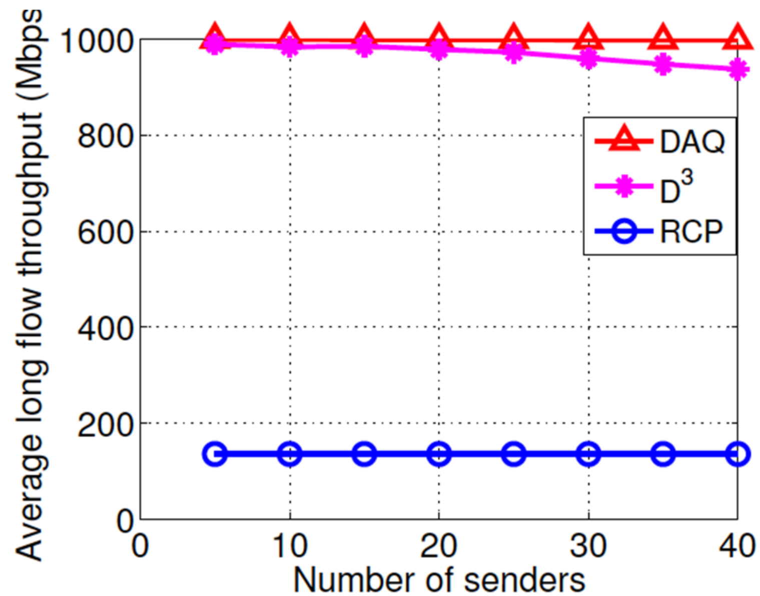


Number of concurrent senders for achieving 99% application throughput with flow size mean of 10 Kbytes and deadlines [20, 40] ms.

Application and Average Throughput



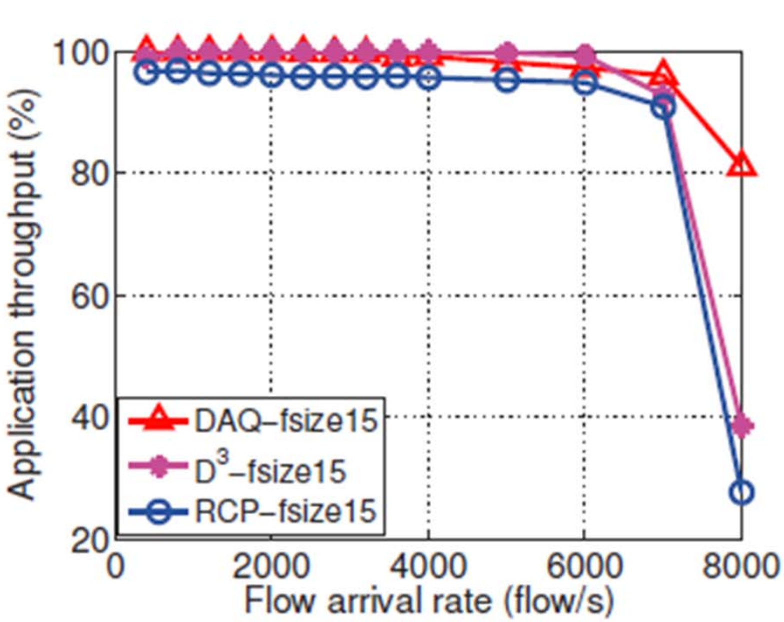
Short flows



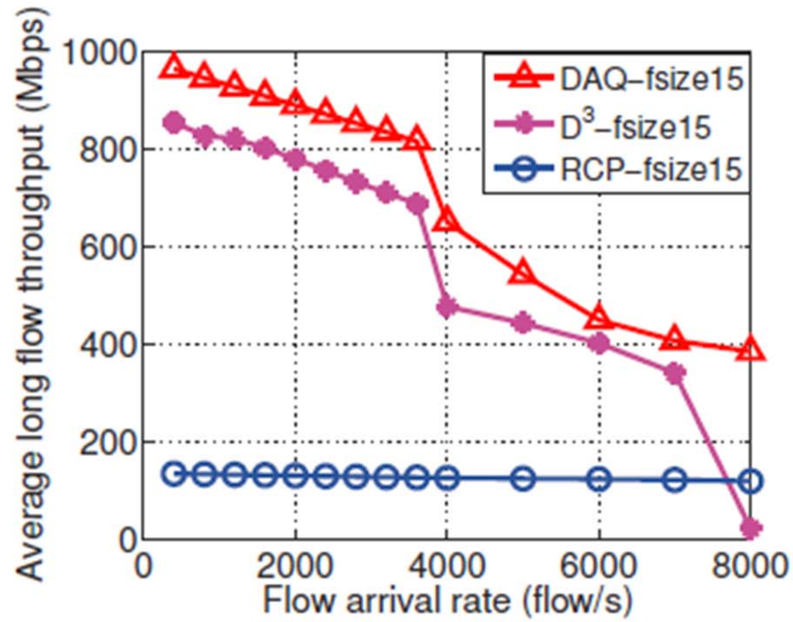
Long flows

Short flow size: 15 Kbyte, long flow size: 100Mbyte (2).
Short flow load: 0.3 %
No. of senders: [5, 40]

Performance under short and long flows



(a) Application throughput for short flows.



(b) Average long flows throughput.

Short flow size: 15KB

Conclusions

- Deadline-oriented approach with small modification to transport layer.
- Urgent flows receive preferential service.
- Few urgent flows speedup transmission.
- DAQ achieves high Application Throughput
- Long flows receive minimum throughput through Weighted Round-Robin

Thank you

rojas@njit.edu