NegotiaToR: Towards A Simple Yet Effective On-demand Reconfigurable Datacenter Network

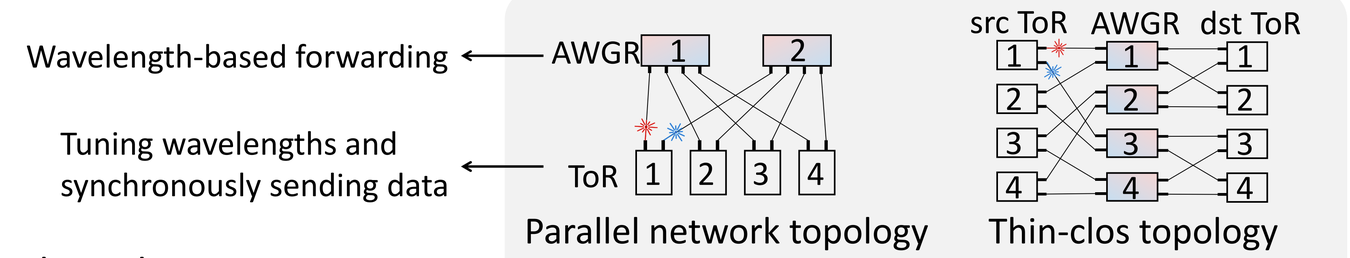
Cong Liang [1], Xiangli Song [1], Jing Cheng [1], Mowei Wang [2], Yashe Liu [2], Zhenhua Liu [2], Shizhen Zhao [3], Yong Cui [1]

[1] Tsinghua University, [2] Huawei Technologies Co., Ltd, [3] Shanghai Jiao Tong University

Background

- In the post Moore's Law era, packet-switched DCNs struggle to fulfill traffic's goodput and latency requirements
- Optical reconfigurable DCNs: high capacity and low cost
  - Fast optical switching hardware is ready for DCNs
  - Scheduling: Rapidly getting non-conflicting paths
    - **Traffic-oblivious:** Practical, but with sacrificed latency and goodput
    - **On-demand:** Potential performance gains, but with practicality concerns…

AWGR-based fast switching with flat topologies



Recent advances

1. ***Low reconfiguration delay*** *within 10 ns* [1]
2. ***DCN scalability*** *using flat topologies above ToRs*

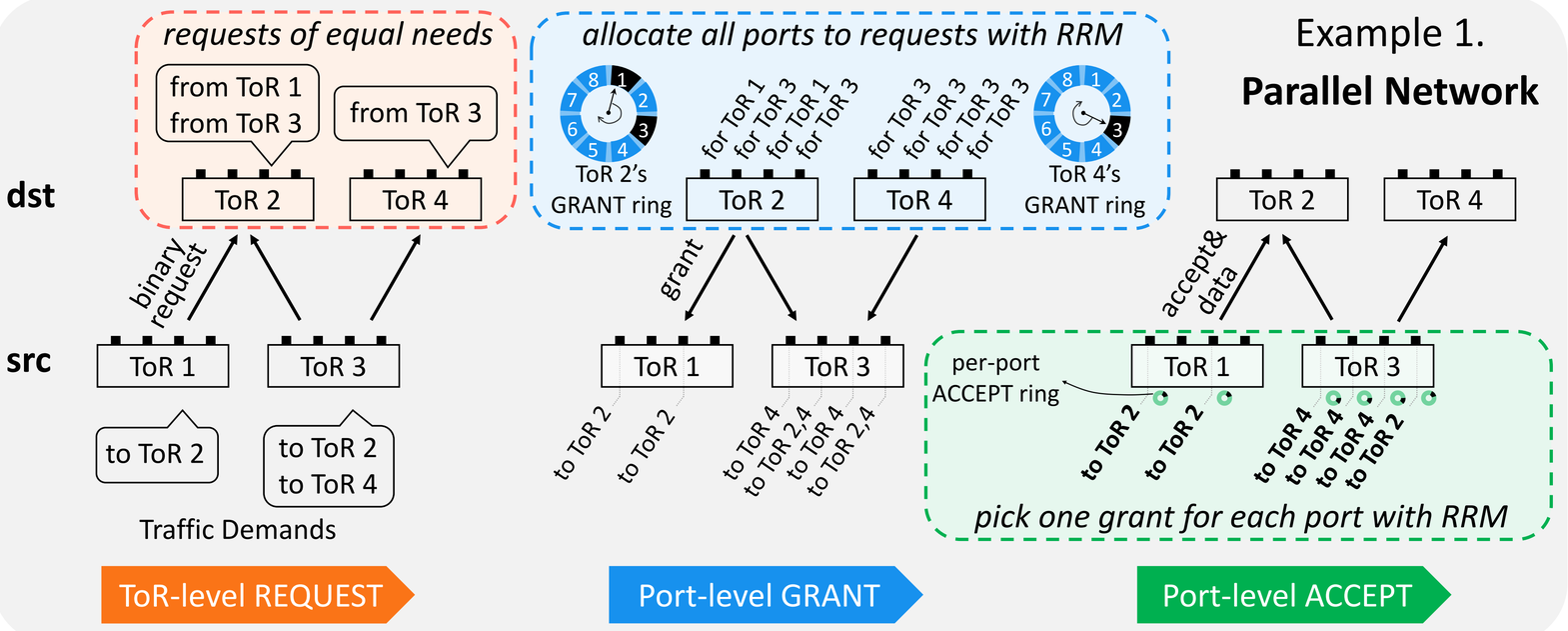
[1] Ballani et al., SIGCOMM '20

Design

◆ **Goal: Practical yet high performance on-demand distributed scheduling**

➢ **NegotiaToR Matching:** On-demand distributed scheduling on flat topologies

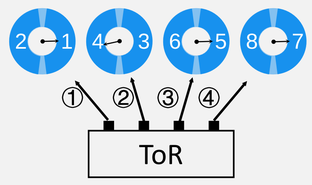
[2] McKeown, PhD Thesis, UC Berkeley '95



Example 1.
Parallel Network

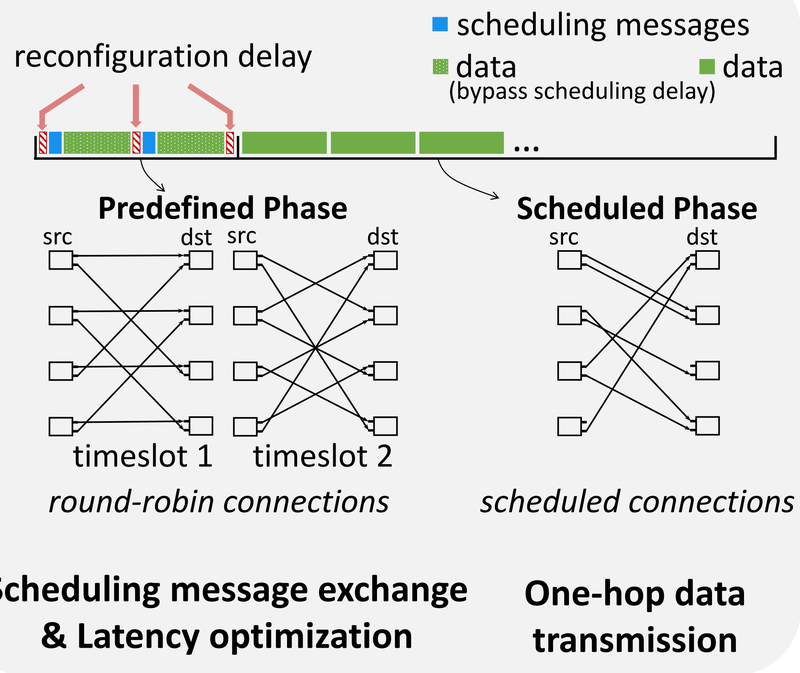
Example 2.
Thin-Clos

Similar with the parallel network, but change the per-ToR GRANT ring to per-port GRANT ring



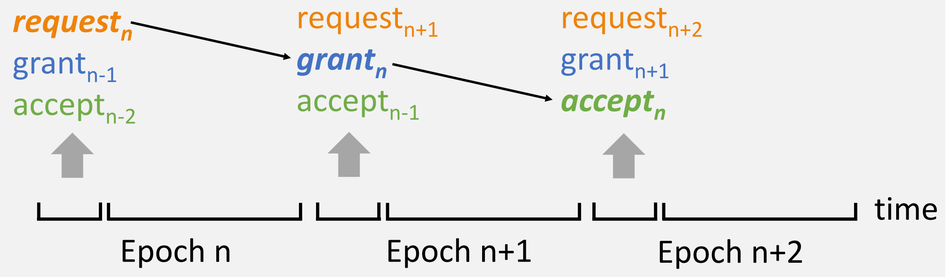
- Inspired by RRM [2] inside crossbar switches
- Tailored towards minimalist on-demand scheduling on flat topologies
  - Non-iterative, binary requests, stateless scheduling, no data relay

➢ **Two-phase epoch**



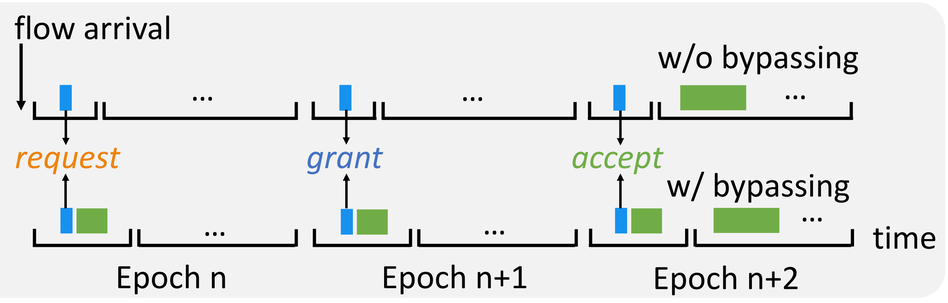
Pipelined scheduling in the predefined phase

- Get a set of accepts in each epoch
- Reconfiguration and data transmission after accepts



Scheduling delay bypassing

- Piggybacking small packets with scheduling messages
- Mice flow latency optimization

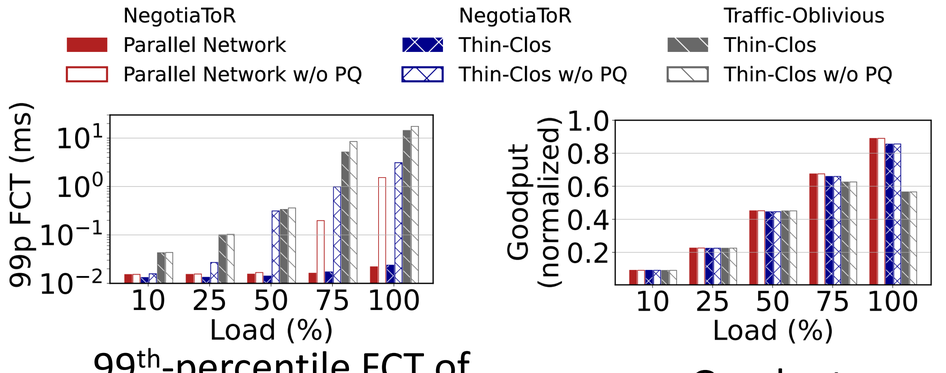


Evaluation

◆ Simulation setup: 128 ToRs, Hadoop trace [3]

- Connected by 2 representative flat topologies
- 10 ns reconfiguration delay, 2x speedup for both NegotiaToR and the traffic-oblivious scheme [4]. Consider ToRs as endpoints
- Results without mice flow prioritization are also shown

NegotiaToR achieves both small mice flow FCT and high goodput on two representative flat topologies



99th-percentile FCT of mice flows (<10KB)

Goodput