RICE

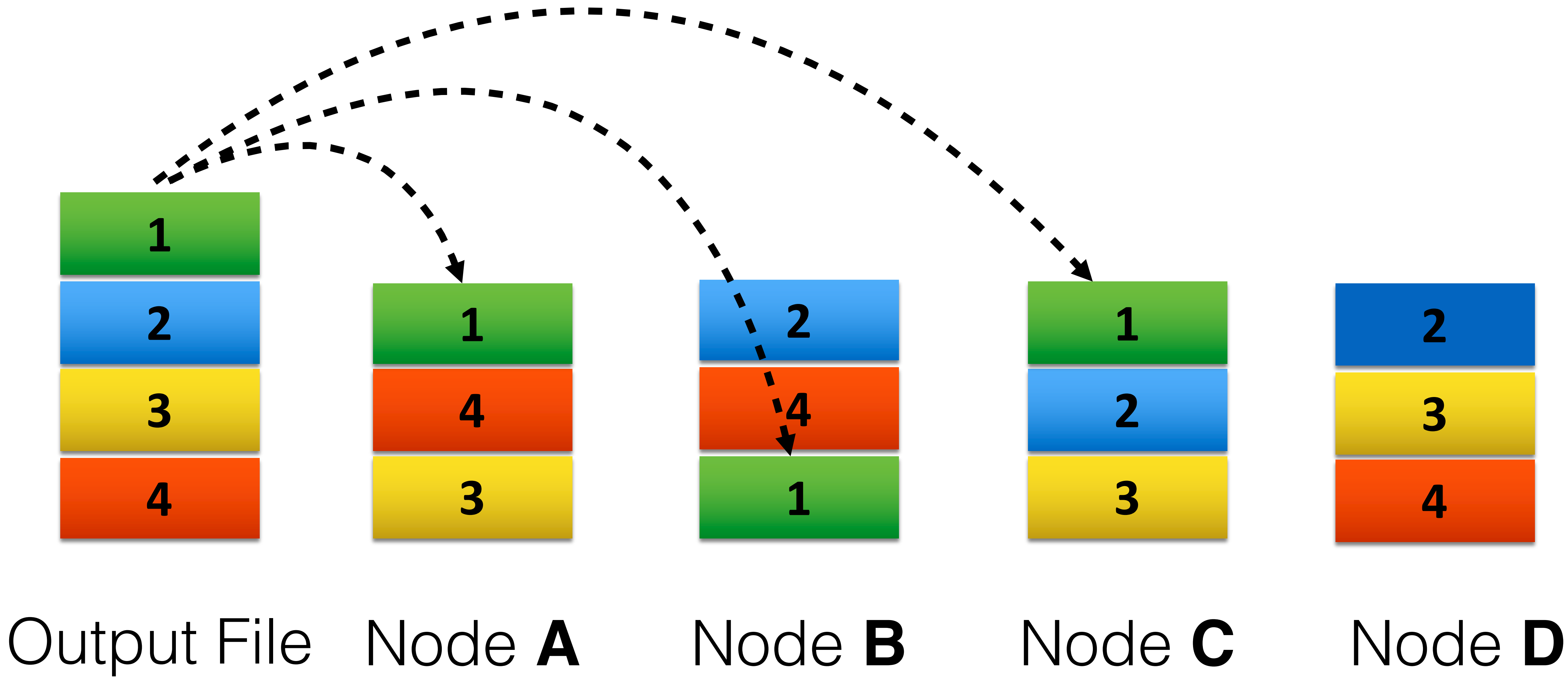
HyperOptics: A High Throughput and Low Latency Multicast Architecture for Data Centers

Dingming Wu, Xiaoye Sun, Yiting Xia, Xin Huang, T. S. Eugene Ng

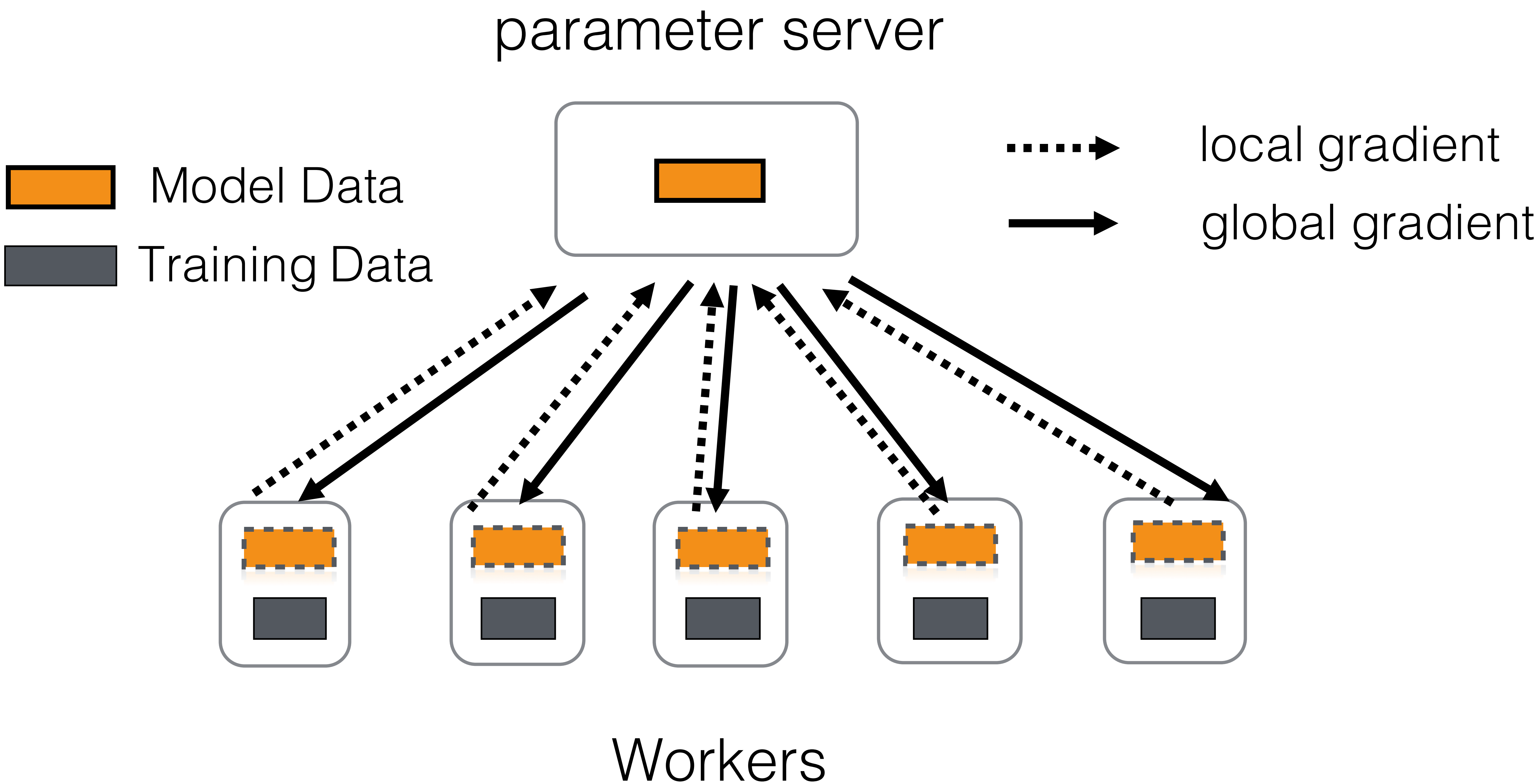
Department of Computer Science
Rice University

Multicast Applications

Data Replication in Distributed File Systems



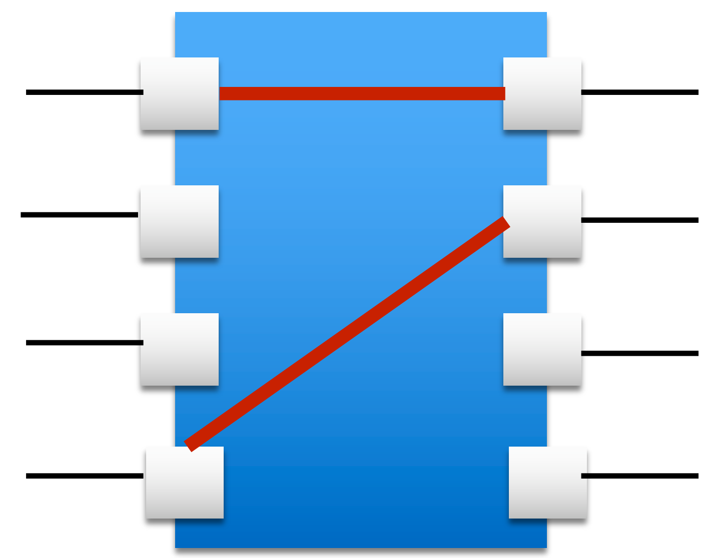
Multicast Applications Distributed Machine Learning



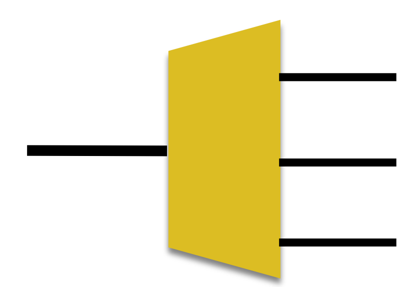
Background

Comparisons of Optical and Electrical Network

	Electrical Network	Optical Network
Principle	Packet Switching	Circuit Switching
Bandwidth Upgrade	Hard	Easy
Energy Efficiency	Low	High
Switching Latency	Low	High

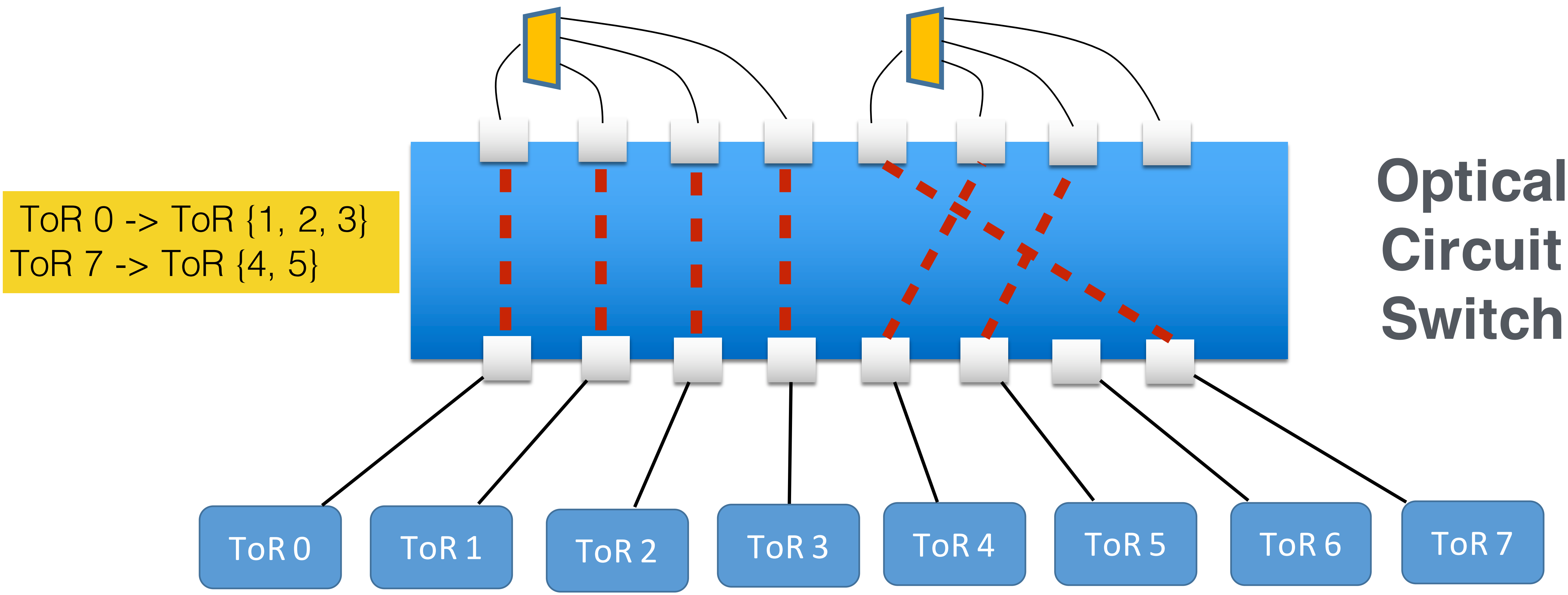


Optical Circuit Switch



Optical Splitter

Existing Optical Networks



e.g. Blast by Xia *et al.* (INFOCOM'15), Work by Wang *et al.* (CCR'13)

- ✓ Flexible topology configuration
- ✓ Data rate transparency
- ✓ Low energy consumption and heat dissipation (<50W)
- ✖ **Poor scalability (limited port-count)**
- ✖ **Slow switching speed (tens of ms)**

Our Solution

Goal: Retain the advantages of optical networks, avoid the port-count limitation and slow switching speed of OCS

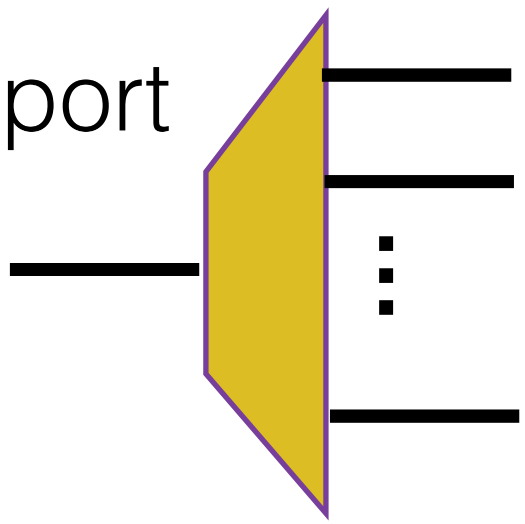
Basic idea:

1. Eliminate the use of OCS
2. Interconnect ToRs statically by optical splitters
3. Non-directly connected ToRs use relays to talk to each other.

Building blocks

Optical Splitter

1 input port



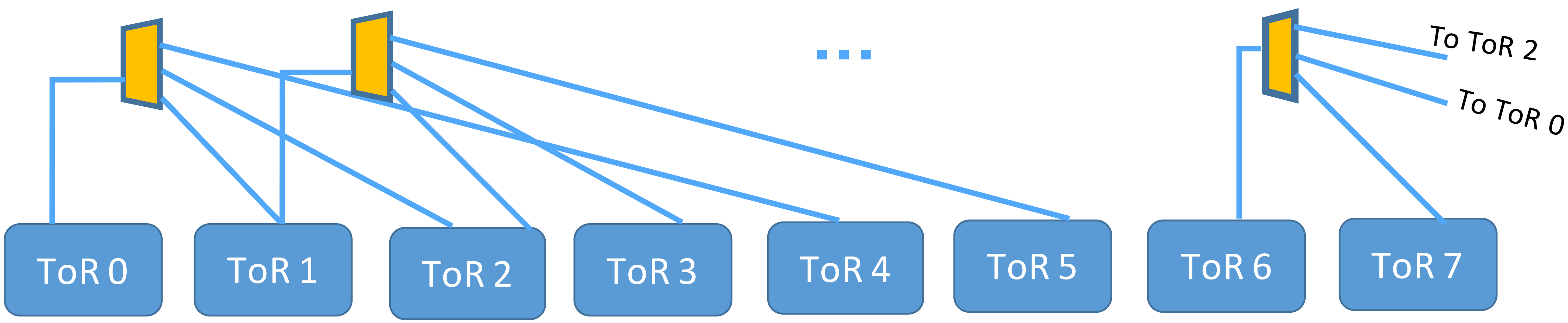
k output ports

1 x k optical splitter, fanout is k

Network Architecture

Assume the number of ToRs n = 2^k, k is the splitter fanout

- **ToR *i* is connected to the input port of splitter *i***
- **The output ports of splitter *i* are connected to ToR *(i+2^0)%n, (i+2^1)%n, … (i +2^{k-1})%n***



Inspired by Chord (sigcomm'01) in overlay networks

Analysis

Maximum path length: log(n)

-average: log(n)/2

Two simultaneous active one-to-all multicasts with full bandwidth

# of occupied ports per ToR

-log(n)

Cost

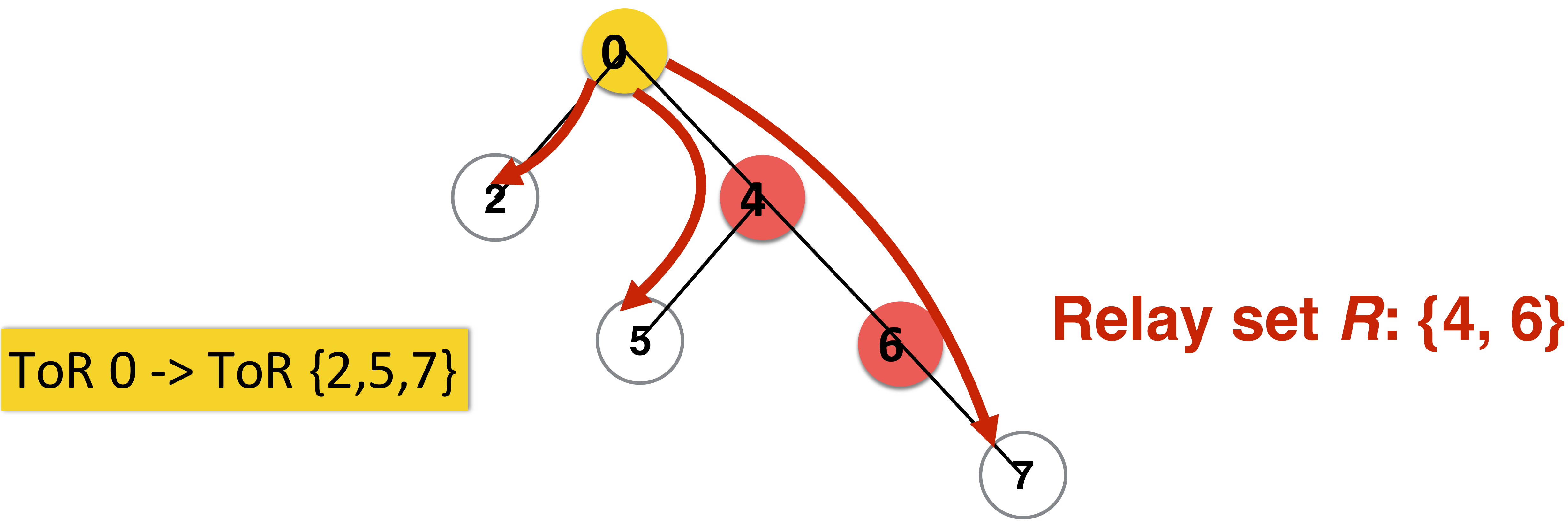
—comparable to the OCS architecture

—cost trend needs discussion

Multicast Tree Building

Multicast Request: (s, D, f)

R = **EmptySet**
for *d* in D
    compute a shortest path *p* from *s* to *d*
    R = R **Union** *p*



Multicast Scheduling

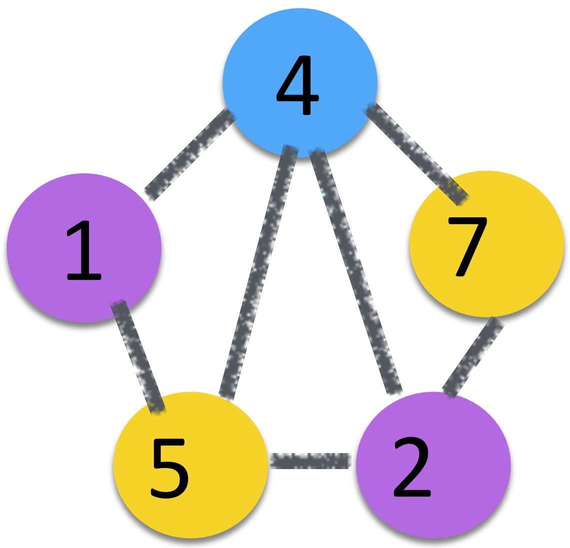
Input: a list of multicast requests

Output: the starting time of each multicast transmission

Goal: minimize the overall flow completion time

Constraints: two multicast trees that share some relays must be serviced sequentially

Multicast Scheduling



Vertices: multicast requests

Edges: conflicts

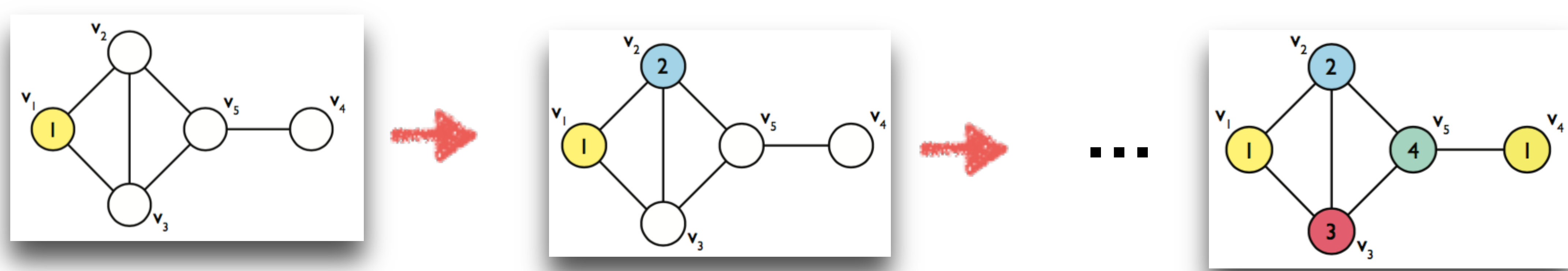
Weights: Flow Durations

Goal: Minimize sum of max weight of each color

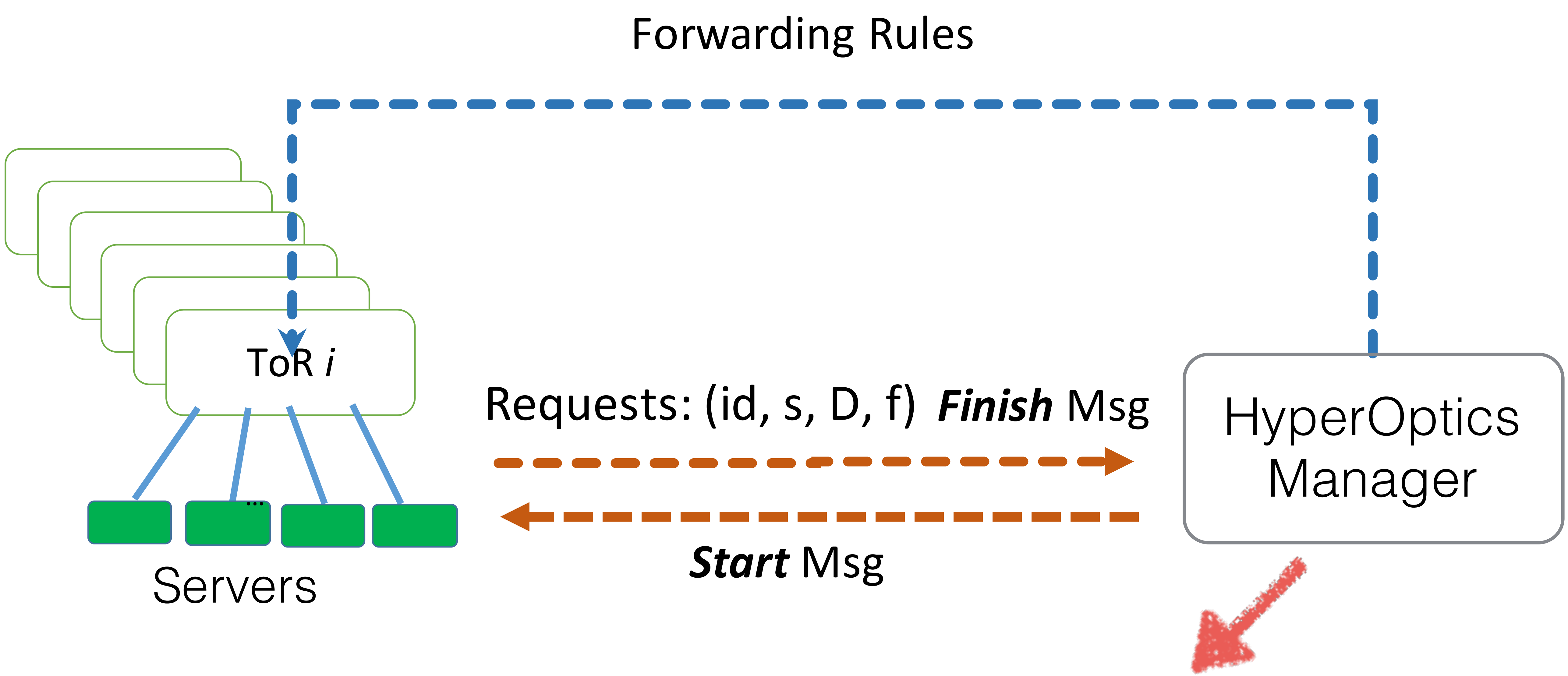
$$+ \max(\quad , \quad) + \max(\quad , \quad) = 13$$

Online Coloring

-A common heuristic to approach graph coloring



System Overview



ID	Start	Finish
1	t0	t1
2	t2	t3
...	...	...

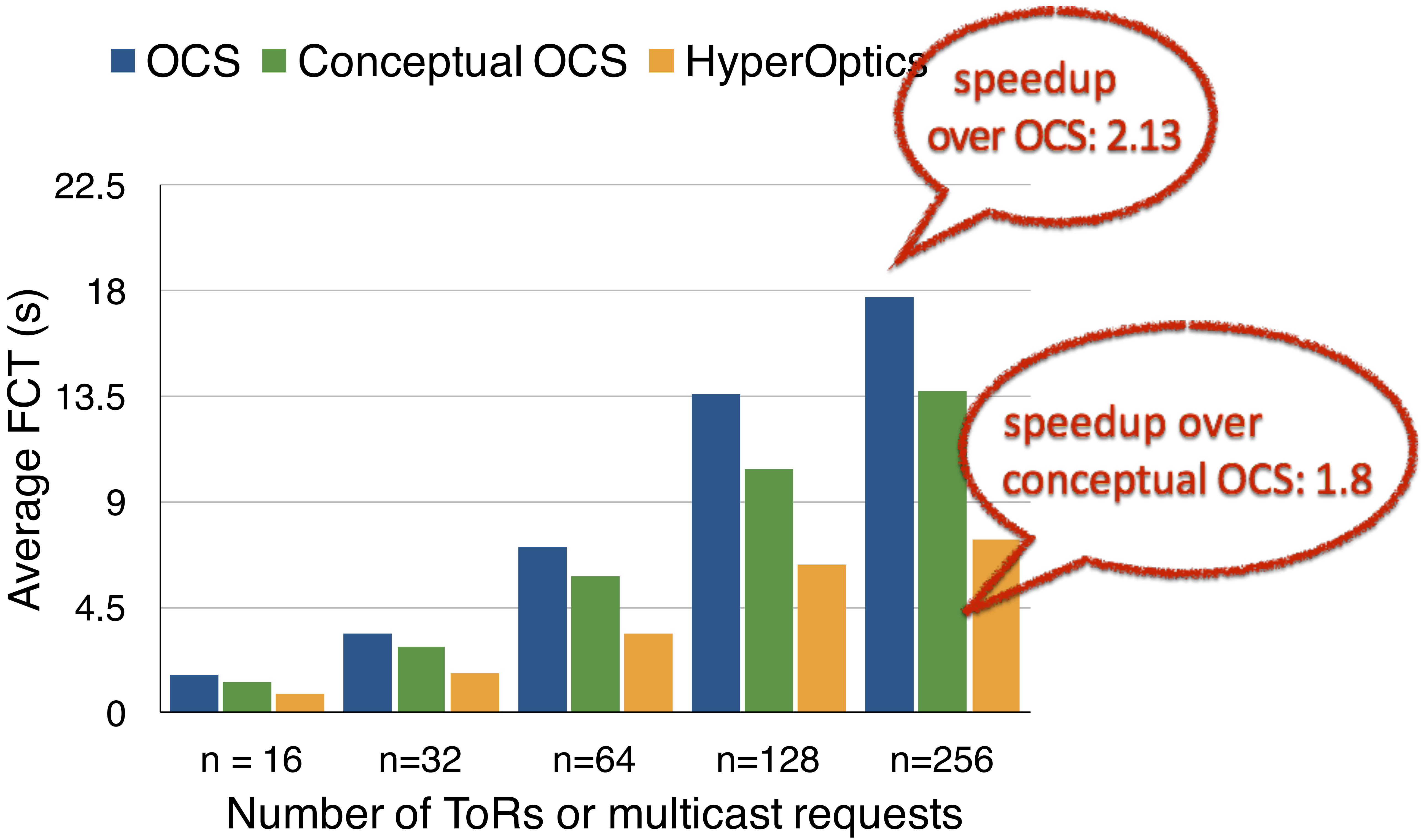
Simulations Setup

1. Synthetic traffic pattern
    - Every rack has one server acting as the multicast source
    - Receivers are a uniform random set of servers in other racks
2. 40Gbps Link Bandwidth
3. Random flow size between 10MB and 1GB

Comparison Bases:

1. OCS network: 320 ports, 25ms topology switching delay
2. Conceptual OCS network: 320 ports, 0 topology switching delay

Results



OCS
Conceptual OCS
HyperOptics
speedup over OCS: 2.13
speedup over conceptual OCS: 1.8
22.5
18
13.5
9
4.5
0
Average FCT (s)
n = 16
n=32
n=64
n=128
n=256
Number of ToRs or multicast requests

Conclusion & Ongoing Work

HyperOptics is an efficient networking architecture for multicast transmissions

- Leverages properties of physical optical technologies
- has novel connectivity structure among ToRs

1. Better Routing & Scheduling Algorithms
2. Fault Tolerance Analysis

Thank you

Dingming.Wu@rice.edu