Programming the Topology of Networks
Technology and Algorithms

Manya Ghobadi

UNIVERSITY OF TORONTO → Google → Microsoft Research → MIT EECS

Pierre Blanche
Jeff Cox
Nikhil Devanur
Mark Filer

Klaus-Tycho Foerster
Jamie Gaudette
Phillipa Gill
Madeleine Glick

Daniel Kilper
Janardhan Kulkarni
Ratul Mahajan
Amar Phanishayee

Gireeja Ranade
Houman Rastegarfar
Stefan Schmid
Rachee Singh

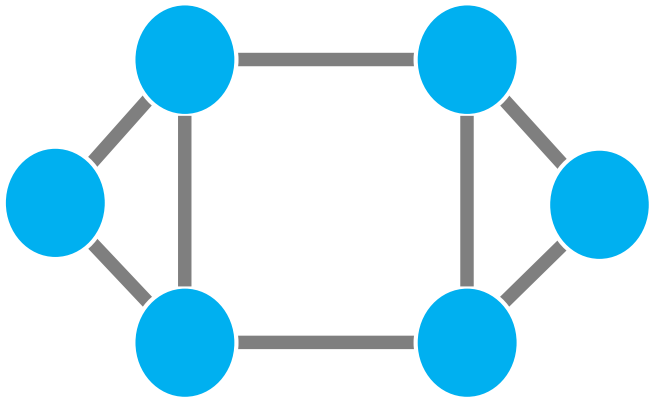
The cloud infrastructure

Inefficiencies and waste in the cloud infrastructure

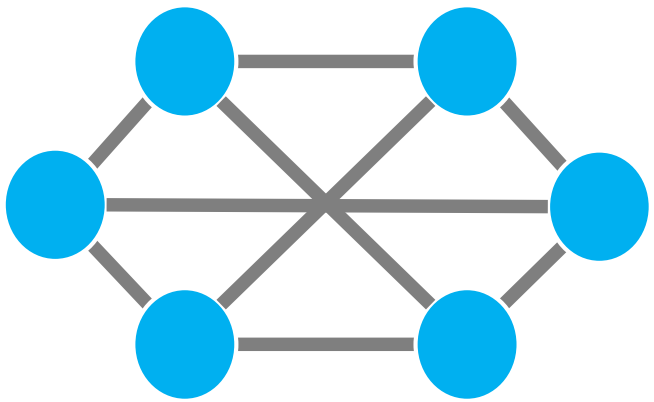
Data centers

Optical cables

Technology and algorithms to optimize network topology



Network topology

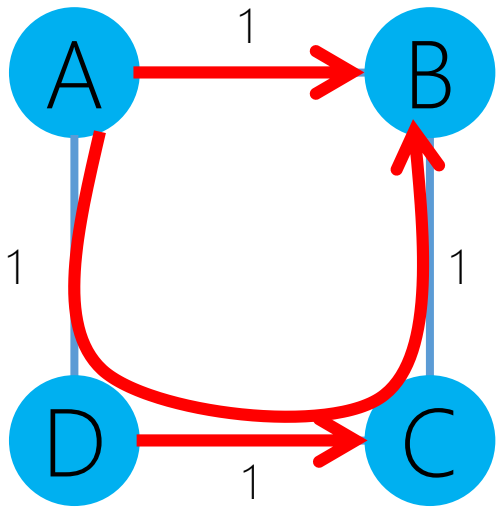


Capacity provisioning

Traffic engineering

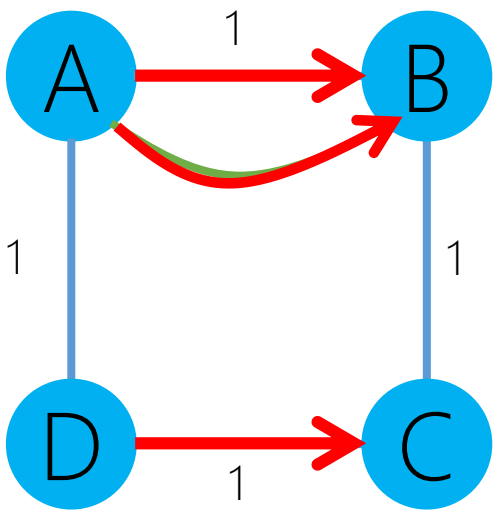
High-level idea

Static topology



Throughput: 2 units

Dynamic topology



Throughput: 3 units

Networks with programmable topologies

Programmable topologies

- Challenging:
  - Requires reconfigurable hardware technology
  - Requires revisiting networking layer algorithms

- Impactful:
  - Cheaper networks
  - Higher throughput

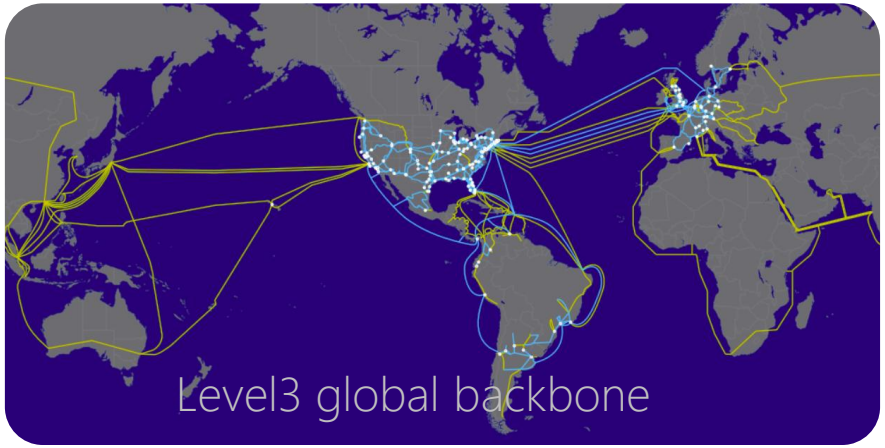
Talk outline

Technology and algorithms to enable programmable topologies in the cloud

Data center networks

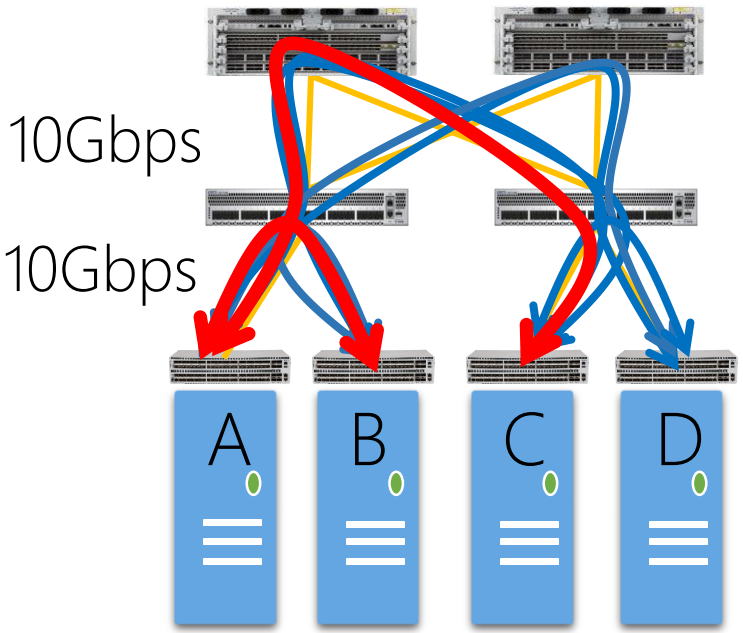
ProjecToR: Programming the network topology [SIGCOMM'16]

Wide-area networks



Programming the capacity of links [SIGCOMM'18, HotNets'17, OFC'16]

Today's data center interconnects



Static capacity between Top-of-Rack (ToR) pairs

	A	B	C	D
A	0	3	3	3
B	3	0	3	3
C	3	3	0	3
D	3	3	3	0

Ideal demand matrix: uniform and static

	A	B	C	D
A	0	6	6	0
B	0	0	0	7
C	8	0	6	7
D	0	12	8	0

Non-ideal demand matrix: skewed and dynamic

Need for a reconfigurable interconnect

Data:

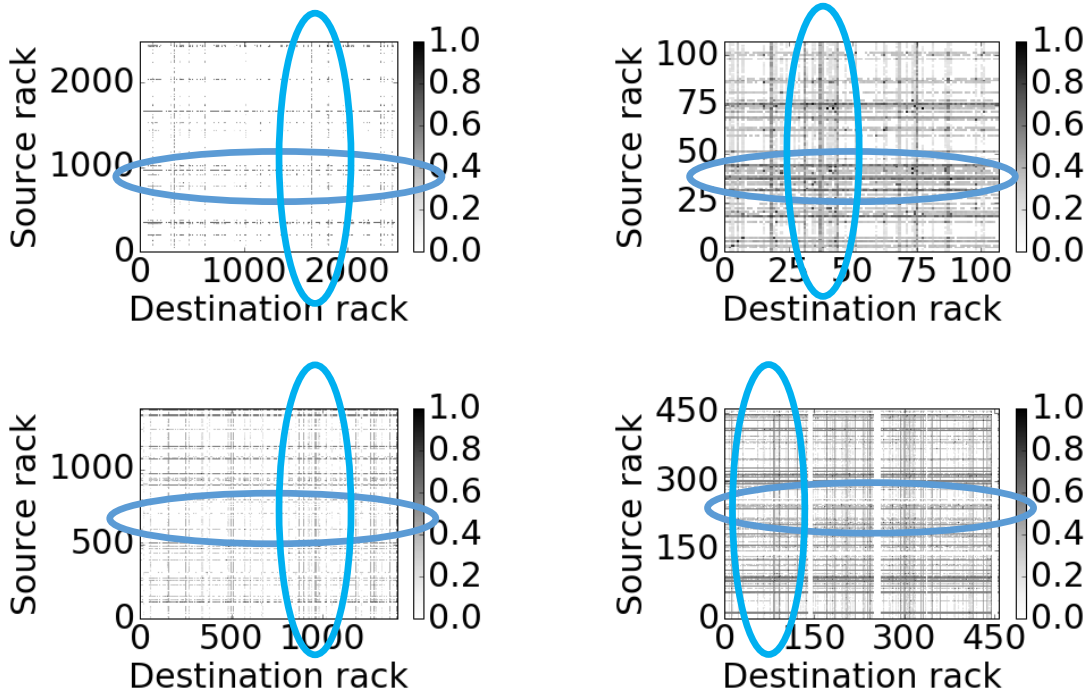
- 200K servers across 4 production clusters
- Cluster sizes: 100 -- 2500 racks

Observation:

- Many rack pairs exchange little traffic
- Only some hot rack pairs are active

Implication:

- Static topology with uniform capacity:
  - Over-provisioned for most rack pairs
  - Under-provisioned for few others

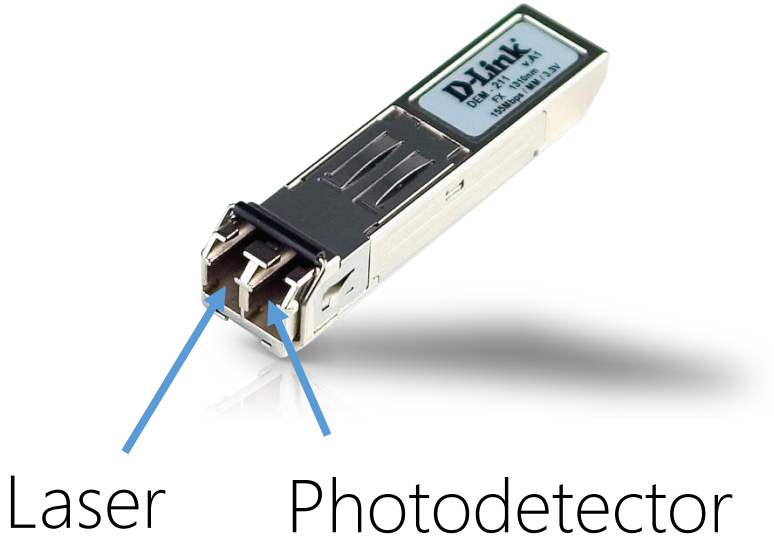


Reconfigurable interconnect:

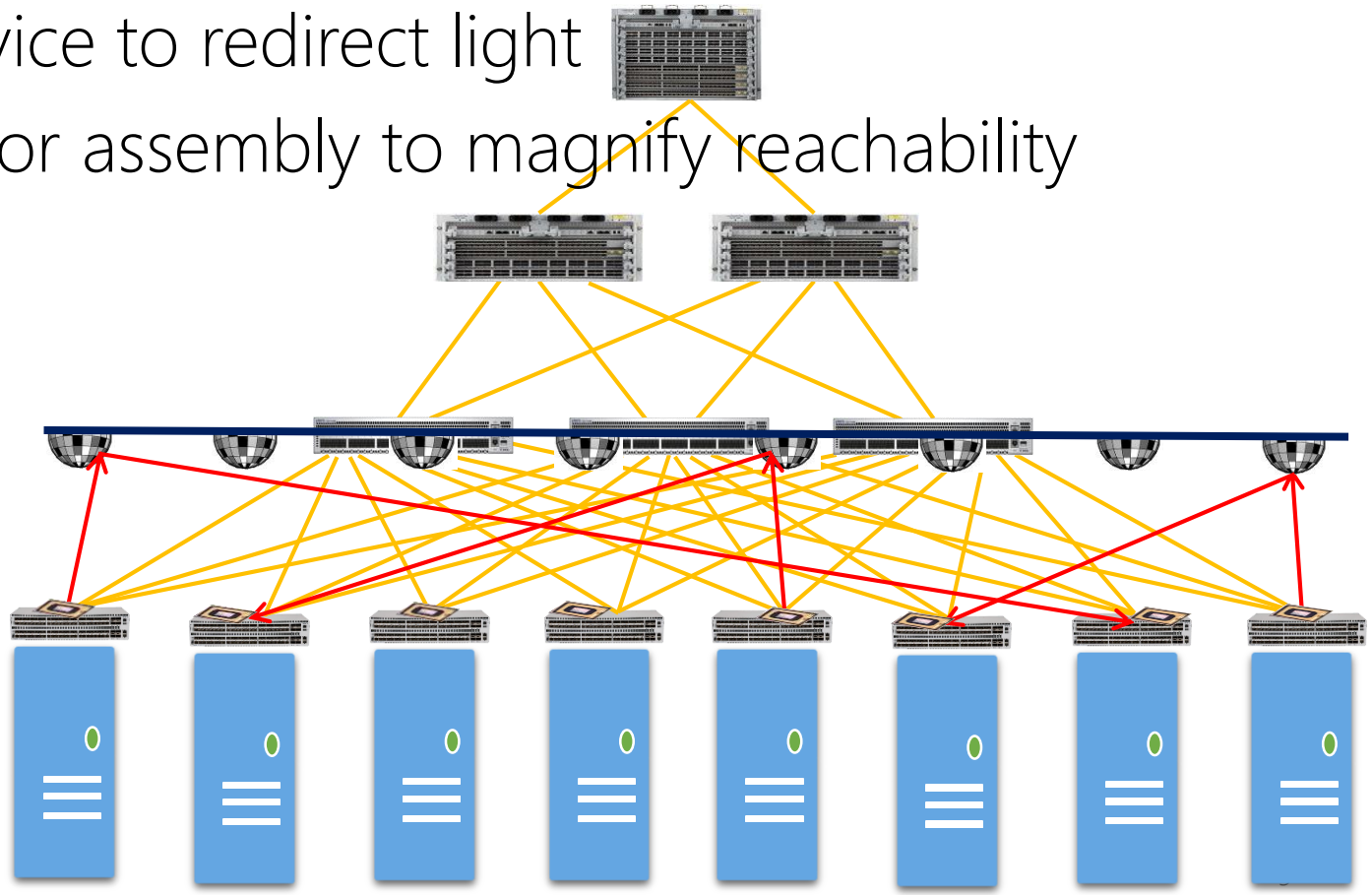
To dynamically provide additional capacity between hot rack pairs

Our proposal: ProjecToR interconnect

- Free-space topology (programmable)
- Digital micromirror device to redirect light
- Disco-ball shaped mirror assembly to magnify reachability

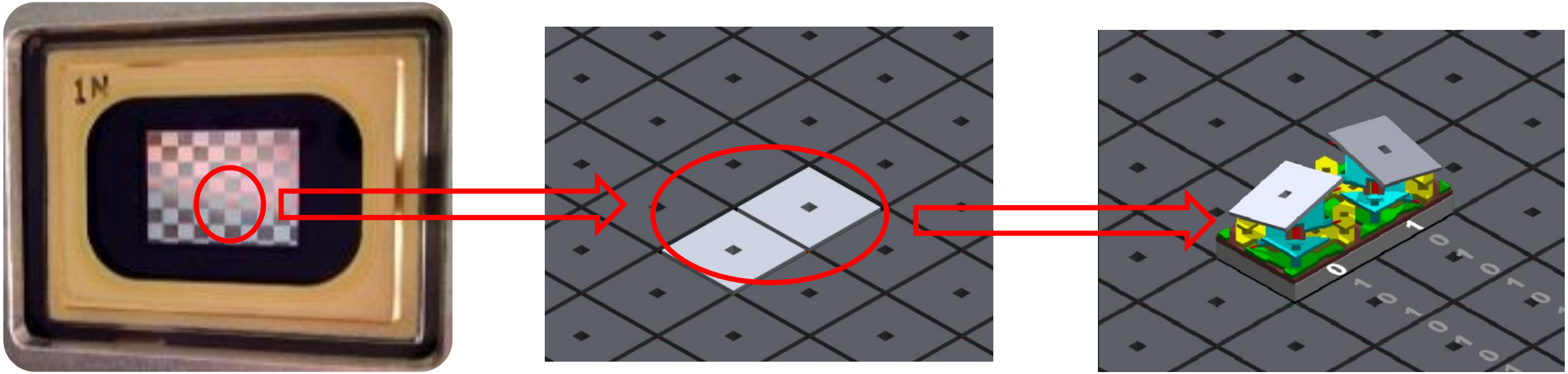


Laser Photodetector



Static topology

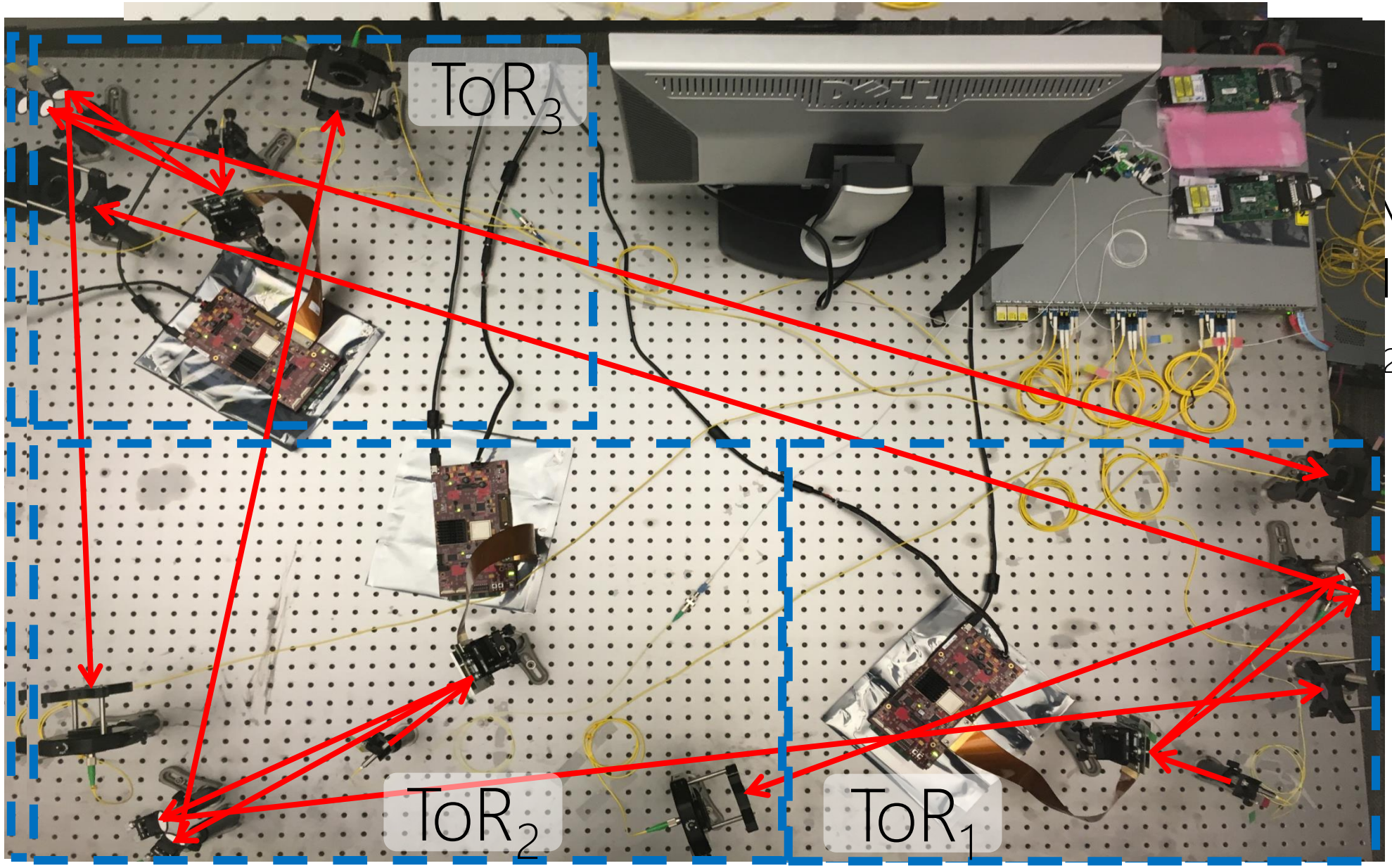
Digital Micromirror Device (DMD)



Array of micromirrors (10 um)

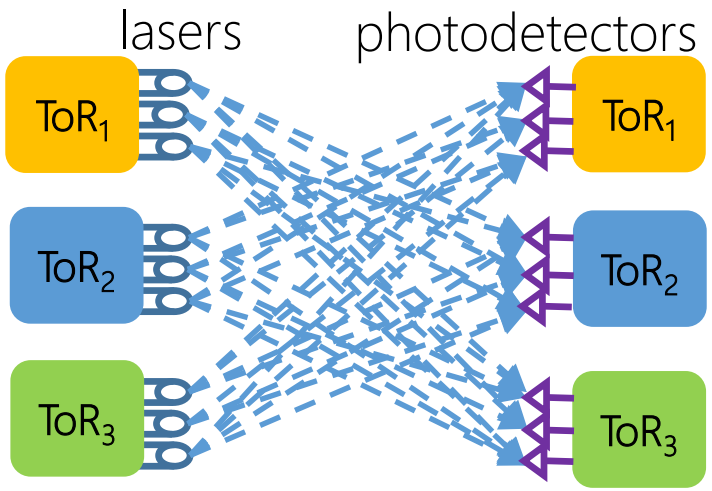
Memory cell

A 3-ToR ProjecToR interconnect prototype



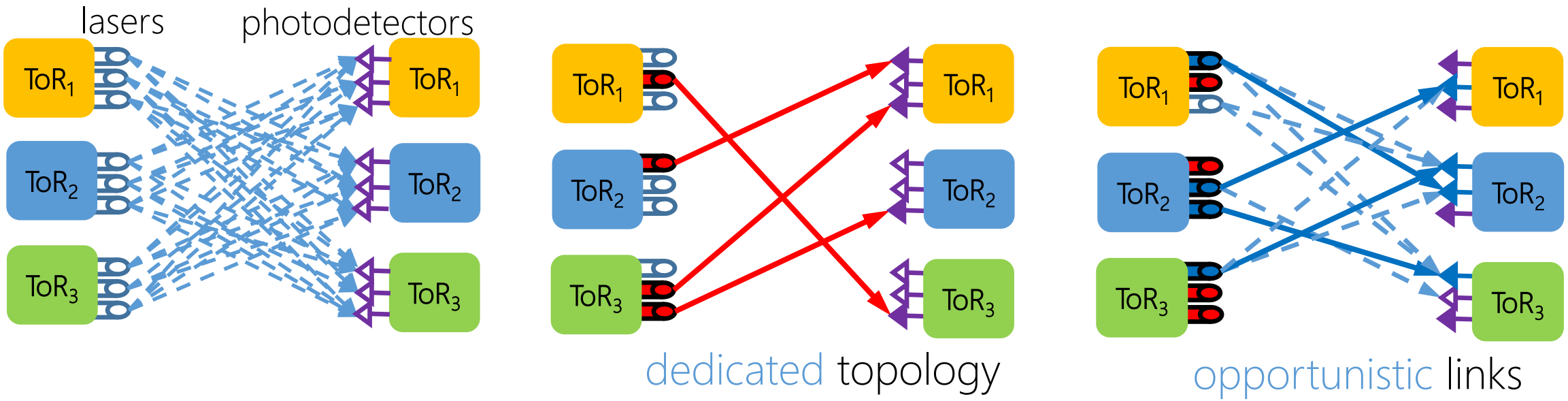
Mirrors
lecting to
$_2$ and ToR$_3$

Routing algorithm



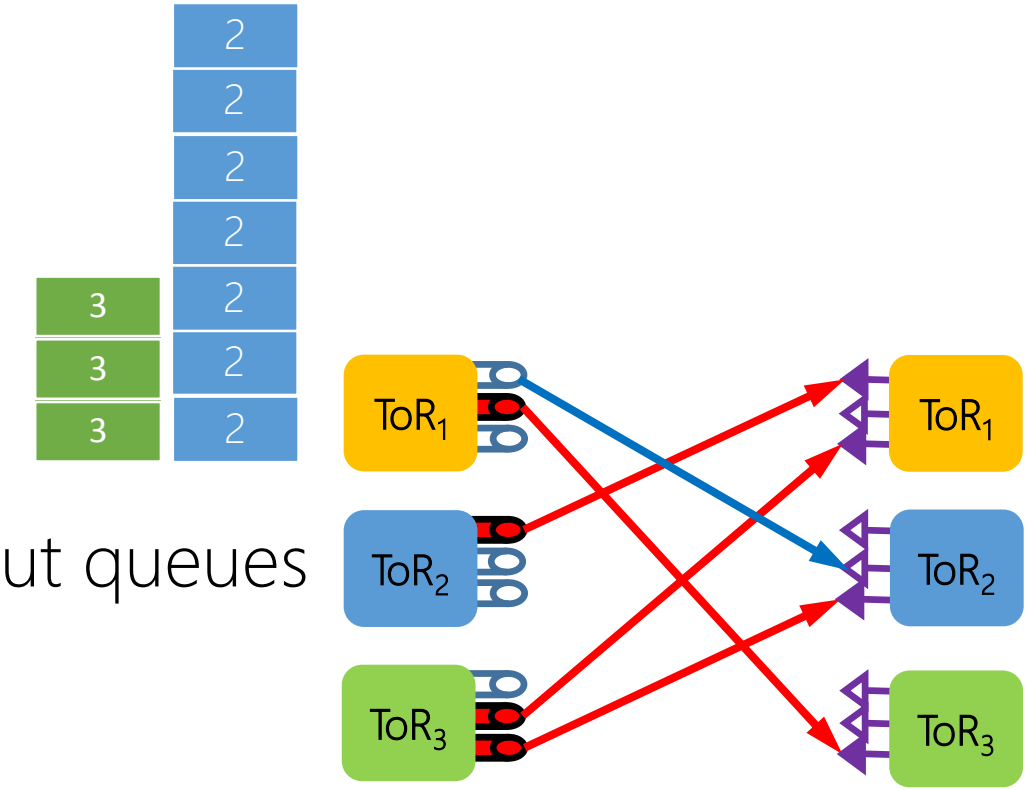
- We have a highly flexible topology allowing for millions of ways to connect lasers to photodetectors
- Ideal solution: fast changing topology to adapt to demand change
- Challenge: It takes 12μs to reprogram a link

Routing algorithm



- Two topology approach:
- Slow switching topology or dedicated topology
- Fast switching links or opportunistic links

Routing packets



Virtual output queues

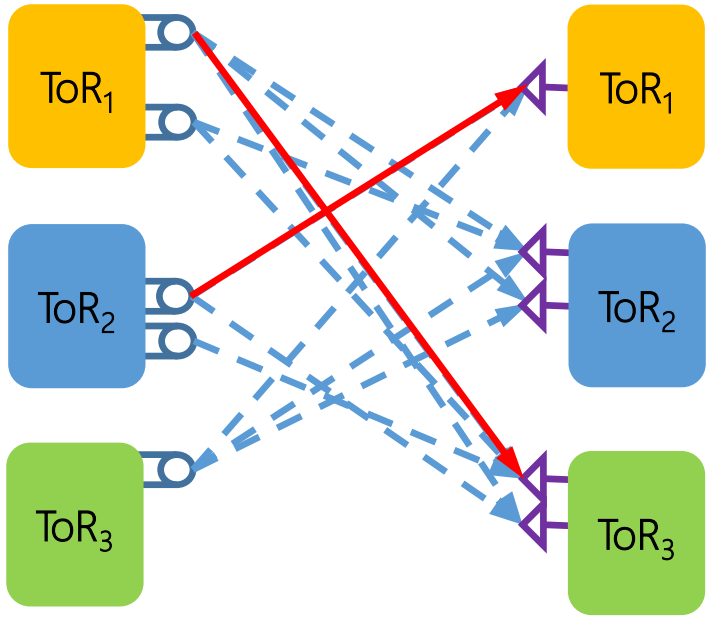
opportunistic link

dedicated topology

K-shortest paths routing

Scheduling opportunistic links

- Given a set of potential links and current traffic demand, find a set of active opportunistic links

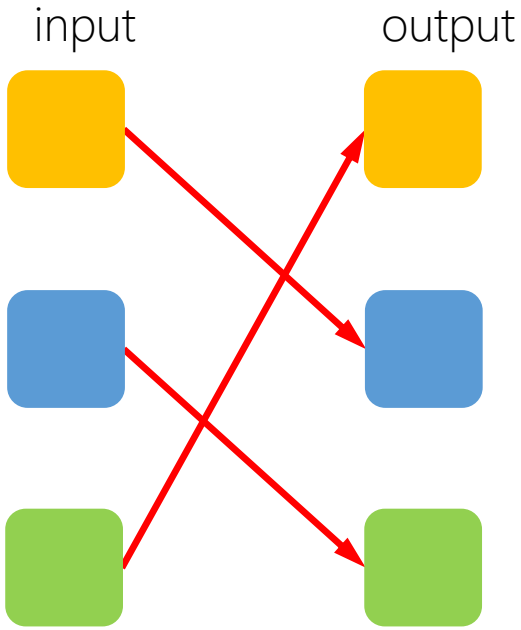


$$\begin{array}{cc} & \text{destination} \\ \text{source} & \begin{bmatrix} 0 & 0 & 100 \\ 100 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{array}$$

Scheduling opportunistic links

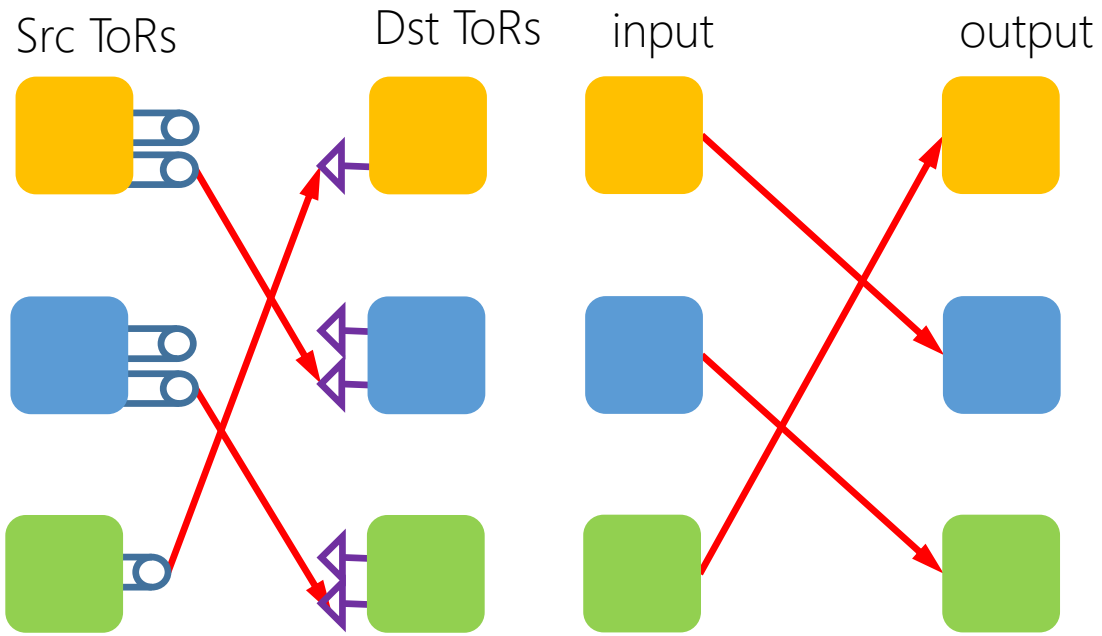
- Standard switch scheduling problem
- Blossom matching
- BvN matrix decomposition
- Centralized scheduler
- Single tiered matching

$$\begin{array}{cc} & \text{destination} \\ \text{source} & \begin{bmatrix} 0 & 100 & 0 \\ 0 & 0 & 100 \\ 100 & 0 & 0 \end{bmatrix} \end{array}$$



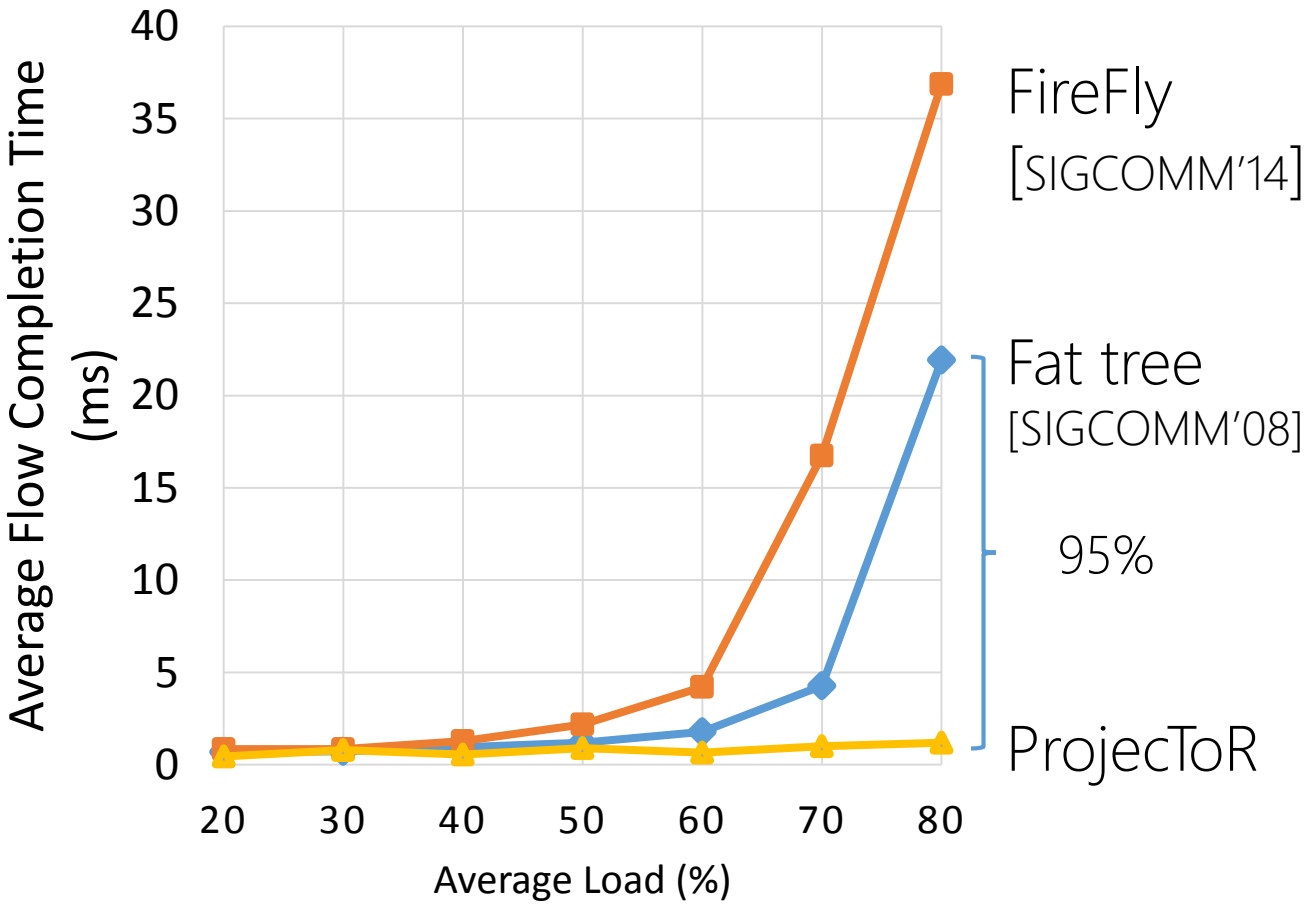
Scheduling opportunistic links

- Standard switch scheduling problem
- Blossom matching
- BvN matrix decomposition
- ~~Centralized~~ Decentralized scheduler
- ~~Single tiered~~ Two-tiered matching



Extended the Gale-Shapely algorithm for finding stable matches [GS-1962]
Constant competitive against an offline optimal allocation

Simulation results



- Slow switching time

- No reconfigurability

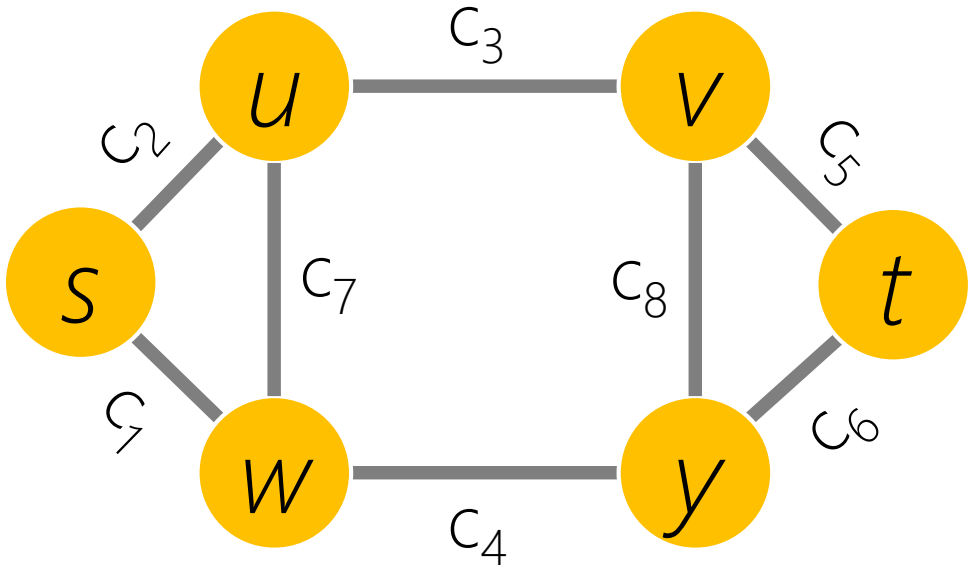
+ Reconfigurable
+ Switching time: 12μs

- Tail flow completion time
- Different traffic matrices
- Impact of switching time
- Impact of fan-out

The key takeaway from this talk

Current assumption: Network topology is fixed

New world: Network topology is dynamic



Problems to solve:

- Scheduling
- Capacity provisioning
- Traffic engineering
- Load-balancing

- **Exciting:** Unusual wealth of algorithms
- **Challenging:** Changes fundamental assumptions
- **Impactful:** Better efficiency ($/Gbps)