ProjecToR: Agile Reconfigurable Data Center Interconnect

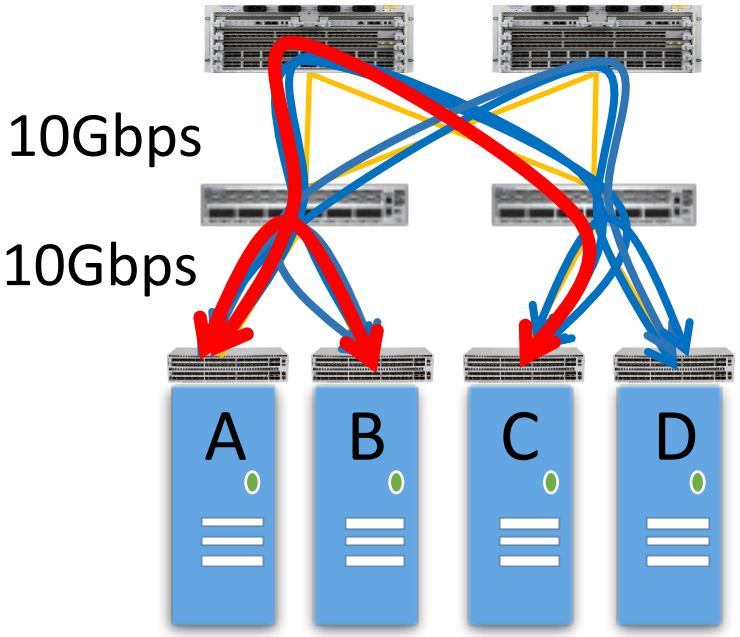
Monia Ghobadi

Ratul Mahajan Amar Phanishayee Pierre Blanche Houman Rastegarfar
Nikhil Devanur Janardhan Kulkarni Madeleine Glick Daniel Kilper
Gireeja Ranade

Microsoft

The University of Arizona

Today's data center interconnects



Static capacity between 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	0
C	0	0	0	0
D	0	12	0	0

Non-ideal demand matrix: skewed and dynamic

Need for a reconfigurable interconnect

Data:

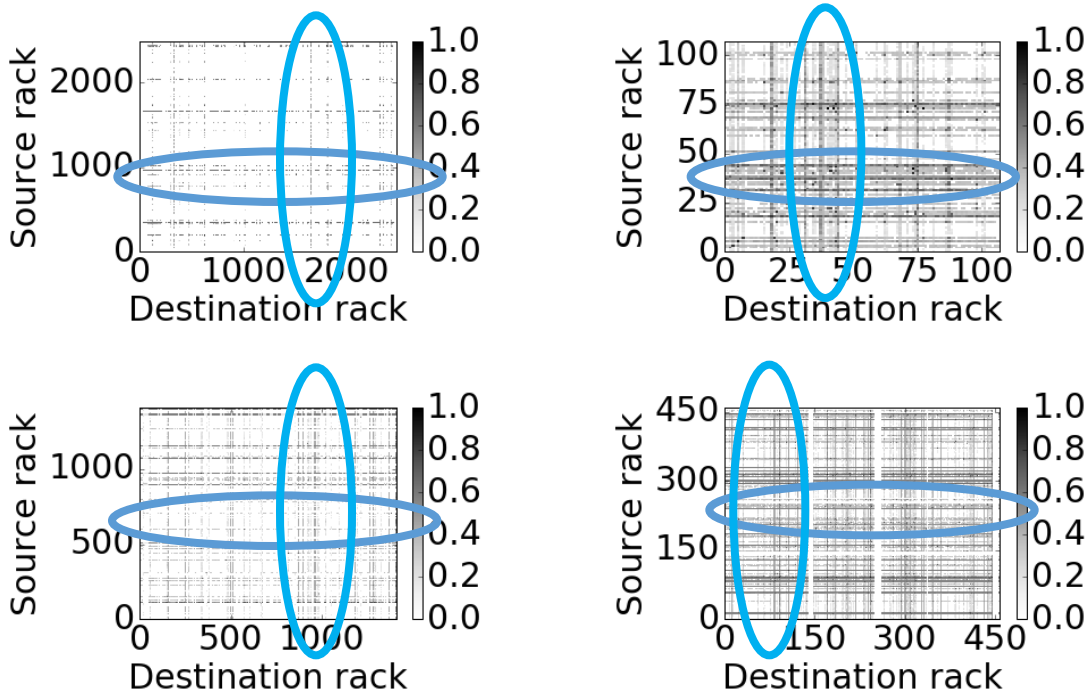
- 200K servers across 4 production clusters
- Cluster sizes: 100 -- 2500 racks

Observation:

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

Implication:

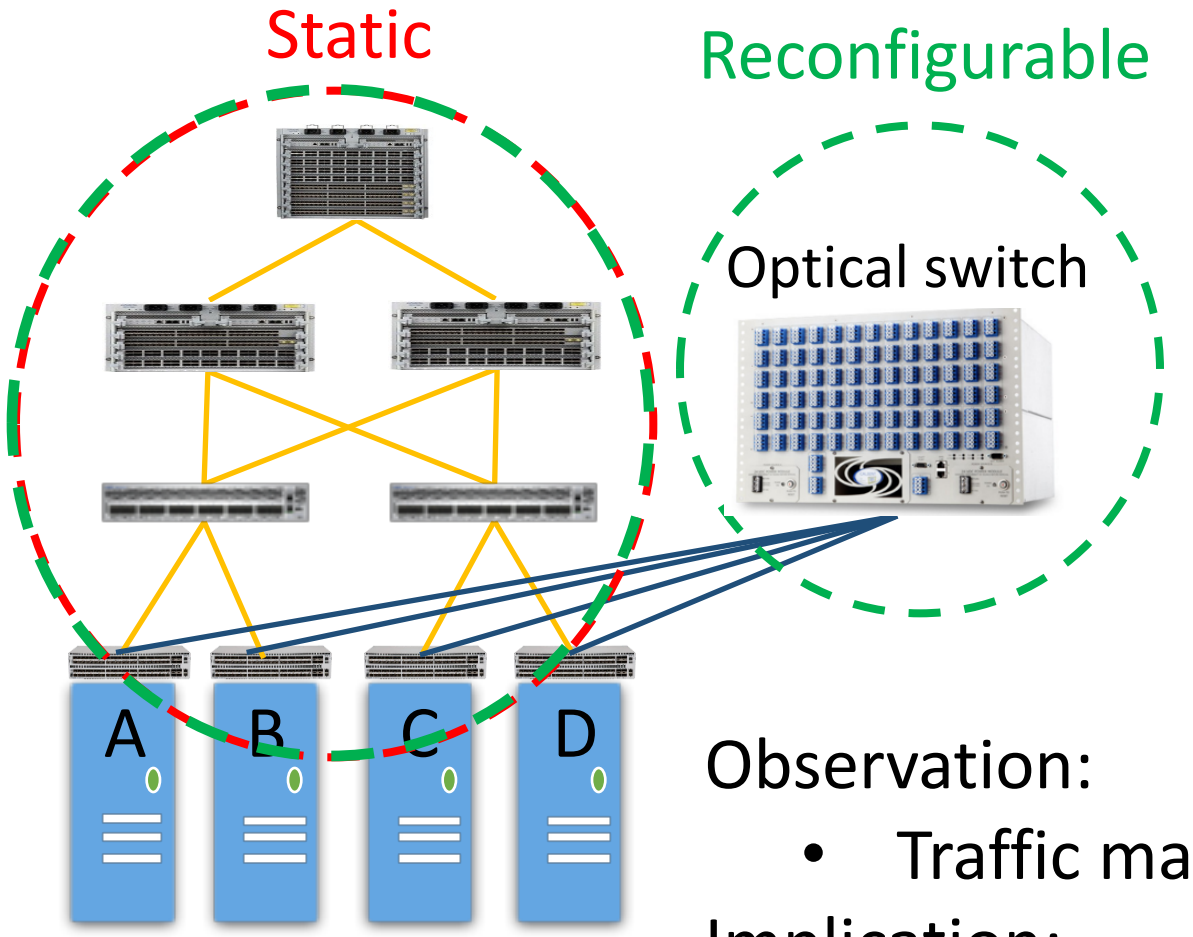
- Static topology with uniform capacity:
  - Over-provisioned for most rack pairs
  - Under-provisioned for few others

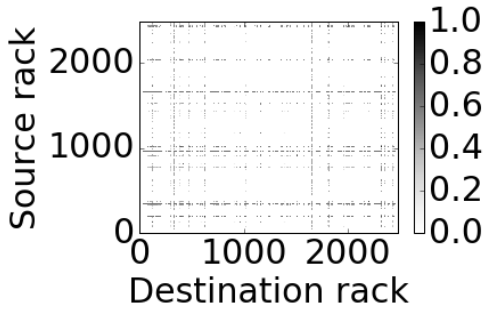


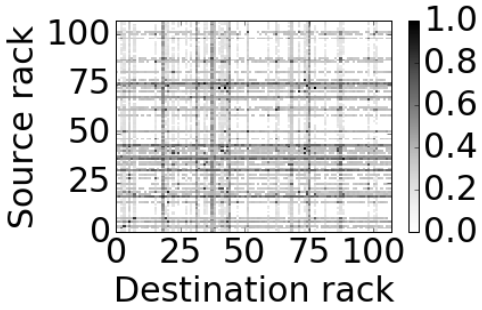
Reconfigurable interconnect:

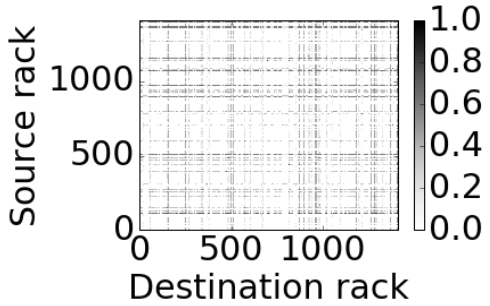
To dynamically provide additional capacity between hot rack pairs

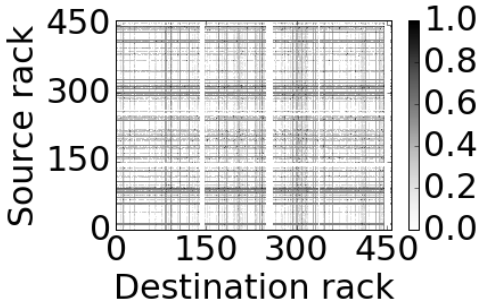
Desirable properties of a reconfigurable interconnect











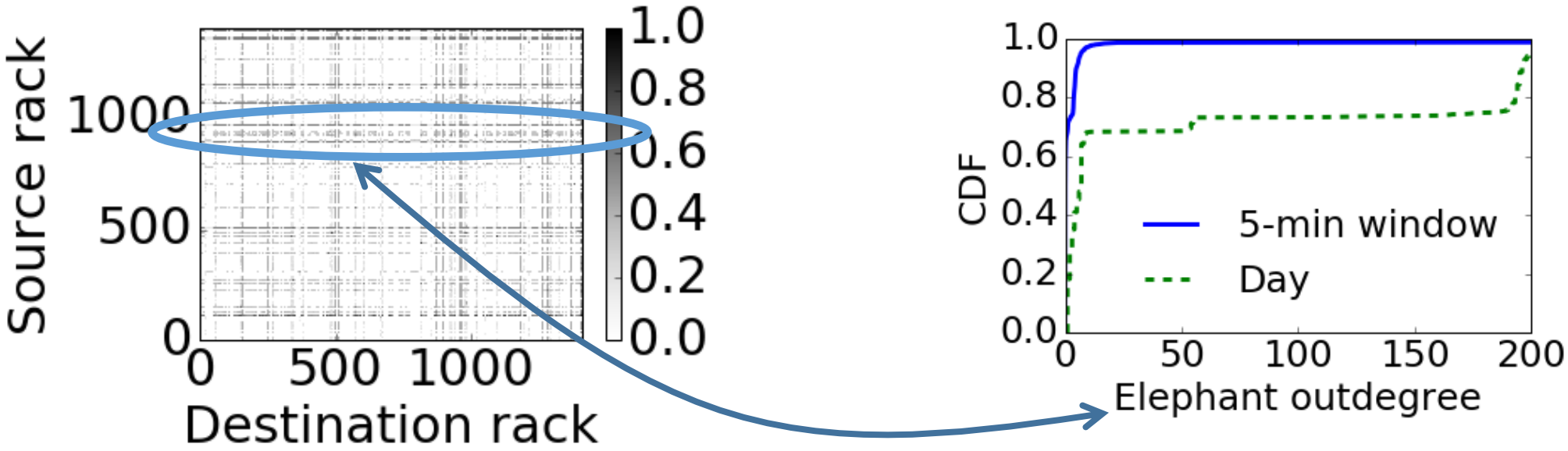
Observation:

- Traffic matrices differ widely

Implication:

- Difficult to determine static vs. reconfigurable divide (Seamless interconnect)

Desirable properties of a reconfigurable interconnect



Observation:

- Source racks send large amounts of traffic to many other racks

Implications:

- Should create direct links to lots of other racks (high fan-out)
- Should switch quickly among destinations (low switching time)

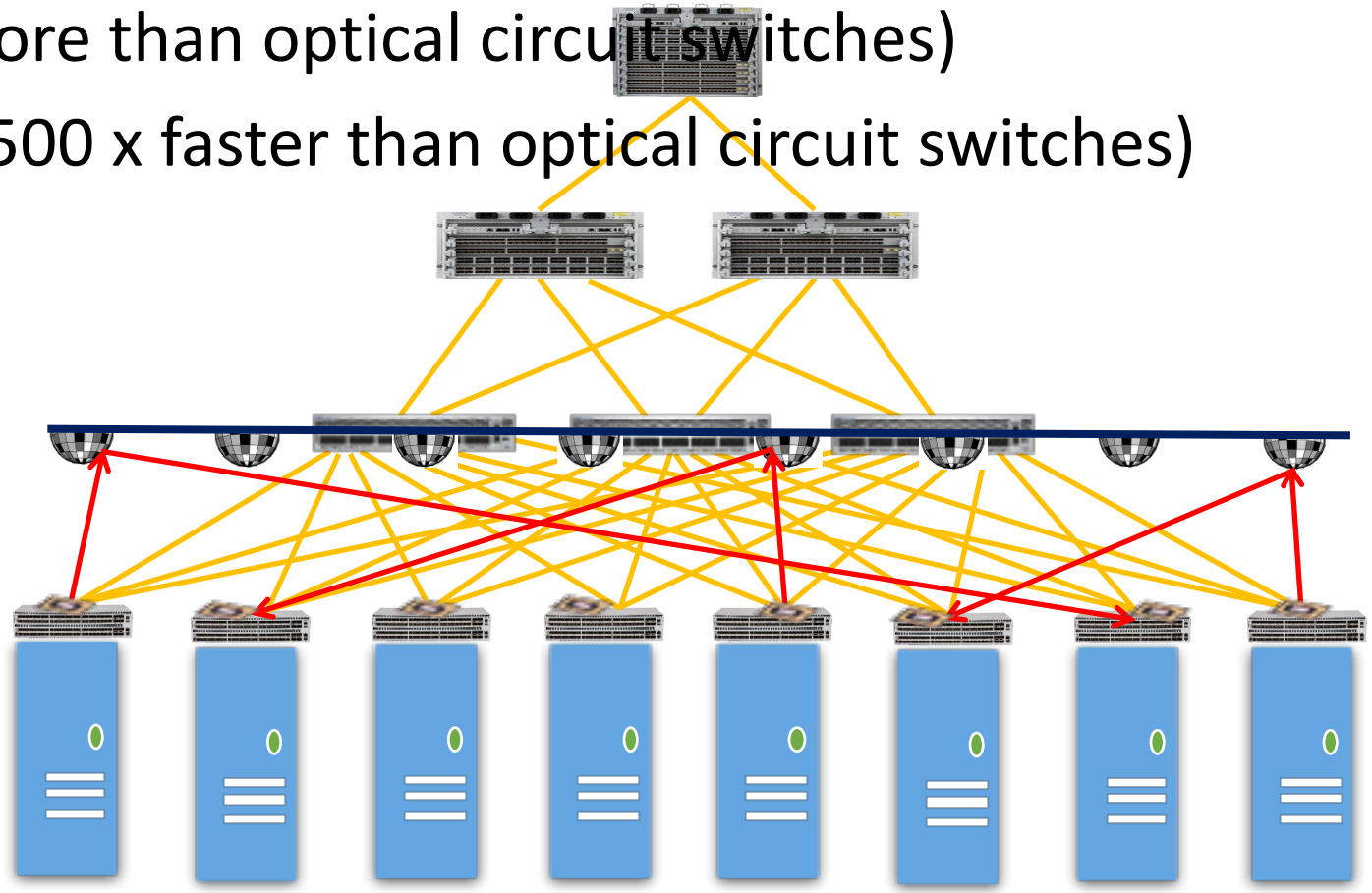
Properties of reconfigurable interconnects

	Enabler technology	Seamless	High Fan-out	Low switching time
Helios, Mordia [sigcomm'10, sigcomm'13]	Optical Circuit Switch	✗	✗	✓
Flyways, 3D Beam forming [sigcomm'11, sigcomm'12]	60GHz	✗	✗	✗
FireFly [sigcomm'14]	Free-Space Optics	✓	✗	✗
ProjecToR	Free-Space Optics	✓	✓	✓

ProjecToR interconnect

- Free-space topology (seamless)
- 18,000 fan-out (60 x more than optical circuit switches)
- 12 us switching time (2500 x faster than optical circuit switches)

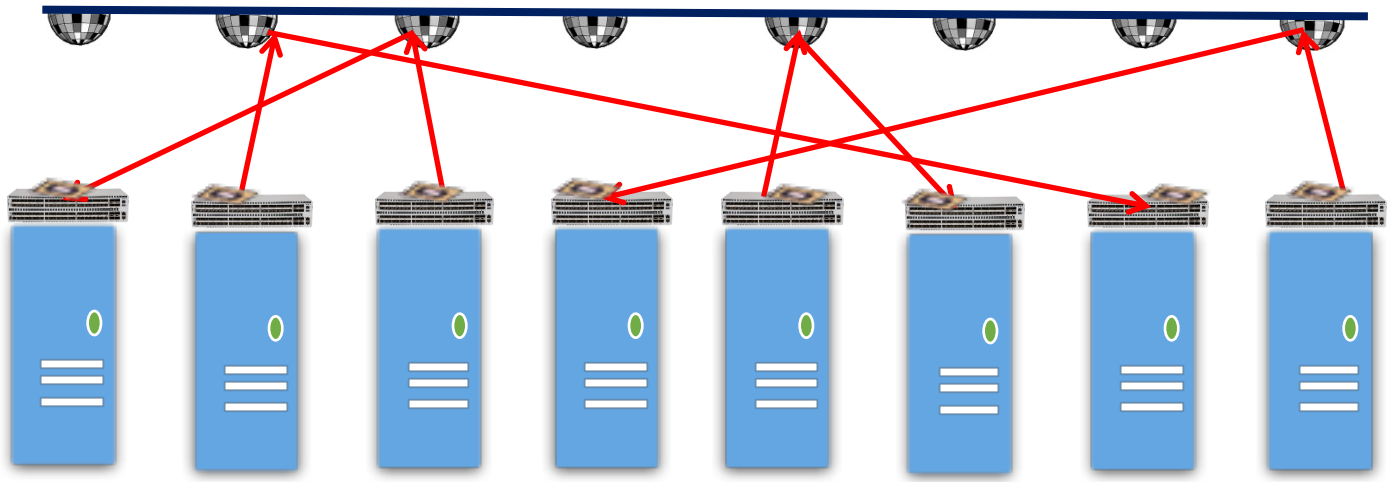
Laser Photodetector



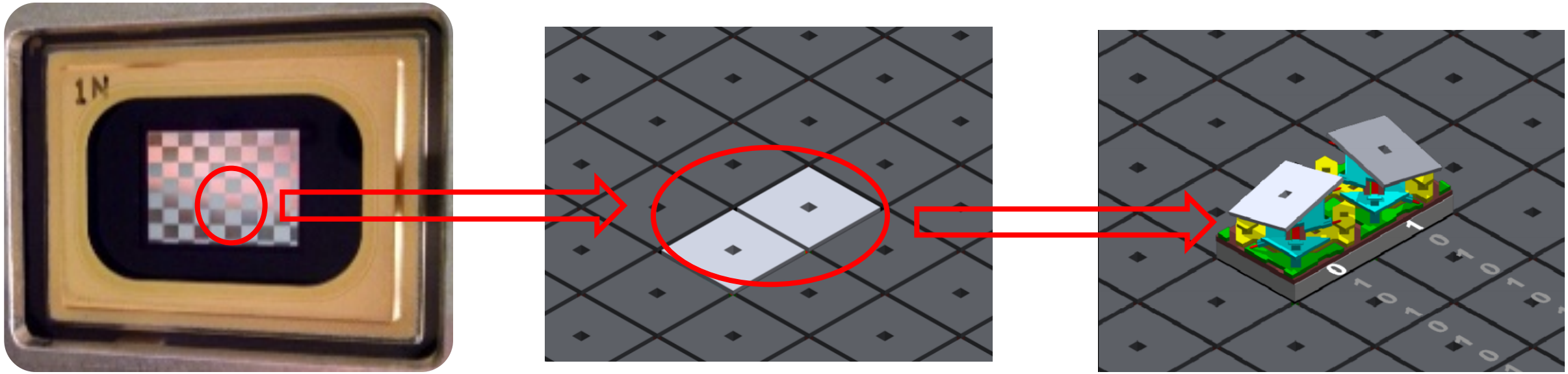
Static topology

Reconfiguration in a ProjecToR interconnect

- Digital micromirror device to redirect light
- Mirror assembly to magnify reach



Digital Micromirror Device (DMD)

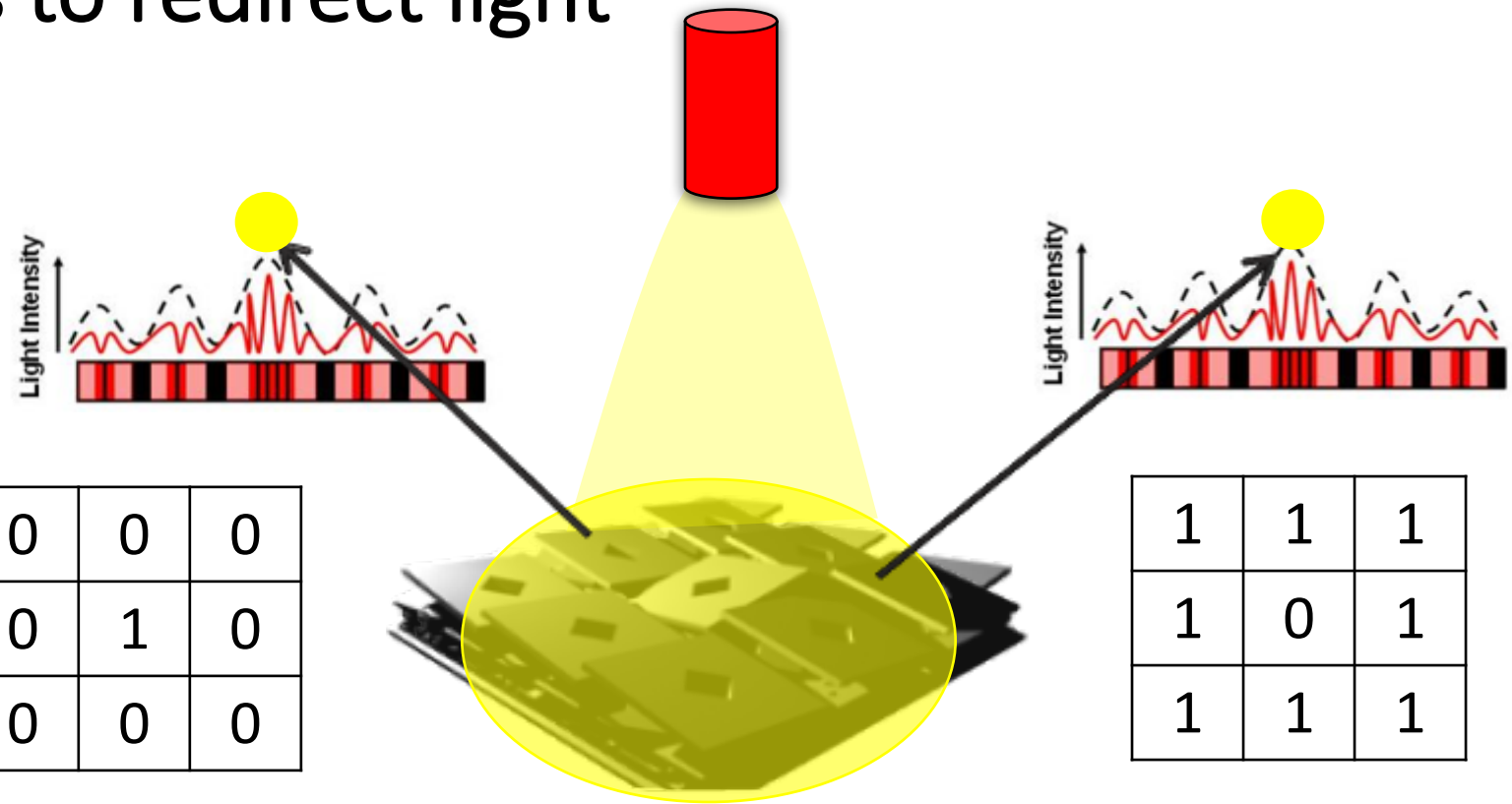


TEXAS INSTRUMENTS

Array of micromirrors (10 um)

Memory cell

Using DMDs to redirect light



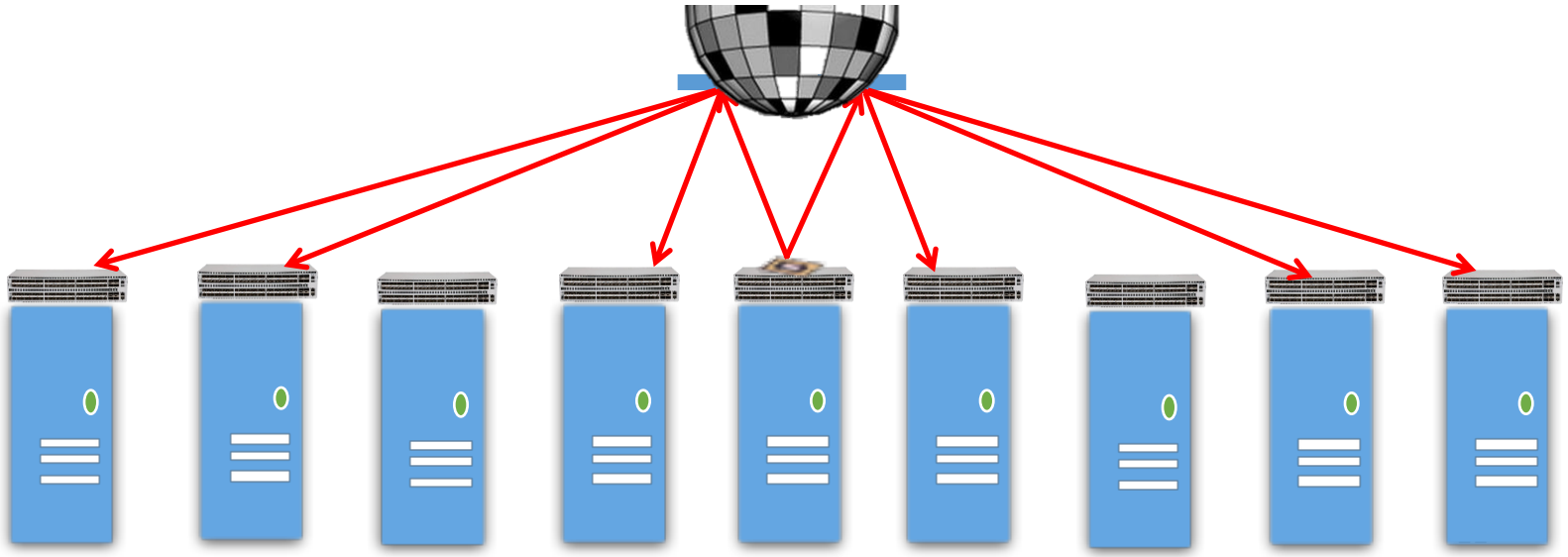
0	0	0
0	1	0
0	0	0

1	1	1
1	0	1
1	1	1

- Theoretical number of accessible locations: total number of micromirrors
  - 768x768 = 589824
- Cross-talk between adjacent locations
- Achievable number of accessible locations
  - 768x768 / 32 = 18,432

Using mirror assemblies to magnify reach

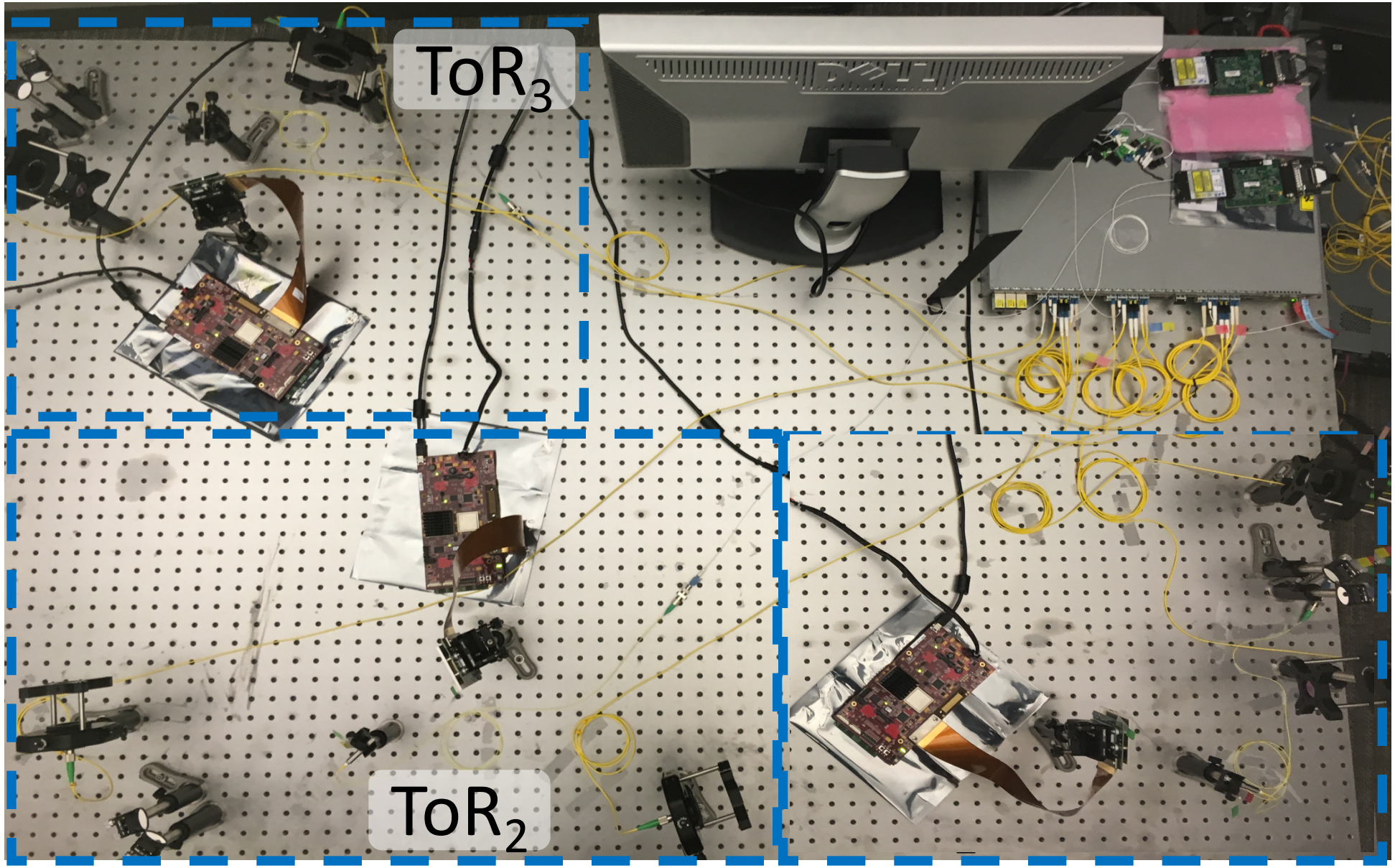
- Challenge: DMDs have a narrow angular reach
- Solution: Coupling DMDs with angled mirrors



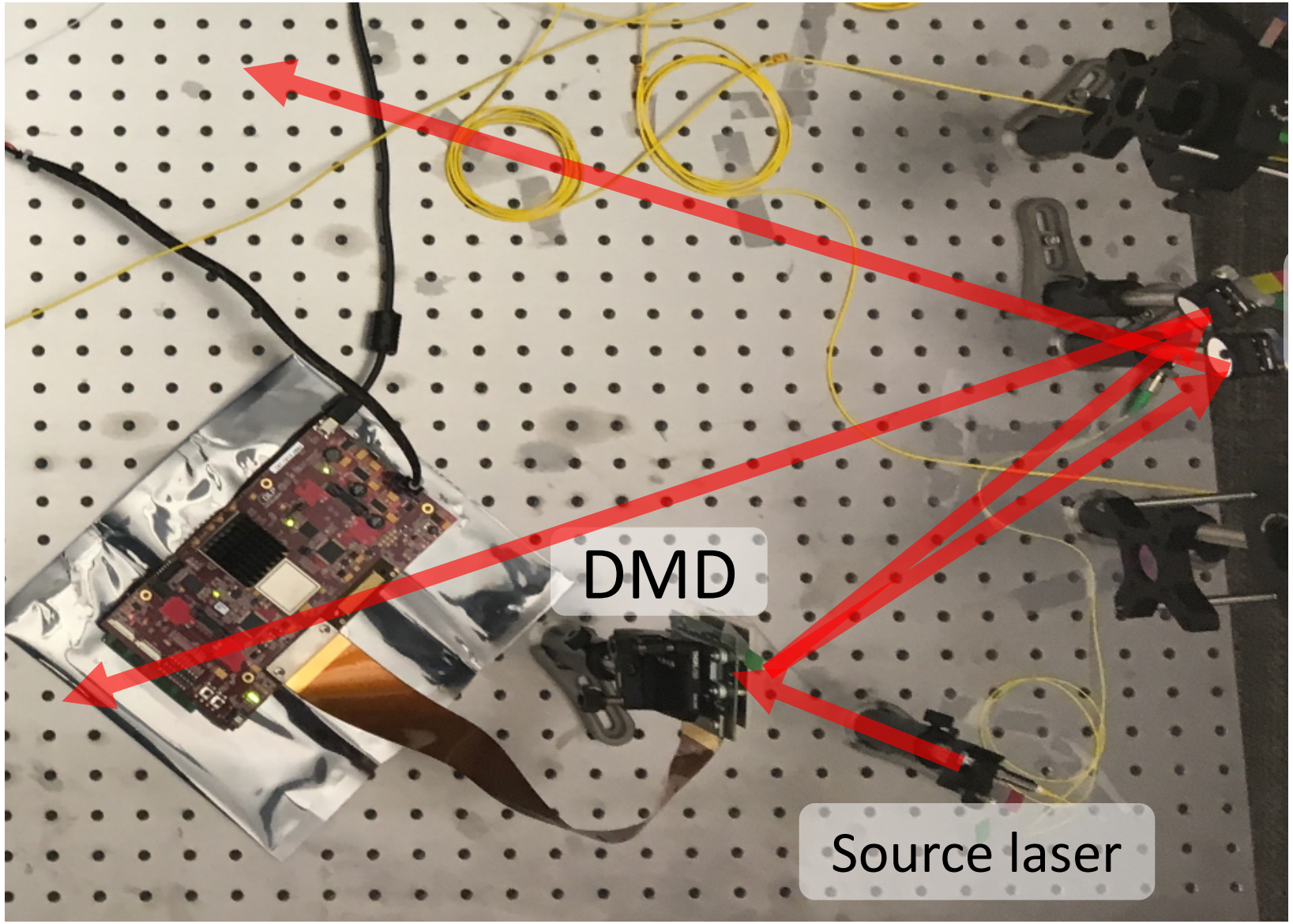
Questions to answer

- How feasible is a ProjecToR interconnect?
  - Built and micro-benchmarked a small ProjecToR prototype
  - Robustness to environmental conditions
- How should packets be routed in a ProjecToR interconnect?
  - Devised a scheduling algorithm and simulated its performance
- How much does a ProjecToR interconnect cost?
  - Estimated cost based on cost break down of each component

Prototype: A 3-ToR ProjecToR interconnect

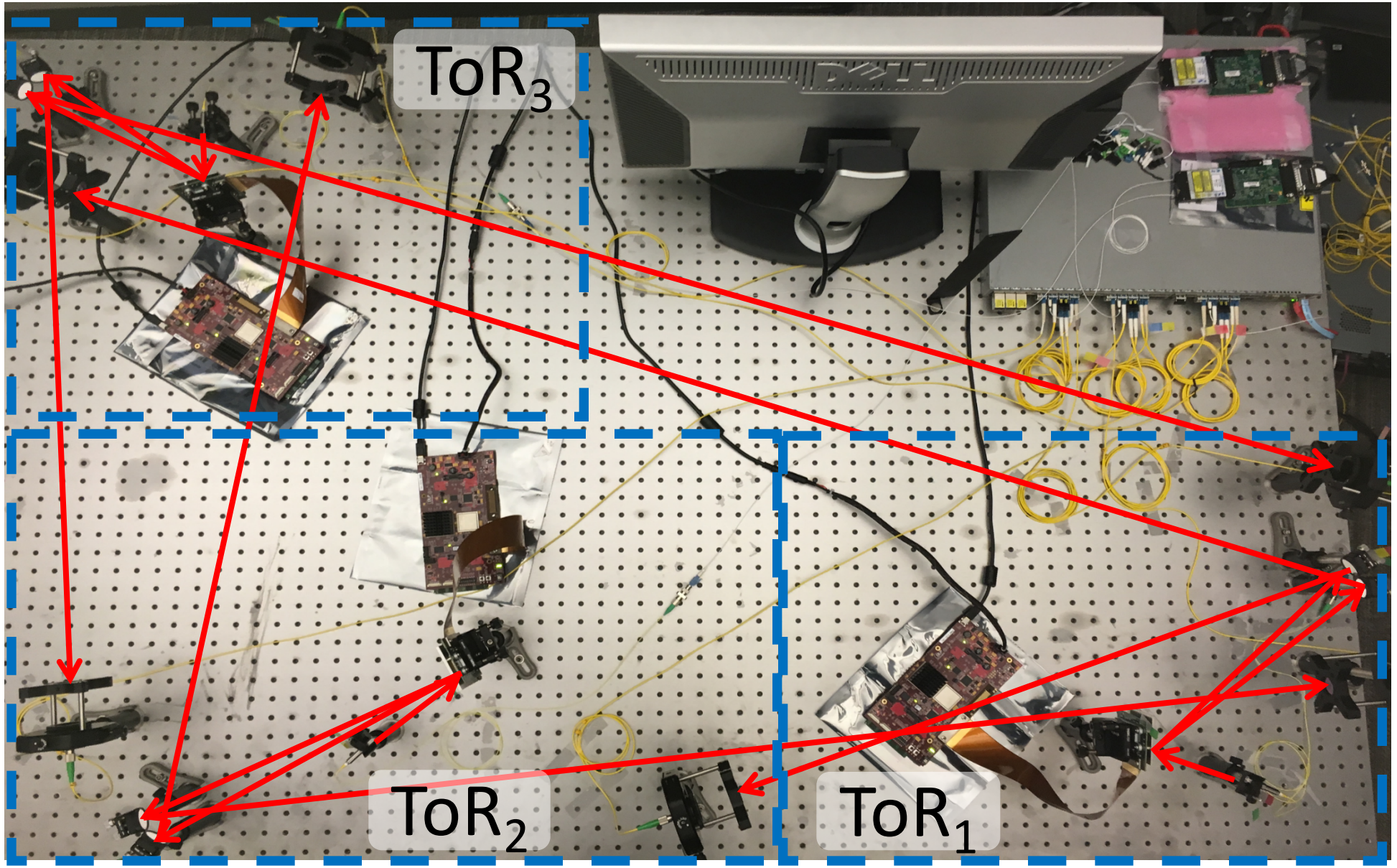


Prototype: A 3-ToR ProjecToR interconnect

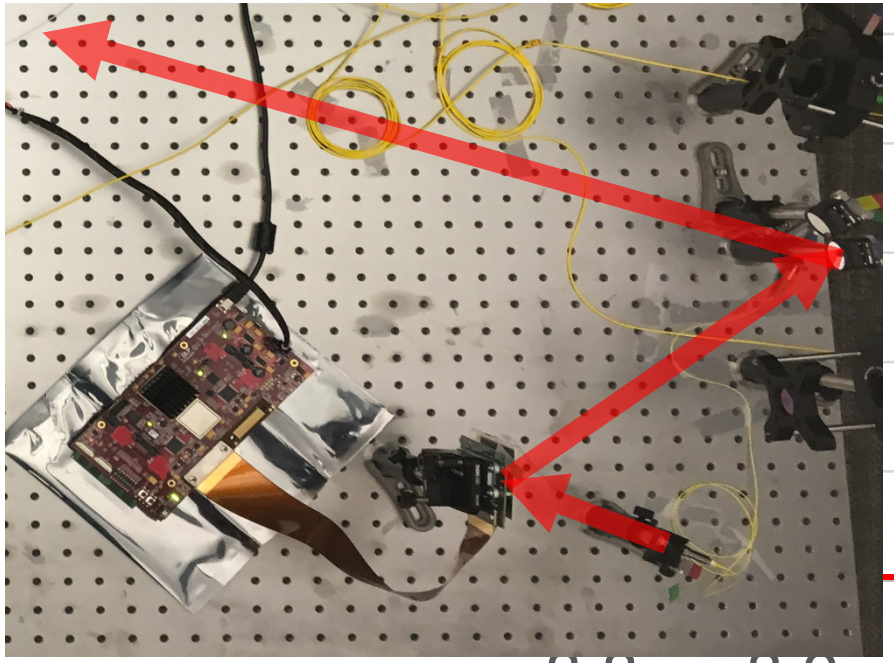


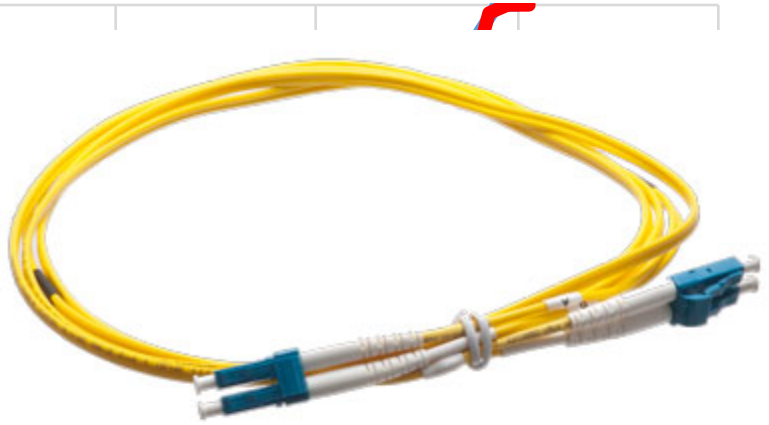
Mirrors reflecting to ToR_2 and ToR_3

Prototype: A 3-ToR ProjecToR interconnect

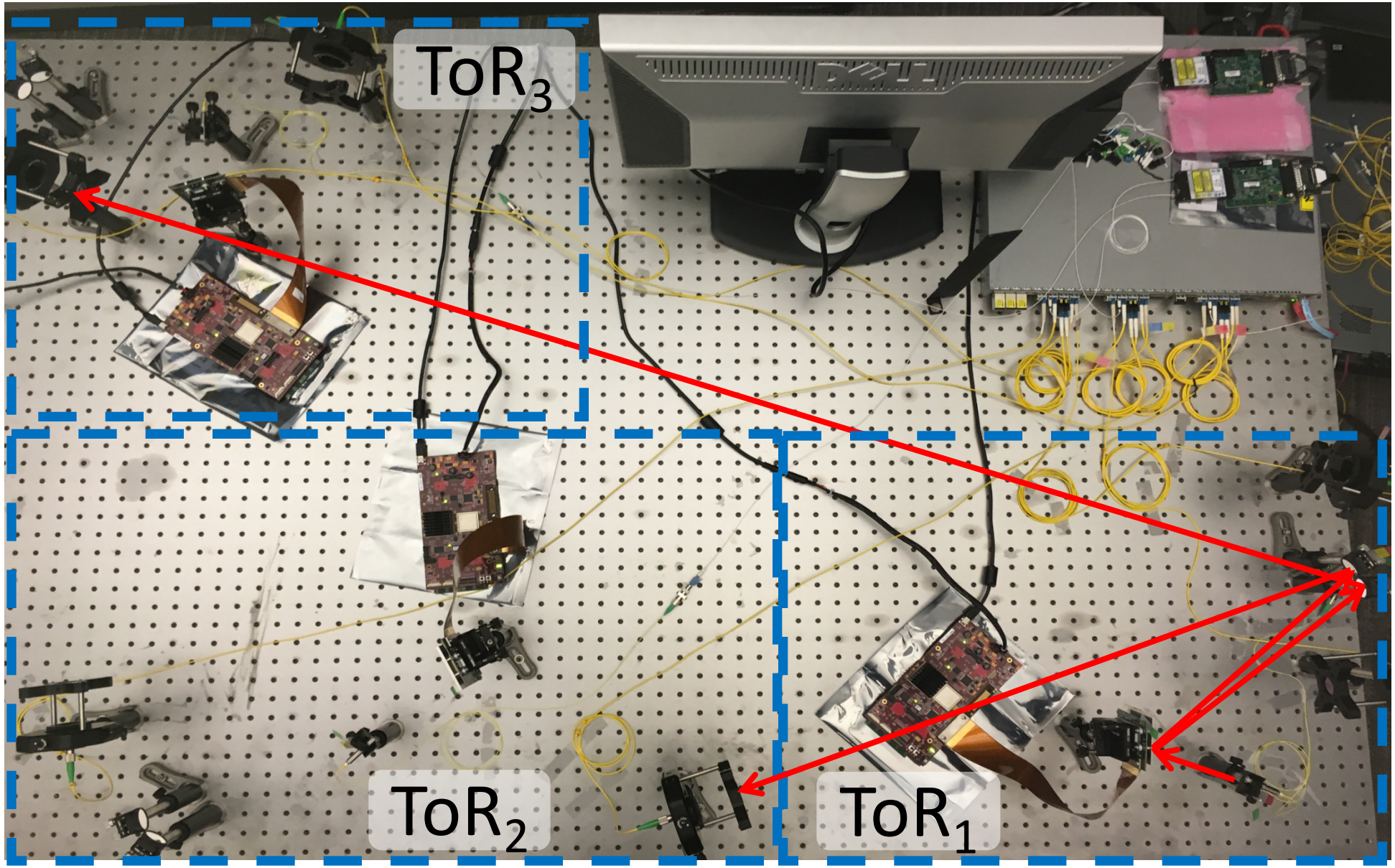


Prototype: throughput

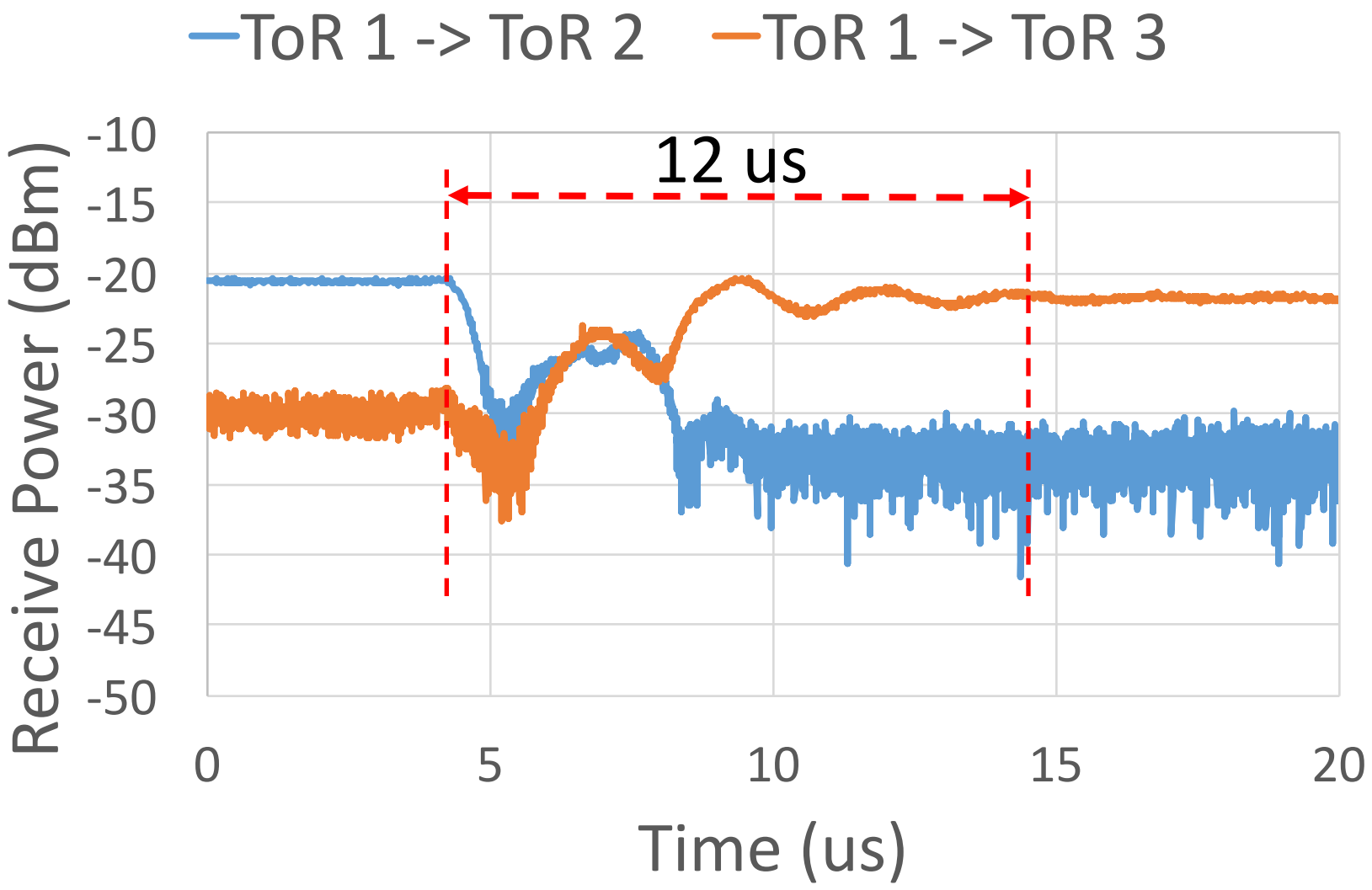




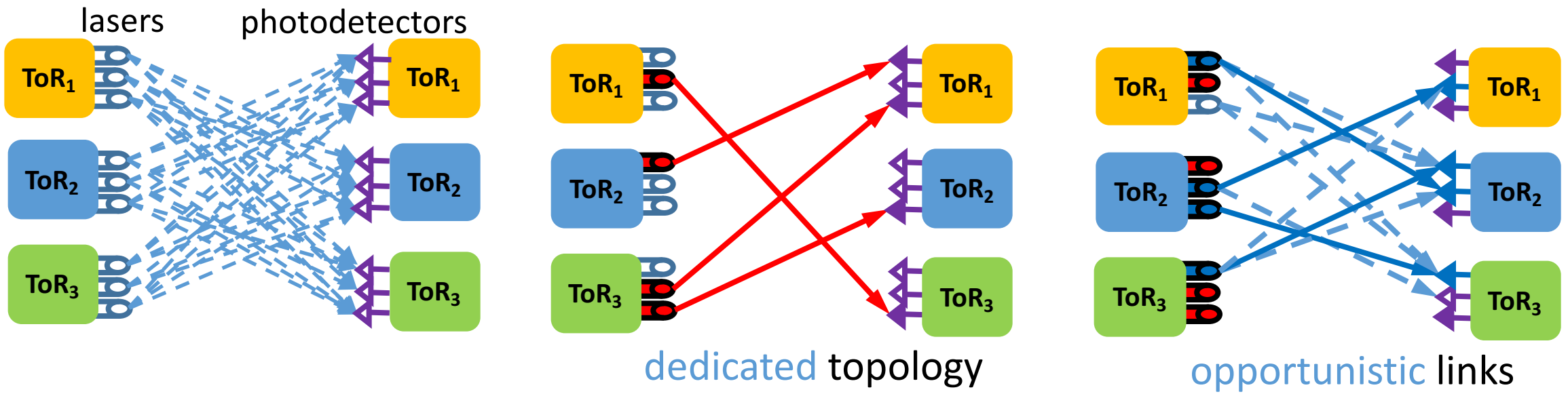
Prototype: switching time



Prototype: switching time

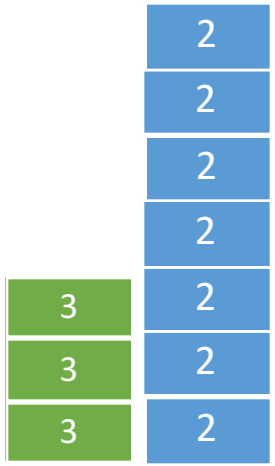


Connecting lasers and photodetectors

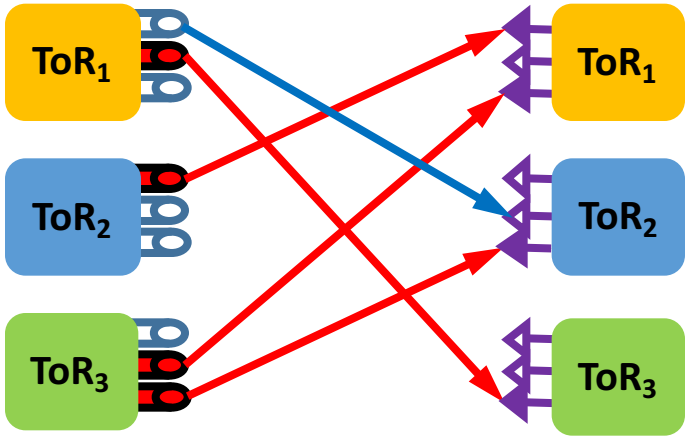


- Two topology approach
  - Slow switching topology or dedicated topology
  - Fast switching links or opportunistic links

Routing packets



Virtual output queues



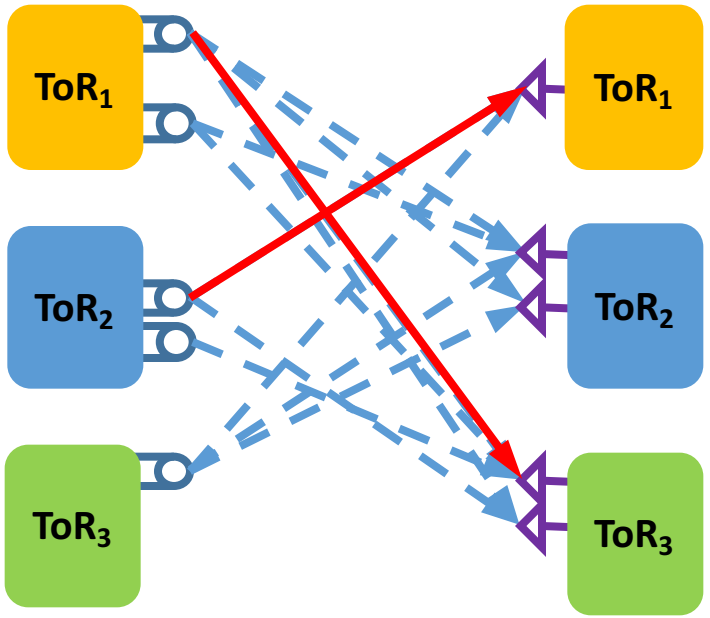
opportunistic link

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

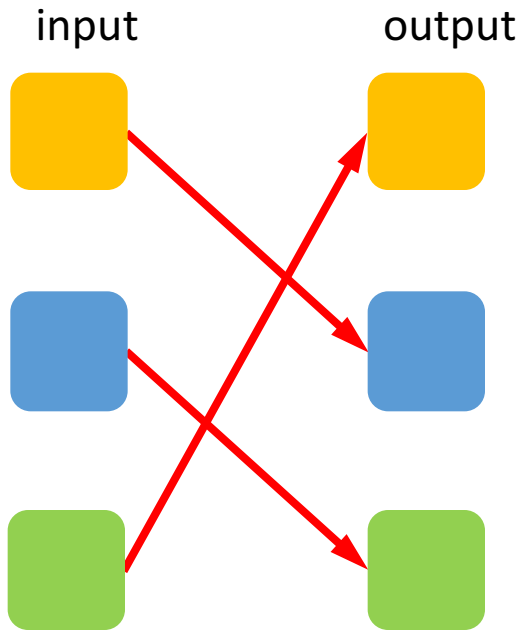


source \ destination			
	0	0	100
	100	0	0
	0	0	0

Scheduling opportunistic links

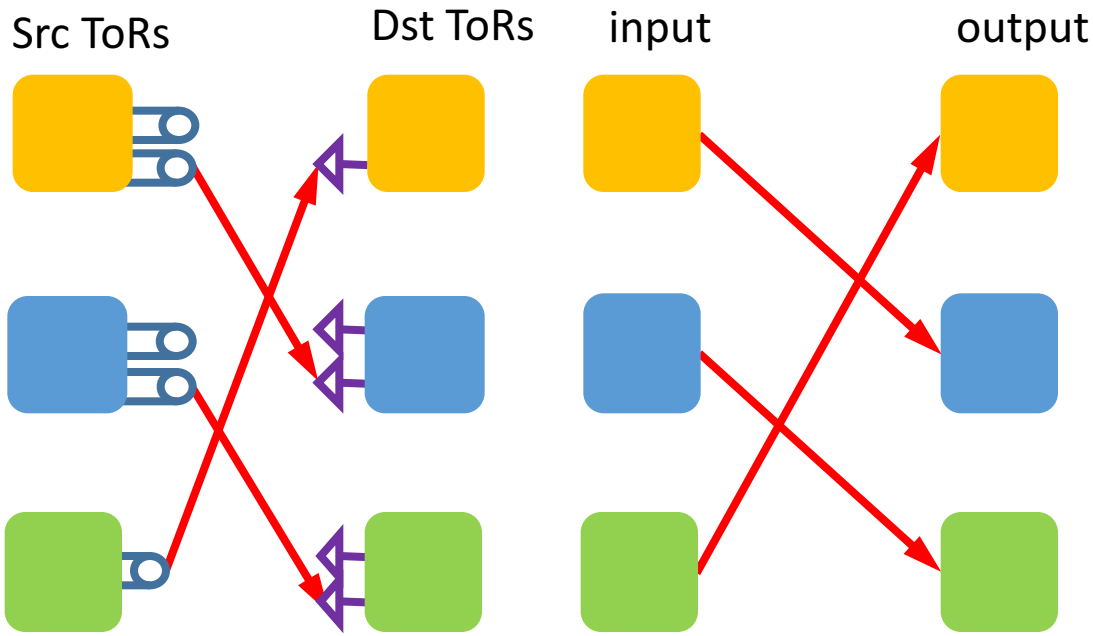
- Standard switch scheduling problem
- Blossom matching
- Matrix decomposition
- Centralized scheduler
- Single tiered matching

source \ destination			
	0	100	0
	0	0	100
	100	0	0



Scheduling opportunistic links

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

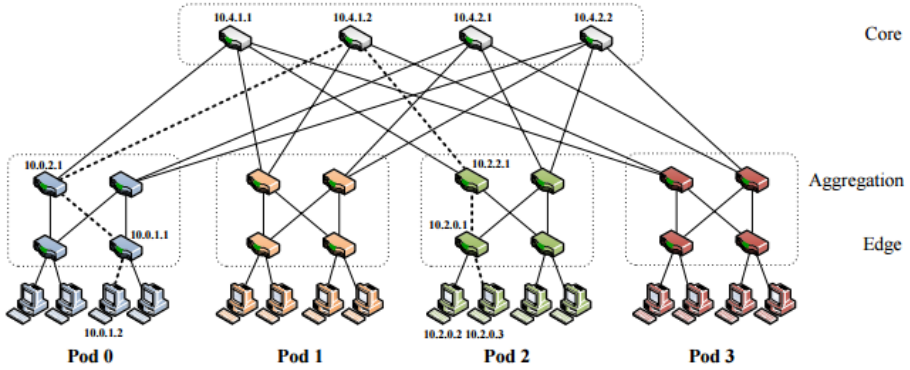


Extended the Gale-Shapely algorithm for finding stable matches [GS-1962]

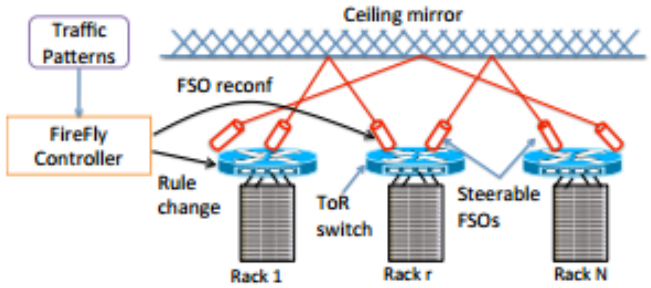
Constant competitive against an offline optimal allocation

Simulations

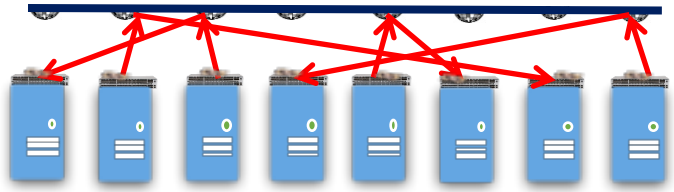
Fat tree



FireFly

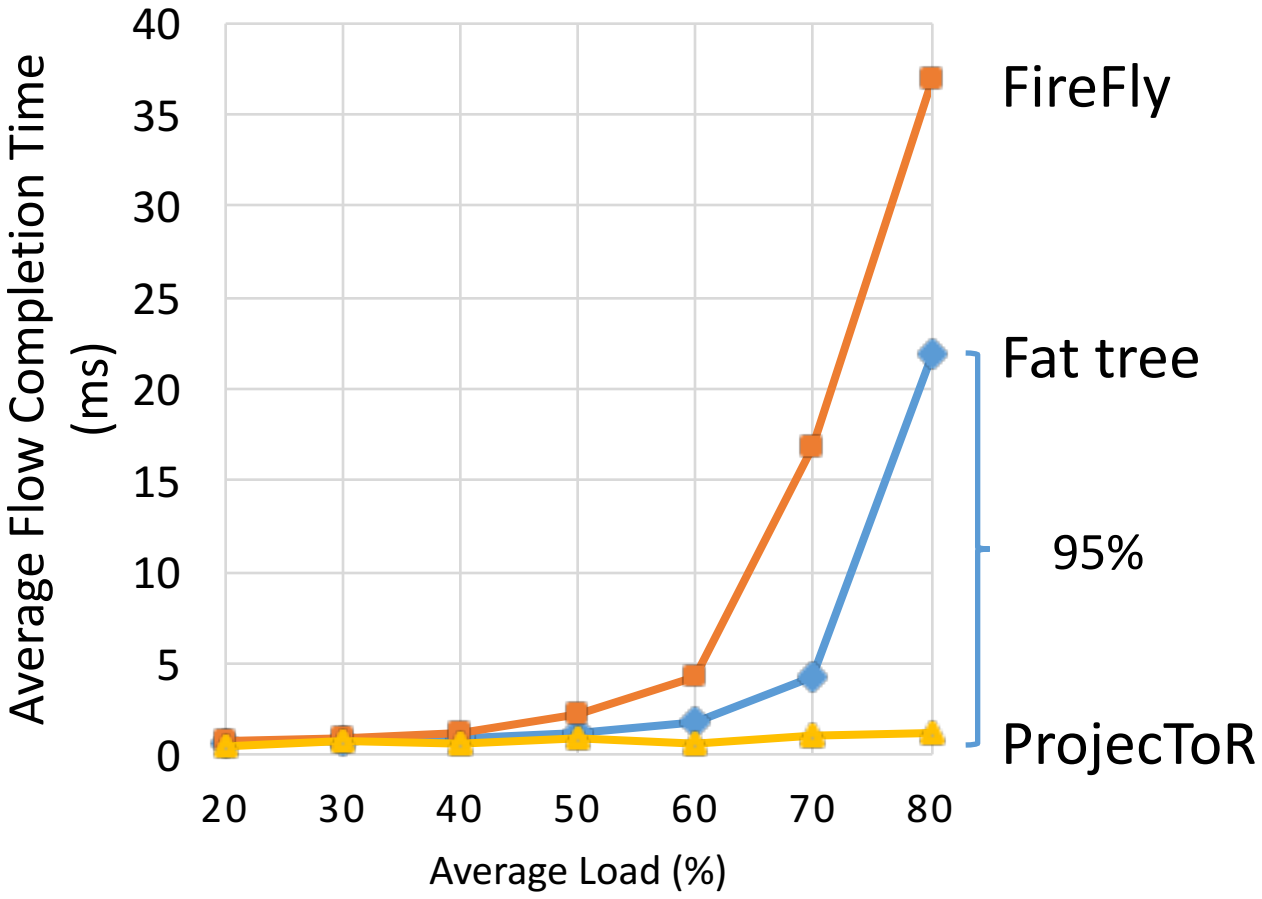


ProjecToR



- 128-ToR (1024 servers) with 16 lasers and photodetectors
- Day-long traffic matrix: to build the dedicated topology
- 5-min traffic matrix: to generate probability of ToR pair communication
- TCP flows arrival with poison arrival rate and realistic flow sizes

Simulation results



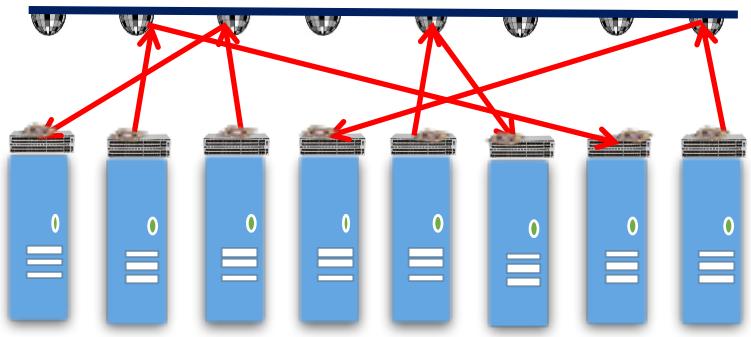
- Slow switching time
- Low fan-out

- No reconfigurability

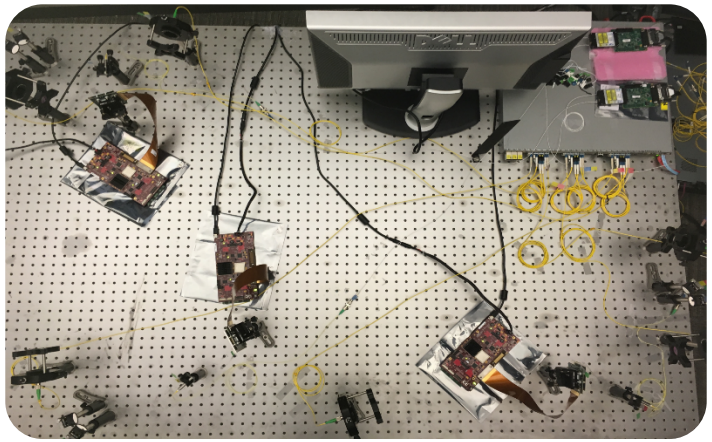
+ Reconfigurable
+ Switching time: 12us
+ high fan-out

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

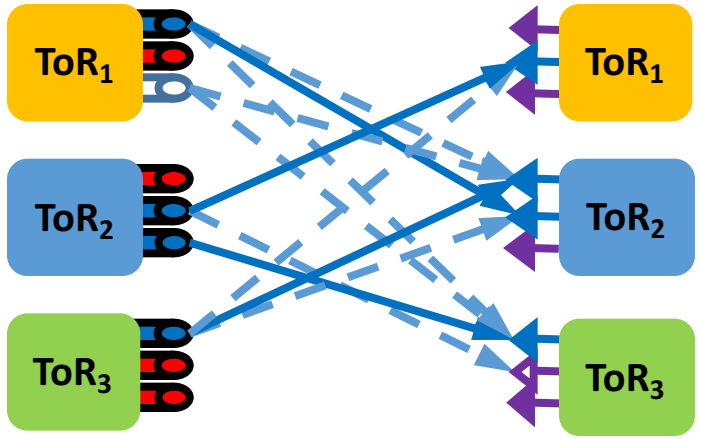
ProjecToR: A reconfigurable data center



Seamless, high fan-out, low switching time interconnect



Small prototype demonstrates feasibility



Decentralized flow scheduling algorithm