UC San Diego

INTRODUCING OPTICAL SWITCHING INTO DATACENTER NETWORKS

George Porter (on behalf of many co-authors!)

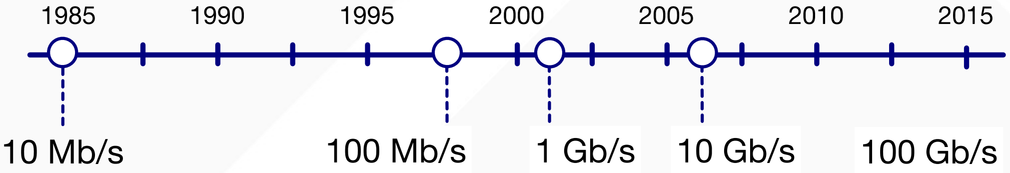
UC San Diego

University of Washington
Feb 8, 2021

GROWTH OF "HYPERSCALE" DATACENTERS

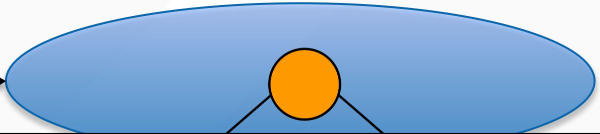
Google facebook YouTube Microsoft YAHOO! amazon.com Spotify salesforce NETFLIX

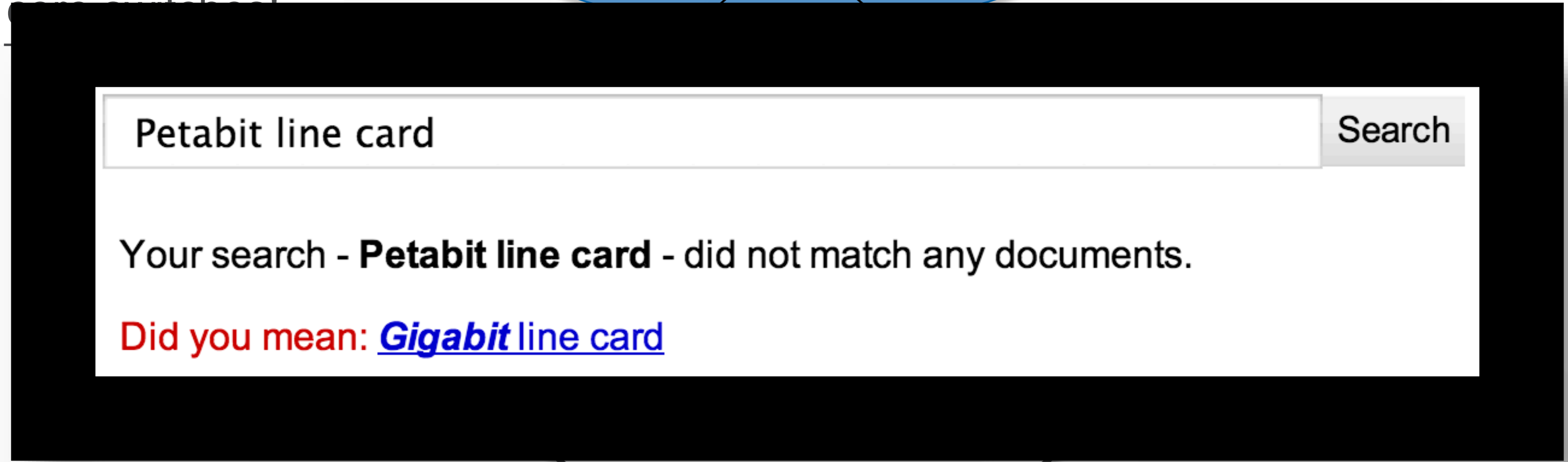
Network problem:
connecting >100,000 servers



HYPERSCALE IMPOSED NEW REQUIREMENTS ON NETWORKS

Can't buy sufficiently fast





100,000 x 10 Gb/s = 1 Pb/s

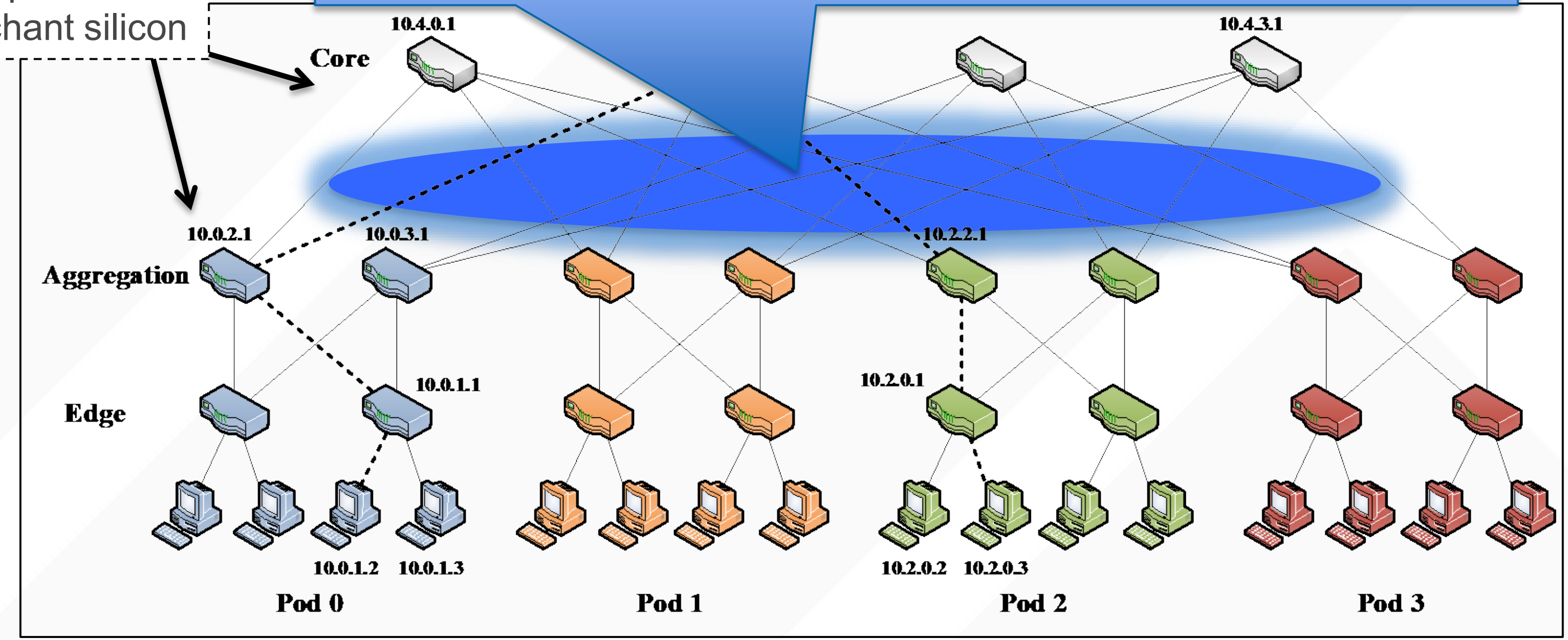
2009: RISE OF “SCALE OUT” NETWORKS

Cheap
Merchant silicon

Scaling bandwidth by scaling:

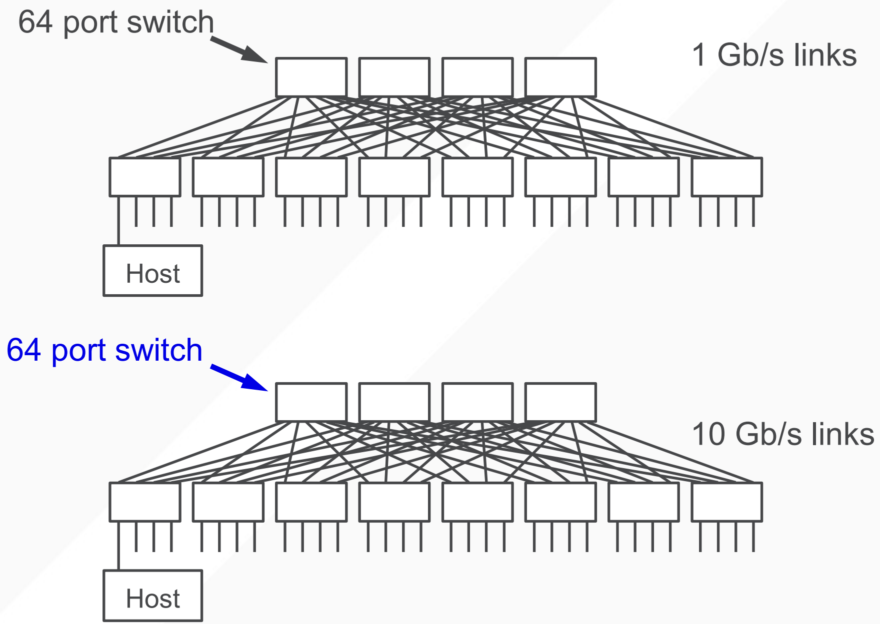
- # switches
- # cables
- # xceivers

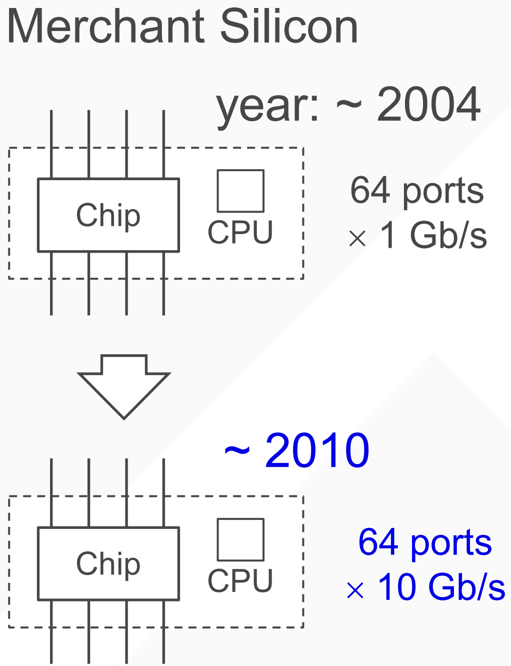
64k servers mean 196,608 links



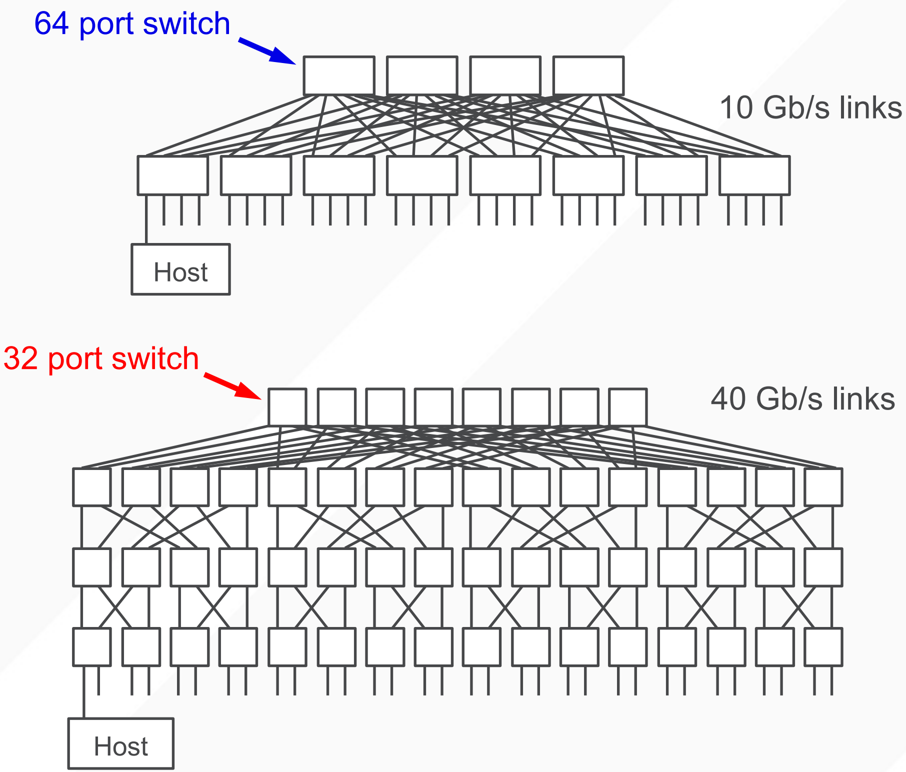
M. Al-Fares, A. Loukissas, and A. Vahdat, “A Scalable, Commodity Data Center Network Architecture,” SIGCOMM 2008.

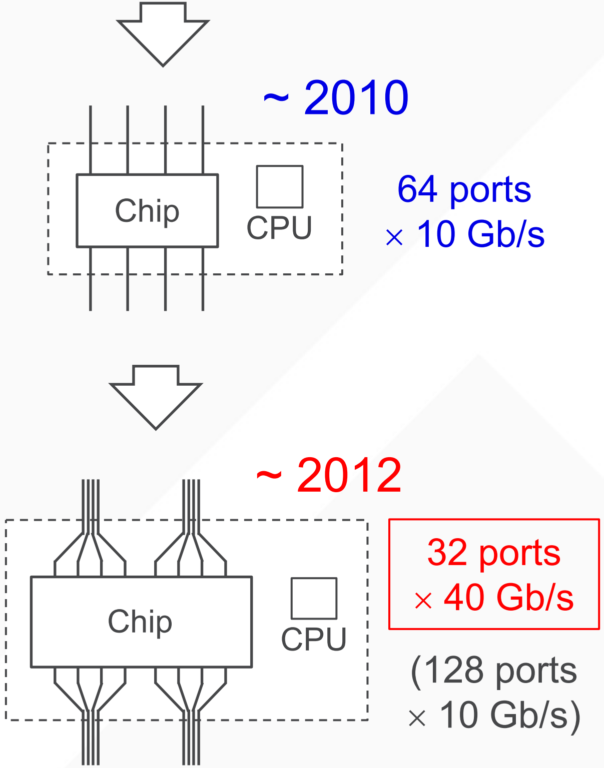
SCALING "TRADITIONAL" FATTREES IS BECOMING EXPENSIVE





SCALING “TRADITIONAL” FATTREES IS BECOMING EXPENSIVE

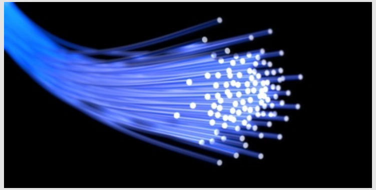




PROLIFERATION OF FAT TREE LAYERS A GROWING PROBLEM

Optical links

1,000 + Gb/s –
1,000 + meters

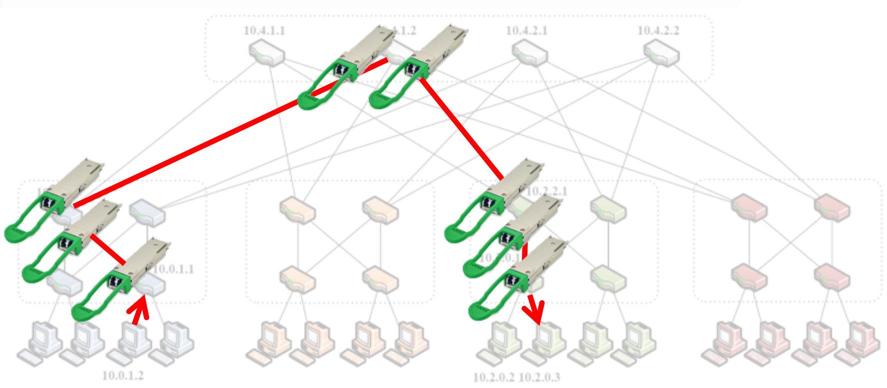


Single mode fiber

10 Gb/s –
2,000 meters

SFP+ transceiver

Datacenter Network



1 Gb/s – 100m

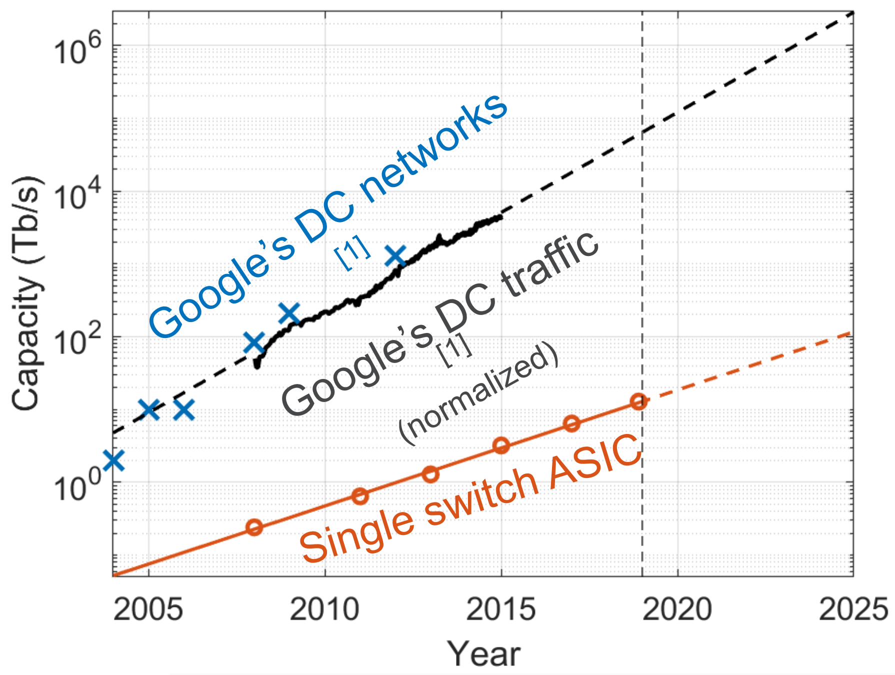
10 Gb/s – 10 meters

Electrical links

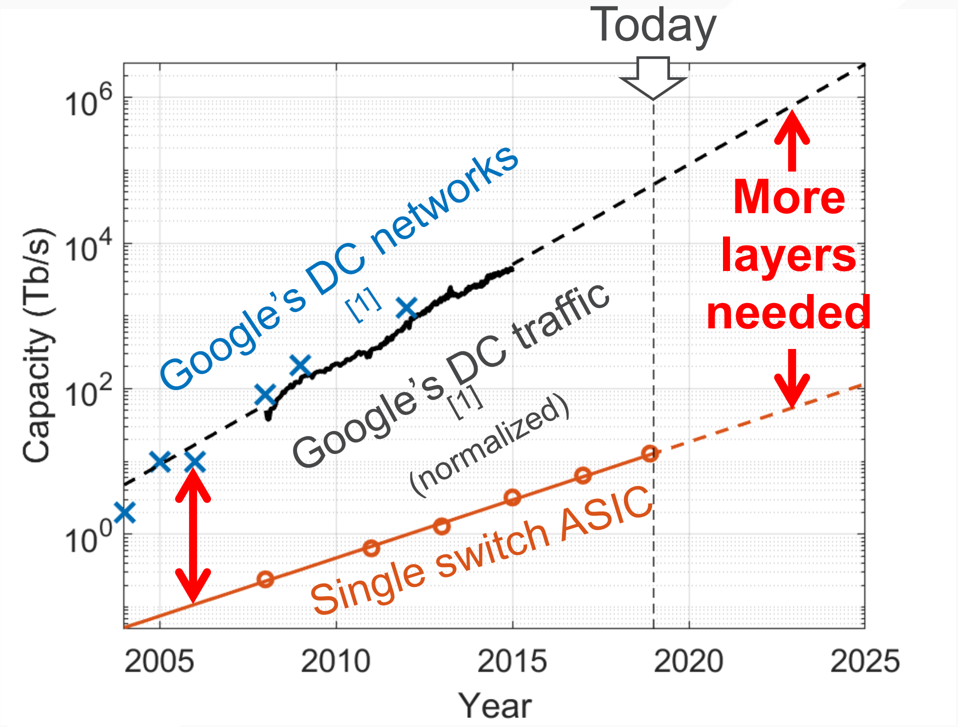
For every device attached to the network, there are multiple transceivers in the network

100k nodes: O(100kW) and O($$$)

BANDWIDTH GROWTH CONTINUES HOWEVER…

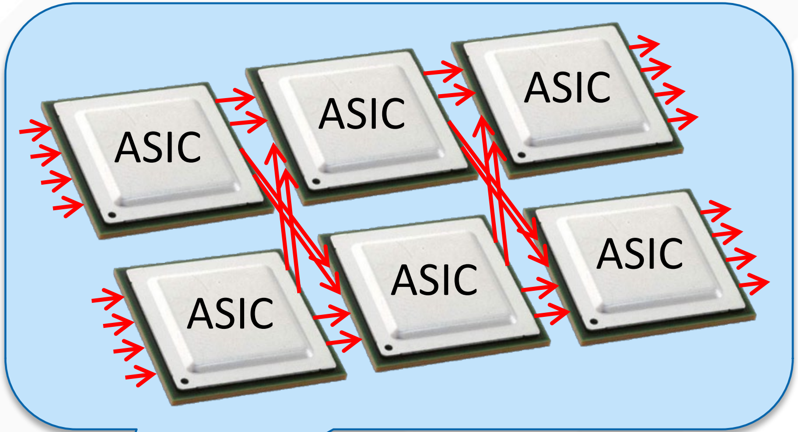


BANDWIDTH GROWTH CONTINUES HOWEVER…



SCALING LIMITATIONS OF CMOS-BASED PACKET SWITCH CHIPS

- Increasing difficulty getting data in/out of the chip
- Divergence between ***link*** rate and ***channel*** rate
  - E.g. 100G vs 4x25G
- More fabric layers = higher cost & power



"Hiding" layers

	0.64 TB/s	5.12 TB/s	12.8 TB/s
Max. chip radix =	**64** x 10G	**128** x 10G	**128** x 25G
Used chip radix =	**16** x 40G	**32** x 40G	**32** x 100G

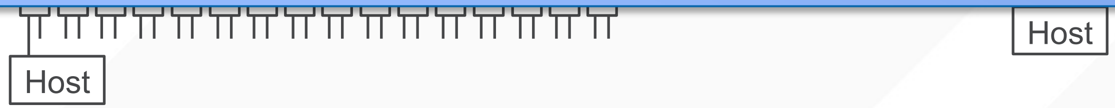
MOVE TO “CHASSIS” BASED FAT TREES (FACEBOOK, GOOGLE)

“Traditional” packet switch

Multistage chassis switch

Fully-provisioned network – 8,192 end hosts

Architecture	# Tiers	# Hops	# Transceivers	# Switch chips	# Switch boxes	# Fibers
Traditional	3	5	49 k	1,280	1,280	25 k
Multistage Chassis	2	9	33 k	2,304	192	16 k
Improvement:			1.5× (cost)		6.7× (cost)	1.6× (cost)
Penalty:		1.8× (latency)		1.8× (power)		



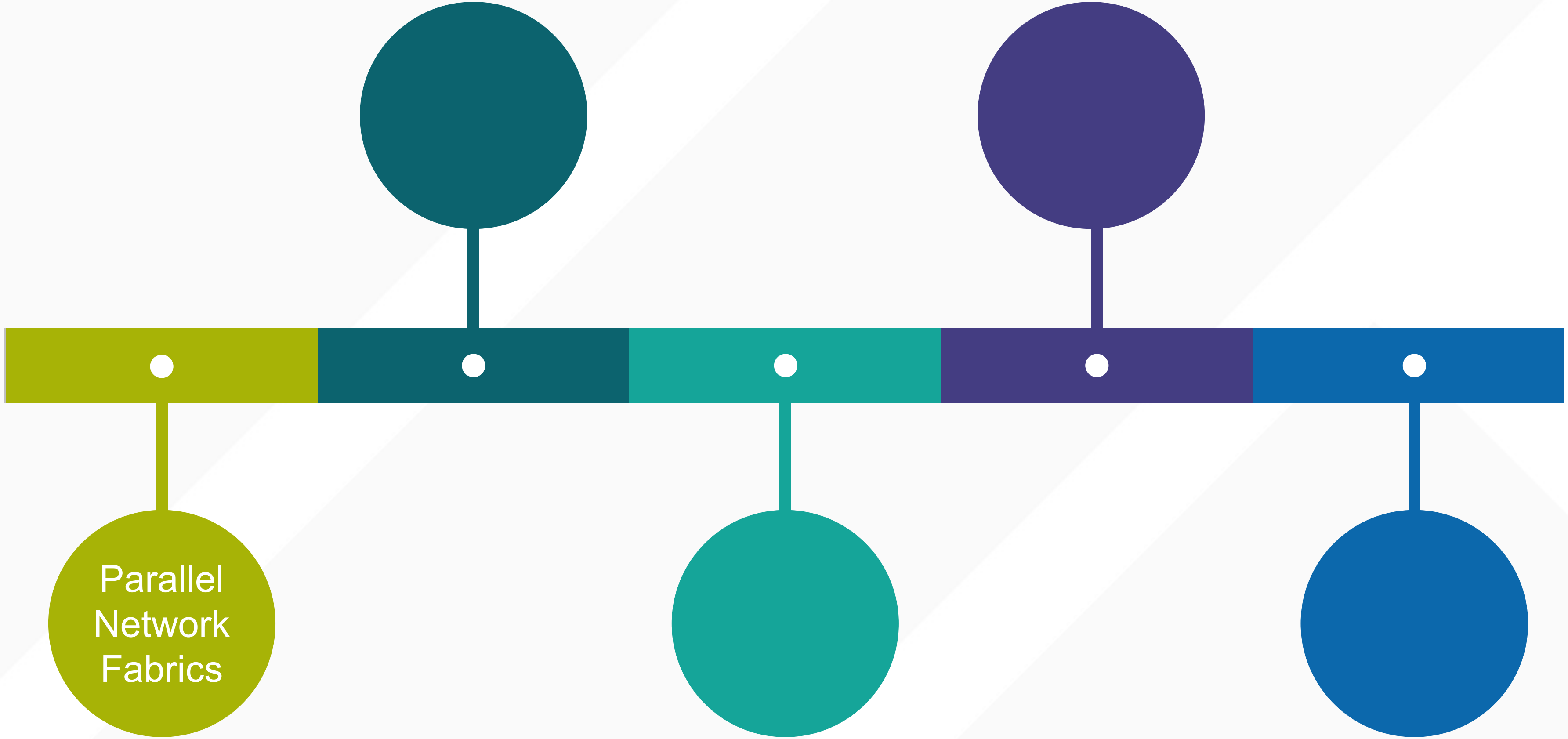
TRENDS

- Conventional datacenter networks facing scaling limitations
  - Largely due to scaling limits of underlying packet switch chips

- Direction 1: Parallel network fabrics
  - Adopted thus far by Facebook and LinkedIn
  - Short-to-medium timeframe

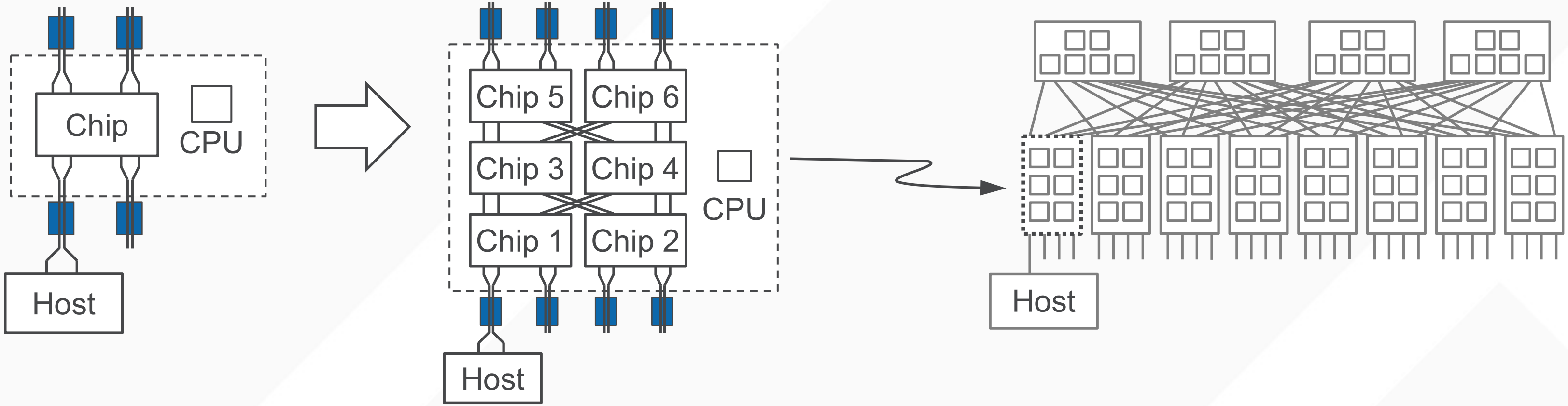
- Direction 2: Replace packet switches with optical switches / circuit switches
  - Medium-to-longer timeframe

RESEARCH TIMELINE: DIRECTION 1: PARALLEL NETWORKS



PARALLEL NETWORK DESIGNS

Conventional architectures:



Chassis architecture still scaling the network *up*… just hiding the tiers in switch chassis.

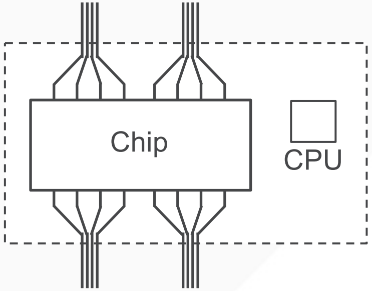
Alternative: Scale out via separate physical data planes

- Benefits: Reduced cost, power, and latency
- Tradeoff: Give up a single “fast” network abstraction

UNDERLYING SWITCH RADIX IS INCREASING

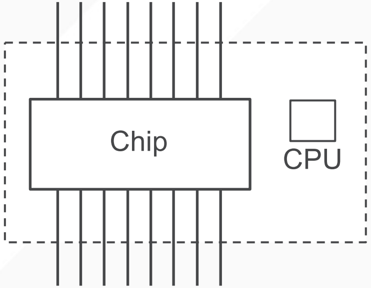
Ex. Broadcom's Tomahawk switch:

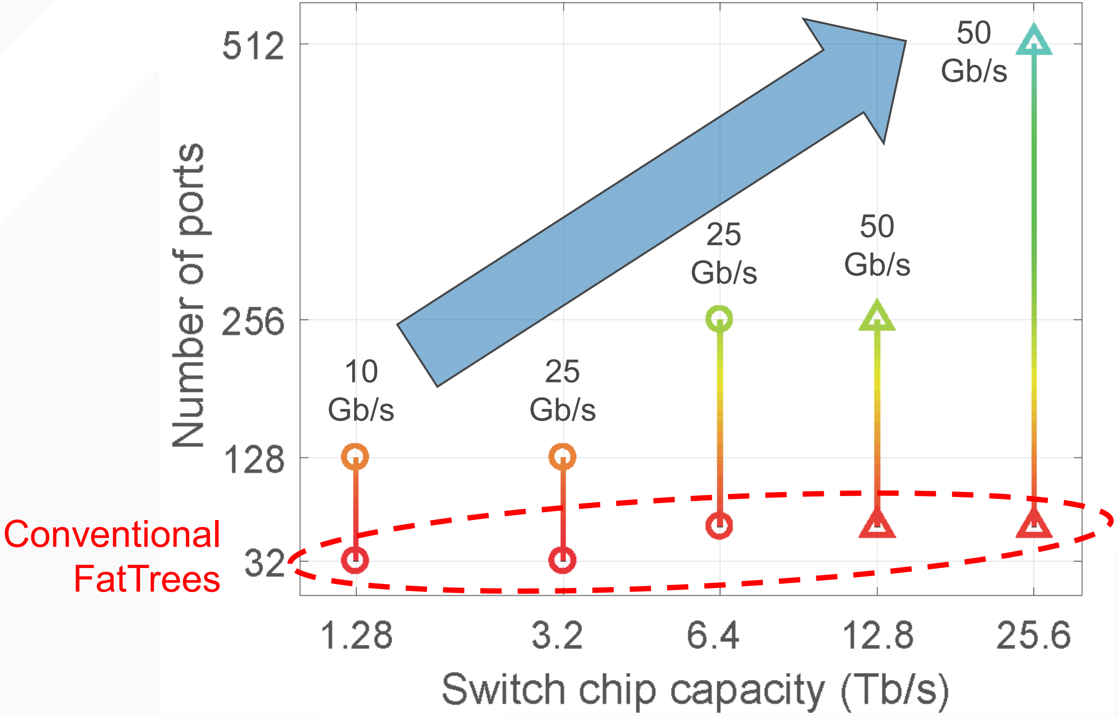
32 ports @ 100 Gb/s



OR

128 ports @ 25 Gb/s

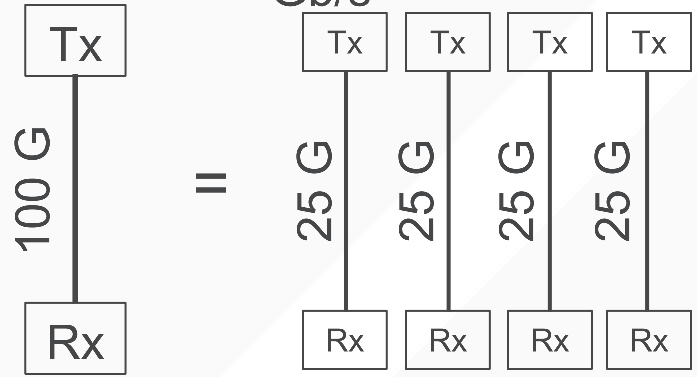


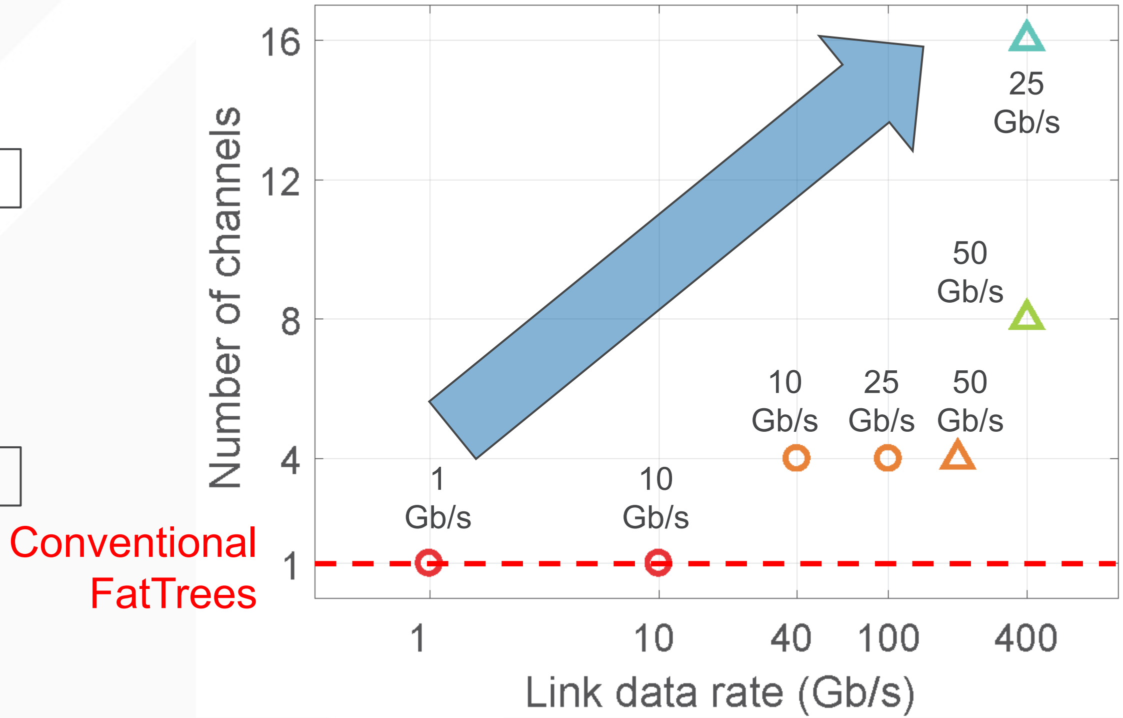


LINK CHANNEL COUNT IS INCREASING

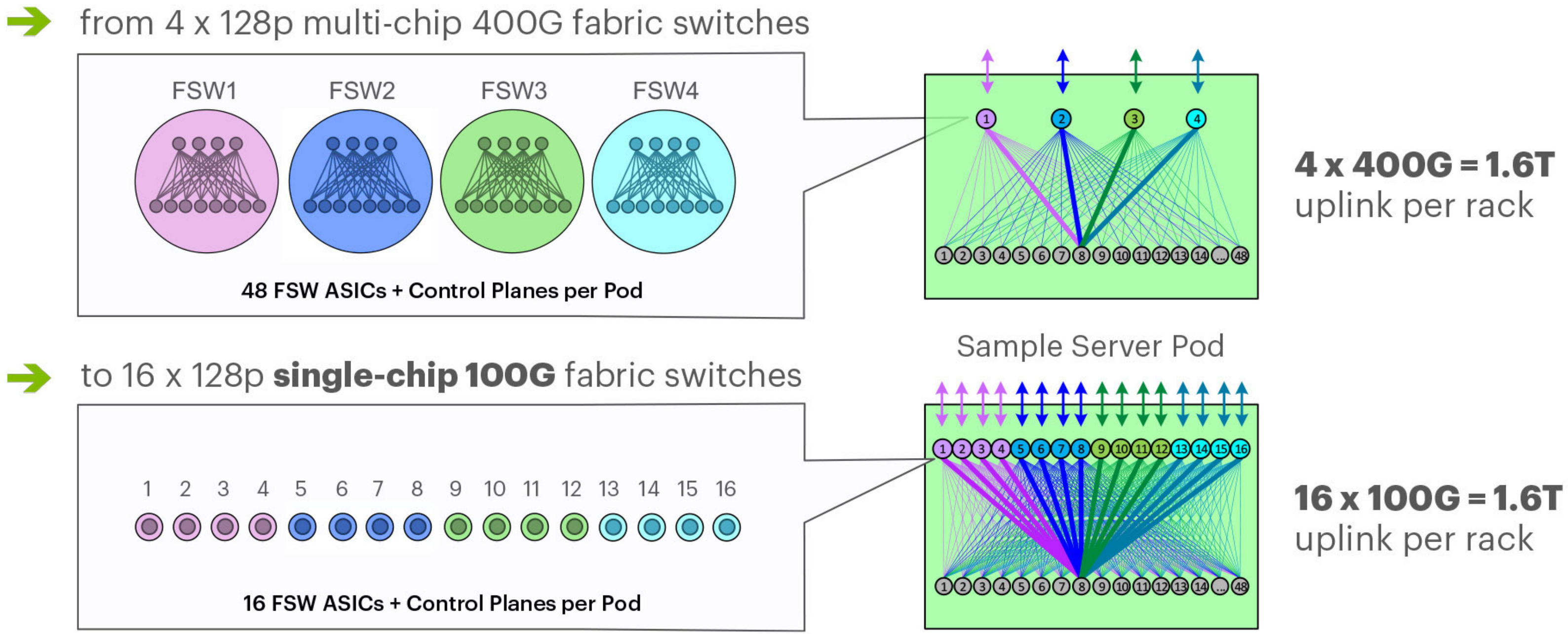
Ex. 100 Gb/s optical link:

= 4× 25 Gb/s





PARALLEL NETWORKS IN INDUSTRY: FACEBOOK

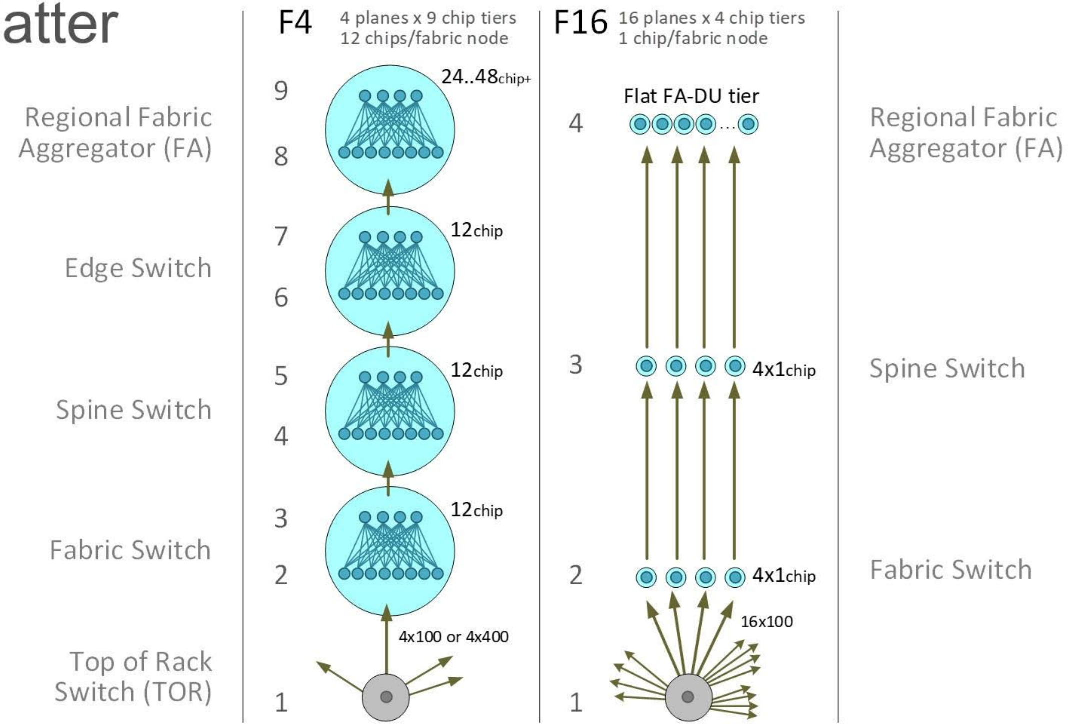


https://engineering.fb.com/data-center-engineering/f16-minipack/

PARALLEL NETWORKS IN INDUSTRY: FACEBOOK

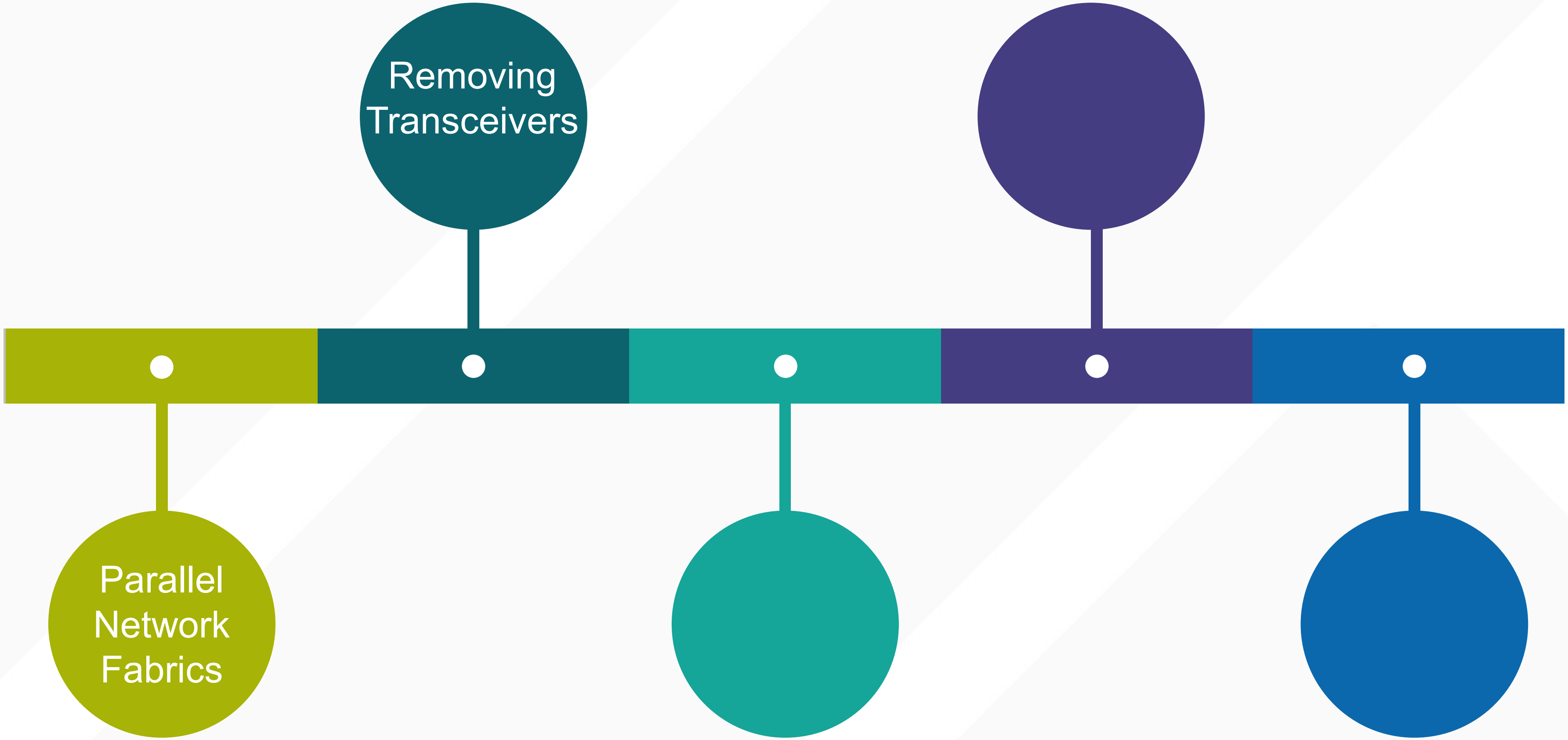
Simpler and Flatter

- Over **3X less** switch ASICs and control planes in fabric
- **2.25X less** tiers of chips in the topology
- Up to **2X less** host-to-host network hops intra-fabric
- Up to **3X less** host-to-host network hops intra-region



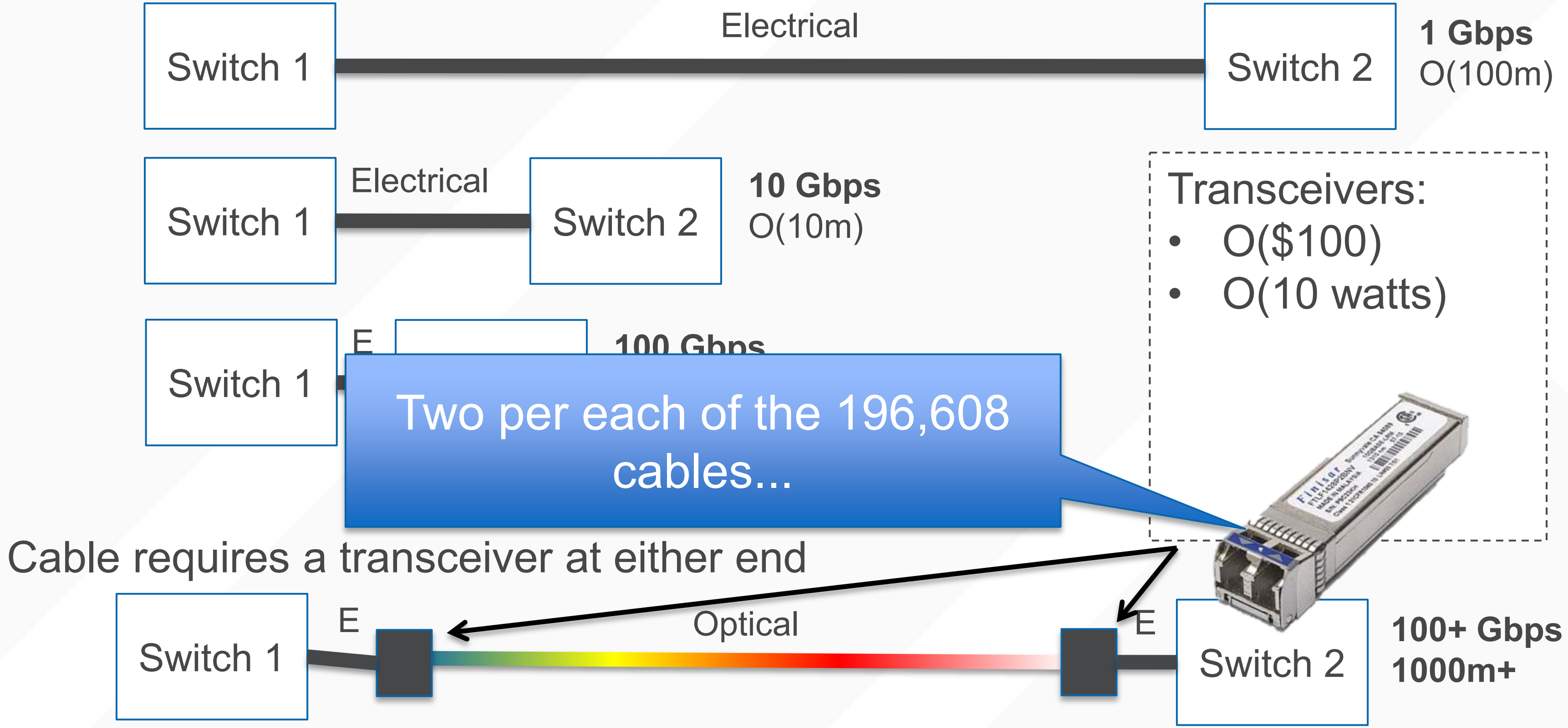
https://engineering.fb.com/data-center-engineering/f16-minipack/

RESEARCH TIMELINE: DIRECTION 2: OPTICAL NETWORKS

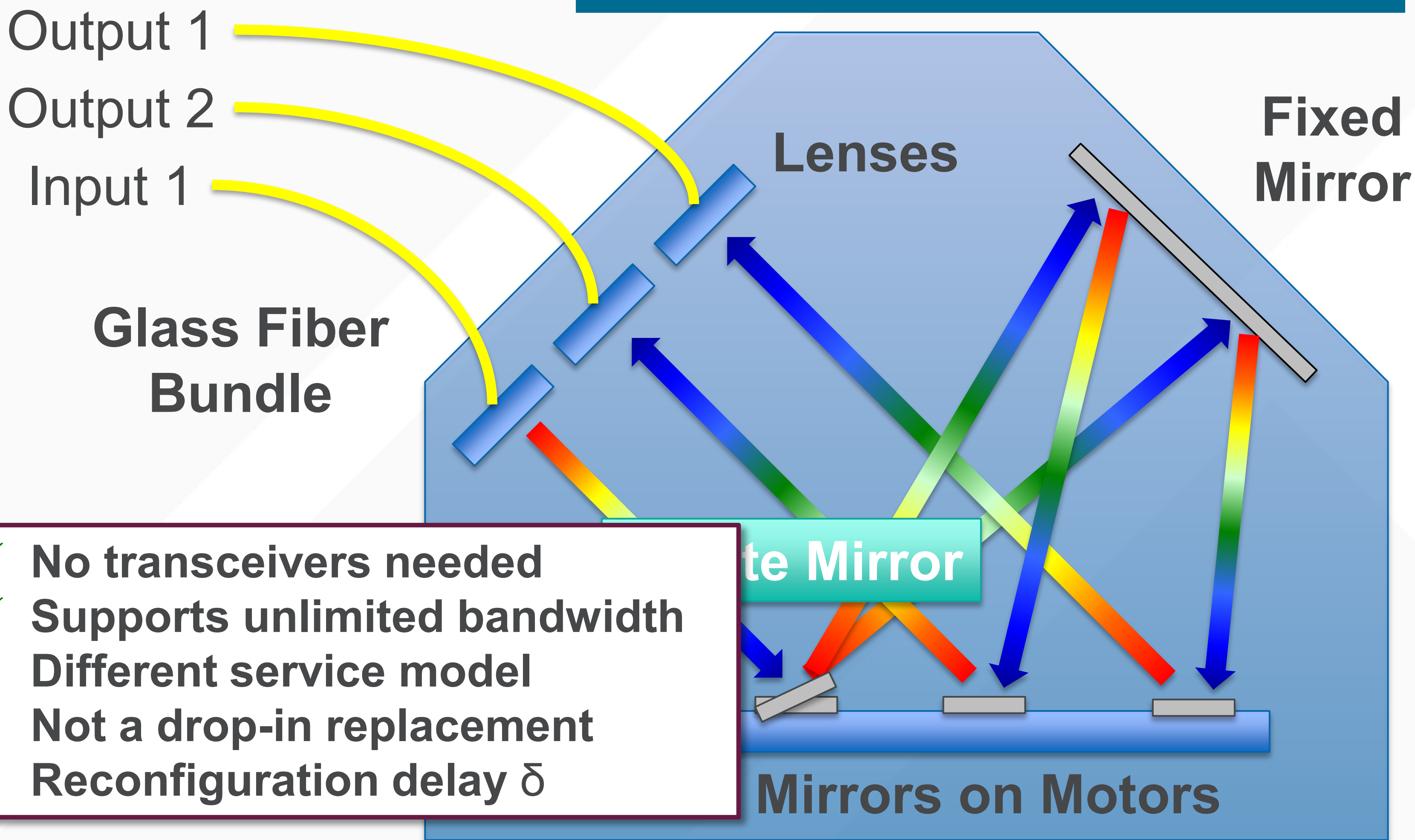


MOTIVATION FOR OPTICAL NETWORKING

The faster the data rate of a cable, the shorter it has to be

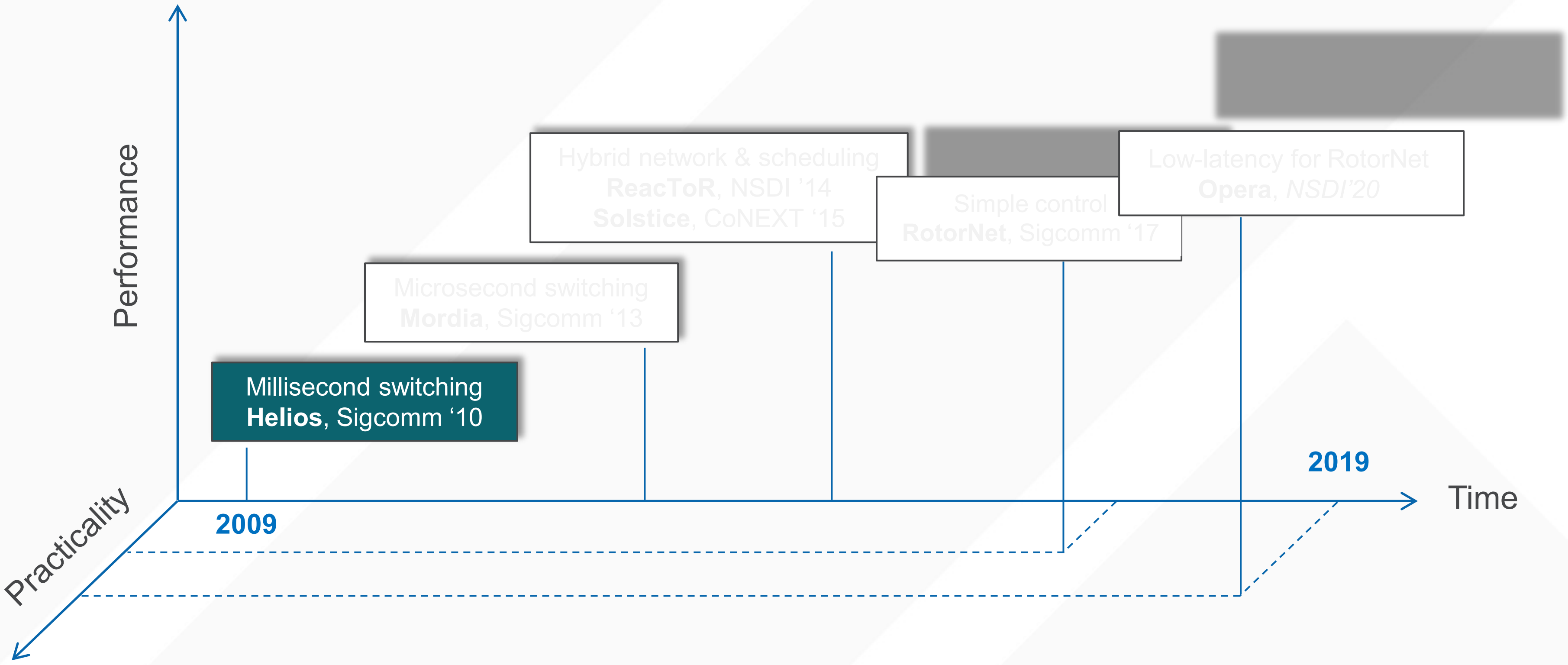


OPTICAL SWITCHES

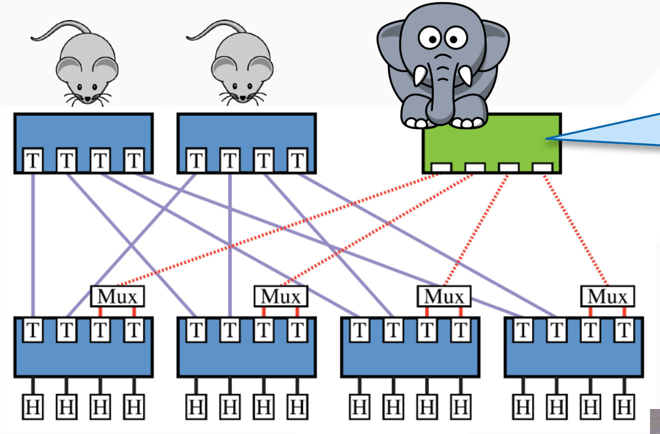


- ✓ No transceivers needed
- ✓ Supports unlimited bandwidth
- \- Different service model
- \- Not a drop-in replacement
- \- Reconfiguration delay δ

REMOVING TRANSCEIVERS

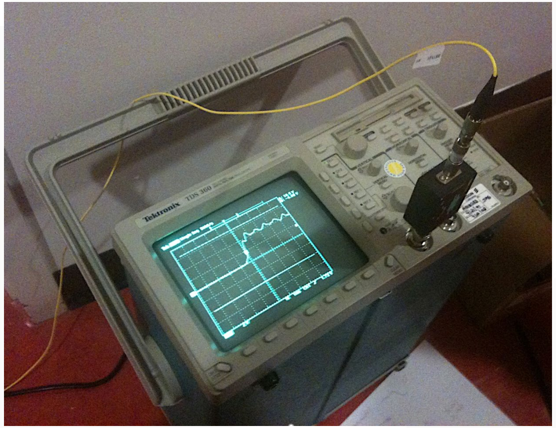


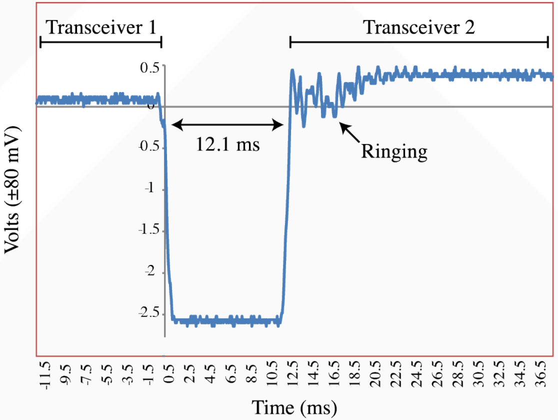
2009 – USING 3D MEMS TO REMOVE TRANSCEIVERS



- Model: 15ms switch time
- Observed: 12.1ms switch time

- Technology: telecom-grade 3D-MEMs
- Scalability: 100s of ports
- Target: Inter-“pod”



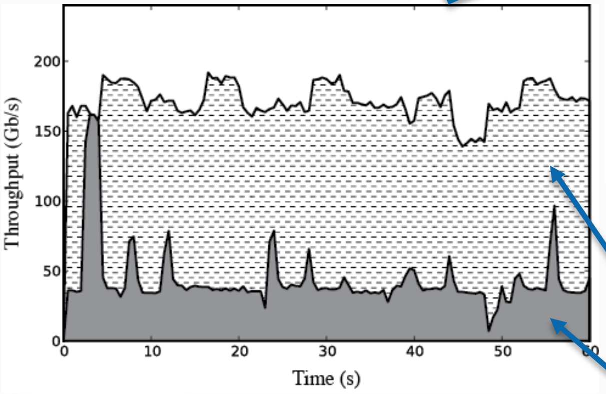


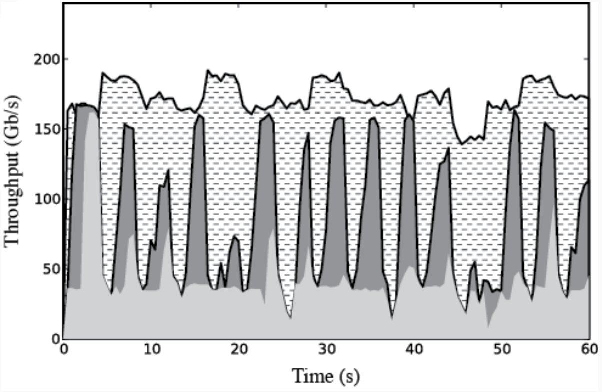
N. Farrington et al., SIGCOMM 2010

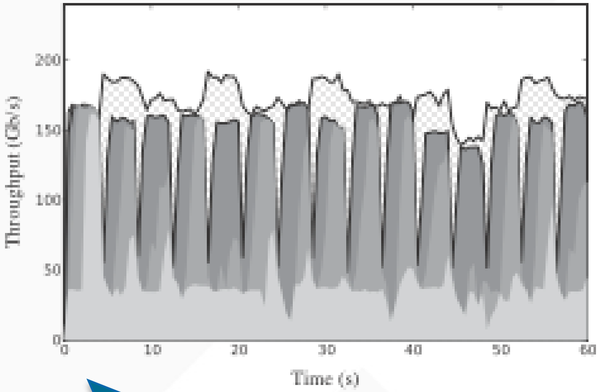
BOTTLENECKS IN NON-SWITCH COMPONENTS

- Telecom not designed for rapid reconfiguration
- Many non-switch bottlenecks in optical components

2-second link flap prevention





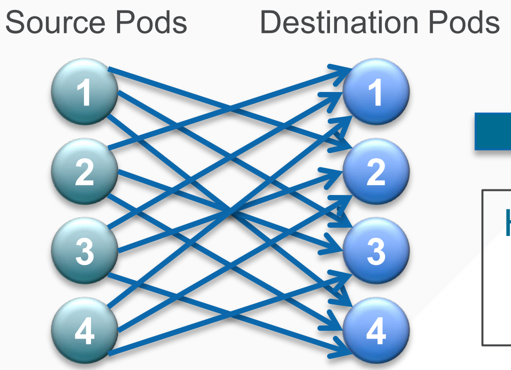


Transceiver Electronic Dispersion Compensation

Packet-switch baseline

Observed circuit-switched bw

CONTROL PLANE 100X SLOWER THAN SWITCH TIME

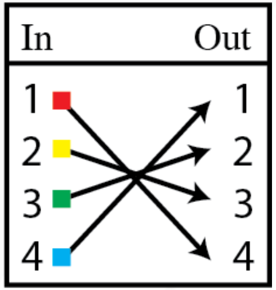


Hedera demand estimator
+
Edmund's Algorithm

Demand Matrix i

	1	2	3	4
1	0	1	1	3
2	7(4)	0	3	1
3	1	3	0	9(4)
4	3	2	1	0

Circuit Switch i

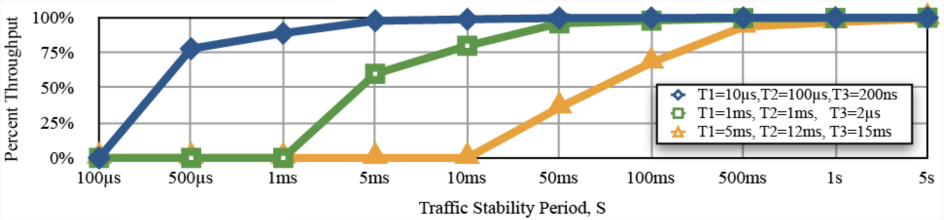


1. Collect counters from packet switches
2. Estimate "true" demand
3. Calculate max-weighted matching
4. Reconfigure packet and optical switches

- One cycle ≈ one second
- Circuits try to "match" current network conditions
- Stateless in between assignments

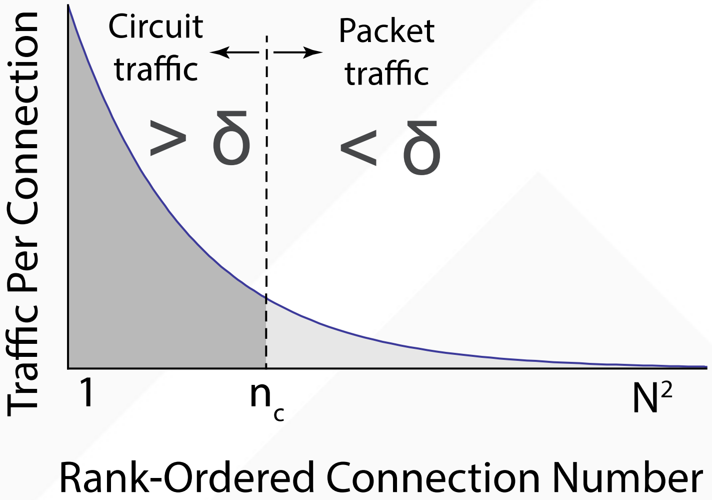
APPLICABILITY LIMITED BY SLOW SWITCH TIME & CONTROL PLANE

- Model: 15ms switch time
- Reality: 1000ms control plane
- To "capture" more of the traffic in optics, need a **faster switch** and **faster control plane**

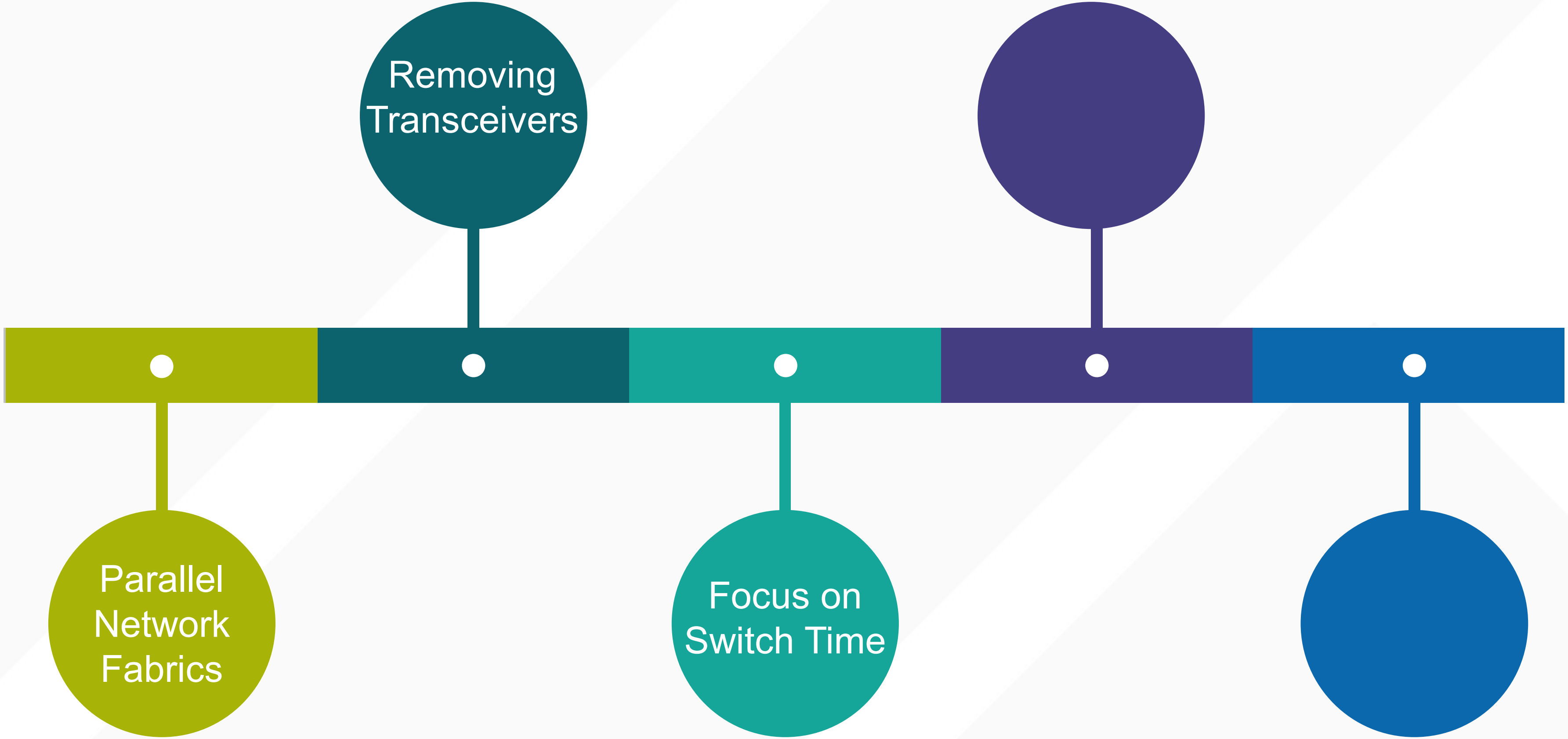


"Hardware Requirements for Optical Circuit Switched Data Center Networks", Farrington et al., OFC 2011

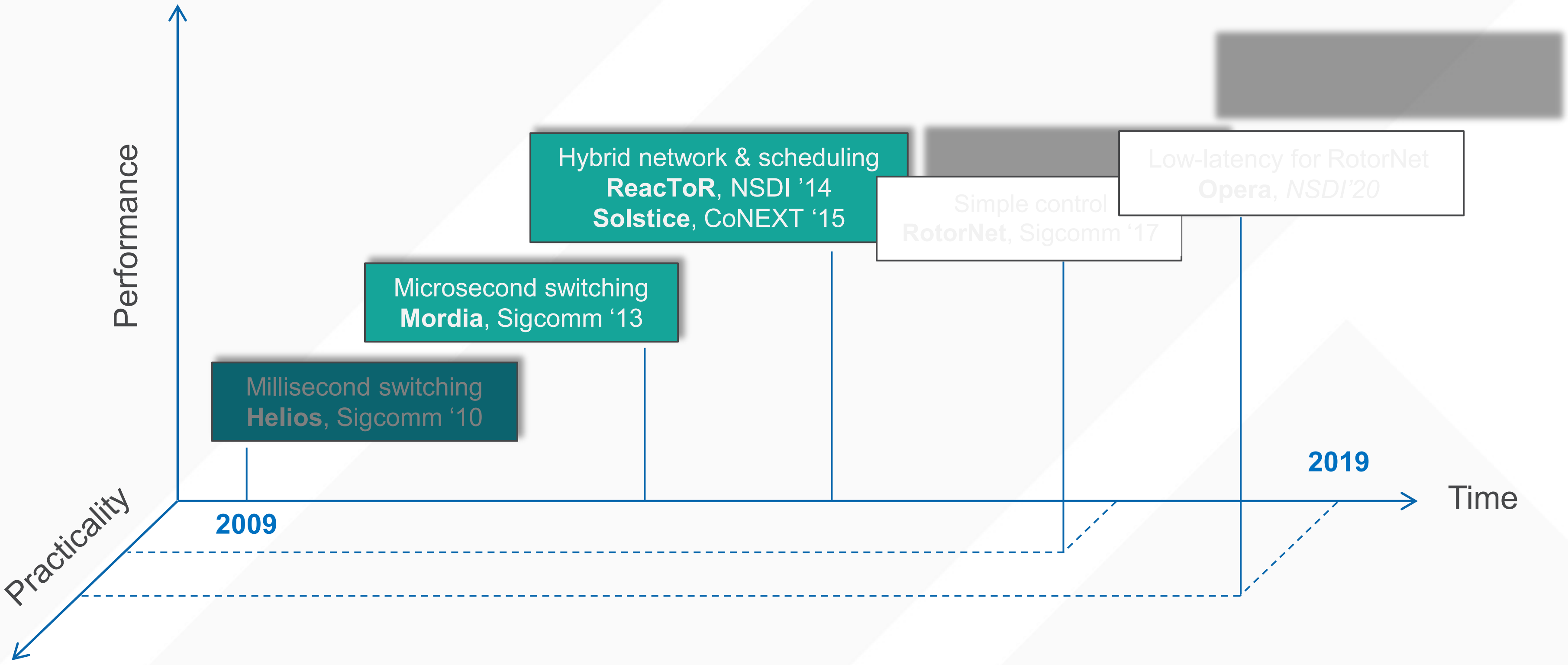
Applicability of circuit switching determined by switch time δ



RESEARCH TIMELINE: DIRECTION 2: OPTICAL NETWORKS



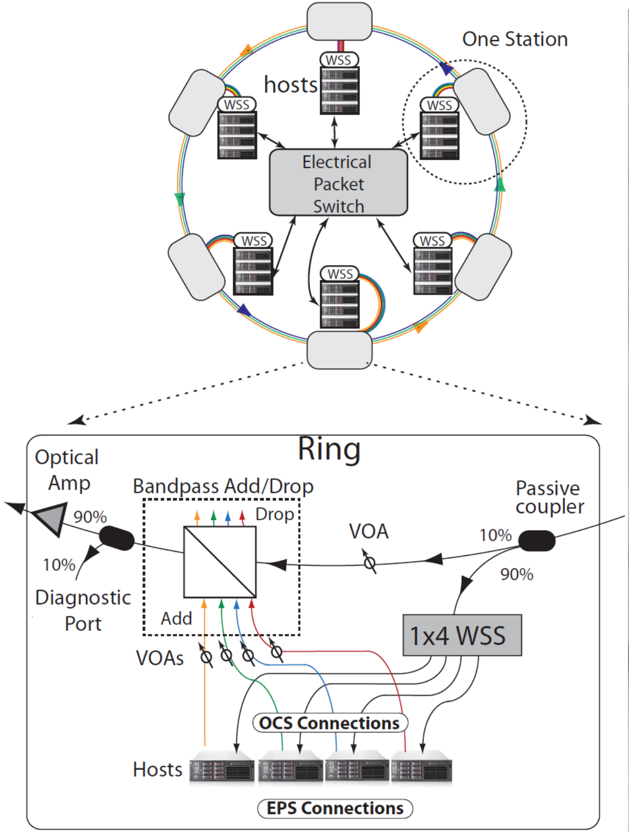
REMOVING TRANSCEIVERS



USING 2D MEMS TO "CHASE MICE"

- Needed a faster switch
- 2D MEMS very fast…
  - 2 µs switch time + ringing
  - Approx 11.5 µs total
- …but not scalable (~24 ports)
  - Lots of ports → slow
  - Few ports → fast

2011 - MORDIA – A 2D-MEMS 24-PORT MICROSECOND SWITCH



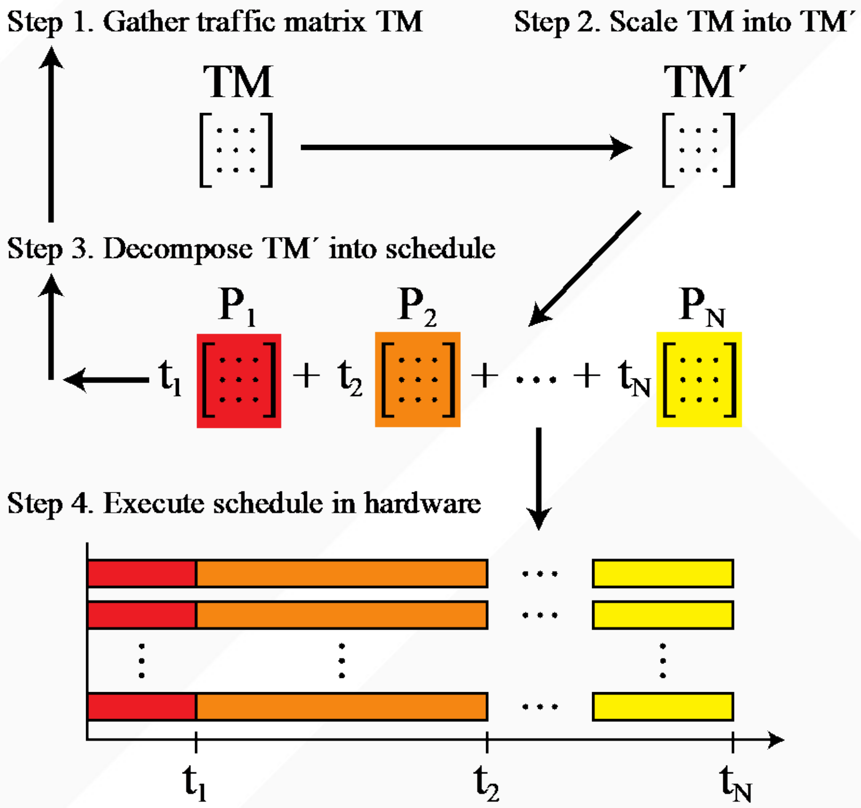
Porter et al., "Integrating Microsecond Circuit Switching into the Data Center", Sigcomm'13.

HOW MICROSECOND SWITCHING CHANGES THE CONTROL PLANE

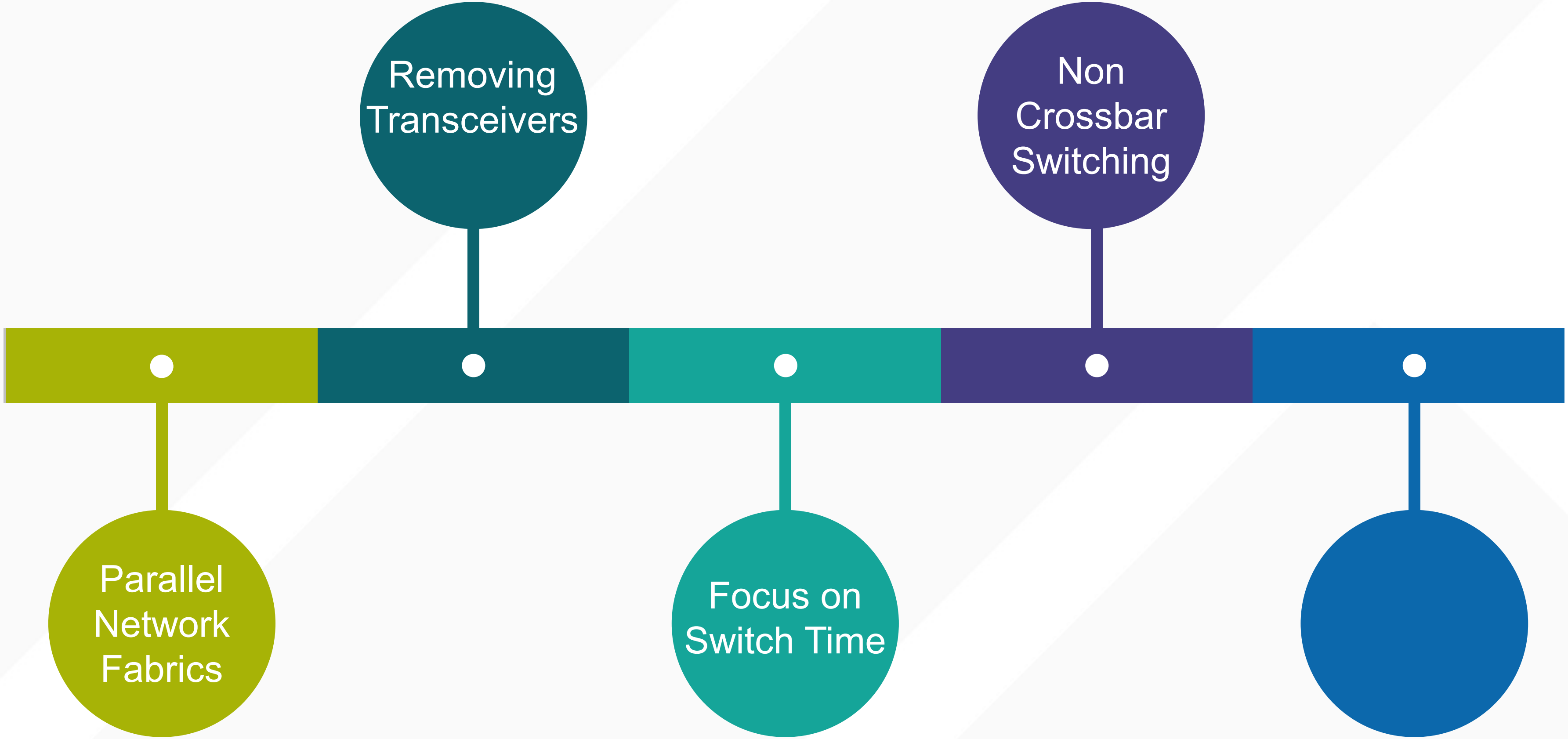
- Microsecond switching prevents scheduling with "fresh" data
  - Collecting demand a bottleneck!
- Insight: amortize series of switch configurations across a single demand estimate:

$$TM' = \sum_{i}^{N} t_i P_i$$

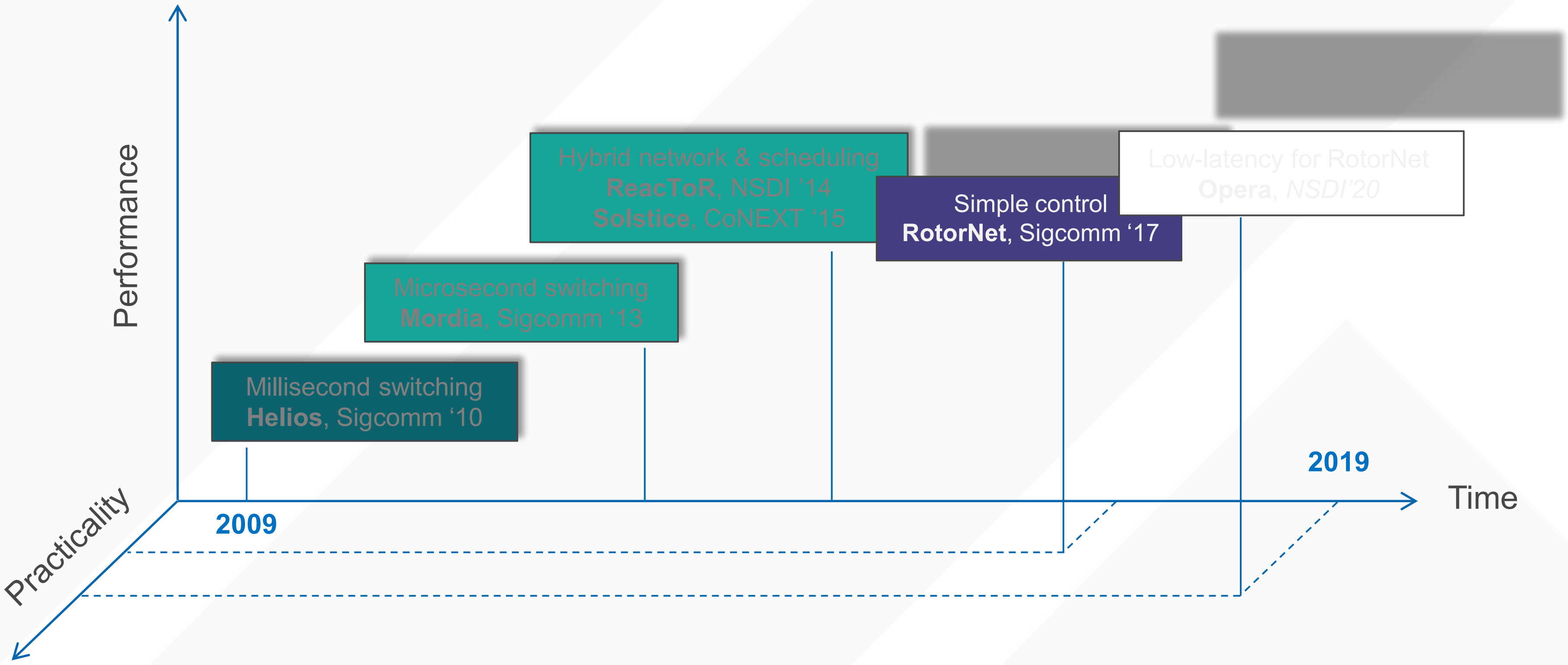
- Embodied by *Solstice* and *Eclipse* algorithms
- Result: "Chasing" demand
  - Reactive and responsive



RESEARCH TIMELINE: DIRECTION 2: OPTICAL NETWORKS



NON-CROSSBAR NETWORKS



Toward 100+ Petabit/second datacenters

UCSDCSE

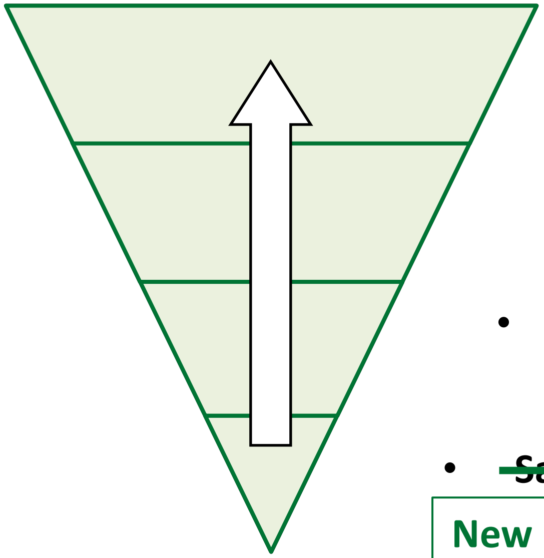
Challenge: deliver (very) low-cost bandwidth at scale

Co-design:

Protocol

Topology

Hardware



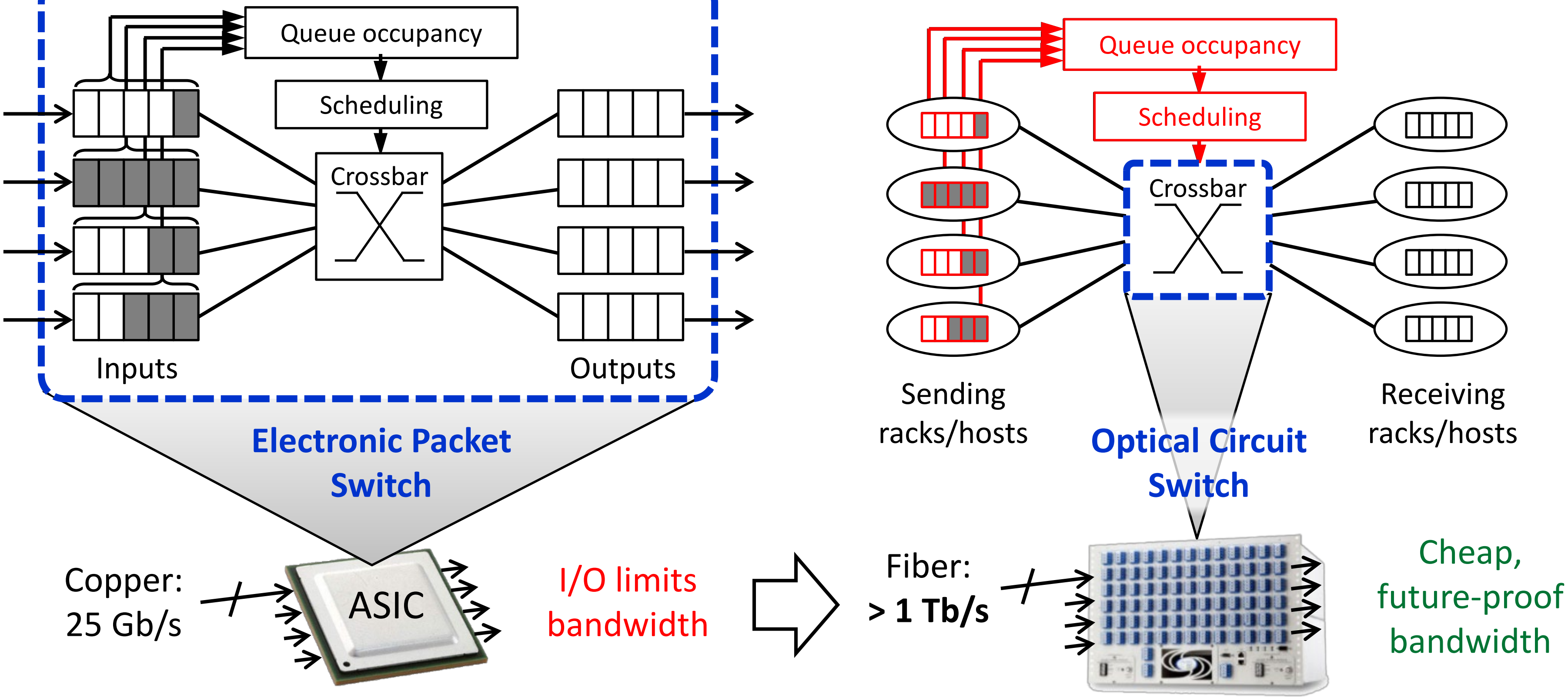
- **New protocols**
  Load balancing, congestion control, …
- **New topologies**
  Jellyfish, Longhop, Slimfly, …
- **New hardware**
  Optical circuit switching, RF/optical wireless, …
- ~~**Same switching model**~~

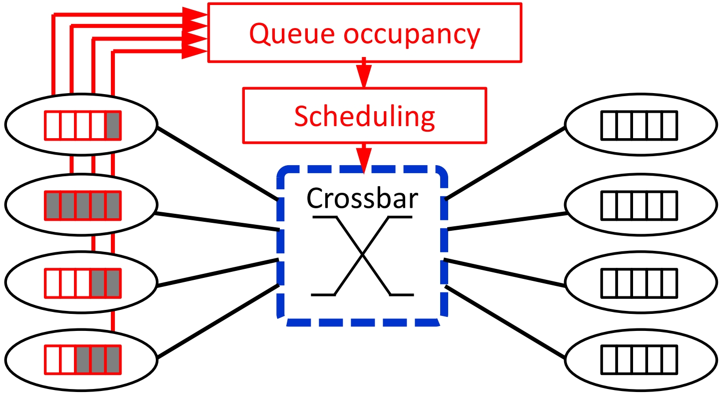
New "Rotor" switching model

RotorNet → "Future-proof" bandwidth (2× today) + simple control + …

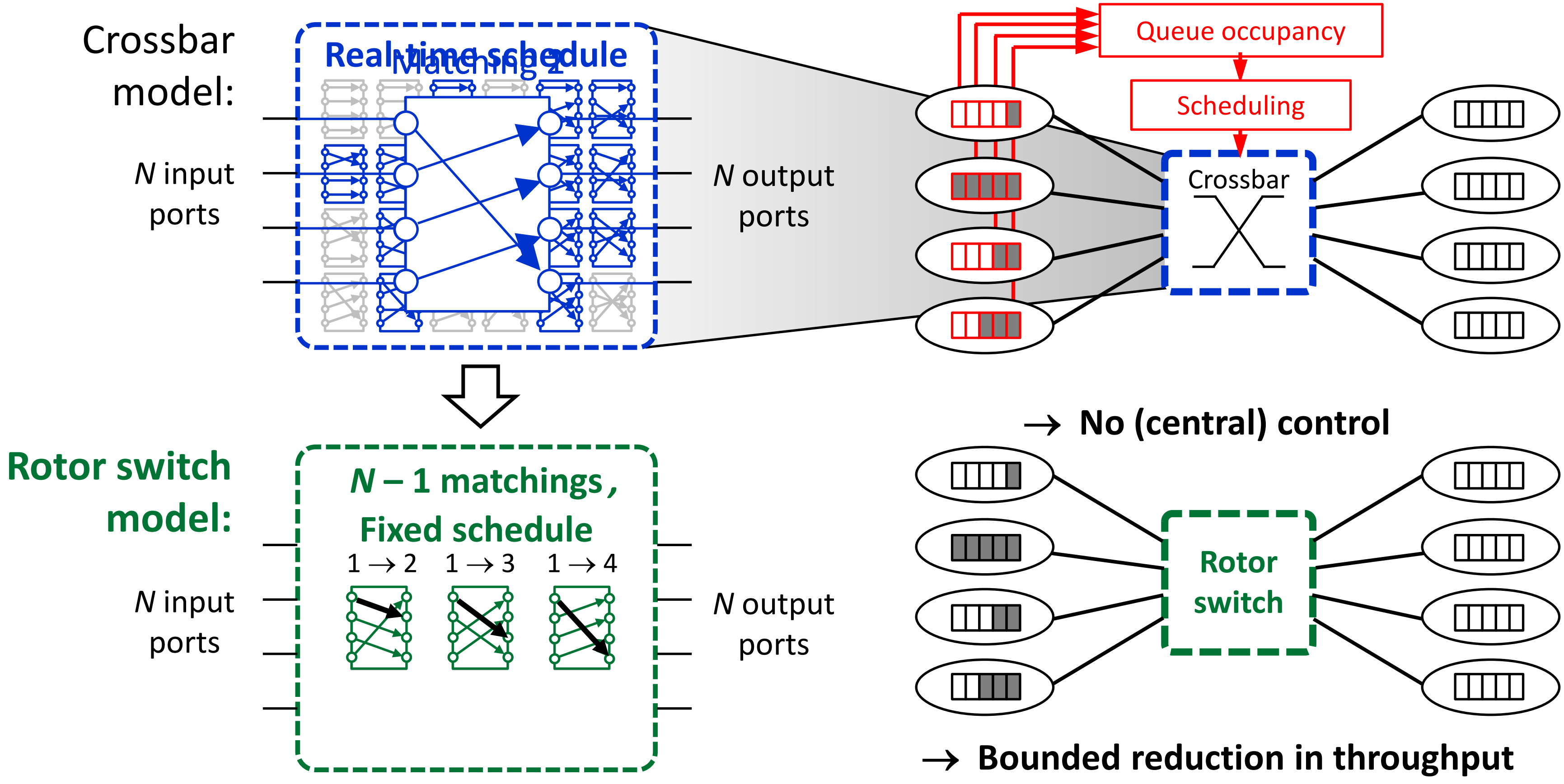
Optical switching – benefits & barriers

Data plane doesn't scale to entire datacenter!




Queue occupancy
Scheduling
Crossbar

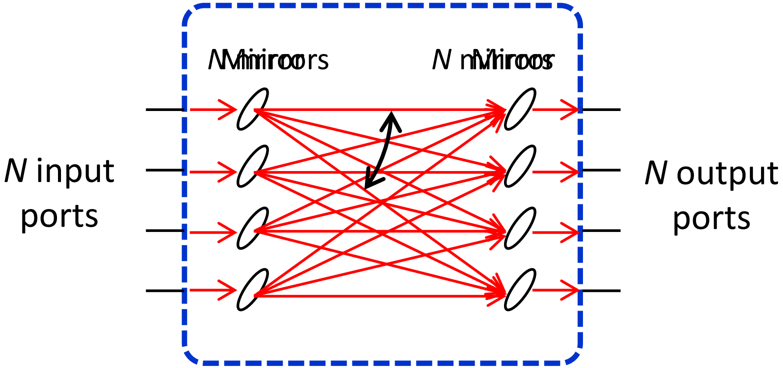
Rotor switching model simplifies control



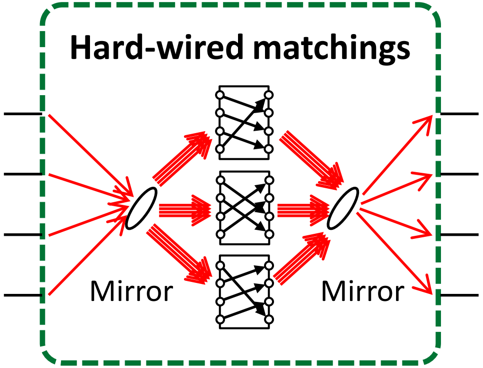
N input
ports
N output
ports

Rotor switches have a simpler implementation

Optical Crossbar:



Optical Rotor switch:



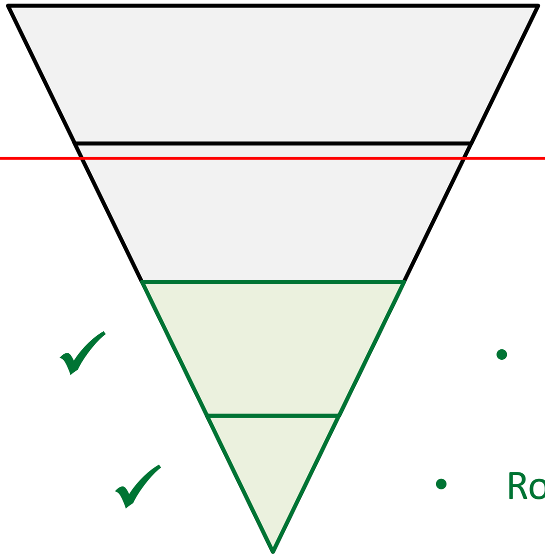
- Cost and complexity scale with:

Ports

Matchings (<< Ports)

Ex. 2,048 ports: 4,096 mirrors, 2,048 directions | 2 mirrors, 16 directions

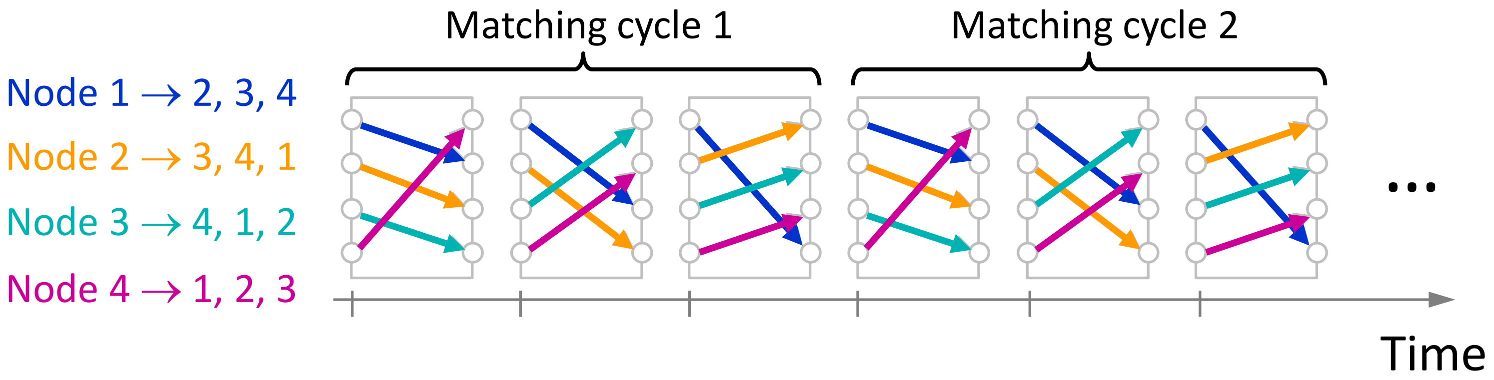
RotorNet architecture overview



- **Forwarding?**
- **Topology?**
- Optical Rotor switch → **More scalable**
- Rotor switching model → **Simpler control**

1-hop forwarding over Rotor switch

- Wait for direct path:

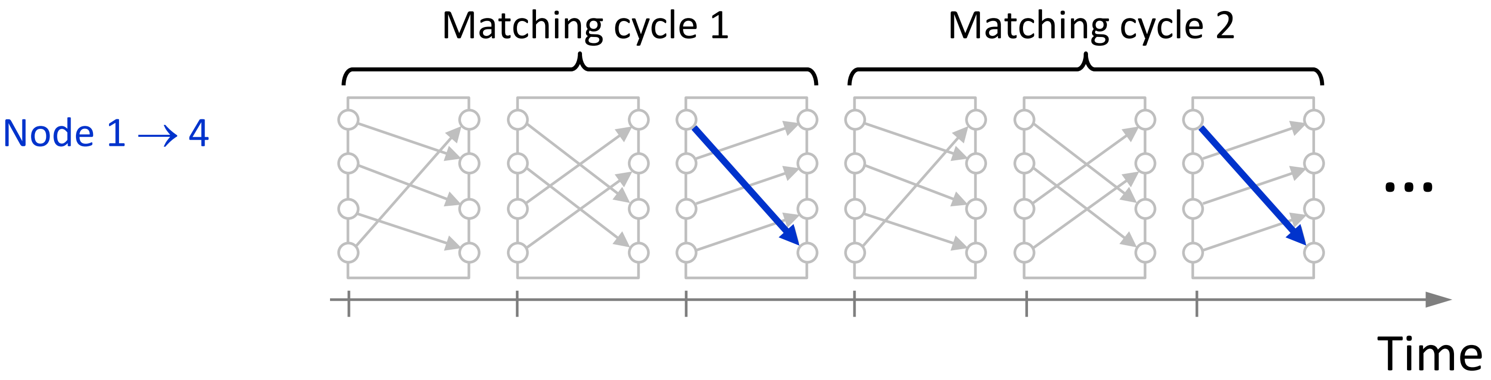


Uniform traffic → 100% throughput

- But datacenter traffic can be sparse …

1-hop forwarding & sparse traffic = low throughput

- Wait for direct path:



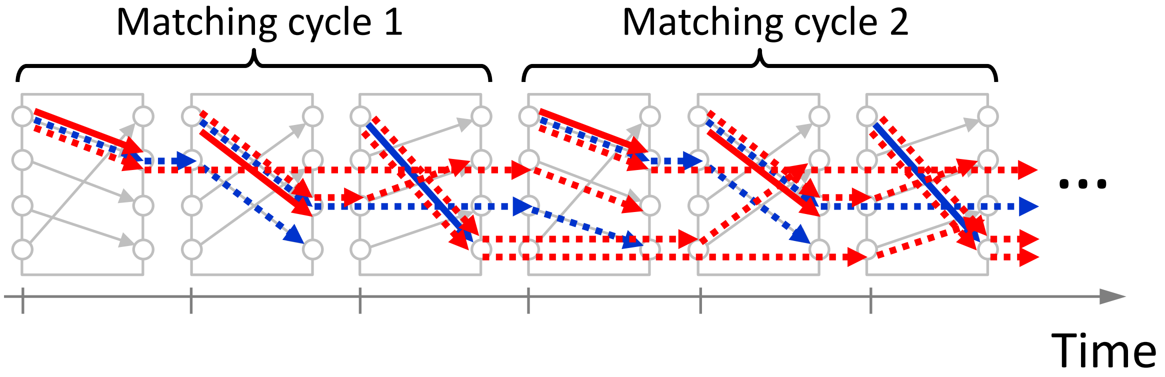
Problem: single flow → 33% throughput

- Hint at improvement: network is underutilized

2-hop forwarding better for sparse traffic

- Not new: Valiant ('82) & Chang et al. ('02)

Node 1 → 4, 3, 2
Node 2 → 3, 4, 1
Node 3 → 4, 1, 2
Node 4 → 1, 2, 3



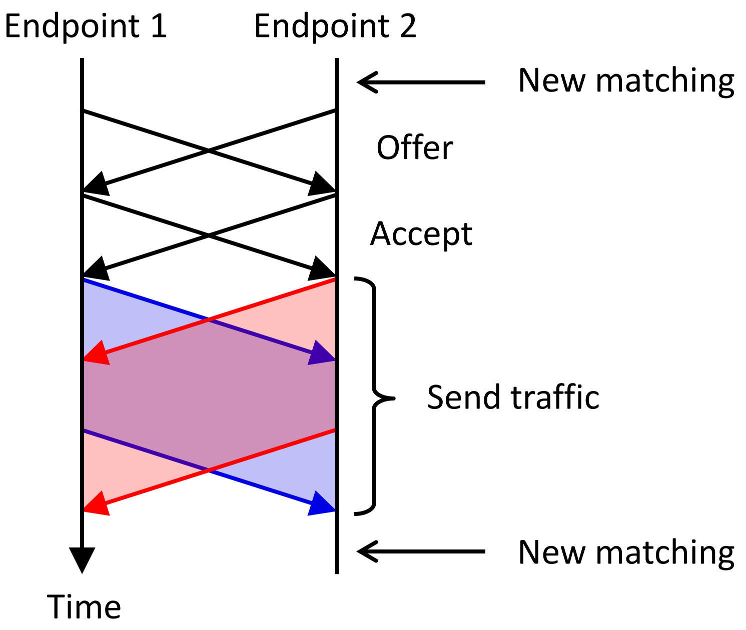
Throughput:	**Single flow**	**33% (1-hop)**	**→ 100% (2-hop)**
	Uniform traffic	**100% (1-hop)**	**→ 50% (2-hop)**

- Optimization: can we adapt between **1-hop** and **2-hop** forwarding?

RotorLB: adapting between 1 & 2-hop forwarding

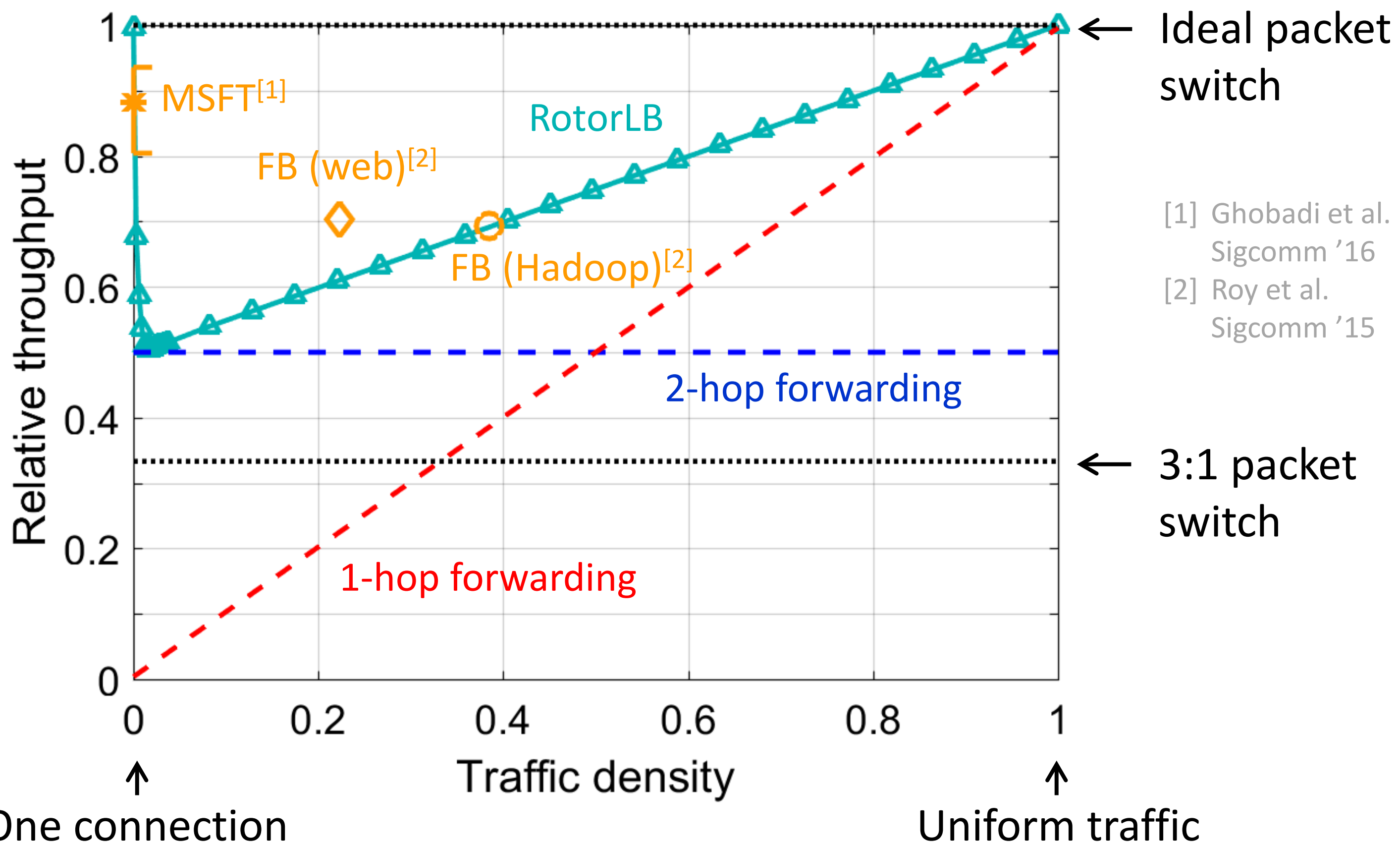
RotorLB (Load Balancing) overview:

- Default to 1-hop forwarding
- Send traffic over 2 hops only when there is extra capacity
- Discover capacity using in-band pairwise protocol:



→ **RotorLB is fully distributed**

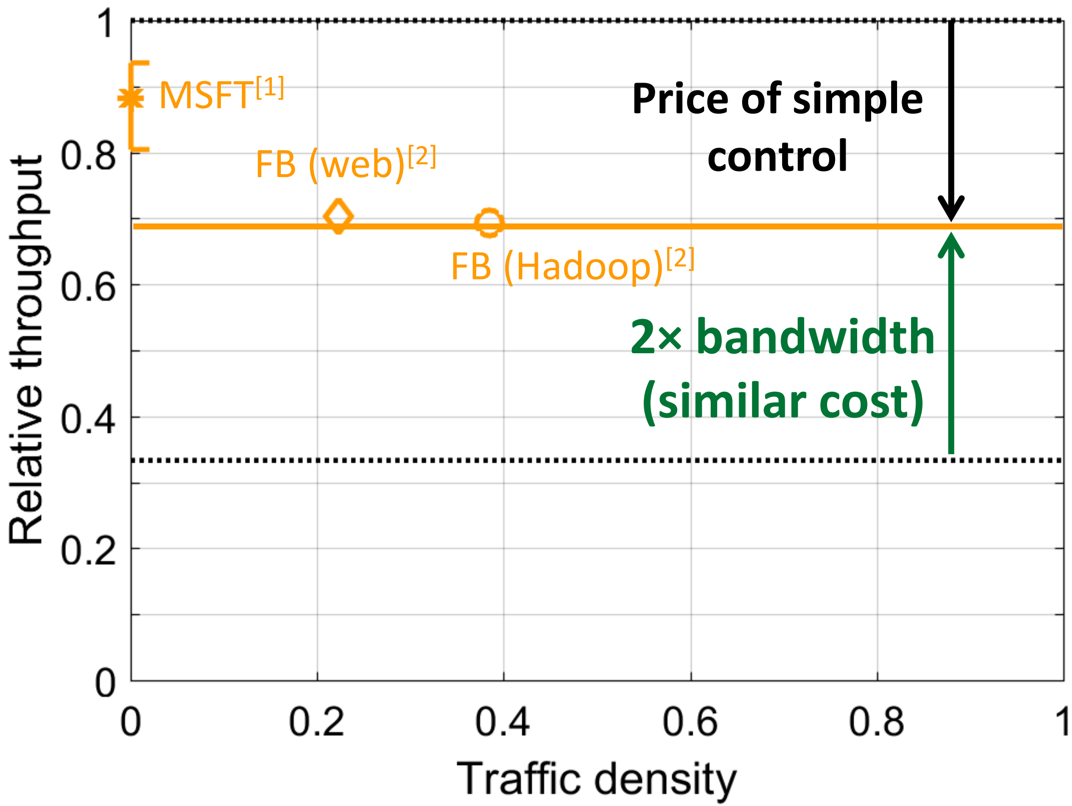
Throughput of forwarding approaches (256 ports)



[1] Ghobadi et al. Sigcomm '16
[2] Roy et al. Sigcomm '15

Throughput of forwarding approaches (256 ports)

UCSDCSE

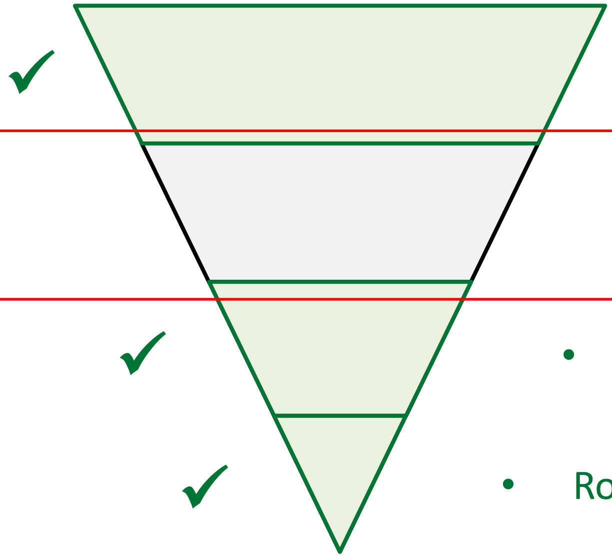


← Ideal packet switch

← 3:1 packet switch

[1] Ghobadi et al. Sigcomm '16
[2] Roy et al. Sigcomm '15

RotorNet architecture overview

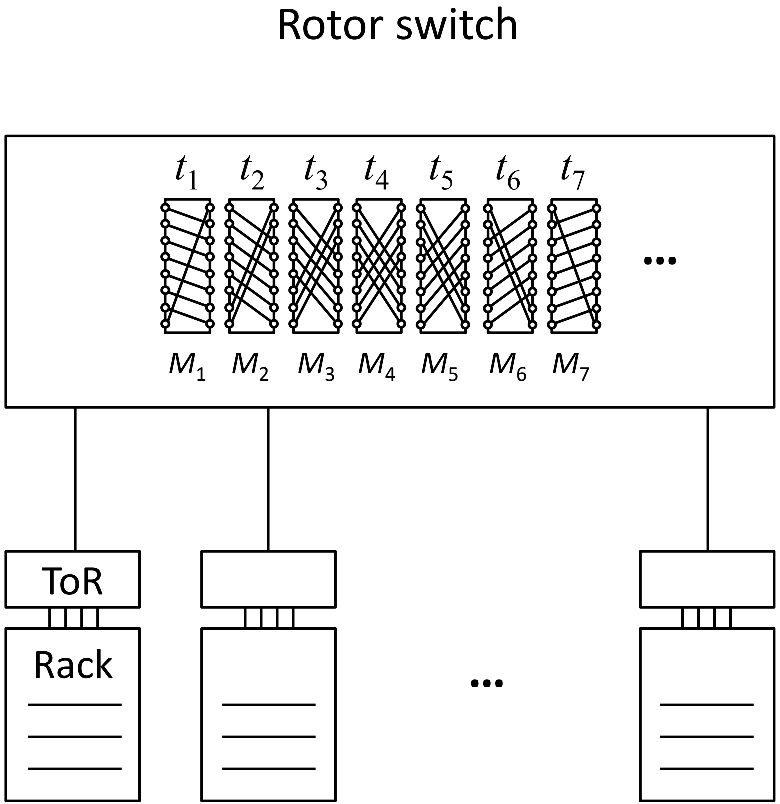


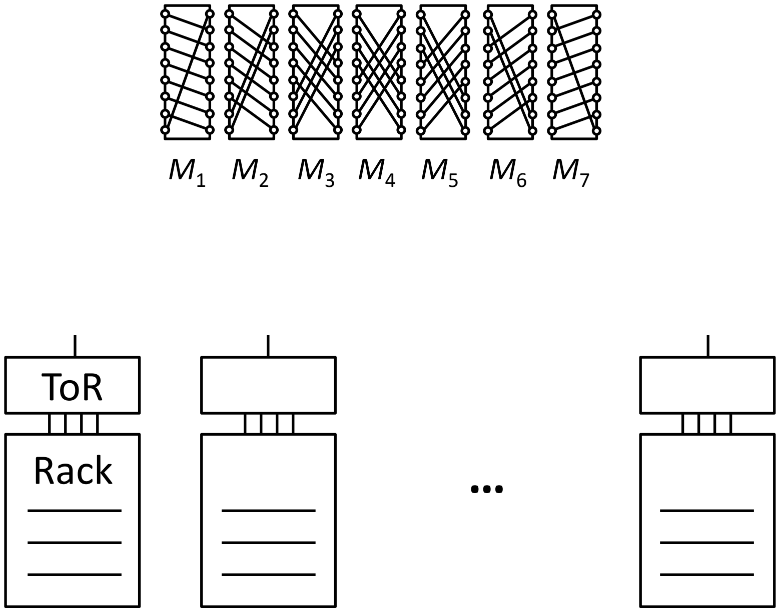
- RotorLB → **Distributed, bounded throughput**
- **Topology?**
- Optical Rotor switch → **More scalable**
- Rotor switching model → **Simpler control**

How should we build a network from Rotor switches?

At large scale:

- **High latency:** Sequentially step through many matchings
- **Fabrication challenge:** Monolithic Rotor switch with many matchings
- **Single point of failure**





M_1 M_2 M_3 M_4 M_5 M_6 M_7
ToR
Rack
…

Distributing Rotor matchings = lower latency

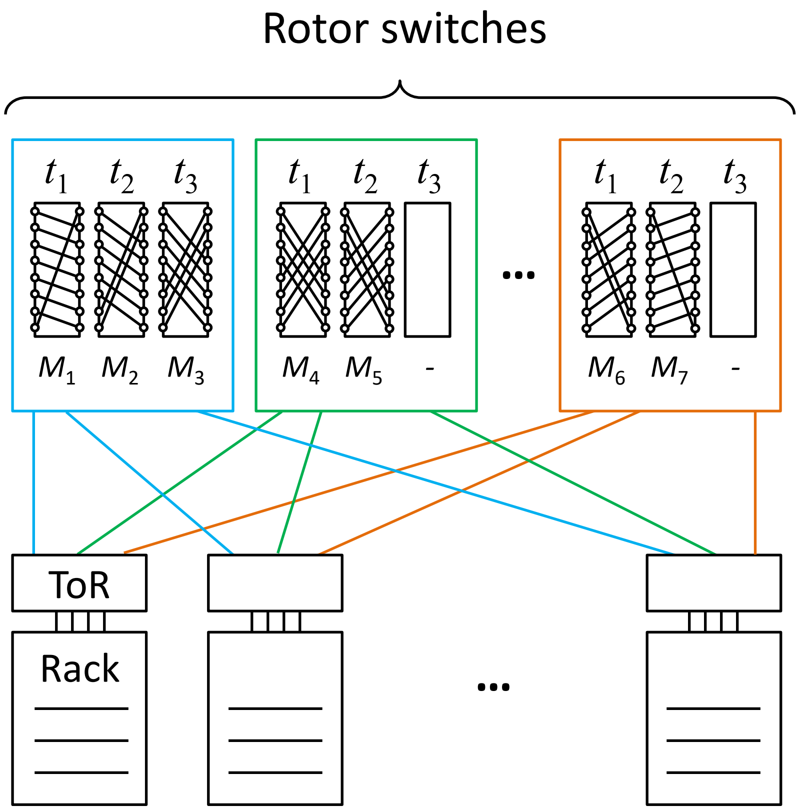
Fault tolerant

Reduced latency:

- Access matchings in parallel

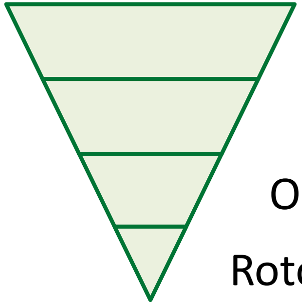
Simplifies Rotor switches:

- Matchings << ports
- More scalable, less expensive



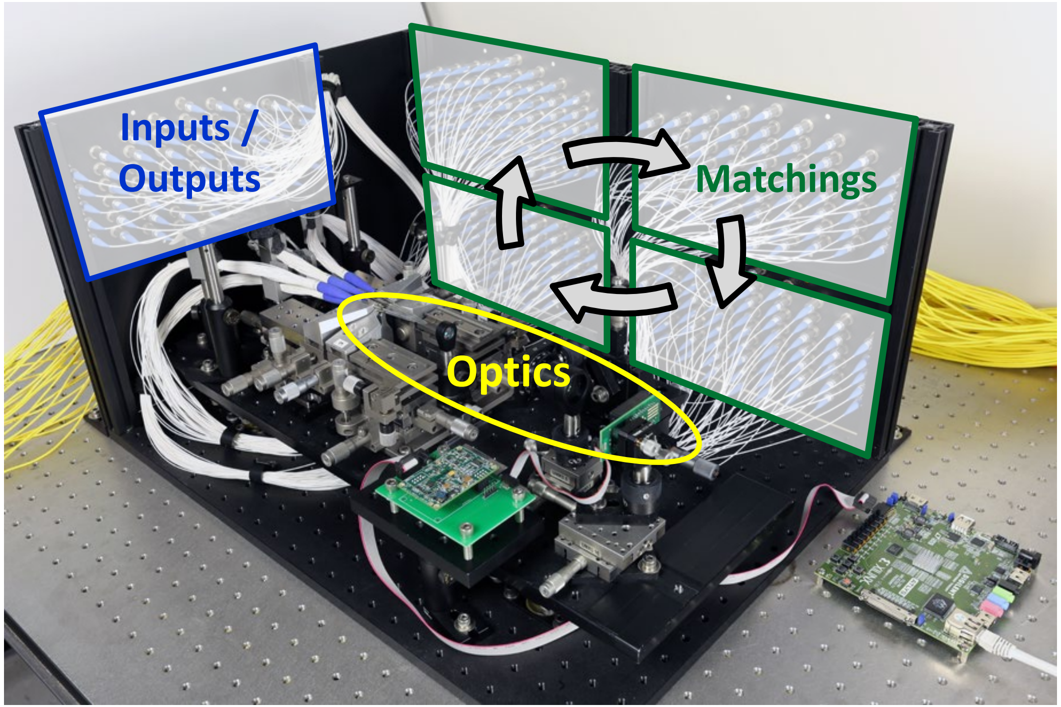
Rotor switching is feasible today

Validated feasibility of entire architecture:
(8 endpoints)



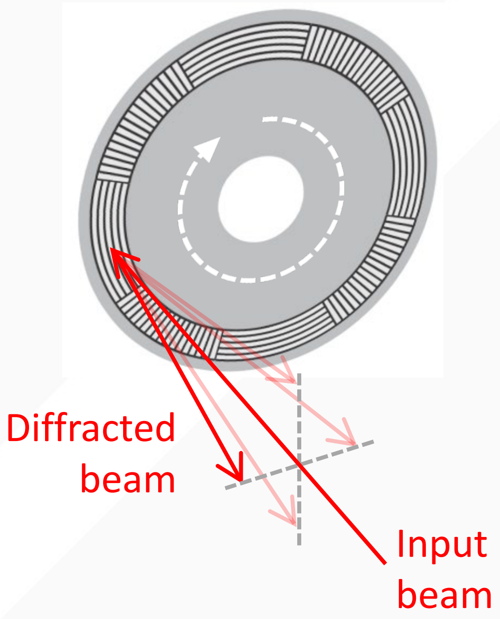
100× faster switching than crossbar

Prototype Rotor switch



SEQUENTIAL SWITCHING ENABLES NEW APPROACH TO BEAMSTEERING

"Pinwheel" sequential beam deflector



=

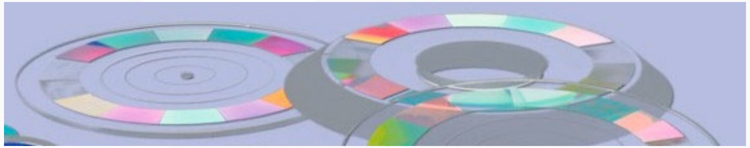
High-speed spindle

(e.g. commercial 3.5” 7200 RPM drive)

+

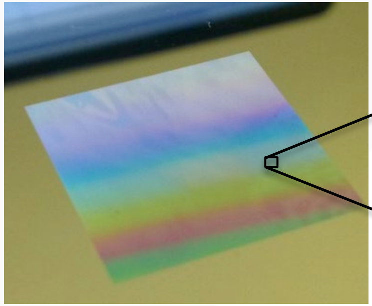
Faceted disk

(custom patterned with diffraction gratings)

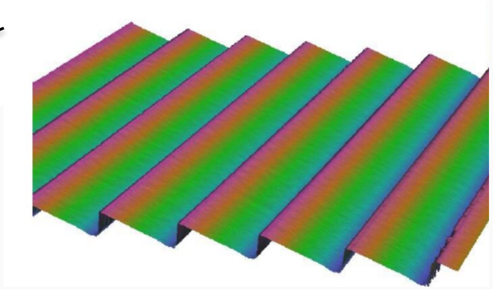


GRATING FABRICATION USING GREYSCALE LASER WRITING

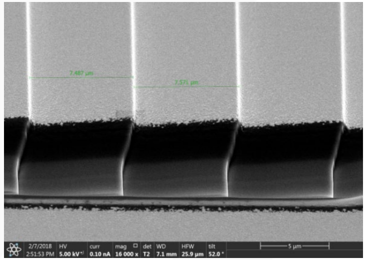
Laser-written photoresist test grating
(with gold coating)

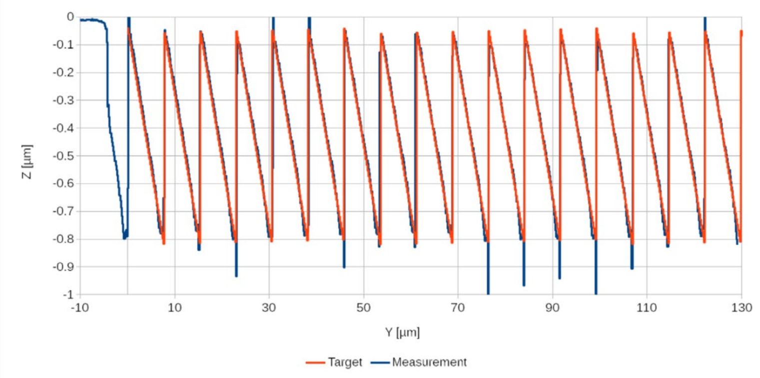


Surface profile of laser-written grating



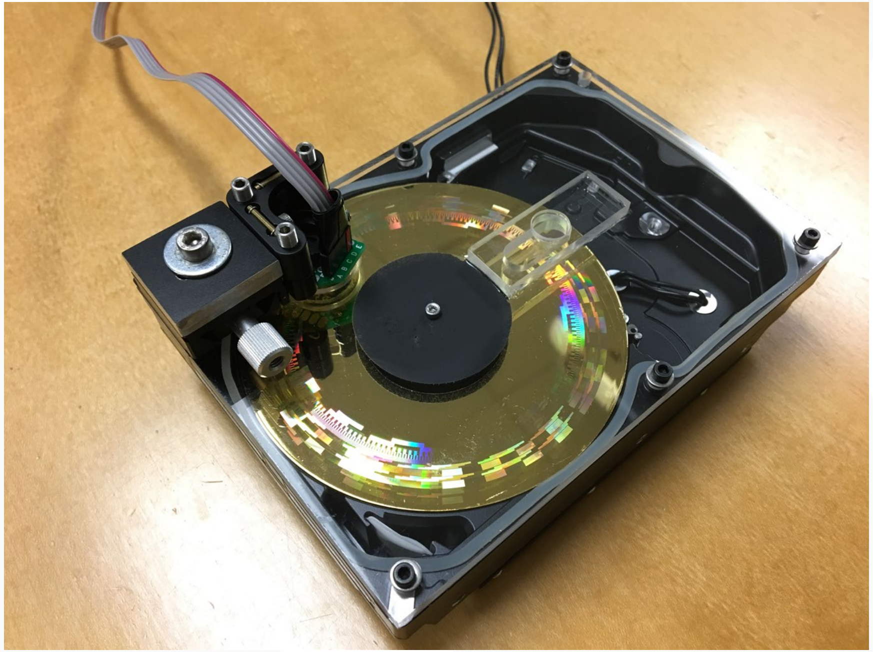
SEM image
Pitch ≈ 6.67 µm, 150 lines / mm





Initial results indicated that laser writing can produce the features needed.

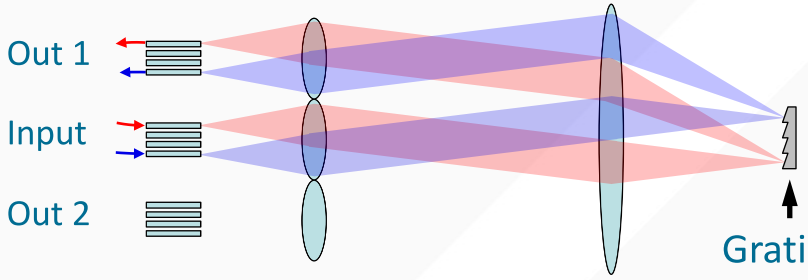
PROTOTYPE PINWHEEL IN 3.5" HGST DESKSTAR NAS DRIVE

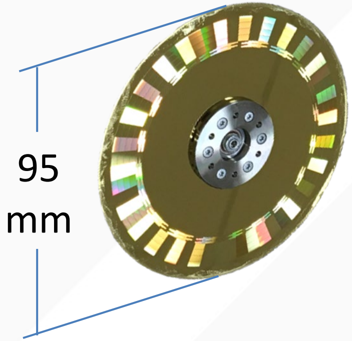


With encoder, encoder tracks, and clear cover

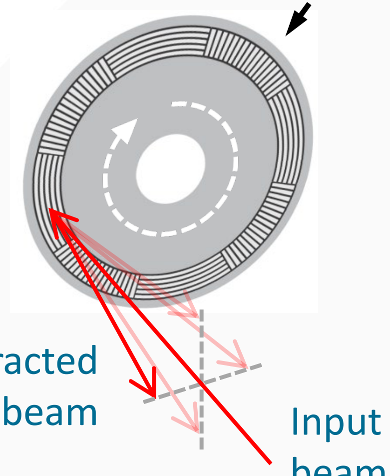
ROTOR SWITCH PROTOTYPE

Optical layout:

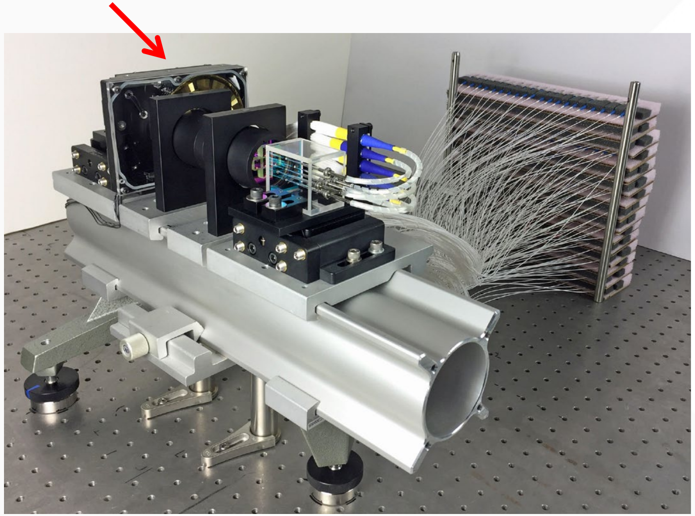




Laser-written
grating pinwheel



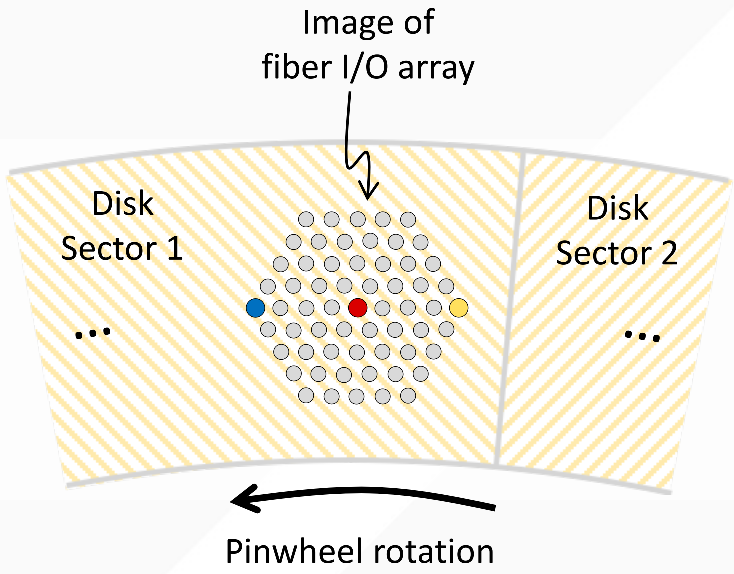
(WD) HGST Deskstar NAS drive



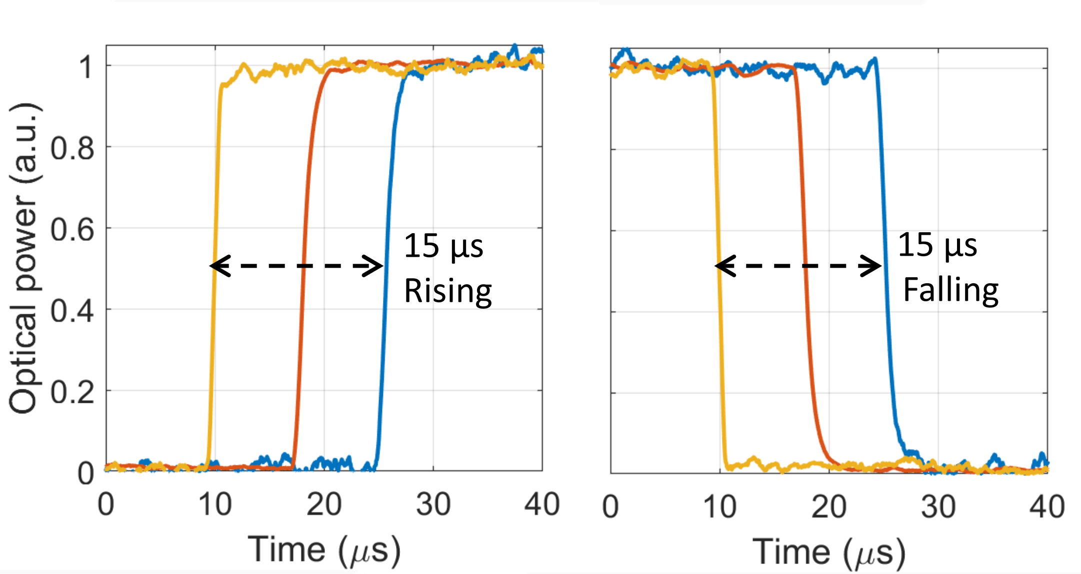
Crosstalk:	< 30 dB
Operating spectrum:	> 120 nm
2-pass insertion loss:	5 – 8 dB*

(*can be improved with better grating)

THE PINWHEEL ENABLES MICROSECOND-SCALE SWITCHING

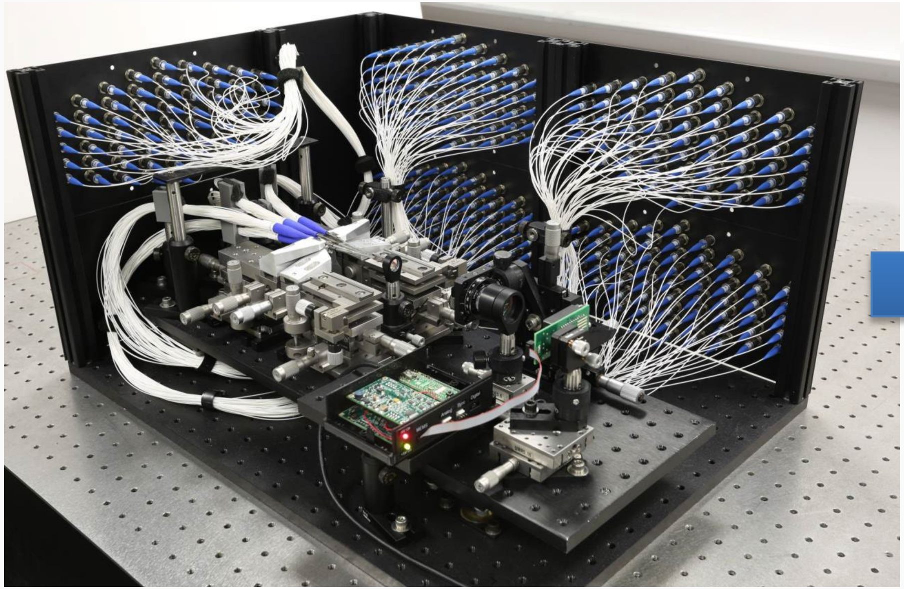


Switching transitions



15 μs reconfiguration @ 7200 RPM
(1,000 x faster than commercial MEMS OXC)

IMPROVED PERFORMANCE WITH NEW PROTOTYPE



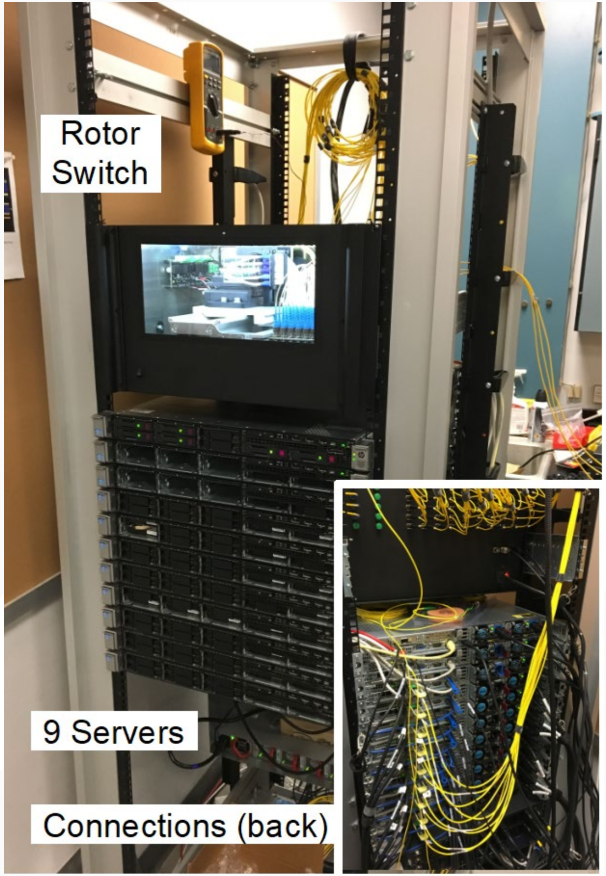
1st Prototype: MEMS selector switch

- *Higher loss optics on enclosed ½" breadboard*
- *150 µs switching*
- *I/O to external connection patch panels*

2nd Prototype: "rotor" switch with pinwheel

- *Lower loss optics mounted on vibration-isolated rail*
- *15 µs switching (@ 7200 RPM)*
- *I/O with 4x internal connection patch panels*

RACK MOUNTED TESTING OF NEW ROTOR SWITCH PROTOTYPE



RotorNet scales to 1,000s of racks

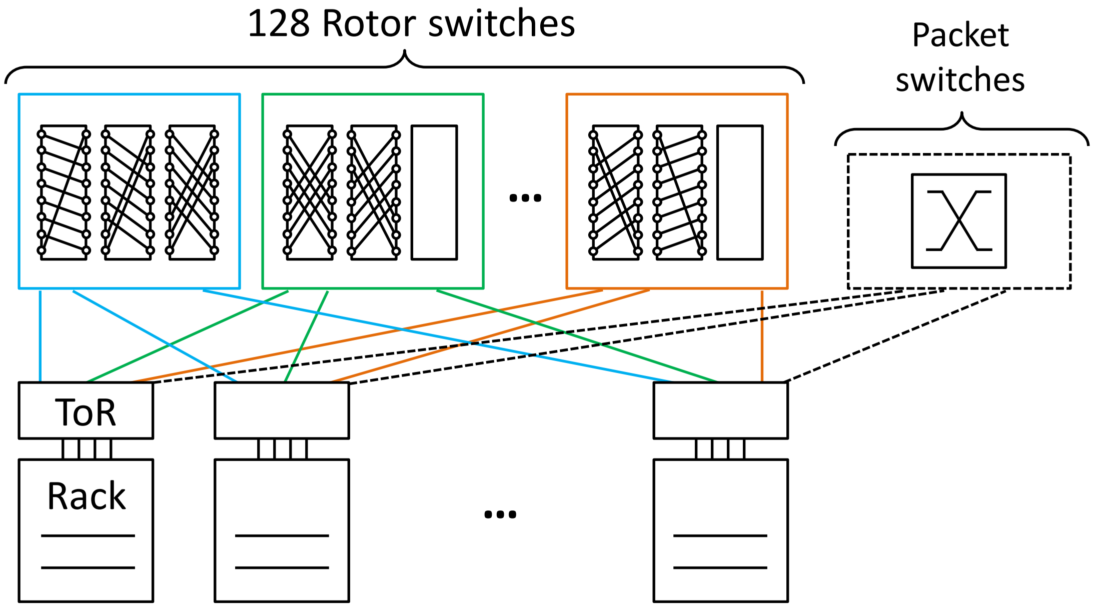
- Rotor switch design point: 2,048 ports, 1,000× faster switching than crossbar

  Details in: W. Mellette et al., *Journal of Lightwave Technology* '16
  W. Mellette et al., *OFC* '16

- 2,048-rack data center:
  → **Latency (cycle time) = 3.2 ms**

- Faster than 10 ms crossbar <u>reconfiguration time</u>

- Hybrid network for low-latency applications



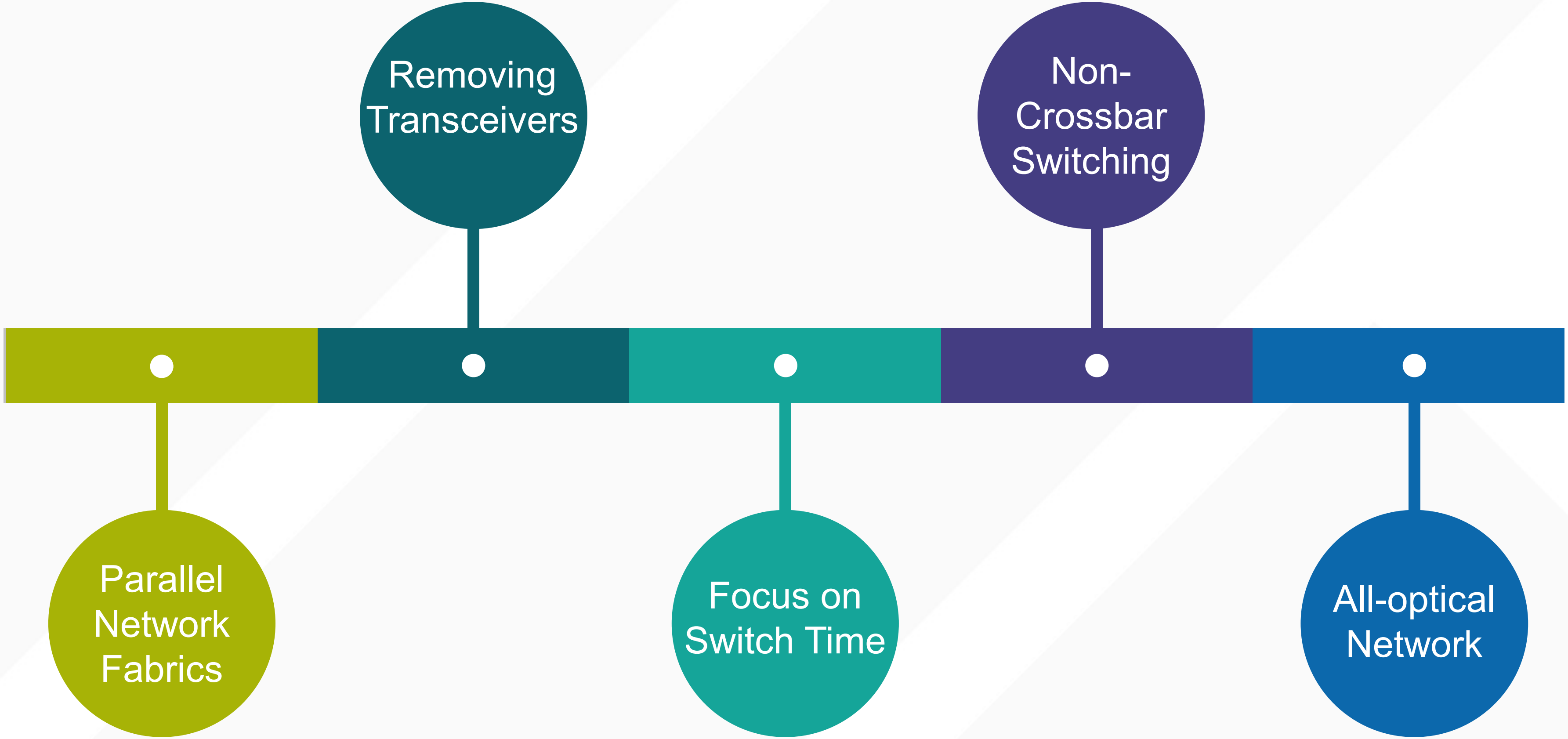
RotorNet component comparison

Network	# Packet switches	# Transceivers	# Rotor switches
3:1 Fat Tree	2.6 k	103 k	0

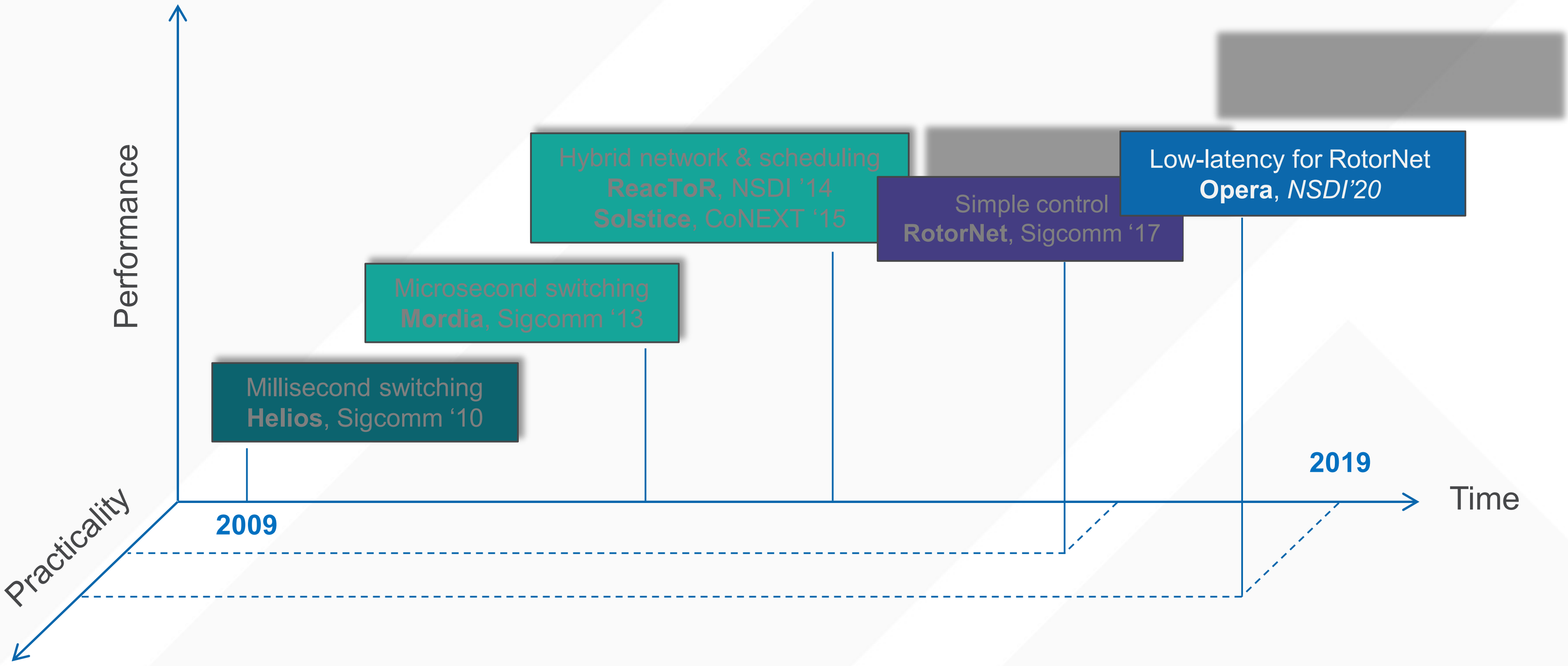
RotorNet delivers:

- Today: Bandwidth 2× less expensive
- Future: Cost advantage grows with bandwidth
- **Benefits of optical switching without control complexity**

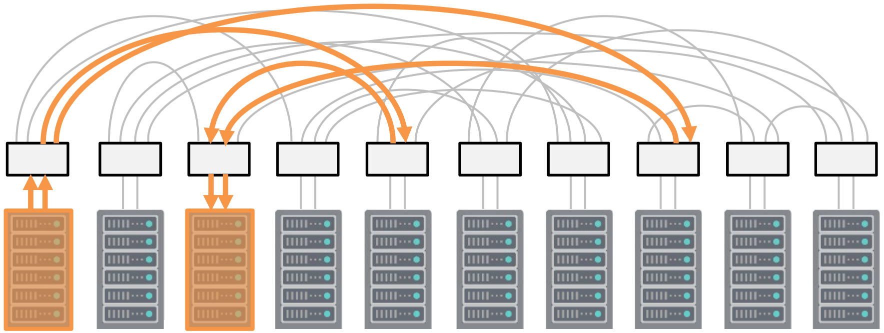
RESEARCH TIMELINE: DIRECTION 2: OPTICAL NETWORKS



REMOVING TRANSCEIVERS



Expander graph networks – an alternative to Fat Tree topologies

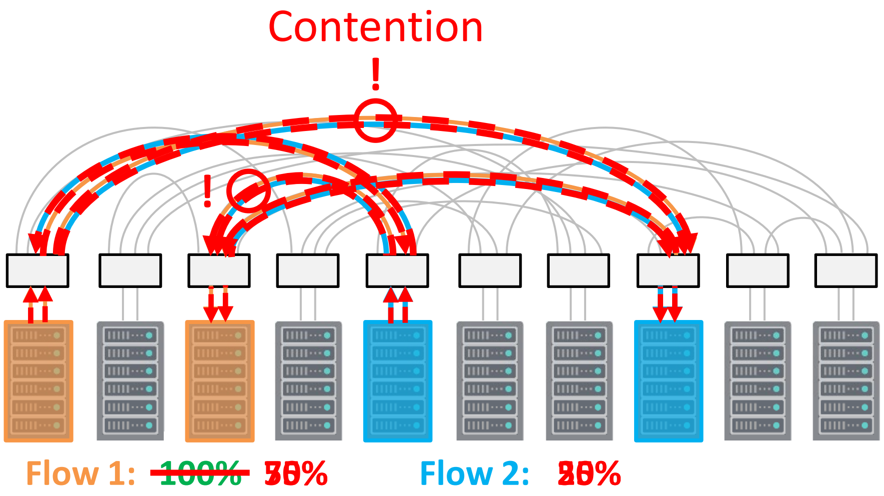


✓ Similar hardware, cost, and power savings to an oversubscribed Fat Tree

✓ Improved throughput vs oversubscribed Fat Tree at low load

“Bandwidth tax” – Reduction in throughput at high traffic loads
– Proportional to average path length

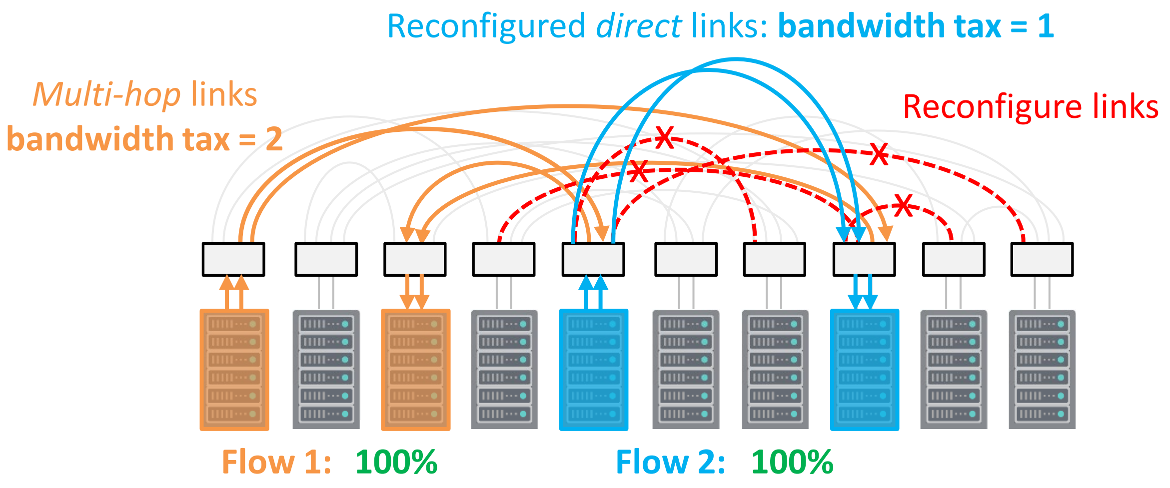
Bandwidth tax limits throughput in expander networks



Bandwidth tax = 2 → Throughput = 50% at high load

→ Is it possible to support high loads while reducing cost and power?

Reconfigurable networks enable higher throughput



✓ Reconfiguration permits high throughput at high load

Added complexity: how do we decide which links to reconfigure and when?

→ "RotorNet" (Sigcomm '17) – fixed schedule of direct circuits

Today's circuit switching technologies reconfigure too slowly → high latency

Our contribution: we can have the best of static *and* reconfigurable

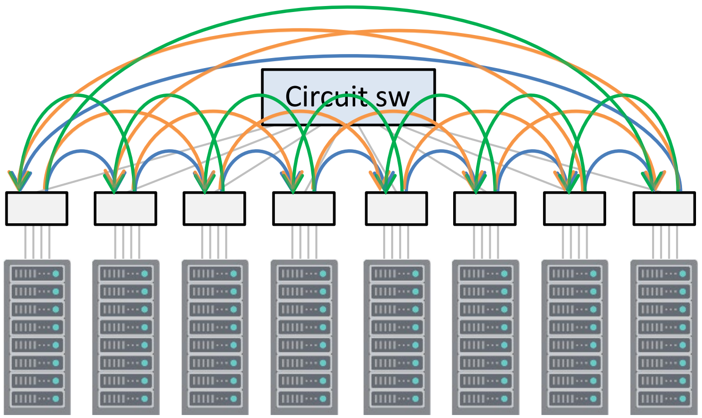
Reconfigurable networks:	high latency	high throughput
Expander networks:	low latency	low throughput

Workload = Short flows + Long flows

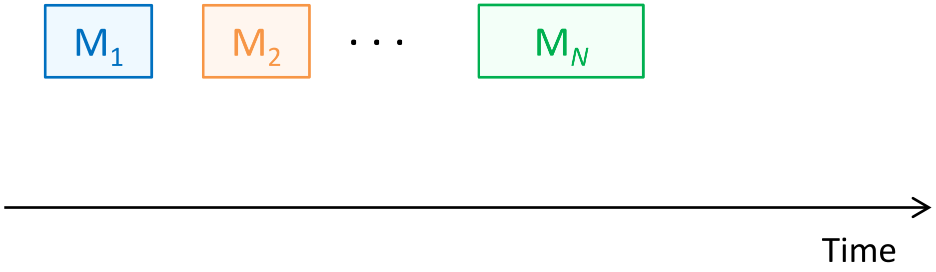
Latency-bound *Throughput-bound*

"Opera" – combining expanders and reconfiguration in a single, unified network

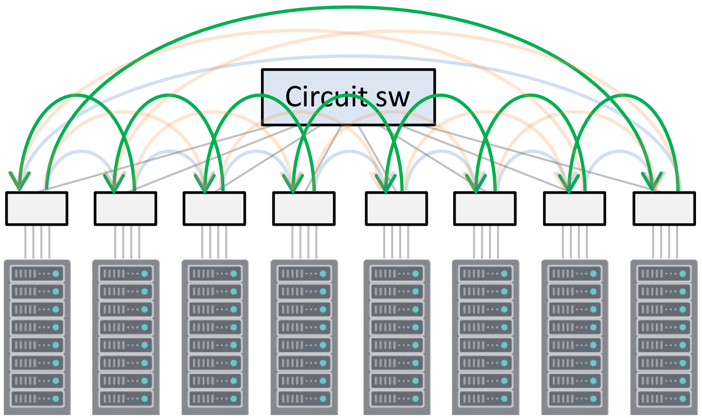
Opera’s design – part 1: providing low-bandwidth-tax connectivity



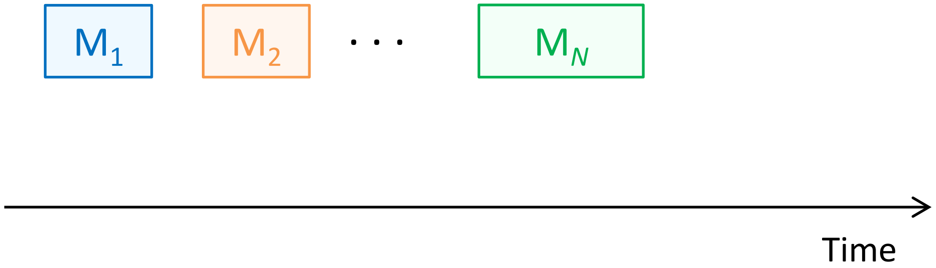
Full, direct inter-rack connectivity with *N* matchings:



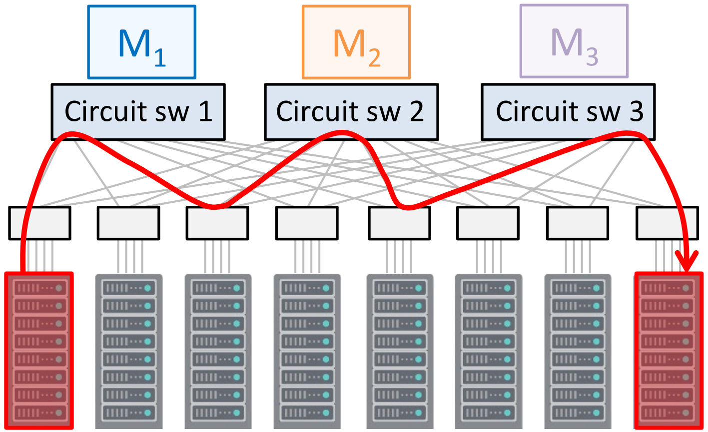
Opera’s design – part 1: providing low-bandwidth-tax connectivity



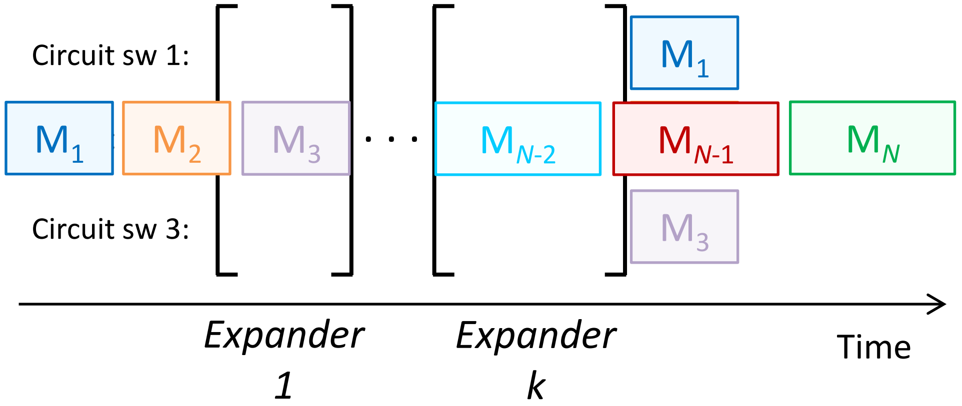
Full, direct inter-rack connectivity with N matchings:



Opera’s design – part 2: providing low-latency connectivity



Full, direct inter-rack connectivity with *N* matchings:



- Short, latency-bound flows can be sent immediately over multi-hop paths (high BW tax)
- Long, throughput-bound flows can wait for direct paths (low BW tax)

Key property: Opera only pays a bandwidth tax for short flows → **lower average tax**

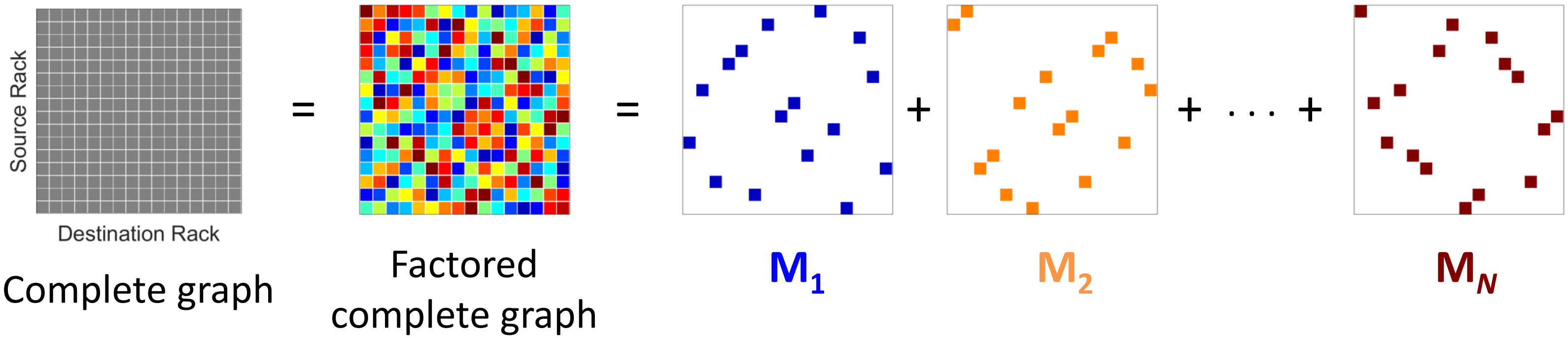
Choosing matchings

i. Expansion — Union of 3 or more randomly-structured matchings is an expander [1]

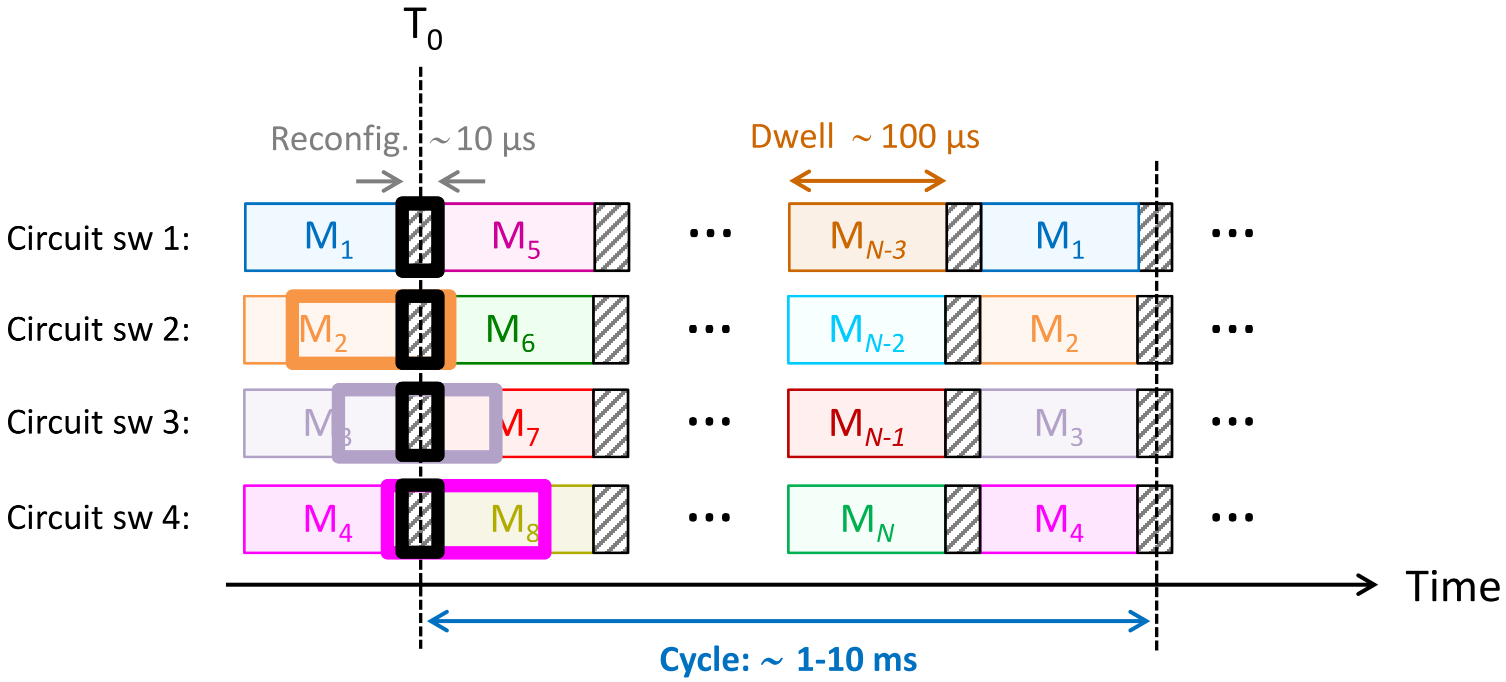
[1] N Alon, "Eigen values and expanders," Combinatorica, 6(2), 1986.

ii. Direct connectivity between all racks over time

Factor complete graph into *N* randomly-structured & disjoint matchings:



Offsetting reconfigurations for continuous connectivity



Time to wait for direct path → cutoff between “short” & “long” flows

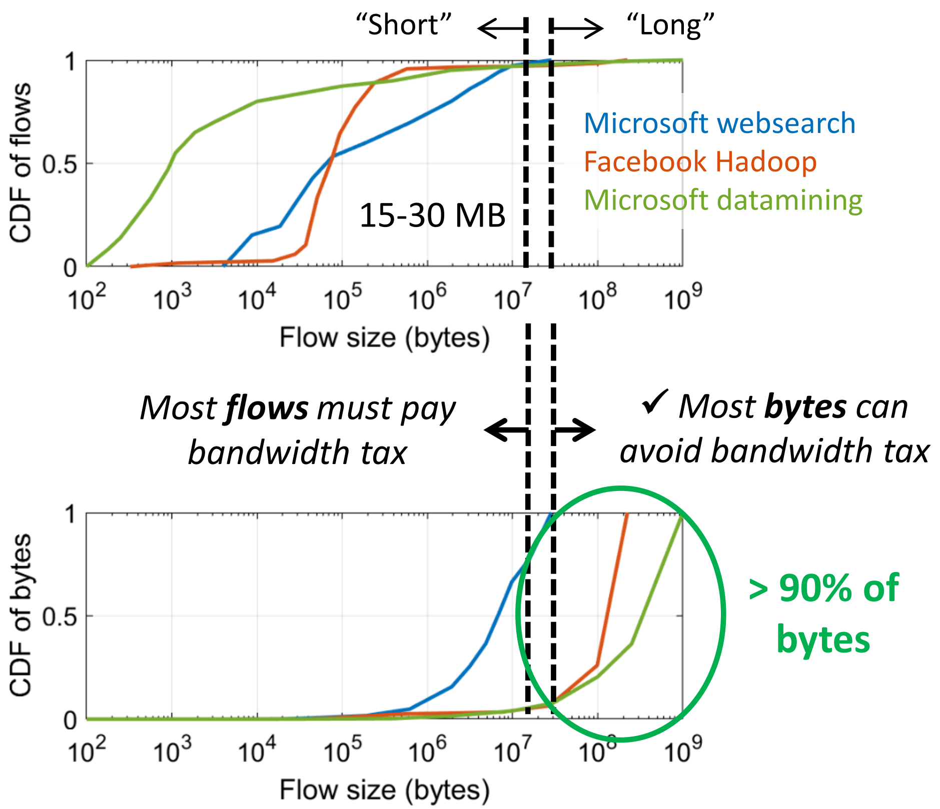
Opera is well-suited for many published data center workloads

<u>Quantifying the cutoff</u>

For 10 Gb/s – 100 Gb/s links:

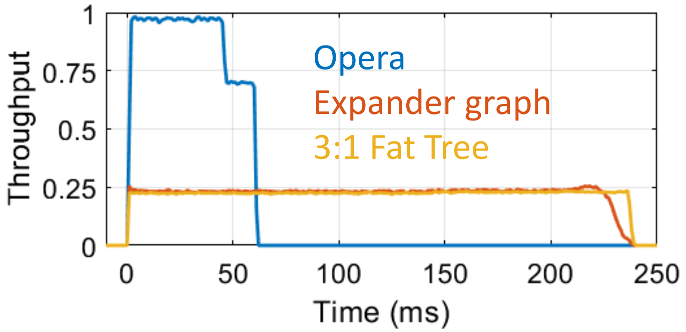
- Long flows ≥ 15-30 MB
  can afford to wait for direct paths
- Short flows < 15-30 MB
  cannot wait for direct paths

Published data center flow distributions:



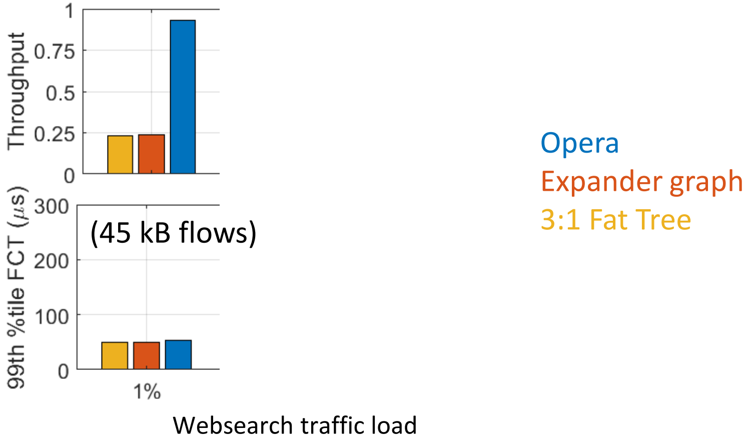
Packet simulations: throughput and latency for 100-rack network

Workload 1: All-to-all shuffle
(favorable)



→ 4x higher throughput & faster completion

Workload 2: Shuffle + MSFT websearch workload
(challenging)



→ 2-4x higher throughput &
equivalent completion times for short flows

Workload 3: MSFT datamining (100 B – 1 GB flows)

→ 60% higher admissible load with equivalent FCTs

Practical considerations

Fault tolerance:

- Full connectivity maintained with 4% of links, 7% of ToRs, or 40% of circuit switches failed
  (Better than oversubscribed Fat Tree, not as good as static expander)
- Failures detected and disseminated within *O*(10 ms)

Prototype implementation:

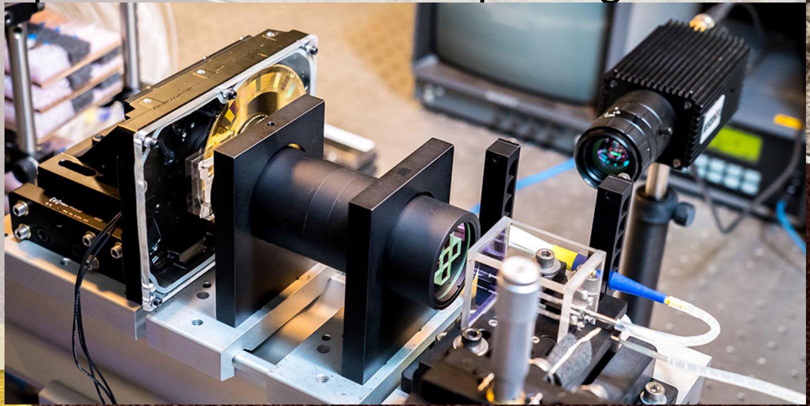
- Time-synchronized routing implemented on programmable Barefoot Tofino switch with P4
- Opera scales to 1,000’s of racks, 10,000’s of servers with commodity switch table sizes

CREDITS AND THANKS TO MY COLLEAGUES/STUDENTS

- Alex C. Snoeren
- Alex Forencich
- Amin Vahdat
- Andrew Grieco
- Anthony Lentine
- Arjun Roy
- Chang-Heng Wang
- Christopher DeRose
- Chunming Qiao
- Conglong Li
- David G. Andersen
- Douglas C. Trotter
- Feng Lu
- Geoffrey M. Voelker
- George C. Papen
- George Papen
- Glenn M. Schuster
- Guohui Wang
- Hamid Bazzaz
- He Liu
- Jordan Davis
- Joseph Ford
- Joshua Zhu
- Kai Chen
- Li Chen
- Malveeka Tewari
- Matthew K. Mukerjee
- Michael A. Kozuch
- Michael Kaminsky
- Minlan Yu
- Nathan Farrington
- Nicolas Feltman
- Pang-Chen Sun
- Rajdeep Das
- Richard Strong
- Rishi Kapoor
- Rob McGuinness
- Ryan Aguinaldo
- Shan Zhong
- Shaya Fainman
- Shayan Mookherjea
- Sivasankar Radhakrishnan
- Srinivasan Seshan
- Stefan Savage
- T. S. Eugene Ng
- Tajana Rosing
- Tara Javidi
- Vikram Subramanya
- William M. Mellette
- Yeshaiahu Fainman
- Yibo Guo

THE FUTURE OF OPTICAL SWITCHING FOR DATACENTERS/HPC

- Reaching the limits of CMOS-based packet switching
  - In terms of cost, power, performance…
- Direction 1: scale bandwidth by adding *parallel* dataplanes
- Direction 2: scale bandwidth by replacing packet switches with optical ones
  - Unique opportunity to incorporate novel optical devices such as spinning pinwheel/hard drive based switches
- Thank you for your time and attention!



UC San Diego