Datacenter Networking in the Era of Plentiful Bandwidth

George Porter

William (Max) Mellette and Alex C. Snoeren

Computer Science & Engineering Department

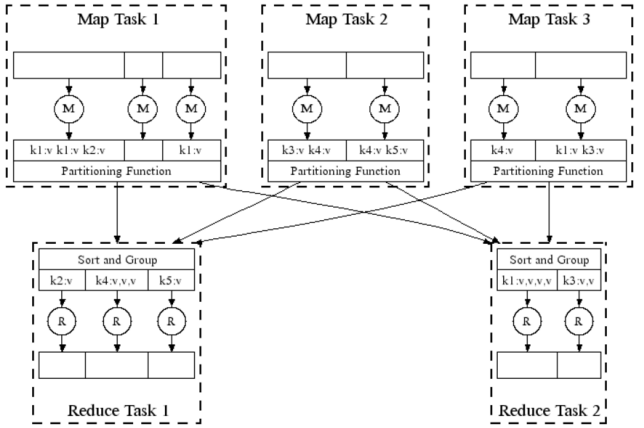
University of California, San Diego

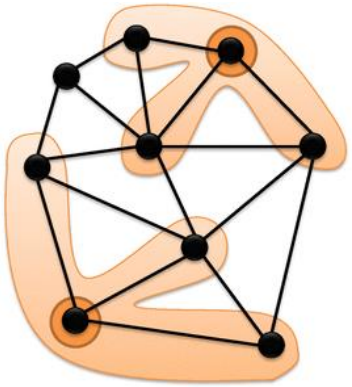
February 4, 2017

UCSDCSE

Parallel processing

Parallel software





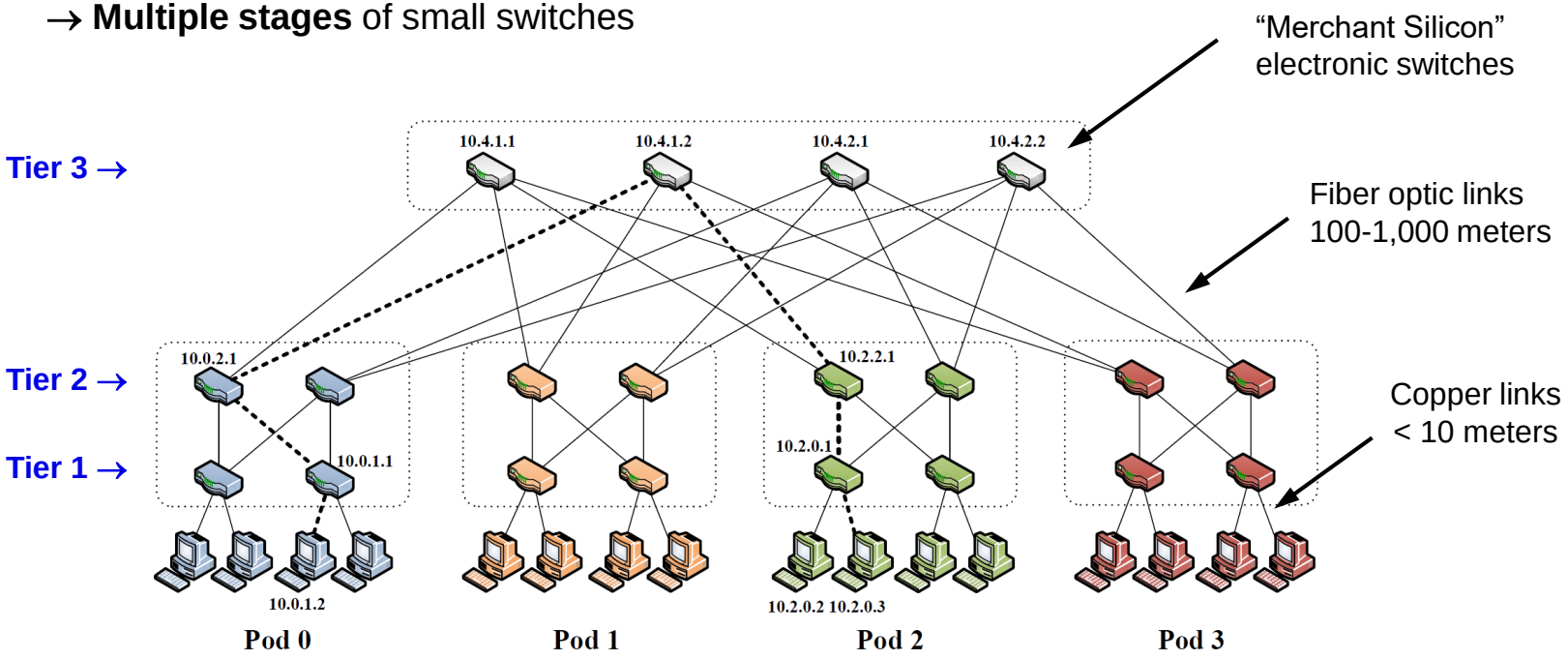
Network?

Scale out networking

Today's Data centers: Folded-Clos "FatTree" topology

Connect 10,000 – 100,000 servers with relatively small port count switches

→ **Multiple stages** of small switches

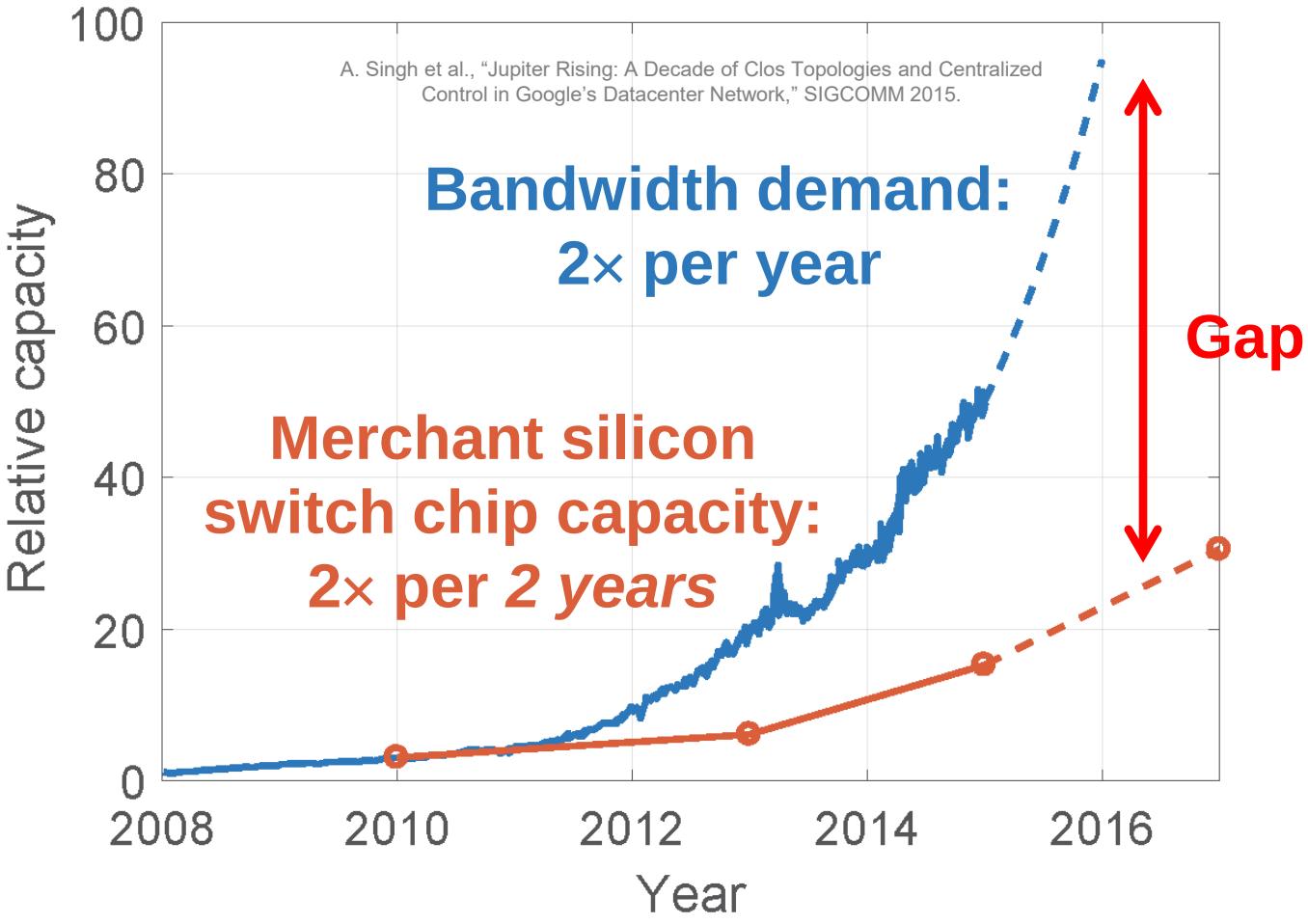


M. Al-Fares, A. Loukissas, and A. Vahdat, "A Scalable, Commodity Data Center Network Architecture," SIGCOMM 2008.

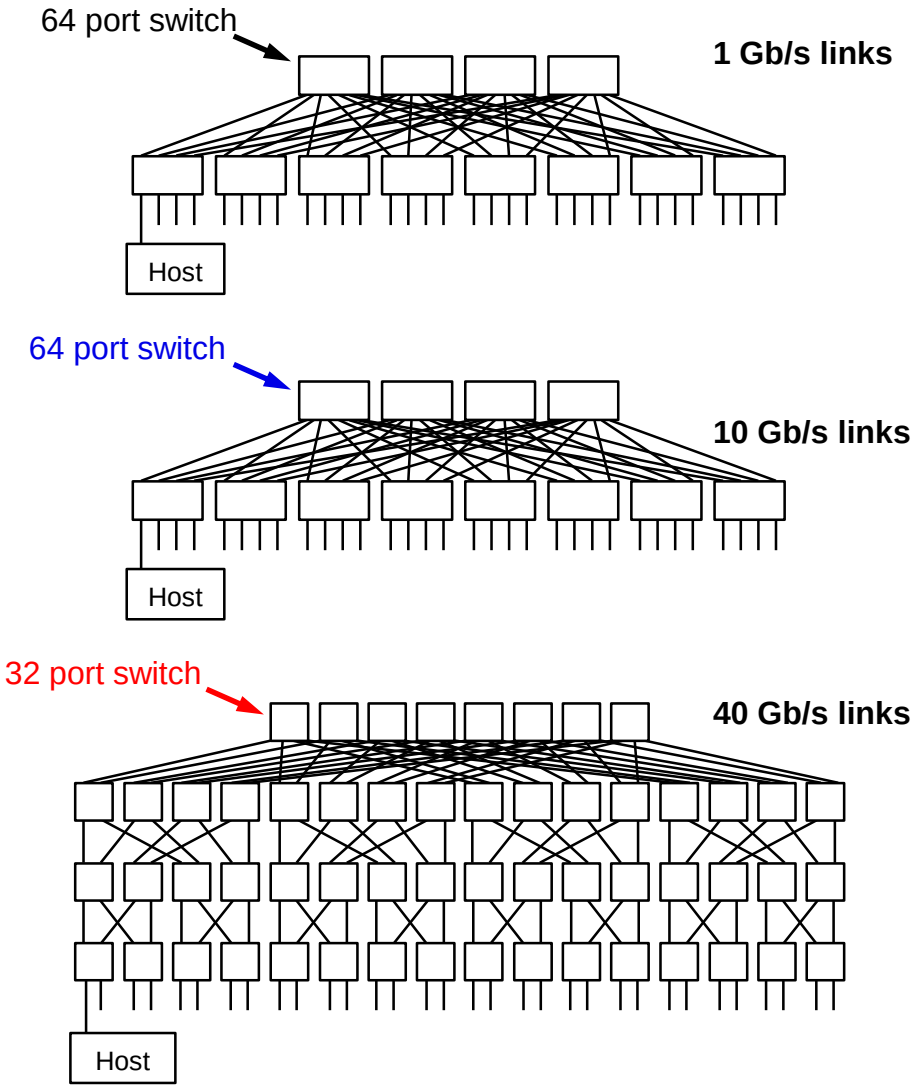
Scaling datacenter networks

UCSDCSE

Google's data center traffic (SIGCOMM 2015)

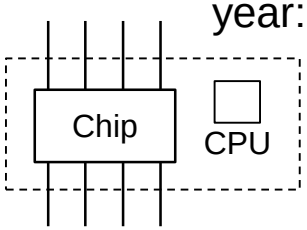


Scaling FatTrees is becoming expensive



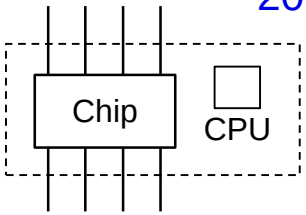
Merchant Silicon

year: ~ 2004



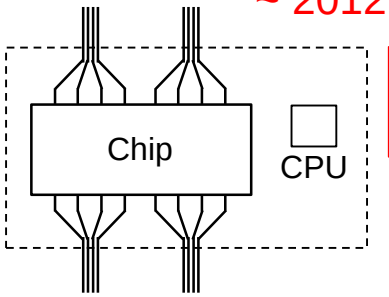
64 ports
× 1 Gb/s

~ 2010



64 ports
× 10 Gb/s

~ 2012



32 ports
× 40 Gb/s

(128 ports
× 10 Gb/s)

Link bandwidth not the issue – the switches are not keeping up.

This talk: two approaches to improve network scalability

UCSDCSE

… How can Silicon photonics help? What are the challenges?

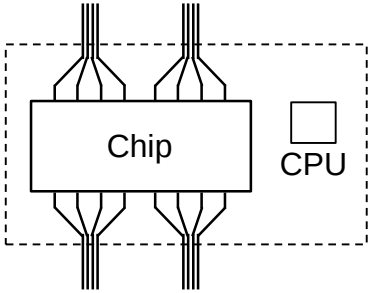
1) Leverage parallelism in electronic switching
   - Need low-cost transceiver arrays
   - Need to move data into and out of electronic switch chips

2) Optical switching and hybrid networks
   - Need low-cost high-bandwidth links
   - Silicon photonic switches

High bandwidth via “ganging” together multiple switch ports

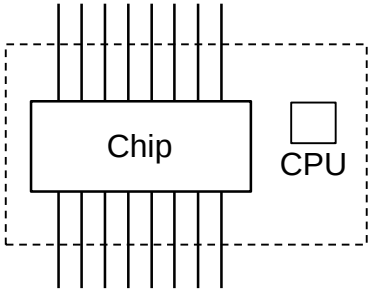
Ex. Broadcom’s Tomahawk switch:

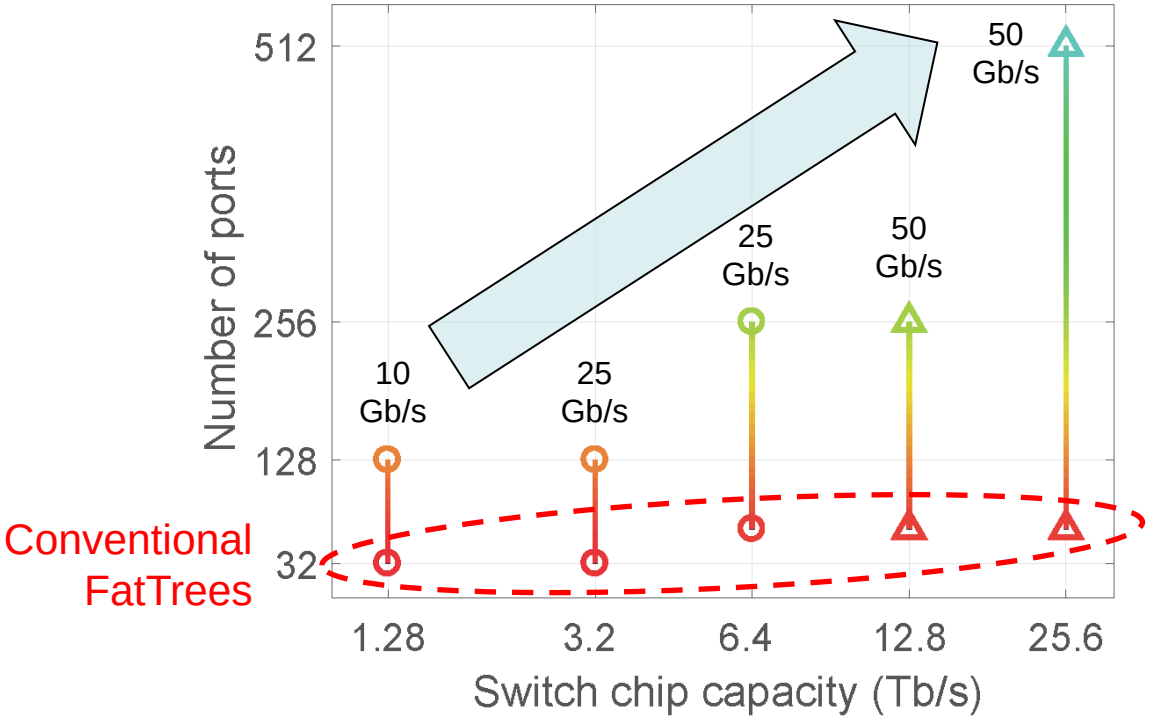
32 ports @ 100 Gb/s



OR

128 ports @ 25 Gb/s



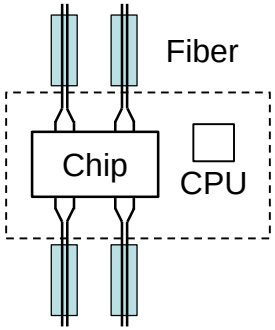


The move to multistage chassis switching

"Traditional" packet switch

16 ports @ 100Gb/s

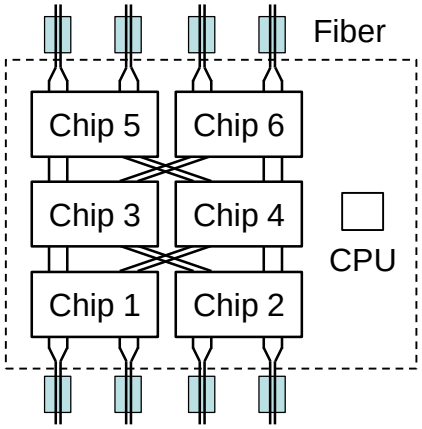
Facebook's "Wedge"

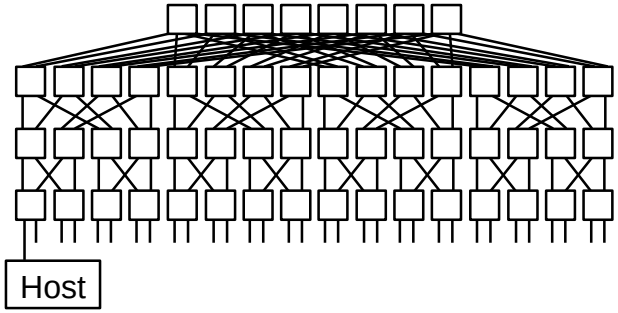


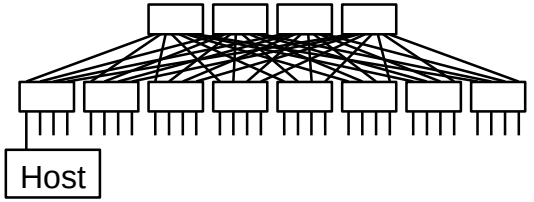
Multistage chassis switch

128 ports @ 100Gb/s

Facebook's "6-pack"





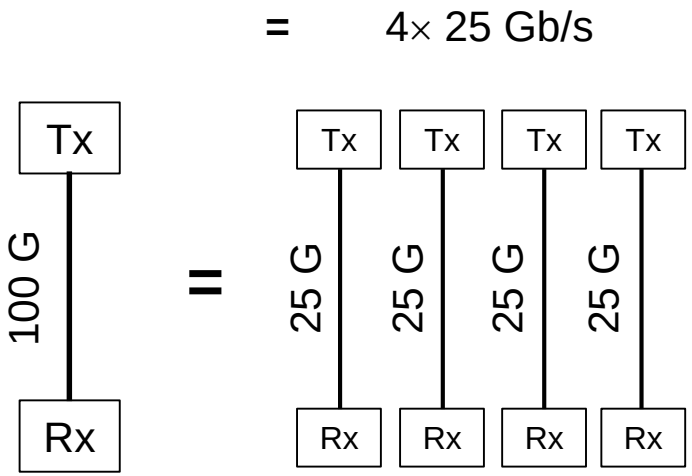


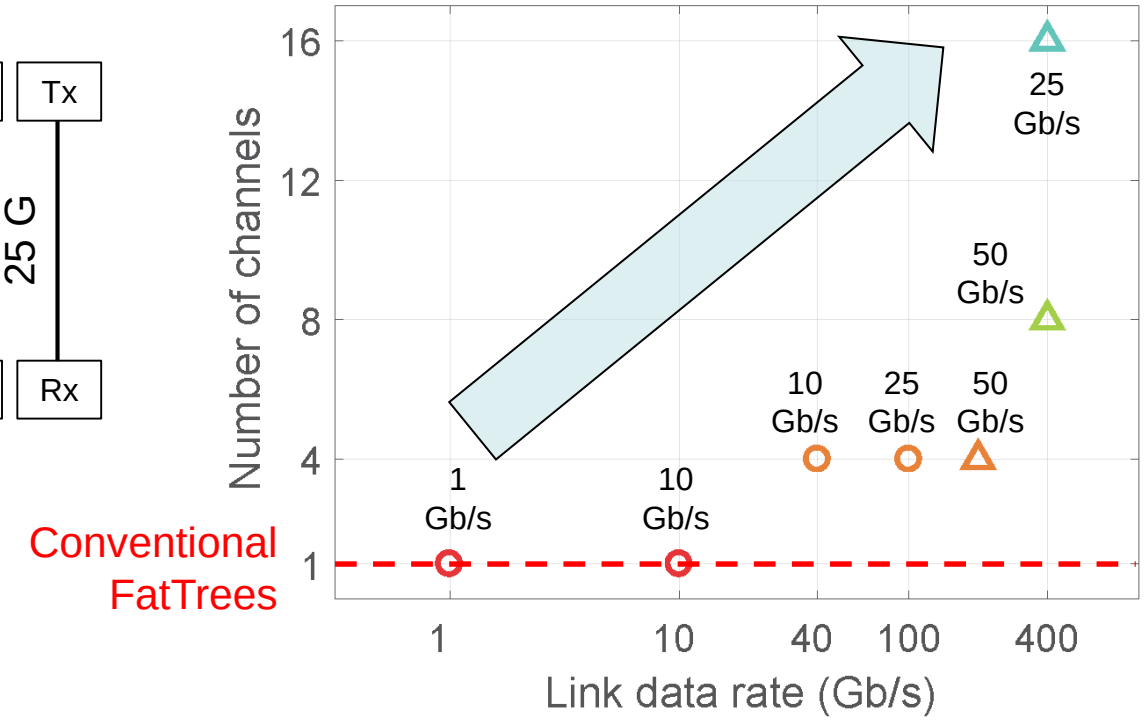
Fully-provisioned network – 8,192 end hosts

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

Link channel count is increasing

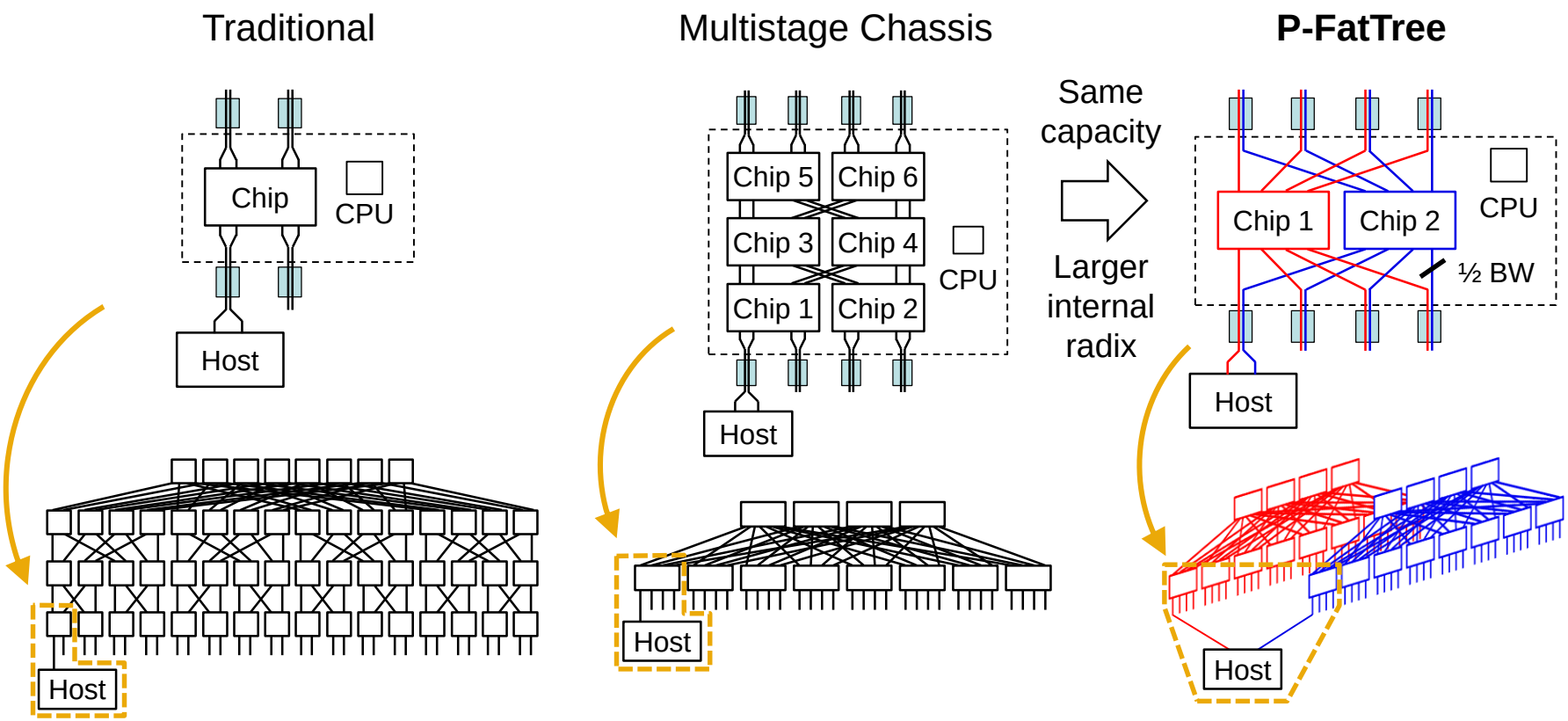
Ex. 100 Gb/s optical link:





→ **Potential idea: Expose hardware parallelism instead of hiding it.**

Potential approach: parallel packet-switched networks



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
P-FatTree	**2**	**3**	**33 k**	**768**	**192**	**16 k**

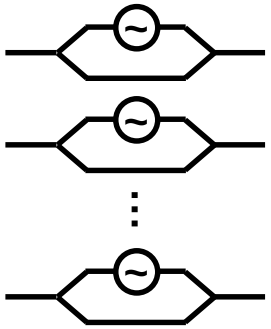
→ Lower cost, reduces switch power consumption by 2-3× and latency by 3×

Silicon photonics: transceiver integration for reduced cost

Today, transceivers account for 50% of data center fabric cost.

Parallelism:
Beyond a single modulator…
Need ***arrays*** of modulators and detectors.

Facebook “6-pack”, code.facebook.com



Pluggable transceivers
- High power
- Large form factor
- Significant cooling
- High cost

Mid board / on board
- Lower power
- Smaller form factor
- Shorter copper traces

On chip / integrated
- Lowest power
- Smallest footprint
- Lowest cost

Data centers have few-year life cycles
→ “Fail in place” may be O.K.

Challenge: data centers require going “off chip”

www.google.com/about/datacenters

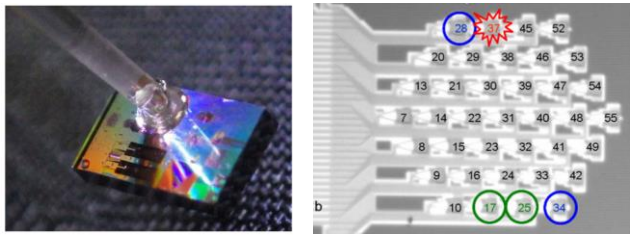
100,000’s of servers
1,000’s of racks

Data center links are 100’s of meters (warehouse) to a few kilometers (campus)

- Multimode and single-mode fiber used today.
- Standard fiber – not polarization maintaining.

Silicon photonics fiber/chip array coupling: (1) low-loss, (2) polarization insensitive, (3) broadband.

- Grating couplers –
  - Easier packaging (top of chip, large chip area)
  - 3 dB loss per coupler, limited bandwidth.
- Edge couplers –
  - Low loss, polarization insensitive, broadband.
  - Harder to package (need clean edge, linear perimeter)



2-D array of grating couplers

V. Kopp et al., “Two-Dimensional, 37-channel, high-bandwidth, ultra-dense silicon photonics optical interface,” JLT 2015

Leveraging environmental characteristics in data centers

Learning from Facebook's CWDM 4 multi-source agreement:

www.finisar.com

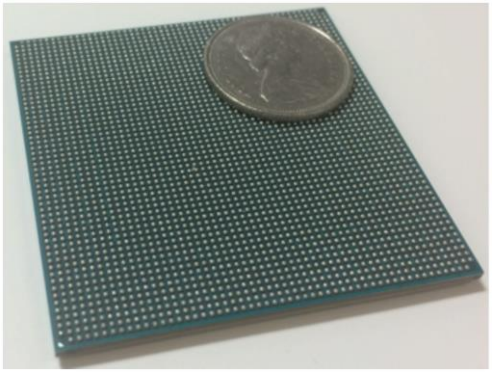
Lowered 100 Gb/s transceiver cost by:

- Reducing operating temperature range (vs. telecom)
- Reducing link budget (2 km reach)
- Reducing reliability requirements (2-3 year lifetime)

Critical for Silicon photonics solutions to adopt similar approach:

- Take advantage of shorter component lifetime requirements.
- Devices and packaging will still need to be robust under temperature swings.
  - Electronic chip temperatures can vary by 10's °C depending on load.
  - Active temperature control increases cost and power consumption.
- Link distance: don't need Telecom reach (i.e. 100's km)
  - Inter-chip centimeters
  - Inter-rack 100's meters
  - Inter-building kilometers

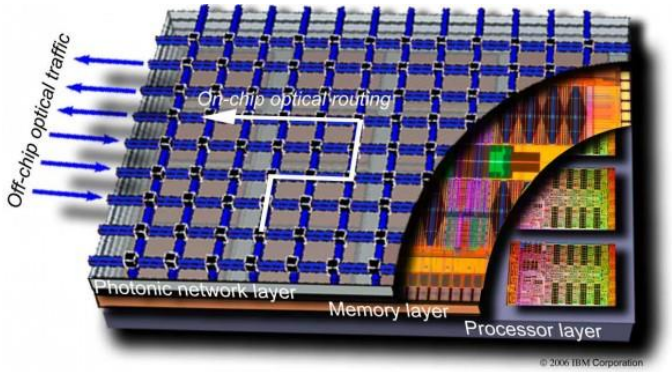
Overcoming pin bandwidth limitations



Electronic switch chips: BGA package

- Capacity = (SerDes BW) × (pins)
- SerDes bandwidth increasing slowly:
  - 10 Gb/s, 25 Gb/s, … 50 Gb/s ?
- Link bandwidth increasing faster:
  - 10 Gb/s, 40 Gb/s, 100 Gb/s, … 400 Gb/s

- How can we move the data onto / off of the chip?
  - Larger package = lower yield = higher cost.



IBM

Photonic interposers

- Optical waveguides have higher data rate than copper traces.

- **Challenges:**
  - Extreme level of integration
  - Optics very close to thermally-varying electronics.
  - Provide enough link budget to get off the board? Rack? Building?

Alternative approach: optical circuit switching

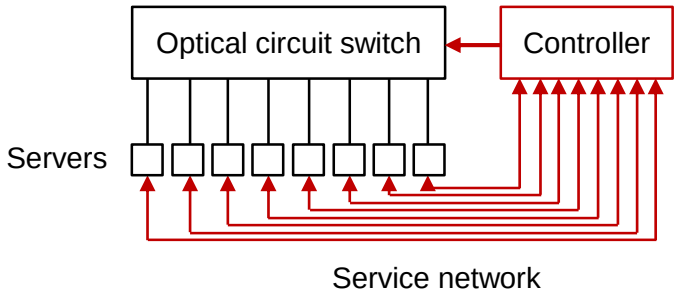
Hybrid networks: combine optical switches (OCS) & electronic packet switches

- OCS not subject to electronic pin bandwidth limitations
- Reduce the number of OEO conversions (fewer transceivers)

Note: Previous discussion on electronic switching still applies to electronic portion of hybrid network.

System-level challenge: Control plane

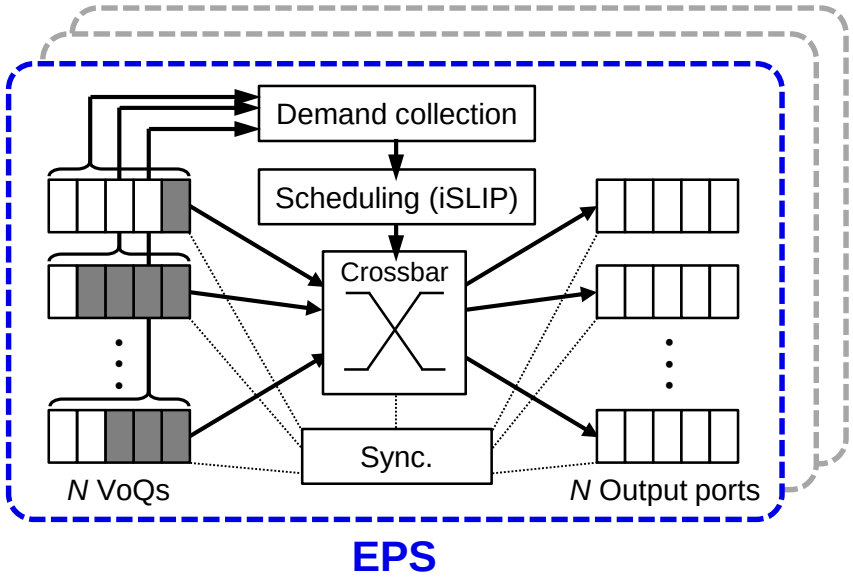
- OCSes do not inspect packets – need external control
- Typical approach: approximate demand, schedule circuits, synchronize transmission



Ongoing work aimed at simplifying control plane.

Challenge: OCS proposals increase control complexity

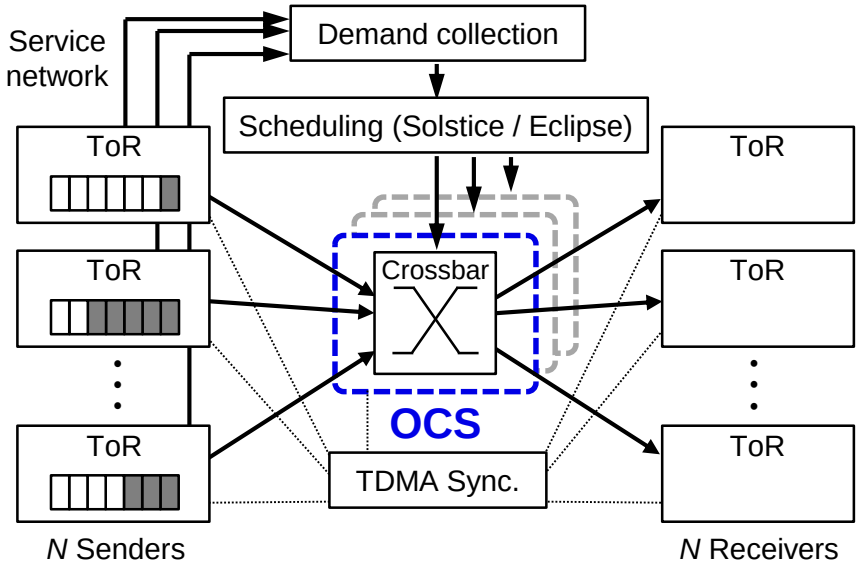
Conventional EPS network:



- Buffer, de-multiplex, and inspect packets.
- Control plane contained within each EPS:
  - VoQ polling
  - Scheduling
  - Synchronization

→ Control plane "hidden" from network

Typical OCS network:

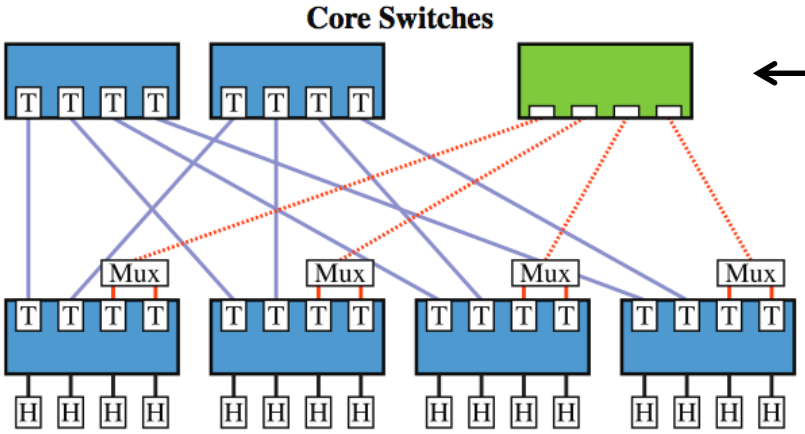


- OCS cannot buffer or inspect packets.
- Control complexity exposed to the network:
  - Demand collection from end points
  - Centralized scheduling
  - Network-wide synchronization

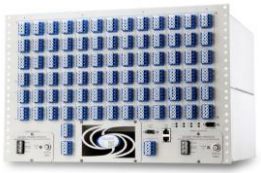
→ Centralized control with sub-millisecond resolution is difficult at scale.

Initial proposal: Helios

Helios Hybrid Datacenter Network



Freespace MEMS switch



Calient:

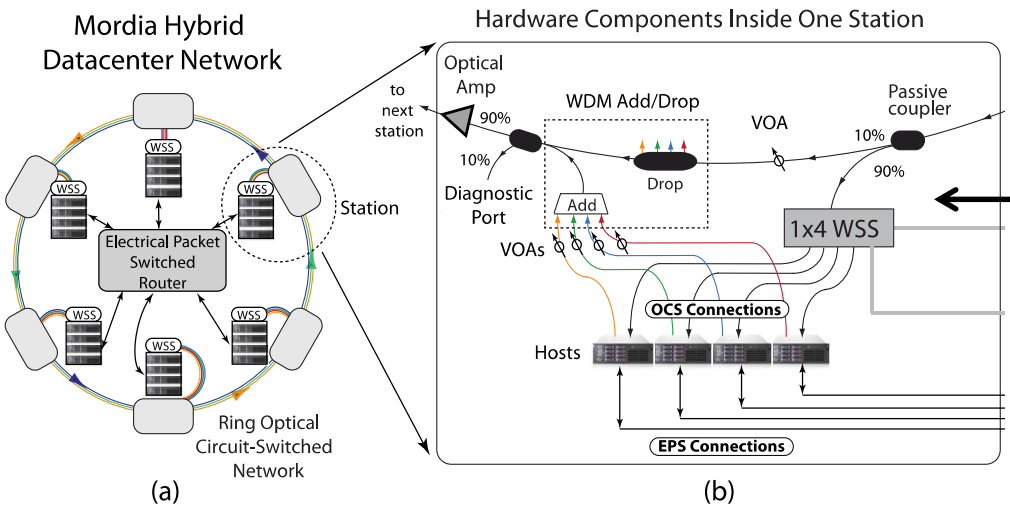
- 300 ports
- 3 dB loss
- 100 ms switching

Central cross-connect topology

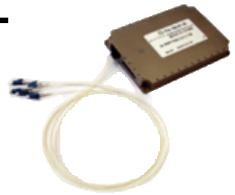
- Scales to hundreds of racks with low loss
- Each OCS port can accept a large number of WDM signals

- Problem: Millisecond switching speed too slow to respond to data center traffic

Faster switching: Mordia



Freespace MEMS
wavelength selective switch



Nistica:

- 4 ports × 24 wavelengths
- 10 μs switching

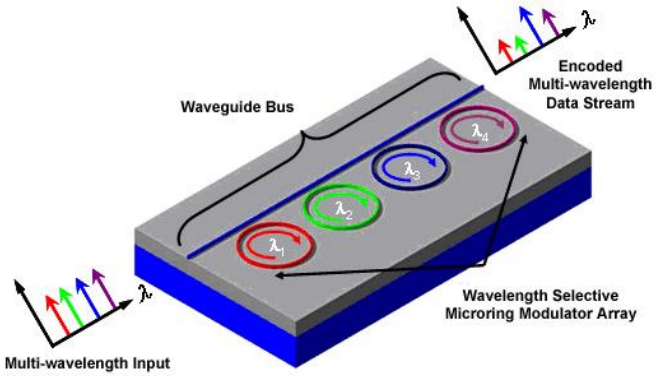
Wavelength-switched ring topology

- Allows for microsecond switching
- More effectively serves data center traffic

- Problem: scalability of rings limited by optical amplifier bandwidth

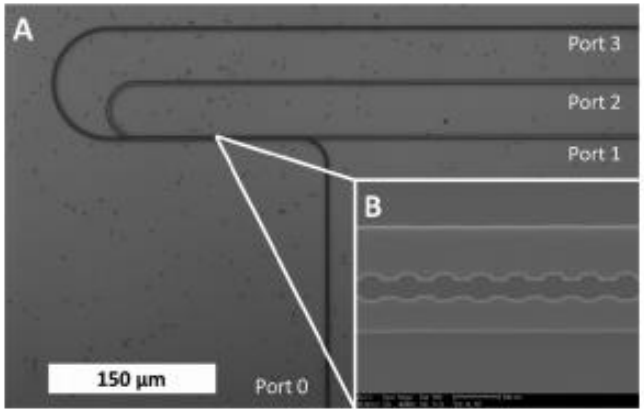
Silicon photonics: high bandwidth links to minimize OCS cost

Wavelength multiplexing:



nanophotonics.eecs.berkeley.edu

Mode multiplexing:



A. Grieco et al., "Integrated space-division multiplexer for application to data center networks," IEEE Selected Topics in Quantum Electronics, 2016.

- Higher bandwidth per fiber reduces cabling complexity & number of OCS ports
- Packet switching expected to reach 400 Gb/s per port
- **Multiple Tb/s per port possible with optical switches**

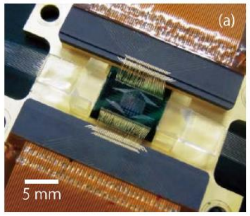
Same challenges we saw before: Temperature, coupling.

+ New challenge: Higher link budget required to transit optical switch(es)

Silicon photonic circuit switches

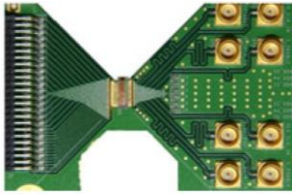
Mach Zehnder (electro / thermo optic)

K. Suzuki et. al., "Ultra-compact 8 x 8 strictly-non-blocking Si-wire PILOSS switch," Optics Express, 2014



Packaged:
8 ports
14 dB loss

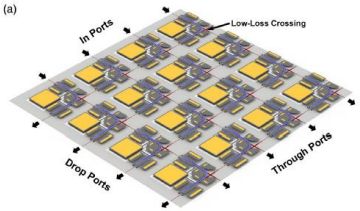
B. Lee et. al., "Monolithic Silicon Integration of Scaled Photonic Switch Fabrics, CMOS Logic, and Device Driver Circuits," JLT2014.



Packaged:
8 ports
20 dB loss

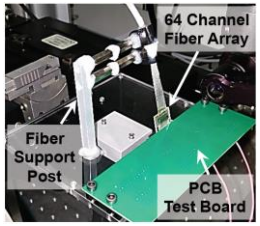
MEMS-actuated waveguide: higher radix

T. Seok et al., "Large-scale broadband digital silicon photonic switches with vertical adiabatic couplers," IEEE Photonics Conference, 2016.



Chip:
64 ports
4 dB loss

T. Seok et al., "12 x 12 packaged digital silicon photonic MEMS switches," IEEE Photonics Conference, 2016.



Packaged:
12 ports
13 - 18 dB loss

The insertion of *any* optical switch degrades the link.

Challenge: minimize switching loss… or larger link budget required.

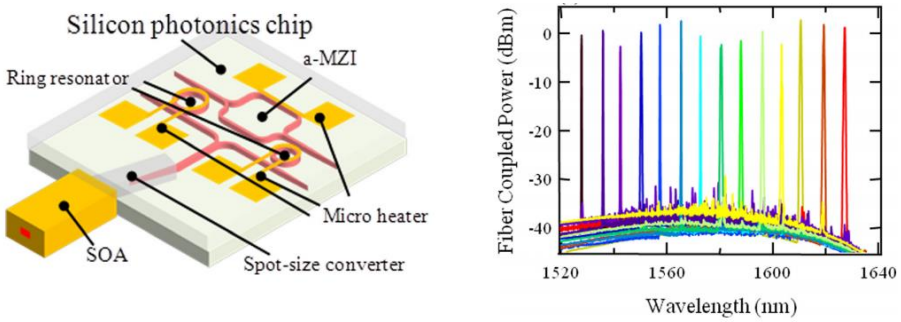
- A small amount of loss can be corrected at higher levels (e.g. FEC).

Challenge: large optical switch radix critical.

- Need to connect 100,000 servers, 2,000 racks, 100 pods
- Switch insertion loss precludes multi-stage optical topologies

Inexpensive wavelength tuning permits new architectures

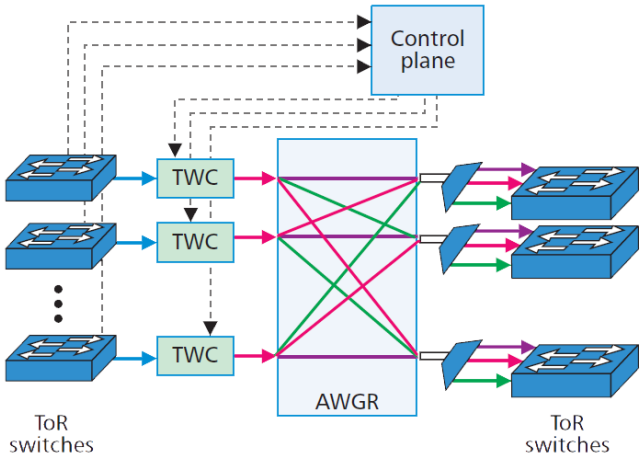
Silicon photonics tunable laser:



T. Kita et al., "Compact silicon photonic wavelength-tunable laser diode with ultra-wide wavelength tuning range," Applied Physics Letters 106, 2015.

Cost needs to be comparable with fixed-wavelength transceivers to be competitive in the data center.

Wavelength routed networks:



C. Kachris et al., "Optical Interconnection Networks in Data Centers: Recent Trends and Future Challenges," IEEE Communications Magazine, 2013.

- Cost of wavelength routed networks critically dependent on cost of tunable elements.
- Potential for fast switching, low loss, and high port count.
- Each network @ 10-25 Gb/s, need multiple networks for high data rate.

Conclusions

Networks likely to scale by **exploiting parallelism** and/or **incorporating optical switching**.

Silicon photonics can help:

- Reduce optical transceiver cost through dense integration
- Surpass electronic pin bandwidth limitations with photonic interposers
- Develop fast, low loss, large port count optical circuit switches
- Reduce cost of device functionality – e.g. wavelength tuning

Potential challenges:

- Arrayed chip/fiber coupling
- Minimizing/mitigating temperature sensitivity
- Scaling optical switch radix while keeping loss low