University of BRISTOL

Lecture 9
Advanced Optical Technologies in DCNs

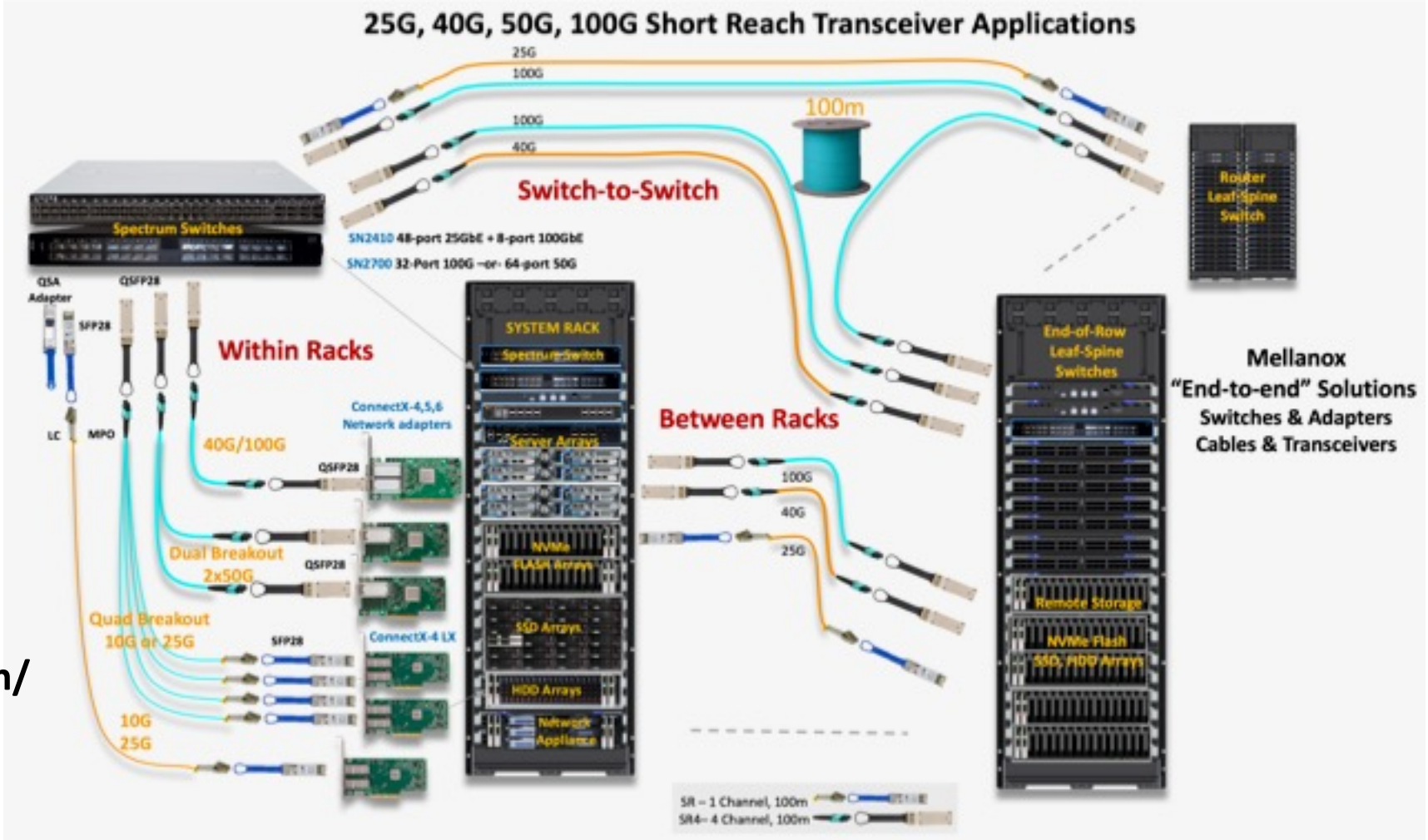
Dr Shuangyi Yan & Dr George T. Kanellos

Optical Technologies in DCNs

Bandwidth
10Gbs
25Gbs
40Gbs
100Gbs

Reach
Multi-Mode
Single-Mode
WDM

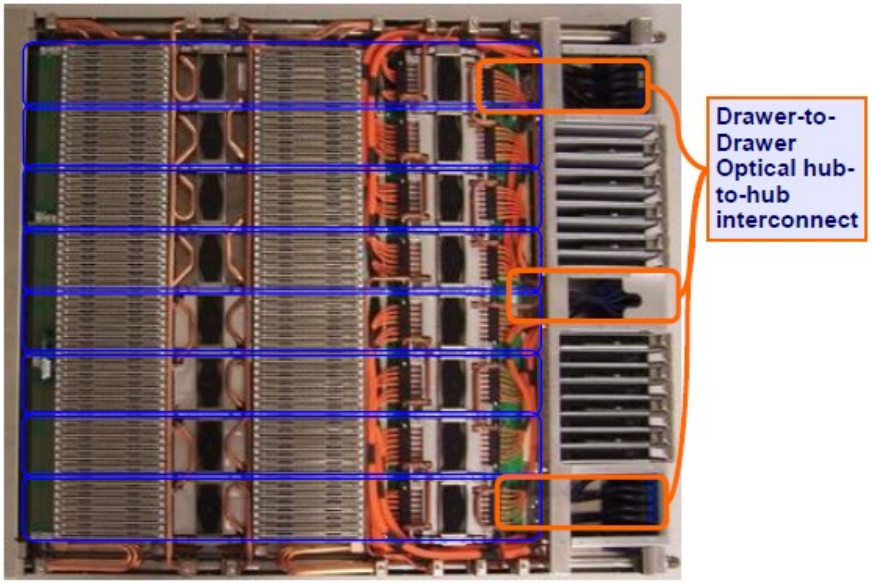
Power consumption/ Size
SFP+
QSFP
CFP



Optical Technologies in DCN

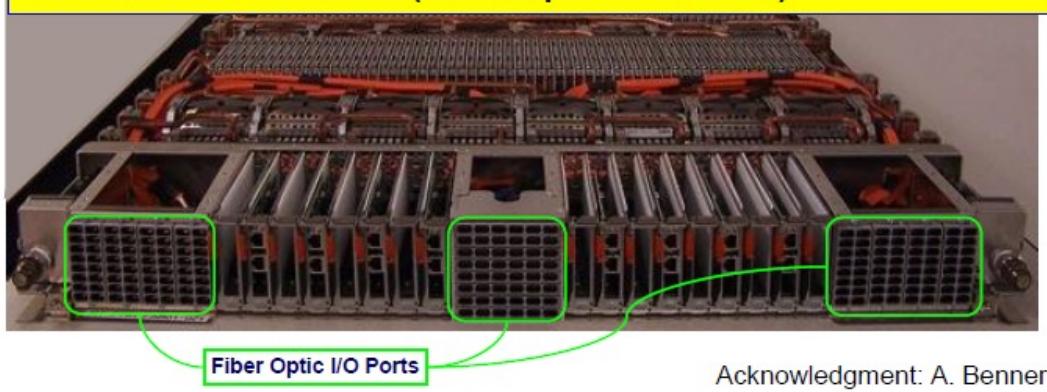
- Transmission medium
- Transmission technologies
- Optical switching technologies (Potential technologies)

Intra-Rack Parallel Optics



256-core Node Drawer

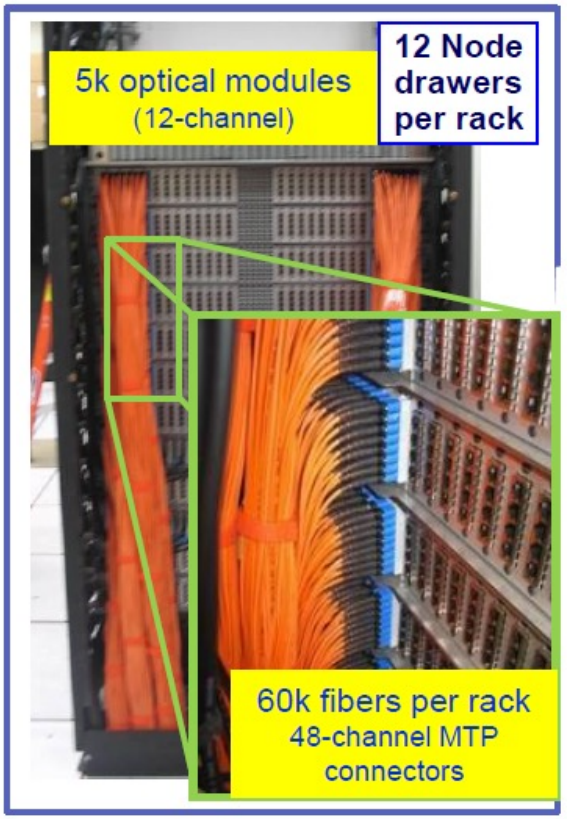
- Optical transceivers tightly integrated, mounted within drawer
- 8 Hub/switch modules (8 x 56 optical modules)



Acknowledgment: A. Benner

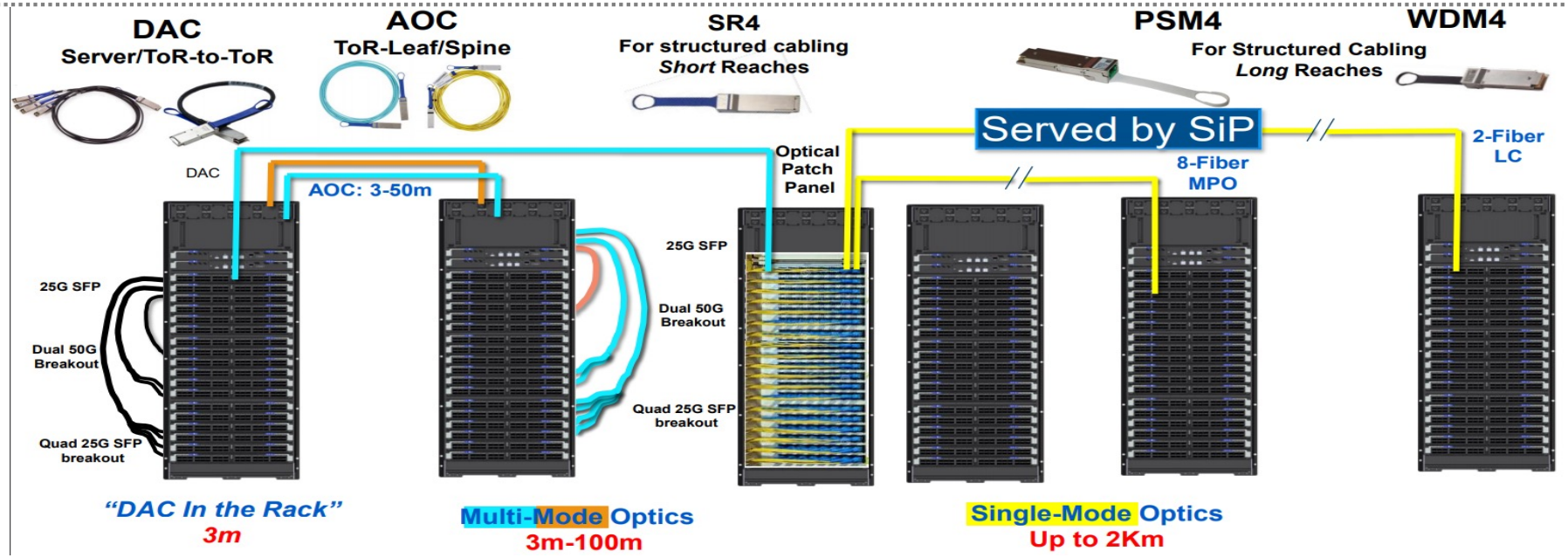
P775 Drawer

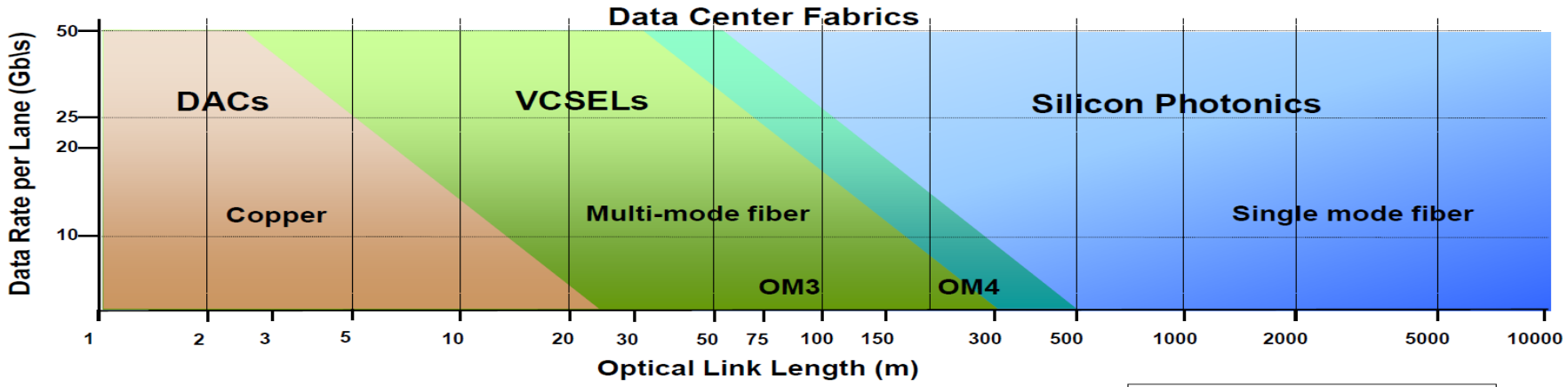
- **8 32-way SMP nodes**
- **Per SMP node:**
  - 1 TF
  - 128 GB DRAM
  - >512 MB/s memory BW
  - **>190 GB/s network BW**



SMP: **Symmetric multiprocessing (SMP)**

Transceivers vs Reach





Direct Attach Copper
- Zero power
- Demo'd 8m at 100G
- Best fit 3m

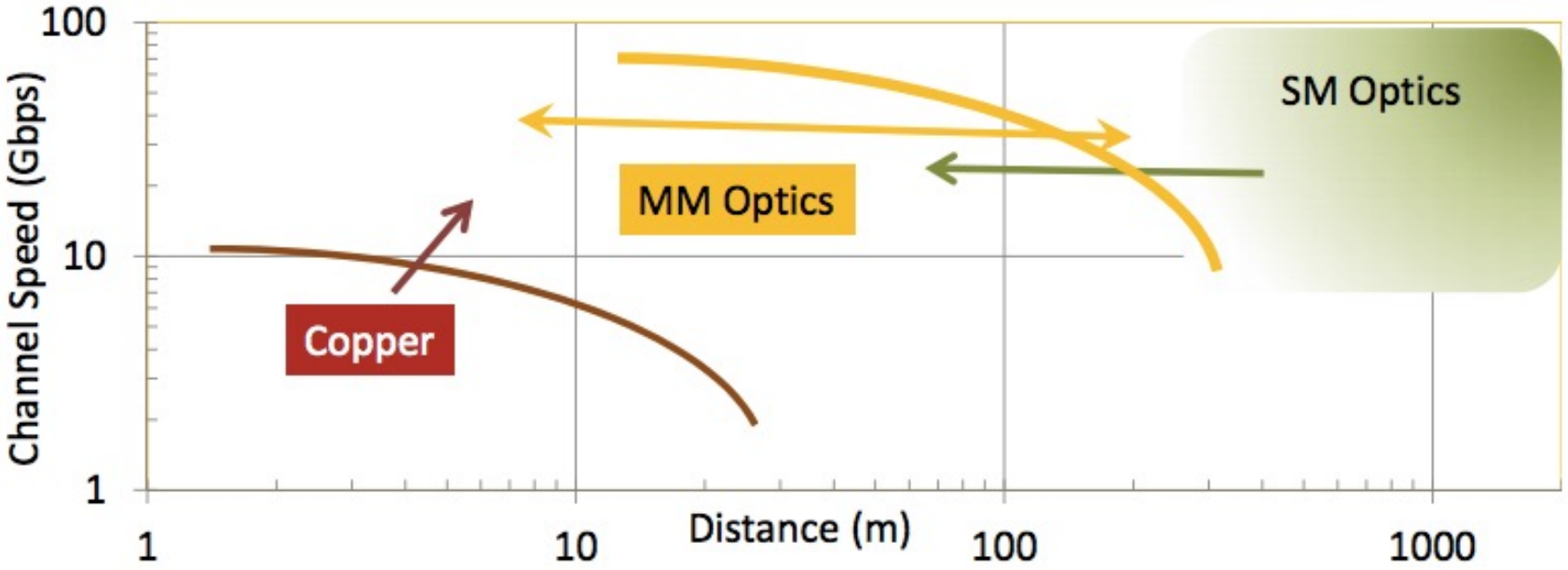
Active Optical Cables
- VCELs or SiP
- Reaches to 200m
- Best fit for 5-20m

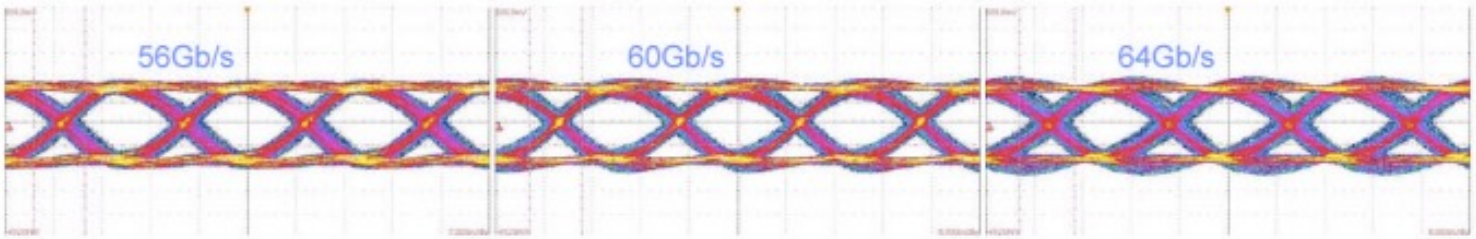
VCSEL Transceivers
- Reaches to 100m
- Best fit for MMF

SiP Transceivers
- Reaches to 2km
- Best fit for SMF
- Parallel or WDM

SiP: System in Package
VCEL: Vertical-cavity surface-emitting laser

Competing transmission medium





OFC 2014 Th3C.2: D. Kuchta et al "64Gb/s Transmission over 57m MMF using an NRZ Modulated 850nm VCSEL" (also >250m at 40G and > 100m at 60G)

MM: Multi-Mode

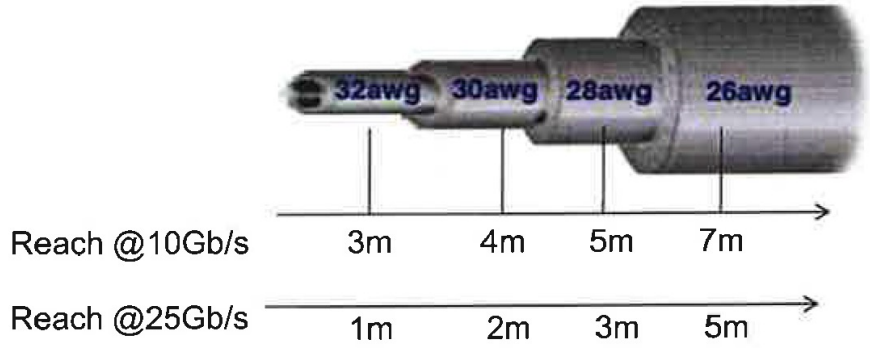
SM: Single-Mode

System packaging: Copper Cables

University of BRISTOL

-

	32AWG	30AWG	28AWG	26AWG
Wire diameter	0.20	0.25	0.32	0.41
8x Cable outer diameter (mm)	4.4	5.5	6.5	7.1
Attenuation @ 5GHz (dB/m)	-4.0	-3.3	-2.4	-2.0
Attenuation @ 12.5GHz (dB/m)	-6.4	-5.2	-3.8	-3.2



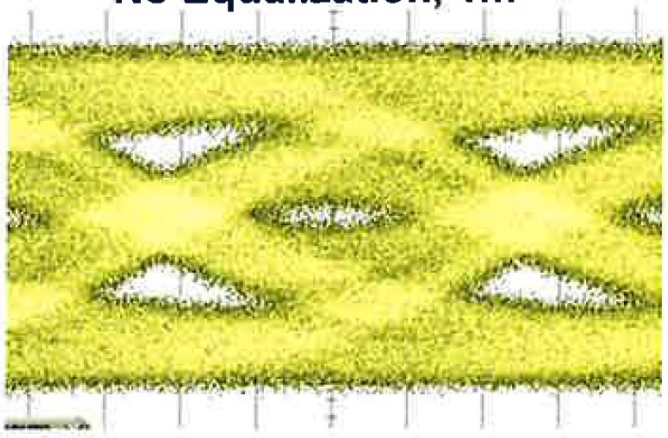
- A single copper cable can support a single channel.
- The attenuation is dependent on the signal bandwidth.
- Copper cables experience 3dB per meter attenuated @ 12.5 GHz.
- They can offer up to 5 meters reach for 25 Gb/s data channels.

System packaging: Copper Cables

- Electrical dispersion compensation continues to drive copper cable to higher data rate and longer reach
- However, it is increasingly difficult and costly in terms of:
  - cost, power and density

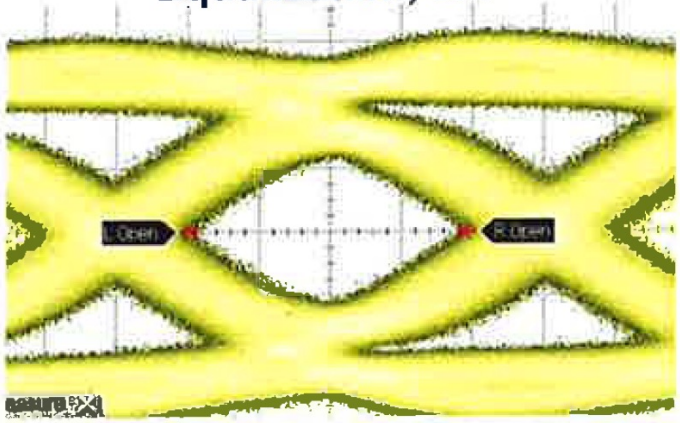
10Gbps Electrical Link in 90nm CMOS over 1m, 5m of 26 Gauge Cable

No Equalization, 1m



Cable + PCB loss: -10dB at 5GHz

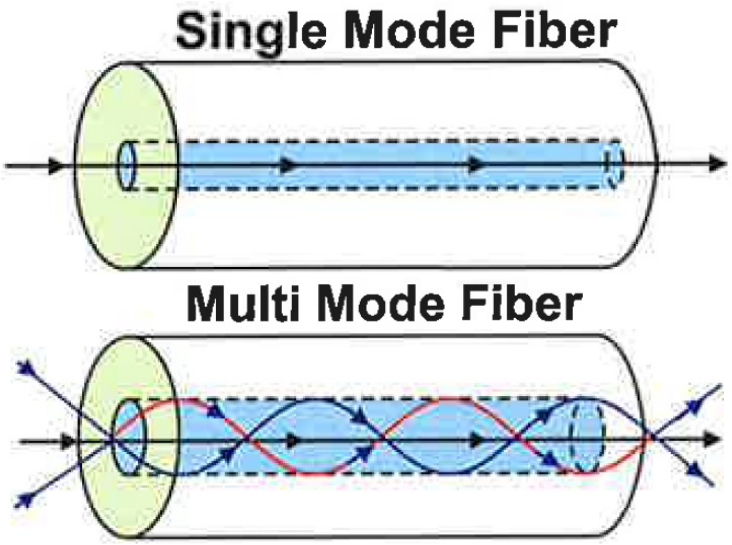
Equalization, 5m

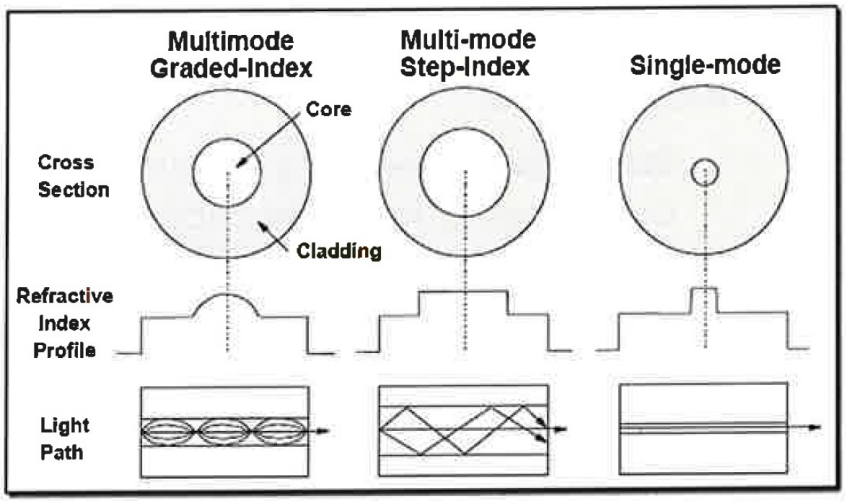


Cable + PCB loss: -20dB at 5GHz
Transmitter has one pre, one post tap

Fibres

- SMF: core diameter: 9 µm (single core)
  - Attenuation: ~ 0.2 dB/Km
- MMF: core diameter: 50 µm, 62.5 µm
- MMF (Multi Mode Fiber) has limited bandwidth
  - Modal dispersion limited: Multiple modes propagate at different velocities causing pulse spreading
  - Graded parabolic index profile reduces modal dispersion
    - Requires careful control of doping concentration layer by layer from the center of the core
- SMF (Single Mode Fiber) has 10s of Tb/s bandwidth by use of WDM (Wavelength division multiplexing)



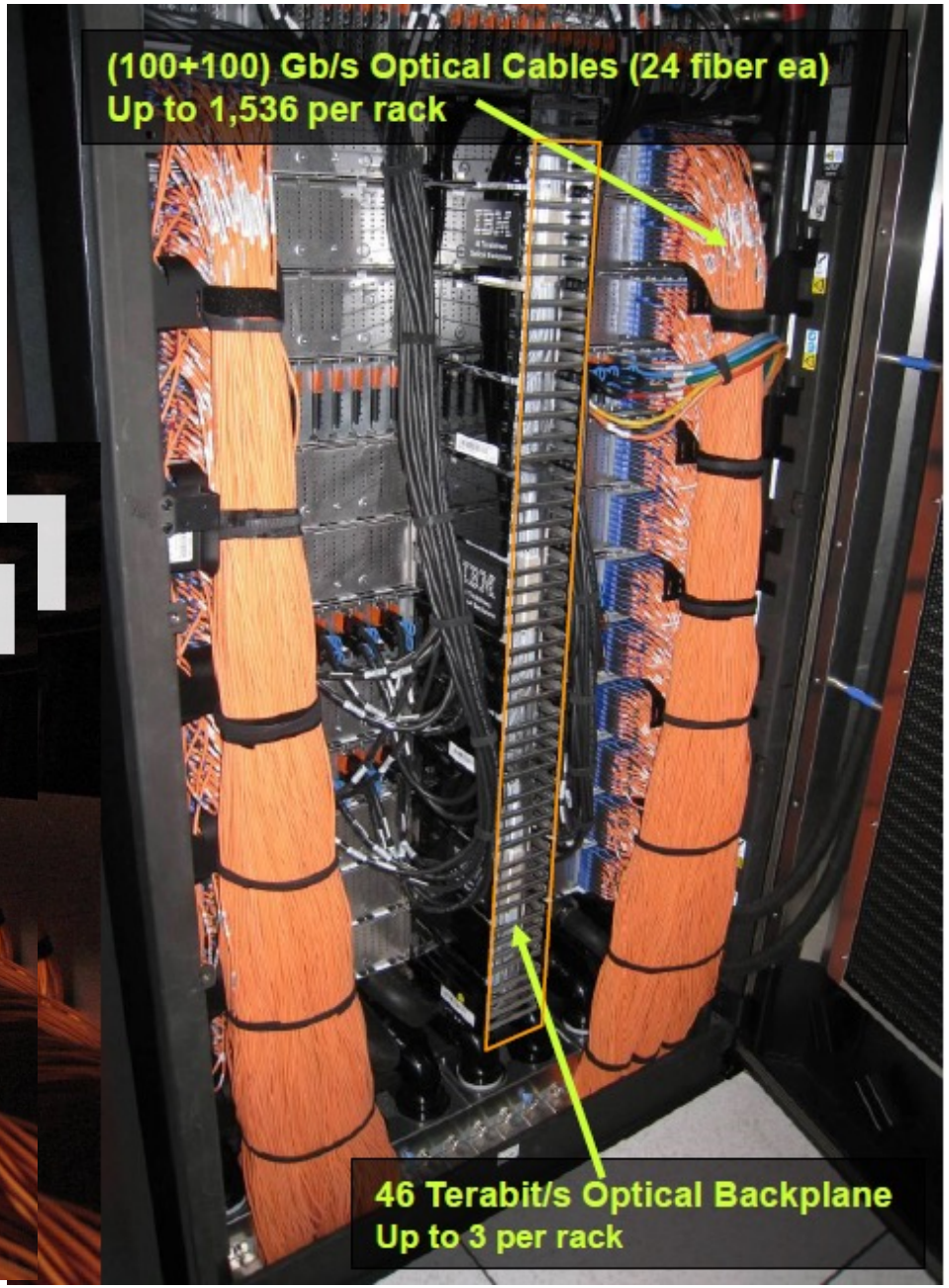


More fibres and bandwidth

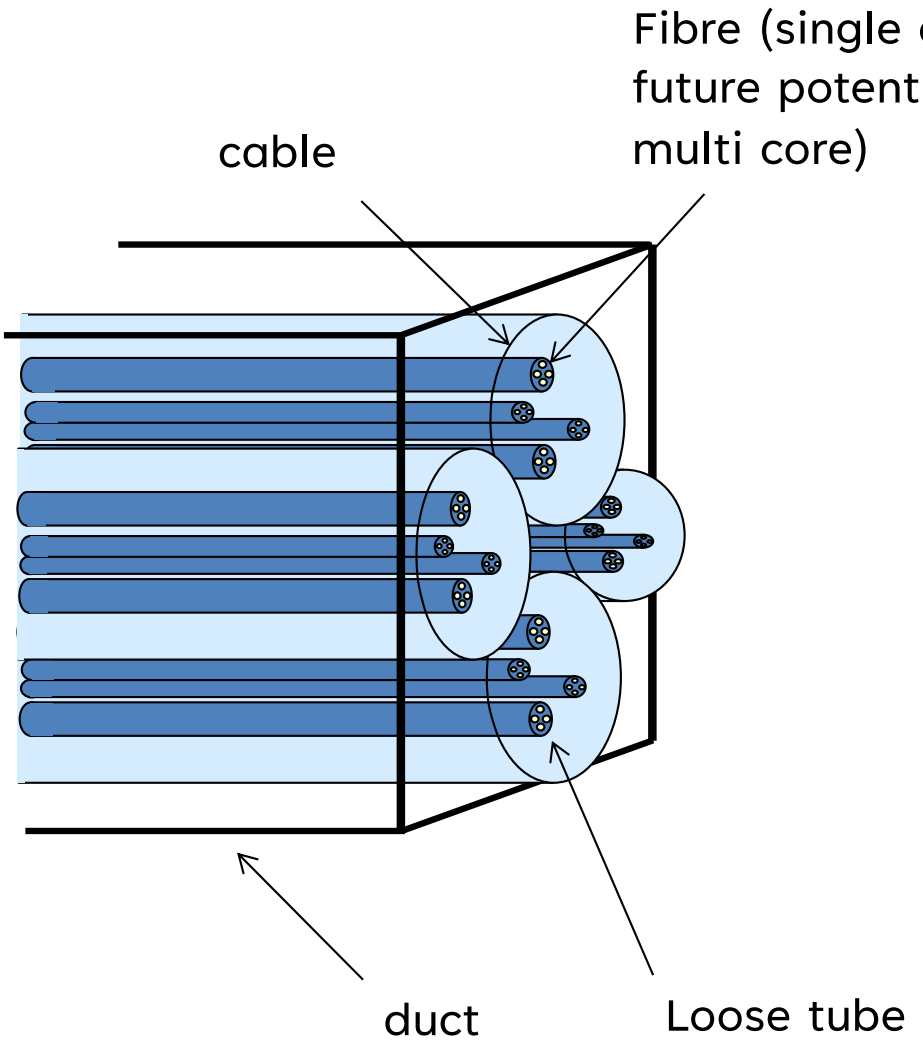
University of BRISTOL

More fibres are needed for connection.

- Wavelength division multiplexing (WDM)
- Space division multiplexing (SDM) based on Multi-core fibres



What is ‘SDM’ (Space Division Multiplexing)?



SDM is any use of multiple cores, irrespective of whether they are in the same fibre, same cable, same duct, to increase the transmission rate of a link in a network

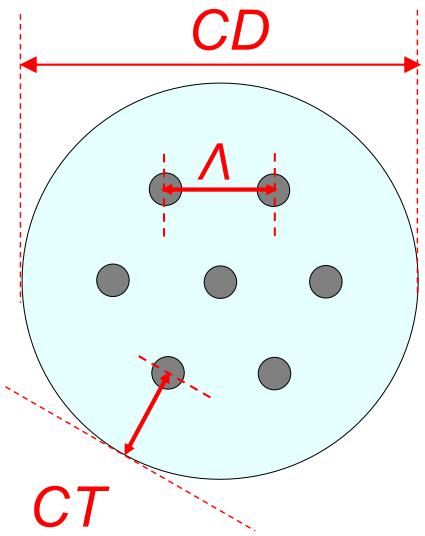
Homogeneous hexagonal Multi-Core Fiber (MCF)

University of BRISTOL

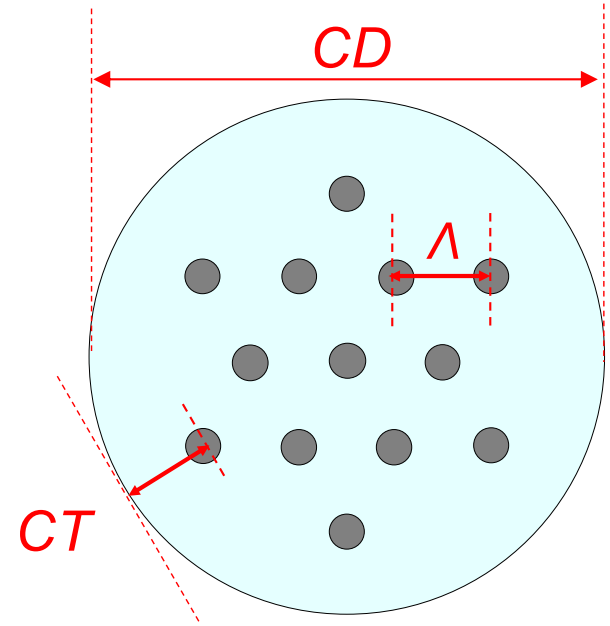
✓Cladding diameter (CD) limit : ~230 μm
✓Outer clad thickness (CT): ~40 μm

1-mm decrement of core pitch results in about 3-dB degradation of XT

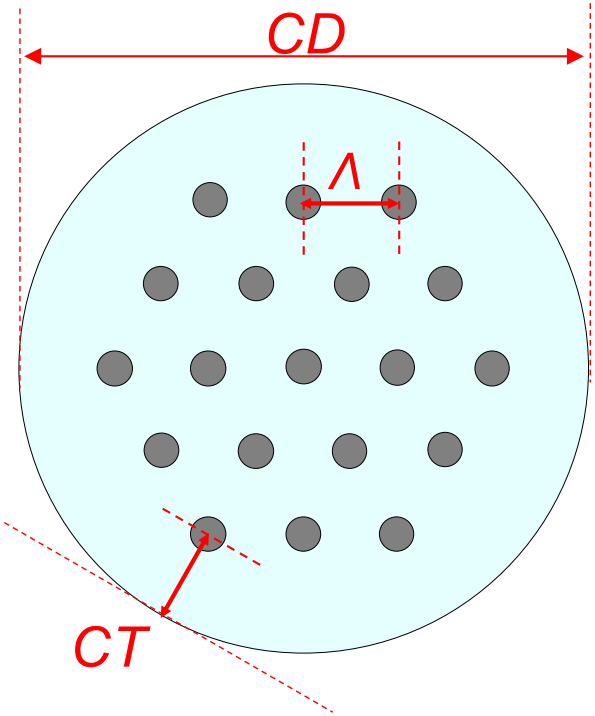
7-core
(Λ < 70 μm)



13-core
(Λ < 43 μm)



19-core
(Λ < 37 μm)



- Cross-section Area
  - $A_{cross} = \pi/4 * (\text{Cladding Diameter})^2$

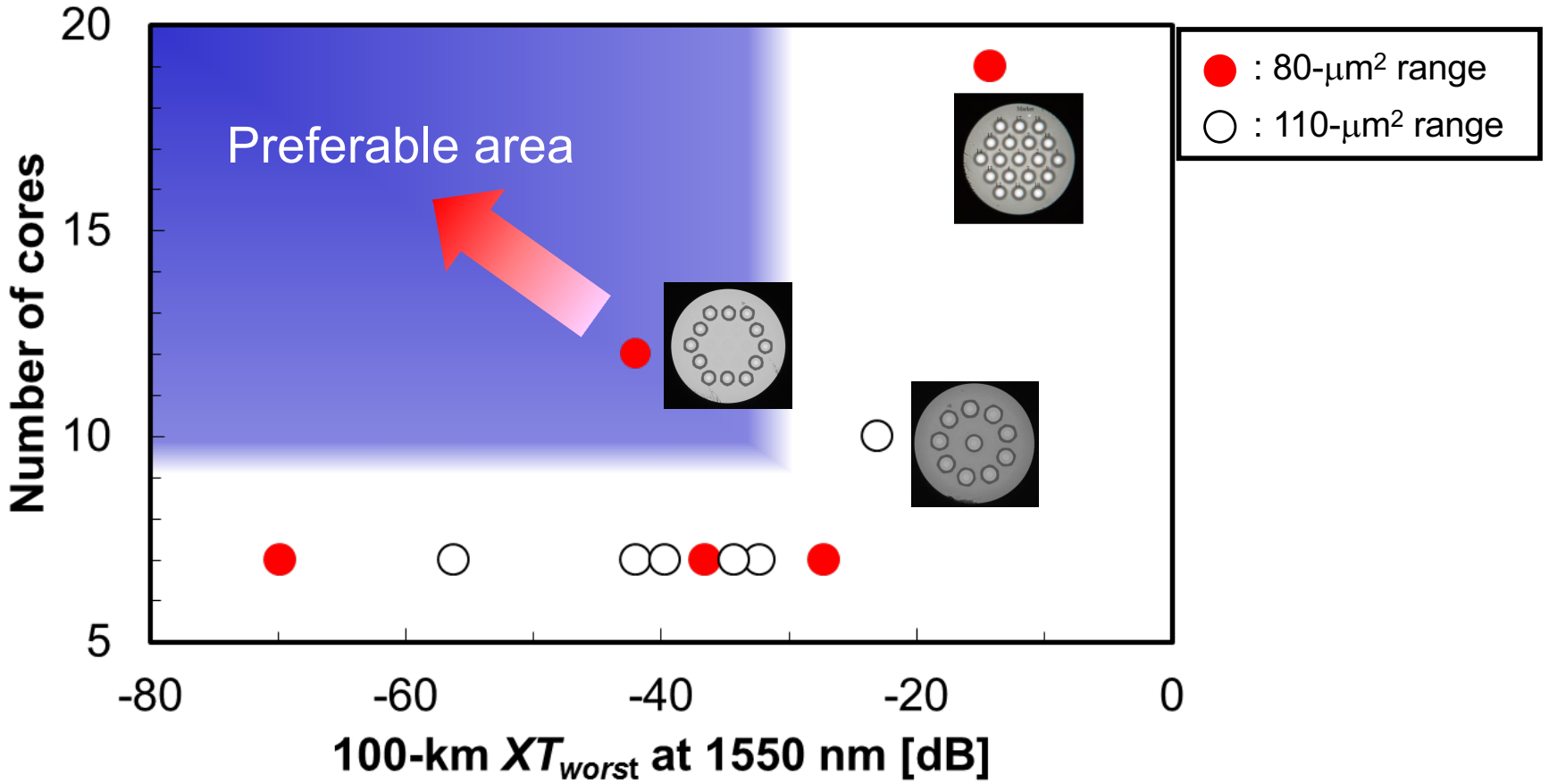
Relation between number of cores and crosstalk level

University of BRISTOL

Core arrangement?
Maximum number of cores?

depend on ...

Required crosstalk level
Maximum cladding size
Target cross-section area (A_{cross})



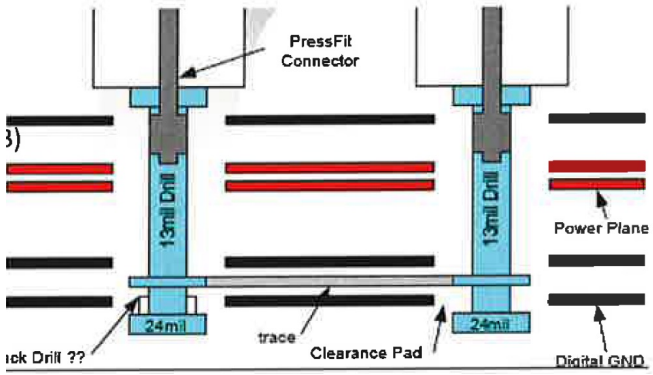
Fabricated single-mode uncoupled MCFs

University of BRISTOL

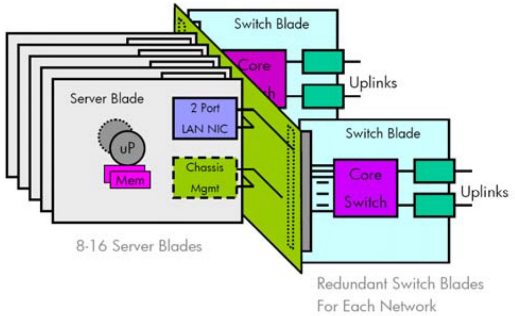
	2008	2009	2010	2011	2012
Fujikura Ltd.	Air-Clad				
Furukawa Elec.				Large A_{eff}	Increased number of core
Sumitomo Elec.					
OFS		All-Solid		Crosstalk suppression with index trench	Large A_{eff} and Low crosstalk
Corning					Inter connection

System packaging: High Speed Backplane

- IEEE P802.3ap – Ethernet Backplane
  - Support up to 1m trace on improved FR-4 PCB + 2 connectors
    - 10GBASE-KR: 10 Gb/s serial
      - 64B/66B encoding
      - Tx pre-emphasis,
      - Adaptive Rx equalization
      - + DFE, FFE , FEC (3dB)
    - 100GBASE-KR4: 25 Gb/s over 4 lanes
      - CAUI 25 Gb/s/lane
      - 64B/66B encoding
      - Tx pre-emphasis,
      - Adaptive Rx equalization
      - + DFE, FFE , FEC (5dB)
- Key considerations
  - Improved design practices for higher speeds
    - Better board material
    - Better SI (Signal Integrity) practices: connector and pin-out, vias, back drill, etc.

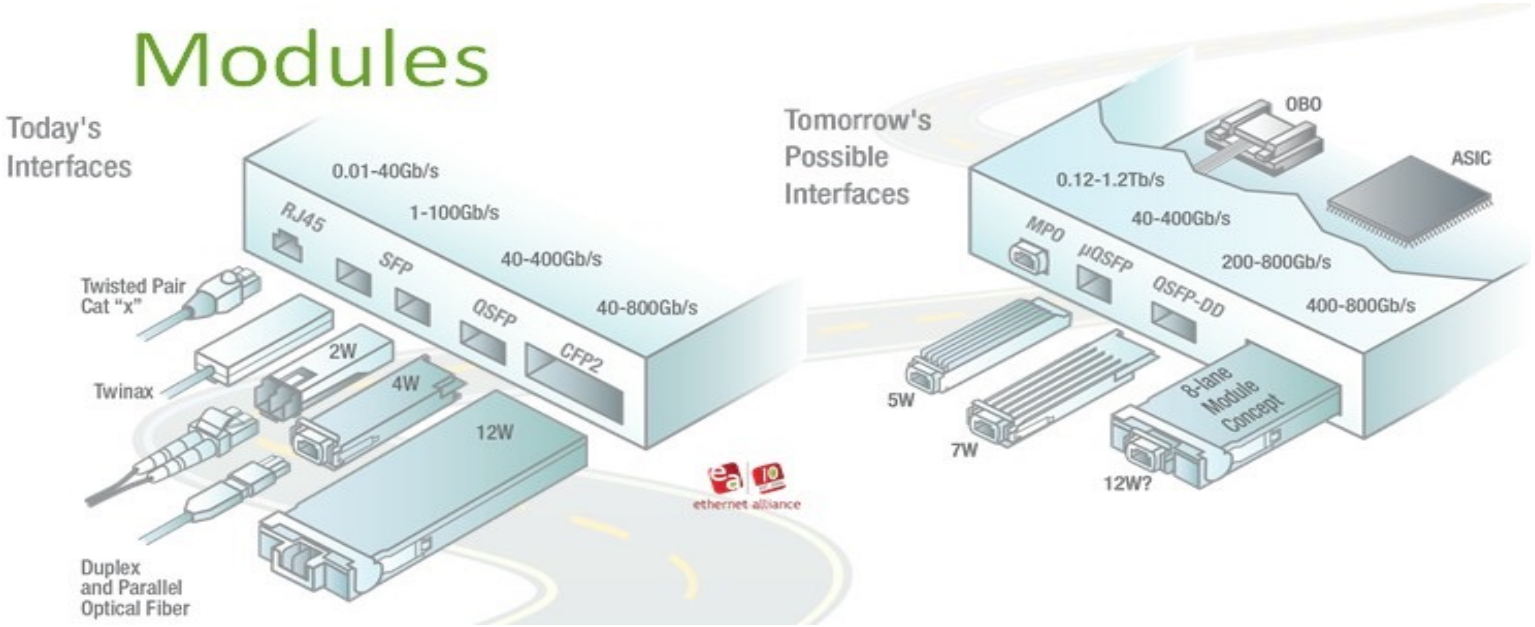


Server Blade & Switch Architecture



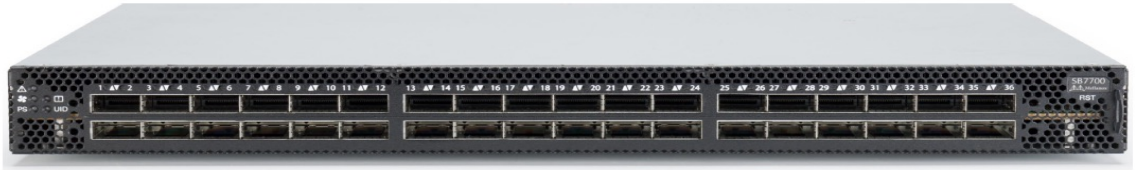
Typical configuration includes a daughter card on the Server Blades

Transceivers size/ form factors

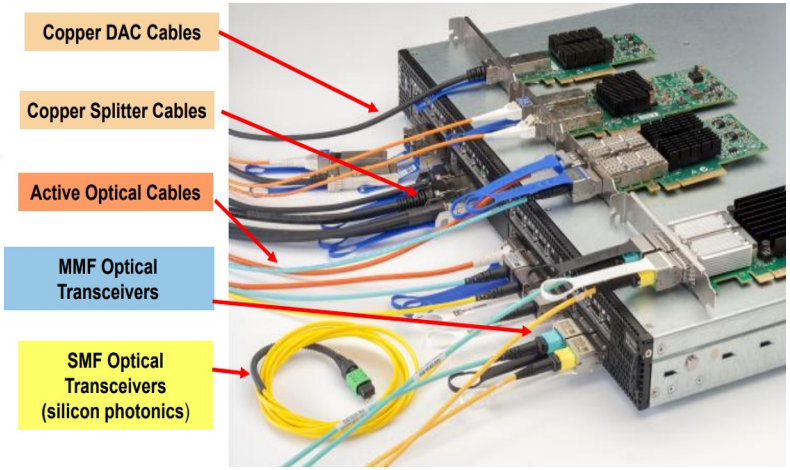


- *SN2700: 32-port Non-blocking 100GbE Open Ethernet Spine Switch System (3.2Tb/s)*

- SB7700 - 36-port EDR 100Gb/s InfiniBand Switch System (3.6Tb/s)

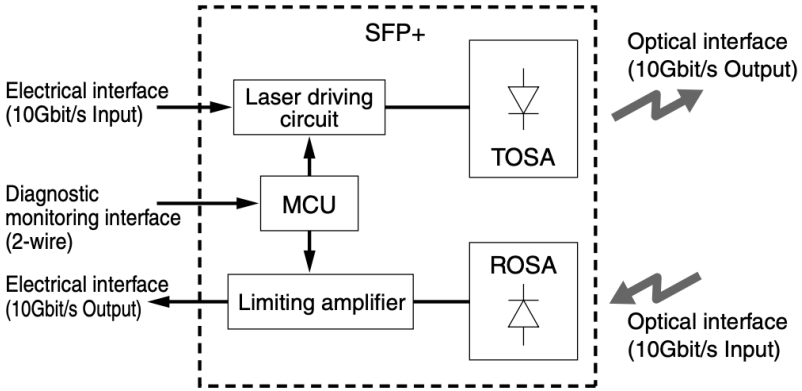


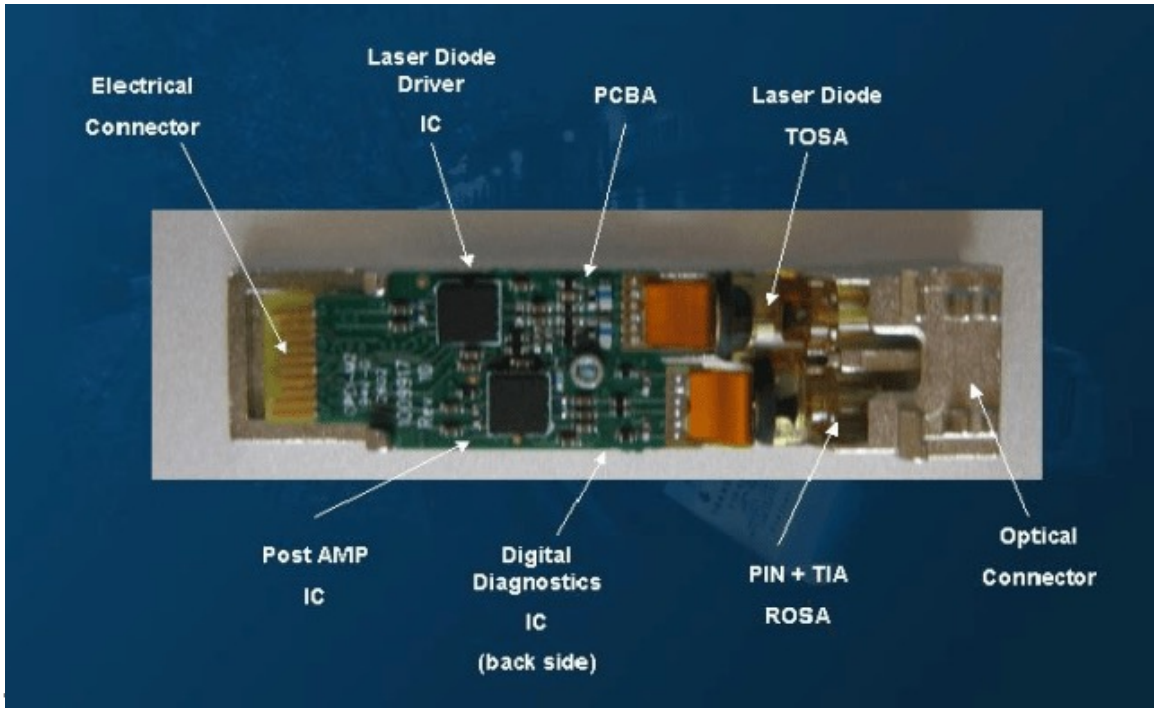
Note: QSFP was a great improvement over SFP; it doubled the density, but we need more



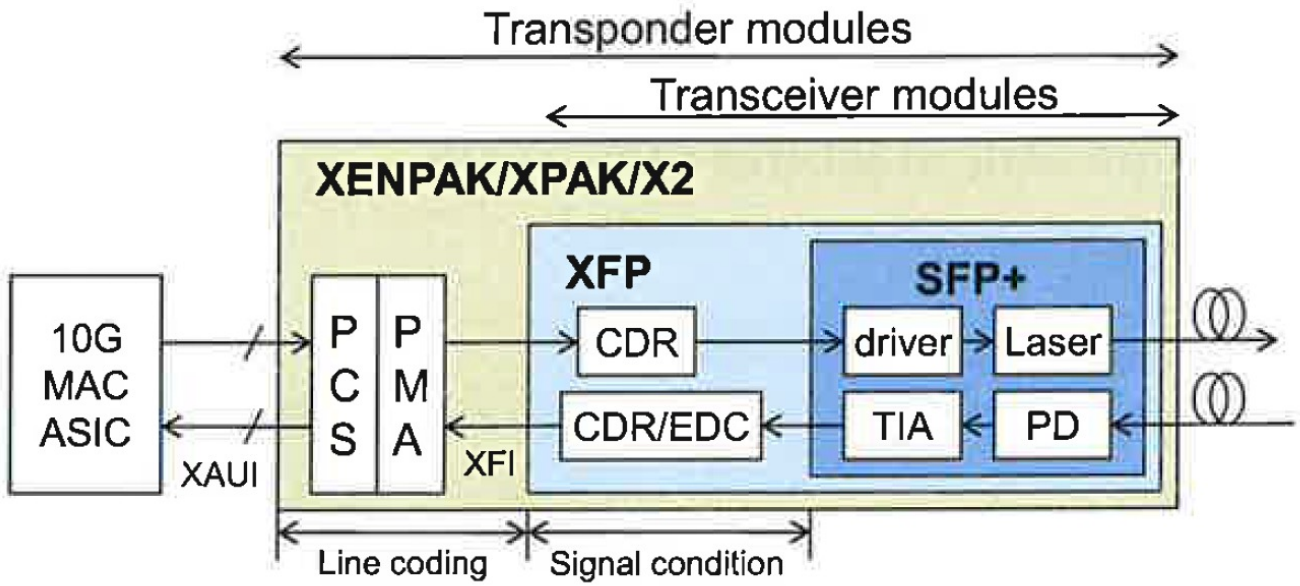
SFP+: Small Form-factor Pluggable Transceiver

SFP+ is an enhanced version of the SFP that supports data rates up to 10 Gbit/s. SFP+ supports 8 Gbit/s Fibre Channel, 10-gigabit Ethernet and Optical Transport Network standard OTU2.



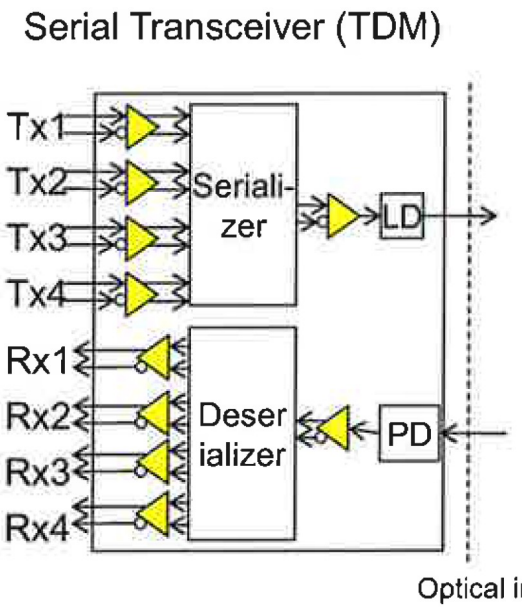


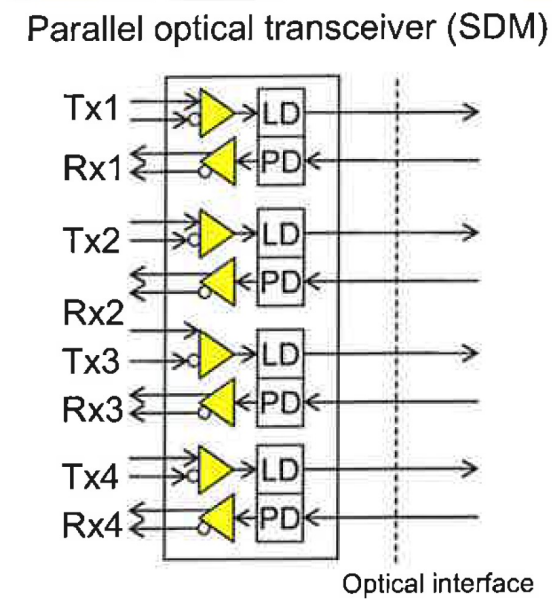
10G Transponders and Transceivers



- Transponders (XENPACK, XPAK, X2) with PHY+ PCS (Physical Coding Sublayer) gave way to transceiver-based solutions
  - Footprint density, power density and price
- Coding and signal conditioning functions can be integrated with ASICs to improve density and power footprint
- Unified transceiver design can be shared with multiple applications: Ethernet, FiberChannel and Infiniband

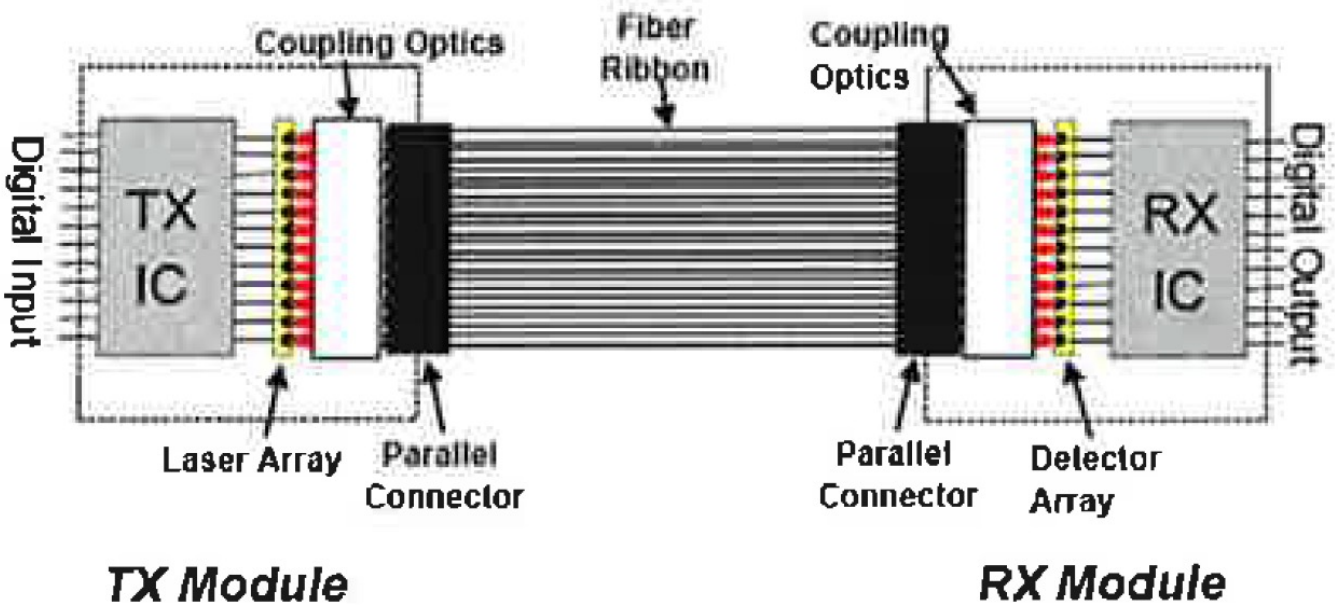
More Bandwidth: multiple channels





- Serial Transceiver: Beyond 20G, electronic and optical performance becomes challenging
  - Gear box consumes power and £
  - Speed mismatch between electrical an optical lanes
- Parallel transceiver takes advantage of parallel data lanes of microprocessor systems
  - Less power consumption due to the absence of SerDes or “gear box”
  - Limited reach and does not solve cabling problem

More Bandwidth: Parallel Optical Tranceiver



VCSEL array

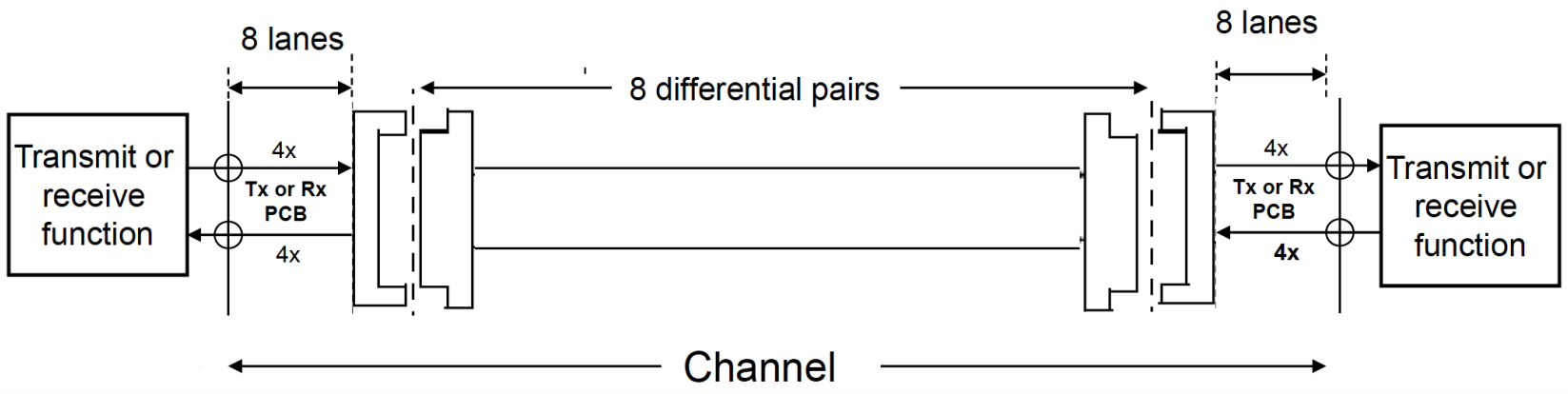
MTP ferrule

Ribbon fiber

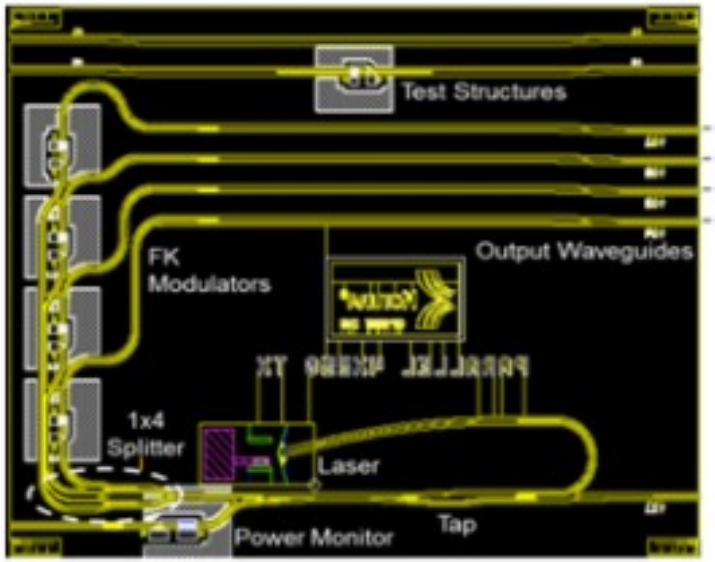
- Challenges for VCSELs as high-speed transmitter
  - Low output power due to short gain region => limited reach of 10
    tight power budget for patch panel and MTP
  - Trade-off between power and speed

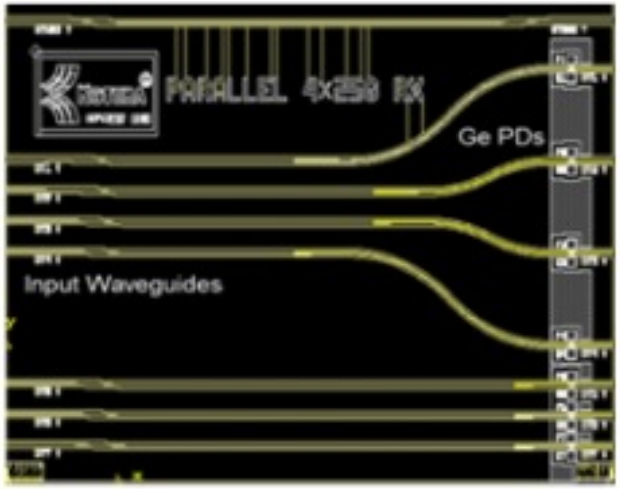
Cabling: 4x25 Gb/s vs 10x10 Gbps

- 8 lanes/pairs versus 20 lanes/pairs
  - Smaller form factor enables higher port densities.
- Cost factors
  - Fewer cable assembly pairs reduces cost.
    - 20-pair assembly ~1.5 times cost of 8-pair assembly.
  - Reduction in number of Tx/Rx lanes simplifies host routing.
  - Reduction in cable assembly pairs enables reduction in cable diameter.
  - Fewer pairs => higher levels of integration => higher port density
- Could define plug compatible port types to enable plug-and-play for 4x10Gb/s and 4x25Gb/s.



4x25G Parallel optical link



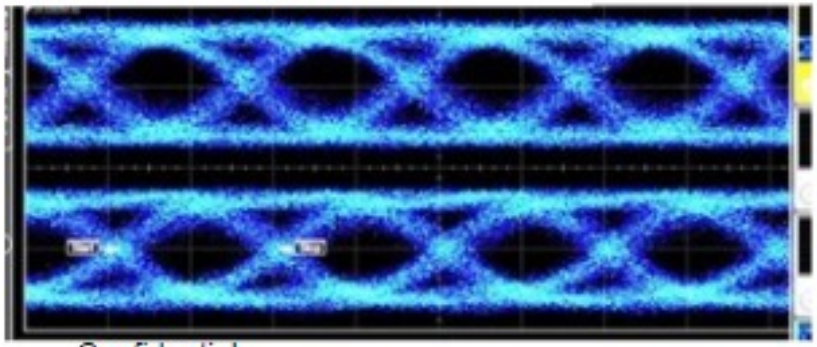


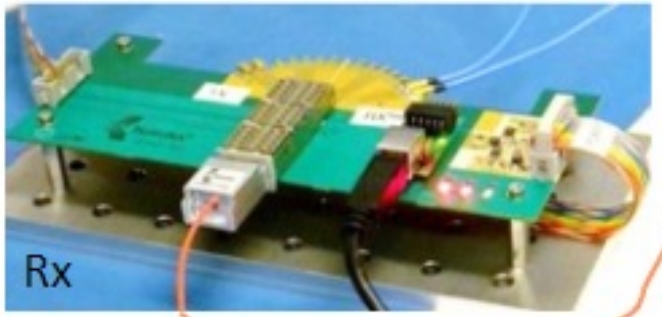
- Current: ~1.5W
- Next gen: 0.8W
- Assuming no CDR
- SMF low jitter

- 4x25G Parallel in QSFP

- Reach independent SR transceiver or AOC solution
- Reliable Edge emitters

Differential Rx Eye at 25Gb



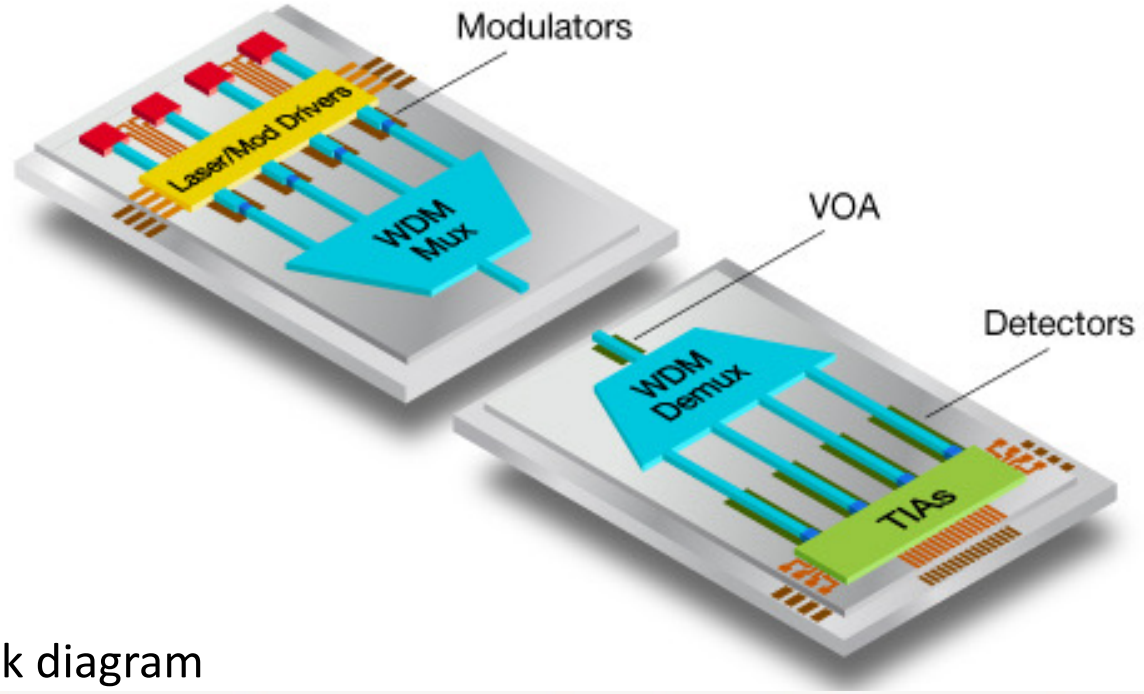


100m

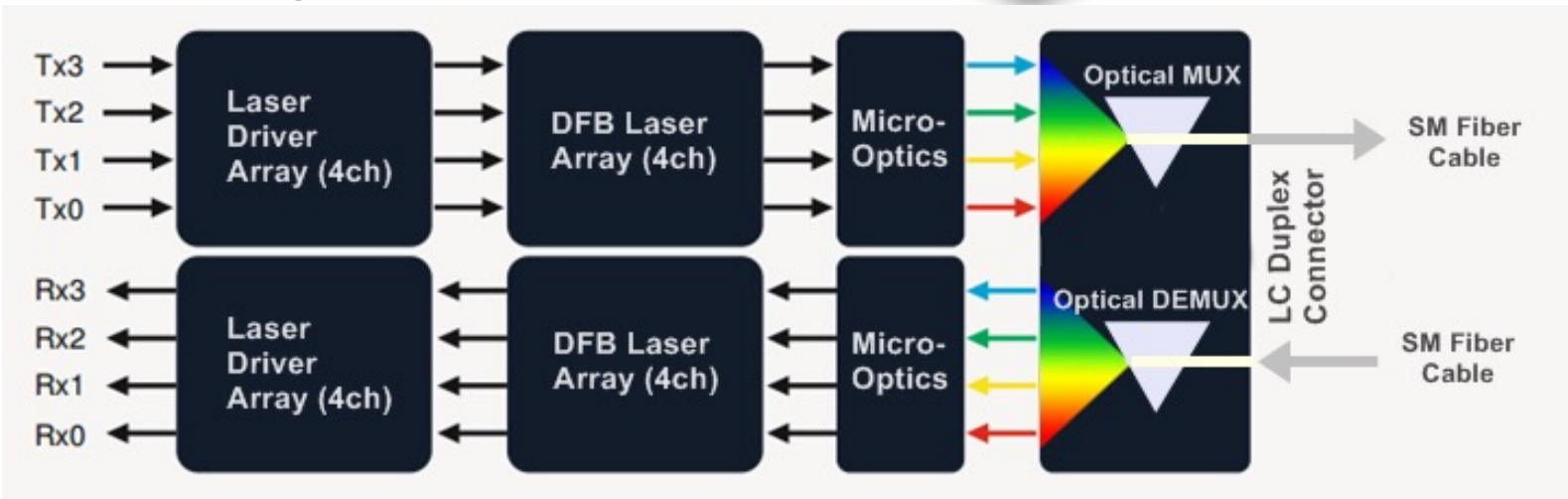
KOTURA

Confidential

WDM transceivers: 4 CWDM channels



QSFP block diagram



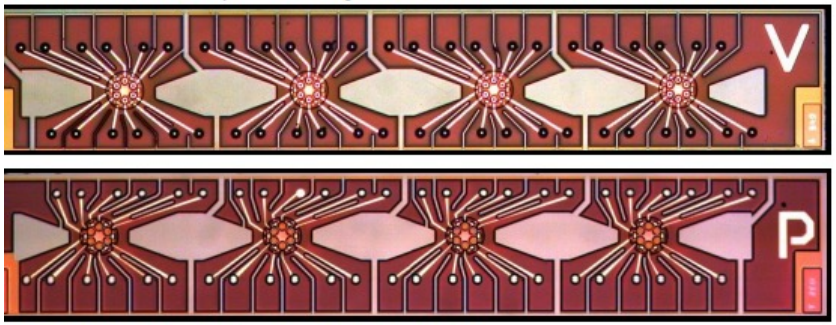
4-Fbier 24-Core Optical Transceiver

University of BRISTOL

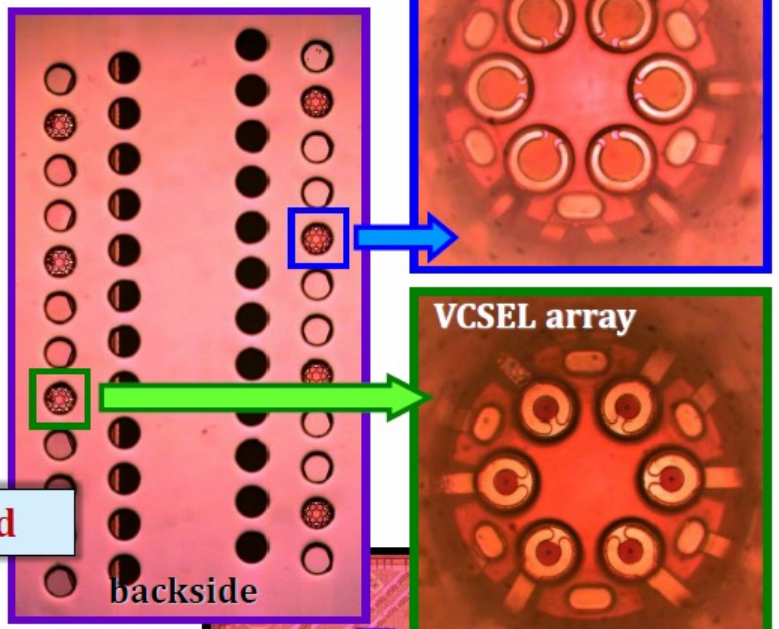
4-Fiber 24-Core Optical Transceiver

IBM

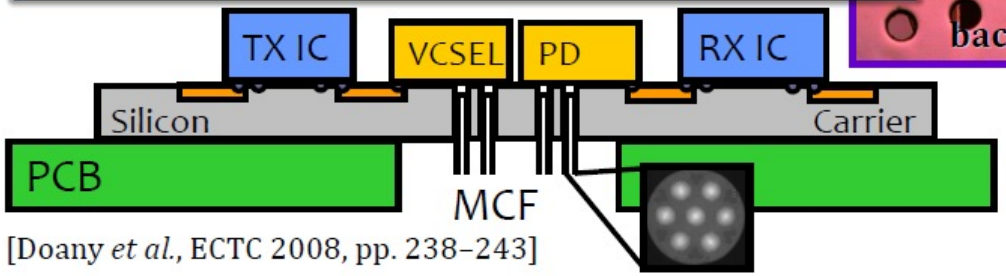
Custom VCSEL/PD Arrays Matched to 4 Multicore Fibers



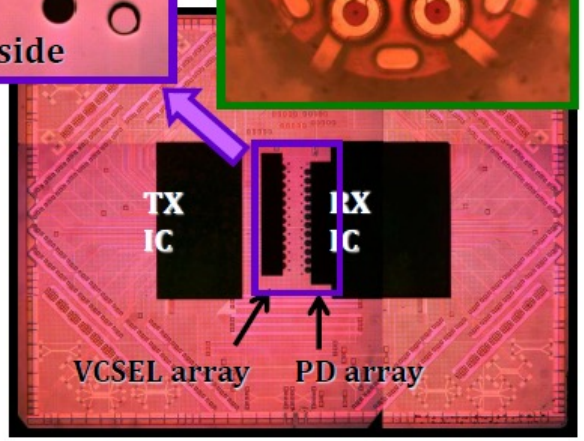
Fabricated by Emcore Corporation



120 Gb/s over 100-m using one MMF strand

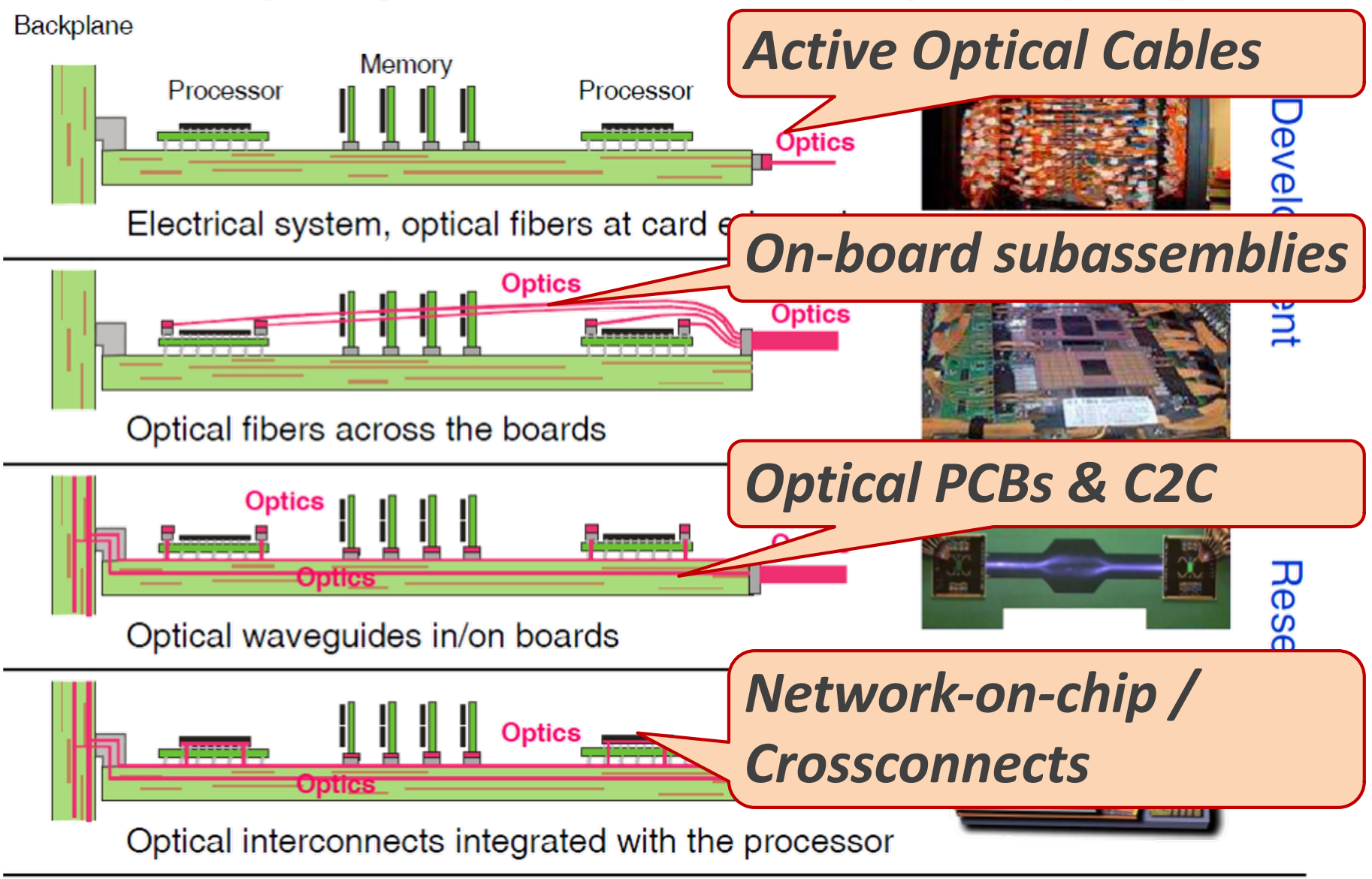


[Doany *et al.*, ECTC 2008, pp. 238–243]



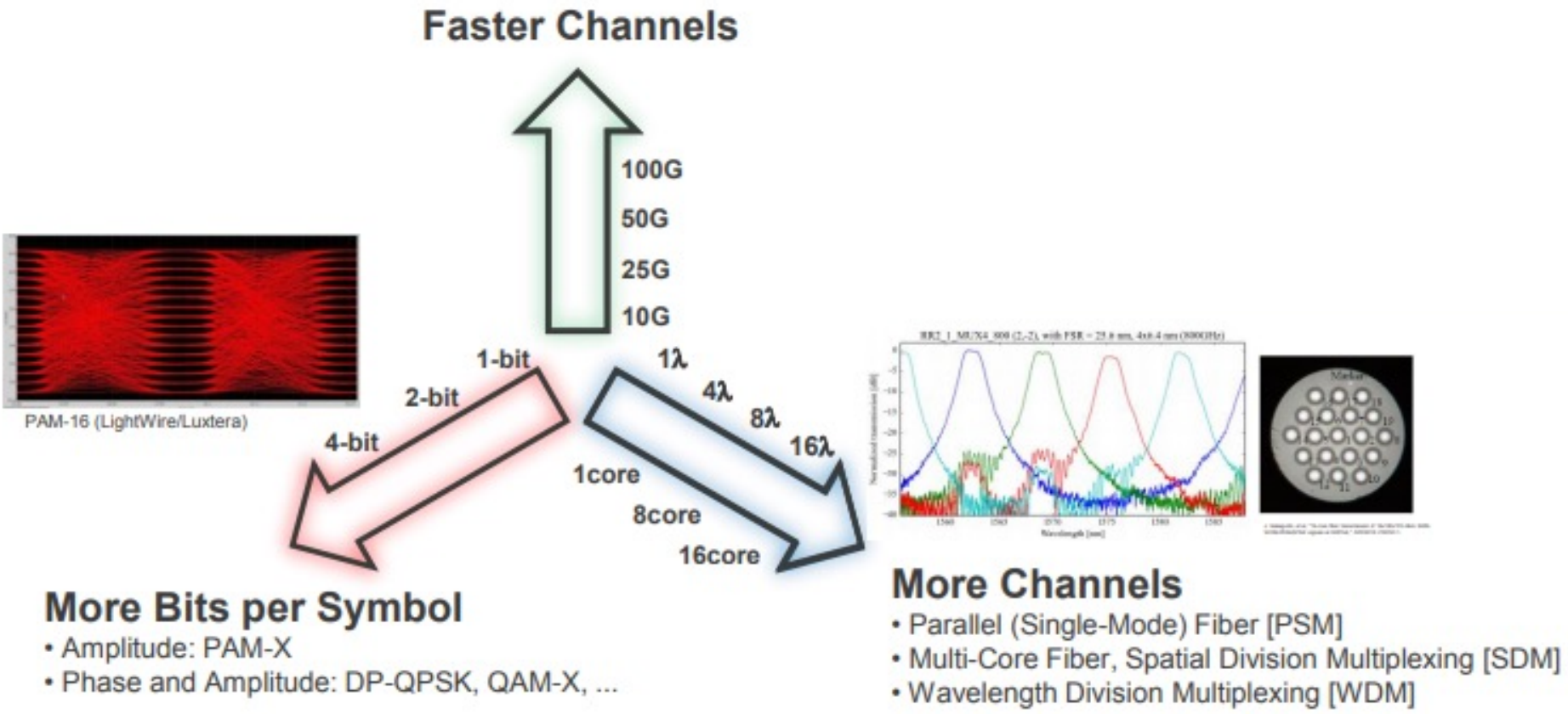
- Custom OE chips designed to fit into existing configuration of Terabus project—Match silicon carrier designed for 24-channel polymer optical waveguide transmitter.

Optical interconnects hierarchy



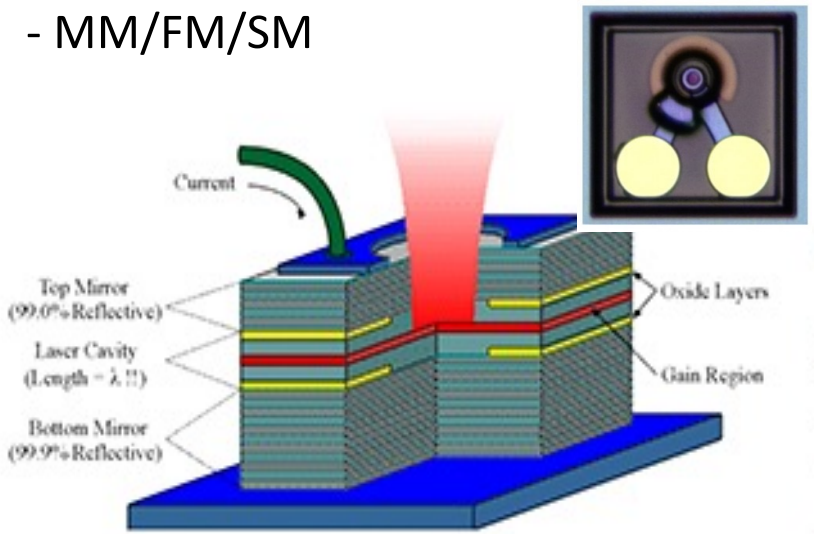
* Source: IBM, B. Jan Offrein, "Silicon Photonics Packaging Requirements", Munich 2011

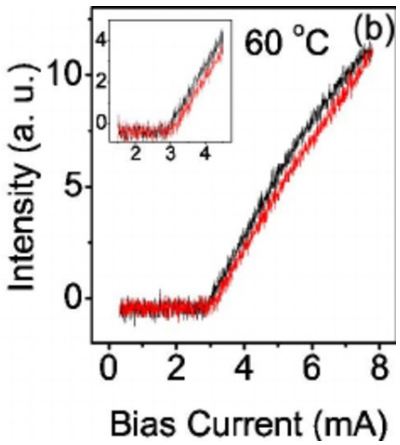
Optical transceivers scaling

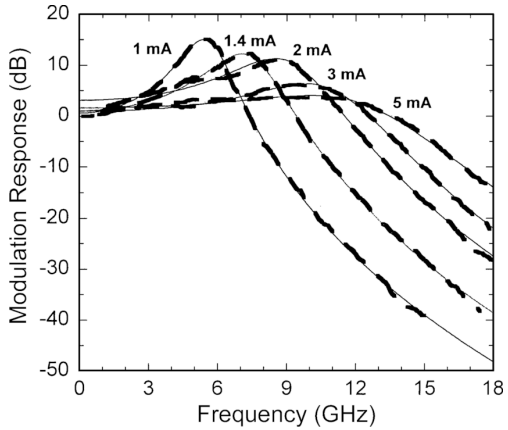


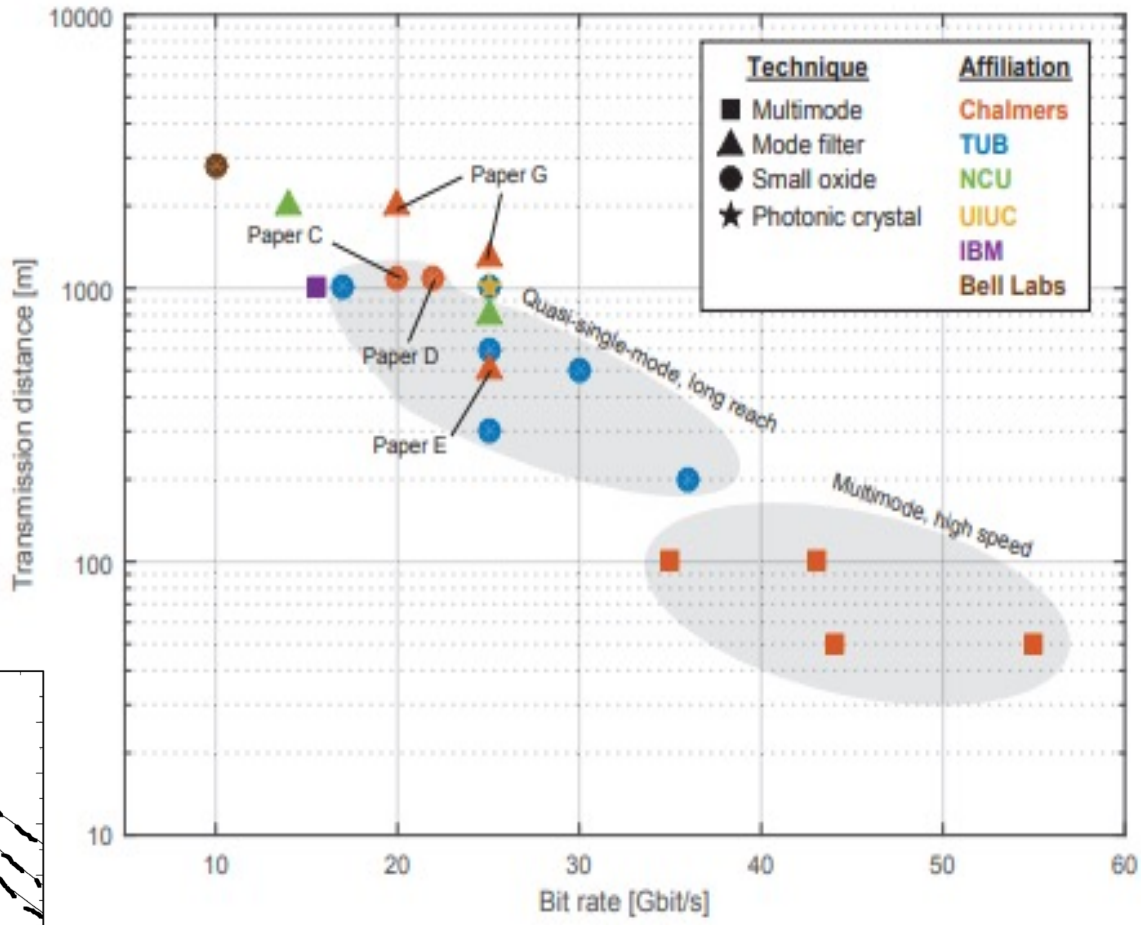
VCSEL lasers

Directly modulated lasers
- MM/FM/SM

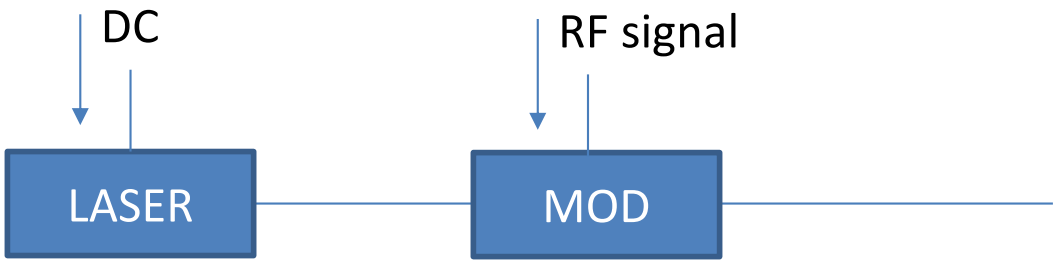






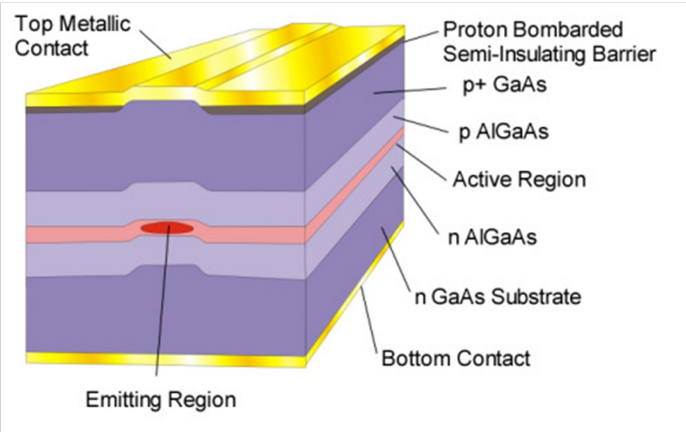


Externally modulated lasers (EML)



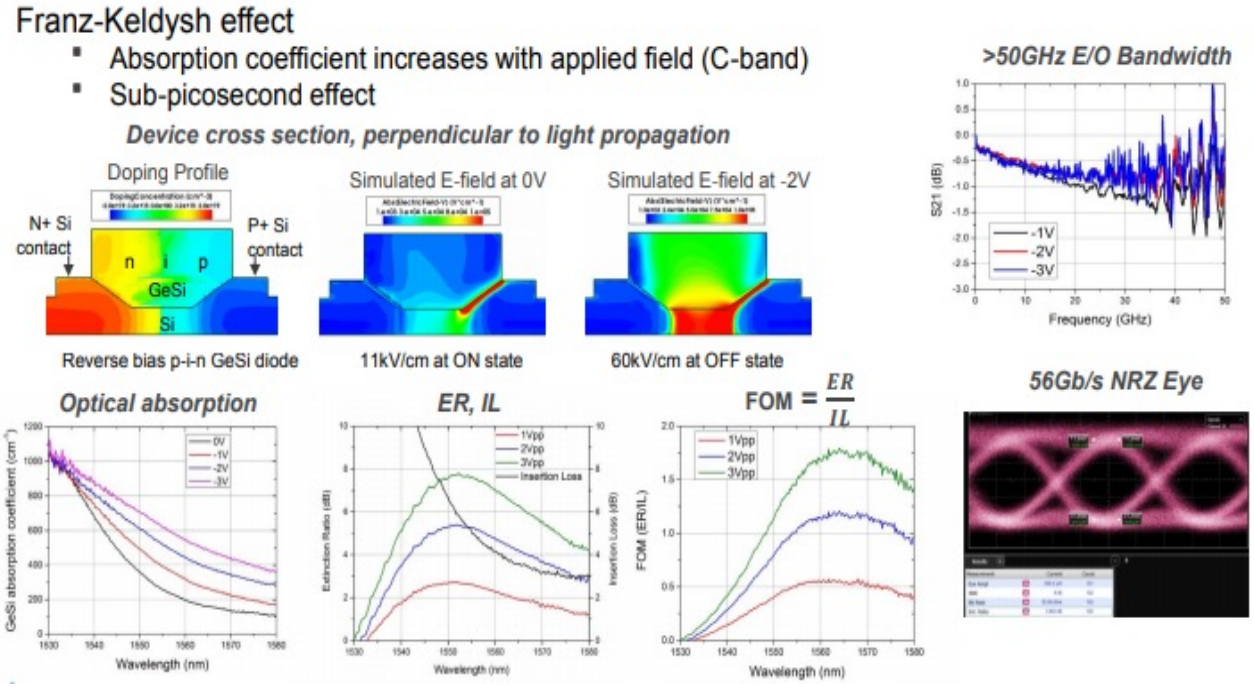
Usually a DFB

Usually an EAM

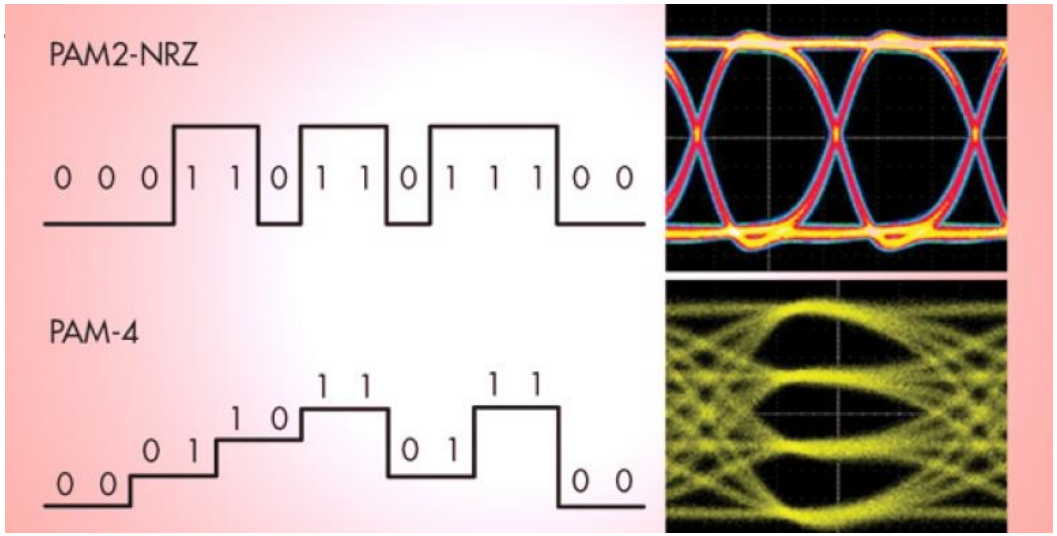


Single-mode

High optical power output

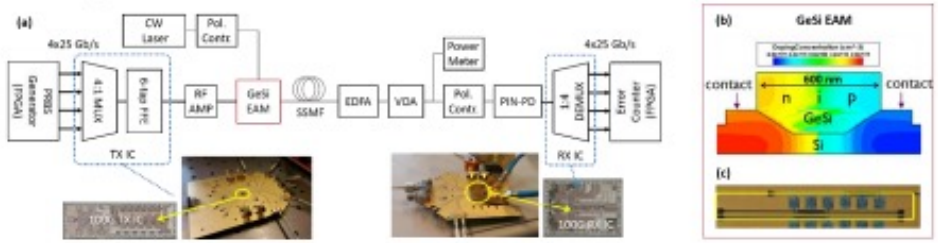


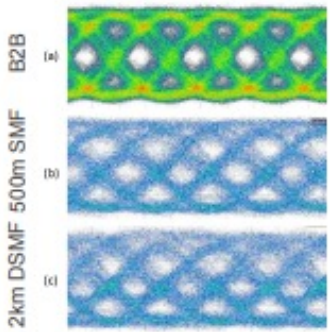
PAM-4 format



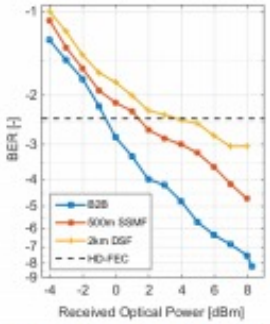
GeSi EAM-PD at 100Gb/s

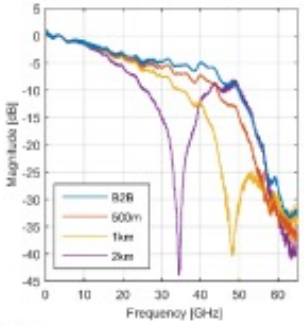
*First single-wavelength **100Gb/s NRZ-OOK** modulation demo in Silicon Photonics*



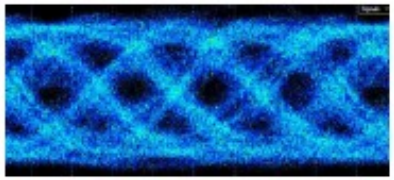


100Gb/s Eye Diagrams (III-V discrete PD)





BER below threshold for FEC



*100Gb/s Eye Diagram for **GeSi EAM-to-GeSi PD***

Collaboration with IDLab

400G Data Center Optics Standards

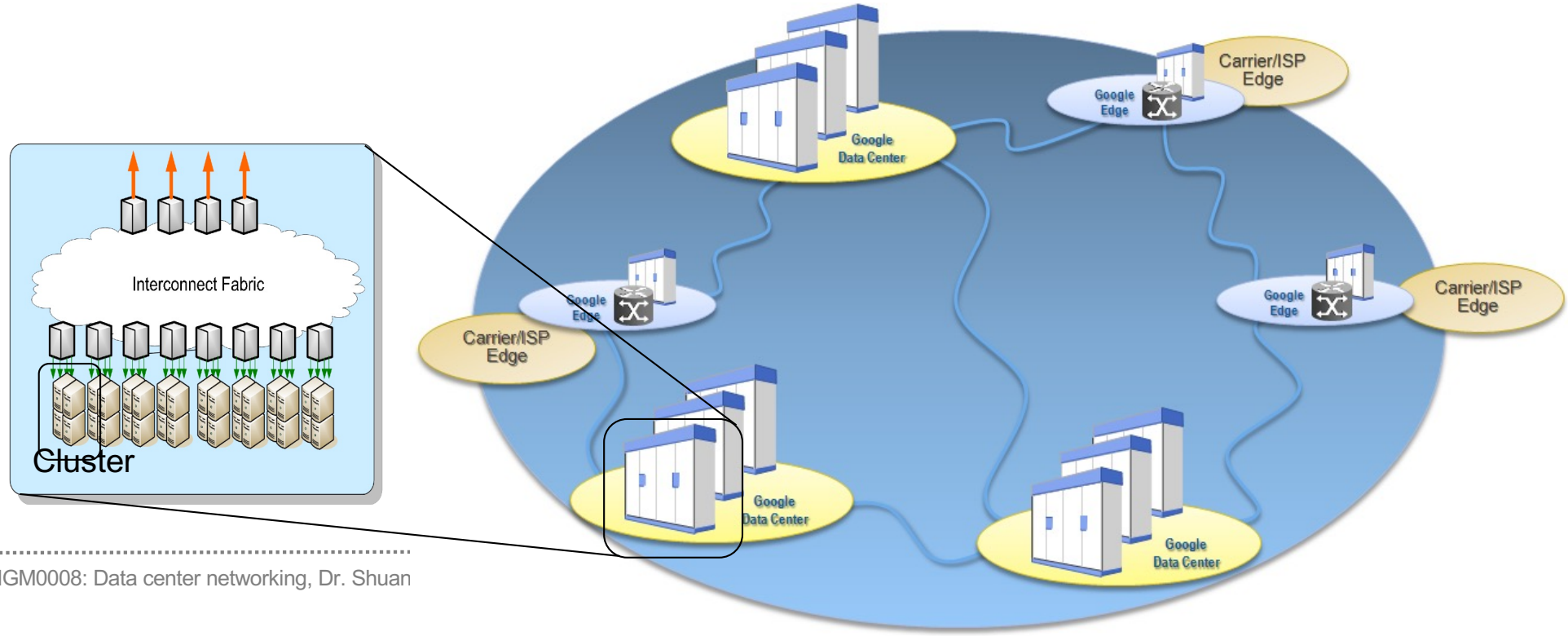
University of BRISTOL

Name	Fiber	Reach	Modulation
400G-ZR/ZR+	Duplex SMF	10km-1000km	16-QAM
400G-FR4/LR4	Duplex SMF	2km/10km	100G-PAM4
400G-DR4	8xSMF	500m/2km	100G-PAM4
400G-SR8	16xMMF	50m	50G-PAM4
400G-CR8	copper	3m	50G-PAM4

Datacenter Optics - Overview

University of BRISTOL

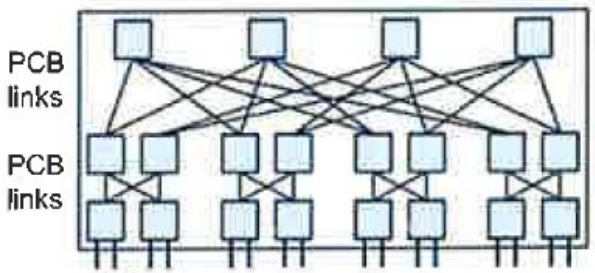
FIGURE OF MERIT	Cost Power	Cost Power	Cost	Spectral efficiency Reach		
TECHNOLOGY	MMF	SMF WDM	WDM	10G / 25G DWDM	100G coherent (QPSK)	QPSK BPSK
REACH	Intra-Datacenter	Campus	Metro	Regional	LH/ULH	Trans-Oceanic



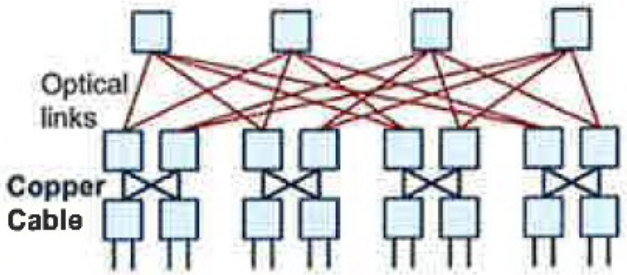
Optical Switching Solutions

A possible solution for large scale switch fabrics

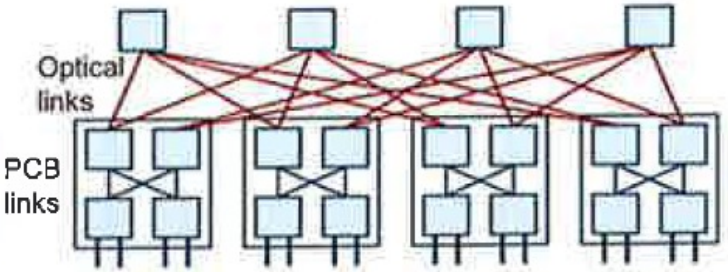
System packaging for switching



Monolithic chassis (Ref:1)



Discrete switch box (Ref:2)
(24-port ToR)

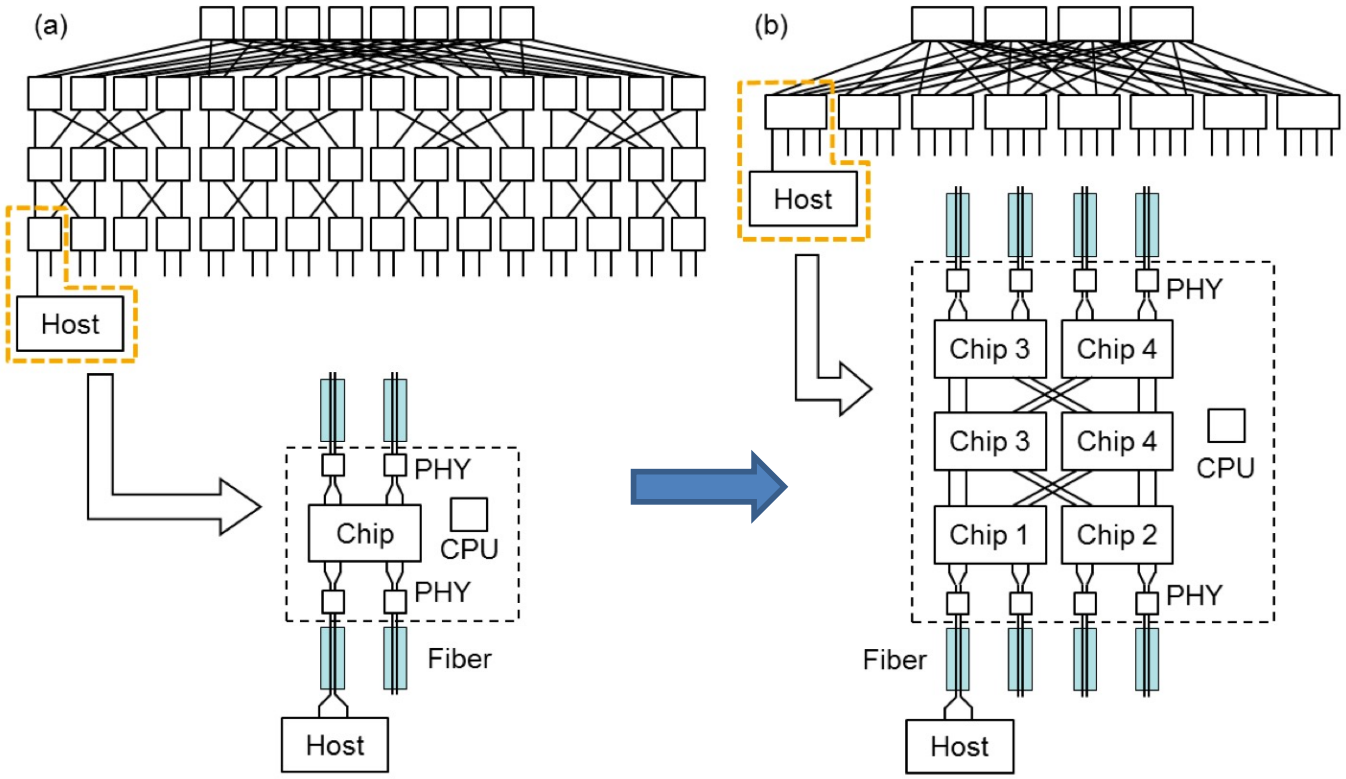


Pod and Core switch (Ref:3)
(Chassis with 48-port line-cards)

	monolithic chassis	720 discrete switch	Pod and core switch
# of boxes	1	720	26
# of rack space	80 RU	720 RU	96 RU
# of rack	2	18	3
# of switch chips	720	720	720
# of PCB links	6912	0	3456
# of copper cable links	0	3456	0
# of optical links	0	3456	3456
Development cost	High	Low	Medium
Deployment cost	Low	High	Medium

1. http://www.sun.com/products/networking/datacenter/ds3456/ds3456_wp.pdf
2. M. Al-Fares, etl, "A Scalable, Commodity Data Center Network Architecture," SIGCOMM'08, Aug. 17-22, 2008, Seattle, Washington, USA
3. N. Farrrington, etl, "Data Center Switch Architecture in the Age of Merchant Silicon," IEEE HotInterconnect '09, New York, USA.

Multi-chip



Advantages:

- Multiple switch chips in a chassis
- Copper backplane connection
- Reduce fiber and optical transceiver

Disadvantages:

- more switch chips in a chassis
- Larger worst-case number of hops (latency)
- Hugh chassis power consumption

Single channel traditional

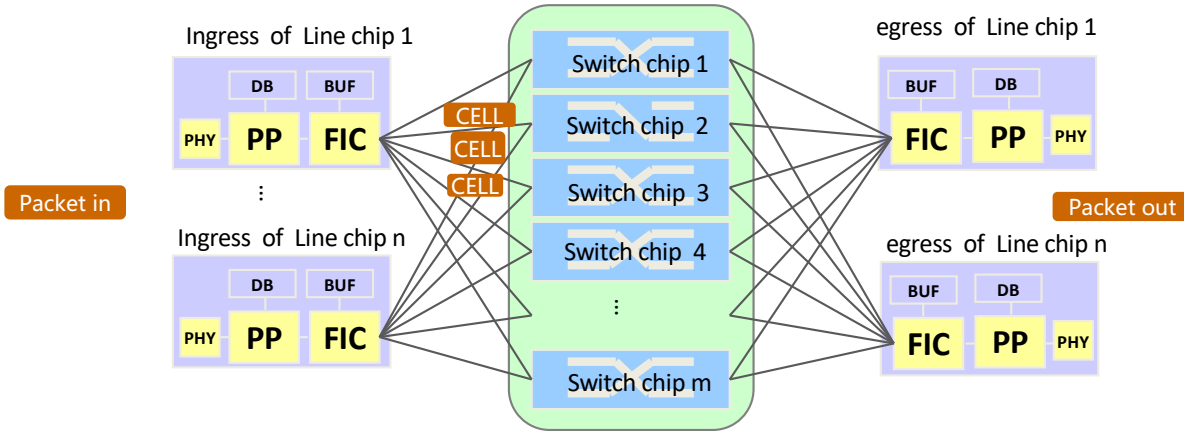
Single channel chassis-based

W. M. Mellette, A. C. Snoeren, and G. Porter, "P-FatTree: A multi-channel datacenter network topology," 2016, pp. 78–84.

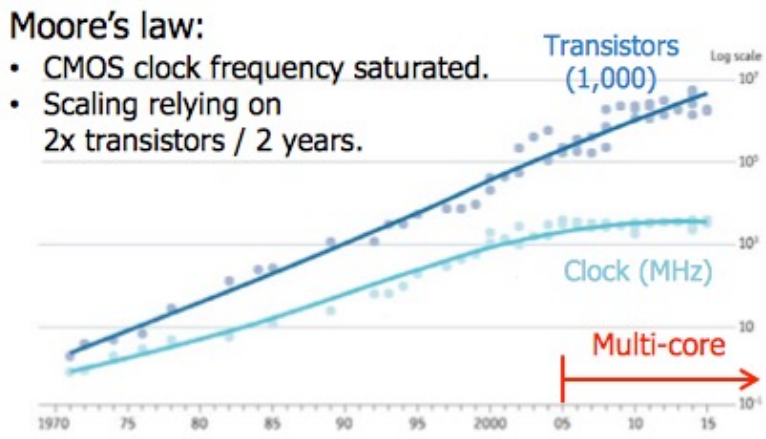
Scale up for large scale DCNs

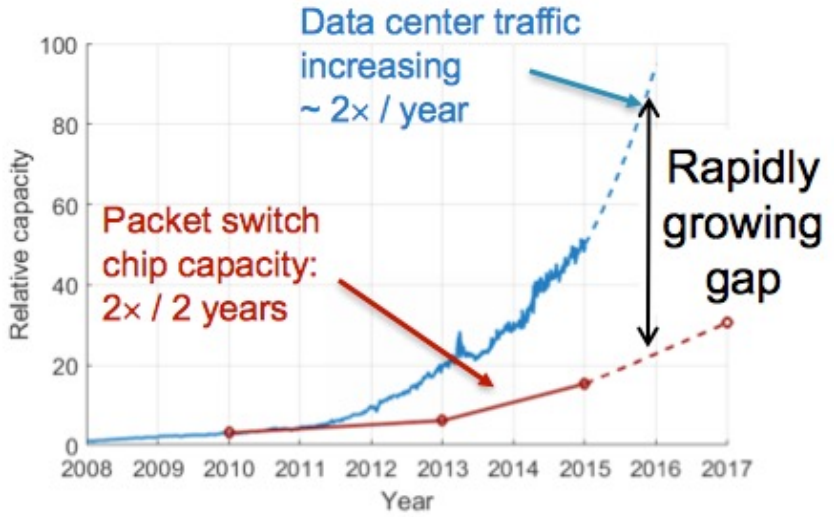
- Scale up electrical switch
  - Switch Capacity
  - Port count (radix)

Source from:
George Rapen, OFC 2017, M3k.1



CLOS architecture for large capacity switch



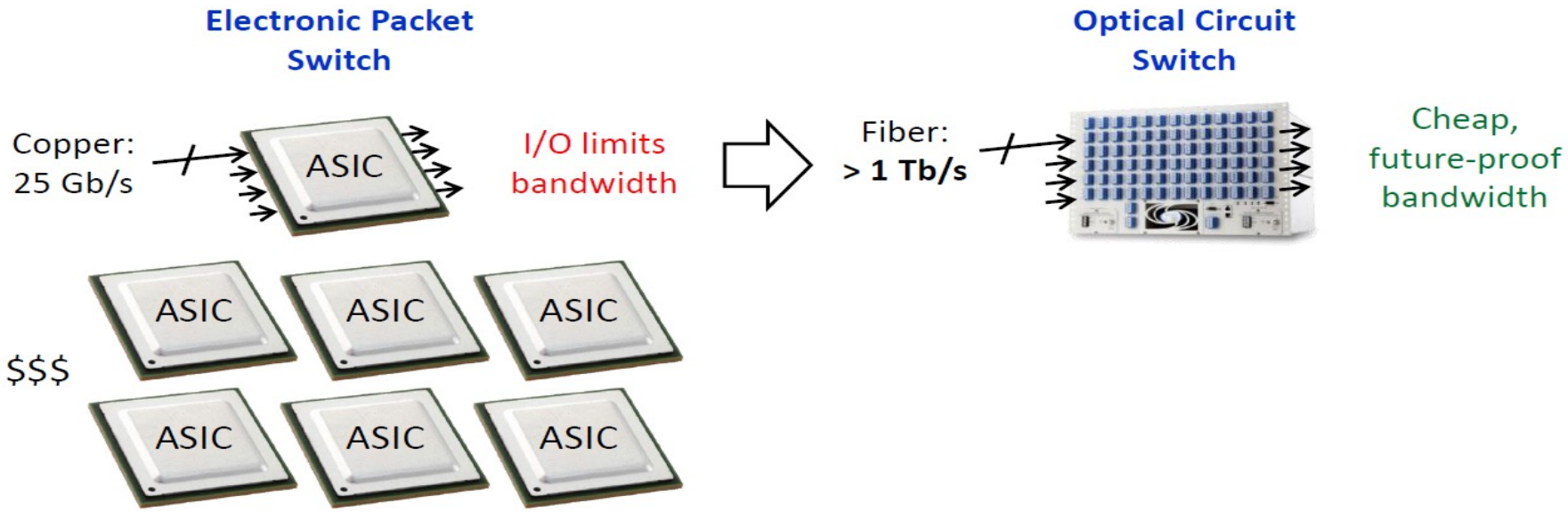


Data center traffic from

A. Singh et al., "Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network," SIGCOMM 2015.

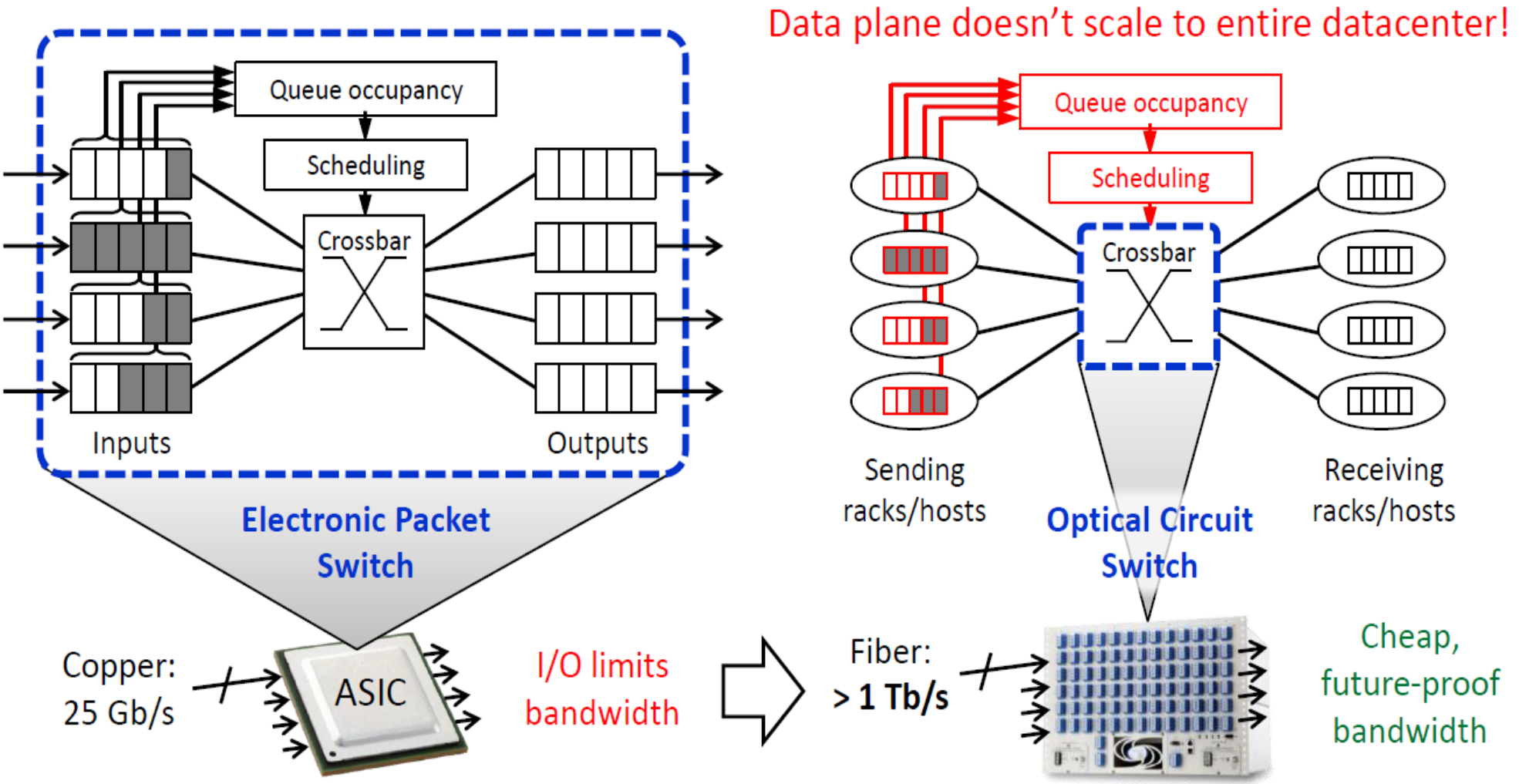
The ITRS projections for signal-pin count and per-pin bandwidth are nearly flat over the next decade. Single-chip bandwidth saturates.

What about optic switches?



- ✓ **No O/E/O**
- ✓ **Practically unlimited Bandwidth per port** (limited by spectral efficiency)
- ✓ **Low Power:** Energy irrespective of throughput (power spent on steering pipes rather than processing/transmitting bits)
- ✓ Potentially **scales to thousands ports**

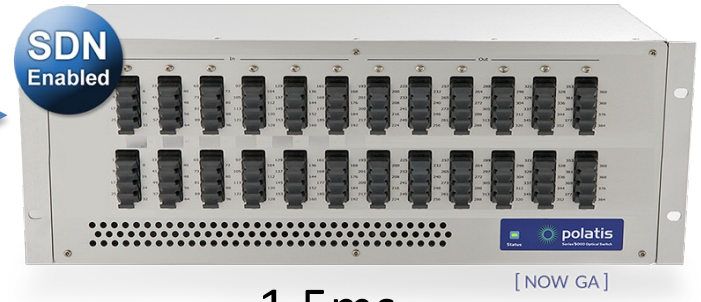
Barrier no1: lack of buffering



Barrier no2: speed

University of BRISTOL

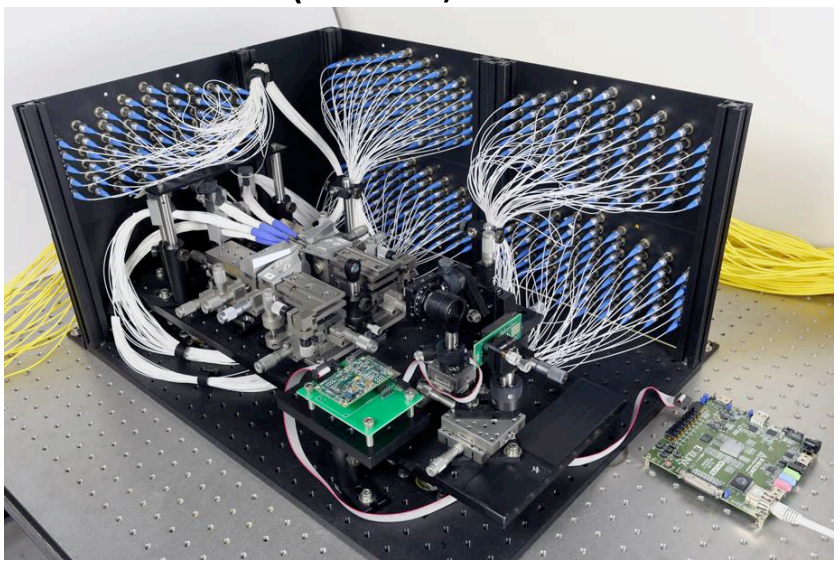
Electromechanical:
- **High port**
- **low speed**



1.5ms

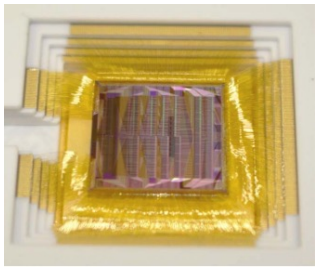
UCSD/BERKELEY 2017

2D MEMS (3.2 ms)

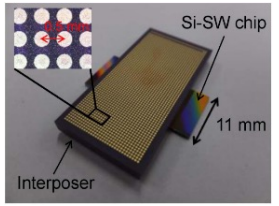


Thermo-Optic

- **Huawei**
  - 32x32, TO, 23 dB, 750 μs



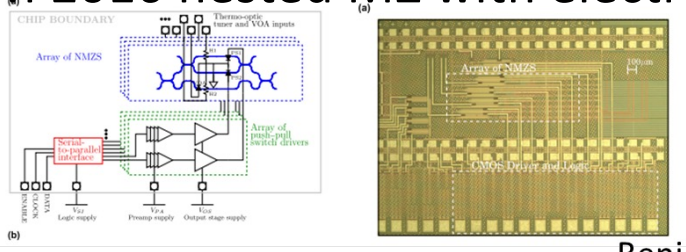
- **AIST**
  - 32x32, TO, 14.5 dB, 30 μs



Photonic integration: Fast Optical switches

University of BRISTOL

IBM 2016 nested MZ with electronic

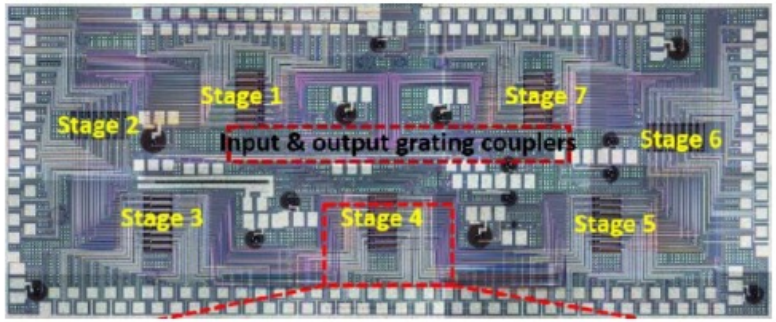


4ns

Benj. Lee

2016

16×16 EO Mach-Zehnder Switch



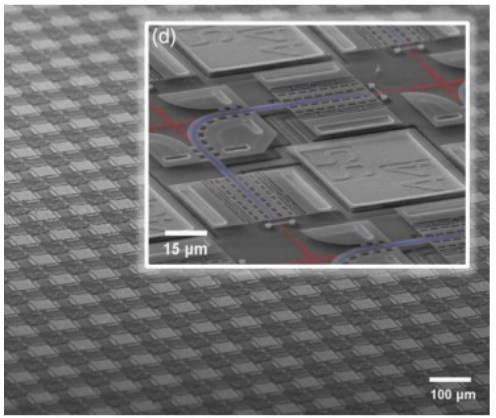
[Lu, Opt. Exp. 2016, 24 (9) 9295]

3.2/2.5ns

Berkeley 2016

0.91 µs

64×64 2D MEMS Switch

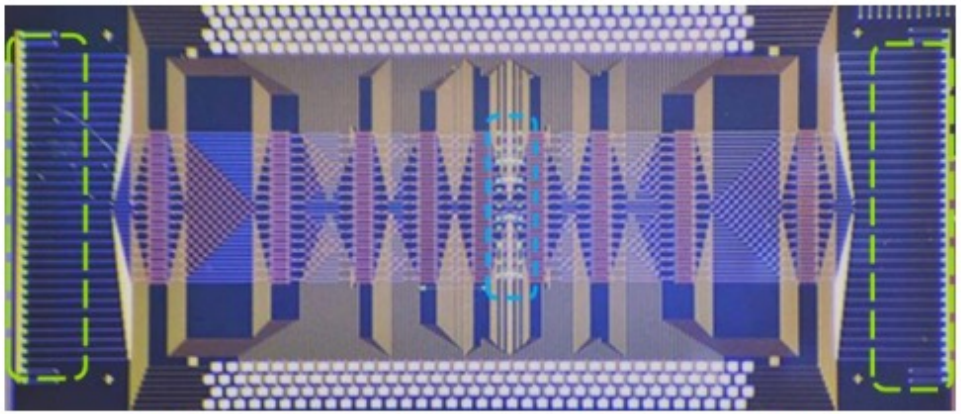


[Seok, Optica, vol. 3, no. 1, pp. 64]

Tae Joon Seok, **Niels Quack,** Sangyoon Han, Richard S. Muller, and Ming C. Wu

2017

32×32 EO Mach-Zehnder Switch



[Qiao, Nature Sci. Rep. 2017, 7 42306]

1/1.2 ns

Optical Switching technologies

- Port to Port switch
  - 3D- MEMS switch
    - Coarse-grained switching technology (Fiber to fiber, port to port)
    - Up-to 320×320 switch fabric
    - Several tens ms switch time

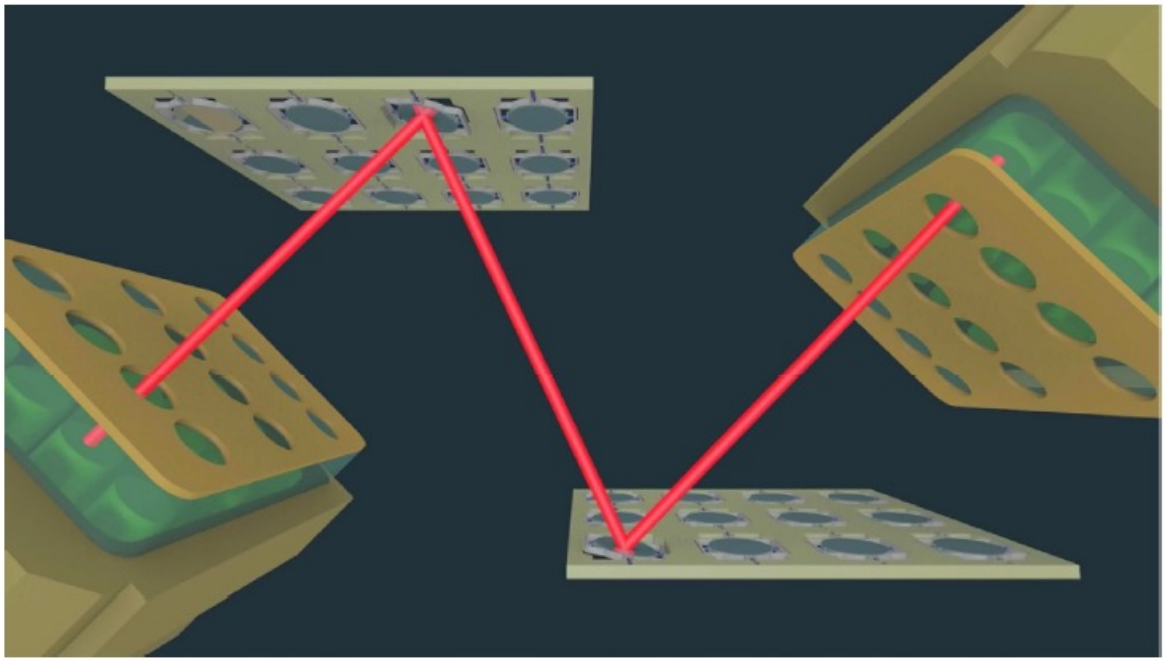
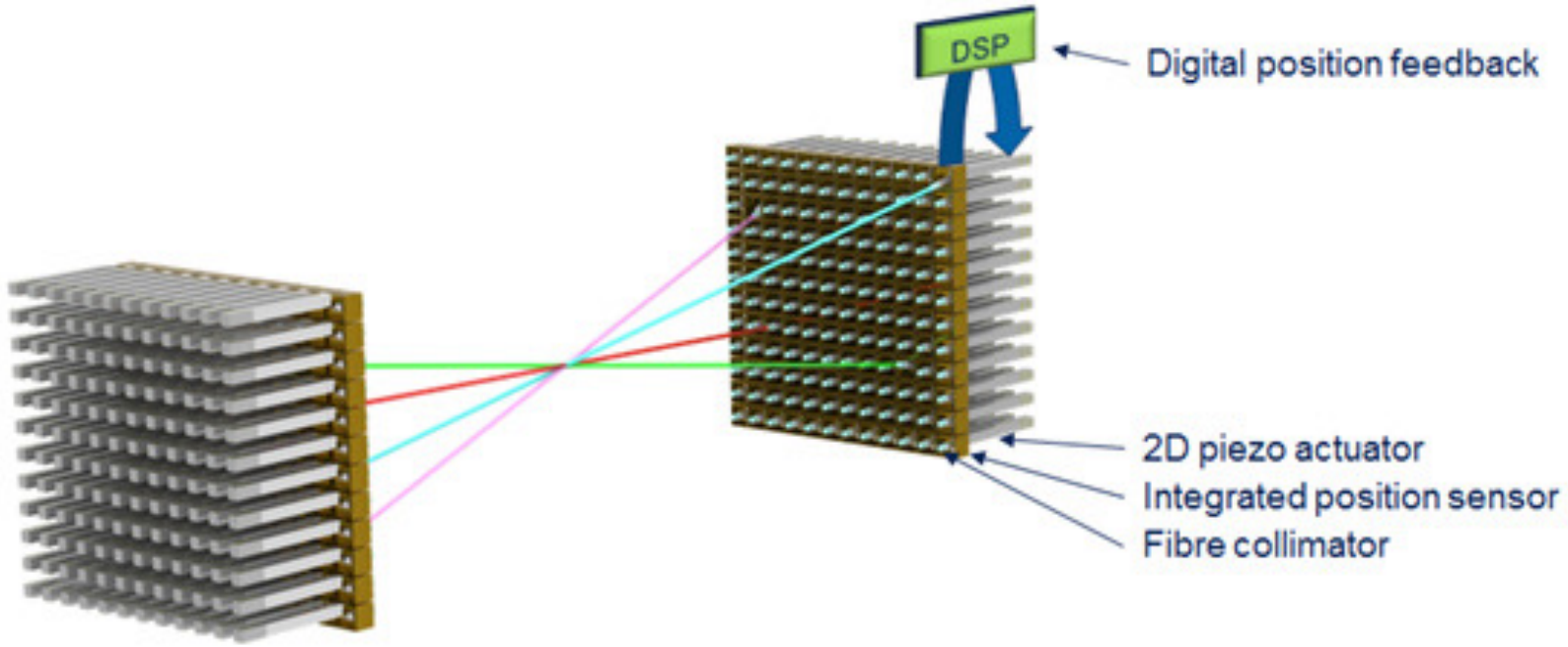


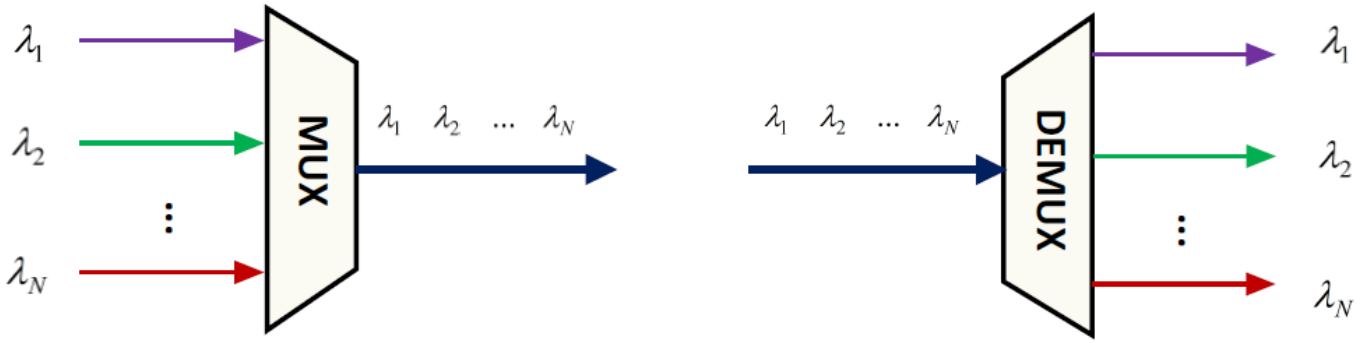
Figure 1.11: A three-dimensional optical MEMS switch.

Optical Switching technologies

- Port to Port switch
  - Beam steering technologies (Polatis)
    - Coarse-grained switching technology (Fiber to fiber, port to port)
    - Up-to 320×320 switch fabric
    - Several tens ms switch time



- Array waveguide grating (AWG) based optical switch
  - MUX and DEMUX



(a) A $N \times 1$ MUX (b) A $1 \times N$ DEMUX

  - AWGR (Arrayed waveguide grating router)

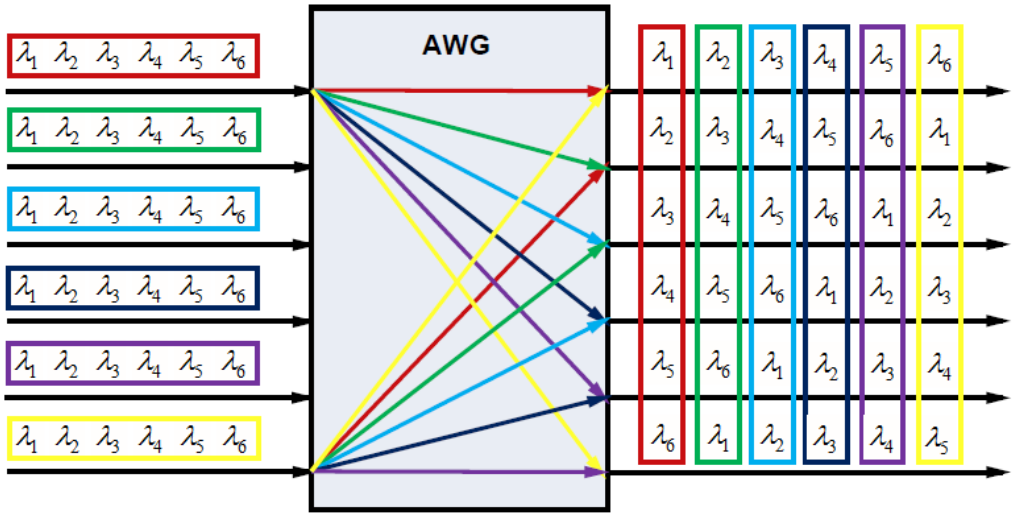
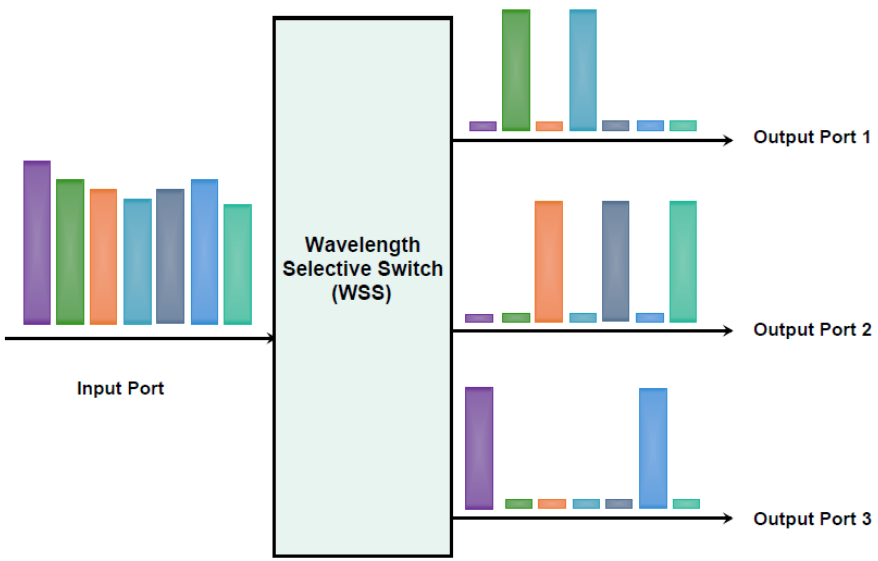
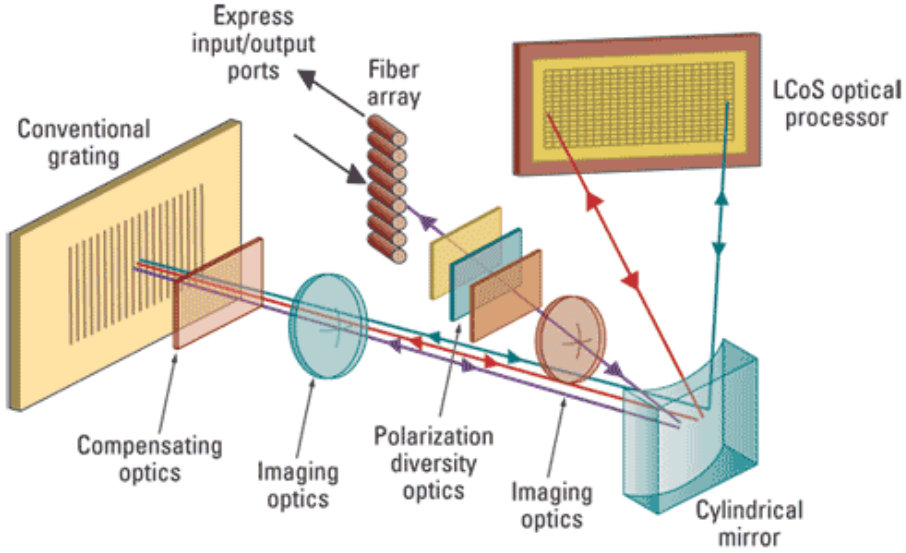


Figure 1.16: The cyclic wavelength routing pattern of a 6×6 AWG switch.

Optical Switching technologies

- Wavelength selective switching (1×3 WSS)





SOA-based time switch

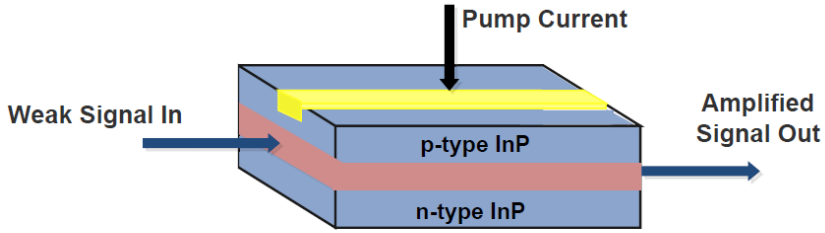
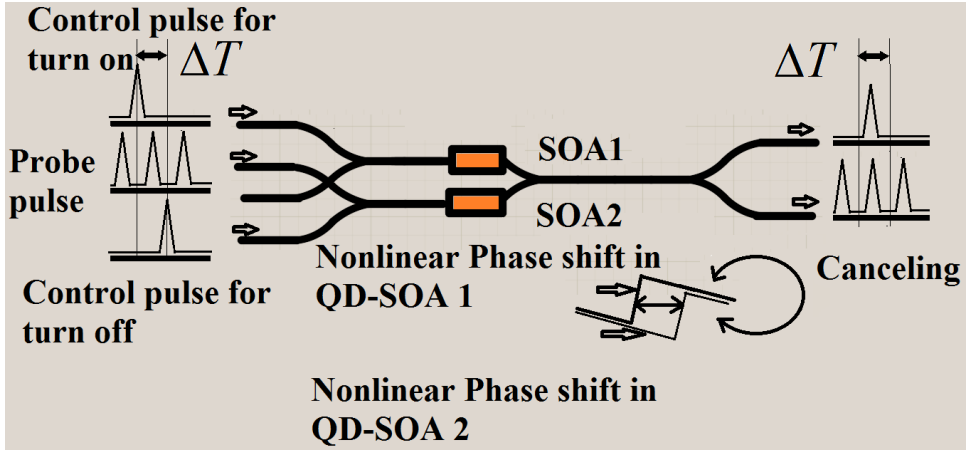
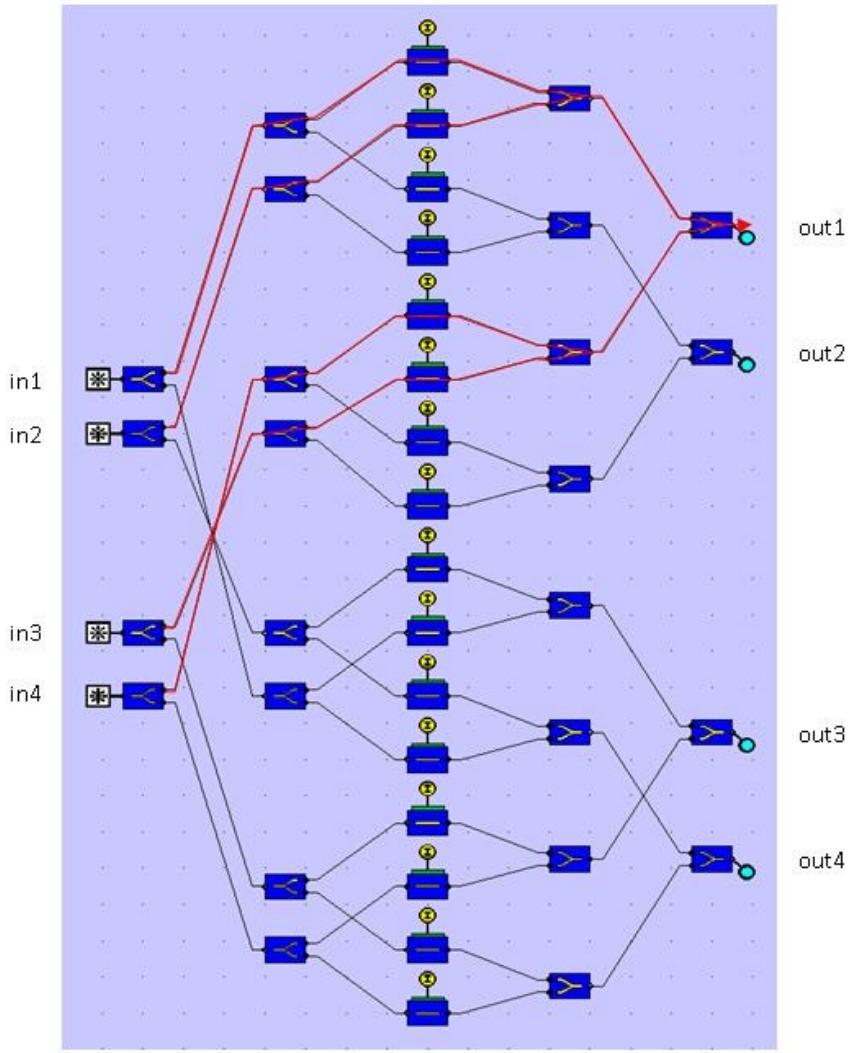


Figure 1.15: A Semiconductor Optical Amplifier (SOA).



All-optical switching

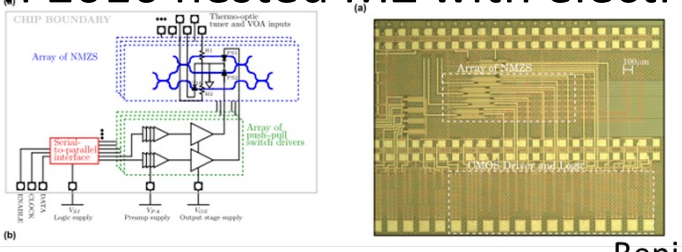


Optical Switch Circuit

Photonic integration: Fast Optical switches

University of BRISTOL

IBM 2016 nested MZ with electronic

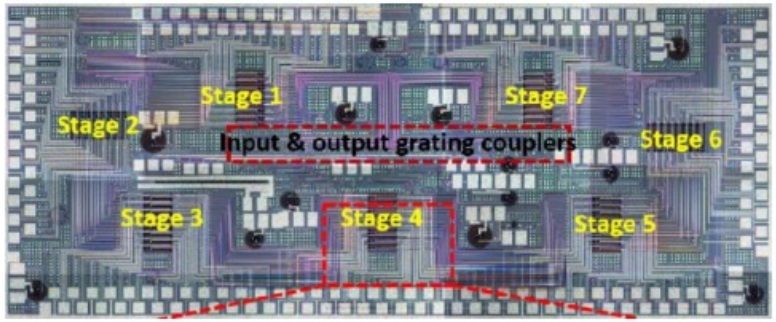


4ns

Benj. Lee

2016

16×16 EO Mach-Zehnder Switch



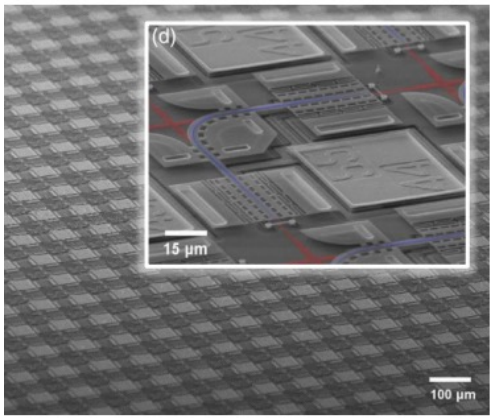
[Lu, Opt. Exp. 2016, 24 (9) 9295]

3.2/2.5ns

Berkeley 2016

0.91 µs

64×64 2D MEMS Switch

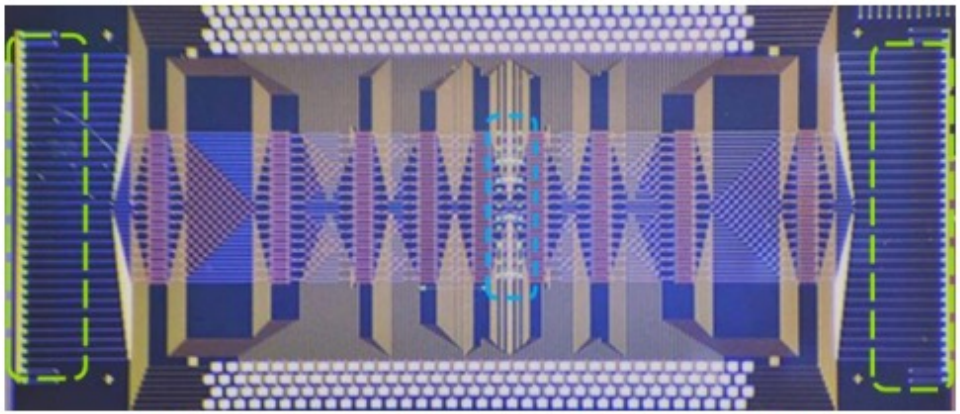


[Seok, Optica, vol. 3, no. 1, pp. 64]

Tae Joon Seok, **Niels Quack,** Sangyoon Han, Richard S. Muller, and Ming C. Wu

2017

32×32 EO Mach-Zehnder Switch



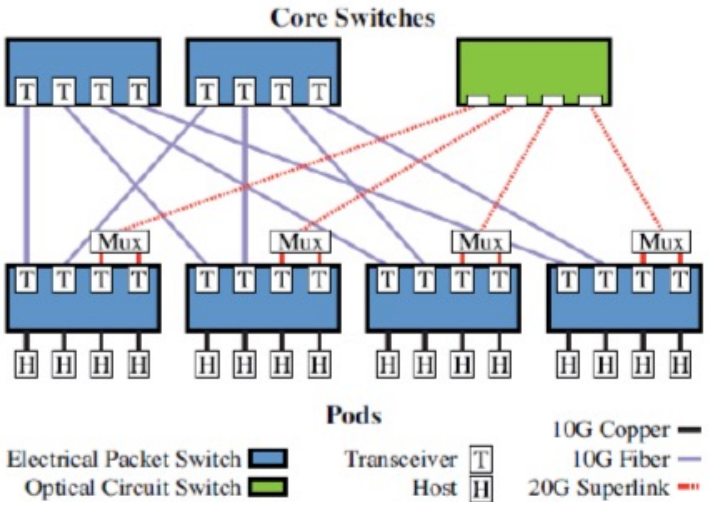
[Qiao, Nature Sci. Rep. 2017, 7 42306]

1/1.2 ns

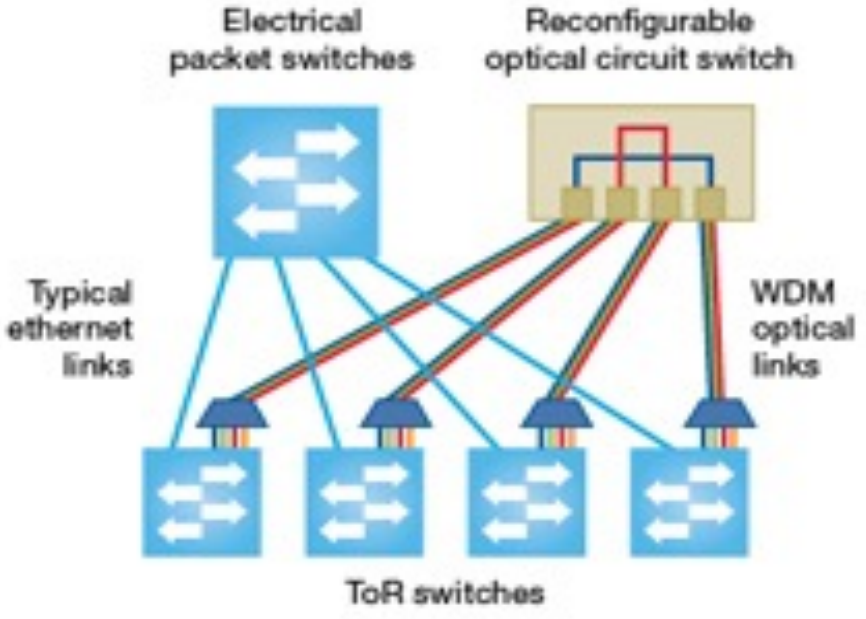
Hybrid architectures

- *Hybrid architectures (Fat tree architecture is enhanced using Optical Circuit Switching)*
  - Easily implemented (commodity switches)
  - Slow switching time (MEMs). Good only for bulky traffic that lasts long
  - Not scalable (constraint by Optical switch ports)

Optical switches handle large elephant flows

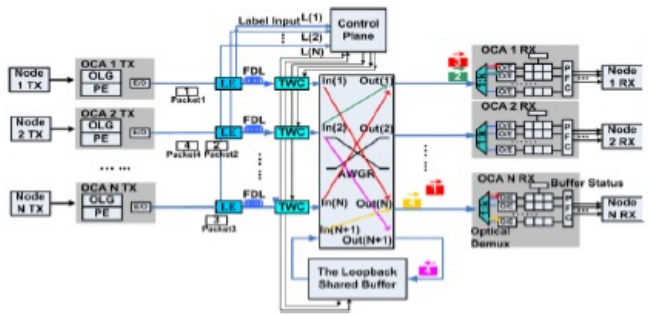


Farrington (SIGCOMM 2011)

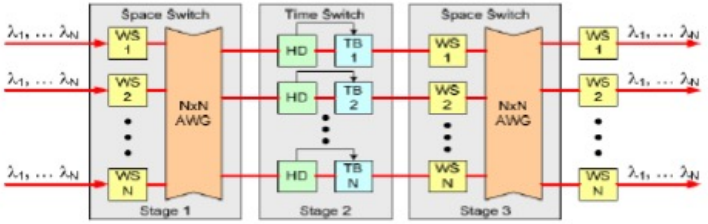


High radix optical architectures

Optical switch architectures with high radices in order to lead to more flat DC architectures (less tiers) by replacing electrical switches in the upper tiers of the fat-tree

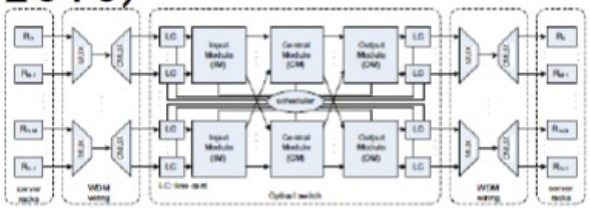


Ye et al (SANCS 2010)



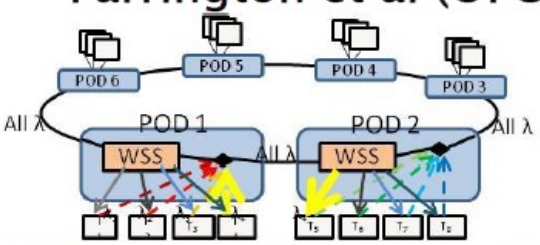
Gripp et al (OFCC 2010)

Xia et al (TR 2010)

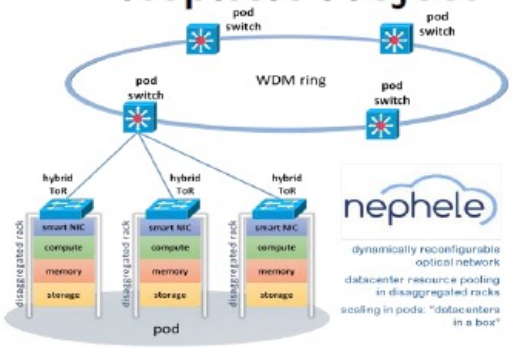


- *Alternatives to fat trees*

Farrington et al (OFCC 2013)

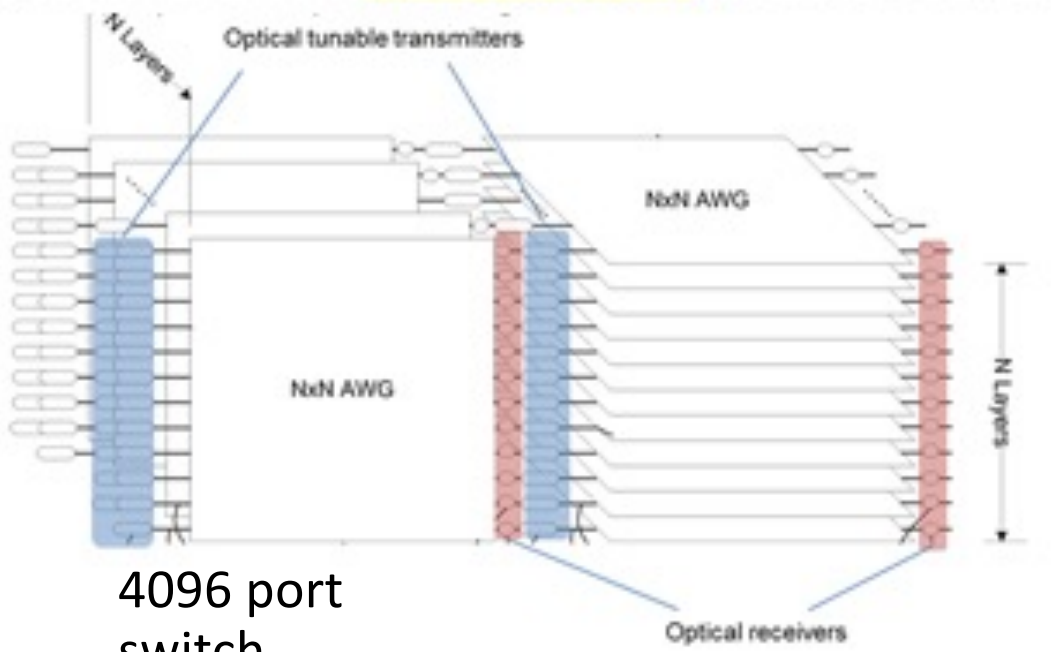


Nephele Project

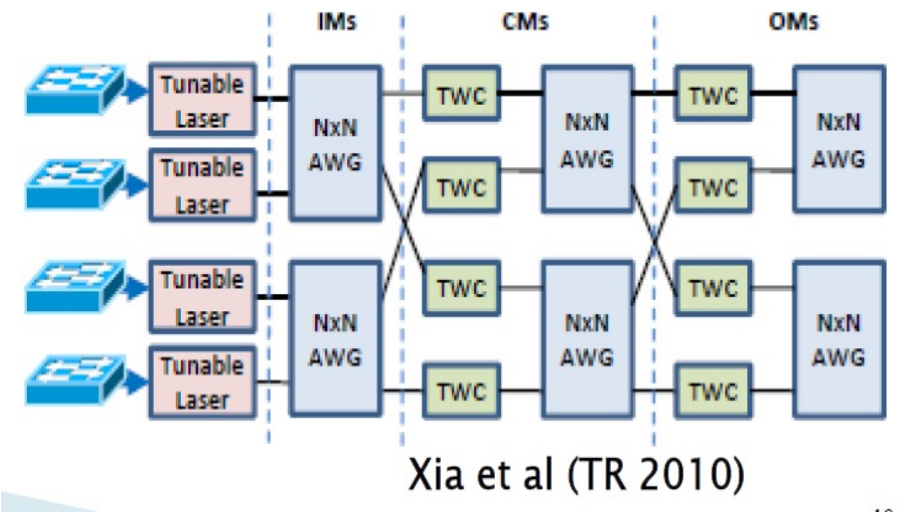


Petabit switch fabric

- Petabit switch fabric: three-stage Closnetwork and each stage consists of an array of AGWRs that are used for the passive routing of packets.
- In the first stage, the tunable lasers are used to route the packets through the AWGRs, while in the second and in the third stage TWC are used to convert the wavelength and route accordingly the packets to destination port.



4096 port switch



Xia et al (TR 2010)

Summary

- Optical fibres as transmission medium
- Transmitter design for high-speed low-cost interconnection
- Possible opportunities for optical switching in DCN