OPTICA
Advancing Optics and Photonics Worldwide

ON2030

Optical Networks toward 2030 Webinar #2

Moderator: David Hillerkuss
June 26, 2024, 9:00-10:00am EST

About ON2030

New bi-monthly webinar series,
"Optical Networks toward 2030 (ON2030)"

The webinar strives to provide an overview of the most important topics in our industry

Key experts explore next generation technologies, including critical aspects such as

- energy efficiency,
- reliability,
- sustainability,
- efficient ecosystem scaling,
- and future-proof solutions.

Update on key advances in international optical network standards (ITU-T, IEEE, OIF and BBF etc.)

Join this series to stay up to date with latest developments and highlights

See the website for regular updates and future instances:

https://www.optica.org/membership/member_programs/optical_networks_toward_2030/

Agenda

Intra-Data-Center Optics – Emerging Applications and Technology Trends

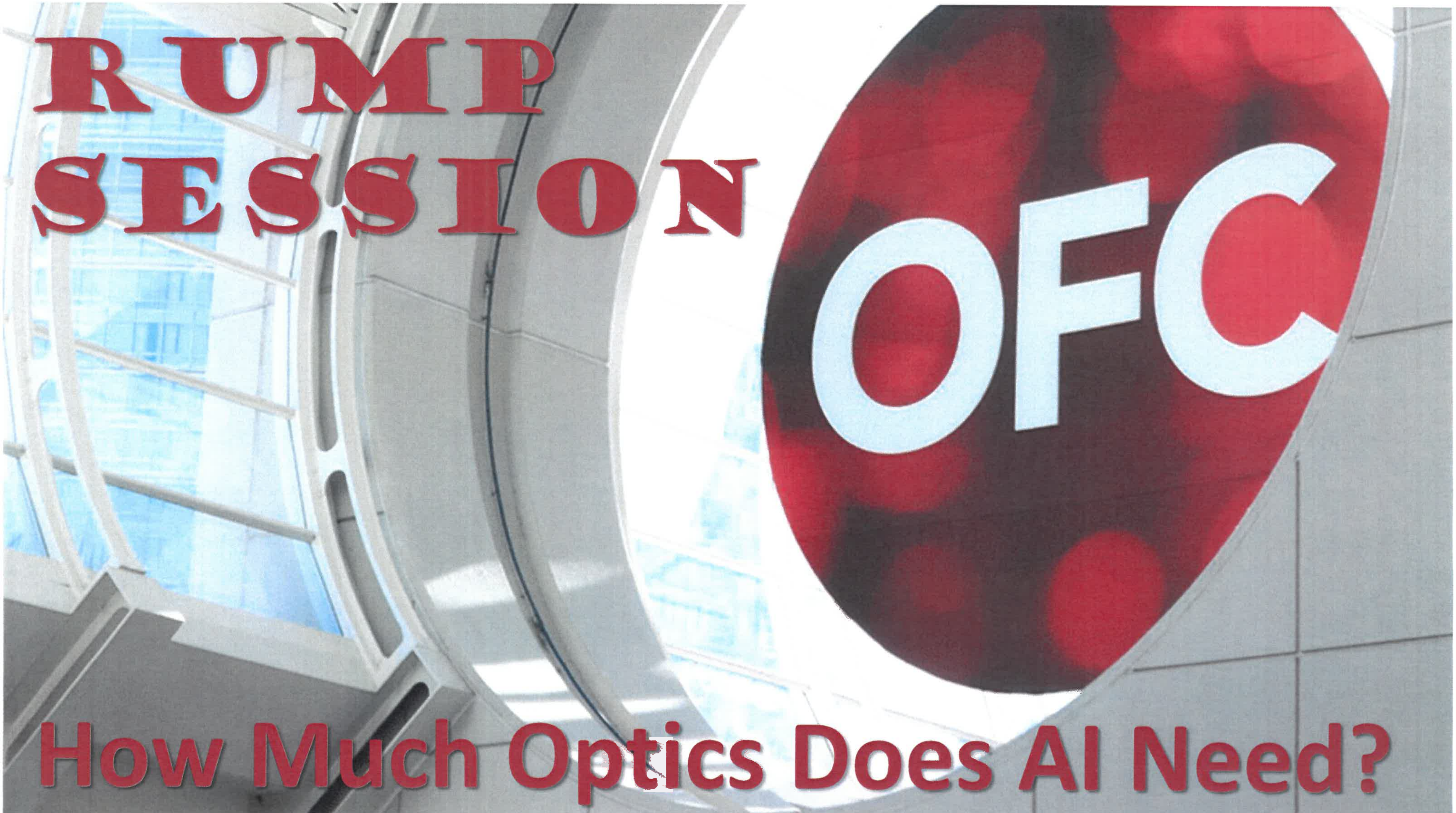
- Ben Lee – Nvidia
- Peter Winzer – Nubis Communications
- Chris Cole – Parallax Group
- Q&A / Panel Discussion

Moderator

- David Hillerkuss – Infinera

Panel Discussion

Thank you


RUMP
SESSION
OFC
How Much Optics Does AI Need?

1. AI Cluster Architectures

   Ashkan Seyedi – Nvidia
   John Shalf – Lawrence Berkeley National Laboratory

2. To Plug or to Co-package in AI?

   Andy Bechtolsheim – Arista
   Near Margalit – Broadcom
   Mark Lutkowitz – FibeReality
   Thomas Liljeberg – Intel
   Chris Pfistner – Avicena

3. Will Optical Switches replace Electronic Switches in AI?

   Ryohei Urata – Google
   Andy Bechtoldsheim – Arista

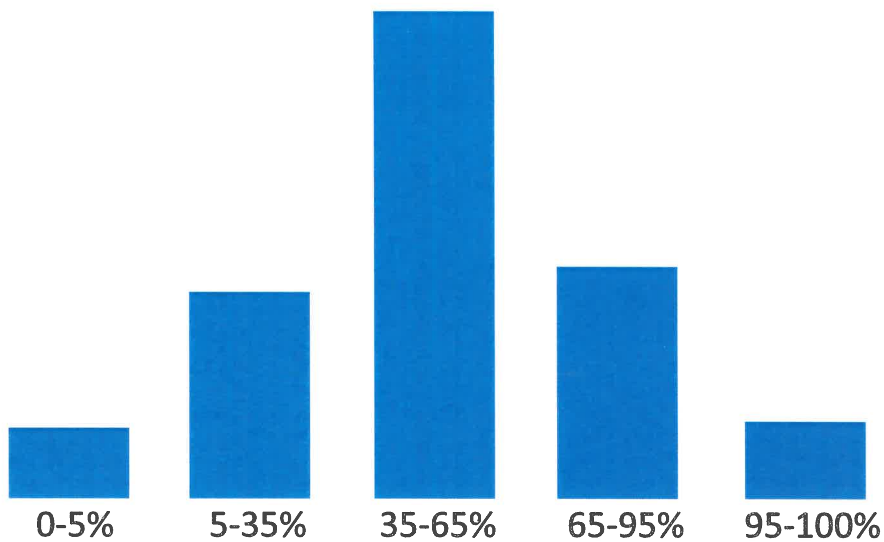
4. Should AI ASIC interposers include optics functionality?

   Nick Harris – LightMatter
   Dave Lazovsky – Celestial.AI
   Chris Cole – Independent Consultant

5. Optical computing – Eternal hype or the next step for AI?

   Patrick Bowen - Neurophos
   Chris Cole – Independent Consultant

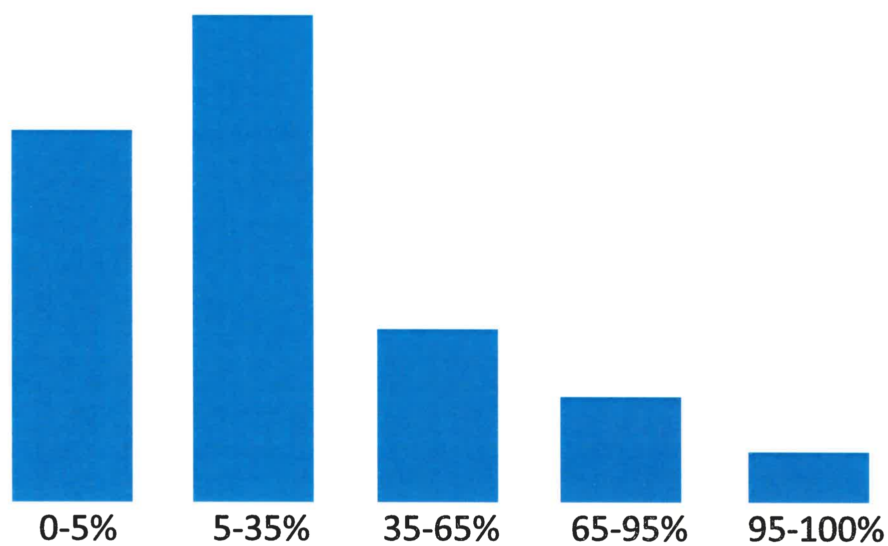
In 5 years from now, what fraction of data centers will be AI clusters?



~250 responses

➔ 50% of data centers will be AI clusters

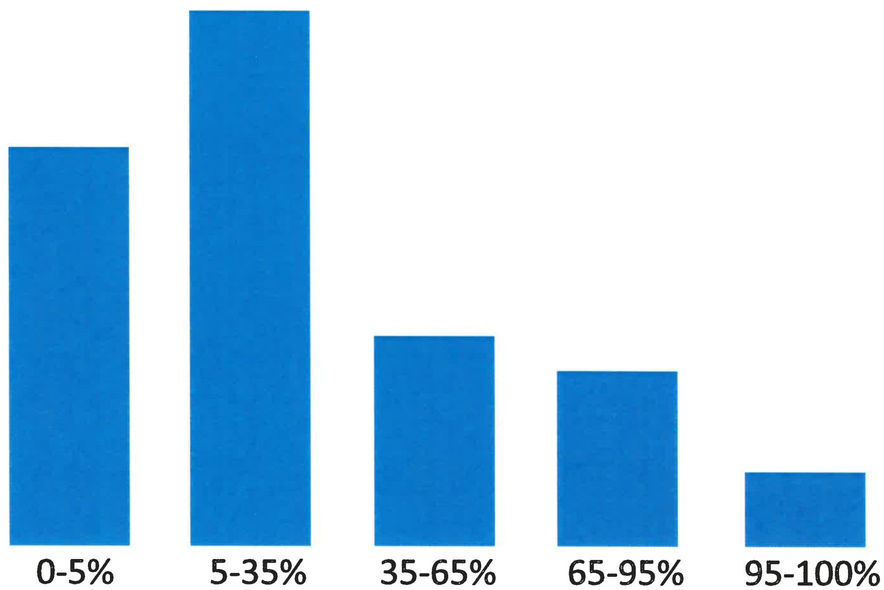
In 5 years from now, what fraction of I/O in AI clusters will be CPOs (vs pluggables)?

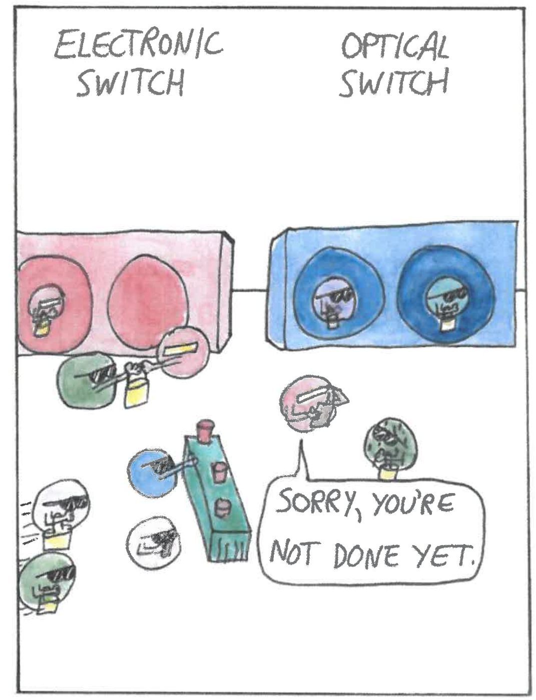


~220 responses

➔ 30% of AI cluster optics will be CPO

In 5 years from now, what fraction of AI cluster switching will be optical?



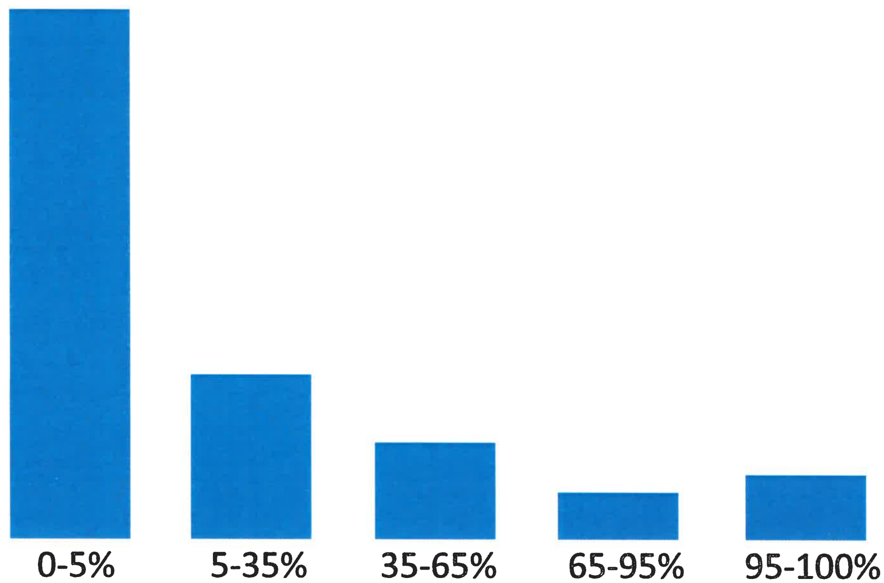


© Chris Doerr

~200 responses

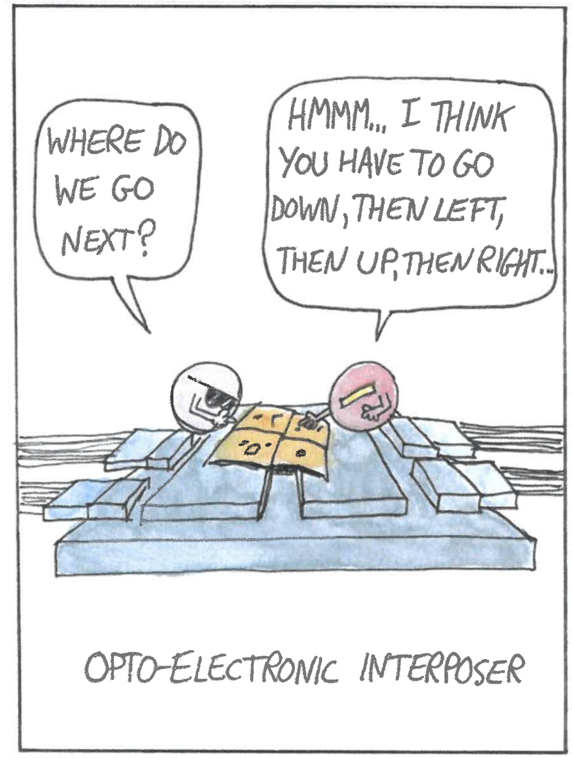
➔ 30% of AI cluster switching will be optical

In 5 years from now, what fraction of AI chips will be on optical interposers?



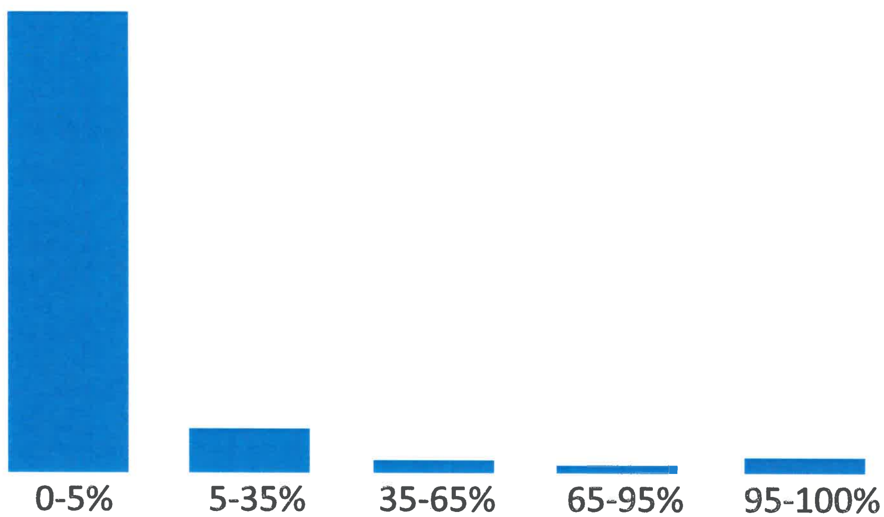
~200 responses

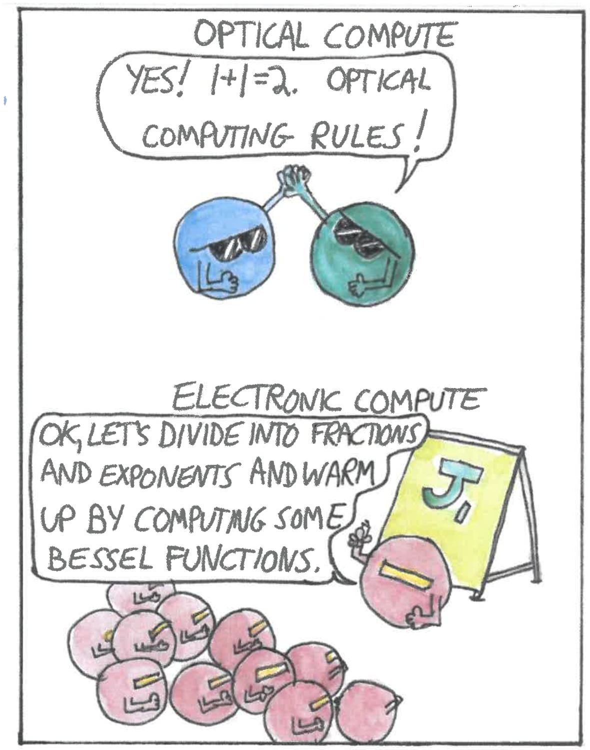
➔ <20% of AI chips will be on optical interposers



© Chris Doerr

In 5 years from now, what fraction of AI processing will use optical compute?





~190 responses

➔ <10% of AI processing will use optical compute

OFC 2024 Rump Session Survey Summary

In 5 years from now:

→ Half the Data Center infrastructure will be ML/AI clusters

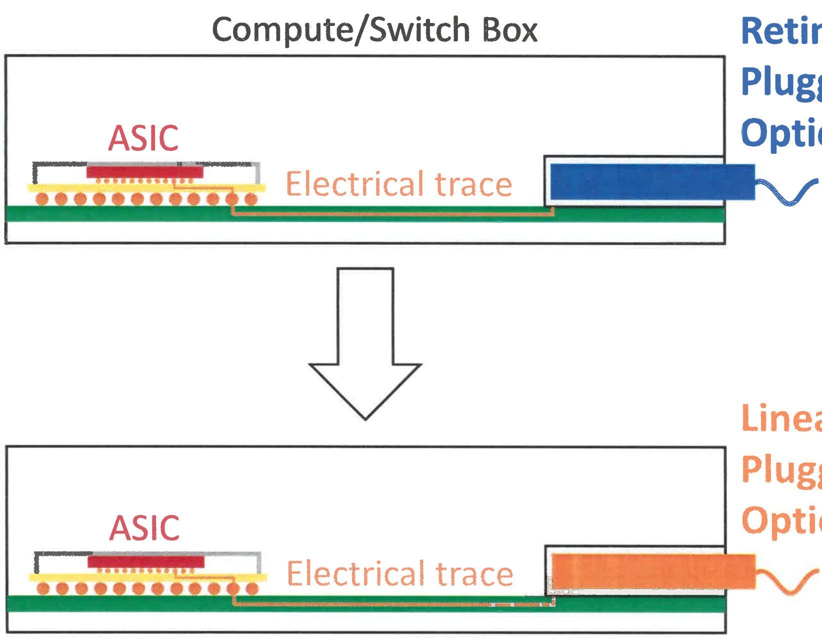
→ But based on fairly traditional technologies

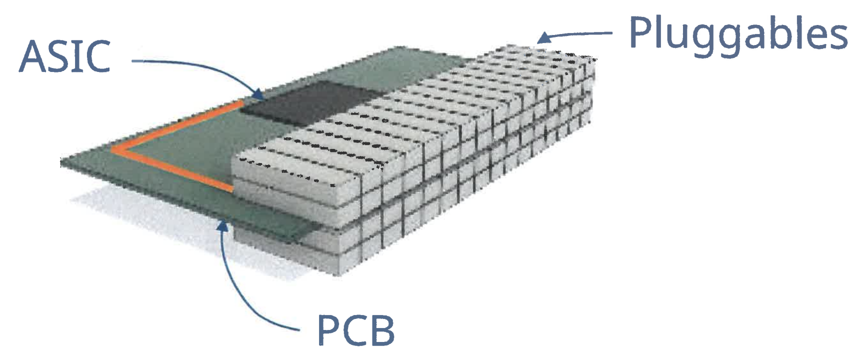
- More pluggables/NPO than chiplets/CPO
- More electronic switching than optical switching
- Not a lot of optical interposers
- Very little optical computing

High-Density Optical I/O for ML/AI Applications

ON2030 Webinar, June 2024

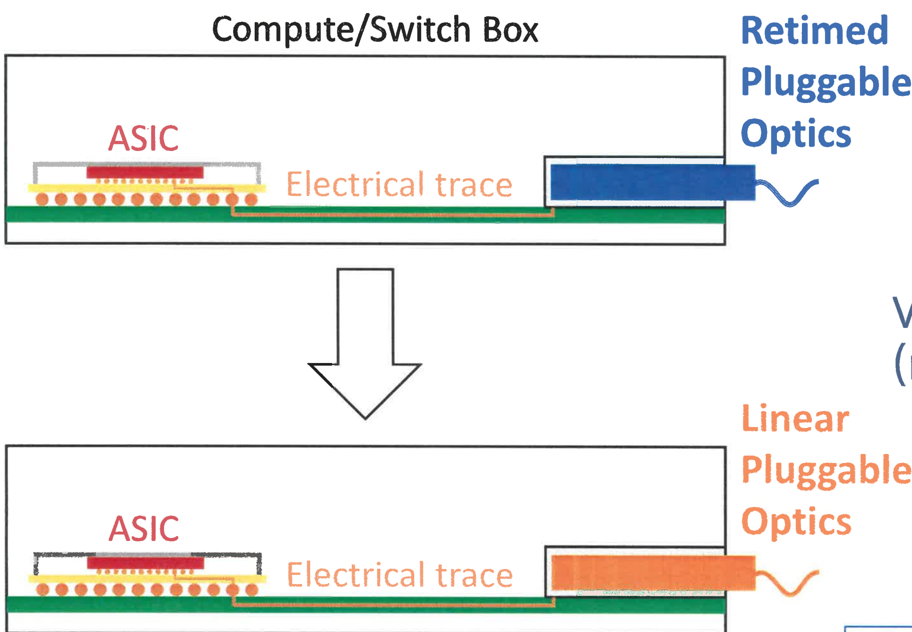
Pluggable Optics: Low Escape Density, High Power



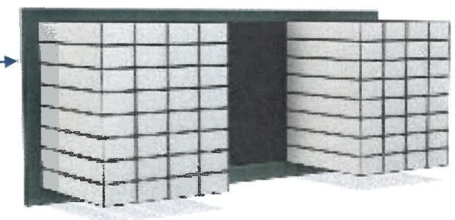


- LPO substantially reduces I/O power & latency
  - But increasingly difficult as bit rates increase

Pluggable Optics: Low Escape Density, High Power

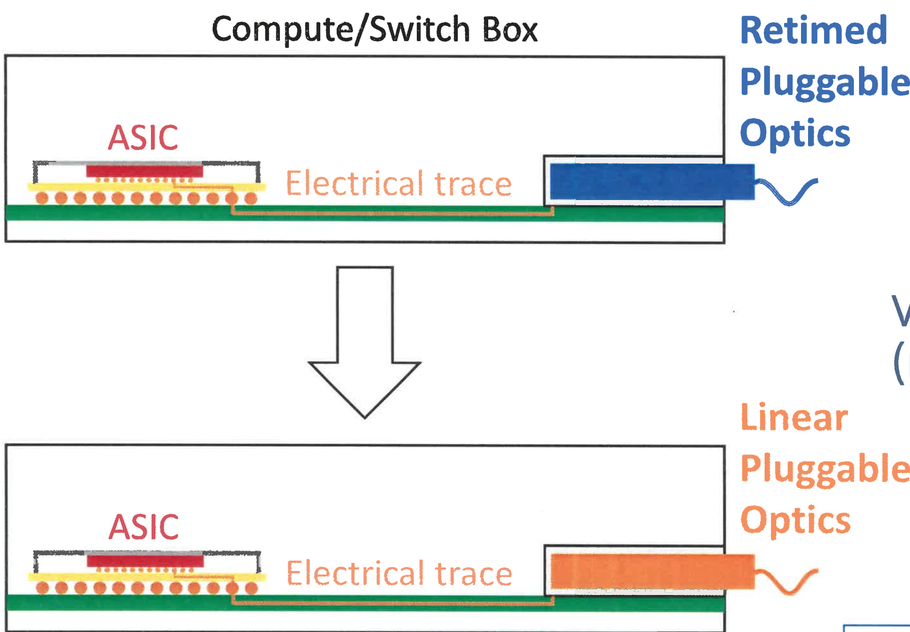


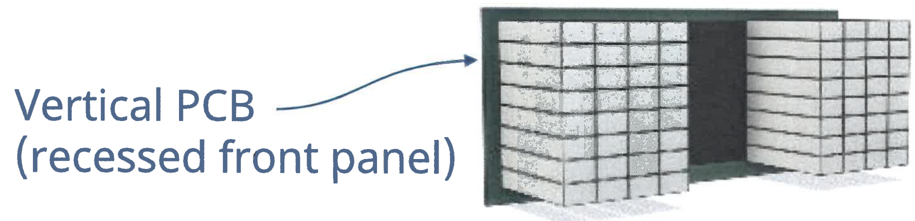
Vertical PCB
(recessed front panel)



- LPO substantially reduces I/O power & latency
  - But increasingly difficult as bit rates increase
  - Vertical Line Cards (VLC) architecture helps

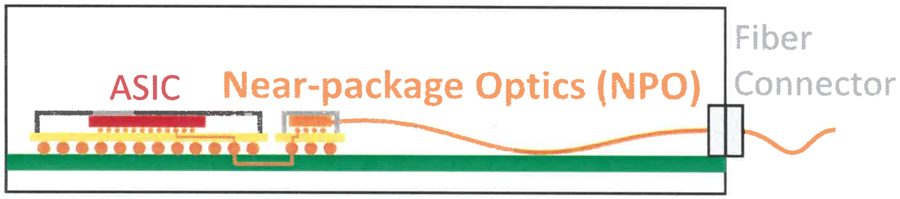
Pluggable Optics: Low Escape Density, High Power





- LPO substantially reduces I/O power & latency
  - But increasingly difficult as bit rates increase
  - Vertical Line Cards (VLC) architecture helps
- LPO does not increase escape density
  - Still fairly long electrical traces, even VLC

Linear Near-Package/Co-Packaged Optics

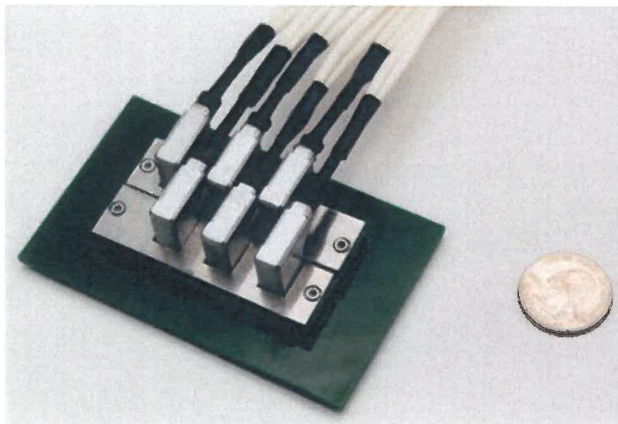


High I/O density	Tbps/mm
Low latency	<1 ns addition wrt passive copper
Low power	<6 pJ/bit (incl. laser, control)
Intra-rack to inter-row	<1 meter to 100's of meters
High Radix	~100 Gbps all-to-all connectivity
Network compatibility	PCIe, Ethernet, Infiniband

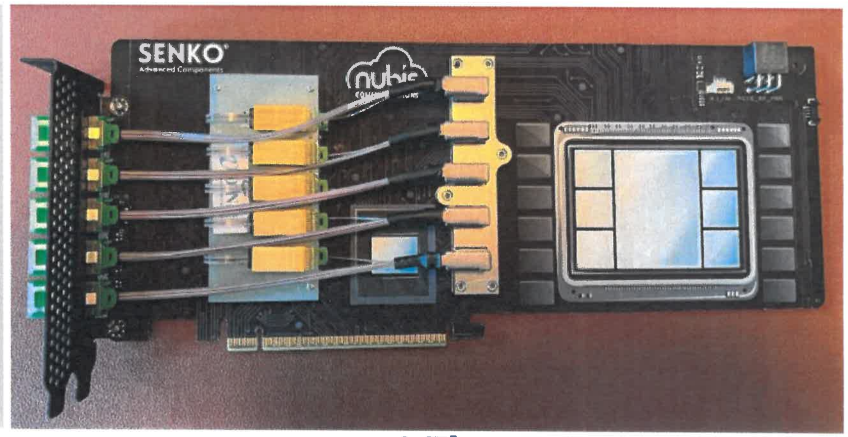
→ **HDI/O** (High-density I/O for ML/AI)

Nubis XT1600™ Near-Package DR+ Optics Modules

16x100 Gbps
(full-duplex)



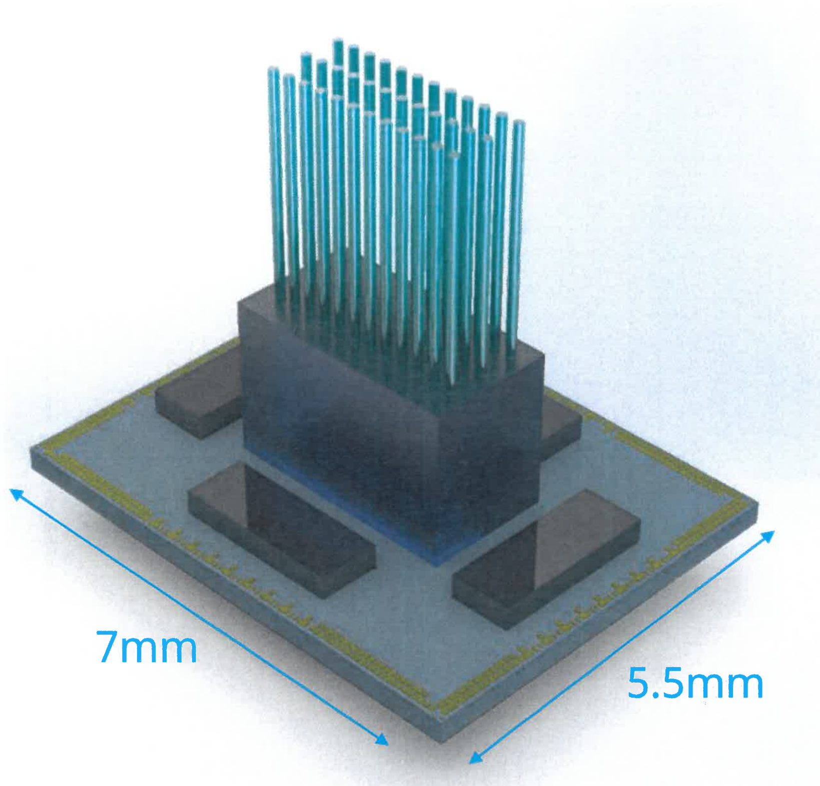
10 Tbps
(on half a business card)



8 Tbps
(ML/AI PCIe Card Interface)

2024 LIGHTWAVE INNOVATION REVIEWS 5.0

Nubis Linear DR+ HDI/O Chiplets



16 x 100 Gbps full-duplex HDI/O chiplet

Protocol agnostic (PCIe, Ethernet, …)

Enables 2D-tiled architectures:

nubis
COMMUNICATIONS

Thank you !

NVIDIA

Intra-Datacenter Optics:
Emerging Applications and Technology Trends

Benjamin Lee, NVIDIA Research

Optica ON2030 Webinar Series

26 June 2024

"Tri-verging" Network Demands

Network requirements differ between DC, HPC, and AI

Traditional Datacenter Fabrics	HPC Fabrics	AI/ML Fabrics
Ethernet	e.g., InfiniBand	e.g., NV Link
Fat-tree topologies	Various topologies	Various topologies
Highly standardized components	Balancing standardization and customization	Proprietary solutions, more tolerant to customization
Optimized for cost and interoperability	Optimized for performance at scale	Optimized for cluster performance

"Tri-verging" Network Demands

Network requirements differ between DC, HPC, and AI

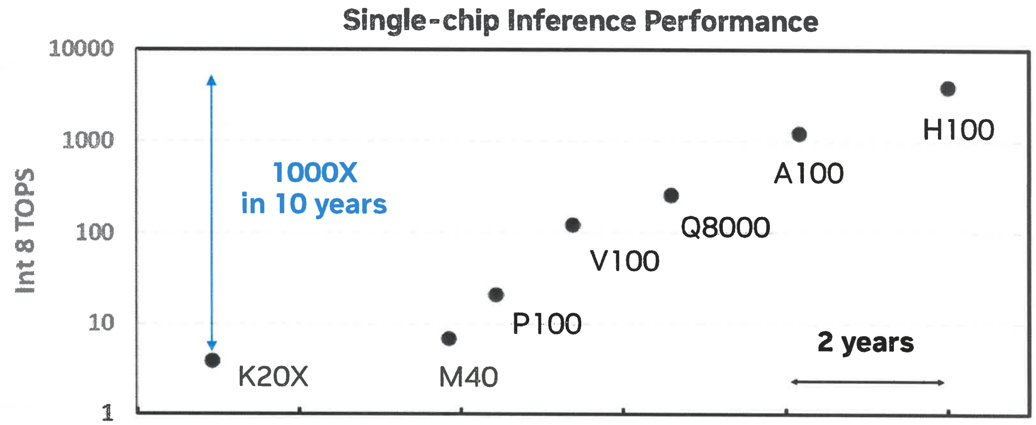
Traditional Datacenter Fabrics	HPC Fabrics	AI/ML Fabrics
Ethernet	e.g., InfiniBand	e.g., NV Link
Fat-tree topologies	Various topologies	Various topologies
Highly standardized components	Balancing standardization and customization	Proprietary solutions, more tolerant to customization
Optimized for cost and interoperability	Optimized for performance at scale	Optimized for cluster performance

GPUs Unlock the AI Revolution

Single-chip GPU performance gains have unleashed the capabilities of AI, but networks are needed to scale it

- **AI is a big deal**

- **Ingredients for AI**
  - Large data sets
  - Algorithms
  - Efficient compute



[B. Dally, HotChips 2023]

- **AI models and AI data sets are large**
  - E.g., it takes ~ 20 GPUs to hold one copy of the GPT4 model parameters
  - Typically, data sets are parallelized across multiple copies of the model (100s)
  - Number of GPUs needed for training and inference of state-of-the-art generative AI models can be in the 10,000s

NVIDIA GH-200 Grace-Hopper Superpod

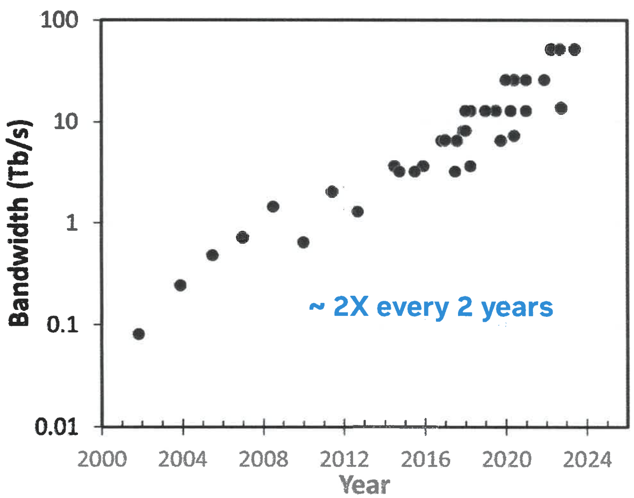
- Eight GH-200 NVL32 compute racks
- 32 Grace CPU / Hopper GPU Super-chips per rack
- 19.5 TB of NVL-addressable memory per rack

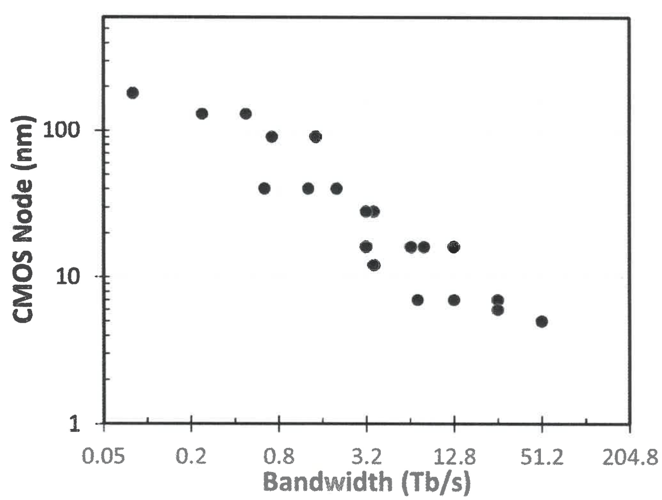
https://www.nvidia.com/en-us/data-center/grace-hopper-superchip/

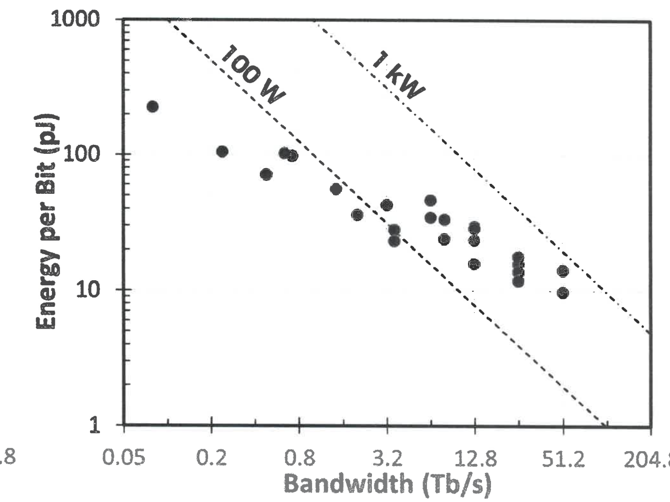
Switch ASIC Scaling

History & projections

Public data from commercial switch ASICs from a variety of vendors over the past 20 years:







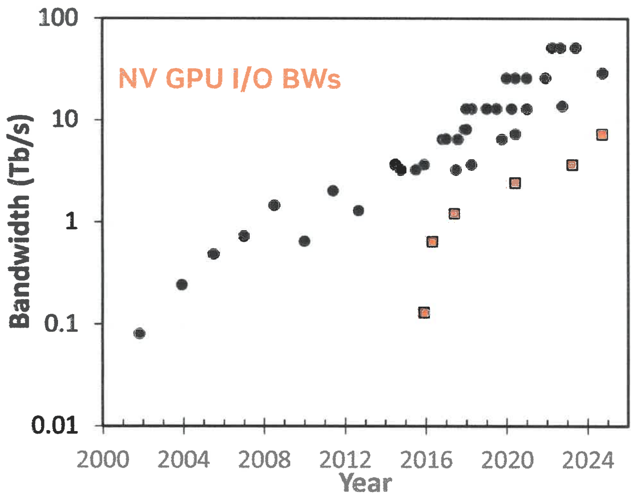
- Energy per bit has decreased due in part to CMOS scaling, but not fast enough to keep power from increasing.
- This is only expected to get worse as CMOS scaling slows.
- I/O power is scaling disproportionately to core power consumption.
- Need a low-power I/O solution, which can be adopted for both switches and GPUs.

All bandwidths are per direction

Switch ASIC Scaling

History & projections

Public data from commercial switch ASICs from a variety of vendors over the past 20 years:



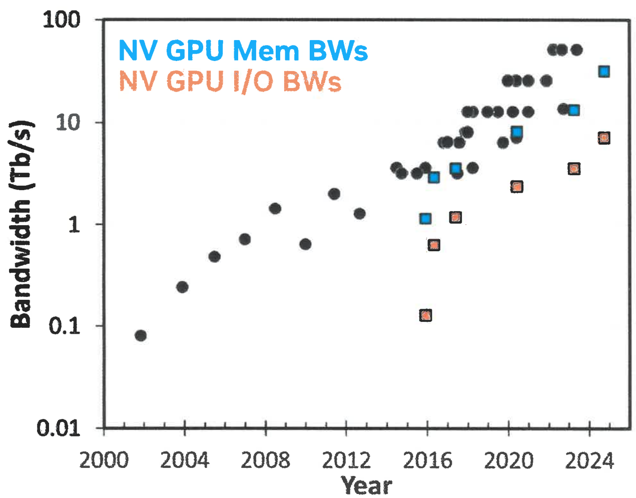
- GPU I/O BW
  - Not far behind switch BW (~ 10x). BW density even closer (single edge). Energy efficiency comparable to switch I/O.

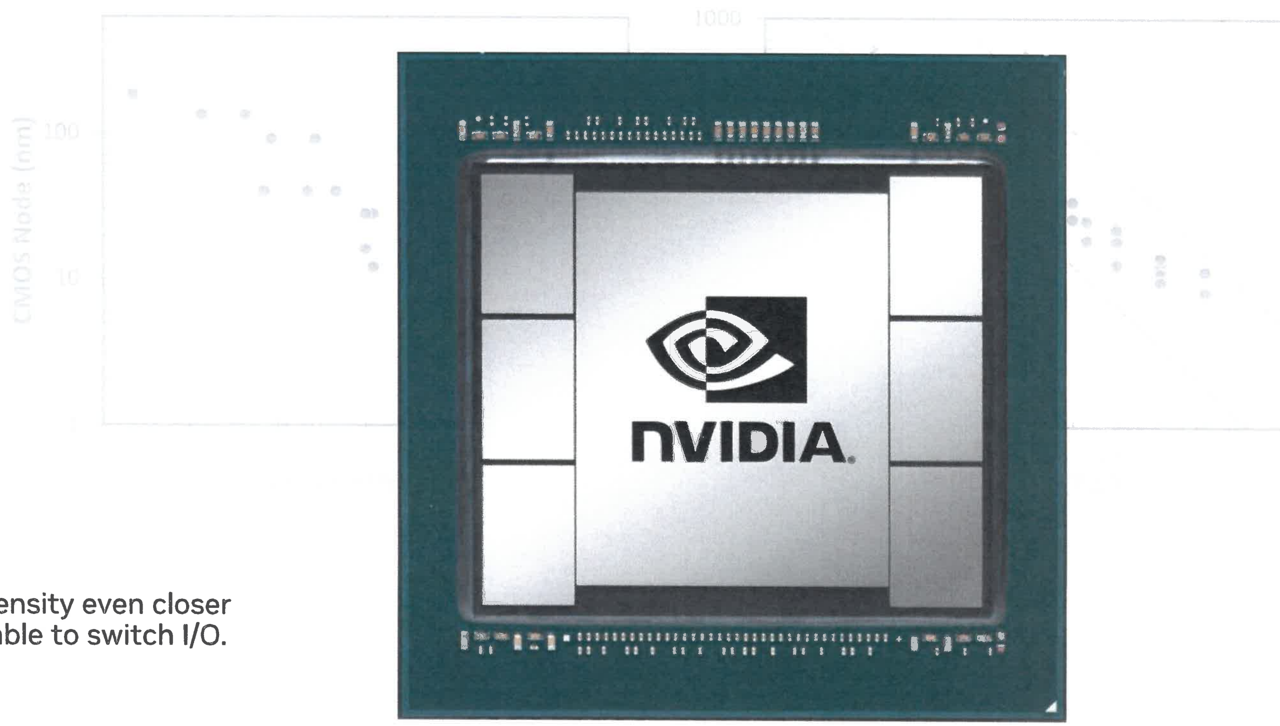
All bandwidths are per direction

Switch ASIC Scaling

History & projections

Public data from commercial switch ASICs from a variety of vendors over the past 20 years:





- GPU I/O BW
  - Not far behind switch BW (~ 10x). BW density even closer (single edge). Energy efficiency comparable to switch I/O.
- GPU Memory BW
  - Connects GPUs to memory on interposer.

All bandwidths are per direction

Electrical Interfaces

Power and density impact of the electrical interface to/from the optics

On-board or edge-of-card optics

LR interfaces over PCB [1-4]

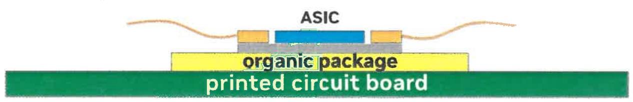
- Pin density requires 100+ Gb/s per trace
- 4.5 - 6.5 pJ/b for 112-Gb/s over ~ 0.5 m

Co-packaged optics

XSR interfaces on organic MCM [5-8]

- 1.2 - 1.7 pJ/b for 112-Gb/s over ~ 100 mm
- 500 - 900 Gb/s/mm

2.5D optics



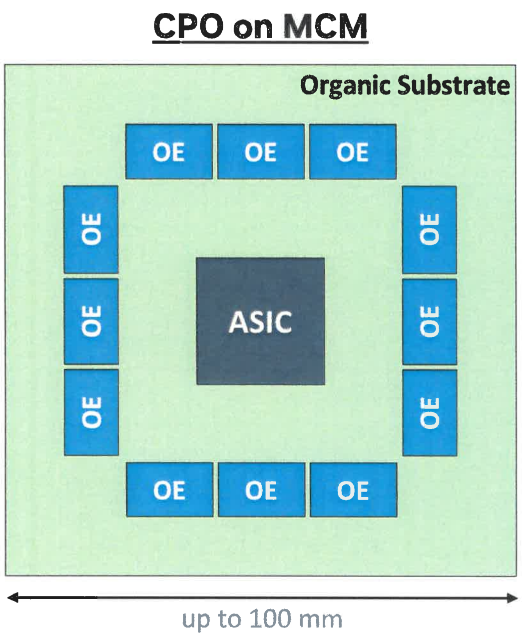
Slow and wide interfaces on interposer [9-10]

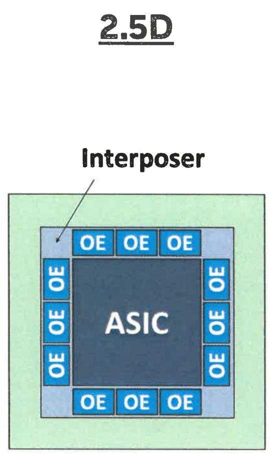
- Higher interconnect density → Lower serial rates → Better efficiency
- 0.2 - 0.3 pJ/b for 25 - 50 Gb/s over 1.2 mm
- 2 - 6 Tb/s/mm

[1] P. Mishra, *ISSCC 2021*, pp. 138-140.
[2] Z. Guo, *ISSCC 2022*, pp. 116-118.
[3] A. Varzaghani, *VLSI 2022*, paper C03-1.
[4] H. Park, *ISSCC 2023*, pp. 5-7.
[5] R. Shivnaraine, *ISSCC 2021*, p. 181-183.
[6] G. Gangasani, *ISSCC 2022*, pp. 122-124.
[7] C. F. Poon, *JSSC 2022*, vol. 57, no. 4, pp. 1199-1210.
[8] R. Yousry, *ISSCC 2021*, pp. 180-182.
[9] Y. Nishi, *JSSC 2023*, vol. 58, no. 4, pp. 1062-1073.
[10] Y. Nishi, *JSSC 2024*, early access.

Requirements for 2.5D-Integrated Optics

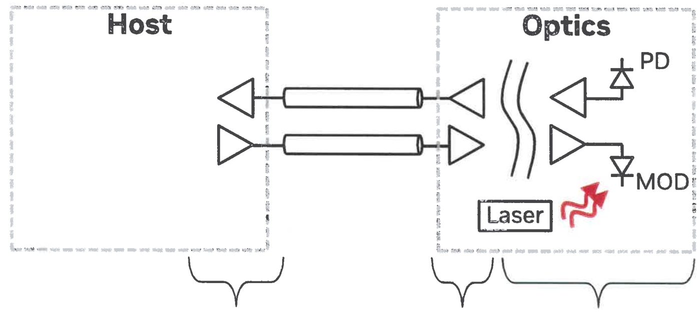
Bandwidth density and energy efficiency





- **No optical beachfront expansion**
  - High optical edge bandwidth densities (~ 2 + 2 Tb/s/mm)
  - Need dense optical connectors
- **Tighter integration**
  - Low-energy optics (~ 1 pJ/b)
  - Need efficient modulator, matched speeds
- **Limited size of interposers**
  - High optical areal bandwidth density
  - Need compact modulator

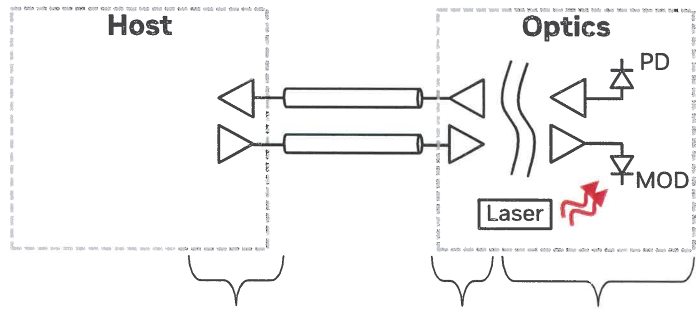
200-Tb/s Switch I/O Power Breakdown



	Energy (pJ/b)	Interface (Host)	Interface (Optics)	Optics	Laser
PCB/LR	5	1000 W	1000 W		
Pluggable optics	10			2000 W	

In-Package I/O Power	Total I/O Power
1000 W	4000 W

200-Tb/s Switch I/O Power Breakdown

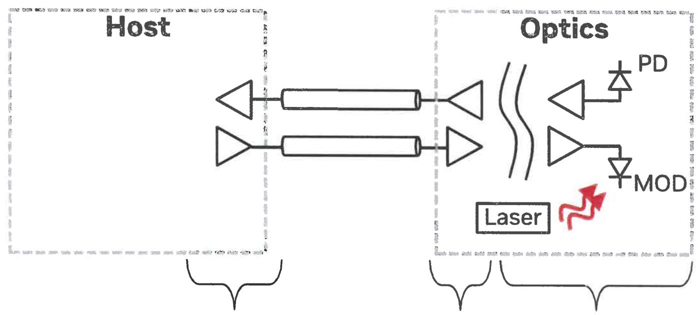


	Energy (pJ/b)	Interface (Host)	Interface (Optics)	Optics	Laser
PCB/LR	5	1000 W	1000 W		
Pluggable optics	10			2000 W	
MCM/XSR	1.5	300 W	300 W		
Co-packaged optics [1-4]	3+2			600 W	400 W

In-Package I/O Power	Total I/O Power
1000 W	4000 W
1200 W	1600 W

[1] C. Schulien, Hot Chips 2022, pp. 1-32.
[2] K. Muth, ECTC 2023, pp. 212-215.
[3] Levy, JSSC 2024, pp. 690-701.
[4] Wade, OFC 2021, pp. 1-3.

200-Tb/s Switch I/O Power Breakdown



	Energy (pJ/b)	Interface (Host)	Interface (Optics)	Optics	Laser
PCB/LR	5	1000 W	1000 W		
Pluggable optics	10			2000 W	
MCM/XSR	1.5	300 W	300 W		
Co-packaged optics [1-4]	3+2			600 W	400 W
Si interposer	0.25	50 W	50 W		
2.5D optics (*aspirational*)	1+2			200 W	400 W

In-Package I/O Power	Total I/O Power
1000 W	4000 W
1200 W	1600 W
300 W	700 W

[1] C. Schulien, Hot Chips 2022, pp. 1-32.
[2] K. Muth, ECTC 2023, pp. 212-215.
[3] Levy, JSSC 2024, pp. 690-701.
[4] Wade, OFC 2021, pp. 1-3.

Conclusions

- Networks are important for the future of AI systems
- Scaling I/O bandwidth is a critical challenge—at the switch and the GPU—calling for major changes to the interfaces.
- Co-packaged optics is poised to deliver a reduction in total I/O power.
- 2.5D optics can further improve efficiency, but it does place challenging constraints on the optics.

NVIDIA.

Optical Computer I/O

Intra-Data-Center Optics
Emerging Applications and Technology Trends
Optica ON2030 Webinar #2
26 June 2024

Chris Cole, Parallax Group

Outline

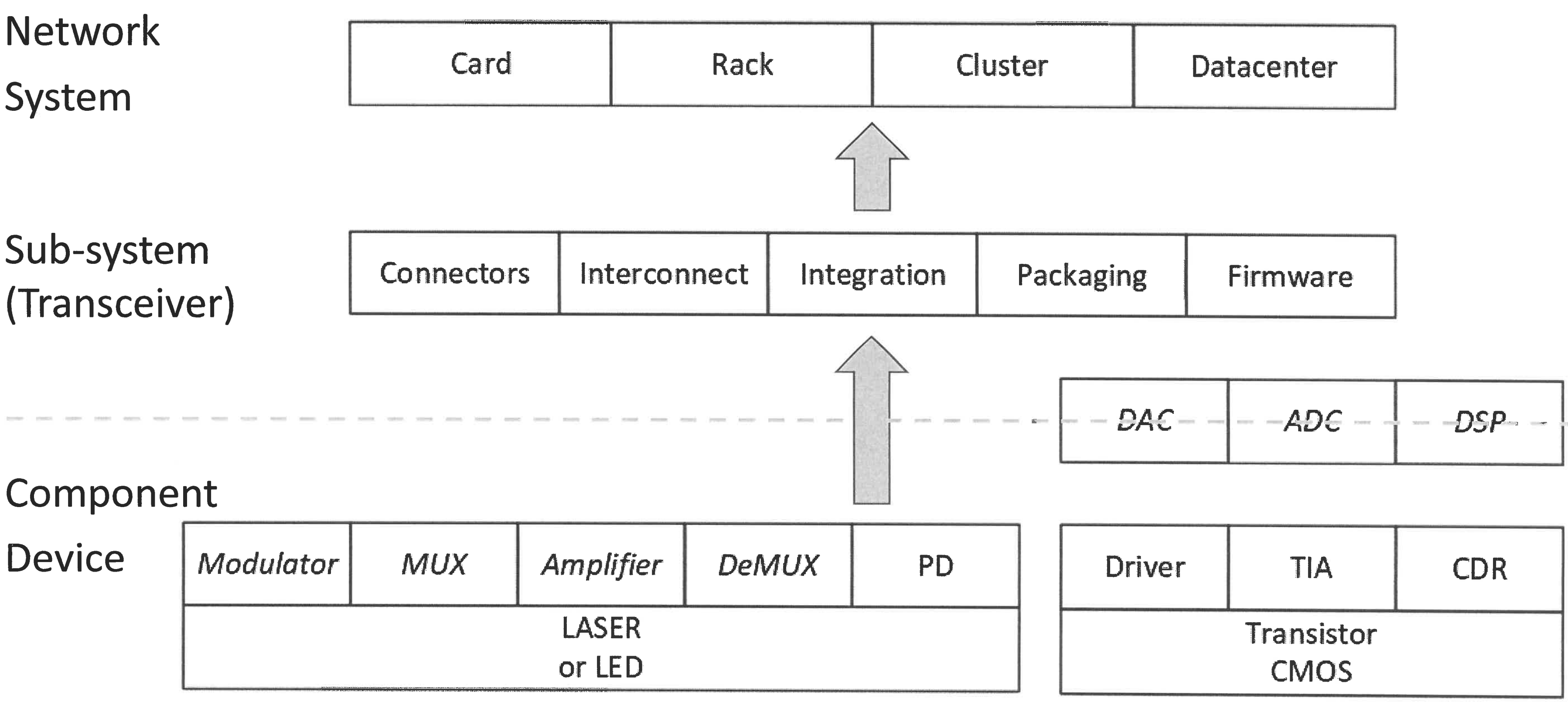
➢ Next Datacom Paradigm Shift

- Optical PCIe
- Half-retimed Optics

Introduction

- Next Datacom Paradigm shift:
  - Optical Computer I/O driven by AI/ML
- Optical Computer I/O requirements:
  - Order(s) of magnitude more stringent than Optical Networking
  - Only met with fundamentally new optical Components and Devices
- Datacom Optics investment priority:
  - Sub-systems and Systems
  - Rearranging and/or aggregating existing technology
- This is in sharp contrast to electronics, which benefit from huge CMOS investment
  - ex. CHIP ACT and matching industry investment: ~$150B

Datacom Optics Hierarchy



Datacom Paradigm Shift Enabling Optical Technologies

Datacenter Paradigm	Network or Computer Link Rate	No. of Lanes	Enabling Component & Device Technology	Enabling Sub-system Technology
Enterprise	100M (ex. Ethernet) 1G 10G	1	VCSEL DFB LASER	LC (Lucent Connector) Pluggable Module
Hyperscale	40G 25/50G 100G	4	EML WDM Si MZM	MT Parallel Connector
	200G 400G 800G	4, 8	DSP	Heatsink 
AI/ML	1T (ex. PCIe) 2T $\geq$4T	$\geq$ 16	Hi-Rel LASER (or LED) DWDM Dense BW Modulator	Dense BW Connector & Packaging

Outline

- Next Datacom Paradigm Shift
- ➢ Optical PCIe
- Half-retimed Optics

Optical PCIe Proposal: External Connectors

- QSFP-DD (x4/x8)
- OSFP-XD (x8/x16)
- CDFP (x4/x8/x16)
- Proposal includes CEM/M.2 with NPO and CPO
- External connector pins to be defined as compliance points, or compatible with PCI-SIG compliance points
- Pin-maps will be standardized
- In-band or out-off band control signaling alternatives

Optical PCIe Proposal: Cabling Configurations

- Implementations not specified (AOC model)
- ECN for PCIe Compliant (Re-timed)
  - "2-retimer" easily compliant
  - "1-retimer" links easily meet link requirements and under logic study
- ECN for Engineered (Not-retimed) Optical Links
  - feasible, but place limitations on compatibility and compliance
- PCIe 6 to 8 optical links meet BER < 1e-7 with robust margin
- PCIe 1 to 5 optical links meet BER < 1e-13 with robust margin
- Worst-case optical link skew is significantly under PCI-SIG limits

Optical PCIe Proposal: Management

- Module advertisement scheme(s) for key implementation attributes
- Implementation(s) not limited by specific form factor, link type, or optical technology
- Framework to build upon collaboration with OIF to utilize CMIS
- Examples of implementation attributes to be advertised
  - Total latency
  - Physical length
  - PCIe Compliant (Re-timed) or Engineered (Not-retimed)
  - For Re-timed; number of logical re-timers

Outline

- Next Datacom Paradigm Shift
- Optical PCIe

➤ **Half-retimed Optics**

Transmit Re-timed Optics (TRO)

- LPO (Linear Pluggable Optics) is Linear Tx Linear Rx
- It has generated a lot of industry interest
  - OIF CEI-112 & 224 Projects
  - Potential for cost, power, latency savings by eliminating Tx and Rx DSP
- Drawback of LPO is two concatenated Cu links
- Retimed Tx Linear Rx (RTLR) cost and power savings similar to LPO
- Advantage of RTLR is Cu link isolation by DSP
- LPO and RTLR address different applications
  - If LPO works, there is no need for RTLR
  - If LPO does not work, then RTLR is a candidate

Adopted OIF RTLR Spec. Objectives

- Full IEEE 802.3 electrical and optical plug-and-play
- 100G/lane
    - 500m SMF DRn
    - 2km SMF FR4
    - 30m MMF SRn link
- 200G/lane
    - 500m SMF DRn
    - 500m SMF FR4

OIF Project Timeline

- Start
  - Discussion in 2023
  - Extensive industry support including major End Users
  - Unanimously approved RTxLRx Project at the OIF Q1 meeting (Jan'24)
- Formalization
  - Objectives expanded
  - Name changed to RTLR
  - Adopted at the OIF Q2 meeting (Apr'24)
- 100G/lane 1st Spec
  - Proposed Jun'24
  - Baseline adoption target: Q3'24 meeting (Aug'24)
- 200G/lane 1st Spec
  - Baseline adoption target: Q4'24 meeting (Nov'24)

Optical Computer I/O

Thank you