Photonic Integrated Subsystems for next Generation Leadership Class HPC

Keren Bergman

Lightwave Research Lab, Columbia University

New York, NY, USA

U.S. DEPARTMENT OF ENERGY | Office of Science

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

High Performance Systems: Trends and Challenges

- SUMMIT (Oak Ridge National Laboratory)
  - Most powerful supercomputer* (June, 2018)
  - Peak performance: 122.3 PetaFLOPS (Linpack)
  - Data Analytics applications up to 3.3 ExaFLOPs

  - Power consumption: 13MW
    - Power efficiency: 13.9 GFLOPs/Watt (#5 Green 500)
  - 4608 Nodes with:
    - 200 G (Dual-rail Mellanox EDR 100G InfiniBand)
    - 9216 IBM Power9 CPUs (2 per node)
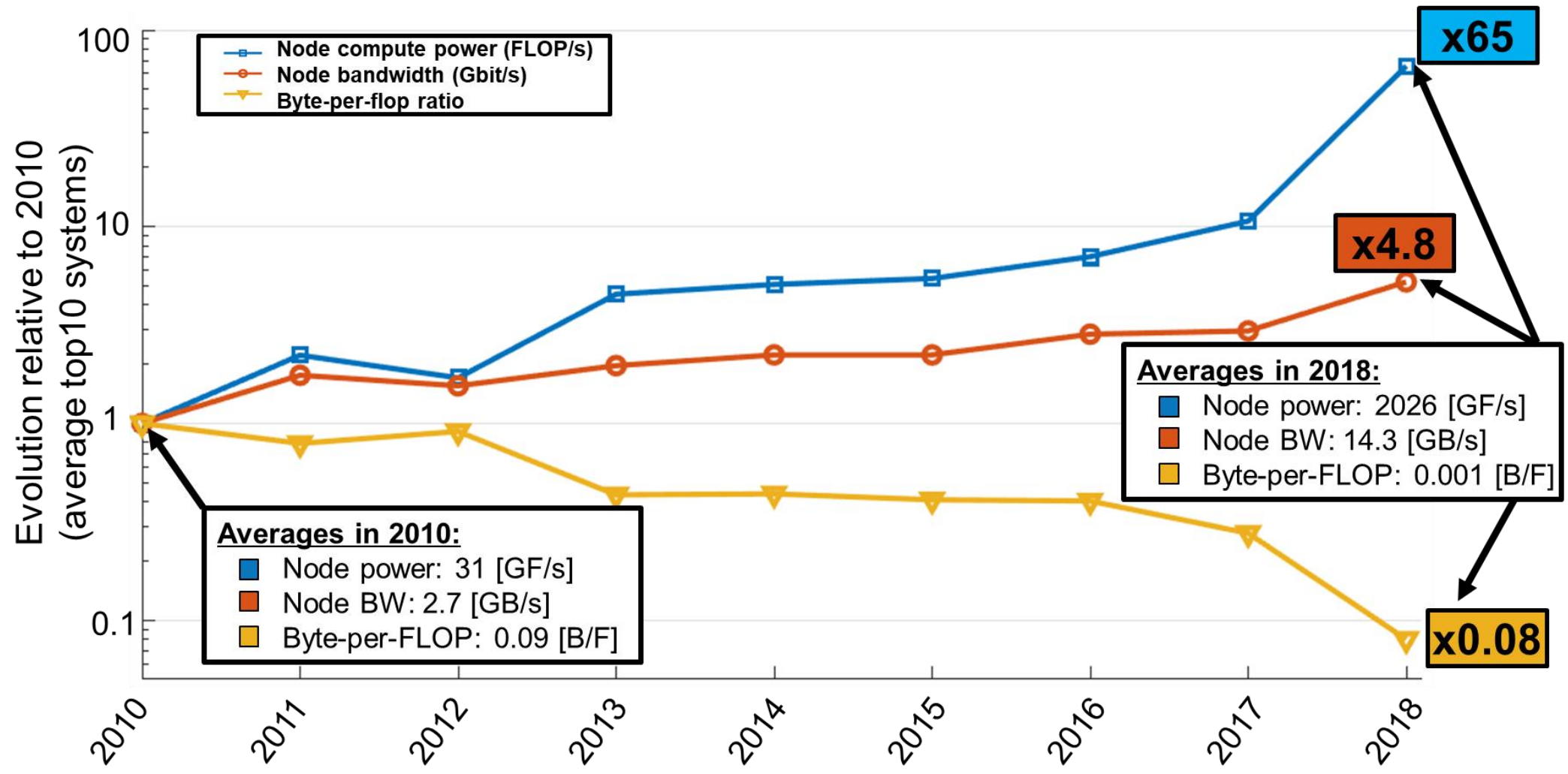   - 27648 Nvidia Volta V100 GPUs (6 per node)

Next challenge:

Reach Exascale+ within 20MW → 50 GFLOPs/watt

Source: www.olcf.ornl.gov/summit/

* at executing High Performance Linpack (top500.org results)

Performance/Communications Trends for Top 10 (2010-2018)

Sunway TaihuLight (Nov 2017) B/F = 0.004; Summit HPC (June 2018) B/F = 0.0005 → **8X decrease**

Performance and the Data Movement Energy Budget

- GFLOPs/Watt = GFlop/second / Joule/second = GFlop/Joule

- 14 GFLOPs/W (Summit)   ⇔ 72 pJ/FLOP

- **Target: 50 GFLOPs/W**   ⇔ 20 pJ/FLOP

- Energy per bit **total budget** (200 bits/FLOP):

14 GFLOPs/W:	72 pJ/FLOP	**0.36 pJ/bit**
50 GFLOPs/W:	20 pJ/FLOP	**0.1 pJ/bit**

| Data Movement Energy:	
– Access SRAM	O(10fJ/bit)
– Access DRAM cell	O(1 pJ/bit)
– Movement to HBM/MCDRAM (few mm)	O(10 pJ/bit)
– Movement to DDR3 off-chip (few cm)	O(100 pJ/bit)

- Scaling performance under ultra-tight energy budget:
  - Raise cache hit rates (expanded caches, more reuse)
  - Improve memory access (read, write) energy efficiency
  - **Improve data movement energy efficiency:**
    - **Novel interconnect technologies and architectures**

Top 500 and "Green 500"

June 2016

Name	Top500 rank	GFlop/W
Shoubu	94	6.7
Satsuki	486	6.2
Sunway TL	1	6.1

November 2016

Name	Top500 rank	GFlop/W
DGX Sat.V	28	9.5
Piz Daint	8	7.5
Shoubu	116	6.7
Sunway TL	1	6.1

Legend:
- Zettascaler 1.6
- Zettascaler 2.0
- Zettascaler 2.2
- Tesla P100
- DGX-1 station + P100
- DGX-1 station + V100
- Zettascaler 1.6 + Tesla P100

June 2017

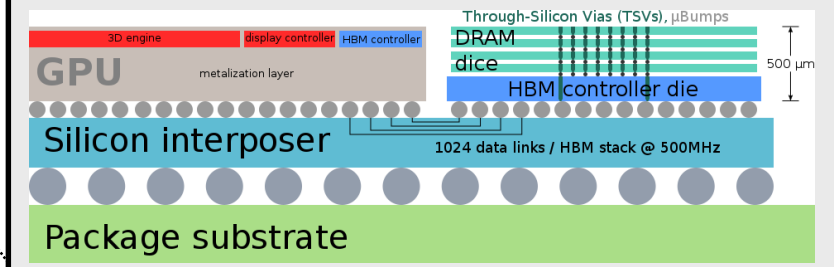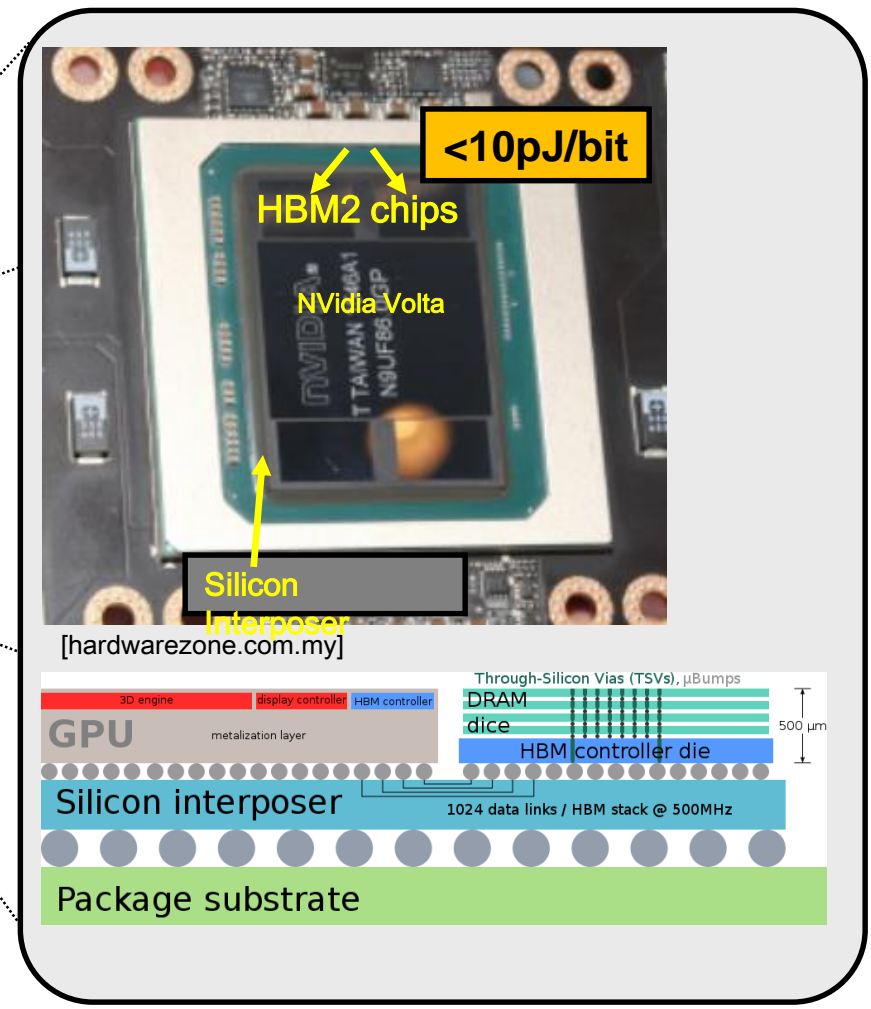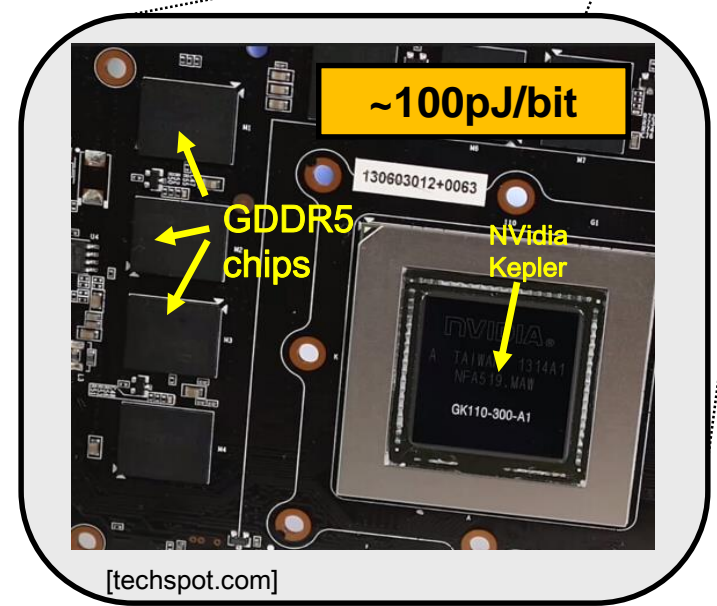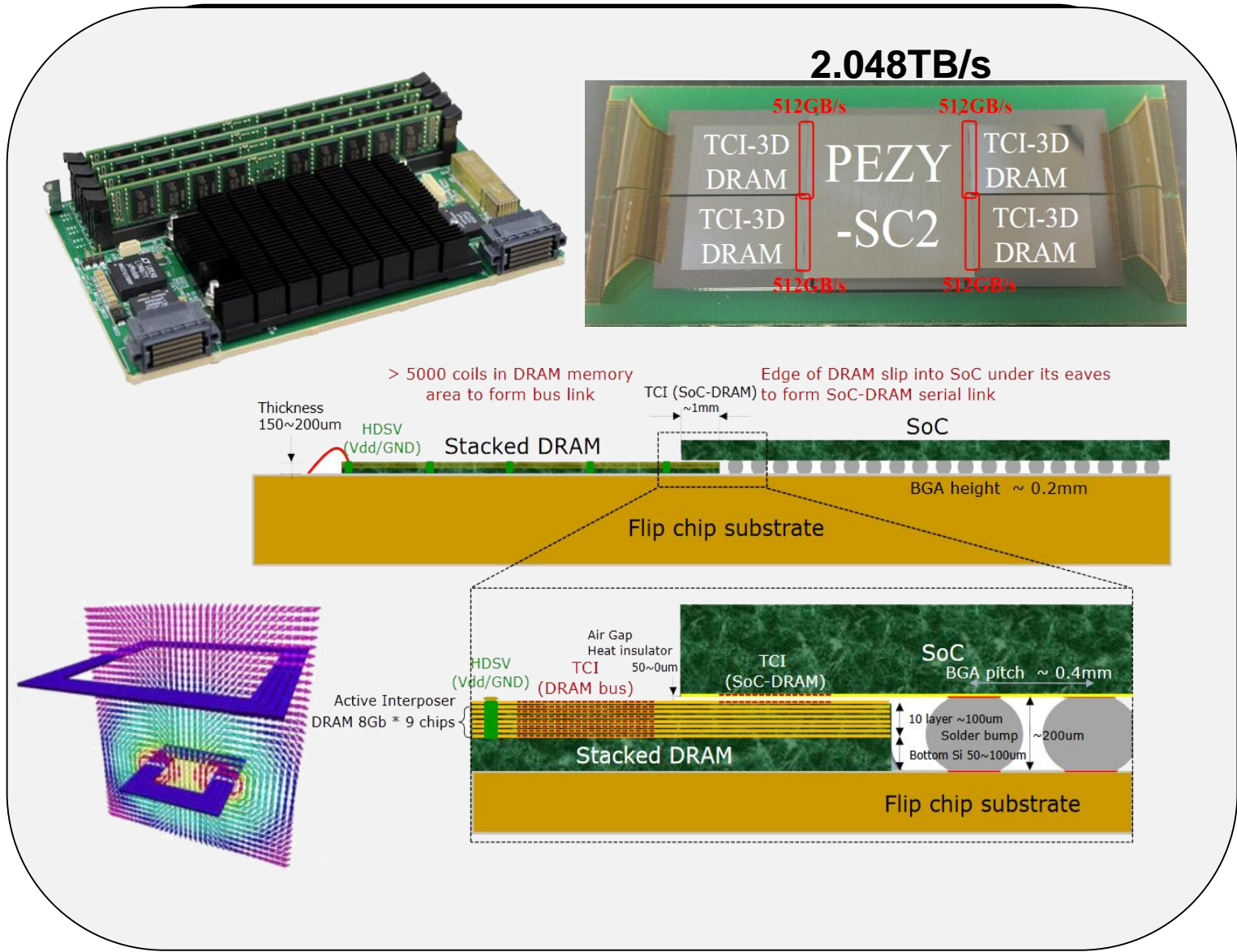| Name	Top500 rank	Gflop/W
TSUBAME3.0	61	14.1
kukai	465	14.0
AIST AI Cloud	148	12.7
RAIDEN	305	10.6
Wilkes-2	100	10.4
Piz Daint	3	10.4
Gyoukou	69	10.2
GOSAT-2	220	9.8
	31	9.5
DGX Sat.V	32	9.5
Reedbush-H	203	8.6
JADE	425	8.4
Cedar	86	8.0
DAVIDE	299	7.7
Shoubu	137	6.7
Hokule'a	466	6.7
Sunway TL	1	6.1

November 2017

Name	Top500 rank	GFlop/W
Shoubu B	259	17.0
Suiren2	307	16.8
Sakura	276	16.7
DGX Volta	149	15.1
Gyoukou	4	14.2
TSUBAME3.0	13	13.7
AIST AI Cloud	195	12.7
RAIDEN	419	10.6
Wilkes-2	115	10.4
Piz Daint	3	10.4
Reedbush-L	291	10.2
GOSAT-2	319	9.8
	35	9.5
DGX Saturn V	36	9.5
Era-AI	109	8.6
Reedbush-H	295	8.6
Cedar	94	8.0
DAVIDE	440	7.9
Shoubu	180	6.7
Sunway TL	1	6.1

June 2018

Name	Top 500	Gflop/W
Shoubu B	359	18.4
Suiren2	419	16.8
Sakura	385	16.7
DGX Volta	227	15.1
Summit	1	13.9
TSUBAME3.0	19	13.7
AIST AI Cloud	287	12.7
Sunway TL	2	6.1 (#23)

NVidia's GPU/memory Integration Assembly

June 2016		
Name	Top500 rank	GFlop/W
Shoubu	94	6.7
Satsuki	486	6.2
Sunway TL	1	6.1
GSI /ASUS	440	5.3
Sugon+K80	446	4.8

June 2017		
Name	Top500 rank	GFlop/W
TSUBAME3.0	61	14.1
kukai	465	14.0
AIST AI Cloud	148	12.7
RAIDEN	305	10.6
Wilkes-2	100	10.4
Piz Daint	3	10.4

<10pJ/bit
HBM2 chips
NVidia Volta
Silicon Interposer
[hardwarezone.com.my]

~100pJ/bit
GDDR5 chips
NVidia Kepler
130603012+0063
GK110-300-A1
[techspot.com]

NVidia major new design

Through-Silicon Vias (TSVs), µBumps
3D engine   display controller   HBM controller
GPU   metalization layer
DRAM dice
HBM controller die
500 µm
Silicon interposer   1024 data links / HBM stack @ 500MHz
Package substrate

- **Memory closer to GPU**
- CoWoS: Chip on wafer on Substrate

ZettaScaler 2.2

November 2017		
Name	Top500 rank	GFlop/W
Shoubu B	259	17.0
Suiren2	307	16.8
Sakura	276	16.7

- ZettaScaler architecture:
  - Modular design
  - Liquid cooled
  - ThruChip Interface (TCI) with sub-pJ/bit efficiency

Architectures big gains in GFlops/Watt: Innovative Data Movement Solutions

2.048TB/s

512GB/s   512GB/s
TCI-3D DRAM   PEZY   TCI-3D DRAM
TCI-3D DRAM   -SC2   TCI-3D DRAM
512GB/s   512GB/s

> 5000 coils in DRAM memory area to form bus link

Edge of DRAM slip into SoC under its eaves to form SoC-DRAM serial link

TCI (SoC-DRAM) ~1mm

Thickness 150~200um   HDSV (Vdd/GND)   Stacked DRAM   SoC

BGA height ~ 0.2mm

Flip chip substrate

Air Gap Heat insulator 50~0um

HDSV (Vdd/GND)   TCI (DRAM bus)   TCI (SoC-DRAM)   SoC   BGA pitch ~ 0.4mm

Active Interposer DRAM 8Gb * 9 chips   Stacked DRAM   10 layer ~100um   Solder bump   ~200um   Bottom Si 50~100um

Flip chip substrate

High Performance Data Centers: Convergence on AI

- Strong interest in **energy efficiency** of Data Centers on AI

- …And not only for "small" systems
  - Training Deep Neural Networks (DNN) takes time!
    - "*Our network takes between five and six days to train on* **two** *GTX 580 3GB* **GPUs**" (Krizhevsky et al., 2012)
    - "*On a system equipped with* **four** *NVIDIA Titan Black* **GPUs**, *training a single net took 2–3 weeks*" (Simonyan et al., 2015)
    - "*our […] system trains ResNet-50 […] on* **256 GPUs** *in one hour*" (Goyal et al., 2017)

- Facebook and NVidia's clusters have 1,000 GPUs (3.3 PFlops)

June 2017		
Name	Top500 rank	GFlop/W
TSUBAME3.0	61	14.1
kukai	465	14.0
AIST AI Cloud	148	12.7
RAIDEN	305	10.6
Wilkes-2	100	10.4
Piz Daint	3	10.4
Gyoukou	69	10.2
GOSAT-2	220	9.8
Facebook	31	9.5
DGX Sat.V	32	9.5
Reedbush-H	203	8.6
JADE	425	8.4
Cedar	86	8.0
DAVIDE	299	7.7
Shoubu	137	6.7
Hokule'a	466	6.7
Sunway TL	1	6.1

November 2017		
Name	Top500 rank	GFlop/W
Shoubu B	259	17.0
Suiren2	307	16.8
Sakura	276	16.7
DGX Volta	149	15.1
Gyoukou	4	14.2
TSUBAME3.0	13	13.7
AIST AI Cloud	195	12.7
RAIDEN	419	10.6
Wilkes-2	115	10.4
Piz Daint	3	10.4
Reedbush-L	291	10.2
GOSAT-2	319	9.8
Facebook	35	9.5
DGX Saturn V	36	9.5
Era-AI	109	8.6
Reedbush-H	295	8.6
Cedar	94	8.0
DAVIDE	440	7.9
Shoubu	180	6.7
Sunway TL	1	6.1

June 2018		
Name	Top 500	Gflop/W
Shoubu B	359	18.4
Suiren2	419	16.8
Sakura	385	16.7
DGX Volta	227	15.1
Summit	1	13.9
TSUBAME3.0	19	13.7
AIST AI Cloud	287	12.7
Sunway TL	2	6.1 (#23)

8

Scaling chip 'escape' bandwidth density

NVIDIA NVSwitch

~22mm
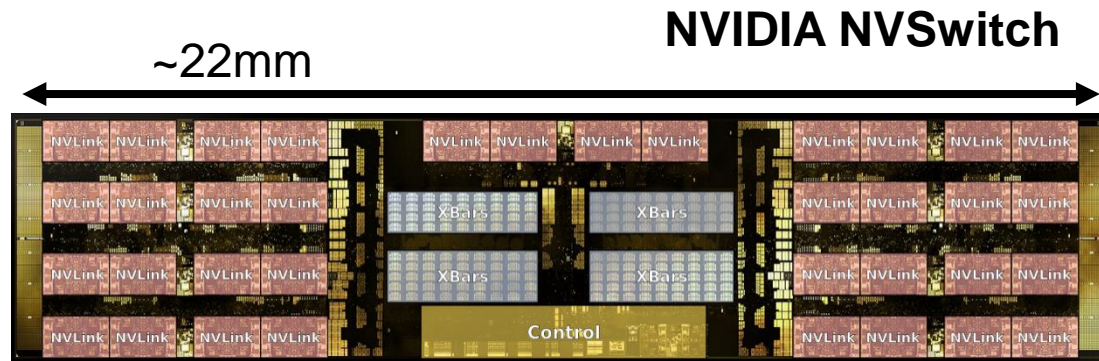
- 18 NVLink 2.0 ports → 9 per long edge top/bottom
- 50GB/s per port (25GB/s each Tx/Rx)
- 1 NVLink ~ 2mm of linear edge
- 50GB/s per 2mm → <u>200Gb/s/mm</u>

Each fiber 'pin' carries
64 λs (scalable)

Dense WDM Silicon Photonic:

- 250um fiber pitch
- 8 fiber links ~ over 2mm linear edge
- 64 λs per fiber link; each λ at 16Gb/s = 1 Tb/s per link
- 8 Tb/s per 2mm → <u>4Tb/s/mm</u>

Fiber pitch
250um
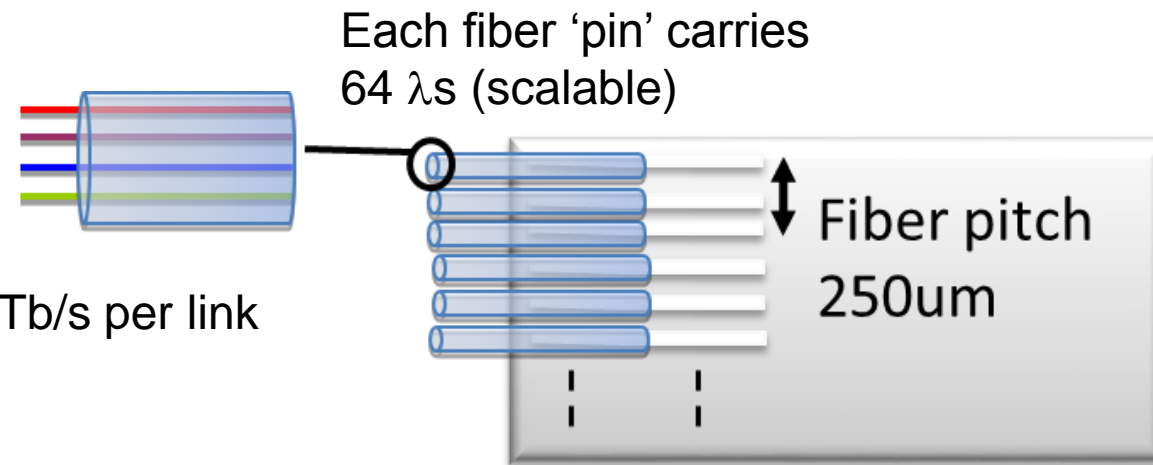
The Photonic Opportunity for Data Movement
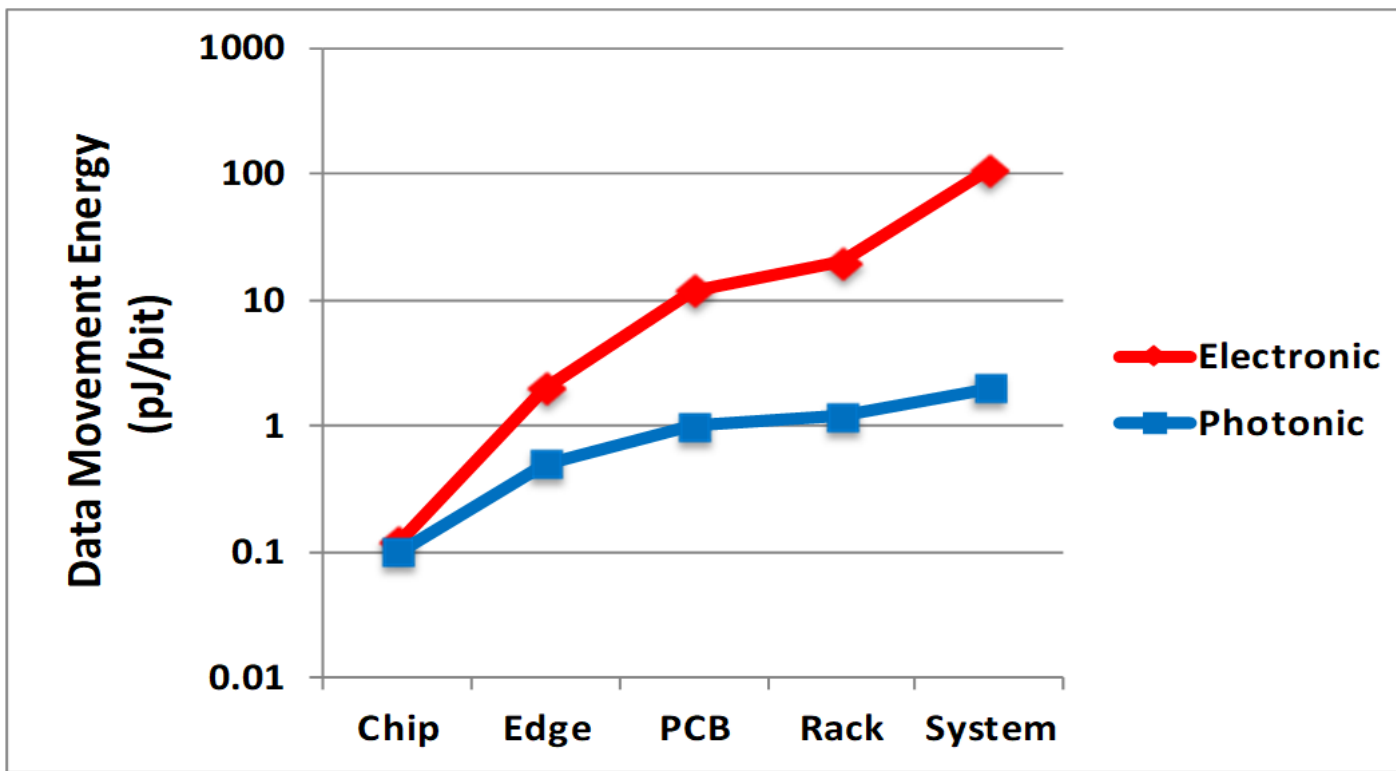
Reduce Energy Consumption
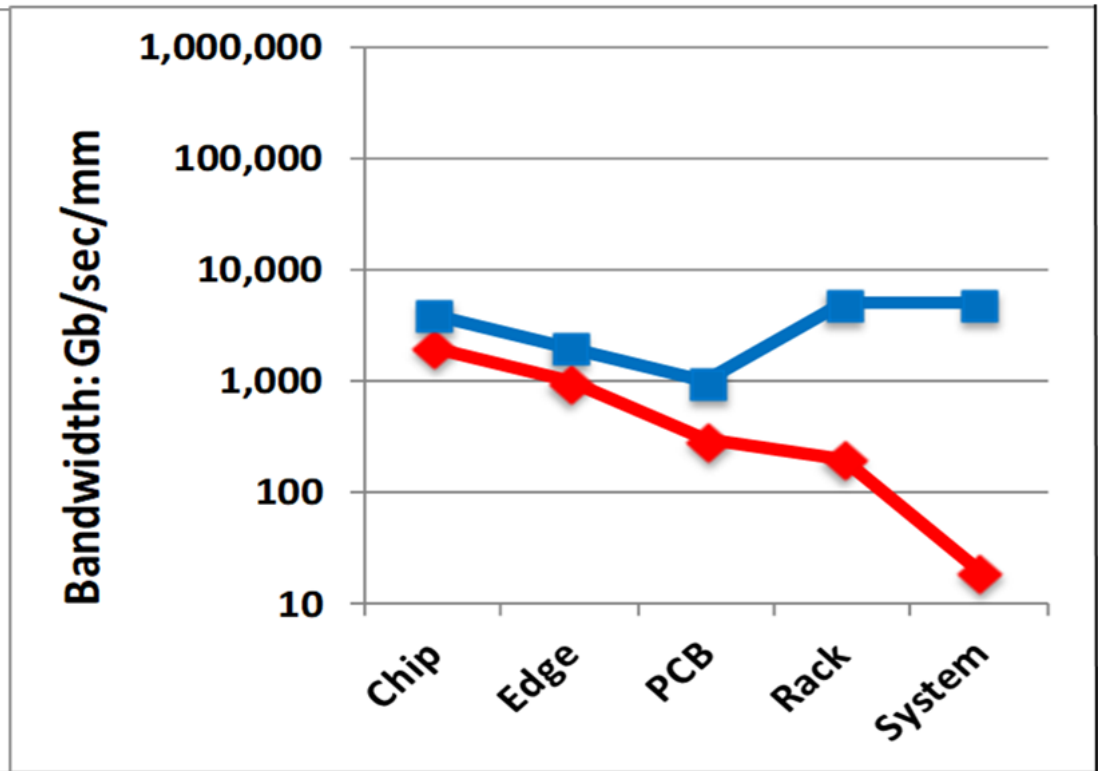
Eliminate Bandwidth Taper

R. Lucas et al., "Top ten exascale research challenges," DOE ASCAC subcommittee Report, 2014

Silicon Photonics Dense-WDM
Scalable, >Tb/s/mm, <1pJ/bit "distance transparent" Optical Interconnect
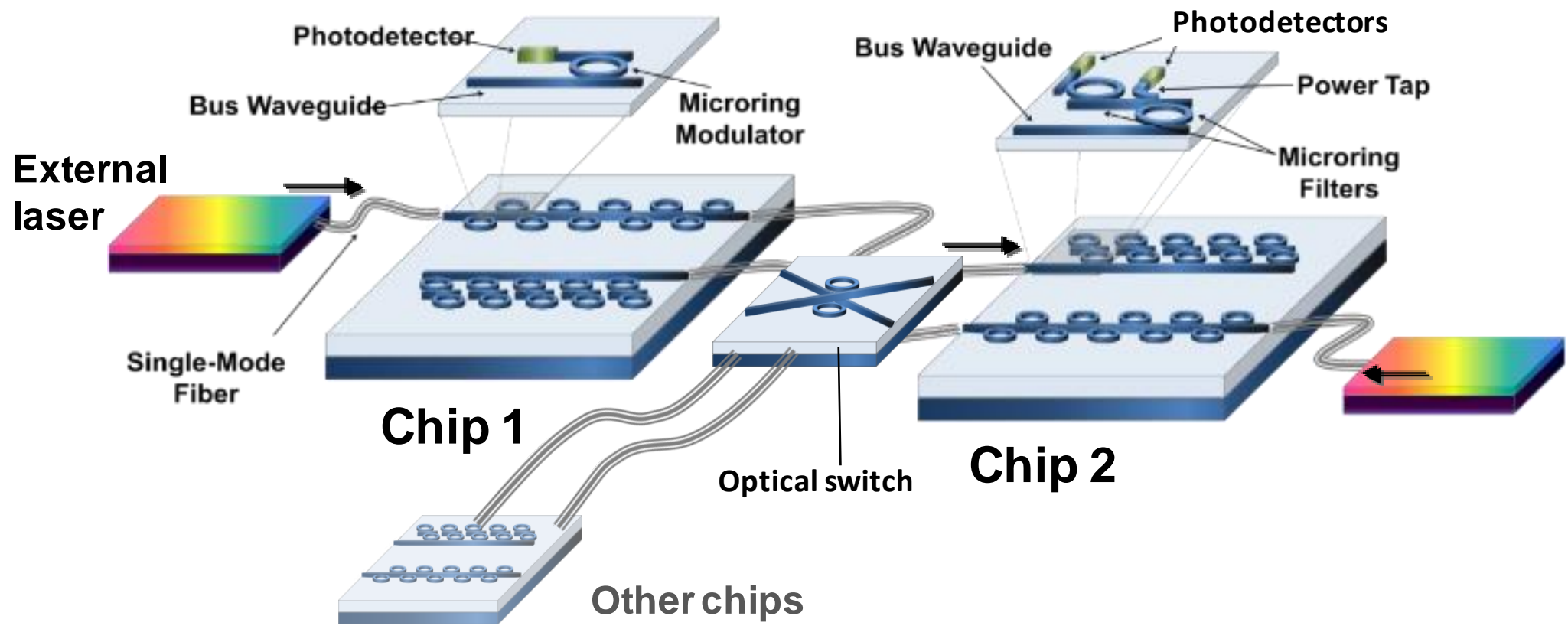
Only "Power Up" Needed Optical Links: Disaggregated Architecture

Current server
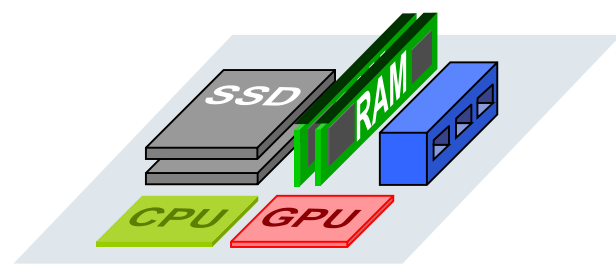
Current rack
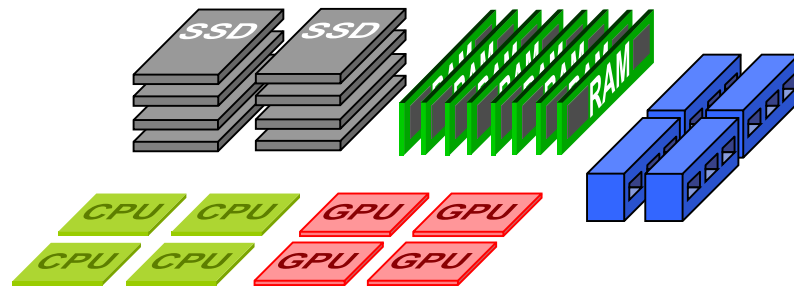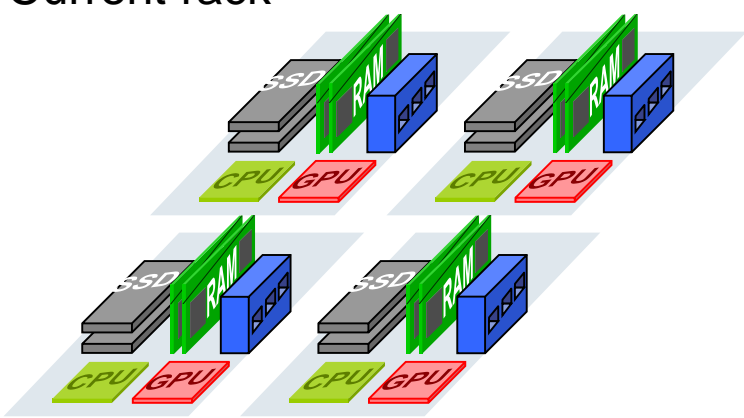
Disaggregated rack
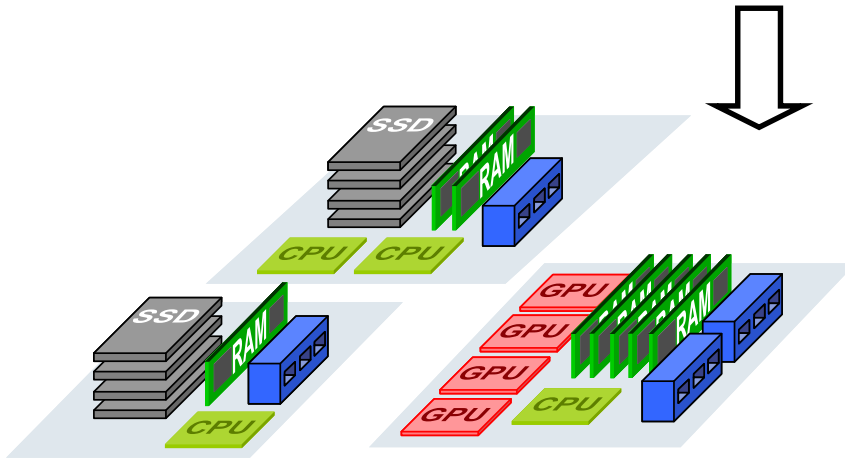
Pool and compose

However… Inter Node Bandwidth (10 GB/s) << needed Embedded Bandwidth (100 GB/s – 1 TB/s)

Disaggregated System Architecture: flexibly interconnected heterogeneous resources

U.S. DEPARTMENT OF **ENERGY**
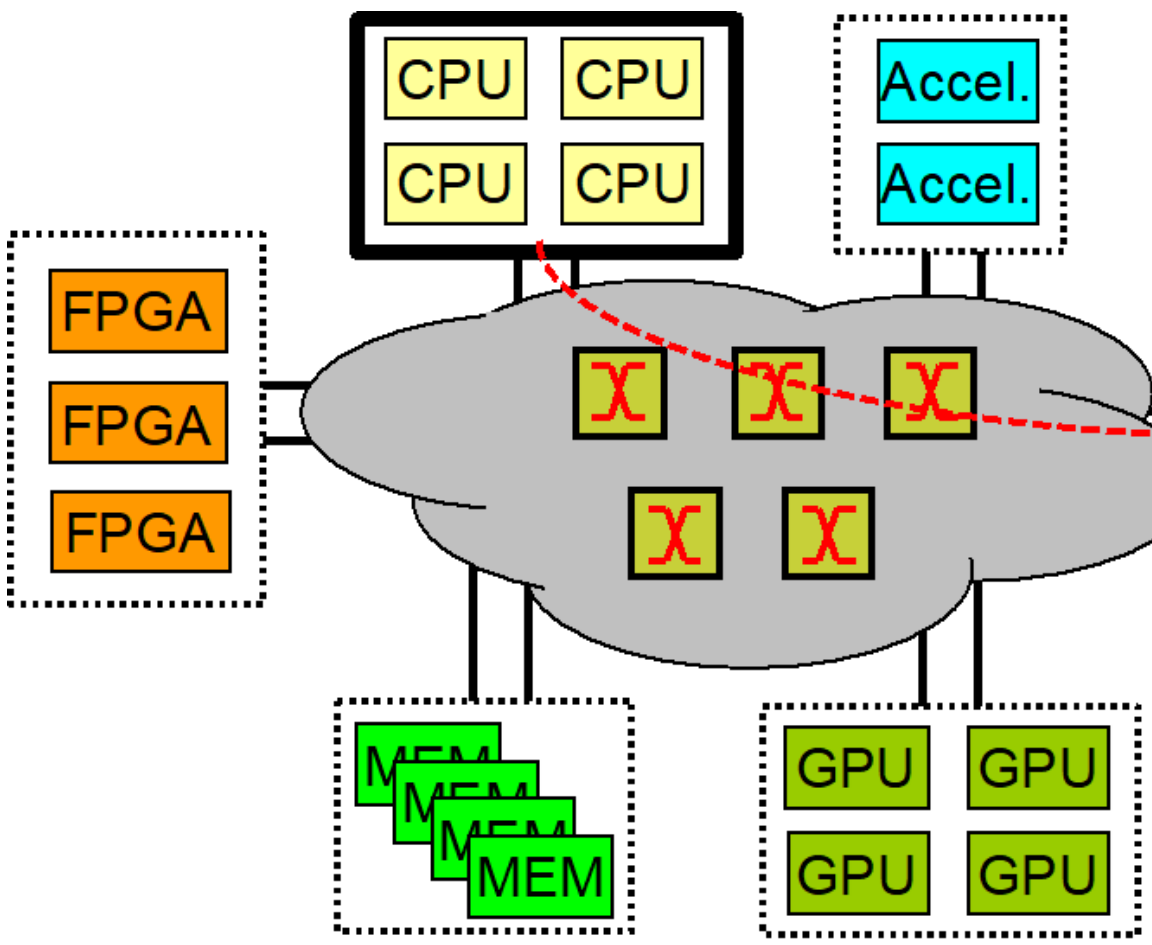
SBIR Collaborative Project (R. Carlson):
Photonic-Storage Subsystem Input/Output (P-SSIO)
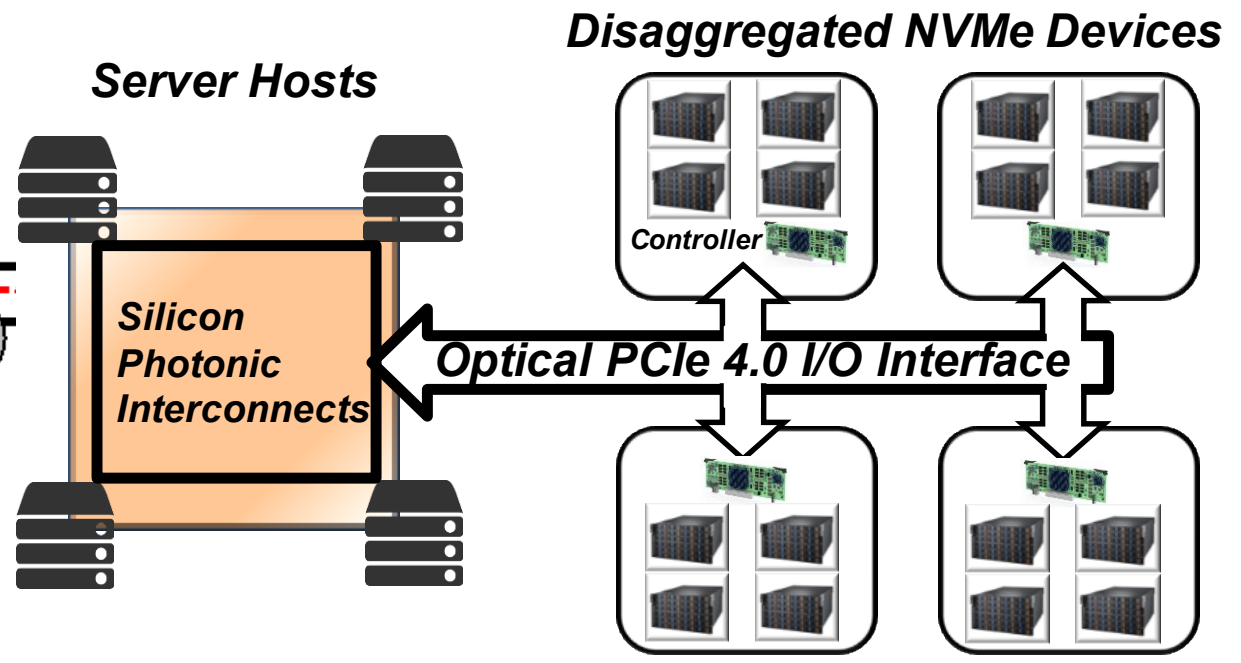
Prototyping photonic IO to storage subsystems and software modules for HPC

Multi-Host/Storage Architecture with Photonic I/Os
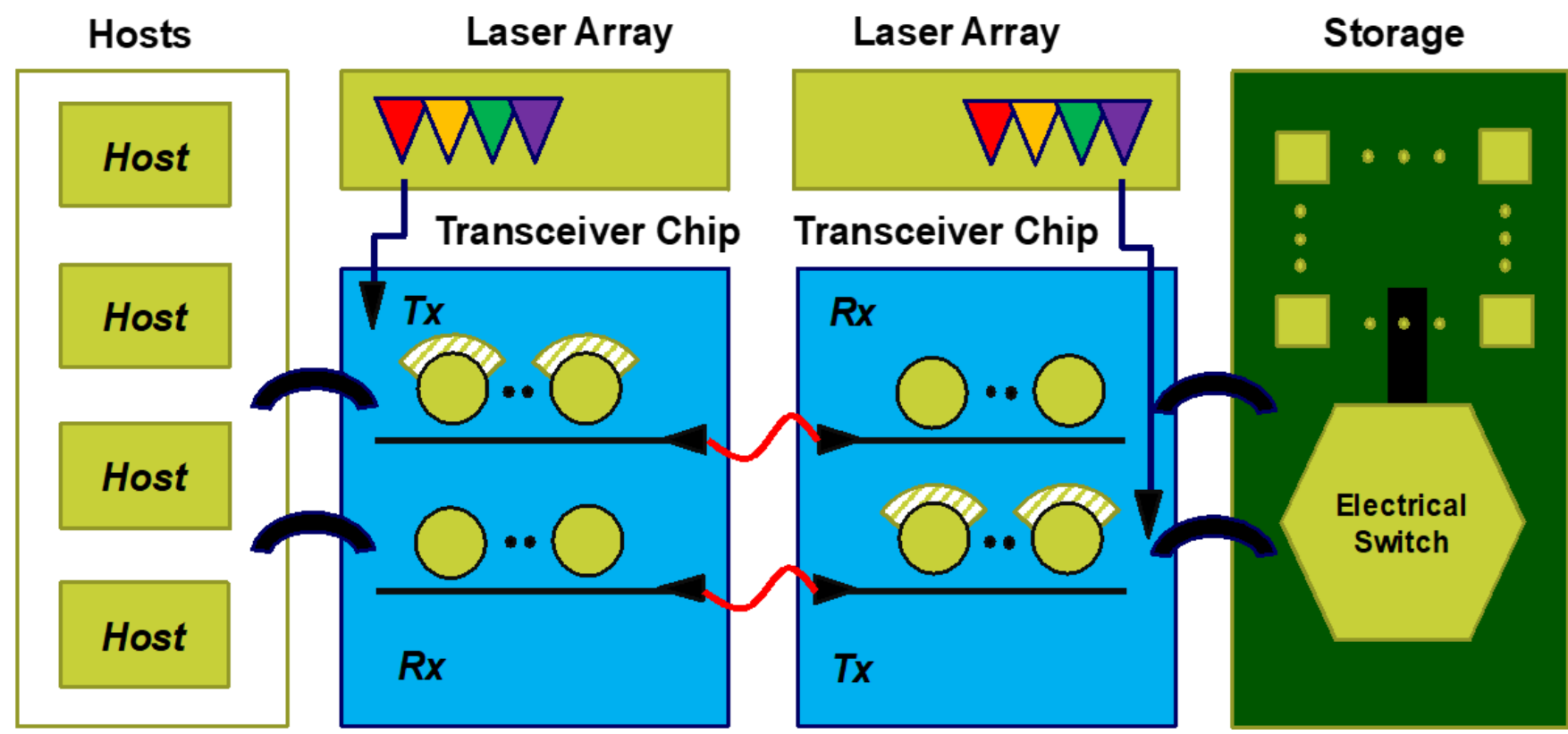
ASCR/SBIR Collaborative Project (R. Carlson): Photonic-Storage Subsystem Input/Output (P-SSIO)

Objectives:

- Energy efficient integrated photonic I/O (0.5pJ/b)

- High bandwidth throughput (256GB/s)

P-SSIO System performance goals

- 4-8 Server class PCIe version 4.0 x 32 controller chips (CPU or dedicated controller)

- 16-32 Non-Volatile Memory Express (NVMe) based Storage Subsystems connected at 16 GT/s I/O rate each

- Simultaneous access from every PCIe controller to multiple NVMe storage devices (256 GB/s aggregate I/O rate with 4 PCIe controllers)

- WDM optical transceivers matched to the PCIe I/O v 4.0 transmission rates

- Reconfigurable optical interconnect fabric

- Low loss Optical connectors and/or integrated Micro Optical Bench assemblies
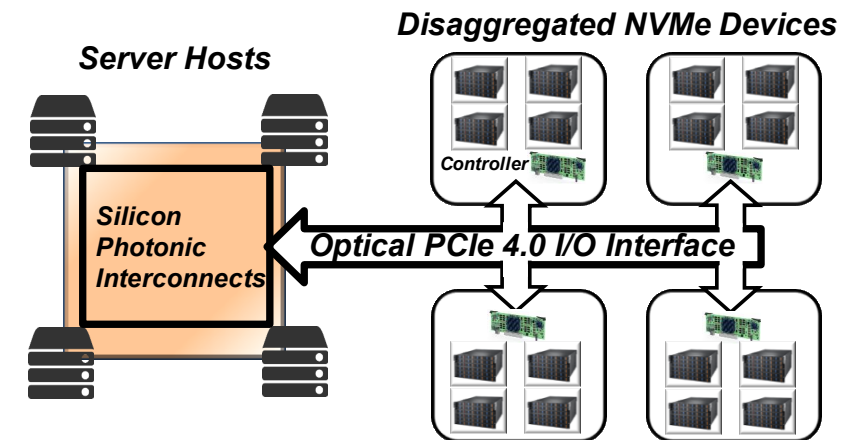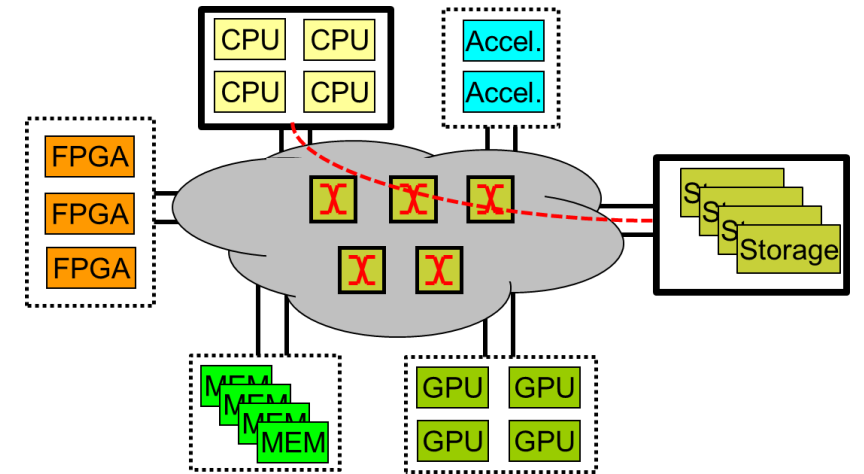
Partners:

Disaggregation: Deeper into the Heterogeneous Hierarchy
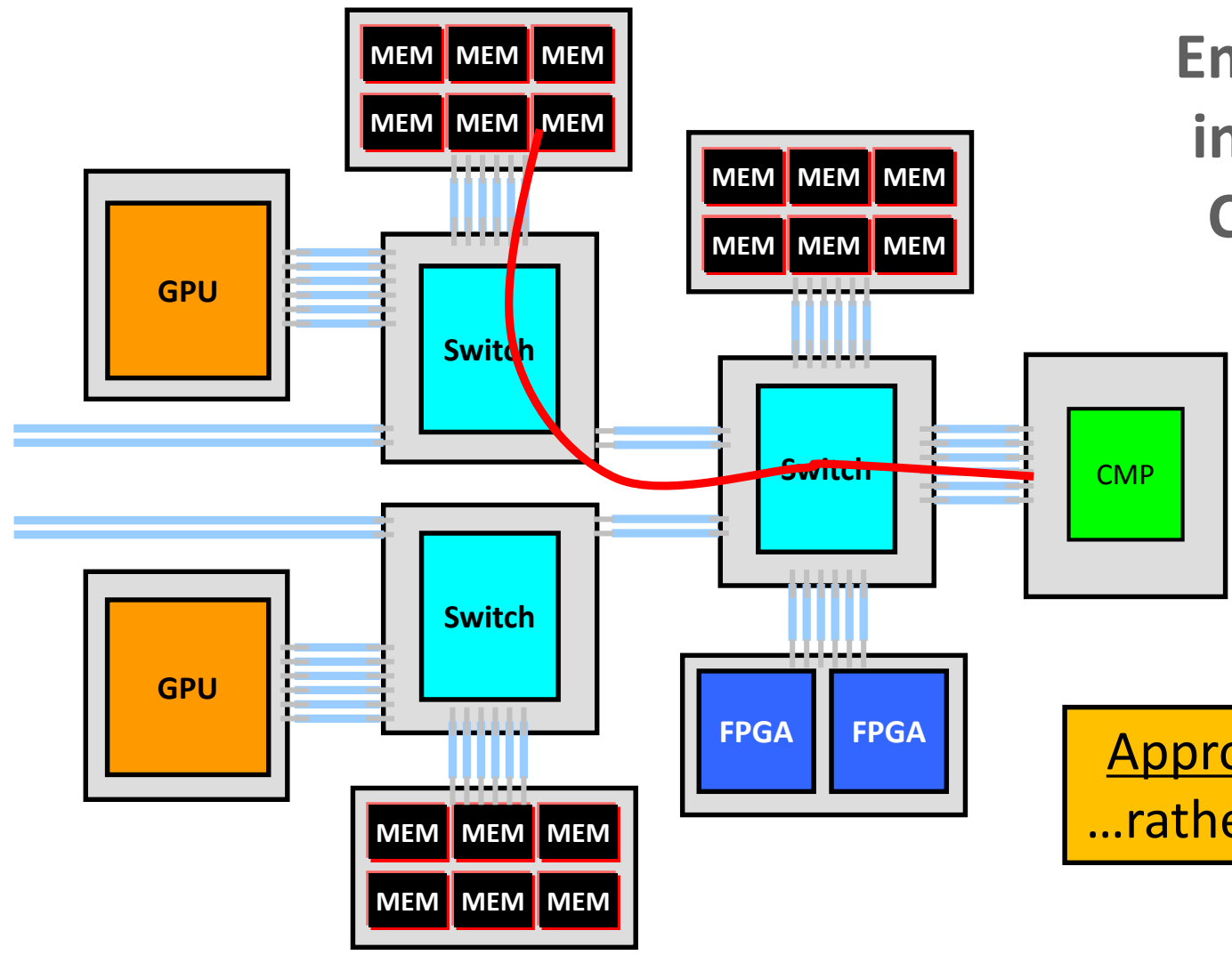
Embedded Photonics into Heterogeneous Compute/Memory

Approach: network of **resources** …rather than a network of servers

Optically-Connected Memory Architecture

2GB/cube

DRAM Layers

4 DRAM layers

Logic Layer

Substrate

WDM transceiver

64 wavelengths X

16 Gb/s/λ

= 1 Tb/s per fiber

Fiber

Each direction: 2 fibers x 1 Tb/s

≈ 256 GB/s (HBM Gen2)

High-Density Fiber Coupling Array (up to 72)

Inverse-Tapered Waveguides

pitch @ 10s of μm

To silicon photonic switch

Silicon Photonic Interposer

24 fibers per coupling assembly

= 12 cubes, 24 GB capacity, 3 TB/s

Photonic interposer

ANALOG PHOTONICS

PLC CONNECTIONS

HBM HBM HBM
HBM HBM HBM

HBM HBM
HBM HBM
HBM HBM

CMP   CMP
SiP Switch
CMP   CMP

HBM HBM
HBM HBM
HBM HBM

HBM HBM HBM
HBM HBM HBM

FREEDOM PHOTONICS

4 coupling assemblies

= 48 cubes, 96 GB capacity, 12 TB/s

U.S. DEPARTMENT OF ENERGY

**SBIR Collaborative Project (R. Carlson):
Photonic Memory Controller Module (P-MCM)**

PHOTONIC MEMORY CONTROLLER MODULE (P-MCM)
SBIR COLLABORATIVE PROJECT PHASE 2
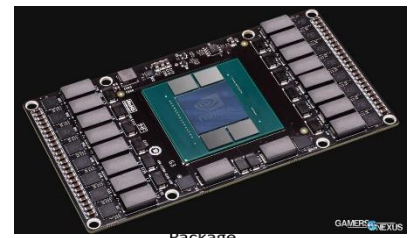
Package
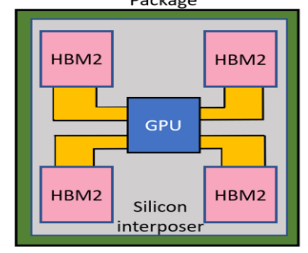
Memory-in-package HBM2 approach with 1024 wires with 1Gb/s links
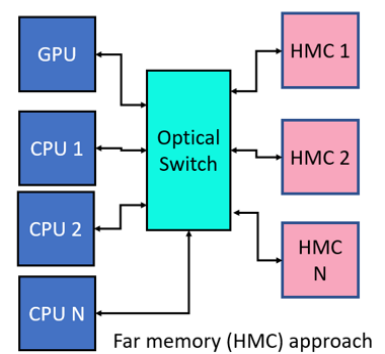
Far memory (HMC) approach in which optical links provide memory banks with 25Gb/s WDM channels
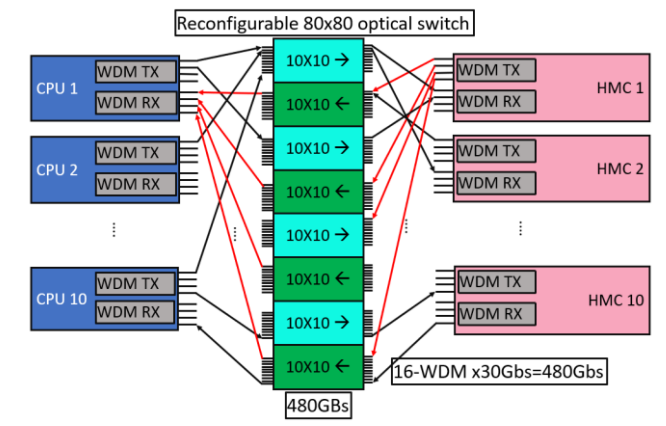
Disaggregated architecture with resource allocation

	HBM2	HMC gen3
Bandwidth	256 GB/s	320 GB/s
IO	8 Parallel (1-2G)128b per channel	30G SerDes 4 links per HMC
Package type	Si-interposer	Discrete (SerDes)
Memory access	DDR	Packet based
Target market	Graphics, Networking, less frequently accessed memory, Small form-factor	High-performance Computing, Networking

10 compute connected to 10 HMC gen3

Aggregate bandwidth:
10*4*16*30Gb/s = 2.4TB/s

Partners:

P-MCM SiP Subsystems

- Leveraging the AIM fabrication facility, Analog Photonics is developing SiP photonic based power efficient microdisk transceivers:
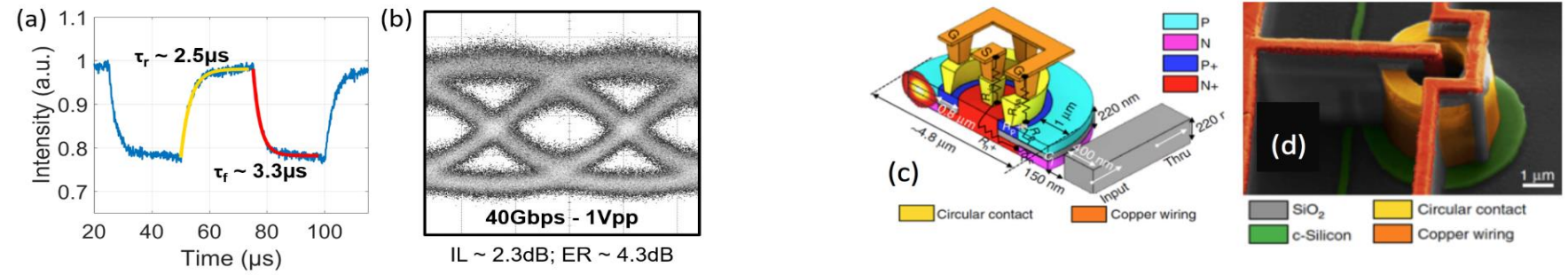
(a) Thermal tuning speed of micro-disk modulator. (b) 40Gbit/s eye diagram achieved from a micro-disk modulator. (c) The doping profile and connection design of our vertical junction microdisk modulator. (d) SEM of the fabricated microdisk.
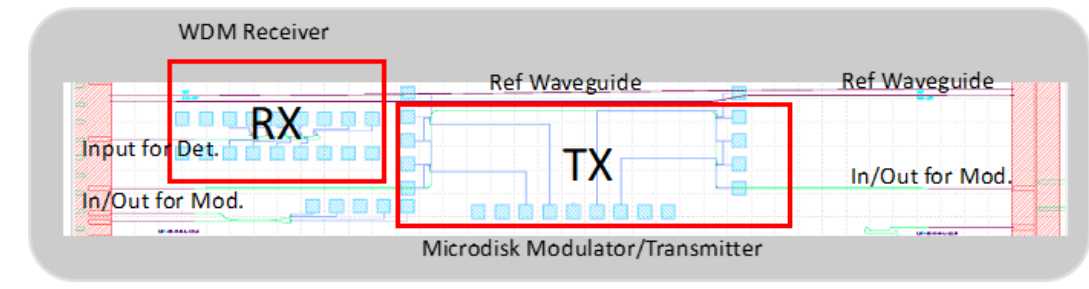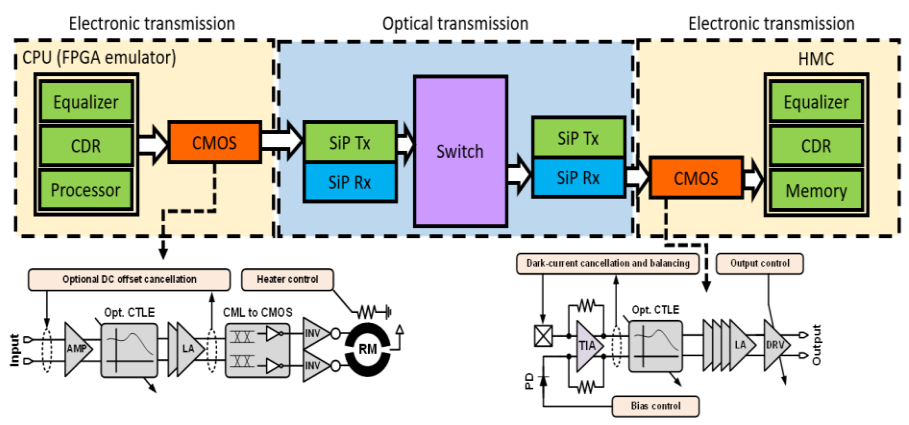
- Delivered 4x25Gbaud low-power WDM transceiver

- Co-integration of front and back-end electronics (Drivers/TIAs)

P-MCM WDM Optical Source
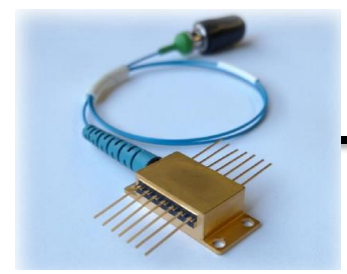
Single DFB laser
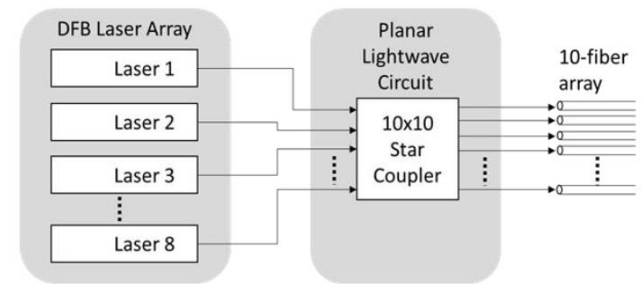
Centralized efficient WDM source for multiple transceiver node
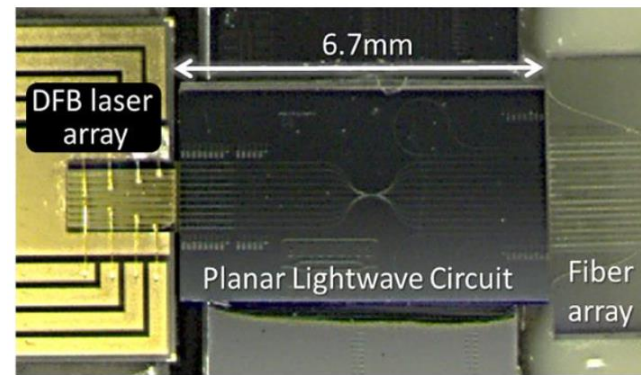
Wall-plug efficiency plot

O-band 8λ source
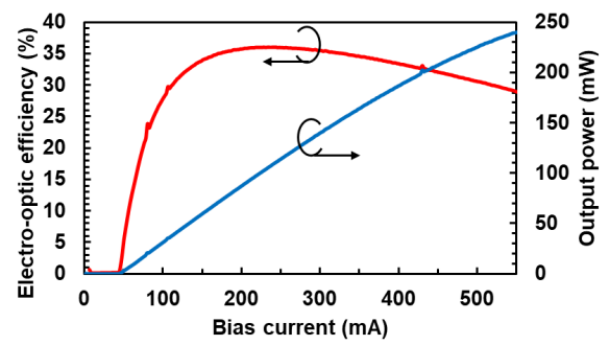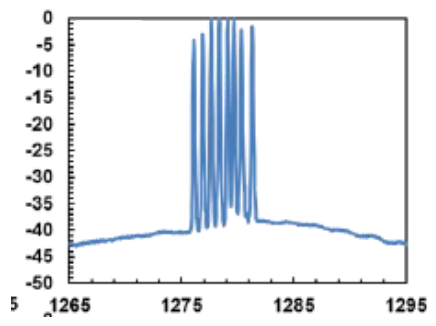
- Transform WDM laser source from a "golden box" form factor to an integrated solution with the SiP transceivers.

- DFB laser technology developed by Freedom exhibit a ultra high efficiency (>35%) when operated at high power.

- Array of DFB laser coupled to a star coupler and distributed as efficient WDM sources.

"WDM Source Based on High Power, Efficient 1280nm DFB Lasers for Terabit Interconnect Technologies" PTL 2018 [in review]

P-MCM SiP Packaging

Low-loss of fiber arrays coupling and Packaging solutions for optical sources to SiP subsystems

PLC CONNECTIONS
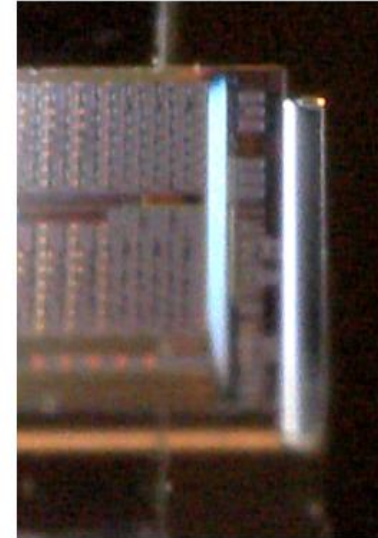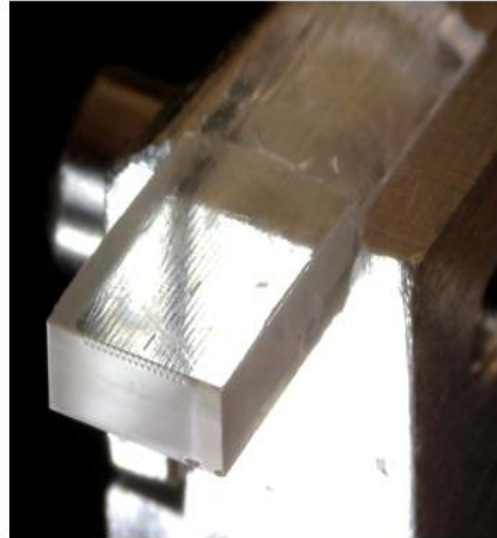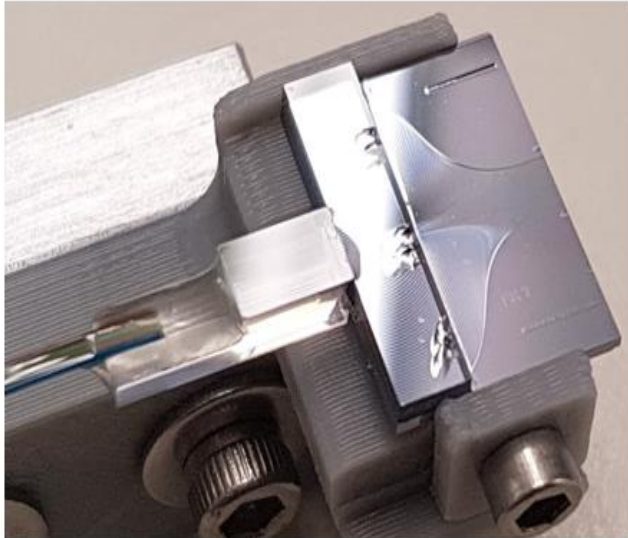
Electrical/Optical fully packaged SiP switch

- 100-port Silica transposer with 20µm output pitch;
- 6-fiber, 127µm-pitch, lid-less fiber array;
- Recessed optical facet on a typical SiP die
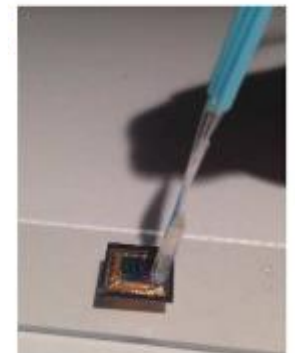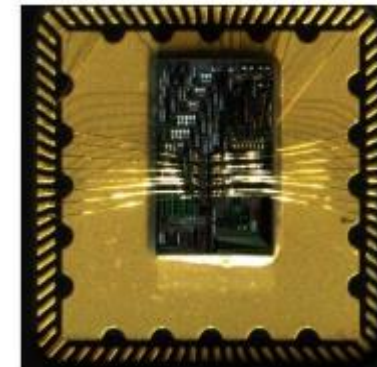
System architecture and testbed for P-MCM
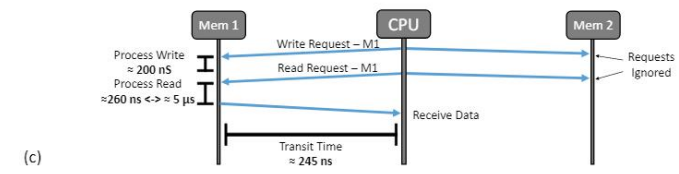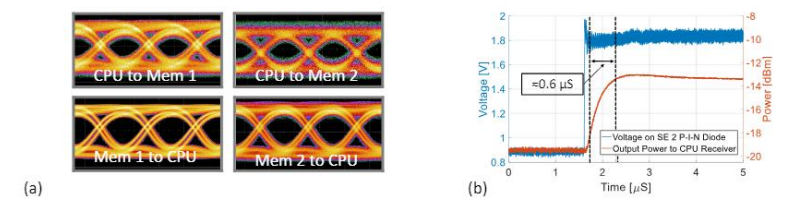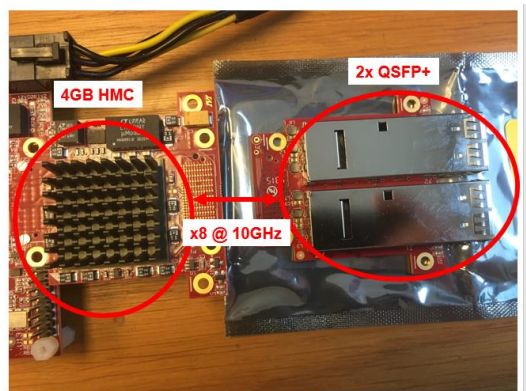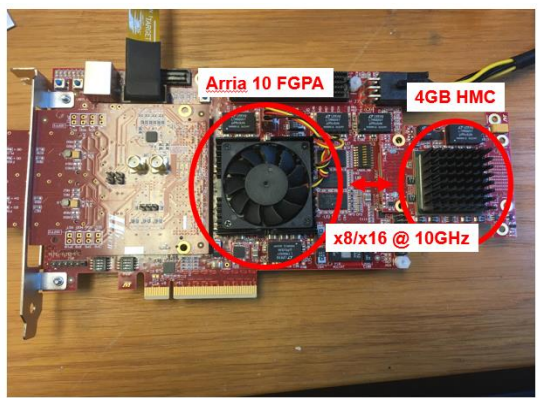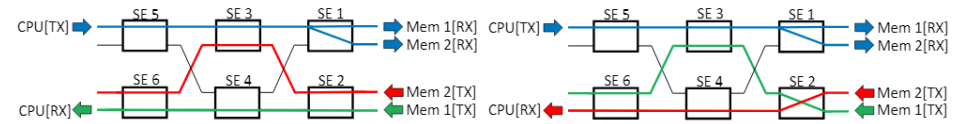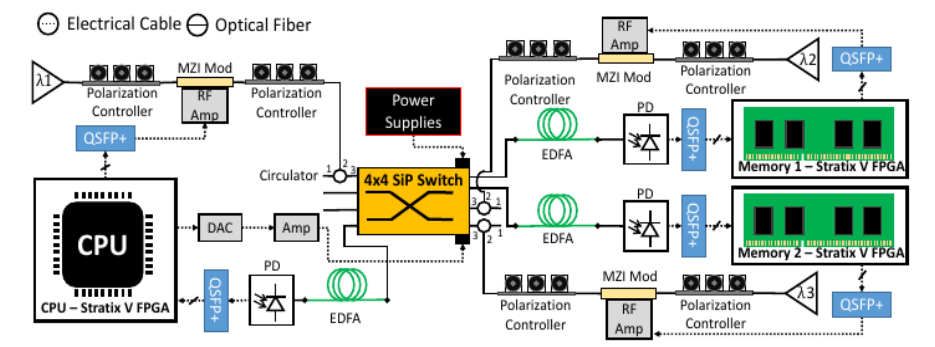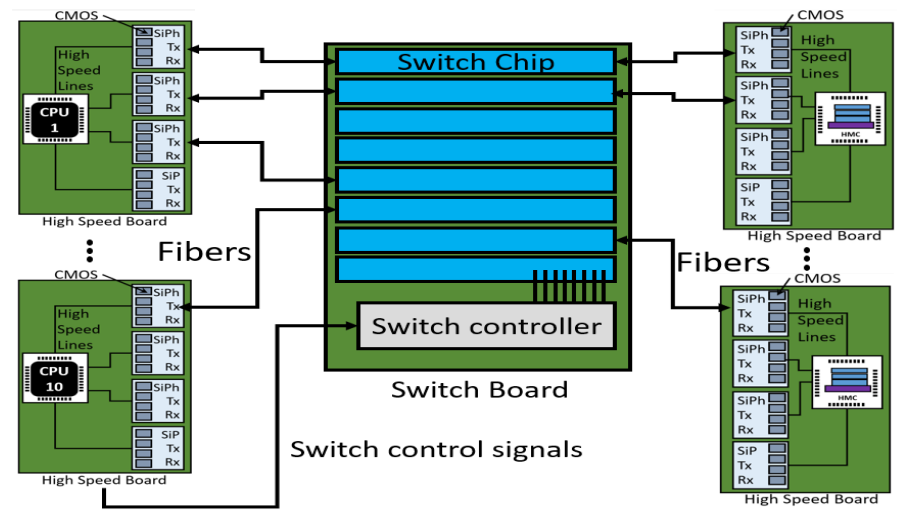
Processor/HMC gen2 testbed FPGA platform

"Reconfigurable Silicon Photonic Platform for Memory Scalability and Disaggregation." *2018 OFC* 2018.

Deep Neural Network (DNN) in the P-MCM PLATFORM

Global Memory/Primary memory

DNN

Compressed DNN

Network

DNN

input layer

hidden layers

output layer

pruning

Computation HW

[S. Han, DAC'18]
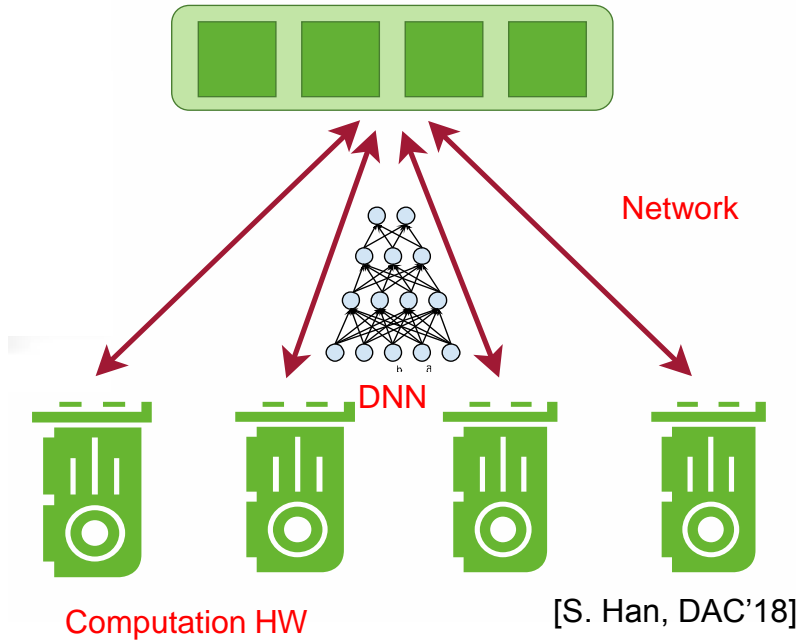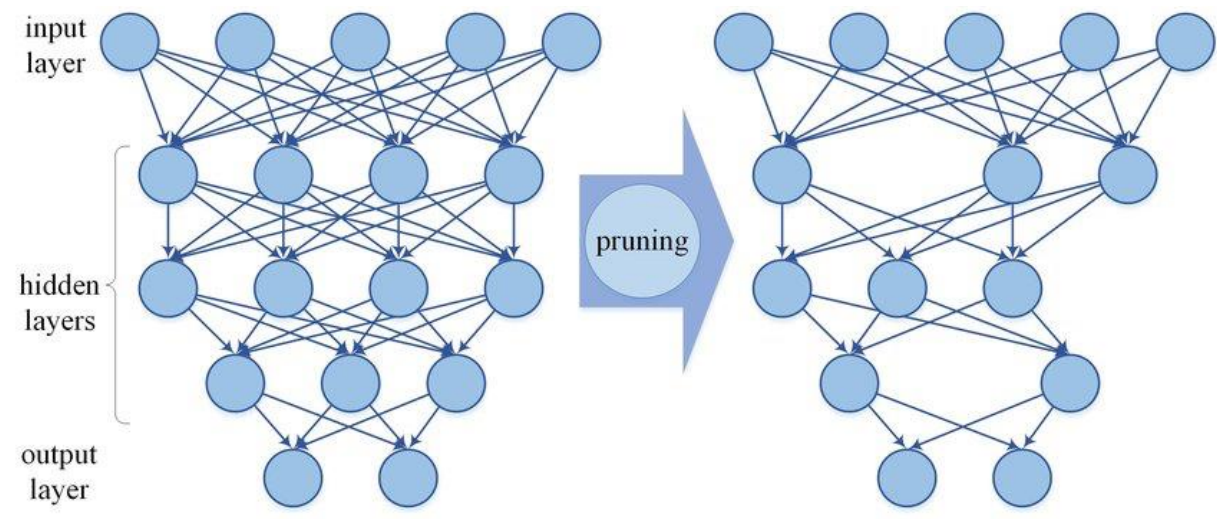
[The Dark Side of Prunning, ISCA'18]

- A DNN model is stored on global memory. It is computed with (GPU, CPU, TPU) where each has its own memory (device memory)
- DNN has three bottlenecks: **network bandwidth**, **memory bandwidth** and engineer bandwidth. (S. Han, W. Dally, DAC' 18)
- **Memory bandwidth (device)** could be saved with model compression: Pruning, Quantization, Decomposition, Distillation.
- **Network bandwidth (global)** could be saved with gradient compression.
- **Problem**: Model matrix with compression becomes sparse = sparse memory access (Yu et. al, ISCA'17)
- **Solution**:
  - Normal… custom accelerators on FPGA design with optimized memory access.
  - Goal … Photonics could enable scalability, disaggregation and increased bandwidth for DNNs.

- ❑ **1.3M ft² facility**
- ❑ **cutting edge 300/450mm toolset**
- ❑ **135k ft² of class 1 capable cleanroom**
- ❑ **processing capability span: 65nm - 7nm**

SUNY POLYTECHNIC INSTITUTE

- ❑ **300mm tools provide unprecedented quality silicon photonics**
- ❑ **unmatched 2.5D/3D stacking w/CMOS**
- ❑ **partnerships drive continuous revitalization investments**

Post S-MLD   IV-IV 65nm Si CMOS

InGaAs

III-V FINs

Undamaged
III-V FIN

300mm Si Photonics Wafer
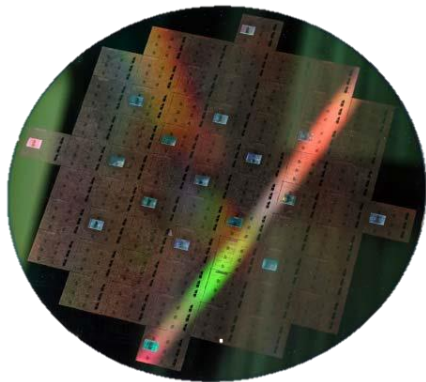
Continuously Tunable
Optical Orbital Angular Momentum
Generator

Erbium Laser

Prior | Post CMP

InP

Oxide
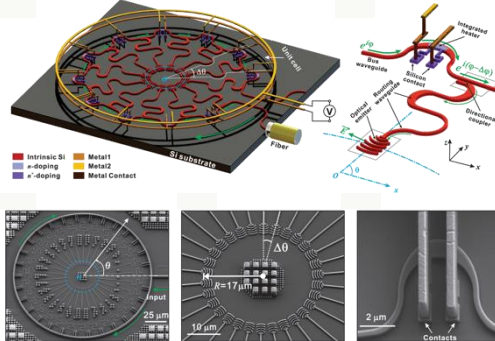
65nm

95nm gap in Si
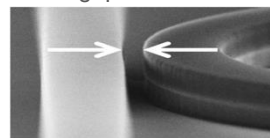
95nm Si_3N_4 Taper on
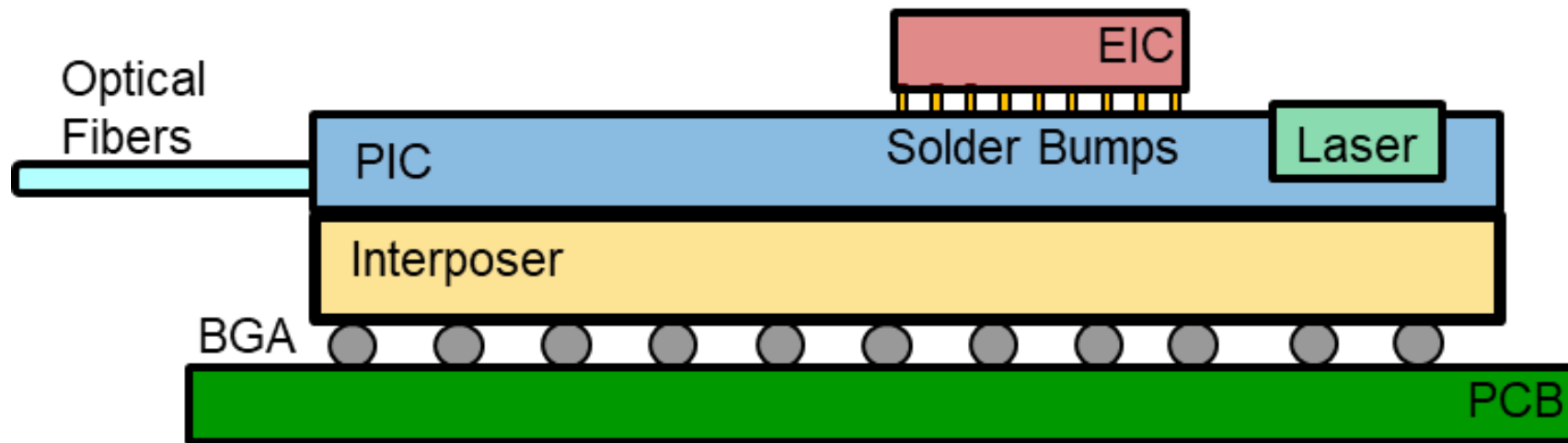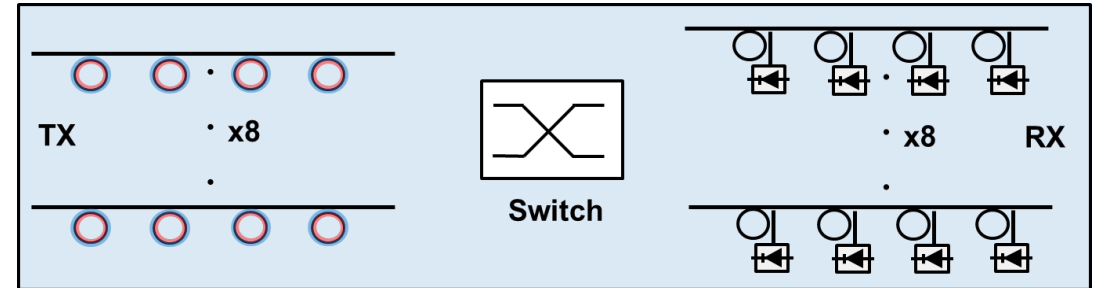Si Waveguide
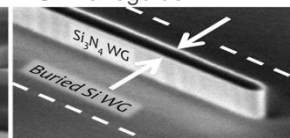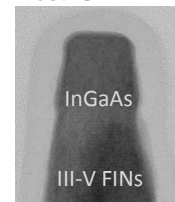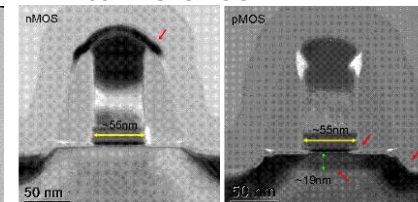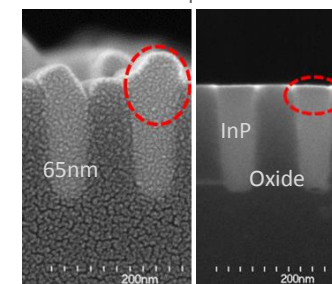
ASIC / Silicon Photonic Interposer Integration
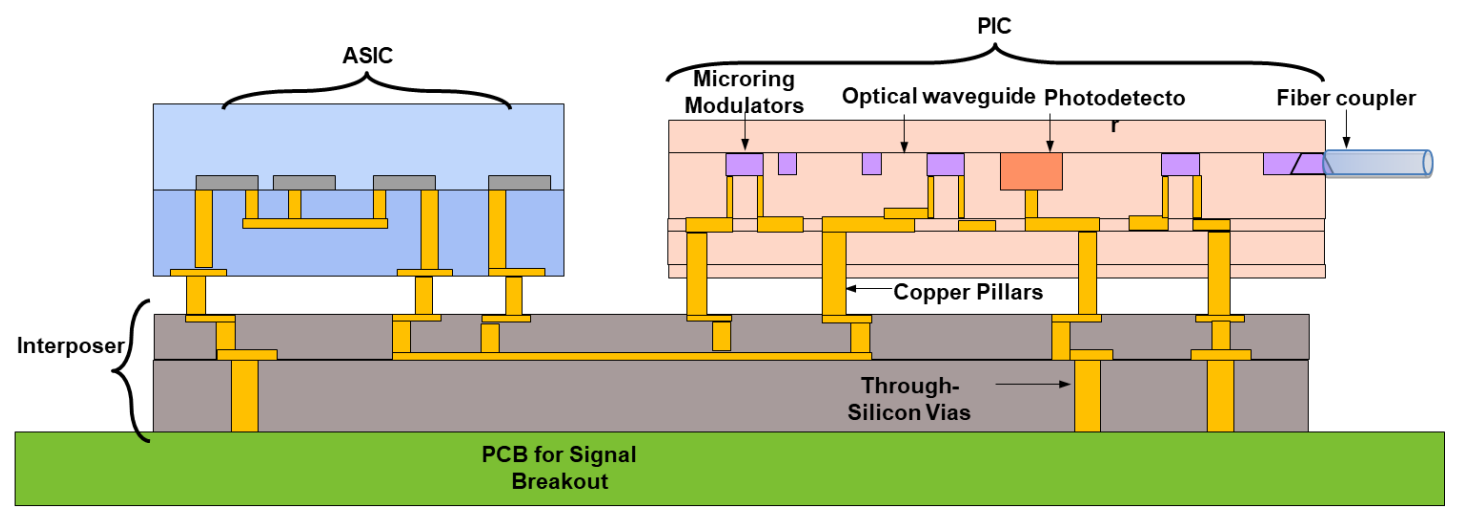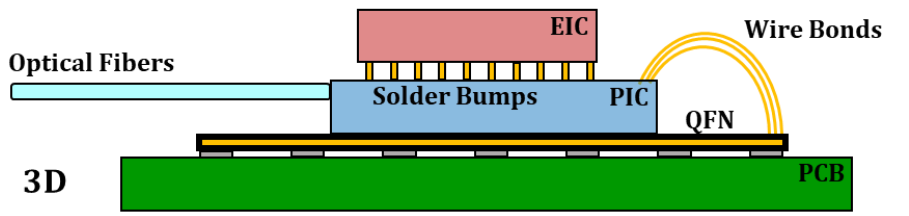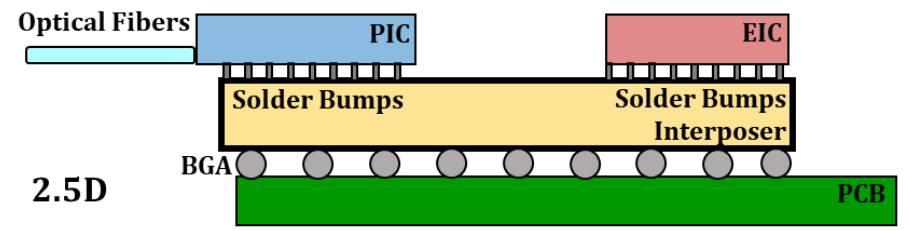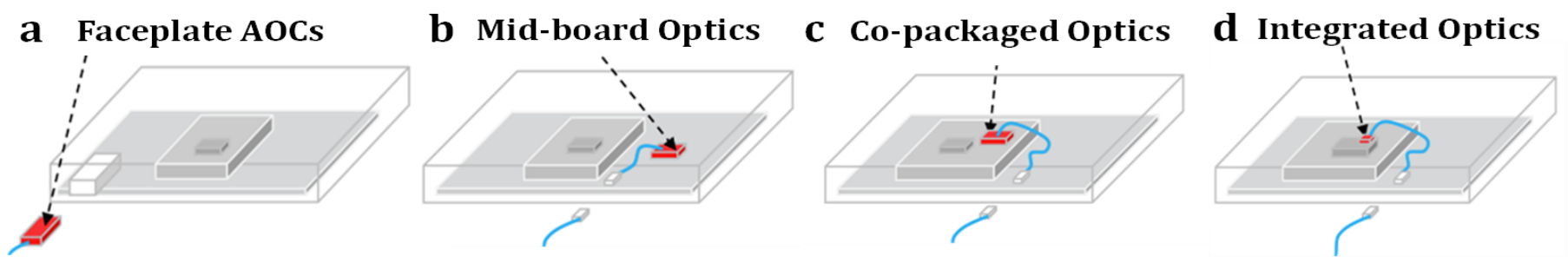

Active Interposer 2.5D

Active Interposer Full Network on Chip (NoC)

- <u>Active</u> interposer combines PIC and interposer
  = single platform
- Allows for laser integration
- EIC chips flip chipped on top of PIC
- Ideal platform for fully integrated network on chip

Active Interposer Full Network-on-Chip (NoC)

- TX and RX: located on ring around outside of active interposer to shorten RF electrical paths and PCB routing

- All signals route out to PCB through BGAs on back of interposer

- **Switch:** 8x8 MZI based switch

- **RX EIC:** TIAs for 11.3 Gbps, single channel, integrated on active interposer

- Additional switch, modulators, and laser integration test structures
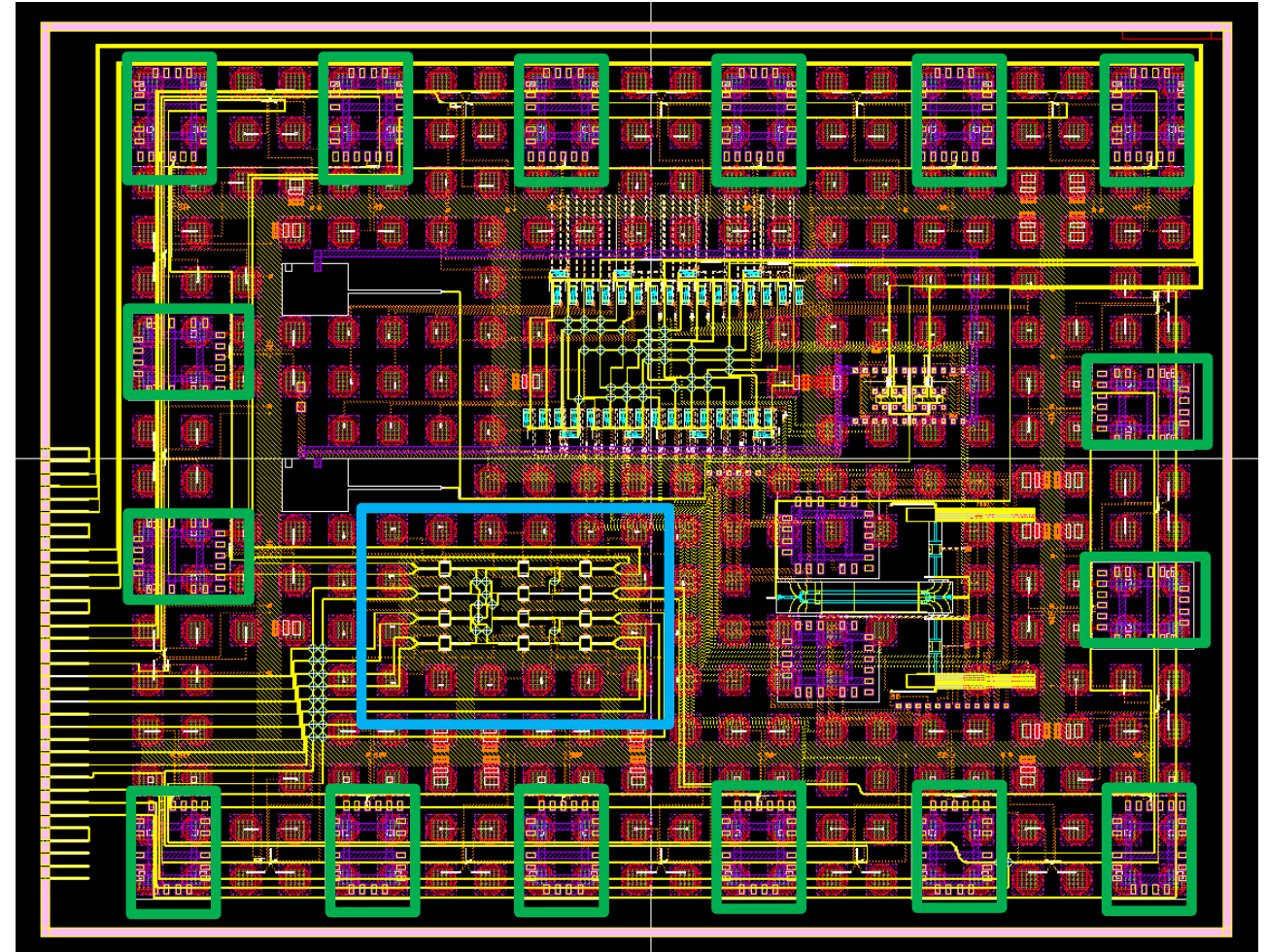
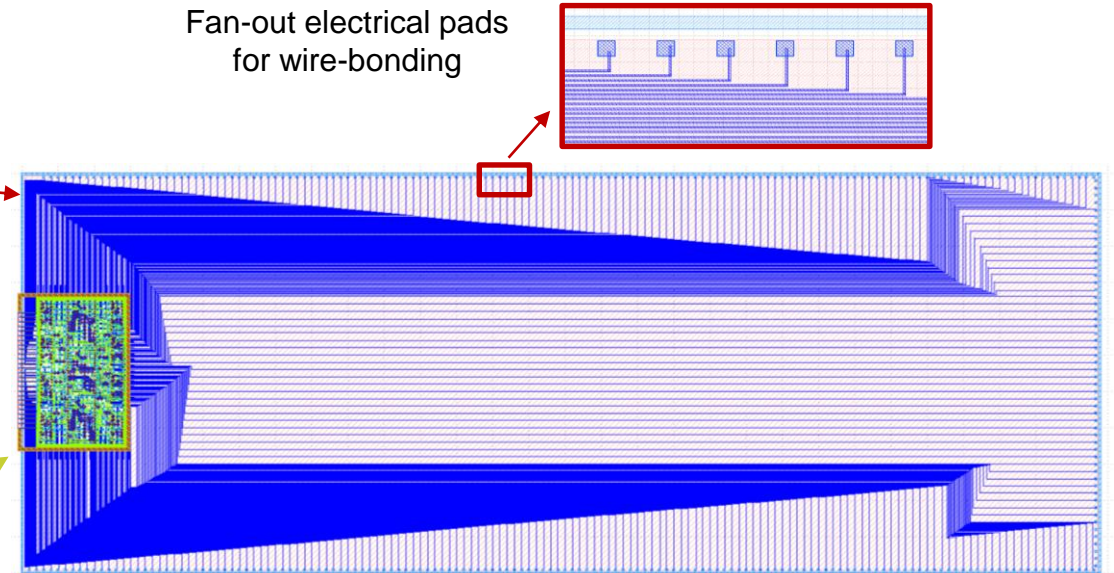# Photonic Switch Fabrication and Packaging – silicon interposer

AIM 3rd Run
12x12 T-O Clos
switch-and-select
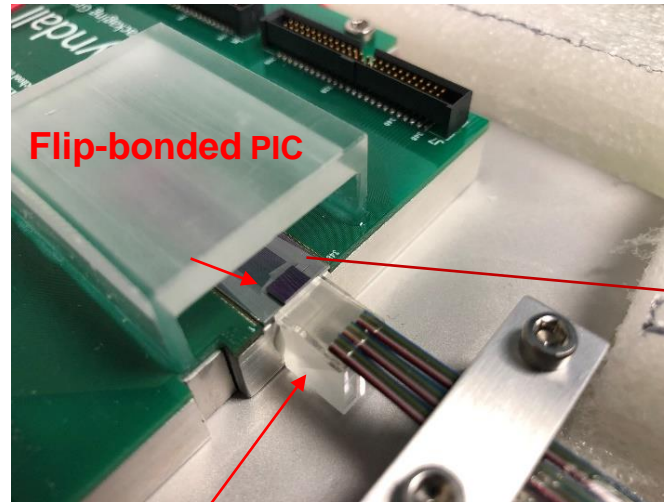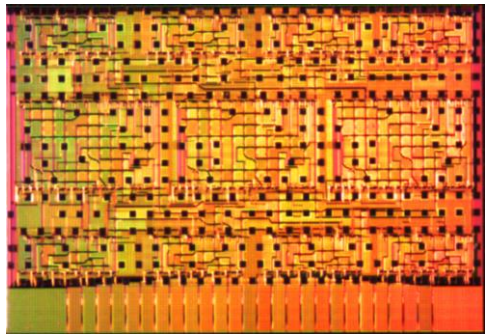
Flip-bonded PIC

Fiber-array holder

Fan-out electrical pads
for wire-bonding

Silicon PIC Switch

Silicon Interposer

- Small pitch of bonding pads on chip with large density: 100um

- Fine Pitch of electrical traces: 8um

- Enabled complex PIC with reduced footprint – improving loss/performance

Adaptive, Flexible Connectivity → Deep Disaggregation

- Universal photonic WDM-switch fabric
  - Extend TB/s photonic connectivity
  - 'anywhere' in the system

- Flexibly assembled topology, direct connectivity of resources
  - Energy efficient usage

- Transparent for packets
  - Low-Latency direct connectivity
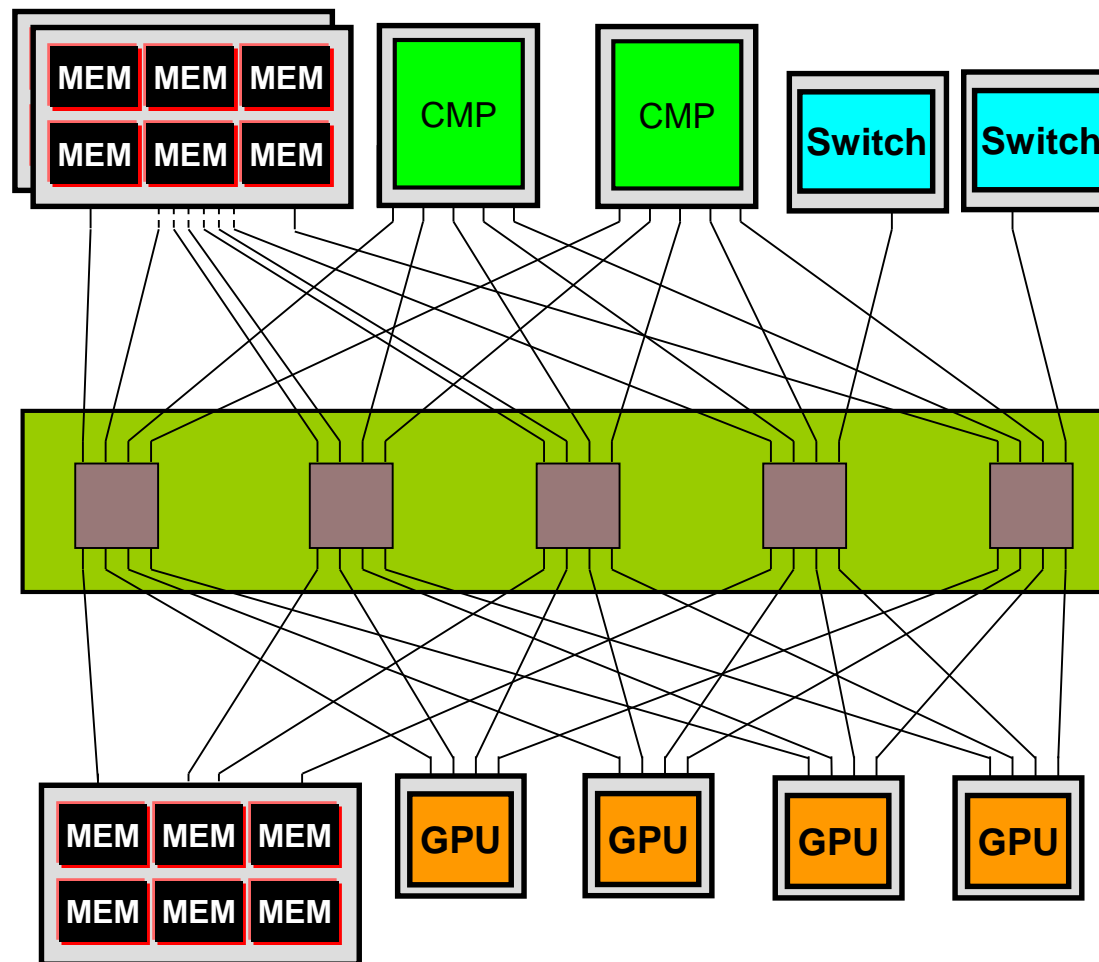
2019 SBIR/STTR Phase 1 Collaborative Development Project:

Photonic - Universal Accelerator Interconnect

- The photonic UAI must deliver >1 TByte/s of bandwidth to the CPU/Memory/Accelerator at bandwidth densities of >800Gb/s per optical I/O channel.

- Accelerator chips may be located up to 100 meters distance from the CPU/memory.

- Any chip (e.g., CPU core, Memory module, or accelerator) must be able to communicate directly with any other chip.
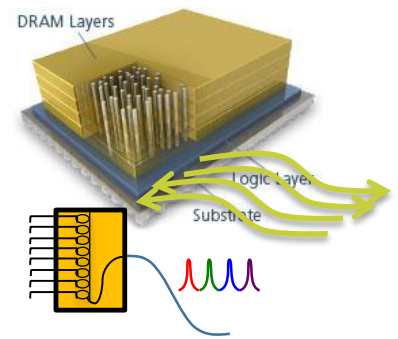
SUMMIT – Node details

- IBM AC922 nodes
- 2x IBM POWER9 + 512GB DDR4
  - 44 cores / 176 threads; 3GB RAM / thread
    - → 240GB/s total BW
- 6x NVIDIA Volta GPU
  - 16GB HBM / GPU → 900GB/s / GPU
  - 96GB HBM total → 5.4 TB/s total BW
- NVLink
  - 2 groups of 1 CPU + 3 GPUs
  - Within a group: all-to-all connected
  - 100GB/s per link
  - 2.4TB/s total BW
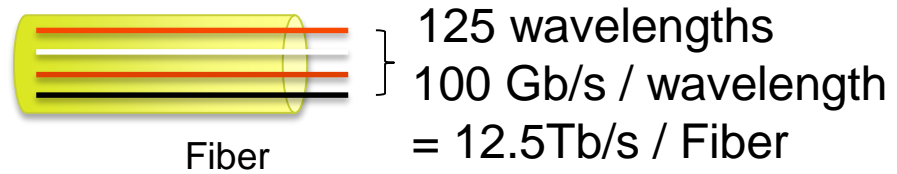- Node Memory: 608GB
- Node compute: ~40TF/s double-precision
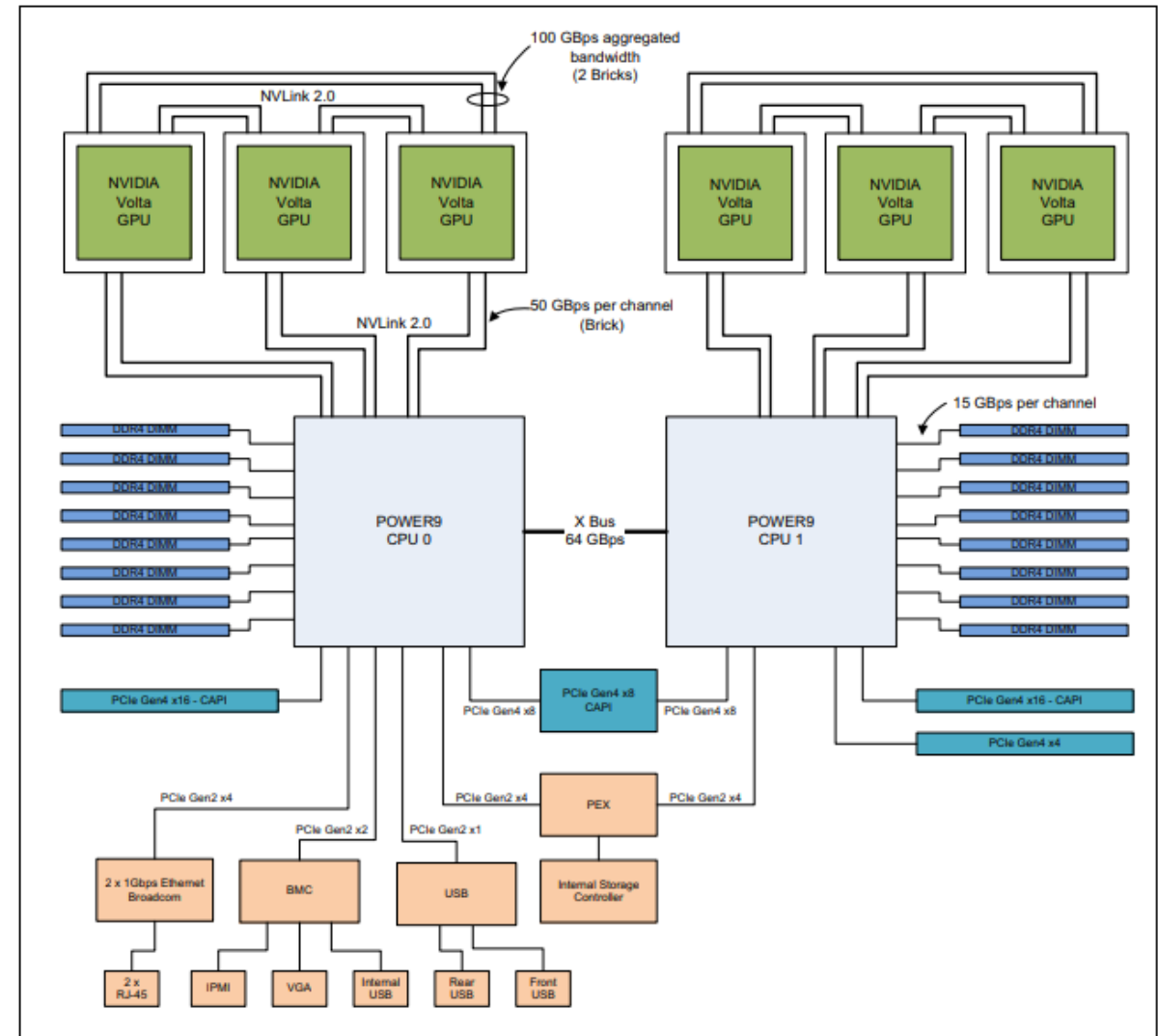
Source: IBM Power System AC922 Introduction and Technical Overview
www.redbooks.ibm.com/redpapers/pdfs/redp5472.pdf

Optically-Connected Heterogeneous Node Architecture

512GB/cube

DRAM Layers

Logic Layer

Substrate

DWDM
transceiver

125 wavelengths
100 Gb/s / wavelength
= 12.5Tb/s / Fiber

Fiber

Multiple fibers (2 TX, 2 RX) = 4 fibers = 6.25 TB/s

High-Density Fiber
Coupling Array

Inverse-Tapered Waveguides

pitch @ 10s of μm

To silicon
photonic switch

Silicon Photonic Interposer
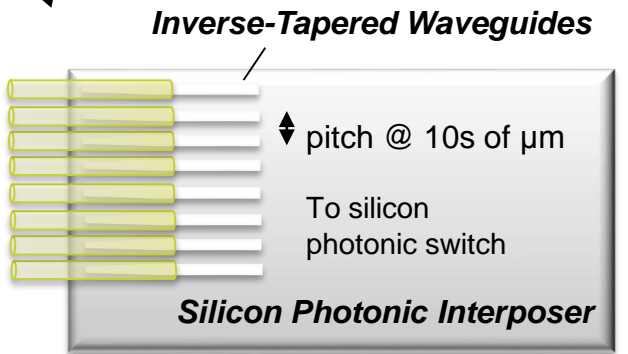
100 fibers per coupling assembly
= 25 cubes, 12.5 TB capacity, 1875 TB/s

HMC	HMC	CMP	CMP	HMC	HMC
HMC	HMC			HMC	HMC
HMC	HMC	SiP	SiP	HMC	HMC
HMC	HMC	SiP	SiP	HMC	HMC
HMC	HMC			HMC	HMC
HMC	HMC	CMP	CMP	HMC	HMC
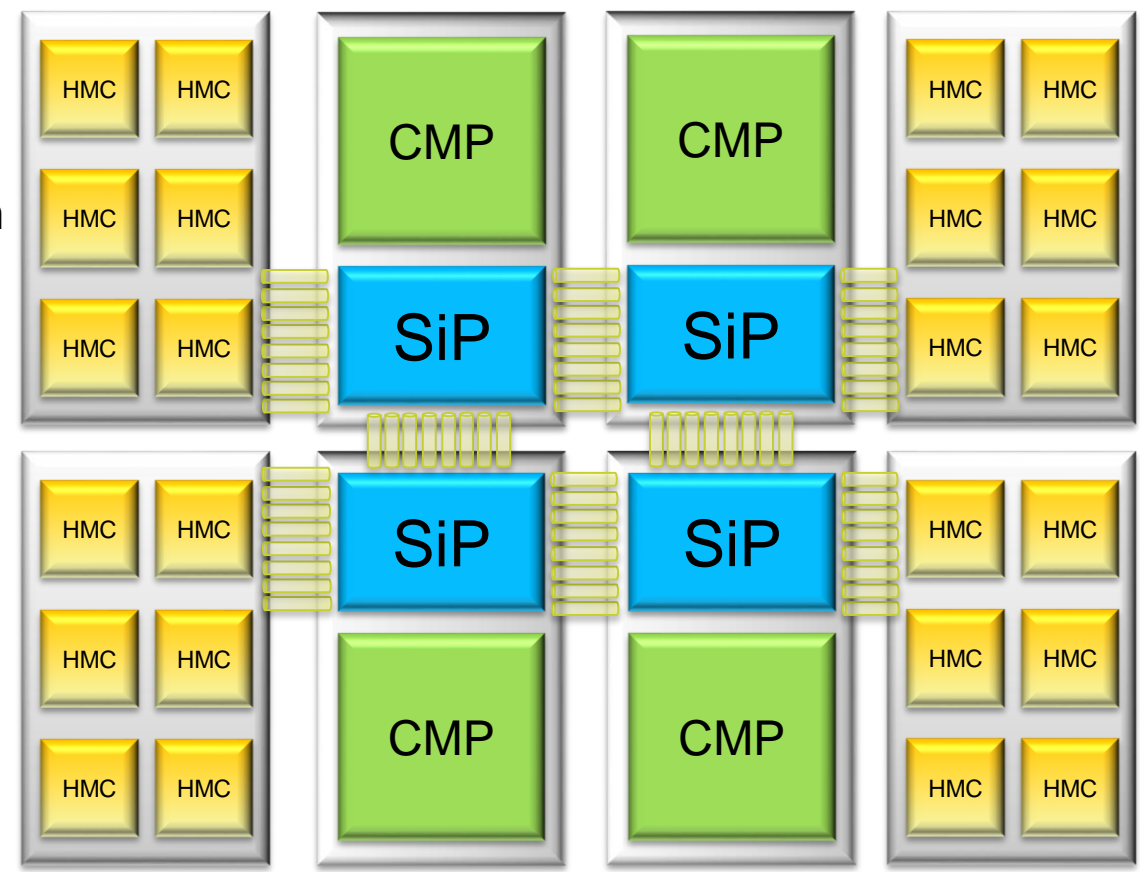
12 memory interposers
= 300 cubes, 150 TB capacity, 1.9 PB/s

Unified Photonic Fabric

Per node:

Compute: ~5PF/s (~125x SUMMIT)

Memory: ~150TB memory (~250x SUMMIT)

Communications: ~2PB/s (~250x SUMMIT)

Internal Node Bandwidth = Node Escape Bandwidth

Large optically connected memory pool

Accessible by all compute nodes

High density, multi core, multi wavelength optical links

Embedded Photonics Potential:
0.4 B/s / FLOP → 800 X SUMMIT

Summary:

- **Data Movement is Critical to any Future Performance Scaling**
  - Power Consumption
  - Bandwidth Density (and Cost)

- **Photonics: System-Wide PB/s Connectivity Bandwidth**
  - 10sTb/s per 'wire' and 1 pJ/bit
  - High bandwidth Optically Connected Heterogenous: Memory/GPUs/CPUs
  - Intra-Node communications bandwidth: PB/s = Inter-Node escape bandwidth

- **Deeply disaggregated Architectures**
  - Optical connectivity for <u>flexibly assembled interconnectivity</u> topologies

- **Computer architecture landscape is changing rapidly - Data Analytics, AI**
  - Optical bandwidth steering, adaptable architectures for scalability
  - Ultimate energy efficiency – use only required resources for needed time period

Extra Slides

P-SSIO System Testbed

Optical PCIe PCIe (Gen2 x1) testbed:
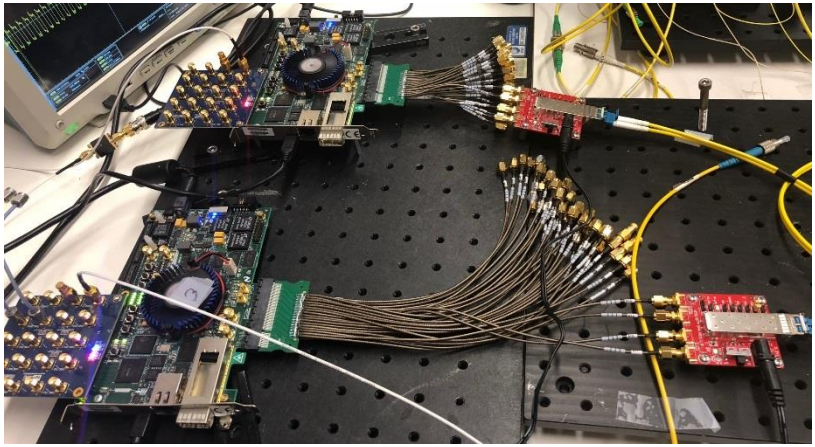
Eye diagram when optical link is on:
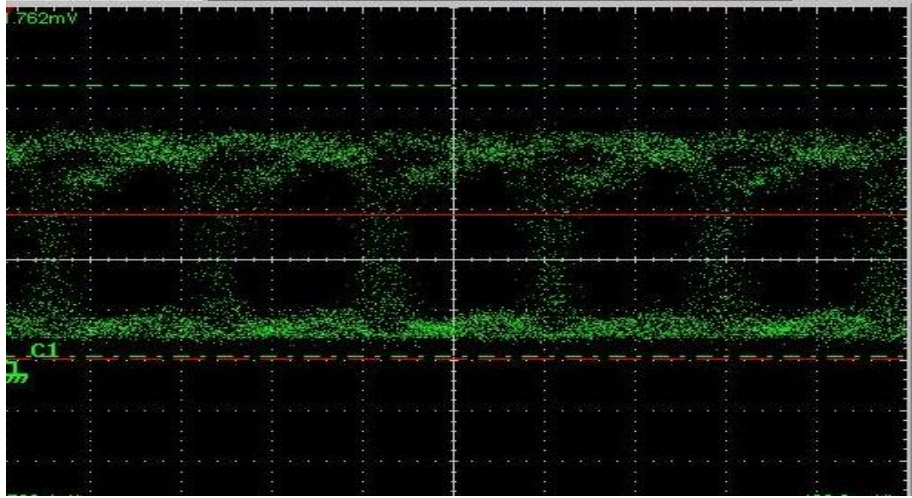
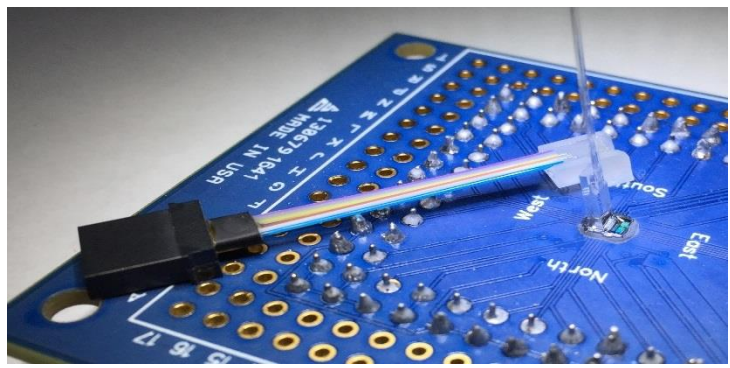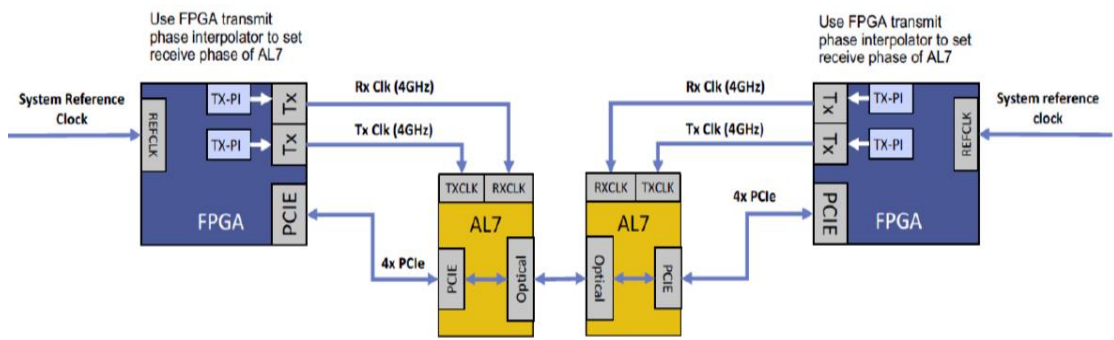**Device configuration and data transaction records at Root-port :**

Device configuration

Data transaction

Transceiver Chip and Laser Source
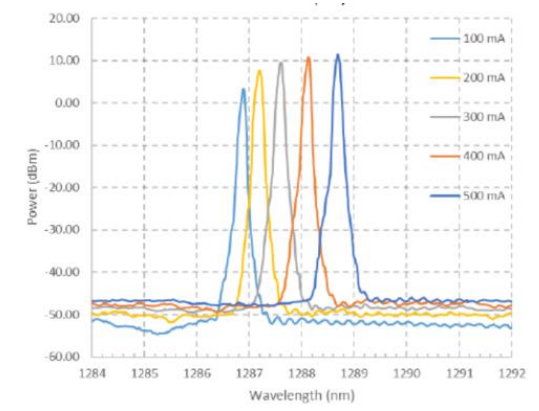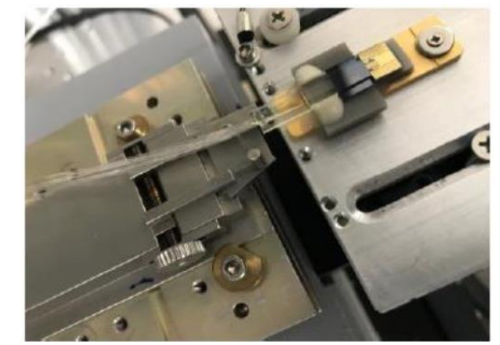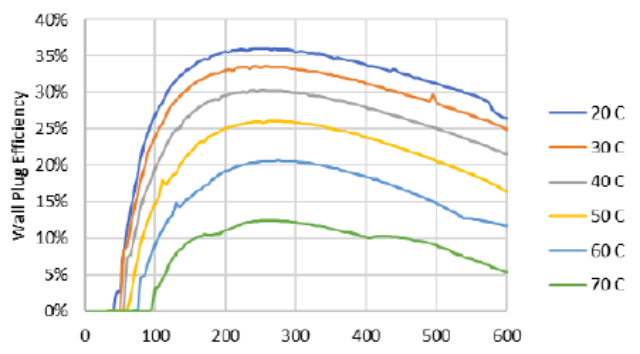
Ayar Labs

FREEDOM PHOTONICS

Ayar Labs:

Freedom Photonics:

- Supply silicon photonic ring-based transceiver chip for system demonstration
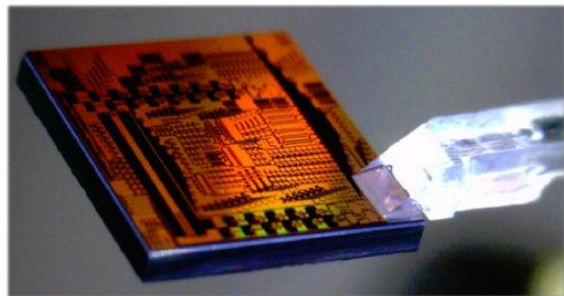- Configure optical TX/RX settings for PCIe data rate

- Delivered Tunable laser source for system demonstration
- Improve temperature performance with uncooled operation
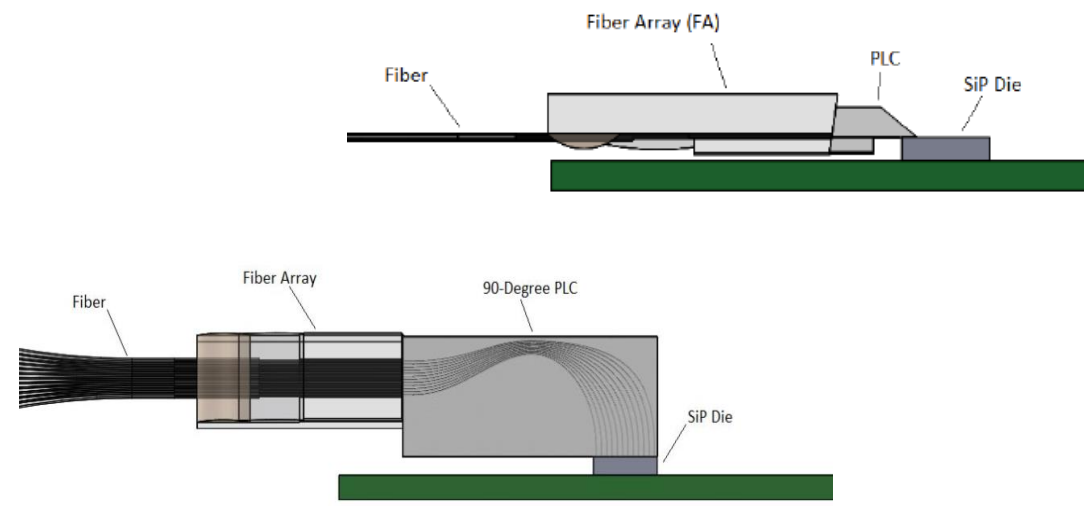- Design WDM DFB laser array for Phase II

Fiber attachment and SiP Packaging

PLC CONNECTIONS   nanoPrecision products

**PLC Connections:**

nanoPresicion:

- In-plane (horizontal) design for reduced height fiber attachment with <u>low coupling loss</u>
- Finalized two designs for fiber attached on-chip testing

Develop a commercially viable (low-cost, high-volume) approach to packaging the P-SSIO components and electro-optical co-assembly

Conventional Architecture → "Assembled" with Flexible Interconnect
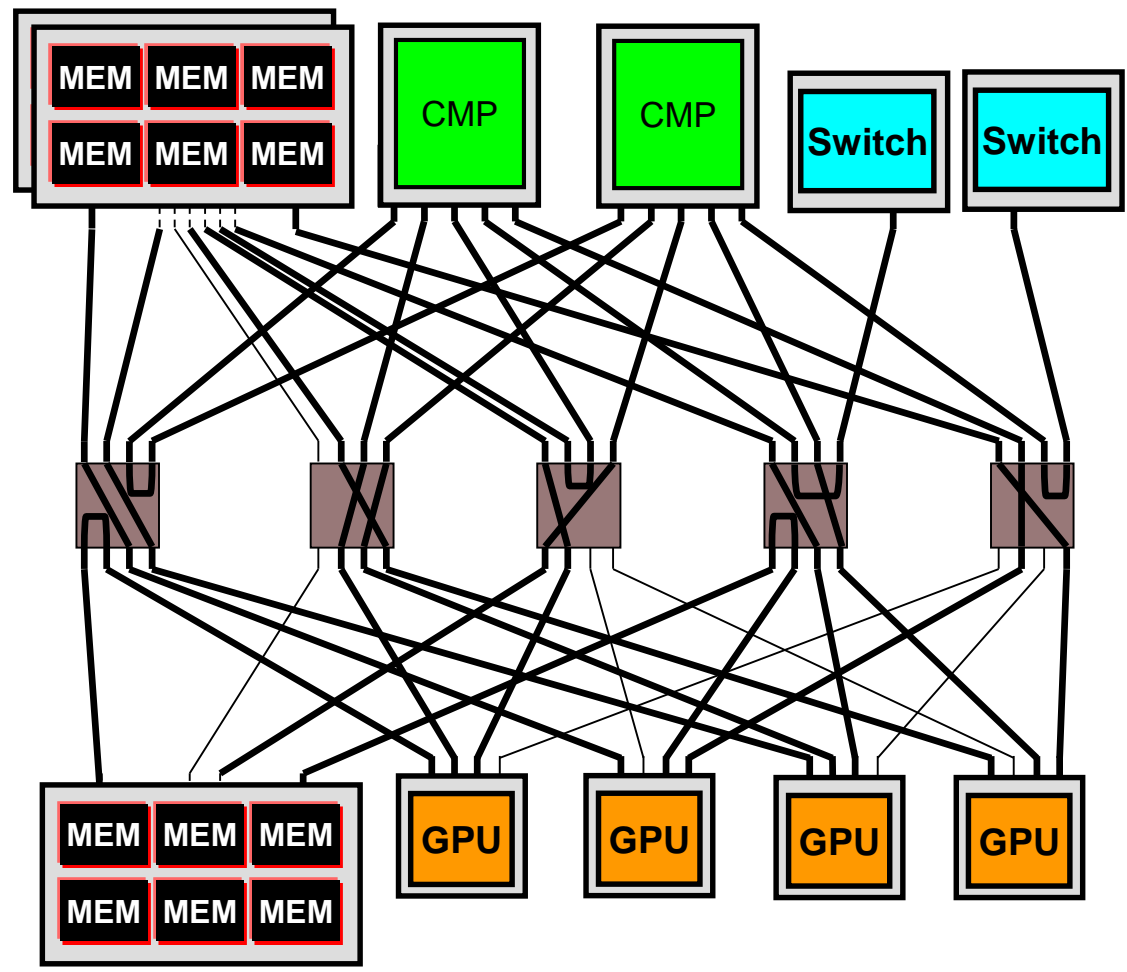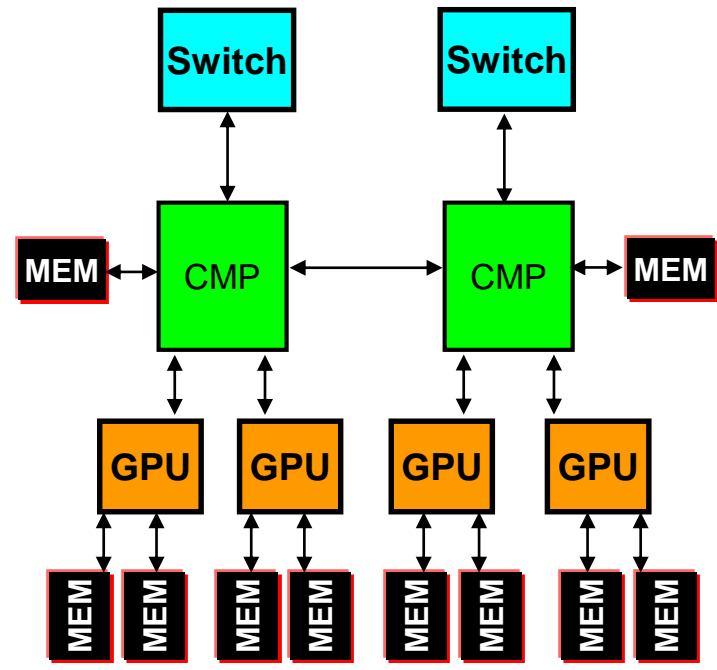
GPU Centric / CMPs Data-Accelerators
→ Node Architecture "Assembled" with Flexible Interconnect