PINE Phase 2: Deeply Disaggregated Computing Systems with Embedded Photonics

Keren Bergman

Lightwave Research Lab

Columbia University, New York, NY, USA

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

PINE: Photonic Integrated Networked Energy efficient datacenters

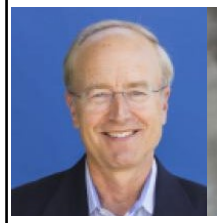
PINE

Bergman

Phase 2 Team

Baiocco

SUNY Polytechnic Institute



Bowers Liu

Q QUINTESSENT

Glick Gaeta Lipson

AIM photonics

Patel

CISCO

Dennison Gray

NVIDIA

Shalf

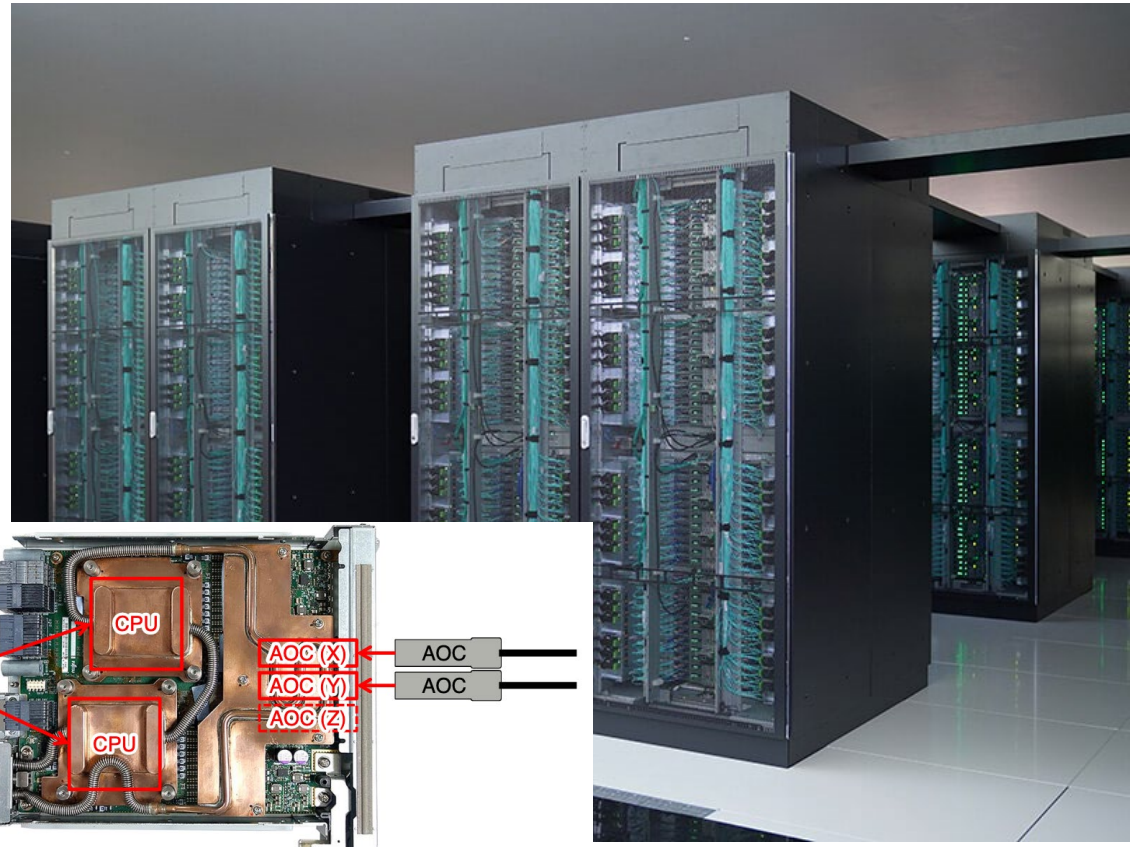
BERKELEY LAB

Ghobadi

MIT

High Performance Systems: Trends and Challenges

- Fugaku (Fujitsu and RIKEN)
  - Most powerful supercomputer* (June, 2020)
  - Performance: 415.5 PetaFLOPS
    - 80% of peak Linpack
    - 2.8X over Summit
  - Power consumption: 28.33MW
    - Efficiency: 14.7 GFLOPs/Watt (#9 Green 500)
  - 158k Nodes (432 racks) with:
    - ARM A64FX 2.6TFLOP/node
    - 3D stacked memory 4x 8GB HBM 1024 GB/s
- 6D Torus Tofu-D
  Each node 10 ports and 400Gb/s BW
  Aggregate network BW 6.5PB (52Pb/s)



Fugaku performance on HPCG = 13PF
This is 2.8% of the Peak Compute

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

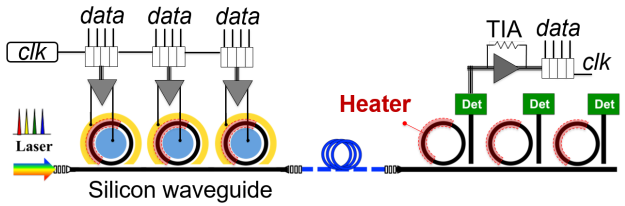
HPCG - high performance conjugate gradient

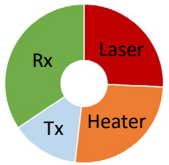
PINE: System-wide High-Bandwidth, Energy Efficient Connectivity

PINE

Energy Optimized Links

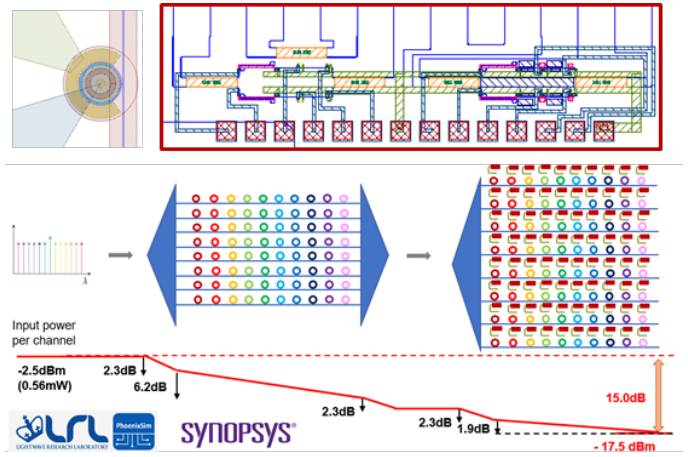
- **800G** aggregate bandwidth per link and **2.2pJ/bit**





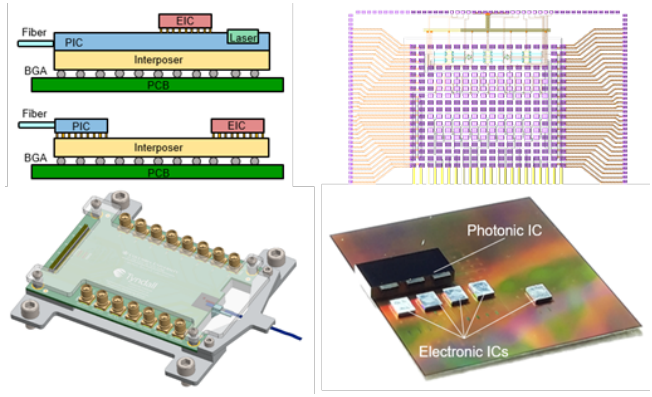
PINE Power Analysis	
Total Link Consumption	2.2 pJ/bit
Link Power Margin	3.0 dB
Link Consumption (no margin)	1.9 pJ/bit

- Complete **electronic/photonic co-design** platform, PhoenixSim/Synopsys integration. Full link PDK all components, fabrication, and measurements

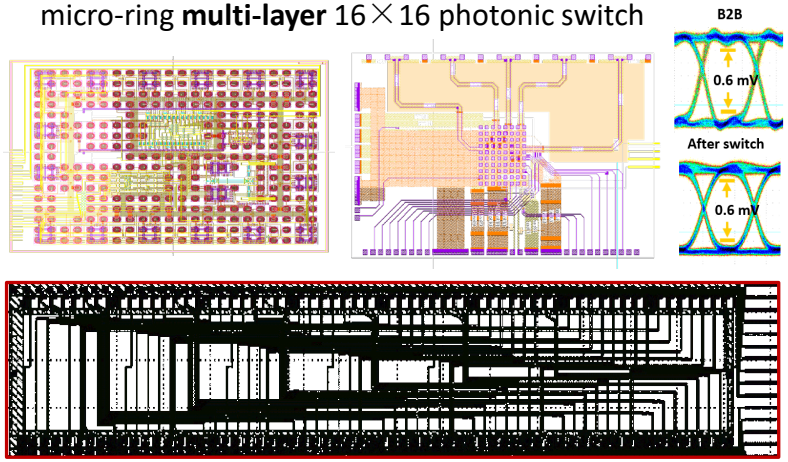


Optically Connected Multi-chip Modules

- Multi-chip module **2.5D and 3D** assembly developed with high-density active interposer **3.2 Tb/s/mm**.

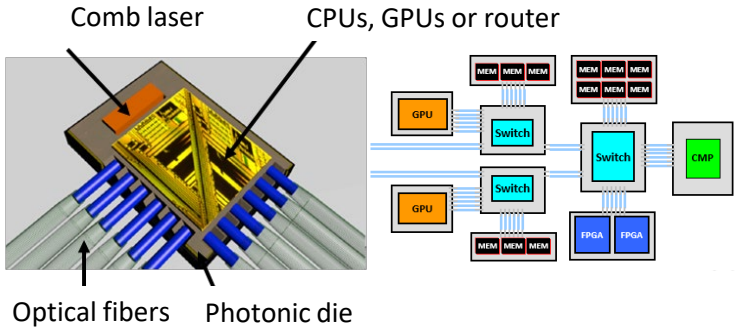


- **3D MCM** active interposer complete **network-on-chip** micro-ring **multi-layer** 16×16 photonic switch

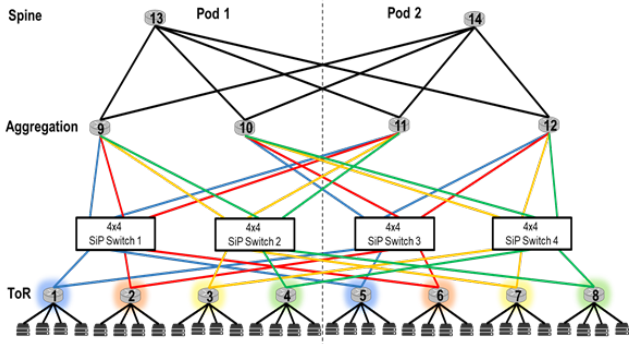


Adaptive Bandwidth Steering Deeply Disaggregated Architecture

- PINE system architecture with flexibly assembled nodes for bandwidth steering.

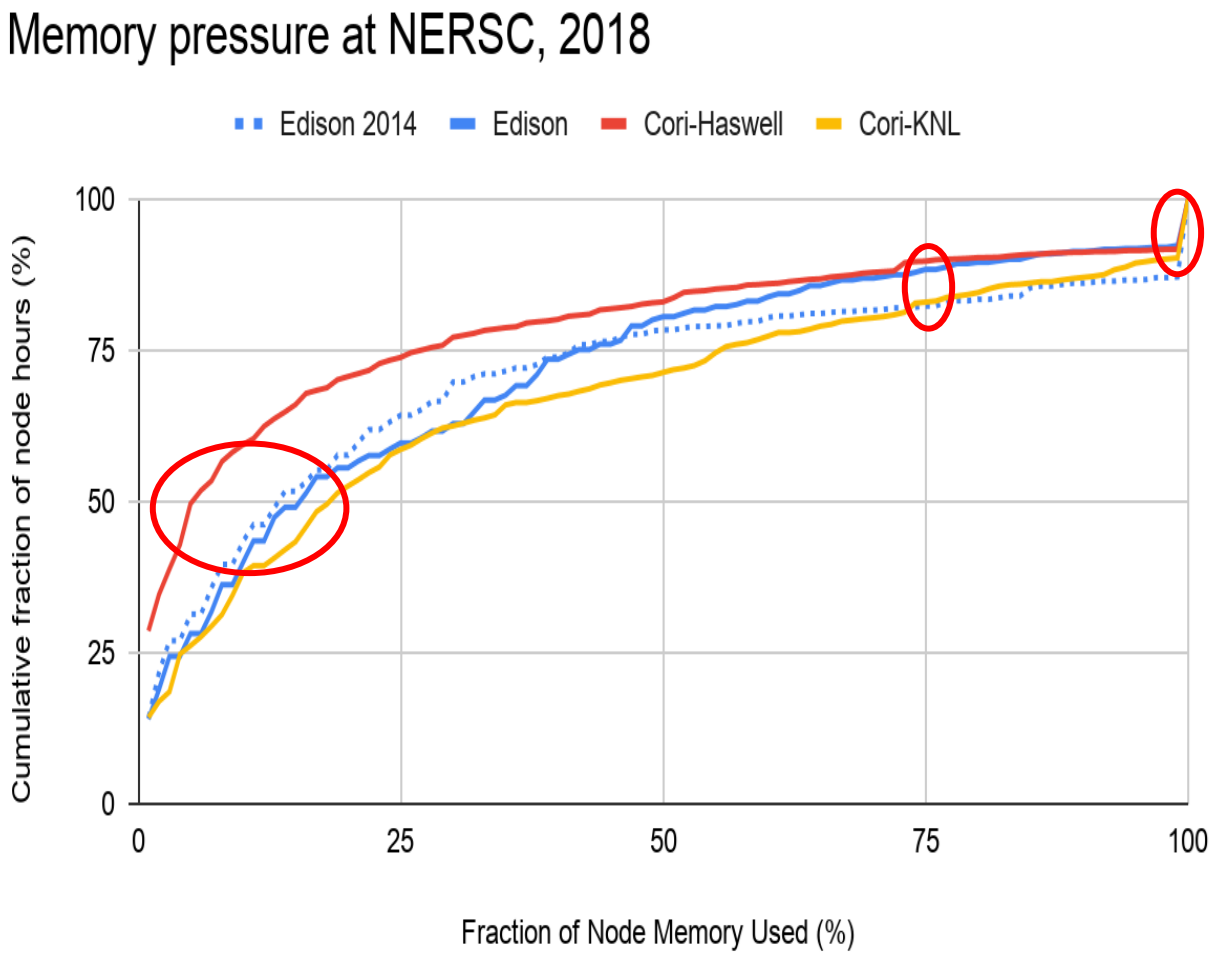


- PINE Flexible Fat Tree architecture: efficient resource utilization, tapering, **reduced power-per-transaction**



- Average **network latency reduces by 87%.**
- Average **throughput (transactions/sec) improve 4.3x**

Diversity of Memory Requirements



Brian Austin, Taylor Groves, NERSC

Why we need <u>Flexible</u> Memory Capacity and BW

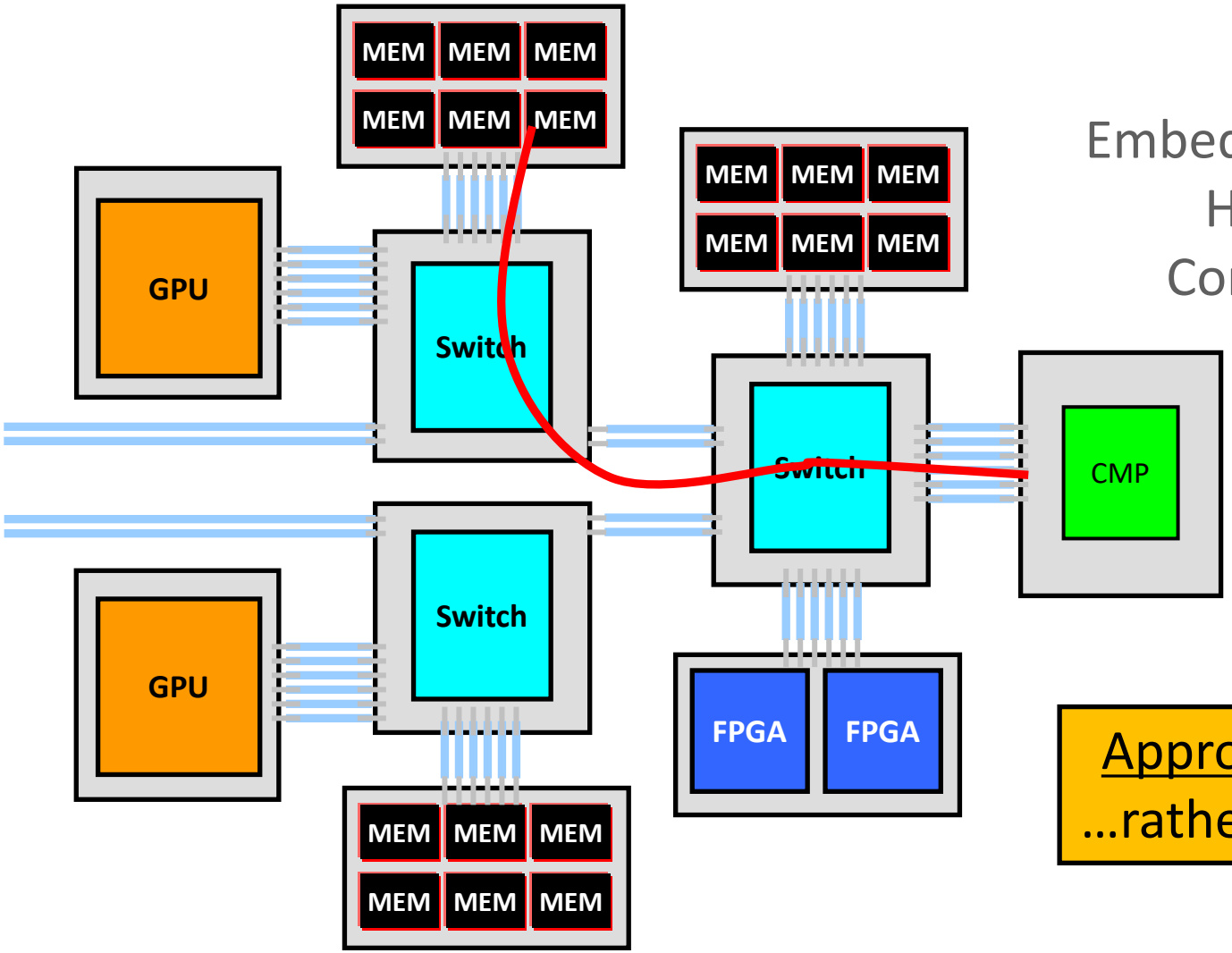
About 15% of NERSC workload uses more than 75% of the available memory per node.

But 50% use LESS than 15% of memory

spike at 100% - these are the jobs that reported memory use that is physically impossible (more than installed memory)

John Shalf, George Michelogiannakis, Brian Austin, Taylor Groves, Manya Ghobadi, Larry Dennison, Tom Gray, Yiwen Shen, Min Yee Teh, Madeleine Glick, and Keren Bergman, "Photonic Memory Disaggregation in Datacenters," Paper PsW1F.5, OSA Advanced Photonics Congress, July 2020.

Flexible Photonic Interconnected Resources



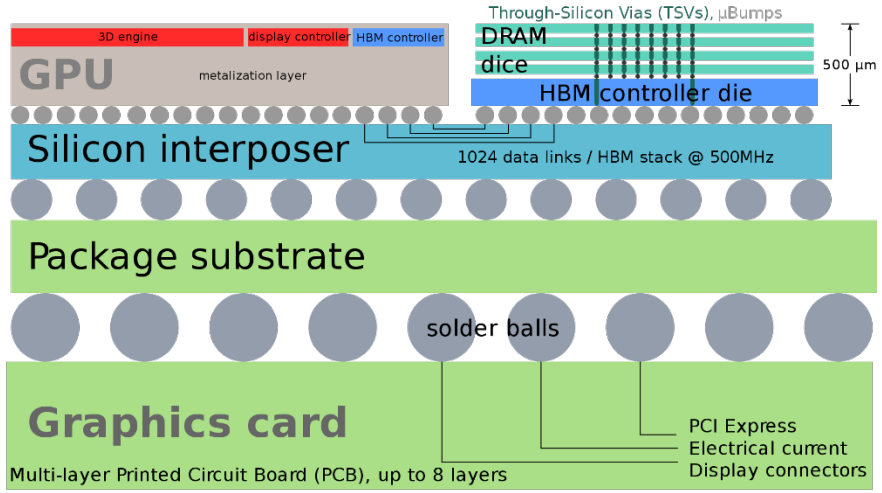
Embedded Photonics into Heterogeneous Compute/Memory

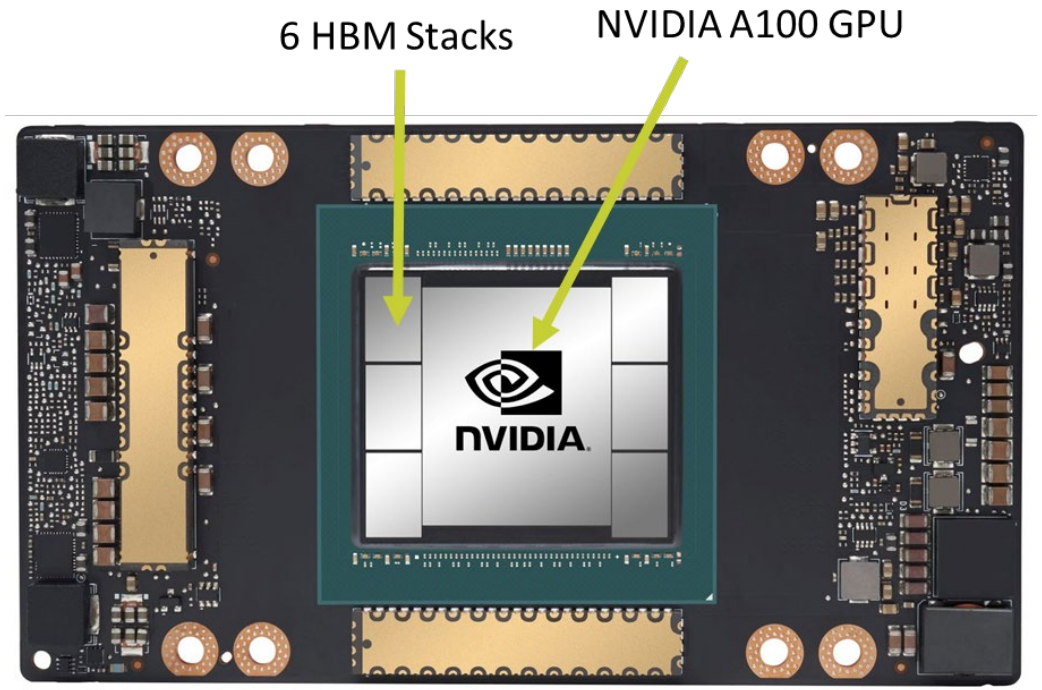
Approach: network of **resources**
…rather than a network of servers

GPU-Memory

GPU-Memory

- Each NVIDIA A100 GPU
- 6 HBM stacks per GPU
- 40 GB HBM / GPU
- 1.6 TB/s / GPU





High BW Memory constrained:

- How much can fit in close proximity to GPU
- Same interposer, wirelength, footprint
- Power density

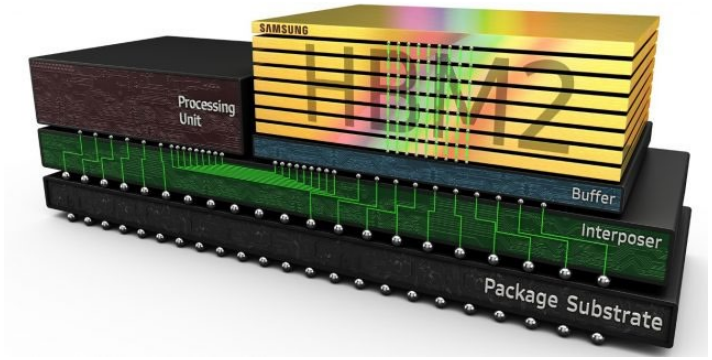
https://www.semiconductor-digest.com/2020/02/04/the-history-and-future-of-dram-architecture-in-different-application-domains-an-analysis/

https://developer.nvidia.com/blog/nvidia-ampere-architecture-in-depth/

Embedded Photonics – Scaling Memory BW

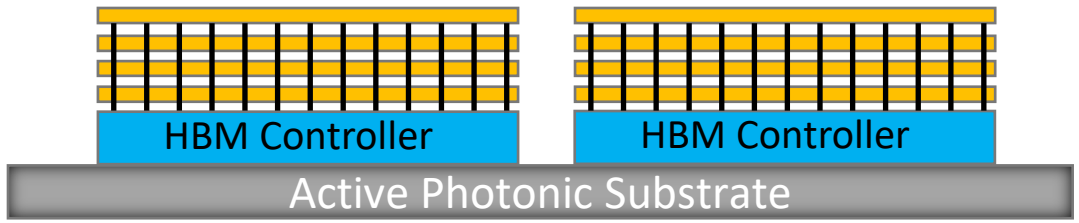
Samsung Flashbolt HBM [†]

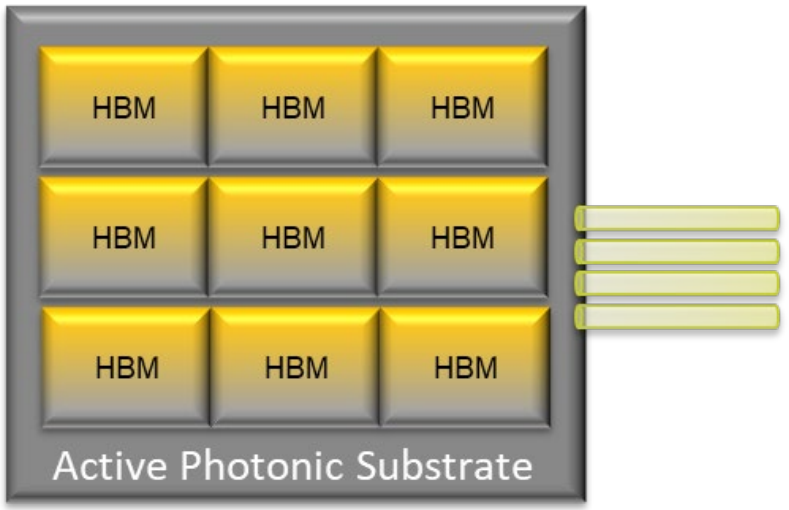
- Capacity 16GB/stack,
- Memory BW ~400GB/s/stack
- Memory BW/capacity ratio: 25x
- 10x11mm = 110mm^2



Scaling HBM over full interposer:

- ~1000mm^2 with 9 stacks
- 144GB per package with current HBM
- Using 25x memory BW/capacity ratio: ~4 TB/s

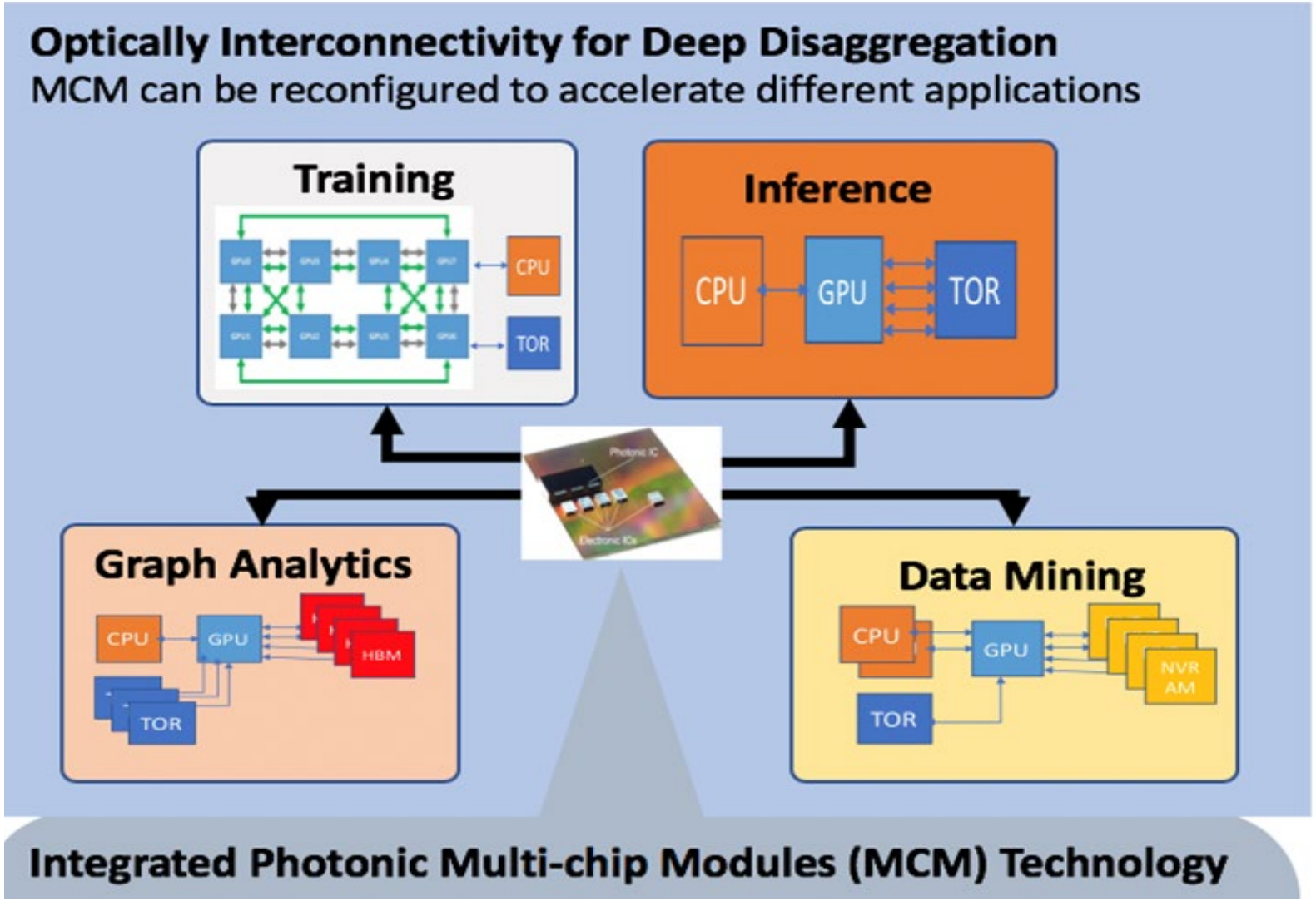




Optical HBM

[†]https://www.samsung.com/semiconductor/dram/hbm-flashbol

Deep Disaggregation: MCM photonic connectivity/switching

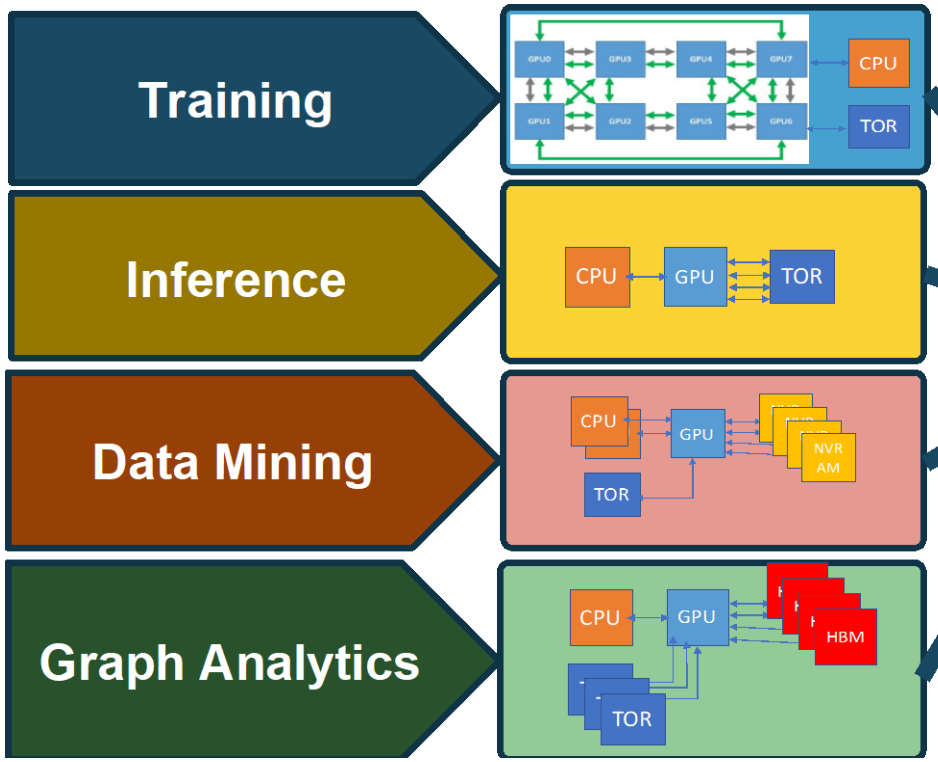


PINE

Deep Disaggregation can provide >10:1 dynamic bandwidth range

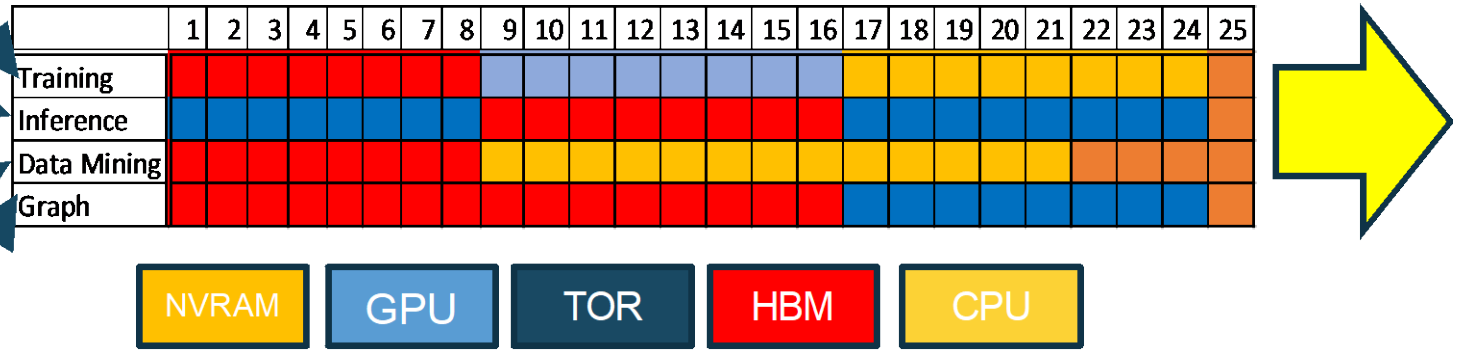
Workload

Logical Node Connectivity



Photonic MCM Connectivity Map

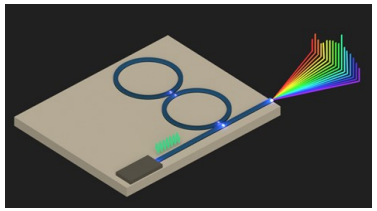
>5X Performance-per-Watt advantage
for applications with diverse node resource demands.

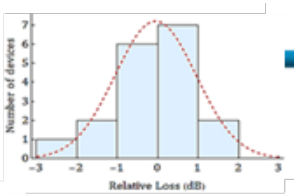


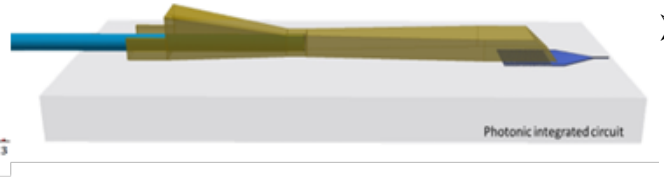
Virtual "Pin" destination for GPU MCM

Data Analytics Workloads

PINE SiP Energy Optimized Dense WDM Link

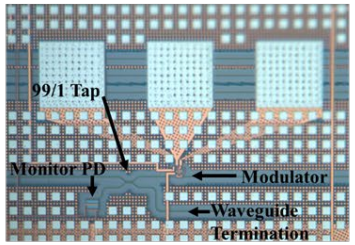


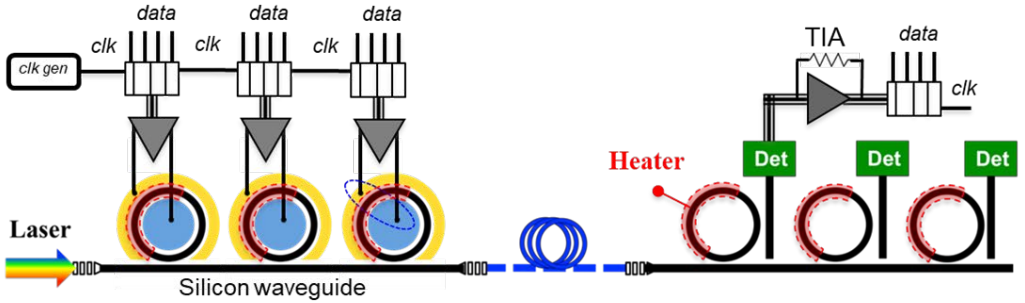




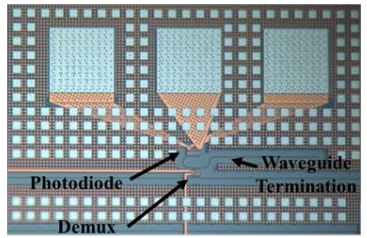
- Passive alignment high-density fiber chip-IO with **<0.6 dB** coupling penalty >**80 degrees**

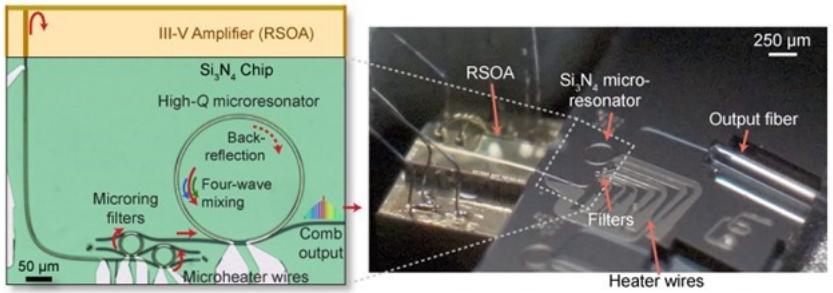
TX



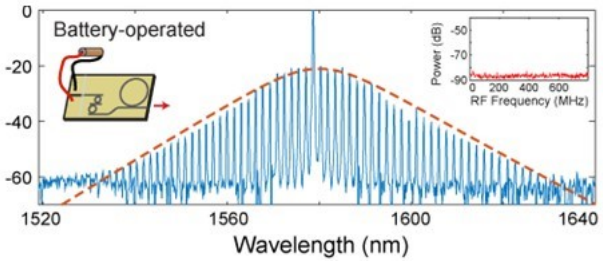


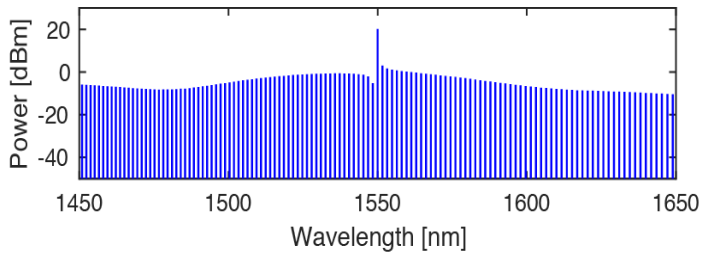
RX



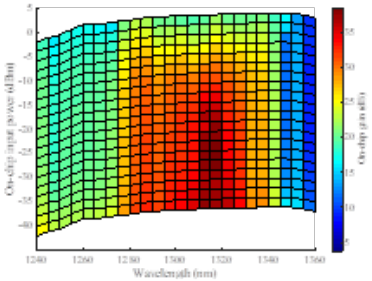


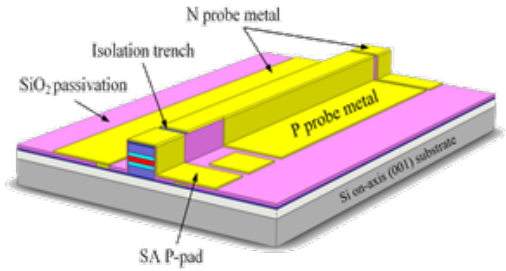
- Energy optimized first fully integrated comb generator with **45%** conversion efficiency.



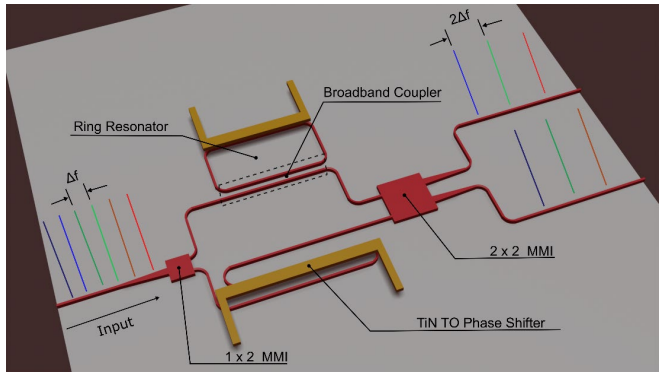


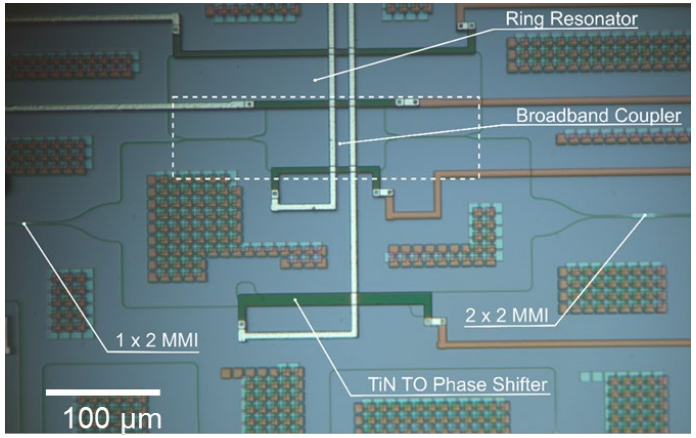
- First monolithic QD SOA on silicon – record WPE at **14.2%** and **39dB on-chip gain**.

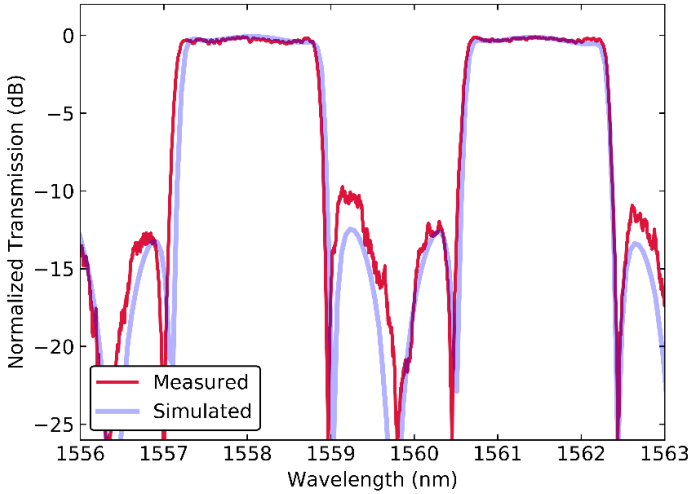




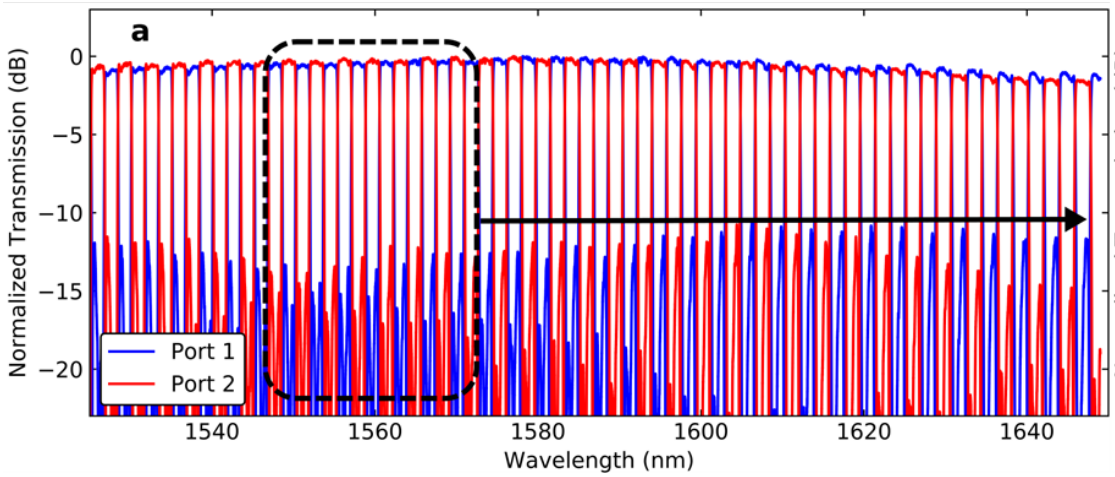
Ultra-Broadband DWDM Filtering

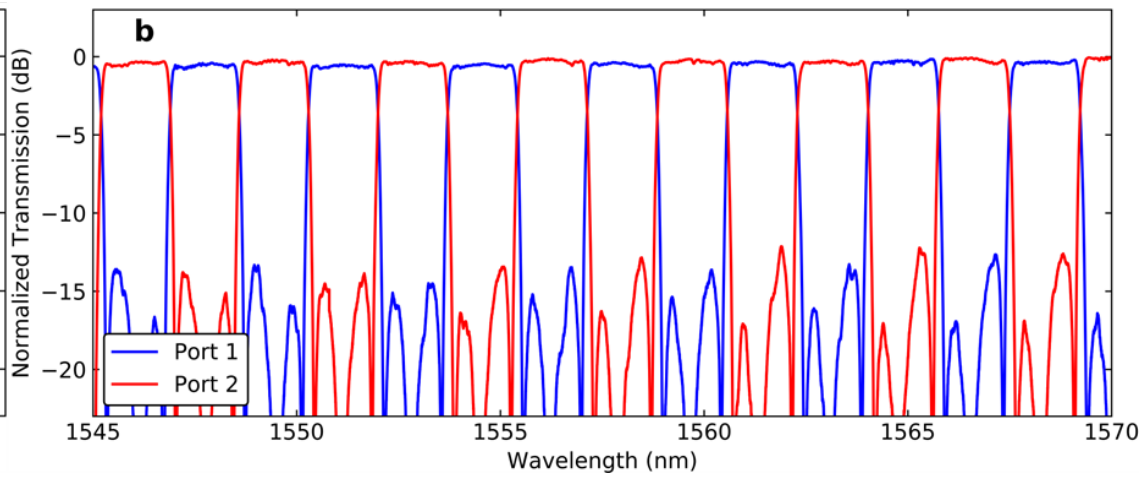






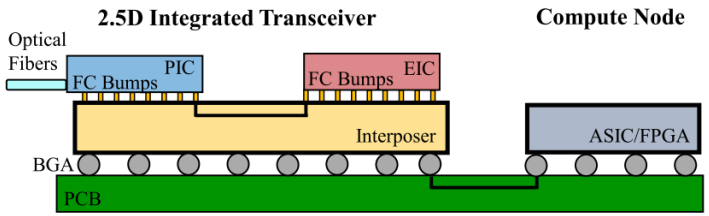
Record usable bandwidth > 125 nm; flat-top response and > 10 dB worst case extinction ratio

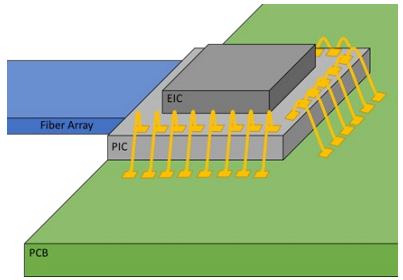


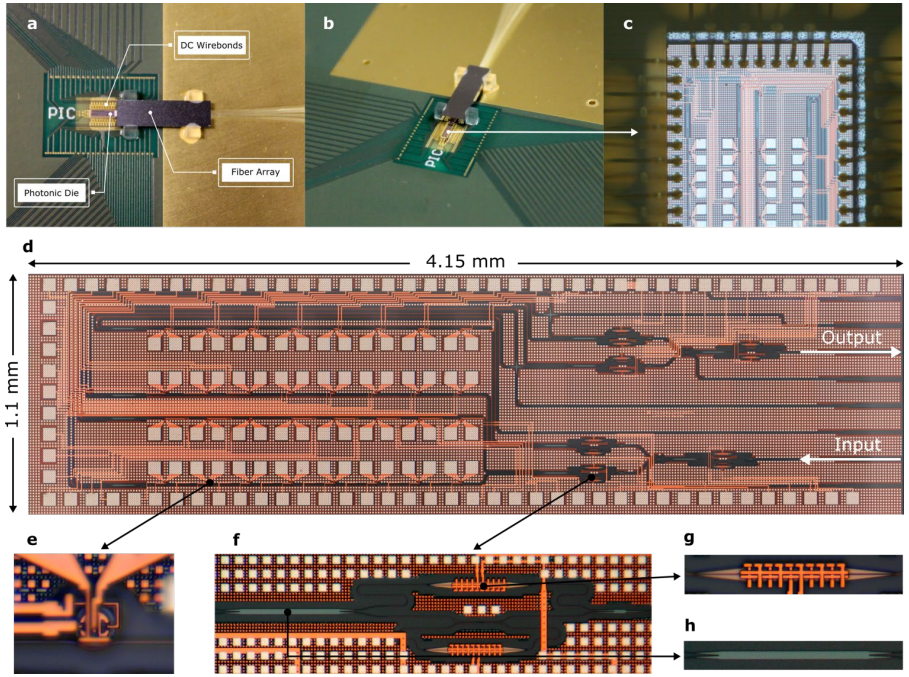


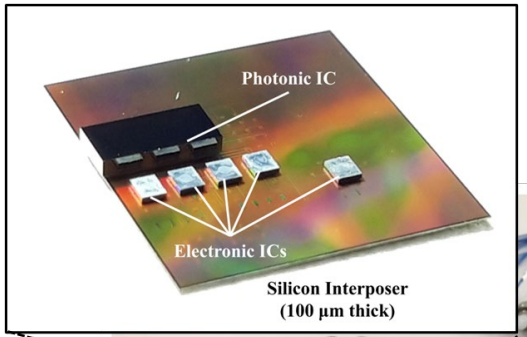
A. Rizzo et al., IEEE PTL 2020

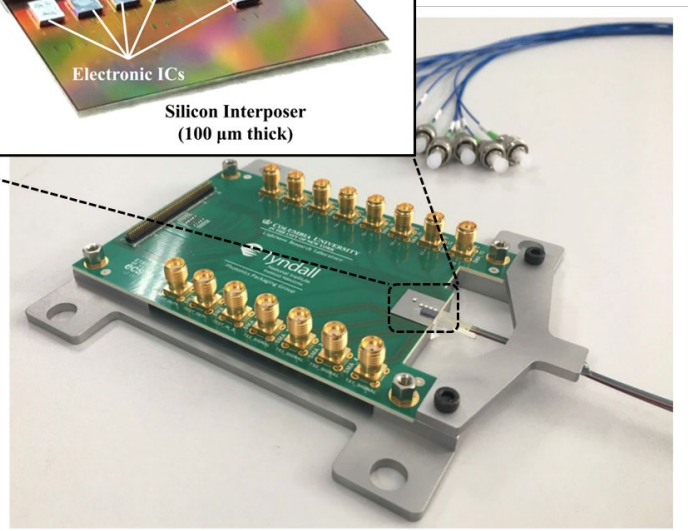
2.5D and 3D High-Density Multi-Chip Modules

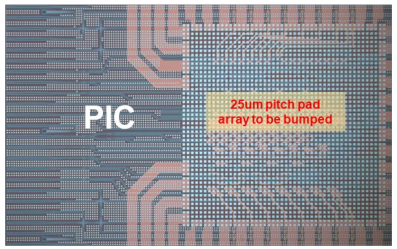


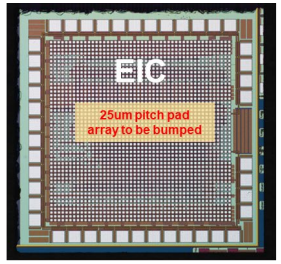


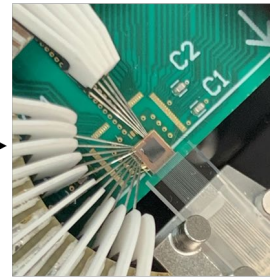












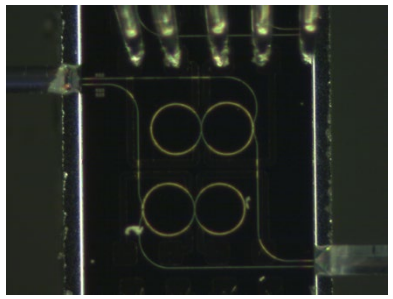
- Fully Packaged 32-channel WDM Transmitter
- 2.5 dB/facet insertion loss, open eye to 16 Gb/s

N. Abrams et al., JLT 2020

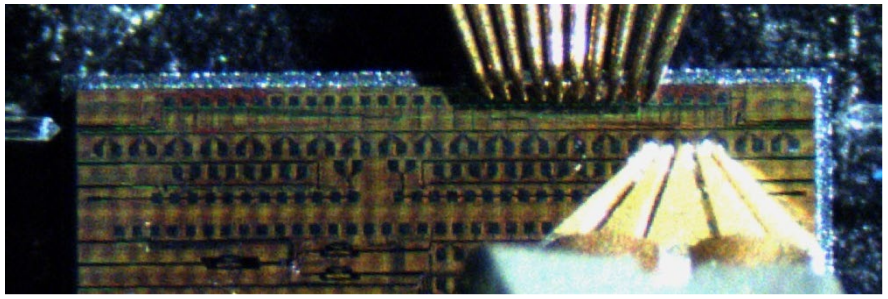
Comb Transmitter Demo

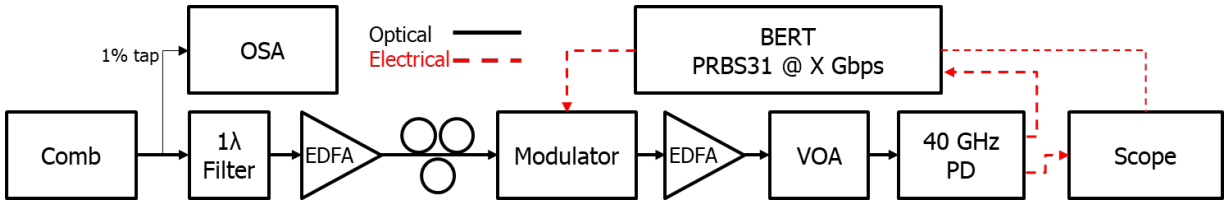
PINE

SiN Comb Chip

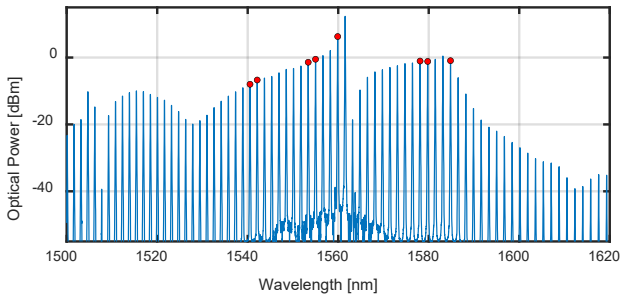


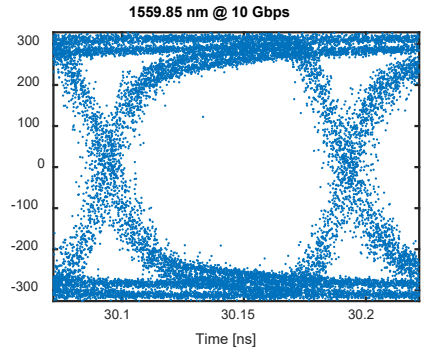
Active SiP Cascaded Modulator Chip

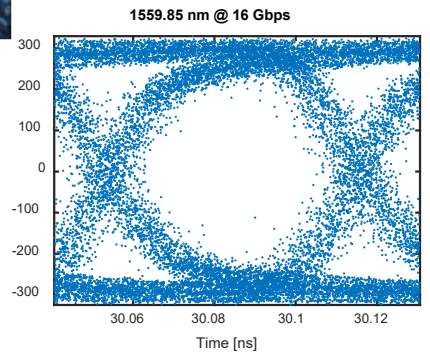


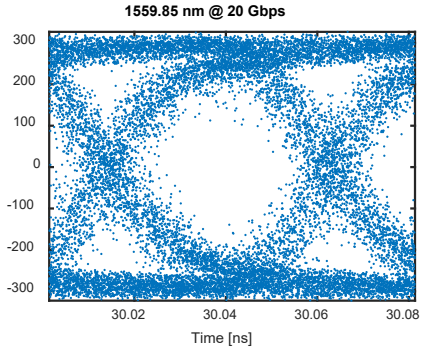


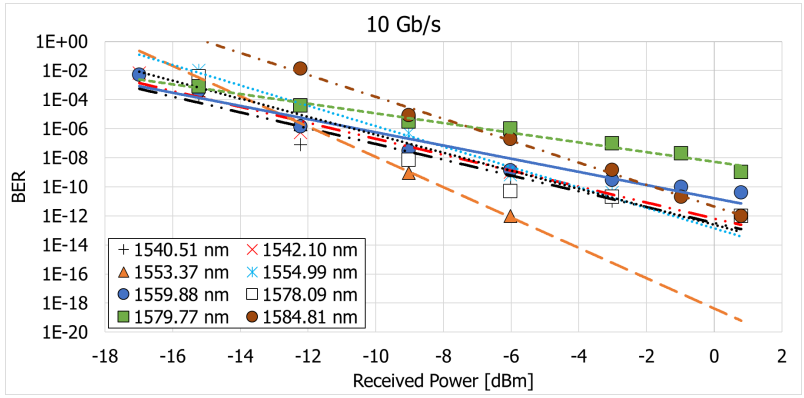
- **First** demonstration of Kerr comb with a SiP modulator
- Bus of 20 cascaded modulators demonstrate scalability

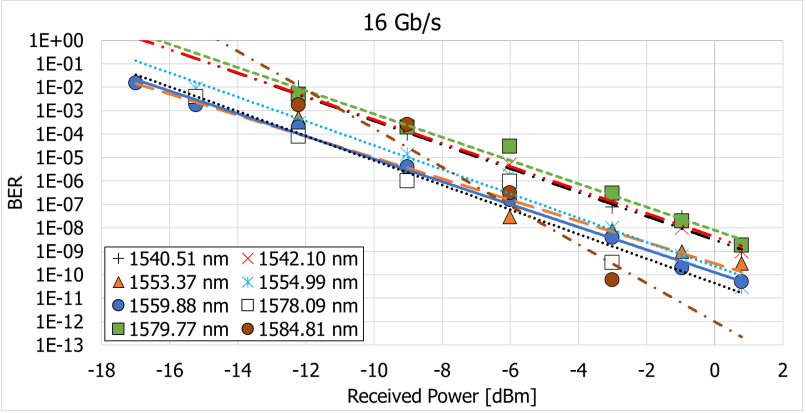


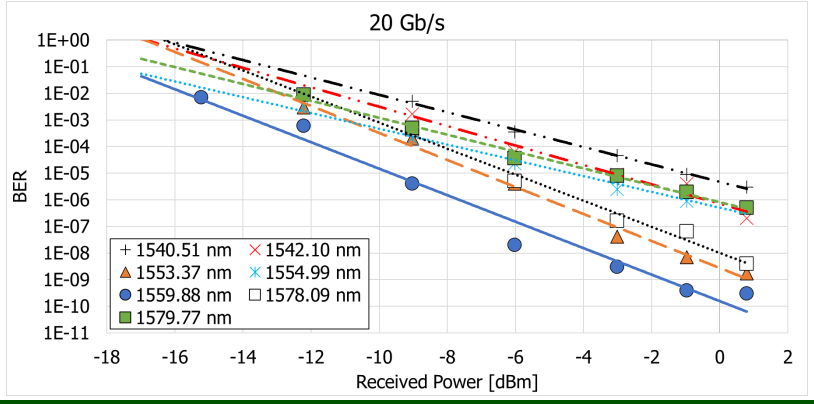












A. Novick et al., CLEO 2021 [Submitted]

PINE Phase 2 T2M

- Our **industry partners** have shown their commitment to the PINE technology and vision by continuing and increasing their involvement in the PINE Phase 2.
- **Transition Path**
  - We are working with PINE Phase 2 partners NVIDIA and LBNL to identify relevant data center and HPC applications that would benefit from the disaggregated architecture
  - **NVIDIA**
    - As a result of collaboration on PINE Phase 1, NVIDIA, in addition to their continuing collaboration on PINE Phase 2 activities, is increasing their internal silicon photonics activities.
    - This path is to build in-house capabilities for fabless design. NVIDIA sees direct use for PINE technologies in the GPU portfolio, in particular with respect to MCM integration for high bandwidth interconnects.
- **Quintessent -** start-up based on ARPA-E technology, is now a PINE Phase 2 partner continuing their contributions based on hybrid quantum dot technologies
- **Start-ups**, based on PINE technology
  - Columbia comb laser - Xcape to appear in Fast Pitch Session
  - Quintessent now PINE Phase 2 partner, UCSB quantum dot based lasers and SOAs

PINE publications (Q10 forward)

Journals

- Madeleine Glick, Nathan C. Abrams, Qixiang Cheng, Min Yee Teh, Yu-Han Hung, Oscar Jimenez, Songtao Liu et al. "PINE: Photonic Integrated Networked Energy efficient datacenters (ENLITENED Program)." IEEE/OSA Journal of Optical Communications and Networking 12, no. 12 (2020): 443-456. (invited)
- Qixiang Cheng, Jihye Kwon, Madeleine Glick, Meisam Bahadori, Luca P. Carloni, and Keren Bergman. "Silicon Photonics Codesign for Deep Learning." Proceedings of the IEEE (2020).
- Anthony Rizzo, Qixiang Cheng, Stuart Daudlin, and Keren Bergman. "Ultra-Broadband Interleaver for Extreme Wavelength Scaling in Silicon Photonic Links." IEEE Photonics Technology Letters 33, no. 1 (2020): 55-58.
- Min Yee Teh, Zhenguo Wu, and Keren Bergman. "Flexspander: augmenting expander networks in high-performance systems with optical bandwidth steering." IEEE/OSA Journal of Optical Communications and Networking 12, no. 4 (2020): B44-B54.
- Min Yee Teh, Shizhen Zhao, and Keren Bergman. "METTEOR: Robust Multi-Traffic Topology Engineering for Commercial Data Center Networks." arXiv preprint arXiv:2002.00473 (2020).
- Nathan C. Abrams, Qixiang Cheng, Madeleine Glick, Moises Jezzini, Padraic Morrissey, Peter O'Brien, and Keren Bergman "Silicon Photonic 2.5D Multi-Chip Module Transceiver for High-Performance Data Centers," Journal of Lightwave Technology 38, no. 13 (2020): 3346-3357. (invited).
- Ziyi Zhu, Giuseppe Di Guglielmo, Qixiang Cheng, Madeleine Glick, Jihye Kwon, Hang Guan, Luca P. Carloni, and Keren Bergman, "Photonic Switched Optically Connected Memory: An Approach to Address Memory Challenges in Deep Learning," Journal of Lightwave Technology, 38(10), pp.2815-2825. (invited)

Conference proceedings

- Min Yee Teh,, Yu-Han Hung, George Michelogiannakis, Shijia Yan, Madeleine Glick, John Shalf, and Keren Bergman. "TAGO: rethinking routing design in high performance reconfigurable networks." SC20
- John Shalf, George Michelogiannakis, Brian Austin, Taylor Groves, Manya Ghobadi, Larry Dennison, Tom Gray et al. "Photonic Memory Disaggregation in Datacenters." In Photonics in Switching and Computing, pp. PsW1F-5. 2020.(invited)
- Nathan C. Abrams, Madeleine Glick, and Keren Bergman. "Silicon Photonic Multi-Chip Module Interconnects for Disaggregated Data Centers." In 2020 International Conference on Optical Network Design and Modeling (ONDM), pp. 1-3. IEEE, 2020. (invited)
- Yu-Han Hung, Shijia Yan, Yiwen Shen, Ziyi Zhu, Min Yee Teh, Madeleine Glick, and Keren Bergman. "A Flexible HyperX Topology using Silicon Photonic Switching for Bandwidth Steering." Optical Interconnect Conference 2020.
- S. Liu, Y. Tong, J. Norman, M. Dumont, A. Gossard, H. Tsang and J. Bowers, "High Efficiency, High Gain and High Saturation Output Power Quantum Dot SOAs Grown on Si and applications", OFC, 2020.
- Liang Yuan Dai, Yu-Han Hung, Qixiang Cheng and Keren Bergman, "Experimental Demonstration of PAM-4 Transmission through Microring Silicon Photonic Clos Switch Fabric," OFC 2020.
- Anthony Rizzo, Qixiang Cheng, Stuart Daudlin, and Keren Bergman, "Ultra-Broadband Silicon Photonic Interleaver for Massive Channel Count Frequency Combs," CLEO 2020.
- Nathan C. Abrams, Qixiang Cheng, Madeleine Glick, Moises Jezzini, Padraic Morrissey, Peter O'Brien, and Keren Bergman, "Silicon Photonic 2.5D Integrated Multi-Chip Module Receiver," CLEO 2020.

Book Chapter

- Cheng, Qixiang, Madeleine Glick, and Keren Bergman. "Optical interconnection networks for high-performance systems." In Optical Fiber Telecommunications VII, pp. 785-825. Academic Press, 2020.