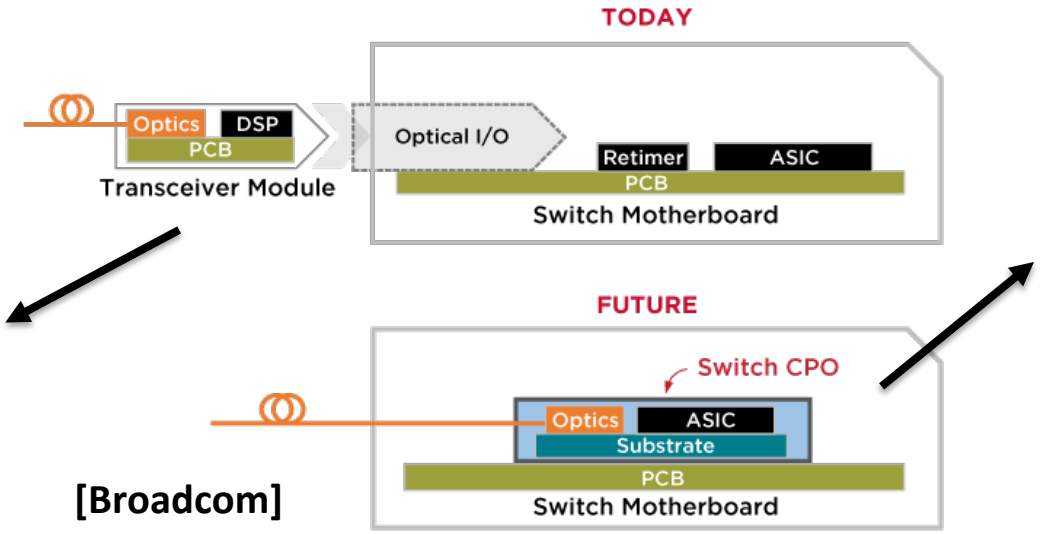
Next-generation Co-Packaged Optics for Future Disaggregated AI Systems

Sajjad Moazeni

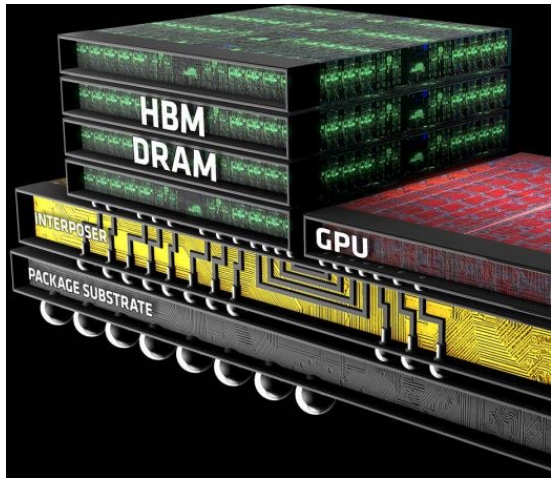
Assistant Professor

University of Washington, Seattle

Co-packaged Optics (CPO)



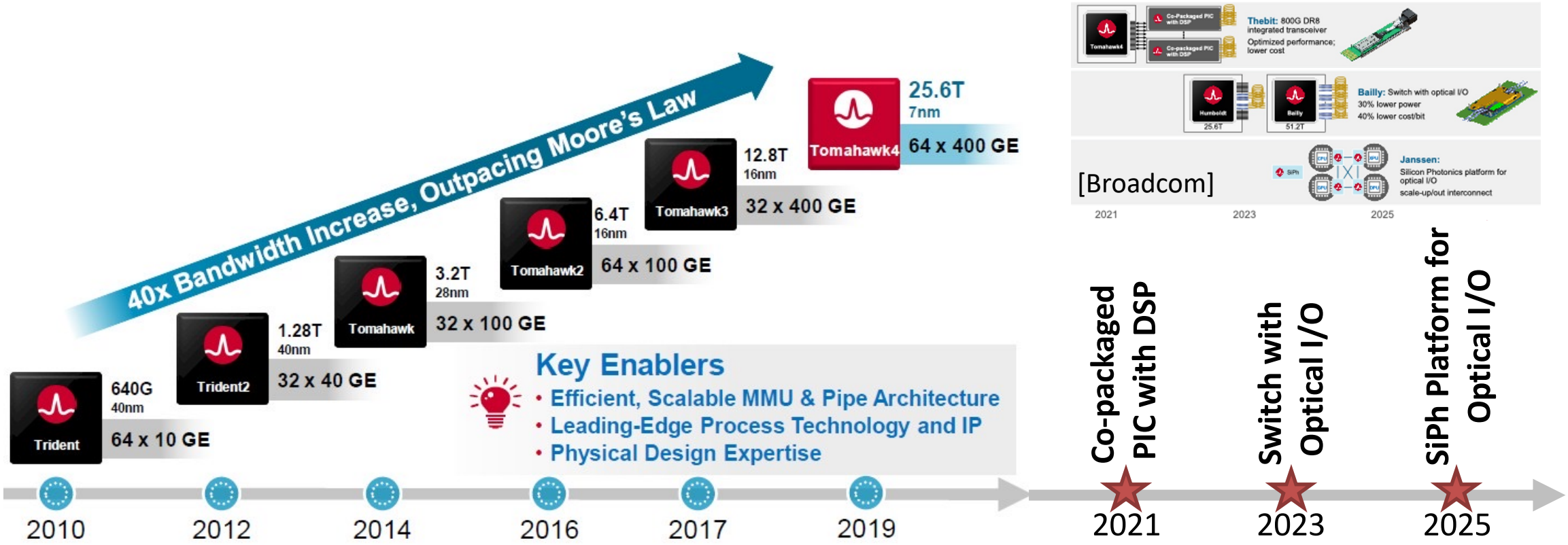
[Broadcom]



Large-scale data-center networking and switches
&
Rise of data-intensive AI/ML applications

Demands significantly larger off-package I/O bandwidths!

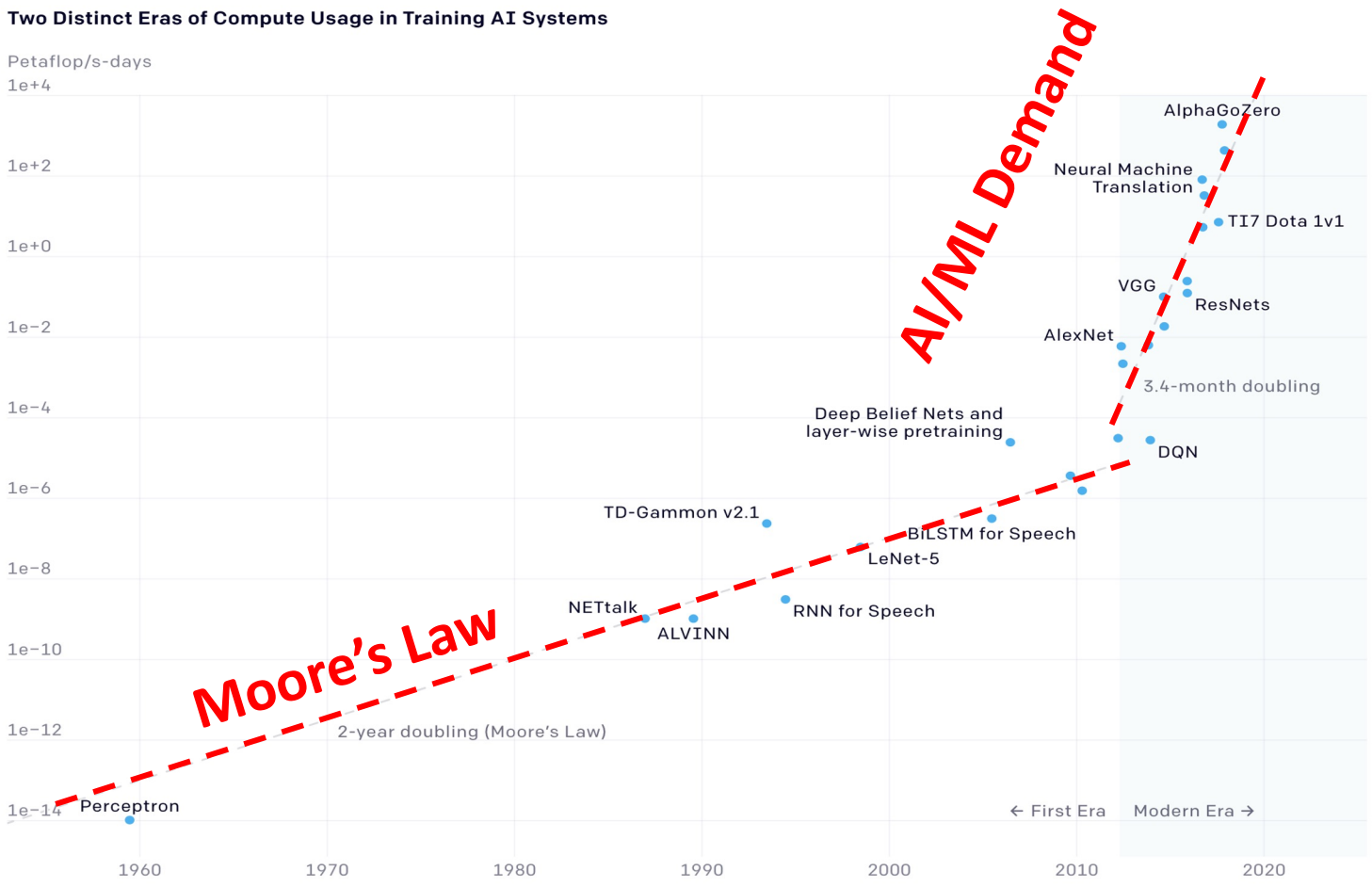
Hyper-scale Data-centers



- **Electrical Links: Energy/BW density scalability issues**
- **Future requirements: multi-Tb/s/mm & Sub-5pJ/b**

Co-packaged Optics for > 51.2Tb/s

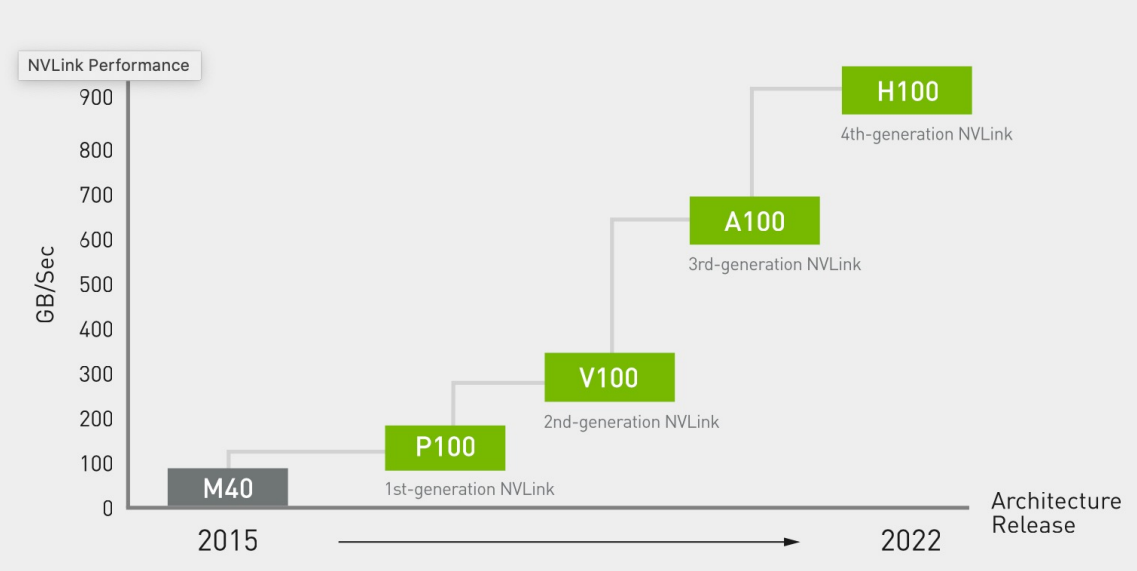
Demanded Bandwidths by AI/ML

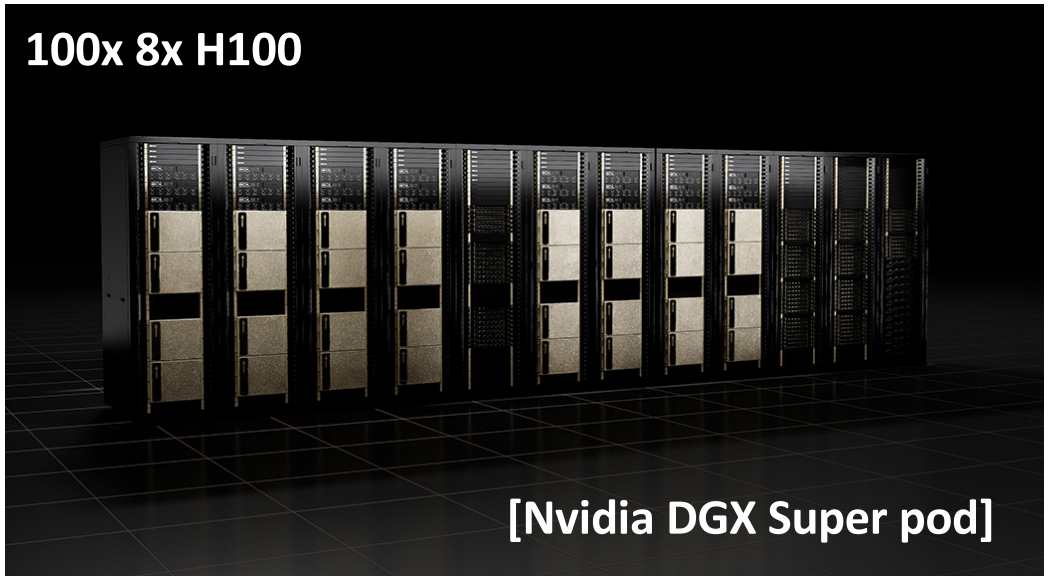


Doubling Every 3.5 Month!

[OpenAI]

Demanded Bandwidths by AI/ML





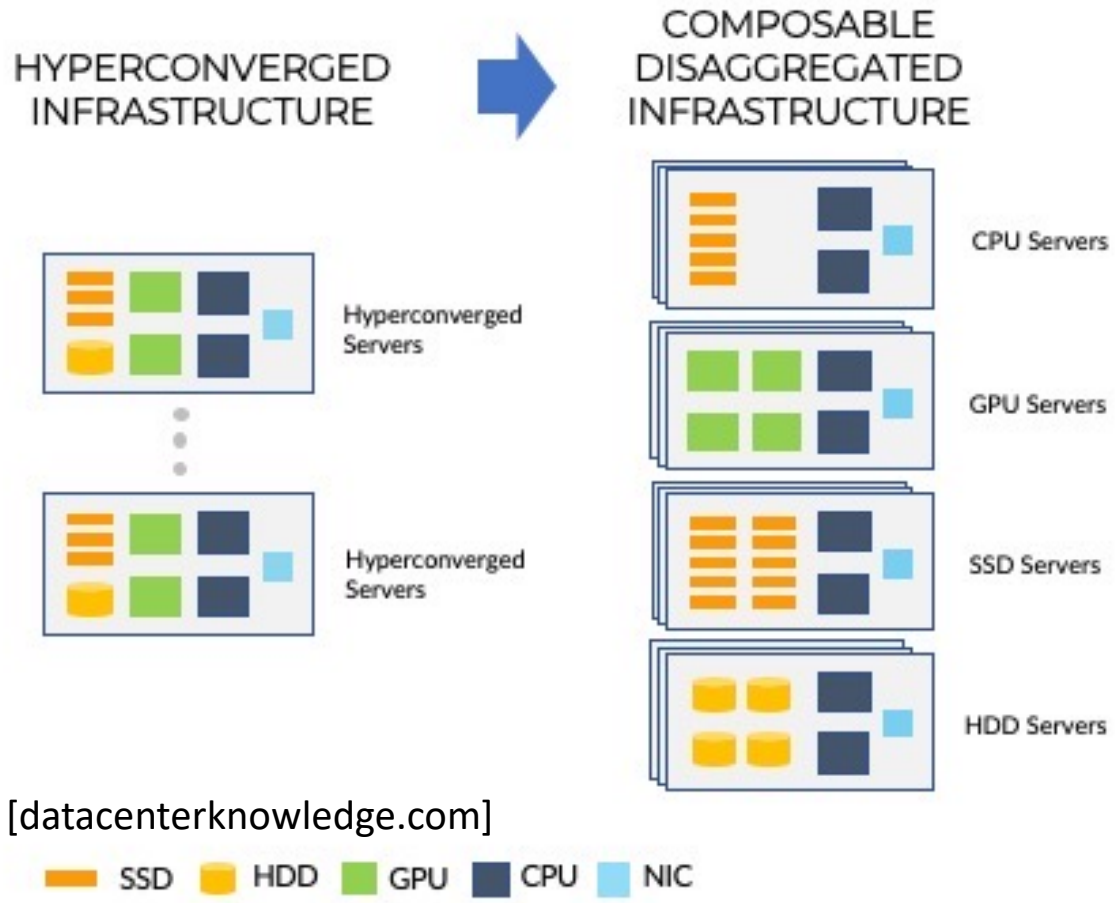
- +10Tb/s off-package bandwidths will be soon required for GPUs
- Optically connected GPU & NVSwitches is the only viable solution
- Scaling out will be also easier with optical interconnects

Disaggregated Compute Systems

- Data-center:
  - Improved utilization and dynamic resource allocations
- AI/ML Workloads:
  - Distributed Training and Parallelism

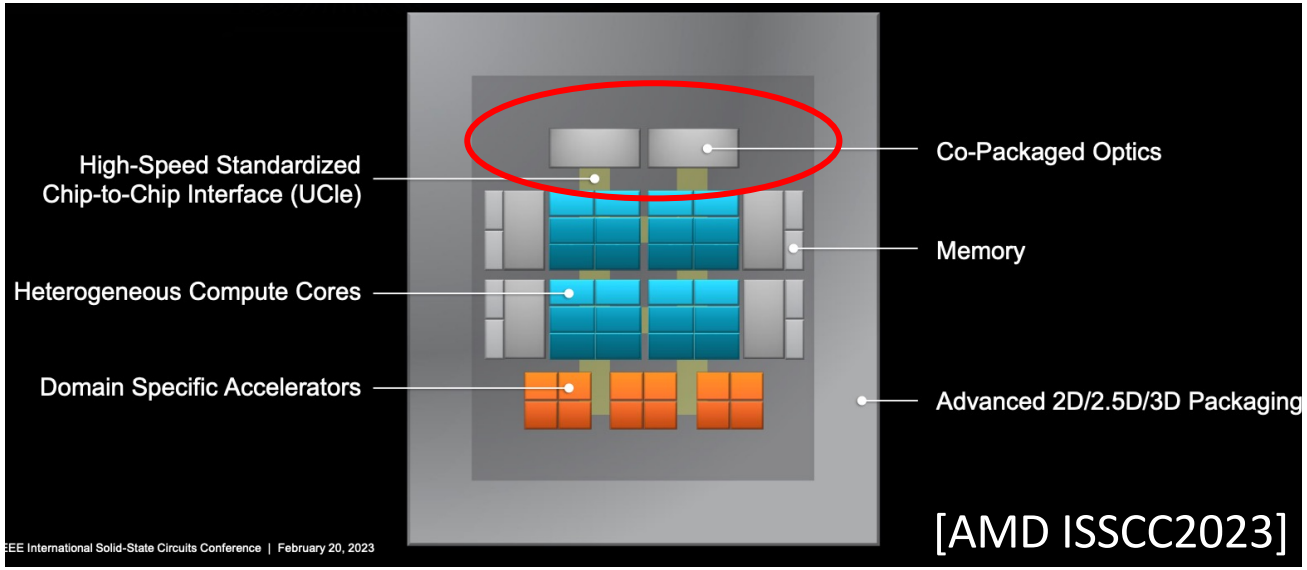
Technology Enablers:

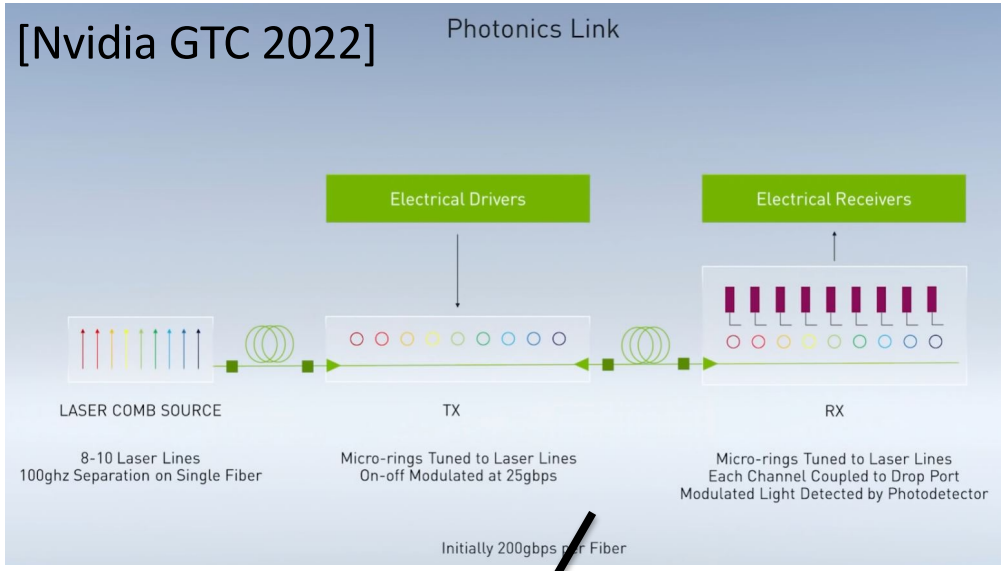
- NVMe-over-fabric, CXL, …
- DPU, etc.
- **<u>Co-packaged Optics (CPO)</u>**



[datacenterknowledge.com]

Goals for Co-packaged Optics (CPO)





Goal for Co-Packaged DWDM

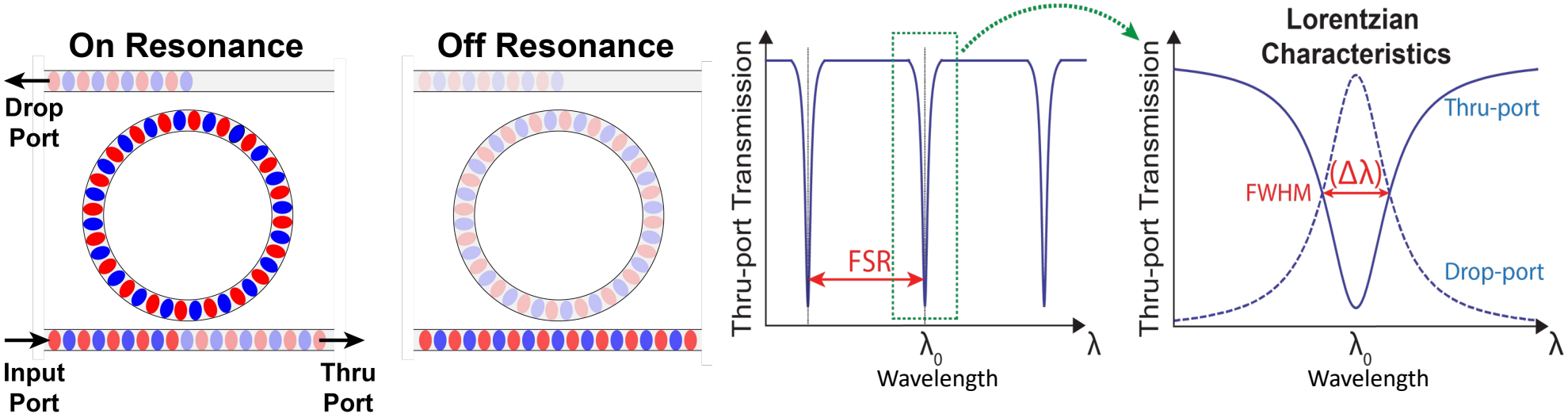
	IPoser	PCB	CPO	Cable	AOC	
Power	10^{-13}	$5x10^{-12}$	10^{-12}	$5x10^{-12}$	10^{-11}	J/b
Cost	10^{-15}	10^{-13}	10^{-10}	10^{-10}	10^{-9}	$-s/b
Density	10^{13}	$5x10^{11}$	$2x10^{12}$	$5x10^{10}$	10^{11}	b/s-mm
Reach	.005	0.5	100	5	100	m

[Dally OFC 2022]

Lower power than cable with comparable cost
Density higher than PCB
Reach comparable to AOC

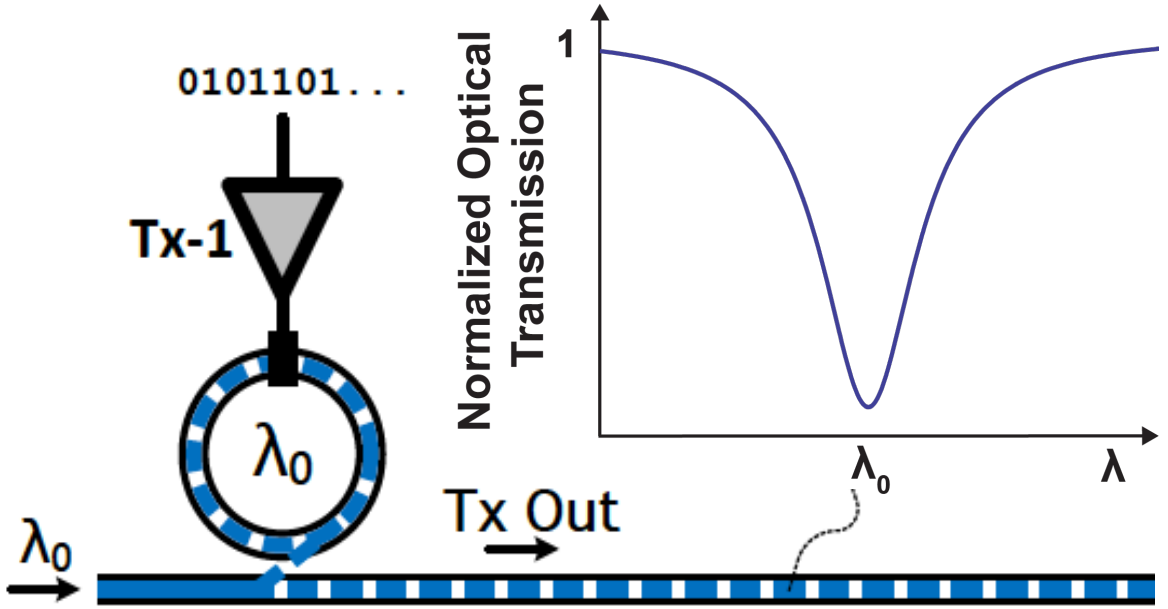
- Silicon Photonics
- Micro-ring resonator (MRM) based optical transceivers (TRx)
- Wavelength division multiplexing (WDM)

Micro-ring Modulator (MRM)



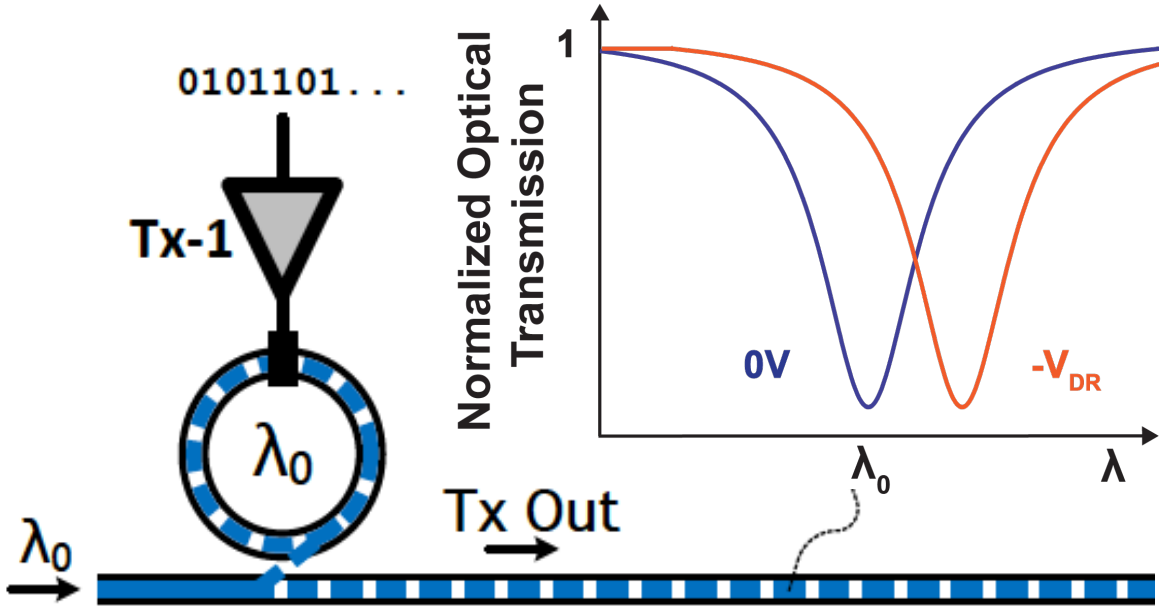
- Resonance wavelength: $\lambda_0 = n_{eff} L/m$, m = 1,2,3,...
  - Q-factor: $Q = \lambda_0 / \Delta\lambda$
- Compact device (radius of 5µm)
  - Energy & area efficient modulator/filter, Only a 20fF load!
- Supporting wavelength division multiplexing (WDM)

MRM based Optical Links



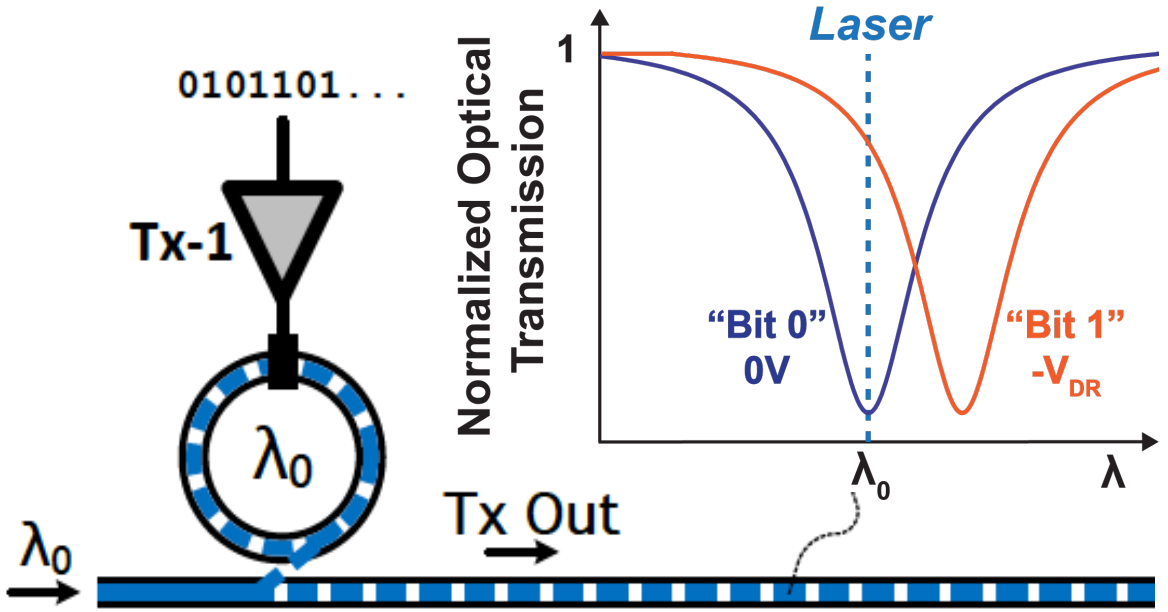
- Modulation Scheme:
  1. Deplete/Inject carriers using PN junctions

MRM based Optical Links



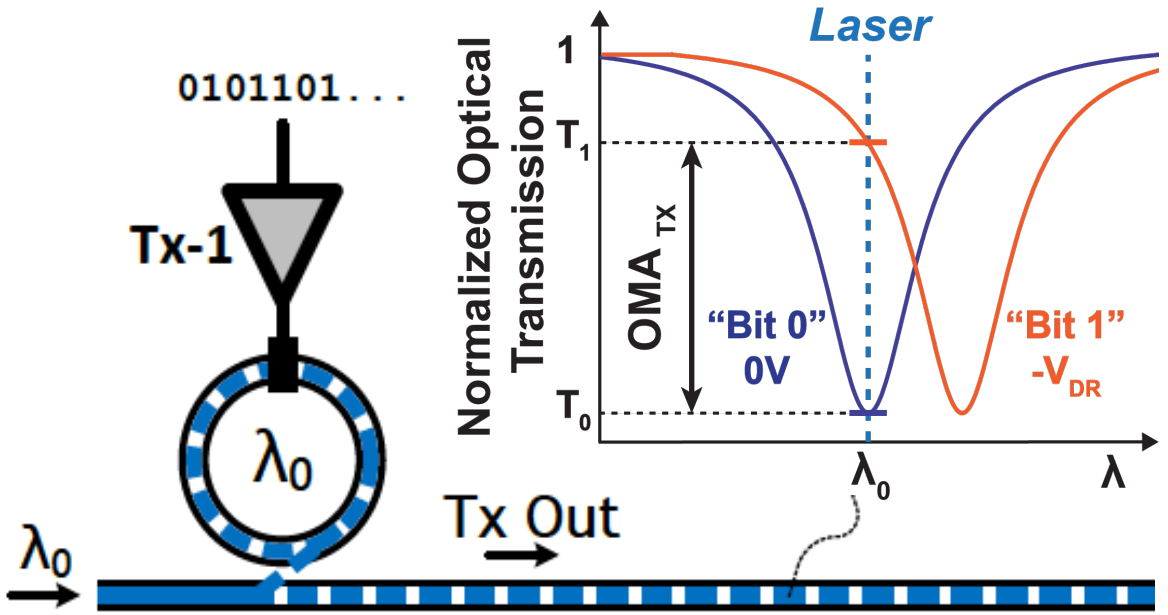
- Modulation Scheme:
  1. Deplete/Inject carriers using PN junctions
  2. Δfree carriers → Δindex of refraction [Carrier-Plasma Effect]

MRM based Optical Links



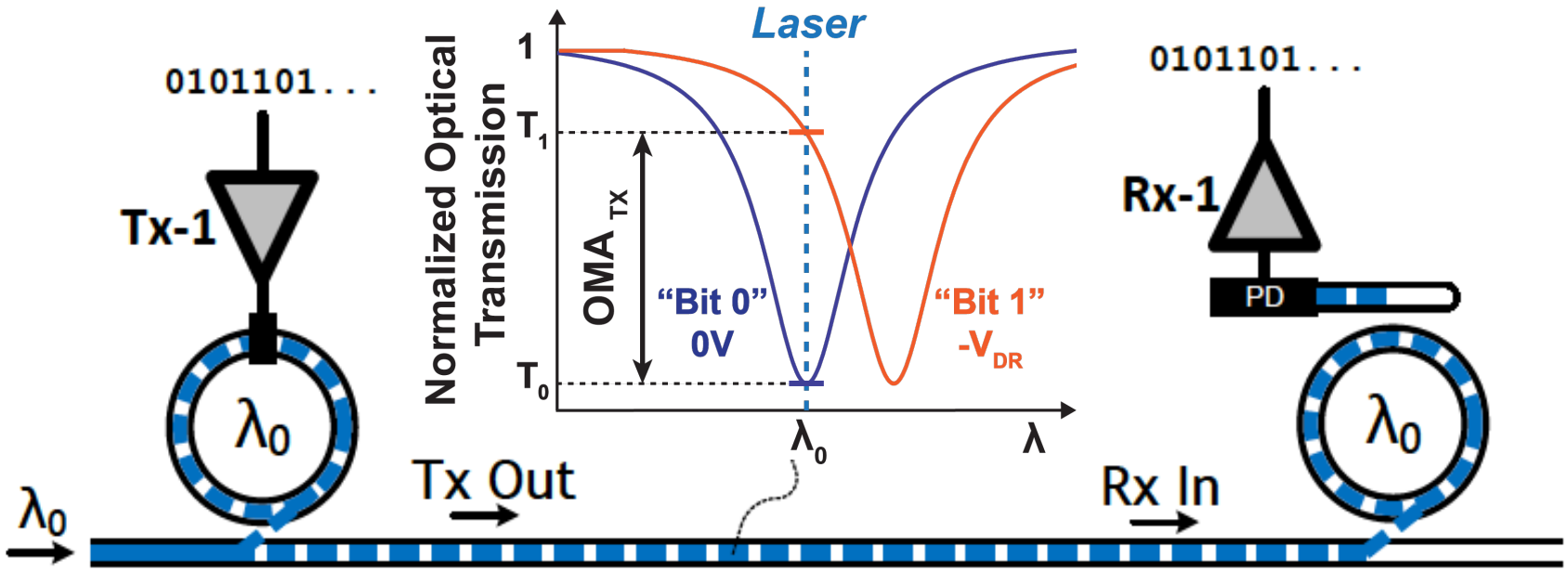
- Modulation Scheme:
  1. Deplete/Inject carriers using PN junctions
  2. Δfree carriers → Δindex of refraction [Carrier-Plasma Effect]
  3. On-Off Keying (OOK) modulation

MRM based Optical Links



- Modulation Scheme:
  1. Deplete/Inject carriers using PN junctions
  2. Δfree carriers → Δindex of refraction [Carrier-Plasma Effect]
  3. On-Off Keying (OOK) modulation
  *. **OMA**: Optical Modulation Amplitude

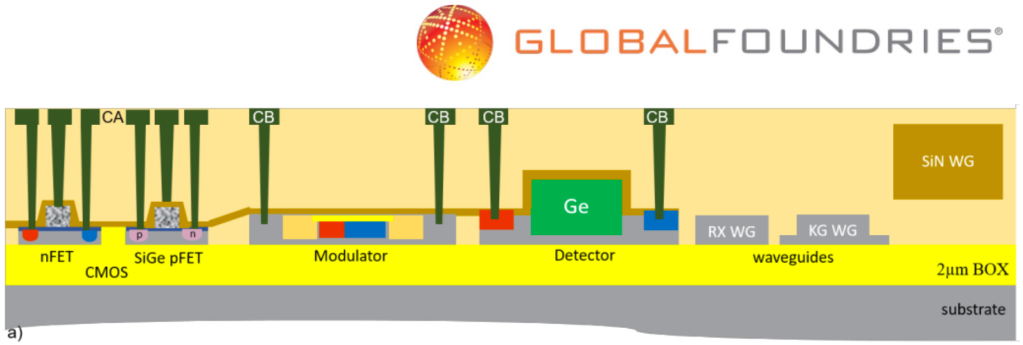
MRM based Optical Links



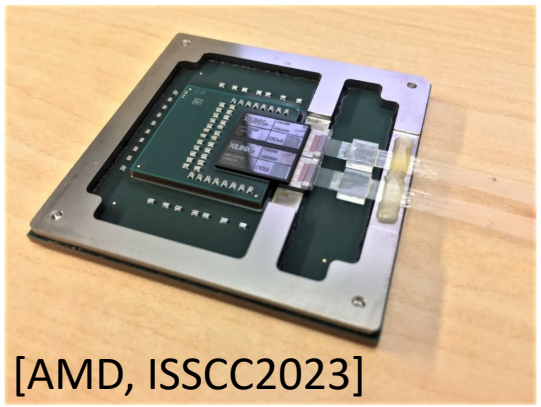
- Modulation Scheme:
  1. Deplete/Inject carriers using PN junctions
  2. Δfree carriers → Δindex of refraction [Carrier-Plasma Effect]
  3. On-Off Keying (OOK) modulation
  *. **OMA**: Optical Modulation Amplitude

Electronic-Photonic Integration

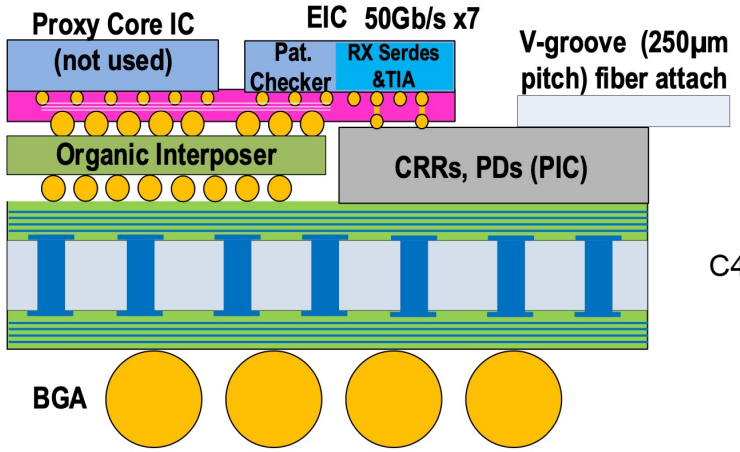
- Electronic-Photonic Integration
  - Monolithic (Low Parasitic, Old CMOS) vs. 2.5D/3D (Large Parasitic, Advanced CMOS)
- Laser integration: off-package or integrated with silicon photonics

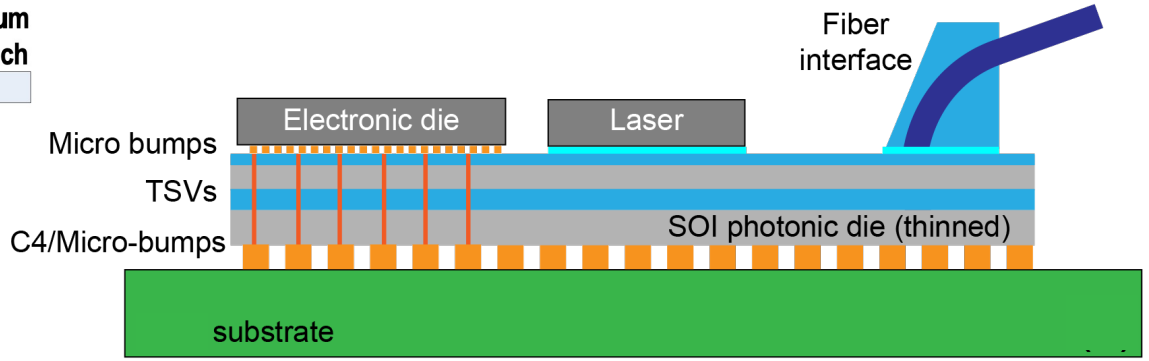


45nm SOI CMOS + Photonics

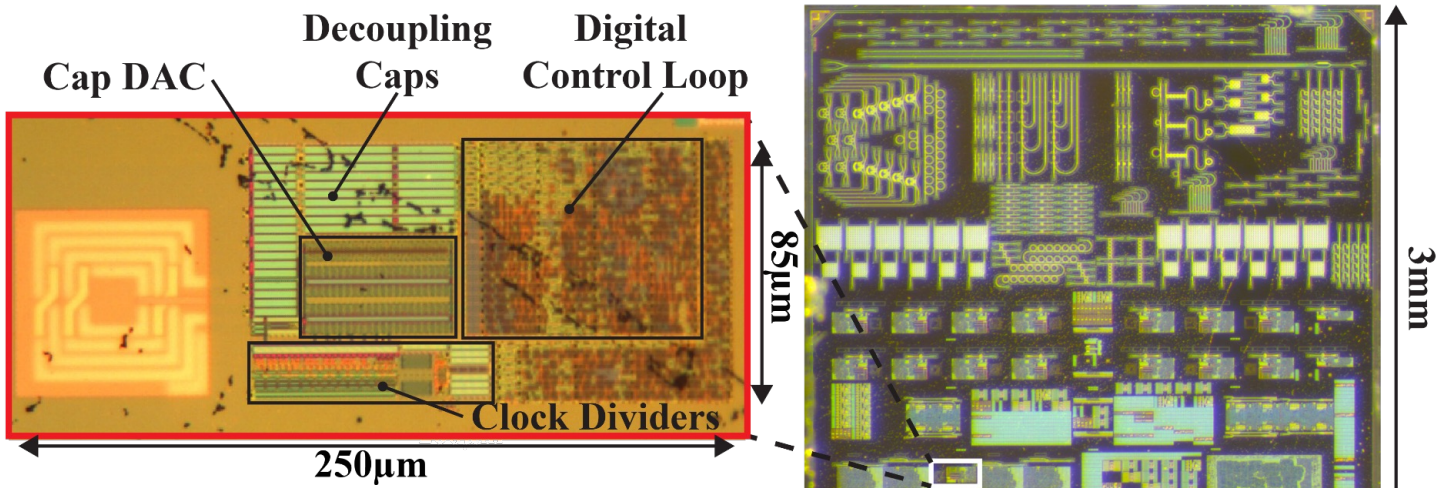


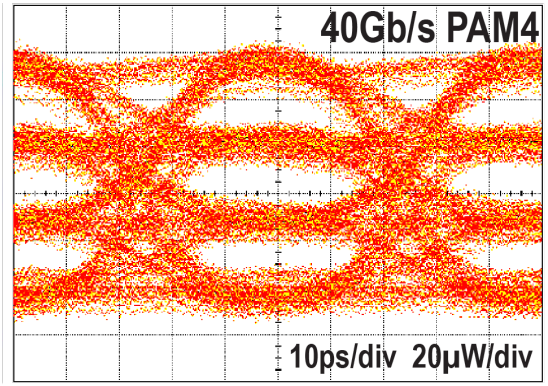
[AMD, ISSCC2023]

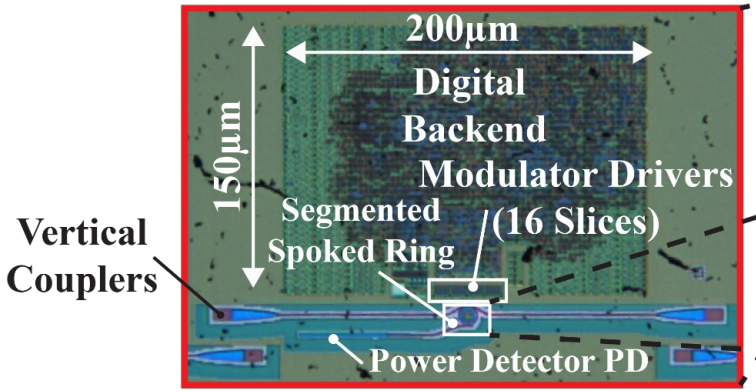


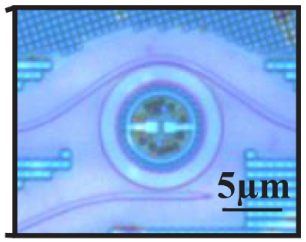


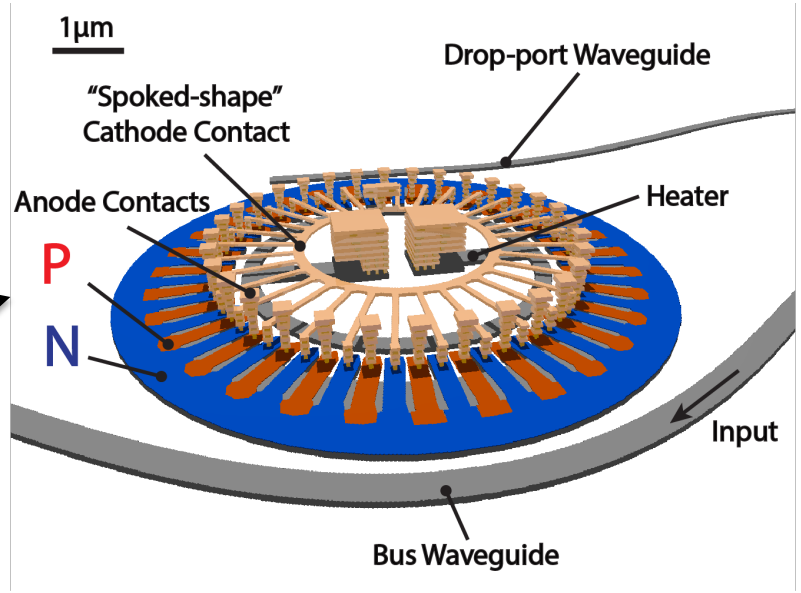
40Gb/s PAM-4 Tx in GF 45nm











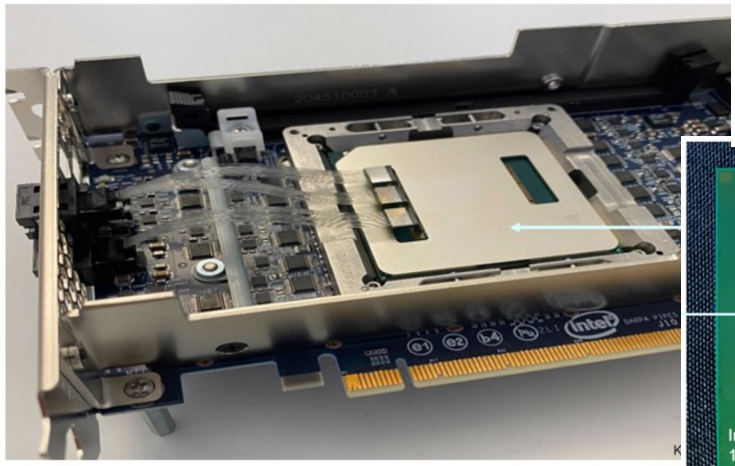
[Moazeni *et al.*, ISSCC/JSSC 17]

State-of-the-Art of CPO

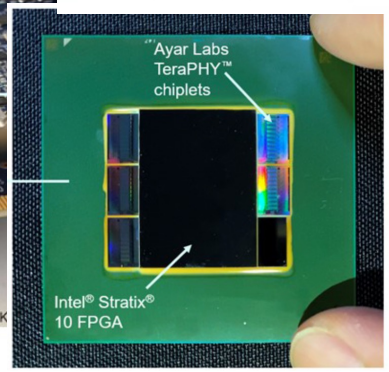
Ranovus + AMD

- Total BW per fiber: 1Tb
- +10pJ/b for 100Gb/s per wavelength
- Integrated Lasers
- 0.5Tb/s/mm

Ayar Labs + Intel



CW WDM MSA



- Total BW per fiber: 800Gb
- +5pj/b for 20Gb/s per wavelength
- External Laser
- 0.5Tb/s/mm

Major Challenges

- **Electronic-Photonic Integration**
  - Monolithic vs. 2.5D/3D
  - Interconnect parasitics > 3x Device cap
- **Energy-efficiency**
  - Need 5-10x improvement
  - Electronic-photonic Co-design
- **Fiber Packaging**
- **DWDM Laser Sources**

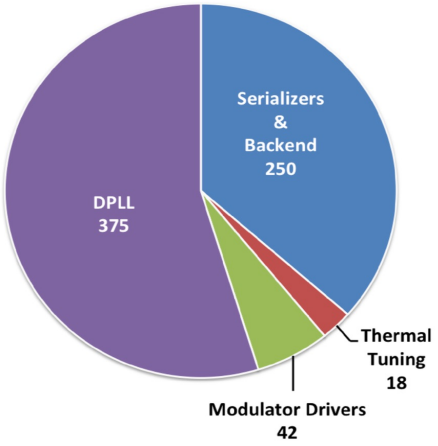
Goal for Co-Packaged DWDM

	IPoser	PCB	CPO	Cable	AOC	
Power	10^{-13}	$5x10^{-12}$	10^{-12}	$5x10^{-12}$	10^{-11}	J/b
Cost	10^{-15}	10^{-13}	10^{-10}	10^{-10}	10^{-9}	$-s/b
Density	10^{13}	$5x10^{11}$	$2x10^{12}$	$5x10^{10}$	10^{11}	b/s-mm
Reach	.005	0.5	100	5	100	m

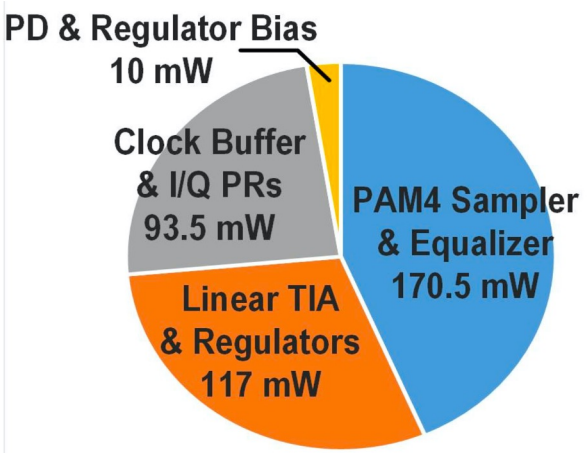
[Dally OFC 2022]

Lower power than cable with comparable cost
Density higher than PCB
Reach comparable to AOC

40Gb/s PAM-4 Tx (0.5pJ/b)
[Moazeni *et al.*, ISSCC17]

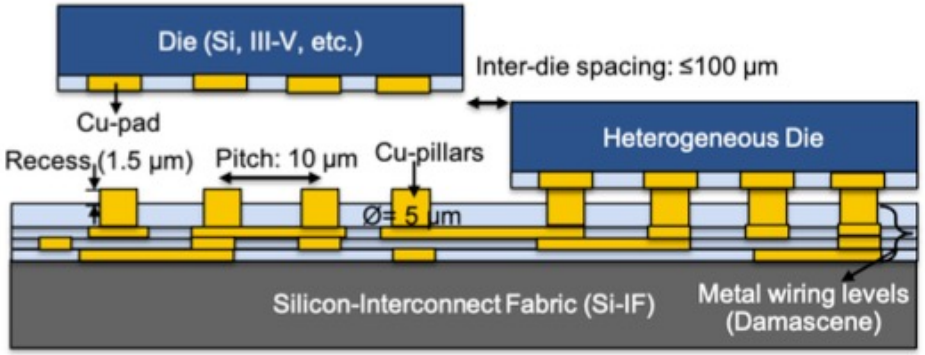


100Gb/s PAM-4 Rx (3.9pJ/b)
[Li *et al.*, ISSCC21]

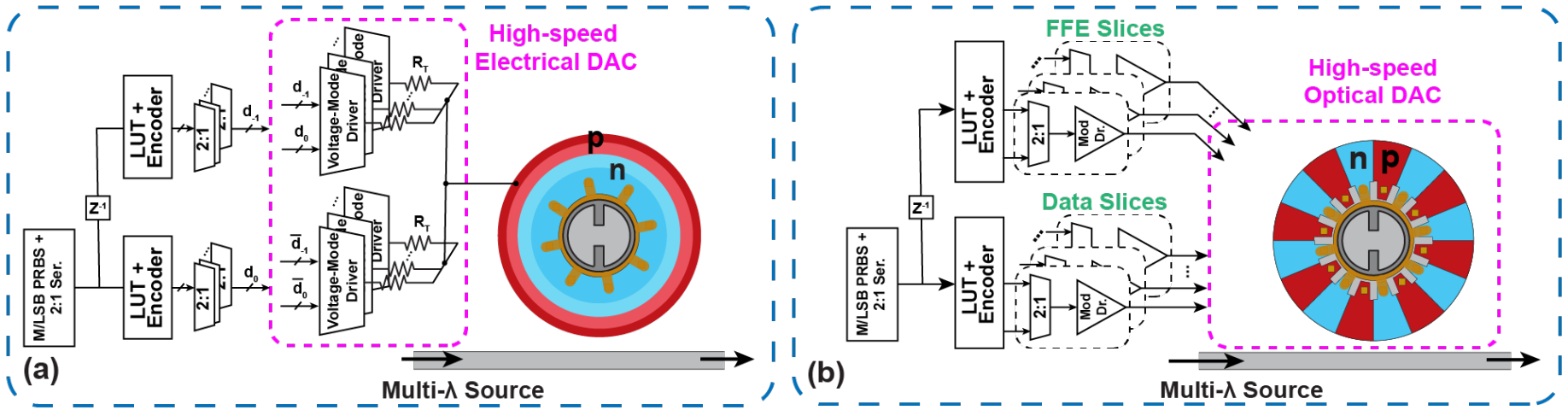


Solutions for Next-generation CPO

- Advanced packaging with direct bonding instead of micro-bumps
- Electronic-photonic Co-design (e.g., Optical DAC)



Silicon-Interconnect Fabric (SiF), UCLA 2021

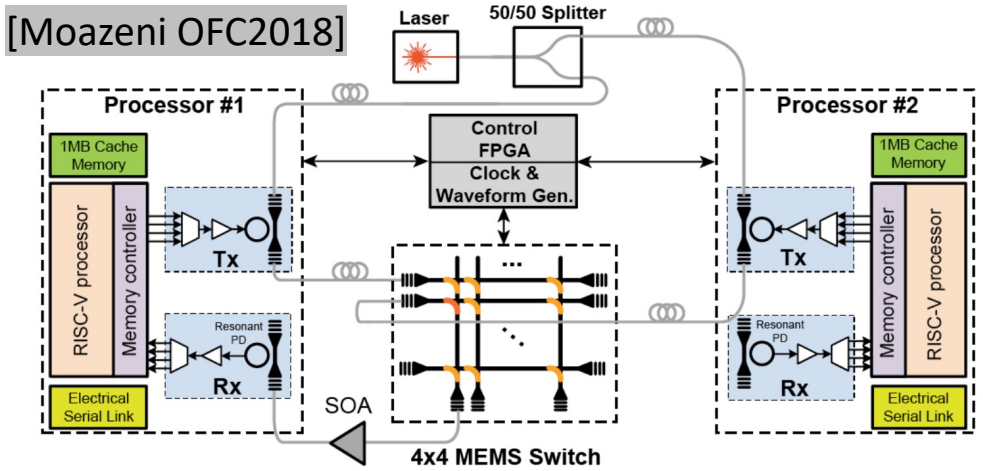


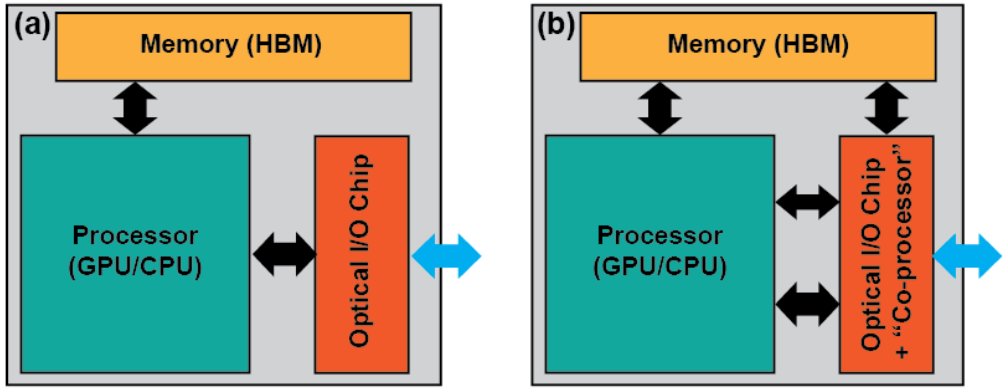
Beyond Just an E/O Bridge

New architectures will be unlocked with CPO …

- *Network-level:*
  - Micro-second optical circuit switching networks
- *Package-level:*
  - Co-processing on the CPO
  - HBM memory access & controller

[Moazeni OFC2018]

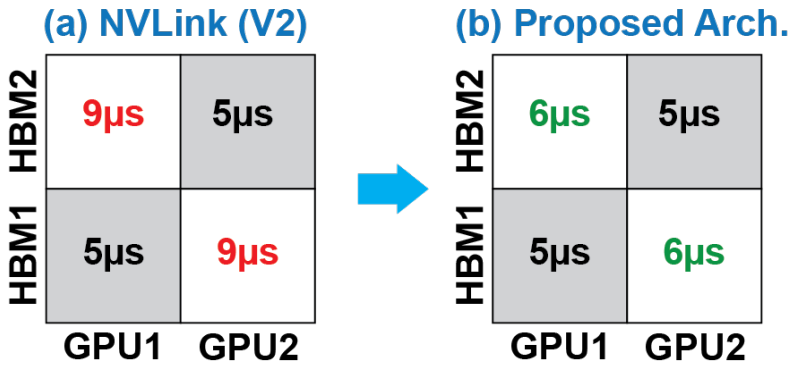




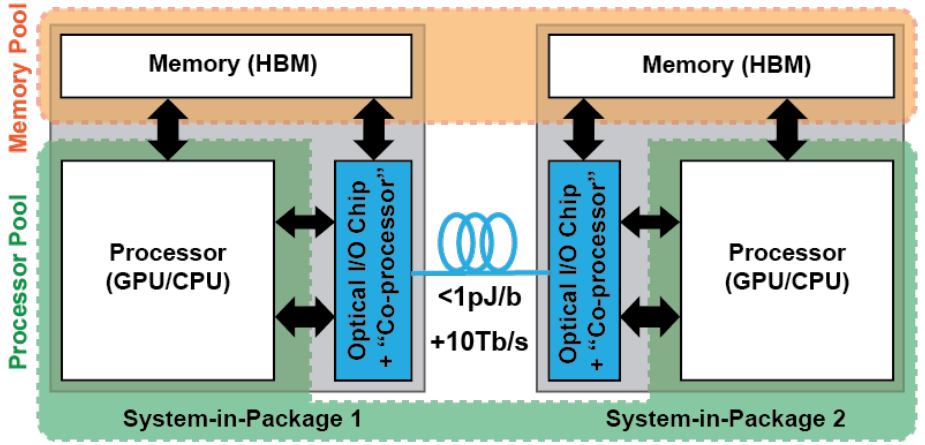
CPOs with Memory Controller

- Bypassing GPU for memory access
  - Add CXL memory controller
- Improving GPUDirect latency

Ultimate CPO-enabled Architecture:



[Ang Li, PNNL 2019]



Conclusion

- Co-packaged Optics can provide the needs of next generation of GPU/Accelerator interconnects
- Next-generation CPO demands +1Tb/s at 1pJ/b
  - Advanced electronic-photonic integration & packaging and co-design
- Co-packaged Optics can bring new opportunities to rearchitecture GPU/Accelerator compute nodes & clusters
  - Disaggregation down to the package level