

Empowering flexible and scalable high performance architectures with embedded photonics

Keren Bergman

*Lightwave Research Lab, Columbia University
New York, NY, USA*

IPDPS 2018



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)



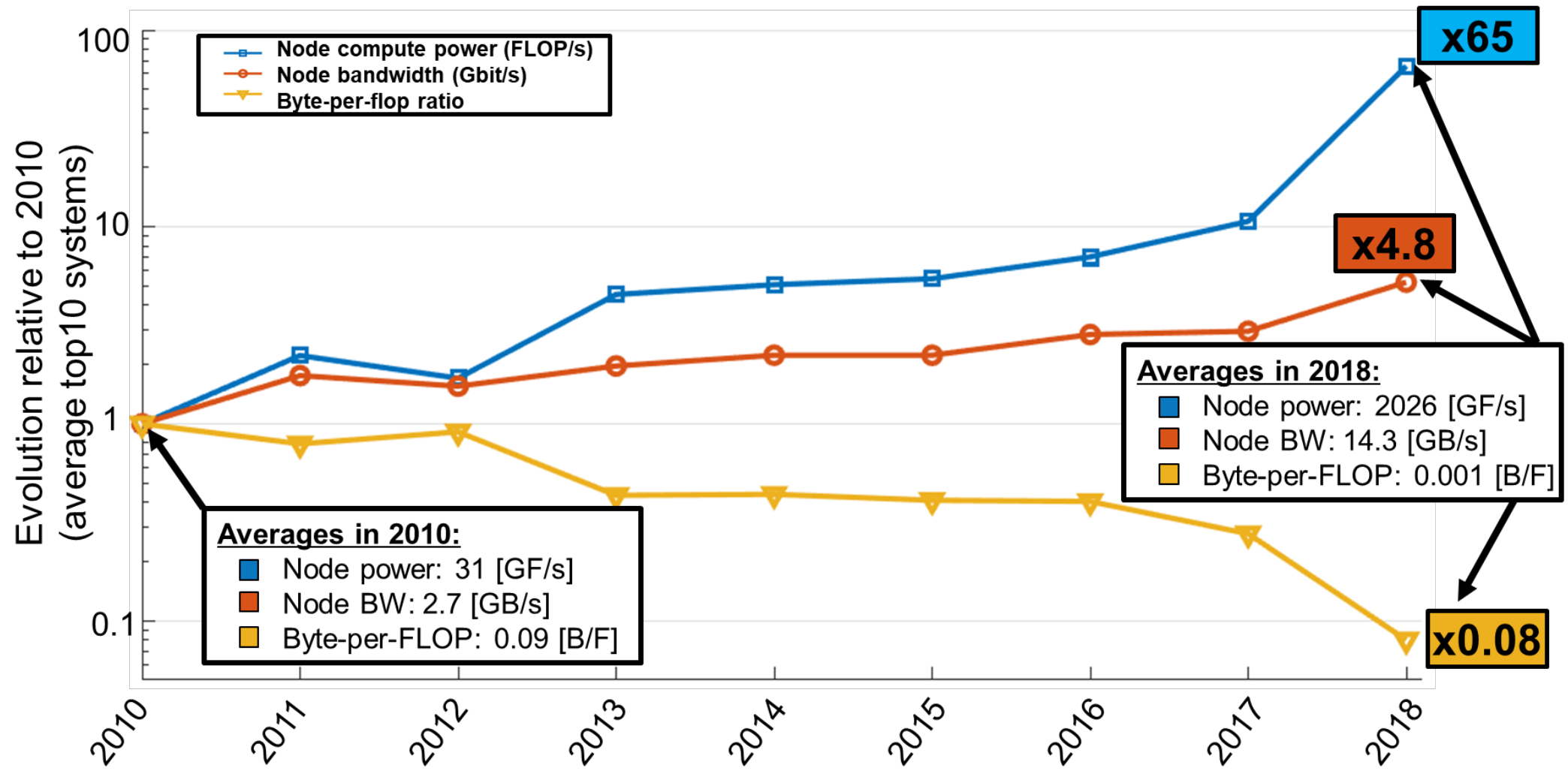
Next challenge:

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

	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
Gyokou	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

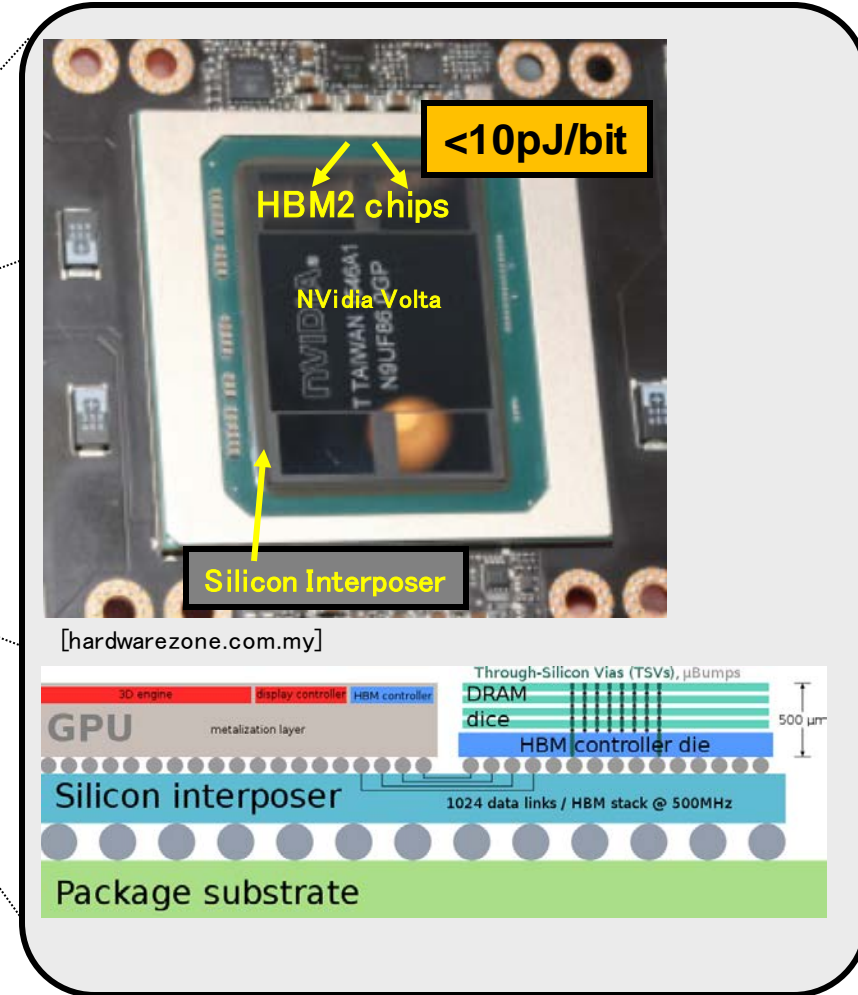
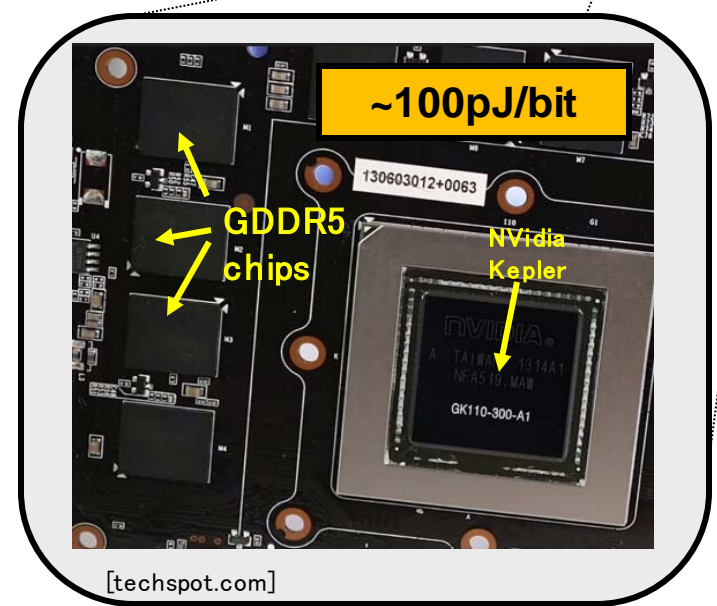
November 2017		
Name	Top500 rank	GFlop/W
Shoubu B	259	17.0
Suiren2	307	16.8
Sakura	276	16.7
DGX Volta	149	15.1
Gyokou	4	14.2
TSUBAME3.0	13	13.7
AIST AI Cloud	195	12.7
RAIDEN	419	10.6
Wilkes-2	115	10.4
Piz Daint	3	10.4
Reedbush-L	291	10.2
GOSAT-2	319	9.8
	35	9.5
DGX Saturn V	36	9.5
Era-AI	109	8.6
Reedbush-H	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



NVidia major new design



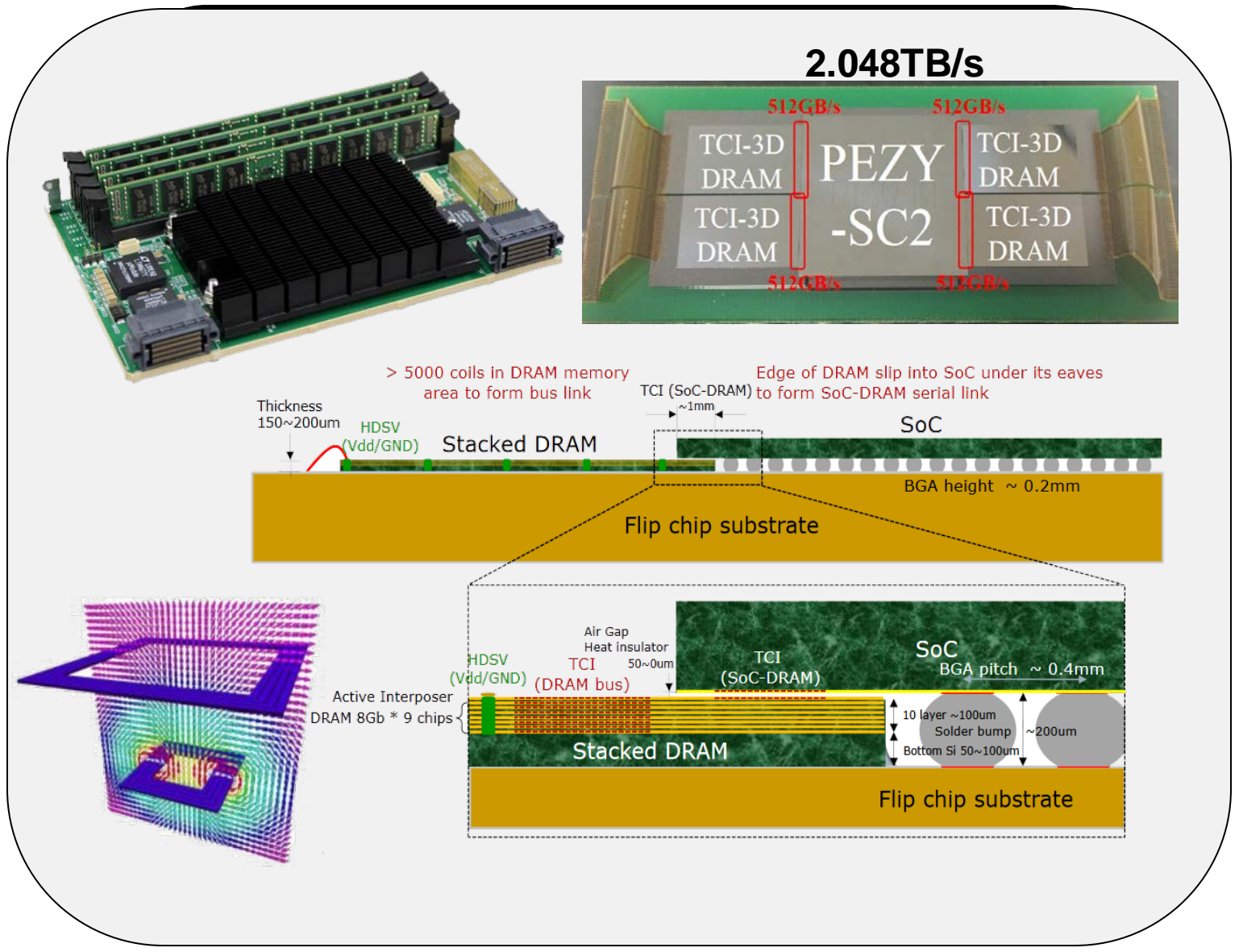
- Memory closer to GPU
- CoWoS: Chip on wafer on Substrate

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



High Performance Data Centers: Convergence on AI

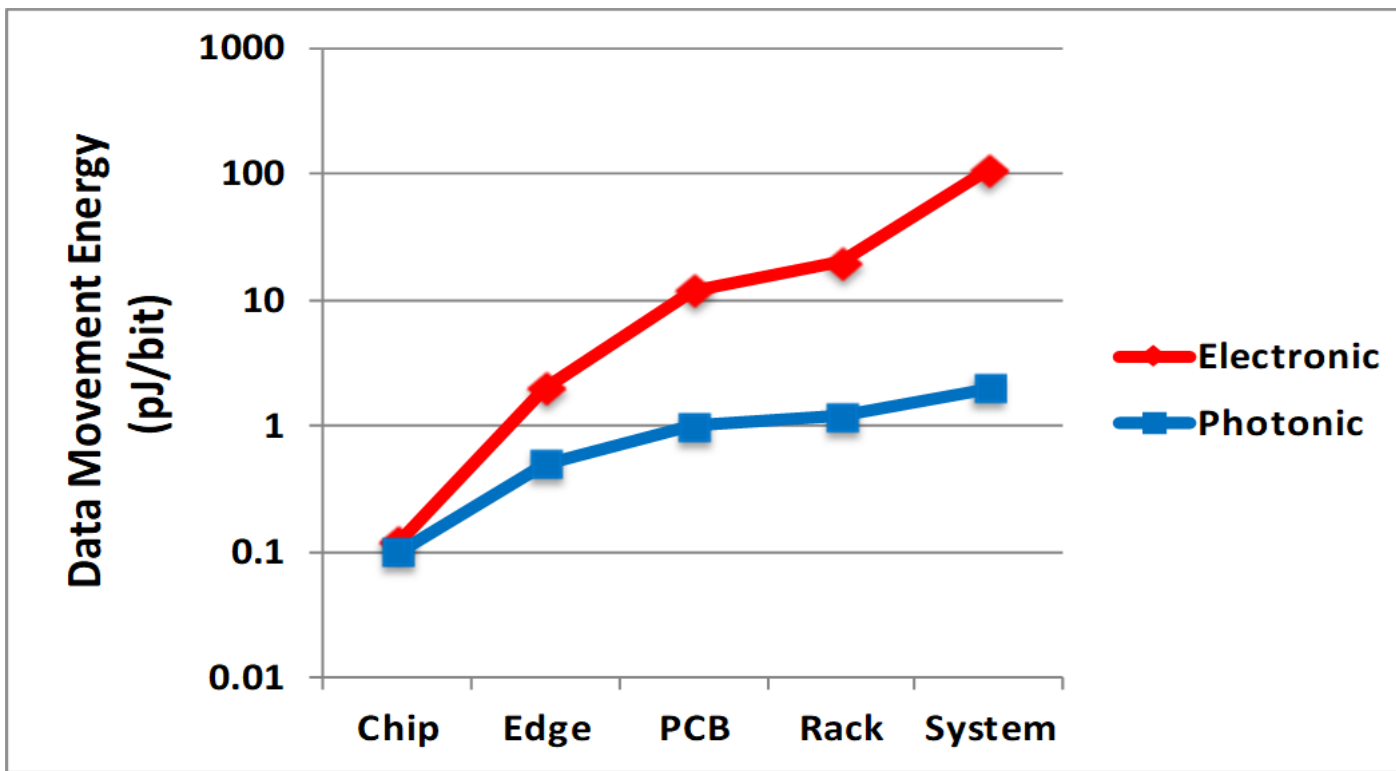
- Strong interest in **energy efficiency** of Data Centers on AI
- ...And not only for “small” systems
 - Training Deep Neural Networks (DNN) takes time!
 - “Our network takes between five and six days to train on **two GTX 580 3GB GPUs**” (Krizhevsky et al., 2012)
 - “On a system equipped with **four 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
Gyokou	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

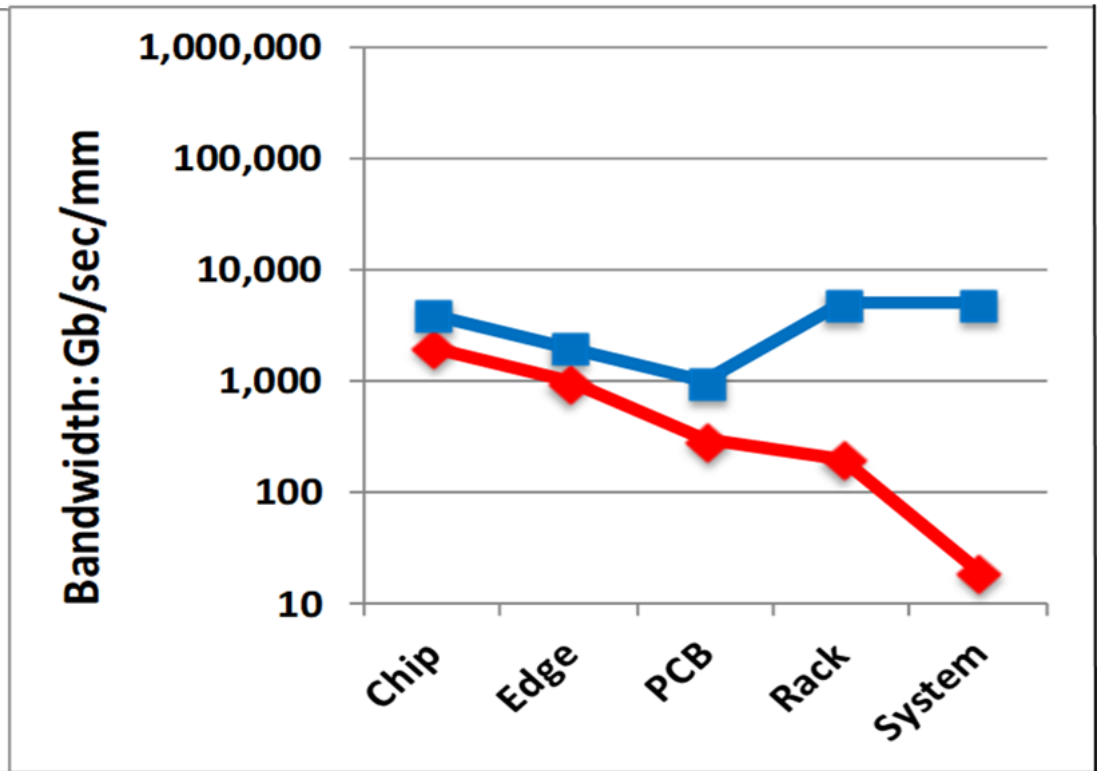
November 2017		
Name	Top500 rank	GFlop/W
Shoubu B	259	17.0
Suiren2	307	16.8
Sakura	276	16.7
DGX Volta	149	15.1
Gyokou	4	14.2
TSUBAME3.0	13	13.7
AIST AI Cloud	195	12.7
RAIDEN	419	10.6
Wilkes-2	115	10.4
Piz Daint	3	10.4
Reedbush-L	291	10.2
GOSAT-2	319	9.8
Facebook	35	9.5
DGX Saturn V	36	9.5
Era-AI	109	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)

The Photonic Opportunity for Data Movement



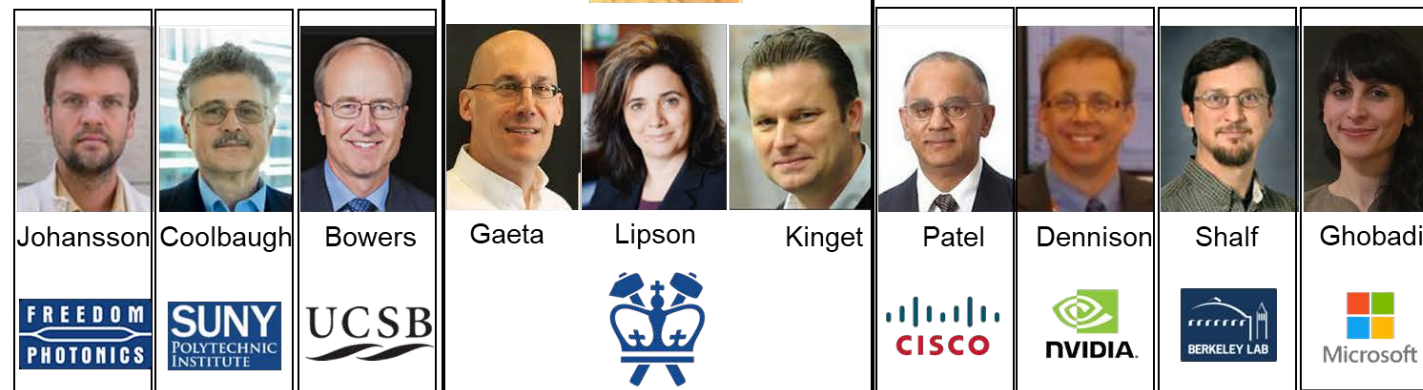
Reduce Energy Consumption



Eliminate Bandwidth Taper

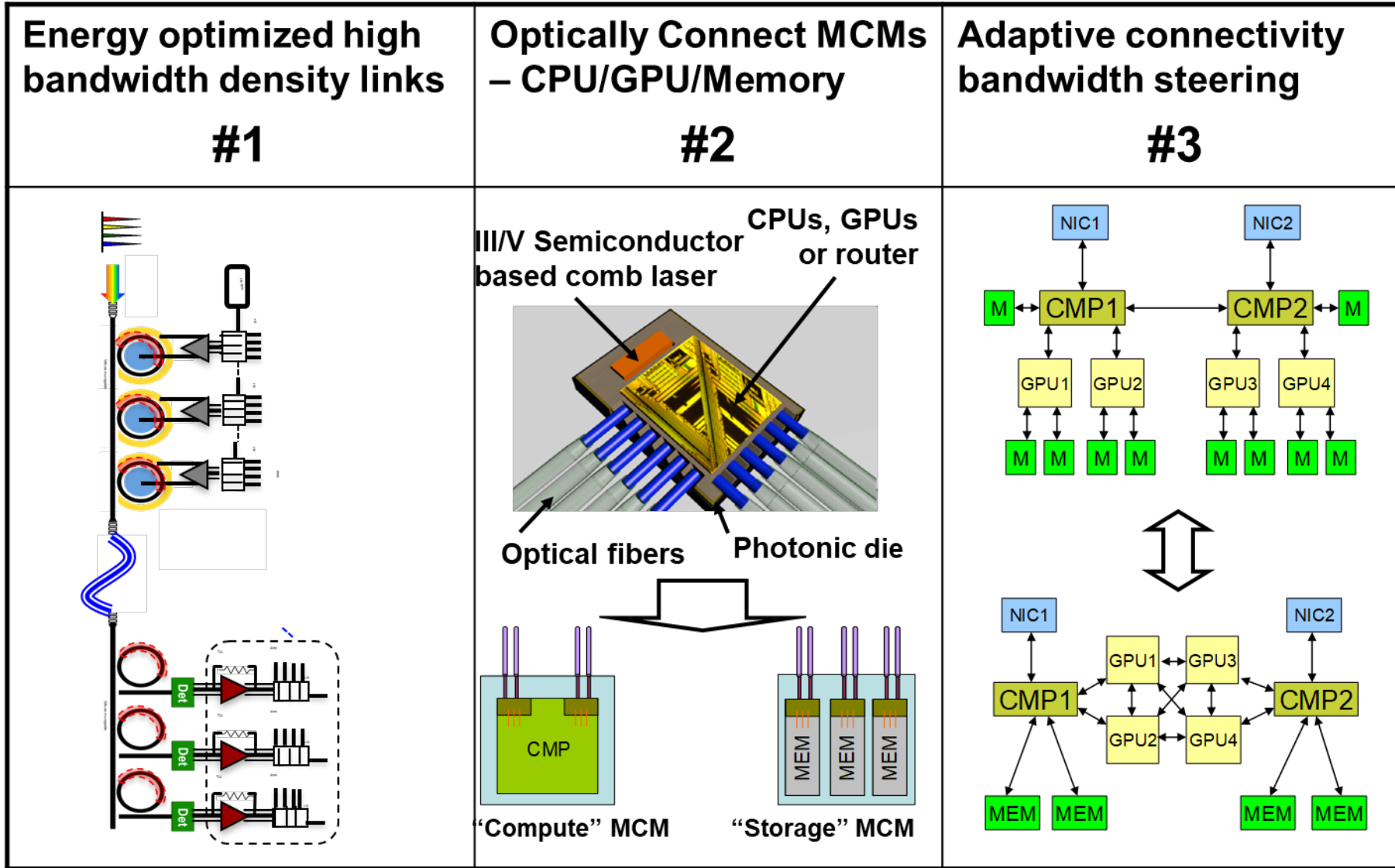
R. Lucas et al., "Top ten exascale research challenges," DOE ASCAC subcommittee Report, 2014

- ENLITENED ARPA-E program
 - Reduce datacenter energy consumption via innovative optically interconnected architectures
- PINE: Photonic Integrated Networked Energy efficient datacenters
 - Over 2X **system-wide** energy reduction



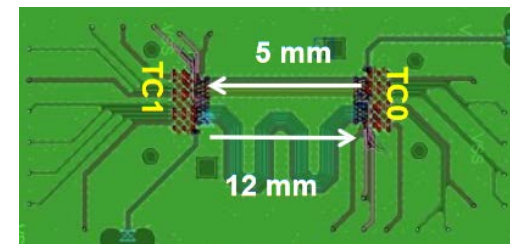
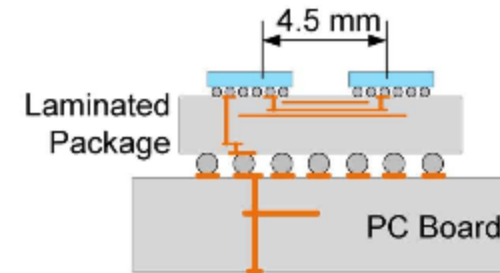
- AIM Multi-Project Wafer runs
 - 300mm fab line at SUNY Poly CNSE

Maximizing the data movement benefits of photonics:



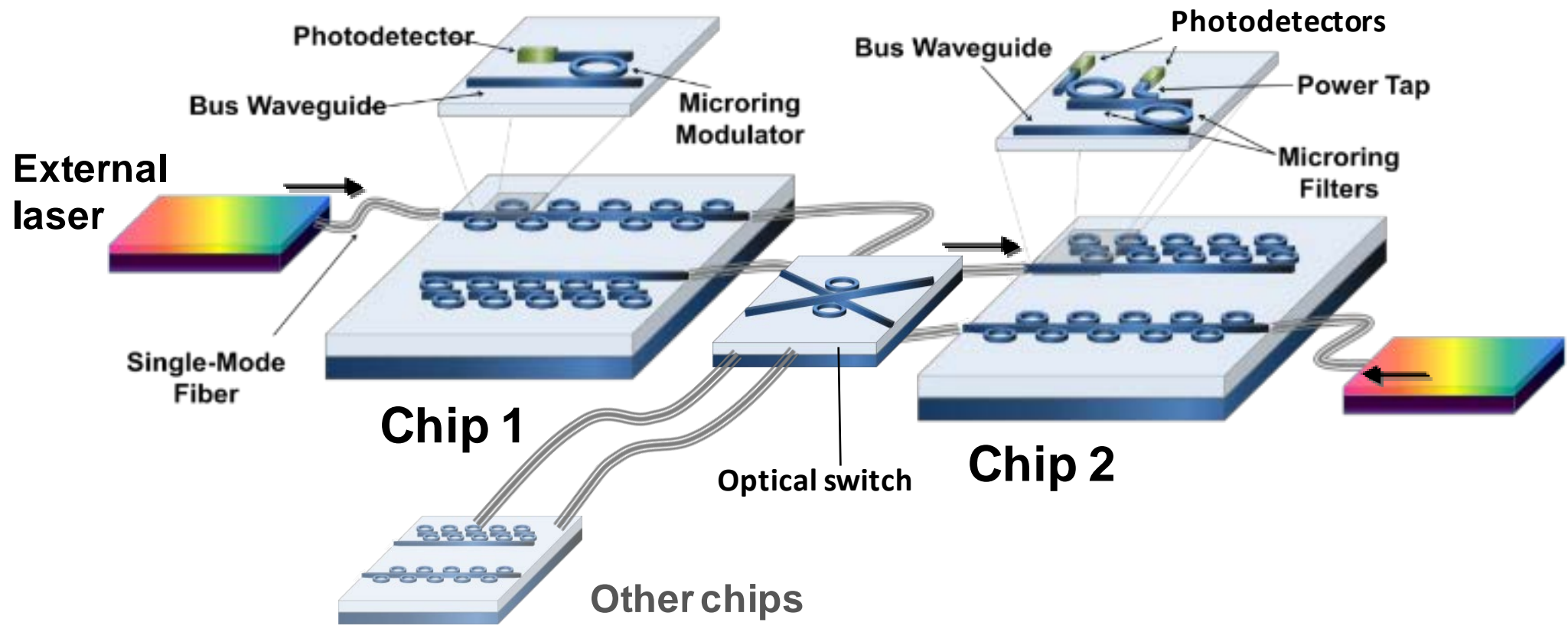
Toward 1 pJ/bit for intra-server communications

- IBM Watson, 2015 (Dickson et al. [3]):
 - 6 Gb/s per pin
 - 1.4 pJ/bit over 19mm
 - IBM Watson, 2015 (Dickson et al. [3]):
 - 16 Gb/s per pin
 - 1.9 pJ/bit over 250mm
 - Nvidia, 2013 (Poulton et al. [4]):
 - 20 Gb/s per pin
 - 0.54 pJ/bit over 4.5mm
 - Kandou bus, 2016 (Shokrollahi et al. [5]):
 - 20.83 Gb/s per pin
 - 0.94 pJ/bit over 12mm
- ➔ <pJ/bit : immediate chip neighborhood only

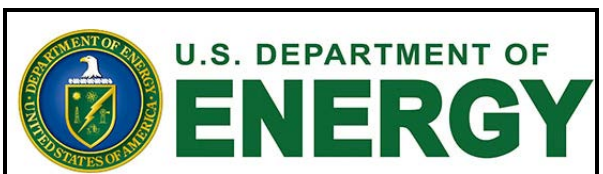
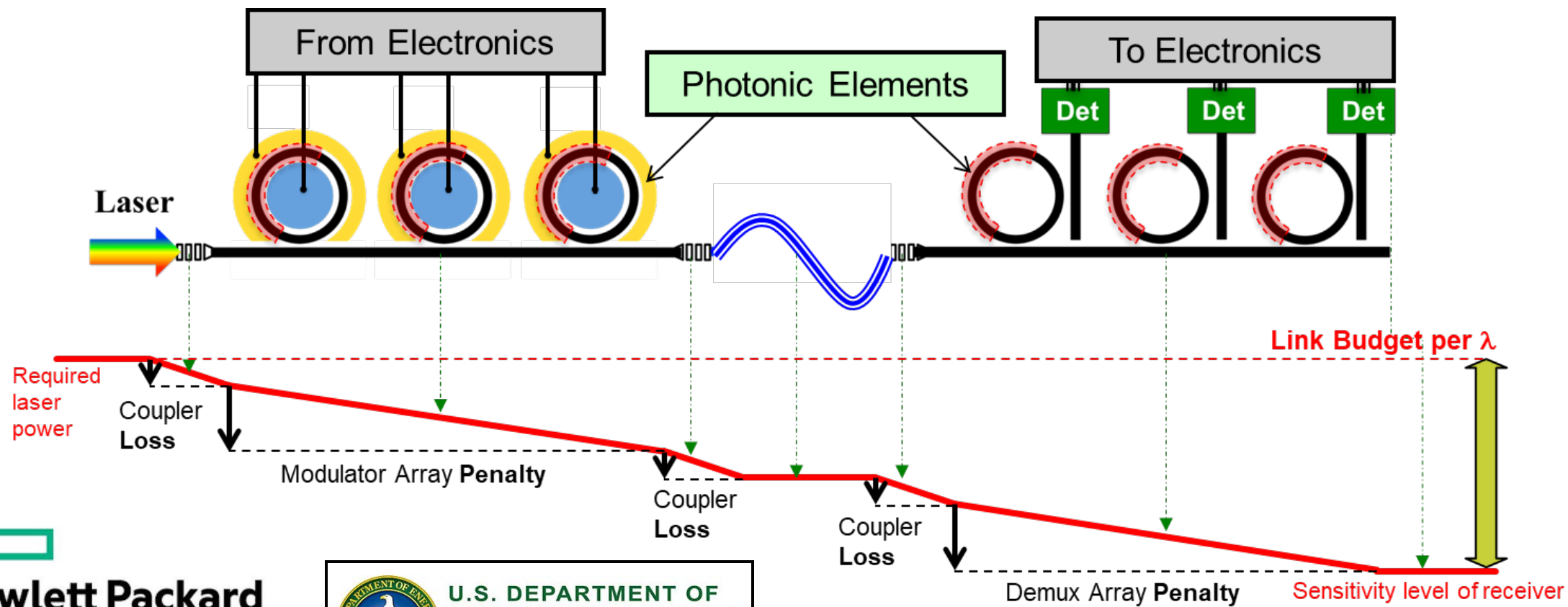


Silicon Photonics Dense-WDM

Scalable, $>Tb/s/mm$, $<1pJ/bit$ “any distance” Optical Interconnect

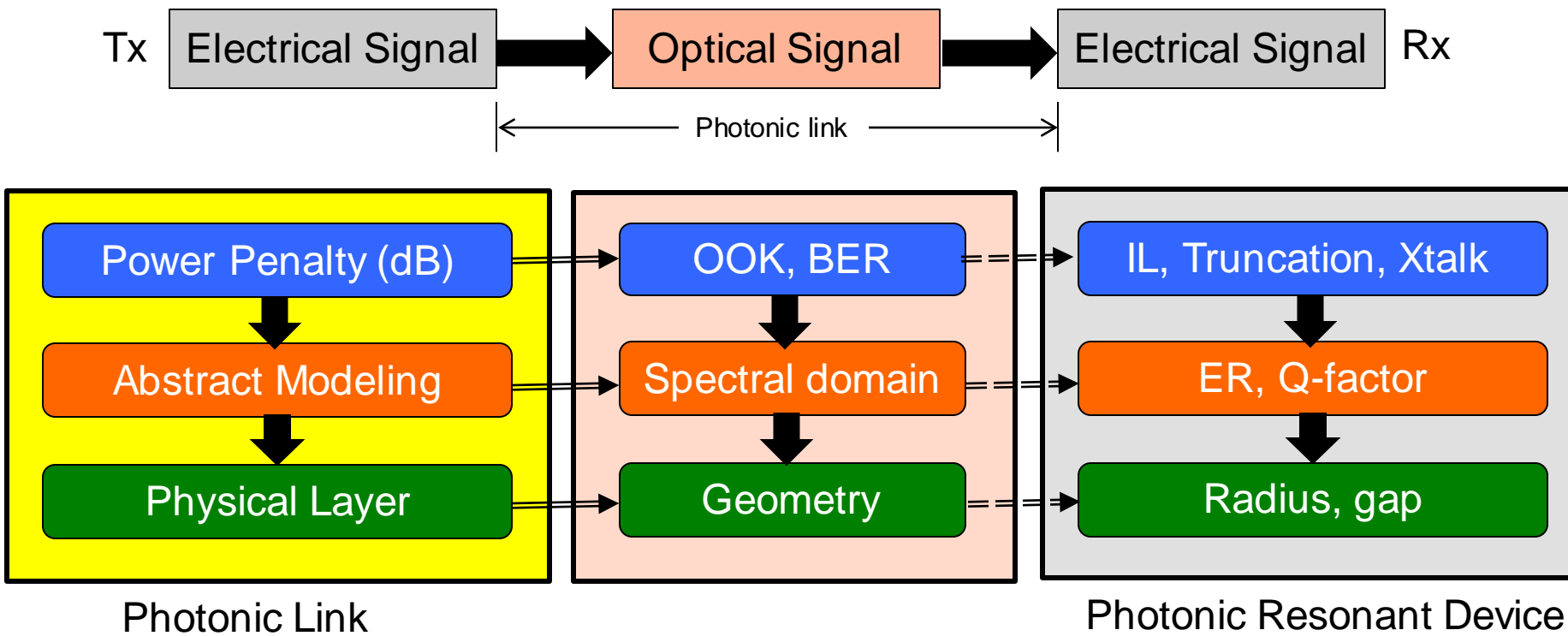


#1 Global Link Optimization: Max Bandwidth Density / Min Energy Design

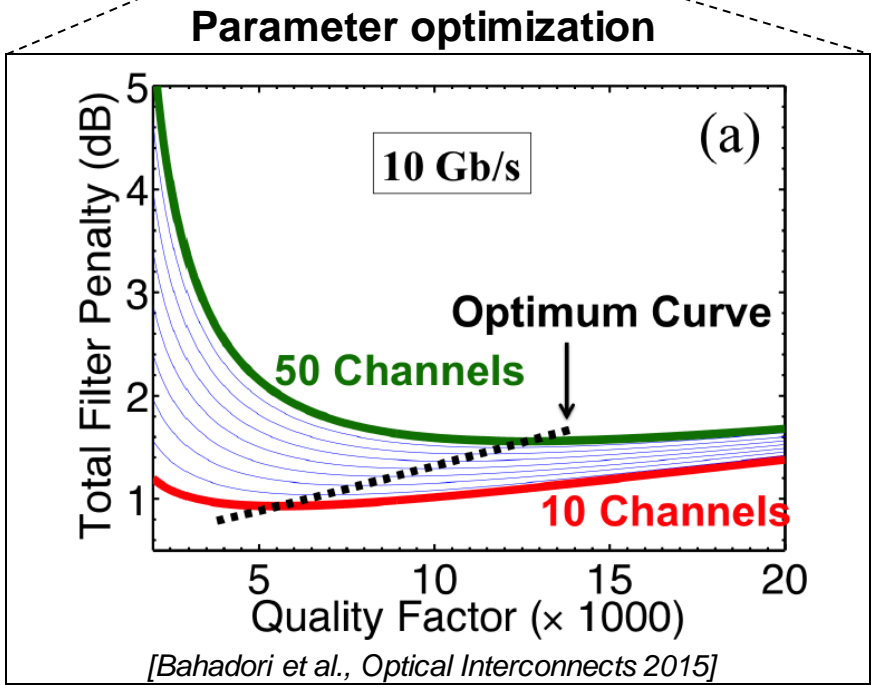
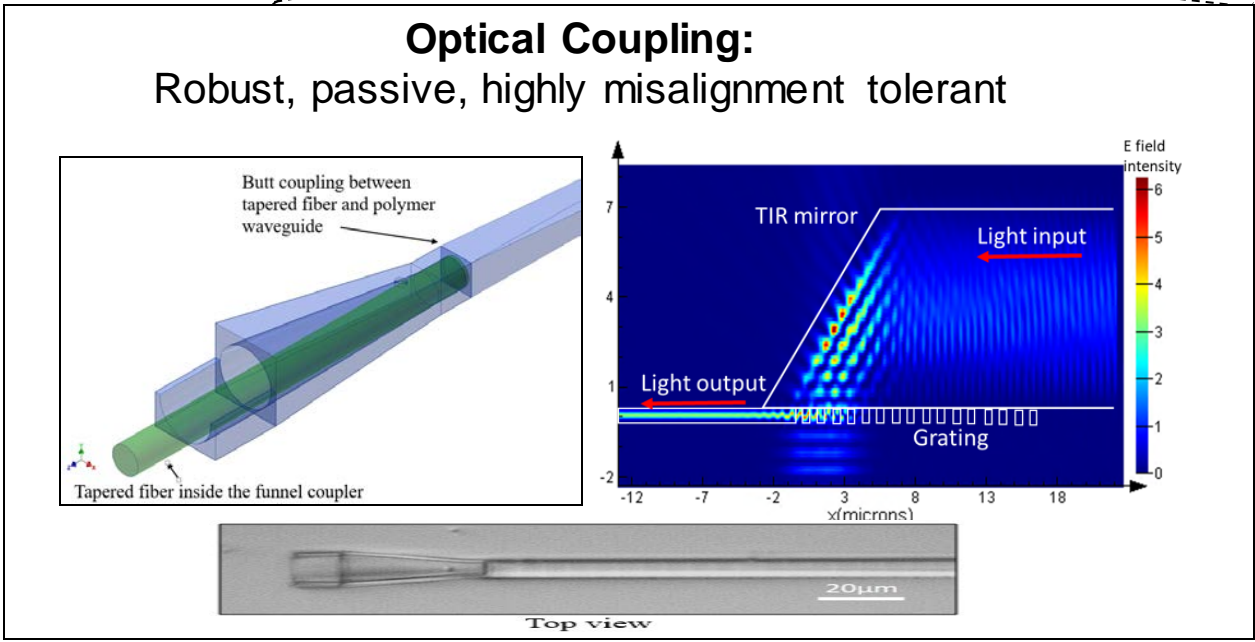
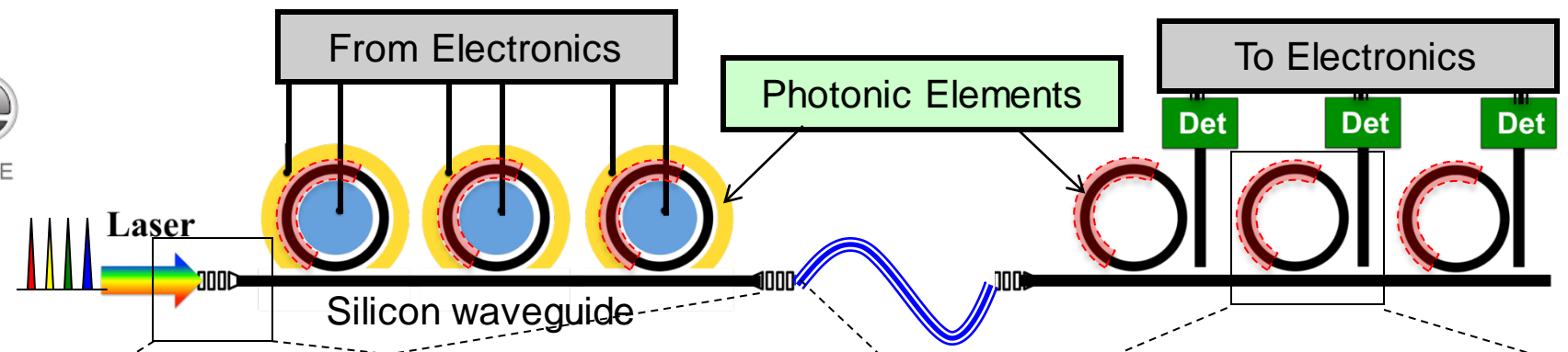


Link Optimization: Top-to-Bottom Approach

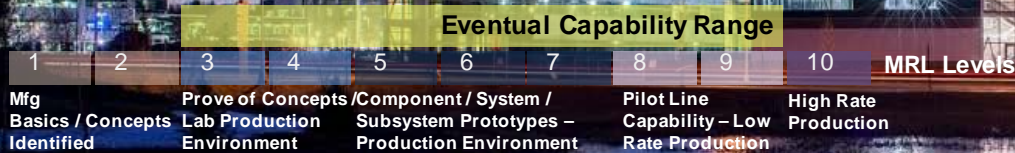
- Top Level analysis includes *optical power penalty* as the *main metric*
- Abstract modeling relates PP to spectral parameters
- Component models relate spectral parameters to physical parameters



Minimizing link losses and penalties



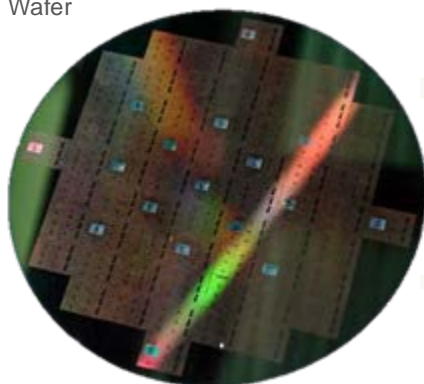
- ❑ 1.3M ft² facility
- ❑ cutting edge 300/450mm toolset
- ❑ 135k ft² of class 1 capable cleanroom
- ❑ processing capability span 65nm - 7nm



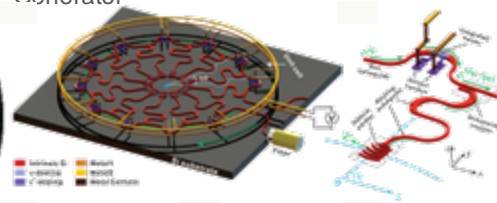
- ❑ years of proven silicon photonics results – multiple government & industry projects
- ❑ 300mm tools provide unprecedented quality photonics
- ❑ unmatched 3D stacking w/CMOS
- ❑ partnerships drive continuous revitalization investments

SUNY POLYTECHNIC INSTITUTE

300mm Si Photonics Wafer



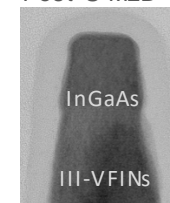
Continuously Tunable Optical Orbital Angular Momentum Generator



Erbium Laser

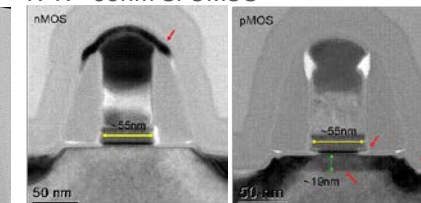


Post S-MLD

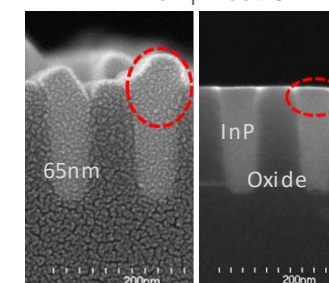


Undamaged III-V FIN

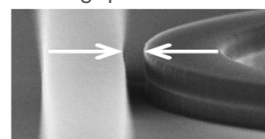
IV-IV 65nm Si CMOS



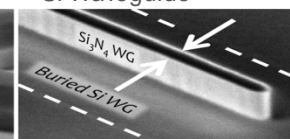
Prior | Post CMP



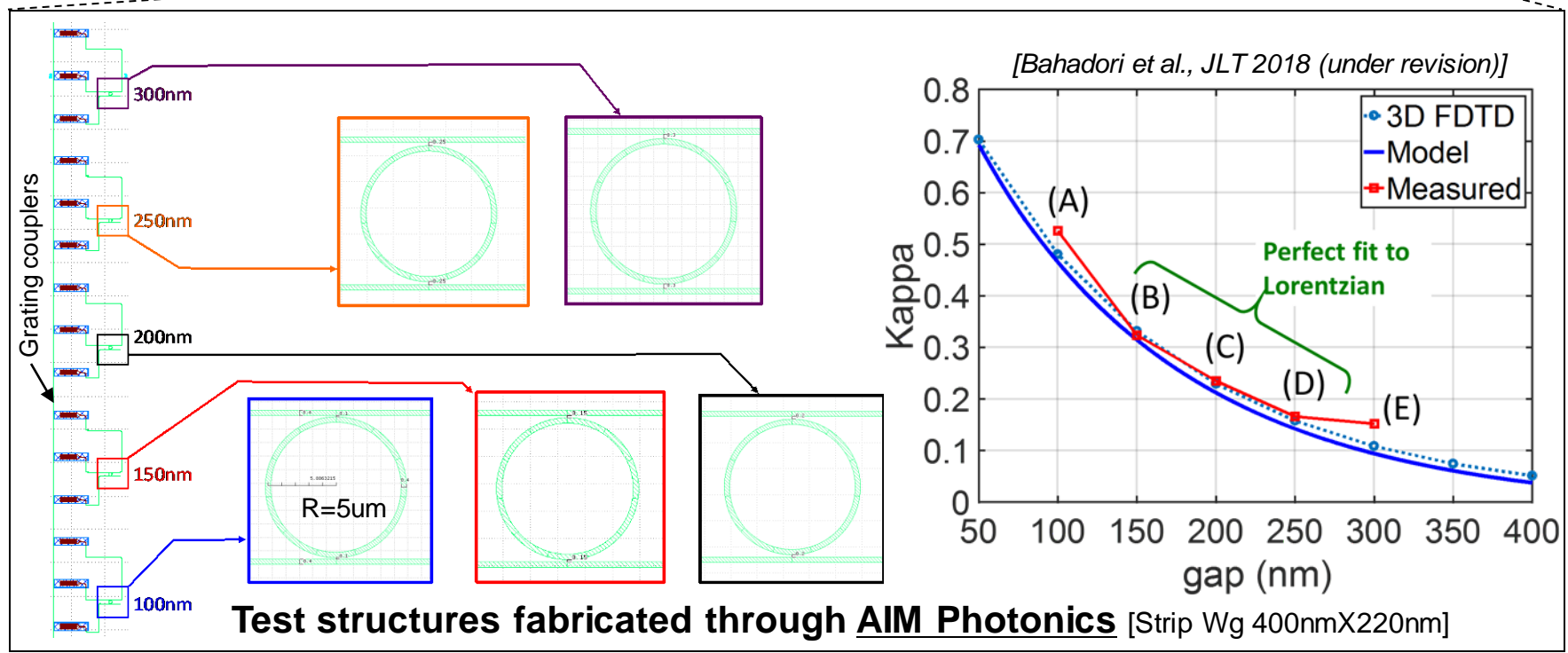
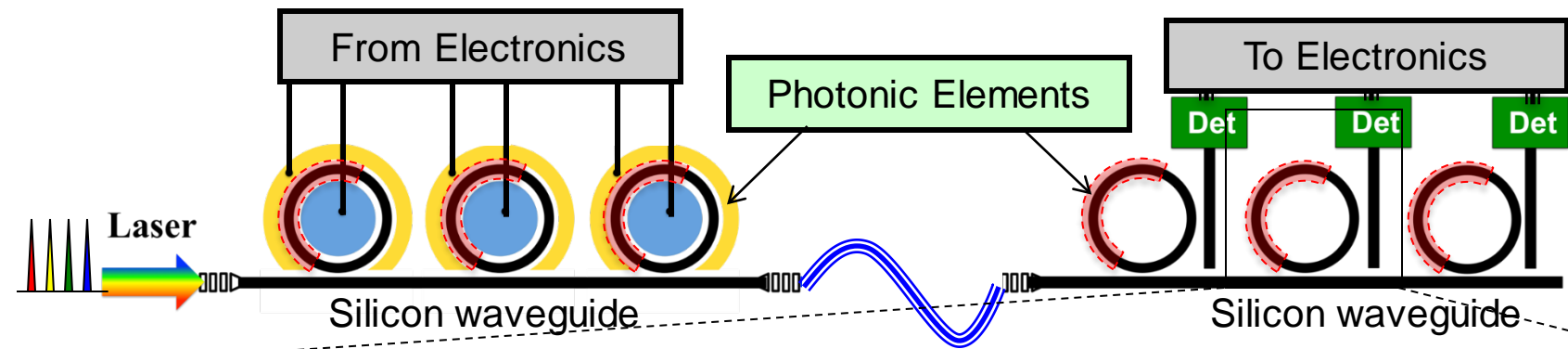
95nm gap in Si



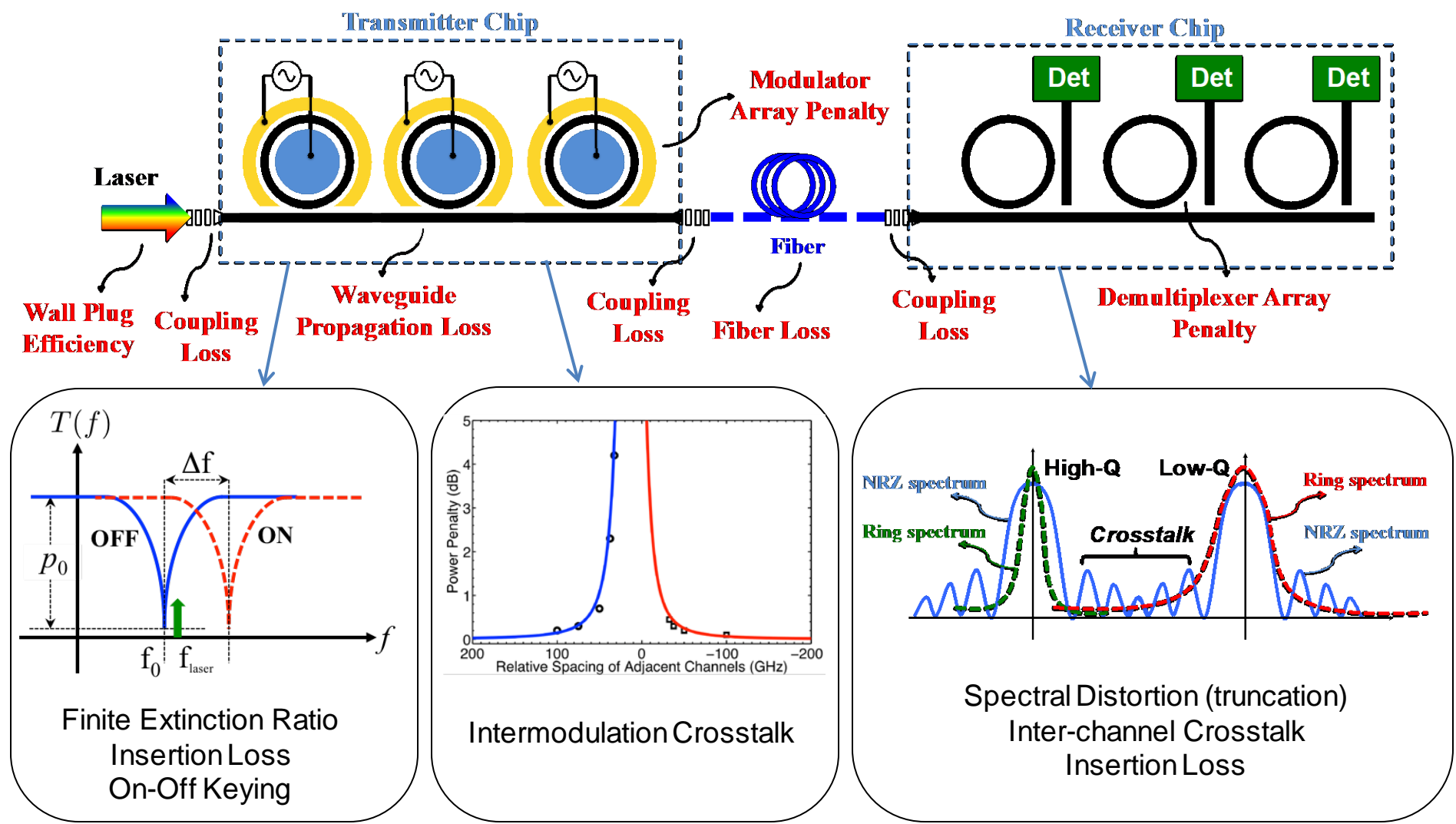
95nm Si₃N₄ Taper on Si Waveguide



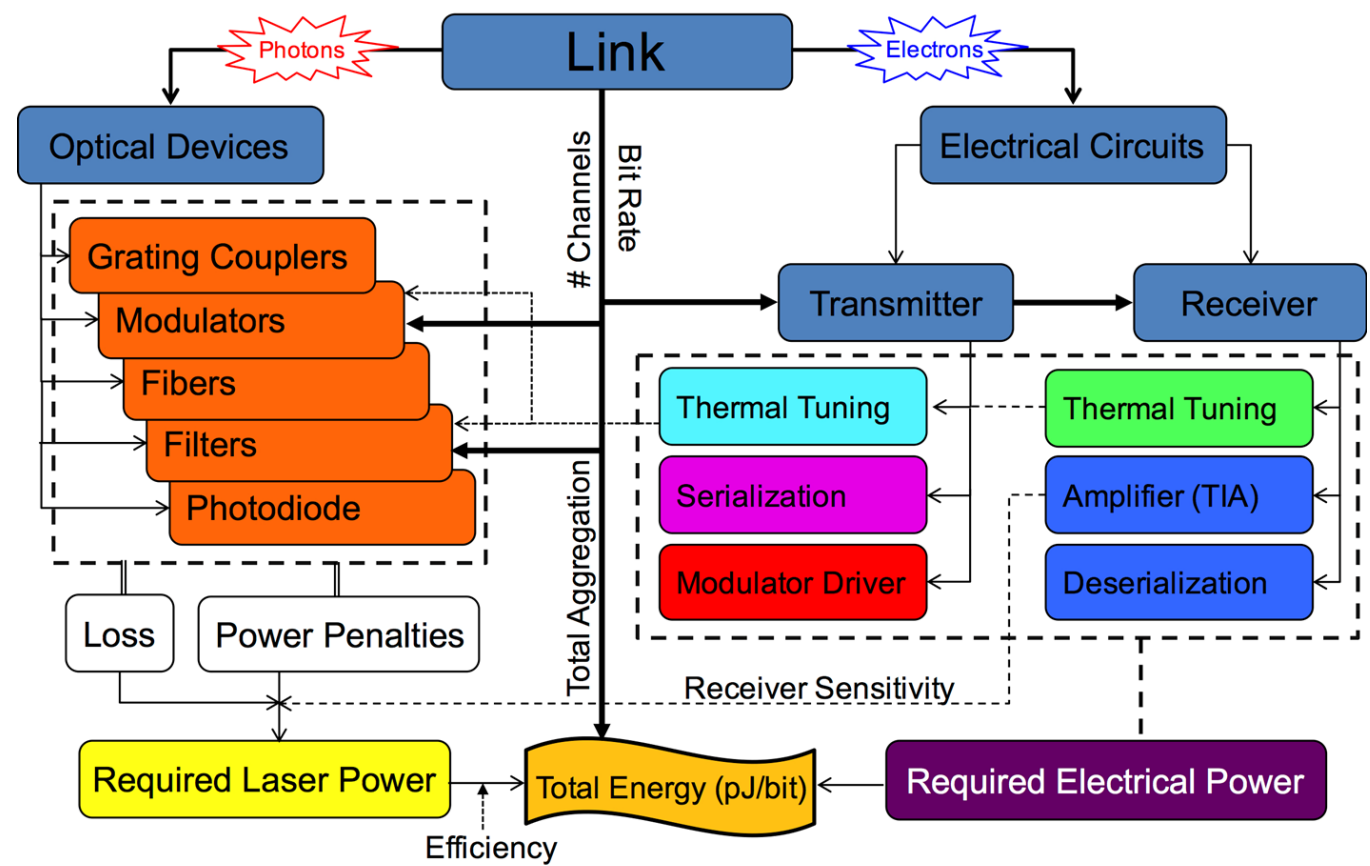
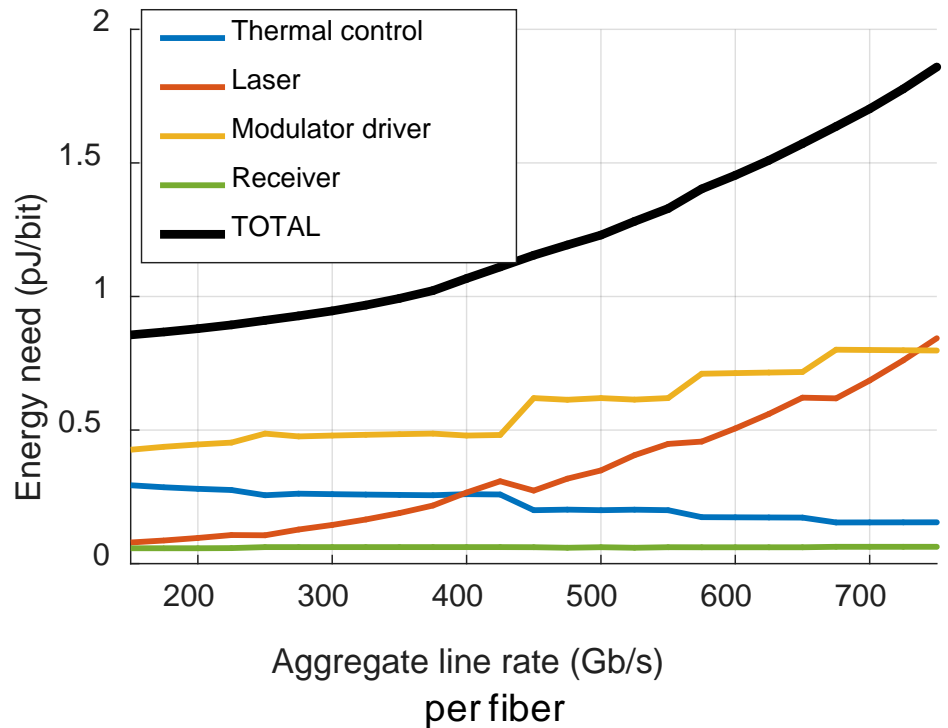
Optimizing ring resonators



WDM Link: optical impairments



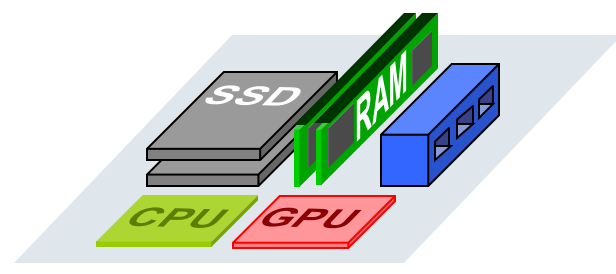
Optimized global power efficiency



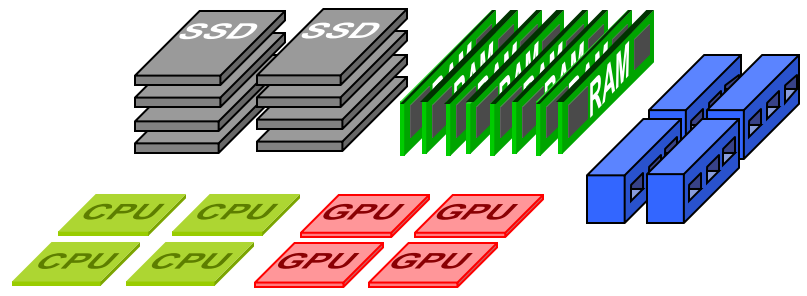
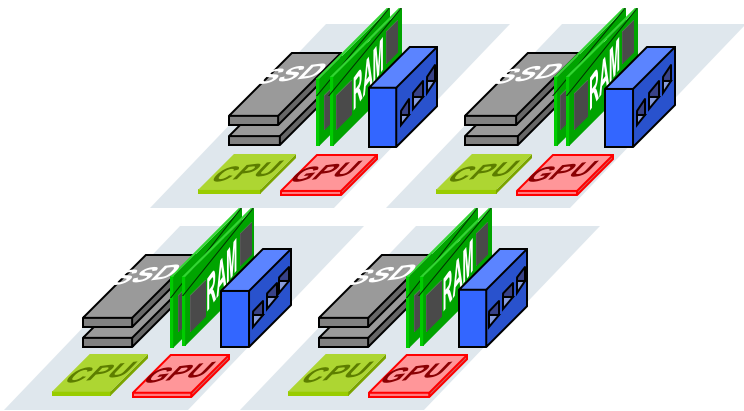
- Complete end-to-end link electronic/photonic models
- Efficiency prediction with all parameters optimized
 - Design space: channel rate / number of channels

Only "Power Up" Needed Optical Links: Disaggregated Data Center Architecture

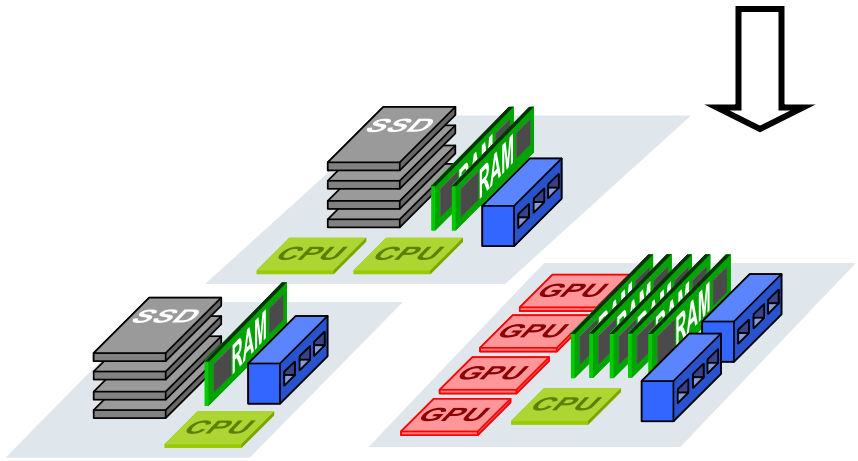
Current server



Current rack



Disaggregated rack



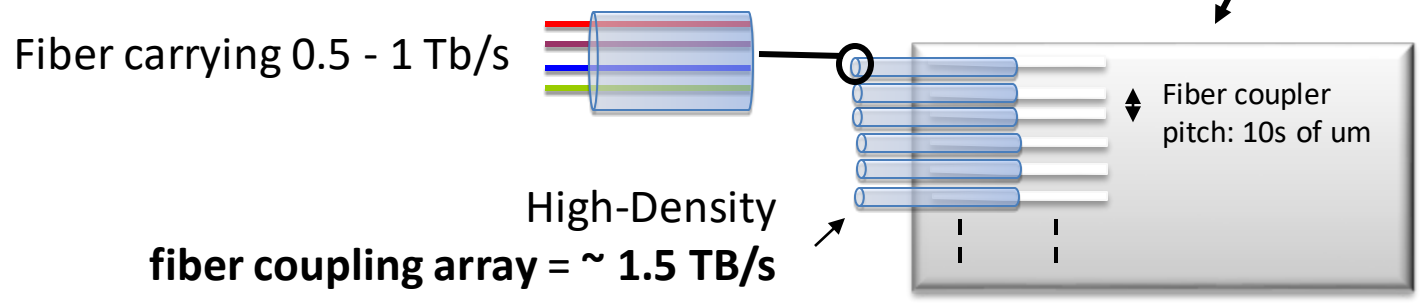
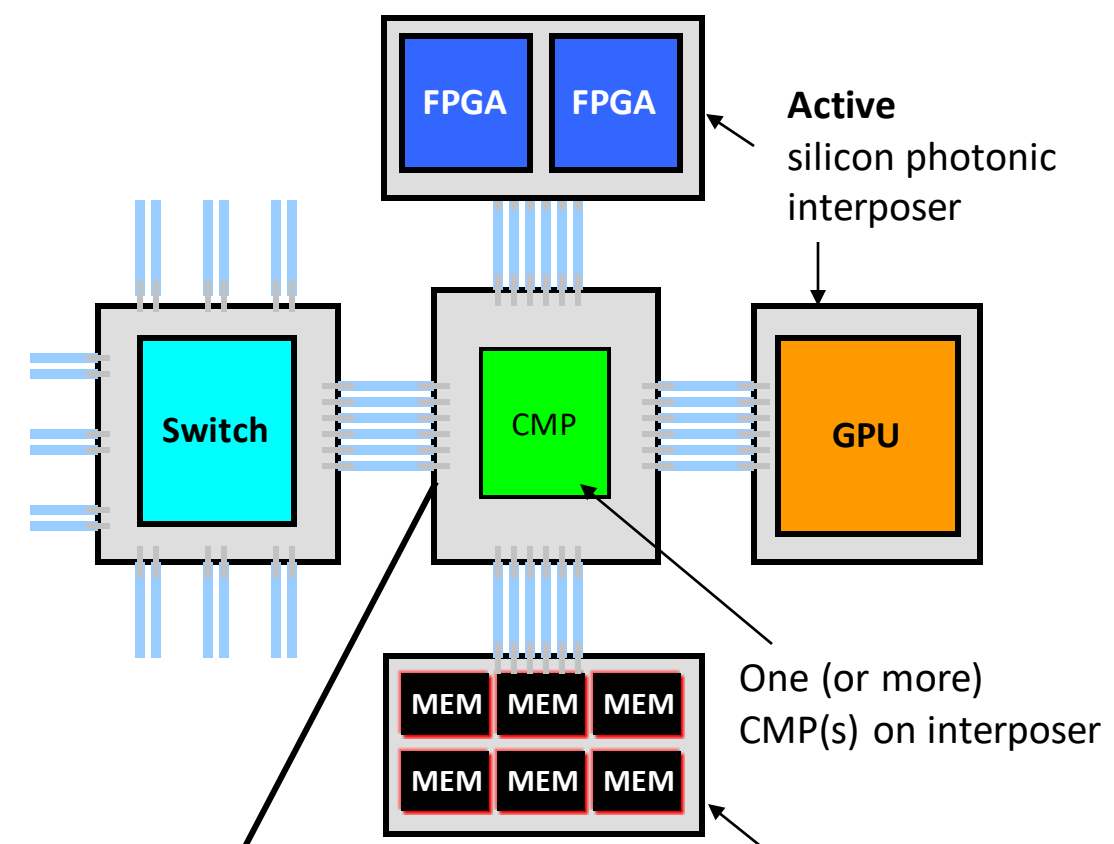
Pool and compose

However... Inter Node Bandwidth (10 GB/s) << RAM Bandwidth (100 GB/s – 1 TB/s)

PINE Concept #2: Ubiquitous Optically Connected-MCM

- OC-MCM: Optically Connected Multi-chip Module
 - Optical communication among interposers/MCM
 - Unified interface
 - Builds on recent industry efforts (Gen-Z, OpenCAPI, CCIX)
 - Enables fully flexible and scalable architectures

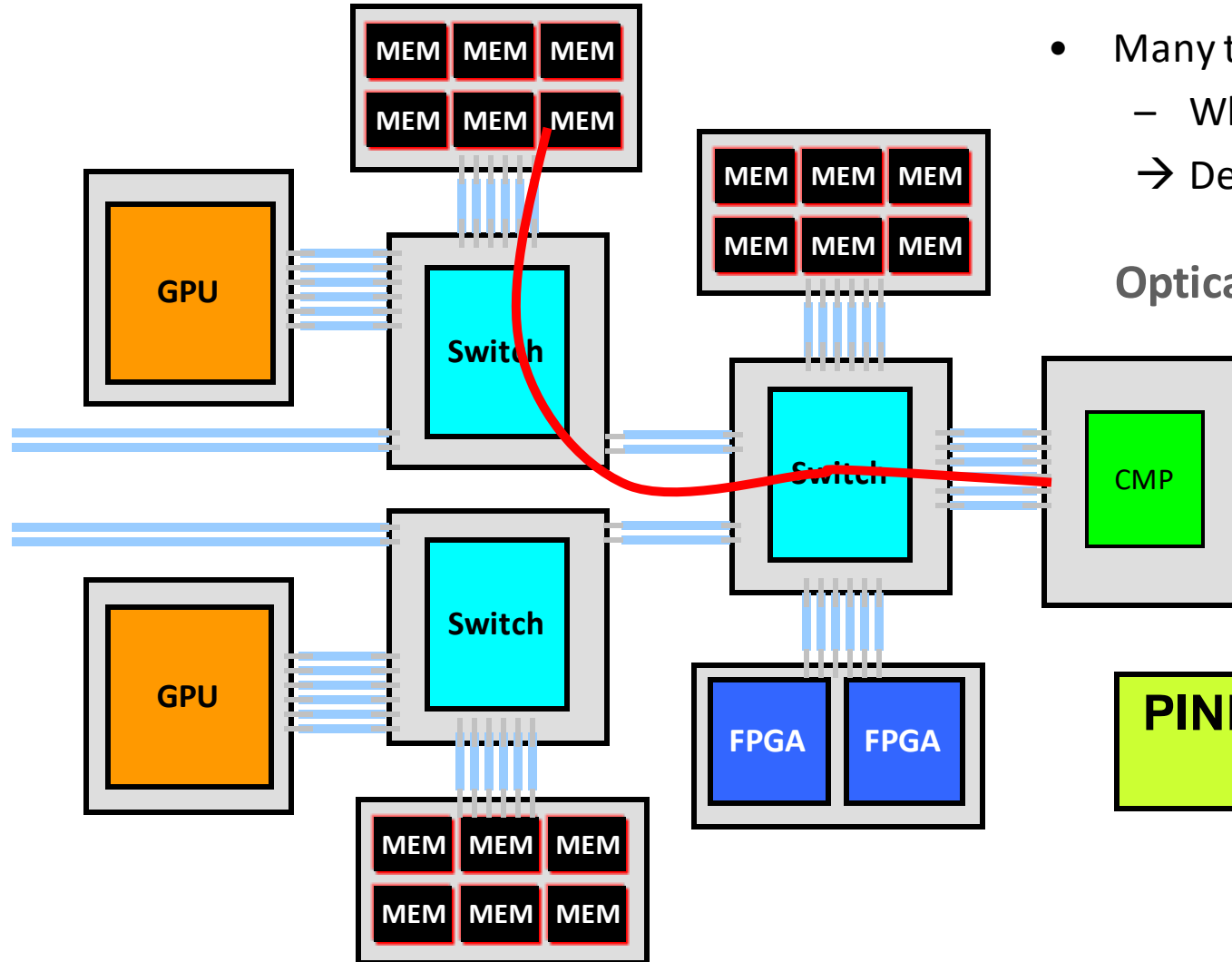
Approach: network of resources
...rather than a network of servers



Example: memory **module**:
6 packages, 48GB, 1.5 TB/s

One package: 8GB, 0.25 TB/s

PINE Disaggregation: Deeper into the Hierarchy



- Many topologies possible
 - Which ones make sense?
 - Design space exploration under loads

Optically Connected MCM (OC-MCM)

PINE: energy efficient bandwidth at all levels, any distance

Need for AI optimized architectures

- Wide parallelism is now everywhere in the datacenter
- The question is not “should we go parallel” anymore, but “how wide” – how many CPUs, GPUs, memories, etc.
- Depend on applications, e.g.:

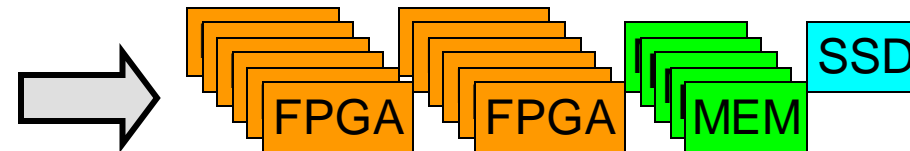
- DNN training:

- Lots of GPUs (Possible Petaflops)
- Some memory (for example 1 TB)
- One CPU to orchestrate



- DNN inferencing

- Many FPGAs or ASICs
- Some high speed memory
- SSDs



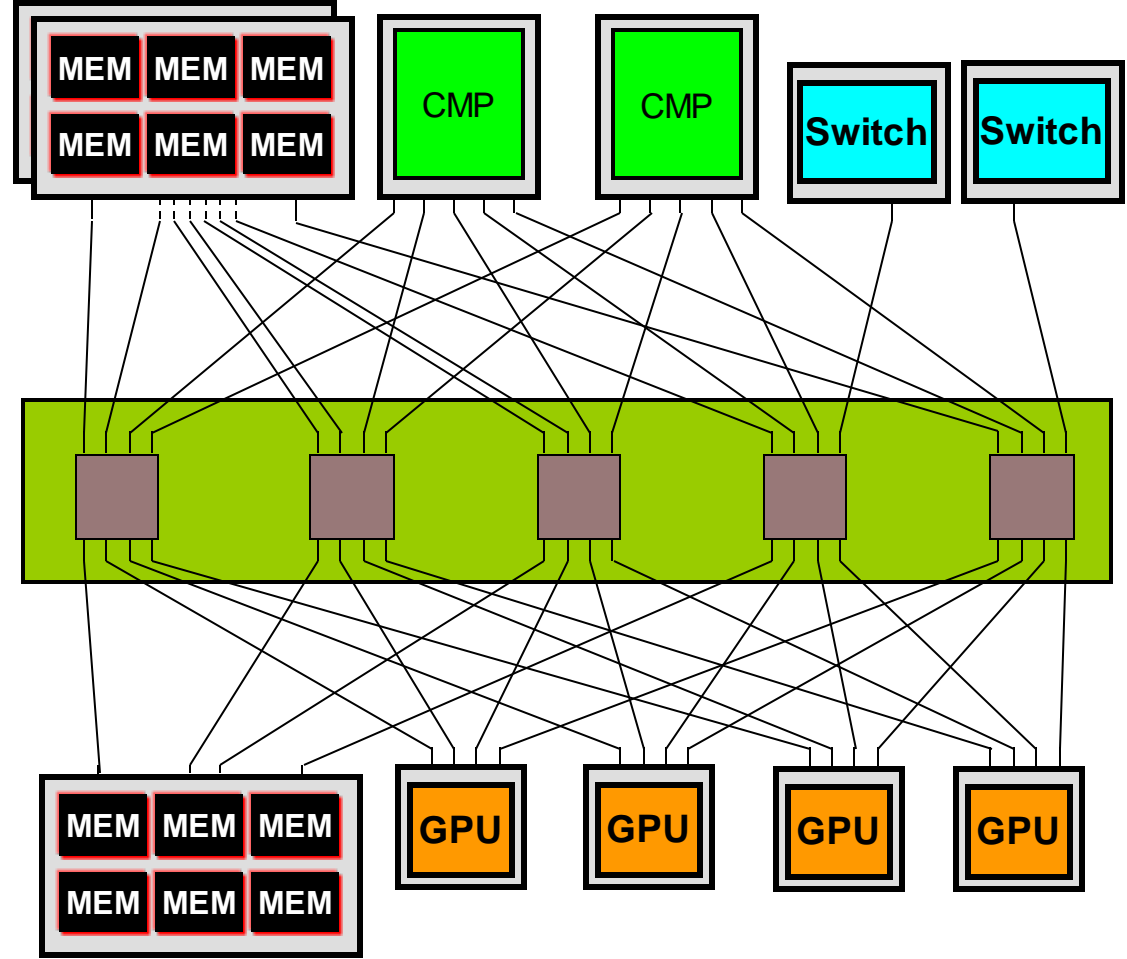
- Etc.

- If each resource can be provisioned in any quantity with flexibility, no more unused resources → **energy savings**

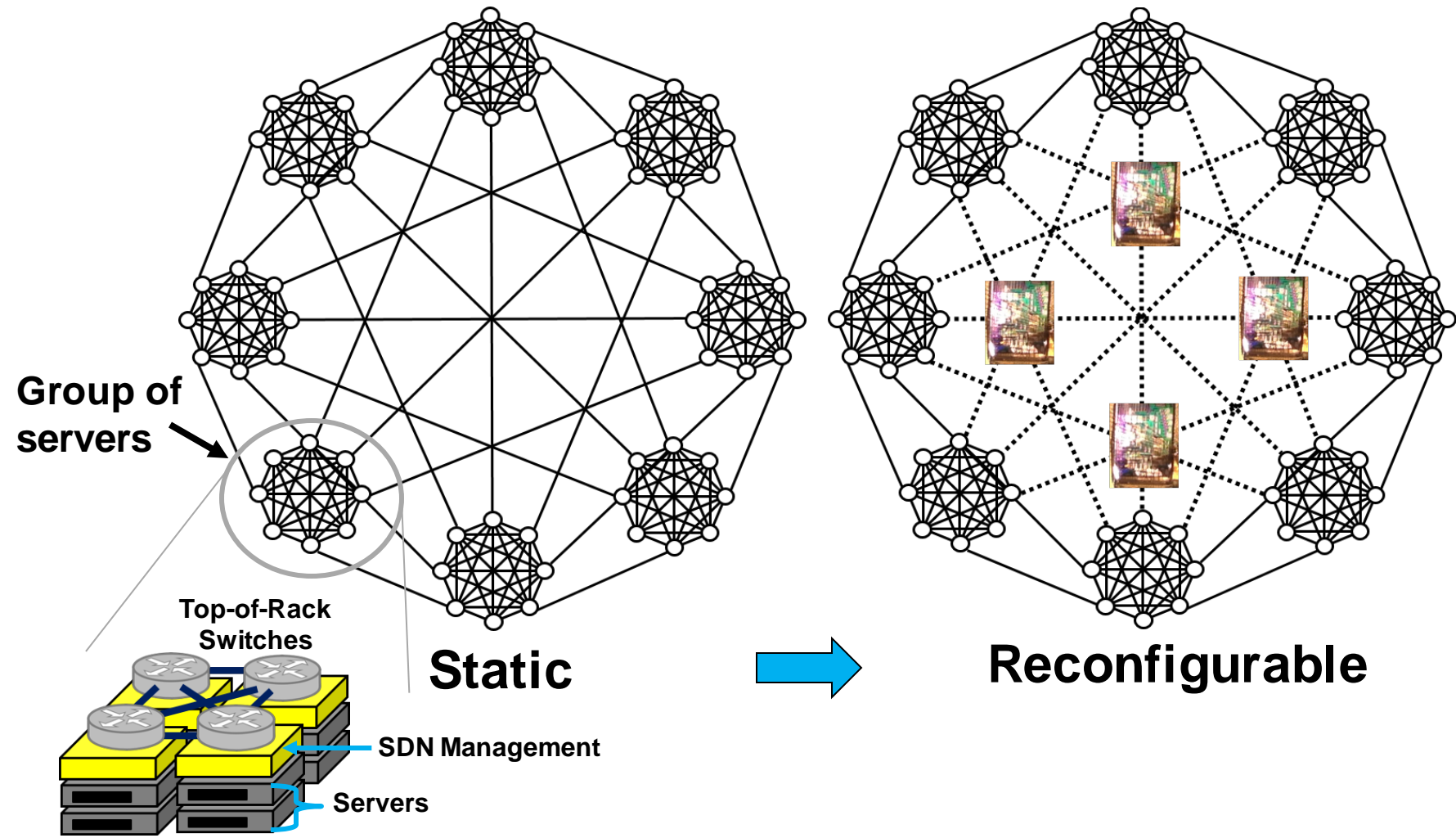
#3 Adaptive, Flexible Connectivity with Bandwidth Steering

→ Deep Disaggregation

- Introduce optical switches in the OC-MCM topology
 - 8x8 realizable with today's technology
 - Tens of switches can be collocated on a single chip
- Flexibly assembled nodes
- Transparent for packets
 - Low-Latency direct connectivity
 - Energy efficient



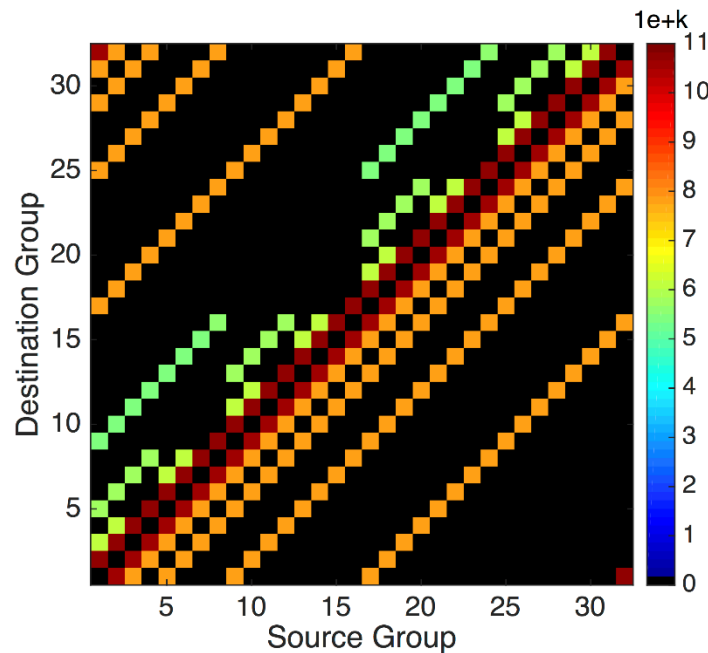
PINE Concept #3: Flexible Bandwidth Steering with Silicon Photonic Switches



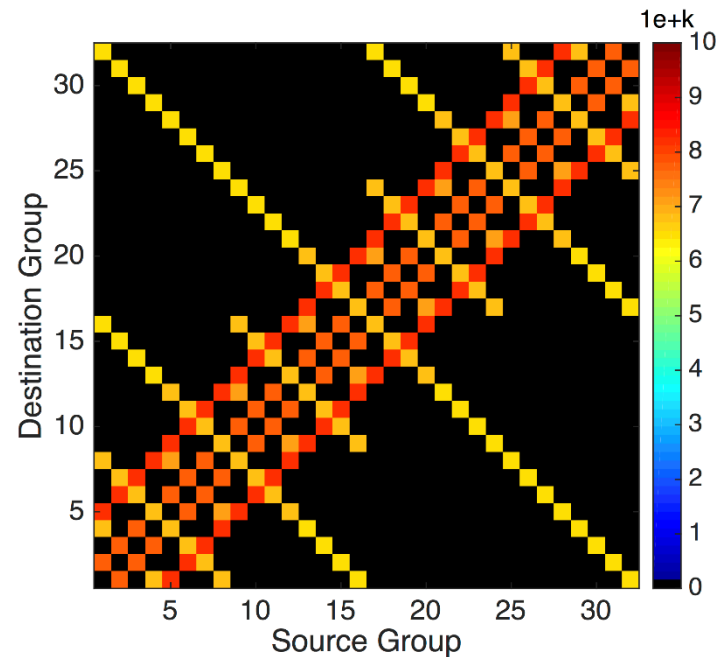
Adapting the Network to HPC Traffic

- Traffic characteristics of HPC applications: **skewed**, **well-defined** and **slowly** varying over runtime
- Reconfigure physical network interconnect topology to match traffic pattern

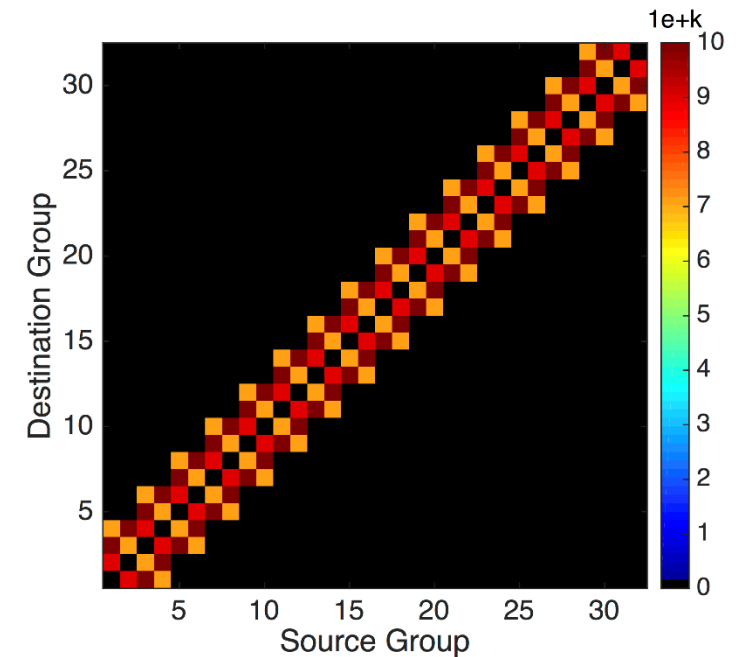
GTC



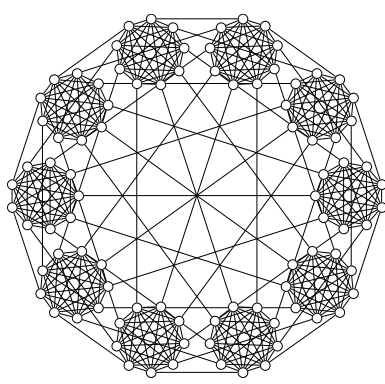
Nekbone



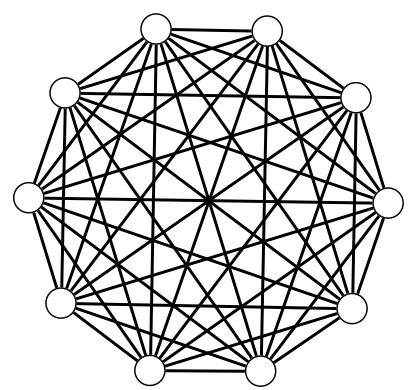
LULESH



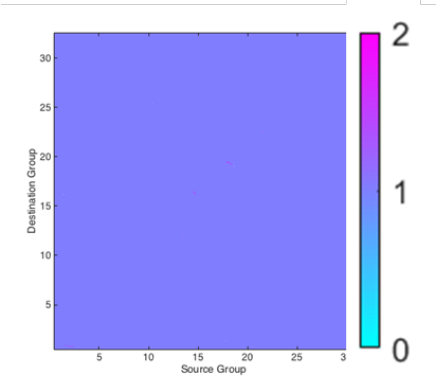
Flexfly: Optical Bandwidth Steering in Dragonfly



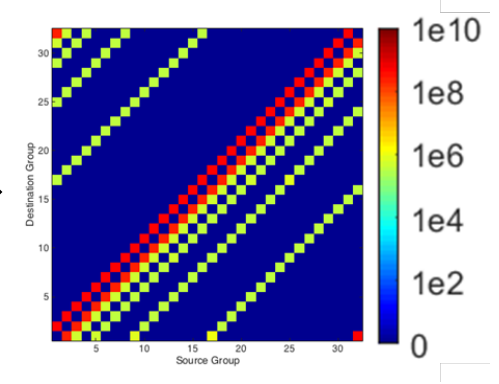
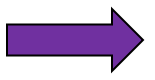
Dragonfly topology



Inter-group topology



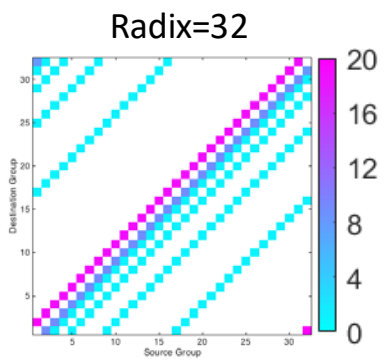
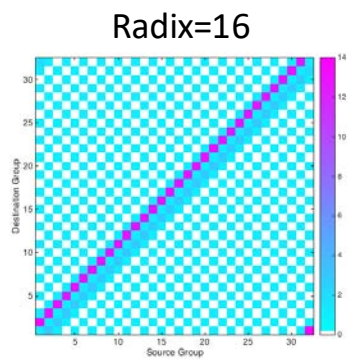
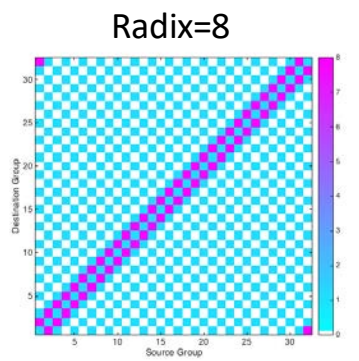
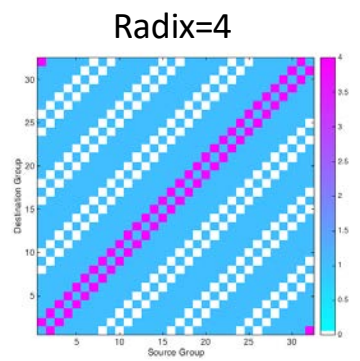
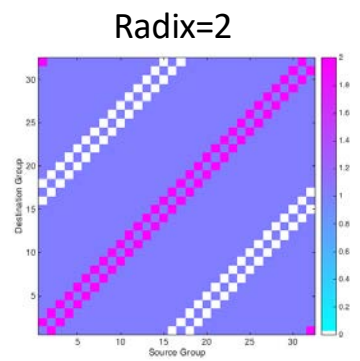
Relative provisioned bandwidth



Inter-group traffic (GTC workload)

Increases relative provisioned bandwidth to match traffic

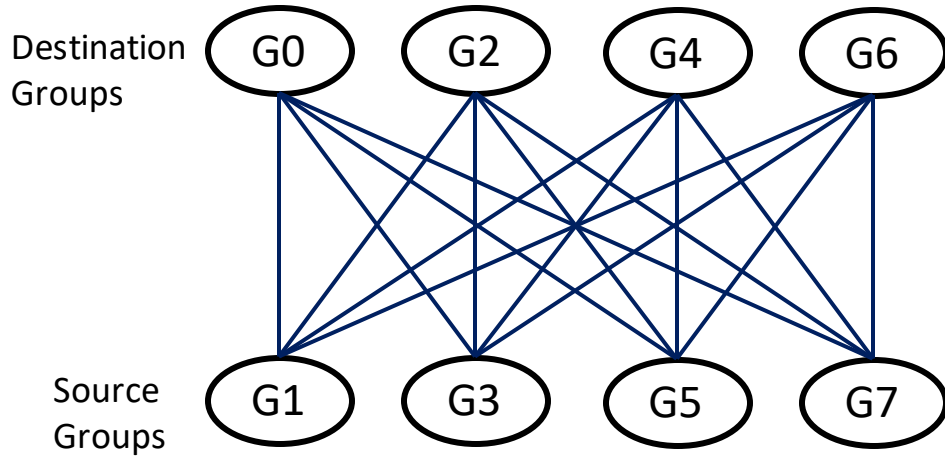
Increasing Photonic Switch Radix →



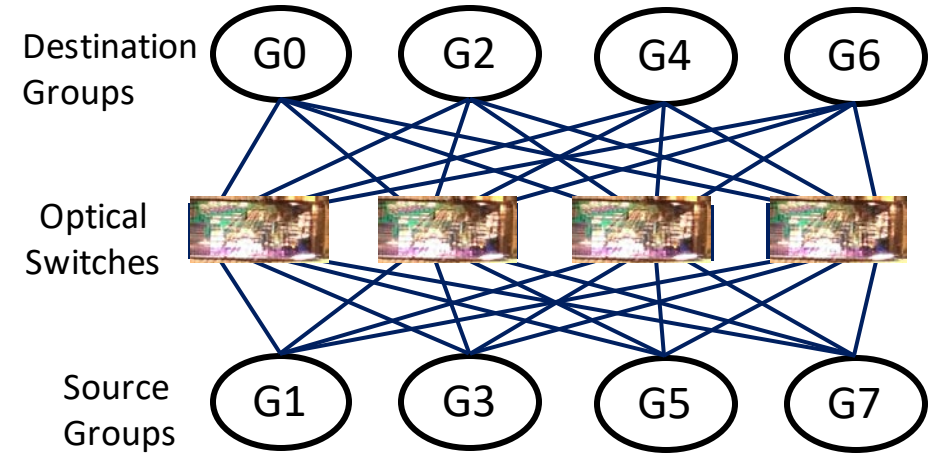
Bandwidth Steering enables “matching” of connectivity matrix to traffic matrix

Photonic Switch Insertion Strategy

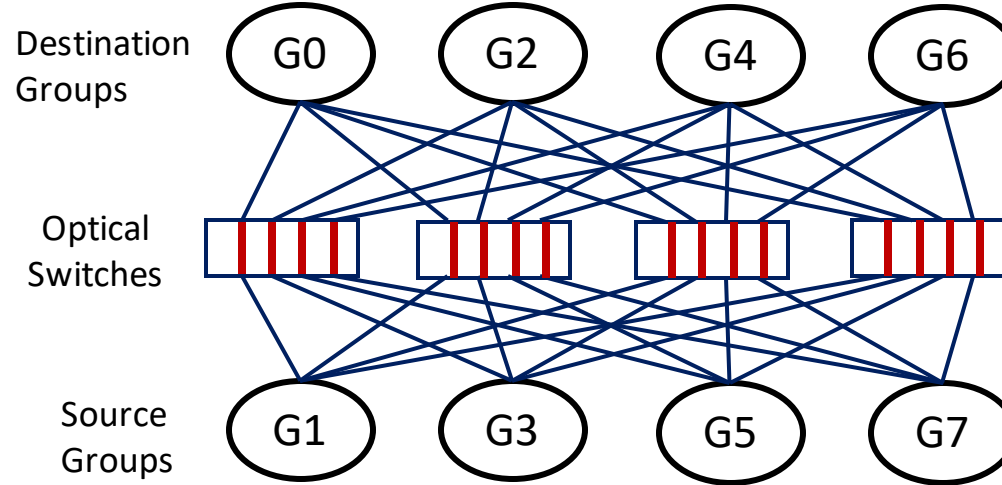
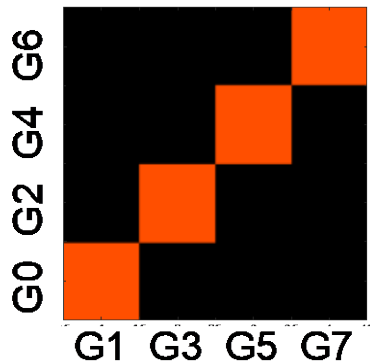
A strided gang in Dragonfly



A strided gang in Flexfly

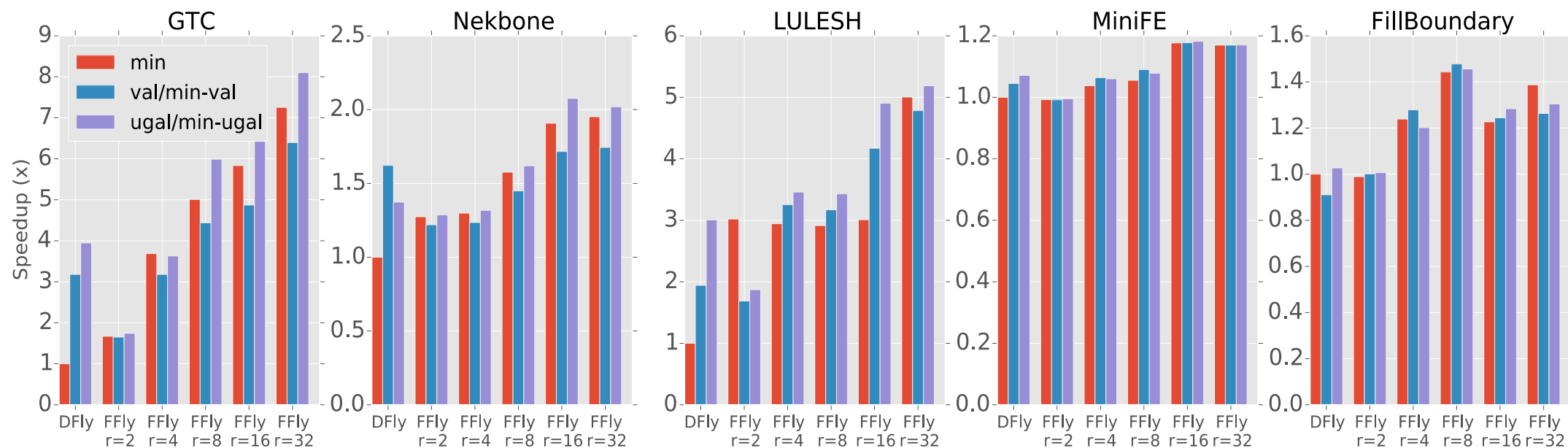


Example extracted traffic matrix



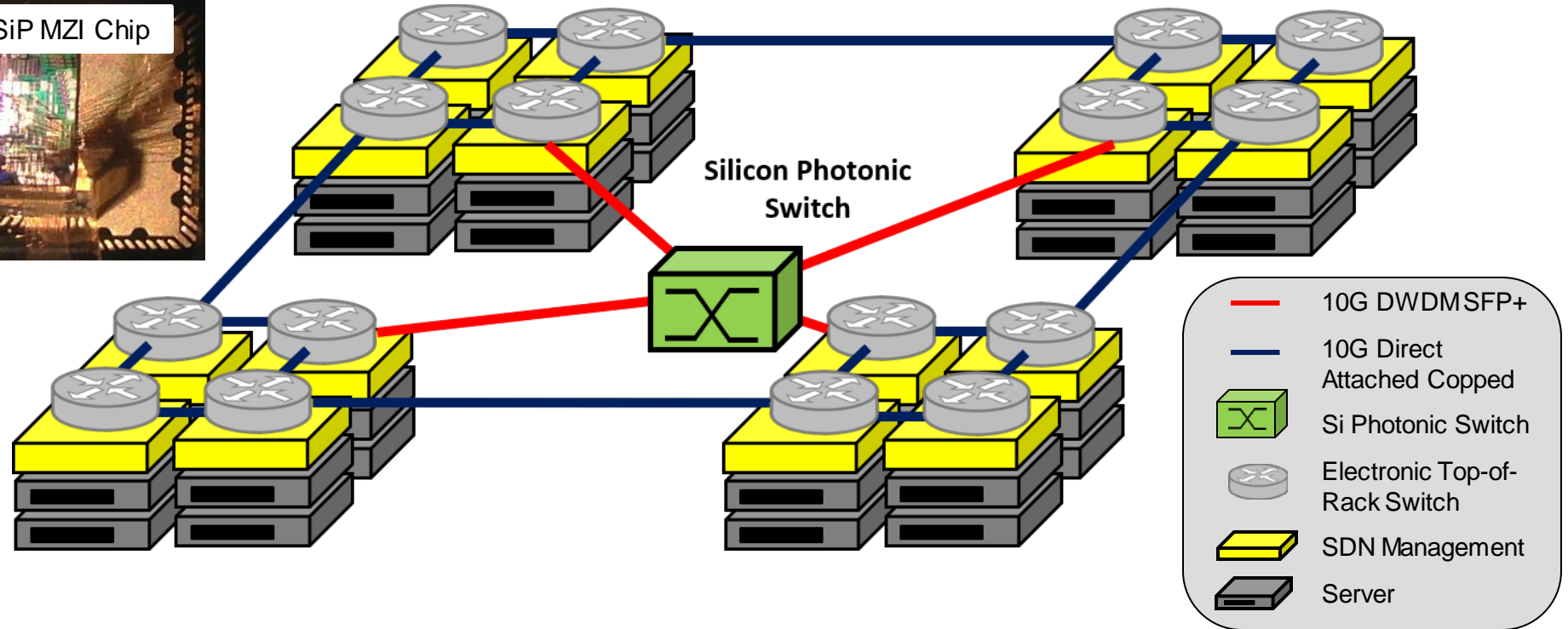
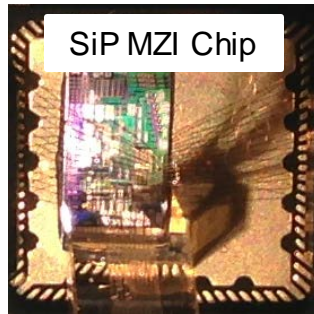
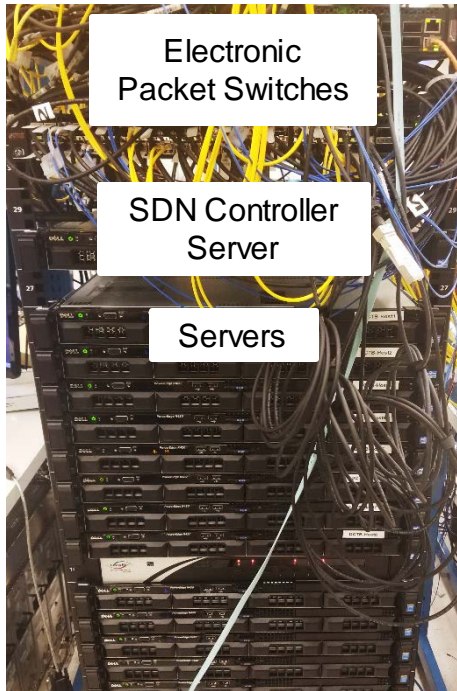
Flexfly: R fold bandwidth increase

Application Speedup



Speedup of	over	GTC	Nekbone	LULESH	MiniFE	FB
Flexfly + Min	Dfly+Min	7.0x	2.0x	5.0x	1.2x	1.4x
	Dfly+ugal	1.8x	1.4x	1.7x	1.1x	1.4x
Flexfly + Min-UGAL	Dfly+Min	8.0x	2.1x	5.2x	1.2x	1.4x
	Dfly+ugal	2.0x	1.5x	1.8x	1.1x	1.4x

Photonic Bandwidth Steering – Prototype System Testbed



- 64- node system arranged in Dragonfly topology.
- Bandwidth steering performed through multiple embedded 4x4 silicon photonic switches
- Operates various HPC benchmarks such as GTC, HPCG, MiniFE

Experimental Results – GTC Benchmark Application

- Testbed operating a skeletonized version of the Gyrokinetic Toroidal Code (GTC) benchmark application with MPI scheduling
- Comparing application execution times between a baseline Dragonfly (all-to-all) topology and bandwidth-steered topology

Number of Ranks	Execution Time (s)		Performance Increase %
	All-to-all Topology	Bandwidth-Steered Topology	
64	30	20	40 %
128	46	30	42 %
256	98	66	41 %
512	252	186	30 %

Summary:

- Power Consumption (and Cost) of IO is Critical Challenge
 - Threatens performance and scalability
- Quest for System-Wide 1 pJ/bit
 - Ultra bandwidth dense photonic links
 - Energy efficiency squeezed from every component
 - Co-integration with electronics
- Deeply disaggregated Data Center Architectures
 - Optical connectivity for flexibly assembled DC nodes
- 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

Thank you!

