Codesign for Energy Efficient Computing

Adaptable Memory Systems for the Future of AI and HPC

John Shalf

Department Head for Computer Science

Lawrence Berkeley National Laboratory

Memsys 2025

Georgetown, Virginia

Is it the end of Moore's Law?

"I predict Moore's Law will never end.

That way I will only be wrong once!"

Alan Kay: Communications of the ACM 1989

The fundamental problem with Wires (and data movement):

Moore's Law undermined by data movement (when smaller is not better)

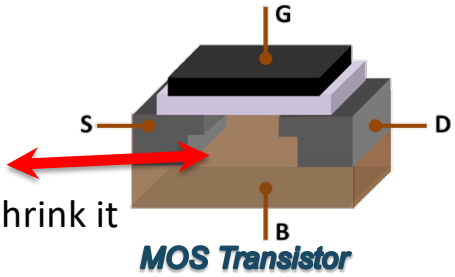
Energy Efficiency of copper wire:

- **Power = Frequency* Length / cross-section-area**

- Wire efficiency *does not improve* as feature size shrinks

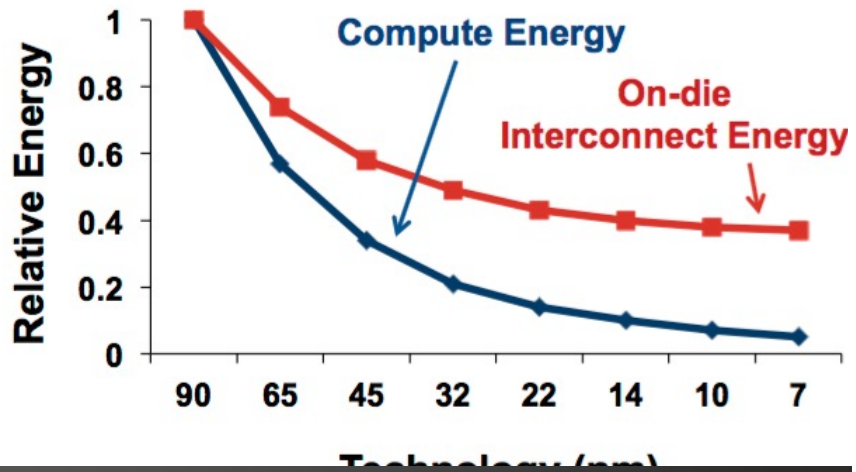
Energy Efficiency of a Transistor:

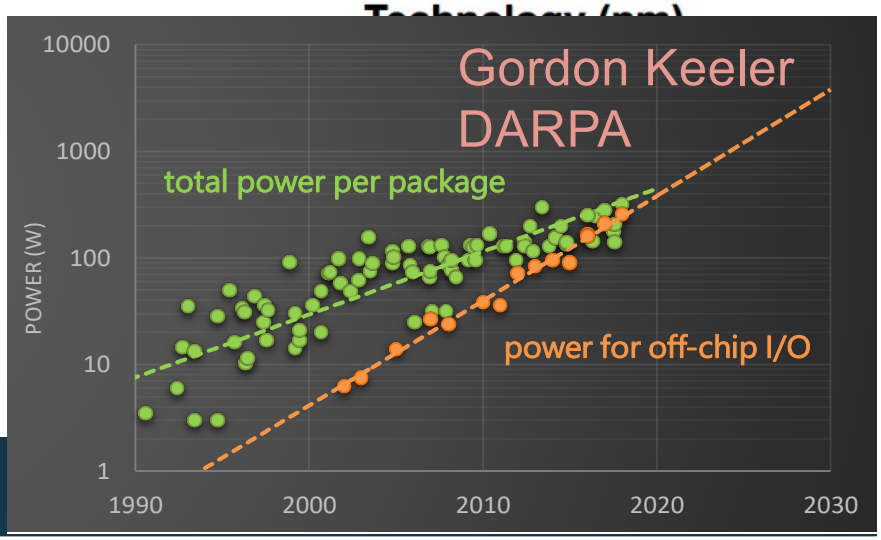
- Power = V^2 * frequency * Capacitance
- Capacitance ~= Area of Transistor
- Transistor efficiency improves as you shrink it



Net result is that moving data on wires is starting to cost more energy than computing on said data

(see also Silicon Photonics)

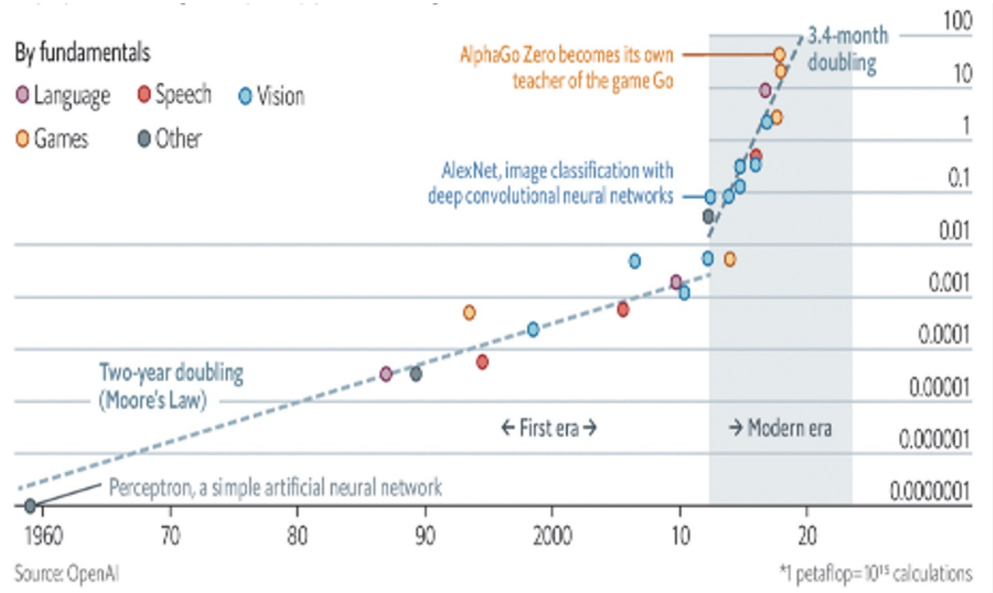




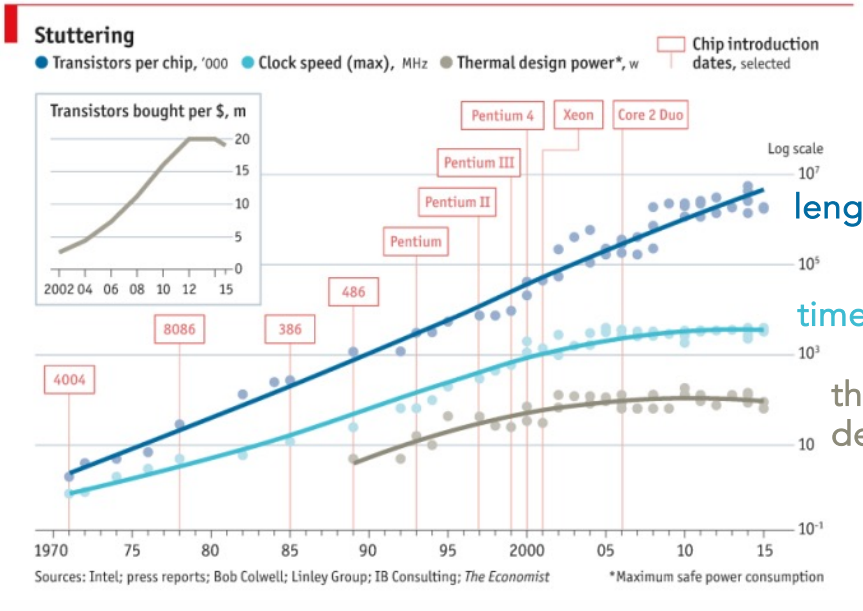
BERKELEY LAB

Explosion of Computing Demand: Driving Need for Hyper-exponential Improvement in Performance, Energy Efficiency and Integration

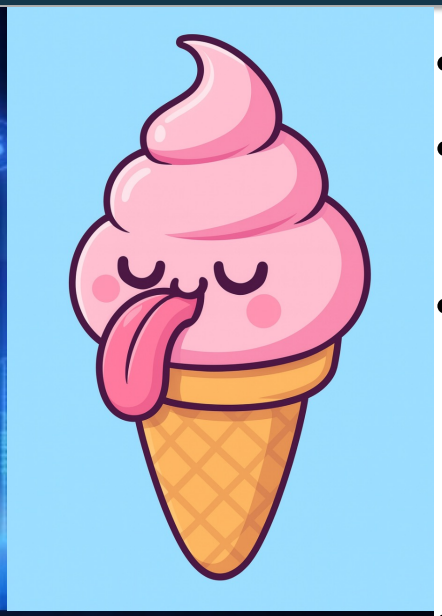
Demand for Computing



Supply for Computing



NVIDIAnomics

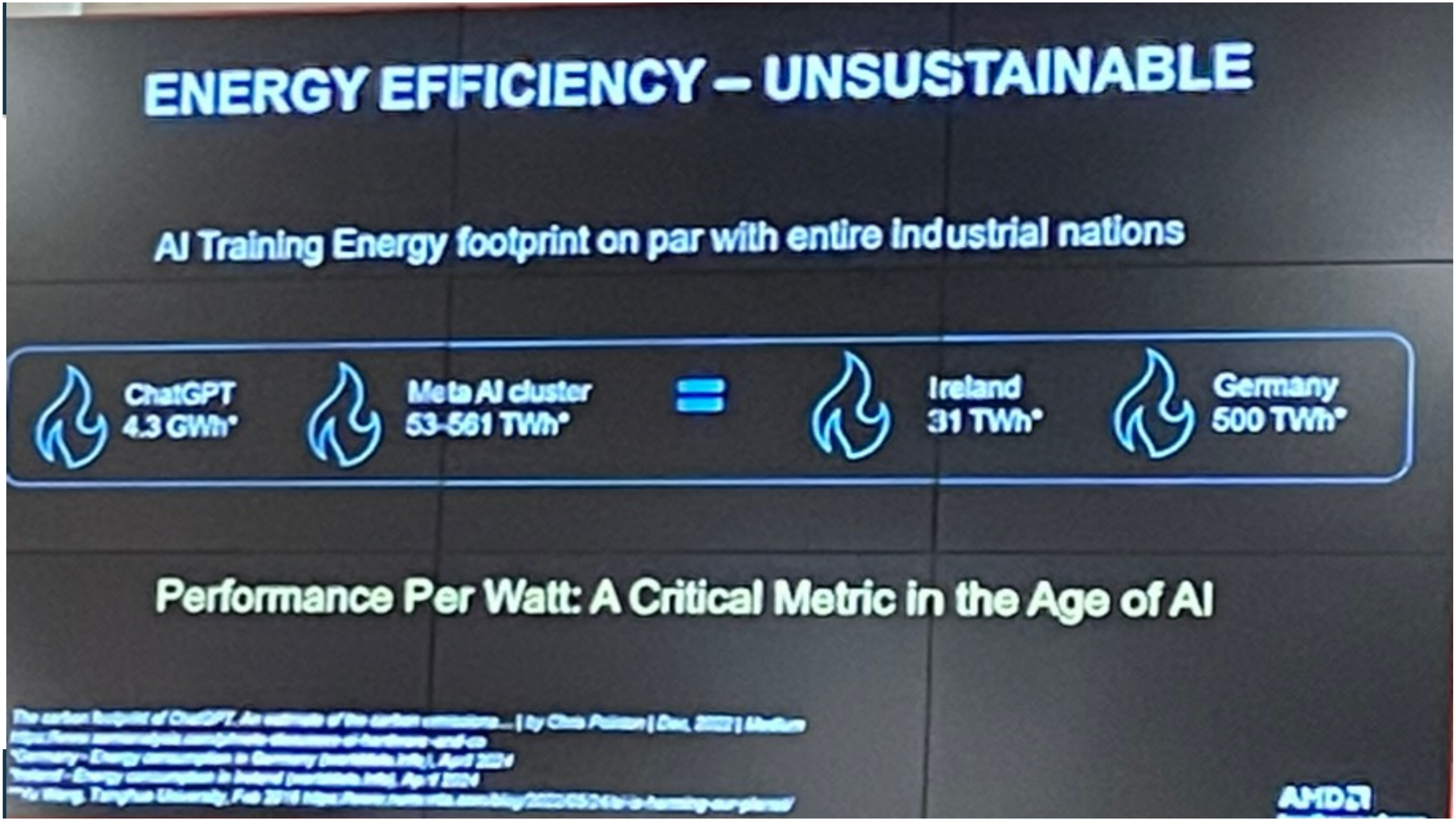


Nvidia Economics: Make $5-$7 for Every $1 Spent on GPUs

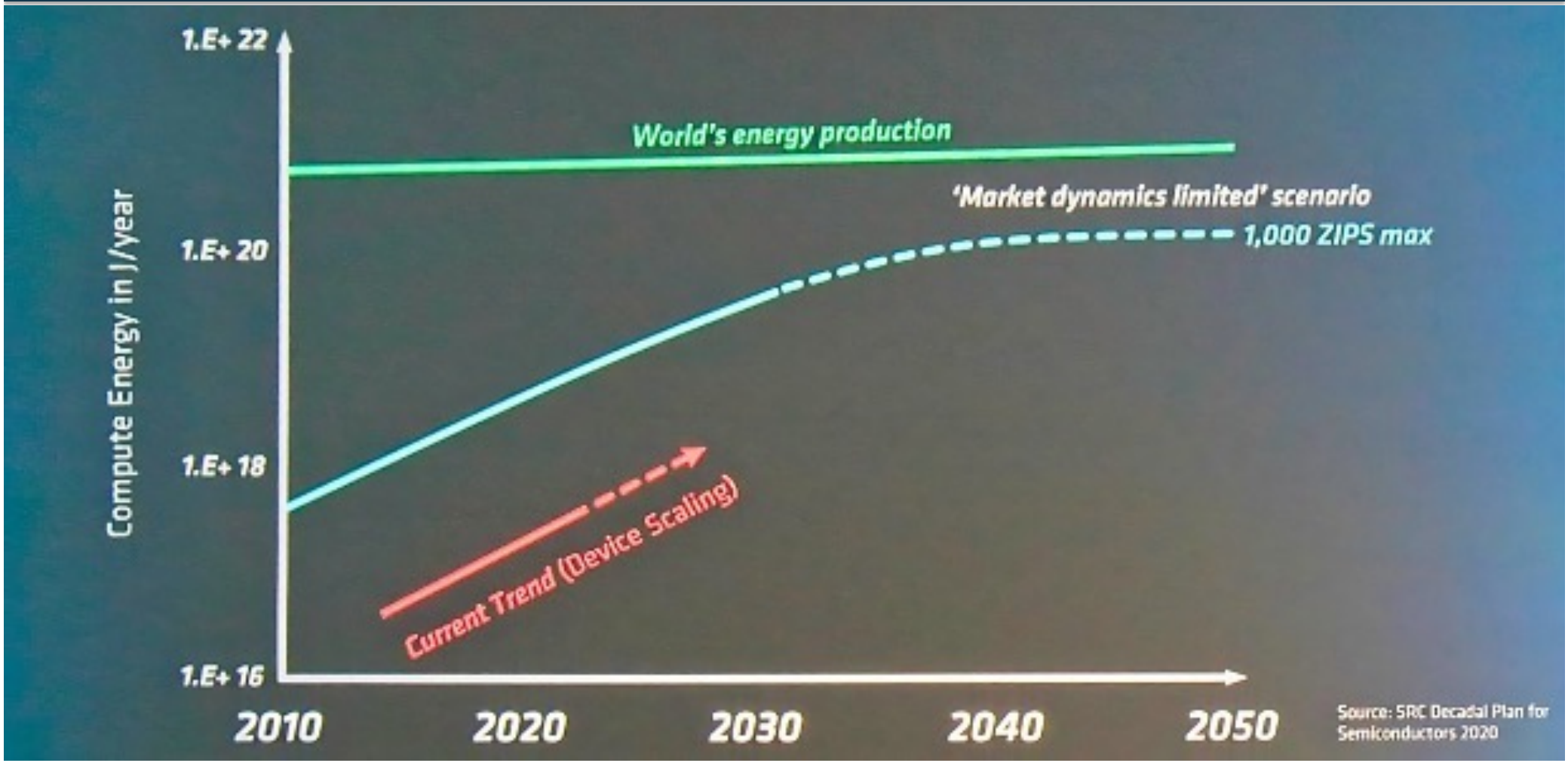
By Agam Shah

- **NVIDIA A100: 250W TDP**
- **NVIDIA H100 SXM has a 700W TDP**
- **Next Generation B100 is projected to consume 1400W TDP**
- **Street price $20k-$30k**
  - Prices lower @ volume
- **And still supply cannot keep up with demand**

BERKELEY LAB


ENERGY EFFICIENCY – UNSUSTAINABLE
AI Training Energy footprint on par with entire industrial nations
ChatGPT
4.3 GWh*
Meta AI cluster
53-561 TWh*
=
Ireland
31 TWh*
Germany
500 TWh*
Performance Per Watt: A Critical Metric in the Age of AI
AMD

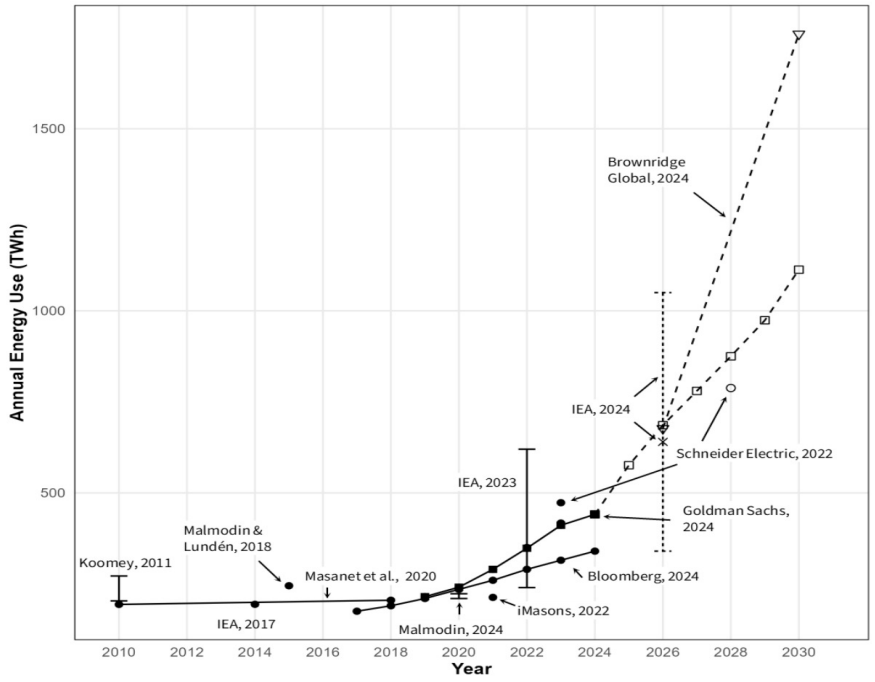
AI Energy Consumption On Pace to Surpass Supply



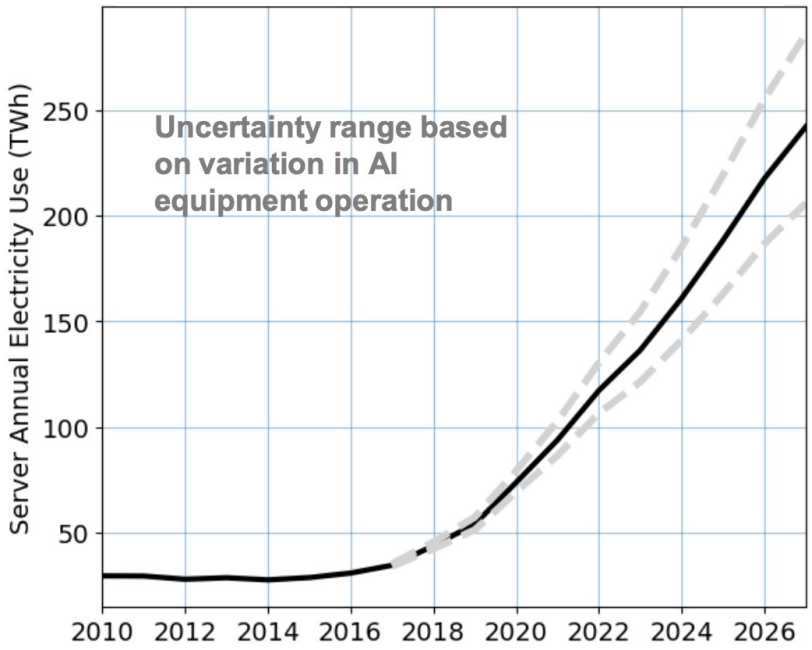
Impact of AI Operational Assumptions on Total Server Estimates

https://datacenters.lbl.gov/sites/default/files/EnergyUsageWebinar12062016.pdf

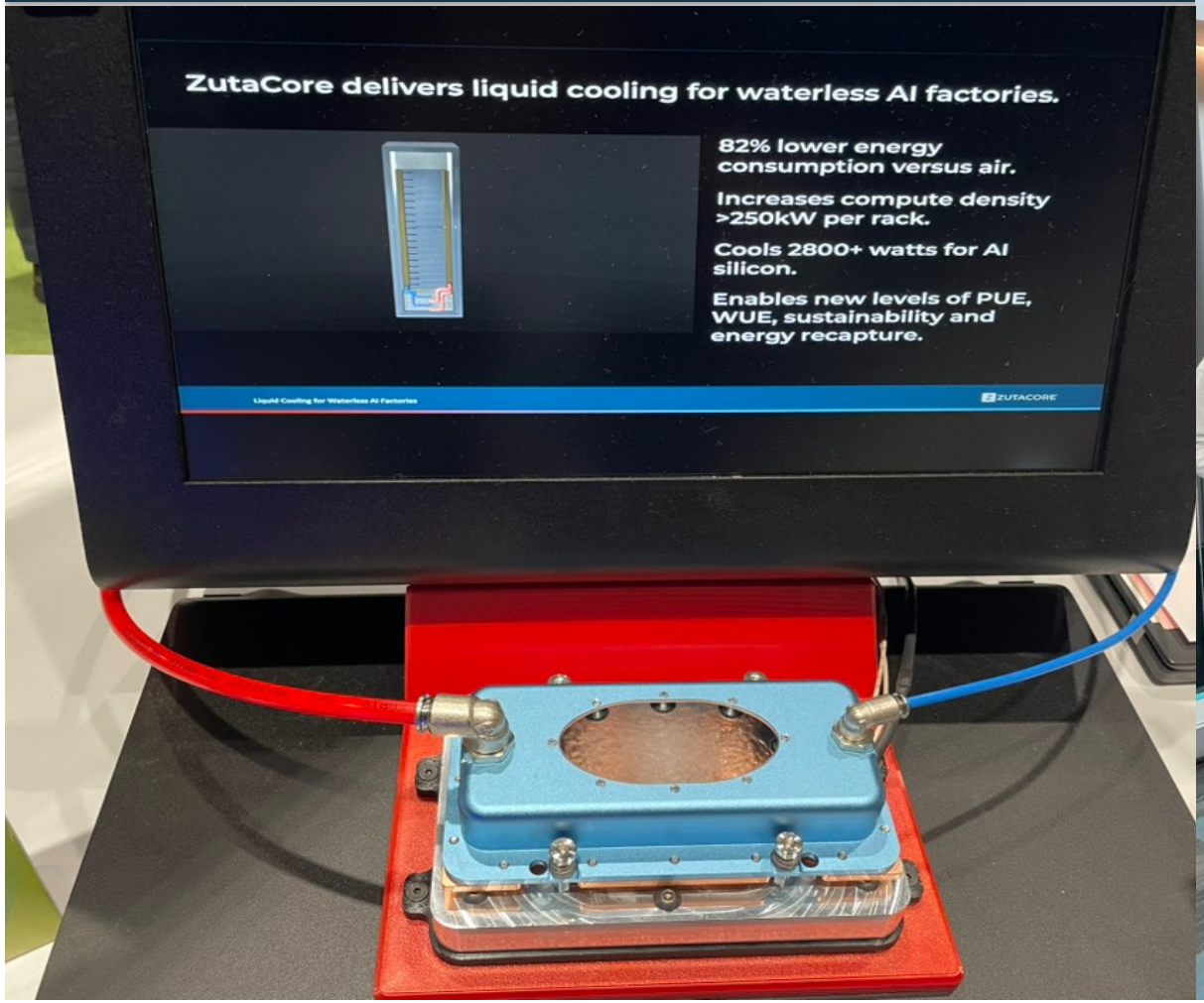
Projected Worldwide AI Energy Use

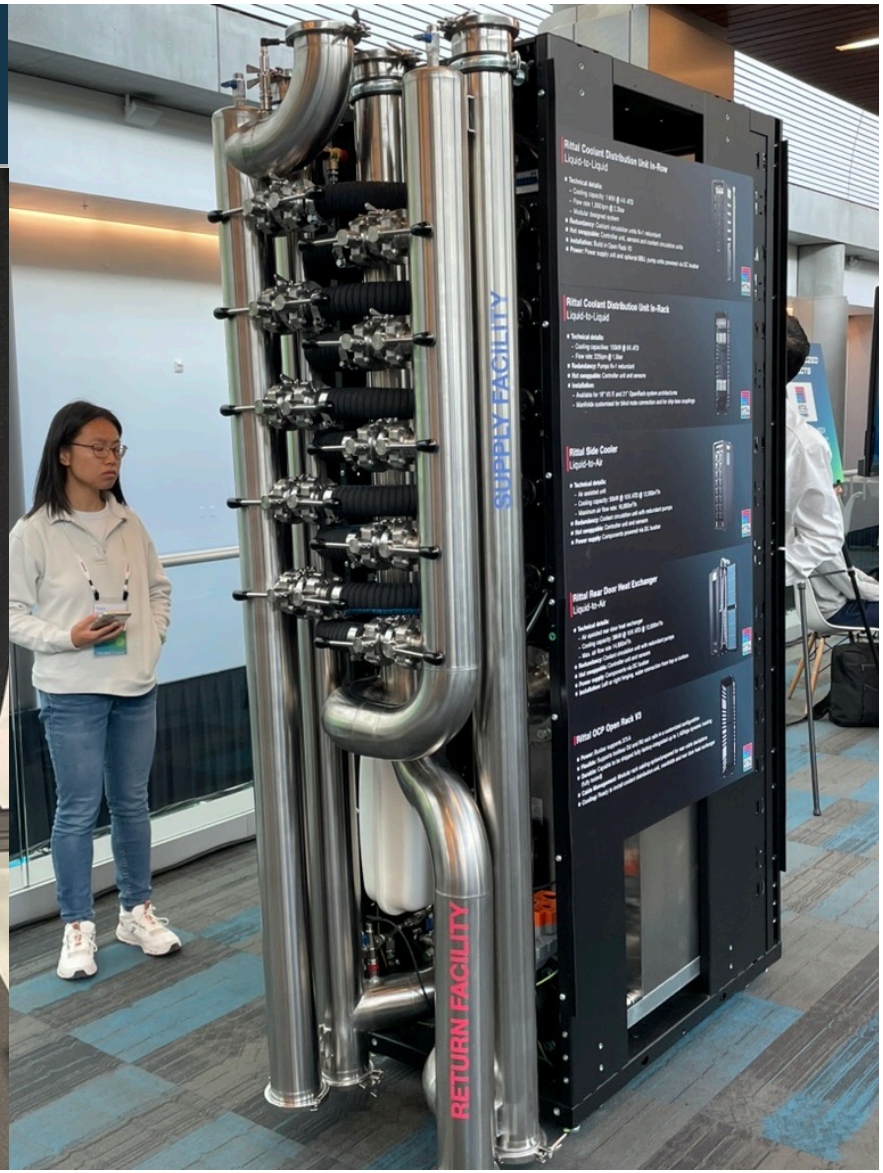


Projected Total US AI Energy Use



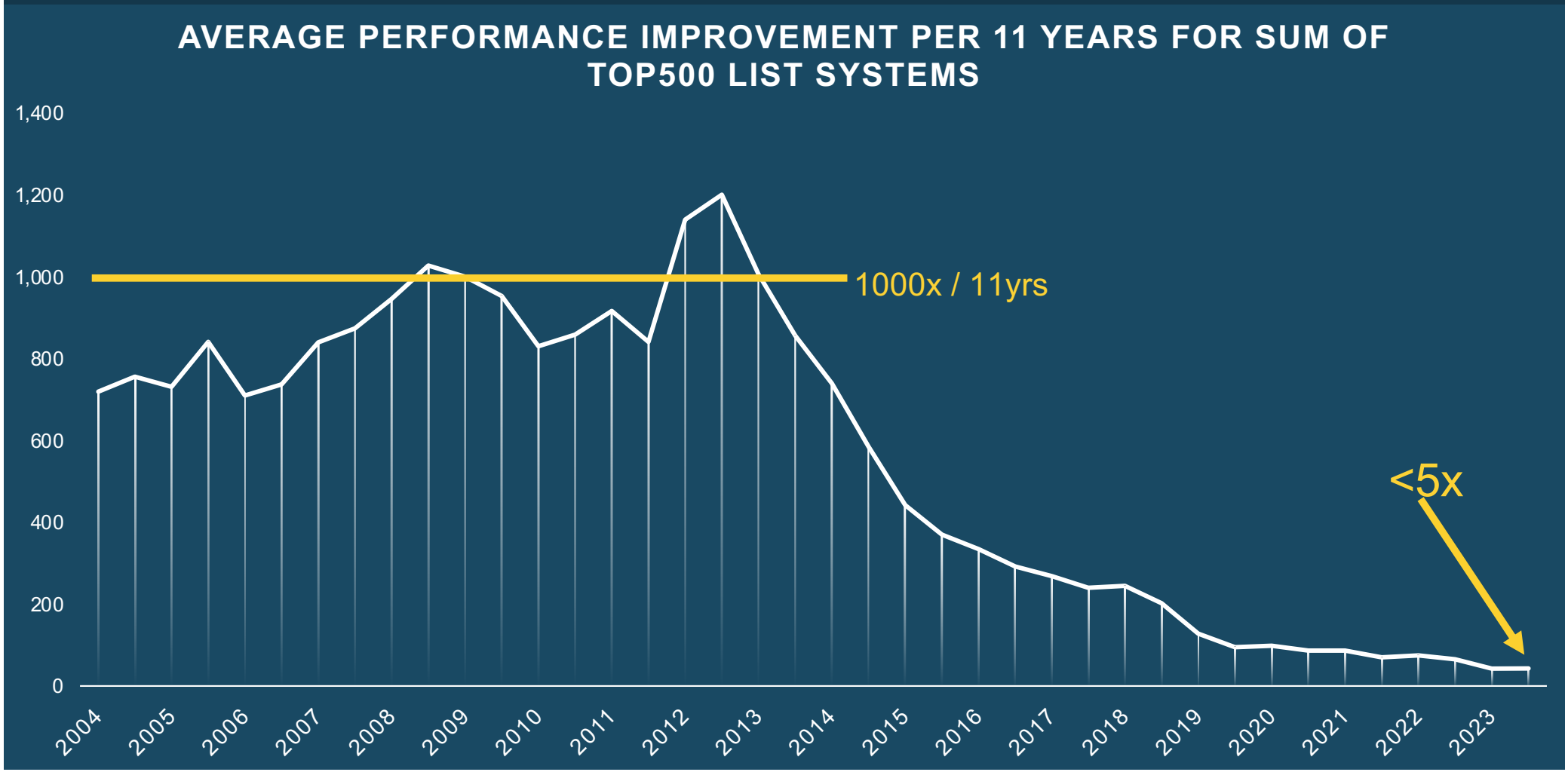
Cooling (3kw socket, 1MW rack)



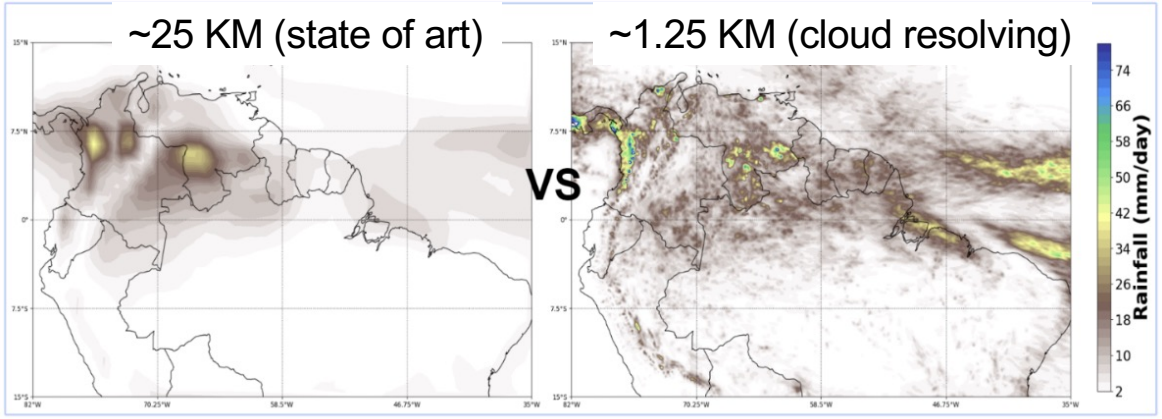


This is HPCs future if we continue business as usual!

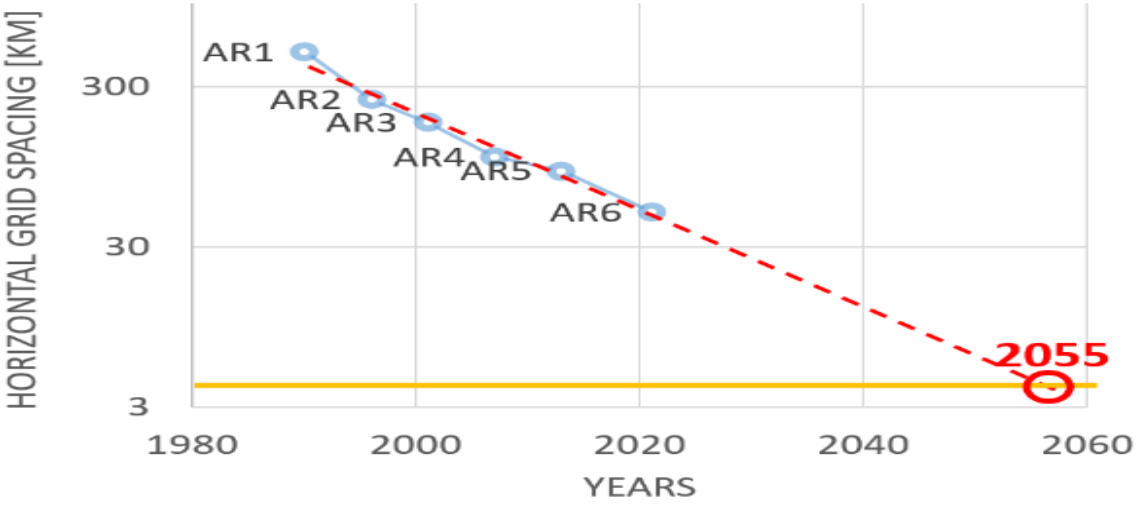
… and scale alone is just power and capital cost…



Example: Kilometer Scale Climate Modeling



Earth System Models



Landmark 3.5KM Simulation on Frontier (exascale) achieved 1.5 simulated years per day performance.

At that rate, for an ensemble calculation it would take ~20 years of dedicated computing to answer important policy questions necessary to achieve 2055 goals.

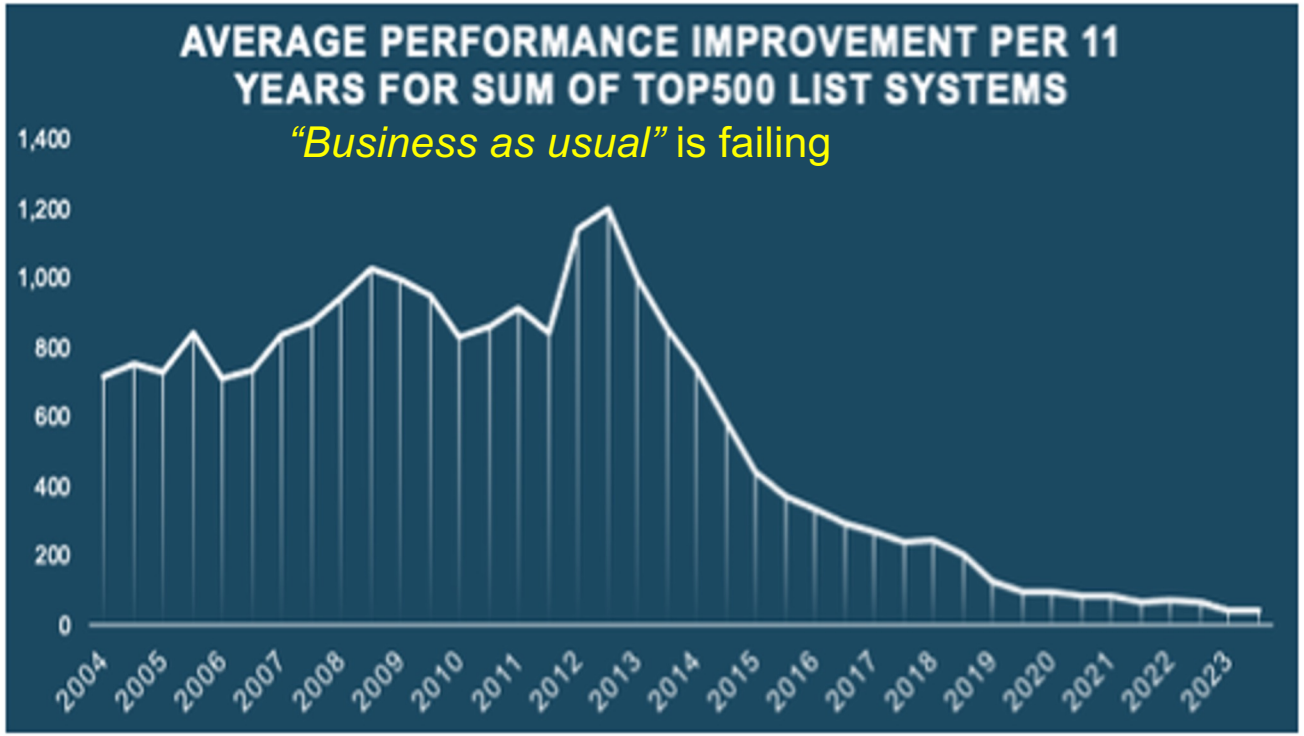
(and that is just for one policy scenario!!!)

Even if we wait for HPC performance improvements, projected 1km modeling goal will be achievable in 2055 at the current rate of progress

This is NOT an acceptable future when there are important scientific imperatives that have ***global societal consequences…***

https://climatecomputer.ncsa.illinois.edu/

Algorithm-Driven Codesign of Specialized Architectures for Energy-Efficient HPC



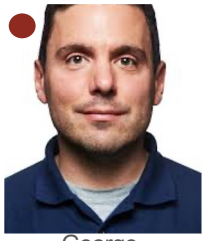
NASEM study on post-Exascale computing *"We must expand (and create where necessary) integrated teams that identify the key algorithmic and data access motifs in its applications and begin collaborative ab-initio hardware development of supporting accelerators,... a first principles approach that considers alternative mathematical models to account for the limitations of weak scaling."*

This is a call for co-design at a much deeper level than we are currently realizing

Addressing cost is crucial: Establish collaborations with industry and the Hyperscale Architects from get-go.

ASCR Competitive Portfolios: Algorithm-Driven Codesign of Specialized Architectures for Energy-Efficient HPC


John Shalf
Thrust lead


George Michelogiannakis


Doru Popovici


Luisa Patricia Gonzalez


Kazutomo Yoshii
ANL


Sophia Shao
UC Berkeley


Xiaokun Yang
U. Houston-Clearlake


Ann Almgren
co-lead


Daniel Martin


Andrew Myers


Weiqun Zhang


Damian Rouson


Alvin Cheung
UC Berkeley

Building on team's expertise in:

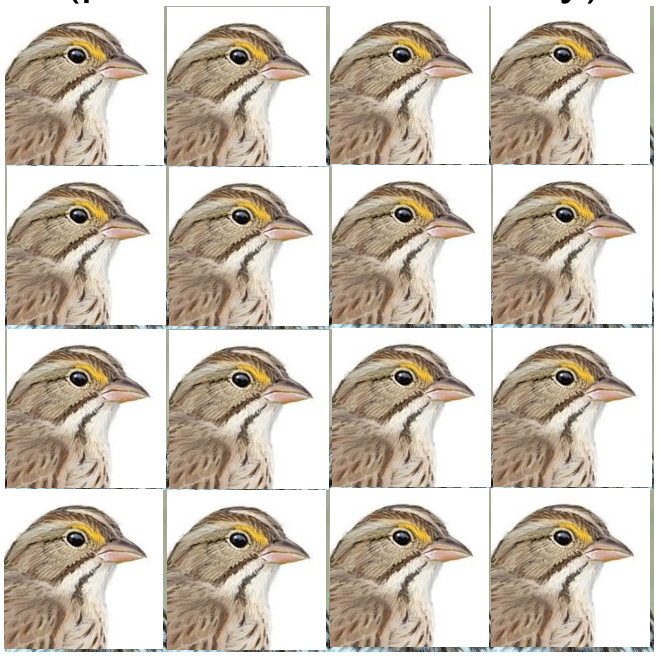
- **Computer Architecture** including hardware design, and hardware generators (CHISEL and GEMMINI accelerator generator), with successful chip tape-outs including compute-in-sensor, mixed-signal and HPC-class prototypes
- **Computer Languages and Compiler Systems** including ML-assisted code generation and Verified Lifting
- **Applied Mathematics** including algorithms coupling models suitable for different scales; Adaptive Mesh and Algorithm Refinement (AMAR) paradigm, PDEs on structured grids, spectral methods (FFTs), particle methods, and scalable solvers

Specialization:
Natures way of Extracting More Performance in Resource Limited Environment

Powerful General Purpose

Xeon, Power

Many Lighter Weight
(post-Dennard scarcity)



KNL AMD, Cavium/Marvell, GPU

Many Different Specialized
(Post-Moore Scarcity)

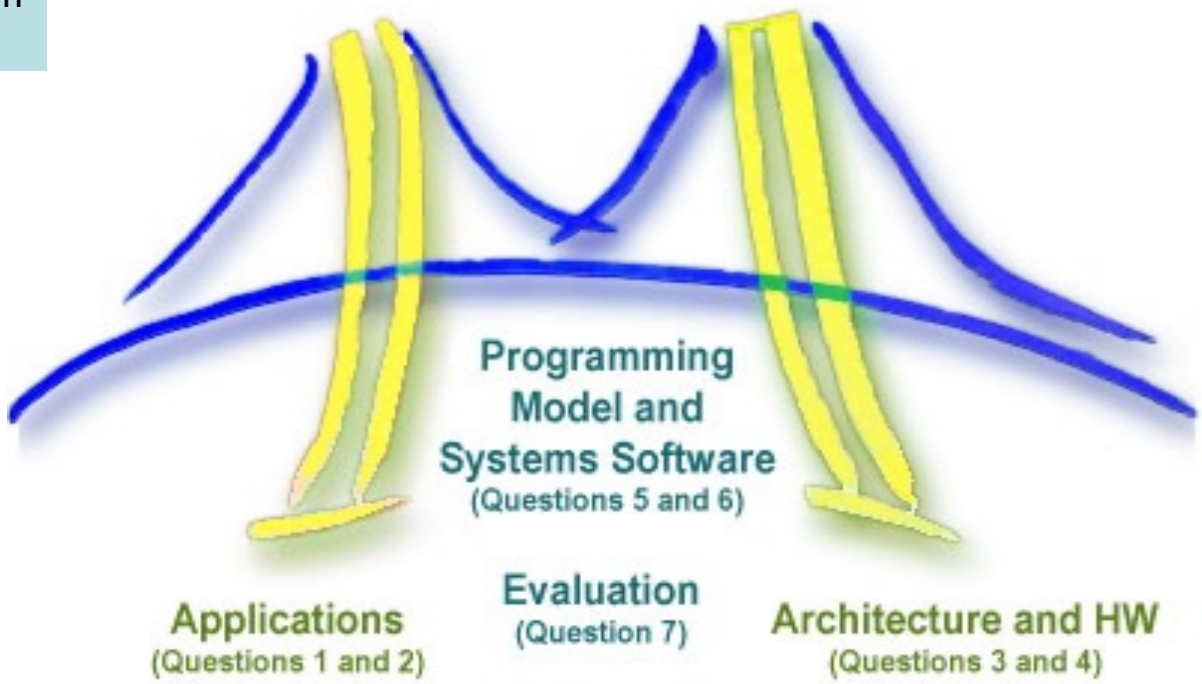
Apple, Google, Amazon, Microsoft Azure

Phil Colella's 7 Dwarfs of Scientific Computing

High-end simulation in the physical sciences = 7 numerical methods:

Exploit the mathematical structure of the problem
design principle for "analogous" computing

1. Structured Grids
2. Unstructured Grids
3. Fast Fourier Transform
4. Dense Linear Algebra
5. Sparse Linear Algebra
6. Particles
7. Monte Carlo



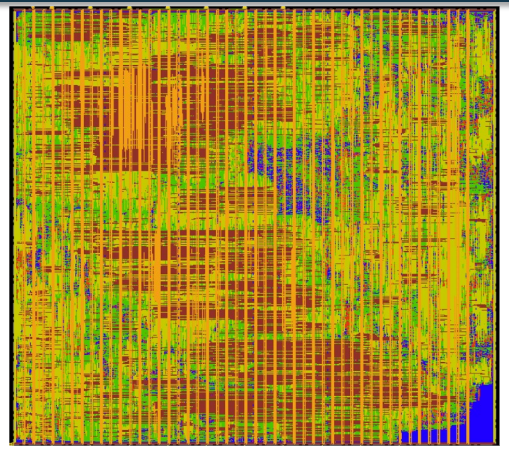
Slide from "Defining Software Requirements for Scientific Computing", Phillip Colella, 2004
Also in "The Landscape of Parallel Computing Architecture: A view from Berkeley" 2008

http://www2.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-183.pdf

Office of Science
U.S. DEPARTMENT OF ENERGY

Example of Mixed-Radix MultiDimensional FFT Accelerator

Initial steps towards a basis set of hardware accelerator primitives for science



FFT96 Accelerator Die: 16nm TSMC

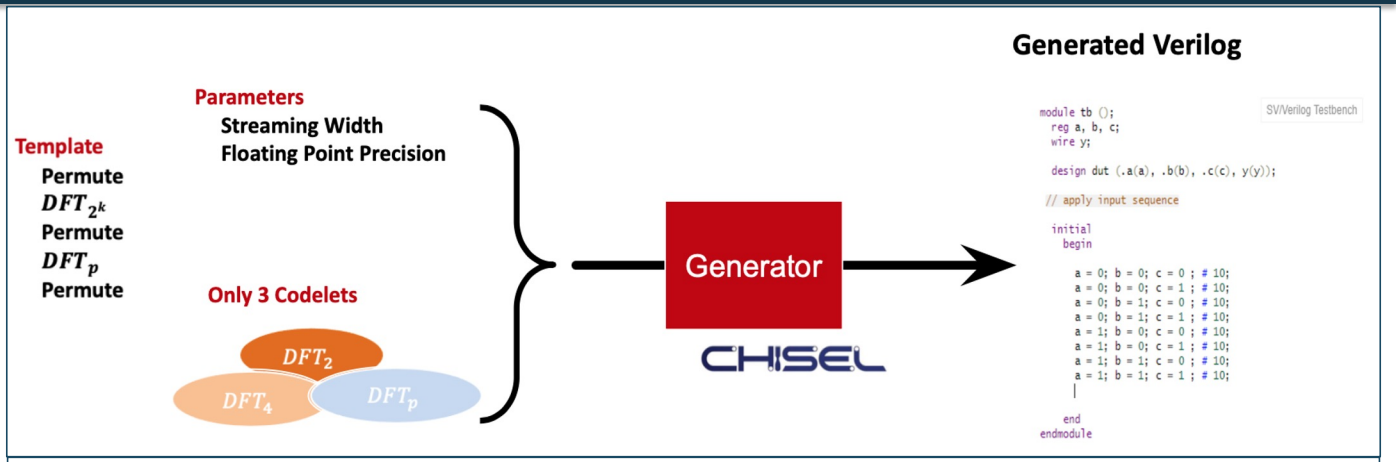
Area eff.: 4.18 TF/mm^2

Energy eff.: 4.8 TF/W

NVIDIA H100 in 4nm TSMC *(10x denser than 16nm)*

Area eff: 0.08 TF/mm^2

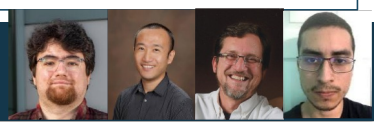
Energy eff: 0.0957 TF/W



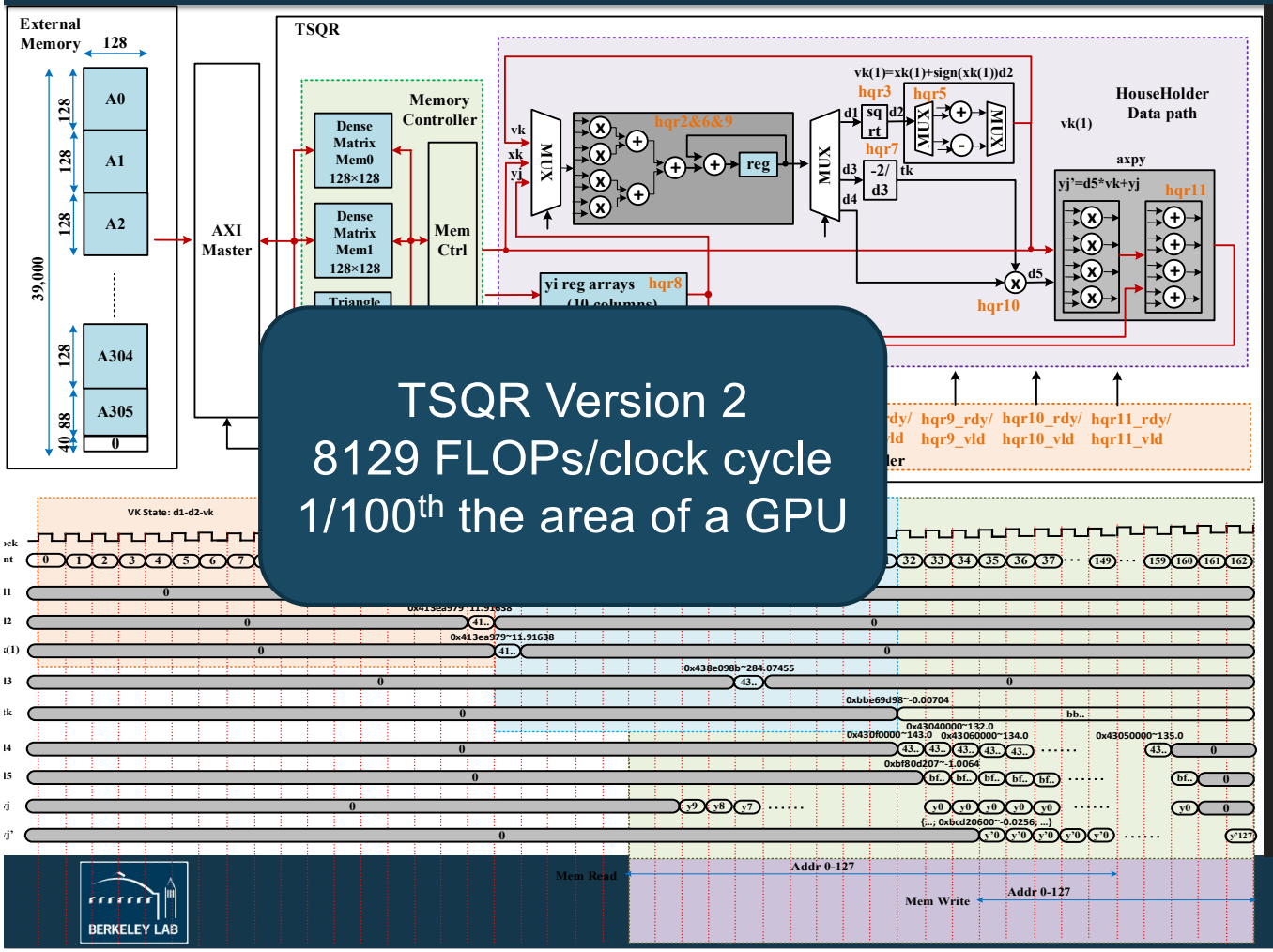
Accomplishment: Demonstrated general FFT Accelerator tile generator in 16nm TSMC that outperforms NVIDIA H100 by 50x in raw performance/area and 50x in energy efficiency/flop

–Required only 3 codelet primitives

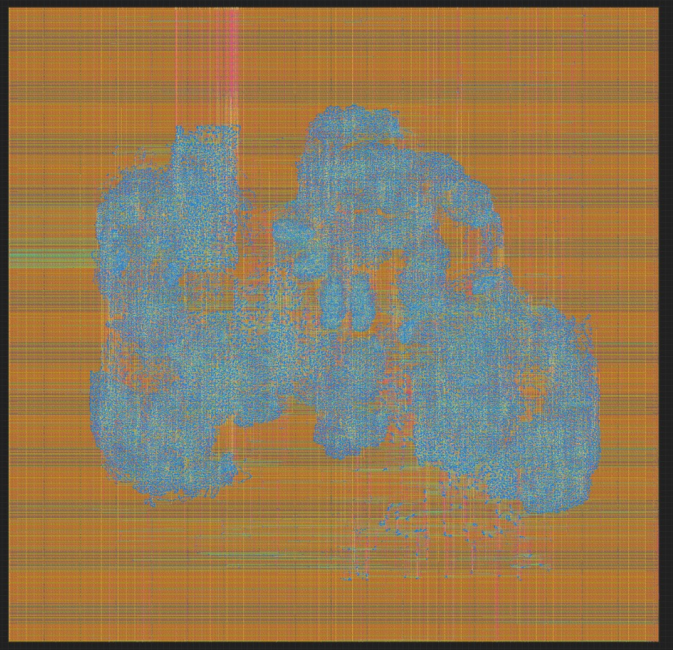
Goal: Generalize this approach to generators for accelerator hardware primitives that cover broad spectrum of algorithms *(e.g., FFT, Dense/Sparse Linear Algebra, particles and PDEs)*



Next Steps – Implement as ASIC (OpenROAD)



TSQR Version 2
8129 FLOPs/clock cycle
1/100th the area of a GPU

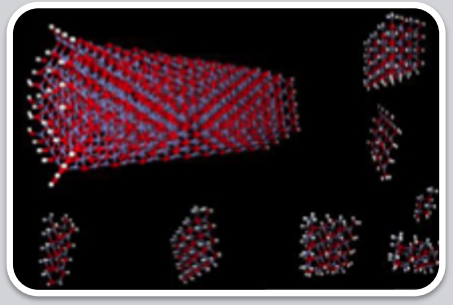


Chip: FFT9_fpuv3
Node: freepdk45
Area: 4723210.000um^2
Fmax: 390.346MHz

BERKELEY LAB

Architecture Specialization for Science

(hardware designed for the algorithms) can't design effective hardware without math



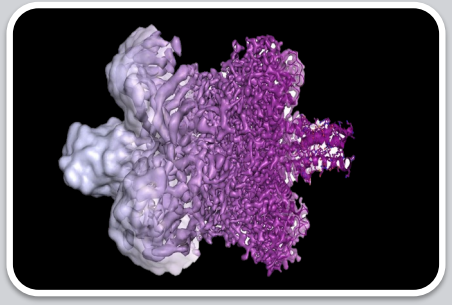
Materials

Density Functional Theory (DFT)

Use O(n) algorithm

Dominated by FFTs

(1/3 of DOE workload)



Smart Sensors

CryoEM detector

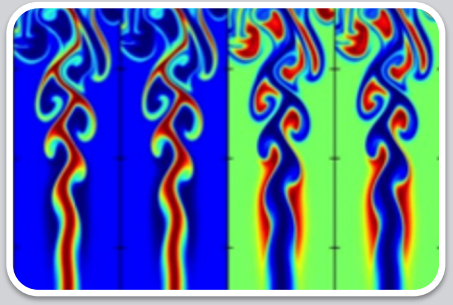
1Terabyte / sec

Compute at data source (data2info)

Genomics

String matching

Ultra-narrow datapath 2-8bit (ACTG)



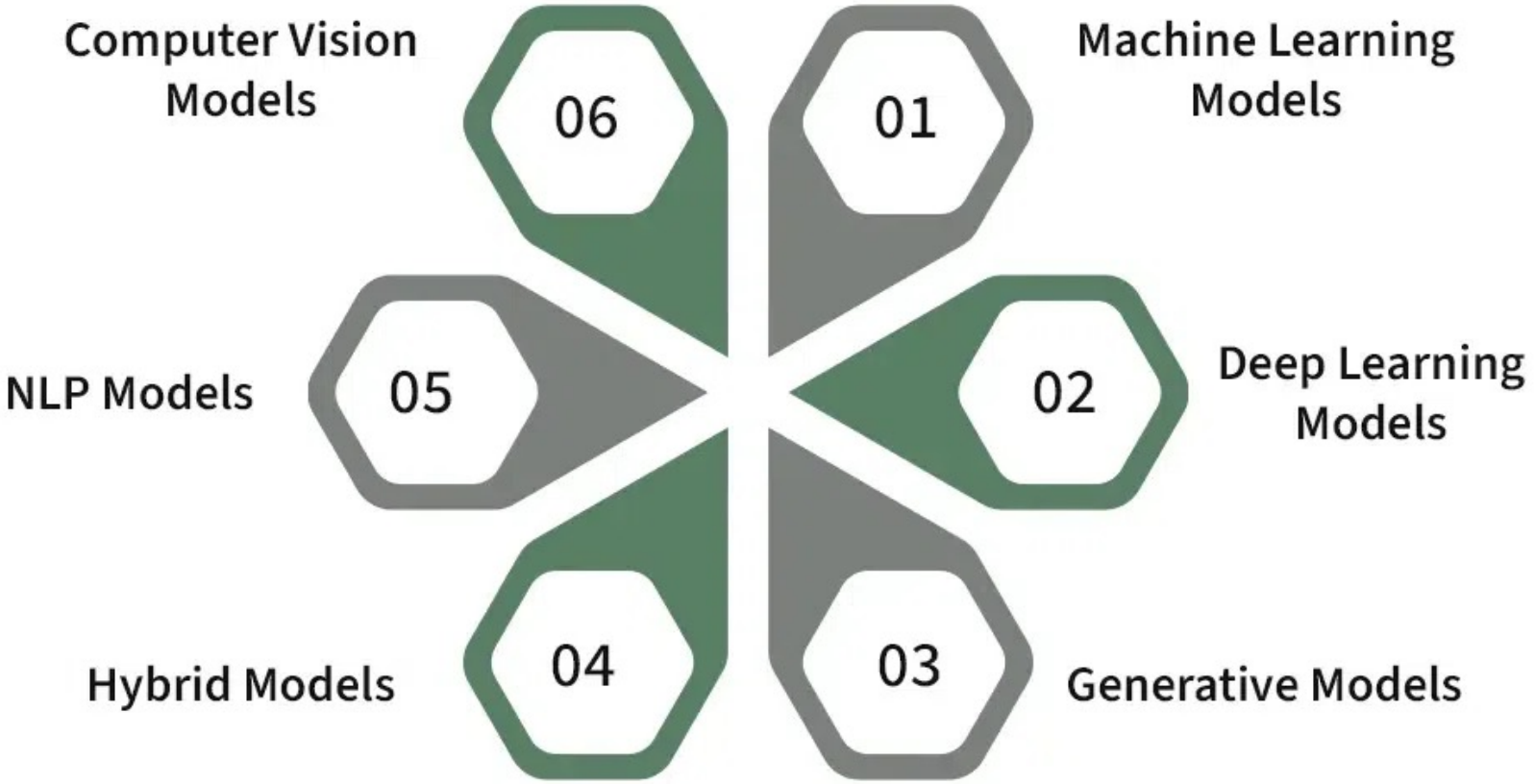
PDEs on Block Struct. Grids

Extreme Strong

Scaling for CFL limited problems

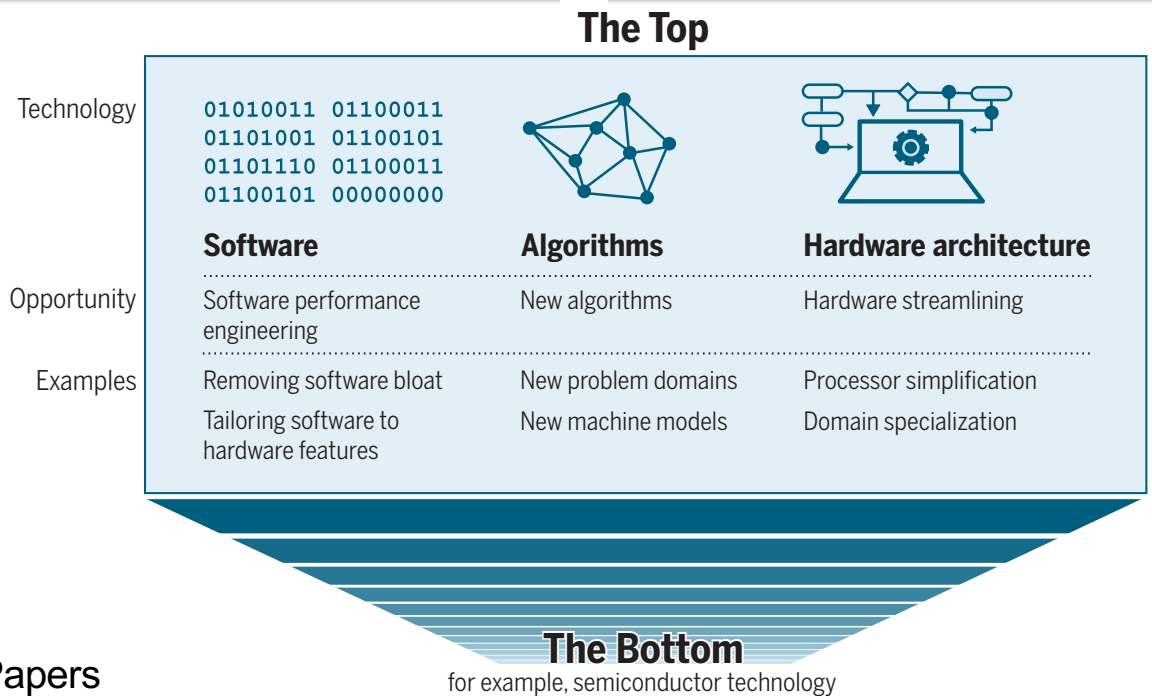
(bigger is not better)

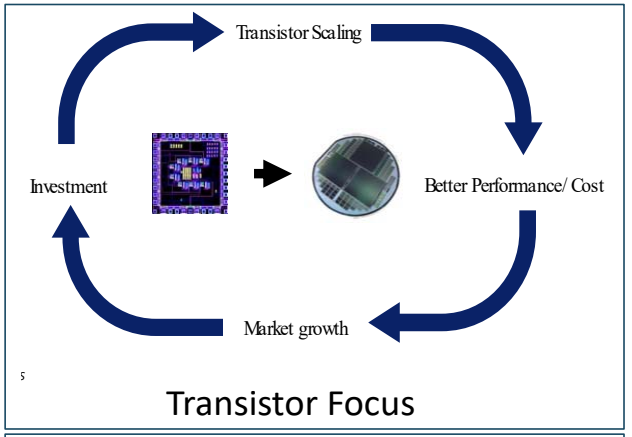
There are in fact *many different kinds* of AI models with significantly different requirements

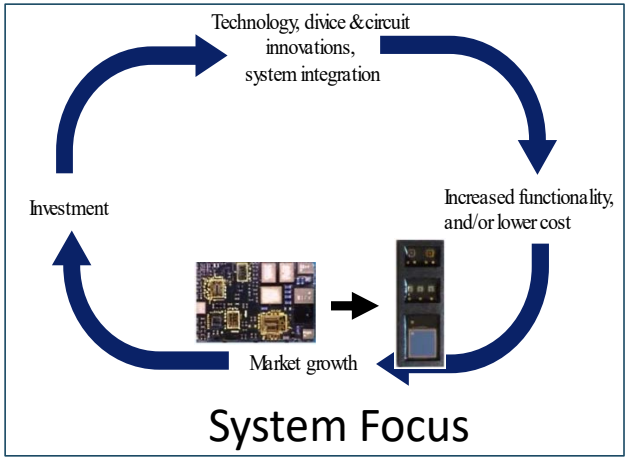


Neil Thompson: Economics of Post-Moore Electronics

http://neil-t.com, MIT CSAIL, MIT Sloan School





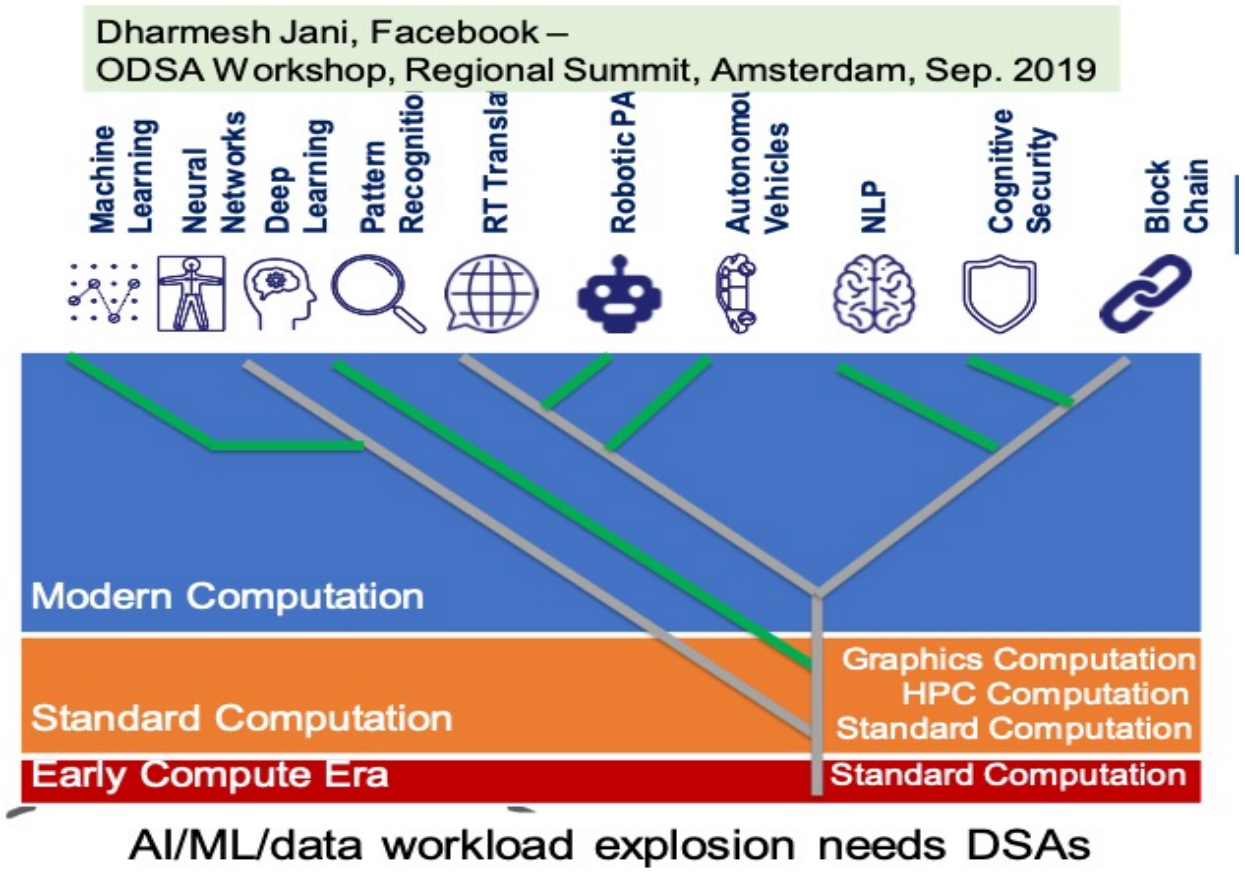


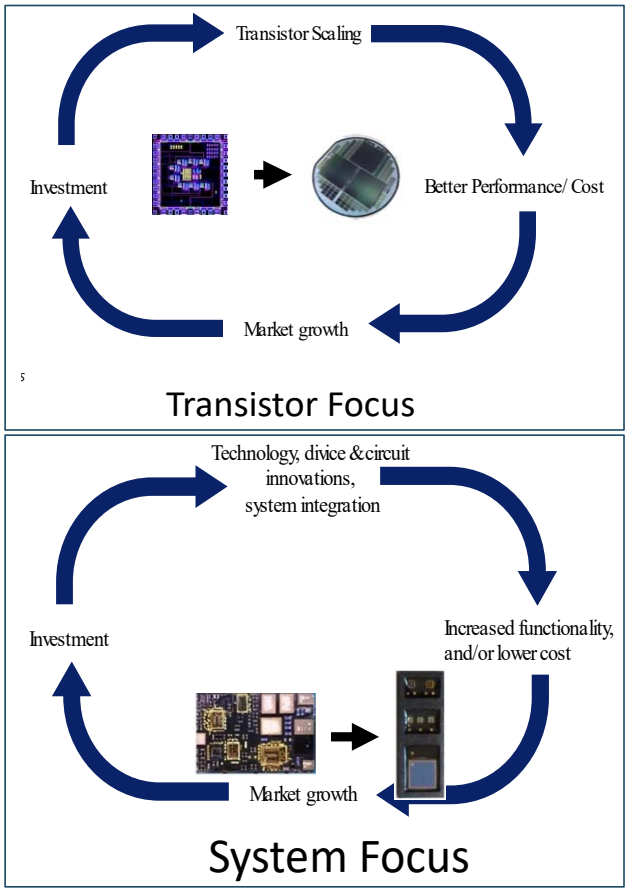
Papers

1. The Economic Impact of Moore's Law
2. There's Plenty of Room at the Top: What will drive computer performance after Moore's Law?
3. The Decline of Computers as a General Purpose Technology

Domain specific Architectures driven by hyperscalers

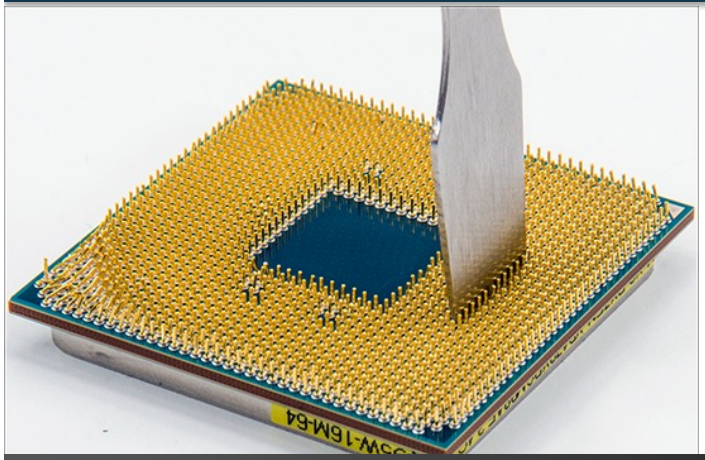
in response to slowing of Moore's Law (switch to systems focus for future scaling)

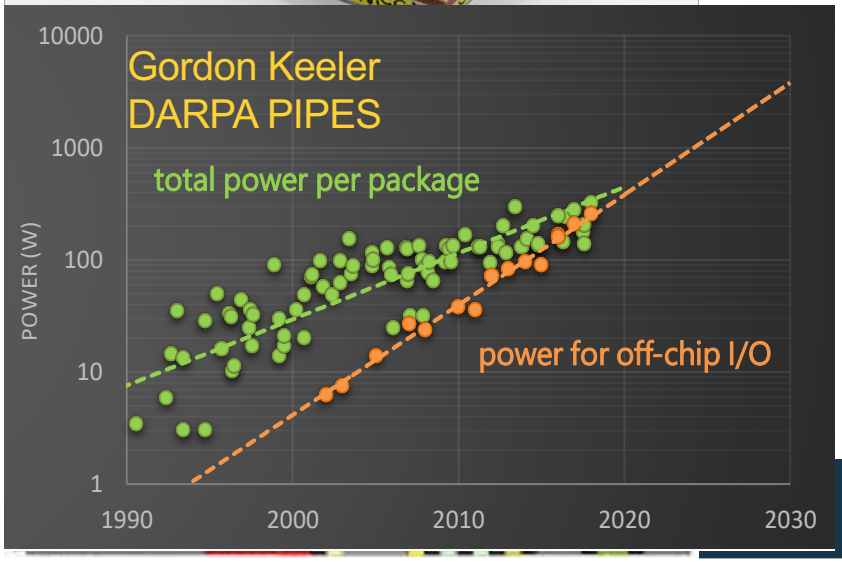




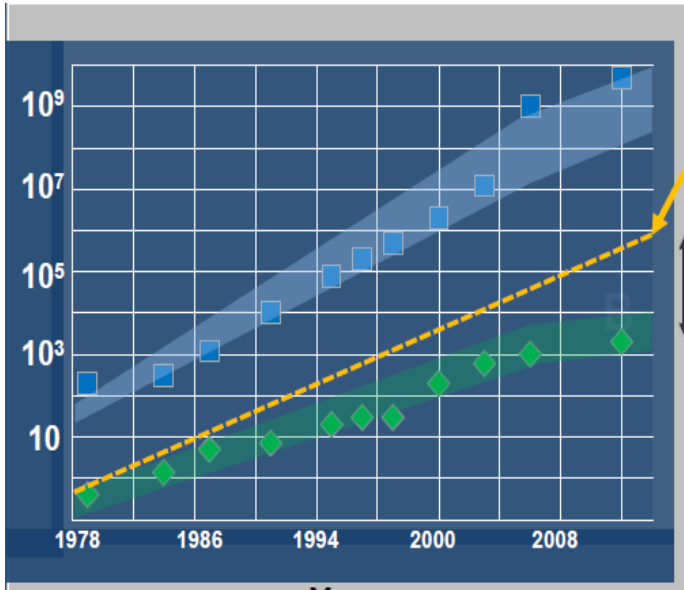
Why Advanced Packaging & Chiplets?

Because Package Performance is Pin Limited





Rent's Rule:
Number of pins = K x Gates^{a} (IBM, 1960)
K = 0.82, a = 0.45 for early Microprocessors



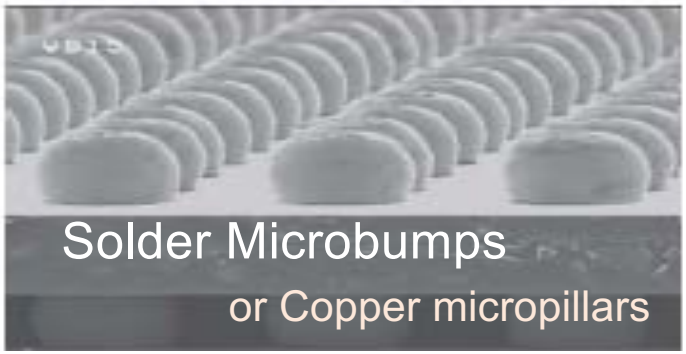
Pins x GHz from Rent's Rule

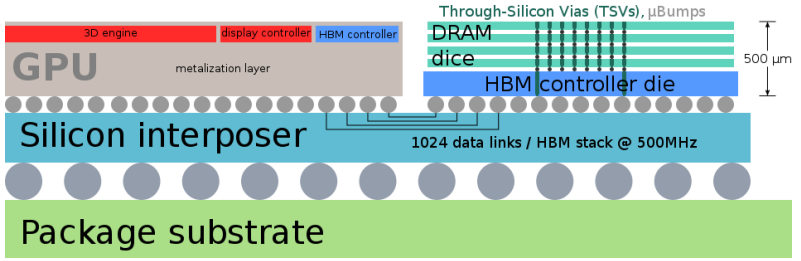
Bandwidth Gap:
~500 x and growing!

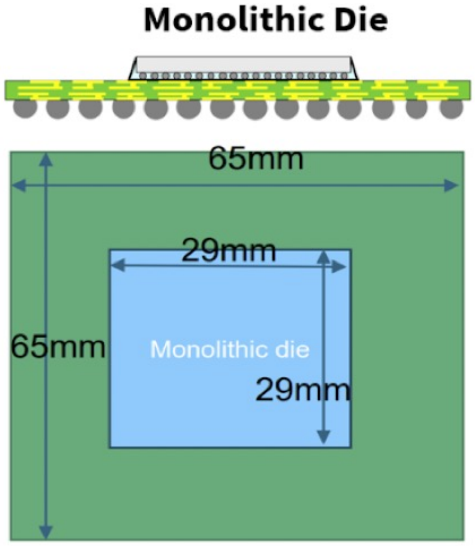
High SERDES rates run counter to end of Dennard Scaling

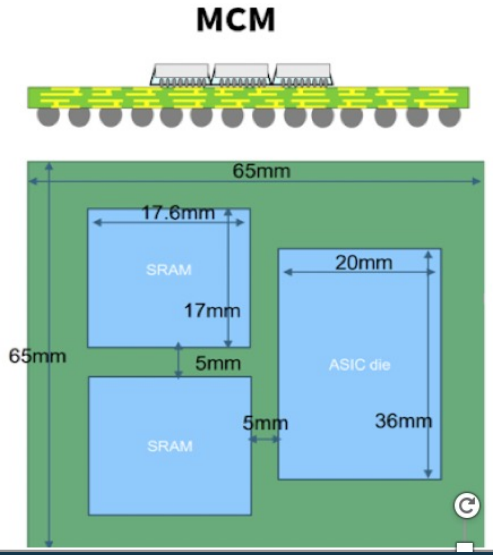
Source: J. Poulton, Nvidia

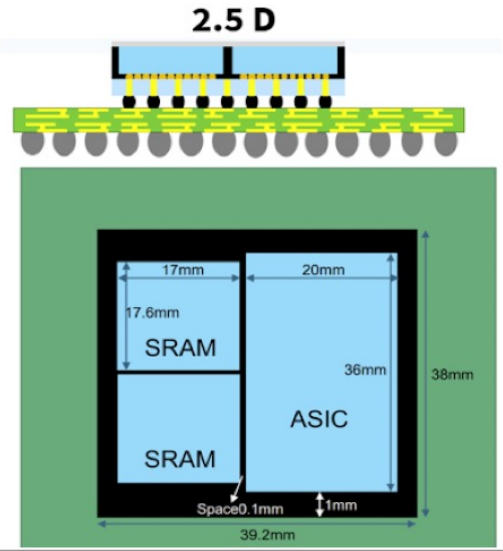
What is a Chiplet?

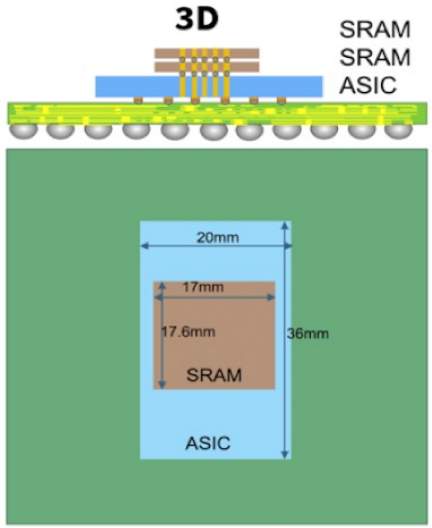






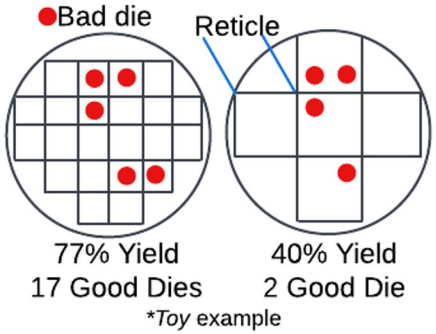






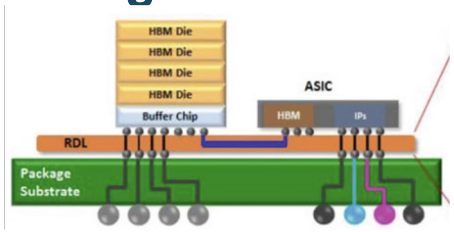
Chiplets have many well known advantages

Yield and cost efficiency



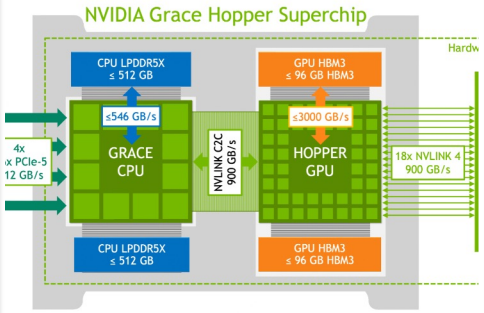
Overcome the memory wall

HBM is 3D integration



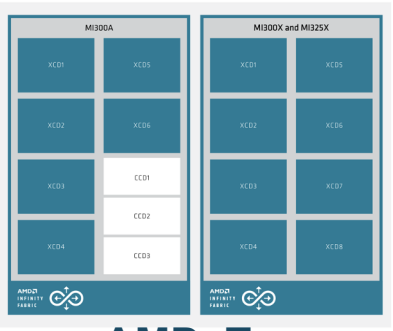
Lau. et al. Chiplet design and heterogeneous integration packaging

Heterogeneous integration



CPU + GPU

Scalability & Modularity/Reuse

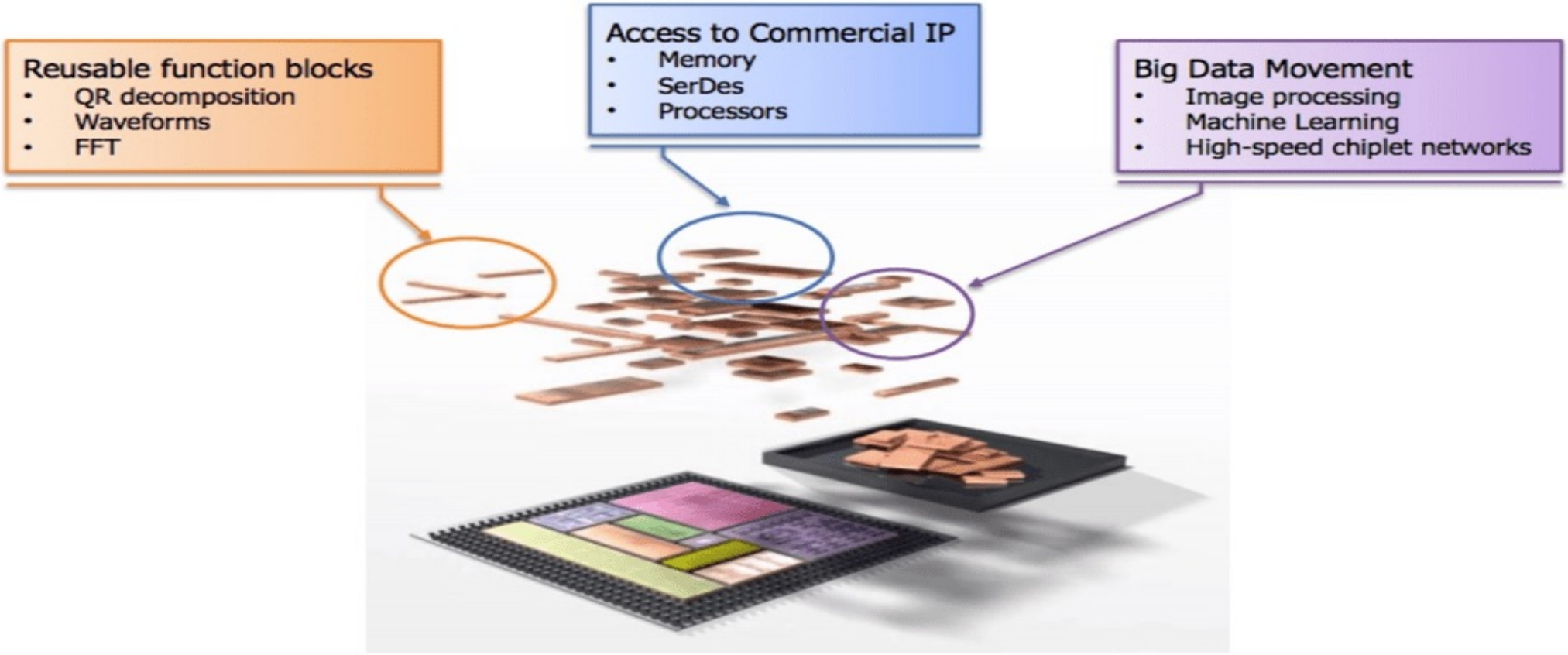


AMD. Two different products. Same chiplets

How do chiplets enable domain specialization?

Lower cost barriers to co-integrating specialization

From DARPA CHIPS

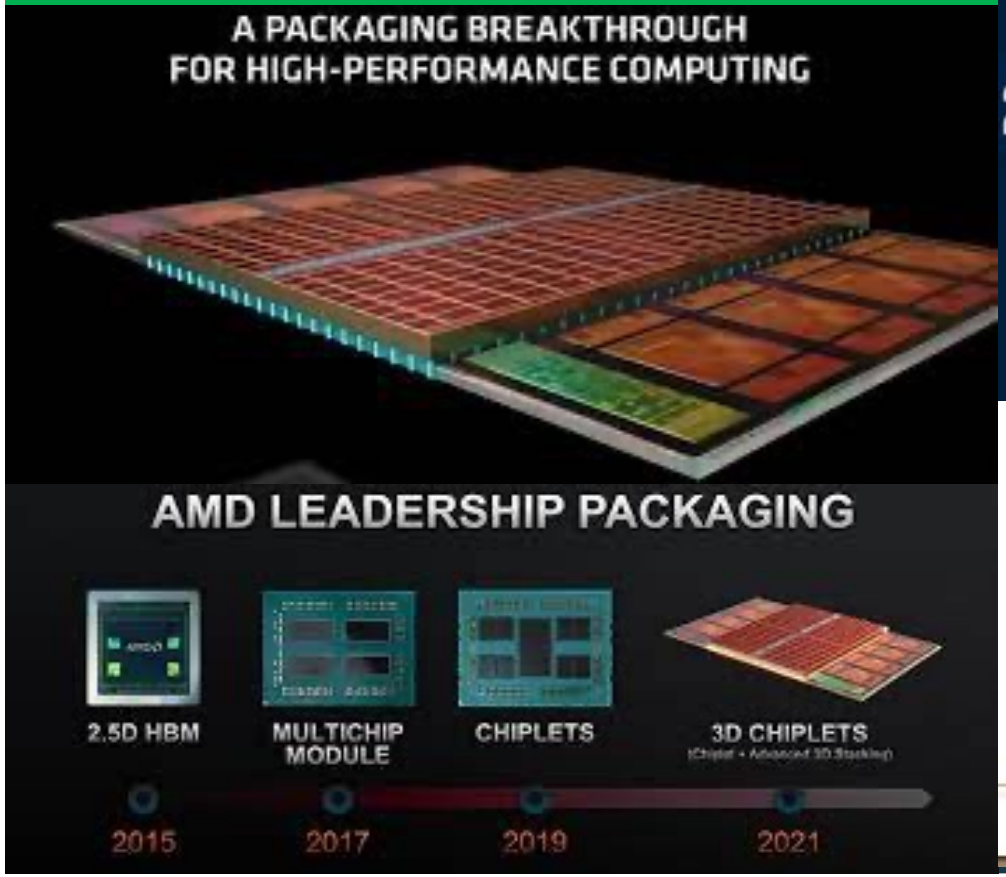


See the multi-agency chiplets workshop at https://sites.google.com/lbl.gov/chiplets-workshop-2023/home

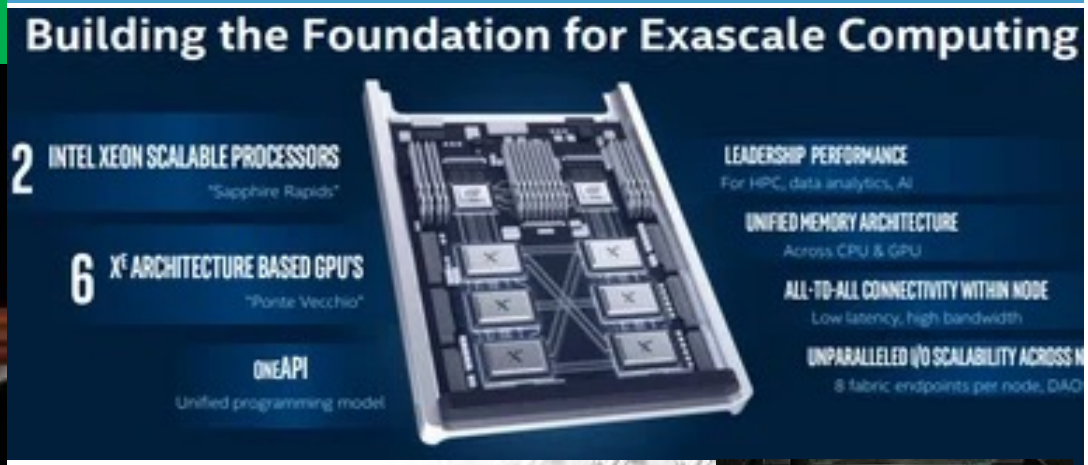
CHIPS modularity targets the enabling of a wide range of custom solutions

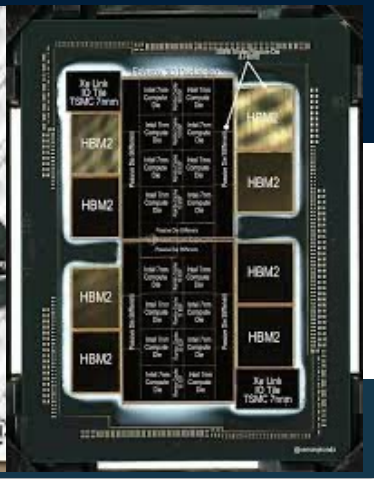
Heterogenous Integration Driving Exascale HPC

AMD Advanced Packaging for OLCF Frontier



Intel Ponte Veccio for ALCF Aurora





Technology Insertion into Mainstream Platforms

AMD, Intel, Arm offer integration path for 3rd party accelerator "chiplets"

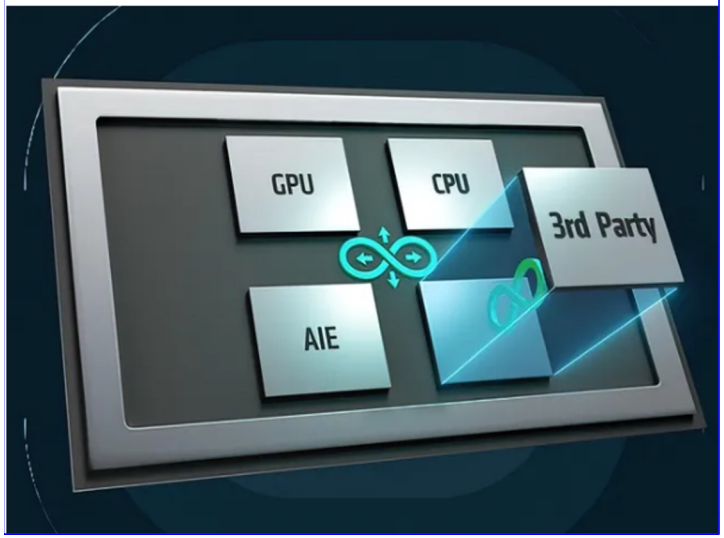
Modular AMD Chips to Embrace Custom 3rd Party Chiplets

News By Francisco Pires last updated June 20, 2022

Supercharging learnings - and earnings - from the console space.

Comments (2)

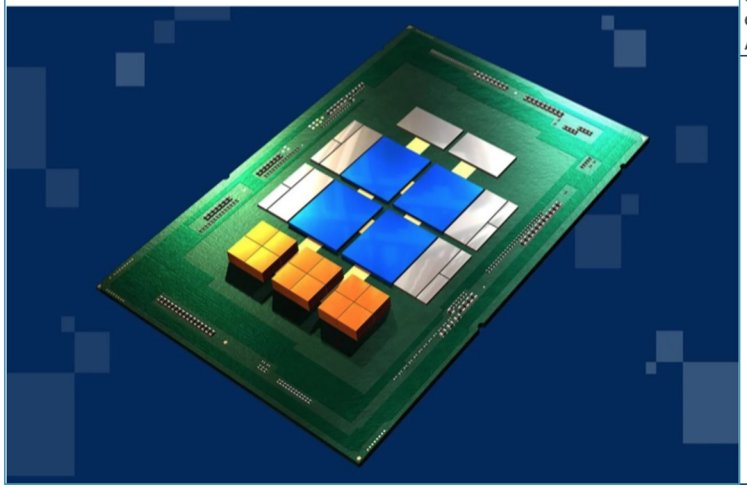
When you purchase through links on our site, we may earn an affiliate commission. Here's how it works.



To 'Meteor Lake' and Beyond: How Intel Plans a New Era of 'Chiplet'-Based CPUs

At the Hot Chips 2022 conference, Intel teased its upcoming 'Meteor Lake' and 'Arrow Lake' processor families, which will use multiple tiny tiles fused together in an attempt to break free of the limits of monolithic chip design. Here's why little tiles are a big deal.

By Michael Justin Allen Sexton August 24, 2022

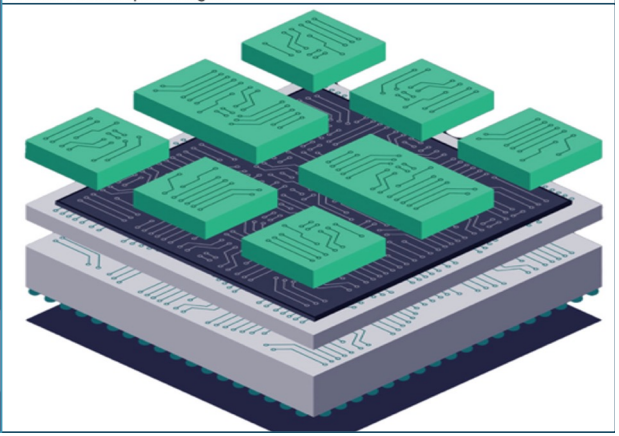


ARM Opens Door to Make Custom Chips for HPC, AI

By Doug Eadline

October 19, 2023

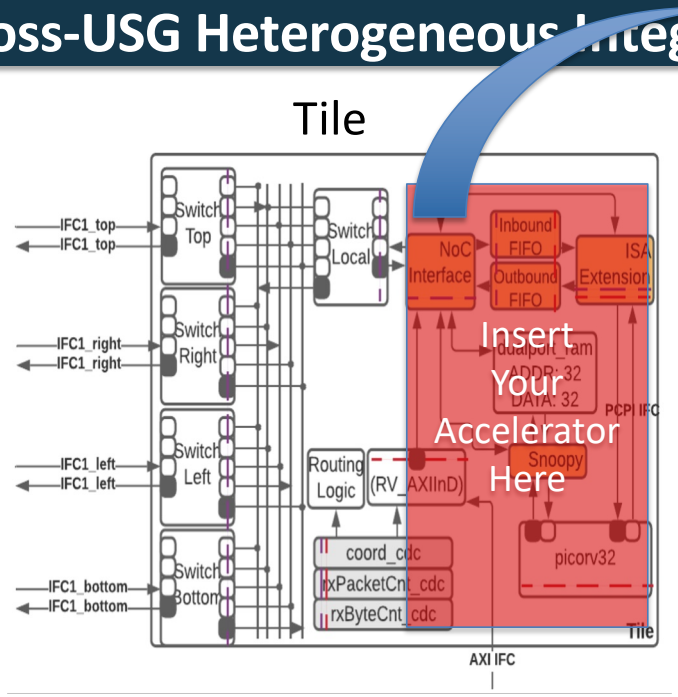
It is safe to say that ARM isn't a scrappy startup that was once the pride of the UK. The US-based IPO made the chip designer a big-game chip player, and the new capital is kickstarting some major initiatives to find more customers for its products. A new effort called Total Design aims at making it easier for companies looking to design chips in-house, an idea gaining ground with the AI boom and chip shortages.

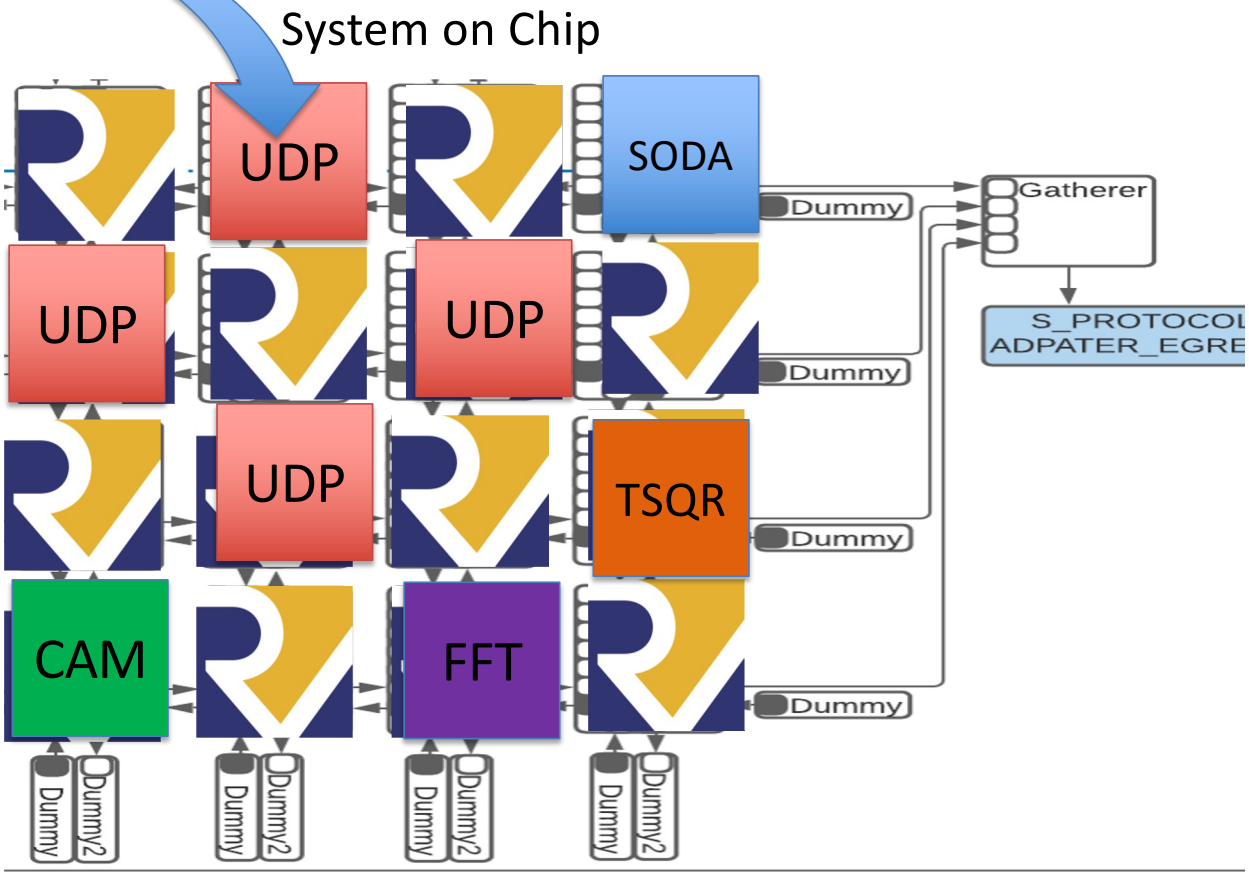


BERKELEY LAB

http://chiplets.lbl.gov/

MoSAIC: Modular System for Accelerator Integration and Communication
Cross-USG Heterogeneous Integration Fabric



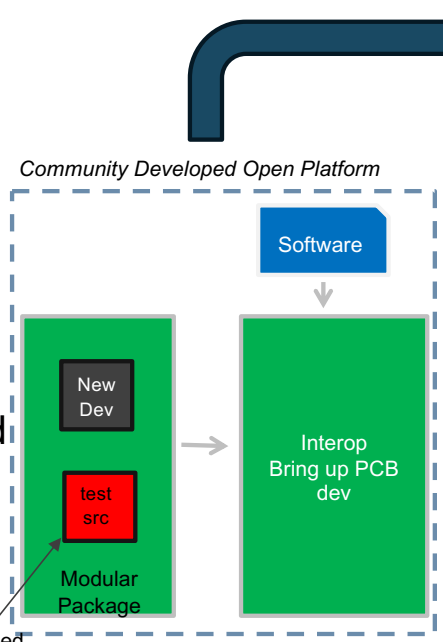


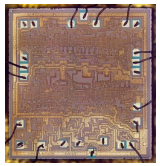
Lean and Mean
Operates at 250MHz (1/4 real-time)
Driverless inter-accelerator interaction
PGAS + **MsgQs for communication**
C++20 software stack

More Efficient Chiplet Development and Integration Path

Platform for open development with path into commercial platform

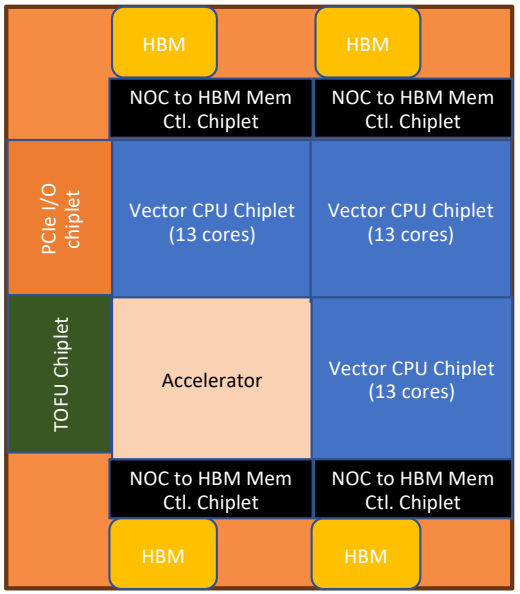
Common Open Chiplet Eval Board for Bring-up



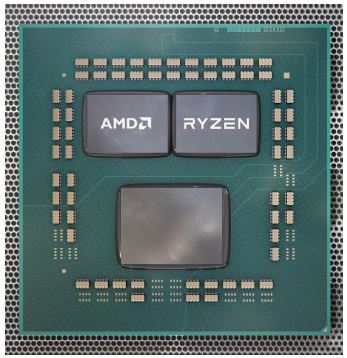


Reused Gold Dev

Chiplet



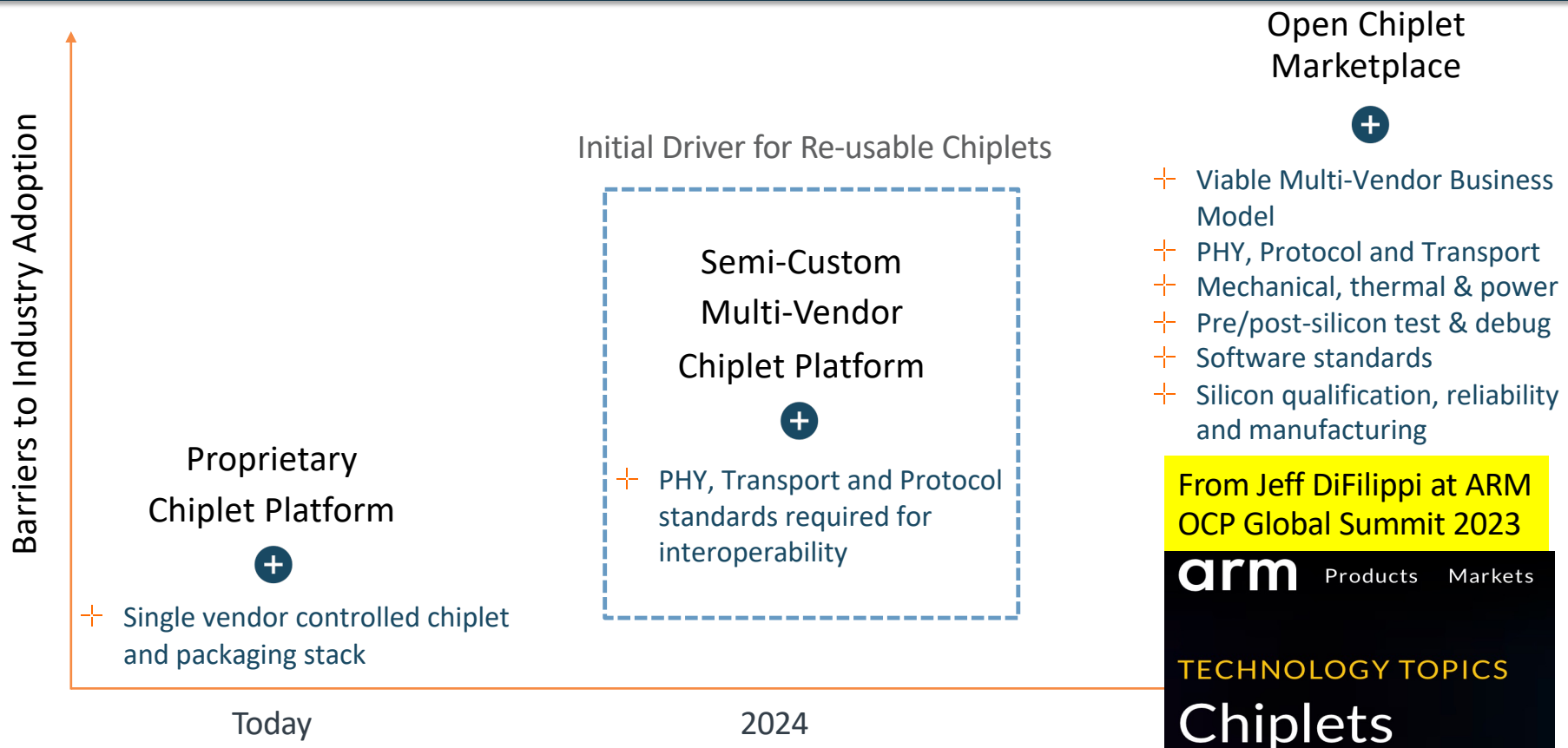
Prototype on Modular Architecture for function and performance integration



Commercial product integration

The Road to an Open Chiplet Marketplace

It's a big vision, but we are not there yet….



From Jeff DiFilippi at ARM
OCP Global Summit 2023

arm Products Markets

TECHNOLOGY TOPICS

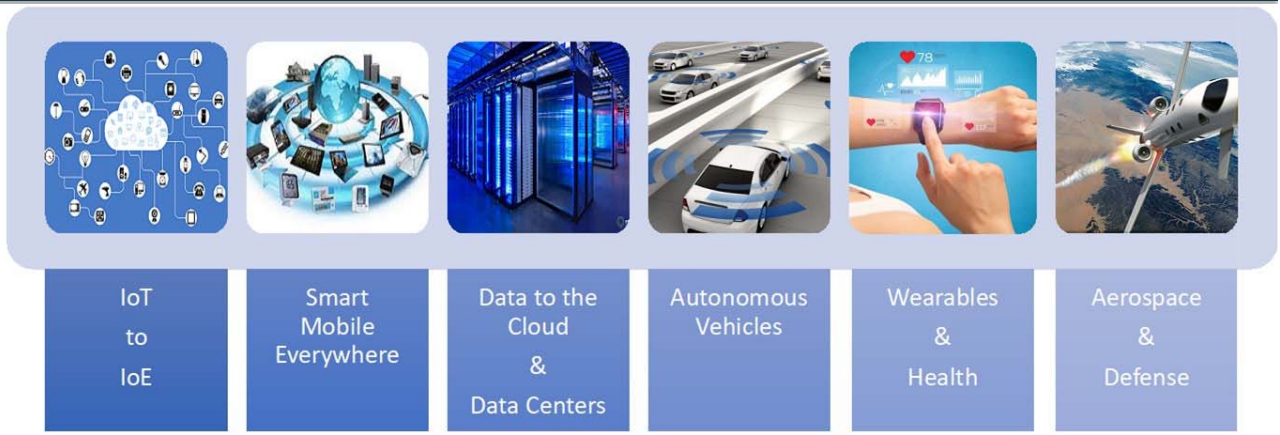
Chiplets

Industry: *Heterogeneous Integration Roadmap*

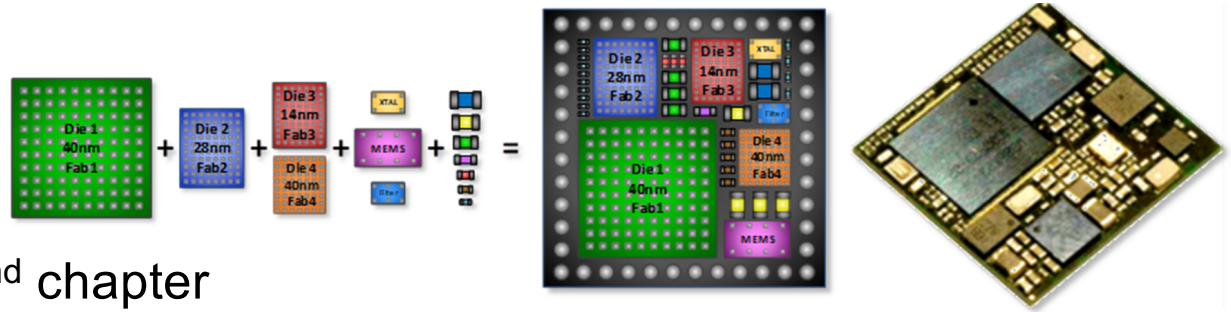
HETEROGENEOUS INTEGRATION ROADMAP

2019 Edition

http://eps.ieee.org/hir



All future applications will be further transformed through the power of AI, VR, and AR.



Die + Heterogeneous

System in Package (SiP)

HPC and Mega-datacenters is 2nd chapter

IEEE ELECTRONICS PACKAGING SOCIETY | ELECTRON DEVICES SOCIETY | IEEE Photonics Society | ASME SETTING THE STANDARD | semi

Chiplets at LBL

Home Register Demos Meeting Venue

http://chiplets.lbl.gov

LBNL/OCP Open Chiplet Economy Experience Center

Hosted by Lawrence Berkeley National Laboratory (LBNL)

Co-organized by the Open Compute Project (OCP)

Date: June 24, 2024

Time: 12:00pm to 5:00pm

Location: Berkeley National Lab, Wang Hall Bldg. 59, Room 59-3101

OPEN Compute Project®

2024 JEDEC and OCP Standards for Chiplet Design with 3DIC Packaging Workshop

JEDEC

OCP Modularity for HPC & AI Workstream needs your input!

Patricia G.
(LBNL)

George M.
(LBNL)

Anu R.
(Microchip Inc.)

luisa.gonzalez@ocproject.net
georgios.michelogiannakis@ocproject.net
tony.gutierrez@ocproject.net
anu@microchip.com

How to participate:

- Subscribe to the Project Mailing List
- Add the Project Meeting Calendar
- Join the Conversation

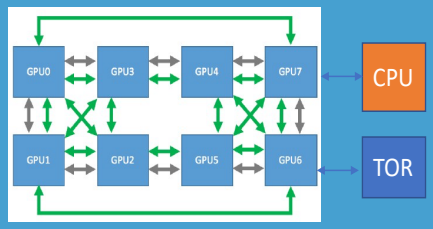
Dynamic Reconfiguration (aka resource Disaggregation)

Ultra-Performance co-packaged optics/networking for resource disaggregation

Diverse Node Configurations for Diverse Workload Resource Requirements

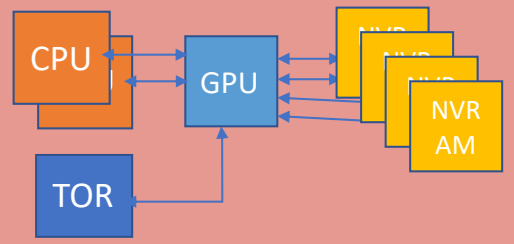
Training

- 8 connections: GPU
- 8 links to HBM (weights)
- 8 links: to NVRAM
- 1 links: to CPU (control)



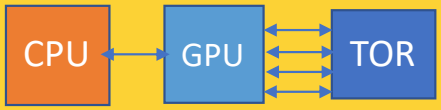
Data Mining

- 6-links: HBM
- 15 links: NVRAM (capacity)
- 4 links: CPU (branchy code)



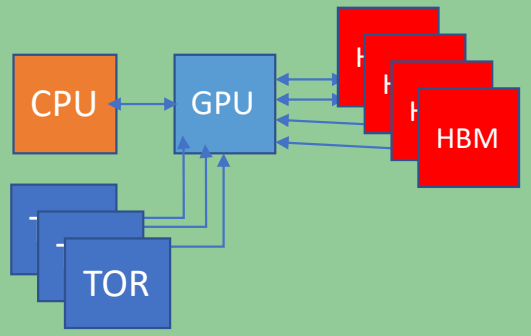
Inference

- 16 links to TOR (streaming data)
- 8 links HBM (weights)
- 1 link: CPU

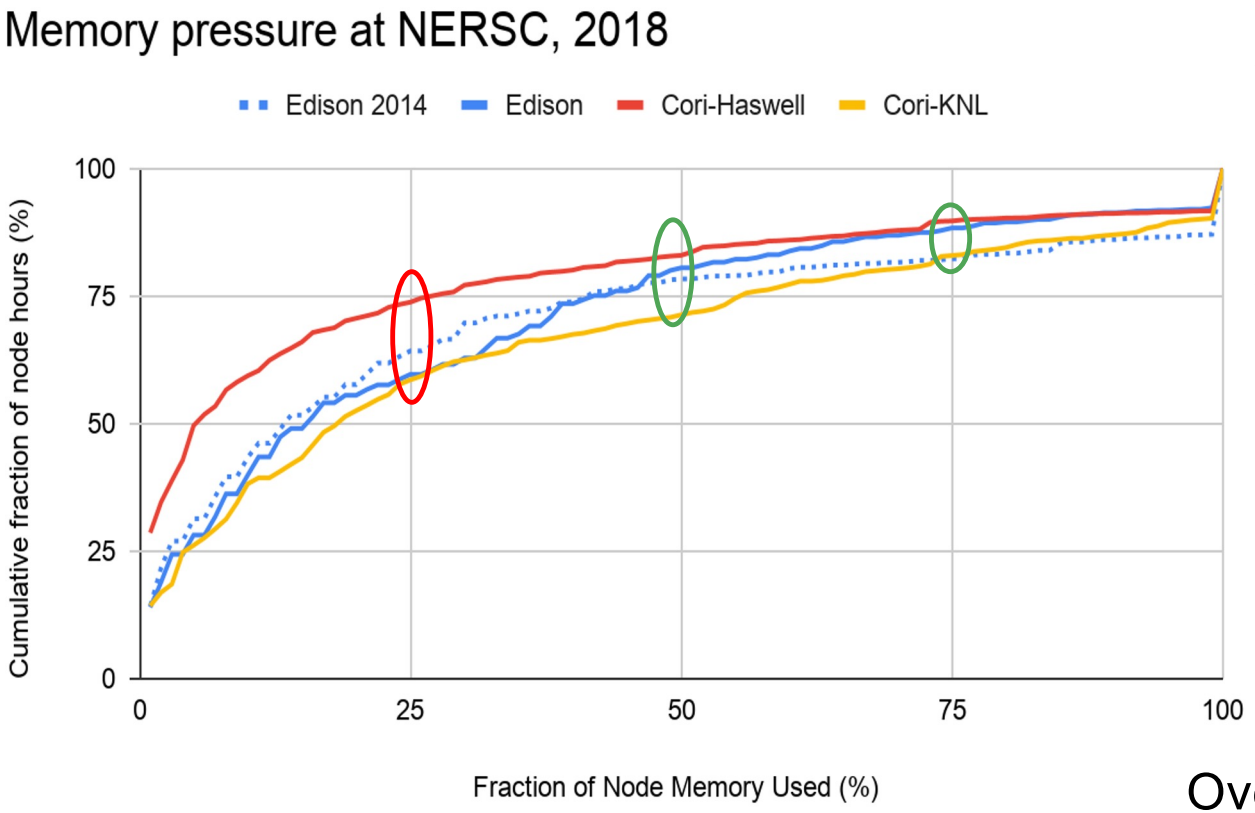


Graph Analytics

- 16 links HBM
- 8 links TOR
- 1 Link CPU



Need for Memory Disaggregation



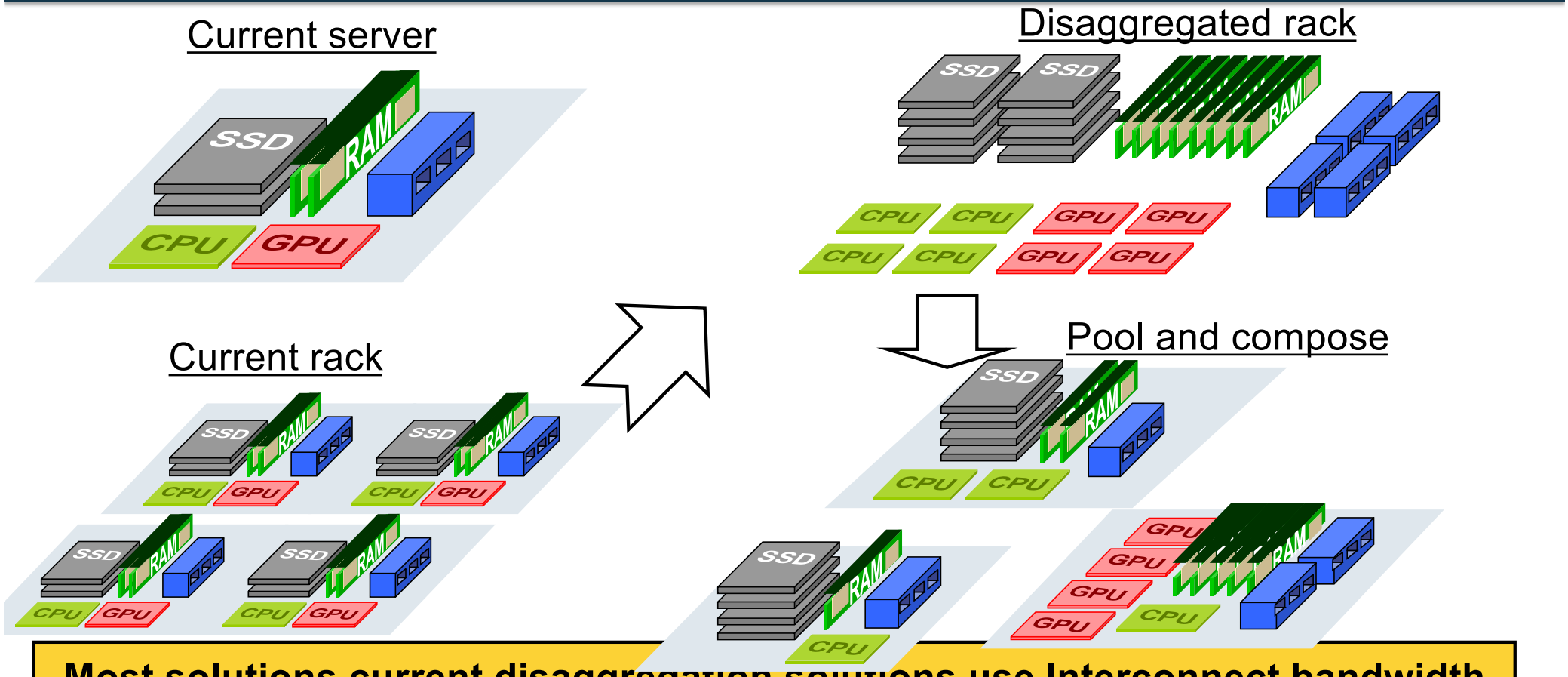
About 15% of NERSC workload uses more than 75% of the available memory per node.

And ~25% uses more than 50% of available memory.

But 75% of Haswell job hours (60% of KNL) use < 25% memory

Overestimate: maxrss x ranks_per_node
Assumes memory balance across

Disaggregated Node/Rack Architecture



Most solutions current disaggregation solutions use Interconnect bandwidth (1 – 10 GB/s)

An AI Chiplets Example in 6 Easy Steps

Memory Disaggregation within the Package

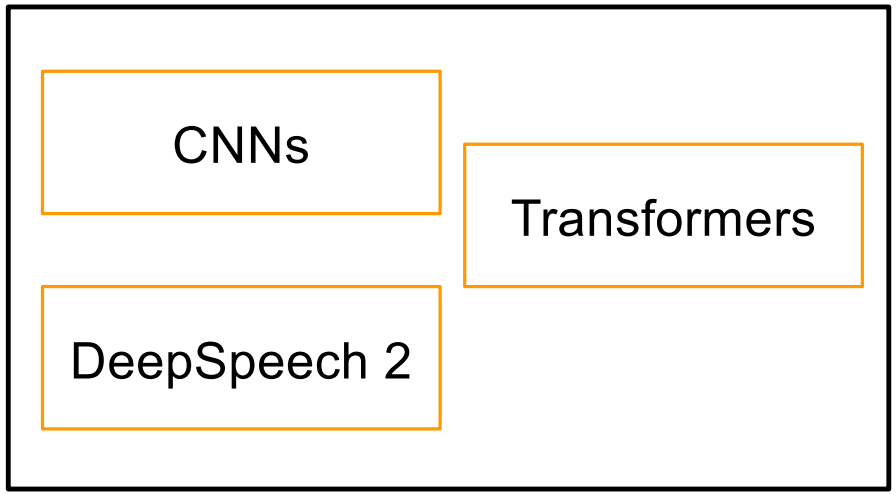
Luisa Patricia Gonzalez Guererro (LBNL)

Recipe for Hardware Specialization enabled by chiplets modularity in 6 steps

Ingredients:

- *4 Chiplets*
- *1 Interposer*

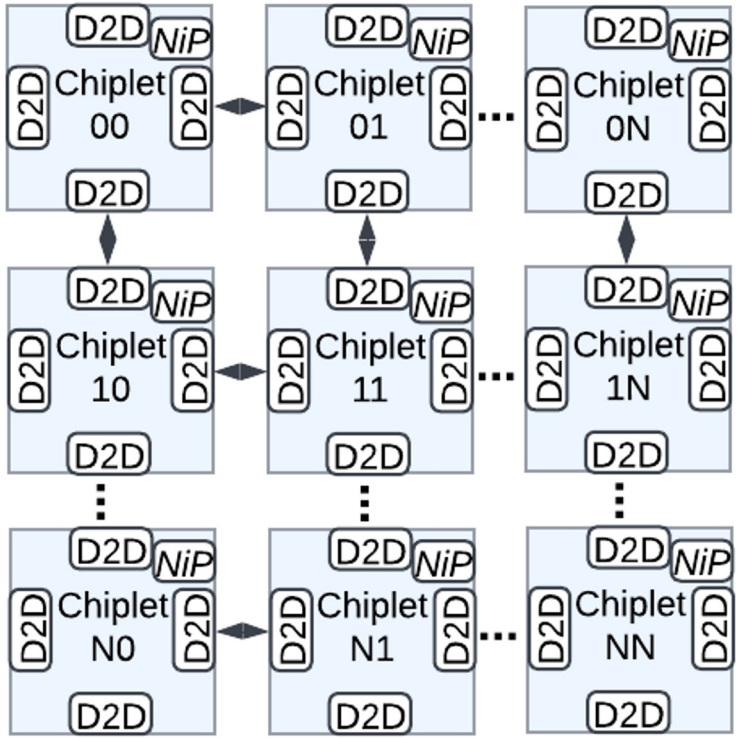
Chiplets Marketplace ;)



Three optimal Systems in Package (SiP) for AI

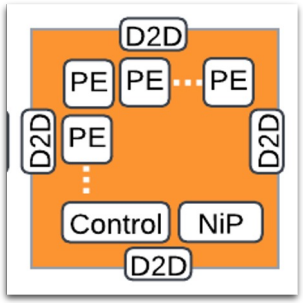
Step 2: Scalable Interposer 2D grid with NxN Chiplets

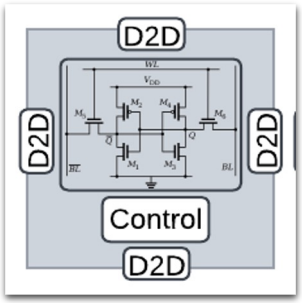
- **Network-In-Package (NiP)**
  - Packet switched
  - Light weight
  - Minimum transmission latency (1CC)
- **No Cache Coherency**
  - Message passing for parallel computing
    - Hardware Message Queues

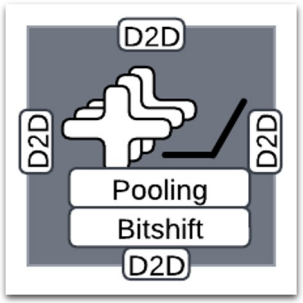


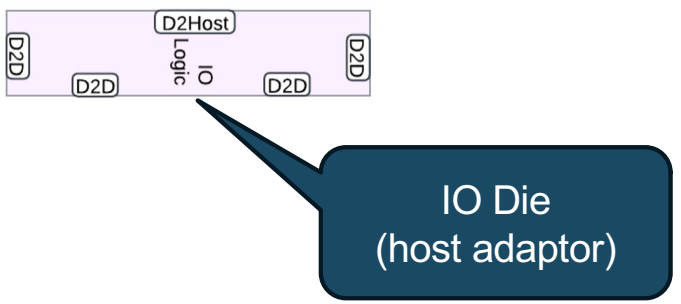
Dataflow is regular and predictable

Step 3: Let's do some math and define our chiplets parameters





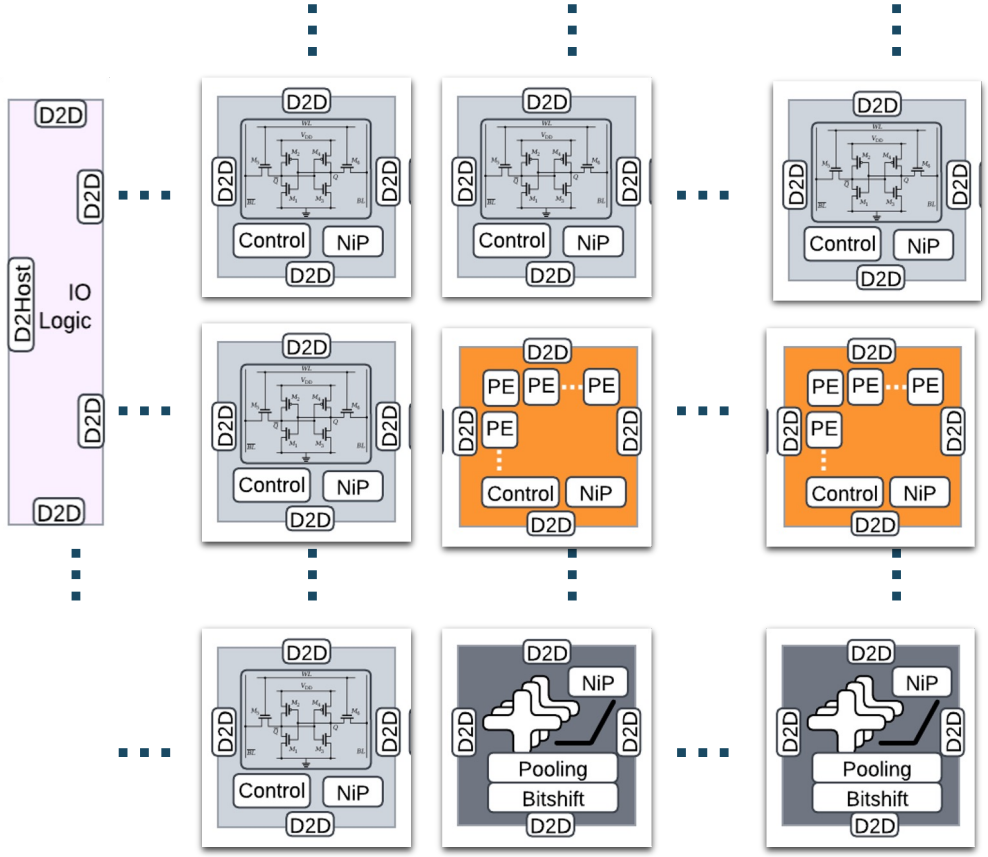




	Chiplet 4 mm x 4mm	Sub 20nm @ 1GHz	Sub 10nm @ 1GHz Sub 5nm @ 2GHz
1	**Systolic Array**	64 x 64 PEs	128 x 128 PEs
2	**Scratchpad**	1MB	9MB
3	**Accumulator**	64 acc	128 acc

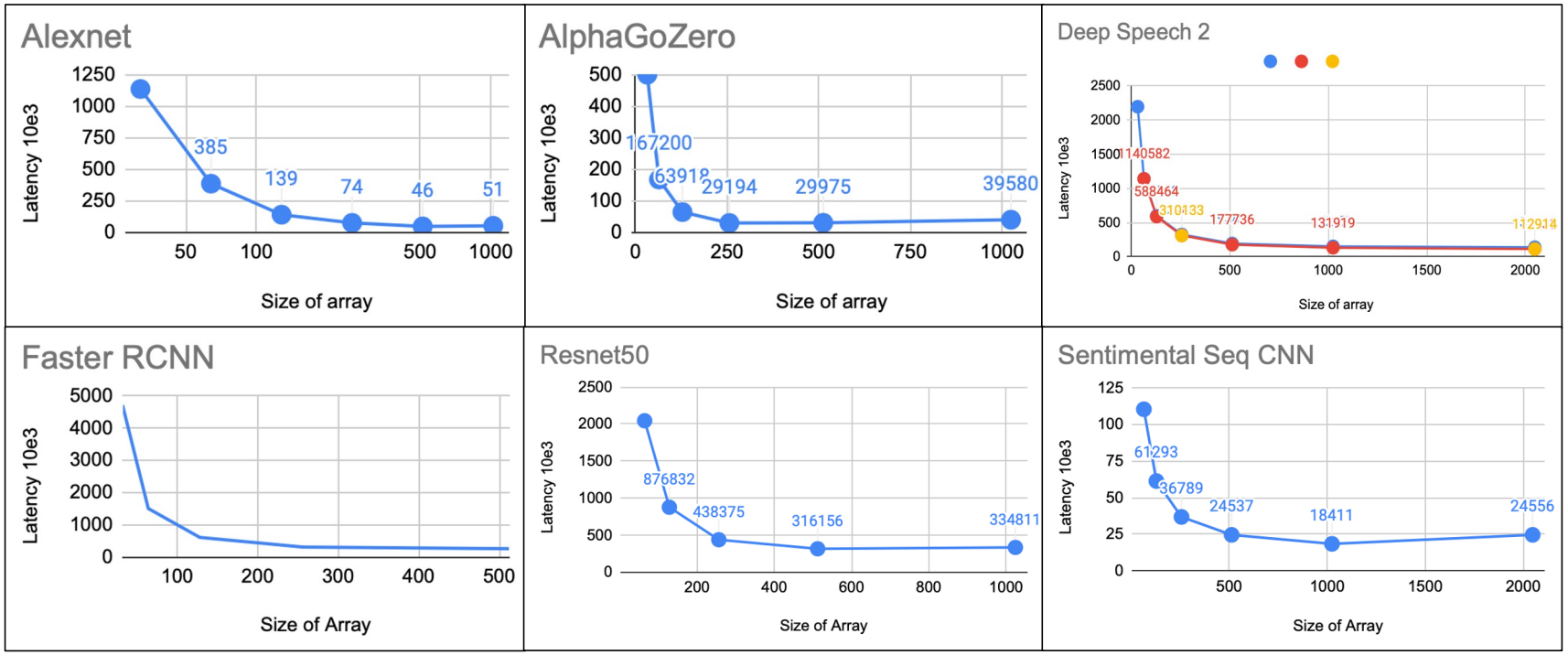
Step 4: Architecture for an AI SiP

- **Scratchpad chiplets**
  - Left: Input/Output Buffer
  - Top: Trained parameters
- **KxK systolic arrays chiplets**
- **Accumulator chiplets : last row**
- **IO Chiplet: From and to host**



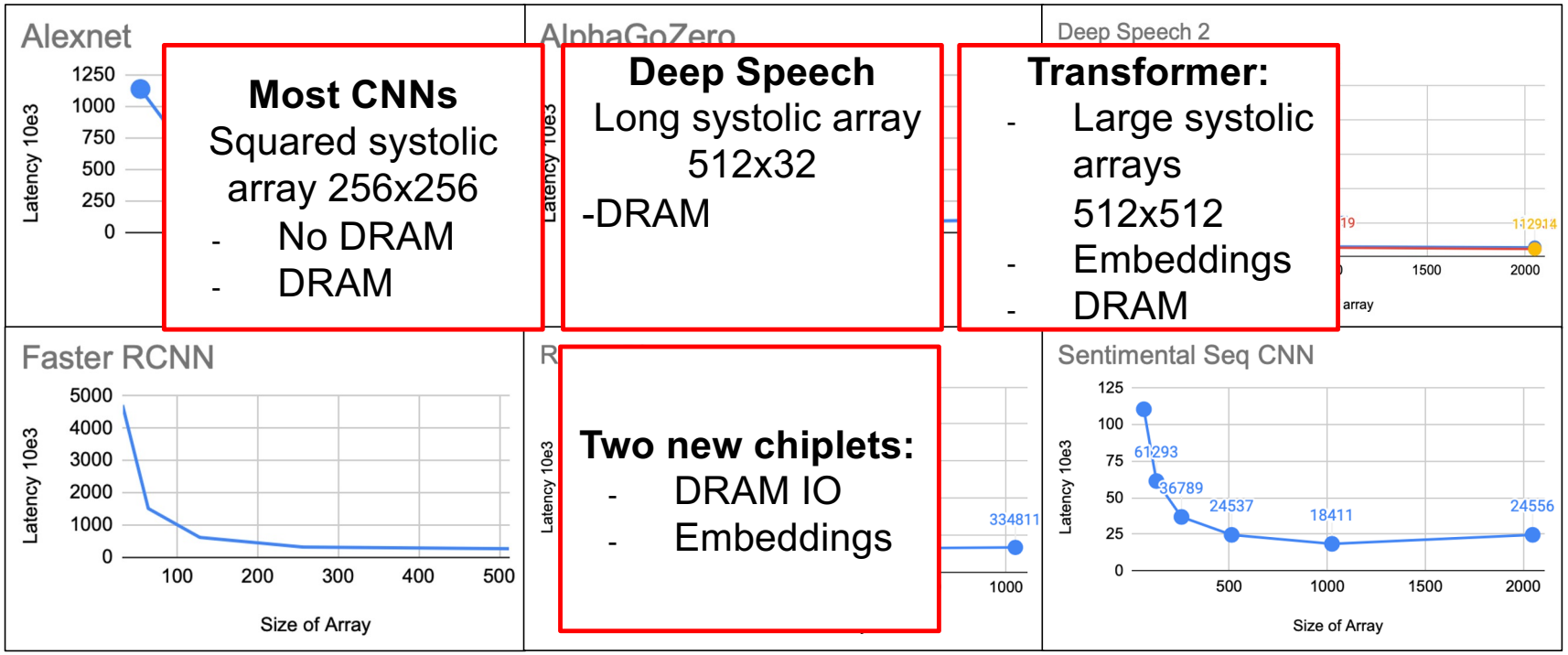
Step 5: Profiling AI algorithms

- **Optimal Size of systolic array**
- **Optimal Scratchpad capacity (MB)**



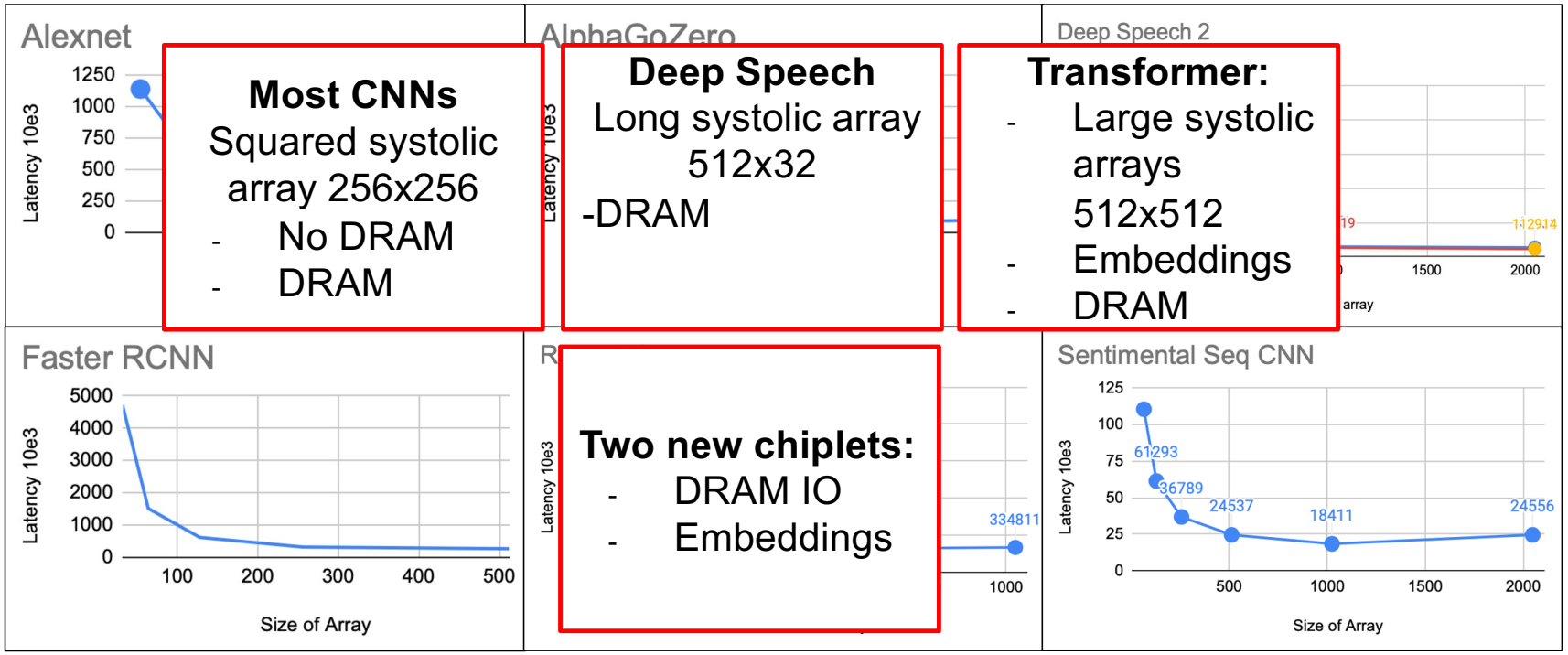
Step 5: Profiling AI algorithms

- **Optimal Size of systolic Array**
- **Optimal Scratchpad capacity (MB)**



Step 5: Profiling AI algorithms

- **Optimal Size of systolic Array**
- **Optimal Scratchpad capacity (MB)**

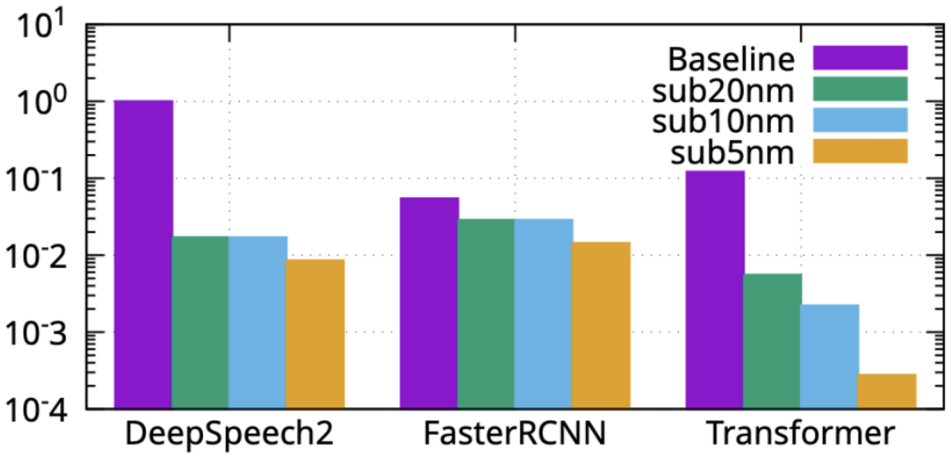


Step 6: Three different SiPs for three dataflows

	Sub 20nm SiP			Sub 10nm/5nm SiP		
	CNN	Deep Speech	Trans.	CNN	Deep Speech	Trans.
Syst. Array PEs	256x256	512x64	512x512	256x256	512x128	512x512
Syst. Array Chiplets Num.	16	8	64	4	8	16
Scratchpad Chiplets Num. (3D)	16	27	20	10	27	12
Accumulator Chiplets Num.	4	1	8	2	1	4
Embeddings Chiplets Num.	0	0	8	0	0	4
BW [TB/s] per Chiplet	1	1	1	2/4	2/4	2/4
Data lanes per Chiplet	8192	8192	8192	16384	8192	16384
Chiplet Grid	6x6	6x6	10x10	4x4	6x6	6x6
Interposer Size* [mm x mm]	38x38	38x38	64x64	26x26	38x38	38x38

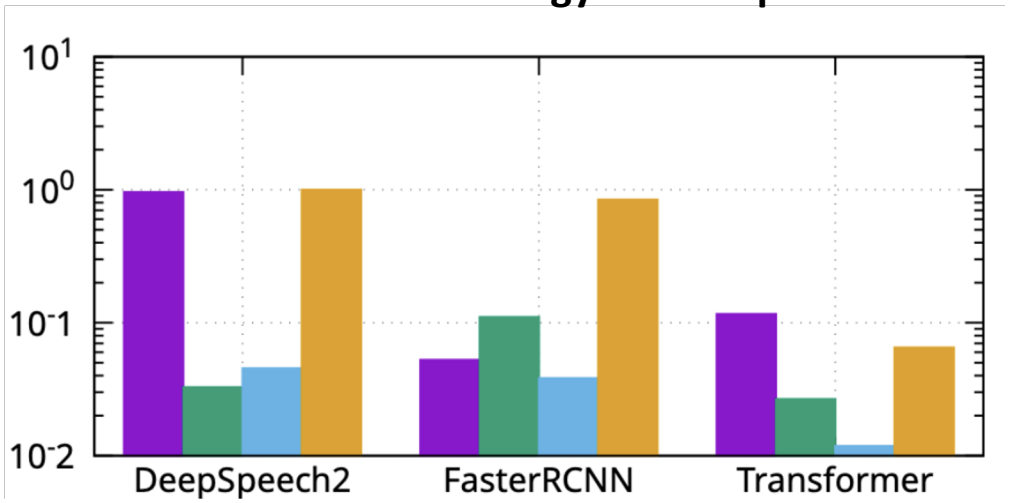
Evaluation: Better Latency and Better Energy consumption

Normalized Latency



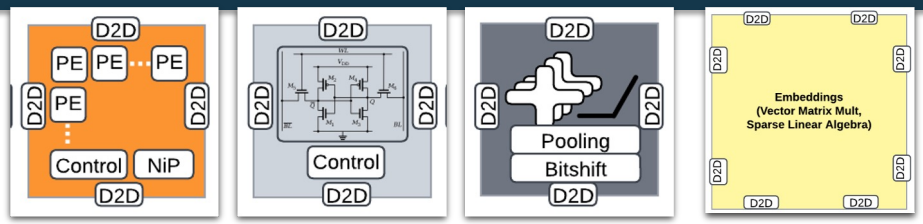
~100x lower latency than H100 baseline

Normalized Energy Consumption



~10x lower energy consumption than H100 baseline

Overspecialization: Fear Not!



4 Chiplets + 1 interposer = 4 highly specialized SiPs for AI

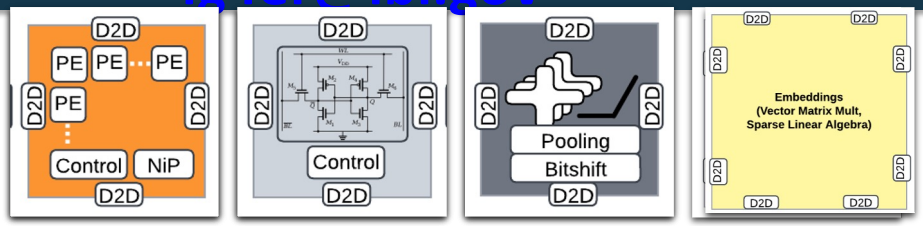
+100x less latency than baseline
+10x better energy consumption than baseline

Final thoughts and next steps

- Chiplet ≠ Application
- Chiplet = Kernel, *i.e. (GEMM)*
- **HPC applications:**
  - Density Functional Theory
  - Molecular Dynamics
  - Climate modeling
  - Processing In Cell
  - NEGFs
  - SpVM

Thank you!

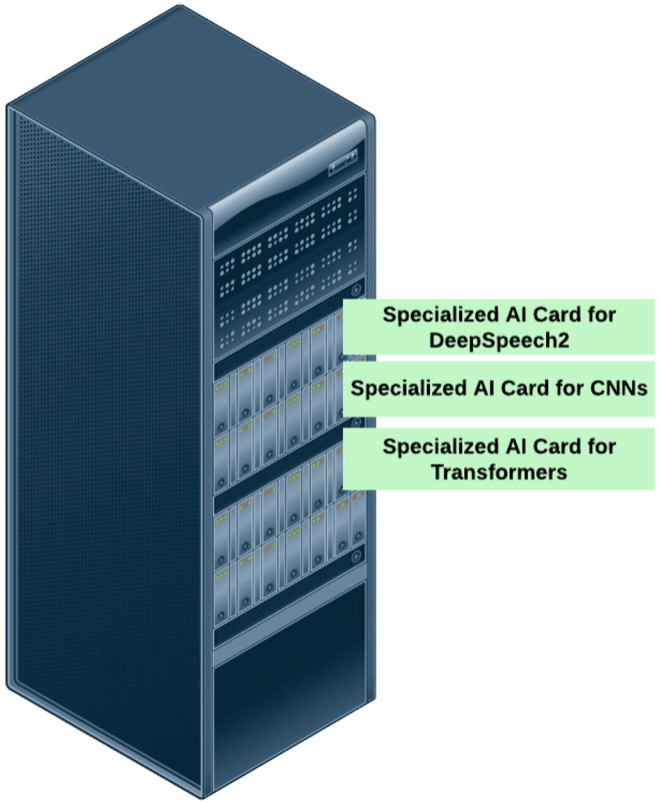
lg4er@lbl.gov



4 Chiplets + 1 interposer = 4 highly specialized SiPs for AI

100x less latency than baseline
10x better energy consumption than baseline

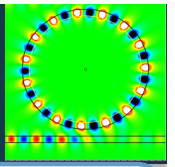
And a lot of that was about memory specialization and memory fabrics

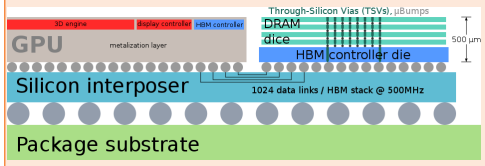


CoPackaged Optics and Photonic Disaggregation

Lets take this to the net level!

Impedance Matching to Packaging Technology

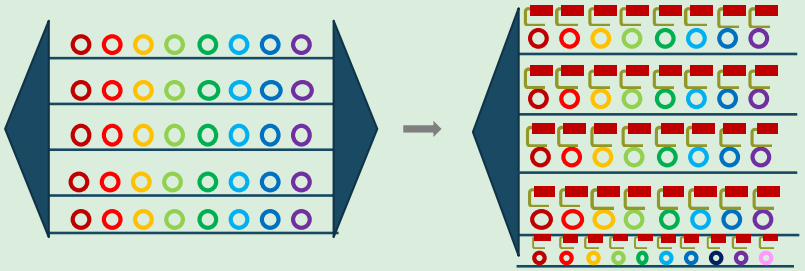


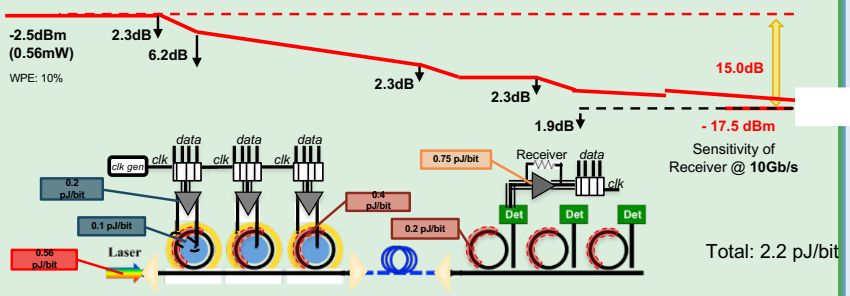


In-package integration

Solder Microbumps
& Copper Pillars@~10Gbps

Wide and Slow!

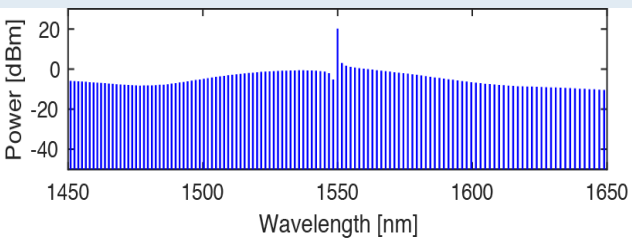


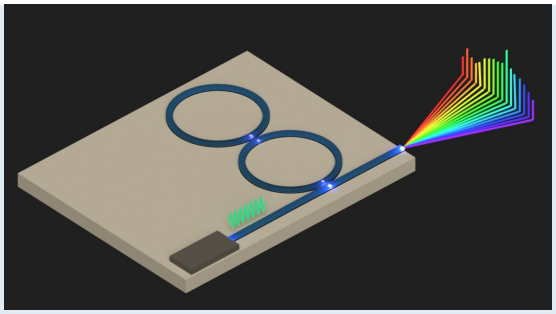


DWDM Using Silicon Photonics

Ring Resonators @ ~10-25 Gb/sec per chan
Many channels to get bandwidth density

Wide and Slow!



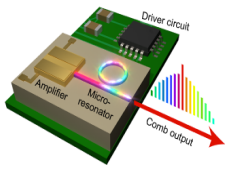


Comb Laser Sources

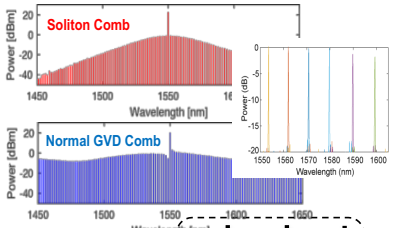
Single laser to efficiently
generate 100s of frequencies

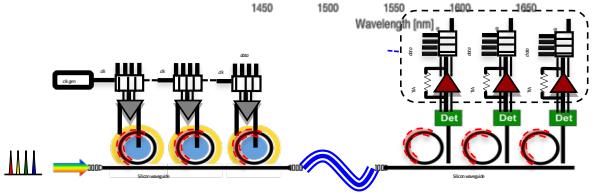
Wide and Slow!

Photonic MCM (Multi-Chip Module)

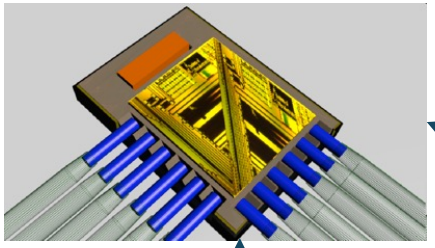


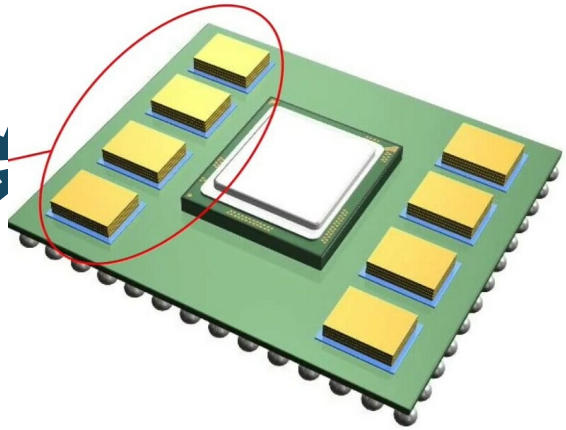
Scales to 100s of λs



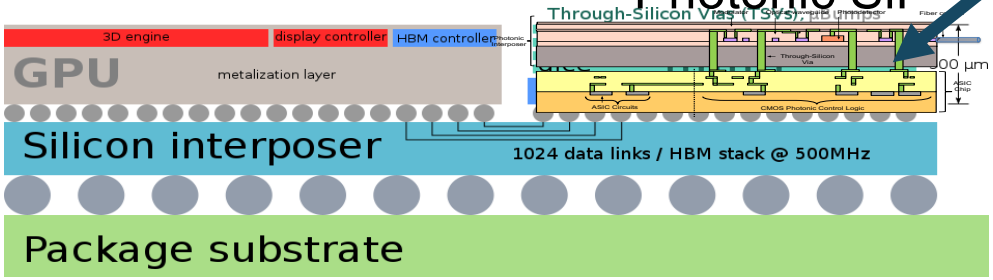


Comb Laser Source with
DWDM Silicon Photonics
Wide-and Slow for high speed links



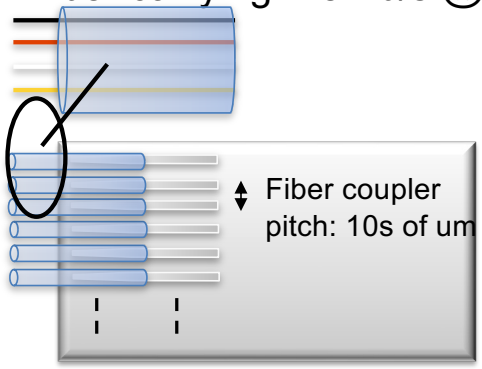


Photonic SiP

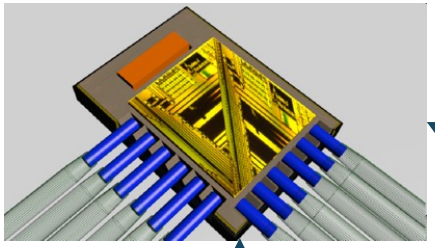


Photonic MCM (Multi-Chip Module)

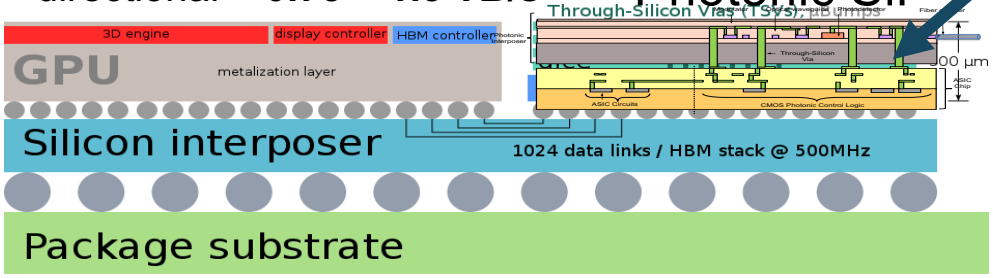
Fiber carrying 1-8 Tb/s @ 0.1pj/bit

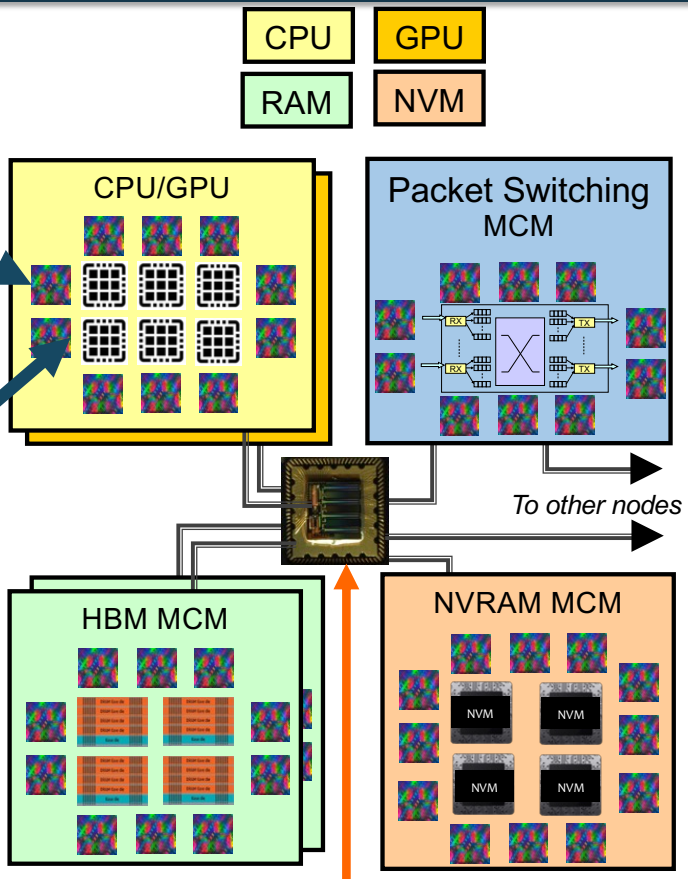


High-Density **fiber coupling array** with 24 fibers = 6-12 Tb/s bi-directional = **0.75 – 1.5 TB/s**



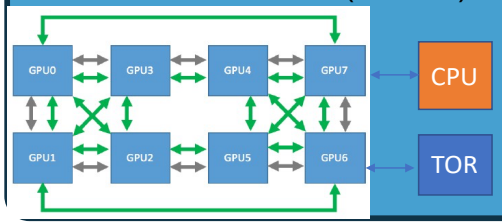
Photonic SiP





Training

- 8 connections: Peer GPU
- 8 links to HBM (weights)
- 8 links: to NVRAM
- 1 links: to CPU (control)

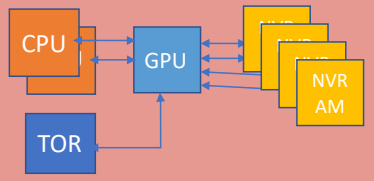


Inference

- 16 links to TOR (streaming data)
- 8 links HBM (weights)
- 1 link: CPU

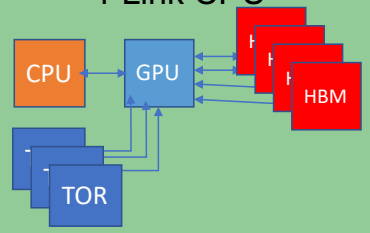
Data Mining

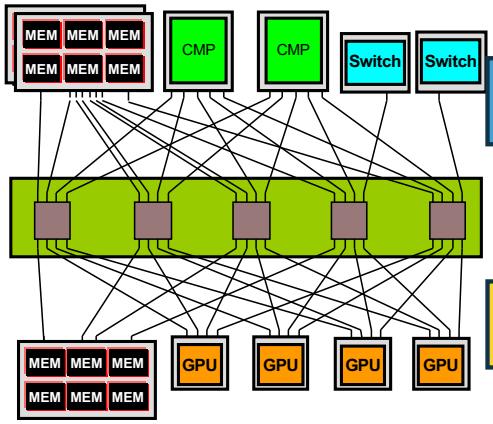
- 6-links: HBM
- 15 links: NVRAM (capacity)
- 4 links: CPU (branchy code)



Graph Analytics

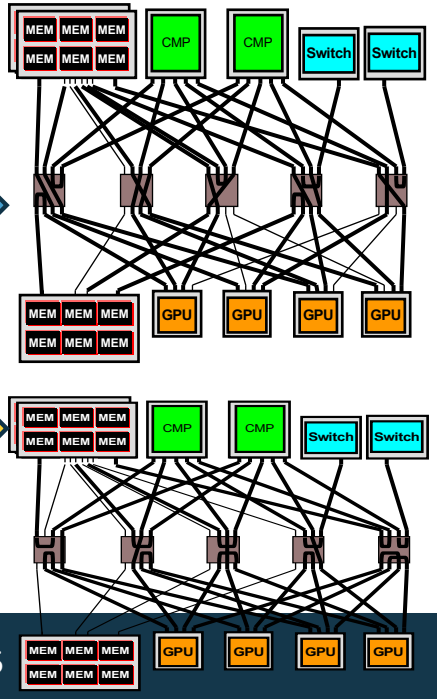
- 16 links HBM
- 8 links TOR
- 1 Link CPU

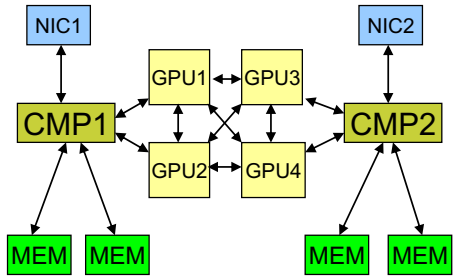


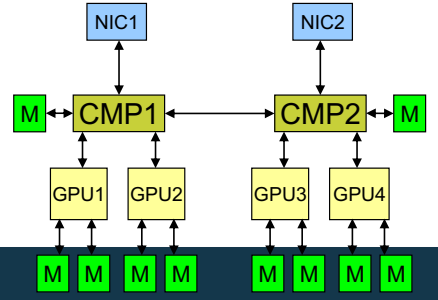


Configure for Training

Configure for Inference







Latency Sensitivity Study: Focus on Single-Hop Networks

Need to minimize network diameter due to latency sensitivity

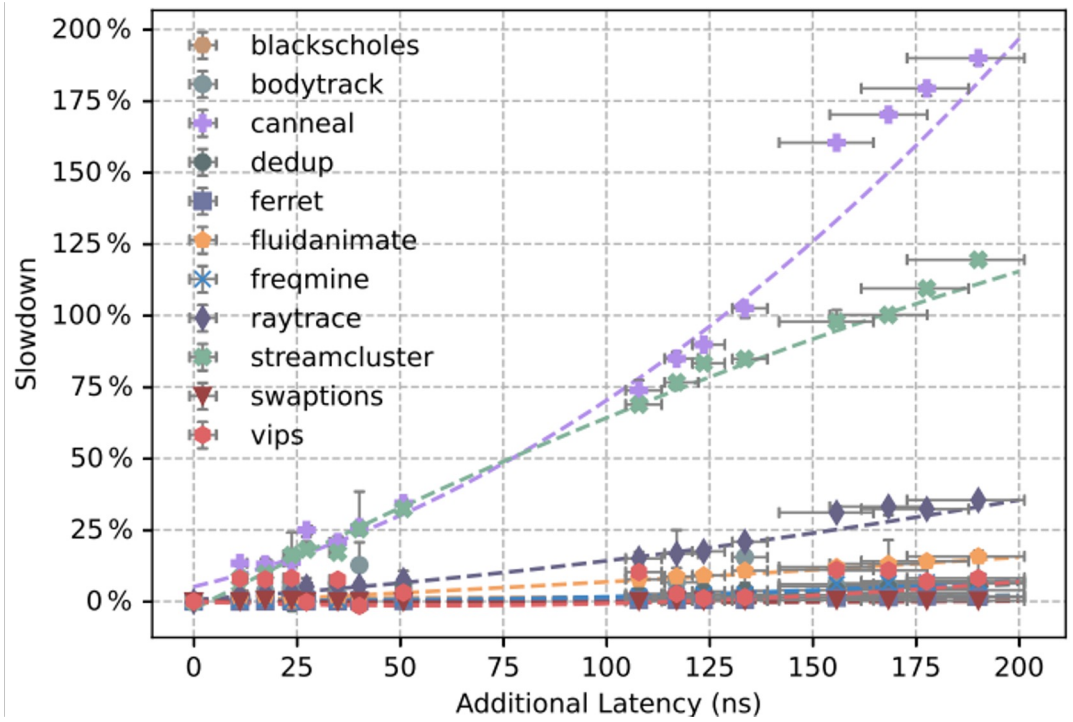
PARSEC 3 on real hardware:

- **Sapphire Rapids HBM nodes**
- **Dual Intel Xeon Max 9462 CPUs**
- **Flat mode**
- **SNC4 clustering**
- **The execution core is constant**
- **We vary memory NUMA domain**

Workload sensitivity to latency:

- **Most show minimal effect**
- **Few (canneal, streamcluster) suffer significant performance degradation with latency increase**

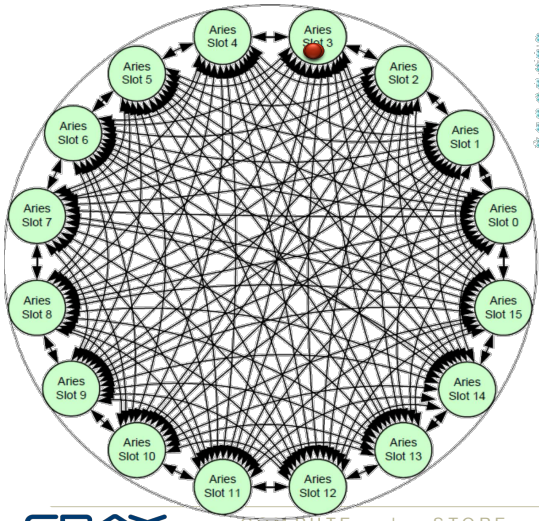
Dots are measurements.
Dashed lines are curve fitting degree 2 polynomials.



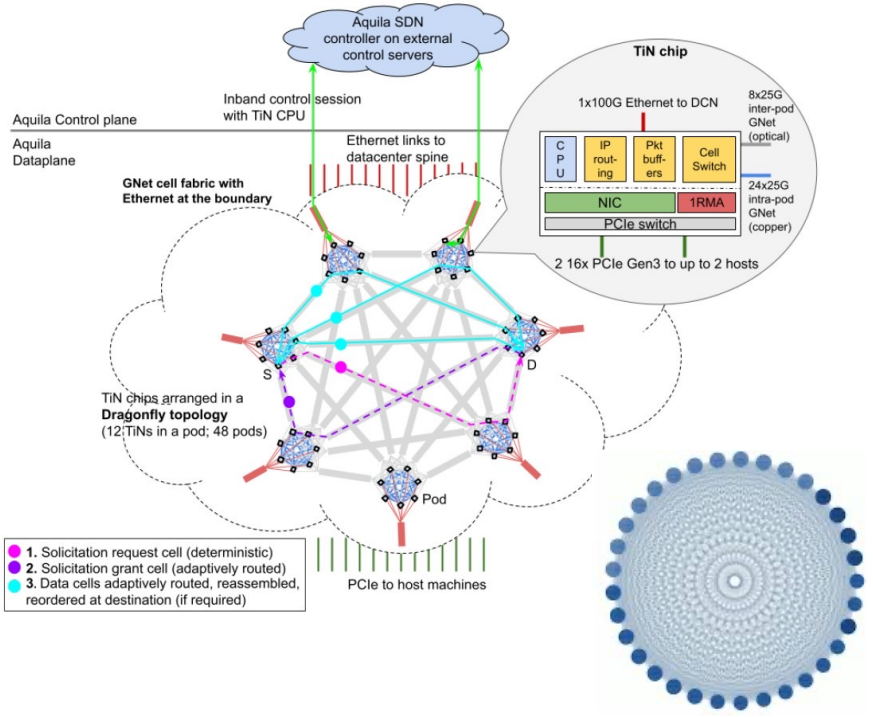
Low-Diameter/Single-Hop Interconnects (Dragonfly or All2All)

Started in HPC, but seeing traction in hyperscale and OCP

Cray Dragonfly – Aries/YARC2 2016

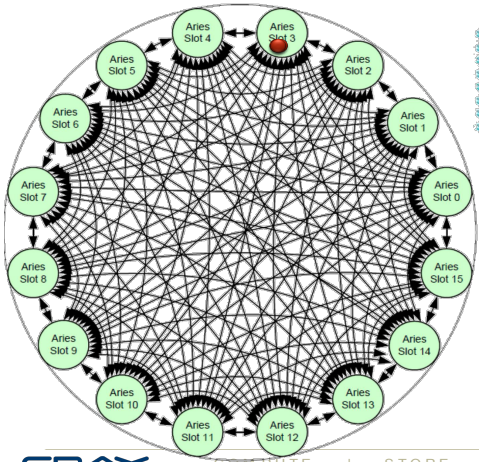


Google Aquila 2022

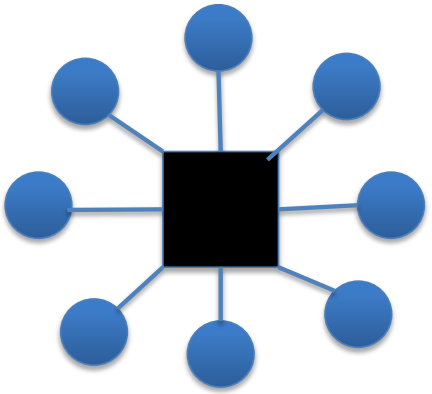


Key Value Proposition for AWGR Lambda Shuffle

- **Electrical All-2-All for single-hop networks**
  - $O(N^2)$ cables
  - O(N) connectors/endpoint

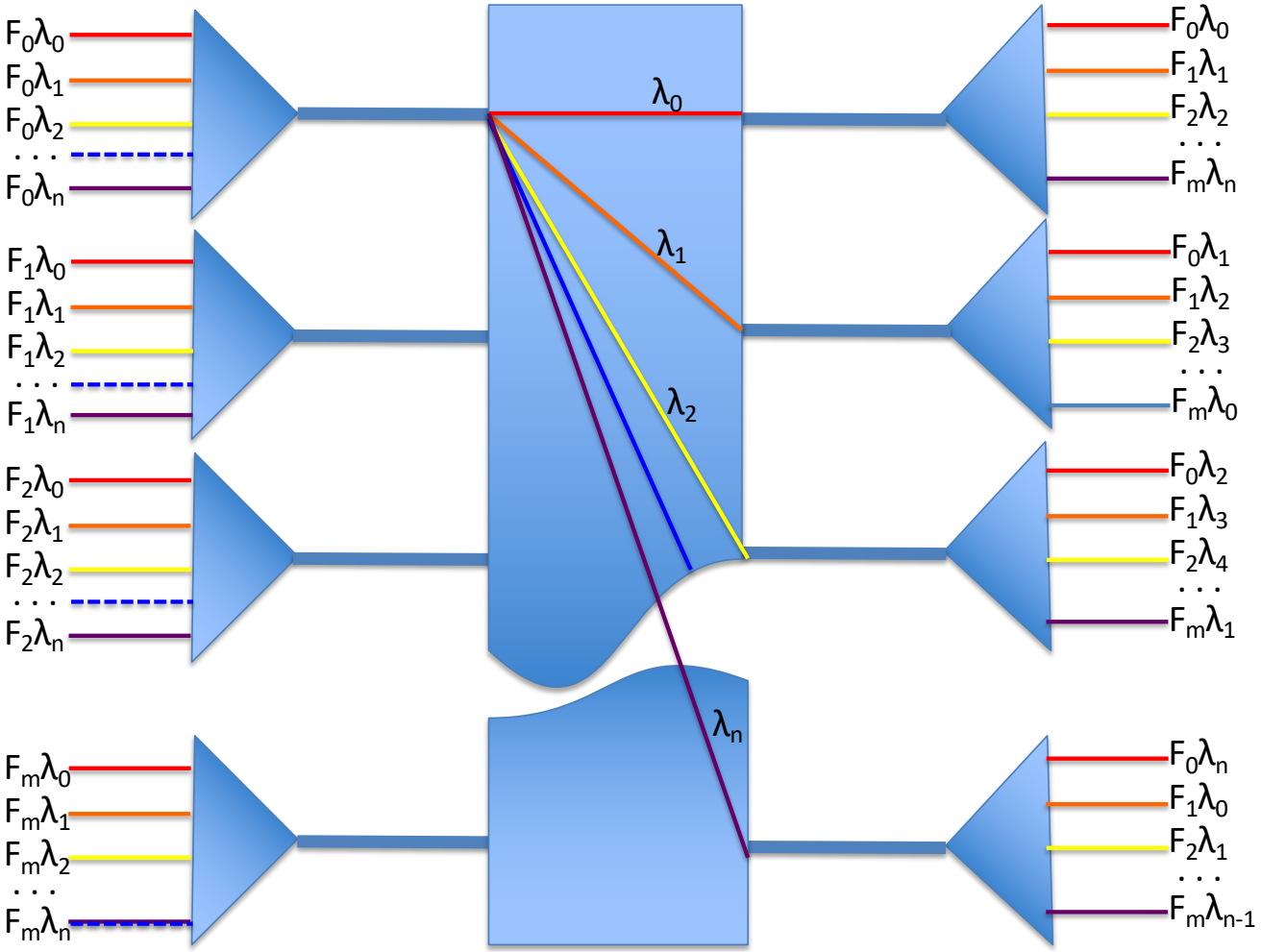


- **Photonics lambda shuffle (unique to photonics)**
  - O(N) cables
  - O(k) connectors/endpoint



Optical connectors are expensive and take space

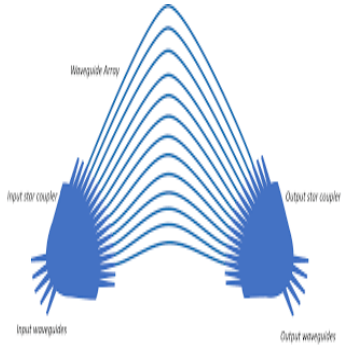
Photonic Lambda Shuffle

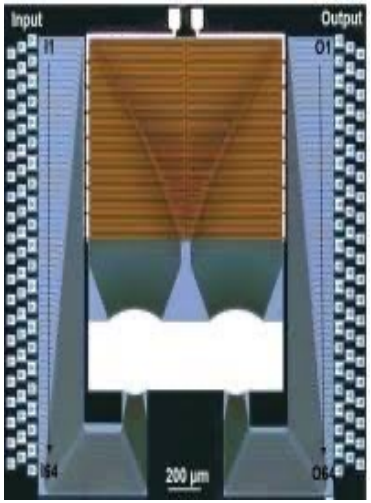


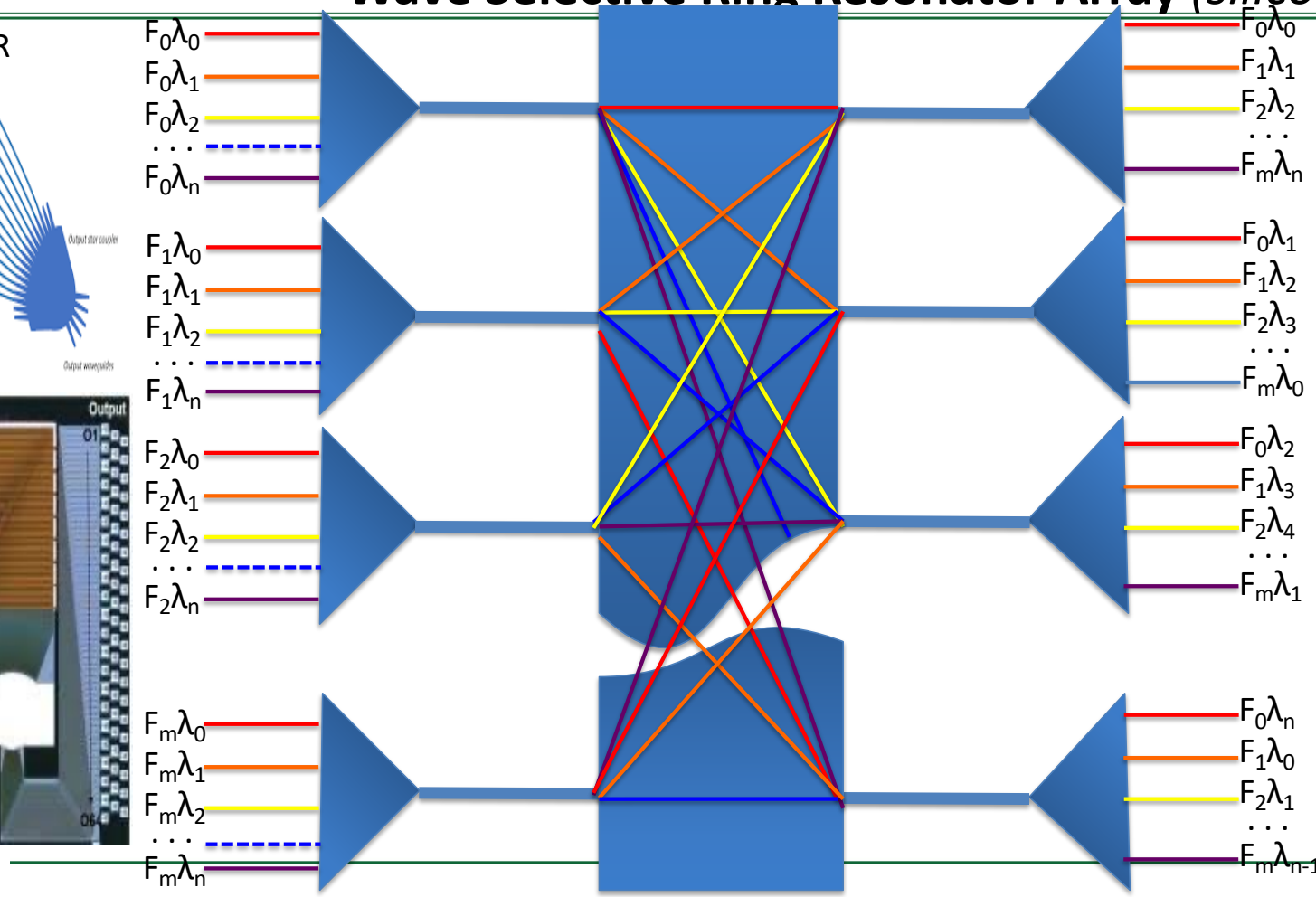
Optica/OSA Industry Incubator 2012

Monolithic AWGR *(stamped in plastic)* or **Wave Selective Ring Resonator Array** *(silicon lithography)*

AWGR

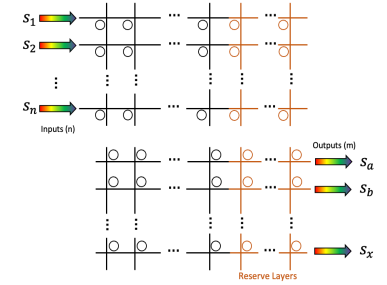


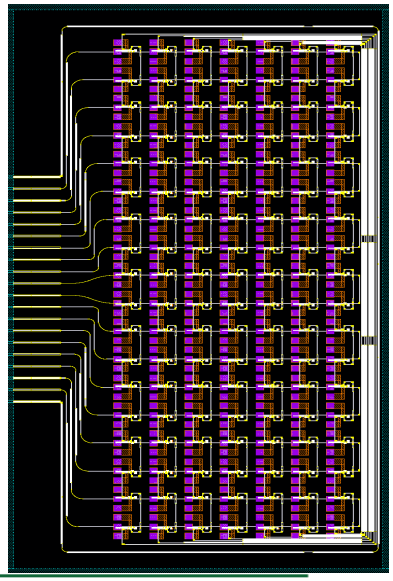




Wave Selective

Spatial/Wavelength-Selective Switch





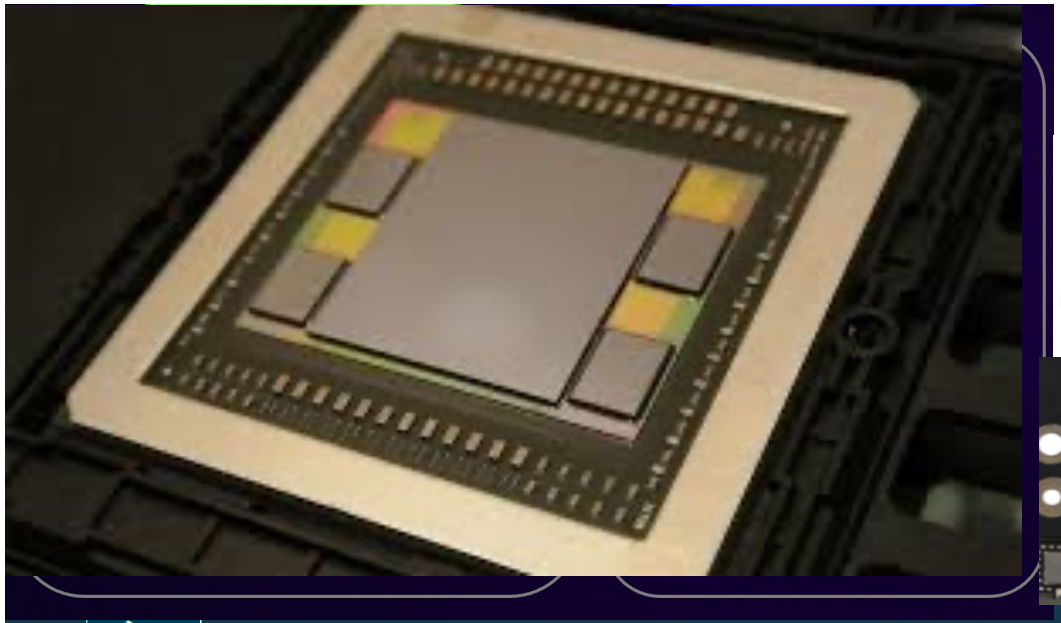
HETEROGENEOUS INTEGRATION ROADMAP

Optica/OSA Industry Incubator 2012

Some other stuff necessary to complete the picture

HBM4: an HMC-like Serial Interface to HBM

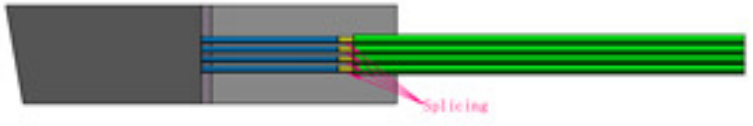
Example Samsung Custom HBM4



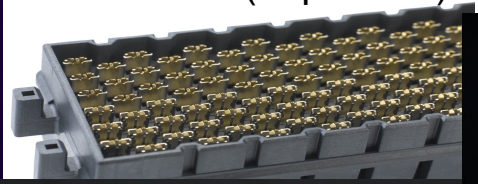
Fiber Attach

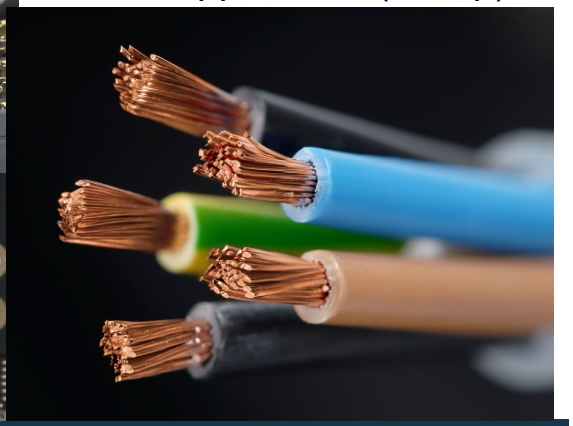
Opportunity with expensive electric connectors

Connector (Expensive) Fiber (cheap)



Connector (Expensive) Copper Wire (cheap)



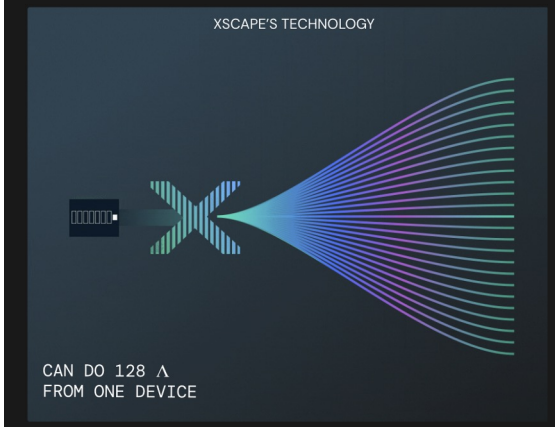


This slide is obsolete as of NVIDIA Hopper

Example

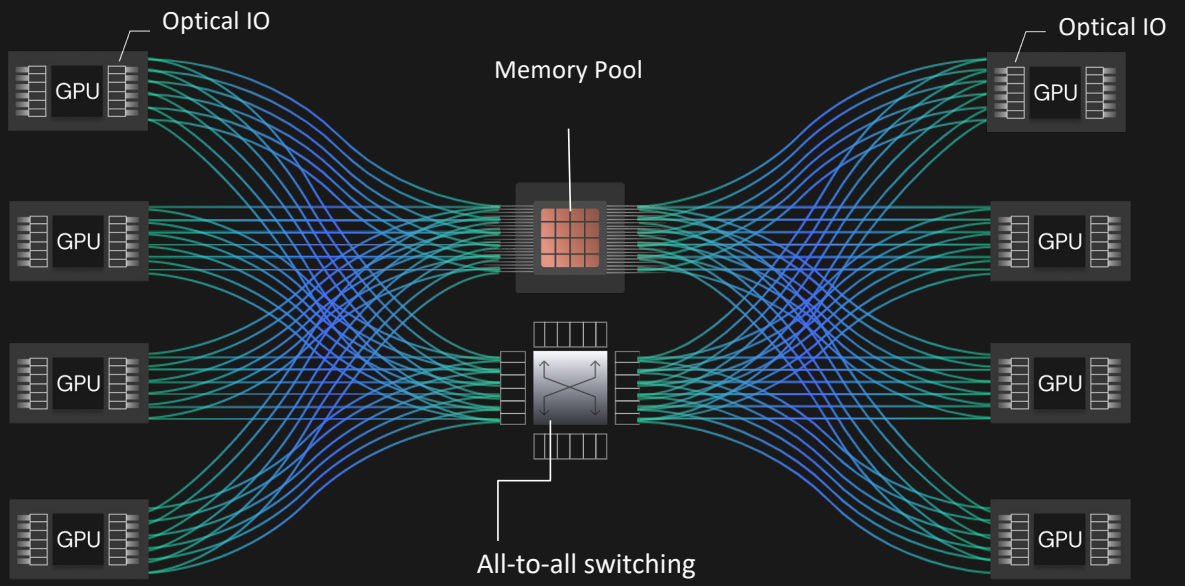
Systems solutions are ***always*** more powerful than any individual component
Comb laser level innovation unlocks value only at Fabric System Level
All solutions depend upon co-packaged optics at the endpoints

Comb driven Optical fabric solution



Xscape 's CombX Laser

Unlocks new fabric system implementations for the customer where the end value is captured



Optical Fabric System

System Throughput per Dollar Spent on Memory

Photonic disaggregation vs. conventional fixed node

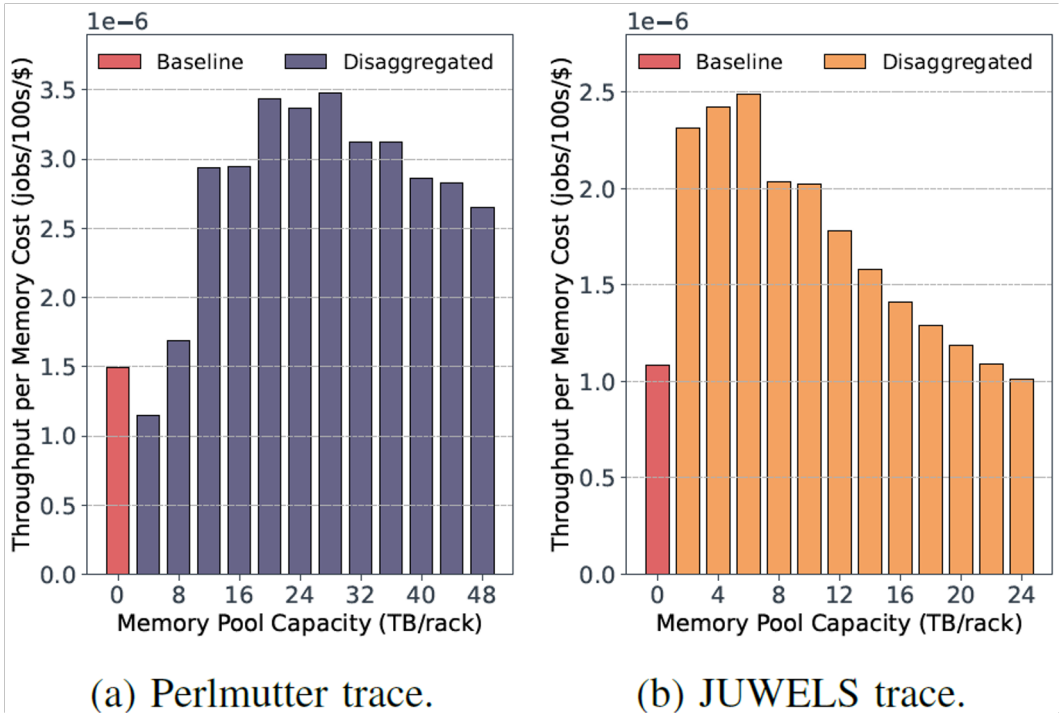
Rack-scale disaggregation with LBNL/Julich Fair Memory scheduling Policy

Assume $4.9 per GB of memory capacity

Baseline: No disaggregation. 512 GB/node

Disaggregation offers 3.5x improvement in system throughput per $

Beyond a certain memory capacity, negligible benefit in throughput but higher CapEx for memory



(a) Perlmutter trace. (b) JUWELS trace.

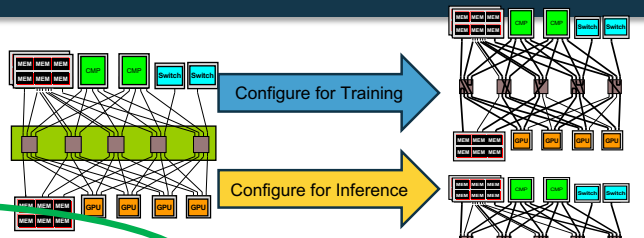
Anatomy of a “Value” Metric

$$\frac{\text{Good Stuff}}{\text{Bad Stuff}}$$

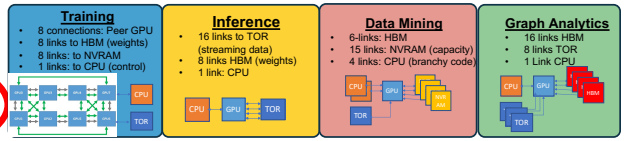
Anatomy of a "Value" Metric

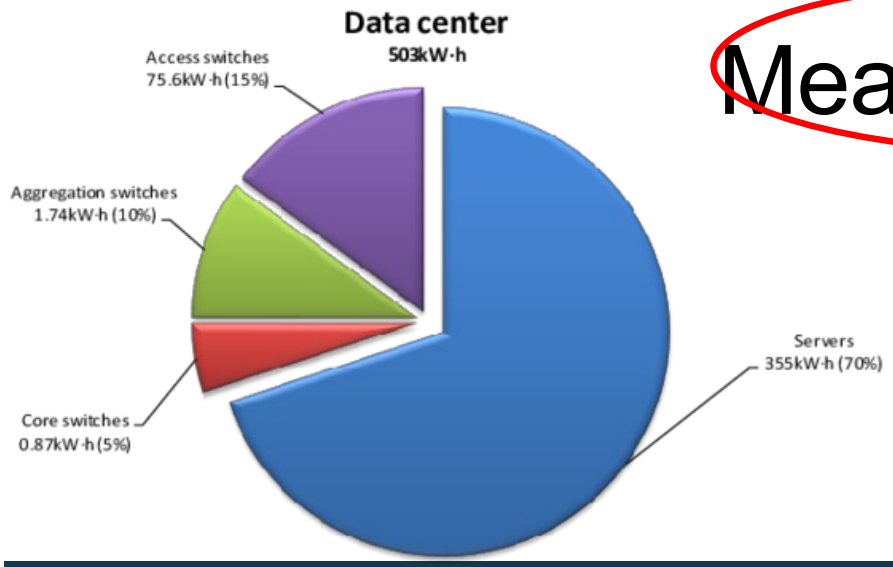
Increase performance with Disaggregation
And Bandwidth Steering!

Deliver bandwidth to where it is needed
By taking it from where it isn't



$$\frac{\text{Performance}}{\text{Measured Watt}}$$





30% of datacenter power goes to network

So max savings by creating infinitely efficient optical interconnect is 30%!

BERKELEY LAB

Exploit the unique properties of photonics to improve both numerator and denominator!

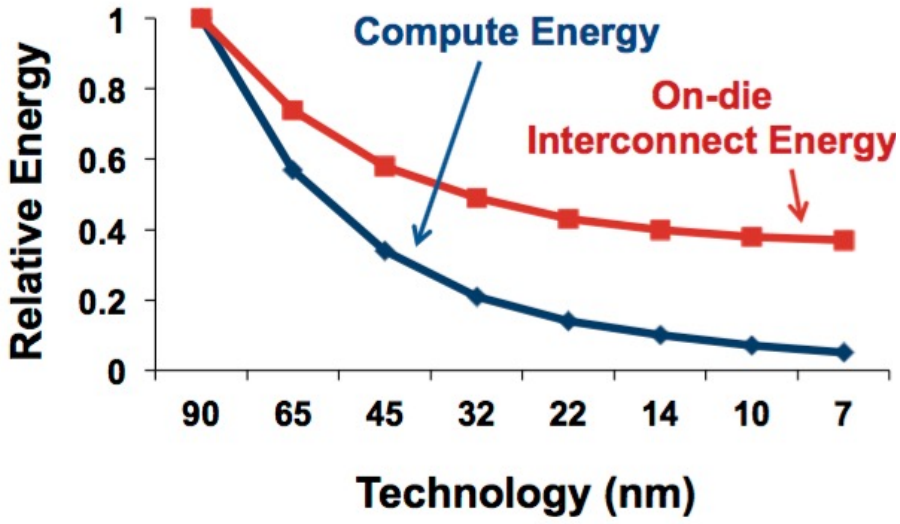
Some final shots and open questions for you to ponder

Question: If data movement is where all the energy is going?

- **Why do we *still* teach students that the order of complexity of algorithms as O(flops)?**

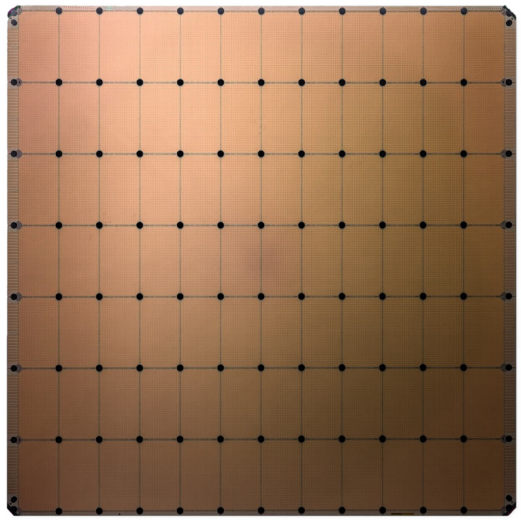
- **Can't we teach order of complexity for data movement?**
  - This has been discussed in SIAM meetings for a couple of decades
  - *Even started a PADAL workshop series with Thomas and Didem on this question*

- **What would a solution look like?**



Compute in Memory? (the elusiveness of NonVon Architecture?

- **Everyone talks about Compute In Memory as if it is easy.**
- **The very first concept Von Neumann had for a computer had no boundary between compute in memory**
  - He got over it
  - Why do we blame the poor guy for finding a solution to what is clearly a very difficult problem!
- **Compute in memory is often proposed to solve this Von Neumann Bottleneck**
  - Yet we have a commercially available compute in memory platform
  - Do you feel all of the problems have been solved?
  - Is it easy now?
- **Also, the fundamentals of our memory is broken**
  - DRAM is densest, but neither cell cycle time nor density has not improved at historical rates since hitting 10nm.

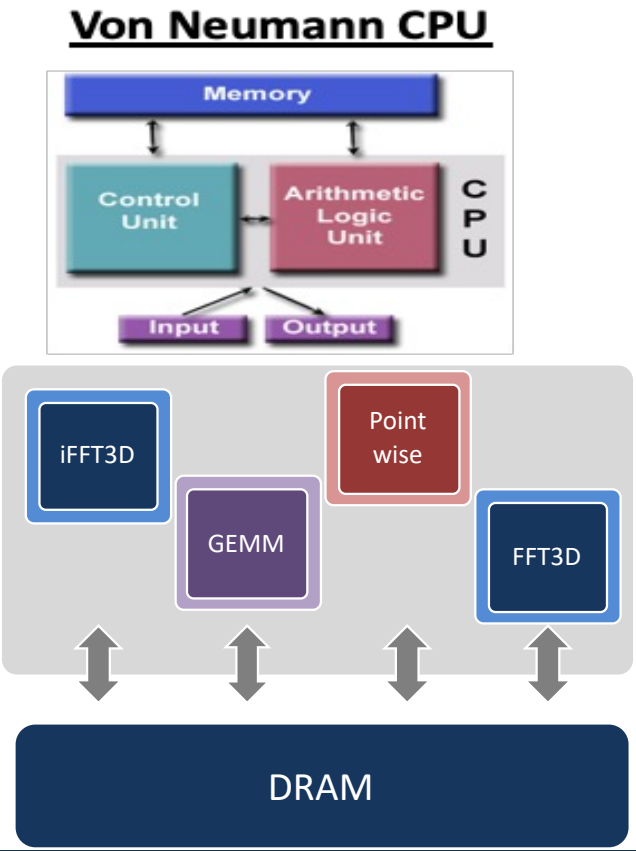


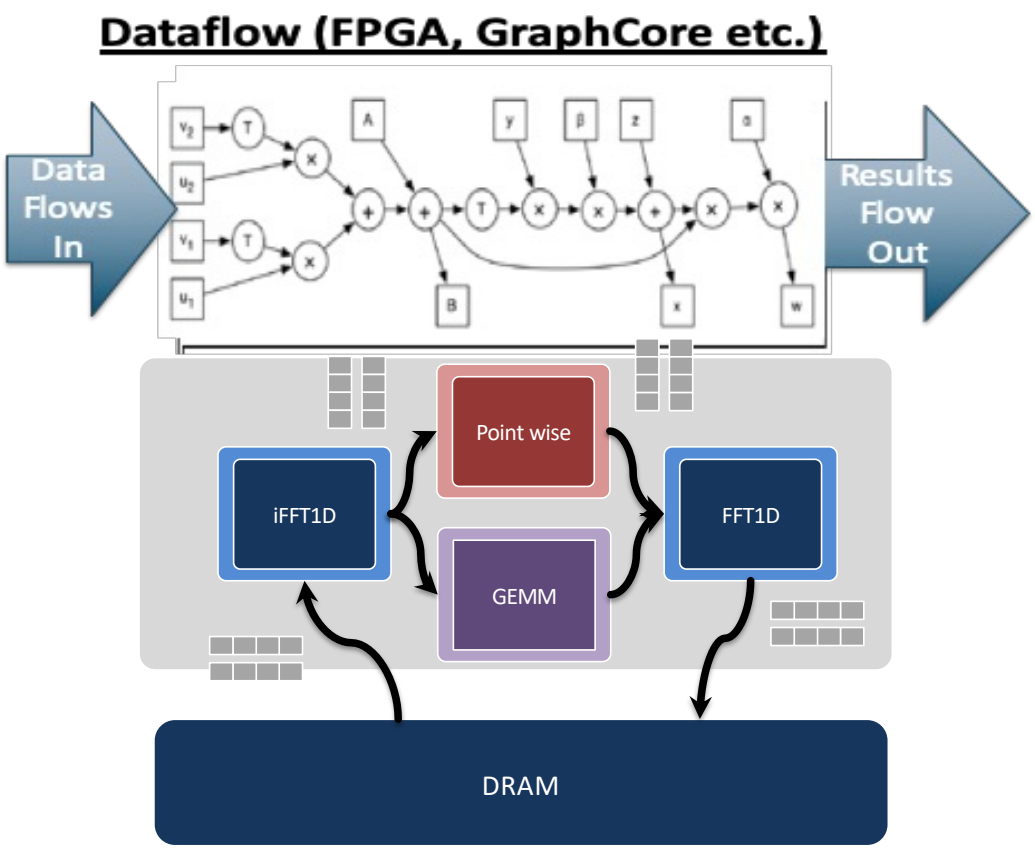
CEREBRAS WSE-3
46,225mm² Silicon
4 Trillion transistors

LARGEST GPU
826mm² Silicon
80 Billion transistors

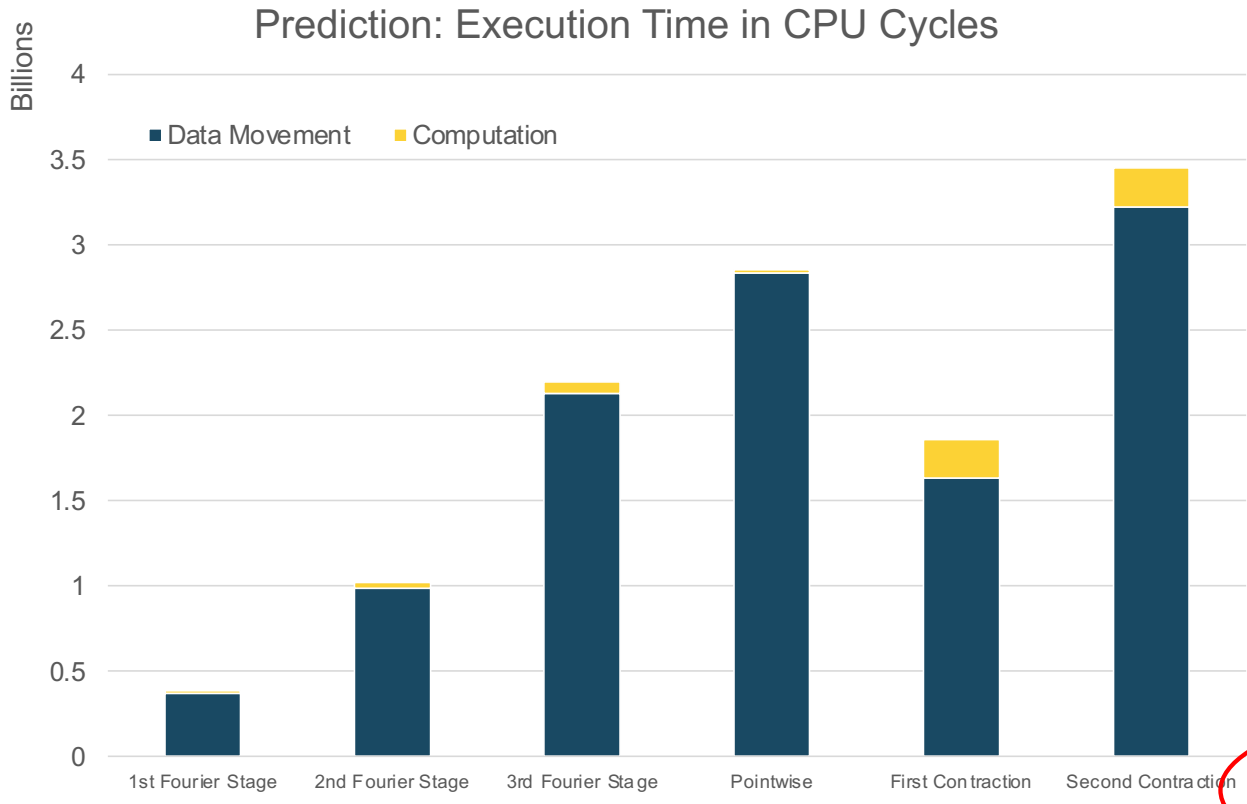
Offload model is an unproductive way to use hetero-acceleration

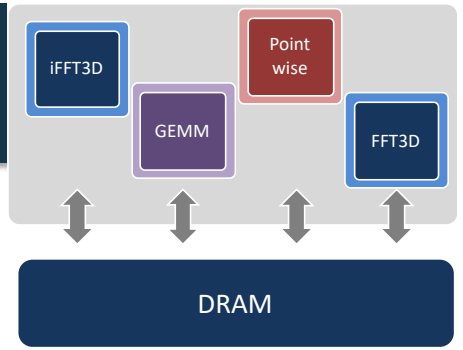
(redesign for static dataflow and deep flow-through pipelines?)





Modeling the Baseline Computation
Limitations of the Offload Model for Accelerators





Assumptions:

A. Hardware

- CPU with 8 cores
- CPU frequency is 3.9 GHz
- DRAM bandwidth 22 GB/s

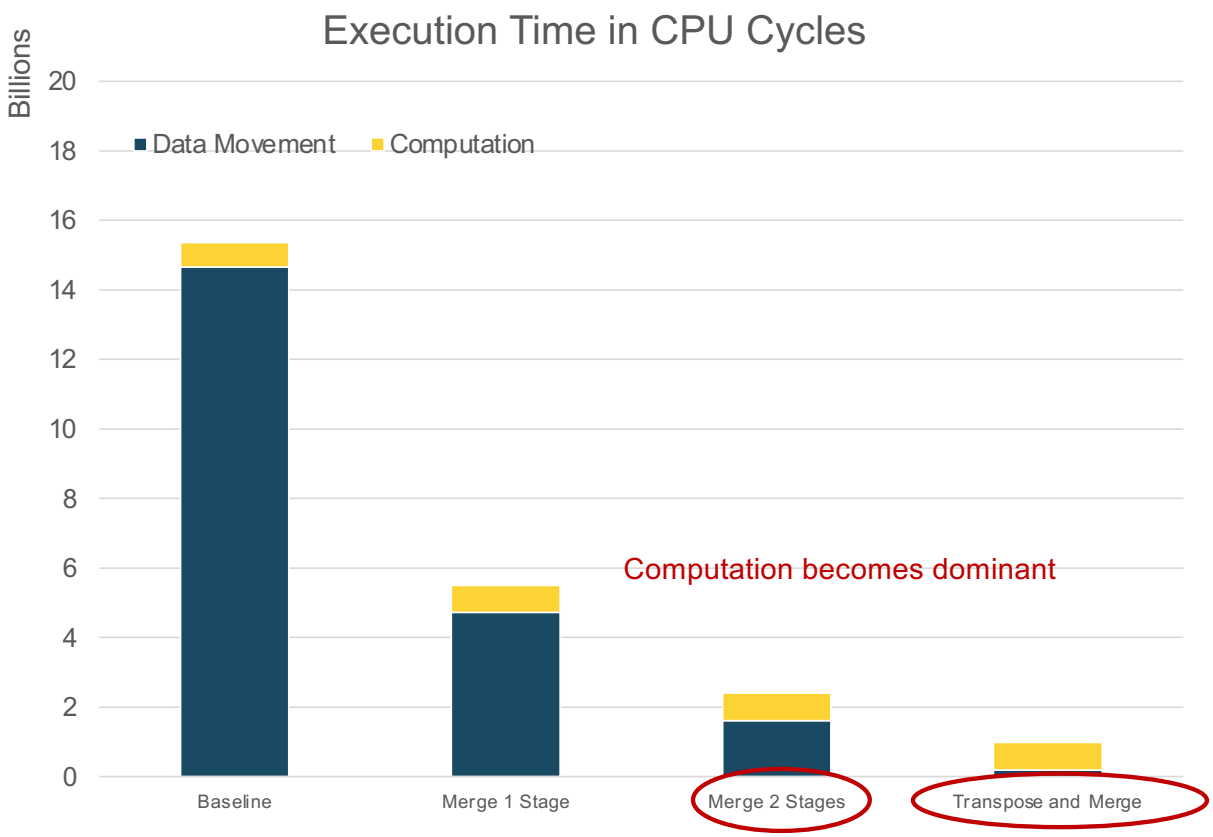
A. Problem Size

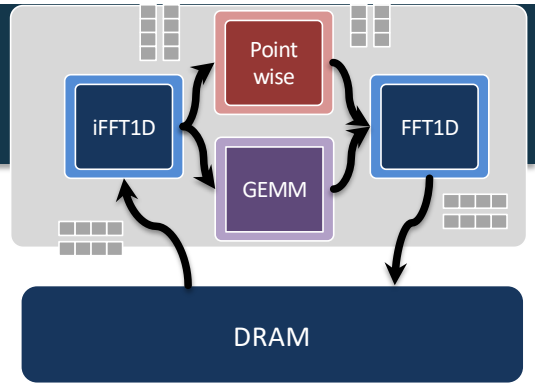
- Sphere diameter 64
- FFT size 128
- Number of bands 256
- Number of atoms 256

Prediction: **15.4 billion cycles**
Actual execution: **17.1 billion cycles**
Accuracy - 90%

Peer Accelerator Model (merged Kernels)

Convert Problem to Compute Bound





Assumptions:

A. Hardware

- CPU with 8 cores
- CPU frequency is 3.9 GHz
- DRAM bandwidth 22 GB/s

A. Problem Size

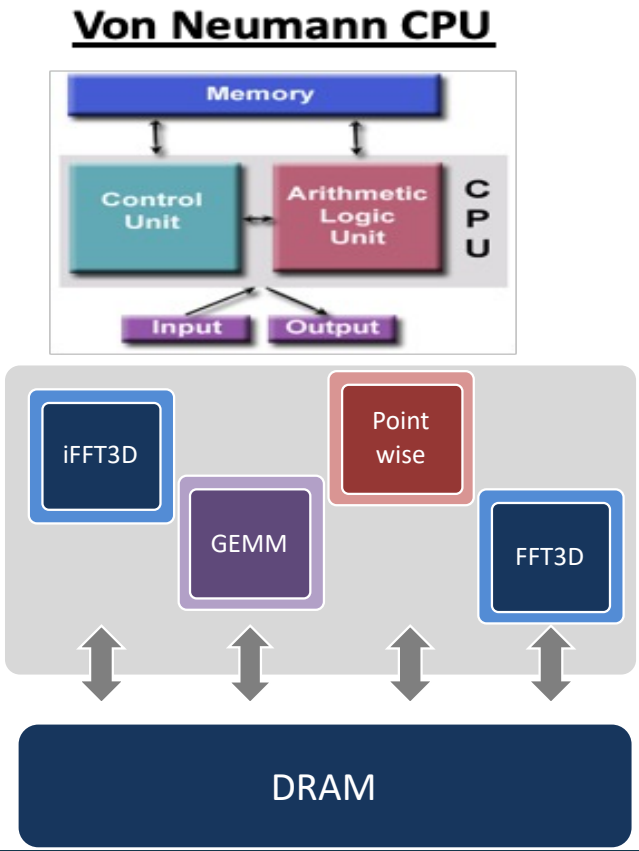
- Sphere diameter 64
- FFT size 128
- Number of bands 256
- Number of atoms 256

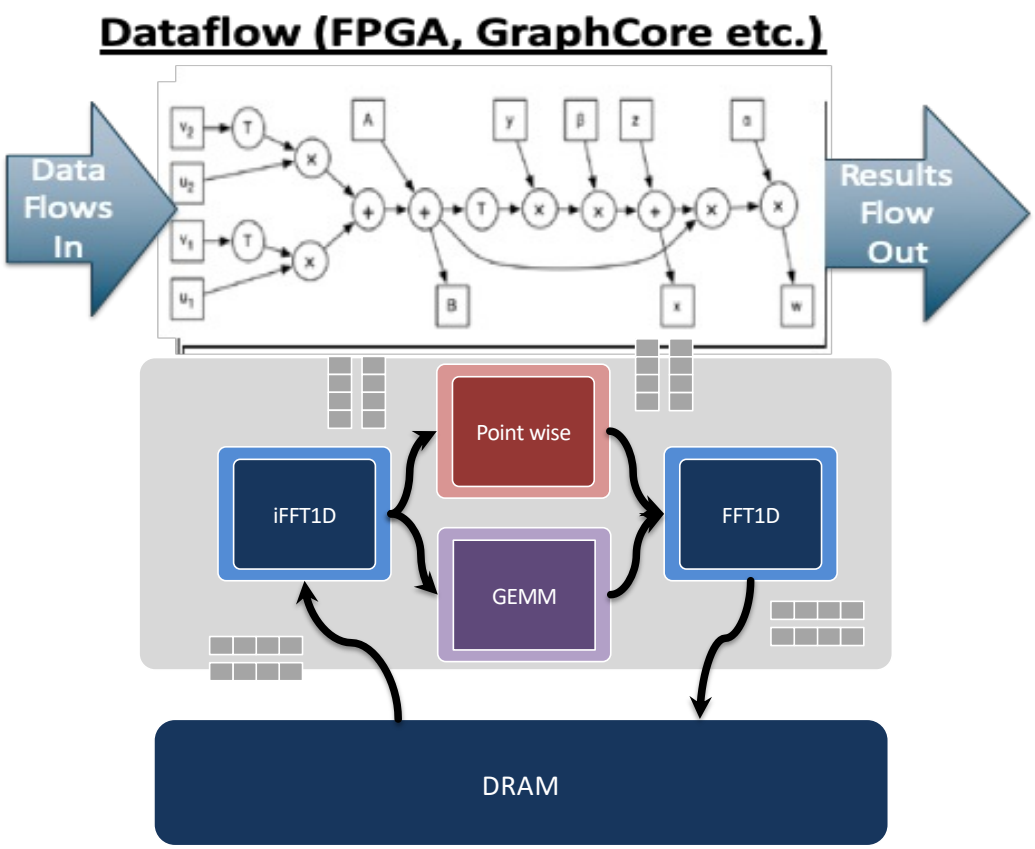
Merging different stages

The Transpose and Merge version requires a lot of on chip memory

Offload model is an unproductive way to exploit hetero-accel

(For p-models, this also breaks encapsulation in subroutines as an abstraction)





BERKELEY LAB

This is an NP-hard graph embedding problem

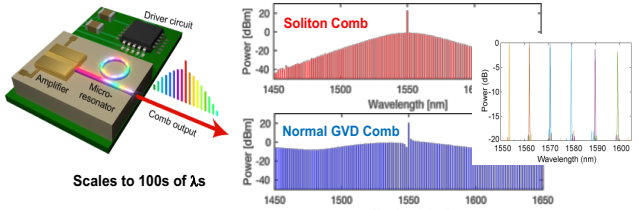
Conclusions

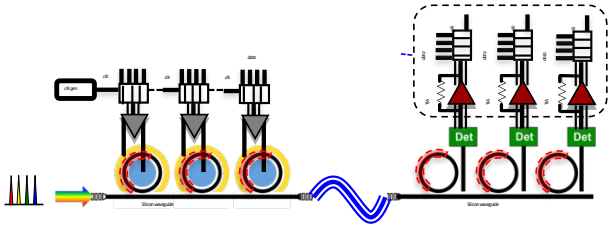
- **Future of energy efficient HPC and AI lies in specialization & Systems**
  - Modularity through advanced packaging
  - Codesign that spans *Algorithms, Software, and Hardware*
- **Systems Performance (not just device performance) is the future**
  - Components are bottom of the value chain
  - Systems enabled by the unique properties of photonic devices are the **top** of the value chain
  - Memory fabrics are the essential glue that enables us to delver such systems
- **Requires a modular approach to deliver systems**
  - Chiplets with baked-in memory fabrics are an opportunity for scalable modularity
  - OCP Workstreams to negotiate standards to promote that modularity

PINE: Photonic Integrated Networked Energy Efficient Datacenters

Addressing the datacenter energy challenge!

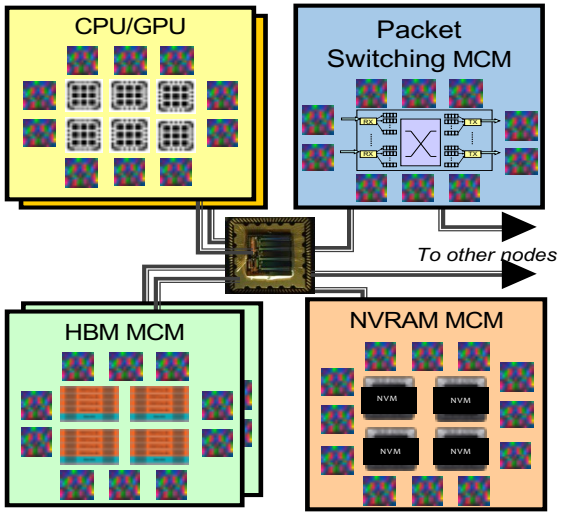
1) Energy-bandwidth optimized optical links



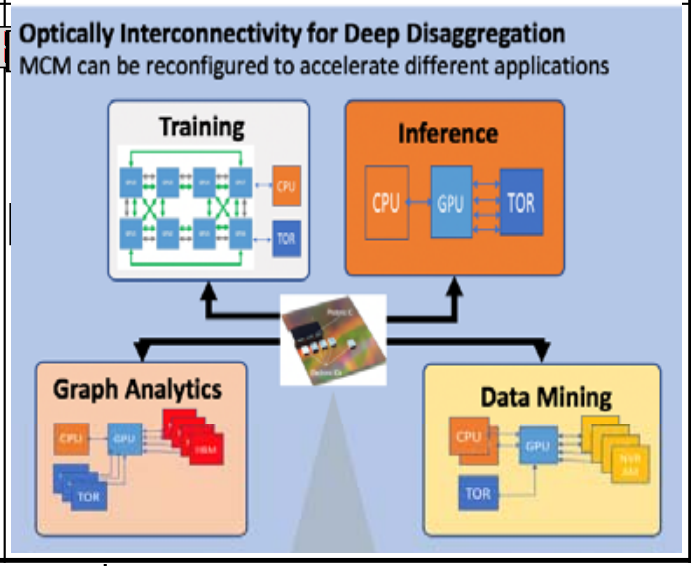


1 Tb/second per fiber

2) Embedded silicon photonics into OC-MCMs



3) Bandwidth steering for Custom Node Connectivity



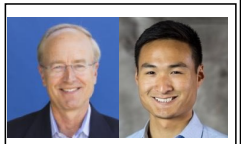
ENLITENED

arpa·e
CHANGING WHAT'S POSSIBLE

Bergman

Baiocco

SUNY Polytechnic Institute



Bowers Liu

Q QUINTESSENT

Glick Gaeta Lipson

AIM photonics

Patel

CISCO

Dennison Gray

NVIDIA

Shalf

BERKELEY LAB

Ghobadi

MIT

OCP Modularity for HPC & AI Workstream needs your input!

Patricia G.
(LBNL)

George M.
(LBNL)

Anu R.
(Microchip)

luisa.gonzalez@ocproject.net
georgios.michelogiannakis@ocproject.net
tony.gutierrez@ocproject.net
anu@microchip.com

How to participate:

- Subscribe to the Project Mailing List
- Add the Project Meeting Calendar
- Join the Conversation

end