Chiplets for HPC

George Michelogiannakis, LBNL
Material credit: John Shalf, LBNL

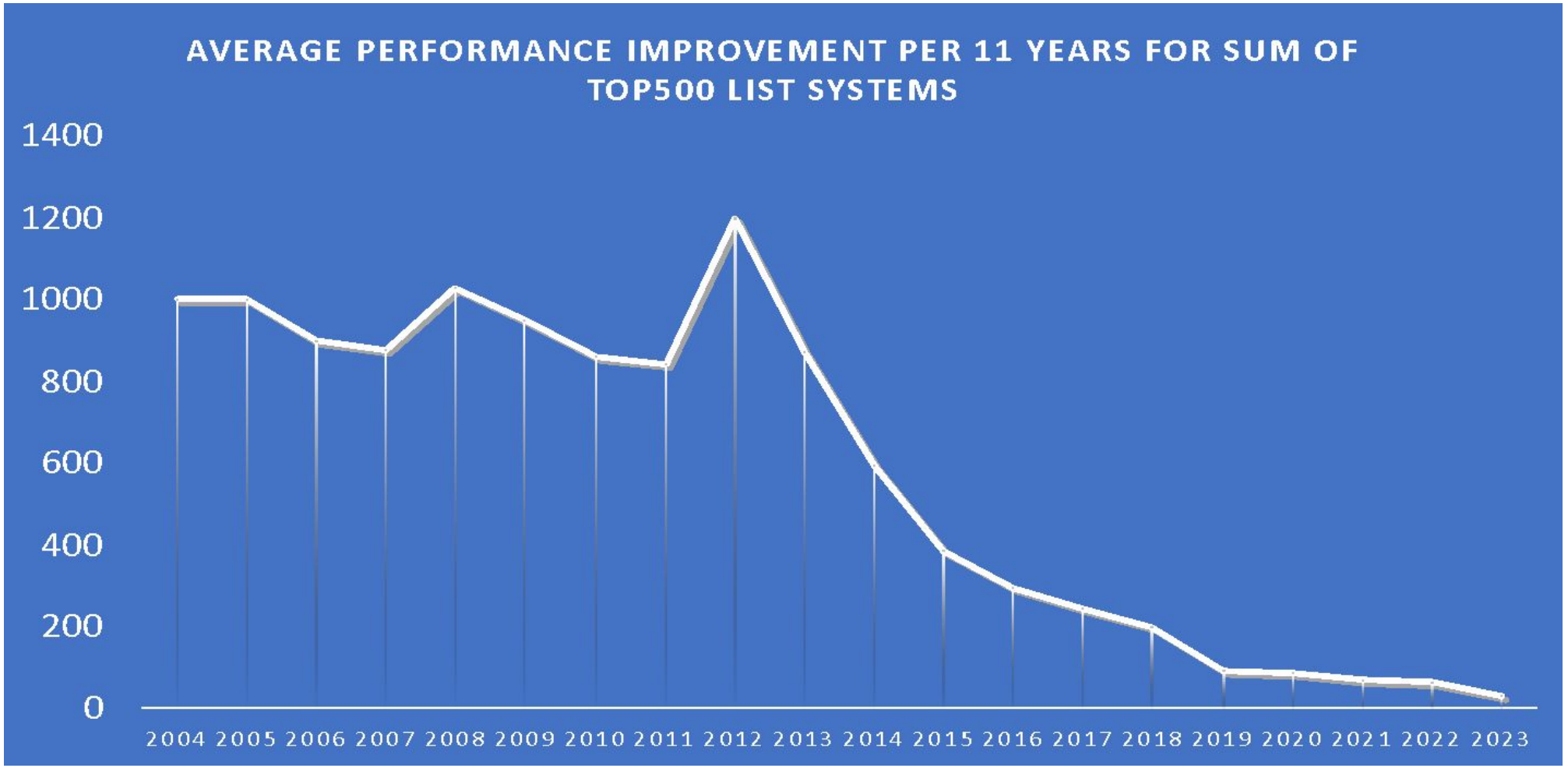
Open Chiplet Economy

OCP Sponsored Tutorial

Chiplet Summit Feb 6th 2024
1:00 pm to 5:00 pm
Santa Clara California, USA

OPEN Compute Project®

HPC's Future if we Don't Change Course



Specialization is Nature's Way

Powerful General Purpose

Xeon, Power

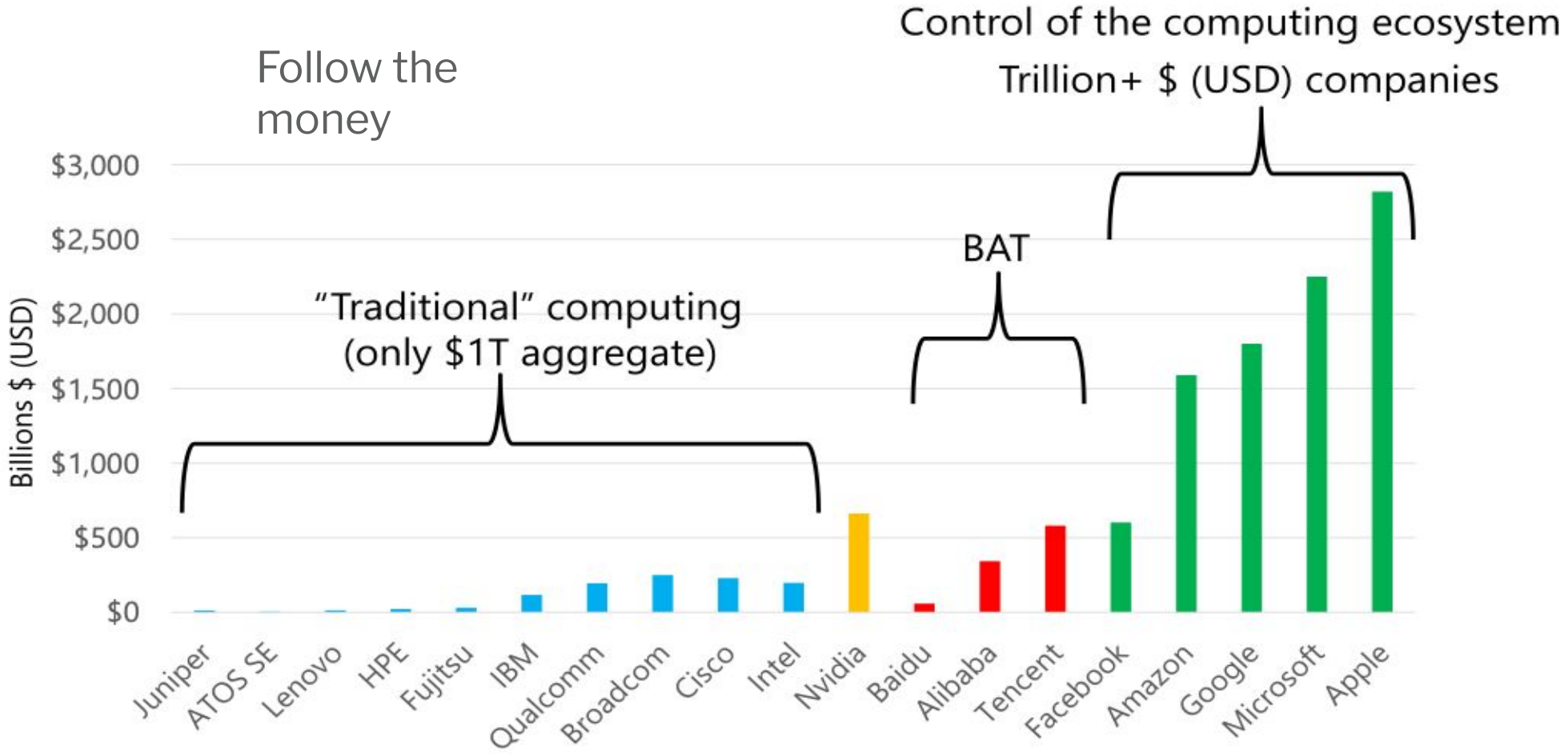
Many Lighter Weight
(post-Dennard scarcity)

Intel KNL, AMD, Cavium/Marvell, GPUs

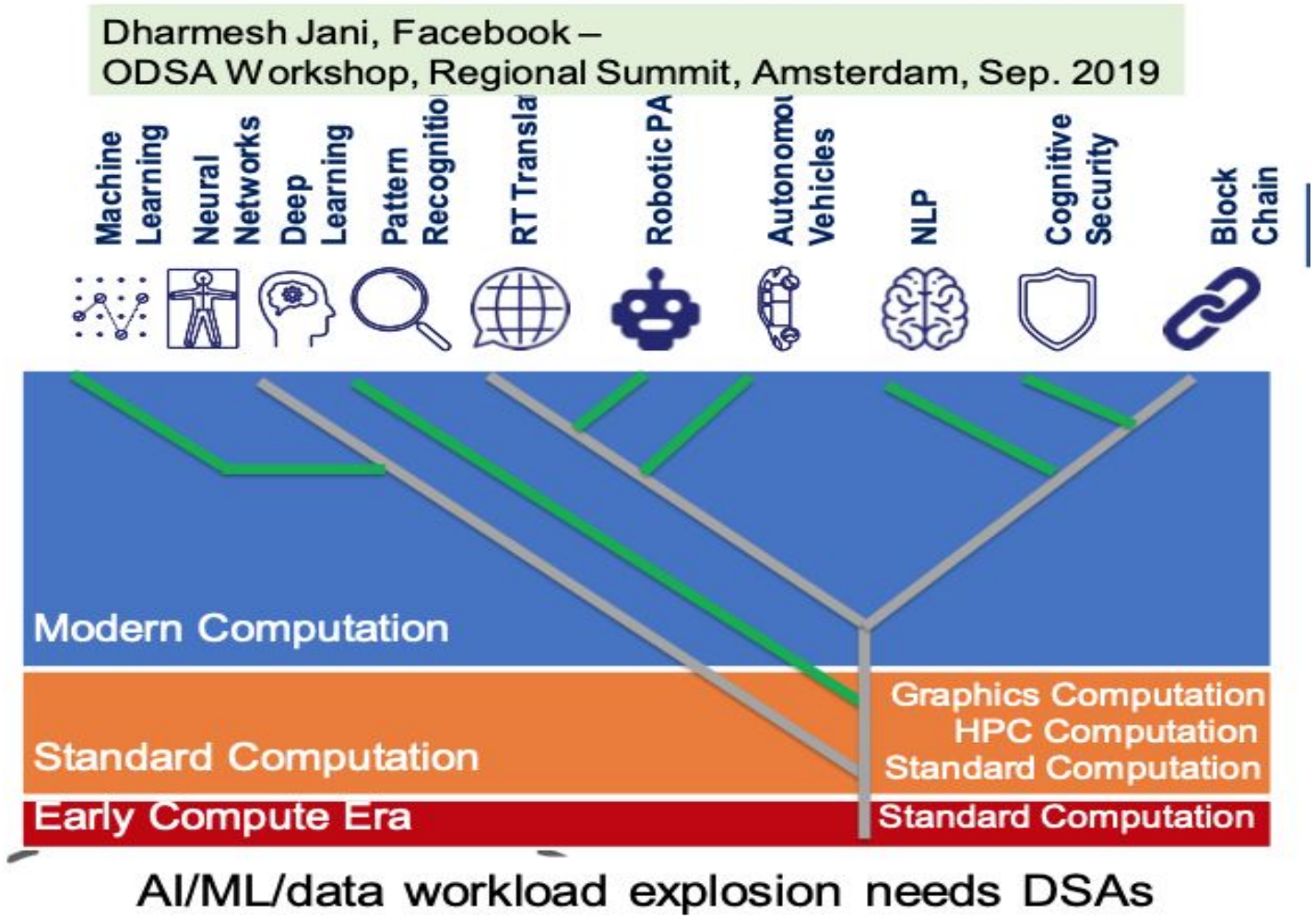
Many Different Specialized
(Post-Moore Scarcity)

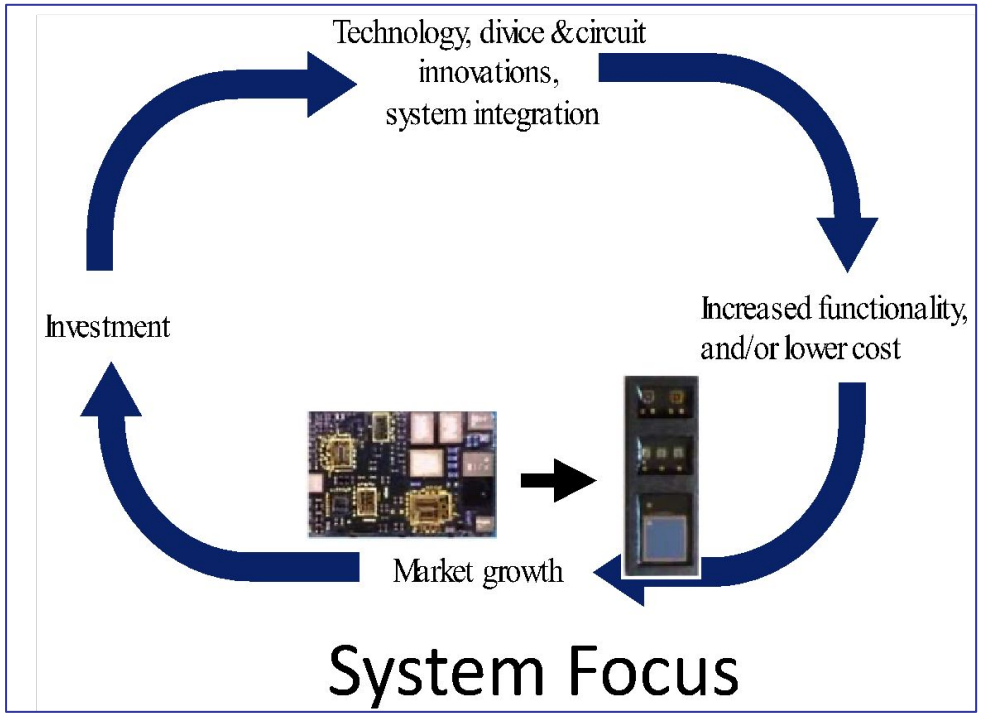
Apple, Google, Amazon, AWS

We Have to Understand The Market



Domain Specific Compute Driven by Hyperscalars





Neil Thompson

Opportunity for HPC: New Economic Model

Open Chiplets Marketplace is forming (ODSA and UCIexpress)

- Licensable IP and assembly by 3rd party lowers that barrier
- Leverage the economic model being created by HyperScale

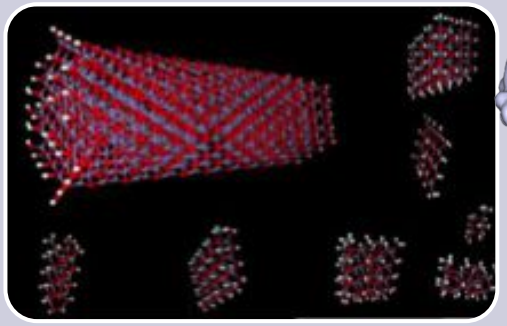
Leverage this baseline and extend to support HPC

- Smaller incremental cost for HPC to "play"
- *HPC has become "too small to attack the city"*

80:20 Rule: Focus open efforts on what uniquely benefits HPC

- Build up a library of reusable accelerators for HPC.
- Interoperability for sustainability: *Interoperate with commercial IP where it exists and focus on open the 20% that doesn't make commercial sense to license*

Architecture Specialization for Science



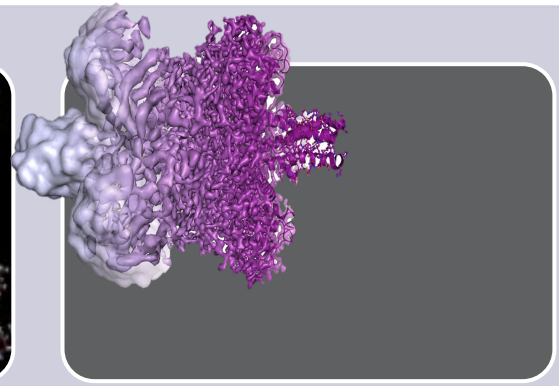
Materials

Density Functional Theory (DFT)

Use O(n) algorithm

Dominated by FFTs

FPGA or ASIC

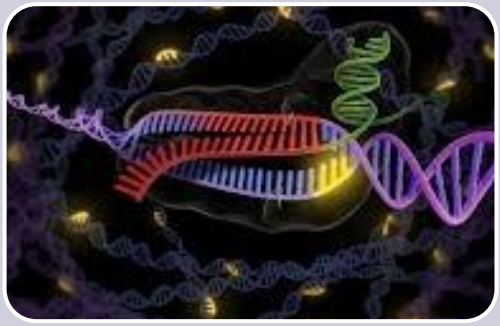


CryoEM Accelerator

LBNL detector

750 GB / sec

Custom ASIC near detector



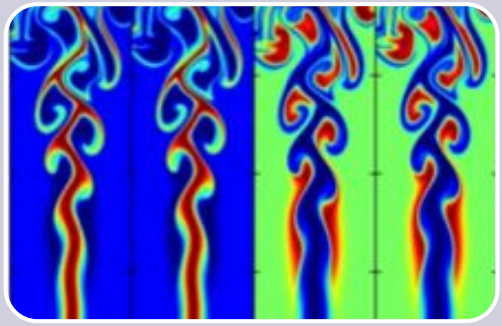
Genomics Accelerator

String matching

Hashing

2-8bit (ACTG)

FPGA



Digital fluid Accelerator

3D integration

Petascale *chip*

1024-layers

General / special HPC solution

Algorithm-Driven Design of Programmable Hardware Accelerators

Example: LS3DF/Density Functional Theory (DFT)

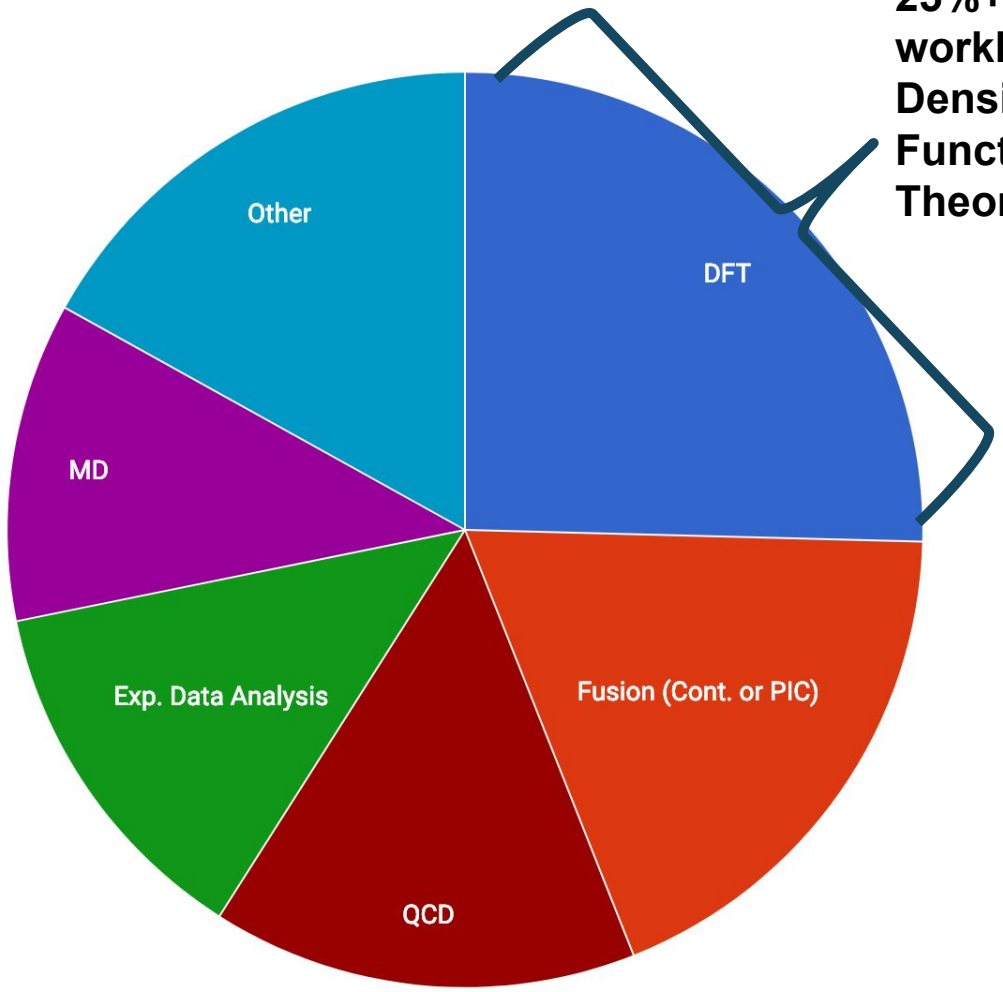
What: Design the hardware acceleration around the target algorithm/application

Why: Huge opportunities to improve performance density and efficiency

- FFT hardware accelerator 50x-100x faster than GPU (using SPIRAL generator)

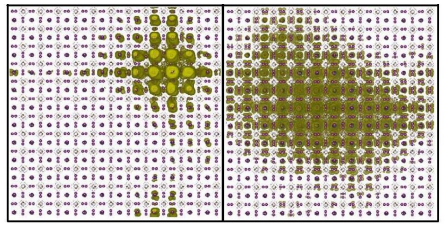
How: Target Density Functional Theory

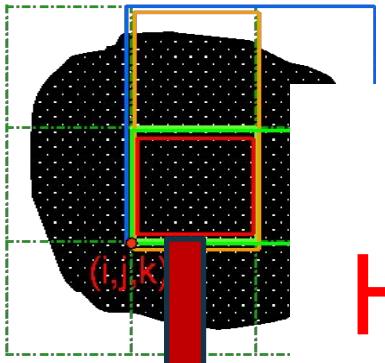
1. Large fraction of the DOE workload
2. Mature code base and algorithm
3. LS3DF formulation minimizes off-chip communication and scales O(N)

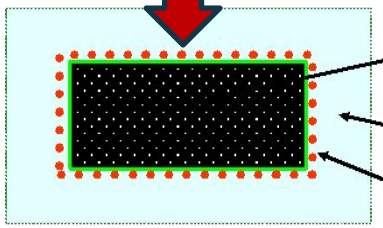


The DFT kernel for each fragment

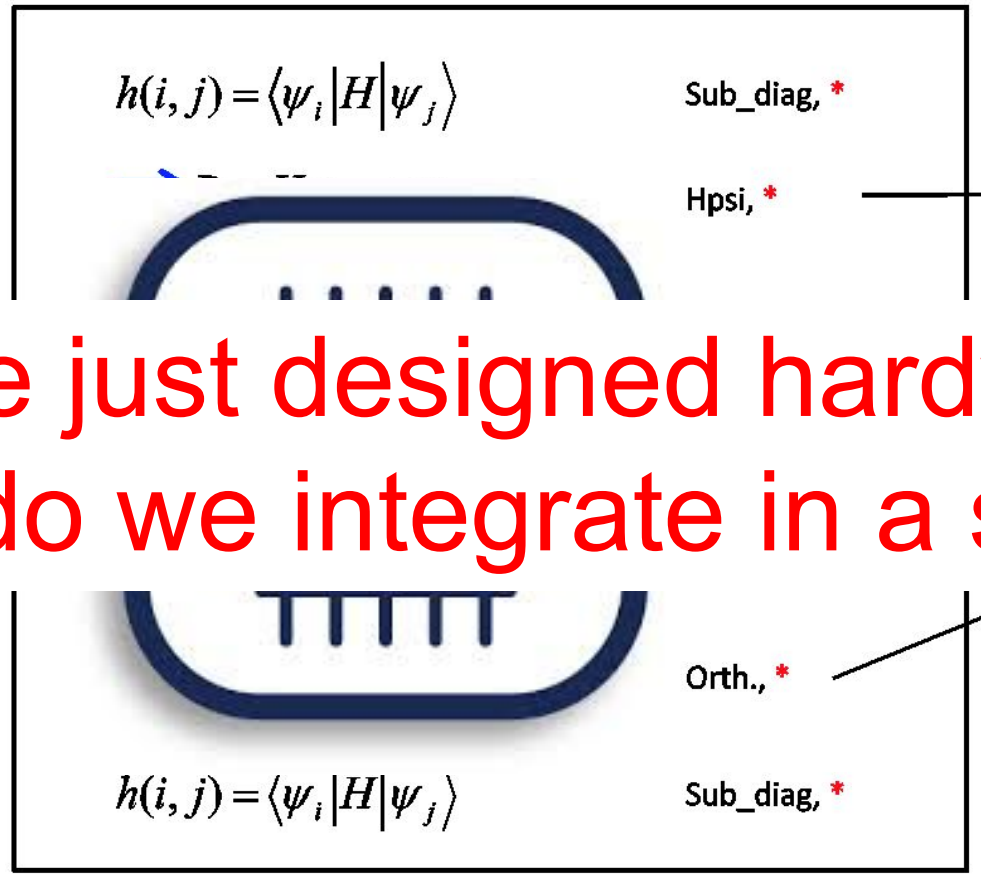
Communication Avoiding LS3DF Formulation – Scales O(N)







Fragment (2x1)



$O(N^2 Log(N))$

Comm bound if non-local

3D parallel FFT

Choelesky

EMM

(N^3)

Compute-bound

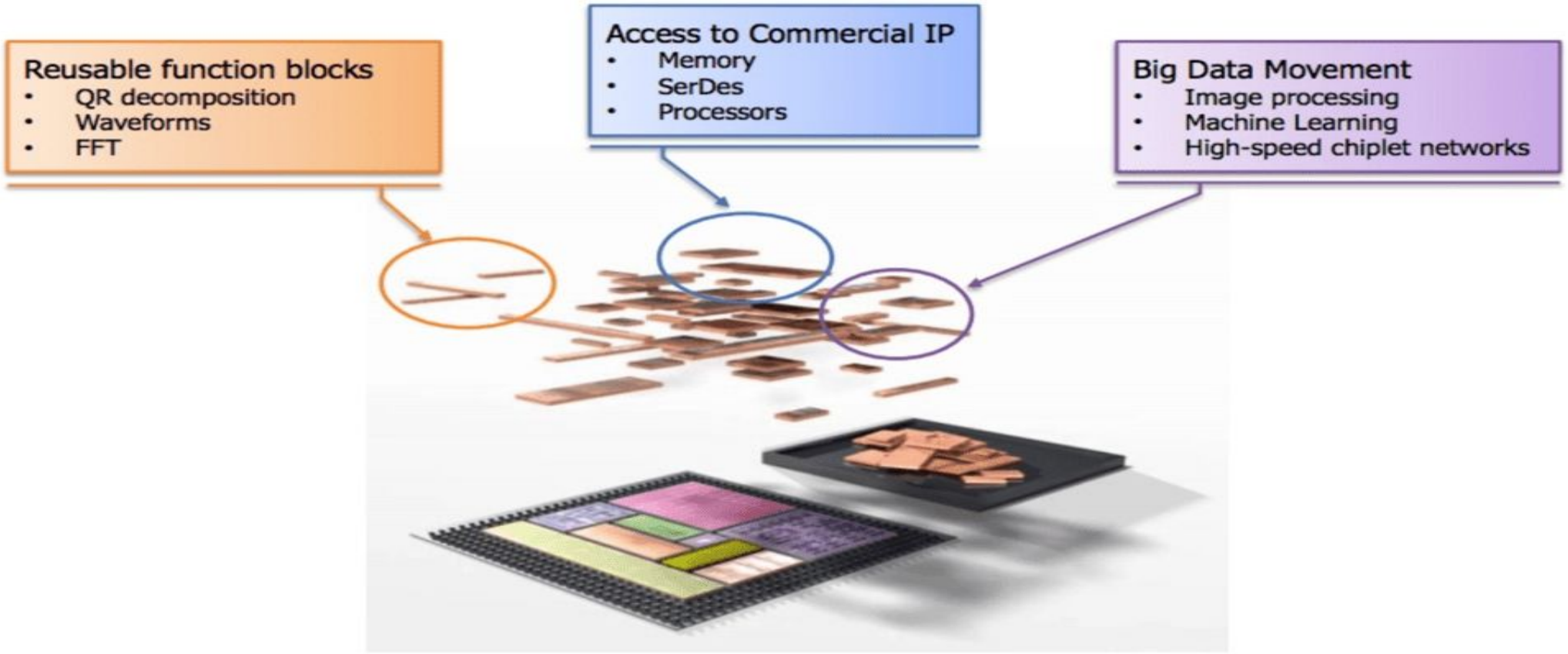
We just designed hardware

How do we integrate in a system?

LS3DF O(N) Algorithm Formulation
Minimizes off-chip Communication

Compute Intensive Kernels
Targeted for HW Specialization

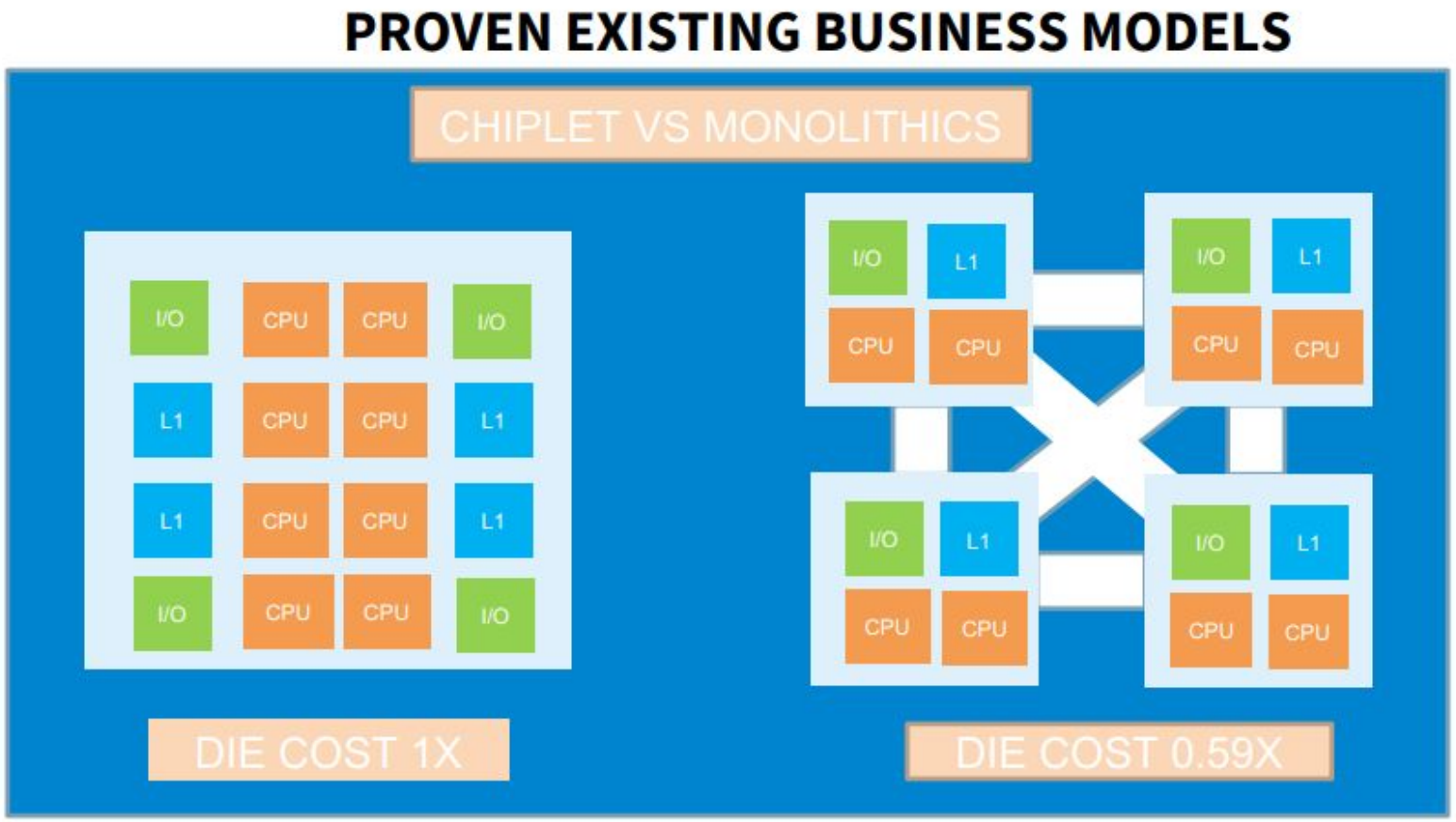
Chiplets Make Specialization Accessible for HPC



From DARPA CHIPS

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

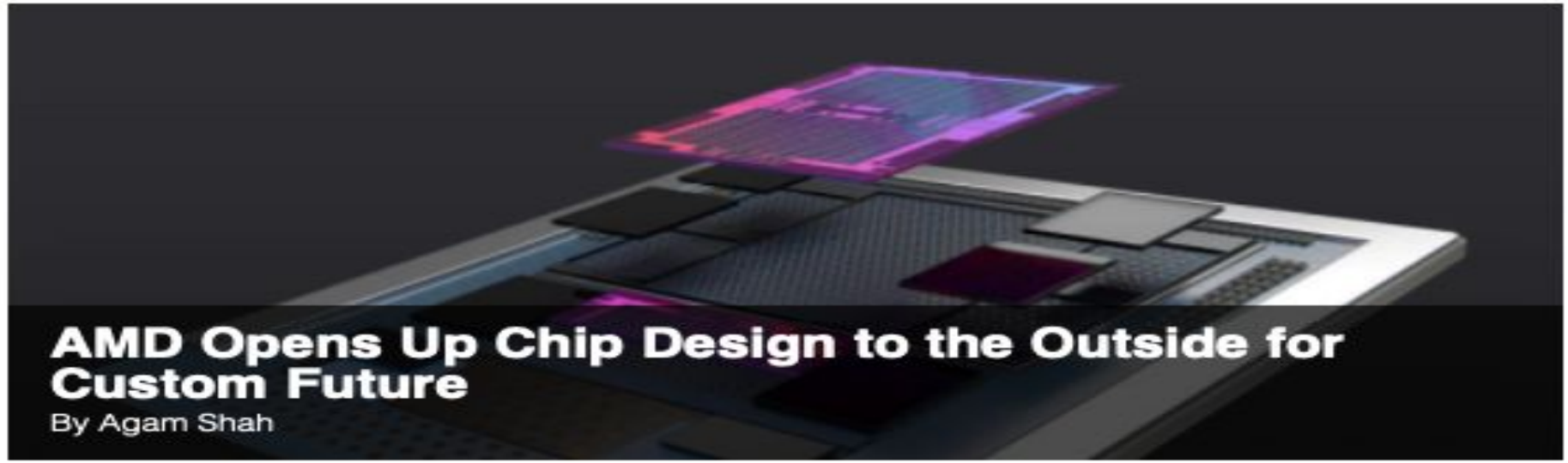
More Flexible and Lower Cost



[L. Su, IEDM'17]

HPC wire

Since 1987 - Covering the Fastest Computers in the World and the People Who Run Them



AMD Opens Up Chip Design to the Outside for Custom Future

By Agam Shah

June 15, 2022

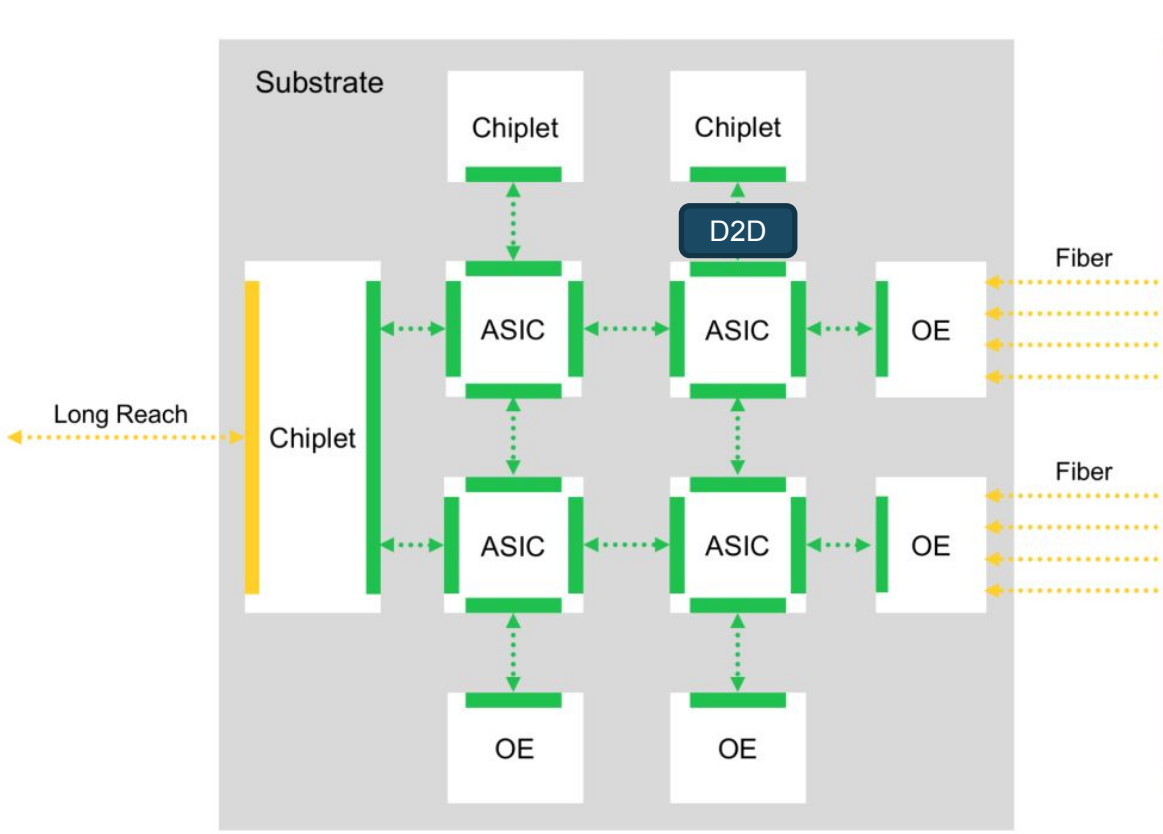
AMD is getting personal with chips as it sets sail to make products more to the liking of its customers.

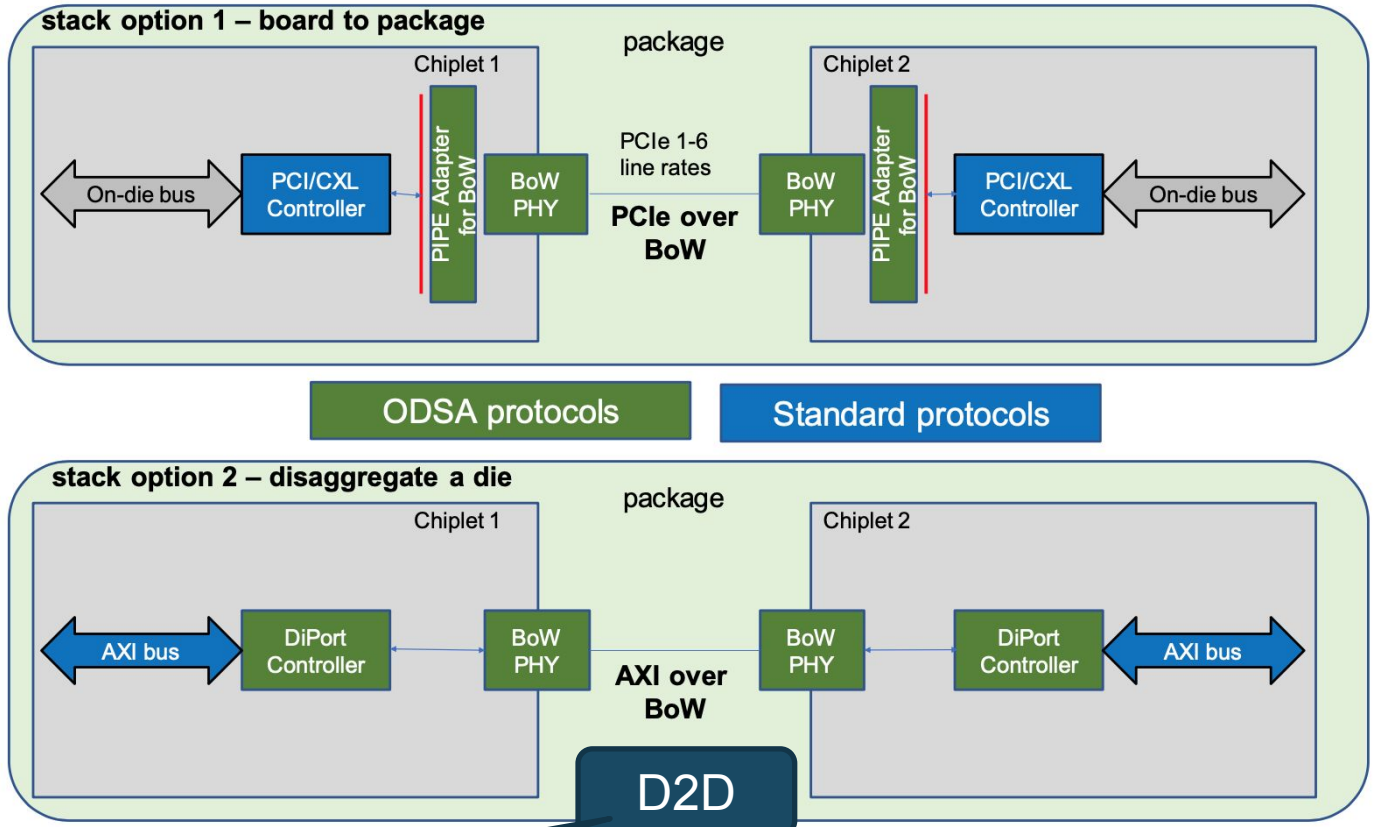
The chipmaker detailed a modular chip future in which customers can mix and match non-AMD processors in a custom chip package.

"We are focused on making it easier to implement chips with more flexibility," said Mark Papermaster, chief technology officer at AMD during the analyst day meeting late last week.

AMD will allow customers to implement multiple dies — also called chiplets or compute tiles — in a tight chip package. AMD already uses tiles, but is now welcoming third parties to make accelerators or other chips to be included in 2D or 3D packages alongside its x86 CPUs and GPUs.

Standardized die-to-die (D2D) Physical Layer Interfaces (ODSA)



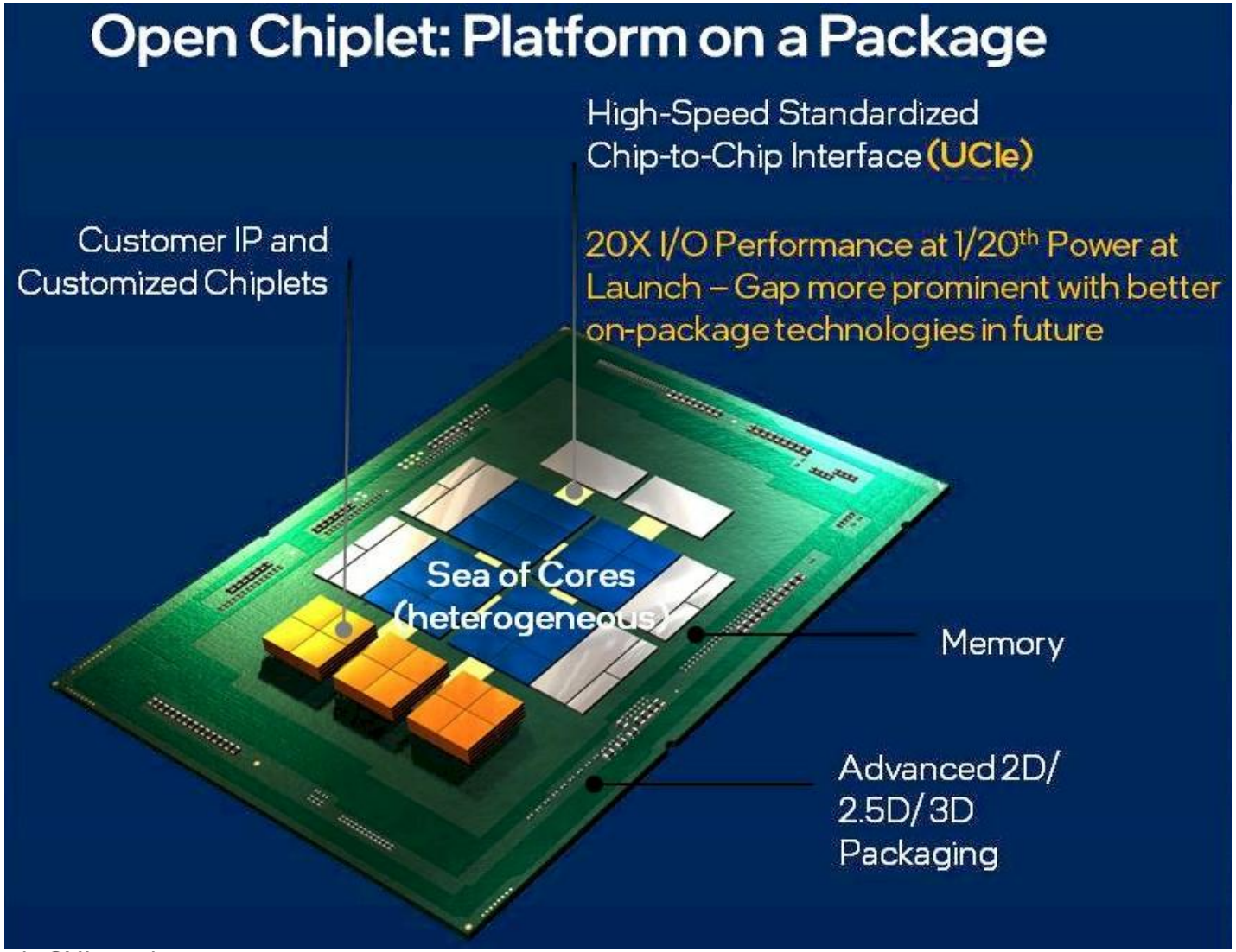


Blue Cheetah supplies the IP for the Die-to-Die (D2D) Phy.

A protocol: UCIe

Uses CXL or PCIe

I/O attach with PCIe/CXL.io
- Memory use cases: CXL.mem
- Accelerator use cases: CXL.cache



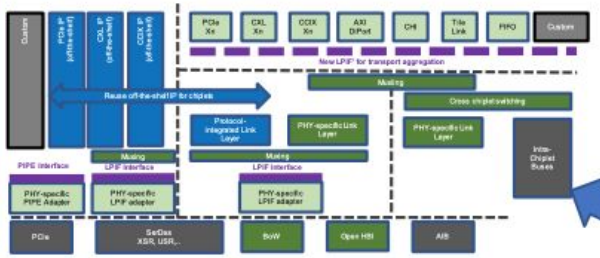
https://www.nextplatform.com/2022/03/02/industry-behemoths-back-intels-universal-chiplet-interconnect/

https://www.snia.org/sites/default/files/PM-Summit/2022/PMCS22-Park-CXL-and-UCIe.pdf

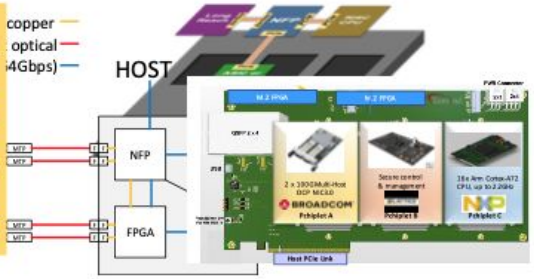
ODSA: Open Domain Specific Architecture

Creating an Open Chiplet Marketplace

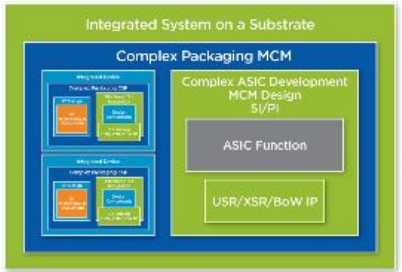
Open D2D Interface
Reduce barrier to interoperation



Reference Designs
Starting point for new designs



Reference Workflows
Reusable, open practices

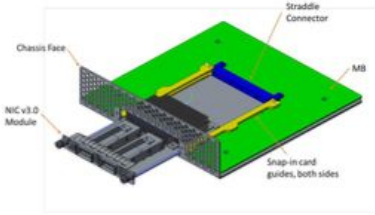


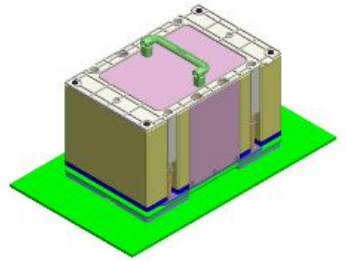
ODSA Activities

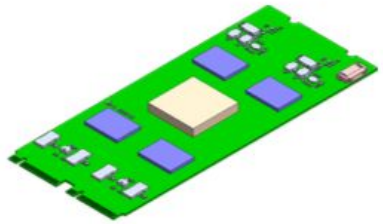
Project

Chiplet Marketplace

Integrate best-in-class chiplets from multiple vendors through open interfaces

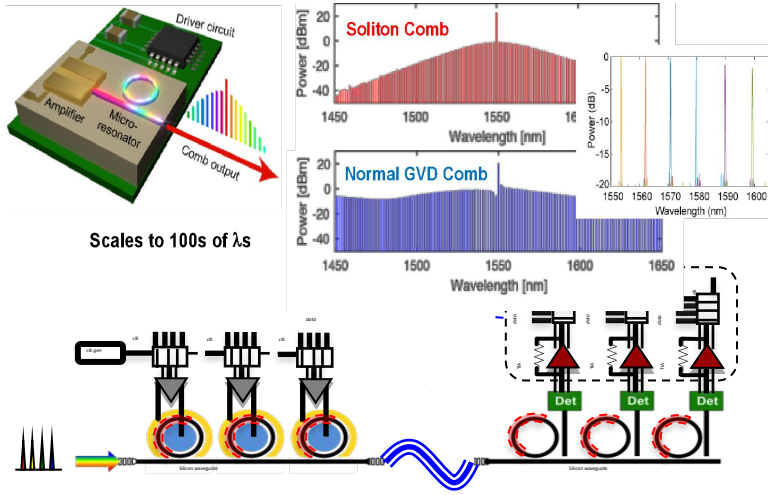






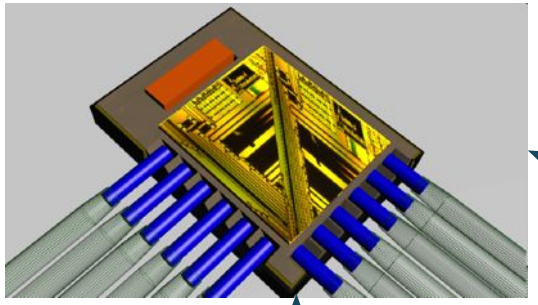
OCP modular form factors

Photonic MCM for High Escape Bandwidth for Remote Memory

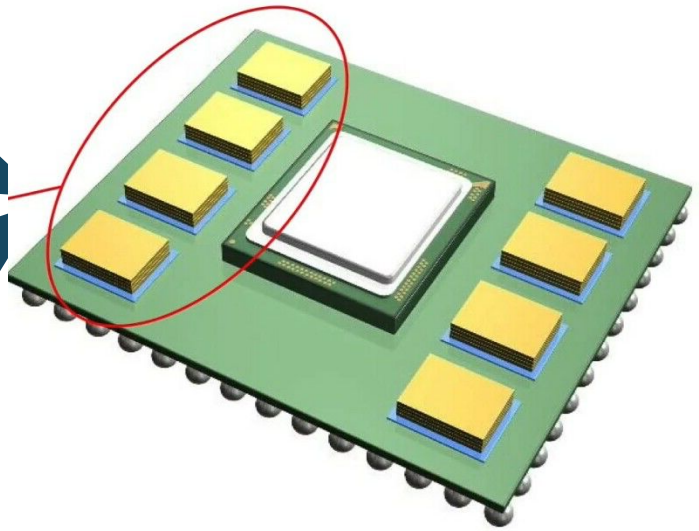


Scales to 100s of λs

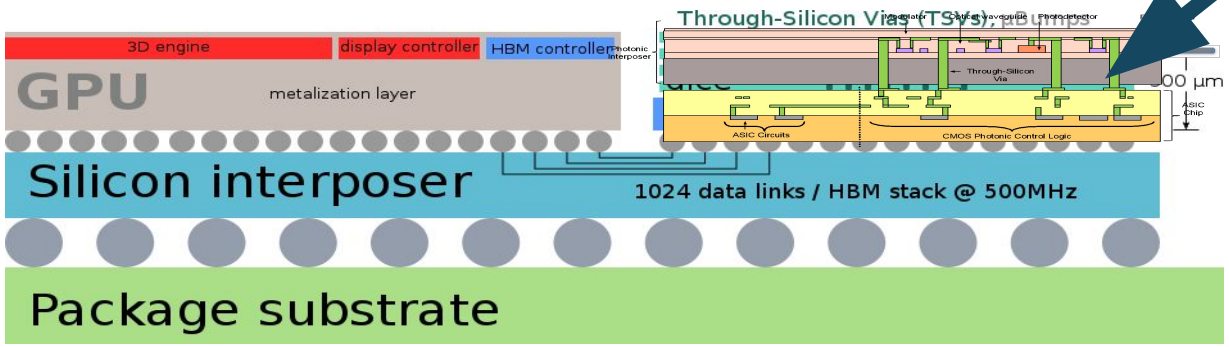
Comb Laser Source with
DWDM Silicon Photonics
Wide-and Slow for high speed links



Photonic SiP



MCM: Multi chip module

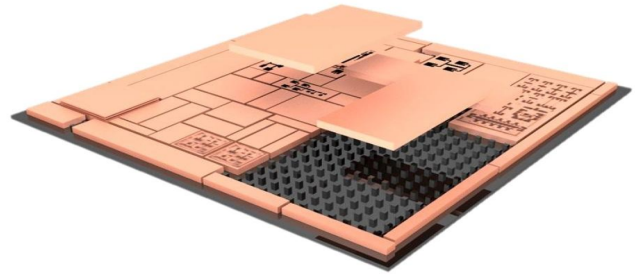


Project38: HPC Improvements Through Innovative Architecture

Cross-agency architectural exploration

Project 38 (P38) is a set of vendor-agnostic architectural explorations involving DOD, the DOE Office of Science, and NNSA

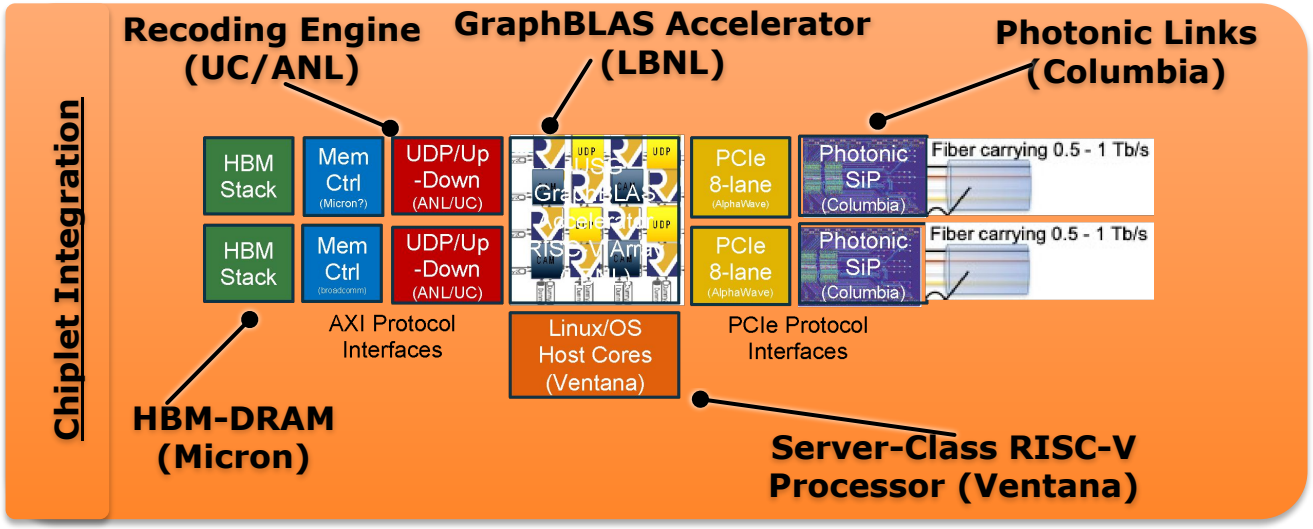
- **Mission**:
- Demonstrate high performance IUSG node -- codesigned to accelerate GraphBLAS
- Demonstrate modular integration of LBNL/ANL IUSG + commercial IP using Open Chiplets
- Create new capability for the USG to rapidly assemble/prototype server-class chip designs



Affordable heterogeneous co-integration using chiplets

Accomplishments thus far

- *Released integration platform (MoSAIC)*
  - *Abstract model to RTL to chiplets or FPGAs*
- *Created end-to-end cost model for chiplet integration*
- *Chisel FFT, sparse matrix multiply, and TSGR generators*
- *GraphBLAS Accelerator ISA for RISC-V (GISA)*
- *AMD collaboration showed benefit of sparse matrix/tensor accel*

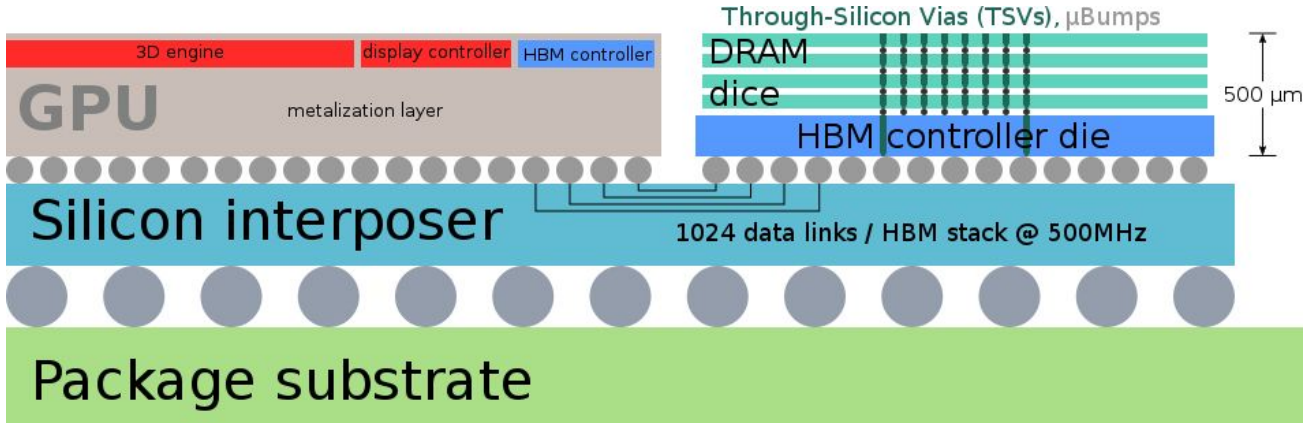


Look for the project 38 poster!

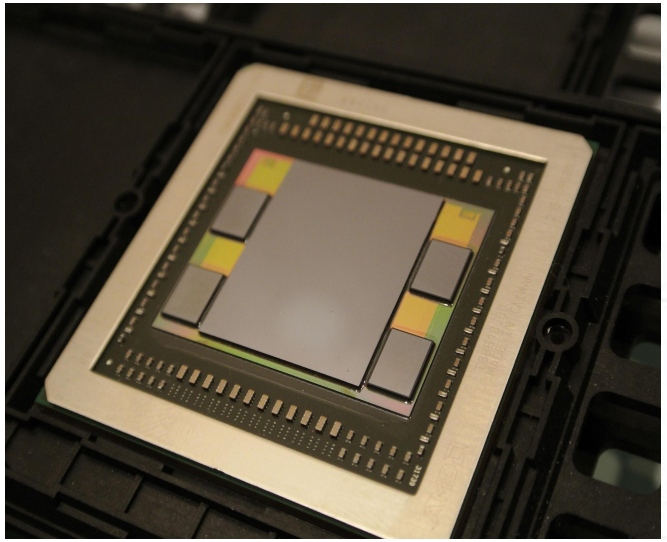
BERKELEY LAB

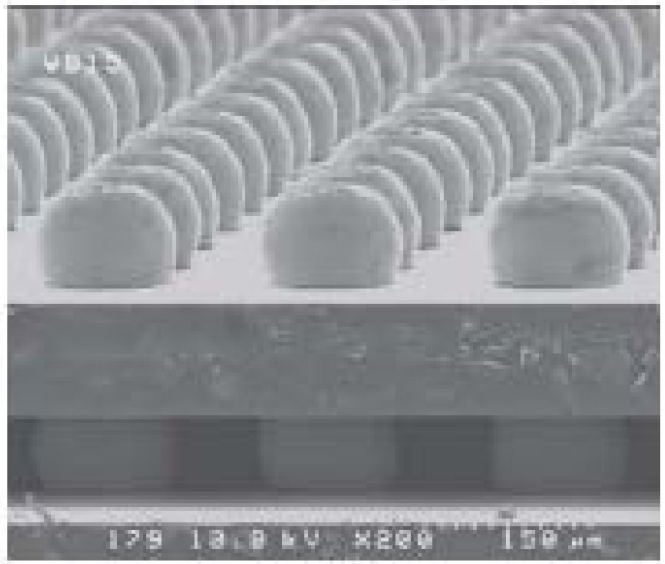
Questions?

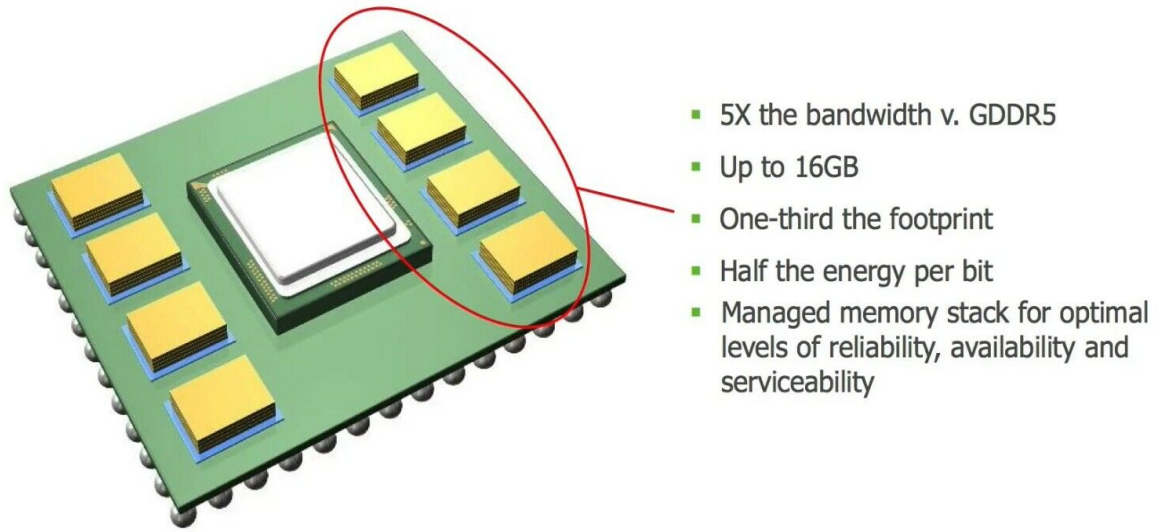
One Challenge is Escape Bandwidth



- **Good News:** Extend bandwidth density and lower power/bit
- **Bad News:** Limited (~2cm) reach
  - Cannot get outside of the package *(**<u>but we need to</u>**)*







Chiplet Bandwidth Roadmap (5 generations of BW doubling)

Table 5: Physical IO Scaling Roadmap for 2D and Enhanced-2D Architectures that use both solder and hybrid interconnects.

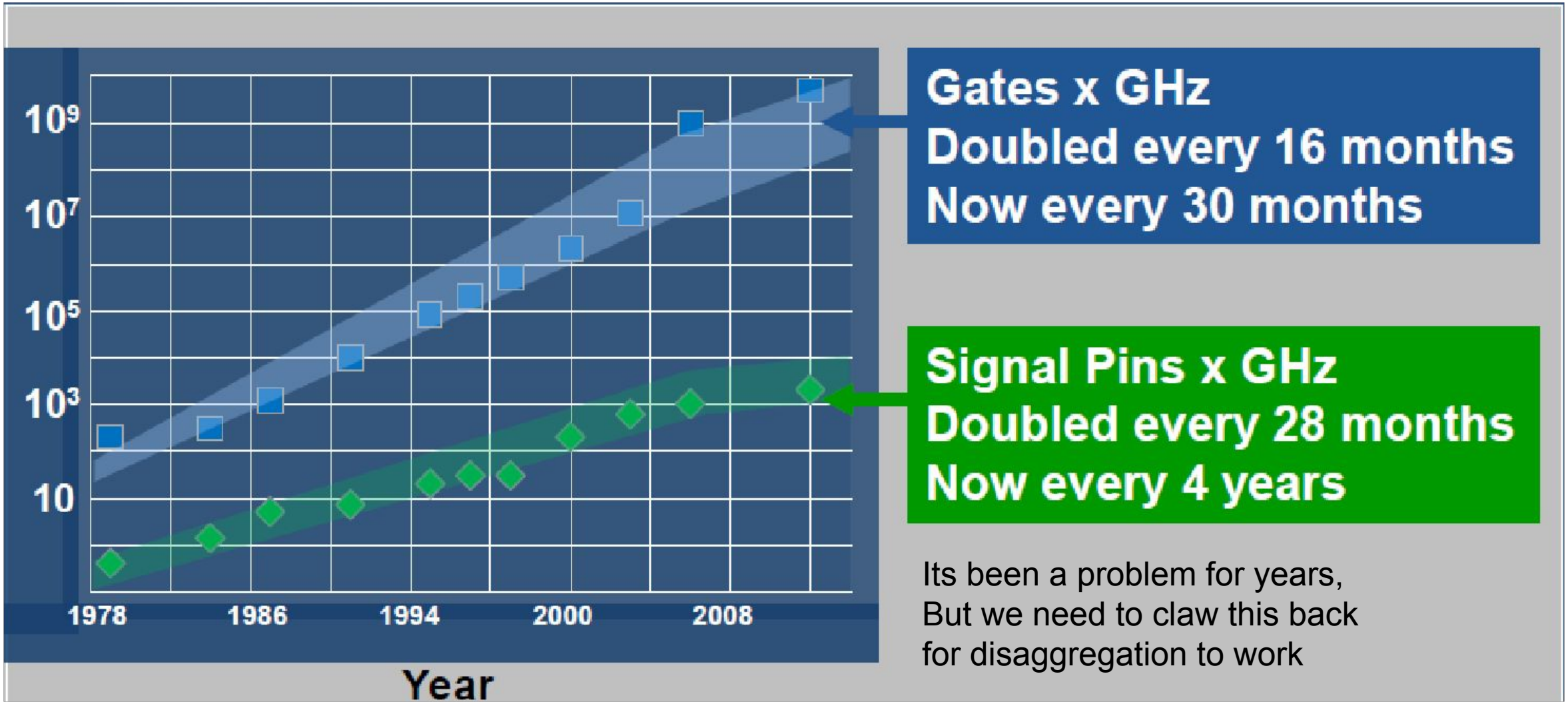
Generation Number →		**1**	**2**	**3**	**4**	**5**
Raw Linear Bandwidth Density (GBps/mm)		125	250	500	1000	2000
Package Technology	Minimum Bump Pitch (µm)[17]	55	40	30	20	10
	Linear Escape Density (IO/mm)	500	667	1000	2000	4000
	Areal Escape Density (IO/mm^2)	331	625	1111	2500	10000
Signaling Speed (Gbps)		2	3	4	4	4

5.1.2 Area Interconnects for 3D Architectures (see Figure 1)

Table 6: Physical IO Scaling Roadmap for 3D architectures that use both solder and hybrid interconnects.

Generation Number →		**1**	**2**	**3**	**4**	**5**
Raw Areal Bandwidth Density (GBps/mm^2)[18]		125	250	500	1000	2000
Package Technology	Minimum Bump Pitch (µm)[19]	40	30	20	15	10
	Areal Escape Density (IO/mm^2)	625	1111	2500	4444	10000
Signaling Speed (Gbps)[20]		1.6	1.8	1.6	1.8	1.6

Package Limited Bandwidth



Source: J. Poulton, Nvidia

Rapid Prototyping of HPC Data Analytics Engine using Open/Modular Chiplets

Motivation

- MPW prototyping of chip designs necessarily have small chips – lower performance
- Many necessary subsystems (memory controllers, PCIe) better supplied from commercial IP.
- **Need in-package integration (2.5D co-packaging) to meet bandwidth requirements**

Our Mission

- Demonstrate high performance IUSG node -- codesigned to accelerate GraphBLAS
- Demonstrate modular integration of LBNL/ANL IUSG + commercial IP using Open Chiplets
- **Create new capability for the USG to rapidly assemble/prototype server-class chip designs**

Our Team

Berkeley Laboratory
John Shalf, Thom Popovici, Anastasiia Butko, Cy Chan, Patricia Gonzalez, George Michelogiannakis, Nirmal Patra

Argonne National Laboratory
Valerie Taylor, Ray Bair, Jose Monsalve Diaz, Dawson Fox

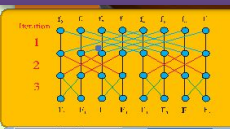
University of Chicago
Andrew Chien

Columbia University
Keren Bergman

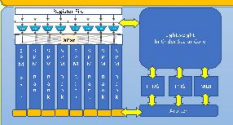
PNNL
Antonino Tumeo, Roberto Gioiosa, Jim Ang

Our Vision: Leveraging the ODSA Open Chiplets Ecosystem for Rapid Prototyping using Mixed IUSG + Commercial Chiplets

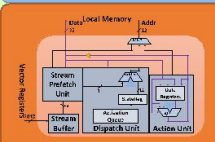
Enabling Technologies



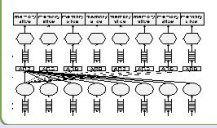
- **Fixed Function Accelerators & COTS IP *(Extreme Heterogeneity)***
  - RISC-V and ARM cores
  - Fixed function FFT (Generated by SPIRAL)



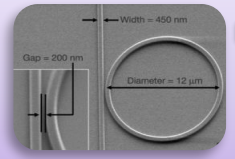
- **Word Granularity Scratchpad Memory (Gather Scatter):**
  - Gather-scatter within processor tile
  - more effective SIMD



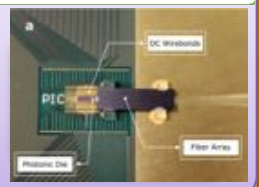
- **Recoding engine (Efficient programmable FSM & data reorg.)**
  - Sub-word granularity and high control irregularity
  - Handles branch-heavy code (avg. 20x improvement over processor core)
  - One lane is 1/100th the size of a x86 processor core



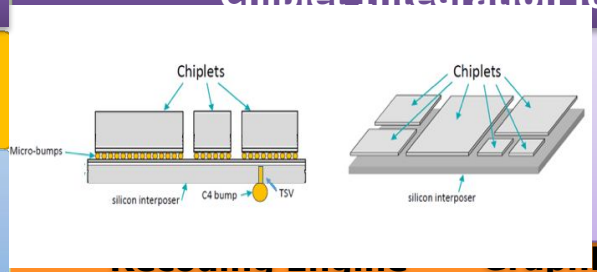
- **Hardware Message Queues (Lightweight Interprocessor Communication)**
  - Gather-scatter between processor tiles
  - Async between tiles to eliminate overhead of barriers



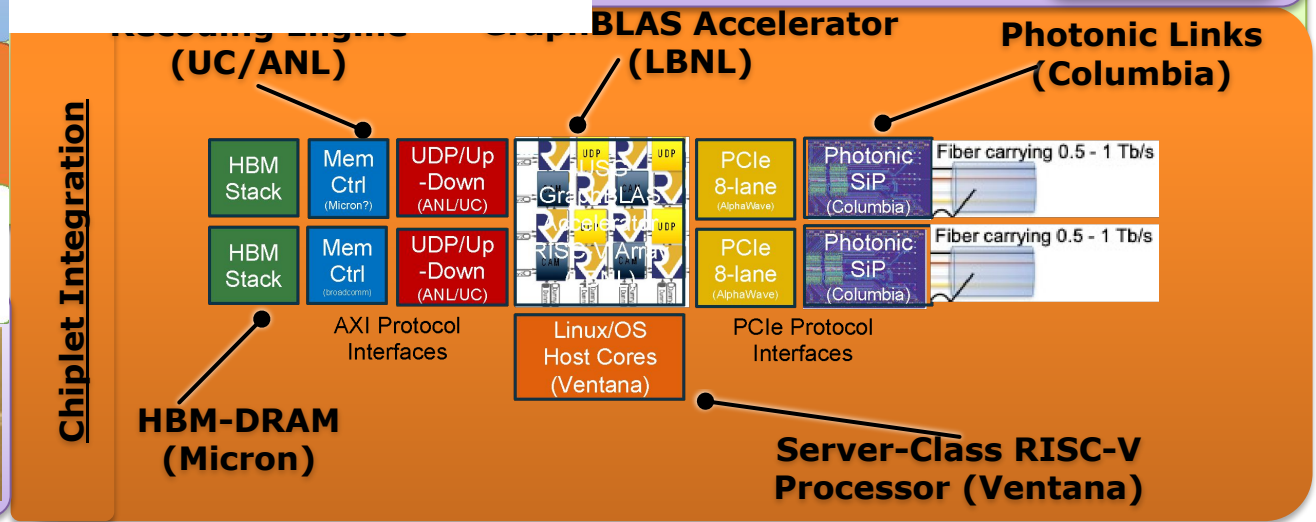
High-bandwidth, energy-efficient silicon photonic building blocks…
Compatible with CMOS microelectronics!



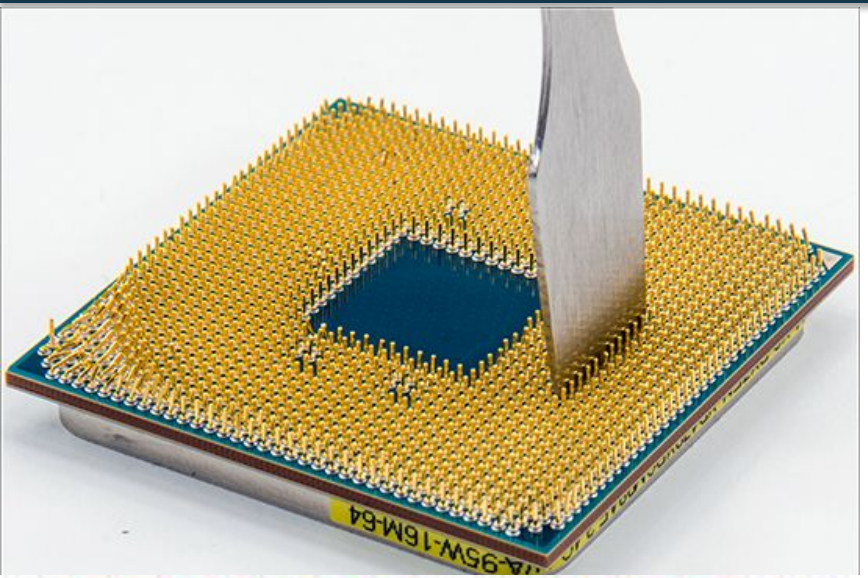
Chiplet Integration for Modularity and Scalability

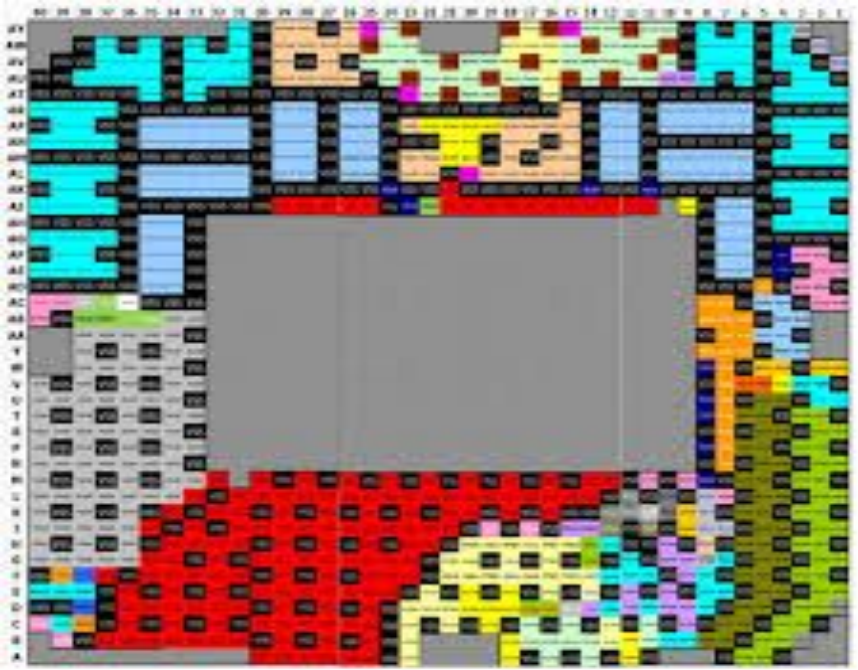


Scalable IUSG computing systems comprised of small chiplet building blocks
Sustained scalability!

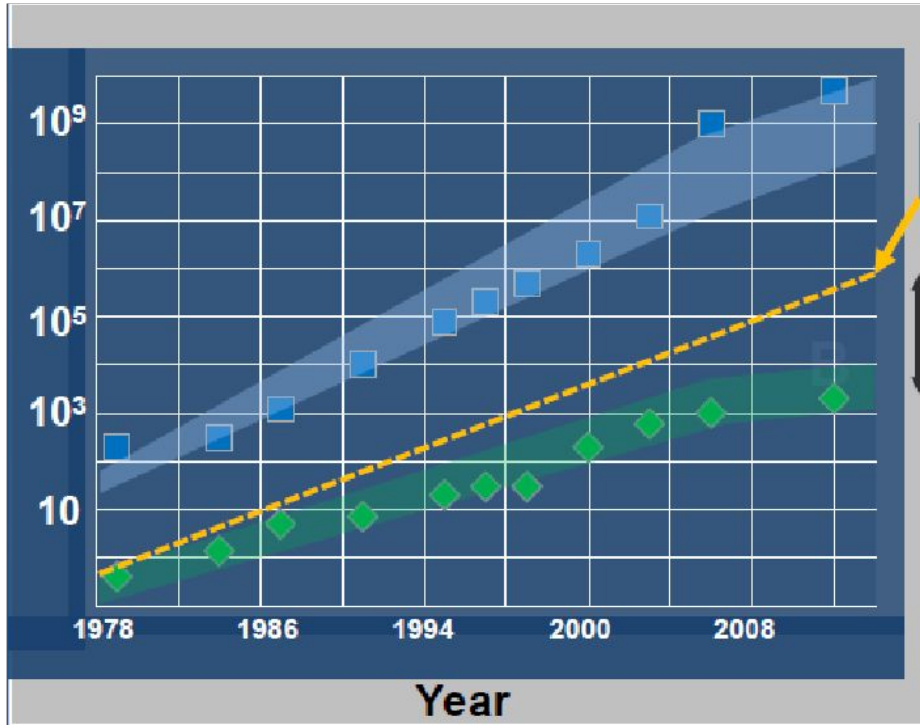


Package Performance is Pin Limited





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



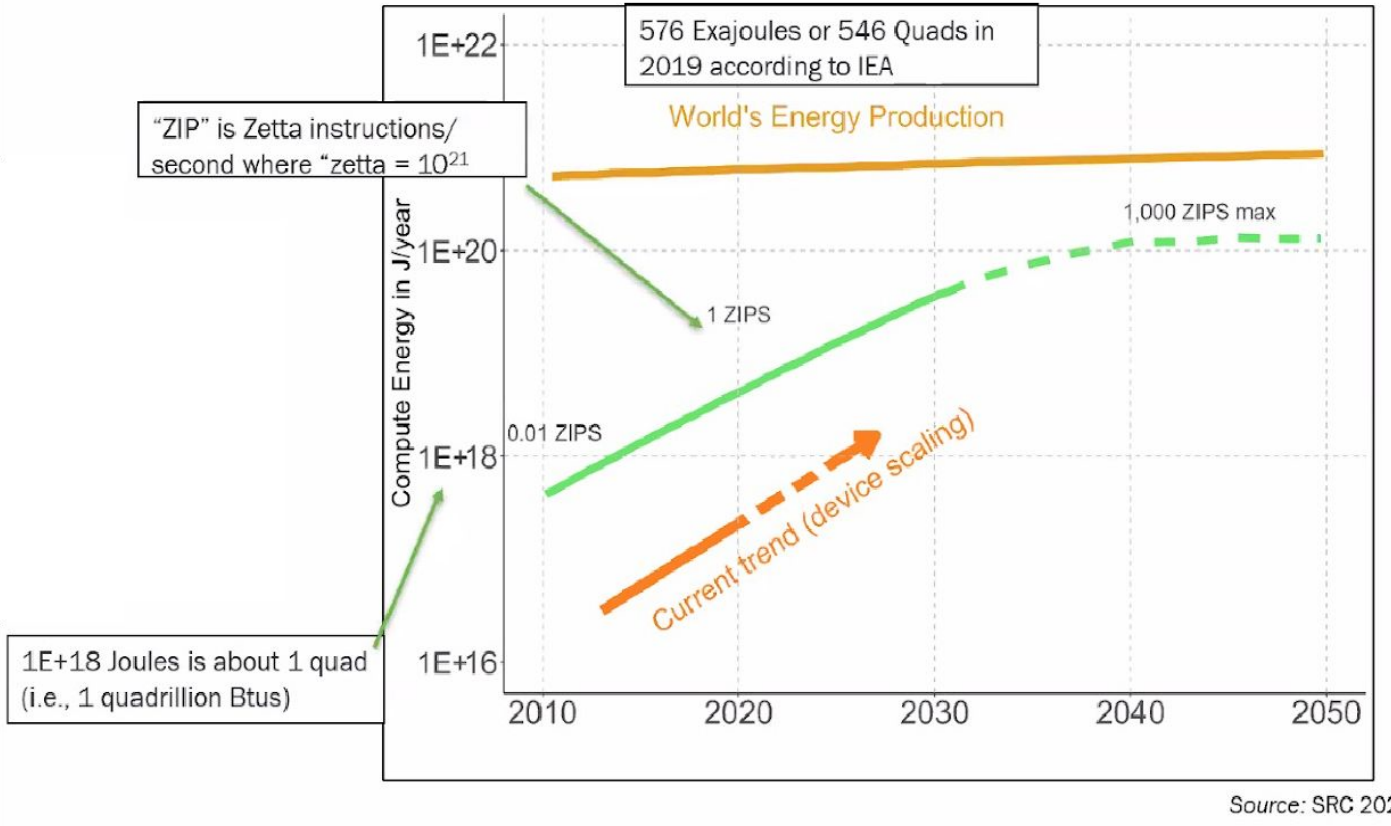
Pins x GHz from Rent's Rule

Bandwidth Gap:
~500 x and growing!

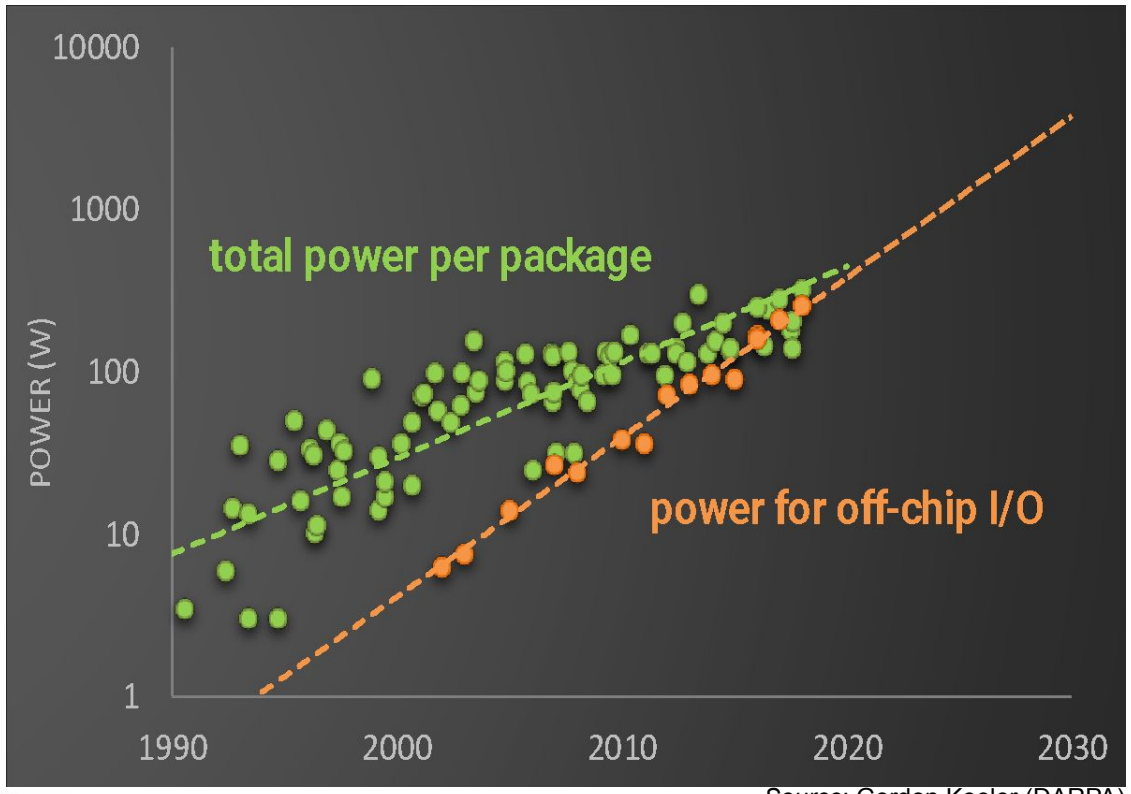
High SERDES rates run counter to end of Dennard Scaling

Source: J. Poulton, Nvidia

Datacenters: Worsten climate change without ultra-energy-efficiency And data movement dominates that power consumption



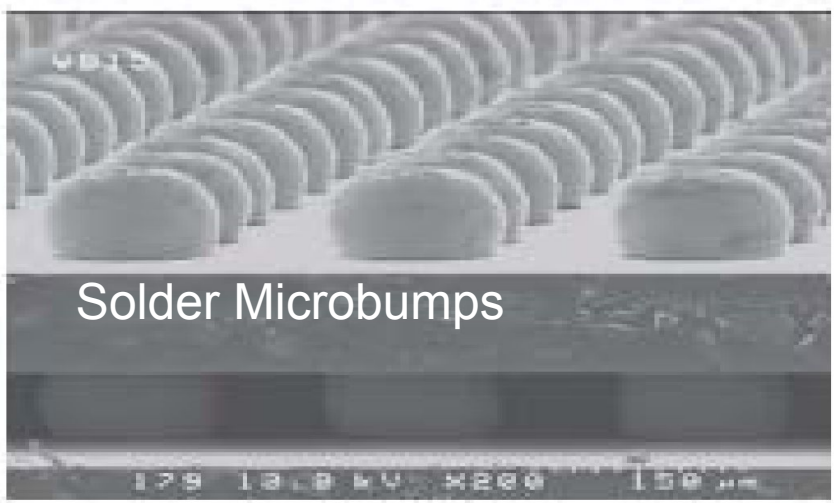
Source: SRC 2021

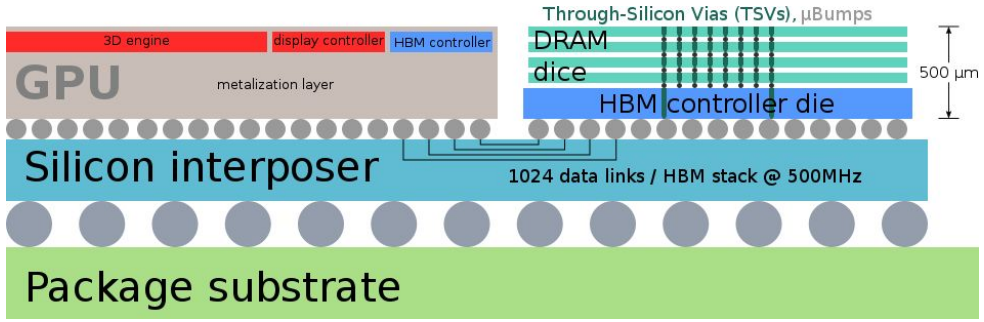


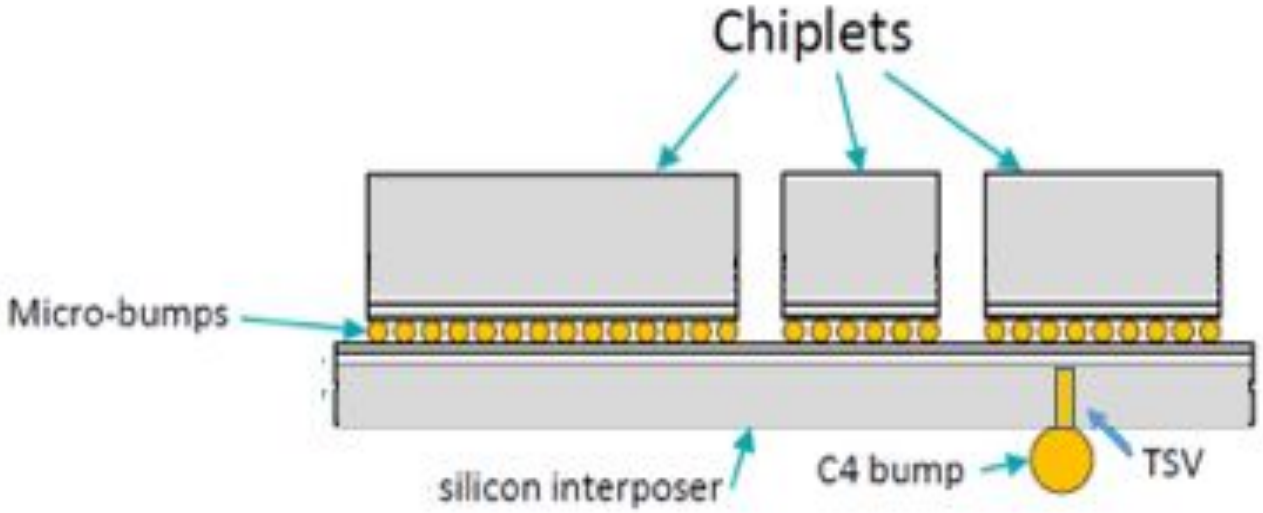
Source: Gordon Keeler (DARPA)

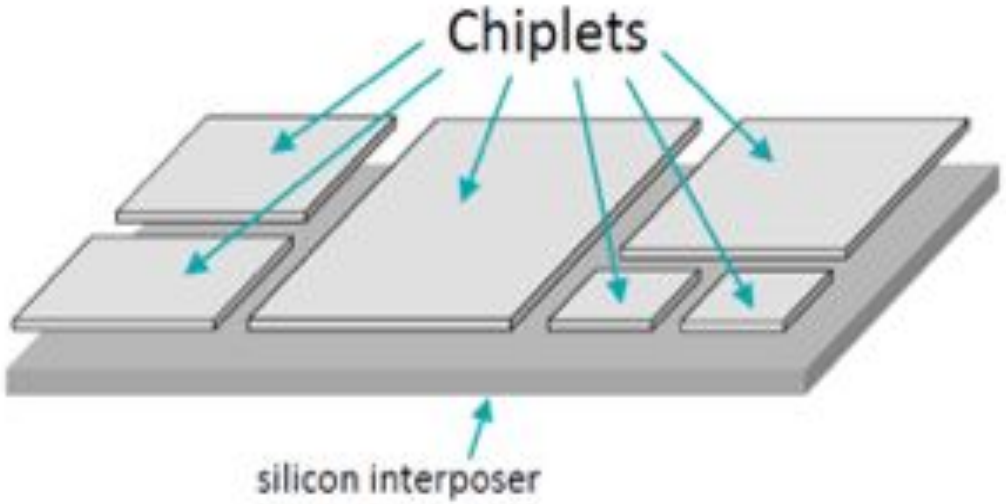
- **January 2021 SRC report projects datacenter energy growth rates will lead to ~25% consumption of planetary energy by 2040.**
- **Data movement is a dominant contributor to that power consumption**

What is a Chiplet?









Different Options