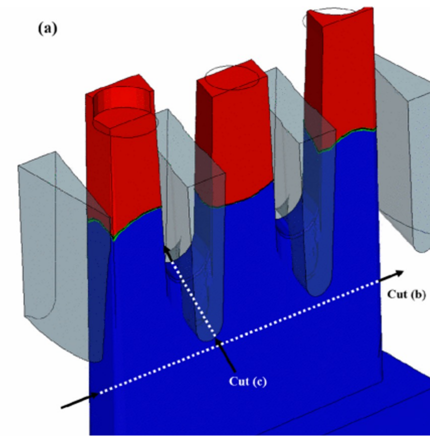
My Thoughts on Memory:

DRAM / Flash

- Highly optimized process
- 3-D already



SRAM

- Std process/planar

Memory speed and power

- Mostly set by wires not cells
- Area set by wires/contacts in planar cells

Power Generality Trade-off

- Scatter data for best worst case
- Concentrate data for best locality
  - But worst worst-case

What is Different Now?

- Cost/bit no longer king
- Advanced packaging
- Liquid cooling

First Steps (already Happening?)

- Replace large SRAM w/ DRAM
- DRAM w/ Flash (when possible)

Creating New Commodity Memories

- Locality first DRAM chiplets
- Stacked SRAM

Overcoming Memory Limitations for On-Device AI and LLM in Wearable AR Systems

Huichu Liu
Silicon Research, Reality Labs, Menlo Park, CA
Jan 10th, 2025

Meta

Custom Silicon Requirements for AR Glasses

- Low power
- High performance
- Small form factor

Memory Challenges for On-Device LLM Use Cases

Electromyography (EMG) Handwriting

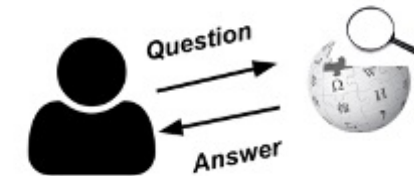
Automatic Speech Recognition (ASR)

User voice ⇒ Text

Machine Translation

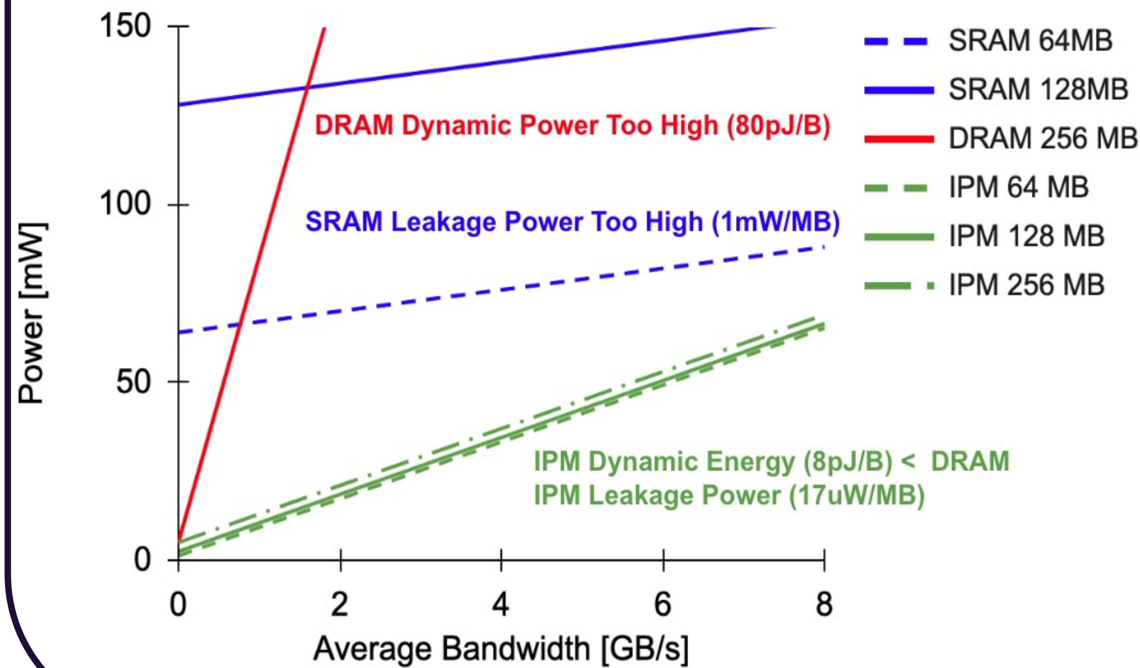
Text (language A)
hello
⇓
Text (language B)
bonjour

Open-domain Question Answering

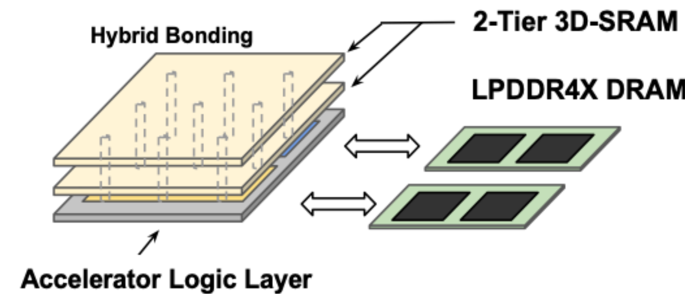


- Large memory capacity
- High memory bandwidth

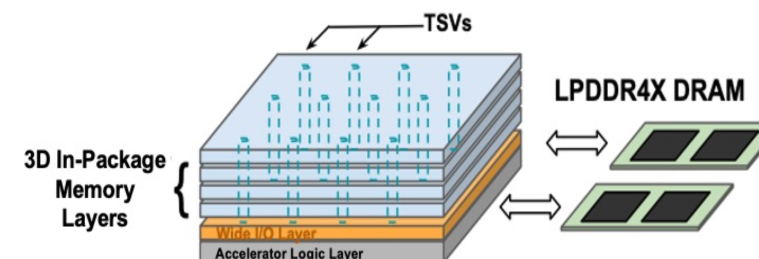
Expand Memory Capacity w/ 3D Integration, In-Package Memory (IPM) Solution & Interface Optimization



3D-Stacked SRAM Architecture



3D-IPM Architecture



- ✓ Integrate logic and memory with 3D technologies
- ✓ Co-optimize memory for workload characteristics
- ✓ Design for end-to-end systems

Meta

EMD Electronics is the electronics business of Merck KGaA, Darmstadt, Germany in the U.S. and Canada.

Inference-RAM

A novel multi-decadal memory category

Addressing the exponential growing demand for AI inference acceleration,
currently stifled by power hungry on-chip communication and off-chip DRAMS

Shridhar Mukund, Chief Systems Architect

January 1st, 2025

In collaboration with:

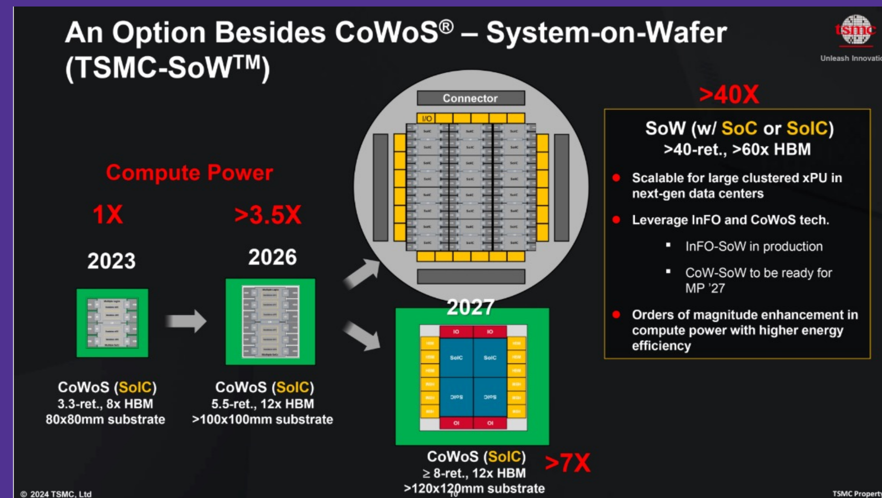
Stanford Differentiated Access Memories Project

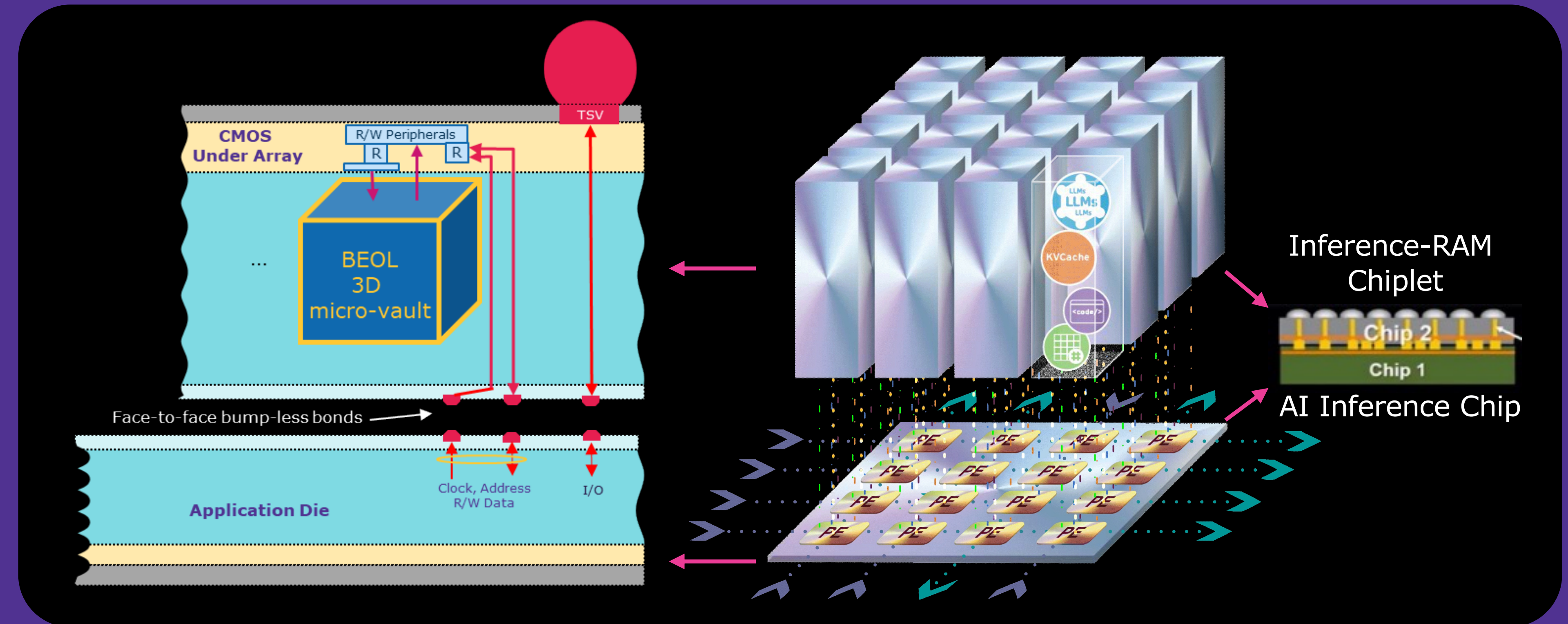
https://MemoryDAX.Stanford.edu

EMD Electronics

Bringing On-Chip SRAM like read bandwidths at DRAM densities

Inference-RAM, A Read-Optimized Differentiated Access Long-Term Memory





Inference-RAM Chiplet houses long-term intelligence,
which is fully determined at compile-time and is largely constant at run-time

- Over-provisioned model parameters,
- Over-provisioned KV caches,
- Network-on-Chip (NoC) route tables
- Program codes, activation function tables, …

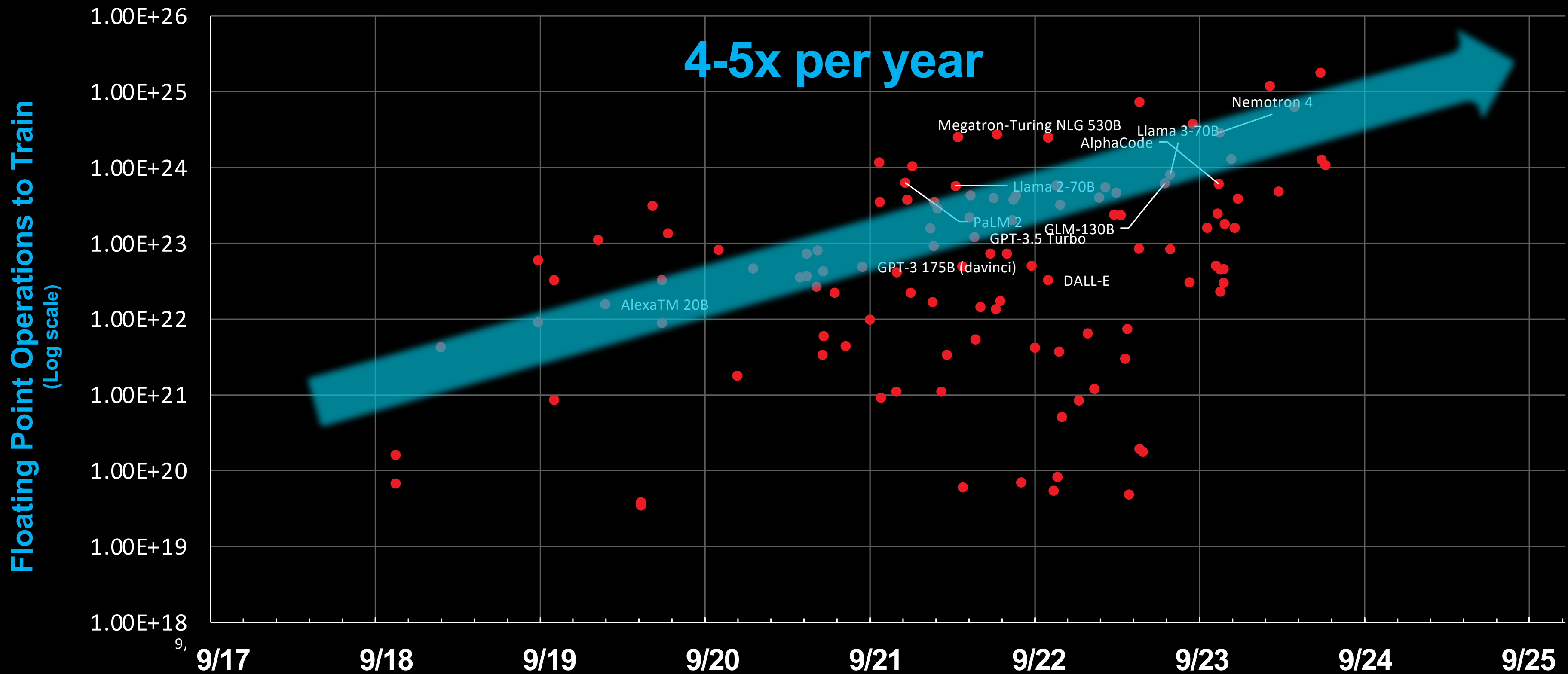
Targeting: >100x advantage in energy-delay product at <10x smaller silicon foot-print

DAM Workshop Panel

Ralph Wittig
Corporate Fellow
Head of AMD Research + Advanced Development

January 10 , 2025

AI is Driving Massive Compute Demand

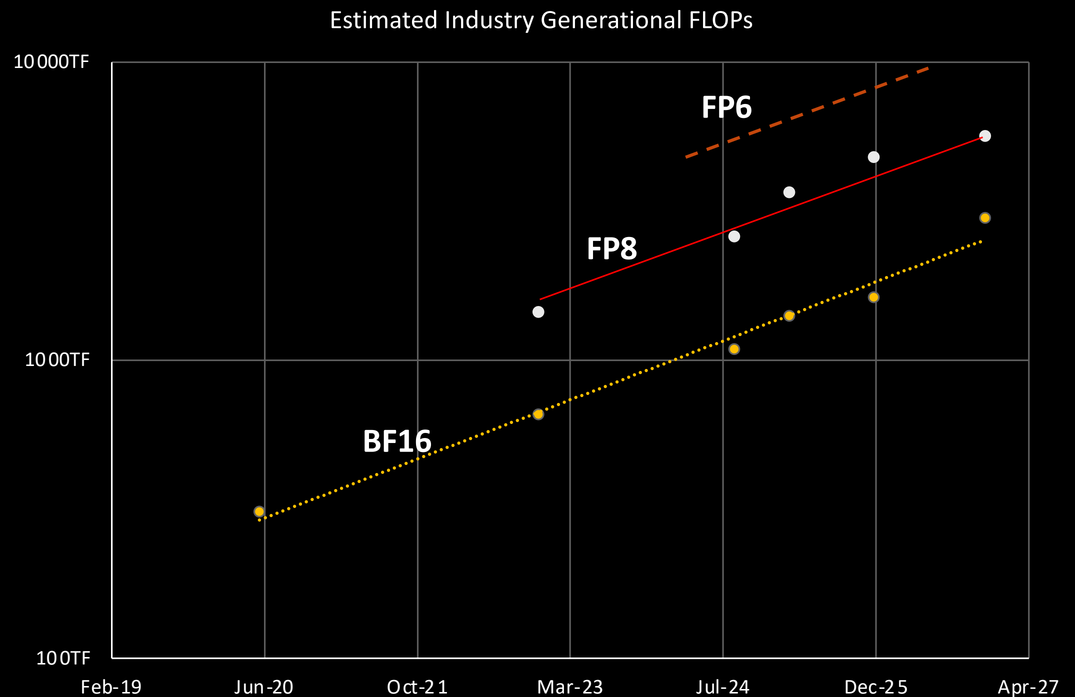


Jaime Sevilla and Edu Roldán, "Training Compute of Frontier AI Models Grows by 4-5x per Year," Epoch AI, May 28, 2024. [Online]. Available: https://epochai.org/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year

FLOP Trends and Requirements

- FLOPs increasing ~2X/2 years
- Dedicated matrix-math datapaths
- AI FLOPs: Reduced precision formats
- With AI FLOPs, get ~2x/1.3 years

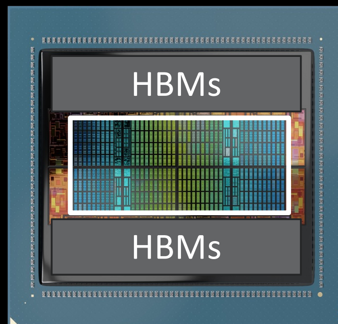
- *Architectural advancements complement technology advancements*

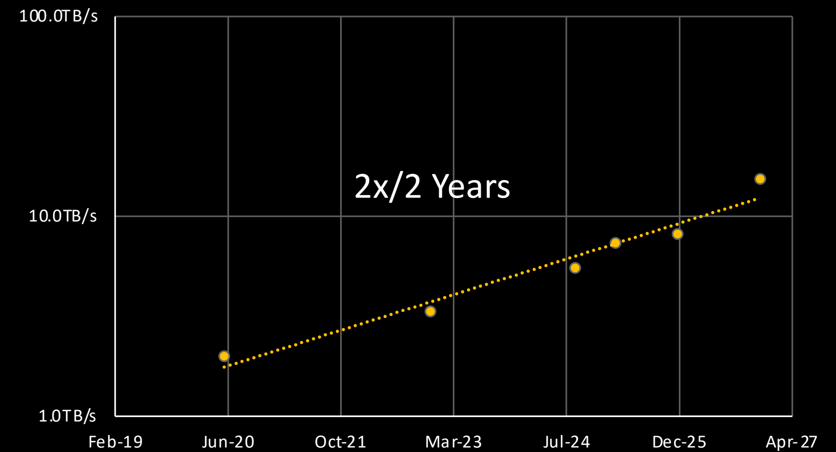


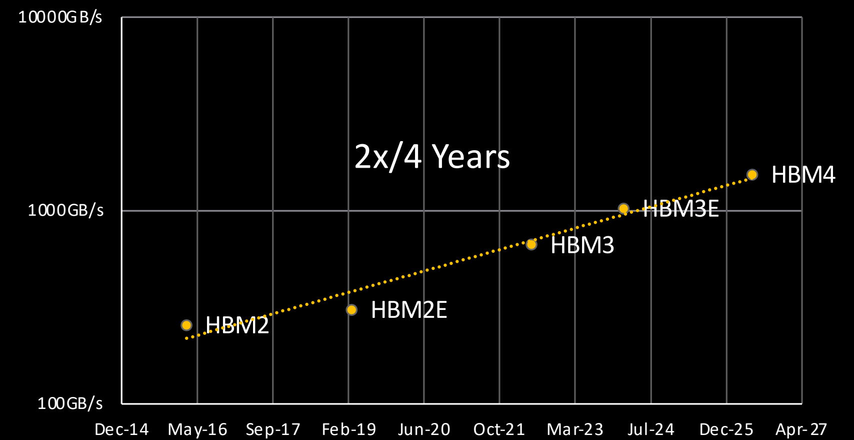
AMD

Memory Bandwidth

- Memory Bandwidth must also double every ~2 years to maintain a consistent bytes/FLOP ratio
- HBM bandwidth doubling only every ~4 years
  - Power per stack has been increasing
- To keep up with demand, HBM stacks per GPU must increase driving ever-larger modules
- *We must find ways to reduce energy/bit*





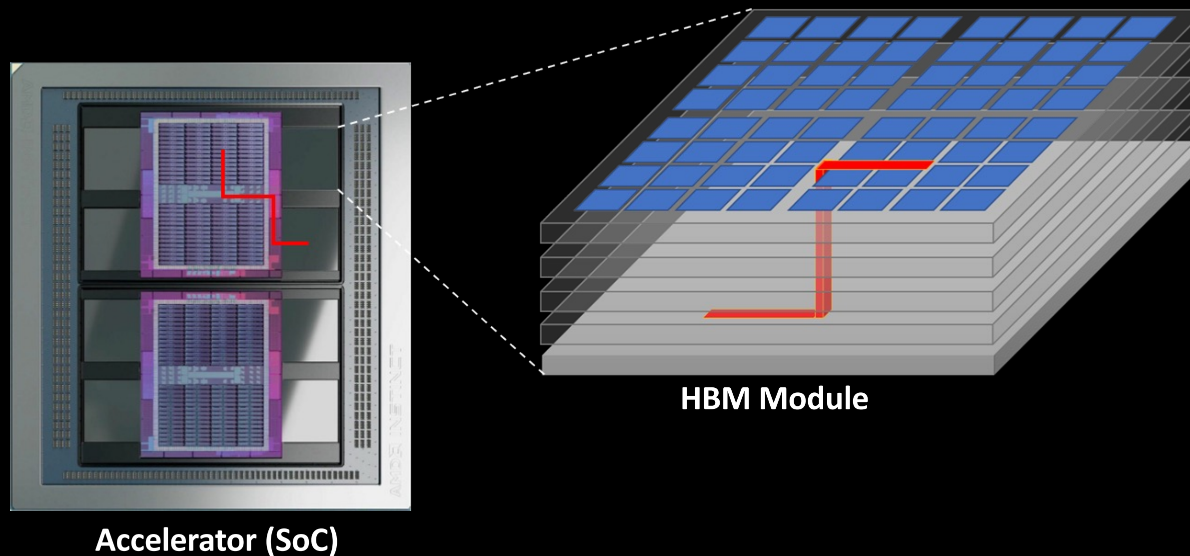


AMD

Datacenter Memory

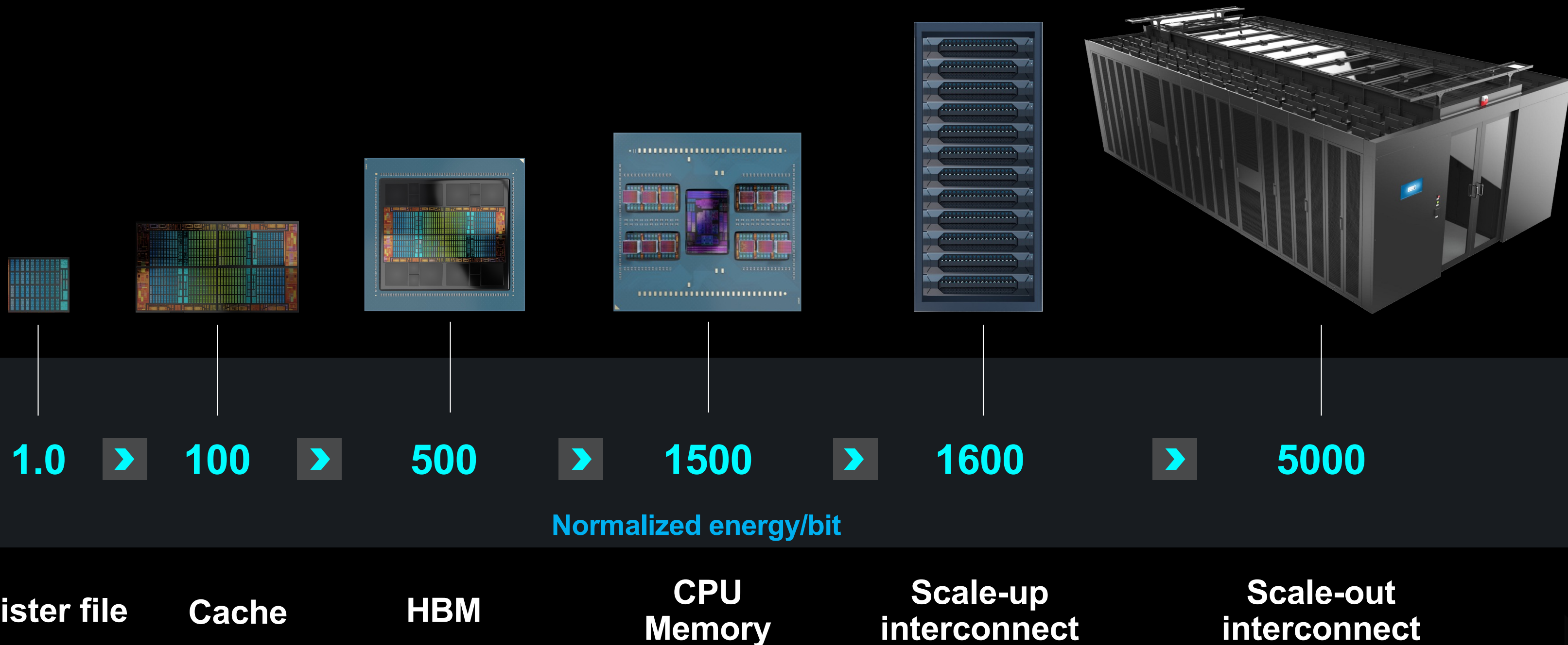
- 2.5D memory (HBM) is the norm
  - Expensive, but maximizes TCO

- Reaching the limits of current HBM organization with centralized TSVs
  - As much as 90% of HBM power can be (largely horizontal) data movement



AMD

Reducing Data Movement Energy

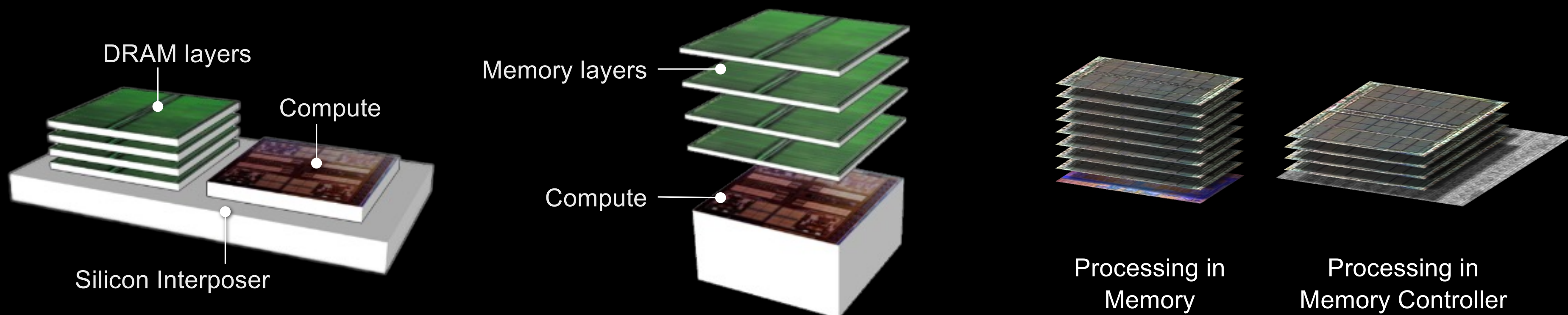


Communication energy grows exponentially with distance

Maximizing locality is key to efficiency

AMD

Even Tighter Integration of Compute and Memory



Higher Levels of Integration Enables Higher Bandwidth at Lower Power

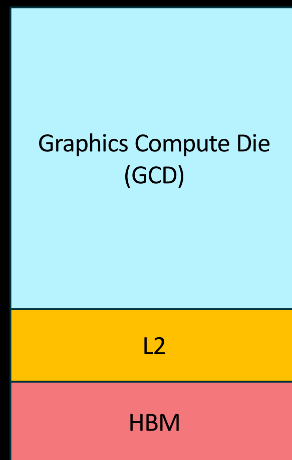
	On Board Memory	2.5D Micro-bumps (HBM)	3D Hybrid Bond
pJ/bit	~12	~3.5	~0.2

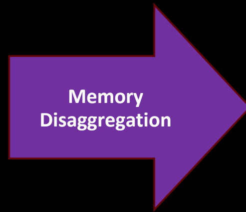
Invest in scaling new logic-memory architectures

Image source: https://commons.wikimedia.org/wiki/File:SDRAM-Modul.jpg, Creative Commons 4.0.

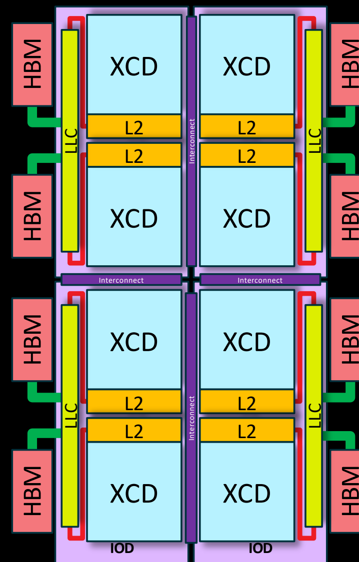
Continual Disaggregation

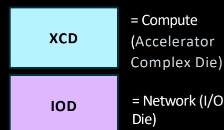
MI200

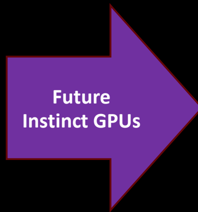




MI300







Aggregated Memory
No L2 Cache NUMA
No HBM NUMA

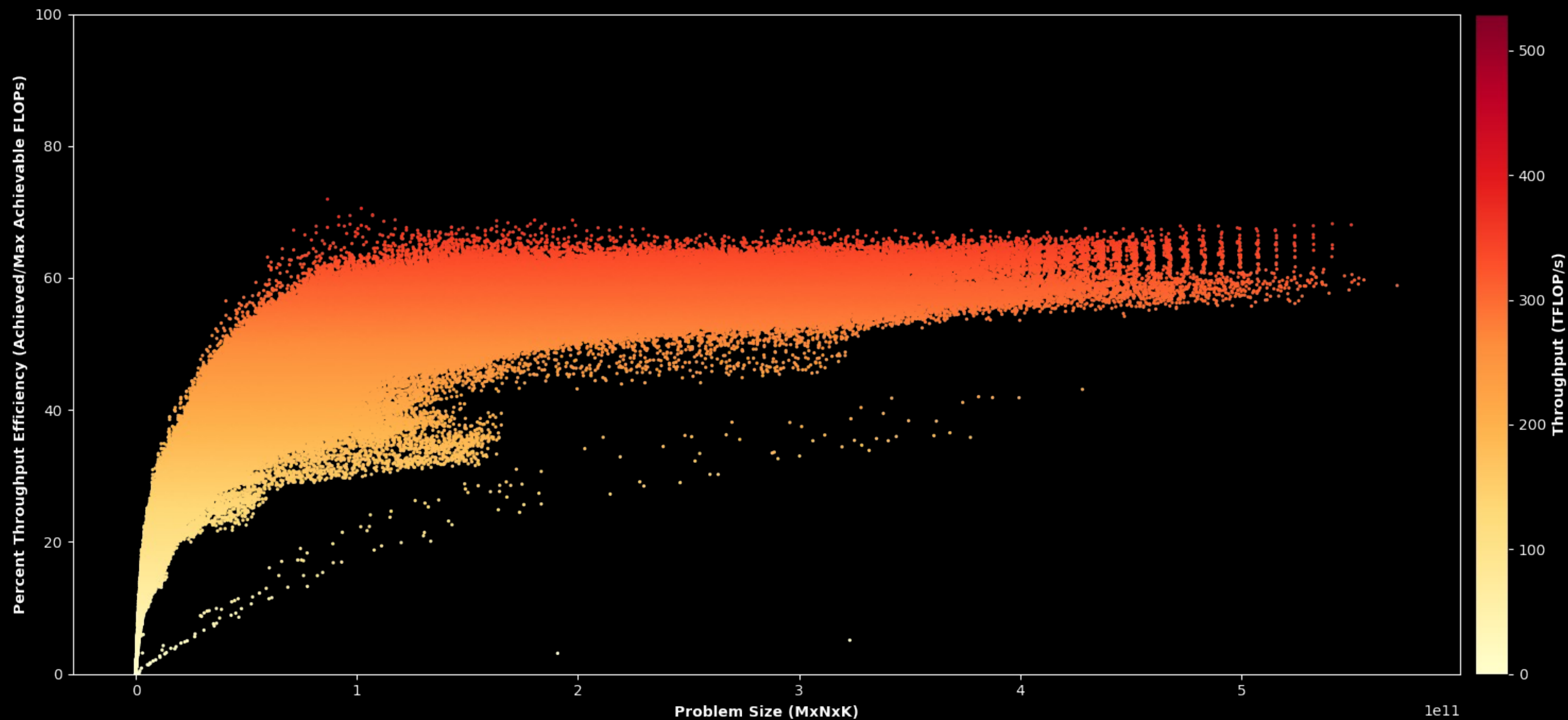
Disaggregated Memory
L2 Cache NUMA
HBM Stack NUMA

Disaggregated Memory
L2 Cache NUMA
DRAM Stack NUMA
Position within XCD NUMA
…

Architectural NUMA effects are inevitable - our algorithms and programming models must evolve to effectively program them

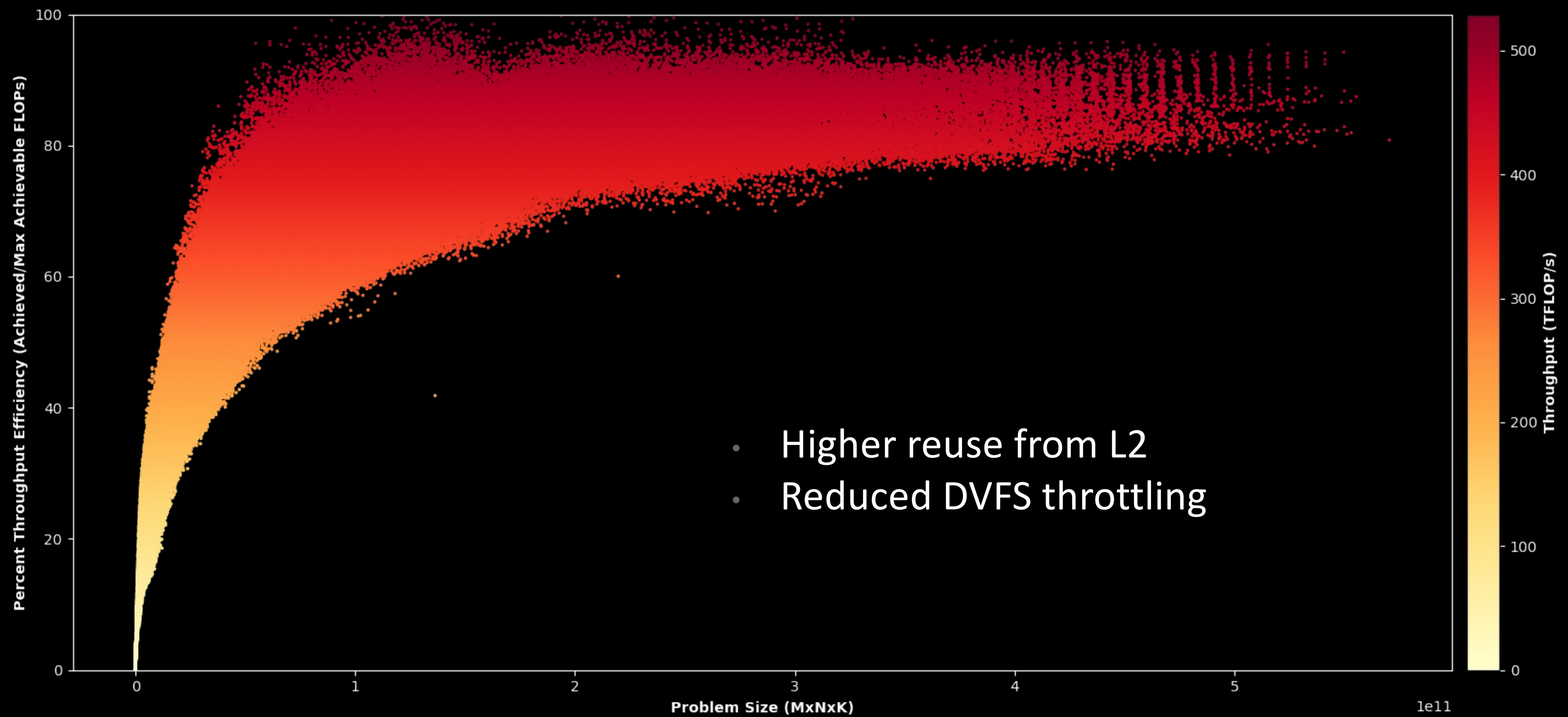
AMD

GEMM Performance with Random Spatially-Unaware Dispatch



AMD

Better Performance with Spatially Aware Dispatch



AMD

Meeting the Challenge Requires Holistic Innovation

- Advanced packaging
- New interconnects and memory
- System level integration
- Spatial computing architectures
- NUMA aware programming models
- Algorithm-software-hardware co-design

AMD