Advancing Energy Efficient AI Communication

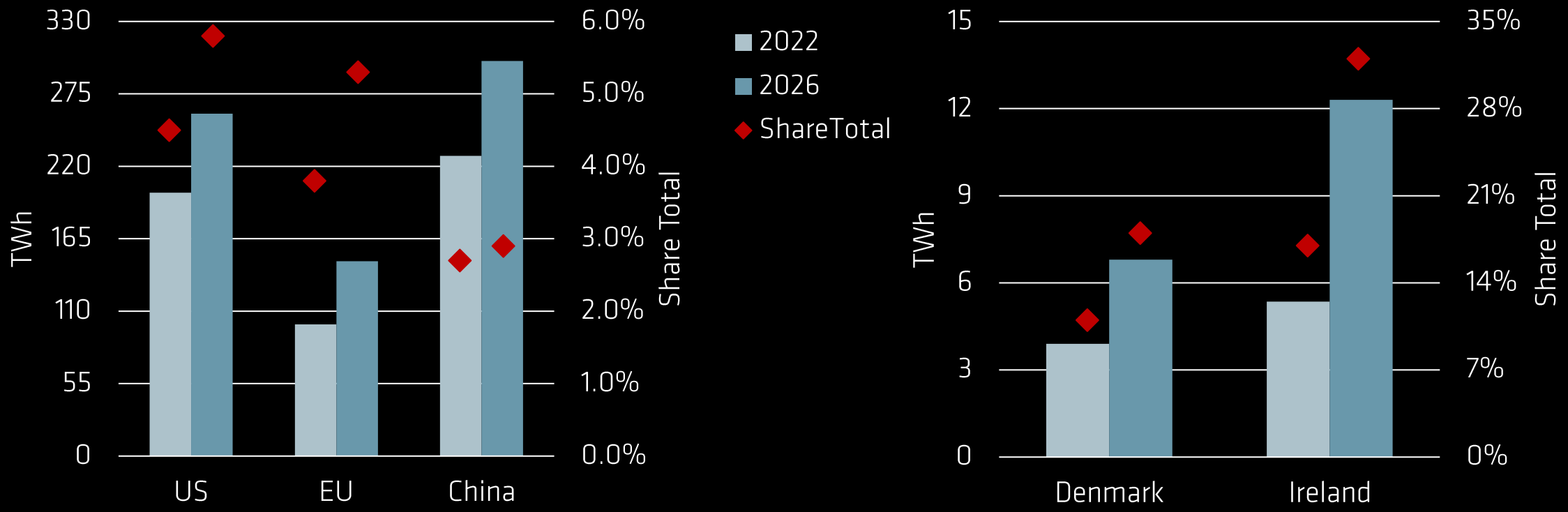
Brad Beckmann (brad.beckmann@amd.com)
Fellow
AMD Research + Advanced Development

February 14, 2025

AMD
together we advance_

The Challenge

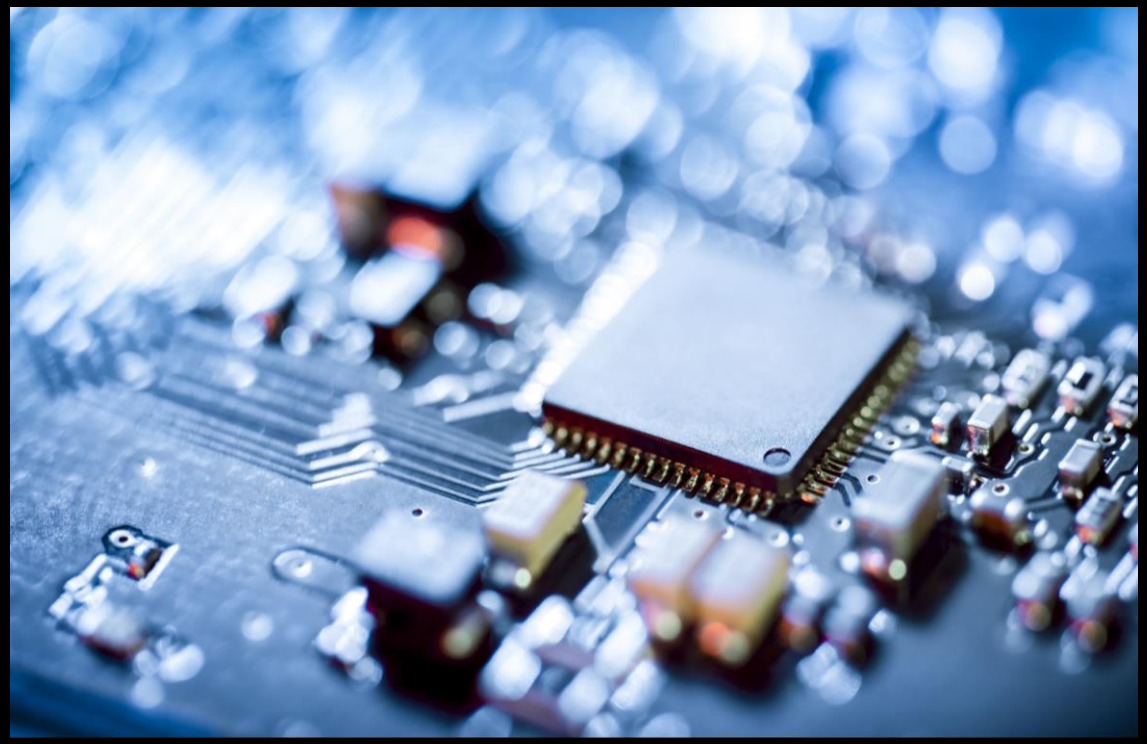
Data Center Power Consumption



Data source: IEA (2024), Electricity 2024, IEA, Paris https://www.iea.org/reports/electricity-2024, License: CC BY 4.0

How We Got Here – 5 Decades of Innovations

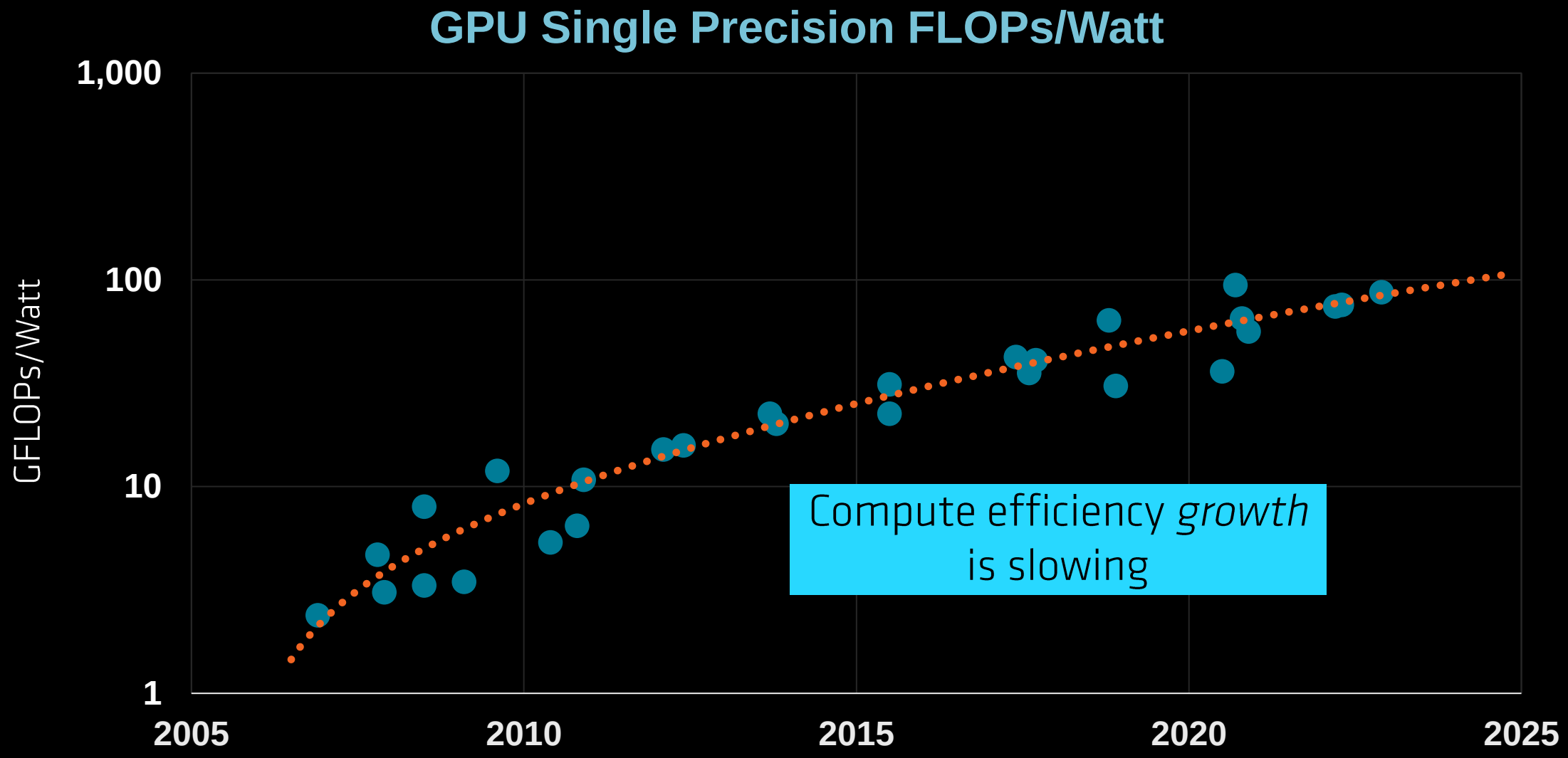
> 2000 times faster



> 3000 times smaller

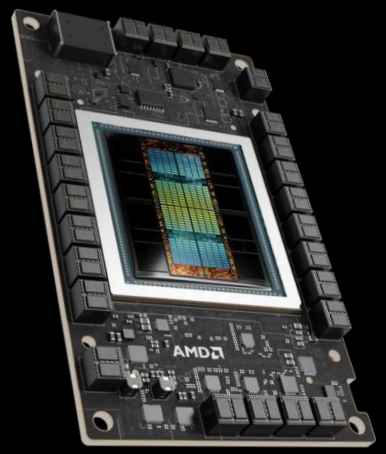
Analysis based on comparing an AMD AM9080 processor running at 2MHz to an AMD Ryzen 9 7950X with base clock frequency of 4.5GHz

Trends – GPU Power Efficiency



Instinct™ MI300 ACCELERATOR OVERVIEW

Architected to deliver maximum HPC and AI capabilities from the latest silicon and advanced packaging technology



304 CDNA 3 CUs

256 GB HBM3E Memory

6 TB/s Memory Bandwidth

AMD Instinct™ MI325X

CPU hosted PCIe® accelerator

228 CDNA 3 CUs | 24 "Zen 4" CPU cores

128 GB HBM3 Memory

5.3 TB/s Memory Bandwidth

AMD Instinct™ MI300A

Self-hosted accelerated processing unit (APU)

AMD

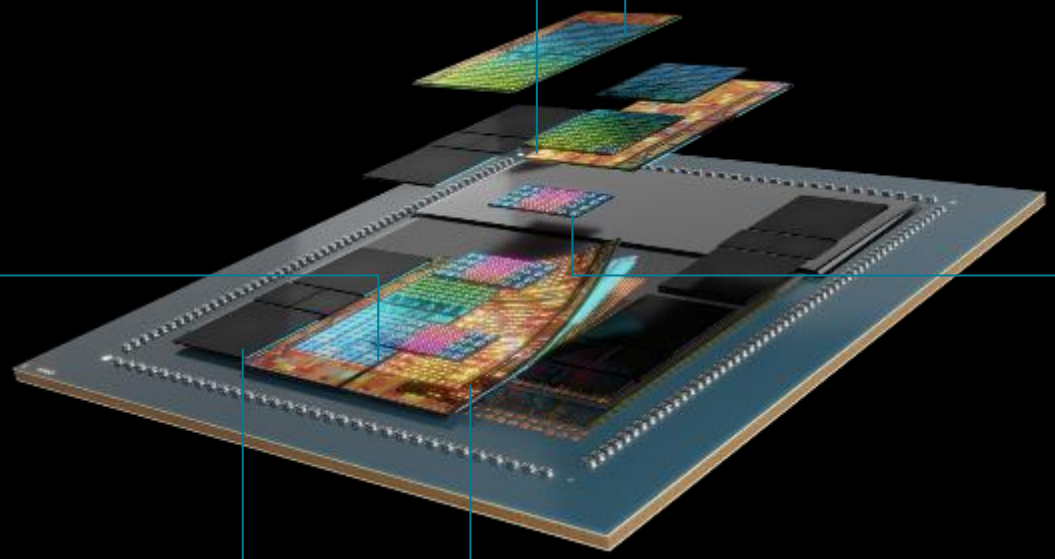
AMD Instinct™ MI300 Series Accelerators

Key Innovations

I/O Die (IOD)
256MB AMD Infinity Cache™
Infinity Fabric Network-on-Chip

AMD Infinity Fabric™ AP Interconnect

8/12 stacks of HBM3/3E
MI300A: 128 GB (8H)
MI300X: 192 GB (12H)
MI325X: 256 GB (12H)



Accelerator Complex Die (XCD)
6 x 38 AMD CDNA™ 3 Compute die

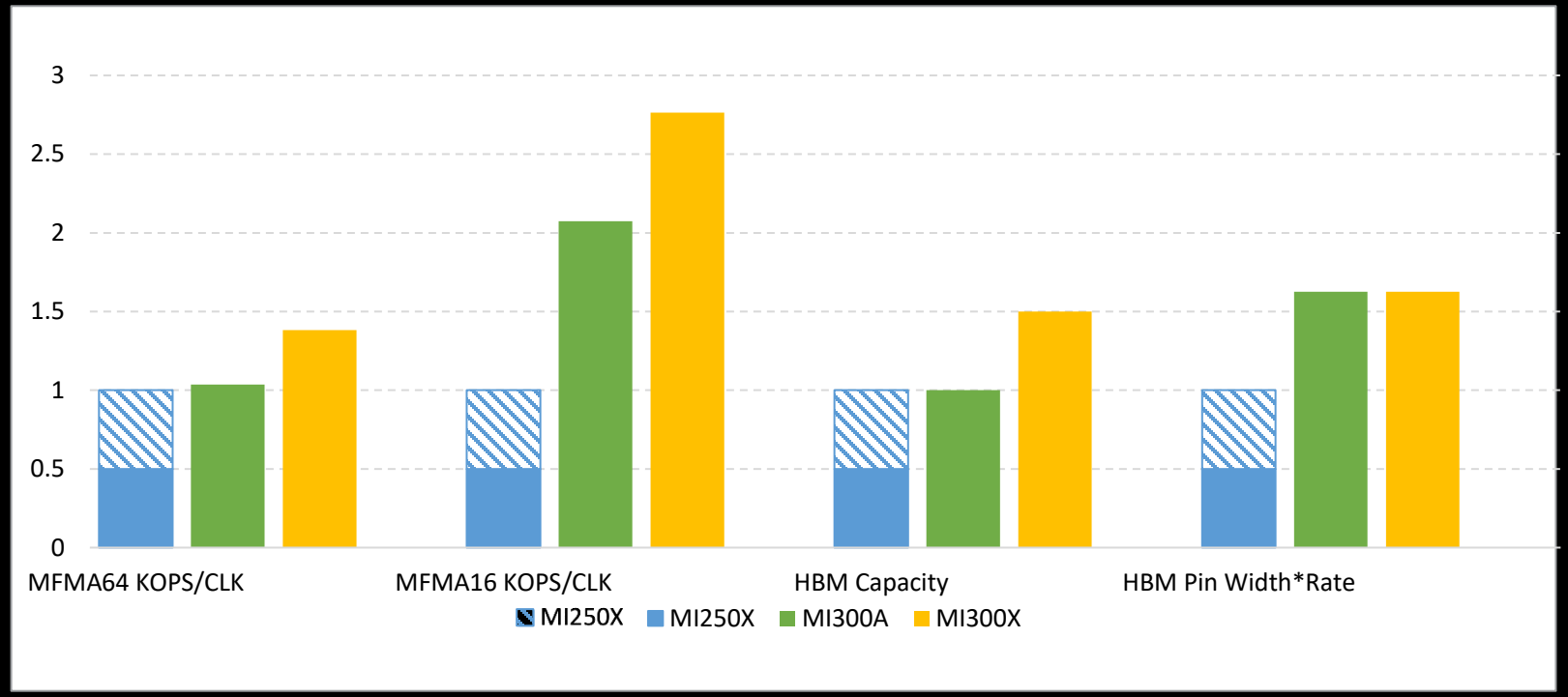
CPU Complex Die (CCD)
3 x 8 “Zen 4” Cores

3.5D Package
3D hybrid bonding
2.5D silicon interposer

Generational Scaling

MI300 Enhancements

- Chiplet technologies
- Advanced packaging
- 5nm and 6nm process nodes
- Single GPU accelerator device
- Unified 8-stack HBM3 memory

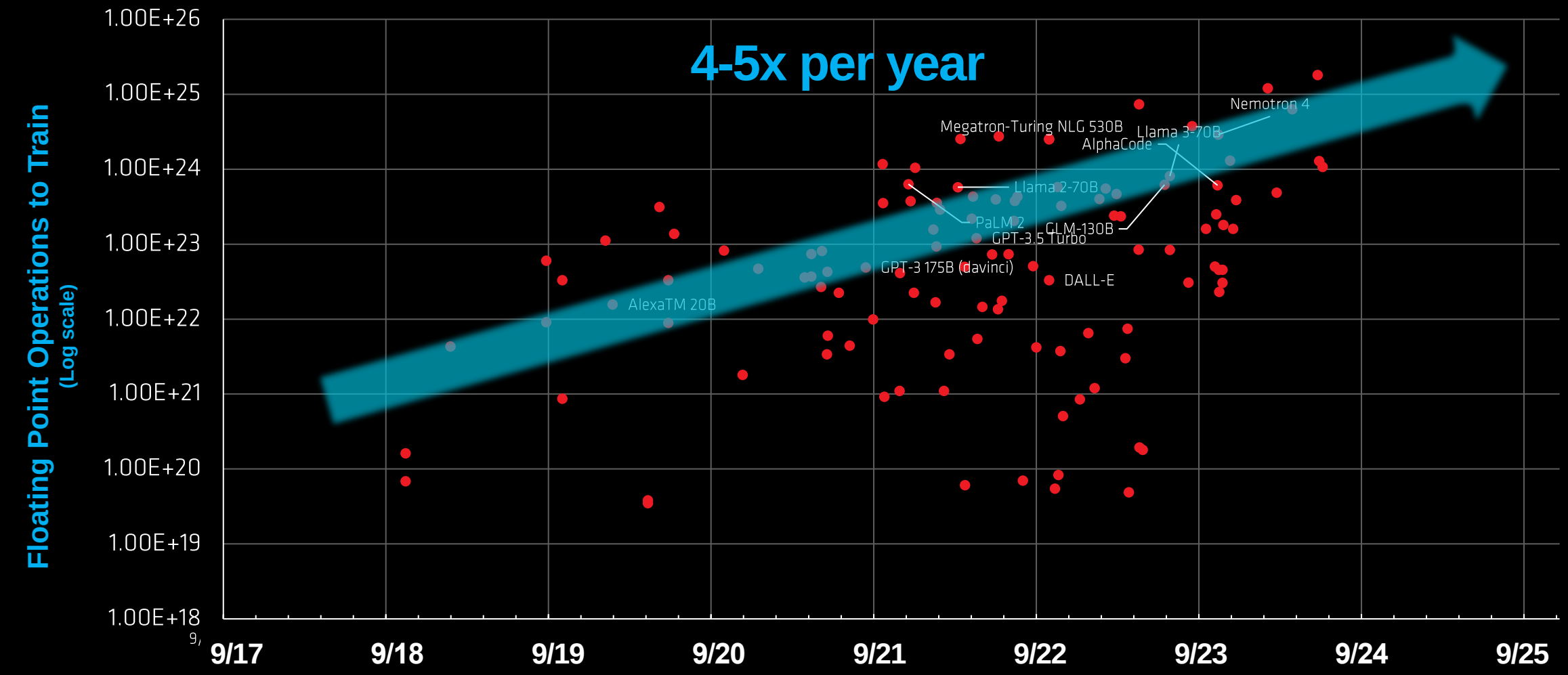


MI300X improvements over MI250X

- 2.75X matrix FMA FP16 OPS/CLK
- 1.5X HBM capacity and peak BW
- 1.2X higher peak engine clocks

Product	AMD Instinct™ MI250X	AMD Instinct™ MI300A	AMD Instinct™ MI300X
GPU Architecture	AMD CDNA™ 2	AMD CDNA™ 3	AMD CDNA™ 3
Lithography	TSMC 6nm FinFET	TSMC 6nm, TSMC 5nm	TSMC 6nm, TSMC 5nm
Power	560W	760W	750W
Peak Engine Clock	1700 MHz	2100 MHz	2100 MHz
Peak DP (FP64) Performance	47.9 TFLOPS	61.3 TFLOPS	81.7 TFLOPS
Peak DP Matrix (FP64) Performance	95.7 TFLOPS	122.6 TFLOPS	163.4 TFLOPS
Peak bfloat16 Matrix Performance	383 TFLOPS	980.6 FLOPS	1307.4 FLOPS
Memory Type	HBM2e	HBM3	HBM3
Memory Clock	1.6 GHz	2.6 GHz	2.6 GHz
Memory Interface	8192-bit	8192-bit	8192-bit
Peak Memory Bandwidth	3276.8 GB/sec	5324.8 GB/sec	5324.8 GB/sec

AI is Driving Massive Compute Demand

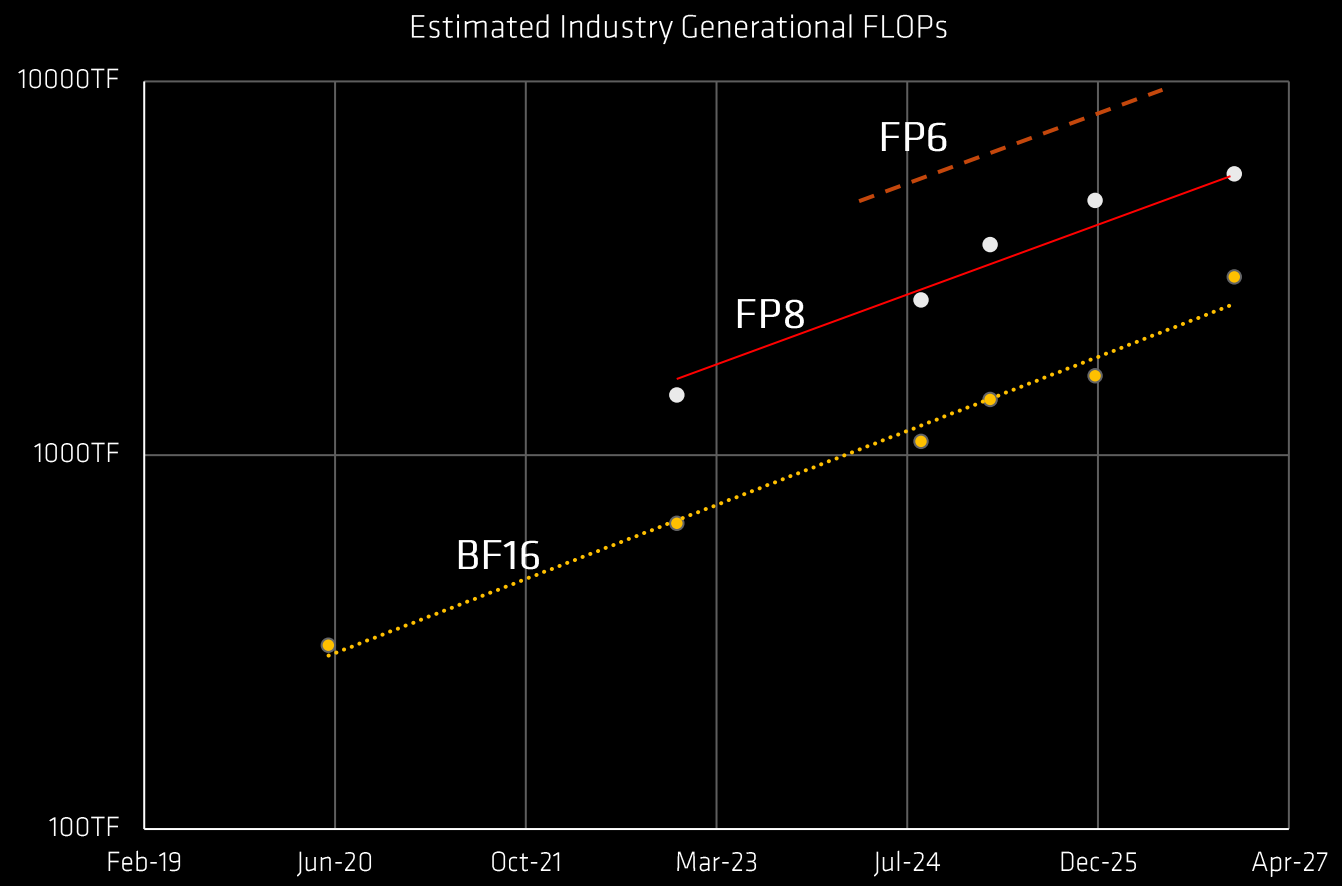


Jaime Sevilla and Edu Roldán, "Training Compute of Frontier AI Models Grows by 4-5x per Year," Epoch AI, May 28, 2024. [Online]. Available: https://epochai.org/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year

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

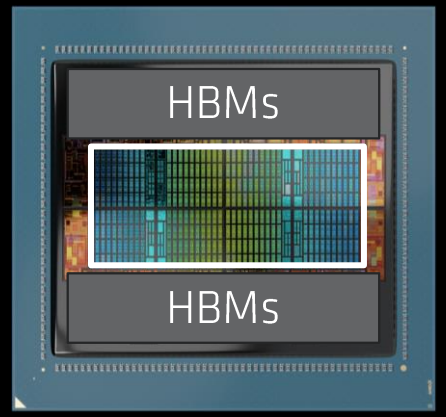
- *Architectural advancements complement technology advancements*

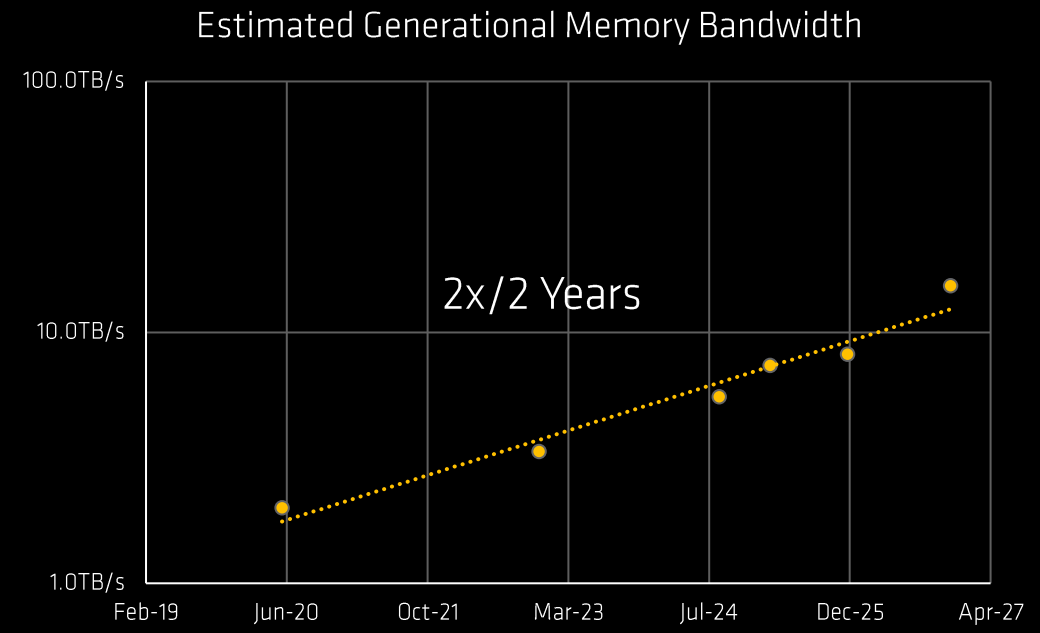


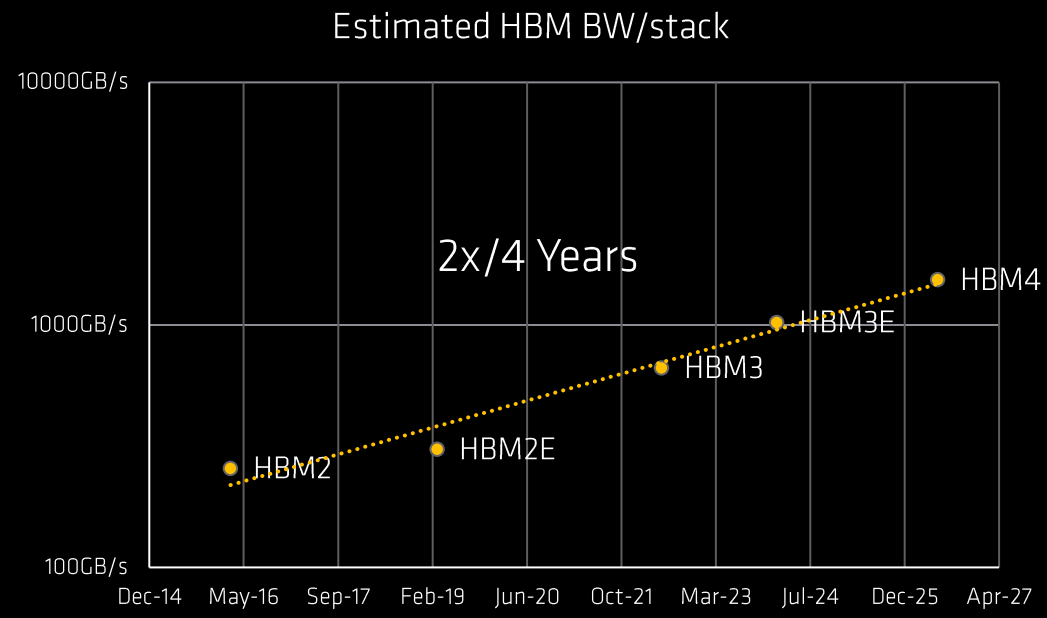
AMD

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*



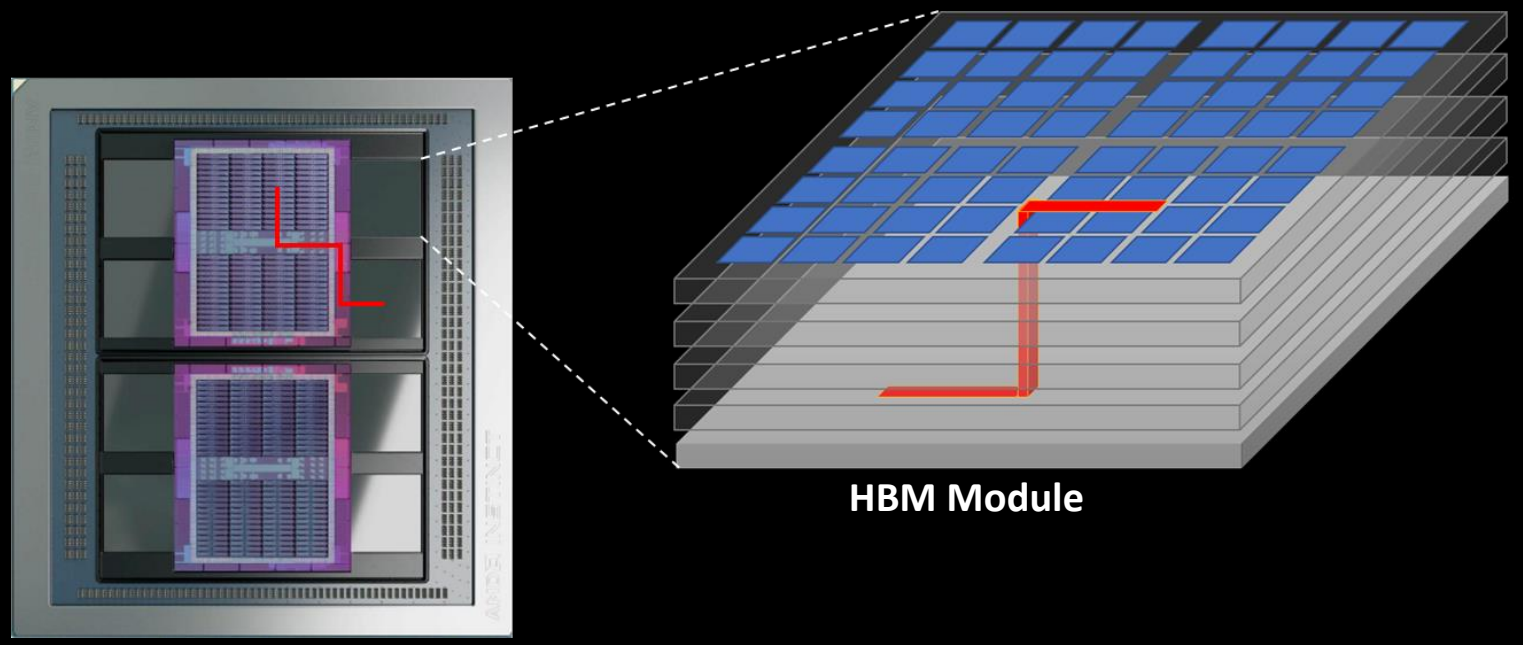




Datacenter Memory

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

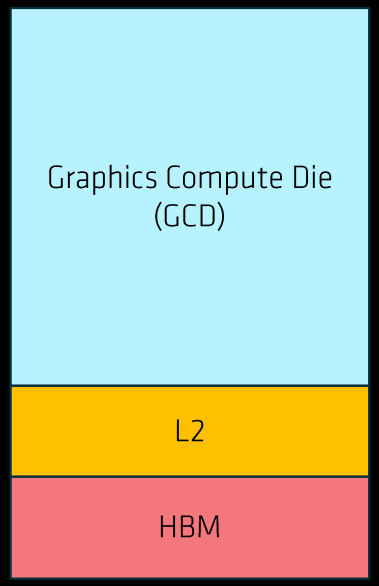
- Reaching the limits of current HBM organization with centralized TSVs
  - As much as 90% of HBM power can be (largely horizontal) data movement



The Opportunity

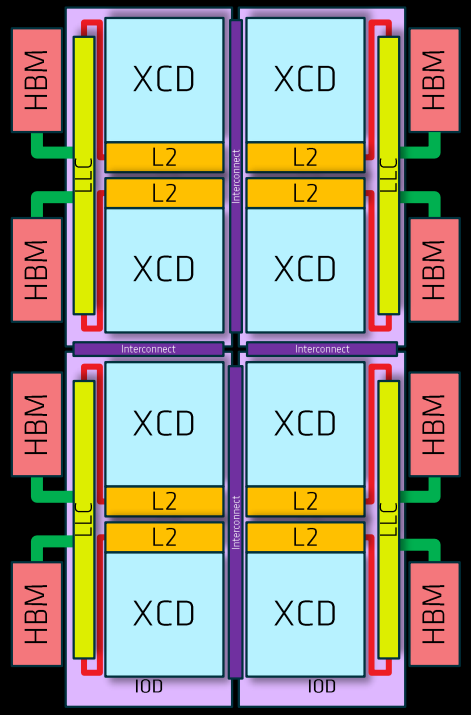
Continual Disaggregation

MI200



Memory Disaggregation

MI300



XCD = Compute (Accelerator Complex Die)

IOD = Network (I/O Die)

Future Instinct GPUs

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

Aggregated Memory
No L2 Cache NUMA
No HBM NUMA

Disaggregated Memory
L2 Cache NUMA
HBM Stack NUMA

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

MI300X – Compute and Memory Partitioning

Compute Partitioning

CPX: 8 Partitions => Each XCD (chiplet) exposed as separate GPU to software

Memory Partitioning

NPS4: 4 Partitions => One partition per IOD (2 HBM channels)

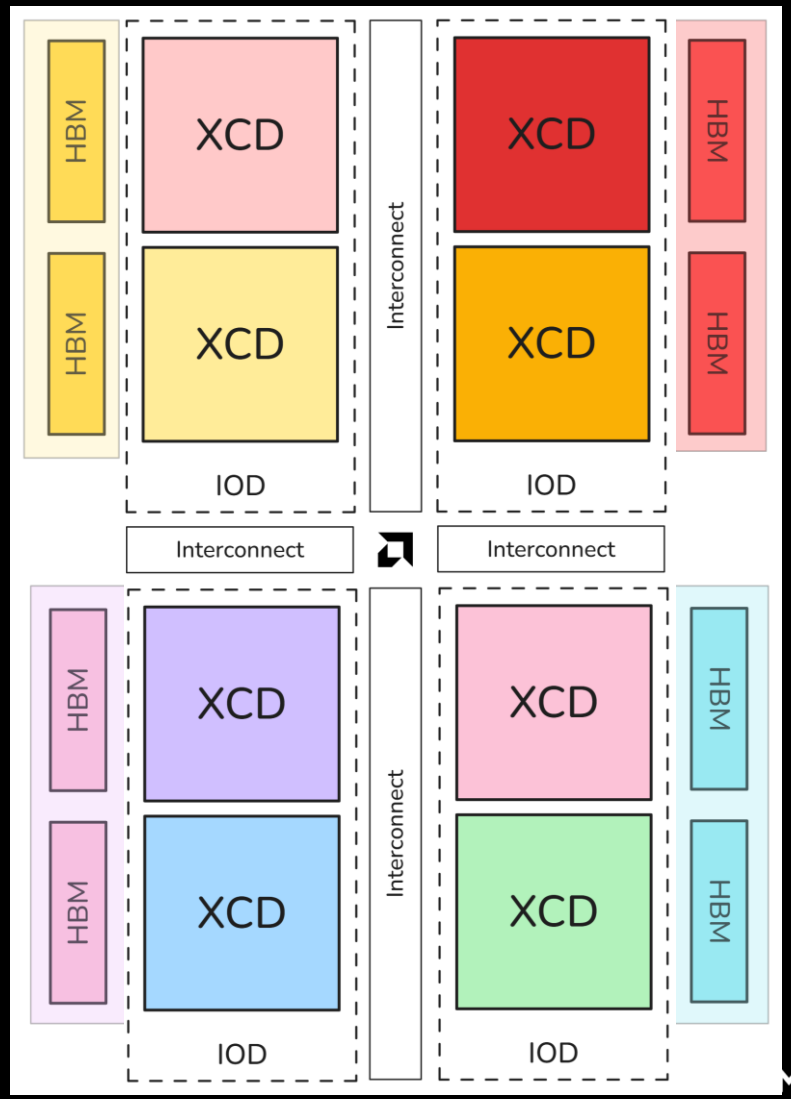
Benefits

- Reduced kernel launch times
- Reducing cross IOD data movement can reduce power consumption => more power available for XCD (compute)

More Details (including performance numbers)

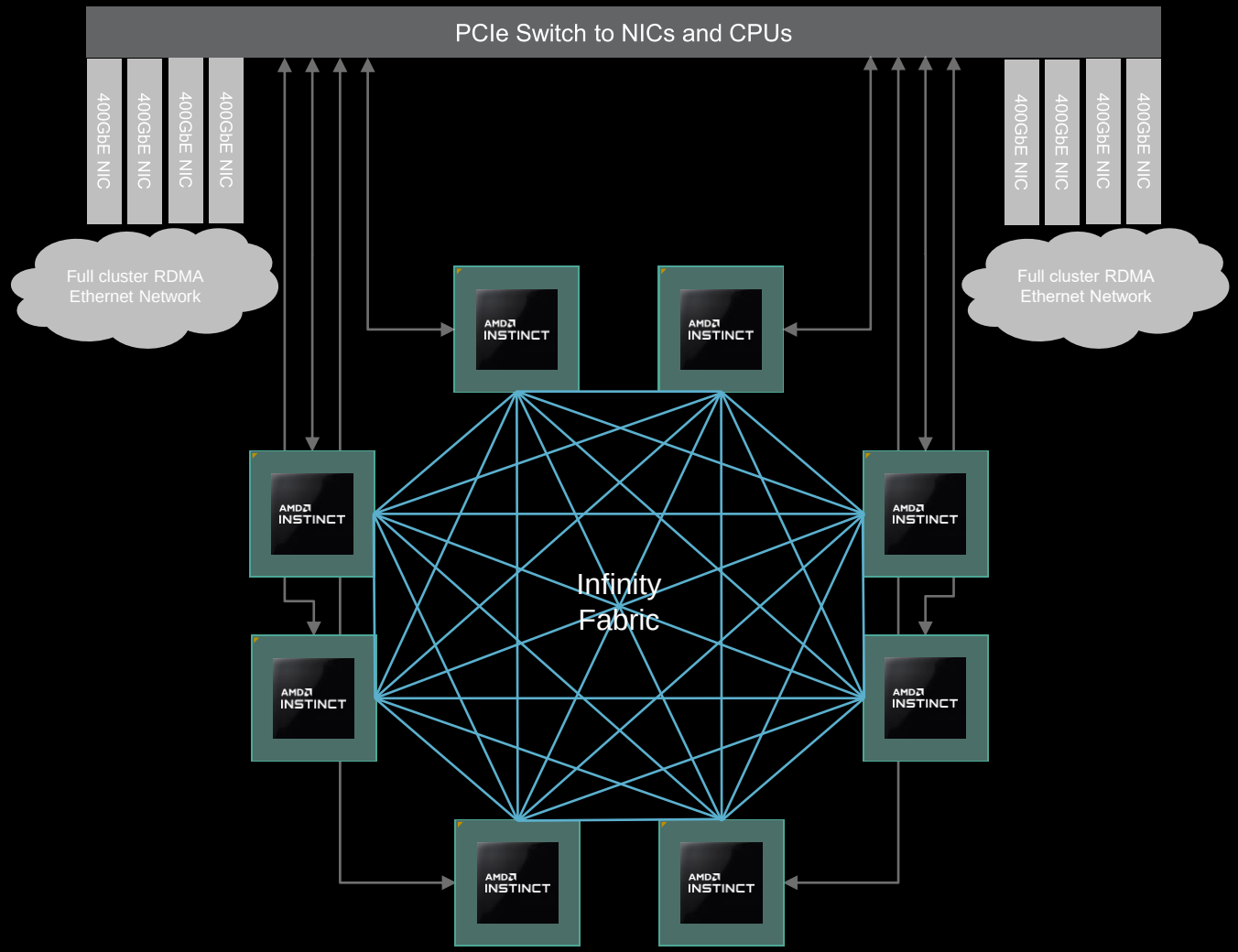
- https://rocm.blogs.amd.com/software-tools-optimization/compute-memory-modes/README.html

CPX & NPS4



Today’s MI300X Scale-up Network based on XGMI

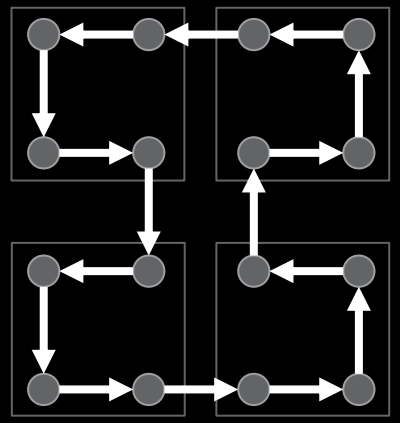
- Low-latency point-to-point XGMI “Infinity Fabric” interconnect
- LD/ST for communication within a GPU and across the 8-GPU hive



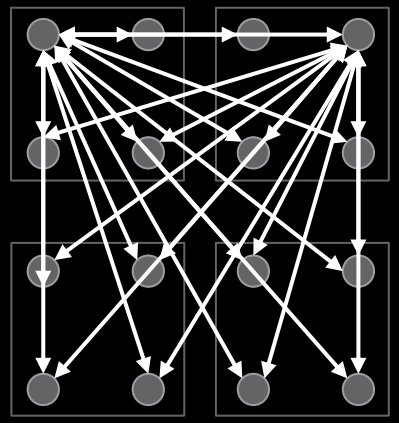
Customizing Collective Kernels for the Hierarchical Network

Microsoft's Collective Communication Library (MSCCL++)* for GPUs allows for customized collectives

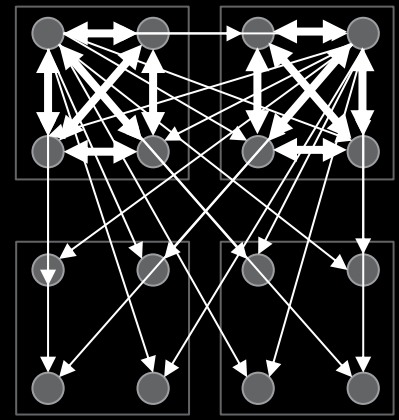
- Addresses overheads in the P2P communication of the original two-sided MSCCL implementation
- New one-sided API on top of which collectives can be implemented in C++ or MSCCLang
  - MSCCLang is Python-based DSL to write collective algorithms and exported in JSON
  - Simplifies the customization of collectives to specific hardware



Non-hierarchical Ring



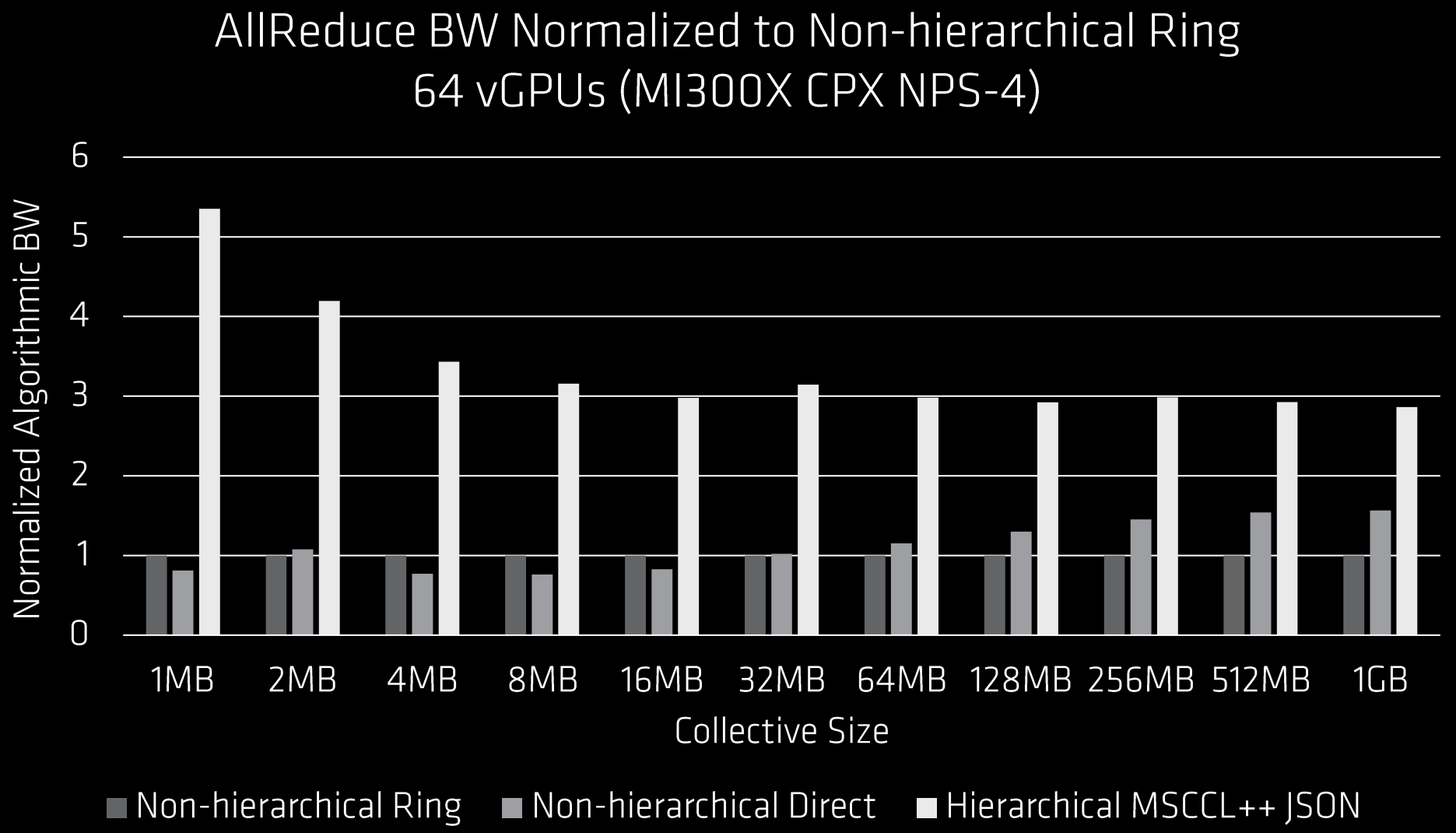
Non-hierarchical Direct



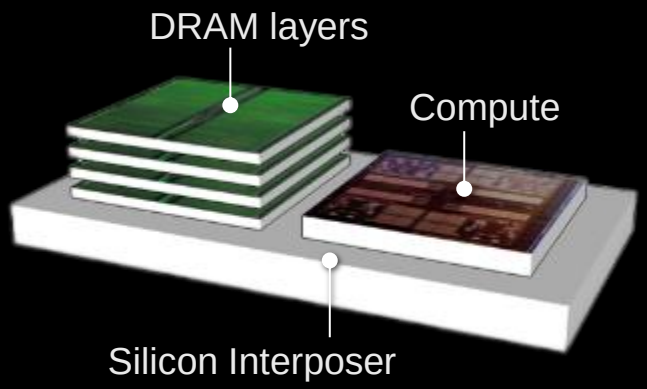
Hierarchical MSCCL++ JSON

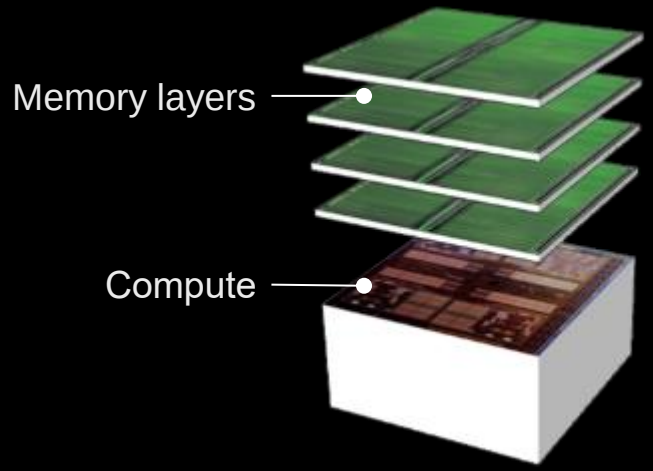
*https://github.com/microsoft/mscclpp/

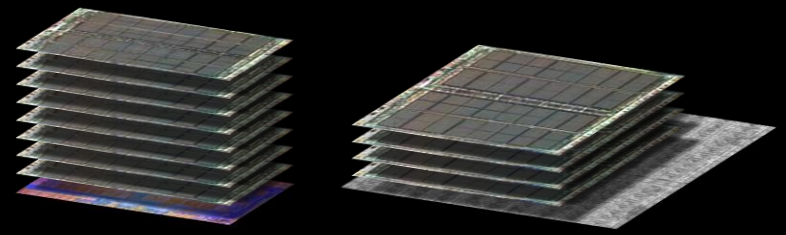
HW Evaluation



Even Tighter Integration of Compute and Memory







Processing in Memory

Processing in Memory Controller

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.

Conclusion: Meeting the Challenge Requires Holistic Innovation

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

Disclaimer

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

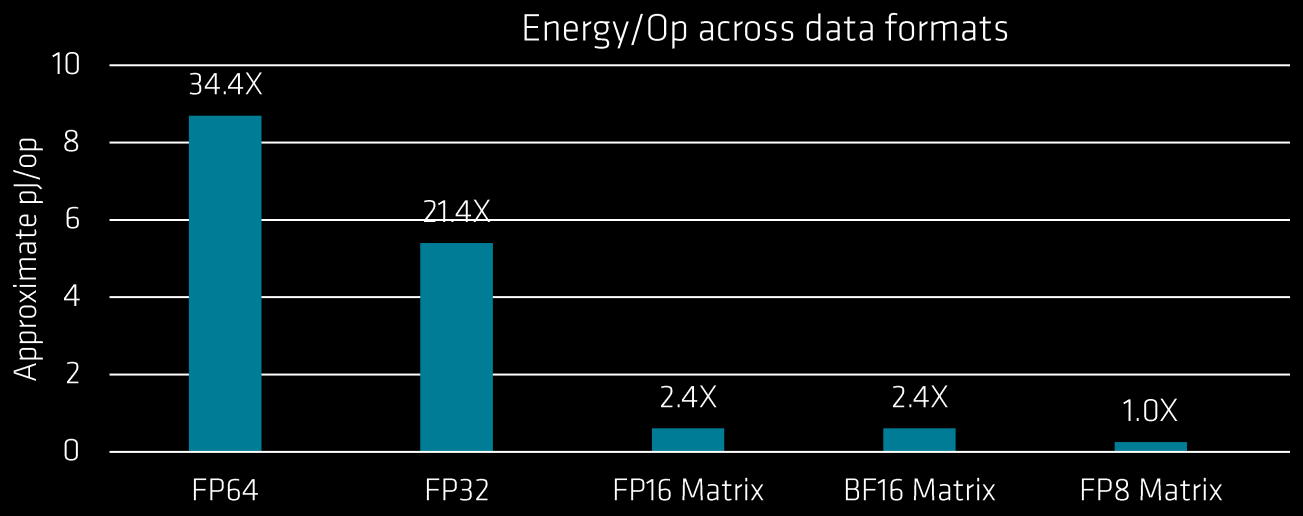
The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

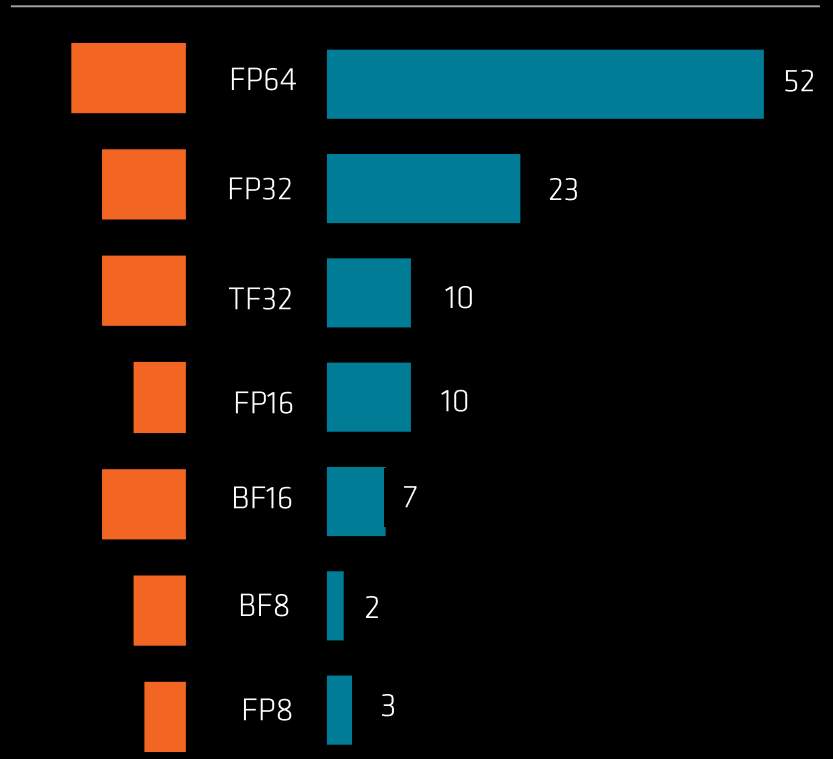
THIS INFORMATION IS PROVIDED ‘AS IS.” AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS, OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION. AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY, OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY RELIANCE, DIRECT, INDIRECT, SPECIAL, OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.


AMD

Compute Power and Reduced Precision

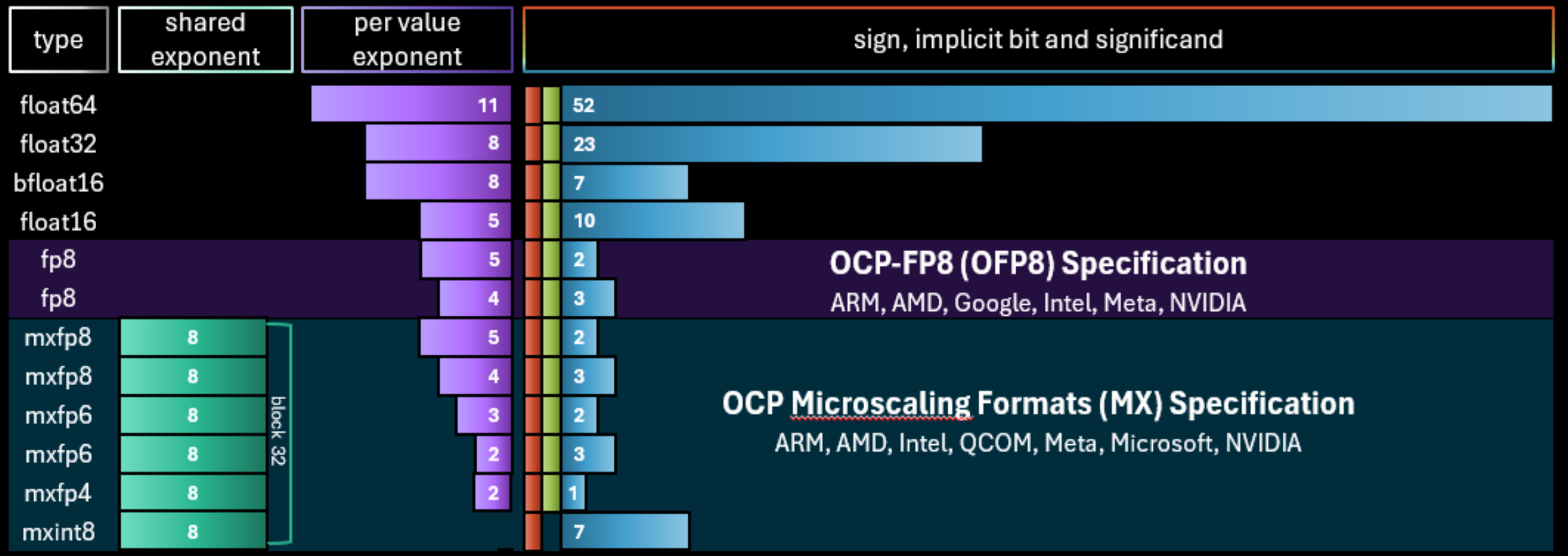
- New AI algorithms that exploit reduced precision arithmetic offer orders improved compute efficiency
  - 64b → 32b →16b → 8b → 4b
- Dedicated matrix math datapaths increase efficiency further





AMD

OCP Microscaling MX Formats



AMD