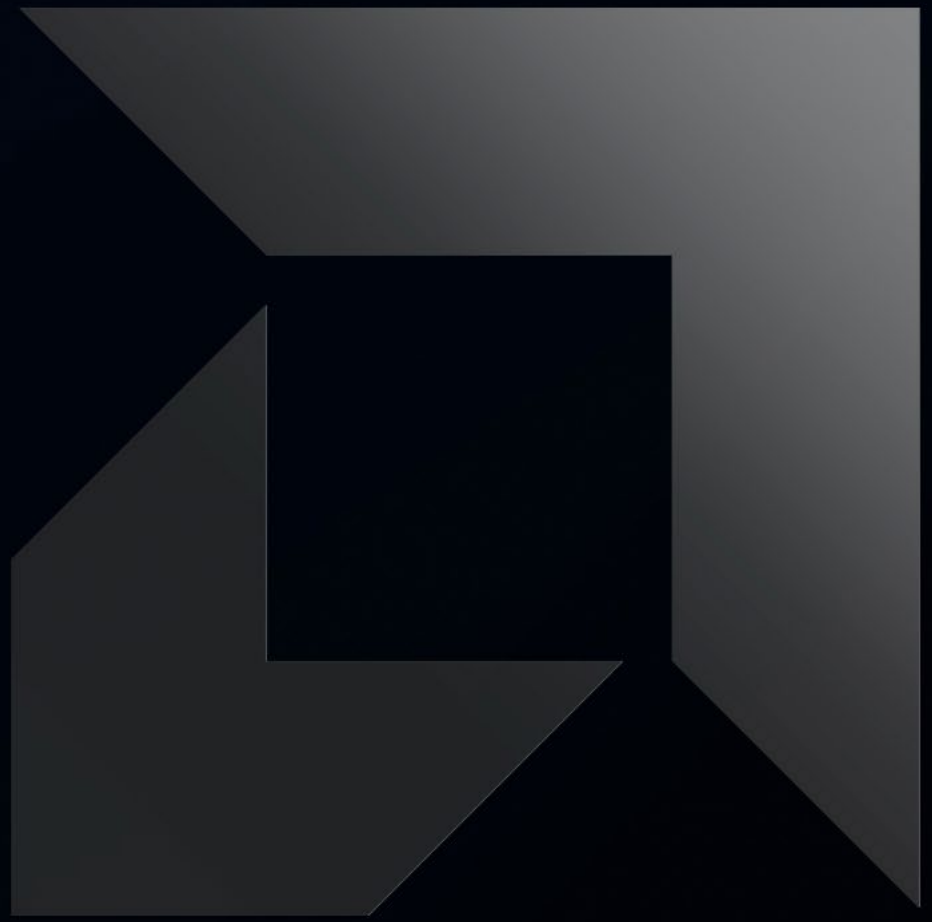


Energy-Efficient Computing for AI and HPC

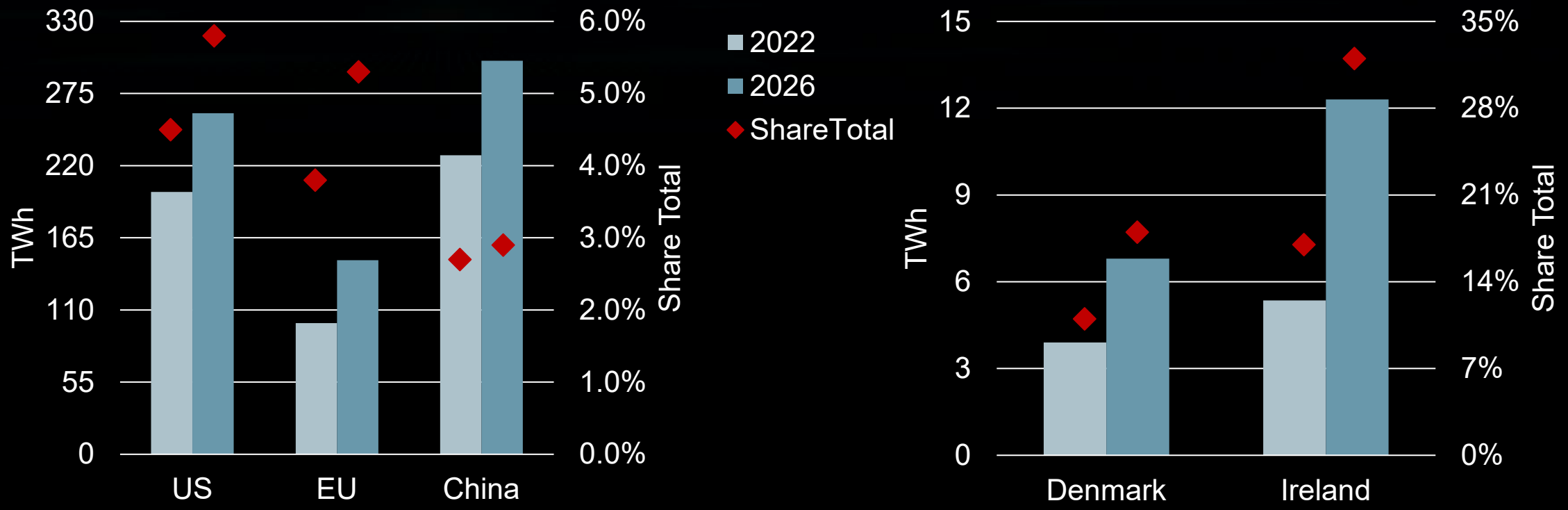
Michael Schulte
Senior Fellow, AMD Research

S0S27
May 18-20, 2025

AMD
together we advance_

The Problem

Data Center Energy Consumption

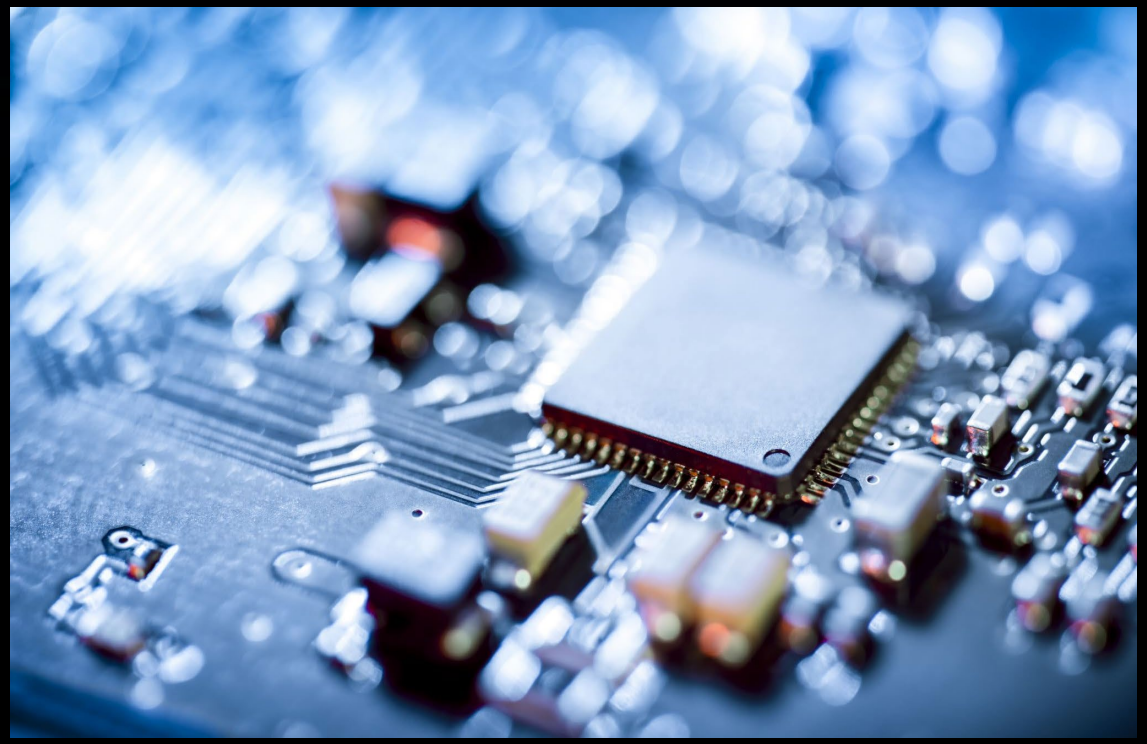


Data source: IEA (2024), Electricity 2024, IEA, Paris https://www.iea.org/reports/electricity-2024, License: CC BY 4.0

How We Got Here – 5 Decades of Innovations

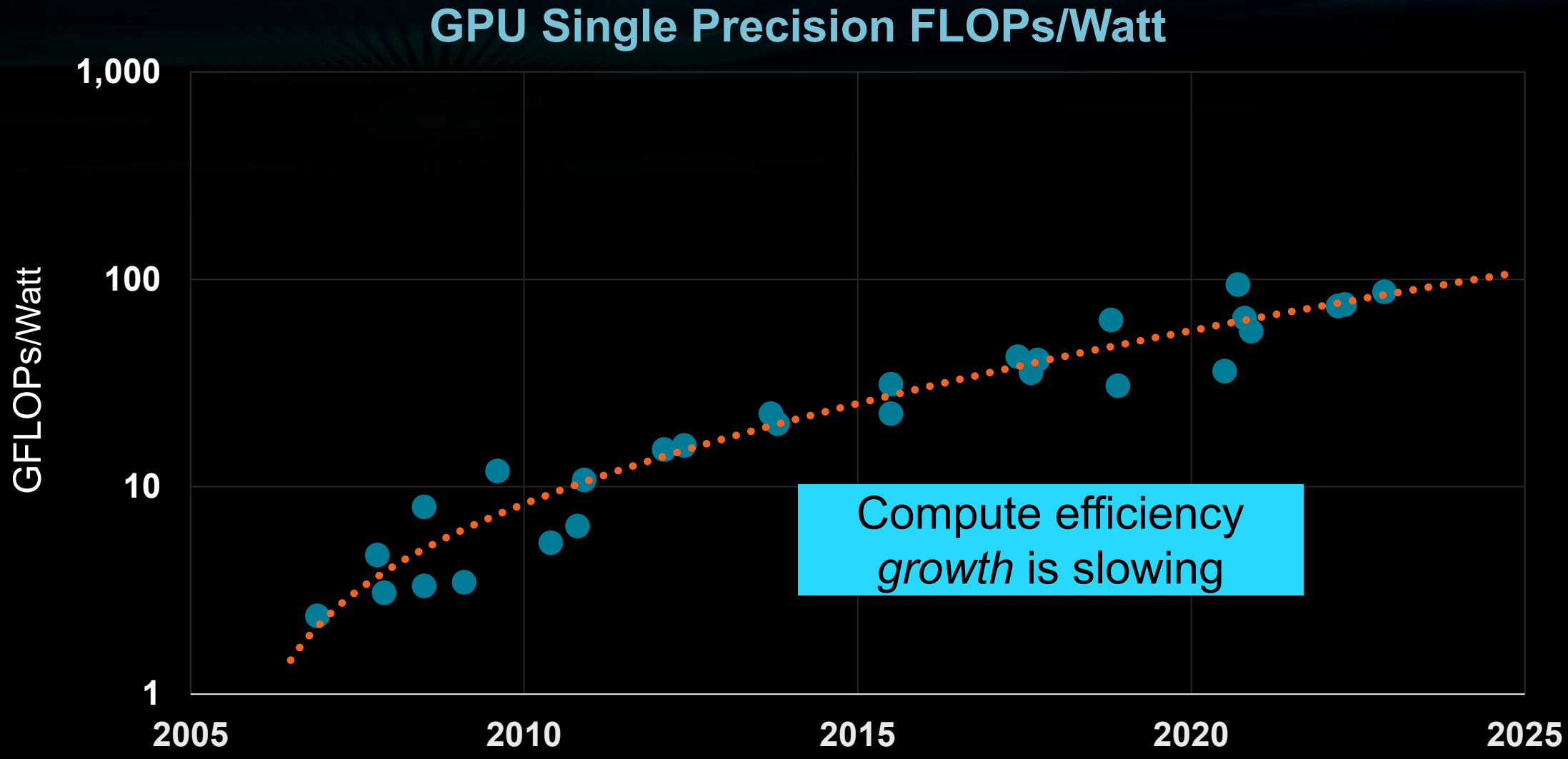
> 2000 times higher frequency

Supercomputers are ~17 billion times faster

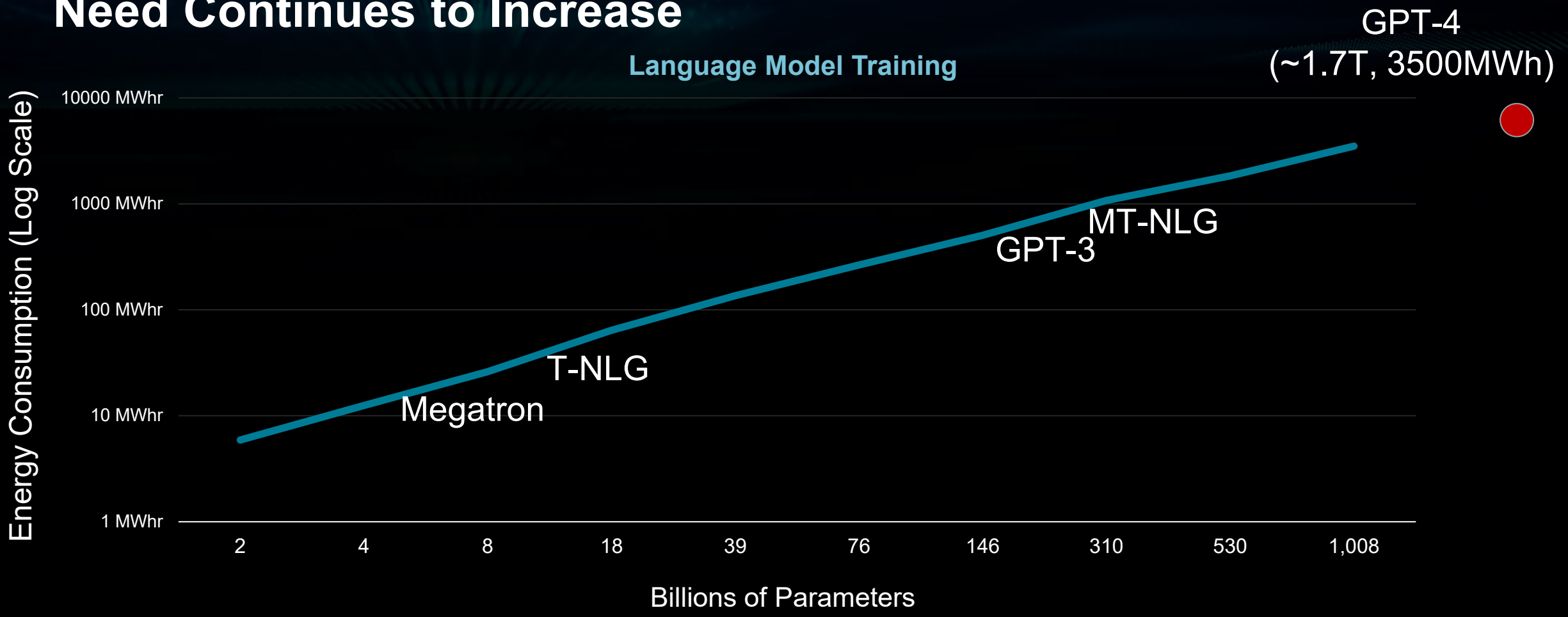


> 3000 times smaller

Trends – GPU Power Efficiency



Need Continues to Increase

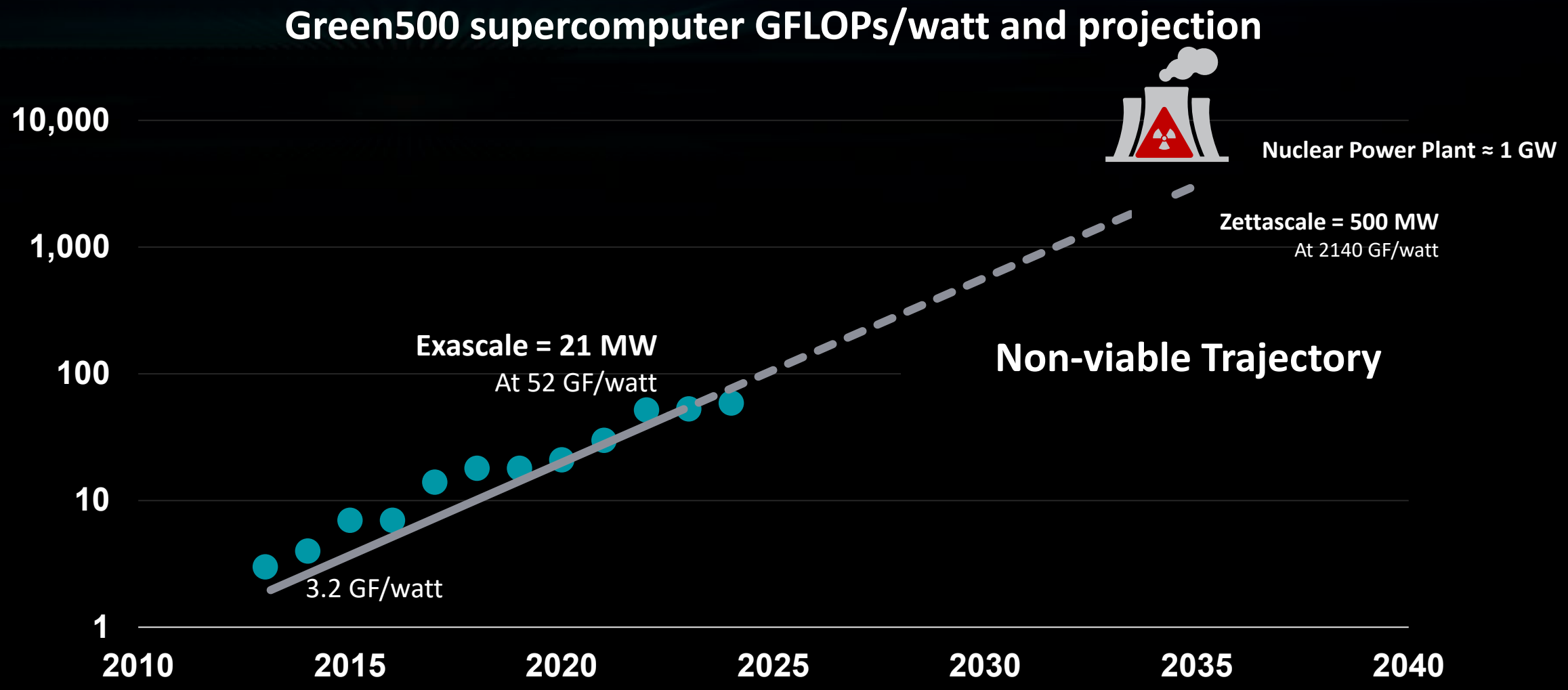


Exponentially growing model sizes drive immense growth in energy for training.

The upper bound on training requirements is yet to be determined.

Based on published parameter counts of leading training models; and AMD internal calculations

Supercomputer Energy Use Trajectory



The Opportunity

Domain-specific computation enables compute efficiency

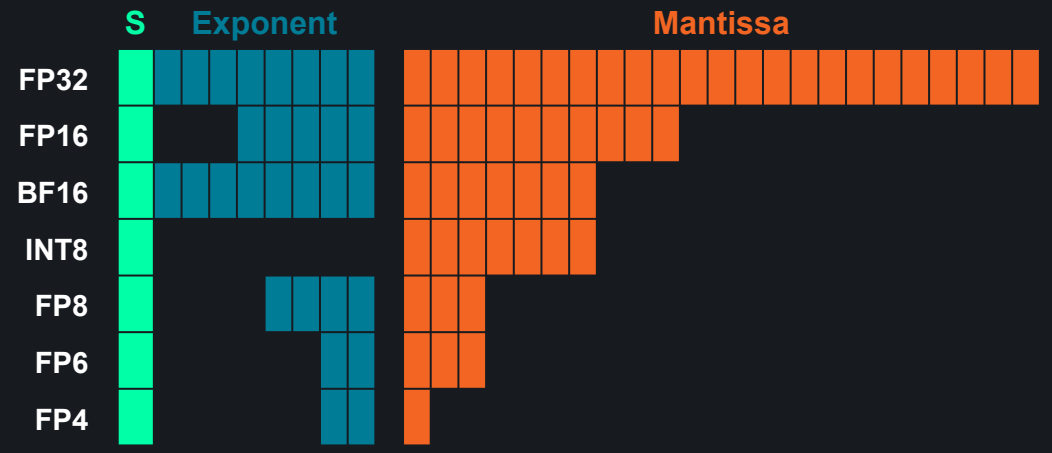
Tailor architecture for application

Adapt algorithms to use lower precision number formats for significant improvements in energy efficiency

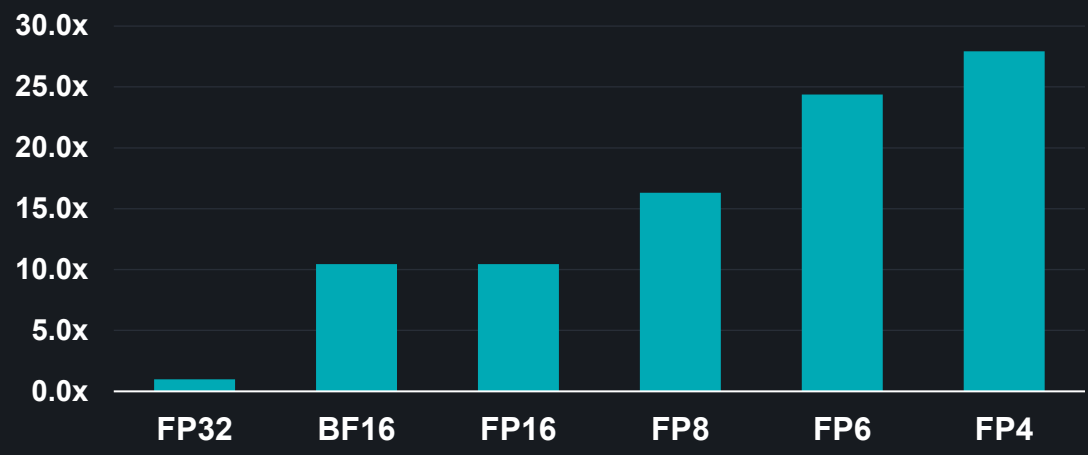
Need to determine appropriate precisions

Based on AMD internal calculations

Bit allocations by format



FLOPs/Joule normalized to FP32



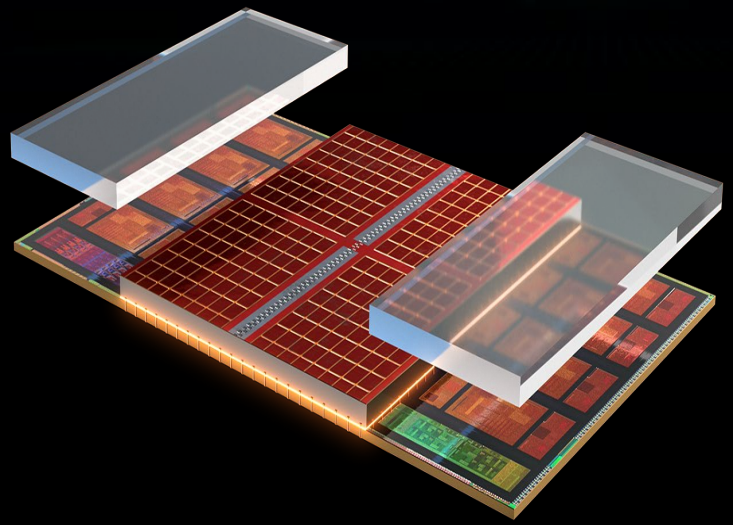
Reducing Data Movement Energy

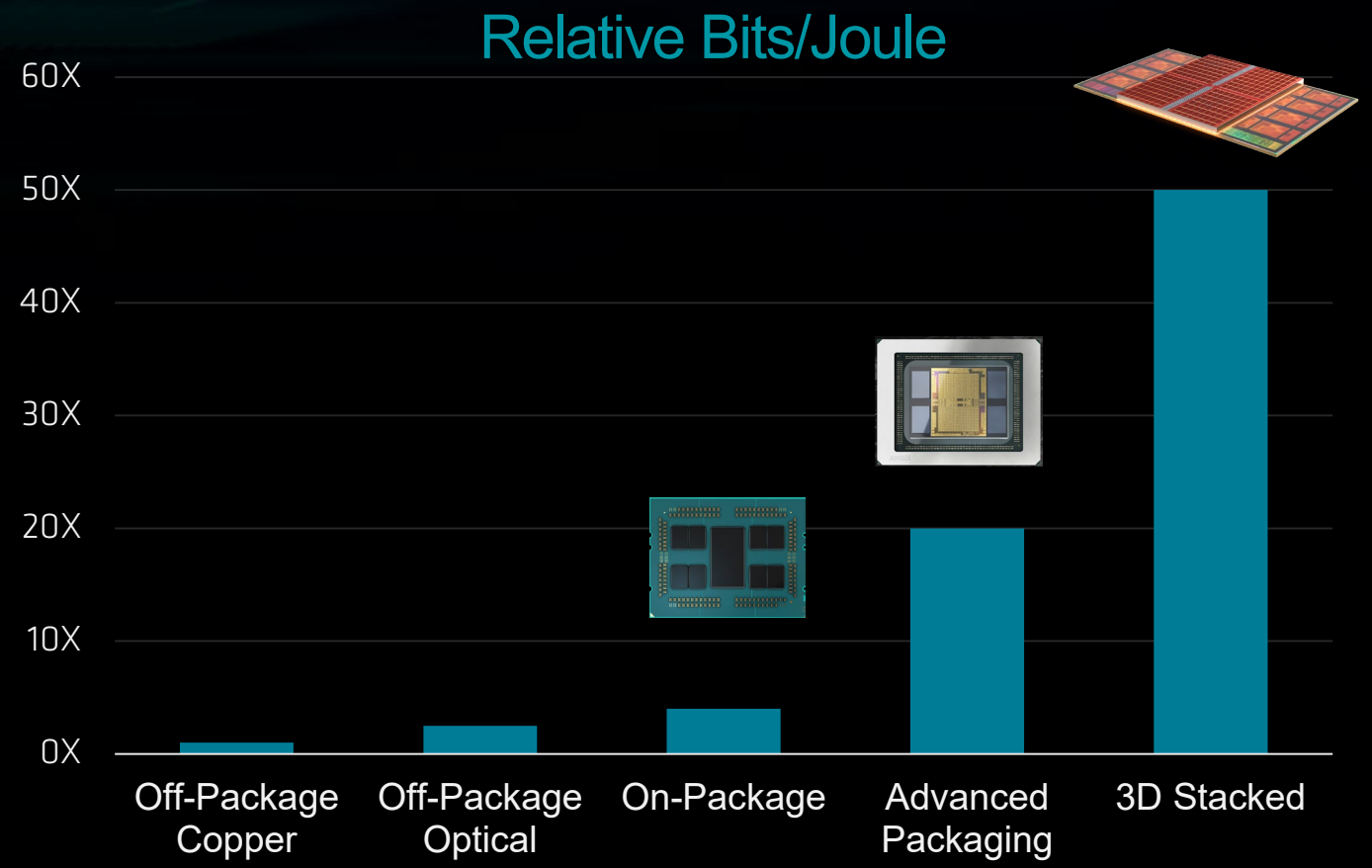
1.0	100	500	1500	1600	5000
Register file	Cache	HBM	CPU Memory	Scale-up interconnect	Scale-out interconnect

Normalized energy/bit

Data movement energy increases with distance → **Maximizing locality is key to efficiency**

3D Chiplets and Communication Energy

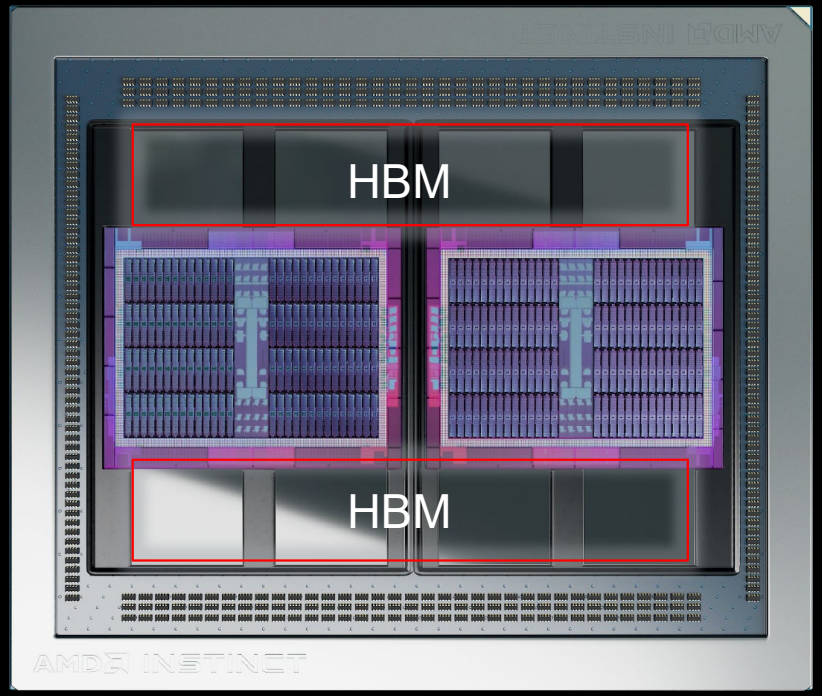


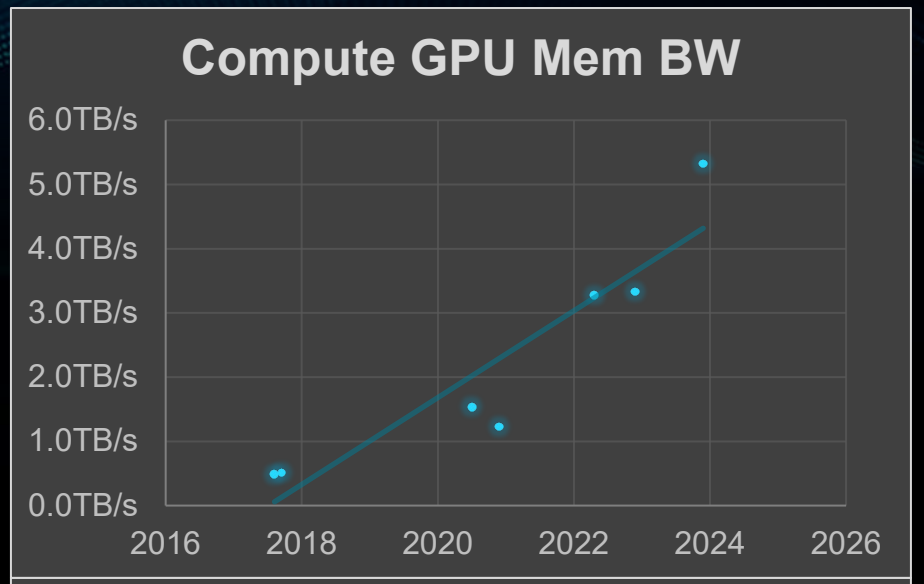


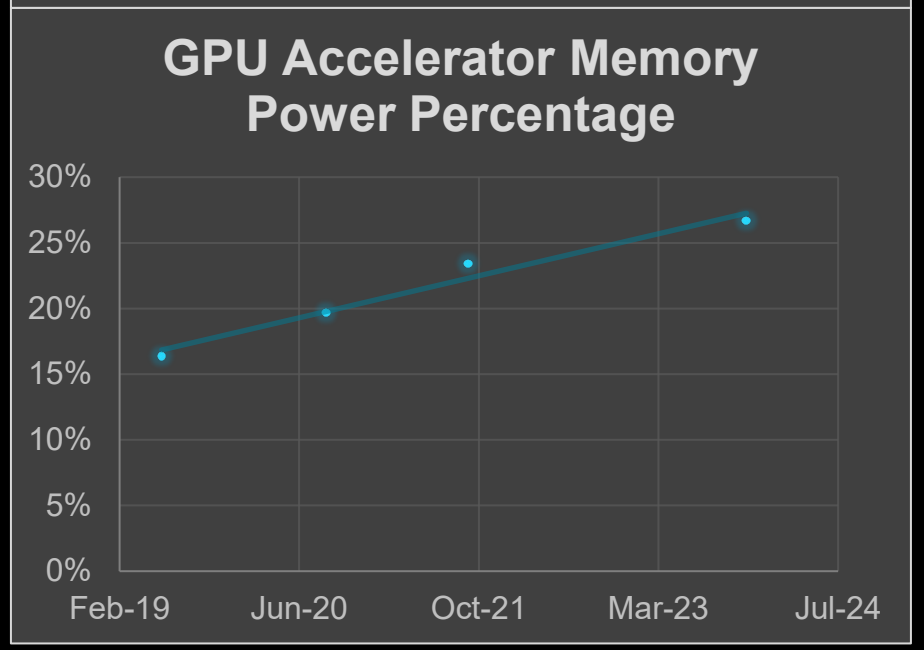
Advanced Packaging Provides up to a 50x Reduction in Communication Power

Thirst for Memory Bandwidth

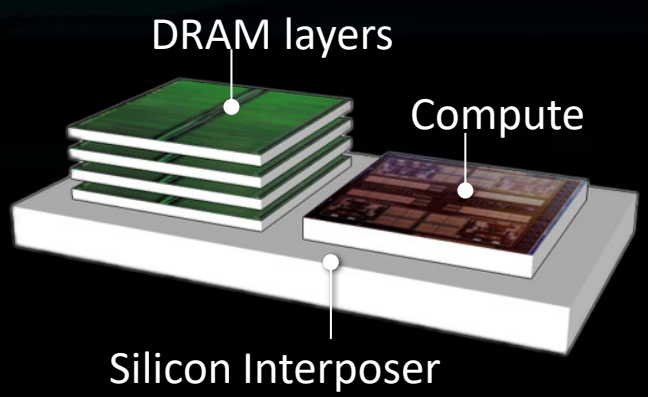
- High bandwidth memory feeds the compute engine providing a key element of performance gains
- Limited efficiency gains combine with demand growth result in higher percentage of power for memory

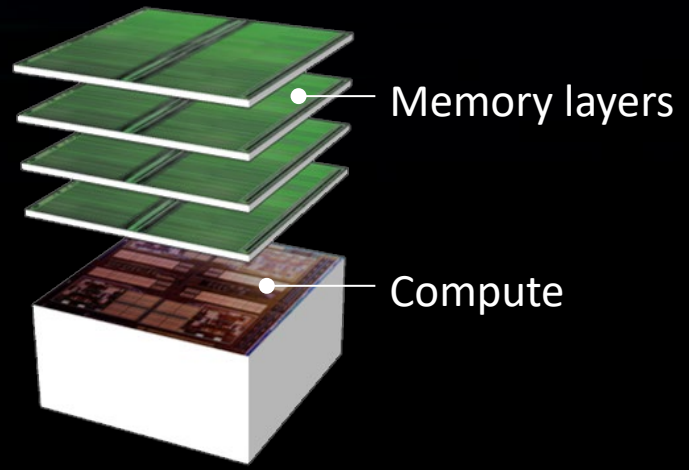


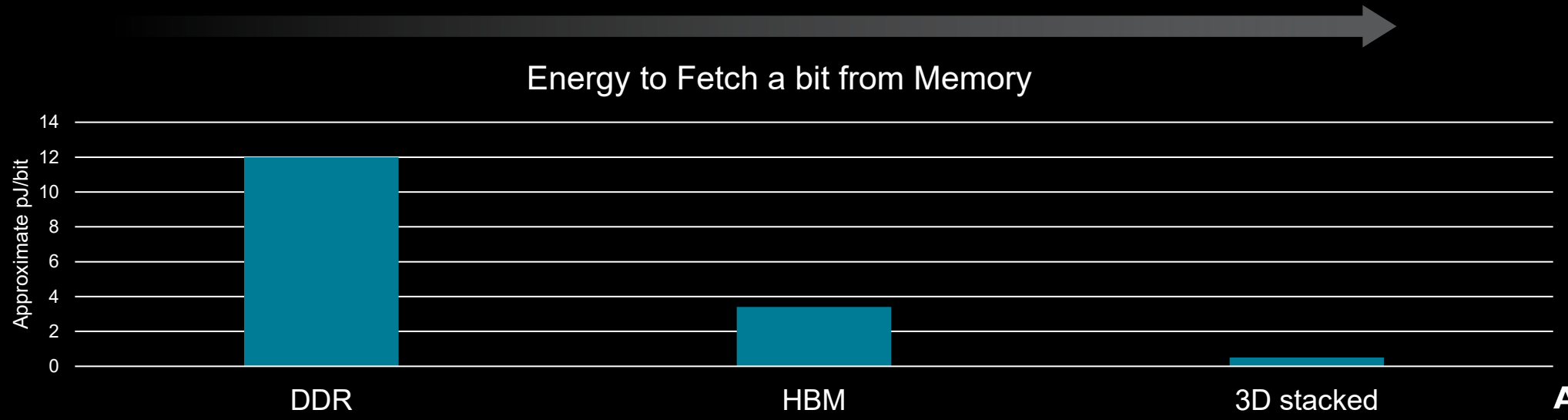




Reducing Memory Energy

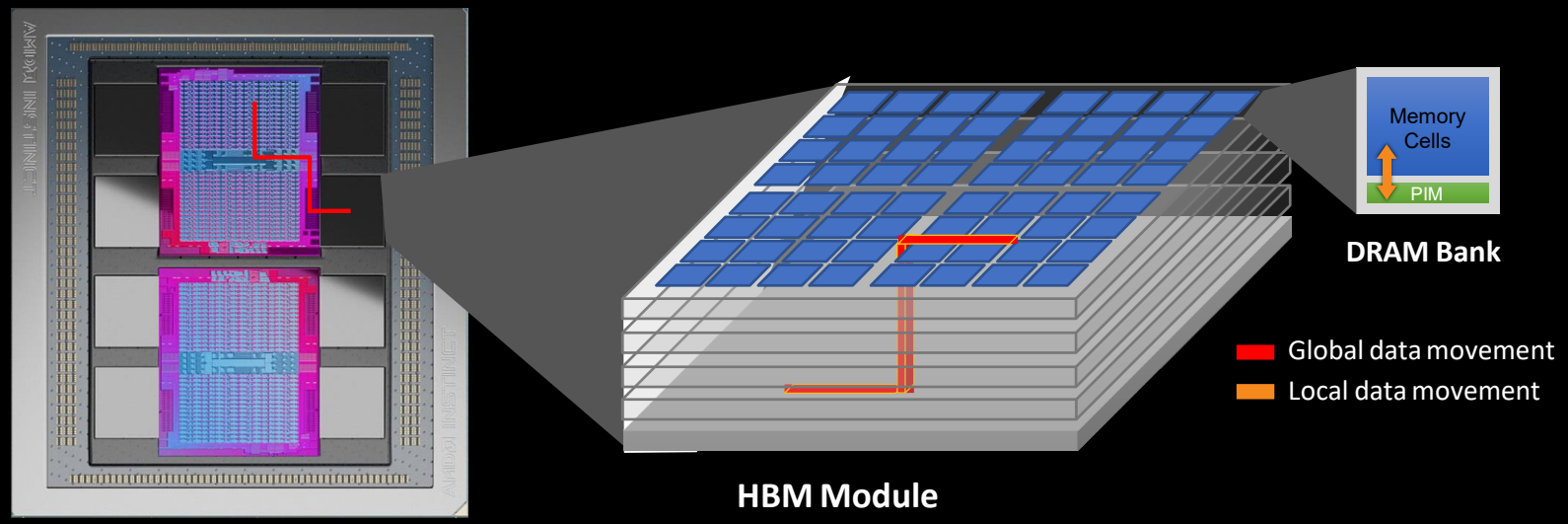


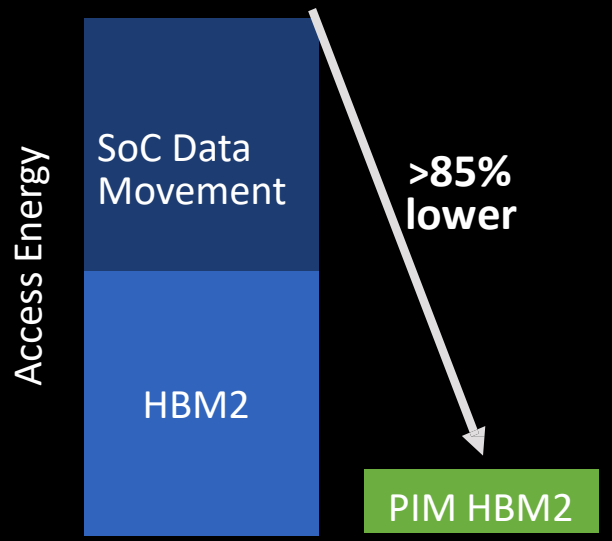




Processing in Memory

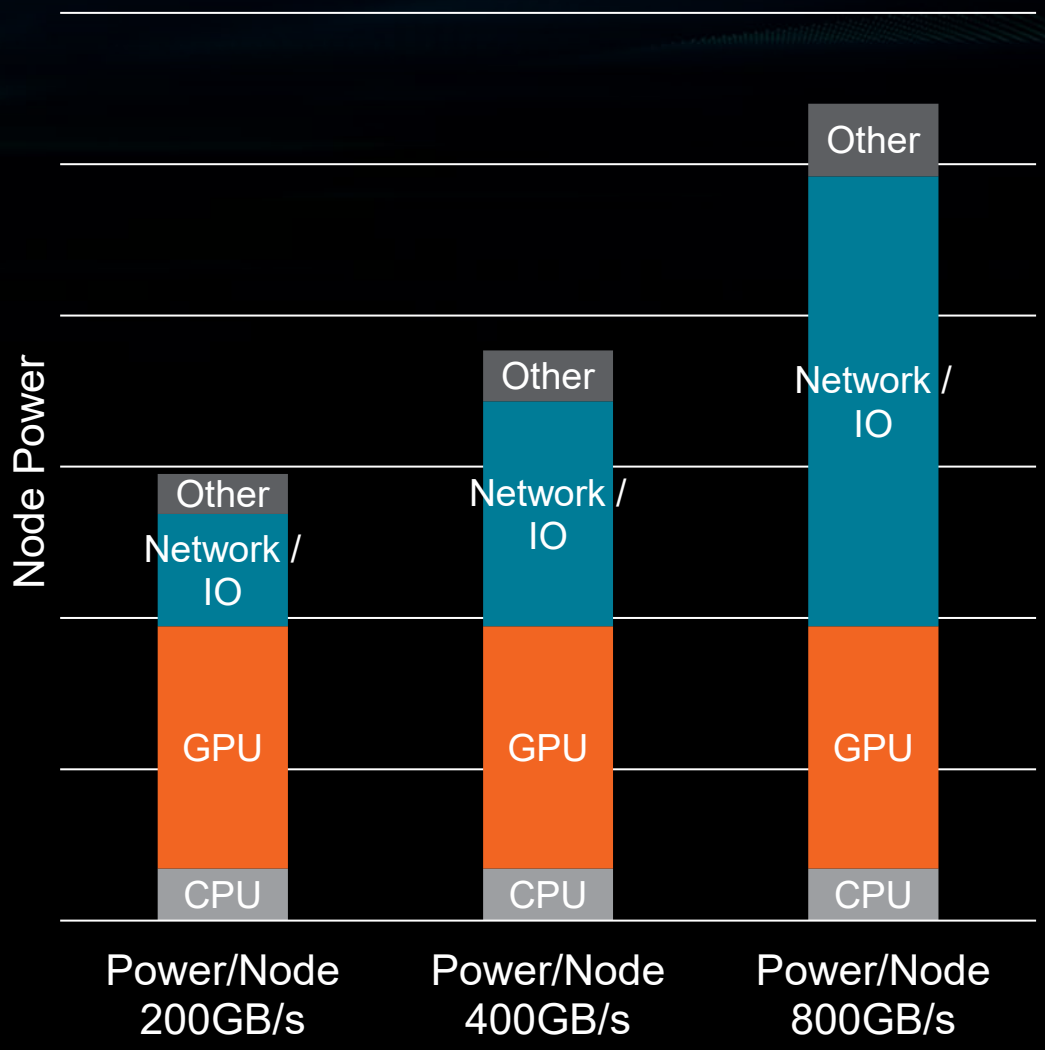
- Key algorithmic kernels can be executed directly in memory, saving precious data movement energy



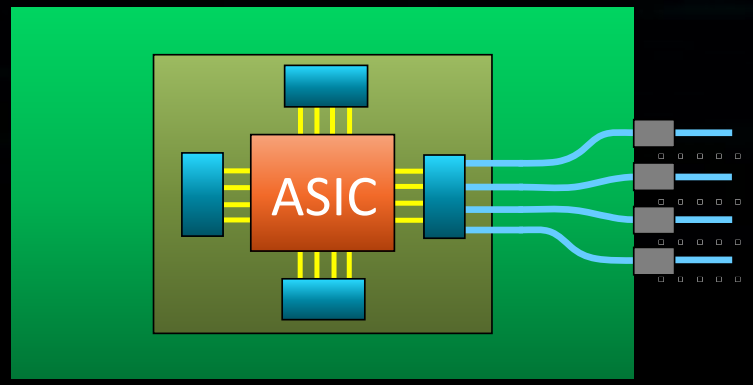


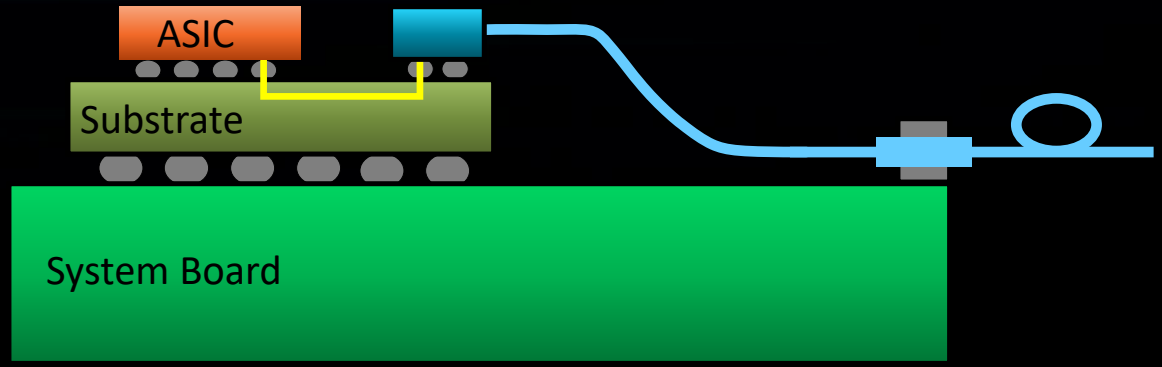
System Power by Function

- Historical trends for model growth and system requirements point to a doubling of network bandwidth every two years
- Even if compute power can be contained, network and IO power will grow
- In two generations, we expect network+IO power to dominate compute node power in AI systems
- Lower power solutions are needed



Optical Communication for Energy Efficient Networks





Co-packaged optics can provide a path forward

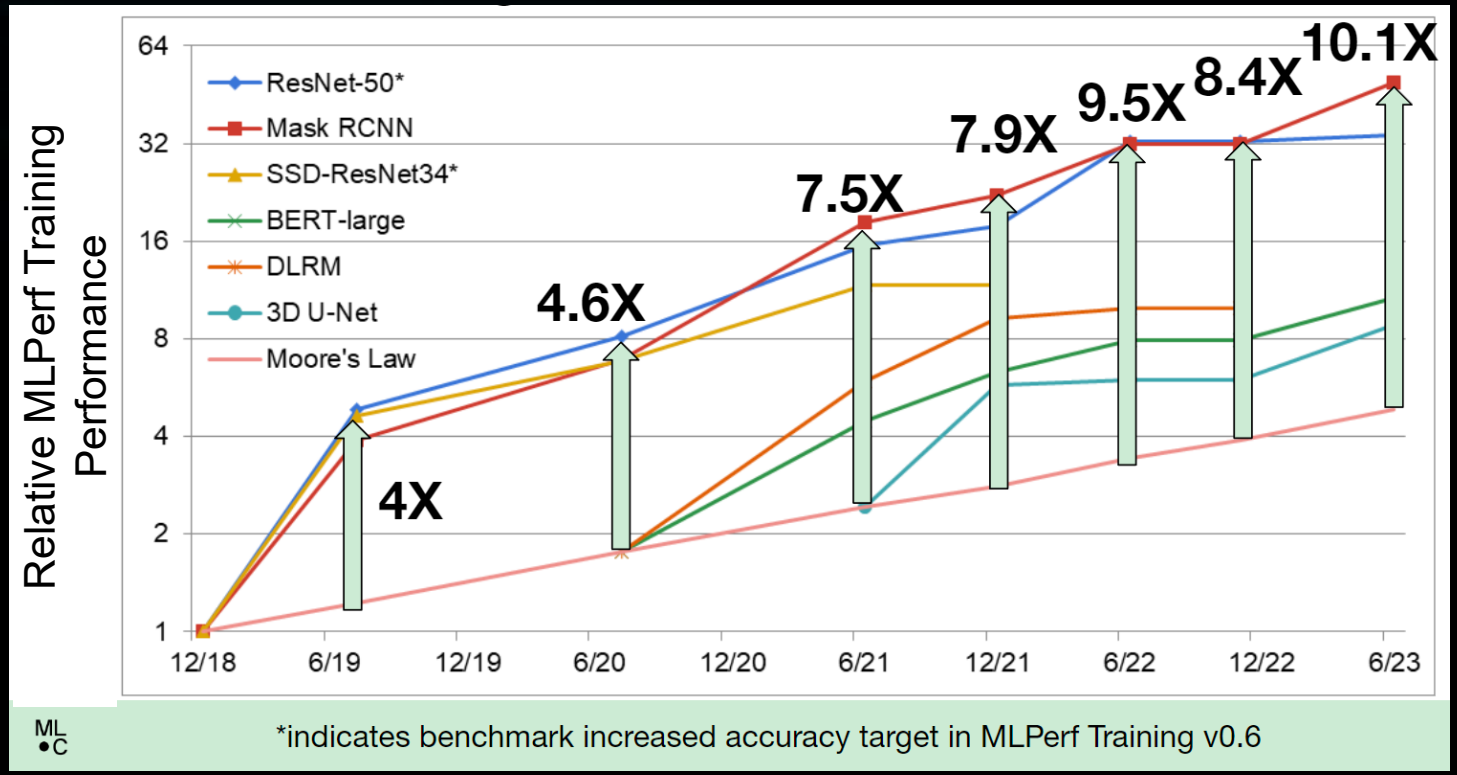
Reach and BW density reduces switch and re-timer power

Path to ~1 pJ/bit and optical circuit switches for greater efficiency

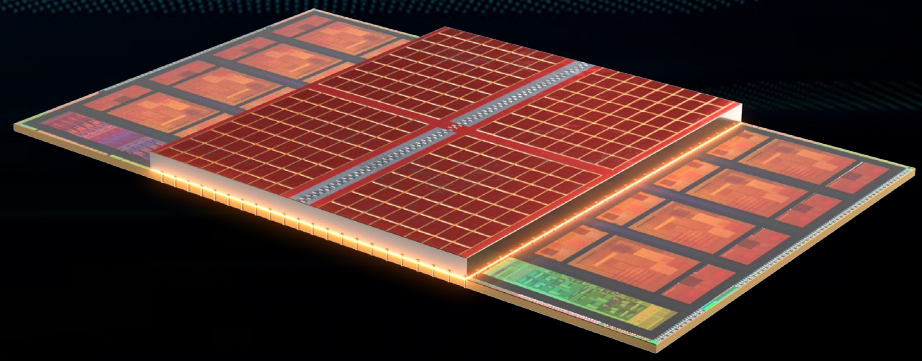
Tight integration of optical transceivers to compute die is a key to efficiency

Algorithm-Software-Hardware Co-Design

- Combination of algorithms, software, and architecture have been and will continue to be a critical lever.
- We can also leverage AI to make the entire system more efficient.



Meeting the Challenge Requires Holistic Innovation



- Hardware architecture
- Advanced packaging
- New interconnects and memory
- System level integration
- SW optimizations
- Intelligent design and management
- Algorithm-software-hardware co-design

18

Final Thought

100+ MW

5,000,000

20W


AMD

Copyright and Disclaimer