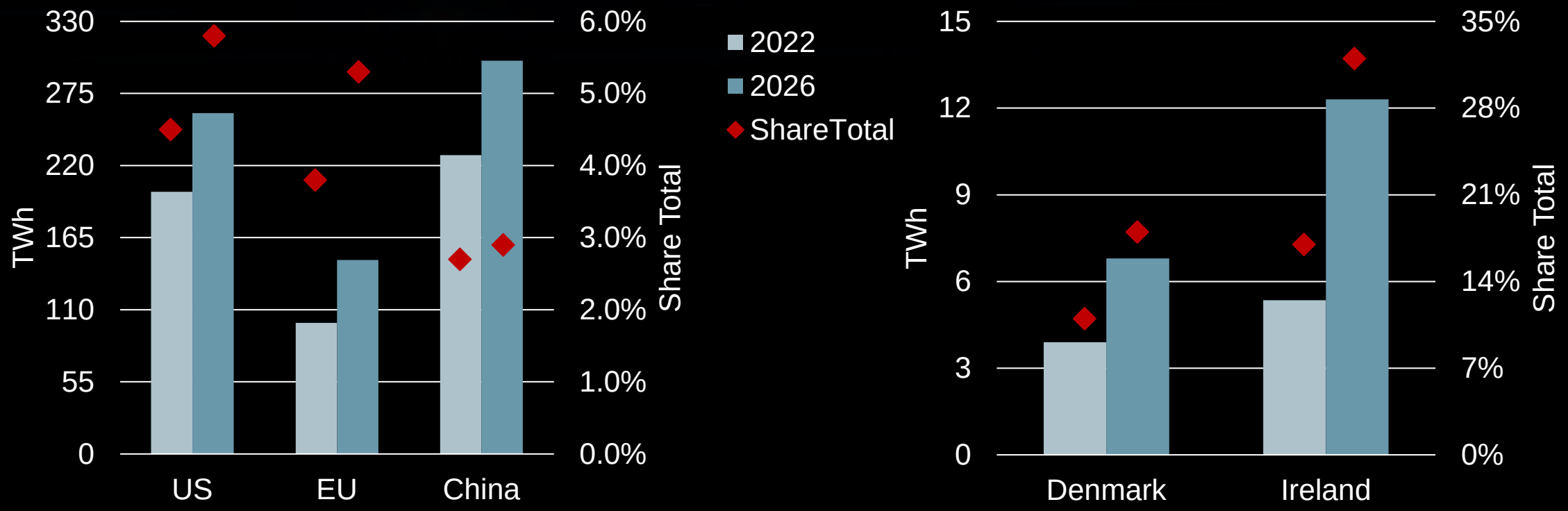
Energy Efficient Computing for Science

Srilatha (Bobbie) Manne
Senior Fellow, AMD

AMD
together we advance_

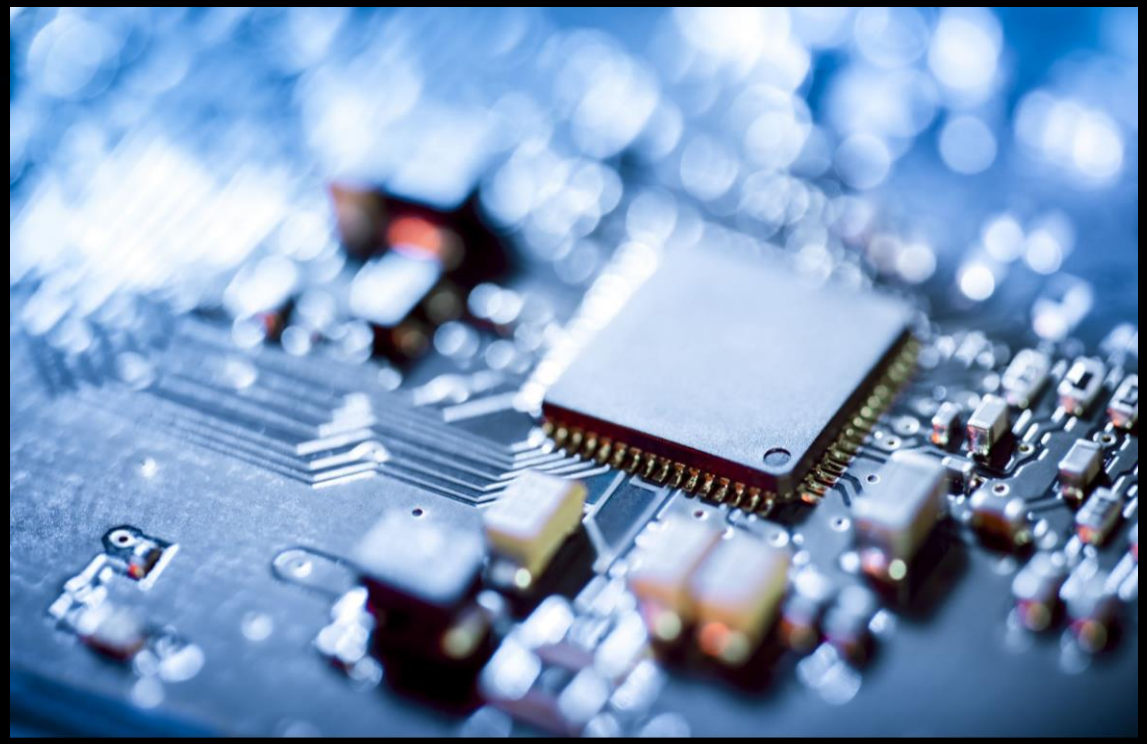
The Problem

Where we are



How we got here – 5 Decades of Innovations

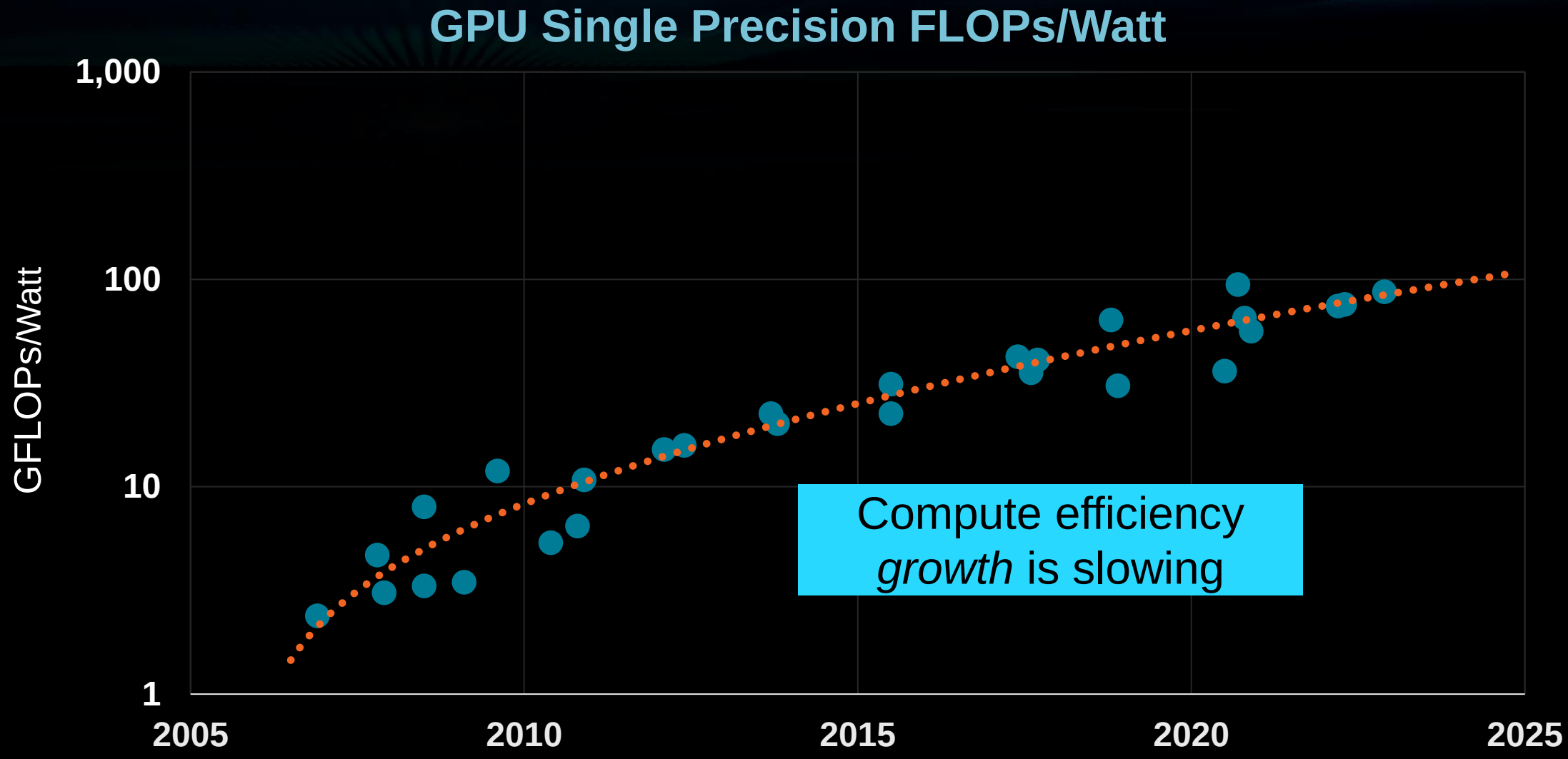
> 2000 times faster



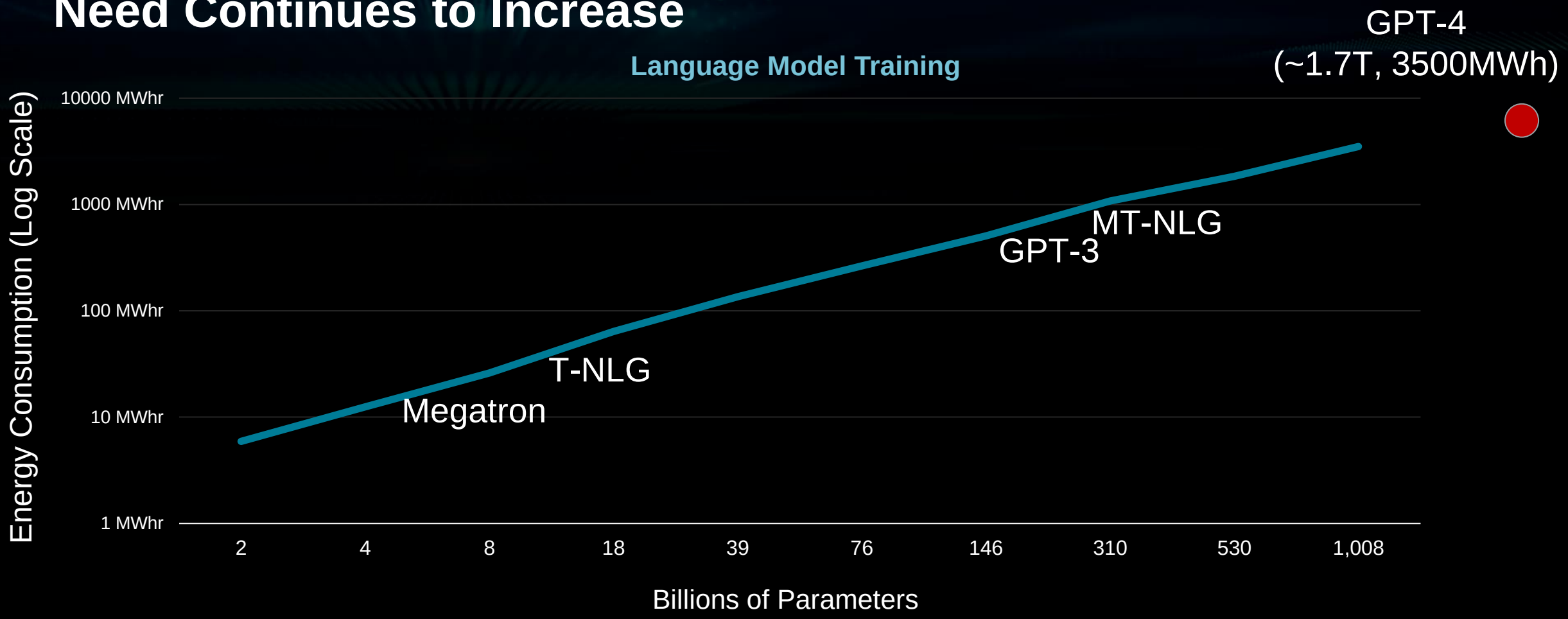
> 3000 times smaller

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 Efficiency



Need Continues to Increase

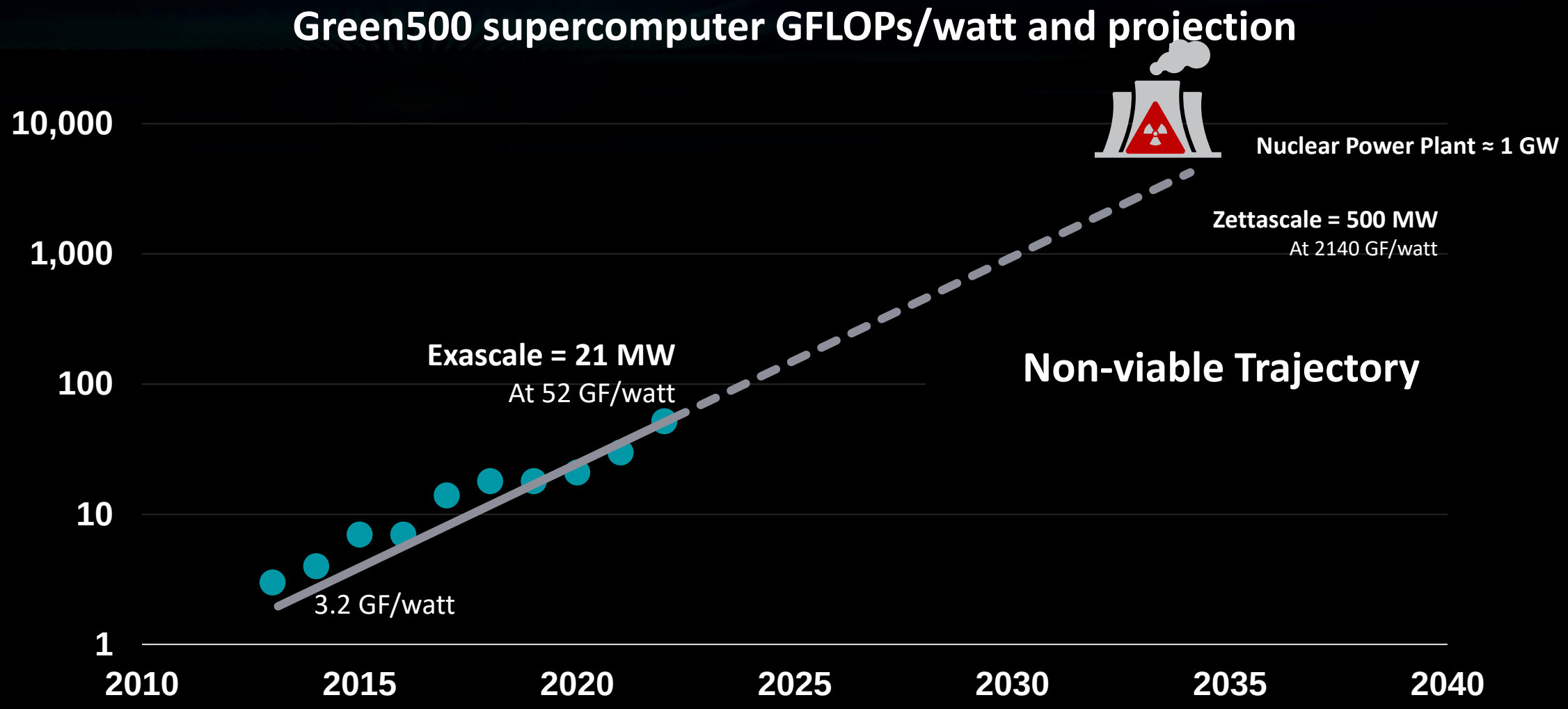


Exponentially growing model sizes drive immense growth in energy for training.

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

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

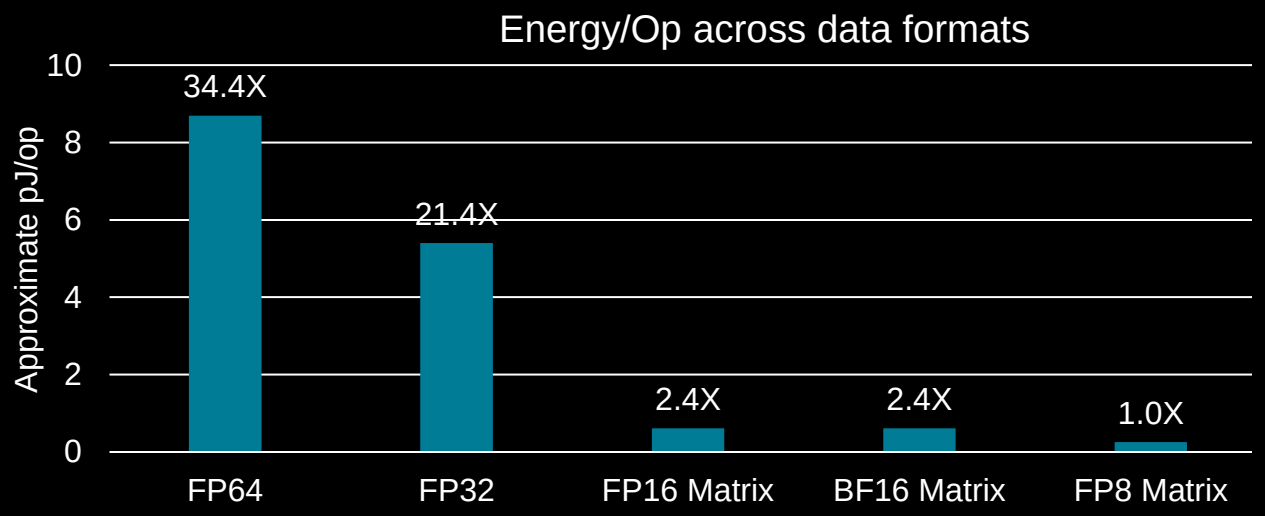
Supercomputer Energy Use Trajectory

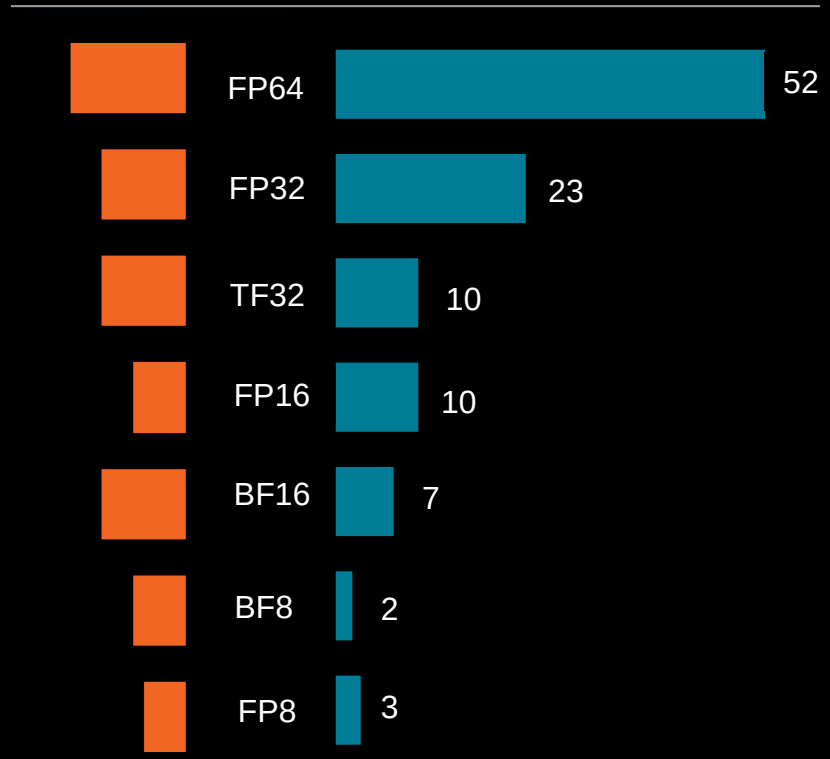


The Opportunity

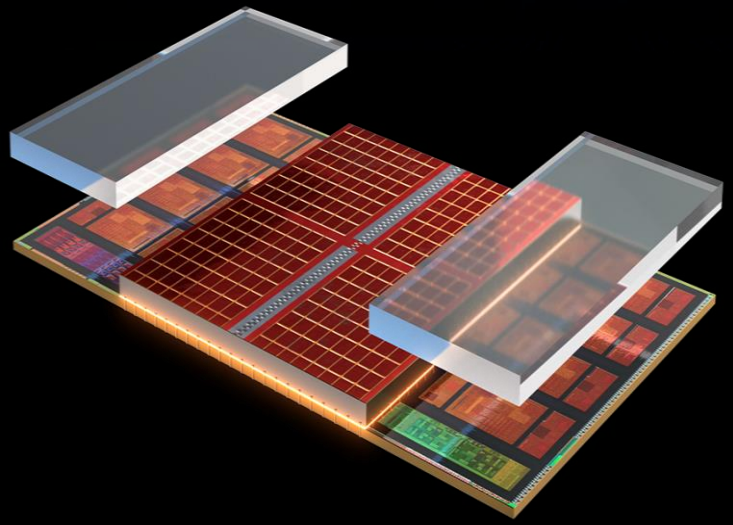
Flop Power and Reduced Precision

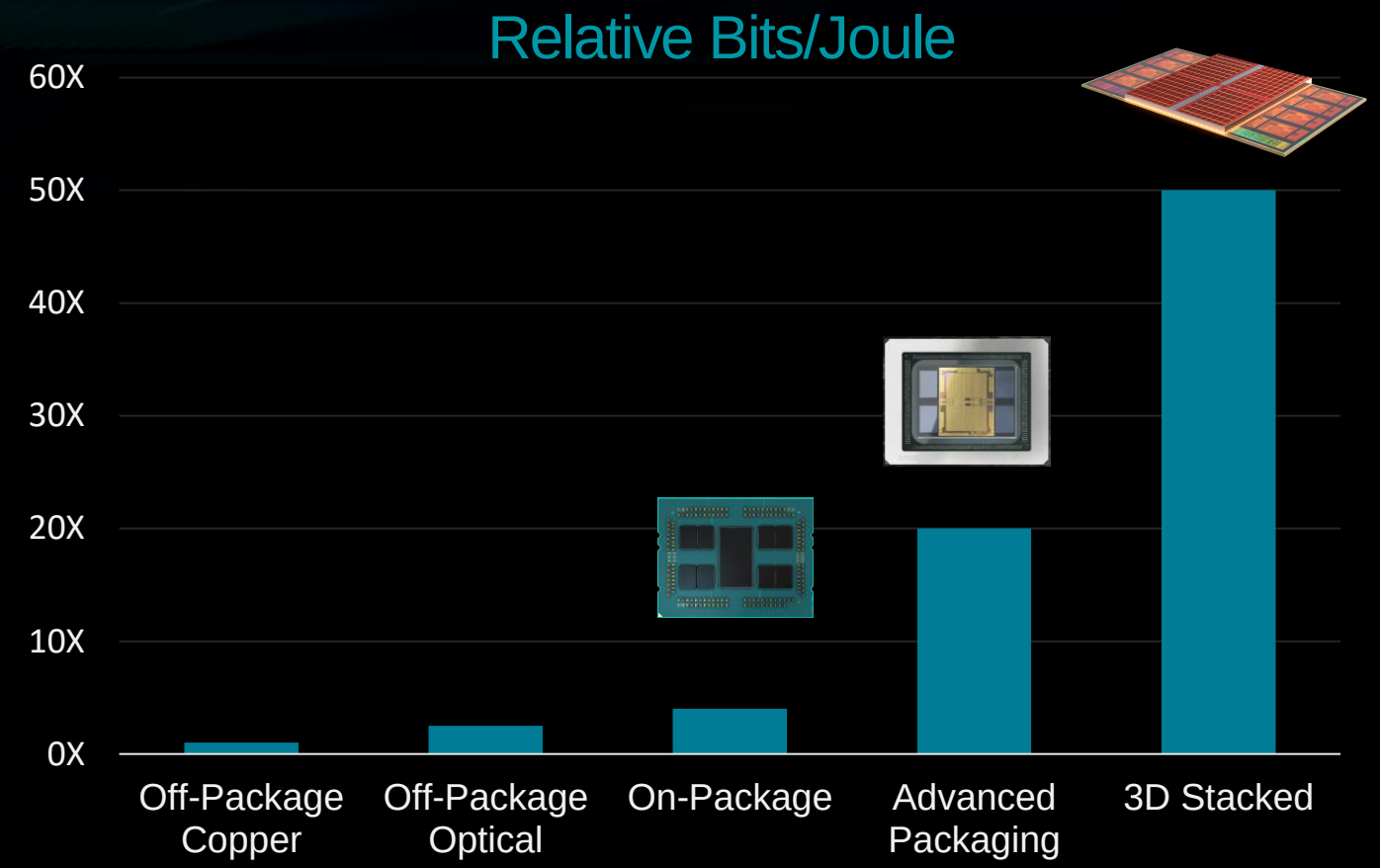
- New algorithms to exploit reduced precision arithmetic offer orders of magnitude improved compute efficiency
  - 32b →16b→8b →6b and 4b formats
- Dedicated matrix math datapaths increase efficiency further





3D Chiplets and Communication Power

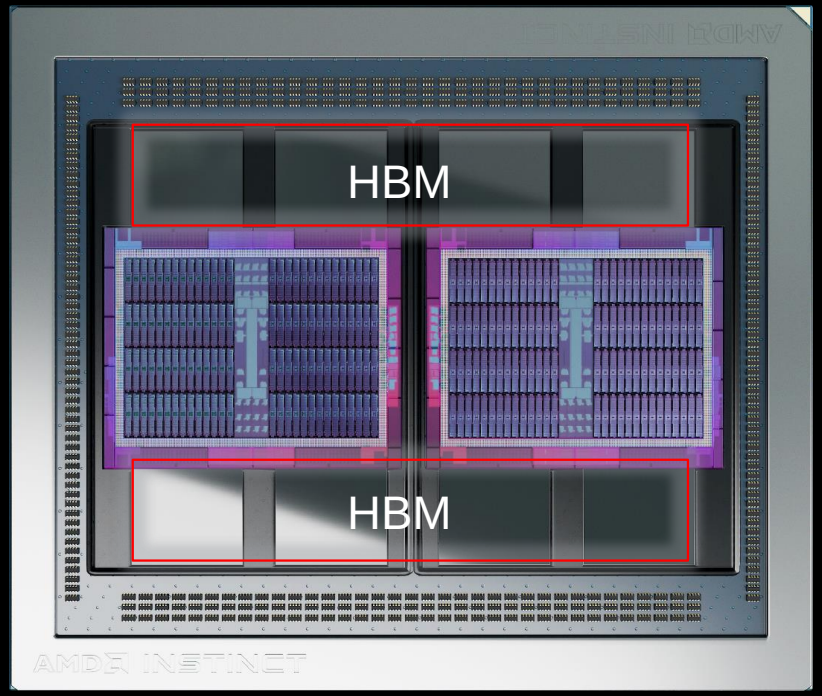


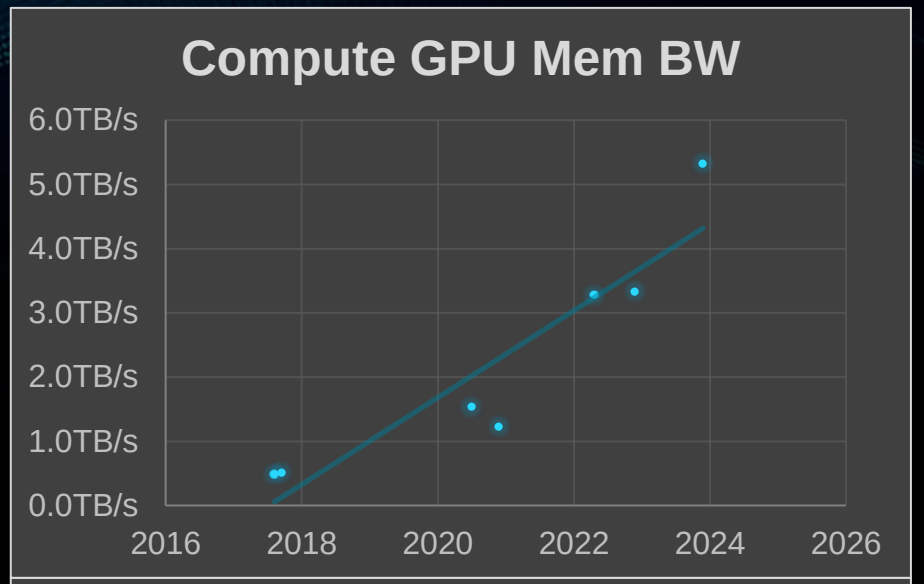


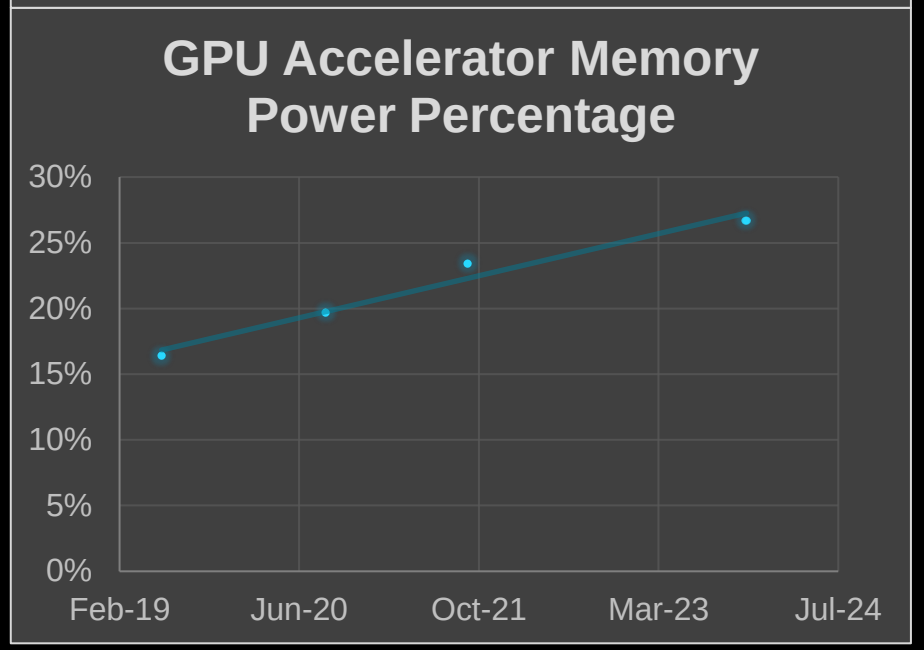
Advanced Packaging Provides up to a 50x Reduction in Communication Power

Thirst for Memory Bandwidth

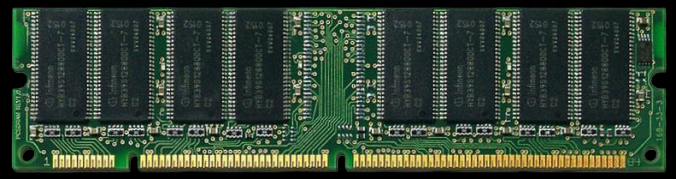
- High bandwidth memory feeds the compute engine providing a key element of performance gains
- Limited efficiency gains combine with demand growth result in higher percentage of power for memory

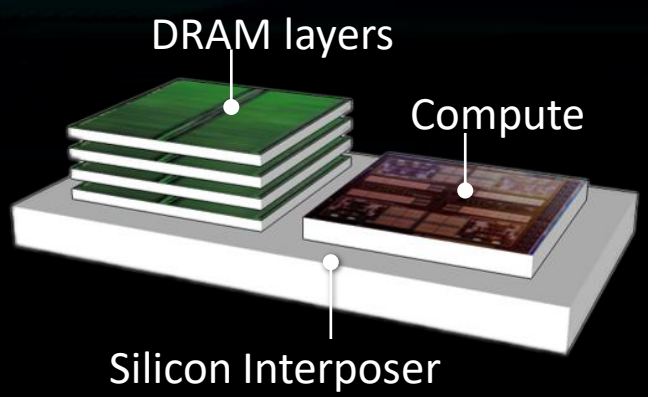


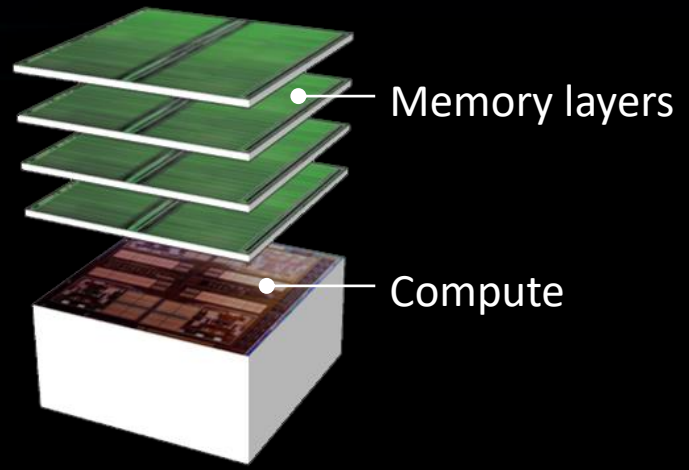


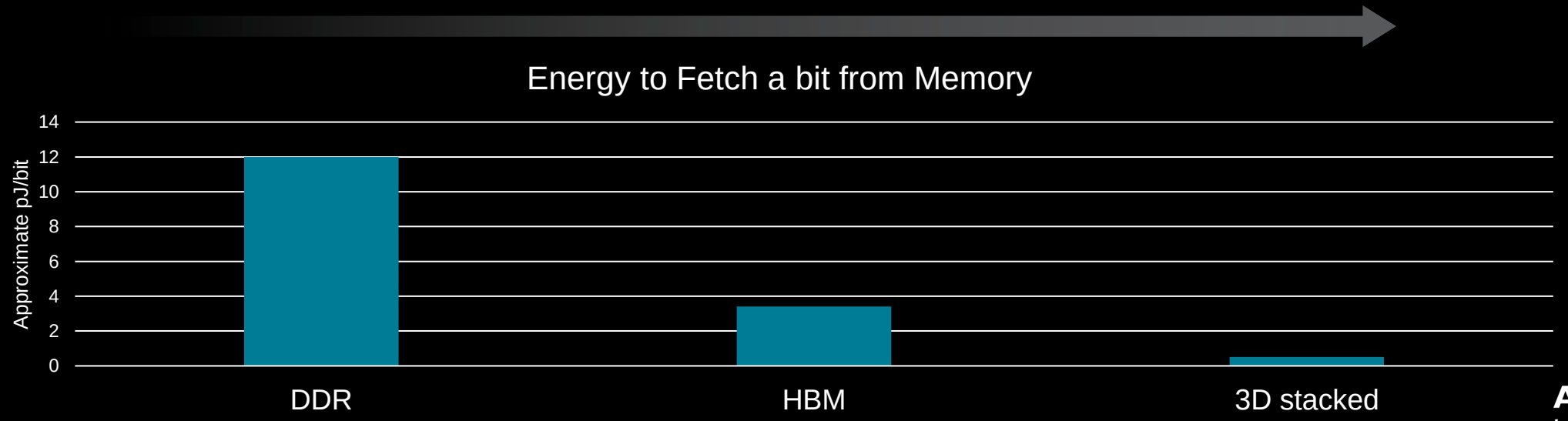


Reducing Memory Power



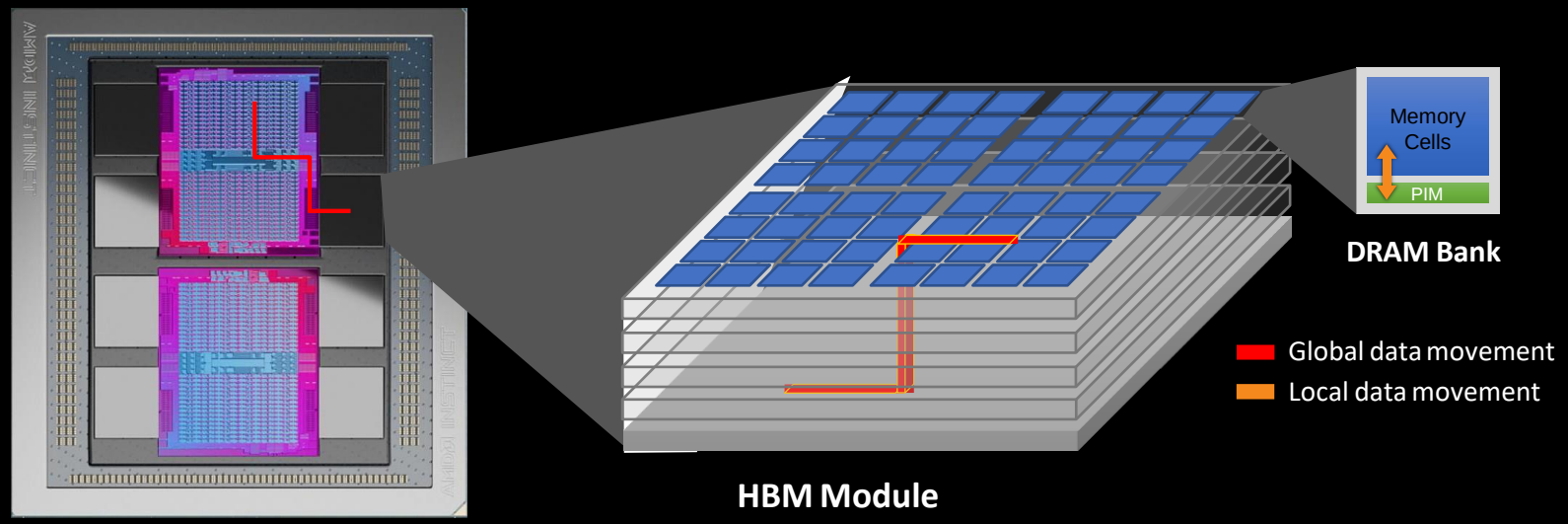


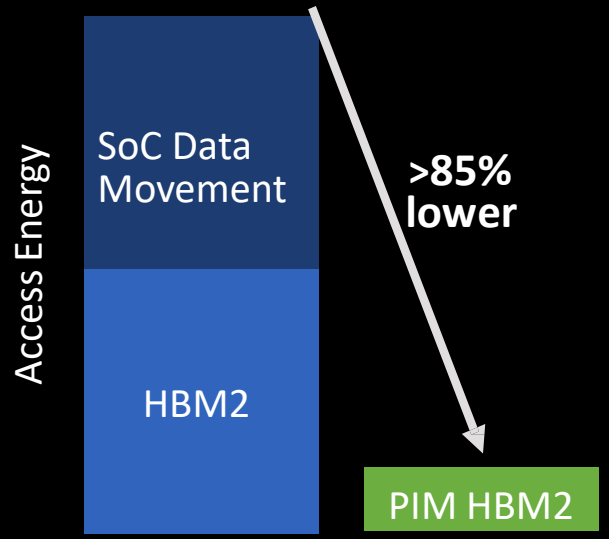




Processing in Memory

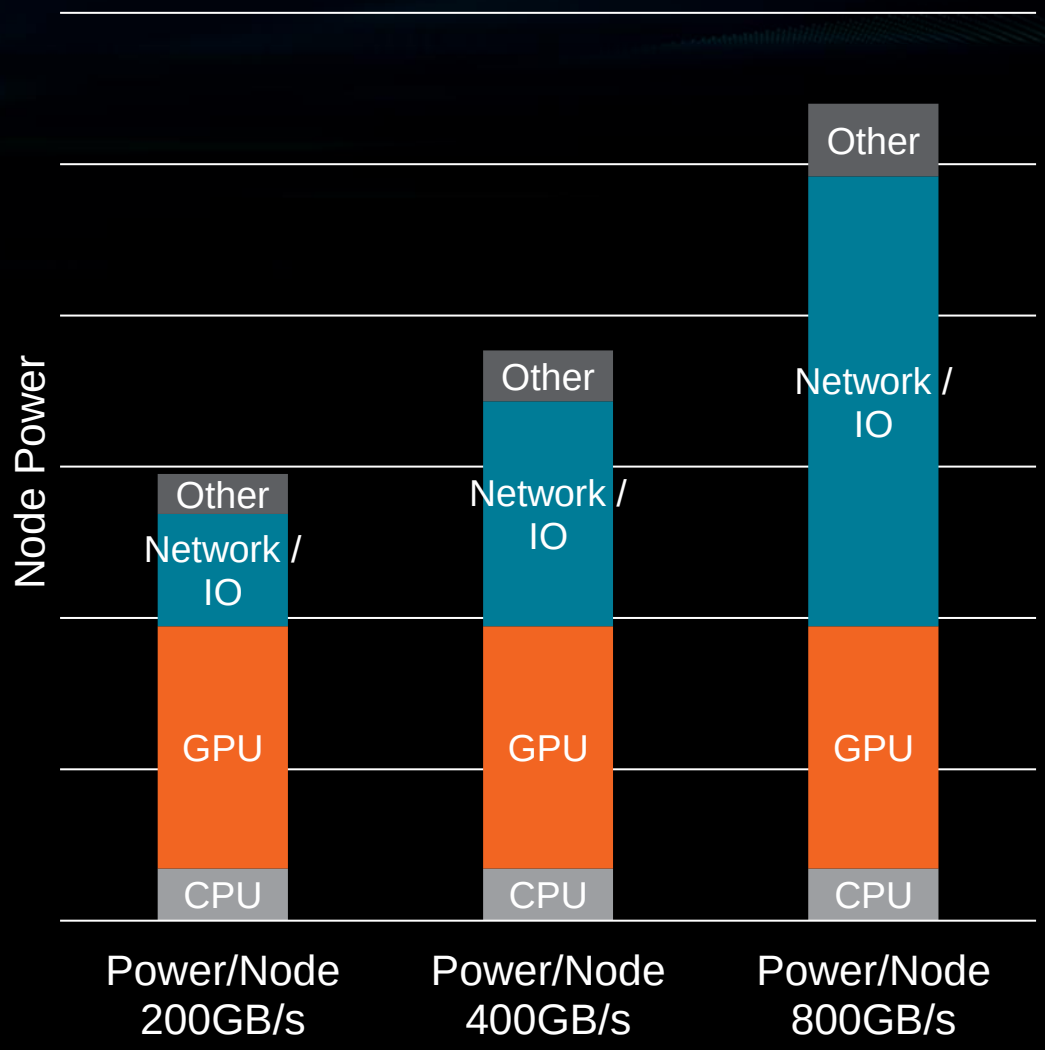
- Key algorithmic kernels can be executed directly in memory, saving precious communication energy



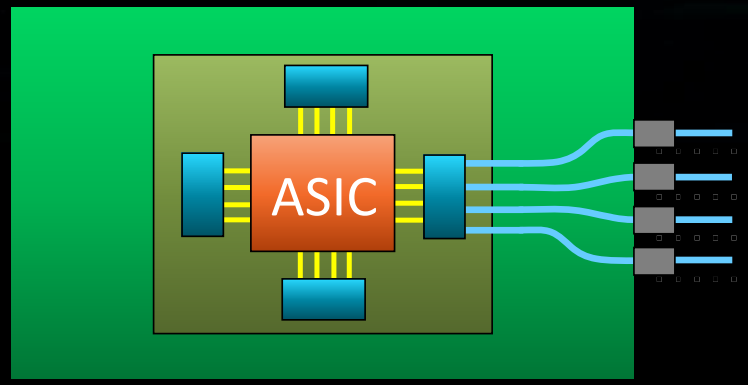


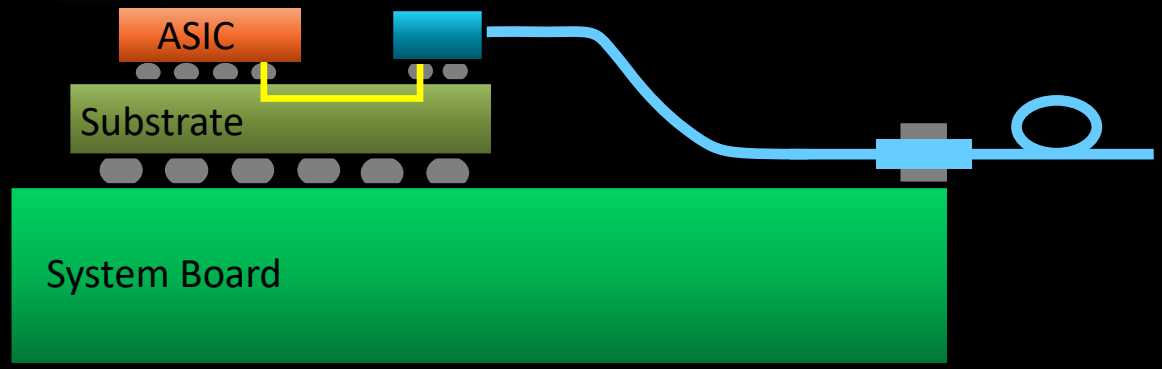
System Power by Function

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



Optical Communication for Energy Efficient Networks





Co-packaged optics can provide a path forward

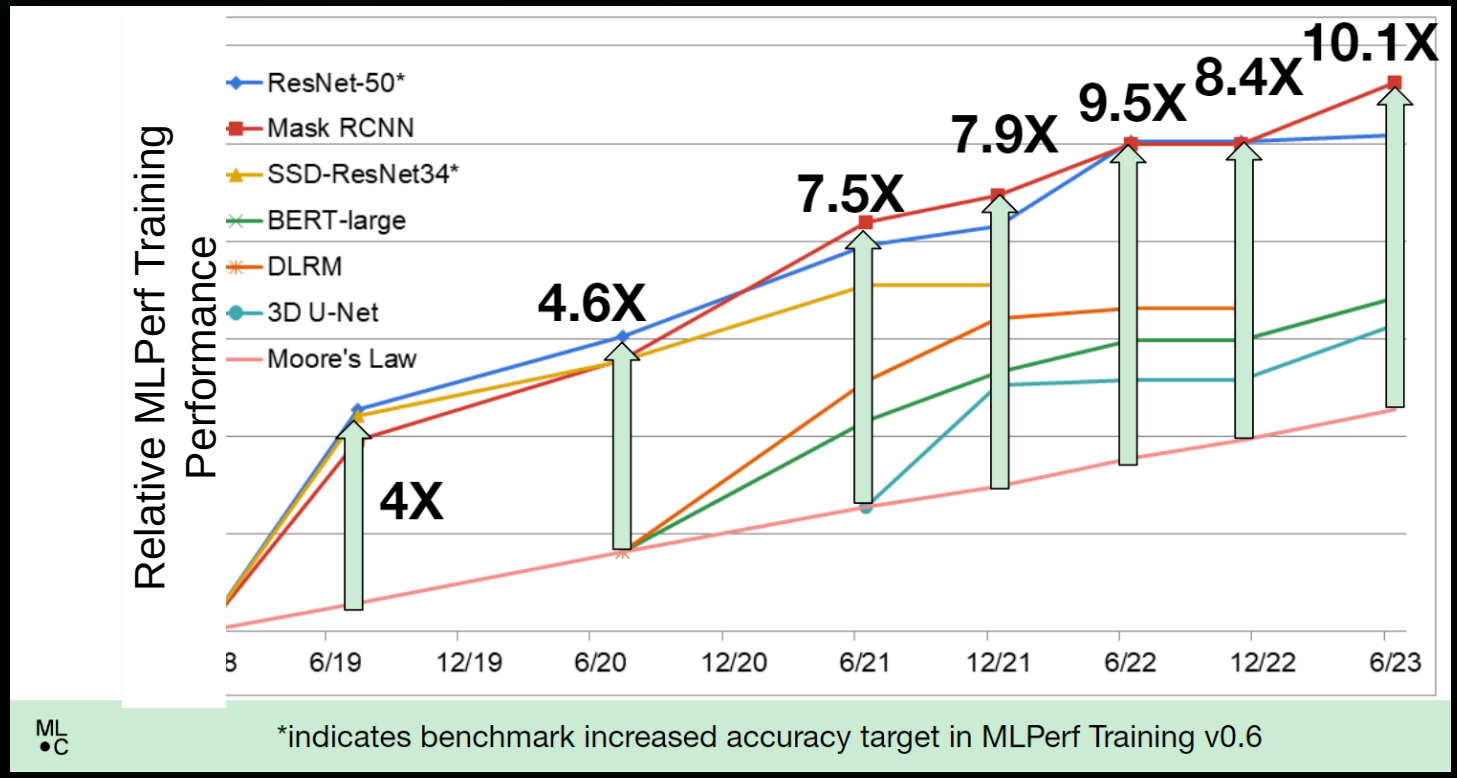
Reach and BW density reduces switch and re-timer power

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

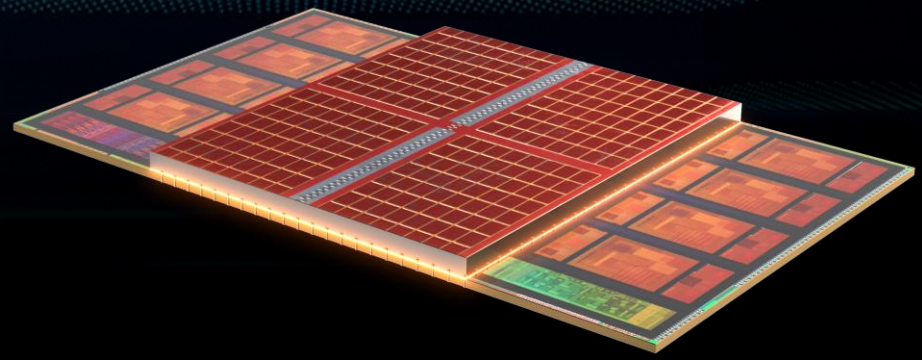
Tight integration of optical transceivers to compute die is a key to efficiency

Software-Hardware Co-Design

- Combination of algorithms and architecture have been and will continue to be a critical lever



Meeting the Challenge Requires Holistic Innovation



- Hardware architecture
- Advanced packaging
- New interconnects and memory
- System level integration
- Intelligent management
- And above all, algorithm-software-hardware co-design

17

Final Thought

100+ MW

5,000,000

20W

Copyright and Disclaimer


AMD