



Energy-Efficient Computing for AI and HPC

Michael Schulte
Senior Fellow, AMD Research

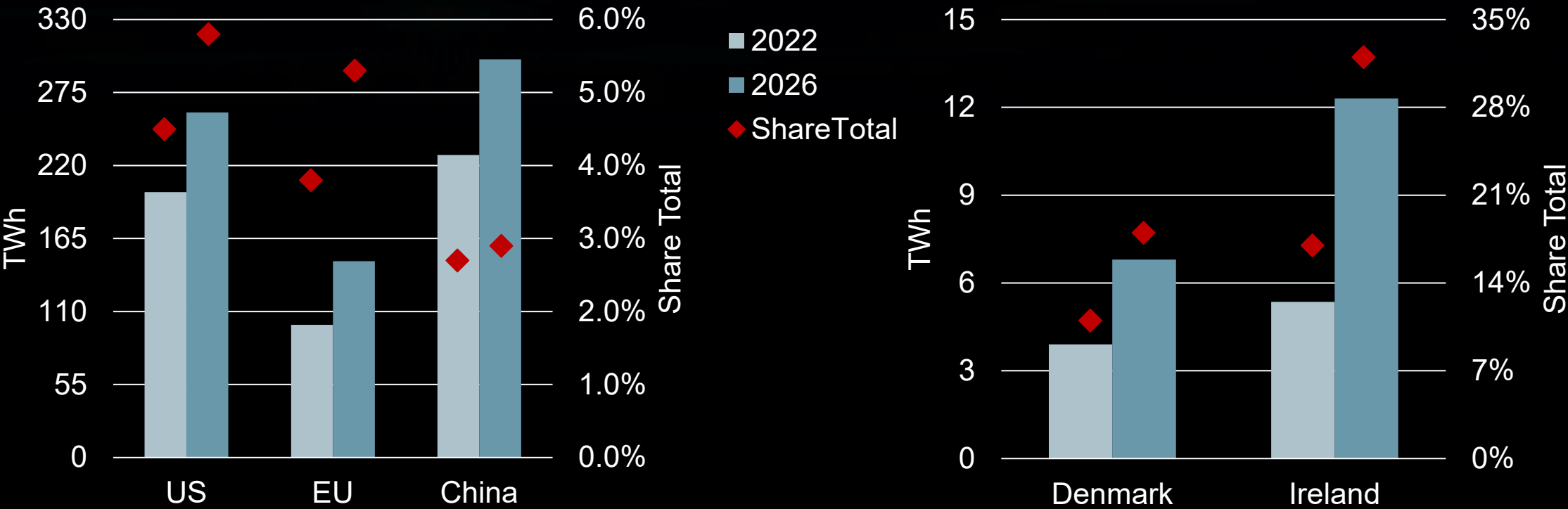
S0S27
May 18-20, 2025

AMD 
together we advance_

The Problem



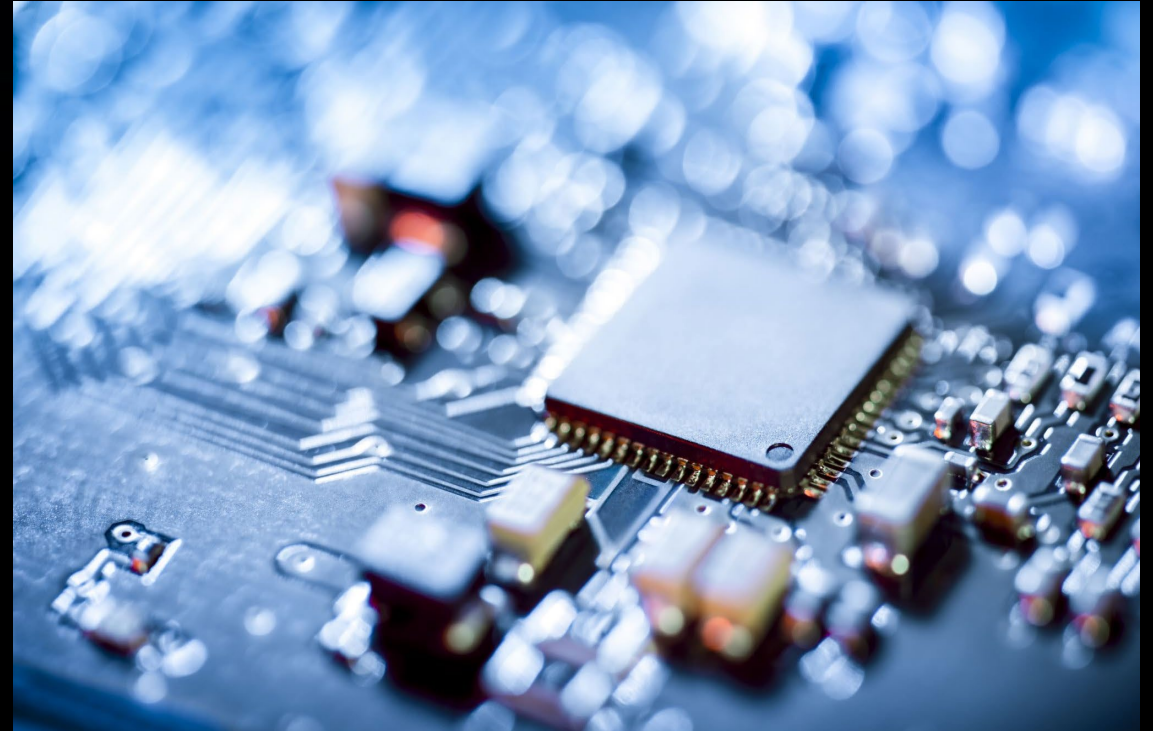
Data Center Energy Consumption



How We Got Here – 5 Decades of Innovations



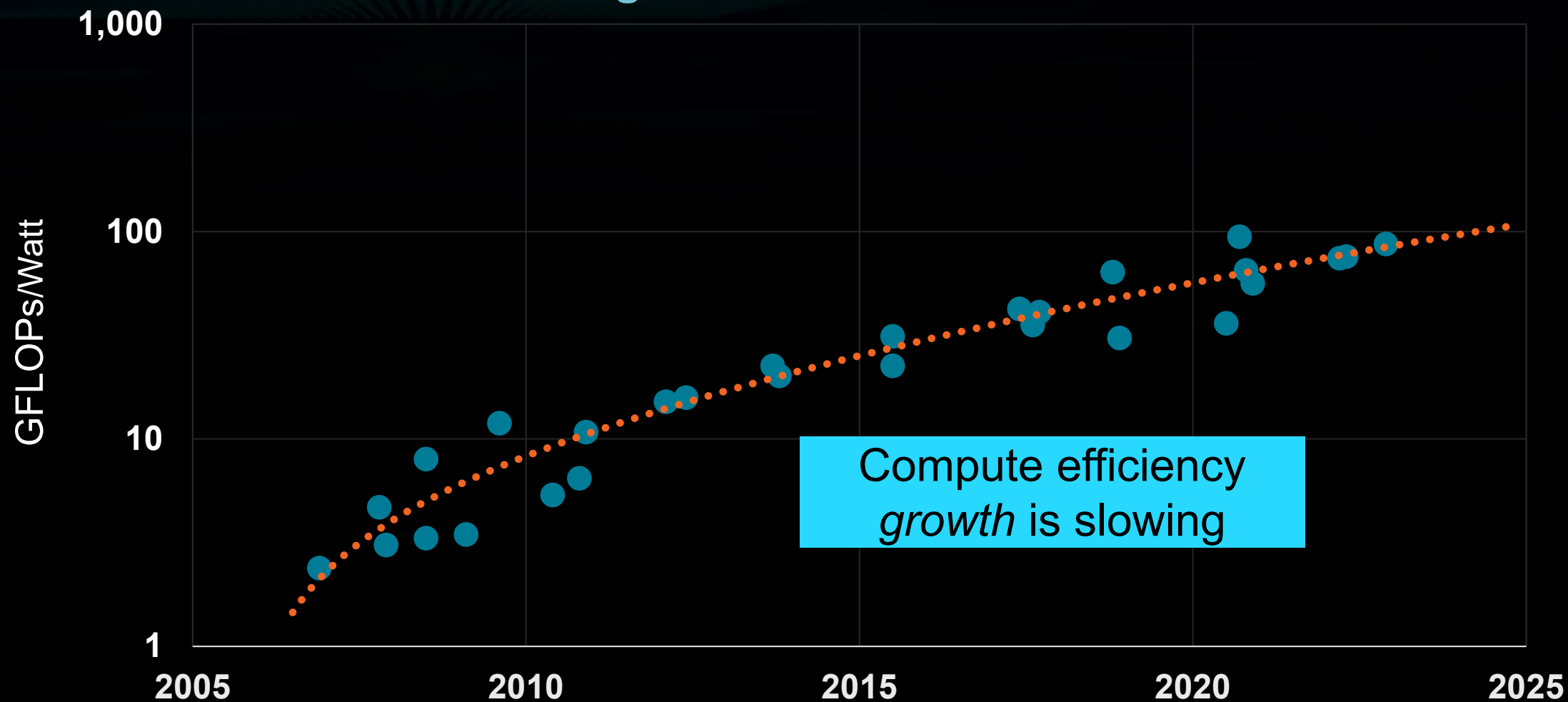
> 2000 times higher frequency
Supercomputers are ~17 billion times faster



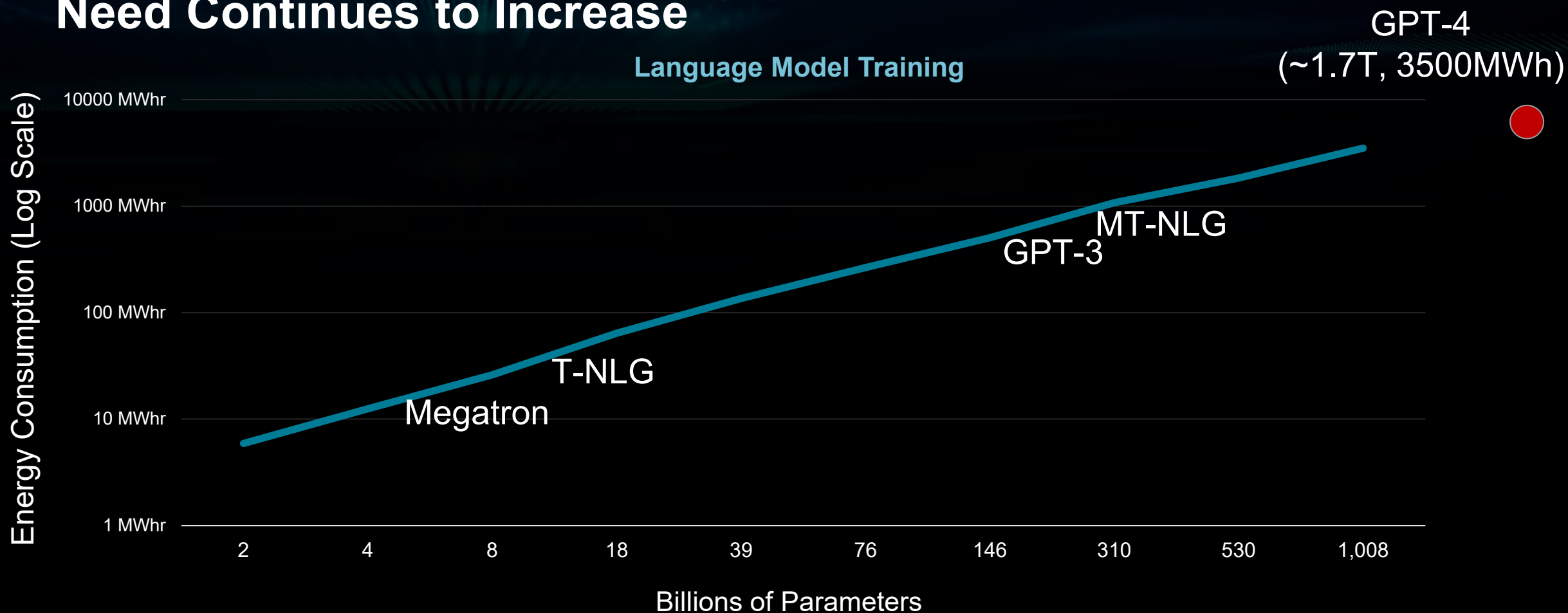
> 3000 times smaller

Trends – GPU Power Efficiency

GPU Single Precision FLOPs/Watt



Need Continues to Increase

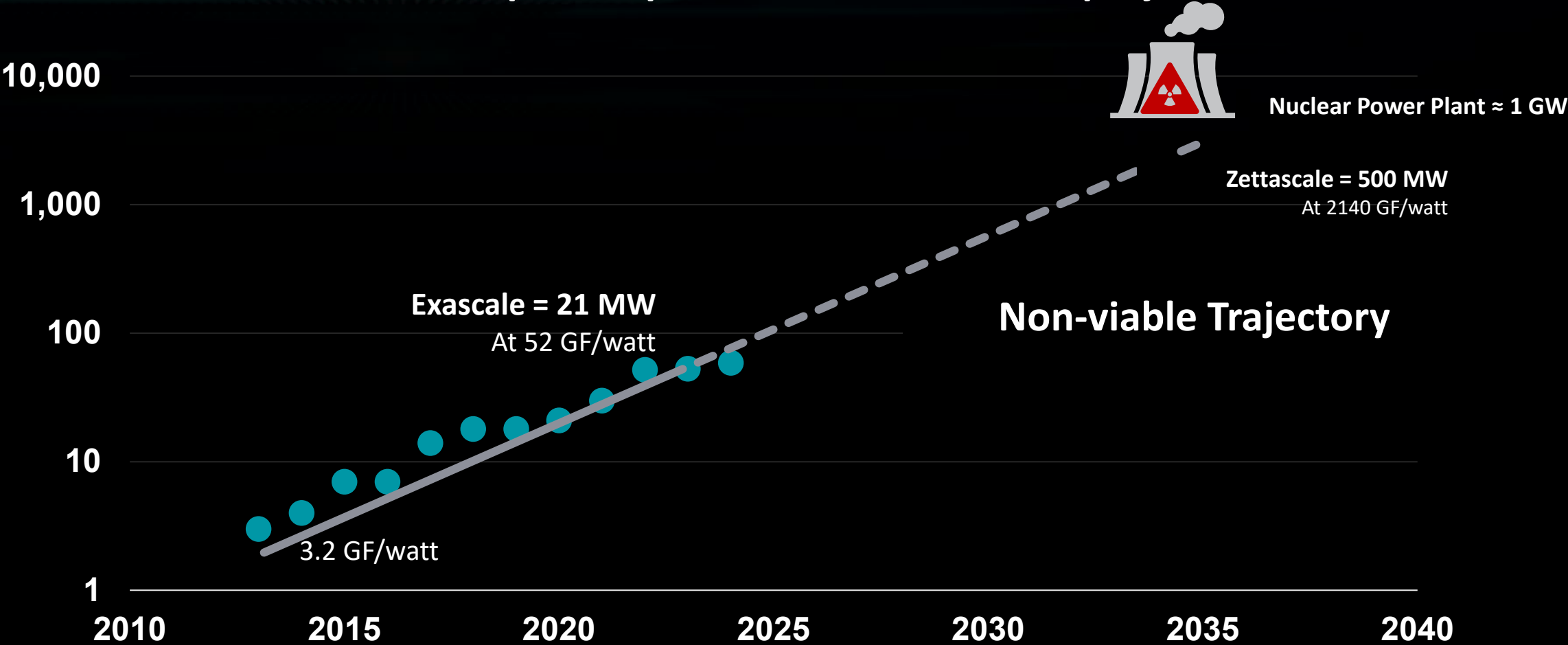


Exponentially growing model sizes drive immense growth in energy for training.

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

Supercomputer Energy Use Trajectory

Green500 supercomputer GFLOPs/watt and projection



The Opportunity



Domain-specific computation enables compute efficiency

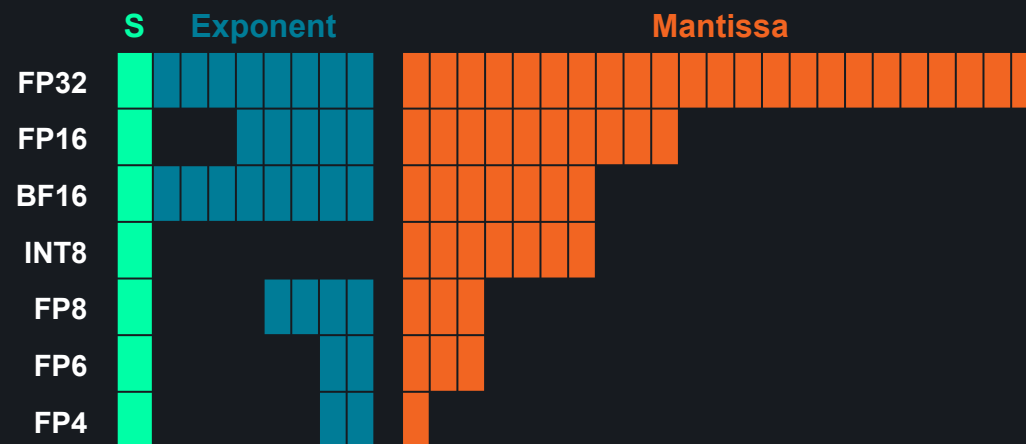
Tailor architecture for application

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

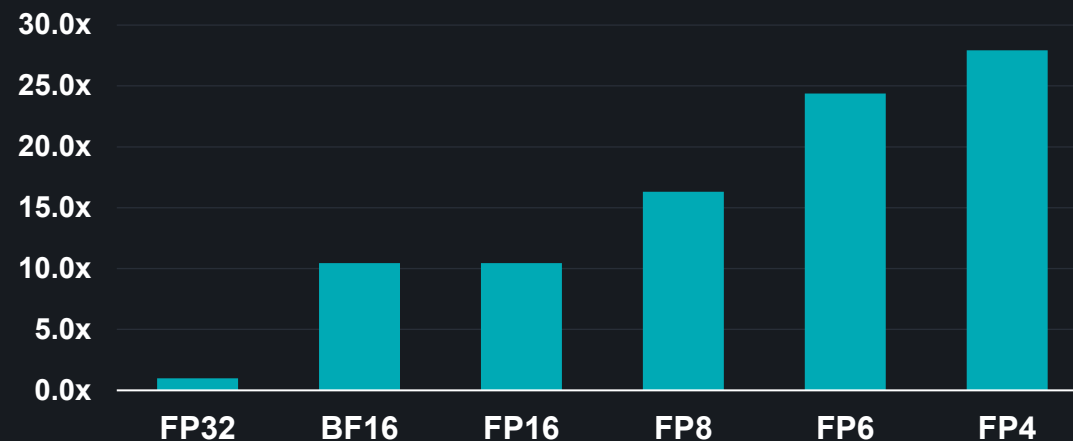
Need to determine appropriate precisions

Based on AMD internal calculations

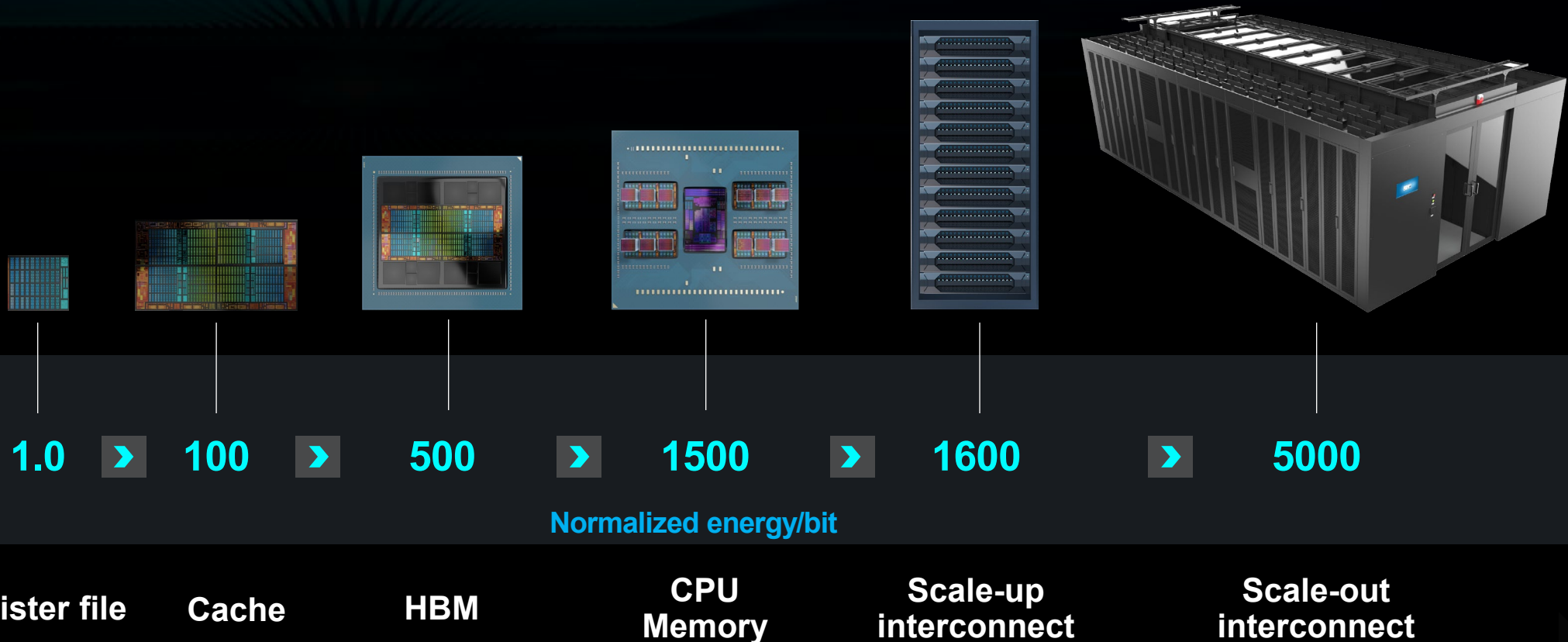
Bit allocations by format



FLOPs/Joule normalized to FP32



Reducing Data Movement Energy

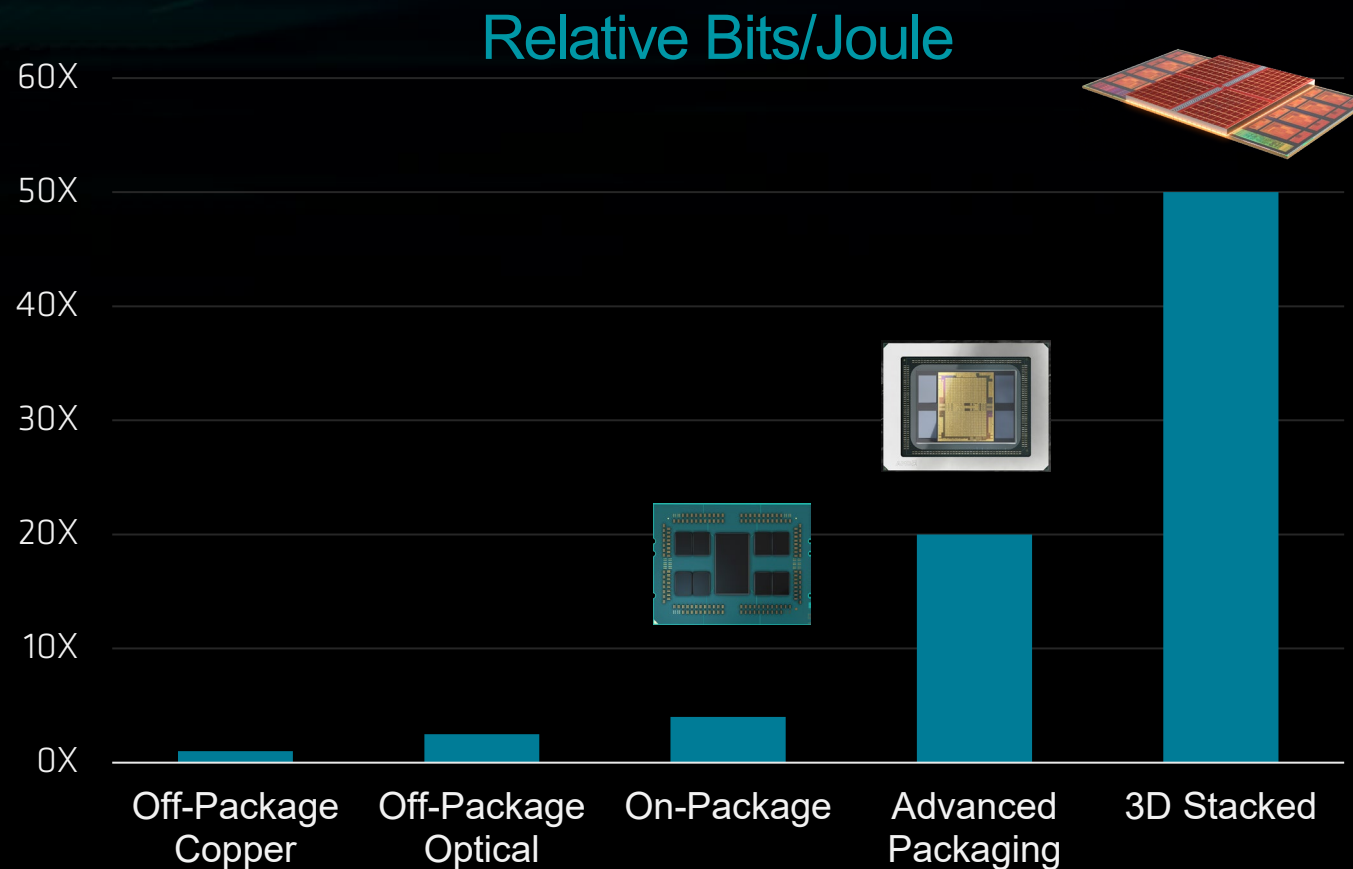
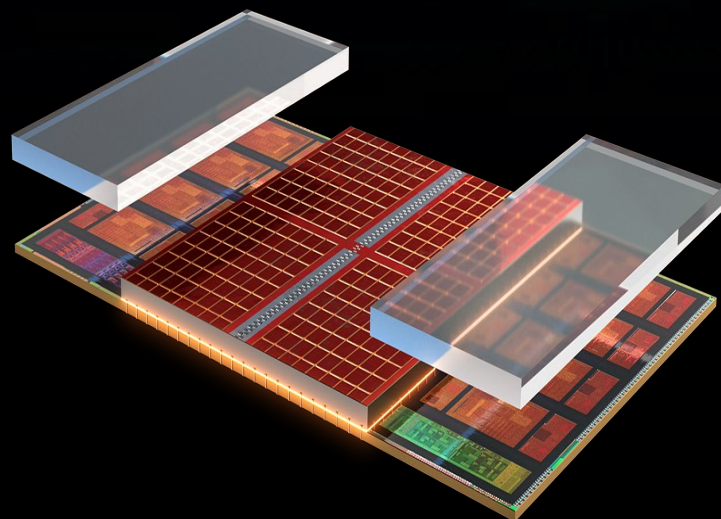


Data movement energy increases with distance



Maximizing locality is key to efficiency

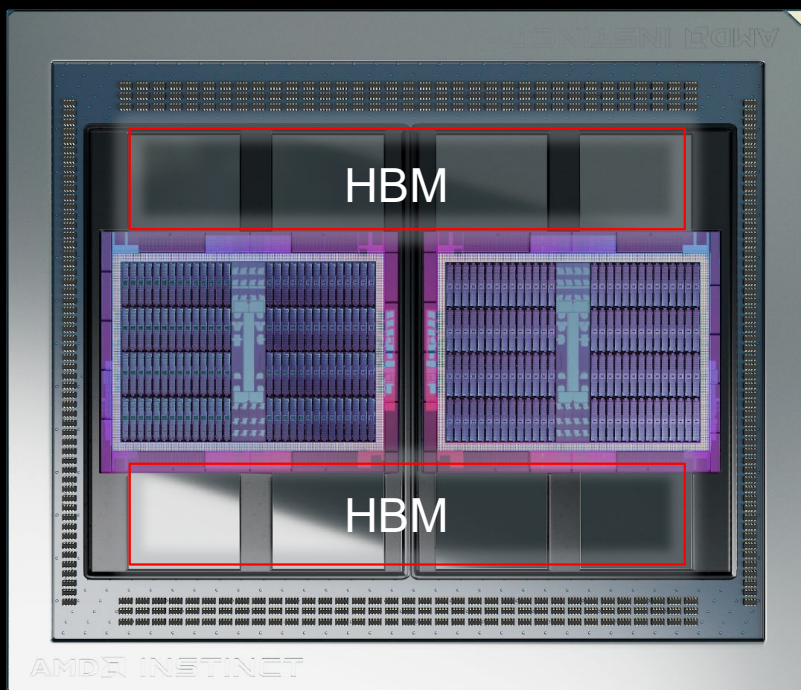
3D Chiplets and Communication Energy



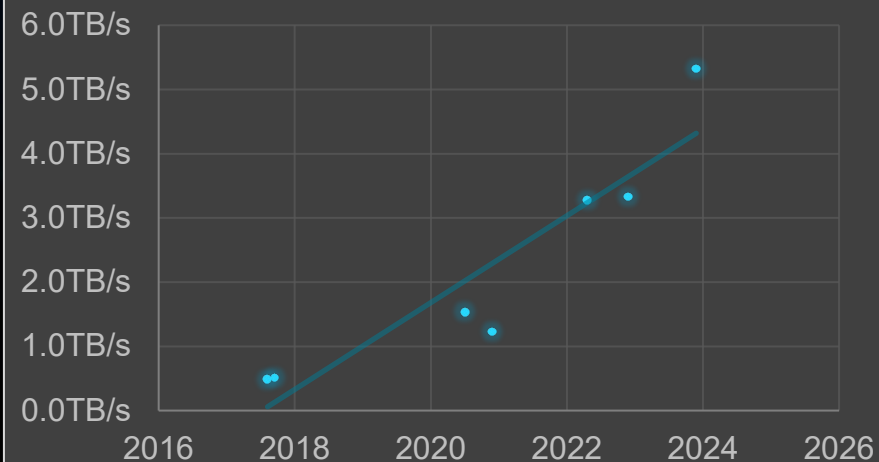
Advanced Packaging Provides up to a 50x Reduction in Communication Power

Thirst for Memory Bandwidth

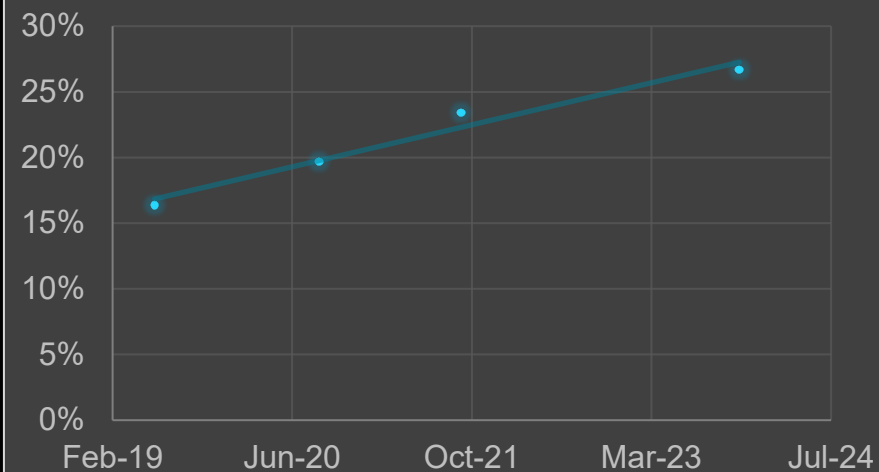
- High bandwidth memory feeds the compute engine providing a key element of performance gains
- Limited efficiency gains combine with demand growth result in higher percentage of power for memory



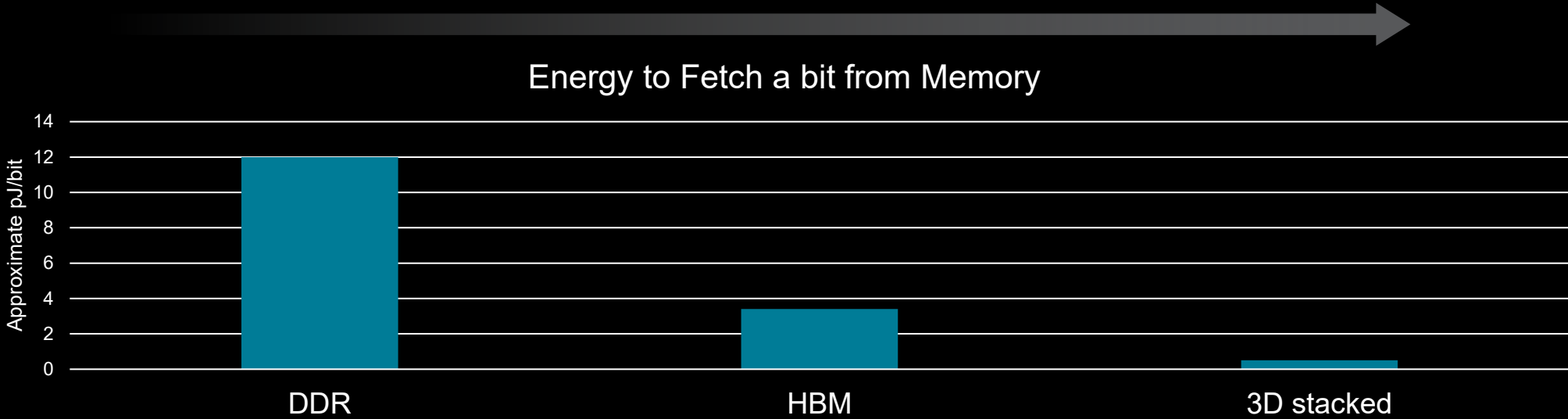
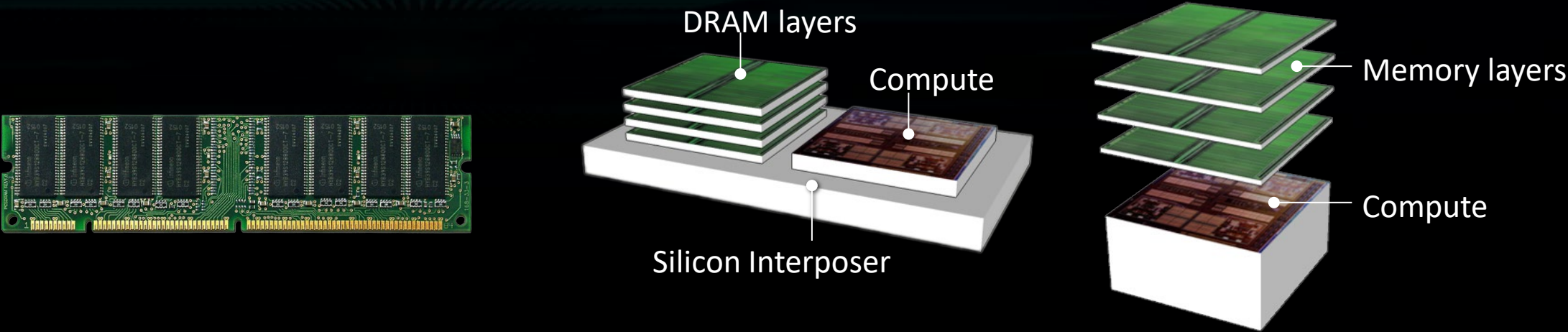
Compute GPU Mem BW



GPU Accelerator Memory Power Percentage

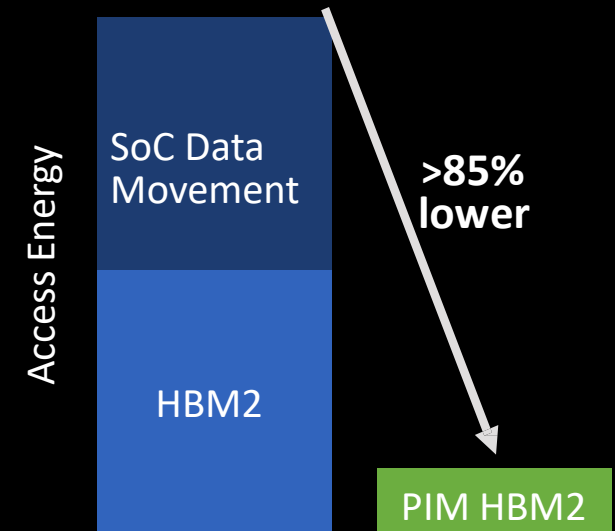
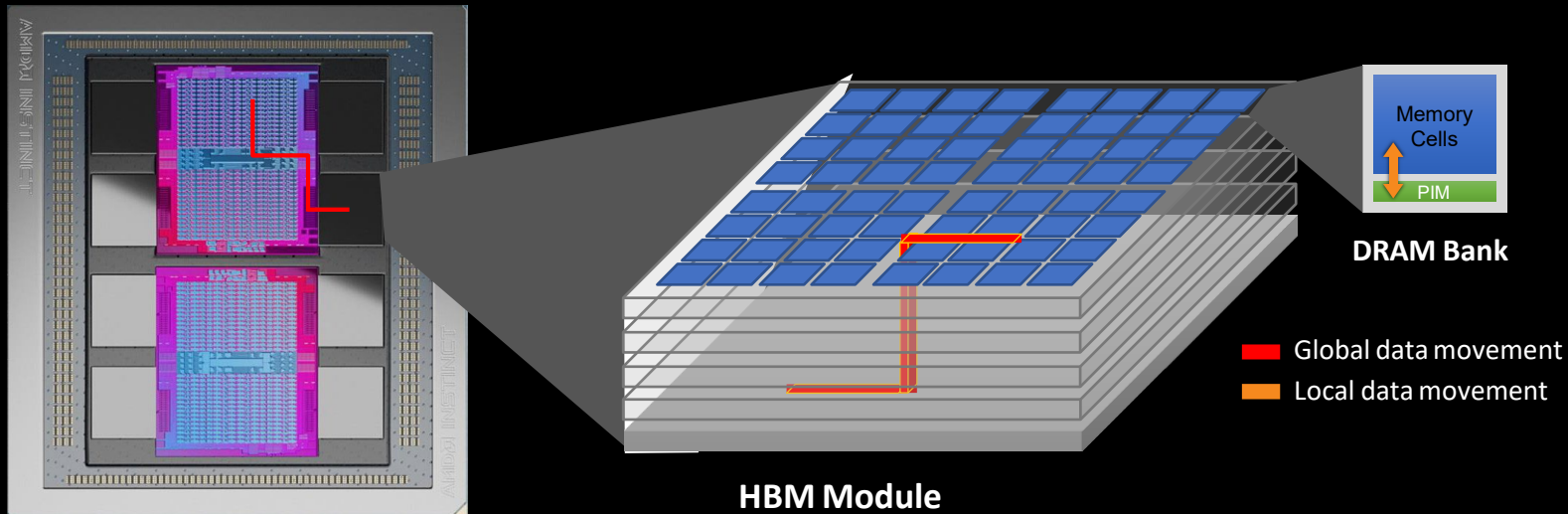


Reducing Memory Energy



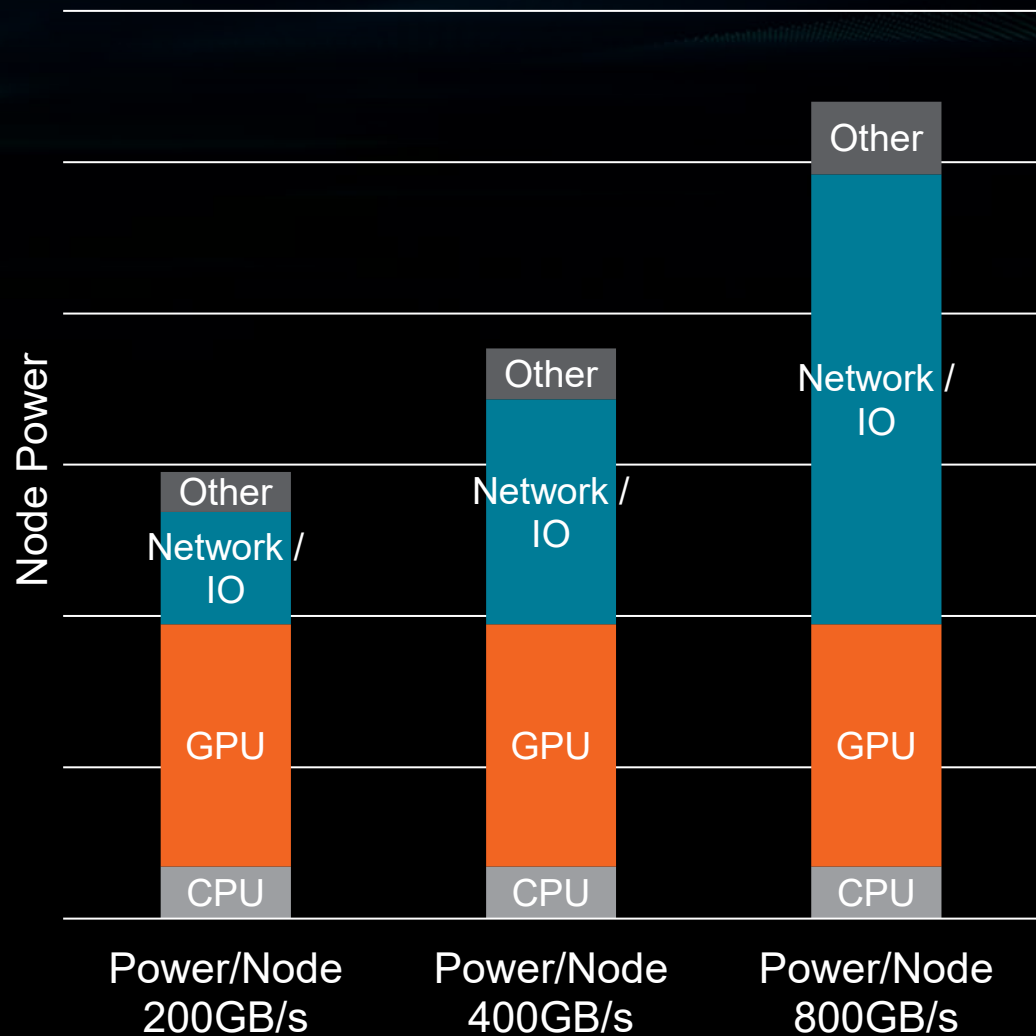
Processing in Memory

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

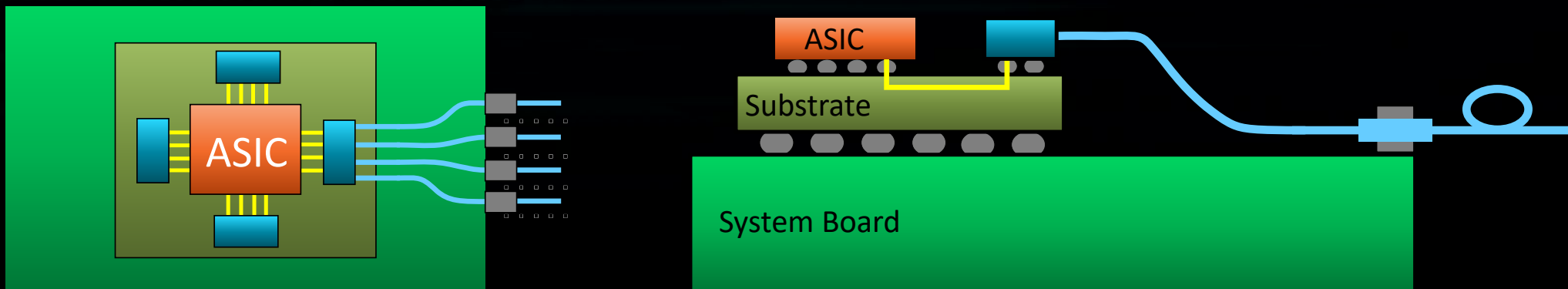


System Power by Function

- 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

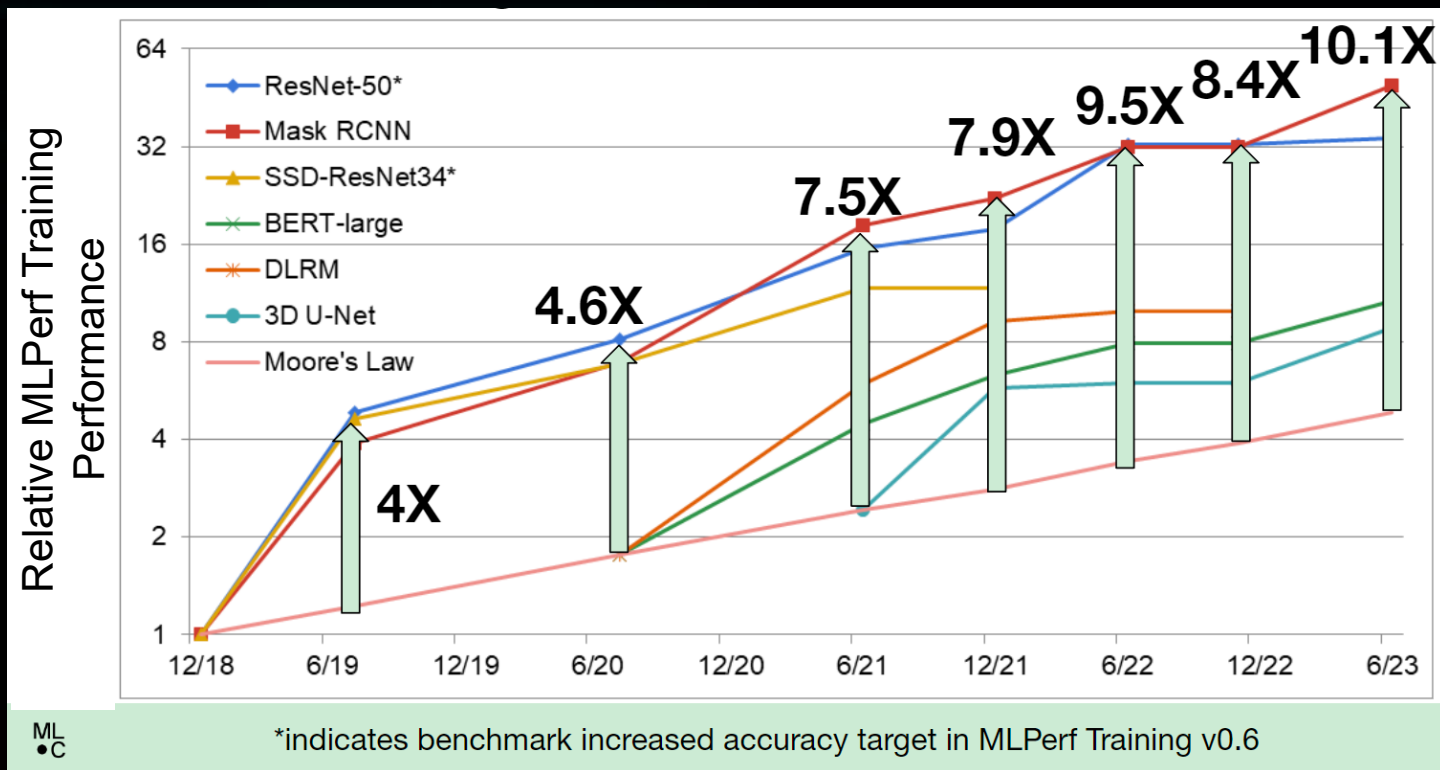
Reach and BW
density reduces
switch and re-timer
power

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

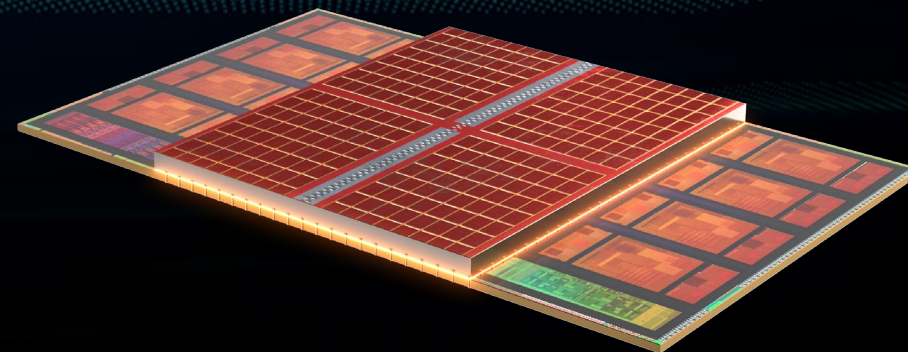
Algorithm-Software-Hardware Co-Design

- 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

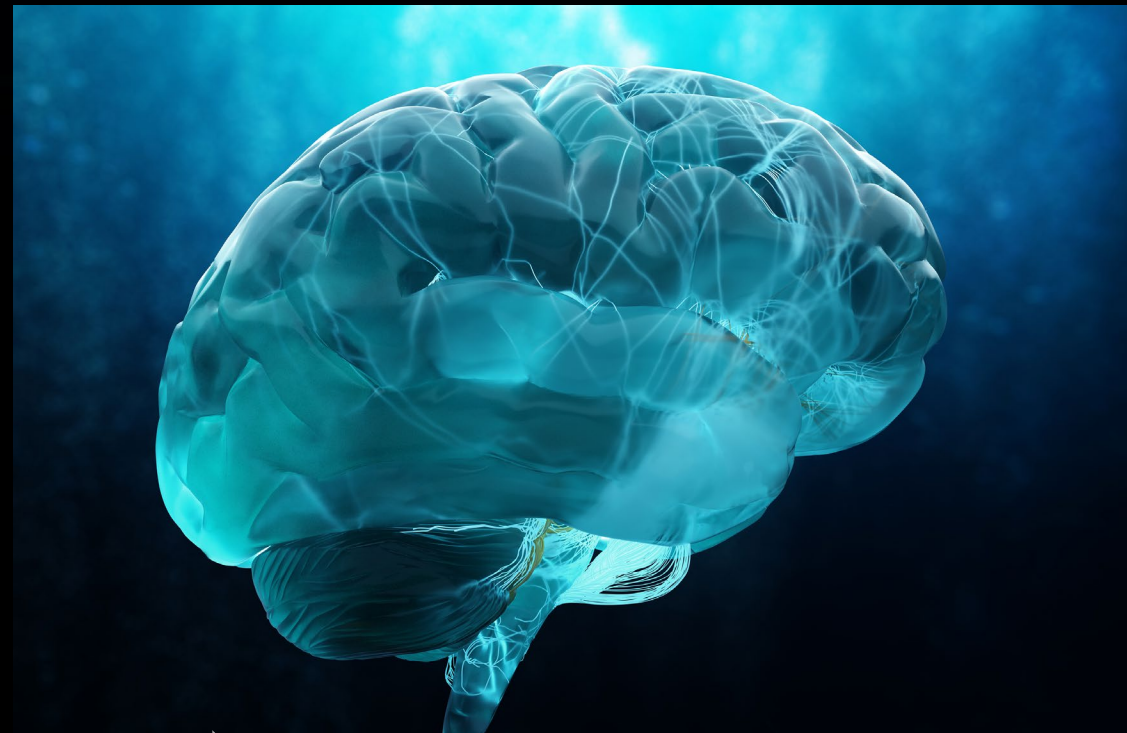
- Hardware architecture
- Advanced packaging
- New interconnects and memory
- System level integration
- SW optimizations
- Intelligent design and management
- Algorithm-software-hardware co-design



Final Thought



100+ MW



5,000,000

20W



Copyright and Disclaimer

The information contained herein is for informational purposes only, and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD's products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18

© 2025 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, Ryzen, EPYC, Instinct, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.