



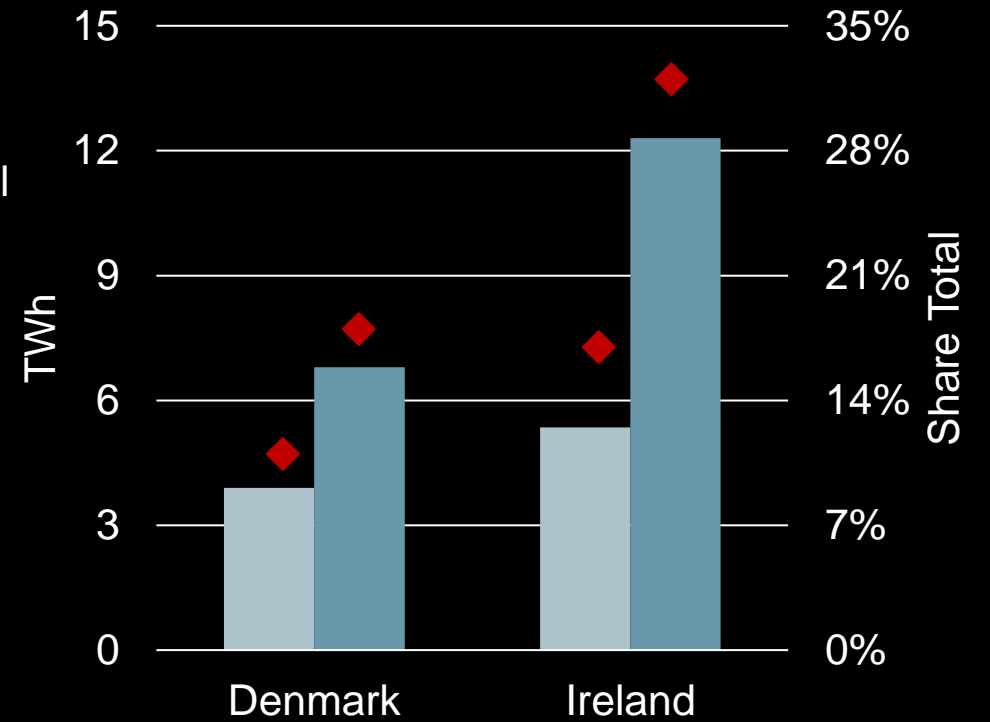
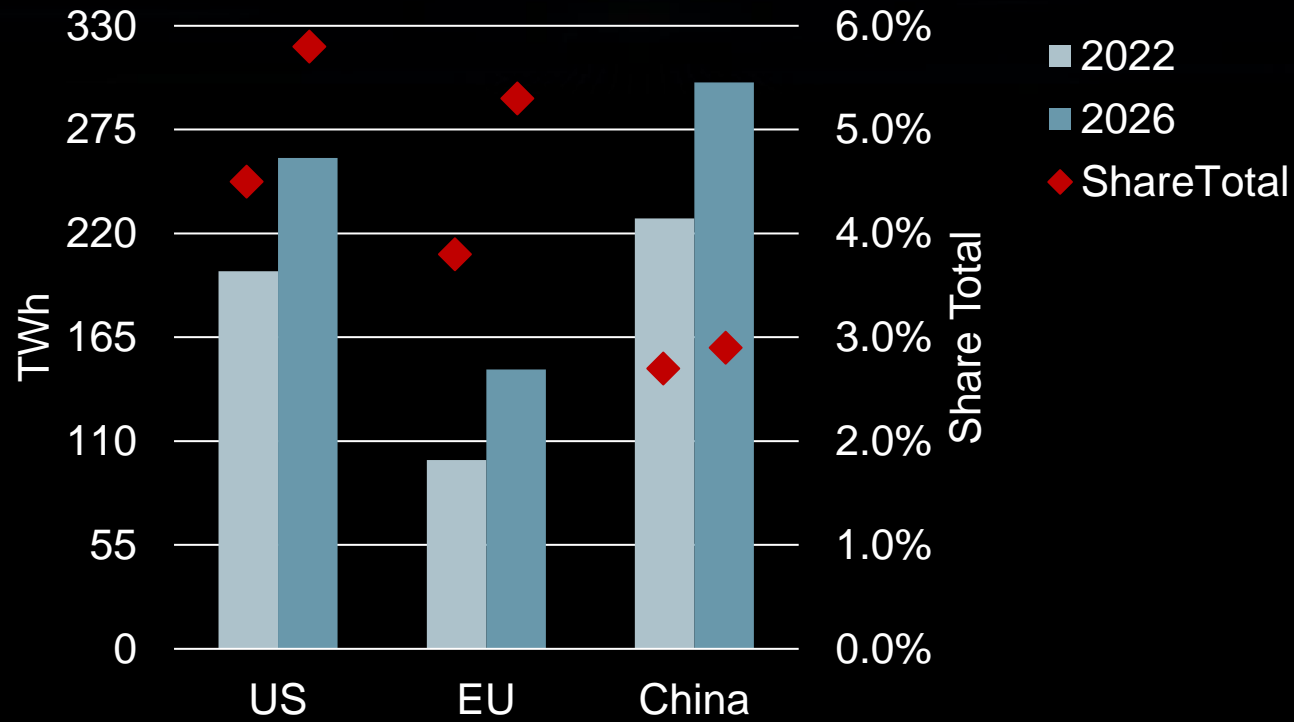
# Energy Efficient Computing for Science

Srilatha (Bobbie) Manne  
Senior Fellow, AMD

# The Problem



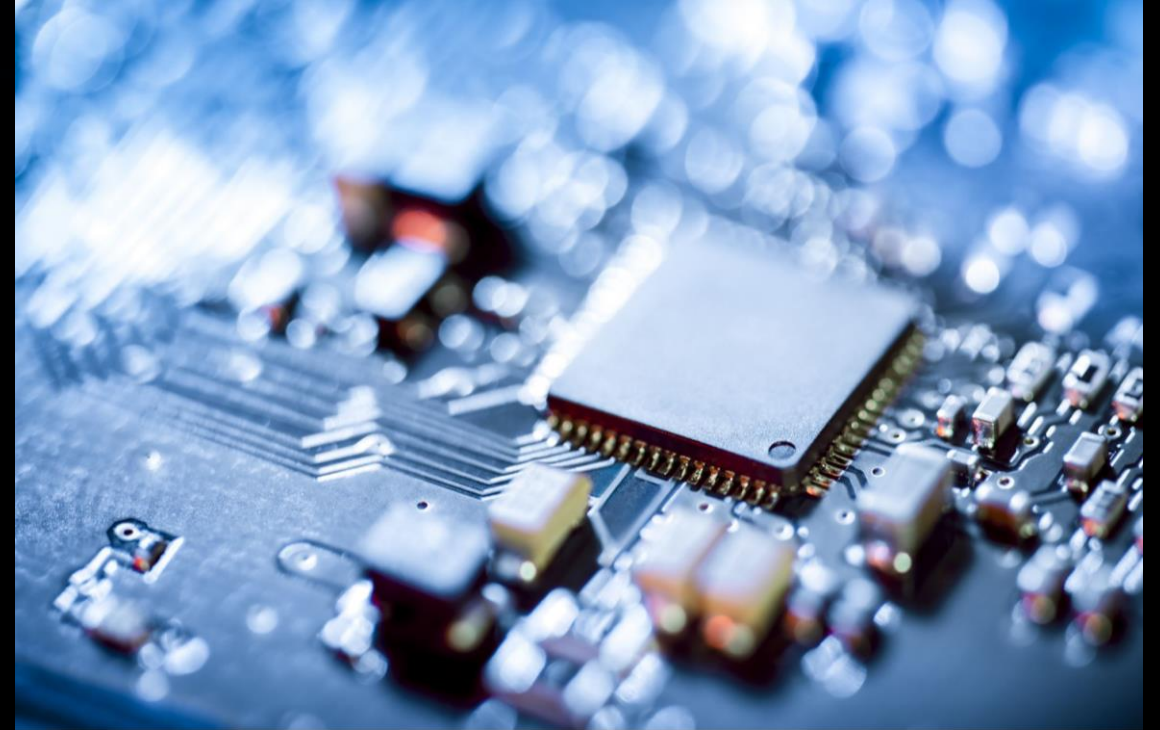
# Where we are



# How we got here – 5 Decades of Innovations



> 2000 times faster



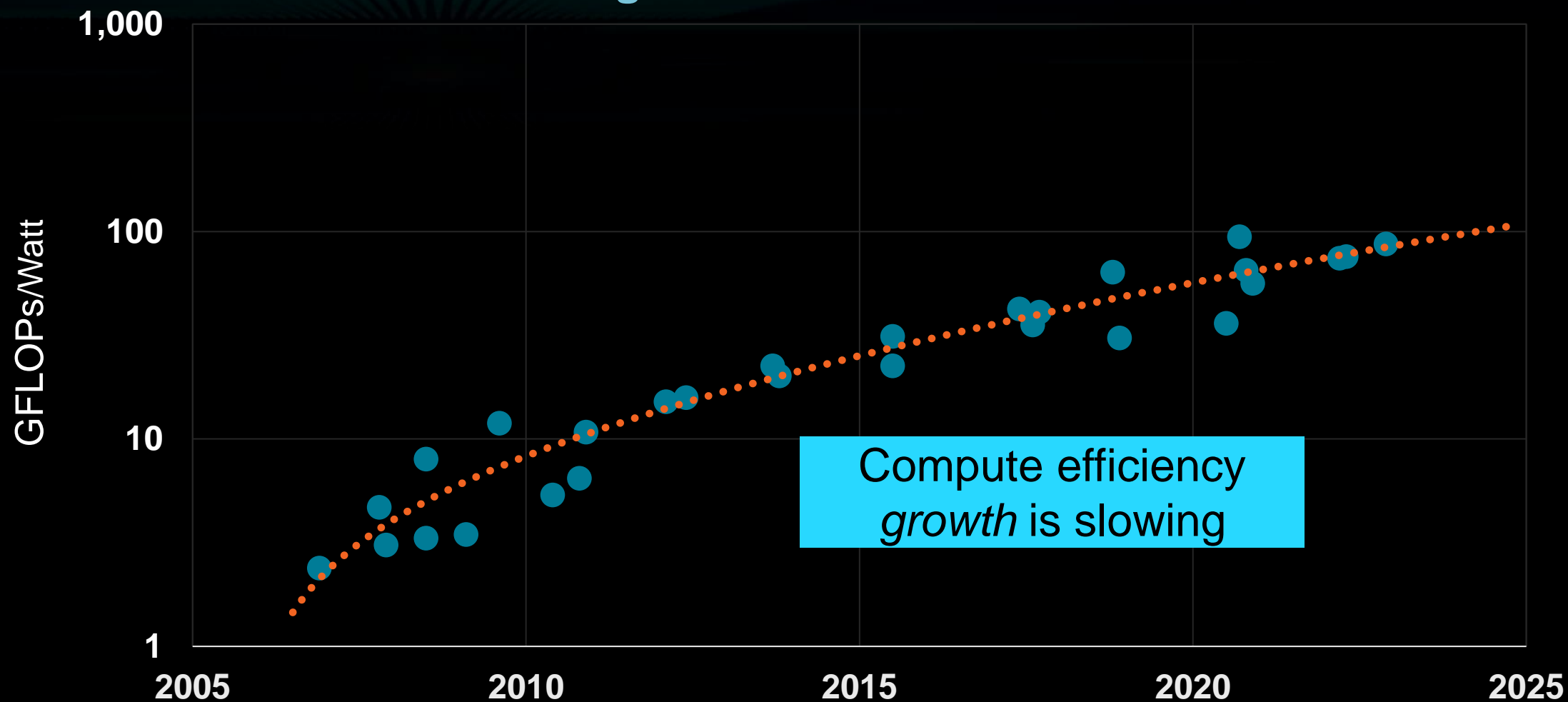
> 3000 times smaller

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

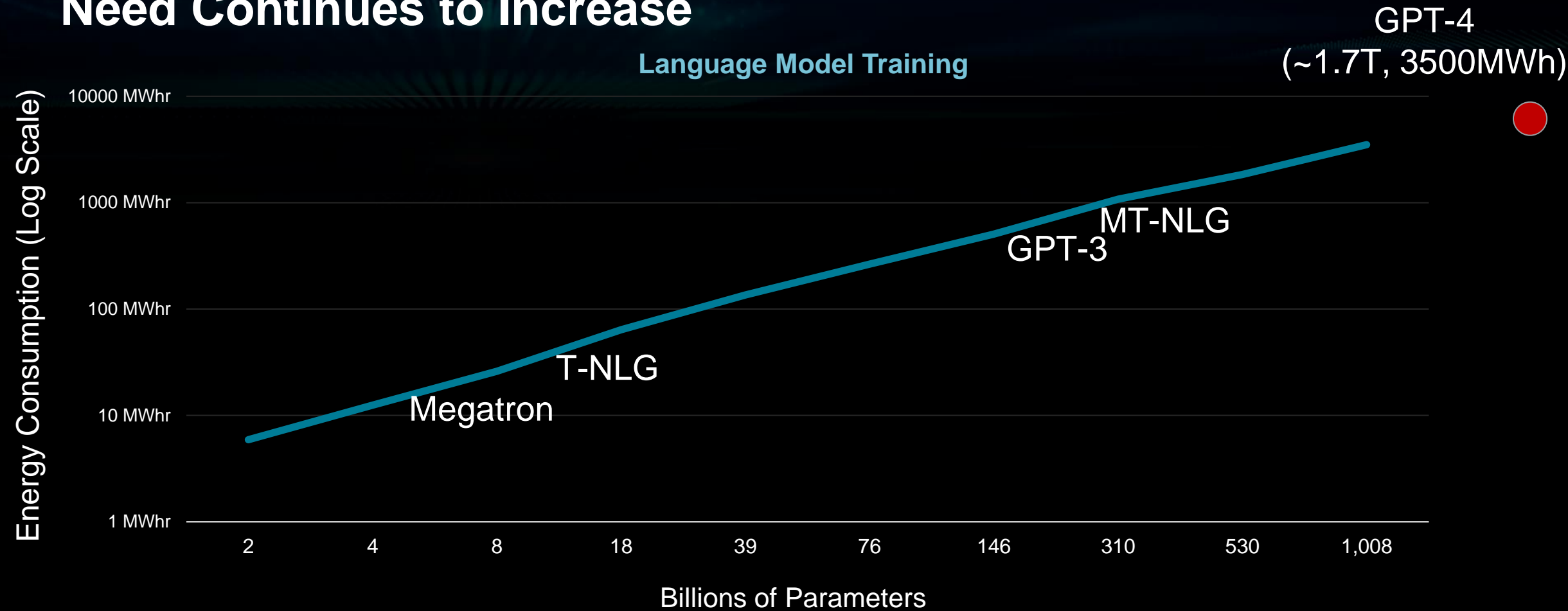


# Trends – GPU Efficiency

## GPU Single Precision FLOPs/Watt



# Need Continues to Increase

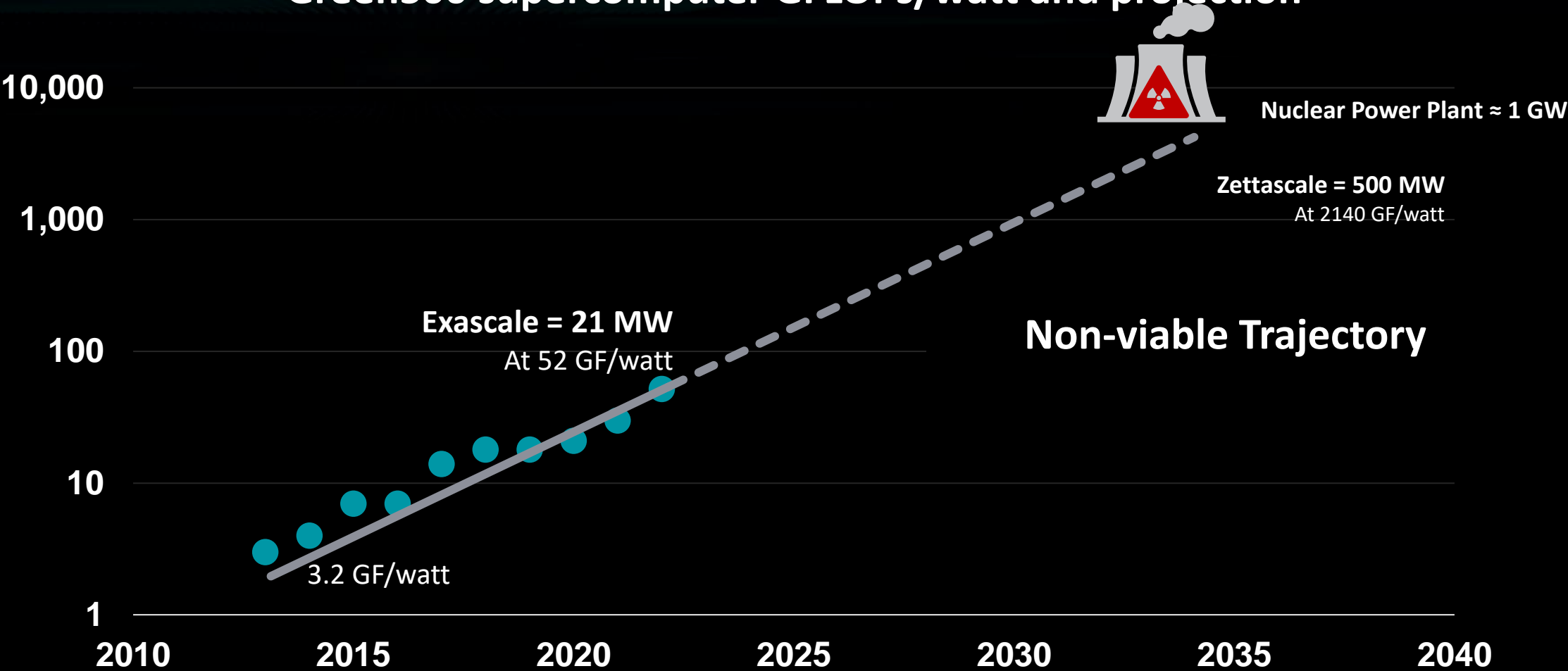


Exponentially growing model sizes drive immense growth in energy for training.

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

# Supercomputer Energy Use Trajectory

Green500 supercomputer GFLOPs/watt and projection



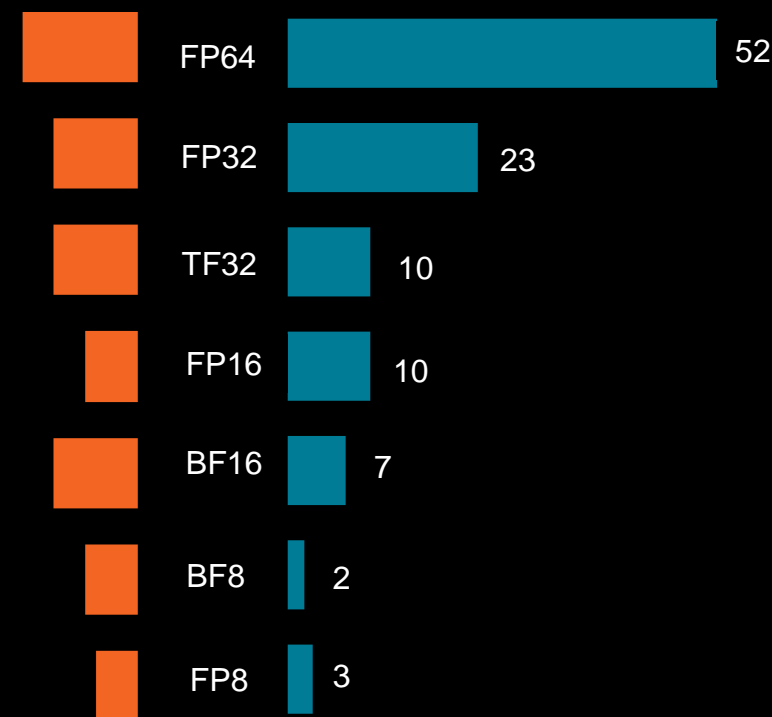
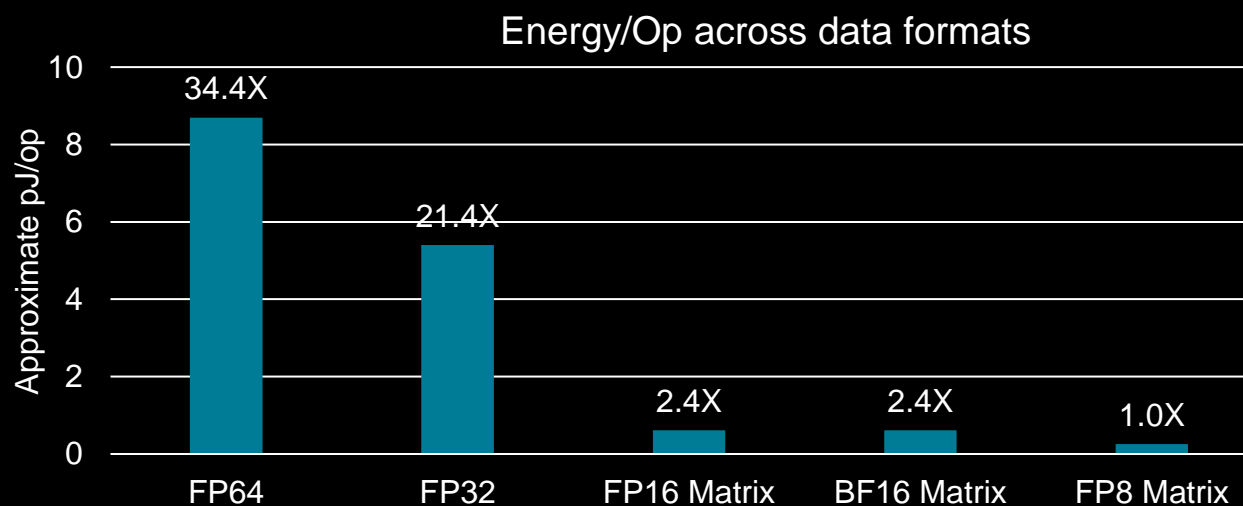
# The Opportunity



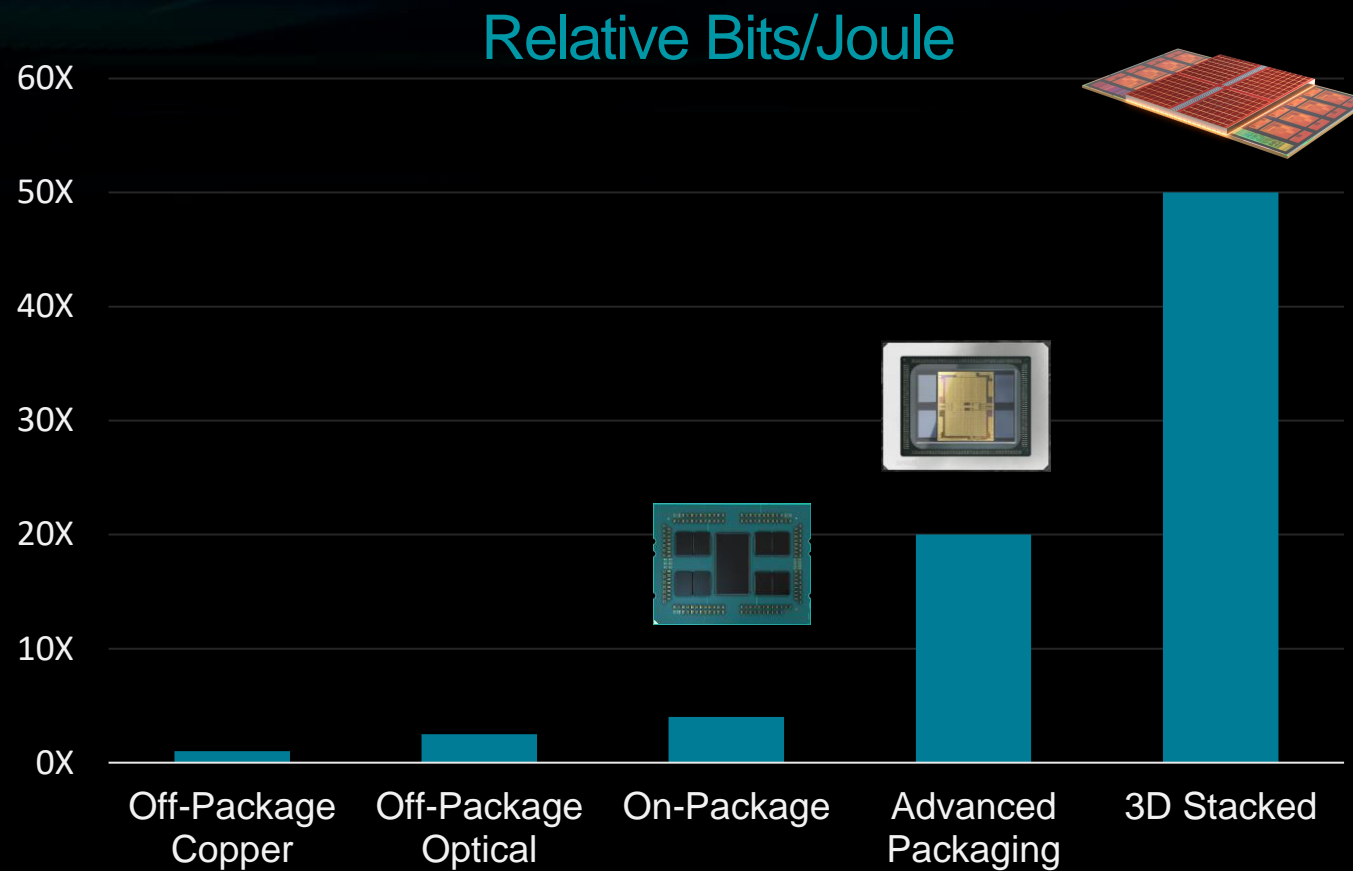
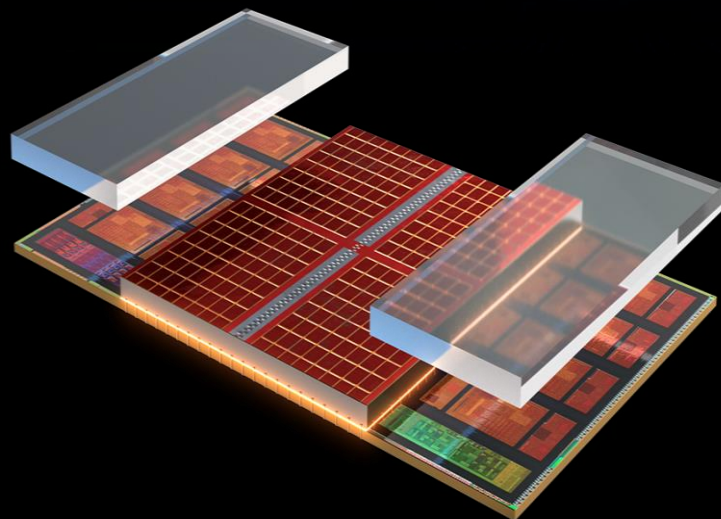


# Flop Power and Reduced Precision

- 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



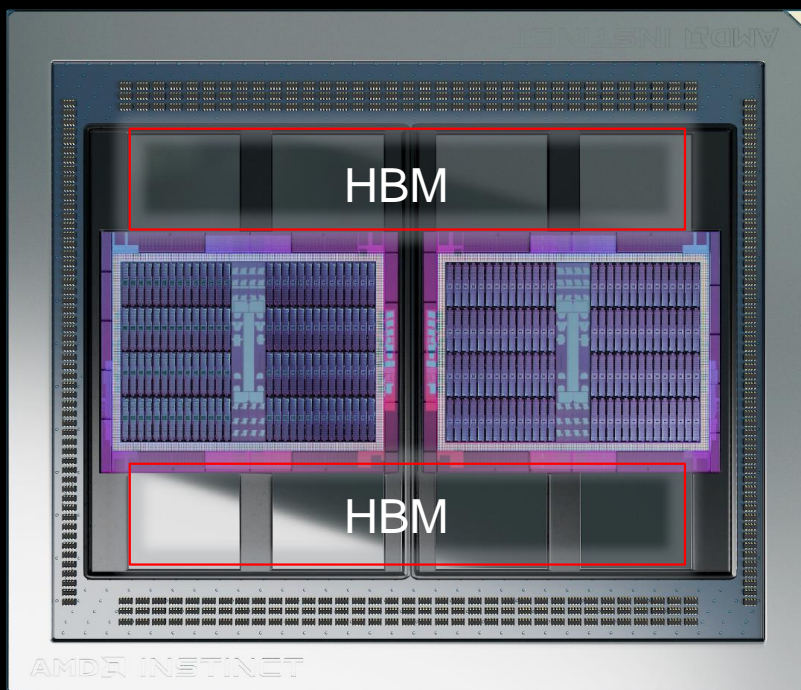
# 3D Chiplets and Communication Power



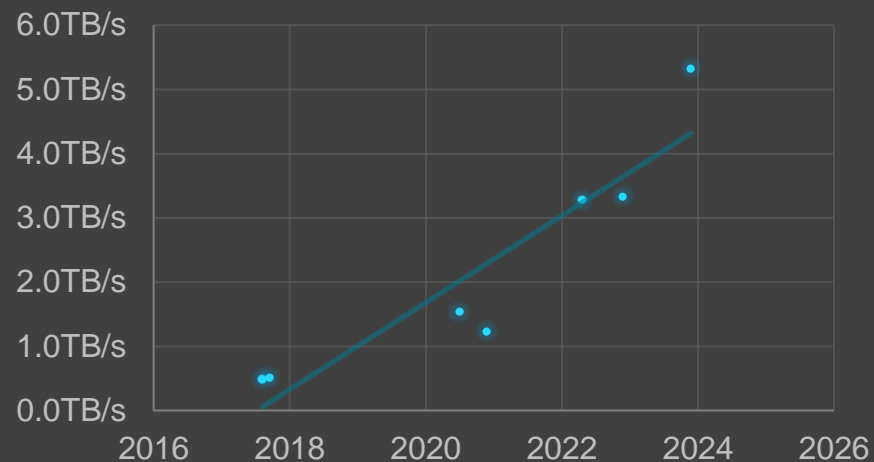
Advanced Packaging Provides up to a 50x Reduction in Communication Power

# Thirst for Memory Bandwidth

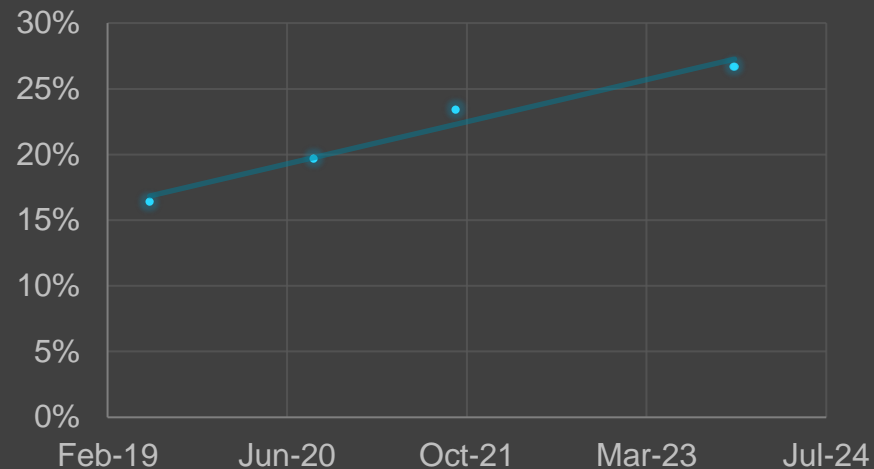
- High bandwidth memory feeds the compute engine providing a key element of performance gains
- Limited efficiency gains combine with demand growth result in higher percentage of power for memory



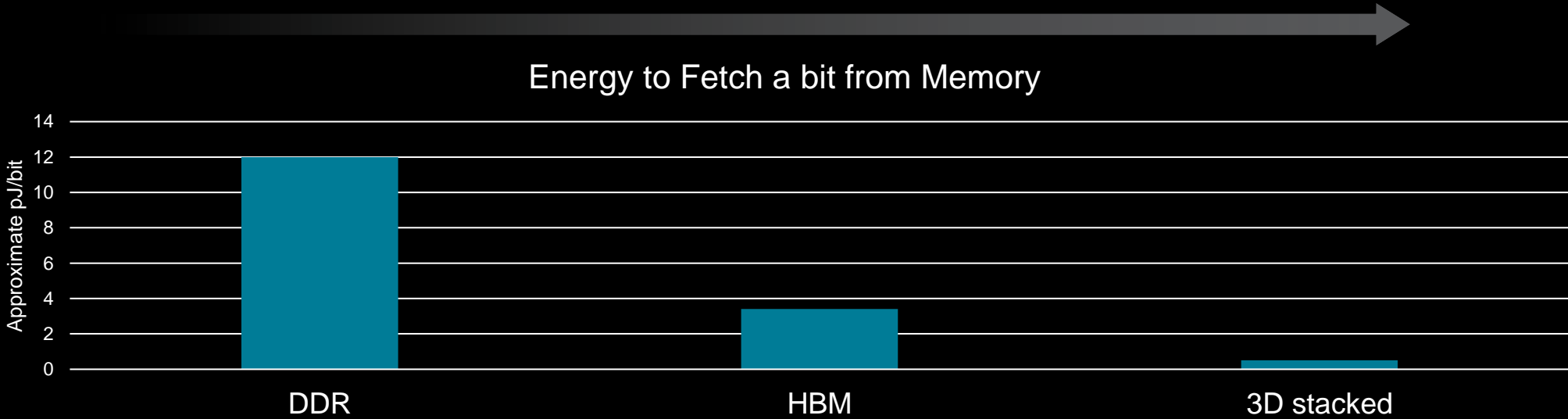
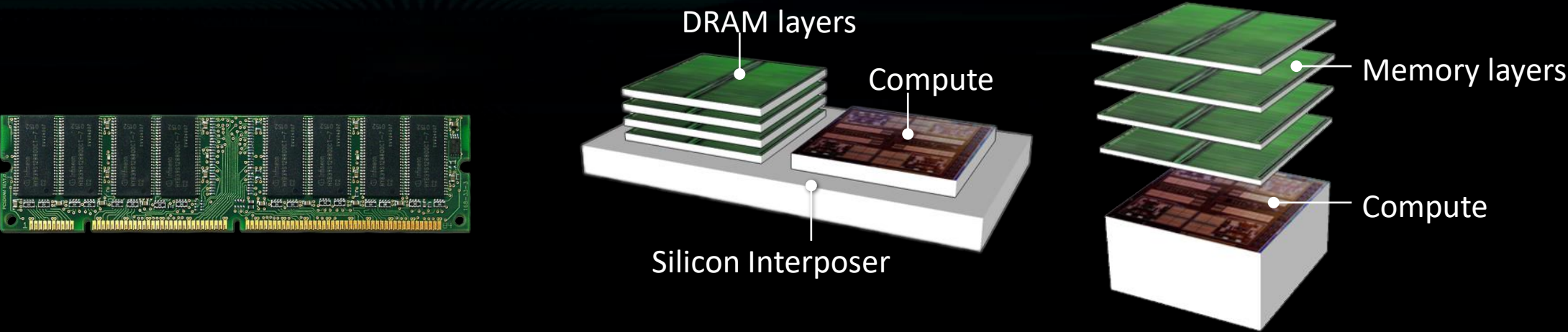
## Compute GPU Mem BW



## GPU Accelerator Memory Power Percentage



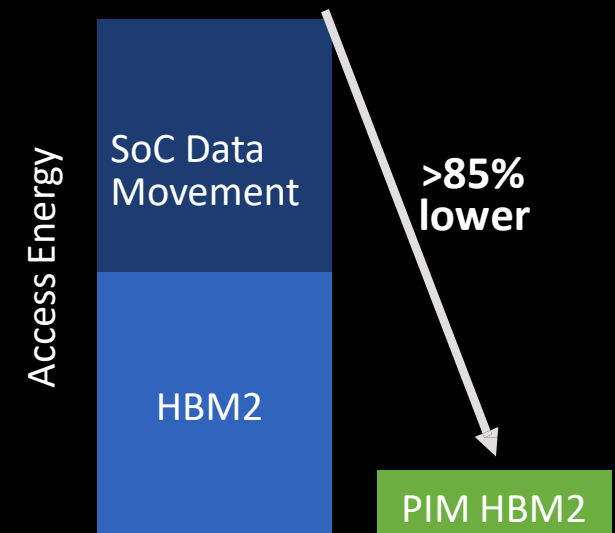
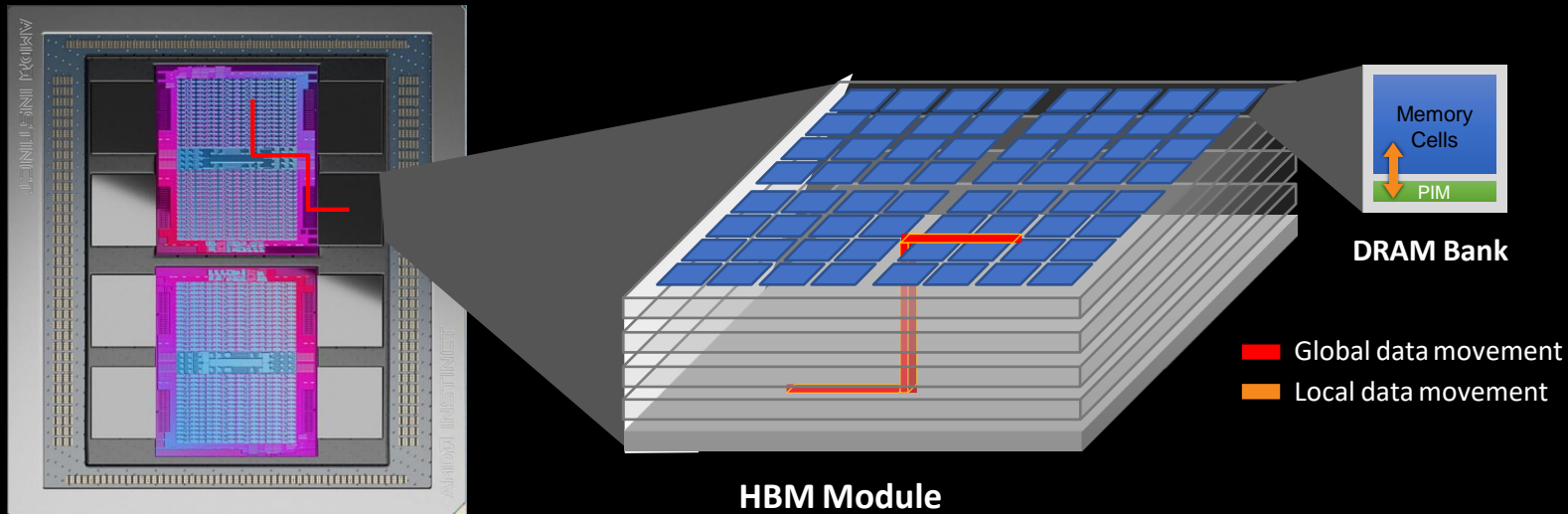
# Reducing Memory Power





# Processing in Memory

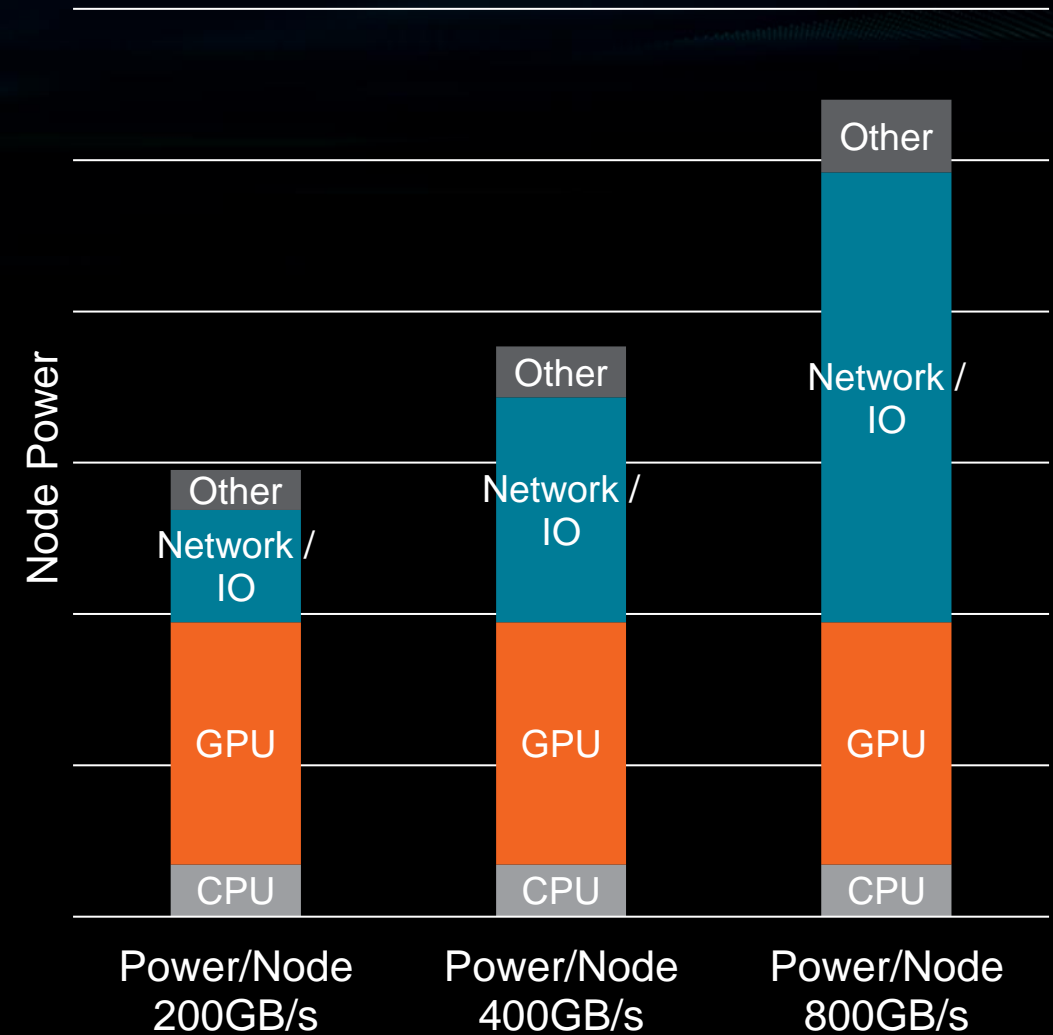
- Key algorithmic kernels can be executed directly in memory, saving precious communication energy



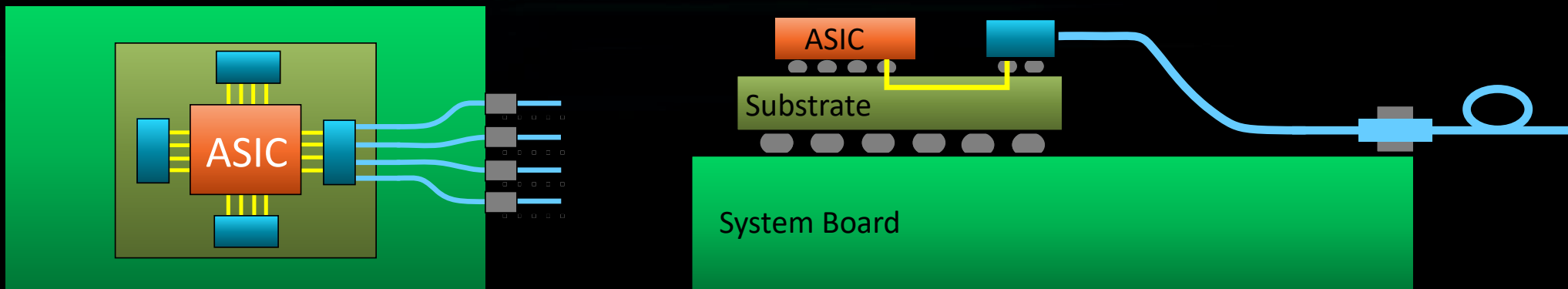


# System Power by Function

- 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

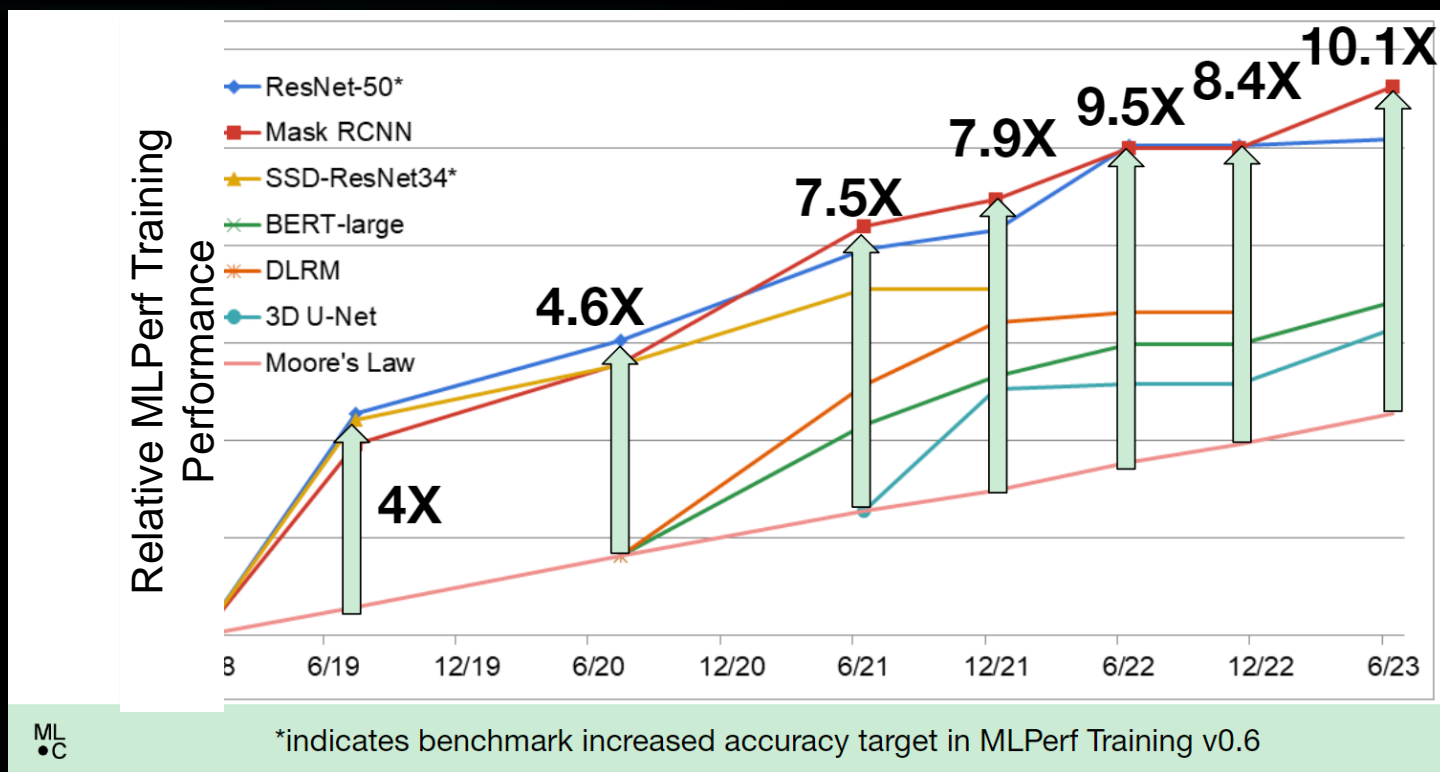
Reach and BW  
density reduces  
switch and re-timer  
power

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

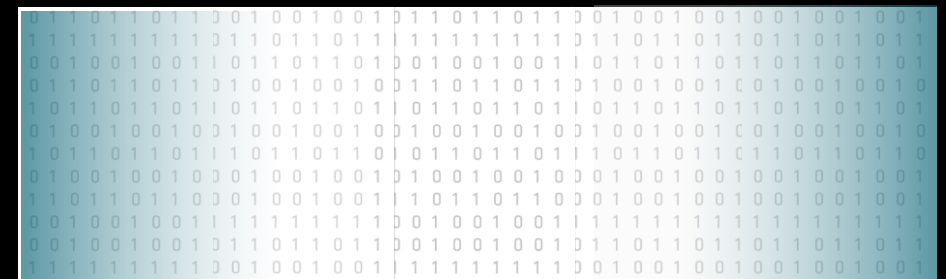
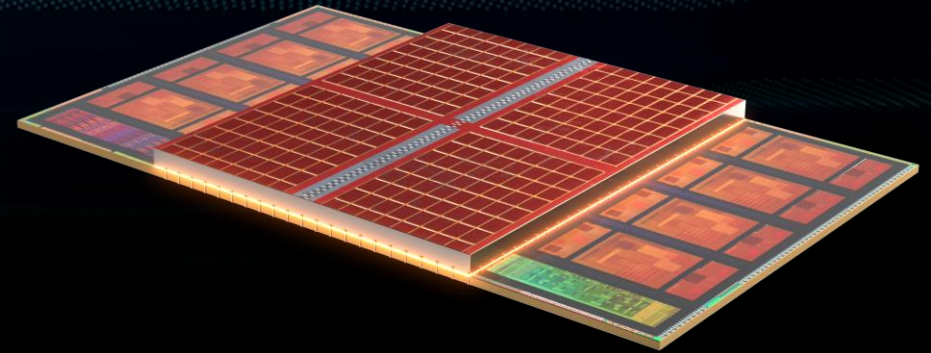
# Software-Hardware Co-Design

- 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





# 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

© 2024 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.

