



Advancing Efficiency

An AMD Overview

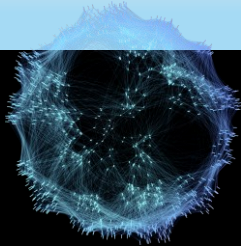
RTC4AO Conference, Garching, 6
November 2023

AMD 
together we advance_

HIGH PERFORMANCE AND ADAPTIVE COMPUTING



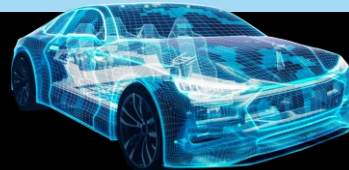
SMARTER CLIENT



AI & ANALYTICS



5G & COMMS



ADAPTABLE &
INTELLIGENT



GAMING,
SIMULATION &



CLOUD, NETWORK,
HYPERSCALE &

POWERING THE SERVICES AND PRODUCTS ESSENTIAL TO OUR DAILY LIVES

AMD recognitions

2021-2022

#1 on the Top500 and four of the top five in the Green500 lists (Nov 2022) of highest performance and most energy efficient supercomputers, respectively

Seven of top ten and 15 of the top 20 most energy efficient supercomputers

Recognized as a 2021 Supplier Engagement Leader for managing climate risks (CDP)

Ranked among top semiconductor companies in 3BL's 100 Best Corporate Citizens (2022)



Performance Trends

The background features a dark blue, almost black, field with intricate, glowing patterns. A central point of light, emitting a soft orange and red glow, radiates outwards, creating a series of horizontal and vertical lines that resemble a digital grid or data flow. The lines are composed of many small, bright points, giving the impression of a complex network or data visualization. The overall effect is futuristic and high-tech.

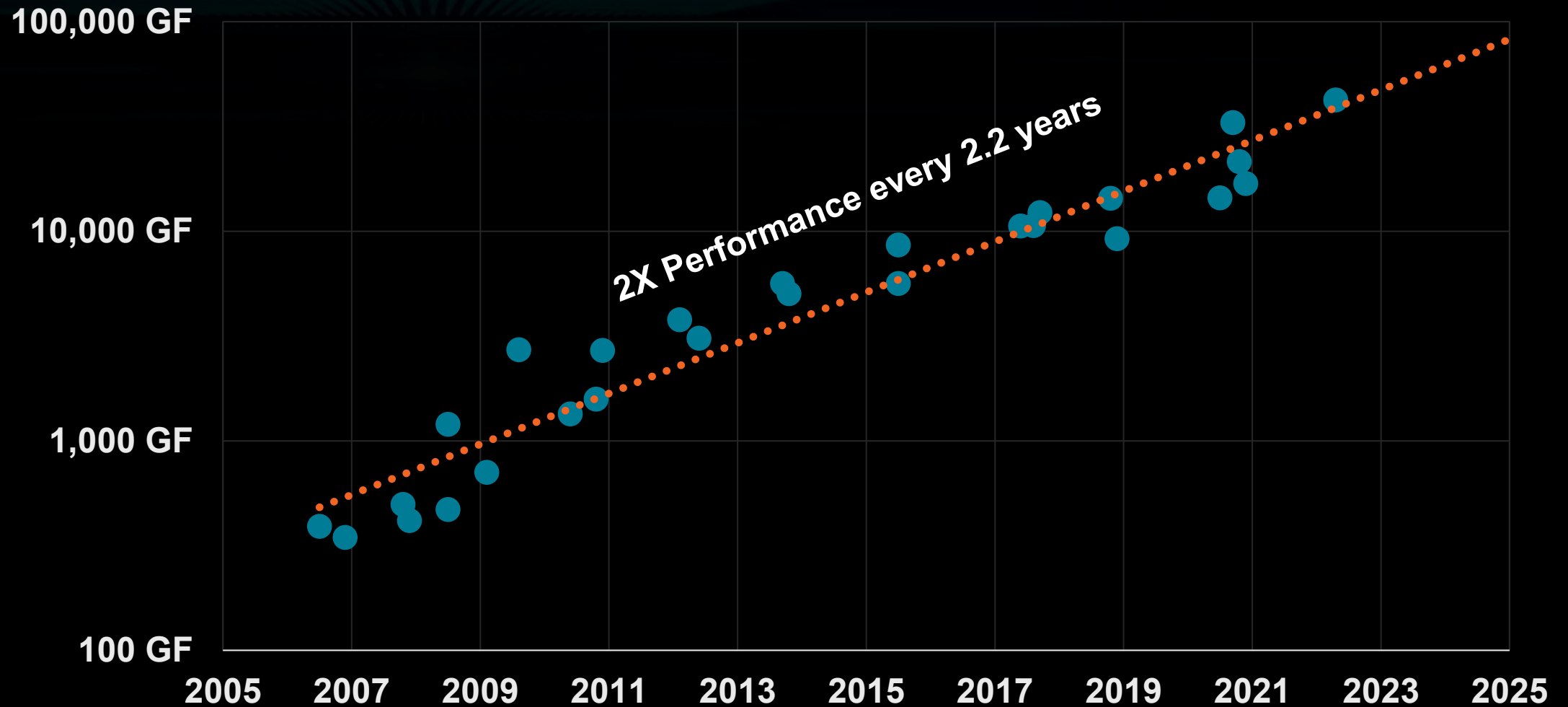
Mainstream Server Performance Trend

Server 2P SpecIntRate Over Time



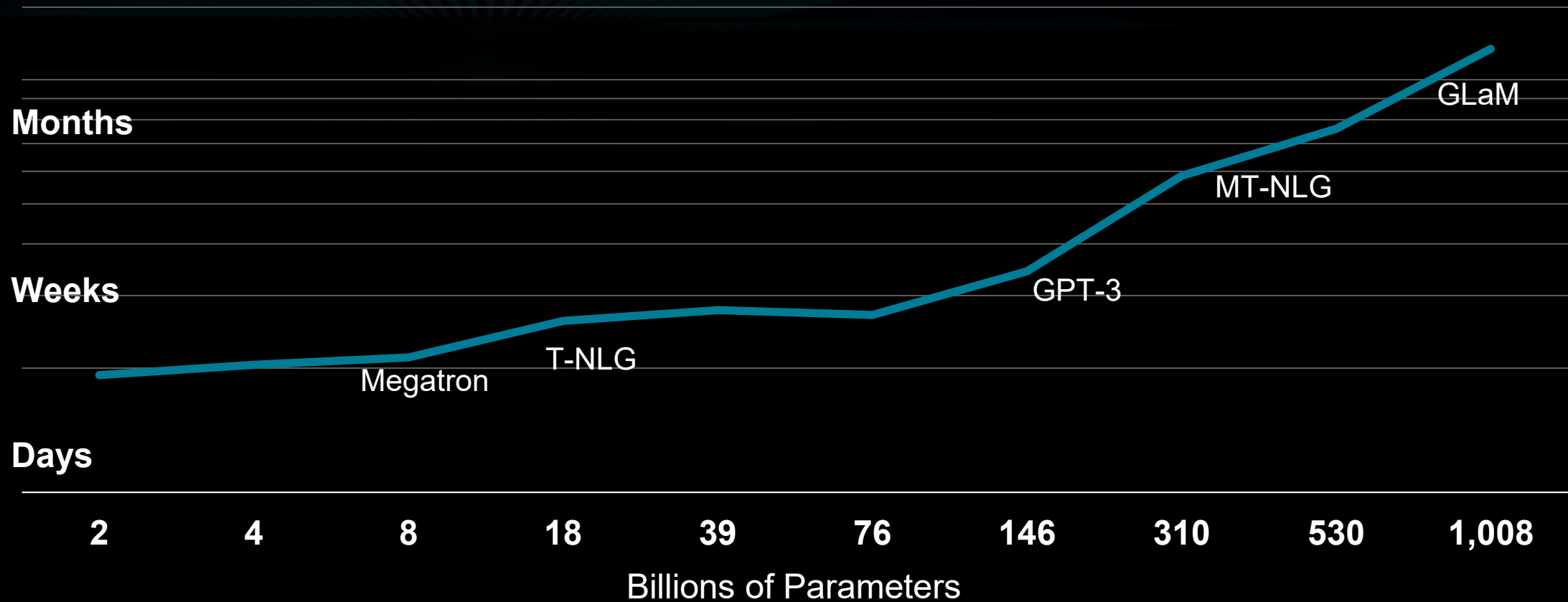
GPU Performance Trend

GPU Single Precision FLOPs Over Time



Time-to-Train Large AI Models

Language Model Training



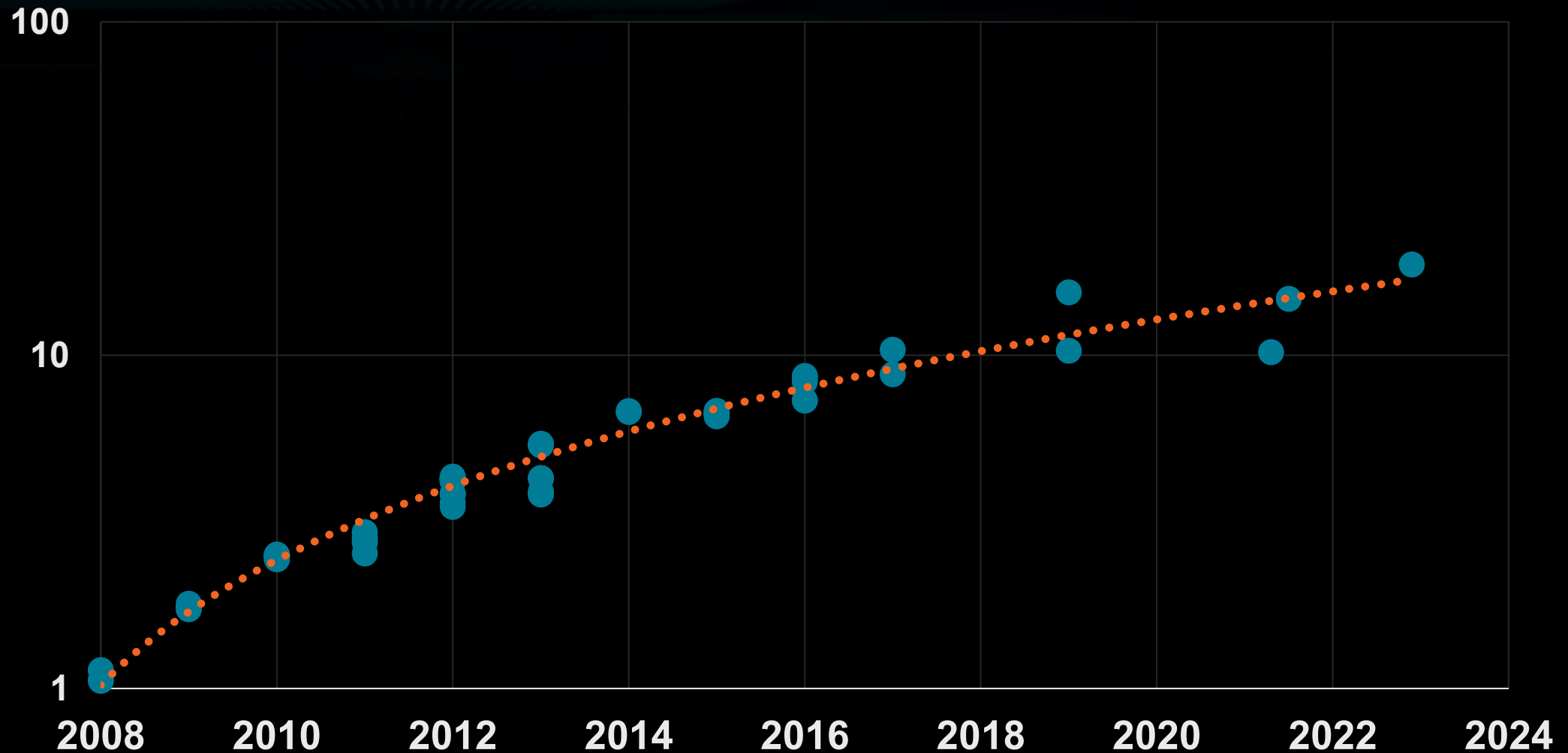
Exponentially growing model sizes driving massive growth in compute and memory

Efficiency Trends

The background features a dark blue, almost black, space filled with intricate digital patterns. A central point of light, glowing with a mix of green and blue, radiates outwards, creating a sense of depth and focus. This light source is surrounded by concentric, slightly blurred lines that resemble a digital signal or a data stream. The overall aesthetic is futuristic and high-tech, with a strong emphasis on light and shadow.

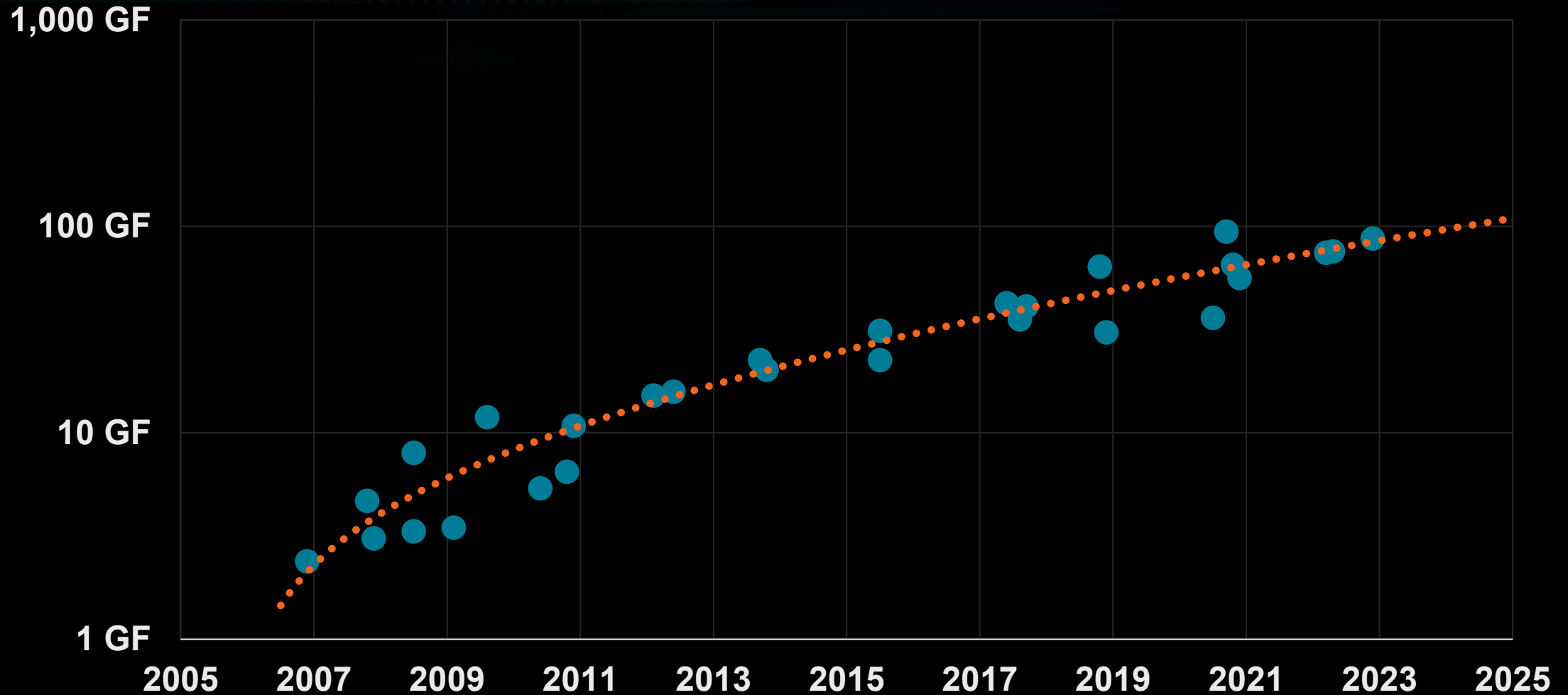
Server Efficiency Trends

Server Perf/Watt



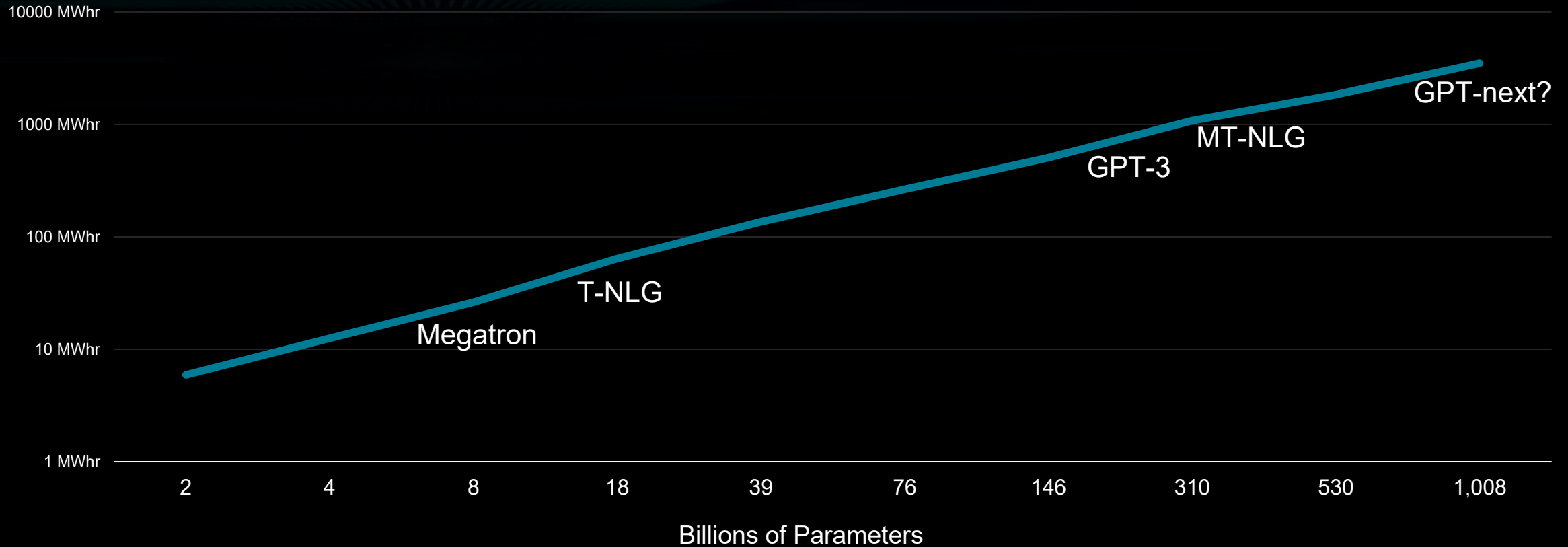
GPU Efficiency Trends

GPU Single Precision FLOPs/Watt



Energy-to-Train Large AI Models

Language Model Training

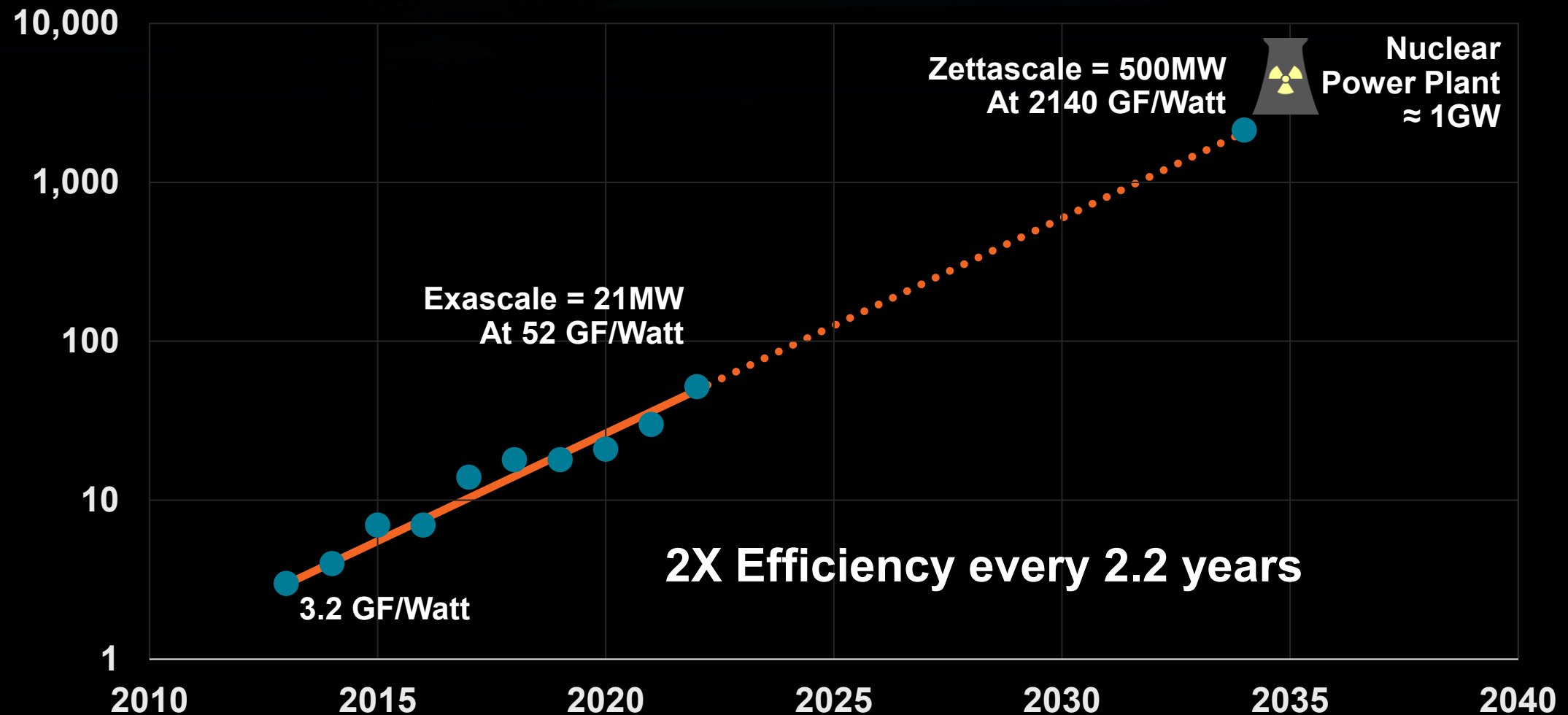


Exponentially growing model sizes driving massive growth in energy for training

The upper bound on training requirements is yet to be determined

Supercomputer Energy Use Trajectory

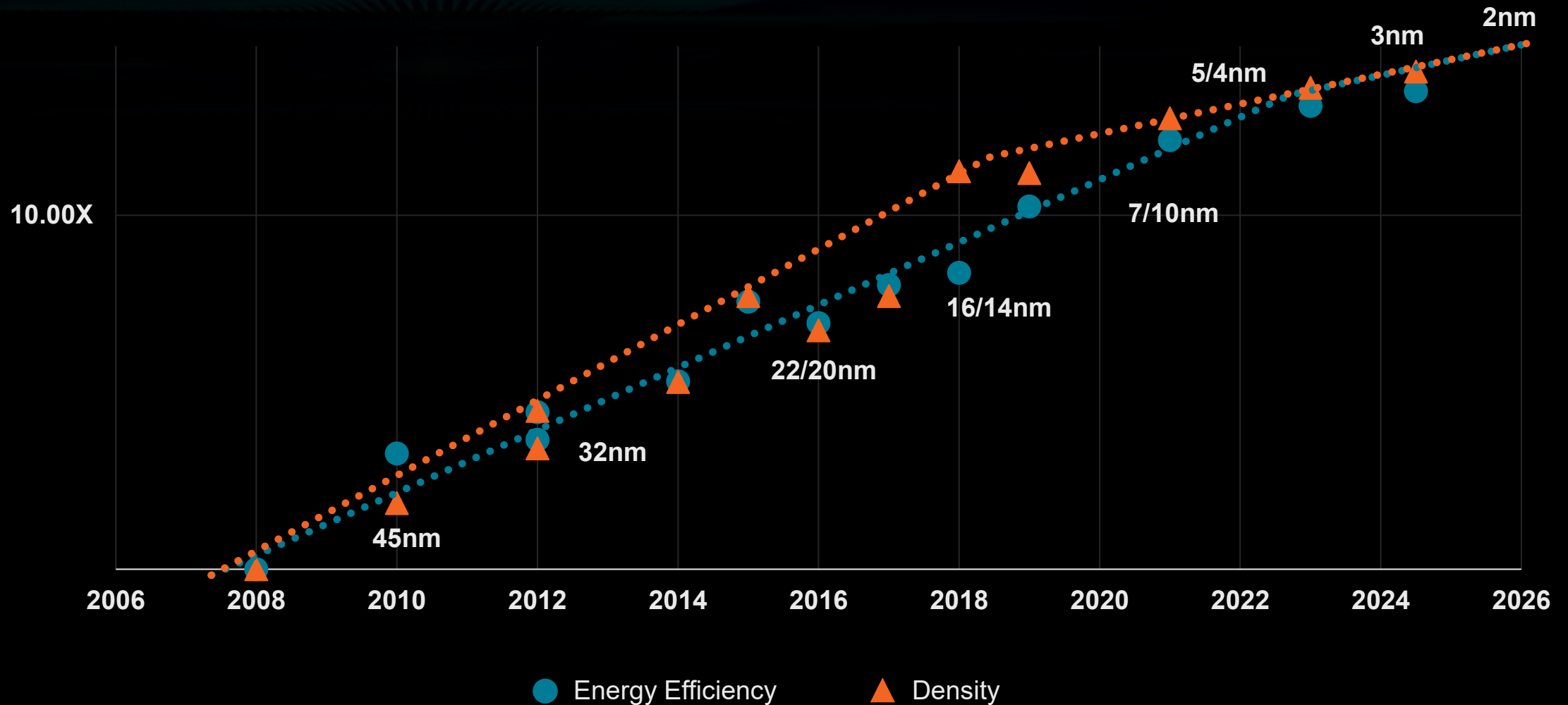
Green500 Supercomputer GFLOPs/Watt and Projection



Technology Trends

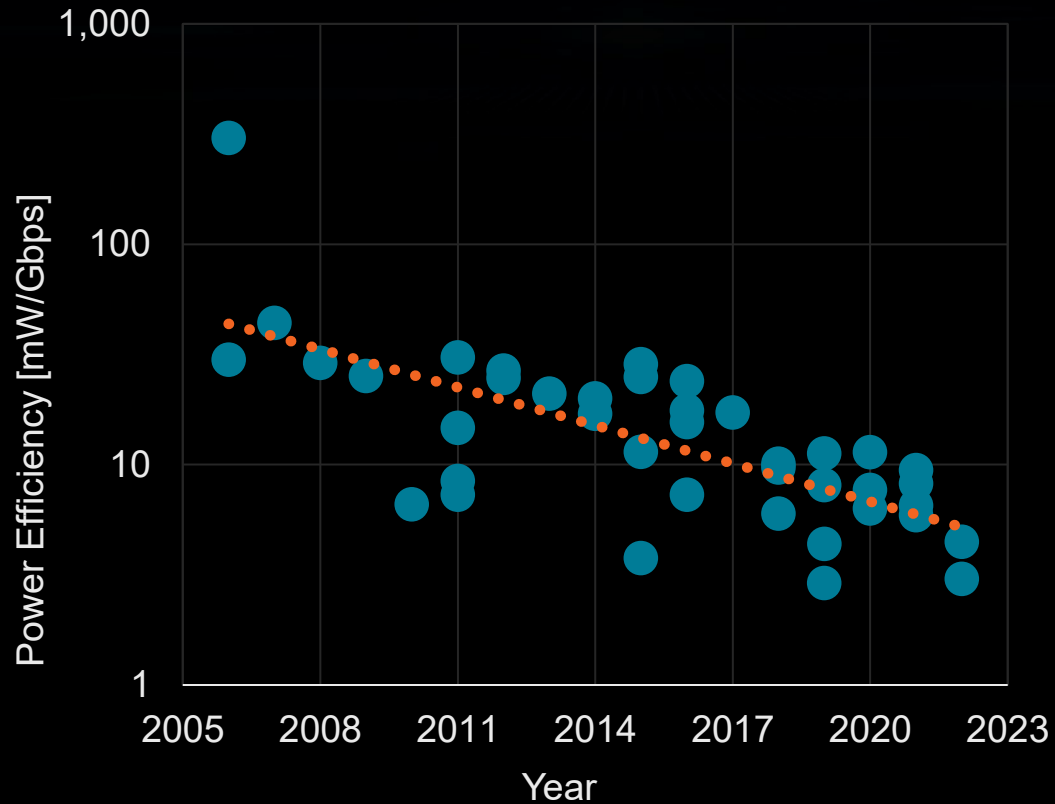
The background features a dark blue, almost black, field with intricate patterns of glowing teal and light blue lines. These lines form a grid-like structure that recedes into the distance, creating a sense of depth and perspective. A bright, glowing point of light is positioned in the center, from which several beams of light radiate outwards, illuminating the surrounding patterns. The overall aesthetic is futuristic and high-tech.

Logic Process Technology Trends

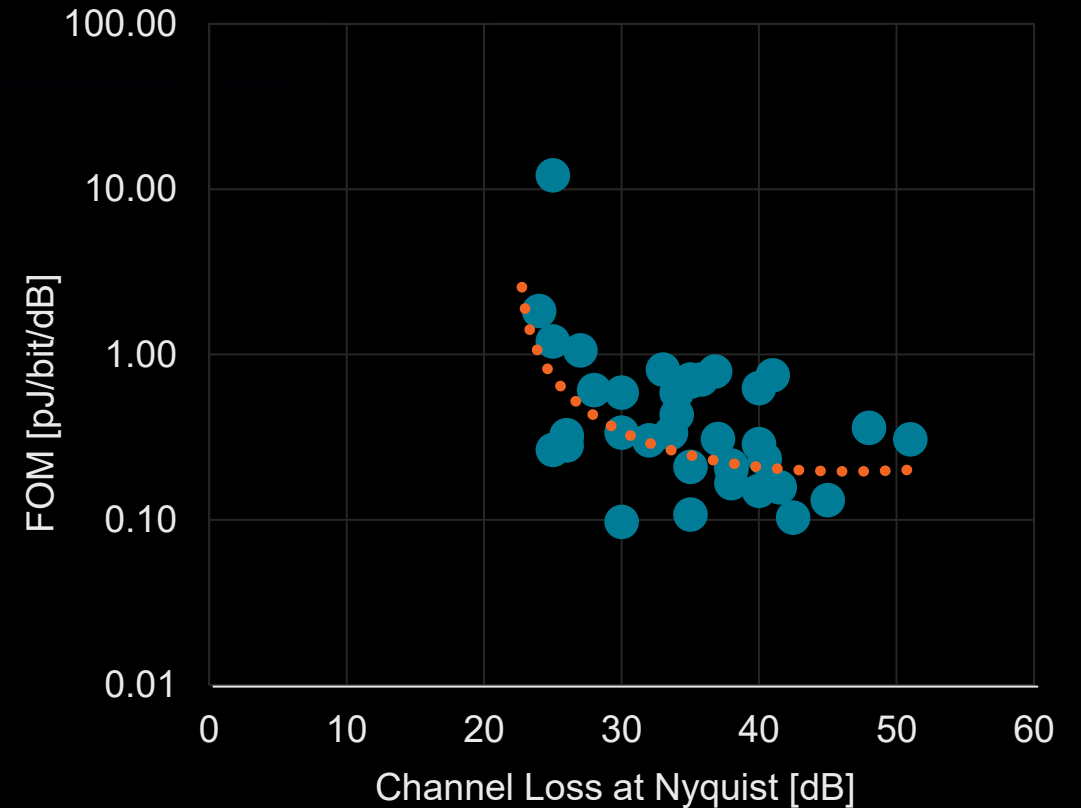


Energy Efficiency of IO Interconnect

Power Efficiency vs. Year



FOM = Power[mW]/(DataRate[Gbps]*
Channel Loss at Nyquist[dB])

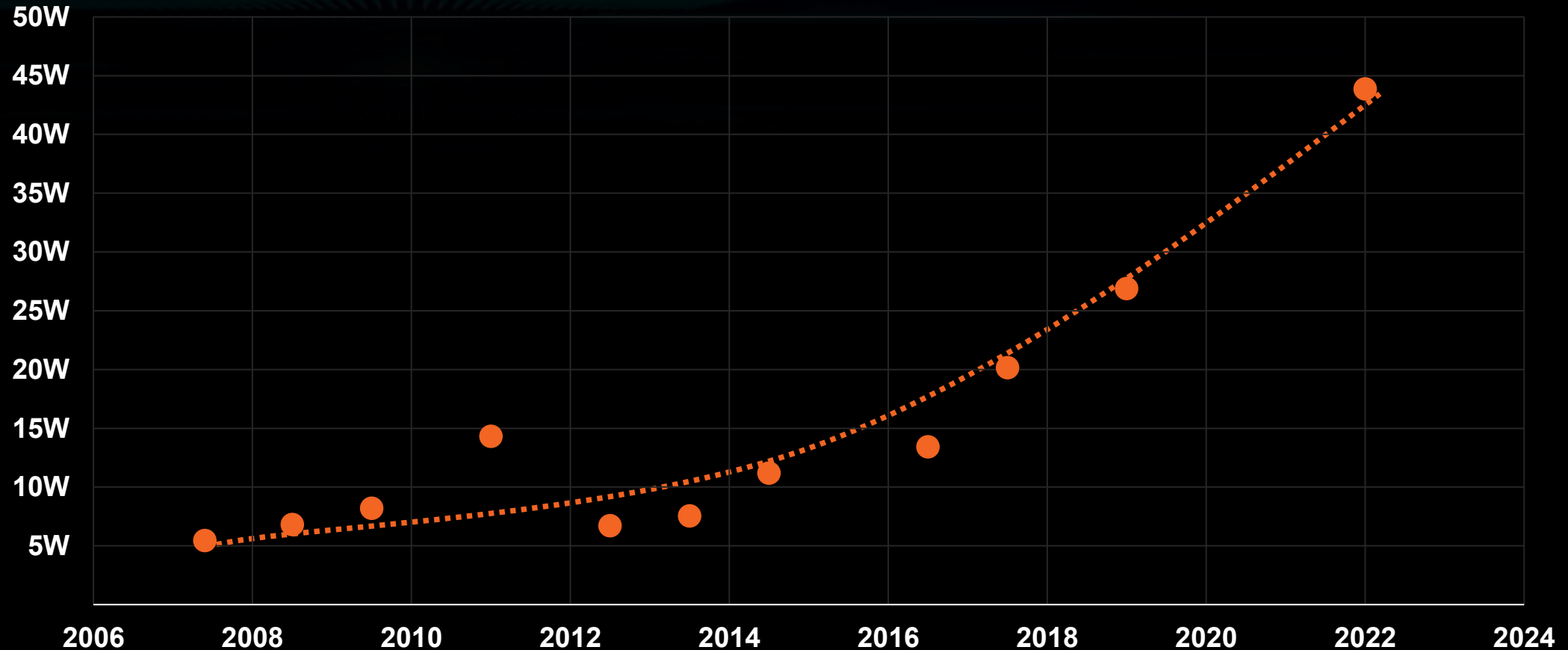


Energy per bit is reducing

But not when normalized to channel quality

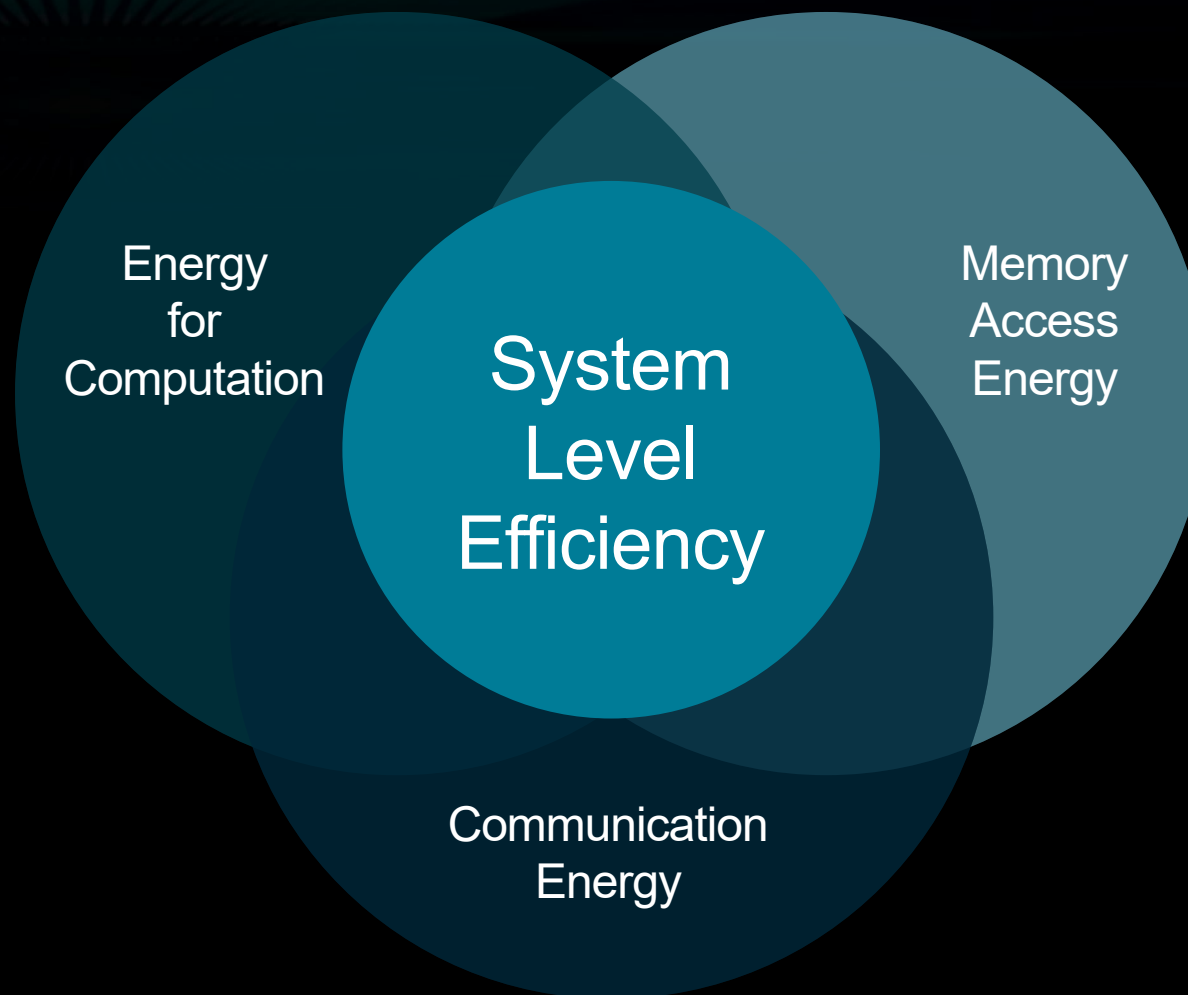
Memory Access Power

Server Memory interface Power



Interface Power Efficiency Improvements are Overwhelmed by Growing Bandwidth Demand

Driving System Level Efficiency Holistically



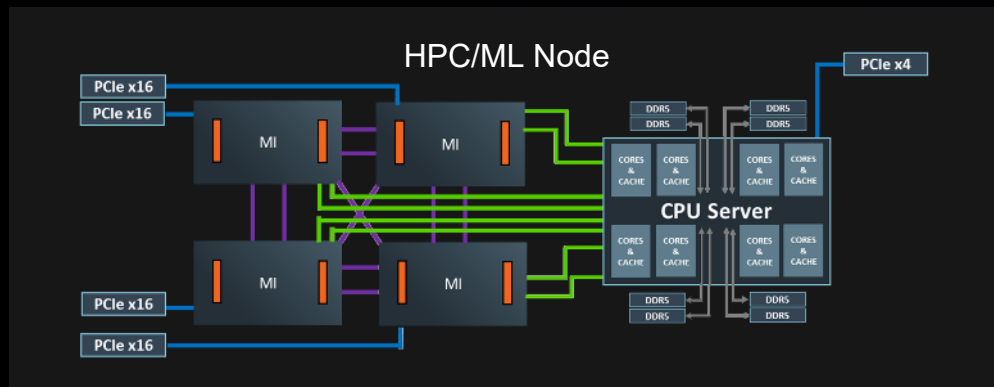
Efficiency Through Advanced Architecture

The background features a dark blue, almost black, field with a complex pattern of glowing teal and light blue lines. These lines form a perspective that converges towards a bright, glowing point in the center, creating a sense of depth and movement. The lines are composed of many thin, parallel segments, giving the overall effect a digital or architectural feel.

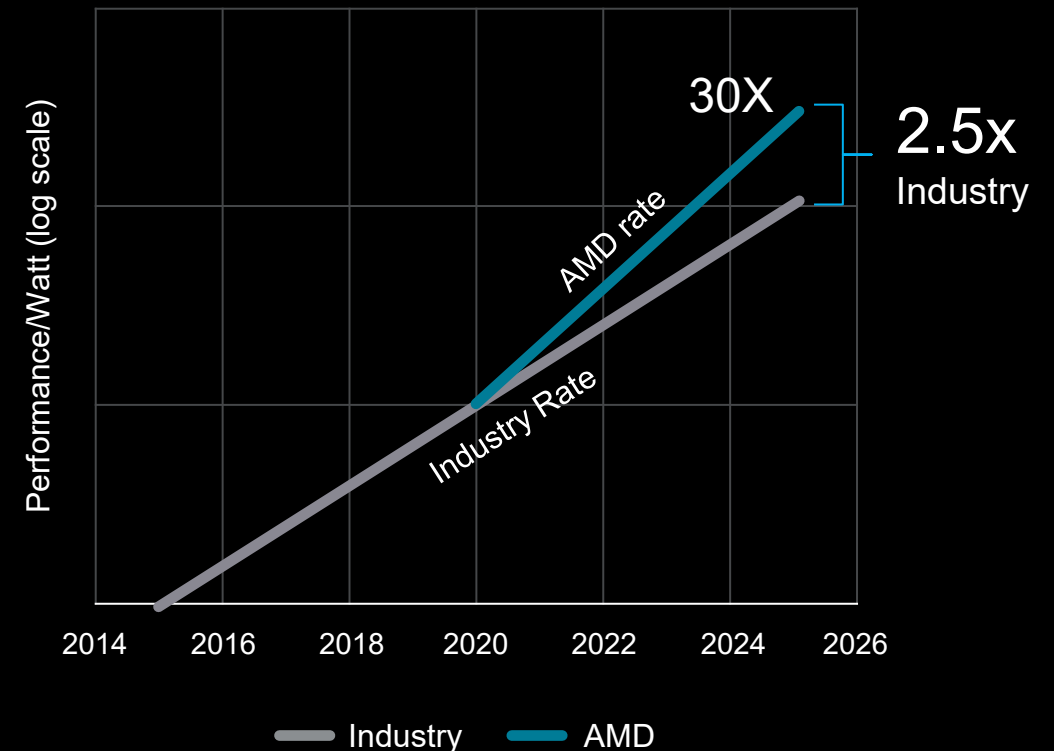
System Level Energy Efficiency Optimization

The AMD 30x by 2025 energy efficiency goal for accelerators set at node level

- Accelerated compute nodes using AMD CPUs and GPUs



Accelerated Computing Performance/Watt Trends



- Exploit architectural innovations, package and silicon technology advances to change the trend
- Industry efficiency improvement rate of ~12x for HPC and AI nodes
- Innovate at the next level up to bend the curve to 30x

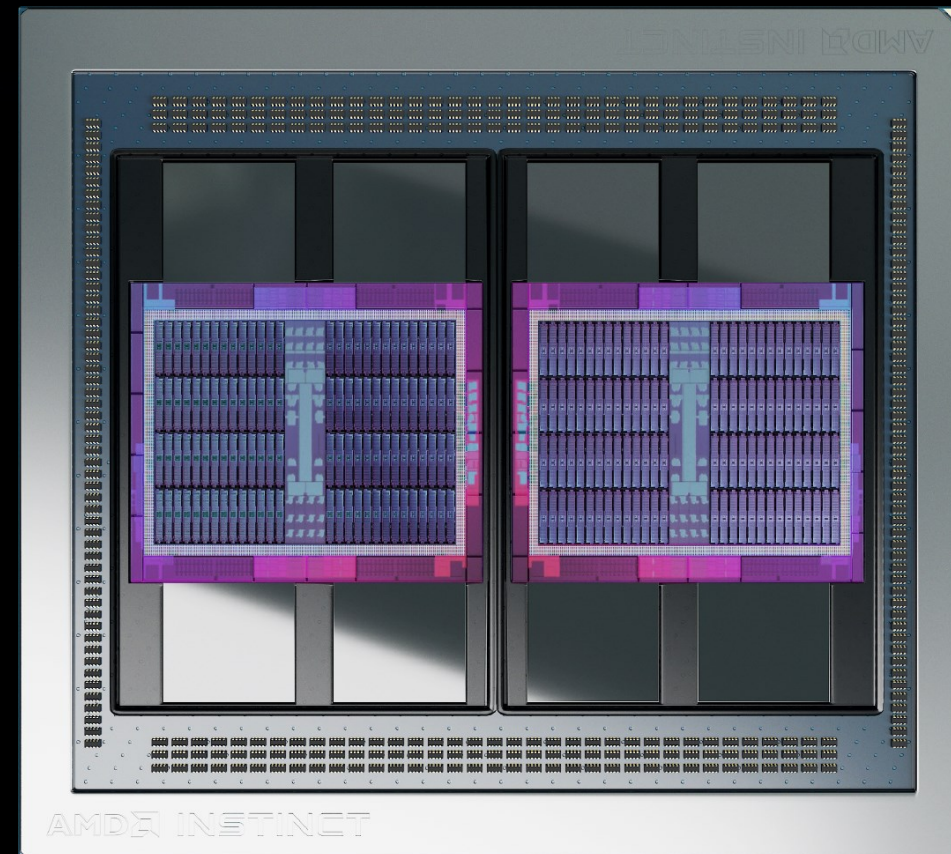
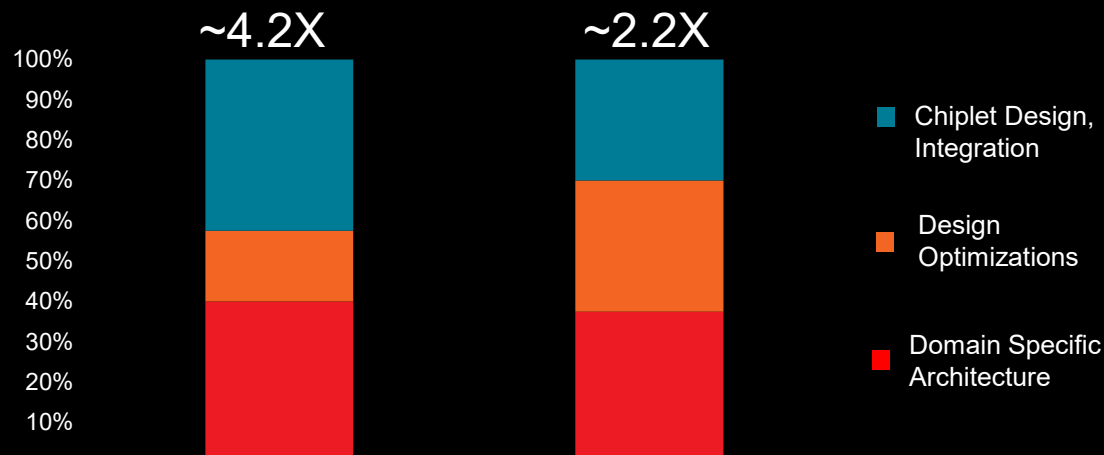
Based on 2015-2020 industry trends in energy efficiency gains and data center energy consumption in 2025.

* Includes AMD high performance CPU and GPU accelerators used for AI training and High-Performance Computing in a 4-Accelerator, CPU hosted configuration. Goal calculations are based on performance scores as measured by standard performance metrics (HPC: Linpack DGEMM kernel FLOPS with 4k matrix size, AI training: lower precision training-focused floating point math GEMM kernels such as FP16 or BF16 FLOPS operating on 4k matrices) divided by the rated power consumption of a representative accelerated compute node including the CPU host + memory, and 4 GPU accelerators.

AMD Instinct™ MI250 Performance and Perf/Watt Gains

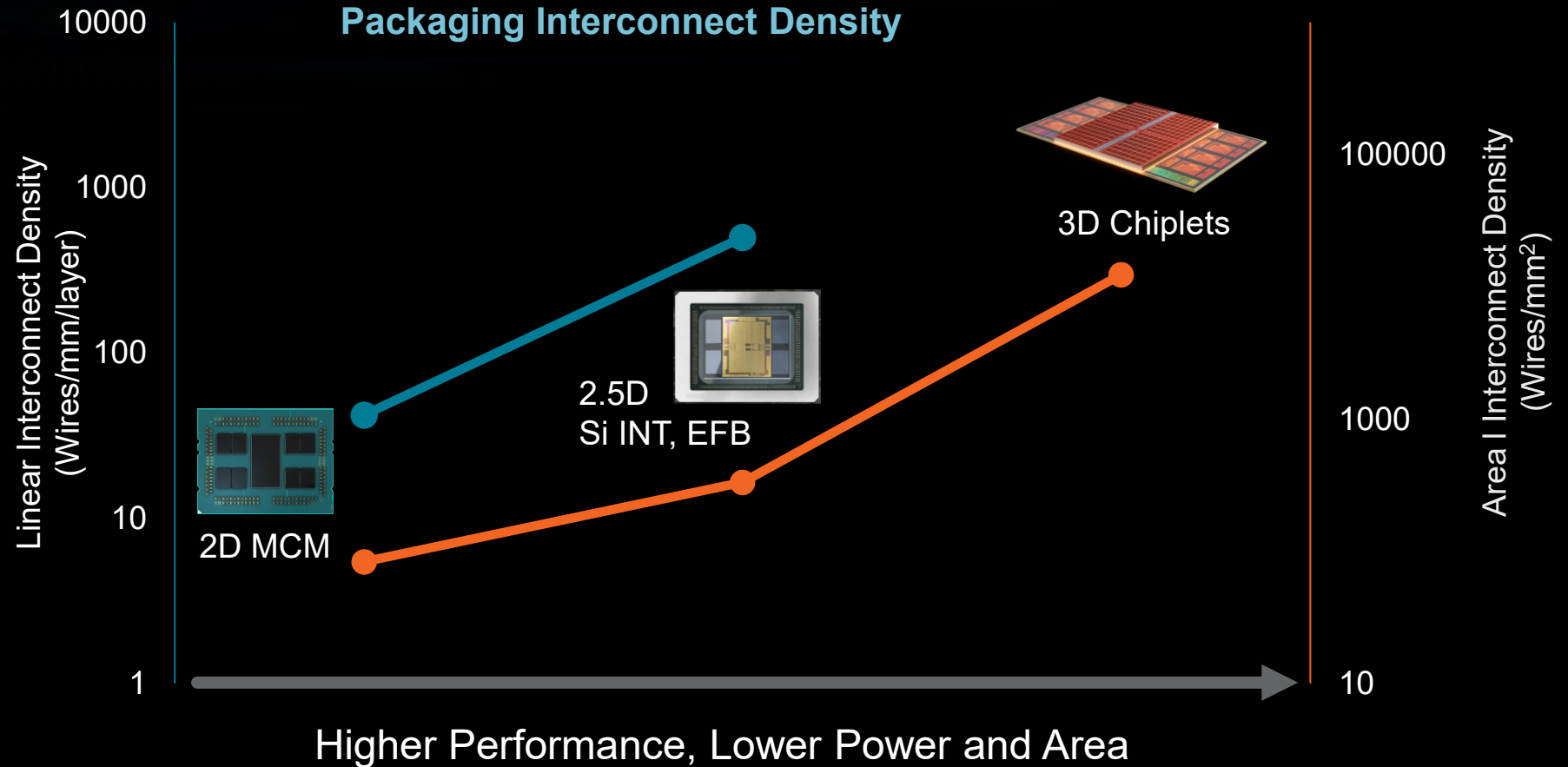
- Instinct MI250, 6nm GPU accelerator delivered significant increases in efficiency and performance
 - Advanced domain-specific architecture
 - Higher levels of integration with chiplets
 - Traditional design optimizations for efficiency
- MI250 systems power 4 out of the top 5 Green500 spots

AMD Instinct™ MI250X vs. MI100



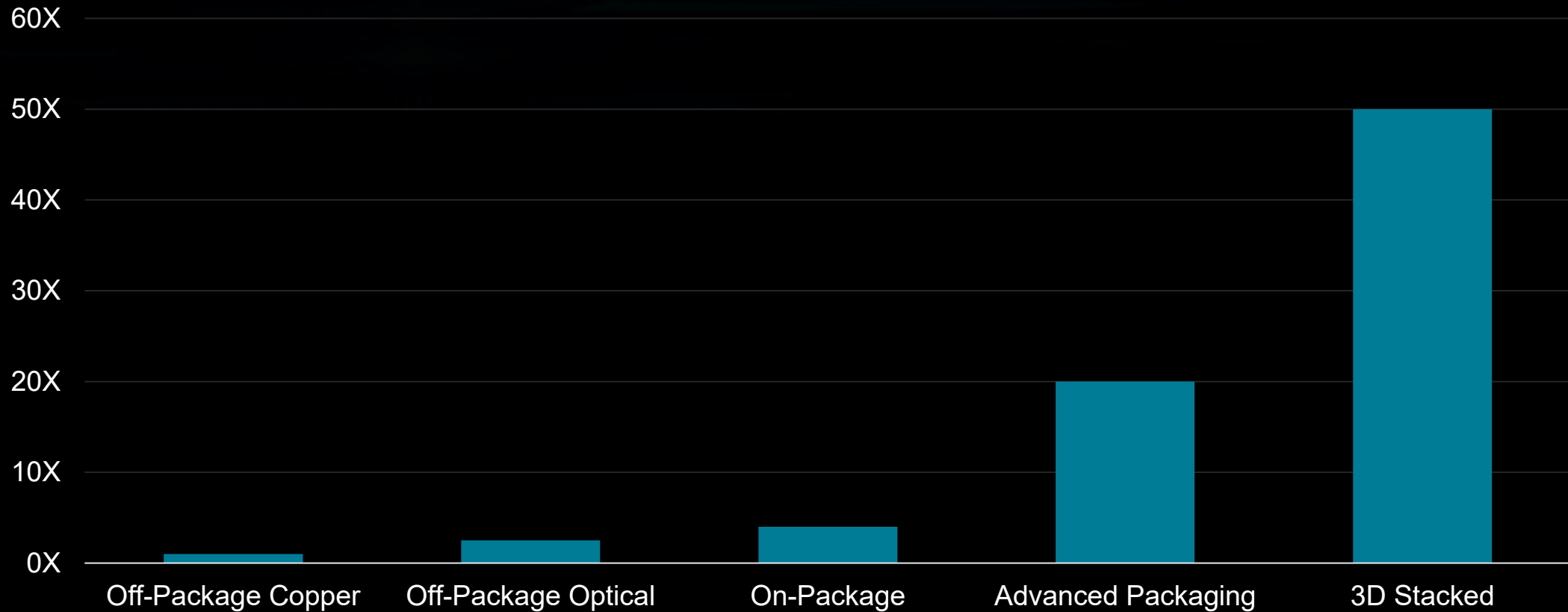
Advanced Packaging Enables Significant Gains in Performance and Efficiency

- High bandwidth between chiplets enables architectural performance gains while lowering total communication energy
- CPU and GPU integration virtually eliminates costly data transfer energy



Reducing The Cost of Communication

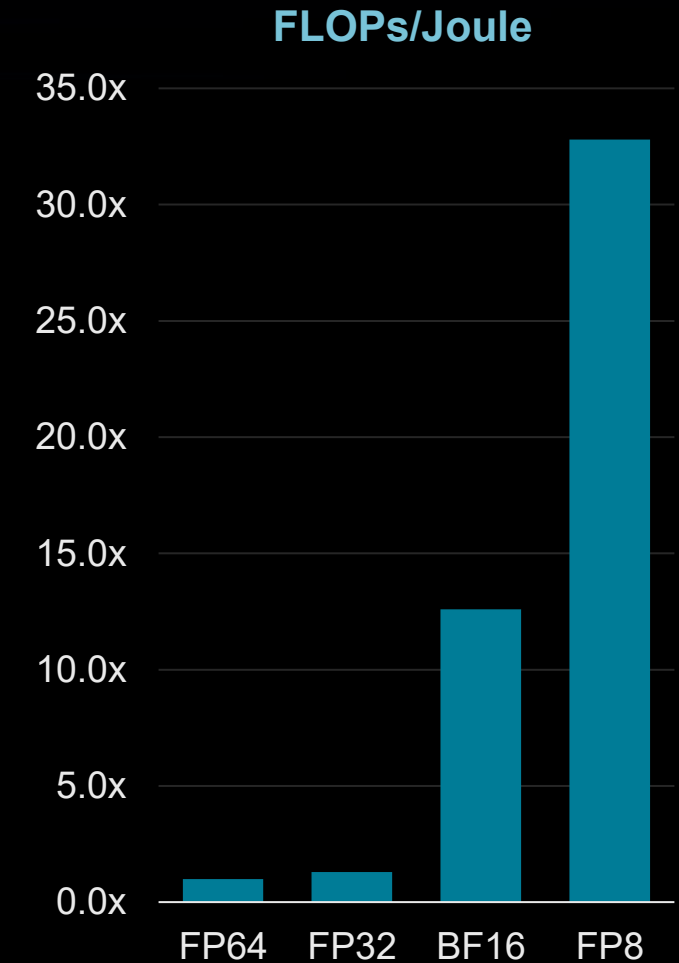
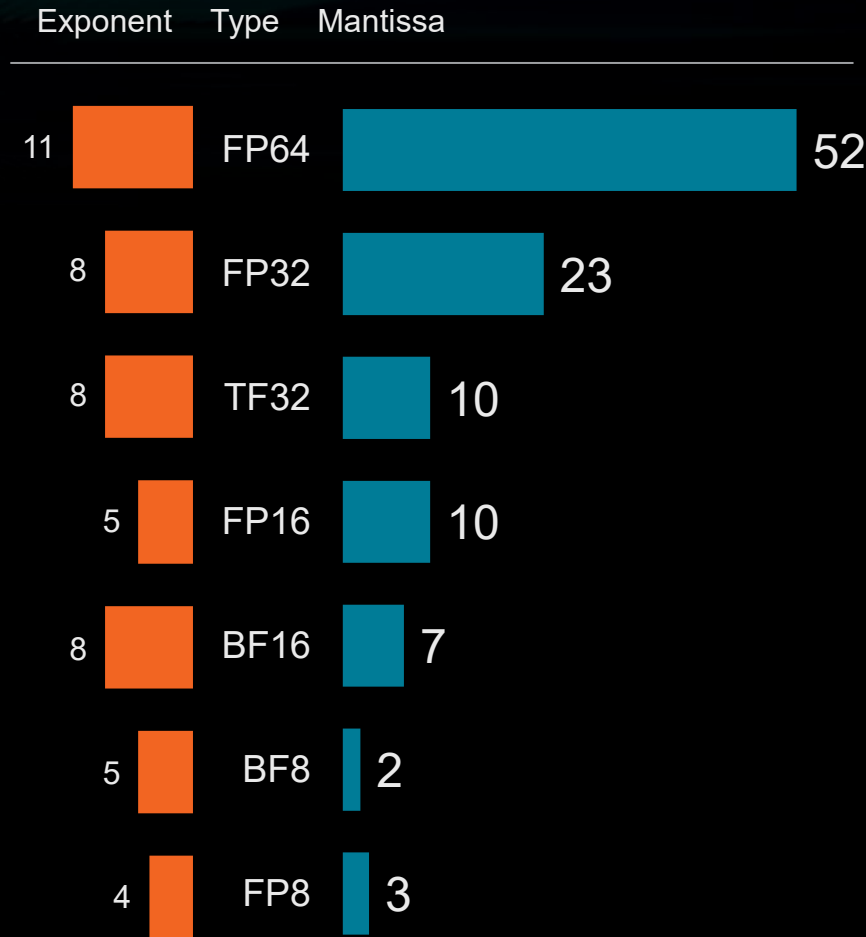
Relative Bits/Joule



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

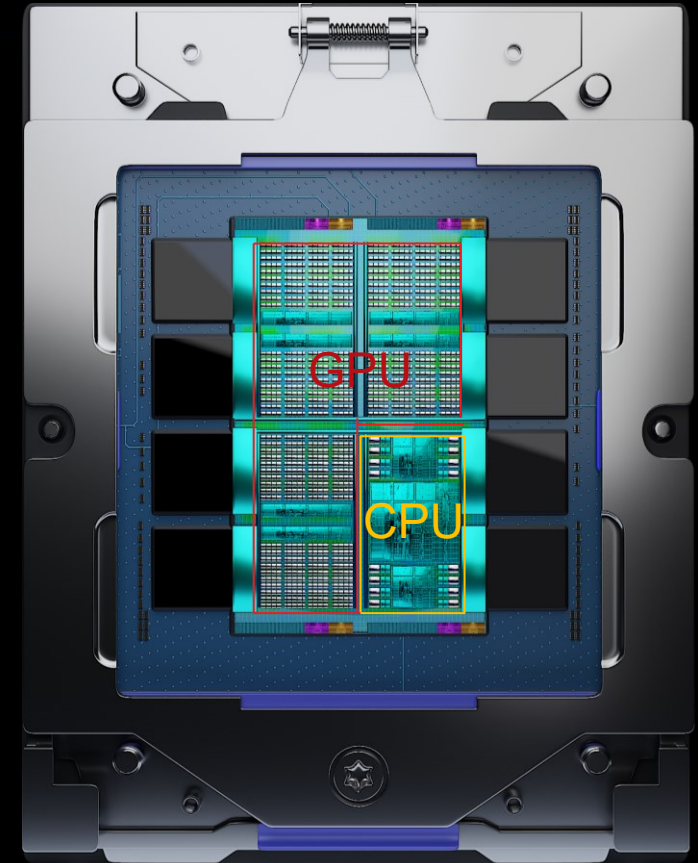
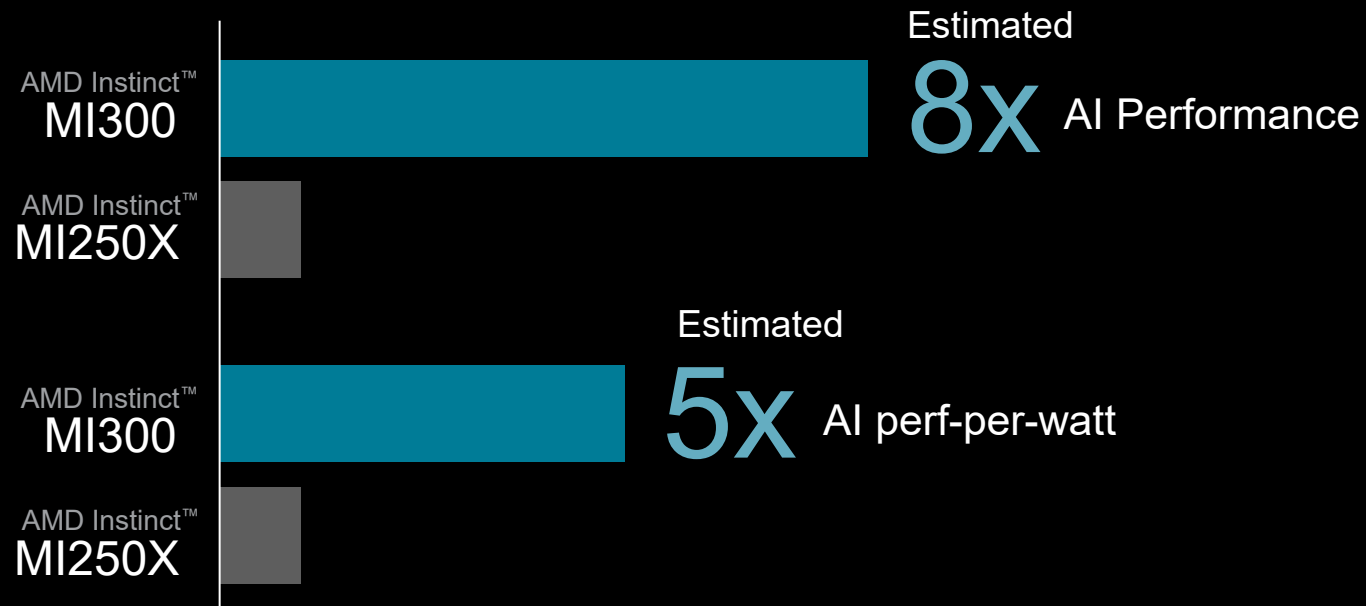
Domain-Specific Computation Enables Workload Optimization which Drives Performance and Efficiency

- Tailor architecture by application
- Adapt algorithms to use lower precision math formats for significant improvements in energy efficiency



MI300 Architectural Innovation at the Next Level

- 5nm process technology with 3D stacking
- Next-gen Infinity Cache™ and 4th Gen Infinity Fabric base die
- New Math formats
- Unified memory APU Architecture



3D CPU+GPU Integration for Next-Level Efficiency

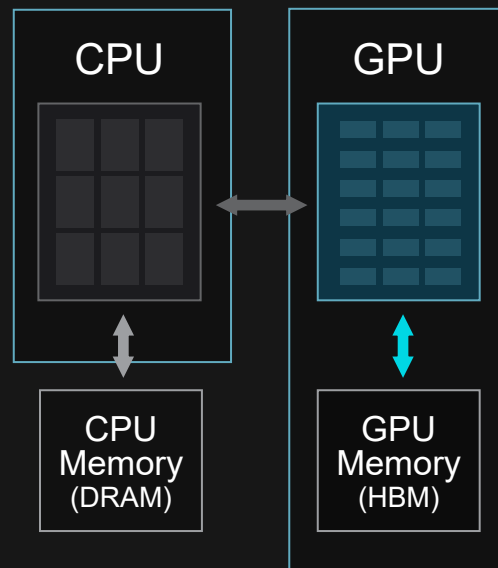
AMD CDNA™ 2 Coherent Memory Architecture



AMD CDNA™ 3 Unified Memory APU Architecture

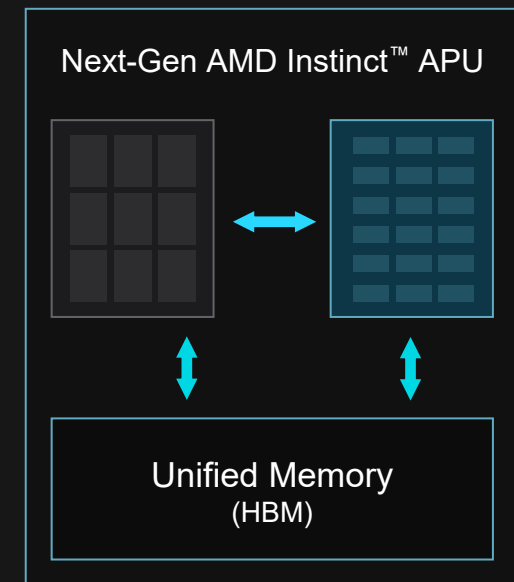
AMD Instinct™ MI250 Accelerator

- Simplifies programming
- Low overhead 3rd Gen Infinity interconnect
- Industry standard modular design



AMD Instinct™ MI300 Accelerator

- Eliminates redundant memory copies
- High bandwidth, low latency communication
- Low TCO with unified memory APU package

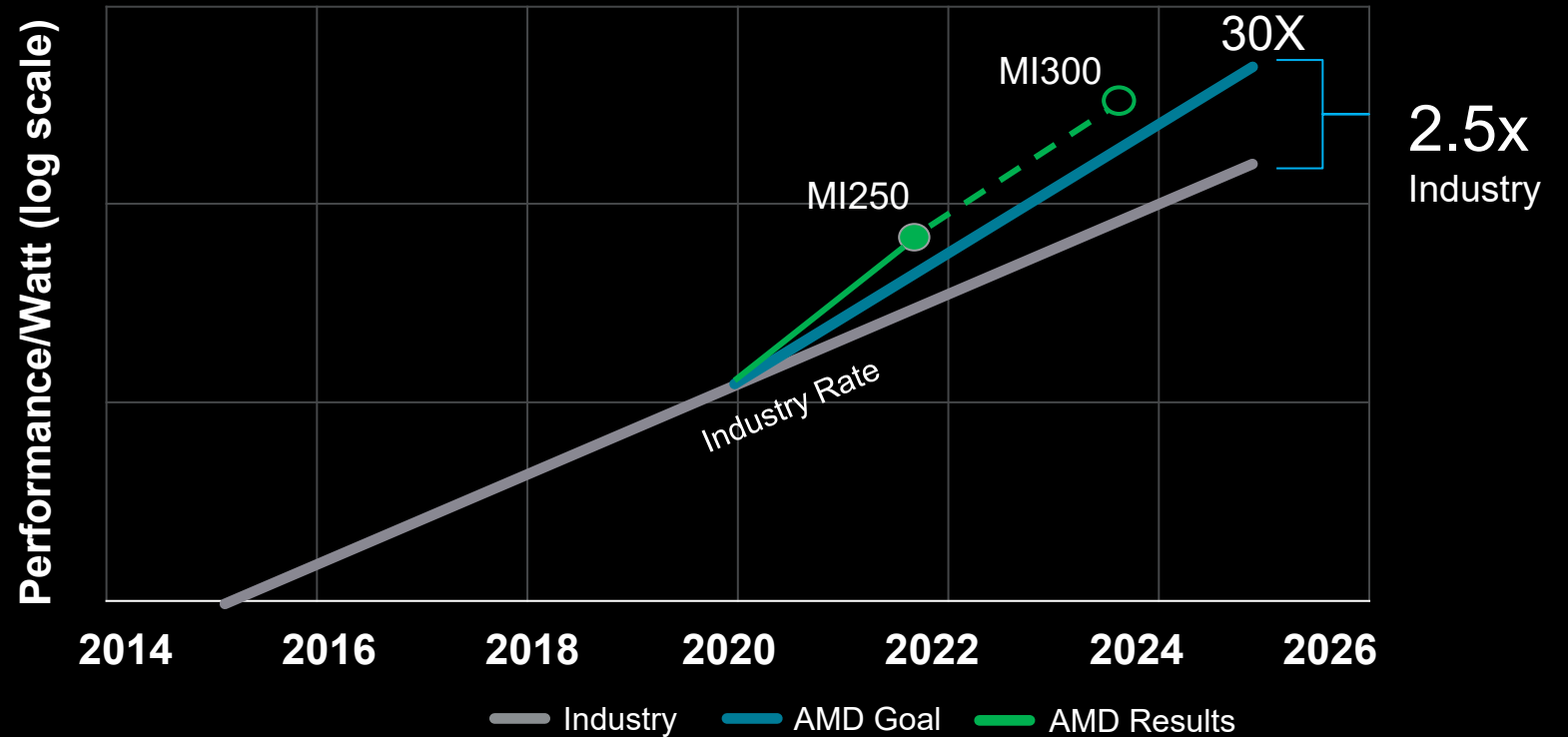


Resulting Node-Level Efficiency Gains

AMD Roadmap On A Trend To Exceed Aggressive 30x Goal

- Architecture, packaging and interconnect innovations pay off
- Chiplet and 3D-enabled architecture put AMD Instinct products on a path to exceed 30x goal

Accelerated Computing Performance/Watt Trends

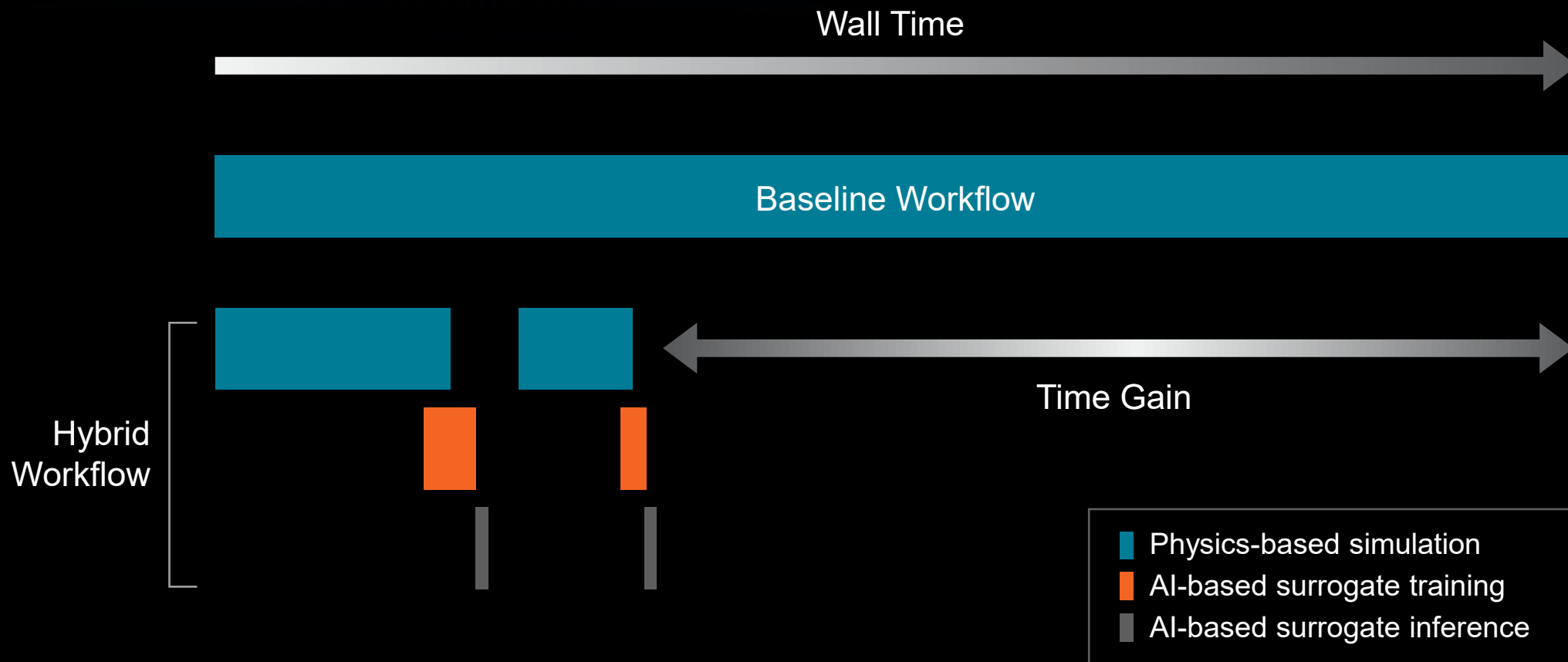


Based on 2015-2020 industry trends in energy efficiency gains and data center energy consumption in 2025.

* Includes AMD high performance CPU and GPU accelerators used for AI training and High-Performance Computing in a 4-Accelerator, CPU hosted configuration. Goal calculations are based on performance scores as measured by standard performance metrics (HPC: Linpack DGEMM kernel FLOPS with 4k matrix size. AI training: lower precision training-focused floating point math GEMM kernels such as FP16 or BF16 FLOPS operating on 4k matrices) divided by the rated power consumption of a representative accelerated compute node including the CPU host + memory, and 4 GPU accelerators.

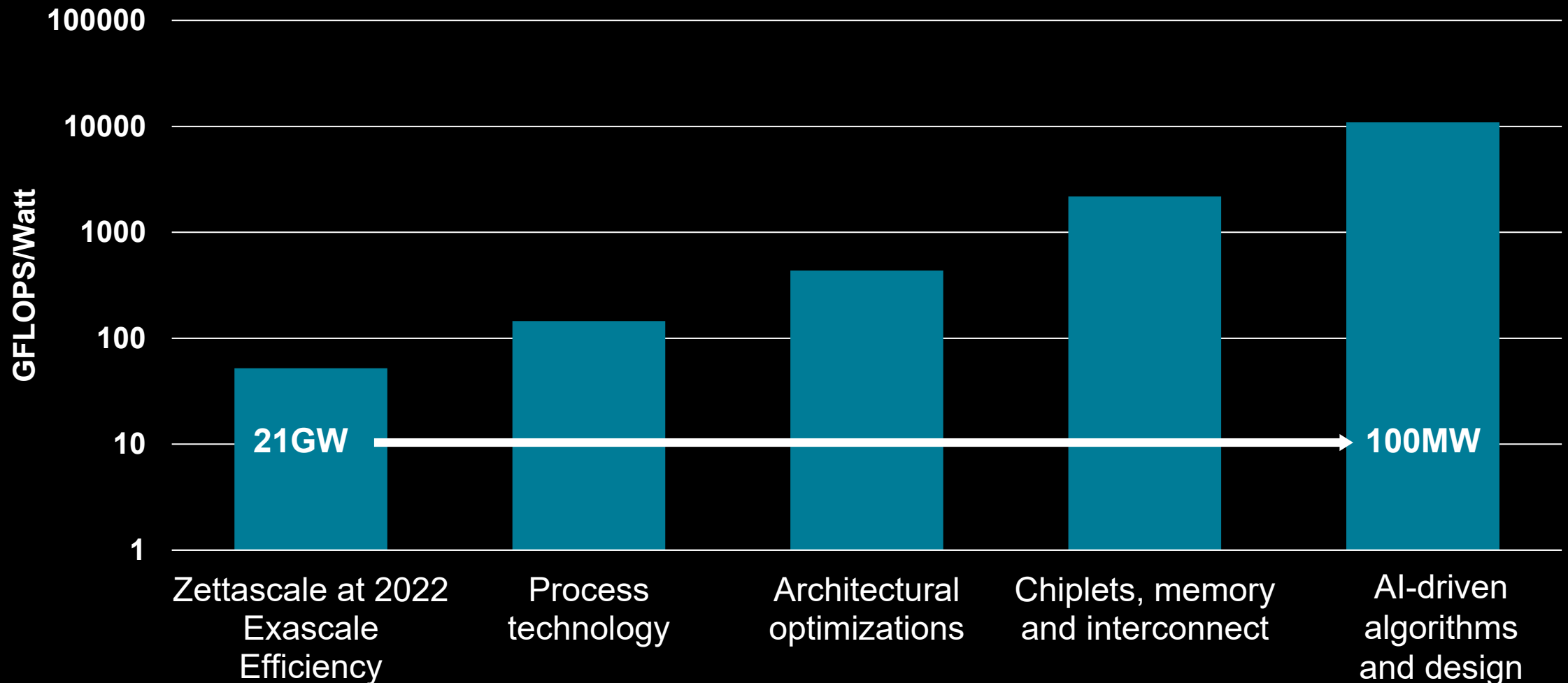
Potential Performance Gains from AI Accelerated HPC

- AI will be the key to breaking past the limits of traditional computing approaches



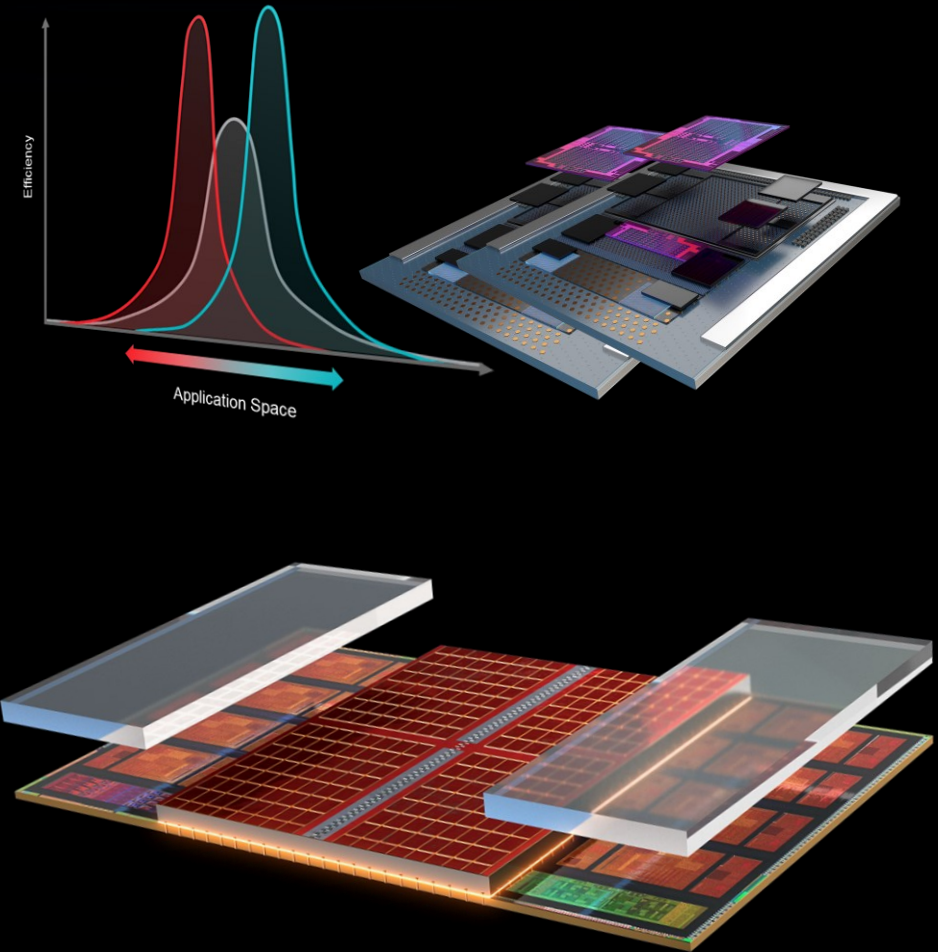
Achieving Zettascale Computing

Efficiency Roadmap to Zettascale Leveraging AI



Driving Performance Gains Over the Next Decade Requires Relentless Focus on Energy Efficiency

- Insatiable demand for more compute
- Energy efficiency is the primary limiter
- We must innovate in new dimensions:
 - System level optimizations
 - Domain specific architectures
 - Tight integration of compute and memory with chiplet architectures, advanced packaging, new interconnects
 - Leveraging AI holistically
- Deep collaboration required across materials, process, circuits, system design, architecture, software, and applications



AMD 

Disclaimer and Attribution

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

© 2023 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, AMD Advantage, AMD Instinct, EPYC, Radeon, Ryzen, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other product names used herein are for identification purposes and may be trademarks of their respective owners.

SPECrate® is a registered trademark of the Standard Performance Evaluation Corporation. For more information, see www.spec.org.

End notes

End note MI300-003. Measurements by AMD Performance Labs June 4, 2022 on current specification and/or estimation for estimated delivered FP8 floating point performance with structure sparsity supported for AMD Instinct™ MI300 vs. MI250X FP16 (306.4 estimated delivered TFLOPS based on 80% of peak theoretical floating-point performance). MI300 performance based on preliminary estimates and expectations. Final performance may vary. MI300-003.

Measurements conducted by AMD Performance Labs as of Jun 7, 2022 on the current specification for the AMD Instinct™ MI300 APU (850W) accelerator designed with AMD CDNA™ 3 5nm FinFET process technology, projected to result in 2,507 TFLOPS estimated delivered FP8 with structured sparsity floating-point performance.

End note: MI300-04. Estimated delivered results calculated for AMD Instinct™ MI250X (560W) GPU designed with AMD CDNA 2 6nm FinFET process technology with 1,700 MHz engine clock resulted in 306.4 TFLOPS (383.0 peak FP16 x 80% = 306.4 delivered) FP16 floating-point performance. Actual results based on production silicon may vary. MI300-04