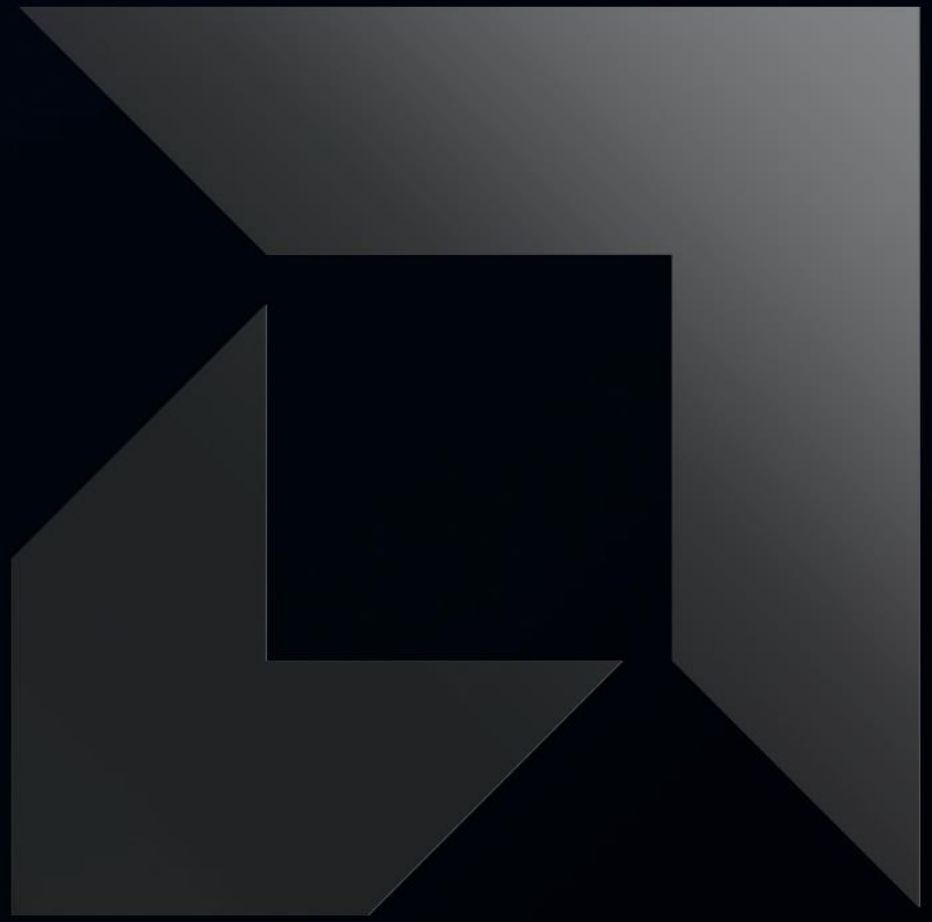


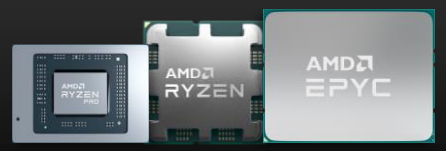
Save Us From Ourselves!
Efficiency in AI

Srilatha (Bobbie) Manne
Senior Fellow
March 15, 2023

AMD
together we advance_

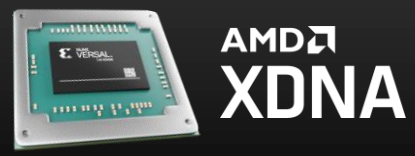
Strategic Pillars

Unmatched Compute Technology

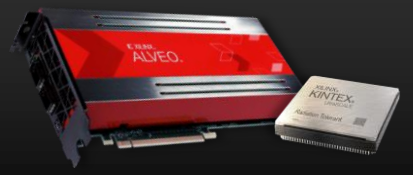


CPUs

GPUs



AI Engine
FPGA Fabric



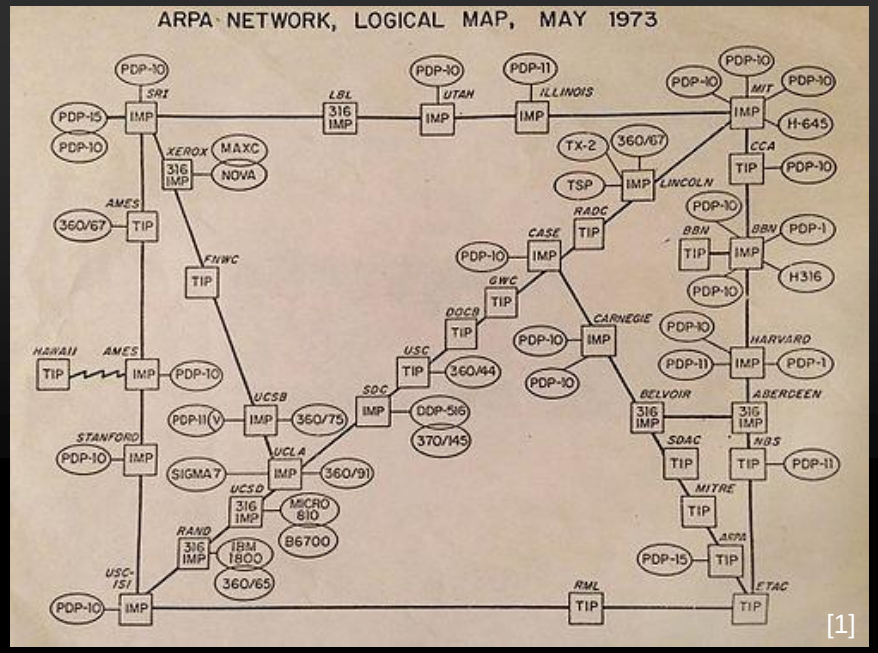
FPGAs and
Adaptive SoCs

SmartNICs
and DPUs

Leadership Technology and Product Portfolio

The Landscape

The Past 50 Years



[1]

1973 ARPANET: 42 computers

2023: Over 42 Billion connected devices[2]

[1] https://commons.wikimedia.org/wiki/File:Arpanet_map_1973.jpg, Public Domain, Wiesoweshalbwarum
[2] https://www.forbes.com/sites/bernardmarr/2022/11/07/the-top-4-internet-of-things-trends-in-2023/?sh=67f33bf62aea

Daily Statistics [1]

500M tweets [2]

5B searches

65B messages

4TB data per connected car

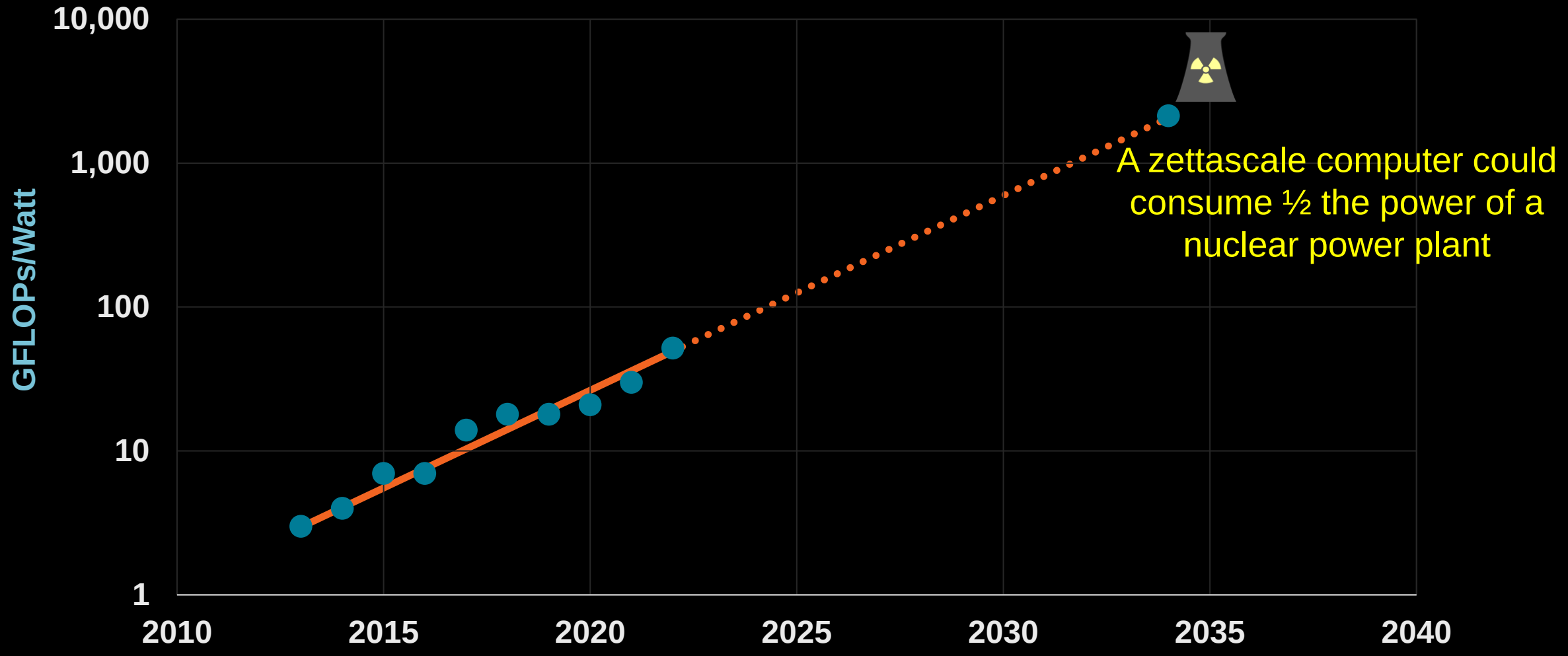
294B emails

4PB of data [3]

[1] https://www.visualcapitalist.com/how-much-data-is-generated-each-day/
[2] https://commons.wikiedia.org/wiki/File:Twitter_logo.png, CC-BY-SA-4.0, Thatgaypigeon
[3] https://commons.wikimedia.org/wiki/File:Facebook_icon_%28black%29.svg,, CC-BY-SA-4.0, Psubhashish

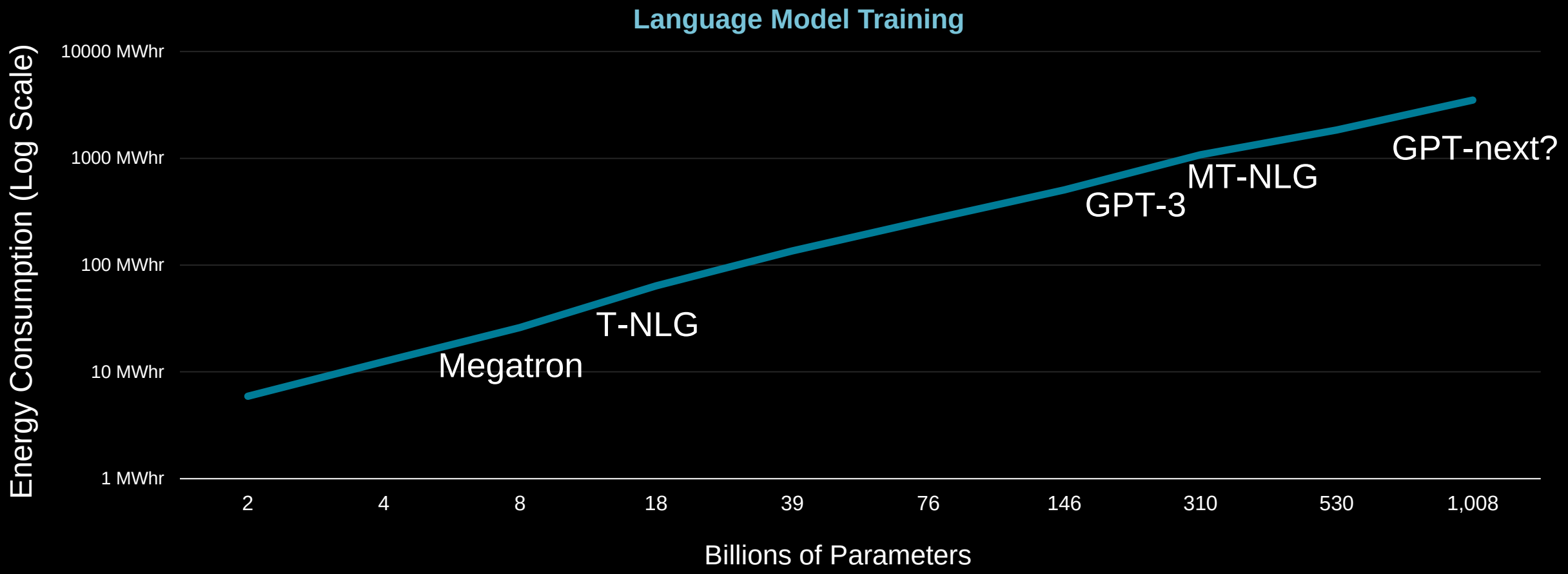
Supercomputer Energy Use Trajectory

Green500 Supercomputers and Extrapolation



Source: “Innovation for the Next Decade of Computer Efficiency”, ISSCC Plenary talk by Lisa Su, AMD (Simplified)

Energy to Train Large AI Models

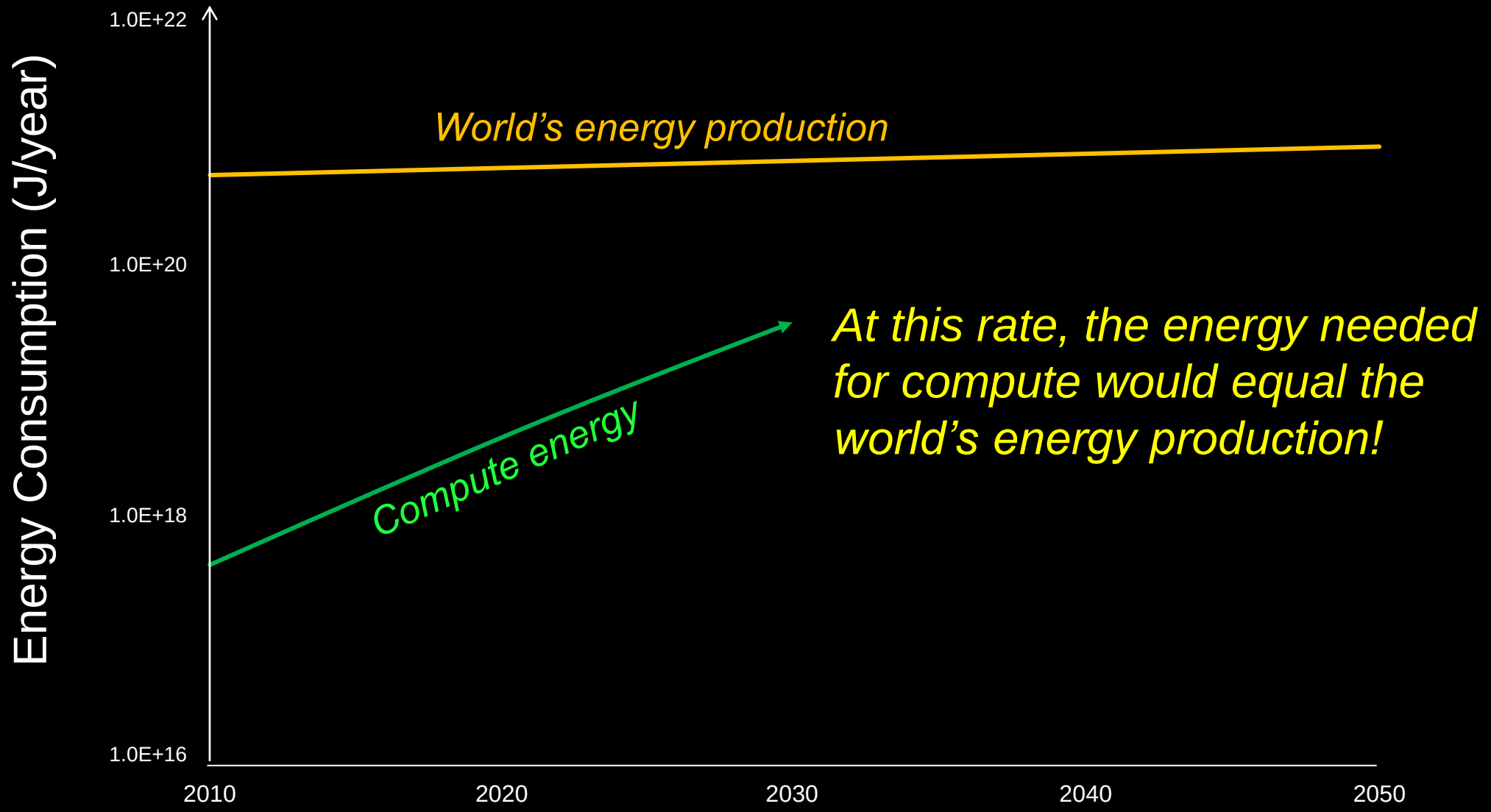


Exponentially growing model sizes drive immense growth in energy for training.

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

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

Compute Energy Trajectory



At this rate, the energy needed for compute would equal the world's energy production!

Source: SRC decadal plan for semi-conductors 2020

“I want the world. I want the whole world. I want to lock it all up in my pocket. It’s my bar of chocolate. Give it to me now!”

- Veruca Salt, from *Willie Wonka and the Chocolate Factory*

A Culture of More and Now

Efficiency

What is Efficiency?

The Jevons Paradox:
Increased Efficiency Leads to Increased Consumption

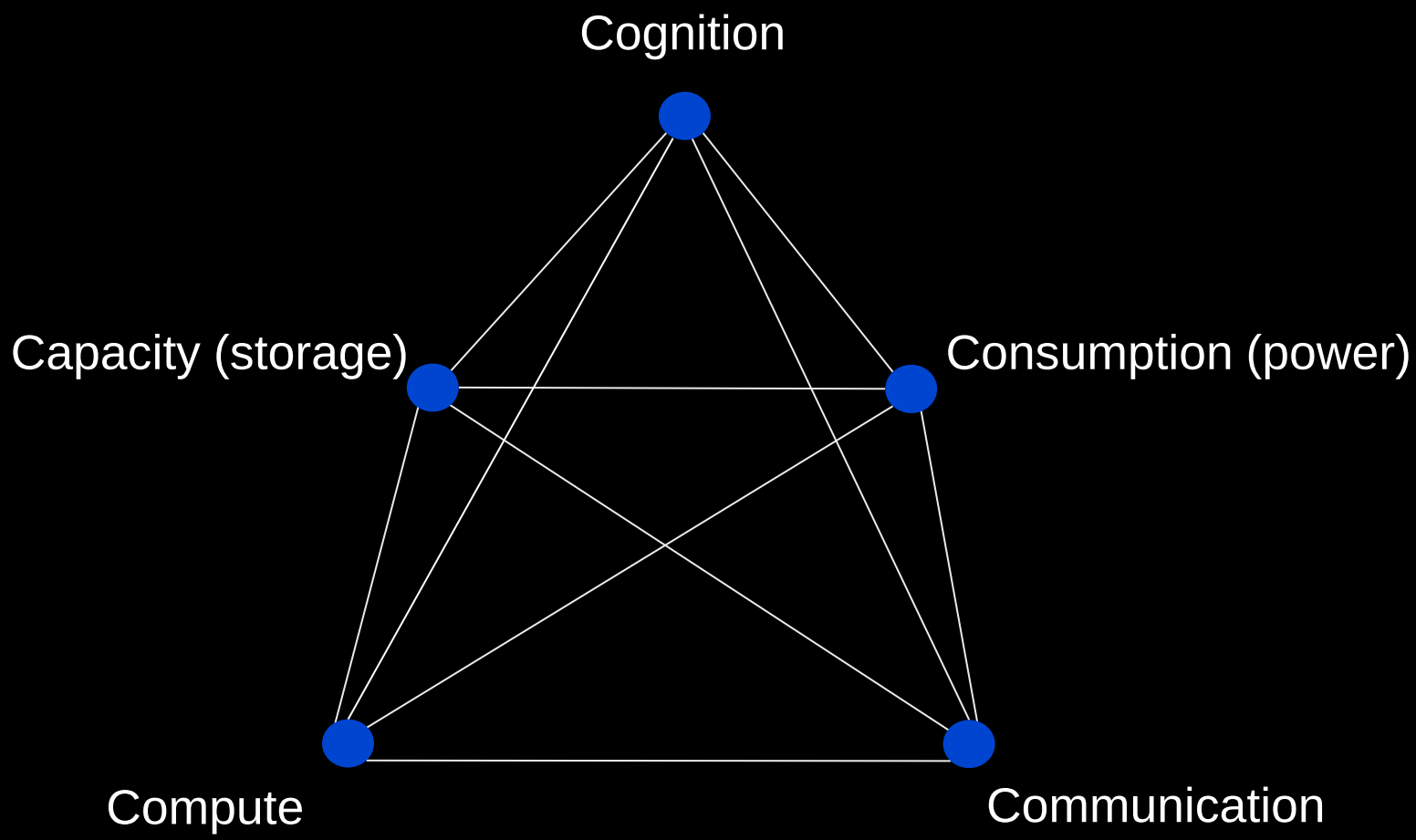
Duality in AI Efficiency

We are making AI computations more efficient

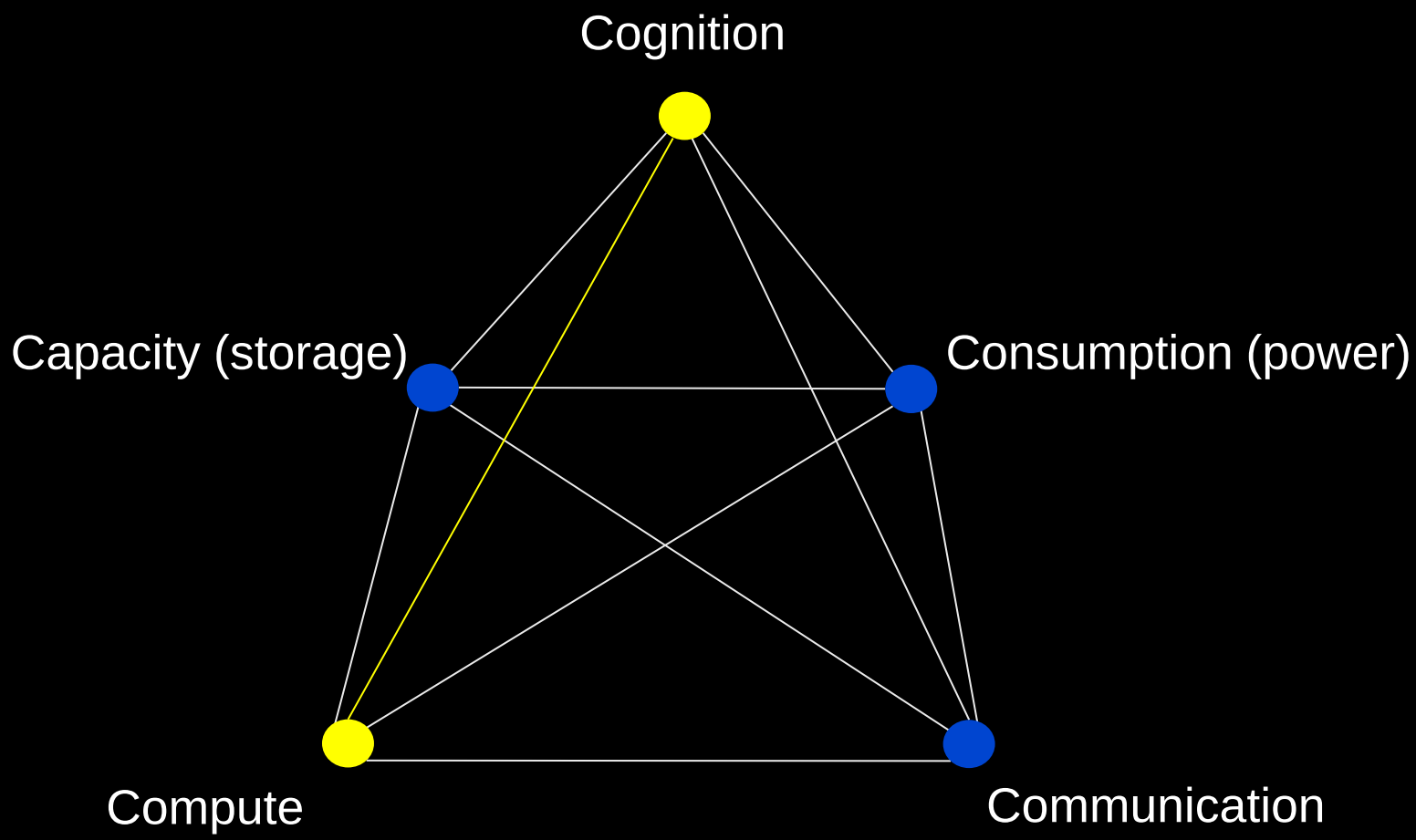
We are using AI to make computations more efficient

The Five “C”s

Thinking About Products: The 5 “C”s



Thinking About Products: The 5 “C”s



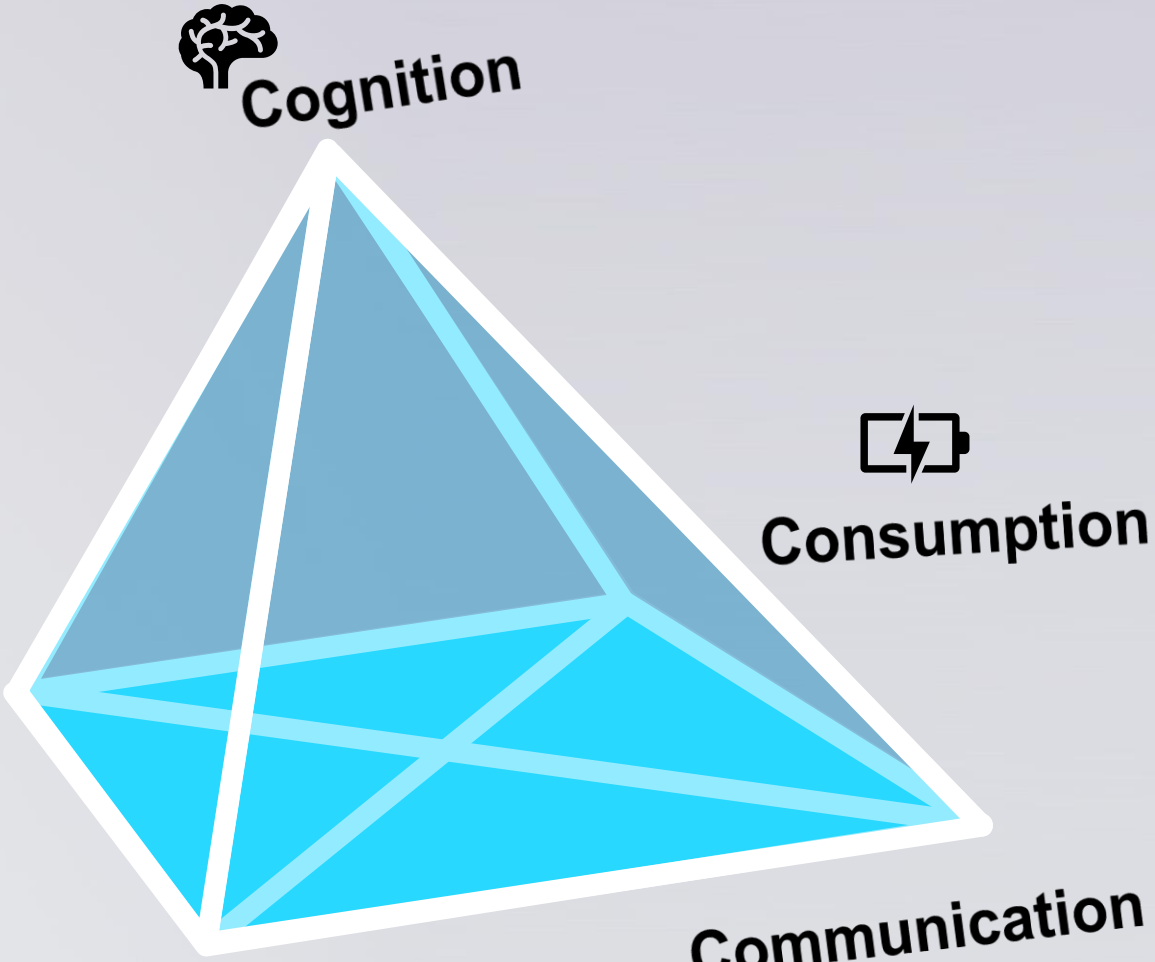
From Data Center to the Edge

AI Compute Server

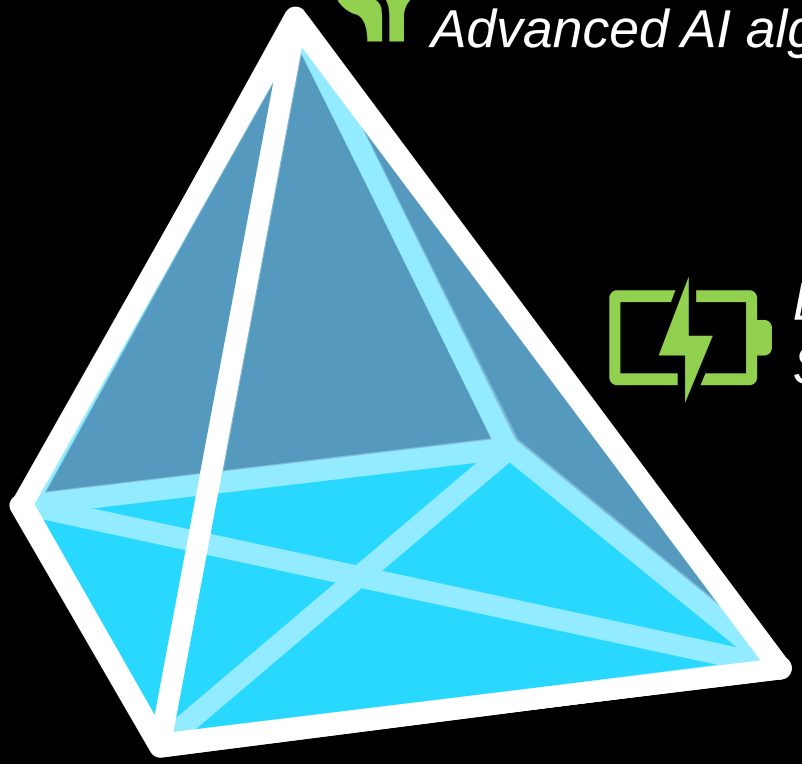
Nanny Cam

CPUs, GPUs, FPGAs	**Compute**	Microcontroller
Network switches, PCIe	**Communication**	Short-range wireless
DRAM, flash, hard disks	**Capacity**	Flash memory
Large training model	**Cognition**	Local inference
Grid connection (megawatts)	**Consumption**	4 AA batteries (milliwatts)

“Where Does Efficiency Come In?”



Efficiency and the 5 “C”s



Advanced AI algorithms

Liquid cooling for heat removal
Sharing of pooled resources

Reduced training data sets
Sparse data representation

Recompute to reduce data movement
Optical interconnect for data movement

Higher compute density per node

Don’t drown…

Surf!

System-Level Efficiencies

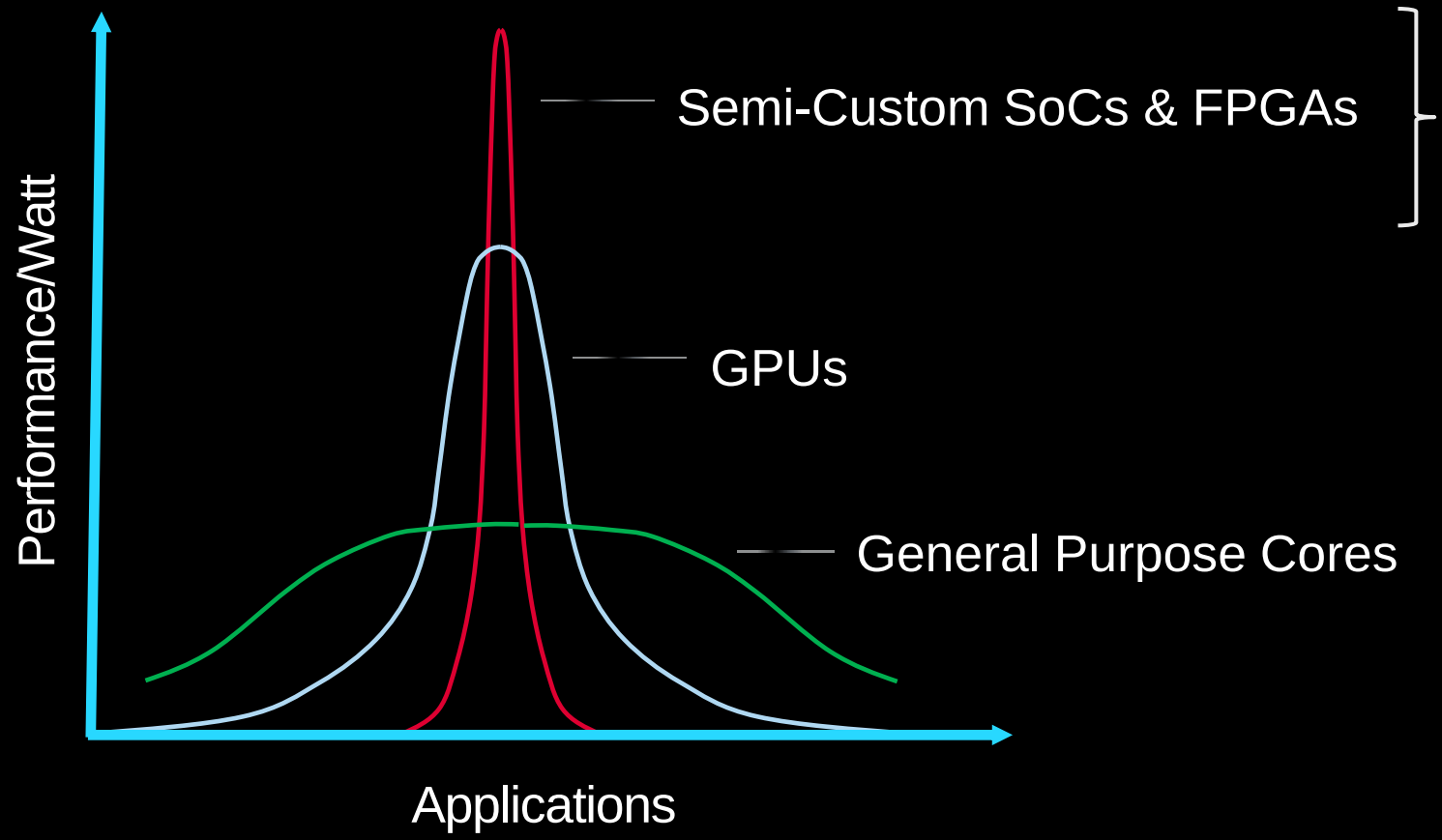
System Efficiency Issues

Large problems can map across 1000s of compute nodes.

It is difficult to move data efficiently within a node.

Extensive communication is required between nodes.

A Partial Solution: Heterogeneous Computing

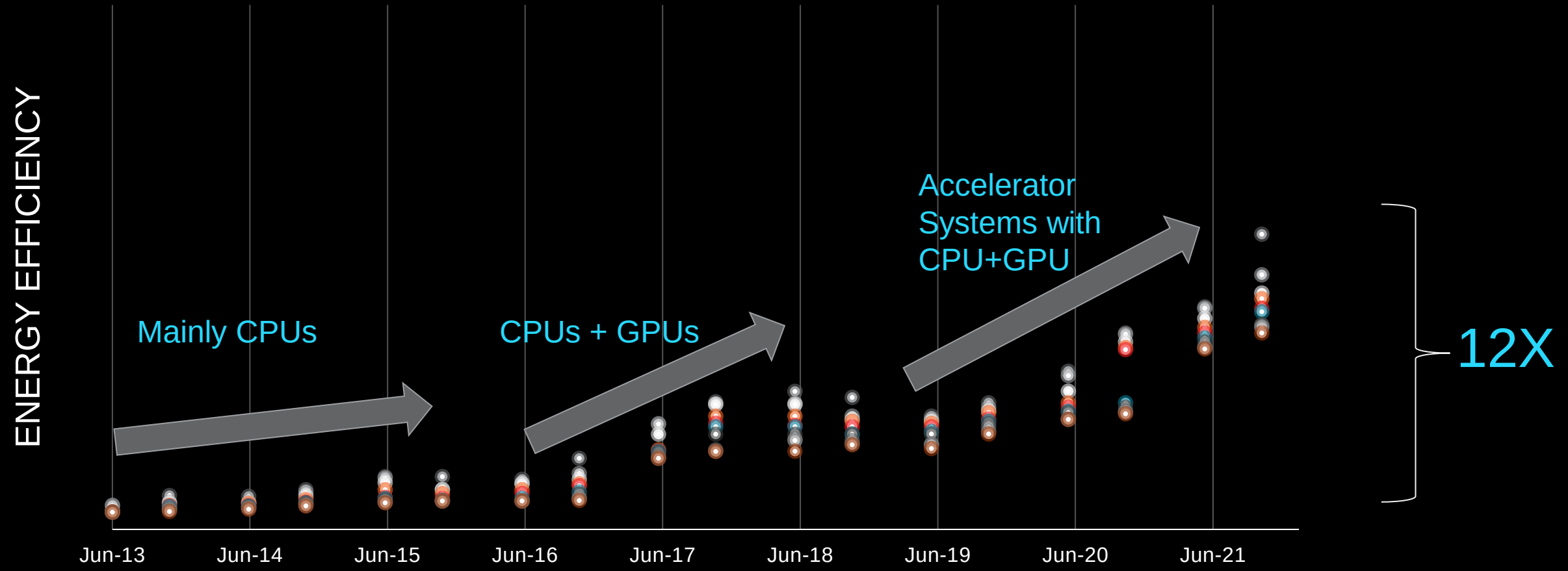


Strong focus on *compute*, but one needs to consider other C's.

Chart for Illustrative Purposes

What Has Worked So Far

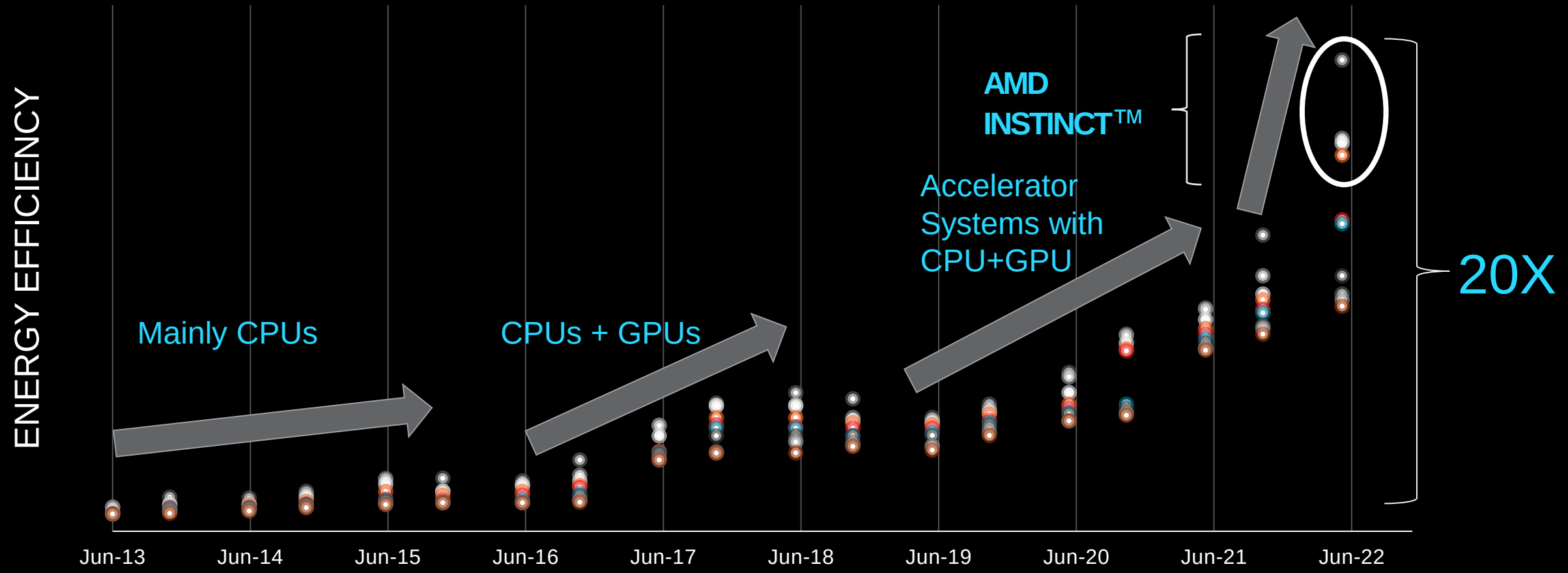
Top 10 Green500 Supercomputers



Source: Green500 list from top500.org

Impact of Multi-pronged Approach

Top 10 Green500 Supercomputers



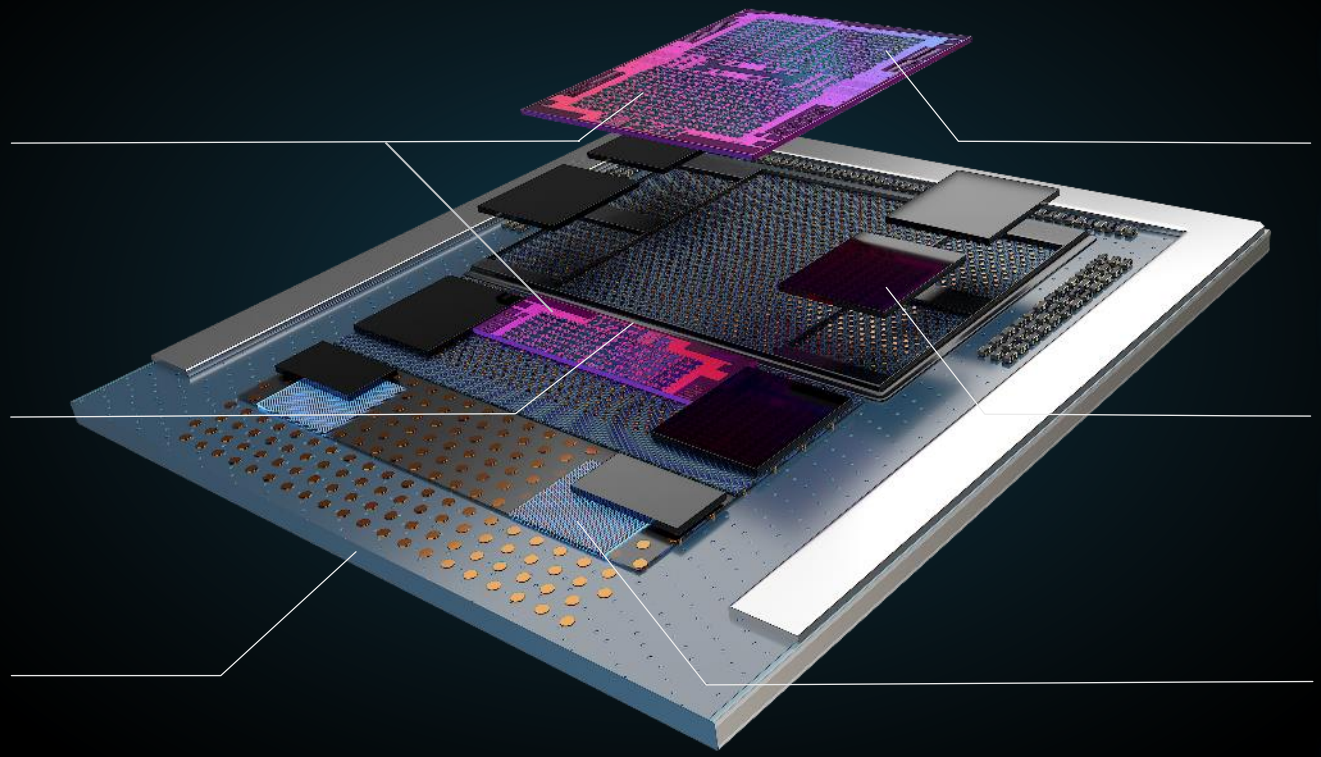
Source: Green500 list from top500.org

AMD Instinct™ MI200 Series

Compute
Two
AMD CDNA™2 Dies

Communication
Ultra High Bandwidth
Die Interconnect

Communication
Coherent CPU-to-GPU
Interconnect



Compute
2nd Gen Matrix Cores
for HPC & AI

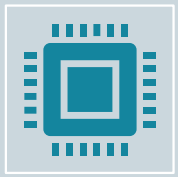
Capacity
Eight Stacks
of HBM2E Memory

Communication
2.5D Elevated
Fanout Bridge (EFB)

AMD Instinct™ MI200 OAM Series

Microarchitectural Efficiency

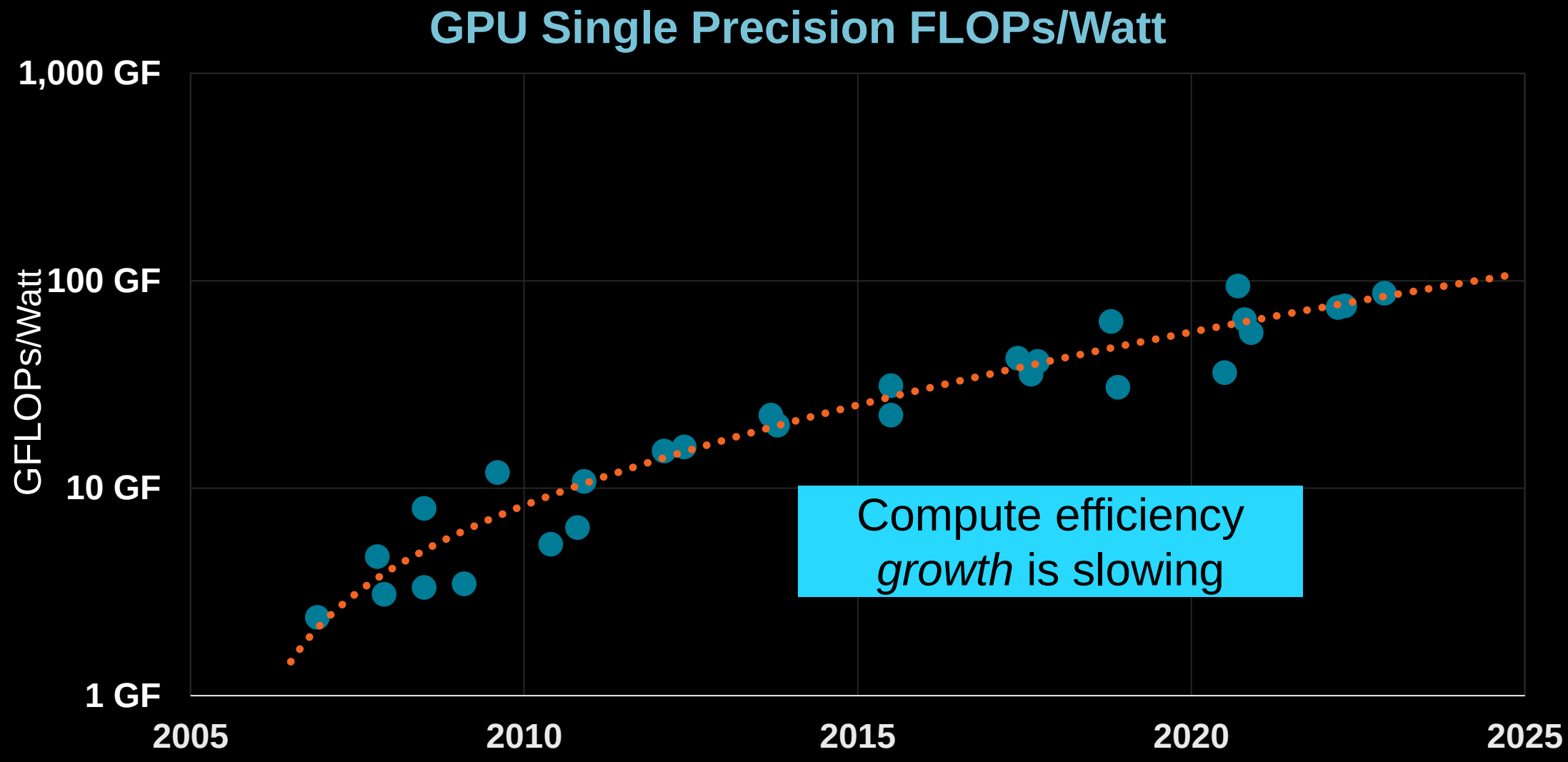
Microarchitectural Efficiency Issues



Each GPU performs trillions of floating-point operations per second in large machine learning and data science applications.

However, GPU floating-point efficiency is not scaling.

Compute Efficiency Trends



Borrow From Nature

Efficient Data Encoding:
Maximize relevant data processing
Minimize compute time

Borrow From Nature

Efficient Data Encoding:
Maximize relevant data processing
Minimize compute time

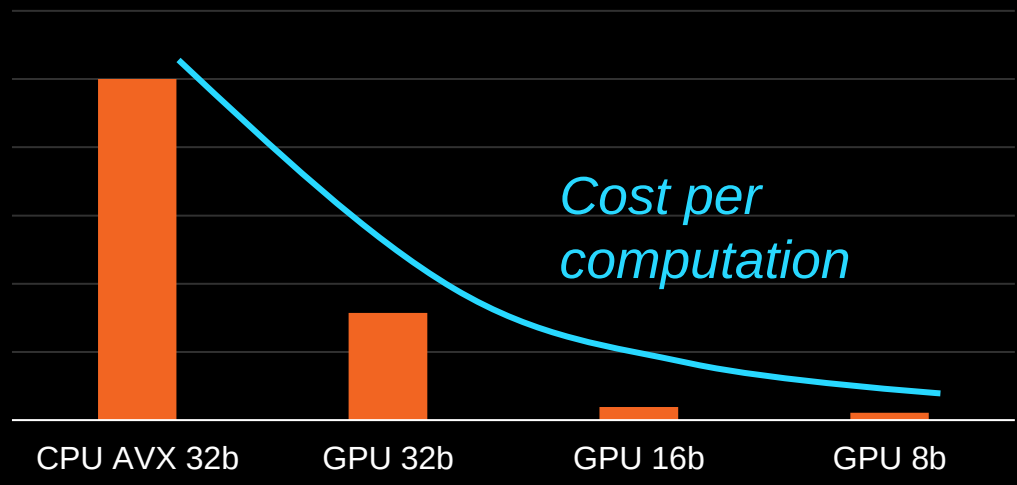
Reduce Precision:
Perceive features, not pixels

Lower-Precision Math Formats

- In AI, lower-precision math is often sufficient for convergence:

  32bit → 16bit → 8bit

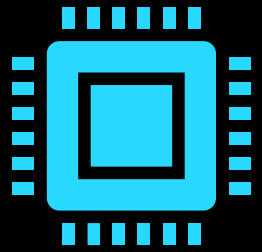
- This can reduce compute, capacity, and communication requirements.
- Energy savings can exceed 10X per operation[1].



1. AMD internal modeling and analysis

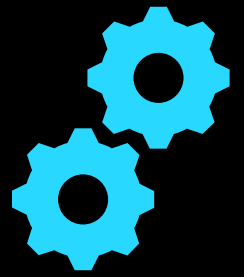
Programmability

Programmability Issues



Hardware is only as efficient as the software written for it.

Heterogeneous hardware complicates programmability.



Hardware efficiencies cannot result in programming inefficiencies.

Taming Programmability Issues

Reduce programming complexity

- Minimizes cognitive load
- Programming abstractions facilitate dynamic runtime adaptivity

Reduce performance optimization effort

- AI-assisted optimization to guide developers toward high-value code changes

Automate performance tuning

- Platform-aware auto-tuning to auto-discover optimal configurations

Reducing Software Complexity

Advanced Runtimes

Guided HW Optimization

Advanced Tooling

Performance Analysis at Scale

Cautionary Statement

This presentation contains forward-looking statements concerning Advanced Micro Devices, Inc. (AMD) such as AMD's vision and mission; AMD's expanded opportunities with acquisitions of Xilinx and Pensando; AMD's long-term growth opportunities and total addressable markets; AMD's next five years and strategic pillars; AMD's technology and architecture roadmaps; the features, functionality, performance, availability, timing and expected benefits of future AMD products and product roadmaps; AMD's path forward in data center, PCs and gaming; and AMD's market and financial momentum, which are made pursuant to the Safe Harbor provisions of the Private Securities Litigation Reform Act of 1995. Forward-looking statements are commonly identified by words such as "would," "may," "expects," "believes," "plans," "intends," "projects" and other terms with similar meaning. Investors are cautioned that the forward-looking statements in this presentation are based on current beliefs, assumptions and expectations, speak only as of the date of this presentation and involve risks and uncertainties that could cause actual results to differ materially from current expectations. Such statements are subject to certain known and unknown risks and uncertainties, many of which are difficult to predict and generally beyond AMD's control, that could cause actual results and other future events to differ materially from those expressed in, or implied or projected by, the forward-looking information and statements. Investors are urged to review in detail the risks and uncertainties in AMD's Securities and Exchange Commission filings, including but not limited to AMD's most recent reports on Forms 10-K and 10-Q. AMD does not assume, and hereby disclaims, any obligation to update forward-looking statements made in this presentation, except as may be required by law.

NON-GAAP FINANCIAL MEASURES In this presentation, in addition to GAAP financial results, AMD has provided non-GAAP earnings per share. AMD uses a normalized tax rate in its computation of the non-GAAP income tax provision to provide better consistency across the reporting periods. For fiscal 2019, 2020 and 2021, AMD used a non-GAAP tax rate of 3%, 3% and 15%, respectively, which excluded the direct tax impacts of pre-tax non-GAAP adjustments. AMD is providing the financial measures because it believes this non-GAAP presentation makes it easier for investors to compare its operating results for current and historical periods and also because AMD believes it assists investors in comparing AMD's performance across reporting periods on a consistent basis by excluding items that it does not believe are indicative of its core operating performance. The non-GAAP financial measures disclosed in this presentation should be viewed in addition to and not as a substitute for or superior to AMD's reported results prepared in accordance with GAAP and should be read only in conjunction with AMD's Consolidated Financial Statements prepared in accordance with GAAP. These non-GAAP financial measures referenced are reconciled to their most directly comparable GAAP financial measures in the Appendices at the end of this presentation.

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


AMD