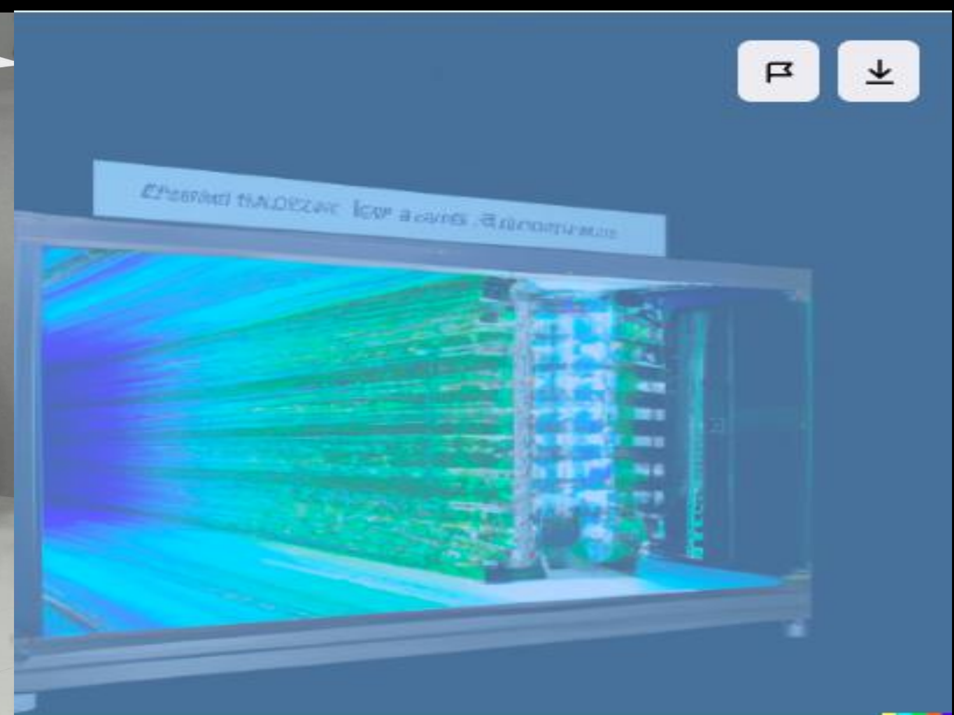
Exponential Technologies: The Role of High-Performance Computing in Shaping the Future of AI

- **Horst Simon, horst.simon@adialab.ae**

- **DP2E-Ai workshop, Paris, June 17, 2025**



ADIA Lab

We aim to **connect** and **collaborate** with the world's brightest minds in date and computational science to promote a culture of scientific thinking.

About ADIA Lab

- ADIA Lab was established on UAE National Day 2022. We play an active role in the growth of the UAE's scientific community and digital ecosystem.
- We engage in and foster basic and applied research in Data Science, Artificial Intelligence (AI) and Machine Learning (ML), High-Performance Computing (HPC) including Quantum Computing. Our research focuses on societally-important applications, and our main areas of interest are Climate Science, Health Sciences, and Digital Finance.
- ADIA Lab is guided by an Advisory Board of global thought leaders, including winners of the Nobel, Turing, Rousseeuw, Godel, Gordon Bell awards, and governed by an Operations Board overseeing operational activities of ADIA Lab.
- Although we are supported by ADIA, we are a separate legal entity that pursues its goals independently from ADIA. Our key goal is to contribute to the continued development of Abu Dhabi's digital ecosystem, with a special emphasis on talent development and on projects that could lead to the creation of startups.

Research Areas and Collaborators

ADIA ⁝ Lab

UNIVERSIDAD DE GRANADA

Universitat de Girona

UNIVERSIDADE DA CORUÑA

جامعة خليفة
Khalifa University

Stanford University

EDHEC BUSINESS SCHOOL

SORBONNE UNIVERSITY ABU DHABI

MBZUAI

THE UNIVERSITY OF TENNESSEE KNOXVILLE

ENS ÉCOLE NORMALE SUPÉRIEURE

INTELLECTUAL VENTURES

Instituto de Salud Carlos III

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

NYU ABU DHABI

ADIA Lab - Activities

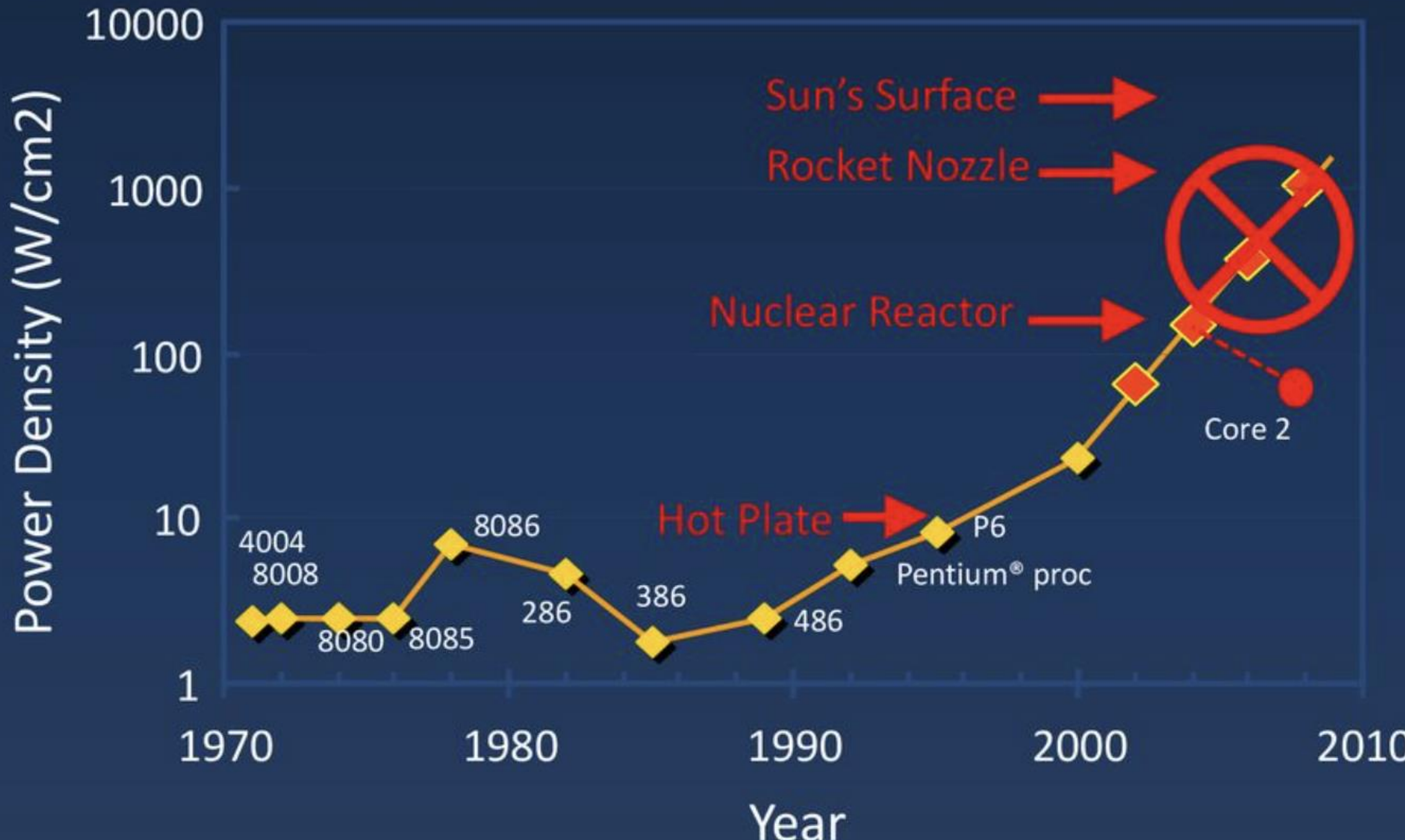
Activities

- Call For Papers for in Data Science for Climate (Award $100k) see https:// www.adialab.ae/call-for-papers
- Monthly Seminars, see https://www.adialab.ae/news-and-events
- ADIA Lab Structural Break Competition (Award pool $100K) , see https://www.crunchdao.com/live/adialab in collaboration with CrunchDAO
- Call for proposals in climate and health data science, deadline June 10, 2025
-

Activities 2025

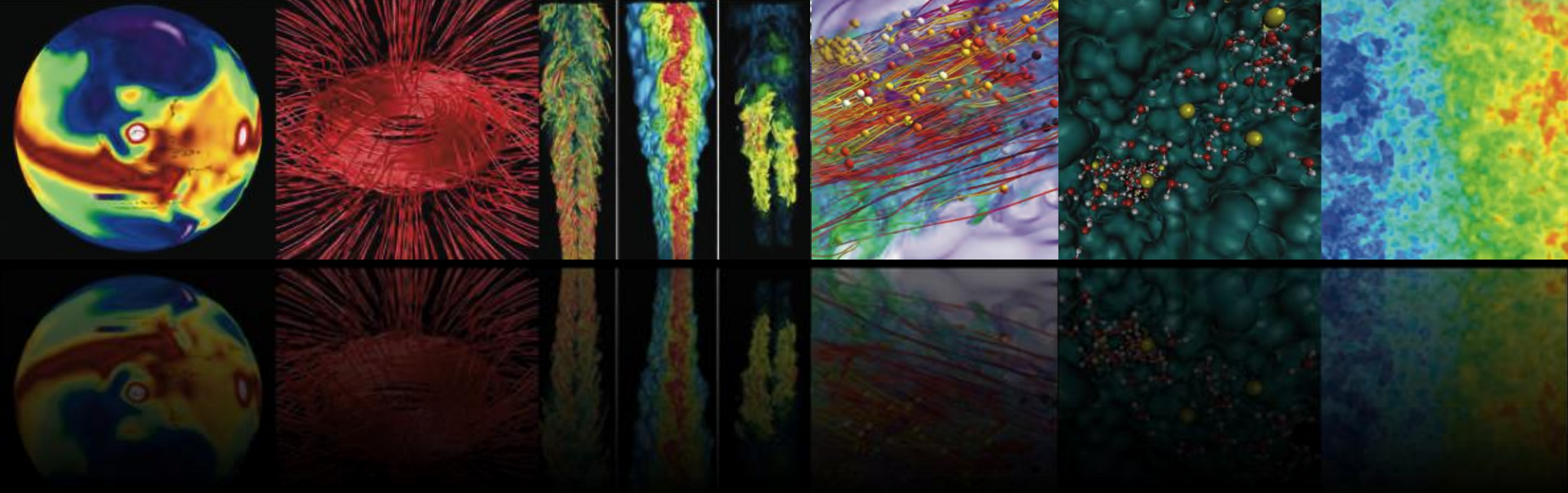
- Call for proposals in climate and health data science, deadline June 10, 2025
- Summer School in Shanghai, Sustainable AI, July 23 – August 1,
- ADIA Lab Symposium, Oct. 28 –Oct. 30, 2025, Abu Dhabi
- Hiring of permanent staff

Power Density Prediction circa 2000



Source: S. Borkar (Intel)

Adapted from UC Berkeley "The Beauty and Joy of Computing"

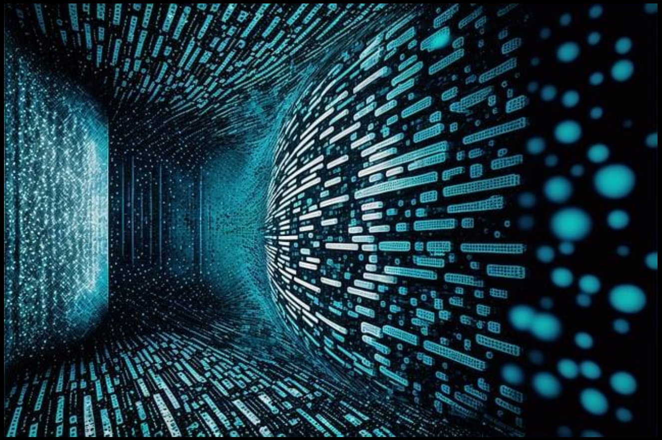


- **1** **How to think about AI**
- **2** HPC ≡ AI
  - The Path to Exascale and beyond
  - AI is the HPC killer app
- **3** What does this mean for benchmarking?

Think about AI

Think about AI again:
That’s what I see

Computers + Algorithms + Data



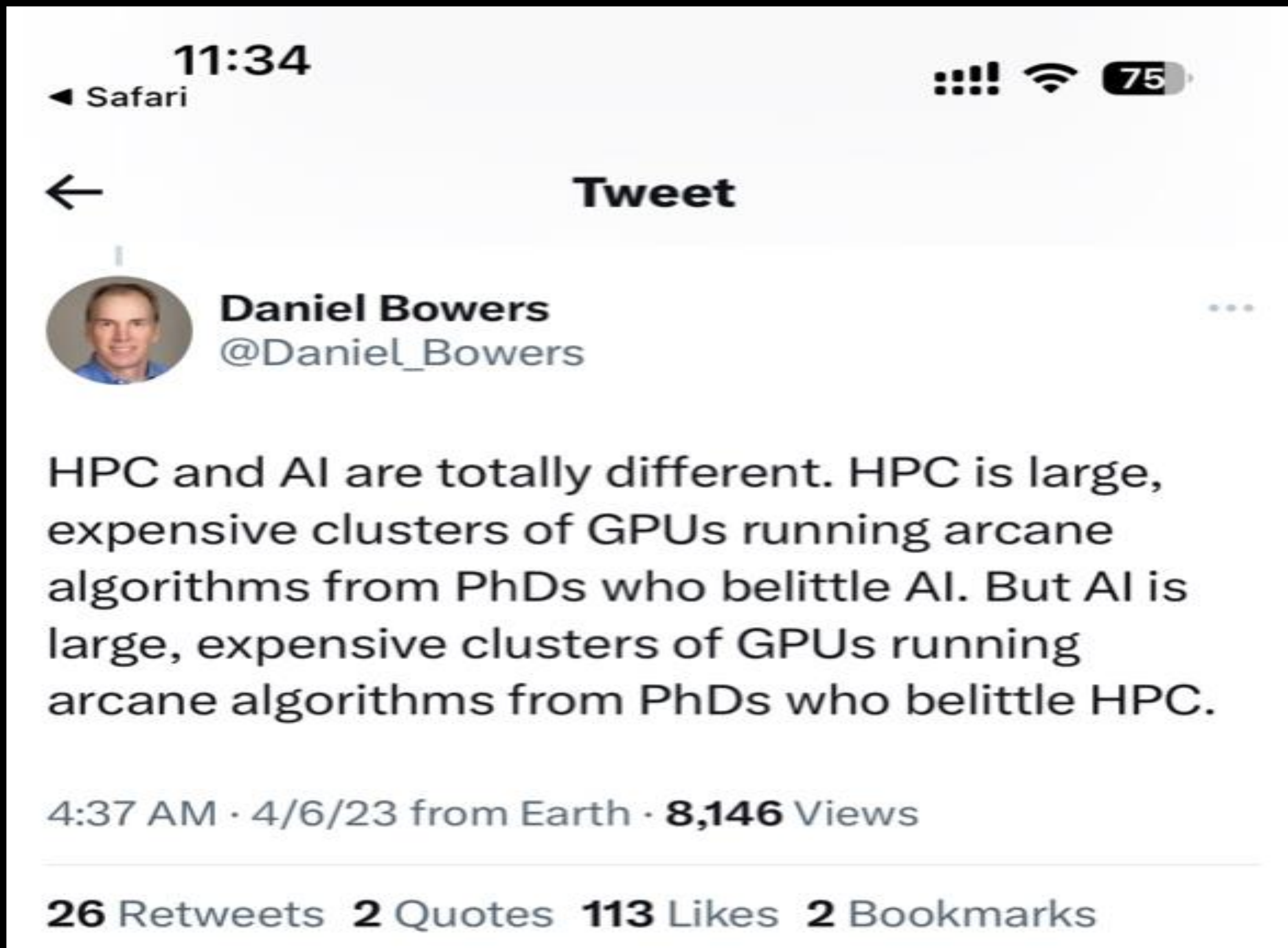
News stories and marketing material about Artificial Intelligence are typically illustrated with clichéd and misleading images.

Such depictions deflect accountability of the developers of AI, misrepresent its capabilities, and obscure the real societal and environmental impacts.

(Neema, “How do we picture Ai in our minds?”, 2023)

HPC ≡ AI

AI is an HPC killer app


◄ Safari 11:34 75

Tweet

Daniel Bowers
@Daniel_Bowers

HPC and AI are totally different. HPC is large, expensive clusters of GPUs running arcane algorithms from PhDs who belittle AI. But AI is large, expensive clusters of GPUs running arcane algorithms from PhDs who belittle HPC.

4:37 AM · 4/6/23 from Earth · **8,146** Views

26 Retweets **2** Quotes **113** Likes **2** Bookmarks

HPC Driven Innovation in AI and Machine learning

HPC has become a transformative force in AI and ML, powering advance by handling masssive data sets snd complex algorithms with unparalled speed and efficiency.

- HPC accelerates deep learning through massive parallel computations
- HPC enables training large language models and computer vision applications, advances in AI algorithms, like reinforemetn learning and generative AI

After Jack Dongarra, presentation at ADIA Lab Symposium

It is impossible to separate “HPC” from “AI”

<u>Traditional HPC</u>

Expanded to include AI workloads

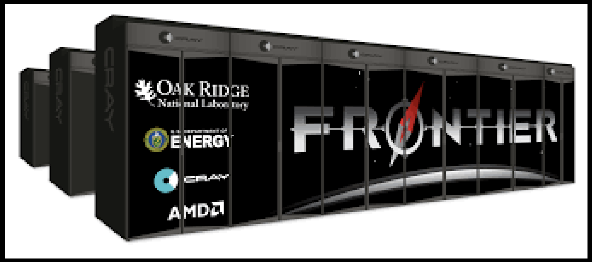
<u>Enterprise AI</u>

New organizations who acquire HPC for AI

<u>Hyperscale AI</u>

Large organizations with AI focused cloud infrastructure

after Intersect360 research



Traditonal HPC:
Mostly HPC, but also AI

Enterprise AI:
Mostly AI, but HPC based

Hyperscale AI:
Cloud environment, increasing AI investments

after Intersect360 research

What does this mean for measuring the progress in AI?

Benchmarking:
the process of evaluating the performance of a HPC system by running standardized tests or workloads to measure key metrics such as processing speed, memory bandwidth, I/O throughput, scalability, and energy efficiency.

Purpose:

A. To compare different HPC systems objectively.
B. To track progress in computational capabilities over time.
C. To focus design of computing architectures and deployments, guide system design, and procurement.
D. To communicate, market, and advertise our community's priorities to sponsors, customers, and the broader society.

The TOP500 Project
by Meuer, Strohmaier, Dongarra, Simon

Supercomputers

Hans Meuer (1936–2014)

- Listing of the 500 most powerful computers in the world
- Yardstick: Rmax of Linpack
  - Solve Ax=b, dense problem, matrix is random
  - Dominated by dense matrix-matrix multiply
- Updated twice a year:
  - ISC’xy in June in Germany
  - SCxy in November in the U.S.
- All information available from the TOP500 web site at: www.top500.org

#	Site	Manufacturer	TOP10 Computer of the TOP500	Country	Cores	Rmax [Pflops]	Power [MW]
1	Lawrence Livermore National Laboratory	HPE	**El Capitan** HPE Cray EX255a, AMD EPYC 24C 1.8GHz, Instinct MI300A, Slingshot-11	USA	11,039,616	1,742	29.6
2	Oak Ridge National Laboratory	HPE	**Frontier** HPE Cray EX235a, AMD EPYC 64C 2.0GHz, Instinct MI250X, Slingshot-11	USA	9,066,176	1,353	24.6
3	Argonne National Laboratory	Intel	**Aurora** HPE Cray EX/Intel Exascale Compute Blade, Xeon Max 9470, Data Center GPU Max, Slingshot-11	USA	4,742,808	1,012	38.7
4	EuroHPC / FZJ	EVIDEN	**JUPITER Booster** BullSequana XH3000, NVIDIA GH200 Superchip, InfiniBand NDR	Germany	4,801,344	793.4	13.1
5	Microsoft Azure	Microsoft	**Eagle** Microsoft NDv5, Xeon Platinum 8480C, NVIDIA H100, Infiniband NDR	USA	1,123,200	561.2	
6	Eni S.p.A. Center for Computational Science	HPE	**HPC6** HPE Cray EX235a, AMD EPYC 64C 2.0GHz, Instinct MI250X, Slingshot-11	Italy	3,143,520	477.9	8.5
7	RIKEN Center for Computational Science	Fujitsu	**Fugaku** Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D	Japan	7,630,848	442.0	29.9
8	Swiss National Supercomputing Centre (CSCS)	HPE	**Alps** HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, GH200, Slingshot-11	Switzerland	2,121,600	434.9	7.1
9	EuroHPC / CSC	HPE	**LUMI** HPE Cray EX235a, AMD EPYC 64C 2.0GHz, Instinct MI250X, Slingshot-11	Finland	2,752,704	379.7	7.1
10	EuroHPC / CINECA	EVIDEN	**Leonardo** Atos BullSequana XH2000, Xeon 32C 2.6GHz, NVIDIA A100, HDR Infiniband	Italy	1,824,768	241.2	7.5

How does TOP500/HPL do?

A. To compare different HPC systems objectively.
B. To track progress in computational capabilities over time.
C. To focus design of computing architectures and deployments, guide system design, and procurement.
D. To communicate, market, and advertise our community's priorities to sponsors, customers, and the broader society.

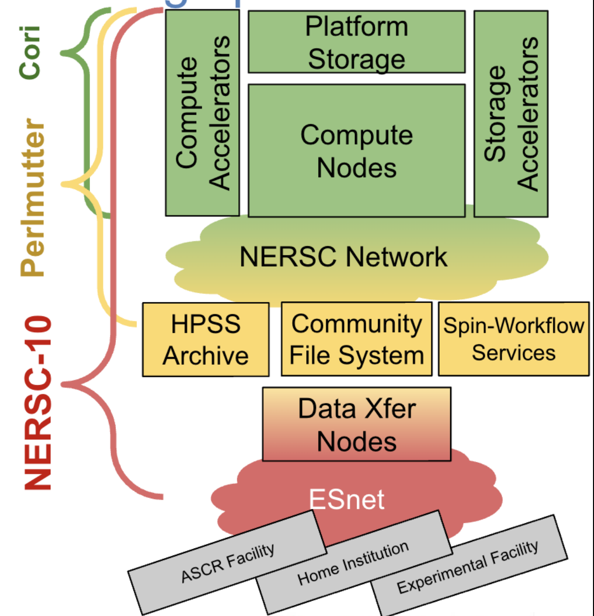
Current HPC benchmarking challenges:
1) Even without AI, the definition of an HPC system is changing

NERSC-10 Architecture: Designed to support complex simulation and data analysis workflows at high performance

NERSC-10 will provide on-demand, dynamically composable, and resilient workflows across heterogeneous elements within NERSC and extending to the edge of experimental facilities and other user endpoints

Complexity and heterogeneity managed using complementary technologies

- **Programmable infrastructure**: avoid downfalls of one-size-fits-all, monolithic architecture
- **AI and automation**: sensible selection of default behaviours to reduce complexity for users



NERSC 6 BERKELEY LAB Bringing Science Solutions to the World U.S. DEPARTMENT OF ENERGY Office of Science

From NERSC, 2022

Current HPC benchmarking challenges:

2) Future procurements will be for merged HPC+AI systems

HPC-AI Budgets

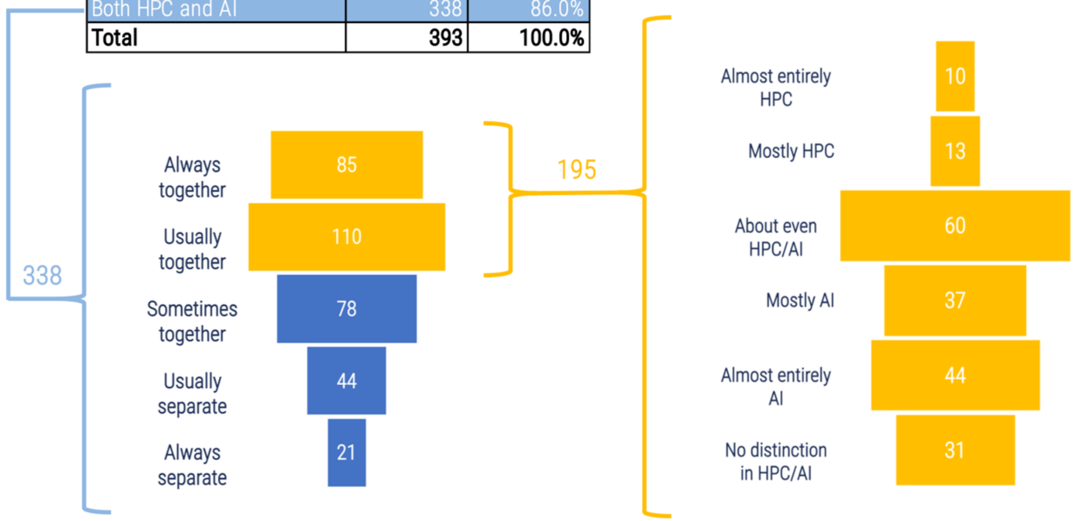
Among commercial (non-hyperscale) use cases, AI is usually found together with HPC

Almost all HPC users are now also AI

More often than not, HPC and AI are merged as part of the same budget

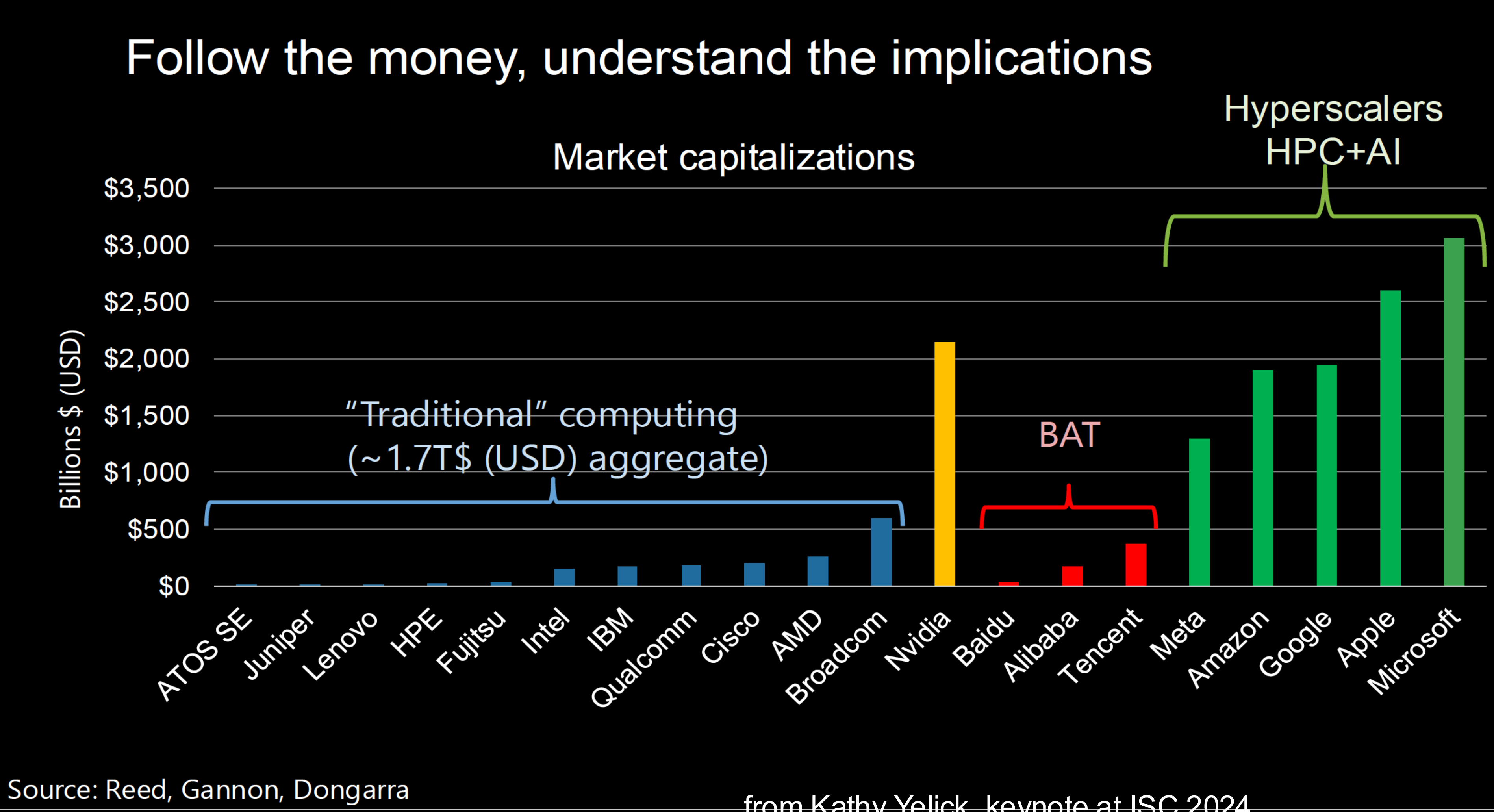
When merged, the budgets are more focused on AI than HPC

Budgets for HPC or AI	Count	Pct.
HPC only	15	3.8%
AI only	40	10.2%
Both HPC and AI	338	86.0%
Total	**393**	**100.0%**

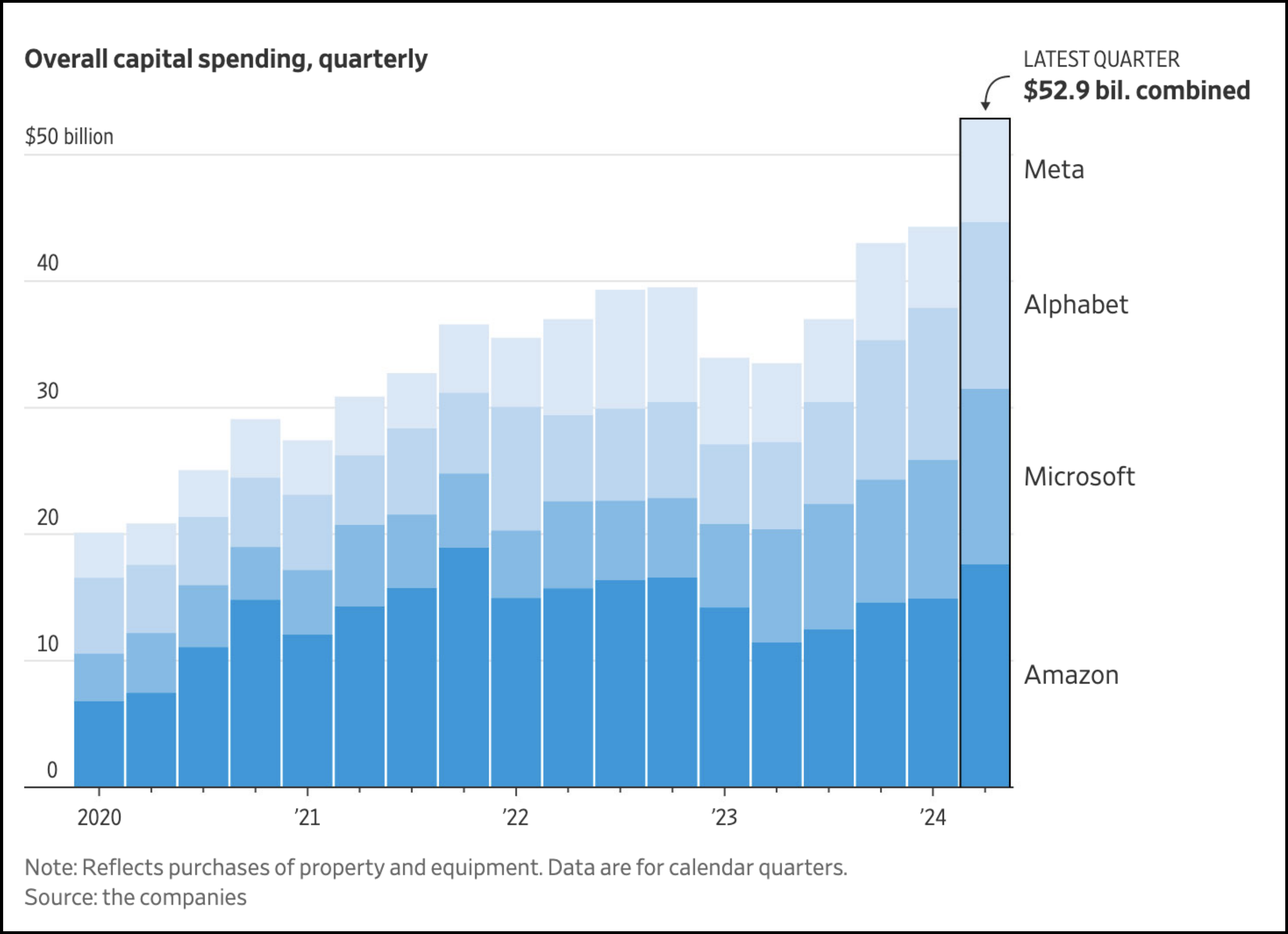


after Intersect360 research

Current HPC benchmarking challenges:
3) Lack of data on hyperscalers and AI factories



from Kathy Yelick, keynote at ISC 2024



Overall AI capital spending, source Wall Street Journal

EPOCH AI

Epoch AI has made significant progress in collecting data on AI Supercomputers and hyperscalers, see epoch.ai

How does Epoch AI define an AI supercomputer？

A computer system that can support training large-scale AI models, deployed on a contiguous campus, and meets **two criteria**:

1. **Chip requirement**: It uses AI-accelerating chips (e.g., NVIDIA V100, A100, H100, Google TPUs, etc.) or meeting technical criteria like support for FP16/INT8, tensor cores, and high-bandwidth memory.
2. **Performance threshold**: It achieves **at least 1% of the performance** of the most powerful AI supercomputer **operational at the time** (using 32-, 16-, or 8-bit FLOP/s as relevant for AI tasks).

This definition is back to the future: in 1985 a supercomputer was defined as being built by Cray, CDC, Hitachi, Fujitsu or NEC, and having vector processing.

How does Epoch AI collect its data on AI supercomputers ?

- Running **Google Search API queries** with terms like “AI supercomputer” or “GPU cluster” over defined date windows (2019–2025).
- Reviewing **company announcements**, **blog posts**, **TOP500 entries** (when GPU-heavy), and internal references from Epoch AI’s own model training dataset.
- Conducting **manual searches** to gather details such as:
  Number and type of chips
  Operational start date
  Theoretical performance (FLOP/s)
  Ownership and location

This process is also back to the future: in 1985 several supercomputer lists were published, based on company announcements and peak performance, like the Mannheim, list a precursor to the TOP500.

The energy challenge of HPC/AI

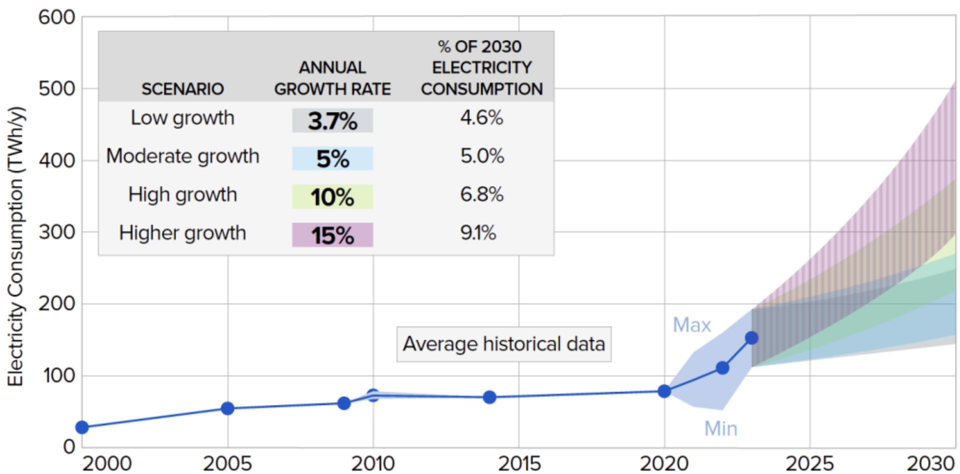


Figure ES-1. Projections of potential electricity consumption by U.S. data centers: 2023–2030 . % of 2030 electricity consumption projections assume that all other (non-data center) load increases at 1% annually.

EPRI White Paper, “Powering Intelligence”, 2024

Microsoft has signed a 20-year deal to exclusively access 835 megawatts of energy from a nuclear plant.

Elon Musk’s AI data center in Memphis Tennessee will need upwards of 70 megawatts of power to run all 100,000 GPUs 14 massive mobile generators to power the facility as he works out power supply agreements with local utilities.

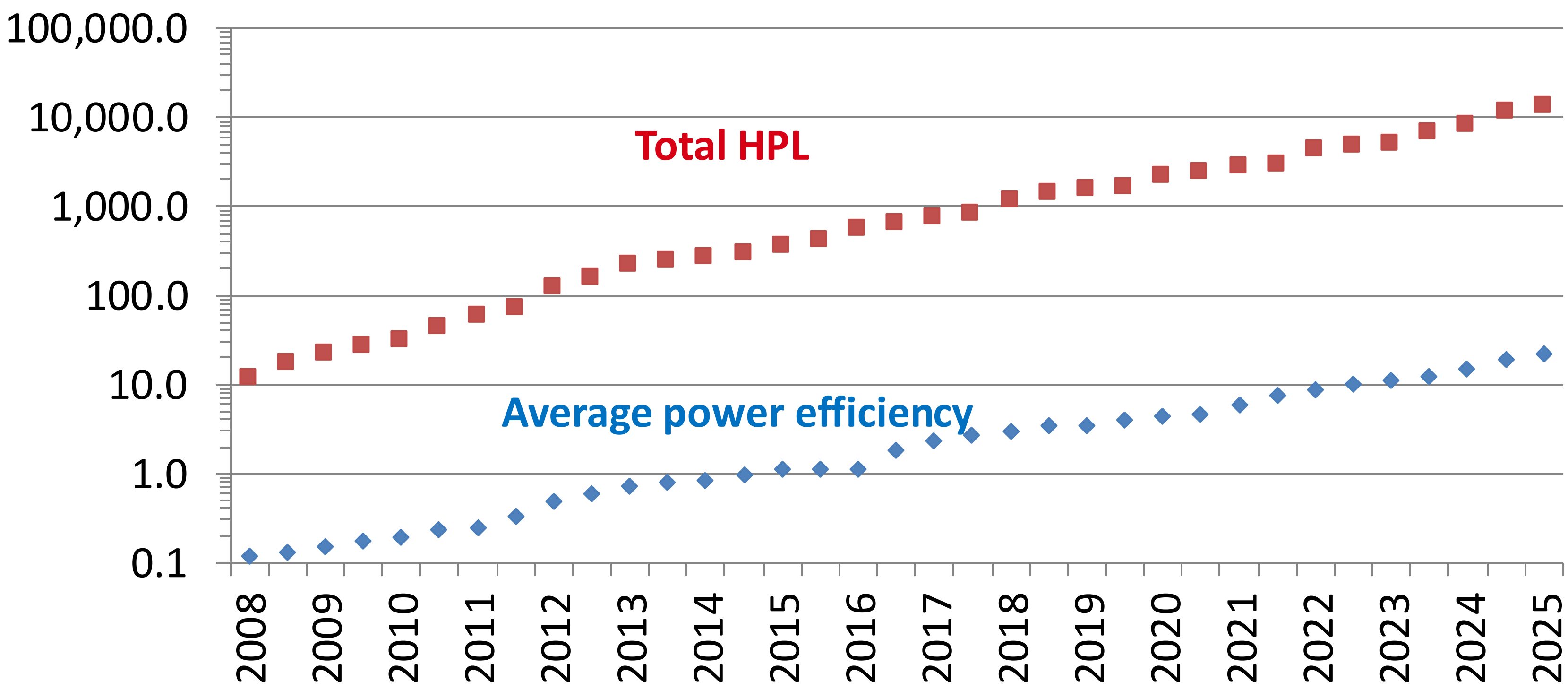
2.5 MW per truck X 14 trucks = 35 MW

They installed the 100,000 GPUs in 122 days and have ordered another 100,000 GPUs.

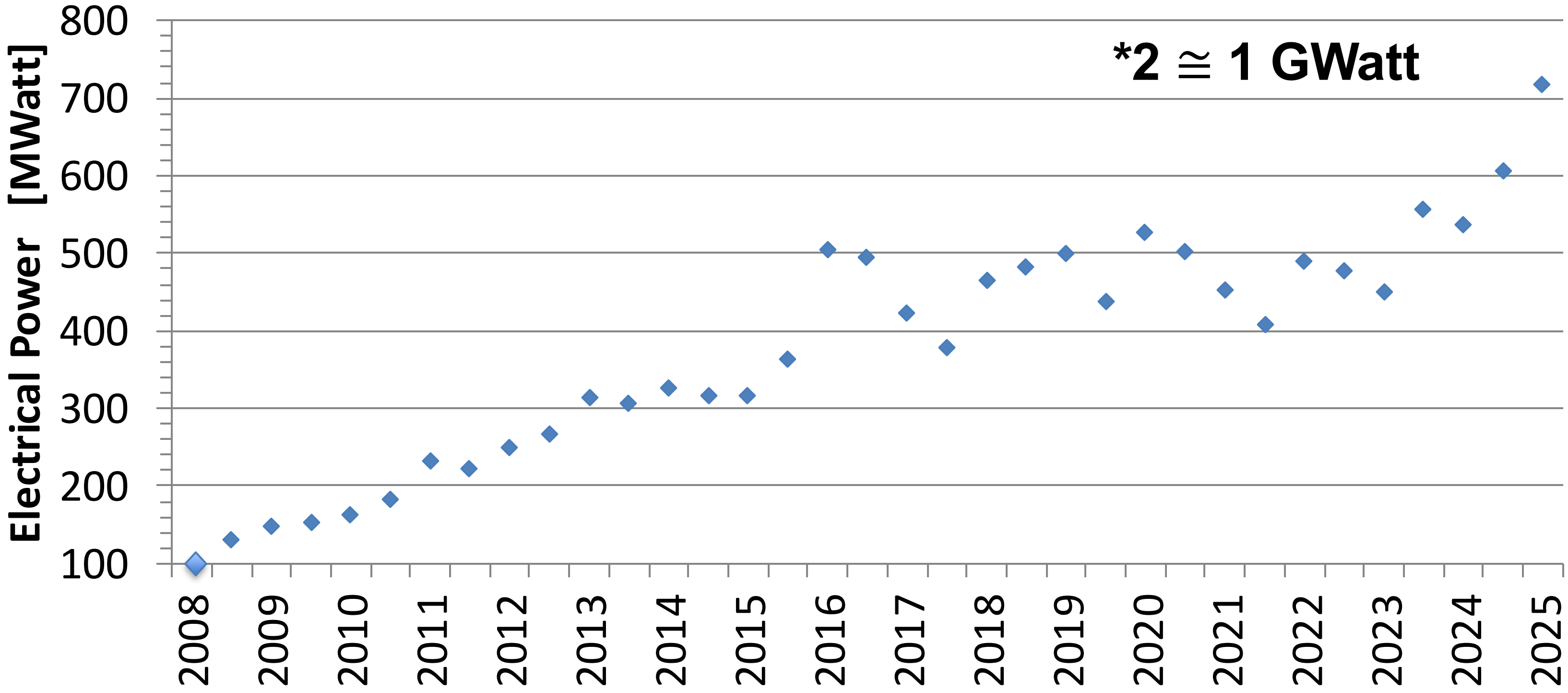
HYPERSCALERS – GIGAWATTS – A TANGENT

- There is talk about Gigawatt sized datacenter in the (near) future!
- Current *hyperscaler center* are typically in the range of 50-200 MWatts
  - TOP500: Highest measured power numbers are in the 30-40 MWatt range
  - But: measured vs name-plate + Storage and Infrastructure! (~Double it?)
  - -> Biggest system in TOP500 would qualify as *hyperscale*
- How "big" is the whole TOP500 in Watt?
  - We take total R_max (an exponential curve)
  - and "average" power efficiency (another exponential!)
  - and divide them to compute approximated total Watts - (be careful ...!)

POWER EFFICIENCY



TOP500 TOTAL POWER (APPROX)

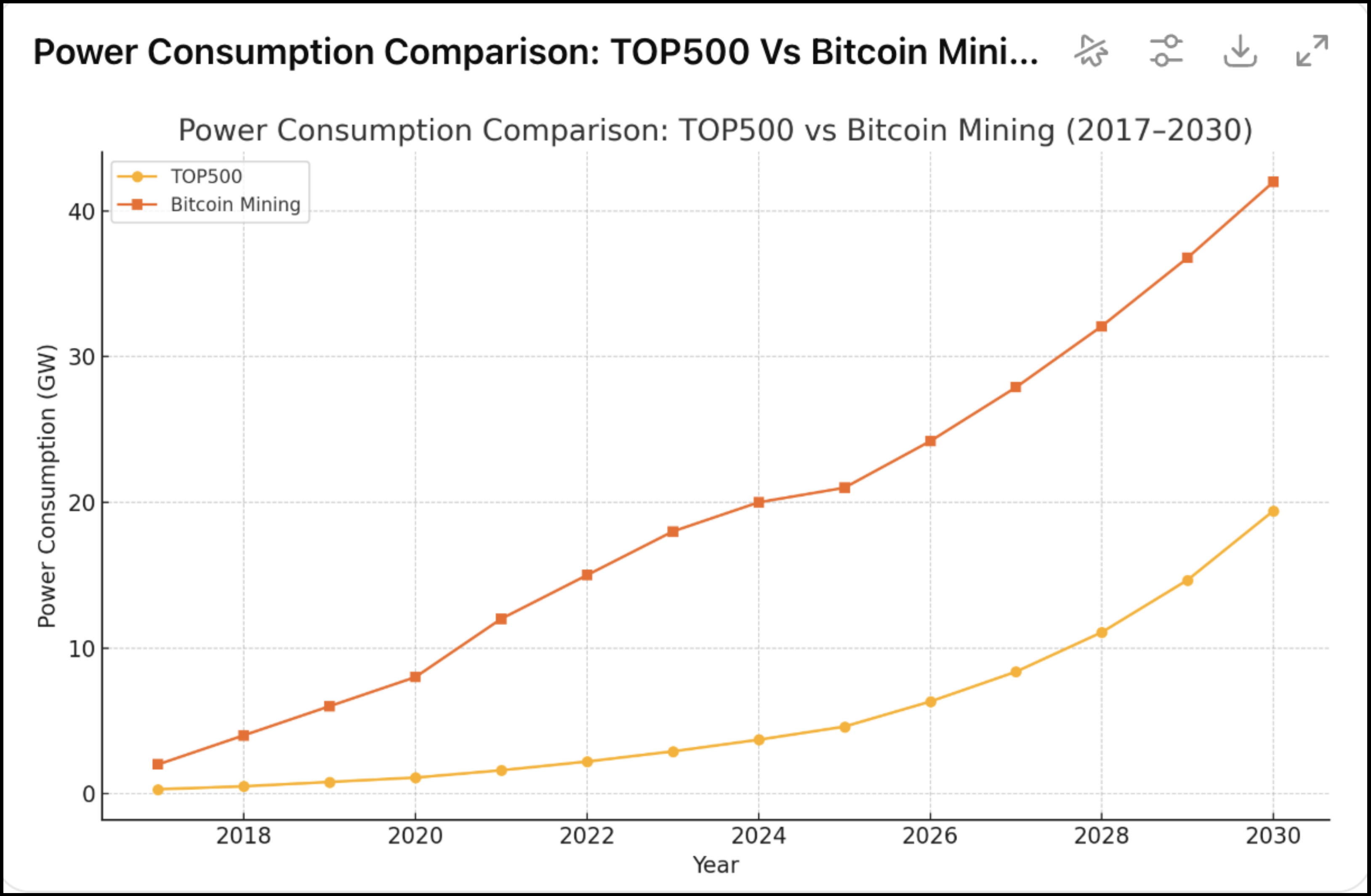


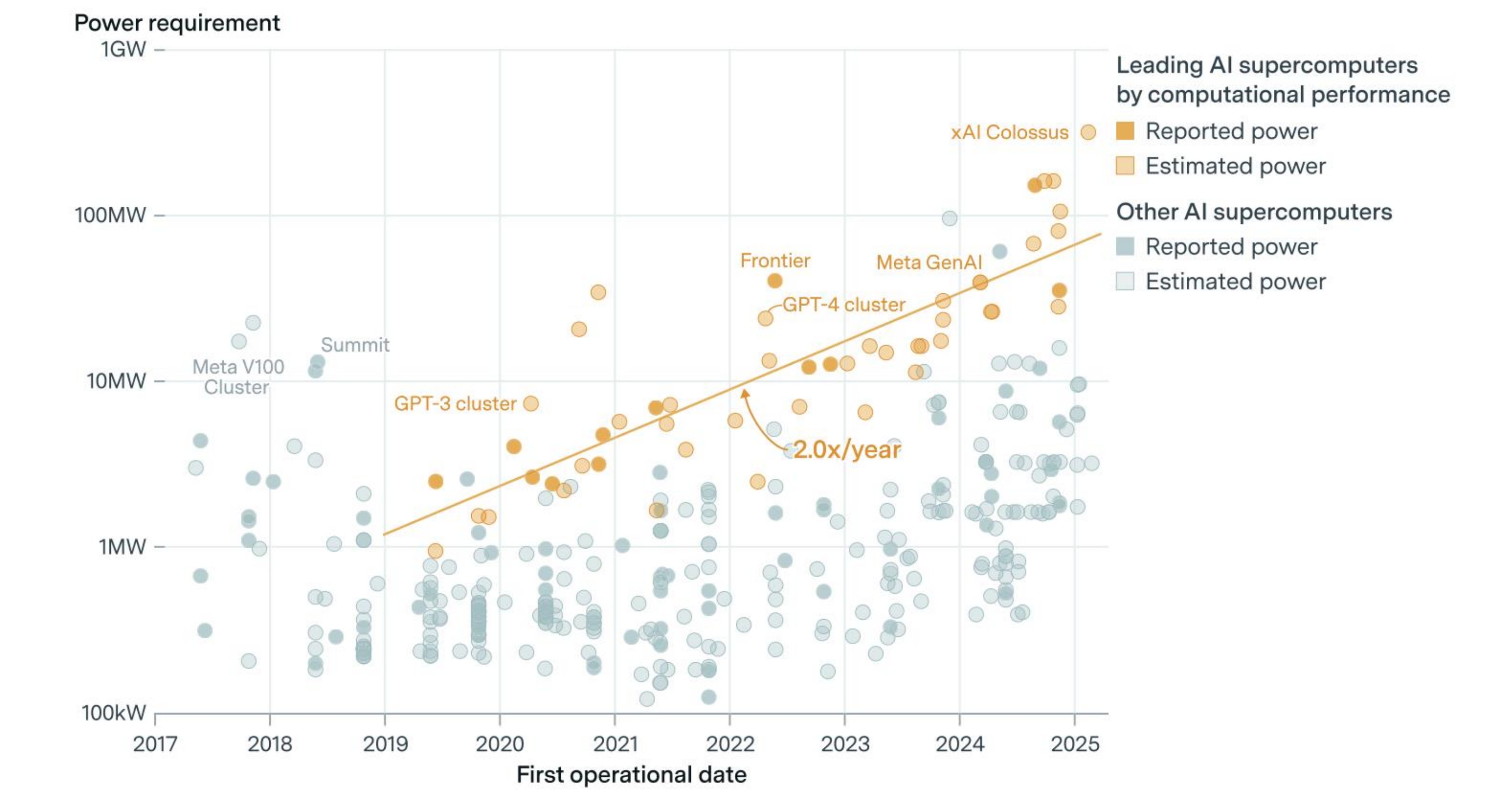
The energy challenge of crypto coin mining

From 14th Crypto Super 500 report:

- Bitcoin mining dominates: ~481 EH/s, far exceeds traditional supercomputers
- Power use > 20 GW, much higher than TOP500 (~4.6 GW in 2025)
- Compute via ASICs, not general-purpose processors
- Mining clustered where electricity is cheapest
- Shows massive compute scale beyond FLOPS (via Hashrate)
- Shift from CPUs/GPUs to ASICs since ~2013
- ASICs now dominate Bitcoin mining due to efficiency

Extrapolation of TOP500 and Crypto Super 500 reports

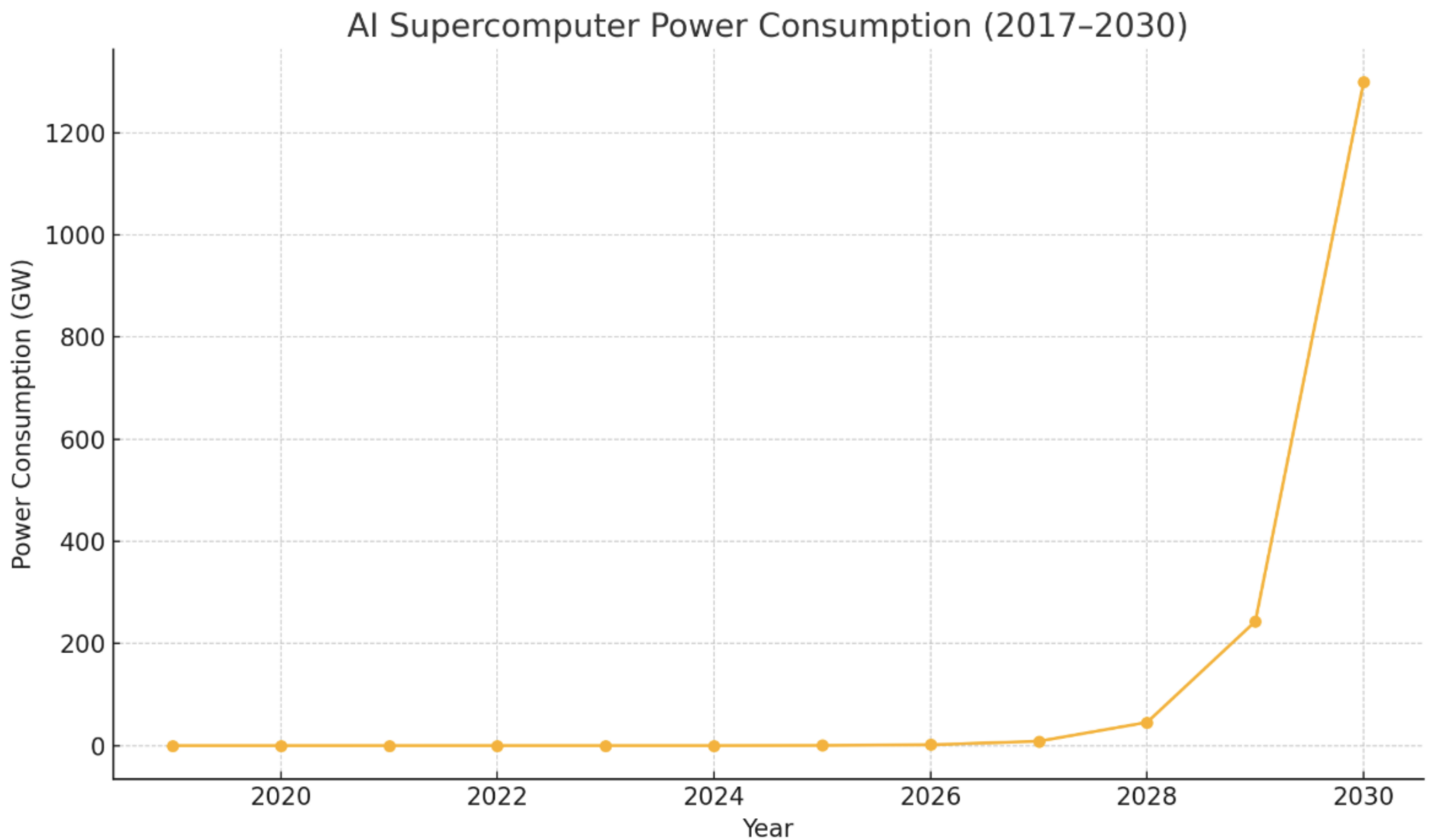


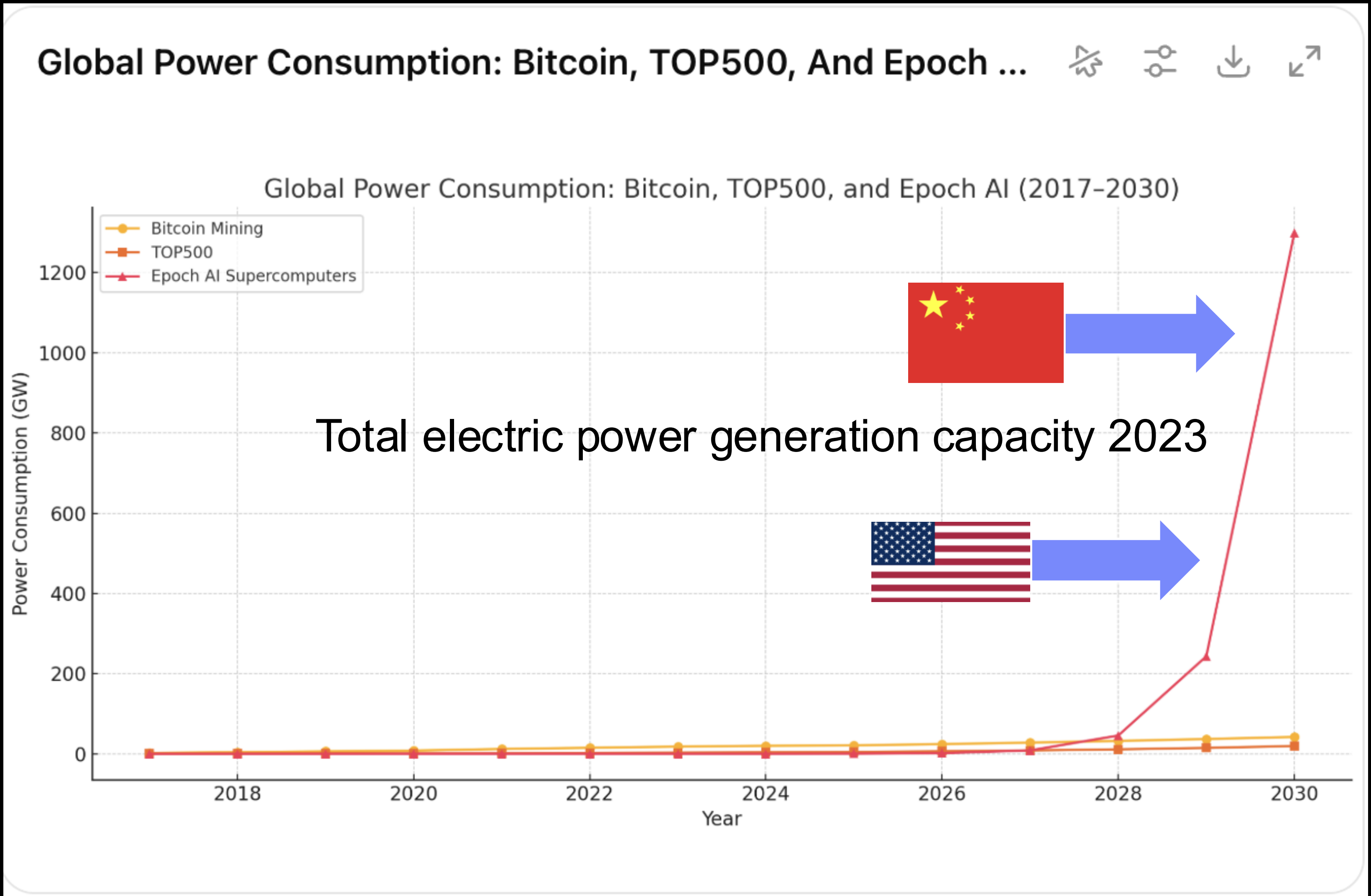


Epoch AI on AI Supercomputers

By 2030: AI supercomputers will demand up to **9 gigawatts (GW)**. This would support a system with around **2 million AI chips**, matching their prediction of a top-tier AI system costing $200 billion

AI Supercomputer Power Consumption (2017–2030)

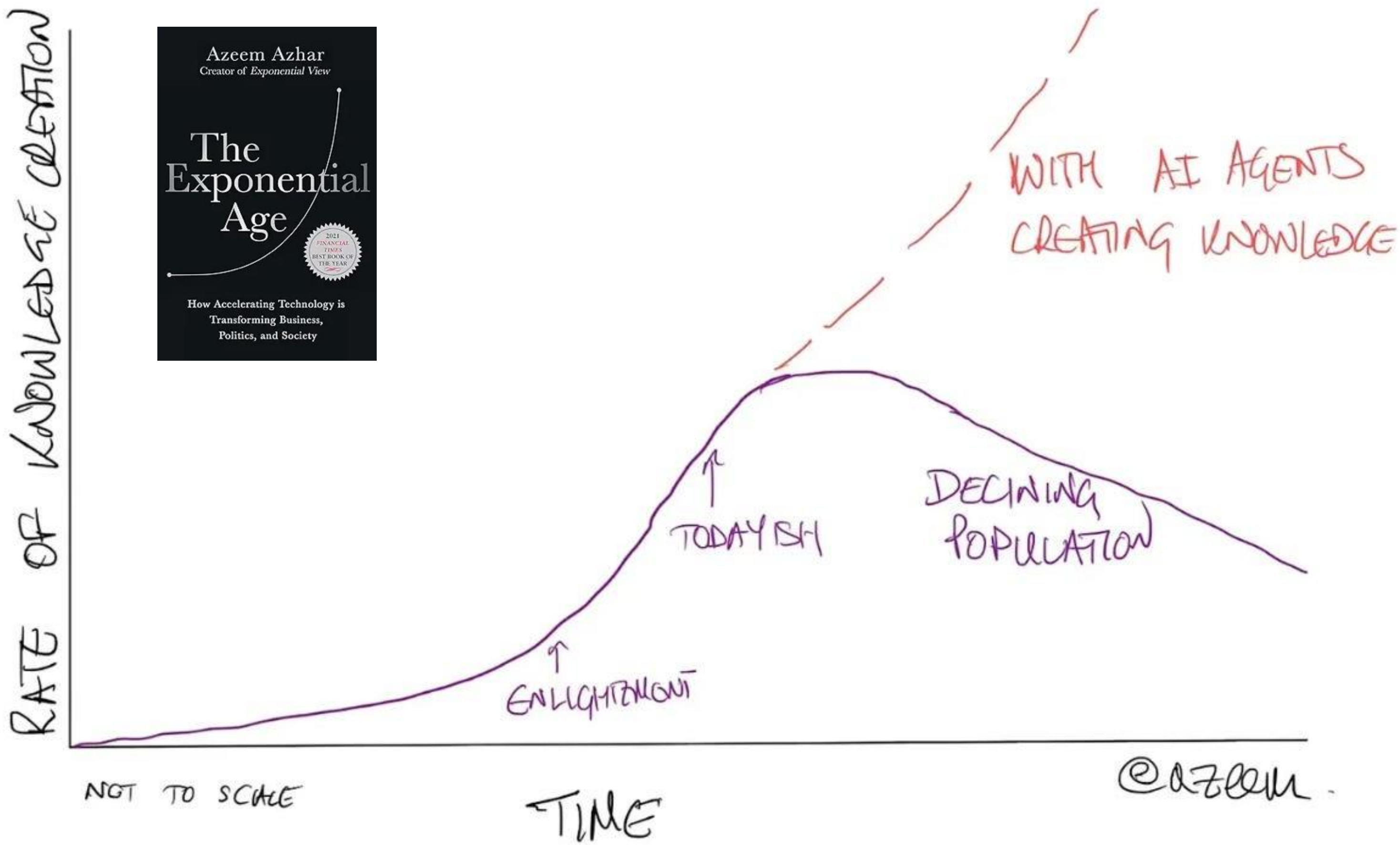



Global Power Consumption: Bitcoin, TOP500, And Epoch ...
Global Power Consumption: Bitcoin, TOP500, and Epoch AI (2017–2030)
Bitcoin Mining
TOP500
Epoch AI Supercomputers
Power Consumption (GW)
0
200
400
600
800
1000
1200
2018
2020
2022
2024
2026
2028
2030
Year
Total electric power generation capacity 2023

In Summary

- HPC will continue to advance in performance, energy efficiencey, AI integration, and hybridization with quantum and edge computing
- AI will continue to grow as exponential technology, leveraging HPC knowledge and investments
- There will be major technology challenges to future growth of AI (architecture, algorithms, and power use) in addition to the poltitical, economic, and social challenges.

Humanity's knowledge creation will depend on AI



From Azeem Azar, The Exponential Age

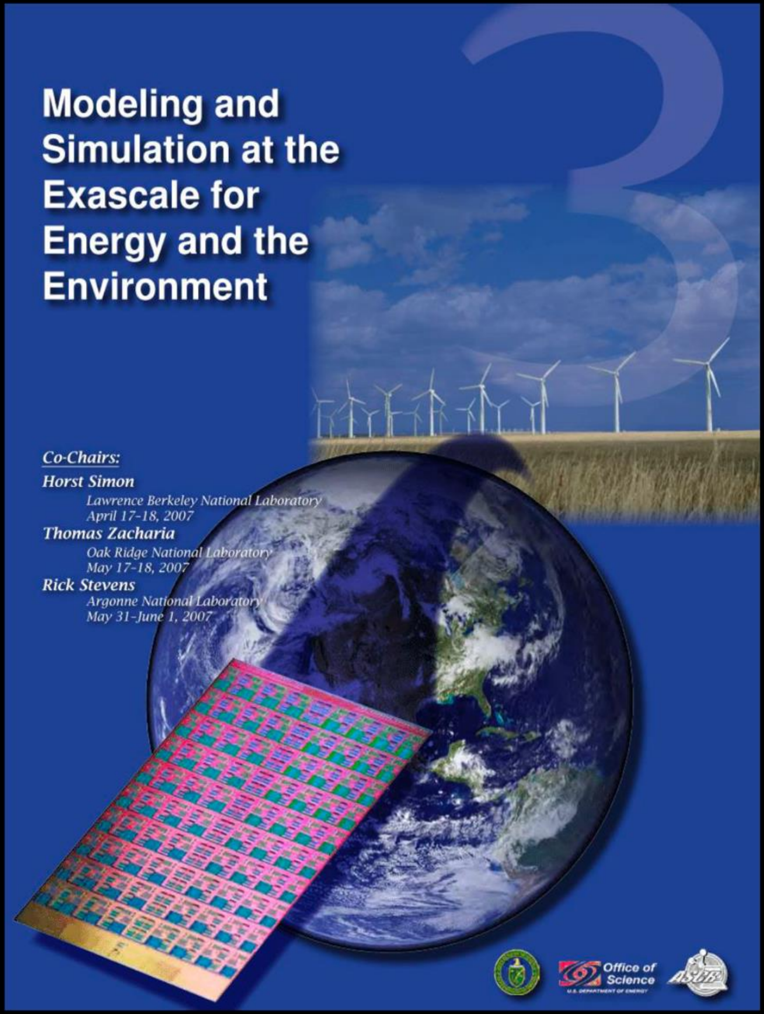
A final thought: predictions are always hard especially about the future

Extra Slides

Traditional HPC

Scientific Computing Circa 2007

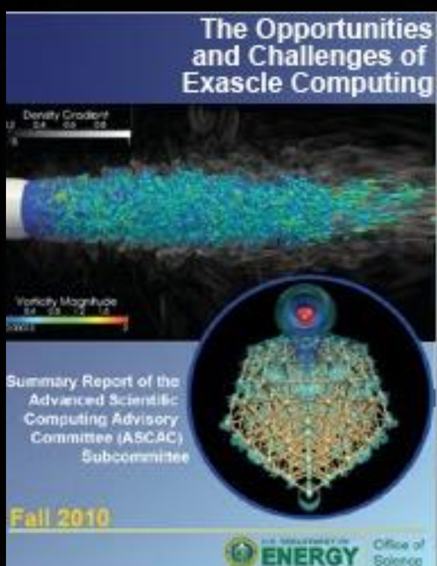
Exascale report from 2007 Town Halls
Entirely focused on modeling and simulation



HPC: Not just for simulation!

Exascale initiatives have been advancing the computational power of supercomputers for the last decade

- **NSCI (National Strategic Computing Initiative) announced by in the US in June 2015**
- **Exascale Computing Project ECP by DOE in the US**
- **Similar initiatives in Europe (EuroHPC), Japan, and China**



English.news.cn

Search | Advanced Sea

Services | Database | Markets | Weather | Site Index | RSS | Feedback

Global Edition | China | World | Business | Culture & Edu | Sports | Entertainment | Science & Tech

Most S

China planning new supercomputer

Source: Xinhua 2016-01-22 19:26:01

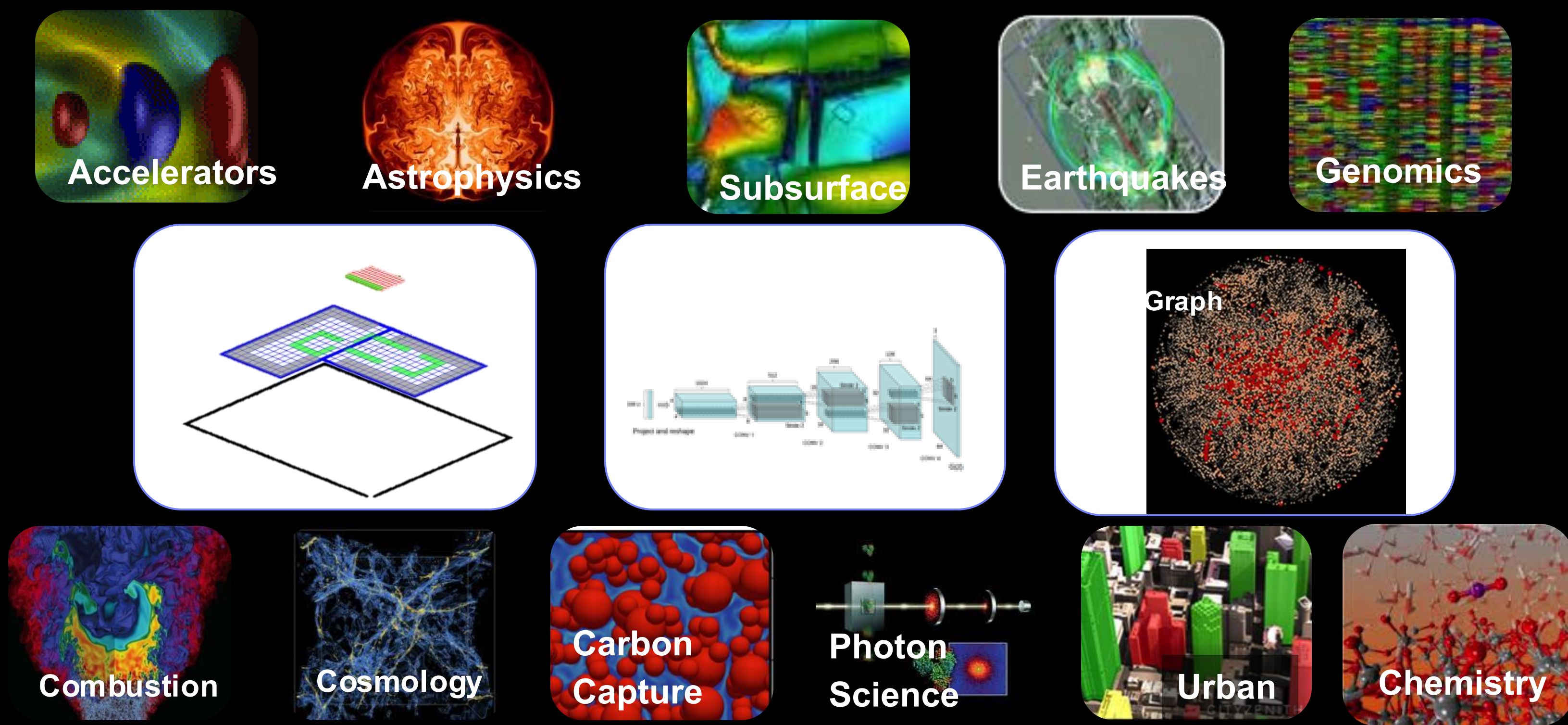
TIANJIN, Jan. 22 (Xinhua) -- China is planning a supercomputer 1,000 times more powerful than its groundbreaking Tianhe-1A as it faces rising demand for next-generation computing.

Meng Xiangfei, head of the applications department of the National Supercomputer Center, said on Friday that the center will release a prototype in 2017 or 2018 of an "exascale" computer -- one capable of at least a billion billion calculations per second

Exascale computing is considered the next frontier in the development of supercomputers.

Over 20 Scientific Applications under Exascale

Each with a science objective, new capability, mathematics or algorithms



Adapted from Kathy Yelick, UCB

The Exascale Market

Over 30 systems and over $12 billion in value

Exascale and Near-Exascale Leadership Systems (2020 to 2028)

Year Accepted	China	Europe	Japan	US	Other Countries*	Total Systems	Total Value
2020			1 near-exascale system ~$1.1B			1	$1.1B
2021	2 exascale ~$350M each	1 pre-exascale system ~$180M	--	1 pre-exascale system ~$200M	--	4	$1.1B
2022	1 exascale ~$350M	2 pre-exascale systems ~$390M total	--	1 exascale system ~$600M (2/3 accepted 2022)	--	4	$1.1B
2023	1 exascale system ~$350M	1 or 2 pre-exascale systems ~$150M each	1 near-exascale system ~$150M	Remaining 1/3 of Frontier system	--	4-5	~$1.0B
2024	1 exascale system ~$350M	2 or 3 pre-exascale systems ~$150M each	?	1 exascale system ~$600M	1 pre-exascale system ~$125M	5-6	~$1.3B
2025	1 or 2 exascale systems ~$300M each	2 exascale systems ~$350M each	?	2 exascale system ~$600M	1 near-exascale system ~$125M	6-9	$1.7B - $2.7B
2026	2 exascale systems ~$300M each	2 or 3 exascale systems ~$325M each	1 exascale system ~$200M	1 or 2 exascale systems ~$325M each	1 or 2 exascale systems ~$150M each	6-9	$1.7B - $2.5B
2027	2 exascale systems ~$275M each	2 or 3 exascale systems ~$300M	1 exascale system ~$150M	1 or 2 exascale systems ~$275M each	2 or 3 exascale systems ~$130M each	8-11	$1.8B - $2.5B
2028	2 exascale systems ~$250M each	2 or 3 exascale systems ~$275M	1 or 2 exascale systems ~$150M each	1 or 2 exascale systems ~$275M each	2 or 3 exascale systems ~$125M each	8-11	$1.7B - $2.6B
Total	**12-13**	**14-19**	**5-6**	**8-11**	**7-10**	**47-61**	**$12.5B - $15.9B**

* Includes S. Korea, Singapore, Australia, Russia, Canada, India, Israel, Saudi Arabia, etc.

Note: After 2023, many exascale systems will be 2-10 exascale.

Source: Hyperion Research, October 2024

From Hyperion Research, SC 2024

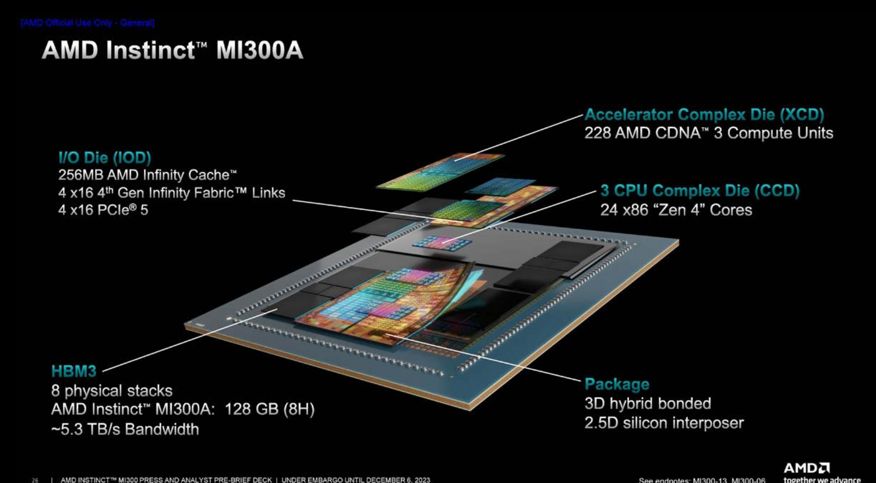
TOP500 Supercomputers

#	Site	Manufacturer	TOP10 Computer of the TOP500	Country	Cores	Rmax [Pflops]	Power [MW]
1	Lawrence Livermore National Laboratory	HPE	El Capitan HPE Cray EX255a, AMD EPYC 24C 1.8GHz, Instinct MI300A, Slingshot-11	USA	11,039,616	1,742	29.6
2	Oak Ridge National Laboratory	HPE	Frontier HPE Cray EX235a, AMD EPYC 64C 2.0GHz, Instinct MI250X, Slingshot-11	USA	9,066,176	1,353	24.6
3	Argonne National Laboratory	Intel	Aurora HPE Cray EX/Intel Exascale Compute Blade, Xeon Max 9470, Data Center GPU Max, Slingshot-11	USA	4,742,808	1,012	38.7
4	Microsoft Azure	Microsoft	Eagle Microsoft NDv5, Xeon Platinum 8480C, NVIDIA H100, Infiniband NDR	USA	1,123,200	561.2	
5	Eni S.p.A. Center for Computational Science	HPE	HPC6 HPE Cray EX235a, AMD EPYC 64C 2.0GHz, Instinct MI250X, Slingshot-11	Italy	3,143,520	477.9	8.5
6	RIKEN Center for Computational Science	Fujitsu	Fugaku Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D	Japan	7,630,848	442.0	29.9
7	Swiss National Supercomputing Centre (CSCS)	HPE	Alps HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, GH200, Slingshot-11	Switzerland	2,121,600	434.9	7.1
8	EuroHPC / CSC	HPE	LUMI HPE Cray EX235a, AMD EPYC 64C 2.0GHz, Instinct MI250X, Slingshot-11	Finland	2,752,704	379.7	7.1
9	EuroHPC / CINECA	EVIDEN	Leonardo Atos BullSequana XH2000, Xeon 32C 2.6GHz, NVIDIA A100, HDR Infiniband	Italy	1,824,768	241.2	7.5
10	Lawrence Livermore National Laboratory	HPE	Tuolumne HPE Cray EX255a, AMD EPYC 24C 1.8GHz, Instinct MI300A, Slingshot-11	USA	1,161,216	208.1	3.4

ADIA Lab

El Capitan @ LLNL

- **System specifications:**
  - Peak 2.793 DP exaflop/s; HPL 1.742 exaflop/s
    - 11,136 HPE Cray EX255a Nodes with 4 MI300A APU, and 512 GiB HBM3 each
    - Per node "DGEMM peak" of 250.8 DP teraflops (achievable)
  - Peak electrical power 34.8 MW; during HPL 28.9 MW
  - Slingshot interconnect
- **MI300A - world's first data center APU**
  - 3D chiplet design with:
  - 3 AMD CDNA 3 GPU dies
  - 3 8-core "Zen 4" CPU dies
  - cache memory, 8*16GiB HBM3
- **Near node local storage: the "Rabbits"**



HPC is an exponential technology

Exponential technology:

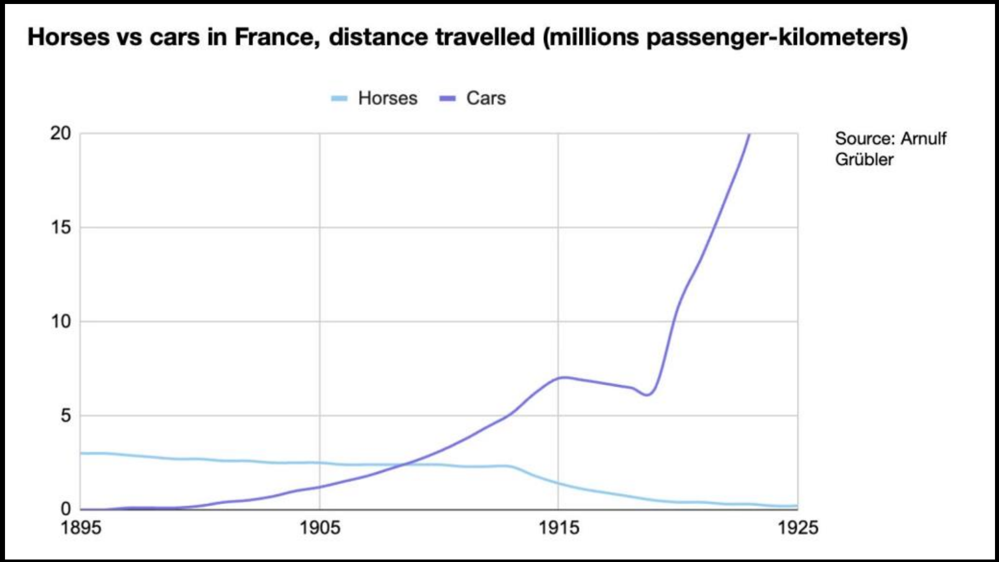
- innovations that experience rapid, accelerating growth, often characterized by **doubling in performance or capacity**
- simultaneously **reducing costs**,
- **transformative impacts** on industries and society.

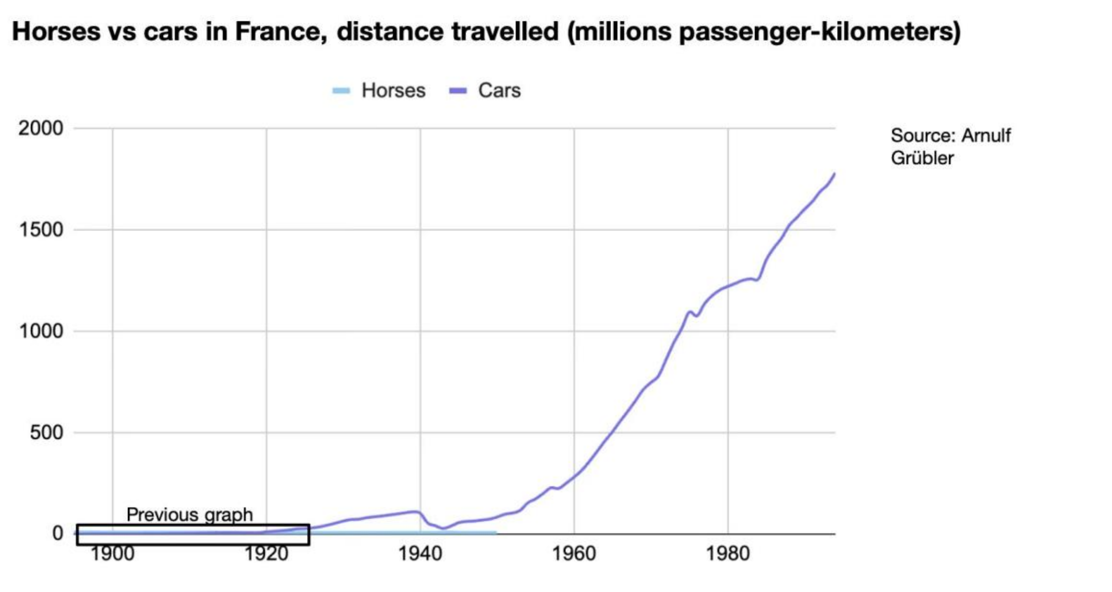
Exponential technology acceleration:

1. **Learning Rates and Cost Reduction**
2. **Network Effects**
3. **Data and Digital Feedback Loops**
4. **Cross-Disciplinary Innovation**
5. **Global Connectivity and Capital**
6. **Adoption Curves and Early Success**

These factors create a virtuous cycle where each improvement accelerates further progress, leading to transformative change at an increasingly rapid pace.

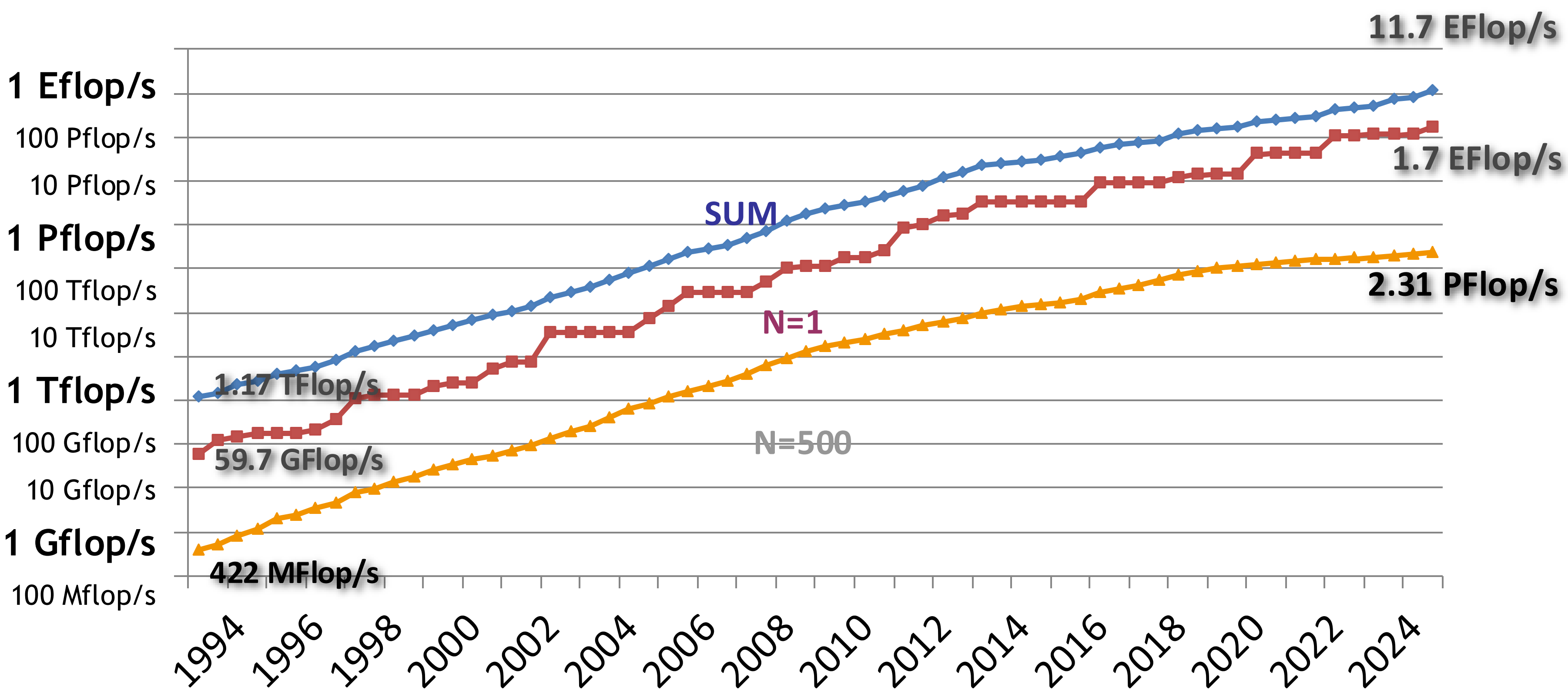
Example: introdcution of cars about 100 years ago



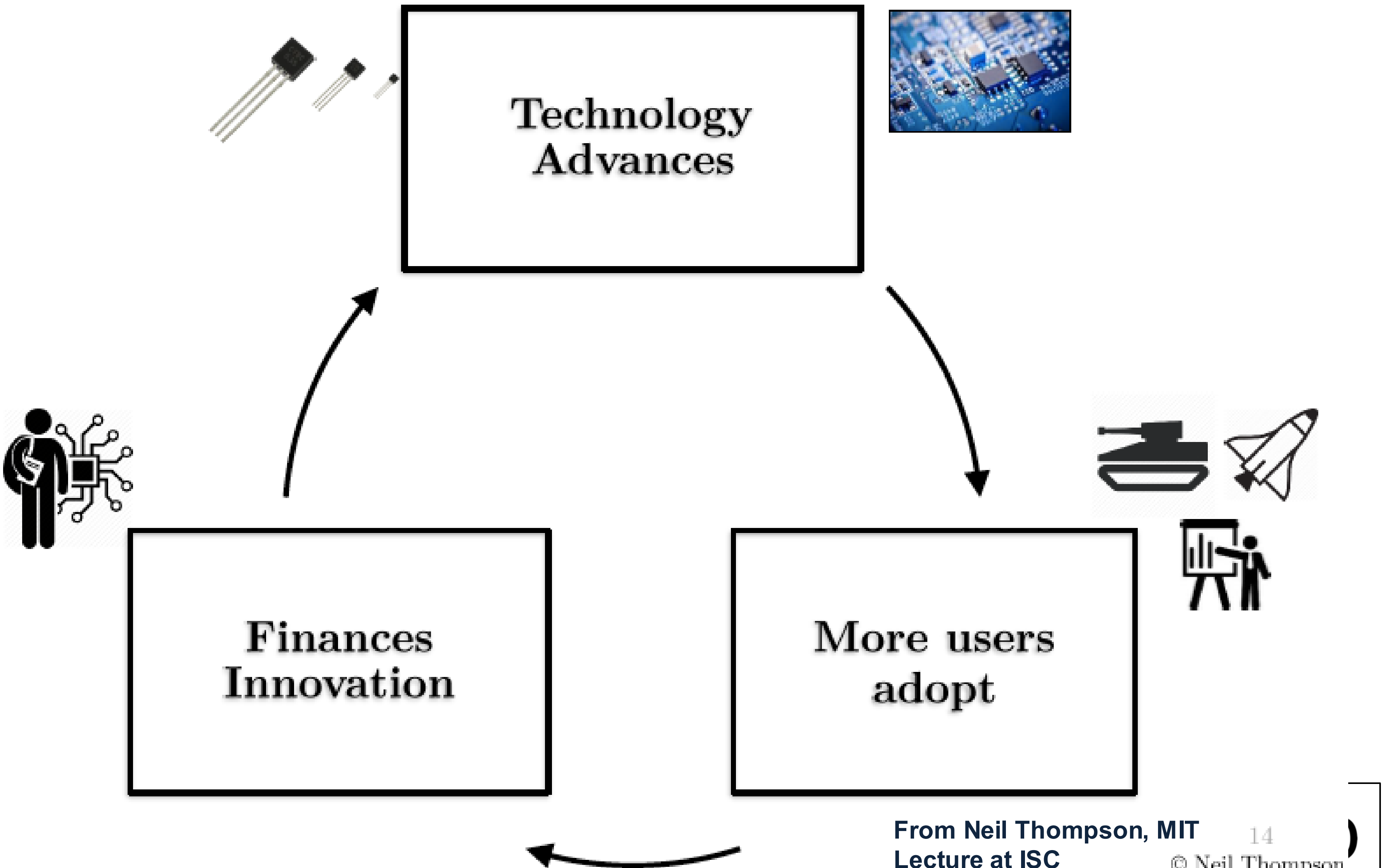


watch Tony Seba on YouTube about EVs

PERFORMANCE DEVELOPMENT

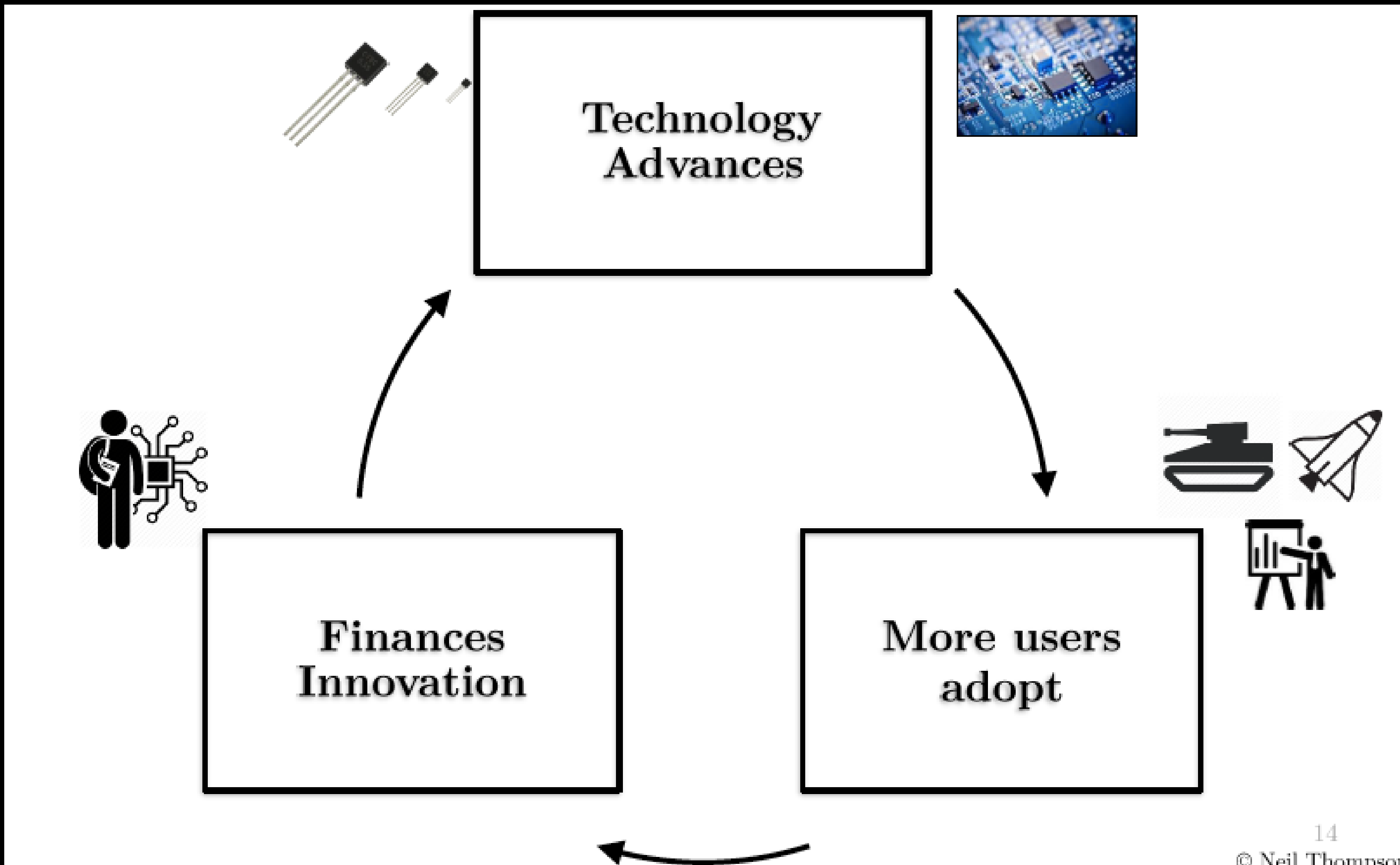


Virtuous Cycle of a General Purpose Technology (semiconductors)

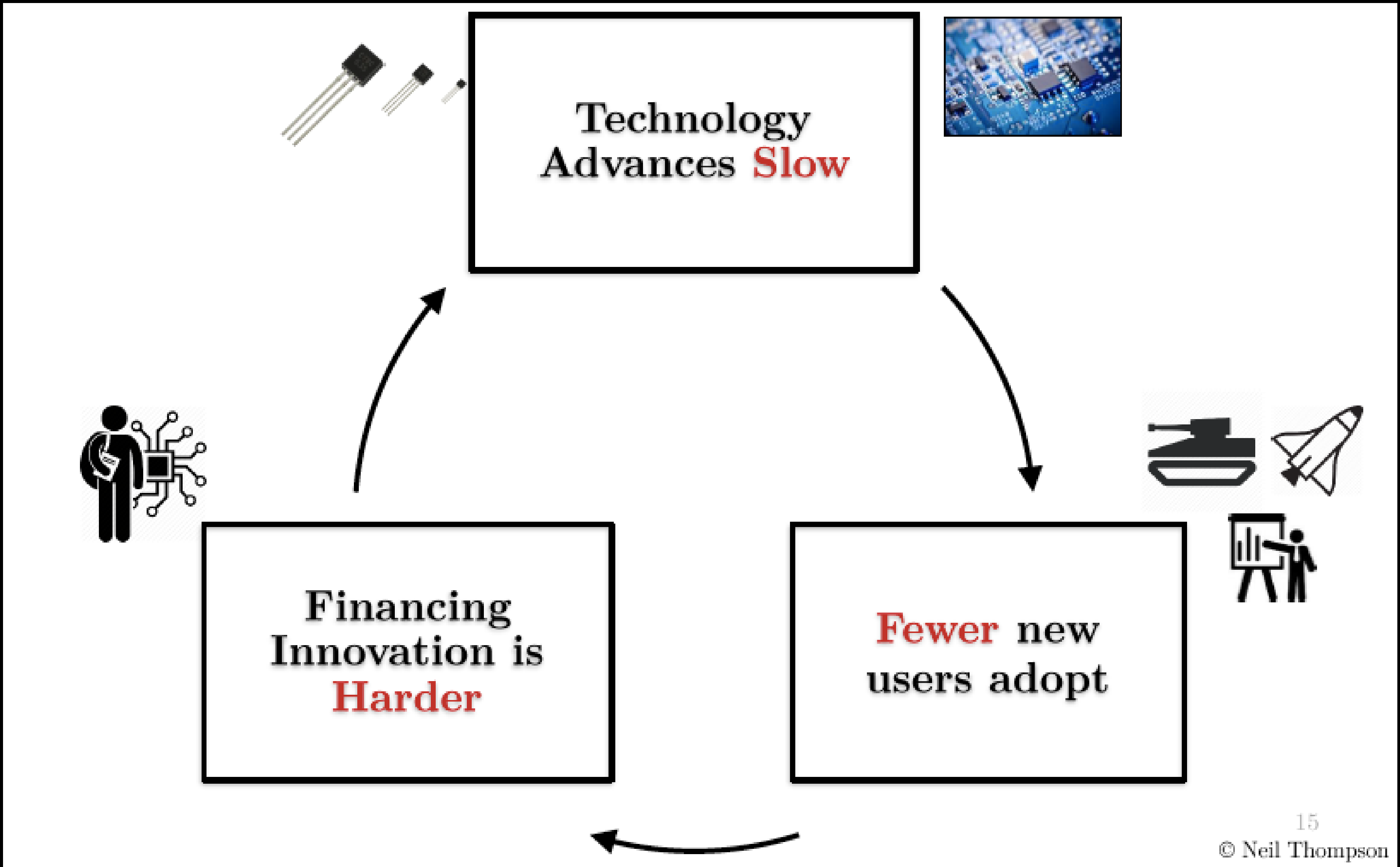


From Neil Thompson, MIT
Lecture at ISC

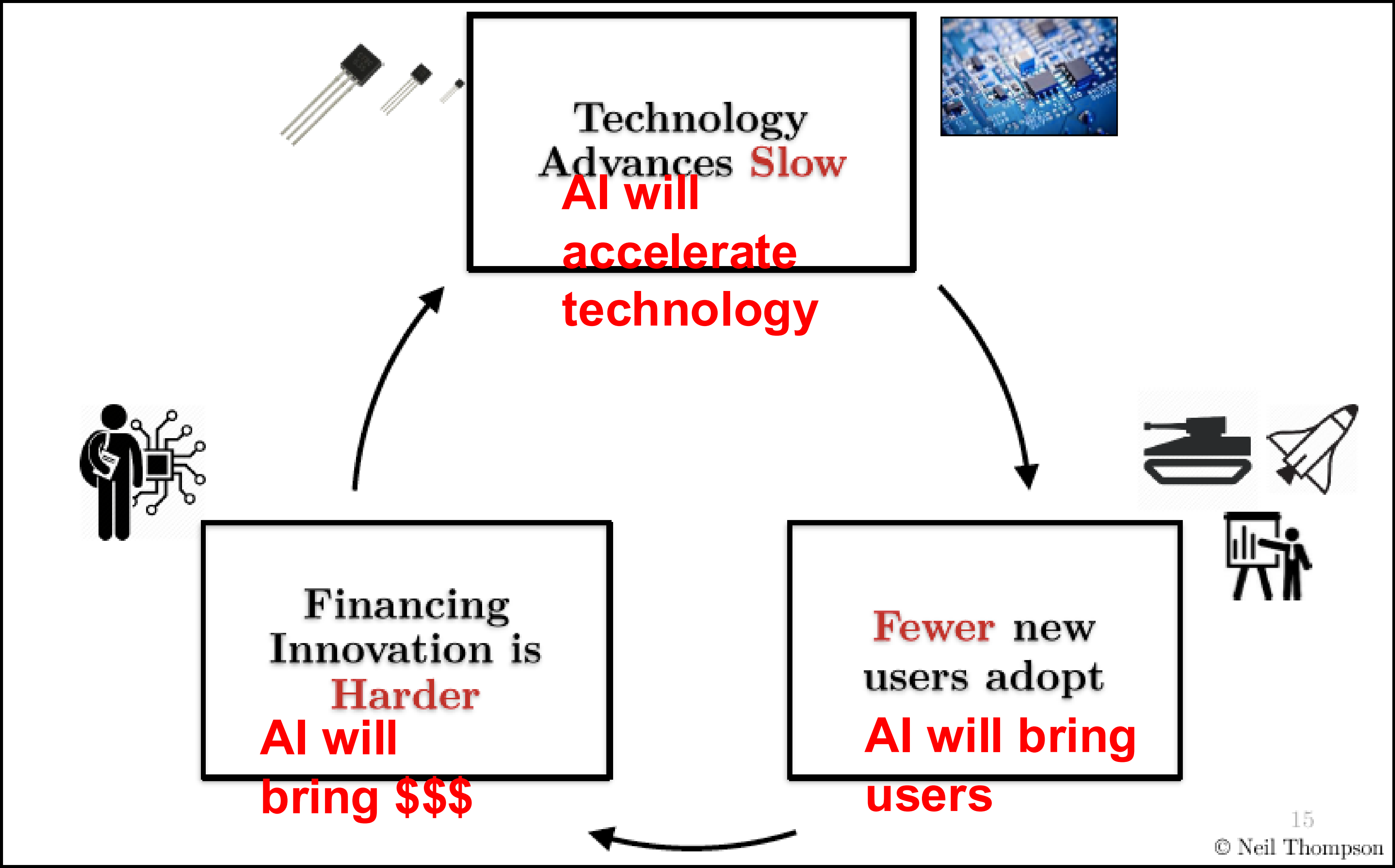
Virtuous Cycle of a General Purpose Technology (semiconductors)



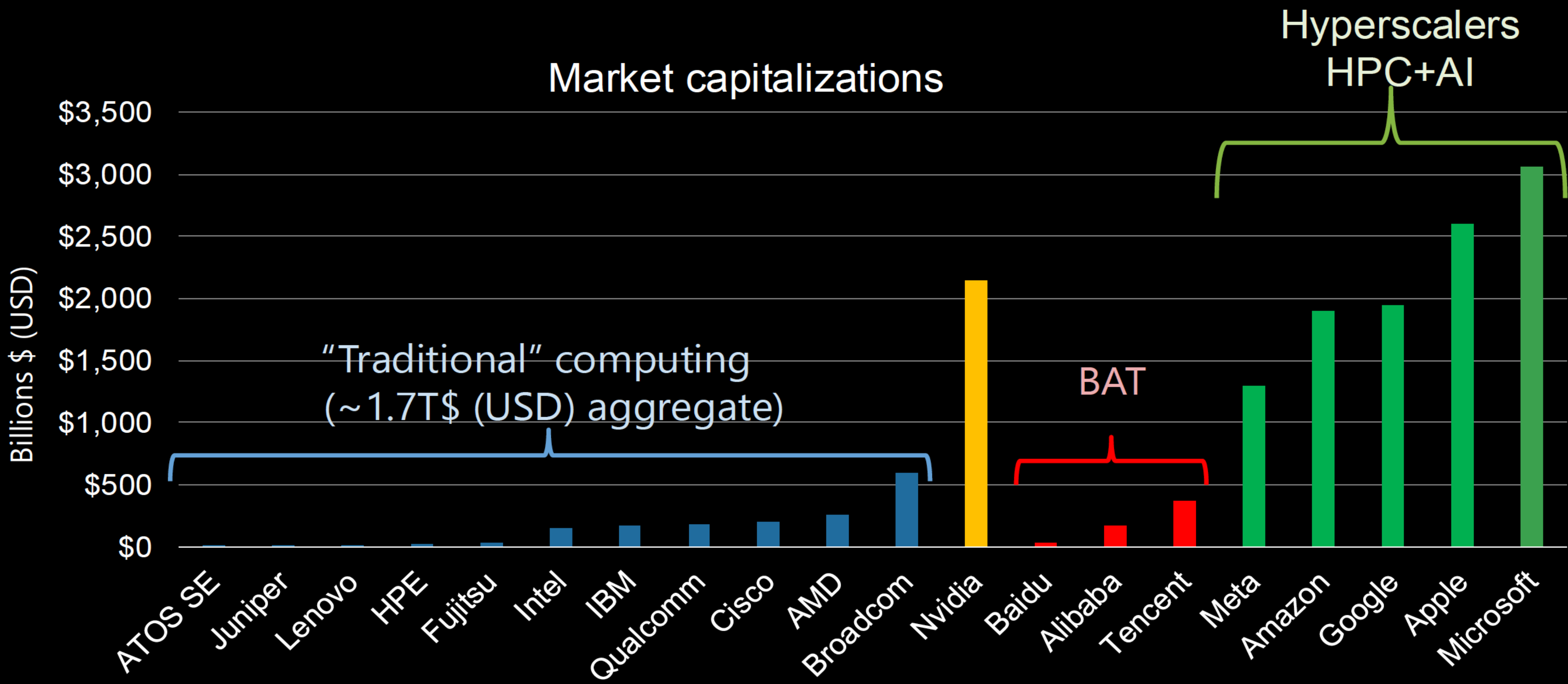
Fragmenting a General Purpose Technology



AI will accelerate HPC

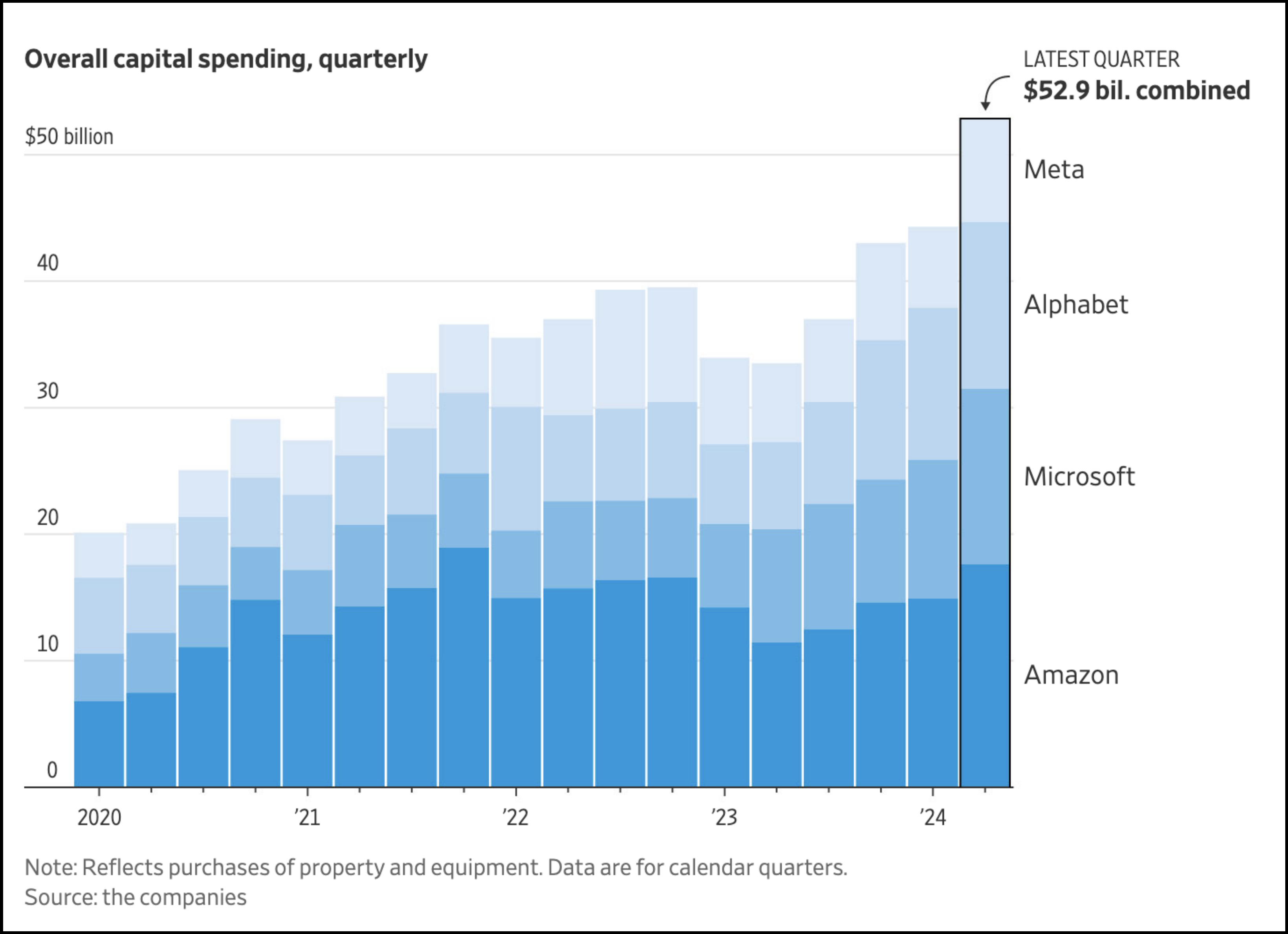


Follow the money, understand the implications



Source: Reed, Gannon, Dongarra

from Kathy Yelick, keynote at ISC 2024



Overall AI capital spending, source Wall Street Journal

HDS: there is only one choice, join them!

Technology and Marketplace: Radically Different than 2008!

What's a post-Exascale strategy for the science community?

Beat them
- Design processors for science
  More Co-Design and don't forget the math and software

Join them
- Leverage AI Hardware
  for (1) AI in Science and (2) Traditional science computations?

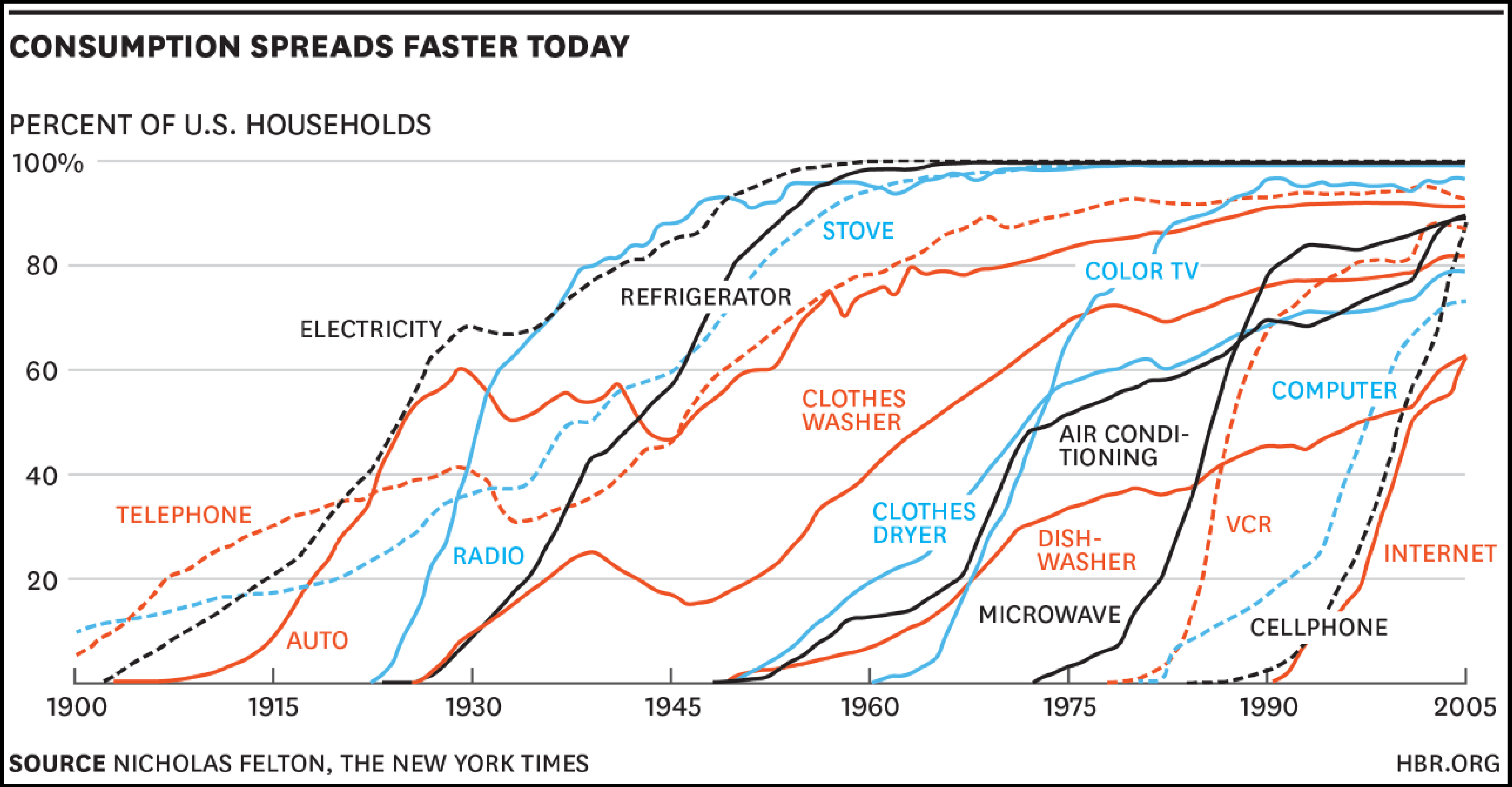
from Kathy Yelick, keynote at ISC 2024

Remember FIFA Football WorldCup 2006?

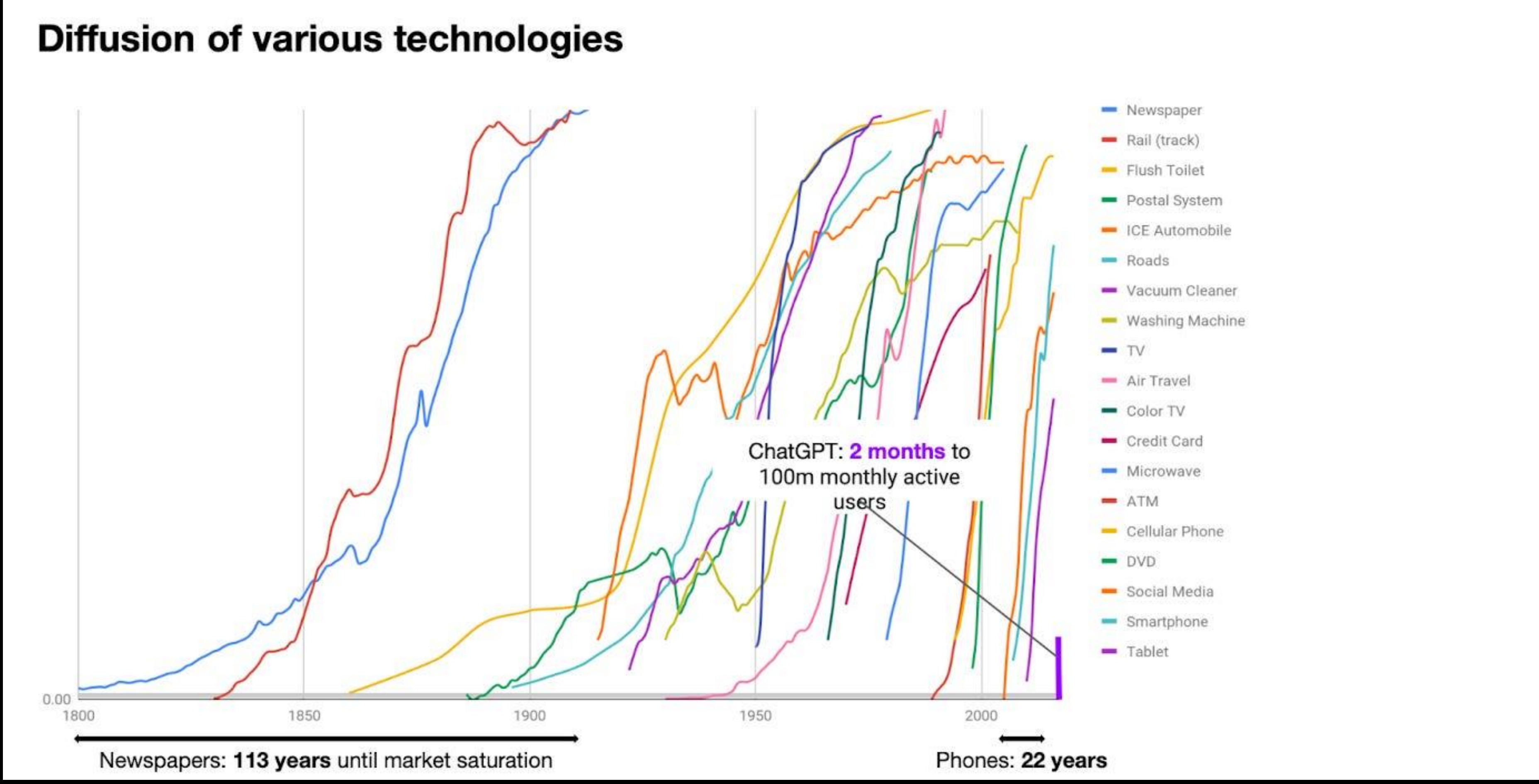
Do you know the most popular app used during WorldCup 2006?

Idea from Raša Karapandža, NYUAD

History of Technology Adoption

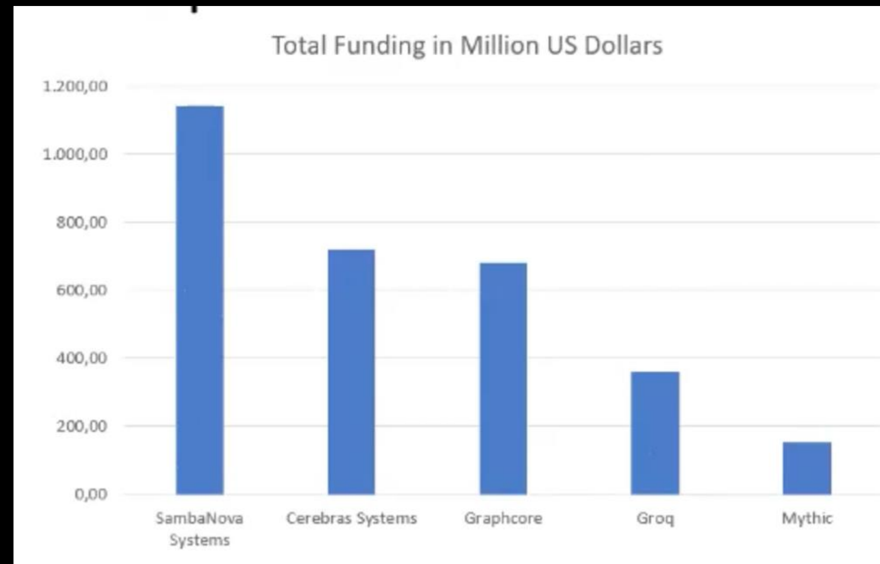


ChatGPT: 100m monthly users within 2 months



Everyone is Making AI Chips

NVIDIA	Traditional chip makers
AMD	
Intel	
IBM	
	"Hyperscaler"
Facebook + Intel	
Amazon (Echo, Oculus)	
Google (TPU, Pixel)	
Apple (SoCs)	
Microsoft ("AI chip")	



Graphcore, Nervana, Wave Computing, Horizon Robotics, Cambricon, DeePhi, Esperanto, SambaNova, Eyeriss, Tenstorrent, Mythic, ThnokForce, Groq, Lightmatter

Not everyone is selling those chips!

After Jack Dongarra, presentation at ADIA Lab Symposium

The energy challenge of AI

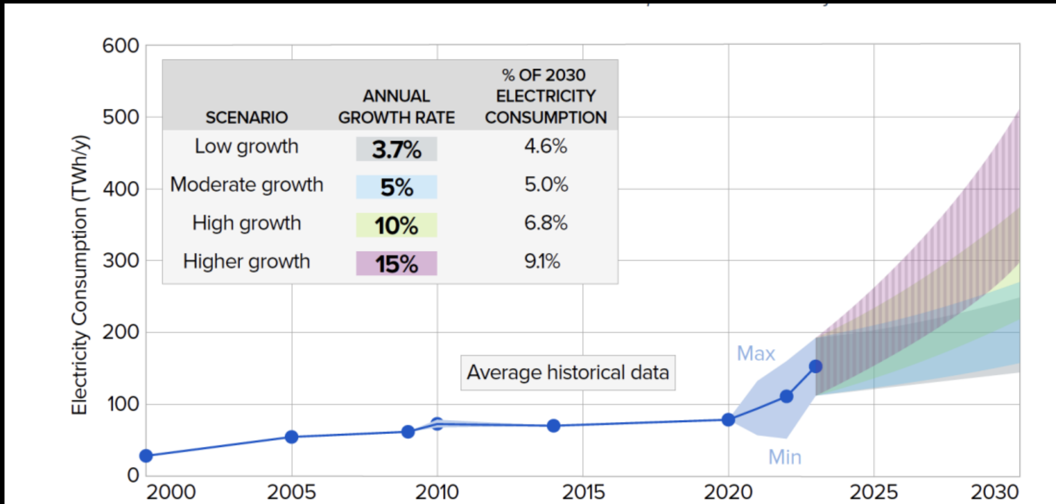


Figure ES-1. *Projections of potential electricity consumption by U.S. data centers: 2023–2030 . % of 2030 electricity consumption projections assume that all other (non-data center) load increases at 1% annually.*

EPRI White Paper, "Powering Intelligence", 2024

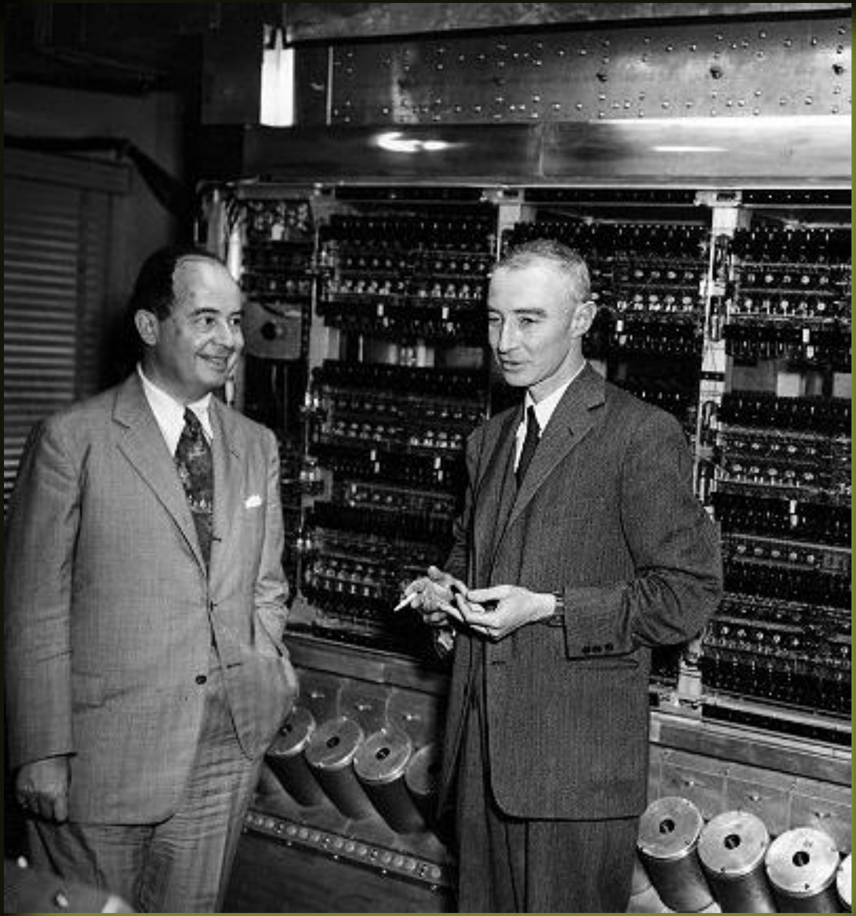
Microsoft has signed a 20-year deal to exclusively access 835 megawatts of energy from a nuclear plant.

Elon Musk's AI data center in Memphis Tennessee will need upwards of 70 megawatts of power to run all 100,000 GPUs 14 massive mobile generators to power the facility as he works out power supply agreements with local utilities.

2.5 MW per truck X 14 trucks = 35 MW

They installed the 100,000 GPUs in 122 days and have ordered another 100,000 GPUs.

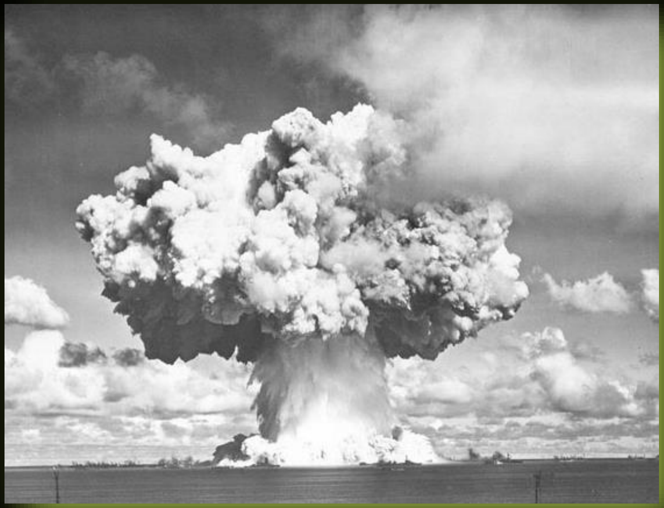
Historical Perspective, or “Why We Are Here Today ….”

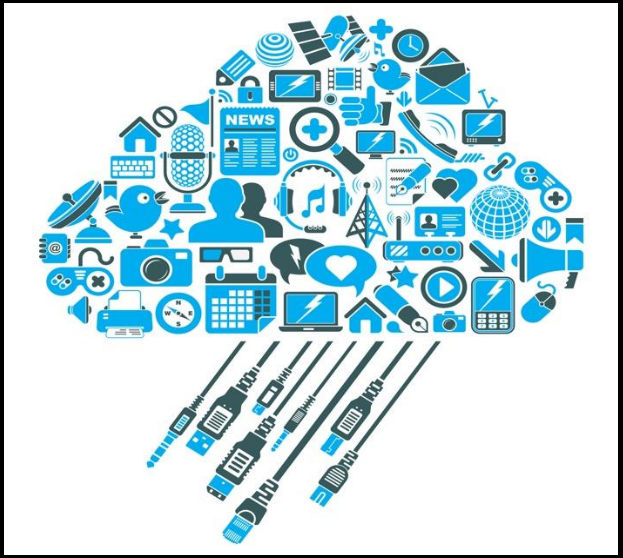


John von Neumann and Robert Oppenheimer, Princeton, IAS, 1952

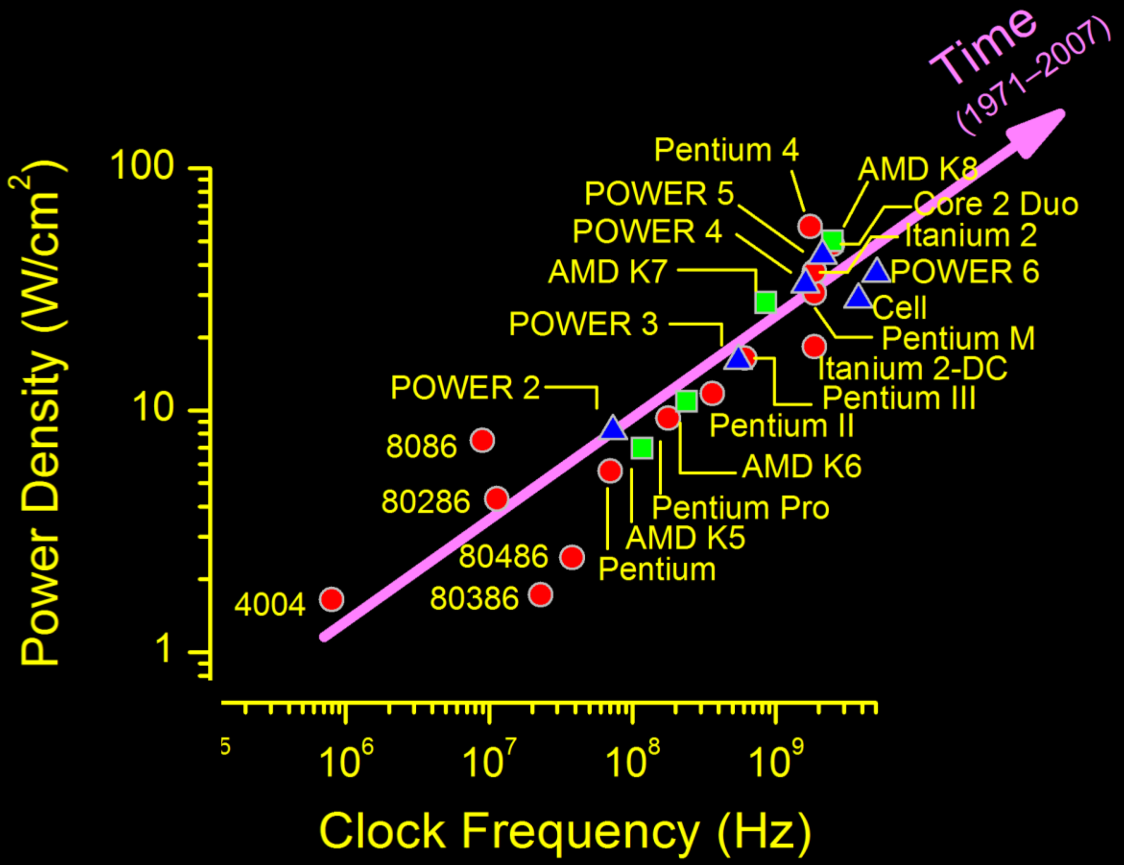
- 70 years of large scale computational physics applications driving computer development
- Remarkable longevity
- Architecture matched to application: Von Neumann architecture and focus on floating point performance

From the Bomb to the Cloud in Sixty Years

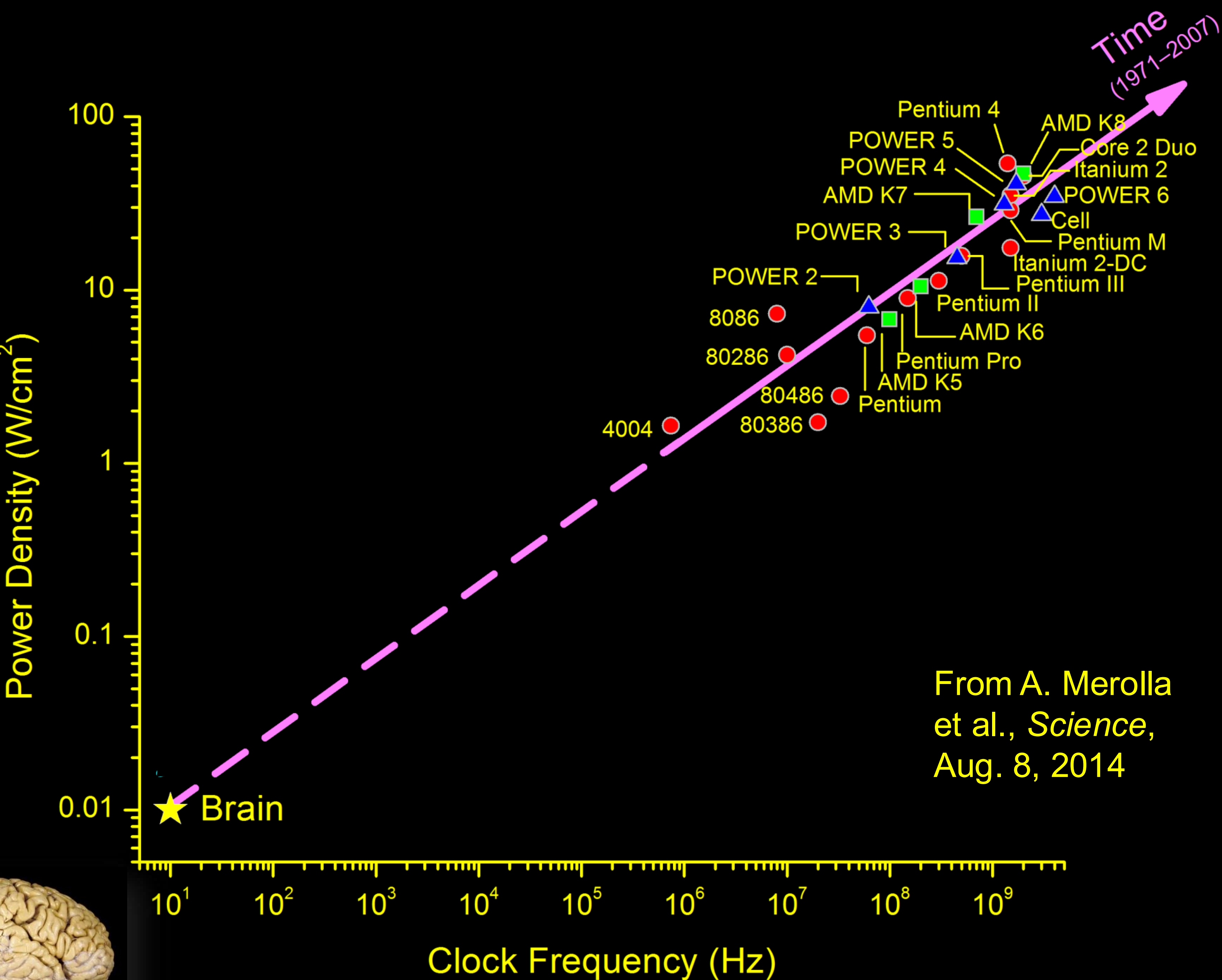




- Von Neumann architecture is ideally suited for large-scale, floating point intense, 2D or 3D grid-based, computational physics applications
- Today we are using the same basic architecture for social networking, Web searches, music, photography, and AI etc.



From A. Merolla et al., *Science*, Aug. 8, 2014



From A. Merolla et al., *Science*, Aug. 8, 2014

In Summary

- HPC will continue to advance in performance, energy efficiencey, AI integration, and hybridization with quantum and edge computing
- AI will continue to grow as exponential technology, leveraging HPC knowledge and investments
- There will be major technology challenges to future growth of AI (architecture, algorithms, and power use) in addition to the poltitical, economic, and social challenges.