Beyond Exascale Computing

Kathy Yelick

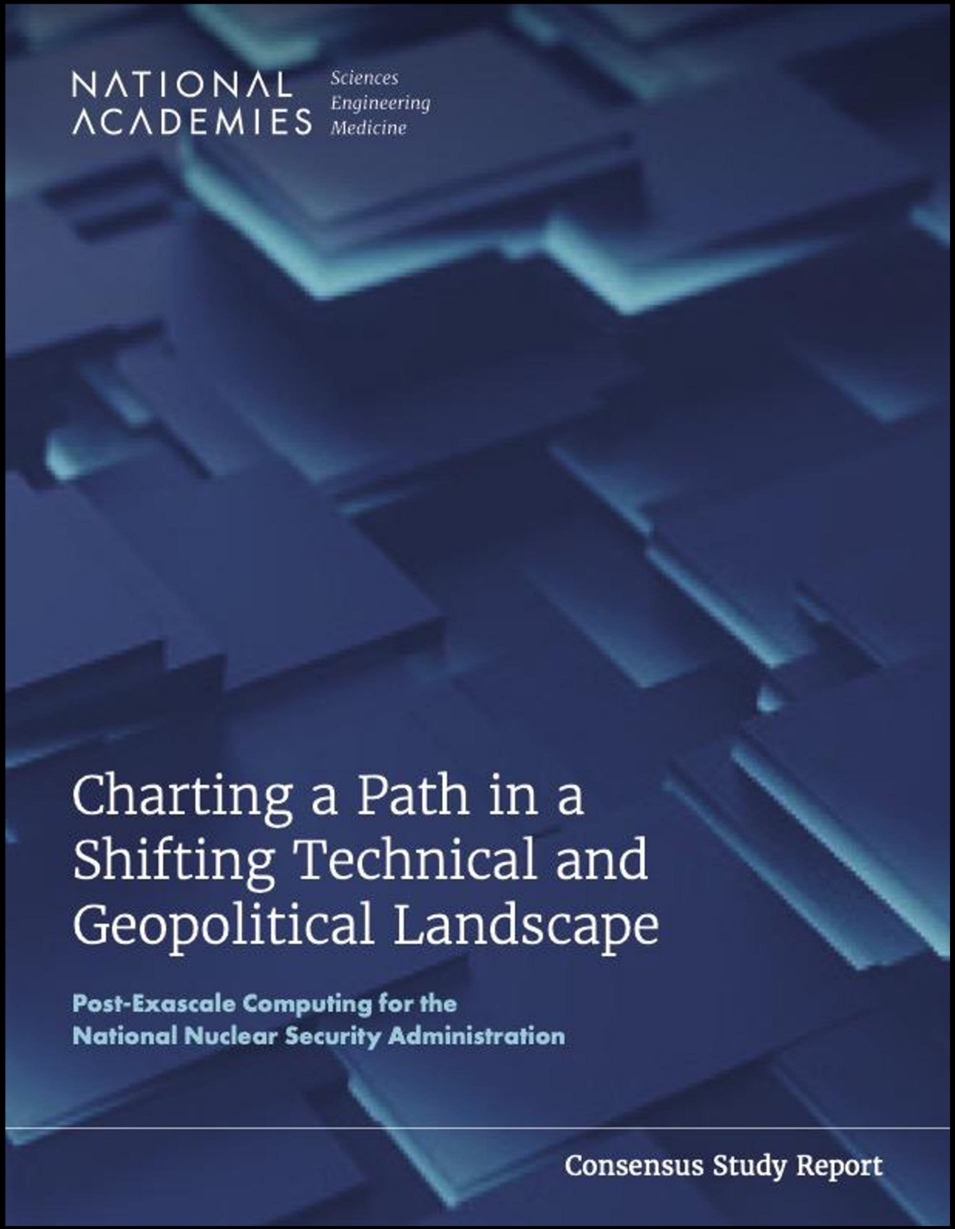
Vice Chancellor for Research
and Robert S. Pepper Professor of
Electrical Engineering and Computer Sciences
U.C. Berkeley

Senior Faculty Scientist
Computing Sciences
Lawrence Berkeley National Laboratory

National Academies Study

Kathy Yelick (Chair)
John Bell
Bill Carlson
Fred Chong
Dona Crawford
Jack Dongarra
Mark Dean
Ian Foster
Charlie McMillan
Dan Meiron
Daniel Reed
Karen Willcox

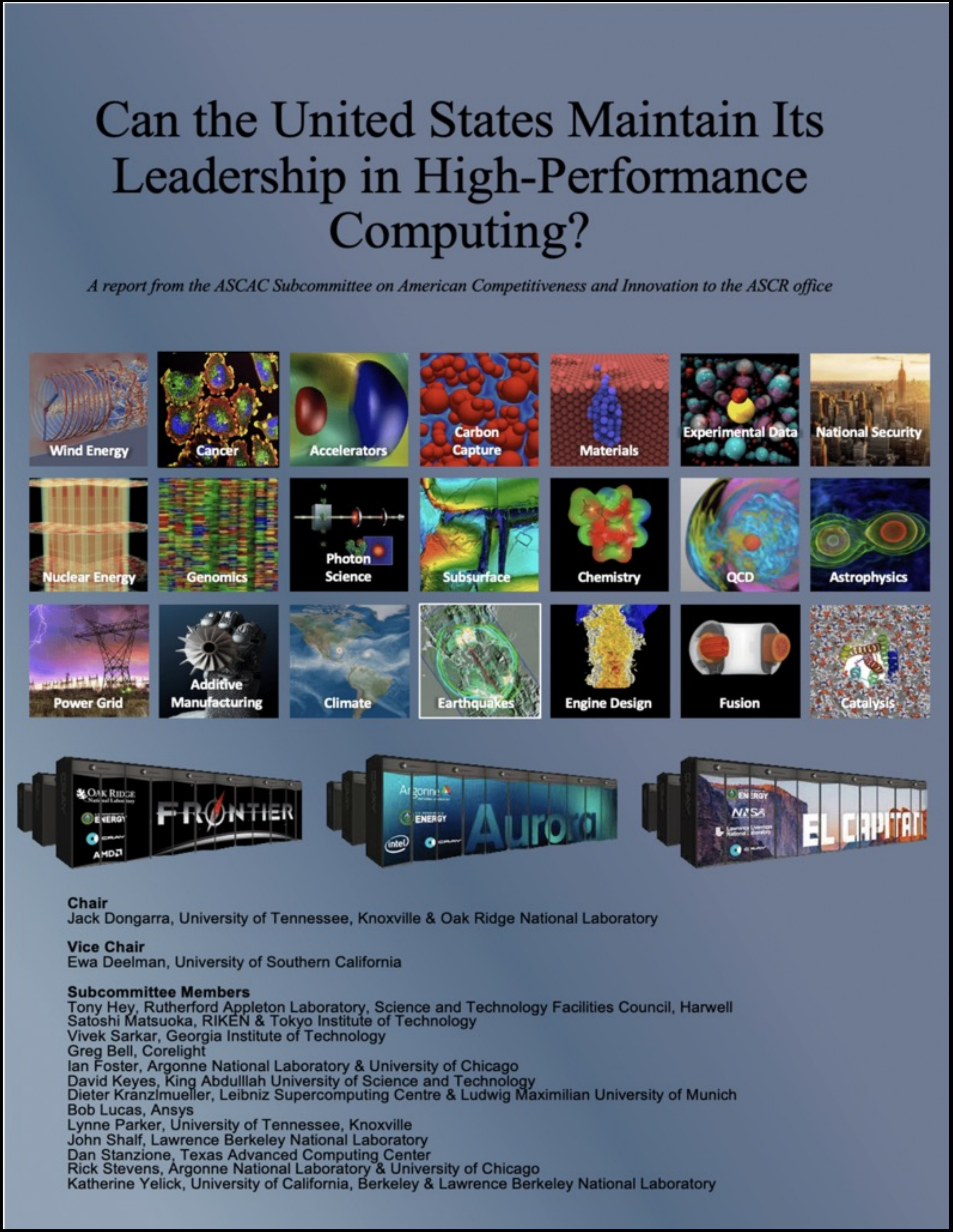
(report at www.cstb.org)



DOE ASCR Report

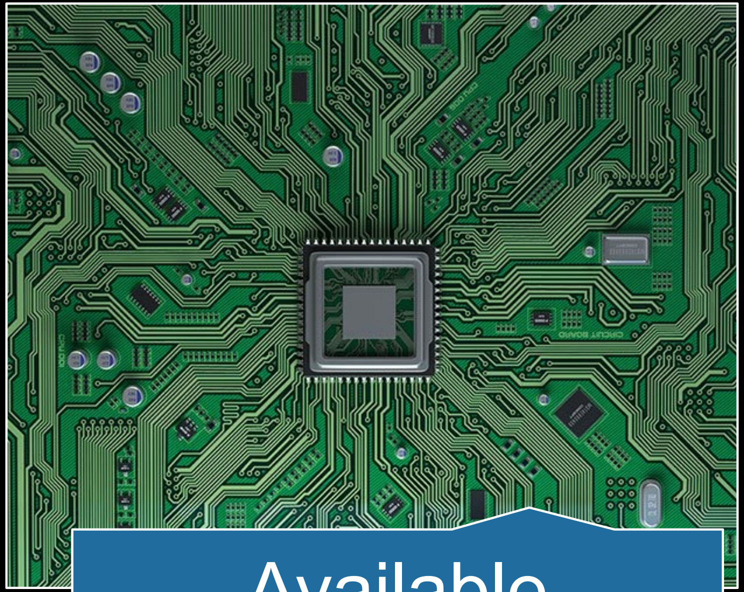
Jack Dongarra (Chair)
Ewa Deelman (Vice Chair)
Tony Hey
Satoshi Matsuoka
Vivek Sarkar
Greg Bell
Ian Foster
David Keyes
Dieter Kranzlmeuller
Bob Lucas
Lynne Parker
John Shalf
Dan Satanzione
Rick Stevens
Katherine Yelick

(report at www.osti.gov/)

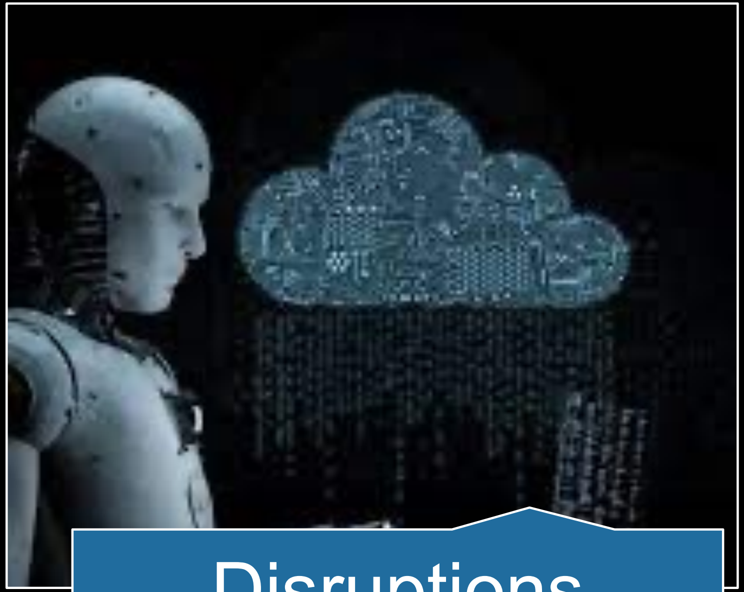


Post-Exascale Computing

Computing demand

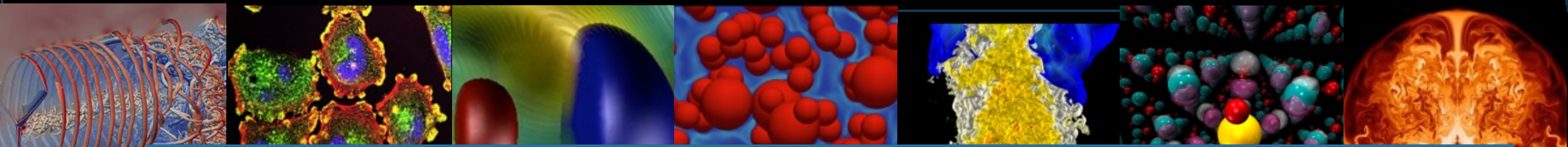


Available technology

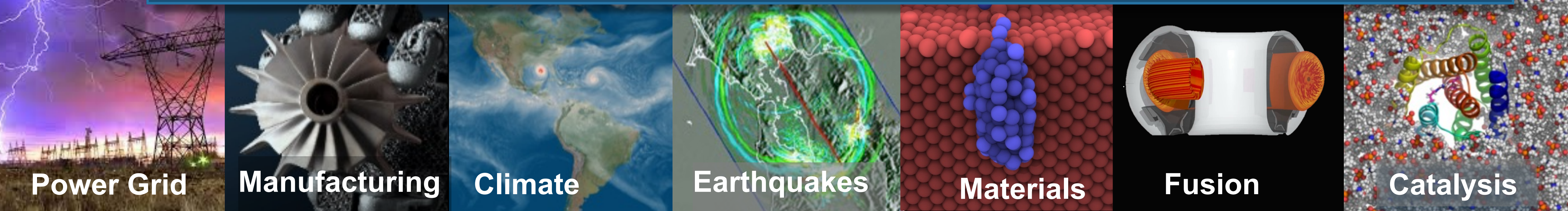


Disruptions

Continue to Rethink Applications



- 24 projects with about 10 people per team
- Rely heavily on hardware features and software teams
- Several new to HPC, all with new capabilities
- We should have another 2 dozen in 10 years!!

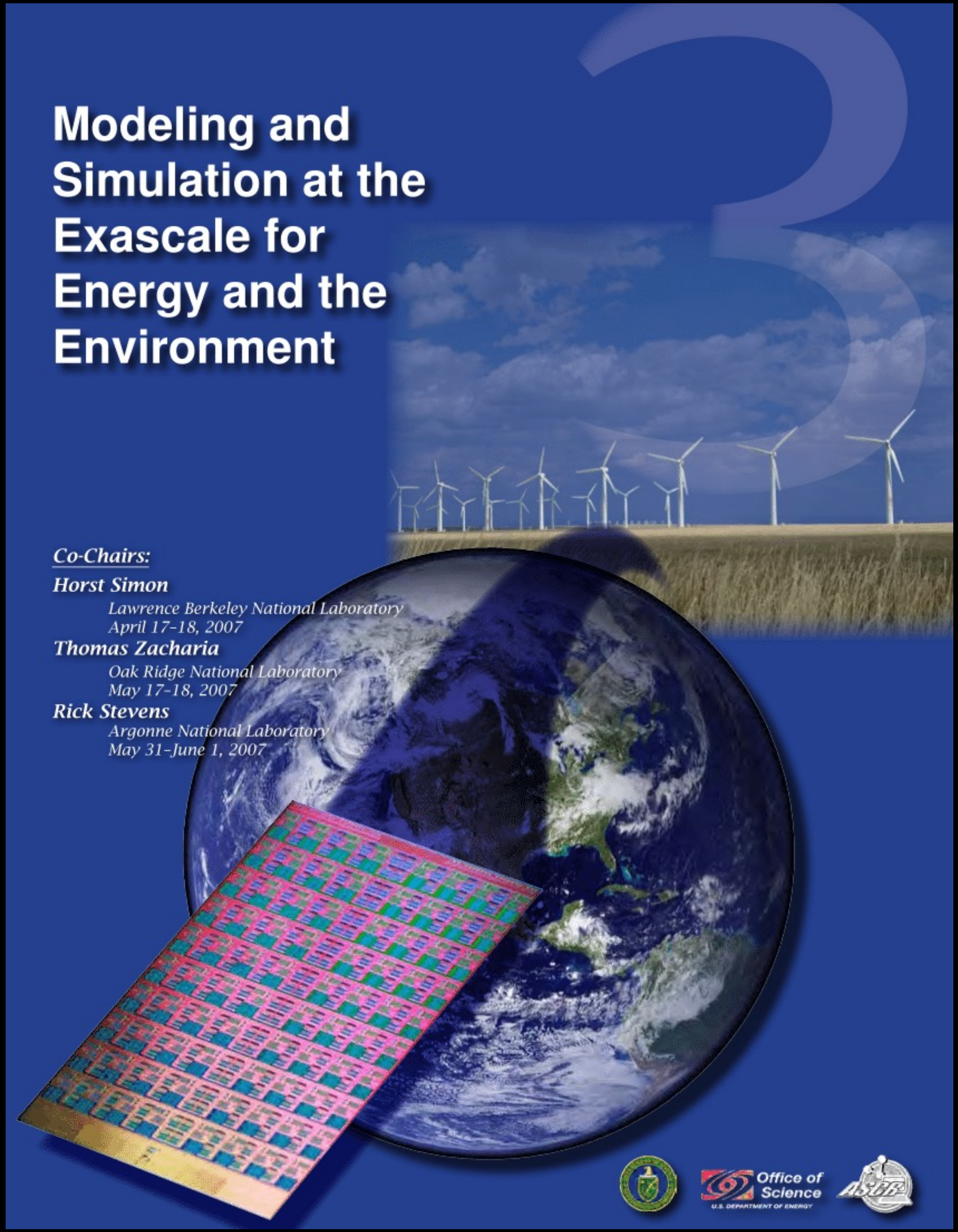


Scientific Computing Circa 2007

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

~~Scientific Computing is often used synonymously with Simulation and HPC~~

Simulation ⊂ Scientific Computing ⊂ HPC



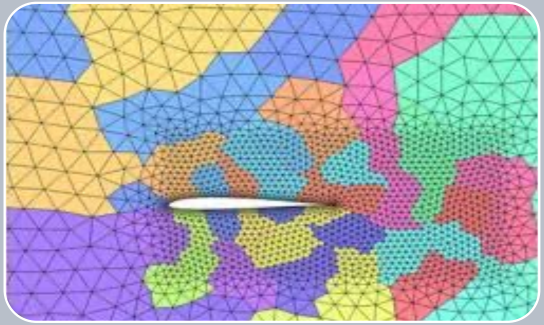
Runtime of “hero” calculations are too long

Number of Nodes	Memory Footprint	Wall-Clock Time
2400	~300–400 TB	6 months
4990	~600 TB	3–4 months
288	~20 TB	1 month
3250	104 TB	5.8 days
512	32.8 TB	2 months

Iterative design does not happen on 6 month cycles

Subset of Application Challenges Beyond Exascale

Reentry

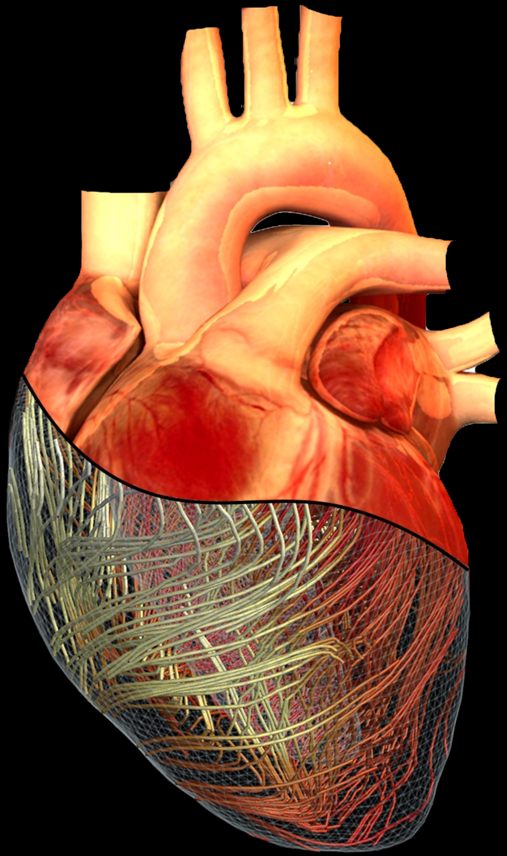


Complex geometry

Combustion

Extreme Environs

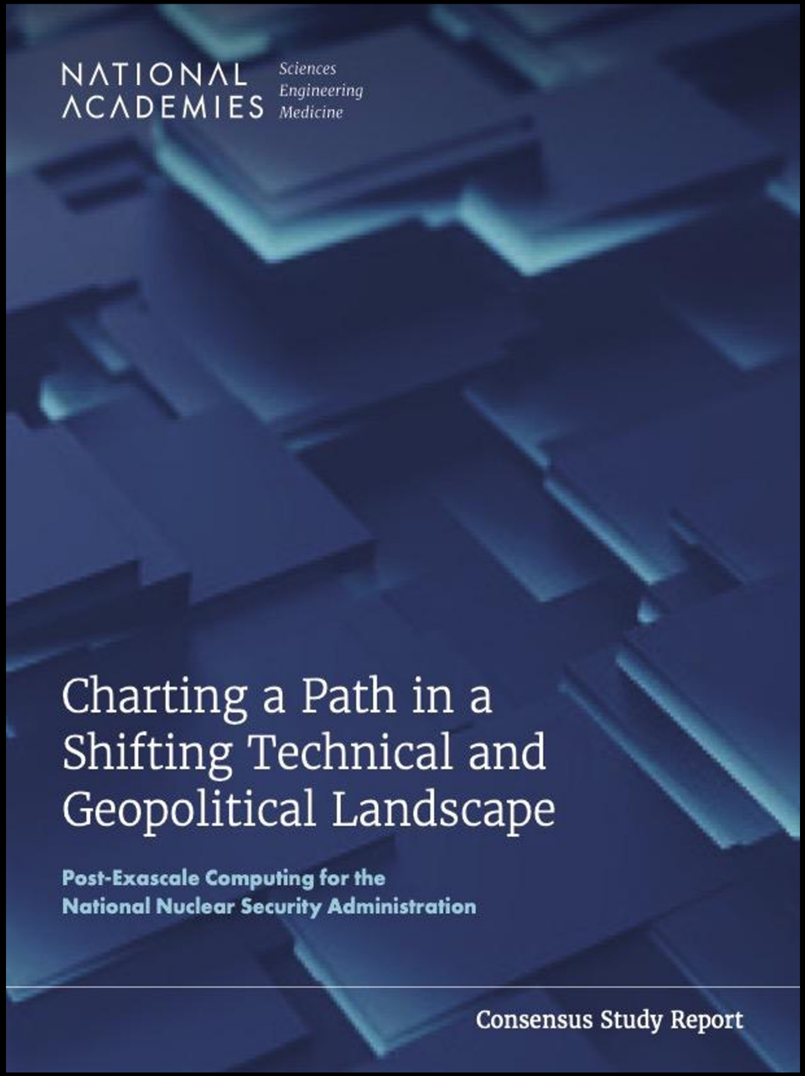
Digital Twins



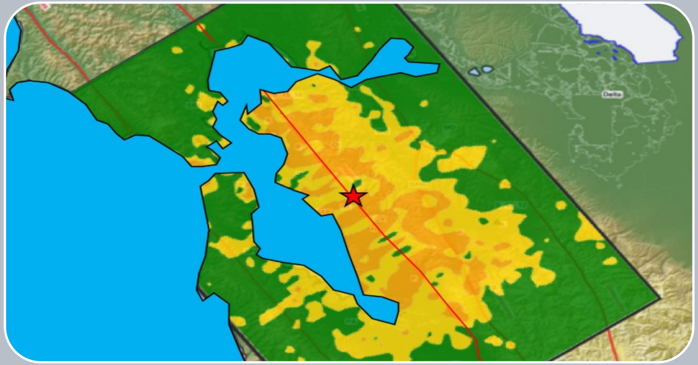
- Simulations
- Sensors / data
- Multi-level
- Real-time

National Academies Study

Finding: The demands for advanced computing continue to grow and will exceed the capabilities of planned upgrades across the NNSA labs.



New demands for HPC in Science



Simulation

From atoms to the universe

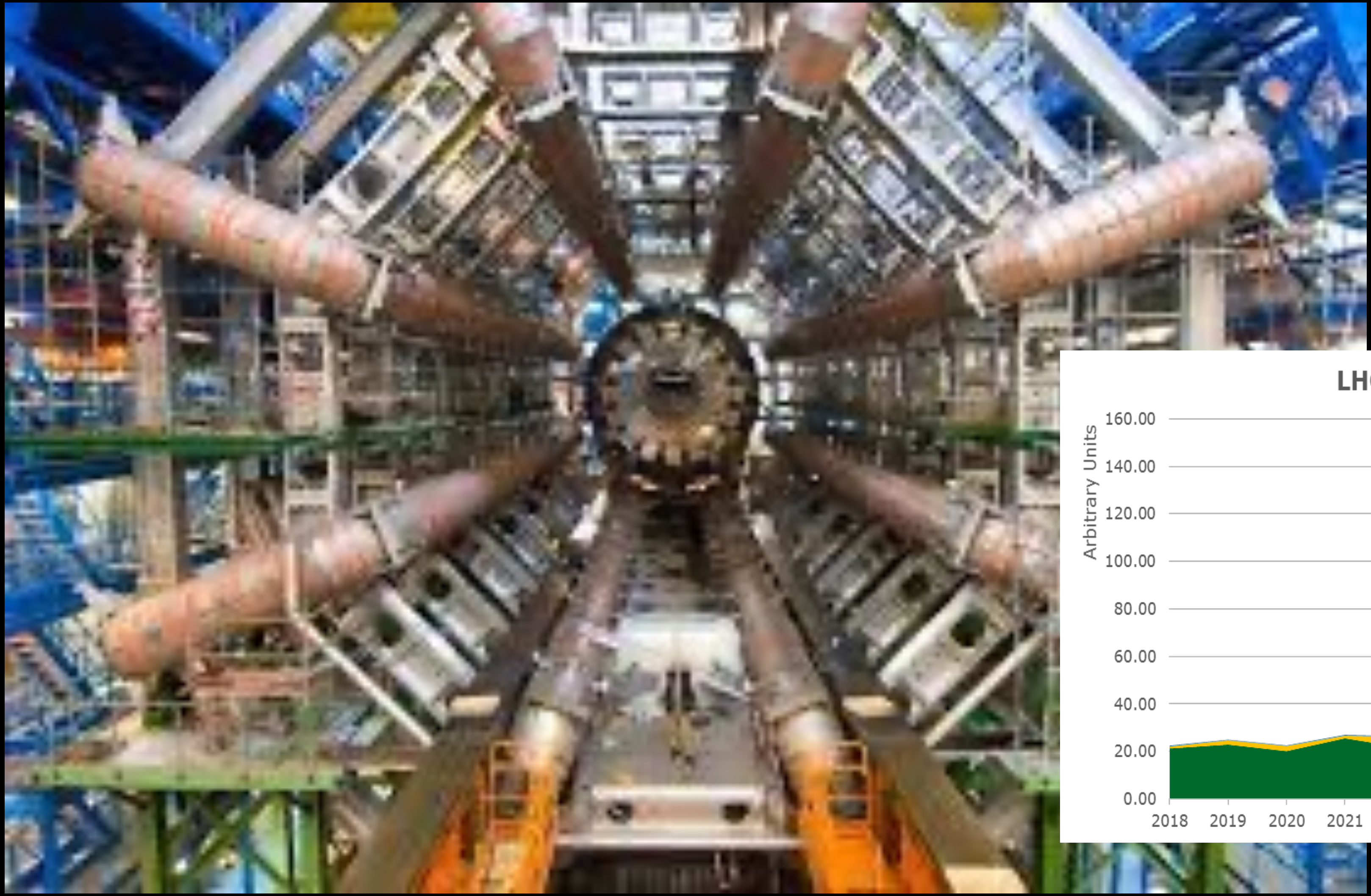
Data

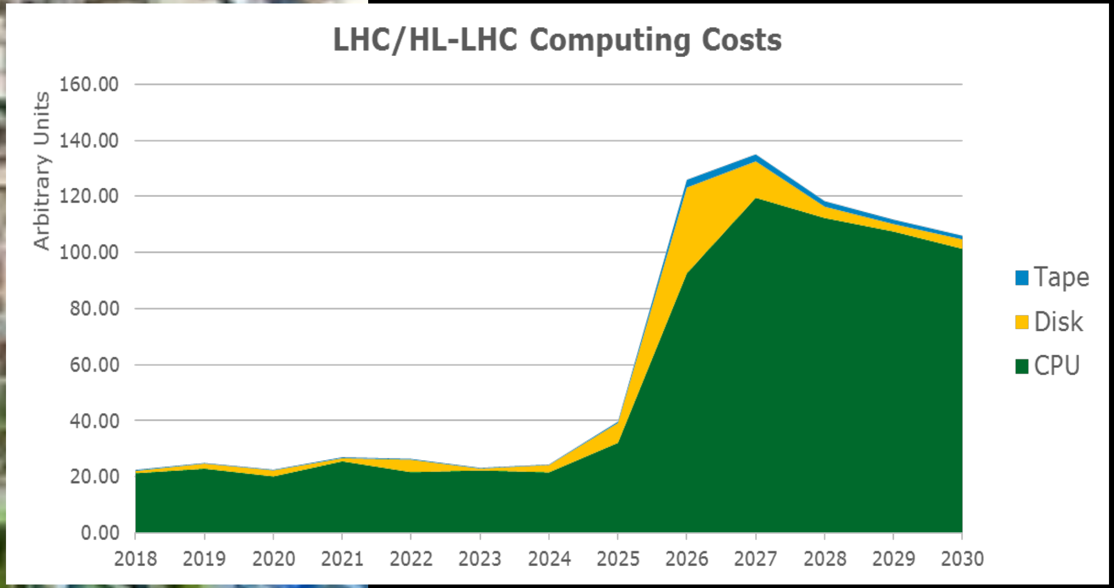
Images, text, to genomes

Learning

Interpret, infer and automate

Prediction of Atlas computing +$1B





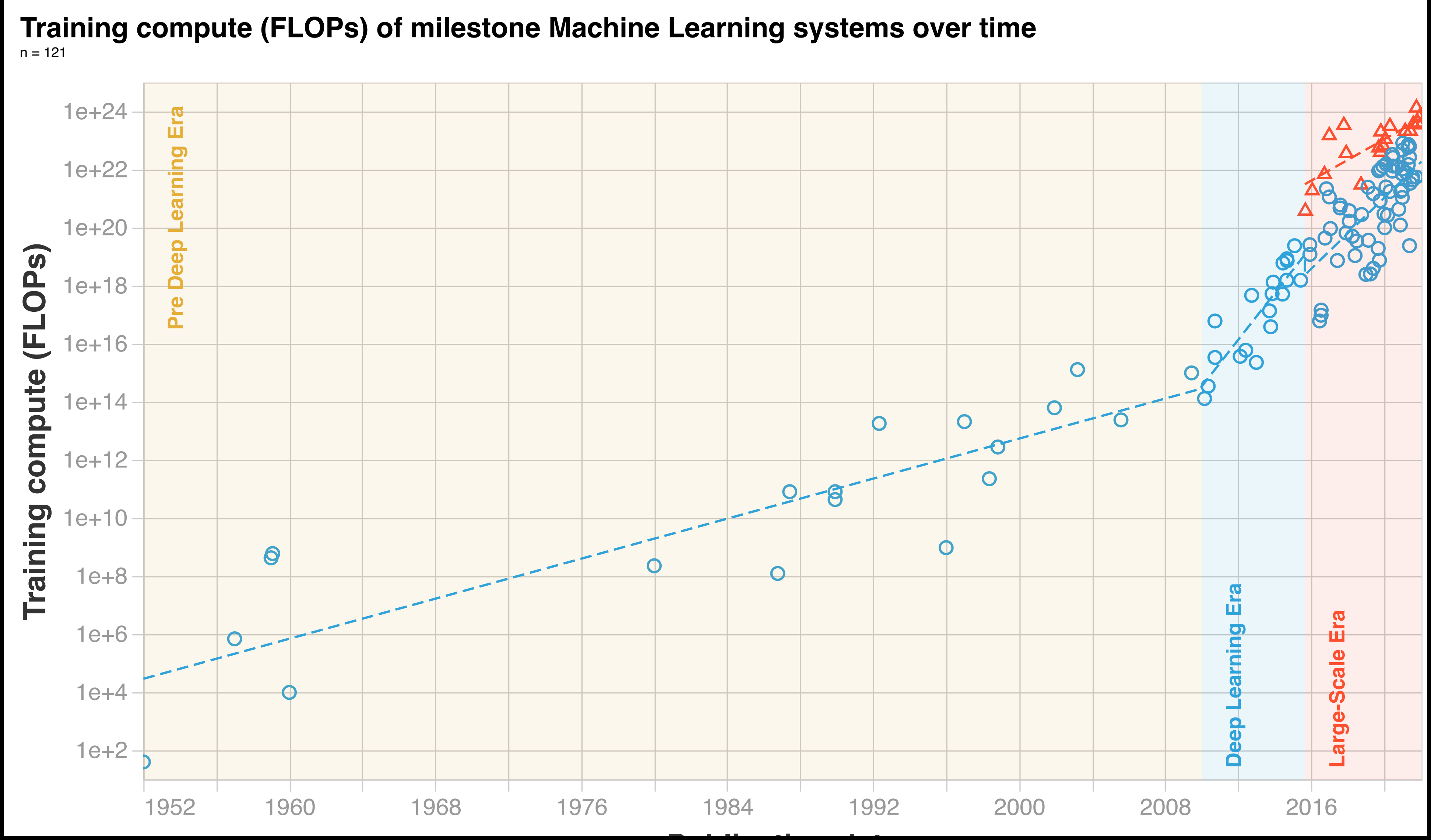
Microbial Data in the Environment

Tara Oceans Microbial data collected from 2009-13

84 Terabytes assembled on 9000 Frontier nodes

HPC changes observational science

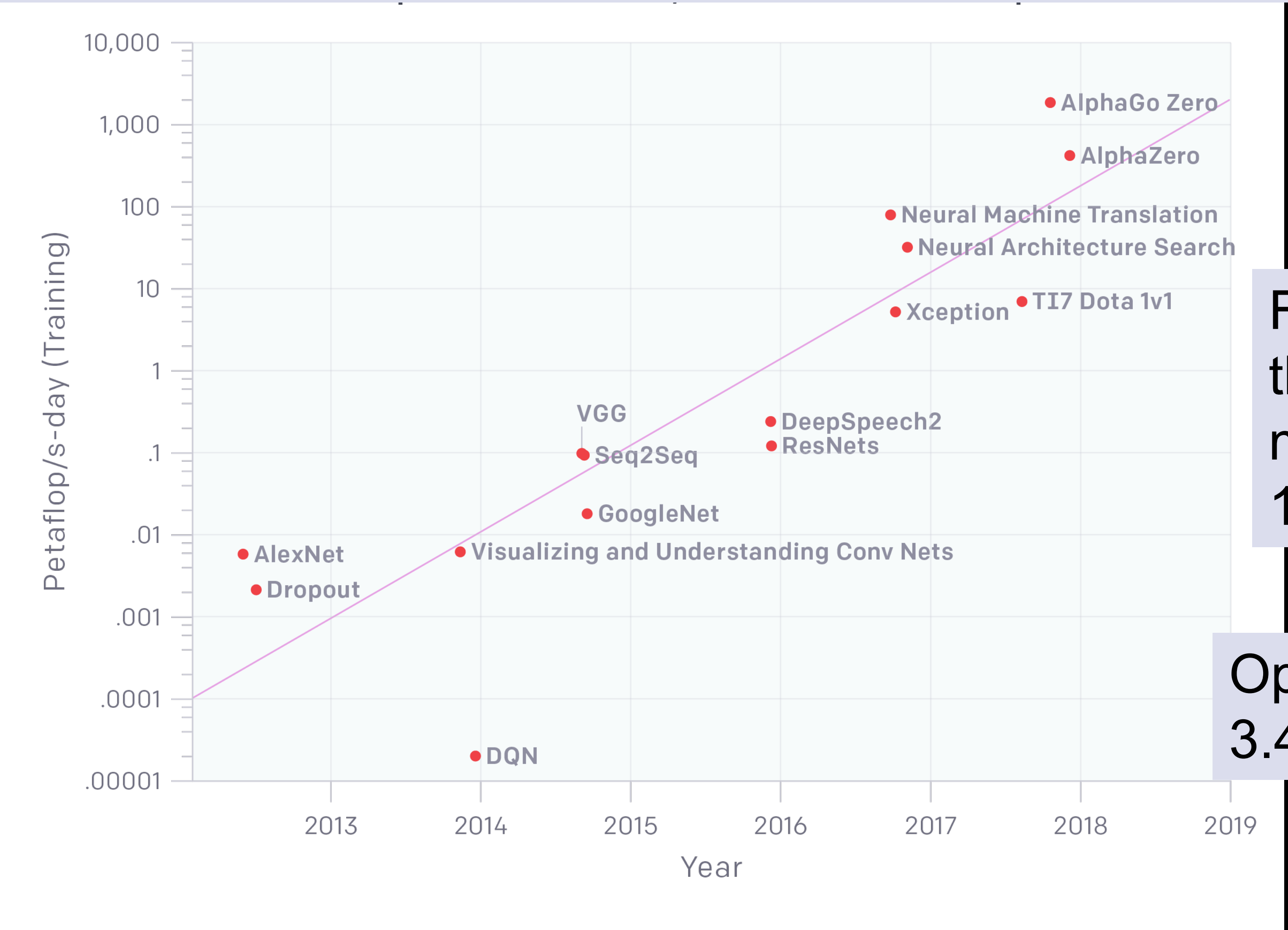
Machine Learning Drives Computational Demand



Computing Requirements in Machine Learning

300,000x increase from 2011 (AlexNet) to 2018 (AlphaGoZero)

A petaflop/s-day = 10^{15} neural net operations per second for one day, ~= 10^{20} operations

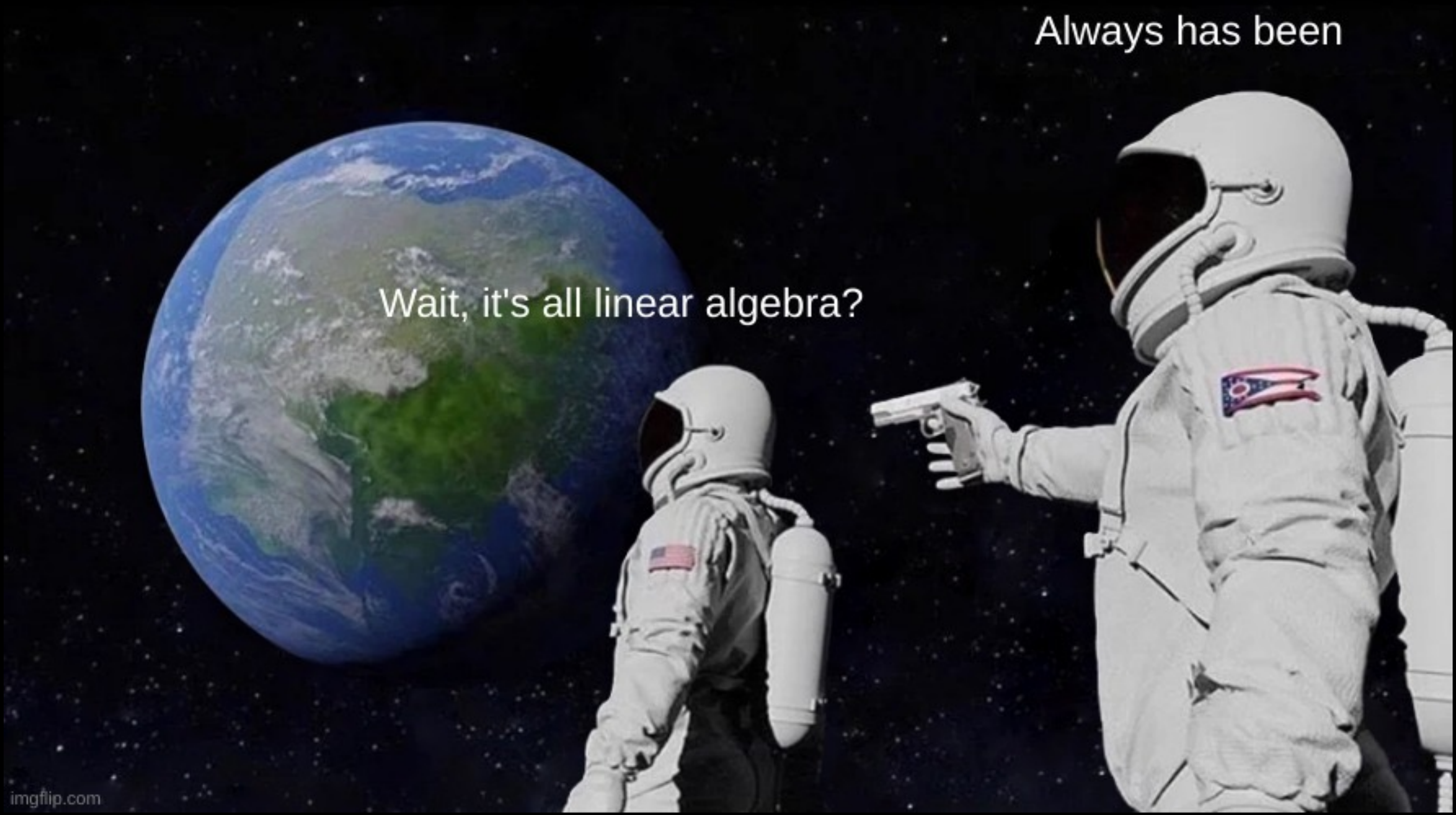


From 2011-2018 the fastest Top500 machine grew < 15x

OpenAI estimates 3.4-month doubling!

https://blog.openai.com/ai-and-compute/

Is there parallelism?

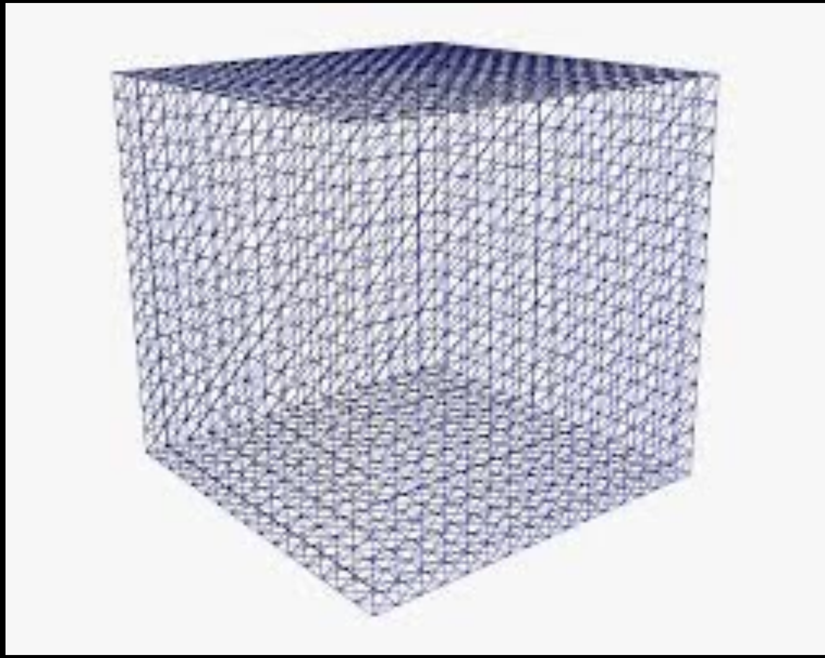

Always has been
Wait, it's all linear algebra?
imgflip.com

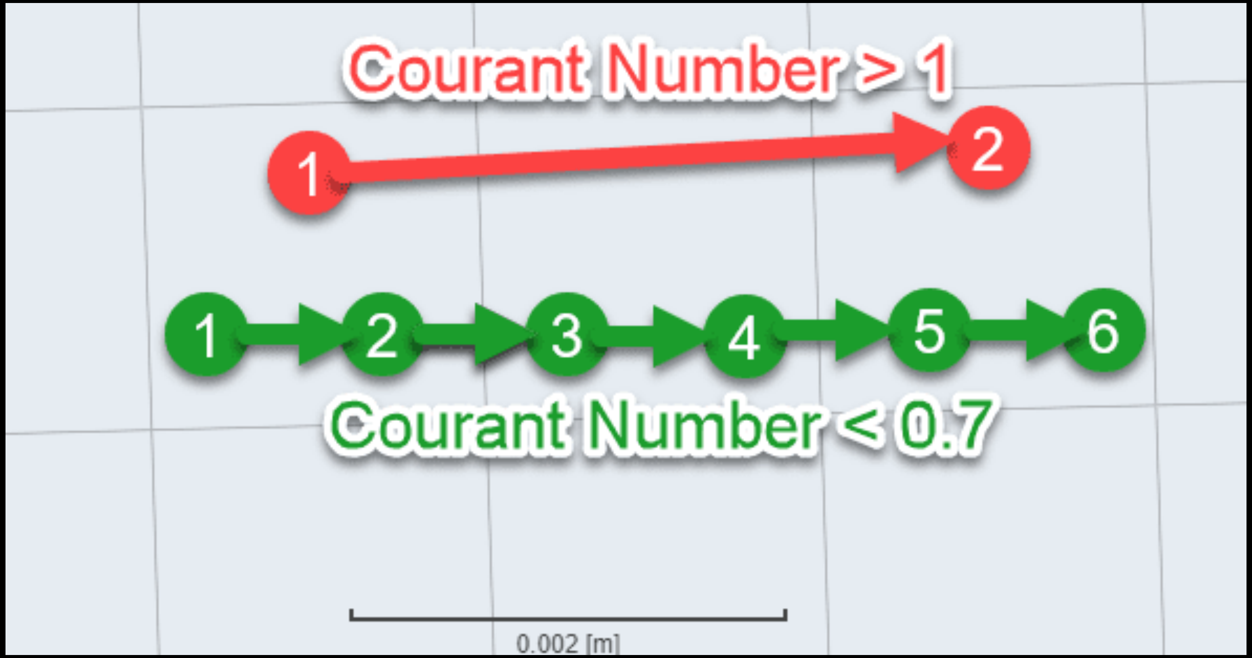
Analytics vs. Simulation Kernels:

7 Dwarfs of Simulation	7 Giants of Big Data
Particle methods	Generalized N-Body
Unstructured meshes	Graph-theory
Dense Linear Algebra	Linear algebra
Sparse Linear Algebra	Hashing
Spectral methods	Sorting
Structured Meshes	Alignment
Monte Carlo methods	Basic Statistics
Phil Colella	NRC Report + our paper

Yelick, et al. “The Parallelism Motifs of Genomic Data Analysis”, Philosophical Transactions A, 2020

Weak Scaling has Diminishing Returns





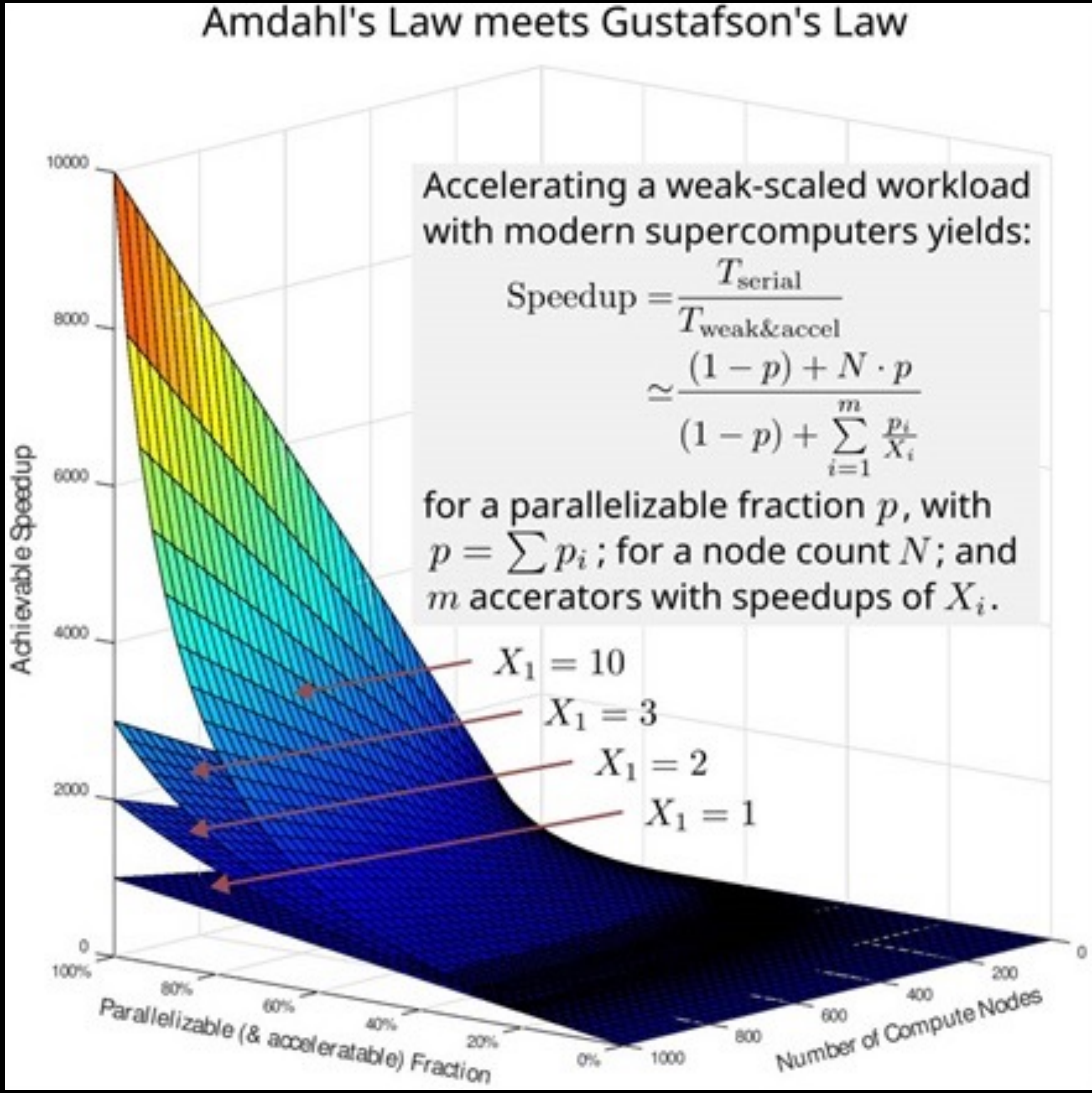
Increase resolution by 10x in each dimension

Increase cores by 1000x

Runtime increases ☹

Strong and weak scaling

- Strong scaling
  - Most desirable for users
  - Harder to find (Amdahl)
- Weak scaling
  - Limited for super-linear algorithms
  - Needs memory capacity to scale
  - Data problems also need I/O

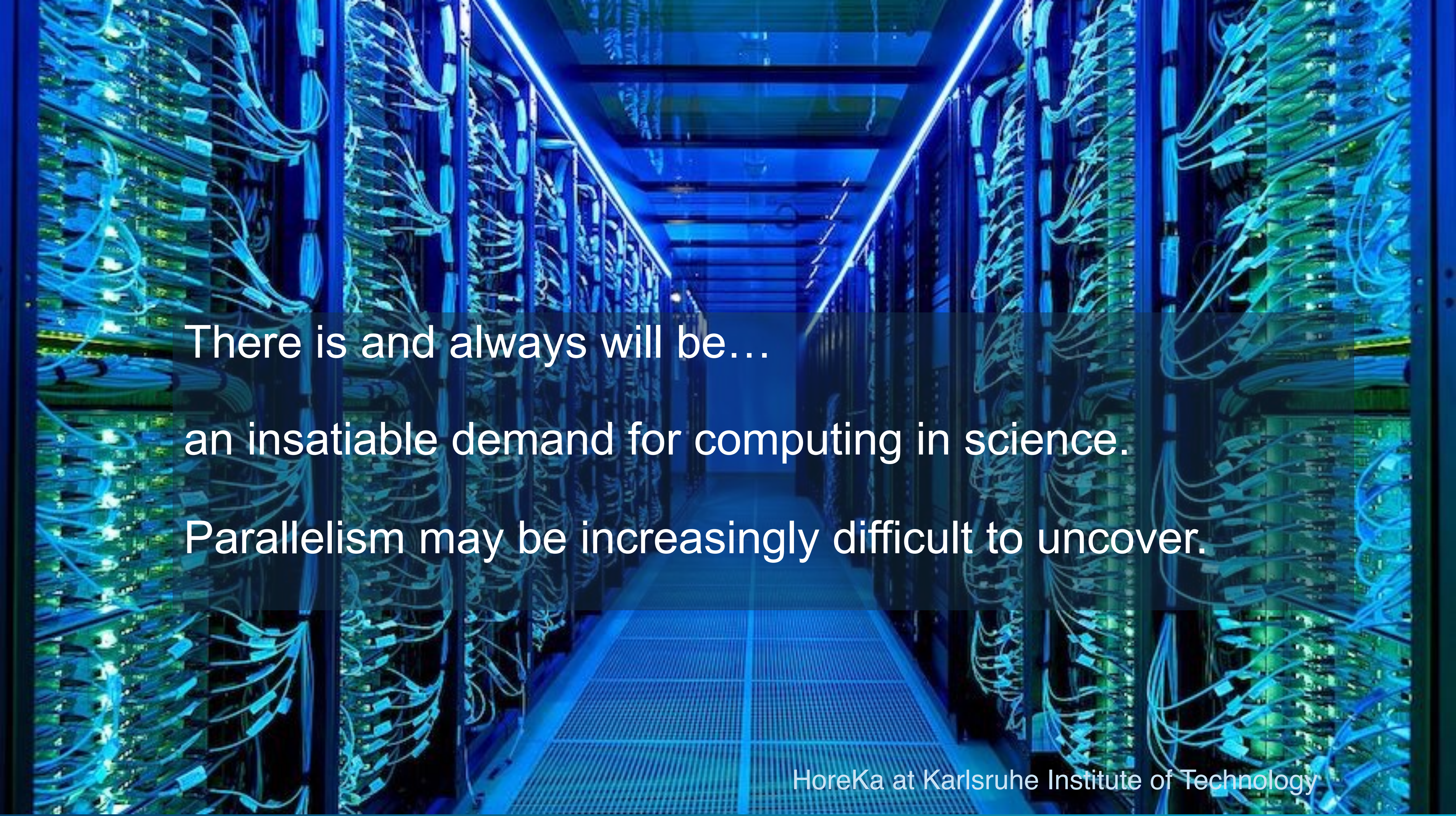


See SIAM News, 9/22 Satoshi Matsuoka and Jens Domke

There is and always will be…

an insatiable demand for computing in science.

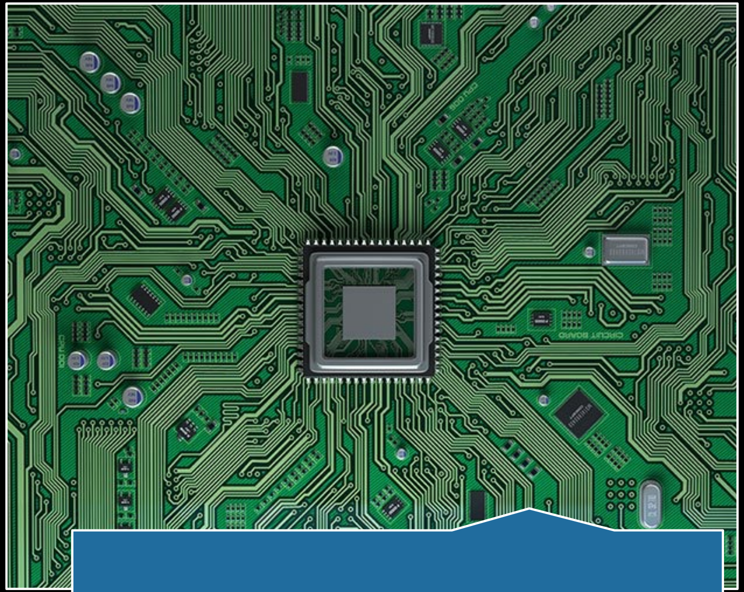
Parallelism may be increasingly difficult to uncover.



HoreKa at Karlsruhe Institute of Technology

Post-Exascale Computing

Computing demand



Disruptions

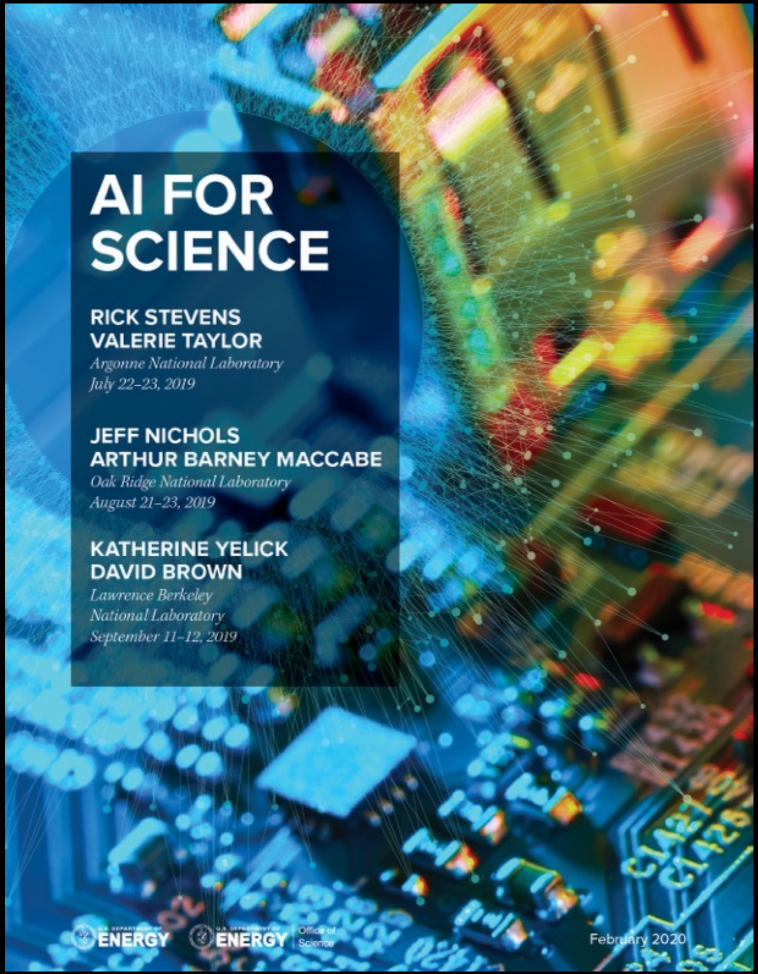
Technology

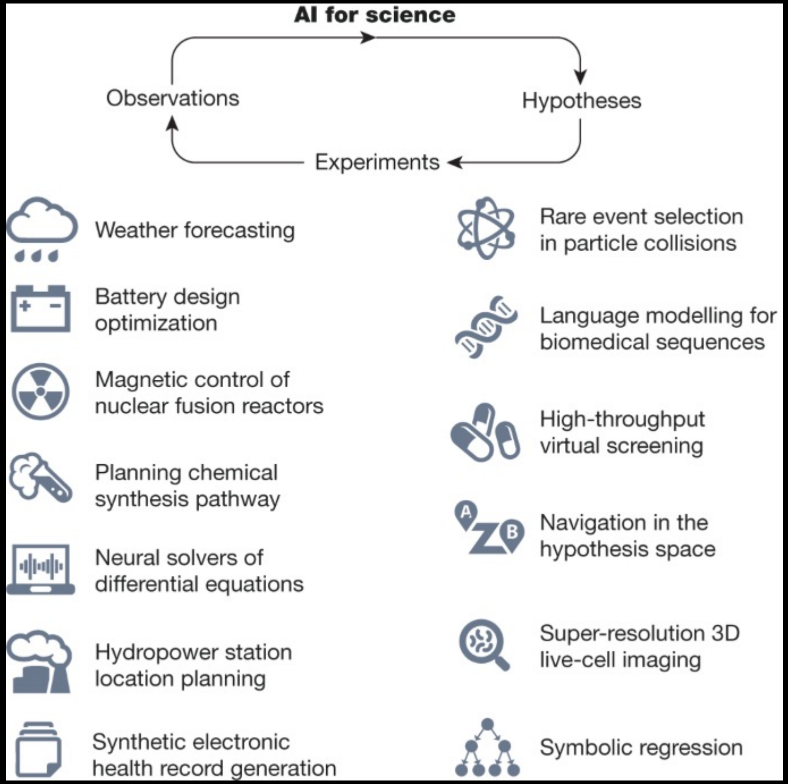
Disruptions

- AI
- Quantum
- Cloud

Implied question: Do these make HPC obsolete?

AI for Science

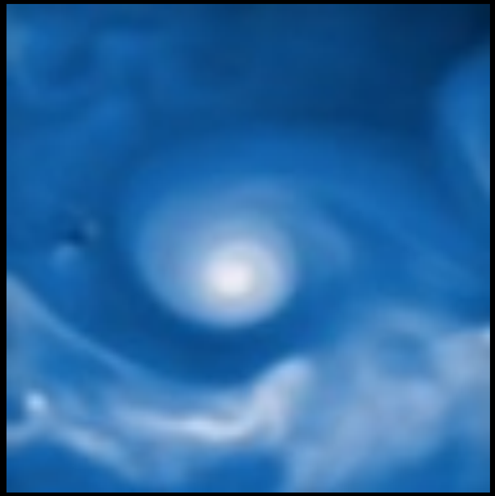




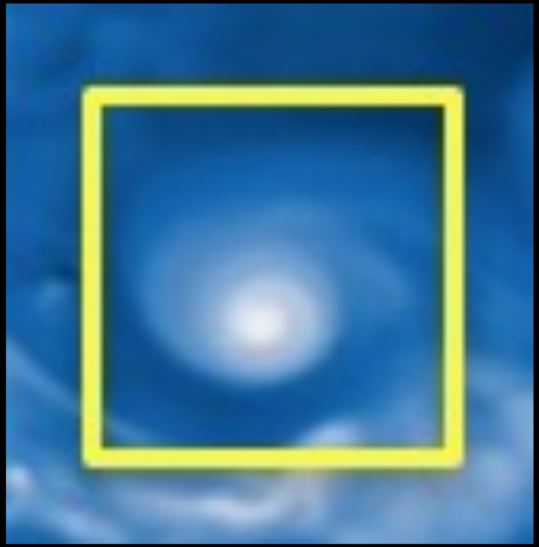
Scientific discovery in the age of artificial intelligence, 2023

Analyze Simulations to Find Hurricanes

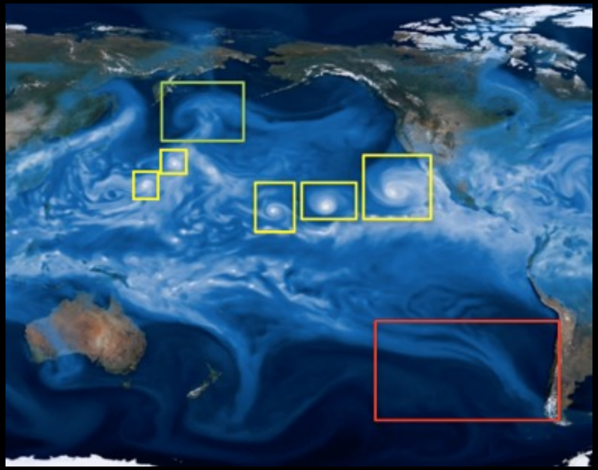
Classification



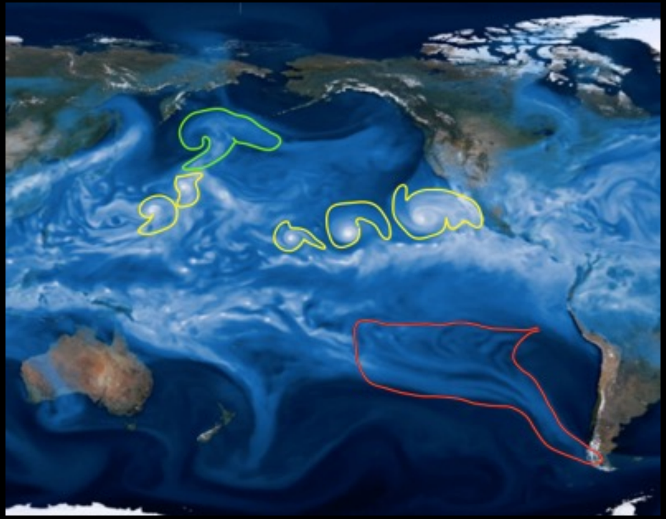
Localization



Detection



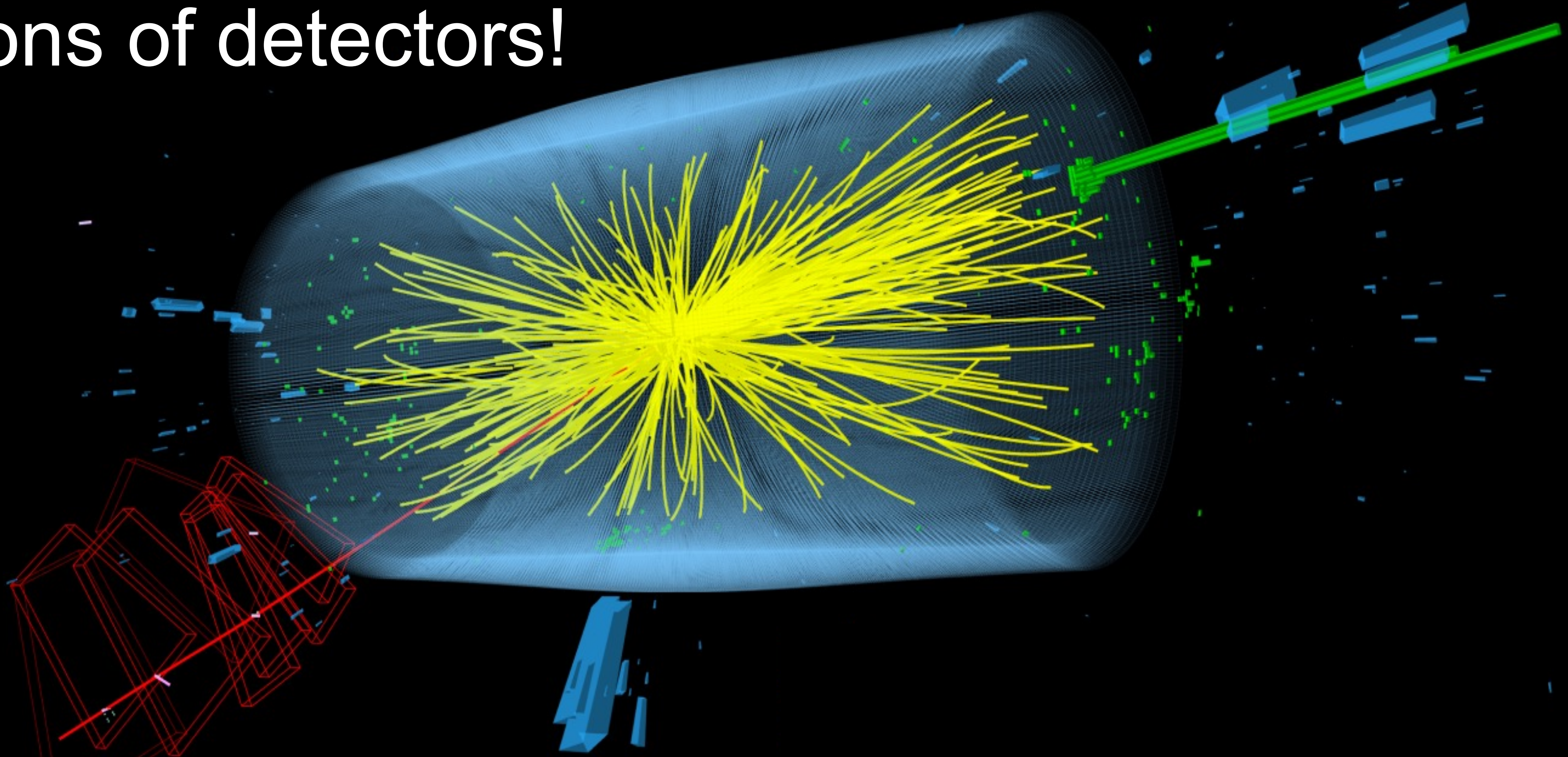
Segmentation



Extending image-based methods to complex, 3D, scientific data sets is non-trivial!

Source: Prabhat

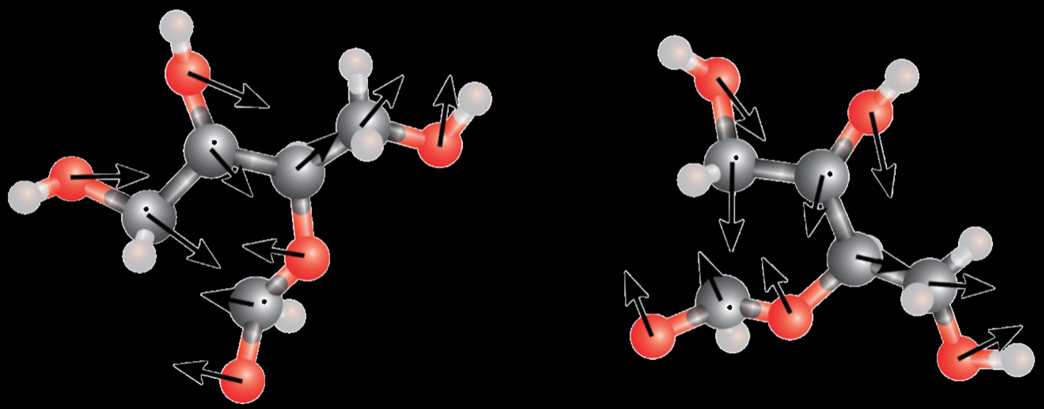
Deep Learning: like adding 4,000 extra tons of detectors!

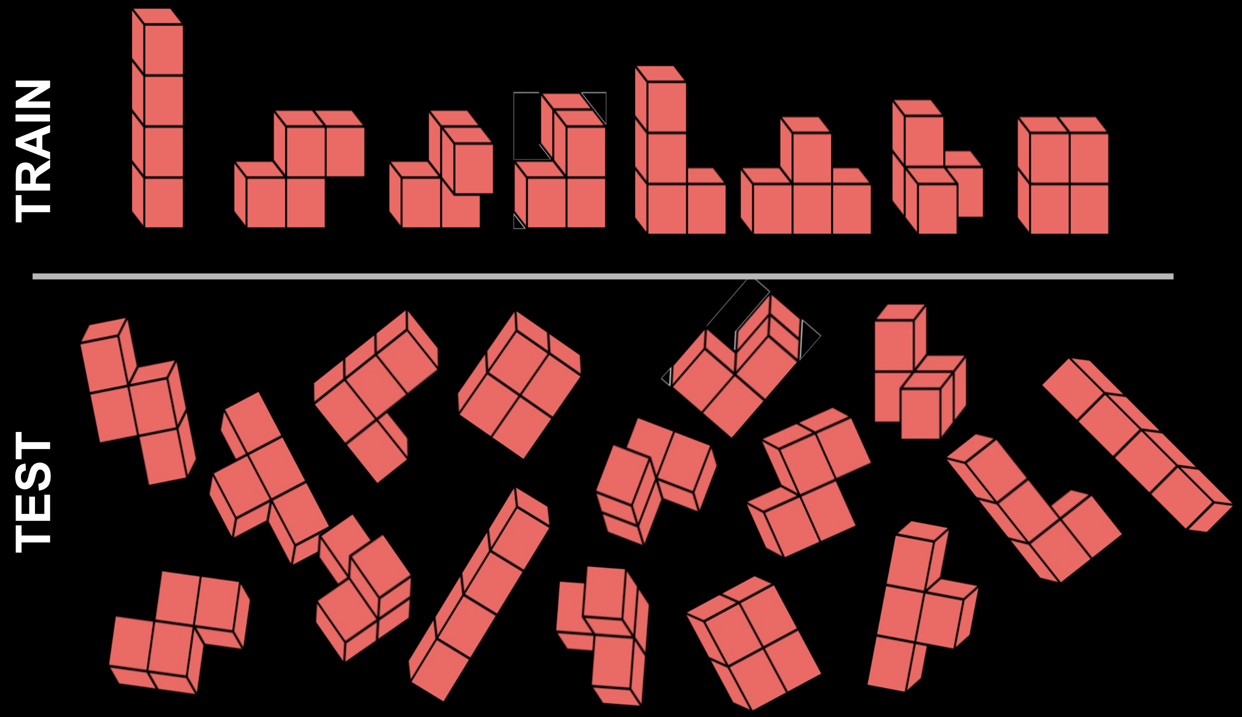


Based on 8/12/2016 slide by Joe Lykken at Fermilab

Deep Learning with Physical Laws

Physics-aware learning





A network with 3D translation- and 3D rotation-equivariance

Slides from Tess Smidt and Risi Condor; E.g., 2018 paper by Thomas, Smidt, Kearnes, Yang, Li, Kohlhoff, Riley

Automation in Self-Driving Laboratories

E.g., Strateos Cloud Lab
14K square feed
200+ instruments

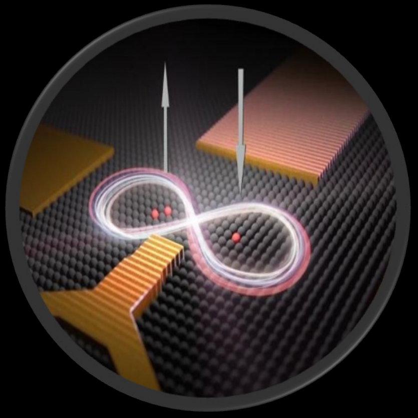
Five Stages of AI

And this includes AI researchers!

AI in Science

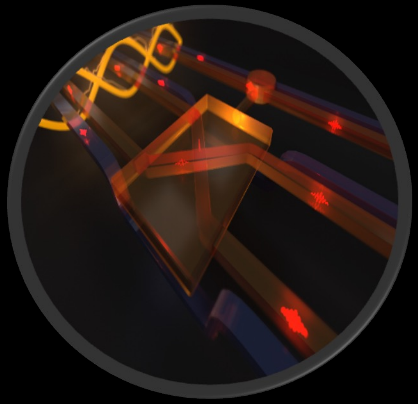
The Computational Science and Engineering community (including NNSA) should have a leadership role in addressing UQ, safety, alignment, and explainability in AI for science and engineering

Types of Quantum Bits Diversity & Progress



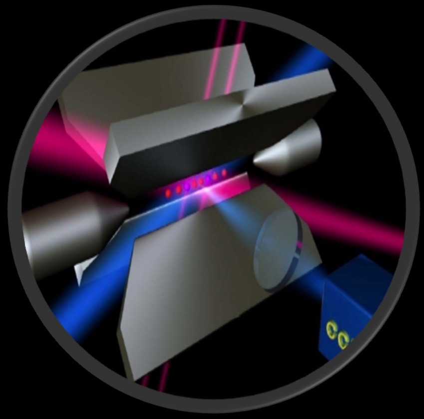
Dopants in Silicon / Diamond

www.sciencedaily.com



Photonic Circuits

www.phys.org

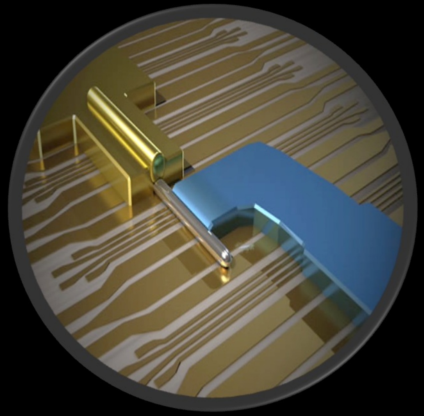


Trapped Ions

www.quantumoptics.at

Superconducting Circuits

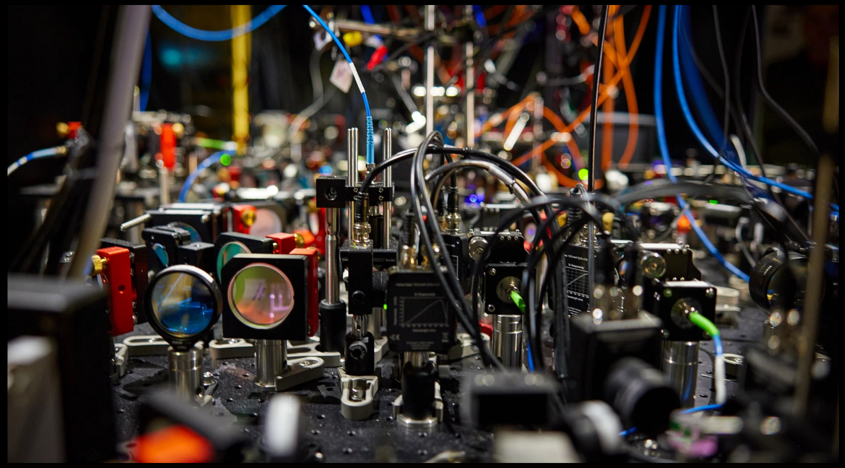
www.qnl.berkeley.edu



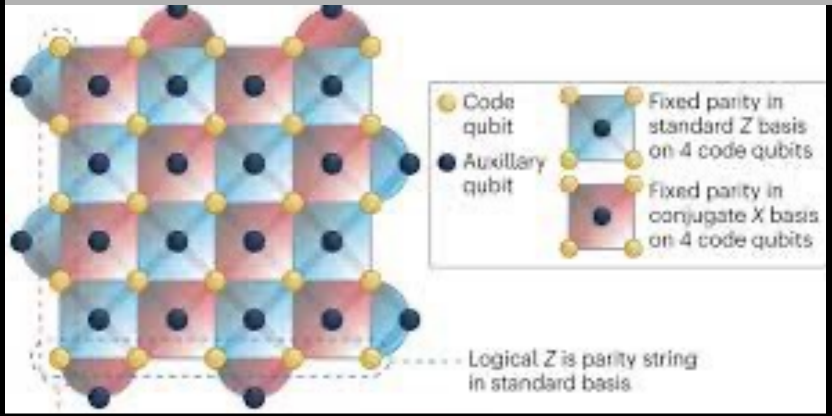
Topological Wires

www.microsoft.com

High-fidelity parallel entangling gates on a neutral-atom quantum computer



A series of fast-paced advances in Quantum Error Correction

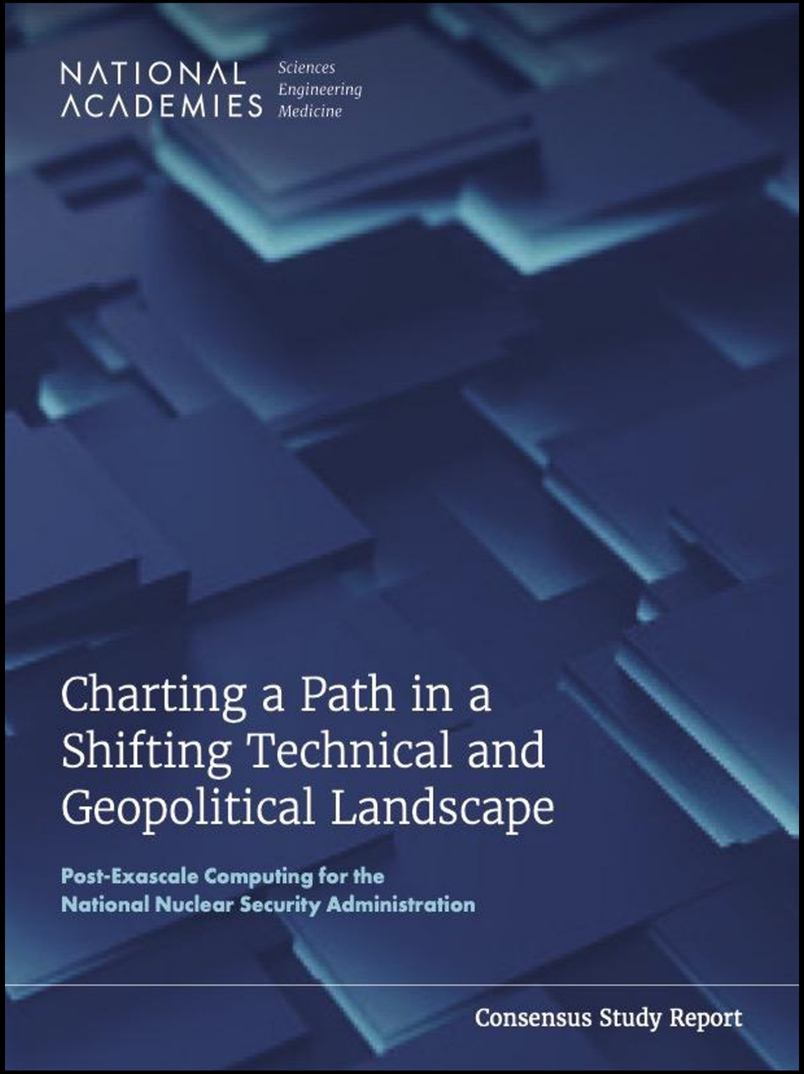


National Academies Study

Finding: **Quantum technology** has the potential to improve the fundamental understanding of material properties.

However, **breakthroughs in quantum algorithms and systems are needed** to make quantum computing practical for multiphysics stockpile modeling.

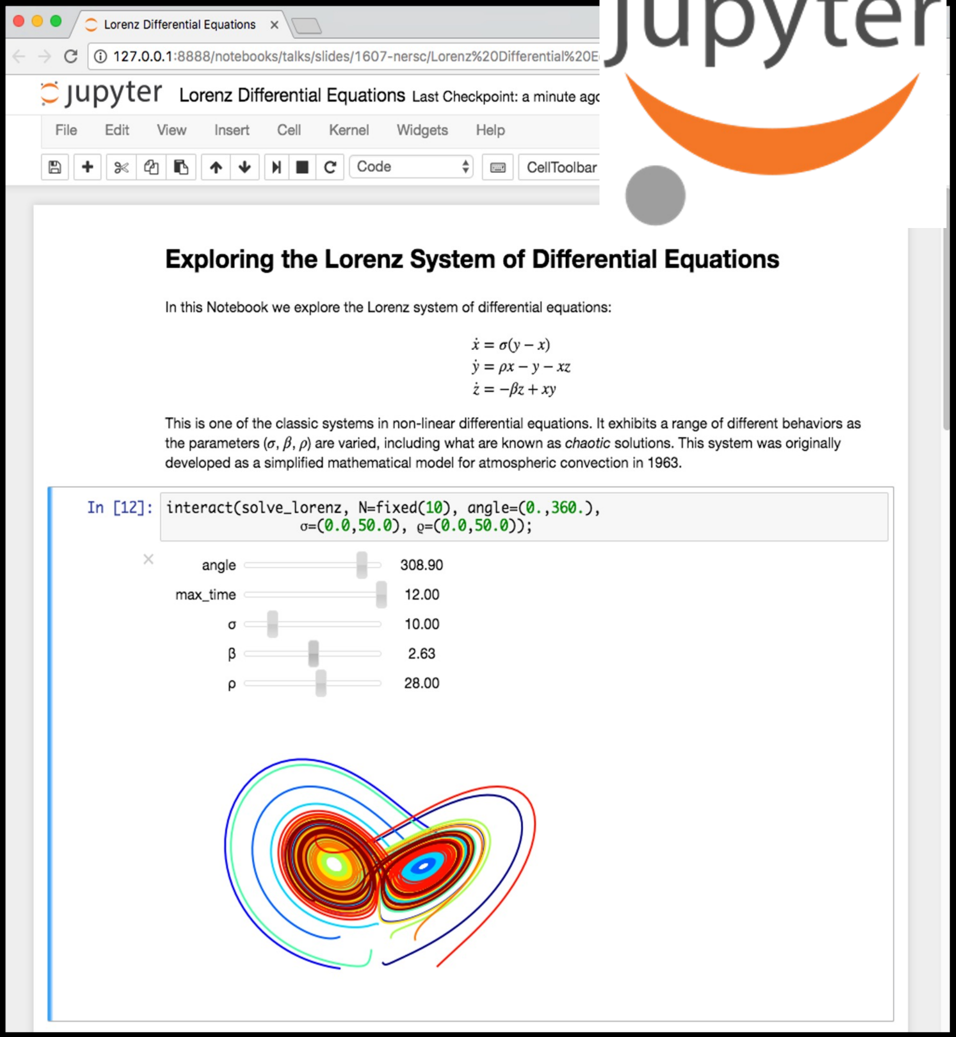
Quantum computing is more likely to serve as a **special-purpose accelerator** than to replace leading-edge computing.



Cloud Computing

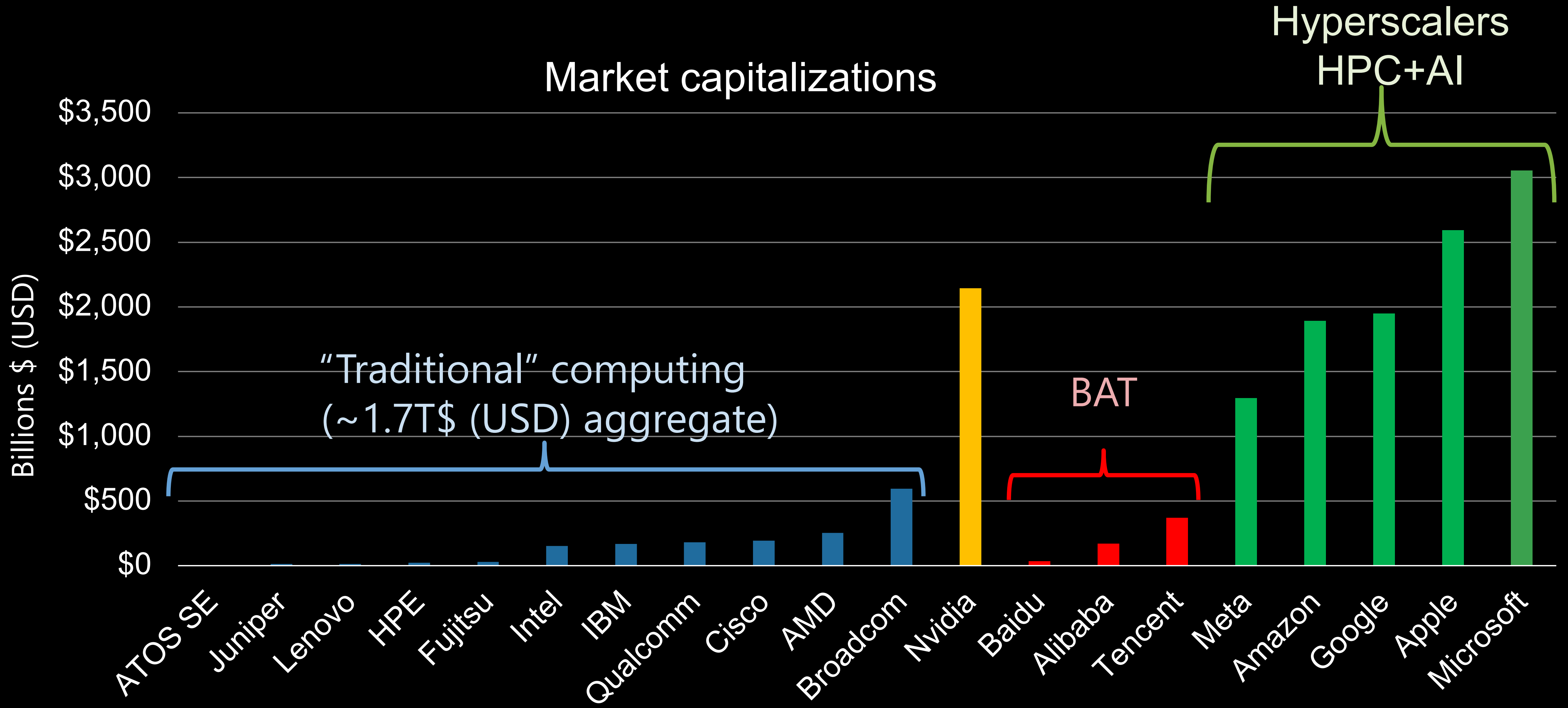
Lessons Learned from Clouds

- Availability
- Cost vs price
- Higher level programming



Old programming models never die, they just get buried under layers!

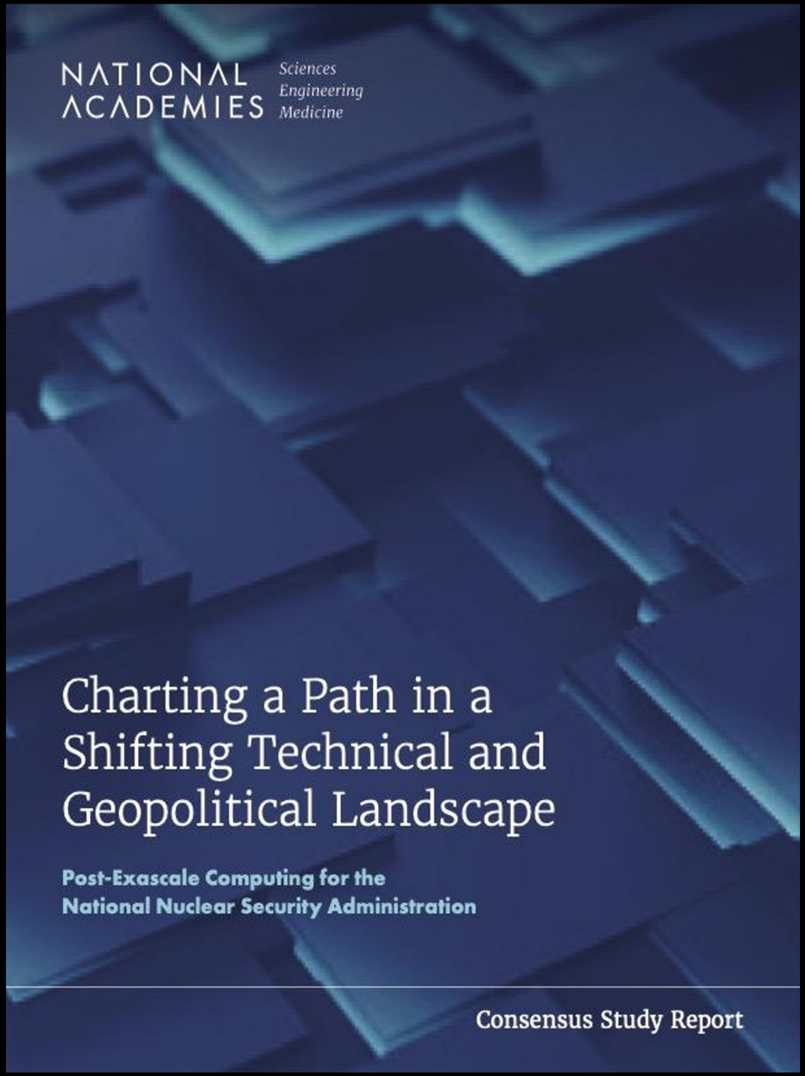
Follow the money, understand the implications



Source: Reed, Gannon, Dongarra

National Academies Study

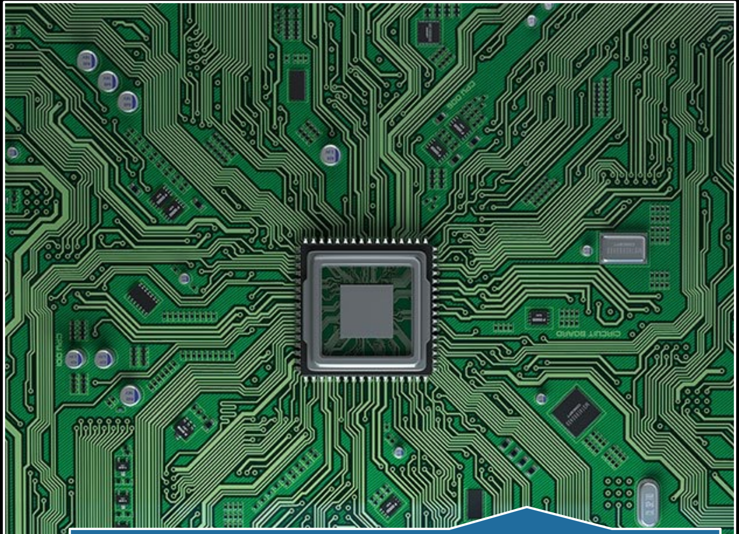
Finding: **Cloud providers are engaged in hardware and software innovations** and will have more market influence in **technology and talent** but are **not aligned with NNSA** requirements.



HPC community has always punched above its weight

Post-Exascale Computing

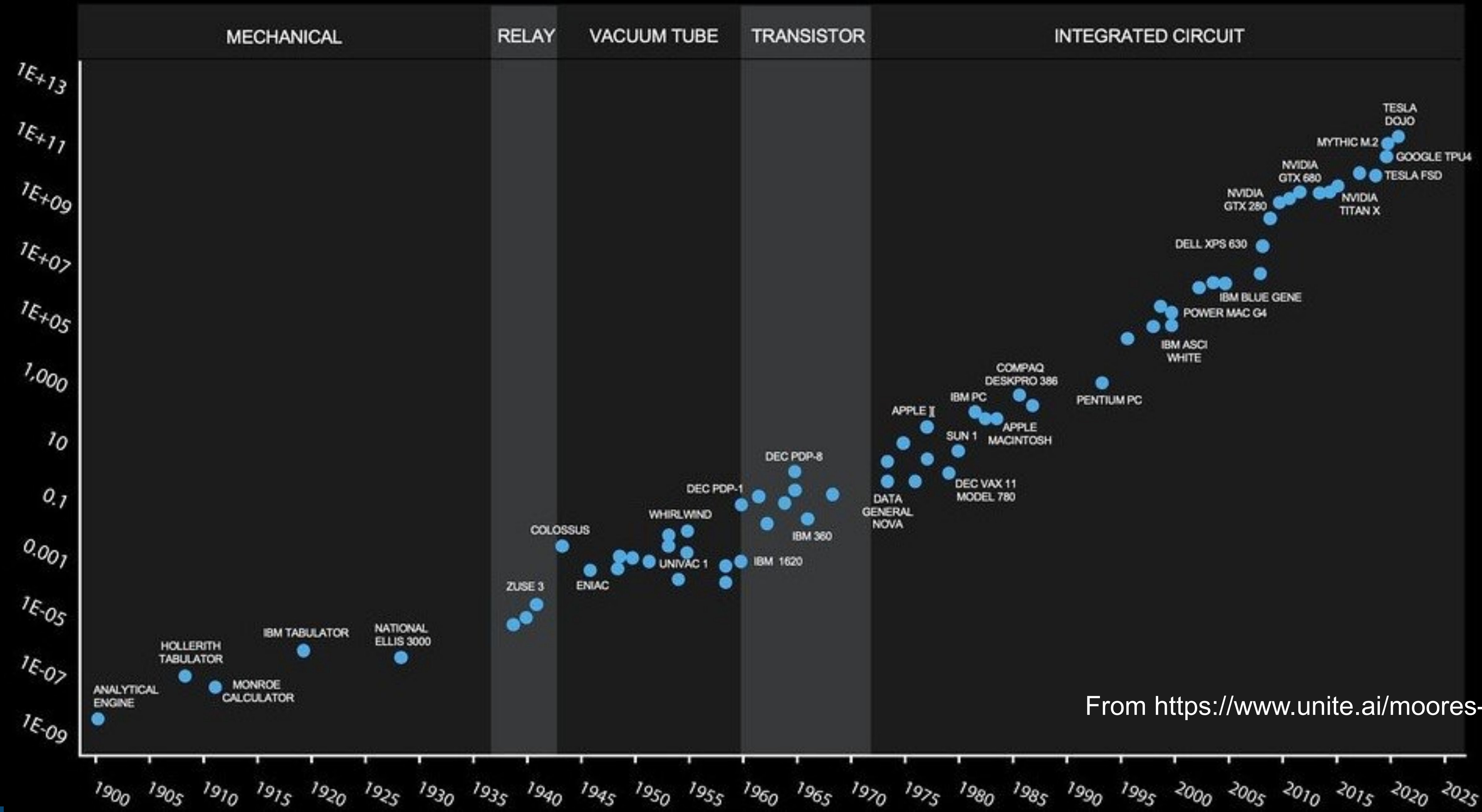
Computing demand



Disruptions

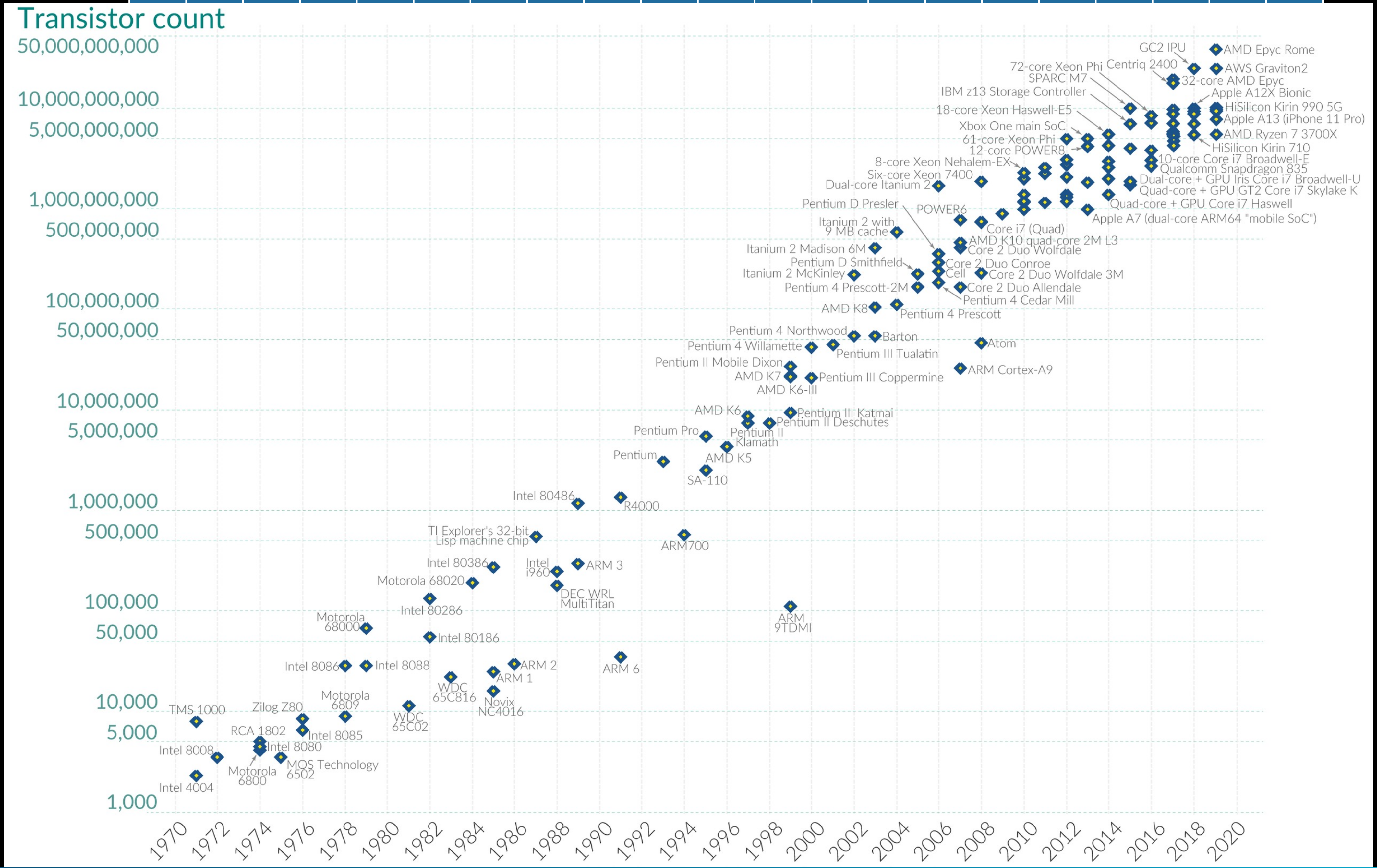
Technology

122 YEARS OF MOORE'S LAW

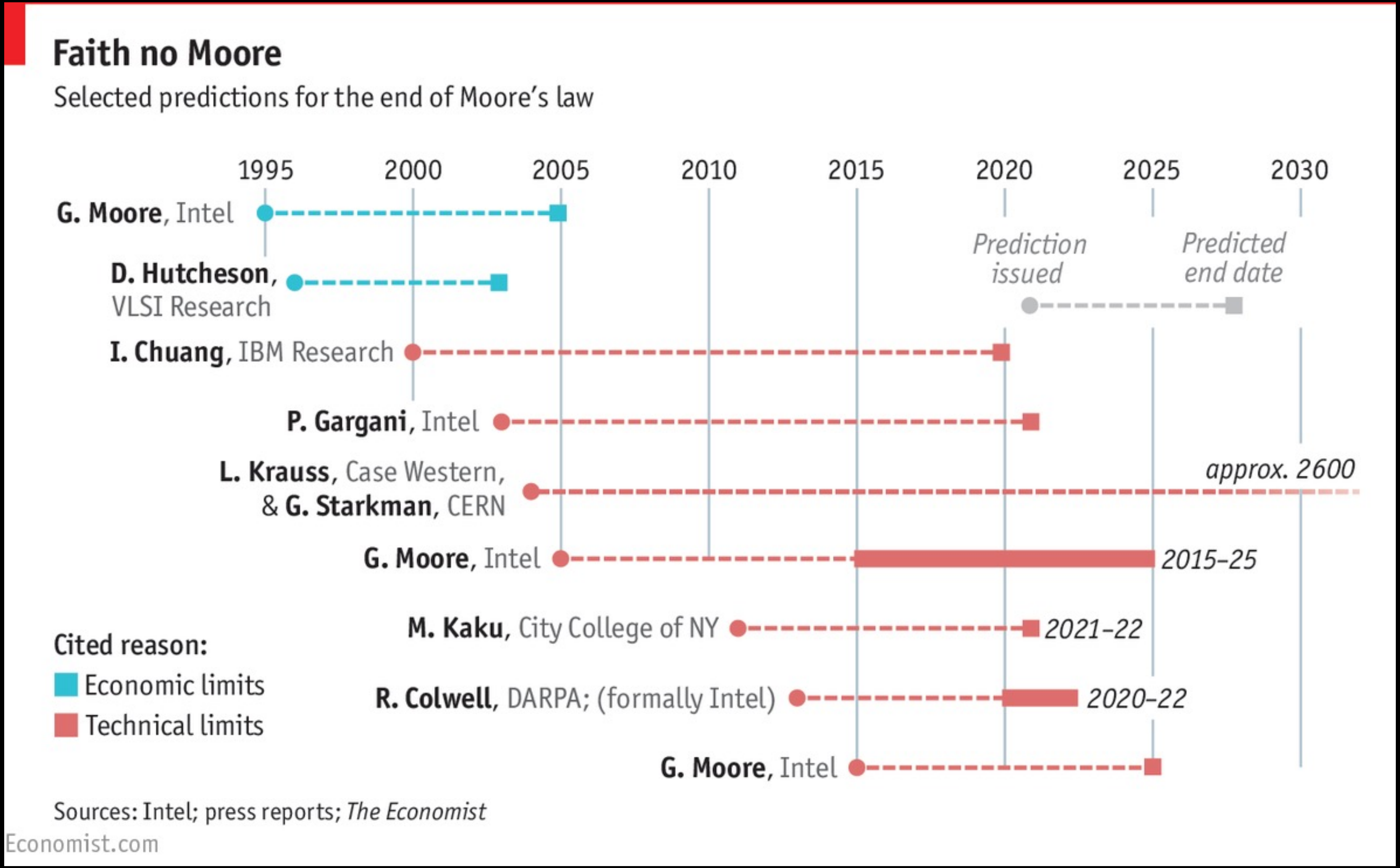


From https://www.unite.ai/moores-law/

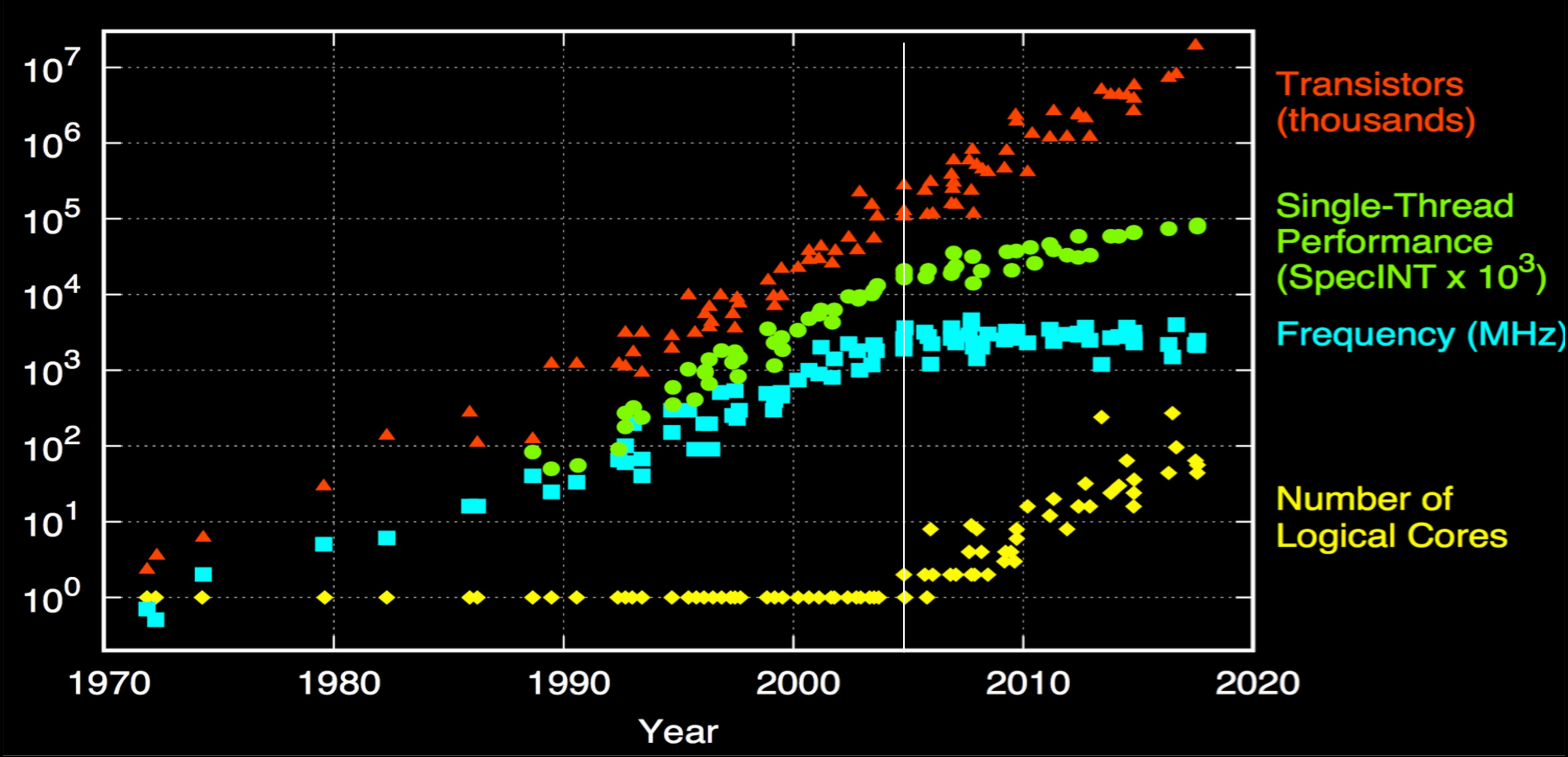
10 µm	6 µm	3 µm	1.5 µm	1 µm	800 nm	600 nm	350 nm	250 nm	180 nm	130 nm	90 nm	65 nm	45 nm	32 nm	22 nm	14 nm	10 nm	7 nm	5 nm	3 nm



https://www.economist.com/technology-quarterly/2016/03/10/horses-for-courses



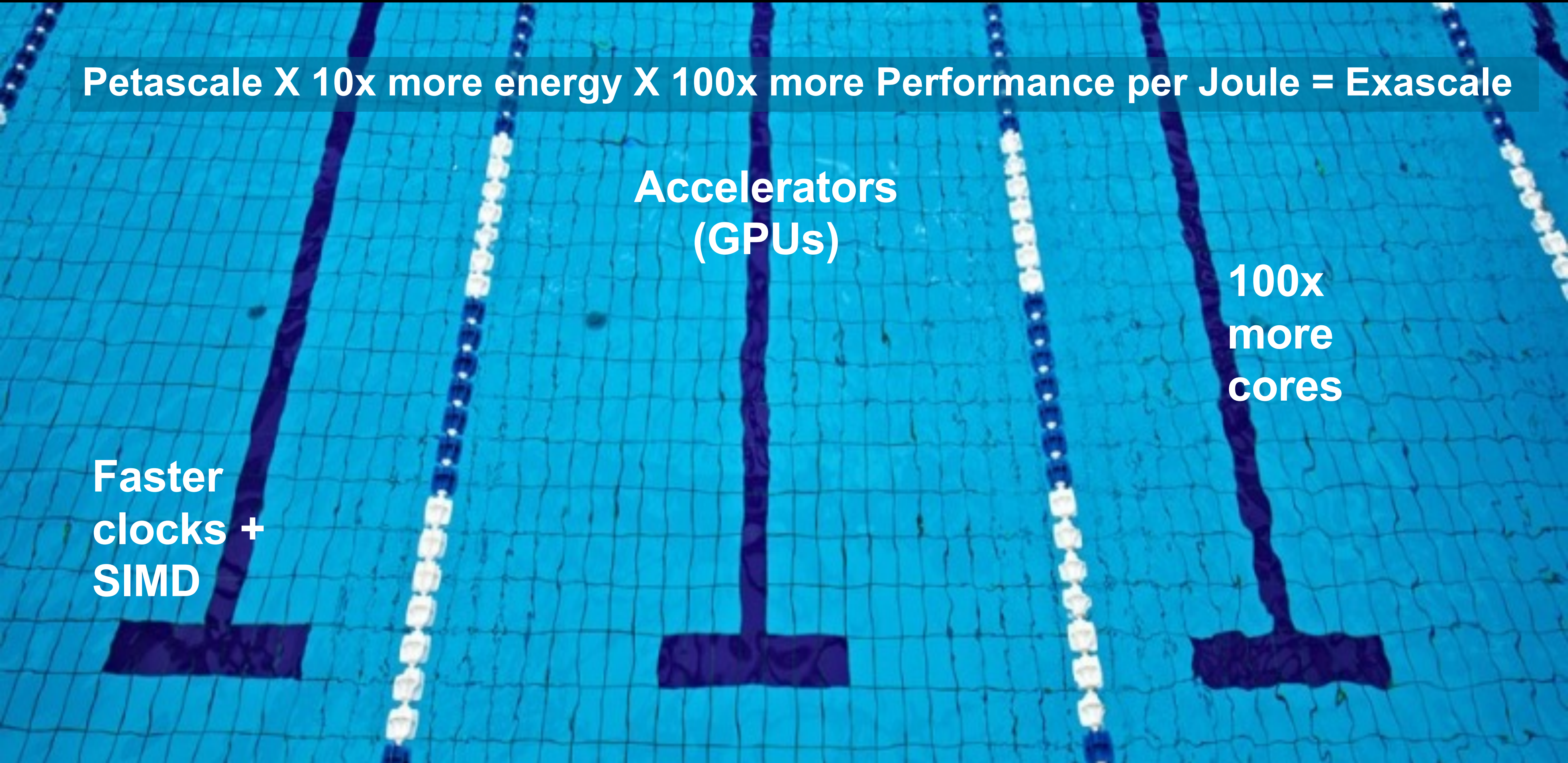
Dennard Scaling is Long Dead; Moore’s Law Will Follow



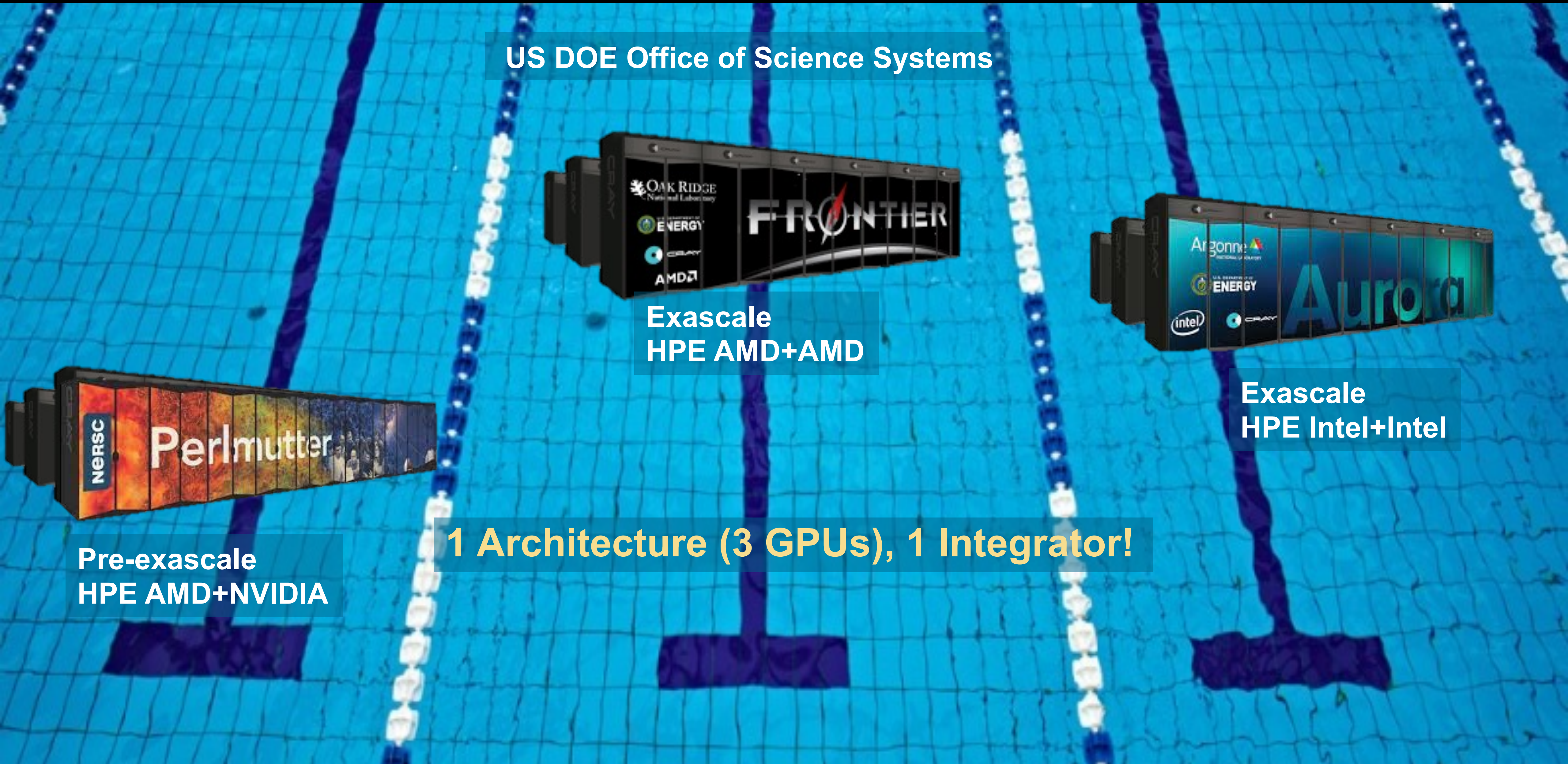
M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, C. Batten, and K. Rupp

Performance Programming pre 2005

Exascale Architecture Plans (2008)



Exascale Era Architectures (US DOE Office of Science)



First-in-Class HPC Systems (Top500)

	First TF	First PF	First EF
	ASCI Red	Roadrunner	Frontier
Month-Year	Jun-97	Jun-08	Jun-22
Best Tech (nm)	500	65	6
Peak (PF/s)	0.001453	1.38	1686
Sustained (PF/s)	0.001068	1.04	1102
Power (MW)	0.85	2.35	21.1
Efficienty (GF/W)	0.00125647	0.44	52.2
Memory (PB)	0.001212	0.04	9.4
FPUs (K)	9	464	534,000
Cabinets	104	296	74
Foorspace (m^2)	150	557	678

Kogge and Dally: Frontier vs the Exascale Report: Why so long? and Are We Really There Yet?
+Wikipedia for ASCI Red

Energy efficiency didn't track technology scaling

Gate Length (nm)	65	32	16	6
Metal 1 pitch (nm)	180	100	64	40
Energy^{-1}	1	1.8	2.8	4.5
Area^{-1}	1	3.2	7.9	20.3

Rumors of 2nm fabs, but how much will it help?

Kogge and Dally: Frontier vs the Exascale Report: Why so long? and Are We Really There Yet?

The “Aggressive” Strawman was a bit early, but close to Summit

	Strawman	Summit
Year	2015	11/2018
Best Tech	32nm	16nm
Peak (PF/s)	2,000	201
Sustained (PF/s)	1,000	148
Power (MW)	67.7	9.8
Efficiency (GF/W)	14.9	14.7
Memory (PB)	3.5	2.8
Bandwidth/flop (B/F)	0.08	0.13
Mem BW (PB/s)	158	27
Bisection(TB/s)	210	105
FPUs (M)	664	144
Cabinets	583	256
Floorspace (m^2)	1195	520

Resilience would have been a bigger problem with a 7x larger Summit

Kogge and Dally: Frontier vs the Exascale Report: Why so long? and Are We Really There Yet?

Post-Exascale Architecture Plans 2024 (Strawperson-v0)

Exascale X 2x more energy X 500x more Performance per Joule ??

GPUs

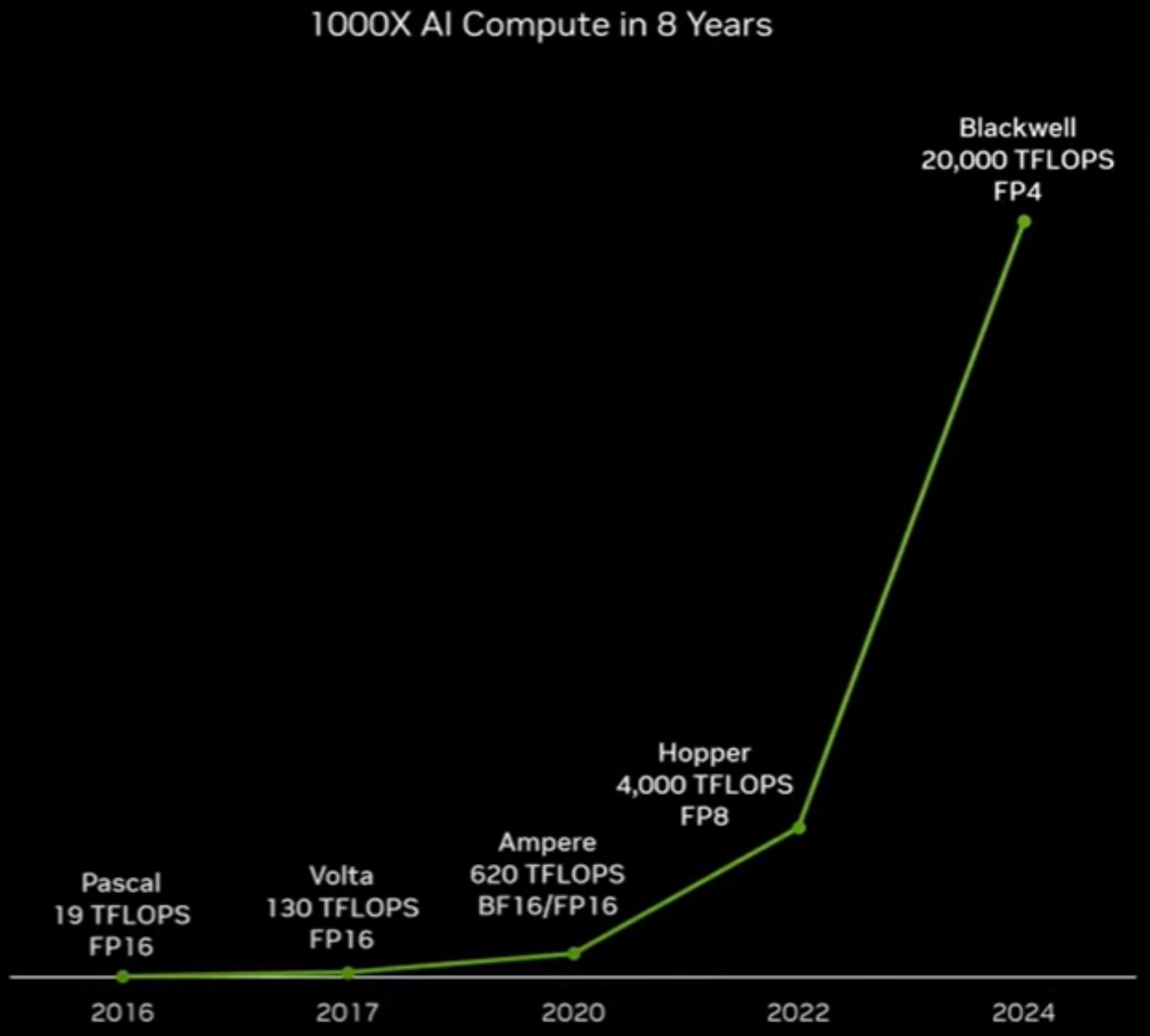
Influenced to make AI better (e.g., sparsity)?

Specialized for AI

Specialized for Simulation

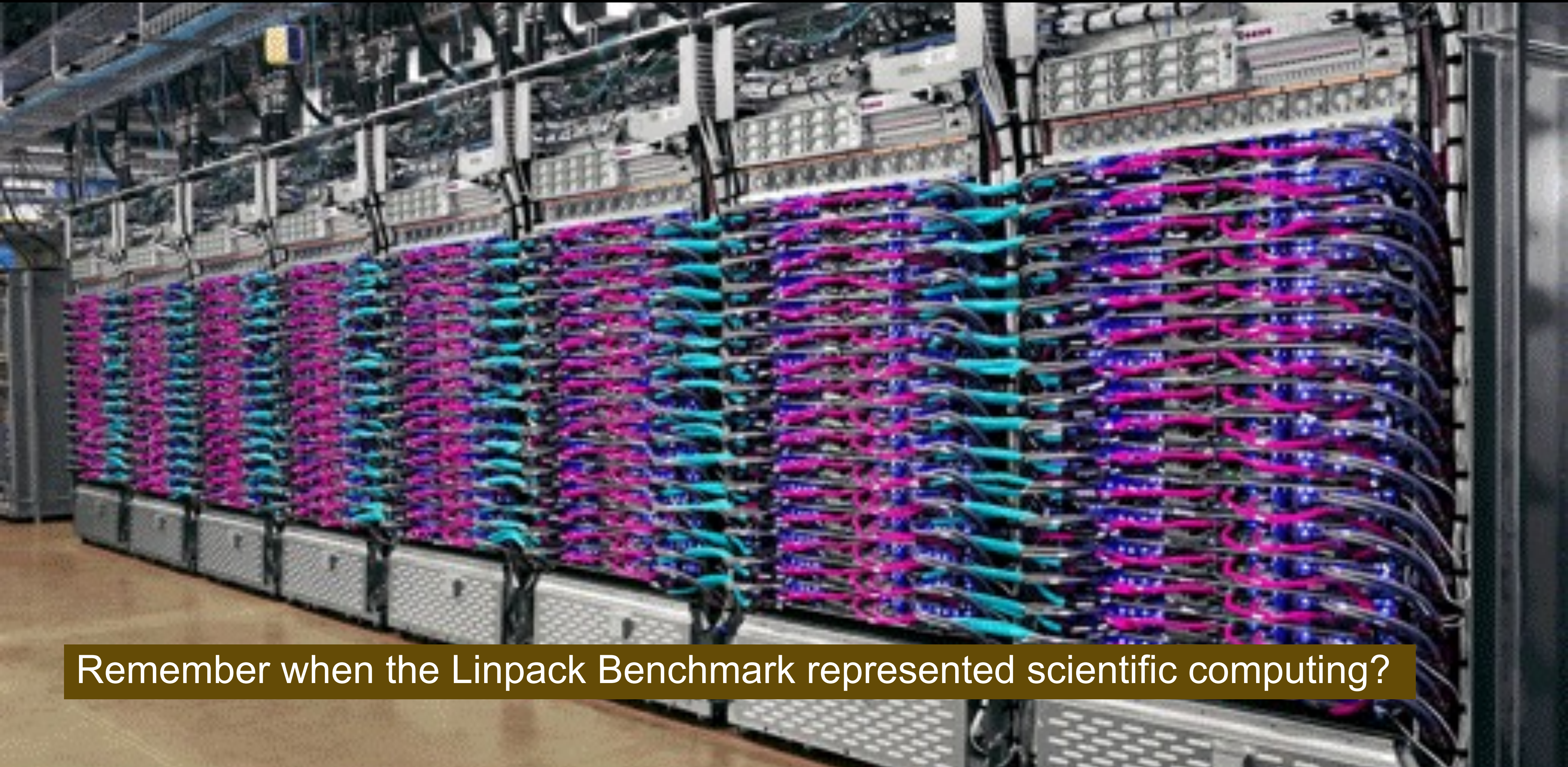
Designed by DOE, DoD, …?

Another Exponential?



Jensen Huang's Nvidia GTC Keynote

Specialization: Is deep learning the only application?



Remember when the Linpack Benchmark represented scientific computing?

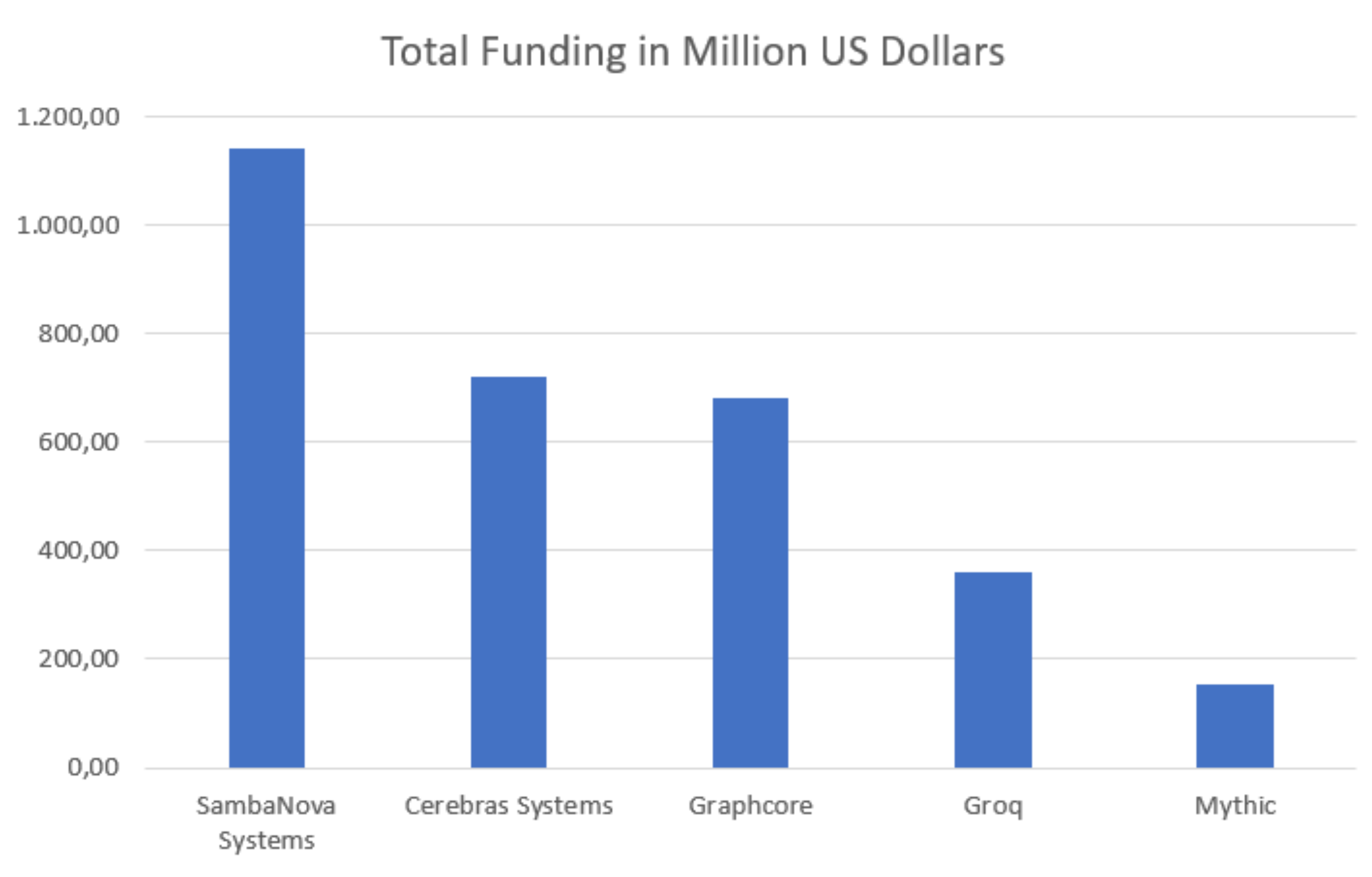
Everyone is Making AI Chips

NVIDIA
AMD
Intel
IBM

Traditional chip makers

"Software" companies

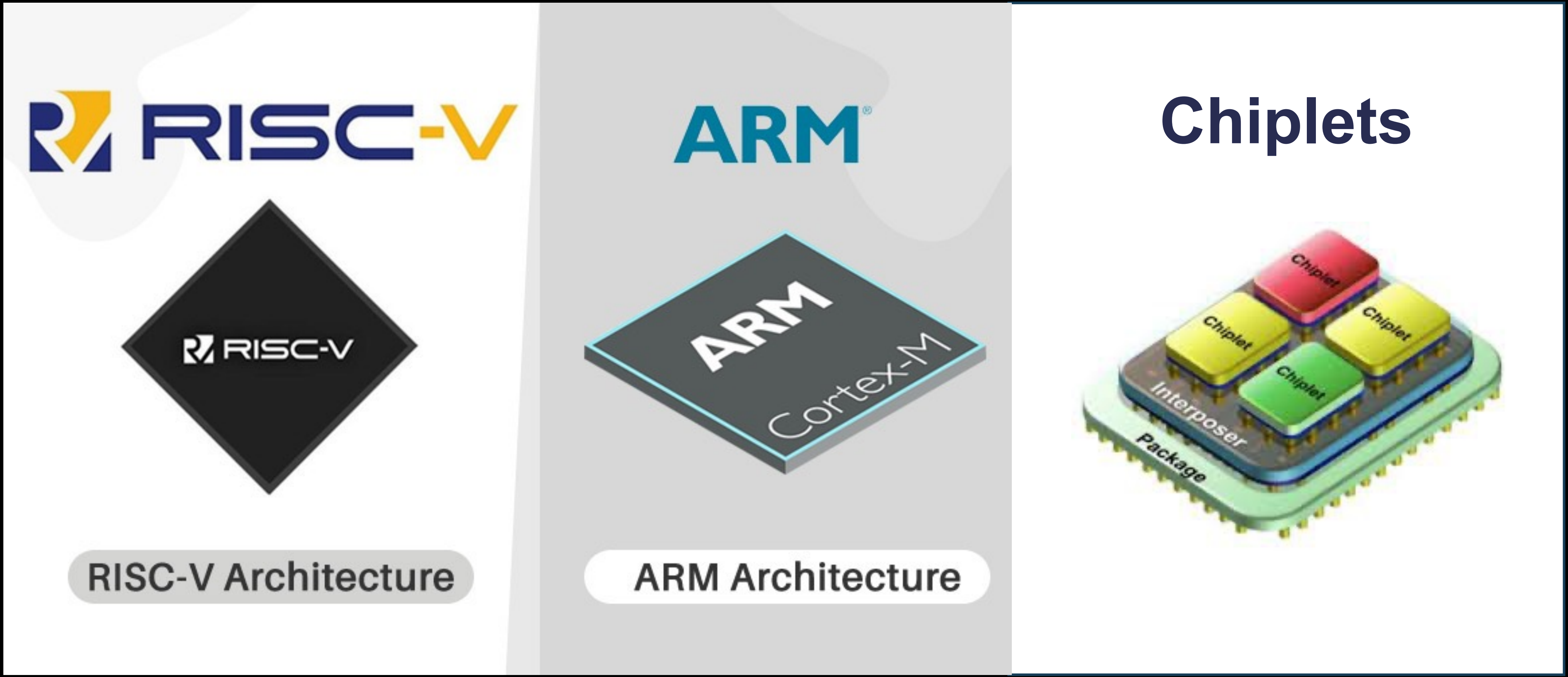
Facebook + Intel
Amazon (Echo, Oculus)
Google (TPU, Pixel)
Apple (SoCs)
Microsoft ("AI chip")



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

Not everyone is selling those chips!

Specialization for the masses?



Technology and Marketplace: Radically Different!

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

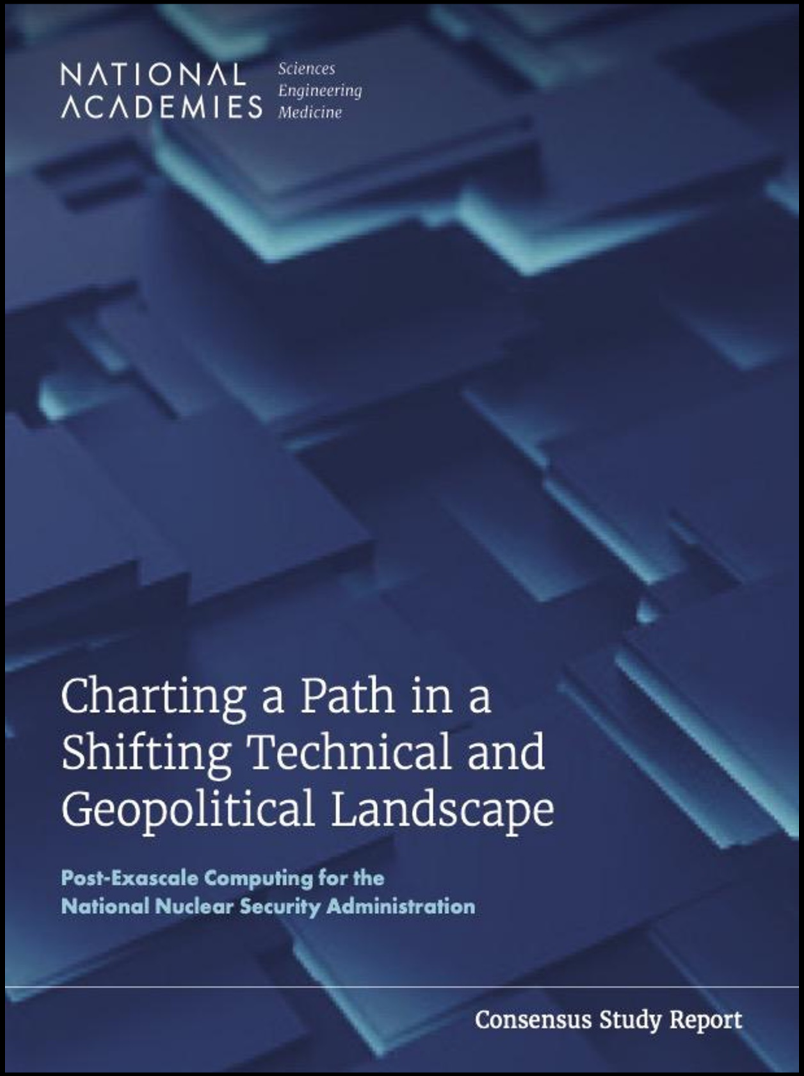
for AI in Science

andSimulation ?

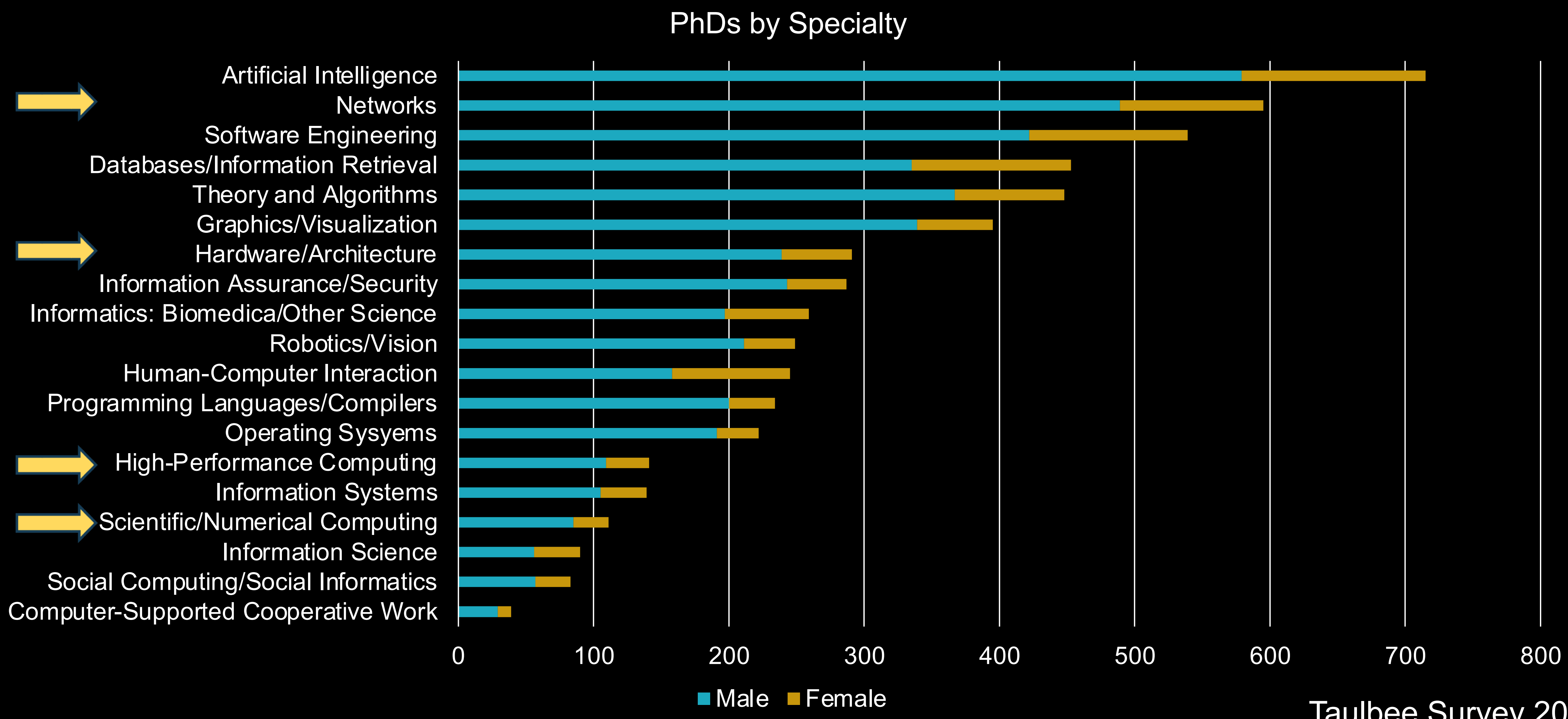
Workforce

Finding: The NNSA laboratories face significant challenges in recruiting and retaining the highly creative workforce that NNSA needs, owing to competition from industry, a shrinking talent pipeline, and **challenges in hiring diverse and international talent**.

- The U.S. national security enterprise has benefited enormously from **inclusion of global talent**.

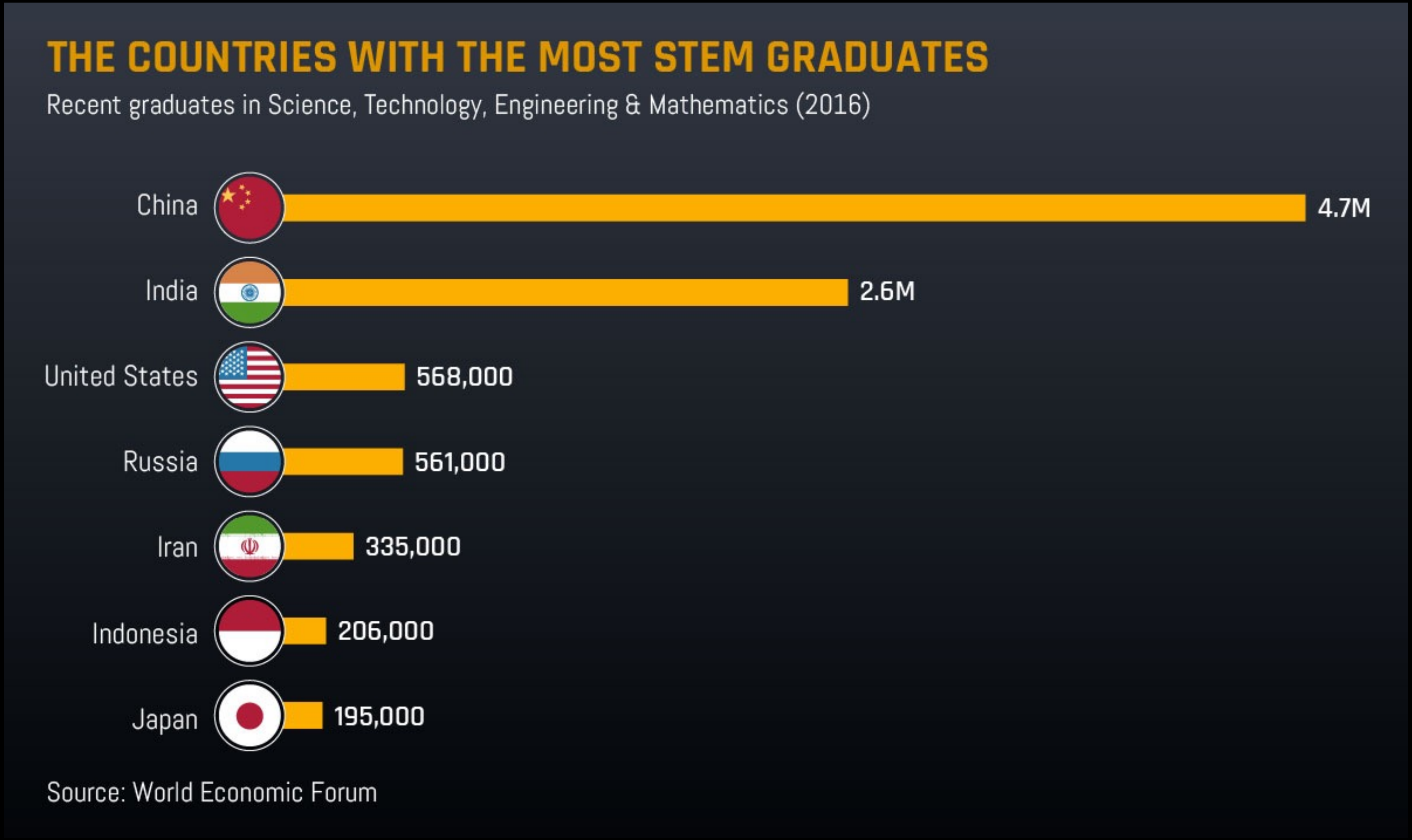


Where are the US Computer Science PhDs student doing?

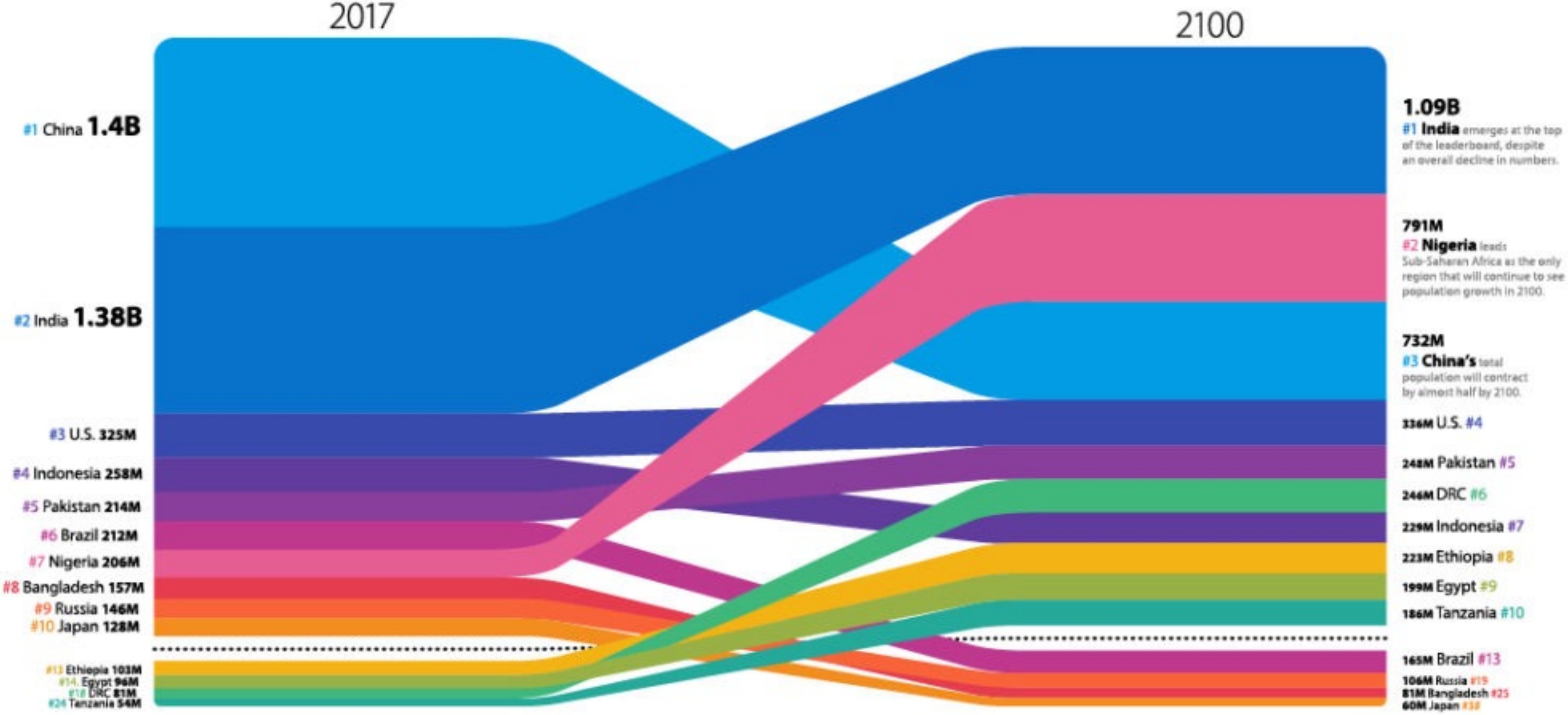


Taulbee Survey 2022

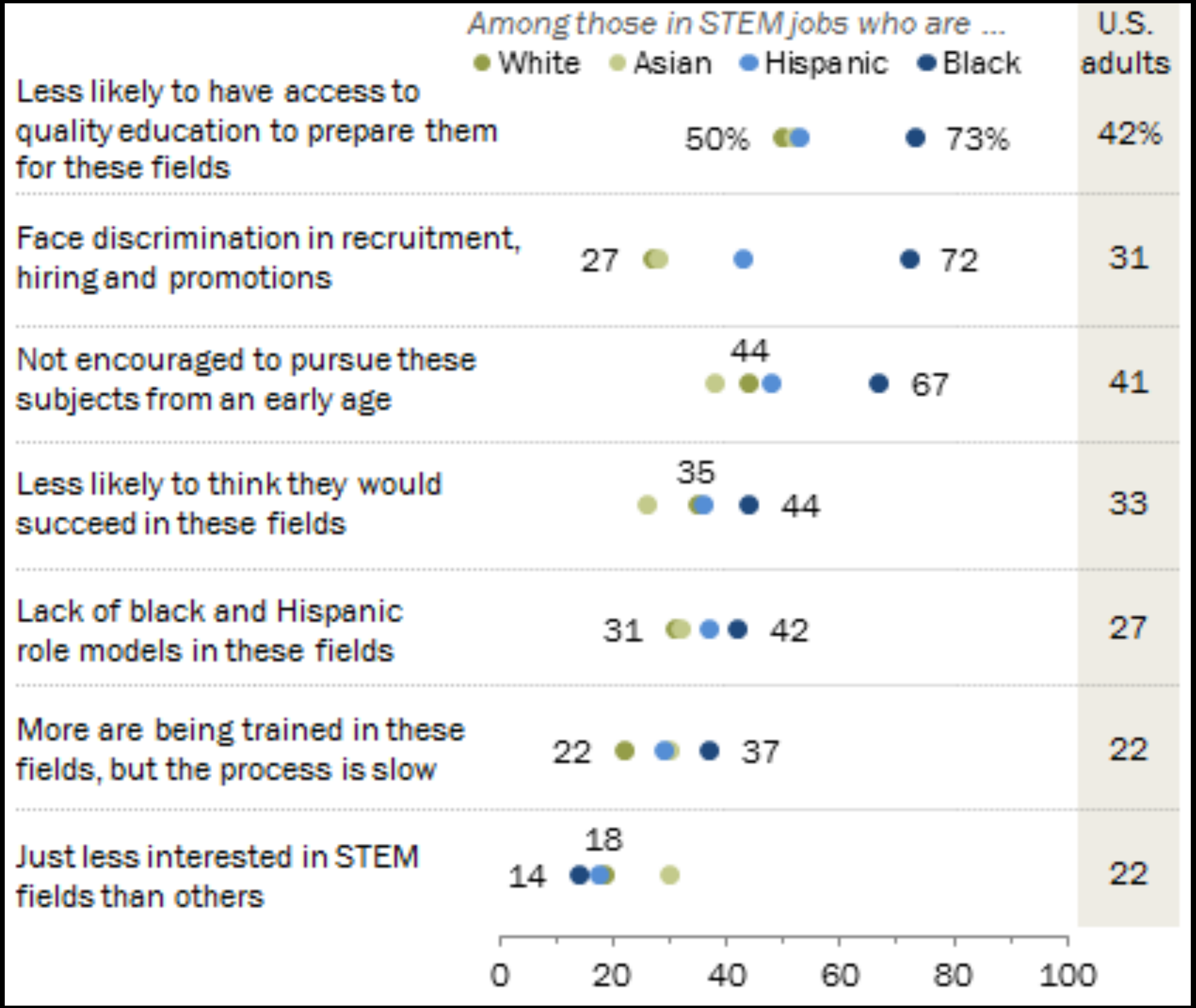
STEM Graduates Around the World



Top 10 Countries by Population



Reasons are Systemic



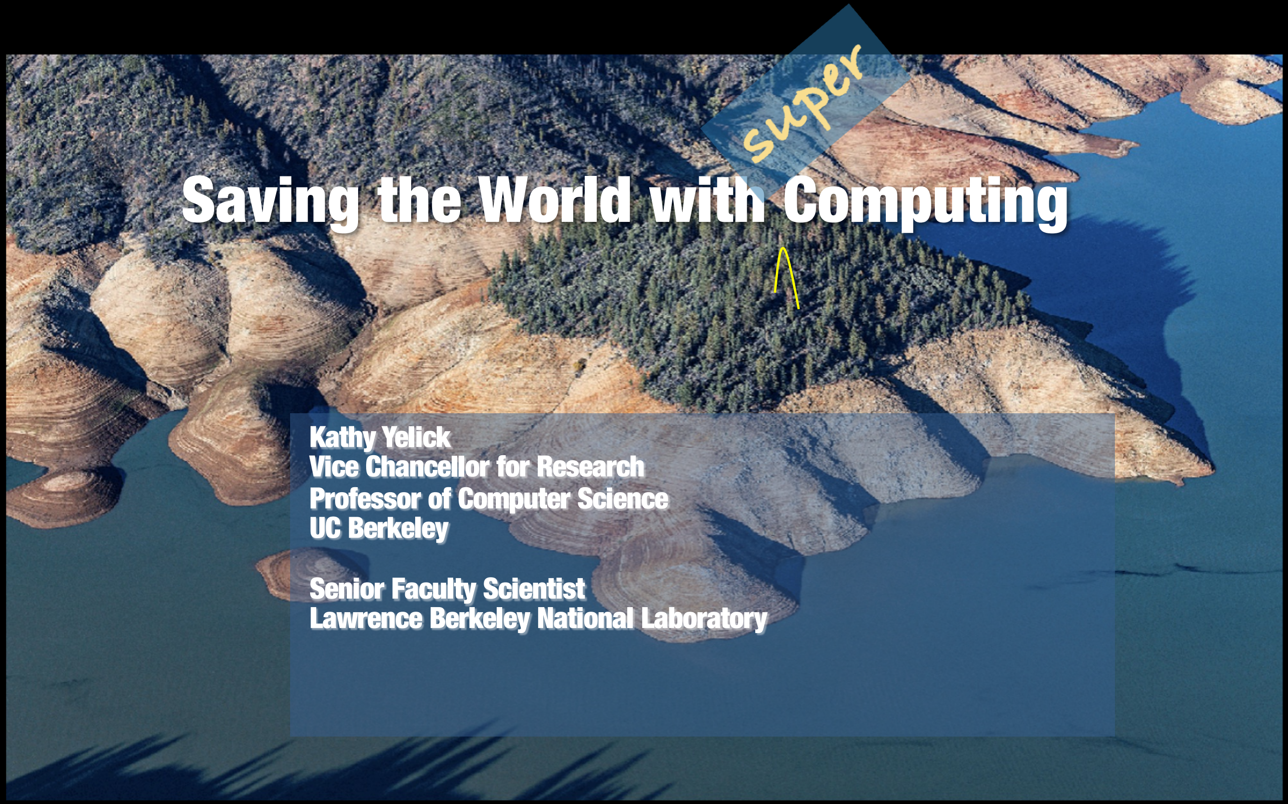
Wide racial and ethnic gaps among STEM employees on why so few blacks and Hispanics work in the field

% of those in science, technology, engineering and math jobs who say each of the following is a ***major reason*** *why blacks and Hispanics are underrepresented in STEM jobs in this country*

Note: Whites, blacks and Asians are non-Hispanic only; Hispanics are of any race. Respondents who gave other responses or who did not give an answer are not shown.
Source: Survey of U.S. adults conducted July 11-Aug. 10, 2017.
"Women and Men in STEM Often at Odds Over Workplace Equity"

PEW RESEARCH CENTER

Using Scientific Computing (Broadly) to Attract and Retain Talent



Often over 50% women with (relatively) high representation of other historically underserved groups

Post Exascale Computing: Not Business at Usual

- **Computing demands** continue to grow
- The benefits of more **weak scaling** are limited
- **Computing technology** has hit several “walls”
- The **computing industry** has changed dramatically
- **AI methods** are having huge impacts elsewhere
- **Quantum computing** potential for science still unknown
- **Cloud computing** is dominating the computing industry
- **Global supply chain** issues present uncertainties