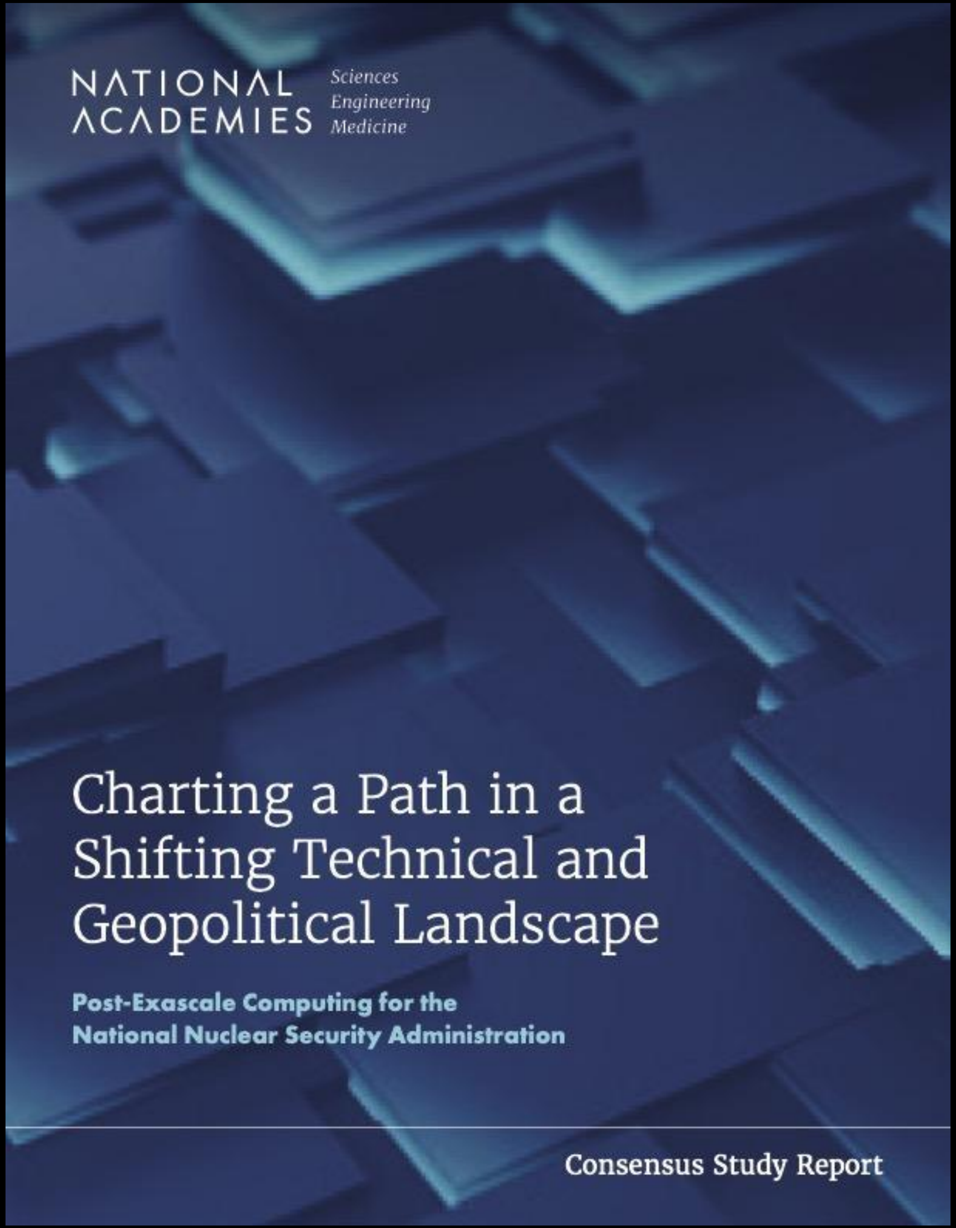
Beyond Exascale Computing

Kathy Yelick

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

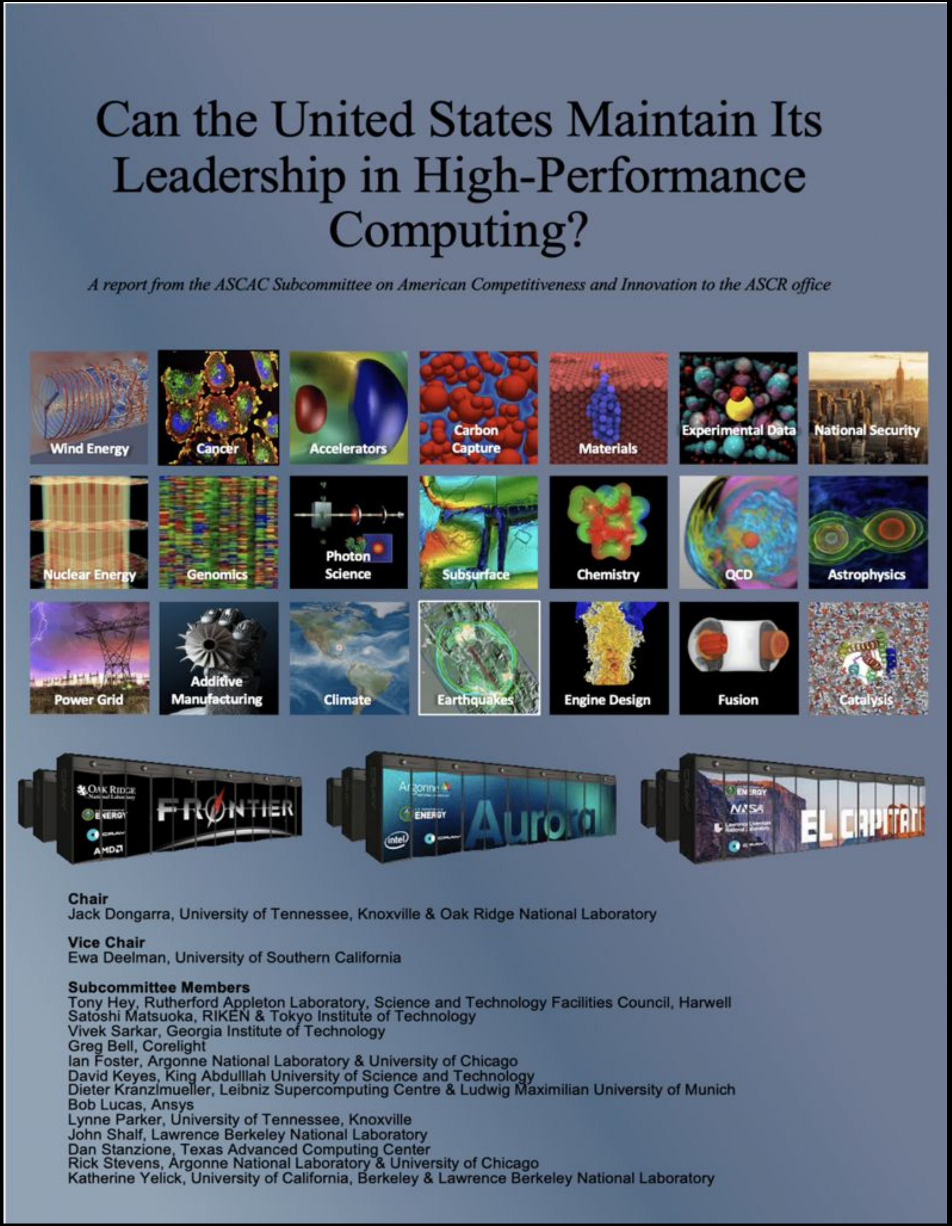
Senior Faculty Scientist
Computing Sciences
Lawrence Berkeley National Laboratory


NATIONAL ACADEMIES *Sciences Engineering Medicine*

Charting a Path in a Shifting Technical and Geopolitical Landscape

Post-Exascale Computing for the National Nuclear Security Administration

Consensus Study Report


Can the United States Maintain Its Leadership in High-Performance Computing?

A report from the ASCAC Subcommittee on American Competitiveness and Innovation to the ASCR office

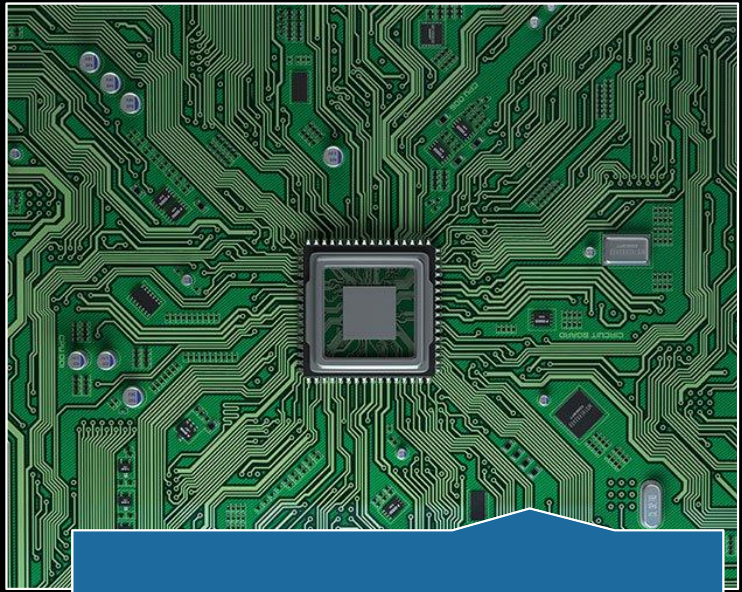
Chair
Jack Dongarra, University of Tennessee, Knoxville & Oak Ridge National Laboratory

Vice Chair
Ewa Deelman, University of Southern California

Subcommittee Members
Tony Hey, Rutherford Appleton Laboratory, Science and Technology Facilities Council, Harwell
Satoshi Matsuoka, RIKEN & Tokyo Institute of Technology
Vivek Sarkar, Georgia Institute of Technology
Greg Bell, Corelight
Ian Foster, Argonne National Laboratory & University of Chicago
David Keyes, King Abdullah University of Science and Technology
Dieter Kranzlmueller, Leibniz Supercomputing Centre & Ludwig Maximilian University of Munich
Bob Lucas, Ansys
Lynne Parker, University of Tennessee, Knoxville
John Shalf, Lawrence Berkeley National Laboratory
Dan Stanzione, Texas Advanced Computing Center
Rick Stevens, Argonne National Laboratory & University of Chicago
Katherine Yelick, University of California, Berkeley & Lawrence Berkeley National Laboratory

Post-Exascale Computing

Computing demand



Disruptions

Technology

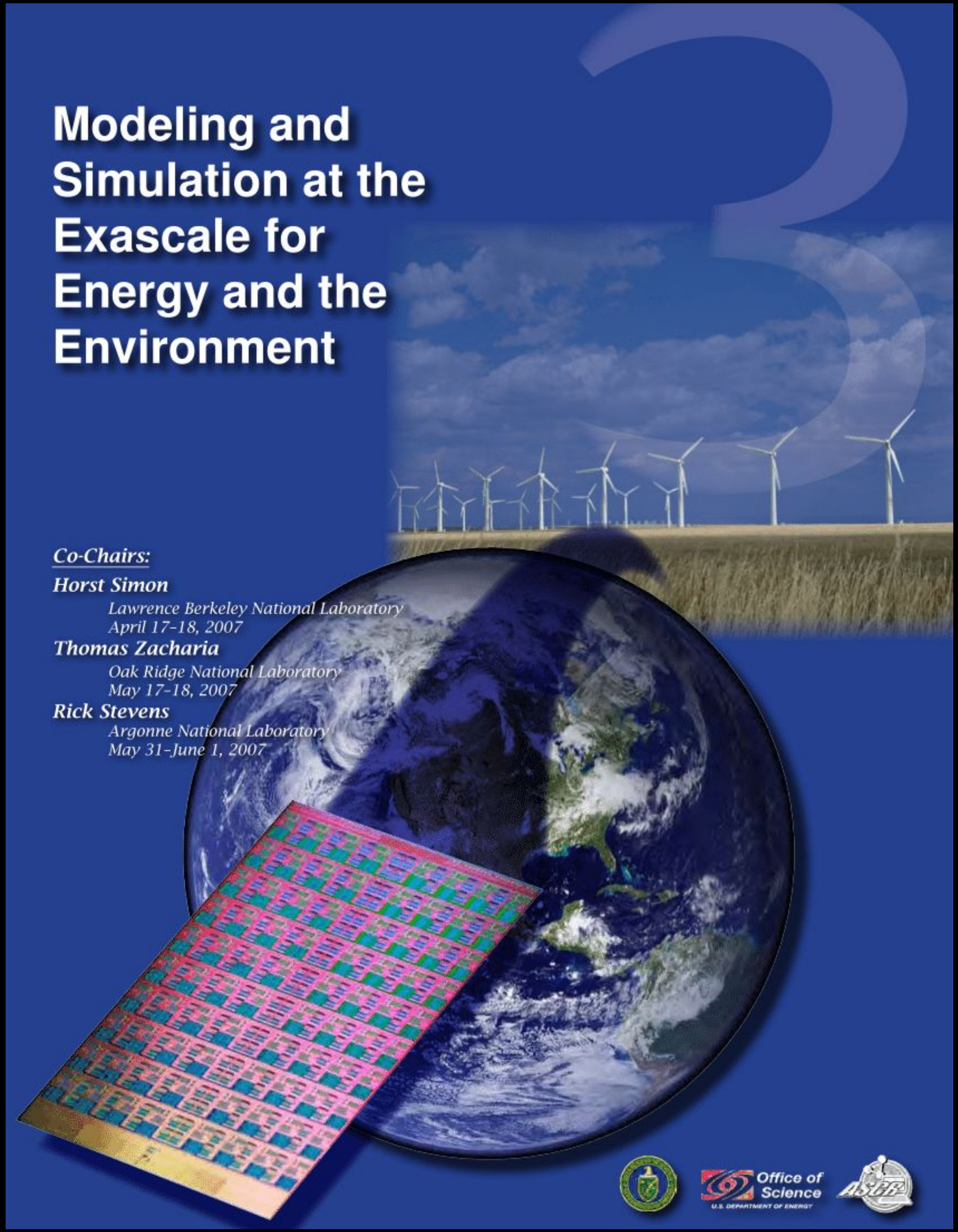
Continue to Rethink Applications

- 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!!

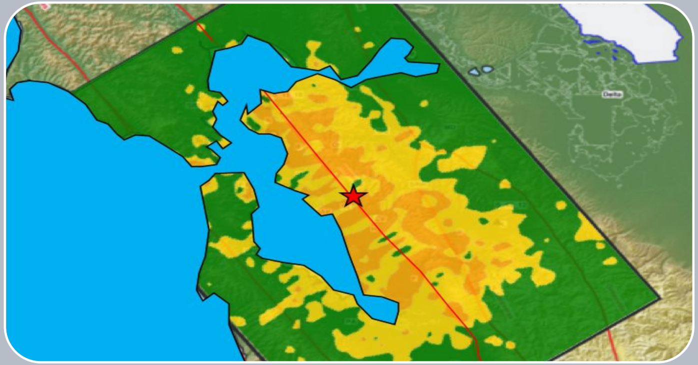
Scientific Computing Circa 2007

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

Simulation ≠ Scientific Computing ≠ HPC



New demands for HPC in Science



Simulation

From atoms to the universe

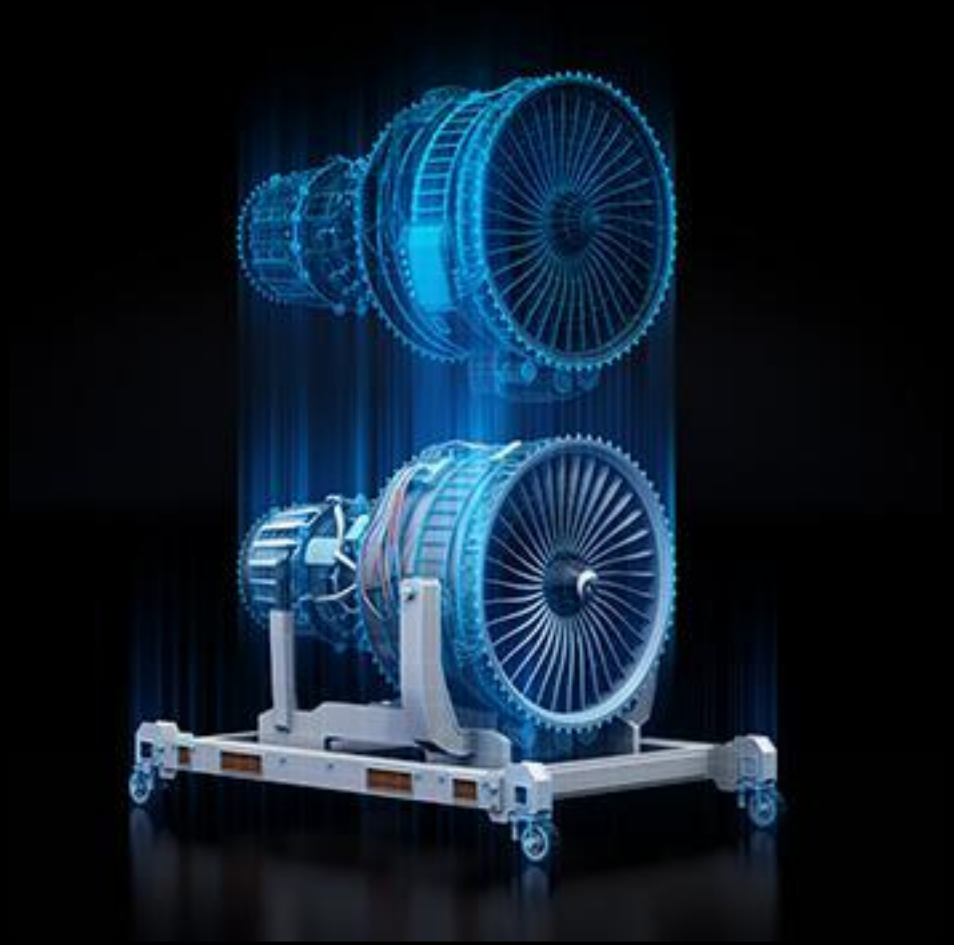
Data

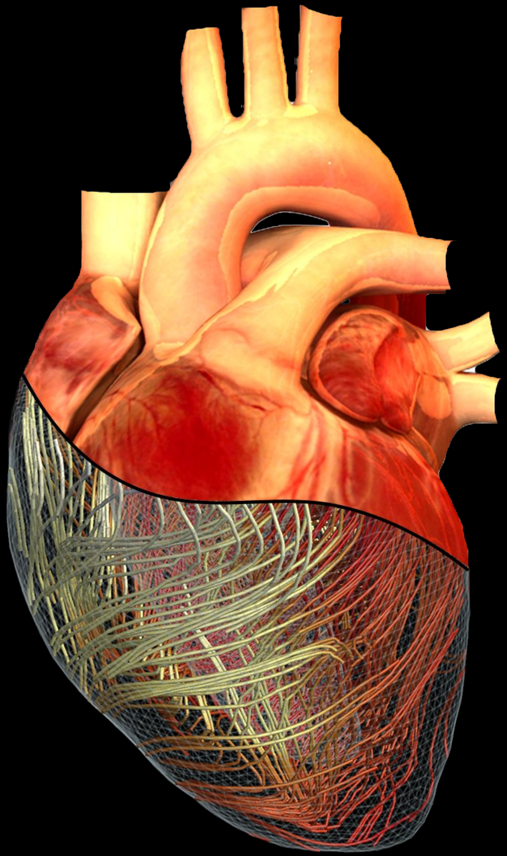
Images, text, to genomes

Learning

Interpret, infer and automate

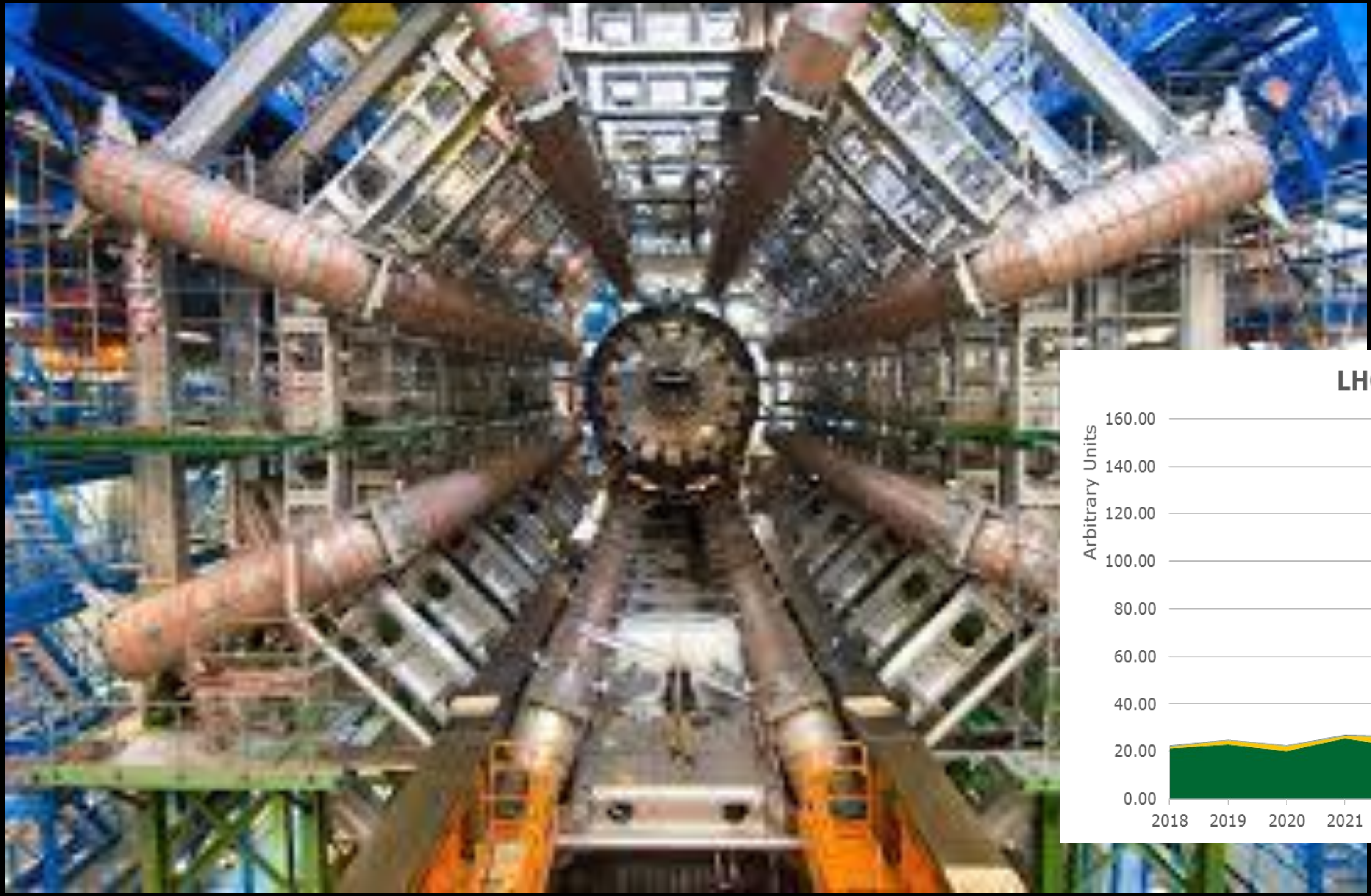
Digital Twins

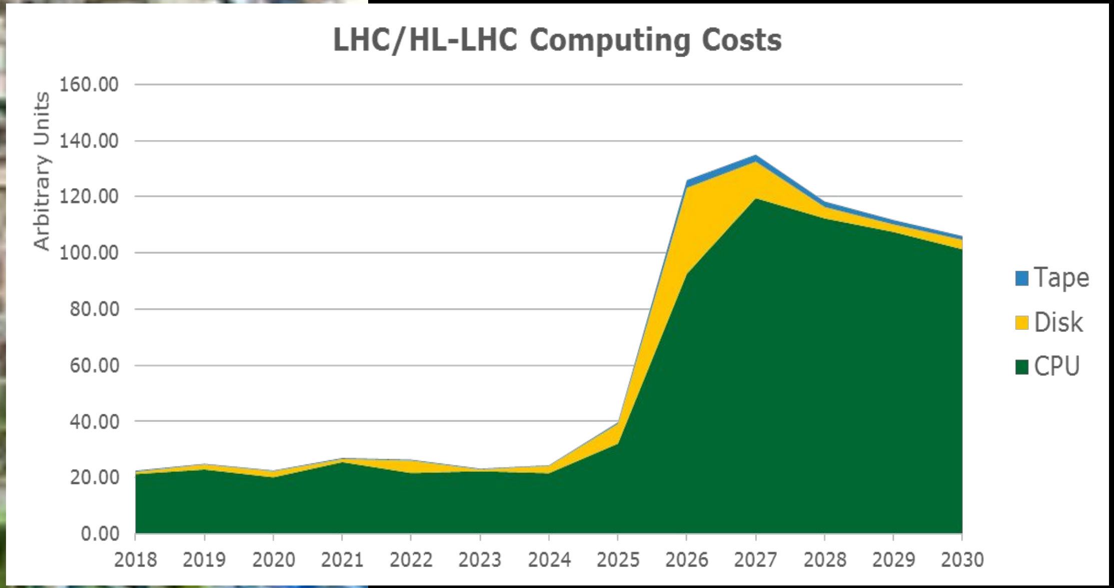




- Simulations
- Sensors / data
- Multi-level
- Real-time

Prediction of Atlas computing +$1B





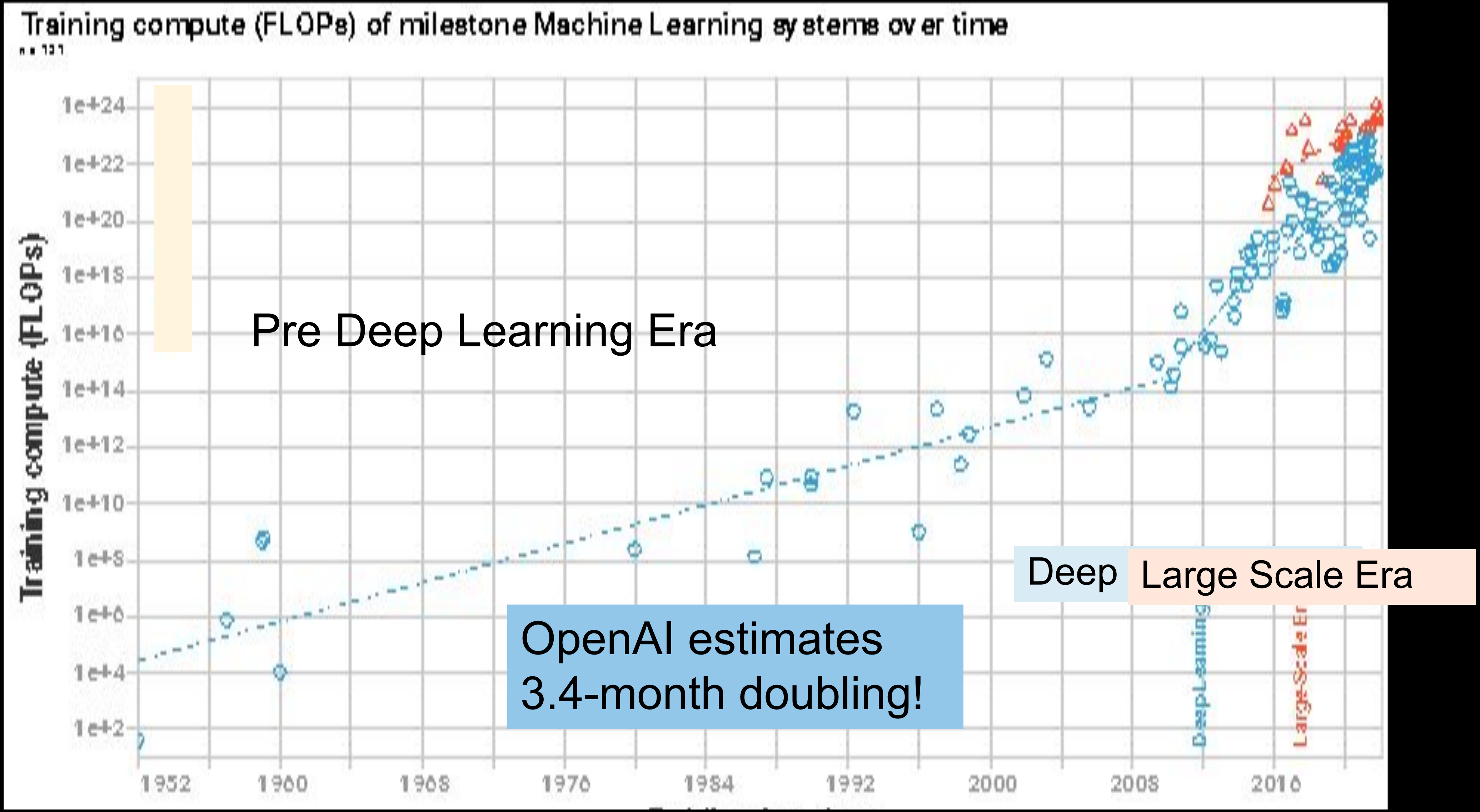
Microbial Data in the Environment

Tara Oceans Microbial data collected from 2009-13

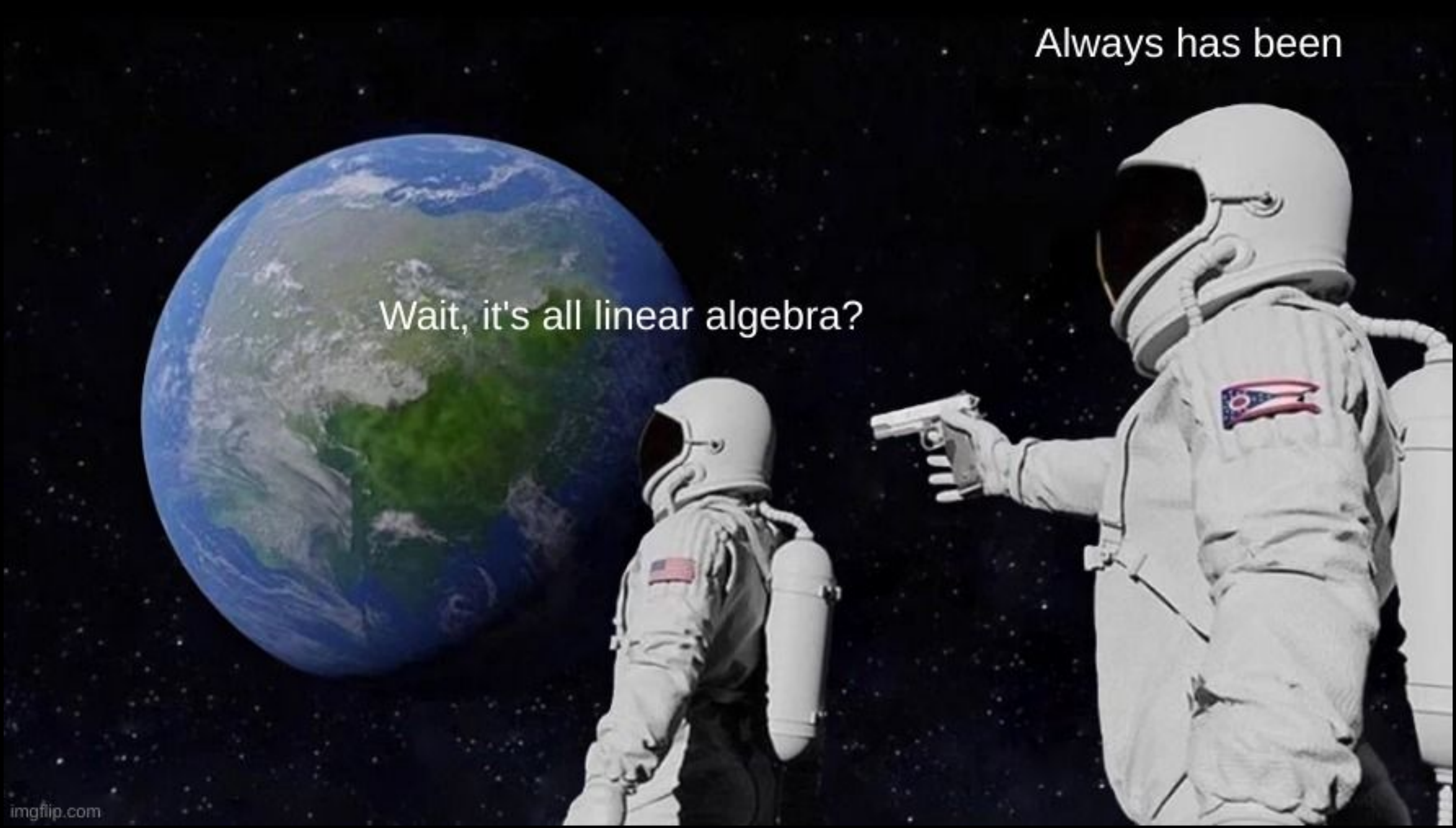
84 Terabytes assembled on 9000 Frontier nodes

HPC changes observational science

Machine Learning Drives Computational Demand



Is there parallelism?

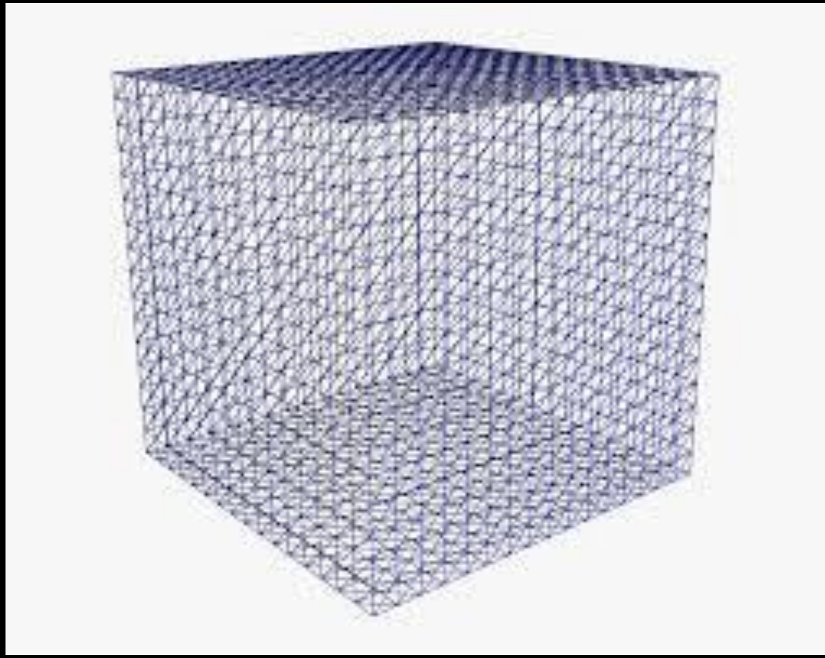

Always has been
Wait, it's all linear algebra?
imgflip.com

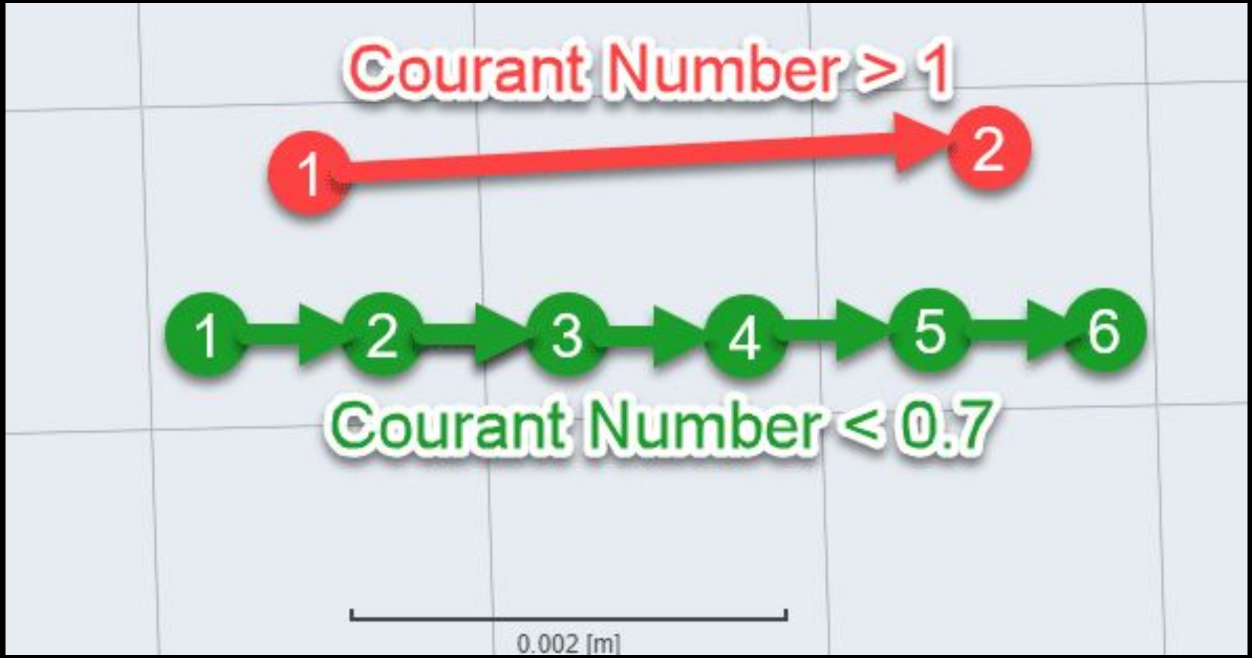
Analytics vs. Simulation Kernels:

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





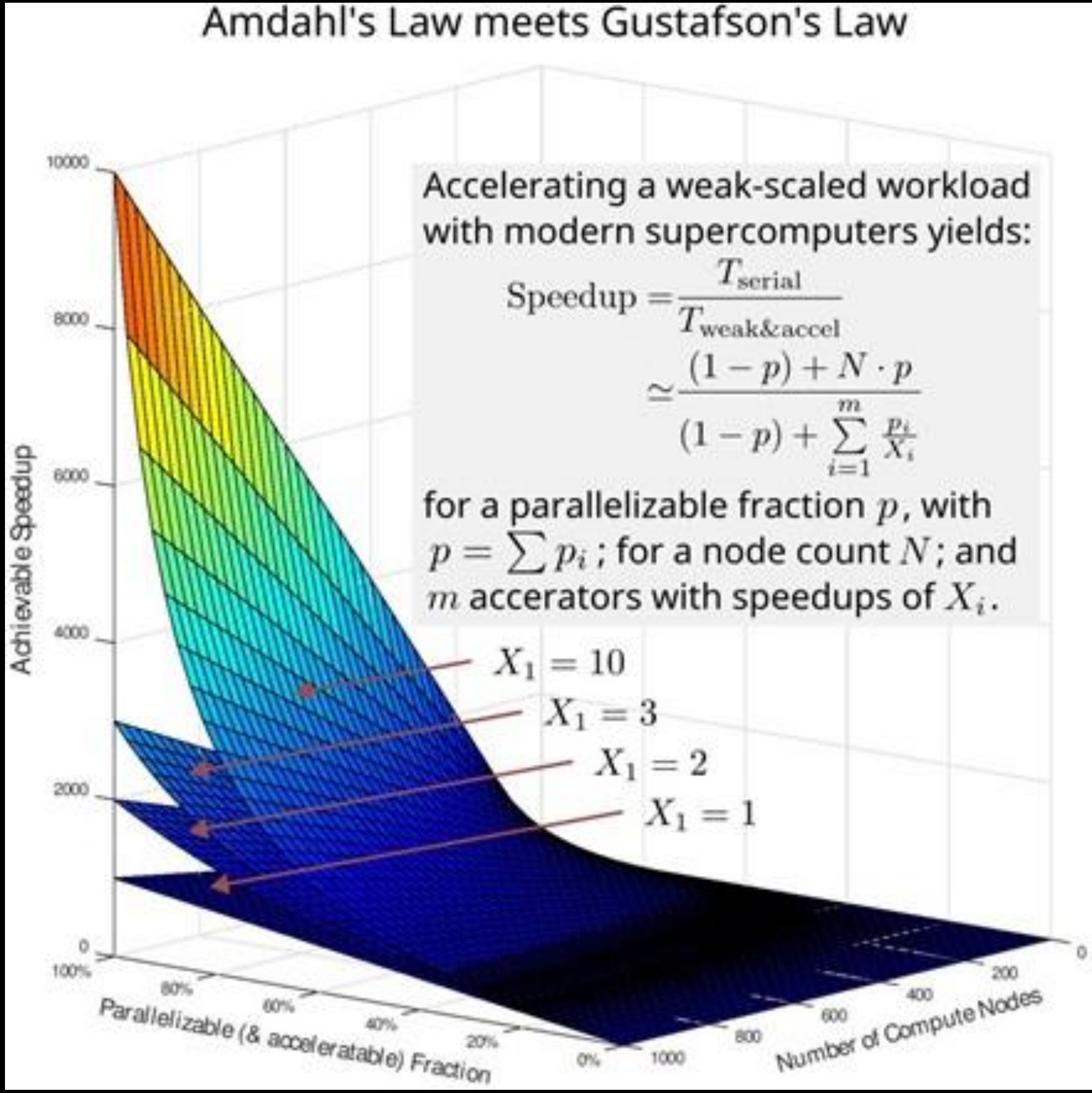
Increase resolution by 10x in each dimension

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

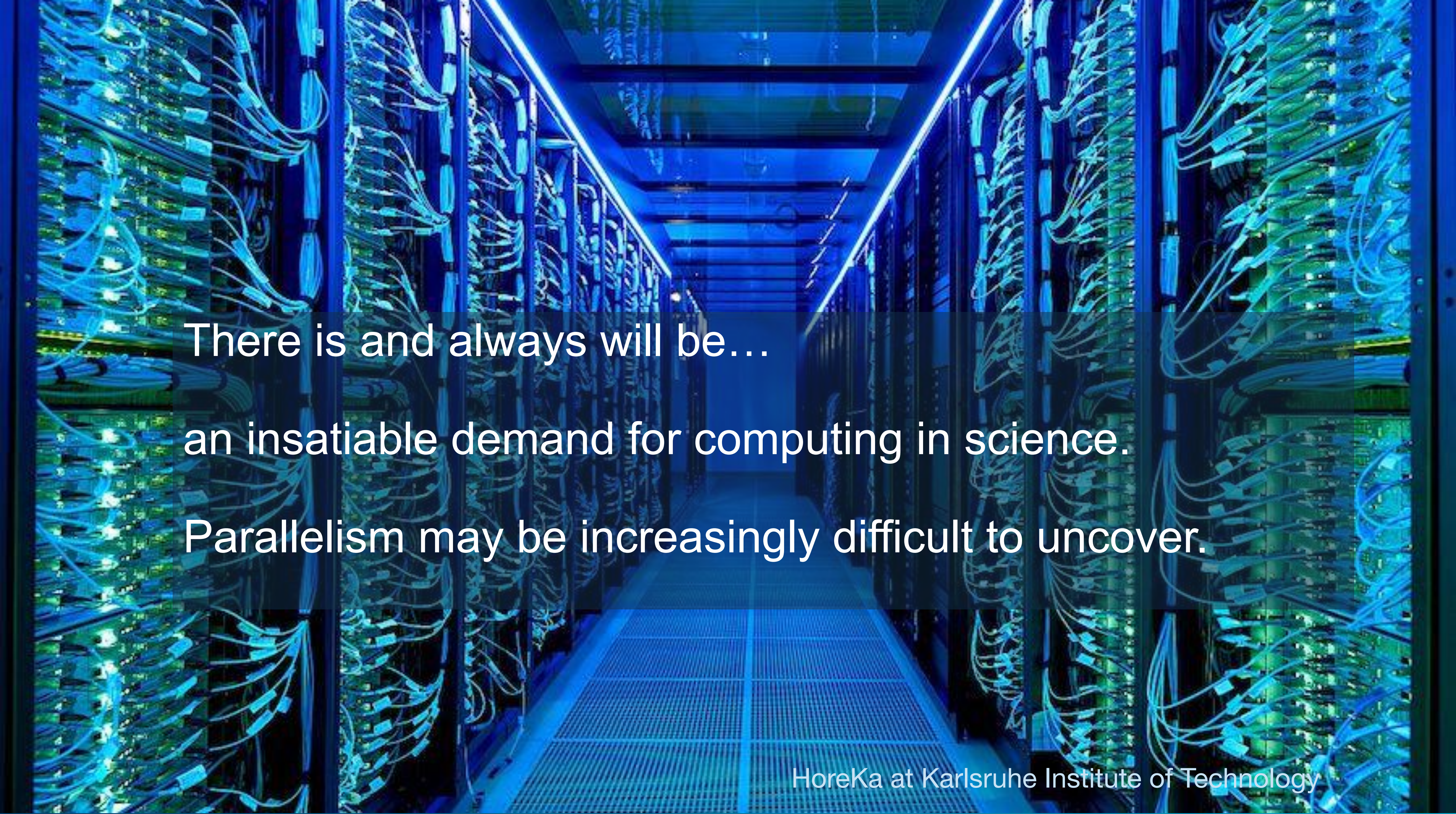


See SIAM News, 9/22 Satoshi Matsuoka and Jens Domke

There is and always will be…

an insatiable demand for computing in science.

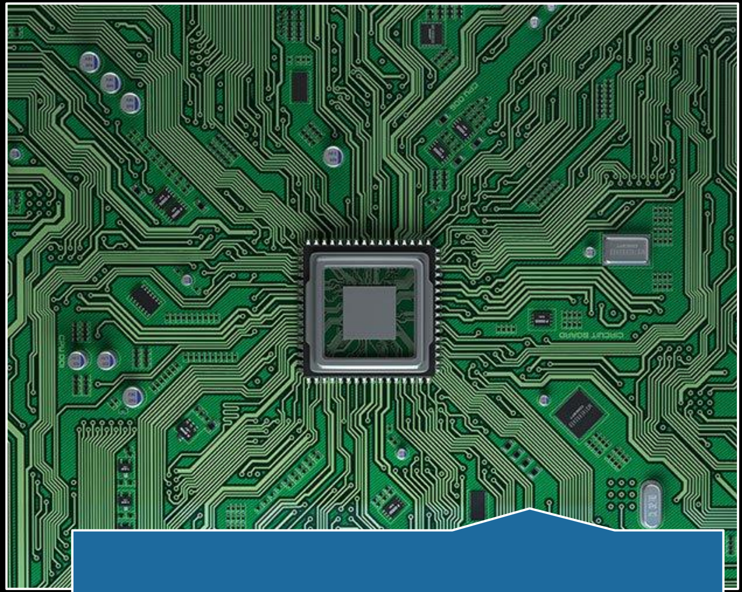
Parallelism may be increasingly difficult to uncover.



HoreKa at Karlsruhe Institute of Technology

Post-Exascale Computing

Computing demand



Disruptions

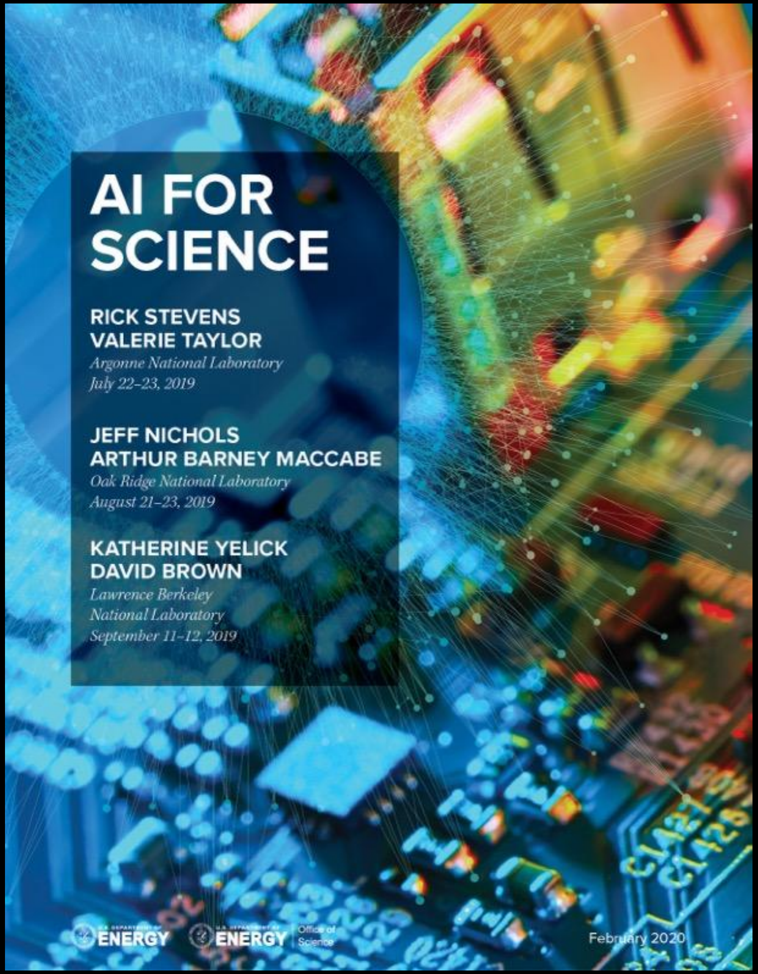
Technology

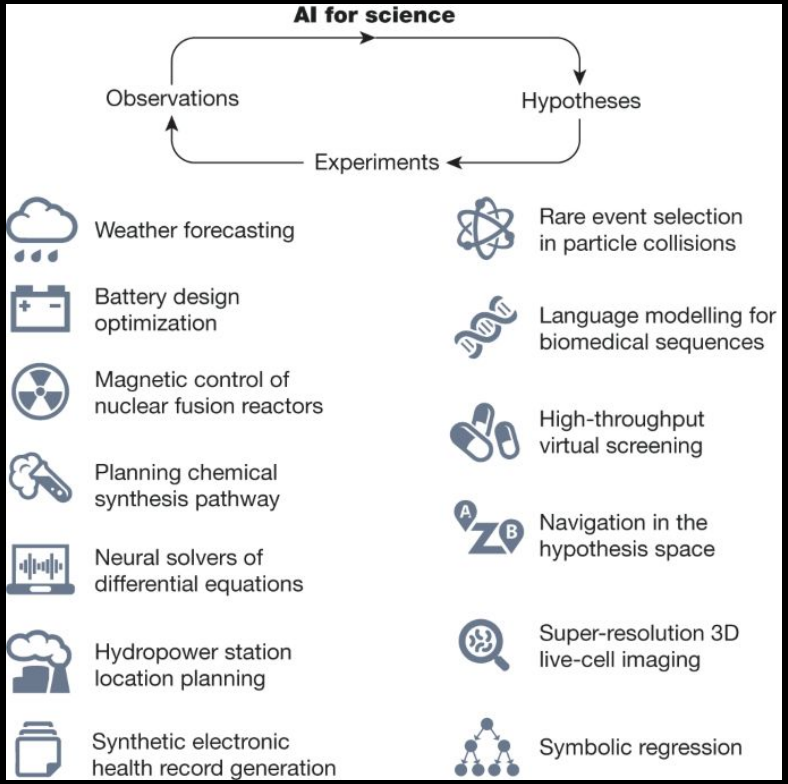
Disruptions

- AI
- Quantum
- Cloud

Implied question: Do these make HPC obsolete?

AI for Science

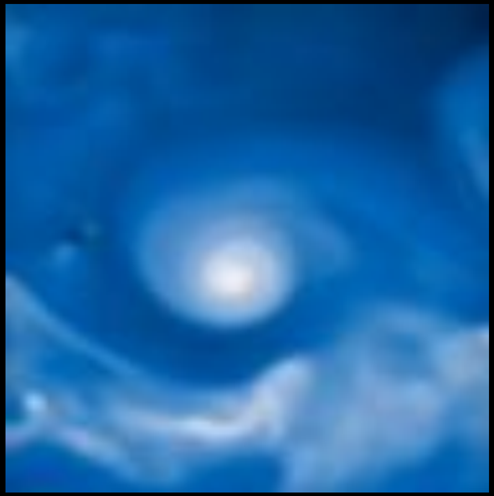




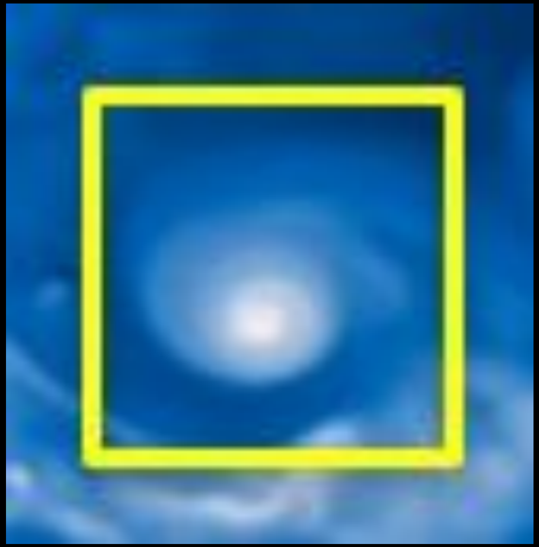
Scientific discovery in the age of artificial intelligence, 2023

Analyze Simulations to Find Hurricanes

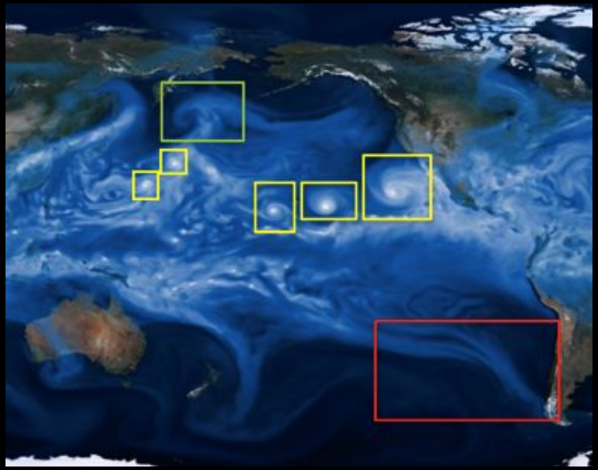
Classification



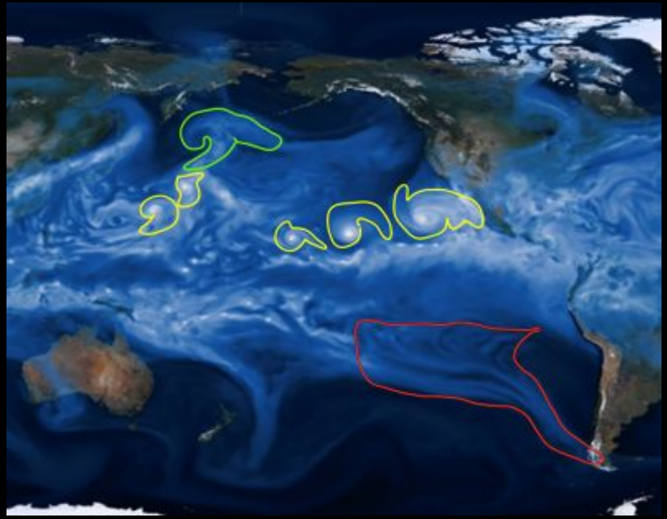
Localization



Detection



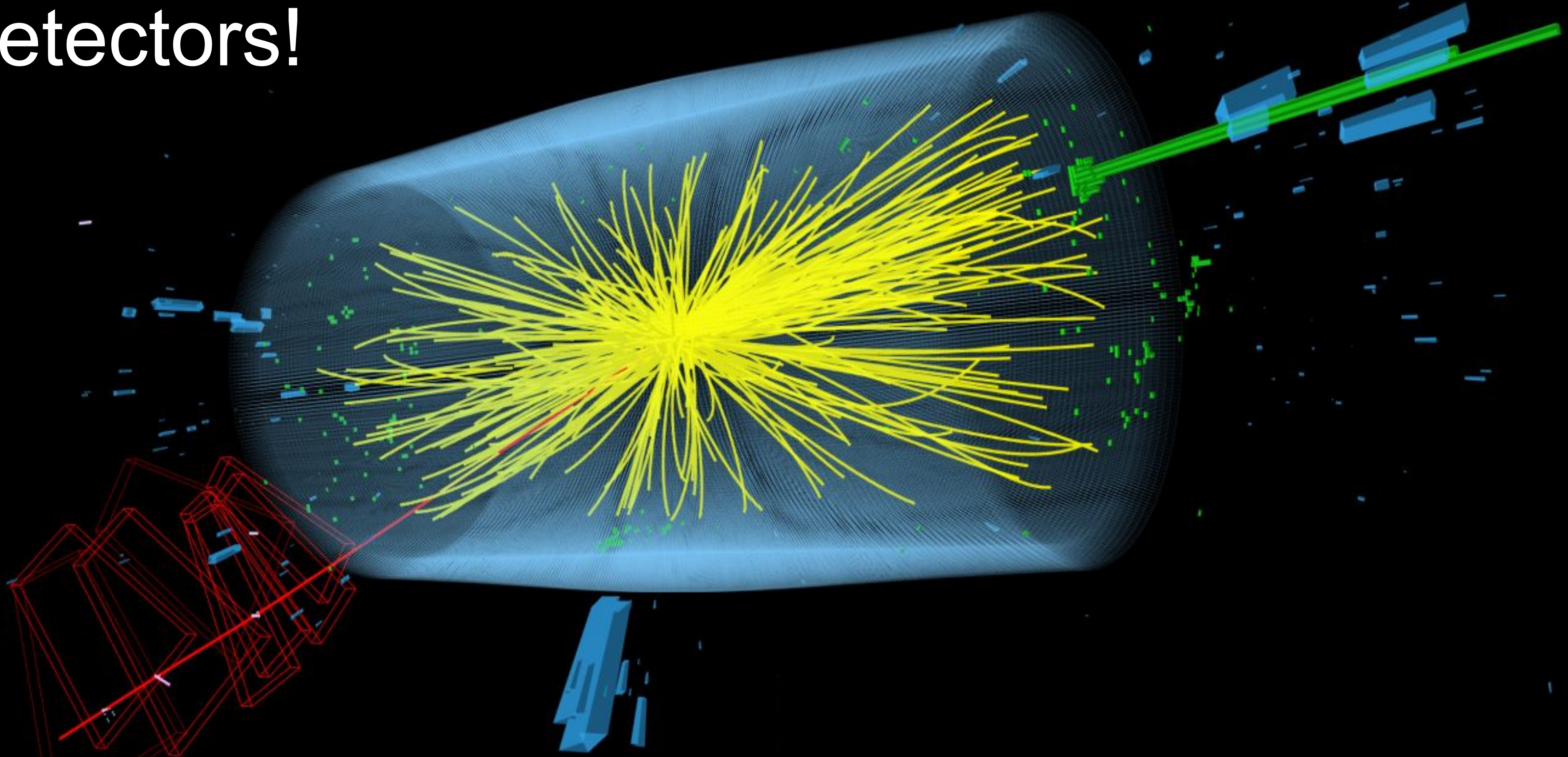
Segmentation



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

Source: Prabhat

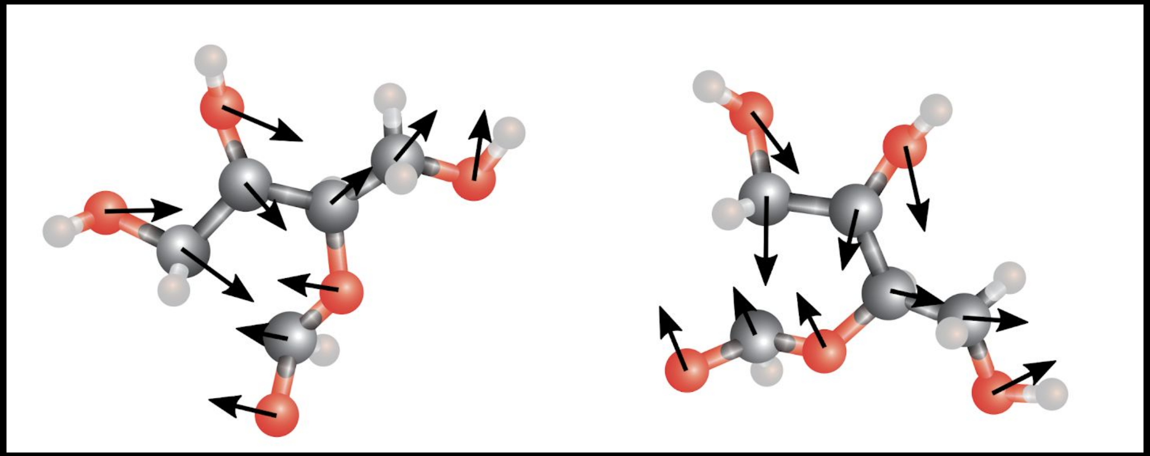
Precision: like adding 4,000 extra tons of detectors!

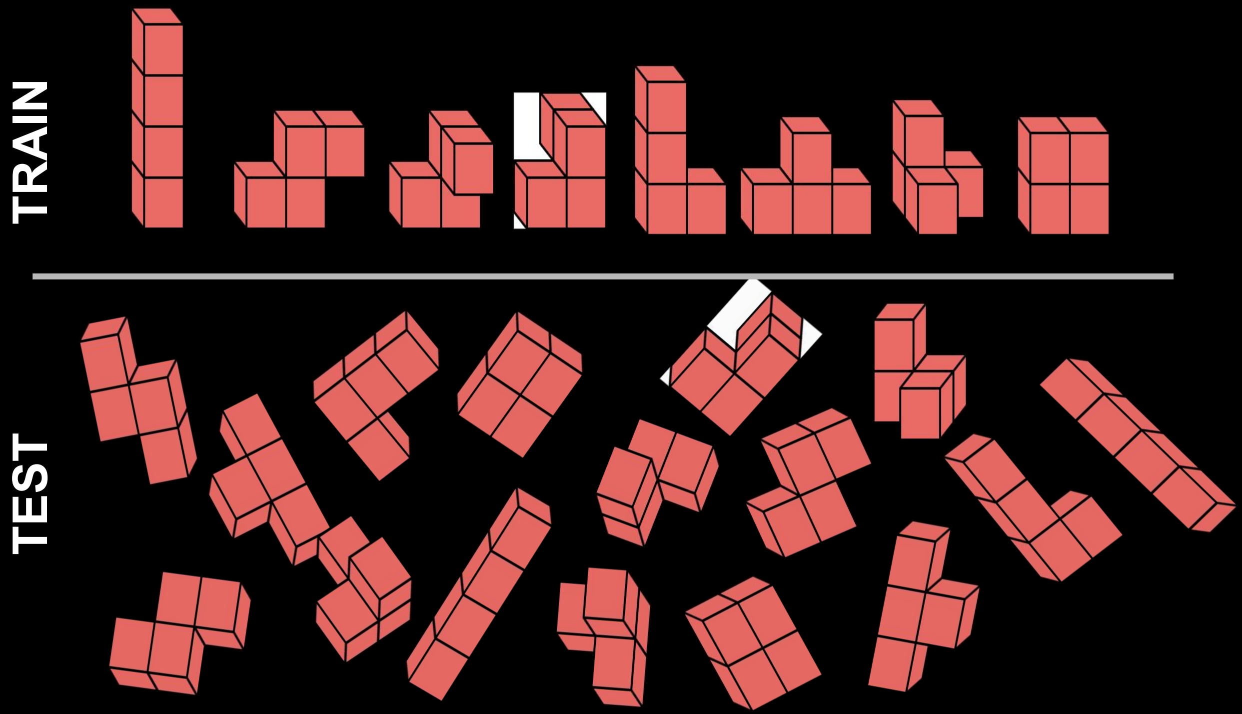


Based on 8/12/2016 slide by Joe Lykken at Fermilab

Design 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

A-Lab at LBNL

Five Stages of AI

Acceptance

Denial

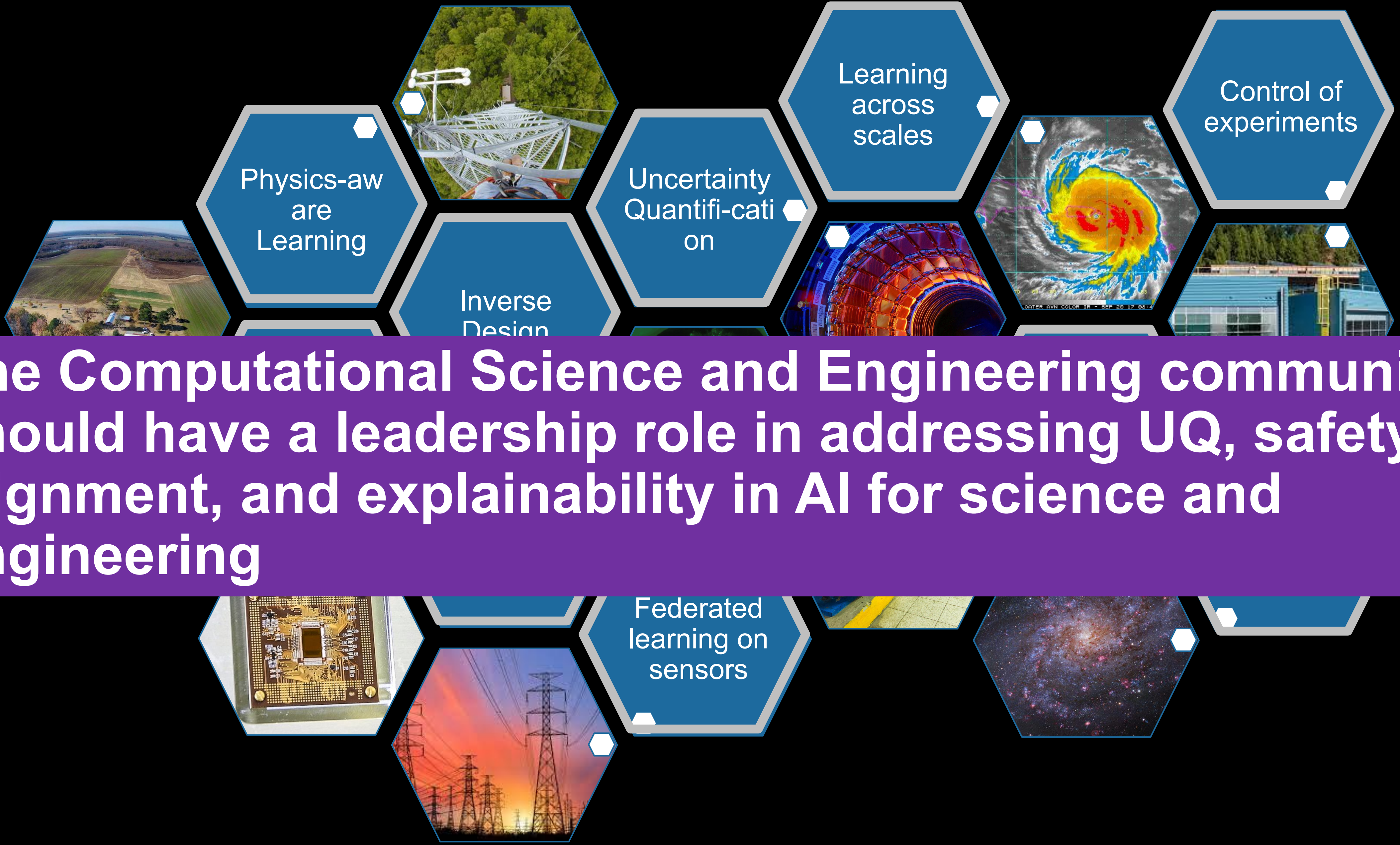
Anger

Depression

Bargaining

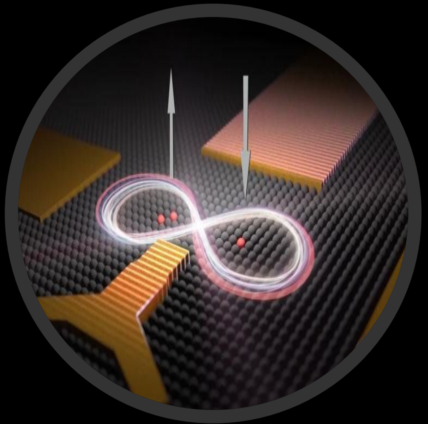
And this includes AI researchers!

AI in Science



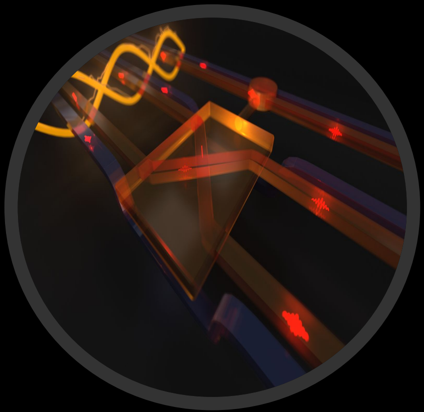
The Computational Science and Engineering community should have a leadership role in addressing UQ, safety, alignment, and explainability in AI for science and engineering

Exciting Progress … But we don't yet have the IC Transitor



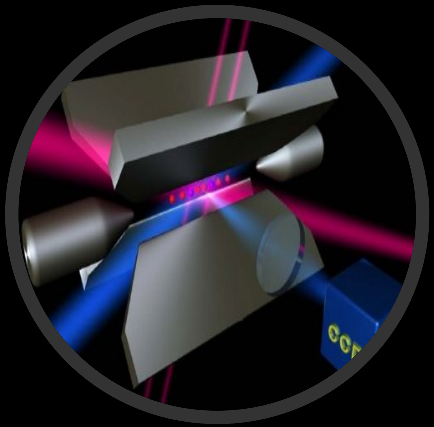
Dopants in Silicon / Diamond

www.sciencedaily.com



Photonic Circuits

www.phys.org

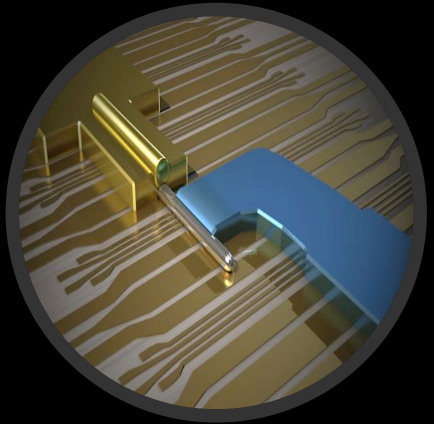


Trapped Ions

www.quantumoptics.at

Superconducting Circuits

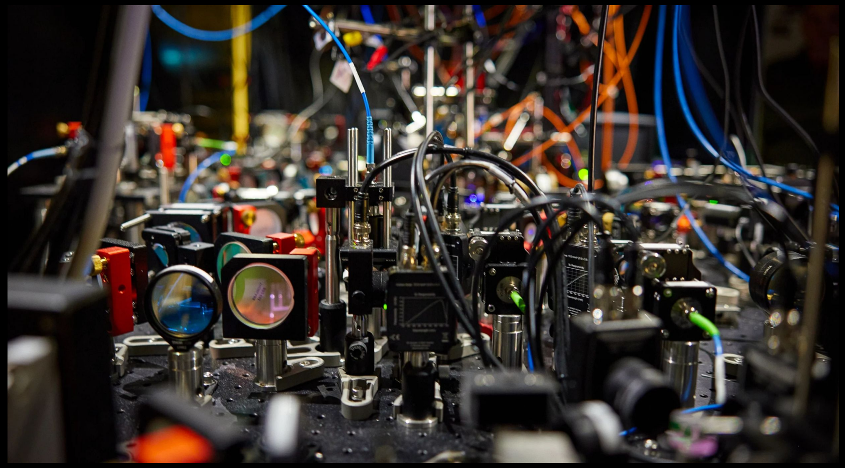
www.qnl.berkeley.edu



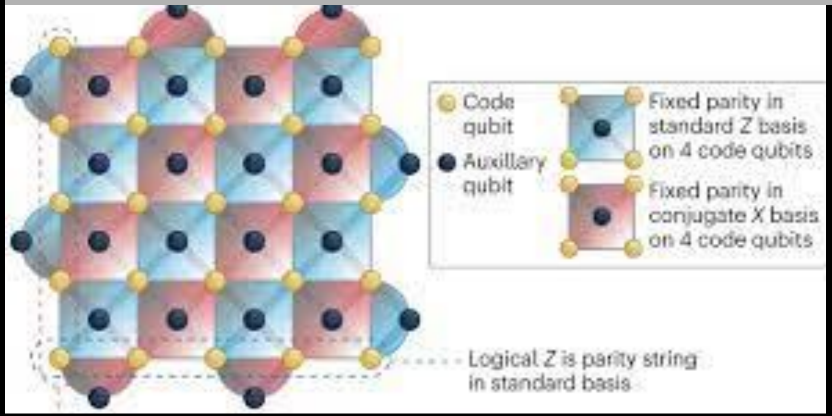
Topological Wires

www.microsoft.com

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



A series of fast-paced advances in Quantum Error Correction

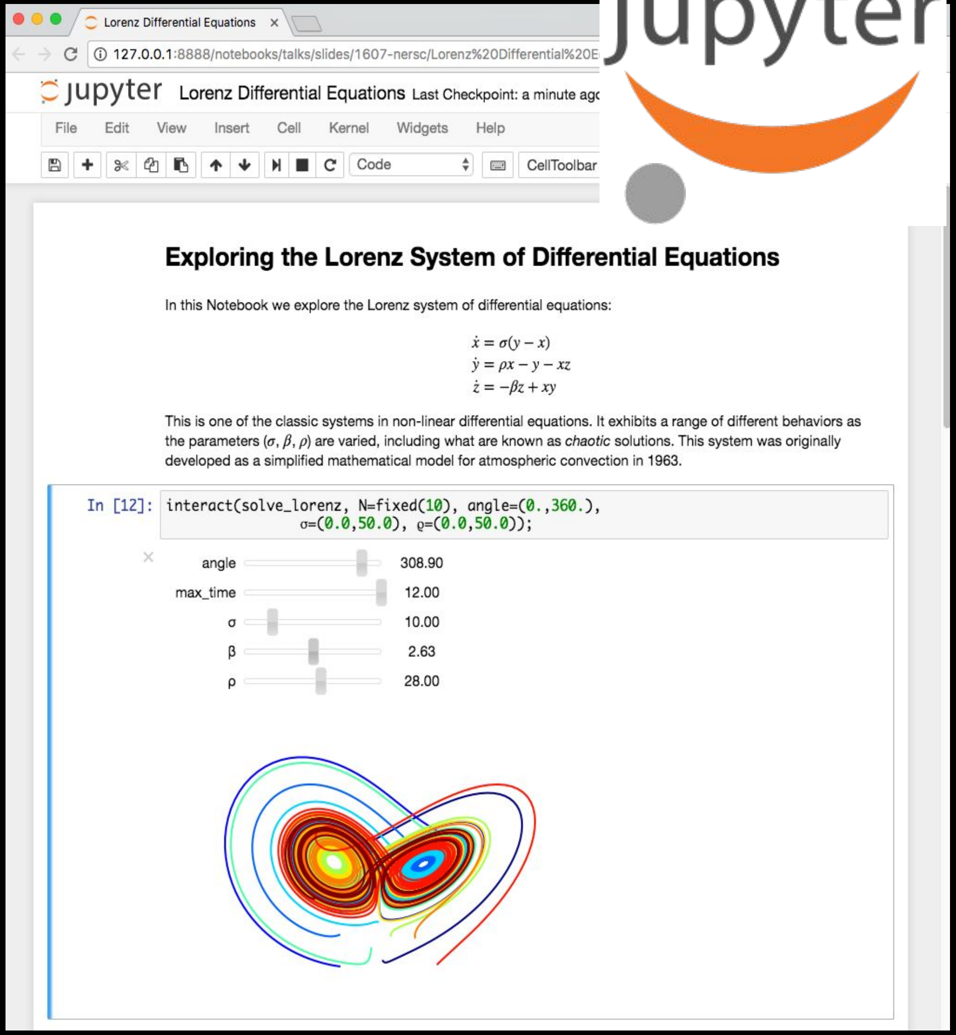


Cloud Computing

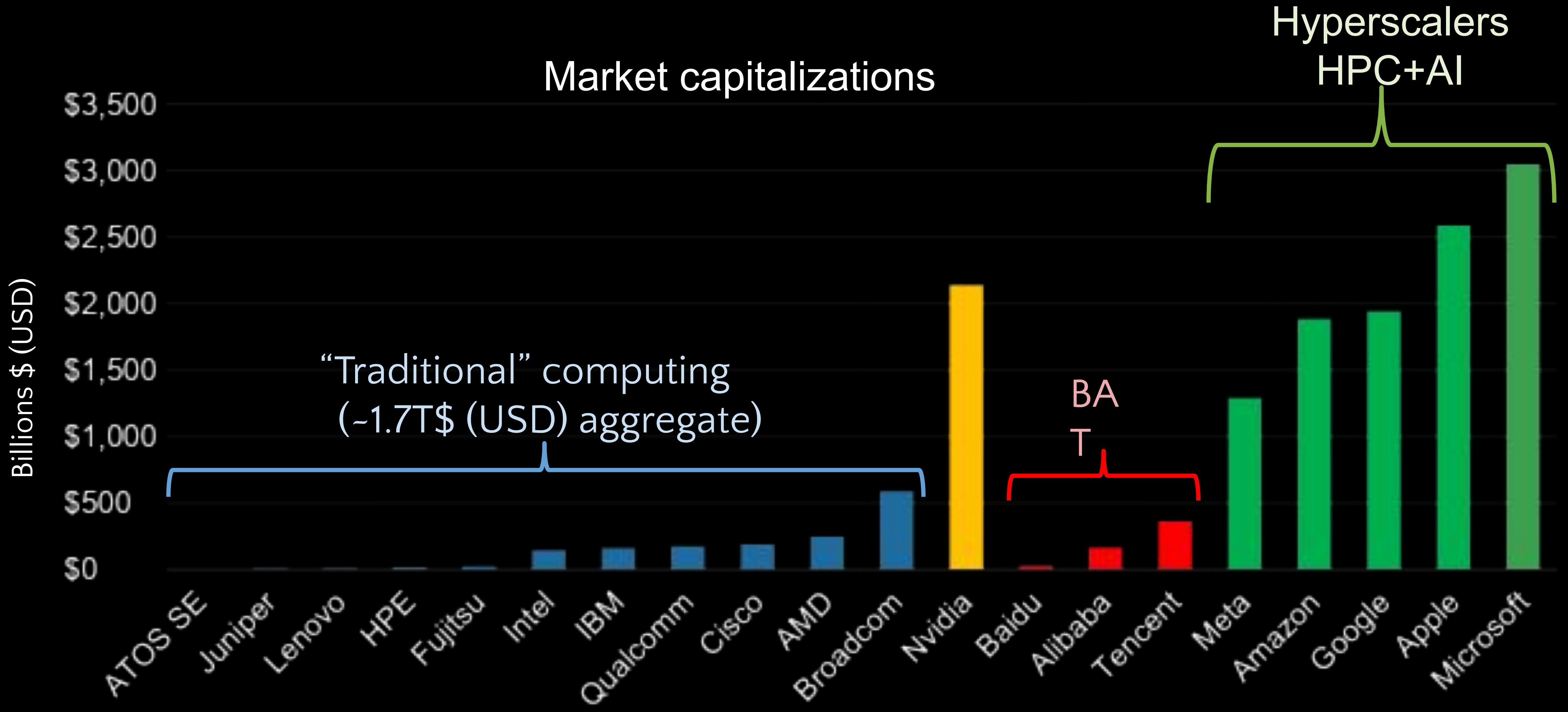
Lessons Learned from Clouds

- Cost vs price
- Availability and resilience
- Higher level programming

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



Follow the money, understand the implications

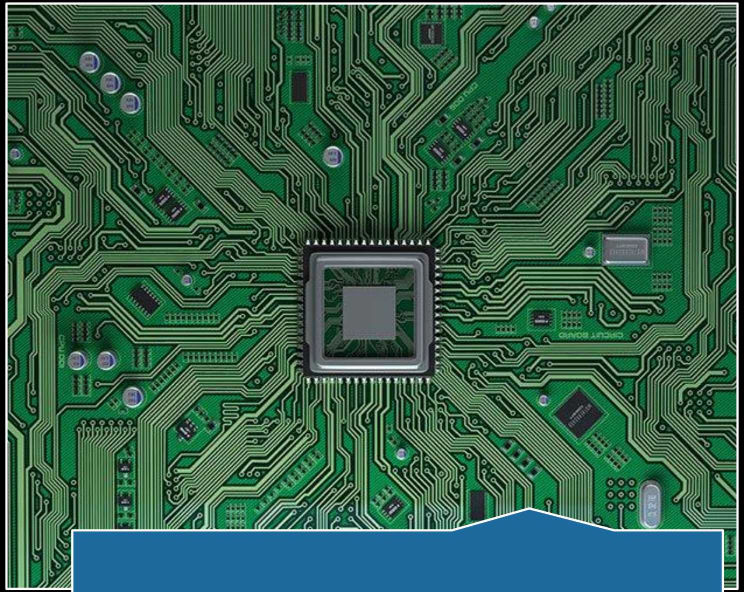


Source: Reed, Gannon, Dongarra

HPC community has always punched above its weight

Post-Exascale Computing

Computing demand

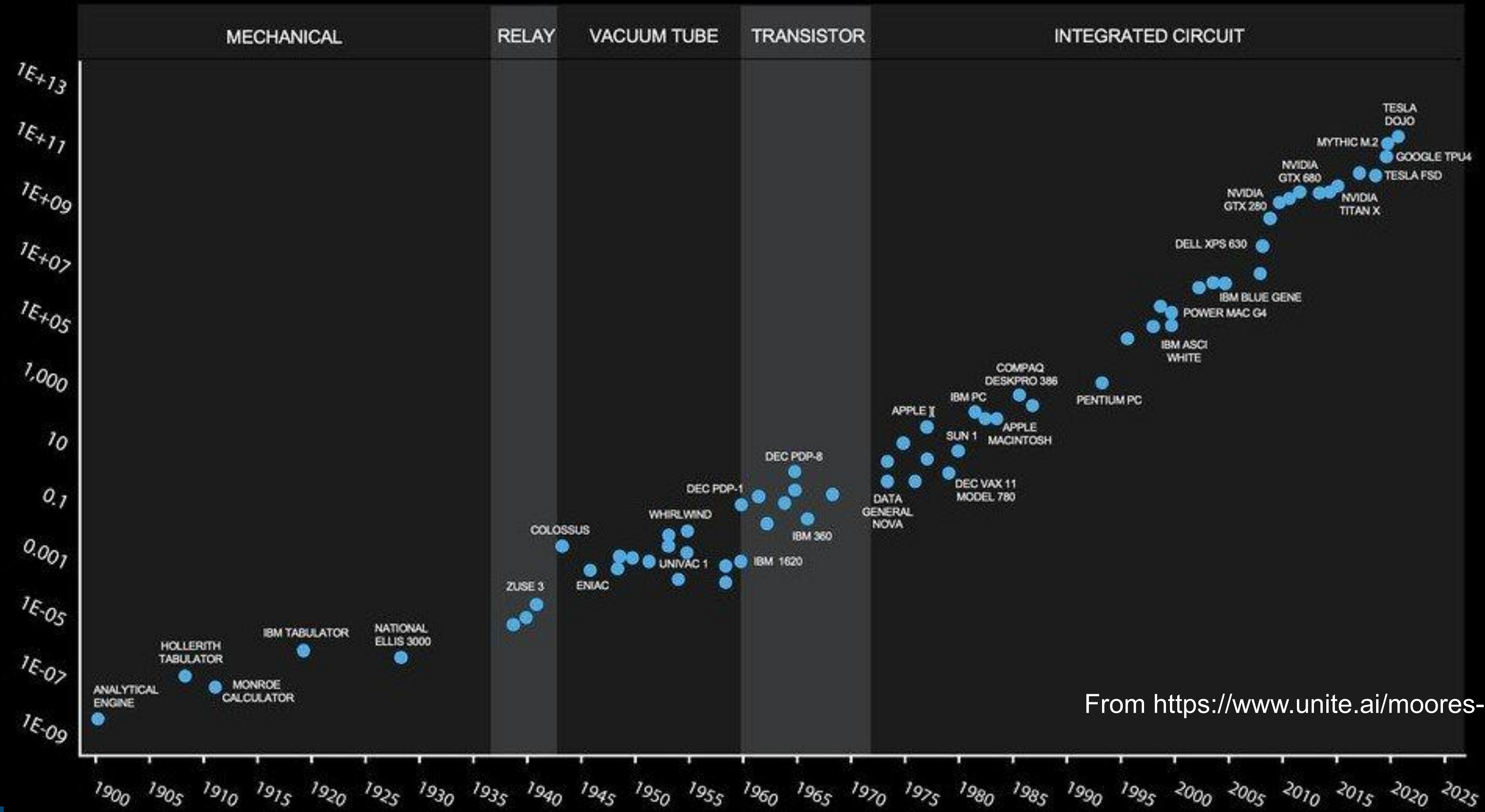


Disruptions

Technology

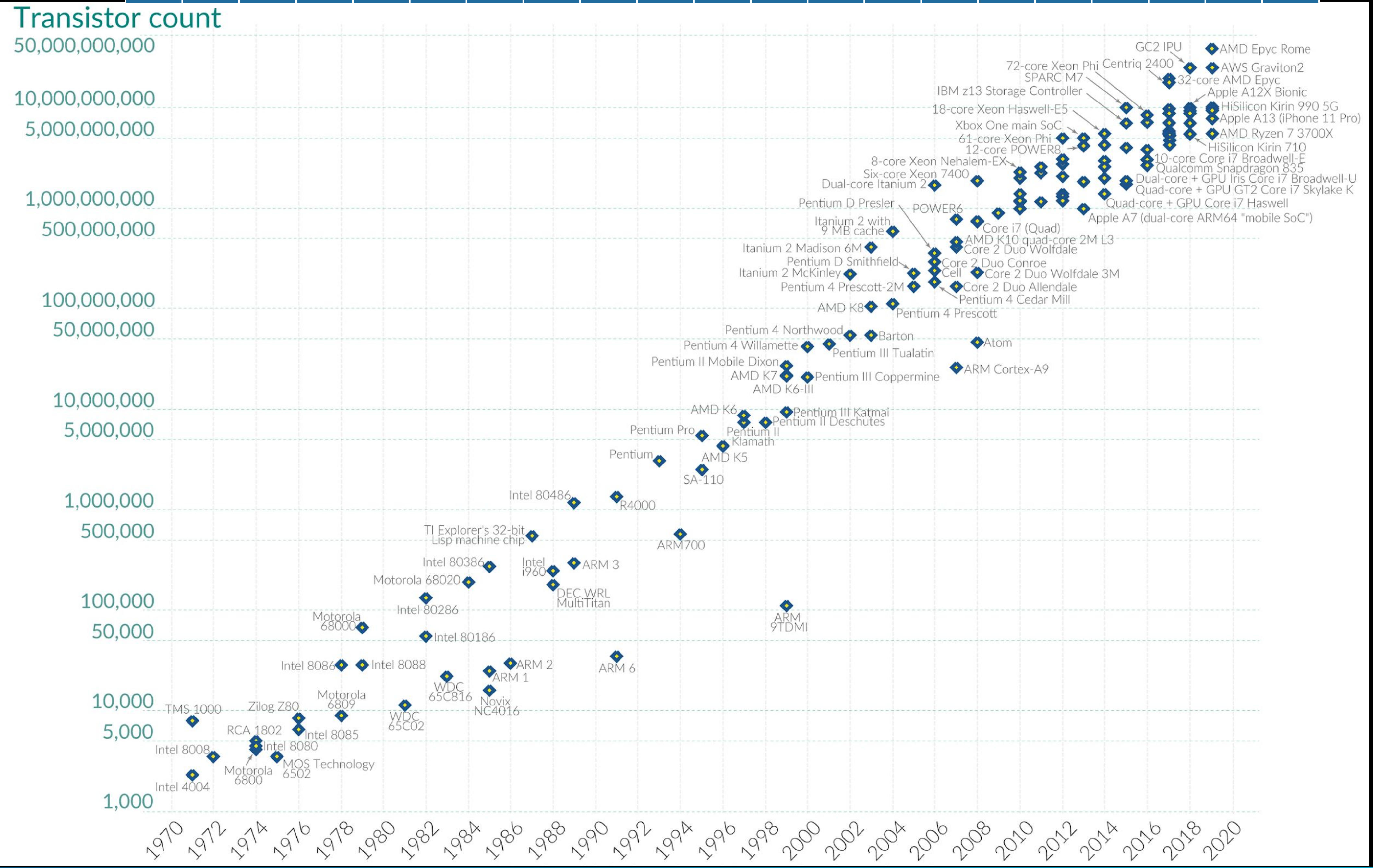
FUTURE VENTURES

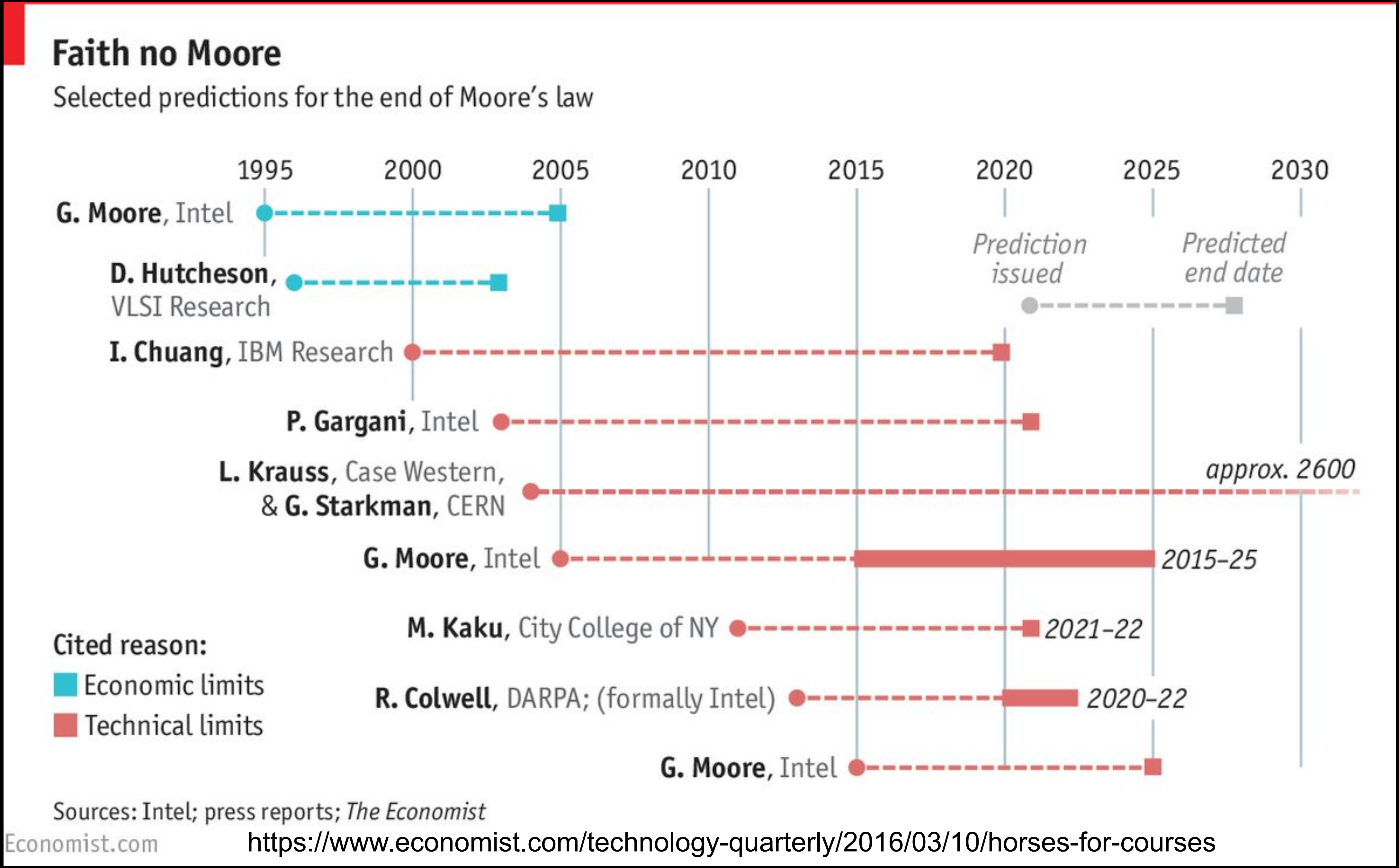
122 YEARS OF MOORE'S LAW



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

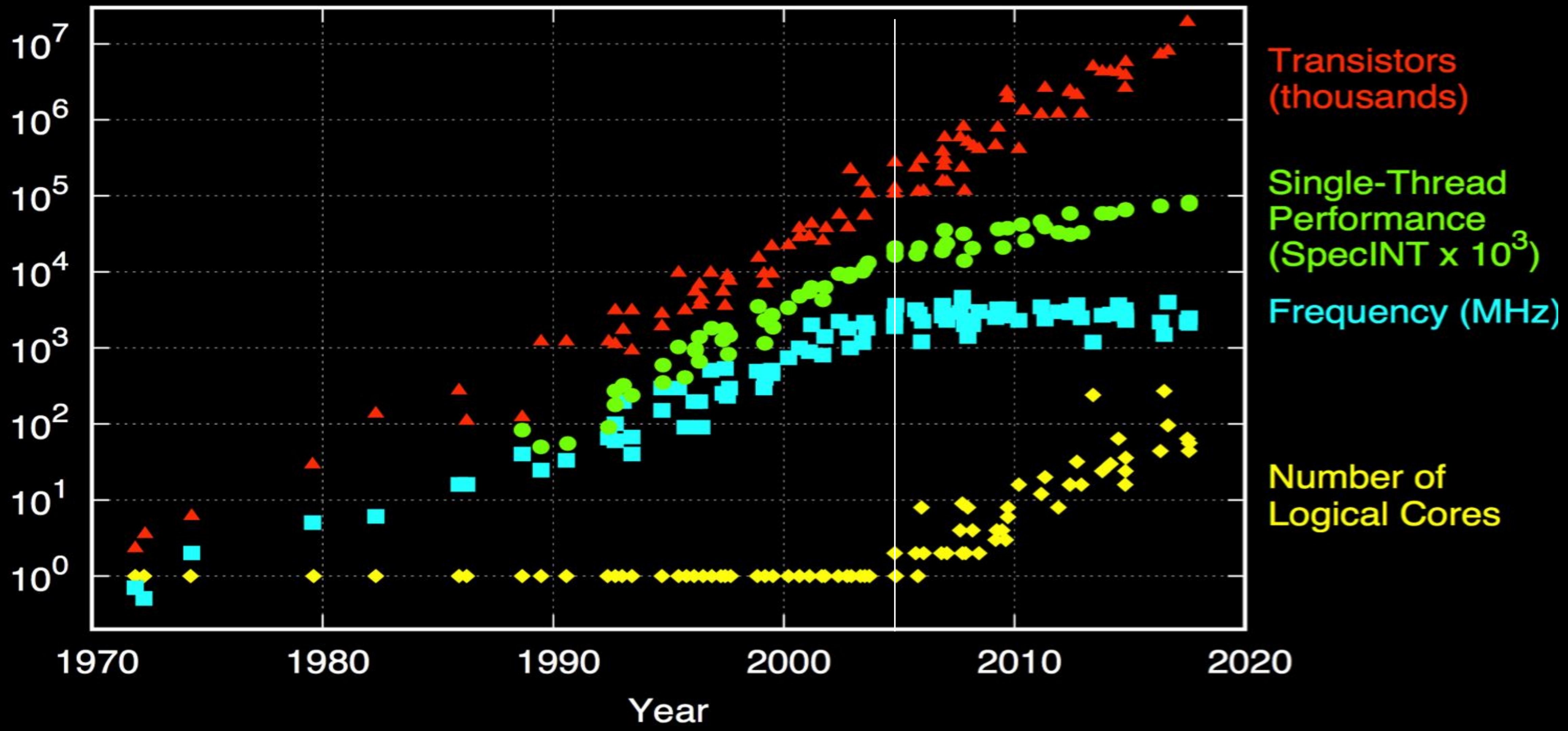
10 µm	6 µm	3 µm	1.5 µm	1 µm	800 nm	600 nm	350 nm	250 nm	180 nm	130 nm	90 nm	65 nm	45 nm	32 nm	22 nm	14 nm	10 nm	7 nm	5 nm	3 nm





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

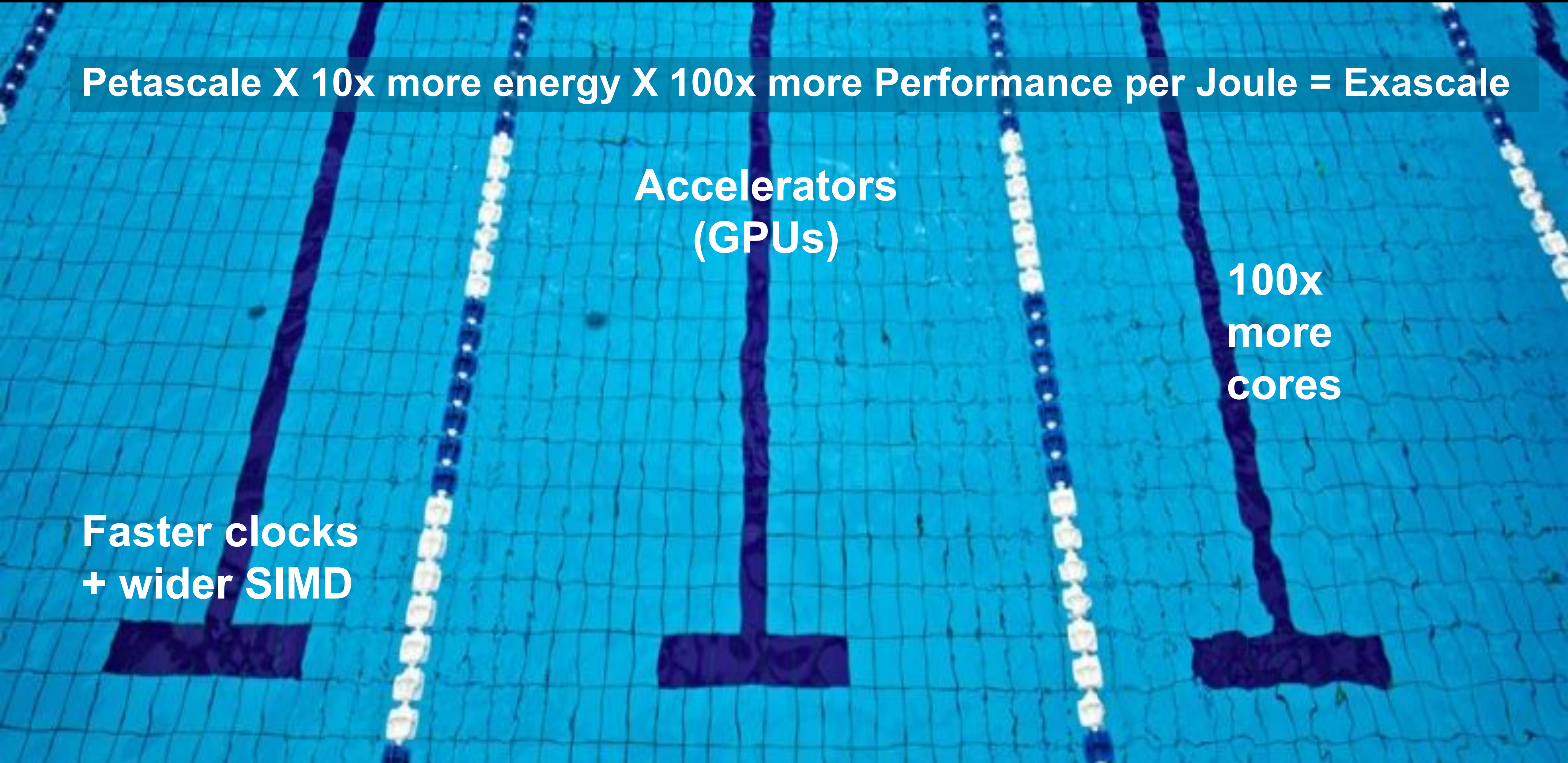
Dennard Scaling is Long Dead; Moore’s Law Will Follow



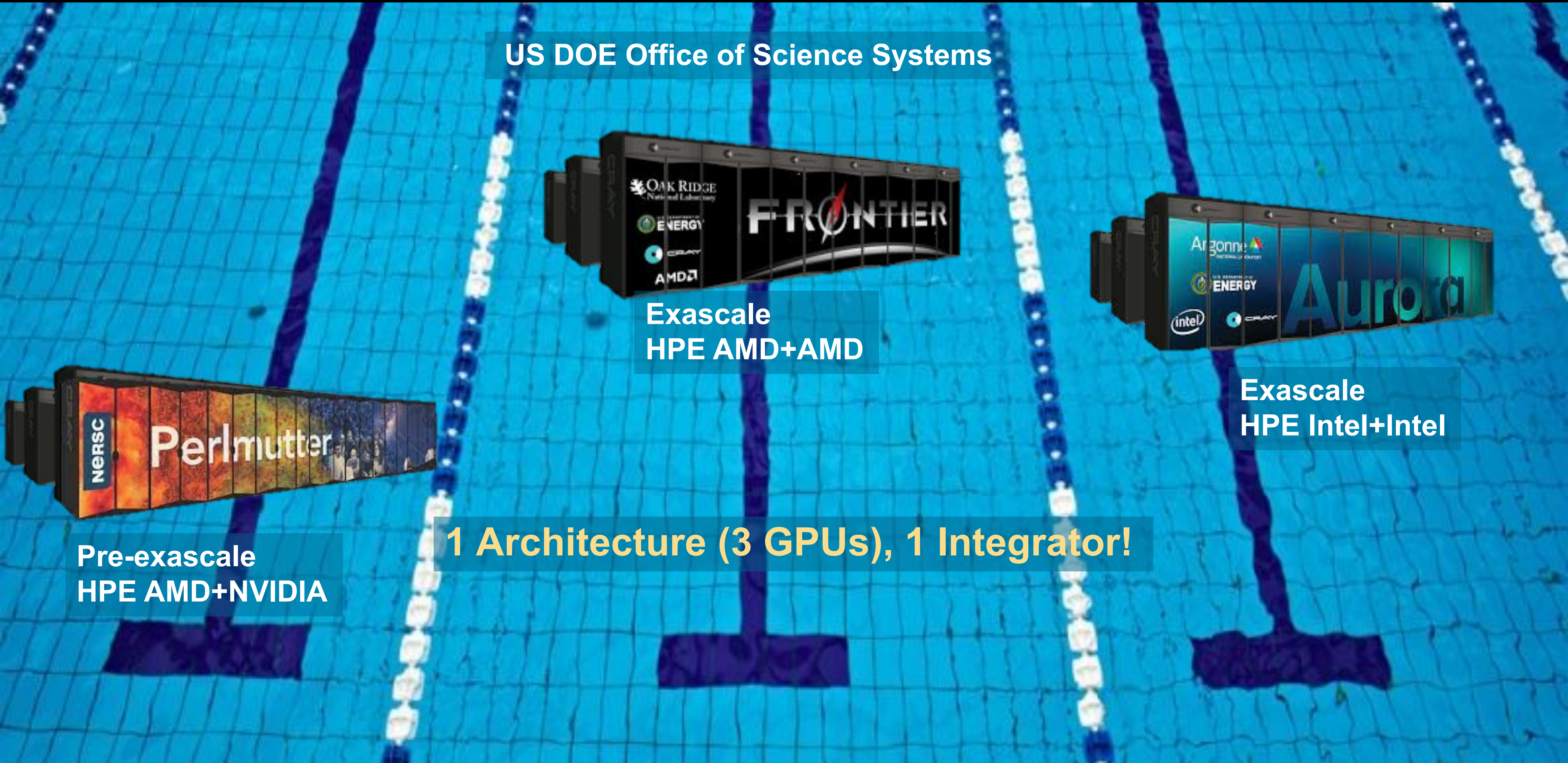
M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, C. Batten, and K. Rupp

Performance Programming pre 2005

Exascale Architecture Plans (2008)



Exascale Era Architectures (US DOE Office of Science)



US DOE Office of Science Systems

Exascale
HPE AMD+AMD

Exascale
HPE Intel+Intel

Pre-exascale
HPE AMD+NVIDIA

1 Architecture (3 GPUs), 1 Integrator!

First-in-Class HPC Systems (Top500)

	First Teraflop		First Petaflop		First Exaflop
	ASCI Red		Roadrunner		Frontier
Year	1997		2008		2022
Best Tech (nm)	500	10x →	65	10x →	6
Power (MW)	0.9	2x →	2.4	10x →	21.1
Efficienty (GF/W)	0.001	400x→	0.4	100x→	52
Memory (PB)	0.001	40x →	0.04	200x→	9
FPUs (K)	9	100x →	464	1000x→	534,000
Foorspace (m^2)	150	4x →	557	1x →	678

Kogge and Dally: Frontier vs the Exascale Report + Wikipedia for ASCI Red

Energy efficiency didn't track technology scaling

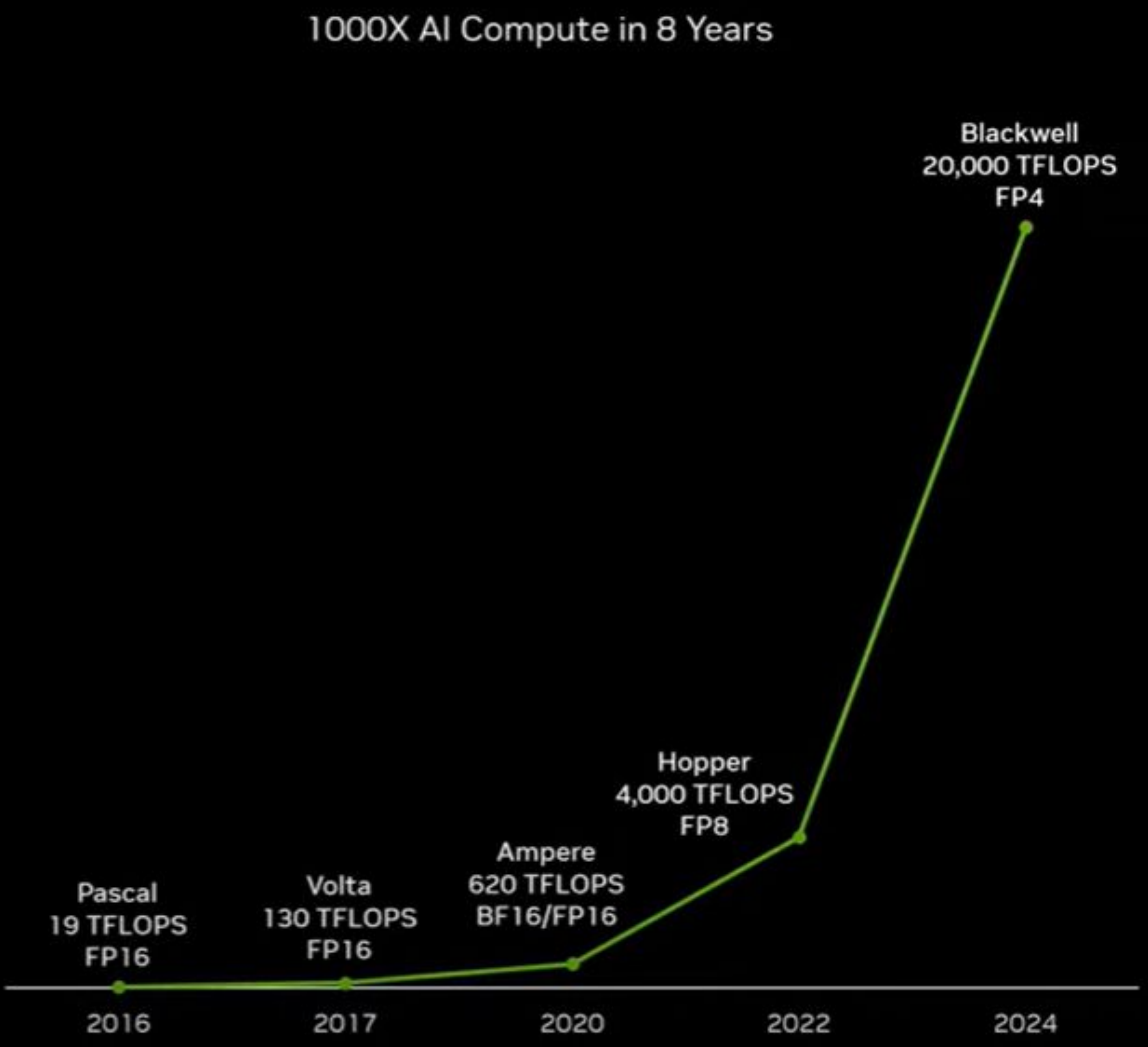
Gate Length (nm)	65	32	16	6
Metal 1 pitch (nm)	180	100	64	40
$Energy^{-1}$	1	1.8	2.8	4.5
$Area^{-1}$	1	3.2	7.9	20.3

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?

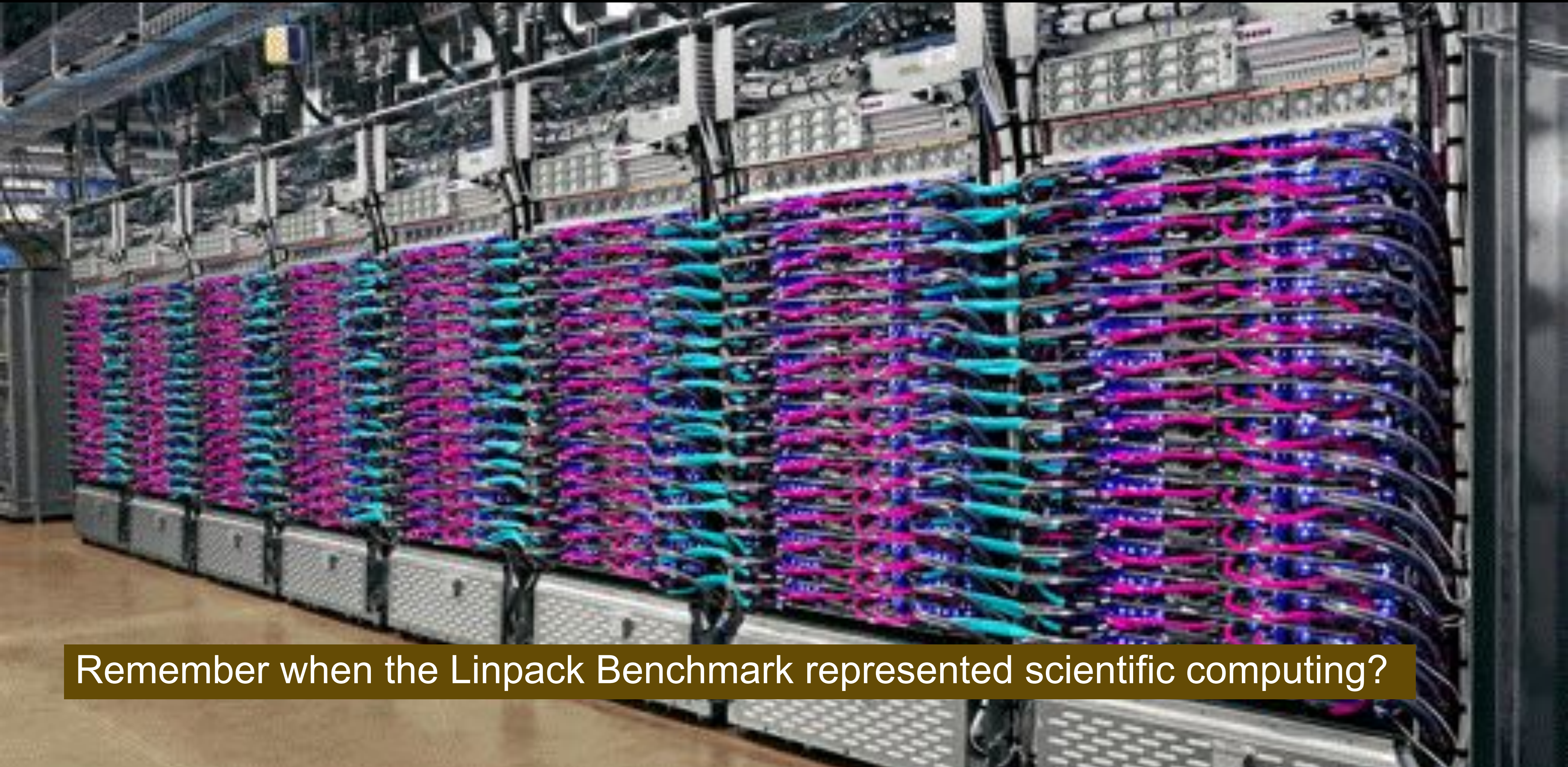
Post-Exascale Architecture Plans 2024 (Strawperson-v0)

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…Not everyone is selling them!

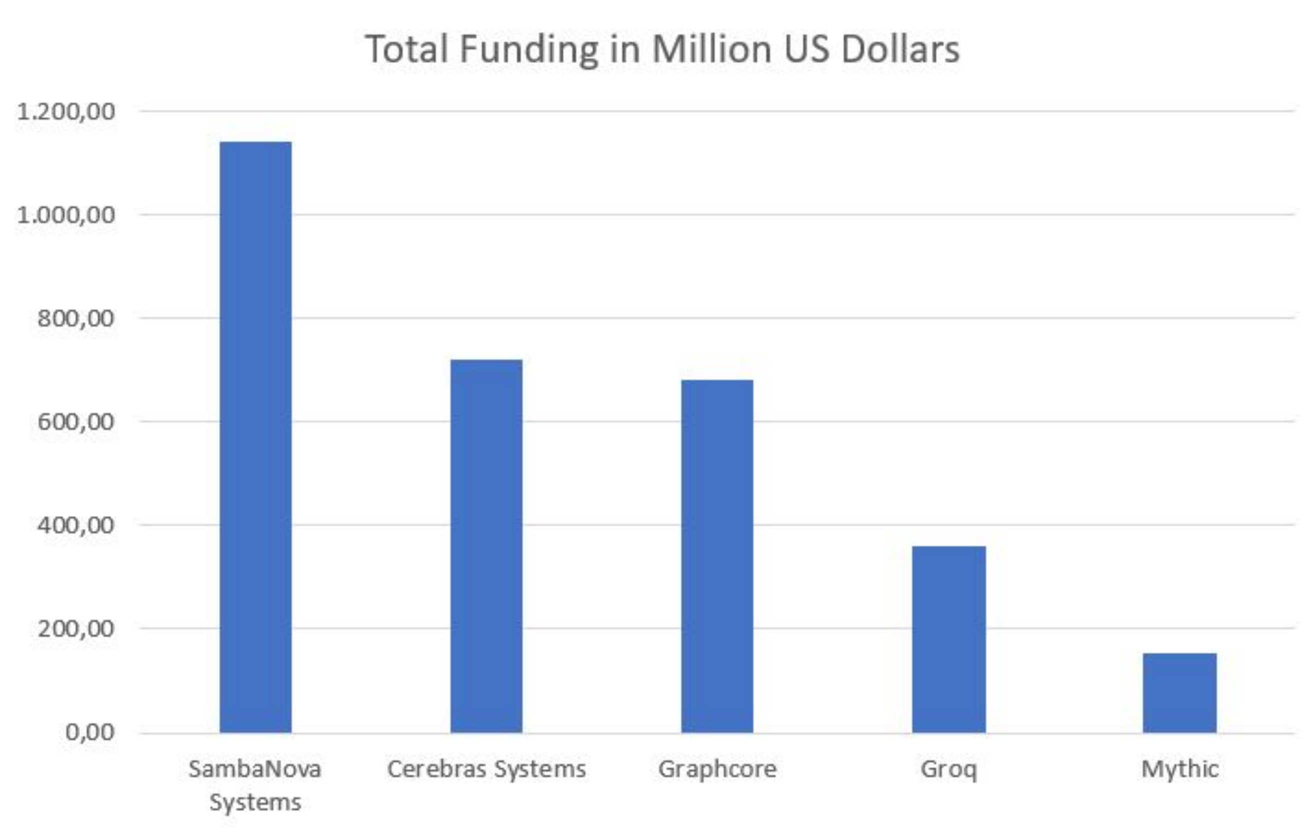
NVIDIA
AMD
Intel
IBM

Traditional chip makers

“Software” companies

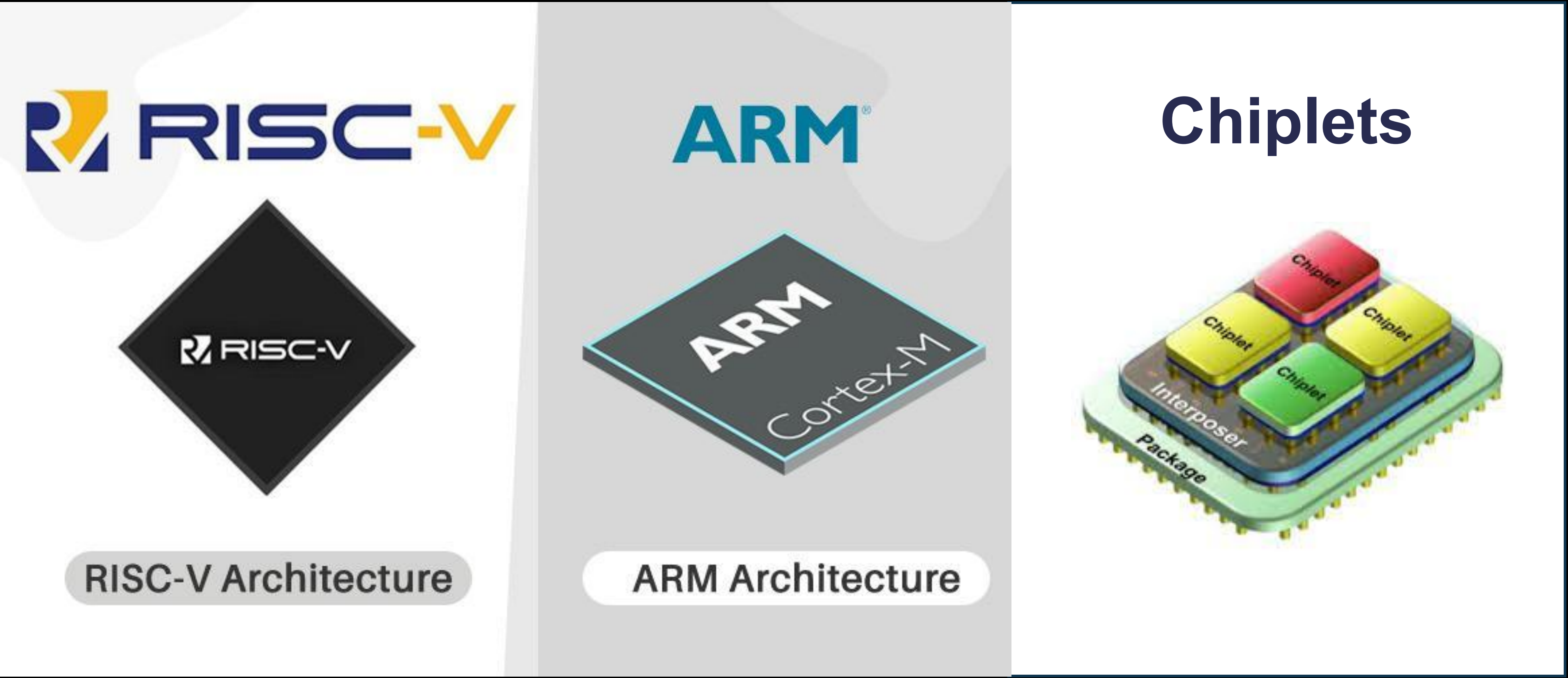
Facebook + Intel
Amazon (Echo, Oculus)
Google (TPU, Pixel)
Apple (SoCs)
Microsoft (“AI chip”)

Startups



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

Specialization for the masses?



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 AI in Science

  and Simulation ?

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