Can Climate Science Win the Hardware Lottery?

Kathy Yelick

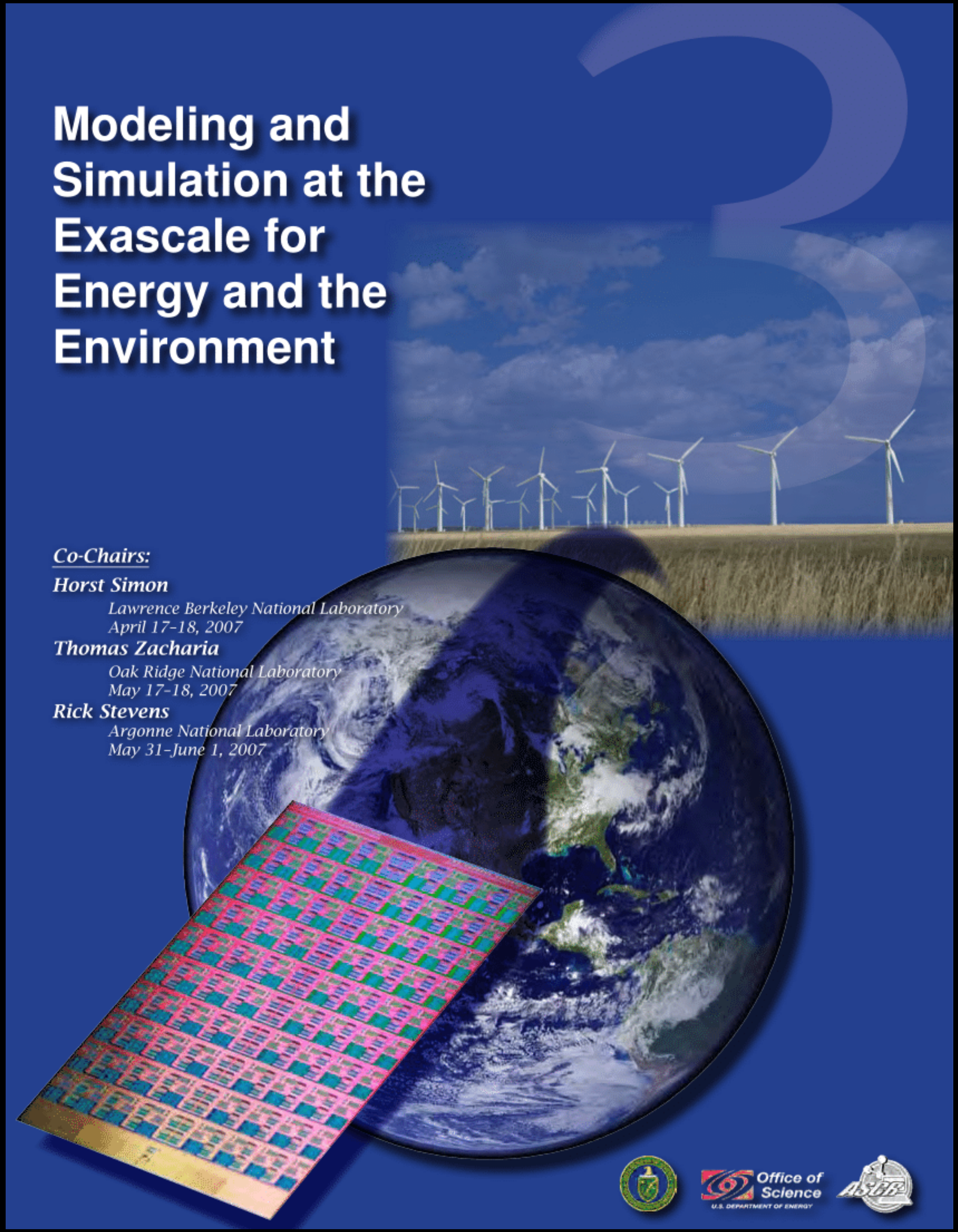
Vice Chancellor for Research
and Robert S. Pepper Professor of
Electrical Engineering and Computer
Sciences
U.C. Berkeley

Senior Faculty Scientist
Computing Sciences

HPC Circa 2007

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

Simulation ≠ Scientific Computing ≠ HPC



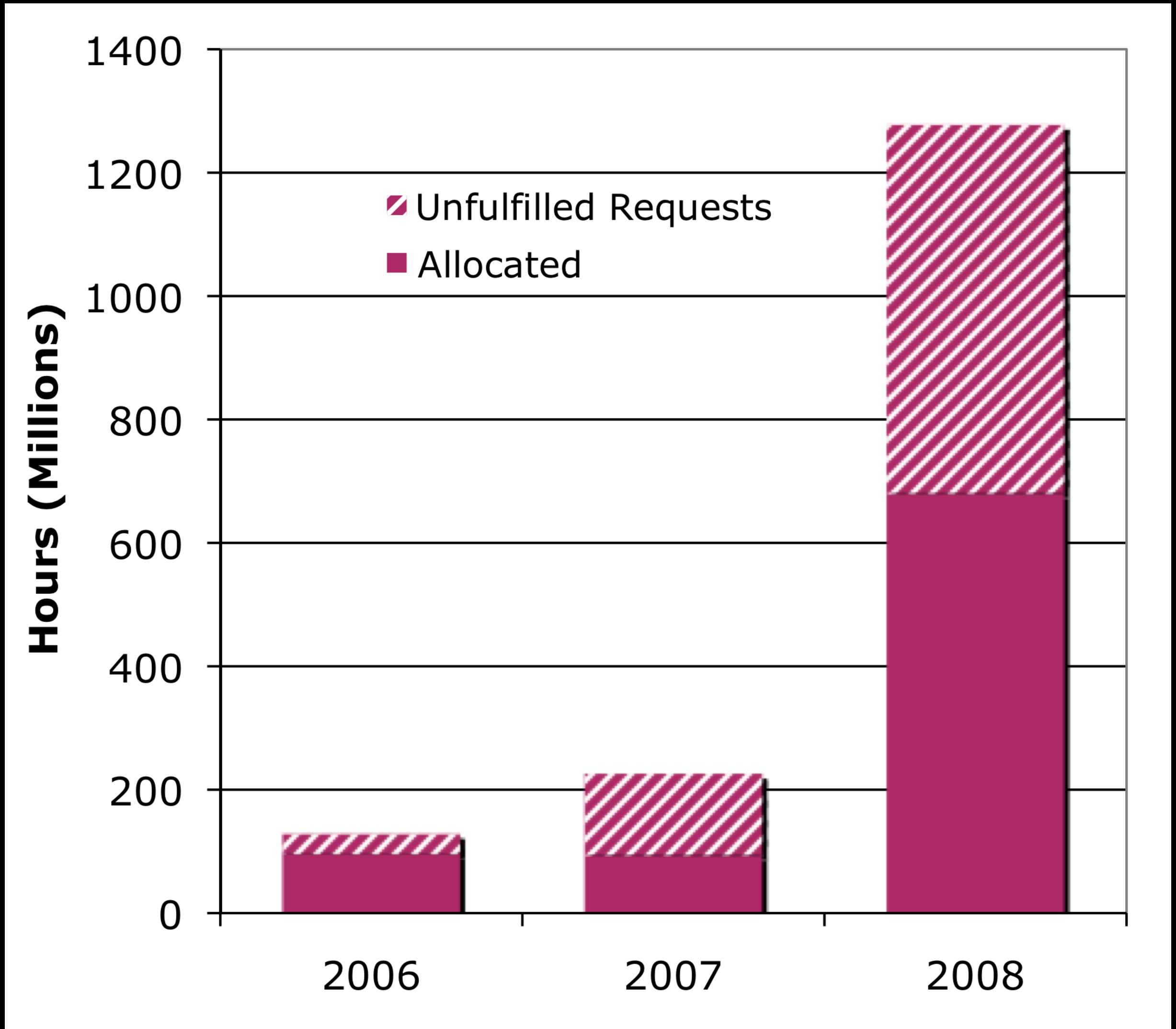
Continue to Rethink Applications

ECP EXASCALE COMPUTING PROJECT

- 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!!

DOE Demand for Computing is Growing (circa 2008)

Compute Hours Requested vs Allocated



- *Each year DOE users requests 2x more hours than allocated*
- *This 2x is artificially constrained by perceived availability*
- *Unfulfilled allocation = hundreds of millions of hours in 2008*
- *When allocation limits are removed, scaling and science increase*

New demands for HPC in Science

Simulation

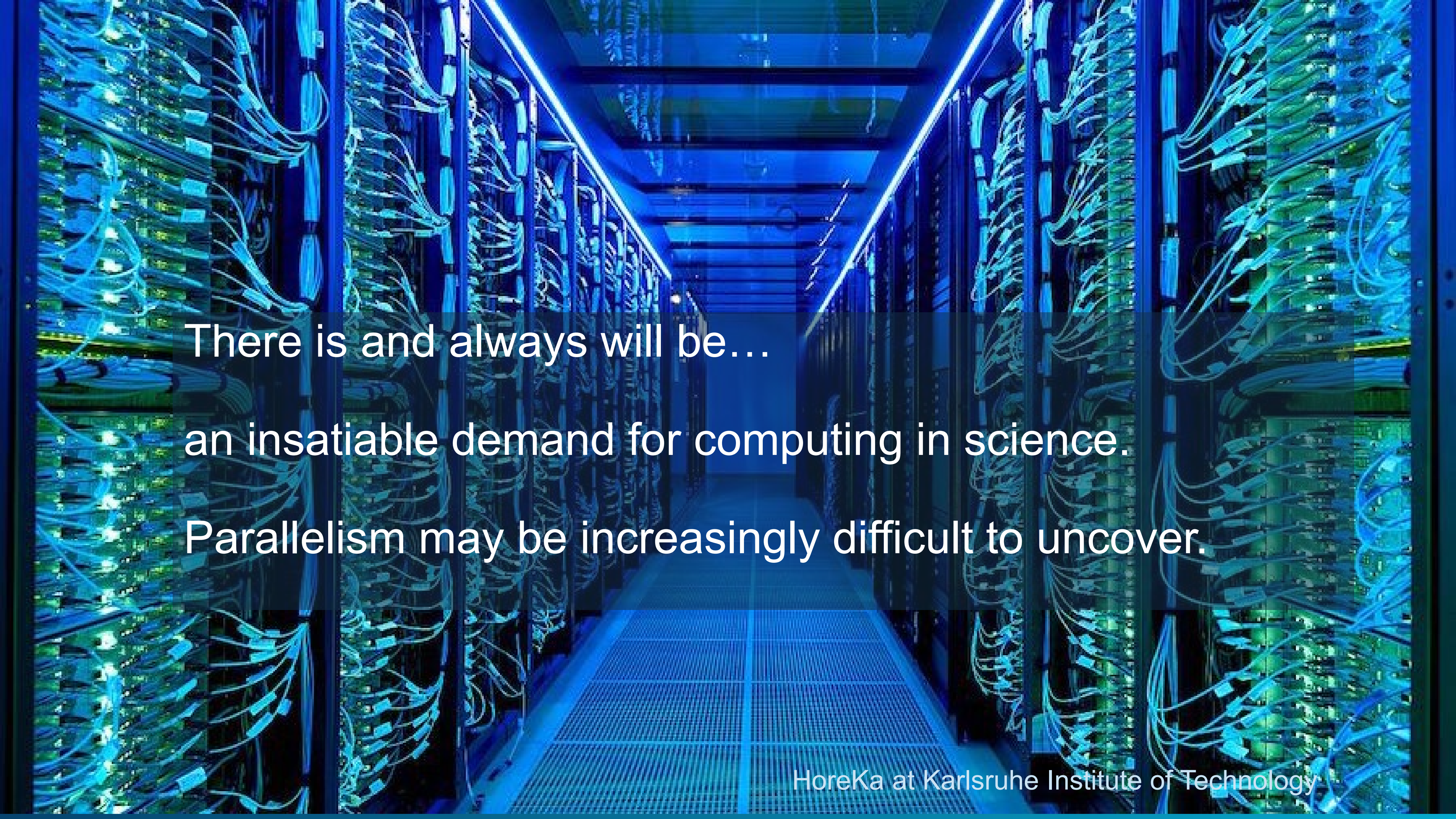
From atoms to the universe

Data

Images, text, to genomes

Learning

Interpret, generate, and automate



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

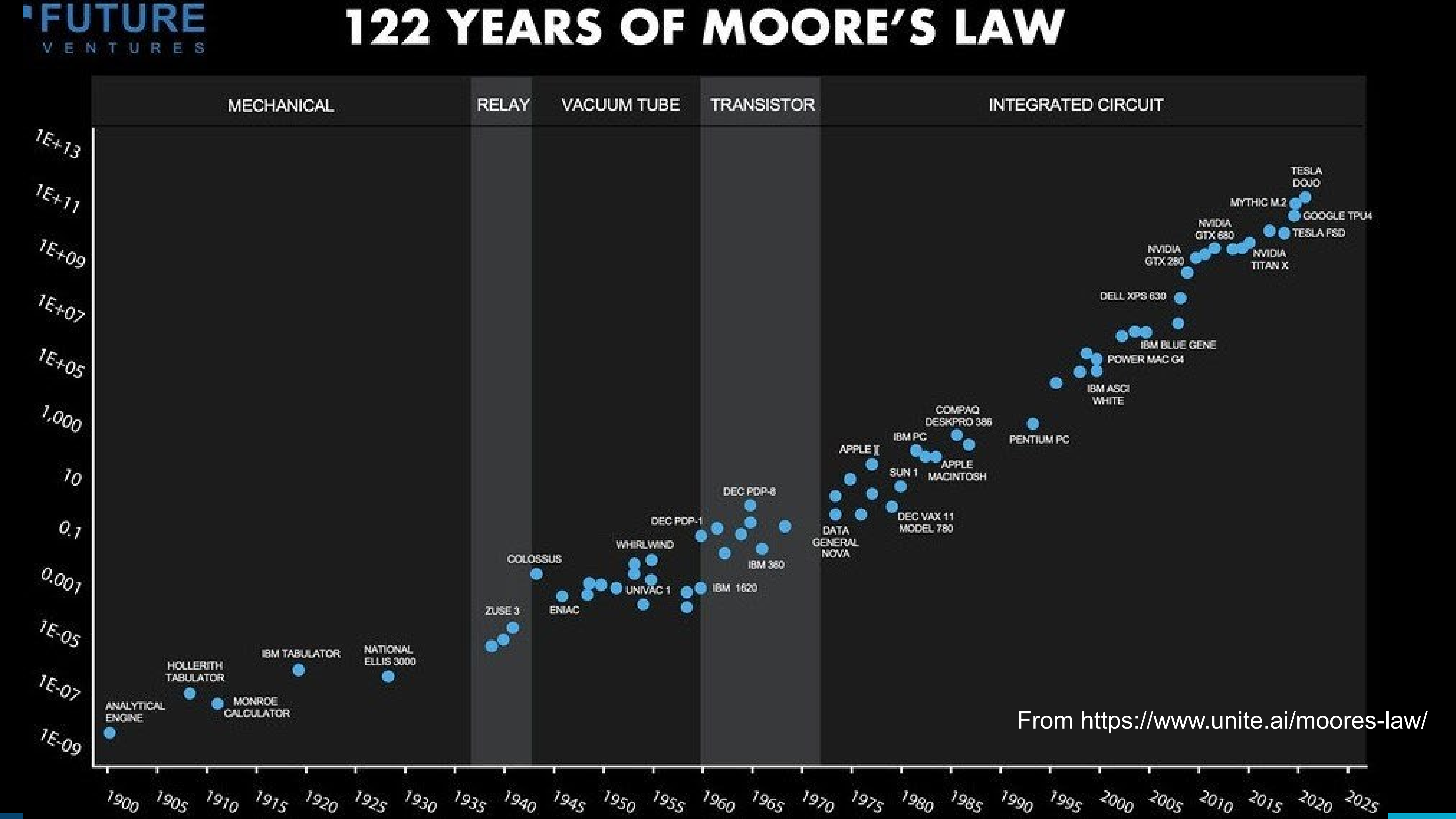
Hardware Lottery

When a research idea wins because it is suited to the available software and hardware and

not because the idea is superior to alternative research directions

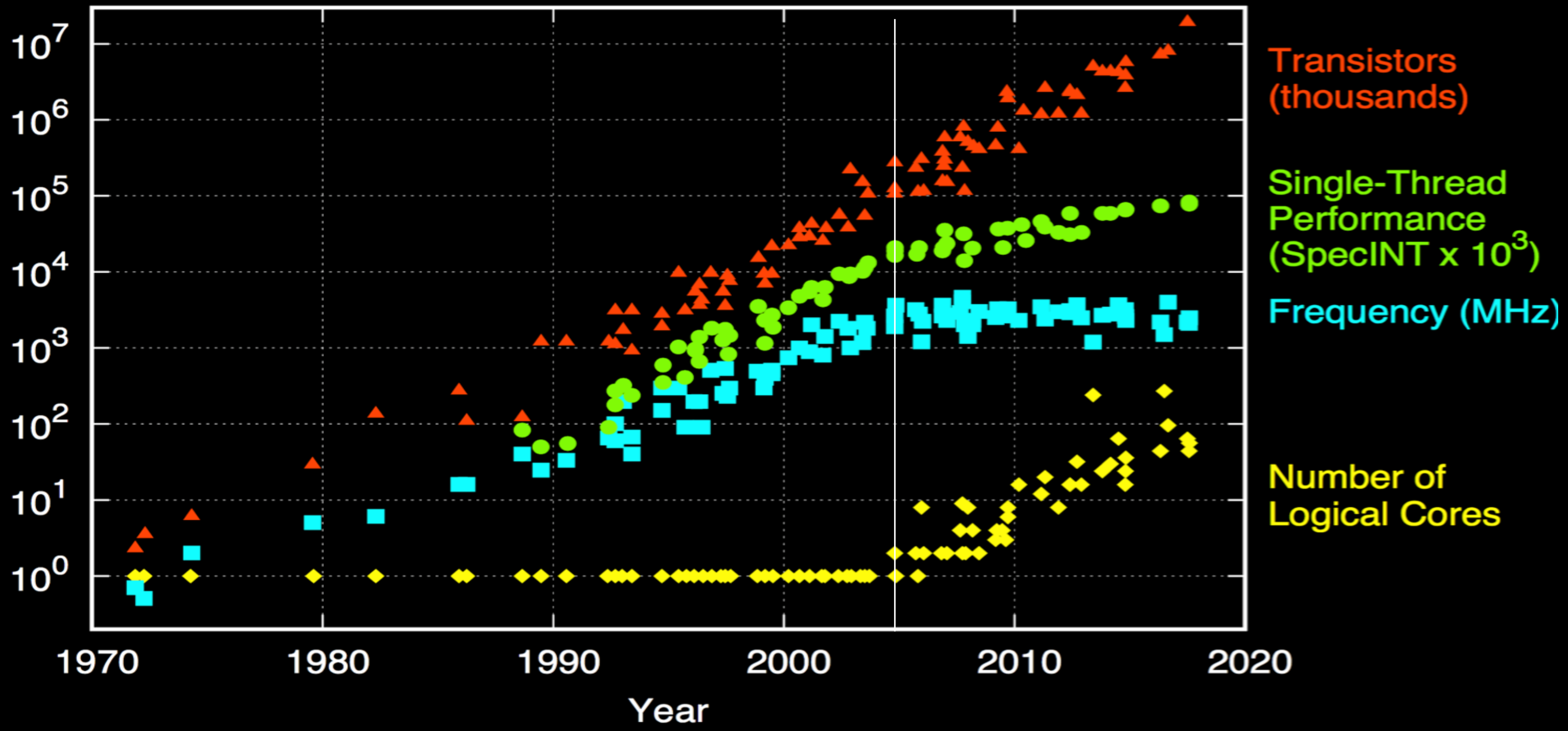
Sara Hooker

122 YEARS OF MOORE'S LAW



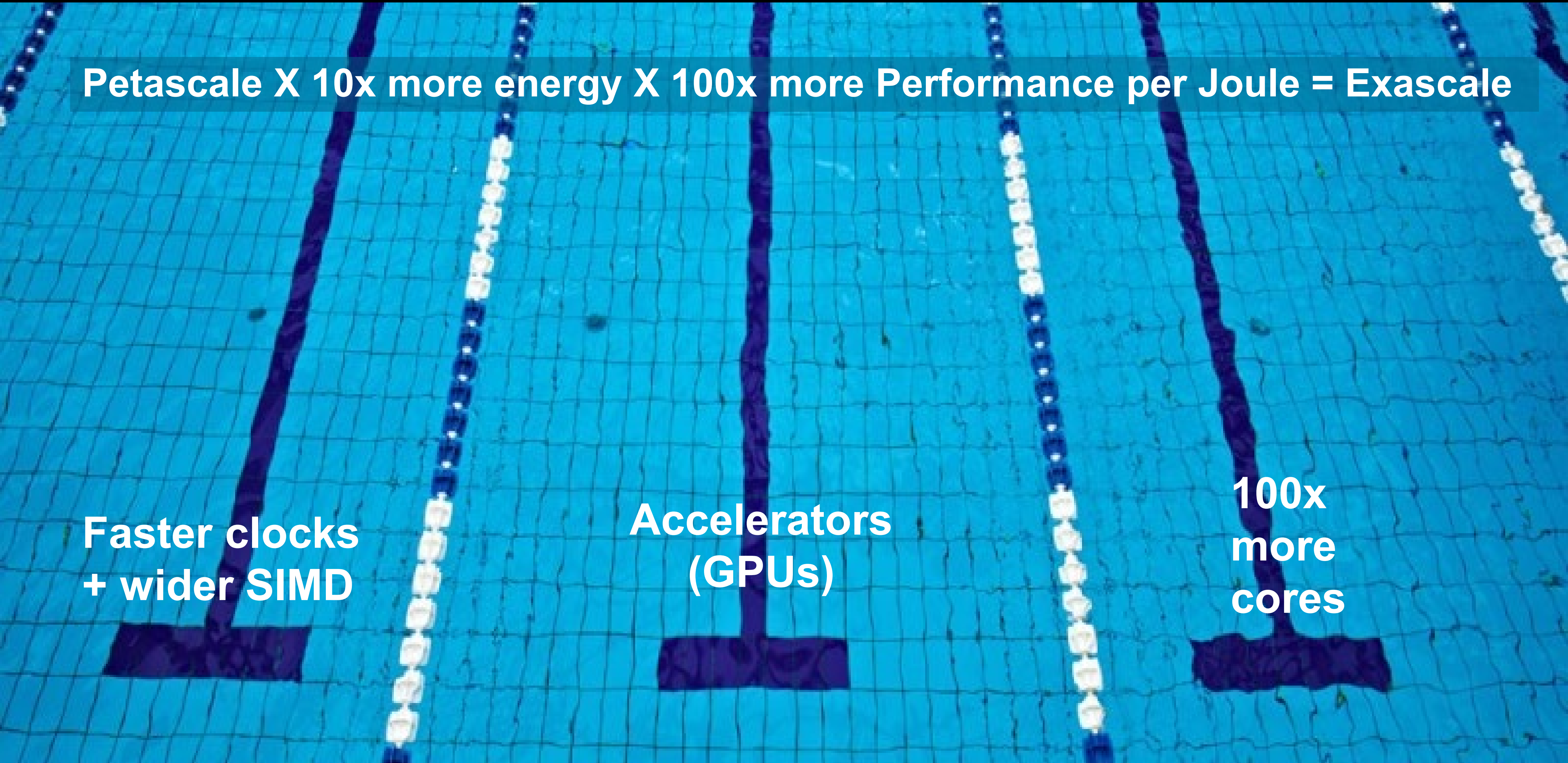
From https://www.unite.ai/moores-law/

Dennard Scaling is Long Dead; Moore's is Dieing



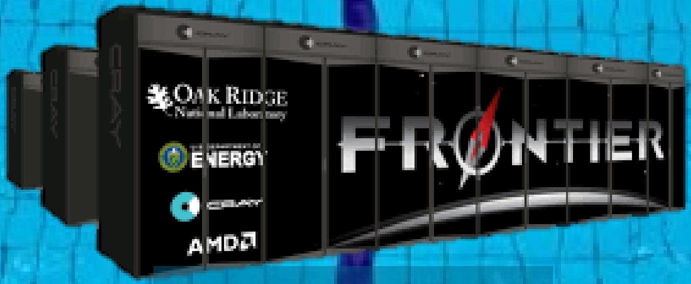
M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, C. Batten, and K. Rupp

Exascale Architecture Plans (2008)



Exascale Era Architectures (US DOE Office of Science)

US DOE Office of Science Systems



Exascale
HPE AMD+AMD

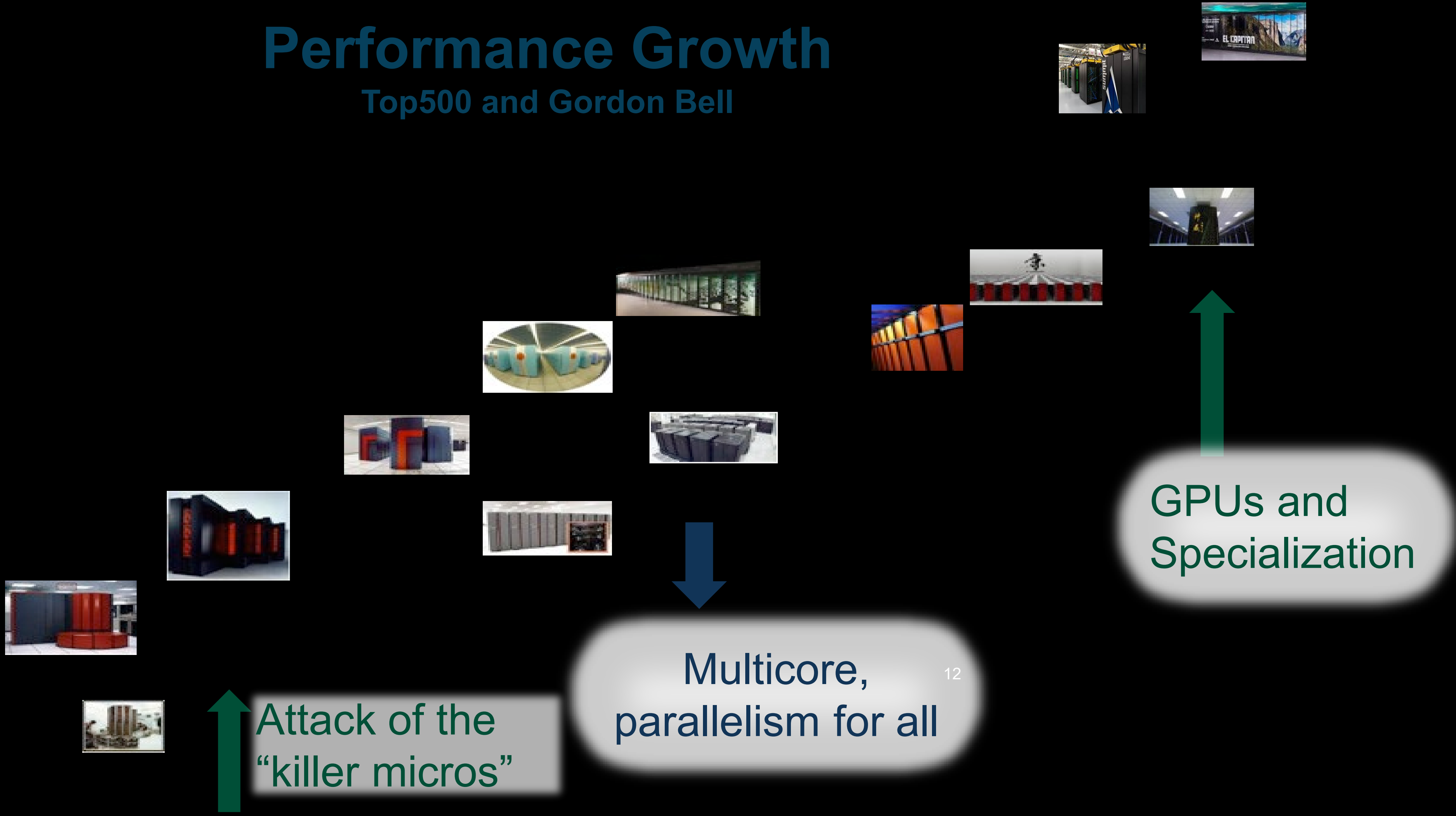
Exascale
HPE Intel+Intel

Pre-exascale
HPE AMD+NVIDIA

1 Architecture (3 GPUs), 1 Integrator!

Performance Growth

Top500 and Gordon Bell



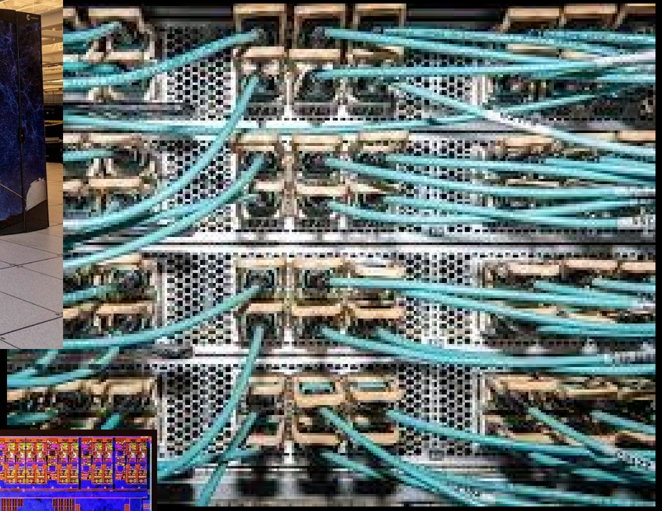
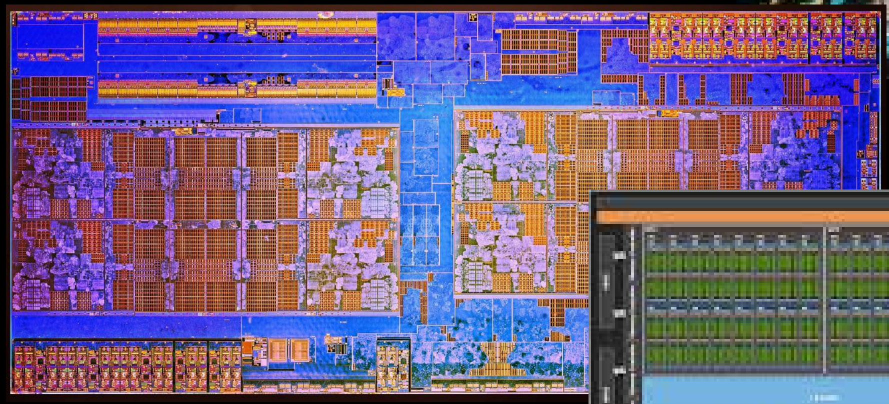
First-in-Class HPC Systems (Top500)

	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

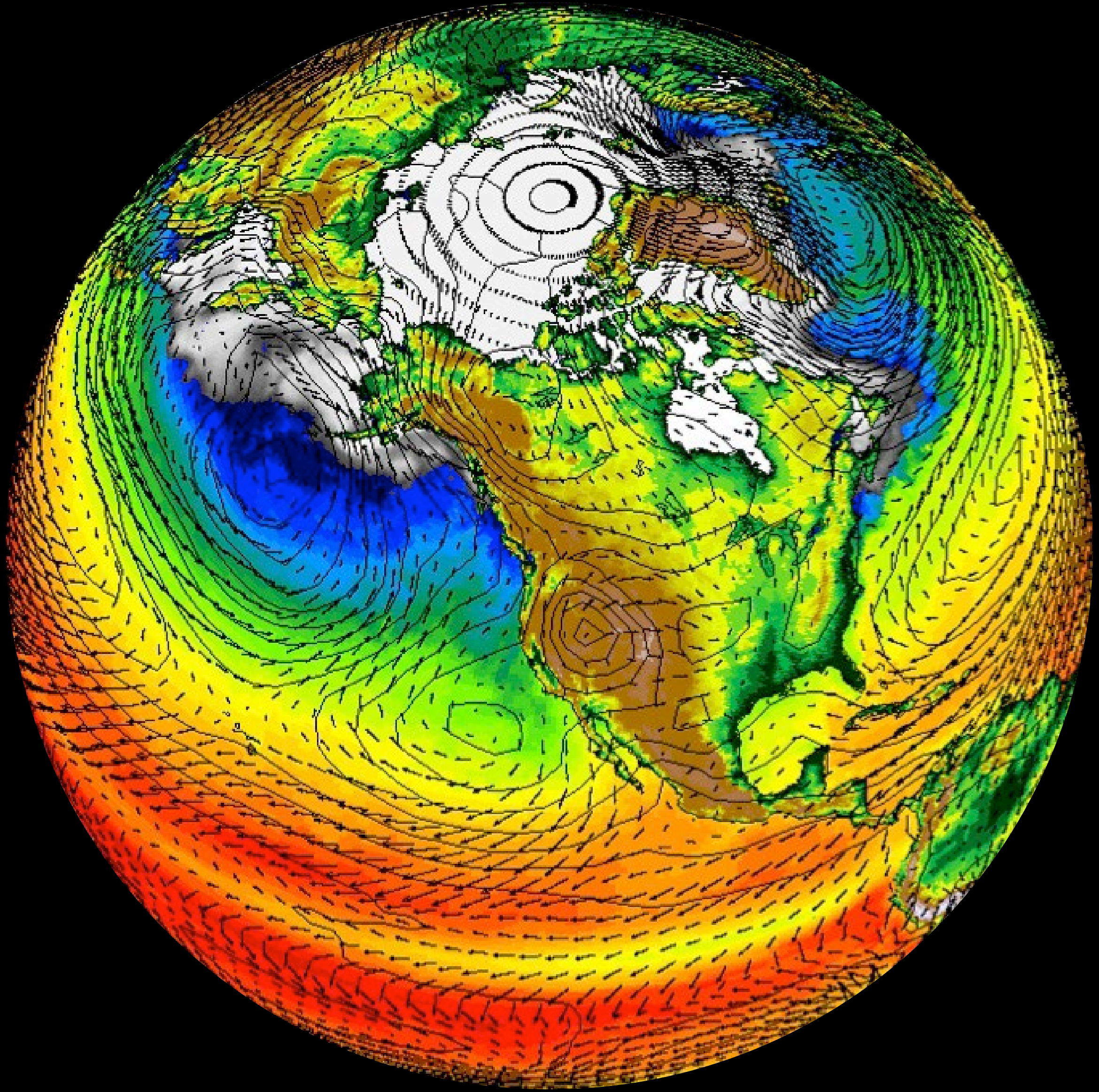
Kogge and Dally: Frontier vs the Exascale Report + Wikipedia for ASCI Red

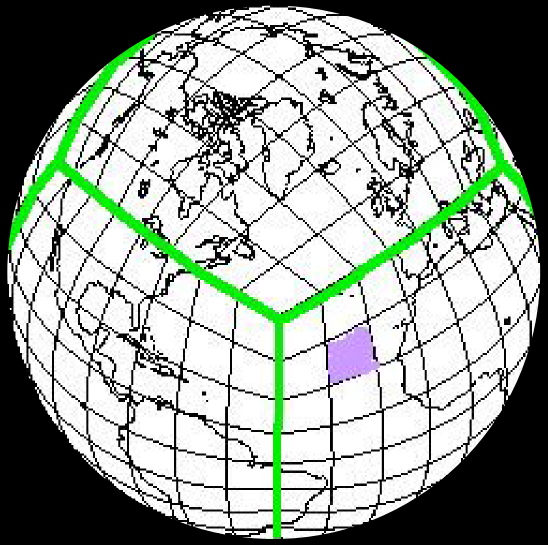
Parallelism: Everything Everywhere All At Once

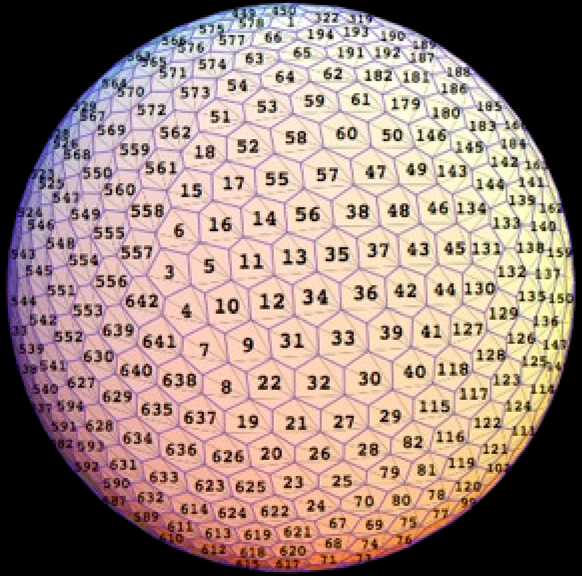
- Between Nodes
- Communication (overlap)
- Avoid synchronization
- Multicore
- GPUs/Accelerators
- SIMD
- Pipelineing


Players in the Climate Modeling Hardware Lottery

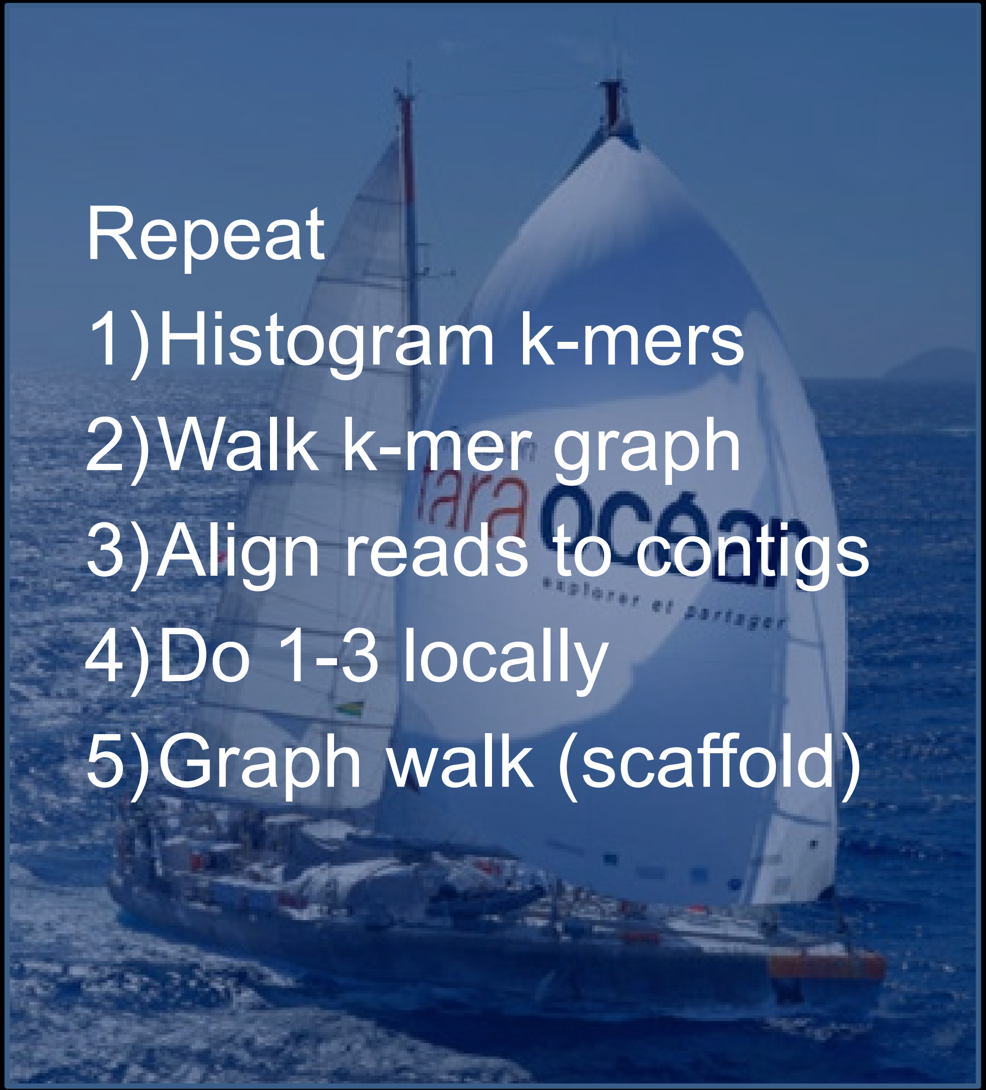






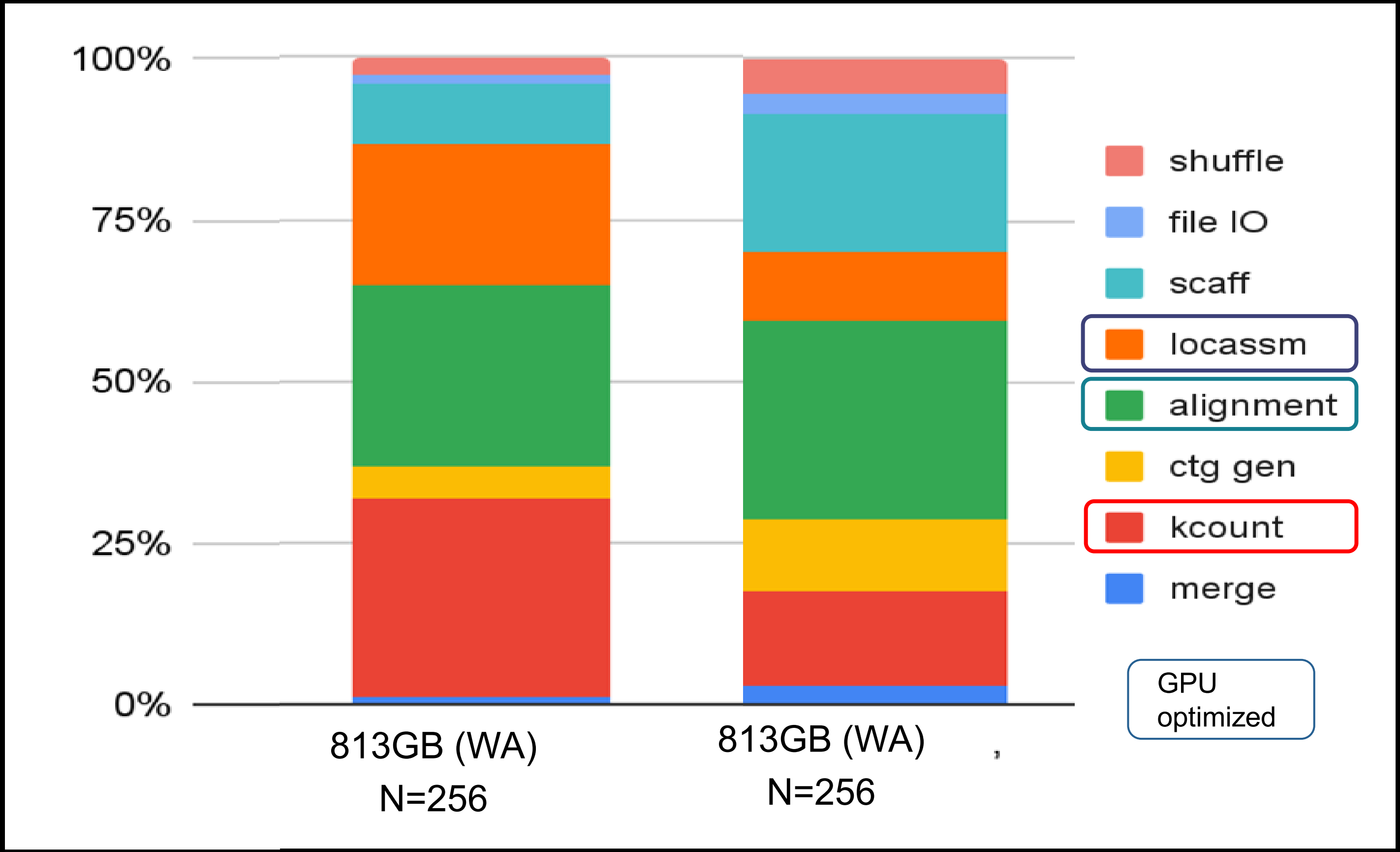
MetaGenome Assembly: Lottery Player?

From shared memory to HPC + GPUs



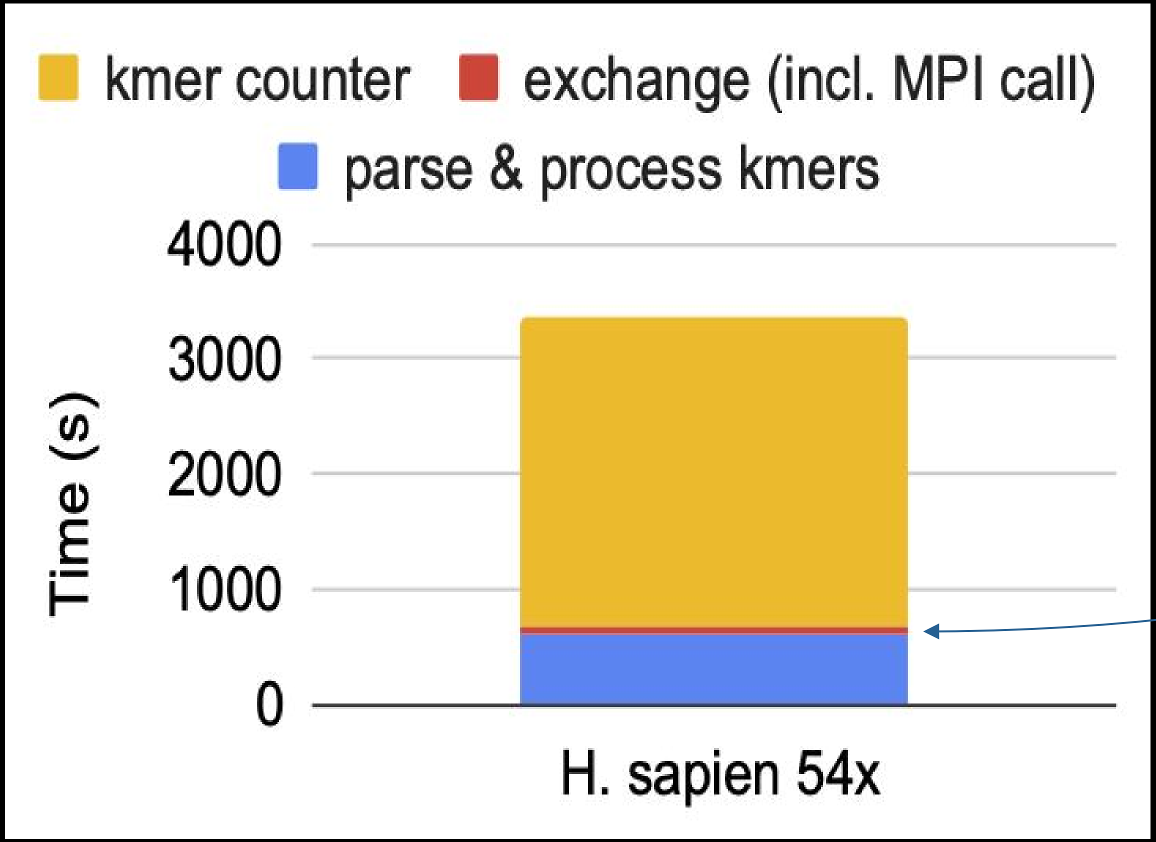
Repeat

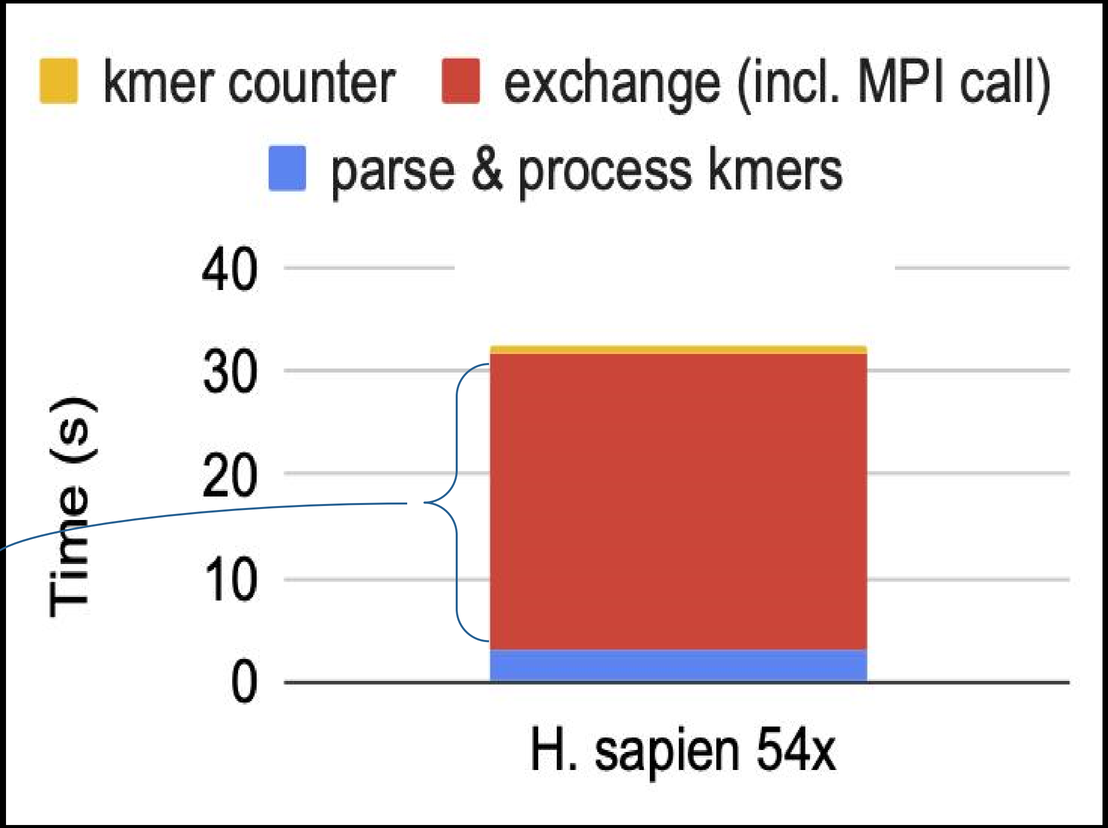
1) Histogram k-mers
2) Walk k-mer graph
3) Align reads to contigs
4) Do 1-3 locally
5) Graph walk (scaffold)



Muaaz Awan, J. Deslippe, S. Hofmeyr, R. Egan, A. Buluç, L. Oliker, K. Yelick, SC21 Best Paper Finalist

K-mer Counting: Finding Data Parallelism

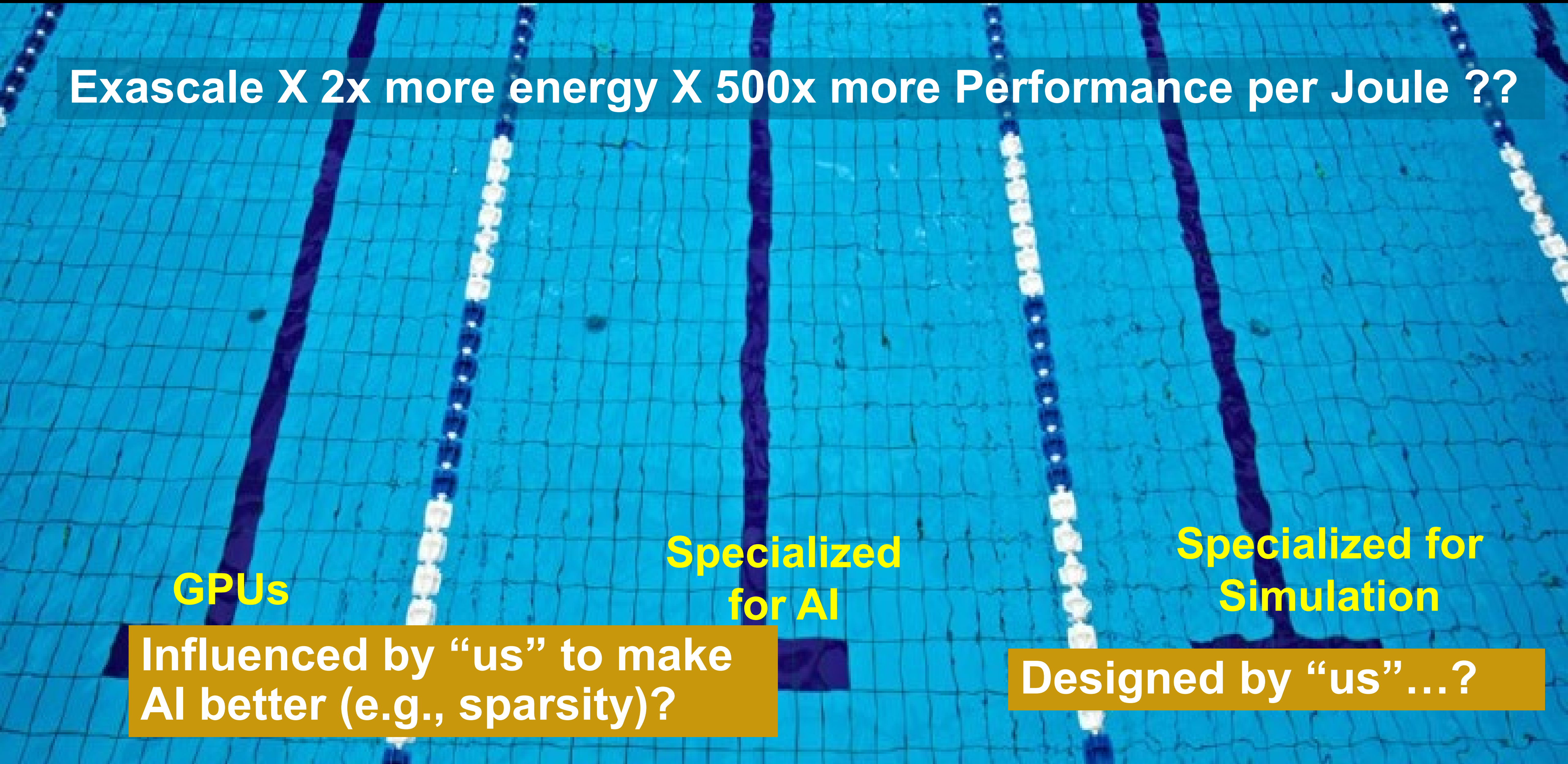




- K-mer counter on Summit. (Note scales -- red k-mer exchange time is roughly equal.)
- Reduce CPU/GPU communication by parsing as well as processing on GPU

Israt Nisa, P. Pandey, M. Ellis, L. Oliker, A. Buluç, K. Yelick. Distributed-Memory k-mer Counting on GPUs. *IPDPS '21 (to appear)*

Post-Exascale Architecture Plans 2024 (Strawperson-v0)



Declining Benefits of Moore's Law Scaling

	2008	2012	2015	2022
Gate Length (nm)	65	32	16	6
Energy^{-1}	1	1.8	2.8	4.5

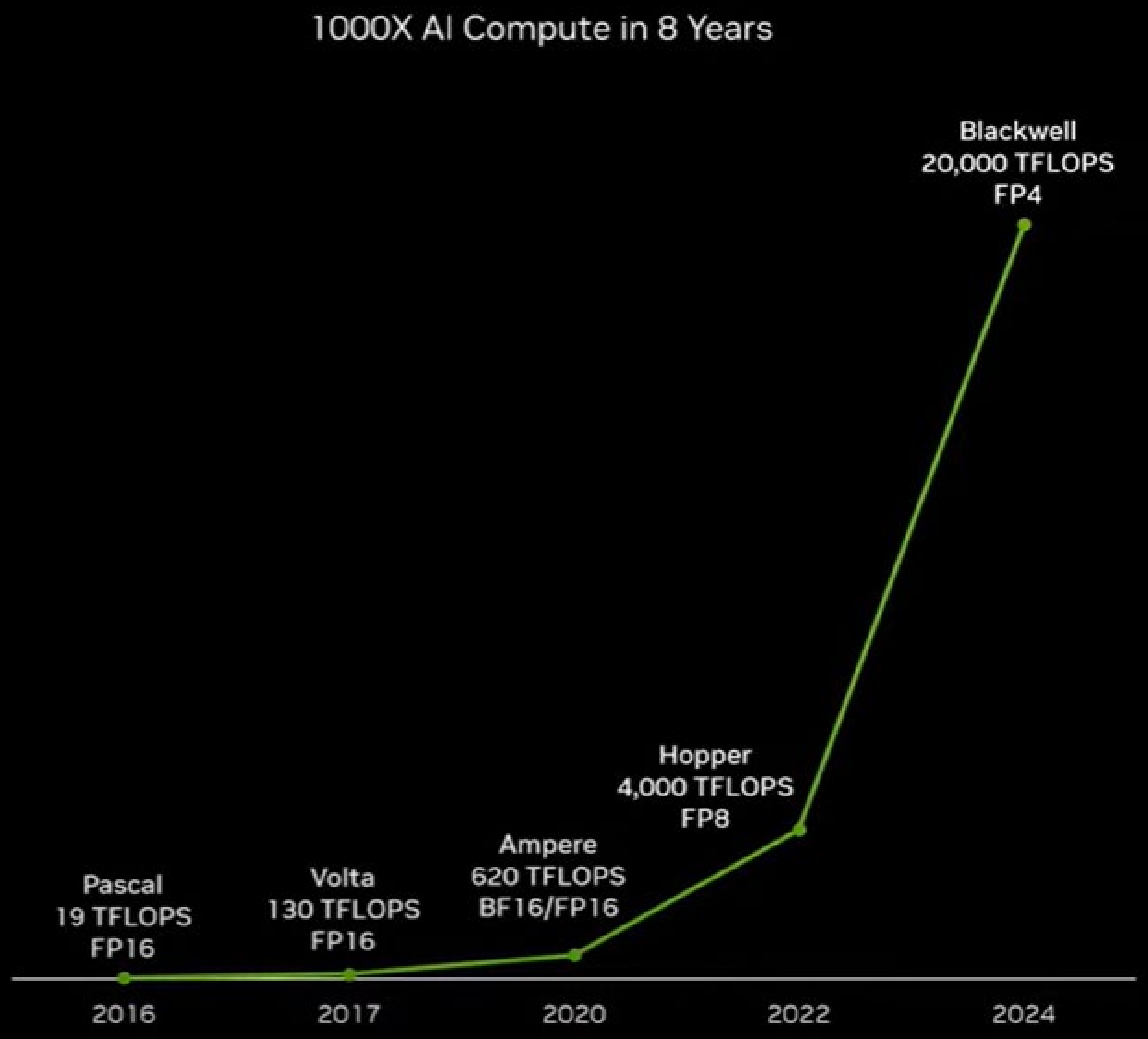
10X

4.5X

2nm fabs coming? But how much will it help?

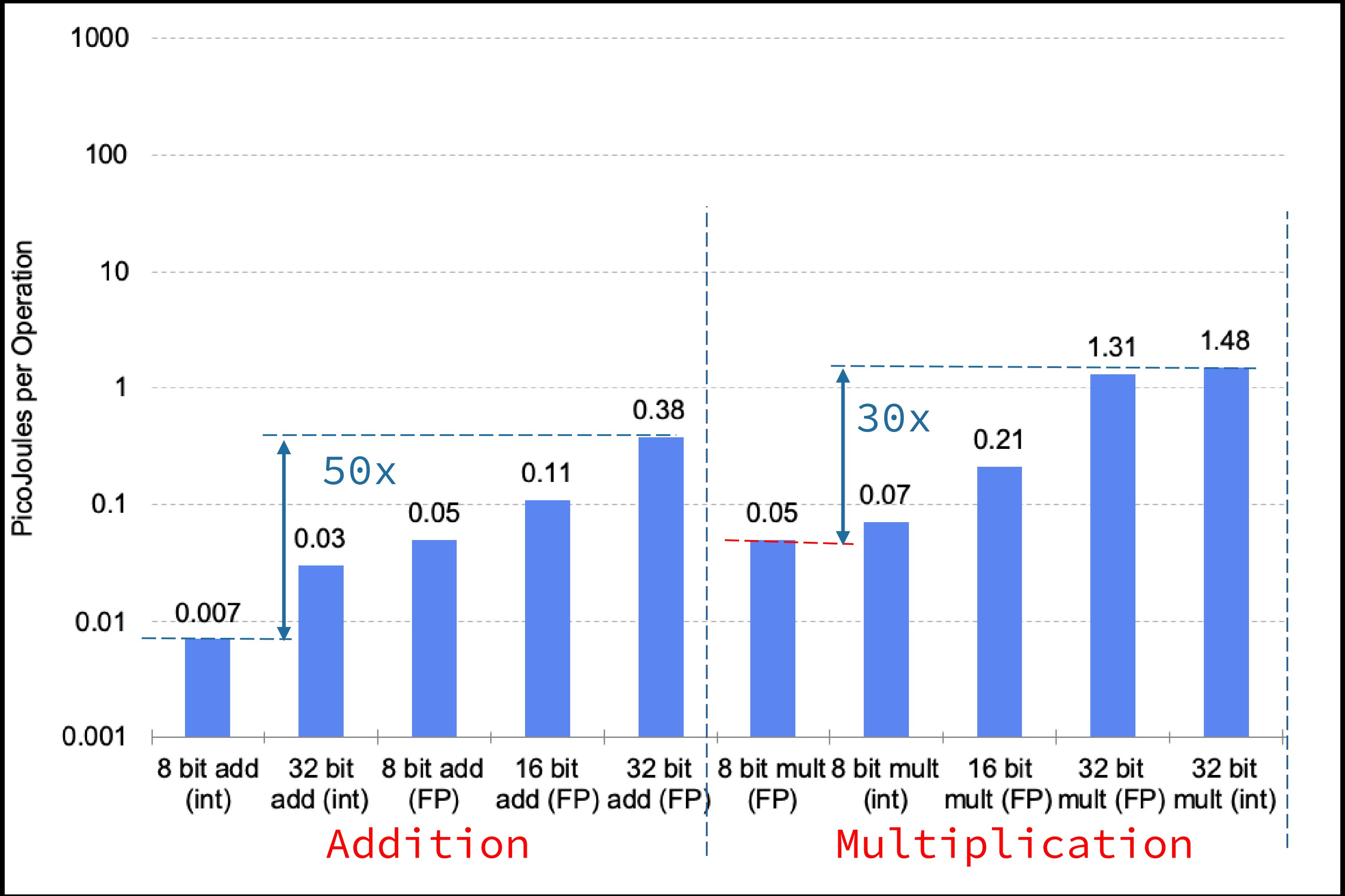
Kogge and Dally: Frontier vs the Exascale Report: Why so long? and Are We Really There Yet?

Another Exponential?



Jensen Huang's Nvidia GTC Keynote

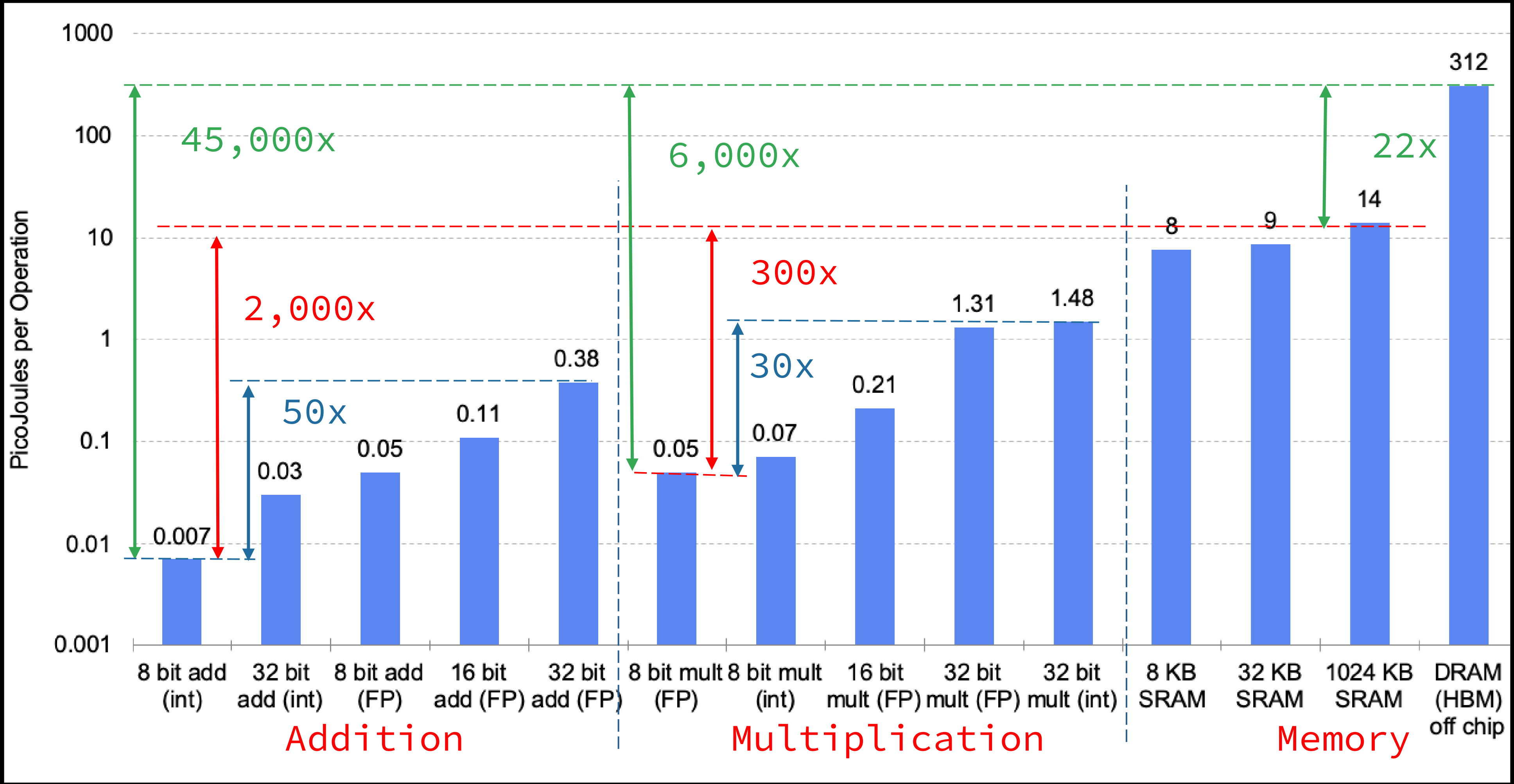
Narrow Data Types Save Energy



Graph from Dave Patterson

Smaller data uses less energy per op

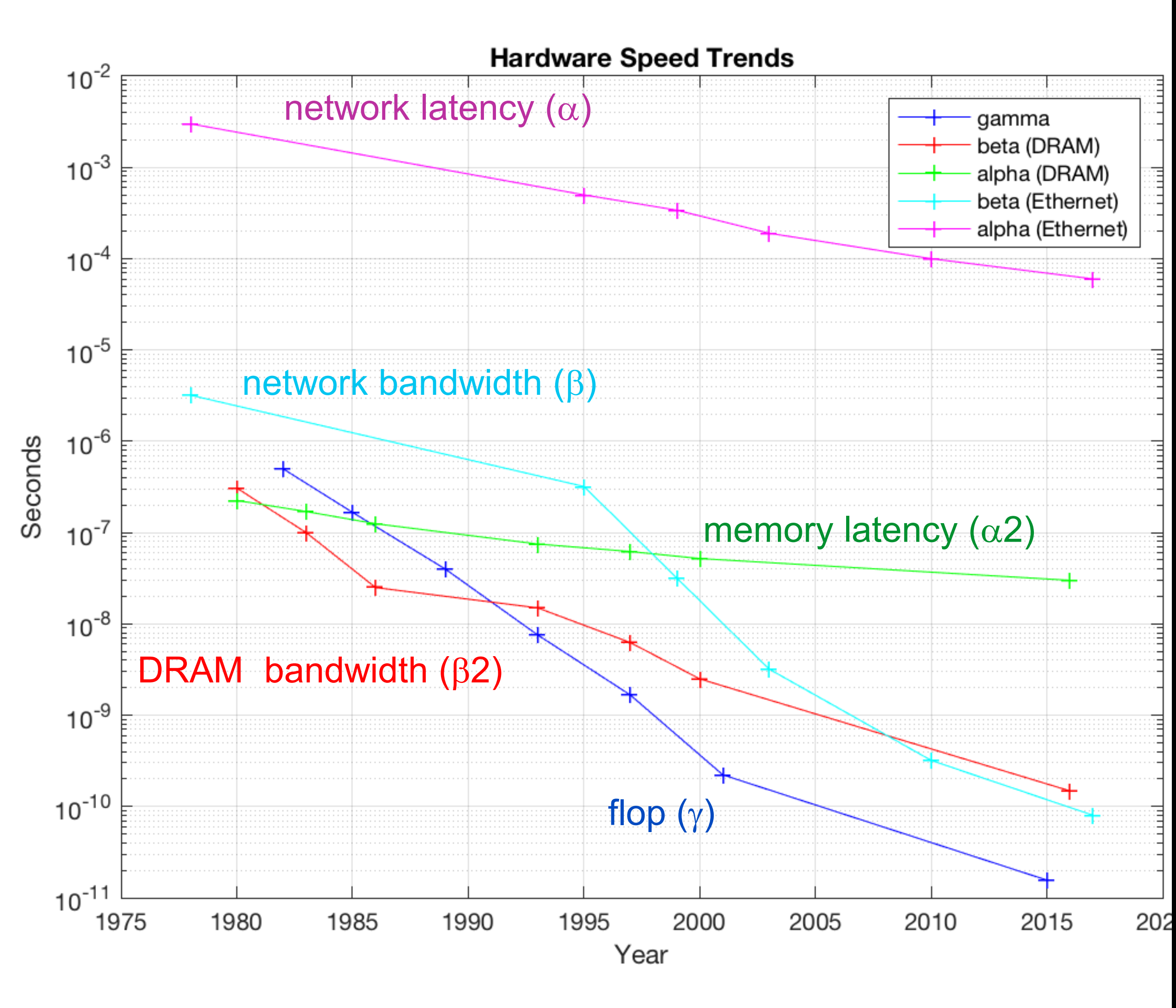
Narrow Data Types Save Energy



Smaller data also uses less less on-chip memory capacity and off-chip memory bandwidth

Graph from Dave Patterson

Data Movement is the Most Expensive

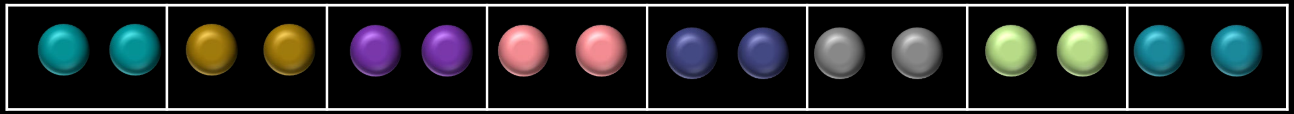


Data from Hennessy / Patterson, Graph from Demmel

Avoid Communication and Maximize Parallelism

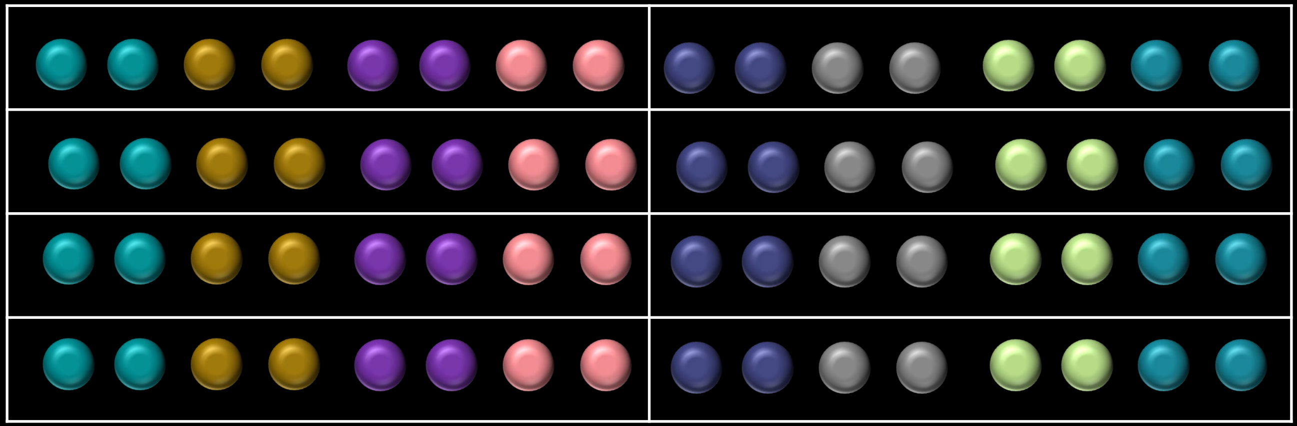
Compute on all pairs of particles, DNA, tokens, or…

Obvious solution



16 particles on 8 processors
Pass all particles around (p steps)

Better solution



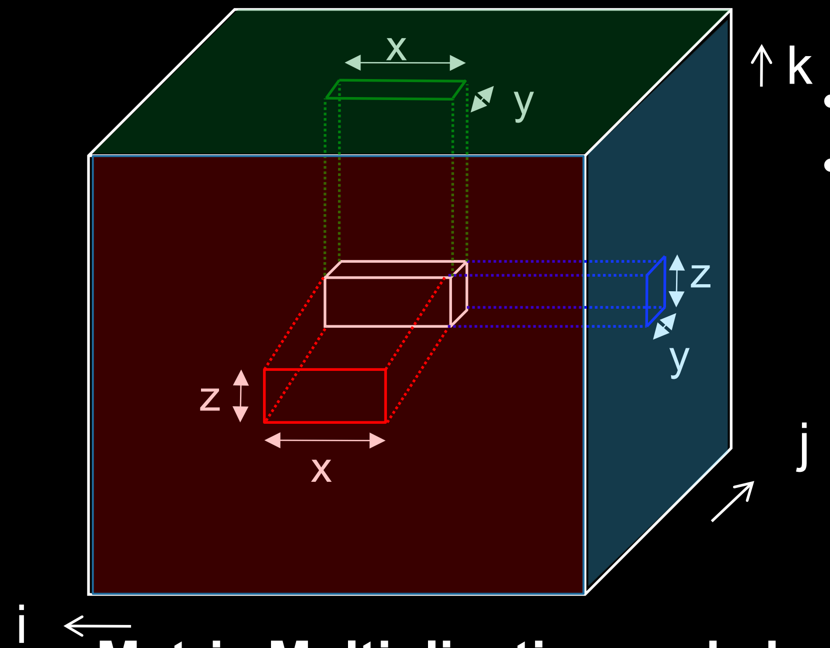
c = 4 copies of particles
8 particles on each

Decreases

- #messages by factor $\mathbf{c^2}$
- #volume sent by factor **c**

Michael Driscoll et al, "A Communication-Optimal N-Body Algorithm for Direct Interactions" IPDPS '13

Communication-Optimal Matrix Multiply



- 2D algorithm: never chop k dim
- 3D: Assume + is associative; chop k, which is → replication of C matrix

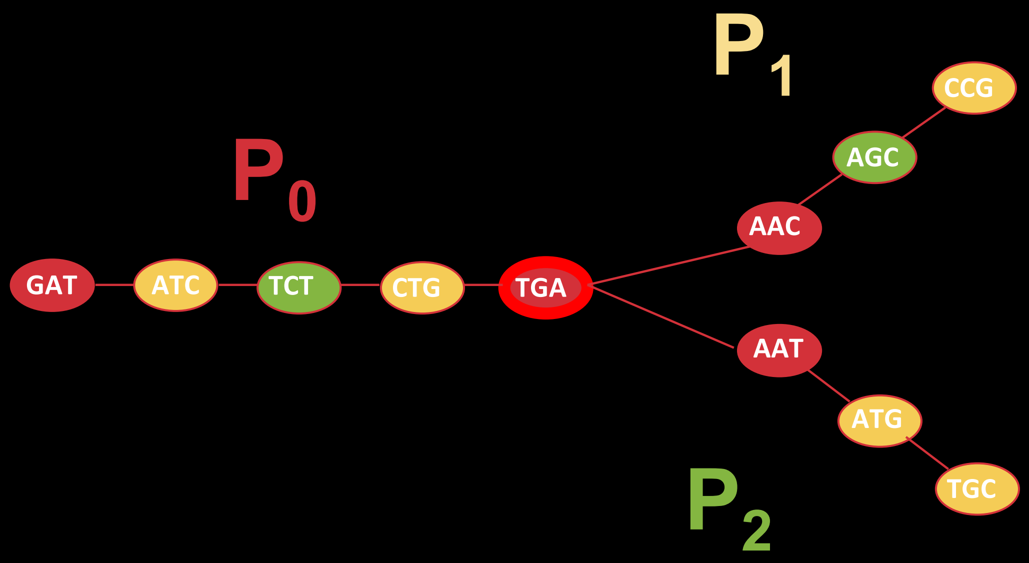
Matrix Multiplication code has a 3D iteration space
Each point in the space is a constant computation (*/+)

```
for i
   for j
      for k
          C[i,j] …  A[i,k] …   B[k,j]  …
```

Solomonik and Demmel, EuroPar 2011

Avoiding Communication in Graph Walk (DFS)

Caching for temporal locality:
depends on graph and layout



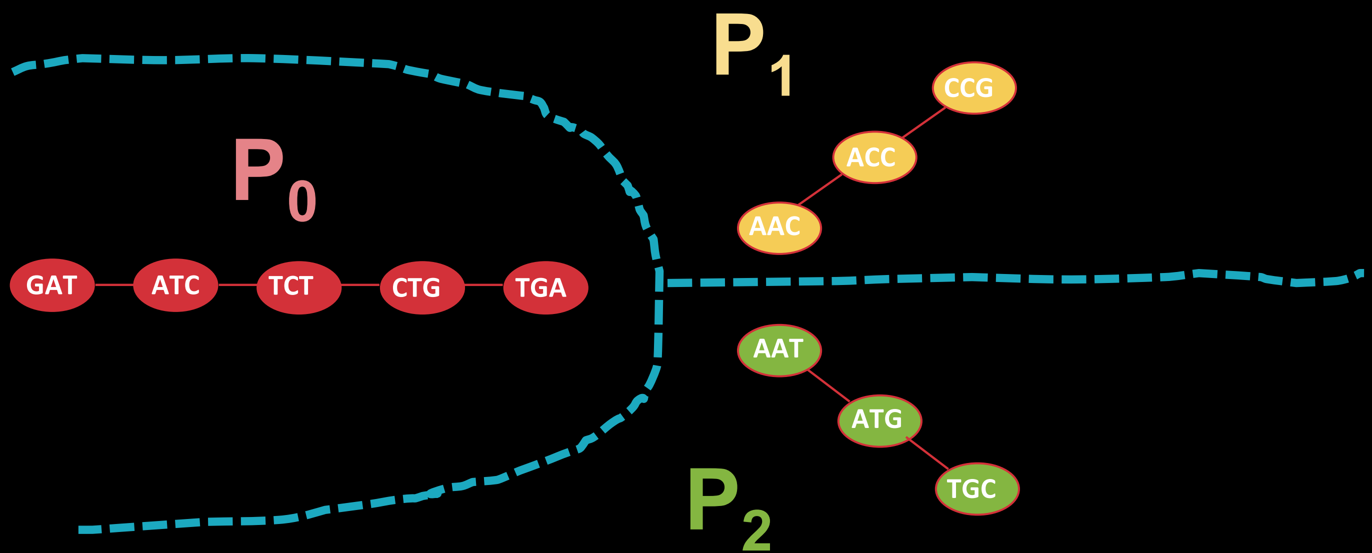
Up to 2.8x faster -- 76% reduction of off-node communication!

Georganas, PhD Thesis and SC18 paper

Avoiding Communication in Graph Walk (DFS)
It's all about the data

Caching for temporal locality: depends on graph and layout

Layout for spatial locality: locality-award hash function, i.e., an "oracle"



Up to 2.8x faster -- 76% reduction of off-node communication!

Georganas, PhD Thesis and SC18 paper

Jump on the AI Bandwagon

If anything will continue to improve in performance, it will be AI

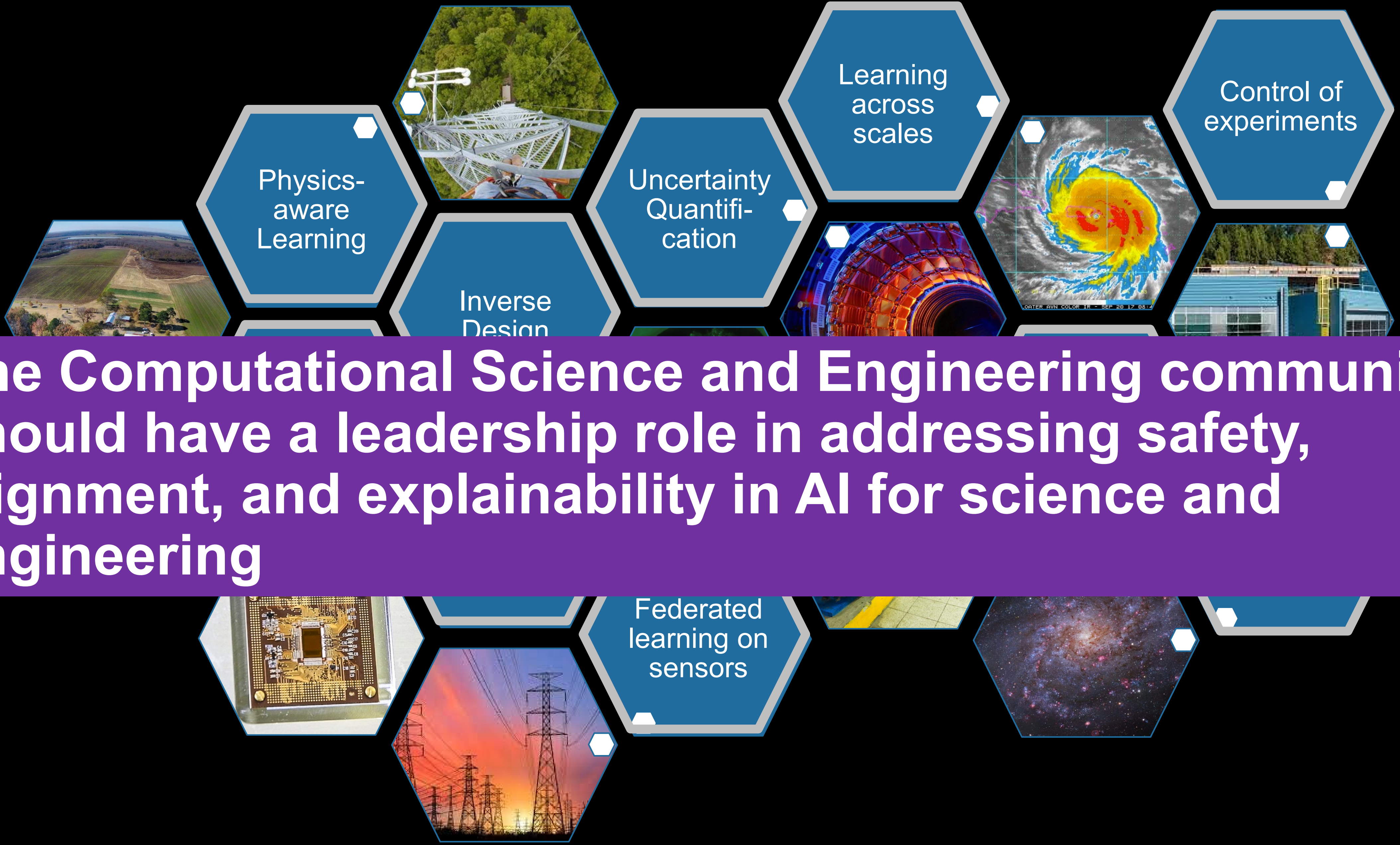
Beat 'em

- Design science processors
- Or co-opt theirs

Join 'em

- Use AI methods for science

AI in Science



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

Five Stages of AI

Acceptance

Denial

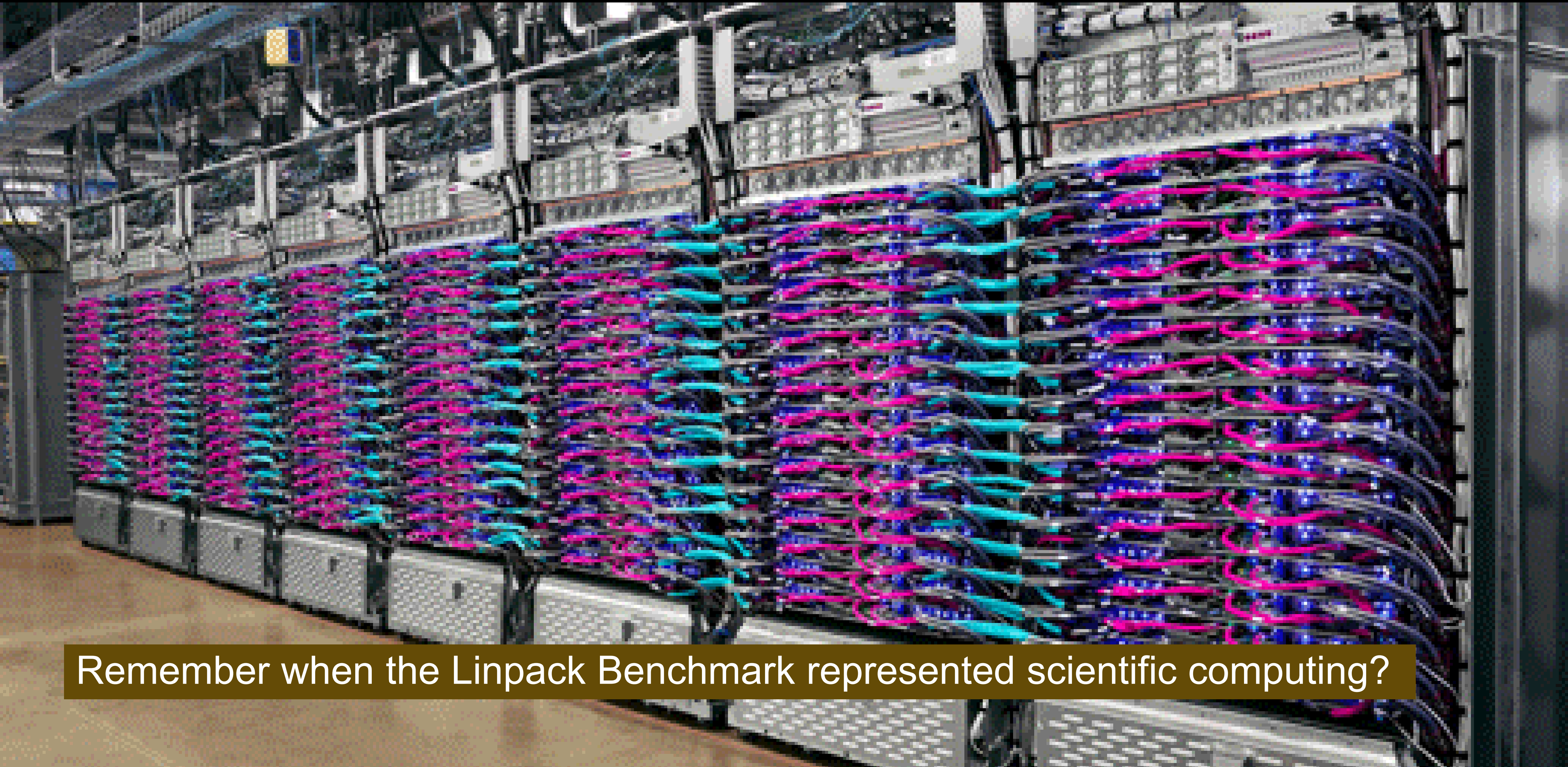
Anger

Depression

Bargaining

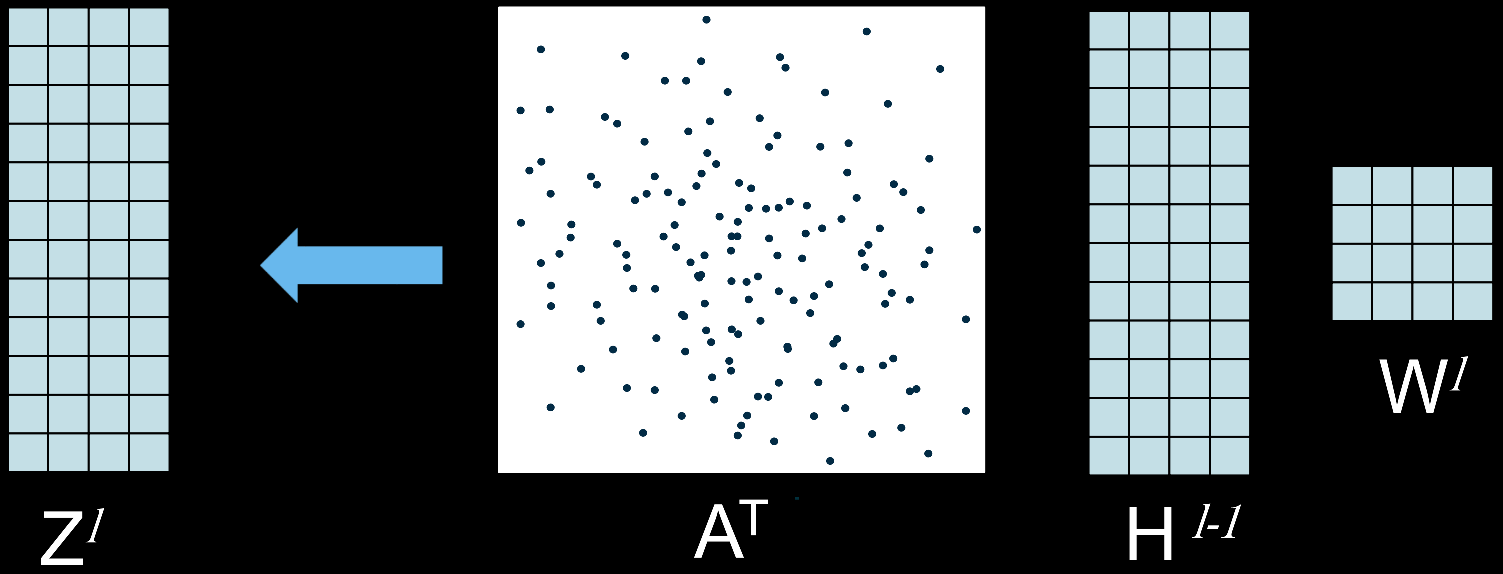
And this includes AI researchers!

Specialization: Are we looking under the lamppost?



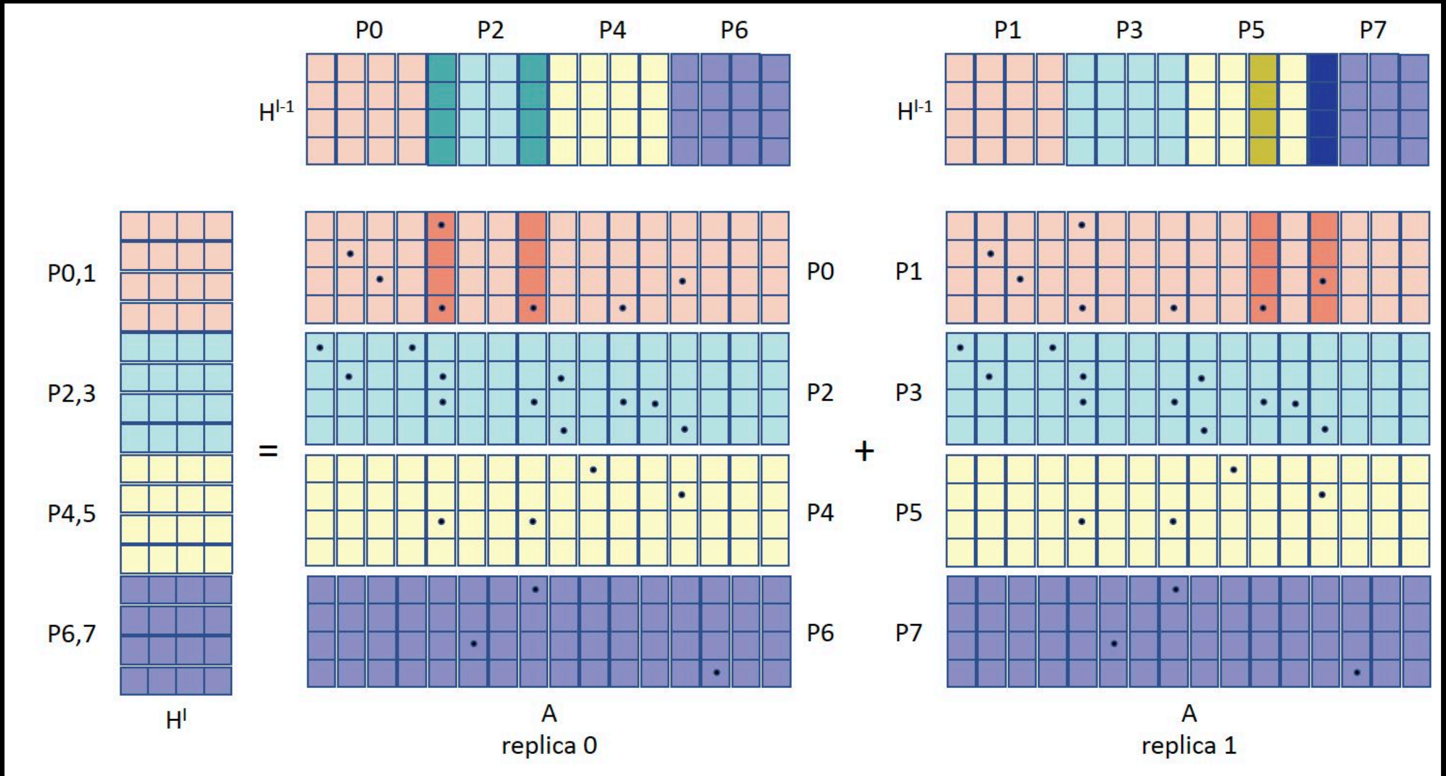
Remember when the Linpack Benchmark represented scientific computing?

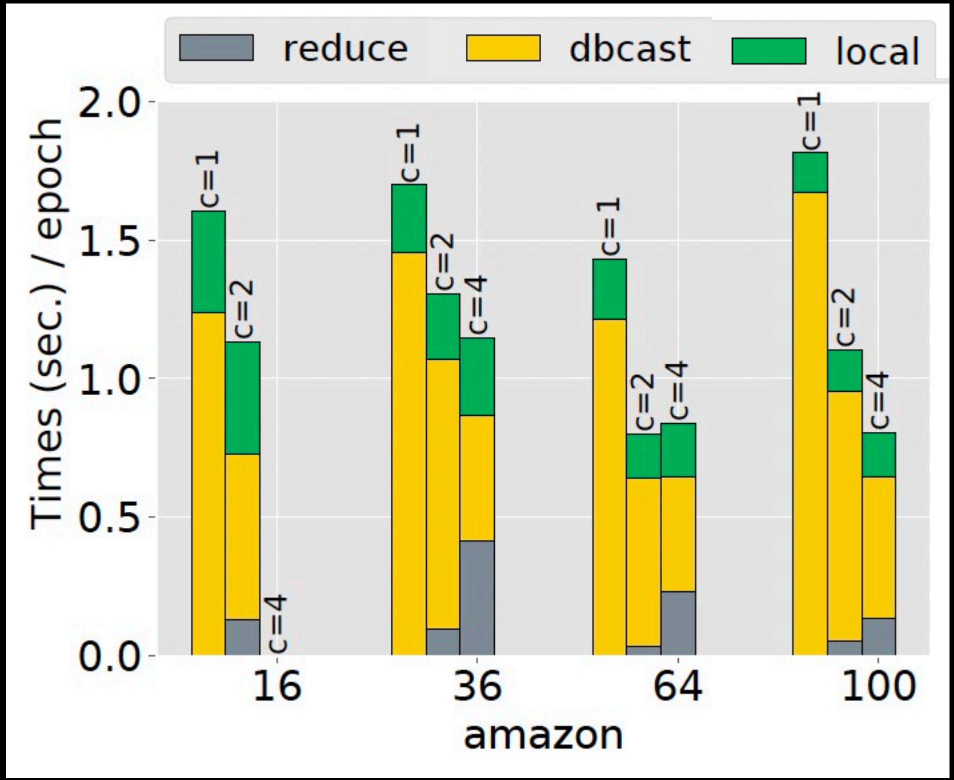
Bottleneck in GNN Training



- $A^T H^{l-1}$ sparse-dense matmul (SpMM)
- $(A^T H^{l-1}) W^l$ dense-dense matmul (DGEMM)
- **SpMM is the bottleneck, not DGEMM!**

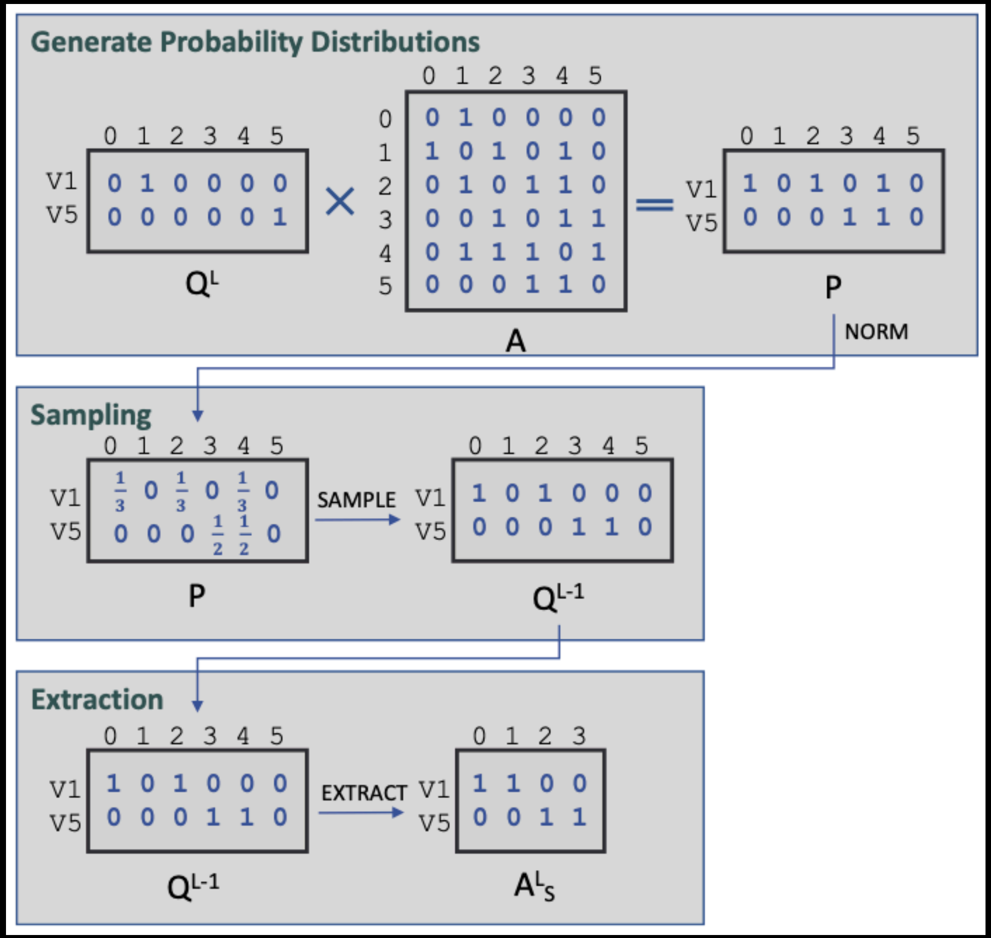
Avoiding Communication in GNNs

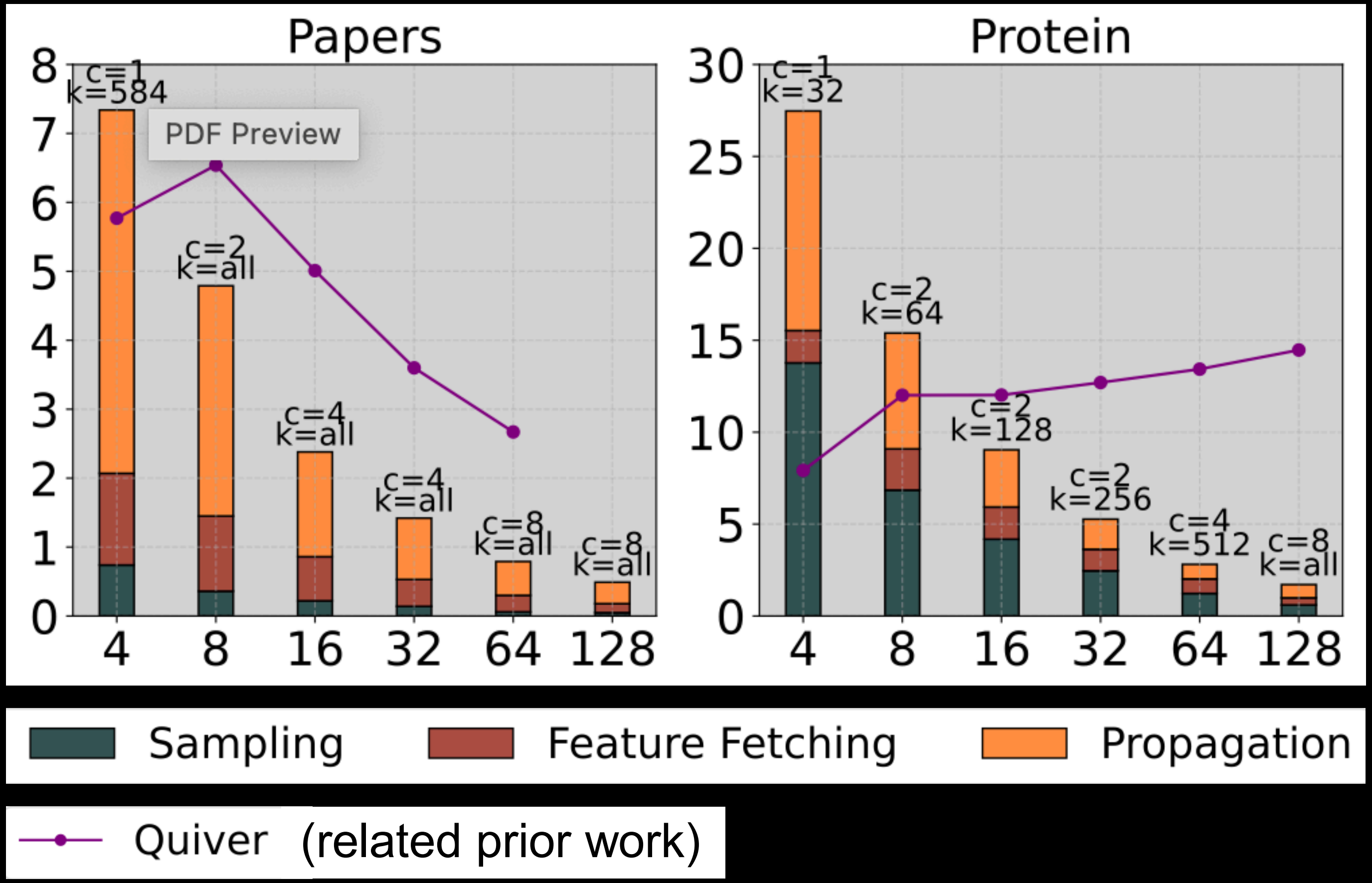




Tripathy, Yelick, Buluc, Reducing Communication in Graph Neural Network Training, SC'20
Mukhopadhyay, Tripathy, Selvitopi, Yelick, Buluc, Sparsity-aware communication …, ICPP '24

Sampling as Communication-Avoiding Sparse Matrix Operations





Distributed Matrix Based Sampling…, Tripathy, Yelick, Buluc, MLSys ‘23

Parallelism Kernels – Dice in the Hardware Lottery

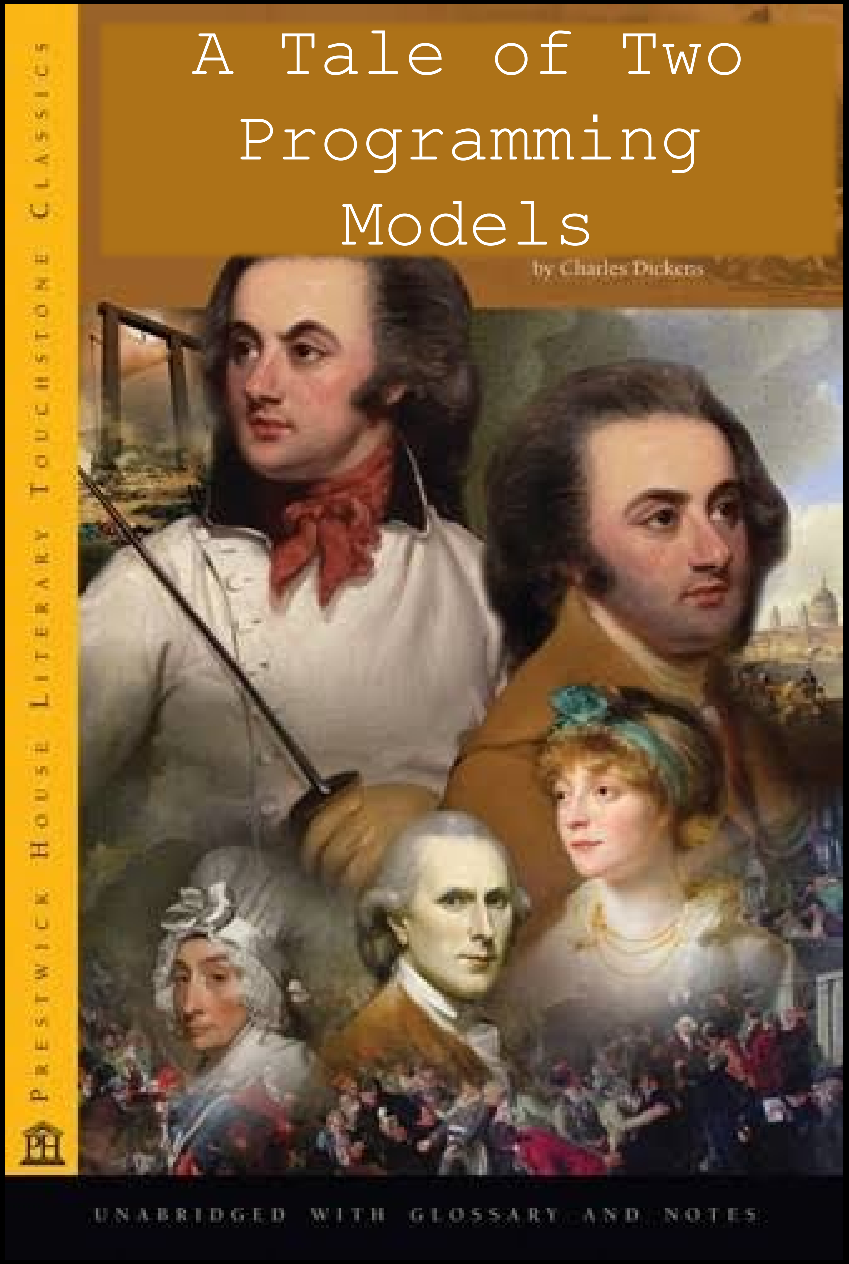
7 Dwarfs of Simulation	7 Giants of Big Data
Particle methods	Generalized N-Body
Unstructured meshes	Graph-theory
Dense Linear Algebra	Linear algebra
Sparse Linear Algebra	Hashing
Spectral methods	Sorting
Structured Meshes	Alignment
Monte Carlo methods	Basic Statistics
Phil Colella	NRC Report + our paper

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

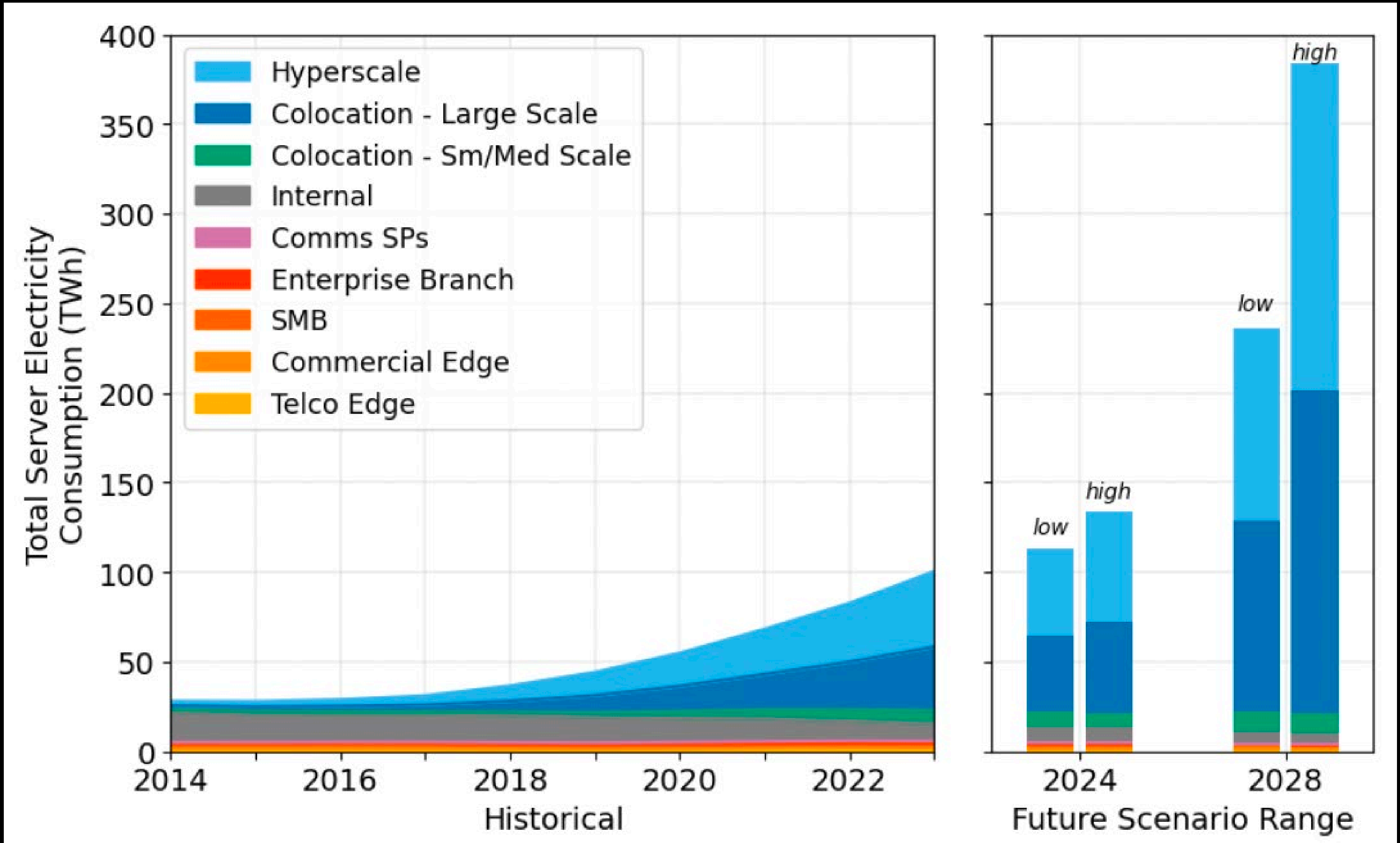
Partitioned Global Address Space

- 1-Sided Communication
- Locality control
- Never cache remote data

- Generalized Sparse Matrices
  - Graph BLAS
  - With Semiring operations

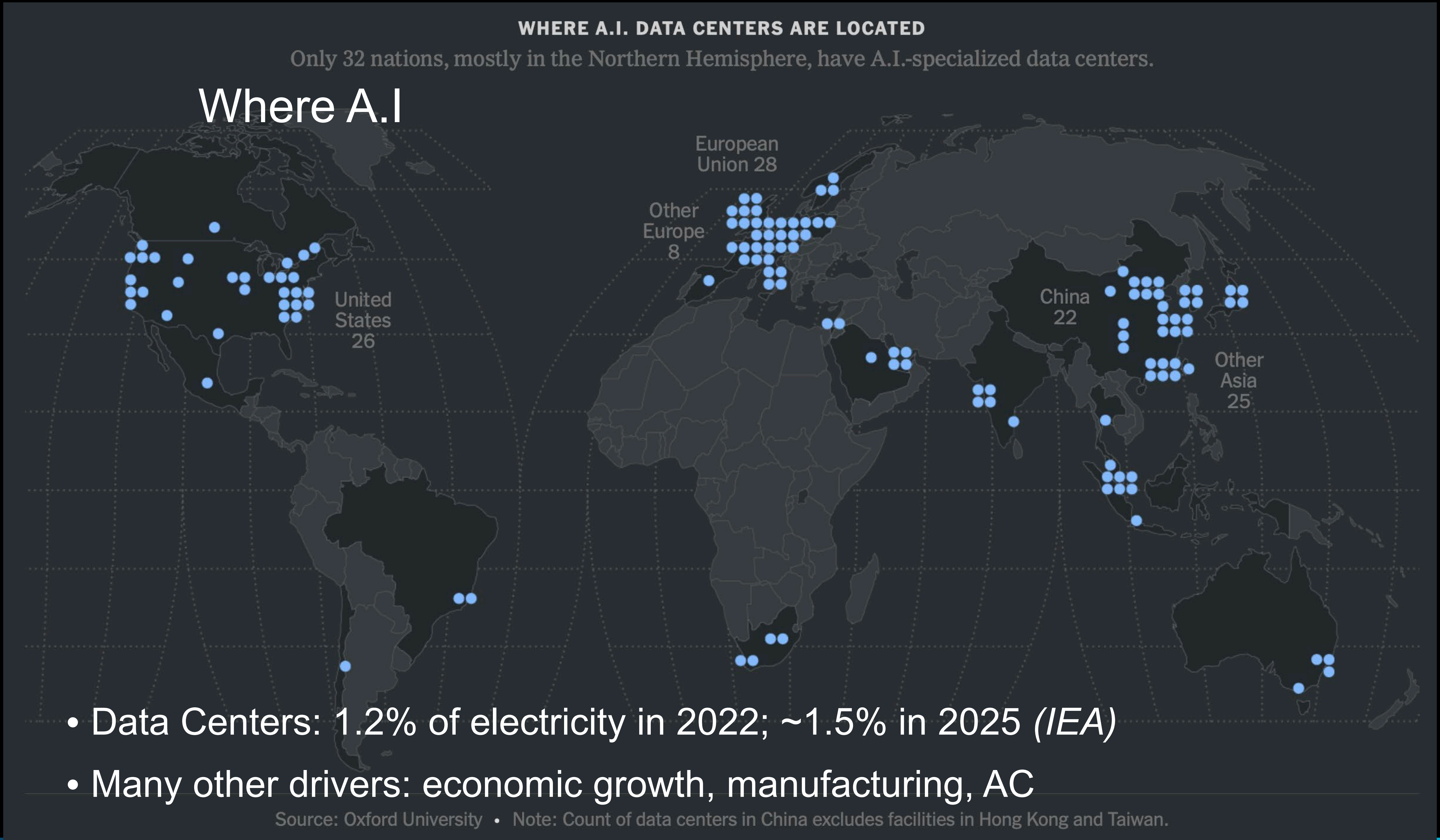


Will Computing and Data Center Power Dominate?



- AI driving growth
- LBNL Reports on US trends

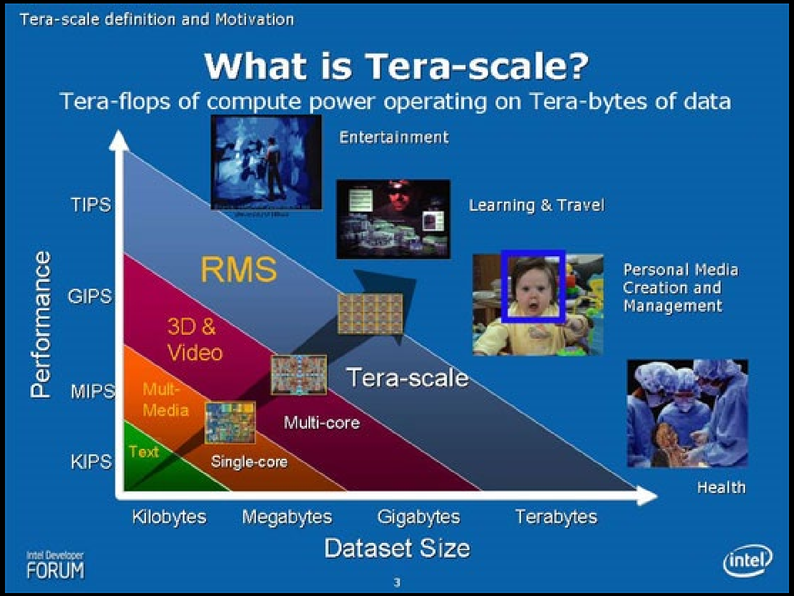
Shehabi et al, 2024 United States Data Center Energy Usage Report, LBNL-2001637

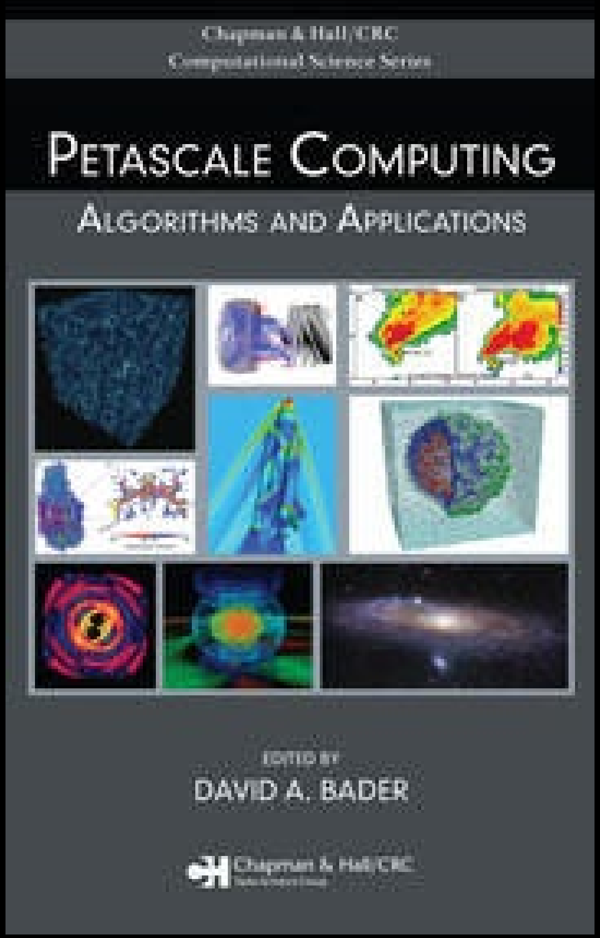


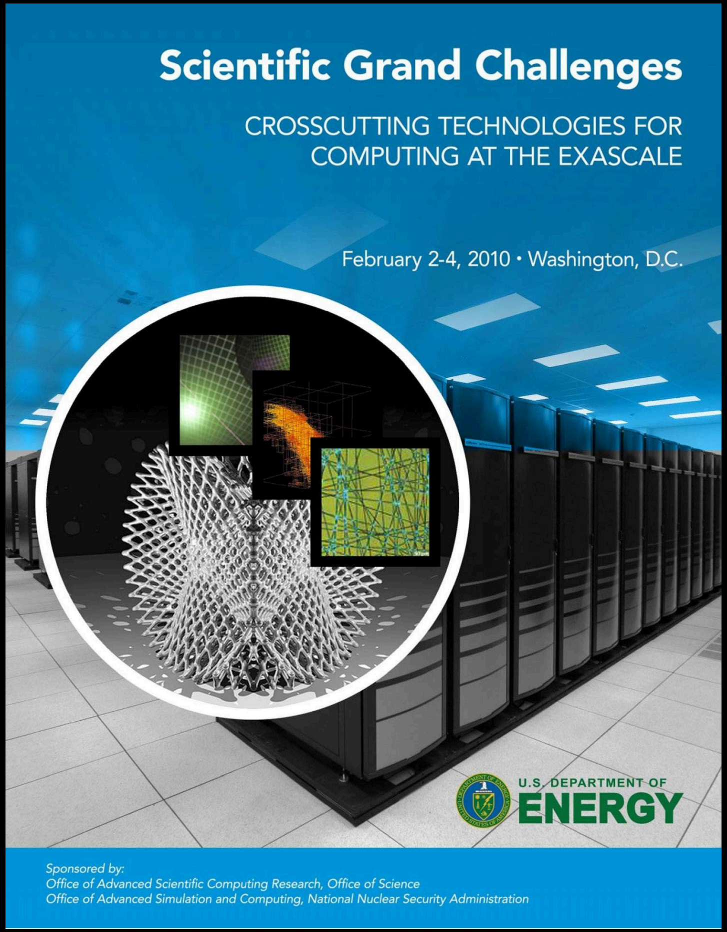
Where A.I

- Data Centers: 1.2% of electricity in 2022; ~1.5% in 2025 *(IEA)*
- Many other drivers: economic growth, manufacturing, AC

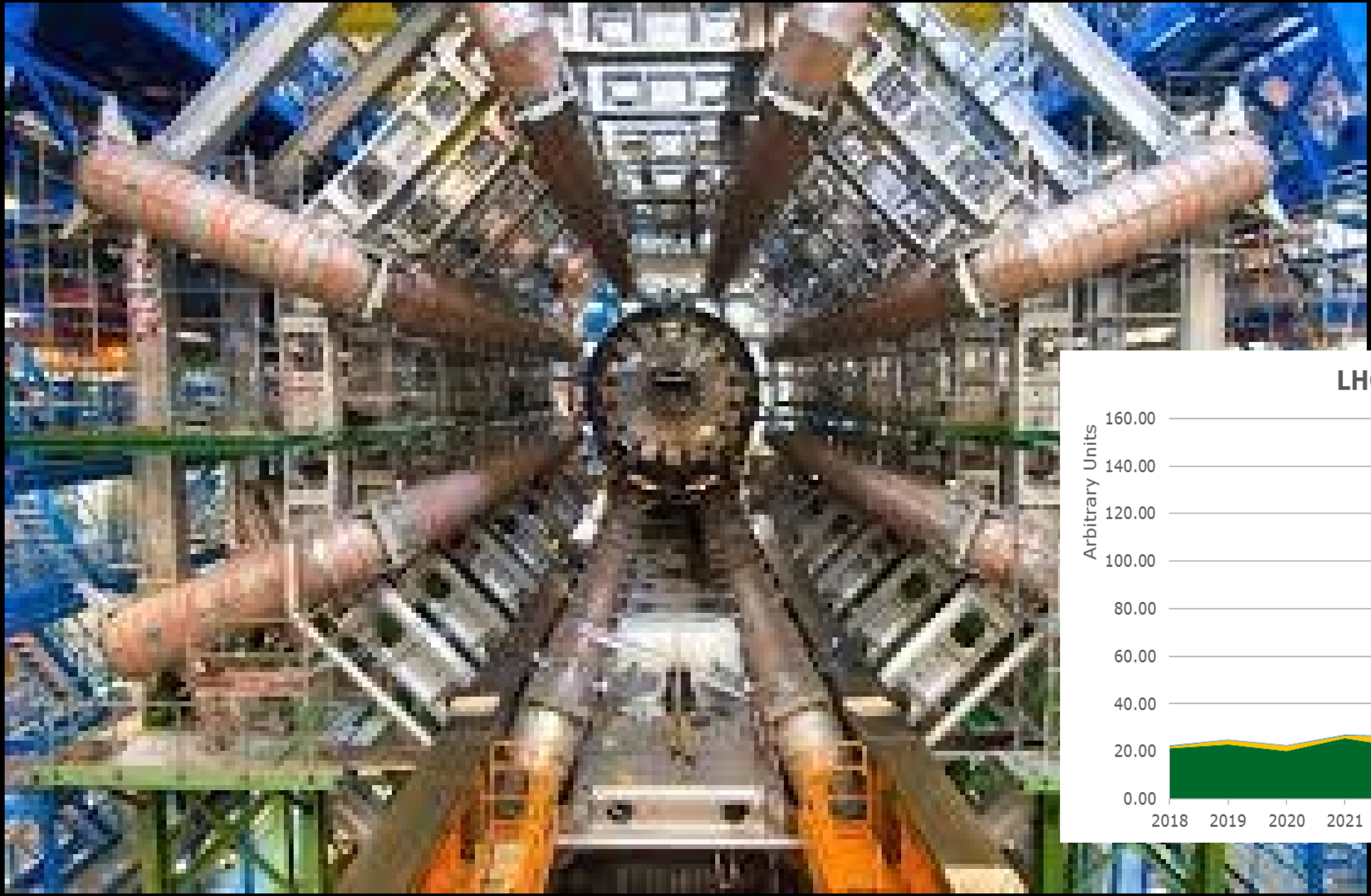
It’s hard to think exponentially

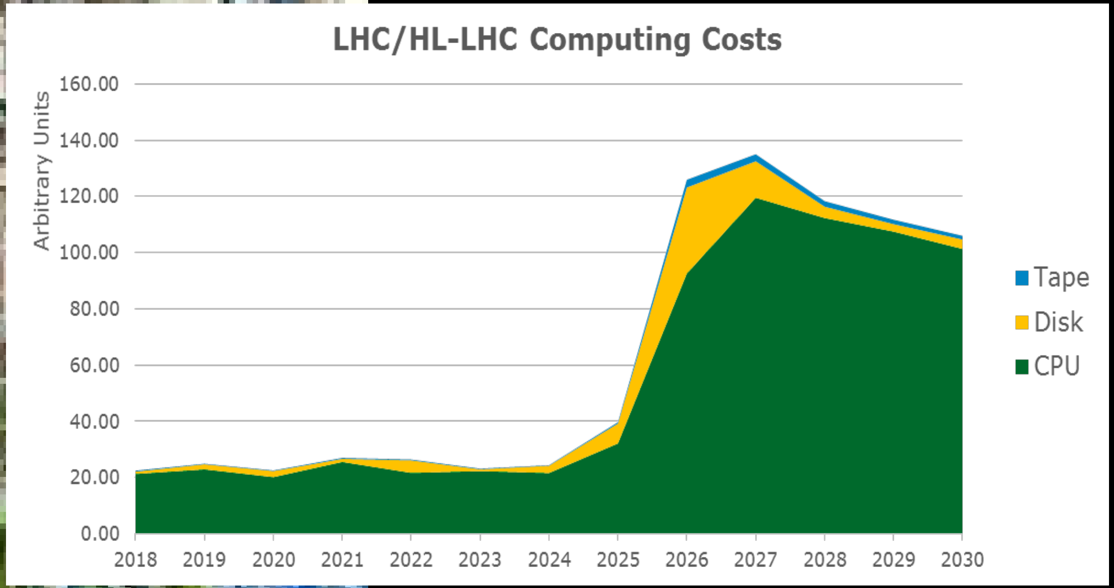






It’s also hard to stop: Prediction of Atlas computing +$1B





Thanks to Moore's Law!

It's been a great run!