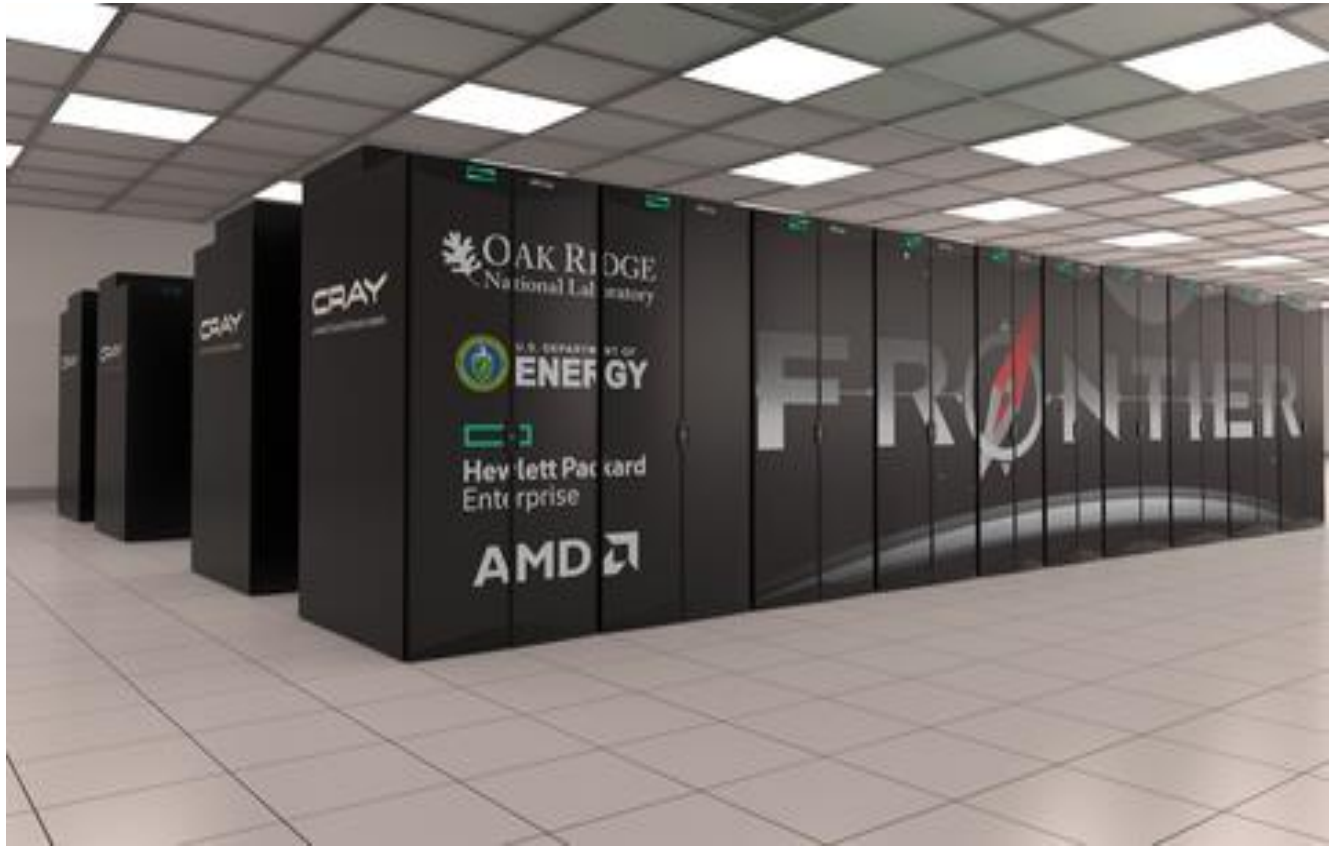
MAX PLANCK
COMPUTING & DATA FACILITY

MPCDF

The Future of HPC - Considering Some Myths

Erwin Laure
29.4.2024

EXASCALE COMPUTING IS HERE (US, CN?)

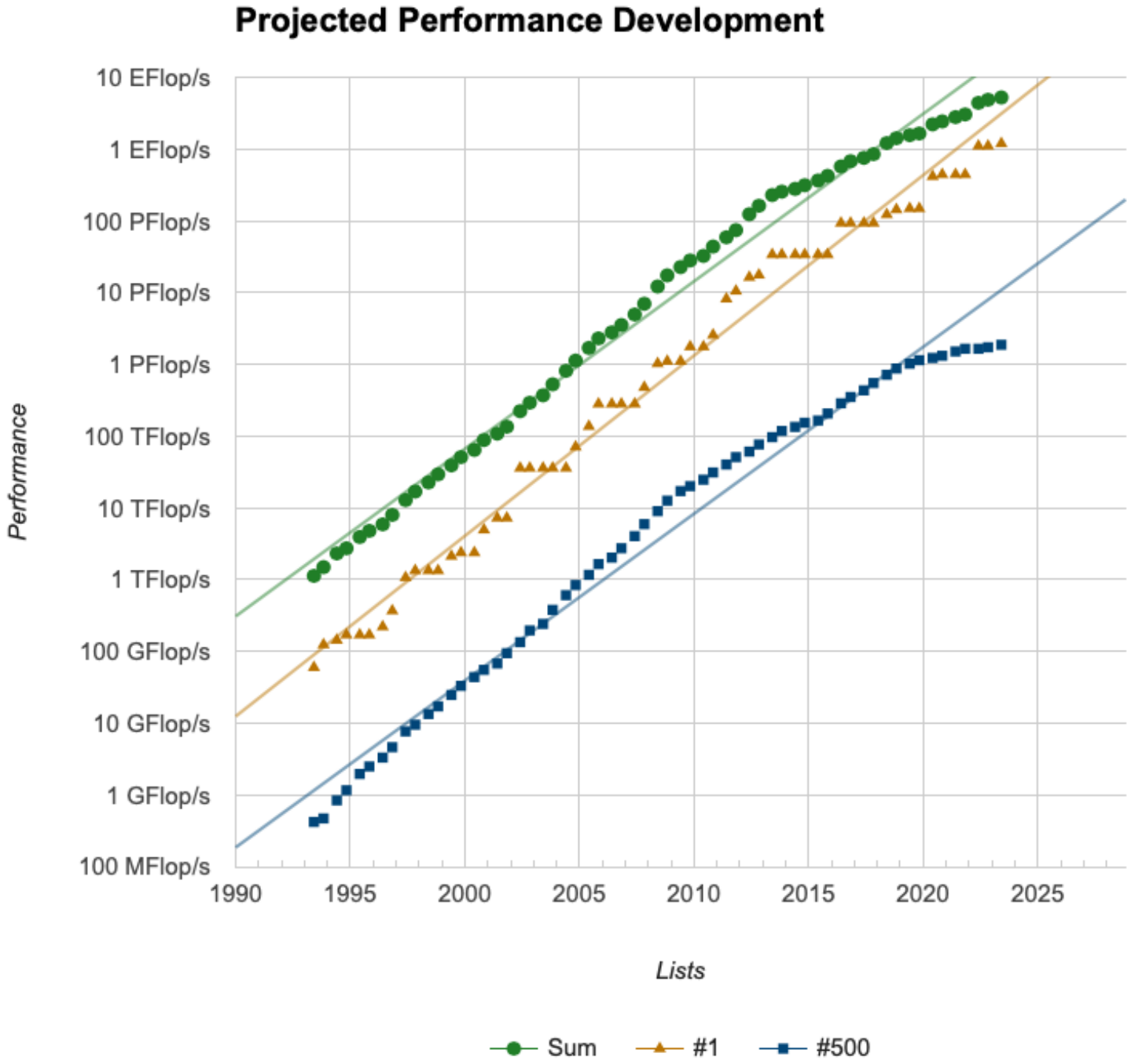


FRONTIER @ OLCF (US): HPE/CRAY

- **AMD EPYC CPUs, AMD MI250 GPUs**
- **8.7 M CPU cores & GPU compute units**
- **52 GFlop/s/W**
- **1102 PFlop/s HPL, rank 1 in Top500 11/2022**

AND SOON IN EUROPE …

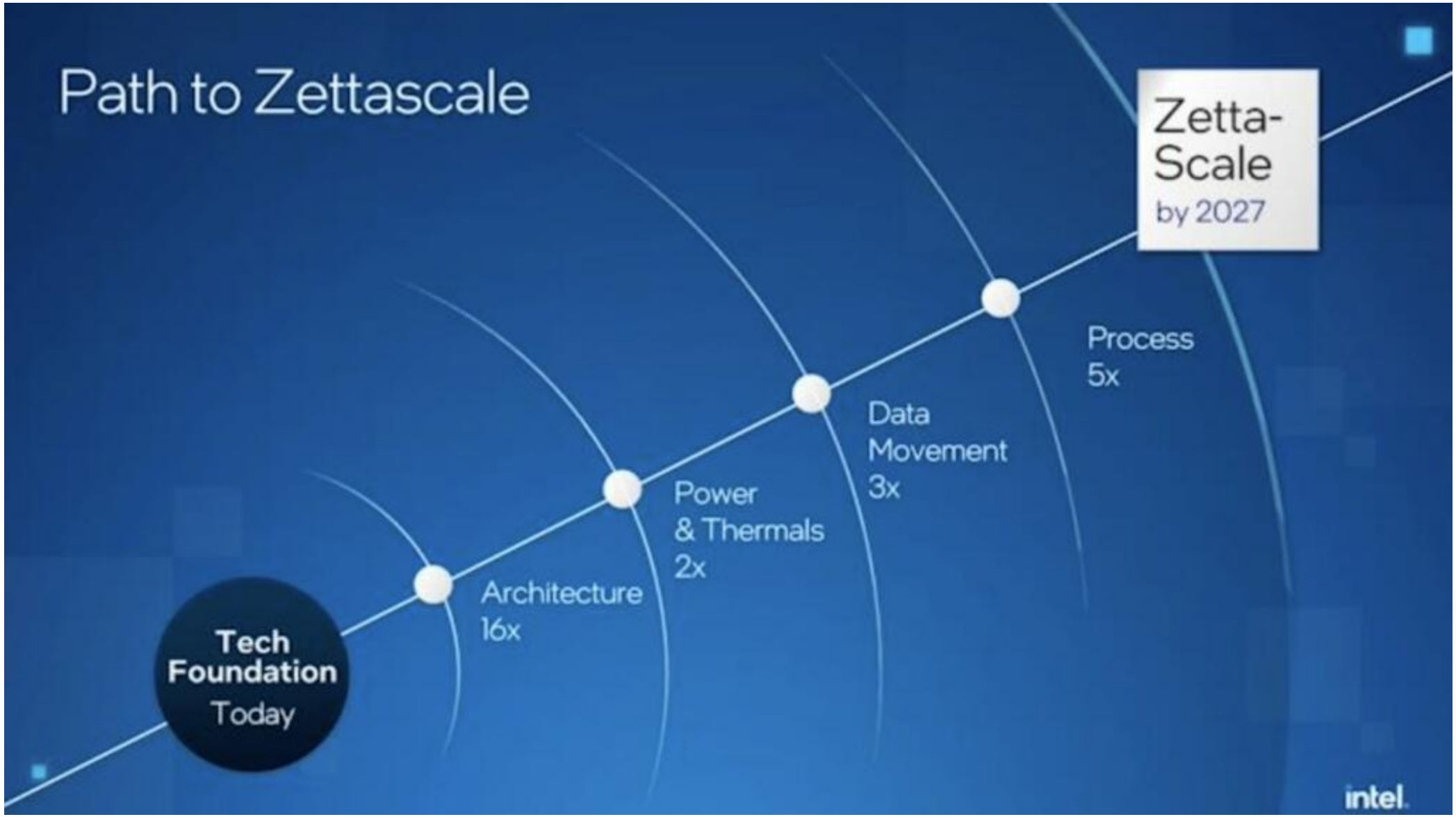
- 1 more pre-exascale @ BSC, 2023
- “Jupiter”@Jülich will be the first European Exascale system (500 M€), 2024



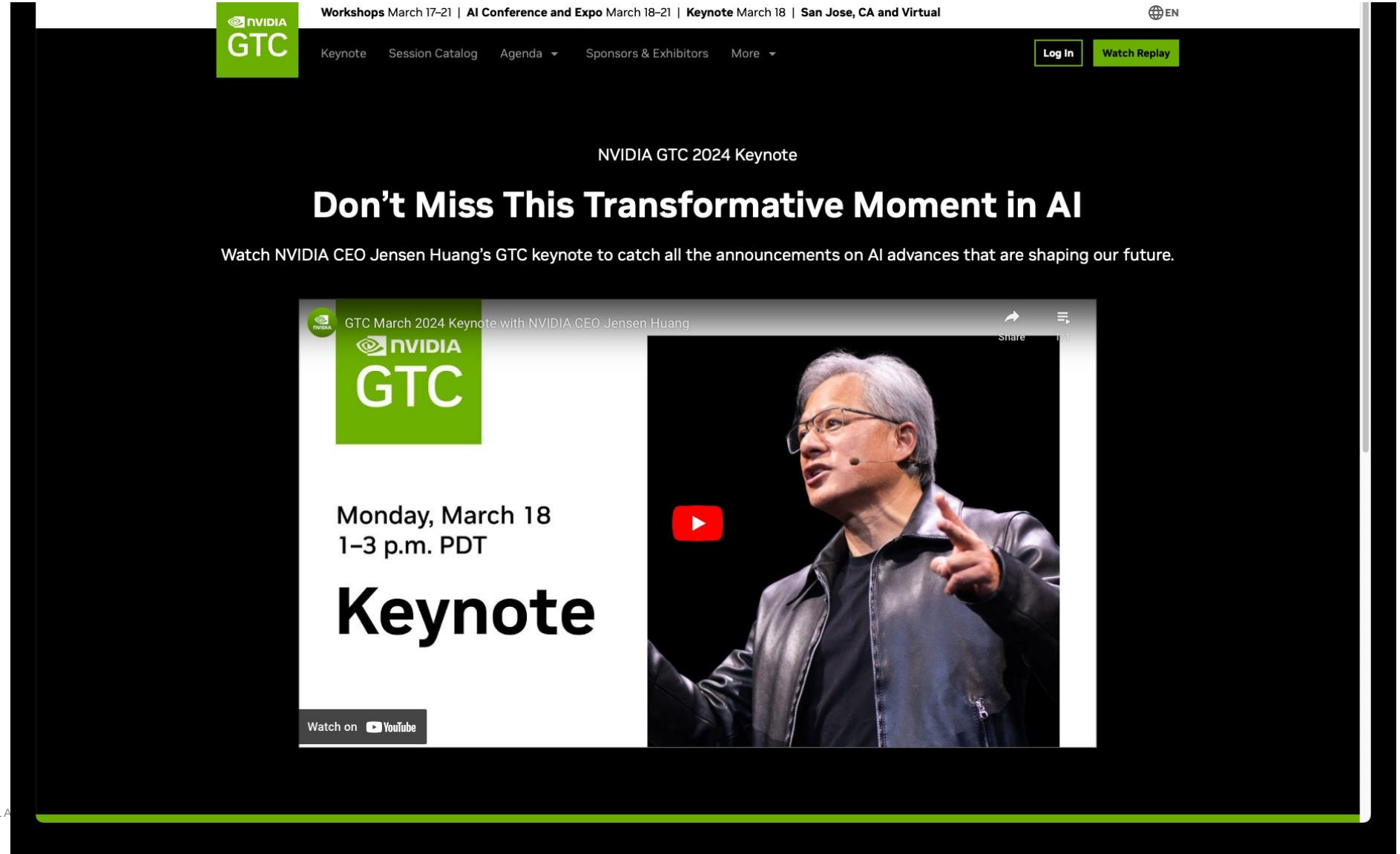
- Globally, the rate of performance increase is diminishing
- Some jumps in Top1 still expected, but what then?

WHAT'S NEXT – THE ZETTAFLOPS SUPERCOMPUTER?

from: https://www.nextbigfuture.com/2023/02/intel-and-amd-path-to-zettaflop-supercomputers.html

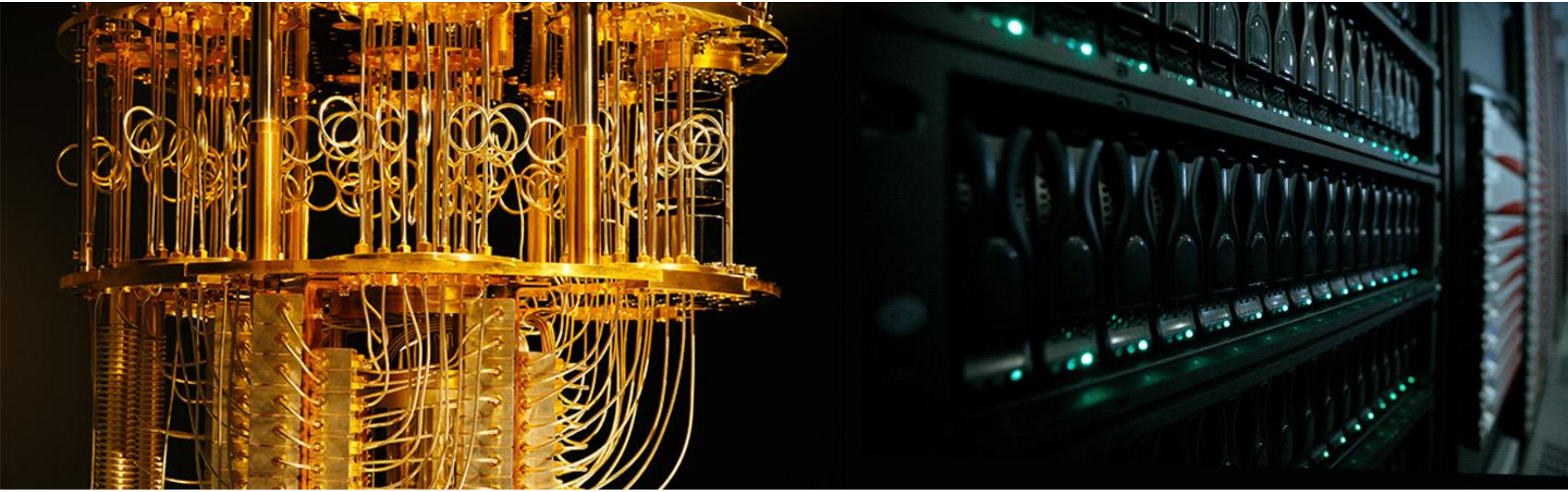


WHAT'S NEXT – AI WILL TAKE OVER?



WHAT'S NEXT – QUANTUM?

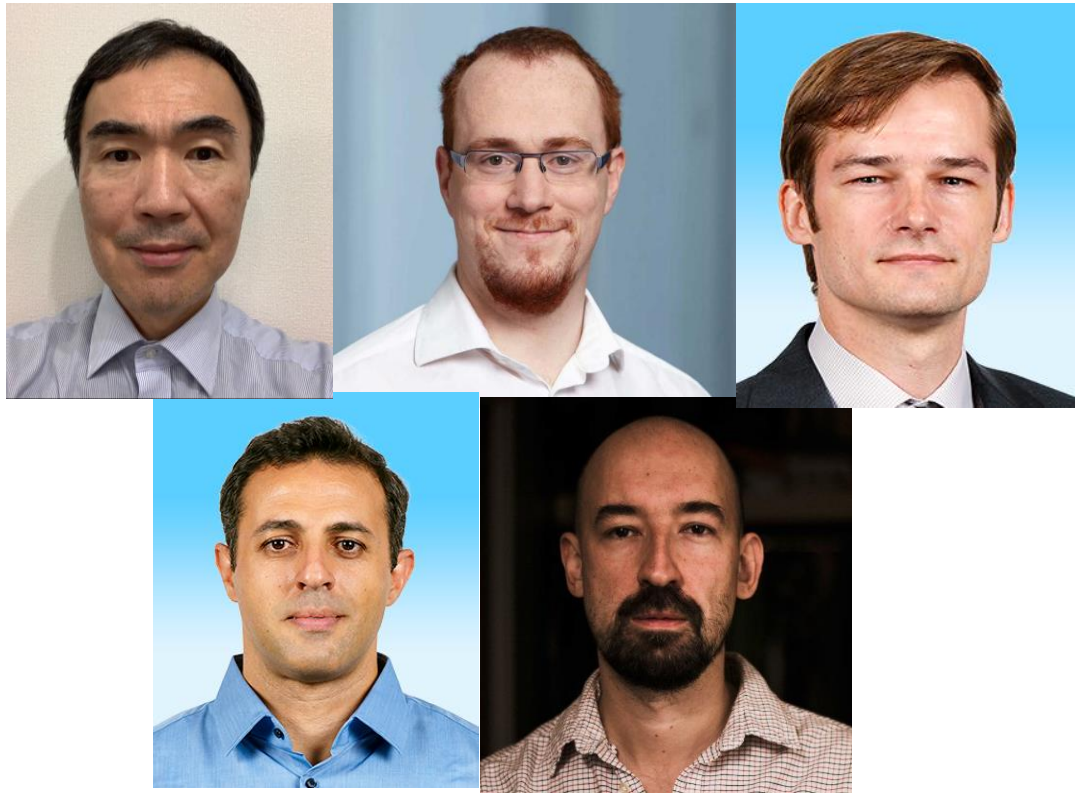
from: https://www.itwm.fraunhofer.de/en/departments/hpc/quantum-computing.html



12 “MYTHS” IN HPC

S. Matsuoka, J. Domke, M. Wahib, A. Drozd, and T. Höfler

https://doi.org/10.48550/arXiv.2301.02432



Myths and Legends in High-Performance Computing

Satoshi Matsuoka[1], Jens Domke[1], Mohamed Wahib[1], Aleksandr Drozd[1], Torsten Hoefler[2]

Abstract

In this humorous and thought provoking article, we discuss certain myths and legends that are folklore among members of the high-performance computing community. We collected those myths from conversations at conferences and meetings, product advertisements, papers, and other communications such as tweets, blogs, and news articles within (and beyond) our community. We believe they represent the zeitgeist of the current era of massive change, driven by the end of many scaling laws such as Dennard scaling and Moore’s law. While some laws end, new directions open up, such as algorithmic scaling or novel architecture research. However, these myths are rarely based on scientific facts but often on some evidence or argumentation. In fact, we believe that this is the very reason for the existence of many myths and why they cannot be answered clearly. While it feels like there should be clear answers for each, some may remain endless philosophical debates such as the question whether Beethoven was better than Mozart. We would like to see our collection of myths as a discussion of possible new directions for research and industry investment.

Keywords

Quantum; zettascale; deep learning; clouds; HPC myths

Introduction

Any human society has their myths and legends—this also applies to the high-performance computing (HPC) community. HPC drives the largest and most powerful computers and latest computing and acceleration technologies forward. One may think that it’s scientific reasoning all the way down in such an advanced field. Yet, we find many persistent myths revolving around trends of the moment.

Myth 1: Quantum Computing Will Take Over HPC!

Numerous articles are hyping the quantum computing revolution affecting nearly all aspects of life ranging from quantum artificial intelligence to even quantum gaming. The whole IT industry is following the quantum trend and conceives quickly growing expectations. The actual development of quantum technologies, algorithms, and use-cases is on a very different time-scale. Most practitioners would not expect quantum computers to outperform classical

THE 12 ‘MYTHS’ IN HPC

Myth 1: Quantum Computing Will Take Over HPC!

Myth 2: Everything Will Be Deep Learning!

Myth 3: Extreme Specialization as Seen in Smartphones Will Push Supercomputers Beyond Moore’s Law!

Myth 4: Everything Will Run on Some Accelerator!

Myth 5: Reconfigurable Hardware Will Give You 100X Speedup!

Myth 6: We Will Soon Run at Zettascale!

Myth 7: Next-Generation Systems Need More Memory per Core!

Myth 8: Everything Will Be Disaggregated!

Myth 9: Applications Continue to Improve, Even on Stagnating Hardware!

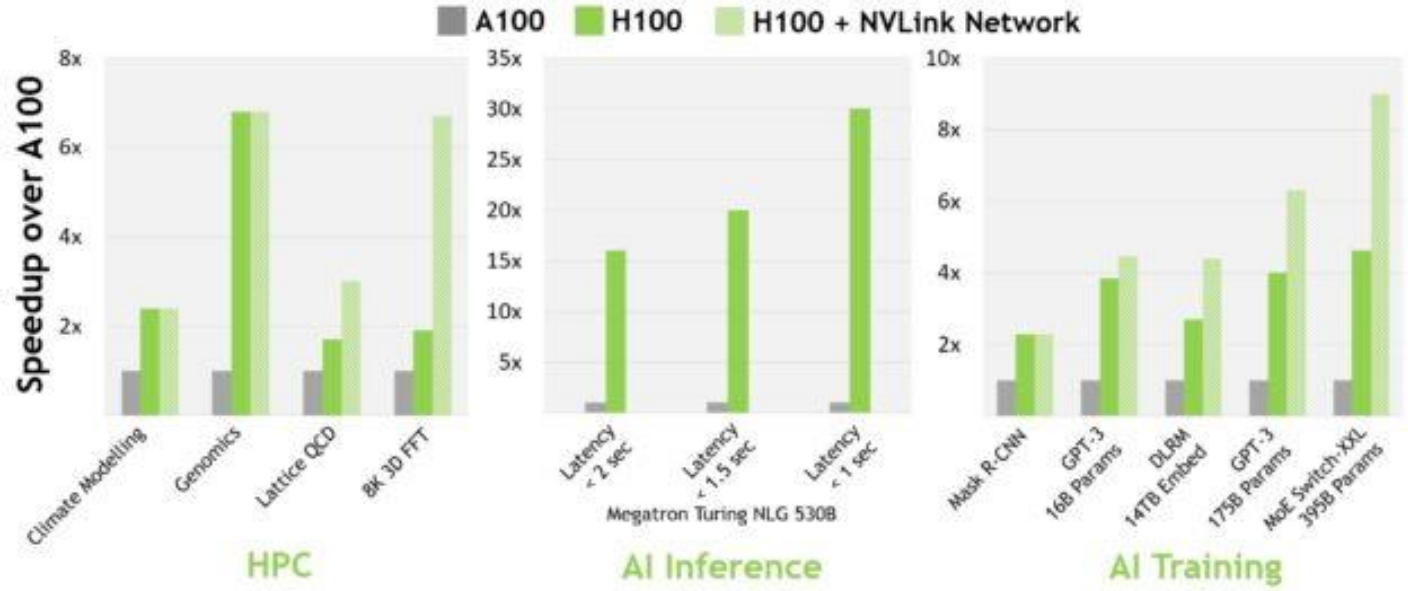
Myth 10: Fortran Is Dead, Long Live the DSL!

Myth 11: HPC Will Pivot to Low or Mixed Precision!

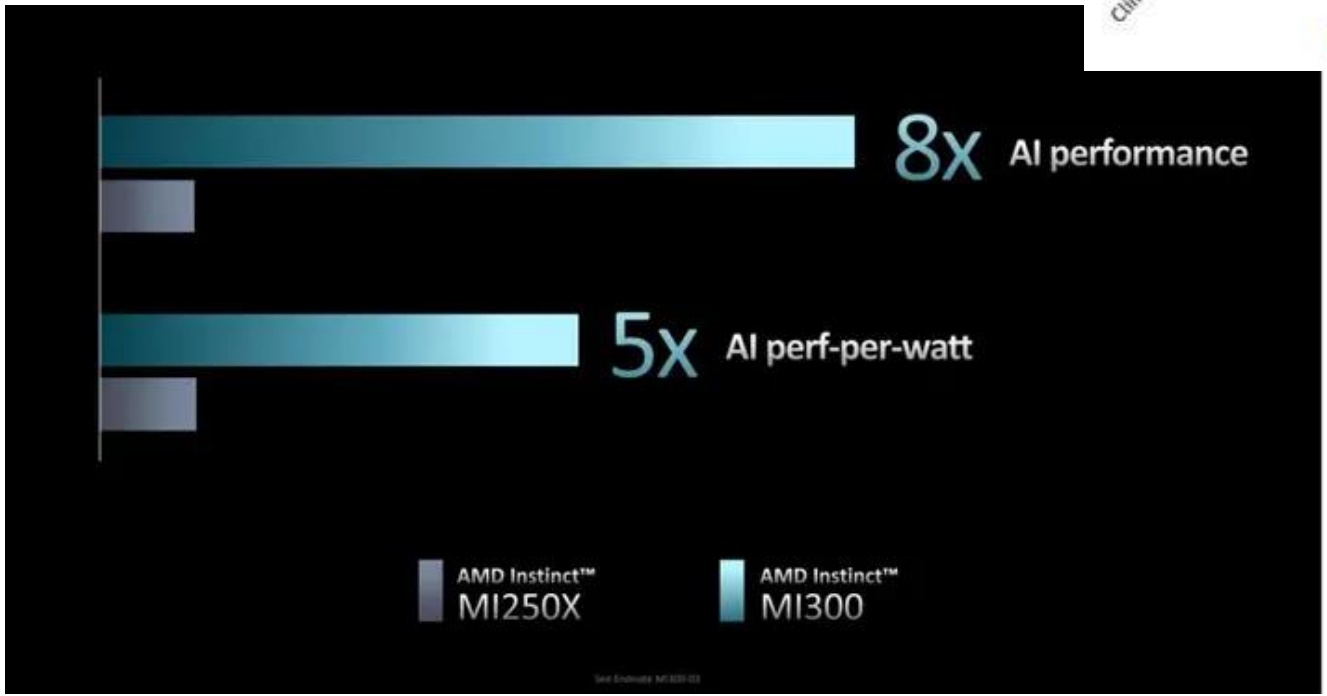
Myth 12: All HPC Will Be Subsumed by the Clouds!

MYTH 9: APPLICATIONS CONTINUE TO IMPROVE EVEN ON STAGNATING HARDWARE

STAGNATING HARDWARE?

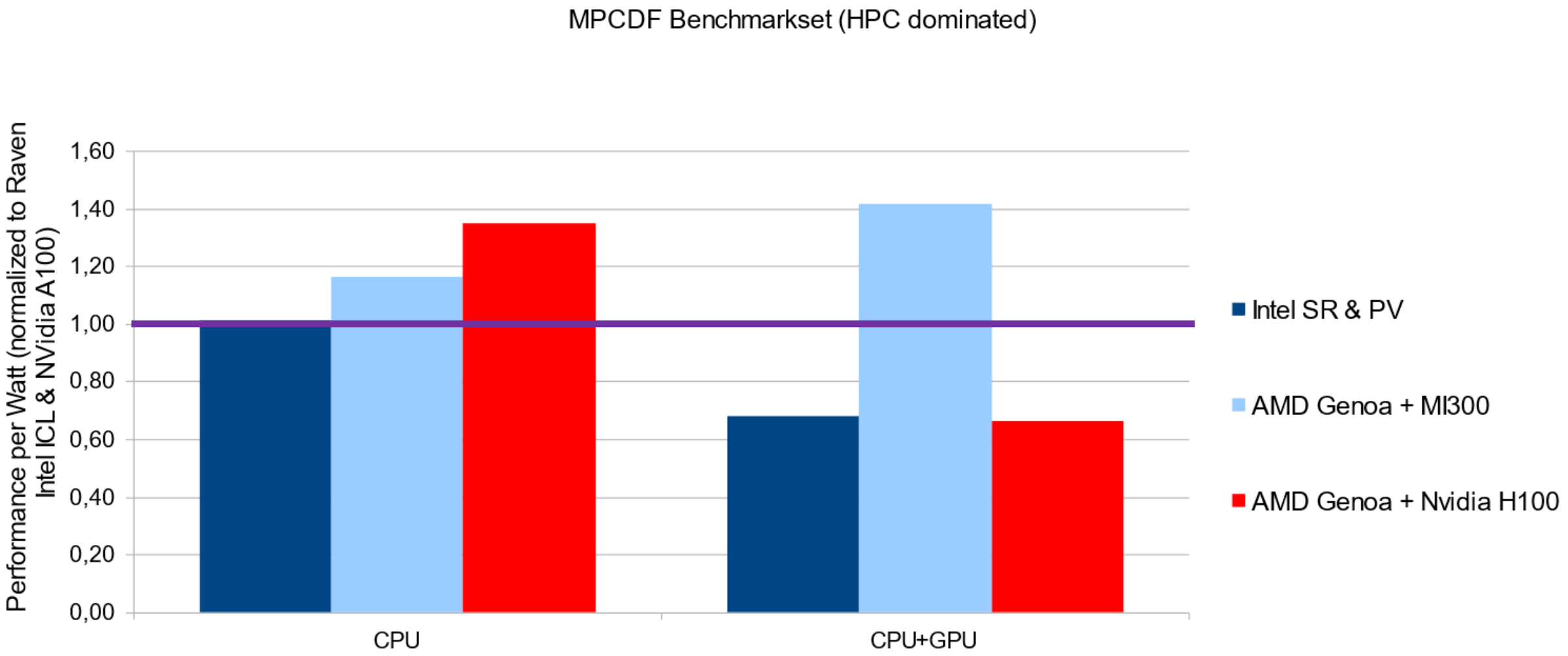


source: Nvidia



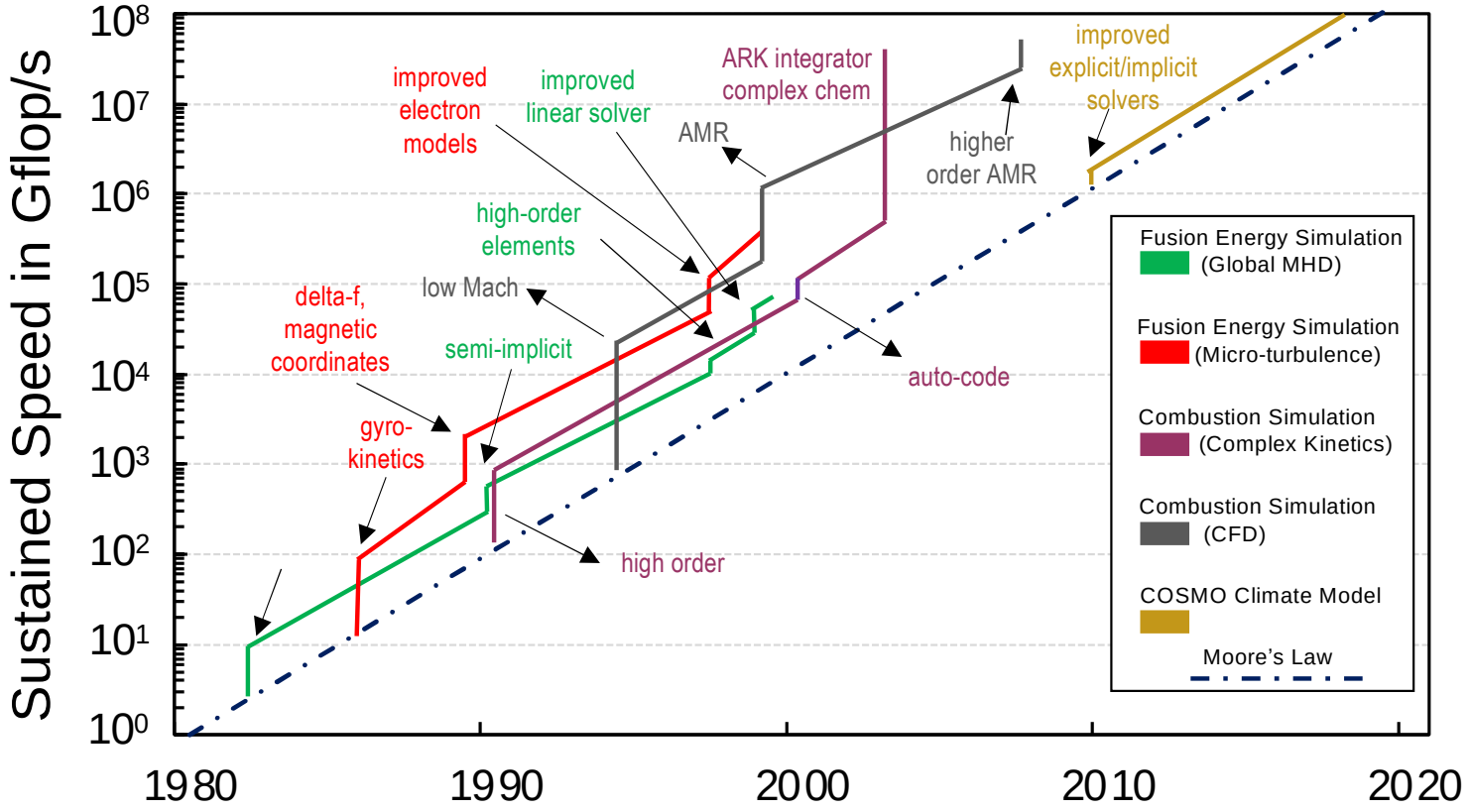
source: AMD

AN INCONVENIENT TRUTH



ALGORITHMIC MOORE'S LAW

arXiv preprints



- Should we dramatically increase investments in software?
- Will the "Algorithmic Moore's Law" end soon as well?
- Are we willing to refactor/rewrite legacy codebases?

Figure 3. Examples of "Algorithmic Moore's Law" for different areas in HPC; Fusion energy and combustion simulations data by Keyes (2022) and climate simulation data by Schulthess (2016)

https://doi.org/10.48550/arXiv.2301.02432

MYTH 4: EVERYTHING WILL RUN ON SOME ACCELERATOR!

LARGE UNEXPLORED TERRITORY – WILL IT BE TAKEN UP?

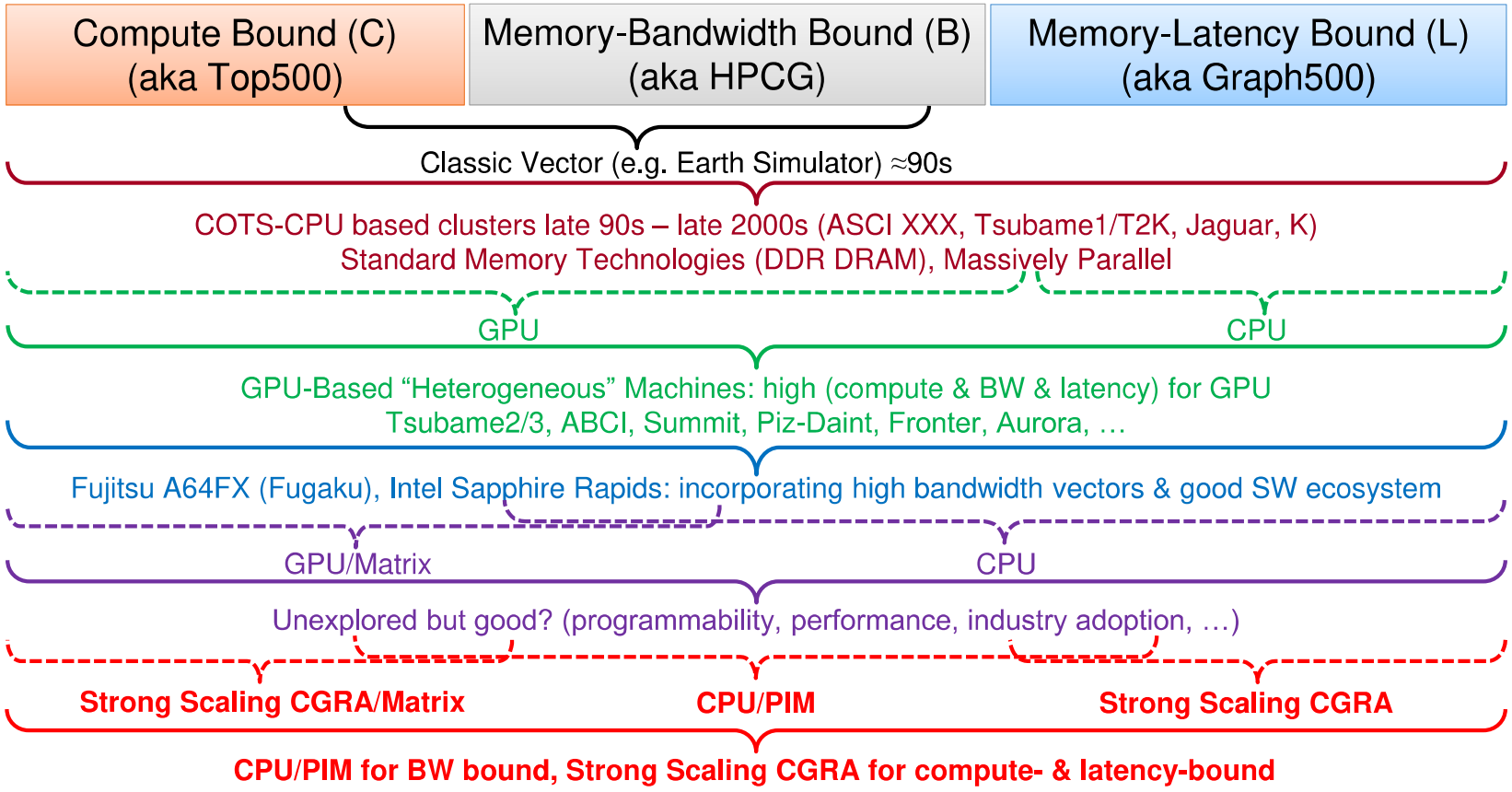


Figure 1. Classification of Compute Kernels and Supercomputing Architecture

https://doi.org/10.48550/arXiv.2301.02432

MYTH 2: EVERYTHING WILL BE DEEP LEARNING!

Original slide courtesy Rick Stevens, ANL, 2023

AI for Science w/HPC Foundations and Applications

Surrogate Models

- Parameter Search
- Augmented Simulations
- Optimization

Foundational Models for Science

- Distillation
- Integration
- Protocols
- Writing
- Synthesis
- Q+A

Developing Scientific Codes

- Debugging
- Optimization
- Security
- Translation
- Coding

Controlling complex systems and simulations of digital twins

- Surgery
- Simulation
- Reactors
- Mobility
- Manufacturing

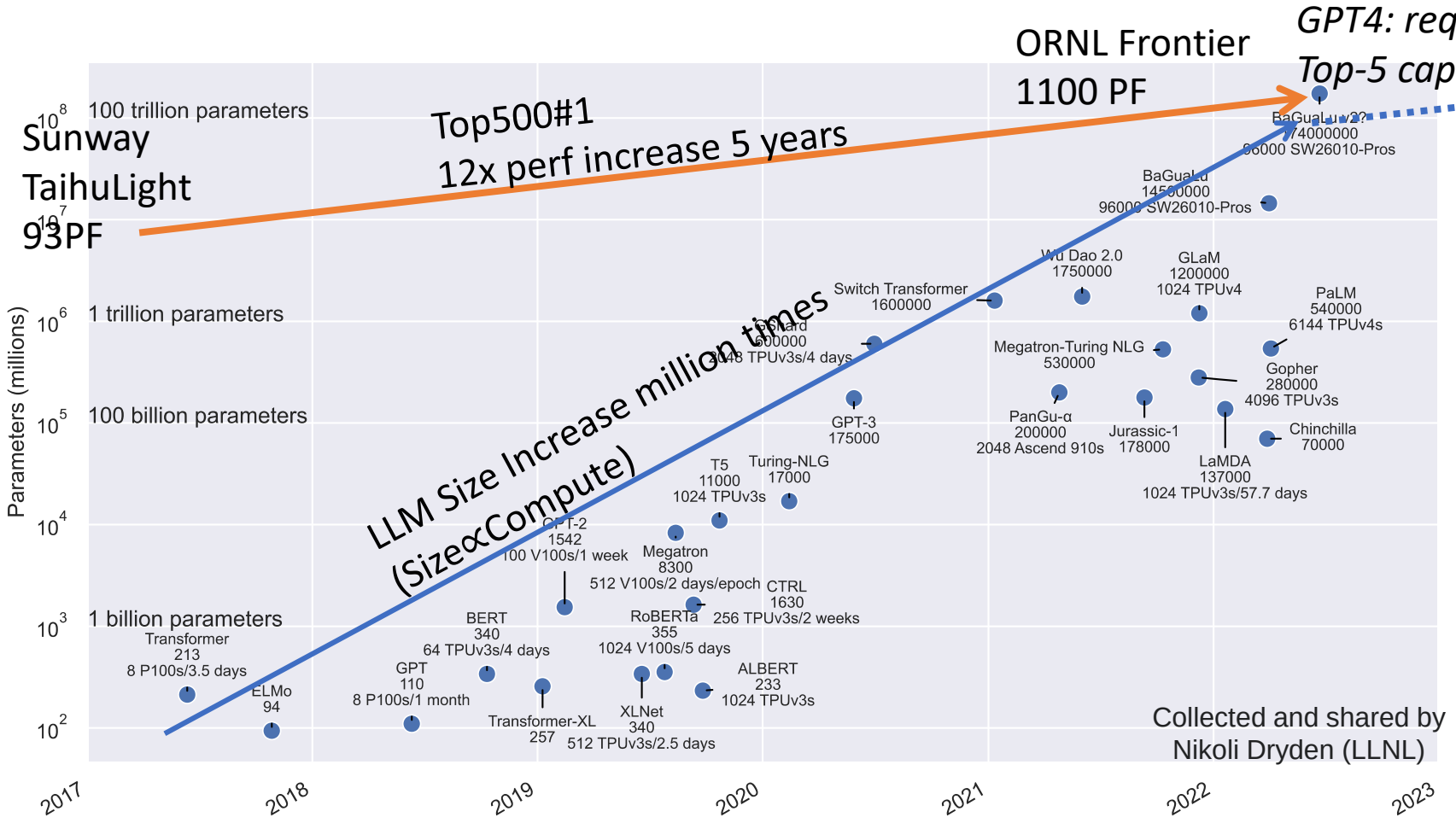
Design & Optimization

- Materials
- Proteins
- Devices
- Industrial Structures

AGI Scientist w/Foundational Models

- Experiments& Simulations Planning
- Instruments & HPC Control
- Discovery & Evaluation

AI Training is Now the Forefront of High End HPC (And thus Free Ride on HPC is Over)



1,000,000x in 5 years!

slide curtesy Satoshi Matsuoka

- The rapid evolution of large language models (LLMs) leading up to GPT-4 can be attributed to scaling, which in turn has been supported by "free ride" or "low-hanging fruit" advancements in supercomputer technologies, such as weak scaling, low-precision arithmetic in GPUs, matrix multiplication engines, high-bandwidth memory (HBM), and high bandwidth interconnects, etc.
- Coincidentally, the GPT-3.5/4.0 revolution occurred when utilizing computational resources equivalent to those of top-tier supercomputers.
- The development of models eg GPT-5 will slow down as the era of "free ride“ ending, causing progress to be proportional to the evolution of supercomputers.
- Moving forward, it is important to focus on research in large-scale supercomputer AI systems, along with how to incorporate domain-specific knowledge in the foundational models

MAX PLANCK
COMPUTING & DATA FACILITY

MPCDF

A FEW EXAMPLES OF AI IN "CLASSICAL" (E)SCIENCE

From Collaborations of MPCDF with various Max Planck Institutes

PREDICION OF 3D PROTEIN STRUCTURES

ENABLING AI SYSTEMS IN COMPUTATIONAL BIOLOGY FOR A BROAD USER BASE

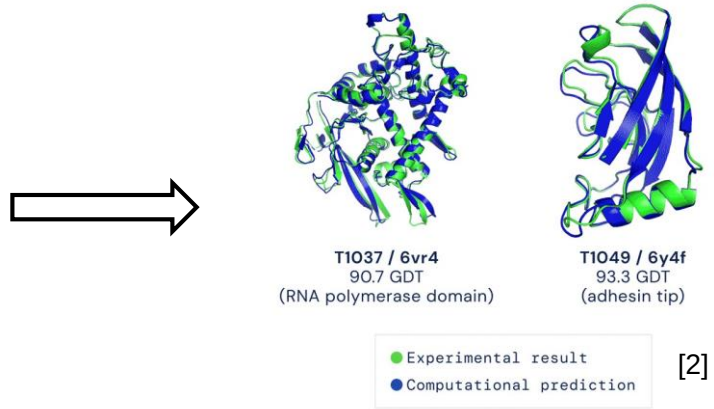
AlphaFold2 [1]

- Deep learning system to predict the 3d structure of proteins based on their linear sequence of amino acids
- Adapted and optimized by MPCDF early on for use on supercomputers with **GPU acceleration**
- High demand and **extreme IO requirements**, mitigated by using dedicated **NVMe-based** storage systems
- Very large and broad user base, encompassing theoretical, interdisciplinary, and experimental groups

Biology

```
>T1037 S0A2C3d4, , 404 residues|

SKINFYTTTIETLETEDQNNTLTTFKVQNVSNASTIFSNGK
TYWNFARPSYISNRINTFKNNPGVLRQLLNTSYGQSSLWAK
HLLGEEKNVTGDFVLAGNARESASENRLKSLELSIFNSLQE
KDKGAEGNDNGSISIVDQLADKLNKVLRGGTKNGTSIYSTV
TPGDKSTLHEIKIDHFIPETISSFSNGTMIFNDKIVNAFTD
HFVSEVNRMKEAYQELETLPESKRVVHYHTDARGNVMKDGK
LAGNAFKSGHILSELSFDQITQDDNEMLKLYNEDGSPINPK
GAVSNEQKILIKQTINKVLNQRIKENIRYFKDQGLVIDTVN
KDGNKGFHFHGLDKSIMSEYTDDIQLTEFDISHVVSDFTLN
SILASIEYTKLFTGDPANYKNMVDFFKRVPATYTN
```



[2]

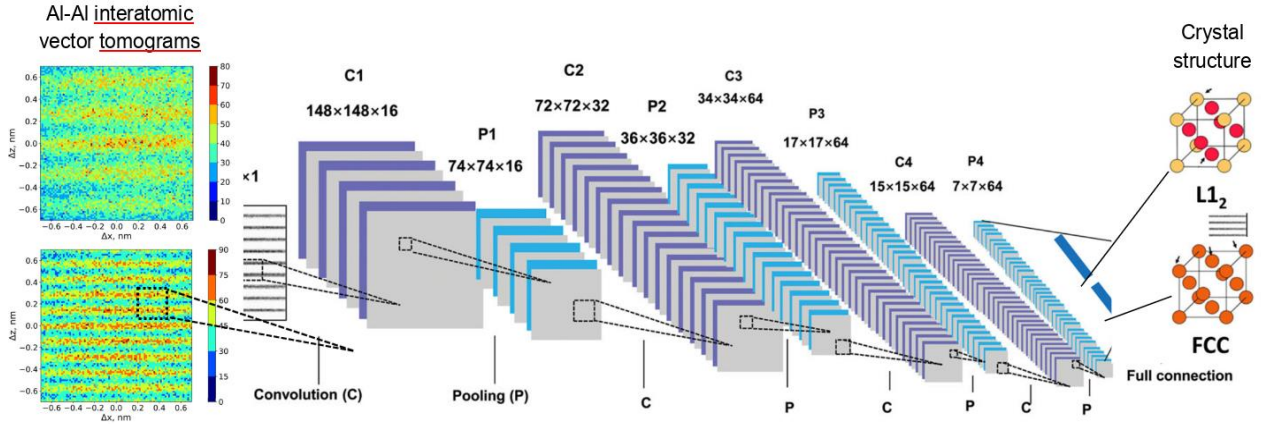
[1] Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." *Nature* 596.7873 (2021): 583-589.
[2] https://github.com/deepmind/alphafold

RECOGNITION OF CRYSTAL STRUCTURES

A COLLABORATION OF MPI FÜR EISENFORSCHUNG AND MPCDF

Automatic analyses of atom probe tomography data

- A **convolutional neuronal network** has been developed which can reconstruct 3D crystal structures from atom probe tomography data
- The method dramatically speeds up the analysis of micrographs
- The method has been extended to reliably detect chemical short-range order (CSRO) in crystalline structures



Y. Li, T. Colnaghi. A. Marek et al. Npj Comput. Mater. 7, 8 (2021)

Materials Science

SISSO++

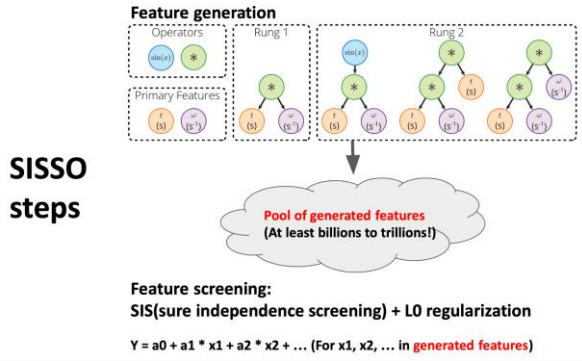
A COLLABORATION OF THE FRITZ-HABER INSTITUTE, MPCDF, EU COE NOMAD

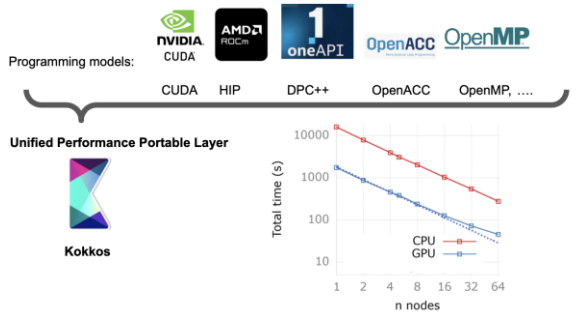
Materials Science

SISSO, a deterministic symbolic regression method

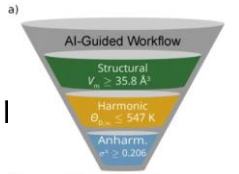
- extracts mathematical expressions directly from data in 2 steps:
  1. create a (huge) pool of analytical expressions through iterative combinations
  2. select optimal candidates for desired properties through (regression) analysis of these expressions and their linear combinations
- SISSO++, open source software (Purcell et al., JOSS, 7(71), 3960, 2022)

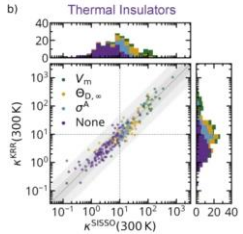
  cross-platform, GPU-acceleration using the Kokkos framework
- scientific application highlight: identification of > 50 strongly thermally insulating materials for thermoelectric elements (devices able to convei otherwise wasted heat into useful electrical voltage)





Y. Yao,S. Eibl,M. Rampp,L. Ghiringhelli, T. Purcell, M. Scheffler (in preparation)





Purcell et al. npj Comput Mater 9, 112 (2023)

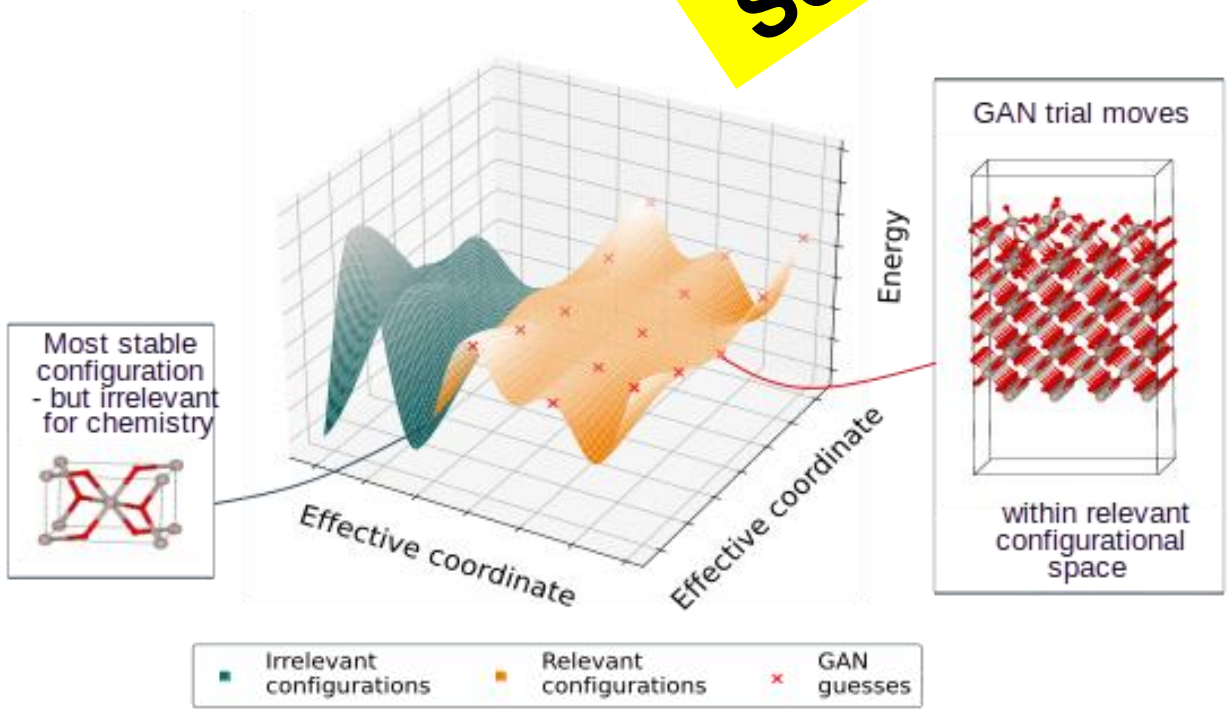
GANS FOR CHEMICAL STRUCTURE GENERATION

A COLLABORATION OF MPI FHI AND MPCDF

Materials Science

Generate relevant chemical structures

- Obtaining chemical structures for interesting configuations is hard, since the most stable (measured) ones are “boring”
- Design and train a **physics informed generative model** which can create physically correct but very interesting structures
- The generated structures will be then used for calculations of material properties



P. König et. al., Presentation at the SKM 2023

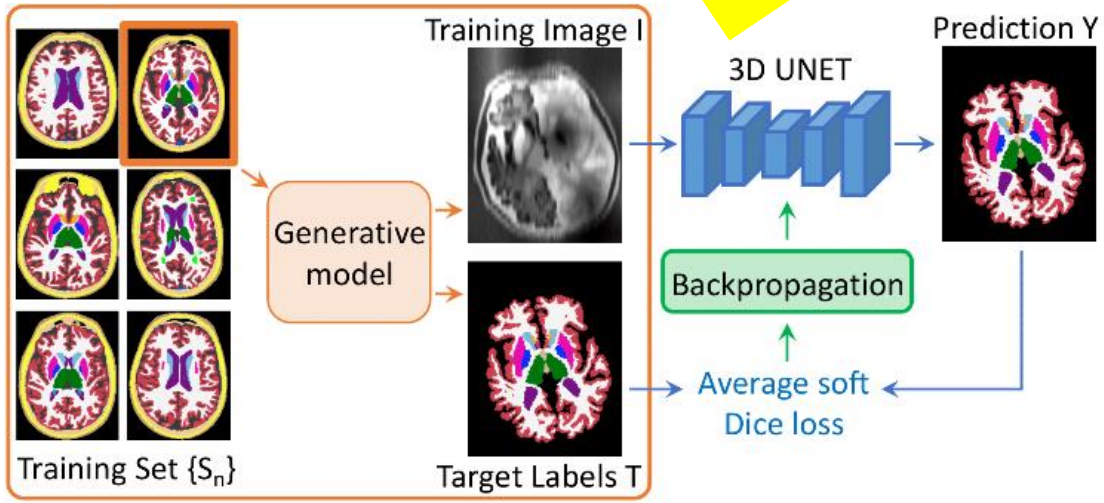
SYNTHSEG

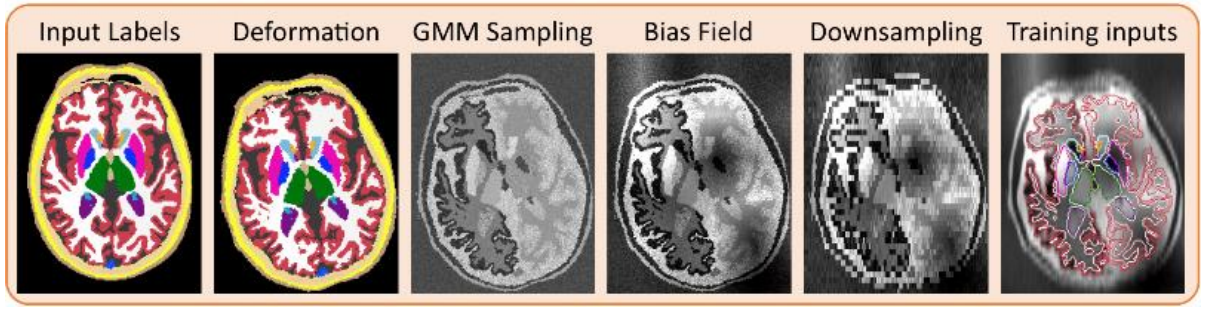
A COLLABORATION OF MPI CBS AND MPCDF

Neuro Science

Synthetic image generation for segmentation networks

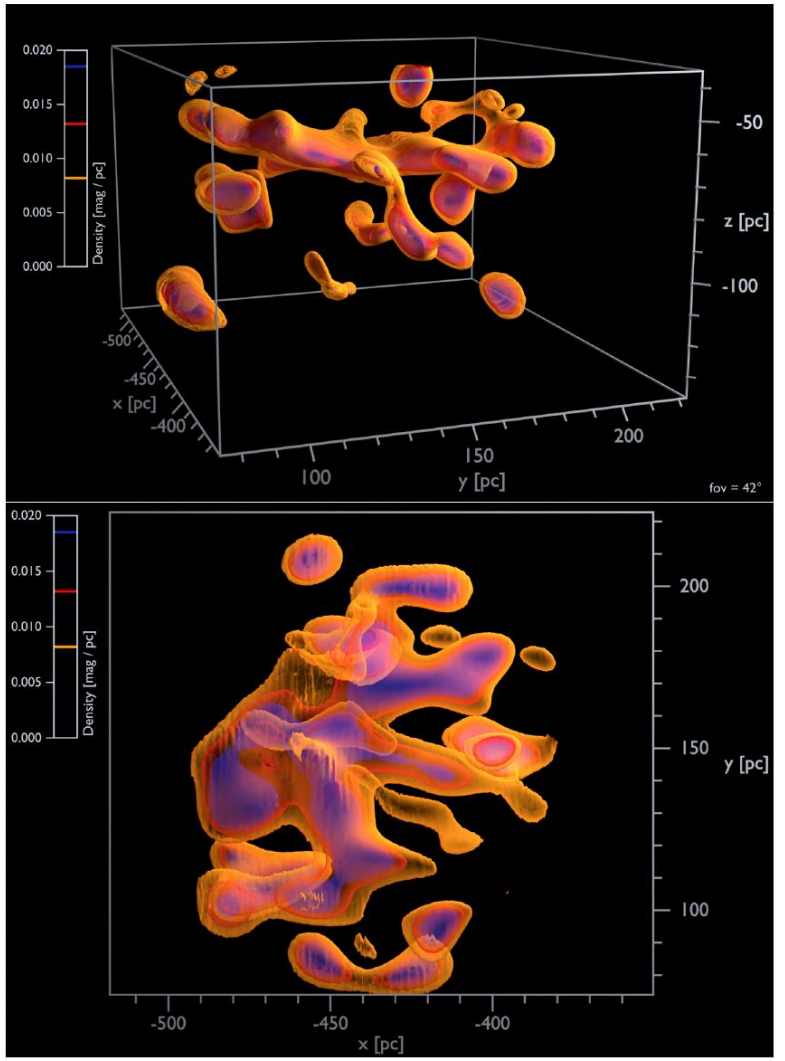
- Instead of training on expensive (and hard to obtain) real MRI scans, a massive and diverse **synthetic dataset** is generated
- The synthetic images are obtainded via a **generative model** that takes as input real exisiting label maps
- The generative model is tuned to produce images that resemble the the real MRI scans
- The final segmentation model (well-proven 3d Unet) is trained with this generated dataset





3D MAPPING OF CLOUD COMPLEXES IN THE MILKY WAY

A COLLABORATION OF MPI ASTRONOMY AND MPCDF



Automatic density reconstruction from distance and optical-IR extinction measurements

- A new algorithm (based on **baysian statistics**) to infer a 3d density distribuition from distance and extinction measurements has been optimized by MPCDF to be able to tackle better resolved inference grids
- A catalog with 16 molecular cloud complexes of the Milky Way a 3d density distribution could be generated

Astrophysics

T. E. Dharmawardena et al., The 3D structure of Galactic molecular cloud complexes out to 2.5 kpc, MNRAS (2022)

SUMMARY

- **AI methods are being explored in many scientific domains**
  - already in production in some
  - “black-box” approach seen critically sometimes
    - effort on validation, trust-worthyness, error-estimation etc.

- **Potential to speed up many tedious tasks**
  - pruning search spaces, creating new study objects via generative models, steering simulations, etc.
  - But will they replace first-principle simulations?
    - and if so, should the physical model be changed?

- **Doubtless, we will see many more (and surprising) adoptons of AI methods in (e)Science**

MYTH 1: QUANTUM COMPUTING WILL TAKE OVER HPC

Scientific Analysis (not Hype) of Utility of Quantum Computing

- **For practical 'quantum supremacy', exponential speedup cf classical algorithm is necessary**
  - Many algorithms only achieve quadratic speedup, thus will lose to classical in practice
    - E.g., Shor's algorithm – exponential => Good
    - E.g., Grover's algorithm – quadratic=>NG
- **For 'pure' quantum algorithms, none exist that exhibit quadratic speedup & can be executed practically on current NISQ machines w/~100 qubits**
  - Shor's algorithm may break RSA 2048 in the far future but will require 20~200mil NISQ qubits https://arxiv.org/pdf/1905.09749.pdf
- **Hybrid algorithms e.g., variational algorithms (e.g. VQE) might be useful in much closer future**
- ***Require platform to conduct scientific analysis of QC, as large qubits as possible, using real state-of-the art real machines and simulators!***

Torsten Hoefler, Thomas Häner, Matthias Troyer
Communications of the ACM, May 2023, Vol. 66 No. 5, Pages 82-87
10.1145/3571725

Disentangling Hype from Practicality: On Realistically Achieving Quantum Advantage

TORSTEN HOEFLER, Microsoft Corporation, USA and ETH Zurich, Switzerland
THOMAS HÄNER and MATTHIAS TROYER, Microsoft Corporation, USA

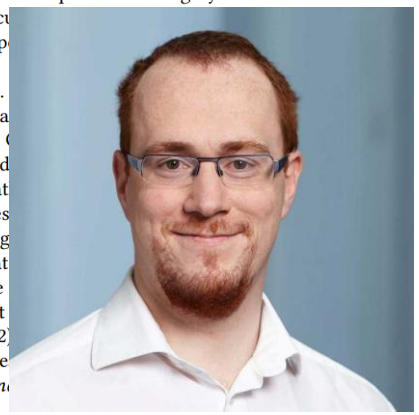
Quantum computers offer a new paradigm of computing with the potential to vastly outperform any imagineable classical computer. This has caused a gold rush towards new quantum algorithms and hardware. In light of the growing expectations and hype surrounding quantum computing we ask the question which are the promising applications to realize quantum advantage. We argue that small data problems and quantum algorithms with super-quadratic speedups are essential to make quantum computers useful in practice. With these guidelines one can separate promising applications for quantum computing from those where classical solutions should be pursued. While most of the proposed quantum algorithms and applications do not achieve the necessary speedups to be considered practical, we already see a huge potential in material science and chemistry. We expect further applications to be developed based on our guidelines.

ACM Reference Format:
Torsten Hoefler, Thomas Häner, and Matthias Troyer. 2022. Disentangling Hype from Practicality: On Realistically Achieving Quantum Advantage. 1, 1 (September 2022), 7 pages. https://doi.org/XXXXXXX.XXXXXXX

Practical and impractical applications. We can now use the above considerations to discuss several classes of applications where our fundamental bounds draw a line for quantum practicality. The most likely problems to allow for a practical quantum advantage are those with exponential quantum speedup. This includes the simulation of quantum systems for problems in chemistry, materials science, and quantum physics, as well as cryptanalysis using Shor's algorithm [13]. The solution of linear systems of equations for highly structured problems [10] also has an exponential speedup, but the I/O limitations discu
and undo this advantage if knowledge of the full solution is required (as opp
obtained by sampling the solution).

Equally importantly, we identify dead ends in the maze of applications.
quadratic quantum speedups, such as many current machine learning tra
design and protein folding with Grover's algorithm, speeding up Monte
walks, as well as more traditional scientific computing simulations includ
systems of equations, such as fluid dynamics in the turbulent regime, weat
achieve quantum advantage with current quantum algorithms in the fores
the identified I/O limits constrain the performance of quantum computing
linear systems, and database search based on Grover's algorithm such that

These considerations help with separating hype from practicality in the
can guide algorithmic developments. Specifically, our analysis shows that
to focus on super-quadratic speedups, ideally exponential speedups and 2
bottlenecks when deriving algorithms to exploit quantum computation be
quantum practicality are small-data problems with exponential speedup, an
problems in chemistry and materials science.



slide curtesy Satoshi Matsuoka

LIKELY/NEEDED QUANTUM DEVELOPMENTS

- More research into algorithms
- QC good for big compute on little data; bad on big data
- QC likely as “accelerator” for certain problems in a classical workflow
  - Most common strategy adopted worldwide today, including EuroHPC
- Commercial viability of QC?

CONCLUSIONS

- **We see a lot of hypes and myths in HPC**
  - some might become reality, some not

- **There is a lot more than (today's) hardware/FLOPS-focussed Exascale computing**
  - scientific approaches need, not hypes

- **Realize that the current hardware market is driven by AI, not HPC**
  - be pragmatic and adopt