MAX PLANCK
COMPUTING & DATA FACILITY

MPCDF

FUTURE AND TRENDS IN SCIENTIFIC HPC

Erwin Laure, Director MPCDF

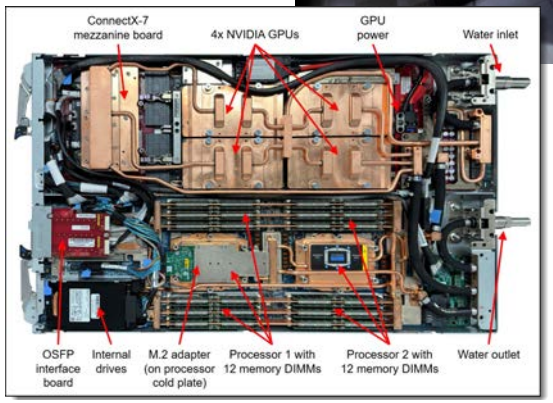
NAISS User Forum – Oct. 1-2, 2024

HPC HAS ALWAYS BEEN A NICHE MARKET

- **Dedicated HPC Systems**
  - Not sustainable in the long run

- **Clusters**
  - Riding the wave of mass consumer CPUs

- **GPU Systems**
  - Riding the wave of gaming industry

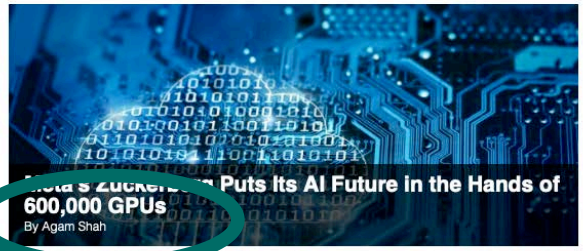


TABOR NETWORK: BIGDATAWIRE AIWIRE HPCWIRE JAPAN QCWIRE HPC & AI WALL STREET

HPCwire

Since 1987 - Covering the Fastest Computers in the World and the People Who Run Them

- Home
- Topics
- Sectors
- Exascale
- Specials
- Resource Library
- Podcast
- Events
- Job Bank
- About
- Subscribe

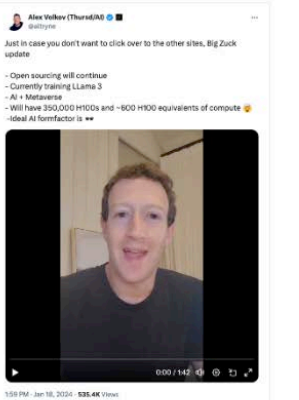


January 25, 2024

In under two minutes, Meta's CEO, Mark Zuckerberg, laid out the company's AI plans, which included a plan to build an artificial intelligence system with the equivalent of 600,000 Nvidia GPUs.

"I'm bringing Meta's AI research efforts closer together to support our long-term goals of building general intelligence, open sourcing it responsibly, and making it available and useful to everyone in all of our daily lives," Zuckerberg said in a video posted on Twitter.

Zuckerberg's announcement was an updated roadmap of Meta's AI plans, which is built around the upcoming Llama3, which is currently being trained. It will succeed last year's Llama2 model weights and tokenizers, which were major successes with just under 2 million combined downloads on Huggingface. Open-source developers have also released thousands of Llama2 forks.



Mark Zuckerberg announces Llama 3 and 360K GPUs on twitter (https://twitter.com/altryne/status/1748057

Llama3 will compete with Google's recently released Gemini model and OpenAI's GPT-4 and upcoming GPT-5 models. OpenAI CEO Sam Altman has not talked about GPT-5 yet but has hinted that it would be much easier to handle text, speech, and images by supporting more data sources.

"We are building an absolutely massive amount of infrastructure to support this by the end of this year. We will have around 350,000 Nvidia H100 or around 600,000 H100 equivalents of compute if you include other GPUs," Zuckerberg said.

MAX PLANCK COMPUTING AND DATA FACILITY

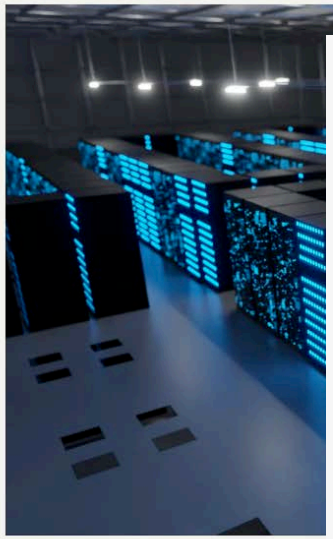
CDO TRENDS
Digital & Data Insights for Business Leaders

A.I. Digital Strategy Data Science Digital Security Data Platform NextGenConnectivity

A.I. DATA CLOUD

Microsoft, OpenAI Planning USD100B AI Supercomputer

By **Paul Mah** April 03, 2024



Microsoft and OpenAI are planning to b
chips to power the next generation of A

According to a report by The Informatio
launch as soon as 2028.

Plans have already been drawn up for th
involvement is contingent on OpenAI fu

Frontier – 40k GPUs

Jupiter – 24k GPUs

TECH & INNOVATION

Google will spend more than $100 billion on AI, exec says

Just last month, Google DeepMind CEO Demis Hassabis said the billio into AI is reminiscent of the cryptocurrency hype

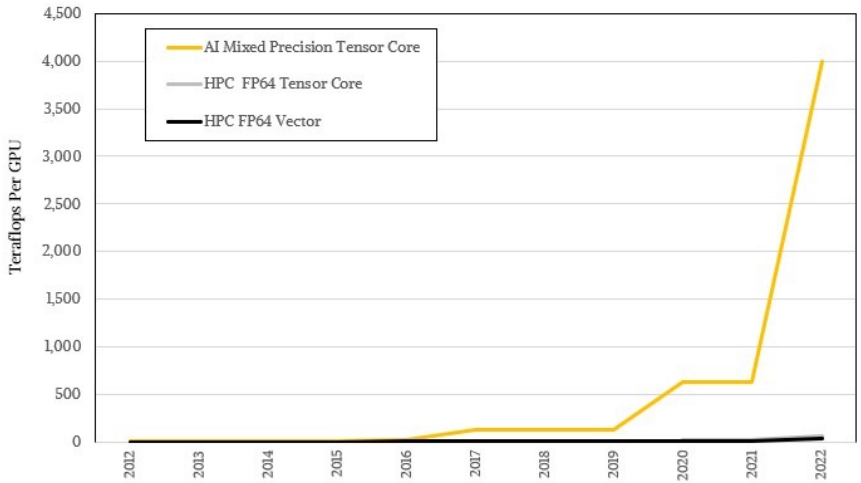
By **Britney Nguyen** Updated April 16, 2024

ARTIFICIAL INTELLIGENCE, MACHINE LEARNING & HPC

Widening gap between AI and HPC (Nvidia GPUs) ?

- **AI vs. traditional floating point (FP64)**
- **focus on lower precision and AI-specific instructions**
  - e.g. tensor cores, bfloat16, …

=> how to leverage in simulation codes?



https://www.nextplatform.com/2022/10/06/the-art-of-system-design-as-hpc-and-ai-applications-diverge

Recall:

- **„first wave“ of GPUs: adopt graphics for HPC simulations**
- **„second wave…“: adopt AI for HPC simulations?**

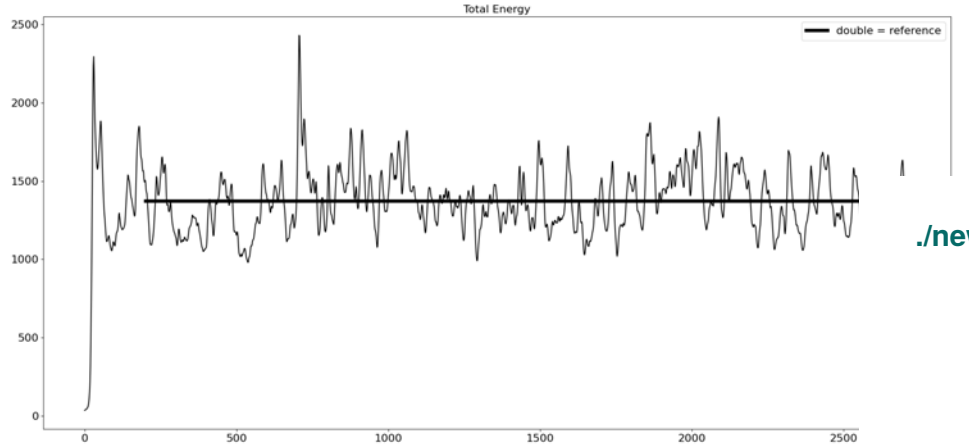
THE (HPC) HARDWARE SITUATION TODAY

- Stagnating CPU performance
- Stagnating/degrading GPU performance (for FP64)
- Small market with only a few players (HPE, Lenovo, Eviden)

- Adopt and/or
- Build own systems?

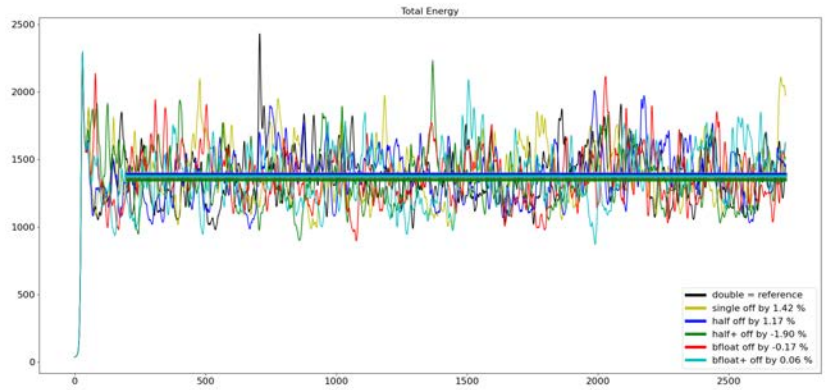
ADOPT – MIXED PRECISION

./newprob run: Total energy profile



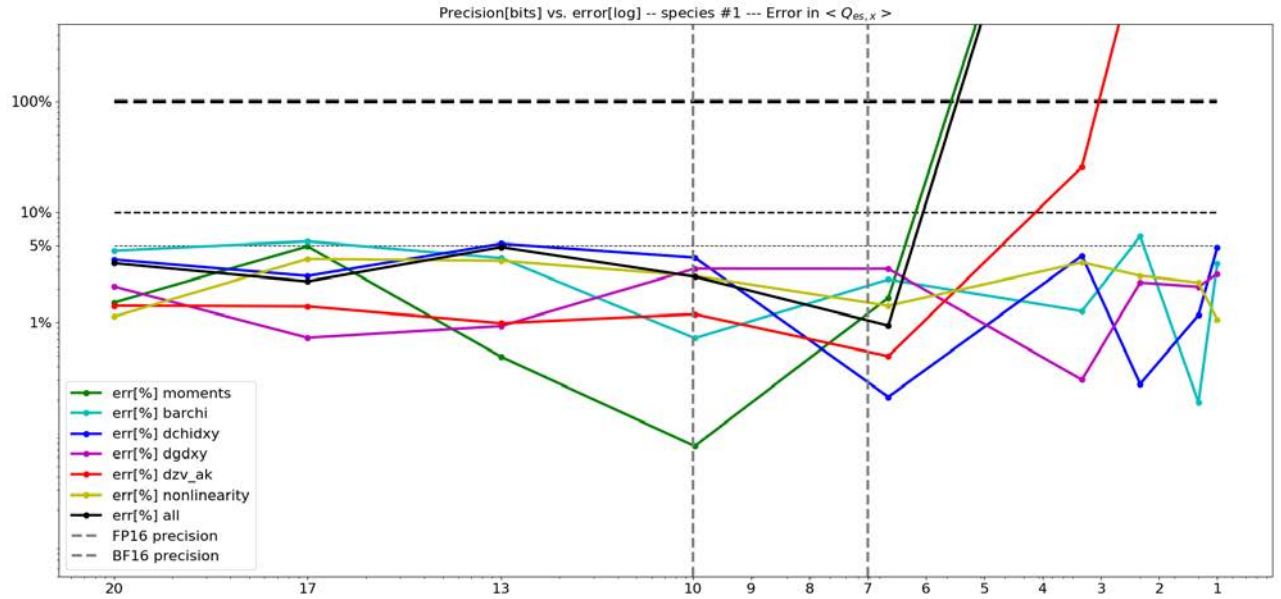
Fusion plasma simulation with GENE

./newprob run: Total energy; mixed precision moments

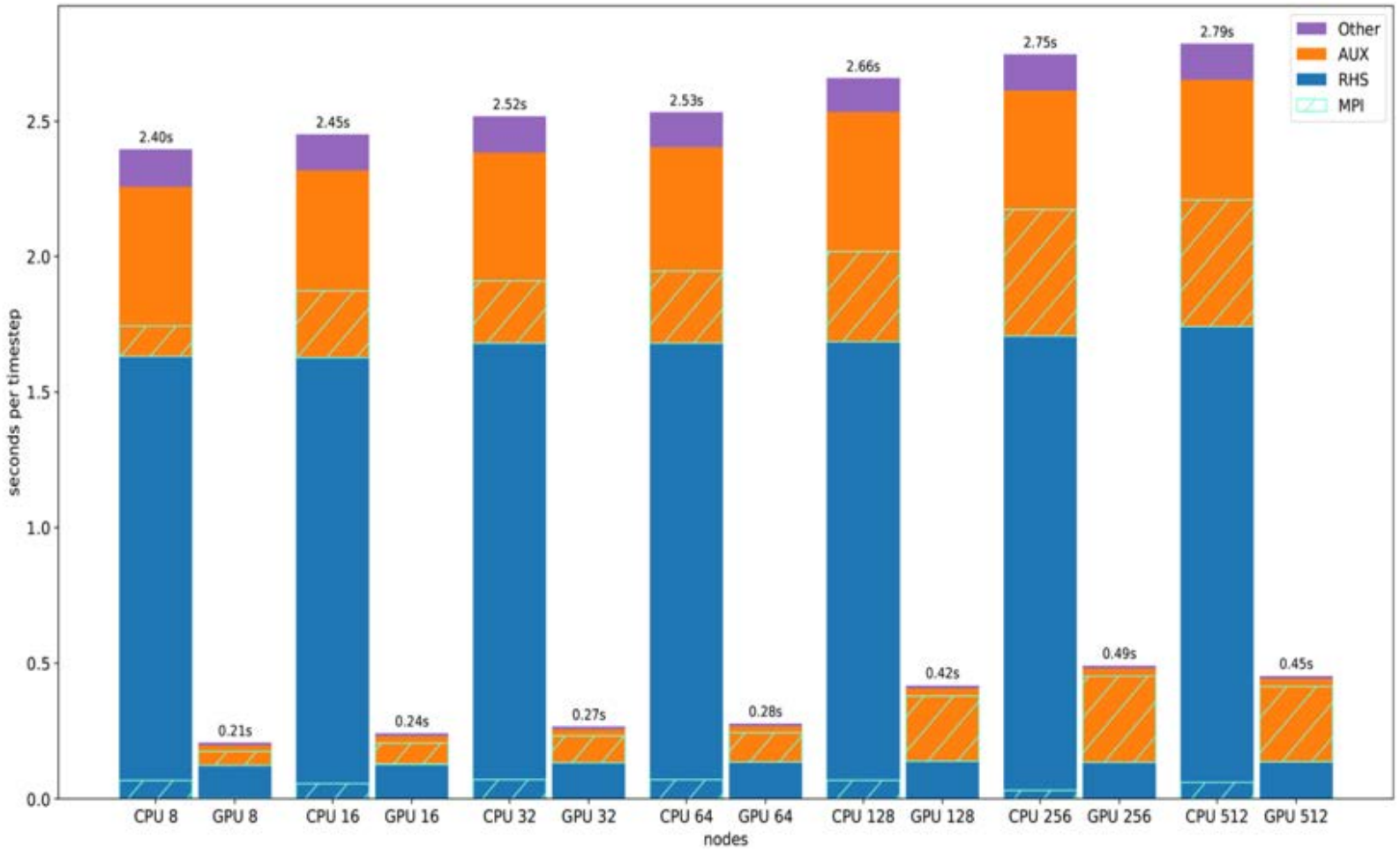


MIXED PRECISION CONT.

Not all terms are sensitive to the reduced precision



GENE SCALING – THE CURSE OF MESSAGE PASSING



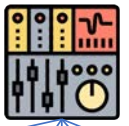
ADOPT – USE AI METHODS

Surrogate Models

Parameter Search
Augmented Simulations
Optimization

Foundational Models for Science

Distillation
Integration
Writing
Protocols
Synthesis
Q+A



Developing Scientific Codes

Debugging
Security
Optimization
Coding
Translation

Controlling complex systems and simulations of digital twins

Surgery
Reactors
Simulation
Manufacturing
Mobility

Design & Optimization

Materials
Devices
Proteins
Industrial Structures

AGI Scientist w/Foundational Models

Experiments& Simulations Planning
Instruments & HPC Control
Discovery & Evaluation

Original slide courtesy Rick Stevens, ANL, 2023

PREDICION OF 3D PROTEIN STRUCTURES

ENABLING AI SYSTEMS IN COMPUTATIONAL BIOLOGY FOR A BROAD USER BASE

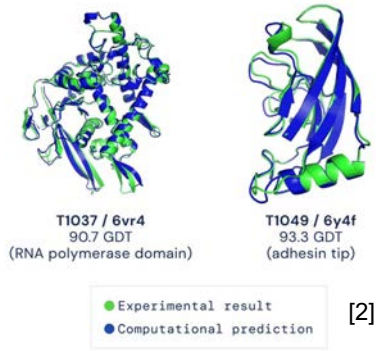
AlphaFold2 [1]

- 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]

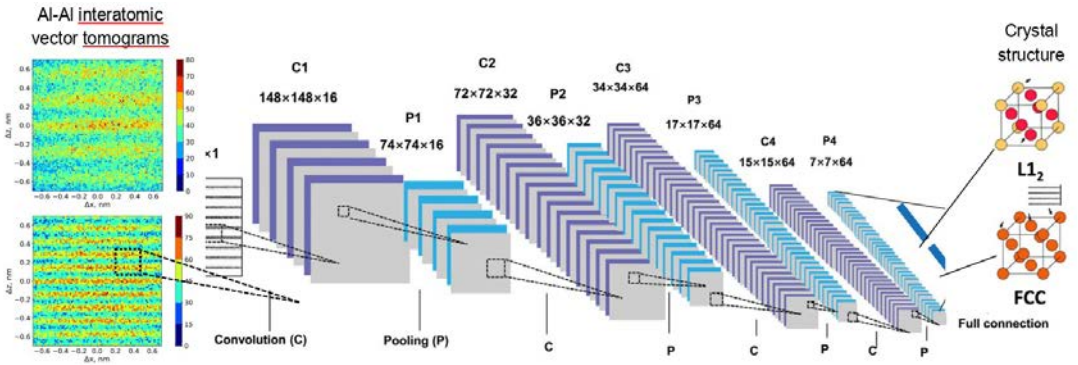
[1] Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." *Nature* 596.7873 (2021): 583-589.
[2] https://github.com/deepmind/alphafold

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++

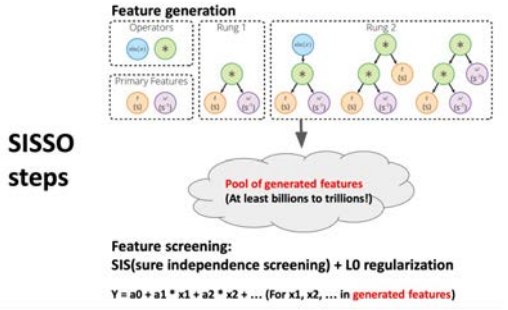
A COLLABORATION OF THE FRITZ-HABER INSTITUTE, MPCDF, EU COE NOMAD

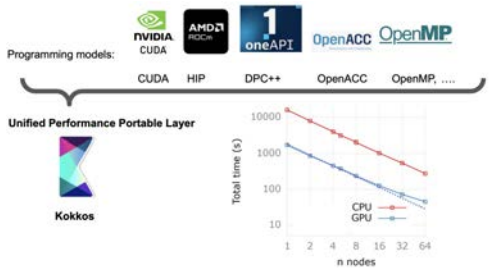
Materials Science

SISSO, a deterministic symbolic regression method

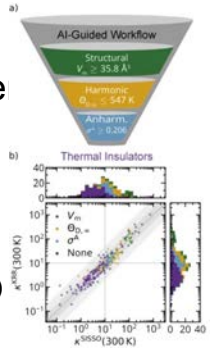
- extracts mathematical expressions directly from data in 2 steps:
  1. create a (huge) pool of analytical expressions through iterative combinations
  2. select optimal candidates for desired properties through (regression) analysis of these expressions and their linear combinations
- SISSO++, open 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 conve 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)

NOMAD

NOVEL MATERIALS DISCOVERY

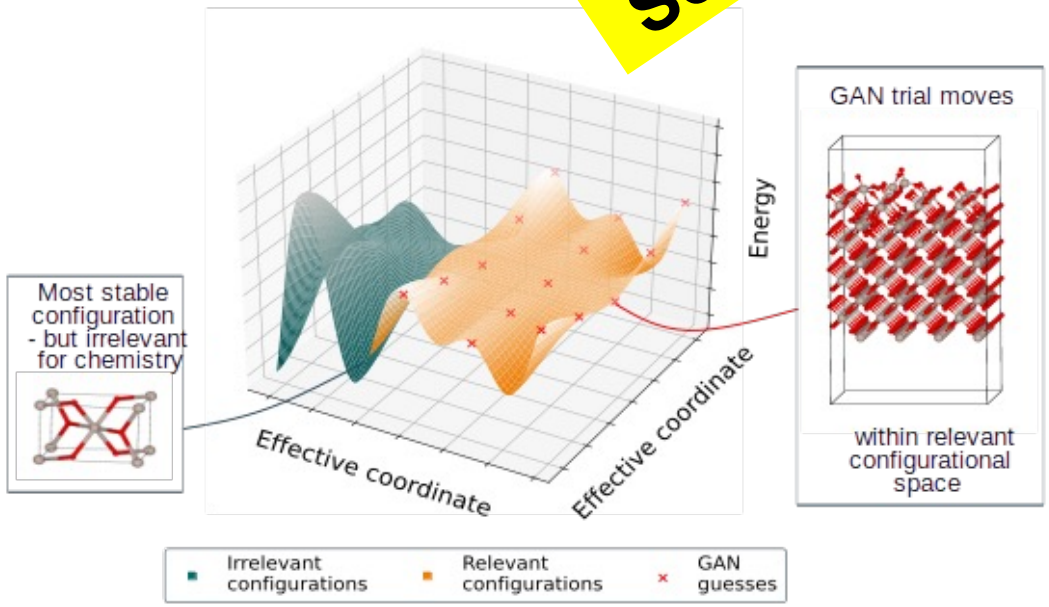
GANS FOR CHEMICAL STRUCTURE GENERATION

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

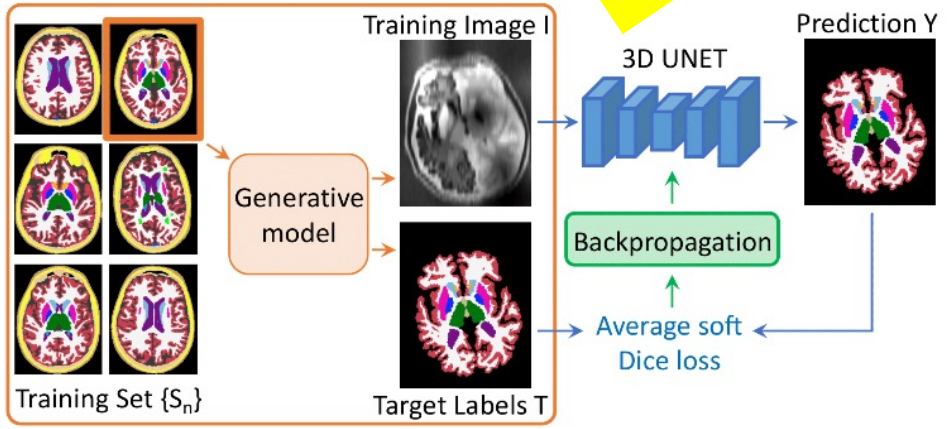
SYNTHSEG

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



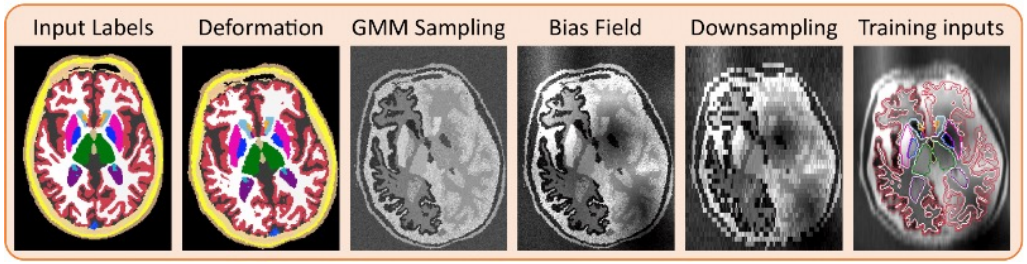


IMAGE SEGMENTATION

A COLLABORATION OF MPI FOR MICROSTRUCTURE PHYSICS AND MPCDF

Automatic segementation of neuron cell images

- Analysing and **segmenting microscop neuron cell images** obtained from tissue samples is very time consuming
- Via manual labeling a data set for cell types has been created
- A **2d Unet neuronal network** is trained on this data set to overcome the tedious manual process of analysing future tissue data

Neuro Science

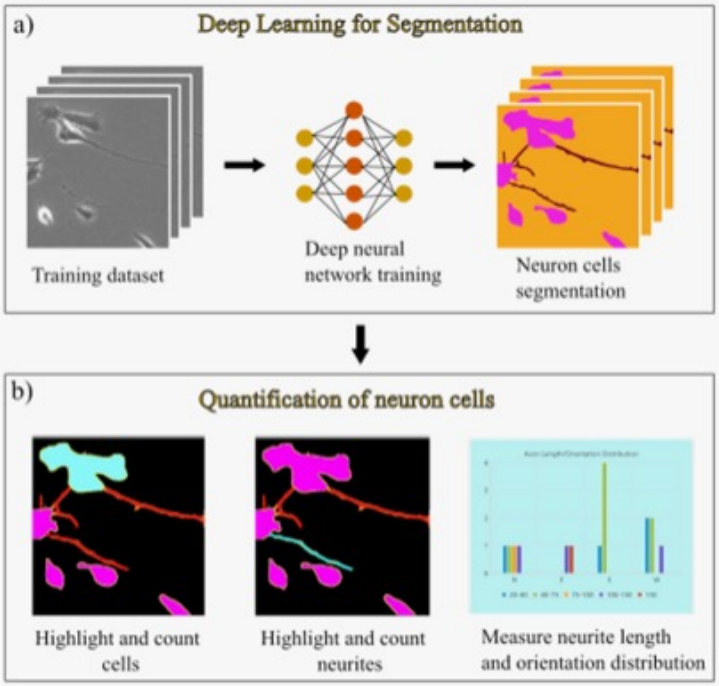


Figure 1: Overview of the computational pipeline of NeuroQuantify, a) Deep learning for cell and neurite segmentation, and b) Quantification number of cells and measurement of neurite lengths and its orientation distribution.

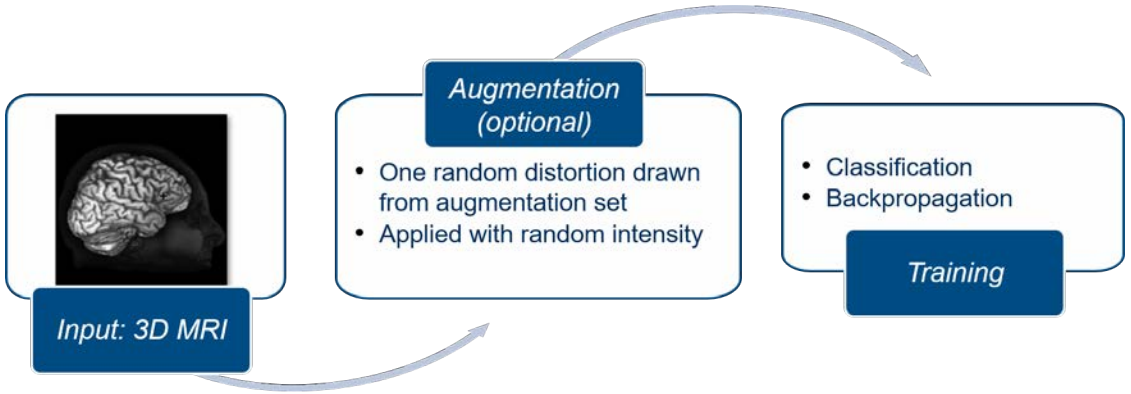
Ka My Dang et al., "NeuroQuantify – A software Package for detection and quantification of Neuron cells and Neurite length-based segmentation", in preparation)

IMAGE SEGMENTATION

A COLLABORATION OF MPI HUMAN COGNITIVE BRAIN SCIENCES AND MPCDF

Robust classification of neuro degeneration

- Ultimate goal: develop a deep neuronal network which can detect from MRI scans the onset of neuro degeneration, such as Alzheimer's disease
- First goal: develop neuronal networks which can descriminate neuro degeneration at **diagnosed** scans vs. "**healty**" scans
- Challenge: obtaining enough MRI scans of early onset neuro generation features



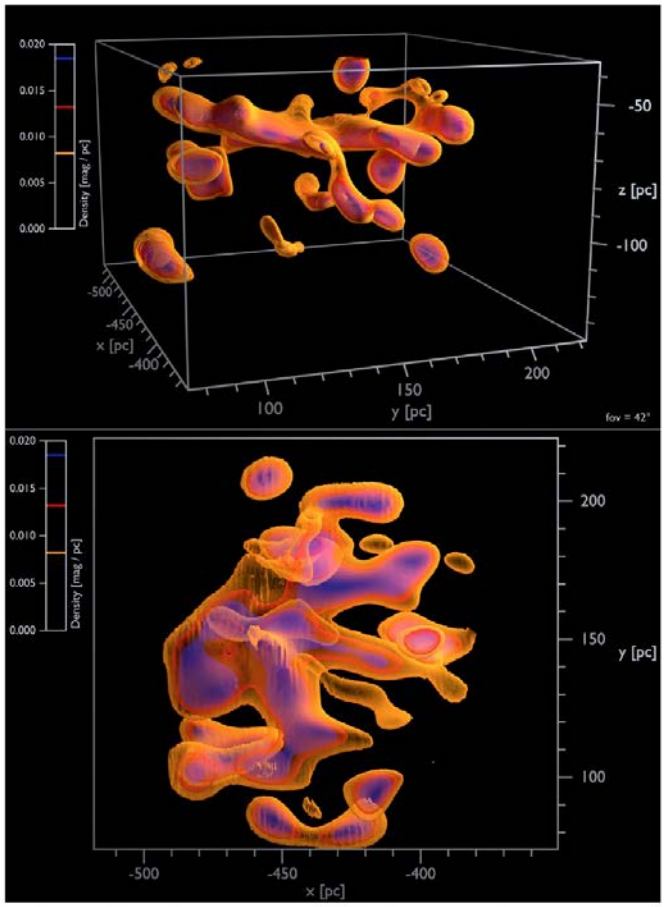
Neuro Science

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

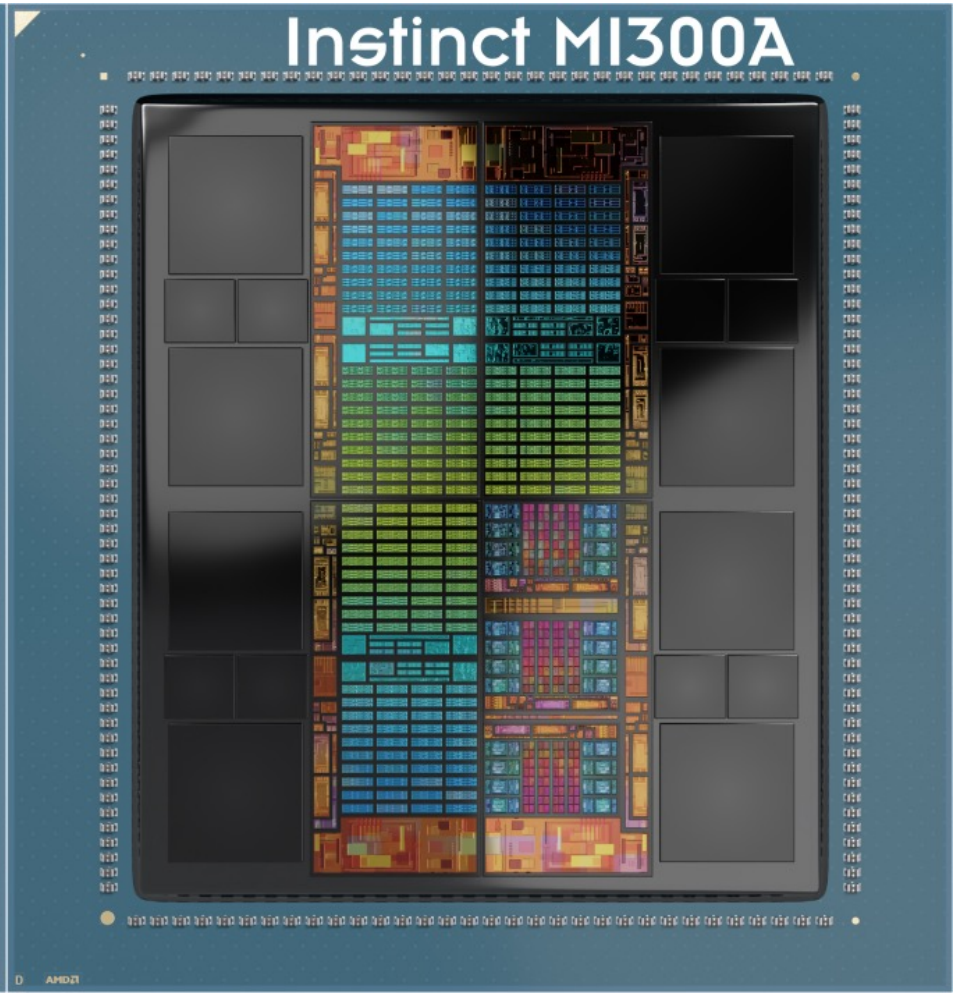
- A new algorithm (based on **baysian statistics**) to infer a 3d density distribuition from distance and extinction measurements has been optimized by MPCDF to be able to tackle better resolved inference grids
- A catalog with 16 molecular cloud complexes of the Milky Way a 3d density distribution could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



Astrophysics

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

OTHER OPTION: BUILD



DISRUPTIVE OPTIONS – QUANTUM?

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

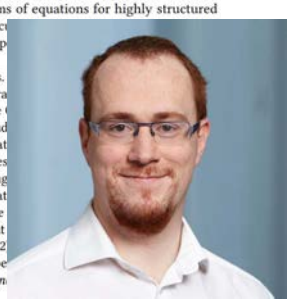
Quantum computers offer a new paradigm of computing with the potential to vastly outperform any imagineable classical computer. This has caused a gold rush towards new quantum algorithms and hardware. In light of the growing expectations and hype surrounding quantum computing we ask the question which 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 disc
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?

FUTURE AND TRENDS IN SCIENTIFIC HPC

- **Stagnating hardware –**
  the importance of algorithmic developments
- **The role of AI**
- **Specialized systems?**
- **Disruptive technologies?**