HPC and Quantum Computing: Impact and Future Trends

HPC and Innovative Computing for EO Workshop | ESA/ESRIN | October 12, 2023

PROF. DR. –ING. GABRIELE CAVALLARO (WWW.GABRIELE-CAVALLARO.COM)
HEAD OF SIMULATION AND DATA LAB “AI AND ML FOR REMOTE SENSING”, JÜLICH SUPERCOMPUTING CENTRE
ADJUNCT ASSOCIATE PROFESSOR, SCHOOL OF ENGINEERING AND NATURAL SCIENCES, UNIVERSITY OF ICELAND

Today's Menu

- Trends in Supercomputing
- The Challenge of Exascale
- Adoption of Innovative Computing Paradigms

Trends in Supercomputing

Supercomputing and its Applications

HPC is for complex calculations that general-purpose computers cannot handle

Digital replica of Earth

...

Drug design

...

Human brain functionality – neural networks

...

Materials design

...

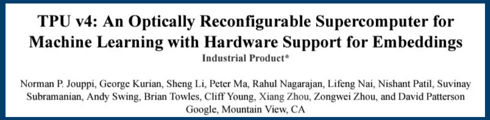
Vegetation evolution – climate change

- High number of processors, vast amounts of memory, and high-speed interconnects

Current Popularity of Supercomputers

Convergence of HPC and AI

Big tech companies are announcing their AI supercomputers


TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings
Industrial Product*
Norman P. Jouppi, George Kurian, Sheng Li, Peter Ma, Rahul Nagarajan, Lifeng Nai, Nishant Patil, Suvinay Subramanian, Andy Swing, Brian Towles, Cliff Young, Xiang Zhou, Zongwei Zhou, and David Patterson
Google, Mountain View, CA


Tech > Science
BABY STEPS Google artificial intelligence supercomputer creates its own 'AI child' that can outperform its human-made rivals
The NASNet system was created by a neural network called AutoML earlier this year
Mark Hodge
Published: 15:22, 5 Dec 2017 | Updated: 11:27, 6 Dec 2017

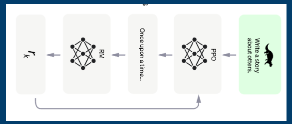

FORBES > INNOVATION > SUSTAINABILITY
Tesla's Biggest News At AI Day Was The Dojo Supercomputer, Not The Optimus Robot
James Morris Contributor
I write about the rapidly growing world of electric vehicles
Follow
0
Oct 6, 2022, 07:23am EDT

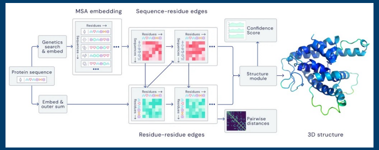

RESEARCH
Introducing the AI Research SuperCluster — Meta's cutting-edge AI supercomputer for AI research
January 24, 2022

- HPC goes far beyond traditional scientific computing, which was driven by large governments
- The field is currently propelled by major industries building highly specialized AI supercomputers

N. P. Jouppi, G. Kurian, et al., "TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings", 2023, https://doi.org/10.48550/arXiv.2304.01433
https://www.forbes.com/sites/jamesmorris/2022/10/06/teslas-biggest-news-at-ai-day-was-the-dojo-supercomputer-not-the-optimus-robot/?sh=22ba4ab780bd
https://ai.facebook.com/blog/ai-rsc/
https://www.thesun.co.uk/tech/5072741/google-nasnet-ai-child-reinforcement-learning/
Torsten Hoefler, "Efficient AI: From supercomputers to smartphones", Scalable Parallel Computing Lab @ ETH Zurich, https://youtu.be/xxwT45ljG4o

Breakthroughs require heavy compute power using many accelerators simultaneously



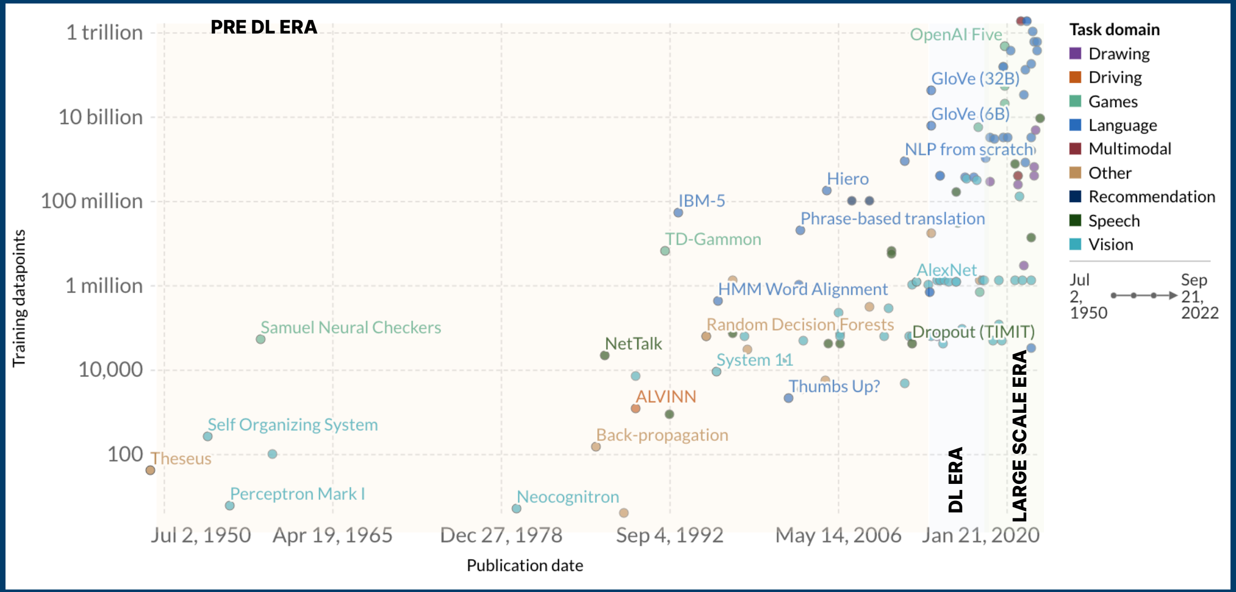


- **GPT**: natural language generation, language understanding
- **CLIP, DALL-E 3**, Stable Diffusion: image understanding and image generation
- **AlphaFold 2**: protein structure prediction
- **AlphaZero, MuZero**: learning control in highly dimensional state-action spaces

https://openai.com/blog/chatgpt
AlphaFold: a solution to a 50-year-old grand challenge in biology, https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology
MuZero: Mastering Go, chess, shogi and Atari without rules, https://www.deepmind.com/blog/muzero-mastering-go-chess-shogi-and-atari-without-rules
Jenia Jitsev, Towards Scalable Deep Learning, Scalable Learning & Multi-Purpose AI Lab, Helmholtz AI, LAION @ JSC

Large-Scale Deep Learning Era

Since around 2015



- In late 2015, a new trend of large-scale models emerged
- Computational capacity significantly higher than that of other models published in the same year (e.g., release of AlphaGo)
- This growth trend is slower than the overall DL trend, with a doubling time of roughly every 8 to 17 months

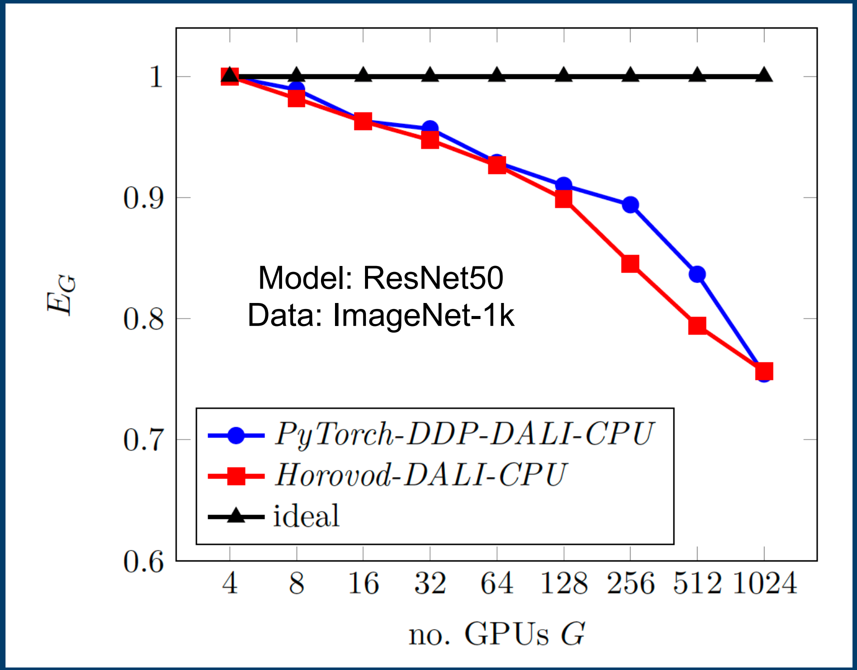
C. Giattino, et al., Artificial Intelligence, https://ourworldindata.org/artificial-intelligence

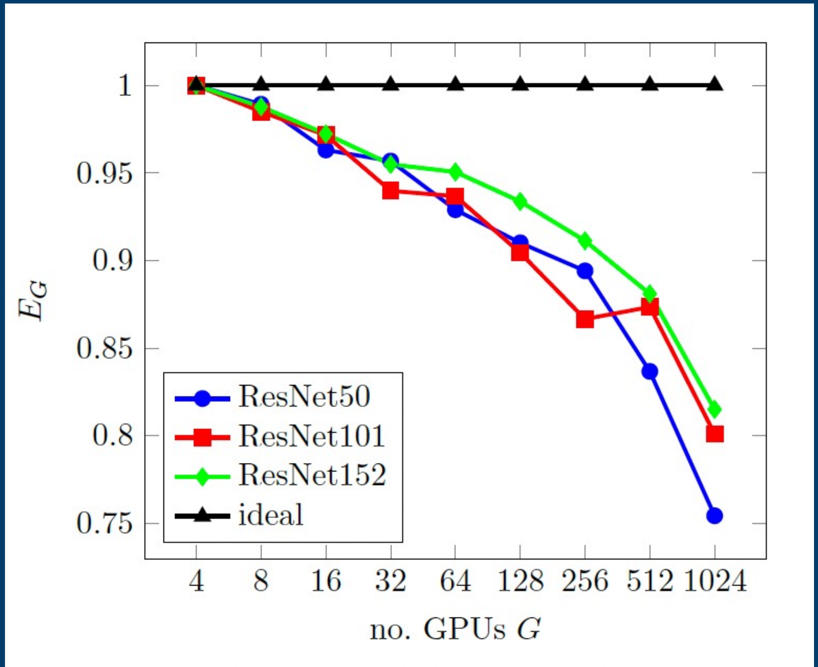
J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn and P. Villalobos, "Compute Trends Across Three Eras of Machine Learning," 2022 International Joint Conference on Neural Networks (IJCNN), pp. 1-8, 2022, https://doi.org/10.1109/IJCNN55064.2022.9891914

D. Silver et al., “Mastering the game of Go without human knowledge,” Nature, vol. 550, pp. 354–359, 2017, https://doi.org/10.1038/nature24270

Distributed Deep Learning at JSC

Large-scale performance analysis of different frameworks (data parallelism)





E_g of (or close to) unity is the ideal scenario with perfect scaling

Parallel Efficiency

$$E_g = \frac{S_g}{G}$$

Speed Up

$$S_g = \frac{T_4}{G}$$

G: number of workers
T_4: reference runtime (4 GPUs)

European Center of Excellence in Exascale Computing "Research on AI- and Simulation-Based Engineering at Exascale" (CoE RAISE), https://www.coe-raise.eu/

Aach, M., Inanc, E., Sarma, R. et al. Large scale performance analysis of distributed deep learning frameworks for convolutional neural networks. J Big Data 10, 96 (2023). https://doi.org/10.1186/s40537-023-00765-w

R. Sedona, G. Cavallaro, et al., "Remote Sensing Big Data Classification with High Performance Distributed Deep Learning", Remote Sensing (MDPI), vol. 11, no. 24, pp. 3056, 2019, https://doi.org/10.3390/rs11243056

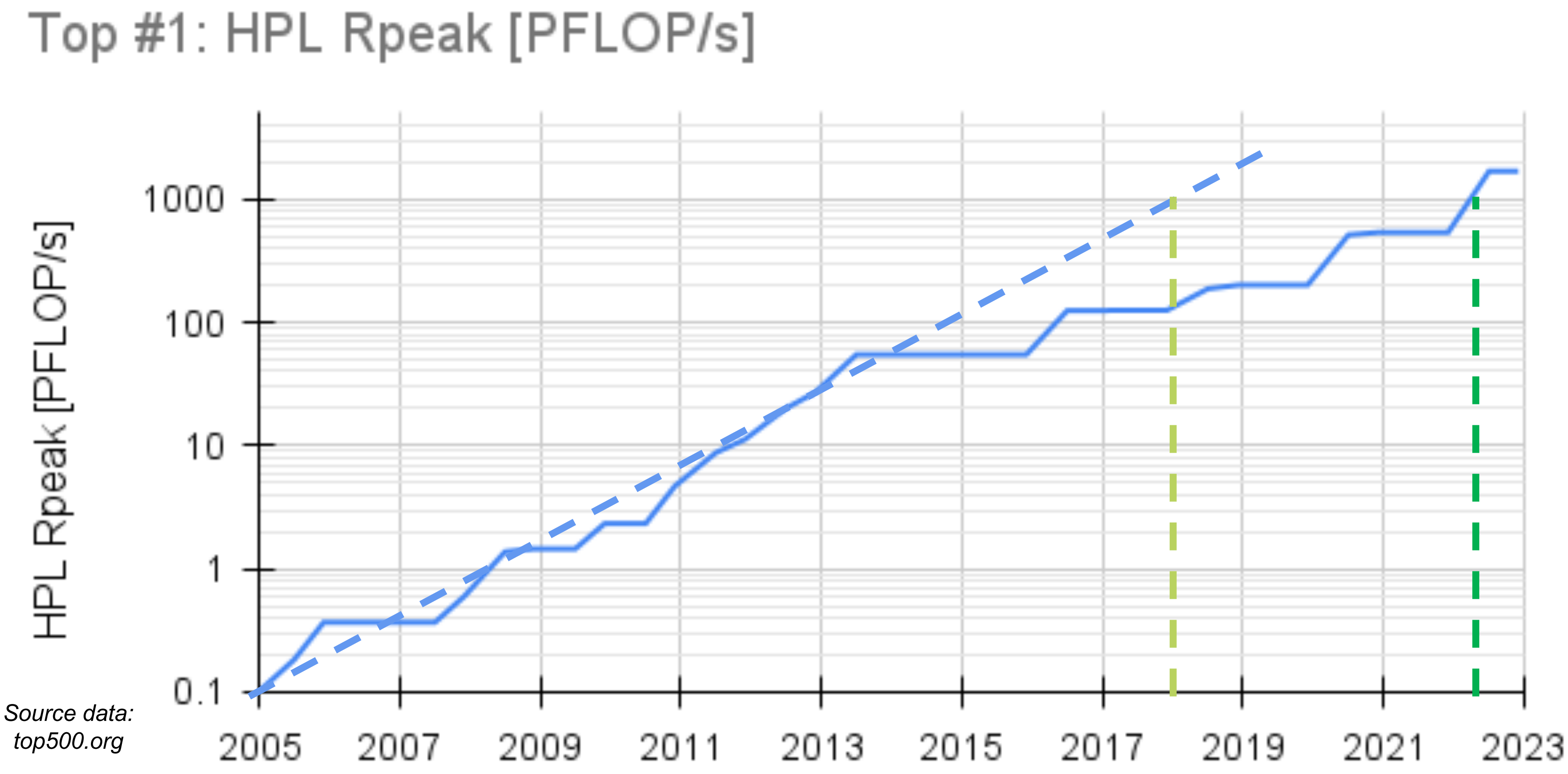
R. Sedona, C. Paris, G. Cavallaro, L. Bruzzone, and M. Riedel, “A High-Performance Multispectral Adaptation GAN for Harmonizing Dense Time Series of Landsat-8 and Sentinel-2 Images,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS), vol. 14, pp. 10134–10146, 2021, https://doi.org/10.1109/JSTARS.2021.3115604

Kesselheim, S. et al. (2021). JUWELS Booster – A Supercomputer for Large-Scale AI Research. In: Jagode, H., Anzt, H., Ltaief, H., Luszczek, P. (eds) High Performance Computing. ISC High Performance 2021. Lecture Notes in Computer Science(), vol 12761. Springer, Cham. https://doi.org/10.1007/978-3-030-90539-2_31

The Challenge of Exascale*

*Exascale computing: capability to perform a billion billion (a quintillion) operations per second (i.e., 10^{18})

Towards Exascale Computing



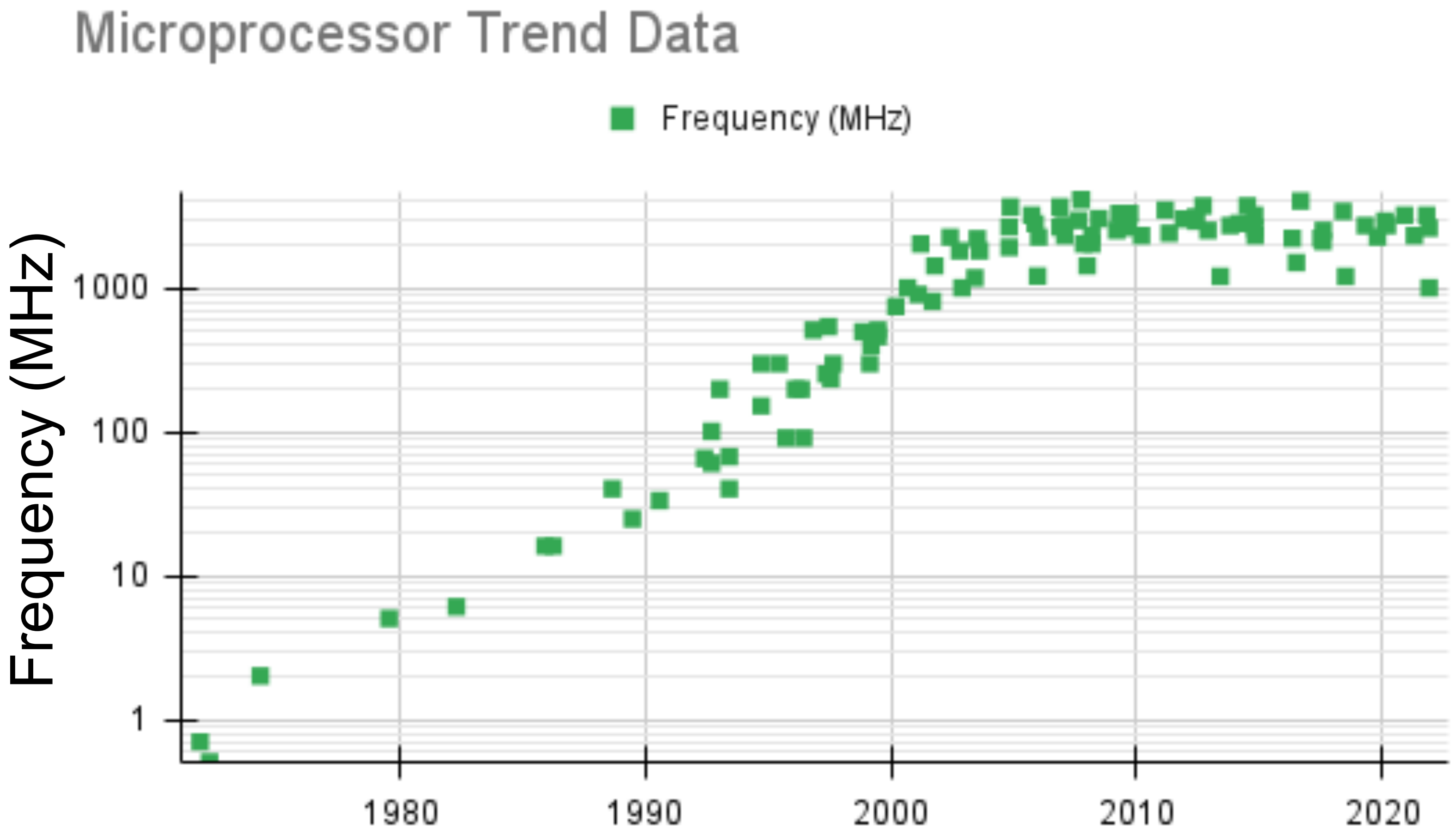
Source data: top500.org

- **1997:** First **1 TFLOP/s** computer: (*Intel ASCI Read/9152)*
- **2008:** First **1 PFLOP/s** computer: (*Roadrunner*)
- So…. First **1 EFLOP/s** computer: **2018 !!**
  - Well… not really
- It took 4 more years… **2022**

- HPL: High-Performance Linpack - solves a (random) dense linear system in double precision (64 bits) arithmetic
- FLOPS: Floating-point operations per second
- Rpeak: Theoretical peak performance (maximum possible performance under optimal conditions)

Estela Suarez, "Trends in modern HPC systems", ESA space weather modelling workshop, 2023

Clock Frequency growth over time

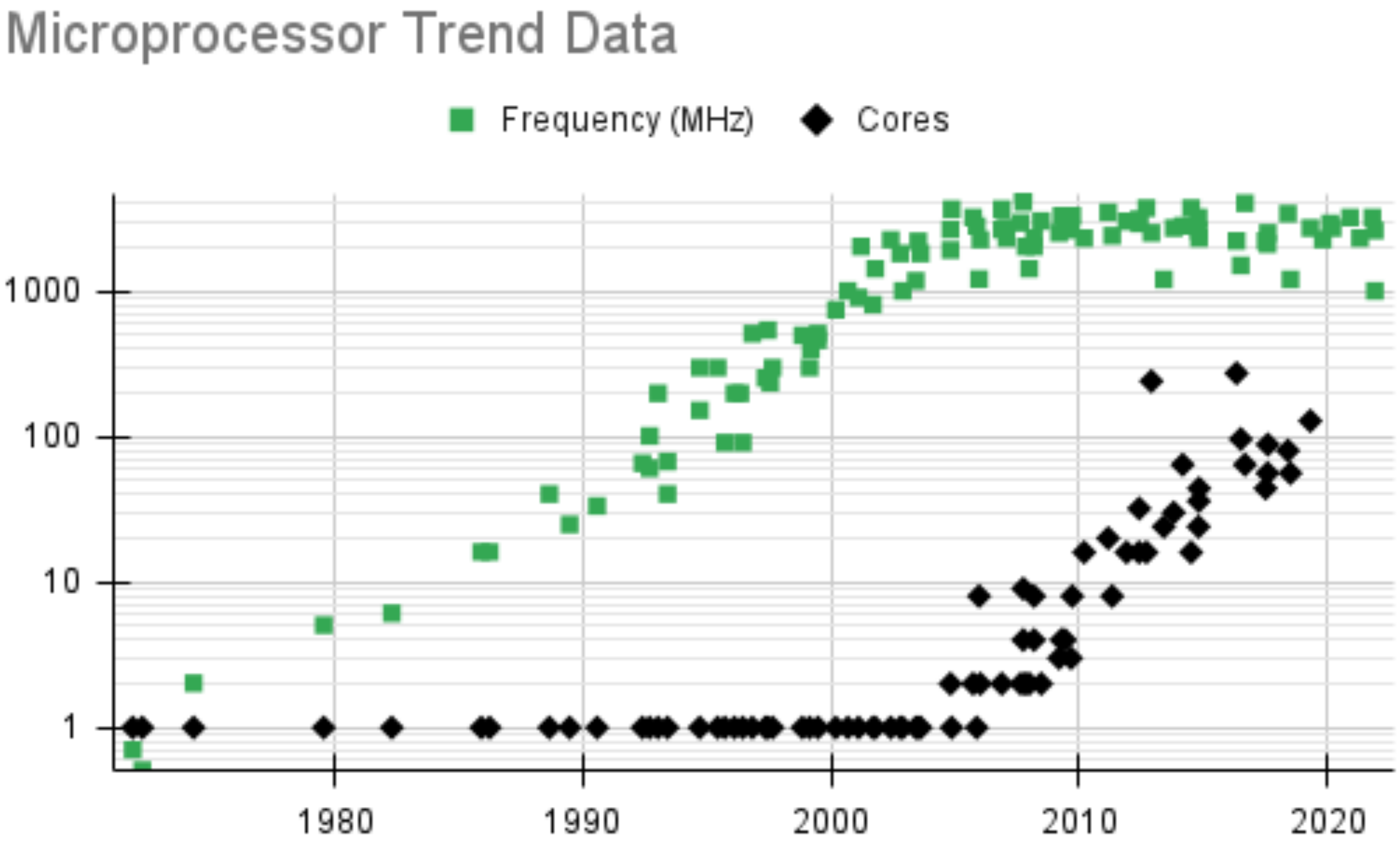


How to keep growing performance ?

Source Data: https://github.com/karlrupp/microprocessor-trend-data

- Moore's law propelled the semiconductor industry to fit an increasing number of transistors and logic into the same volume
- Dennard scaling limits: Performance Plateau in Early 2000s

Clock Frequency and Core Count growth over time



How to keep growing performance ?

➔ integrate more (and more complex) cores per processor

➔ higher concurrency

Even more with GPUs !

Source Data: https://github.com/karlrupp/microprocessor-trend-data

- Multi-core era: still takes advantage of Moore's law
- Emergence of domain specific architectures

Main Challenges for Exascale Computing

- **Concurrency:** Applications must support billions of individual threads
- **Energy efficiency:** Impossible with traditional CPUs alone
- **Memory and storage:** Gap between compute performance and memory bandwidth
- **Communication:** Very large amount of devices need to exchange data with each other

➔ **Limited application scalability**

➔ **Processor heterogeneity**
Caveat: Difficult to program & lack of performance portability

➔ **Deeper hierarchies with high memory and storage heterogeneity** (DDR, HBM, NVM, etc.)

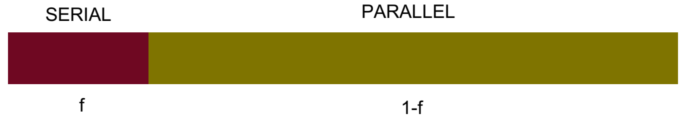
➔ **Need low-latency, high-bandwidth technology and advanced features** (dynamic routing, in-network processing, etc.)

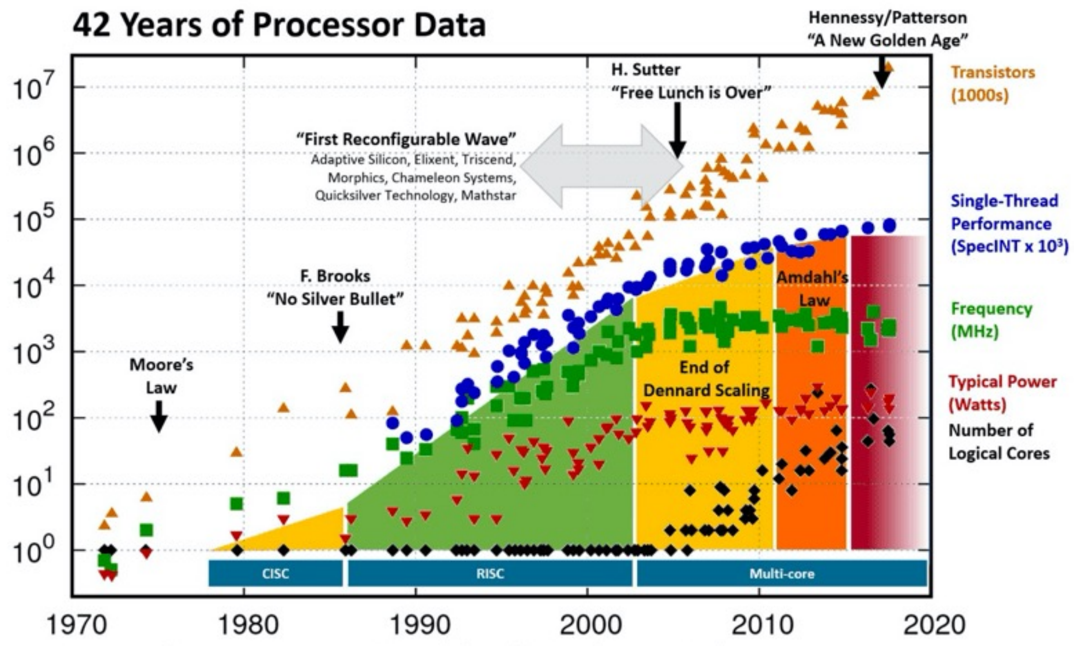
Limits Imposed by Amdahl’s law on Parallelizability

Scalability limited by sequential & low scaling code part(s)

Maximum speedup is constrained by the fraction of the program that cannot be parallelized (i.e., serial portion)

As the number of processing elements increases, the impact of the serial portion becomes more significant





Hennessy and Patterson, Turing Lecture 2018, overlaid over “42 Years of Processors Data” https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/; “First Wave” added by Les Wilson, Frank Schirrmeister
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

R. Muralidhar, R. Borovica-Gajic, R. Buyya, “Energy Efficient Computing Systems: Architectures, Abstractions and Modeling to Techniques and Standards”, in ACM Computing Surveys, vol. 54, no. 11s, 2022, https://doi.org/10.1145/3511094

G. M. Amdahl, "Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities, Reprinted from the AFIPS Conference Proceedings, Vol. 30 (Atlantic City, N.J., Apr. 18–20), AFIPS Press, Reston, Va., 1967, pp. 483–485, when Dr. Amdahl was at International Business Machines Corporation, Sunnyvale, California," in IEEE Solid-State Circuits Society Newsletter, vol. 12, no. 3, pp. 19-20, Summer 2007, doi: 10.1109/N-SSC.2007.4785615.

Limited Application Scalability

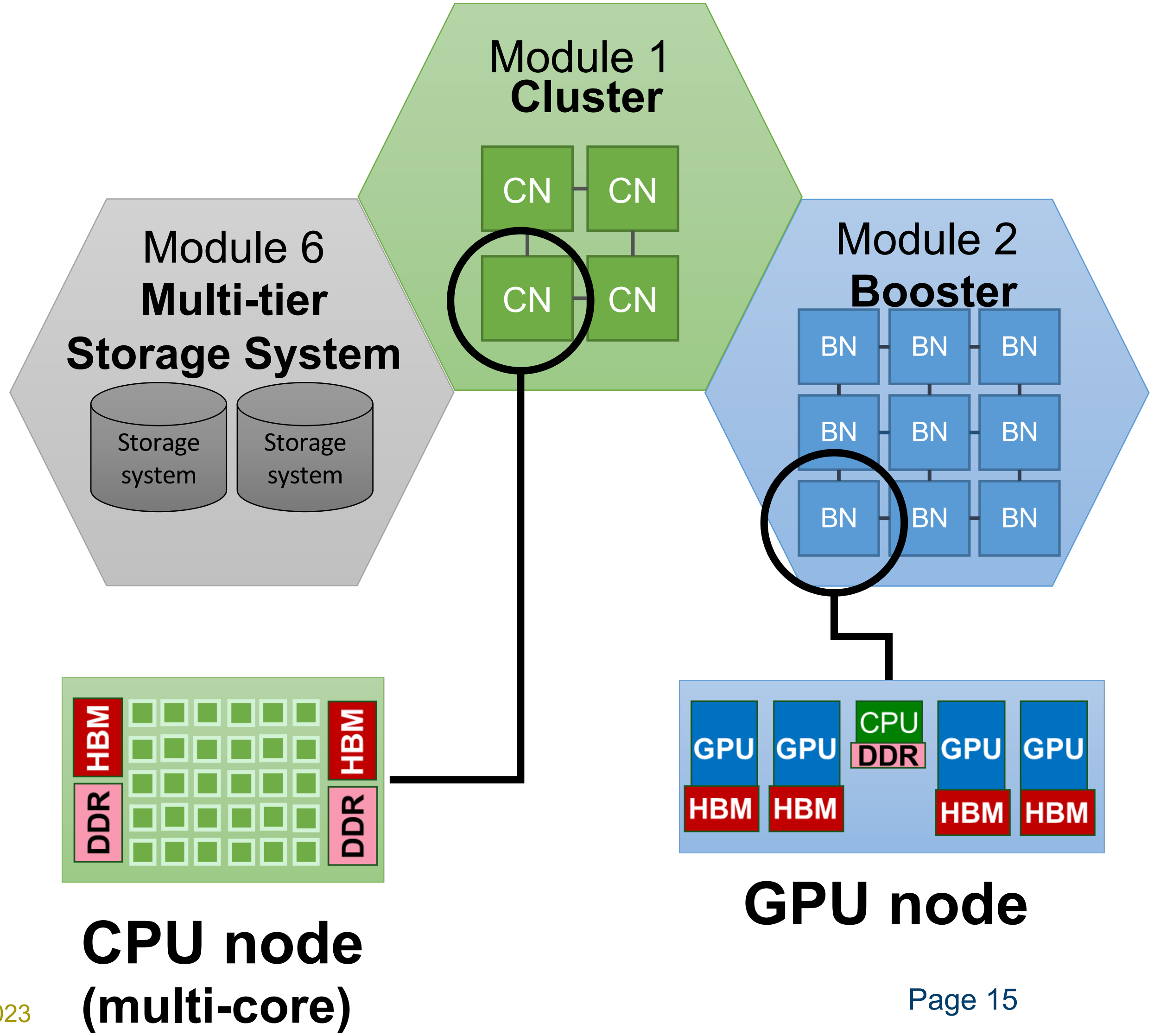
- Scaling an application up to Exascale is very hard
- Users often end up running ensembles of many small jobs
- Is this still HPC ?

Machine ✓	Great Science ?
Scaling Codes ✗	Large Jobs ✗

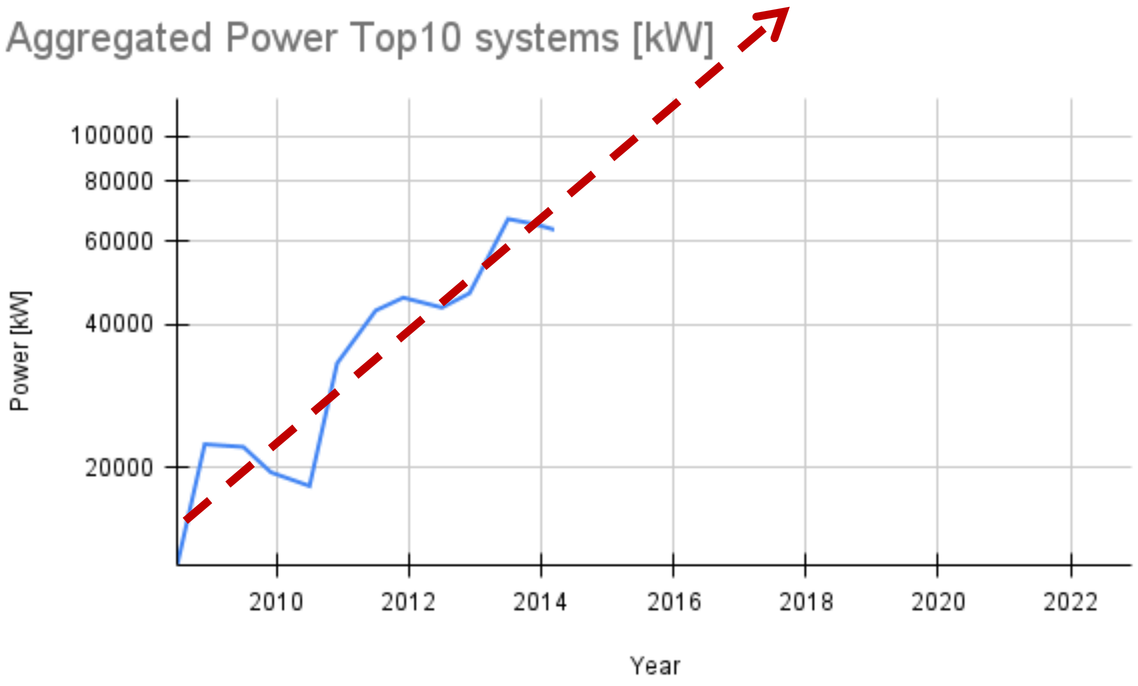
One possible approach: specific hardware to scale each part of the application at the necessary pace

One Cluster-Booster architecture

- **Cluster**: high single-thread perform.
- **Booster**: high throughput



Energy Efficiency
Impossible with traditional CPUs alone

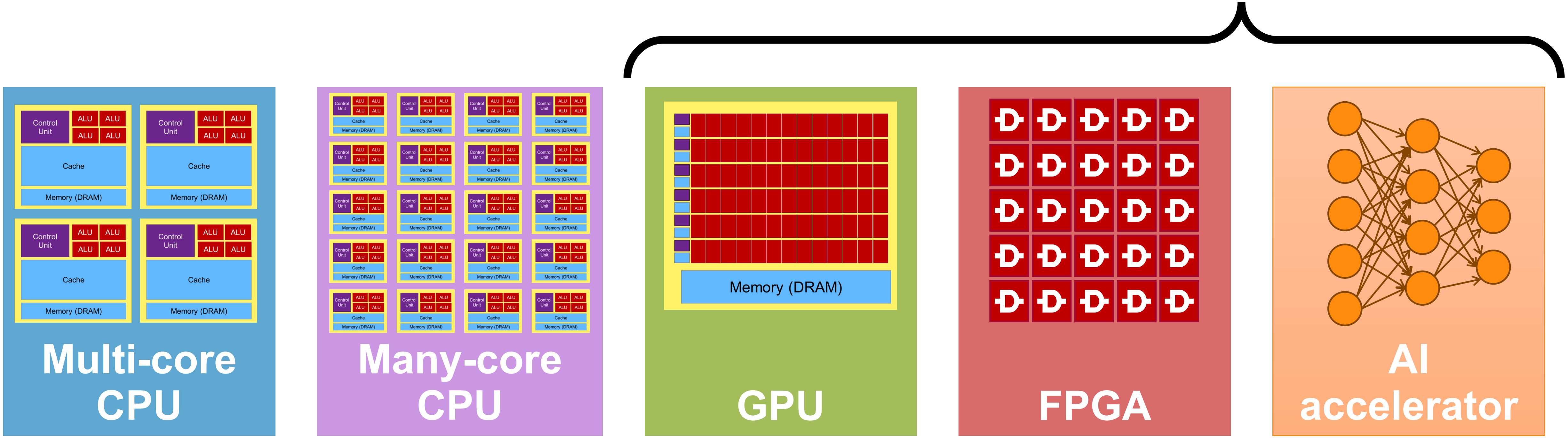


Source data: top500.org

Processor Heterogeneity

Different trade-offs in the design ➔ different processing units

Accelerators



10’s strong cores

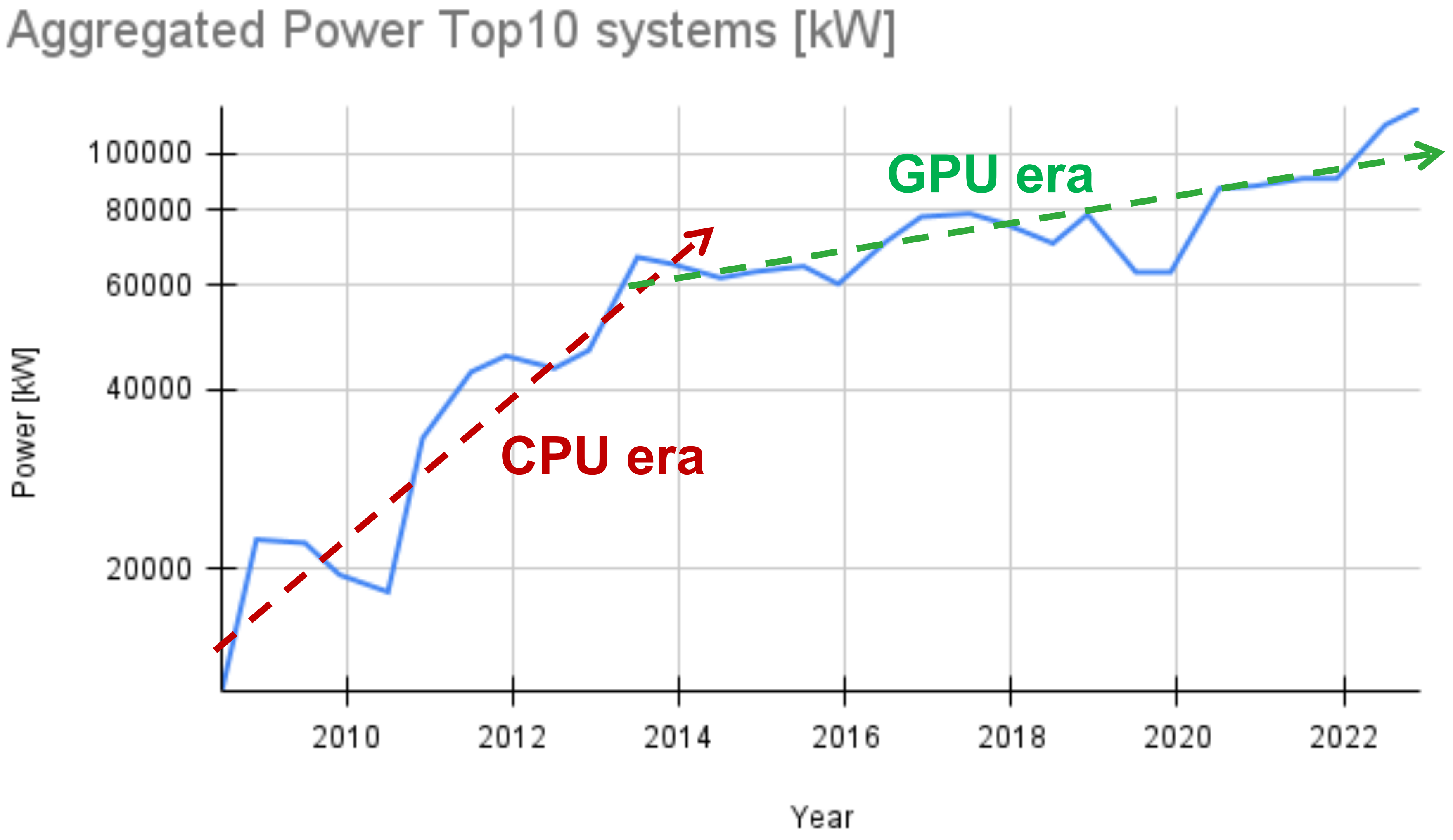
100’s “weak” cores

1000’s functional units

1000’s programmable gates

custom ASIC implementations

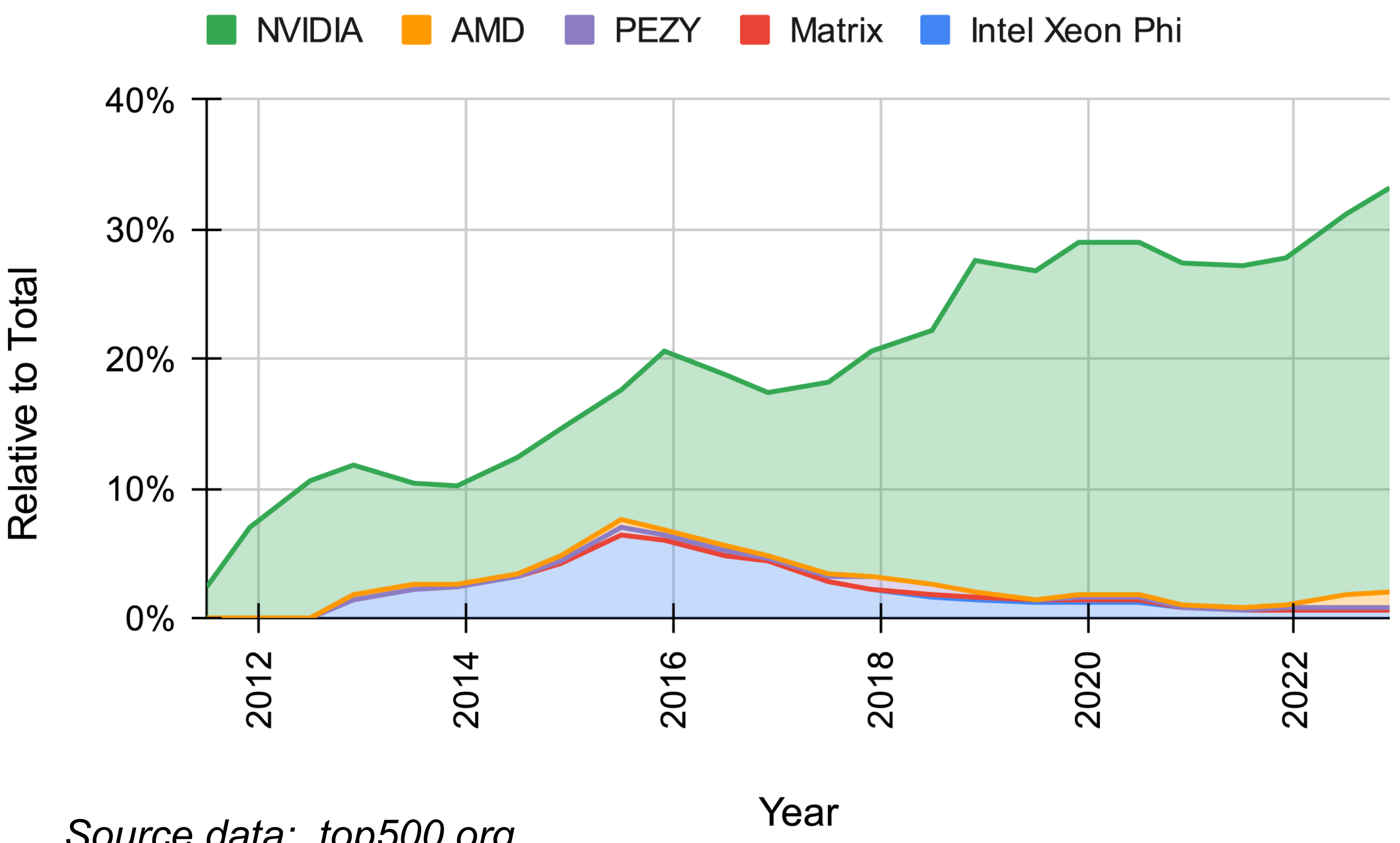
Energy Efficiency



Source data: top500.org

Systems using Accelerators in the Top500

Share of Top500 systems using some kind of accelerator

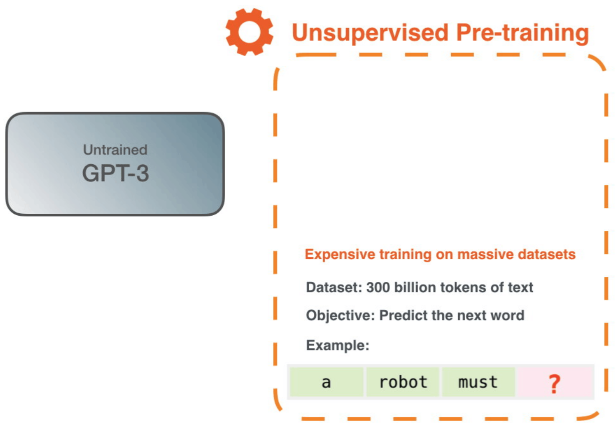


- Amount of HPC systems using acceleration devices is continuously growing
- Mostly general-purpose graphic cards (GPU)
  - Mostly NVIDIA (till now)

Source data: top500.org

How much does it cost to pre-train AI foundation models?

GPT-3



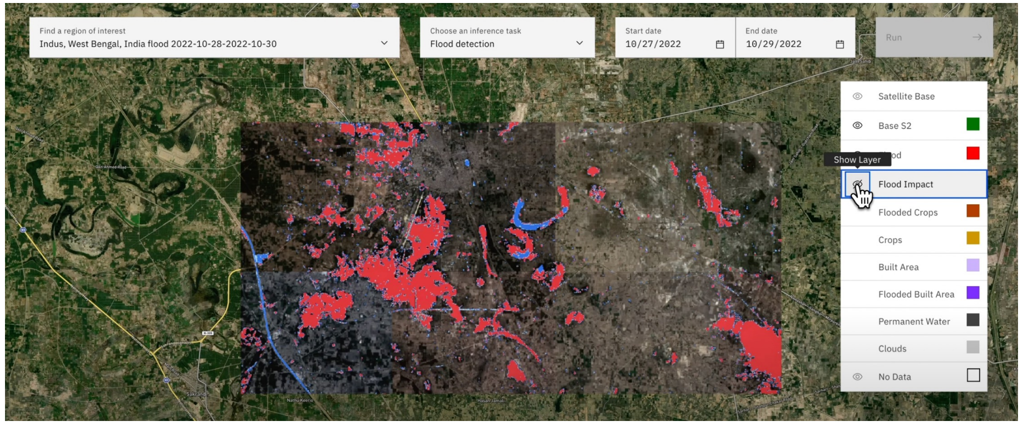
- Model size: 175B parameters
- Time: 100 years (one Nvidia A100 GPU)
- Cost*: >1M€
- Power consumption= ~385 $[MWh]$

  CO2 footprint: >100 $[tCO_2eq]$ = lifecycle of ~5 cars

https://openai.com/blog/gpt-3-apps

Jay Alammar, How GPT3 Works - Visualizations and Animations, http://jalammar.github.io/how-gpt3-works-visualizations-animations/

NASA/IBM Prithvi



- Model size: 100M parameters
- Time: 1 year (one Nvidia A100 GPU)
- Cost*: >10,000€
- Power Consumption= ~3,85 $[MWh]$

https://www.earthdata.nasa.gov/news/impact-ibm-hls-foundation-model

IBM Research, IBM geospatial foundation model, https://youtu.be/9bU9eJxFwWc?si=0by1WdkFT23o0vY5

*Cost for 8x A100 = 12 $/hour on AWS (membership with best deal)

Stefan Kesselheim, "Large Language Models Training in Practice", Helmholtz AI, 2023

H. Touvron, et al, "LLaMA: Open and Efficient Foundation Language Models", 2023, https://doi.org/10.48550/arXiv.2302.13971

PRESS RELEASE | 3 October 2023 | European High-Performance Computing Joint Undertaking

Procurement contract for JUPITER, the first European exascale supercomputer, is signed

The procurement contract for JUPITER, the first EuroHPC exascale supercomputer, has been signed by the European High Performance Computing Joint Undertaking (EuroHPC JU) and a consortium comprising of Eviden and ParTec.

JSC

Schedule

- 17.12.2021: Call for Expression of Interest (EoI) for Hosting Entity
- 14.02.2022: Deadline EoI Submission
- 16.05.2022: Hearings
- 15.06.2022: Hosting site decision and announcement
- **03.10.2023: Procurement contract signed**

Budget

- Total: **500 Mio. €** for purchase and operation
- Funding partners: EuroHPC JU (250 Mio. €),
  Germany/BMBF* (125 Mio. €),
  MKW NRW** (125 Mio. €)

EuroHPC JU, Procurement contract for JUPITER, the first European exascale supercomputer, is signed, https://eurohpc-ju.europa.eu/procurement-contract-jupiter-first-european-exascale-supercomputer-signed-2023-10-03_en

JSC, Europe’s Exascale Supercomputer in Its Starting Blocks, https://www.fz-juelich.de/en/news/archive/press-release/2023/europes-exascale-supercomputer-in-its-starting-blocks

GCS Centre Jülich Supercomputing Centre Set to Operate Europe’s First Exascale Supercomputer, https://www.gauss-centre.eu/news/gcs-centre-juelich-supercomputing-centre-set-to-operate-europes-first-exascale-supercomputer

EVIDEN, An Eviden led consortium to build Europe’s first exascale supercomputer, https://eviden.com/insights/press-releases/an-eviden-led-consortium-to-build-europes-first-exascale-supercomputer/

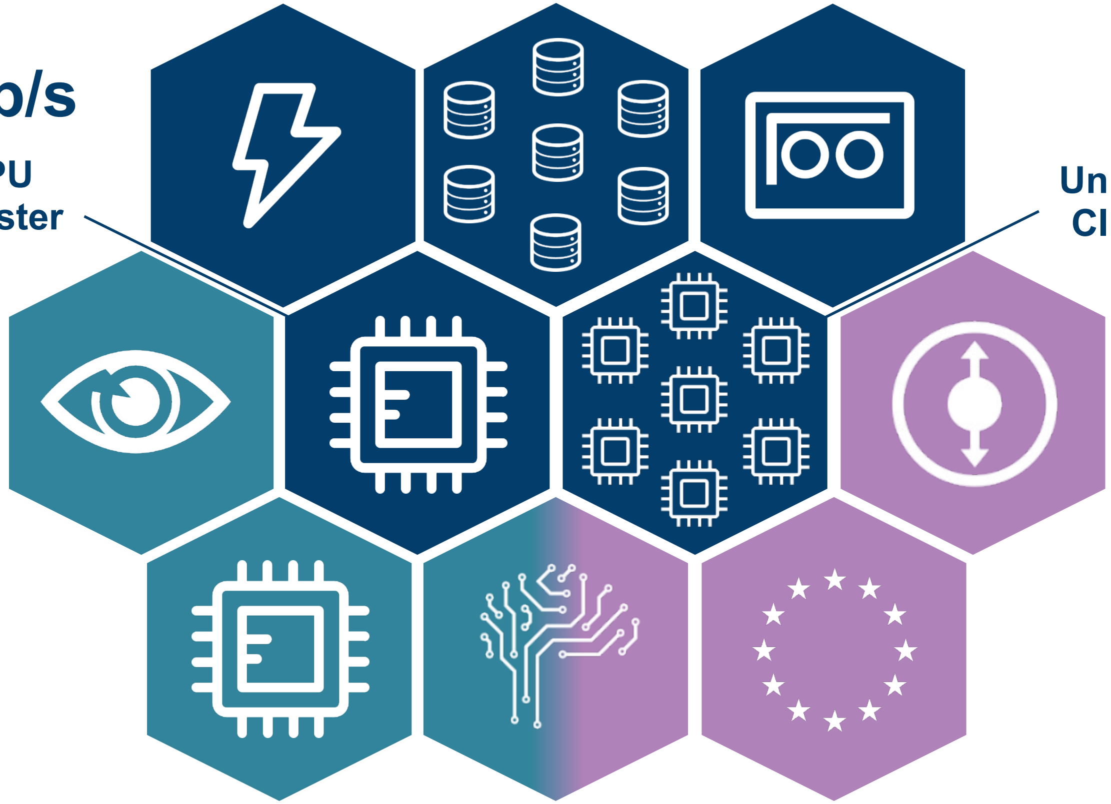
European Processor Initiative (EPI), https://www.european-processor-initiative.eu/

*German Federal Ministry of Education and Research (BMBF)
**Ministry of Culture and Science of the State of North Rhine-Westphalia (MKW NRW)

JUPITER – Modular Exascale Computer

> 1 Exabyte (EB)

> 1 Exaflop/s
GPU



Target >20×
application performance compared to JUWELS Booster

>0.4 B/FLOP**
CPU

Basis Configuration

Optional Modules

Future Technology Modules

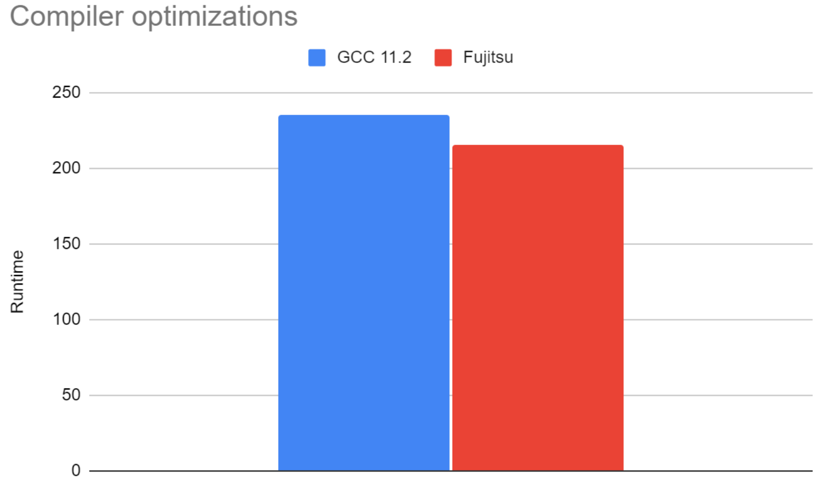
fz-juelich.de/jupiter

*Based on SiPearl's Rhea processor designed in Europe (CPU with high memory bandwidth), developed in the framework of the European Processor Initiative (EPI)
**byte-per-flop (B/FLOP) ratio: amount of data a system can transfer from/to memory for each floating-point operation it performs. Ideal value varies based on the specific workloads and applications (0.4 relatively high, good balance between computation and memory bandwidth)

Prepare Code Porting on European Future Processors

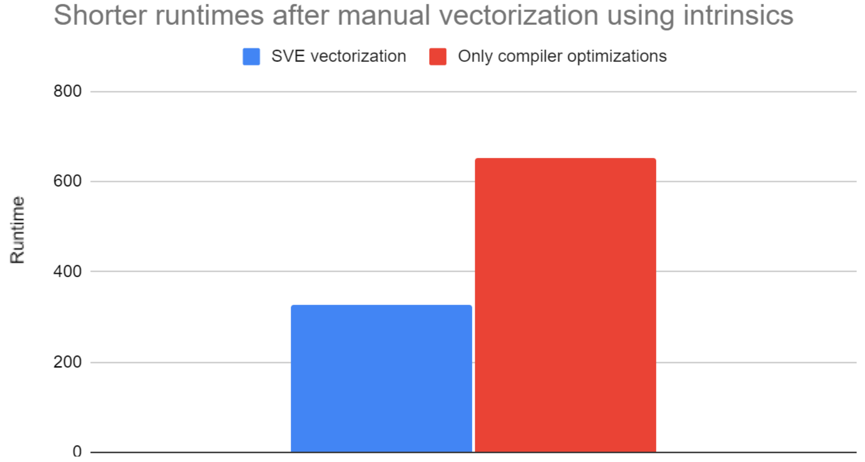
Eg. Highly Parallel Density-based spatial clustering of applications with noise (HPDBSCAN)*

Compiler optimizations (choosing the right compiler)



FUJISTU compiler with right flags
(up to 8% improvement)

Vectorization Developer intervention needed



Change memory layout of the data, use SVE against vectorization by compiler, etc
(100% improvement in runtimes)

European Pilot for Exascale (EUPEX), https://eupex.eu/

*Götz, M., Bodenstein, C., Riedel M., HPDBSCAN: highly parallel DBSCAN, Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments, ACM, 2015.

EUPEX European Pilot for Exascale

- A64FX is a 64-bit ARM architecture microprocessor designed by Fujitsu with different features (Scalable Vector Extension (SVE) vector instruction set, High Bandwidth Memory 2 (HBM2))
- JUPITER will rely on SiPearl's ARM processors

Adoption of Innovative Computing Paradigms

Hybrid use of HPC & Quantum Computing

For practical quantum computing …

the best of both computer technologies must be combined

Kristel Michielsen, "Perspectives of Quantum Computing at the Jülich Supercomputing Centre", JSC seminar, 2023

Hybrid use of HPC & Quantum Computing

- Linking conventional high-performance computers and quantum computers
  - may become common practice in computer centers

- HPC simulations of quantum computers
  - provide essential insight in their operation
  - enable benchmarking and contribute to their design

- Hybrid simulations
  - create new opportunities for challenging computational problems in science and industry

JUNIQ @ Jülich

JSC's Quantum Computing Strategy

HELMHOLTZ
RESEARCH FOR GRAND CHALLENGES
HELMHOLTZ
QUANTUM

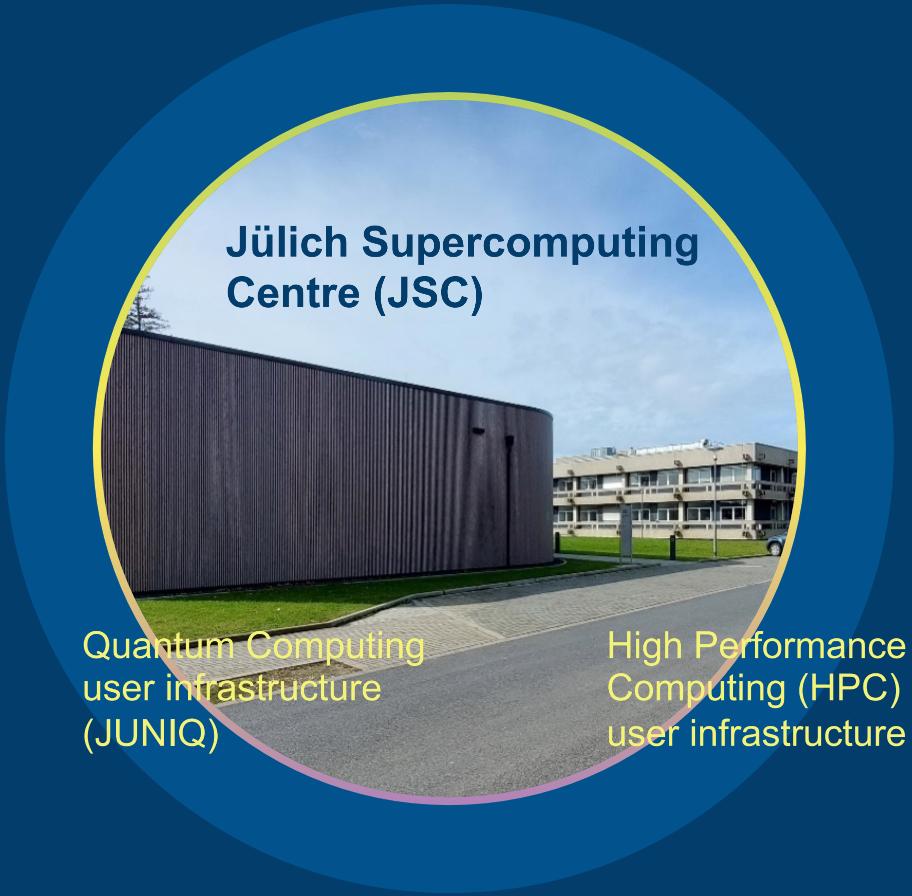
Four Pillars

I. Modeling and emulation (since 2004)

II. Provision of QC systems (since 2016)

III. HPC-QC integration (since 2017)

IV. Creation of a quantum computing user infrastructure (since 2016)



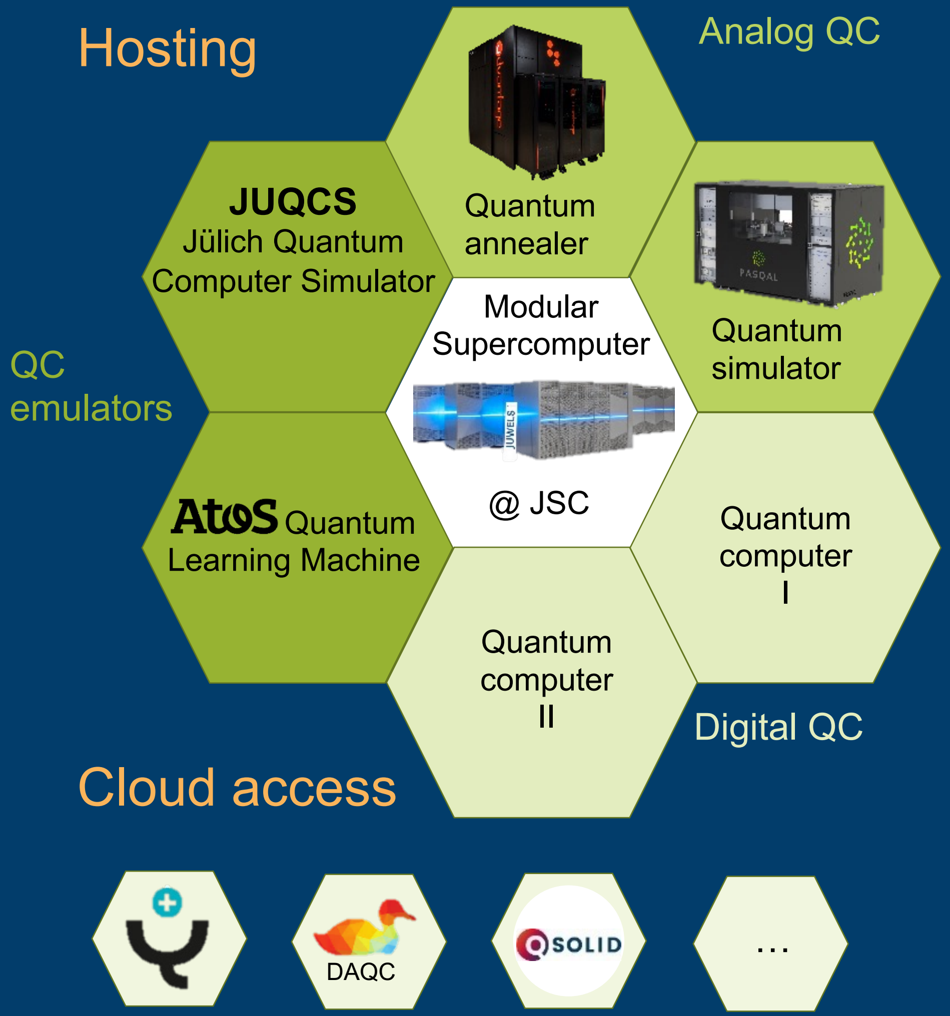
JUNIQ – Jülich UNified Infrastructure for Quantum computing

JÜLICH Forschungszentrum | JUNIQ QUANTUM USER FACILITY

Jülich UNified Infrastructure for Quantum computing (JUNIQ)

Bundesministerium für Bildung und Forschung

Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen



1. QC user facility for science and industry
2. Installation, operation and provision of QCs
3. Unified portal for access to QC emulators and to QC devices at different levels of technological maturity (QC-PaaS)
4. Development of algorithms and prototype applications
5. Services, training and user support
6. Modular quantum-HPC hybrid computing

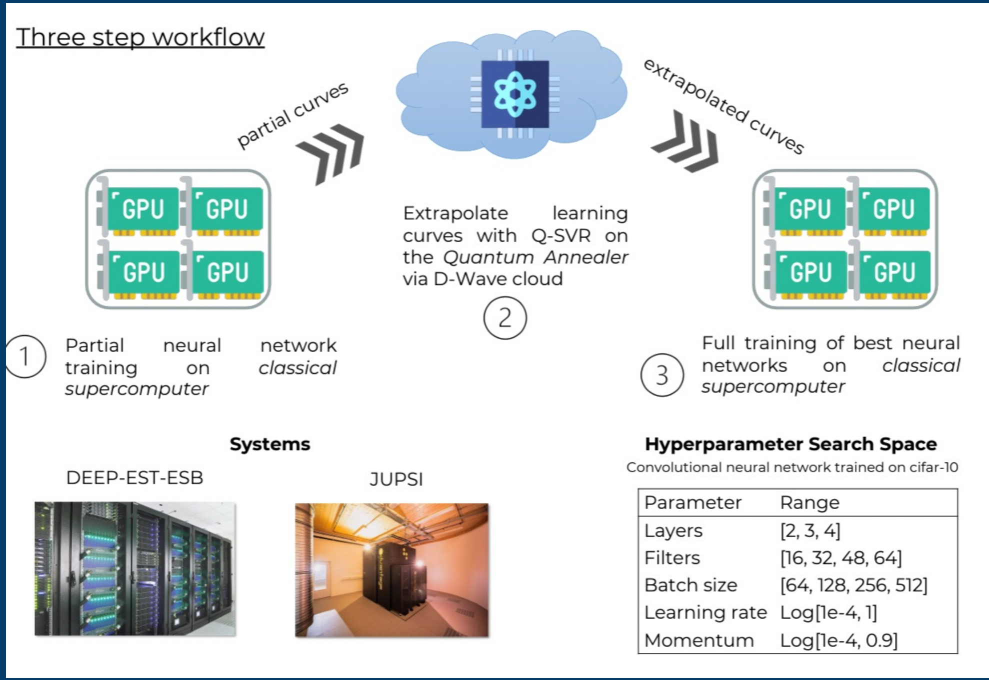
Rolling call for peer-reviewed access:

https://www.fz-juelich.de/ias/jsc/juniq

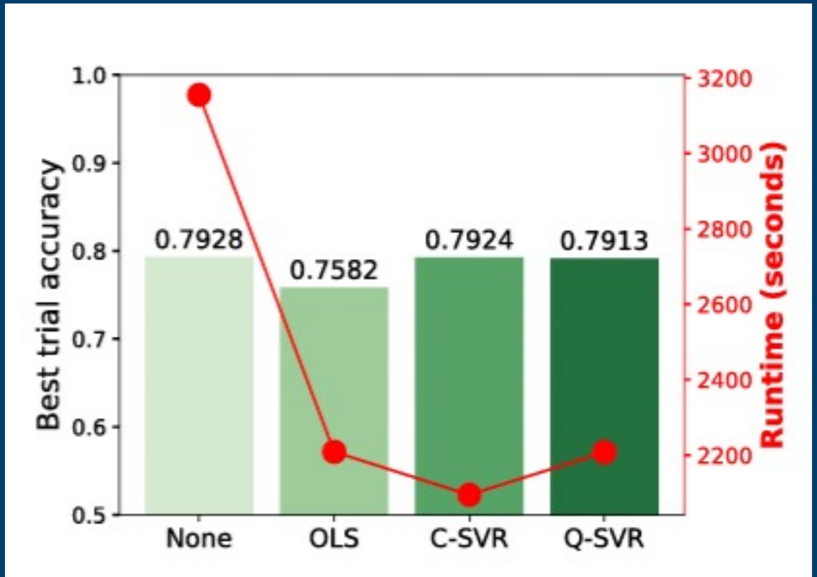
JÜLICH Forschungszentrum | JUNIQ QUANTUM USER FACILITY

Hybrid Quantum-Classical Workflows

Example: Hyperparameter Optimization of Neural Networks



Parameter	Range
Layers	[2, 3, 4]
Filters	[16, 32, 48, 64]
Batch size	[64, 128, 256, 512]
Learning rate	Log[1e-4, 1]
Momentum	Log[1e-4, 0.9]



- Neural network training on classical machine
- Performance prediction on quantum machine
- Save 30% compute resources

European Center of Excellence in Exascale Computing "Research on AI- and Simulation-Based Engineering at Exascale" (CoE RAISE), https://www.coe-raise.eu/

Aach, M, Wulff, E., Pasetto, E., Delilbasic, A., Sarma, R., Inanc, E., Girone, M., Riedel, M. & Lintermann, A., “A Hybrid Quantum-Classical Workflow for Hyperparameter Optimization of Neural Networks”, ISC High Performance 2023, ISC2023, http://hdl.handle.net/2128/34520

Delilbasic, B. Le Saux, M. Riedel, K. Michielsen, G. Cavallaro, "A Single-Step Multiclass SVM based on Quantum Annealing for Remote Sensing Data Classification," 2023, https://doi.org/10.48550/arXiv.2303.11705

E. Pasetto, M. Riedel, K. Michielsen, G. Cavallaro, "Kernel Approximation on a Quantum Annealer for Remote Sensing Regression Tasks", 2023, https://doi.org/10.36227/techrxiv.22794146.v1

E. Pasetto, M. Riedel, F. Melgani, K. Michielsen and G. Cavallaro, "Quantum SVR for Chlorophyll Concentration Estimation in Water With Remote Sensing," in IEEE Geoscience and Remote Sensing Letters (GRSL), vol. 19, pp. 1-5, 2022, https://doi.org/10.1109/LGRS.2022.3200325

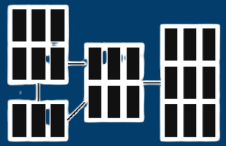
G. Cavallaro, M. Riedel, T. Lippert and K. Michielsen, "Hybrid Quantum-Classical Workflows in Modular Supercomputing Architectures with the Jülich Unified Infrastructure for Quantum Computing," in Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 4149-4152, 2022, https://doi.org/10.1109/IGARSS46834.2022.9883225

Cornerstones of Next-Generation Computing

Develop supercomputing at Exascale

Introduce innovative and unconventional computing technologies

- Quantum Computing, Neuromorphic Computing, In-Memory Computing, etc



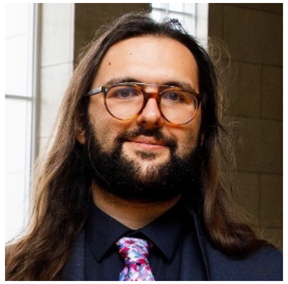
Federated HPC, cloud and data infrastructures for data analytics and AI serving different communities

Comprehensive and efficient support structures and tools

Educate a new interdisciplinary generation of simulation and data science specialists

With thanks to

- Member of the Simulation and Data Lab Remote Sensing
  - At Jülich Supercomputing Centre (JSC) [1] and University of Iceland [2]



Dr. Rocco Sedona

Surbhi Sharma

Amer Delilbasic

Joseph Xavier Arnold

Edoardo Pasetto

Liang Tian

- At JSC
  Estela Suarez, Jens Henrik Göbbert, Kristel Michielsen, Morris Riedel, Stefan Kesselheim, Thomas Lippert, Andreas Lintermann
- Projects and research activities with:

[1] https://www.fz-juelich.de/en/ias/jsc/about-us/structure/simulation-and-data-labs/sdl-ai-ml-remote-sensing
[2] https://ihpc.is/simulation-and-data-lab-remote-sensing/