Modular Supercomputing and its Role in Europe's Exascale Computing Strategy

Sarah M. Neuwirth
Goethe University Frankfurt, Germany
Jülich Supercomputing Centre, FZJ, Germany

Lattice 2022, Bonn, August 2022

Motivation
Performance Development – Top500

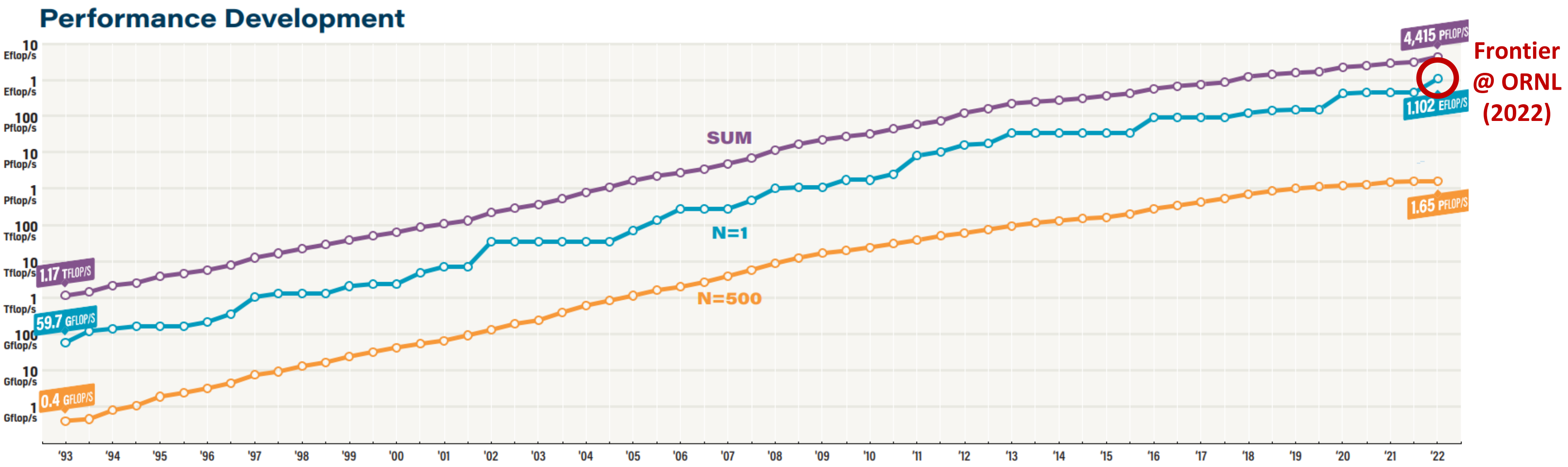


Image source:
top500.org

- 1 Eflop/s (exaFLOP/s) = [10^{18} Floating Point Operations per Second]
- The first exascale supercomputer was expected to enter operation in 2021.
- ***Observation:*** Decline in performance growth => *new architecture paradigm needed* (?)

Motivation
European HPC Ecosystem

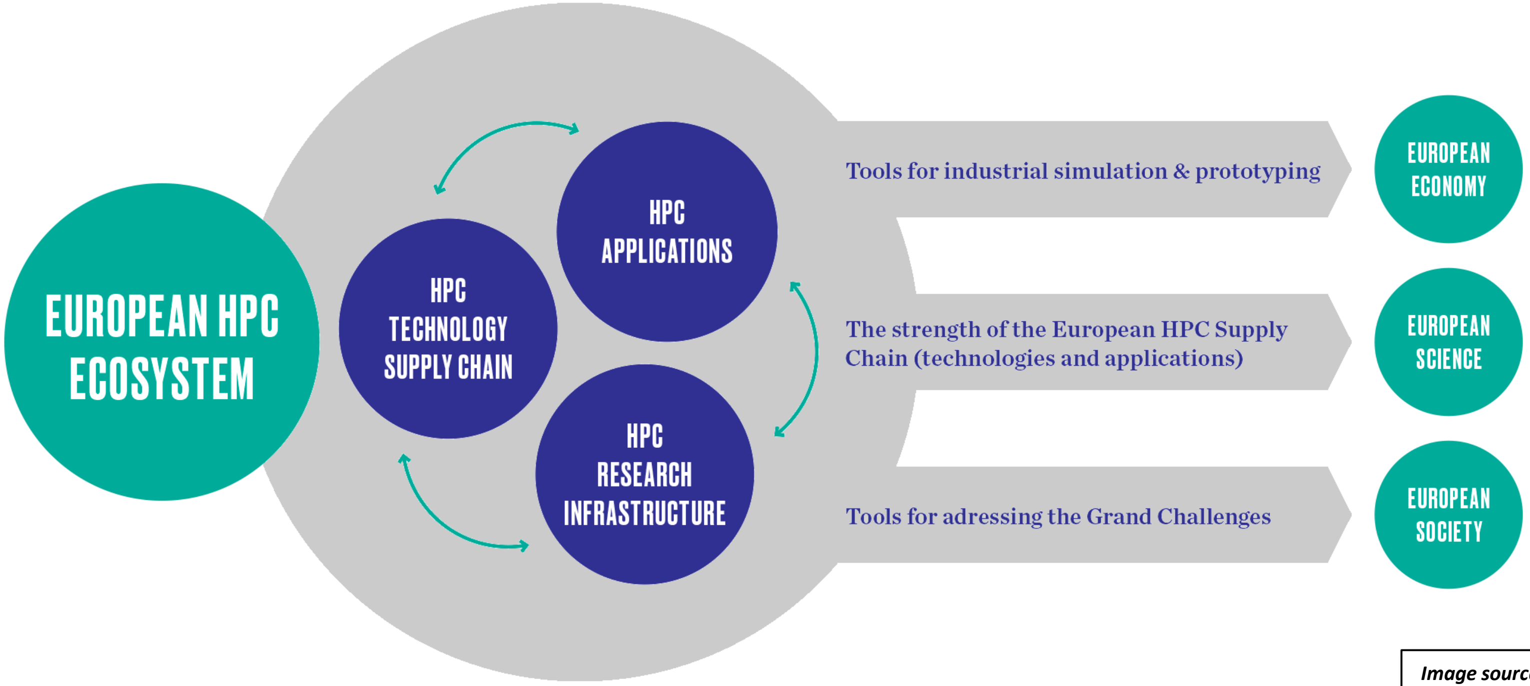
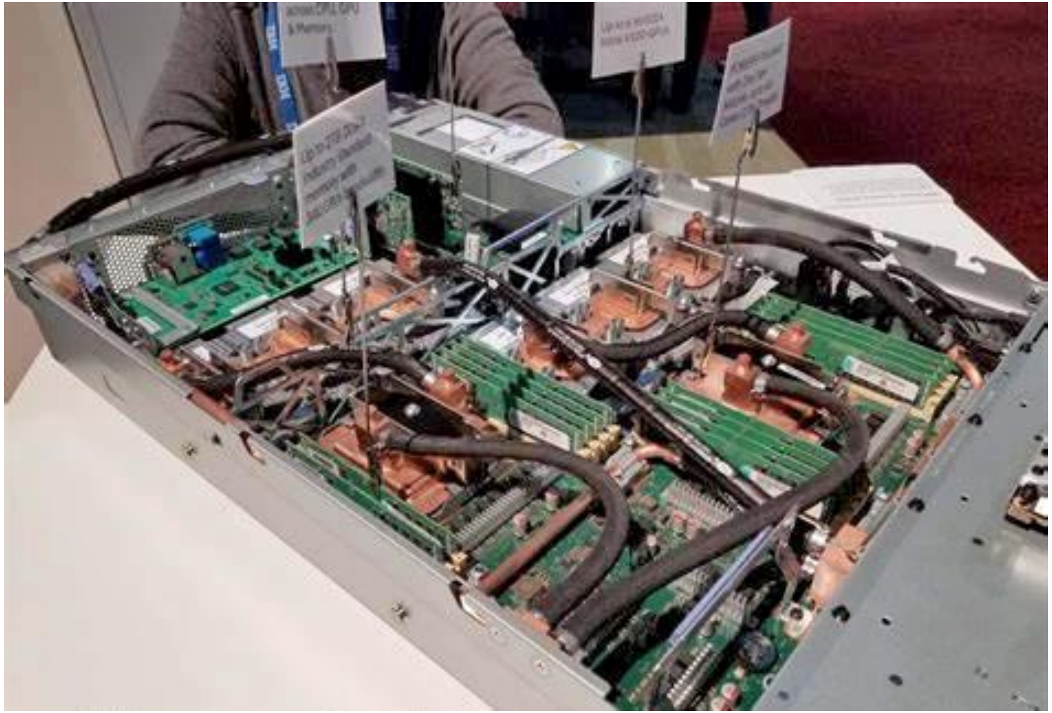


Image source: ETP4HPC

Modular Supercomputing

Modular Supercomputing Architecture

Traditional HPC Concept – Fat Nodes



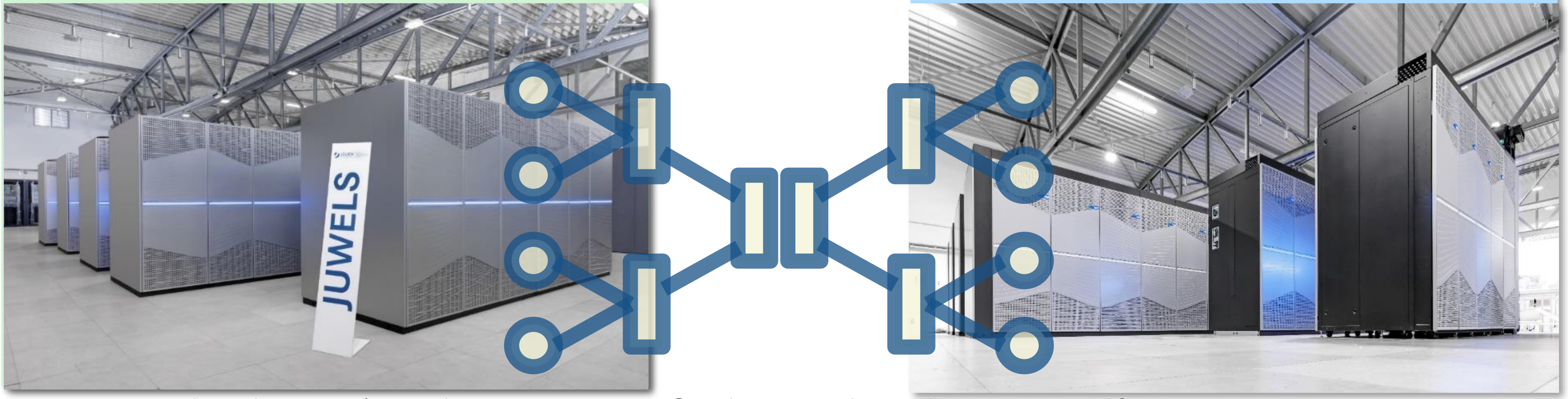
- ***Underutilization***
  - Difficult: CPUs, GPUs, FPGAs etc, on one board concurrently
  - I/O "shared"
- ***Scalability***
  - Computing complex problems collides with scale out
  - Costly: energy and # of nodes
- ***Composability***
  - How to include future computing elements?
  - Quantum, neuromorphic?

Modular Supercomputing Architecture

State-of-the-Art: JUWELS Exascale Pathfinder

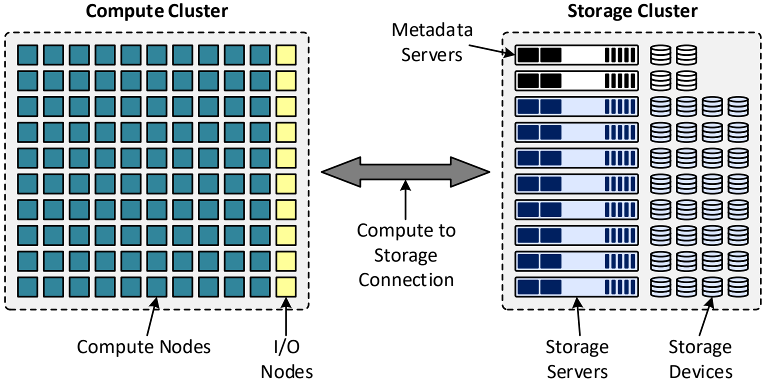
GOETHE UNIVERSITÄT FRANKFURT AM MAIN

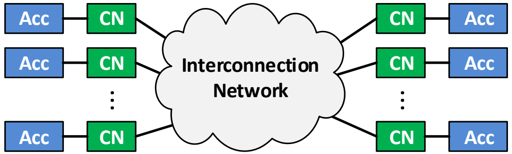
JUWELS Cluster	JUWELS Booster
Intel Xeon (Skylake) processor	AMD EPYC Rome 7402 processor
InfiniBand EDR network	3,700 NVIDIA A100 GPUs
2,500 compute nodes	InfiniBand HDR DragonFly+
10.4 (CPU) + 1.6 (GPU) PFLOP/s peak	73 PFLOP/s peak



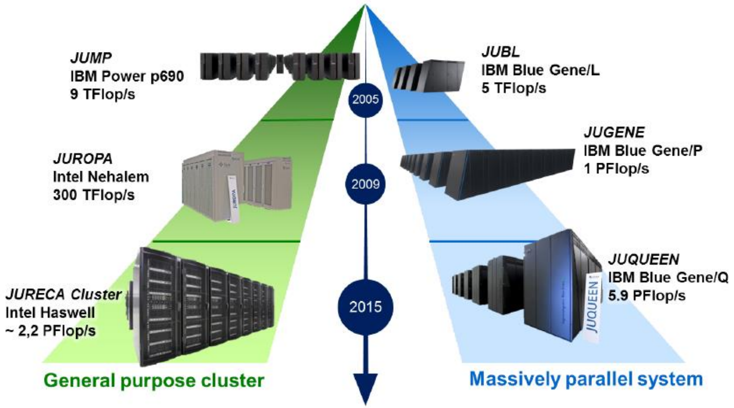
Modular Supercomputing Architecture
Evolution of a European HPC Paradigm

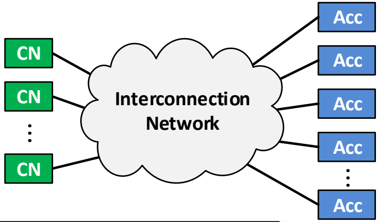
Traditional HPC Systems – Static Concept:





Cluster-Booster Concept* – Dynamic Solution:

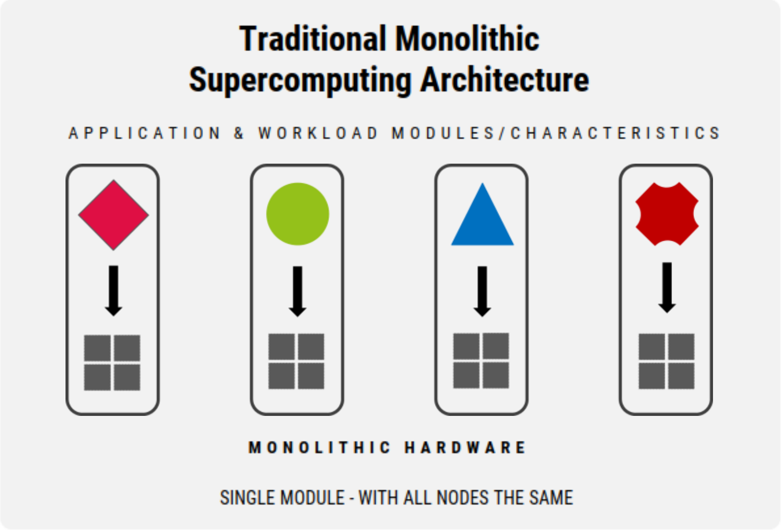


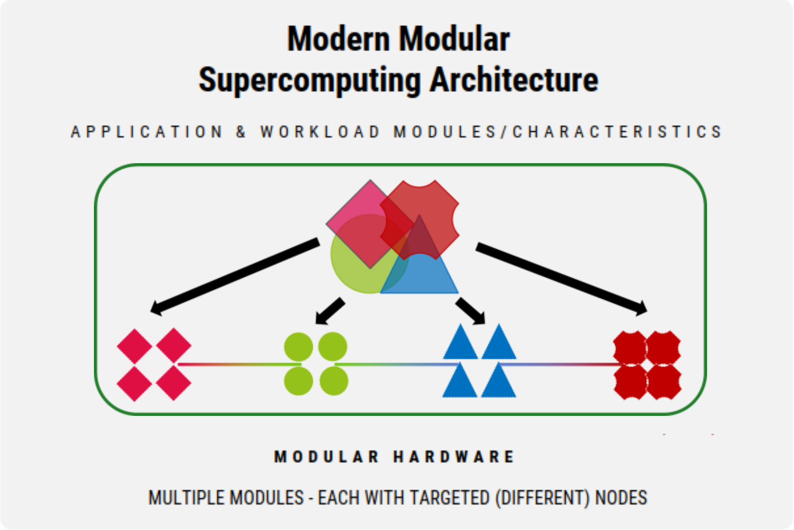


*Eicker, N., Lippert, T., Moschny, T., Suarez, E., *The DEEP Project An alternative approach to heterogeneous cluster-computing in the many-core era*. Concurrency and Computation: Practice and Experience Vol. 28, Issue 8, 2394–2411 (2016).

Modular Supercomputing Architecture
General Concept of MSA

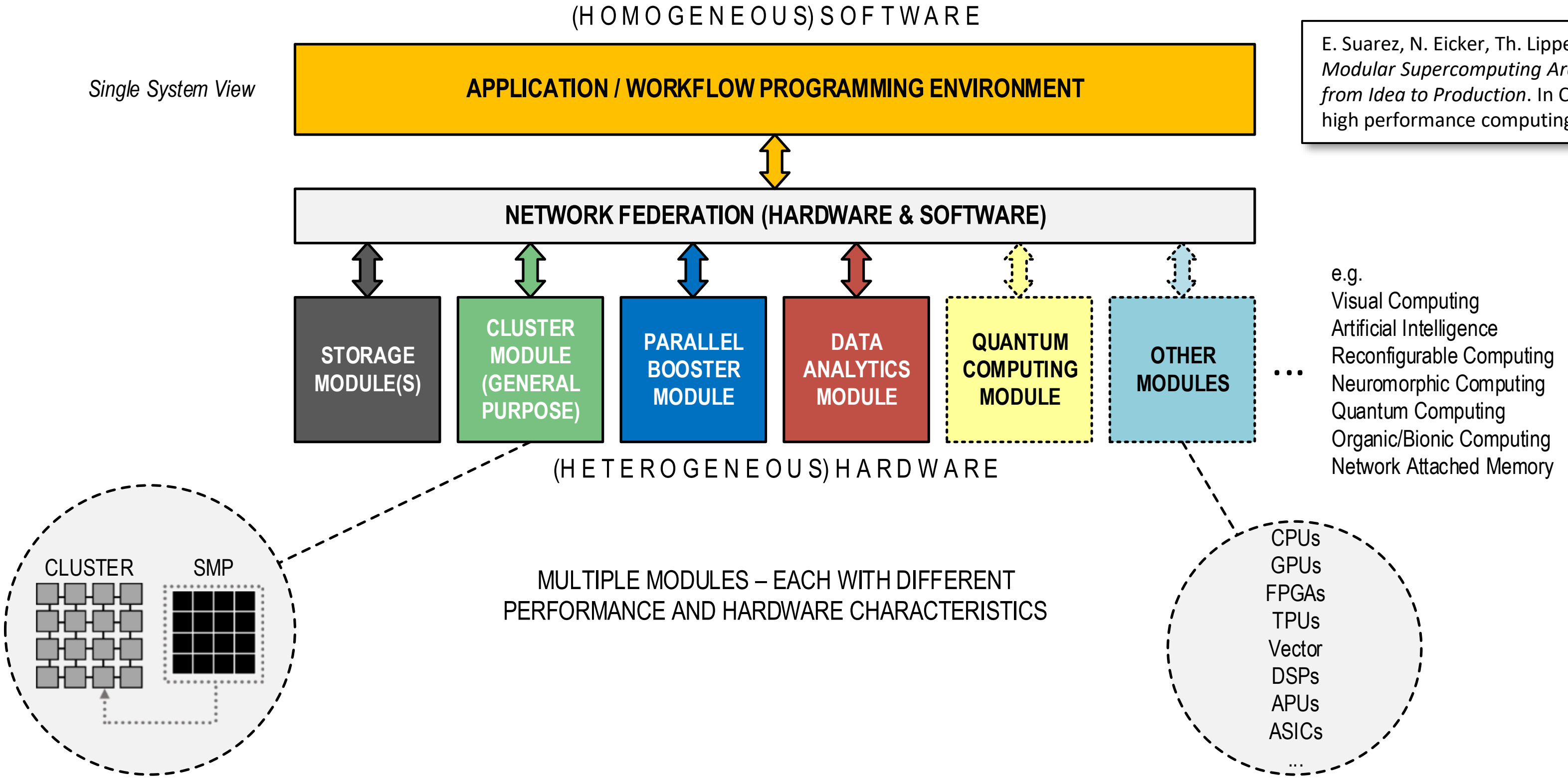
The ***Modular Supercomputing Architecture*** (MSA) concept was invented by the ***Jülich Supercomputing Centre*** (JSC) as a generalization of the previously implemented ***Cluster-Booster*** concept.





Modular Supercomputing Architecture
High-level Illustration of the System Architecture

GOETHE UNIVERSITÄT FRANKFURT AM MAIN



E. Suarez, N. Eicker, Th. Lippert (2019). *Modular Supercomputing Architecture: from Idea to Production*. In Contemporary high performance computing, CRC Press.

Modular Supercomputing Architecture
Amdahl's Law – The Simple Case

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

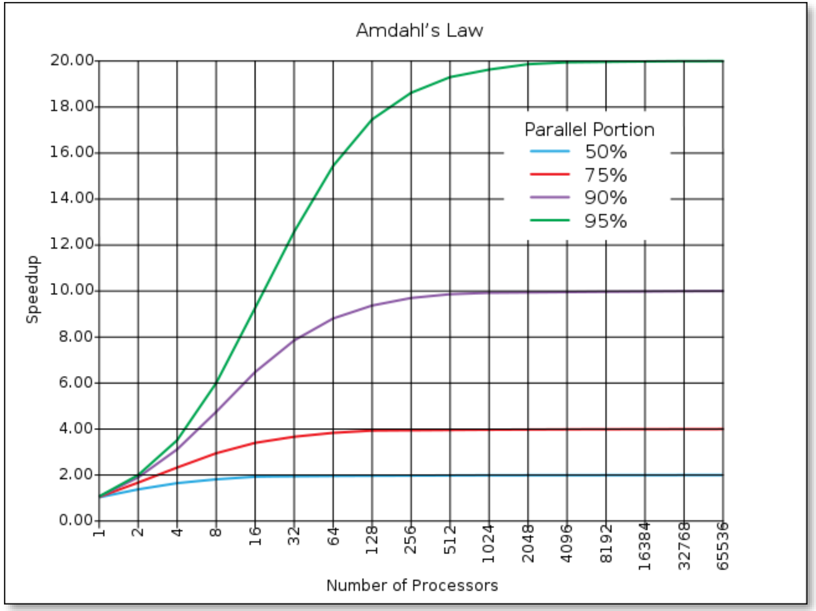
- ***Amdahl's Law**** states that the speedup of a code is always limited by the sequential part
- Considers a time-to-solution problem for a fixed problem size => *Strong Scaling*
- Speedup S with N processors is given by

$$S = \frac{1}{s + p/N}$$

- Scaling on infinite nr. of processors $N \rightarrow \infty$:

$$S_\infty = \lim_{N\rightarrow\infty} \frac{1}{s + p/N} = \frac{1}{s}$$

- ***But:*** The number of useful processors is limited by the critical path!



*G. M. Amdahl, "*Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities*," in Proceedings of the April 18-20, 1967, Spring Joint Computer Conference, AFIPS '67 (Spring), p. 483–485, Association for Computing Machinery, 1967.

Modular Supercomputing Architecture
Generalized Amdahl's Law

- Let a code have $r-1$ parts with different concurrencies $N_i < N_h = N_r$
  - The r^{th} code part has concurrency $N_h = N_r$, that Amdahl likes to scale to infnity
  - For simplicity, all N_i are defined for the r code parts in consideration
    - as the numbers where 80 % of the maximal parallel speed-up has been achieved
    - or where the critical path does not allow for more processors
- In total, all portions add up to 1: $\sum_{i=1}^{r} p_i = 1$
- The ***Generalized Amdahl's Law (GAL)*** becomes

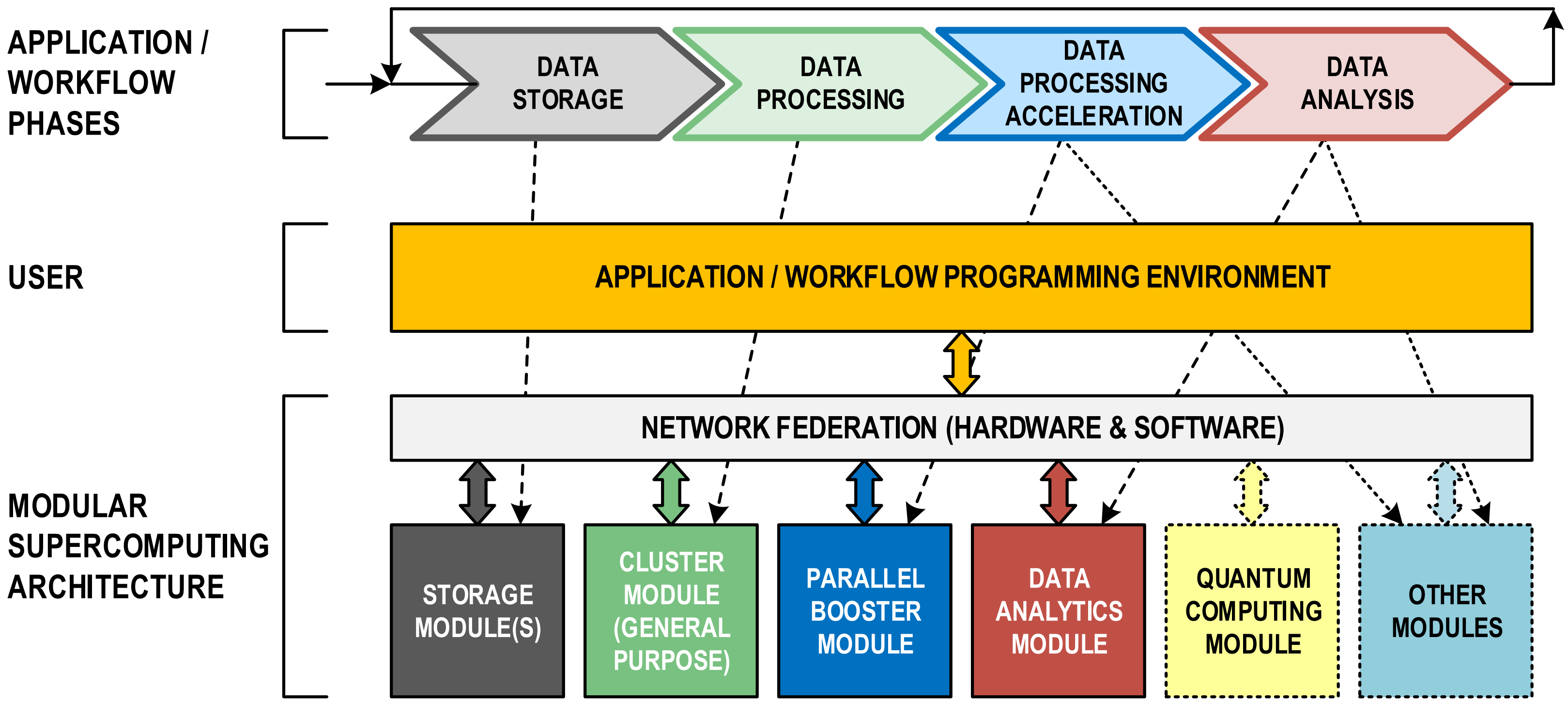
$$S^{GAL} = \frac{T_{single}}{T_{parallel}} = \frac{\sum_{i=1}^{r} p_i}{\sum_{i=1}^{r} \frac{p_i}{N_i}} = \frac{1}{\sum_{i=1}^{r} \frac{p_i}{N_i}}$$

- Asymptotically, the speedup is dominated by a lower concurrency N_d:

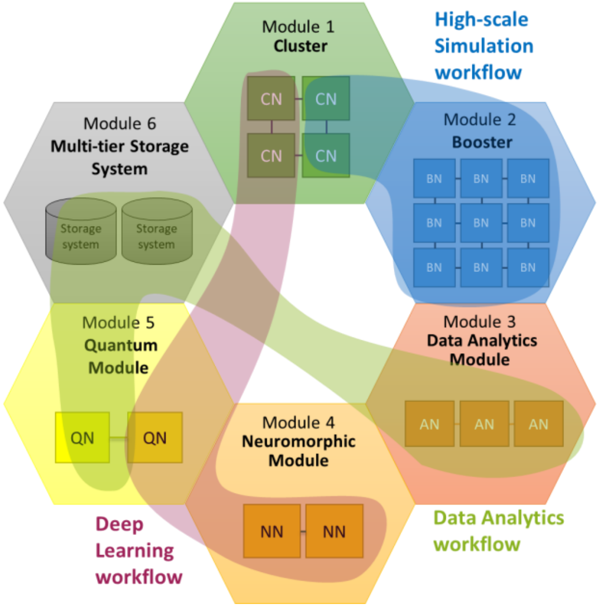
$$S^{GAL} = \frac{1}{\sum_{i=1}^{d-1} \frac{p_i}{N_i} + \frac{p_d}{N_d} + \frac{p_h}{N_h}} \rightarrow \frac{1}{\frac{p_d}{N_d} + \frac{p_h}{N_h}}$$

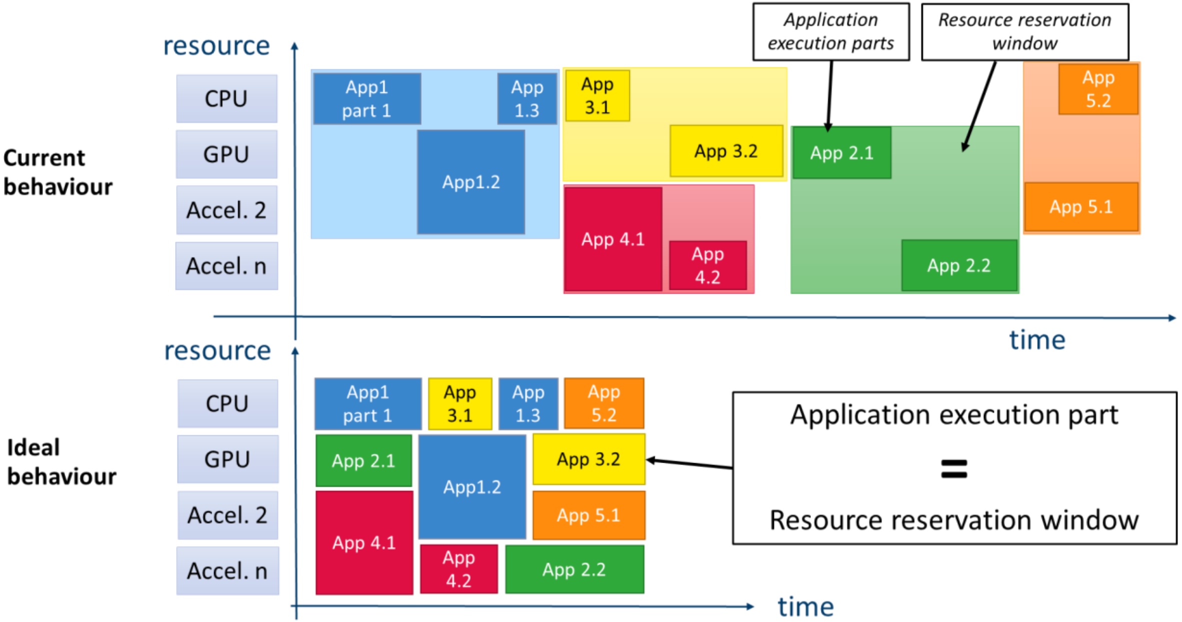
$$S_{N_n \to \infty}^{GAL} = \frac{N_d}{p_d}$$

Modular Supercomputing Architecture
Application and Workflow Mapping



Modular Supercomputing Architecture
Resource Management

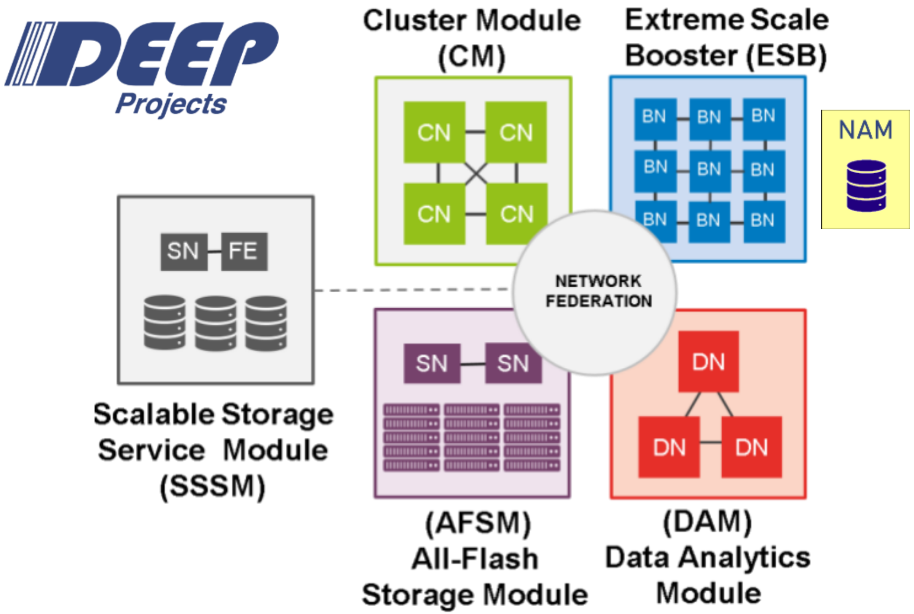




DEEP-EST Prototype System

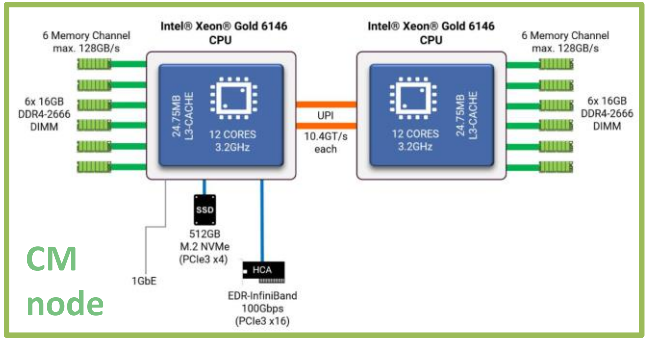
DEEP-EST Prototype System
First MSA Prototype @ JSC

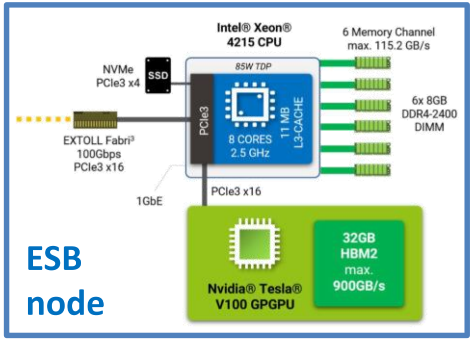
Modular Supercomputing prototype developed within the DEEP-EST project.

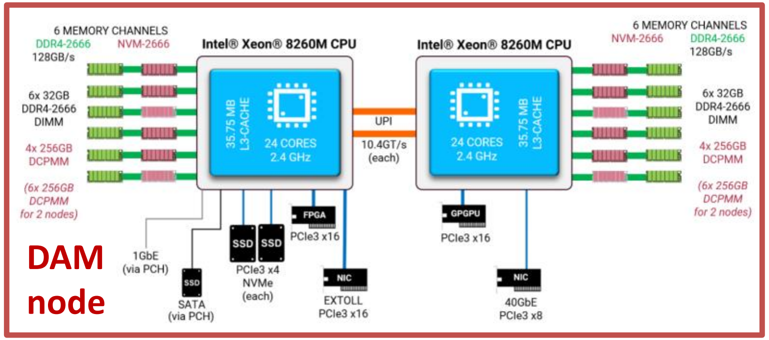


System architecture from the DEEP system, implementing the Modular Supercomputing Architecture (MSA).

DEEP-EST Prototype System
System Overview and Usage Targets







Cluster Module:
Applications and code parts requiring high single-thread performance and a modest amount of memory.
=> *typically moderate scalability*

Extreme Scale Booster:
Compute intensive applications and code parts with regular control and data structures.
=> *high parallel scalability*

Data Analytics Module:
Data-intensive analytics and machine learning applications and code parts requiring large memory capacity, data streaming, bit- or small datatype processing.

Suarez, E., Kreuzer, A., Eicker, N., Lippert, Th., *The DEEP-EST project*, 2021, In Porting applications to a Modular Supercomputer - Experiences from the DEEP-EST project. Online: https://juser.fz-juelich.de/record/905812

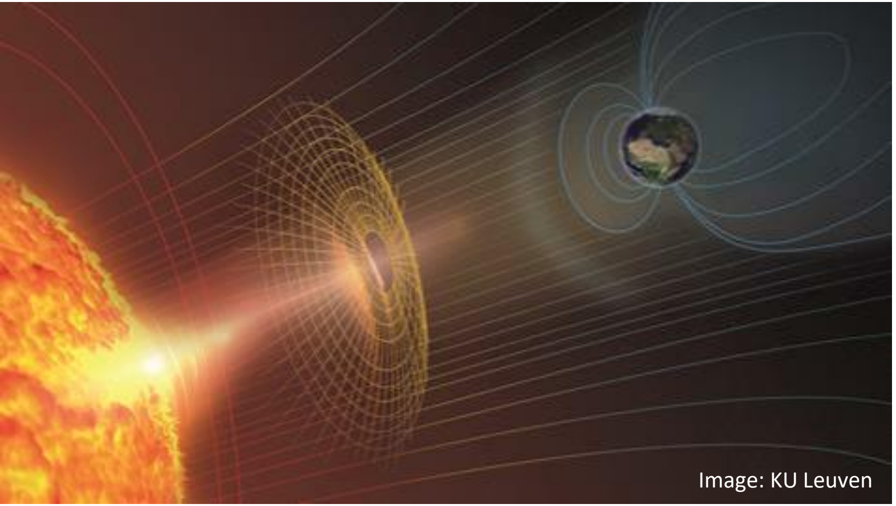
DEEP-EST Prototype System
*Set of Six Applications Ported**

- ***Neuroscience***
  - *NEST:* Simulation of point-like neurons
  - *Arbor:* Simulation of detailed neurons
  - *Elephant:* Analysis of electrophysiological experiments
- Molecular Dynamics – *the most widely used code*
- Radio Astronomy
  - *LOFAR* correlator and imager
  - *SKA* – most important astronomy project to come
- ***Space Weather***
  - CBA-application with modular extension: xPic
- Data Analytics in Earth Science
  - Clustering of big data by PiSVM
- CERN: High Energy Physics
  - Reconstruction workflows on GPUs, FPGAs

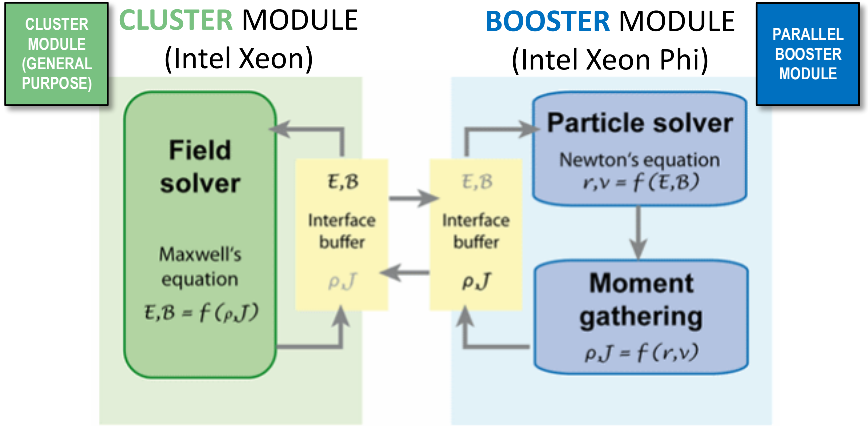
*A. Kreuzer, Th. Lippert, E. Suarez, and N. Eicker (2021). *Porting Applications to a Modular Supercomputer – Experiences from the DEEP-EST Project*. http://hdl.handle.net/2128/30498

DEEP-EST Prototype System

Example: Space Weather Simulation



- Simulates plasma produced in solar eruptions and its interaction with the Earth magnetosphere
- Particle-in-Cell code: ***xPIC***
- Author: ***KU Leuven***, Belgium



computes electromagnetic (EM) field evolution
(low/medium scalable code)

calculates motion of charged particles in EM field
(highly scalable code)

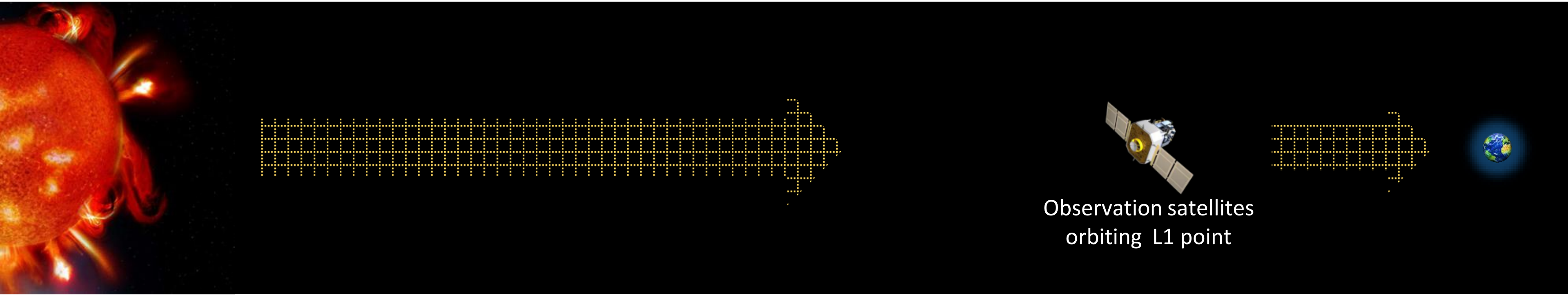
35% gain in performance* for ***combined Cluster and Booster*** system compared to same size homogenous system

*Kreuzer, A.; Eicker, N.; Amaya, J.; Suarez, E., "*Application Performance on a Cluster-Booster System*", 2018 IEEE IPDPS Workshops (IPDPSW), Vancouver, Canada, pp. 69 - 78 (2018).

DEEP-EST Prototype System
Example: Space Weather Simulation on MSA

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Sun — 1.5×10^8 km — **L1 Satellites** — 1.5×10^6 km — **Earth**



Data Analytics
Forecast solar wind conditions at L1 from remote image analysis of the Sun

Space Weather Prediction
Detailed physics simulations of the Earth environment given the solar wind conditions at L1

Pre-Processing — **DATA ANALYTICS MODULE** — Training — DLMOS APPLICATION — Inference — **PARALLEL BOOSTER MODULE**

DL analysis triggers full simulation

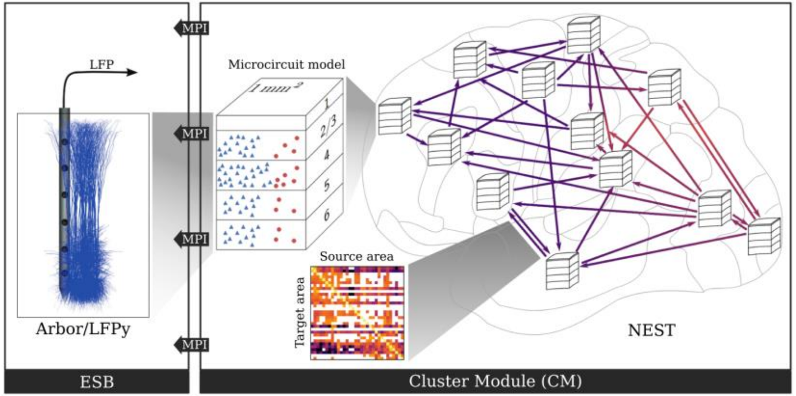
CLUSTER MODULE (GENERAL PURPOSE) — Field Solver — xPic APPLICATION — Practice Solver — **PARALLEL BOOSTER MODULE**

DEEP-EST Prototype System
Example: Simulation of Large-scale Brain Activity

- ***Scientific goal:*** multi-scale simulation of the brain
  - Understand brain as extremely energy-efficient processor
  - Cure mental health problems
- ***Scientific computing challenges in neuroscience***
  - Simulating networks of 10^{11} neurons with 10^4 synapses each
  - Embed simulations of detailed neuron models
  - Analyze data from millions of point processes
- ***Applications:*** NEST, Arbor, Elephant

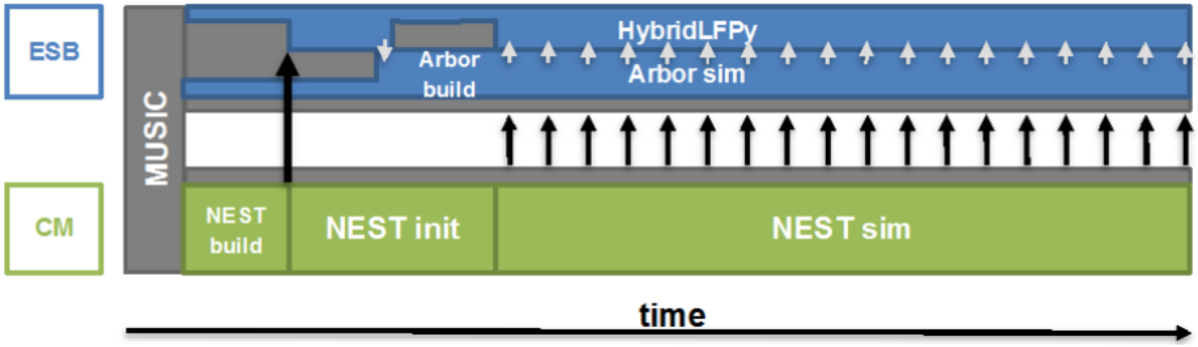
=> The long-term goal of the neuroscience work on MSA is to provide an optimized setup for the integrated simulation and analysis of large-scale brain activity.

Image Source: Suarez, Estela, Susanne Kunkel, Hans Ekkehard Plesser, Anne Küsters, and Thomas Lippert. "*Modular Supercomputing for Neuroscience*." In BrainComp 2019-Workshop on Brain-Inspired Computing, no. FZJ-2019-06093. Jülich Supercomputing Center, 2021.

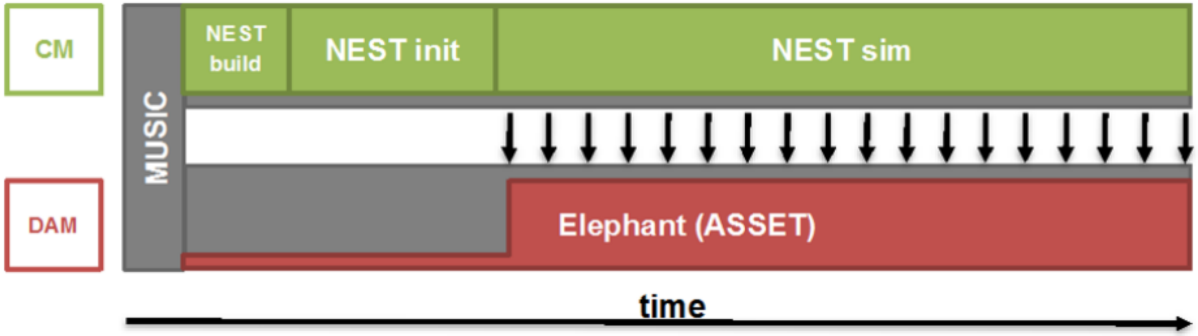


DEEP-EST Prototype System

Example: Simulation of Brain Activity on MSA



Schematic workflow of NEST and Arbor/HybridLFPY in the MSA.

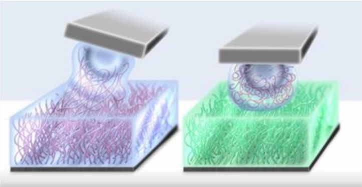


Schematic workflow of NEST and Elephant (ASSET) in the MSA.

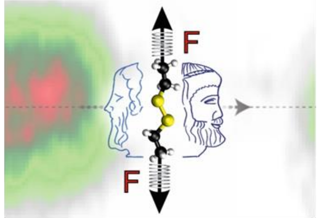
Europe’s Exascale Computing Strategy

Europe's Exascale Computing Strategy
Breakthroughs @ Petascale

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

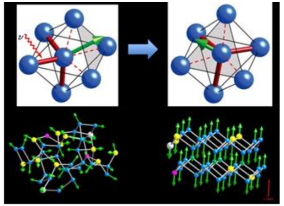


Sissi de Beer, **Solvent-induced immiscibility of polymer brushes eliminates dissipation channels,** Nature Communications 5, 4781

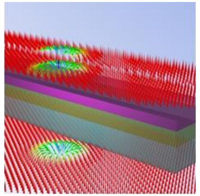


D. Marx et al., **Force induced conformational change** Nature Chemistry 5 (2013) 685

Z. Fodor, JK. Szabo, et al, **Calculation of the axion mass**. *Nature* 539, 69-71 (2016).

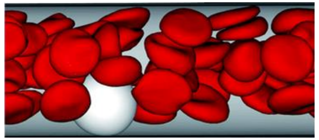


R.O. Jones et al., **One order of magnitude faster phase change at reduced power in Ti-Sb-Te,** Nature Materials 10 (2011) 129

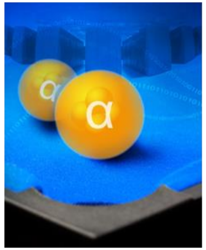


B. Dupé, et al., **Engineering skyrmions in transition-metal multilayers for spintronics**, Nat. Commun.

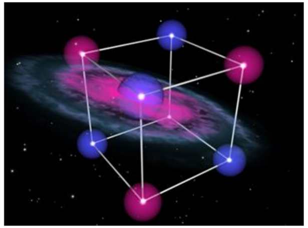
Kalmán Szabo et al., **Ab initio calculation of the neutron-proton mass difference,** Science 347 (2015) 6229



D.A. Fedosov, G. Gompper, **White blood cell margination in microcirculation**. Soft matter 10(17), 2961 -2970 (2014)



S. Elhatisari, et al., **Ab initio αα scattering**, Nature 528, 111 (2015). [do10.1038/nature16067].



M. Lezaic et al., **A multiferroic material to search for the permanent electric dipole moment of the electron,** Nature Materials 9 (2010) 649

Amunts, K. et al., **BigBrain - an ultra-high resolution 3D human brain model**. Science Vol. 340 (2013)no. 6139 pp. 1472-1475.

Europe’s Exascale Computing Strategy

The Exascale Race

2021/22 (1?)

Cost for a system: n/a

Integrator: China
Processors: China

2022 (1)

Cost for a system: 600 M$

Integrator: US
Processors: US

2022/23 (2)

Cost for a system: 900 M$

Integrator: Japan
Processors: Japan

2023/25 (3)

Cost for a system: 600 M€

Integrator: TBD
Processors: EU,US

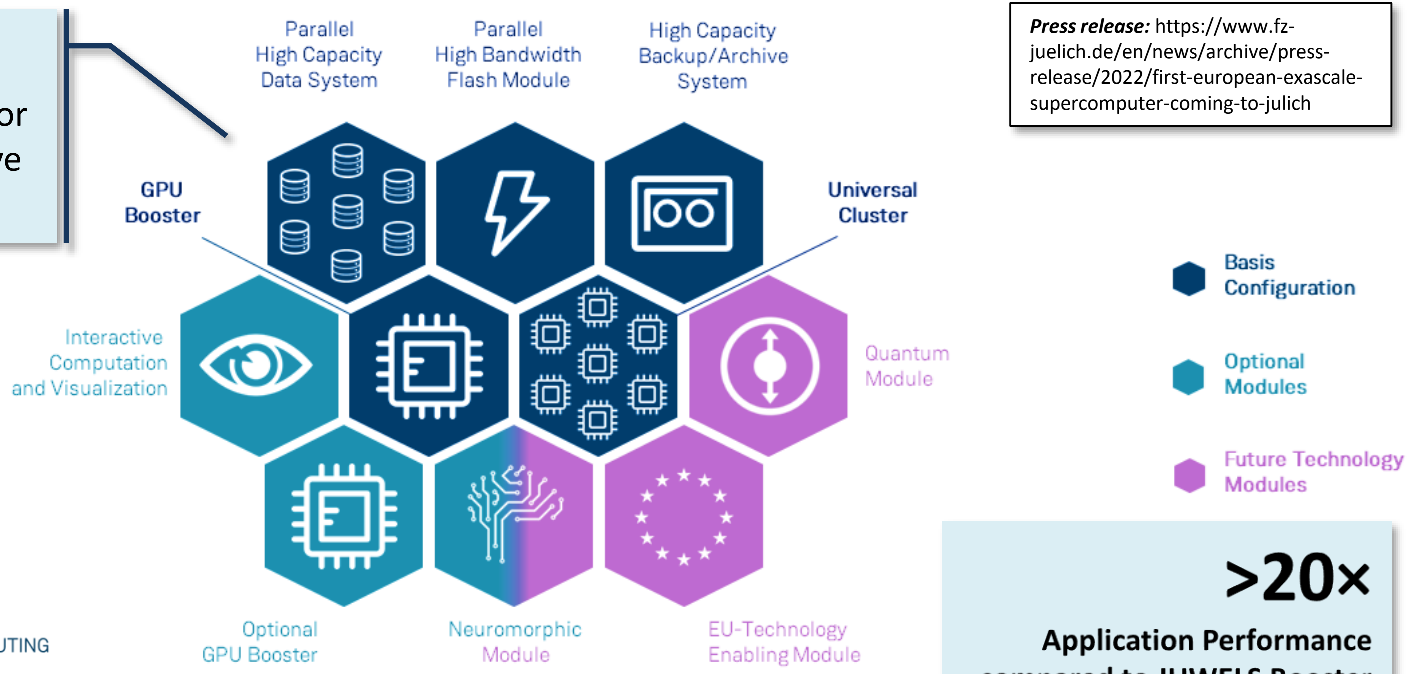
Europe's Exascale Computing Strategy
Expected Breakthroughs by Exascale

- Fundamental Sciences
  - ➢ Astrophysics, Cosmology, Particle Physics
- Climate, Weather, and Earth Sciences
  - ➢ Climate Change, Meteorology, Oceanography, Solid Earth Sciences
- Life Sciences
  - ➢ Bioinformatics, Systems& Structural Biology, Neuroscience
- Energy
  - ➢ Renewable Energy, Fusion Energy, Sustainable Energy
- Infrastructure & Manufacturing
  - ➢ Engineering, Integrative Design, Manufacturing
- Future Materials
  - ➢ Atomic and Electronic Structure, Data-driven Materials Design
- Complexity & Data
  - ➢ AI, Deep Learning, GANs, Convergence with Simulation

PRACE Scientific Steering Committee, "The Scientific Case for Computing in Europe 2018-2026." (2018). Online: https://prace-ri.eu/about/scientific-case/

Europe’s Exascale Computing Strategy
A Modular Exascale Concept

JUPITER @ JSC
(“Joint Undertaking Pioneer for Innovative and Transformative Exascale Research”)



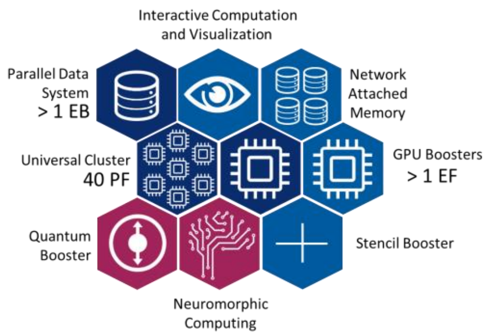
Press release: https://www.fz-juelich.de/en/news/archive/press-release/2022/first-european-exascale-supercomputer-coming-to-julich

>20×
Application Performance compared to JUWELS Booster

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

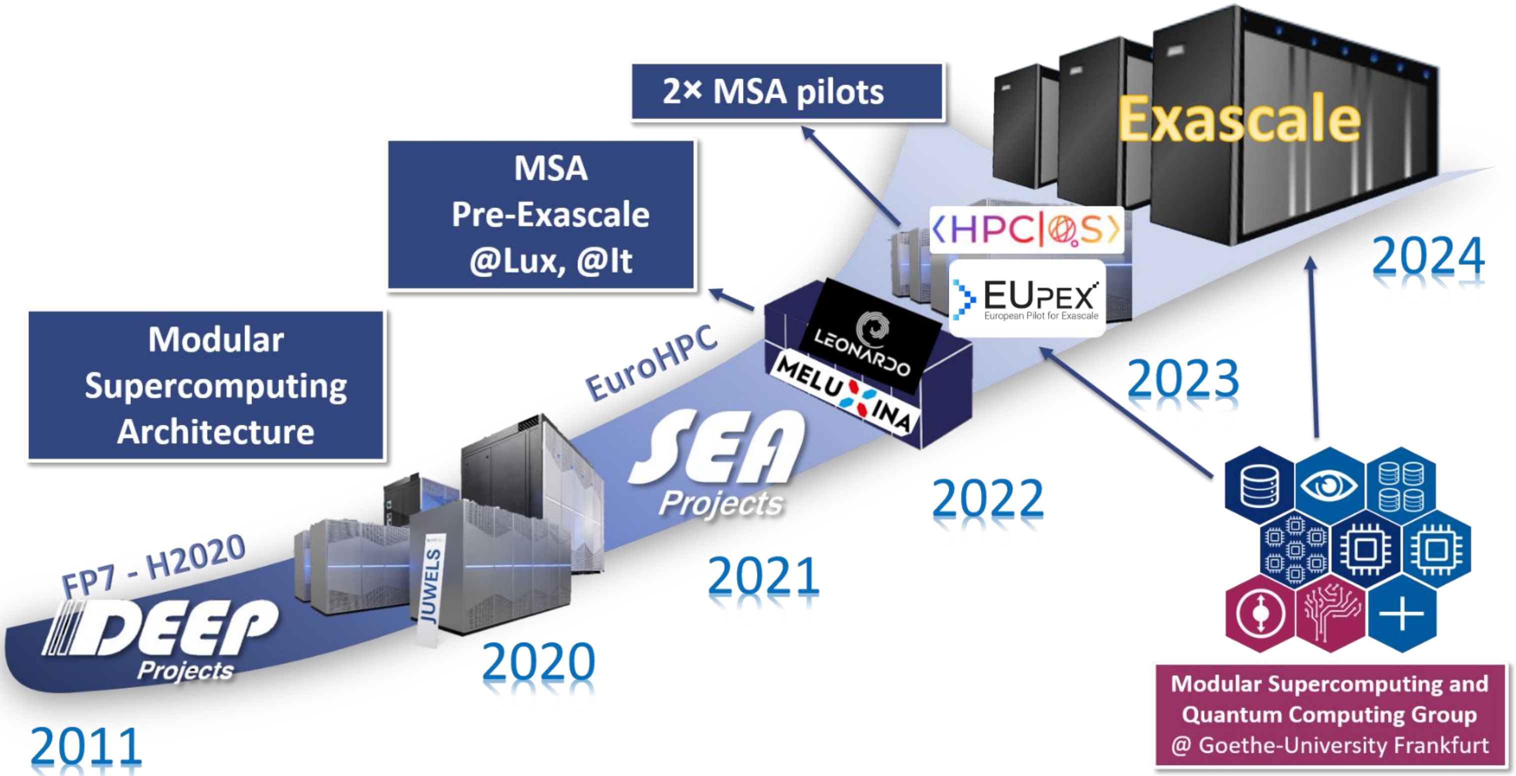
Europe's Exascale Computing Strategy
Target Definition Exascale

- ***Basis:*** JUWELS Booster
  - Basic unit: partition with 50 PetaFLOP/s peak performance on JUWELS
  - Each benchmark application $i = 1, \ldots, N$ characterized by a value P_i
  - Criteria: for example, Performance / Watt, Energy to Solution, TCO to Solution
- ***Target:*** Exascale Supercomputer Booster
  - Each benchmark application i reaches a value $P_{exa,i} = 20P_i$
  - Requires higher peak performance than ExaFLOP/s …



Europe's Exascale Computing Strategy
Contributions of the Goethe University

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

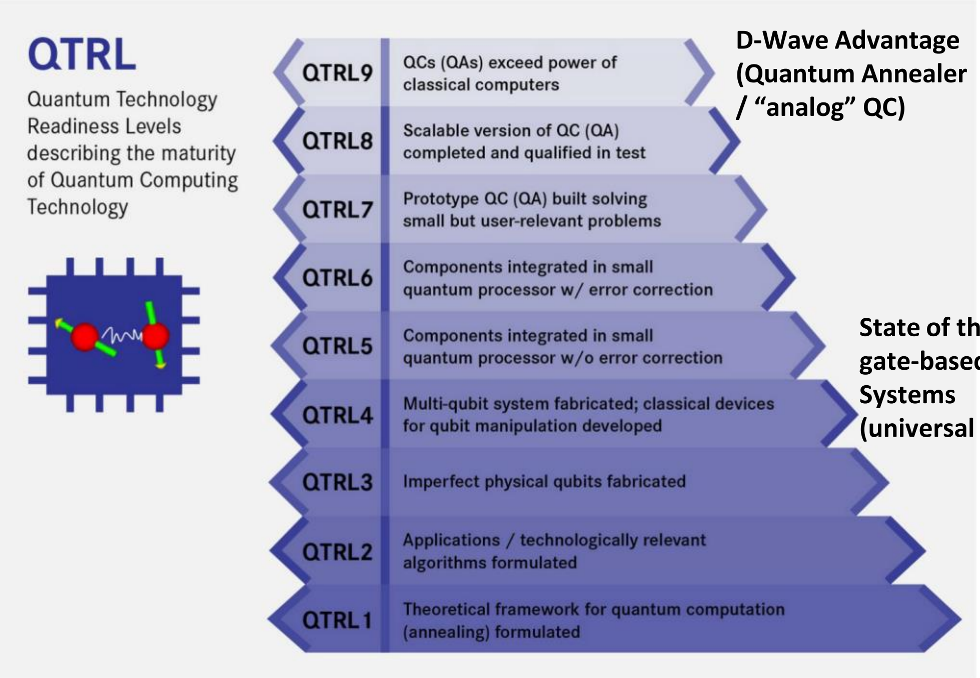


A Quantum Future for HPC?

A Quantum Future for HPC?
Quantum Technology Readiness Levels (QTRL)

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

- Number of *qubits* in *QPUs* has been growing exponentially
- *Variational Quantum Algorithms** (VQA) are designed to keep the circuit depths low
- VQAs are *hybrid algorithms* with classical and quantum part
  - Quantum processor evaluates a cost function which represents the problem
  - Classical part optimizes the parameters of the quantum circuit to find an optimal solution
  - Alternating use of classical and quantum processor requires a close connection



D-Wave Advantage (Quantum Annealer / "analog" QC)

State of the art gate-based Systems (universal QCs)

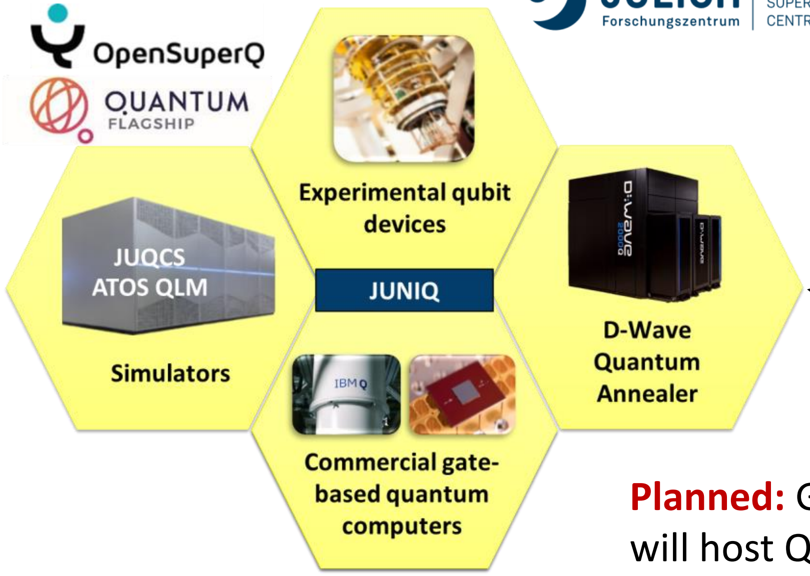
*M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles, "*Variational quantum algorithms*," Nature Reviews Physics, vol. 3, pp. 625–644, Sep 2021.

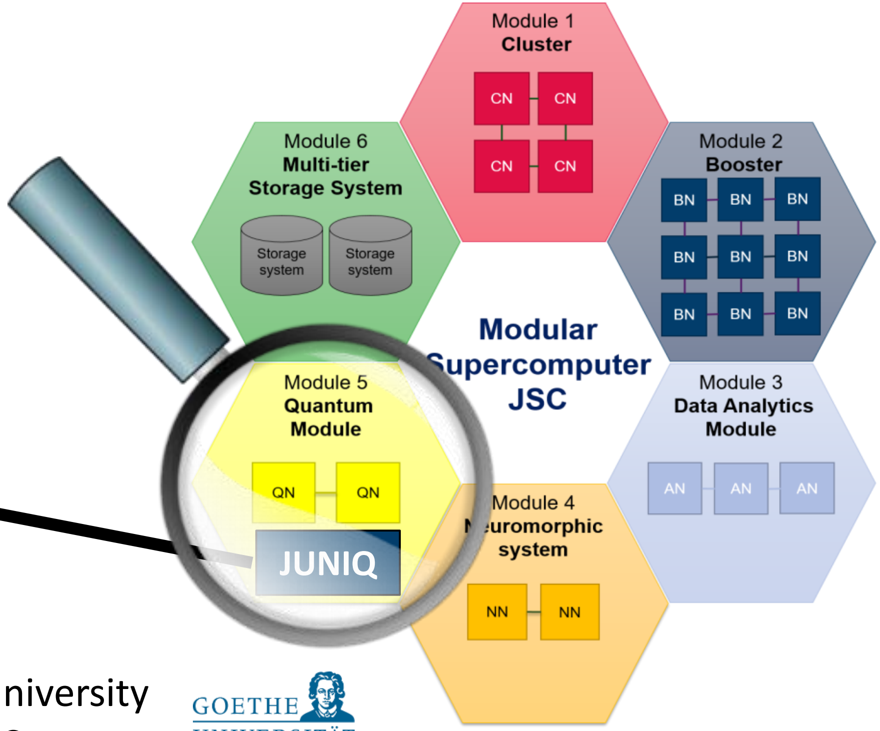
A Quantum Future for HPC?

JUNIQ @ JSC: User Facility for Quantum Computing

JUNIQ - Jülich User INfrastructure for Quantum computing

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE





Planned: Goethe University will host Quantum Computer

Open Research Challenges and Ongoing Work at Goethe University Frankfurt

Challenges and Future Directions
Research Projects Addressing MSA

DEEP-SEA: DEEP Software for Exascale Architectures

- Better manage and program compute and memory heterogeneity
- Targets easier programming for modular supercomputers
- Continuation of the DEEP project series

IO-SEA: Input/Output Software for Exascale Architectures

- Improve I/O and data management in large-scale MSA systems
- Builds upon results of SAGE 1-2 projects and MAESTRO

RED-SEA: Network Solution for Exascale Architectures

- Develop European network solution
- Focus on BXI (Bull eXascale Interconnect)

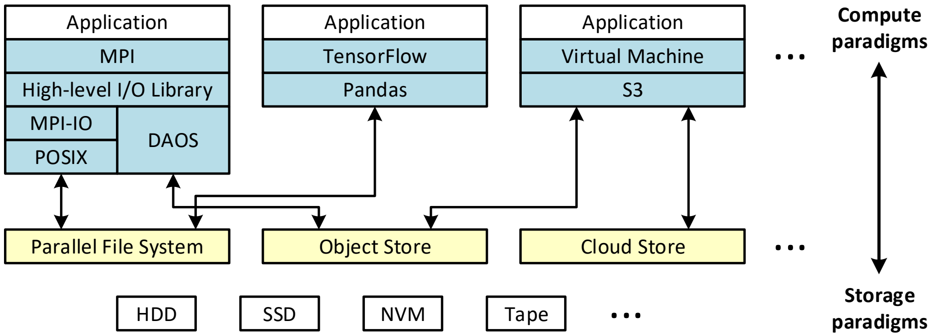
<u>Coordinated with on-going EU projects:</u>

EUPEX European Pilot for Exascale

epi European Processor Initiative

Challenges and Future Directions
Exascale Data Challenges

❶ **System scalability:**
Future supercomputers may include hundreds of thousands of nodes and data is to be accessed by ~10^6 clients. Traditional parallel file systems cannot operate efficiently at this scale.
=> *Access to data becomes a critical issue*



❷ **Data scalability:** bigger machines mean more data, which means more records or files.
=> *I/O systems should be able to store hundreds of exabytes or even zettabytes*

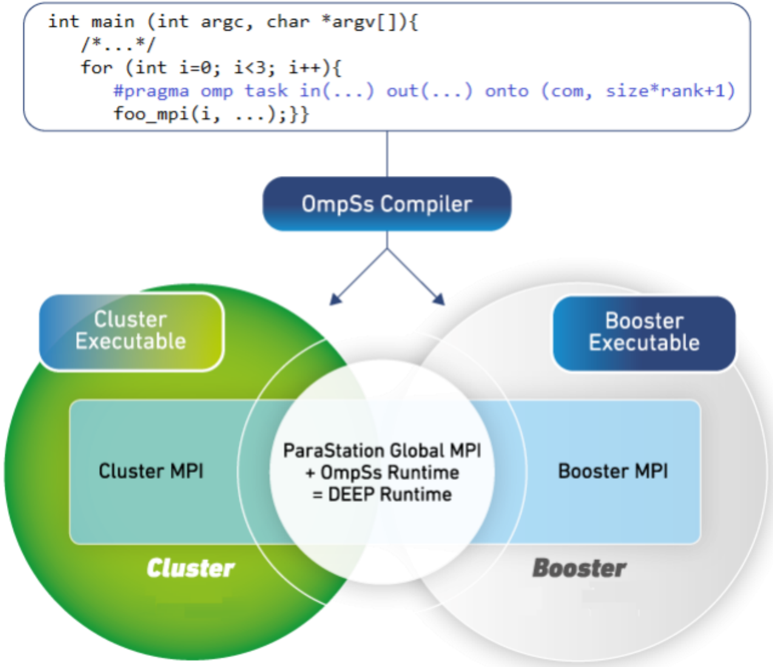
❸ **Data heterogeneity:** Small vs. large files, sequential vs. random access, access.
=> *Quite complex "data taxonomy" needs to be supported*

❹ **Data placement:** To use complex supercomputing architectures such as MSA best, data must be used and produced as close as possible to the place where the simulation code runs.

Challenges and Future Directions
Integrating Quantum Processors into the MSA

- ***Idea:*** Tightly couple with HPC systems and run every part of the code on the best suited resource
- ***Goal:*** Extend the *OmpSs programming model**
  - ✓ Already used in the MSA environment to parallelize code and map it to accelerators like GPUs and TPUs
  - ❑ Use already existing toolchains and formats for quantum computing
  - ❑ Dataflow between classical and quantum processor should be modeled with `pragmas`
  - ❑ OmpSs scheduler will be able to execute tasks on CPUs and accelerators, including QPU, asynchronously

=> *Concept integrates naturally into the already existing modular supercomputing architecture*



*Duran, Alejandro, et al. "*OmpSs: a proposal for programming heterogeneous multi-core architectures.*" Parallel processing letters 21, no. 02 (2011): 173-193.

Questions?