HPC Processing Technology in the AI Era

22.04.2024 I SALISHAN 2024 I Bernd Mohr (FZJ-JSC)

Based on input from
G.Cavallaro, E.Suarez, S.Kesselheim,
A.Lintermann (all from FZJ-JSC)

A BIG THANK YOU TO …

- **Prof. Dr.
Estela Suarez**
- Co-Lead of JSC Division Novel System Architecture Design
- Head of RG Next Generation Architectures and Prototypes

- **Dr. Stefan Kesselheim**
- Head of SDL Applied Machine Learning
- Head of AI Consultant Team

- **Dr. Andreas Lintermann**
- Head of SDL Highly Scalable Fluids & Solids Engineering
- Coordinator of CoE RAISE

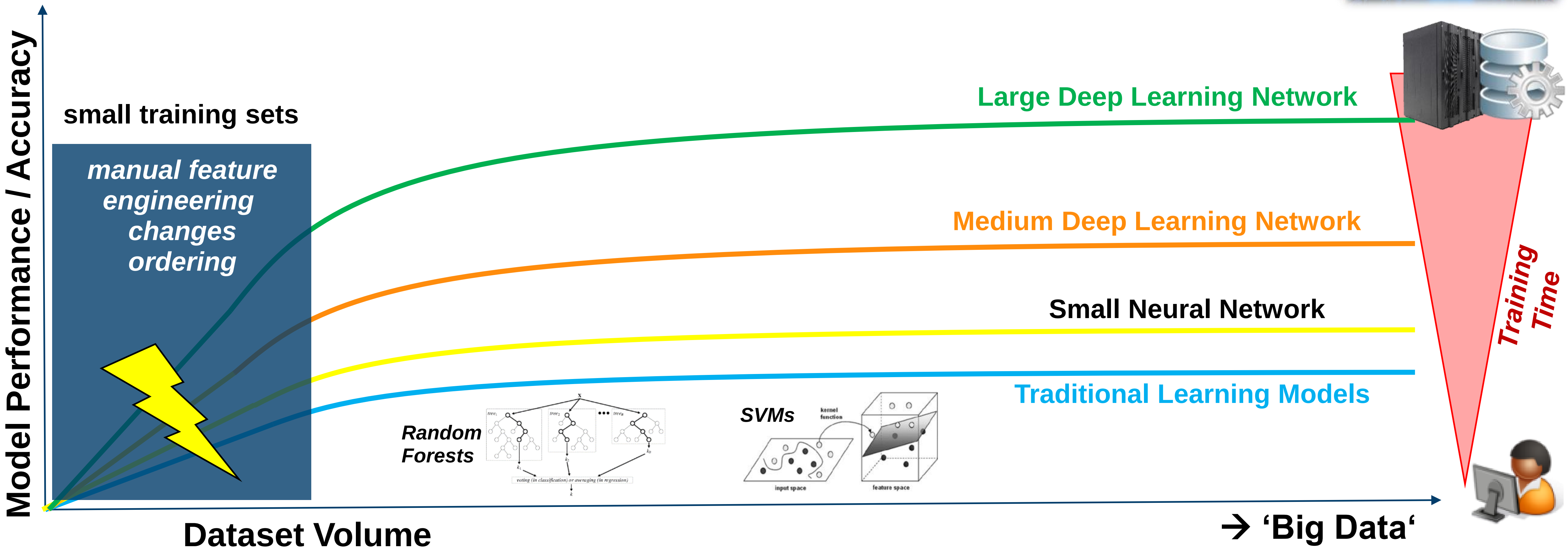
- **Prof. Dr. -Ing.
Gabriele Cavallaro**
- Head of SDL Artificial Intelligence and Machine Learning for Remote Sensing

OUTLINE

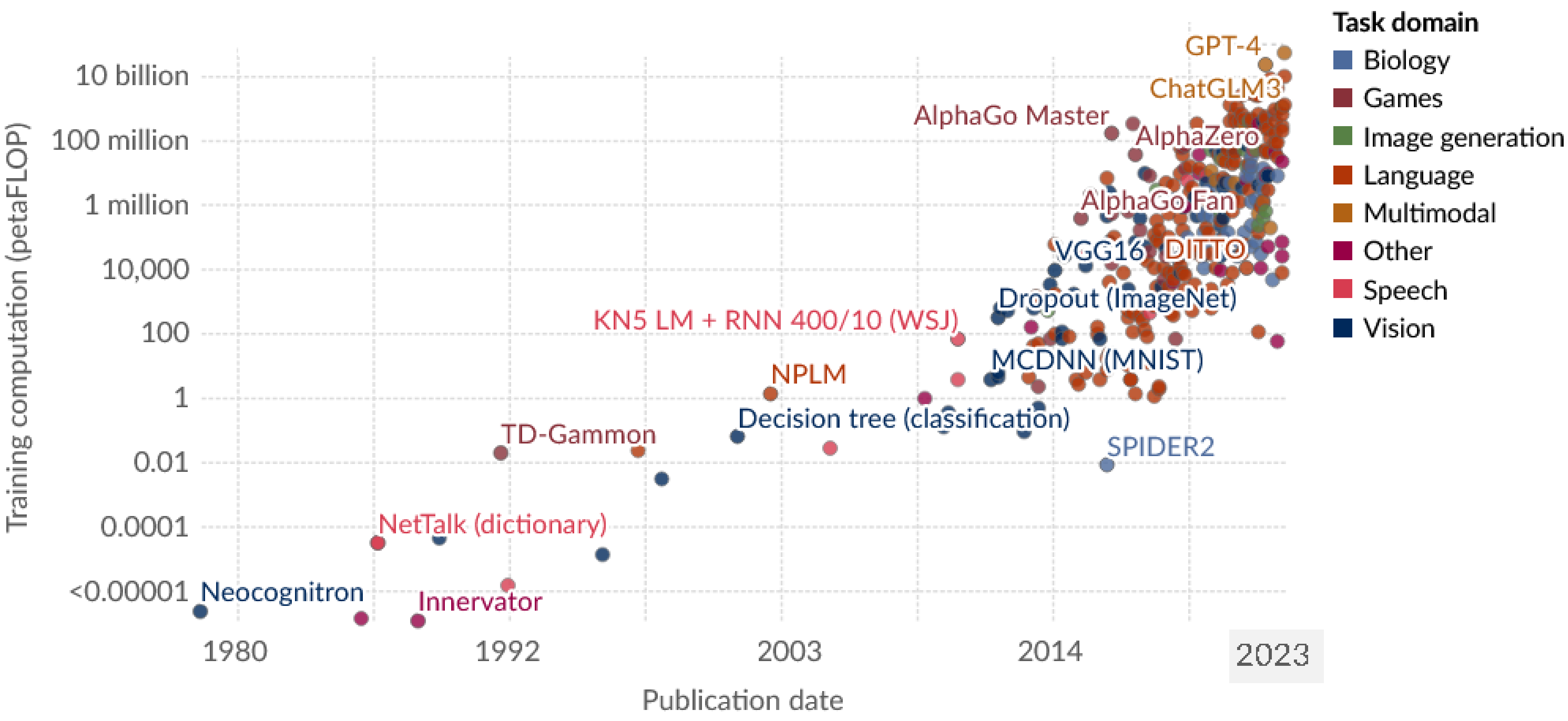
- AI in HPC
- AI computational requirements
- How to fit both AI and HPC users?
- HPC processing technology in the AI era

HPC and AI relationship

Source: Prof. Morris Riedel – Univ. Iceland & FZJ-JSC

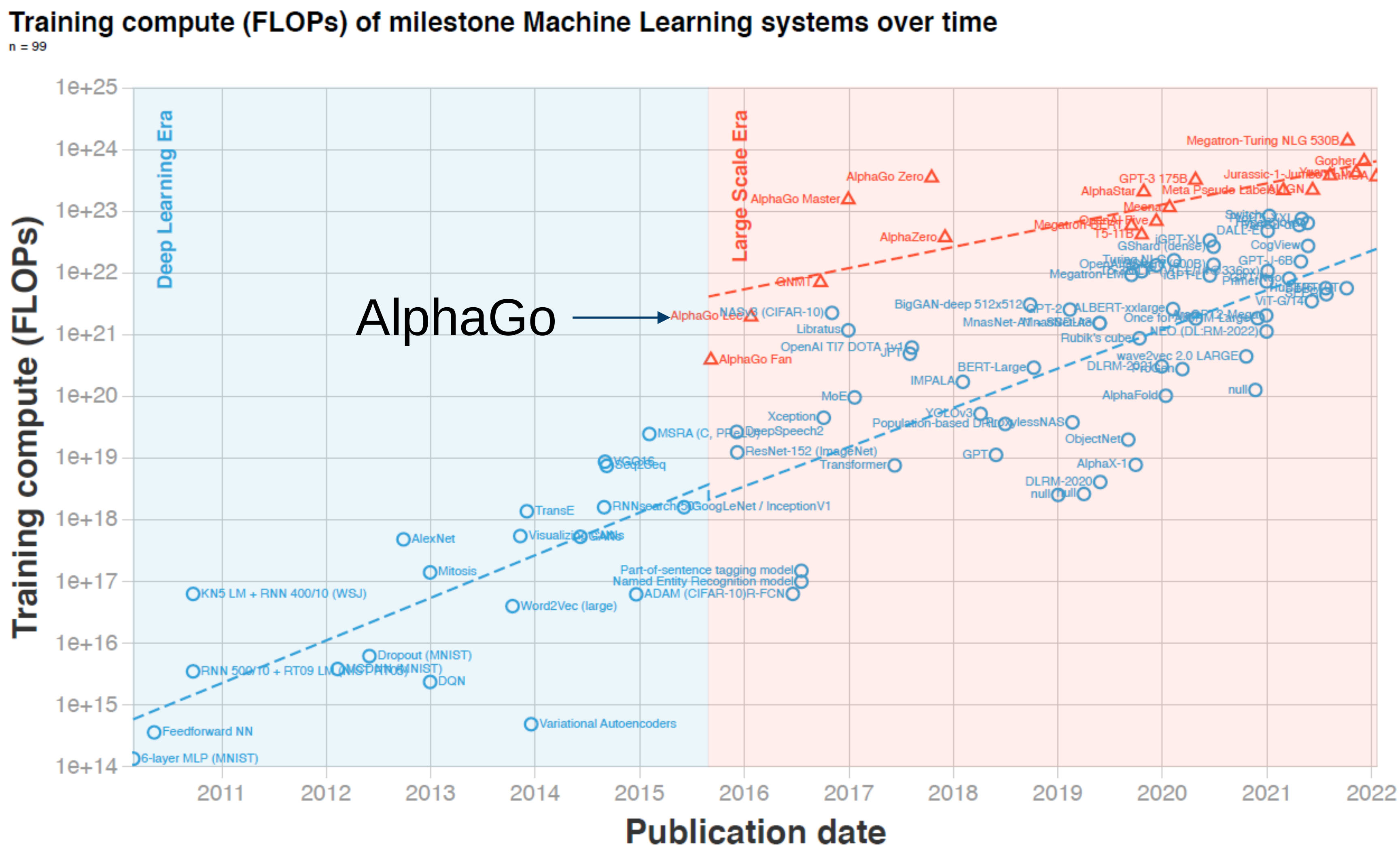


Growth in Training Computation (1980-2023)



Adapted from: Charlie Giattino, Edouard Mathieu, Veronika Samborska and Max Roser (2023), ourworldindata.org

Growth in Training Computation (2010-2022)



- **2015**: a new trend of large-scale models
- Computational capacity significantly higher (e.g., AlphaGo) than other models published in the same year
- Slower growth than the overall DL trend
  - doubling time ~8-17 months

Source: Sevilla et al. IJCNN 2022, https://doi.org/10.1109/IJCNN55064.2022.9891914

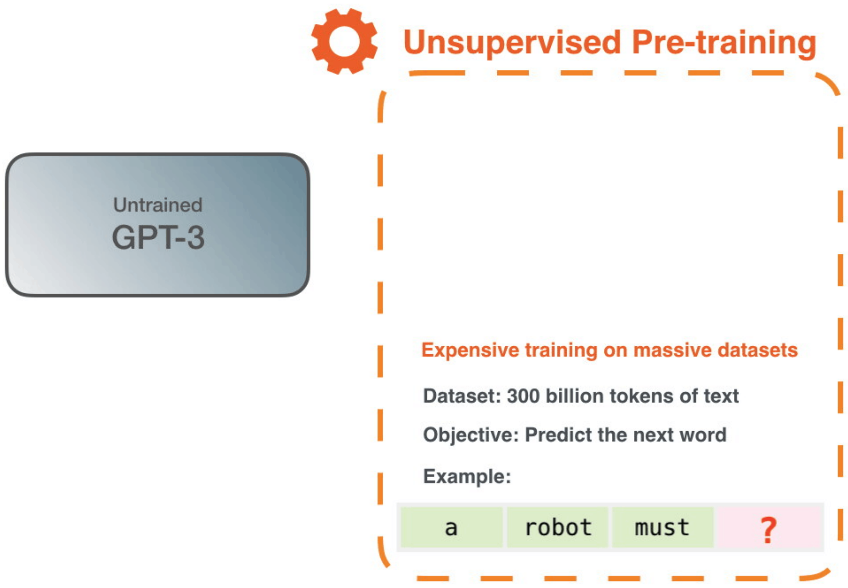
New Trends in AI-Foundation Models (FMs)

- **Strong trend towards FMs trained on extensive domain-agnostic datasets, using:**
  - unsupervised learning
  - self-supervised representation learning,
  - multimodal learning

- **Deliver more robust insights and decision-making, and bring advances in:**
  - Mainstream problems, e.g.: Natural Language Processing (NLP), Computer Vision
  - But also to many scientific fields, e.g. Earth observation [Jakubik et al, 2023].

Foundation Models

- Large deep learning models trained on a vast amount of data at scale
  - by self-supervised learning, or
  - semi-supervised learning
- They can be adapted to a wide range of downstream tasks
- Early examples of foundation models
  - pre-trained large language models,
  - e.g., GPT foundation models



Source: Jay Alammar, How GTP3 works

How to create a Foundation Model?

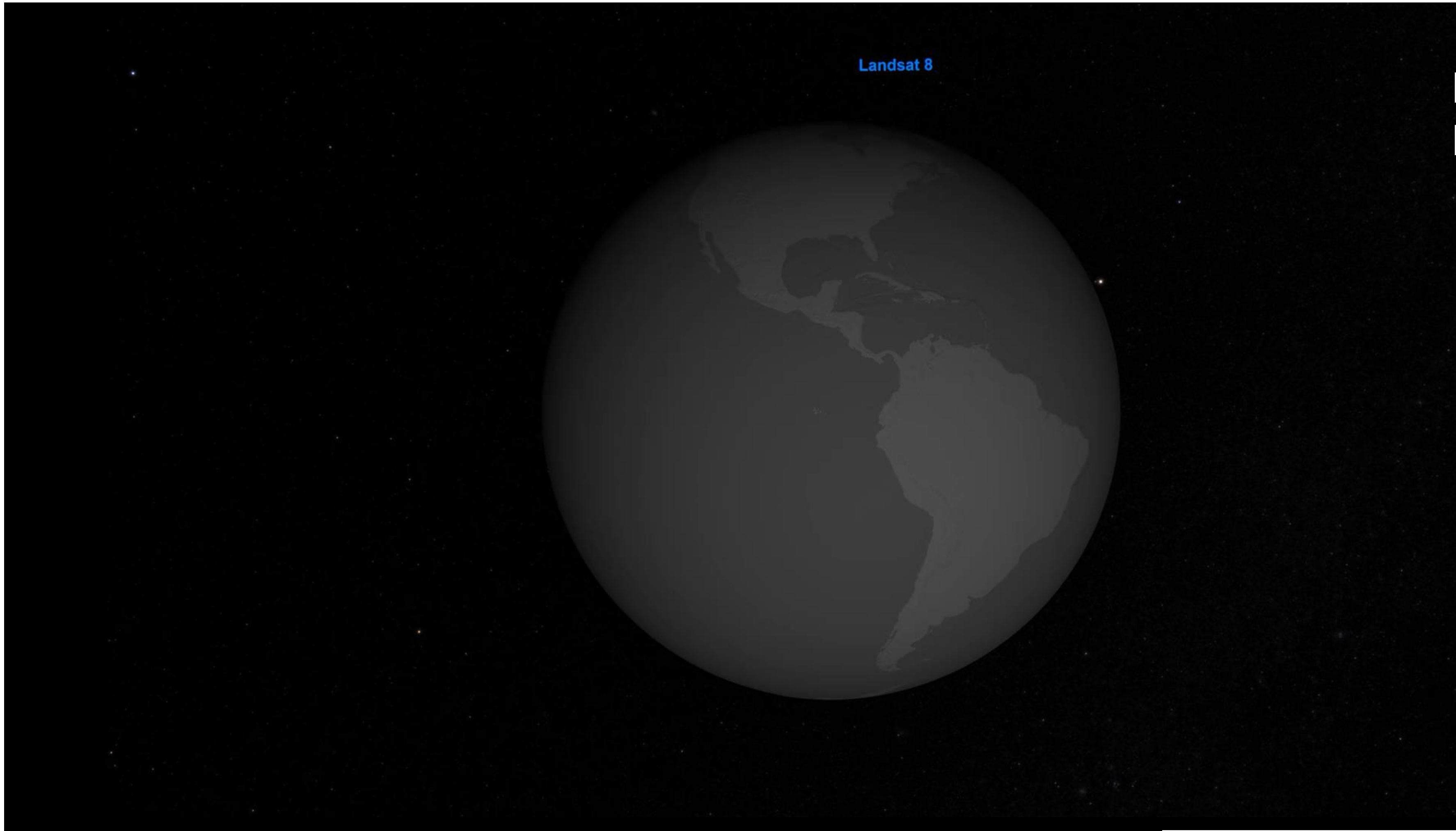
1) Gather data at scale

2) Train model once and evaluate

3) Fine-tune model for multiple downstream tasks

4) Inference (operational)

1). Gather Data at Scale

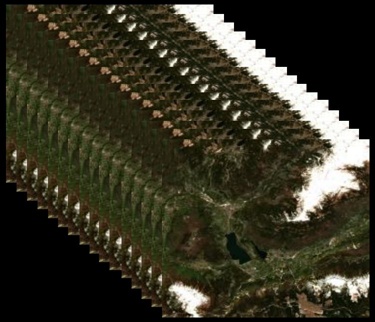


E.g., NASA's Harmonized Landsat Sentinel-2

Source: Landsat with Sentinel - Global Coverage, NASA SVS, https://svs.gsfc.nasa.gov/4745

2). Train Foundation Model once and Evaluate

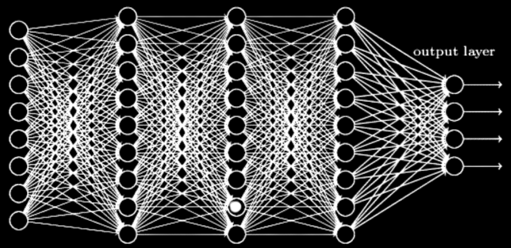
Big Data



Satellite Image Time Series

MAIN TRAINING

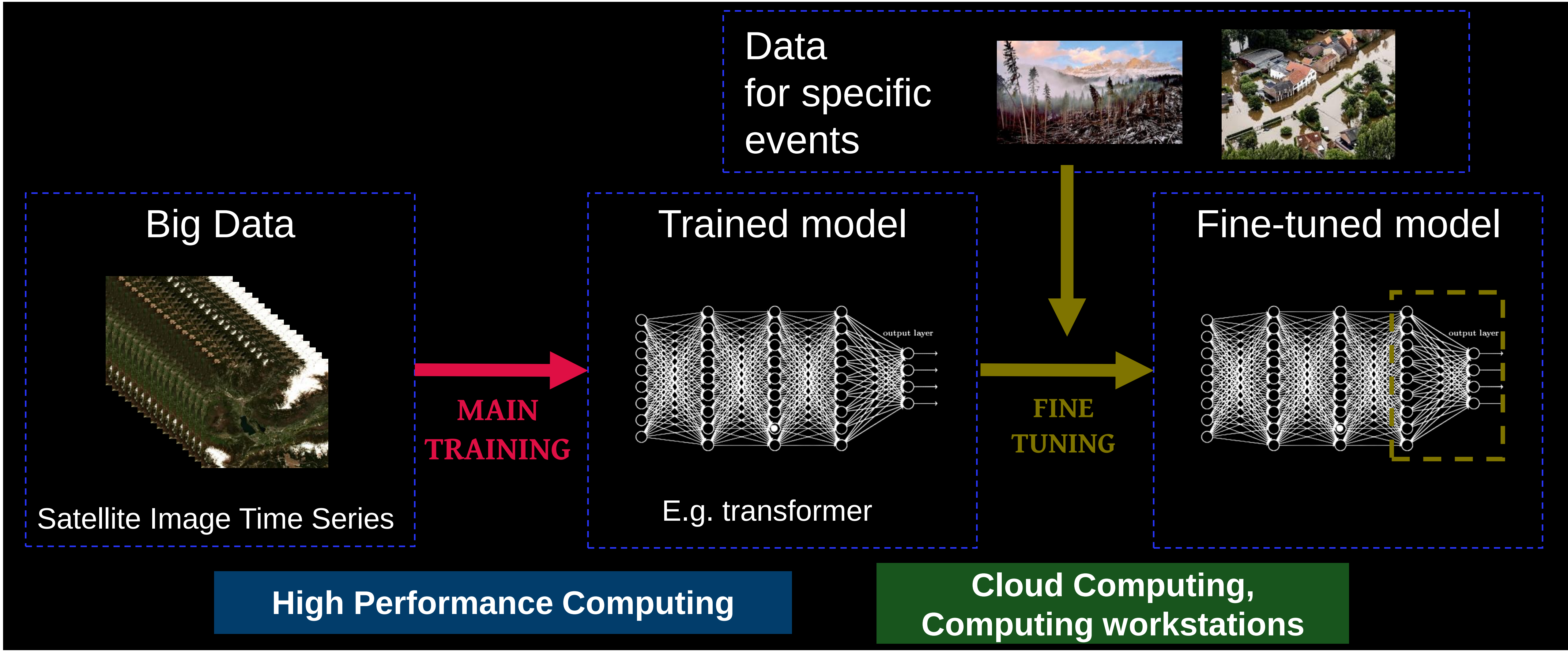
Trained model



E.g. transformer

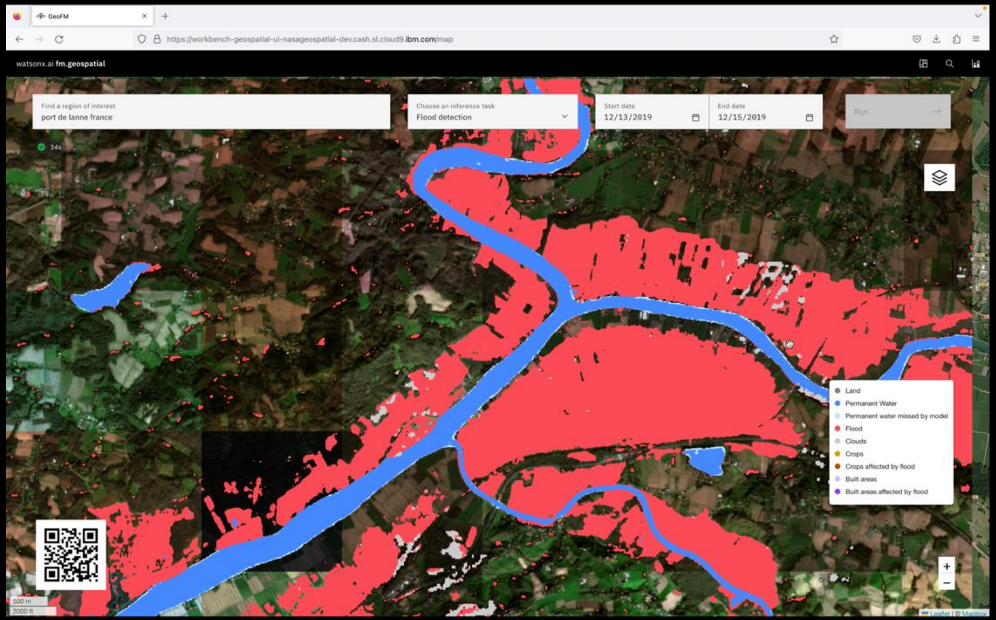
High Performance Computing

3). Fine-tune model for multiple Downstream Uses

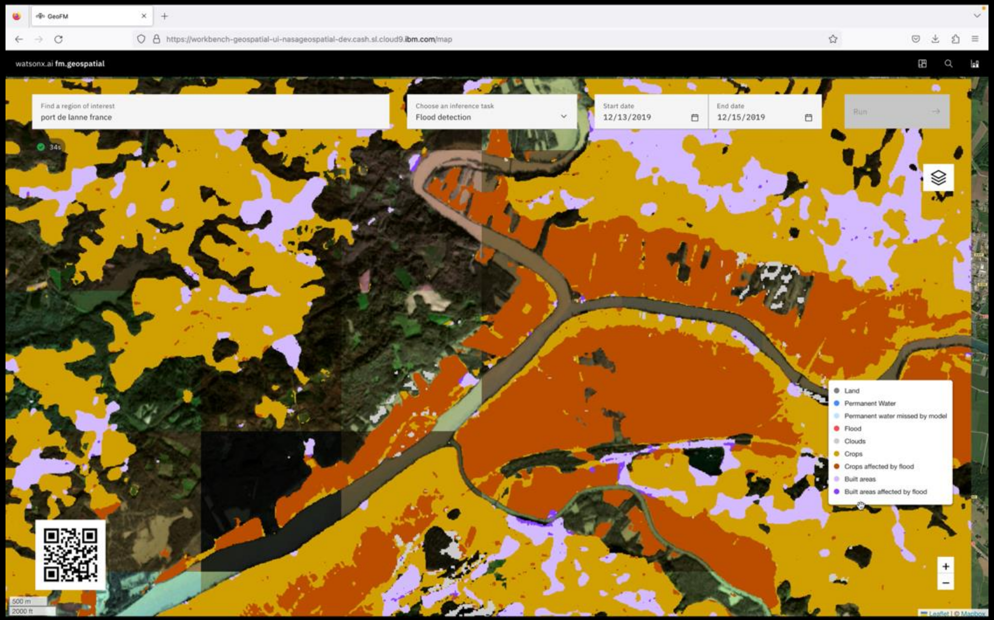


4). Inference Downstream Task: flood mapping

Flood detection



Flood impact



Maskey, et al., IEEE GRSM 2023, https://doi.org/10.1109/MGRS.2023.3302813

GPT-3: Time Required for Full Training

175 Billion weight parameters

JUWELS Booster @ Jülich

- 1× Nvidia Ampere100 ≈ 90 years
- 1× Nvidia Hopper100 ≈ 15-30 years
- 2,000× Nvidia Ampere100 ≈ 16 days
  (if scaled well on JUWELS Booster)

GPT-4: Time Required for Full Training

1,8 Trillion weight parameters

JUWELS Booster @ Jülich

- 1× Nvidia Ampere100 ≈ 1,200 years
- 1× Nvidia Hopper100 ≈ 200-600 years
- 2,000× Nvidia Ampere100 ≈ 900 days

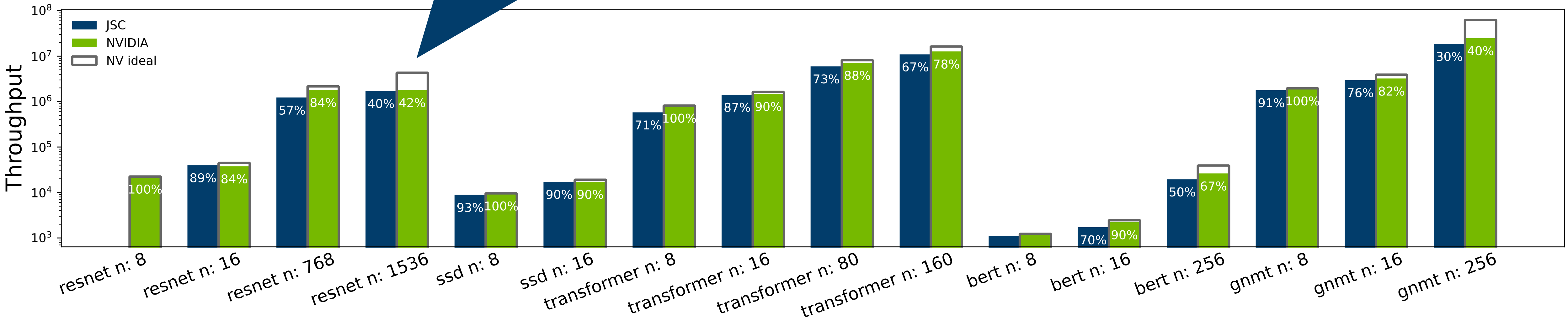
JUWELS BOOSTER

Benchmark Result

Example
- Task: Train ResNet50 on ImageNet
- GPUs: 1536
- Throughput: 1.7 Million images / sec
- Training complete after 43 seconds!
- Parallelization efficiency: 40%

- Benchmark: NVIDIA's submission to MLPerf Training v0.7
- Metric: Throughput in Samples/sec
- 5 Benchmarks on up to 1536 GPUs
- Reference: NVIDIA's results on Selene

Source: Kesselheim et al. ISC 2021



n: number of GPUs

OUTLINE

- AI in HPC
- AI computational requirements
- How to fit both AI and HPC users?
- HPC processing technology in the AI era

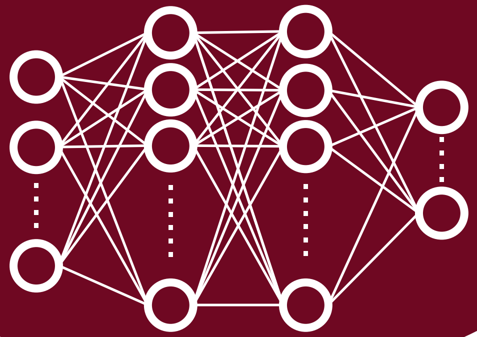
Deep Learning computational characteristics

- **Networks with 100s/1000s layers**:
  - each having numerous parameters
  - adjusted during training

- **Training models**
  - parallelisation via data parallelism
  - large-scale matrix and tensor operations
    ➔ computationally intensive
  - complexity increases exponentially with size of the model and the data
  - preferred precision Bfloat16

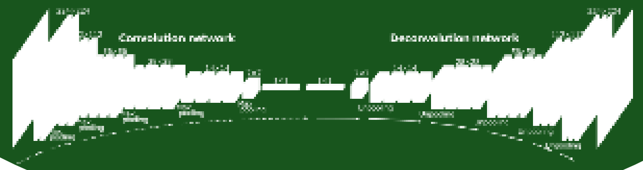
Neural Networks

Layered arrangement of differentiable units (neurons) trained by backpropagation



Deep Learning

Artificial neural networks adapt and learn from vast amounts of data



Deep Learning computational characteristics

- **Networks with 100s/1000s layers**:
  - each having numerous parameters
  - adjusted during training

- **Training models**
  - parallelisation via data parallelism
  - large-scale matrix and tensor operations ➔ computationally intensive
  - complexity increases exponentially with size of the model and the data
  - preferred precision Bfloat16

- **Accelerators can do this very well**
  - parallelism ➔ distributed training, replicate model on several GPUs
  - high memory bandwidth ➔ large data volumes
  - specialized hardware ➔ cost effective
  - reduced precision ➔ higher performance

Training Deep Learning Models Requires Accelerators

- **GPUs**: generic deep learning hardware (parallelizing matrix/tensor operations via vectorization)
- **Specialized hardware**, eg.
  - ASICs, e.g. TPUs (Google)
  - in-memory computing chips
  - Graphcore IPU: Colossus MK2,
  - Cerebras Wafer Scale Engine 2 (850k cores)

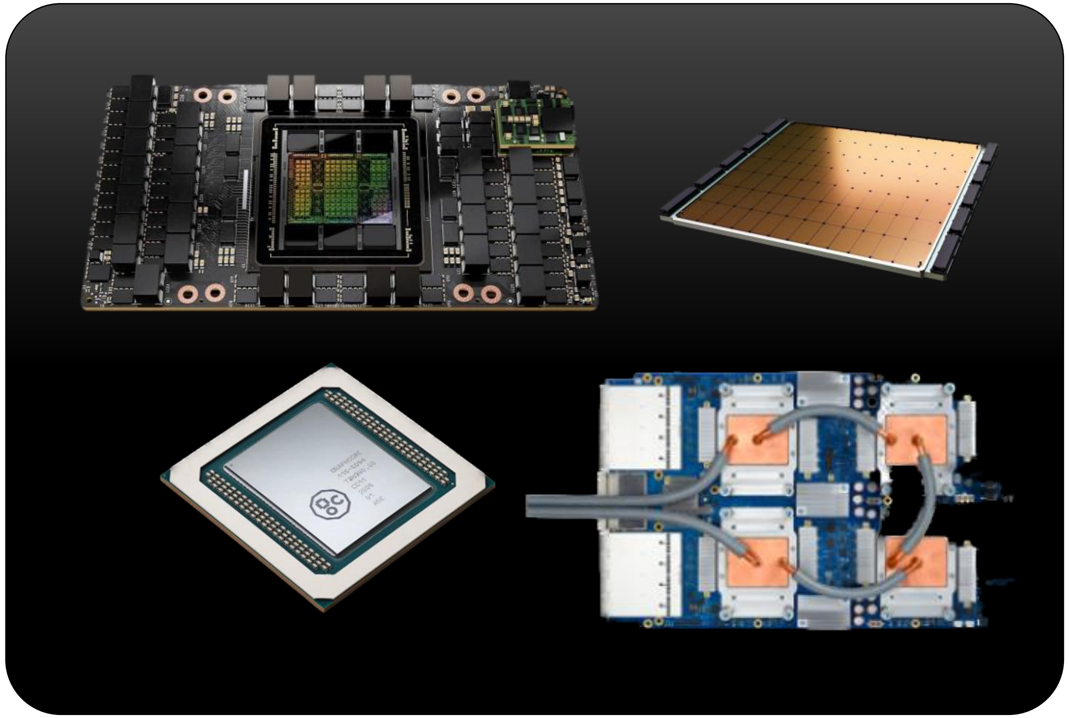
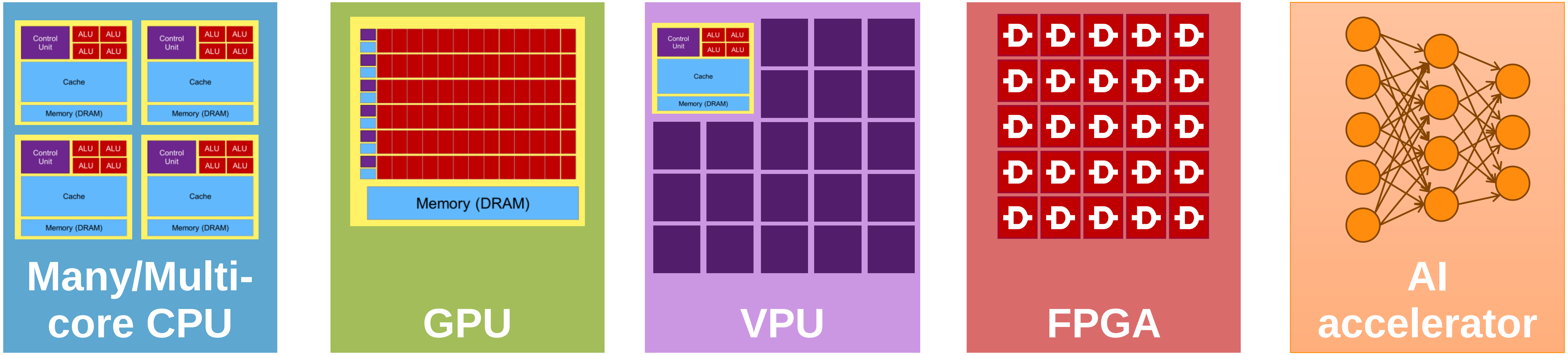


Image sources: NVIDIA, Google, Graphcore, Cerebras

Increasing Processor Diversity

Different trade-offs in the design ➔ different processing units

Accelerators



10's strong cores

1000's functional units

100's vector arithmetic units

1.000.000's programmable gates

custom ASIC implementations (e.g. TPUs)

HPC SW Stack

		Examples
Application Layer	Application	Climate & Meteorology, Drug design, QCD, Astrophysics, Protein Dynamics,…
Programming Environment	Language	C/C++, Fortran, Python, CUDA
	Parallel programming	MPI, Open MP, Open ACC, CUDA, DSL
	Libraries	Math libraries, I/O libraries, checkpointing libraries, …
Tools	Compilers	icc, gcc, llvm
	Debuggers	TotalView, Allinea DDT, PGI, GNU GDB,…
	Performance analysis tools	Score-P, Scalasca, Vampir, V-Tunes, Extrae/Paraver,…
Cluster SW	Resource Management/ Job Scheduling	SLURM, Torque/Maui, IBM LSF, PBS pro
	File system	Lustre, NFS, GPFS, BeeGFS
	Cluster Management	ParaStation, Monitoring tools, SW installation tools, Containers…
System SW	Operating system	Linux OS (RedHat, CentOS,…)
Hardware	Hardware	Server, Storage, Switch, Infrastructure

AI SW Stack

- Application Layer
- Programming Environment
- Frameworks
- Cluster SW
- System SW
- Hardware

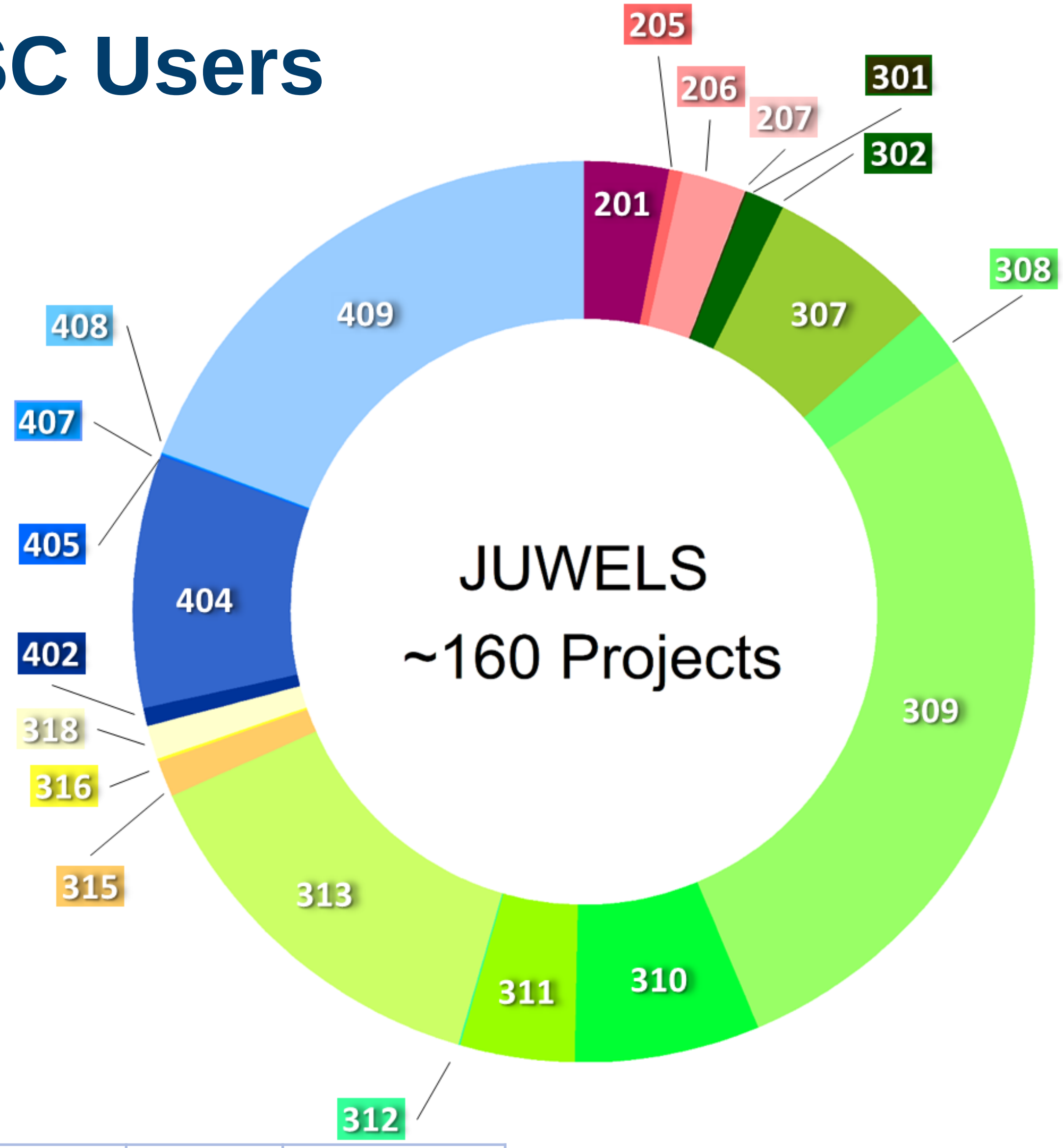
	Examples
Application	Climate & Meteorology, Drug design, QCD, Astrophysics, Protein Dynamics,…
Language	Python
AI Frameworks	PyTorch, TensorFlow, Horovod,…
	Math libraries, I/O libraries, parallel libraries…
Resource Management/ Job Scheduling	SLURM, Torque/Maui, IBM LSF, PBS pro
File system	Lustre, NFS, GPFS, BeeGFS
Cluster Management	ParaStation, Monitoring tools, SW installation tools, Containers…
Operating system	Linux OS (RedHat, CentOS,…)
Hardware	Server, Storage, Switch, Infrastructure

OUTLINE

- AI in HPC
- AI computational requirements
- How to fit both AI and HPC users?
- HPC processing technology in the AI era

JSC Users

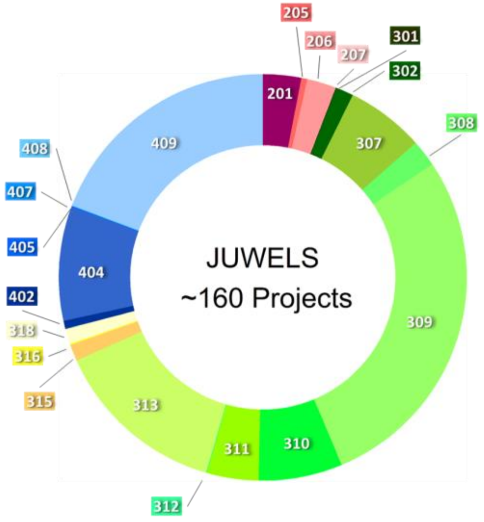
May - October 2023



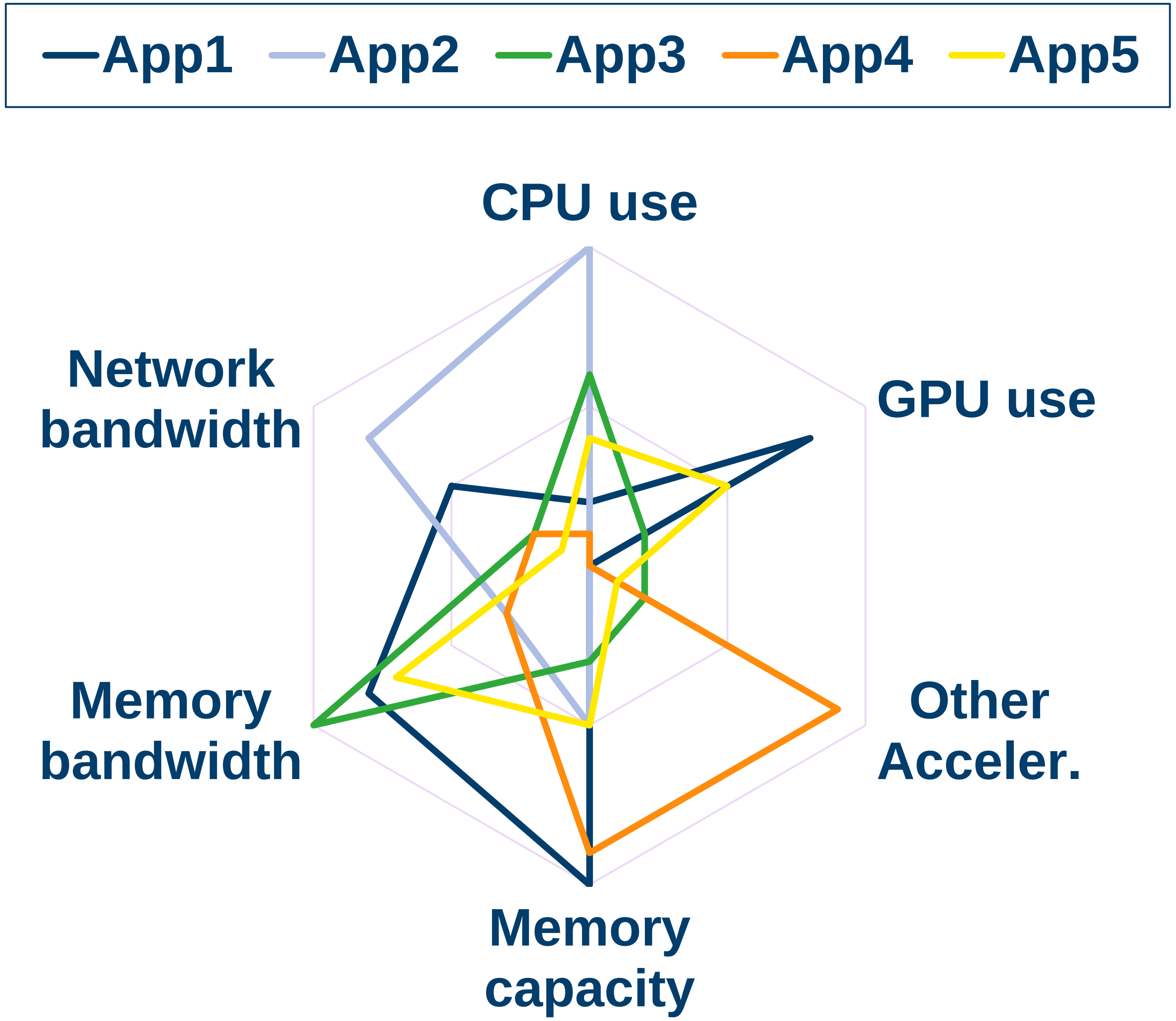
6-month	Mcoreh	EFLOP
JUWELS	1,180	2,24 M

201	Basic Biological and Medical Research
204	Microbiology, Virology and Immunology
205	Medicine
206	Neurosciences
207	Agriculture, Forestry and Veterinary Medicine
301	Molecular Chemistry
302	Chemical Solid State and Surface Research
303	Physical and Theoretical Chemistry
307	Condensed Matter Physics
308	Optics, Quantum Optics and Physics of Atoms, Molecules and Plasmas
309	Particles, Nuclei and Fields
310	Statistical Physics, Soft Matter, Biological Physics, Nonlinear Dynamics
311	Astrophysics and Astronomy
312	Mathematics
313	Atmospheric Science, Oceanography and Climate Research
315	Geophysics and Geodesy
316	Geochemistry, Mineralogy and Crystallography
318	Water Research
402	Mechanics and Constructive Mechanical Engineering
403	Process Engineering, Technical Chemistry
404	Heat Energy Technology, Thermal Machines, Fluid Mechanics
405	Materials Engineering
406	Materials Science
407	Systems Engineering
408	Electrical Engineering and Information Technology
409	Computer Science

How to serve diverse requirements with one single system?



Diverse Requirements



Node design

?

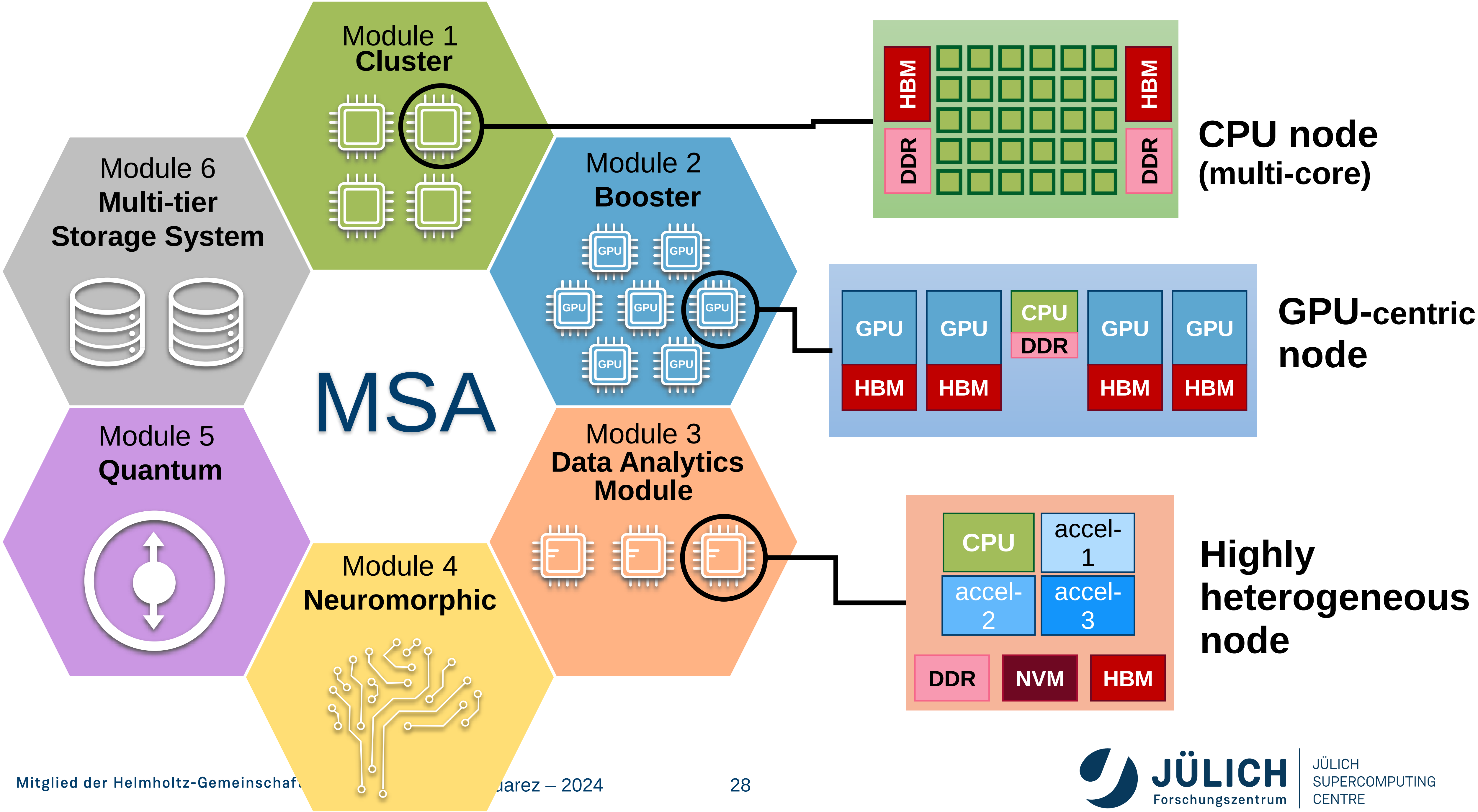
BACKGROUND

2011-2021: The DEEP projects

- **DEEP** (2011 – 2015)
  - Introduced Cluster-Booster architecture
- **DEEP-ER** (2013 – 2017)
  - Added I/O and resiliency functionalities
- **DEEP-EST** (2017 – 2021)
  - **Modular Supercomputer Architecture**

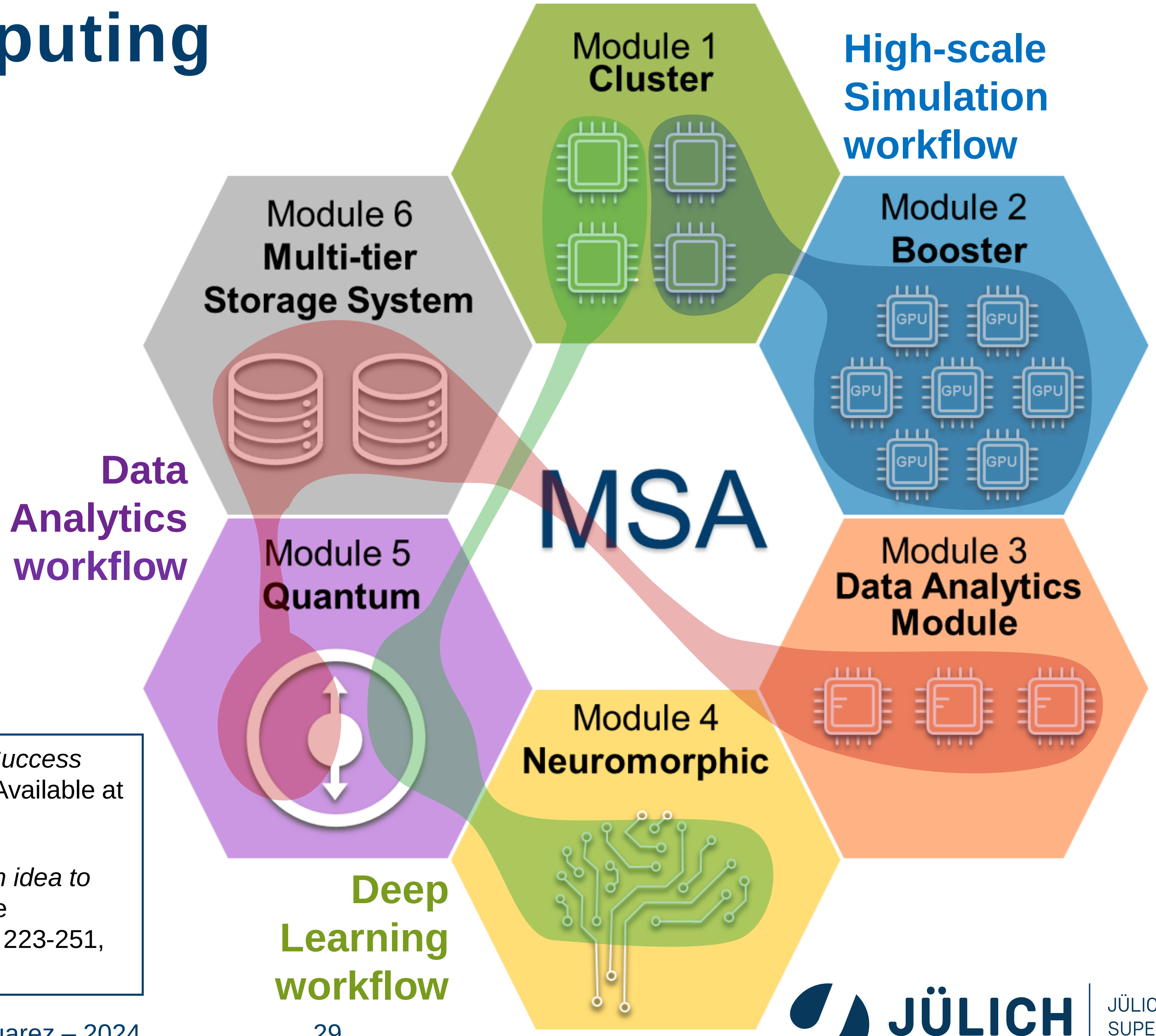
2021-2024: The SEA projects

- **DEEP-SEA**
  - **Software for Exascale Architectures**
- Also: **IO-SEA, RED-SEA**


Module 1
Cluster
Module 2
Booster
GPU
Module 3
Data Analytics Module
Module 4
Neuromorphic
Module 5
Quantum
Module 6
Multi-tier Storage System
MSA
HBM
DDR
CPU node (multi-core)
GPU
CPU
DDR
HBM
GPU-centric node
CPU
accel-1
accel-2
accel-3
DDR
NVM
HBM
Highly heterogeneous node

Modular Supercomputing Architecture

Serve HPC and AI applications with composable heterogeneous resources



- **Suarez** et al. *"Modular Supercomputing Architecture – A Success Story of European R&D",* ETP4HPC White Paper. (2022) Available at https://www.etp4hpc.eu/white-papers.html#msa.
- **Suarez** et al., "*Modular Supercomputing Architecture: from idea to production*", Chapter 9 in Contemporary High Performance Computing: from Petascale toward Exascale, Volume 3, p 223-251, CRC Press. (2019)

Modular Supercomputer JUWELS

Entry in Nov'20 TOP500 The List.

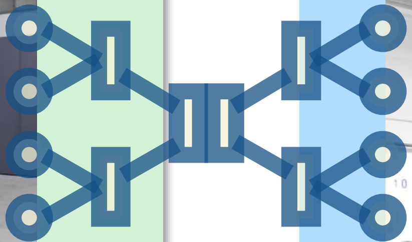
JUWELS Cluster #44

Intel Xeon (Skylake) processor

InfiniBand EDR network

2,500 compute nodes

10 PFLOP/s peak (CPU-based)



JUWELS Booster #7

AMD EPYC Rome 7402 processor

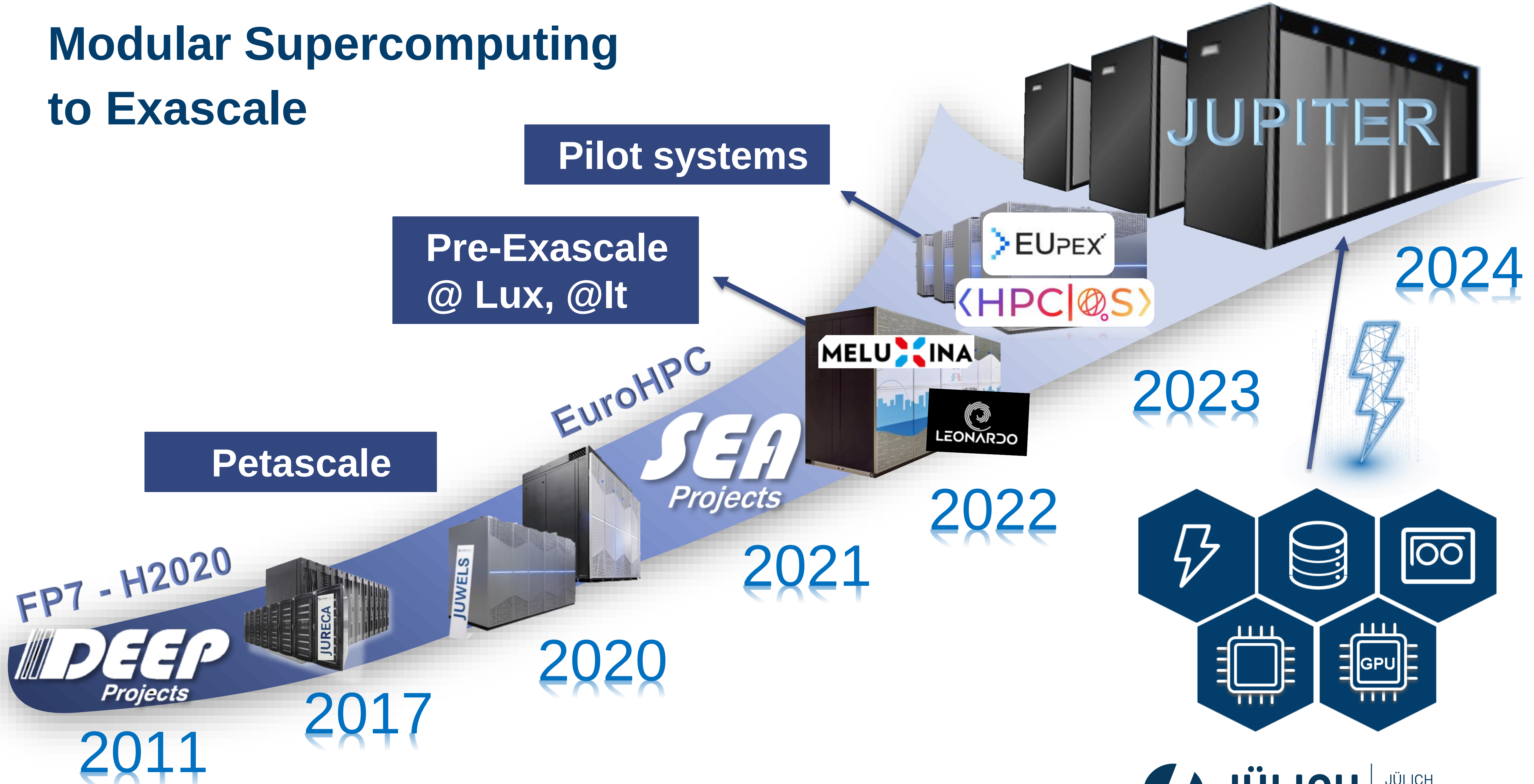
3,700 NVIDIA A100 GPUs

InfiniBand HDR DragonFly+

70 PFLOP/s peak (GPU-based)

Funded through SiVeGCS (BMBF, MWK-NRW)

Modular Supercomputing to Exascale



JUPITER Modular Heterogeneous Architecture

Universal Cluster

>5 PetaFLOP/s (FP64, HPL)

SiPearl Rhea1 (ARM Neoverse Zeus)

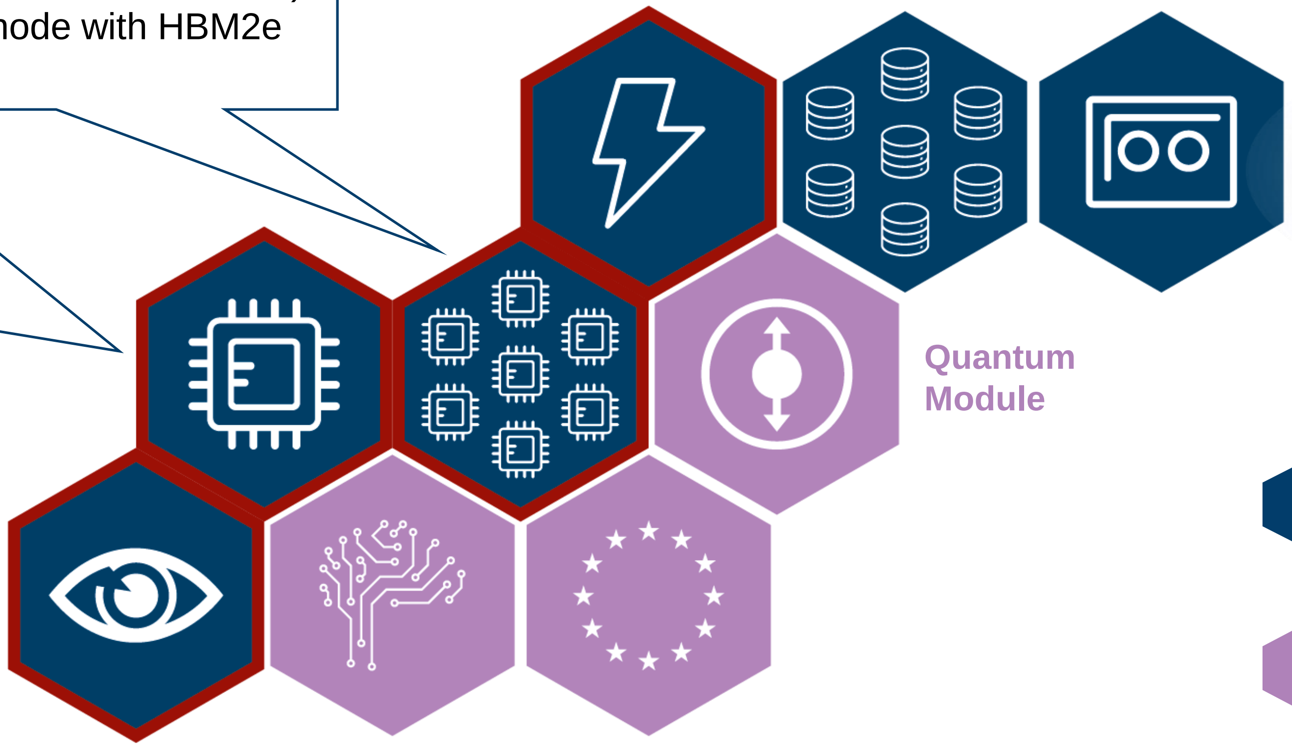
- 2 x CPUs per node with HBM2e
- >1,300 nodes

GPU Booster

1 ExaFLOP/s (FP64, HPL)

NVIDIA Grace-Hopper

- 4 x chips per compute node
- 72 cores per Grace CPU
- Hopper H100 GPU with HBM3
- ~6,000 nodes



Core Configuration

Future Modules

EuroHPC Joint Undertaking

Bundesministerium für Bildung und Forschung

Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

OUTLINE

- AI in HPC
- AI computational requirements
- How to fit both AI and HPC users?
- HPC processing technology in the AI era

Role of Supercomputers

- Big tech companies deploy their AI supercomputers
- Supercomputing now goes far beyond traditional scientific computing, which was driven by large governments
- Major industries building highly specialized supercomputers are taking the lead

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

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

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

RESEARCH

Introducing the AI Research SuperCluster — Meta's cutting-edge AI supercomputer for AI research

January 24, 2022

Sources: Jouppi et al. 2023, Forbes, Fabebook, TheSun, Hoefler@ETHZ

SUPERCOMPUTING EVOLUTION

Architecture paradigms

- **1940 – 1950**: first computers are Supercomputers (specialized, expensive)
- **1960 – 1980**: vector computers dominate HPC,
  while general purpose computers come to market at much lower prices
  - Focus: floating operations (linear Algebra)
  - **Special purpose technologies** (fast vector processors, parallel architectures)
  - Only few machines produced → **expensive!**
- **1990 – 2000**: cluster computers are born
  - Integrate general purpose CPUs in HPC → **more economic approach**
  - Many „computers“ connected through fast network
  - Distributed memory ➔ MPI
- **2010 – 2020**: heterogeneous cluster systems
  - CPU + Accelerator technologies (mostly GPUs) → **more FLOPS/Watt**
  - Intel / AMD + NVIDIA / AMD / Intel
- **2020 – today**: very large GPU-based systems in HPC,
  while hyperscalers dominate AI-market, drive GPU design (and price), and build their own processors for their clouds

Which are the options for HPC?

A) **Build own technology**, e.g. Fugaku

Source: RIKEN

- ***Challenges:***
  - Multi-year, multi-million investment, not possible for every HPC site
    - Will chiplet designs help?
    - Will free-licence ISAs (RISC-V) help?

or

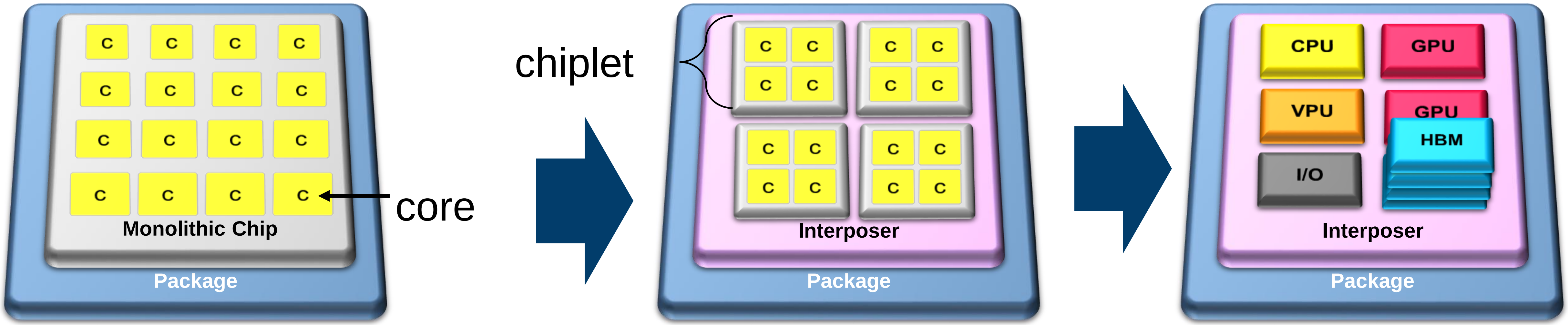
B) Adapt to AI market

- Use same hardware
- Ok for AI-training workloads
- HPC workloads must be ported

- ***Challenges:***
  - Reduced precision
    - can legacy HPC codes adapt?
  - HPC is needed for training:
    - what when inference grows over training?
  - Hardware accessibility:
    - will hyperscalers sell their processors?

A) Build Own Technology: using chiplets

Towards chiplet-based designs



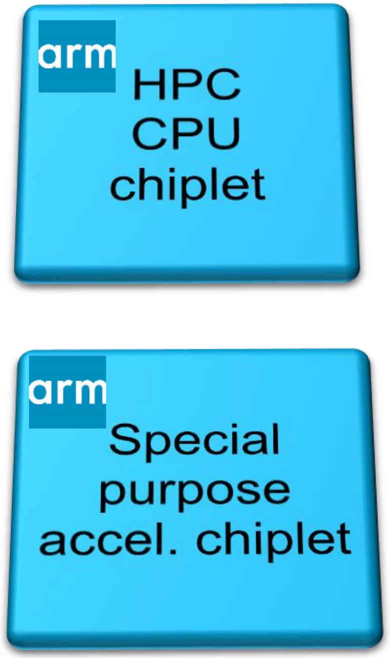
Monolithic Die

Multiple Dies – 2D/2.5D
(homogeneous chip)

Multiple Dies – 3D
(heterogeneous chip)

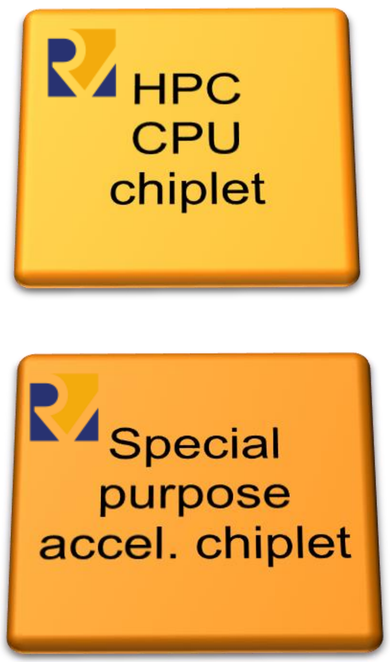
A) Build Own Technology: with licensable ISA

- **Rely on well supported software environment**
- **Several companies with HPC products already**
  - Fujitsu: A64Fx
  - NVIDIA: Grace
  - Amazon: Graviton
- **+: Good software basis, low portability effort**
- **- : High licence costs**



A) Build Own Technology: with open ISA

- **Open Instruction Set Architecture** (ISA) – BSD licence
- **Rich development community in industry and academia**
  - Open standard enables new players bring new ideas
  - Used mainly in low-power embedded market
  - But also HPC designs are in development (e.g. supporting scalable vector formats (RVV; similar to Arm SVE)
- **+: Lowers entry barrier (cost) for new developments**
- **- : Lacks software support, danger of ‘proliferation’**



B) Adapt to the trends in AI market

- **Computer Technology and Architecture perspective**
  - Integrate new developments into HPC environment (e.g. with MSA)
  - Enter in co-design with new players in chip development
    - *Both new start-ups and hyperscalers*
- **Application and Programming environment perspective**
  - Further develop and rely on portable programming models (e.g. Kokkos)
  - Heavily invest on software engineering for applications
    - Maybe Foundation Models can help in code porting
  - Develop/Adapt algorithms for mixed/lower precision

Which are the options for HPC?

A) Build own technology and **B) Adapt to AI market**