Supercomputing for AI

Highest Compute Performance for Sovereign AI in the European Union

Eric Schnepf – Senior HPC Advisor for the ISC Group

Agenda

- Supercomputing - Introduction
- Supercomputing for AI
- Gigawatt AI Datacenters
- AI@SC25 – AI@ISC26
- Recent Initiatives by the European Union
  - The EuroHPC Supercomputing Ecosystem
  - EuroHPC quantum computers
  - AI Factories
  - AI Gigafactories

Supercomputing - Introduction

How does Supercomputing and HPC relate with AI?

Supercomputer – computer with highest possible compute performance at a given time (measured in FLoating-point Operations Per Second – FLOPS)

Supercomputing - Introduction

How does Supercomputing and HPC relate with AI?

Supercomputer – computer with highest possible compute performance at a given time (measured in FLoating-point Operations Per Second – FLOPS)

High Performance Computing – hardware and software technologies to achieve highest possible compute performance for various applications (vectorization, parallelization, accelerators, high-speed interconnect…)

Supercomputing - Introduction

How does Supercomputing and HPC relate with AI?

Supercomputer – computer with highest possible compute performance at a given time (measured in FLoating-point Operations Per Second – FLOPS)

High Performance Computing – hardware and software technologies to achieve highest possible compute performance for various applications (vectorization, parallelization, accelerators, high-speed interconnect…)

Artificial Intelligence – application class (methods, algorithms, models, tools, workflows) that needs highest performance for compute, data transfer and I/O

>>> HPC is the engine for AI

Supercomputing - Introduction

Source:
https://en.wikipedia.org/wiki/Cray-2

- Selected milestones since the 70th
  - 1976 - First Supercomputer **CRAY-1**
    -> **Vector** computer with **160 MFLOPS** Peak
  - 1986 – **CRAY-2** with **2 GFLOPS** Peak, ~150-200 kW
    -> Liquid Cooling – Flourinert-cooling „waterfall“
  - 1993 – Massive parallel computer (**MPP**) Thinking Machines Corporation **CM-5**
    -> No.1 in the first TOP500 edition with **59.7 GFLOPS** HPL
  - 1997 – **ASCI Red** – MPP system based on 7,256 Intel Pentium Pro CPUs (**1.068 TFLOPS**)
    -> No. 1 in TOP500 from 1997/06-2000/06 (incl. upgrades)
  - 2008 – **Roadrunner** – MPP system based on IBM QS22 blades using the PowerXCell processor of Sony PlayStation 3 (**1.026 PFLOPS**)
    -> No. 1 in TOP500 from 2008/06-2009/06
  - 2010 – **Tianhe-1A** – **GPU-accelerated** MPP system based on 14,336 Intel Xeon + 7,168 NVIDIA Tesla GPUs (**2.57 PFLOPS**)
    -> No.1 in TOP500 in 2010/11
  - 2020 – **Fugaku** – **Arm-based** MPP system (CPU-only)
    -> No.1 in TOP500 from 2020/06-2021/11 (415.5 - **440 PFLOPS**)
  - 2026 – **Lineshine** – CPU-only architecture
    -> ~13 Mio CPU cores, ~42 MW
    -> No.1 in TOP500 with **2.19 EFLOPS** HPL

Source: Presentation by Yutong Lu at ISC 2026

An early computer dedicated for AI (1994)

Neurocomputer Synapse-1 developed by Siemens and Siemens Nixdorf

- Neuro signal processor MA16 for training neural networks
- 610,000 transistors, 800 M ops/s
- Synapse-1 with 8xMA16 (systolic array)
- 8000x faster than Sparc-2 workstation
- Advantage disappeared in the coming years !

Innovationspreis '95

„The Winner is… Siemens Nixdorf"

Das Urteil über den neuen Computer klingt scheinbar vernichtend. „Konventionelles Rechnen", so Dr. Ulrich Ramacher, Professor an der Technischen Universität Dresden, „ist nicht seine Stärke." Doch der vermeintliche Mangel ist gewollt. Die Maschine, die der Physiker – noch bei Siemens – federführend entwickelt hat, zeigt andere Qualitäten. Synapse-1 ist ein Neurocomputer, der ähnlich arbeitet wie das menschliche Gehirn und darauf spezialisiert ist, Muster zu erkennen. Außerdem ist er lernfähig und „Europas erster seiner Art" – ein Hinweis, den Dr. Jürgen Pampus nie vergißt. Dr. Pampus ist für den Verkauf des 350 000 DM teuren Systems zuständig, das von der SNI-Tochter EFG in Dresden gefertigt wird.

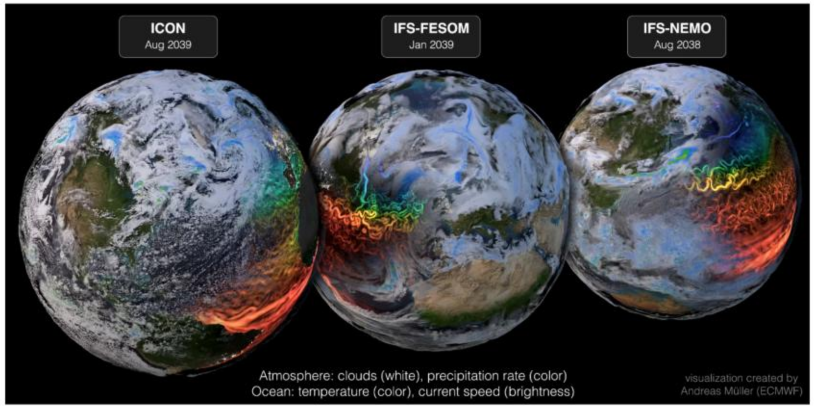
Das Forschungs- und Vertriebsteam mit „ihrem" Neurocomputer – dem schnellsten seiner Art.

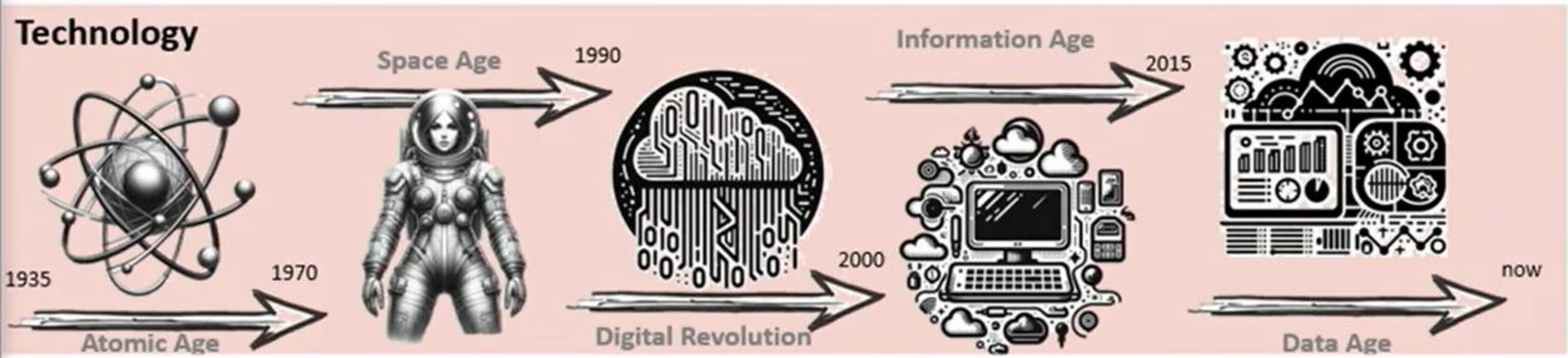
Synapse-1

Supercomputing - Introduction

- Initial applications with extreme need for compute power
  - Technical / scientific simulations (weather, crash tests, fluid dynamics…)
  - So-called „Traditional HPC“, “Mod/Sim”
- Todays applications with highest need for compute power
  - Training neural networks for leading AI models (LLMs, GenAI)
  - AI for Science
  - AI-enhanced simulation applications
  - AIFS – ECMWF’s machine-learned weather forecast model
  - ICON – earth system model

Source: Destination Earth news release on May 18



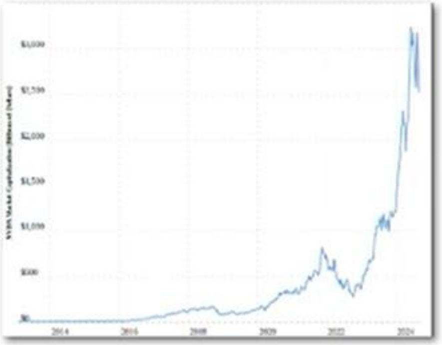


We are now entering the

Age of Computation

NVIDIA

compute accelerators + networks not data storage!



You reposted
Sam Altman
@sama
openai now generates about 100 billion words per day.

Researchers warn we could run out of data to train AI by 2026. What then?

ICML'24
Will we run out of data? Limits of LLM scaling based on human-generated data

FORBES > INNOVATION
The Pros And Cons Of Using Synthetic Data For Training AI

OUTLOOK | 27 April 2023
Synthetic data could be better than real data
Machine-generated data sets have the potential to improve privacy and representation in artificial intelligence, if researchers can find the right balance between accuracy and fakery.

The Data Age is ending

NEWS | 30 August 2024 nature
Researchers built an 'AI Scientist' – what can it do?
The large language model does everything from reading the literature to writing and reviewing its own papers, but it has limited applicability.

Graph of Thoughts

- invoke LLMs iteratively to solve complex tasks -

AI needs highest compute performance

- Looking at Large Language Models (LLM, Transformer)
  - >95% of computing done in matrix multiplication kernel
    - Great fit for GPUs
    - Easy to parallelize
  - Strong increase of numbers of parameters (weights) -> billions
  - Even stronger increase in demand for compute performance
- -> Supercomputers fuel Modern AI
  (Amazon, Meta, Tesla, Google, Microsoft, Oracle)
- AI Factories are reaching Gigawatt scale

- <u>Recommendation</u>
- Torsten Hoefler‘s (CSCS, ETH Zürich) presentations about AI and the age of computation
- YouTube Channel:
  https://www.youtube.com/playlist?list=PLerS07GSe_l7cL7kkaGuY5cLgyDg8wVlw

Supercomputing for AI - Requirements

- AI Training needs extremely high compute performance
  - AI model is trained with huge amount of data
  - Iterative process to optimize the weight parameters in the neural net
  - Matrix multiplications (tensor operations) dominate the training phase
  - Reduced precision is sufficient
- AI Inferencing needs very short latency (response time)
  - Example object recognition: data are received at edge device (e.g. camera, sensor)
  - Inference by processing this input by the trained AI model
  - Perform inference as close as possible to the edge device
- Use Case / Number of parameters in AI model define the requirements
  - Compute performance
  - Memory
  - Interconnect (bandwidth / latency)

Supercomputing for AI

- Typical components of AI factories with AI-optimized Supercomputers
  - Data center with warm water liquid cooling and re-use of waste heat
  - Modular architecture for different workloads (training, inferencing, reasoning)
  - >10.000 compute nodes with 1-2 CPUs plus 4-8 GPUs per node
  - High-speed interconnect (Infiniband, Ultra Ethernet)
  - High-performance data storage
  - 100 MW – 1 GW power consumption

- Comparison with JUPITER:
  - Warm water liquid cooling
  - Modular architecture
  - “Booster module”:
    ~6000 compute nodes with 4 NVIDIA GH200 CPU/GPUs per node
  - High-speed interconnect (Infiniband NDR)
  - High-performance data storage
  - 1 EFLOPS HPL at 18 MW

Supercomputing for AI vs. Simulation

- What‘s common?
  - Cluster Architecture using many GPU-based compute nodes with high-speed interconnect
  - Extreme Scaling and integration density
  - Challenges regarding power supply and cooling infrastructure

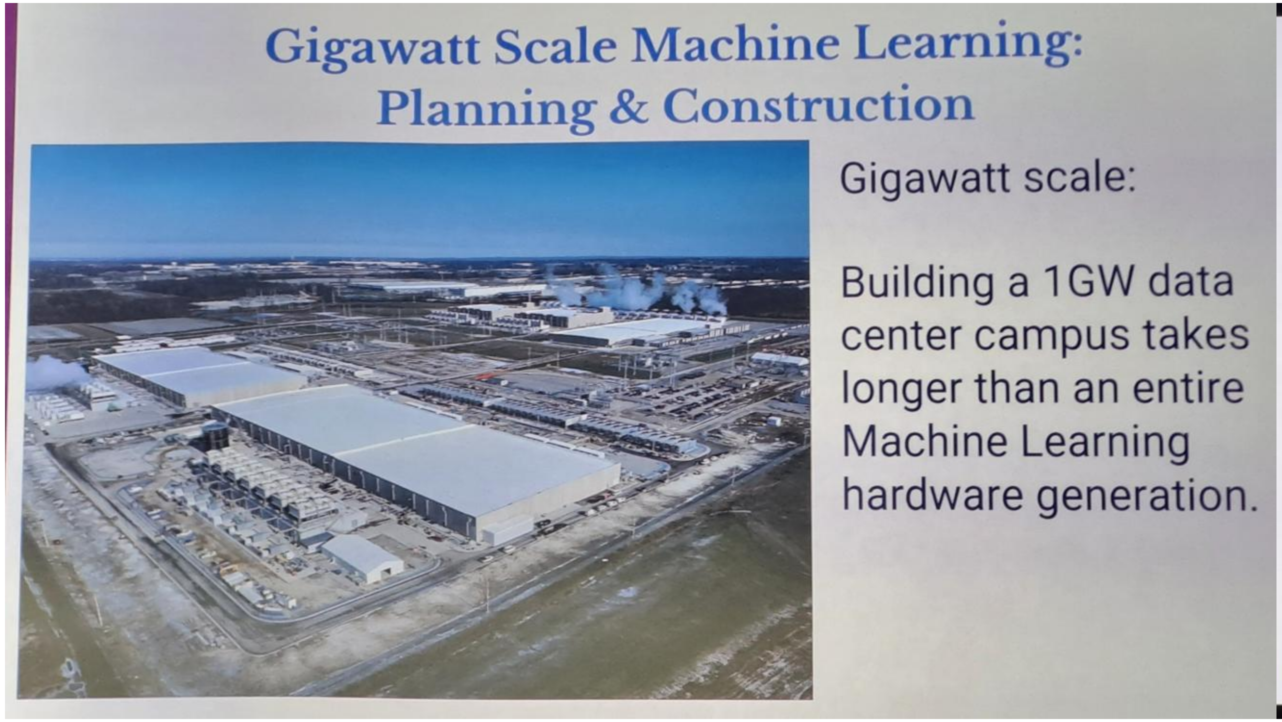
- What‘s different?
  - AI needs less compute precision – but higher number of compute operations
  - New accelerators focus on AI requirements (INT4, FP8, FP16….)
  - Technical/scientific simulation applications continue to need 64bit precision
    -> mixed precision !

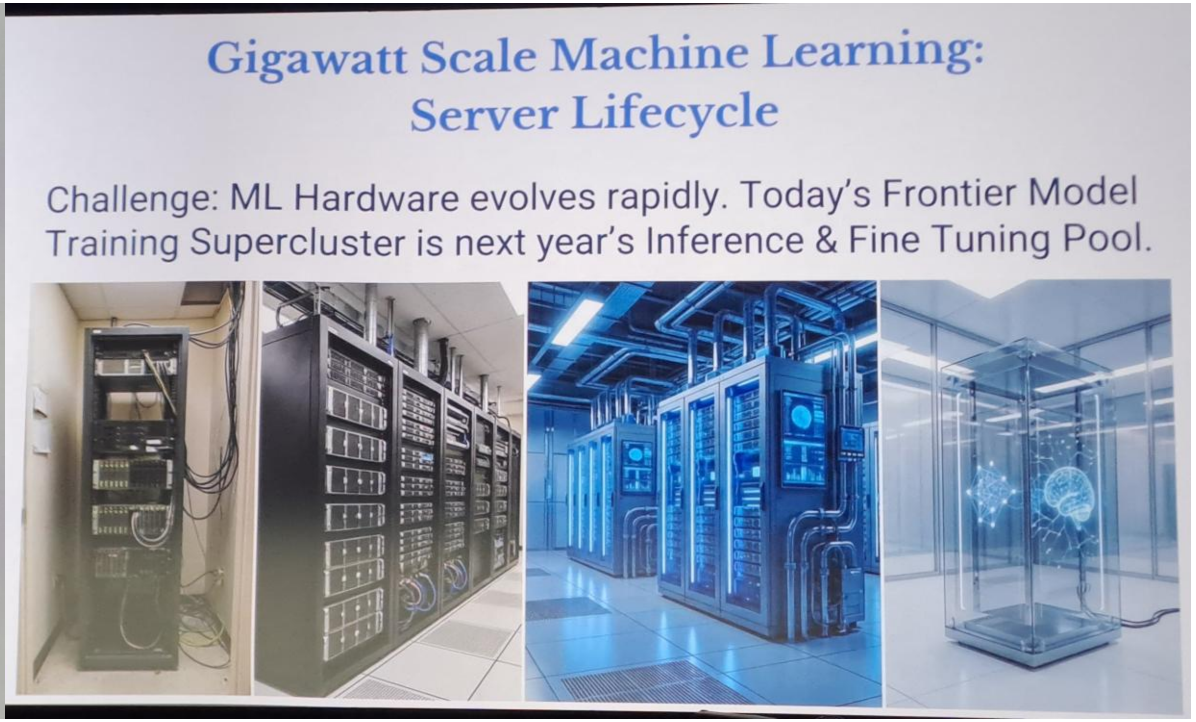
Example: Microsoft Azure AI Superfactory

Source: https://blogs.microsoft.com/blog/2025/11/12/infinite-scale-the-architecture-behind-the-azure-ai-superfactory/

SC25 Invited Talk by Google:
“Gigawatt Scale Machine Learning – Deployment & Operation”

- Main Topics: Planning & Construction, Deployment, Network,
  Server Lifecycle, Reliability
- Quote: „With every 10x increase in scale,
  you need to re-do everything“





SC25 Panel:
“AI Factory Supercomputers Are Not HPC Supercomputers”

- AI Supercomputers
  - Purpose built
  - Bigger in scale
  - Need new considerations
  - Might span across cities
- AI Factories
  - AI workflows
  - Producing tokens / “intelligence”
  - Should be automated
- Power is the main constraint !

- Data for continuous training:
  - Huge amount of data input
  - Object metadata, real-time
  - Significant effort for data preparation/cleaning
- Storage:
  - Optimized for random access (IOPs)
  - Fast local storage per node
  - Asynchronous checkpointing (every 15-30 min.)

ISC High Performance 2026

Keynotes

- Prof. Martin Schulz - HPC: A Heterogeneous Future
- Prof. Amanda Randles - HPC for Vascular Digital Twins
- Prof. Jack Dongarra - HPC In Transition
- Recordings: https://www.youtube.com/playlist?list=PLfC_0SkiYlaU

AI Sessions

- AI for Science
- AI Factories and Gigawatt Datacenters
- Physical AI & Robotics
- Recordings: https://www.youtube.com/playlist?list=PLeUAPtaQ36dc

ISC26 Panel on AI Factories – Future Challenges

Heterogeneous accelerators

- Training, inference, modelling/simulation, quantum
- Management complexity of different interconnected workloads

Orchestration of jobs and data

- Data movement between accelerator memory

Performance evaluation

- Selection of best performing accelerator for a job
- Integrated benchmarking

Discussion items from Jason Haga, AIST, Japan

Recent Initiatives by the European Union

- The following slides are based on presentations by the EuroHPC Joint Undertaking
- Sources:
  - https://www.eurohpc-ju.europa.eu/media-events/events/eurohpc-ai-factories-update-webinar-2025-10-17_en
  - https://www.eurohpc-ju.europa.eu/media-events/events/eurohpc-user-days-2025-2025-09-30_en
  - https://www.eurohpc-ju.europa.eu/eurohpc-quantum-computers_en
  - https://www.eurohpc-ju.europa.eu/ai-factories_en
  - Presentation by Anders Jensen at ISC High Performance 2026
  - Press Release by European Commission on July 30th, 2026 https://ec.europa.eu/commission/presscorner/detail/en/ip_26_1708

ABOUT THE EUROHPC JOINT UNDERTAKING

EuroHPC Joint Undertaking

- An EU body and funding entity based in Luxembourg, established in 2018 and autonomous since 2020
- Governed by a Board composed of the European Commission, 38 Participating States and 3 Private Members

Digital Europe Programme

- Infrastructure
- Federation of supercomputing services
- Widening usage and skills

Horizon Europe Programme

- Technology
- Applications
- International Cooperation
- Quantum R&I

Connecting Europe Facility

- Hyperconnectivity
- Data Connectivity

- Procure, deploy and maintain **High Performance Computing, AI Factories, AI Gigafactories and Quantum infrastructure** in Europe
- **Fund innovative Research and Innovation projects**, to develop European skills, applications, software and hardware and foster a European supply chain
- Provide **access to High Performance Computing, AI Factories, AI Gigafactories and Quantum resources** and user support

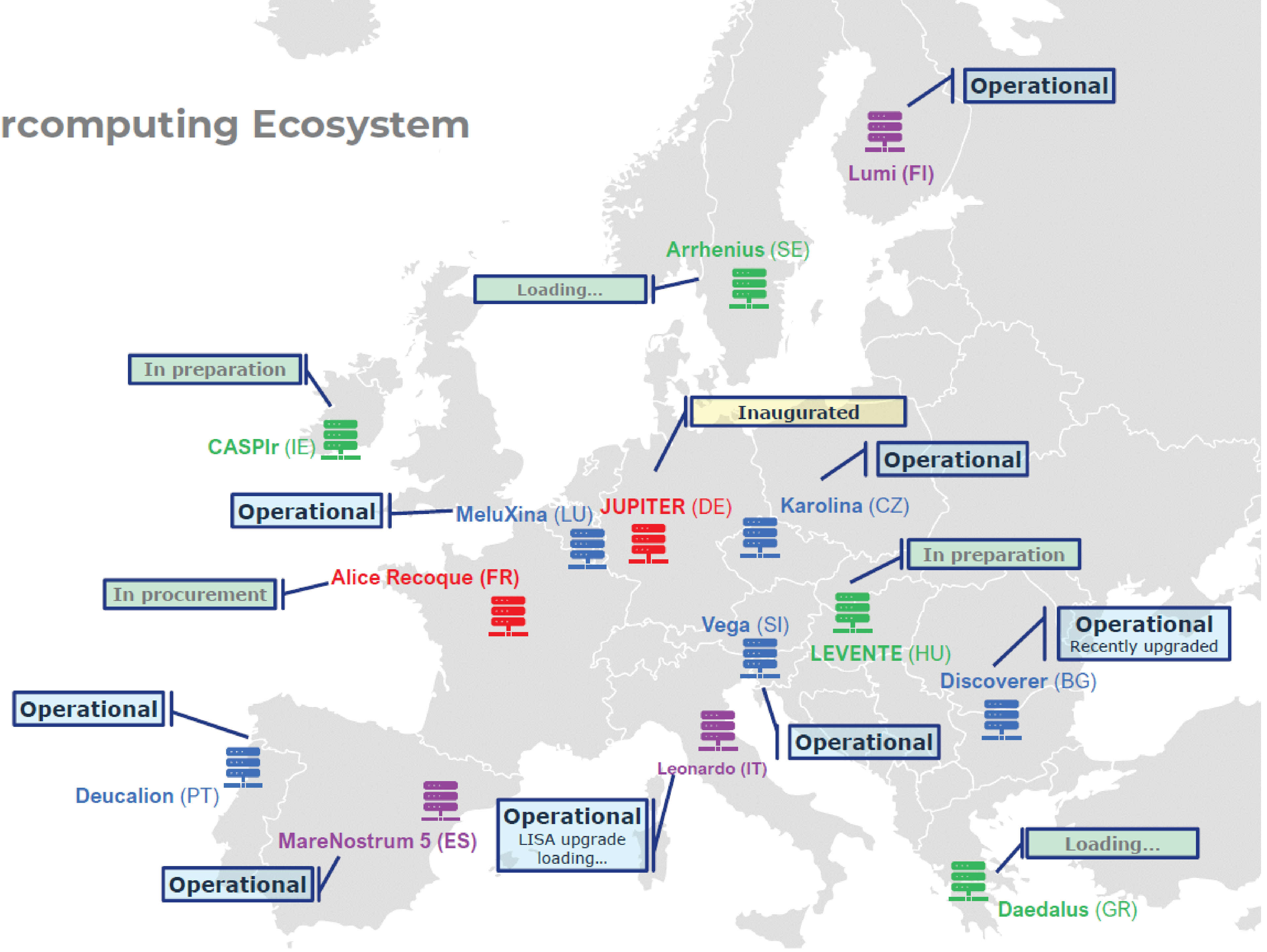
The EuroHPC Supercomputing Ecosystem

EXASCALE

PRE-EXASCALE

PETASCALE

MID-RANGE



FEDERATION PLATFORM & HYPERCONNECTIVITY

EuroHPC
Joint Undertaking

"EuroHPC JU aim to develop, deploy, extend and maintain in the EU a world-leading federated, secure and hyper-connected supercomputing, quantum computing, service and data infrastructure ecosystem"

Federation Platform

Designed to support an expanding range of workflows and enable innovative approaches to complex challenges.

Timeline

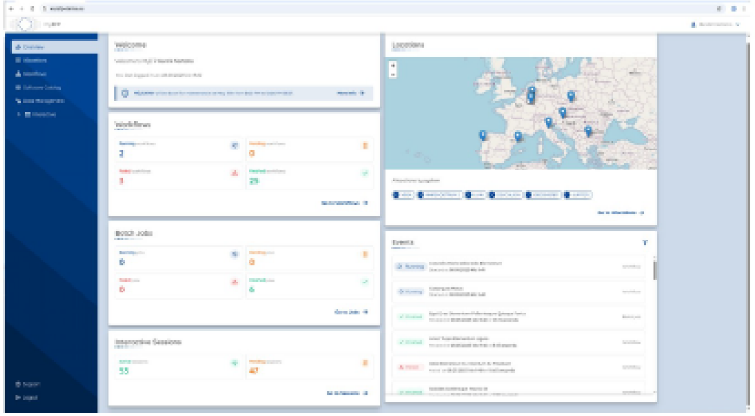
- Jan 2025: Start of development
- Sept 2025: Collaboration agreements with EuroHPC Hosting Entities to streamline integration
- April 2026: Initial release

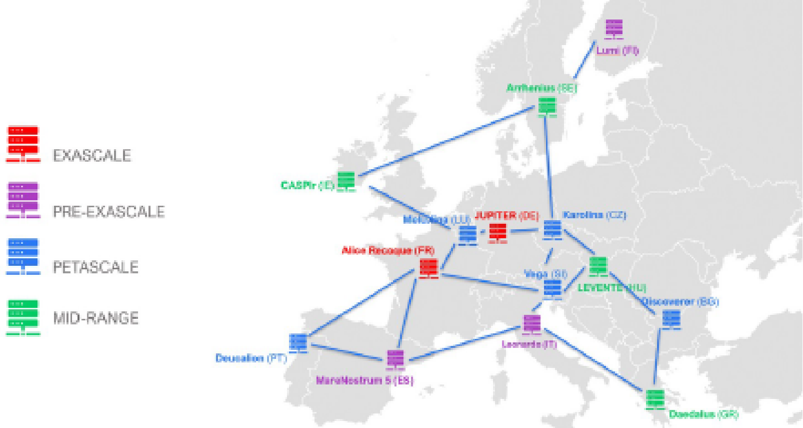
HyperConnectivity

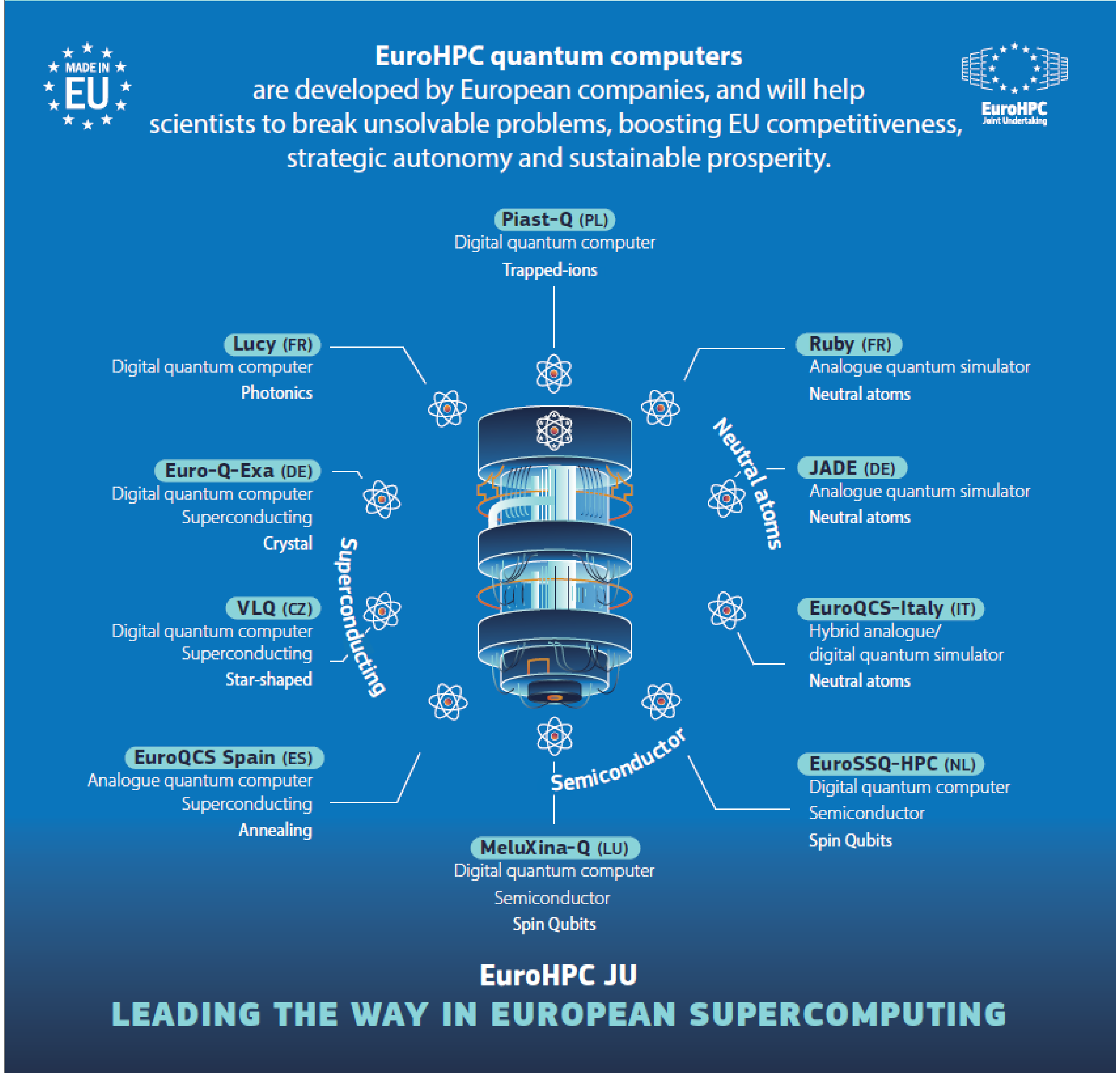
Deliver terabit-per-second capacity with robust security flexible, and future-proof services.

Timeline

- Sep 2025: Framework Contract Signed
- 2026: High Level Design (HLD) and HyperConnectivit Service Area (HCSA) delivered
- 2029: Fully hyperconnected ecosystem in operation

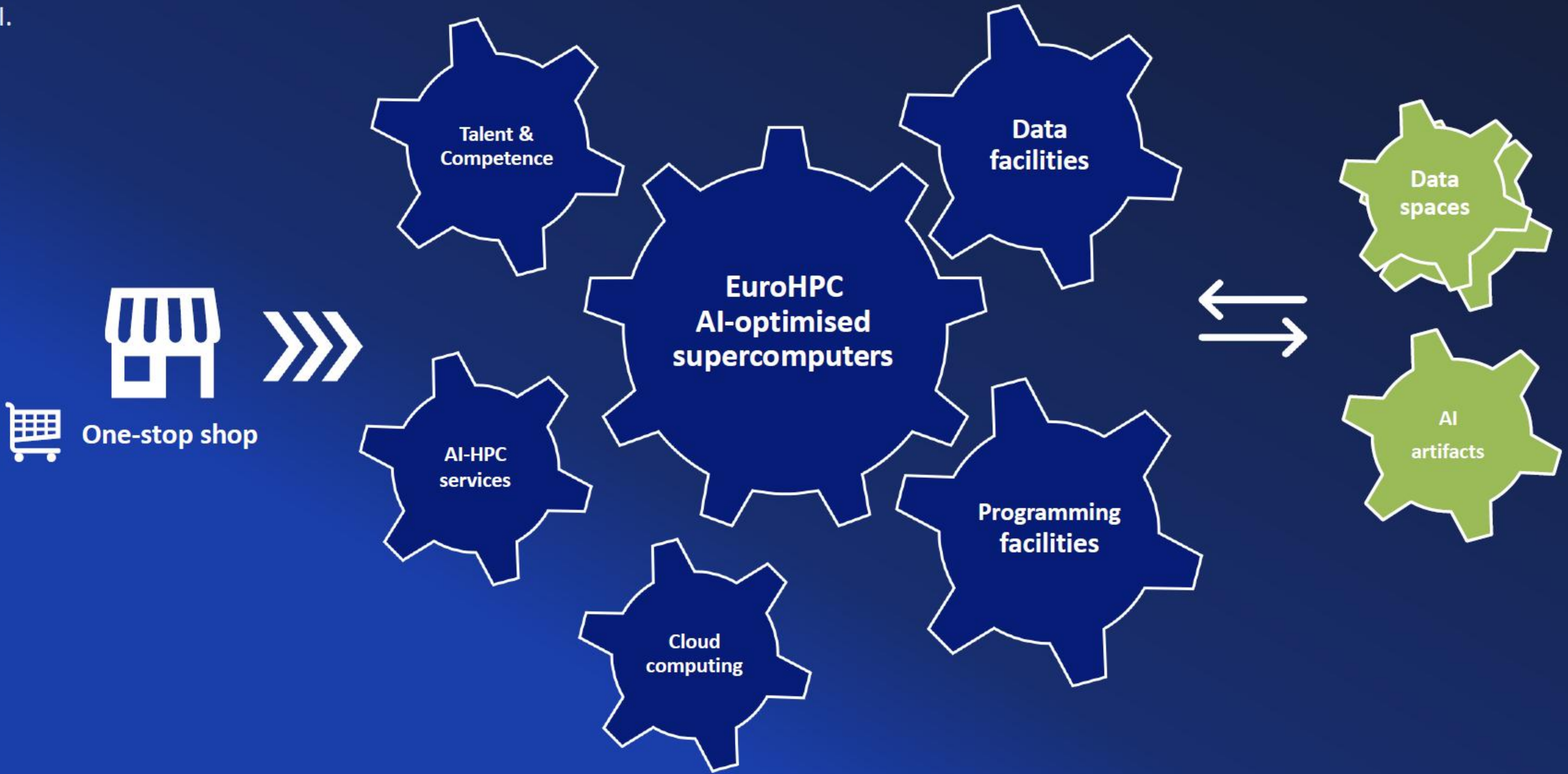


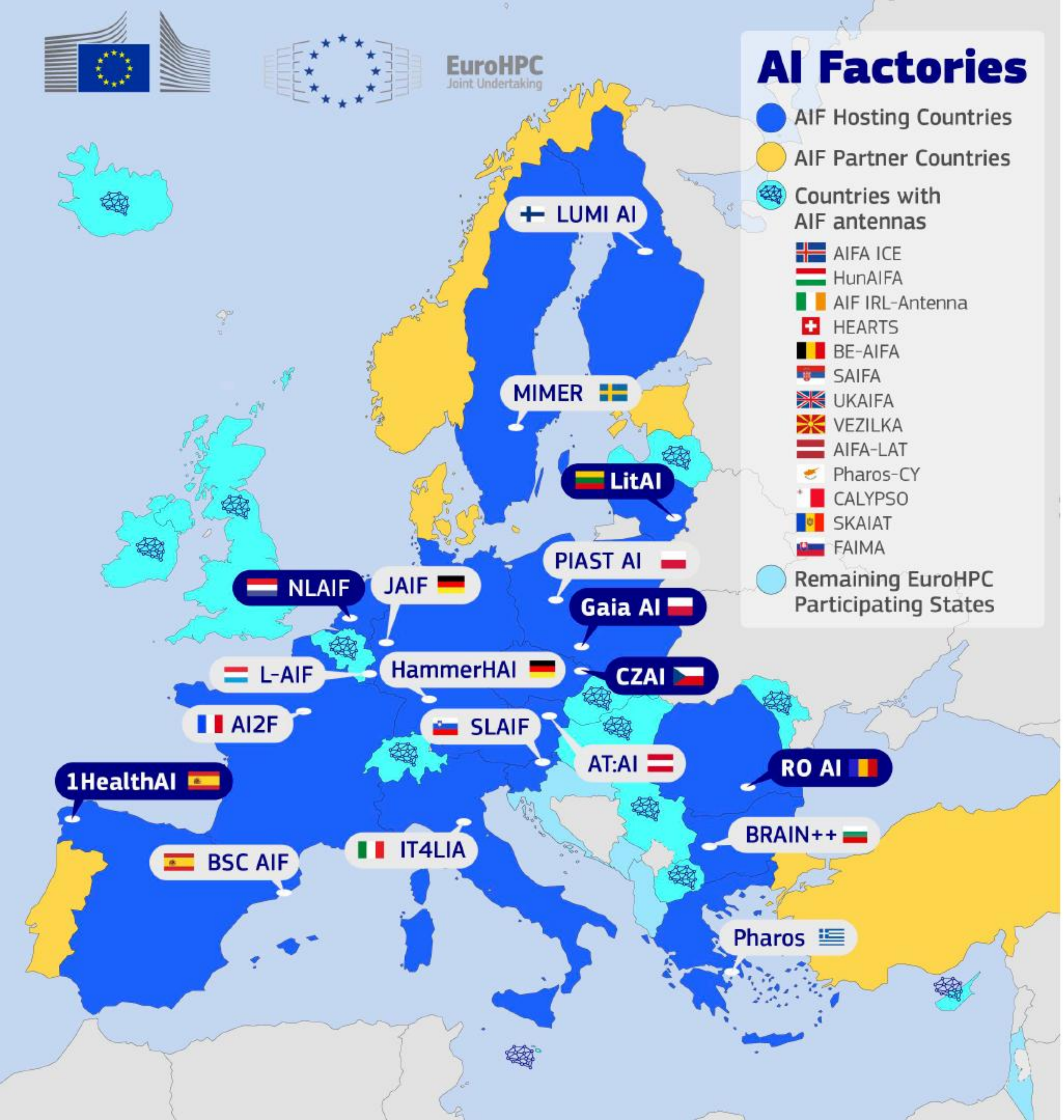



EuroHPC quantum computers
are developed by European companies, and will help scientists to break unsolvable problems, boosting EU competitiveness, strategic autonomy and sustainable prosperity.
MADE IN EU
EuroHPC Joint Undertaking
Piast-Q (PL)
Digital quantum computer
Trapped-ions
Lucy (FR)
Digital quantum computer
Photonics
Ruby (FR)
Analogue quantum simulator
Neutral atoms
Euro-Q-Exa (DE)
Digital quantum computer
Superconducting
Crystal
JADE (DE)
Analogue quantum simulator
Neutral atoms
Neutral atoms
Superconducting
VLQ (CZ)
Digital quantum computer
Superconducting
Star-shaped
EuroQCS-Italy (IT)
Hybrid analogue/
digital quantum simulator
Neutral atoms
EuroQCS Spain (ES)
Analogue quantum computer
Superconducting
Annealing
Semiconductor
EuroSSQ-HPC (NL)
Digital quantum computer
Semiconductor
Spin Qubits
MeluXina-Q (LU)
Digital quantum computer
Semiconductor
Spin Qubits
EuroHPC JU
LEADING THE WAY IN EUROPEAN SUPERCOMPUTING

AI Factories

AI Factories are dynamic ecosystems build around AI-optimised supercomputers, offering computing resources and support services to the European industry, as well as to the European scientific users for the development of large AI models to take advantage of AI technology capabilities in the European Union, and for the development of skills and knowledge in the domain of AI.





EuroHPC JU AI Ecosystem

19 AI Factories

Austria – AT:AI
Bulgaria – Brain++
Czechia - CZAI
Finland – Lumi AI
France –AI2F
Germany – HammerHAI
Germany – JAIF
Greece - Pharos
Italy – IT4IA
Lithuania – LitAI
Luxembourg – LAIF
Netherlands – NLAIF
Poland – Gaia AI
Poland – PIAST AI
Romania – RO AI
Slovenia - SLAIF
Spain – BSC AIF
Spain - HealthAI
Sweden – Mimer AI

13 AI Antennas

Belgium – BE-AIFA
Cyprus – Pharos CY
Hungary - HunAIFA
Iceland – AIFA ICE
Ireland – AIF-IRE-Antenna
Lativa – AIFA-LAT
Malta – Calypso
Moldova - FAIMA
North Macedonia – VEZILKA
Serbia - SAIFA
Slovakia - SKAIAT
Switzerland – HEARTS
United Kingdom – UKAIFA

AI System procurements

Tenders evaluated and awarded
- **MN5** Upgrade (BSC-AIF | BSC) (Signed)
- **HammerHAI** (HLRS) (Signed)
- **MIMER** (Linköping University) (Awarded, pending contract signature)
- **IT4LIA** (CINECA) (Awarded, pending contract signature)

In Tendering stage
- **LUMI AI** (LAIF | CSC)

In Selection stage
- **AI:AT** (AIF Austria | ACA)
- **NLAIF** (AIFNL)

Under preparation
- **PIAST AIF** (PSNC)
- **Vega-AI** (SLAIF | IZUM)
- **Discoverer++** (Sofia Tech Park)
- **CZAI** (IT4I)

AIFs relying on systems from previous calls
- **JUPITER** (JAIF | JSC)
- **Alice Recoque** (AI2F | GENCI)
- **Daedalus** (Pharos | GRNET)

Relaunched as open procedure
- **MeluXina AI** (L-AIF | LuxProvide)

All AIF systems from 3rd cut off at preliminary stage!

https://www.eurohpc-ju.europa.eu/ai-factories_en

What to expect from the EuroHPC AI-optimised supercomputers

EuroHPC Joint Undertaking

- AI optimised GPU architectures: Emphasis on **FP32**, **FP16** and **FP8** performance
- Tightly integrated on node level – Memory coherency
  - NV Link, UALink, Infinity Fabric
- **Revised benchmarks**
  - HPL and Top500 not the driving metric.
- Dense interconnect on the partition level
  - At least one link per GPU
  - Software friendly
  - Traditional Infiniband but also Ethernet (e.g. Ultra Fast Ethernet)

- Richer software/middleware stacks
  - Slurm is not the king
  - Kubernetes, containers, virtualisation and relevant workflow and resource management tools
- **MLOPs**
  - Support for the complete AI training/inference lifecycle: Preparation, fine tuning, augmentation, model catalogues.
- **Multitenancy** and advanced security.
  - User isolation on compute, storage and network layer

EuroHPC AI supercomputers will offer diverse architectures and design choices suited for the specific goals of the various AI Factories

Other AI Factories key ingredients

Data Labs

- Access to **high-quality** data
- Wide availability of **open public AI-ready** data
- Access to **Common European Data Spaces**
- **Trusted/secured** access to and reuse of **industrial data**

Skills and talent

- Development and retention of **talent in Europe**
- Comprehensive **AIF's skills plan**, including tailored activities and educational initiatives (workshops, MOOCs, hands-on training, training sandboxes, etc.)

Cloud

- Access to **sovereign cloud capabilities**
- Crucial for **further model development (AI inference), finetuning, and deployment, application development and for scaling up**

Networking

- Collaboration and networking **among AIF Factories** (knowledge sharing, resource optimisation, domain specialization, etc.)
- Links to **EU and National (AI) initiatives**

AI GIGAFACTORIES

Taking AI Factories to the next level by **integrating coherently** massive computing power, **beyond 100 000 advanced AI processors**.

Large-scale facilities designed to **develop, train, and deploy the next generation most complex AI models at an unprecedented scale.**

Federated with the **EuroHPC network of AI Factories** and sovereign cloud.

Essential for Europe to be able to **compete on the global level** and ensure its **strategic autonomy in science and in critical industrial sectors**.

Focus on power capacity, supply chain, cutting-edge networks, energy-efficiency, and AI-driven automation.

Need for **public-private partnerships** given the magnitude of the required investments.

July 30: EU launches AI Gigafactories call

European Commission

European Commission - Press release

EU launches AI Gigafactories call to boost Europe's computing capacity and unlock more than €30 billion in investment

Brussels, 30 July 2026

The EU today launched a call for tenders to establish up to seven AI Gigafactories across Europe, as part of its latest major push to accelerate Europe's technological sovereignty and ambition of becoming the AI Continent.

Led by industry and supported by up to €10 billion in EU and national funding, the initiative is expected to unlock at least €20 billion in private investment across the Union. It will expand Europe's AI computing capacity and give start-ups, scale-ups, small and medium-sized enterprises, industry, academia and public authorities access to infrastructure for training, inference and fine-tuning advanced frontier AI models.

The AI Gigafactories will combine advanced AI processors, software and cloud technology stacks, high-speed connectivity and energy-efficient data centres. Together with Europe's network of 19 AI Factories, they will strengthen Europe's technological leadership, resilience and strategic autonomy.

The initiative will ensure that Europe can develop advanced AI on its own infrastructure, in line with EU rules and values. AI technologies developed in the AI Gigafactories will follow EU standards on data protection, safety, security and ethics.

AI Gigafactories – Definition

- „AIGF refers to state-of-the-art large-scale facility with sufficient capacity to handle the complete lifecycle, from development to large-scale inference, of very large AI models and applications,

  providing a supercomputing service infrastructure which is composed of AI-optimised computing capacity, a supporting data centre infrastructure including high-capacity storage and networking,

  dedicated secure cloud user access environments, and specialised secure AI-oriented support services for its advanced operations,

  all of which are supported by an environmentally sustainable infrastructure, in particular for energy and water supply systems."

AI Gigafactories Call – Main Facts

- Deadline for proposals: November 12th
- Award decisions are expected by early 2027
- Start of operation at latest 18 months after signature
- 2 phase approach
- Lot 1: €4 Billion (EU + MS) for up to 4 Medium-Scale AI Gigafactories
  - €100 M from EU for phase 1 per project
  - €400 M from EU for phase 2 per project
- Lot 2: €6 Billion (EU + MS) for up to 3 Large-Scale AI Gigafactories
  - €200 M from EU for phase 1 per project
  - €800 M from EU for phase 2 per project

AI Gigafactories Call – Metrics

- Lot 1: Medium-Scale AI Gigafactories
  - Phase 1: matches Europe's most powerful AI factory
  - Phase 2: increase to at least three times that level
- Lot 2: Large-Scale AI Gigafactories
  - Phase 1: 2x compared with Europe's most powerful AI factory
  - Phase 2: increase to at least four times that level
- AI Accelerator Benchmark Methodology
  - Per-Precision Compute Throughput
  - Memory Bandwidth
  - Memory Efficiency Factor
  - -> Global Performance Score (normalized against NVIDIA H100)

$$\mathbf{P[p]} = \frac{\mathbf{1}}{\frac{\mathbf{1}}{\mathbf{I\_stream[p]} \times \frac{\mathbf{BW}}{\mathbf{f}}} + \frac{\mathbf{1}}{\mathbf{C[p]}}}$$

$$\mathbf{GP = \Sigma p\ X[p] \times P[p]}$$

where $p \in \{FP32, BF16, FP16, INT8, FP8, INT4, FP4\}$.

The weighting profile ($X[p]$) is defined as follows:

p	FP32	BF16	FP16	INT8	FP8	INT4	FP4
X[p]	0.05	0.175	0.175	0.1875	0.1875	0.1125	0.1125

AI Gigafactories Call – Links

- Press Release: https://ec.europa.eu/commission/presscorner/detail/en/ip_26_1708
- Tender Details: https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/tender-details/bc712395-692d-41ab-aa25-8e254ae86eb2-CN
- FAQ AI Gigafactories: https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/tender-details/docs/bc712395-692d-41ab-aa25-8e254ae86eb2-CN/FAQ%20AIF%20Gigafactories_v300726_V1.pdf

Questions?
Comments?