SLAIF

SLOVENIAN ARTIFICIAL INTELLIGENCE FACTORY

Slovenska tovarna umetne intelligence

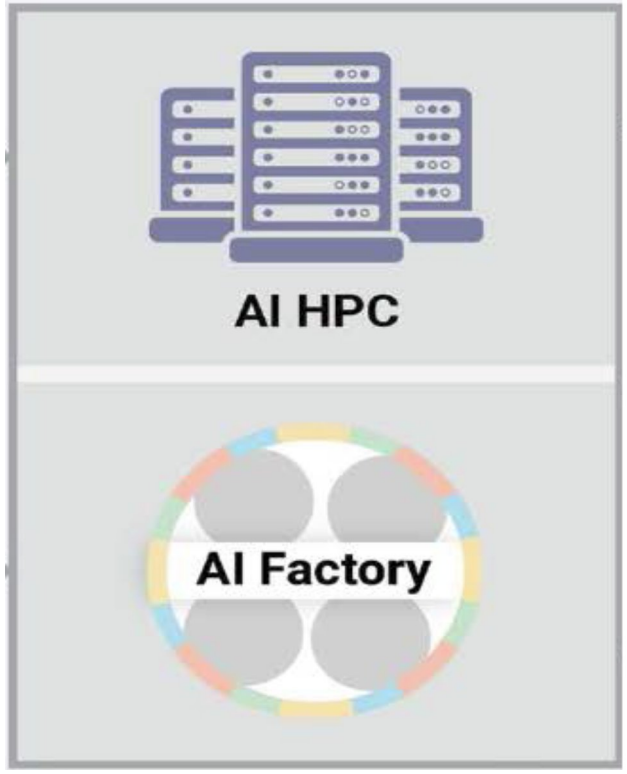
Sašo Džeroski, Dragi Kocev
Andrej Filipčič, Jan Jona Javoršek

ARTIFICIAL INTELLIGENCE FACTORIES: WHAT ARE THEY?

AI factories have two main components

- AI-optimized supercomputer (AI-HPC)
- AI-factory activities and services

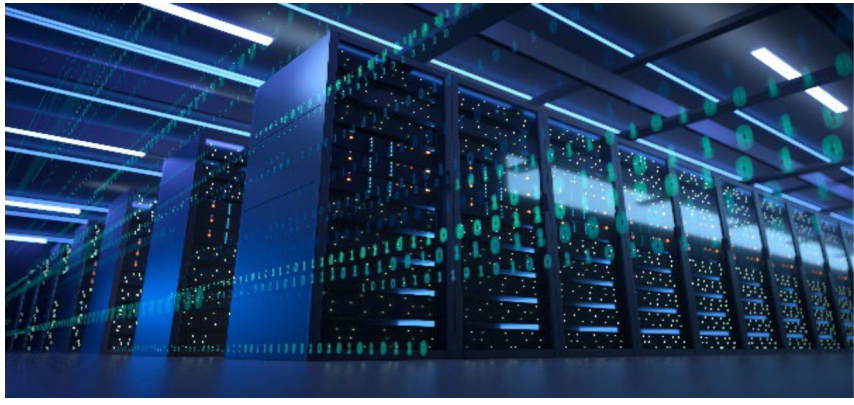
European Parliament



The European Commission describes **AI factories** as **dynamic ecosystems** formed around European supercomputers **that** bring together critical components that **facilitate the development of cutting-edge generative AI models**. These **components** are the following: **AI-dedicated supercomputers**, associated **data centres** in proximity or connected via high-speed networks, and **talent**, including data specialists, researchers and SMEs. AI factories will also provide training. Supercomputing support services centres are also an essential part of AI factories, facilitating access to supercomputers and to dedicated supercomputer-friendly programming facilities and algorithmic support.

Supercomputers are high-performing computing (HPC) systems with very high computational power capable of performing complex and large-scale computational tasks. They can process and analyse large volumes of data at a rate that exceeds other computers by far. Supercomputers require a scaled-up software that organises, assigns, stores and processes data in a particular way. A typical supercomputer has many general-purpose central processing units (CPUs), similar to those found in personal computers. In supercomputers, these general-purpose CPUs are connected via ultrafast networks. Supercomputers adapted to AI development also contain graphic-processing units (GPUs).

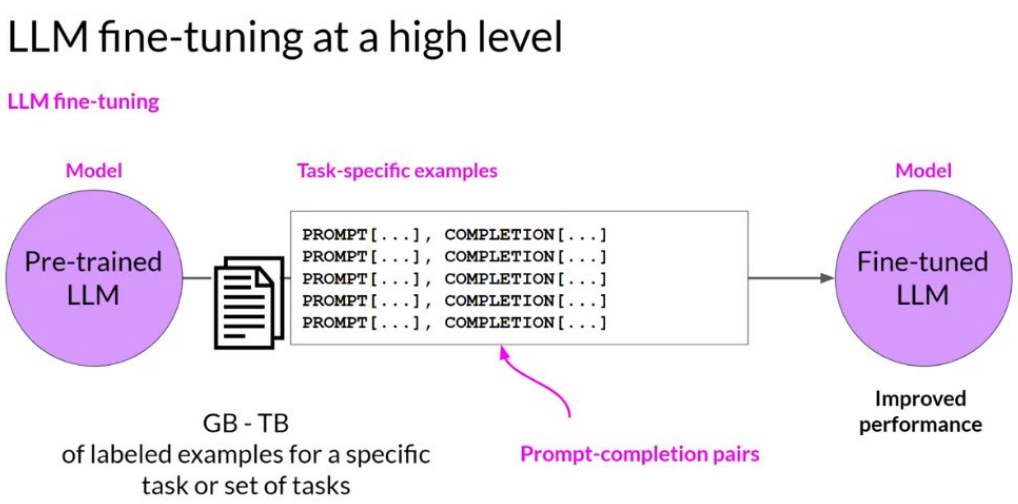
Sources: European Commission and *The Economist*.

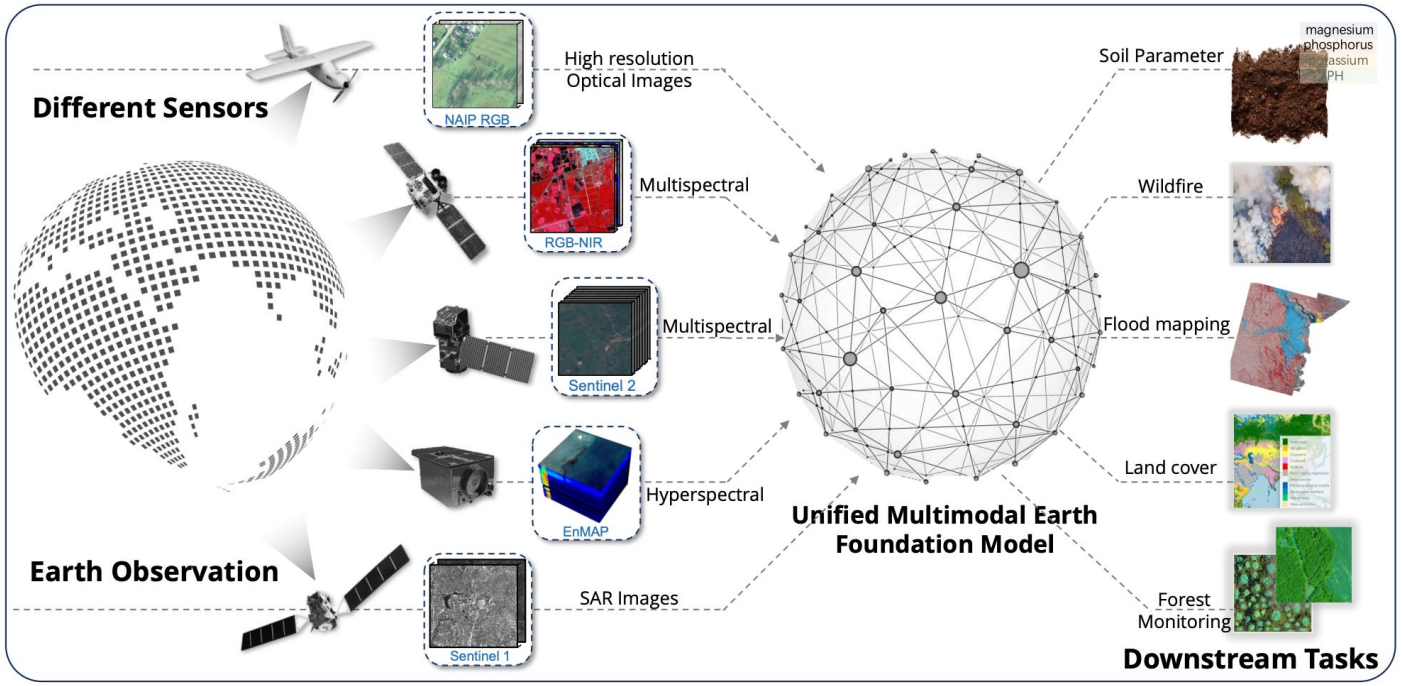


WHY DO WE NEED ARTIFICIAL INTELLIGENCE FACTORIES?

Mostly to **pre-train** and **fine-tune** LLMs and other generative AI models, incl. multi-modal foundation models (working on, e.g., both text and images)

- *Foundation models (FMs) are large models, generated by applying ML (deep learning) to a broad collection of data at scale. Can be **adapted** for use in **a wide range of downstream tasks.** LLMs are a prime example of FMs.*

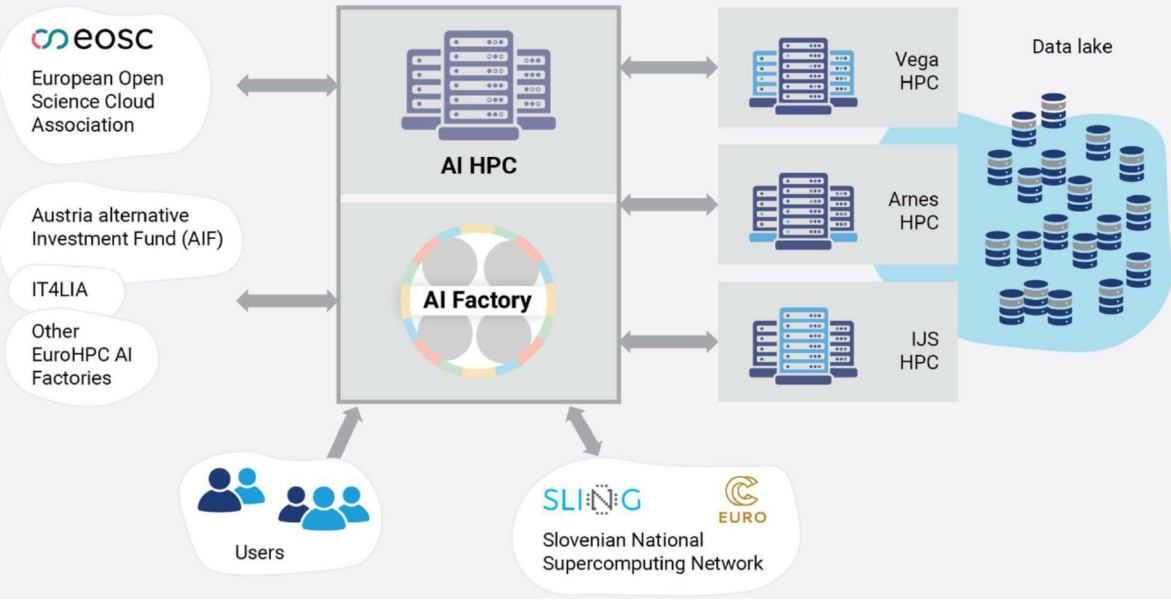




- ***Multi-modal FMs (e.g., with different RSI modalities)***

SLAIF: AI AS INFRASTRUCTURE FOR SOCIETY AS A WHOLE

The Slovenian AI factory will support the growth of the Slovenian & regional AI ecosystem



EuroHPC Joint Undertaking

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA VISOKO ŠOLSTVO, ZNANOST IN INOVACIJE
REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PREOBRAZBO

*At JSI, we view **AI** as **infrastructure** for science for quite a while*
*SLAIF will support industry, public sector, science: AI infrastructure **for the whole society***
*AI HPC Consortium: **IZUM (Hosting Entity)**, JSI, ARNES*
*AI Factory Consortium: **JSI (technical coordinator)**, UniLj, UniMb, FIS NM, UniNG, UniPr, IZUM (administrative coordinator), ARNES, GZS, TPLj*

THE AIF CONSORTIUM

The SLAIF consortium brings together Slovenia's leading institutions in research & education, HPC infrastructure, and industry support, creating a powerful alliance to implement the AI Factory. In collaboration with **IT4LIA** and **AI:AT**.

Research Powerhouses

Jožef Stefan Institute (JSI) leads the consortium with strong expertise in HPC, AI, and technology transfer. **University of Ljubljana** (UL) and **University of Maribor** (UM) additionally contribute educational capabilities.

Infrastructure Specialists

IZUM serves as the EuroHPC JU Hosting Entity for HPC Vega, while **ARNES** provides advanced internet services and coordinates **SLING**. Faculty of Information Studies (**FIS**) has experience in both AI and HPC infrastructure.

Industry Connectors

Technology Park Ljubljana (TPLJ) supports SMEs and startups in adopting advanced technologies, while the **Chamber of Commerce and Industry** of Slovenia (GZS) represents over 5,200 member companies.

Regional Expertise

University of Nova Gorica (UNG) and **University of Primorska** (UP) provide additional research capabilities and international connections, strengthening the consortium's regional coverage. Others through **SLING**.

AI FACTORY WORK PACKAGES

The SLAIF implementation is organized into seven interconnected work packages, each with specific objectives and responsibilities to comprehensively develop the AI Factory.

WP1: Project Management (JSI)

Ensures smooth coordination across consortium partners and stakeholders, establishes inclusive policies for service access, and maintains quality assurance to meet project milestones effectively.

WP2: Workflow and Data Orchestration Infrastructure (JSI)

Develops robust infrastructure integrating HPC and cloud resources for scalable execution of AI workflows and establishes a national data lake enabling automated data exchange.

WP3: Core AI Platform (UL)

Delivers core AI platform workflows as generic horizontal services, maintains a semantic catalogue of AI components, and supports development and deployment of AI workflows.

WP4: User-Centric AI Services (UM)

Focuses on making HPC/AI clusters more user-friendly for first-time users, providing libraries, tools, and interfaces for seamless connection to services.

Complemented by **WP5** (Vertical Services for Sectorial Applications), **WP6** (AI Skills and Training), and **WP7** (Outreach and Dissemination).

VERTICAL SERVICES FOR SECTORIAL APPLICATIONS

Demonstrate the impact of AI across multiple sectors, developing specialized applications that address specific needs of industry, society and science. Leverage core AI platform to deliver tailored solutions.

AI for Green Transition

Revolutionizing agriculture with precision farming, environmental monitoring with EO data, optimizing energy systems, and enhancing smart manufacturing for sustainability.

AI for Health and Biotechnology

Analyzing complex biosignals for early disease detection, delivering personalized treatment plans, accelerating drug discovery, and developing tools for medical decision-making.

AI for Digital Society

Adapting language models for Slovene and other less-resourced languages, empowering creative industries, streamlining public administration, and transforming education.

AI for Science

Automating scientific model discovery, accelerating life sciences research, facilitating materials science innovation, supporting environmental sciences, and advancing digital humanities.

AI SKILLS AND TRAINING

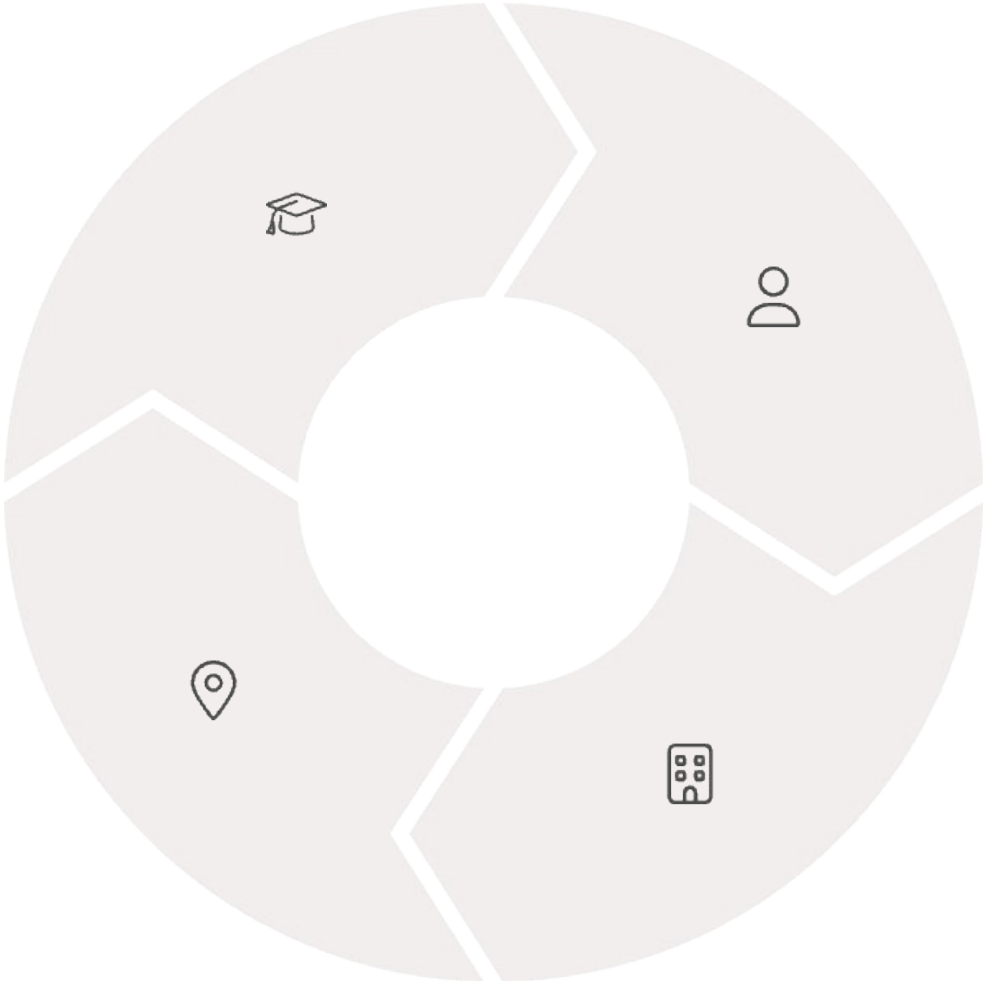
We will develop AI and HPC talent through **customized training programs** that address the needs of diverse users, from students to professionals and SMEs. This will ensure that Slovenia builds the necessary skills for AI adoption and develops the talent needed to fully leverage AI technologies.

AI Talent Development

Expanding formal education pathways, offering accelerated learning opportunities, and establishing a national HPC/AI certification registry.

Regional Onboarding

Establishing hubs in Ljubljana, Maribor, and Novo Mesto to provide workshops, consultancy, and help-desk services for new users.



Customized Training

Tailoring programs for different user groups, from foundational modules to advanced topics, with sector-specific training and regional outreach.

SME Engagement

Empowering businesses with workshops on practical AI applications, user-friendly platforms, and case studies showcasing successful implementations.

AI FOR SCIENCE

AI for Materials Science

Accelerating discovery of new materials, optimizing properties, and simulating complex molecular interactions

AI for Climate Science

Enhancing climate models, analyzing complex atmospheric data, and improving predictions for climate change impacts

AI for Fundamental Research

Supporting physics, chemistry, and astronomy with advanced pattern recognition and data analysis capabilities

Gravity ARIS projects on:

- Artificial Intelligence for Science
- Large Language Models for Digital Humanities

AI FOR SCIENCE

AI for Materials Science

Accelerating discovery of new materials, optimizing properties, and simulating complex molecular interactions

AI for Climate Science

Enhancing climate models, analyzing complex atmospheric data, and improving predictions for climate change impacts

AI for Fundamental Research

Supporting physics, chemistry, and astronomy with advanced pattern recognition and data analysis capabilities

Gravity ARIS projects on:

- Artificial Intelligence for Science
- Large Language Models for Digital Humanities

LLMs4EU & ERA Chair Projects

LLMs4EU – Large Language Models for the European Union

(ALT EDIC, in Slovenija JSI, Uni Lj and others)

WP2. Data Collection and Data Infrastructure
- T2.1 Identification and Specifications for Language Data **(JSI lead)**

WP3. Language Technology Tools catalogue
- T3.1 Definition of requirements for a large language and multimodal model catalogue **(JSI lead)**

WP4. Methodologies, algorithms and infrastructure for model fine-tuning

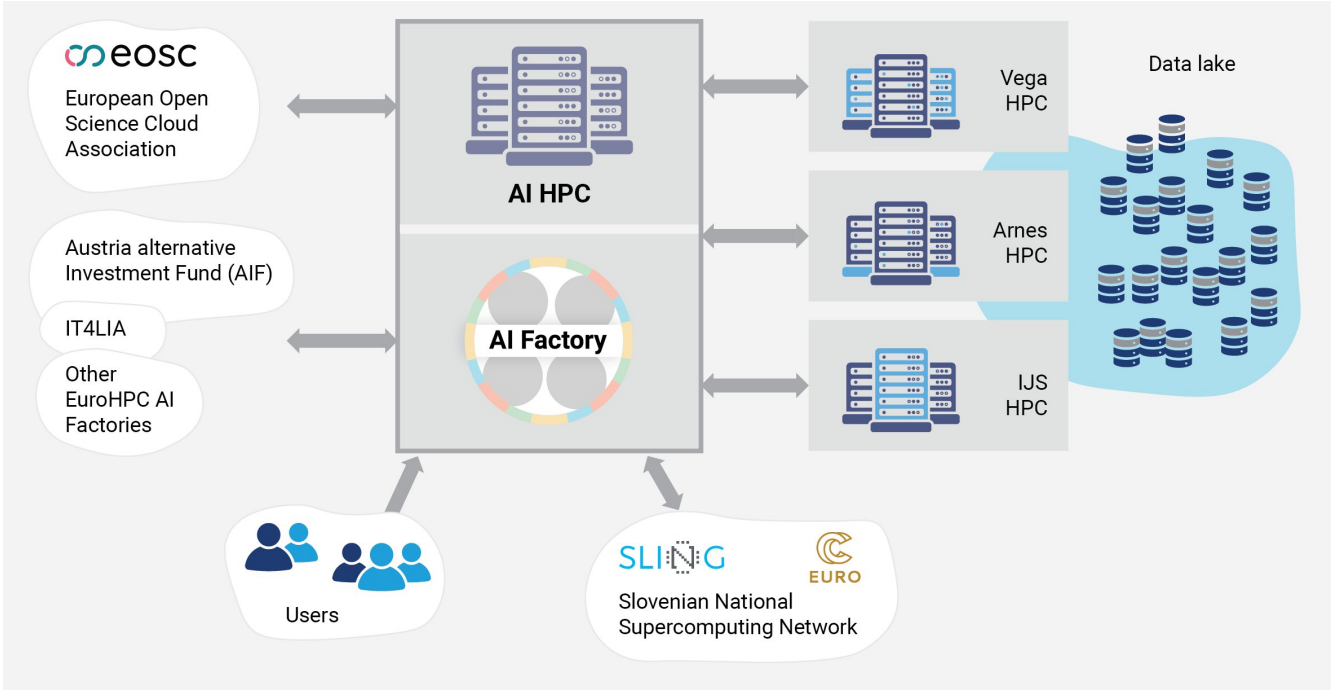
WP5. Distributed Evaluation Centre for LLMs

European Research Area Chair research groups:

AutoLearn-SI: Leveraging Benchmarking Data for Automated Machine Learning and Optimization: Automated Machine Learning (AutoML) and Automated Optimisation (AutoOPT) / Tome Eftimov, JSI

Artificial Intelligence 4 Digital Humanities: methods for collecting and analyzing large-scale textual data in various languages / Marko Robnik Sikonja Uni Lj, Antoine Doucet, @ La Rochelle University, France

SLAIF: Supercomputer (temporary name AI-HPC)



SLING consortium is the main driving force behind the success of the Vega EuroHPC and SLAIF acceptance its latest achievement.

AI-HPC

IZUM
Institut informacijskih znanosti, Maribor

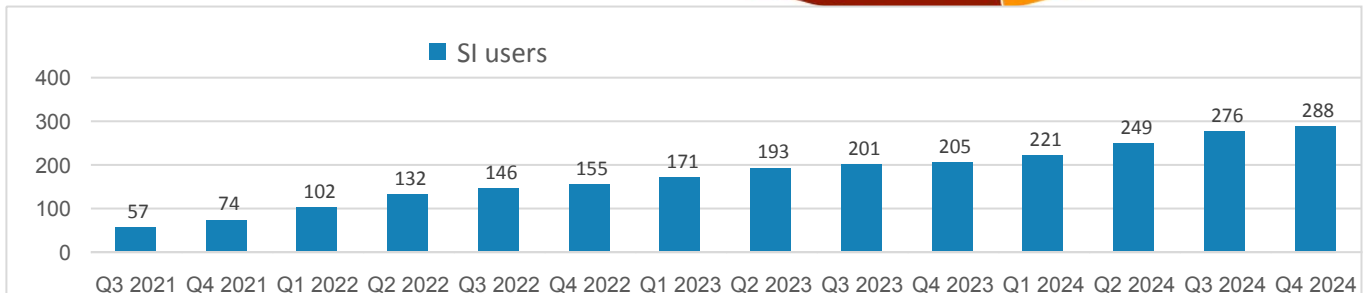
Jožef Stefan Institute
Ljubljana, Slovenia

arnes

- Consortium: IZUM, IJS, Arnes
  - Provides AI services and solutions
  - EuroHPC Joint Undertaking – to boost small and medium enterprise growth
  - Common development within SLAIF and SLING with NCC
- Estimated specifications:
  - Computing power : 100 PFLOPS FP64 / 10 ExaFLOPS FP4/FP8
  - Storage capacity: 100 PB
  - Connectivity (ARNES/GÉANT): ~ 1.2 Tb/s
  - Peak power: 5-6 MW
  - Integration with Open Data Infrastructure and other computing centres

Requirements and Goals

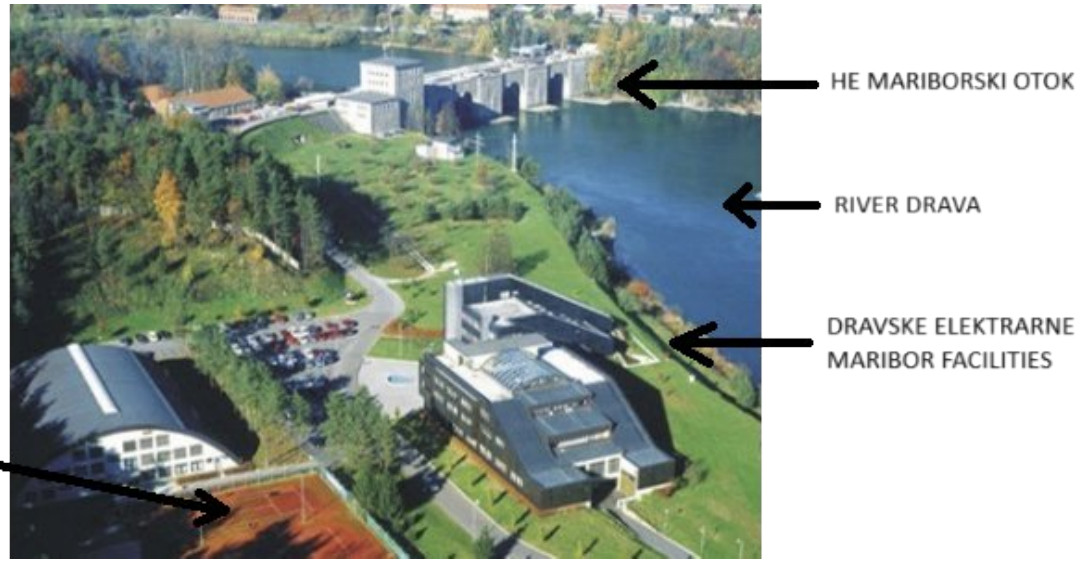
- Vega EOL in 2027, AI-HPC provides existing functionality:
  - support applications using CPU
  - (scientific) applications using GPU
  - permanent data storage (international scientific collaboration)
  - open data infrastructure
- Most of funding for AI-optimised equipment
  - Support double precision compute if possible
- Efficient resource usage and simplicity of access
  - Scheduling of tasks/jobs (SLURM) - fast and efficient execution
  - Cloud infrastructure (OpenStack, K18s) ~ commercial clouds
  - Efficient and fast user support, training, education and outreach
- Integration with other centres:
  - transparent data access to external data and transfers
  - transparent usage of partner centres (IT4LIA HPC, AT HPC,...)
- System, services, functionality and security flexibility for users (DevOps)

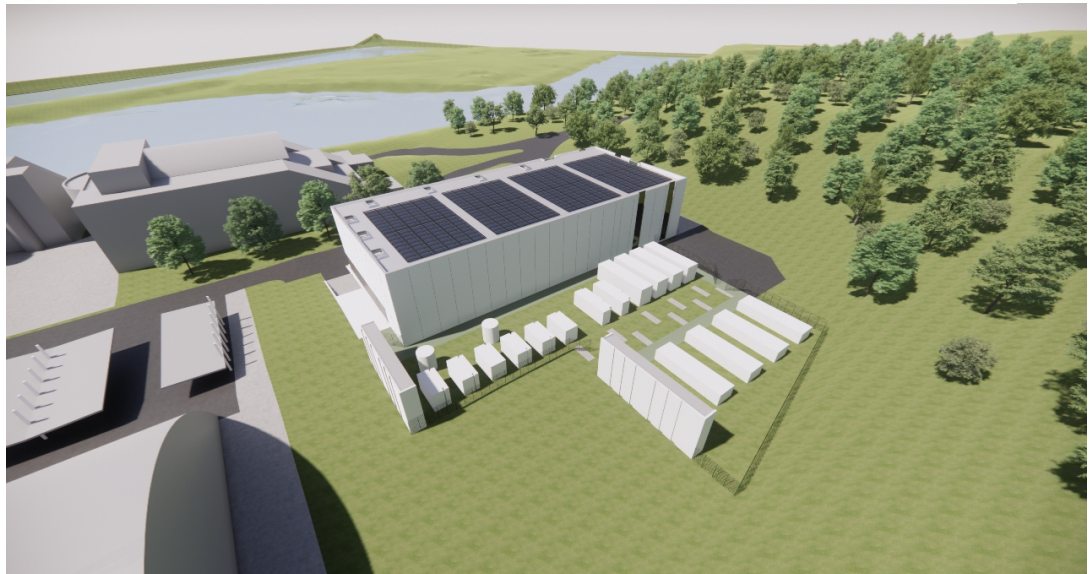


New Data Centre

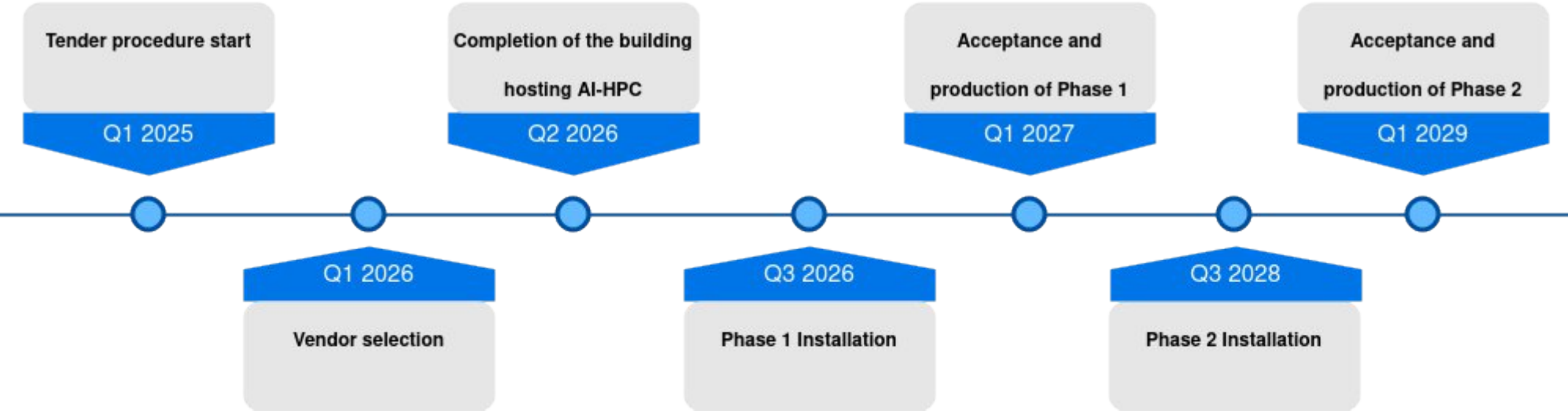
Strategic position: power source, heat reuse, HE location

Construction started May 6th 2025



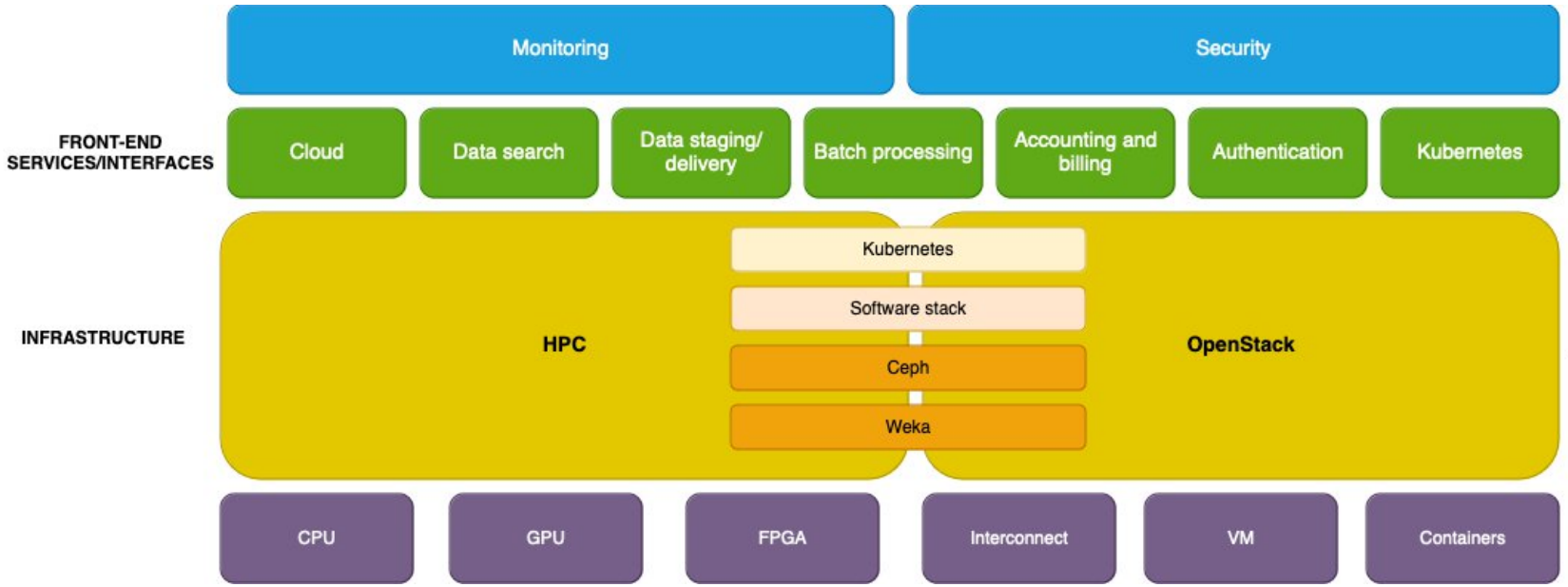


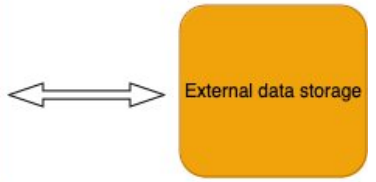
AI-HPC timeline



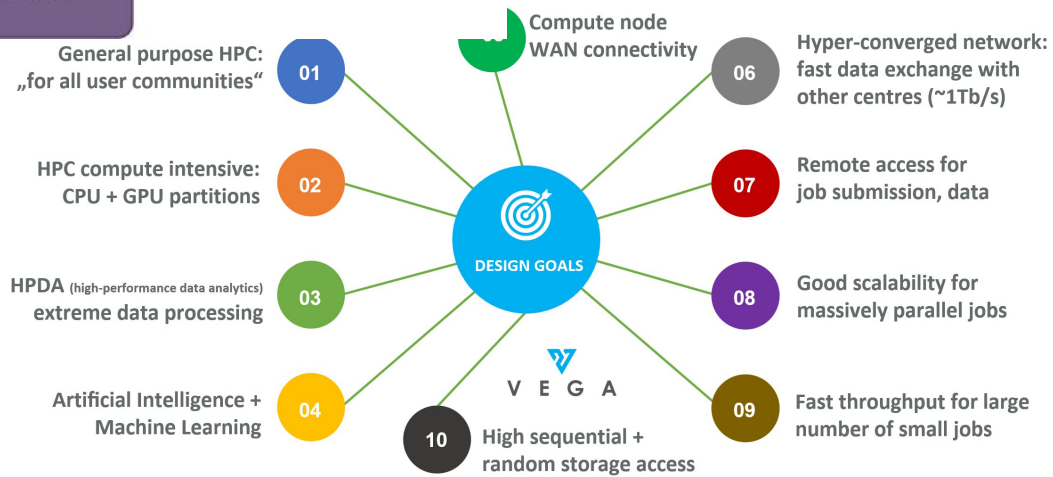
Phase 1: 7-year operation (2034) **Phase 2**: 5-year operation

AI-HPC Hybrid Architecture



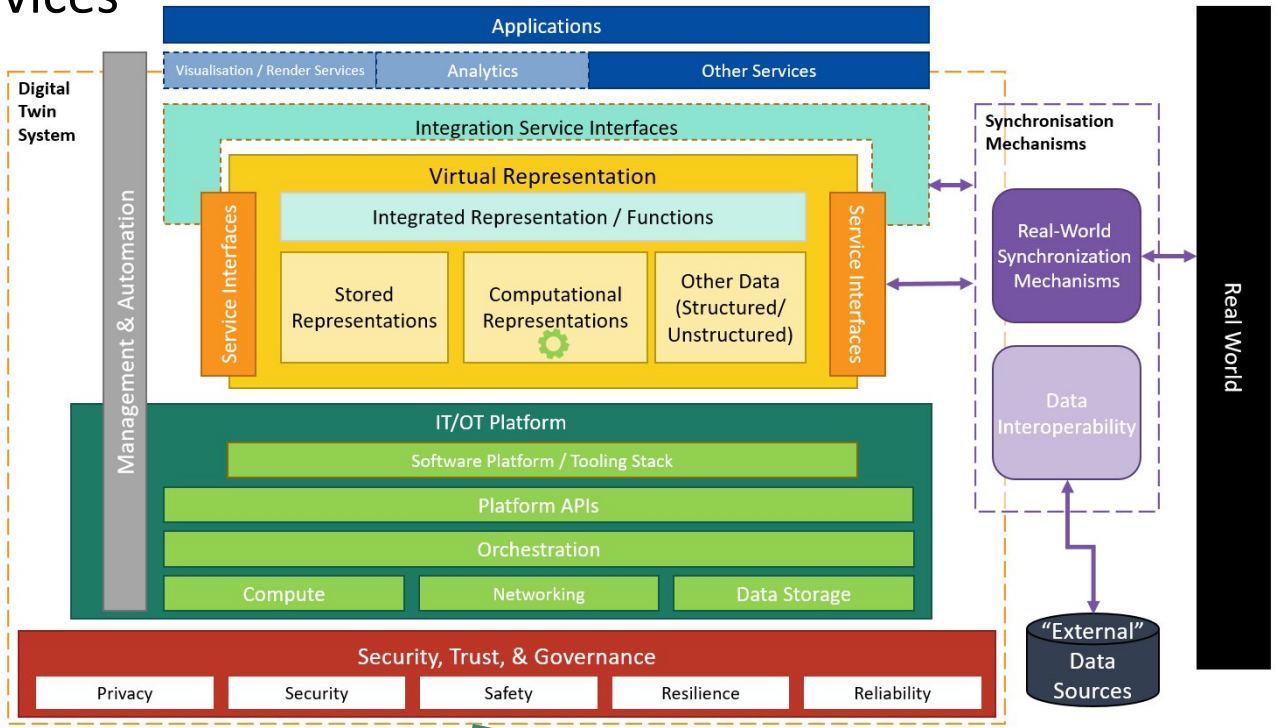


- HPC/queue for efficiency
- Cloud interfaces for interoperability
- Automated configuration and provisioning
- Support for workflow deployment
- Scientific software and AI stack
  (EESSI, CVMFS, registries and repositories)



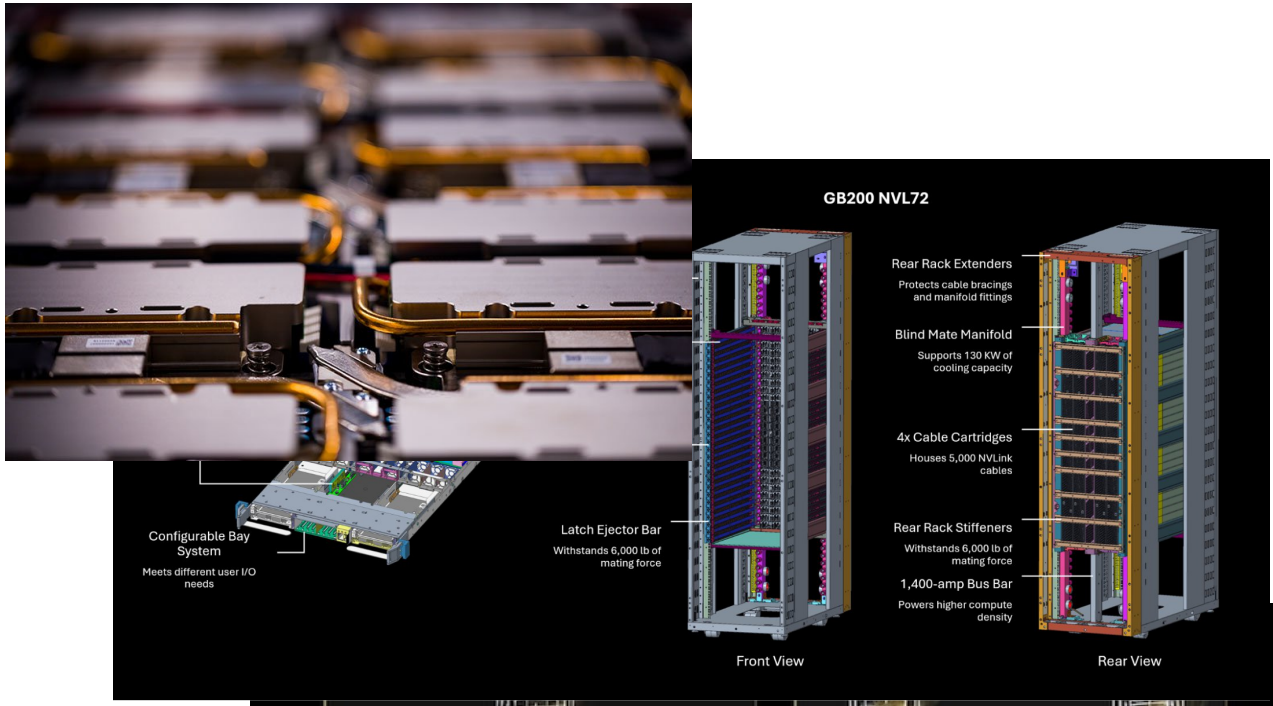
Flexibility of AI-HPC Infrastructure

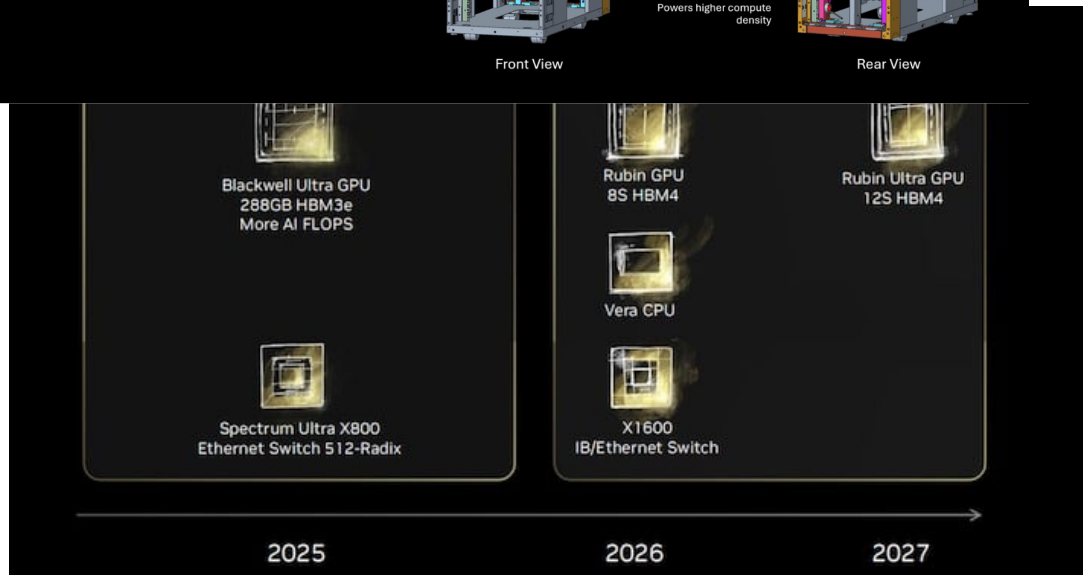
- Agile hybrid and modular architecture
  - transparent resource allocation for computational tasks, interactive work, development and services
  - system and environment configuration on user demand
- Usage modes:
  - classic HPC, containers, virtualisation
  - user-driven service environment
- Unified data access
  - direct (posix)
  - remote (https/davs, xrootd, S3 protocols ...)
  - federated access to (AI) datasets, central catalogue
  - transparent data sharing in federated data centres
- Services for workflow and data orchestration
  - Fully automated processing chains



Hardware Options

- CPU: 300-400k cores
  - AMD (Venice)
  - ARM64 (Grace, RHEA-2, Vera)
- GPU: 2-3k cards
  - NVidia GB200/300, Rubin
  - AMD MI450x
- Storage:
  - High throughput: Weka, VAST, DDN Infinia...
  - Large capacity: Ceph (FS, S3,...)
- Interconnect:
  - Infiniband, BX3, Slingshot - 800Gb/s, 1.6Gb/s
  - NVLink6/7 - 4/8/72





Questions?

With thanks to,
among many:

Andrej Filipčič

Barbara Krašovec

Damjan Harisch

Dejan Lesjak

Dejan Valh

Dragi Kocev

Iztok Lebar Bajec

Panče Panov

Uroš Lotrič

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PREOBRAZBO

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA VISOKO ŠOLSTVO ZNANOST IN INOVACIJE

Funding

	Cost [MEUR]
HPC (CAPEX + OPEX)	*123.3*
AI Factory	*11.7*
HPC + AI Factory	***135.0***
HPC (IT4LIA)	***10.0***
Total	**145.0**

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PREOBRAZBO

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA VISOKO ŠOLSTVO ZNANOST IN INOVACIJE