

EPI Forum

Barcelona, 9-10.10.2024.





Enabling AI Nations

Rod Evans, EMEA VP – Supercomputing, HER & AI



Le Monde

'Artificial intelligence is having an 'iPhone moment' and will shake up society the way Apple did in 2007'

Far from mocking the blunders of ChatGPT or Bard, Wall Street understands that a revolution is taking place, writes New York correspondent Arnaud Leparmentier.

Published on March 4, 2023, at 3:00 am (Paris) | ⌚ 2 min. • [Lire en français](#)

ChatGPT is the “iPhone Moment for AI”

Nations must act to protect their sovereignty and stay competitive



FORBES > INNOVATION > ENTERPRISE TECH

The New Global AI Arms Race: How Nations Must Compete On Artificial Intelligence

Bernard Marr Contributor ①

Forbes

“The overall objective of the strategy is to use artificial intelligence for economic growth, employment and a better quality of life..”

Dr. Mihailo Jovanovic, eGovernment Leader, Serbia

“AI isn't an option for countries.
It's a must for development.”

Ali Taha Koc,
Head of Digital Transformation Office, Turkiye

Every nation must become an AI Nation

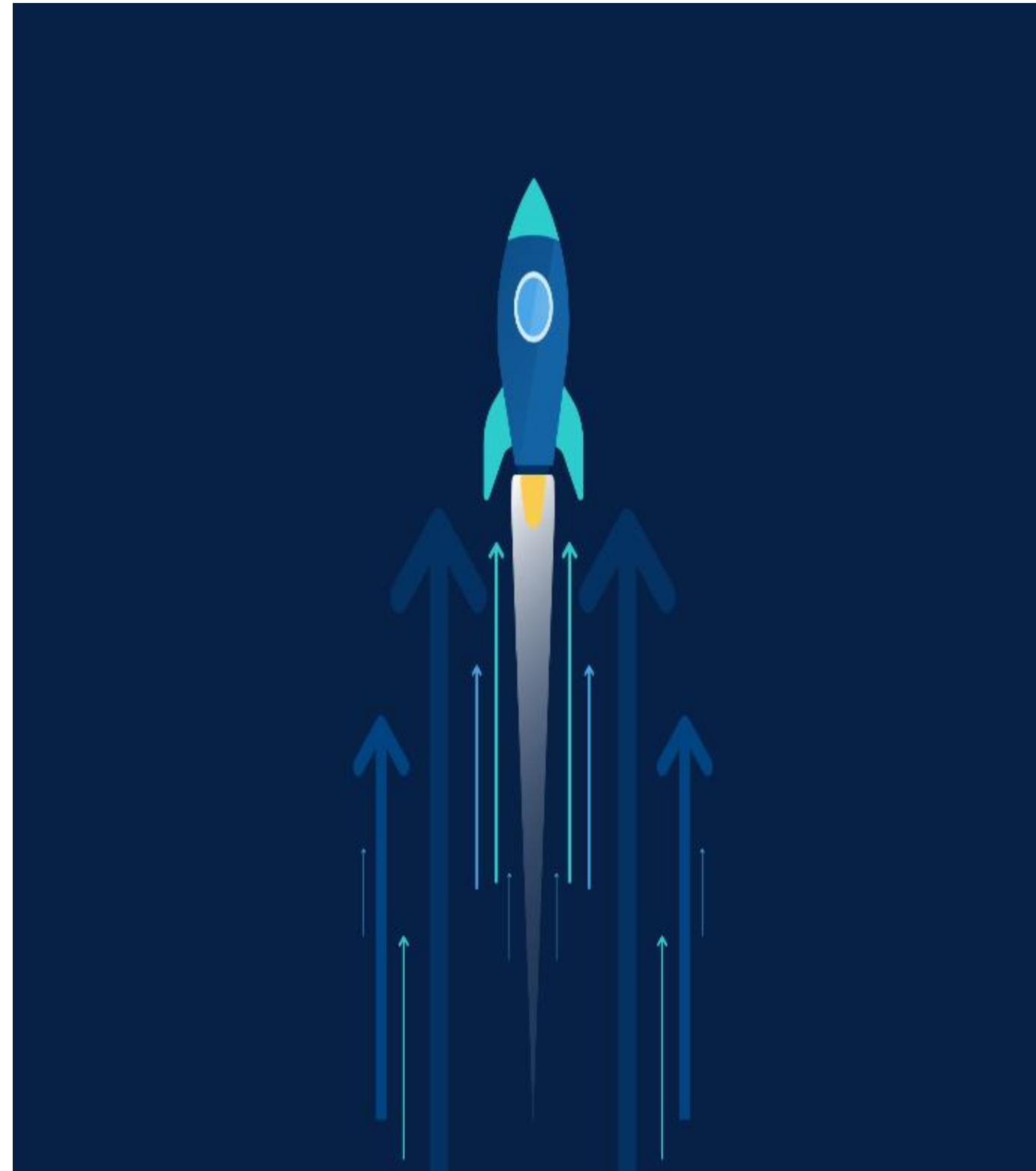
Nations must act to protect their sovereignty and stay competitive



First-ever OECD blueprint linking compute to economic growth

Compute capacity is now essential public infrastructure

OECD sees a growing “compute divide” across and within nations



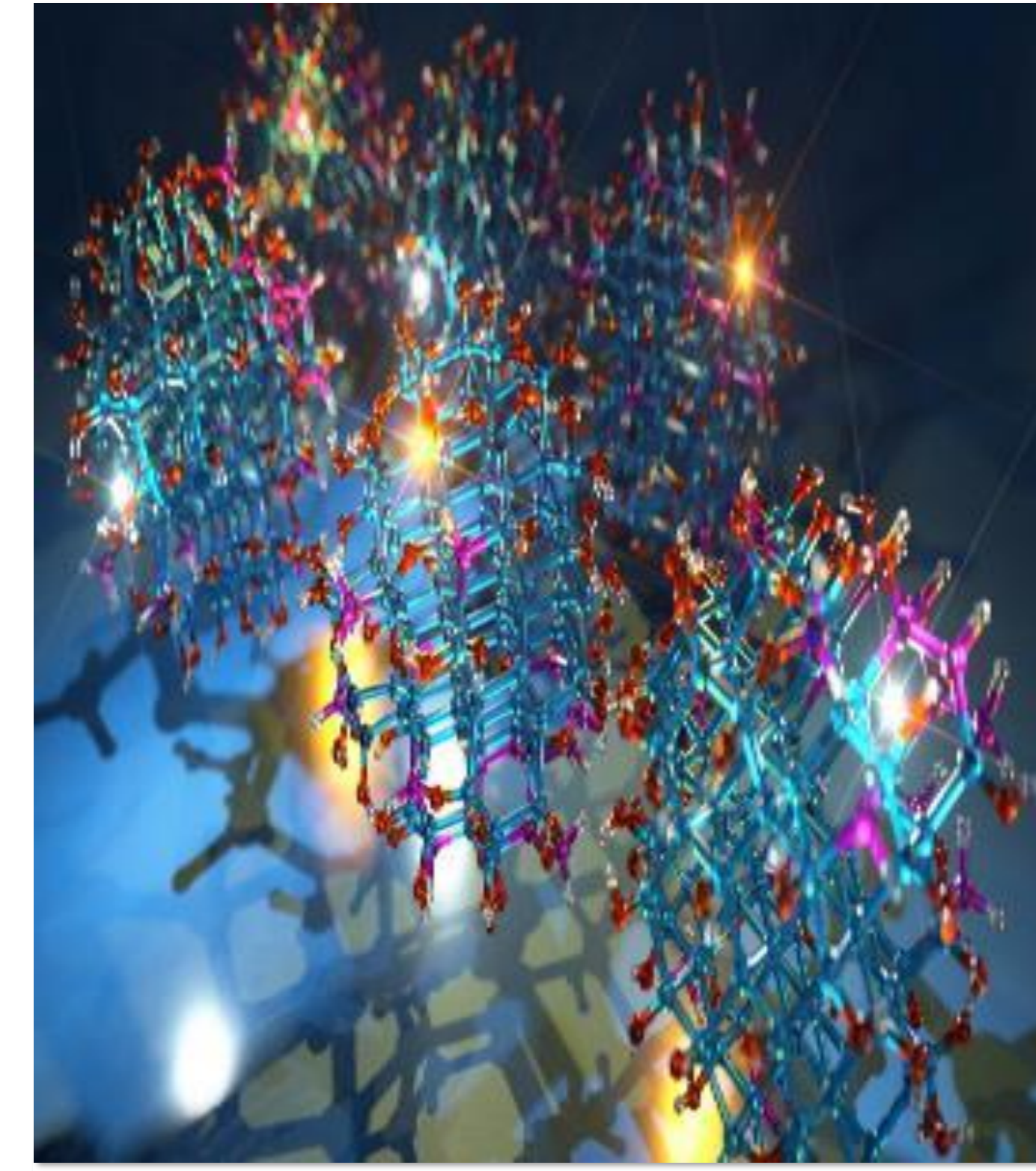
INNOVATION

Unleashes innovation across industries, start-ups, academia and Ministries



PRODUCTIVITY

Helps reinvent work, create new jobs, bring a new competitive advantage



SCIENCE

Transforms scientific research, drug discovery, genomics, & clinician productivity



CULTURE

Ensures that local language, culture, history and values are protected and represented

Generative AI is unleashing innovation across economies

For many nations, developing sovereign foundation models is a new priority AI initiative

NVIDIA's Role in Building AI Nations

- NVIDIA offers its expertise in AI infrastructure to countries looking to harness AI
- The AI Nations initiative involves helping countries develop an array of capacity-building programs targeting talent development and upskilling, entrepreneurship, and workforce-readiness.
- Collaboration in Building National Language LLMs
- **Objective – Enable Data & Digital Sovereignty**



Sovereign AI Narrative

NVIDIA LEADING THE DIALOGUE

NVIDIA CEO: Every Country Needs Sovereign AI

Jensen Huang describes transformative potential of sovereign AI at World Governments Summit.
February 12, 2024 by [Brian Caulfield](#)



Get ready for the age of sovereign AI | Jensen Huang interview



Nvidia CEO Jensen Huang and PM Modi discussed India's tech potential, emphasizing AI during a meeting with top U.S. tech leaders.



Nvidia CEO Jensen Huang with Indian PM Narendra Modi

Nvidia CEO Huang says countries must build sovereign AI infrastructure

By Reuters
February 11, 2024 10:58 PM PST · Updated 3 months ago



NVIDIA's CEO Jensen Huang attend a session of the World Governments Summit, in Dubai, United Arab Emirates, February 12, 2024. REUTERS/Amr Alfiky [Purchase Licensing Rights](#)



Nvidia chief urges 'sovereign AI'

By Emma W. Thorne, Editor at LinkedIn News

Updated 2 months ago

Share

Every country needs to build and control its own artificial intelligence infrastructure, Nvidia CEO Jensen Huang **said Monday**. This "**sovereign AI**" would allow for both cultural and economic benefits, he said at a summit in Dubai, adding that he thinks the broader dangers of AI are overblown. Huang said his company — which **currently dominates** the AI chip market — is "democratizing" access to AI, but "the rest of it is really up to you to take initiative, activate your industry, build the infrastructure."

Saudi Arabia accelerates digital economy growth through Nvidia partnership



The Kingdom's Minister of Communications and Information Technology Abdullah Al-Swaha's meeting with Nvidia CEO Jensen Huang aims to support and strengthen the region's digital economy. Supplied

Building an AI Nation

National Policy Imperatives

Economic Growth

Scientific Discovery

Urban Renewal

National Security

Citizen Services

INFRASTRUCTURE FOR DOMESTIC AI PRODUCTION + CONSUMPTION

HUMAN + COMPUTE

PUBLIC BENEFIT / OUTCOMES

Competitive industries
Home-grown innovation
Robust startup community

Vaccine development
New materials
Climate prediction

Efficient public transit
Resilient housing
Increased livability

Secure borders
Early threat detection
Energy and food security

Digitised public services
Up-skilled workforce
Public data exchanges

AI Initiatives

- Generative AI / LLMS
- Digital Twins + Simulation
- Climate Science + Resilience
- Autonomous Systems + Machines
- National Security + Cyber defense
- Public Health / Life Sciences

AI Workforce

- Advanced AI upskilling
- Quantum Developer Curriculum
- Jetson Edge AI Curriculum
- Developer Hackathons
- Developer Bootcamps

AI Ecosystems

- Start-ups
- Global System Integrators
- Solution Development Partners
- Investors, VCs
- Independent Software Vendors
- Universities + Research Institutes
- National Supercomputing Centers
- NGOs

NVIDIA AI Nations Next Partnerships are Democratizing AI Across the World

NVIDIA Helps Drive AI Adoption and Research in Thailand



BANGKOK — Dec. 3, 2020 — NVIDIA today signed a memorandum of understanding (MoU) with a consortium of universities in Thailand to drive research and accelerate scientific breakthroughs in artificial intelligence (AI) and high performance computing (HPC).

Nvidia accelerates Australia's AI roadmap with CSIRO partnership

The chip maker's technology will be used to create digital twins and trial quantum computing initiatives



Italy Forges AI Future in Partnership with NVIDIA

Collaboration begins with a research hub at AlmageLab in Modena.

January 15, 2020 by [FREDERIC PARENTTE](#)



NVIDIA Partners with NHS Trusts to Deploy AI Platform in UK Hospitals

November 28, 2022

Nov. 28, 2022 — A consortium of 10 National Health Service Trusts — the publicly funded healthcare system in England — is now deploying the MONAI-based AIDE platform across four of its hospitals, providing AI-enabled disease-detection tools to healthcare professionals serving 5 million patients a year.

AIDE, short for AI Deployment Engine, is expected to be rolled out next year across 11 NHS hospitals serving 18 million patients, bringing AI capabilities to clinicians. It's built on [MONAI](#), an open-source medical imaging AI framework co-developed by [NVIDIA](#) and the AI Centre, which allows AI applications to interface with hospital systems.



Value added capabilities that advance National AI programs

NVIDIA's full-stack collaboration approach is unique within the global tech industry

Context for Change to AI

- **Generative AI** and **AI foundation models** are advancing at unprecedented pace and are set to play a pivotal role in shaping the future of technology and society.
- **AI** is a key policy area of the **EU digital strategy**. AI in combination with **HPC** can contribute to a more innovative, efficient, sustainable and competitive economy, while also improving safety, education and healthcare for citizens.



- In her 2023 State of the Union address, President von der Leyen announced that **the supercomputing resources of the EuroHPC JU will be made available to European AI startups to train their large-scale models**, contributing to the EU's aim of leading global advances in AI and of achieving responsible and ethical innovation.

Why the Focus change to AI Factories?

- Recognition that Supercomputer centers deliver key services for Scientific research and are a good way to consolidate resources with structure and process
 - User support and service delivery
 - High degree of knowledge on building and managing large systems
 - They could assist National AI Strategies by providing guidance on missing elements for implementation in HPC environments, parallelisation techniques etc

BUT....

- Many Countries have cancelled Post-Exascale projects focused solely on HPC
- AI innovation largely happening outside Academic research
 - >50% of French “Jean Zay 2” platform is now from AI Start-ups
 - iGenius – working on Italian LLM with Cineca now building their own AI Factory
- Wide scale belief in most National circles that Exascale is done. What’s next?
- Perception that HPC is hard – limited new adoption outside fundamental research
 - Impact on wider economy of HPC adoption is seen as narrow and limited
- AI has demonstrated high value to improve and augment numerical simulations
- Stagnating economies – Government’s focus on growth
- Access to scalable resources is fundamental for the development of National AI plan and ecosystem

AI FACTORY IS THE NEW CRITICAL INFRASTRUCTURE

INSTRUMENT FOR **SCIENTIFIC DISCOVERY**

- Human condition
- Fundamental research
- Industrial innovation

ENGINE FOR **ECONOMIC GROWTH**

- Start-up ecosystem
- New jobs/sectors
- Retain talent; Workforce productivity

PLATFORM FOR **PUBLIC SECTOR INNOVATION**

- Improve access to services
- Expand citizen services
- Defend national interests

NVIDIA AI Nations NEXT FRAMEWORK

Full-Stack Collaboration Approach

National AI Program

AI Initiatives

NVIDIA helps nations advance AI R&D workloads and applied use-cases across every industry

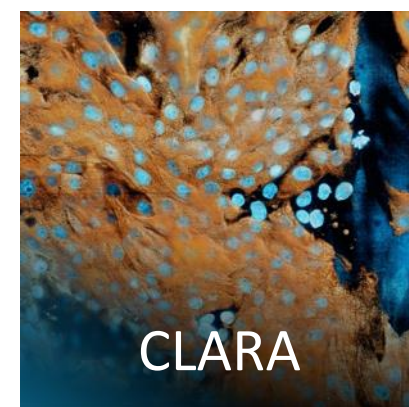
AI Workforce

NVIDIA helps nations upskill local talent and develop the AI-ready workforce

AI Ecosystems

NVIDIA helps nations strengthen their local AI ecosystem and learn from global leaders

NVIDIA AI Enterprise



Medical Imaging



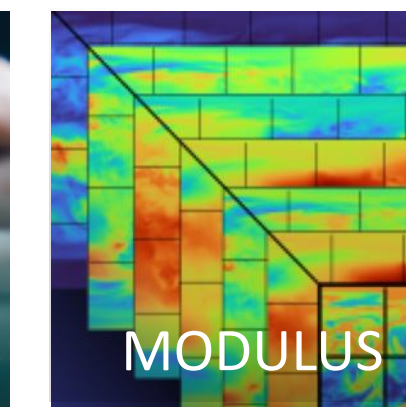
Speech AI



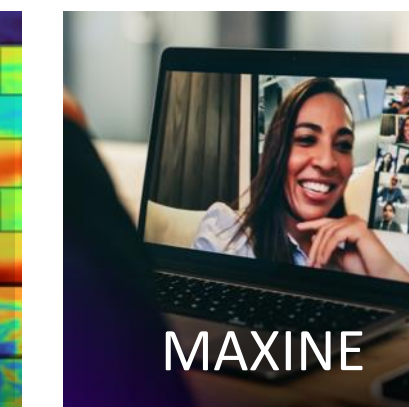
Customer Service



Recommenders



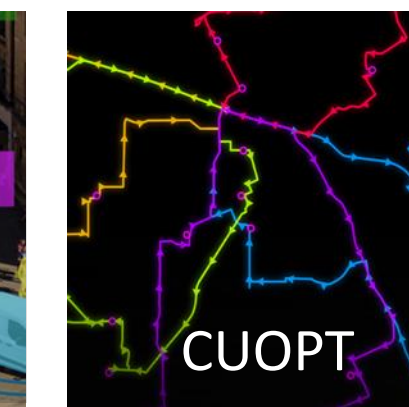
Physics ML



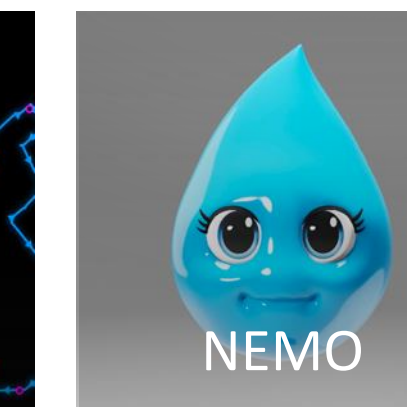
Video



Video Analytics



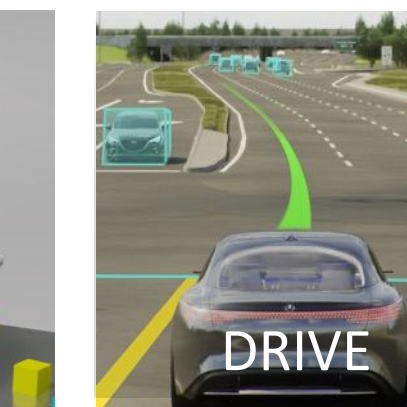
Logistics



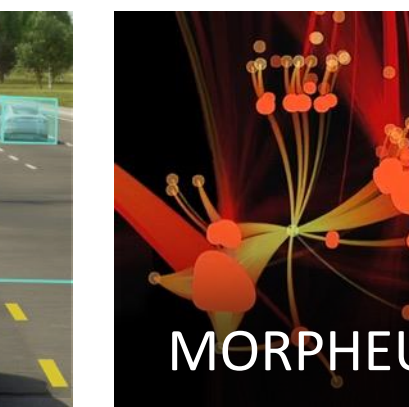
Conversational AI



Robotics



Autonomous Vehicles



Cybersecurity

Inception for Start-ups, Deep Learning Institute, Hackathons/Bootcamps, GTC

NV PS, Industry/Domain Experts, HPC + Technical Support, AI Tech Center Research Collaborations

National AI Infrastructure



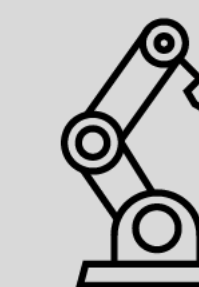
Cloud



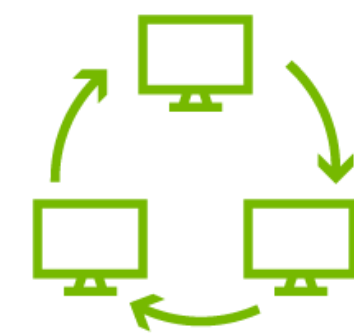
Data Center



Edge



Embedded



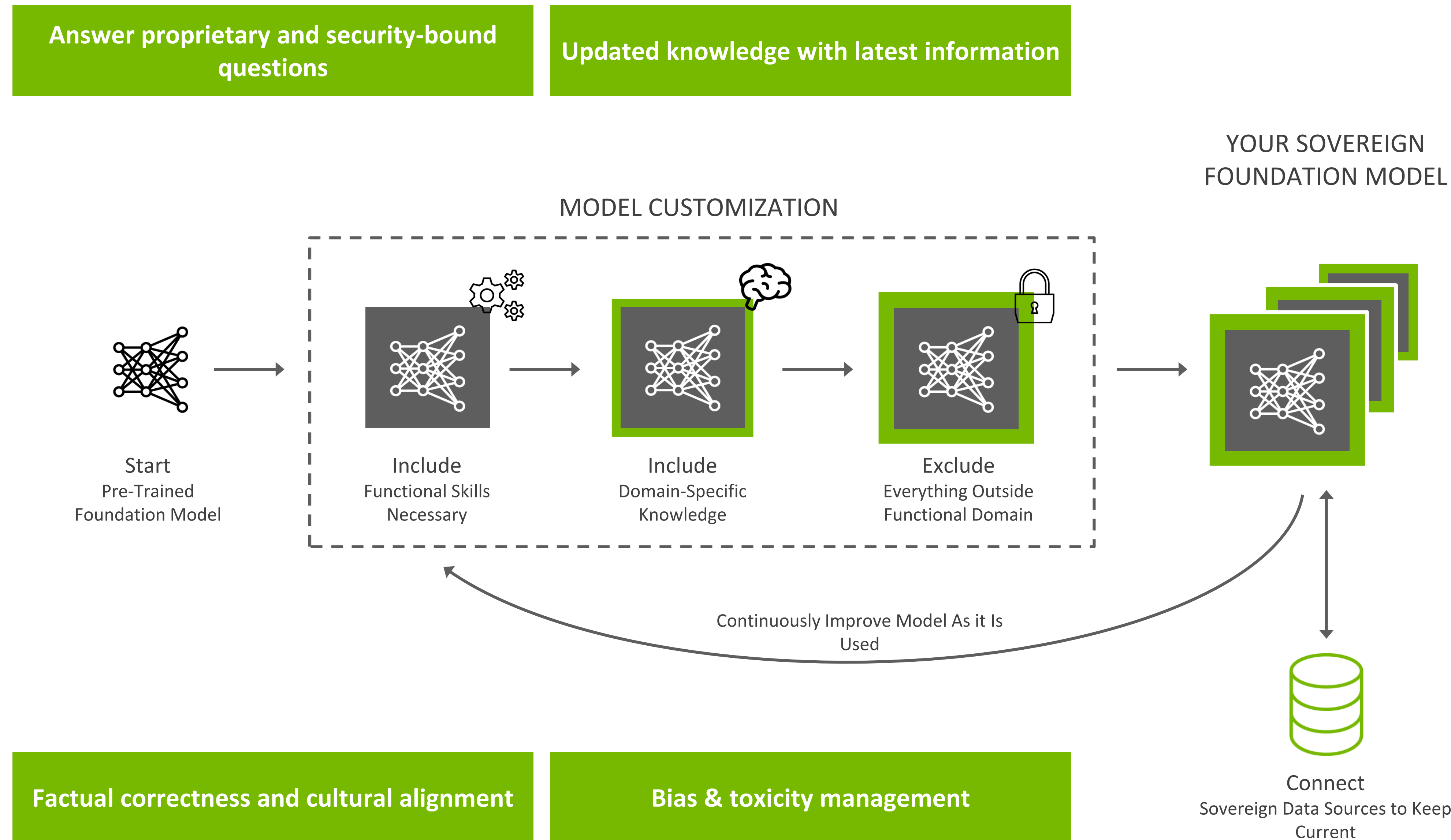
Sovereignty



Sustainability



Safety



Multi-Cloud. On-Prem. Run Anywhere.

Nations can leverage NVIDIA Nemo (LLM) in public clouds or sovereign cloud/data centers, including a National Super Computing Center or Domestic Commercial Data Center Provider.

Public Clouds

Microsoft Azure

aws

ORACLE®

Sovereign Cloud/DC

NVIDIA DGX SuperPODs

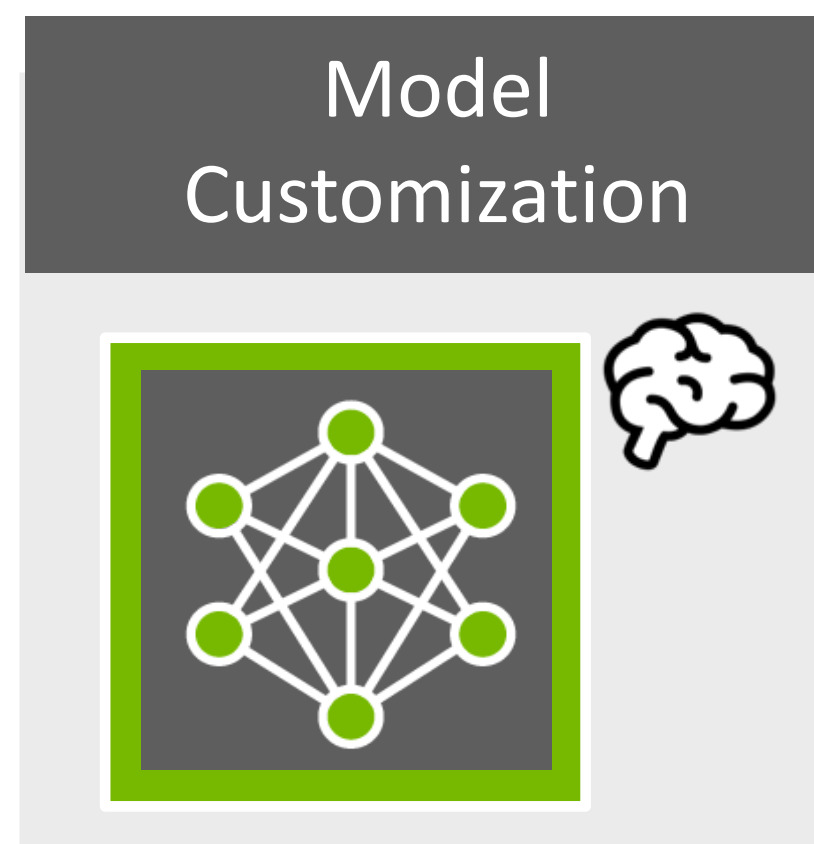
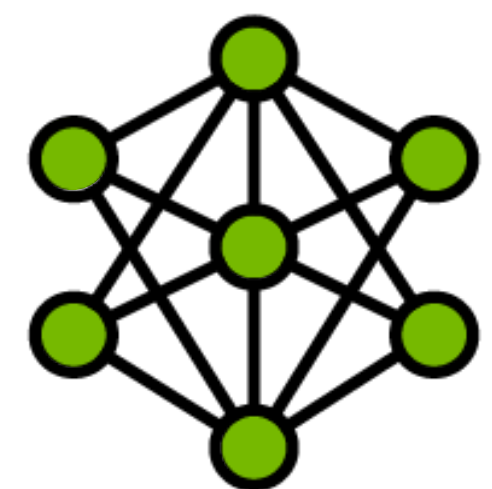
Helping nations customize + deploy sovereign foundation models

NVIDIA Nemo is an end-to-end, open-source approach to training and deploying LLMs with billions of parameters

Sovereign Foundation Models

Continual pre-training and finetuning foundation models to absorb national culture and values

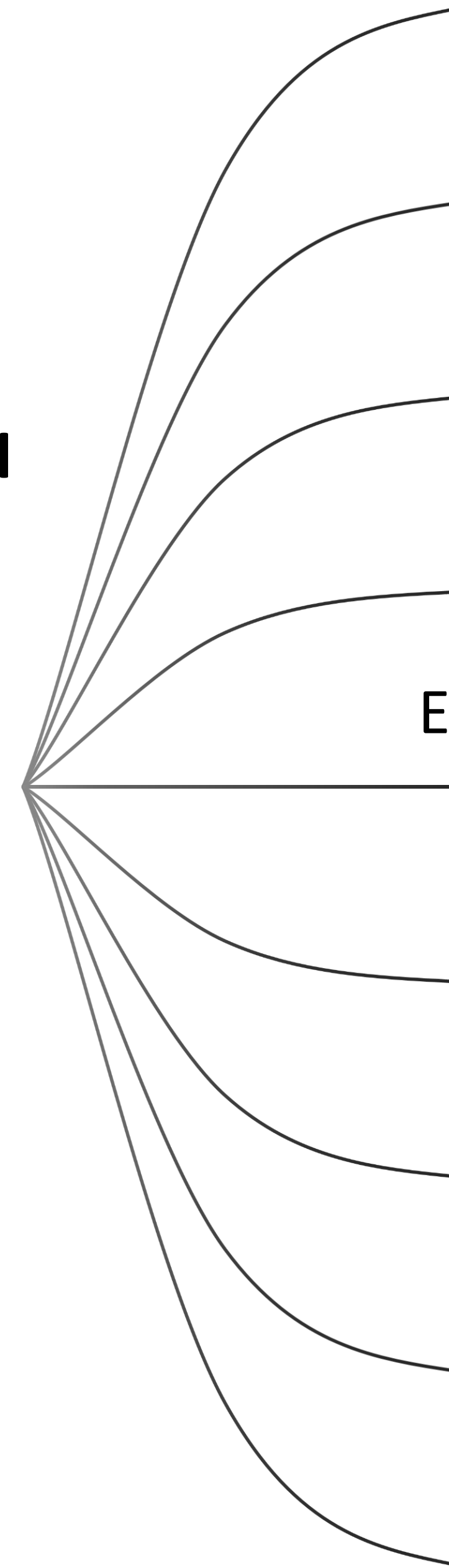
Foundation Model



Continual pre-training
Vocabulary Expansion



Sovereign Model



Personalized Citizen Services



Education and Skills Development



Regulatory Compliance and Enforcement



Urban Planning and Smart Cities

NVIDIA Launches NIM Microservices for Generative AI in Japan, Taiwan

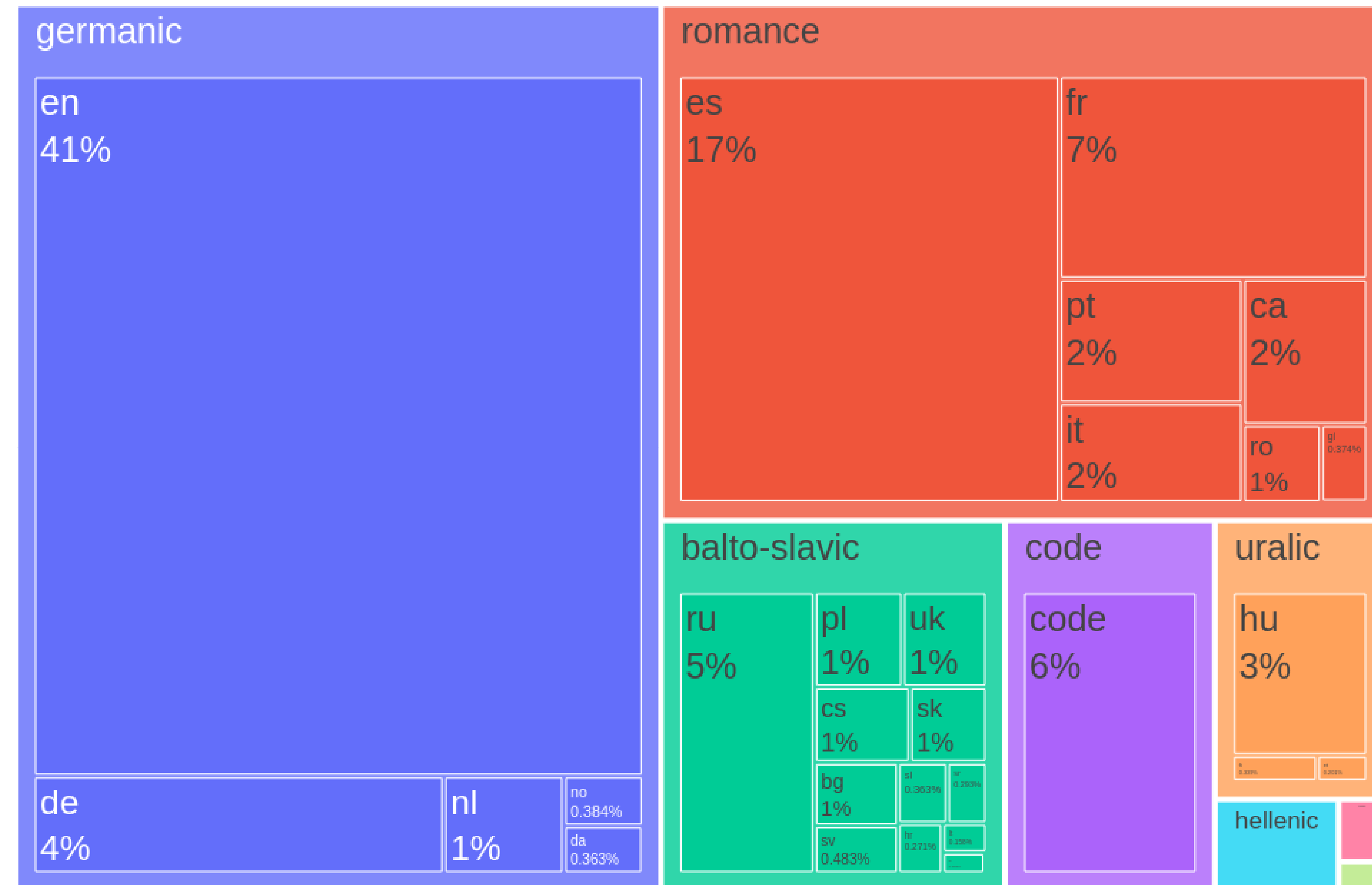
Four new microservices help accelerate deployment of sovereign AI applications that offer advanced cultural and language fluency.

August 26, 2024 by [Karl Briski](#)



An Example of a Sovereign Foundation Model: Salamandra 2B, 7B & 40B Models

NeMo Framework 🤝 MareNostrum5 at BSC



- **Training Framework:**

Pre-training was conducted using NVIDIA's NeMo Framework, which leverages PyTorch Lightning for efficient model training in highly distributed settings.

- **Compute Infrastructure:**

All models were trained on MareNostrum 5, a pre-exascale EuroHPC supercomputer hosted and operated by Barcelona Supercomputing Center.

- The accelerated partition is composed of 1,120 nodes with the following specifications:
 - 4x Nvidia Hopper GPUs with 64GB HBM2 memory
 - 2x Intel Sapphire Rapids 8460Y+ at 2.3Ghz and 32c each (64 cores)
 - 4x NDR200 (BW per node 800Gb/s)
 - 512 GB of Main memory (DDR5)
 - 460GB on NVMe storage

SOVEREIGN AI

Domestic capacity to produce AI at scale



👁️ Sovereign AI is critical to economic resilience + national security

SOVEREIGN AI

👁️ Ecosystem of investors, developers, scientists, entrepreneurs, creators, Ministries, customers

LOCAL ECOSYSTEM

👁️ Local talent trained from basic AI awareness to specialized skills to deliver AI

AI-READY WORKFORCE

👁️ Models that are trained and/or fine-tuned on local data, models hosted and run on local infrastructure

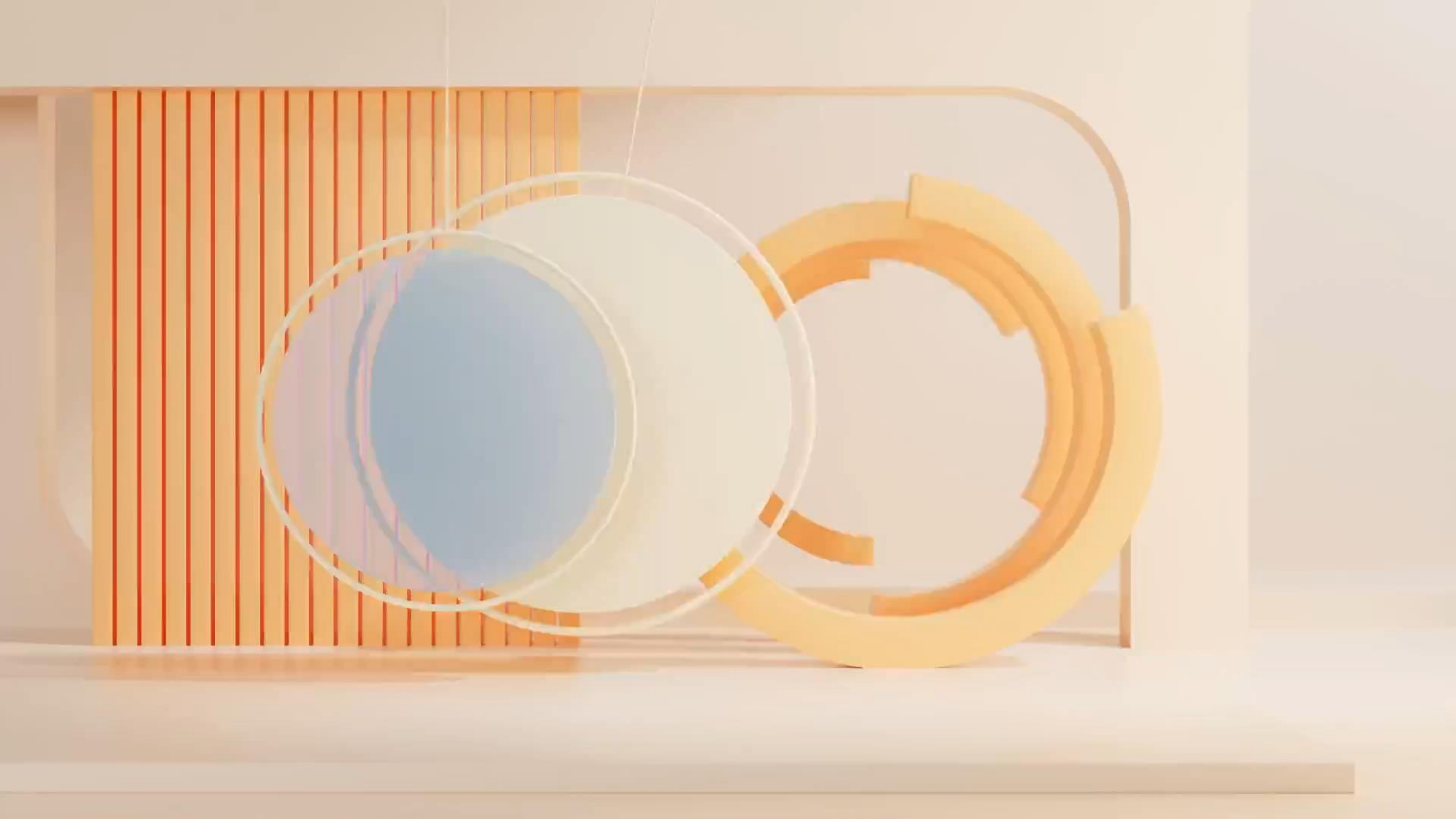
SOVEREIGN AI MODELS

👁️ Sovereign AI compute capacity must be owned + operated locally

LOCALLY OWNED COMPUTE INFRASTRUCTURE

👁️ EPI – Leading Europe's Technology future

INTERGRATION OF EUROPEAN TECHNOLOGY



France:

GENCI/Jean Zay 4
National AI Platform

Spain:

BSC
LLM & Ai
Collaboration

Germany:

ALL 10 AI centres
standardizing on
NVIDIA -

Serbia:

National AI Alliance
since 2020

EUROPEAN NATIONAL AI PROJECT Collaborations

Finland:

National AI
Technology Centre

Slovenia:

National AI Alliance
since 2021

Italy:

AI Technology Centre
with CINI & Cineca

Commercial/ National AI:

Novo Nordisk
Foundation (Denmark)
WASP/Linkoping
University (Sweden)

Example of National AI Focus – Germany:

- Accelerate AI Transformation with Shared AI Infrastructure

Working with Governments to “*cross the cavern*” with AI

- Germany – almost every University is State owned – either by Federal or Regional Lander
- Federal Government provides funding through “BMBF” – Federal Ministry of Education and Research
- Specific Programs for HPC & AI
- AI Plan Action Plan launched in November 2023
- Consortiums of Universities delivering AI Research across Germany



Federal Ministry
of Education
and Research



Bettina Stark-Watzinger
Member of the German Bundestag
Federal Minister of Education and
Research

Driving AI through 10 Centres of Excellence

KI-Kompetenzzentren

- 6 x Machine Learning/ National AI Competency Centres

KI-Servicezentren

- 4 x National AI Services Centres



Foundations & GEN AI models

Public Administration

SAVIA

Project : Chatbot to support legal professionals, lawmakers, and citizens in accessing information on the current and past legislation of the Emilia-Romagna region in Italy.

Who : Assemblea Legislativa Emilia Romagna & CINECA

Type: LLM – GenAI (FT + RAG)

Pre-trained LLMs : llama2-13B, Mistral-7B, Mixtral-8x7B-Instruct-v0.1
(
3-400B

Infrastructure Leonardo – EuroHPC/CINECA

Training : Max 1 node – 4 NVIDIA A100 SXM 64 GB – 400 GPUh – 4 days (FT/RAG)

Inference Live demo Oct'24 <https://assemblealegislativa.github.io/savia/>

Support from NVAITC Italy

MarlAnne

Project : Build, train LLM for information sensitive and highly regulated industries; 1st model fully compliant with the AI act as it is taught exclusively on content with a permissive license, multilingual and focused on European open data .

For public services and large companies concerned with compliance & ethics as environmental impact.

Who : Pleias & Sciences Po – French Government support

Type: LLM – FM

Infrastructure Jean Zay – GENCI/IDRIS

Training : NVIDIA H100 SXM 80 GB

Frameworks : Nemo-Megatron & Nanotron

Inference: next step

Support from NVIDIA DevTechs

AI 4 Science in Healthcare

Domain: Healthcare / Biotechnology

Project : Predict a more accurate patient's response to a therapy for digital pathology, agnostic to experimental environment by removing the impact of the usage of different instruments or prep conditions used for digital images scans – Improve FM robustness.

Who: Owkin startup

Type: GenAI

How to: Learn through GAN to encode an image and output 2 images by separating what relates to the structure of the image (sematic content) and its texture (color, grain, etc.) - Mostly public (TGSA)

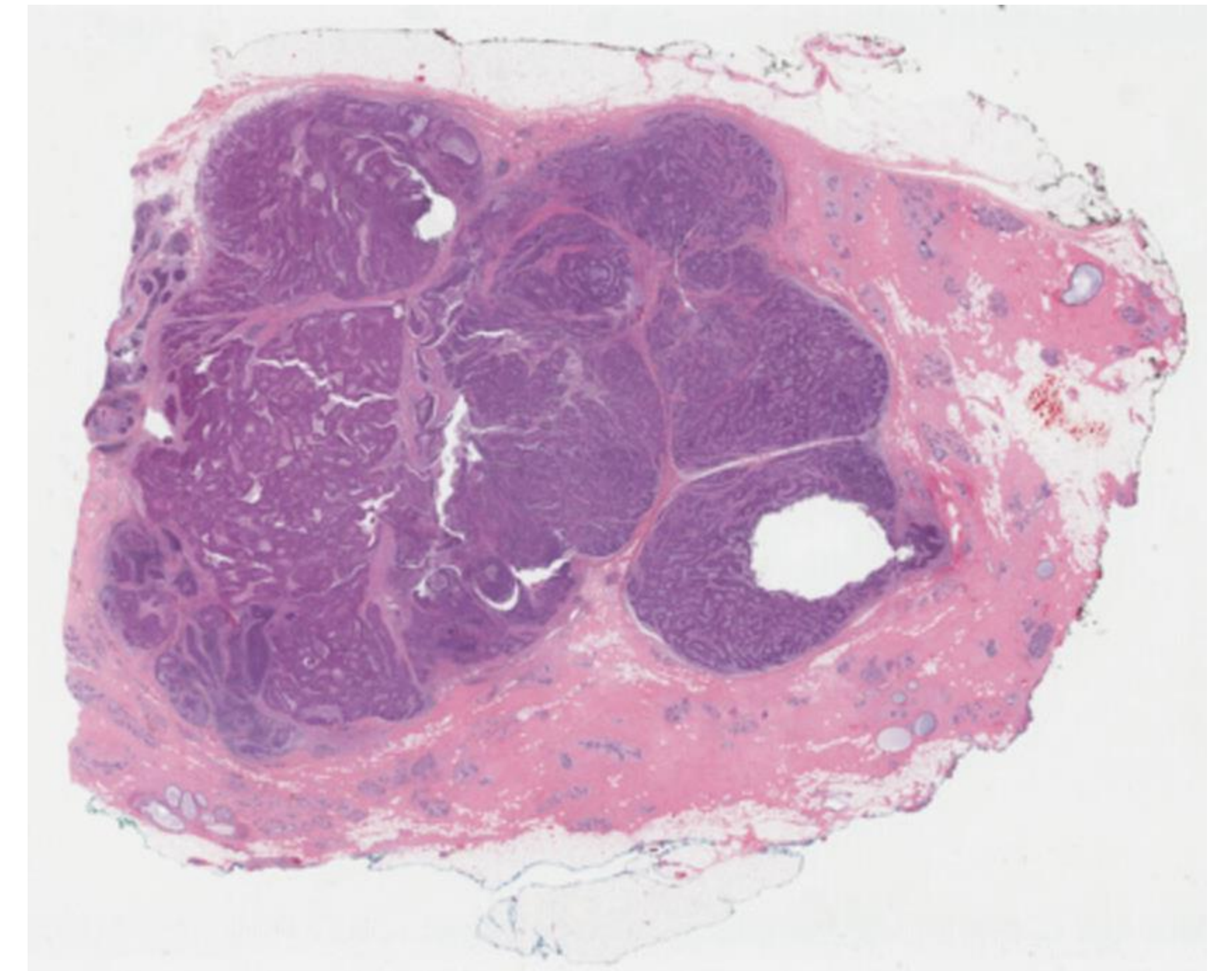
Infrastructure : Jean Zay – GENCI/IDRIS

Training: NVIDIA H100 SXM 80 GB

Frameworks : PyTorch - Step 1: Model : StyleGAN2/3 ; step 2: Stable diffusion implementation to increase the fidelity of input image and the realism of a synthesized image (increase FM robustness)

100% open-source from model to be released.

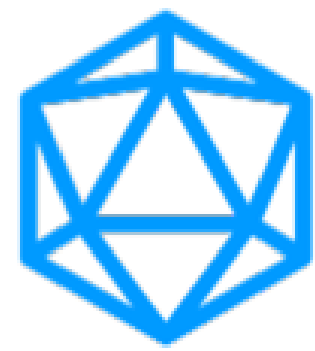
Support from NVIDIA DevTechs, Healthcare Specialists & LLM Solution Architects



A glowing 3D GPU chip is centered on a blue circuit board background. The chip is a square with a metallic, reflective surface that glows with a warm orange light. It is surrounded by a complex network of yellow circuit traces that branch out across the blue background. The background is also filled with faint, glowing binary code (0s and 1s) and hexadecimal characters (A-F, 0-9) in a light blue color, creating a digital atmosphere. The overall composition is dynamic and futuristic, emphasizing high-tech computing.

NVIDIA & EPI

WORKING TO BUILD EUROPE'S NEXT SUPERCOMPUTERS



SIPEARL

partnership – bringing GPU’s to RHEA

The **A** Register



Home-grown Euro chipmaker SiPearl signs deal with HPE, Nvidia

Claims partnerships will drive development and adoption of exascale computing in Europe

 Dan Robinson

Mon 30 May 2022 // 14:58 UTC

European microprocessor designer SiPearl revealed deals with Nvidia and HPE today, saying they would up the development of high-performance compute (HPC) and exascale systems on the continent.

Announced to coincide with the [ISC 2022](#) High Performance conference in Hamburg this week, the agreements see [SiPearl](#) working with two big dogs in the HPC market: HPE is the owner of supercomputing pioneer Cray and Nvidia is a leader in GPU acceleration.

With HPE, SiPearl said it is working to jointly develop a supercomputer platform that combines HPE's technology and SiPearl's upcoming Rhea processor. Rhea is an Arm-based chip with RISC-V controllers, planned to appear in next-generation exascale computers.

The partnership will expand heterogeneous computing options for supercomputing, according to SiPearl, and is intended to support and accelerate adoption of exascale systems in Europe.



[Press release](#)

SiPearl Collaborates with NVIDIA on Enabling Accelerated Computing Solutions with European Microprocessor

SiPearl, the European company designing the high-performance and low-power microprocessor for exascale[®] supercomputing, and NVIDIA announce a collaboration to provide a joint offering combining SiPearl HPC microprocessors with NVIDIA accelerated computing and networking portfolio.

SEMICONDUCTOR **DIGEST**

NEWS AND INDUSTRY TRENDS

SEMICONDUCTORS

SiPearl Collaborates with NVIDIA on Enabling Accelerated Computing Solutions with European Microprocessor

SHANNON DAVIS • MAY 31, 2022



SiPearl, the company designing the high-performance, low-power microprocessor for European exascale supercomputers, has entered into a strategic collaboration agreement with NVIDIA for joint technical and business developments aiming to combine both companies' portfolio of hardware and software solutions.

NVIDIA and SiPearl will develop a proxy platform for porting activities and SVE workload analysis combining the strengths of SiPearl CPU (such as HBM memory) and NVIDIA GPU (including massive parallelism and throughput). The collaboration will include joint efforts with third-party European research institutions on elements such as SoC and NoC simulation capabilities in open-source and research-oriented modeling



NVIDIA®

SC23—NVIDIA today announced that JUPITER — which launches a new class of supercomputers for AI-driven scientific breakthroughs — will be powered by the NVIDIA Grace Hopper™ accelerated computing architecture to deliver extreme-scale computing power for AI and simulation workloads.

Hosted at the Forschungszentrum Jülich facility in Germany, JUPITER — which is owned by the EuroHPC Joint Undertaking and contracted to Eviden and ParTec — is being built in collaboration with NVIDIA, ParTec, Eviden and SiPearl to accelerate the creation of foundational AI models in climate and weather research, material science, or discovery, industrial engineering and quantum computing.



[Home](#) / [Dissemination Communication Press Repository](#) / [Home-grown Euro chipmaker SiPearl signs deal with HPE, Nvidia](#)

Home-grown Euro chipmaker SiPearl signs deal with HPE, Nvidia



Implementing our CPU in GPU accelerated nodes requires a solid development environment, domain specific libraries, development kits and HPC application tuning. NVIDIA investment and expertise across all of these areas is driving sustained growth of the ecosystem and the market

Philippe Notton, SiPearl founder and CEO



Advancing the State-of-the-Art in Compilers

NVIDIA is investing in open source and commercial compilers for ARM

- **NVHPC**

- Continuously improving performance
 - **23.3 release adds support for Neoverse V2 CPUs**

- **LLVM/Clang**

- NVIDIA will provide supported & optimized builds of LLVM/Clang as drop-in replacements for mainline
- NVIDIA contributing directly to LLVM to enhance Grace performance, especially in applications we prioritize
- Engaging with Arm (Inc.) to leverage their compiler team to improve key CPU benchmark performance e.g. SPEC

- **GCC**

- Engaging with Arm (Inc.) to leverage their compiler team
- Working with Arm to improve mainline GCC

NVHPC

- NVIDIA's innovation space
- Focus on value for workloads

LLVM / Clang

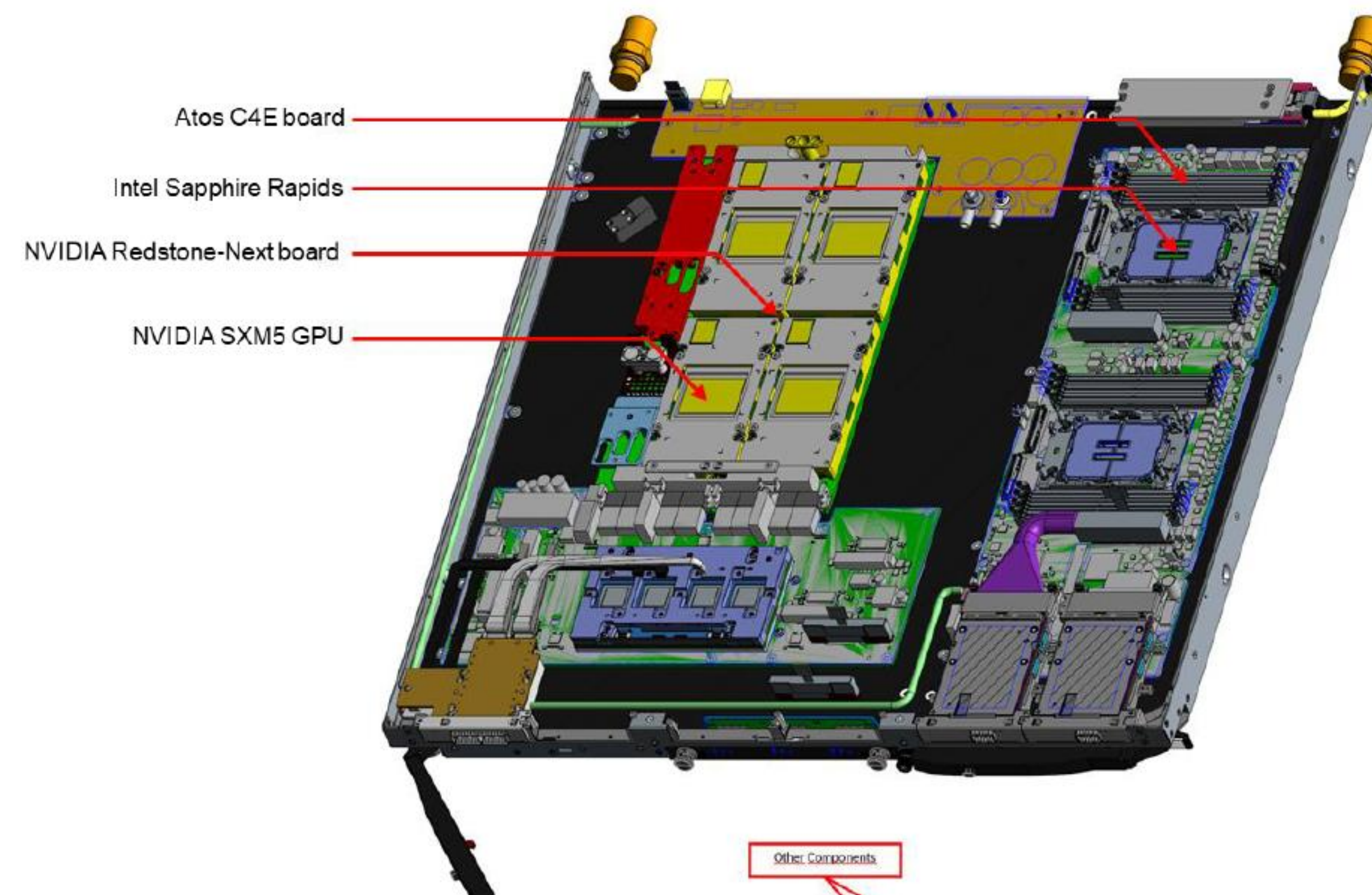
- Required for an excellent CPU software ecosystem
- A foundation for potential compiler innovations

GCC

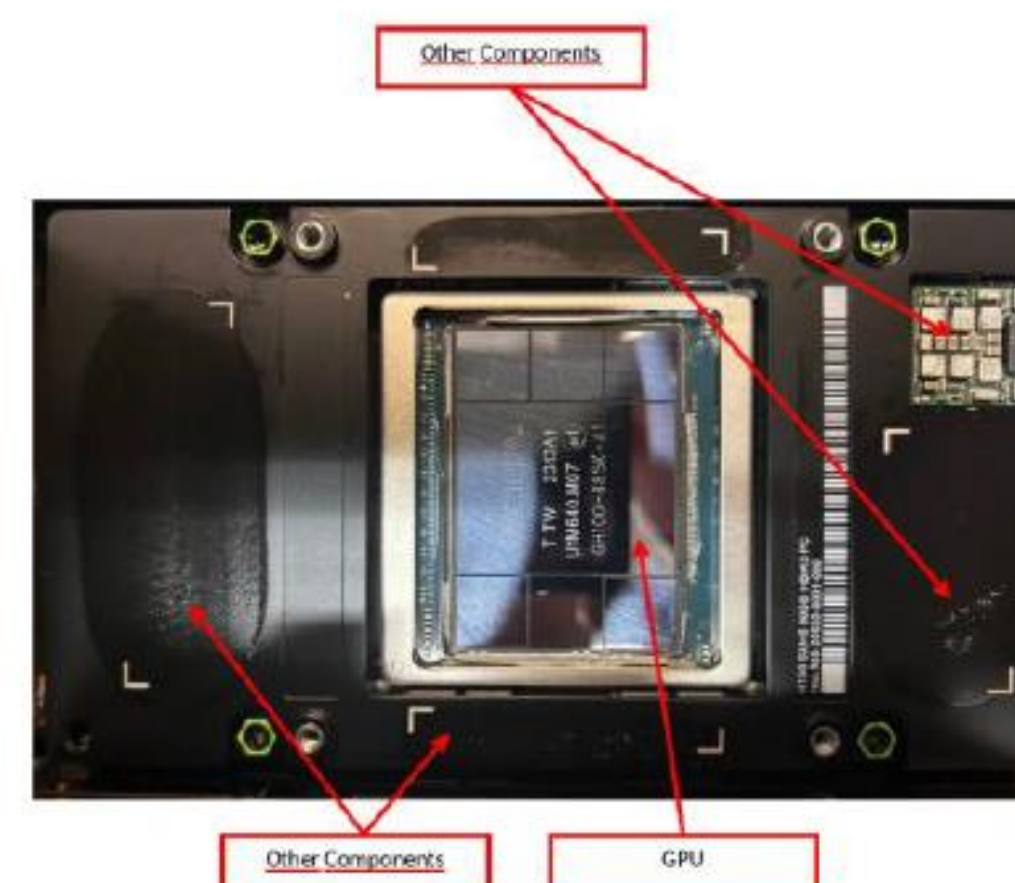
- Required for an excellent CPU software ecosystem
- Often the default; sets performance expectations

ARM + GPU Prototype

EVIDEN



GPU H100



IDV-A (Atos) based on 2-socket Intel Sapphire Rapids connected to 4 Nvidia Hopper GPUs

Thermal Requirements

CPU: Intel Sapphire Rapids

TDP: 350 W

N.: 2

GPU: Nvidia Hopper

TDP: 700 W

N.: 4

RAM Modules: 4 x 8Gb

VR: CPU Voltage Regulator

Thermal power: 150 W

Total thermal power: 3800 W

Maximum Water Temperature of the primary loop at rack level: 45°C

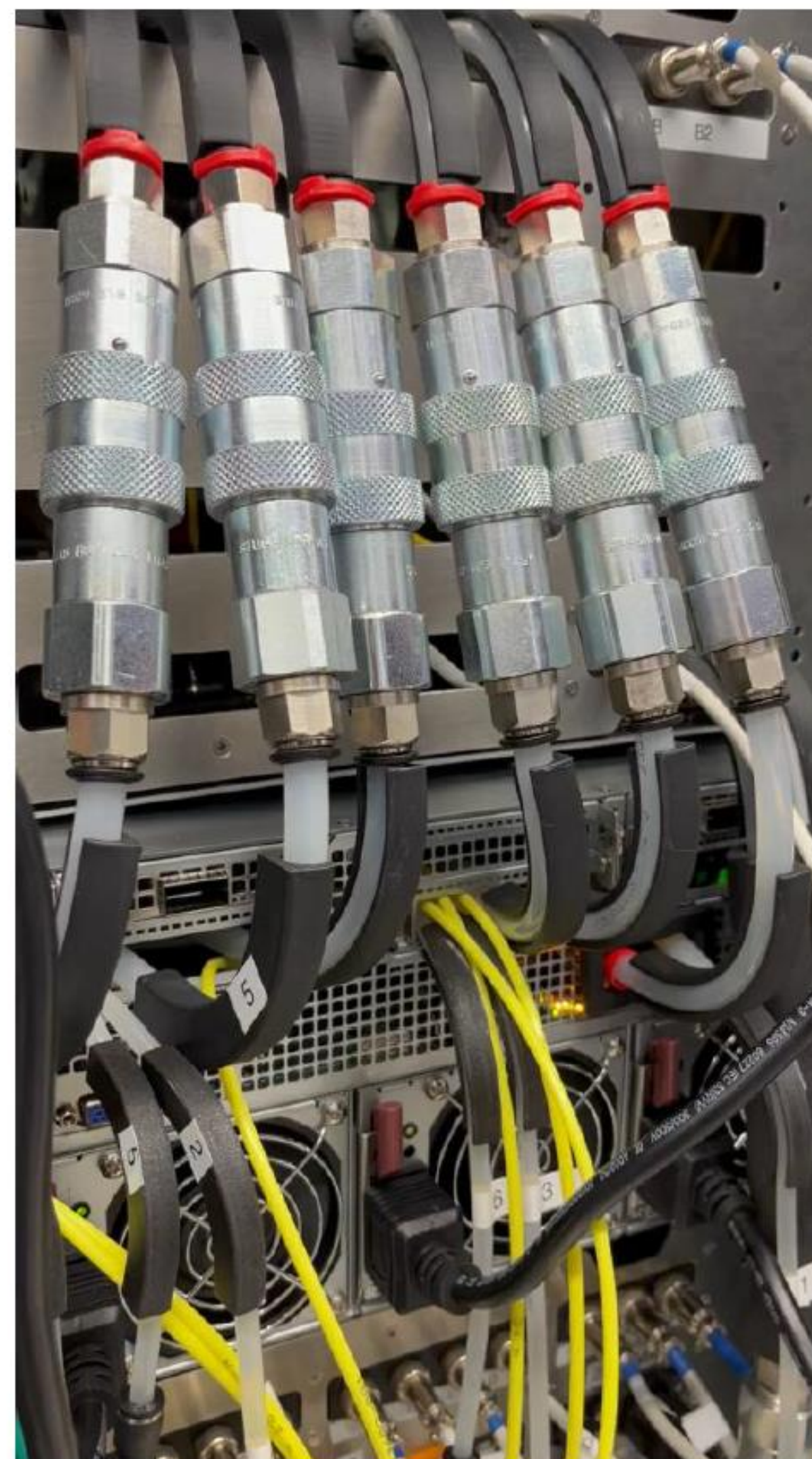
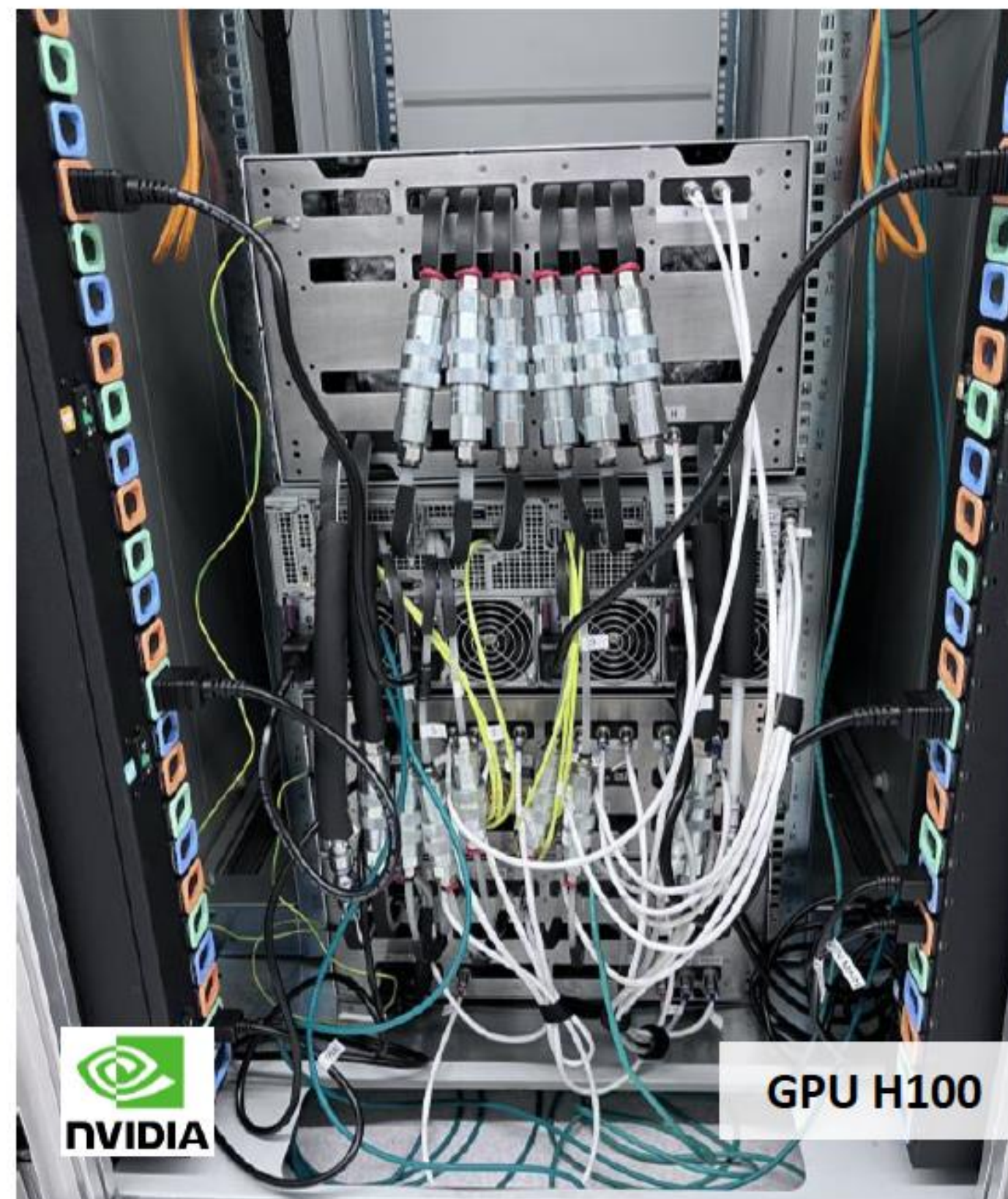


EP-Too Pilot

EPI-TO (University of Torino – CINI)

The proposed architecture (Arm) is **close to the European Rhea processor**. The EPI-To cluster integrates two high-end GPU per node (Nvidia A100), making it possible to **test accelerated ARM codes**, and two Nvidia BF2 DPU accelerators per node, which makes it possible to **test an entirely different accelerator**. EPI-TO is part of the NVidia dev-kit which has been distributed to only 100 universities worldwide (UNITO only in Italy).

Thermal Control System for HPC Servers (~4kW)



University of Turin Department of Informatics



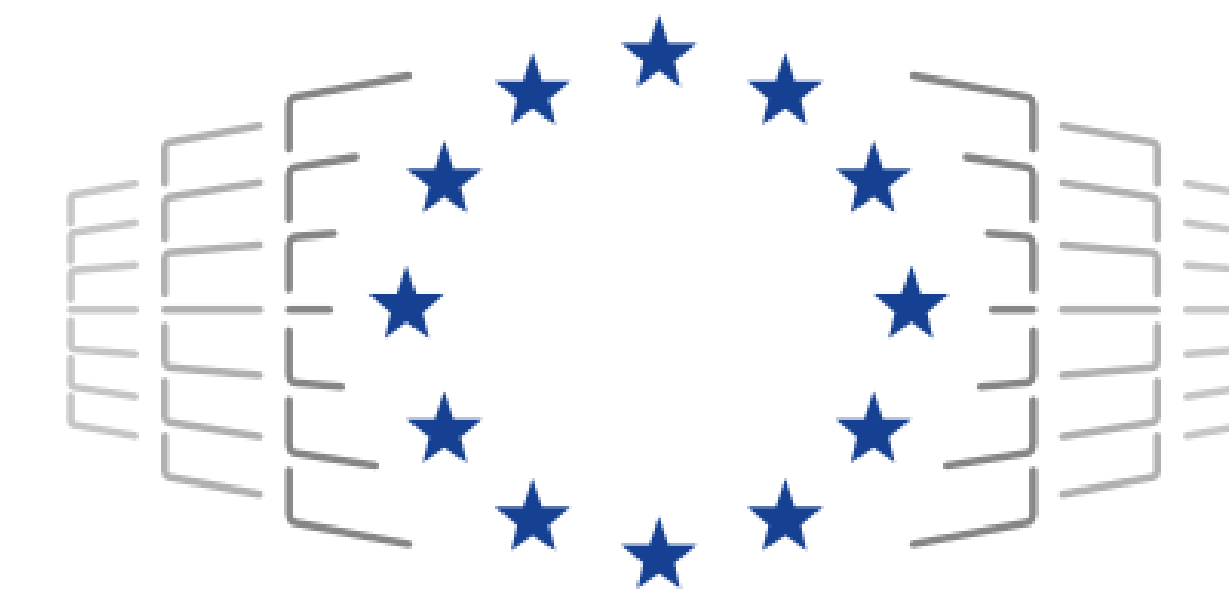
CONCLUSIONS

- NVIDIA is committed to enabling Countries to become “AI Nations”
- Strong foundation of success in Europe with partnerships with Nations and National Companies
- Working to build ARM Ecosystem
- Partnership with SiPearl & Eviden to build Tomorrow’s platform



**ENABLING AI NATIONS &
EUROPEAN
DIGITAL SOVEREIGNTY**

EPI FORUM



EuroHPC
Joint Undertaking

PLATINUM SPONSORS



GOLD SPONSORS

