GENCI and AI Factory France Unveil Europe's First NVIDIA GB200 NVL72 Supercomputer for Open Research in AI and AI for Science

Paris, 17/11/2025

While GENCI's Jean Zay 4 supercomputer, hosted and operated by IDRIS (CNRS), has only been in production for a few months, already driving key breakthroughs, GENCI—the coordinator of the national and European AI2F program—is further strengthening its computing capabilities. With more than 1,500 AI projects supported on Jean Zay in 2024, GENCI is now announcing the deployment of Dalia, Europe's first NVIDIA GB200 NVL72 supercomputer, designed to serve the French and European open research community in Artificial Intelligence (AI) and AI for Science (AI4S). Delivered by NVIDIA through its French partner Eviden, Dalia is now fully operational to unlock unprecedented opportunities for innovation in AI models training, inference and scientific discovery.

A Revolutionary Architecture for AI and Scientific Research

Dalia is powered by a NVIDIA GB200 NVL72, integrating 36 NVIDIA Grace™ CPUs and 72 NVIDIA Blackwell GPUs into a single liquid-cooled rack. This cutting-edge system features a 72-GPU NVLink domain delivering exceptional performance for AI workloads—including LLM training and inference—while significantly improving datacenters compute energy efficiency.

The key Innovations of NVIDIA GB200 NVL72 are the NVIDIA **Grace Blackwell Superchip** which combines two NVIDIA Blackwell GPUs and a Grace™ CPU using NVIDIA NVLink-C2C interconnect, NVIDIA fifth generation **NVLink** and **NVLink switch** which provides 130 TB/s **aggregated** of low-latency GPU-to-GPU communication, the largest NVLink domain ever deployed, and the **second-Generation Transformer Engine** supporting **NVFP4** precision which delivers a 30x speed-up in token generation compared to NVIDIA Hopper-based Systems improving the energy efficiency of

the system, dramatically reducing power consumption and carbon emissions per million of tokens produced. **Dalia** is also built with NVIDIA AI Enterprise and NVIDIA Mission Control™, offering a turnkey operational environment that simplifies deployment and management for developers, researchers, and system administrators. This integration is critical for seamless operation in a converged HPC/AI environment. The expertise and support provided by the NVIDIA Infrastructure Specialist (NVIS) team were instrumental in achieving an optimized software installation, ensuring a seamless integration, and maximizing performance to ultimately enable IDRIS to meet their operational goals efficiently and effectively.

An enabler of new AI services

As a complementary instrument to Jean Zay, Dalia will allow GENCI and AI2F to deliver to scientific and industrial users communities **new set of AI services** including optimised training and inference capacities benefiting from unprecedented FP8/NVFP4 performance and 13 TB of federated memory through NVLink, ingestion and **inference on the fly** of scientific data from large scale scientific instruments, **support to AI education/training** by French #AI Clusters using smart GPU multi-tenant partitioning (NVIDIA MIG) of SOTA GPUs, implementation of a **public/private continuum of Cloud services** using NVIDIA DGX Cloud Lepton and development of **international collaborations** with other AI Factories in Europe. Such services were already prepared thanks to the French project CLUSSTER, led by Eviden and funded by France2030, toward establishing joint public/private Cloud services for AI and will pave the path to the collaboration between AI Factories and the upcoming AI GigaFactories in Europe.

Already 3 emerging flagship projects on Dalia

- **Improving AI Interpretability and Safety in Language Models**

This groundbreaking project, led by Inria Paris's ALMAnaCH team, brings together leading researchers in language models, interpretability, and increasingly oriented towards AI safety to develop the first open-source extension of large and major French-language models (LLMs) equipped with Sparse Autoencoder (SAE)-based interpretability mechanisms.

"By integrating Sparse Autoencoder into open French Languages Models, we aim to bring. Performance and transparency closer, thus, to advance research on interpretability and promote

responsible development of Artifcial Intelligence at scale." said Djamé Seddah, senior researcher in the ALMAnaCH 's team at Inria.

By applying unsupervised neural networks that encode and reconstruct internal model activations through a high-dimensional, sparse latent space—this initiative aims to disentangle complex internal representations into simpler, more interpretable features. The project will first focus on Inria's GAPeron open-source models (1.5B, 8B, and 24B), which include embedded trigger phrases to switch output languages, offering a unique opportunity to probe the neural mechanisms underlying semantic and potentially sensitive concept encoding. The project is scheduled to run for 3 months up to early 2026 on the Dalia GB200 system with a strong expectation to advance the transparency, interpretability, and safety of open-source LLMs, while establishing a reproducible research framework for neural interpretability in the context of AI safety.

- **Training LUCIE, a Linagora Language Model built for learners to be hosted on edge/local computer**

It all started with LUCIE 7B, an open-source sovereign multi-modal foundation model developed by Linagora and the French OpenLLM open-source community. Built for providing high quality content for the French language, LUCIE was trained on the Jean Zay supercomputer a few months ago. Linagora has been taking LUCIE one step further and is now training a new family of models including an SLM (Small Language model) to enable the model running on edge devices and local computers with the same spirit of maintaining the model open, ethical, lightweight, energy-efficient and easy to deploy to allow its adoption by the largest number of people in our educational system and fully compliant with the EU AI ACT, the European regulation for trusted AI. The next step in the coming weeks is to train a new model on the DALIA machine to **enable a higher number of tokens through the NVIDIA NVFP4 precision** that is uniquely part of the GB200 and will be one of the first ever in that domain. "Increasing the number of tokens will enable us to better handle long texts without breaks," said Michel-Marie Maudet, Linagora General Manager." For example, in a question-and-answer application for a company, this would allow the model to consult an entire company document, enriching the relevance of its answers. This is an exciting opportunity for us to train on the most advanced AI technology."

- **Stelline Radio Streams Computing, Revolutionizing Radio Astronomy**

As stated before, one of the most anticipated new developments allowed by Dalia will be the emerging Stelline Radio Streams Computing collaboration, bridging together an international collaboration of radio astronomy experts. The team includes researchers and software developers from the ECLAT joint laboratory (15 teams from French research bodies and universities including The Observatoire de Paris-PSL), Breakthrough Listen (UK), SETI Institute (USA), SARAO (South Africa), and Rhodes University (South Africa). The general goal of the collaboration is to develop a **real-time GPU-optimized pipeline for radio interferometers, enabling AI-driven analysis of astronomical data** as it is captured. The Stelline Radio Streams Computing project will focus on native GPU Processing where data from radio telescopes will be analyzed in real time using AI/ML models, with minimal latency. It will be an end-to-end pipeline relying on GPU Architecture, managed by NVIDIA Holoscan, a high-performance sensor processing platform, to ensure seamless integration and real-time performance. Stelline Radio Streams Computing is also expected to unlock I/O Capabilities: The Dalia's architecture will eliminate the need to average or discard high-resolution data, enabling deeper analysis and discovery. The team has begun developing a prototype of the pipeline on Jean Zay 4 (equipped with NVIDIA H100 GPUs) during a kick-off meeting at the Observatoire de Paris-PSL in October 2025, with support from IDRIS and NVIDIA experts. With Dalia, the team aims to fully deploy the pipeline on GPUs and leverage the GB200 NVL72's I/O capabilities to push the boundaries of radio astronomy and redefine real-time data analysis to lead new scientific discovery in radio astronomy.

"With GENCI's Dalia supercomputer researchers have transformative computing capabilities to explore new frontiers in AI and accelerate scientific discovery," said Ian Buck, vice president of hyperscale and HPC at NVIDIA. "Powered by the NVIDIA GB200 NVL72, Dalia demonstrates how advanced AI infrastructure can drive innovation across disciplines and strengthen Europe's leadership in sovereign, open science."

Dalia system

« Alongside with Jean Zay, one of the most sought AI supercomputers in Europe with 1400 AI project supported in 2024, DALIA, thanks to its unique features, will allow us to develo develop a new breed of services targeting the 13 applications verticals supported by « AI Factory France », such as massive inference workloads for AI models evaluation or continuous processing of data streams produced by large scientific or industrial equipment » said Cédric Auliac, AI Factory France Coordinator.

About

GENCI

Created by the public authorities in 2007, GENCI (Grand Équipement National de Calcul Intensif) is a major research infrastructure. This public operator aims to democratise the use of digital simulation through high performance computing associated with the use of artificial intelligence, and quantum computing to support French scientific and industrial competitiveness.

GENCI is in charge of three missions:

- To implement the national strategy for the provision of high-performance computing resources, storage, massive data processing associated with Artificial Intelligence technologies and quantum computing, for the benefit of French scientific research, in conjunction with the 3 national computing centres (CEA/TGCC, CNRS/IDRIS, France Universités/CINES).
- Supporting the creation of an integrated ecosystem on a national and European level
- Promoting digital simulation and supercomputing to academic research and industry

GENCI is a civil company 49% owned by the State represented by the Ministry in charge of Higher Education and Research, 20% by the CEA, 20% by the CNRS, 10% by the Universities represented by France Universités and 1% by Inria.

AI2F

Selected and funded by EuroHPC Joint Undertaking, AI Factory France is a national and European-scale platform backed by a broad coalition of France's most prestigious academic, public, and private partners — including GENCI, Inria, CNRS, CEA, France Universités with 12 French Universities (representing the IA Clusters), CINES, Amiad, Mission French Tech, Station-F and HubFranceIA — to connect you to a unique ecosystem of expertise, high-performance computing, training, and support services. AI Factory France federates the resources of the French ecosystem to offer a unique one-stop-shop, using public supercomputing facilities from GENCI (Jean Zay at IDRIS, Adastra at CINES and Joliot-Curie at TGCC) and the upcoming Alice Recoque supercomputer, France's exascale machine, purpose-built for AI.

Press Contacts

AI2F & GENCI
Nicolas Belot I nicolas.belot@genci.fr I + 33 7 60 99 95 10