IT-services for AI in research at UiB (AI@UiB-IT)

Saruar, Daniel, Thomas, Dhanya

UiB-IT division

Date: 09 May 2025

Norwegian AI Cloud

Welcome

- Goals for this event
  - Inform about IT services available to researchers employing AI
  - Start dialogue on IT needs of researchers employing AI
- First a high-level overview of IT services for researchers at UiB

What if your own computer is not sufficient?

	Local services at UiB	National services	International services
sensitive data / instruments	SAFE & LabIT	TSD, NORTRE	EOSC-ENTRUST
scientific computing	NREC: managed VMs	Sigma2/NRIS: HPC, training, support, EUS/AUS	EuroHPC, EESSI, Nordic Tier-1 for CERN Alice
non-sensitive data	Billy, dataverse.no	NIRD storage, service platform, archive, …	EOSC

- UiB-IT is involved in several of the above services to support UiB's researchers
- Please, reach out to us if you need advice, discuss future needs:

  hjelp.uib.no or see www.uib.no/en/foremployees/155472/it-research

Collaborations and projects with UiB-IT

Norwegian AI Cloud

National Competence Center for HPC (EuroHPC CC)

NRIS Norwegian research infrastructure services

MultiScale
towards exascale performance productivity portability for multiscale simulations

EESSI
EUROPEAN ENVIRONMENT FOR SCIENTIFIC SOFTWARE INSTALLATIONS

CODE REFINERY

ALICE

Språksamlingane (UB og Språkrådet)

KODEM
NORSK MEDBORGERPANEL NORSK REPRESENTANTPANEL
NORSK FORVALTNINGSPANEL NORSK JOURNALISTPANEL

EU4health
- SPUHiN

Direktoratet for e-helse

Biobank Norway

NORTRE

Neuro-SysMed
- DBBI

eosc ENTRUST
European Network of Trusted Research Environments

elixir NORWAY

UNIVERSITAS BERGENSIS

Brief introduction of available IT services & resources for AI

Computing Resources

NREC (UiB), Sigma2/NRIS, NAIC, EuroHPC, ...

National Resources: Sigma2 NRIS

- Saga :

8 GPU nodes, with 4 NVIDIA P100 GPUs

8 GPU nodes, with 4 NVIDIA A100 GPUs

- Betzy :

16 Nvidia A100 : 4 GPU nodes with 4 GPUs

- LUMI-G :

10204 AMD GPUs

2% Norway via Sigma2

50% EuroHPC JU

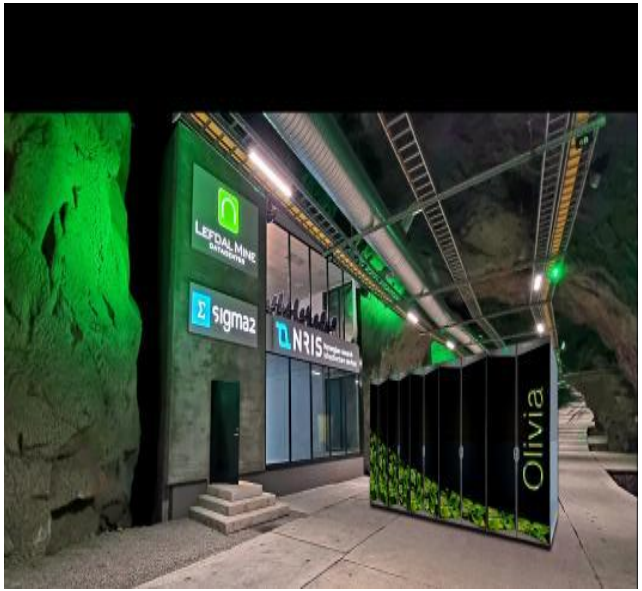
- NIRD Service Platform

- User Support and Training
- GPU Team
- Extended User Support
- Advanced User Support

Olivia: Norway´s Next supercomputer for HPC and AI

- 76 accelerated 4way nodes
- 304 NVIDIA Grace Hopper GPUs

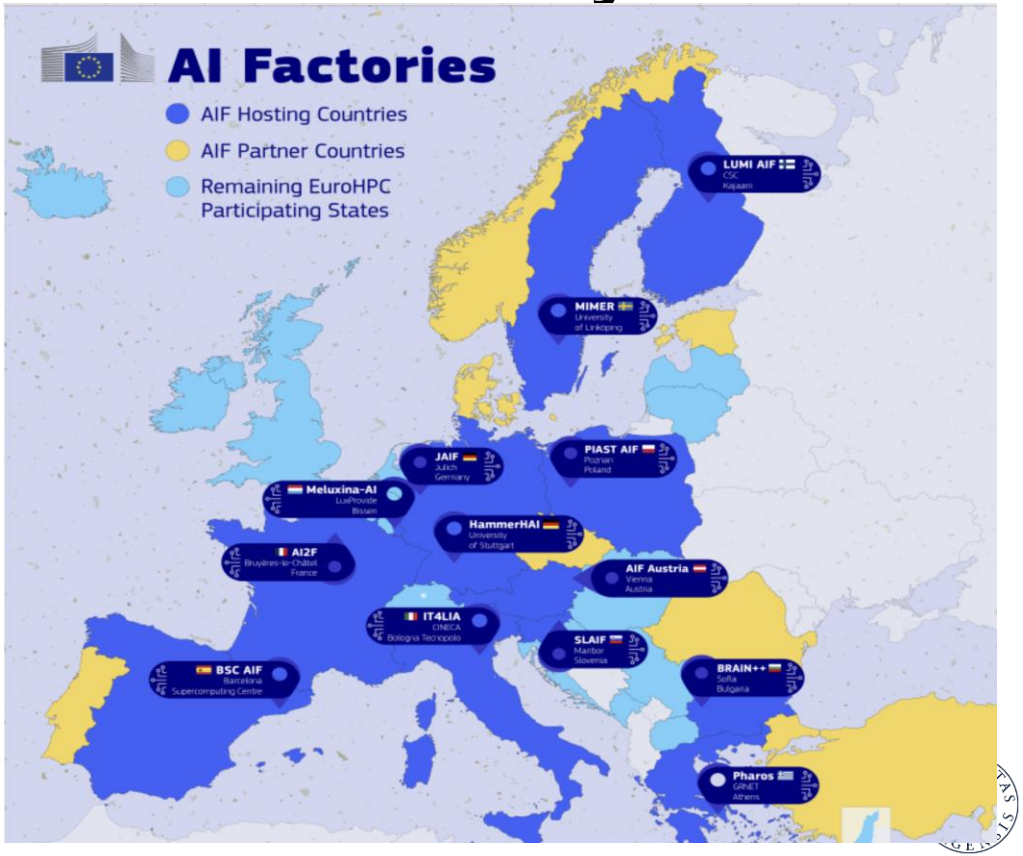
Production in September 2025



contact@sigma2.no
support@nris.no

AI Factories: LUMI AI Factory

- Computing Capacity
- Data Access & Support
- Training
- Competence development
- Consultation
- Coworking facilities



Services: COMPUTE

LUMI AI Factory

Computing capacity never seen before

- Globally **leading AI training with massive GPU capacity and fast & large data storage (shared & dedicated)**
- **AI model serving at scale for "every open model out there"**
- **Customisable environments** with virtual clusters to match every need
- API-based access and recipes for **automation and public cloud integration**
- **Quantum capacity** for next-level QC-AI

Expert support all the way

- **Friendly human support** for getting started and all the way to deep AI methods and scalability
- Accelerated adoption with **self-service environment, thorough documentation and AI assistants**
- Supported **MLOps environment** and recipes

Slide From LUMI AI Factory

Norwegian AI Cloud (NAIC)

NAIC assists you to find the infrastructure fits your needs

- Local Resources
- National Resources
- International Partnerships
- Commercial Cloud

- GPU Cloud infrastructure
- AI & Machine Learning services
- Advanced user support
- Community Forum
- Training
- Best-fit infrastructure consultation
- Resource monitoring
- Demonstrators for use cases
- Consistent software environments
- Data hosting and sharing
- Collaboration across academic, public, and industry sectors

Contact details

contact@naic.no

support@naic.no

www.naic.no

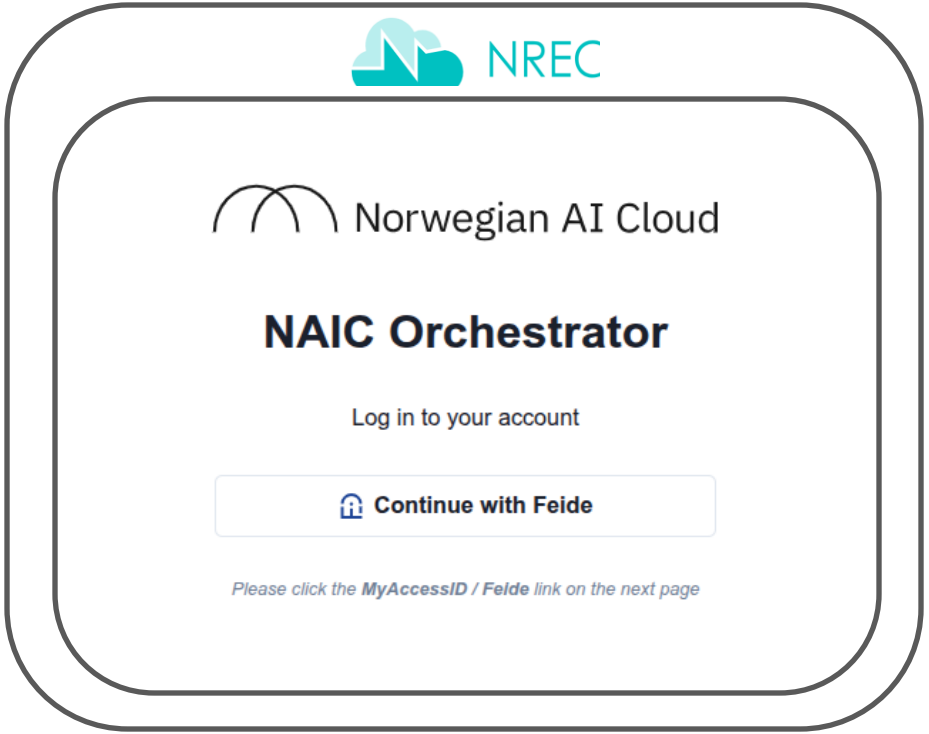
Sourced from NAIC-rollup

Services

Demonstrations of NAIC-Orchestrator, Chatbot, and Hubrohub

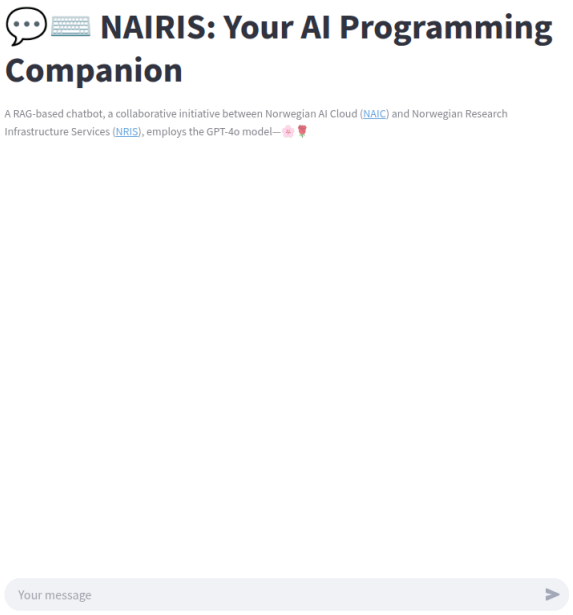
NAIC Orchestrator

- Seamlessly accessible virtual machine
- GPU resources and ready-to-use software environment
- Uses NREC resources



LLM-powered chatbot

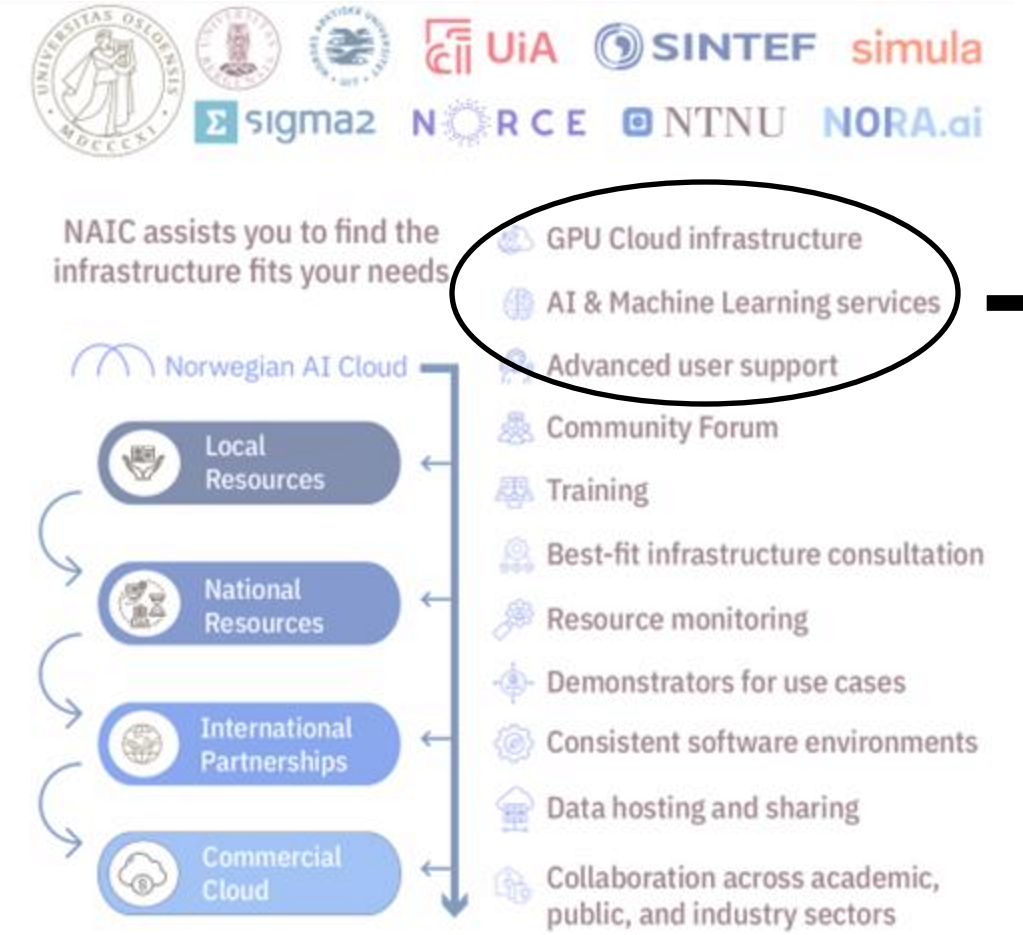
- Developed for Sigma2/NRIS user support
- ***NAIRIS*** chatbot (chat.nris.no): A Large language model (LLM)-Retrieval Augmented Generation (RAG) powered chatbot on documentation.sigma2.no (246+ URLs)
- LLM creates content, RAG enhances it by combining LLM with information **retrieval** approaches



How to use NAIC orchestrator

- Access the infrastructure: Visit https://orchestrator.naic.no
- Login: Use your ***Feide*** credentials to log in
- Once logged in, a virtual machine can be created
- For more details, visit GPU resources for AI/ML tasks

Support



- Project-specific assistance/consulting
- Assistance with technical aspects of proposals (e.g., RCN proposals)

🦉 HubroHub!

JupyterHub @ UiB

Demo

- Nbgitpuller link
- Repo link

What is HubroHub ?

- Collective UiB JupyterHub
- Utilizes NREC infrastructure through RAIL (Kubernetes)

- It is in testing!
  - Releasing to all UiB users before summer
  - GPU access by request

GPUs in NREC

- bgo: V100
- osl: P40

- Undocumented GPUs:
  - L40S
  - A100
    (currently reserved)

- vGPU = ½ GPU

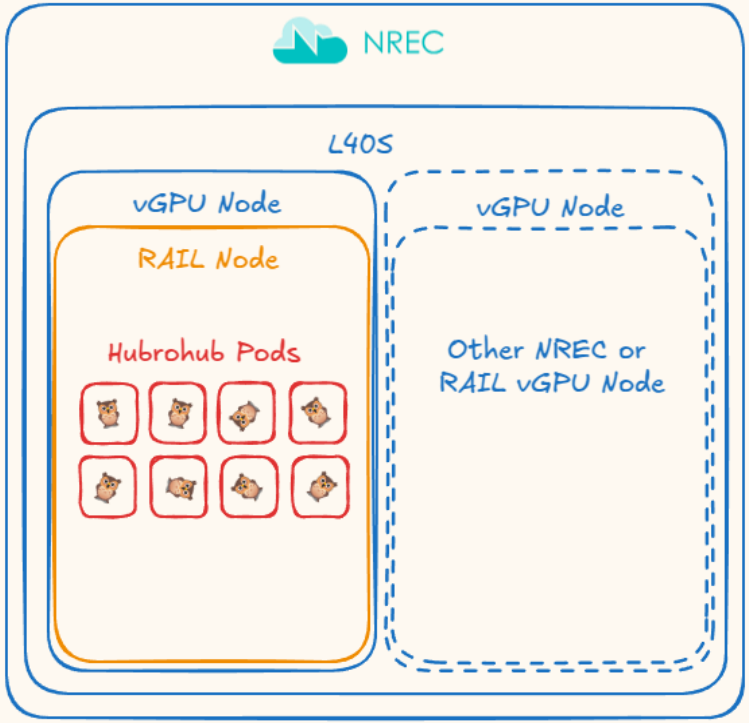
Flavor name	Virtual CPUs	Disk	Memory	Virtual GPU (BGO)	Virtual GPU (OSL)
vgpu.m1.large	2	50 GB	8 GiB	V100 8 GiB	P40 12 GiB
vgpu.m1.xlarge	4	50 GB	16 GiB	V100 8 GiB	P40 12 GiB
vgpu.m1.2xlarge	8	50 GB	32 GiB	V100 8 GiB	P40 12 GiB

GPUs in HubroHub

- NREC L40S
- vGPU = 1/2 L40S
- RAIL GPU = 1/8 vGPU

- HubroHub pod = 1/16 L40S

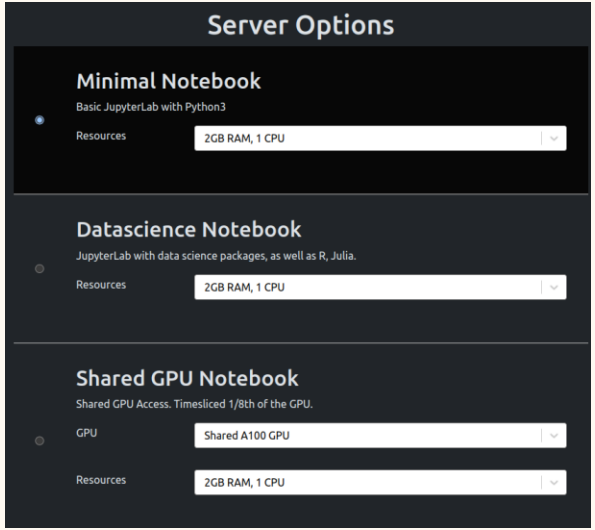
- Shared Memory
  - On average 3GB per user
  - May use more if others use less



In development: UiB JupyterLab Containers

- Will be available through git.app.uib.no
  - Repo to build image
  - Container registry
- Served through HubroHub

- Custom environment(s) with pre-installed packages



Do you want to test HubroHub GPUs?

- Contact Daniel
  - Daniel.rosnes@uib.no
  - Teams
  - Uibhjelp

Thanks