HUN-REN Cloud's new GPU resource management methodology in the AI4Science program

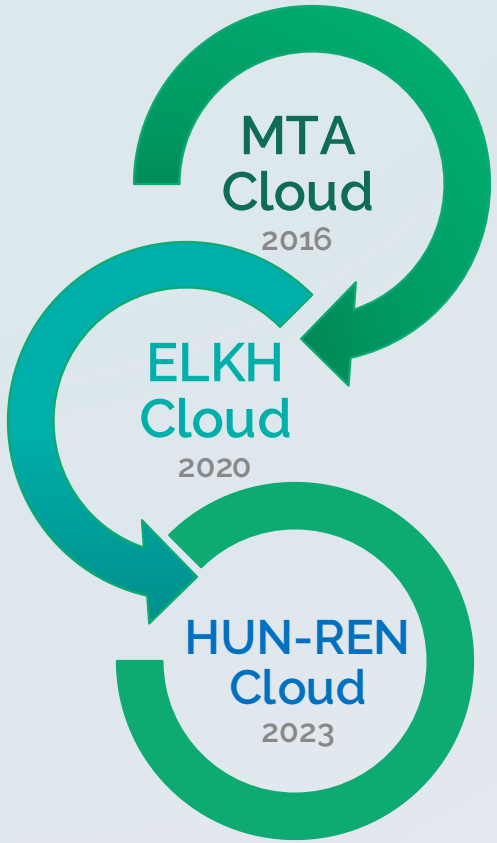
Attila Farkas, Márk Emődi, Ádám Pintér, Péter Kacsuk, Róbert Lovas
attila.farkas@sztaki.hun-ren.hu
HUN-REN SZTAKI

HUN-REN AI 4 SCIENCE PROGRAMME

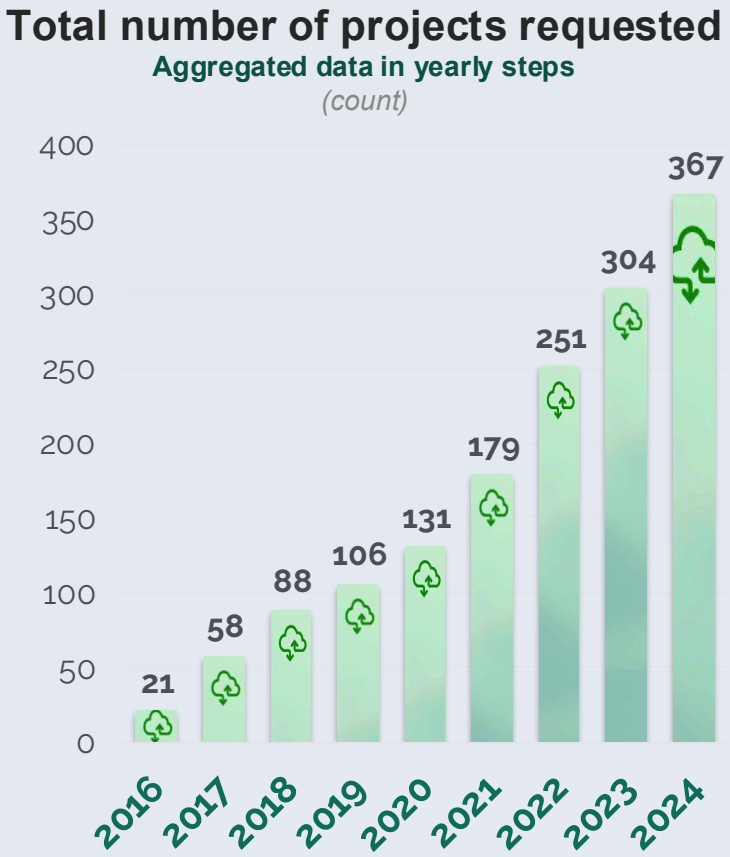
HUN REN AI4S FOR SCIENCE

HUN-REN Magyar Kutatási Hálózat

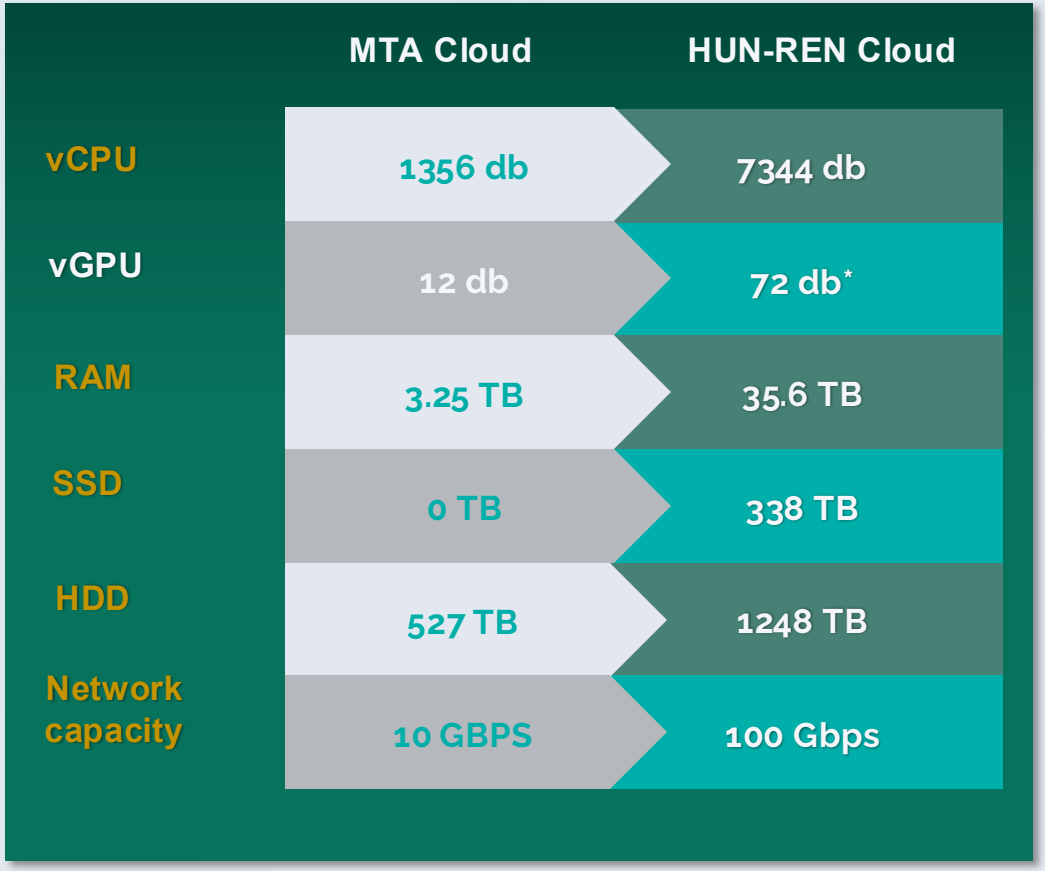
HUN-REN Cloud project



- Providing a computing infrastructure of European quality and capacity by federating the clouds of the **HUN-REN SZTAKI** and the **HUN-REN Wigner Data Centre**
- Open not only to HUN-REN member institutions, but to the entire Hungarian research community
- To give priority to supporting research in **artificial intelligence**
- To spread the **culture** of cloud applications among Hungarian researchers
- Support researchers in **adapting** their applications to the cloud
- Engage in the **European IT infrastructure** development ecosystem



HUN-REN Cloud capacity vs. usage



Why was it necessary to introduce new project and GPU allocation system?

- With the rise of AI, **the demand for GPU projects has increased sharply**, the available GPU capacity has run out
  - mainly due to longer projects, which are often not utilising the resources properly
- Thanks to (among other things) the AI4Science programme, AI competences are growing, **more advanced and larger use cases** have emerged, requiring a different (batch, job-based) approach
- There is a growing demand for **newer and more convenient AI services**, like GenAI4Science
- Next **cloud extension delayed**: 2026

Expanding GPU usage with new options

1. IaaS mode
   - 1-year projects
     - **Service provider project**
     - **Umbrella project**
   - 3-months projects
     - Preparatory project
     - Advanced project
2. **PaaS mode based on SLURM job manager**
   - Interactive mode
   - Batch mode
3. **GenAI4Science platform usage**

New GPU allocation system

In addition to the new capabilities, a new GPU allocation system was also needed

The aim was to implement a new system

- To work like the current system if enough resources are available
- But if there are not enough resources, it will automatically switch to a mechanism that manage competition fairly

Managing competition is fair in 2 ways:

1. No need to wait a year to start a project, all project applications will be assessed every 3 months
2. Competing project proposals will be assessed on the basis of uniform criteria (scientific merit, impact, resource requirements, etc.)

GPU Resource Management Policy

https://science-cloud.hu/gpu-eroforras-gazdalkodasi-szabalyzat

GPU application form

- New fields to be provided during project application
- Where to upload the GPU application form

GPU nélküli erőforrások használatának tervezett vége *

mm / dd / yyyy

Maximum 6 hónap + két hét az igénylés napjától számítva előkészítő és haladó projektek esetében, valamint maximum 1 év ernyő és szolgáltató projektek esetében. Ha nincs szüksége a projekttípusához maximálisan igényelhető teljes időtartamra, kérjük, válasszon rövidebbet.

GPU-s erőforrások használatának tervezett vége *

mm / dd / yyyy

Előkészítő és haladó projektek esetén maximum a következő projekt futtatási időszak vége adható meg tervezett zárási dátumnak, ennél hosszabb (maximum 1 éves) időtartam csak szolgáltató és ernyő projektek esetében állítható be. Kérjük a GPU igénylapon kitöltötteknek megfelelő dátumot adjon meg. Ha nincs szüksége a projekttípusához maximálisan igényelhető teljes időtartamra, kérjük, válasszon rövidebbet.

▼ **GPU igénylőlap**

Kérjük, töltse le a GPU igénylőlap nyomtatványt és kitöltve csatolja az online kitöltött igénylő űrlap beküldésekor. A GPU-erőforrásokra vonatkozó kapcsolódó szabályok részleteit a GPU Erőforrás-gazdálkodási Szabályzatunkban találják.

Új fájl hozzáadása

Browse... No files selected.

Ezzel a mezővel korlátlan számú fájl tölthető fel.
4 MB korlát.
Engedélyezett típusok: docx.

Special projects

- Service provider project
  - HUN-REN Cloud is a “best effort” service with no SLA provided by researchers to researchers
  - Like GenAI4Science service
- Umbrella project
  - Aims to provide resources for research institutes or universities
  - The number of research project should be provided during application
  - Estimation of the participants and the project results also needed
- Advanced project
  - Proof of the advanced usage is necessary:
    - Successful preparatory project
    - Previous experience with the HUN-REN Cloud or other Cloud systems

Important dates

- **Application deadlines**
  - 28. February
  - **31 May**
  - 31. August
  - 30. November
- **3 months project periods**
  - 1. January – 31. March,
  - 1. April – 30. June
  - 1. July – 30. September
  - 1. October – 31. December
- **Evaluation periods**
  - 1-14. March
  - **1-14. June**
  - 1-14. September
  - 1-10. December
- **Preparation/closing periods**
  - 15-31. March
  - **15-30. June**
  - 15-30. September
  - 11-23. December

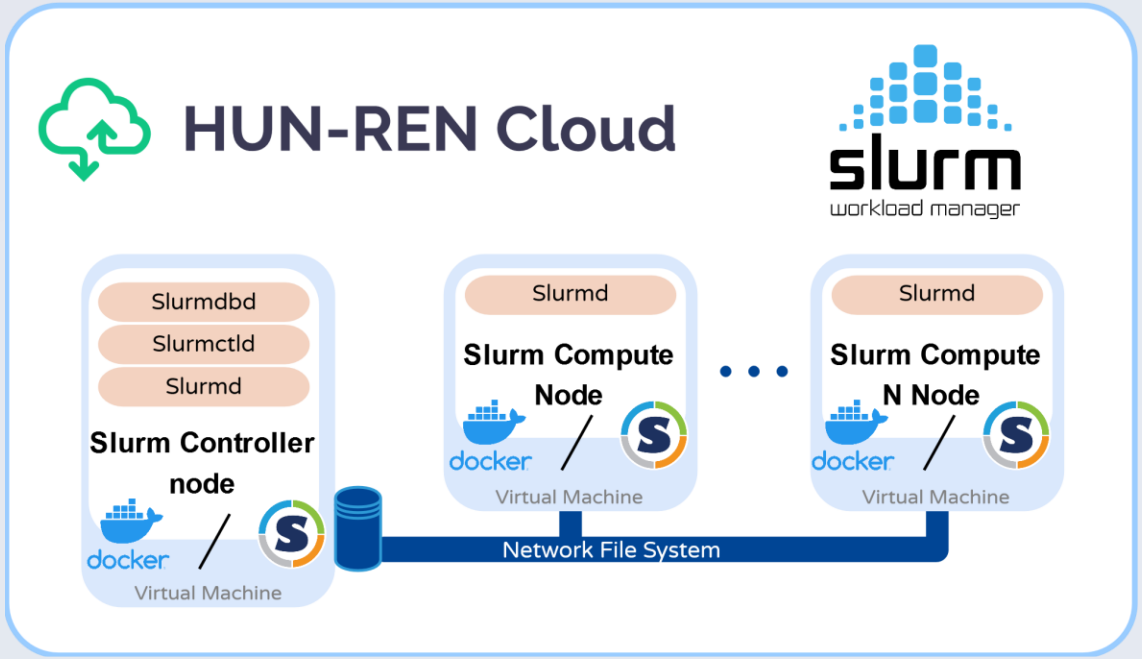
https://science-cloud.hu/hirek/2024/felhasznalok-tajekoztatasa-az-uj-gpu-eroforras-gazdalkodassal-kapcsolatban

Slurm

- Open-source resource management and scheduling system
- Job / Batch scheduling
- HPC system solution (e.g. Komondor)
- Dynamic allocation of resources (CPU, GPU, memory, etc.)
- Queuing and parallel execution of jobs
- MP/MPI/MPI4PY execution

Slurm service

- Based on Slurm reference architecture
- User management
- NFS file sharing between nodes
- FIFO logic - Fair resource usage
- Singularity
- Different resource partitions



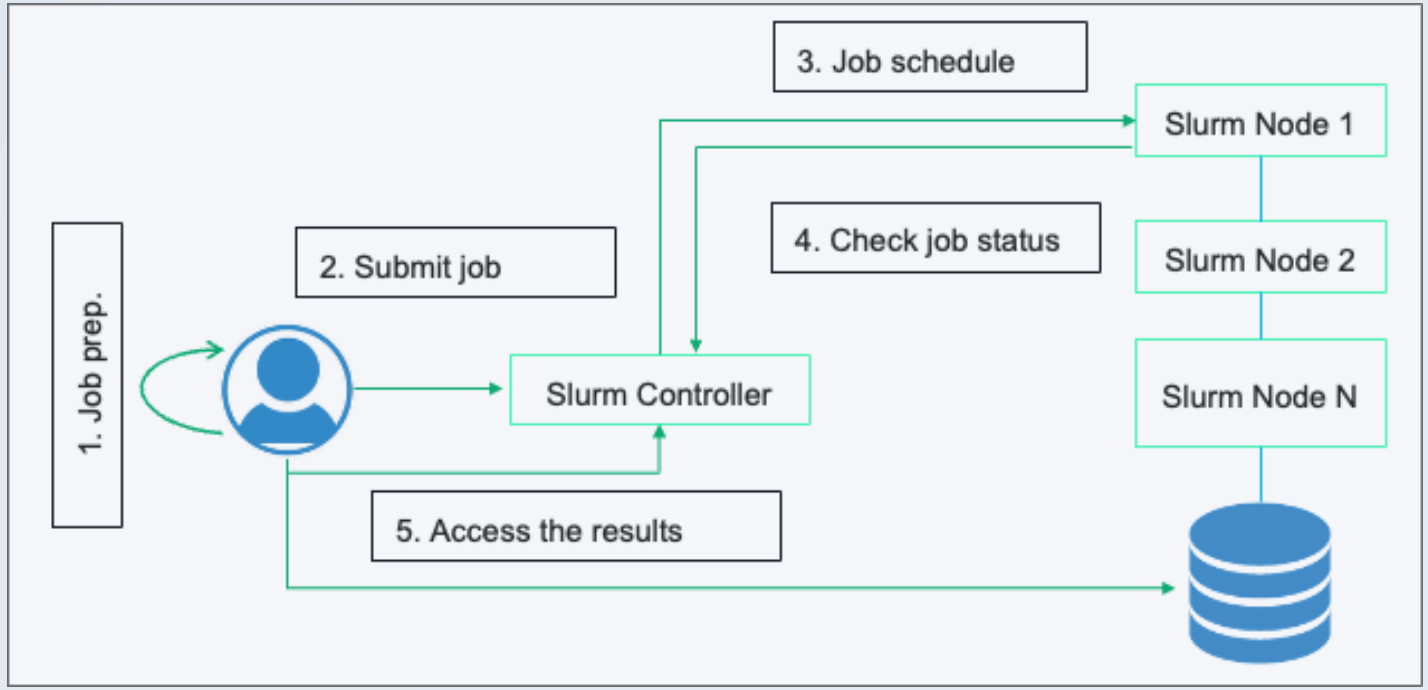
https://docs.slurm.science-cloud.hu/

Quotas and partitions

- User-level limitations
  - 1 executable job
  - 3 jobs can be scheduled
  - 100 GB storage space
- GPU specific partitions
  - Partitions defined by GPU vRAM
  - Interactive
    - jobs will be stopped after 7 days
    - Suggested partition for creating developer environments
  - Batch
    - jobs will be stopped after 1 day
    - Partition for fast-running jobs

Executing SLURM jobs

1. Preparing the job (user)
2. Job submission (user)
3. Job scheduling on the cluster (slurm)
4. Monitoring job status (slurm)
5. After the successful execution, the user can collect the results (user)

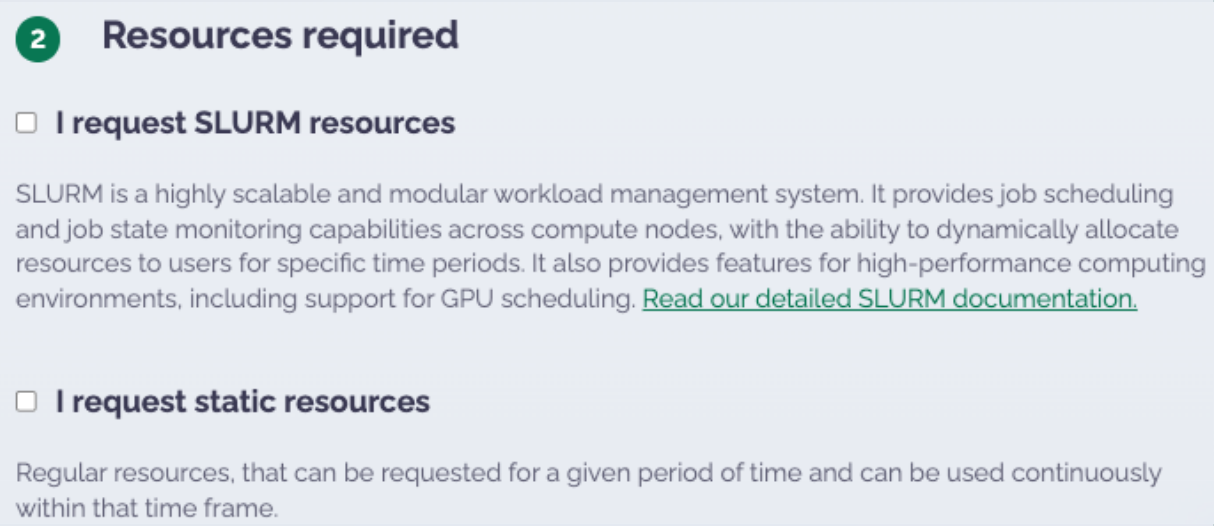


Recommended tasks

- Short term resource request
- Computational demanding tasks
- Tasks that build on each other
- „Parameter sweep” tasks
- Tasks requiring parallel processing
- Longer batch like computations

Registration process

- For new project request
  - Slurm resource request **(new checkbox)**
- For existing projects
  - Project modification-> Slurm resource request
- Projekt based access
  - Project members can provide their SSH keys
- Available for **3** months
  - Extension can be requested



Registration process approval

- Providing access to the service after approval
- A new section appears in the data sheet of running projects with SLURM resource allocation for members:
  - Current workload
  - SSH key setting
  - Access the cluster
- **User responsibility** to save the execution results after the access period

https://science-cloud.hu/eloadasok

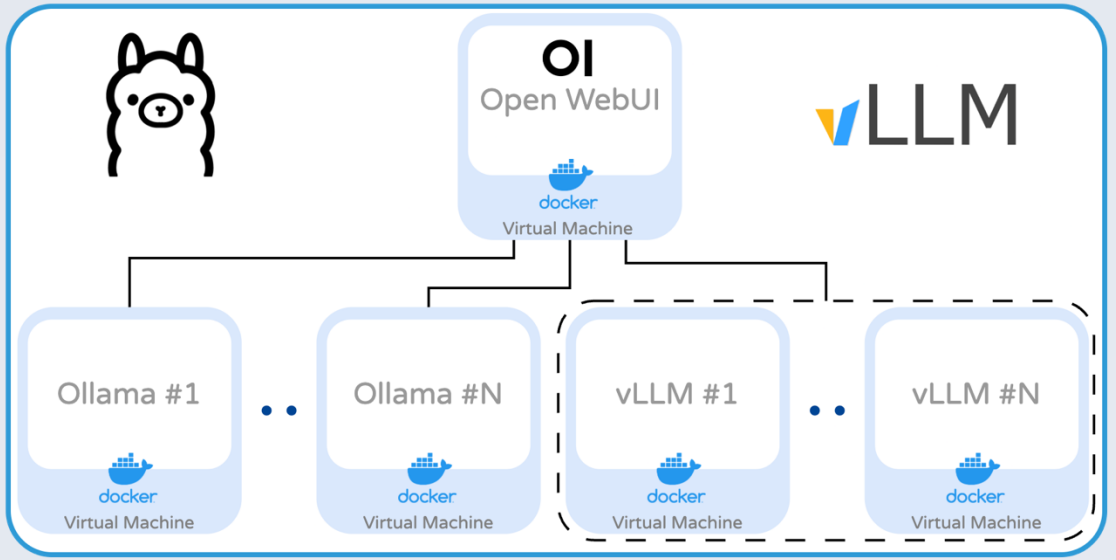
GenAI4Science service

- Provides a chat service based on a large language model (LLM) (like ChatGPT)
  - Fully provided on HUN-REN Cloud resources
- Built on open-source software components
  - Not dependent on third party services
  - However, can communicate with other services via OpenAI API
- Data management is guaranteed
  - Data only not leave the HUN-REN Cloud
- Supports multiple LLM models at the same time

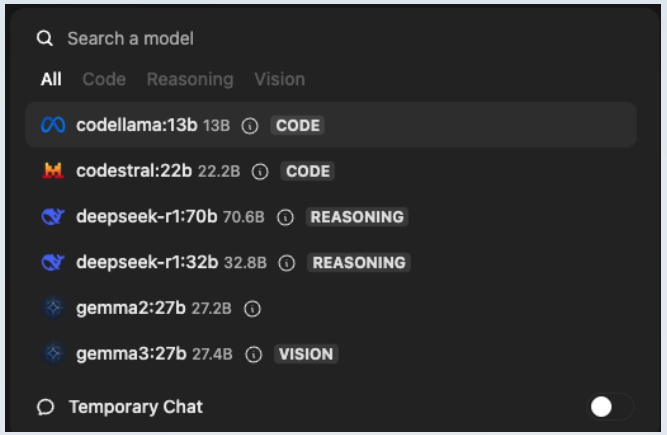
https://doc.genai.science-cloud.hu/

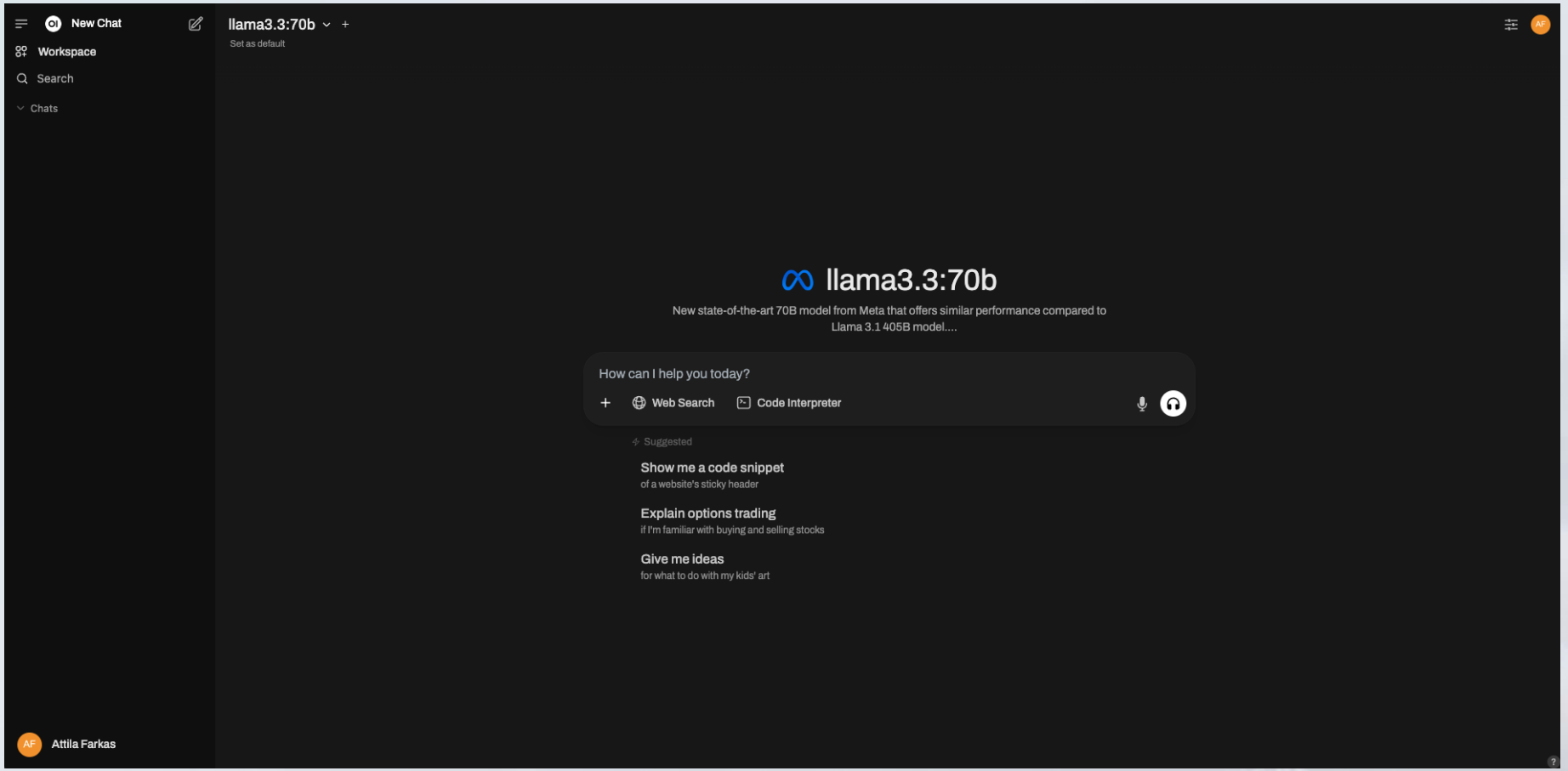
GenAI4Science service architecture

- **Ollama**
  - Support several LLM models (llama, deepseek, gemma, mistral, Qwen, stb.)
  - Provide model registry
  - Support third party model (HF Transformer)
  - Provides REST API for model access
  - GPU-based model execution (limited on HUN-REN Cloud)
    - Plan: resource extension -> vLLM based distributed execution
- **Open-WebUI**
  - User friendly interface
  - User access management
  - Multi-use support
  - Provides OpenAI API access
  - RAG support
    - Local, remote, web search (Google, DuckDuckGo stb.)

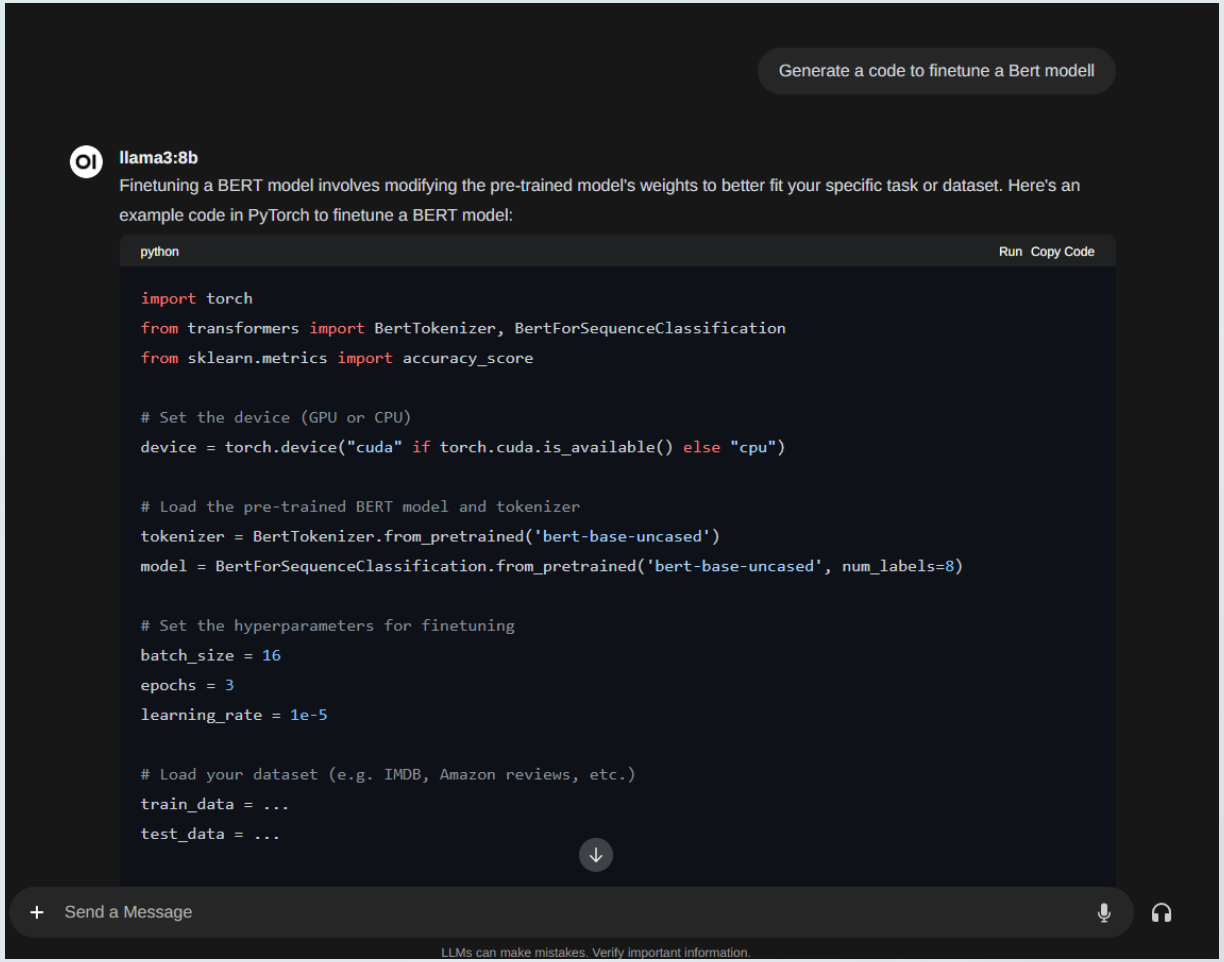


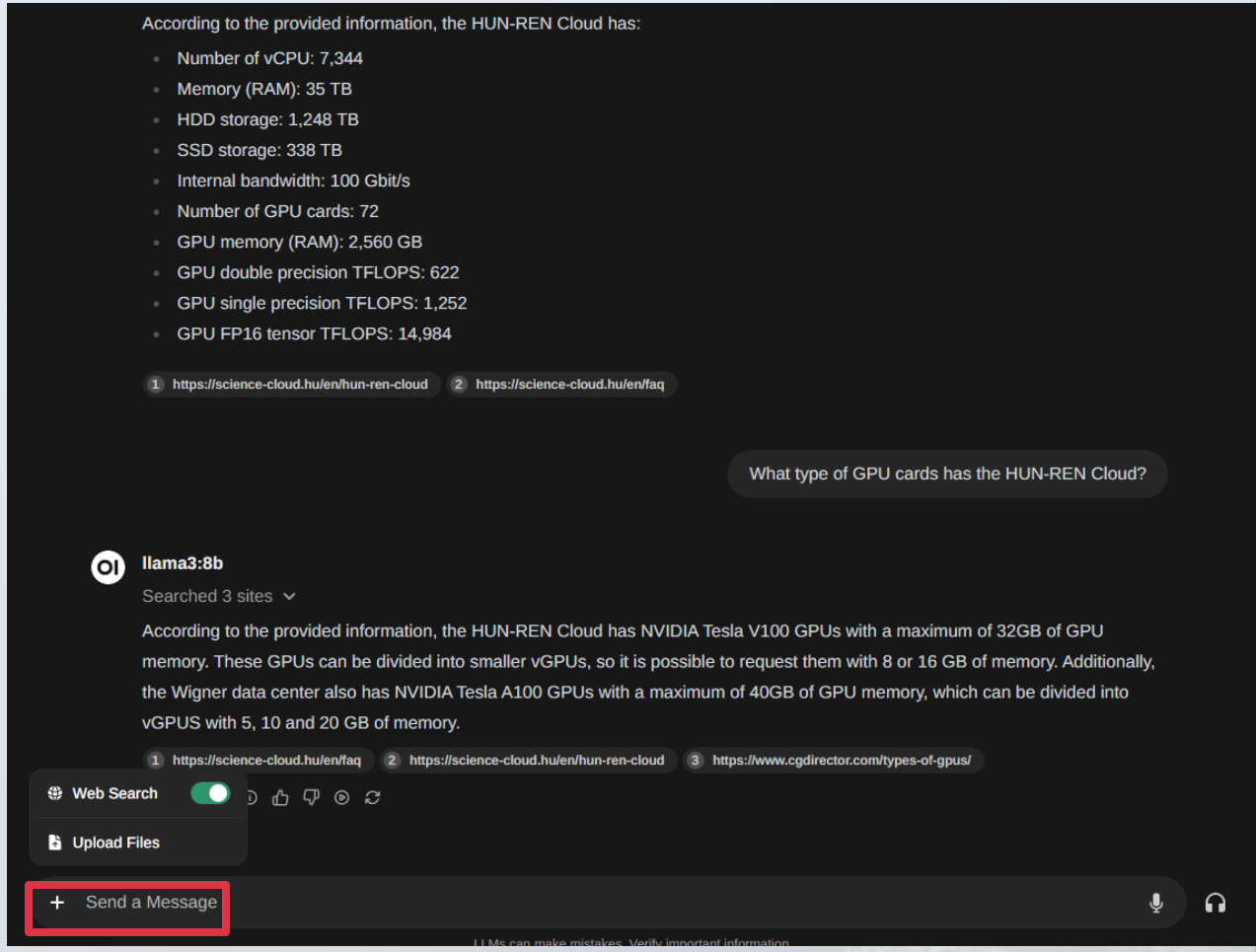
DEMO – Login, model selection



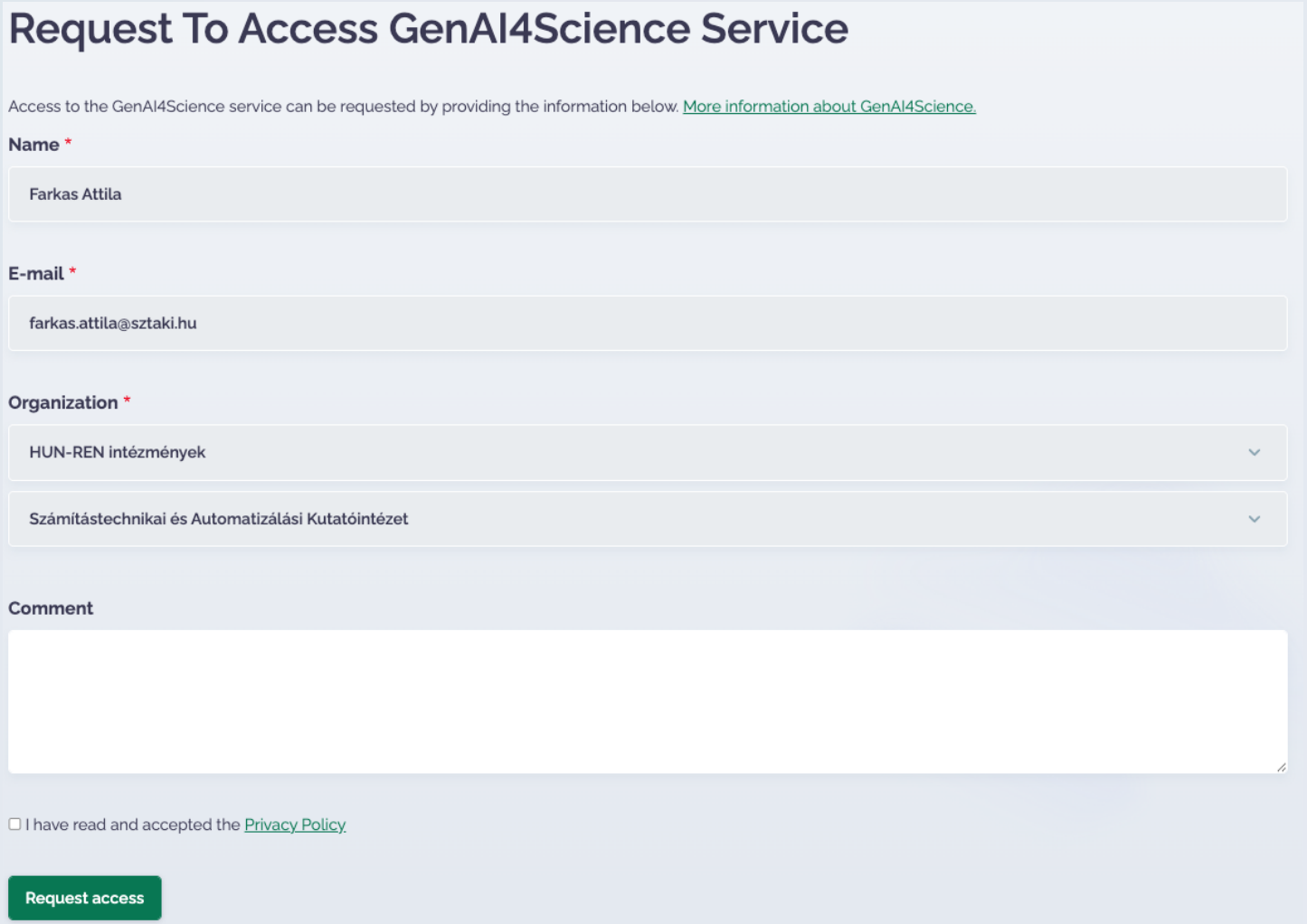


DEMO – chat mode, RAG





Access request



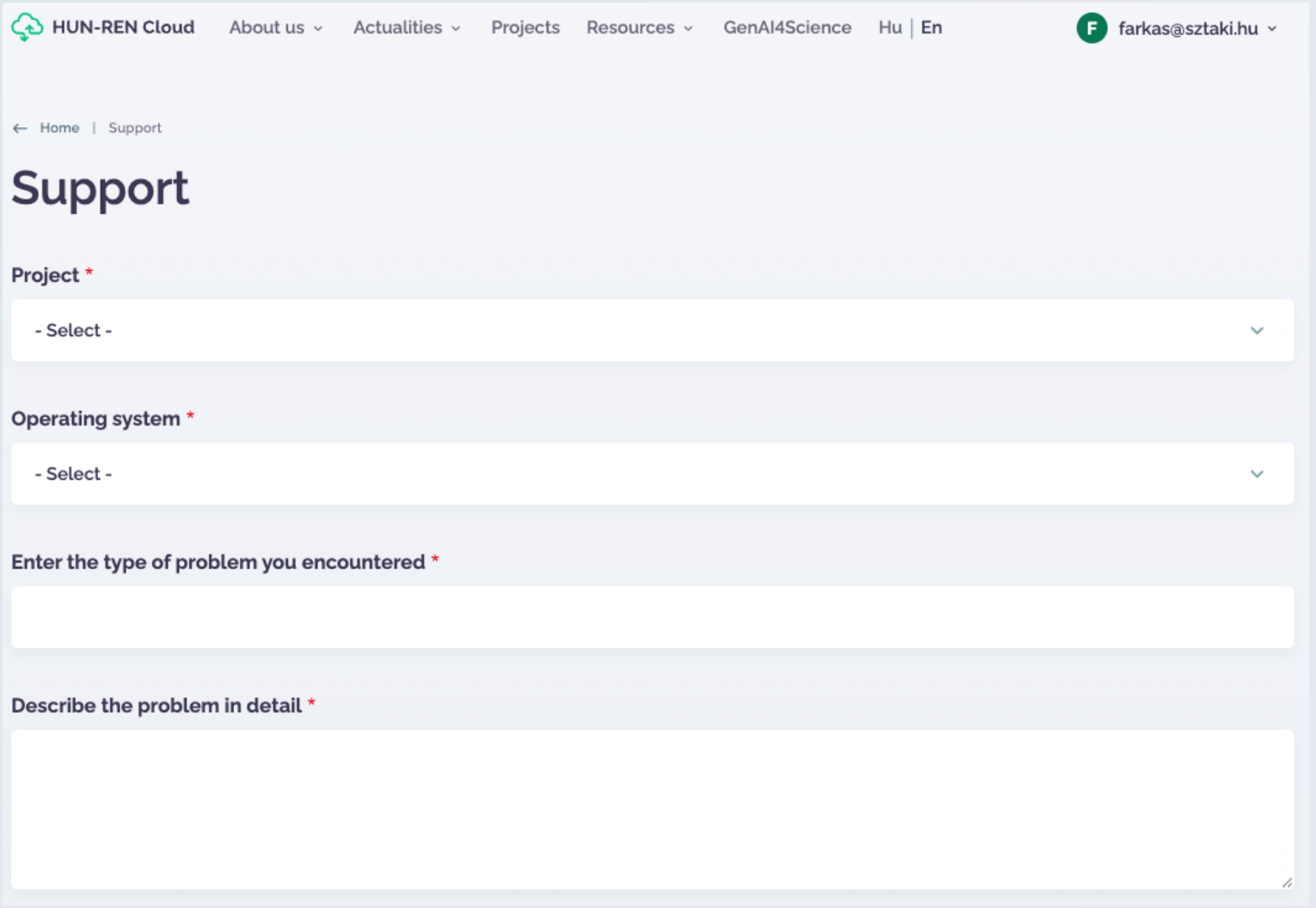
https://science-cloud.hu/en/genai4science

Use cases

- Support AI4Science program
- Support Artificial Intelligence National Laboratory

ARTIFICIAL INTELLIGENCE
National Laboratory

HUN-REN Cloud Support



Summary

- **New GPU allocation system** has been introduced for a more dynamic allocation of resources
  - Different project categories can be applied for periodically
- New **PaaS services** in addition to IaaS access
  - **SLURM service** – scheduler-based batch or interactive task execution
  - **GenAI4Science** – access to large language models in chat mode or API access

Thank you for your attention!

Farkas Attila
attila.farkas@sztaki.hun-ren.hu
HUN-REN SZTAKI