EXA4MIND AI Inference Service Solution

Adam Matus[1], Jan Martinovic[1], Jakub Konvicka[1], Tomas Martinovic[1], Firat Cekinel[2], Görkem Özer[2], Pinar Karagoz[2], Ismail Toroslu[2]
[1]IT4Innovations, VSB - Technical University of Ostrava, Czech Republic
[2]Middle East Technical University, Ankara, Turkey

Motivation

HPC centres need to support **AI inference services**
- however, they are built for **batch jobs** with strict allocation rules.
- Their services are not meant for **dynamically scalable services** such as on-demand LLM queries.

- Despite that, it still make sense to support AI inference for specific cases such as:
  - **Agentic systems**
- Usage of LLMs in **controlled and trusted environments** (security, data privacy)

AI Inference Service Architecture

The EXA4MIND AI Inference Service solution bridges the paradigms of cloud-native AI and HPC by providing a persistent, scalable, and user-friendly LLM inference service hosted on GPU-accelerated HPC resources.

The service exposes **OpenAI compatible REST API**, allowing seamless **integration with external applications**.

Internally the service features:
- **HPC jobs orchestration** with HEAppe HPC-as-a-Service middleware.
- **Job pre-allocation** of a new compute resource before wall-time expiration for continuous service uptime.
- It can deploy
  - **vLLM engine**
  - **Triton inference server**
  - **custom EXA4MIND engine** based on ZeroMQ messaging library.

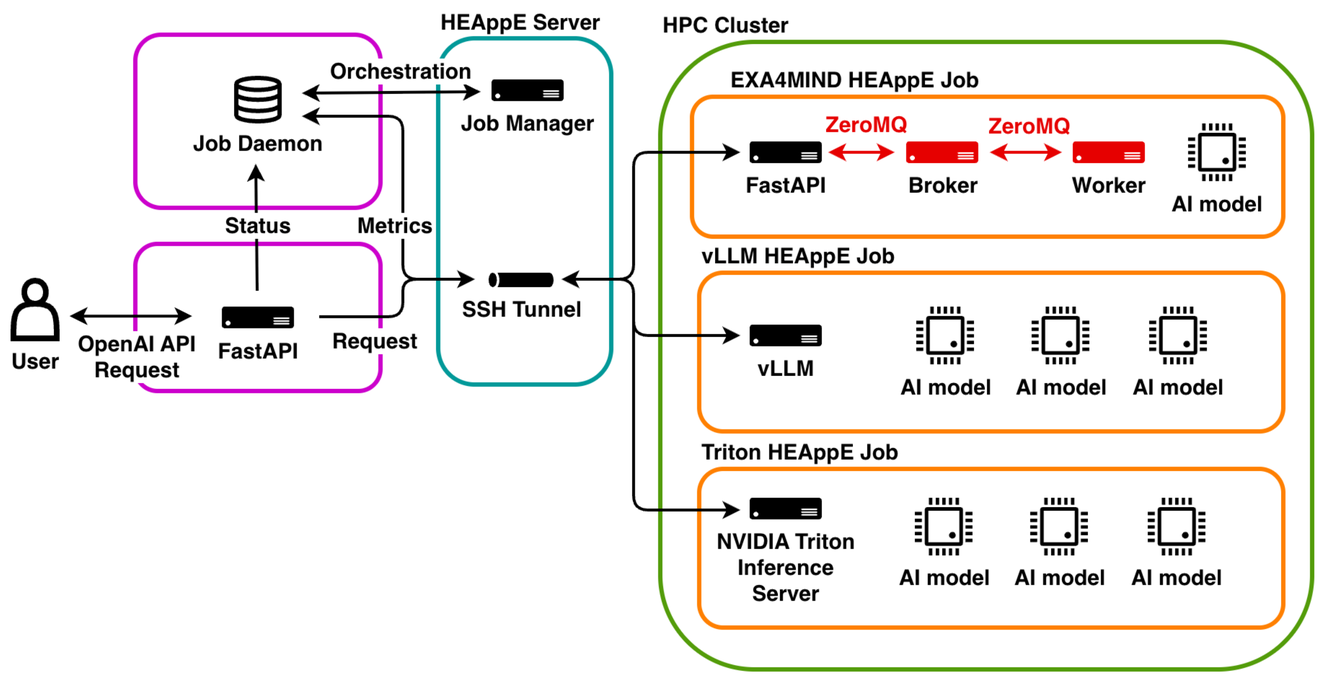

Figure 1: EXA4MIND AI Inference Service Architecture

Results

We tested the sever with **HuggingFace Inference benchmarker**[1] on LUMI supercomputer running Qwen2.5-Coder-7B-Instruct in two settings increasing the number of **Queries per Second (QPS)**:

1. vLLM on baremetal compute node
2. vLLM on compute node queried with AI inference service

The measured results show:
- **End to end (E2E) latency** overhead of 12% to 23%
- **Time to First Token (TTFT)** is large due to HTTP overhead
- Overall solution **scales well up to limits of GPU** similarly to baremetal

QPS	TTFT (P90) (ms)	E2E (P90) (ms)	Throughput (tokens/s)
0.50	68.36 1365.65	3334.70 3758.64	91.00 87.21
0.75	65.67 1503.20	3607.30 4166.15	139.14 129.12
1.12	67.68 1793.84	4049.61 4981.15	202.34 178.01
1.68	73.65 1951.60	4958.47 5707.67	311.66 233.18
2.53	73.26 2674.75	5409.09 6588.92	471.48 401.44
3.79	76.80 3432.89	6224.82 7588.16	704.91 628.73
5.69	81.98 3393.31	7909.88 9036.41	993.93 880.58

Table 1: Benchmark results. Top option 1., bottom option 2.

Conclusion

By unifying the resource predictability and high utilisation of traditional HPC environments with the agility and ease of access typical of cloud services, this architecture **eliminates operational barriers for AI users** while **adhering to the strict scheduling and governance** requirements of supercomputing facilities. It enables HPC centres to deliver **reliable, production-grade AI inference at scale**, positioning them at the forefront of the **accelerating convergence of Cloud, AI, and HPC technologies**.

https://heappe.eu

EXA4MIND AI Service documentation https://inference.exa4mind.eu

https://exa4mind.eu

[1]https://github.com/huggingface/inference-benchmarker

This research received the support of the EXA4MIND project, funded by the European Union's Horizon Europe Research and Innovation Programme, under Grant Agreement N° 101092944. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. This work was also supported by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90254).

Funded by the European Union