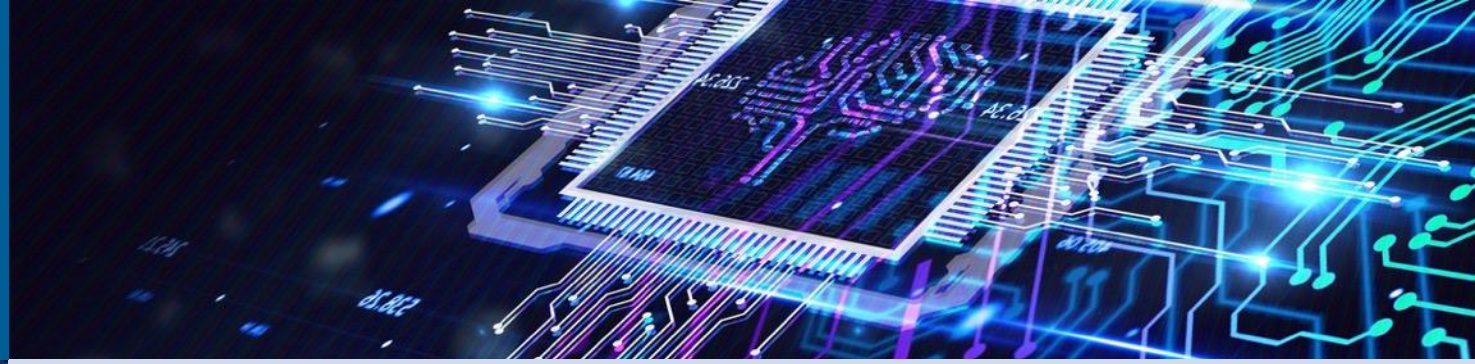
CSC

ICT Solutions for Brilliant Minds



LUMI and LLMs

25.4.2024, AIoD TCB Meeting, Markus Koskela, CSC

LUMI is an HPE Cray EX Supercomputer

LUMI

Modern architecture

LUMI-C:
x86 Partition

Supplementary CPU partition:
over 262,000
AMD EPYC CPU cores.

LUMI-K:
Container Cloud Service

LUMI-O:
Object Storage Service

30 PB
encrypted object storage (Ceph) for storing, sharing and staging data.

LUMI-Q:
Quantum Computing

High-speed interconnect

Possibility for combining different resources within a single run. HPE Slingshot technology.

LUMI-G:
GPU Partition

Sustained performance
380
Pflop/s powered by 11 912 AMD Radeon Instinct™ MI250X GPUs.

LUMI-D:
Data Analytics Partition

Interactive partition with
32 TB
of memory and graphics GPUs for data analytics and visualization.

LUMI-F:
Accelerated Storage

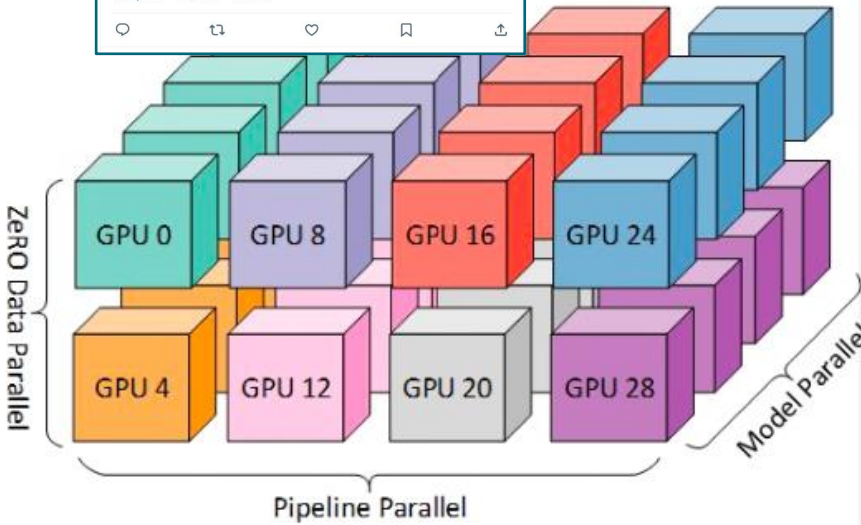
10 PB
Flash-based storage layer with extreme I/O bandwidth of 2 TB/s and IOPS capability.

LUMI-P:
Lustre Storage

80 PB
parallel file system.

LLM training on LUMI

- **Possible** as there are AMD/ROCm versions of Pytorch and other relevant libraries
- **Challenging** as the models don't fit into a single GPU memory
- Small and medium sized model training and model fine-tuning rather easy
  - Pytorch DDP
  - HF Accelerate
- Large-scale pre-training is more complicated
  - Megatron-LM, Megatron-Deepspeed
  - FSDP



Poro 34B and the Blessing of Multilinguality

Risto Luukkonen[1,2] Jonathan Burdge[2] Elaine Zosa[2]
Aarne Talman[2,3] Ville Komulainen[1] Väinö Hatanpää[4]
Peter Sarlin[2] Sampo Pyysalo[1]

[1] TurkuNLP Group, University of Turku [2] Silo AI
[3] University of Helsinki [4] CSC – IT Center for Science

risto.m.luukkonen@utu.fi jonathan.burdge@silo.ai
peter@silo.ai sampo.pyysalo@utu.fi

Abstract

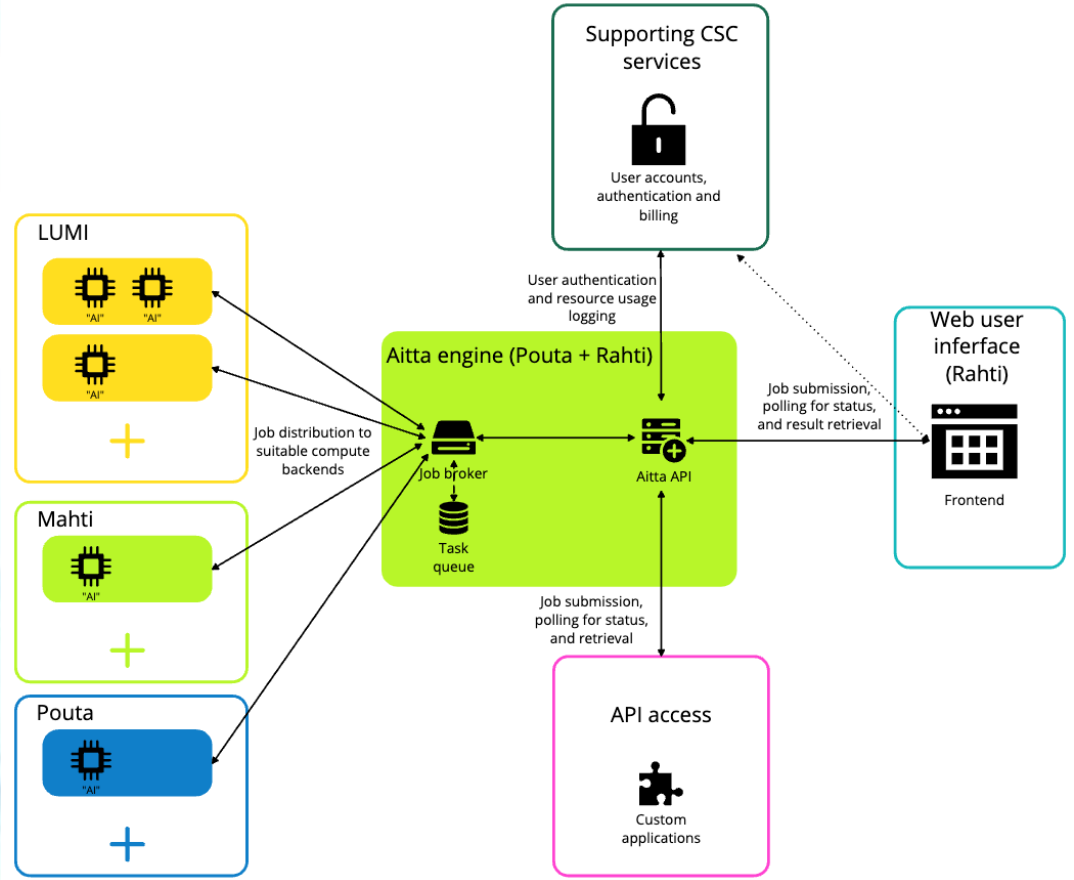
The pretraining of state-of-the-art large language models now requires trillions of words of text, which is orders of magnitude more than available for the vast majority of languages. While including text in more than one language is an obvious way to acquire more pretraining data, multilinguality is often seen as a curse, and most model training efforts continue to focus near-exclusively on individual large languages. We believe that multilinguality can be a blessing and that it should be possible to substantially improve over the capabilities of monolingual models for small languages through multilingual training. In this study, we introduce Poro 34B, a 34 billion parameter model trained for 1 trillion tokens of Finnish, English, and programming languages, and demonstrate that a multilingual training approach can produce a model that not only substantially advances over the capabilities of existing models for Finnish, but also excels in translation and is competitive in its class in generating English and programming languages. We release the model parameters, scripts, and data under open licenses at https://huggingface.co/LumiOpen/Poro-34B.

Open Language Model: OLMo

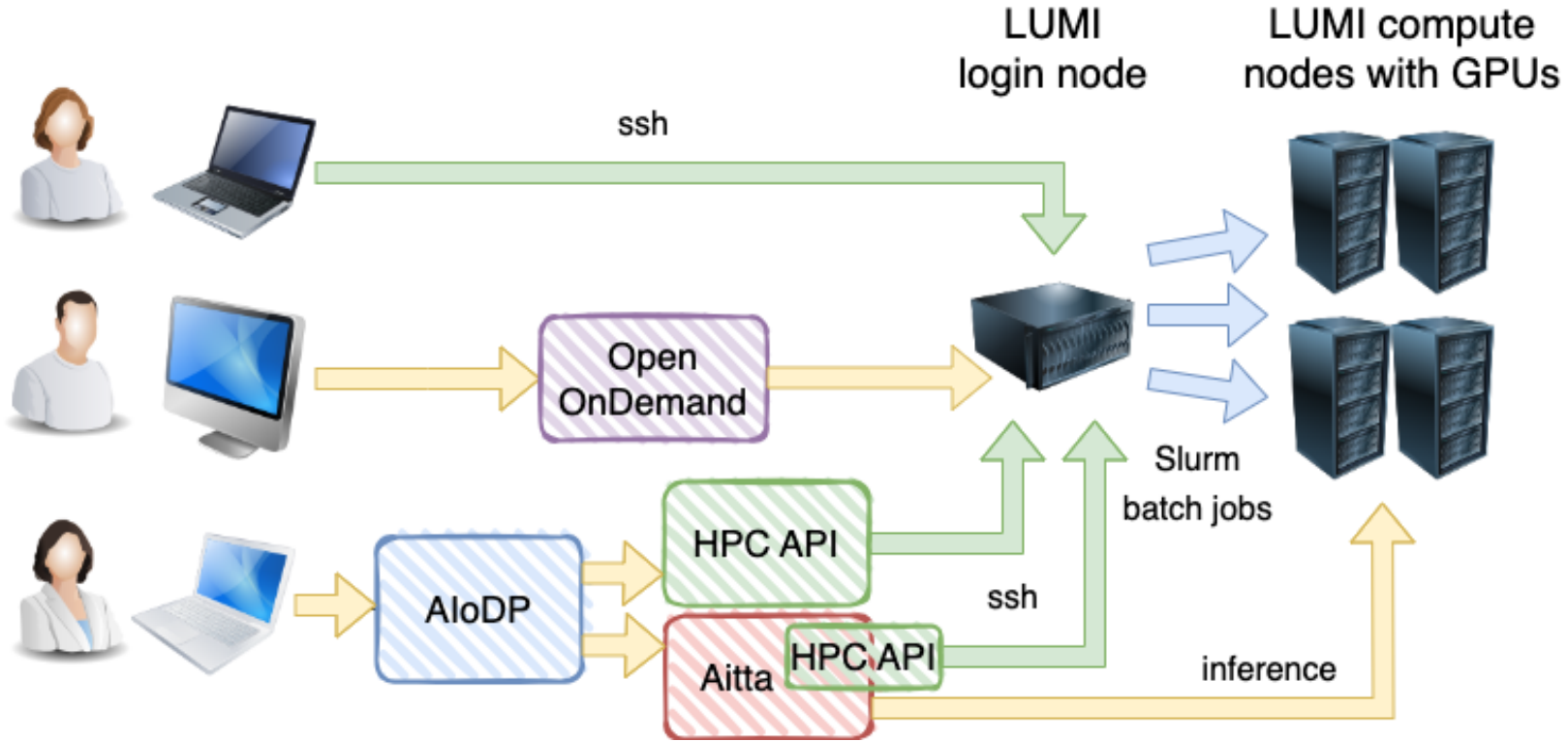
State-of-the-Art, Truly Open LLM and Framework

Aitta inference platform

- Scalable AI inference platform for R&D purposes
- Web UI and API
- Scalable, distributed architecture
- Models run in various backends depending on the need
  - EuroHPC supercomputer LUMI
  - National supercomputer Mahti
  - IaaS cloud, container cloud



Access to LUMI



HEAppE HPC API

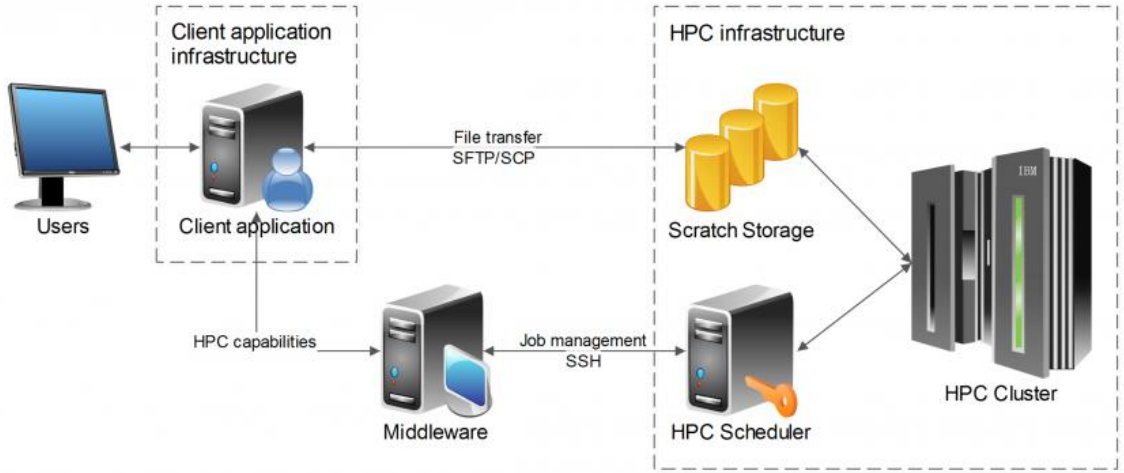


Image from: https://heappe.eu/about_heappe/

- middleware between client and HPC developed at IT4Innovations
- set up separately for each computing project; uses service accounts
- work ongoing to integrate with Aitta

Some issues to consider

- LUMI computing projects
  - EuroHPC JU / Finnish national quota
  - Open research / industry
  - AIoD computing project / separate projects?
- User accounts
  - Personal accounts (including Project manager)
  - Service accounts
- Service level and redundancy
  - LUMI is not 24/7 and has service breaks
- Sensitive and confidential data
  - LUMI is not suitable for sensitive personal data
- Types of use
  - Batch processing: model finetuning,model training, batch inference
  - Interactive usage: Jupyter, VS Code, MLflow, AI inference platform (Aitta)

Markus Koskela

D.Sc. (Tech.), Senior ML specialist
tel. +358 50 381 8676
markus.koskela@csc.fi

facebook.com/CSCfi

twitter.com/CSCfi

linkedin.com/company/csc---it-center-for-science

github.com/CSCfi