Infrastructure for AI

Staff meeting on AI

Roel Aaij

March 4th, 2026

Nikhef

At Nikhef

- Stoomboot has 20 NVIDIA L40s GPUs (48 GB)
  - 4 nodes with 4 L40 per node
  - 2 nodes with 2 L40 per node
  - Can request jobs with up to 4 GPUs
  - Sufficient to train smaller models
- Plofkip
  - Experimental server, for test cases and benchmarking
  - 2 x L40, 1 x AMD MI250x (64 GB), 4 x Intel B60 pro (24 GB)
  - Ask me or Emily for access
  - Also used to serve some LLMs
- Capibara (tentatively)
  - New server to arrive soon (coming weeks)
  - 4x AMD MI300a (512 GB total memory, 384 GB usable by GPU)
  - This is an APU, so memory bandwidth is very high (no PCIe)
  - Intended for training medium-sized models
  - will be tested for 1-2 months, then part of stoomboot

At Snellius

- Dutch HPC
- Sizeable GPU partitions
  - 352 H100 (94 GB per GPU), 4 per node
  - 288 A100 (40 GB per GPU), 4 per node
- Fast interconnect allows multi-node training
- Need to request an allocation; we can help you with this
- eInfra:
  - test, show things work and show things scale
  - ~always approved within a week
  - 1 hour of work
  - 1M SBU: ~5200 H100 hours OR 7800 A100 hours
- Rekentijd:
  - ~1-2 day(s) of work to get up to (rekentijd)
  - >95% approval probability, if you uses an eInfra for due diligence
  - approval takes ~2 months, but you quickly get 10% allocated
  - up to 30M SBU