AI factory hardware design

Robert Jan Schlimbach

01 May, 2025

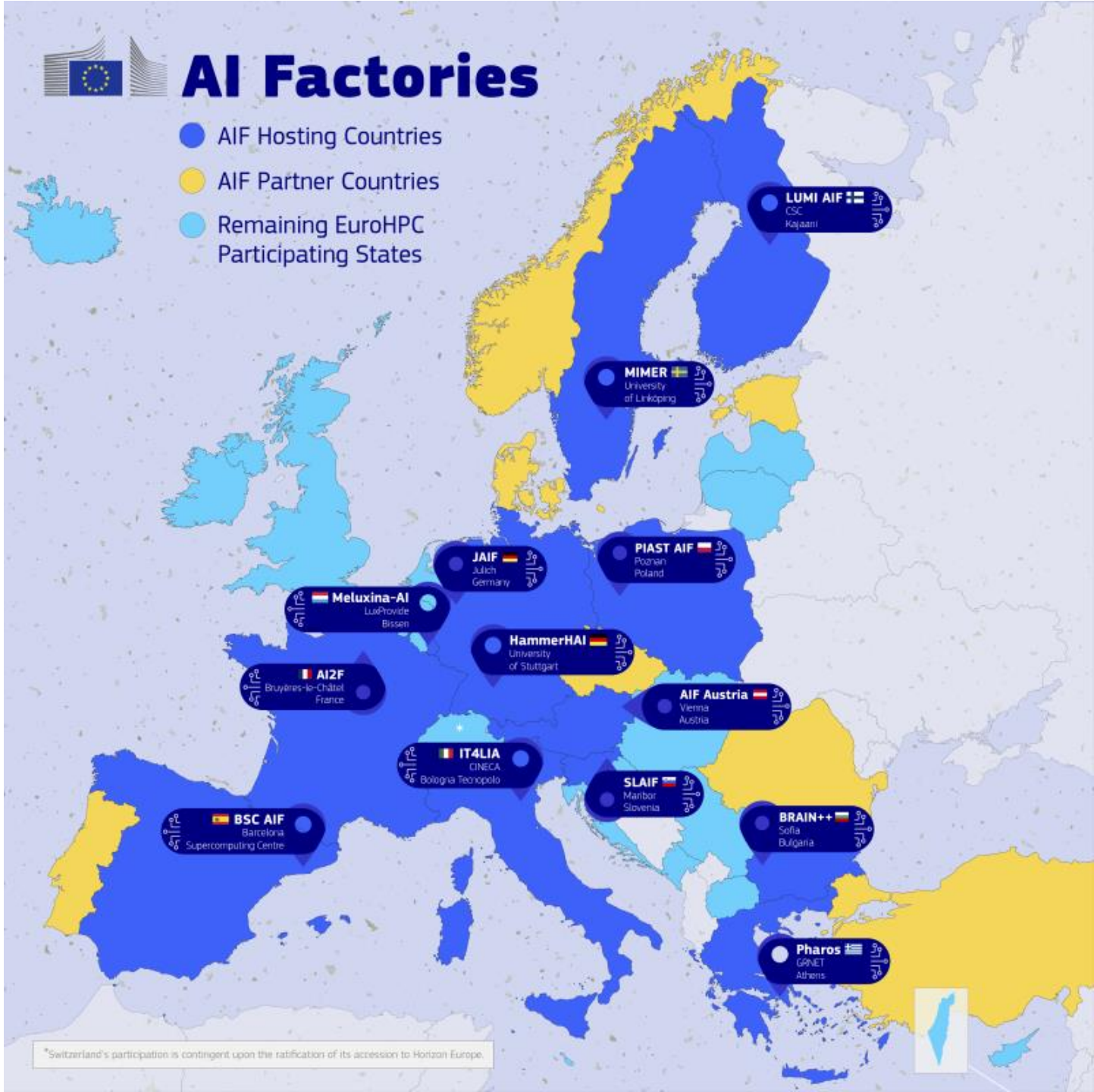
Who am I

- Computer Science + AI at UvA
- HPC + ML advisor in the HPML group at SURF
  (High-Performance Machine-Learning)
- I do optimization in the broad sense of the word
- First OS that I remember is Windows 95
  First OS that I really used was Windows XP
- First GPU was an ATi Radeon 9700 (128MB of RAM)
- First HPC system was DAS4
  (DAS 5 was already online)
- I like busy slides (but you'll fine that out soon enough)

SURF

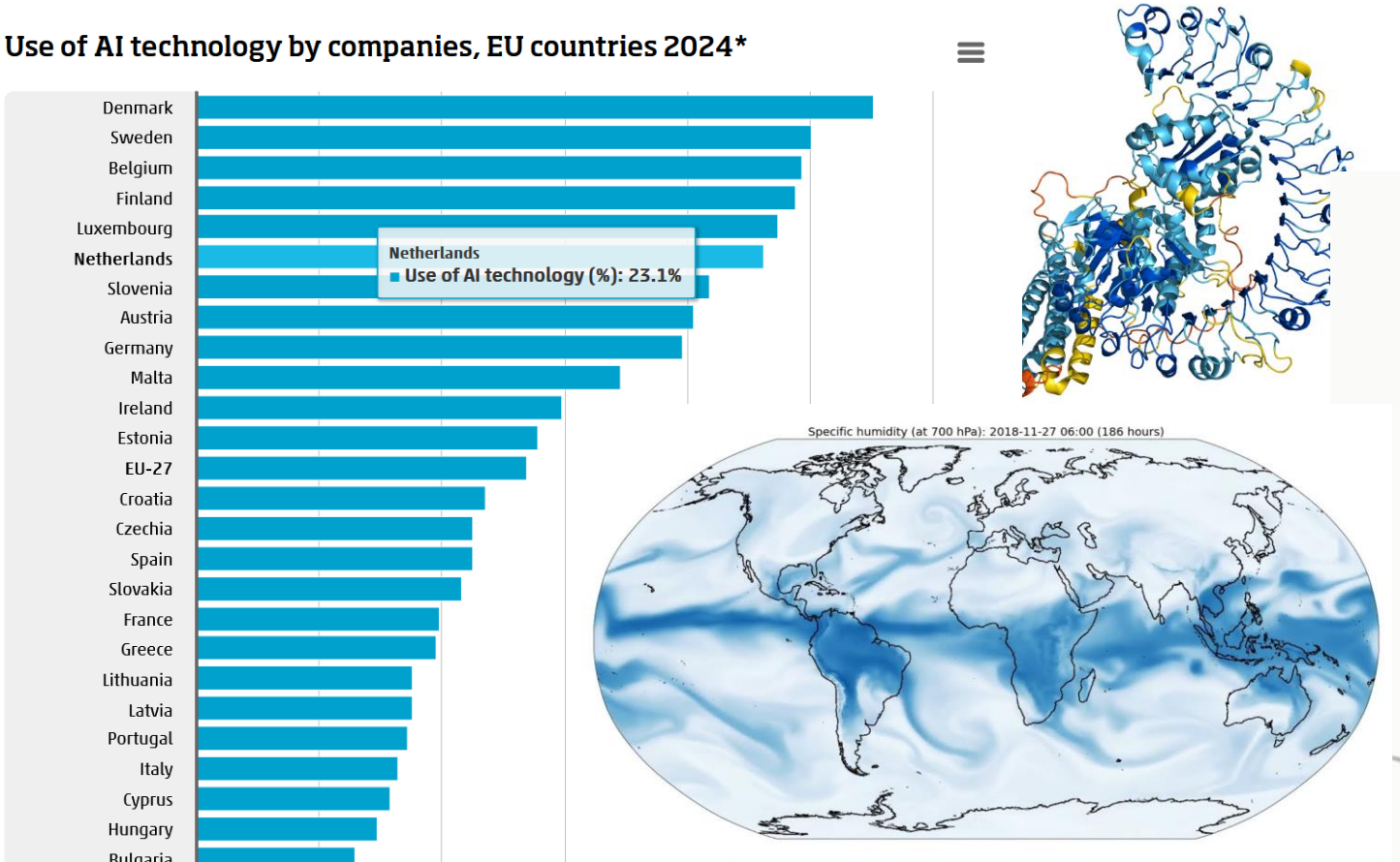
AI factories recap

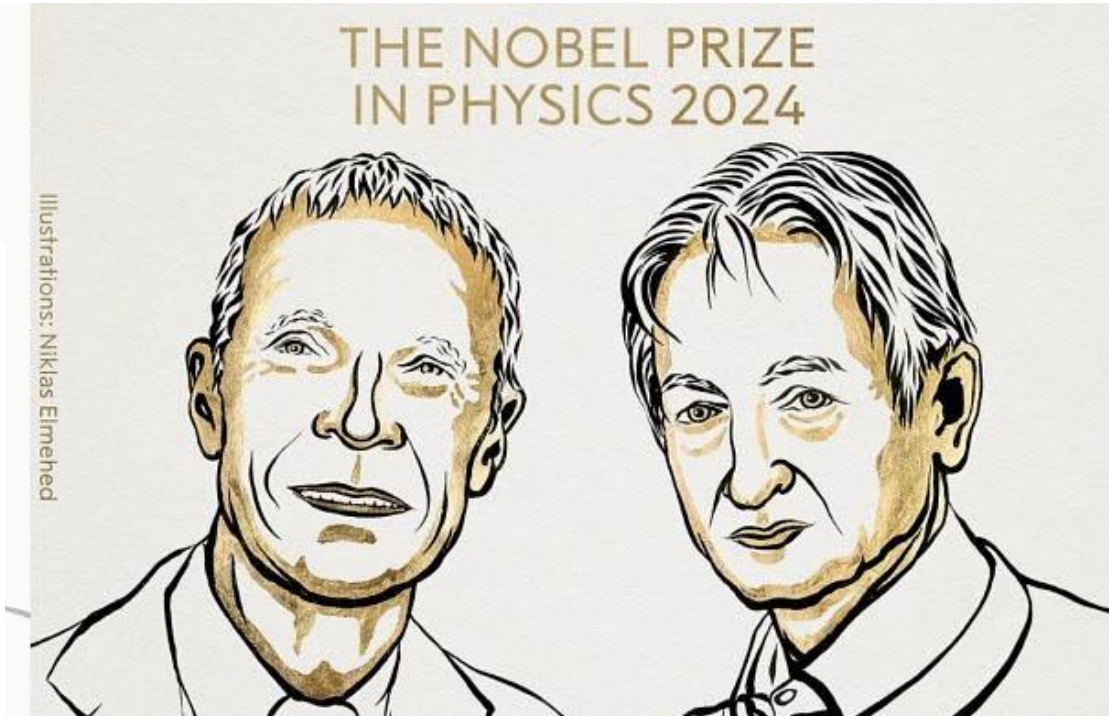
- Initiative by the EU to close the AI-gap
  to our ‘friends’ to the west
  and to our ‘friends’ to the east
- Bootstrapping AI development
  and AI uptake in Europe
- Budget ‘fluid’; up to EC discretion
  So far ~€1.5 billion granted
  Up to €200 million matched per site
- AI Gigafactories also announced:
  ~€20 billion available
- ‘Create AI aligned with EU values’
- We want to build an AI factory,
  ready end of 2026 in The Netherlands



Do we even need an AI factory?

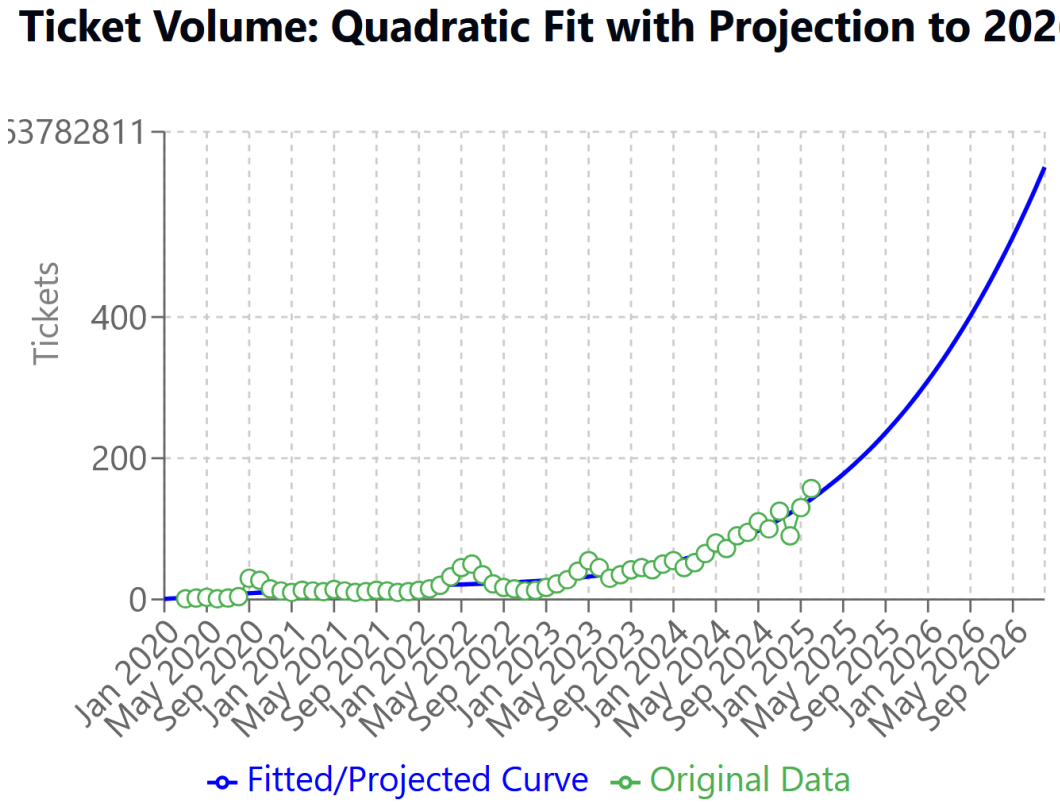
- Do I need to convince you AI is here to stay?
- Many AI workloads you can only do on big systems
- @Ana: everyone their own efficiency agent





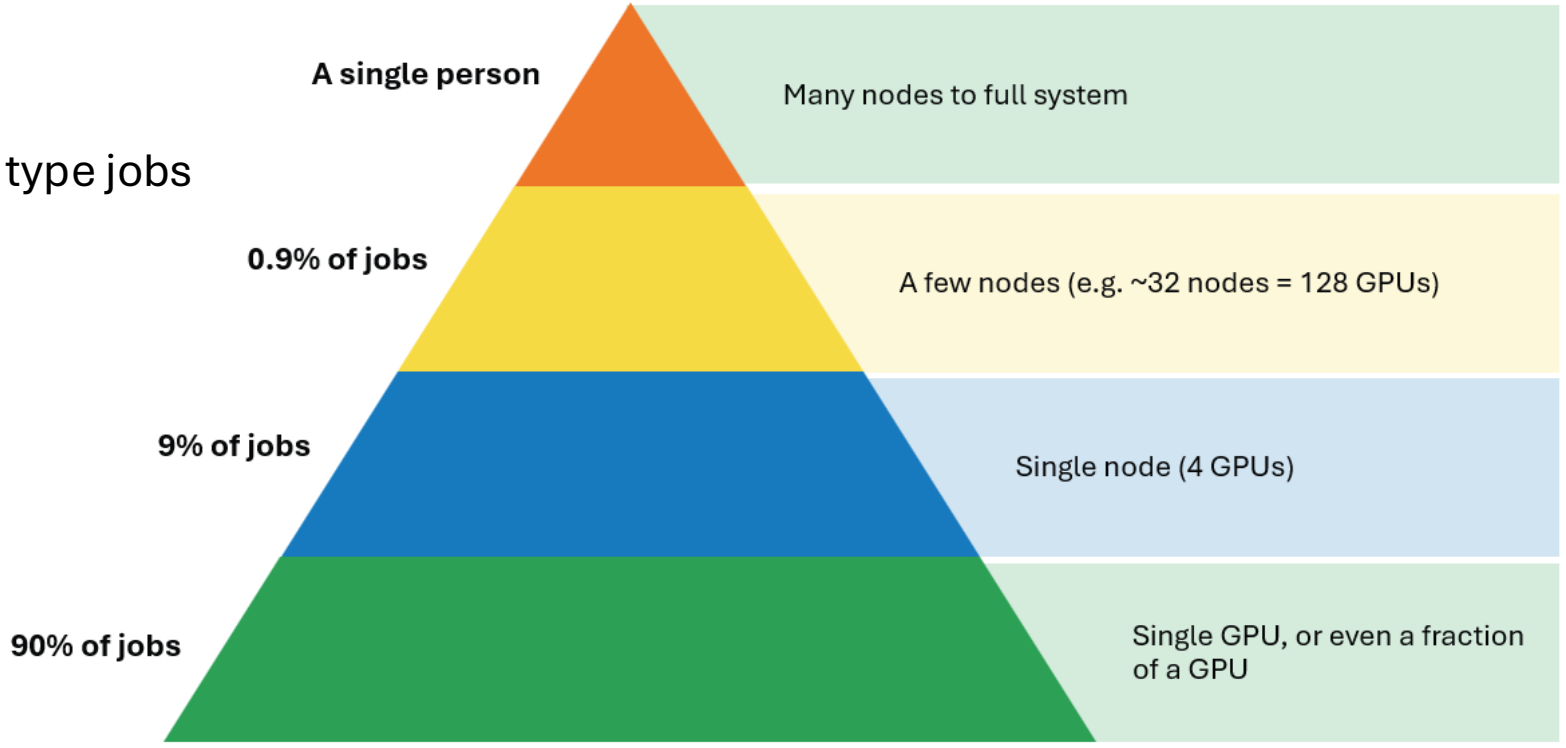
Do we even need an AI factory?

- Our perspective:
  - On average, the HPML teams gets 3-4 tickets per day regarding AI
  - Application support and compute resource requests, on average taking 1-4hrs (one person fulltime)
  - AI says we practically get infinite tickets a day at the end of 2026
    (Actually ~15, extrapolated, 5 people fulltime)
  - Not withstanding larger application support
- Industry perspective:
  - AI has firmly taken over the HPC world
  - “AI is the new HPC” ->
    AMD: “in 2025 ~30M datacenter GPUs datacenter were sold, of which ~500k for HPC applications”
- Public perspective:
  - Data and compute sovereignity is of paramount importance
  - AI factories and Gigafactories



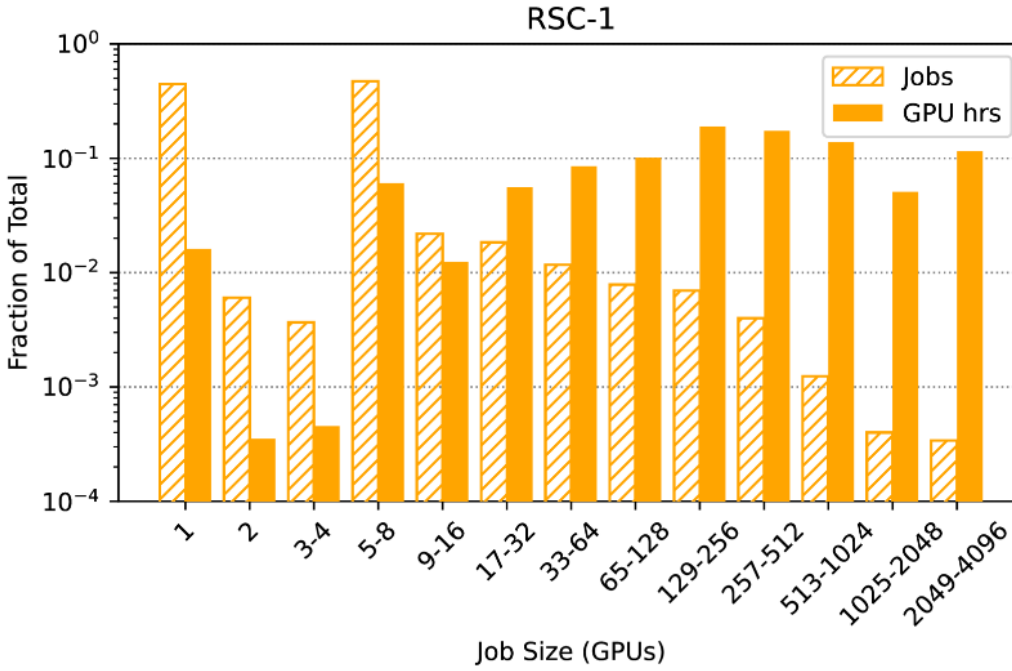
Who to design the AI factory for?

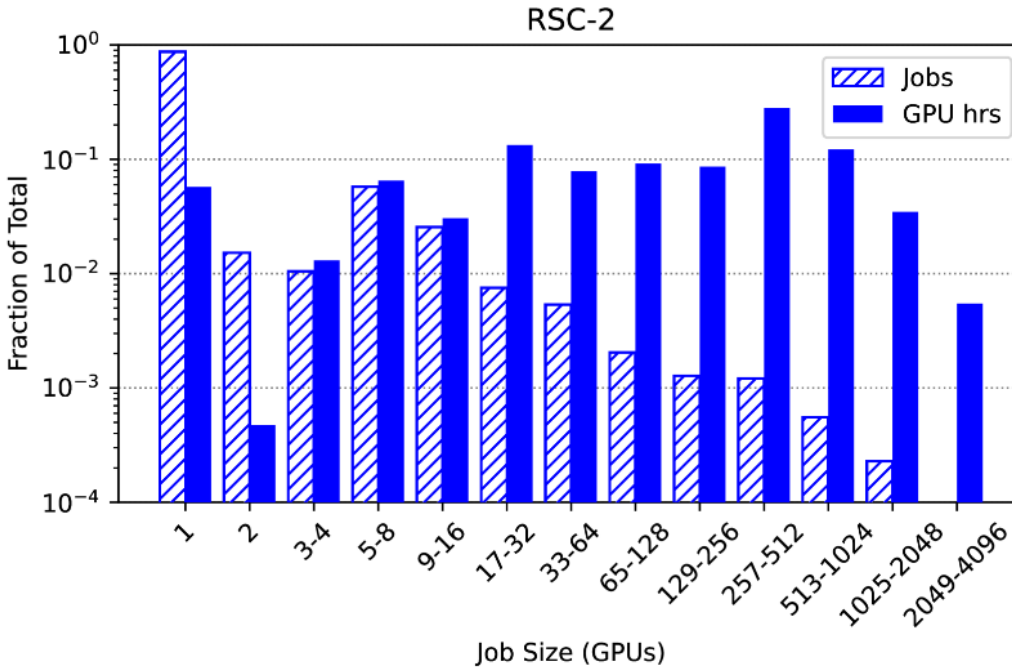
- 50/50 split between EuroHPC JU and The Netherlands
  - EuroHPC jobs almost per definition are large applications
- Dutch Research?
  - Primarily ‘just’ need many GPUs, less emphasis on large jobs
- SME?
  - Data security very important
  - More inference / always online type jobs
- Public?
  - Again, data security
  - Small jobs?



Who to design the AI factory for?

- Again, what to optimize for?
- Do we want/need 10 vans or 1 Ferrari?
- Critical scale for certain training runs (‘small’ Large-LMs)
  - Probably in the order of ~2k GPUs IFF tightly coupled (https://arxiv.org/pdf/2412.19437)





Source: https://arxiv.org/html/2410.21680v1

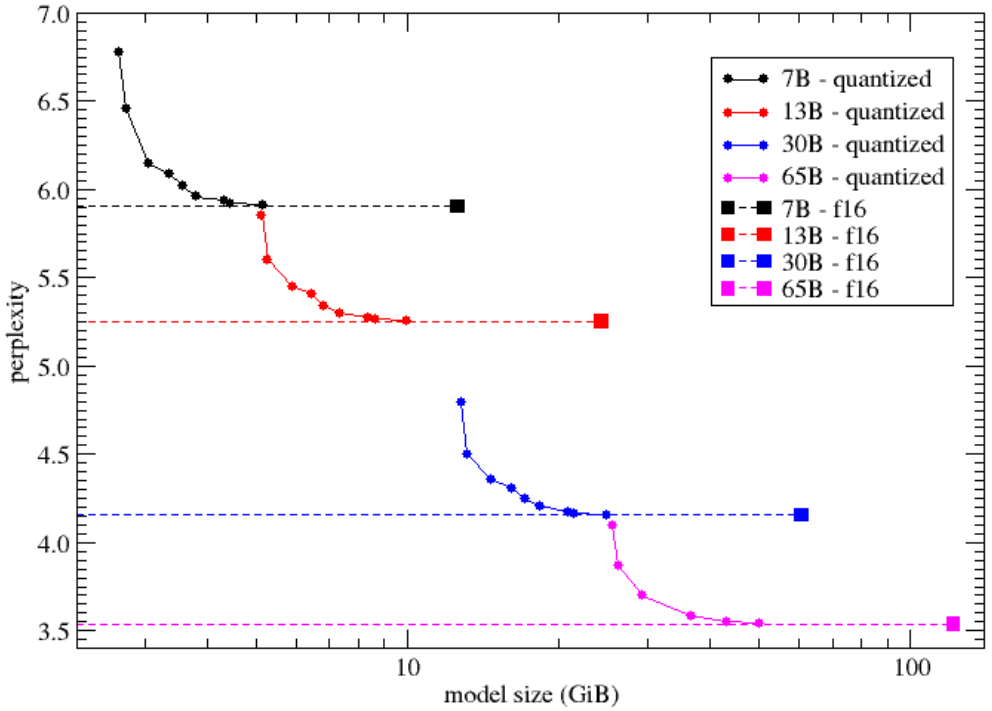
Critical innovations

- Reduced precision (FP64 is dead)
  - double whammy of increased FLOPS and reduced memory footprint (many workloads are membw bound)

Technical Specifications

	DGX GB200
FP4 AI	1,440 PFLOPS
FP8 AI	720 PFLOPS
FP16 AI	360 PFLOPS

- PODs
  - order-of-magnitude faster interconnect than current interconnect
  - Enables model sharding (FSDP, Tensor-Parallel, MoE)
    Bigger models are simply better (given enough data)



So where does that leave us?

Disclaimer time:

- None of this is final
- Good ideas are always appreciated
- Fully dependent on funding

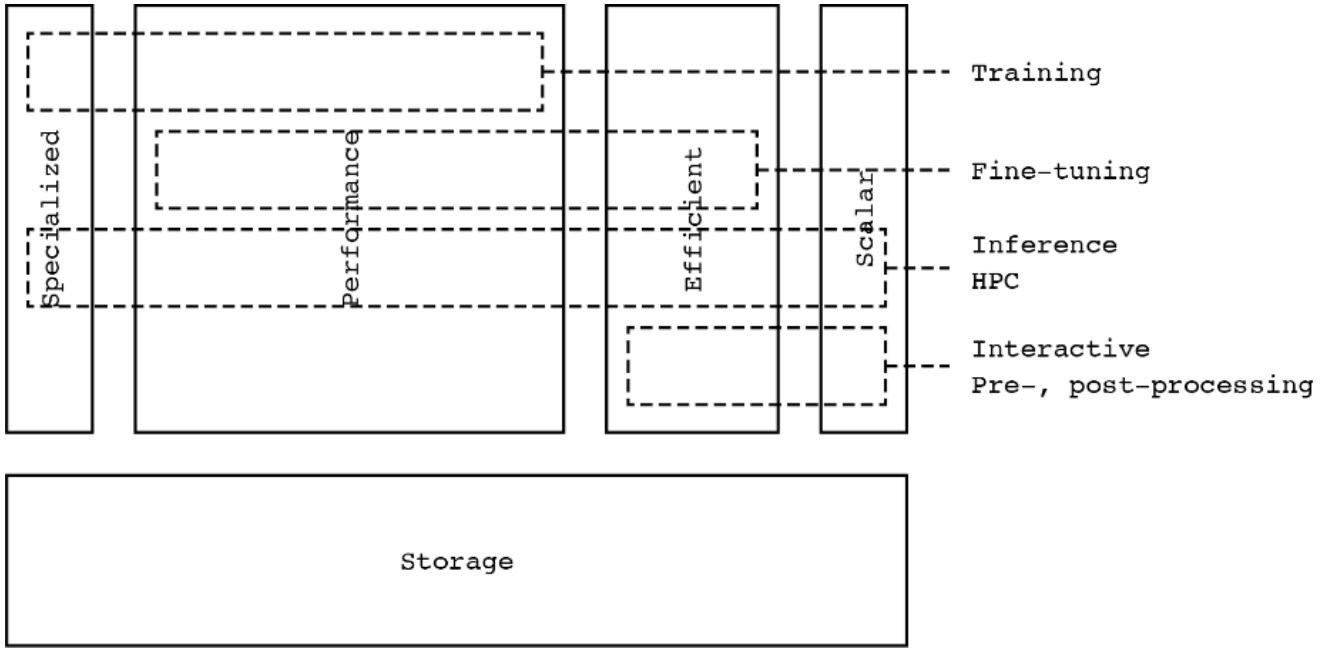
HW design TL;DR: (Goes without saying) optimize both for capability and cost-effectiveness

- Exascale 16-bit TFLOPS
- Power envelope: 5-15MW
- Compute: various accelerated partitions
- Storage: tiered
- Network: as fast as possible when necessary

SURF

Compute

- Performance partition
  - Big GPUs, Big interconnect
    PFLOPS per GPU, 200GB+ HBM VRAM
  - Focus on training or big-model inference
  - 120-240kW+ per rack
- Efficient partition
  - Cost effective GPUs
  - No interconnect
  - Online inference
- CPU partition
  - Pre/Post processing and CPU-based ML
- Specialized partition
  - Exotic hardware for specific use cases



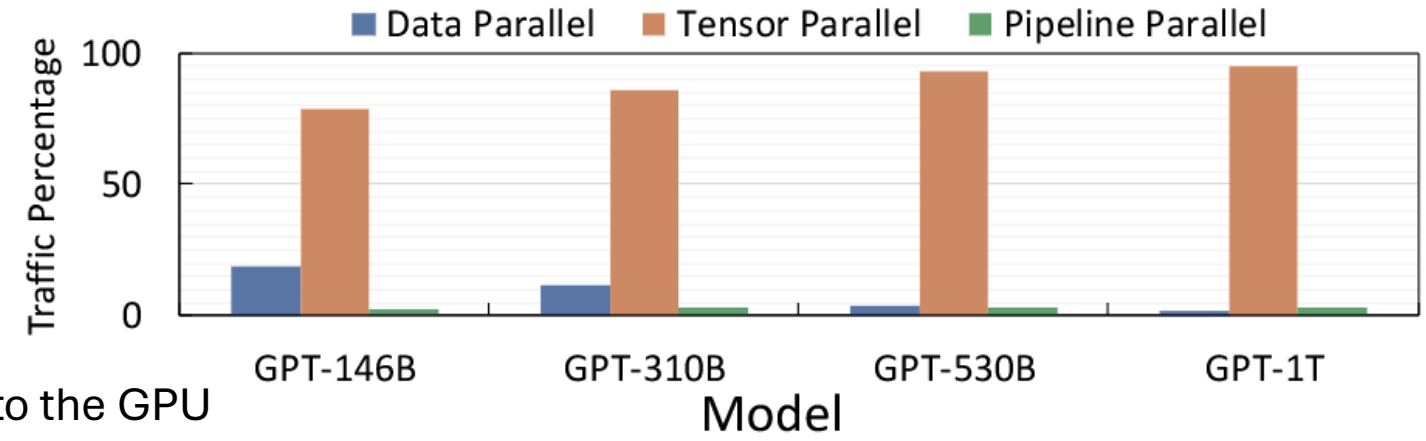
Storage

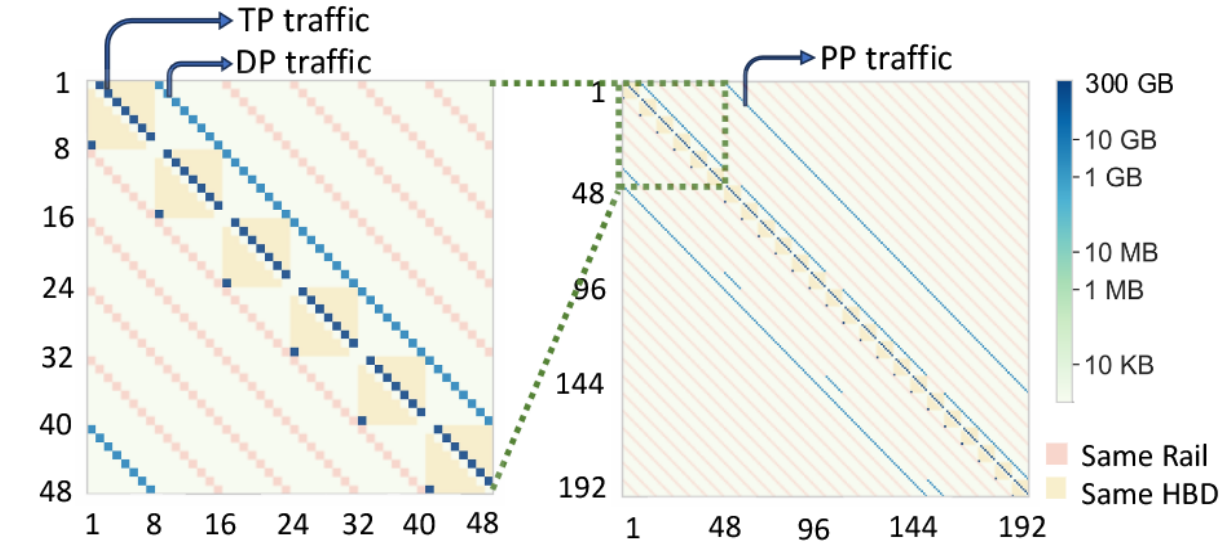
- Tiered:
  - ‘Hot’ storage: all-NVMe SSD scratch
    - ML workloads are very often continuous-load high-IOps
    - No node-local storage!
  - ‘Warm’ storage: performant spinning disk parallel fs
    - Backbone of the system
  - ‘Lukewarm’ storage: object store
    - Data ingress, and integration with cloud and other EU systems
  - ‘Cold’ storage: tape
    - Not generally available to users, but used as cost-effective backend for data storage and archival purposes

SURF

Interconnect

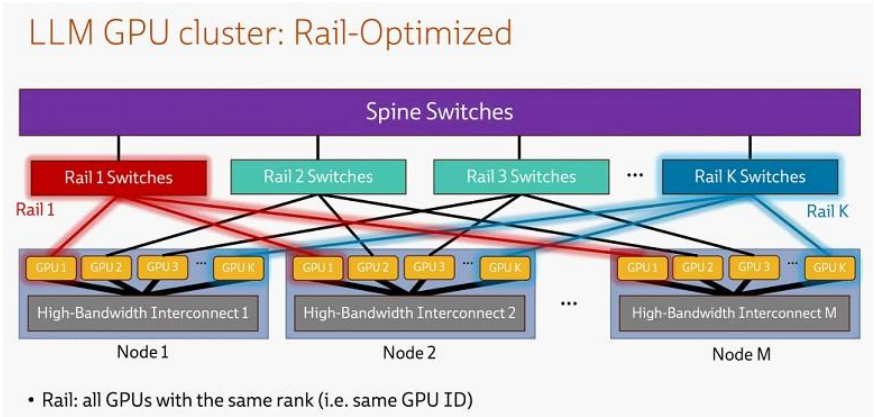
- All-copper ‘scale-up’
  - 8Tbps+ **per GPU** BW, connected directly to the GPU
  - For reference: Snellius GPU nodes = 2x200Gpbs divided over 4 GPUs
  - Single layer of switches connect up-to 256 GPUs
- Big ‘scale-out’ network is probably not necessary
  - We are too small for that
  - E.g.: single supplementary 800Gbps NIC per node for dataloading or scale-out





(a) GPT-1T MegatronLM traffic matrix GPU 1 to 48 (one pipeline stage)

(b) GPT-1T MegatronLM traffic matrix GPU 1 to 192 (four pipeline stages)



Do we even need an AI factory?

Personal opinion time

I see no way of the AI genie going back into the bottle

If we (as The Netherlands) don’t get onto the (EU) boat now, we will fade into obscurity

And there is no saying when the next boat comes

SURF

Not mentioned here at all: expertise center

- Huge part of the AI factory
- Data management, expertise centers, tooling