Distributed and Decentralised Training: Technical Governance Challenges in a Shifting AI Landscape

Jakub Kryś Yashvardhan Sharma Janet Egan

[2507.07765]

Hyperscale LLM training

- Traditionally, LLM training has taken place in huge individual clusters
- Convenient for compute governance:
  - Detectable, quantifiable, excludable, concentrated (Sastry et al., 2024)
- Not only raw FLOPS matter – fast interconnects are super important

Low-communication training algorithms

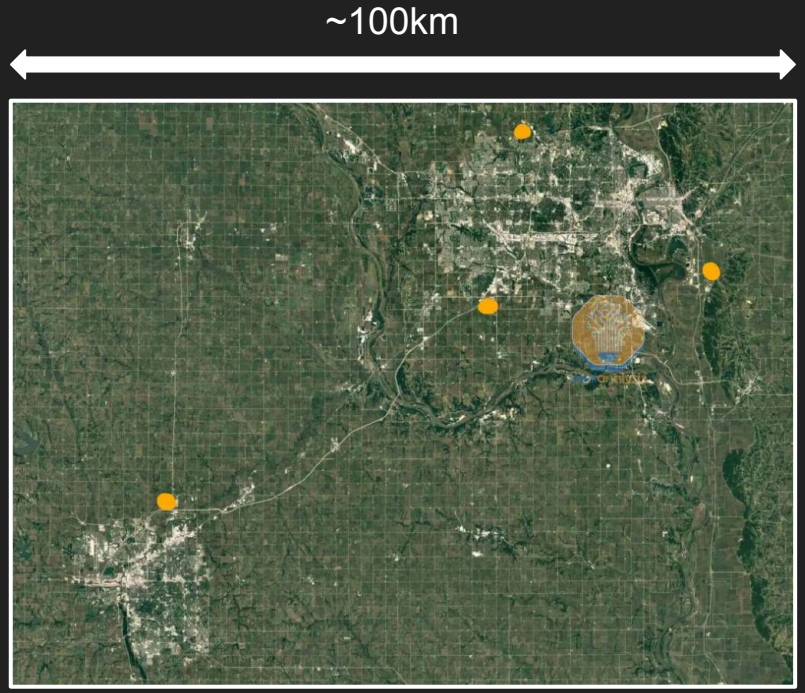
(SWARM, DiLoCo, DeMo and more)

Reduce inter-GPU communication by 3 orders of magnitude or more(!)

Enable two new training paradigms: *distributed* and *decentralised*

Multi-data centre *distributed* training

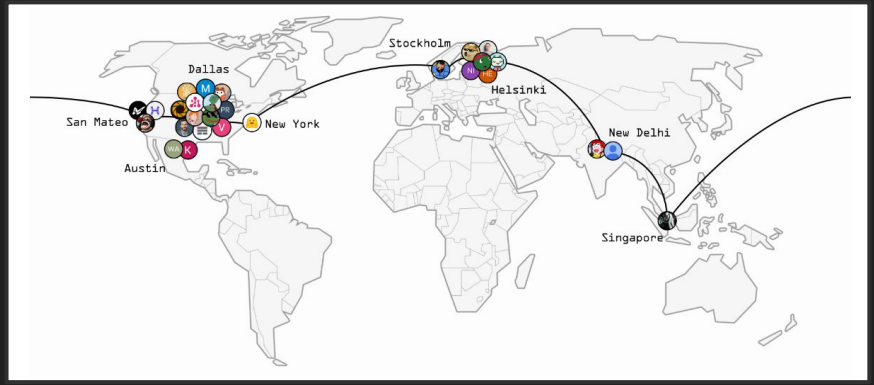
- Combining multiple clusters to orchestrate one training run
- Motivated by power constraints (?)
- GPT-4.5, Gemini-1.5, more?



Google's multi-data centre region in Iowa/Nebraska (SemiAnalysis)

Decentralised training

- Community-driven compute
- No central coordinating entity (to be achieved)
- Motivated mainly by ‘ideological’ factors (open-source, preventing power concentration, blockchain-like)
- Photon, INTELLECT-1, Consilience
  - INTELLECT-2 (post-training)



INTELLECT-1 (10B) global pre-training run (Prime Intellect)

Why does this matter?

- Hyperscalers can train across sites to reach even bigger model sizes (up to 450T parameters?)
- A broader group of actors will be able to produce (sub-frontier?) models
- The risk of compute structuring is higher → KYC schemes more important
- (speculative) Full decentralisation could lead to ungovernable, unshutdownable AIs

Open questions

- What is the true performance of distributed/decentralised models on downstream tasks?
- What is the scalability of such training?
- What is the smallest ‘useful unit of compute’ for decentralised training?
- How can decentralised training be monetised?
- To what extent should we be concerned about dangerous capability proliferation?
    - How do we balance this with concentration of power?