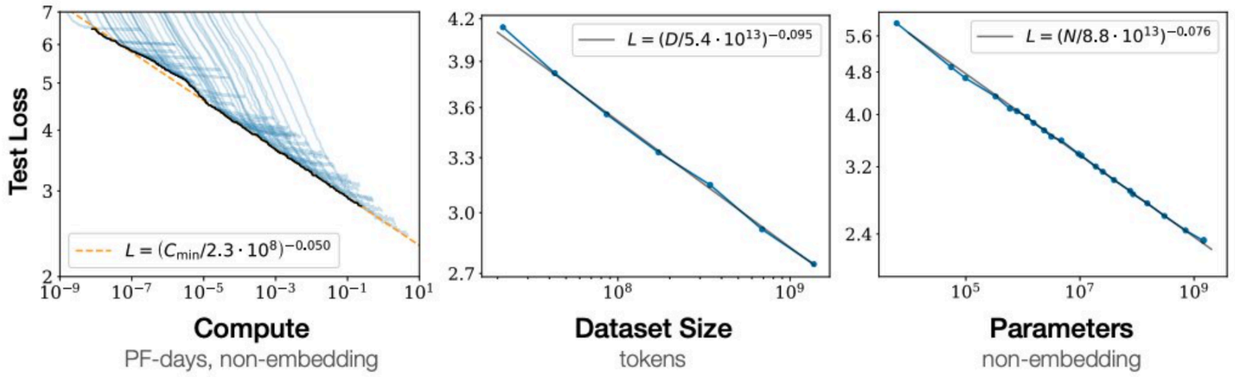
Decentralised Learning

AKA How to train really large models on consumer devices over the Internet

MIKA SENGHAAS

EPFL

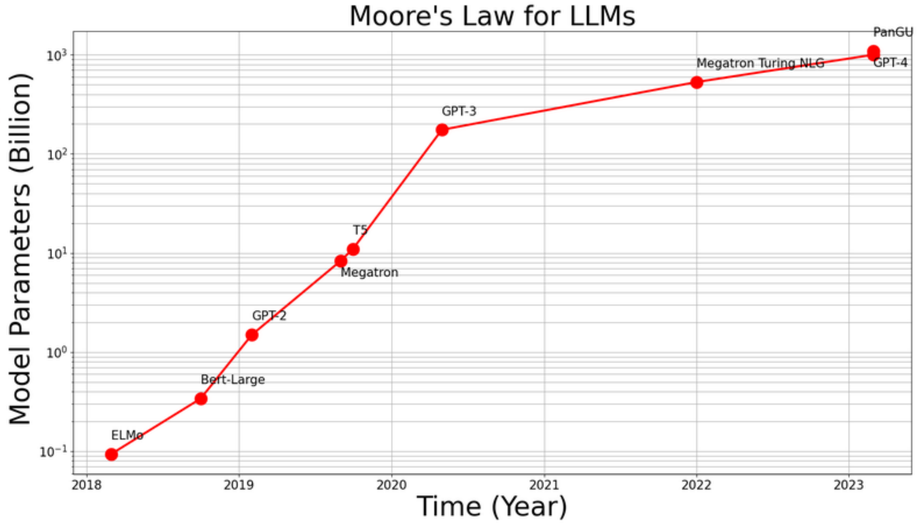
Scale



Kaplan et.al (2020),
Hoffmann et.al (2022)

We train larger models

↳ 100B+ (1000B+?)



LLMs improve as function of scale

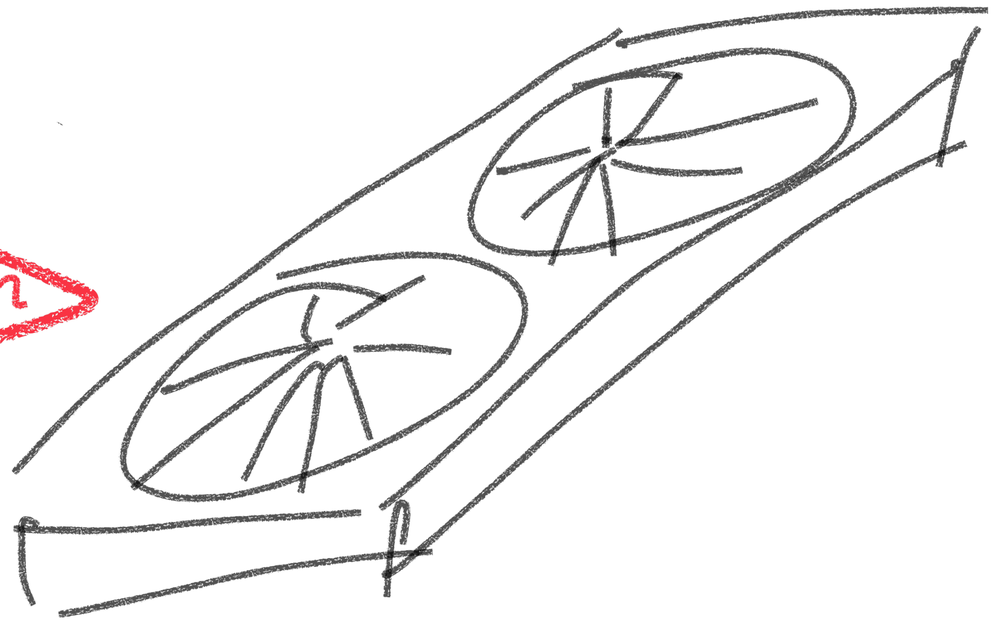
↳ predictably

Scale

10B at BF/FP16 ≈ 20GB · 3 ≈60GB+
+Activations

Llama-3.1
405B

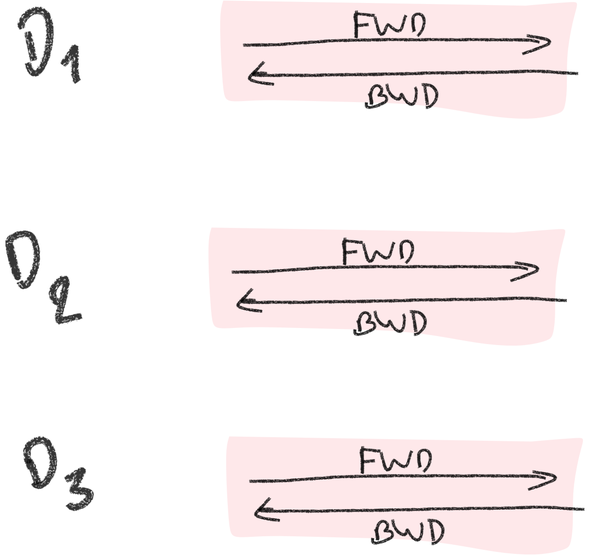
Doesn't fit



A100 / H100

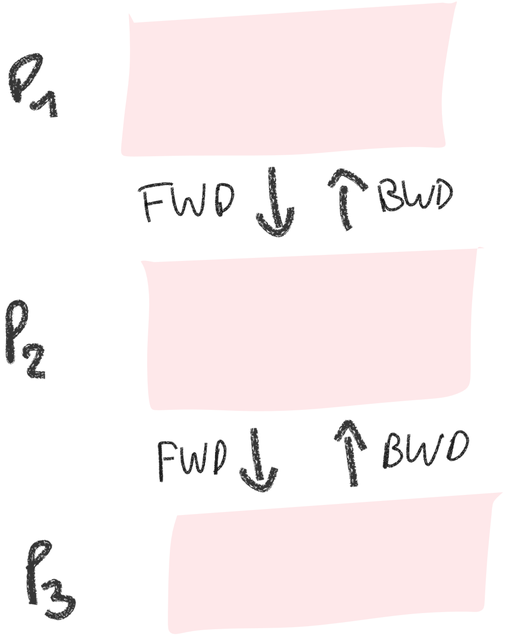
But really
anything >10B

Parallelism Paradigms



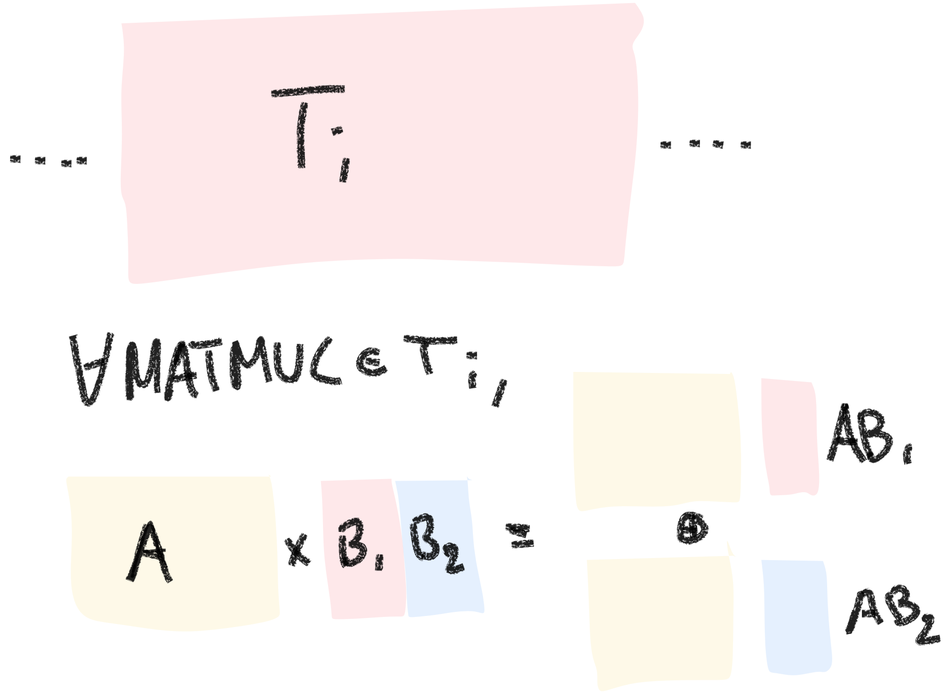
Data Parallel

DP



Pipeline Parallel

PP



Tensor Parallel

TP

Parallelism Paradigms

Data Parallel
DP

Pipeline Parallel
PP

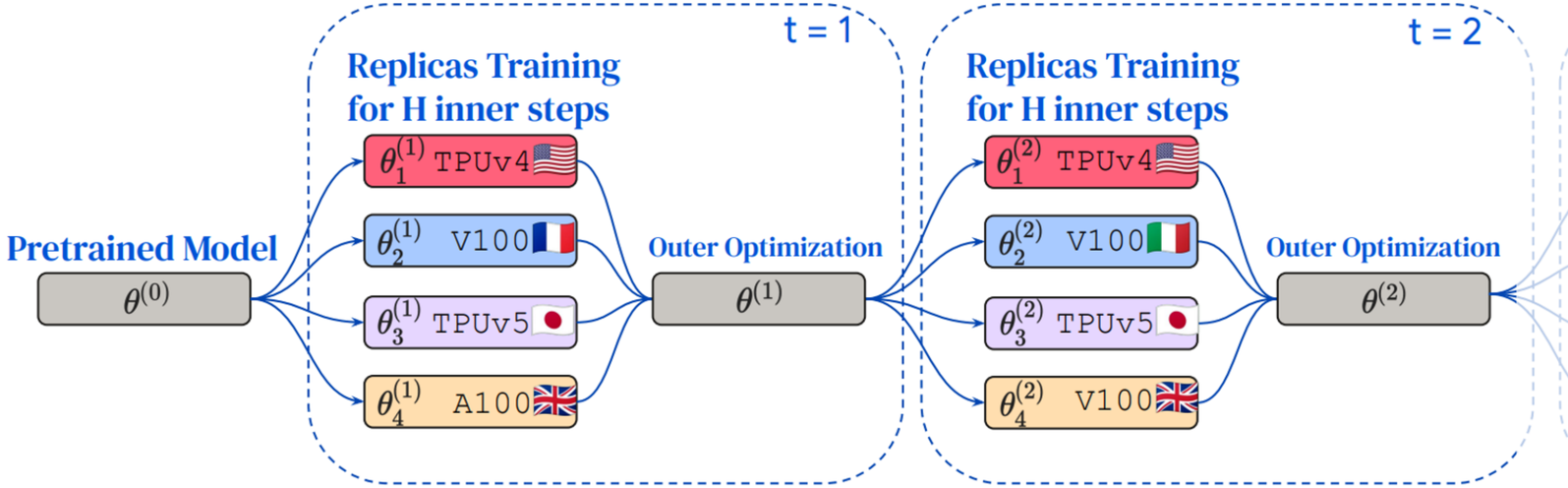
Tensor Parallel
TP

WORK IN DATA CENTRE
BUT NOT
DECENTRALISED

Decentralised Setting

1. Heterogeneous Hardware
2. Heterogeneous Network
3. Reliability
4. Scale

DiLoCo (Douillard, 2023)



Low-Bandwidth Data Parallel

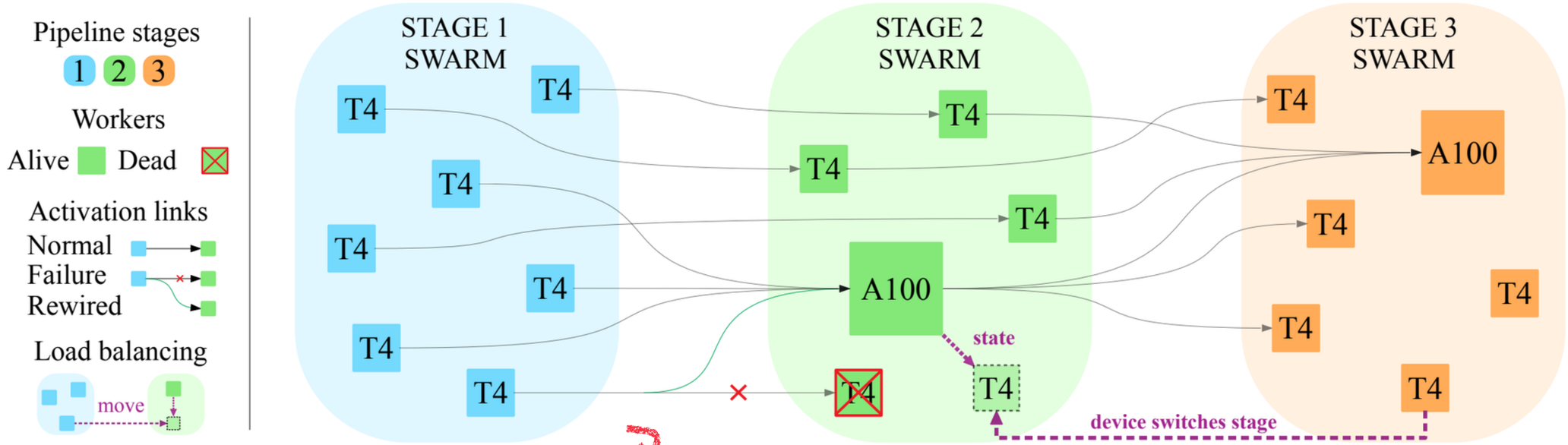
⊕ Works for islands of data centres

⊕ Open source

⊖ Model has to fit in min. GPU mem

⊖ Synchronous global update (≠ heterogeneity)

SWARM (Ryabinin et. al, 2023)



"Dynamic pp"

Stochastic Wiring

Adaptive Swarm Balancing

⊕ Adaptive + suits heterogeneity

Research Directions

1. Benchmarking
2. Battle-test SWARM
3. Async DiLoCo
4. Multi-parallel Training