
PyTorch FSDP

Fully Sharded Data Parallel

Yanli Zhao, et al. Meta AI

Presenters: Jaeyeon Kim, Kayden Kehe, Theo
Lebryk, and Aditya Palaparthi

Presentation Advisors: Qitong Wang and
Konstantinos Kopsinis

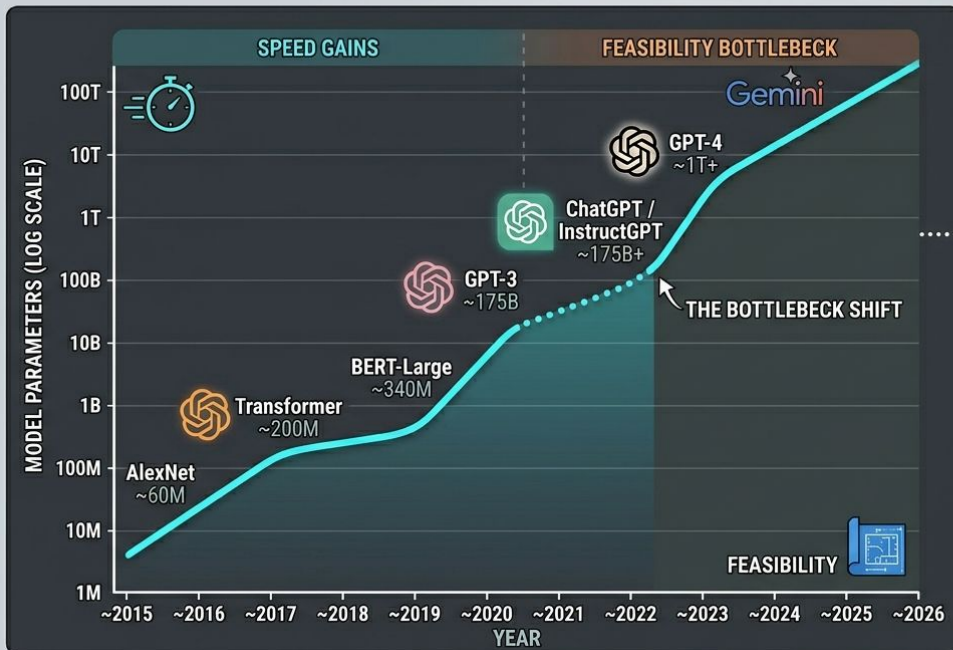


Why We Need a New System



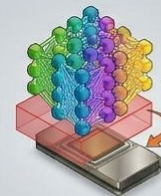
Why Model Scaling Matters

FROM SPEED TO FEASIBILITY: THE RISE OF DISTRIBUTED TRAINING

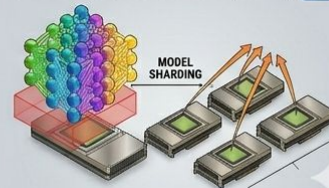


DISTRIBUTED TRAINING IS NOW ABOUT FEASIBILITY, NOT JUST SPEED.

PROBLEM: MODELS > SINGLE GPU HBM



SOLUTION: DISTRIBUTED SHARDING & ASSEMBLY (FSDP)



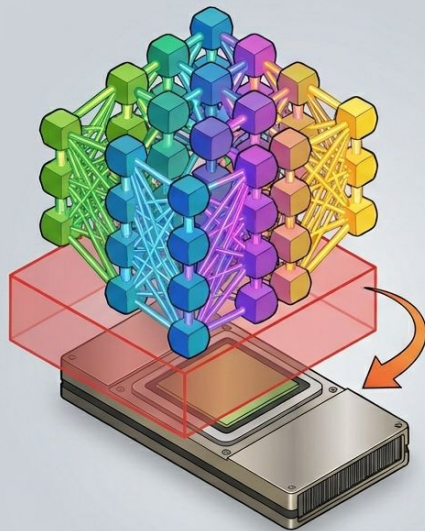
The slide features decorative hexagonal shapes in the corners. The top-left corner has a cyan hexagon overlapping a light blue one. The top-right corner has a light blue hexagon overlapping a medium blue one. The bottom-left corner has a medium blue hexagon overlapping a cyan one. The bottom-right corner has a cyan hexagon overlapping a light blue one.

Discussion

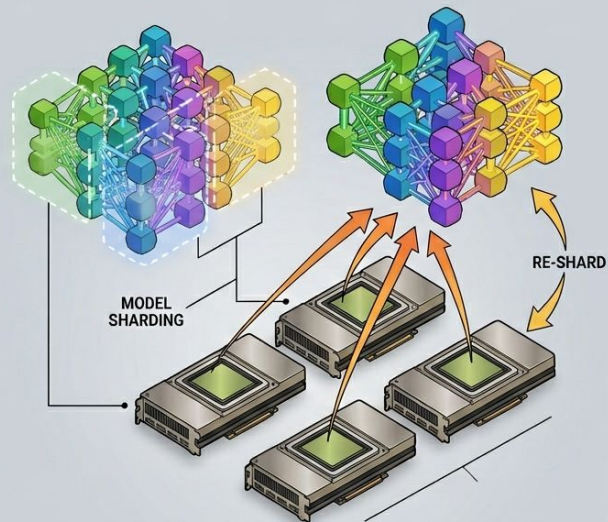
What do you think breaks first when models scale – compute, memory, or communication?

Big Picture

THE PROBLEM



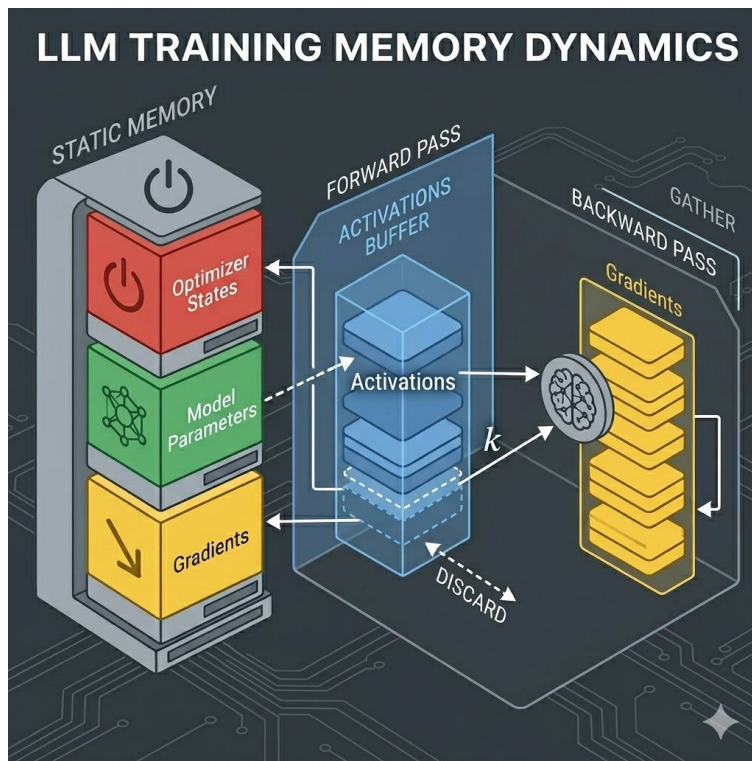
THE SOLUTION



Why Do Existing Methods Break?

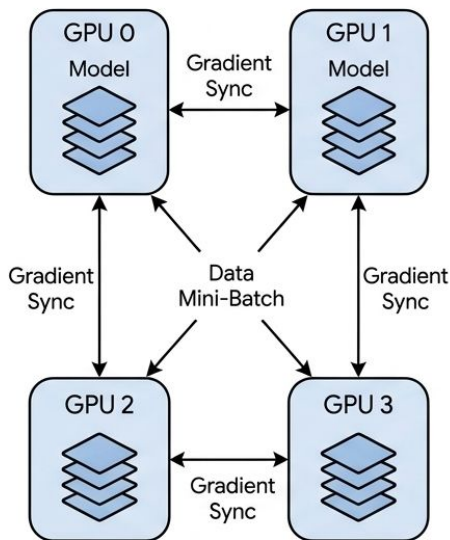


Recap on Deep Learning Training



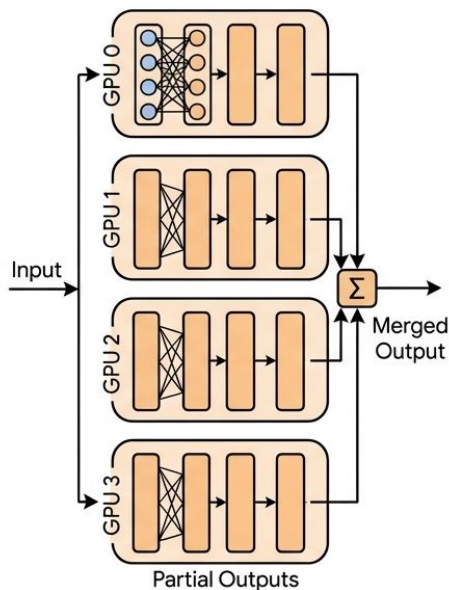
Previous Approaches Break

Data Parallelism (DP)



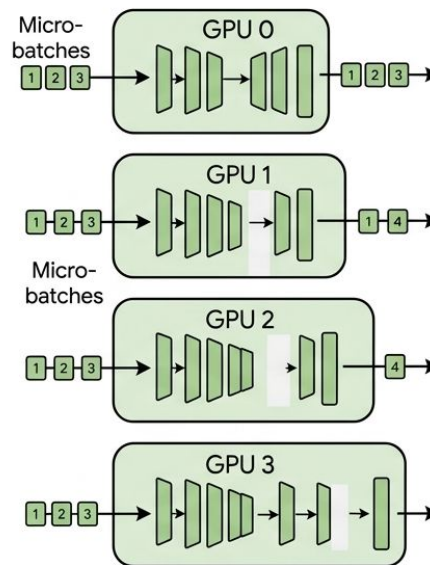
Replicate model, split data

Tensor Parallelism (TP)



Split one layer across GPUs

Pipeline Parallelism (PP)



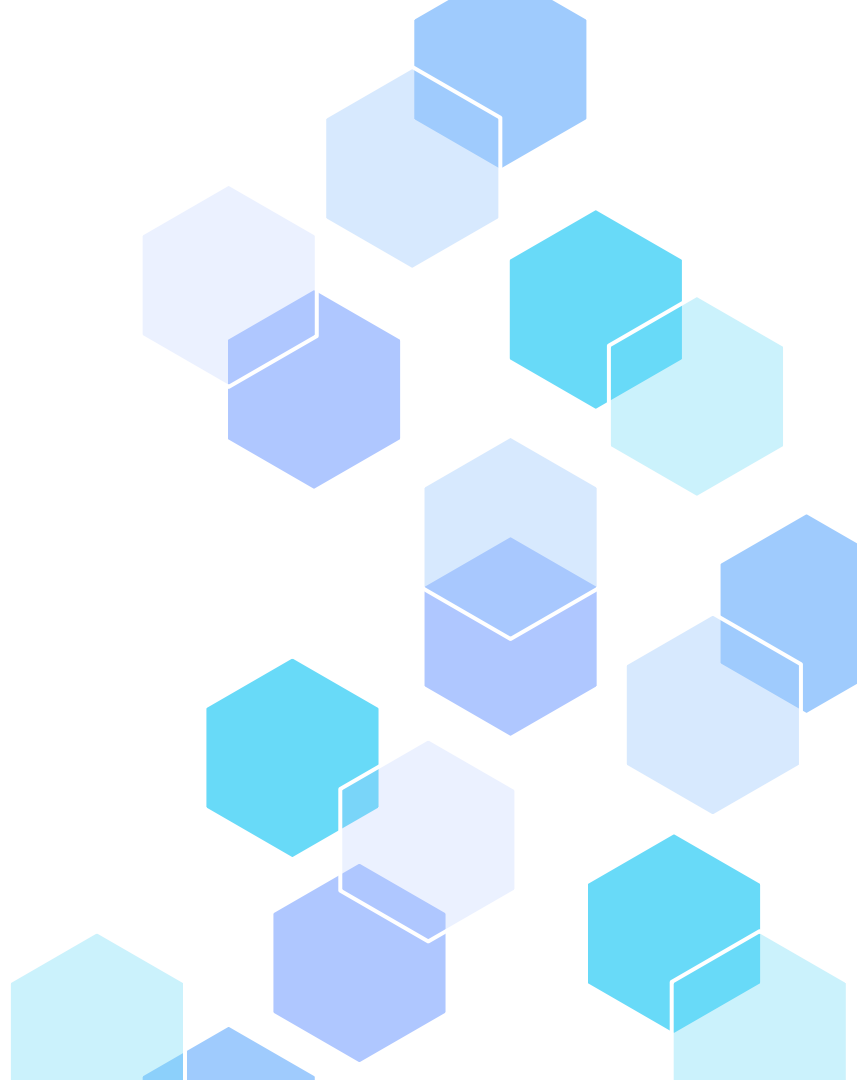
Split layers by stage

The slide features decorative hexagonal shapes in the corners. The top-left and bottom-right corners have overlapping cyan and light blue hexagons. The top-right and bottom-left corners have overlapping light blue and cyan hexagons. The word "Discussion" is centered at the top in a bold, dark blue font.

Discussion

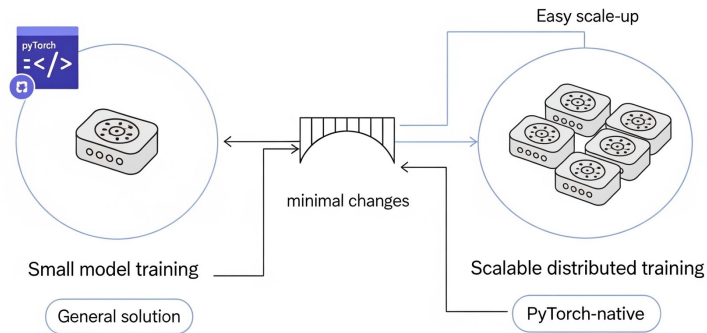
If you had to choose one of these approaches for trillion-parameter models, which would fail first and why?

What Solution Do We Need?

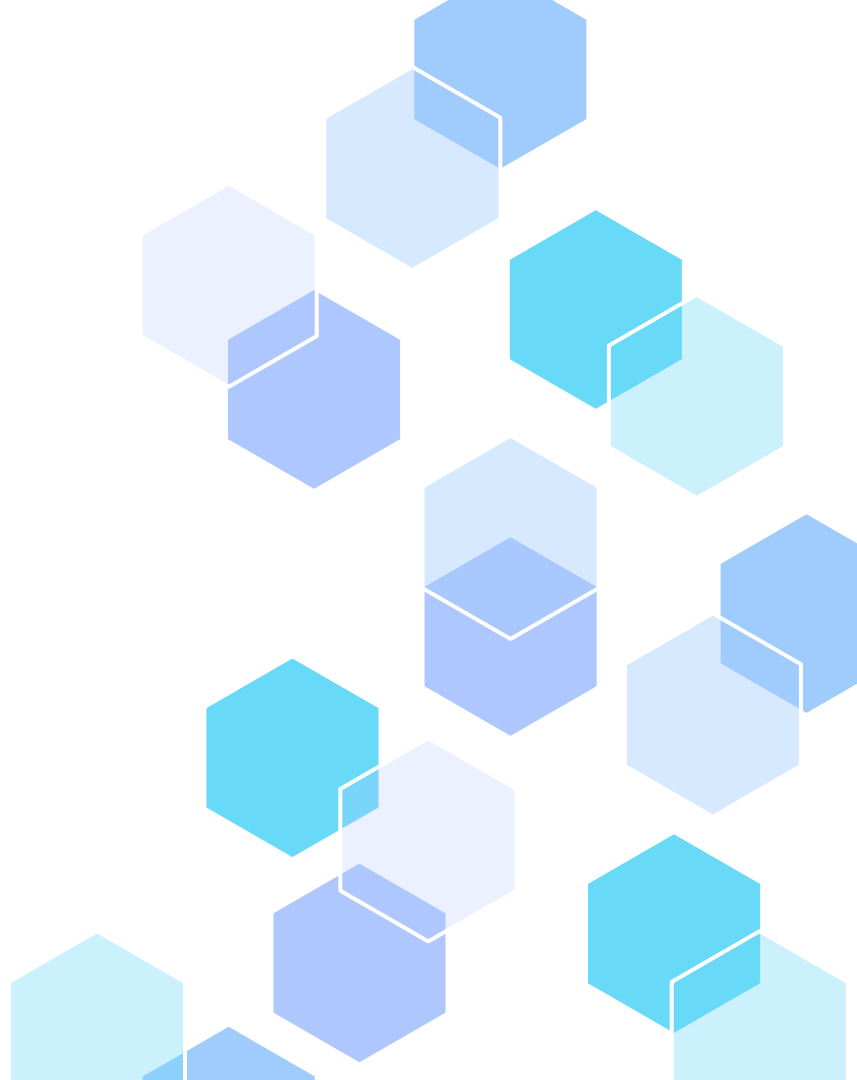


Native Scaling in PyTorch

PyTorch-native



Why Is This Difficult?



Scaling Up

Training a 1B parameter model (FP32 + Adam) — 21 GB total



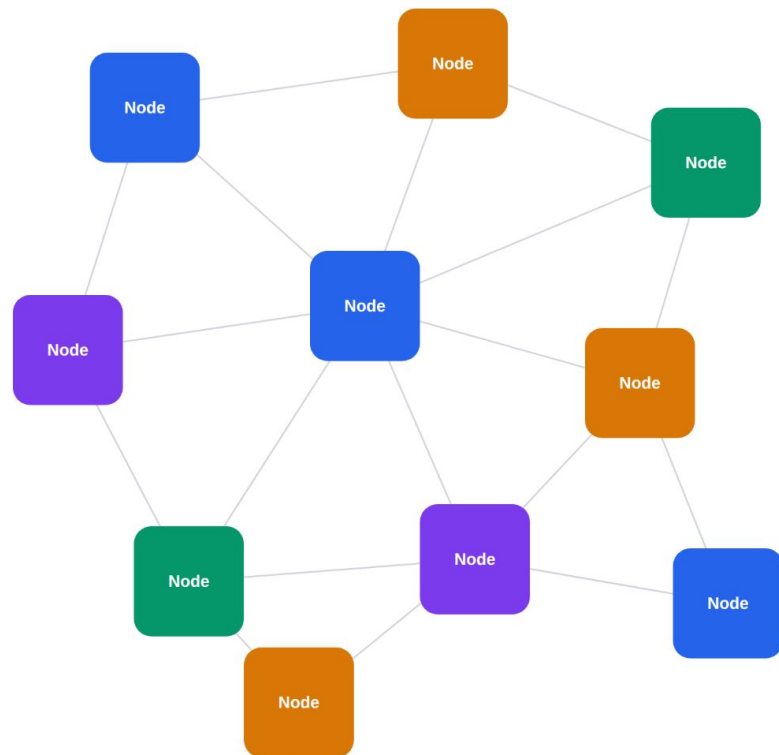
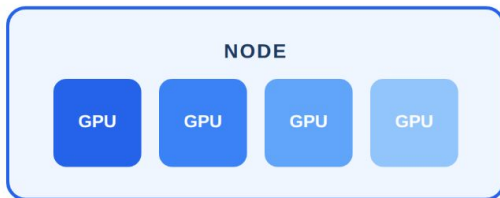
Training a 10B parameter model (FP32 + Adam) — ~190 GB total



80 GB A100 — only fits this much

Some modern frontier LLMs have over **1 trillion** parameters

Scaling Up



The slide features a white background with decorative hexagonal shapes in the corners. The top-left corner has a cyan hexagon overlapping a light blue one. The top-right corner has a light blue hexagon overlapping a medium blue one. The bottom-left corner has a medium blue hexagon overlapping a cyan one. The bottom-right corner has a cyan hexagon overlapping a light blue one.

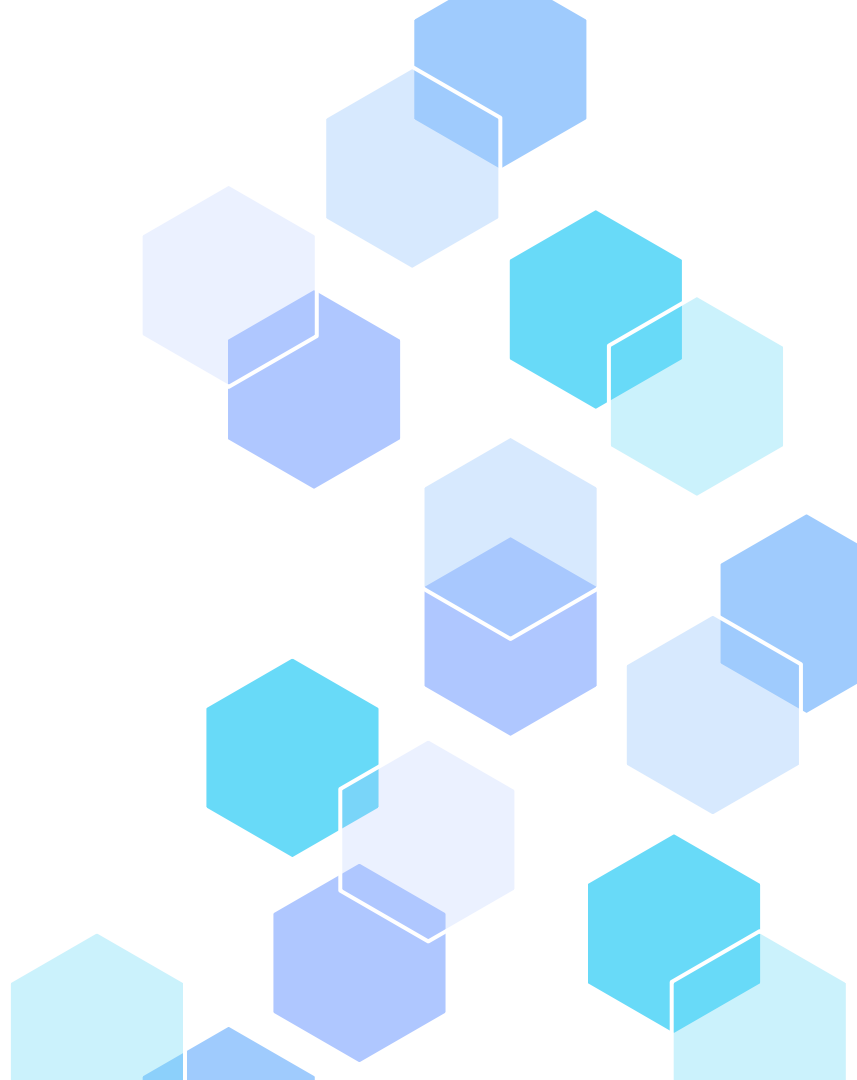
Discussion

What kind of problems might we face when scaling up the number of GPUs?

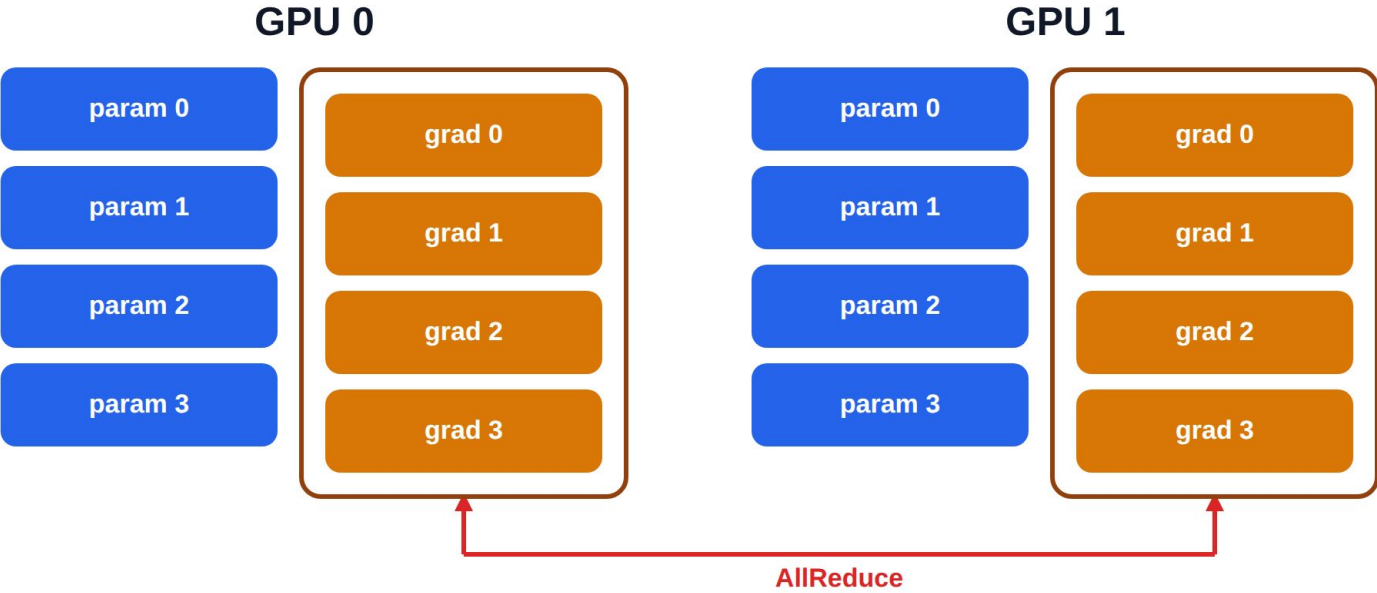
Usability



Existing Solutions



Data Parallelism

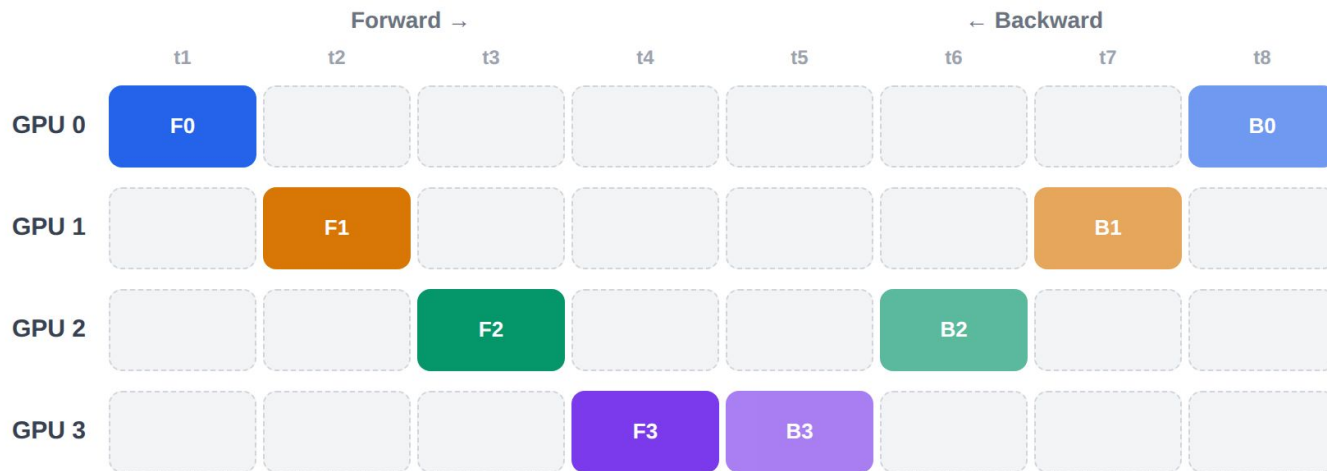


DDP – Distributed Data Parallel

Pipeline Parallelism

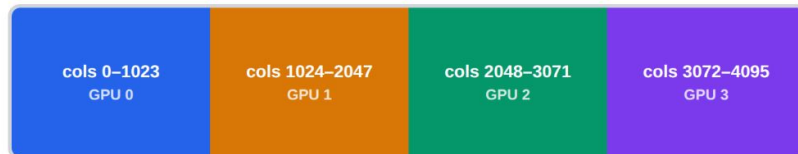


Pipeline Parallelism

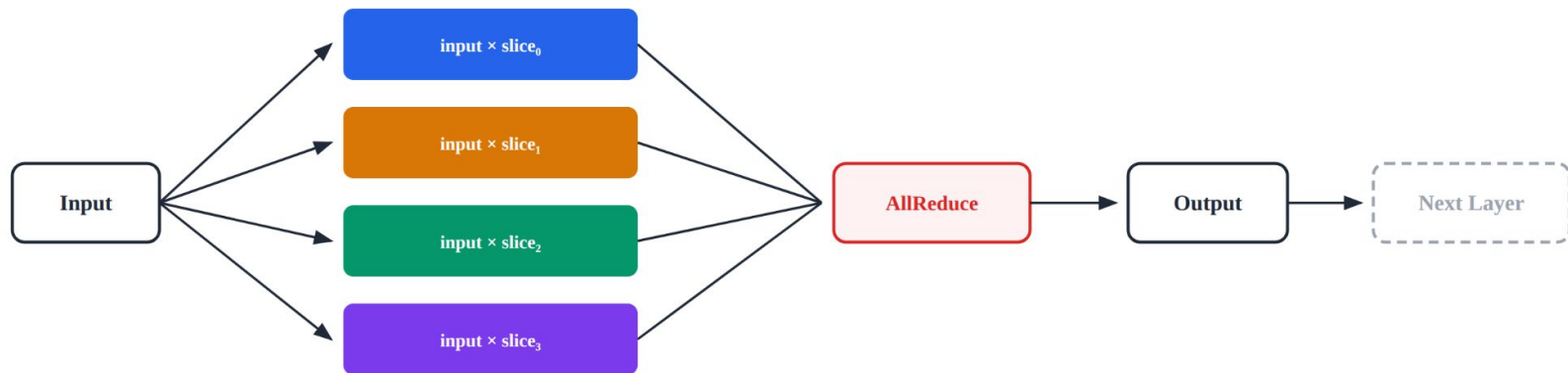
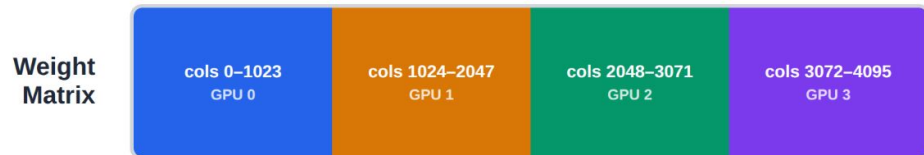


Tensor Parallelism

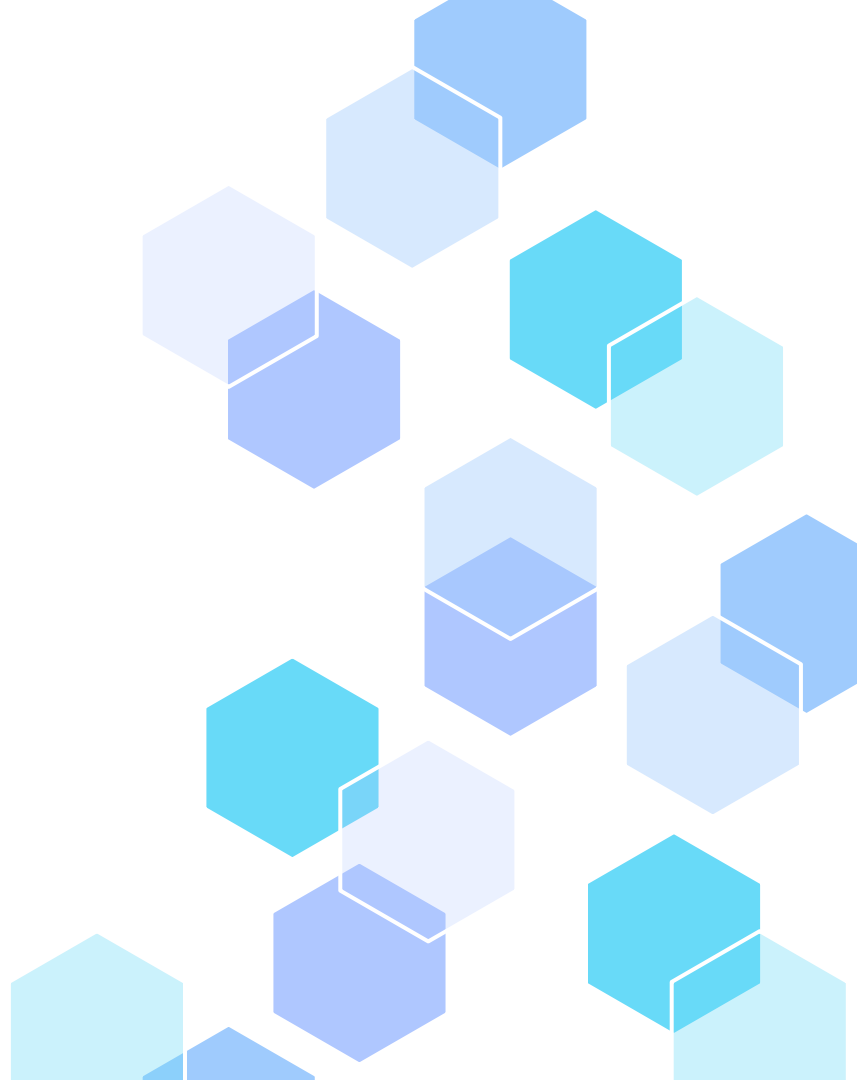
**Weight
Matrix**



Tensor Parallelism



Core Intuition



Overview

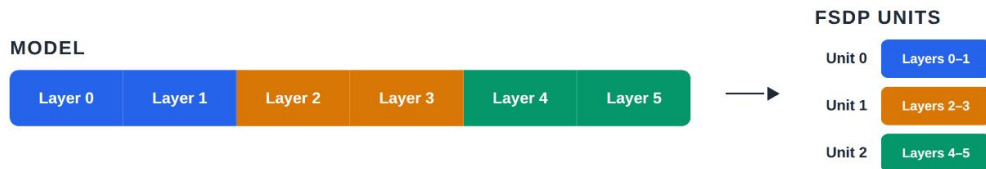
MODEL



FSDP UNITS



Overview



1 At Rest



2 Compute



NCCL Operations



The slide features decorative hexagonal shapes in the corners. The top-left corner has a cyan hexagon and a light blue hexagon. The top-right corner has a light blue hexagon and a medium blue hexagon. The bottom-left corner has a medium blue hexagon and a cyan hexagon. The bottom-right corner has a light blue hexagon and a cyan hexagon.

Comprehension Check

Why does FSDP rebuild parameters (AllGather) instead of keeping full copies like DDP?

Trading Latency for Memory

Ways to decrease memory pressure

- ↓ unit size
- ↑ Sharding Factor (full vs hybrid)
- Reduce After Forward
- Gradient Accumulation with communication
- No prefetching

Trading Memory for Latency

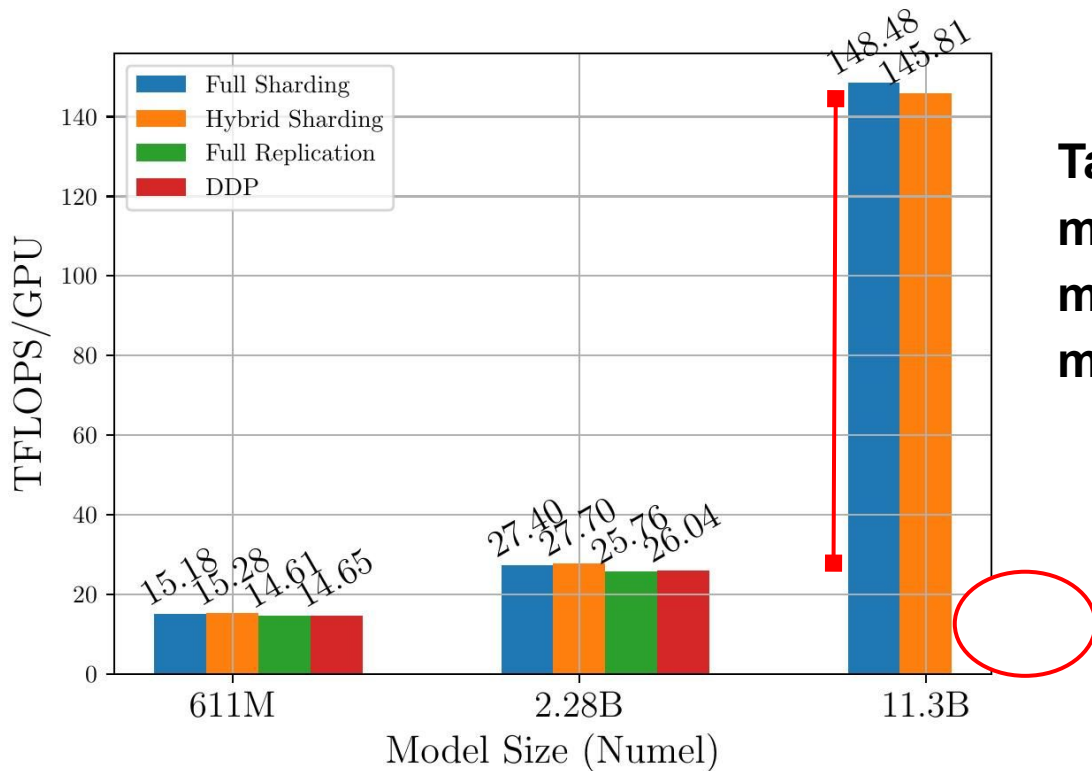
Ways to decrease latency

- ↑ unit size
- ↓ Sharding Factor (full vs hybrid)
- No Reduce After Forward
- Gradient Accumulation with no communication
- Prefetching

Experimental Results & Analysis

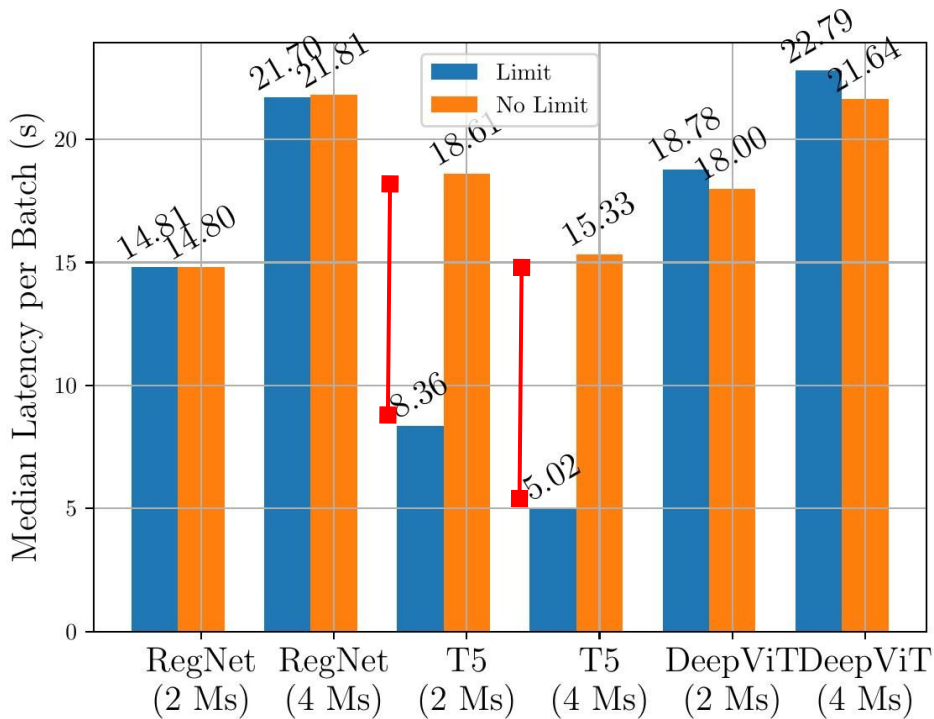


Training Efficiency



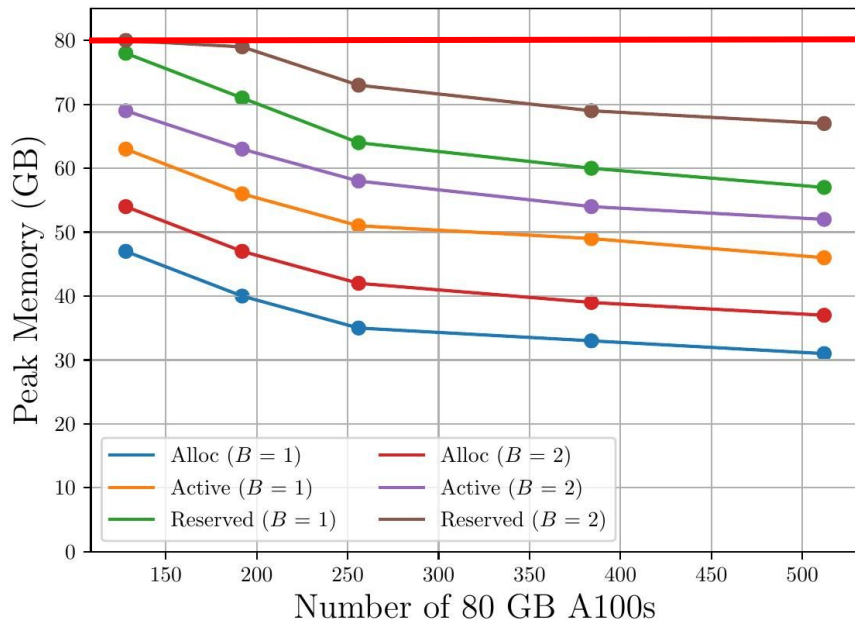
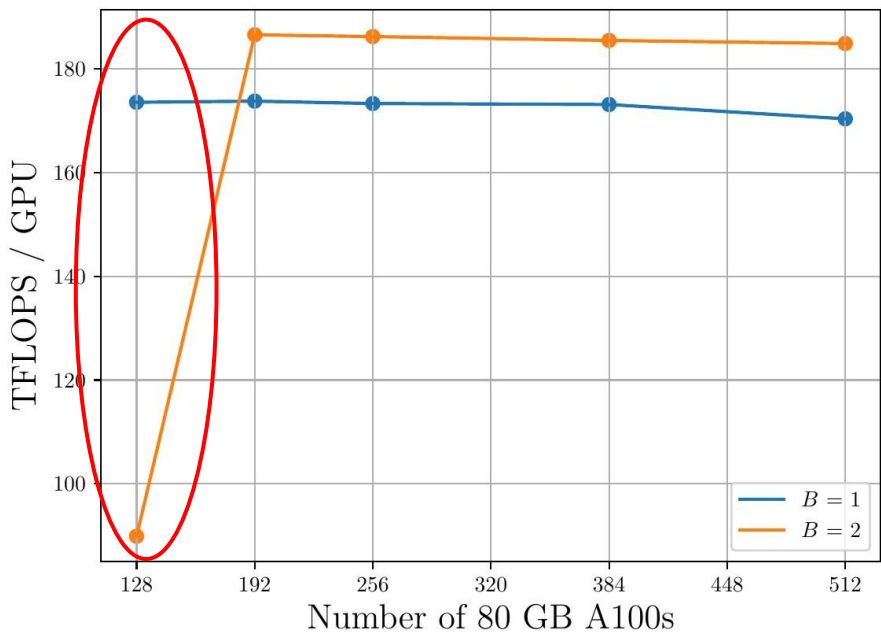
Takeaway: FSDP really matters for multi-billion parameter models

Rate Limiting



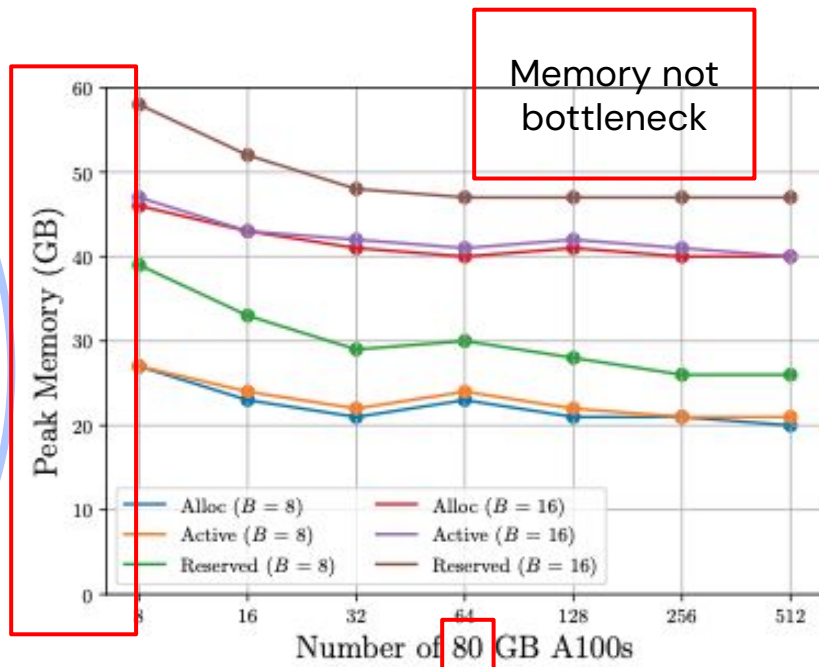
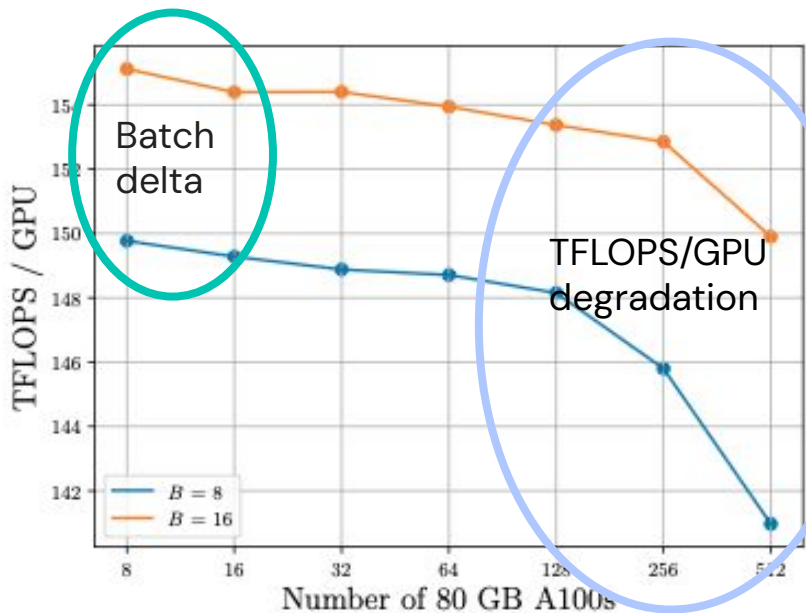
Takeaway: Verify if memory defragmentation occurs before rate limiting

GPT 175B Performance



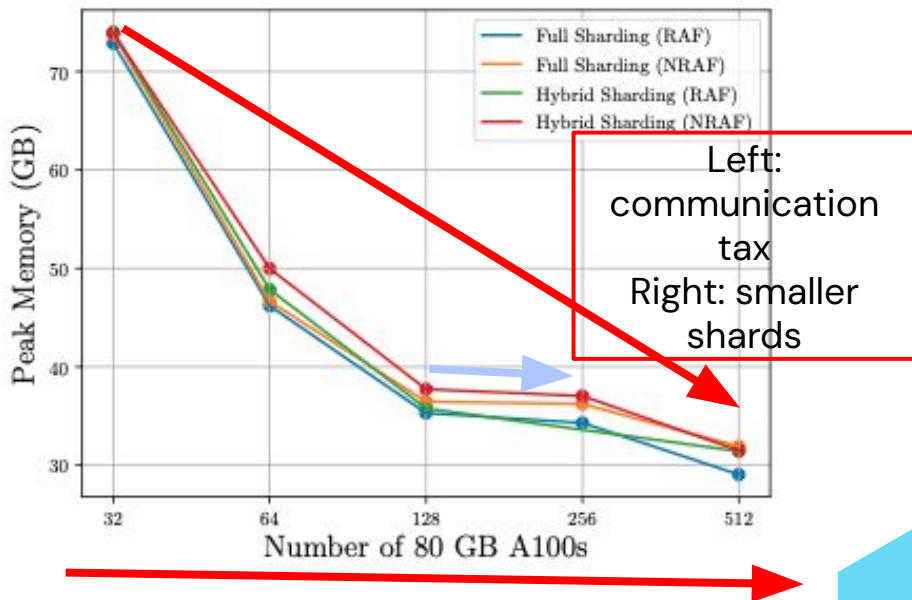
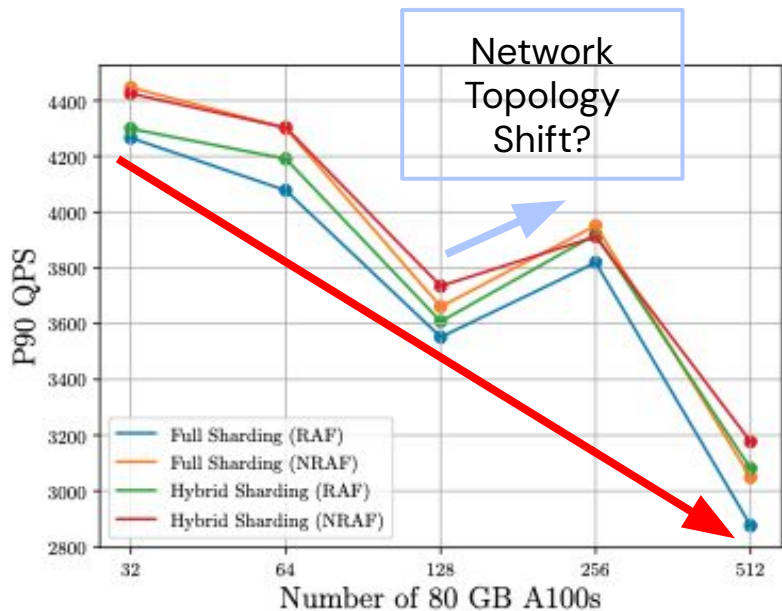
Takeaway: As peak memory approaches 80GB, defragmentation occurs

T5-11B Performance



- Despite ample memory headroom, per-GPU compute efficiency drops ~7% at 512 GPUs because communication overhead outpaces communication overlap at scale

DHEN (768B Sparse, 550M Dense) Performance



- FSDP offers a tunable memory-throughput trade-off: RAF minimizes memory, while Hybrid NRAF maximizes throughput

Are Experiments Sufficient?

Takeaway

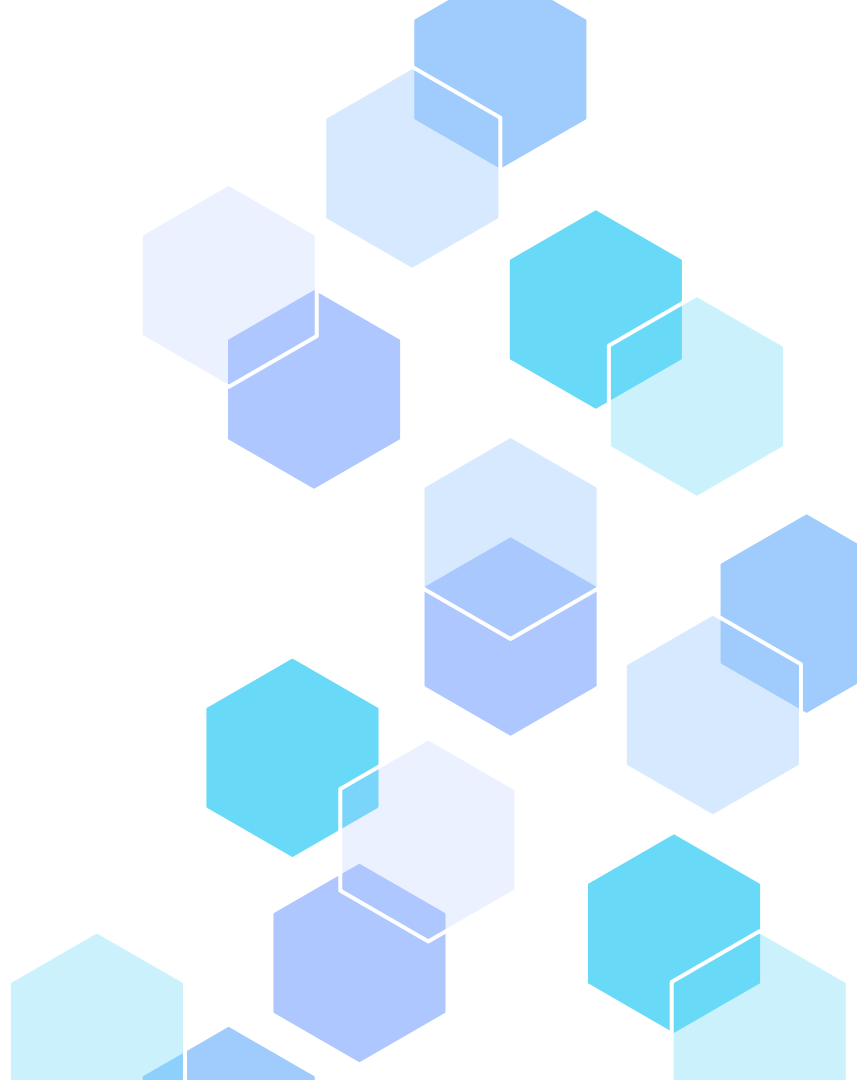
- **Convincing for:**
 - **Scaling**
 - **Memory efficiency**
- **Not sufficient for:**
 - **Generality**
 - **System comparisons**

The slide features a white background with decorative hexagonal shapes in the corners. The top-left corner has a cyan hexagon overlapping a light blue one. The top-right corner has a light blue hexagon overlapping a medium blue one. The bottom-left corner has a medium blue hexagon overlapping a cyan one. The bottom-right corner has a cyan hexagon overlapping a light blue one.

Discussion

Is FSDP a net win for scaling when considering memory vs. communication tradeoffs?

Gaps in the Paper



Potential Limitations

Narrow Evaluation Scope



Transformer

Text



CNN

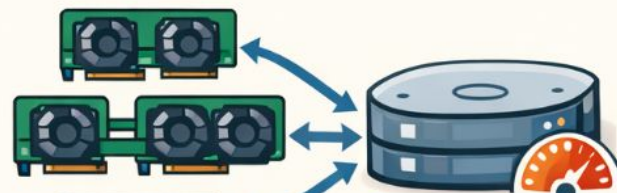


Multi-Task



Deep & Wide Models

Network Communication Bottleneck

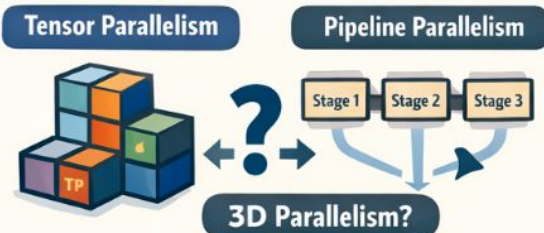


Many GPUs

Slow Network

Bandwidth Limits

Hybrid Parallelism Gaps



Tensor Parallelism

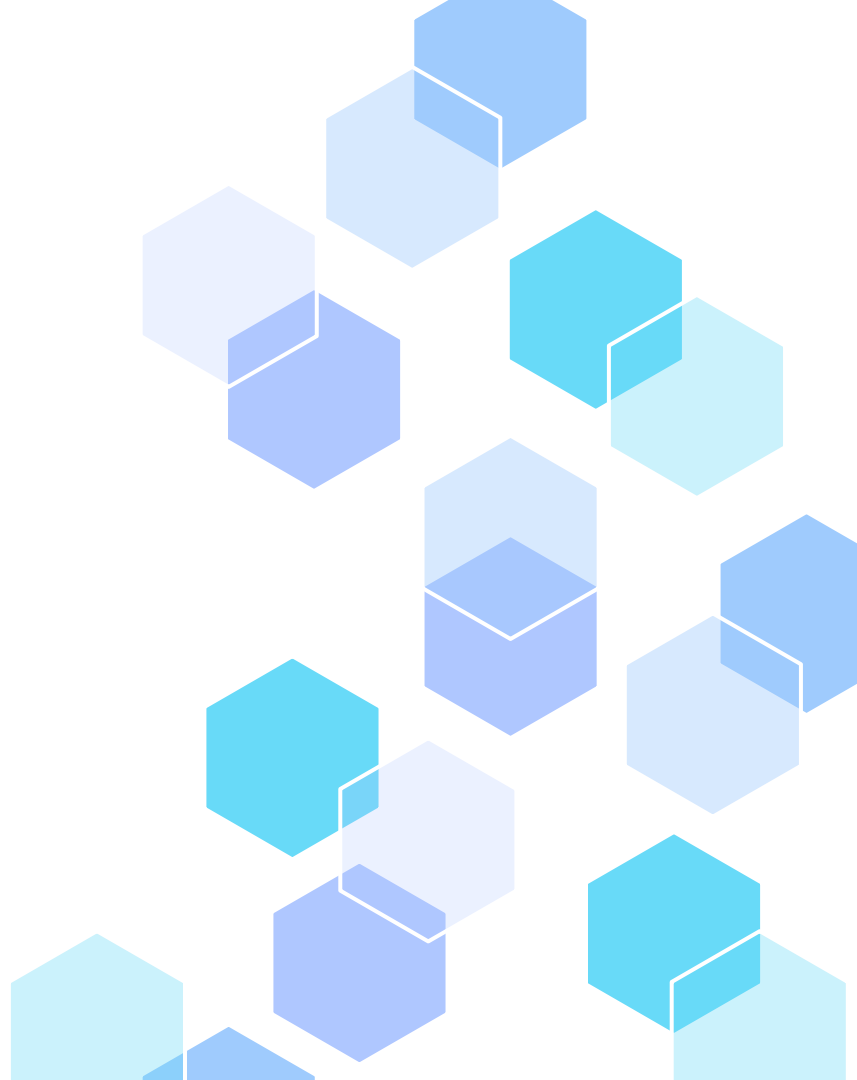
Pipeline Parallelism

3D Parallelism?

Checkpointing & Fault Tolerance

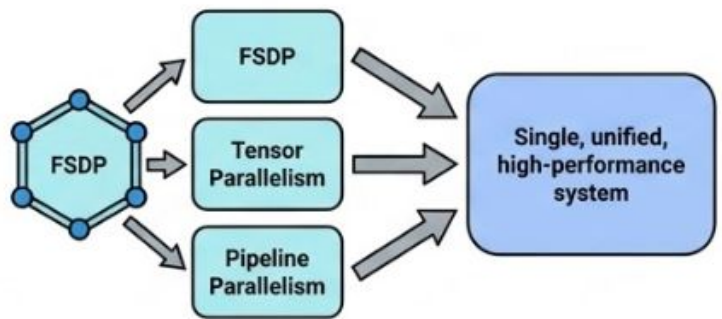


Next Steps

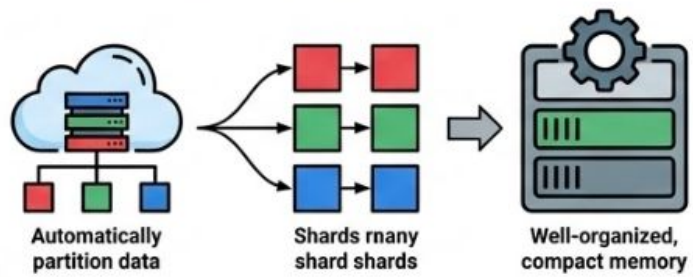


Future Systems

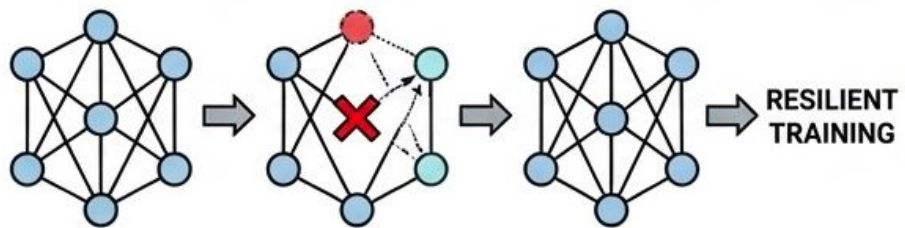
Hybrid Parallelism Systems



Automatic Sharding & Memory Optimization

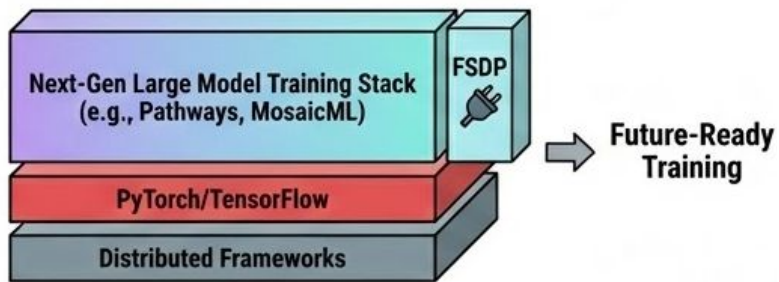


Elastic & Fault-Tolerant Training

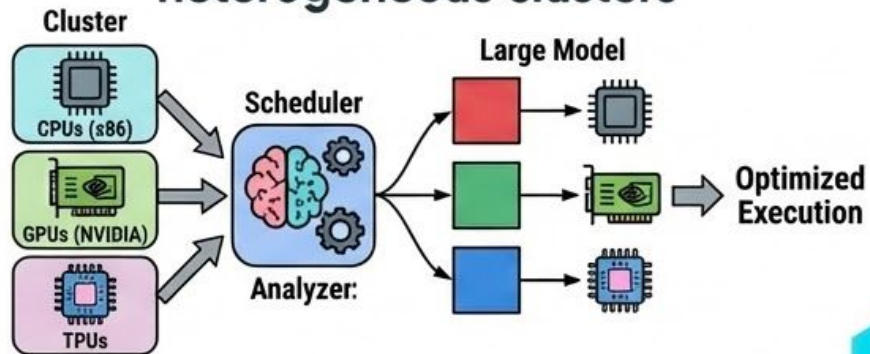


Future Systems

Integration with next-generation large model training stacks



Hardware-aware scheduling for heterogeneous clusters



Thanks!

CREDITS: This presentation template was created by [Slidesgo](#), and includes icons by [Flaticon](#), and infographics & images by [Freepik](#)

