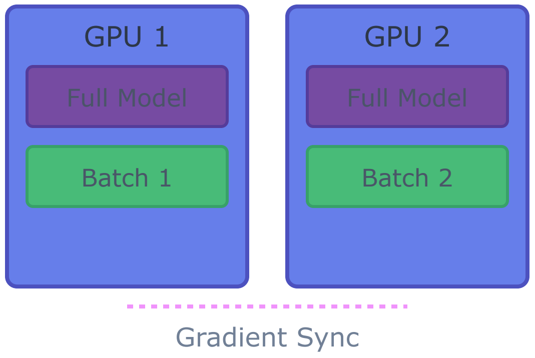
Distributed Training Paradigms

Data Parallelism



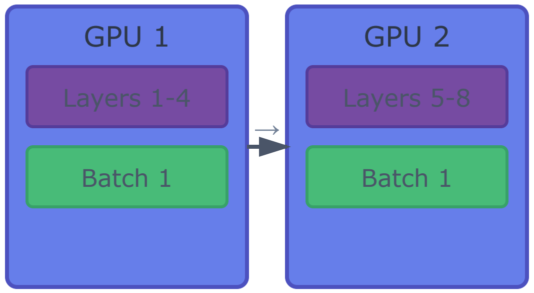
Characteristics

- Model replicated
- Different data per GPU
- Gradients synchronized
- Scales to 64-128 GPUs

Efficiency: 75-90%

Best for: Models fitting in GPU

Pipeline Parallelism



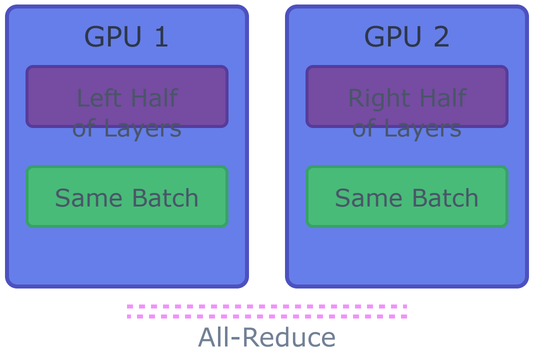
Characteristics

- Model partitioned by layers
- Sequential data flow
- Pipeline bubbles reduce efficiency
- Scales to 16-32 stages

Efficiency: 60-80%

Best for: Very large models

Tensor Parallelism



Characteristics

- Operations partitioned
- Same data per GPU
- Frequent communication
- Requires fast interconnect

Efficiency: 80-90%

Best for: Huge single layers

Hybrid Approach: GPT-3 Training (1,024 GPUs)

8-way Tensor Parallelism

... (8 GPUs per layer)

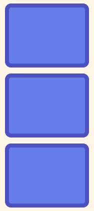
↓ Each layer split across 8 GPUs

16-way Pipeline Parallelism

... (16 stages)

↓ 16 layer groups in pipeline

8-way Data Parallelism



8 replicas
of entire
pipeline

Total: 8 × 16 × 8 = 1,024 GPUs

Combines strengths of all three approaches for maximum scale

Overall efficiency: ~50-60% (communication overhead at this scale)