

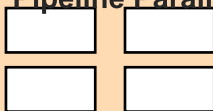
3D Parallelism for Large Model Training

Data Parallel



Split batch
across GPUs

Pipeline Parallel



Split layers
across GPUs

Tensor Parallel



Split matrices
across GPUs

3D Combined
Optimal Training