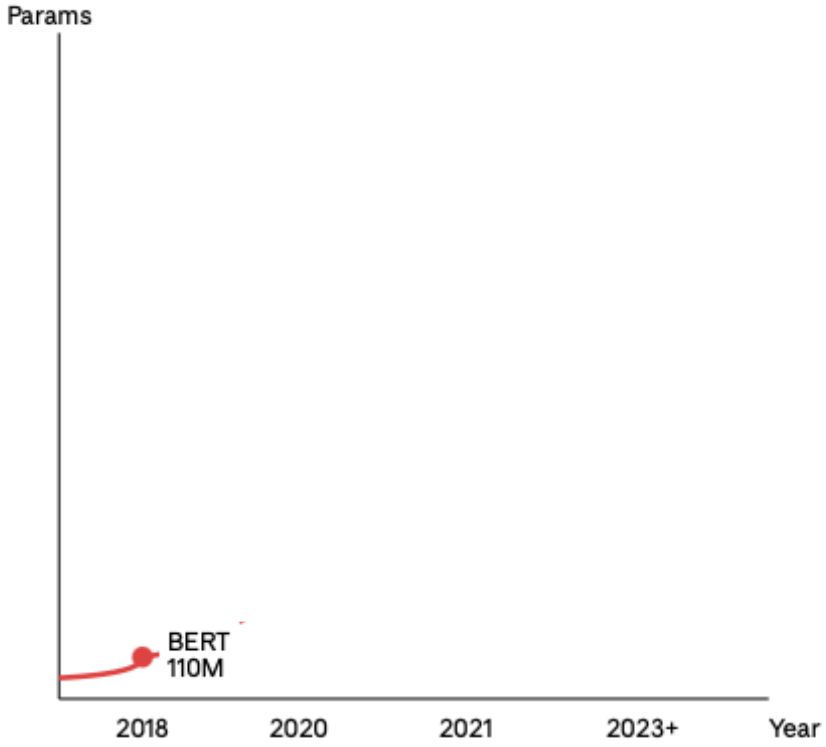
ZeRO: Memory Optimizations Toward Training Trillion Parameter Models

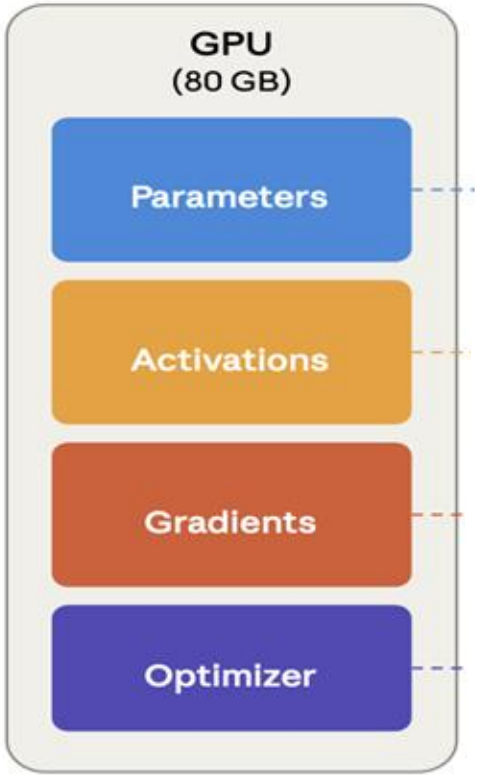
Samyam Rajbhandari∗, Jeff Rasley∗ , Olatunji Ruwase, Yuxiong He

Model size is growing exponentially, but GPU memory is not

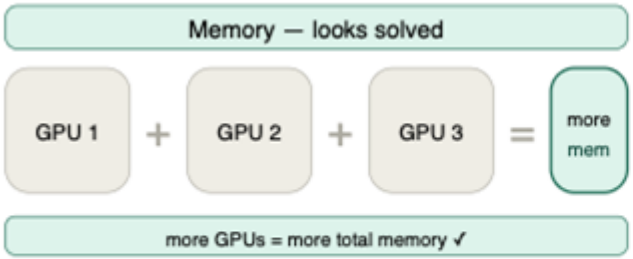


How is a model trained?

What is stored in a single GPU?

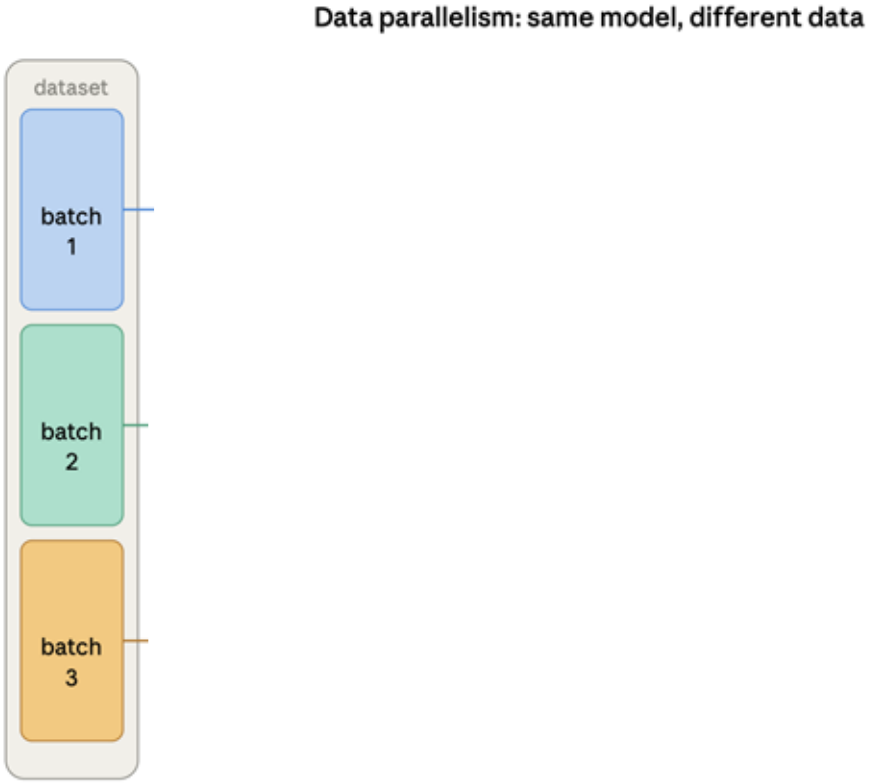


Naive Solution

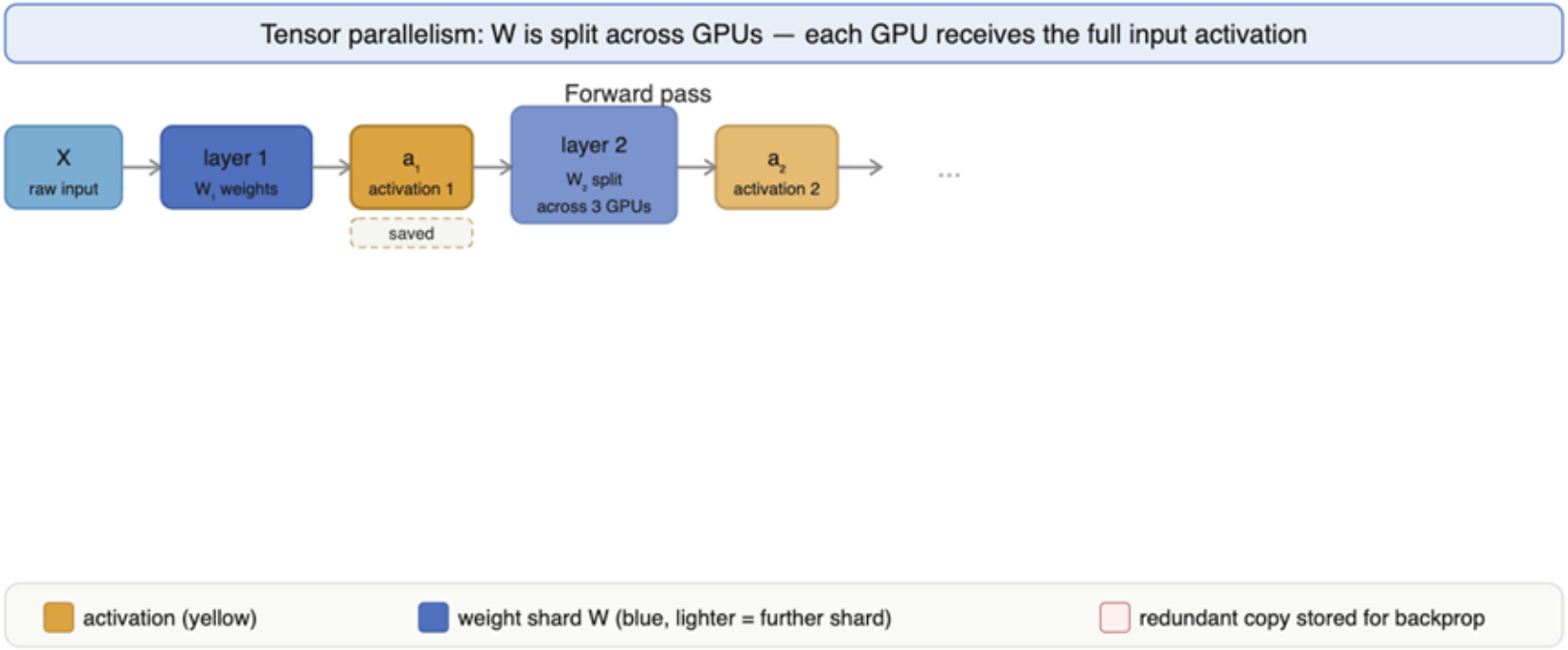


How do existing methods attempted to solve this problem?

1. Data Parallelism

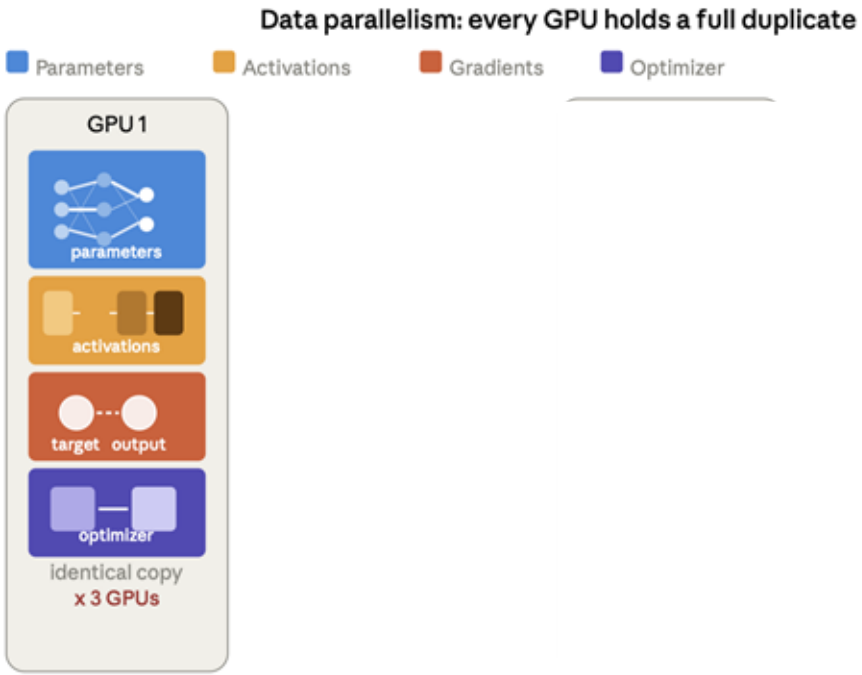


2. Model Parallelism

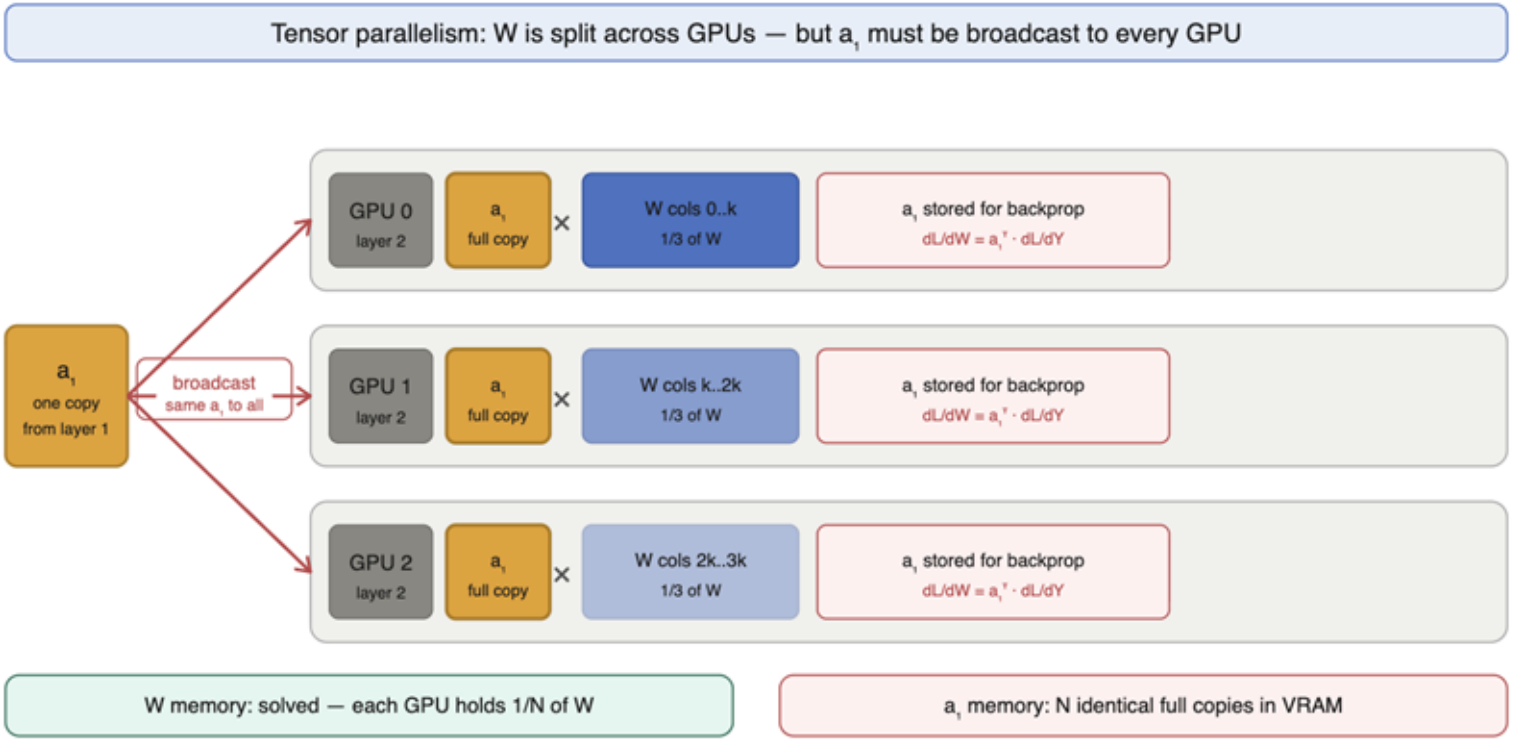


Where do Existing methods Fail?

1. Problem with Data Parallelism:

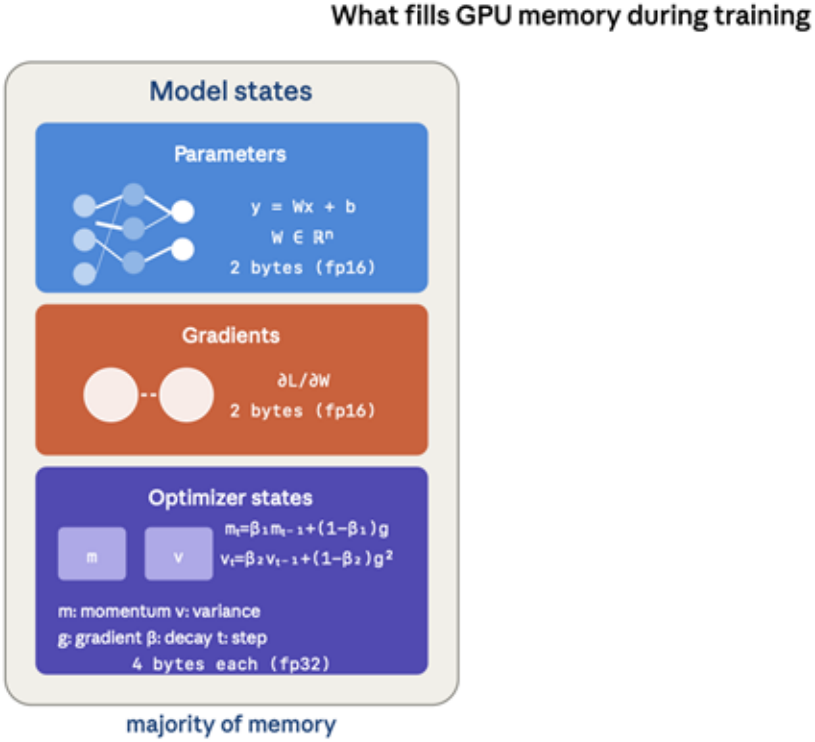


2. Problem with Model Parallelism



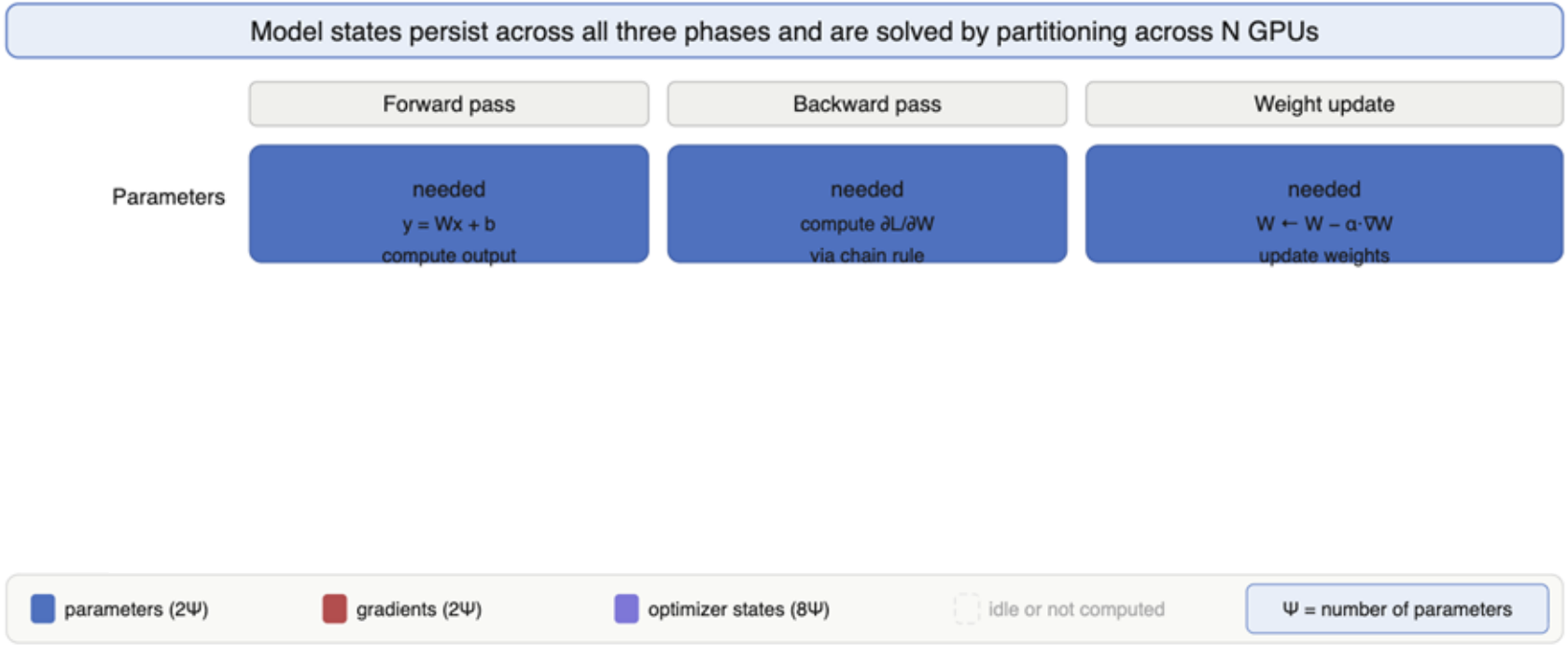
Redundancy is the key Issue

ZeRO Full Spectrum Analysis



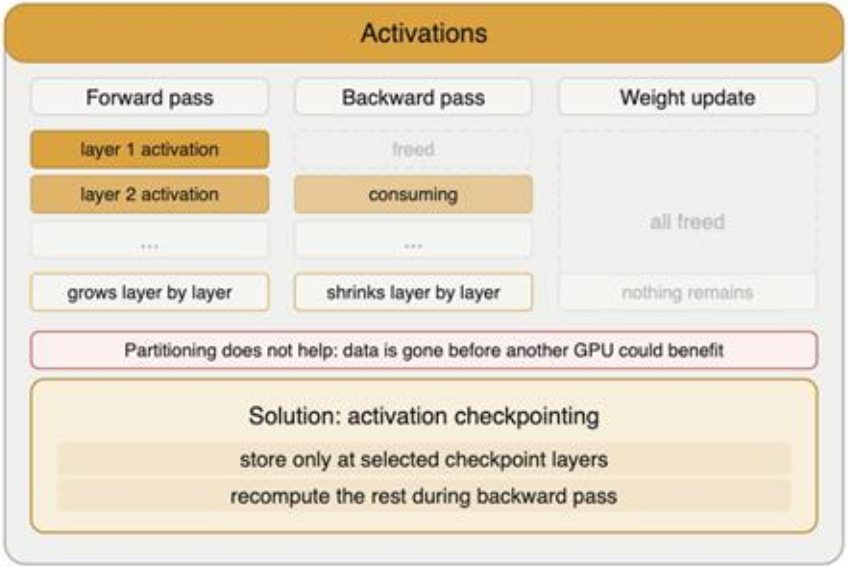
The Main Intuition is: Partition What persists, Checkpoint What Disappears

Model States Persist



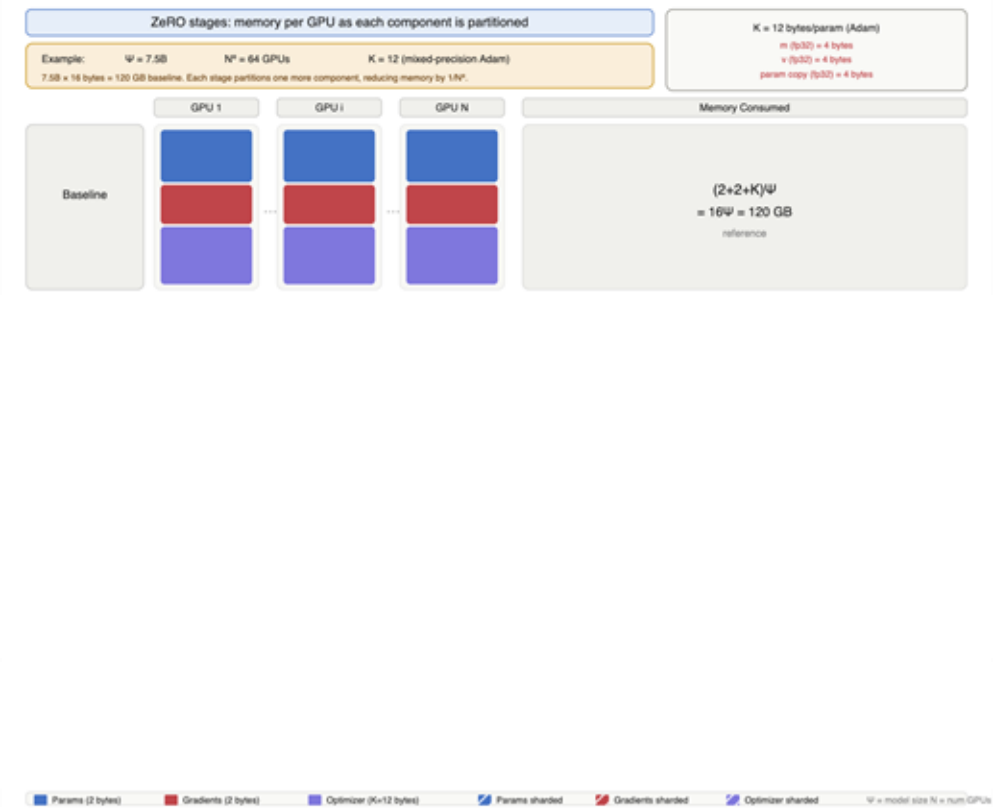
Residual States Disappears

Residual states are transient within a phase and cannot be solved by partitioning

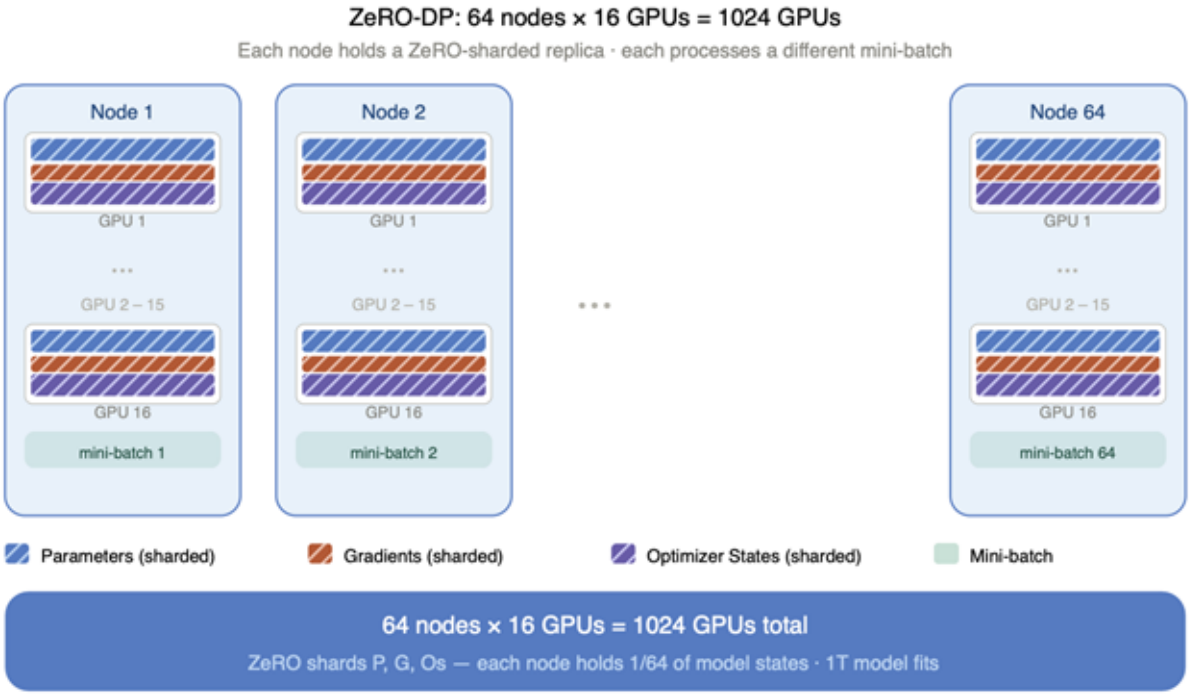


Both residual state problems are within a single GPU and within a single phase, partitioning across GPUs cannot help

ZeRO-DP: Optimizing Model State Memory



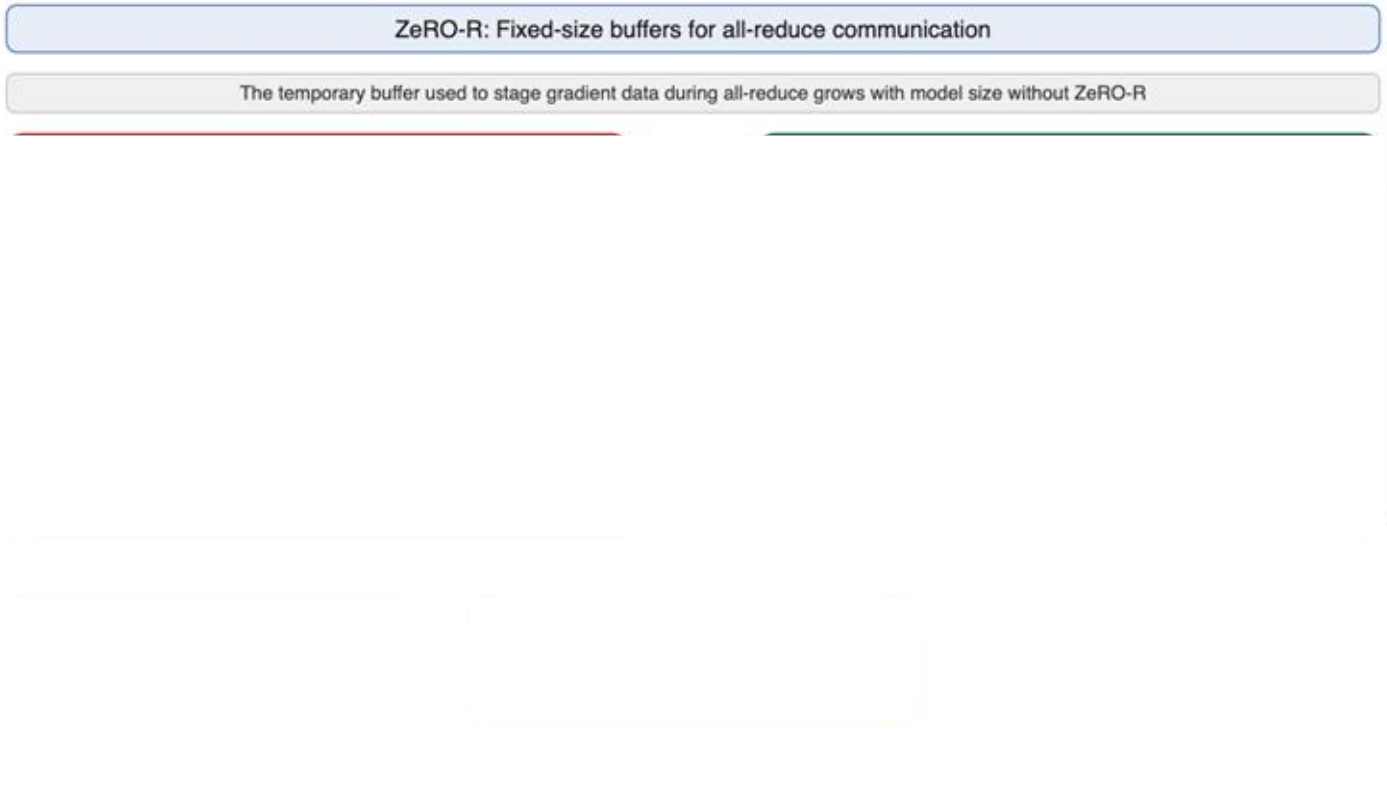
Result: Enables 1T-parameter Model Training



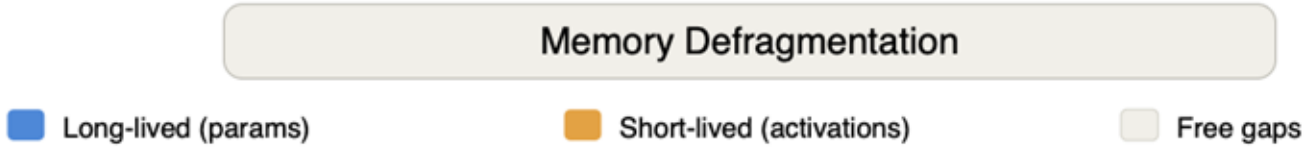
ZeRO-R: Optimizing Activations

Activation Partitioning

ZeRO-R: Optimizing Residual State Memory



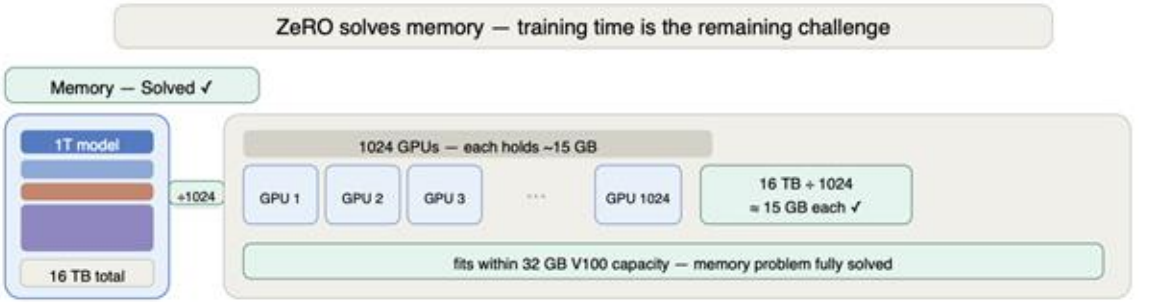
ZeRO-R: Optimizing Residual State Memory



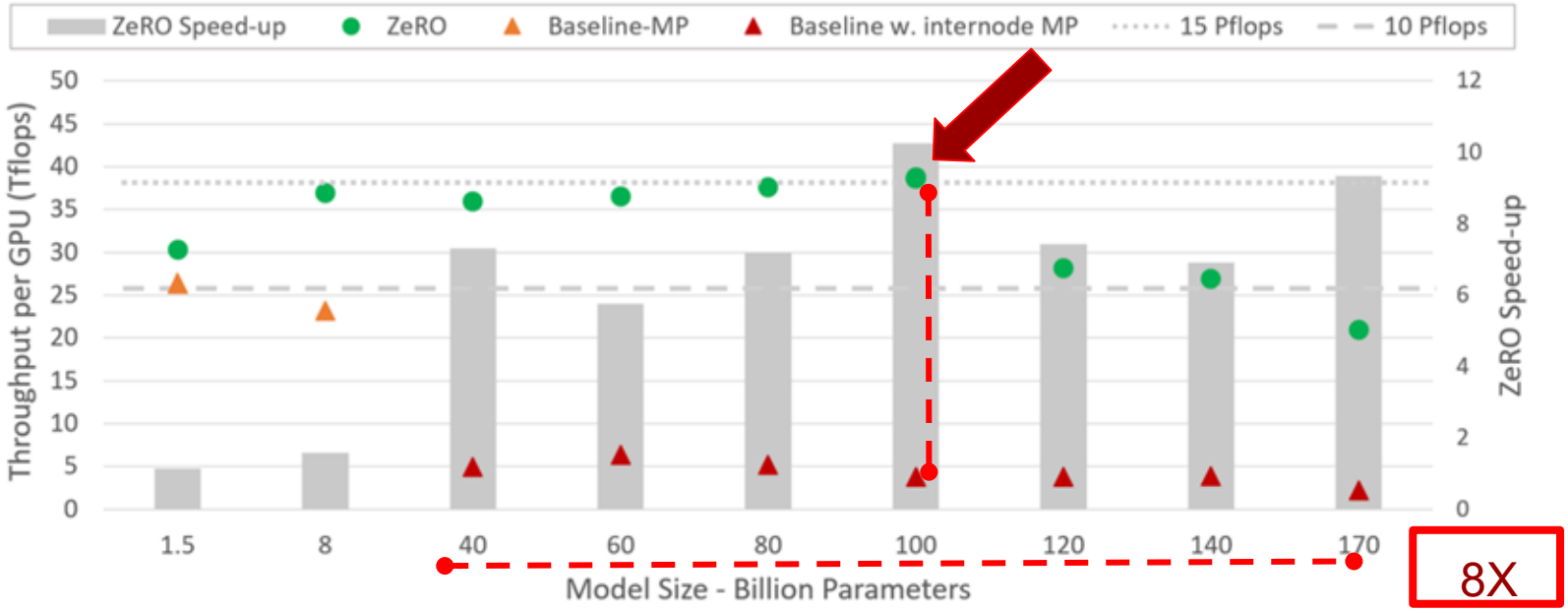
Zero-DP + Zero-R = ZeRo

Evaluation

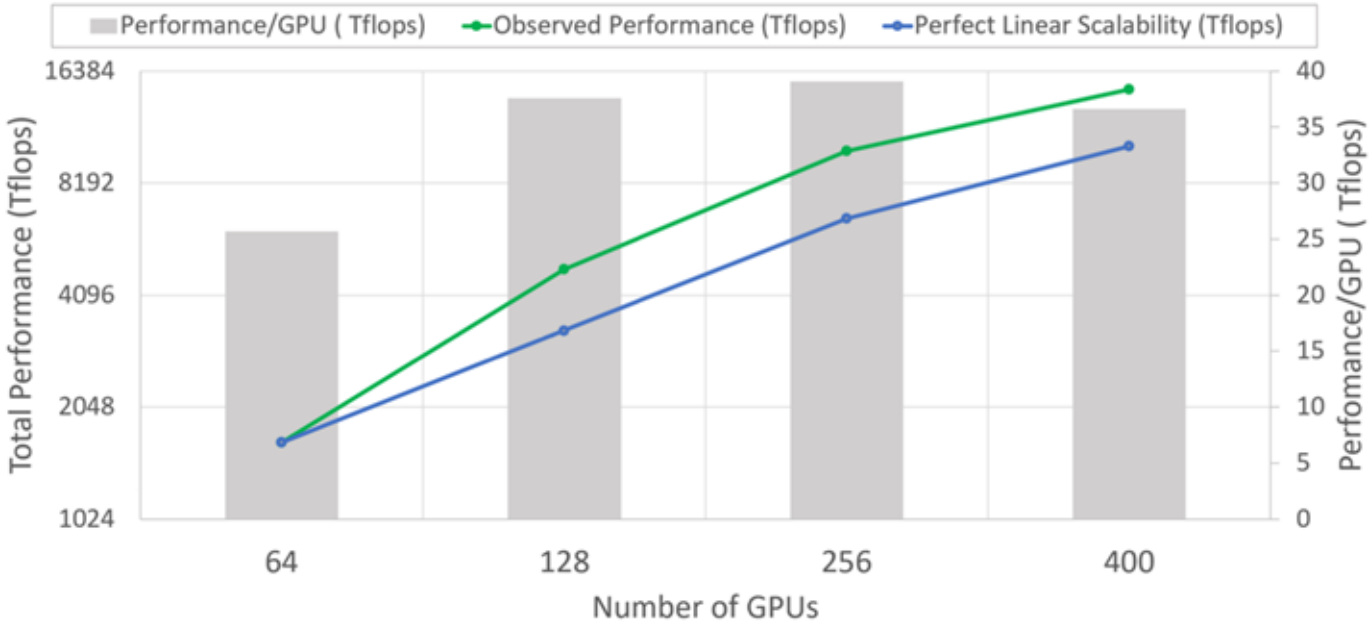
Reshift Target: 1 Trillion Parameter Model



Memory Efficiency and Speed with model size



Scalability



Drawbacks

- **Higher communication overhead:** Bottleneck on slow interconnects (e.g., Ethernet vs NVLink)

states; harder to maintain

Thank you!