1. Model scaling hits GPU memory ceilings fast—even 1.5B params can overload a 32 GB GPU.
2. Data Parallelism (DP) duplicates all model states—wastes memory on every GPU.
3. Optimizers like Adam multiply memory with extra states—memory bloat grows rapidly.
4. Residuals like activations and buffers dominate memory—especially for long sequences.
5. Model Parallelism (MP) splits models across GPUs but adds costly inter-GPU communication.
6. Pipeline Parallelism (PP) staggers micro-batches but needs large batches and adds complexity.
7. CPU offloading saves GPU memory but drags down performance due to slow transfers.
8. Activation checkpointing trades compute for memory—helps only to a point.
9. ZeRO partitions optimizer states, gradients, and weights across GPUs to eliminate redundancy.
10. ZeRO-DP gives 4× to N× memory savings—no model rewrite needed.
11. ZeRO-R cuts memory by optimizing residuals—activations, buffers, fragmentation.
12. ZeRO alone trains 100B+ models fast—no need for MP unless absolutely necessary.
13. Combine ZeRO + MP when activations bottleneck or batch size must shrink.
14. ZeRO enables trillion-parameter scale—by combining DP, MP, and activation partitioning.
15. In tests, ZeRO showed 10× speedup and super-linear GPU scaling up to 400 GPUs.
16. ZeRO maintains 30–40% GPU peak FLOPS—even on 100B+ parameter models.