LLM training in practice: insights from 85,000 checkpoints

Glenn K. Lockwood
VAST Data, Inc.
glenn.lockwood@vastdata.com

1 Abstract

Training large language models (LLMs) at scale places enormous demands on I/O infrastructure, and vendor guidance often emphasizes the *supply side* of this challenge: the peak I/O bandwidth required to prevent workloads from being unduly limited by storage. Such specifications tend to overstate I/O performance requirements, because they reflect scenarios where GPUs are all running at perfect, maximum utilization. Real-world workloads are not as efficient, but little has been said about the I/O requirements of LLM training from the demand side.

VAST operates the data platform for many of the world's largest AI training clusters and has visibility into this *demand side* of I/O performance. To understand how different supply-side and demand-side I/O requirements are, we analyzed more than 85,000 checkpoints written by 40 different production LLM training jobs across 18 AI training clusters. We found that the bandwidth required to checkpoint efficiently at even trillion-parameter scale is very modest, with the largest models never exceeding several hundred GB/s. We generalize these findings into a simple performance model that, given a model size and desired checkpoint frequency, provides a demand-size perspective on the global bandwidth required to support model training.

2 Methods

We analyzed "phone-home" telemetry that samples a wide array of I/O metrics every ten seconds across every deployed VAST cluster. We specifically examined data from clusters that (1) opted in to sharing cluster telemetry with VAST, (2) were deployed at AI cloud providers whose customers train AI models, and (3) were mounted by at least 128 GPU nodes (1,024 GPUs).

From these data streams, we developed a process of identifying checkpoints that first identifies all peaks in the write bandwidth timeseries data, then correlate those intense write periods with the used capacity of the cluster to confirm that the peaks correspond to large volumes of data being written and retained. For each identified checkpoint, we then estimated several workload metrics, including total model parameter count (based on the assumption that checkpoints consume 14 bytes per parameter[1, 2]), checkpoint duration, and effective and peak write bandwidth. We then apply kernel density estimation to cluster checkpoints based on model parameter count to characterize how training jobs perform checkpointing over time.

These methods identified 40 large, production training jobs, covering more than 85,000 checkpoints across 18 VAST clusters. Model sizes ranged from 45 billion to over 1 trillion parameters, representing modest- to frontier-scale LLMs.

3 Preliminary findings

Perhaps counterintuitively, we observe that training larger models across more GPUs does not require proportionally larger I/O performance for checkpointing. Instead, write bandwidth has no meaningful correlation with model size at scale, because state of the art models rely on asynchronous, hierarchical checkpointing to node-local storage to achieve scalable checkpoint performance. As a result, the I/O workload experienced by globally shared storage does not represent the time that GPUs are idle, but rather, the time required to drain a checkpoint from node-local storage asynchronously after training has resumed.

From this observation, we identify *checkpoint overlap* as the most relevant metric for understanding the I/O performance of checkpointing during training. This metric is the fraction of time between successive checkpoints that overlaps with the checkpoint being drained to shared storage. For example, if a training job drains checkpoints every 30 minutes, and each drain takes 3 minutes, the checkpoint overlap is 10%. Our analysis shows that virtually every training job has less than 10% checkpoint overlap regardless of model size, as shown in Figure 1.

We then propose that the performance of a storage system for LLM training should be sized to ensure that checkpoint overlap is a reasonable value, not sized to match the number of GPUs being used. Quantitatively, this model can be represented as

$$\text{Checkpoint bandwidth (GB/s)} = \frac{N_{params} \times b_{pp}}{t_{interval} \times f_{overlap}} \times 10^{-9}$$

where N_{params} is the number of model parameters, b_{pp} is the bytes checkpointed per model parameter, $t_{interval}$ is the desired checkpoint frequency, and $f_{overlap}$ is the desired checkpoint overlap.

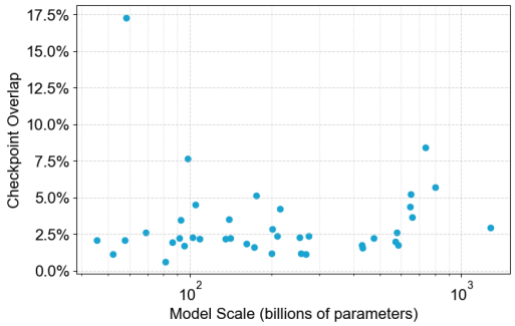

Fig. 1. Fraction of the checkpoint interval that overlaps with the checkpoint draining process as observed across 40 different model training jobs.

REFERENCES

[1] S. Rajbhandari, J. Rasley, O. Ruwase and Y. He, "ZeRO: Memory optimizations Toward Training Trillion Parameter Models," *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, Atlanta, GA, USA, 2020, pp. 1-16, doi: 10.1109/SC41405.2020.00024.

[2] S. Dash *et al*., "Optimizing Distributed Training on Frontier for Large Language Models," *ISC High Performance 2024 Research Paper Proceedings (39th International Conference)*, Hamburg, Germany, 2024, pp. 1-11, doi: 10.23919/ISC.2024.10528939.