LLM training in practice

Insights from 85,000 checkpoints

Glenn K. Lockwood, Ph.D.
Principal Technical Strategist
VAST Data

Nobody knows how much I/O performance is required by AI

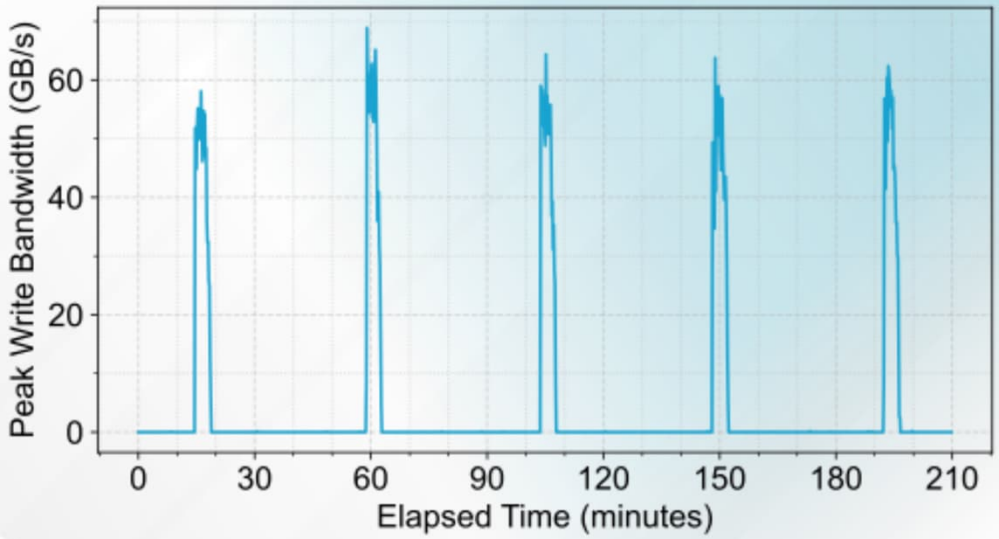
But plenty will tell you how much performance you should buy

The study: Figure out how much performance is actually used

Longitudinal study of 85K checkpoints, 40 models trained, 18 VAST clusters, 1 year, at 10-second intervals

1. Figure out how to pull checkpoints from thousands of GPU-years of VAST cluster telemetry
2. Characterize 85,619 individual checkpoints
3. Cluster checkpoints into training jobs
4. Characterize training jobs (model size, checkpoint interval, etc)

Example of one (very clean) trace on a VAST cluster during training



#1: Checkpoint bandwidth *decreases* when training larger models

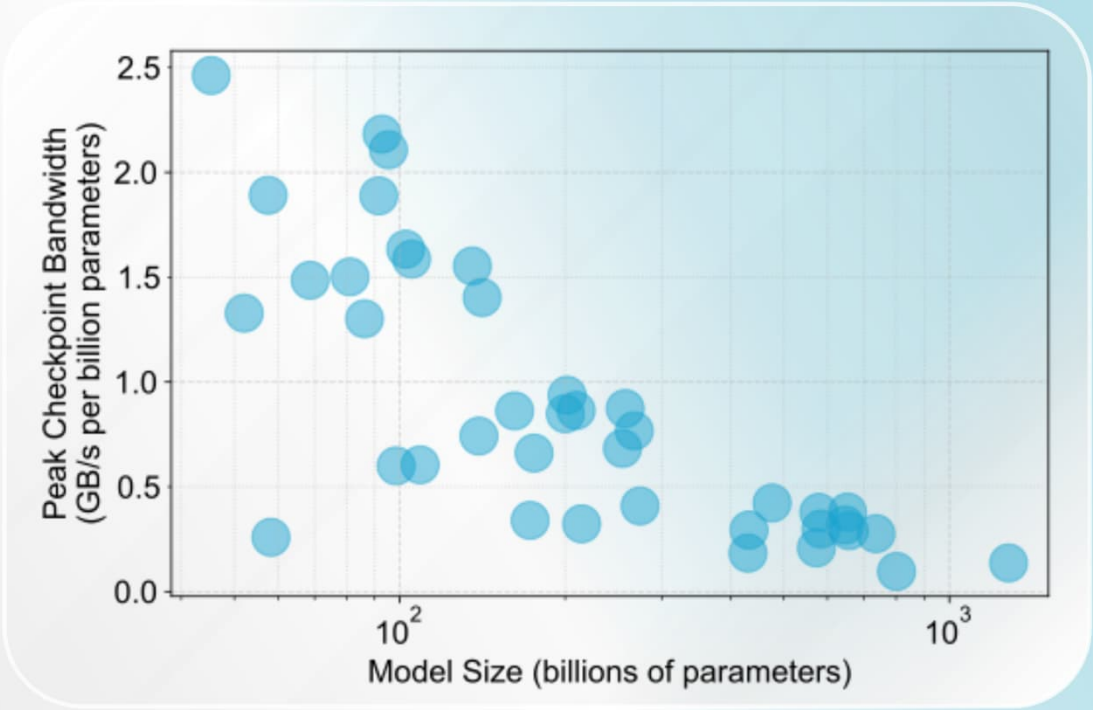
Checkpoint volume is constant for a model, even if job size is not

Large models use **data parallelism**

Data parallelism uses **more GPUs** to process more **input data**

Input data is independent of **checkpoint size**

∴ **Large models** need **more GPUs**, not more **checkpoint bandwidth**



#2: Bandwidth/GPU is independent of model size

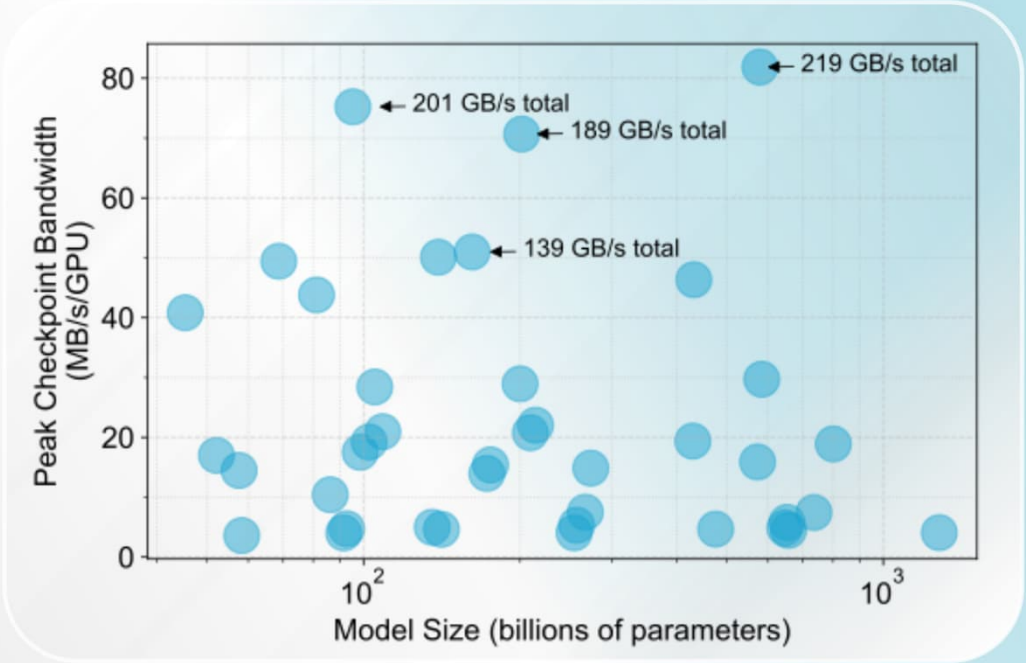
Checkpoints are buffered using local SSDs in production training

Reasonable assumption:

- Big models → Lots of GPUs
- Lots of GPUs → Lots of crashes
- Lots of crashes → Fast checkpointing
- Fast checkpointing → High bandwidth

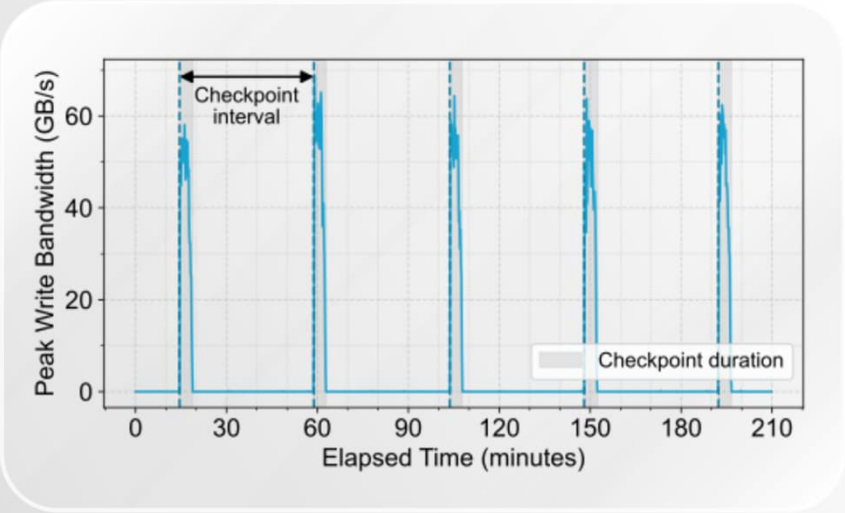
Not so!

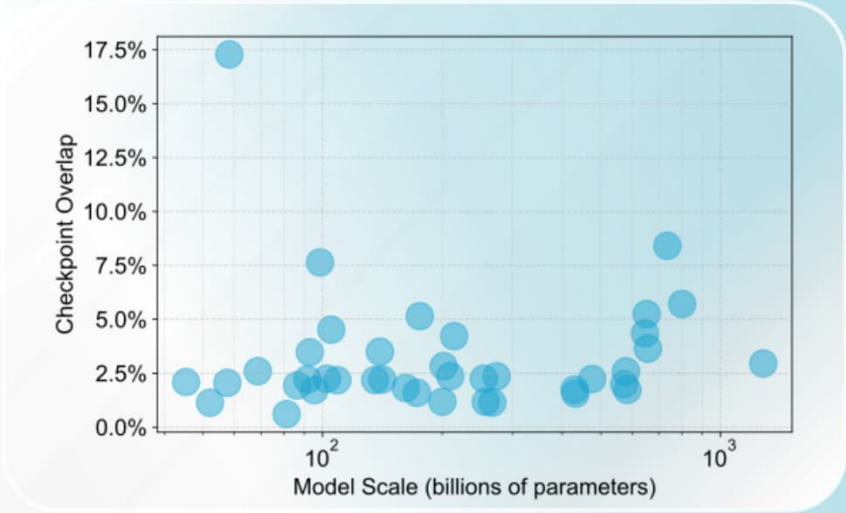
- Async checkpointing uses node-local SSD for high-bandwidth
- Global storage just drains last checkpoint before next one begins
- MB/s/GPU is fine for models ≥ 1T parameters in practice



#3: Performance is driven by checkpoint interval, not model size

Since checkpoint drains don't block GPUs, they can be "just fast enough" instead of "as fast as possible"





In practice, most model builders keep checkpoint overlap under 10%

Our goldilocks model for training performance

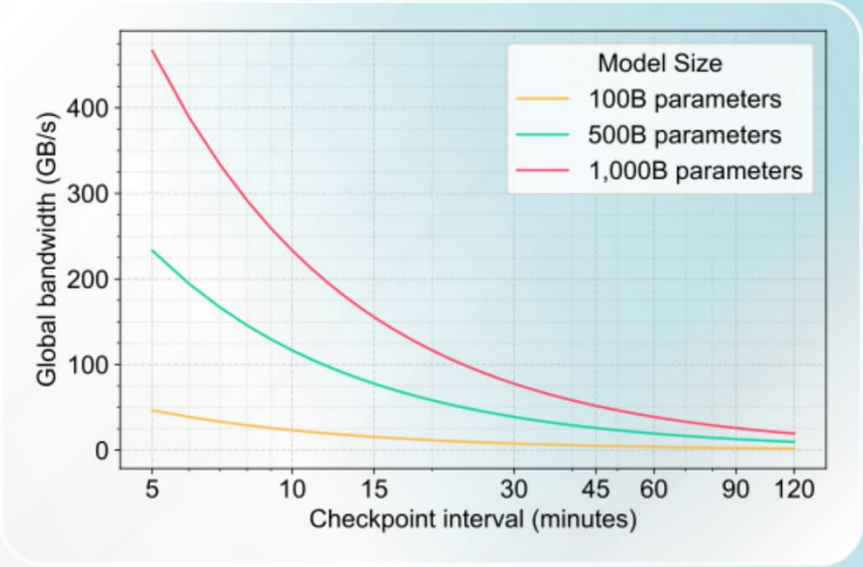
There is an optimal global bandwidth for checkpointing during model training

Faster is only better when you aren't limited by space, power, cooling, or cost.

- Less storage bandwidth → less storage power
- Less storage power → more GPU power
- More GPU power → more GPUs
- More GPUs → faster training

∴ Less storage bandwidth = faster training

(until checkpoint overlap gets too large)



Further reading:
If this project seems useful, read the write-up.

The idea

Analyze the data from the VAST fleet

- VAST powers AI infrastructure at every leading AI cloud provider
- Look at our **raw telemetry** (no customer logs, etc.) to see what's really happening in production
- Perform **longitudinal study** of training jobs' I/O patterns
  - How do they really checkpoint?
  - How much GB/s/GPU are they using?
  - Can we define how much bandwidth is "enough?"