THE GUTS OF LARGE LANGUAGE MODEL CHECKPOINTING

HPC I/O in the Data Center Workshop,
Hamburg, May 16, 2024

Glenn K. Lockwood, Microsoft
Subramanian Kartik, VAST Data
(Proxy: Sven Breuner, VAST Data)

LLM Checkpointing – Megatron-LM Deployment Model

GPT-3 175B Parameter Model – Example for 128 DGX Superpod (4 DGX-H100 SUs)

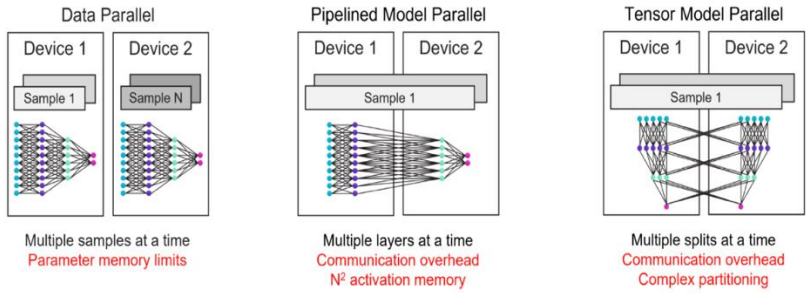


Figure 5 Existing scaling techniques on distributed GPU clusters and their challenges. Scaling on GPU clusters requires a complex combination of all forms of parallelism.

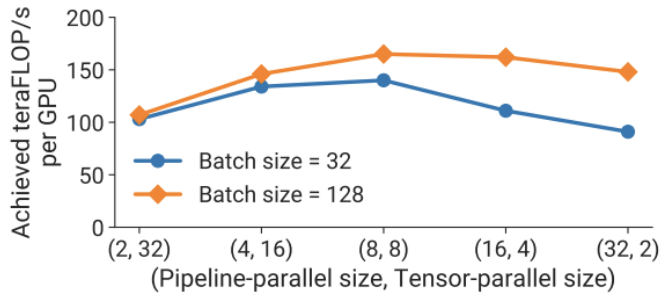


Figure 13: Throughput per GPU of various parallel configurations that combine pipeline and tensor model parallelism using a GPT model with 162.2 billion parameters and 64 A100 GPUs.

- Model Size Exceeds GPU VRAM
  - GPT-3 is ~350 GB (2 bytes/parameter)
  - Gradients, optimizer state etc increase in-mem state size by 7x
- Model Is Very Deep (~90 Layers)

- Tensor Model Parallel
  - Shard Model Across 8 GPUs In A DGX
- Pipeline Parallel – N DGXs in a Pipeline Parallel Group
  - N=16 typically for GPT-3 sized models
- Data Parallel Across the pipeline parallel groups

ONLY ONE PIPELINE PARALLEL GROUP of GPUs NEED TO BE CHECKPOINTED
RESTORE NEEDS ALL GPUs TO BE REPOPULATED

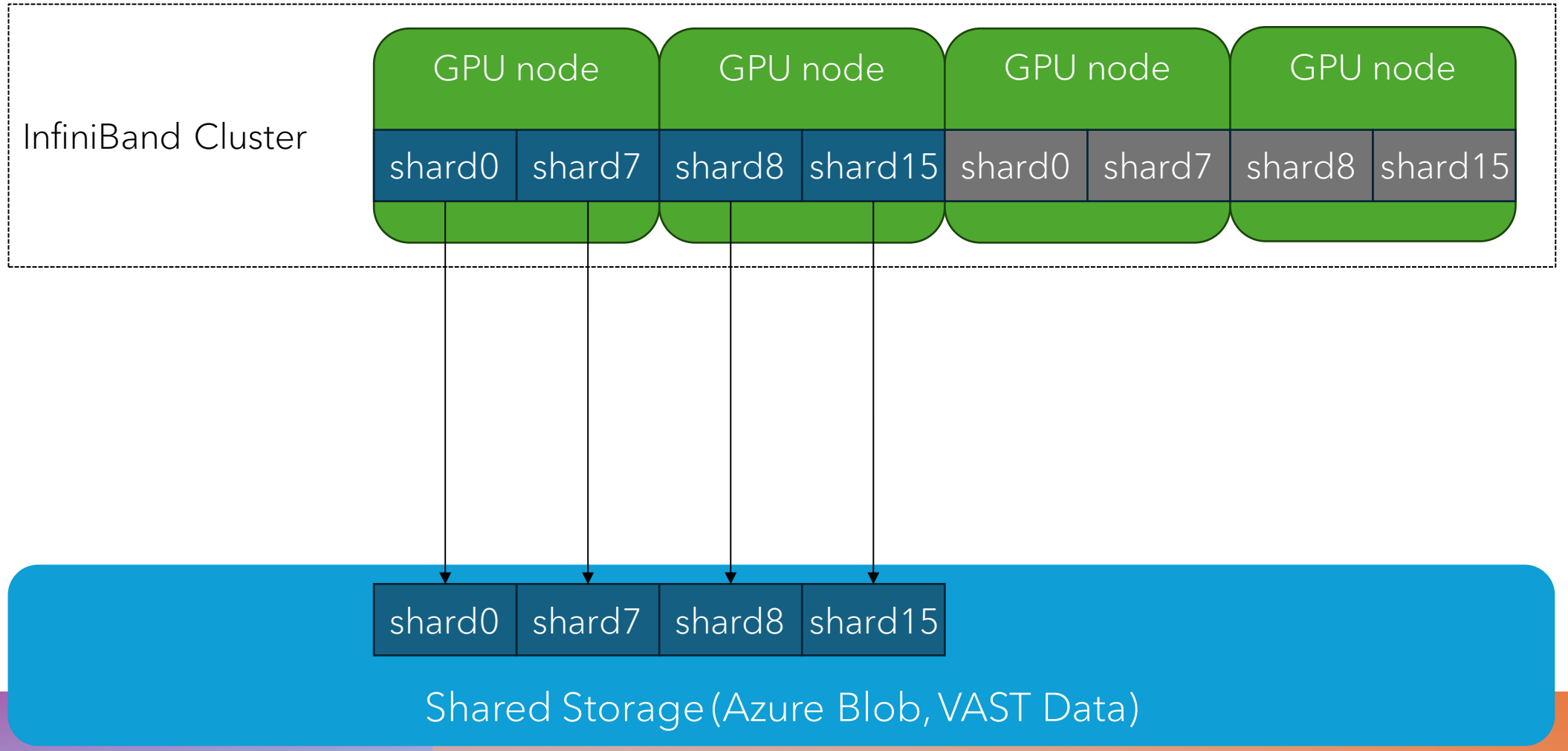
https://arxiv.org/pdf/2104.04473.pdf

The LLM Checkpointing Sizer

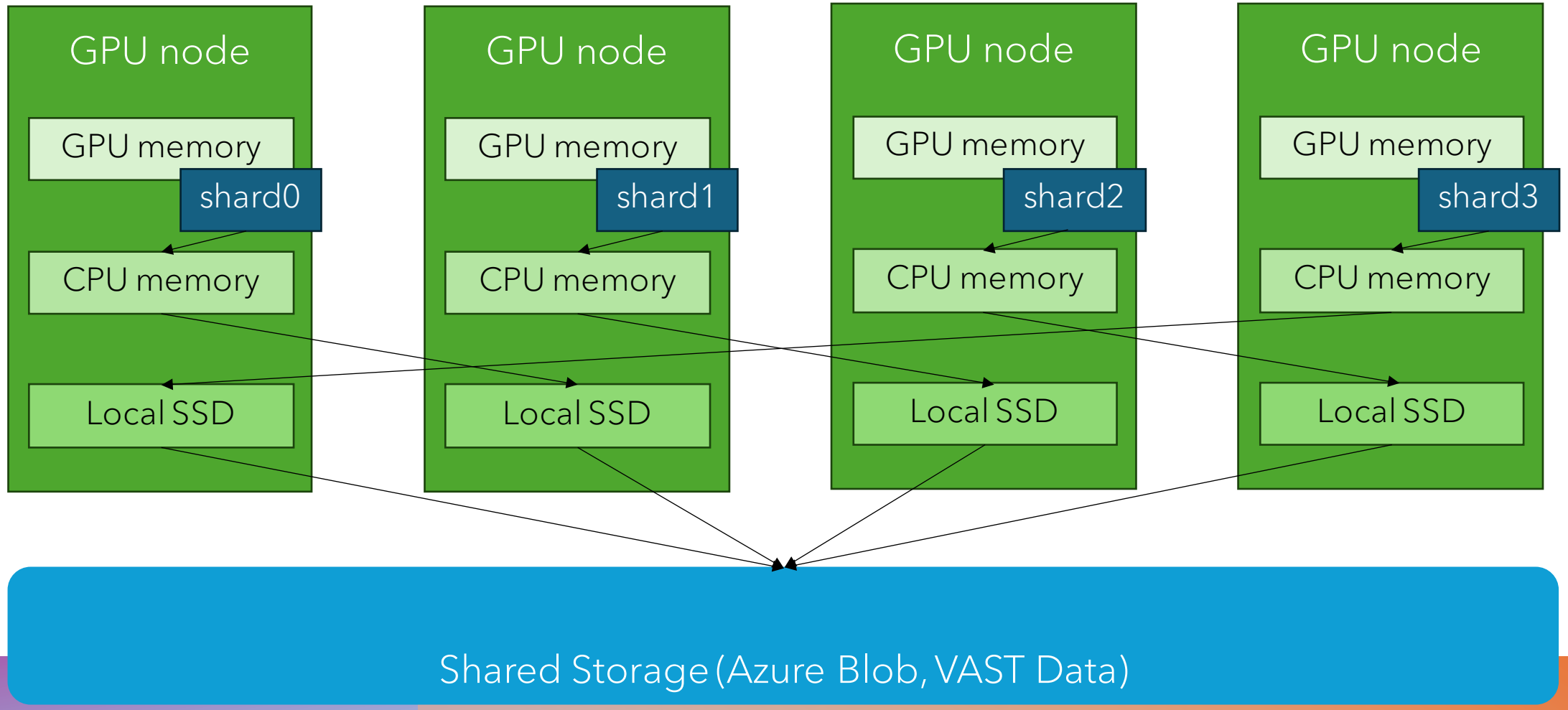
https://shorturl.at/gmzZ7

LLM CHECKPOINTING IN PRACTICE

CHECKPOINTING DIRECTLY TO SHARED STORAGE

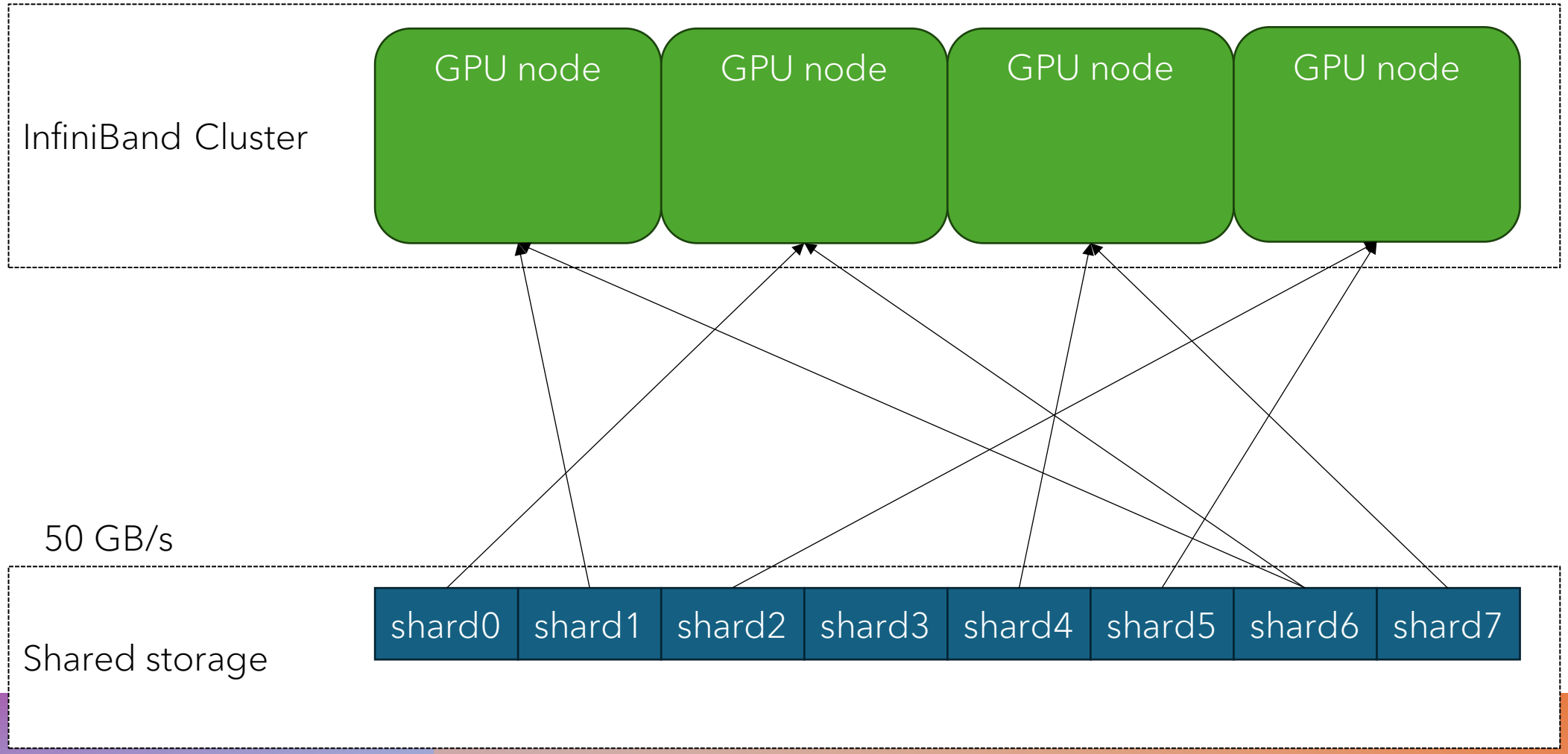


HIERARCHICAL CHECKPOINTING

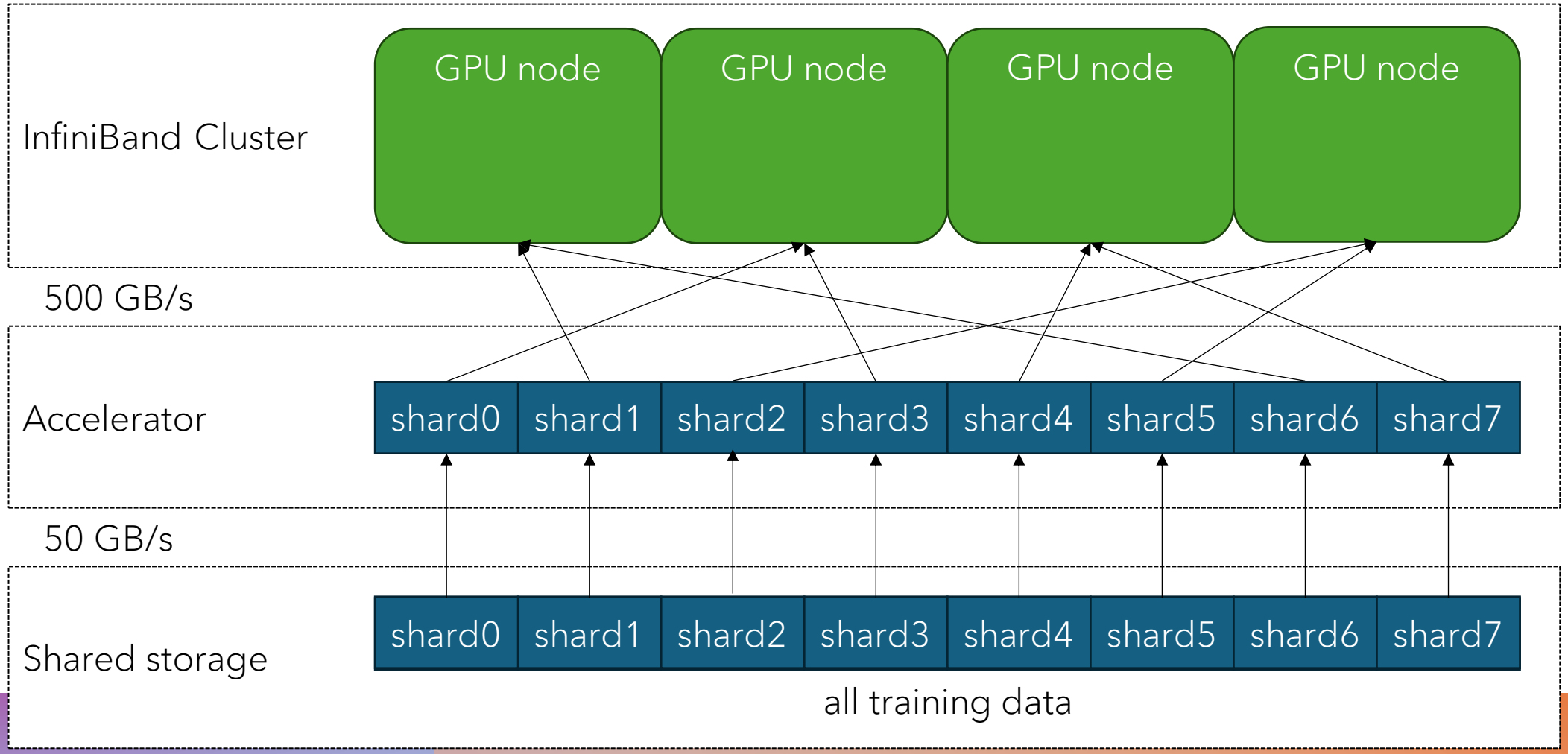


ISN’T TRAINING READ-HEAVY?

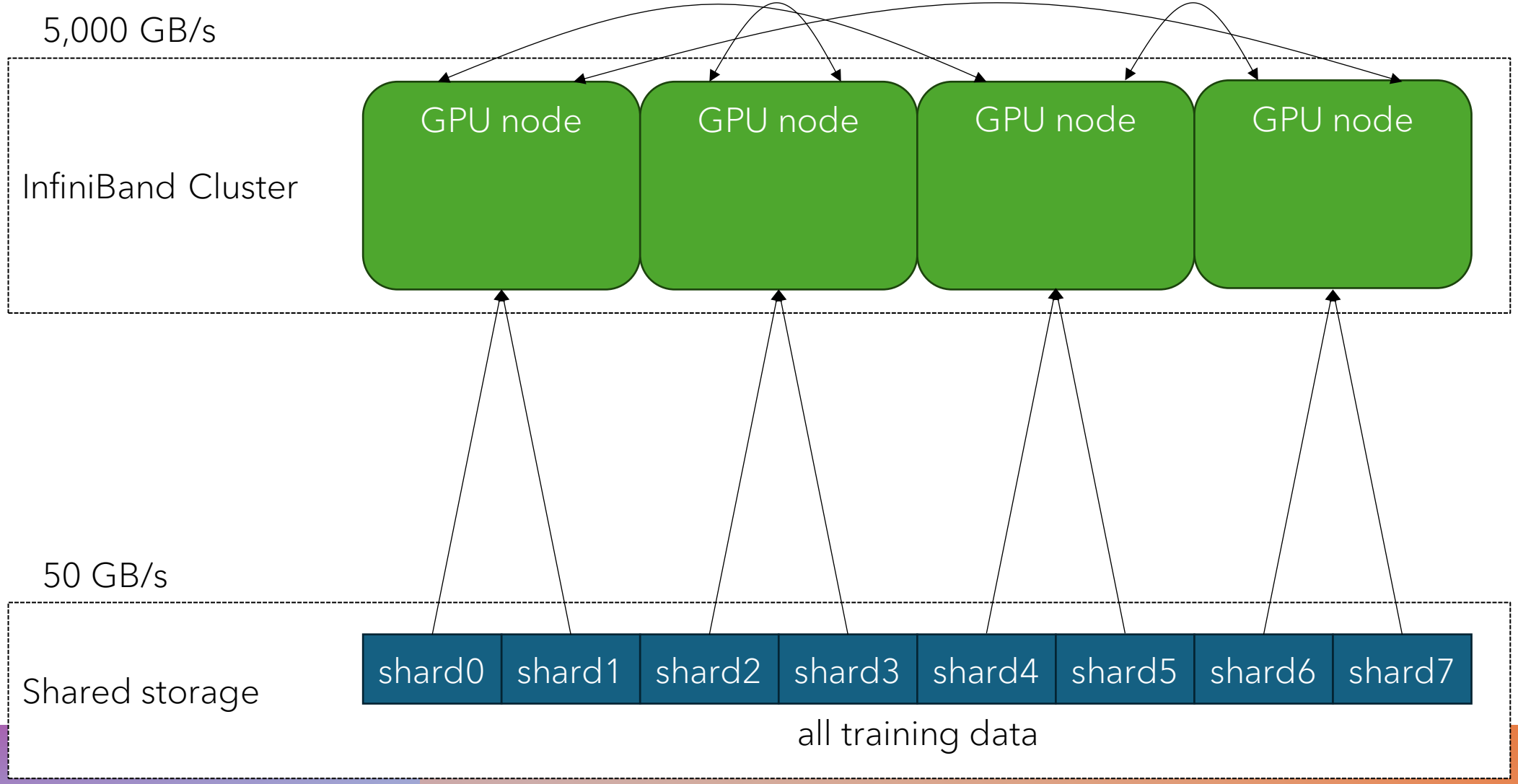
TRAINING MODELS DIRECTLY FROM SHARED STORAGE



TRAINING MODELS WITH AN INTERMEDIATE CACHE

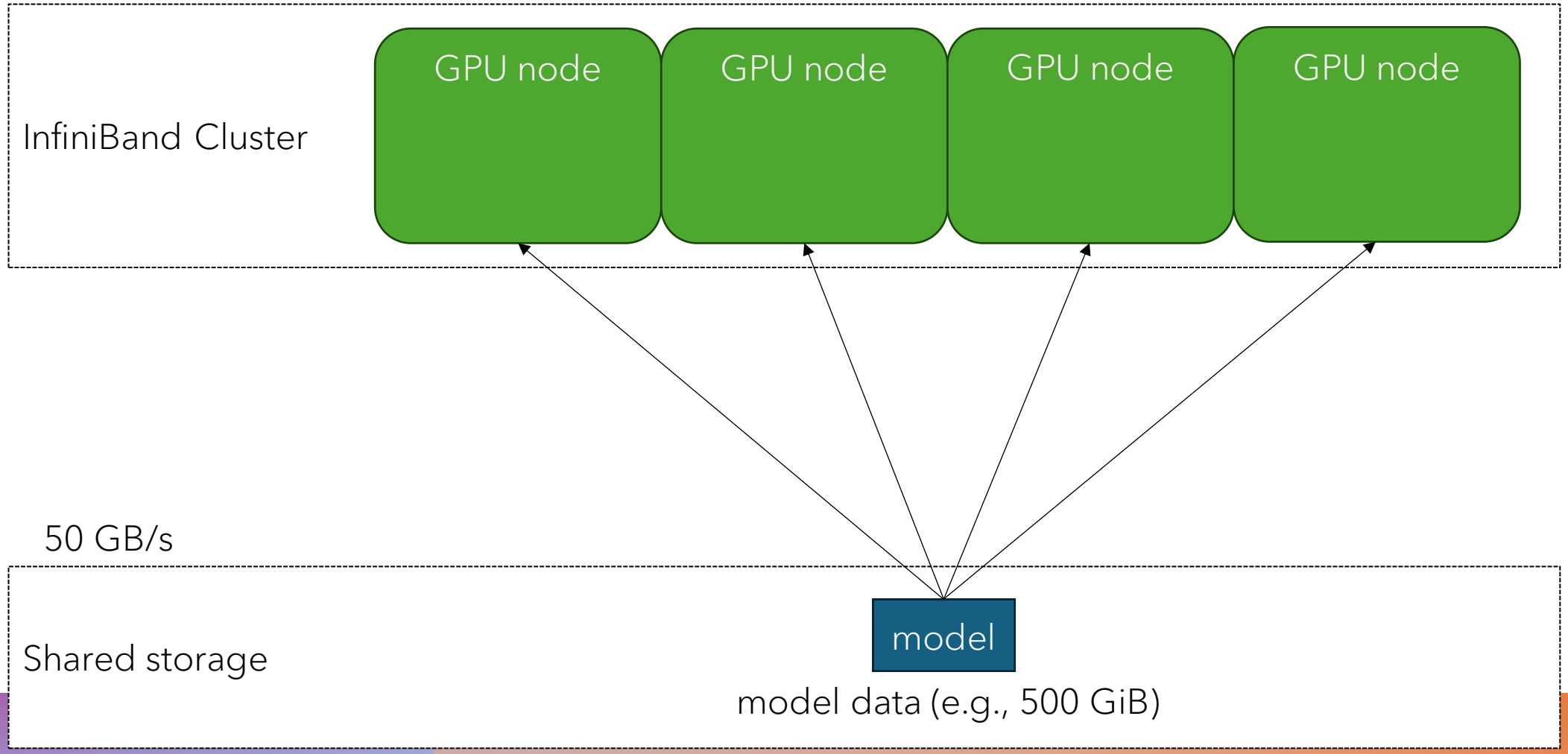


TRAINING WITH CLIENT-SIDE ACCELERATION

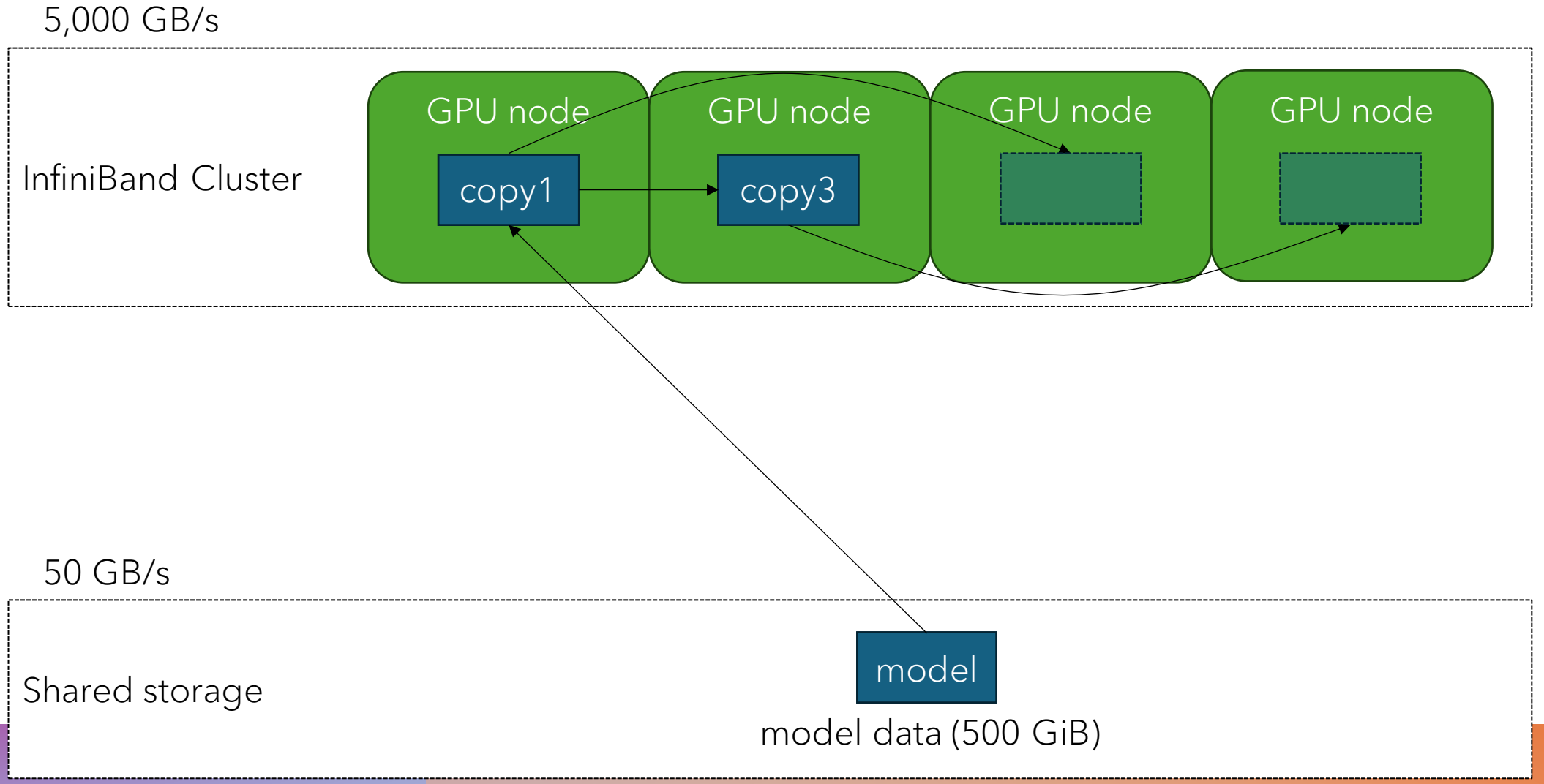


MODEL LOADING FOR INFERENCING

LOADING MODELS DIRECTLY FROM SHARED STORAGE



HIERARCHICAL MODEL LOADING



WHERE DOES HIERARCHICAL DATA MOVEMENT BREAK DOWN?

Multimodal models that train on low-density data **modalities**

Services built around inferencing that rely on state or external data

CONCLUSION

- We Challenge One-Size-Fits-All Advice
  - Let the data drive what performance and capacity requirements LLMs really need to handle Checkpointing
  - Advocate for decisions based on data, not dogma
- Understanding LLM Behavior
  - Emphasizes calculating LLM behavior from first principles and real data
  - Rejects rationale-less guidance for LLM training requirements

REFERENCES

Academic References and Their Online Sources		
Author(s)	Year	Source Link
He et al.	2023	Link
Narayanan et al.	2021	Link
Dash et al.	2023	Link
Kaplan et al.	2020	Link
Hoffmann et al.	2022	Link
Maurya et al.	2023	Link
Wang et al.	2023	Link