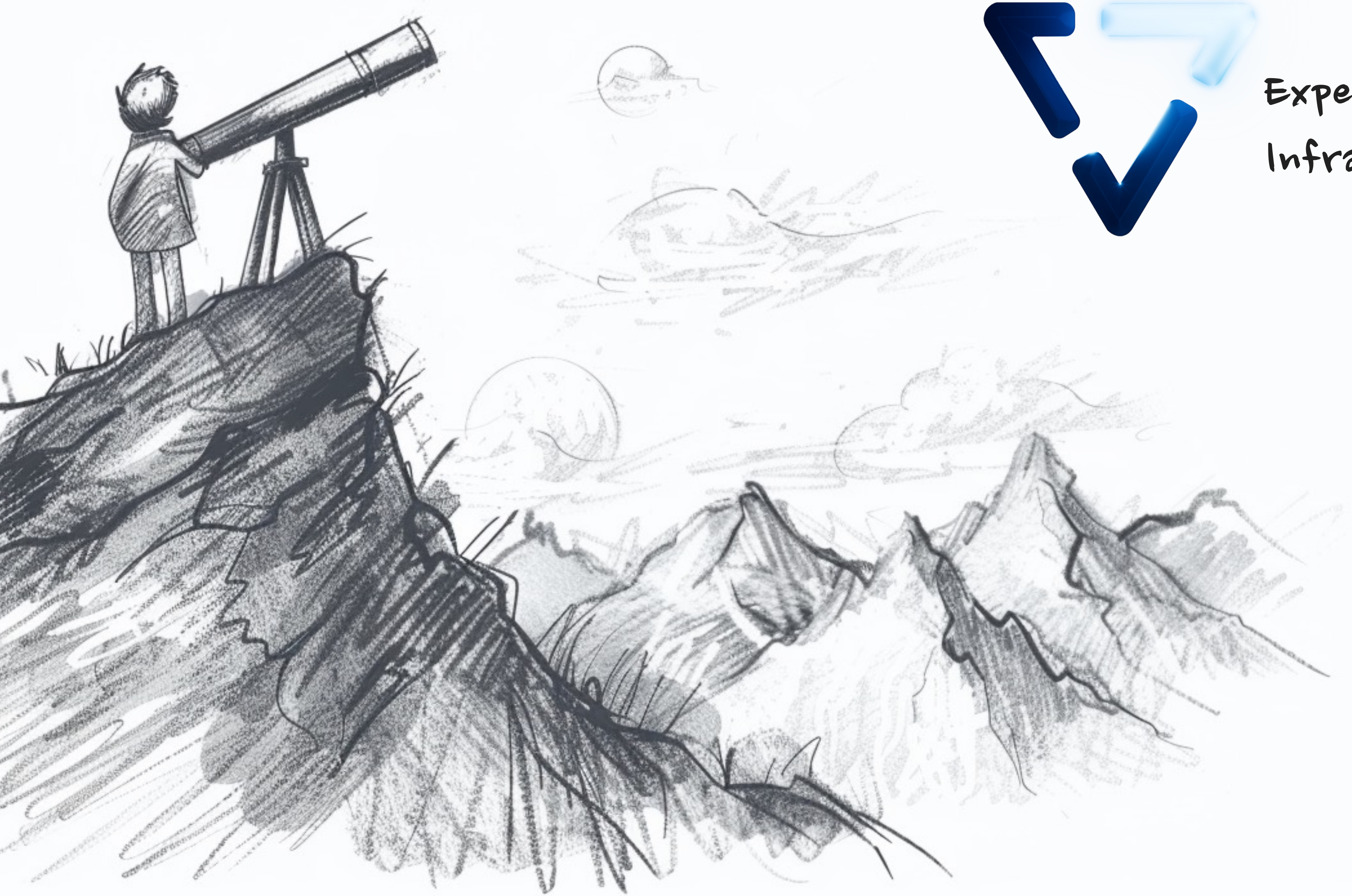
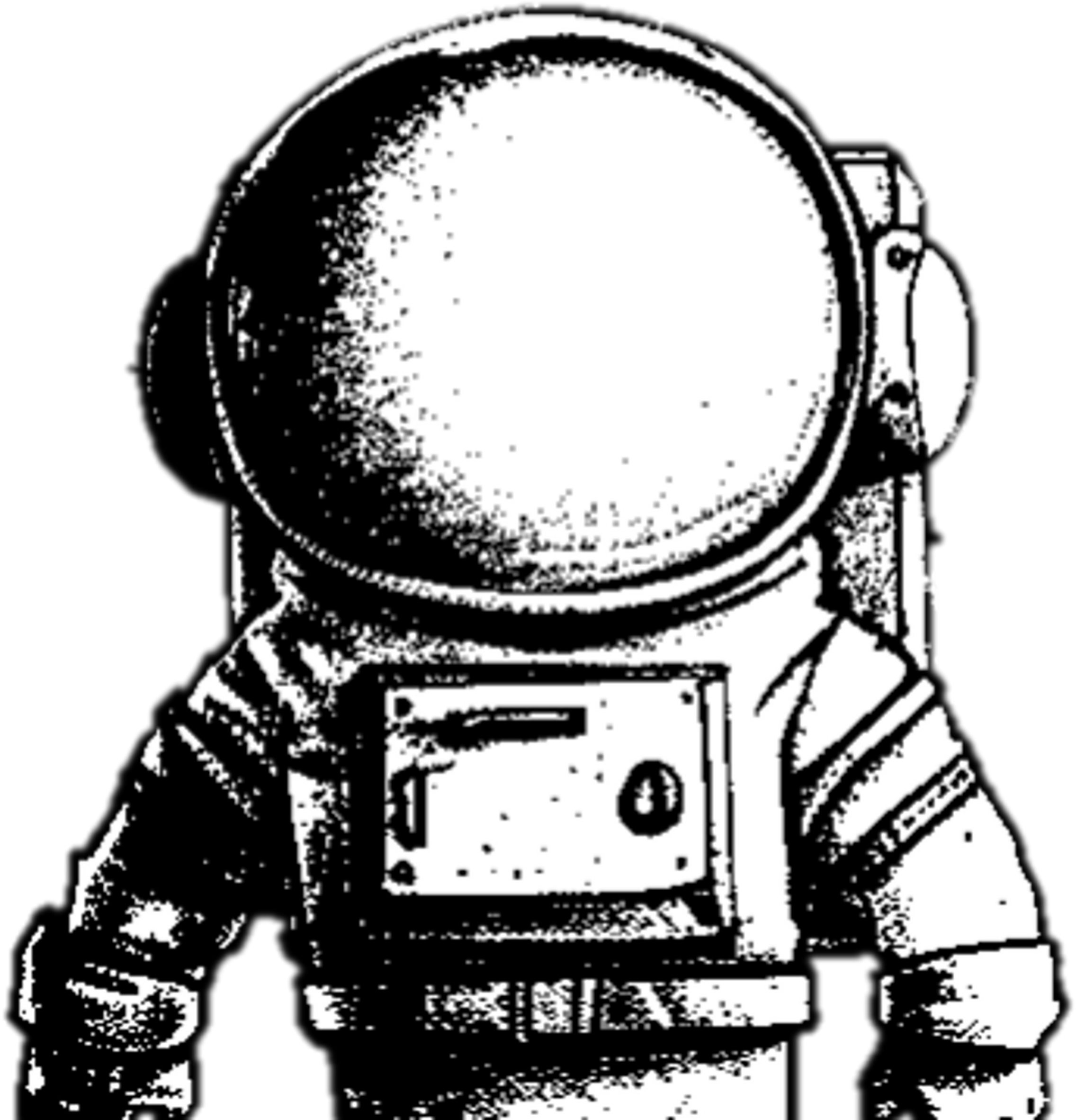


Experiences Building AI Data Infrastructure at Exascale

Forum TERATEC 24
Unlock the future

Jeff

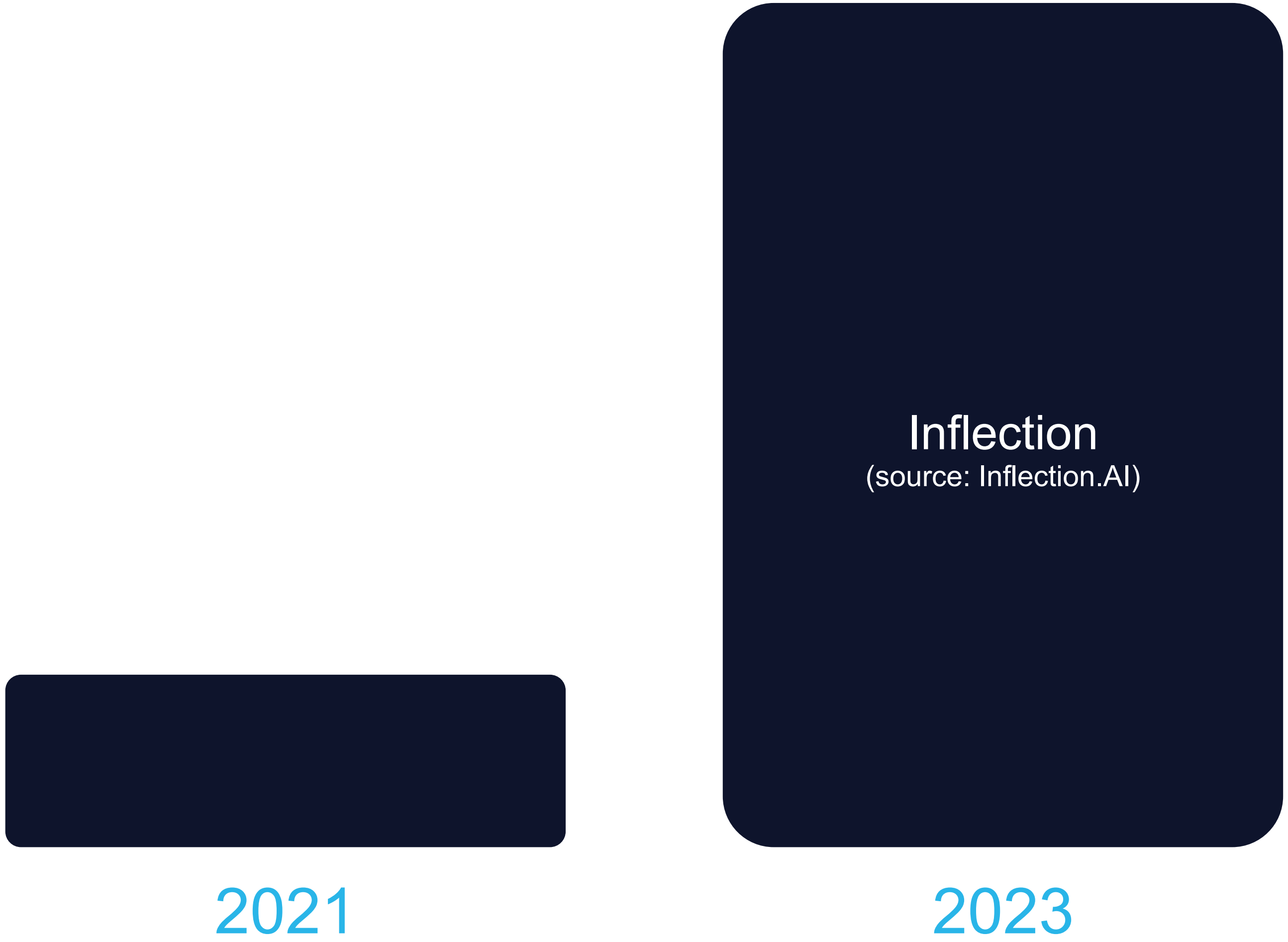
Co-Founder: VAST Data



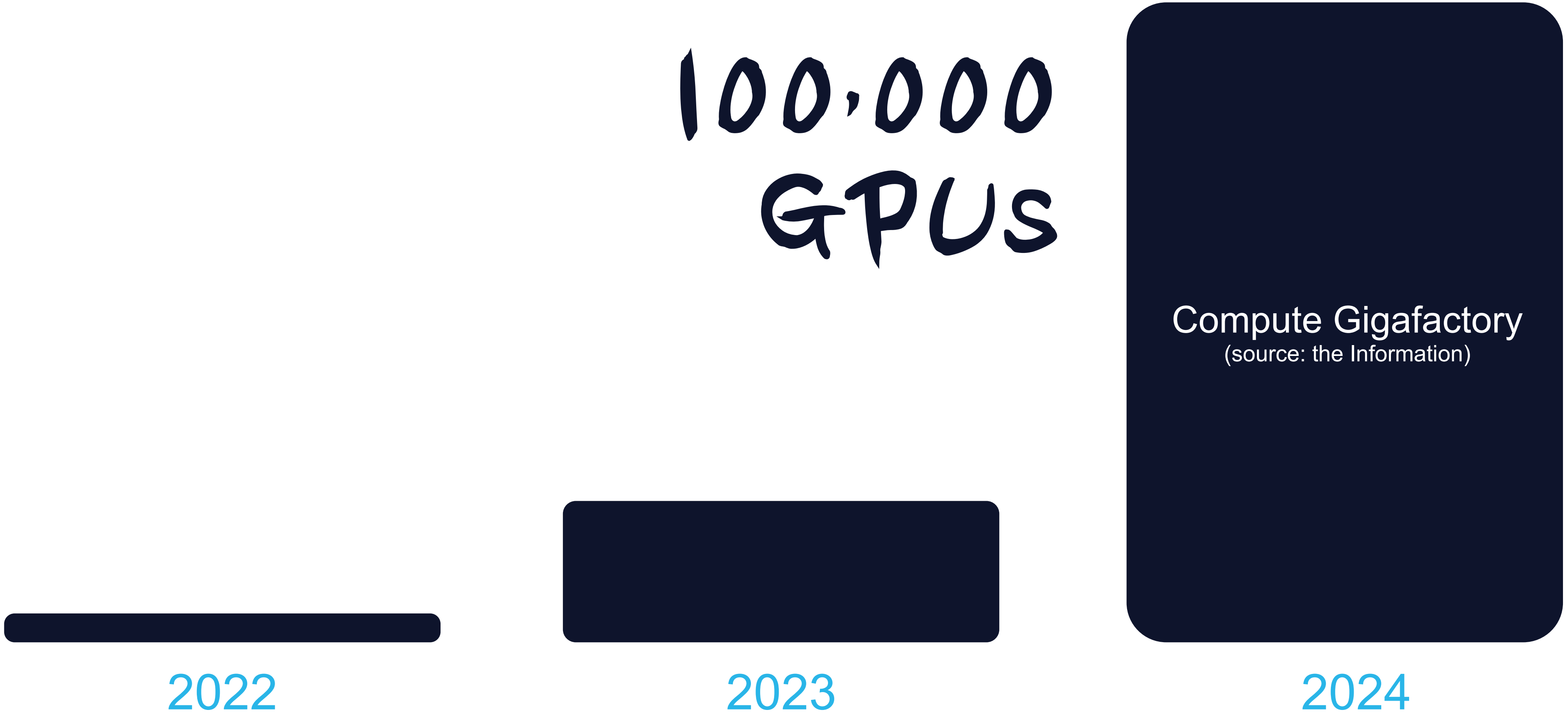
Selene
(source: NVIDIA)

2022

4,500 GPUs



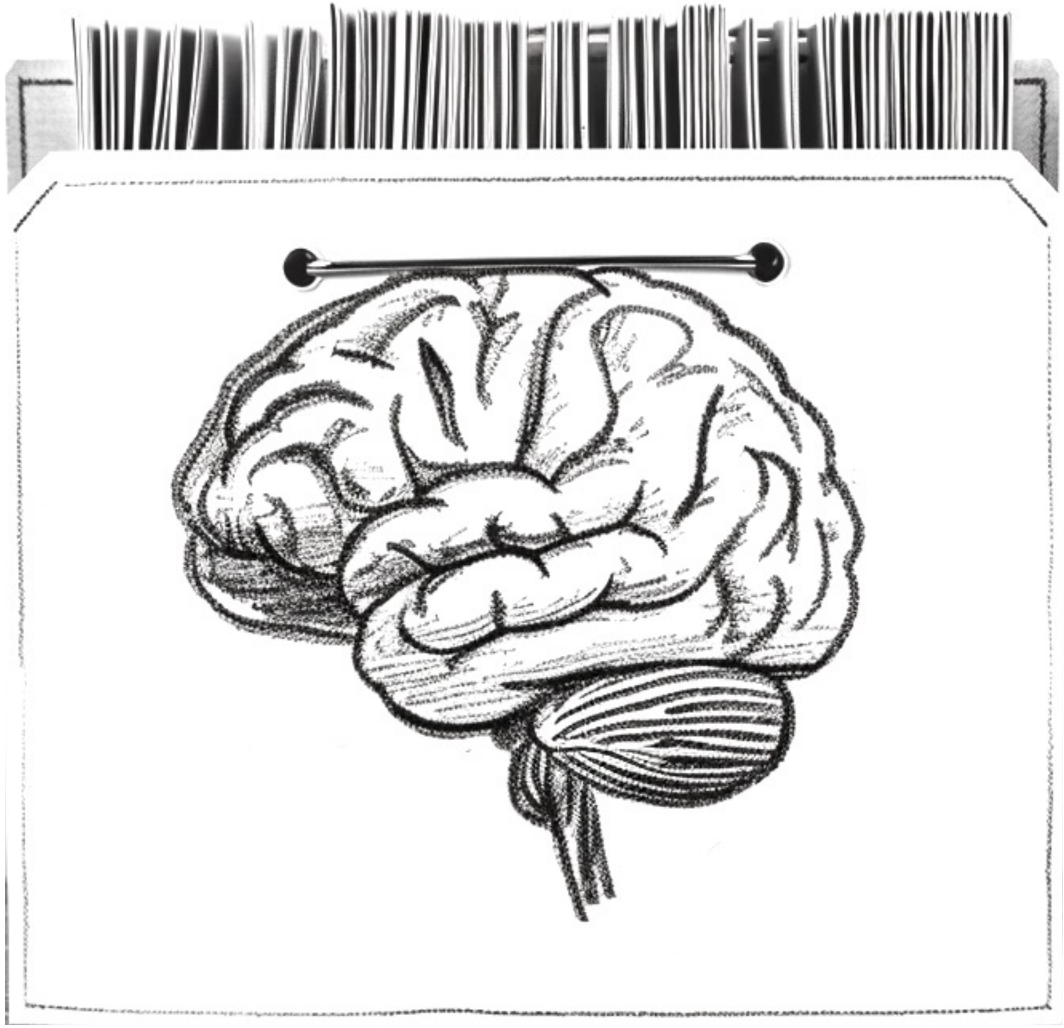
22,000 GPUs

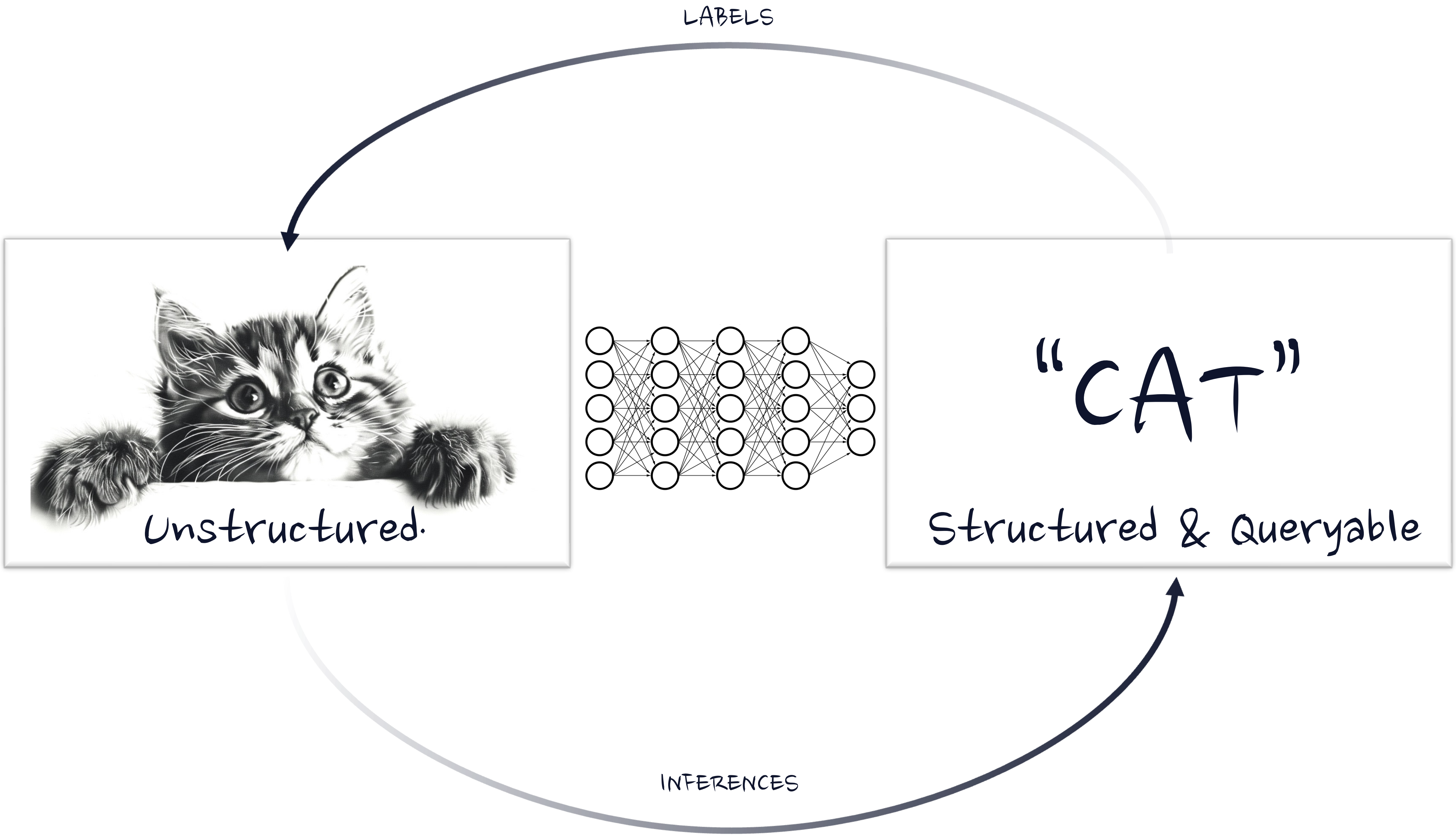

100,000
GPUs
Compute Gigafactory
(source: the Information)
2022
2023
2024

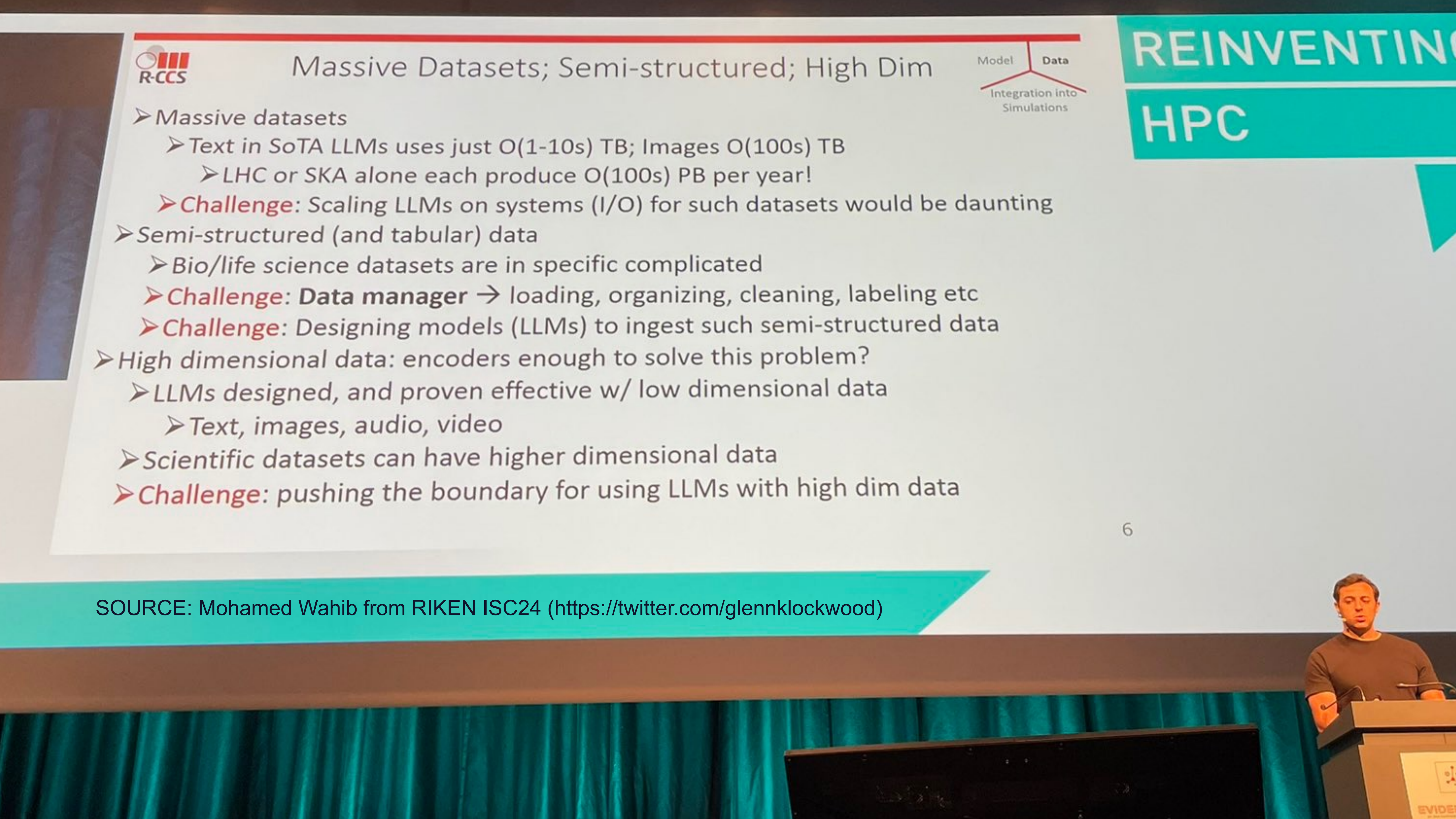
$10Bs of AI Infrastructure • 100Ks of GPUs

CoreWeave
core42
Lambda
Genesis Cloud
Crusoe
DigitalOcean
TAIGA CLOUD BY NORTHERN DATA GROUP
ampz
HYPERSCALE NEXUS
NEBUL
GREENNODE
ETC.

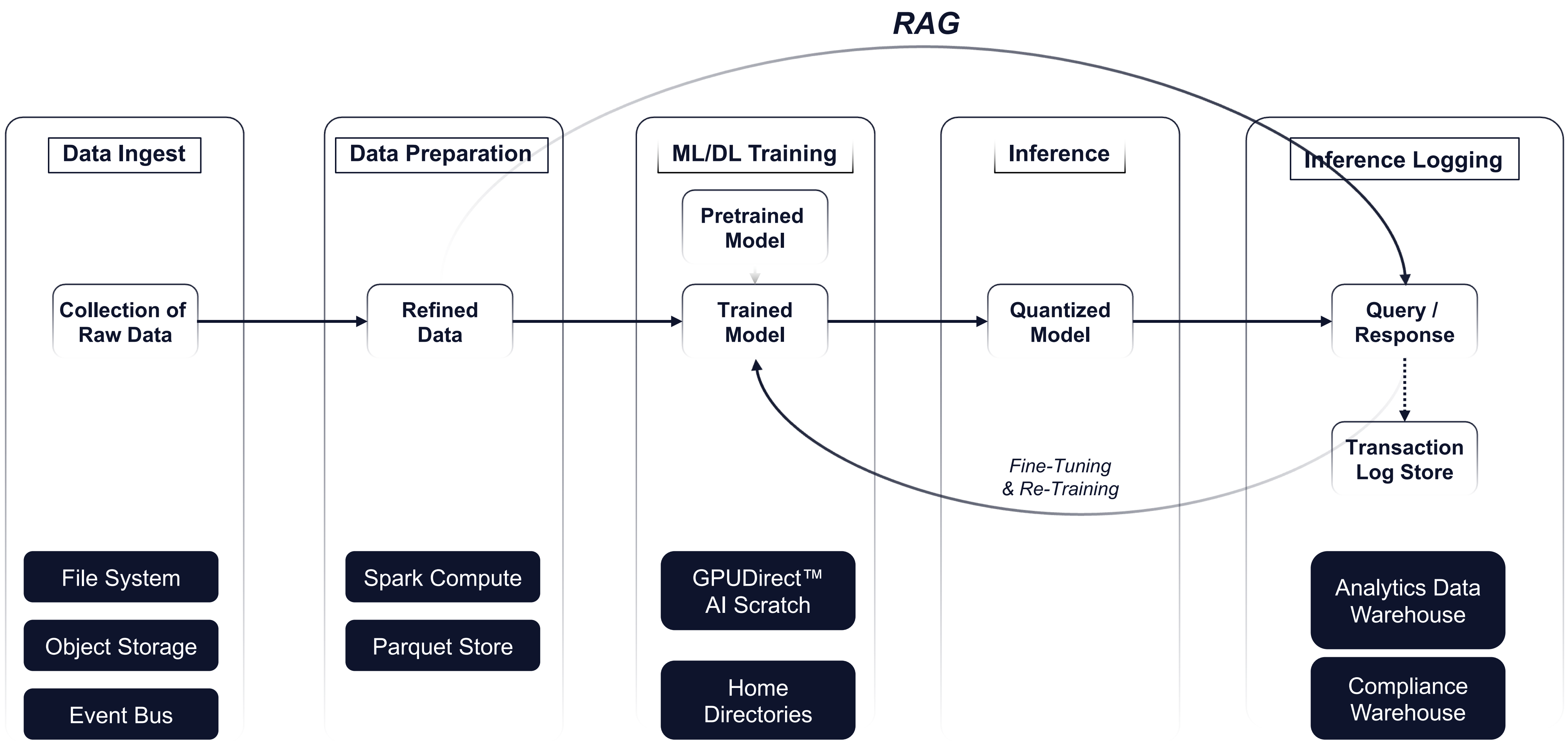
Today's AI prescription lacks depth.

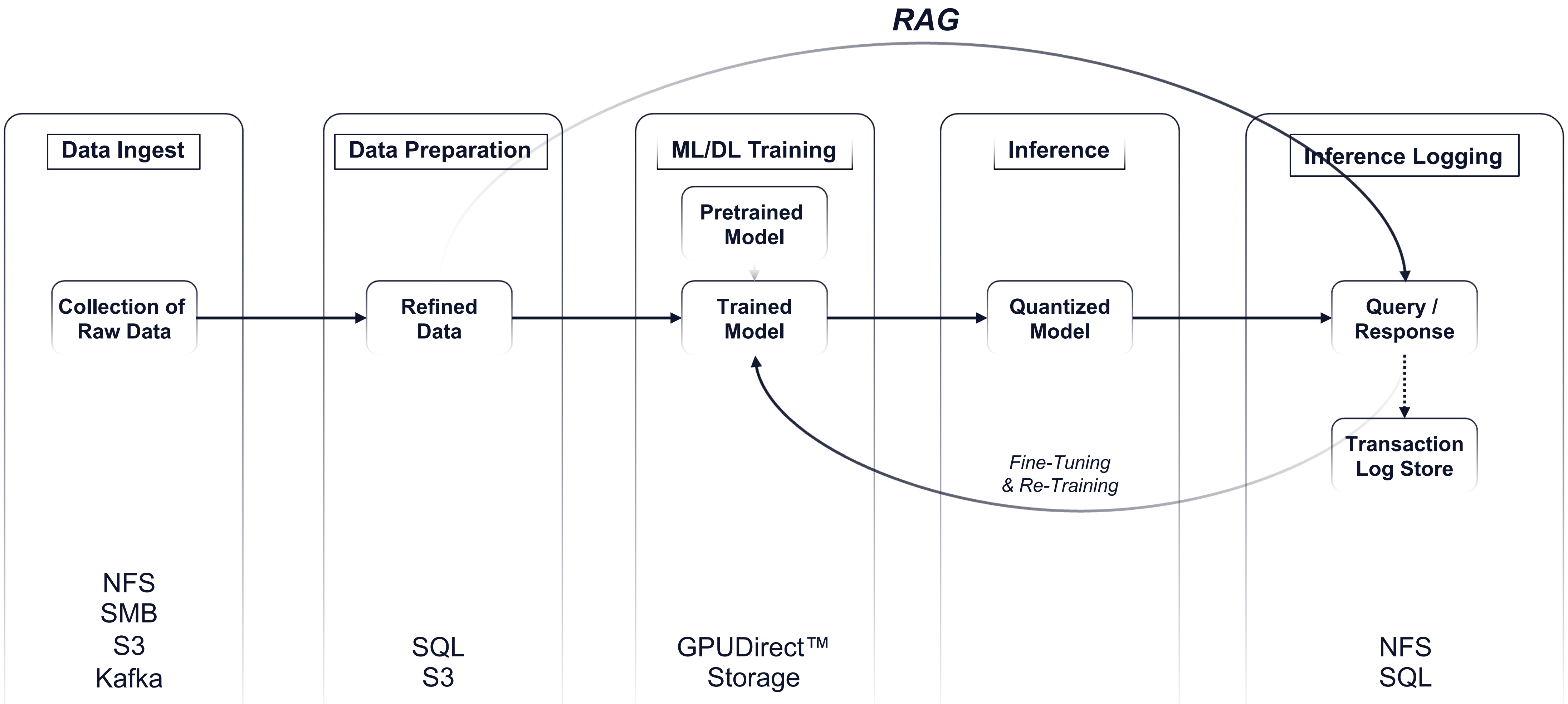



LABELS
Unstructured.
"CAT"
Structured & Queryable
INFERENCES



SOURCE: Mohamed Wahib from RIKEN ISC24 (https://twitter.com/glennklockwood)


RAG
Data Ingest
Data Preparation
ML/DL Training
Inference
Inference Logging
Pretrained Model
Collection of Raw Data
Refined Data
Trained Model
Quantized Model
Query / Response
Transaction Log Store
Fine-Tuning & Re-Training
File System
Object Storage
Event Bus
Spark Compute
Parquet Store
GPUDirect™ AI Scratch
Home Directories
Analytics Data Warehouse
Compliance Warehouse



VAST DATA PLATFORM: ACCELERATE ACCESS ACROSS PROTOCOLS + LANGUAGES

Power will change infra consumption.

← 100MW

The Rise of Cloud AI

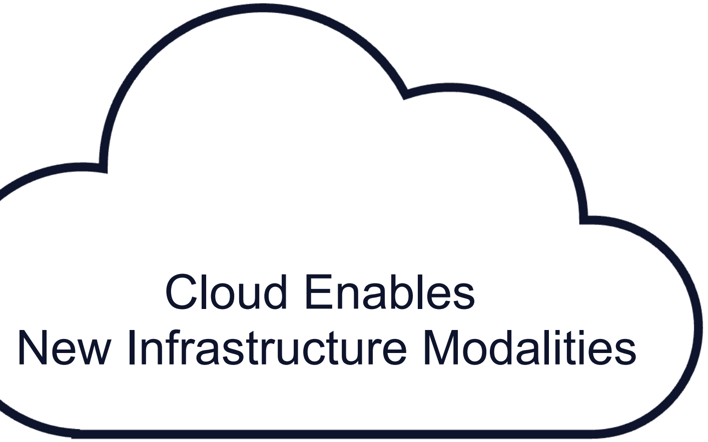
Confidential computing infrastructure

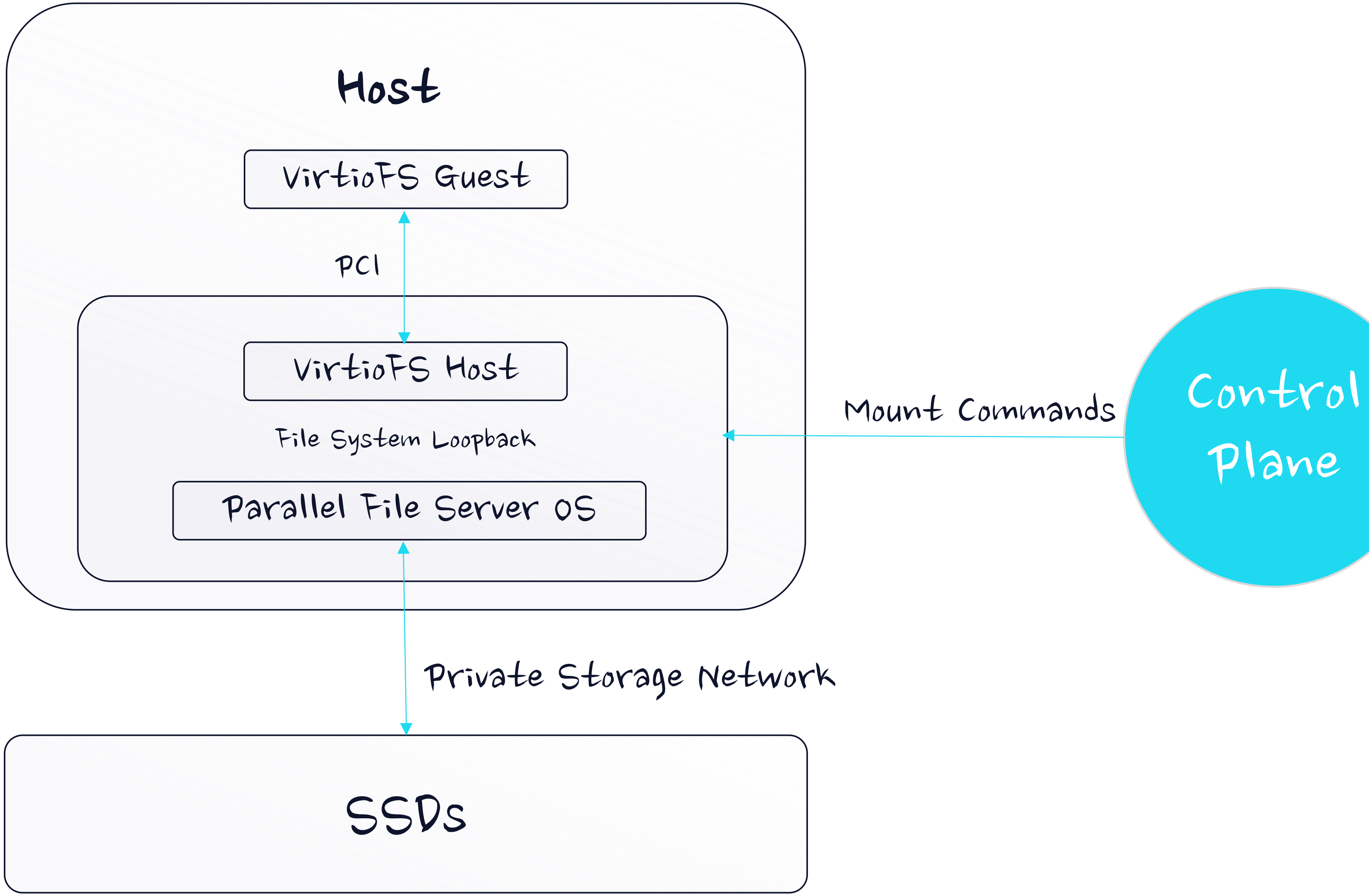
Multi-tenant, host-based isolation, EKM

Data Mobility

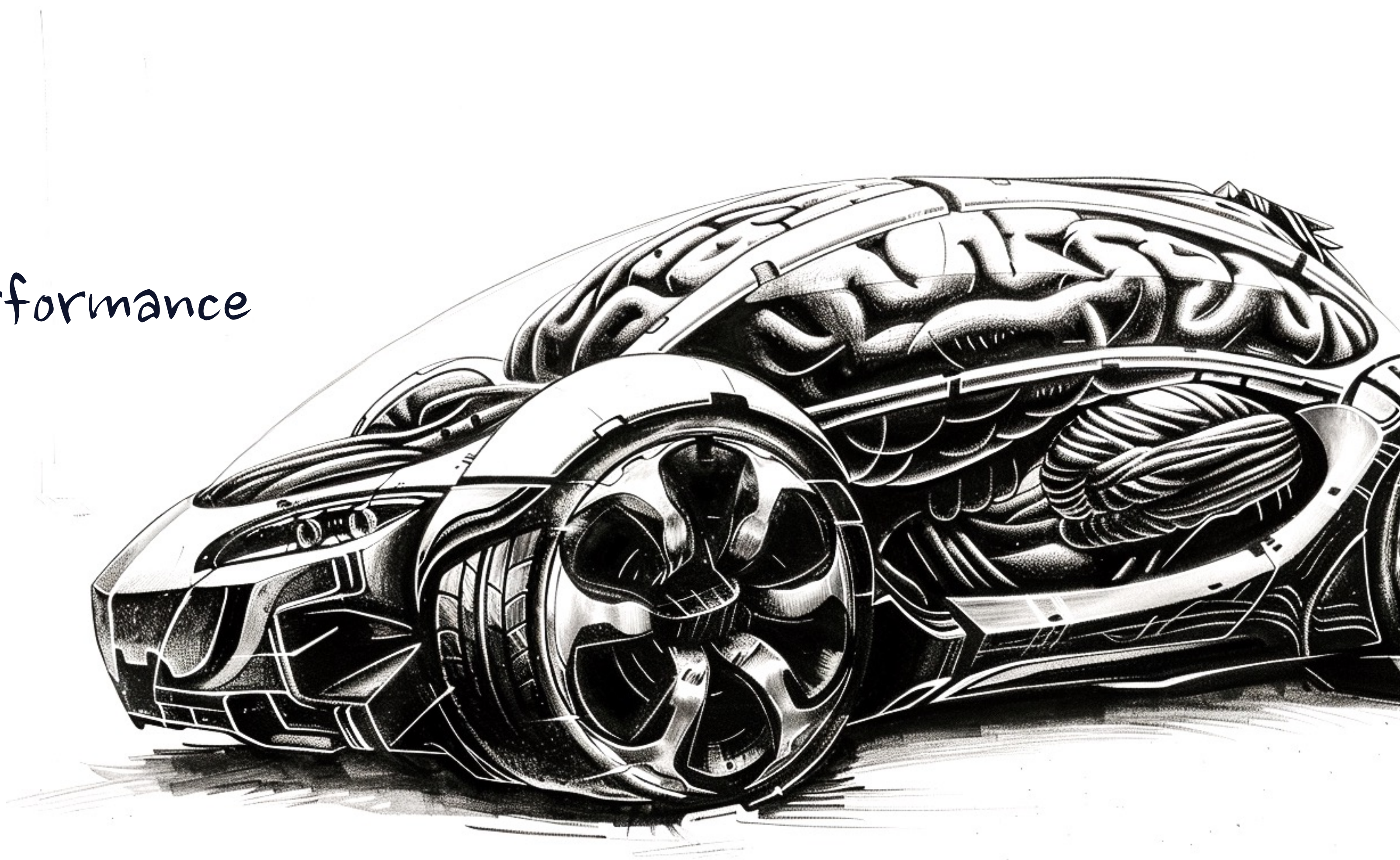
Solve for processor and power gravity

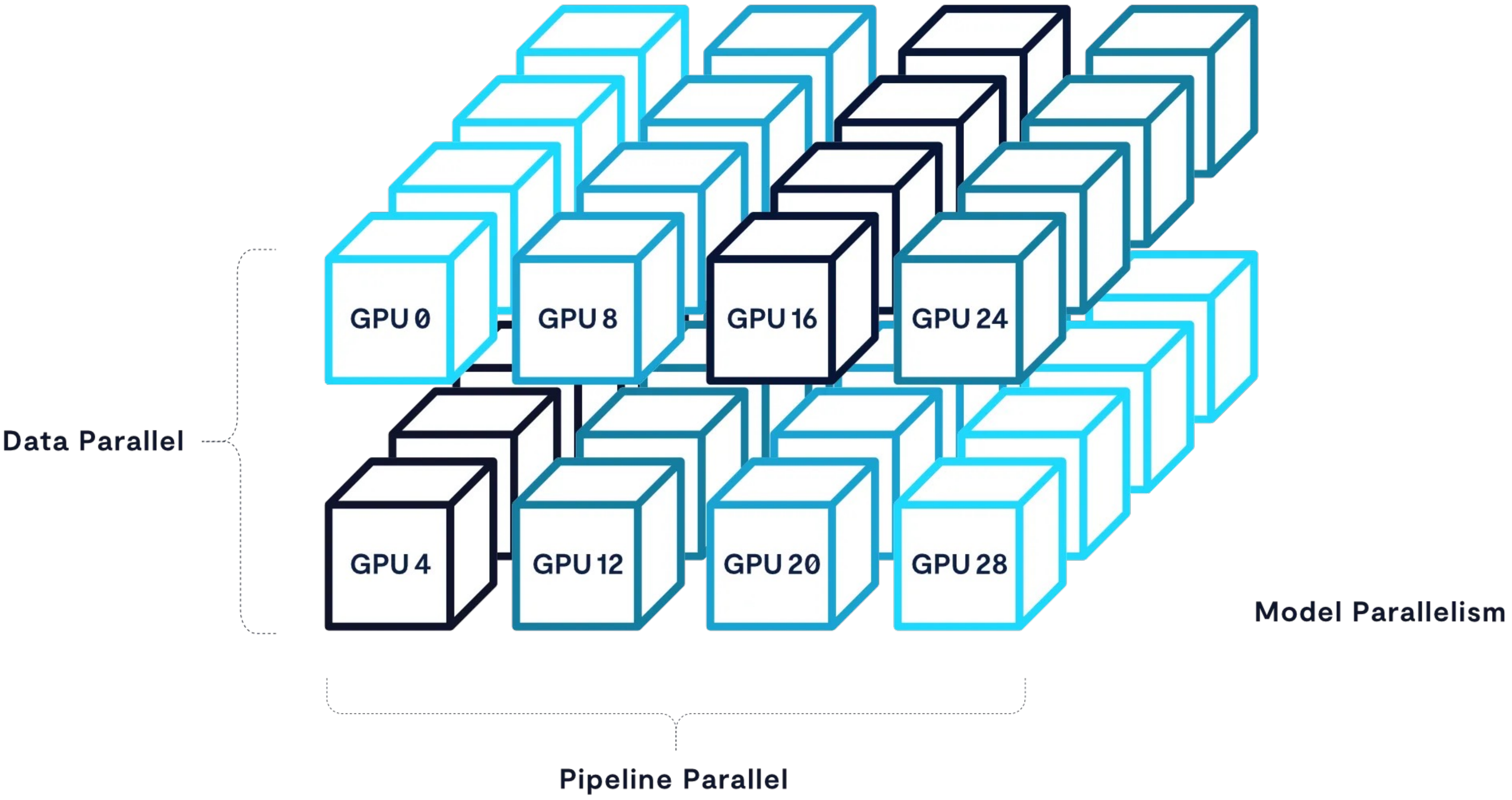
AI Factory Architecture





AI performance ≠ HPC performance



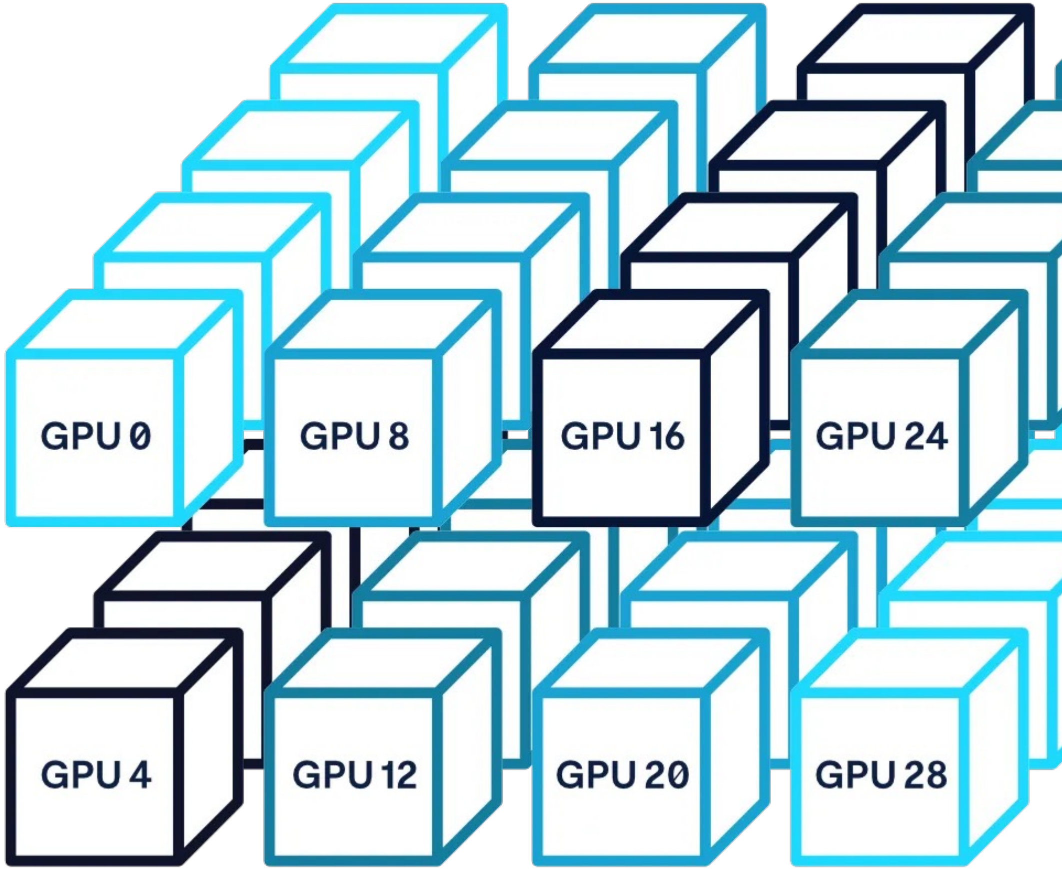


GPU 0
GPU 8
GPU 16
GPU 24
GPU 4
GPU 12
GPU 20
GPU 28
Data Parallel
Model Parallelism
Pipeline Parallel

Pipeline Parallel octets are self-contained and includes the entire model and pipeline for an LLM, and hence **only one octet needs to be checkpointed per system**, regardless of the overall size of the cluster because it is a complete representation of the system.

One thread per GPU.
Eight threads per 8-way GPU system.

Restores are **data parallel**.



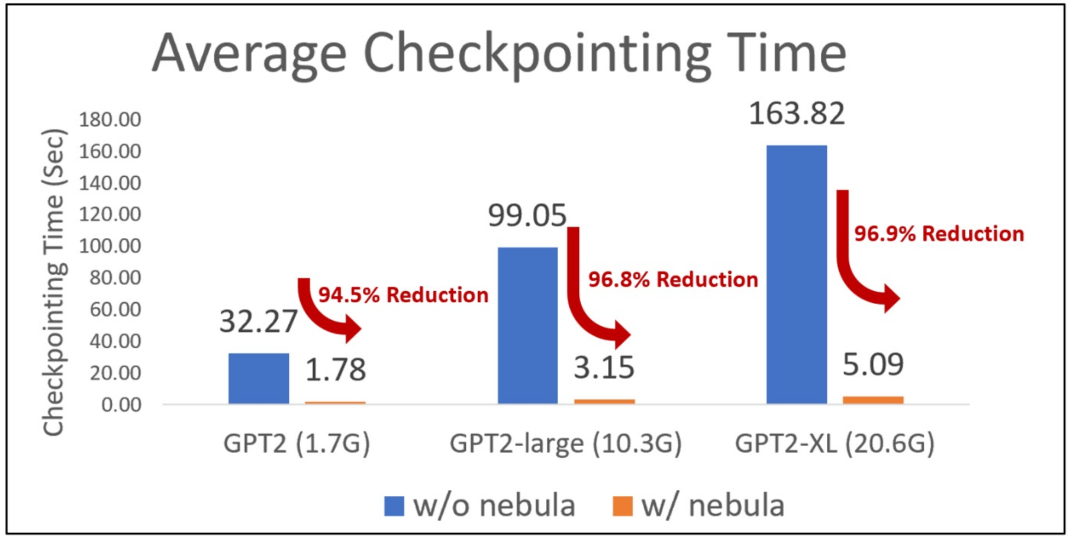
5.10 Checkpoint Loading and Saving

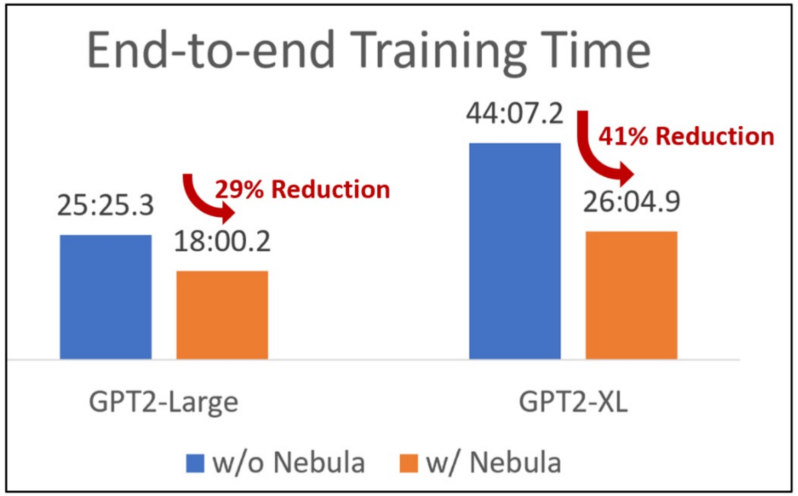
An important practical consideration for the training of large models is loading and saving model checkpoints, which are especially large for the models considered in this paper. For example, the trillion-parameter model has a checkpoint of size 13.8 terabytes. The initial load of checkpoints for the trillion-parameter model by all 384 nodes (3072 GPUs) reaches a peak read bandwidth of 1TB/s, the maximum read throughput possible from the parallel filesystem. Checkpoint saves reach 40% of peak write bandwidth (273 GB/s).

Source: NVIDIA + MICROSOFT + STANFORD Megatron-LM paper

Checkpoints are going asynchronous.

Nebula Saving with DeepSpeed and ORT





PS: Saving 4 checkpoints in an end-to-end training, with saving every checkpoint by 500 steps.

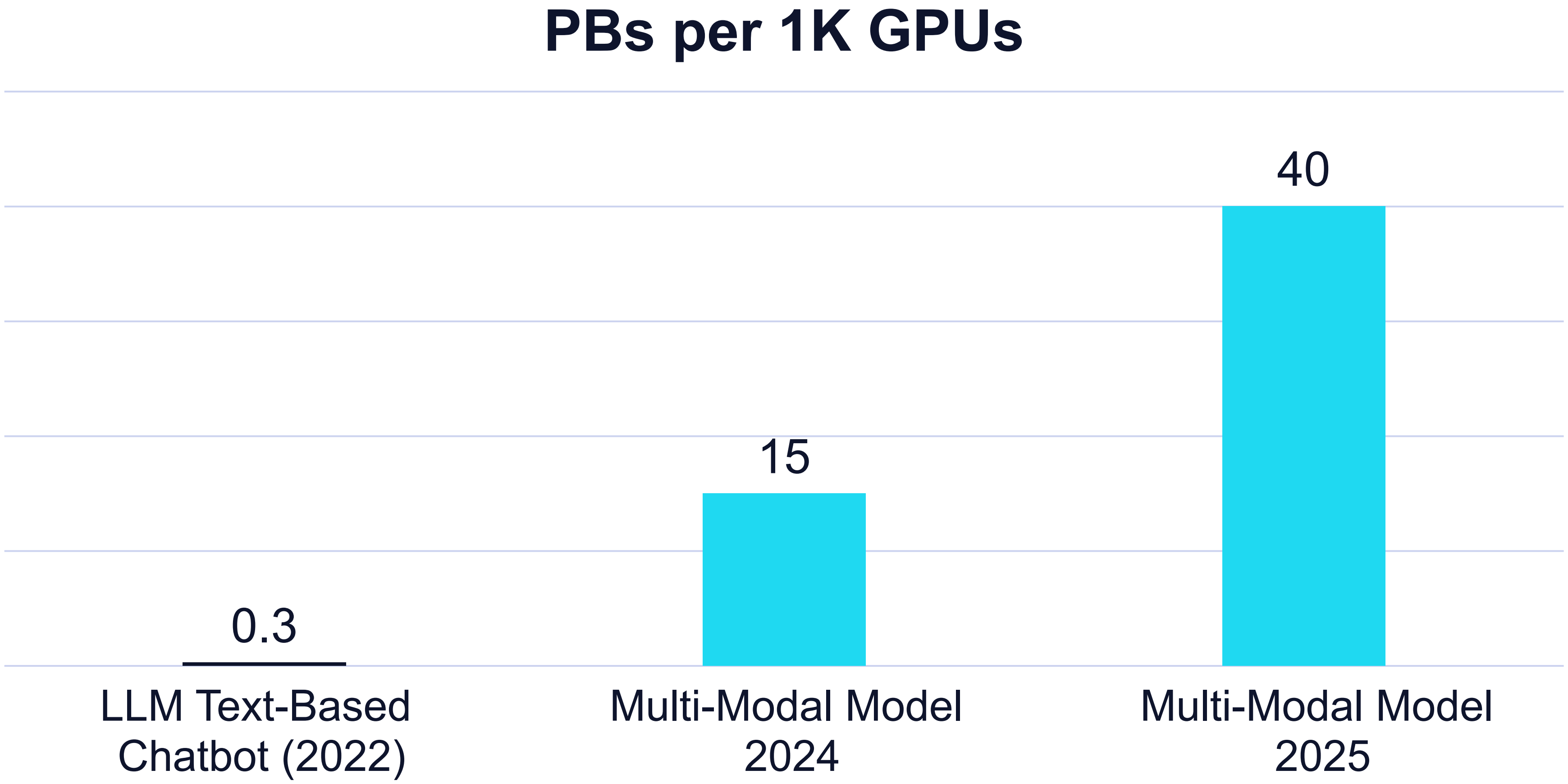
w/o Nebula: DeepSpeed and ORT enabled
w/ Nebula: DeepSpeed, ORT, and Nebula enabled

Environment: V100, 1 node, 8 GPUs
Image: ptebic.azurecr.io/public/azureml/aifx/stable-ubuntu2004-cu117-py38-torch1131:latest
DeepSpeed 0.8.0， ORT: 1.13.1， Nebula: 0.15.9

Source: Microsoft

New Trend: Multi-Modal

Video & Imagery are driving up LLM capacities quickly.



Conclusion #1:

Growth of system size & power increases will drive up MTBF

Conclusion #2:

Async makes write bandwidth, checkpoint-speed a non-issue

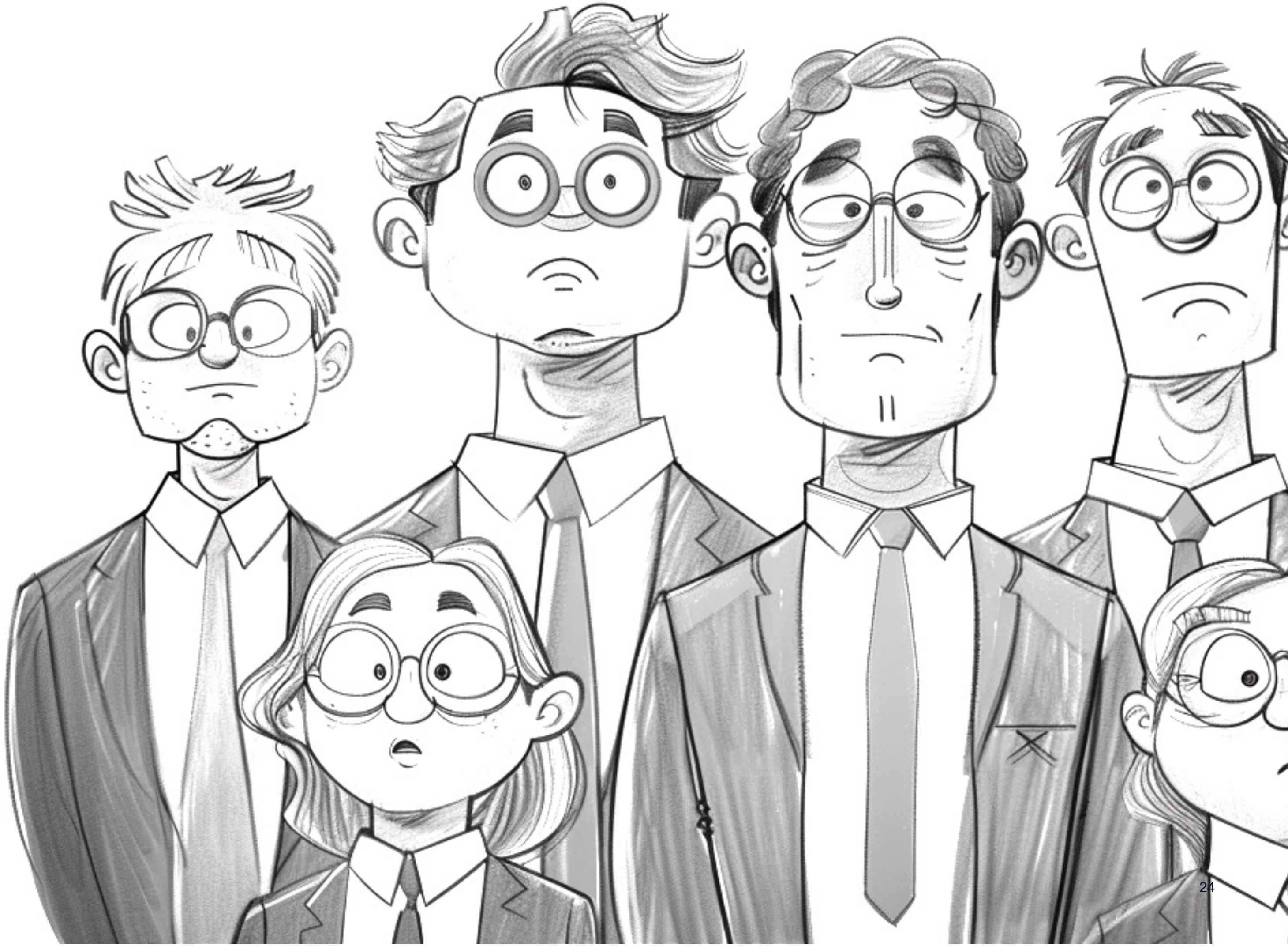
Conclusion #3:

Recovery Speeds will scale much larger as systems grow in size

Conclusion #4:

Multi-Modal: Capacity = Performance with All-NVME

AI's Sarbanes-Oxley Moment



> We think that regulatory intervention by governments will be critical to mitigate the risks of increasingly powerful models

Sam Altman –2023

EU AI Act: Risk-based approach

Article 72

Post-Market Monitoring by Providers and Post-Market Monitoring Plan for High-Risk AI Systems

2. The post-market monitoring system shall actively and systematically collect, document and analyse relevant data ***which may be*** provided by ***deployers or which may be*** collected through other sources on the performance of high-risk AI systems throughout their lifetime, and which allow the provider to evaluate the continuous compliance of AI systems with the requirements set out in Chapter III, Section 2

The “it” in AI models is the dataset.

Posted on June 10, 2023 by jbetker

Trained on the same dataset for long enough, pretty much every model with enough weights and training time converges to the same point… Model behavior is not determined by architecture, hyperparameters, or optimizer choices. **It’s determined by your dataset, nothing else.**

When you refer to “Lambda”, “ChatGPT”, “Bard”, or “Claude” then, it’s not the model weights that you are referring to. **It’s the dataset.**

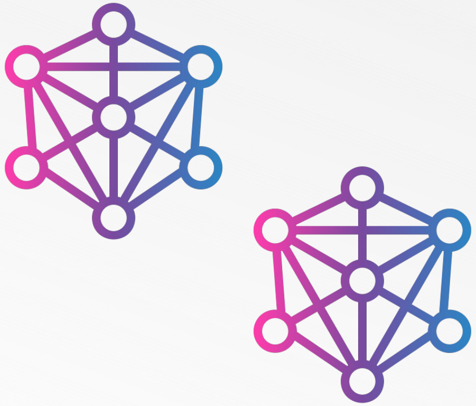
Cyber Resilience

Ransomware Protection

Data Provenance

Data Tampering Protection

Model Reproducibility



Pipeline Versioning

Cyber Resilience

<u>Ransomware Protection</u>

Indestructible Snapshots

Ransomware Detection

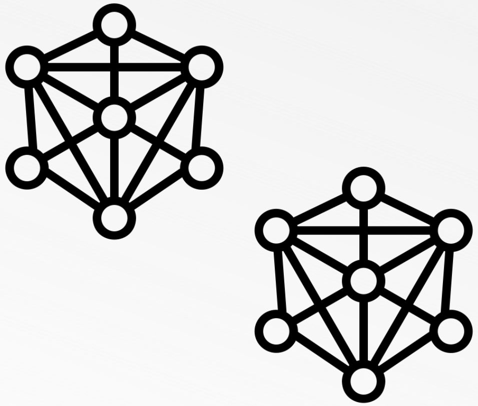
Data Provenance

<u>Data Tampering Protection</u>

WORM Data Retention

SQL-Accessible Audit Logs

Model Reproducibility



<u>Pipeline Versioning</u>

An enterprise-grade python dataset with immutability

Point #1:

Consider the whole pipeline

Conclusion #2:

Cloud is coming, whether we like it or not

Conclusion #3:

HPC Performance ≠ AI Performance, Multi-Modal Solves All

Conclusion #4:

Enterprise Data Management is Having Its Moment

Fastest-Growing in Infrastructure History

Enterprise-Grade Massively-Scalable Radically-Affordable Cloud-Native

Thank you!
www.vastdata.com