AI Data Pipeline

SOLIDIGM™

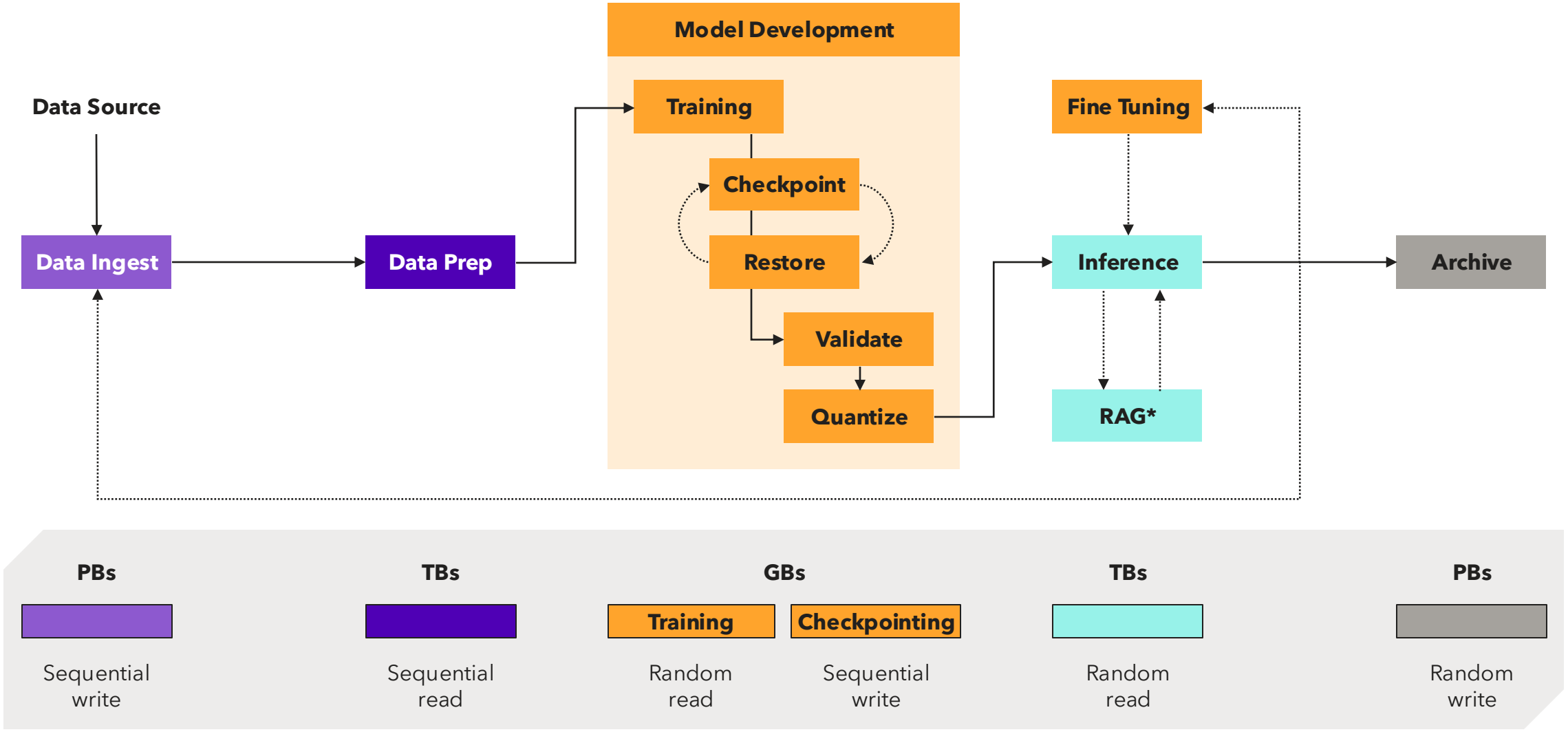
"An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem."

– John Tukey,
The Future of Data Analysis (1962)

"So come on and chickety-check yo self before you wreck yo self."

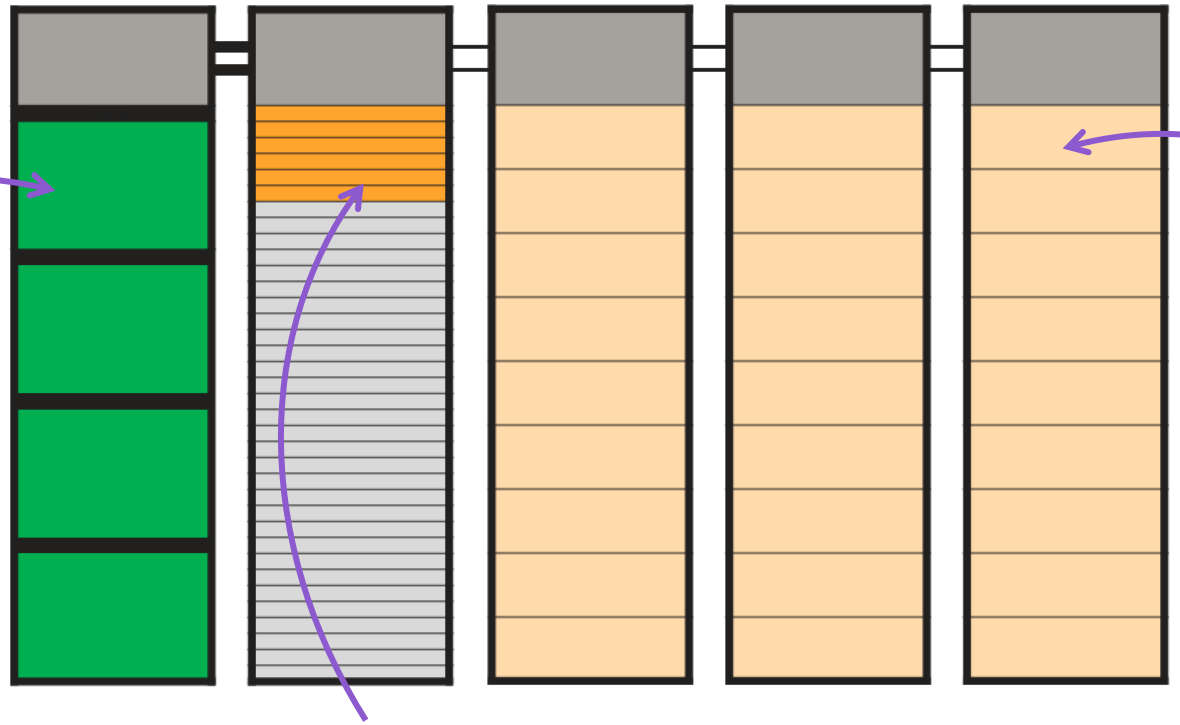
– O'Shea Jackson Sr,
Check Yo Self (1992)

Data Is Everywhere in the AI Pipeline



* Retrieval Augmented Generation

Storage Anatomy of a Typical AI Cluster



GPU Servers

Limited storage (usually 8x U.2 slots)

Object Storage Tier

Storage servers or JBODs containing many storage devices (usually HDDs today)

All-Flash Performance Tier

Servers containing fast storage devices (usually TLC NAND today) to offset HDD performance shortcomings

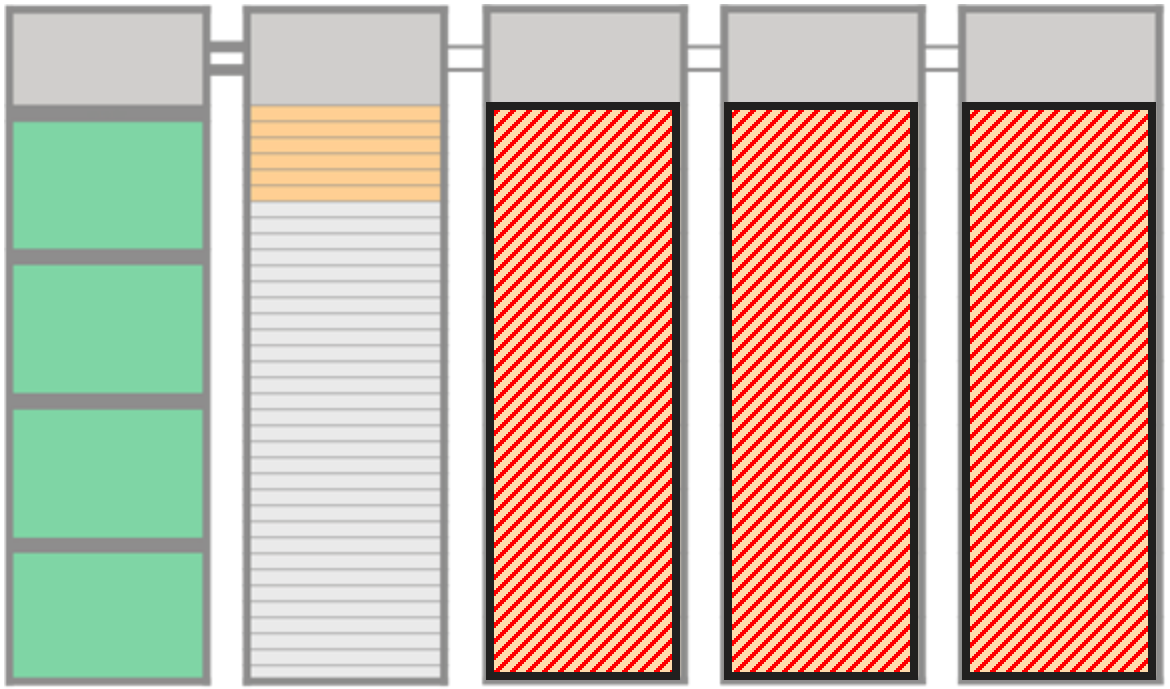
Tier performance and capacity estimates: NVIDIA.

Data Movement in an AI Cluster

GPU Server
All-Flash Tier
Object Storage Tier
Networking

Sequential Read
Random Read
Sequential Write
Random Write

1. Data Ingest



Raw data is written sequentially to the object storage tier.

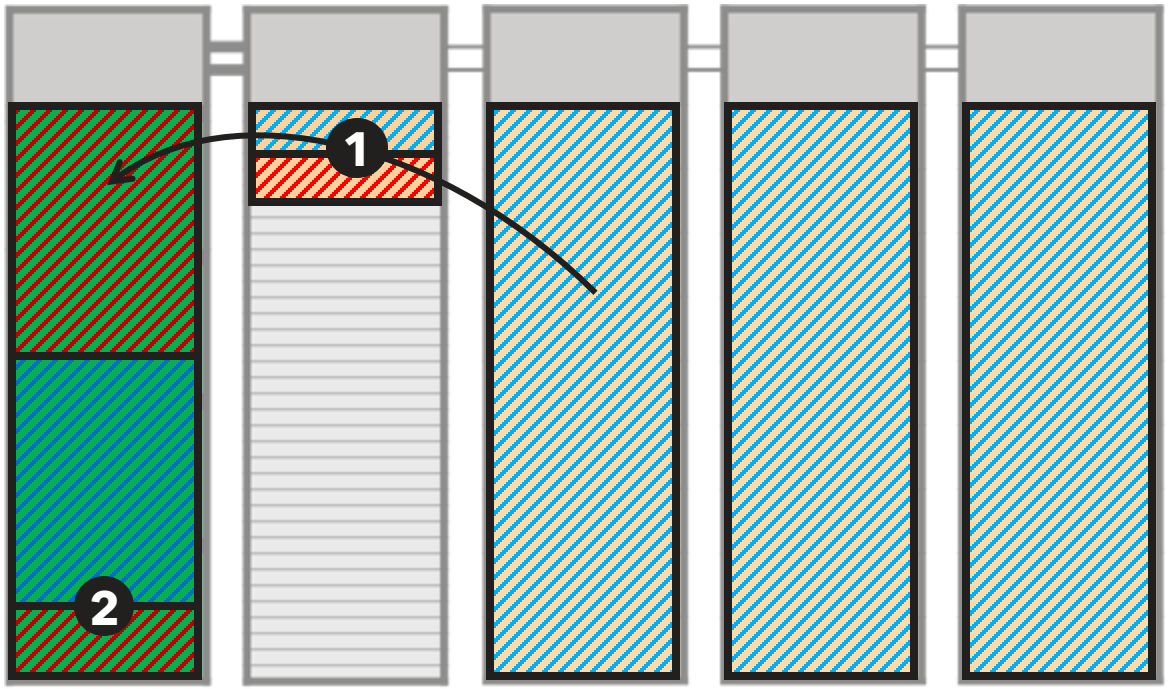
Data Ingest → Data Prep → Training → Checkpointing → Inference → Archive

Data Movement in an AI Cluster

GPU Server
All-Flash Tier
Object Storage Tier
Networking

Sequential Read | Random Read | Sequential Write | Random Write

2. Data Prep



1. Data is read from object storage and written to the compute servers.

2. CPUs pre-process the raw data, reading it and writing clean data afterward.

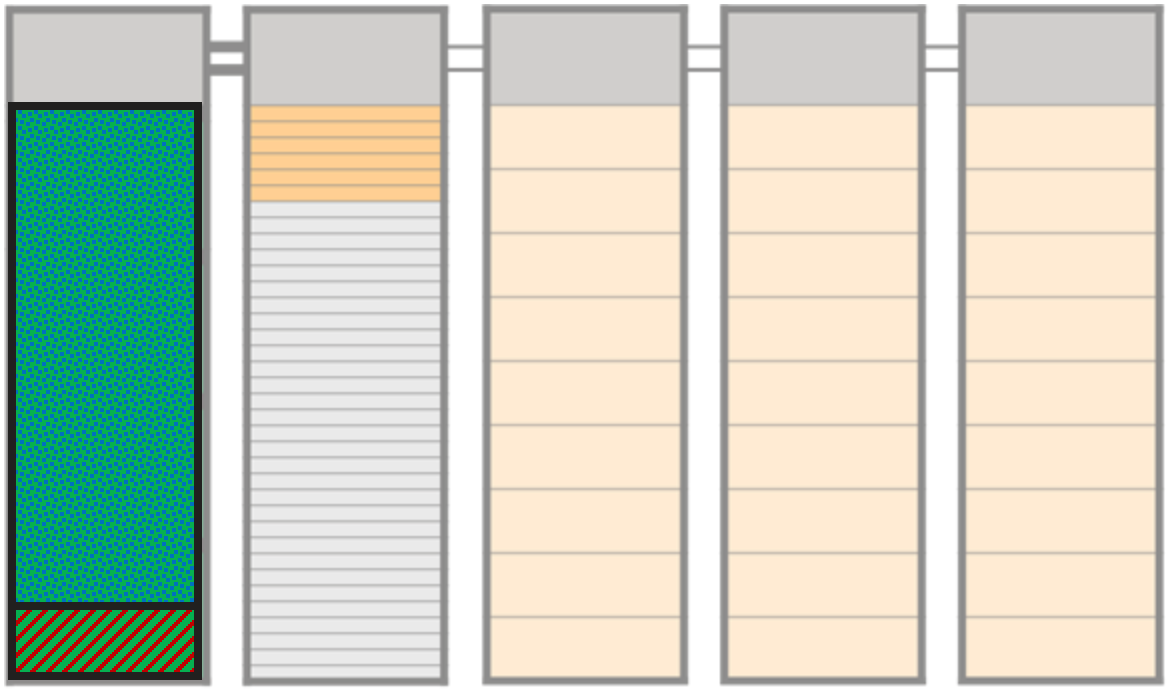
Data Ingest → **Data Prep** → Training → Checkpointing → Inference → Archive

Data Movement in an AI Cluster

GPU Server
All-Flash Tier
Object Storage Tier
Networking

Sequential Read | Random Read | Sequential Write | Random Write

3. Training



GPUs train the model by exposing data in random order.

The resulting trained model is written to disk.

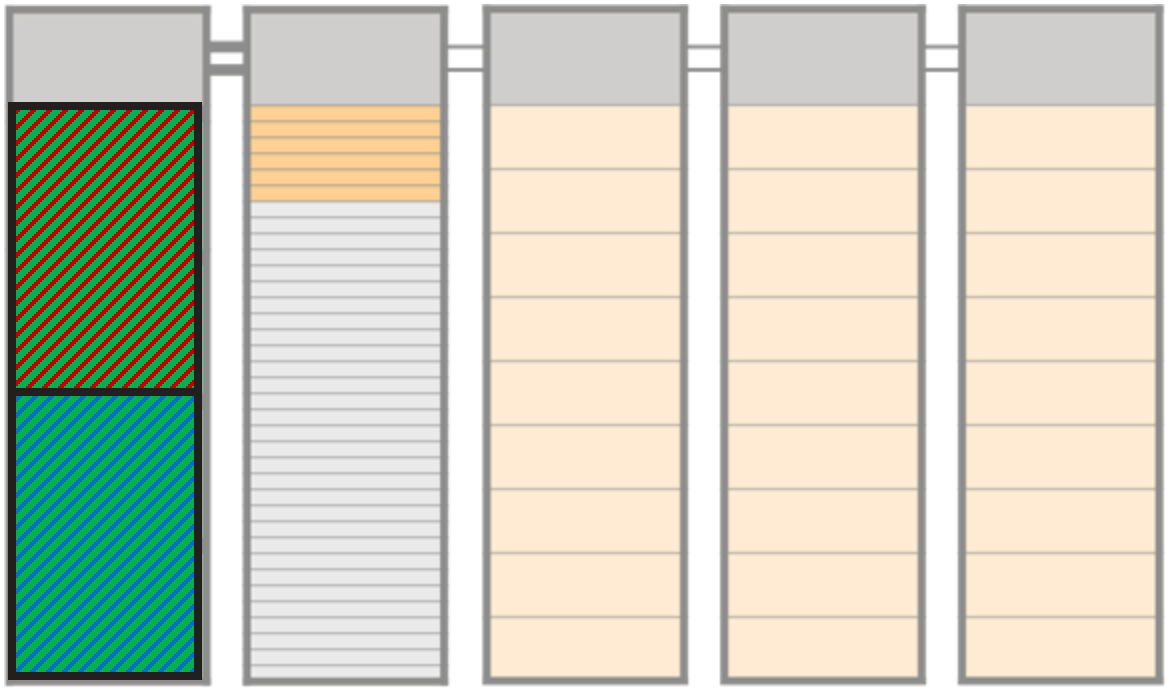
Data Ingest → Data Prep → **Training** → Checkpointing → Inference → Archive

Data Movement in an AI Cluster

GPU Server
All-Flash Tier
Object Storage Tier
Networking

Sequential Read
Random Read
Sequential Write
Random Write

3a. Checkpointing



The model-in-training is written periodically to disk and read back as needed.

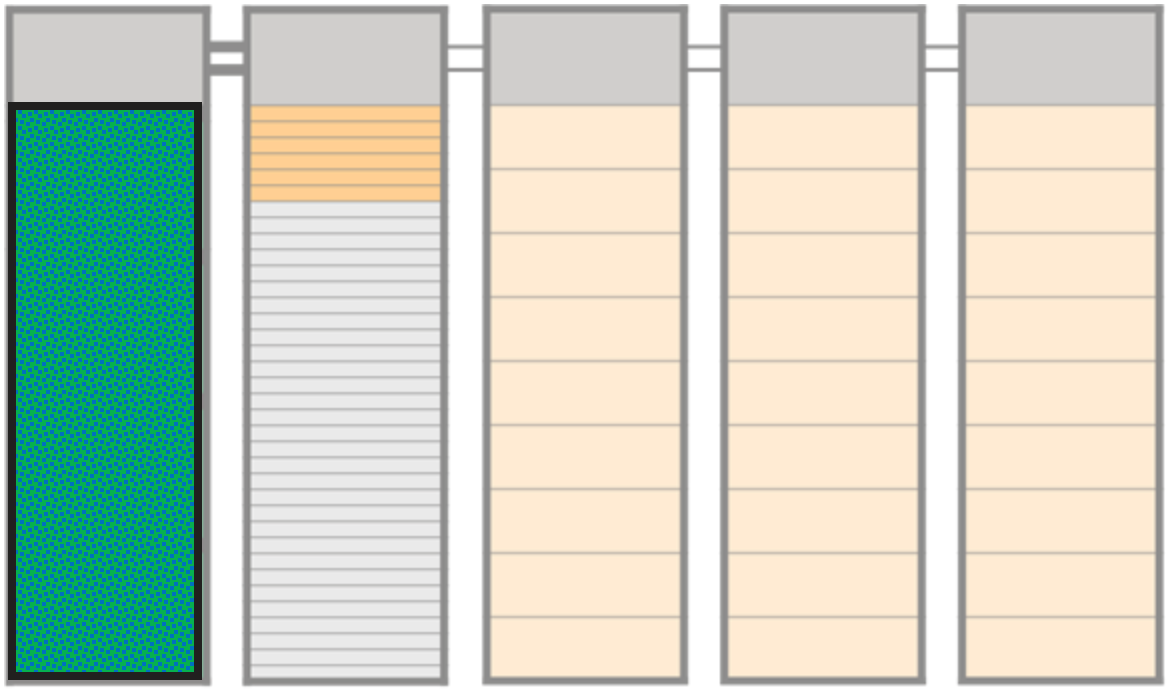
Data Ingest → Data Prep → Training → **Checkpointing** → Inference → Archive

Data Movement in an AI Cluster

GPU Server
All-Flash Tier
Object Storage Tier
Networking

Sequential Read | Random Read | Sequential Write | Random Write

4. Inference



The model is deployed and begins receiving inputs, generating random read activity in the GPU servers.

Optionally, RAG creates additional I/O activity.

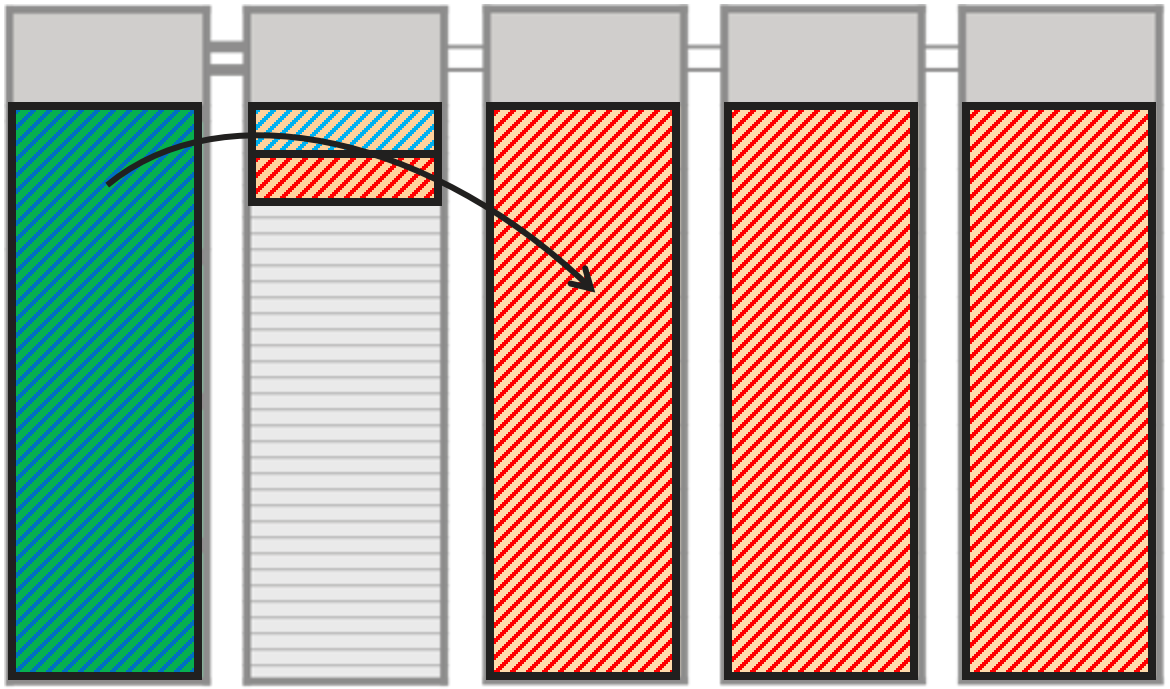
Data Ingest → Data Prep → Training → Checkpointing → **Inference** → Archive

Data Movement in an AI Cluster

GPU Server
All-Flash Tier
Object Storage Tier
Networking

Sequential Read
Random Read
Sequential Write
Random Write

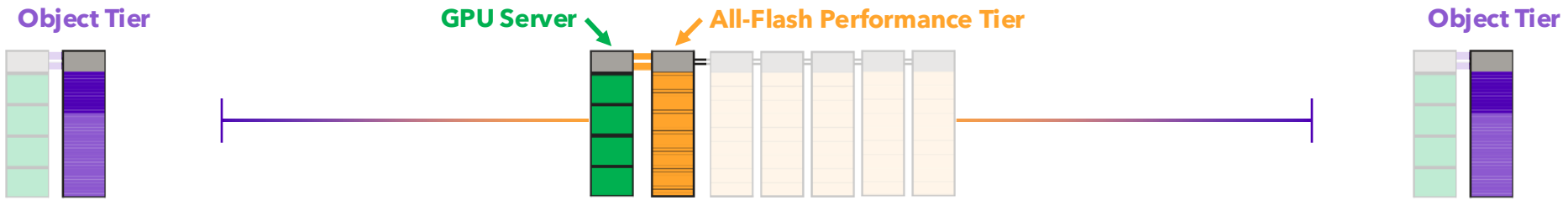
5. Archive



Model inputs and outputs are captured and written to disk in the object storage tier.

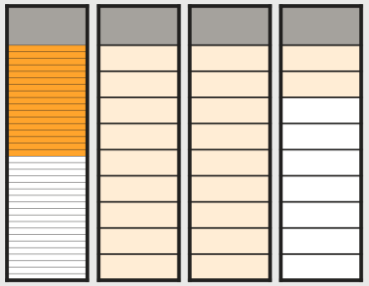
Data Ingest → Data Prep → Training → Checkpointing → Inference → **Archive**

A Portfolio Designed to Optimize AI Storage Efficiency



Stage	Data Ingest	Data Prep	Training	Checkpointing	Inference	Archive
Storage Requirements	High capacity and sequential write performance	Sequential read and write performance	Random read performance	Sequential write performance	Random read performance	High capacity
Recommended Solution	**Solidigm™ D5-P5336** PCIe 4.0 QLC SSD Capacity Read Write	**Maximize Performance / Watt** **Solidigm™ D7-P5520** PCIe 4.0 TLC SSD Capacity Read Write — or — **Solidigm™ D5-P5430** PCIe 4.0 QLC SSD Capacity Read Write				**Solidigm D5-P5336** PCIe 4.0 QLC SSD Capacity Read Write
	Maximize TB / Watt					

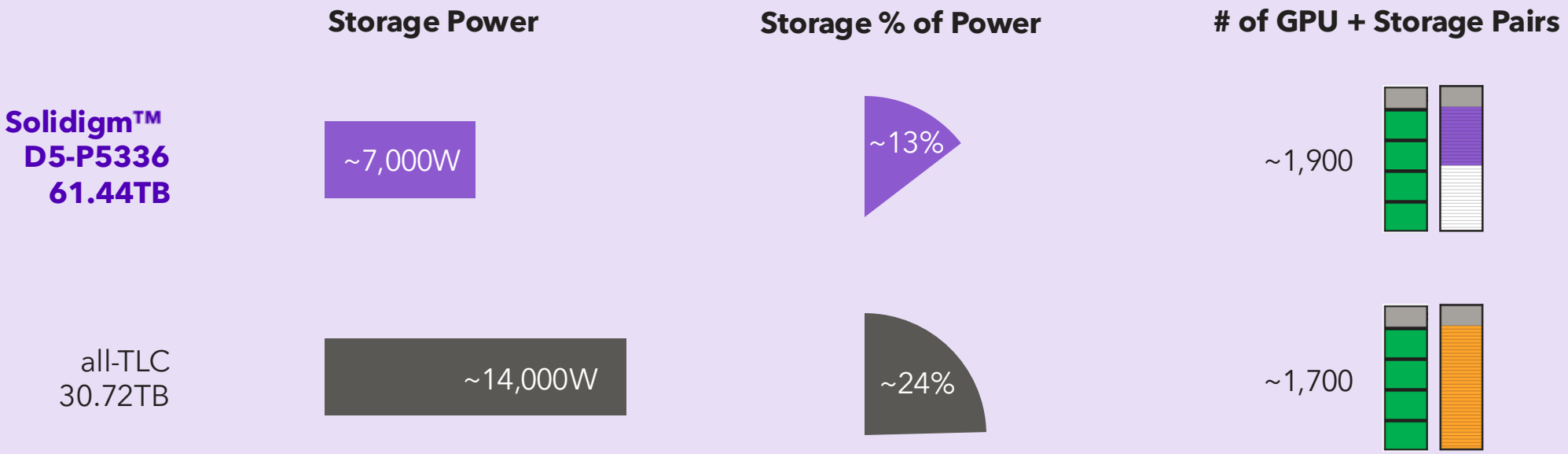
QLC improves power efficiency for new AI DC builds

	GPU Server	+	NAS: **TLC Performance Tier + HDD Object Tier**		or	NAS: **all-QLC**	**= the all-QLC advantage**
	DGX H100 Rack 4x GPU Servers		**TLC** 7.68TB	**all-HDD JBOF** 24TB		**Storage Server + JBOF** **Solidigm™ D5-P5336, 61.44TB**	
Solve for 100MW Total Available Power with 16PB per DGX Rack							**4X** fewer storage racks
Power Elements	~45,000W		~33,000W			**7,000W**	**~80%** less storage power
Storage as % of Total Power			~45%			**~13%**	
DGX + Storage Rack Pairs per DC			~1,300			**1,900**	**~50%** more GPU + storage pairs

Source – Solidigm, Aug 2024. Power consumption analysis assumes a green field (new) bottom-range Hyper-scaler / Tier 2 AI DC implementation utilizing leading-edge power and space optimizations. See Appendix 'QLC Power Efficiency vs HDD' for modeling details.

Higher density also improves **power efficiency vs. TLC**

Solving for 100MW Total Available Power with 16PB per DGX Rack



Source – Solidigm, Aug 2024. Power consumption analysis assumes a green field (new) bottom-range Hyper-scaler / Tier 2 AI DC implementation utilizing leading-edge power and space optimizations. See Appendix 'QLC Power Efficiency vs TLC' for modeling details.

Solidigm's Portfolio Optimized for the AI Era

Trusted by the Industry's Most Innovative Companies

Cheetah RAID STORAGE

"Today's primary edge constraint is bandwidth. Solidigm QLC SSDs offer an **impressive combination** of **capacity**, **performance**, and **reliability** to overcome this challenge. Using these SSDs, Cheetah's high-performance servers make them **highly suitable for efficient edge solutions**."

~Doug Emby, VP of Operations

ddn

"DDN is the world's **leading Data Intelligence platform** for AI and HPC. Many of the largest AI and SC installations globally rely on DDN, including **NVidia's largest SuperPOD**. Using **Solidigm 61.44 TB QLC SSDs**, DDN continues to deliver **industry-leading scale, power efficiency, performance and reliability** to some of the largest AI installations on the planet."

~James Coomer, Senior Vice President of Product

GIGABYTE™

"We have a wealth of enterprise servers and boards that are qualified for Solidigm's NVMe drives, including form factors E1.S and U.2. Because of the strong demand from our data center customers, we are able to support all **diverse storage workloads** with Solidigm's complete drive portfolio."

~Vincent Wang, Sales VP

PEAK AIO

"With Solidigm's technology, PEAK:AIO can achieve 2PB of storage per 2U, offering **exceptional power efficiency** with the winning combination of **SLC and Solidigm QLC SSDs**. This powerful partnership enables our customers to overcome AI's infrastructure challenges by providing **outstanding performance** in a compact, **energy-efficient design**, thus removing barriers to innovation."

~Roger Cummings, CEO, PEAK:AIO

(OCIENT)™

"There's no time like now for **fast, efficient, high-capacity storage**. Only Solidigm has a technology that helps check all the boxes. **Without Solidigm there would be no Ocient."**

~Shantan Kethireddy, VP of Customer Solutions

VAST

"VAST Data systems start at over 300 TB of Flash – they lean on **high density** Solidigm QLC SSDs for a variety of customers. Solidigm's QLC SSDs provide up to 61.44TB of storage which makes the design of the **system highly scalable to meet the needs of AI-era applications**, today and in the future.

~Kartik Subramanian, Global Systems Engineering Lead

xINNOR

"The combination of high-capacity QLC SSDs from Solidigm and the data integrity and performance assured by xiRAID is an **ideal solution for providing large, fast and reliable storage** to GPUs running AI models."

~Davide Villa, Chief Revenue Officer