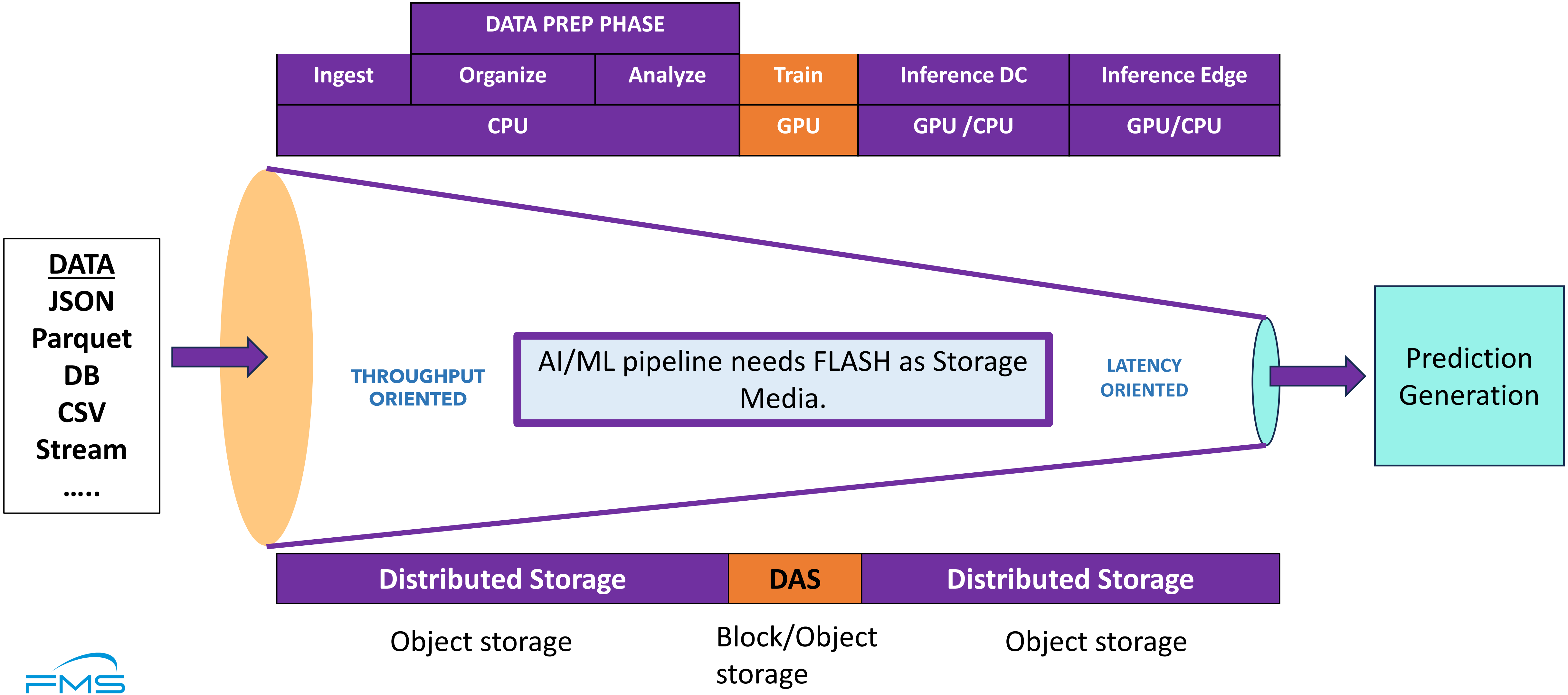
IO Characteristics of AI models and Workloads

Aug 7th 2025

Kiran Bhat, Product Manager, Solidigm

Alessandro Goncalves – Solution Architect, Solidigm

Data pipeline and Storage



	DATA PREP PHASE				
Ingest	Organize	Analyze	Train	Inference DC	Inference Edge
CPU			GPU	GPU /CPU	GPU/CPU

DATA
JSON
Parquet
DB
CSV
Stream
.....

THROUGHPUT ORIENTED

AI/ML pipeline needs FLASH as Storage Media.

LATENCY ORIENTED

Prediction Generation

Distributed Storage	DAS	Distributed Storage
Object storage	Block/Object storage	Object storage

Video Streamer - Pipeline

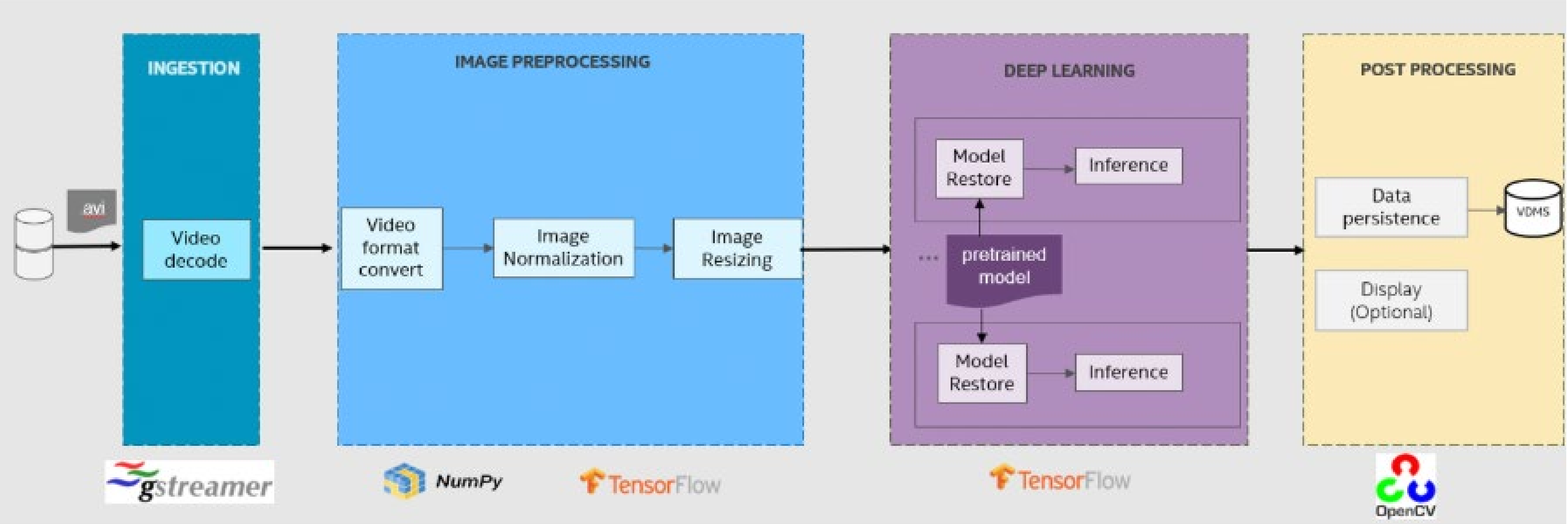


Fig: Video streamer Application Pipeline

Source: AI Pipeline Optimization on Xeon® Processors | Intel®

Face Recognition Pipeline

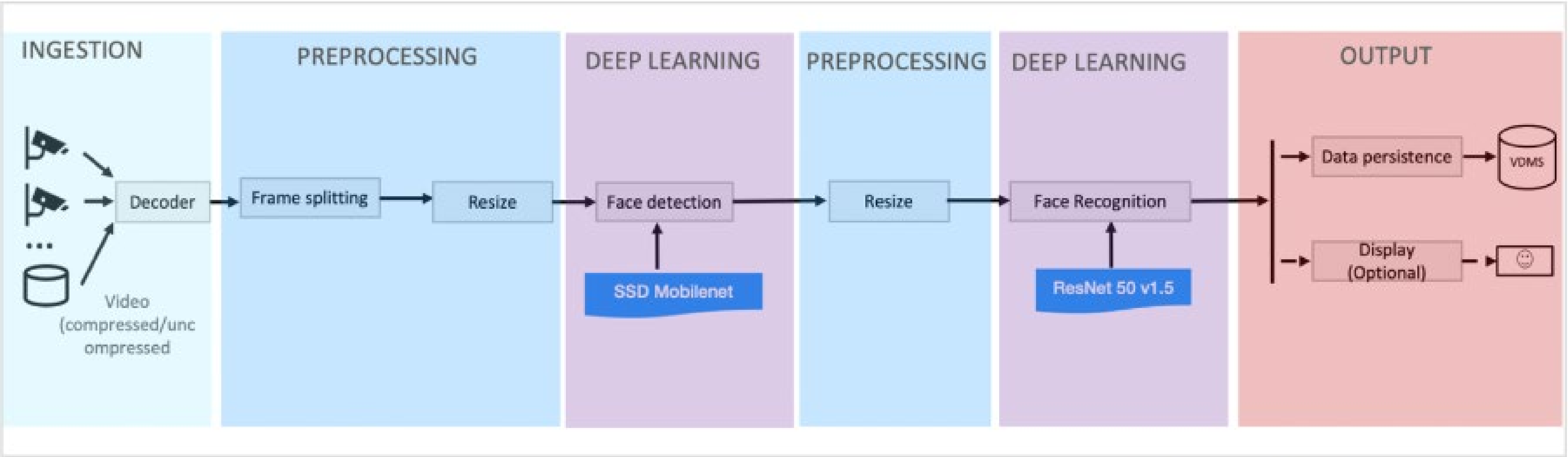


Fig: Face Recognition Application Pipeline

Source: AI Pipeline Optimization on Xeon® Processors | Intel®

Census - Pipeline

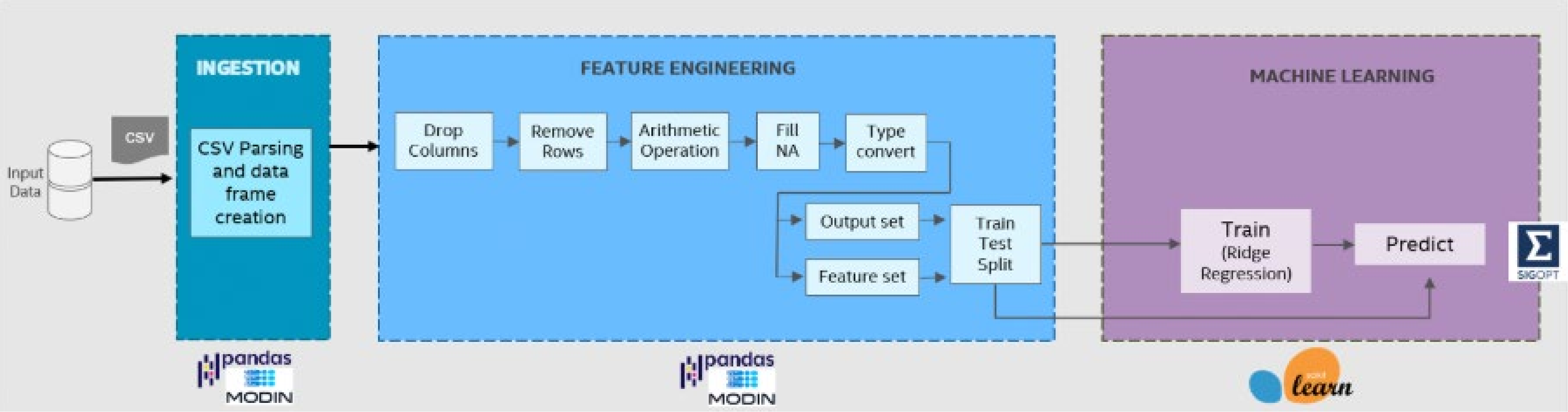
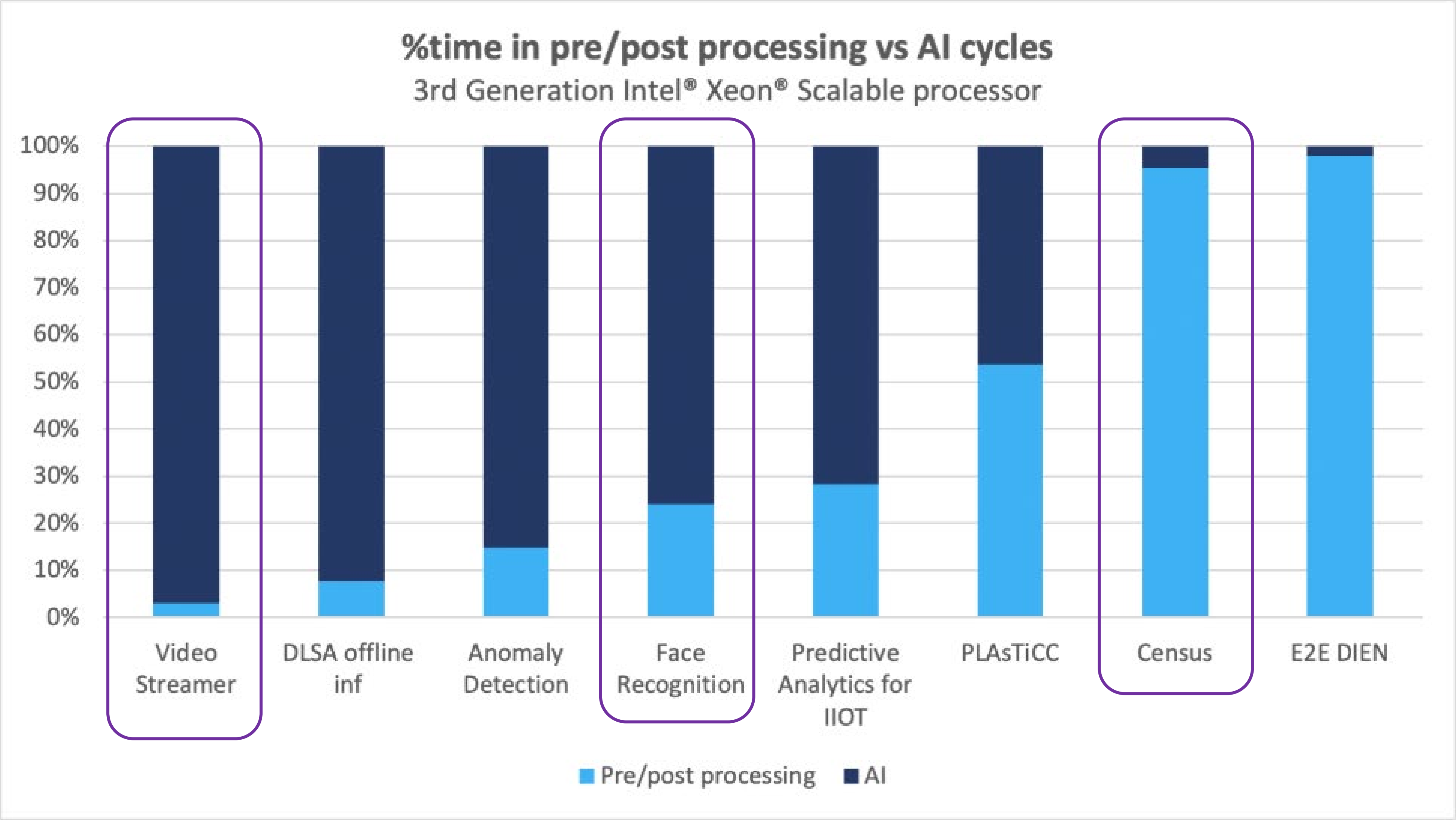


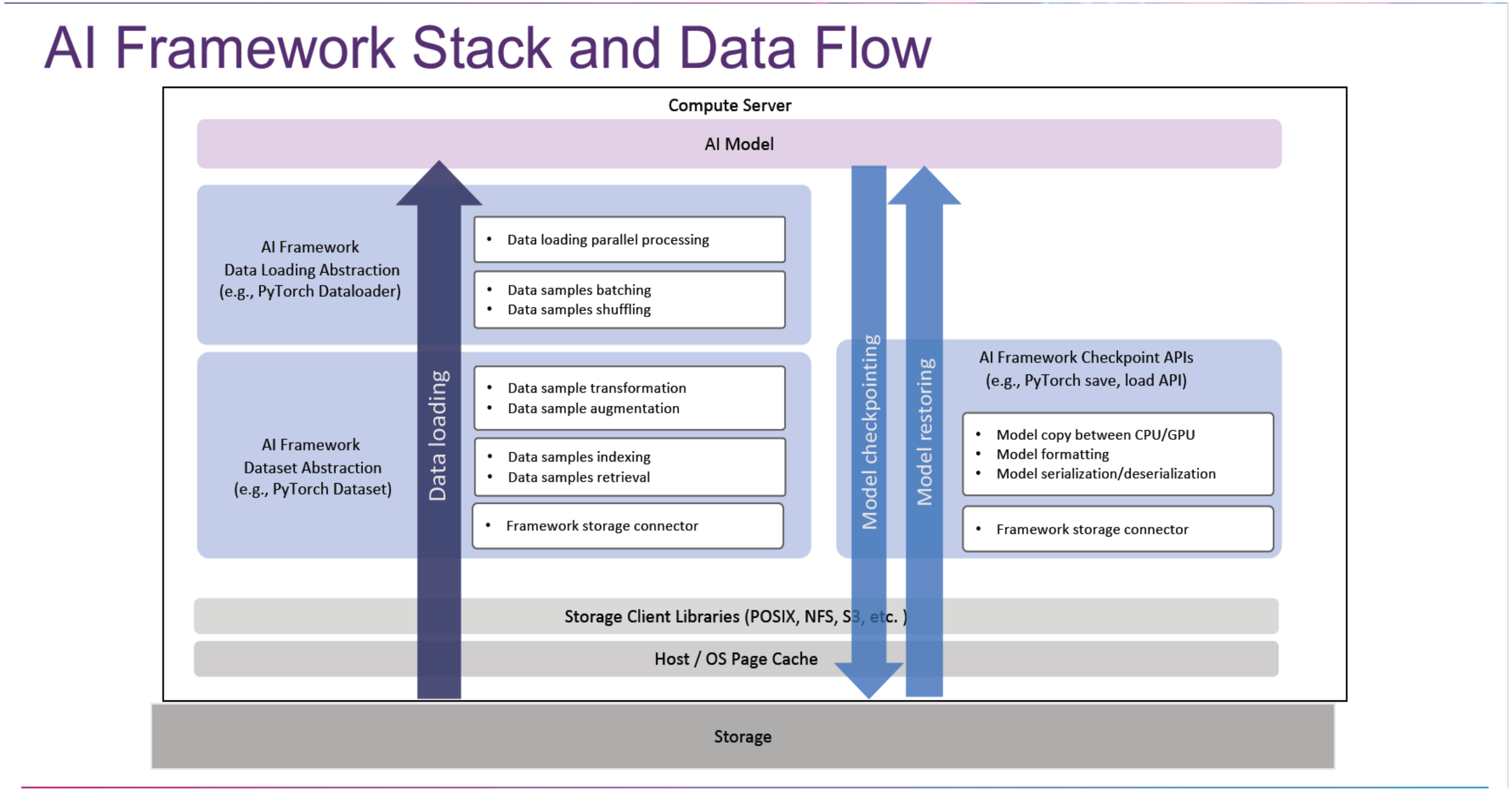
Fig: Census application pipeline

Source: AI Pipeline Optimization on Xeon® Processors | Intel®



Source: AI Pipeline Optimization on Xeon® Processors | Intel®

Complex Solution impacts Storage Performance



AI Framework Stack and Data Flow

Compute Server

AI Model

AI Framework
Data Loading Abstraction
(e.g., PyTorch Dataloader)

- Data loading parallel processing

- Data samples batching
- Data samples shuffling

AI Framework
Dataset Abstraction
(e.g., PyTorch Dataset)

- Data sample transformation
- Data sample augmentation

- Data samples indexing
- Data samples retrieval

- Framework storage connector

AI Framework Checkpoint APIs
(e.g., PyTorch save, load API)

- Model copy between CPU/GPU
- Model formatting
- Model serialization/deserialization

- Framework storage connector

Storage Client Libraries (POSIX, NFS, S3, etc.)

Host / OS Page Cache

Storage

Source: https://snia.org/sites/default/files/ESF/AI-Storage-The-Critical-Role-of-Storage-in-Optimizing-AI-Training-Workloads.pdf

It’s All About the Use Case

- AI/ML storage systems encompass a wide range of requirements and architectural complexities.
- Data Pipeline varies by each use case and model used
- Each use case has different overall architecture.
  - Software Stack – framework, application and library
  - Data set
  - Model
  - Training Parallelism

Ingest Phase

- Source data is different for each use case.
  - File / Block / Object Storage.
  - Parquet, CSV, JSON,image files, etc…
- Source data can be static, based in very slow tiers.
  - LLM , Image Search, etc…
  - Data is prepared Offline and put into the pipeline.
- Source also can be real time devices.
  - Facial / Defect Recognition

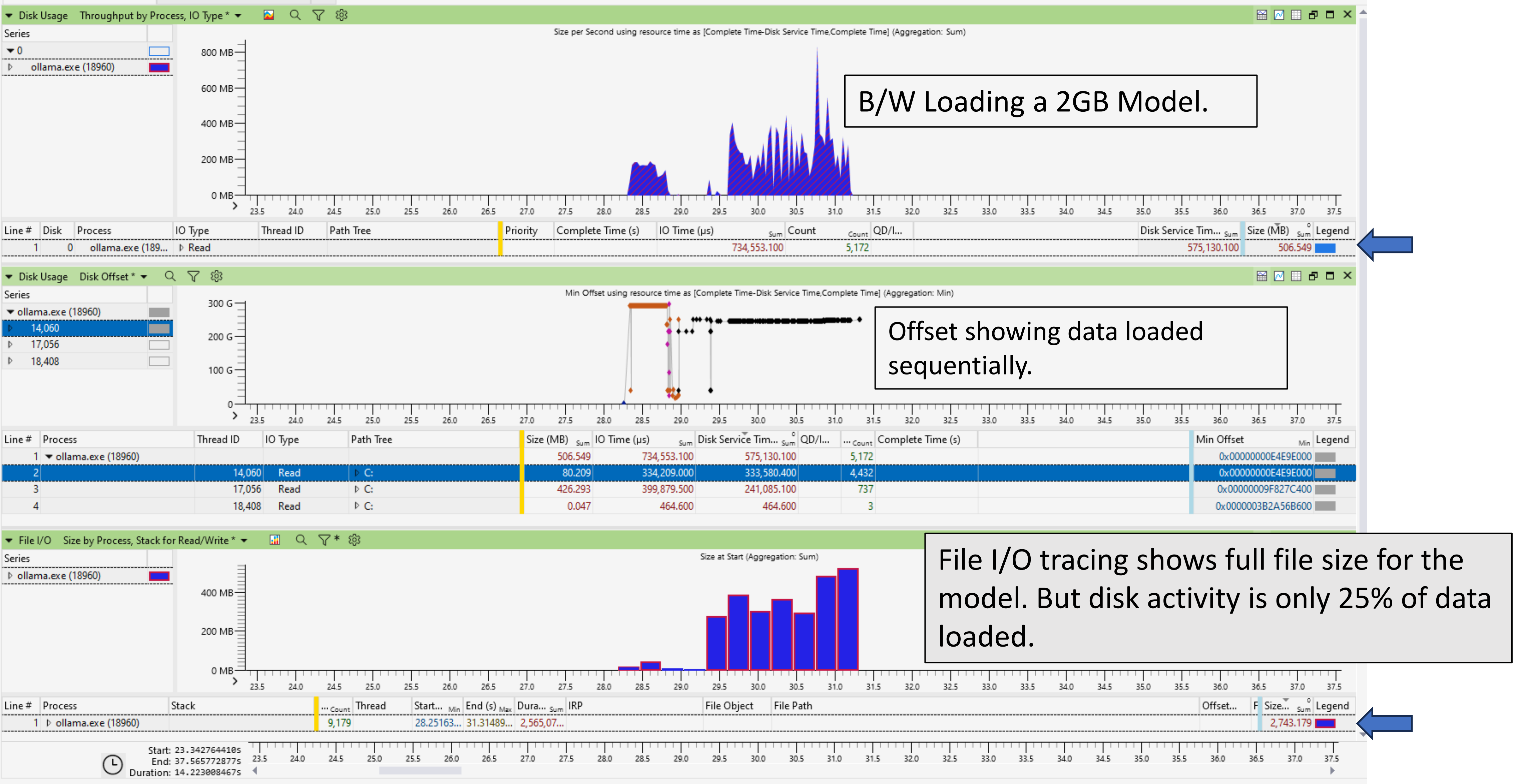
Train - Model Sharding / Parallelism

- Tensor
  - Splits individual weight tensors into multiple chunks on different devices.
- Pipeline
  - Partitions the model vertically into stage by layers. Different devices can process in parallel different stages of the full model.
- Context
  - Divides the input context into segments, reducing memory bottleneck for very long sequence length inputs.
- Data parallelism
  - Full model is inserted in each device HBM, and data is processed in parallel in multiple devices. Synchronization happens after each train step.

TRAIN - Checkpoint

- Checkpoint size is the model size plus model and Optimizer state.
- Model size is set by parameter and precision.
  - FP32 = 4Bytes
  - FP16 = 2Bytes
- Model State is added to checkpoint size.
  - Normally 2 states are saved.
  - Each state has the same number of parameters and normally with the same precision.
- Checkpointing saved as one or more files is based on the model parallelism and implementation
- The higher the GPU count utilized, the higher the checkpoint frequency.

Inference IO Profile – LLM –Llama 3



B/W Loading a 2GB Model.

Line #	Disk	Process	IO Type	Thread ID	Path Tree	Priority	Complete Time (s)	IO Time (µs) Sum	Count Count	QD/I...		Disk Service Tim... Sum	Size (MB) Sum	Legend
1	0	ollama.exe (189...	▷ Read					734,553.100	5,172			575,130.100	506.549	🟦

Offset showing data loaded sequentially.

Line #	Process	Thread ID	IO Type	Path Tree	Size (MB) Sum	IO Time (µs) Sum	Disk Service Tim... Sum	QD/I...	... Count	Complete Time (s)		Min Offset Min	Legend
1	▾ ollama.exe (18960)				506.549	734,553.100	575,130.100		5,172			0x00000000E4E9E000	▨
2		14,060	Read	▷ C:	80.209	334,209.000	333,580.400		4,432			0x00000000E4E9E000	▨
3		17,056	Read	▷ C:	426.293	399,879.500	241,085.100		737			0x00000009F827C400	▨
4		18,408	Read	▷ C:	0.047	464.600	464.600		3			0x0000003B2A56B600	▨

File I/O tracing shows full file size for the model. But disk activity is only 25% of data loaded.

Line #	Process	Stack	... Count	Thread	Start... Min	End (s) Max	Dura... Sum	IRP	File Object	File Path	Offset...	F	Size... Sum	Legend
1	▷ ollama.exe (18960)		9,179		28.25163...	31.31489...	2,565,07...						2,743.179	🟦

Start: 23.342764410s
End: 37.565772877s
Duration: 14.223008467s

the Future of Memory and Storage

Why Storage Matters in Inference

Mainstream

- Data Ingest and Archive
- RAG Database Scaling

Emerging

- KV Cache (Context) Offload
- Model Weight Offload

Network Storage

GB / W | **GB / in^3** | GB / s

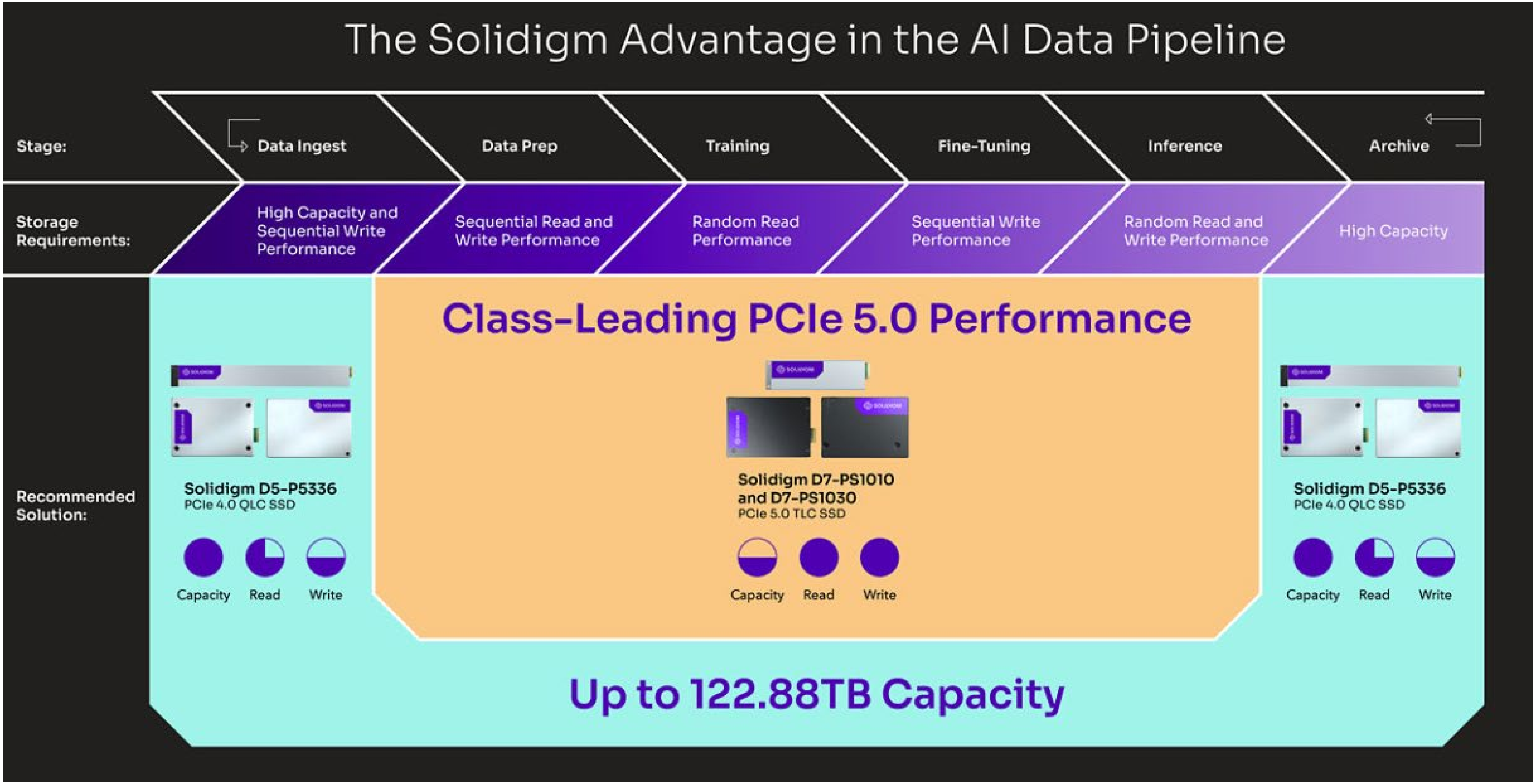
Direct Attach Storage

IOPS | IOPS / W | IOPS / $

Modern storage overcomes memory constraints to enable **larger models, longer interactions, and better outputs**

Summary

- AI/ML storage systems encompass a wide range of requirements and architectural complexities
- IO characteristics (IO size, Read/write ratio, Queue Depth, threads) depends on the model, framework, Data set and Training parallelism, type of storage
- There is no one size fits all; storage needs to be selected based on the usage case scenario

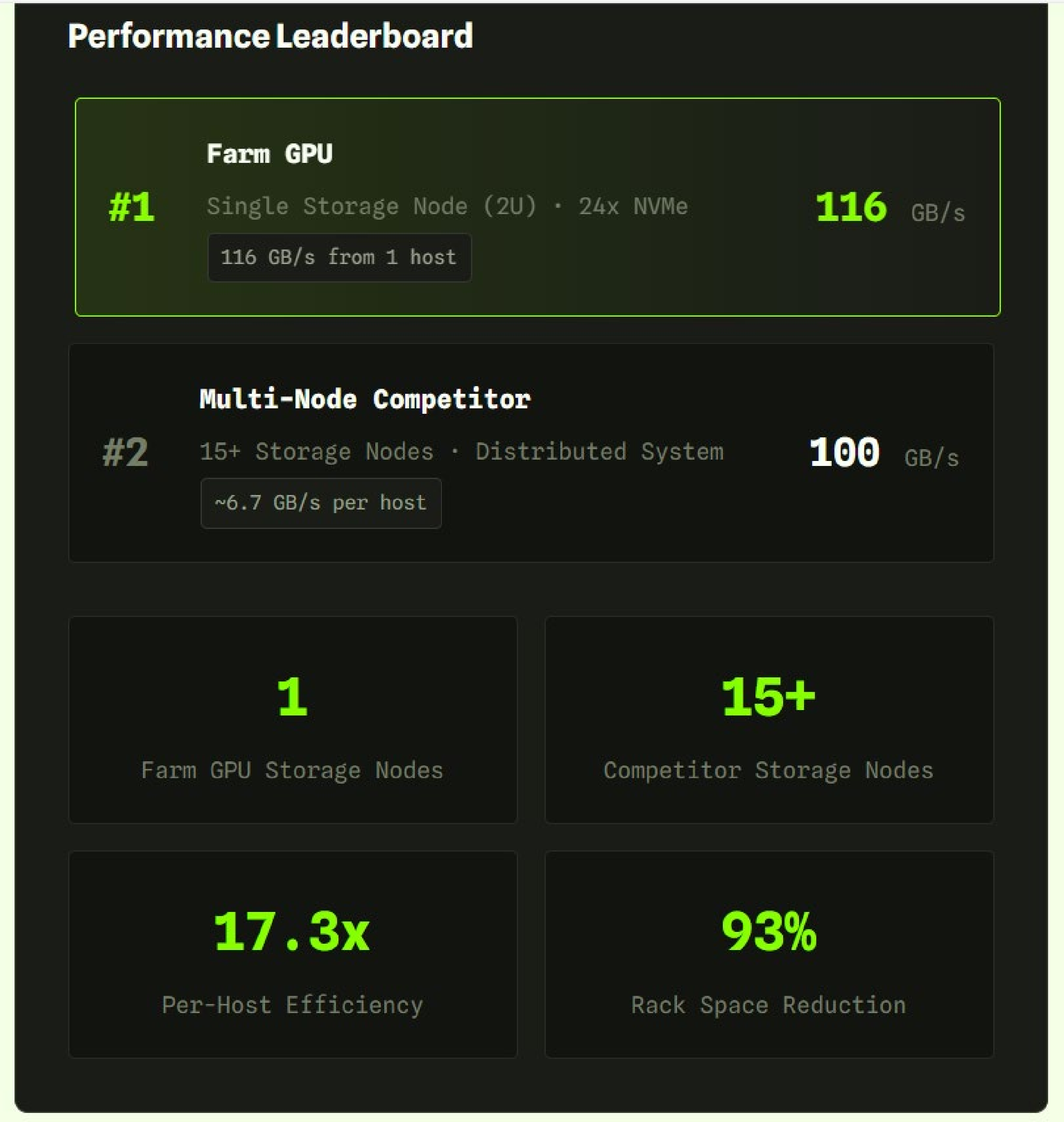


The Solidigm Advantage in the AI Data Pipeline

Stage:	Data Ingest	Data Prep	Training	Fine-Tuning	Inference	Archive
Storage Requirements:	High Capacity and Sequential Write Performance	Sequential Read and Write Performance	Random Read Performance	Sequential Write Performance	Random Read and Write Performance	High Capacity
Recommended Solution:	Solidigm D5-P5336 PCIe 4.0 QLC SSD ● Capacity ◕ Read ◒ Write	Solidigm D7-PS1010 and D7-PS1030 PCIe 5.0 TLC SSD ◒ Capacity ● Read ● Write				Solidigm D5-P5336 PCIe 4.0 QLC SSD ● Capacity ◕ Read ◒ Write

Class-Leading PCIe 5.0 Performance

Up to 122.88TB Capacity



Performance Leaderboard

#1 **Farm GPU**
Single Storage Node (2U) · 24x NVMe
116 GB/s from 1 host
116 GB/s

#2 **Multi-Node Competitor**
15+ Storage Nodes · Distributed System
~6.7 GB/s per host
100 GB/s

1 Farm GPU Storage Nodes	**15+** Competitor Storage Nodes
17.3x Per-Host Efficiency	**93%** Rack Space Reduction

Solidigm D7-PS1010 offers the highest throughput based on the latest MLPerf 2.0 measurement results submitted, for 3D-UNET model training carried out by farm GPU.

[Redefining AI Storage Economics: A Deep Dive into Single-Node Performance and System-Level Optimization]()

Backup

References

- AI Pipeline Optimization on Xeon® Processors | Intel®
- [2108.09373] Understanding Data Storage and Ingestion for Large-Scale Deep Recommendation Model Training