# KubeCon | CloudNativeCon

## North America 2024

KubeCon | CloudNativeCon
North America 2024

# AIStore as a Fast Tier Storage Solution

## Enhancing Petascale Deep Learning Across Remote Cloud Backends

Abhishek Gaikwad, Aaron Wilson, Alex Aizman

# Model Training Challenges

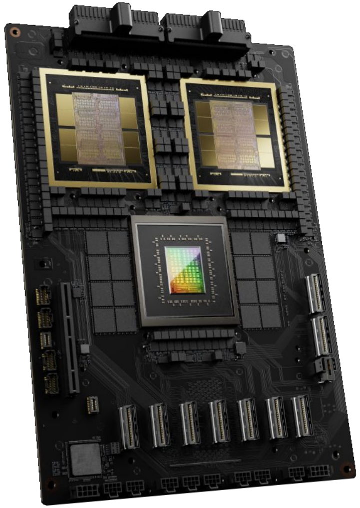


## Fast GPUs, Slow I/O

- GPU advancements are outpacing I/O capabilities

## Cloud Storage Limitations

- Petascale storage is expensive and slow
- Cross-region traffic incurs extra charges (and latency)
- Multiple epochs might require re-downloading data



*NVIDIA GB200 GPU*

**Multiple Backends -** AWS S3, GCP, Azure, on-prem

Training on petabytes of data is difficult

**Storage bottleneck** from random reads over multiple epochs

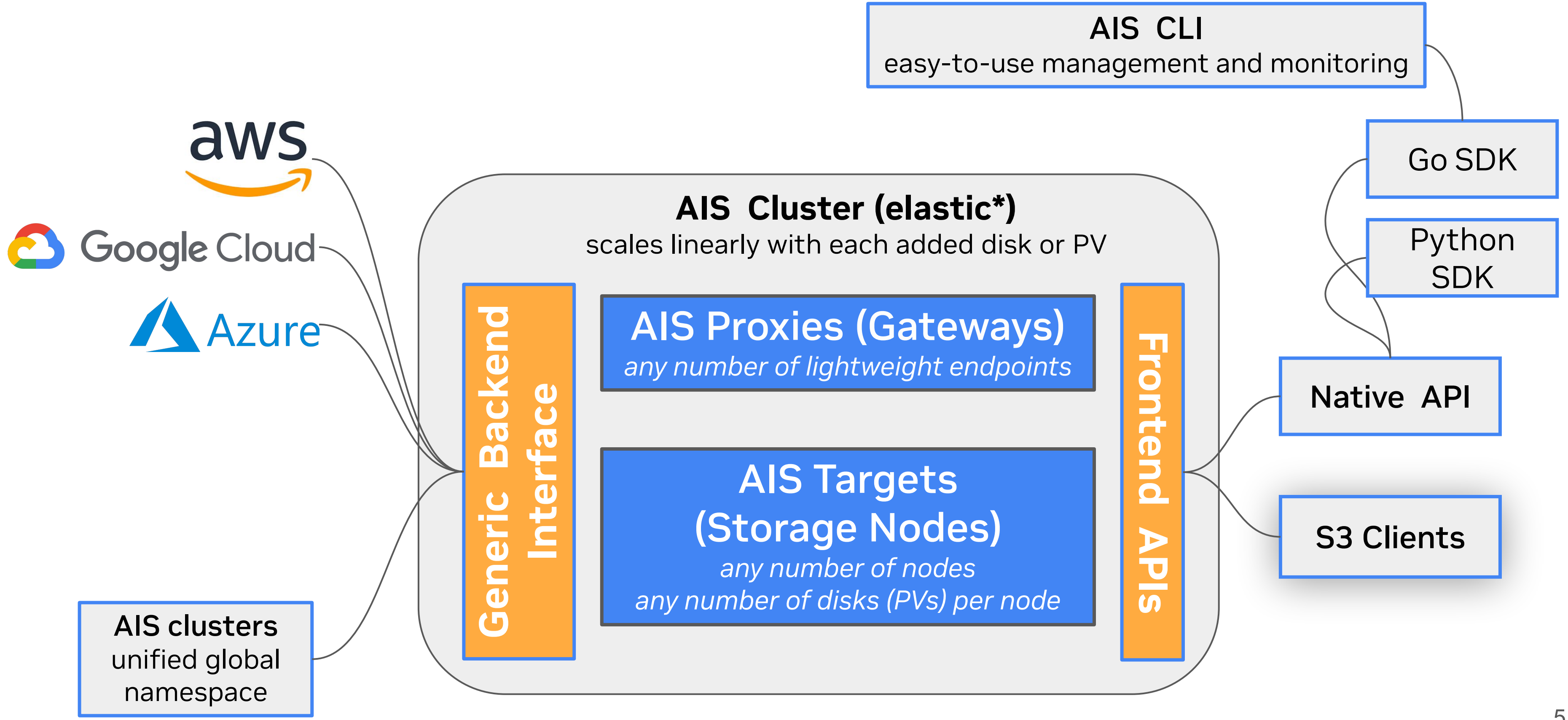
# AIStore (AIS): Scalable Object Storage for AI Applications



- An **open-source** (totally, from day one), lightweight, built-from-scratch object storage system tailored for AI and deep learning workloads
- Offers **linear scalability** with each added storage disk or node, ensuring balanced I/O distribution
- Features an **elastic cluster** architecture
- **Deployable anywhere**—from a single Linux machine to multi-petabyte Kubernetes clusters
- Over 7 years of development and **used in production** at NVIDIA

# AIStore Overview

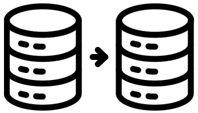




# AIStore Features

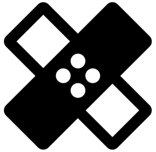


## High Availability & Data Protection

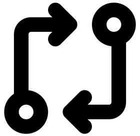


N-Way Mirroring

Erasure Coding



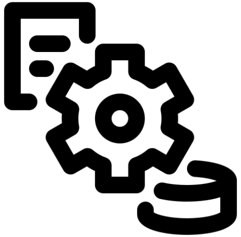
Self Healing



Lifecycle management

## ETL Offload

Run I/O intensive data transformations close to data (both offline and inline)



# AIStore Features

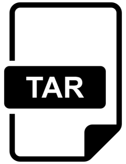
## Read-After-Write Consistency (*) and Write-Through Caching

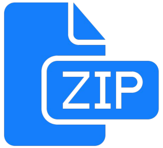
## Kubernetes Integration

- Easy deployment via AIS Operator


## Small File Datasets

- Sharding (original datasets)
- Resharding
- Appending
- Reading matching files without extracting



# AIStore Features

## Authentication and Access Control

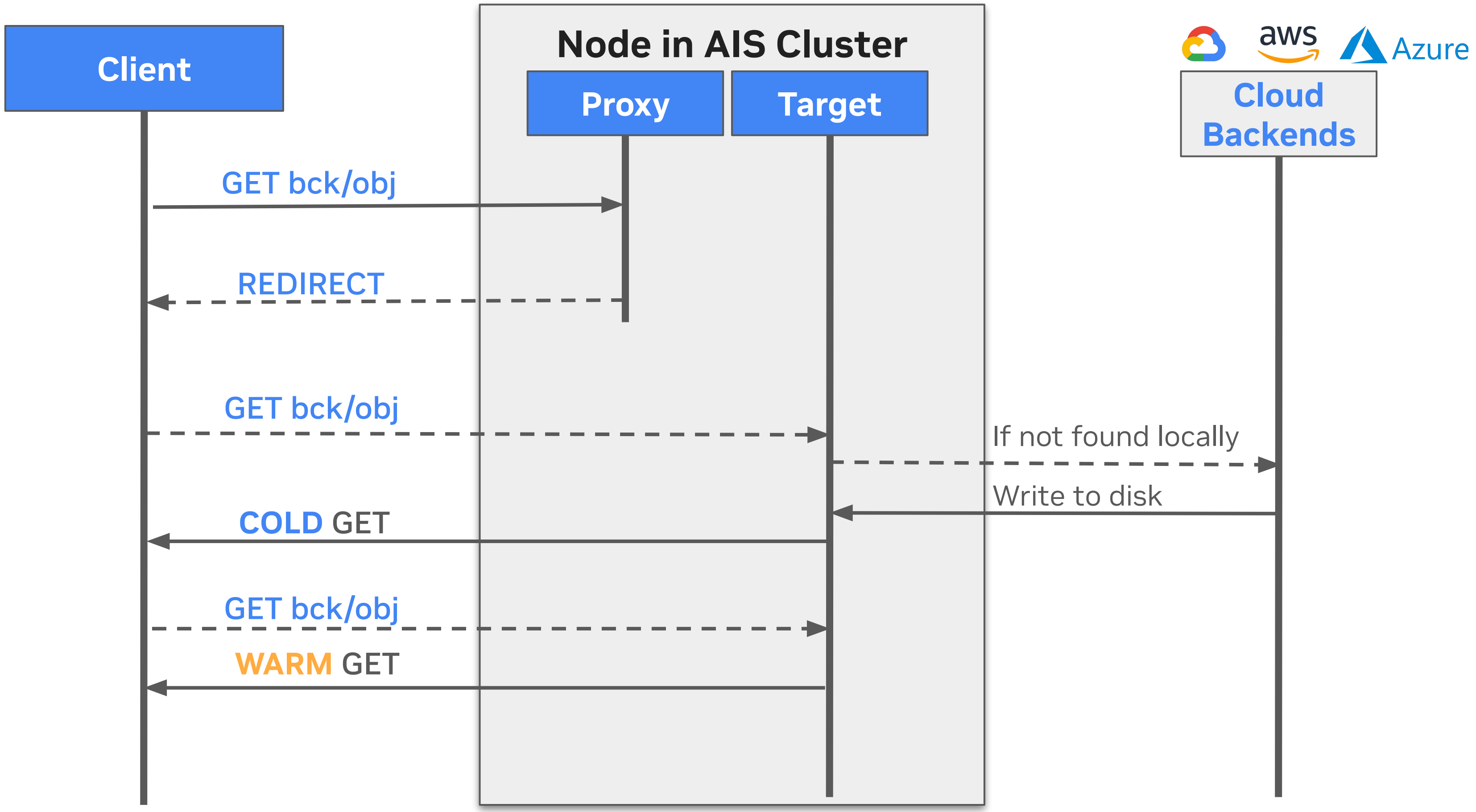
- OAuth 2.0 compliant Authentication Server (AuthN)

## Batch Jobs

- Prefetch
- Download
- Copy
- Transform

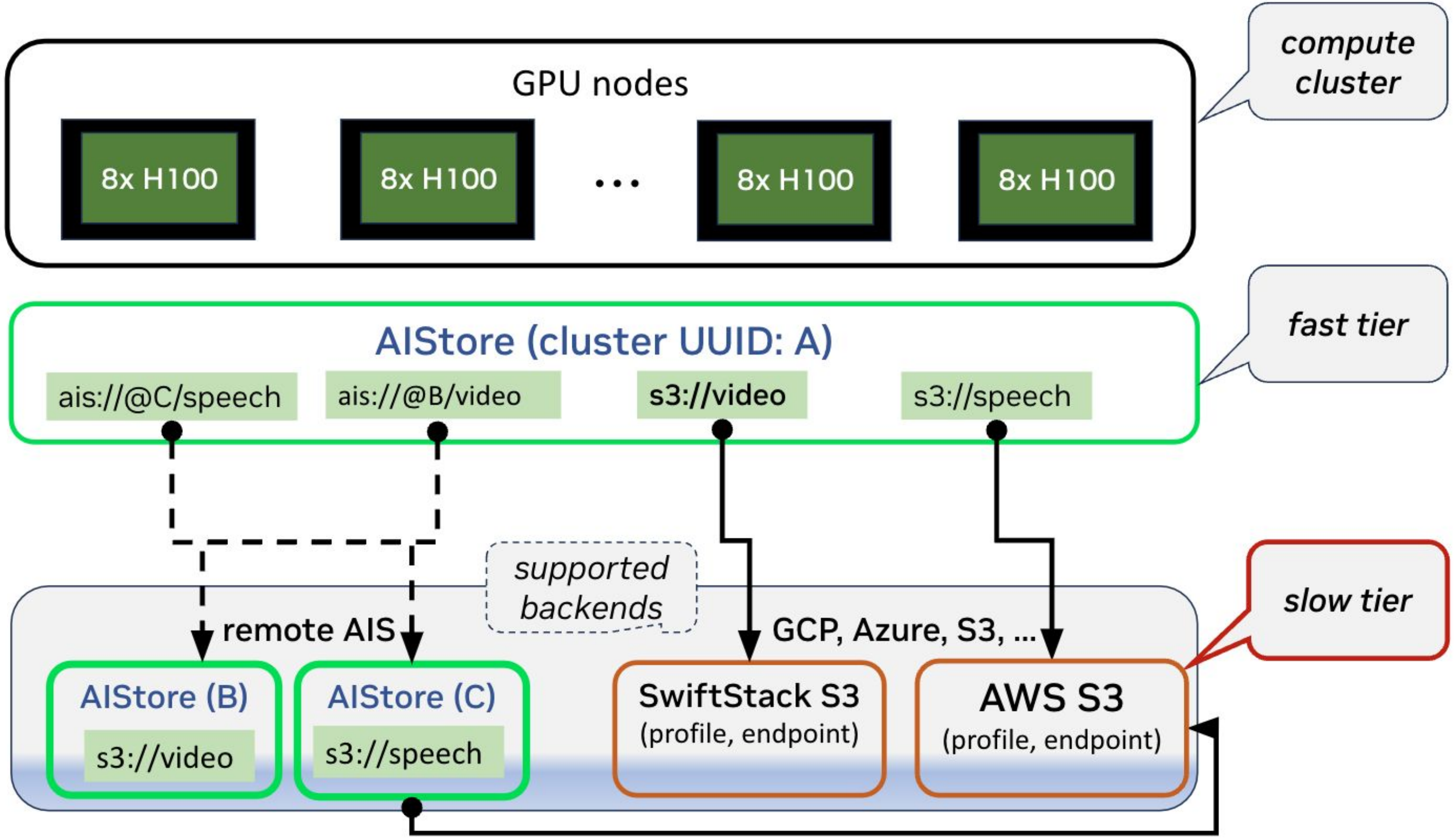
# AIStore Simplified Read Flow





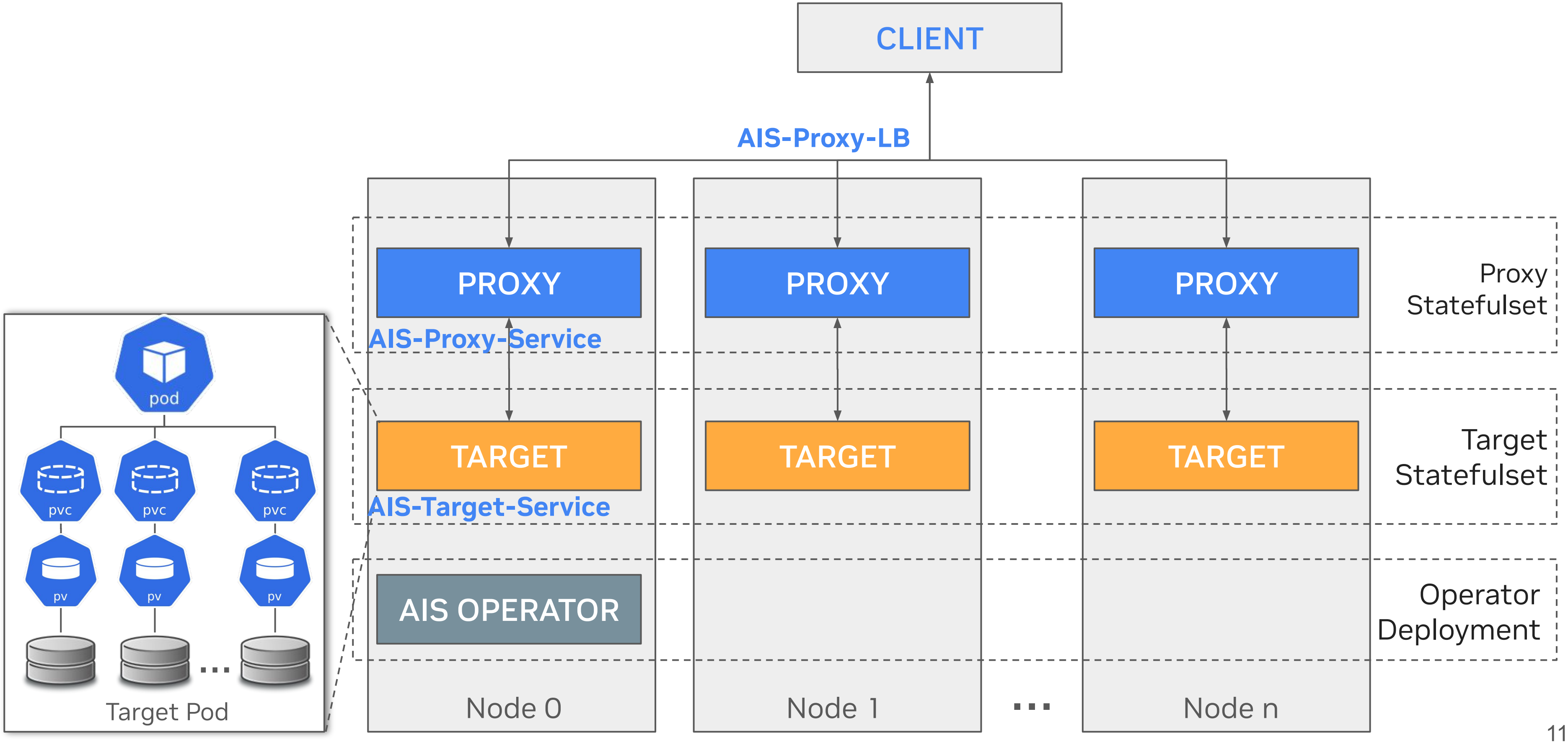
# AIStore Fast-Tier





# AIStore in Kubernetes





# AIStore Fast-Tier

## Synthetic Benchmarks: Measuring Performance for AI Workloads

# Benchmarks - Environment Overview



## AIStore Cluster Configuration – 16 node

- **Provider:** Oracle Cloud Infrastructure

  **Node Specifications**

  - **Instance Type:** BM.DenseIO.E5.128
  - **Memory:** 1536 GB
  - **CPU Cores (OCPU):** 128
  - **Storage:** 81.6 TB NVMe SSD (12 x 6.8 TB drives)
  - **Network Bandwidth:** 100 Gbps

# Benchmarks - Environment Overview



## Cluster Overview

- **K8s Node Scheduling:** 1 Target pod & 1 Proxy (gateway) pod per node
- **Node Count:** 16
- **Total Storage Capacity:** 1.16 PiB (1.31 PB), 192 drives

## Benchmark Setup

- **Benchmark Tool:** AIS Loader
  - Load generator for benchmarking AIStore and S3-compatible backends
- **Worker Configuration:**
  - **16** nodes running AIS Loader, same spec as AIS nodes
  - 1280 client threads, 80 per node

# Benchmarks - Data Retrieval Comparison



## Benchmark Types

- **Direct GET:**
  - Direct retrieval from S3, **without** AIStore
- **Cold GET:**
  - Initial retrieval from S3 **through** AIStore, persists objects
- **Warm GET:**
  - Subsequent retrieval of objects **with** AIStore, read from AIS disks

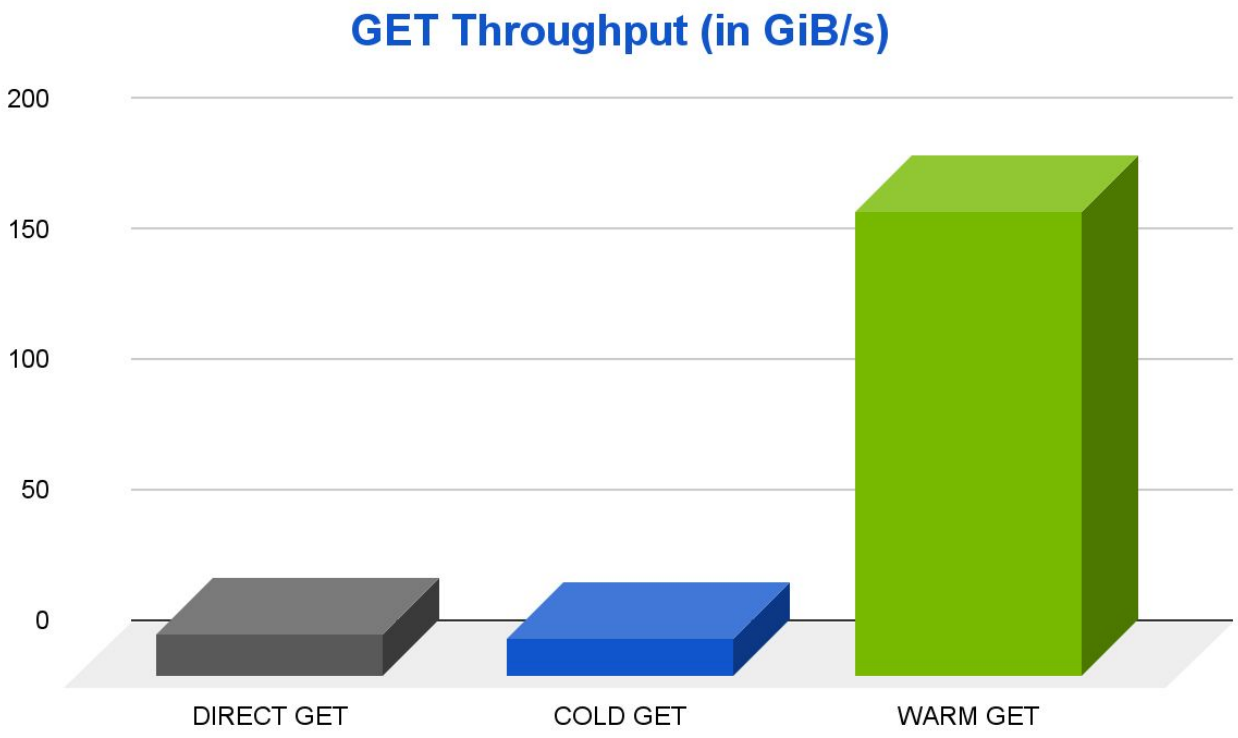
## Note

- Tests were conducted using a SwiftStack Object Store (S3-compatible), chosen over AWS S3 for superior connectivity (bandwidth)

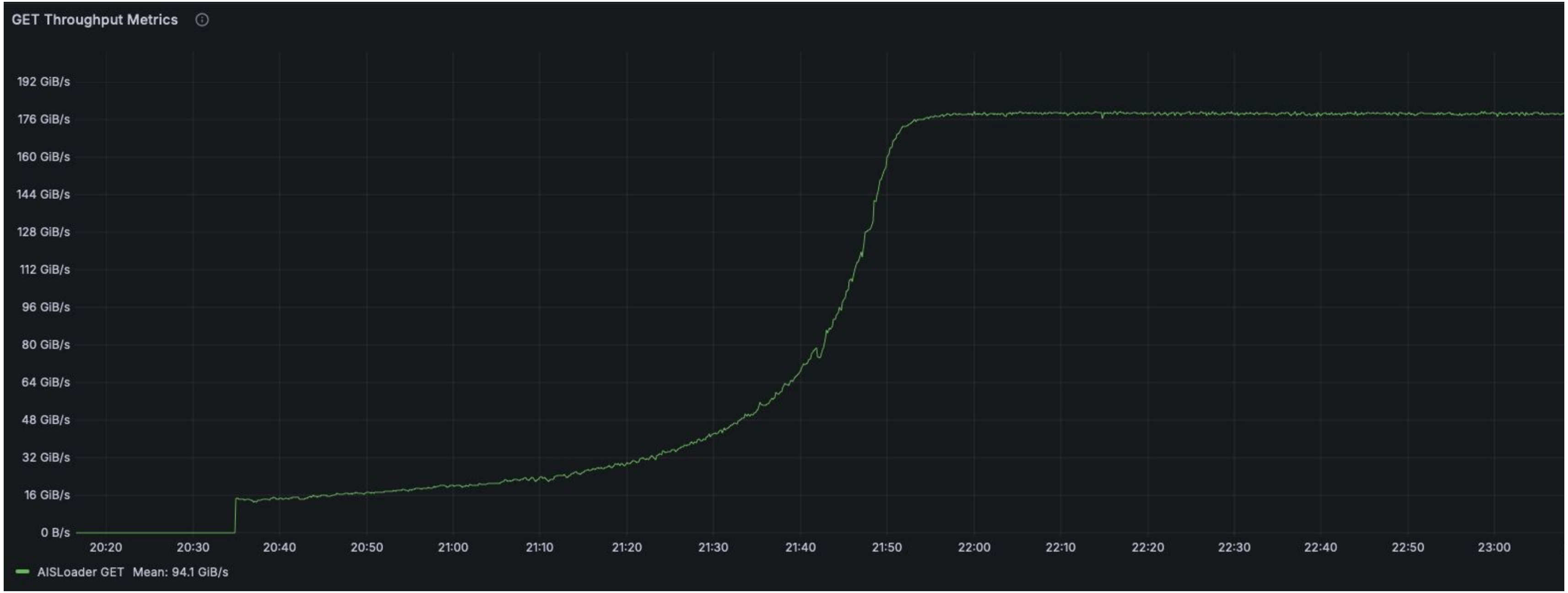
# Benchmarks - GET Throughput

KubeCon CloudNativeCon North America 2024

| Bench Type | Relative Performance |
|---|---|
| **DIRECT** GET | 1 |
| **COLD** GET | 0.86 |
| **WARM** GET | 10.7 |



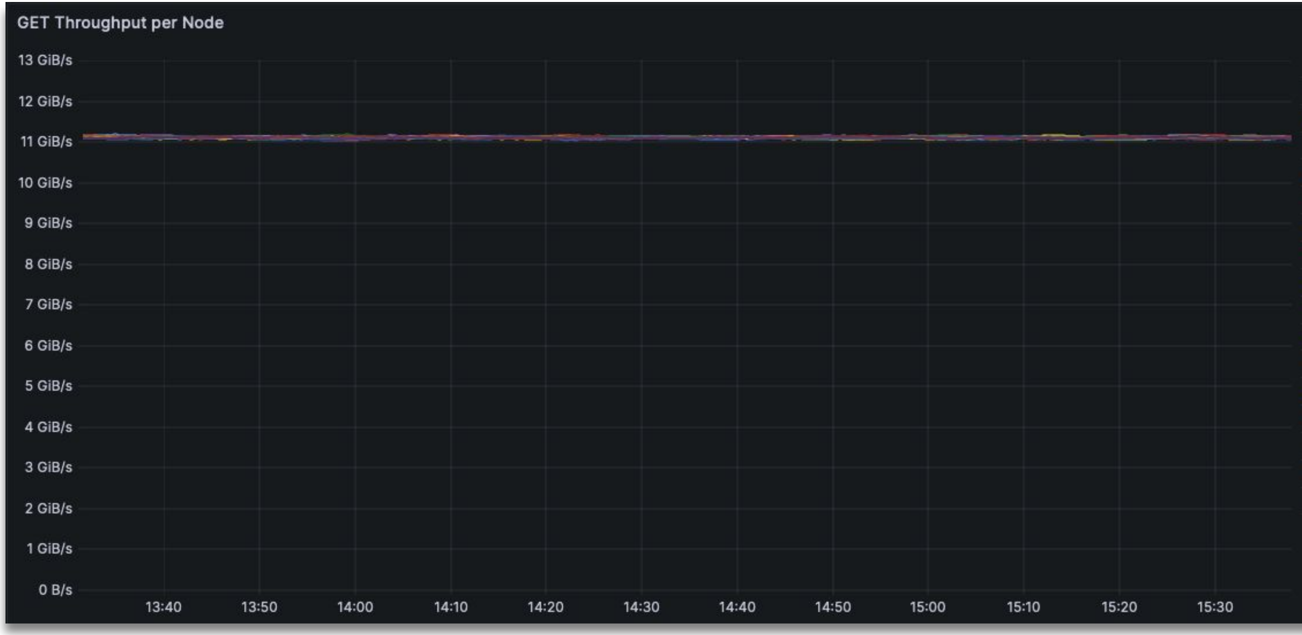
# Benchmarks - First Epoch Throughput




As clients request the same data, requests are satisfied locally

# Benchmarks - WARM GET Node Scaling






| Name | Mean | Last * | Max | Min |
|---|---|---|---|---|
| GaBDKWLN | 11.1 GiB/s | 11.1 GiB/s | 11.2 GiB/s | 11.1 GiB/s |
| IWWqtBSP | 11.1 GiB/s | 11.1 GiB/s | 11.1 GiB/s | 11.1 GiB/s |
| IsOwgWIB | 11.1 GiB/s | 11.1 GiB/s | 11.2 GiB/s | 11.1 GiB/s |
| MZazDpym | 11.1 GiB/s | 11.1 GiB/s | 11.2 GiB/s | 11.1 GiB/s |
| MmHXpmpl | 11.1 GiB/s | 11.2 GiB/s | 11.2 GiB/s | 11.1 GiB/s |
| PCbfqVTC | 11.1 GiB/s | 11.1 GiB/s | 11.2 GiB/s | 11.1 GiB/s |
| PqWxfxgX | 11.1 GiB/s | 11.1 GiB/s | 11.2 GiB/s | 11.1 GiB/s |
| PsKxnkJc | 11.1 GiB/s | 11.1 GiB/s | 11.1 GiB/s | 11.0 GiB/s |
| QEFVhUwY | 11.1 GiB/s | 11.1 GiB/s | 11.2 GiB/s | 11.1 GiB/s |
| WoDEYCOS | 11.1 GiB/s | 11.1 GiB/s | 11.2 GiB/s | 11.1 GiB/s |
| nuXRwSwO | 11.1 GiB/s | 11.1 GiB/s | 11.1 GiB/s | 11.1 GiB/s |
| oUEJODRF | 11.1 GiB/s | 11.1 GiB/s | 11.2 GiB/s | 11.1 GiB/s |
| ufCpyzBW | 11.1 GiB/s | 11.1 GiB/s | 11.2 GiB/s | 11.1 GiB/s |
| uxdVyeCH | 11.1 GiB/s | 11.1 GiB/s | 11.2 GiB/s | 11.0 GiB/s |
| xDmJgdJV | 11.1 GiB/s | 11.1 GiB/s | 11.2 GiB/s | 11.1 GiB/s |
| xlwjtPal | 11.1 GiB/s | 11.1 GiB/s | 11.2 GiB/s | 11.1 GiB/s |

Equal distribution of load over 16 nodes

# Benchmarks - WARM GET Disk Scaling




| Name | Mean | Last * | Max | Min |
|---|---|---|---|---|
| nvme0n1 | 95.0% | 96% | 98% | 91% |
| nvme10n1 | 95.3% | 95% | 98% | 92% |
| nvme11n1 | 95.3% | 96% | 98% | 91% |
| nvme1n1 | 95.2% | 95% | 98% | 92% |
| nvme2n1 | 95.1% | 93% | 98% | 92% |
| nvme3n1 | 95.4% | 95% | 98% | 92% |
| nvme4n1 | 95.1% | 96% | 98% | 92% |
| nvme5n1 | 95.3% | 96% | 98% | 92% |
| nvme6n1 | 95.1% | 95% | 98% | 92% |
| nvme7n1 | 95.3% | 94% | 98% | 93% |
| nvme8n1 | 95.0% | 95% | 98% | 91% |
| nvme9n1 | 95.1% | 95% | 98% | 92% |

Equal utilization of all 12 drives on any given target

# Benchmarks - Network Utilization



- **Network Utilization: >95%**
  - 1529 Gb/s (178 GiB/s) data transfer
  - 1549 Gb/s including HTTPS and auth overhead (<2%)
  - Theoretical advertised physical limit of 1600 Gb/s


Total network data sent: 1549 Gb/s

# Benchmarks - Disk Performance



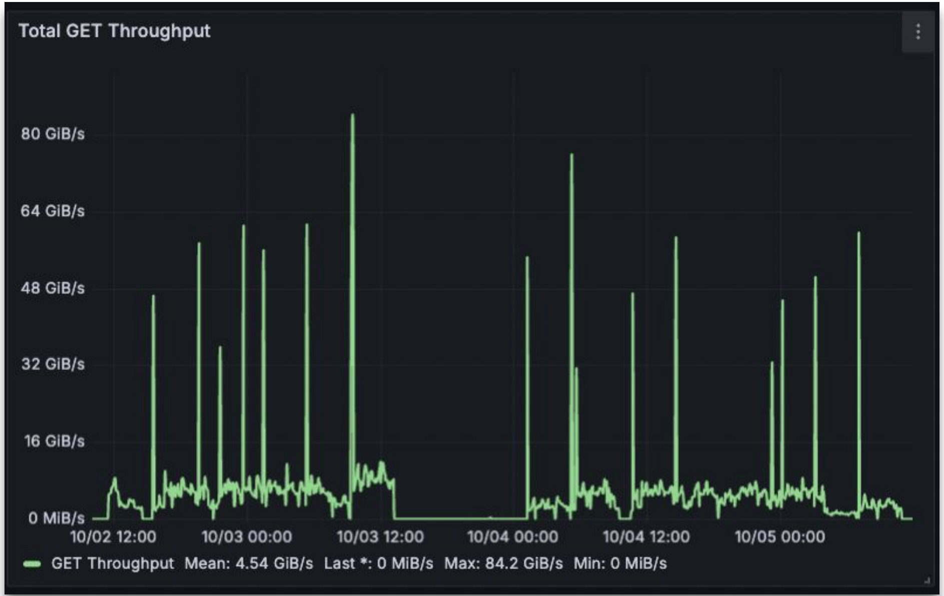
- **FIO Benchmark:**
  - Expected performance of ~3.3 GiB/s per drive for 10 MiB objects
  - 192 drive cluster – theoretical **633 GIB/s** >178 GiB/s observed
- **Disk Read Reduction:**
  - 1.5 TB of memory per node, 35% reads from page cache, reducing disk I/O
- **Scaling**
  - These drives **saturate the network**! (for this workload)
  - Options: scale up with nodes, optimize node specs


Actual disk read: ~35% below total throughput, 26% of theoretical bandwidth

# Production Training Workload on Speech Data

KubeCon CloudNativeCon North America 2024

- 16 worker nodes
  - Each node:
    - 8 NVIDIA H100 GPUs
    - 64 dataloader workers
  - **1024 parallel PyTorch DataLoaders**
- Peak Performance
  - ~12,000 I/O requests per second
  - Data throughput: 84.2 GiB/s
  - 84.2 MiB/s per client thread
- Similar clients to bench, so why lower?
  - 30s min stats interval, due to prometheus scraping
  - Not granular enough to capture spike
  - Busy training after data pull (good!)




2 jobs run over the course of 3 days

# Deployment Demo



# Future Work



- ETL Optimization
  - Improve usability, performance, scaling
- AuthN Extended Features
  - Highly-available deployment
  - Interoperability with other IAM
- Cloud credential management (single-tenant design)
- Experiment with cluster variations
  - Larger
  - Hyper-converged
  - Multi-tier (AIS with AIS backend)

# Conclusion



- AIStore
  - Easy to deploy and get started
  - Long list of features
  - Unlock **scalable performance**
- Come talk to us at the Nvidia booth!
- Resources:
  - github.com/NVIDIA/aistore
  - github.com/NVIDIA/ais-k8s
  - aistore.nvidia.com/

KubeCon | CloudNativeCon

North America 2024

# Questions?

Email us at
aistore@nvidia.com

Leave Feedback!

# NVIDIA & Community Talks

KubeCon CloudNativeCon North America 2024

A Tale of 2 Drivers: GPU Configuration on the Fly Using DRA | **Wednesday** 3:25pm - 4:00pm

Tutorial: Get the Most Out of Your GPUs on Kubernetes with the GPU Operator | **Wednesday** 4:30pm - 6:00pm

From Silicon to Service: Ensuring Confidentiality in Serverless GPU Cloud Functions | **Thursday** 11:00am - 11:35am

Unlocking Potential of Large Models in Production | **Thursday** 2:30pm - 3:05pm

Which GPU Sharing Strategy Is Right for You? a Comprehensive Benchmark Study Using DRA | **Thursday** 4:30pm - 5:05pm

Engaging the KServe Community, The Impact of Integrating a Solutions with Standardized CNCF Projects | **Thursday** 5:25pm - 6:00pm

Maintainer Track: WG Serving: Accelerating AI/ML Inference Workloads on Kubernetes | **Friday** 11:55am - 12:30pm

From Vectors to Pods: Integrating AI with Cloud Native | **Friday** 2:00pm - 2:35pm

Enabling Fault Tolerance for GPU Accelerated AI Workloads in Kubernetes | **Friday** 2:55pm - 3:30pm

Thousands of Gamers, One Kubernetes Network | **Friday** 2:55pm - 3:30pm

Best Practices for Deploying LLM Inference, RAG and Fine Tuning Pipelines on K8s | **Friday** 4:00pm - 4:35pm

Best of Both Worlds: Integrating Slurm with Kubernetes in a Kubernetes Native Way | **Friday** 4:55pm - 5:30pm