SDC | StorageAI™

A SNIA Event

April 29, 2026 • Denver, Colorado

# An Update on Accelerated Object Storage for AI/ML

Jason Goldschmidt (Dell Technologies)
Shrikant Mether (HPE)

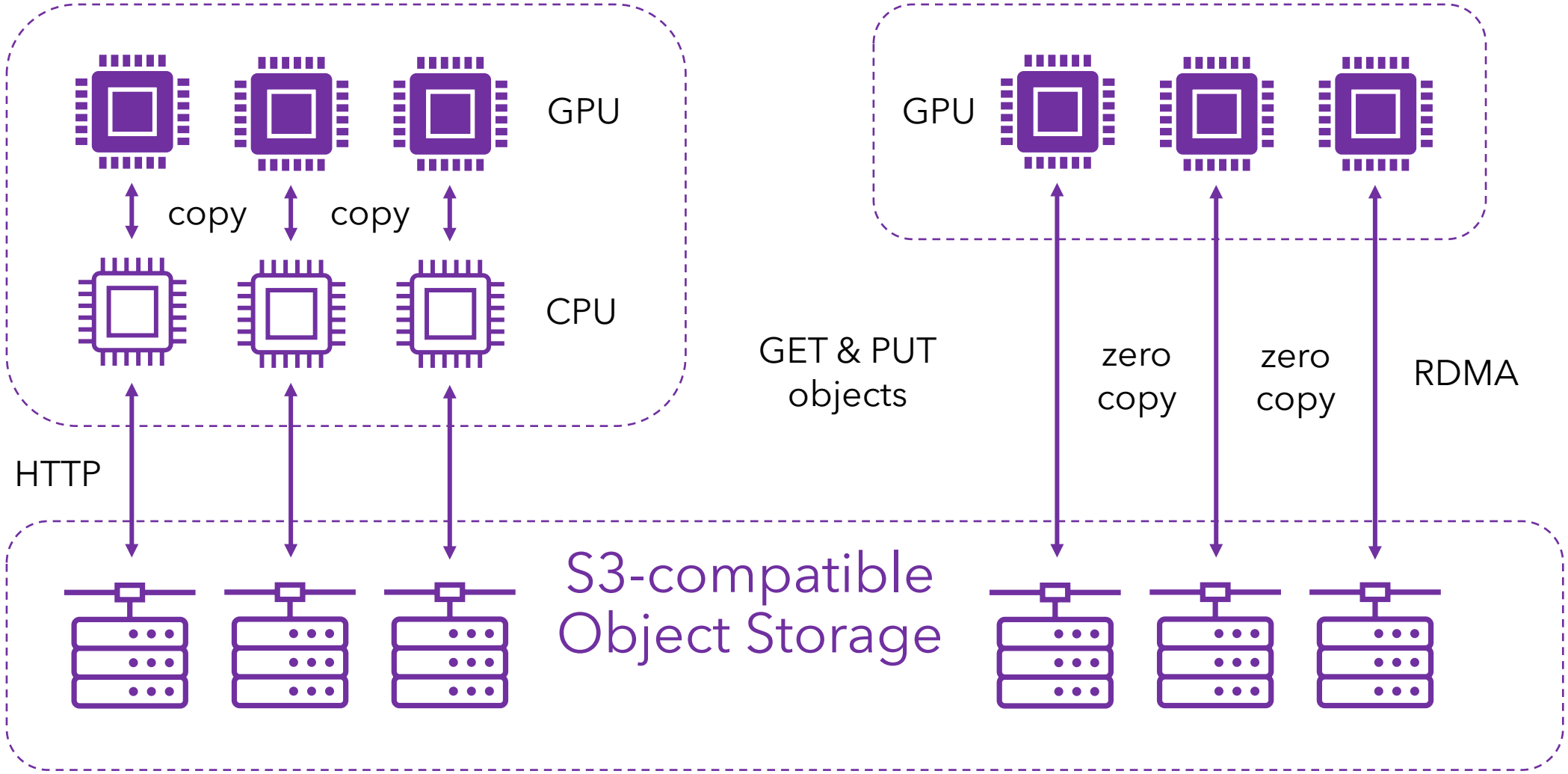
# Why Accelerated Object Access Matters for AI

- AI workloads stress the storage–compute boundary
  - GPU-centric pipelines demand low-latency, high-bandwidth, predictable data delivery
- Object storage is a natural fit but conventional object access becomes a limiter
  - CPU-mediated data paths, protocol overhead, and extra copies reduce efficiency and determinism
- An emerging industry direction
  - RDMA-Accelerated object storage enables direct, zero-copy data movement into application and GPU memory
  - NVIDIA has released accelerated CUDA libraries for object storage, and is collaborating with object storage vendors
  - SNIA Accelerated Object I/O TWG is evangelizing and working toward interoperability for this class of access

# Traditional S3

# RDMA Accelerated S3



# Use Cases benefiting from RDMA Acceleration

- Many S3-compatible object storage systems optimize for aggregate throughput at scale with large number of clients
- AI/ML workloads instead demand high per-thread throughput and low latency from a small number of clients
  These workloads include:
  - Inferencing (KVCache)
  - Data Loading
  - Indexing (E.g Vector Databases)
  - Checkpointing for training and inference
- RDMA Acceleration for S3-compatible storage can meet the requirements for AI/ML workloads
  - Achieving zero-copy data movement from object store into GPU/CPU memory via RDMA, with CPU-managed control path

# A Vendor Perspective – Dell Technologies

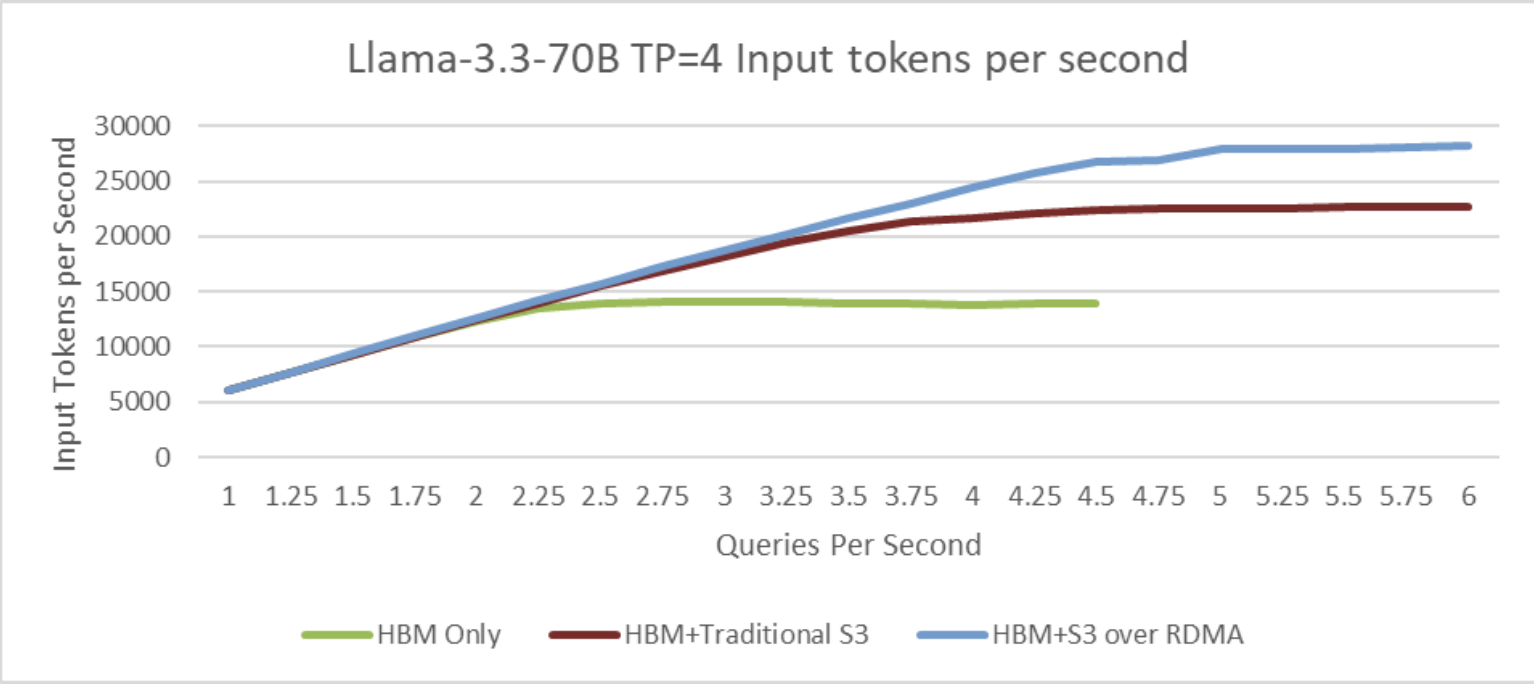
- What Dell customers are telling us
  - AI infrastructure is increasingly constrained by data delivery, not compute
  - GPUs must stay busy – requiring low-latency, high-bandwidth, predictable access to data
  - Object storage is favored for scale, economics, and governance – if performance gaps can be closed
- How Dell is responding
  - By utilizing the NVIDIA cuObject libraries, ObjectScale 4.3 preserves S3 compatibility while enabling accelerated data paths with RDMA
  - Opensource contributions to popular AI Inference frameworks and KVCache connectors (LMCache, NIXL)
  - Dell Storage Performance Tool: open-source benchmark for S3-compatible storage with RDMA acceleration
- Why SNIA matters to us as a vendor
  - Customers expect interoperability, portability, and ecosystem choice
  - Open specifications enable broad adoption without vendor or silicon lock-in
  - SNIA provides a neutral forum to align object semantics, accelerated access, and diverse hardware platforms

*Accelerated object access isn't about a faster protocol – it's about preserving the value of object storage as AI becomes the dominant workload, while doing it in an open and interoperable way.*

# Case Study: Dell ObjectScale and Multi-turn AI Inference

- AI-Server: Dell XE9680 8xH100 NVIDIA GPUs running vLLM utilizing LMCache and NIXL
- Simulating 80 users performing multi-turn long-input/short-reply chatbot queries
  - HBM-only (2.25 QPS),
  - ObjectScale with traditional S3 API-based storage (3.75 QPS)
  - ObjectScale with RDMA accelerated S3 (5.25 QPS) – **doubles token throughput compared to HBM-only KVCache**



# A Vendor Perspective - HPE

- What HPE customers are telling us
  - Enterprise AI infrastructure is rapidly adopting optimizations to enhance “effective” GPU utilization
    - Costumers expect GPUs to be busy generating new tokens rather than waste compute reprocessing input prompts
  - In the AI/ML world, dominated by unstructured data, object store is foundational for storing petabytes of data. Making this data directly available in GPU memory at high-throughput, low-latency would significantly improve efficiency of enterprise AI infrastructure

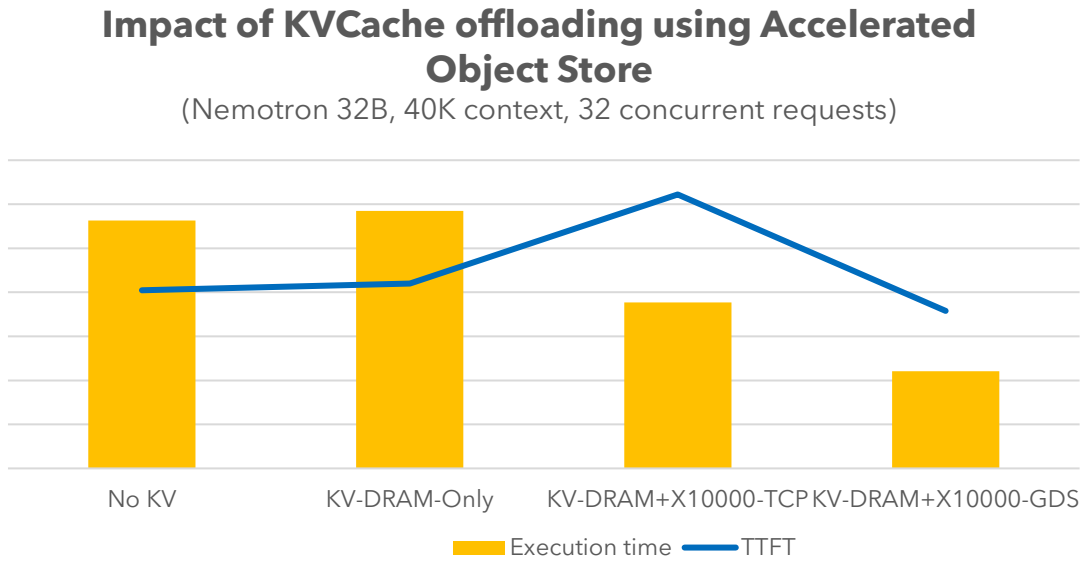
- HPE’s Experience
  - HPE explored accelerated object access using NVIDIA’s cuObject Accelerated CUDA libraries for object store
  - Partnering with NVIDIA, HPE developed accelerated object access support into HPE Alletra X10000 product and HPE S3 client SDK. We observed significantly improved throughput, lower latency and reduced client resource utilization.
  - To unlock full potential of accelerated object access for AI/ML applications, HPE is integrating RDMA capable S3 SDK into frameworks powering AI/ML applications like PyTorch, Nvidia Inference Transfer Library (NIXL), LMCache
  - HPE believes that having an interoperable specification for accelerated object access will help gain wider adoption for this emerging technology

# Case Study: KVCache offload using Accelerated Object Store

- **AI Server:** HPE Proliant DL380a Gen12, 4xH200 NVIDIA GPUs, vLLM + LMCache
- **Experiment**: Nemotron-terminal 32B, 40K context, 256 parallel users, 32 active at a time.

- KV Cache offload using RDMA Accelerated object store can provide significant improvements in inference throughput as compared to offload to CPU DRAM or object store over traditional S3




*Note: These are preliminary results. Production systems are expected to perform better.

# An Industry Response: SNIA Accelerated Object IO TWG

- Enable broader adoption of fast object to address emerging AI/ML workload needs
- Per the charter, the TWG aims to produce an interoperable SNIA specification for RDMA accelerated S3-compatible object storage
  - Driven by AI/ML use cases
  - Opensource reference software implementations
- Produce and deliver educational material for wider adoption
- Current work item: Interoperable RDMA Accelerated S3-compatible client and server definition
  - Discovery mechanism for clients to determine server capabilities
  - Metadata definitions for client-server interaction and RDMA configuration
  - Encourage client-server decoupling through an open protocol specification

# An Industry Response: SNIA Accelerated Object IO TWG

- Contact TWG co-chairs with questions and how you can contribute
  - acc-obj-io-twg-chair@snia.org
    - Jason Goldschmidt jason.goldschmidt@dell.com
    - Nick Connolly nick.connolly@arm.com
- Now is a perfect time to join the TWG and participate in the adoption of this emerging protocol definition
  - Define requirements, answer open questions, broaden use-cases
  - Align with other TWGs in the StorageAI community
  - Participate in future object plug-fests


SDC | StorageAI™
A SNIA Event
Thank You