SNIA DEVELOPER CONFERENCE
SDC 25
By Developers FOR Developers

Hyatt Regency Santa Clara, CA
September 15-17, 2025

Accelerating Object Storage for AI/ML with S3 RDMA

Jason Goldschmidt
Dell Technologies
Member of Technical Staff – Office of Storage CTO

www.sniadeveloper.org

Objective and Postulates

- Customers want a faster Object service to support AI workloads
  - Previous, S3 not used as primary data source due to latency
  - Performance has prevented S3 adoption for Gen AI *on-prem*
  - Object presents better control mechanism than files
  - Trends suggests a slow, but steady, increase in Object adoption as primary storage for AI
- A faster object storage
  - High performance, low latency storage system for an application developed for S3
  - S3 semantics, but with significant and measurable performance improvements for data transfers
  - Utilize RDMA to facilitate high-throughput direct data placement with ultra-low latency
- S3 over RDMA represents an opportunity for performance sensitive applications

Goal: Achieve file-like performance for Objects within a datacenter environment

- Involve RDMA operations for S3 data transfers
- Objects are transferred memory to memory between S3 server and client
- RDMA provides high throughput and low latency data transfer mechanisms
- RoCEv2 allows use of existing Ethernet infrastructure
- Dell Technologies is collaborating with NVIDIA to integrate S3 RDMA acceleration technology into Dell ObjectScale solutions

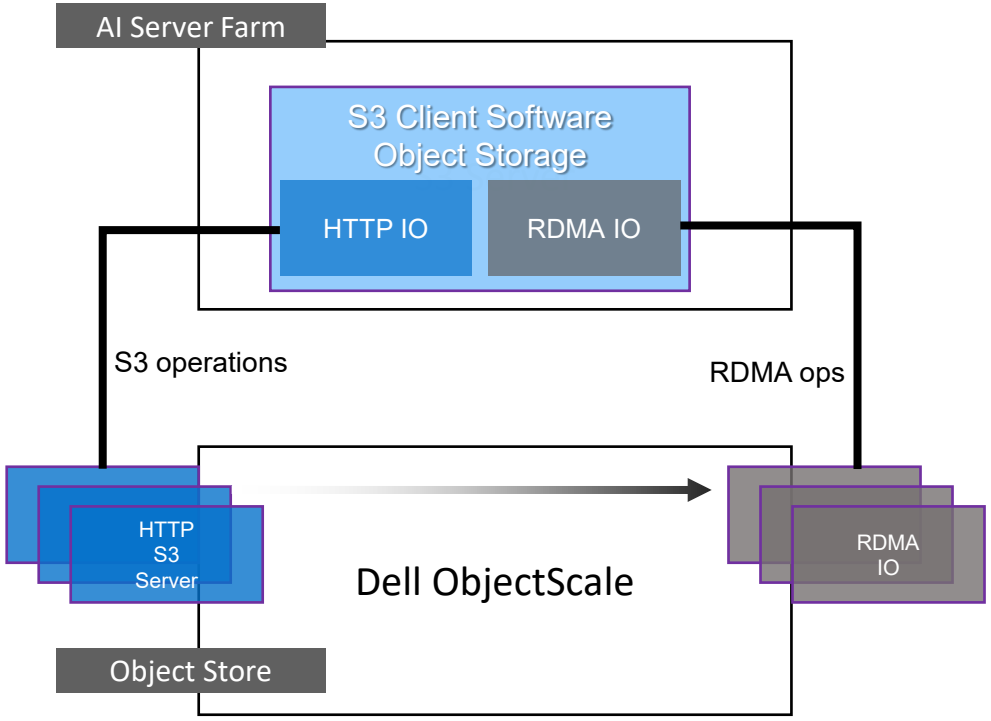
S3 RDMA Use Cases

- Traditionally, S3 is very well suited for high throughput transfers involving many threads and clients
- AI/ML workloads are, however, requiring high throughput transfers for very few threads and very resource dense systems (AI Servers) These workloads include:
  - Data Loading
  - Indexing
  - Checkpointing
  - Inferencing (KVCache)
- S3 RDMA can meet the requirements for AI/ML workloads

Case Study: Dell ObjectScale v4.1 with S3 over RDMA

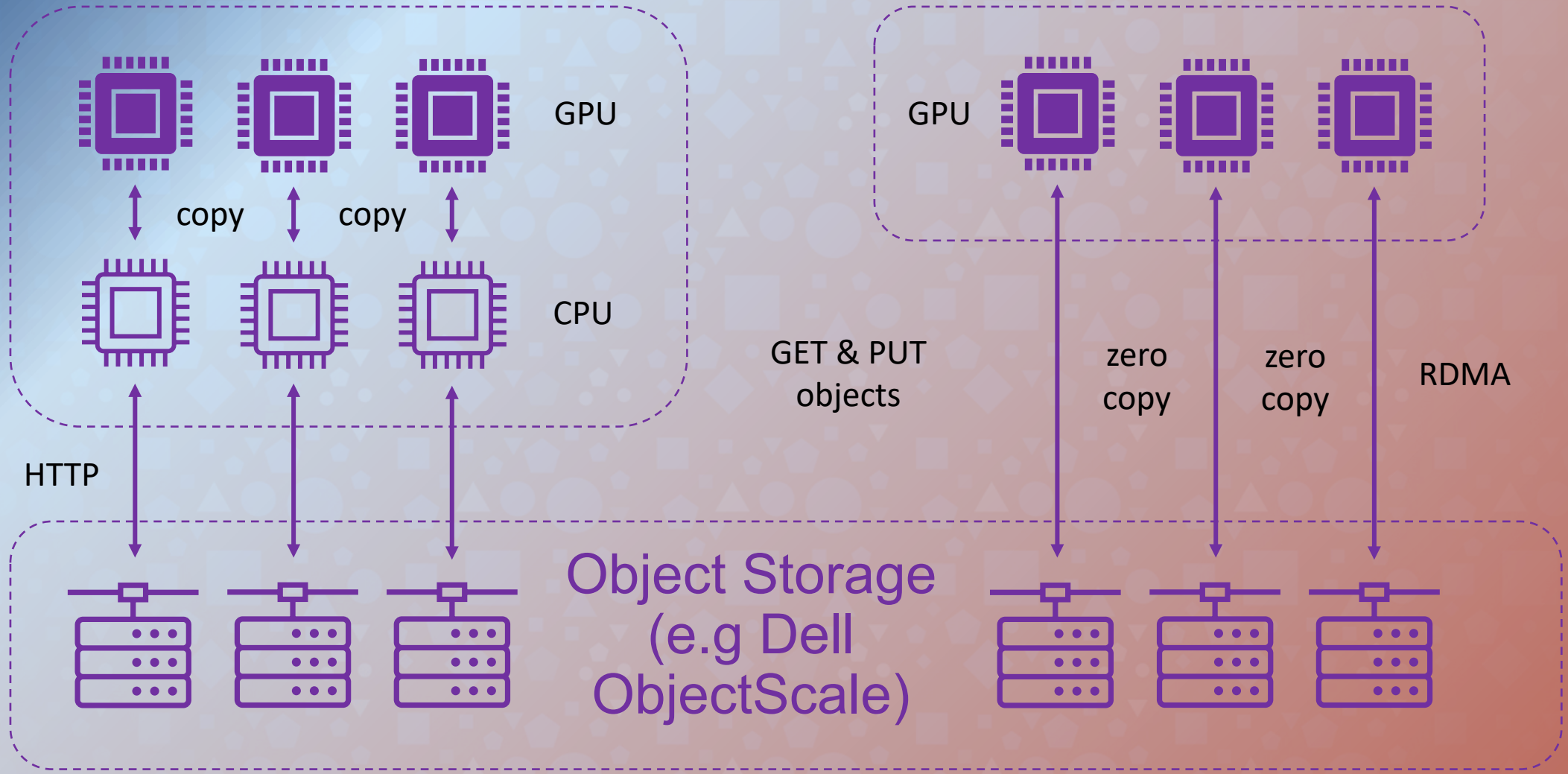
S3 RDMA Model Developed with NVIDIA & Partners

- Reduced AI/ML training and inferencing time
- AI/ML GPU dedicated
- Lower CPU utilization
- Faster transfers under CPU load
- Lower latency

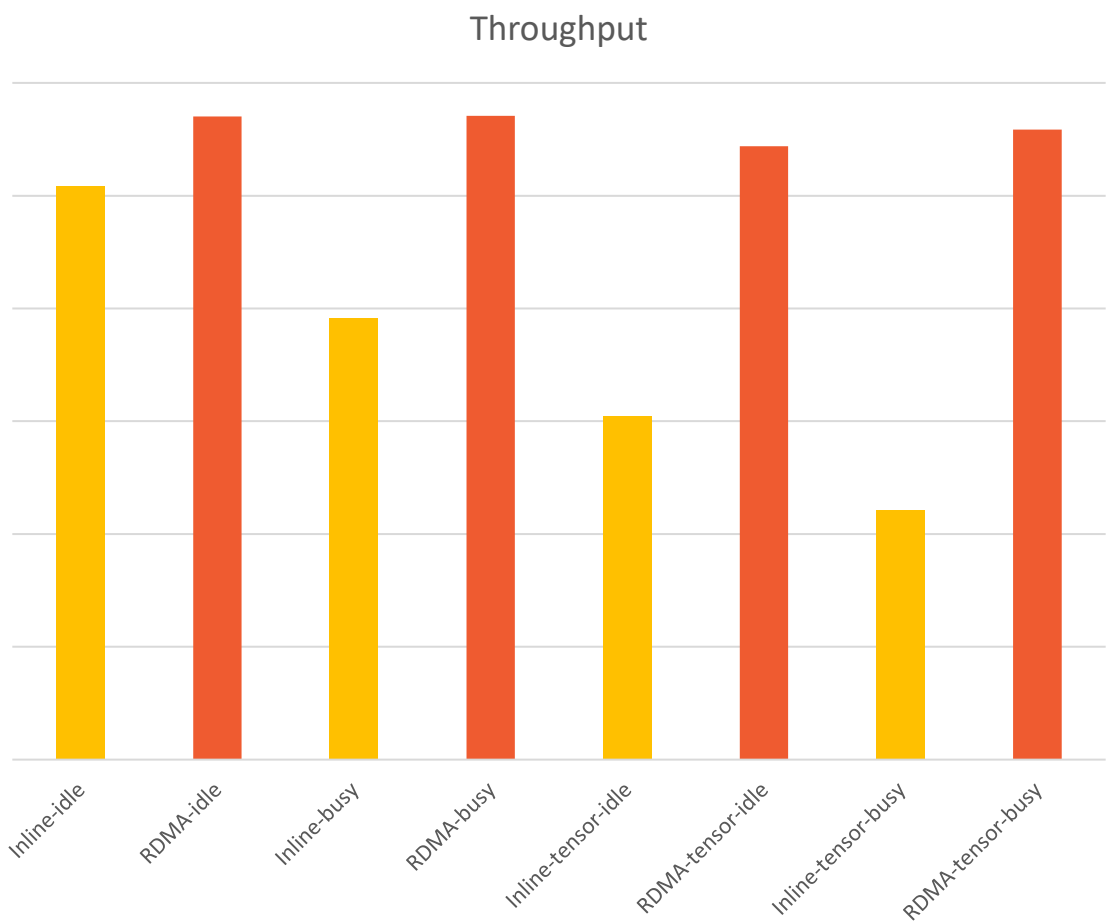


Traditional S3

S3 RDMA with GPUDirect

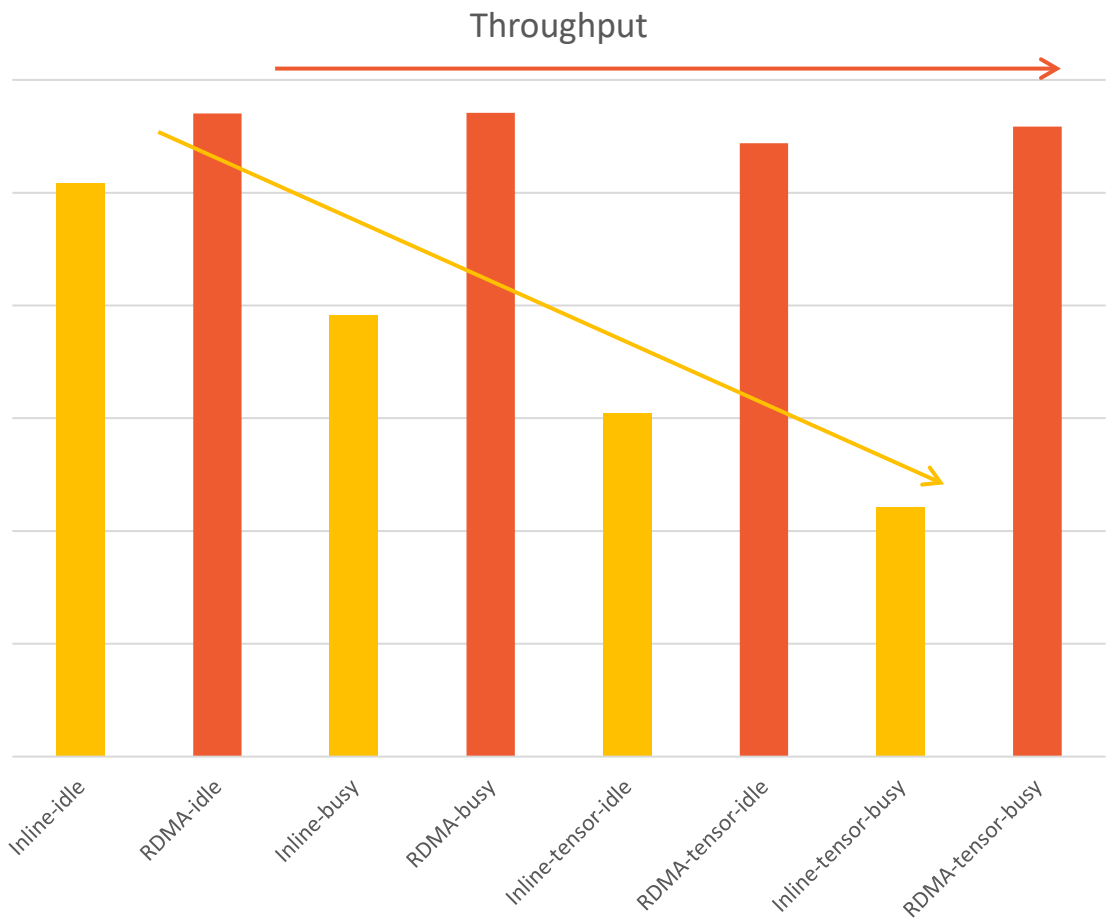


Single GPU client test



- Comparing Inline vs RDMA Large Object GET
- Scenarios:
  - Idle client object transfer
  - Busy client object transfer
  - Idle client tensor load into GPU
  - Busy client tensor load into GPU

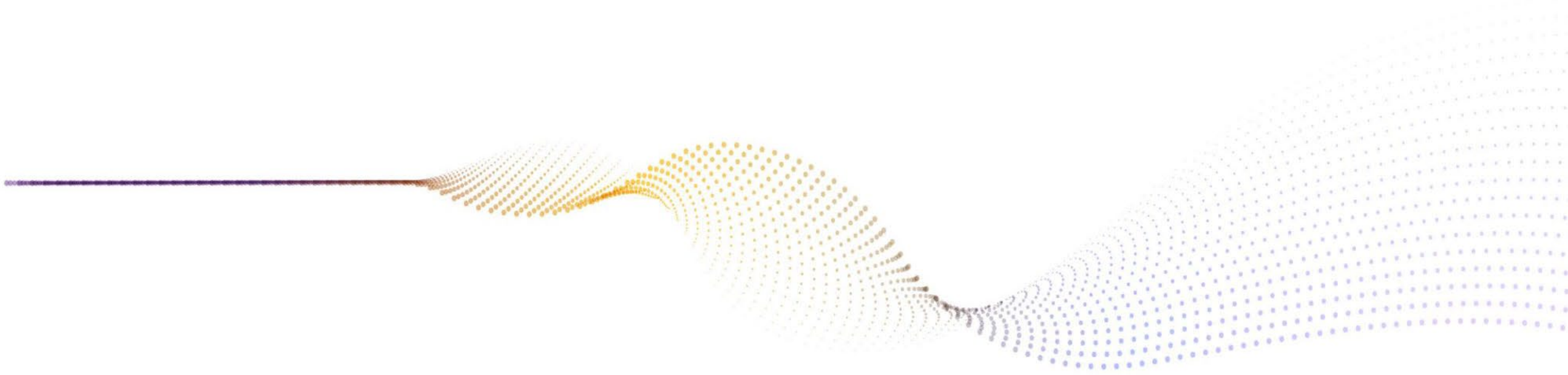
Single GPU client test



- Comparing Inline vs RDMA Large Object GET
- Scenarios:
  - Idle client object transfer
  - Busy client object transfer
  - Idle client tensor load into GPU
  - Busy client tensor load into GPU

Performance and Scale Results

- S3 over RDMA and inline throughput are similar; inline requires far more client-side resources (20-30x CPU utilization) to perform the data transfer
- S3 over RDMA performs zero-copy GPUDirect data transfers; inline requires an additional data copy, which requires 40% additional GPU utilization
- S3 over RDMA maintains throughput on a busy client system; inline performance degrades by half
- **S3 over RDMA is uniquely positioned for AI/ML workloads requiring GPUDirect operation and consistent throughput/latency when client-side resources are heavily utilized**



Thank you for attending!

Please remember to rate this session. You get access the presentations at
http://sniadeveloper.org/conference

