SNIA | Data, Storage & Networking

AI Meets Storage:
Comparing On-Prem, Cloud, and Hybrid Architectures Across the AI Lifecycle

Live Webinar
June 11, 2026
11:00 am PT / 2:00 pm ET

Today's Presenters

Erik Smith
Distinguished Engineer
Dell Technologies
Moderator

Rohan Mehta
Member of Technical Staff -
Systems Performance Architect
Micron Technology

Himabindu Tummala
Distinguished Engineer
Dell

The SNIA Community

200
industry leading organizations

2,000
active contributing members

50,000
IT end users & storage pros worldwide

SNIA | DSN A SNIA Community

SNIA | Data, Storage & Networking

What We Do

Drive the awareness and adoption of a broad set of technologies, including:

- Storage Protocols (Block, File, Object)
- Traditional and software-defined storage
- Disaggregated, virtualized and hyperconverged
- AI, including storage and networking considerations
- Edge implementation opportunities and factors
- Storage and networking security
- Acceleration and offloads
- Programming frameworks
- Sustainability

How We Do It

By delivering:

- Expert webinars and podcasts
- White papers
- Articles in trade journals
- Blogs
- Social Media
- Presentations at industry events

Logistics

- The slides are available under the attachments tab at the bottom of your console.
- Questions are welcome!
- Please rate the session and provide feedback!
- Want more sessions like this or other topics, let us know!
  - JOIN US! We meet on Thursday mornings at 11:00 AM eastern.
  - Email dsn-chair@snia.com if you have questions.

SNIA Legal Notice

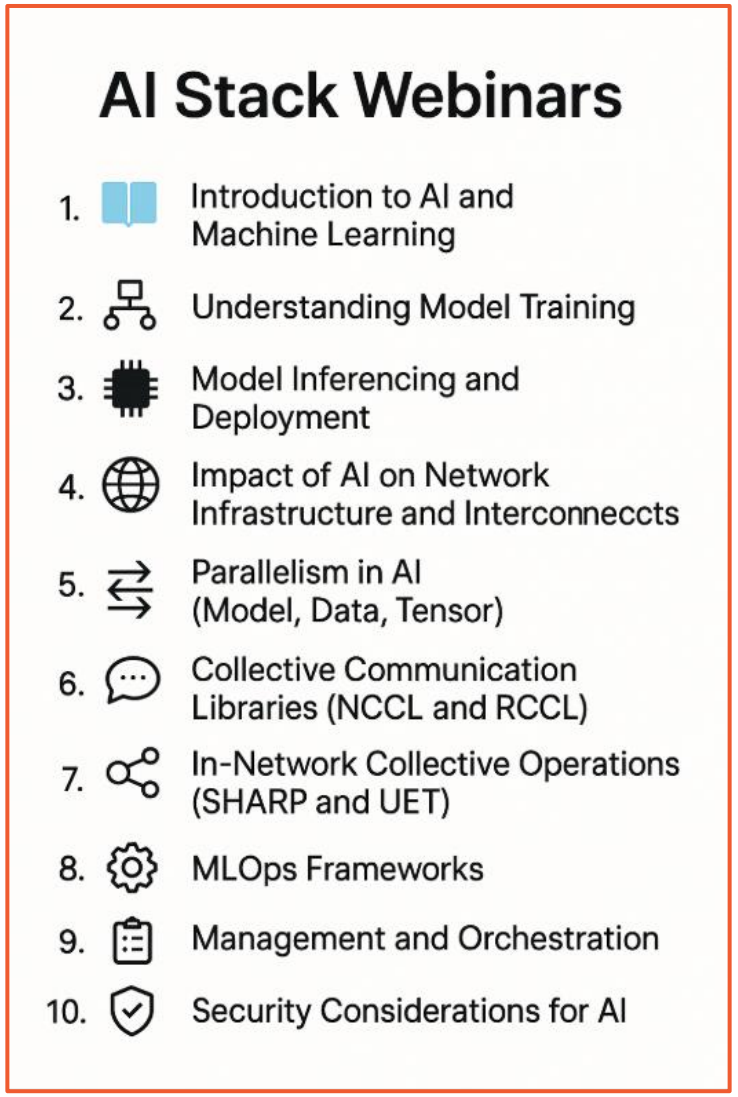
- The material contained in this presentation is copyrighted by SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

  NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

The “AI Stack” Webinar Series

- Building a Strong Foundation for All Experience Levels:
  - Starting from the basics
  - Building steps-by-step
  - Connecting theory to practice
  - Demonstrations
  - Preparing for real-world challenges

Watch all “AI Stack” Webinars at the SNIAVideo YouTube Channel:
https://www.youtube.com/playlist?list=PLH_ag5Km-YUZaWla60wr-s3vqX0M40-t-



SNIA | DSN
A SNIA Community

Introduction

Agenda

Overview of Storage solutions available today	Brief overview of the 3 different types of Storage solutions available today.
Recap of Storage SW Stack concepts	Core concepts that drive decision making behind which Storage solution to use for given use-case.
Deep Dive into Cloud and On-prem	
Requirements of the killer-app of today: AI/ML	Introduction to the specific storage requirements driven by AI workloads and data complexity.
Use Case Matrix & Decision Framework	Guidance tool featuring use case scenarios and a framework for selecting storage solutions based on needs.

Quick Poll

1. **How familiar are you with this topic?**
   A. I’m new at this
   B. I'm working with storage architecture, but looking forward to learning more
   C. I live with this day and night

SNIA | DSN
A SNIA Community

Difference between Cloud, On-Prem, and Hybrid Storage Systems

Cloud vs On-Prem vs Hybrid

Cloud Storage

- Scalability and elasticity
- Remote accessibility for data and tools
- Dynamic workload demands
- Pay-as-you-go

On-Prem Storage

- Direct control and compliance over data and infrastructure
- Security and customization
- High operational overhead
- Low latency

Hybrid Storage

- Part on-prem, part cloud
- Workload dependent
- Cost, security, compliance, and performance
- Balance on-prem expertise and scale

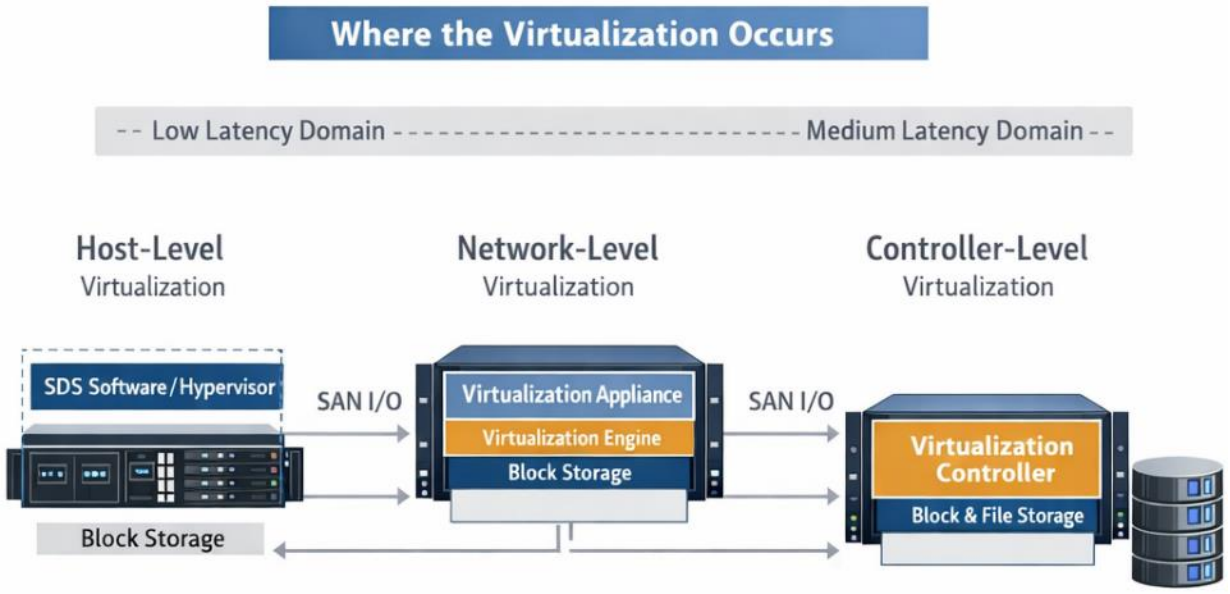
ISO Definitions vary a bit…

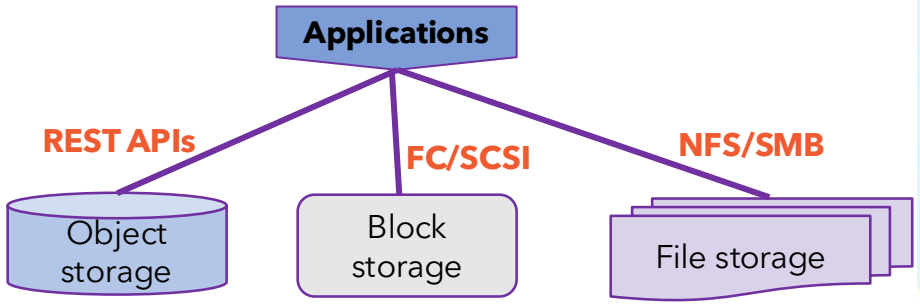
- ISO/IEC 22123 uses the terms “public”, “private”, and “hybrid” for defining cloud deployment models.
- Distinguishing factor is primarily from the perspective of users’ access levels rather than ownership or physical location of resources.
  - For example: A **private** cloud can be owned, managed, or operated by the you (customer) or a third party and can exist on premises or off premises. If it is hosted off premises, the cloud resources can be owned, managed, and operated by a **public** cloud service provider.

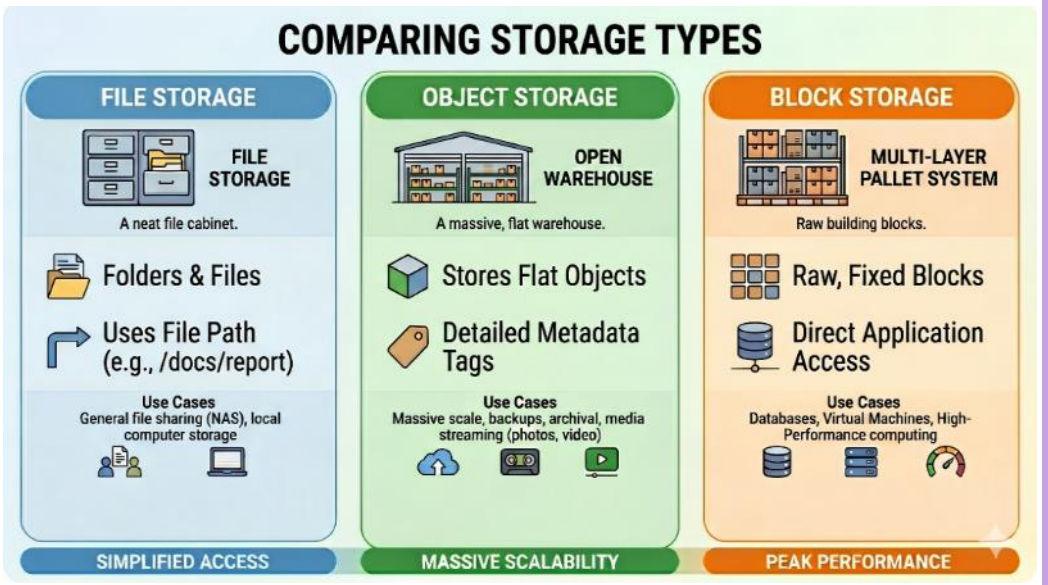
Recap of Storage Software Stack Webinar On-demand at:

https://www.snia.org/educational-library/deconstructing-storage-software-stack-legacy-silos-ai-scale-architecture

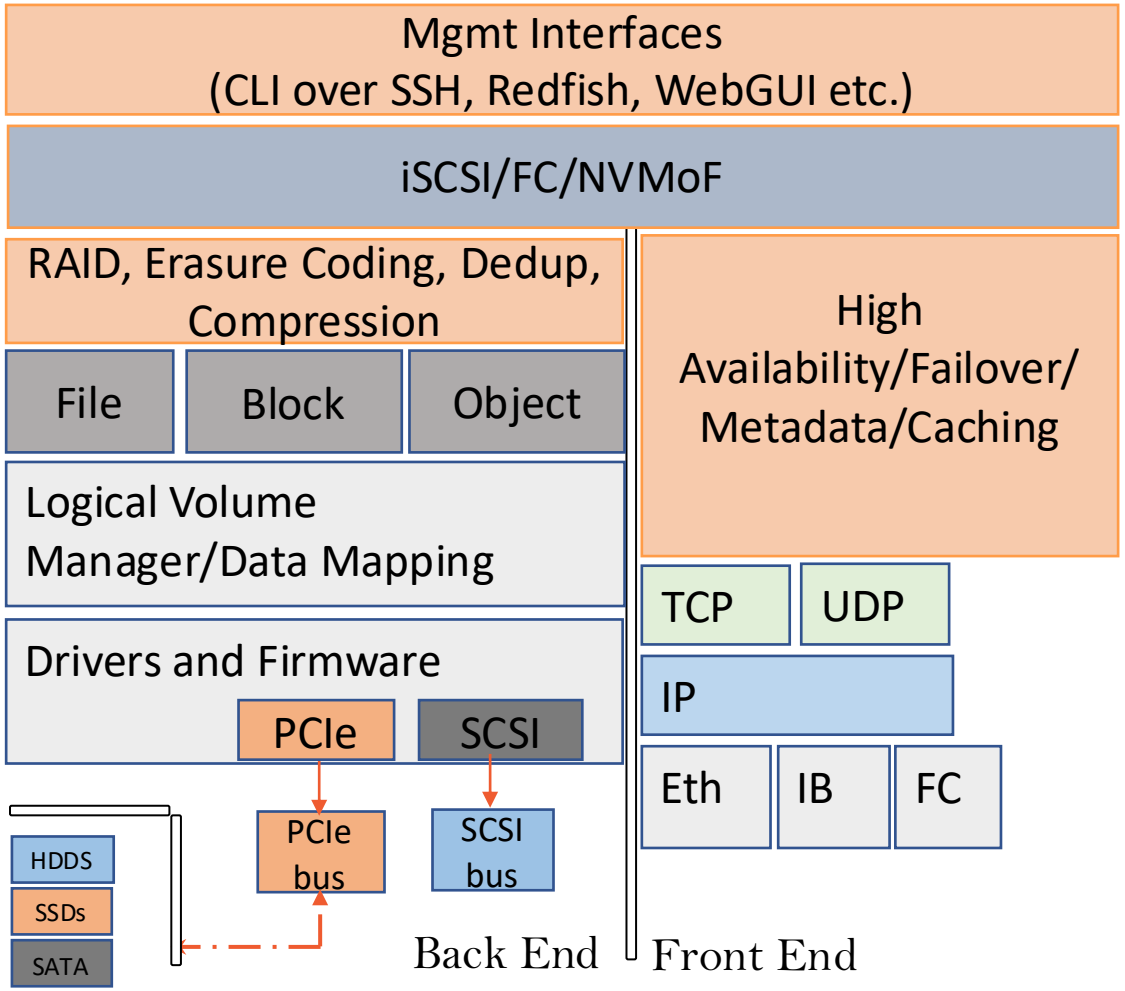
Storage Virtualization Techniques







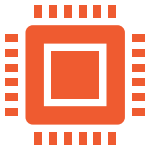
Storage Controller



SW Stack Functionalities

- **Front-End :** Connectivity, Low Latency, and Reliable Access.
  - **Services:** Multipathing, QoS, Access Control, and Metadata management.
- **Back-End :** Data Integrity, Security, and Media Longevity.
  - **Services:** Device discovery, health monitoring, and media-specific optimizations.
- **Unified Virtual Pool:** Location-agnostic storage volume
- **Advanced Resilience:**
  - **RAID:** Redundancy *within* the local storage domain.
  - **Erasure Coding:** Data resilience *across* distributed domains.
- **Efficiency Layer:** Deduplication, Compression, and **Auto-tiering** (Hot/Cold data)

Indirection and Redirection



Indirection

Accessing data through mapping layers that provide flexibility and abstraction.

The mapping layer is your centralized “brain” or “database”.

Key component for an append-only system.

Introduces amplification factor for both Reads and Writes.

Complex custom GC.

Easier snapshotting.

Low Perf

Redirection

Rerouting data requests to enable failover, load balancing, and optimized performance.

Just do some math to figure out where data is; doesn’t actually handle the data.

Great for update-in-place systems.

Low Write and Read amplification

Simpler GC.

Complex snapshotting.

High Perf.

Storage Abstraction Layers

Storage abstraction layers vary in complexity and integration, affecting flexibility and management.

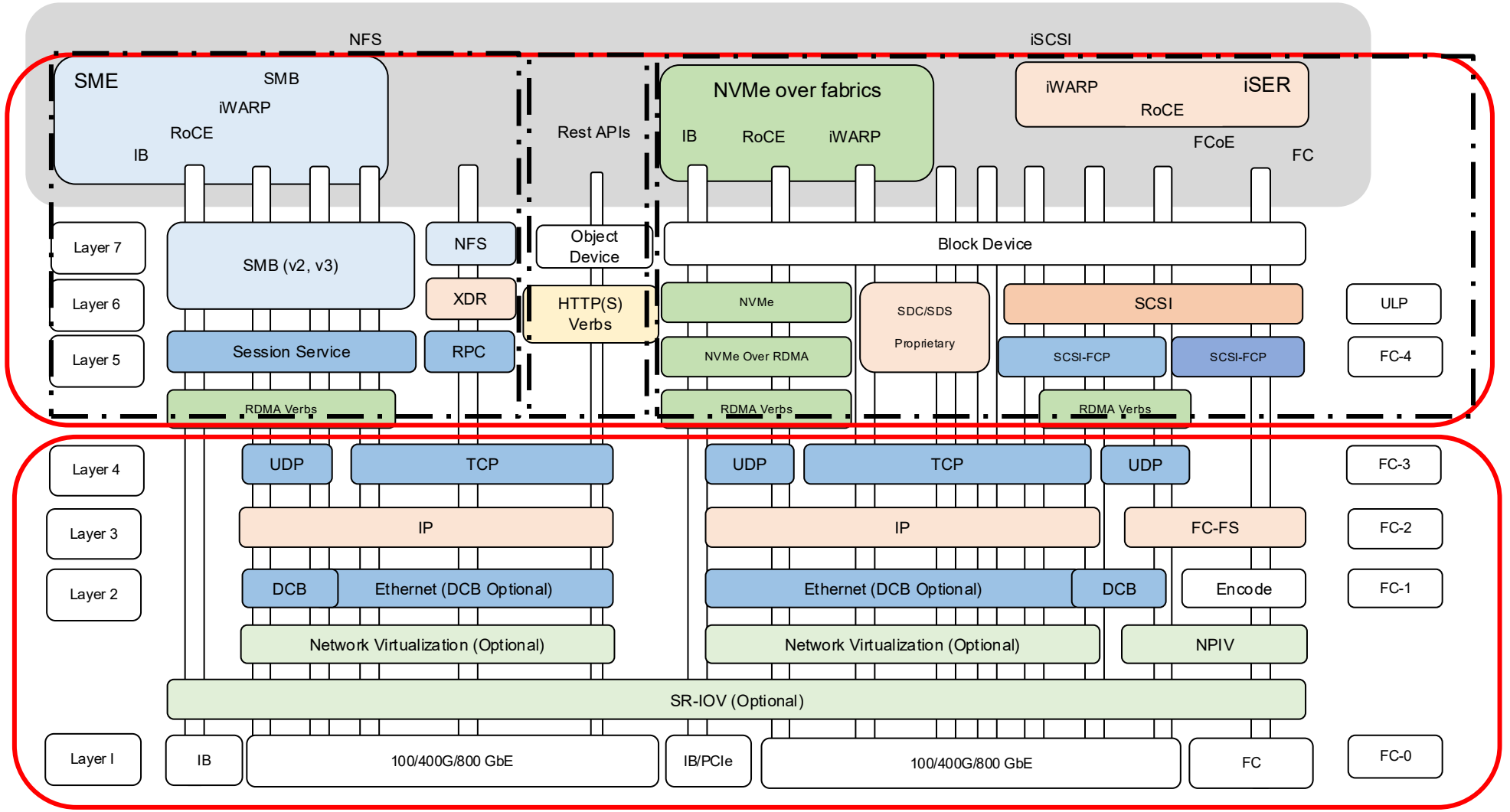
Where the durability/replication layer is present makes all the difference

Service Models Overview

Service models like IaaS, PaaS, and SaaS define control levels and responsibilities in the storage solution you are choosing.

Not a new concept. Introduced at SNIA SDC 2024 talk on “Efficient Utilization of Storage Media”.

Host Server - Software Stack



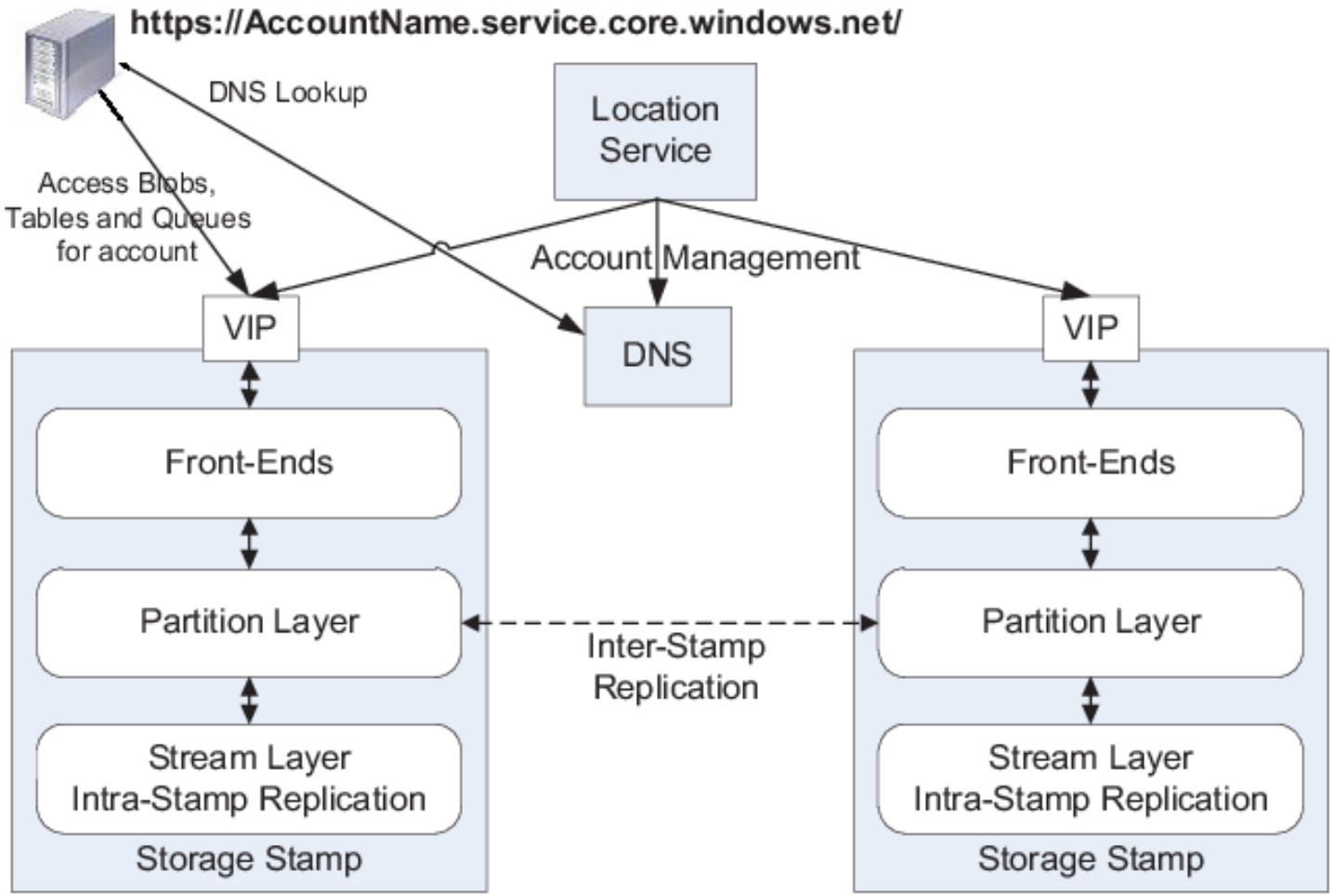
Thanks to Erik Smith for this diagram; also available here.

SNIA DSN A SNIA Community

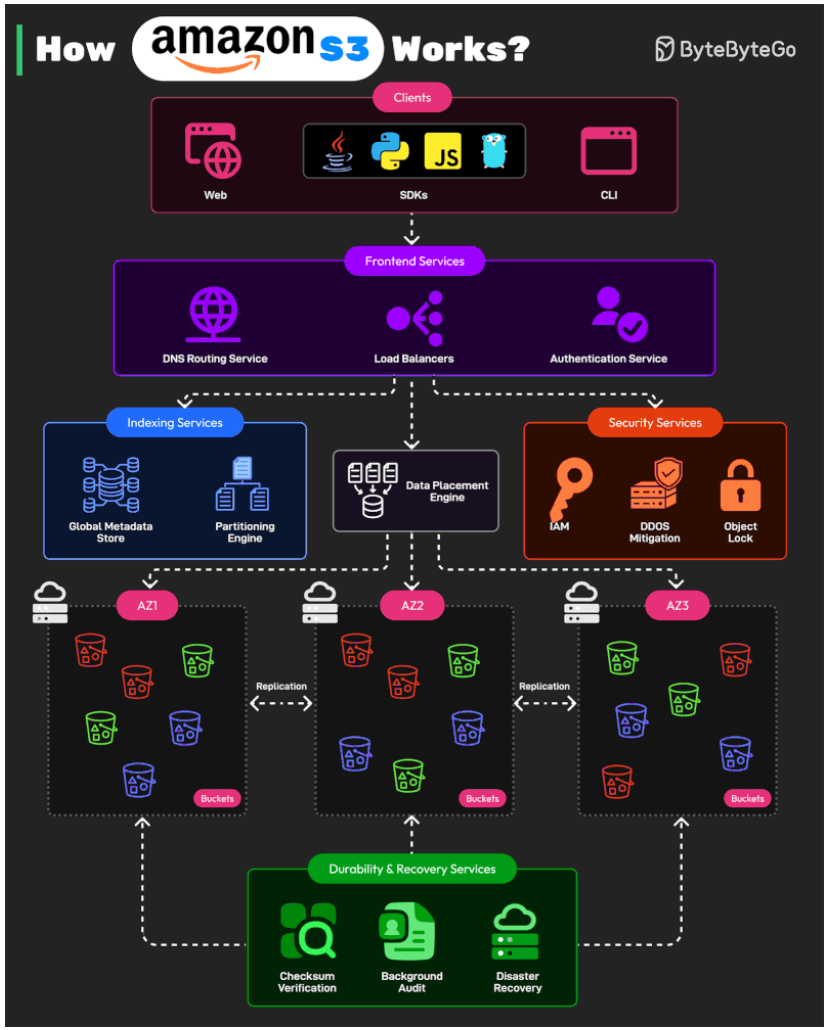
SNIA | DSN
A SNIA Community

Cloud Storage Architecture

Typical Cloud Storage Architecture



Microsoft SOSP paper



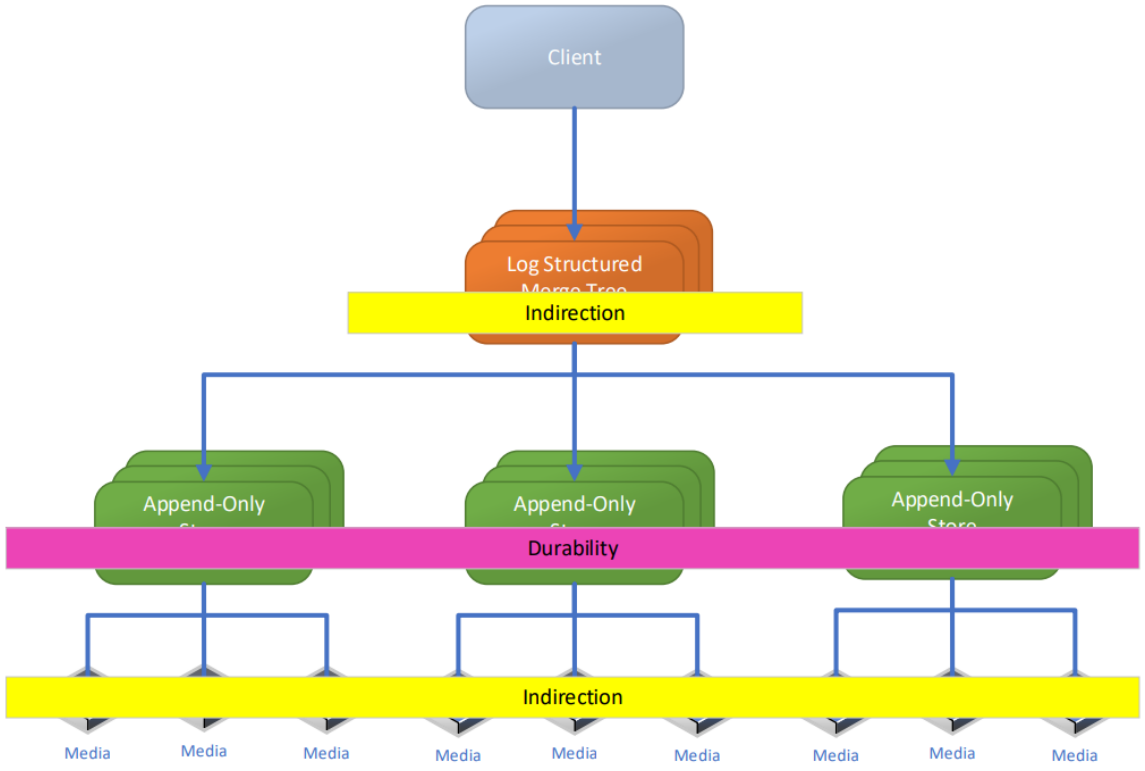
ByteByteGo post by Alex Xu

Tracing a Request: PUT and GET

- Let’s trace a single PUT request — say, you’re uploading a 100 MB training shard to an object bucket.
- DNS and Load balancing → FE (authn/authz) → Service Categorization → Metadata Lookup → Data placement → Durability → Background services

- Can you think in reverse for a GET request?
  - What would happen if 100s of simultaneous GET requests come in for the same data that you PUT earlier?

Durability, Consistency, and Placement



- Hardest problems to solve!
  - Durability – replication and erasure coding
  - Consistency – the only tweak-able factor in CAP theorem, historically traded against Availability
  - Placement – partitioning, sharding, merging, metadata load-balancing

Ref: "Efficient Utilization of Storage Media"

“Everything as a Service” Model

Stage of AI Pipeline	Size	Type of Storage
Raw Input Data	Tens of petabytes	*Capacity-based file or object storage*
Refined AI Training Data	Petabytes	*High-performance file storage*
Training Checkpoints	Terabytes	*High-perf file storage (recent); capacity-based file/object (older)*
Trained AI Model	Megabytes to terabytes	*High-performance file storage*
Quantized Model	Megabytes to gigabytes	*High-performance file or object storage*
Alerts Metadata	Terabytes (over time)	*High-performance file or object storage*
Vector Database for RAG	Petabytes	*High-perf file + capacity-based object storage*
Inference & Query Logs (Compliance)	Petabytes	*Capacity-based object storage*

- Every interaction is a service.
- Users get virtualized access to storage; implementation is managed by the cloud provider.
- Objects, blocks, files – available to all AI phases; best options are evolving as workloads keep evolving.
  - Training seems to have settled down to object and file storage.
  - AI focused managed file systems (e.g.: Lustre) are gaining prominence.
- **Everything (typically) runs on the same infrastructure!**

Quick Poll

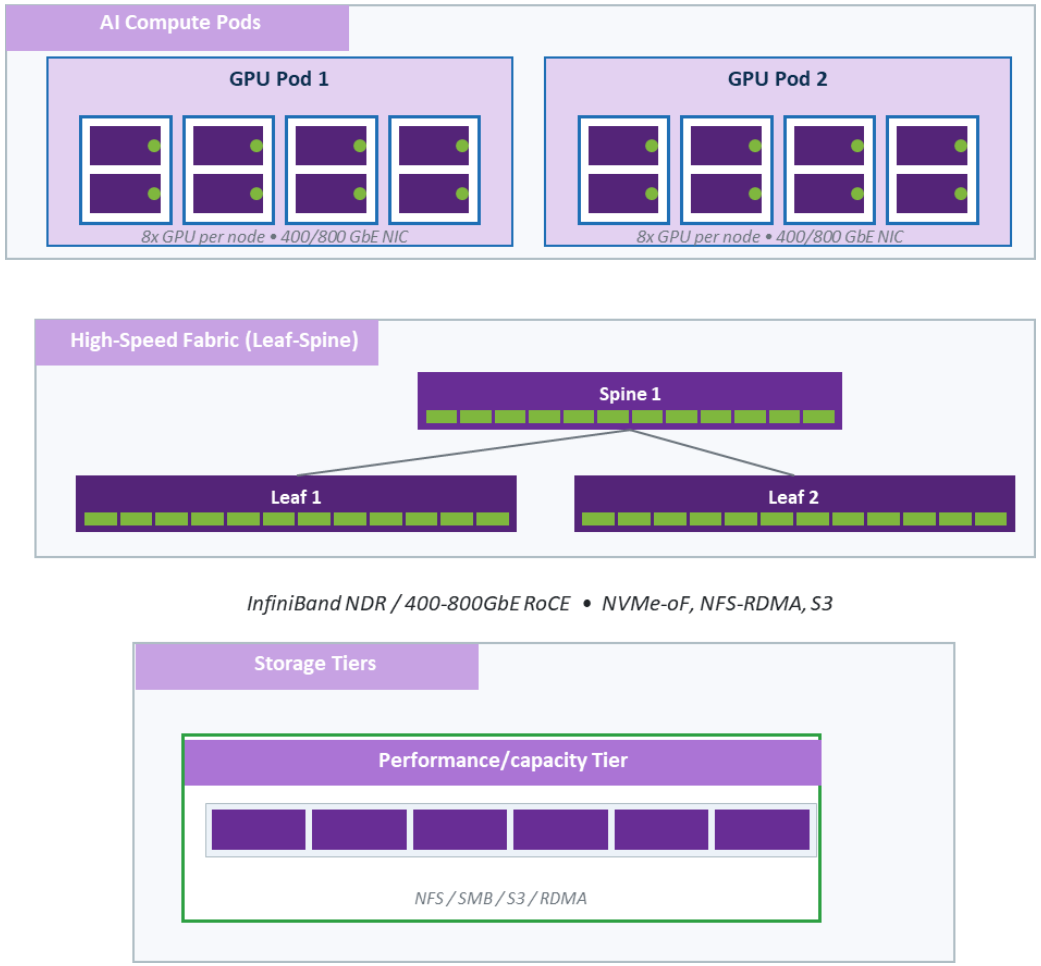
2. How helpful was the summary of storage virtualization techniques and protocol stack in improving your understanding of cloud storage architecture?

A. Very helpful
B. Somewhat helpful
C. Not at all helpful

SNIA | DSN
A SNIA Community

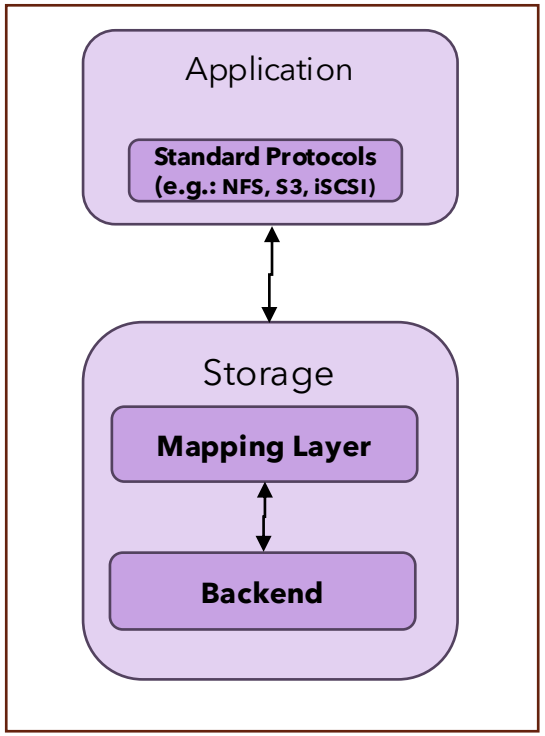
On-Prem Storage Architecture

On-Premises AI Reference Architecture Overview

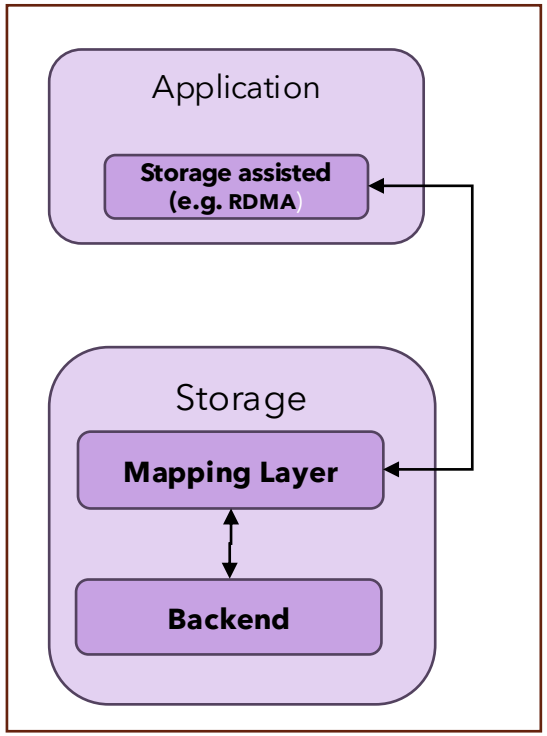


- GPU Compute Pods
  - Accelerator-dense GPU compute pods form the core, connected via high-speed network interfaces.
- Leaf-Spine Network Fabric
  - Non-blocking leaf-spine fabric using InfiniBand or high-speed Ethernet enables efficient east-west GPU and storage traffic.
- Tiered Storage Systems
  - Storage is tiered with performance file systems, object storage, and low-latency NVMe-over-Fabrics coexisting on parallel fabrics.
- Centralized Control Plane
  - Provides identity integration, encryption, monitoring, snapshots, replication, and policy-driven tiering.

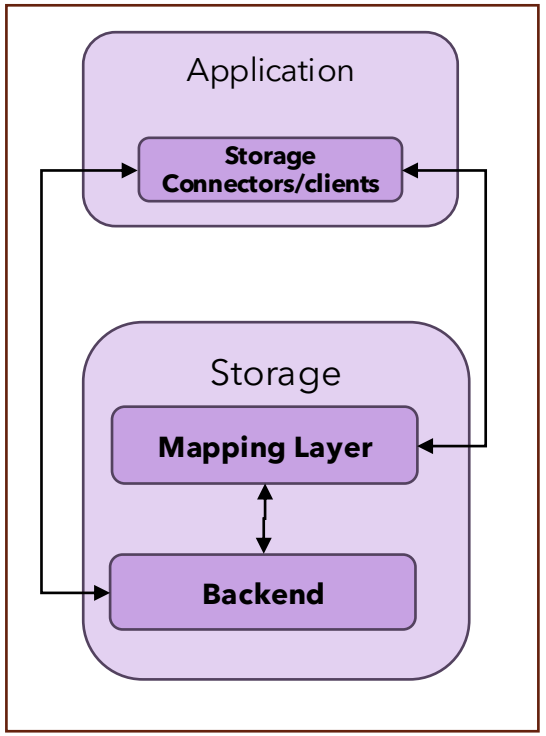
Different Storage Architectures



Centralized
Data access is managed through a centralized mapping layer that abstracts physical storage. Clients use standard protocols (NFS, S3, iSCSI), while the storage system handles data placement, metadata, and optimization.

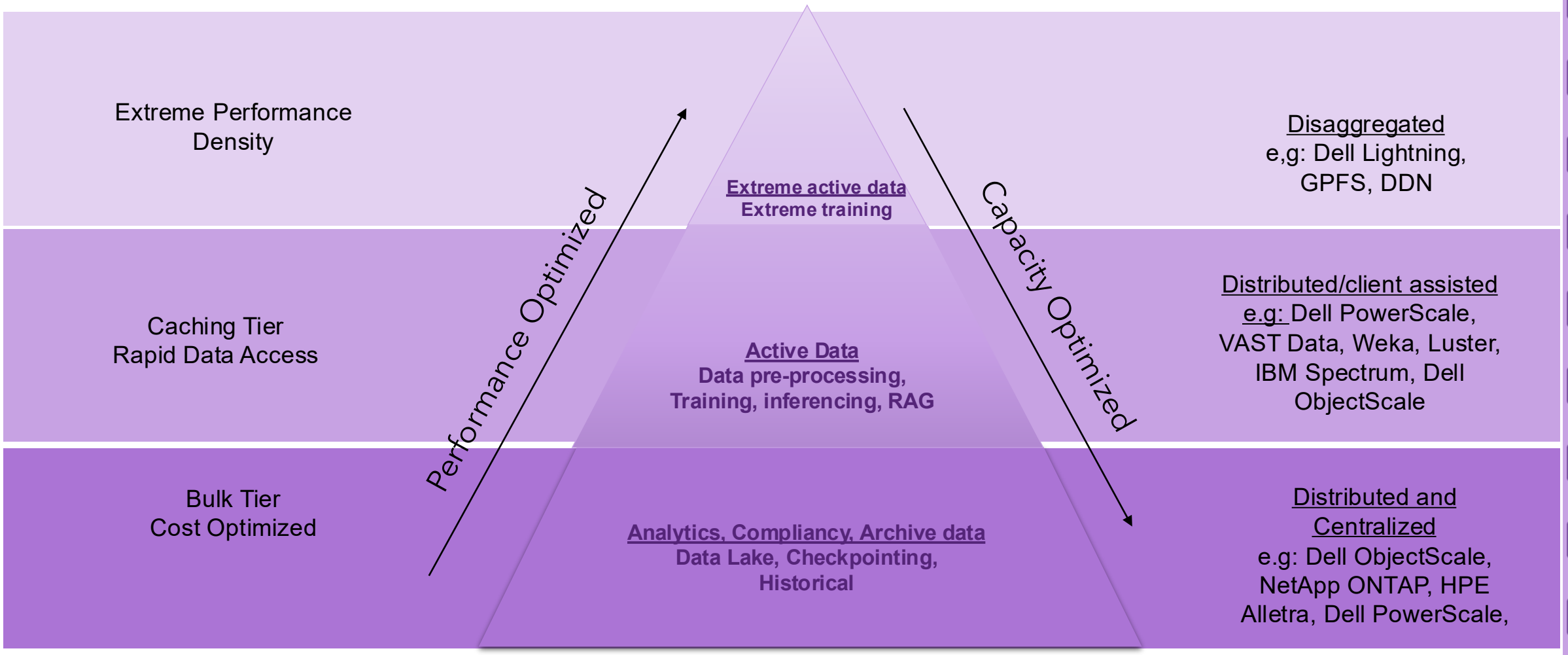


Distributed/client assisted
Data access is optimized by pushing some mapping and data placement intelligence to the client side. Clients participate in caching, load balancing, and parallel I/O, reducing reliance on centralized control.



Disaggregated
Storage intelligence is separated from data paths, enabling clients to directly access data in the backend using metadata-driven routing. The mapping layer guides where data resides without being in the data path.

Storage Architecture Pyramid for AI workloads



SNIA | DSN A SNIA Community

SNIA | DSN
A SNIA Community

Storage Fundamentals for AI Workloads

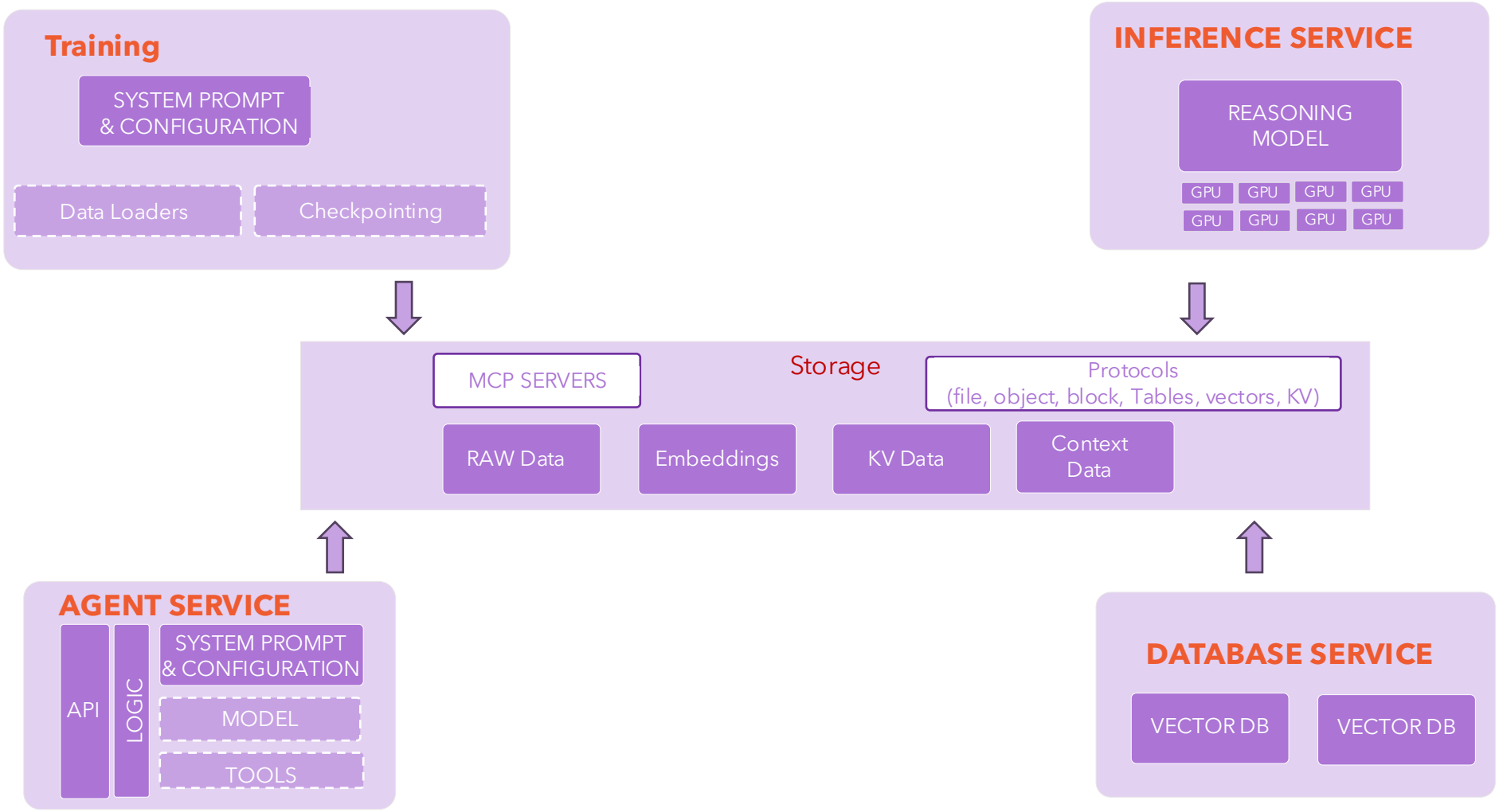
AI Processing Stages

Stages	Ingest Stage Storage	Preparation and ETL Patterns	Training/Fine Tunning	Inference/RAG/Agentic
Behavior	Ingest captures raw data at scale using object or file protocols tolerant to variable latency.	Preparation refines data into training-ready formats with metadata-heavy, mixed I/O patterns.	Training uses sustained high-throughput reads and periodic large checkpoint writes for model state.	Inference requires minimal delay access to model weights and caches, ensuring low latency.

AI Workloads (Training → Inference → Agentic)

AI Phases	Workflow (Collapsed)	IO Workload Patterns (Collapsed)
Large Model Training	Data preprocessing, data loading, checkpointing, and model finalization forming an end-to-end training pipeline from raw data to inference-ready models	Extremely large-scale (PB–EB) data processing; highly concurrent read/write during preprocessing; write-once, massively parallel read access to feed 1000s of GPUs (latency critical); bursty, high-throughput sequential writes for checkpointing (latency critical, often cached via DRAM/NVMe); low-IO long-term storage for finalized models
Fine Tuning (less intensive training)	Smaller-scale data preparation, loading of pretrained models, lightweight checkpointing, and finalization for domain-specific adaptation	Smaller datasets (GB–TB); lower concurrency (1–64 GPUs); moderate read/write activity; reduced checkpoint size (GB-scale weights or adapters); overall lower IO intensity with shorter training cycles and long-term storage for results
Distillation (training for smaller model)	Teacher model generates soft targets and synthetic datasets, followed by student model training to emulate teacher behavior	Write-heavy generation of synthetic datasets; reuse of original or generated datasets for training; mixed read/write workloads with emphasis on large data generation and reuse
Reinforcement Learning	Iterative loop of experience collection, sampling/training, and checkpoint/state updates	Append-heavy writes for replay buffer generation (GB–TB); small random reads for sampling; large sequential writes for checkpointing; mixed IO patterns combining streaming, random, and sequential access
Inferencing	Prompt/context preparation, model deployment, KV cache management, and model execution	Latency-sensitive operations; prompt/context reads; model loading via async or batched IO; KV cache reads/writes with large IO sizes; execution dependent on fast cache access or recomputation
Agentic AI	Continuous loop of information collection (vector DB, external sources, context, events) and inference with decision/action cycles	Highly latency-sensitive; large volumes of small random reads; near real-time interactions; KV cache-driven inference; continuous, bursty access patterns

Storage Types & Protocols: Storage providing raw data to transformed data to semantic data



SNIA | DSN
A SNIA Community

Decision Frameworks

Selection Drivers: On-Prem vs. Cloud vs. Hybrid

Match the architecture to workload, data, and operating model

Driver	Lean On-Prem	Lean Cloud	Lean Hybrid
Data gravity	PB-scale data already on-site; high data egress costs	Data predominantly generated/consumed in cloud; minimal gravity to on-prem; egress awareness for large-scale datasets	Data sources mixed across edge, partners, cloud
Performance	Sustained high GPU utilization, predictable latency	Elastic performance; optimized for burst and scale-out; higher latency variability vs on-prem	Variable demand; bursty large training runs
Sovereignty	Strict residency, regulated or sensitive workloads	Cloud-native controls; region-based compliance; may require additional governance for strict sovereignty	Mixed sensitivity; isolate regulated subset on-prem
Cost profile	24x7 utilization; capital-friendly TCO	Fully OPEX model; pay-as-you-go; cost-efficient for bursty/variable workloads but requires cost control (egress, IOPS)	Spiky utilization; OPEX preferred for bursts
Operations	Mature on-site storage and network teams	Provider-managed infrastructure; minimal ops overhead; DevOps / platform teams drive usage	Shared with cloud / managed services teams
Time-to-capacity	Multi-quarter procurement acceptable	Instant elasticity; capacity available in minutes	Need elastic capacity in days / hours
Resilience	Cross-site DR within the data center estate	Built-in multi-region resilience; geo-redundancy and managed DR options	Cloud region or second site as DR target

Conclusion and Next Steps

Summary

Storage Strategy Alignment

- Emphasizing the importance of aligning storage strategies with specific AI workload requirements ensures optimal performance and efficiency.

Understanding Storage Concepts

- Understanding processing stages, storage types, protocols, and conceptual differences helps in making informed AI storage decisions.

Decision-Making Tools

- Use case matrix and decision flow charts provide practical frameworks for selecting appropriate AI storage environments.

Quick Poll

3. Do you feel better equipped and prepared to make storage architectural decisions now in the world of AI?

A. Yes

B. Somewhat

C. No

Join the StorageAI effort at SNIA to Solve AI Data Challenges

StorageAI™ is an open standards project for efficient data services related to AI workloads. See how industry leaders are combining forces to create an open ecosystem for efficient data services to address the most difficult challenges related to AI workloads.

Learn more at snia.org/storage.ai

SNIA | DSN
A SNIA Community

Q&A

After this Webinar

- Please rate this webinar and provide us with your feedback
- This webinar and a copy of the slides are available at the SNIA Educational Library snia.org/educational-library
- A Q&A from this webinar, including answers to questions we couldn't get to today, will be posted on our blog at sniablog.org
- Follow us on LinkedIn and X @SNIA

SNIA | DSN
A SNIA Community

Thank You