Agentic AI for Advanced Research

Storage Architecture, Data Pipelines, and Grant-Ready Operations

RMACC HPC Symposium 2026 · 90-minute technical session

Earl J. Dodd
Global Leader — HPC / Supercomputing
World Wide Technology
Earl.Dodd@WWT.com

Audience: Faculty, PIs, research computing leaders, HPC architects, storage teams, data stewards

Storage is the control plane for agentic research workflows

Technical thesis

Agentic workflows are dependency graphs over data, not single jobs over compute. They iterate, branch, retrieve, validate, checkpoint, and rerun. The dominant failure mode is workflow friction — metadata pressure, small-file churn, checkpoint bursts, vector pipeline drift, and governance gaps — not raw FLOPS.

Storage shapes velocity

GPU and CPU utilization is bounded by the data plane: locality, prefetch, cache residency, checkpoint path, and metadata throughput.

Storage shapes reproducibility

Snapshots, immutable copies, lineage, and policy-driven placement determine whether a result can be re-run, audited, or shared.

Storage shapes efficiency

Tiering, zero-copy, and unified namespaces remove duplicate datasets and idle compute waiting on transfers between silos.

Implication: Architect storage as a workflow control plane first; pick controllers, tiers, and cloud surfaces to fit the phases — not the other way around.

Workflow phase I/O signatures

What each phase does to storage

Phase	Dominant I/O pattern	What hurts performance / correctness
Ingest	Sequential writes; high file-create rate from instruments, repos, collaborators	Namespace hot spots, inode exhaustion, ingestion bottleneck on shared metadata
Curate / normalize	Mixed read/write; many small-file ops; tag and validate passes	Metadata storms; serialized rename/attr ops on directory trees
Index / embed	Random small reads; vector writes; embedding model fan-out	IOPS ceiling on shared tier; vector DB write amplification; pipeline drift
Train / fine-tune	Bursty reads of shards; periodic checkpoint writes (large, synchronous)	Checkpoint storms collapse fairness; data-loader I/O wait stalls GPUs
RAG / inference	Latency-sensitive small reads; KV-cache reuse; vector search	Tail latency on metadata + vector lookups directly hits user-perceived response time
Publish / rerun	Snapshots, exports, lineage capture, archival writes	Provenance gaps; manual copies break reproducibility and DMP commitments

Design rule: Map every phase to a storage behavior before choosing a platform. Phases with conflicting behaviors should not share a single tier.

Failure modes that break default storage designs

Metadata storms

Many open/stat/getattr operations from agent steps and orchestrators saturate metadata services before bandwidth ever bottlenecks.

Small-file churn

Curation, embedding, and tool-use pipelines create high-rate small writes that punish bulk-throughput-optimized tiers.

Random read I/O

Retrieval, indexing, and shuffle steps generate random small reads that defeat sequential prefetch heuristics.

Checkpoint bursts

Synchronous, multi-GB checkpoint writes from training collapse fairness on a shared tier and stall queued reads.

Vector pipeline drift

Embedding model versions, chunking strategies, and index parameters change; without lineage, RAG quality silently regresses.

Governance gaps

Copies multiply across stages; access controls and retention diverge from DMP/IRB commitments.

Diagnostic: If GPUs are under-utilized, instrument the data path before adding compute. The bottleneck is usually metadata, small-file rate, or checkpoint contention — not FLOPS.

Building the unified data plane

Block, file, and object across hybrid environments

Capabilities

- Unified storage OS spanning hybrid cloud; block, file, and object served.
- Manage, protect, and move data across on-prem and cloud environments under a single policy model.
- Built-in snapshots, replication, DR, business continuity, and backups.
- QoS, tiering, and policy-driven mobility, managed via a Console.

Why this matters for agentic AI

- One namespace and one policy plane reduce the number of copies an agent has to chase across stages.
- Snapshotting underpins reproducible data products and disaster recovery for research outputs.
- QoS isolates training, inference, and curation tenants on shared infrastructure.
- Tiering moves cold corpora and provenance archives to lower-cost media without breaking access paths.

Architecture role: Storage solutions like ONTAP provides the data-management substrate; specialized platforms (AFX for AI scale, EF-Series for extreme block I/O) extend it for specific phases.

Source: NetApp ONTAP

Protocol and workload fit

Matching access pattern to access path

Workload / phase	NFS	SMB	Block	Object	Cloud Volumes	SnapMirror
Ingest	**Primary**	**Primary**	—	**Primary**	Optional	Replicate in
Curate, normalize, tag	**Primary**	Optional	—	Optional	—	—
Vector index / embedding store	**Primary**	—	Optional	Optional	Optional	Mirror corpus
Training data shards	**Primary**	—	**Primary**	Optional	Burst	Cache
Checkpoint write tier	**Primary**	—	**Primary**	—	—	—
RAG retrieval / serving	**Primary**	—	Optional	Optional	Optional	—
Archive, lineage, publish	—	—	—	**Primary**	Tier-out	Replicate out

Read this table as design constraints: Each "Primary" entry is a hard requirement on protocol availability and behavior; "Optional" entries indicate paths the architecture should support without forcing them.

Source: NetApp ONTAP

Disaggregate for AI @ scale

Independent linear scaling of compute and storage

4 TB/s

Single-cluster throughput

1+ EB

FabricPool tiered capacity

128

Storage controllers (max)

52

Storage enclosures (max)

Disaggregated architecture

- Extends with independent scaling of storage controllers and storage capacity.
- Inherits enterprise data management, security, and protocol stack.
- Designed for data-intensive AI workloads and AI pipelines (training, fine-tuning, RAG, inference).
- Certified for NVIDIA DGX SuperPOD; aligned with NVIDIA AI Data Platform reference design.
- Pool tiers cold data to lower-cost media without breaking access paths.

Why this shape: Decoupling controller and capacity scale removes the classic AFA tradeoff: you can grow throughput without overbuying capacity, or grow capacity without stalling on controller IOPS.

Source: NetApp AFX, NetApp Q2 FY2026

AIDE pipeline: from ingest to retrievable knowledge

Integrate the AI Data Engine (ONTAP and AFX)

Step 1

Ingest

Files into ONTAP volumes via NFS/SMB; remote ONTAP syncs to local AFX via SnapMirror.

Step 2

Workspaces

Workspaces created in ONTAP System Manager; access controls bound to data scopes.

Step 3

Metadata + classify

Continuous metadata extraction; classification and governance with policy-driven guardrails.

Step 4

Collections

Collections + Vectors built from the metadata catalog; embeddings stored in a vector DB in place.

Step 5

Consume

Direct file/object access, semantic search, RAG endpoints, REST API.

Design property: Metadata is enriched in place — the corpus is not copied between systems to be indexed, classified, or made searchable.

Source: NetApp AIDE docs, NetApp & NVIDIA

Metadata fabric: keep semantic context next to the data

Avoid the duplicate-index, duplicate-corpus pattern

Common anti-pattern

- Copy corpus into a separate indexing service.
- Re-extract metadata; build a parallel catalog.
- Generate embeddings into a third store.
- Two or three divergent sets of access controls.
- Re-runs require re-copying; lineage is fragile.
- Storage cost ≈ 2–3× of the source corpus.

Metadata-fabric pattern

- Metadata is continuously extracted from source volumes.
- Single catalog and policy boundary; ACLs follow data.
- Embeddings stored in a vector DB attached to the data plane.
- Re-runs use snapshots and SnapMirror, not re-copies.
- Lineage and provenance are inherent, not bolt-on.
- Storage cost stays near 1× plus index overhead.

Test for any RAG or agentic deployment: Count the number of full-corpus copies the system maintains. If the answer is greater than one, the architecture has not yet adopted a metadata fabric.

Source: NetApp AIDE, NetApp & NVIDIA

Block tier for extreme I/O @ scale

All-flash NVMe arrays for HPC and AI scratch / parallel filesystems

EF80

5,000,000	110 GB/s	57 GB/s
IOPS	Read bandwidth	Write bandwidth

All-flash NVMe; sustained ultra-low latency under 100 µs

Protocols: NVMe/IB, NVMe/RoCE, NVMe/FC, FC

EF50

1,700,000	41 GB/s	30 GB/s
IOPS	Read bandwidth	Write bandwidth

All-flash NVMe; sustained ultra-low latency under 100 µs

Protocols: NVMe/IB, NVMe/RoCE, NVMe/FC, FC

HPC integration

EF-Series integrates with parallel filesystems including Lustre and BeeGFS, providing a high-IOPS, low-latency block tier behind the parallel namespace. Typical role: scratch, training data shards, and checkpoint write tier where tail latency and IOPS dominate over capacity.

> **When to choose EF-Series:** Phase needs sub-100 µs response time, very high IOPS, or parallel filesystem backing. ONTAP/AFX is not the right tier for these specific phases.

Source: NetApp EF-Series datasheet (PDF)

Storage phase-to-platform fit

One workflow, multiple specialized tiers under policy

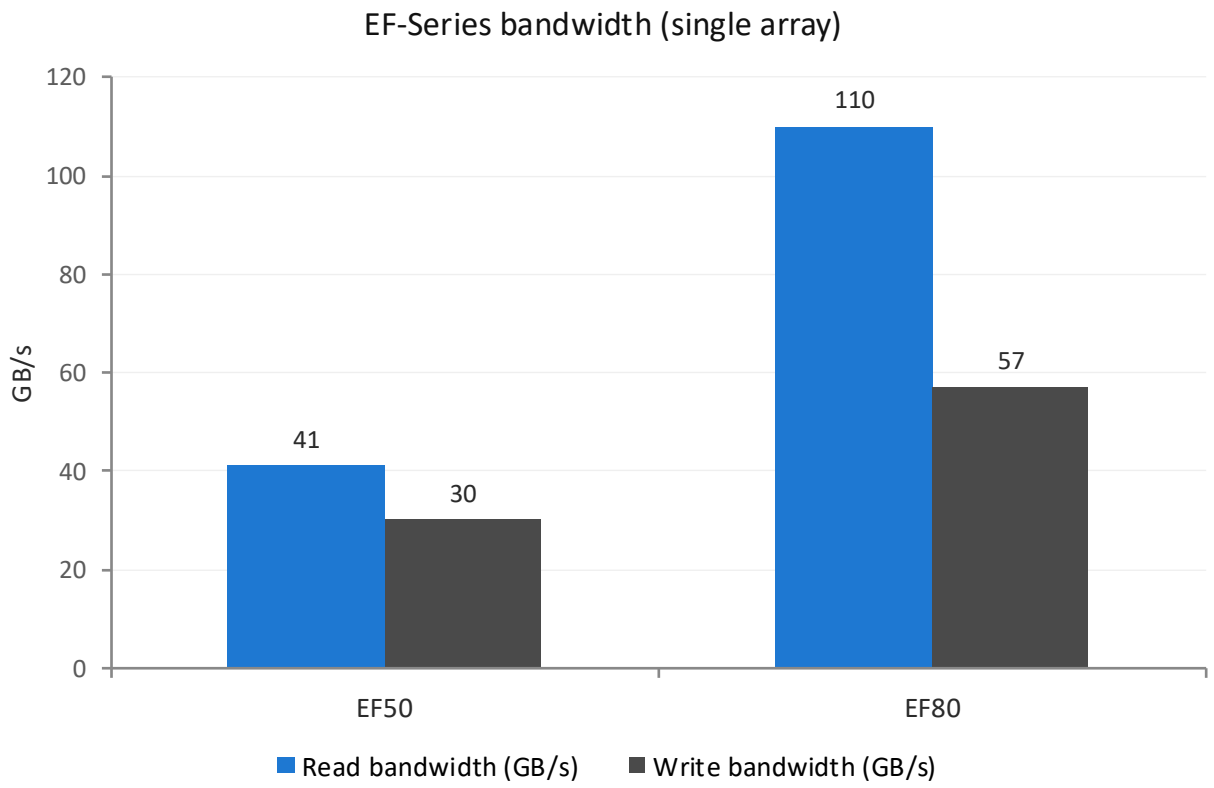
Workflow phase	ONTAP / AFF	AFX + AIDE	EF-Series	StorageGRID	Cloud Volumes
Ingest	**Primary**	Mirror in	—	Object landing	Cross-region
Curate / classify	**Primary**	**Primary**	—	—	Optional
Index / vectorize	Optional	**Primary**	—	—	Optional
Train / fine-tune	Optional	Hot data	Scratch / shards	—	Burst
Checkpoint	—	Optional	**Primary**	—	—
RAG / inference	Optional	**Primary**	—	—	Edge / region
Archive / publish	Optional	—	—	**Primary**	Tier-out

Operating principle: A single policy plane should span the row; the columns are deployment choices that follow the phase's I/O signature.

Source: NetApp ONTAP, NetApp AFX, NetApp EF-Series

AI/HPC storage performance envelope (illustrative)

EF80 and EF50 spec points; AFX cluster throughput ceiling



IOPS and cluster ceiling

5,000,000

EF80 IOPS

1,700,000

EF50 IOPS

4 TB/s

AFX single-cluster throughput ceiling

Disclaimer: values are vendor-published peaks. Size against sustained, protocol-specific figures from your own POCs.

Source: NetApp EF-Series datasheet, NetApp AFX

GPU feed model: where storage stalls compute

Locality, prefetch, caching, checkpoint path

Common stall points

- Data-loader read of training shards waiting on shared metadata service.
- First-touch cache miss for a new epoch on cold capacity tier.
- Checkpoint write fsync serialized across rank-0; collective stalls.
- KV-cache evicted under memory pressure; recompute cost on retrieval.
- Vector-search tail latency adds to per-token RAG response time.
- Cross-zone or cross-region read paths inserted by orchestration.

Design responses

- Place hot training shards on the scratch tier near the cluster.
- Use SnapMirror to land remote corpora on local AFX before training begins.
- Separate checkpoint tier (block, low-latency) from read tier; size for burst write.
- Offload KV-cache and vector handling to a dedicated memory tier (NVIDIA STX).
- Co-locate vector DB with source corpus through AIDE; avoid cross-system hops.
- Pin orchestration so data-loader workers run in the same fabric as the storage.

Instrument before architecting: Measure GPU idle time and attribute it to a specific stall point. Architecture changes that do not target an observed stall will not improve utilization.

NVIDIA collaboration: AIDE, AI Data Platform, and STX

Co-engineered storage and acceleration for agentic AI

AIDE × AI Data Platform

AIDE is co-engineered with NVIDIA and integrated with the NVIDIA AI Data Platform reference design, with a global metadata catalog and semantically enriched metadata kept in place.

STX reference architecture

Modular rack-scale storage reference architecture for agentic AI built with NVIDIA Vera Rubin and BlueField-4 DPUs; uses Spectrum-X networking; centralizes intelligent data handling.

KV-cache and vector offload

STX includes a specialized memory tier for KV-cache storage; BlueField-4 offloads KV-cache and vector tasks; optimized for RAG and agentic AI workflows. AIDE will support STX.

Why this matters for research: For RAG and agentic inference, KV-cache and vector handling dominate the data path. Pushing those tasks to a memory tier and DPU offload removes them from the critical inference path.

Source: NetApp & NVIDIA press release, NVIDIA STX

Cloud collaboration: hybrid and air-gapped paths

Cloud Volumes Flex Unified and Google Distributed Cloud

Google Cloud NetApp Volumes Flex Unified

- Generally available; single storage pool for file and block in all Google Cloud regions.
- Run enterprise applications, databases, HPC, EDA, VMware, and AI workloads without changing applications.
- Data is consumable by Google Cloud AI services without moving or duplicating it again.
- Block storage capability added in Q2 FY2026 — relevant for database and structured-data workflows.

GDC air-gapped (delivered by WWT)

Google Distributed Cloud

- Expanded NetApp + Google Cloud collaboration for sovereign / private AI deployments.
- NetApp AFF, StorageGRID, and Trident enable zero-trust security, local data residency, and encryption key management.
- Supports AI workloads, including Gemini capabilities, in disconnected environments.
- Use case: research environments with regulatory, IRB, or export-control constraints that prevent public-cloud egress.

Architecture choice: Choose the cloud surface by data residency and connectivity constraints, not by feature checklist. ONTAP semantics carry across both surfaces.

Source: GC NetApp Volumes Flex Unified, NetApp + GDC air-gapped

Adoption signal: Q2 FY2026

What the install base is buying for AI

~200

AI / data-lake modernization deals in the quarter

$1.71B

Q2 revenue, up 3% YoY

3

Recurring use-case clusters: data prep, training, RAG / inference

Technical reading

- AFX positioned as ultra-scalable, extreme-performance disaggregated storage; certified for NVIDIA SuperPOD.
- AIDE is positioned as an end-to-end AI data service integrated into ONTAP — discovery, curation, policy-driven guardrails, and real-time vectorization.
- Zero-copy caching and native cloud connectivity are emphasized to unify data across sites, clouds, and models.
- Google Cloud NetApp Volumes added block storage in Q2 — relevant for structured data and database workloads alongside AI.

Why surface this in a technical talk: The deal mix — data-lake modernization, training/fine-tuning, RAG/inference — is a useful map of which research workloads are actually being deployed at scale today.

Source: NetApp Q2 FY2026 earnings call (2025-11-25)

Measurement plan

What to instrument before architecture changes

Per-phase metrics

- Ingest: file-create rate, bytes/sec, inode growth, ingestion lag.
- Curate: ops/sec by op type (open, stat, getattr, setxattr, rename), small-file size distribution.
- Index/embed: vector write rate, embedding model version, chunking parameters, drift signals.
- Train: data-loader wait fraction, GPU idle attribution, checkpoint write latency p50/p99.
- RAG / inference: vector-search p99, KV-cache hit rate, end-to-end token latency.
- Publish / rerun: snapshot/replication SLA hit rate, lineage capture coverage.

Cross-cutting controls

- GPU idle attribution by cause: data-loader, sync, orchestration, network.
- Number of full-corpus copies maintained across systems.
- Metadata-op p99 vs. read/write throughput under multi-tenant load.
- QoS policy hit/miss rates per tenant per phase.
- Snapshot, SnapMirror, and immutable-copy SLAs vs. commitments in DMP/DMS.
- Cost per stage: storage, transfer, idle compute attributable to I/O wait.

Operating rule: Adopt one observability backbone across phases. If each phase has its own metrics surface, attribution becomes guesswork.

Security and provenance model

Reproducible, auditable, ransomware-aware

Snapshots and immutable copies

Point-in-time recovery for datasets and workspaces; immutable / WORM copies underpin reproducible data products and audit obligations.

Ransomware detection and response

Behavioral anomaly detection on volumes; rapid restore from snapshots is the primary recovery path.

Access controls and policy guardrails

ACLs, RBAC, and AIDE policy-driven guardrails on classification, retention, and use; ACLs follow the data through SnapMirror and tiering.

Lineage and provenance

Metadata catalog and snapshot history capture which data, which transformations, which model versions produced which outputs.

Reproducibility test: For any published result, you should be able to identify the snapshot, dataset version, model version, and policy state used to produce it.

Source: NetApp ONTAP, NetApp AIDE

Grant-ready architecture mapping

DMP / DMS language to concrete storage controls

Grant / DMP requirement	Storage or data-management control	Where it lives
Data preservation and access for the project lifetime + retention period	Snapshots, replication, immutable copies; tiered archive	ONTAP, StorageGRID, FabricPool tier
Reproducibility of analyses and reruns	Versioned datasets, lineage capture, workspace snapshots	ONTAP snapshots, AIDE workspaces and catalog
Controlled access (PHI, IRB, export-controlled)	RBAC, encryption at rest and in flight, key management; air-gapped option	ONTAP, GDC air-gapped + StorageGRID + Trident
Sharing with collaborators / repositories	Object access, time-bounded credentials, scoped exports	StorageGRID, ONTAP S3, SnapMirror
Cost transparency for budgets / NSF / NIH reports	Per-tenant QoS, tiering and tier-cost reporting	ONTAP QoS, console, FabricPool reporting
Disaster recovery and business continuity	SnapMirror, cross-site / cross-region replication	ONTAP, Cloud Volumes regions

Mapping rule: Every DMP / DMS sentence should resolve to a specific control on a specific platform — not a paragraph of intent.

Source: NetApp ONTAP, NetApp AIDE

Decision matrix and checklist

Use to evaluate architecture proposals

Phase need	First-choice platform	Trigger to add
Bulk file/object data plane, snapshots, replication	ONTAP / AFF	Always
AI ingest, metadata, classification, vectorization	AFX + AIDE	In-place RAG, training corpora, governed data products
Extreme IOPS / sub-100 µs / parallel FS scratch	EF-Series	Lustre/BeeGFS scratch, checkpoint bursts
Object archive, publish, lineage tier	StorageGRID	Long-tail retention, DMP archive obligations
Hybrid burst, cloud-side AI consumption	Cloud Volumes Flex Unified	Cross-region collaborators, Google Cloud AI
Sovereign / air-gapped AI	GDC air-gapped + AFF + StorageGRID + Trident	Residency, export control, IRB

Architecture review checklist

- Each phase mapped to a tier with documented I/O signature.
- Single ONTAP policy plane covers all tiers.
- GPU idle attribution instrumented end-to-end.
- Full-corpus copy count = 1 (plus index overhead).
- Snapshots, SnapMirror, immutable copies sized to DMP commitments.
- Air-gapped or sovereign path tested, not just designed.
- One owner per dataset on storage side and DMP side.
- KV-cache and vector path planned (STX or equivalent).

Use this matrix as a gate: Any architecture proposal that does not satisfy the right-hand checklist is not yet ready for procurement.

Source: NetApp ONTAP, NetApp AFX, NetApp EF-Series

Operating model

Who owns what across the storage and data plane

Storage administrators

Owns the deployment, capacity, performance, snapshot/replication SLAs, QoS policy enforcement, observability backbone.

Data engineers / data wranglers

Own AIDE pipelines, metadata schemas, classification rules, vector indexing parameters, dataset versioning and lineage capture.

Research computing / HPC team

Own scheduler integration, data staging, GPU feed, parallel filesystem health, cross-tier movement and tiering policies.

Faculty / PIs / data stewards

Own DMP commitments, IRB/regulatory scope, dataset definitions, retention and sharing decisions, validation of reproducibility for publications.

Boundary rule: Every stored dataset has exactly one DMP owner and one storage owner. Ambiguity is the leading cause of governance gaps.

Source: NetApp ONTAP

Technical takeaways

1 **Treat storage as a workflow control plane**

Map every phase to an I/O signature and a tier before choosing platforms.

2 **Unify policy, specialize tiers**

ONTAP carries policy across AFX, EF-Series, StorageGRID, and Cloud Volumes.

3 **Keep metadata next to data**

AIDE-style in-place metadata and embedding avoid duplicate-corpus architectures.

4 **Decouple controllers from capacity**

AFX scales throughput and capacity independently; EF-Series serves extreme block I/O.

5 **Push KV-cache and vectors off the GPU path**

NVIDIA STX with BlueField-4 offload reduces RAG and agentic tail latency.

6 **Match the cloud surface to constraints**

Cloud Volumes Flex Unified for hybrid; GDC air-gapped for sovereign / private AI.

7 **Instrument before architecting**

GPU idle attribution and full-corpus copy count are the two most useful starting metrics.

8 **Make DMPs executable**

Map every DMP sentence to a specific control on a specific platform.

One-line summary: Architect the data plane to fit the workflow shape; then let the workflow run faster, cheaper, and reproducibly.

Thank You

Remember, “Data at rest is a Cost. Data in motion is a Currency”