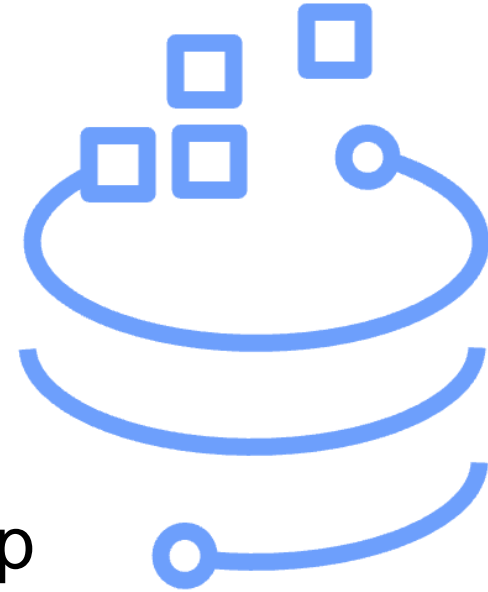


Data Acceleration Tier (DAT)

Aug 7, 2025

Storage Scale West Coast User Group

Storage Scale



John Lewars (Storage Scale Performance Architect)

John Divirgilio (Storage Scale Performance)

Olaf Weiser (Storage Scale Performance)



IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice and at IBM's sole discretion.



Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.



The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.



The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.



Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.

Data Acceleration Tier (DAT) based on Asymmetric Data Replication*

Accelerates AI and Analytics by allowing data to be stored in two pools at once, enabling concurrent use of:

a Storage Scale RAID (formerly known as GNR) encoded **reliable** copy (replica)

AND

a **performance** copy (replica) which can be:

- (a) a replica physically local to the compute client
- or -
- (b) a replica stored on an NVMeoF drive

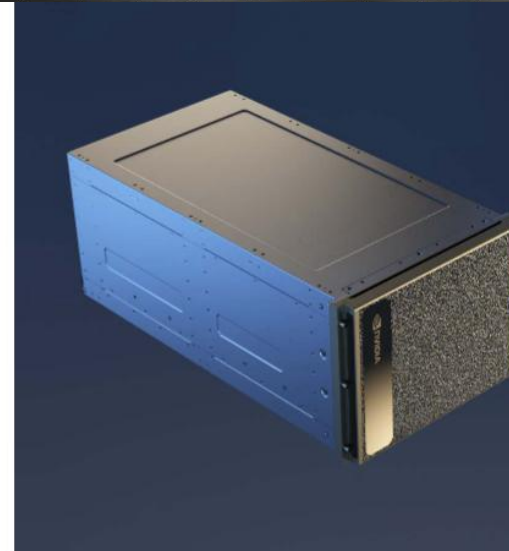
This provides high performance without compromising reliability. IOPs are significantly improved over existing offerings.

In the Fall 2025 release, IBM intends to support the NVMeoF-based and local-storage offerings (note these offerings may not be configured at the same time).

NVIDIA DGX SuperPOD with IBM ESS/ISS



* Asymmetric Data Replication has been referred to in past presentations by its IBM internal name, 'Ustore'.



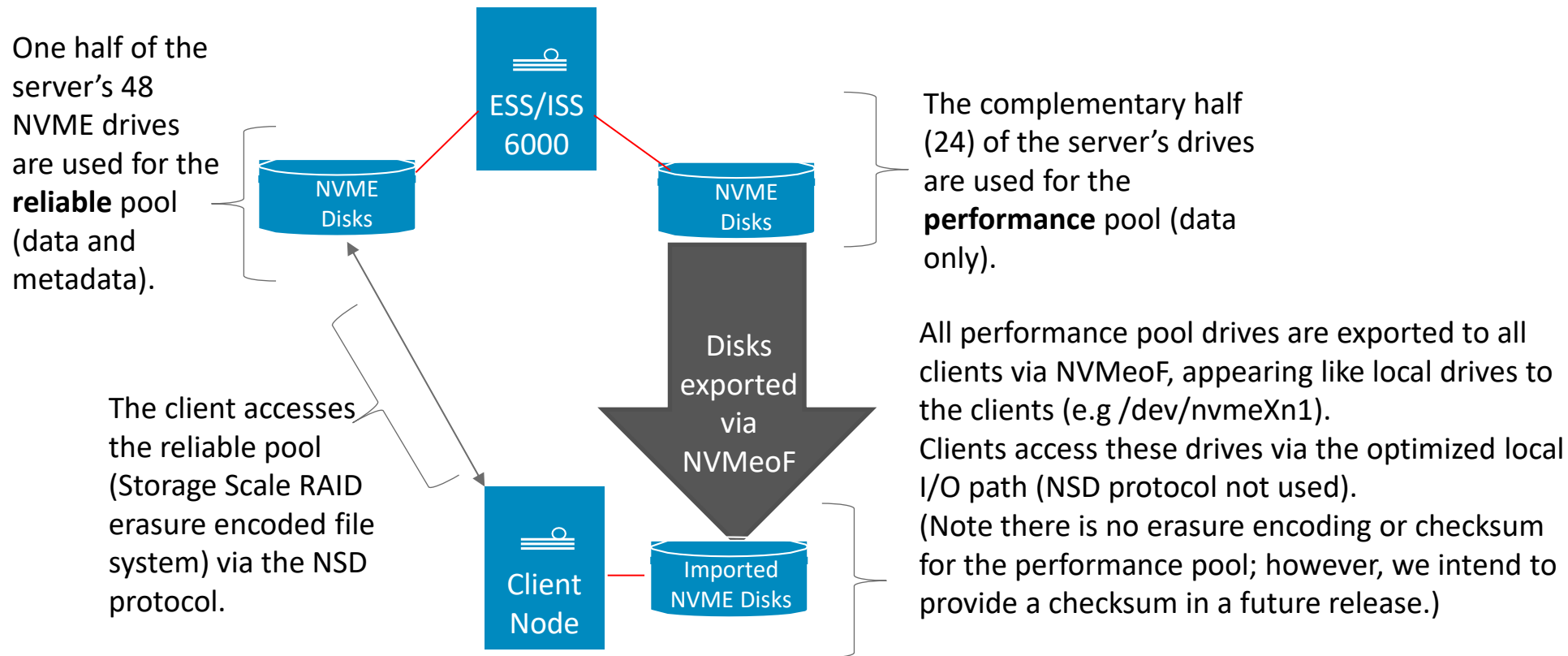
Specifications

GPU	8x NVIDIA H100 Tensor Core GPUs
GPU memory	640GB total
Performance	32 petaFLOPS FP8
NVIDIA® NVSwitch™	4x

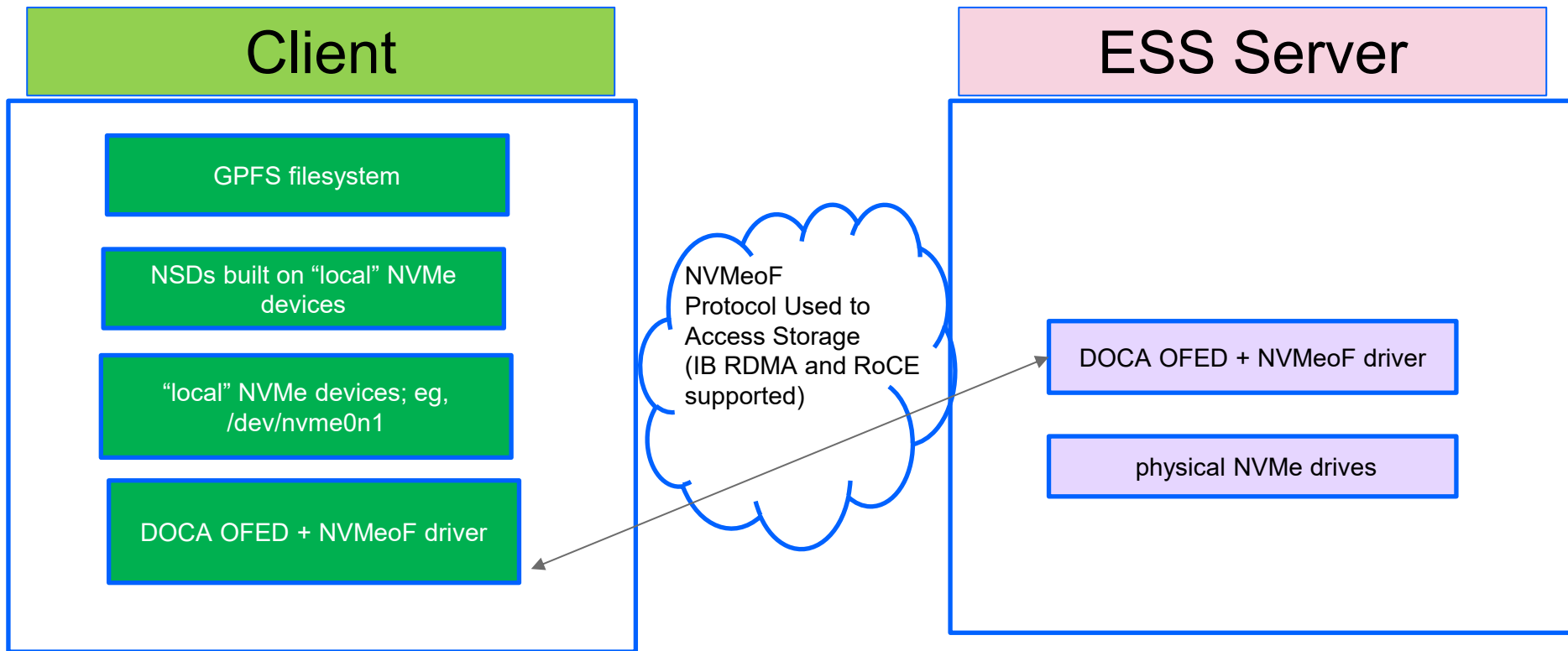
DAT / Asymmetric Data Replication Concepts

- GPFS automatically maintains the consistency of the reliable and performance replicas, just like it does for multiple data replicas.
- DAT filesystem reads are from the performance replica for optimal performance.
- Data and metadata are stored in the reliable pool but the performance pool is for data only.
- The local storage offering provides policy and mm command features to get the performance replicas affinitized on the clients that read the data.
- DAT filesystem writes are concurrent to the reliable and performance replicas, and control is returned to the user after the reliable replica has been updated. In a future release, we intend to support a mode where control is returned after the performance replica is updated. The reliable replica (if one) can be updated asynchronously; for instance, after a file is closed.

Asymmetric Data Replication File System use case: DAT exploiting shared storage via **NVMeoF** for the performance pool



Using NVMeoF for the DAT Performance Pool



The Data Acceleration Tier uses the NVMeoF protocol to "export" NVMe drives from the servers to the clients. All clients see all of the server's performance pool drives.

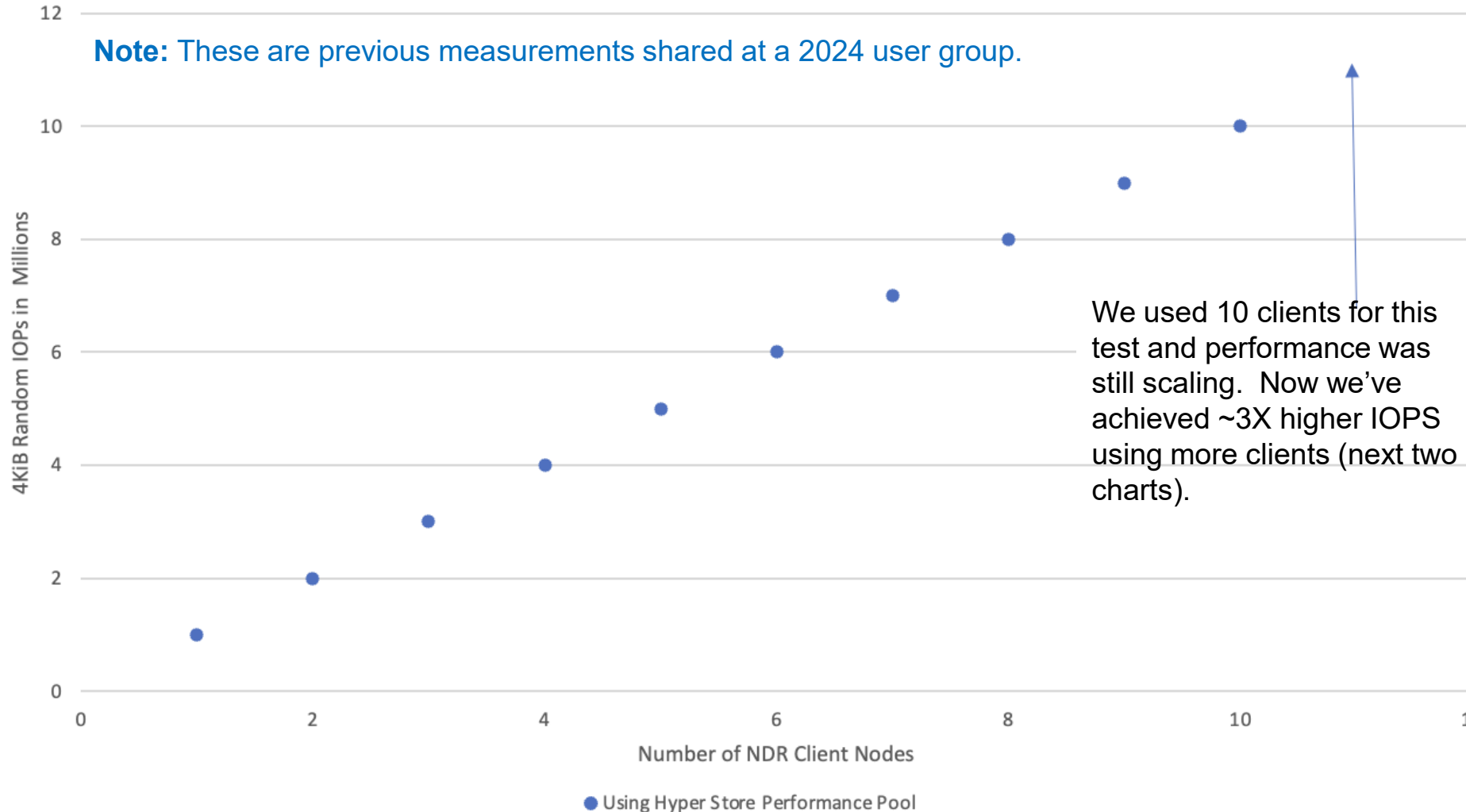
Configuration Details for Performance Measurements

InfiniBand Connected ESS/ISS 6000 Configuration Evaluated

NVMe Drives	48x 3.84TB NVMe Tier-1 Flash (Gen 4)
Memory per canister	1536 GB (24x 64GB dimms)
IB adapters per canister	4x CX7 adapters, each with 2x 200Gb ports
NVMeoF Offload	NVMeoF Hardware Offload Enabled
Clients	10x clients each with 2 cpu chips/48 cores and 2x CX7 400Gb/s IB 1-port adapters
Switch	NDR 400Gib MQM9700 - Infiniband

4 KiB Random Read IOPS on 10 NDR clients and 1 SSS 6000 Default Config compared to Data Acceleration Tier

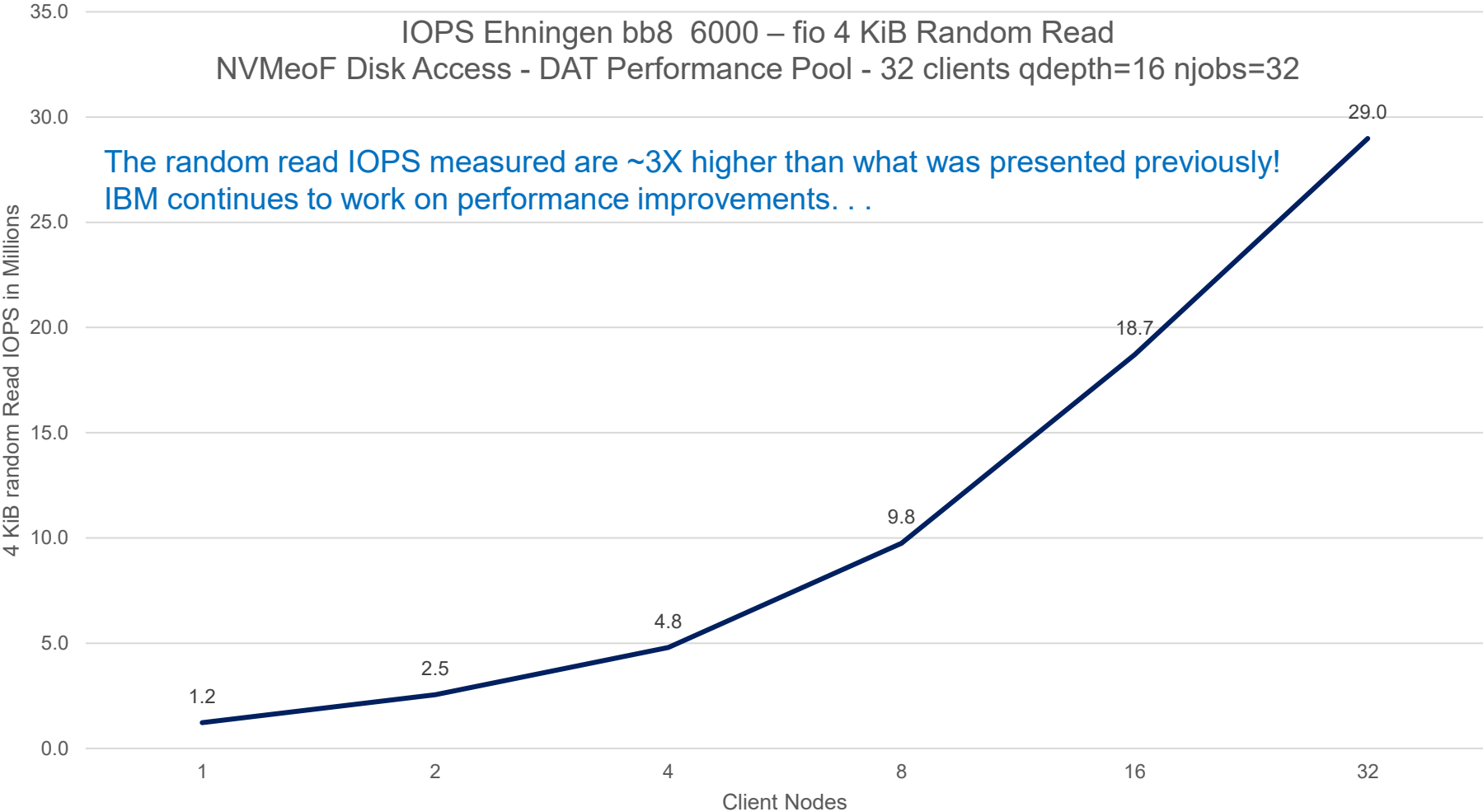
Note: These are previous measurements shared at a 2024 user group.



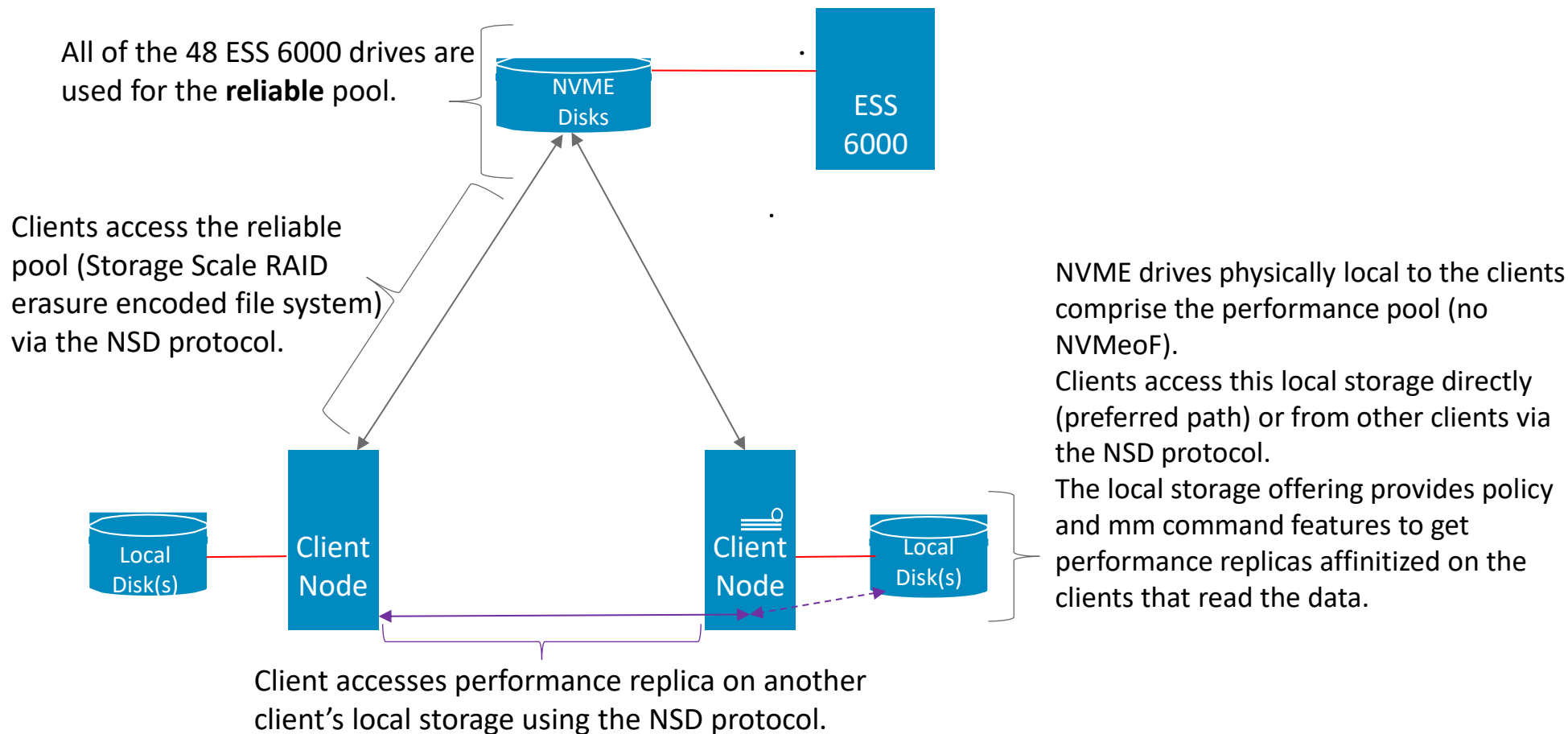
Configuration Details for Performance Measurements

Ethernet Connected ISS 6000 Configuration Evaluated	
NVMe Drives	48x 3.84TB NVMe Tier-1 Flash(Gen 5)
Memory per canister	768GB (24x 32GB dimms)
IB adapters per canister	4x CX7 adapters, each with 2x 200Gb ports
NVMeoF Offload	NVMeoF Hardware Offload Not Enabled
Clients	32x clients each with 1 cpu chip/24 cores (AMD EPYC 9274F) 400 Gb/s CX7 RoCE (half nodes have 1 port adapters and the other half use 2 port adapters)
Switch	400Gib SN 5600 - Ethernet

DAT 4 KiB Rand. Read with NVMeoF (RoCE) solution: **29 million IOPs at 32 nodes**

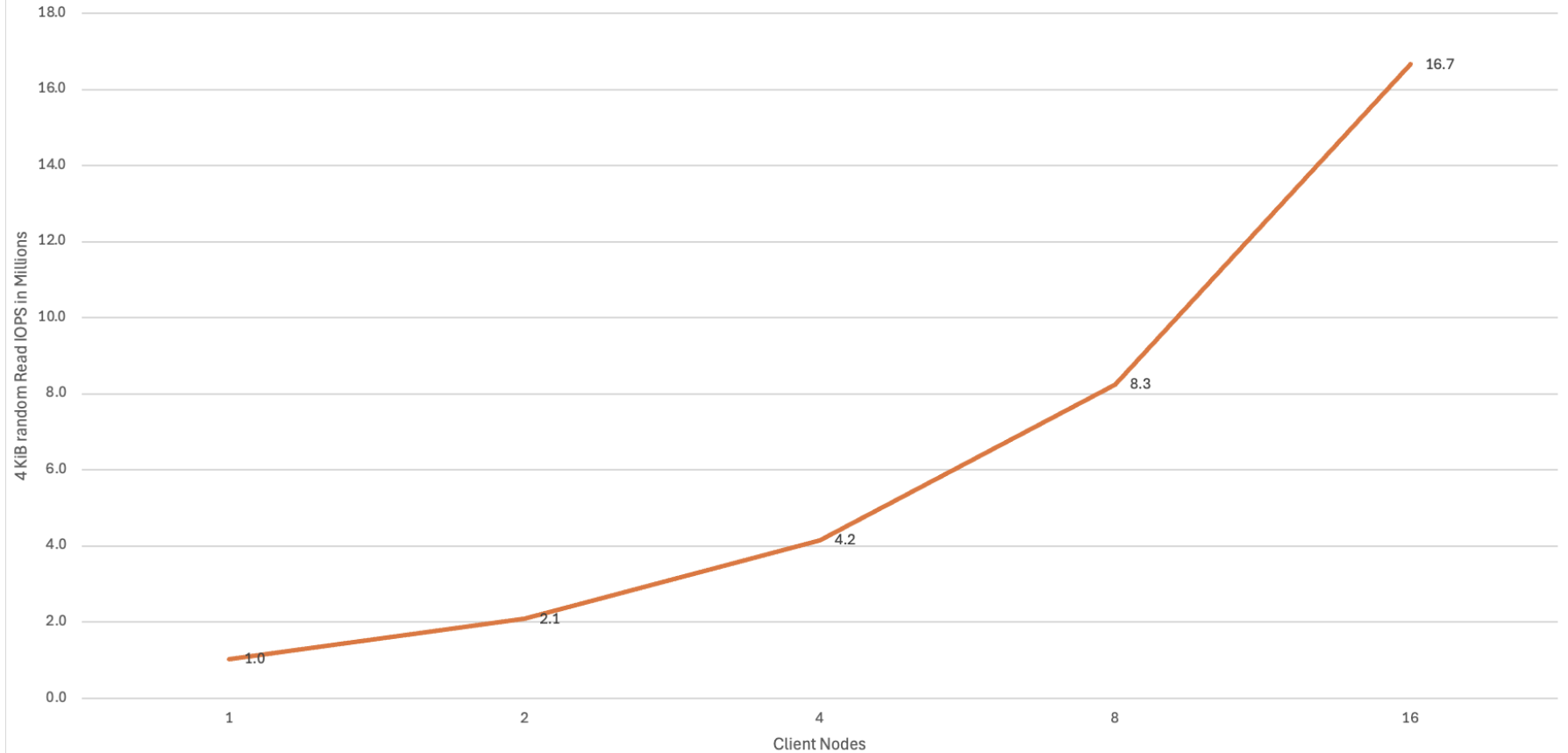


Asymmetric Data Replication Use Case: Exploiting **client-local storage** for performance pool



4 KiB Rand Read IOPS with Local Storage Solution: **16 million IOPs at 16 nodes**

IOPS Ehningen bb8 ISS 6000 + 2 NVME drives per client - 4 KiB Random Read IOPS (**Millions**)
Local Drive Variant of DAT - 32 clients qdepth=256 njobs=32 - Scaling Up Number of Clients



Asymmetric Data Replication Value Proposition: Exploiting local storage

Local storage configurations (particularly NVMe drives in client nodes) have become more common and IBM Storage Scale provides multiple solutions that leverage local storage:

Existing solutions that use local storage to create a filesystem:

- SNC (Shared-nothing cluster)
- FPO (File Placement Optimizer: SNC with write-affinity)
- ECE (Erasure Code Edition)
- DAT (Data Acceleration Tier)

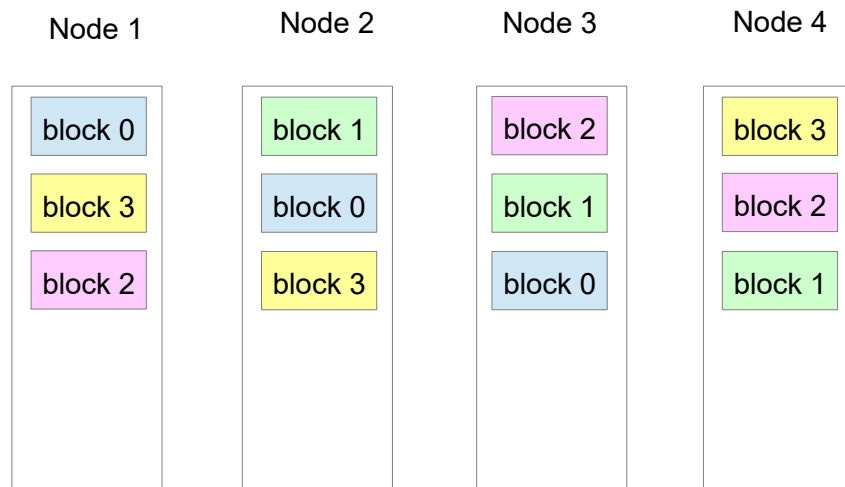
Existing solution that uses local storage to cache data for existing file systems:

- LROC (Local Read-Only Cache)

Using Asymmetric Data Replication, DAT combines key features of these existing local storage solutions to create a filesystem and cache data in a persistent, non-redundant manner.

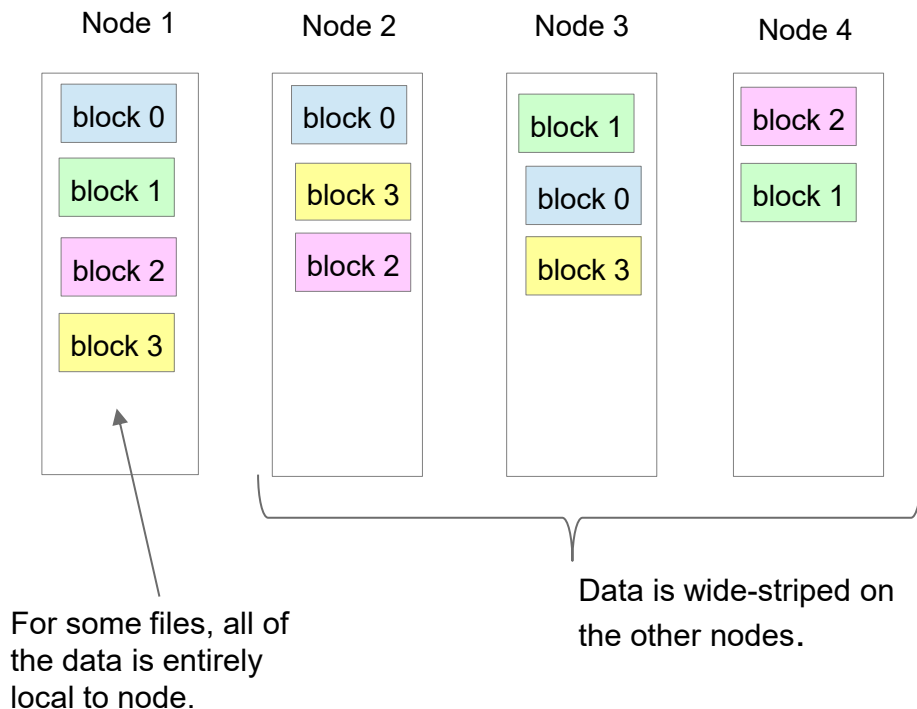
Workloads on local storage with SNC (Shared Nothing Cluster)

- Data Access: Workloads that use data-aware schedulers (e.g., Hadoop) can take advantage of locally stored data.
- Data Reliability/Availability: 3-way replication. Metadata scan needed after a failed node rejoins the cluster.
- Cluster Resources: Each unit of user-consumable storage uses **3x** units of storage for replication overhead.



Workloads on local storage with FPO (File Placement Optimizer)

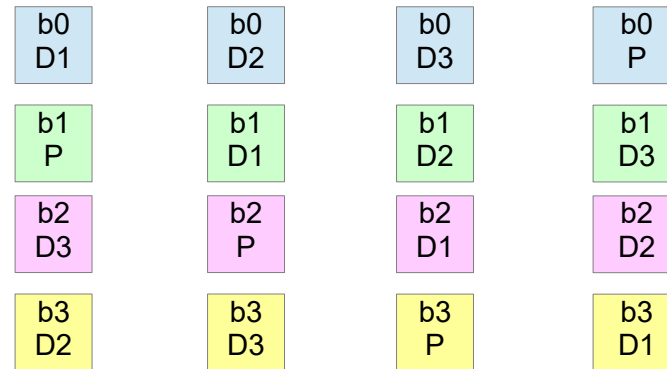
- Data access: Provides a method to affinitize data so a wider range of applications can exploit data locality; e.g., databases can ensure that one replica of a target fileset is entirely stored on a given node.
- Data reliability/availability: 3-way replication. Metadata scan needed after a failed node rejoins the cluster.
- Cluster resources: Each unit of user-consumable storage uses **3x** units of storage for replication overhead.



Workloads with ECE (Erasure Code Edition) on Storage-Rich Servers

Employs Storage Scale RAID software across local storage in the cluster. Each data block is split into N data strips and M parity strips, and these strips are spread across nodes.

- Data Access: Most data not local and there is no control over data placement (data chunks too small for schedulers like Hadoop to manage). A read typically requires gathering strips from other ECE servers (multiple network hops).



- Data Reliability/Availability: Uses erasure encoding and parity strips to provide data reliability: more efficient than using traditional replication.
- Cluster Resources: Each unit of user-consumable storage uses **1.25x** of storage resources for parity overhead (assuming 8+2P).

Workloads on local storage with an Asymmetric Data Replication File System

Maintain two copies (replicas) of each data block in two different pools with asymmetric replication.

- “reliable copy”: protected by ECE erasure encoding
- “performance copy”: stored entirely (as one unit) on a single node, providing fast reads for workloads that leverage data locality

The performance copy is a persistent, non-redundant cache.

Data Access: The performance copy of data is local to the client (physically present or via NVMeoF).

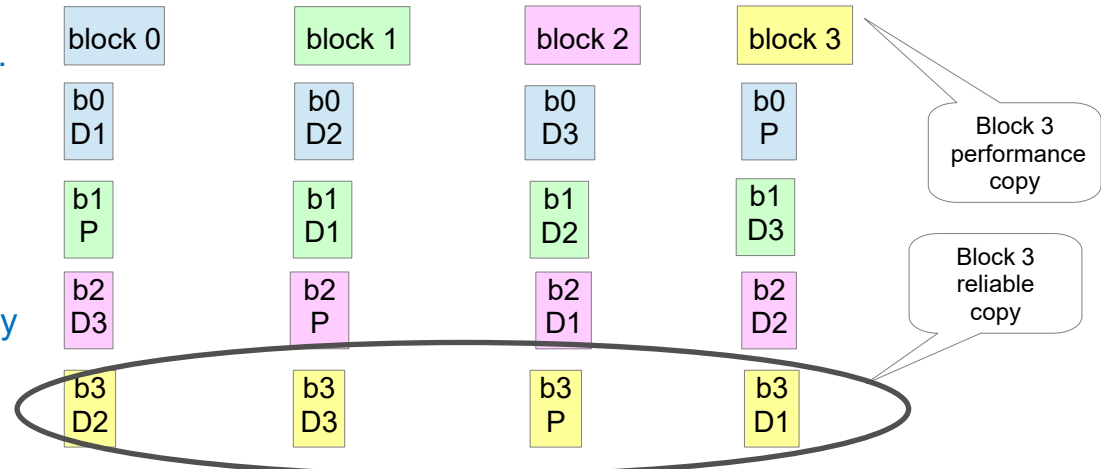
Data Reliability/Availability:

The reliable copy is erasure-encoded. The performance copy is maintained with a feature called “lazy consistency”. Missed updates (eg, performance disk offline) will be corrected on the next read after the disk is back online.

A full metadata scan is not needed when a performance disk is brought back online (full metadata scans can take some time for large filesystems).

Also, when a missing update is repaired, all missed updates in the same indirect block are fixed using an asynchronous thread.

Cluster resources: Each unit of user-consumable storage uses **2.25x** of storage resources, lower than 3x FPO.



Workloads on local storage with LROC

LROC uses local storage (typically SSD/NVMe) to support caching.

Asymmetric Data Replication/DAT has advantages over LROC:

- LROC cache is potentially redundant because a data replica can be cached on multiple nodes. (No more than a single performance replica will exist in a DAT filesystem.)
- LROC is not persistent because a reload is needed after a failed node is back up. With DAT, the performance copy is persistent except in the case of missed updates (the reliable copy is used until the performance disk is back up and missed updates are repaired).
- LROC is a true cache which means it won't help Direct I/O workloads. DAT provides a high-performance tier that can accelerate Direct I/O workloads.

Upgrading an existing filesystem to a DAT filesystem using the local storage solution.

1. Install latest Scale/ESS software on clients and 6000 (Fall 2025 release)
2. “mmchconfig release=LATEST” (upgrades daemon version)
3. “mmchfs -V full” (upgrades filesystem format version)
4. Add performance pool using mm* and mmvdisk commands

In the Fall 2025 release, we will provide enhancements to the commands for the creation of a DAT filesystem.

Supported configurations for the Fall 2025 release

There is a 1-to-1 mapping between a DAT filesystem and an ESS 6000.

- An ESS 6000 supports one DAT filesystem.
- A DAT filesystem may not use more than one ESS 6000. **Note:** We intend to remove this restriction in a future release.

For optimal performance, DAT filesystems use the fully-configured ESS 6000 performance model with NVMe drives (eg, the hybrid model is not supported).

DAT filesystems do not support a remote storage cluster. **Note:** We intend to remove this restriction in a future release.

Thanks!