Seagate DNE Architecture and Performance

John Fragalla
HPC Principal Engineer

SEAGATE

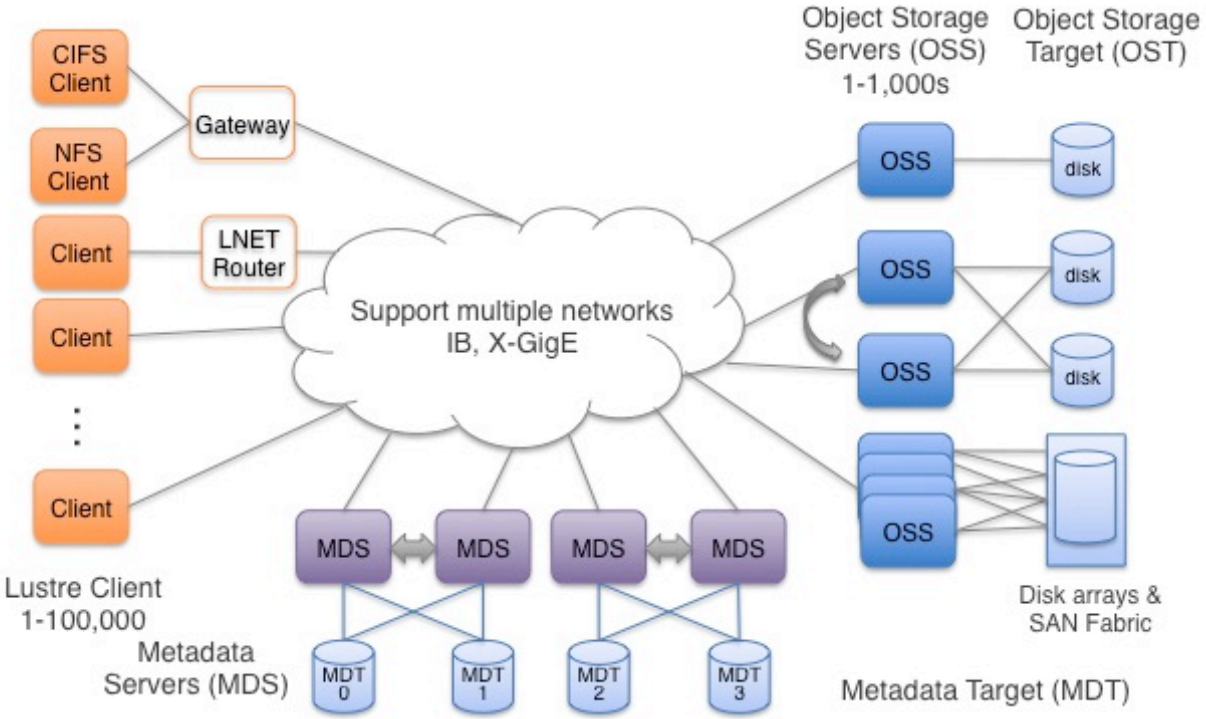
Agenda

- DNE Phase 1
- DNE Phase 2
- Seagate MDT Architecture
- MDT Performance Results
- DNE Use Cases

Lustre Parallel File System

Lustre is an open source, distributed parallel file system

- Object-based design provides extreme scalability
- Compute clients interact directly with storage servers
- Comprised of:
  - Clients
  - Metadata Servers and Targets
  - Storage Servers and Targets

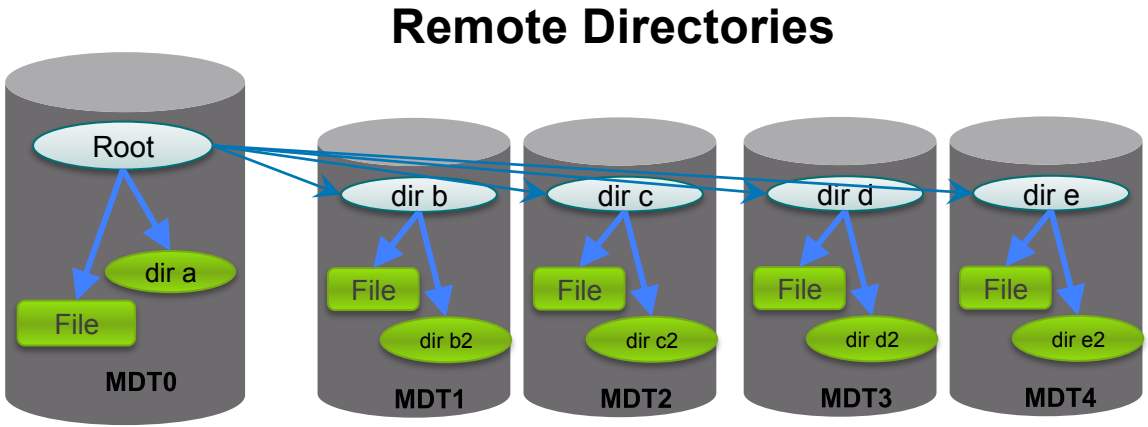


Lustre Distributed NamespacE (DNE)

- Distributed NamespacE (DNE) is a new feature available in Lustre 2.5 that allows multiple MDS / MDT components to participate in a single file system.
- DNE allows the namespace to be divided across multiple metadata servers.
- Enables the size of the namespace and metadata throughput to be scaled with the number of servers.
- The Lustre DNE project is comprised of 2 phases.

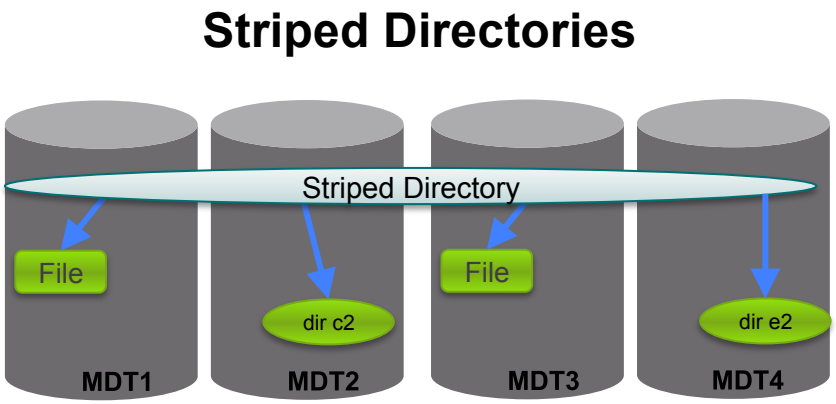
Phase 1, Lustre 2.5 Release

Remote Directories -- Lustre sub-directories are distributed over multiple metadata targets (MDTs). Sub-directory distribution is defined by an administrator.



Phase 2, Future Lustre Release

Striped Directories -- The contents of a given directory are distributed over multiple MDTs.



Seagate System Management Unit: Management and Metadata (MDS/MDT)

CSM Manager and MDS/MGS Nodes

- 2RU 4-node Sandy Bridge Servers
  - Server 1: CSM Mgmt
  - Server 2: Boot
  - Server 3: MGS
  - Server 4: MDS

Fault Tolerance (active/passive)

Serviceability

2U24 EBOD – MDT

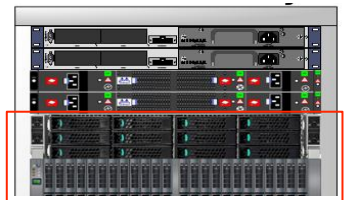
- SAS EBOD for MDS/MGS/ Management
- Disk Configuration
  - Qty 4 Lustre Management (MGS)
  - Qty 4 System Management and NFS
  - Qty 2 Global Hot spares
  - **Qty 14 Drives for MDT**

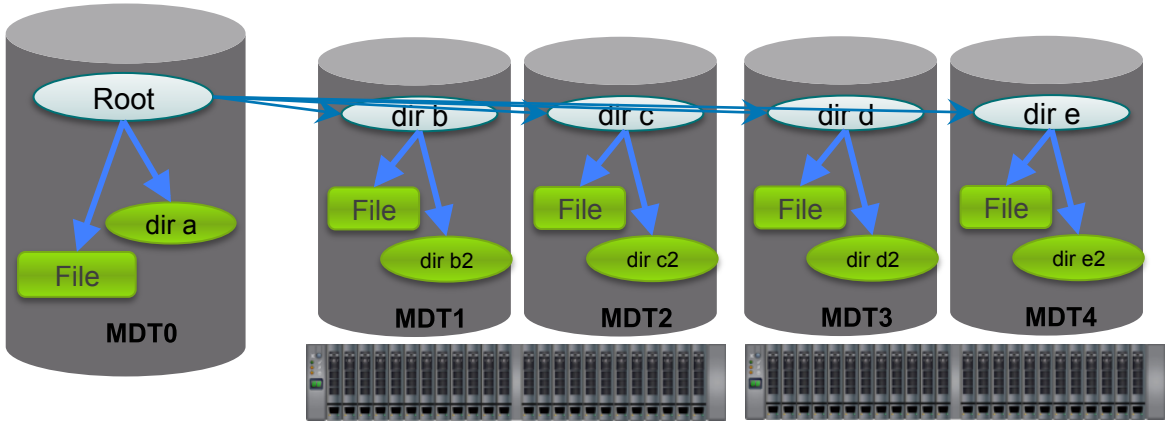
Auxiliary DNE Unit (ADU)

- 2U24 MDS/MDS node pair + storage
  - Dual CS9000 ESM(Sati)s in 2U24
  - Dual MDS diskless node pairs
  - Normal operation 1 MDT per MDS/ESM
  - Failover 2 MDTs on remaining MDS/ESM
  - ESM canister SSDs uses for persistent state information
- HDD configuration
  - 20 HDDs comprise 2 x RAID10 (5+5) MDTs
  - 2 HDDs comprise 2 x Hot Spares
  - 2 Blank Trays

Seagate & Lustre 2.5 DNE Hardware

- DNE is available in **Software Release v2.0**
  - MDT0 is master and default in DNE environment
- DNE Servers are configured in active / active pairs
  - Seagate 2U24 with 2 MDS embedded server modules
- Scale Metadata Capacity / Performance with DNE Server pairs, up to 16 ADU Unit (16 DNE Unit)





Lustre DNE Phase 1 Directory Management

lfs allows you to assign mdt to directory.

lfs mkdir -i <mdtindex> <remotedir>

- An administrator can allocate remote sub-directories to separate MDTs.
- Creating remote sub-directories in parent directories not hosted on MDT0 is not recommended.
- This is because the failure of the parent MDT will leave the namespace below it inaccessible.
- For this reason, by default it is only possible to create remote sub-directories off MDT0.
- To relax this restriction and enable remote sub-directories off any MDT, an administrator must issue the command:
  - lctl set_param mdd.*.enable_remote_dir=1.

Seagate MDT System Configuration and Setup

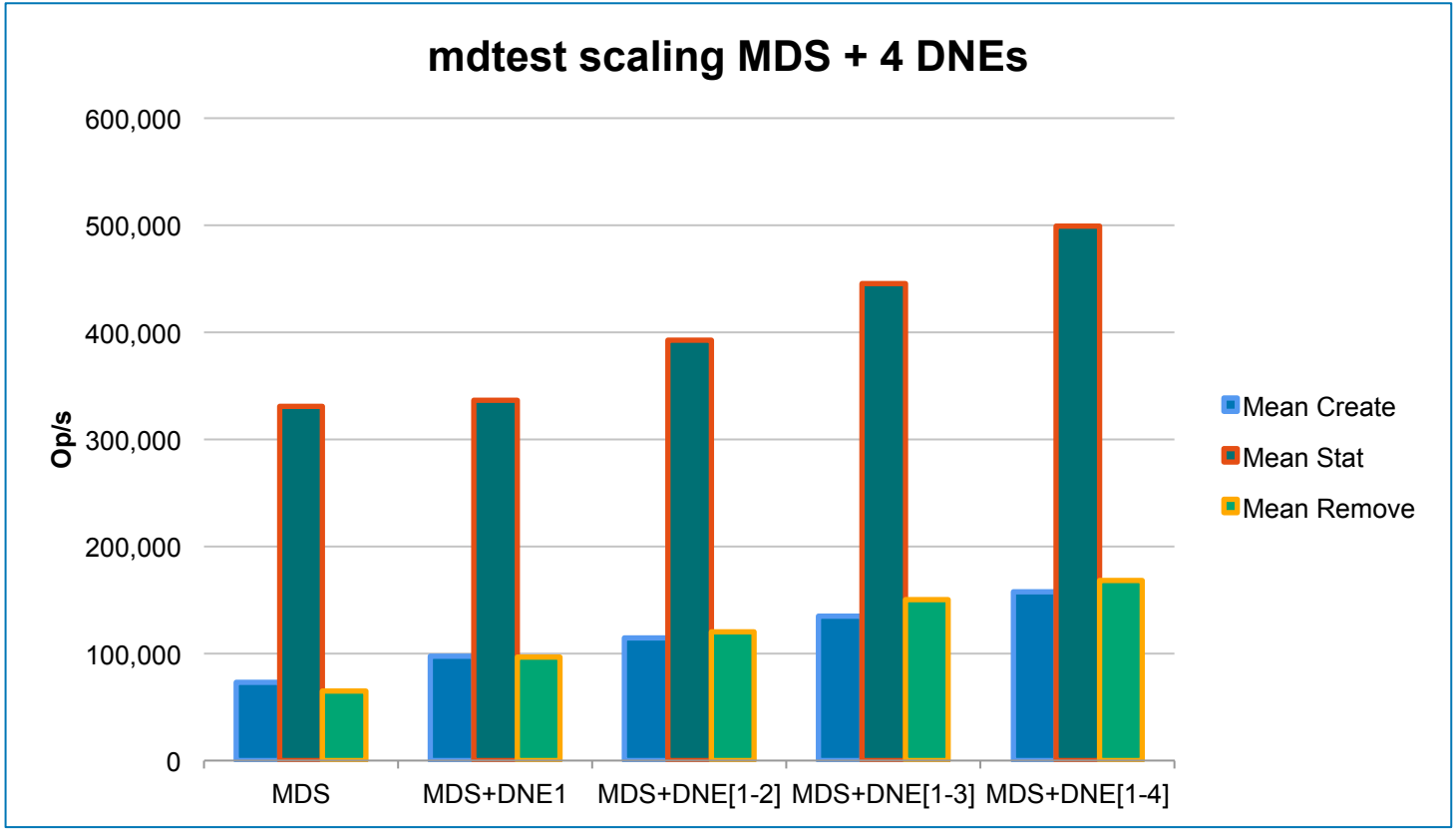
- Base MDS/MGS System Configuration (HA Pair)
  - 2x E5-2680 CPU (2.8 GHz 10-Core, 20M Cache)
  - 64GB Memory
  - 14x900GB 10K RPM Drives in R10
- Each DNE System Configuration
  - 1x 2609 v2 CPU (2.5GHz 4-Core 10M Cache)
  - 32GB Memory
  - 10x900GB 10K RPM Drives in R10
  - Note: Each ADU has 2 DNE Systems (HA Pair)
- 32 client nodes accessing MDT0 on Base MDS Node
- 8 clients accessing each ADU DNE MDTx Target

MDTest Benchmark Setup: Balance

- MDTest 1.9.3 Tool: Default, it does Creates, Reads, Stats, and Deletes
- Base MDT has 2x Higher Performance than each ADU
  - Another way to Say this: Base MDT is 4x higher performing than each DNE Target
- Benchmark Setup Example for Balanced Performance:
  - 1 MDT and 2 DNE
  - lfs mkdir -i 0 /lustre/mdt0
  - lfs mkdir -i 1 /lustre/mdt1
  - lfs mkdir -i 2 /lustre/mdt2
  - ./mdtest-1.9.3/mdtest -i 5 -n 2048 -u -F –C -T -r -d // /lustre/mdt0/0/test@/lustre/mdt0/1/test@/lustre/mdt0/2/test@/lustre/mdt0/3/test@/lustre/mdt1/0/test@/lustre/mdt2/0/test

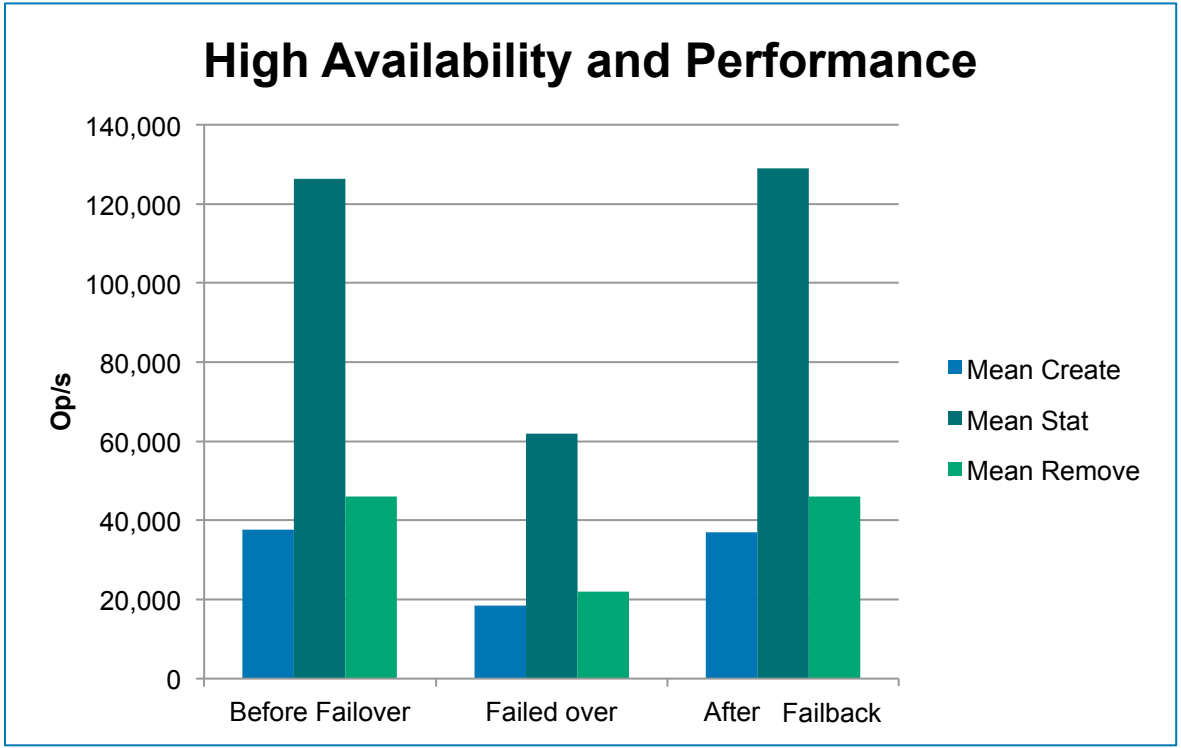
Scaling MDS and DNEs

- Base MDT + 4 DNE Servers (2 ADUs)
- mdtest create/stat/del
- Mean of 5 iterations



Metadata High Availability (ADU HA)

- MDT failover will ensure that the Lustre filesystem remains available in the face of MDS node failure
- Based on existing OSS pair failover model
- Failover is graceful, quick, and non-disruptive
- Failback is automatic and non-disruptive



Other DNE Use Cases

- Single File System exceeds 4 Billion Inodes with DNE
  - 4 Billion is a EXT4 Limitation
- Isolate different workloads on separate DNE Targets
- Creating Storage Pools with dedicated DNE Targets
- Support multiple functions in a single file system, e.g. Home, Project, and Scratch
- MDT Load Distribution

Thank You

john.fragalla@seagate.com