Optimizations and Strategies for Managing Exabyte AI Data Environments and Accelerated Computing Demands

Morris Skupinsky
Senior Solutions Architect

ddn

Lustre cliches from 2014

It’s a science project

It's a research filesystem

It’s just scratch

It can't handle flash

Sources for Lustre cliches from 2014

Maintenance and upgrades were a highly manual process and usually required downtime

Many changes required downtime to implement

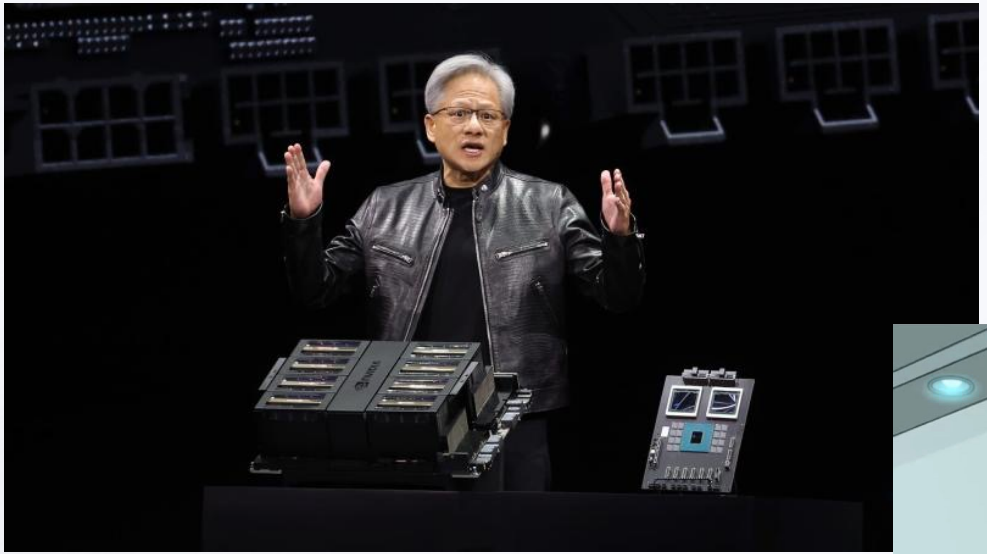
Monitoring tools required their own expertise to use

Every install was somehow unique

Requires a resident Lustre expert

Very flexible but complex – lots of knobs and no presets/guidance
Rocket scientists were literally managing Lustre filesystems

….and then



….

Step 4: PROFIT!

New environments mean new requirements

UPTIME

Time to production

Online maintenance

Simplified management and monitoring

Dynamic configurations

Expansion and migration to new hardware

New workloads at larger scale

Not enough rocket scientists to go around

DDN Reduces Complexity, Cost and Risk

Fully-integrated and optimized building blocks are easy to deploy and manage at-scale!

- Turnkey appliance: DDN has spent ten years collapsing the infrastructure and virtualizing services to provide the smallest footprint and most efficient appliances
- Full redundancy: all components are redundant and cache mirroring is done via internal PCIe links to avoid reliance on external networks for cache coherency
- Online updates: appliance software and all component firmware are updateable online

DDN EXA6 (2.14) – EXAScaler Management Framework

Up to 10X Faster and More Repeatable Deployments, Upgrades and Expansions

Orchestration Manager

Monitor cluster “state”
run/track distributed command plans

Service Mesh

- Secure comms across infrastructure
- Servers, clients and utility nodes

User Interfaces

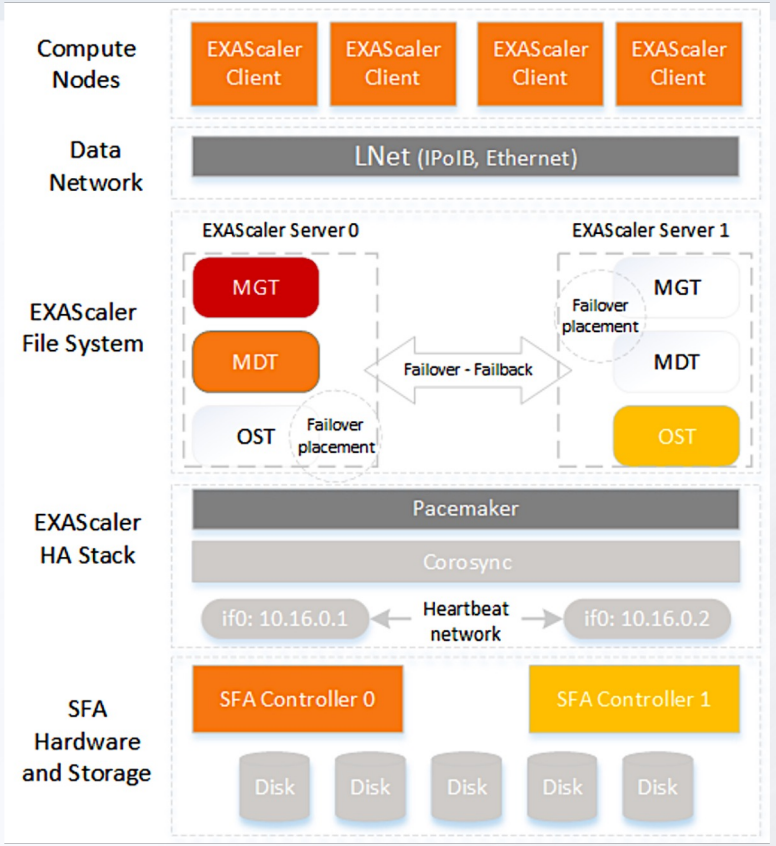
- API (GraphQL) as entry point
- Enhanced CLI and GUI for new functionality

Enhanced Security infrastructure

- Role Based Access Control
- Audit logging

HA Reliability Improvements

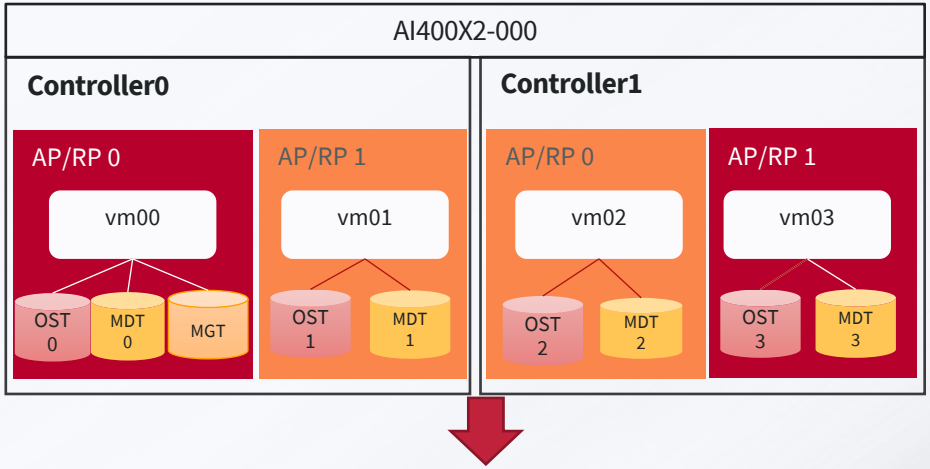
- Tolerate reduced HCA speed (for failed or degraded port)
- Ordered shutdown and startup of EXA targets
- Orderly shutdown of Stratagem EXA client
- Numerous minor improvements and bug fixes
- `clusterctl` maintenance mode (pacemaker)



Automatic Config Generation

- '`emf config generate`' tool added to generate the minimal EXAScaler config given properly configured SFA appliance
- Network details and SFA IPs are the only required user input

- Avoid potentially error prone task of creating configuration from example file
  - Doesn't require deep knowledge of available configuration fields
- Typo-free configuration file
- Implement best practices:
  - One HA group per SFA appliance
  - Avoid IO forwarding




```
[host.vm00]
oid = "0x10000000"
sfa = "AI400X2-000"

[host.vm00.fs.ai400x]
mdt_list = [0]
ost_list = [0]
```

```
[host.vm01]
oid = "0x10000001"
sfa = "AI400X2-000"

[host.vm01.fs.ai400x]
mdt_list = [1]
ost_list = [1]
```

```
[host.vm02]
oid = "0x10008000"
sfa = "AI400X2-000"

[host.vm02.fs.ai400x]
mdt_list = [2]
ost_list = [2]
```

```
[host.vm03]
oid = "0x10008001"
sfa = "AI400X2-000"

[host.vm03.fs.ai400x]
mdt_list = [3]
ost_list = [3]
```


More information in Section 2.7.1 Automated File System Cluster Configuration

EXAScaler Health

Filesystem
- targets
- evictions
- recovery

Network
- interfaces
- pings
- error counters

Cluster
- HA
- Resource placement

SFA Appliance
- controllers
- PDs/VDs
- batteries, fans, temperature

EMF Node Manager

EMF Alert Manager

Command-line tools

```
# emf health filesystem
# emf health network
# emf health cluster
# emf health sfa
```

Prometheus health exporters

DDN Insight

Customer Prometheus Infrastructure

EXAScaler Prometheus Exporters

Integrating with BCM

- Enables customer to integrate EXAScaler telemetry with their own Prometheus Infrastructure:
  - EXA and SFA exporter
  - Node exporter and system exporter from prometheus.io
- API documentation and User Guide
- Samples “starter” dashboards

EXA 6.3.1

- Lustre exporter
- SFA exporter
- Node exporter
- systemd exporter

Customer Prometheus infrastructure

Initial release

DDN Insight

Q3 2024

qa-14kxe / Historical Jobstat

DDN Insight: Historical JobStats

400 total

Job Name	User	Client	Read Throughput	Write Throughput	Read IOPS	Write IOPS	Start Time(MM/dd/yy...	Elapsed Time
job82591	0	es14ke-1-srv	1.47 MiB	906.68 KiB	15	6	01/16/2024 21:20:24	2 minutes
job82644	0	es14ke-1-srv	1.04 MiB	-	2	1	01/16/2024 21:25:23	2 minutes
job82665	0	es14ke-1-srv	1.02 MiB	-	2	0	01/16/2024 21:27:23	1 minute
job82635	0	es14ke-1-srv	1.02 MiB	634.68 KiB	7	1	01/16/2024 21:24:23	2 minutes
job82633	0	es14ke-1-srv	1017.73 KiB	11 B	2	0	01/16/2024 21:24:23	2 minutes
job82632	0	es14ke-1-srv	1017.73 KiB	-	3	1	01/16/2024 21:24:23	2 minutes
job82614	0	es14ke-1-srv	1017.73 KiB	-	5	0	01/16/2024 21:22:23	2 minutes
job82601	0	es14ke-1-srv	1017.73 KiB	11 B	2	3	01/16/2024 21:21:23	2 minutes
job82593	0	es14ke-1-srv	1017.73 KiB	11 B	3	1	01/16/2024 21:20:24	2 minutes
job82397	0	es14ke-1-srv	1017.73 KiB	616.53 KiB	3	1	01/16/2024 21:02:23	2 minutes
job82634	0	es14ke-1-srv	1001 KiB	-	4	2	01/16/2024 21:24:23	2 minutes

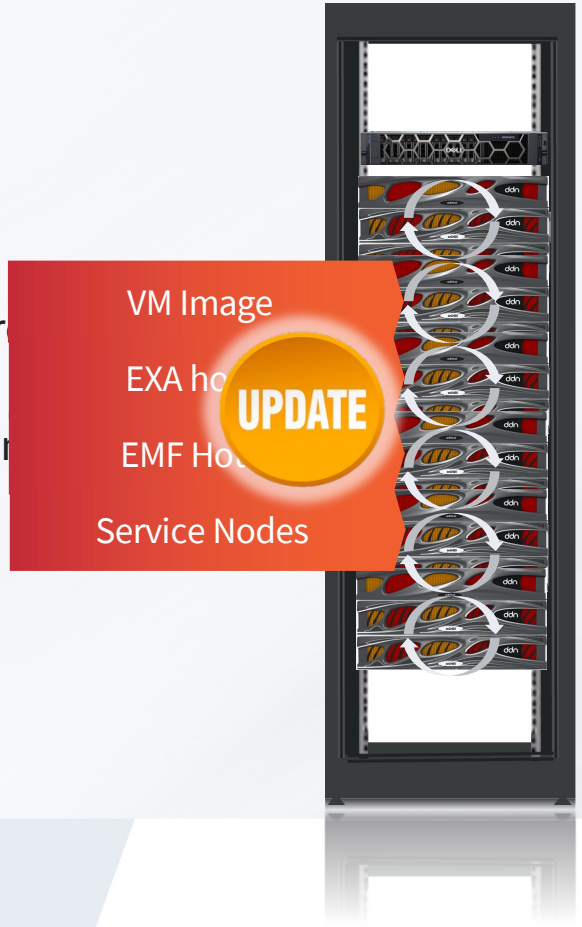
The Top 100 Consumer widget displays the Throughput, IOPS, and Metadata operations for only the top 100 user jobs running on the filesystem(s) for the last hour.

Using the Historical JobStats feature, DDN Insight retains job statistics for the past year. The user can view the job performance based on Throughput, IOPS, and other metadata.

EXAScaler Online Upgrades

ONLINE UPGRADES

- Online orchestrated updates (failover / fail back): EMF softwar
  EXAScaler software + SFA firmware
- Push-Button Upgrades (one EMF command for 100's of applia
- Updates staged with no client or server shutdowns needed



Lustre today

EXAScaler, based on Lustre, is now deployed across many enterprise environments

- Installations with hundreds of appliances
- Fast deployment (networking and compute usually take longer to deploy than Exascaler)
- Online software upgrade and expansion
- High uptime expectations
- Can monitor entire storage environment from hardware to filesystem to client

Research and therapeutic medicine
- Medical imaging
- Gene sequencing personalized therapy
- Drug discovery

Financial services
- Fraud detection
- High frequency trading/backtesting

Manufacturing
- Autonomous vehicles
- Virtual prototyping/testing
- Production quality control

AI accelerated computing and Generative AI
- LLM training
- Model checkpointing
- Edge data collection and inference

The use case

- Extremely large datasets
- Many DDN appliances
- Throughput in the TB/s for reads and writes
- 24/7 operation
- Spark writes data by creating files in a "`_temporary`" directory, then renames the file(s) to the parent directory
- In Lustre, this type of rename operation needed a BigFilesystemLock (BFL) on the MDT, serializing all rename operations (to avoid multiple clients renaming a subdirectory into itself), which also blocked lookup+open for writes in that directory
- This is further complicated when quotas are enabled. If a rename is performed into a target directory with a different projid, the kernel returns EXDEV, which forces a userspace copy of the file to the target directory rather than a rename
- The problem only became evident at large scale with tens of thousands of these operations happening every second
- These analytics/search workloads are not common for Lustre filesystems, but at the scale of the workload, Lustre performed poorly, while other storage platforms simply fail

The solution

https://jira.whamcloud.com/browse/LU-17426

- Do not hold BFL for rename of regular files, even if to a different directory on the same MDT

https://jira.whamcloud.com/browse/LU-17434

- Ensure Spark `_temporary` subdirectory is always created on same MDT as parent
- Allows previous optimization to always be possible

https://jira.whamcloud.com/browse/LU-17441

- Process rename operations with different MDS threads to avoid blocking other MDT operations

https://jira.whamcloud.com/browse/LU-13176

- Change projid before rename to avoid EXDEV when using rename to move a file into a target directory with a different projid

Whamcloud

The solution

- These changes were implemented within a few days by experienced Lustre engineers
- Backported to an EXAScaler hotfix after two weeks of intensive testing
- Customer was then able to go into full production and generating revenue
- The changes also landed in Community Lustre 2.16
- Thanks to this experience, Lustre is now better suited for Apache Spark workloads at extremely large scale

Whamcloud

ddn

INTELLIGENT DATA SOLUTIONS

WE ARE HIRING!

Join the DDN Team

APPLY NOW

#Hiring


ddn