DDN’s Vision for the Future of Lustre

LUG2015
Robert Triendl
Topics

1. The Changing Markets for Lustre
2. A Vision for Lustre that isn’t Exascale
3. Building Lustre for the Future
4. Peak vs. Operational Performance
5. Application Optimized Lustre
6. Why Conventional Storage Still Matters
# Hyperscale Storage Markets

## HPC
- Scratch
- Petabytes
- Streaming Write
- Large Files
- Infiniband
- Single Location

## Cloud
- WORM(N)
- Billions of Files
- Random Read
- Small Files
- Ethernet
- Distributed

---

**Big Data & Data Analytics**
Lustre Markets Today

- Data: 33%
- Work: 31%
- Mixed: 22%
- Cloud: 2%
- Archive: 12%
Market Diversification

- Work
- Data
- Mixed Use
- Archive

- Weather Climate
- HPC Work
- CAE Chemical
- Genomics
- Big Data Science
- Energy
- Security
- General Academic
- Finance
- HPC Cloud
- Tier 2
- Cloud

ddn.com
# Lustre Futures Beyond Exascale

<table>
<thead>
<tr>
<th>Category</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Manufacturing</strong></td>
<td>CIFS/NFS Export, AD Integration, RAS Features, Snapshots, Data Management, etc.</td>
</tr>
<tr>
<td><strong>Genomics</strong></td>
<td>Random Performance, Small File &amp; Metadata Performance, Data Management, Security, etc.</td>
</tr>
<tr>
<td><strong>General</strong></td>
<td>Broad Application Support, Connectors, User Monitoring, User Access to Snapshot, etc.</td>
</tr>
<tr>
<td><strong>Cloud</strong></td>
<td>Virtualization, Snapshots, Small File Read Performance, Data Distribution, etc.</td>
</tr>
<tr>
<td><strong>Archive</strong></td>
<td>Data Management Features, SMR Drive Use, Data Scrubs, Data Distribution, etc.</td>
</tr>
</tbody>
</table>
Market Evolution

- 2012:
  - Archive: 20%
  - Cloud: 10%
  - Mixed: 30%
  - Data: 40%
  - Work: 80%

- 2013:
  - Archive: 20%
  - Cloud: 10%
  - Mixed: 30%
  - Data: 40%
  - Work: 80%

- 2014:
  - Archive: 20%
  - Cloud: 10%
  - Mixed: 30%
  - Data: 40%
  - Work: 80%
Market Segments

<table>
<thead>
<tr>
<th>Year</th>
<th>Industry</th>
<th>Government</th>
<th>University</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td></td>
<td>0%</td>
<td>60%</td>
</tr>
<tr>
<td>2012</td>
<td></td>
<td>10%</td>
<td>50%</td>
</tr>
<tr>
<td>2013</td>
<td></td>
<td>20%</td>
<td>40%</td>
</tr>
<tr>
<td>2014</td>
<td></td>
<td>30%</td>
<td>30%</td>
</tr>
</tbody>
</table>
Disks: Throughput vs. IOPS

- Lustre 1.8
- Lustre 2.4
- ExaScaler 2.2
- Lustre with Btrfs
- Next Gen SAS Drives

![Chart showing throughput vs. IOPS for different storage systems.](chart.png)
Lustre Development at DDN

▶ Lustre Usability Features
▶ Build-in Reliability and Availability
▶ Lustre Recovery
▶ Features for a Broader Market
▶ Performance for Broad Set of Applications
▶ Application-optimized Lustre
Why Btrfs?

- Standard Local Filesystem in RHEL7
- Better Throughput Performance than ZFS
- Similar Feature Set, but all Linux
- No Possible Patent Infringement
- Simple Integration and Deployment
Application-Optimized Lustre

- Lustre for Specific Applications
- Workload Profiling
- Optimization Across I/O Calls
- Optimizing Application Runtime
- Working with Customers
Genome Pipeline Benchmarks

Samtools 20% faster with DDN Lustre optimizations

Lustre 2.5 Client Performance

Human Genetics samtools workflow

Runtime (Hours)

<table>
<thead>
<tr>
<th>Version</th>
<th>Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5.1</td>
<td>40</td>
</tr>
<tr>
<td>DDN Branch</td>
<td>30</td>
</tr>
</tbody>
</table>
SSD Pools and Caching

- DSS to Link File Layer to Block Layer
- Build into the File System
- Better use of SSDs for I/O Optimization
- Increased Small File Performance
- Increased Random Read to Large Files
- Additional specificity with `fadvise()`
ExaScaler Monitoring

- Filesystem, OSS, MDS, OST, MDT, etc.
- JOB ID, UID/GID, application stats, etc.
- Archive of data by policy

- Lightweight
- Near real-time
- Massive scale

Burst Buffer

OSS, MDS Storage

Monitoring Server
- collectd
- Graphite plugin
- graphite

UDP(TCP)/IP based small text message transfer
64 Billion of Lustre Stats in 15 days!
Why Block-Level Raid?

- Best Mixed I/O Performance
- Consistent Performance
- Hardware-optimized Performance
- Best Performance During Failure
- Integrated Storage Services
SFA RAID Stack Performance

- Above 1 Million 4K IOPS per 8 CPU Cores
- Above 10 GB/sec per 8 CPU Cores
- 8 Cores Sufficient for PCI Infrastructure
- More Cores for File System Services
- Additional Cores for More Functionality
SFA Random Read

MB/sec

512K I/O Size
1M I/O Size
2M I/O Size
4M I/O Size

DDN STORAGE
ddn.com
Flexible SSU Design
SFA14K Performance SSU

- Up to 45 GB/sec
- Up to 2950 TB
- External MDT

- 4-6 OSS
- OST Storage

- Monitoring
- 2-4 MDS
- MDT Storage
“Wolfcreek” Hardware
“Wolfcreek” Hardware