Benchmarking Lustre
Setting Realistic Performance Expectations

Torben Kling Petersen PhD
Distinguished Technologist
Lead HPC Storage Architect - EMEA & APAC

John Fragalla
Distinguished Technologist
Lead HPC Storage Architect - Americas
Problem Statement

Hardware complexity
- AMD Rome architecture means complex NUMA settings
  - CPU mapping to NVMe drives
  - CPU to Network adapters
  - CPU to SAS HBAs
- Network options
  - HDR InfiniBand
  - 200G Ethernet (both TCP and RoCE)
  - 200G SlingShot
    - (OmniPath)
- Client hardware

Software
- Lustre release
  - Point release, backports, tunables
  - Ldiskfs vs OpenZFS
- OS release(s)
Benchmarking methodology ??

Benchmark purpose:
• Marketing
• Specific customer requirements
• Product consistency
• Other ??

**IO500**
• Full
• 10 node challenge

• Issues
  • Not limited to production systems
  • Weighting between runs skews results
  • Limited tunings/modification
  • Consistency over time

**“Old School”**
• IOR
• MDtest
• FIO
• IOzone

• Issues
  • Consistency over time
Test Environments

- **Ethernet and HDR Testing**
  - 21 Clients
    - 1 MLX HCA in HDR Mode or Ethernet Mode
    - CentOS 8.4 (kernel 4.18.0-305.25.1.el8_4.x86_64) Lustre 2.12 and 2.15 Clients
  - E1000 HDR-200 System
    - 1 MDU 2 GridRAID Flash Unit 1 Flash-10 Unit 1 D2 (LDISKFS)
  - E1000 HW RoCE / TCP 200GigE System
    - 1 MDU 1 GridRAID Flash Unit 1 D2 (LDISKFS or dRAID OpenZFS)

- **Slingshot-11**
  - 21 Clients
    - CentOS 8.4 (kernel 4.18.0-305.25.1.el8_4.x86_64) Lustre 2.12 and 2.15 Clients
    - 1 CXI HCA Adapter
    - 1 MLX HCA in Ethernet Mode
  - E1000 Cassini/KFI System
    - 1 MDU 1 GridRAID Flash Unit 1 D2 LDISKFS (klibfabric)

Lustre LNET drivers used:
- ko2iblnd
  - RDMA driver used for InfiniBand HDR and 200GigE RoCE
- ksocklnd
  - TCP/IP driver used for 200GigE
- kfilnd (klibfabric)
  - RDMA driver used on Cassini (CXI) adapters with Slingshot-11
Scalable Storage Units Benchmark - Defined

<table>
<thead>
<tr>
<th>Metadata Unit (MDU)</th>
<th>Extreme Performance (SSU-F) Flash</th>
<th>IOPS Performance (SSU-F) Flash</th>
<th>HDD Performance (SSU-D2) HDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDISKFS RAID Layout</td>
<td>2 x GridRaid 12[(8+2)+1]</td>
<td>2 x GridRaid 12[(8+2)+1]</td>
<td>4 x GridRaid 53[(8+2)+2]</td>
</tr>
<tr>
<td>ZFS dRAID Layout</td>
<td>2x draid1:1d:12c:1s</td>
<td>2x draid1:1d:12c:1s</td>
<td>4x draid2:53d:16c:2s</td>
</tr>
<tr>
<td>Network ports</td>
<td>4 x 200 Gbps</td>
<td>4 x 200 Gbps</td>
<td>2 x 200 Gbps</td>
</tr>
<tr>
<td>Height Rack Units</td>
<td>2</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Number of Lustre Servers</td>
<td>2 MDS Nodes</td>
<td>2 OSS Nodes</td>
<td>2 OSS Nodes</td>
</tr>
<tr>
<td>Number of Lustre Targets</td>
<td>2 MDTs</td>
<td>2 OSTs</td>
<td>2 OSTs</td>
</tr>
</tbody>
</table>

- ClusterStor E1000 was launched in 2019 with LTS Lustre version 2.12.
- Multiple software stack updates over the years (extract below used for this presentation):
  - Neo 4.1 CentOS 7.6 (kernel 3.10.0-957.1.3957) and Lustre 2.12.0.5
  - Neo 4.4 CentOS 7.6 (kernel 3.10.0-957.1.3957) and Lustre 2.12.4.3
  - Neo 6.x Rocky Linux 8.4 OS (kernel 4.18.0-305.10) and Lustre 2.15.0.3
Benchmark Details - Standard

- **IOR Throughput**
  - Direct-IO (DIO) and Buffered-IO (BIO)
  - File-Per-Process (FPP) and Single Shared File (SSF)
  - 64M transfer size for DIO  1M transfer size for BIO
  - 16 ranks per node
  - **Fixed time results** used (provided peak performance results across different protocols)
  - **Fixed data results** collected for consistency
  - Flush caches on the clients between writes and reads

- **IOR IOPS**
  - Buffered-IO (BIO)
  - 4K transfer size with random operation
  - FPP using 8GB Files
  - 64 ranks per node
  - Fixed time results
  - Flush caches on the clients between writes and reads

- **MDTEST**
  - Unique directory operation
  - 1 Million objects per MDT
  - 16 ranks per node
  - Directory and File operations
  - Mean of 3 iterations
  - OK and 32K File sizes
  - Non-DOM Results

- **obdfilter-survey**

- **Versions**
  - IOR: 3.3.0
  - MDTEST: 1.9.3
Standard Benchmark Sweep

For each new release of patch the following benchmarks are collected:

- **Throughput** (both SSF and FPP)
  - IOR (R/W) on All Flash System (2 OSTs) with GridRAID or OpenZFS
  - IOR (R/W) on Single JBOD (2 OSTs) with GridRAID or OpenZFS
  - IOR (R/W) on 2x JBODs (4 OSTs) with GridRAID or OpenZFS
  - IOR (R/W) - Single client single thread (with and without over-striping) and multi-thread
  - IOR (R+W) – “Bi-directional” Write performance during simultaneous read (50% R / 50% W)

- **IOPS**
  - IOR (R/W) on All Flash System (2 OSTs) with GridRAID or OpenZFS
  - IOR (R/W) - Single client single thread (with and without over-striping) and multi-thread
  - IOR (R/W) on Single JBOD (2 OSTs) with GridRAID or OpenZFS
  - IOR (R/W) on 2x JBODs (4 OSTs) with GridRAID or OpenZFS

- **Metadata**
  - MDtest – Full sweep on single and dual MDTs
    - 0K and 32K files without DoM
    - 0K and 32K files with DoM

A full sweep is run for every minor and major software release or patch delivered by engineering. Full sweep takes the better part of 2 days and is fully scripted for consistency. NB all page caches are flushed between each part (e.g. W and R).
Test Permutations

Client side

- Network Checksums
  - HPE use disabled ..
- Max RPCs in Flight
  - Default is 64
  - HPE use 256
- Max Dirty MB
  - Default is 2000
  - HPE use default
- Max Pages per RPC
  - Default is 256 (1MB),
  - HPE use 4MB for flash and 16MB for HDD based systems
- Max Read Ahead MB
  - Default is 64 MiB
  - HPE use 512MiB
- Max Read Ahead per File MB
  - Default is 64 MiB
  - HPE use 512MiB

Server side

- System tunings
  - NPS (1, 2, 4)
  - HPE use
    - NPS4 for MDS
    - NPS2 for LDISKFS OSS
    - NPS1 for OpenZFS OSS
    - CPT=8 for all systems
- Failover testing
All Flash Array with PD-RAID LDISKFS Performance

All Flash Array Throughput - ko2iblnd

Throughput (GB/s)

Neo release

4.1
4.4
6.1

Fix Time DIO Write (GB/s)
Fix Time BIO Write (GB/s)
Fix Time DIO Reads (GB/s)
Fix Time BIO Reads (GB/s)

24%
30%
18%
6%
Clusterstor E1000 SSU Flash Gridraid LDISKFS Performance

**All Flash Array IOPS - ko2iblnd**

- BIO Write IOPS

**122%**

**All Flash Array IOPS - ko2iblnd**

- BIO Read IOPS

**54%**
HDD Performance – 2x JBODs PD-RAID on LDISKFS
Performance gain from new firmware BIOS Change from NPS4 to NPS2 Lustre 2.15 & new kernel.

HDD Throughput (2x JBODs) - ko2iblnd

- Fix Time DIO Write (GB/s)
- Fix Time BIO Write (GB/s)
- Fix Time DIO Reads (GB/s)
- Fix Time BIO Reads (GB/s)

Throughput (GB/s)

Neuro Release

- 16% 21% 17% 10%
All Flash RAID-10 LDISKFS IOPS Performance

Performance gain from Lustre 2.15 & new kernel.

![Graph showing performance gains in IOPS](image-url)
Single All Flash MDT LDISKFS Performance
Performance gain from Lustre 2.15 & new kernel.
All Flash OST - Single Stream IO Performance

Performance gain from Lustre 2.15 & new kernel.

- Single Stream IO Performance
  - 21%
  - 75%
  - 115%
  - 100%
  - 21%
HDD TCP Performance - TCP on 2x JBODs with LDISKFS
Performance gain from BIOS Change from NPS4 to NPS2

HDD Throughput (2x JBODs) TCP - ksockInd

- 158%
- 126%
- 120%
- 158%
- 158%

ClusterStor Neo Release

Fix Time DIO Write (GB/s)
Fix Time BIO Write (GB/s)
Fix Time DIO Reads (GB/s)
Fix Time BIO Reads (GB/s)
Comparing RDMA Fabrics Protocols
**PERFORMANCE UPDATE E1000 SSU-F**

- Samsung PM1733
- 20+ clients, Stonewalling IOR, GridRAID-12, Lustre 2.15 etc ....

### GridRAID

<table>
<thead>
<tr>
<th></th>
<th>IO</th>
<th>IB (HDR) 6.1-010.39</th>
<th>Ethernet (TCP) 6.1-010</th>
<th>Ethernet (RoCE) 6.1-010</th>
<th>KFI 6.1-010.40</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DIO 64PPN</strong></td>
<td>Write</td>
<td>58.9</td>
<td>35.7</td>
<td>57.9</td>
<td>54.2</td>
</tr>
<tr>
<td></td>
<td>Read</td>
<td>85.3</td>
<td>81.2</td>
<td>85.4</td>
<td>82.2</td>
</tr>
<tr>
<td><strong>BIO 64PPN</strong></td>
<td>Write</td>
<td>63.4</td>
<td>36.5</td>
<td>64.4</td>
<td>59.1</td>
</tr>
<tr>
<td></td>
<td>Read</td>
<td>83.3</td>
<td>70.7</td>
<td>83.1</td>
<td>82.1</td>
</tr>
<tr>
<td><strong>IOR Buffered IO</strong></td>
<td><strong>4K Random IOPS</strong></td>
<td>Write</td>
<td>84,389</td>
<td>83,747</td>
<td>85,590</td>
</tr>
<tr>
<td></td>
<td>Re-Write</td>
<td>53,085</td>
<td>53,903</td>
<td>48,542</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Read</td>
<td>1,217,062</td>
<td>676,030</td>
<td>1,217,216</td>
<td>735,866</td>
</tr>
</tbody>
</table>
Throughput Comparison HDD based OSTs
20+ clients, Stonewalling IOR, GridRAID-53, Lustre 2.15 etc ....

<table>
<thead>
<tr>
<th></th>
<th>IO</th>
<th>IB (HDR) 6.1-010.39</th>
<th>Ethernet (RoCE) 6.1-010.48</th>
<th>KFI 6.1-010.40</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single JBOD (2 OSTs)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIO 64PPN</td>
<td>Write</td>
<td>18.7</td>
<td>18.5</td>
<td>21.4</td>
</tr>
<tr>
<td></td>
<td>Read</td>
<td>19.2</td>
<td>18.9</td>
<td>20.2</td>
</tr>
<tr>
<td>BIO 64PPN</td>
<td>Write</td>
<td>17.7</td>
<td>17.8</td>
<td>20.3</td>
</tr>
<tr>
<td></td>
<td>Read</td>
<td>16.2</td>
<td>16.2</td>
<td>17.9</td>
</tr>
<tr>
<td><strong>2x JBODs (4 OSTs)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIO 64PPN</td>
<td>Write</td>
<td>35.9</td>
<td>35.8</td>
<td>33.9</td>
</tr>
<tr>
<td></td>
<td>Read</td>
<td>35.9</td>
<td>34.7</td>
<td>33.0</td>
</tr>
<tr>
<td>BIO 64PPN</td>
<td>Write</td>
<td>34.9</td>
<td>34.9</td>
<td>36.7</td>
</tr>
<tr>
<td></td>
<td>Read</td>
<td>33.6</td>
<td>31.2</td>
<td>36.2</td>
</tr>
</tbody>
</table>
### E1000 SINGLE MDT RAID-10

- 20+ clients, Stonewalling IOR, GridRAID-53, Lustre 2.15 etc ....

<table>
<thead>
<tr>
<th>MDTest (Single MDT)</th>
<th>IB (HDR) 6.1-010.39</th>
<th>Ethernet (RoCE) 6.1-010.48</th>
<th>KFI 6.1-010.40</th>
</tr>
</thead>
<tbody>
<tr>
<td>File Creates per Second</td>
<td>155,771</td>
<td>156,851</td>
<td>100,699</td>
</tr>
<tr>
<td>File Stats per Second</td>
<td>682,054</td>
<td>683,682</td>
<td>524,510</td>
</tr>
<tr>
<td>File Reads per Second</td>
<td>301,053</td>
<td>293,466</td>
<td>210,164</td>
</tr>
<tr>
<td>File Removes per Second</td>
<td>143,286</td>
<td>152,239</td>
<td>113,314</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Single MDT</th>
<th>IB (HDR) 6.1-010.39</th>
<th>Ethernet (RoCE) 6.1-010.48</th>
<th>KFI 6.1-010.40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Directory Creates/sec</td>
<td>109,971</td>
<td>101,232</td>
<td>83,437</td>
</tr>
<tr>
<td>Directory Stats/sec</td>
<td>406,336</td>
<td>406,419</td>
<td>351,709</td>
</tr>
<tr>
<td>Directory Removes/sec</td>
<td>175,605</td>
<td>180,357</td>
<td>136,213</td>
</tr>
<tr>
<td>File Creates/sec</td>
<td>156,606</td>
<td>159,457</td>
<td>101,253</td>
</tr>
<tr>
<td>File Stats/sec</td>
<td>679,259</td>
<td>683,807</td>
<td>526,747</td>
</tr>
<tr>
<td>File Reads/sec</td>
<td>284,603</td>
<td>255,884</td>
<td>212,517</td>
</tr>
<tr>
<td>File Removes/sec</td>
<td>143,981</td>
<td>155,846</td>
<td>122,003</td>
</tr>
</tbody>
</table>
Lustre MDT performance - TCP/IP vs HDR IB RAID-10 LDISKFS

<table>
<thead>
<tr>
<th>MDT0 0K Files Unique Directory</th>
<th>1M objects</th>
<th>Operation</th>
<th>ksocklnd</th>
<th>ko2iblnd</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dir creation</td>
<td>91 841</td>
<td>109 971</td>
<td>109 971</td>
<td>84%</td>
<td></td>
</tr>
<tr>
<td>Dir stats</td>
<td>319 945</td>
<td>406 336</td>
<td>406 336</td>
<td>79%</td>
<td></td>
</tr>
<tr>
<td>Dir removes</td>
<td>140 888</td>
<td>175 605</td>
<td>175 605</td>
<td>80%</td>
<td></td>
</tr>
<tr>
<td>File creation</td>
<td>108 772</td>
<td>156 606</td>
<td>156 606</td>
<td>69%</td>
<td></td>
</tr>
<tr>
<td>File stats</td>
<td>326 335</td>
<td>679 259</td>
<td>679 259</td>
<td>48%</td>
<td></td>
</tr>
<tr>
<td>File reads</td>
<td>184 085</td>
<td>284 603</td>
<td>284 603</td>
<td>65%</td>
<td></td>
</tr>
<tr>
<td>File removes</td>
<td>115 581</td>
<td>143 981</td>
<td>143 981</td>
<td>80%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MDT0 32K Files Unique Directory</th>
<th>1M objects</th>
<th>Operation</th>
<th>ksocklnd</th>
<th>ko2iblnd</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dir creation</td>
<td>97 151</td>
<td>109 271</td>
<td>109 271</td>
<td>89%</td>
<td></td>
</tr>
<tr>
<td>Dir stats</td>
<td>318 725</td>
<td>408 859</td>
<td>408 859</td>
<td>78%</td>
<td></td>
</tr>
<tr>
<td>Dir removes</td>
<td>139 571</td>
<td>193 565</td>
<td>193 565</td>
<td>72%</td>
<td></td>
</tr>
<tr>
<td>File creation</td>
<td>107 753</td>
<td>152 572</td>
<td>152 572</td>
<td>71%</td>
<td></td>
</tr>
<tr>
<td>File stats</td>
<td>339 420</td>
<td>698 665</td>
<td>698 665</td>
<td>49%</td>
<td></td>
</tr>
<tr>
<td>File reads</td>
<td>182 009</td>
<td>216 228</td>
<td>216 228</td>
<td>84%</td>
<td></td>
</tr>
<tr>
<td>File Removes</td>
<td>112 215</td>
<td>149 733</td>
<td>149 733</td>
<td>75%</td>
<td></td>
</tr>
</tbody>
</table>

ksocklnd IOPS limited by TCP/IP latency and CPU utilization.
Process and Lessons Learnt

Switching from Intel to AMD Rome

• Started with BIOS setting with 4 NUMA domains (NPS4) and changed to 2 NUMA domains (NPS2) plus tuning Lustre CPU Partition Tables to 8 provided an increase in performance on throughput and IOPS

Not all NVMe drives are the same

• Different NVMe drive vendors do not perform equal despite similar specs (e.g. Samsung PM1733 and Kioxia CM6)
  • New NVMe Firmware improved performance

Keeping up with Linux enhancements takes a lot of work

• Moving from RHEL based 7.8 kernel with Lustre 2.12 to RHEL based 8.4 Kernel (e.g. Rocky Linux 8.4) with Lustre 2.15 provided additional improvement on DIO path
  • Significant experimentation of Lustre tunables required

New versions (client, server, ofed, OS etc.) often introduce regressions

• Repeated baseline testing is paramount to deliver consistency
SUMMARY

• Since 2019 ClusterStor E1000 has improved all facets of performance up to 400% depending on the I/O operation.
• Changes were due to constant tunings on the platform improvements with new kernels or adopting Lustre 2.15.
• ko2iblnd performance is identical for HDR InfiniBand and HW RoCE 200 GigE.
• ksocklnq performance is limited on peak writes or due to TCP/IP Latency.
• Lustre 2.15 brings big performance improvement on MDT File Stats/s and single stream performance.

• RDMA based protocols perform essentially the same regardless of type (e.g., Infiniband, RoCE or SlingShot)...
• Continuous benchmarking is important for any product.
  • Performance regressions can and will occur frequently....
  • The ability to confidently propose sizing to meet a future deployment is key.
• Peak performance is repeatable but ONLY in the lab.
  • No NOT expect hero numbers in customer environments (but we can get close) ...
THANK YOU

(for listening to a madmans ramblings ....)

tkp@hpe.com