HPC storage @ CSCS

Stefano Gorini and Colin McMurtrie, CSCS
June 23rd, 2016
Online Filesystems
Mission Critical Filesystems

- optimized for small files
- GPFS 3.5
- blocksize: 256 KiB
- metadata on SSD in double copy
- files < 128 KiB on inodes

<table>
<thead>
<tr>
<th>Filesystem</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>/users</td>
<td>86 TiB</td>
</tr>
<tr>
<td>/apps</td>
<td>58 TiB</td>
</tr>
</tbody>
</table>

Net: 144 TiB  Raw: 200 TB  50
Online Filesystems with Backup - /project

<table>
<thead>
<tr>
<th>Filesystem</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>/project</td>
<td>5.8 PiB</td>
</tr>
</tbody>
</table>

- optimized for big files
- GPFS 3.5
- blocksize: 1 MiB
- metadata on SSD in double copy
- files < 256 KiB on inodes
- 2015: + 2 PiB
- Quota based on research proposals

- 2 EMC VNX8000
- 37 disk enc. (60 slot each)
- RAID6 + Hot spares
- 4 TB NL-SAS disks
- SSDs for Metadata

Net: 5.94 PiB  Raw: 8.65 PB  2163
Online Filesystems with Backup - /store

- GPFS 4.1
- 2 Tiers
- Policy based on access time
- Quota based on contracts
- Previous size: 2.6 PiB

<table>
<thead>
<tr>
<th>Filesystem</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>/store</td>
<td>4.4 PiB</td>
</tr>
</tbody>
</table>

### EMC VNX8000
- 1 EMC VNX8000
- 8 disk enc. (60 slot each)
- RAID6 + Hot spares
- 4 TB NL-SAS disks
- SSDs for Metadata

### NetApp E5600
- 3 NetApp E5600
- 18 disk enc. (60 slot each)
- parity-declustering RAID
- 6 TB NL-SAS disks

Net: 10.34 PiB  Raw: 15.21 PB
Online Filesystems - `/scratch-shared`

<table>
<thead>
<tr>
<th>Filesystem</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>/scratch-shared</code></td>
<td>1.2 PiB</td>
</tr>
</tbody>
</table>

- GPFS 4.1
- No backup
- Prev. size: 642 TiB

Net: 11.54 PiB  Raw: 17.01 PB  3712
Online Filesystems - /scratch/daint

<table>
<thead>
<tr>
<th>Filesystem</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>/scratch/daint</td>
<td>2.7 PiB</td>
</tr>
</tbody>
</table>

- optimized for very big files
- optimized for writes
- 116 GiB/s as peak performance
- Lustre 2.1
- 6582 client nodes (dora+daint)
- Robinhood for cleaning policies

Net: 14.24 PiB  Raw: 20.95 PB  5680

Lustre Scratch FS
24 SSU, 48 OSSes,
1 MMU, 1 MGS
2.7 PiB
Other Storage Systems

<table>
<thead>
<tr>
<th>Filesystem</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>/scratch/santis</td>
<td>167 TiB</td>
</tr>
<tr>
<td>/scratch/dora</td>
<td>904 TiB</td>
</tr>
<tr>
<td>/scratch/brisi</td>
<td>226 TiB</td>
</tr>
</tbody>
</table>

- Test and Development Systems for Cray Sonexion 1600

- Cray Sonexion 2000 for Dora and its TDS systems
  - Lustre 2.1
  - Declustered RAID (GridRAID)
  - New Expansion Storage Units
  - 4 OSSs with 2 OSTs each one
  - 41 disks (113 TiB) per OST
  - stripe_count=1

- Management Infrastructure (Nagios, Ganglia, Puppet, Greylog, custom solutions…)

Net: 15.51 PiB  Raw: 22.84 PB  6172
Data Services
Data Transfer Service

- Data Movers Services
  - GridFTP (4 nodes cluster)
  - GPFS AFM for HBP
    - Between CSCS and EPFL
    - To be extended to Juelich, Cineca and BSC.
Backup/Archive Service

- 3 TSM Servers + 1 spare
- IBM TS3500 Tape Library (18257 slots)
- 28 drives (24 LTO5 + 4 LTO6)
- 12510 LTO5 + 100 LTO6 cartridges
- Mainly used with mmbackup for GPFS
- 5 Storage Agents
- Big DB2 databases (~ 400GB) to keep metadata infos

Raw: 19.01 PB
Backup/Archive Service
Customized Solutions
MeteoCH

- Cray Sonexion 1300 for old /workspace
- Cray Lustre for old /opr and new /scratch
- built on NetApp hardware

<table>
<thead>
<tr>
<th>Filesystem</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>/workspace</td>
<td>223 TiB</td>
</tr>
<tr>
<td>Albis /opr</td>
<td>18 TiB</td>
</tr>
<tr>
<td>Lema /opr</td>
<td>18 TiB</td>
</tr>
<tr>
<td>Escha /scratch</td>
<td>73 TiB</td>
</tr>
<tr>
<td>Kesch /scratch</td>
<td>73 TiB</td>
</tr>
</tbody>
</table>

Net: 22.85 PiB  Raw: 32.82 PB  9636
Description of the Problem

- **FIELDEXTRA (pre/post processing Fortran tool) slowdown**

<table>
<thead>
<tr>
<th>Condition</th>
<th>zone_reclaim_mode</th>
<th>Number of Runs</th>
<th>Average [s]</th>
<th>Standard dev [s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
<td>15</td>
<td>198.533</td>
<td>12.928</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>38</td>
<td>440.921</td>
<td>337.741</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>62</td>
<td>193.677</td>
<td>27.617</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>161</td>
<td>499.379</td>
<td>1133.936</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>173</td>
<td>199.08</td>
<td>11.316</td>
</tr>
</tbody>
</table>
Is it the FS?

- Lets try GPFS…..
  - No Variation FIELDEXTRA always perform the same
- So is it Lustre FS storage HW? ….No
Dedicated Test and Analysis Session

- All the problems are not related to an high load on the Lustre file system
- The kernel parameter reclaim \texttt{vm.zone_reclaim_mode} has a significant effect on the slowdown ("condition 5")
- Running the suite on the same node mitigates the slowdown

Important Remark:

During the analysis of the Fieldextra process with \texttt{perf}, in case of slowdown, Fieldextra was spending a lot of time with the kernel function \texttt{clear_page_c_e}:

Samples: 1M of event 'cycles', Event count (approx.): 854374192198
13.12% Fieldextra [kernel.kallsyms] [k] \texttt{clear_page_c_e}
7.58% fieldextra fieldextra_12.2.0_gnu4.9.3_opt_omp [] spumb_c_
7.35% fieldextra [kernel.kallsyms] [k] compaction_alloc
Solution

- MCH redesigned the initialization of data arrays (~40 GB on disk) by doing this initialization stepwise.
- With this new version of fieldextra no significant performance fluctuation has been seen. More testing is underway to confirm these results.
- Running the test case during more than 12 hours without cache cleaning on all nodes (“condition 5”).
- The new initialization even improves the performance on top of that:

  “The test case ~30% faster than the fastest runtime with the current operational executable”
Q & A

sgorini@cscs.ch - colin@cscs.ch