Lustre on ZFS

Ricardo M. Correia
Lustre Group
ZFS features

- Immense capacity
- End-to-end checksumming
- Self-healing
- Pooled storage model
- Lightweight snapshots, clones
- Built-in compression
- Easy administration
ZFS / ldiskfs comparison

• Advantages of ZFS/DMU:
  > Can run in userspace, more portable
  > Protection from data corruption
  > Larger limits
  > Good stress tester (ztest/lztest)
  > Many useful features
  > No zfsck

• Disadvantages of ZFS/DMU:
  > No zfsck
  > More CPU and IO overhead
Current Lustre-ZFS status

• Compared to ldiskfs:
  > Performance not there yet
  > No user/group quotas
  > Failover with ZFS not working yet
  > No multi-mount protection

• Compared to Solaris ZFS:
  > Almost everything works, but:
    – No easy way to change tunables yet
    – No FMA, hot spares not working

• FUSE no longer required for Lustre
ZFS performance

- Performance not as good as ldiskfs yet

![RAID-0 streamed write throughput graph]

- ext3
- Lustre ZFS
- Solaris ZFS
- Idiskfs
- raw disk BW
ZFS performance

• ..but it's looking much better now

<table>
<thead>
<tr>
<th>Config</th>
<th>PIOS threads</th>
<th>Chunk size</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>42</td>
<td>1M</td>
</tr>
<tr>
<td>B</td>
<td>42</td>
<td>2M</td>
</tr>
<tr>
<td>C</td>
<td>84</td>
<td>128K</td>
</tr>
</tbody>
</table>

![Graph showing throughput (MB/s) for different configurations and chunk sizes]
How to get good performance

• Lots of room for improvements:
  > ZIO pipeline optimizations (async I/O, ...)
  > Zero-copy
  > Larger IOs to disk
  > More intelligent block allocator
  > Cache size/txg size tuning
  > ZAP improvements
  > Checksum offload

• For metadata only:
  > EAs in the dnode + larger dnodes
  > Disable ditto blocks
Other things that need to be done

• User/group quotas
• Multi-mount protection
  > Even more important than with ldiskfs
• ACLs
  > Lustre uses POSIX ACLs
  > ZFS uses NFSv4 ACLs
• Ext3-like feature flags
Thanks

Ricardo.M.Correia@Sun.COM

http://opensolaris.org/os/community/zfs/