The ONE MILLION IOPS
TU Dresden Lustre FS
21-03-2016
>> Agenda
Agenda
▶ Benchmark Requirement
▶ Storage System
  – Disk IOPS
  – Controller Performance
  – Server CPU & Memory
▶ IO Cell
  – OSS IO Cell
  – OSS & MDS IO Cell
Benchmark Requirement
Benchmark Requirement

▶ Committed to:
  - 1 Miops/s Random Read

▶ Measurement on FAT:
  - 1.3 Miops Random Read

▶ What we have done to reach it:
  - N=609
  - ppn=24
  - n=14616
  - Runtime=266s
  - Score=1,320,137 iops

Dear All,

I’m proud to announce you that we finally reached our commitments for the randomops benchmark.

<table>
<thead>
<tr>
<th>JOBID</th>
<th>N</th>
<th>n</th>
<th>Time</th>
<th># Ops/task</th>
<th>IOPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>12501</td>
<td>609</td>
<td>14616</td>
<td>266</td>
<td>20000</td>
<td>1,820,137</td>
</tr>
<tr>
<td>12504</td>
<td>609</td>
<td>14616</td>
<td>37</td>
<td>2000</td>
<td>1,342,102</td>
</tr>
<tr>
<td>12515</td>
<td>609</td>
<td>14616</td>
<td>37</td>
<td>2000</td>
<td>1,353,712</td>
</tr>
</tbody>
</table>

Regards,

Diego Moreno
>> Storage System
Storage System Chosen
Iops Performance of Disk Drives

- Toshiba PX02SMF080 800GB 2.5 SAS
  - Random Read
    • 120 000 Iops 4K
  - Random Write
    • 25 000 Iops 4K
  - Sequential Read
    • 900 MB/s
  - Sequential Write
    • 400 MB/s
Storage System Chosen
Netapp EF560

- Per controller
  - 1xIvy Bridge 6 Core
  - 4 port SAS3 12Gb/s
  - 24 Gib of Ram
  - 12 GiB of Mirrored data cache

- For each dual controller
  - 20xSAS SSD Toshiba
Storage System Chosen
Netapp EF560

- **EF560 Duplex System**
  - 3x8GiB DIMM
  - 4xSAS3 port
  - 6 core Ivy Bridge CPU

- **Netapp EF560**
  - 9x4xSAS3 port
  - 6 core
  - Ivy Bridge
  - CPU
  - 3x8GiB DIMM
Storage System Chosen
Netapp EF560 controller Iops and Sequential Performance

- Random Read
  - 825Kiops 8KiB Raid5
  - 900Kiops cached 4k

- Random Write
  - 130Kiops 4KiB

- Sustained Read 512KiB
  - 12 000MB/s

- Sustained Write 512KiB
  - 6 000MB/s CME
  - 9 000MB/s CMD

Netapp EF560 - 48 SSDs - 8K RAID 5

<table>
<thead>
<tr>
<th>% Reads</th>
<th>Under 1ms</th>
<th>Under 0.5ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>650,000</td>
<td>628,000</td>
</tr>
<tr>
<td>75%</td>
<td>314,000</td>
<td>281,000</td>
</tr>
<tr>
<td>50%</td>
<td>195,000</td>
<td>168,000</td>
</tr>
<tr>
<td>25%</td>
<td>144,000</td>
<td>133,000</td>
</tr>
<tr>
<td>0%</td>
<td>117,000</td>
<td>102,000</td>
</tr>
</tbody>
</table>

Tests were run with Iometer & Toshiba drives

X IOPs Mix read/write
Y latency in ms
Storage System Chosen
R423e3 IO Server CPU and Memory

- **CPU**
  - 2xIvy Bridge E5-2650v2@2.6GHz
  - 8 core no HT

- **Ram**
  - 2x4x8GiB@1600MT/s

- **Infiniband**
  - 2xIB FDR Card 1xCard by Socket
  - 6GB/s fullduplex

- **SAS Card**
  - 4xSAS3 Card 2xCard by Socket
  - 2.4GB/s fullduplex
Storage System Chosen
R423e3 IO Server CPU and Memory

- Ivy Bridge 8 core
- 4 DIMM Channel
- Dual Socket
- 3 PCIeG3 Card
>> IO Cell
IO Cell
OSS IO Cell

IO Cell = 2xOSS + 1xDual Controller Storage
IO Cell
Three OSS IO Cell & One MDS IO Cell
Thanks
For more information please contact:
johann.peyrard@atos.net

Atos, the Atos logo, Atos Consulting, Atos Worldgrid, Worldline, BlueKiwi, Canopy the Open Cloud Company, Yunano, Zero Email, Zero Email Certified and The Zero Email Company are registered trademarks of Atos. July 2014. © 2014 Atos. Confidential information owned by Atos, to be used by the recipient only. This document, or any part of it, may not be reproduced, copied, circulated and/or distributed nor quoted without prior written approval from Atos.

21-03-2016
Performance Measurements
Of a Global SSD Lustre File System

Lustre User Group 2016, Portland, Oregon

Zellescher Weg 12
Willers-Bau A 207
Tel. +49 351 - 463 – 34217

Michael Kluge (michael.kluge@tu-dresden.de)
Measurement Setup

- no read caches on the server side:

  root@taurusadmin3:~> pdsh -w oss[21-26] lctl get_param obdfilter.*.read_cache_enable
  oss26: obdfilter.highiops-OST0009.read_cache_enable=0
  ...

- how files are opened

  file_fd = open( filename, O_RDWR|O_CREAT, 0600 );
  posix_fadvise( file_fd, 0, 0,
                     POSIX_FADV_RANDOM | POSIX_FADV_NOREUSE | POSIX_FADV_DONTNEED);
Measurement Setup

- Size on disk at least 4x size of server RAM
- Rotation of MPI ranks
- Data written before the test
- Always started at LUN 0
- Size of one IOP: 4 KiB to 1 MiB
- Data collected NOT in exclusive mode
- Data presented as maximum at least three measurements
- Each run was about 5 minutes
- 1 individual file per process, always used pread/pwrite
Single Process / Single LUN for Our SATA Disks (writes)
Single Process / 1 - 12 LUNs

Performance of one LUN, one process

- Single process becomes I/O bound immediately
Testing Alls LUNs Of The SSD File System

Utilized all cores

Writes to a single LUN are faster than reads

Performance variations > 40%
One Node / Many LUNs / Different Stripes / Reads+Writes

4K read IOPS

- stripe 1
- stripe 2
- stripe 4
- stripe 8
- stripe 12

process count

IOPS

4K write IOPS

- stripe 1
- stripe 2
- stripe 4
- stripe 8
- stripe 12

process count

IOPS
**One Node / Different Block Size (Writes)**

- Different block sizes from 4 KiB to 1 MiB
- Measurement done twice
  - 4K steps
  - Only powers of 2
- Peak IOPS still the same
- Peak bandwidth at 10 GB/s
- There is some pattern in the bandwidth curve
One Node / Ratio between Reads and Writes

- Changing the mixture between reads and writes
- From 0% to 100% writes in steps of 5%
- For 24 processes there is a sweet spot …
Many Nodes / Many LUNs

- Stripe 12 (all files have one stripe on all SSDs)
- Up to 256 nodes (out of ~1500)
- 24 processes/node
- Reads are still cached?
- 1 Mio. IOPS was measured with:
  - 40 TB data
  - 1500 nodes, 24 ppn
  - reads only
What to take home

- Single process can issue about 30,000 IOPS (CPU bound)
- One node can issue > 100,000 write IOPS
- (close to) Peak IOPS of the file system can be reached with only a few nodes
- Performance remains stable as node numbers increase
- Writes appear to be faster as long as the performance capacity of the underlying hardware is not maxed out (writes on most SSDs are generally slower)