IO500 on Bancho Cluster of Exascaler/Lustre

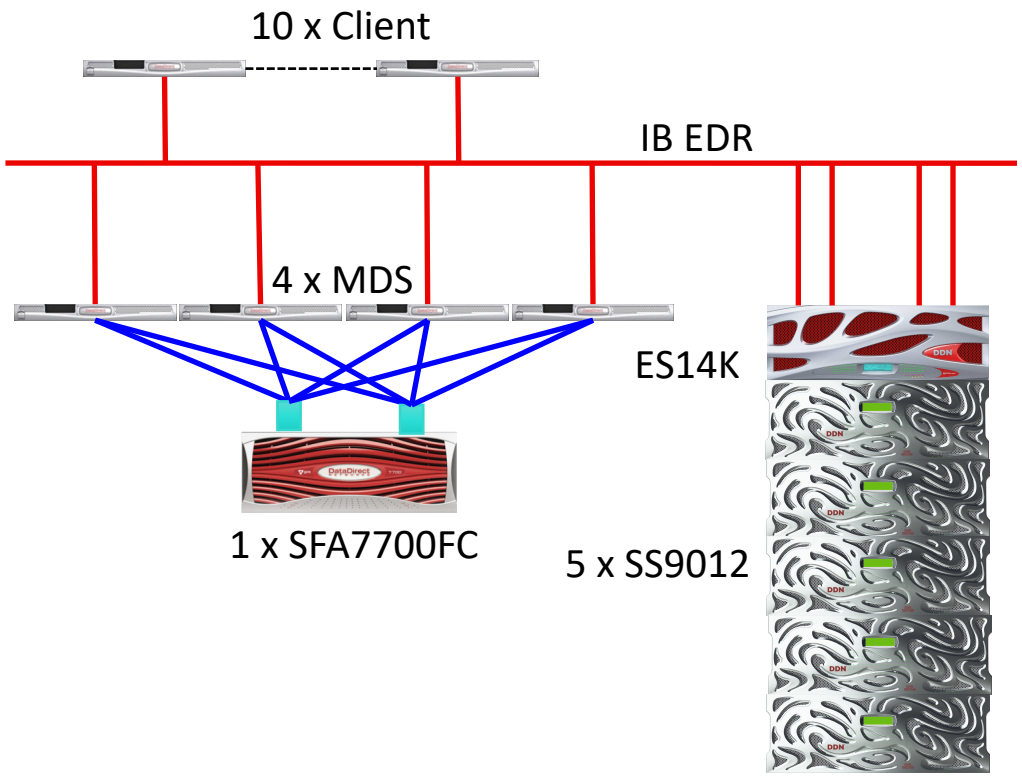
Shuichi Ihara

About “Bancho” Cluster

Whamcloud

- Cluster locates at DDN Tokyo office in Japan
  - Benchmark lab opened in 2010 after started DDN business in Japan, 2008
  - Various Storage, Server and network configuration
  - Co-located Whamcloud Lustre Performance engineering
- Focused on IO benchmark
  - Customer IO Benchmark and POC
  - Strategic IO performance benchmark
    - New H/W technology and new version of software
  - IO Performance analysis and optimization

DDN IO-500 Exascaler 5.0 (PRE) Test Configuration



MDS/MDT	4 x MDS 1x Platinum 8160, 96GB RAM, 1 x IB EDR 1 x SFA7700 FC 2 x MDT(2 x RAID1 SSD) per MDS
OSS/OST	1 x ES14K + 5 x SS9012 408 x NL-SAS 10TB 8 x DCR POOL (51/MR=1) 4 x vOSS(8 CPU Core, 90GB RAM, 1x IB EDR
Client	10 x Intel Server 2 x E5-2650v4, 128GB RAM, 1x IB EDR CentOS7.4
Software	Pre-Exascaler5.0 master branch for lustre-2.12

Hardware resource ratio of Bancho cluster

Whamcloud

- Total aggregate hardware resource on server and client

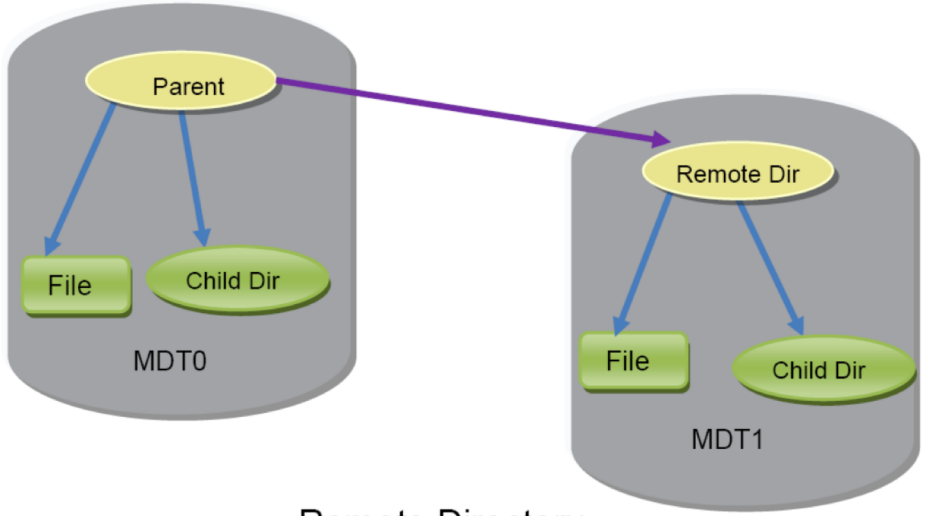
	Client	Sever	Server/Client
CPU Core	240 (24 x 10 Client)	128 (8 x 4 OSS, 24 x 4 MDS)	50%
RAM(GB)	1280 (128 x 10 Client)	744 (90 x 4 OSS, 96 x 4 MDS)	58%
Network Bandwidth(Gbps)	1000 (100 x 10 Client)	400 + 400(100 x 4 OSS, 100 x 4 MDS)	40%

- Aggregate storage device bandwidth against storage network bandwidth
  - Storage devices 101GB/s(254MB/s x 400 HDD) > storage network 50GB/sec (100Gbps x 4 OSS)
  - Bottleneck on storage network bandwidth

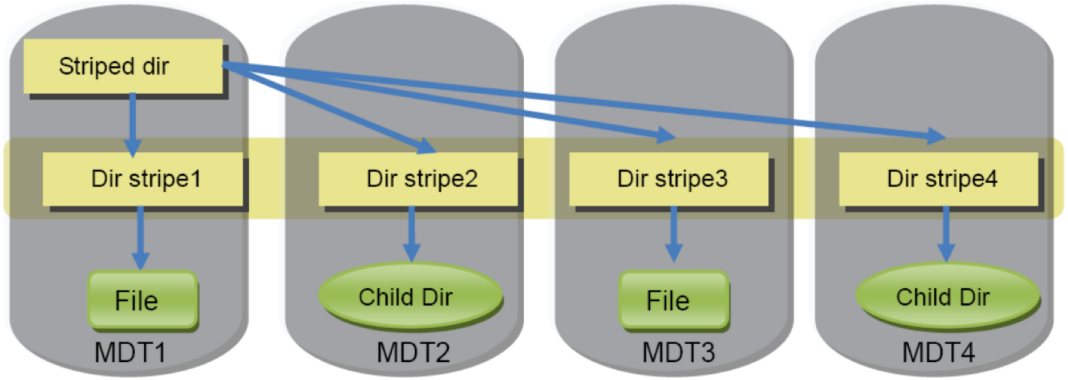
Lustre Tuning/Configuration
Striped Directory for metadata performance

Whamcloud

- Lustre supports two type of Distributed metadata
  - Remote Directory(DNE1) and Striped Directory(DNE2)



Remote Directory



Striped Directory

  - Single MDS and MDT by default
  - Enable DNE with “lfs setdirstripe” command

    # lfs setdirstripe -c 4 /scratch0/io500

    # lfs setdirstripe -c 4 -D /scratch0/io500

Lustre Tuning/Configuration
Other Tips

- Client Side setting

```
# lctl set_param \
osc.*.max_pages_per_rpc=16M \
osc.*.max_rpcs_in_flight=16 \
llite.*.max_read_ahead_mb=2048 \
osc.*.checksums=0
```

Sending more aggressive RPCs to server and readahead

- Server Side setting

```
# lctl set_param \
osd-ldiskfs.*.read_cache_enable=0
obdfilter.*.writethrough_cache_enable=0
```

Avoiding Page Cache on OSS when it flush data to flush

That’s all changed parameters and configuration from default Exascaler setting.

DDN IO-500 Exascaler 5.0 (PRE)

Score for 10 Client Configuration Challenge

Whamcloud

```
[RESULT] BW   phase 1            ior_easy_write              37.540 GB/s : time 343.38 seconds
[RESULT] IOPS phase 1         mdtest_easy_write             199.685 kiops : time 325.87 seconds
[RESULT] BW   phase 2            ior_hard_write               0.262 GB/s : time 300.21 seconds
[RESULT] IOPS phase 2         mdtest_hard_write              24.348 kiops : time 395.55 seconds
[RESULT] IOPS phase 3                      find            3332.110 kiops : time  21.96 seconds
[RESULT] BW   phase 3             ior_easy_read              35.374 GB/s : time 364.41 seconds
[RESULT] IOPS phase 4          mdtest_easy_stat             527.669 kiops : time 124.08 seconds
[RESULT] BW   phase 4             ior_hard_read               4.627 GB/s : time  17.03 seconds
[RESULT] IOPS phase 5          mdtest_hard_stat              79.476 kiops : time 106.64 seconds
[RESULT] IOPS phase 6        mdtest_easy_delete             226.094 kiops : time 288.22 seconds
[RESULT] IOPS phase 7          mdtest_hard_read              46.141 kiops : time 182.72 seconds
[RESULT] IOPS phase 8        mdtest_hard_delete              58.842 kiops : time 143.64 seconds
[SCORE] Bandwidth 6.33725 GB/s : IOPS 159.413 kiops : TOTAL 31.7843
```

Conclusions

Whamcloud

- IO-500 on Exascaler/Lustre was one of good exercise
  - Good opportunity try Lustre DNE2 out and confirmed it worked very well.
  - Happy to test with DoM(Data on Metadata) or other new Lustre features next time. ☺
  - Might be good a regression benchmark toolset as well.
- More normalization would be nice to have
  - 10 client challenge was good start, but doesn't see storage side configuration detail.
  - At least, we may see "score / Total aggregated H/W resource"
  - Good motivation on improving efficiency and eventually it could be good score/cost.