OAK RIDGE National Laboratory

Tuning the IO-500 benchmark on Summit

George S. Markomanolis,

Dustin Leverman,

Sarp Oral

Oak Ridge National Laboratory

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

Summit – Spectrum Scale

- 154 data servers which are also metadata servers
- 211 NLSAS drives and only 4GB non volatile memory per data server
- Spectrum Scale 5.x

Tuning I

- IOR hard cases provide low results
- IOR easy and the metadata results seem good
- Trying to tune IO-500 for better IOR hard results

Tuning II – Initial effort

- **160** compute nodes, **16** MPI processes per node

[RESULT] BW phase 1	ior_easy_write	**1786.080** GB/s : time 365.16 seconds
[RESULT] BW phase 2	ior_hard_write	**0.021** GB/s : time 1066.83 seconds
[RESULT] BW phase 3	ior_easy_read	1212.820 GB/s : time 537.75 seconds
[RESULT] BW phase 4	ior_hard_read	**3.506** GB/s : time 6.39 seconds
[RESULT] IOPS phase 1	mdtest_easy_write	**13342.900** kiops : time 357.11 seconds
[RESULT] IOPS phase 2	mdtest_hard_write	**10553.100** kiops : time 308.06 seconds
[RESULT] IOPS phase 3	find	3029.010 kiops : time 41.21 seconds
[RESULT] IOPS phase 4	mdtest_easy_stat	**19455.700** kiops : time 10.96 seconds
[RESULT] IOPS phase 5	mdtest_hard_stat	6020.080 kiops : time 4.48 seconds
[RESULT] IOPS phase 6	mdtest_easy_delete	1746.550 kiops : time 89.84 seconds
[RESULT] IOPS phase 7	mdtest_hard_read	**18049.300** kiops : time 3.98 seconds
[RESULT] IOPS phase 8	mdtest_hard_delete	46.578 kiops : time 206.28 seconds

[SCORE] Bandwidth **19.9883** GB/s : IOPS **4056.73** kiops : TOTAL **284.758**

Tuning III – IOR Hard

- Increasing number of segments helps a bit
- Decreasing the number of processes per node gives a significant boost

Final results

- **504** compute nodes, **2** MPI processes per node

[RESULT] BW phase 1	ior_easy_write	**2158.700** GB/s : time 362.34 seconds
[RESULT] BW phase 2	ior_hard_write	**0.572** GB/s : time 462.76 seconds
[RESULT] BW phase 3	ior_easy_read	1788.320 GB/s : time 437.39 seconds
[RESULT] BW phase 4	ior_hard_read	**27.403** GB/s : time 9.66 seconds
[RESULT] IOPS phase 1	mdtest_easy_write	3071.260 kiops : time 352.36 seconds |
[RESULT] IOPS phase 2	mdtest_hard_write	24.375 kiops : time 327.20 seconds
[RESULT] IOPS phase 3	find	**21780.030** kiops : time 46.35 seconds
[RESULT] IOPS phase 4	mdtest_easy_stat	**28769.400** kiops : time 52.52 seconds
[RESULT] IOPS phase 5	mdtest_hard_stat	560.886 kiops : time 19.36 seconds
[RESULT] IOPS phase 6	mdtest_easy_delete	2699.000 kiops : time 398.34 seconds
[RESULT] IOPS phase 7	mdtest_hard_read	15354.700 kiops : time 17.02 seconds
[RESULT] IOPS phase 8	mdtest_hard_delete	26.504 kiops : time 181.79 seconds

[SCORE] Bandwidth **88.2049** GB/s : IOPS **1522.68** kiops : TOTAL **366.48**

Conclusions

- Although metadata performance decreased with less MPI processes per node, the increase of IOR hard results was significant
- SSDs are important for IOR Hard
- Calibrating the MPI processes per node, led to the best result
- Whole procedure lasted for 10 hours
- A lot of data from testing cases, ~10 PB

Acknowledgement

This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725.

OAK RIDGE National Laboratory