IO500 Performance Report

Silverton Consulting, Inc. StorInt™ Dispatch

This Storage Intelligence (StorInt™) dispatch covers the IO500 benchmark[1]. As you should recall, the IO500 is focused on HPC (high performance computing) workload file IO. Unlike other file benchmarks (using NFS and SMB), most IO500 submissions use POSIX file systems which require client software to access their file systems. The IO500 supports two rankings one that allows submissions with any number of (client) nodes and the other with only 10 client nodes.

The Virtual Institute for IO (VI4IO), the group that organizes the IO500 benchmarks, ranks submissions using a score that is a function of 4 IOR bandwidth intensive workloads (easy read, easy write, hard read, and hard write) and 8 mdtest metadata intensive workloads (easy write, stat, & delete, hard write, stat, delete, & read, and easy find). TheIO500 IOR set of benchmarks simulate big block, bandwidth intensive (traditional HPC) file IO activity and mdtest set of benchmarks simulate small block, IO activity. Both are used to rank systems IO performance.

IO500's overall score is a composite (geomean) of scores on IOR bandwidth (easy and hard) and mdtest (easy, hard and easy find) metadata intensive workloads. Similarly, the individual IOR and mdtest scores are composites of their individual workloads

IO500 any number of client node results

We start our discussion with the overall composite scores as reported by IO500 for submissions with any number of client nodes, in Figure 1.

[1] All IO500 information is available at https://www.vi4io.org/std/io500/start as of 27 July 2021

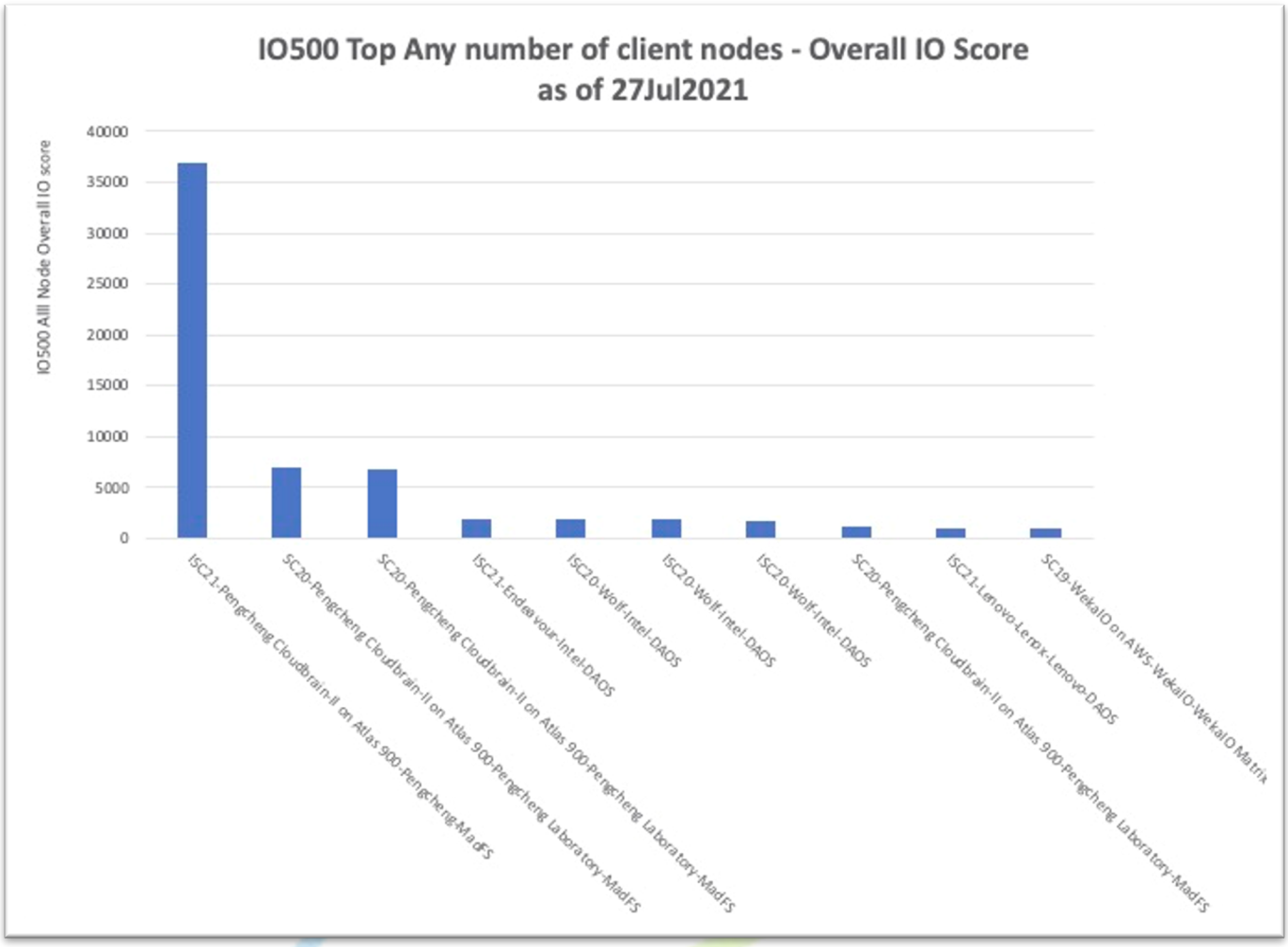

Figure 1 Top 10 IO500 Any number of client nodes overall score results

In Figure 1, the Pengcheng CloudBrain-II MadFS system came in first, with a composite score of ~37K, ~5.2X the nearest (last year's Pengcheng) competitor, using 512 client nodes and 512 storage nodes over 100GbE networking. There was no information provided on the storage media used in this submission. We can only assume that it was either substantially faster than last year's media (6-2TB NVMe SSDs/storage node), possibly Optane PMEM or there were more drives per storage node.

The #2 and #3 results were also Pengcheng submissions (from SC20) using 256 and 255 client nodes with 254 and 256 storage nodes, respectively and 100GbE. In addition, the #8 result was also Pengcheng (from SC20) and was their 10-client node system we discussed in our November IO500 performance report.

After the top three, we have a new Endeavour-Intel-DAOS submission at #4 which used 10 client nodes and 40 storage nodes with 1TB Optane PMEM per node over (200Gbps) Infiniband networking. After that, there are 4 more Wolf-Intel-DAOS submissions, the first 3 used 512GB Optane PMEM with 30 storage nodes each and 10, 52, and 52 client nodes over

Omnipath networking. The last DAOS submission here used 36 client nodes and 12 storage nodes with 256GB of Optane PMEM per storage node over (100Gbps) Infiniband networking

The only other new vendor on this ranking was WekaIO coming in at #10, running 345 client nodes and 120 storage nodes with NVMe SSDs all running on AWS. It's very unusual to see any storage benchmark submissions using AWS infrastructure (all the other submissions were in national labs). This just points to some interesting characteristics of WekaIO and its ability to manage high levels of IO response time and bandwidth variability.

In Figure 2 we show the top 10 IOR scores for any number of client nodes.

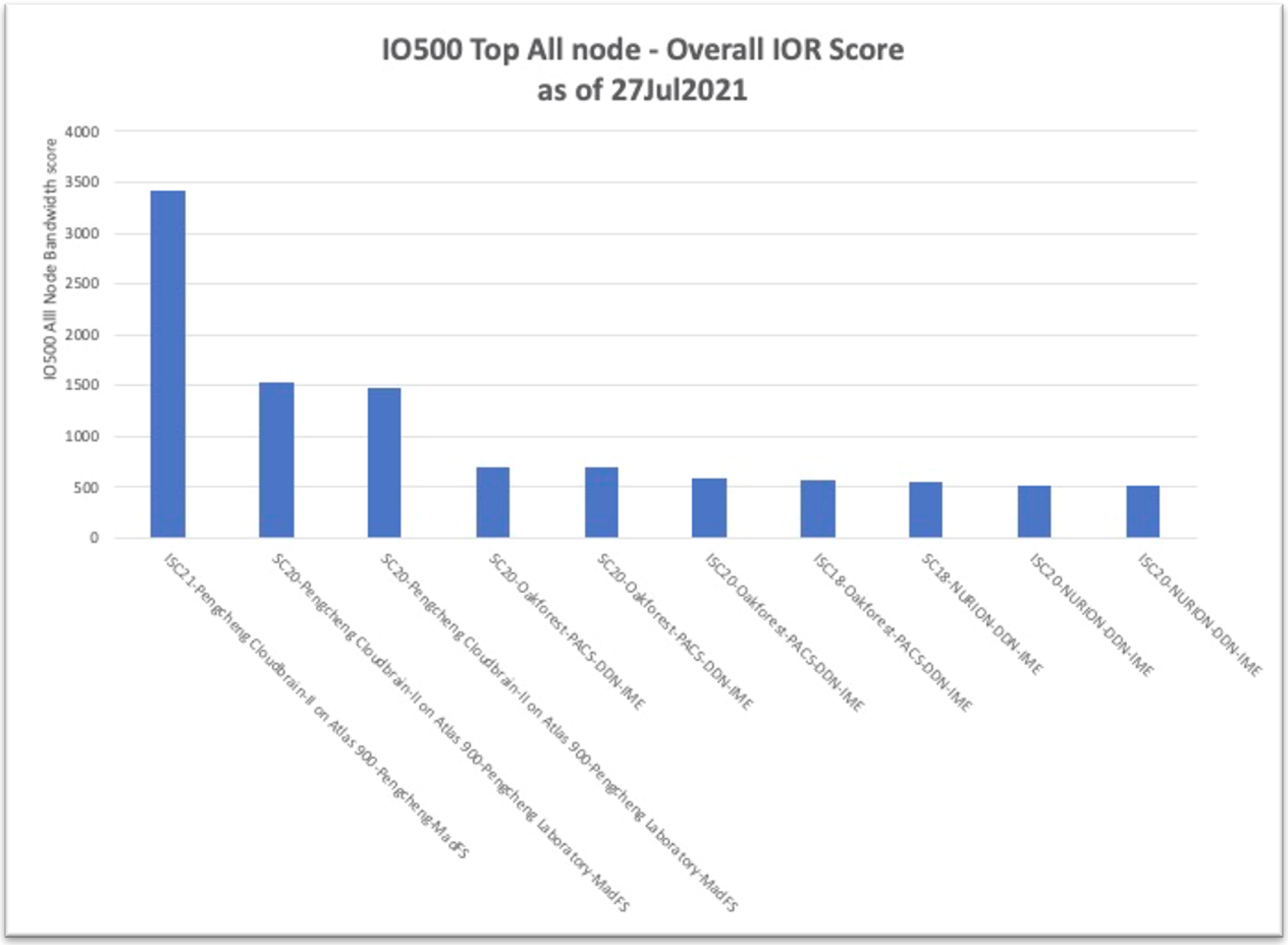

Figure 2 IO500 top 10 IOR score results

Here too, the latest Pengcheng submission came in at #1 with a IOR score of ~3.4K. It is interesting that the #1 Pengcheng system performed 4.9X the nearest non-Pengcheng submission.

Similarly, #2 and #3 IOR ranked submissions were older Pengcheng systems with IOR scores of ~1.5K each. We have discussed the system characteristics above with one difference the #2

submission had 256 clients with 256 storage nodes and #3 system had 254 client and 255 storage nodes (opposite of their respective rankings in Figure 1s overall score).

The remainder of these rankings all used DDN-IME filesystems at either Oakforest or NURIAN (South Korea). All these systems used 2048 client nodes over 25 (Oakforest) or 48 (NURIAN) storage nodes with NVMe SSDS over Omnipath networking. (Note, the storage node, media and networking information for the #8 Oakforest and #9-10 NURIAN systems are missing but here we assume they are the same as the other submissions from these labs). All the DDN IME submissions seemed to perform similarly on the IOR benchmarks all under 700 IOR score. Moreover, none of these DDN-IME submissions ranked in the top ten for the overall score.

In Figure 3, we show the IO500 top 10 mdtest overall score rankings.

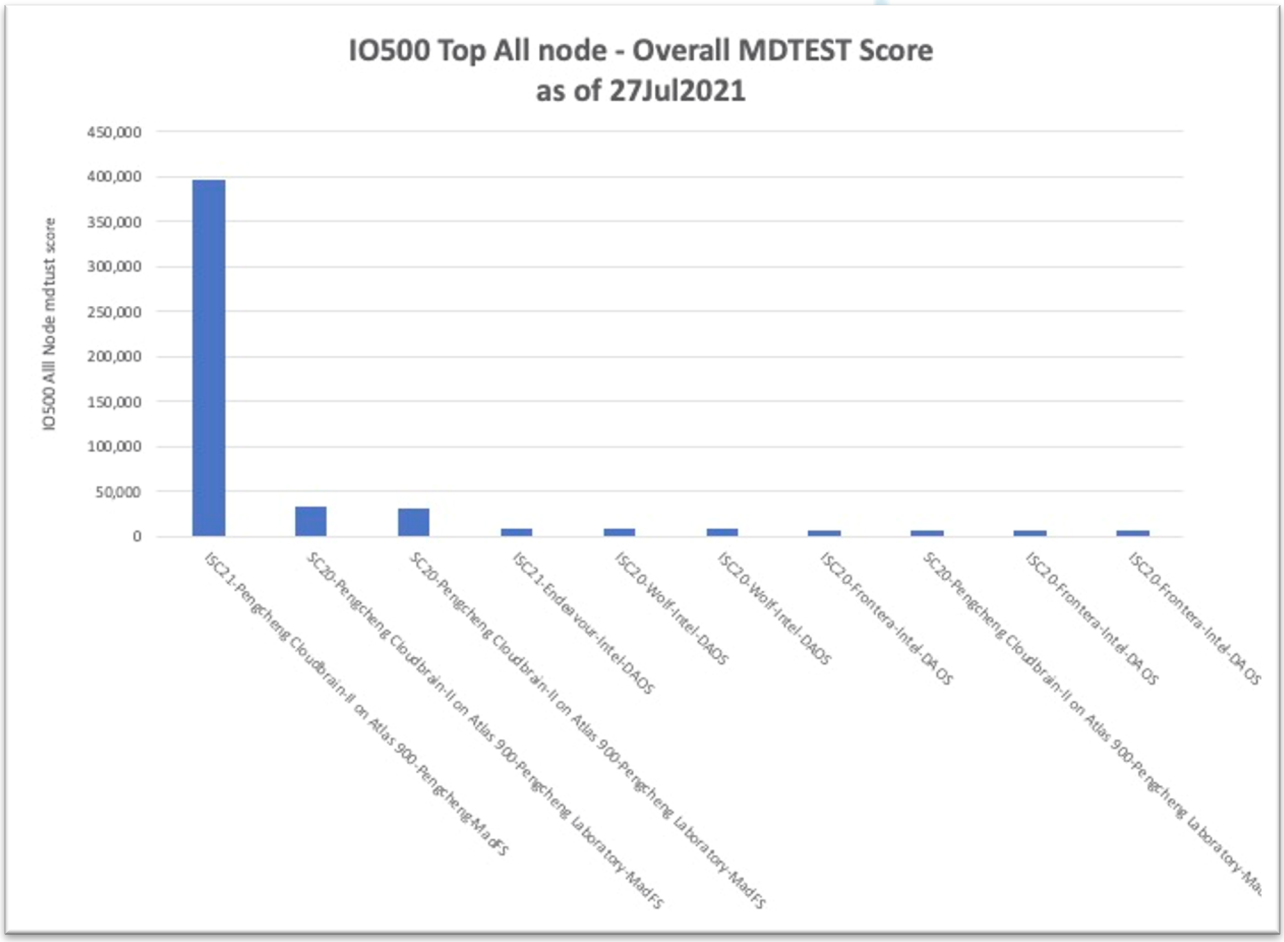

Figure 3 IO500 Top 10 mdtest score results

We almost need a logarithmic scale for the y-axis here. Again, the #1 system is the latest Pengcheng submission (ISC21) with an overall mdtest score of ~37.0K or ~19.8X better than the nearest non-Pengcheng competitor (Intel-DAOS @#4). We have discussed all these systems earlier. But whatever Pengcheng MadFS is doing, it is working exceptionally well for small block IO.

IO500 10 client node results

Alas, there were no new 10 client node submissions, so there is no news to talk about here. We would say we see a similar picture in the 10 client node results from last year. That is, the Pengcheng 10 client node was #1 in overall and mdtest rankings. In the mdtest overall score it only managed 2.2X better than the non-Pengcheng, nearest DAOS competitor. It failed to come in as #1 in the IOR benchmark rankings, which was a Frontera-DDN-IME submission, but the 10-client node Pengcheng system was close (168 vs 176) to the #1 IOR 10 client node system.

Significance

IO500 HPC focused benchmarks differ substantially from other file system benchmarks we report on. For one thing, client node counts become important and for another file bandwidth is an equal factor to file IOPS. IO500 updates their rankings, twice a year (during ISC & SC conferences) so we may see some new results later in 2021.

This is our third attempt at analyzing IO500 results. We continue to think it best to report standard IO500 metrics. However, we have identified a few different metrics to add to their numbers, using SCI computed numbers on their results, but will save those for a later report.

This report was sent out to subscribers as part our **free, monthly Storage Intelligence e-newsletter**. If you are interested in receiving future storage performance analyses along with recent product announcement summaries, please use the QR code (below right) to sign up for your own copy.

Silverton Consulting, Inc., is a U.S.-based Storage, Strategy & Systems consulting firm offering products and services to the data storage community

Newsletter signup QRcode