IO500: The High-Performance Storage Community

Committee

- Andreas Dilger - Whamcloud/DDN
- **Dean Hildebrand - Google**
- George Markomanolis - AMD
- Jay Lofstead - Sandia National Laboratories
- Jean Luca Bez - Lawrence Berkeley Lab
- Julian Kunkel - Georg-August-Universität Göttingen/GWDG

SC24
Atlanta, GA | hpc creates.

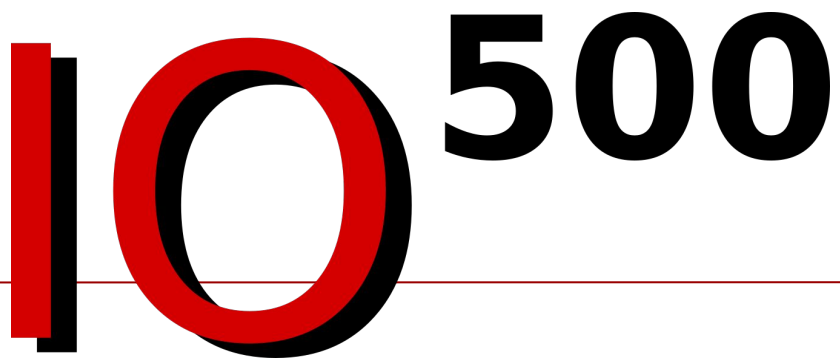
BoF Agenda

1. **Welcome** – Dean Hildebrand
2. **Award Presentations** – Dean Hildebrand
3. **New IO500 List Analysis** – George Markomanolis
4. **Community Talk**
   - *How the IO500 is improving Ceph*
     - Mark Nelson, Head of R&D at Clyso
5. **Updates**
   - **Website** - Jean Luca Bez
   - **Random Read and Steps for Inclusion** – Andreas Dilger
6. **Community Discussion** – Jay Lofstead

IO500 Organization Status

- A US non-profit, public charity organization: IO500 Foundation
  - Domain, mailing list, servers, GitHub belongs to IO500 Foundation
- Website contains results with links to details, CFS, BoF slides, etc.
  - io500.org
  - Contribute fixes at github.com/IO500/webpage
- Please join our mailing list for announcements:
  - io500.org/contact
- Please join our Slack for discussions:
  - io500workspace.slack.com/
  - Join link: rb.gy/sn8esm

Award Ceremony



No New Awards

11 New Production Submissions!

List of Awarded Systems in the Ranked Lists

10 Client Production	Bandwidth Metadata Overall	**Argonne National Laboratory**	DAOS	734,50 11,336.72 2,885.57	GB/s KIOPS/s score
10 Client Research	Bandwidth Metadata Overall	**JNIST and HUST PDSL**	OceanFS2	2,439.37 7,705,448.04 137,100.00	GB/s KIOPS/s score
Production	Bandwidth Metadata Overall	**Argonne National Laboratory**	DAOS	10,066.09 102,785.11 32,165.90	GB/s KIOPS/s score
Research	Bandwidth Metadata Overall	**Argonne National Laboratory** **Pengcheng Laboratory** **Pengcheng Laboratory**	DAOS SuperFS SuperFS	6,048.49 9,119,612.35 210,255	GB/s KIOPS/s score

10 Client Node Production - Overall Winner

	# ↑	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE ↑	BW (GIB/S)	MD (KIOP/S)
	1	SC23	Aurora	Argonne National Laboratory	DAOS	2,885.57	734.50	11,336.27
	2	ISC23	SuperMUC-NG-Phase2-EC-10	LRZ	DAOS	1,008.81	218.38	4,660.23
	3	ISC24	Lise	Zuse Institute Berlin	DAOS	324.54	65.01	1,620.13
New	4	SC24	GEFION	Danish Centre for AI innovation AS	EXAScaler	314.03	154.70	637.43
New	5	SC24	CHIE-2	SoftBank Corp	EXAScaler	299.32	159.93	560.19
New	6	SC24	HiPerGator AI	University of Florida	EXAScaler	243.61	124.89	475.20
	7	ISC24	NHN CLOUD GWANGJU AI	NHN Cloud Corporation	EXAScaler	176.57	62.58	498.22
	8	ISC24	Afton	University of Virginia	WEKA	105.94	33.28	337.29
	9	SC23	Earth Simulator 4	Japan Agency for Marine-Earth Science and Technology	EXAScaler	101.88	48.19	215.38
New	10	SC24	Proxima	Poznan Supercomputing and Networking Center	Lustre	77.76	17.94	337.05

Certificate

IO500 Performance Certification

This Certificate is awarded to:

Argonne National Laboratory
(Aurora DAOS EC)

#1 in the 10 Client Node Production Overall, Bandwidth and Metadata Score

IO500

Nov 2025

IO500 Steering Board

https://io500.org/list/SC24/ten-production

10 Client Node Production - Bandwidth Winner

Sort by BW

	#	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	BW ↑ (GIB/S)
	1	SC23	Aurora	Argonne National Laboratory	DAOS	734.50
	2	ISC23	SuperMUC-NG-Phase2-EC-10	LRZ	DAOS	218.38
New	3	SC24	CHIE-2	SoftBank Corp	EXAScaler	159.93
New	4	SC24	GEFION	Danish Centre for AI innovation AS	EXAScaler	154.70
New	5	SC24	HiPerGator AI	University of Florida	EXAScaler	124.89
	6	ISC24	Lise	Zuse Institute Berlin	DAOS	65.01
	7	ISC24	NHN CLOUD GWANGJU AI	NHN Cloud Corporation	EXAScaler	62.58
	8	SC23	Earth Simulator 4	Japan Agency for Marine-Earth Science and Technology	EXAScaler	48.19
	9	ISC24	Afton	University of Virginia	WEKA	33.28
	10	SC23	Randi	Center for Research Informatics at University of Chicago	Spectrum Scale	31.05

Certificate

IO500 Performance Certification

This Certificate is awarded to:

JNIST and HUST PDSL (Cheeloo-1)
with OceanStor Pacific from Huawei
#1 in the 10 Client Node Research Overall Score

IO500

Nov 2024

IO500 Steering Board

https://io500.org/list/SC24/ten

IO500 Production List - 8 New Entries from 4 FSs

	# ↑	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE ↑	BW (GIB/S)	MD (KIOP/S)
	1	SC23	Aurora	Argonne National Laboratory	DAOS	32,165.90	10,066.09	102,785.41
	2	SC23	SuperMUC-NG-Phase2-EC	LRZ	DAOS	2,508.85	742.90	8,472.60
	3	SC23	Shaheen III	King Abdullah University of Science and Technology	Lustre	797.04	709.52	895.35
New	4	SC24	IRIS	MSKCC	WekaIO	665.49	252.54	1,753.69
	5	ISC23	Leonardo	EuroHPC-CINECA	EXAScaler	648.96	807.12	521.79
New	6	SC24	CHIE-3	SoftBank Corp	EXAScaler	500.20	331.66	754.41
New	7	SC24	GEFION	Danish Centre for AI innovation AS	EXAScaler	368.56	209.06	649.73
	8	ISC24	Lise	Zuse Institute Berlin	DAOS	324.54	65.01	1,620.13
New	9	SC24	HiPerGator AI	University of Florida	EXAScaler	243.61	124.89	475.20
	10	ISC22	CTPAI	China Telecom Research Institute	DAOS	187.84	25.29	1,395.01

IO500 Production List - Bandwidth

Sorted by BW

	#	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	BW ↑ (GIB/S)
	1	SC23	Aurora	Argonne National Laboratory	DAOS	10,066.09
	2	ISC23	Leonardo	EuroHPC-CINECA	EXAScaler	807.12
	3	SC23	SuperMUC-NG-Phase2-EC	LRZ	DAOS	742.90
	4	SC23	Shaheen III	King Abdullah University of Science and Technology	Chandelier	709.52
New	5	SC24	CHIE-3	SoftBank Corp	EXAScaler	331.66
New	6	SC24	IRIS	MSKCC	WekaIO	252.54
New	7	SC24	GEFION	Danish Center for AI innovation AS	EXAScaler	209.06
New	8	SC24	HiPerGator AI	University of Florida	EXAScaler	124.89
	9	ISC24	Helios	ACC Cyfronet AGH	Chandelier	122.31
	10	ISC24	Lise	Zuse Institute Berlin	DAOS	65.01

Certificate

IO500 Performance Certification

This Certificate is awarded to:

Argonne National Laboratory
(Aurora DAOS EC)

#1 in the IO500 Production Overall, Bandwidth, and Metadata Score

IO500

Nov 2024

IO500 Steering Board

https://io500.org/list/SC24/production

IO500 Research List - Bandwidth Winner

Sorted by BW

#	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE	BW ↑ (GIB/S)	MD (KIOP/S)
1	SC22	Aurora Storage	Argonne National Laboratory	DAOS		6,048.69	
2	ISC23	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng Laboratory	SuperFS		4,847.48	
3	ISC23	Cheeloo-1 with OceanStor Pacific	JNIST and HUST PDSL	OceanFS2		2,439.37	
4	ISC23	Leonardo	EuroHPC-CINECA	EXA6		807.12	
5	SC22	ParaStor	Sugon Cloud Storage Laboratory	ParaStor		718.11	
6	SC20	Oakforest-PACS	JCAHPC	IME		697.20	
7	ISC20	NURION	Korea Institute of Science and Technology Information (KISTI)	IME		515.59	
8	SC22	StarStor	SuPro Storteck	StarStor		515.15	
9	ISC23	SuperMUC-NG-Phase2	LRZ	DAOS		433.05	
10	ISC21	Endeavour	Intel	DAOS		398.77	

IO500 Research List - Overall Winner

# ↑	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE ↑	BW (GIB/S)	MD (KIOP/S)
1	ISC23	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng Laboratory	SuperFS	210,255.00	4,847.48	9,119,612.35
2	ISC23	Cheeloo-1 with OceanStor Pacific	JNIST and HUST PDSL	OceanFS2	137,100.00	2,439.37	7,705,448.04
3	SC22	Aurora Storage	Argonne National Laboratory	DAOS	20,694.50	6,048.69	70,802.51
4	SC22	ParaStor	Sugon Cloud Storage Laboratory	ParaStor	8,726.42	718.11	106,042.93
5	SC22	StarStor	SuPro Storteck	StarStor	6,751.75	515.15	88,491.65
6	SC22	SuperStore	Tsinghua Storage Research Group	SuperFS	5,517.73	179.60	169,515.95
7	ISC22	Shanhe	National Supercomputing Center in Jinan	flashfs	3,534.42	207.79	60,119.50
8	SC22	HPC-OCI	Cloudam HPC on OCI	BurstFS	3,033.03	278.48	33,033.54
9	SC21	Athena	Huawei HPDA Lab	OceanFS	2,395.03	314.56	18,235.71
10	SC21	OceanStor Pacific	Olympus Lab	OceanFS	2,298.69	317.07	16,664.88

Certificate

IO500 Performance Certification

This Certificate is awarded to:

Pengcheng Laboratory (Cloudbrain-II)
with SuperFS from Tsinghua University
#1 in the IO500 Research Overall Score

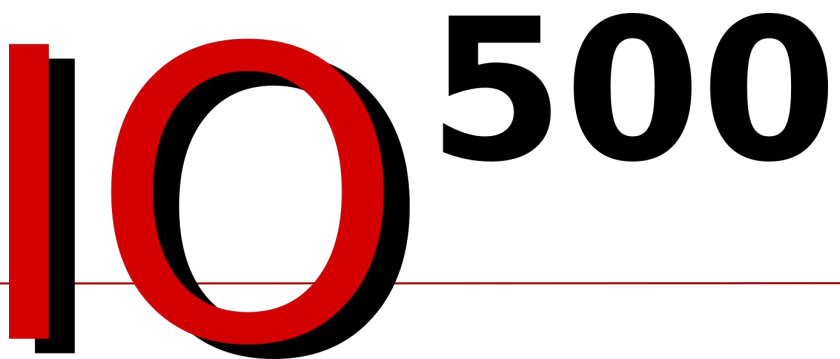
IO500

Nov 2024

IO500 Steering Board

https://io500.org/list/SC24/io500

IO500 List Analysis



IO500 List - Growth in Entries and Institutions

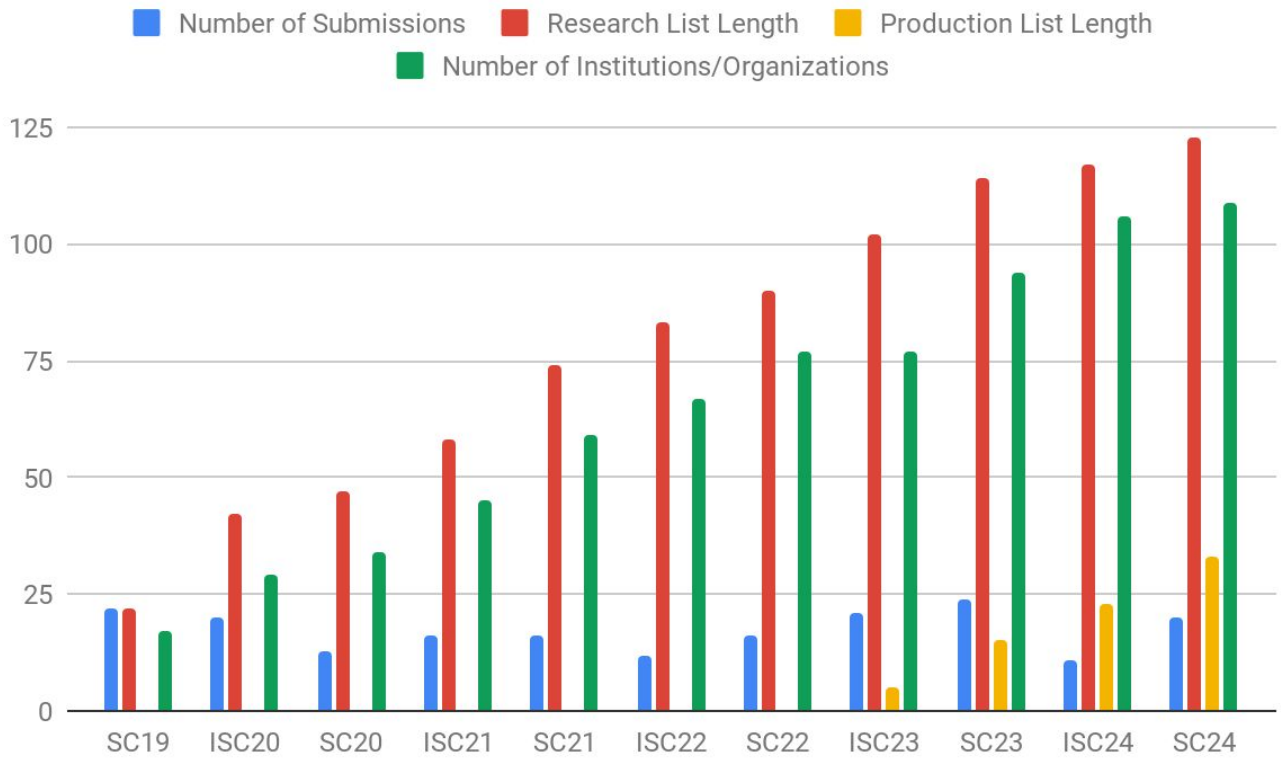
SC24

25 submissions (4 rejected)

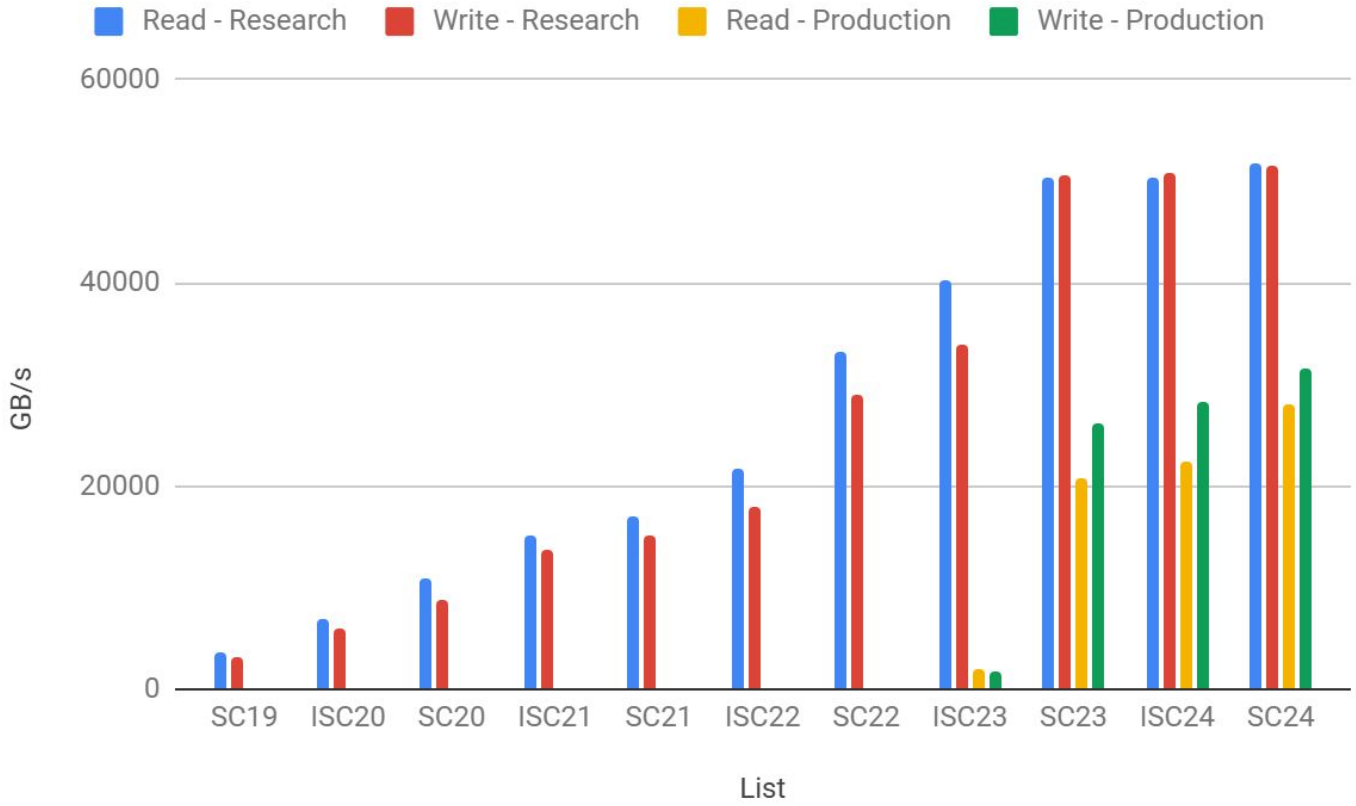
- 5 for 10-Client Research
- 7 for 10-Client Production
- 6 for IO500 Research
- 11 for IO500 Production
- 1 for Full (< 10 client nodes)

Around 268 list entries

More than 100 institutions

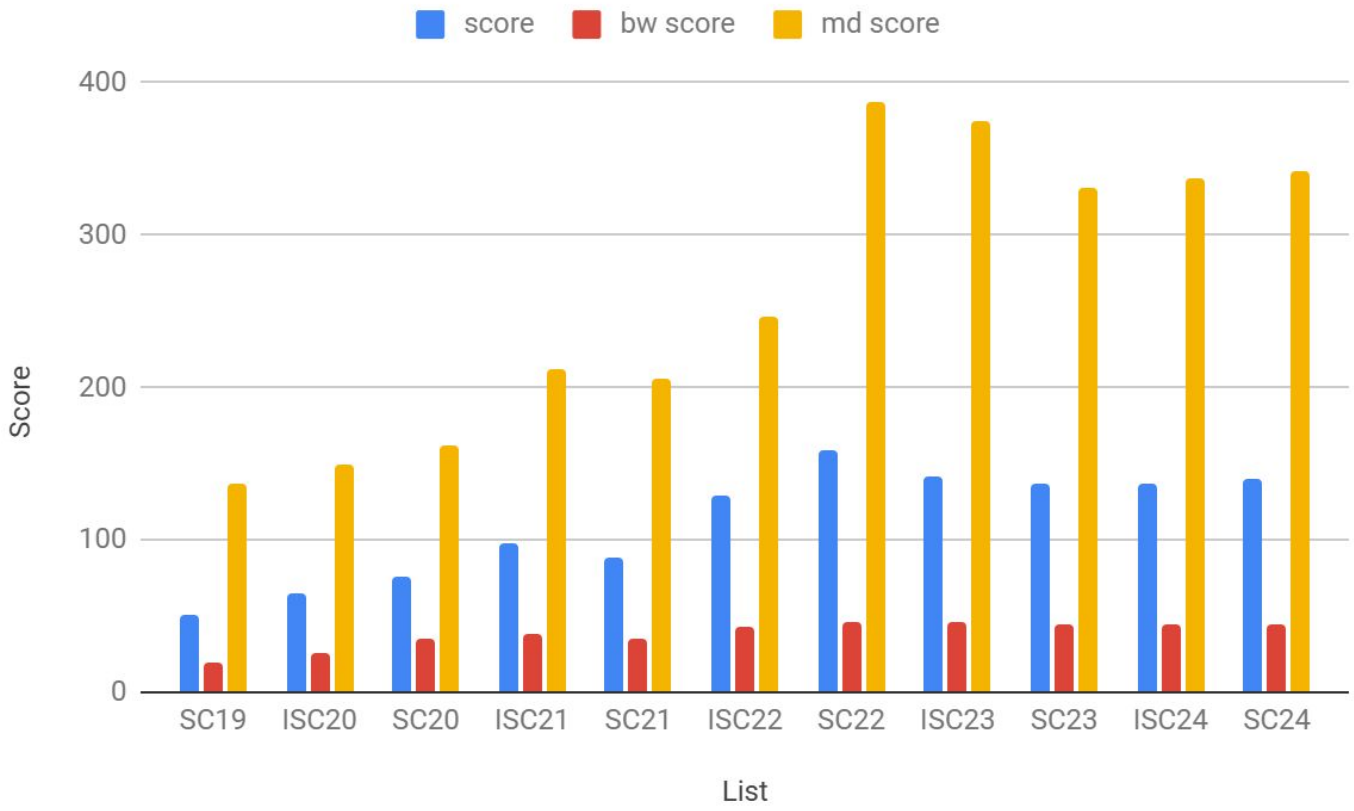


IO500 List - Aggregate List Bandwidth

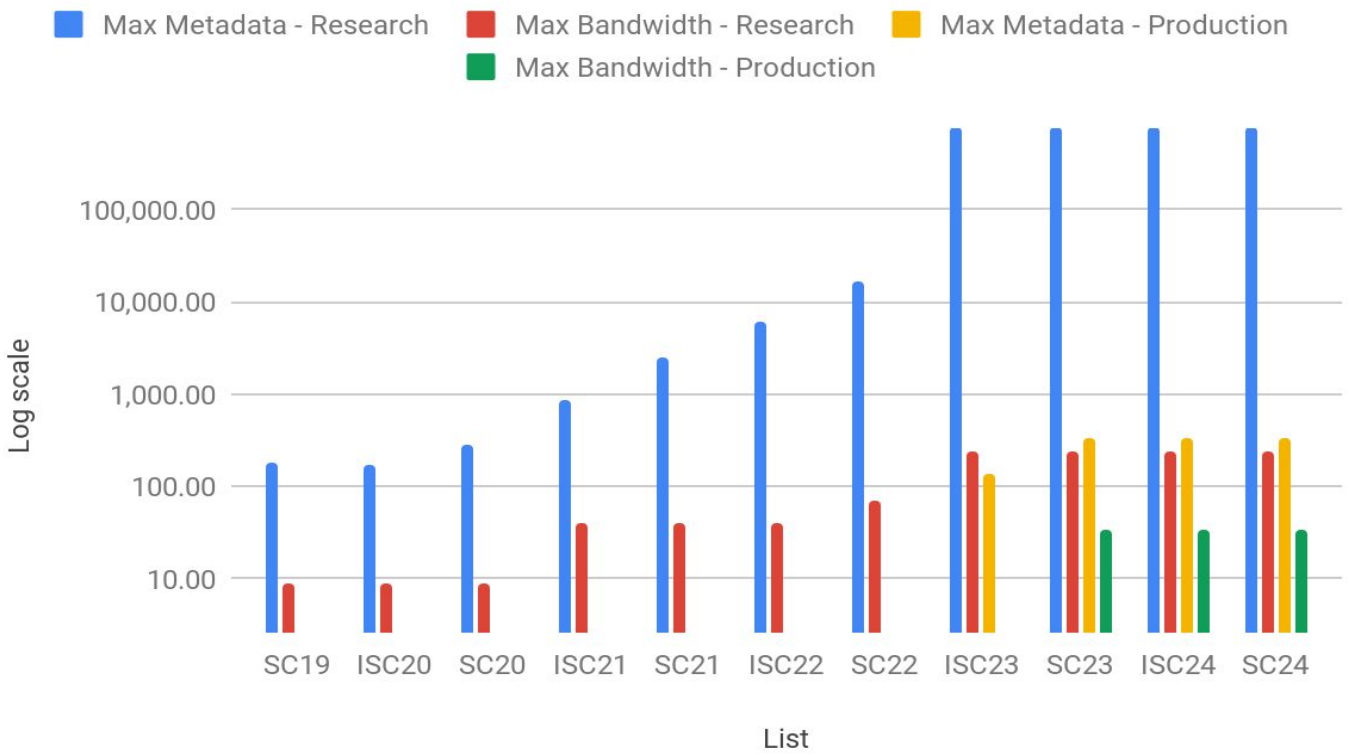


IO500 List - Median Scores

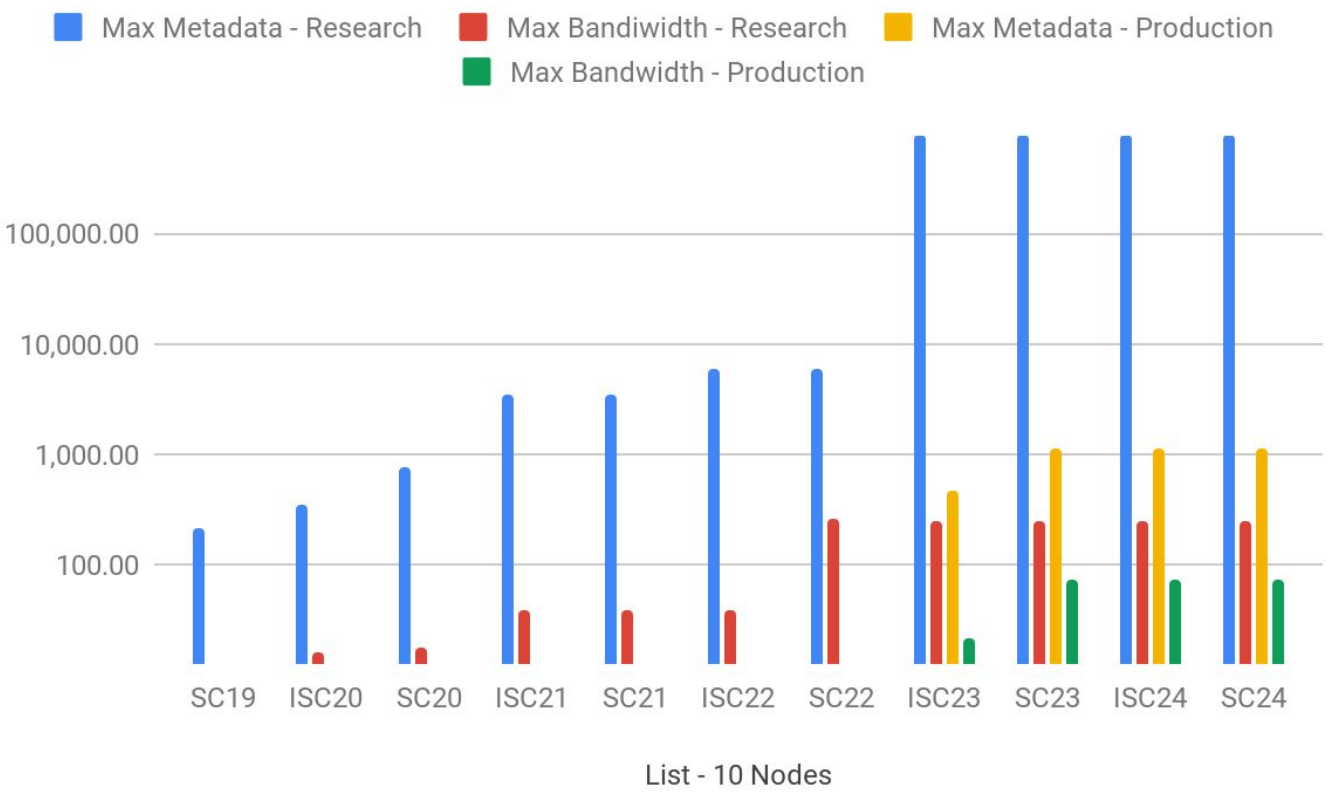
Median scores are mixed compared to SC22



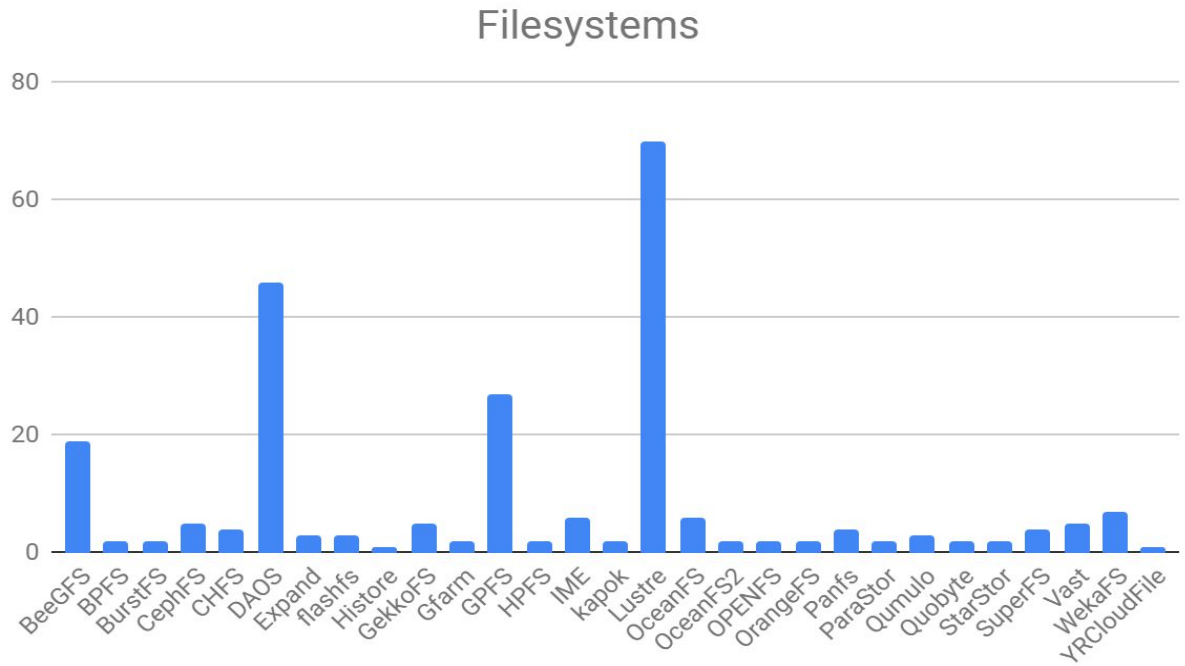
IO500 List - Growth in Max Score per Client



10-Client List - Growth in Max Scores per Client

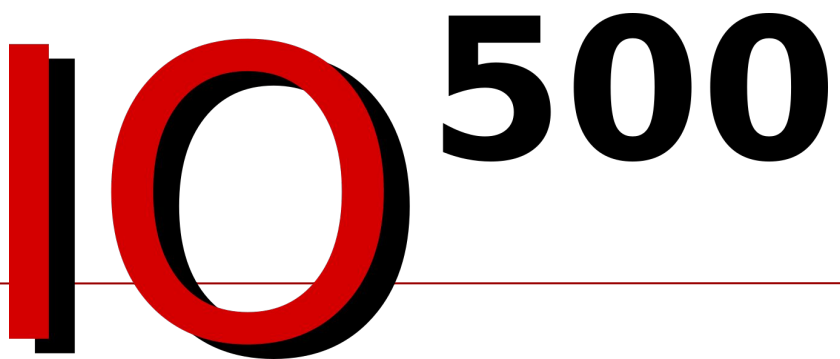


IO500 List - Number of File System Entries



Lustre and DAOS among many systems, followed by GPFS and BeeGFS

Community Talk



How the IO500 is improving Ceph

Mark Nelson mark.nelson@clyso.com

11/19/2024

CLYSO

How do we typically analyze ceph performance?

- ✓ Modular cluster deployment tool and remote-execution framework called “CBT”.
- ✓ Collection of benchmarks ranging from fio, elbencho, hsbench, and others.
- ✓ CPU profiling via perf, custom libunwind / libdw wallclock profiling, and “uncore” profiling via Intel and AMD tools.

CL'SO

What makes the IO500 special?

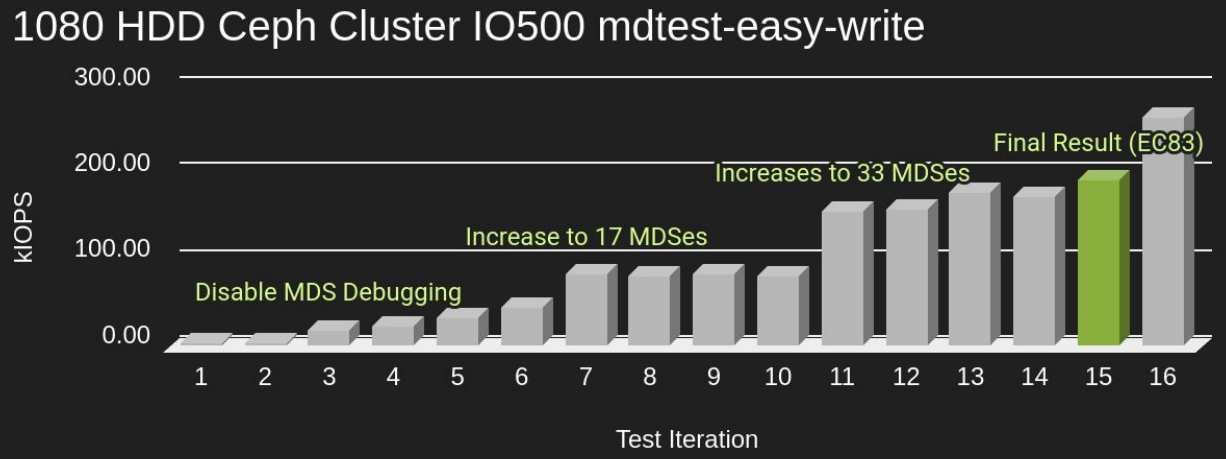
- ✓ Both throughput (IOR) and metadata (mdtest) orchestrated via MPI.
- ✓ Unique focus on difficult situations including client contention and unaligned writes.
- ✓ Able to test POSIX via CephFS kernel or fuse clients and libcephfs API access via a custom AIORI backend similar to DAOS.

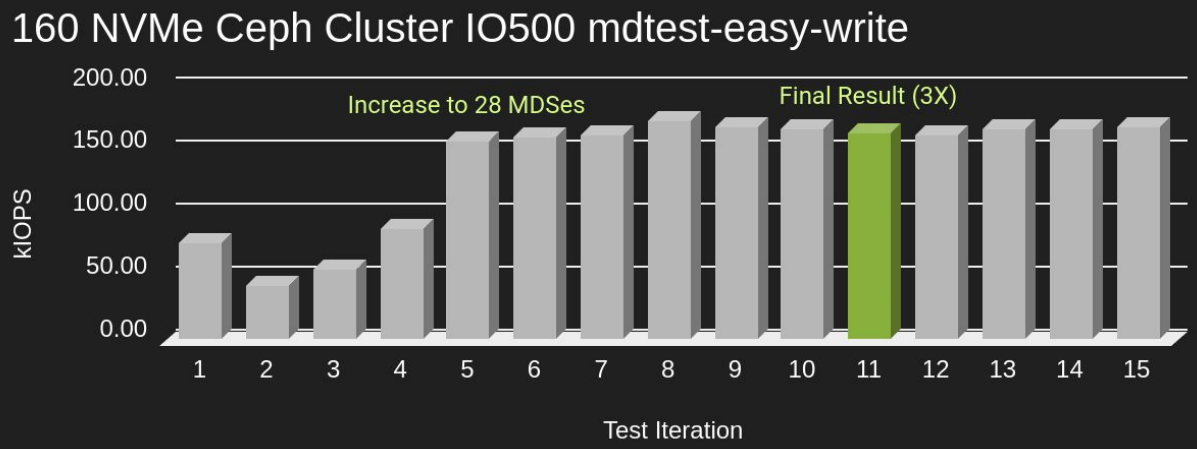
CL`I`SO

How do we use it?

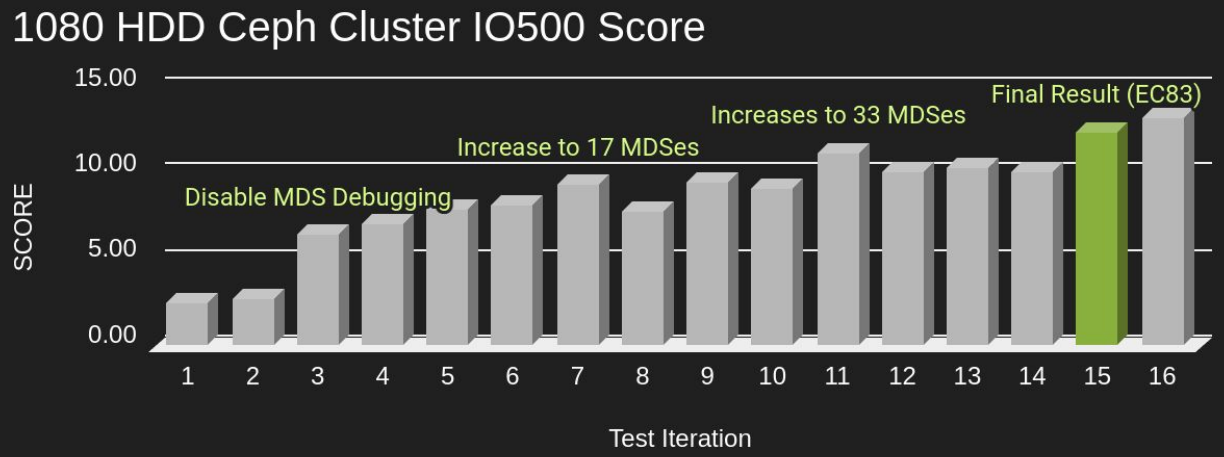
1. Tuning file system performance (especially around metadata) for customers.
2. Planning the development and prototyping of new performance features.
3. Evaluating and potentially bisecting performance regressions in new releases.

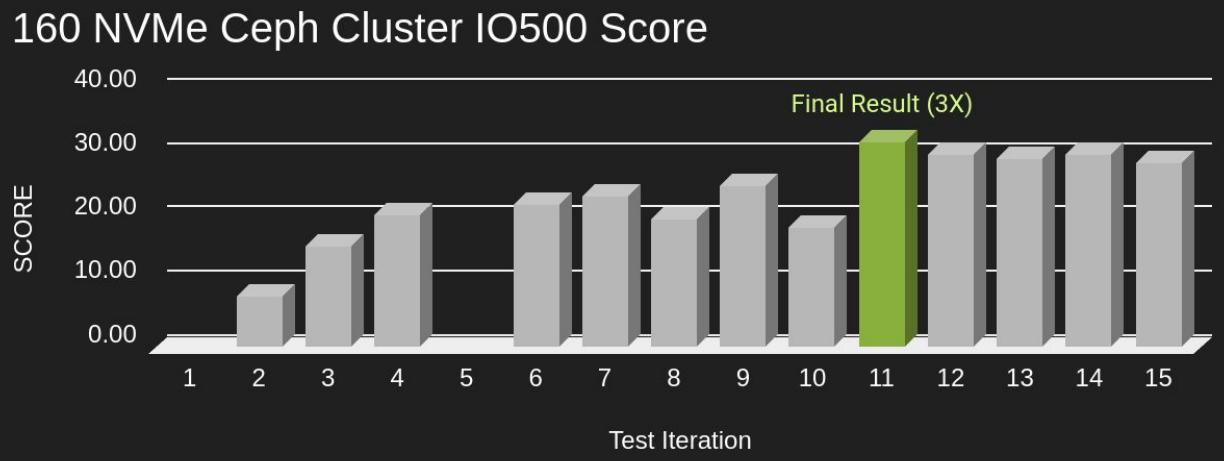
Case Study: GWDG





Case Study: GWDG





Tricky IO500 Corner-case

- ✓ Ceph scales well in mdtest-easy by simply adding more MDSes to pin directories to.
- ✓ To achieve scaling in mdtest-hard, ceph must use dynamic subtree partitioning.
- X Alternating between easy and hard tests can cause performance problems! MDSes may migrate responsibility for existing hard data during subsequent easy tests.

CL`,`SO

New feature in Ceph Reef: bal_rank_mask

Split the active metadata daemons between pinning and dynamic subtree partitioning.

Static Pinning (Good for mdtest-easy)	Dynamic Subtree Partitioning (Good for mdtest-hard)
MDS 2	MDS N+1
MDS 3	MDS N+2
MDS 4	MDS N+3
MDS N	MDS N+M

CL'SO

Regression testing for Squid

Using the IO500, we identified areas where CephFS regressed in the Squid RC release:

Test	Reef	Squid RC
mdtest-easy-write	127.72	38.00
mdtest-hard-write	22.09	6.09
mdtest-hard-stat	119.67	35.68
mdtest-hard-delete	36.11	5.58

CL'SO

Bisecting the regression

Version	Step	mdtest-easy-write
18.2.2	1	127.72
v19.0.0-898	2	130.89
v19.0.0-1700	3	136.30
v19.0.0-1762	7	133.60
v19.0.0-1794	6	Stalled (fast?)
v19.0.0-1806	11	Stalled (fast?)
v19.0.0-1807	14	39.03
…	…	…

CL'SO

Uncovering the bug

A wallclock profile during the test showed increased lock contention, which was ultimately identified in the offending commit and fixed:

```
Thread 140520 (quiesce_db_mgr) - 1000 samples
+ 100.00% clone
 + 100.00% start_thread
  + 100.00% QuiesceDbManager::quiesce_db_thread_main()
   + 89.80% pthread_cond_timedwait
   + 6.40% __pthread_mutex_unlock_usercnt
   |+ 6.40% __lll_unlock_wake
```

CL`SO

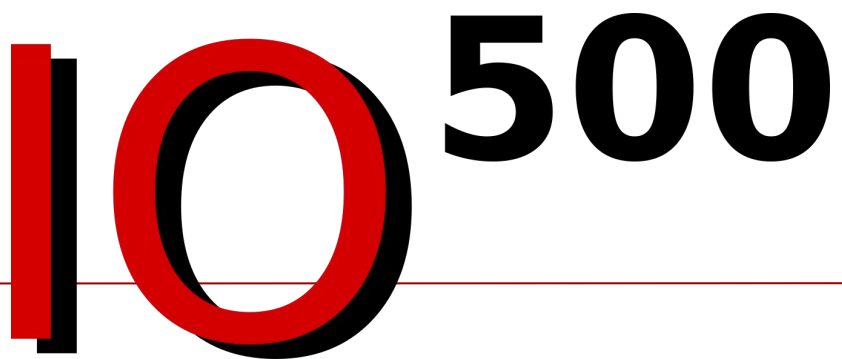
The IO500 is very useful for testing and improving Ceph!

Thank You!

Contact: mark.nelson@clyso.com

CLYSO

Updates



Latest IO500 Updates

- New Phases
  - Released random read proposal (discussed later in this session)
  - Still trying to define a ‘hard’ find phase
    - Need community input on what is ‘hard’
  - Will be removing all current optional phases when we add in random read phase
- Scoring
  - Metadata scores getting very large and overshadowing bandwidth due to geometric mean
    - Considering rebasing metadata scores from kIOPs to mIOPs, but affects rankings
- List Download
  - Some fields missing
  - Per-FS fields makes comparisons difficult, can we map to a common flat schema?
- `io500.org` submissions page
  - Please continue to give feedback
- Community meeting
  - Can be scheduled upon request

ISC 25 (June 10-13, 2025)

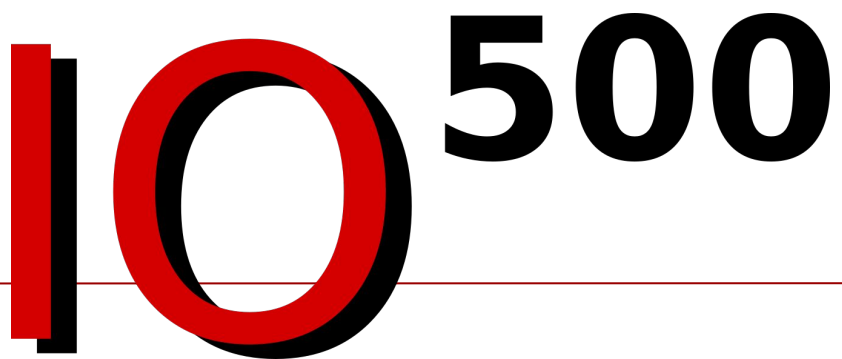
- Call for submission: April 7th
- Submission deadline: June 1st
- List release: BoF date TBD

Website Updates

Website Updates

- Migration to new framework version
  - Improve system stability and security
- Additional options on selection fields
  - e.g. interconnect, architecture, and file system
  - Reach out if you noticed something missing!
- Key sections like storage schema given higher visibility
- Working to address raised issues:
  - Complex validation of submission form
- Required input from community:
  - How to handle edits on previous submissions?
    - For entries before the new submission system
    - For current submissions from a given institution

Benchmark Phases and Extended Access Patterns



Random Read Phase - Motivation

- Want to measure fundamental property of the underlying storage
- Random IO pattern common for AI/ML training workloads
  - Random data subsampling is fundamental to how training is done
- Also seen in various HPC workloads
  - Sparse or transverse grid/matrix access
  - Adaptive Mesh Refinement
  - Genomic analysis
  - Financial modelling
- Prior survey results show support for adding a random IO phase
- Want to add new phase without invalidating existing scores

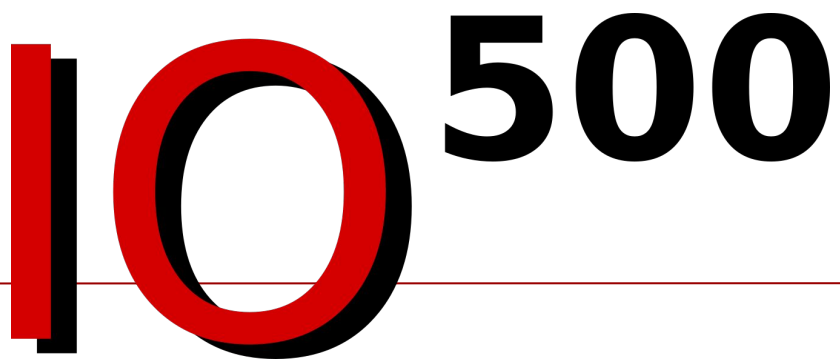
`ior-rand-read` Phase - Implementation

- New random 4KB read phase to be added for next release
  - Reuse existing `ior-easy-write` files for input to avoid writing new files
  - Total data size is the largest available from previous phases to avoid caching
  - No data verification needed, was done during `ior-easy-read` already
- Run at end of other phases to avoid conflicting with other phases/scores
  - Hard stonewall at 300 s (no wearout) to limit increase in runtime
- Existing score is kept, add new score with `ior-rand-read` phase
  - Open question whether to include in MD score with IOPS or BW score as GB/s?
  - Is `ior-rand-read` an IOP or a bandwidth?
- Next steps are to include `ior-rand-read` into benchmark runs/score
  - ISC25 - phase is run by default (unless disabled), but result is not part of official score
  - SC25/ISC26? - new ranked list using new score, when there are enough results
    - Propose when 6+ of Top-10 list entries have new score trigger move to new ranking

Other Potential Access Patterns

- Should a `ior-random-`**`write`** phase also be added?
  - Relatively few HPC workloads have purely random writes
- Want `find-hard` to be “harder” than just “`find` in `mdtest-hard/` dir”
  - Existing `find` score is totally unbalancing the other results
  - Output `find` filename(s) into a file in the storage system for review?
  - Extra attributes, something other than filename (string) comparison?
  - Geometric mean of `find-hard` and `find-easy` to replace existing `find`?
- Should a directory-level `mdtest-rename` phase be added?
  - Rename `mdtest-easy` files and/or directories in a cycle?
  - Check if a hierarchical namespace, or flat strings with ‘`/`’ in them?
- Expect runtime would increase by about 20 min if other phases added

Voice of the Community & Open Discussion



Open Floor

- Downloading and comparing submissions
- Collecting storage system metadata
- Submission form is still hard
- How to make ‘Find’ really hard