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

CONNECTING THE DOTS

ISC High Performance

IO500

BoF Agenda

1. **Welcome** – Jay Lofstead
2. **Award Presentations** – Jay Lofstead
3. **New IO500 List Analysis** – George Markomanolis
4. **Community Talk**
    - Explainable IO: Understanding Why rather than just What.
        - Sarah Neuwirth
5. **Updates**
    - **Website** - Jean Luca Bez
    - **New `ior-rnd4k-easy-read` Phase** – Julian Kunkel
6. **Community Discussion** – Andreas Dilger

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

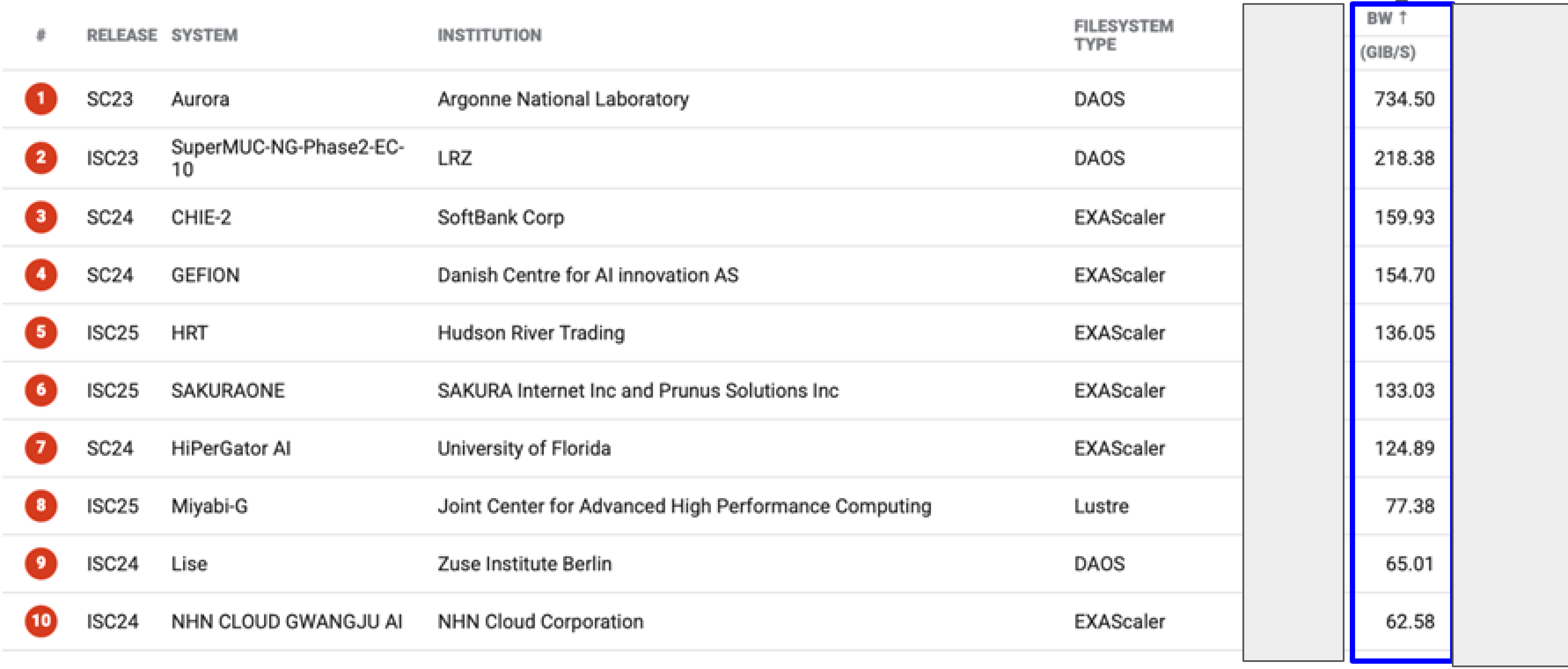
Award Ceremony

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

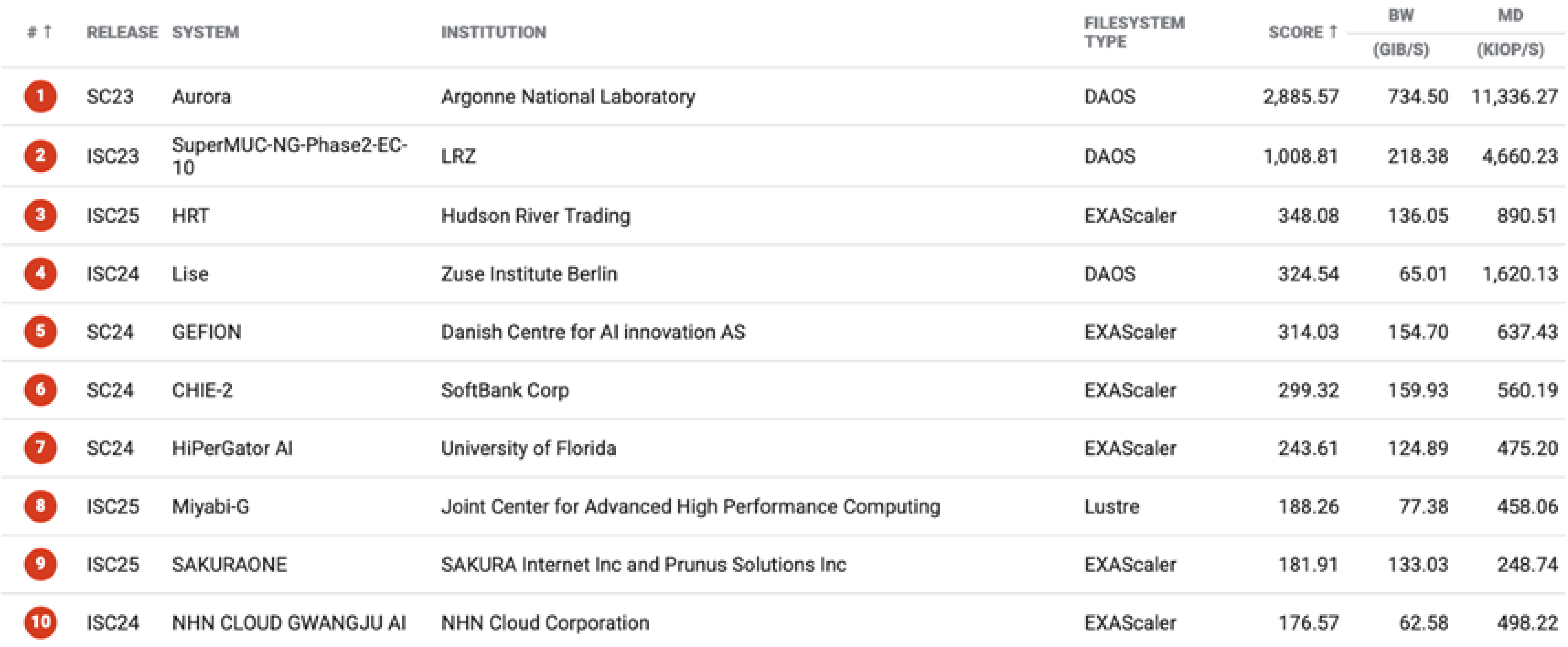
10 Client Node Production - Bandwidth Winner

Sort by BW



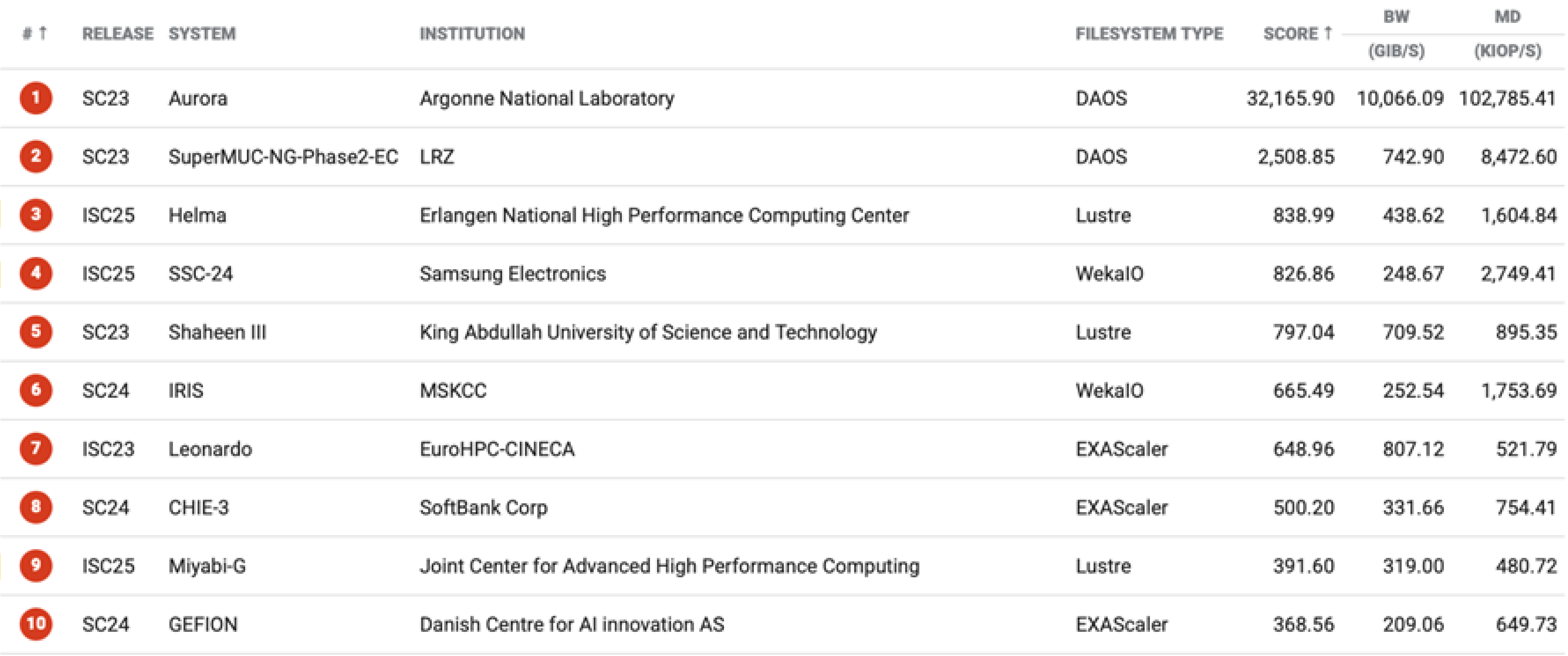
	#	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	BW ↑ (GIB/S)
	1	SC23	Aurora	Argonne National Laboratory	DAOS	734.50
	2	ISC23	SuperMUC-NG-Phase2-EC-10	LRZ	DAOS	218.38
	3	SC24	CHIE-2	SoftBank Corp	EXAScaler	159.93
	4	SC24	GEFION	Danish Centre for AI innovation AS	EXAScaler	154.70
New	5	ISC25	HRT	Hudson River Trading	EXAScaler	136.05
New	6	ISC25	SAKURAONE	SAKURA Internet Inc and Prunus Solutions Inc	EXAScaler	133.03
	7	SC24	HiPerGator AI	University of Florida	EXAScaler	124.89
New	8	ISC25	Miyabi-G	Joint Center for Advanced High Performance Computing	Lustre	77.38
	9	ISC24	Lise	Zuse Institute Berlin	DAOS	65.01
	10	ISC24	NHN CLOUD GWANGJU AI	NHN Cloud Corporation	EXAScaler	62.58

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
New	3	ISC25	HRT	Hudson River Trading	EXAScaler	348.08	136.05	890.51
	4	ISC24	Lise	Zuse Institute Berlin	DAOS	324.54	65.01	1,620.13
	5	SC24	GEFION	Danish Centre for AI innovation AS	EXAScaler	314.03	154.70	637.43
	6	SC24	CHIE-2	SoftBank Corp	EXAScaler	299.32	159.93	560.19
	7	SC24	HiPerGator AI	University of Florida	EXAScaler	243.61	124.89	475.20
New	8	ISC25	Miyabi-G	Joint Center for Advanced High Performance Computing	Lustre	188.26	77.38	458.06
New	9	ISC25	SAKURAONE	SAKURA Internet Inc and Prunus Solutions Inc	EXAScaler	181.91	133.03	248.74
	10	ISC24	NHN CLOUD GWANGJU AI	NHN Cloud Corporation	EXAScaler	176.57	62.58	498.22

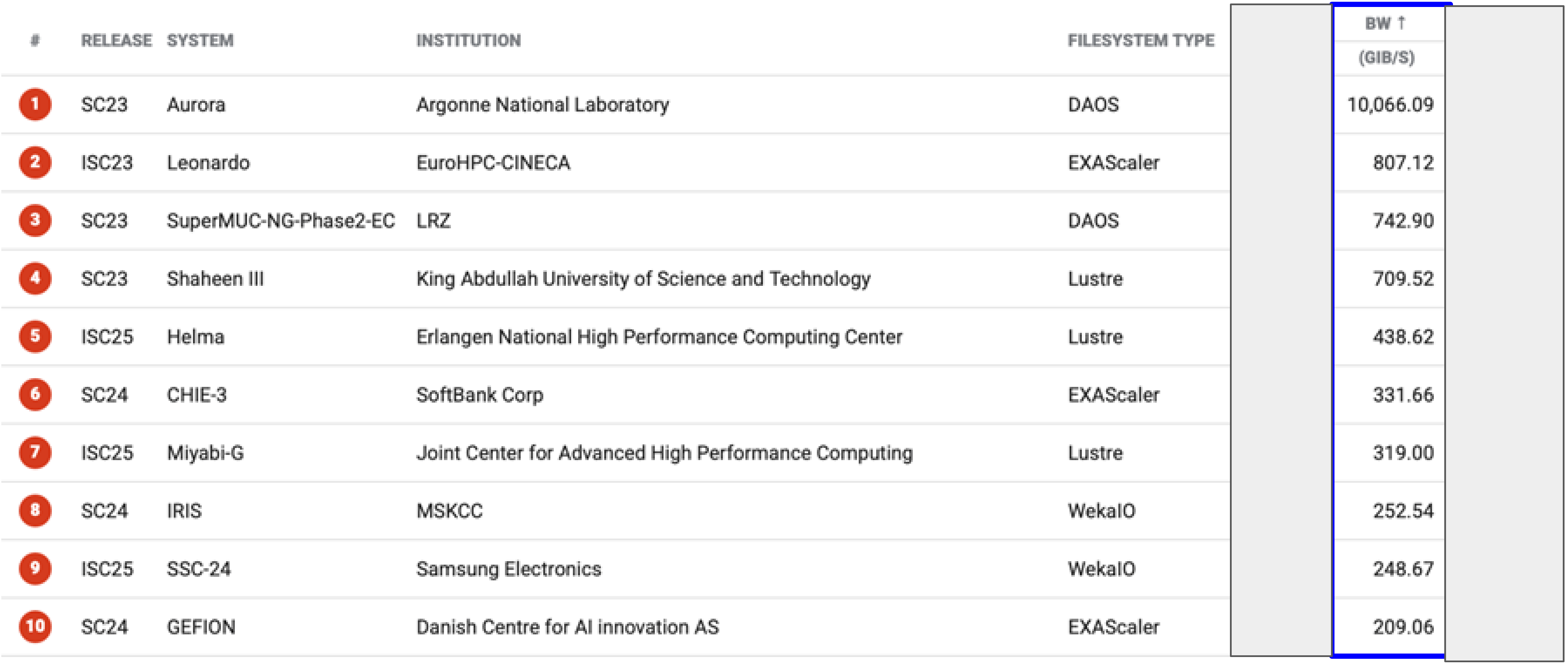
IO500 Production List - 3 New Entries



	# ↑	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE ↑	BW	MD
							(GIB/S)	(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
New	3	ISC25	Helma	Erlangen National High Performance Computing Center	Lustre	838.99	438.62	1,604.84
New	4	ISC25	SSC-24	Samsung Electronics	WekaIO	826.86	248.67	2,749.41
	5	SC23	Shaheen III	King Abdullah University of Science and Technology	Lustre	797.04	709.52	895.35
	6	SC24	IRIS	MSKCC	WekaIO	665.49	252.54	1,753.69
	7	ISC23	Leonardo	EuroHPC-CINECA	EXAScaler	648.96	807.12	521.79
	8	SC24	CHIE-3	SoftBank Corp	EXAScaler	500.20	331.66	754.41
New	9	ISC25	Miyabi-G	Joint Center for Advanced High Performance Computing	Lustre	391.60	319.00	480.72
	10	SC24	GEFION	Danish Centre for AI innovation AS	EXAScaler	368.56	209.06	649.73

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	Lustre	709.52
New	5	ISC25	Helma	Erlangen National High Performance Computing Center	Lustre	438.62
	6	SC24	CHIE-3	SoftBank Corp	EXAScaler	331.66
New	7	ISC25	Miyabi-G	Joint Center for Advanced High Performance Computing	Lustre	319.00
	8	SC24	IRIS	MSKCC	WekaIO	252.54
New	9	ISC25	SSC-24	Samsung Electronics	WekaIO	248.67
	10	SC24	GEFION	Danish Centre for AI innovation AS	EXAScaler	209.06

IO500 List Analysis

IO500 List - Growth in Entries and Institutions

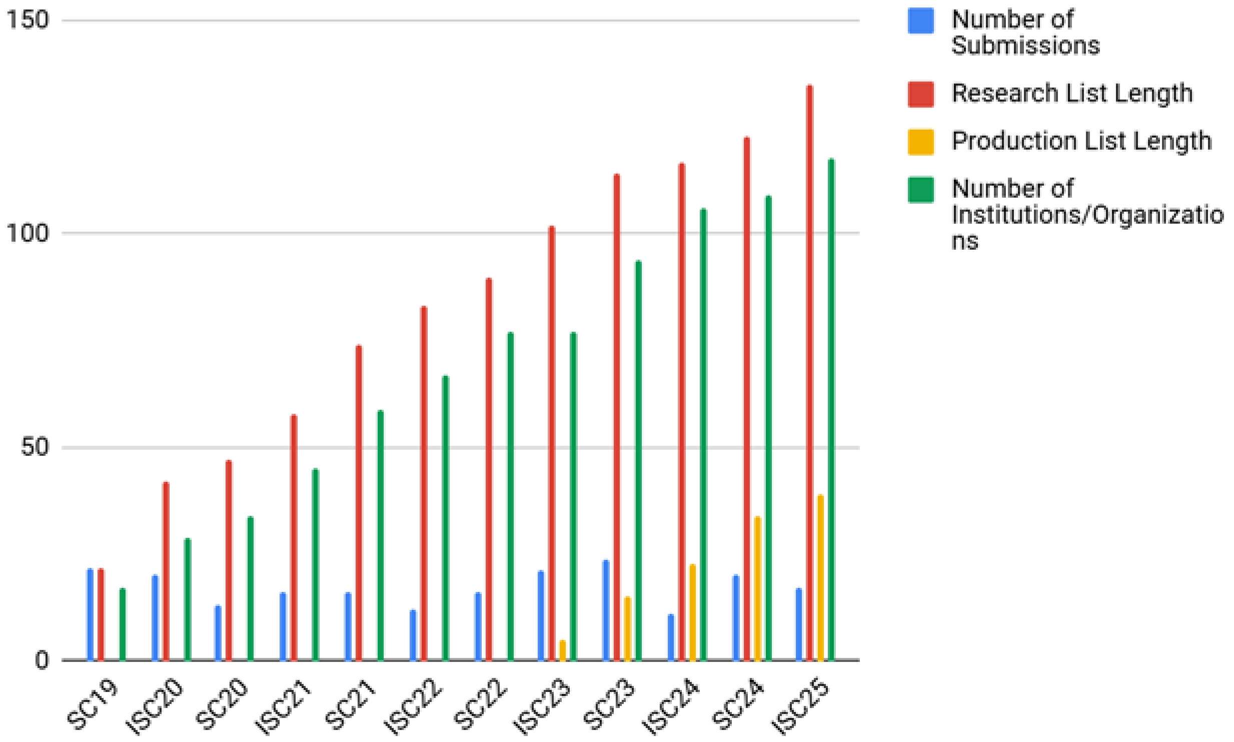
ISC25

19 submissions (2 rejected)

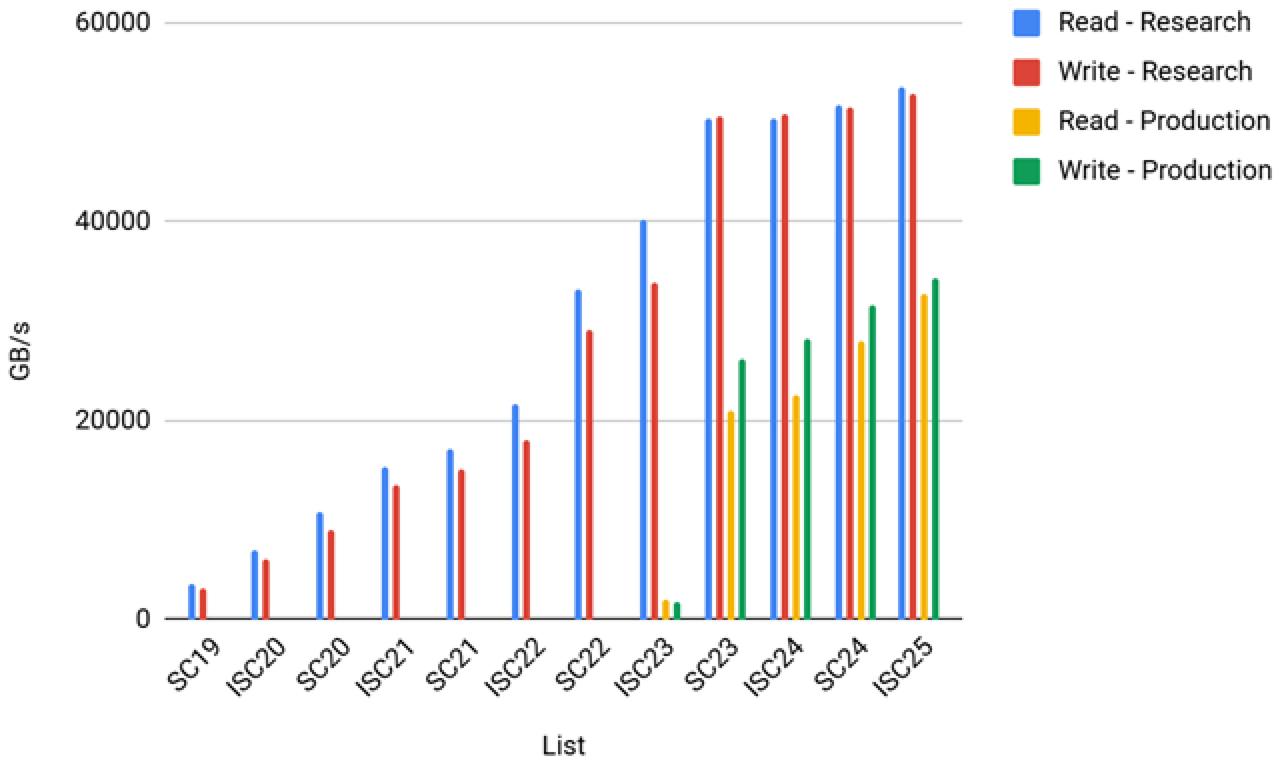
- 10 for 10-Client Research
- 3 for 10-Client Production
- 12 for IO500 Research
- 5 for IO500 Production
- 1 for Full (< 10 client nodes)

Around 285 list entries

More than 100 institutions

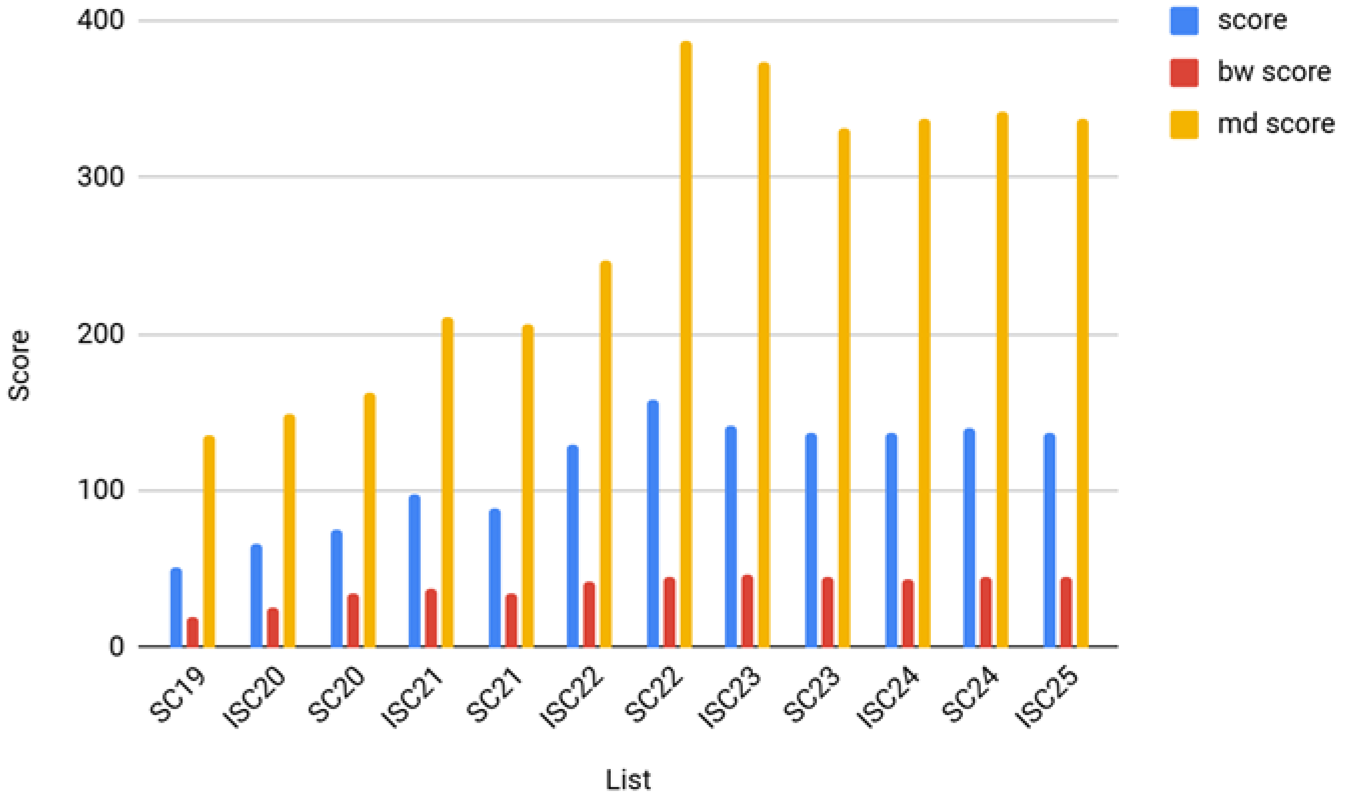


IO500 List - Aggregate List Bandwidth

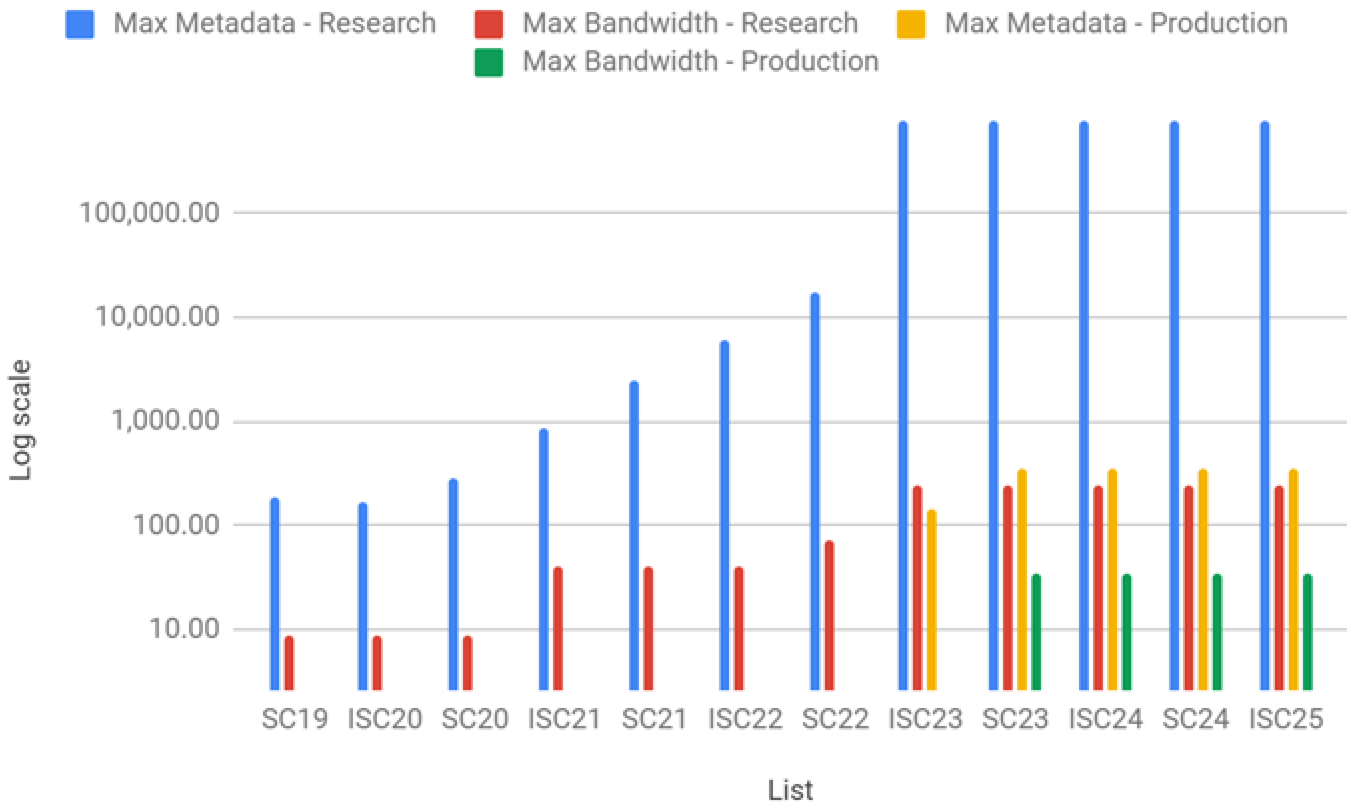


IO500 List - Median Scores

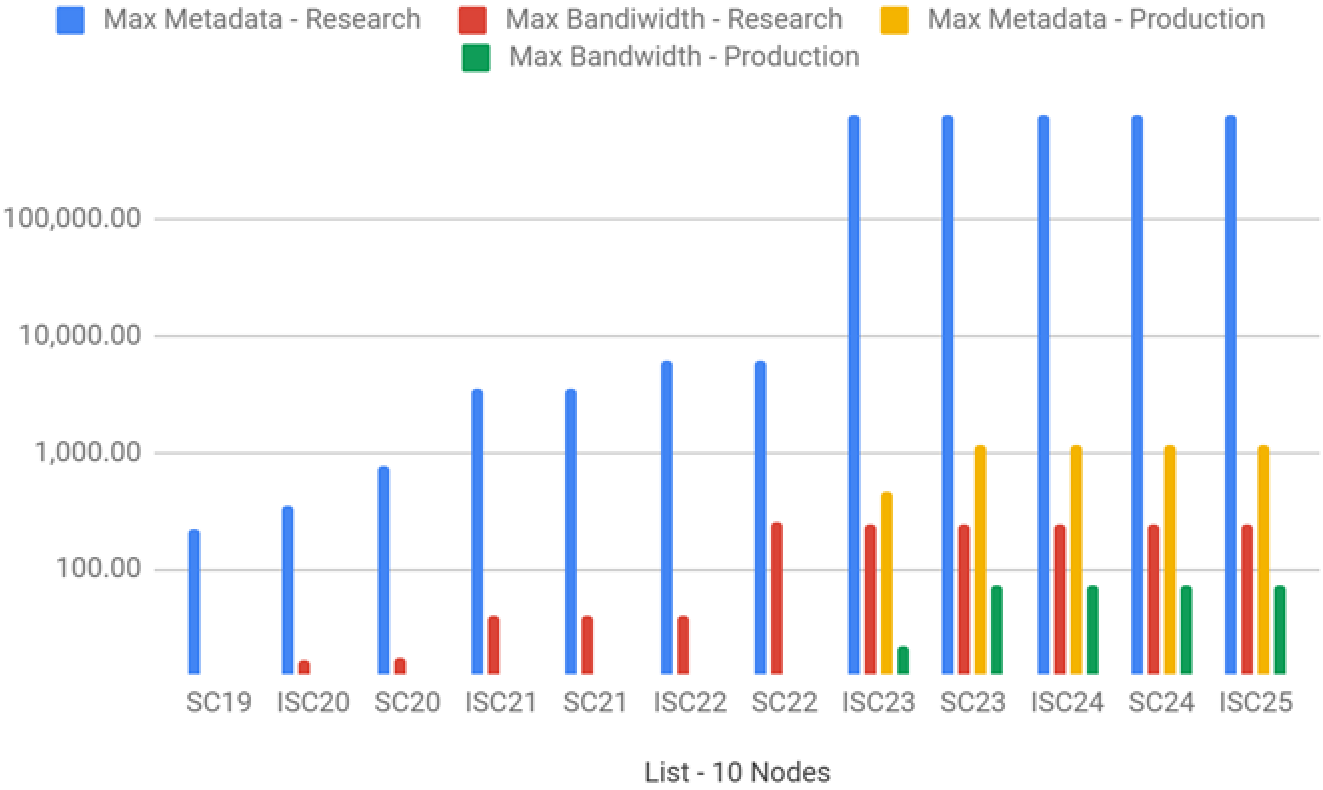
Median scores are mixed compared to SC24



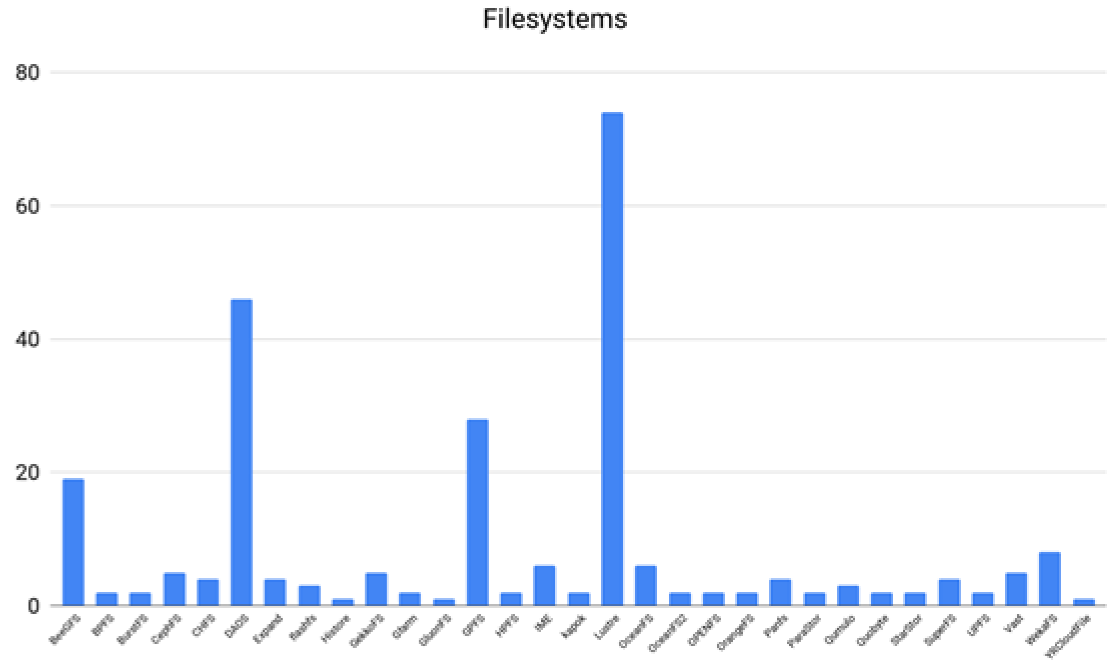
IO500 List - Growth in Max Score per Client



10-Client List - Growth in Max Scores per Client



IO500 List - Number of File System Entries



Lustre and DAOS have the most submissions, followed by GPFS and BeeGFS

Community Talk

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 4K Random Read Phase

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 showed support for adding a random IO phase
- Want to add new phase without invalidating existing scores

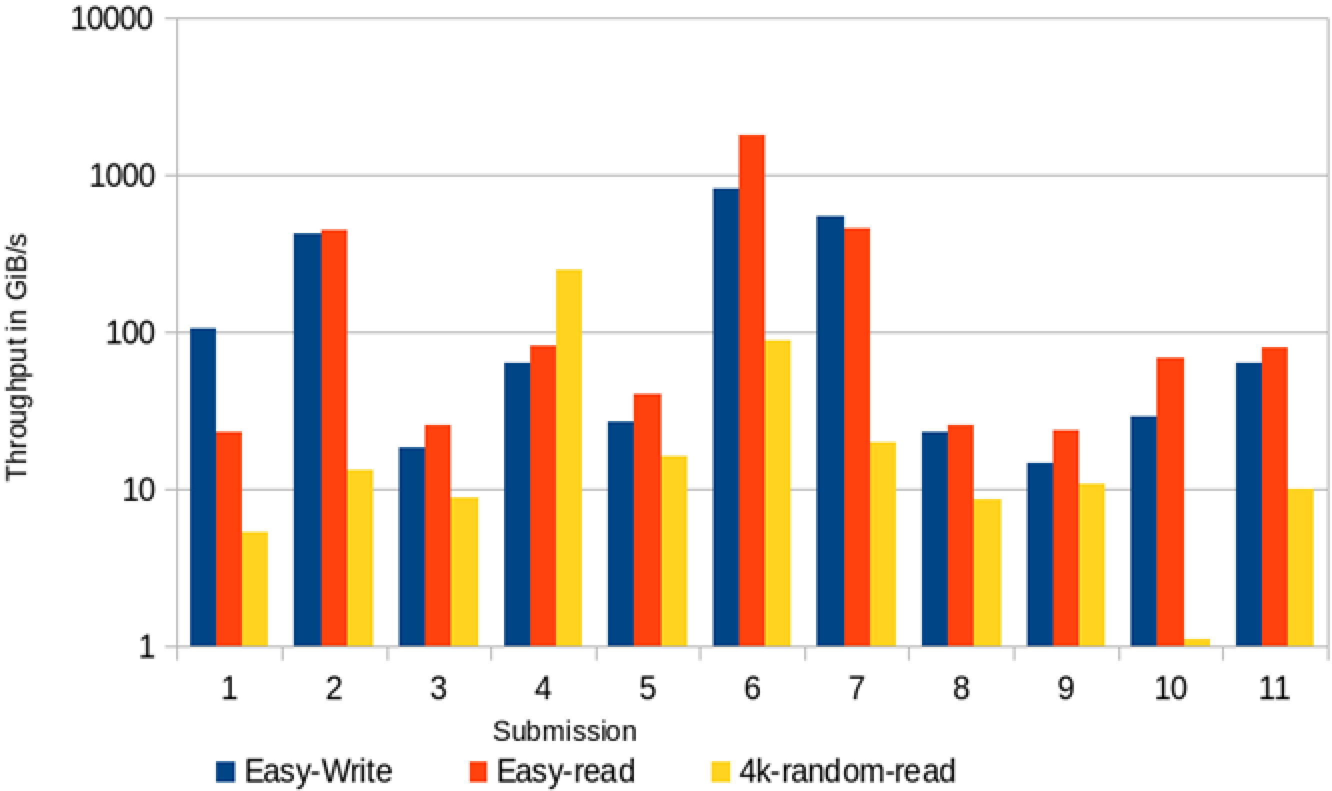
Latest IO500 Updates

- 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 “`find-easy`”
    - 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

`ior-rnd4k-easy-read` Phase - Implementation

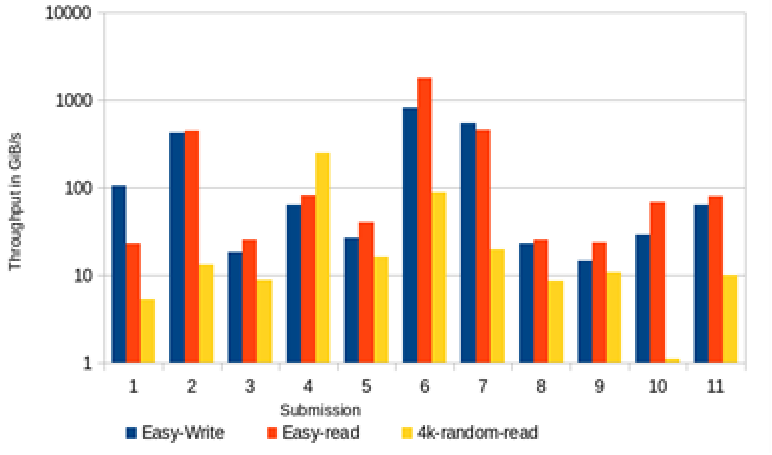
- New random 4KB read phase was added (unranked)
  - Reuse existing `ior-easy-write` files for input to avoid writing new files
  - Total data size is the largest available from previous phases to avoid caching
  - No data verification needed, was done during `ior-easy-read` already
- Run at end of other phases to avoid conflicting with other phases/scores
  - Hard stonewall at 300 s (with wearout) to limit increase in runtime
- Existing score is kept, add new score with `ior-rnd4K-easy-read` phase
  - Reported as bandwidth to allow comparison to other ior phaes
- Next steps to include `it` into benchmark runs/score
  - ISC25 - phase was run by default (unless disabled), but result is not part of official score
  - 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
    - How to rank previous submissions without invalidating all entries?

Results from submissions



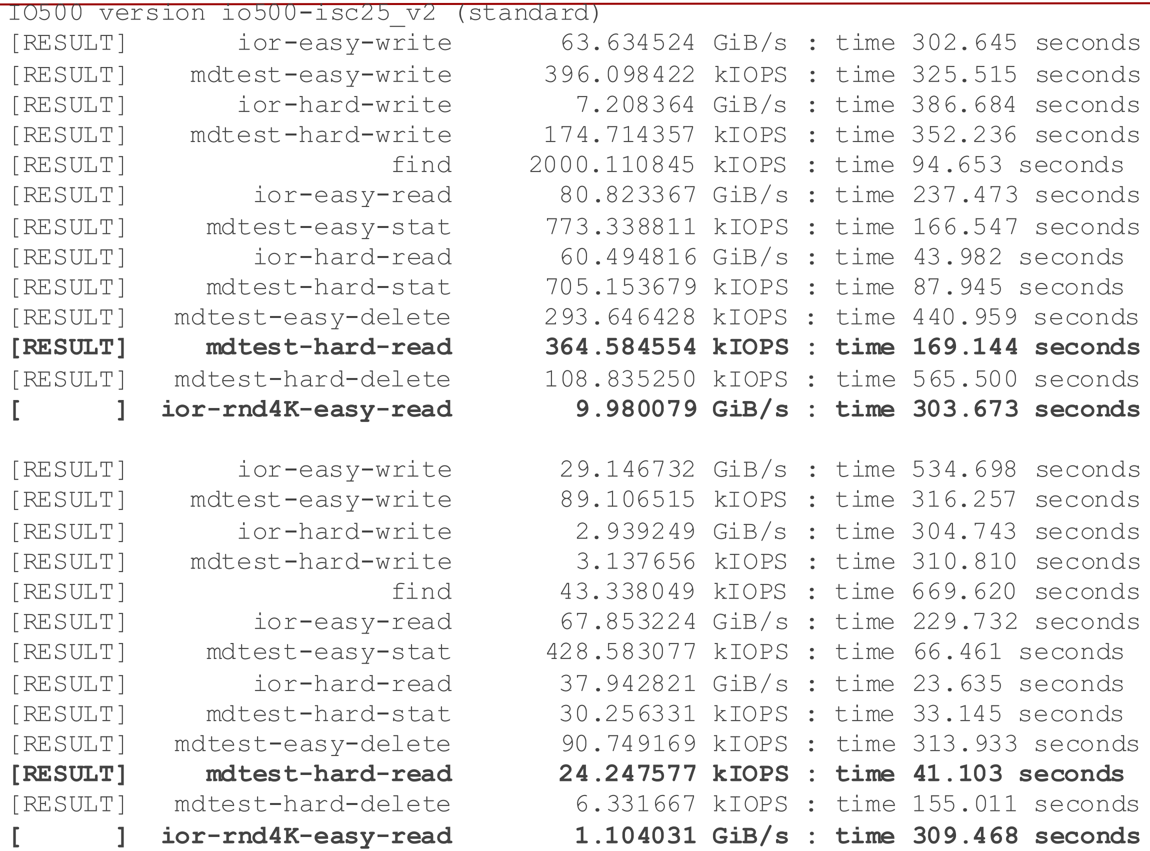
Results from submissions

- Some numbers exceeded expectations
  - Caching took place
  - Discussion + code reviews
- Bug in v1
  - file size was wrongly determined
  - random with 31 bits of offsets
- Both remedied in v2
  - 64 bit random with 128 bit entropy
- Submission 10 yields low hard scores
  - write - 2.9 GiB/s - read - 38.0 GiB/s
- Submission 4 - has high metadata rates
  - 337 kOPS/s MD hard read
  - would be about 1.3 GiB/s throughput



- Code base
  - v1: 1-9
  - v2: 10-11

Detailed Submissions 10-11 - kOPS vs. BW



```
IO500 version io500-isc25_v2 (standard)
[RESULT]       ior-easy-write        63.634524 GiB/s : time 302.645 seconds
[RESULT]    mdtest-easy-write       396.098422 kIOPS : time 325.515 seconds
[RESULT]       ior-hard-write         7.208364 GiB/s : time 386.684 seconds
[RESULT]    mdtest-hard-write       174.714357 kIOPS : time 352.236 seconds
[RESULT]                 find      2000.110845 kIOPS : time 94.653 seconds
[RESULT]        ior-easy-read        80.823367 GiB/s : time 237.473 seconds
[RESULT]     mdtest-easy-stat       773.338811 kIOPS : time 166.547 seconds
[RESULT]        ior-hard-read        60.494816 GiB/s : time 43.982 seconds
[RESULT]     mdtest-hard-stat       705.153679 kIOPS : time 87.945 seconds
[RESULT]   mdtest-easy-delete       293.646428 kIOPS : time 440.959 seconds
[RESULT]     mdtest-hard-read       364.584554 kIOPS : time 169.144 seconds
[RESULT]   mdtest-hard-delete       108.835250 kIOPS : time 565.500 seconds
[      ]  ior-rnd4K-easy-read         9.980079 GiB/s : time 303.673 seconds

[RESULT]       ior-easy-write        29.146732 GiB/s : time 534.698 seconds
[RESULT]    mdtest-easy-write        89.106515 kIOPS : time 316.257 seconds
[RESULT]       ior-hard-write         2.939249 GiB/s : time 304.743 seconds
[RESULT]    mdtest-hard-write         3.137656 kIOPS : time 310.810 seconds
[RESULT]                 find        43.338049 kIOPS : time 669.620 seconds
[RESULT]        ior-easy-read        67.853224 GiB/s : time 229.732 seconds
[RESULT]     mdtest-easy-stat       428.583077 kIOPS : time 66.461 seconds
[RESULT]        ior-hard-read        37.942821 GiB/s : time 23.635 seconds
[RESULT]     mdtest-hard-stat        30.256331 kIOPS : time 33.145 seconds
[RESULT]   mdtest-easy-delete        90.749169 kIOPS : time 313.933 seconds
[RESULT]     mdtest-hard-read        24.247577 kIOPS : time 41.103 seconds
[RESULT]   mdtest-hard-delete         6.331667 kIOPS : time 155.011 seconds
[      ]  ior-rnd4K-easy-read         1.104031 GiB/s : time 309.468 seconds
```

From analysis,
ior-rand-read looks like
it matches expectations
=> new read pattern
=> as MD value not useful

would be 2616.2 kOPS

would be 288.3 kOPS

Selection as BW is meaningful

Voice of the Community & Open Discussion

Other Potential Access Patterns

- Should a `ior-rnd4k-`**`write`** phase also be added?
  - Relatively few HPC workloads have purely random writes
- Should we add the 1MiB random read/write?
  - From the extended mode
- 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`?
- Expect runtime would increase by about 20 min if other phases added

SC 25 (Nov 16-21, 2025)

- Call for submission: Oct 1st
- Submission deadline: Nov 10th
- List release: BoF date TBD

SC25

St. Louis, MO | hpc ignites.

Open Floor

- Downloading and comparing submissions
- Collecting storage system metadata automatically
- Is the submission form getting better?
- How to make ‘`find-hard`’ really hard