BoF Agenda

1. **Welcome** – Jay Lofstead
2. **The New IO500 List Analysis** – Andreas Dilger
3. **Award Presentations** – Jay Lofstead
4. **Roadmap**
   - **Website Update** - Andreas Dilger
   - **Benchmark Phases and Extended Access Patterns -** Julian Kunkel
   - **List Split and Reproducibility -** George Markomanolis
5. **Community Discussion**

IO500: The High-Performance Storage Community

Committee

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

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

IO500 List Analysis

IO500 List - Growth in Entries and Institutions

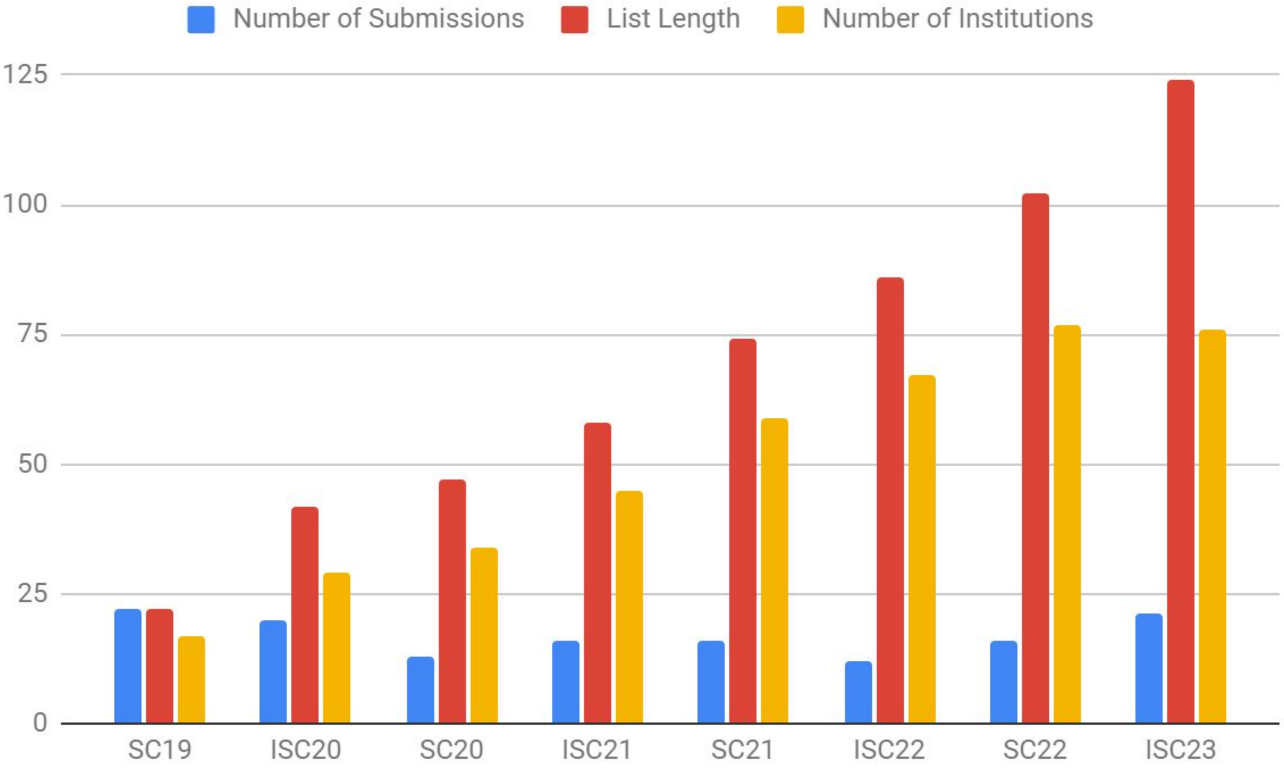
ISC23

21 submissions

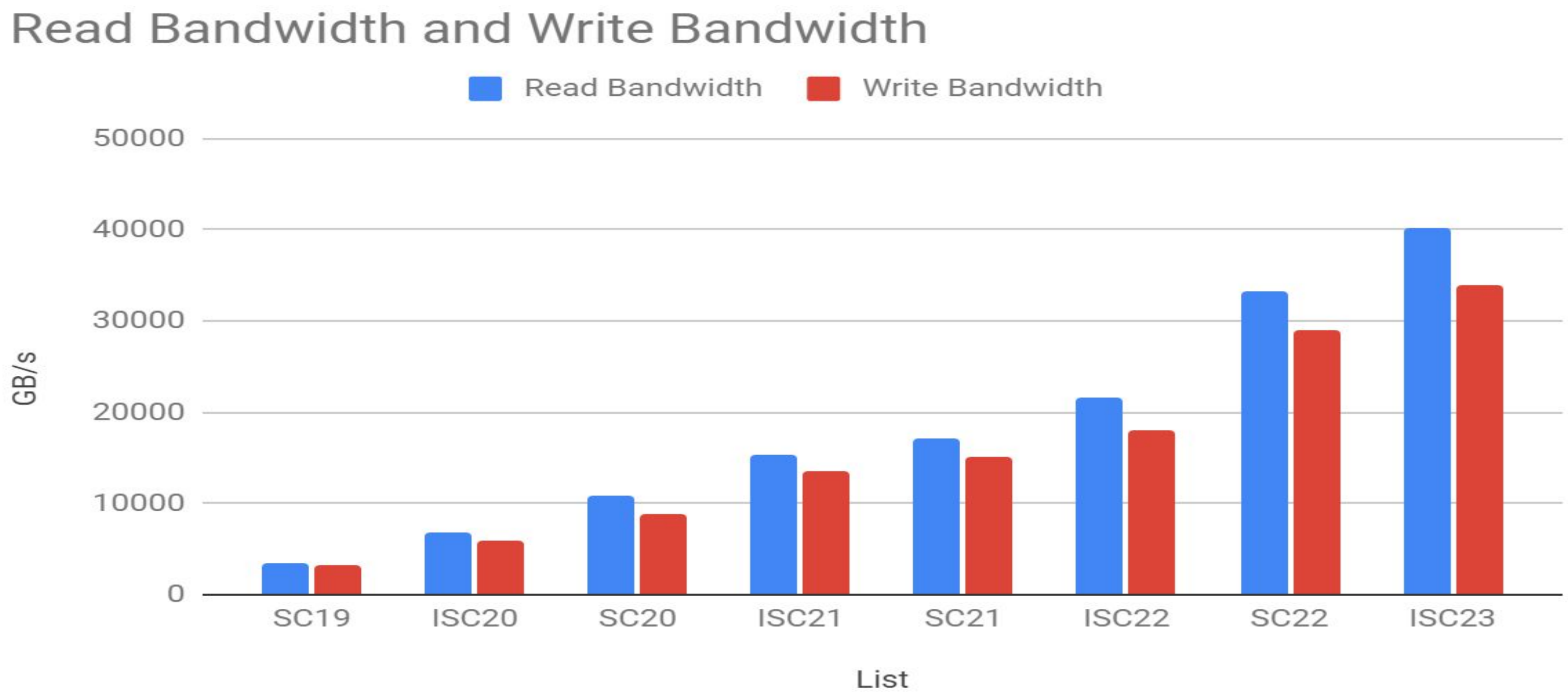
- 14 for 10-Client Research
- 2 for 10-Client Production
- 16 for IO500 Research
- 4 for IO500 Production

124 list entries

76 institutions

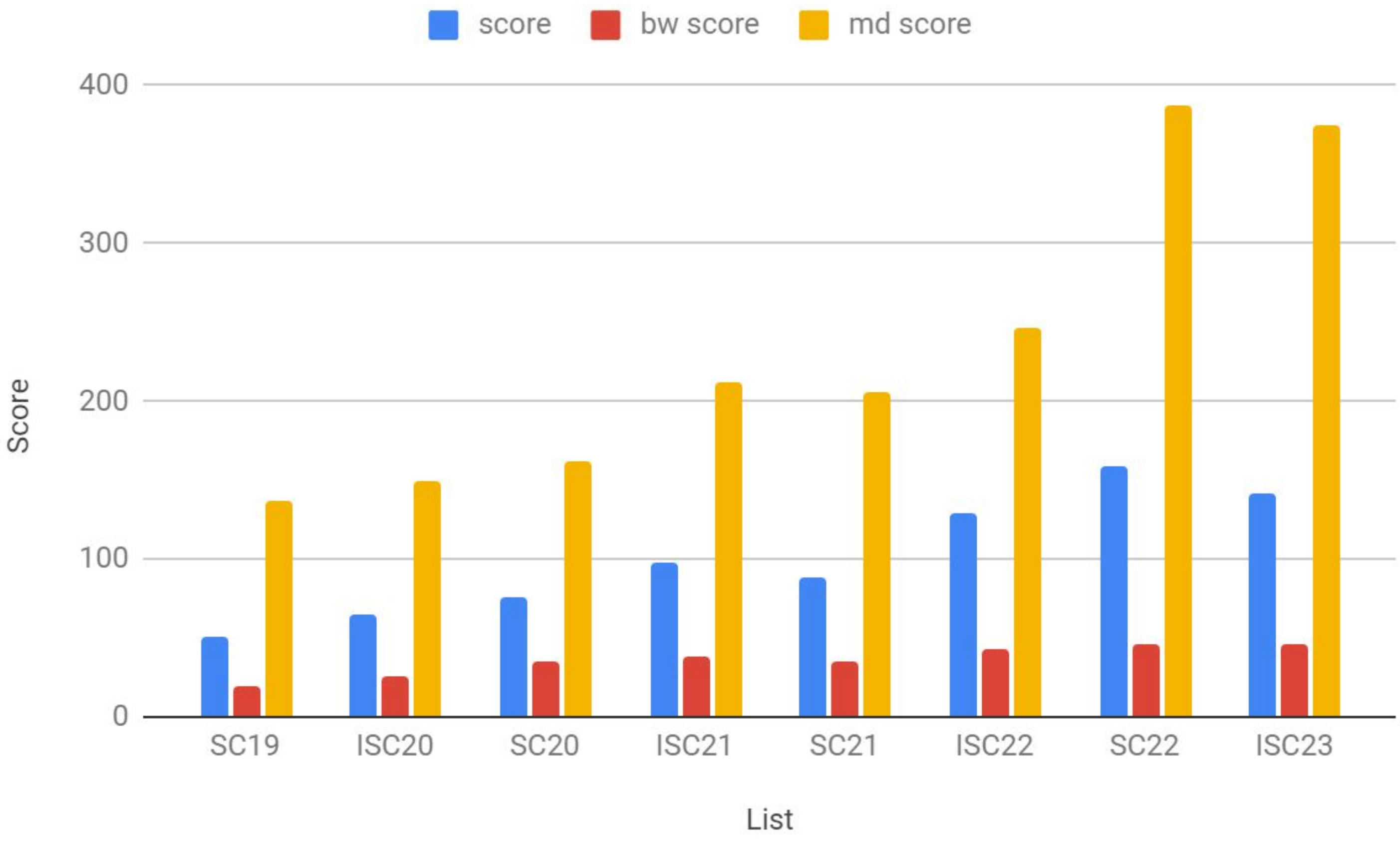


IO500 List - Aggregate List Bandwidth

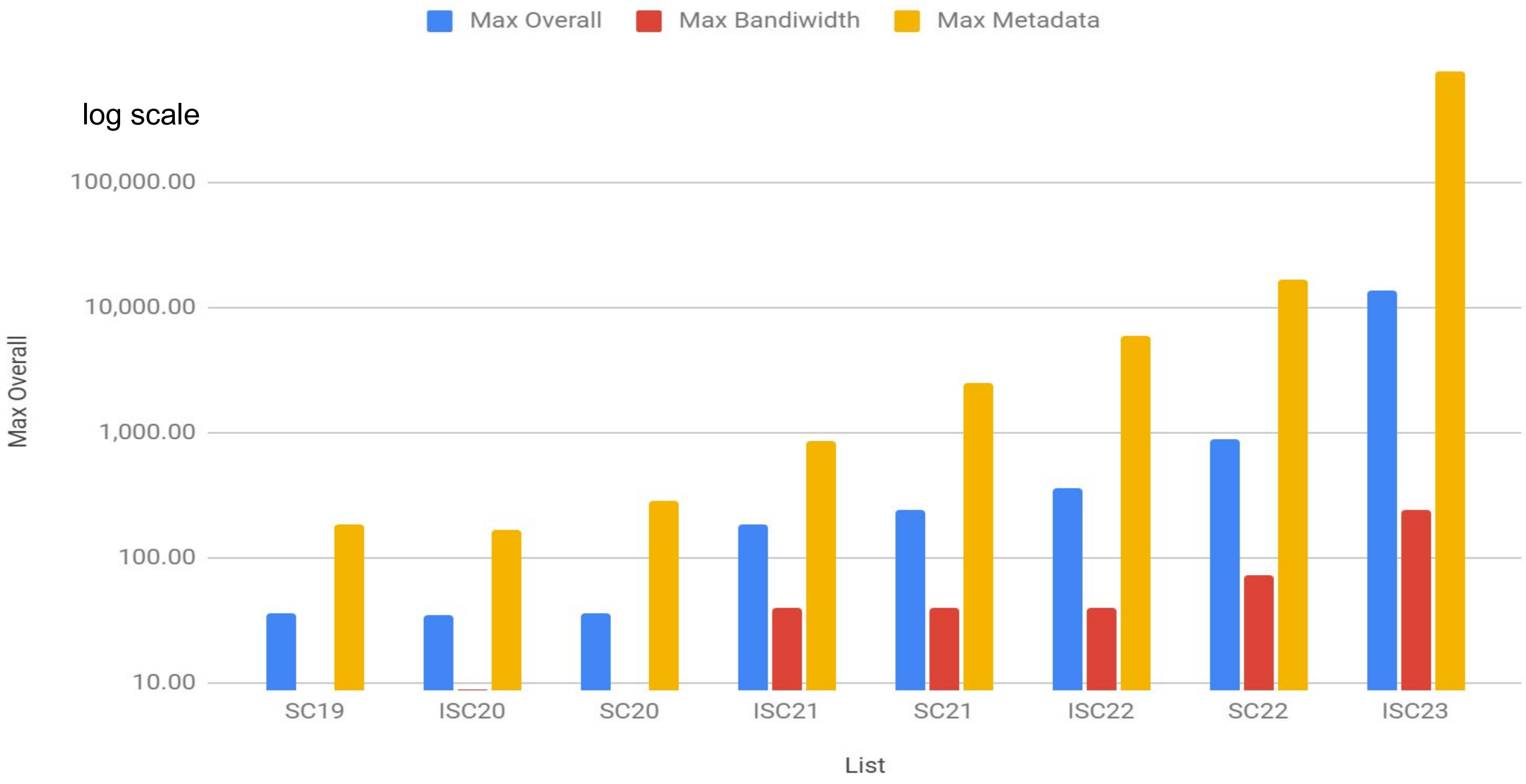


IO500 List - Median Scores

Median scores are mixed compared to SC22

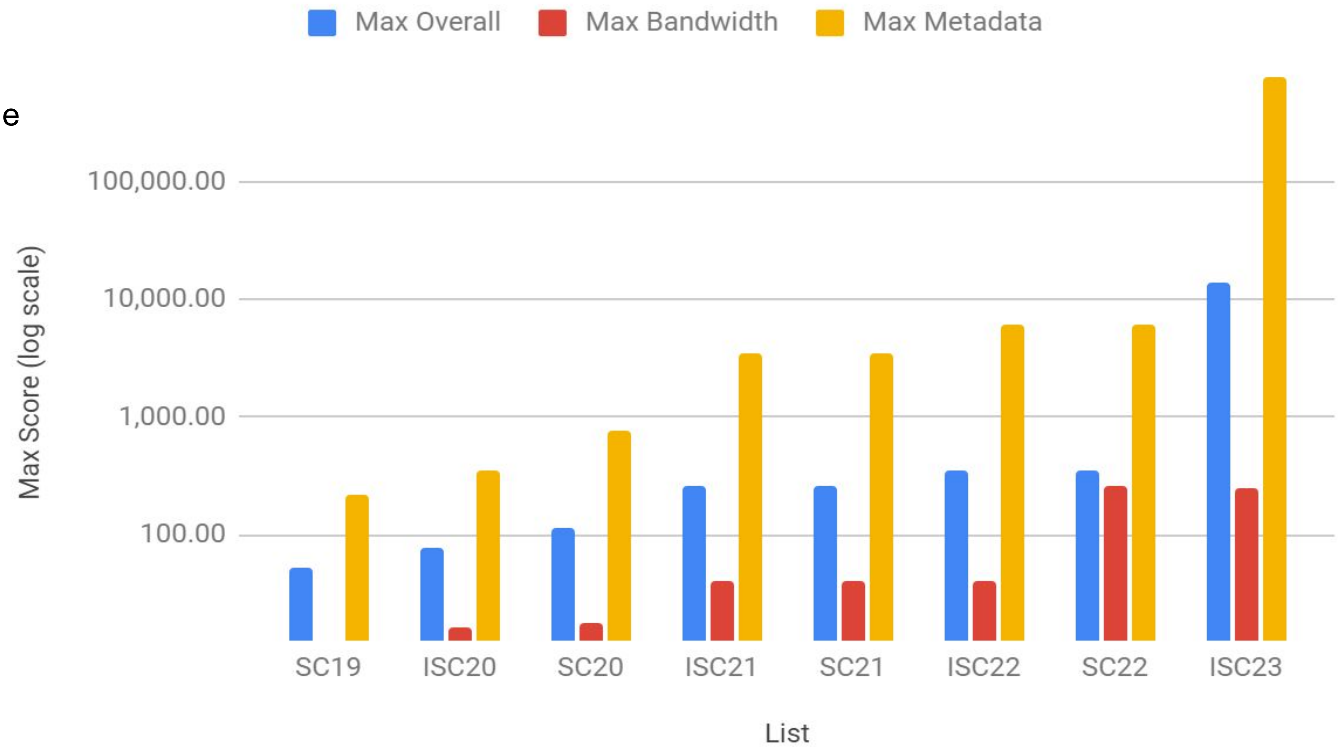


IO500 List - Growth in Max Score per Client



10-Client List - Growth in Max Scores per Client

log scale



No growth in bandwidth

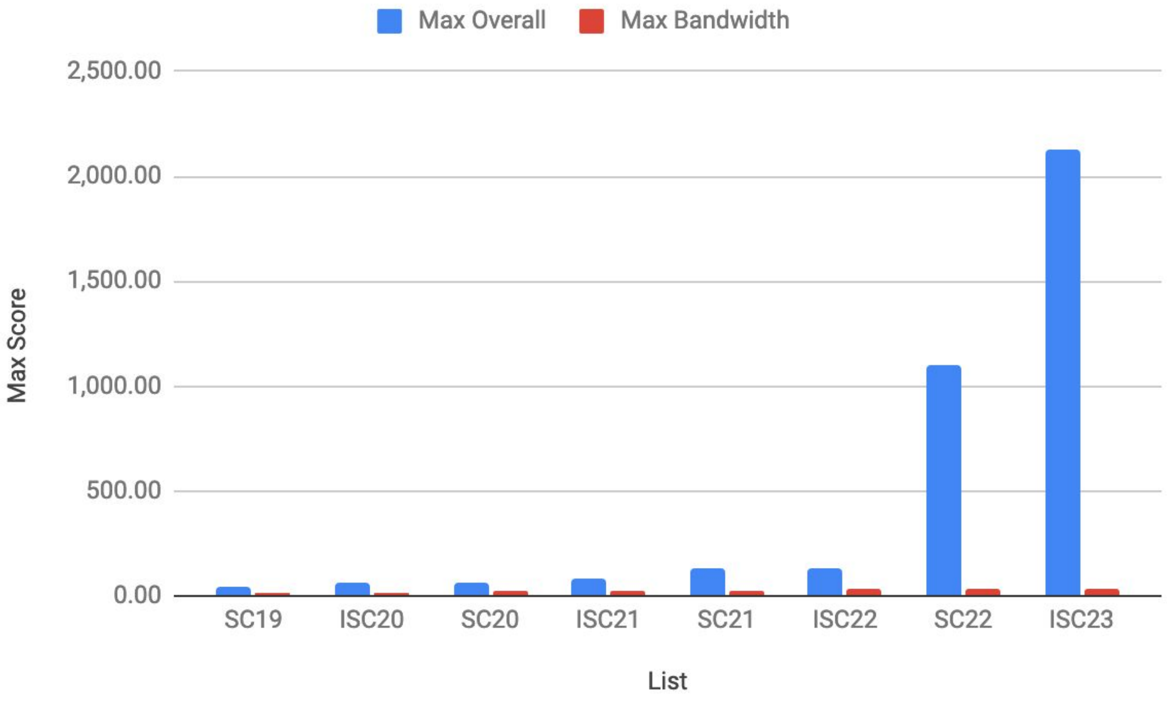
Good growth in metadata this year

IO500 List - Growth in Max Score per Storage Server

Per-client scores are flat

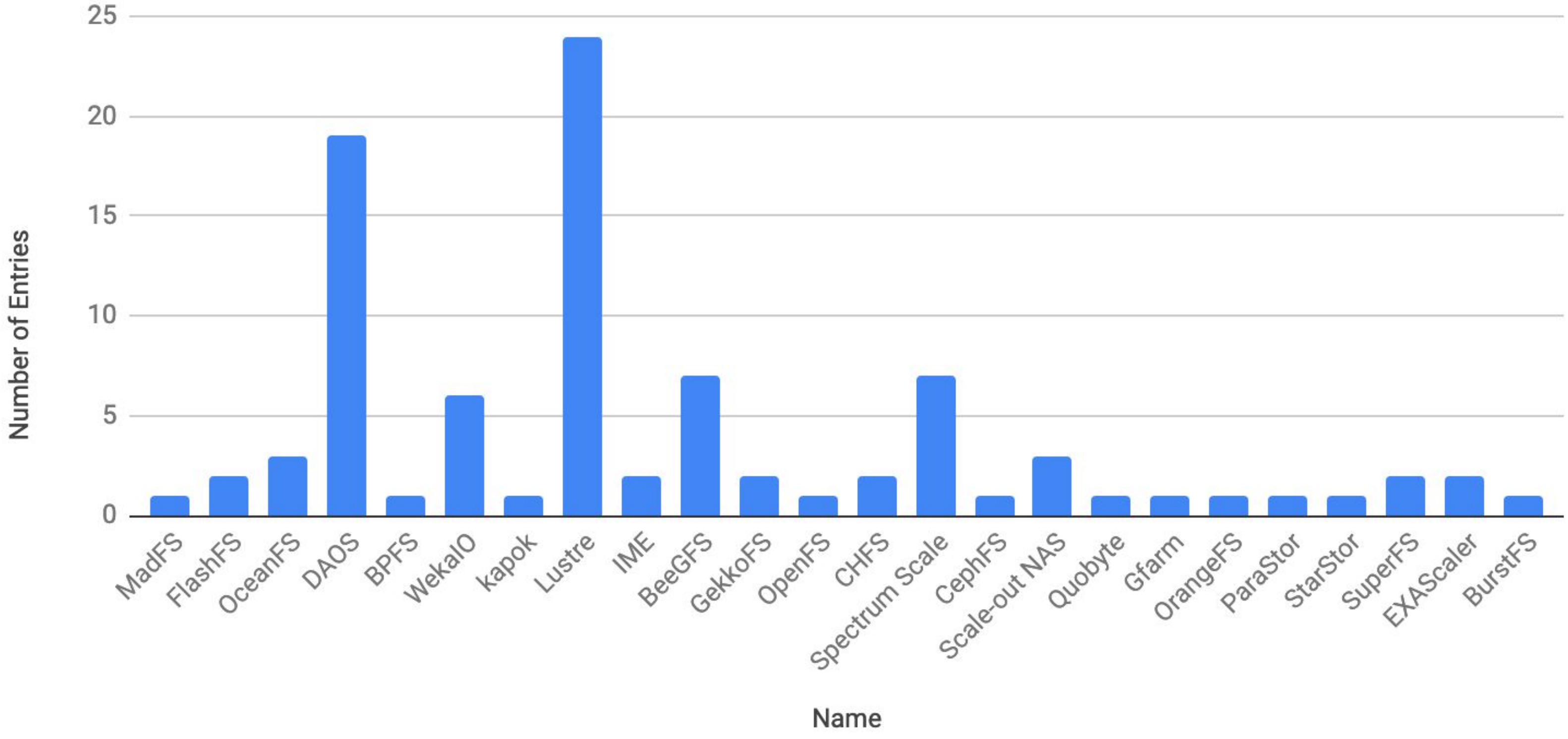
Per-storage server scores are growing much slower

- bandwidth flat for 5 lists
- metadata flat for 3 lists



Note: metadata score per server growth reflected in overall score

IO500 List - Number of File System Entries



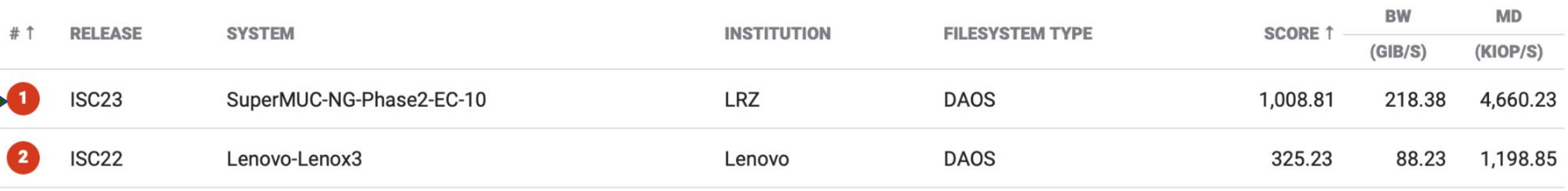
Award Ceremony

IO500

Seven Awards

- 10 Client Production List
- 10 Client Research List
    - Bandwidth
    - Overall
- IO500 Production List
    - Bandwidth
    - Overall
- IO500 Research List
    - Bandwidth
    - Overall

10 Client Node Production - Overall Winner



# ↑	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE ↑	BW	MD
						(GIB/S)	(KIOP/S)
1	ISC23	SuperMUC-NG-Phase2-EC-10	LRZ	DAOS	1,008.81	218.38	4,660.23
2	ISC22	Lenovo-Lenox3	Lenovo	DAOS	325.23	88.23	1,198.85

Certificate

IO500 Performance Certification

This Certificate is awarded to:

Leibniz-Rechenzentrum (SuperMUC Phase2)

#1 in the 10 Client Node Production Overall Score

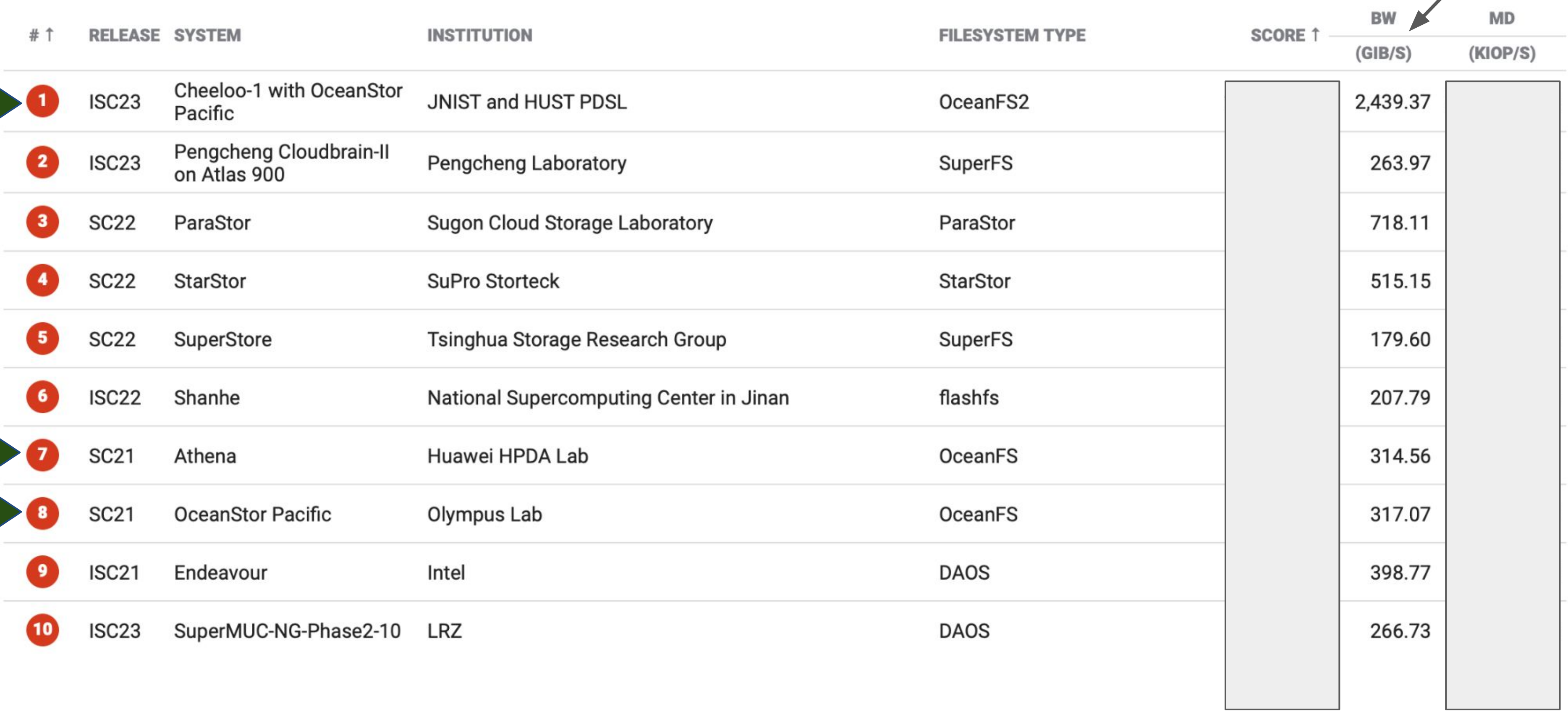
May 2023

IO500 Steering Board

https://io500.org/list/ISC23/ten-production

10 Client Node Research - Bandwidth Winner

Sort by BW



	# ↑	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE ↑	BW (GIB/S)	MD (KIOP/S)
◆	1	ISC23	Cheeloo-1 with OceanStor Pacific	JNIST and HUST PDSL	OceanFS2		2,439.37	
	2	ISC23	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng Laboratory	SuperFS		263.97	
	3	SC22	ParaStor	Sugon Cloud Storage Laboratory	ParaStor		718.11	
	4	SC22	StarStor	SuPro Storteck	StarStor		515.15	
	5	SC22	SuperStore	Tsinghua Storage Research Group	SuperFS		179.60	
	6	ISC22	Shanhe	National Supercomputing Center in Jinan	flashfs		207.79	
◆	7	SC21	Athena	Huawei HPDA Lab	OceanFS		314.56	
◆	8	SC21	OceanStor Pacific	Olympus Lab	OceanFS		317.07	
	9	ISC21	Endeavour	Intel	DAOS		398.77	
	10	ISC23	SuperMUC-NG-Phase2-10	LRZ	DAOS		266.73	

Certificate

IO500 Performance Certification

This Certificate is awarded to:

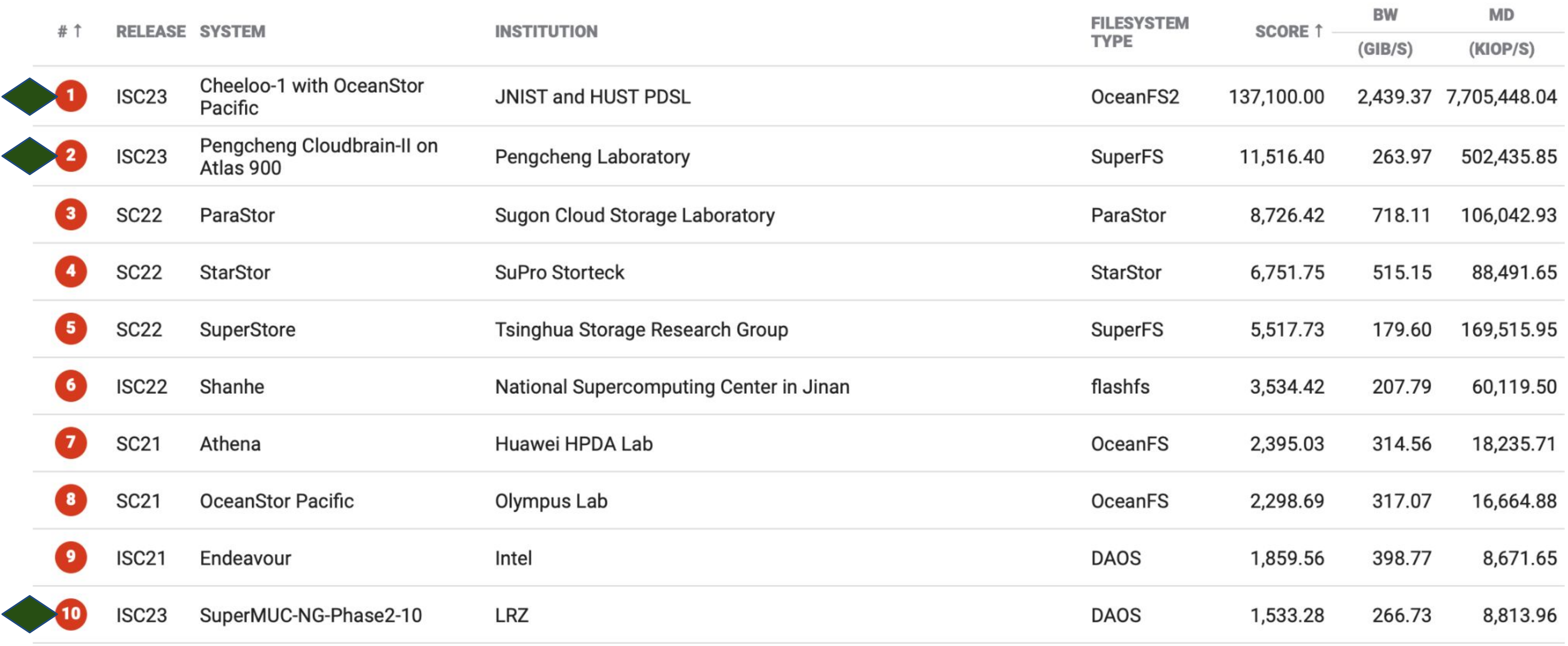
JNIST and HUST PDSL (Cheeloo-1)
with OceanStor Pacific from Huawei
#1 in the 10 Client Node Research Bandwidth Score

May 2023

IO500 Steering Board

https://io500.org/list/ISC23/ten

10 Client Node Research - Overall Winner



# ↑	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE ↑	BW	MD
						(GIB/S)	(KIOP/S)
◆ 1	ISC23	Cheeloo-1 with OceanStor Pacific	JNIST and HUST PDSL	OceanFS2	137,100.00	2,439.37	7,705,448.04
◆ 2	ISC23	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng Laboratory	SuperFS	11,516.40	263.97	502,435.85
3	SC22	ParaStor	Sugon Cloud Storage Laboratory	ParaStor	8,726.42	718.11	106,042.93
4	SC22	StarStor	SuPro Storteck	StarStor	6,751.75	515.15	88,491.65
5	SC22	SuperStore	Tsinghua Storage Research Group	SuperFS	5,517.73	179.60	169,515.95
6	ISC22	Shanhe	National Supercomputing Center in Jinan	flashfs	3,534.42	207.79	60,119.50
7	SC21	Athena	Huawei HPDA Lab	OceanFS	2,395.03	314.56	18,235.71
8	SC21	OceanStor Pacific	Olympus Lab	OceanFS	2,298.69	317.07	16,664.88
9	ISC21	Endeavour	Intel	DAOS	1,859.56	398.77	8,671.65
◆ 10	ISC23	SuperMUC-NG-Phase2-10	LRZ	DAOS	1,533.28	266.73	8,813.96

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

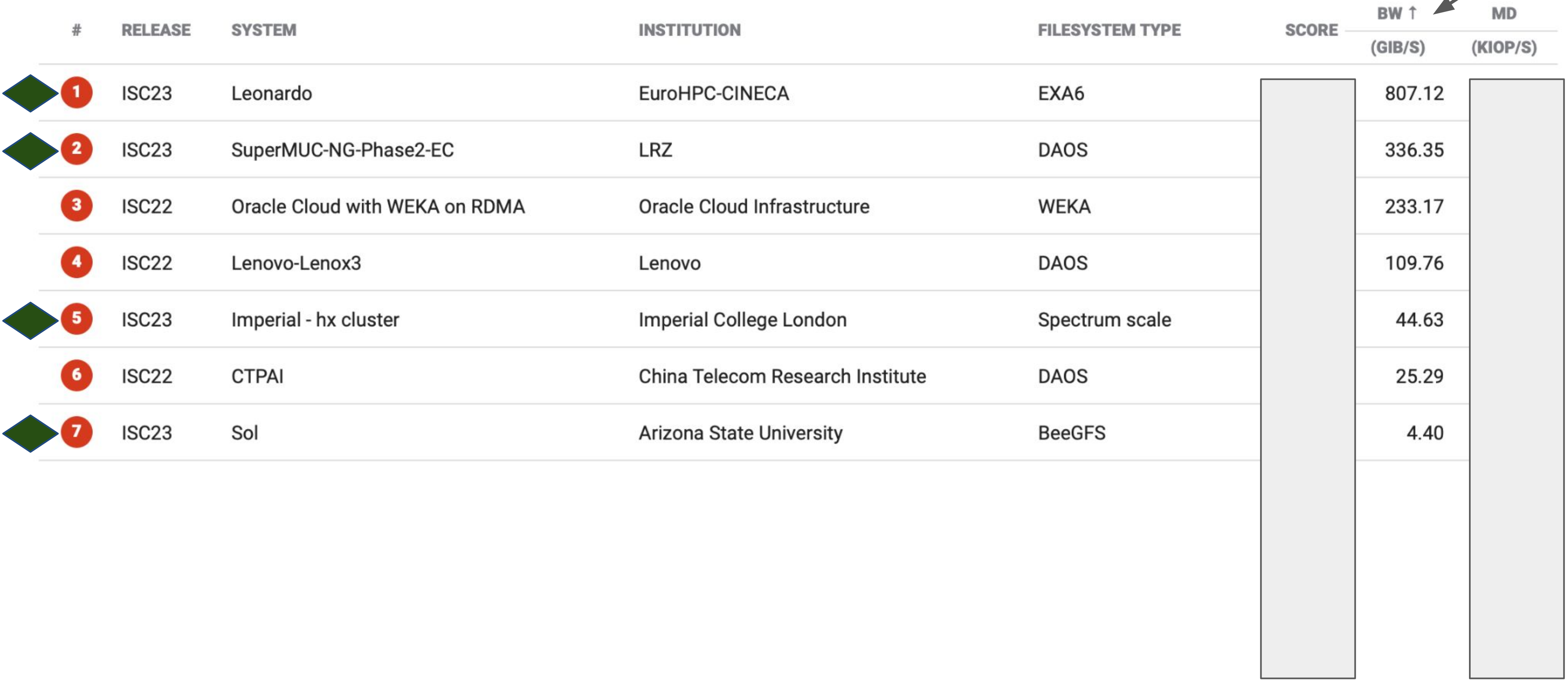
May 2023

IO500 Steering Board

https://io500.org/list/ISC23/ten

IO500 Production List - Bandwidth Winner

Sorted by BW



#	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE	BW ↑ (GIB/S)	MD (KIOP/S)
◆ 1	ISC23	Leonardo	EuroHPC-CINECA	EXA6		807.12	
◆ 2	ISC23	SuperMUC-NG-Phase2-EC	LRZ	DAOS		336.35	
3	ISC22	Oracle Cloud with WEKA on RDMA	Oracle Cloud Infrastructure	WEKA		233.17	
4	ISC22	Lenovo-Lenox3	Lenovo	DAOS		109.76	
◆ 5	ISC23	Imperial - hx cluster	Imperial College London	Spectrum scale		44.63	
6	ISC22	CTPAI	China Telecom Research Institute	DAOS		25.29	
◆ 7	ISC23	Sol	Arizona State University	BeeGFS		4.40	

Certificate

IO500 Performance Certification

This Certificate is awarded to:

EuroHPC-CINECA (Leonardo)

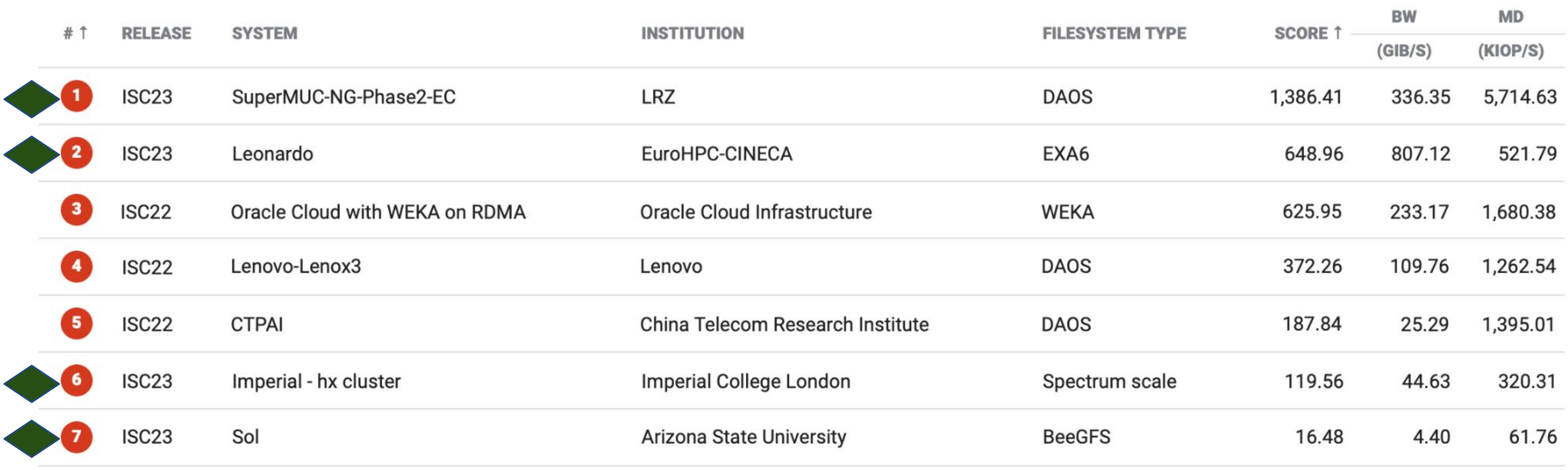
#1 in the IO500 Production Bandwidth Score

May 2023

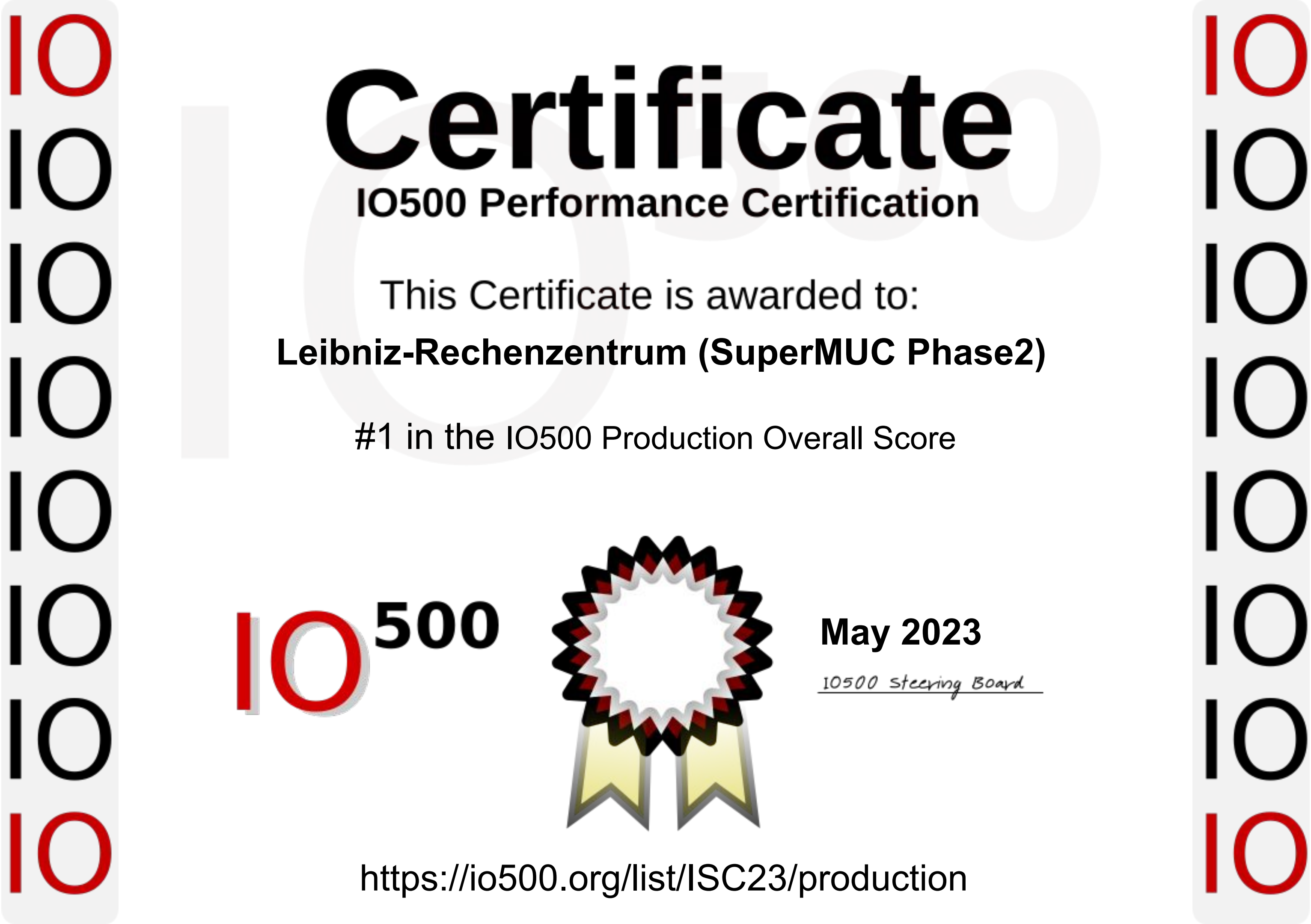
IO500 Steering Board

https://io500.org/list/ISC23/production

IO500 Production List - Overall Winner



# ↑	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE ↑	BW	MD
						(GIB/S)	(KIOP/S)
◆ 1	ISC23	SuperMUC-NG-Phase2-EC	LRZ	DAOS	1,386.41	336.35	5,714.63
◆ 2	ISC23	Leonardo	EuroHPC-CINECA	EXA6	648.96	807.12	521.79
3	ISC22	Oracle Cloud with WEKA on RDMA	Oracle Cloud Infrastructure	WEKA	625.95	233.17	1,680.38
4	ISC22	Lenovo-Lenox3	Lenovo	DAOS	372.26	109.76	1,262.54
5	ISC22	CTPAI	China Telecom Research Institute	DAOS	187.84	25.29	1,395.01
◆ 6	ISC23	Imperial - hx cluster	Imperial College London	Spectrum scale	119.56	44.63	320.31
◆ 7	ISC23	Sol	Arizona State University	BeeGFS	16.48	4.40	61.76



Certificate

IO500 Performance Certification

This Certificate is awarded to:

Leibniz-Rechenzentrum (SuperMUC Phase2)

#1 in the IO500 Production Overall Score

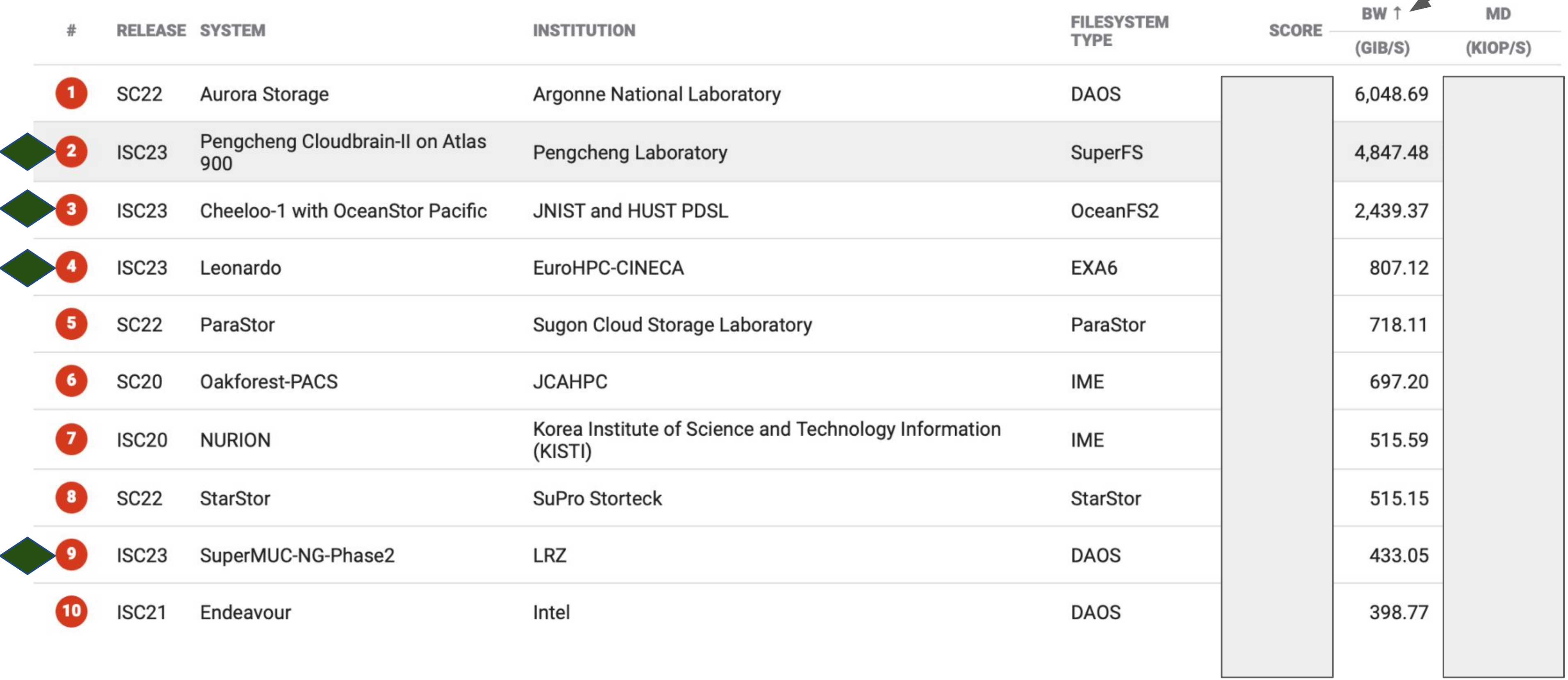
May 2023

IO500 Steering Board

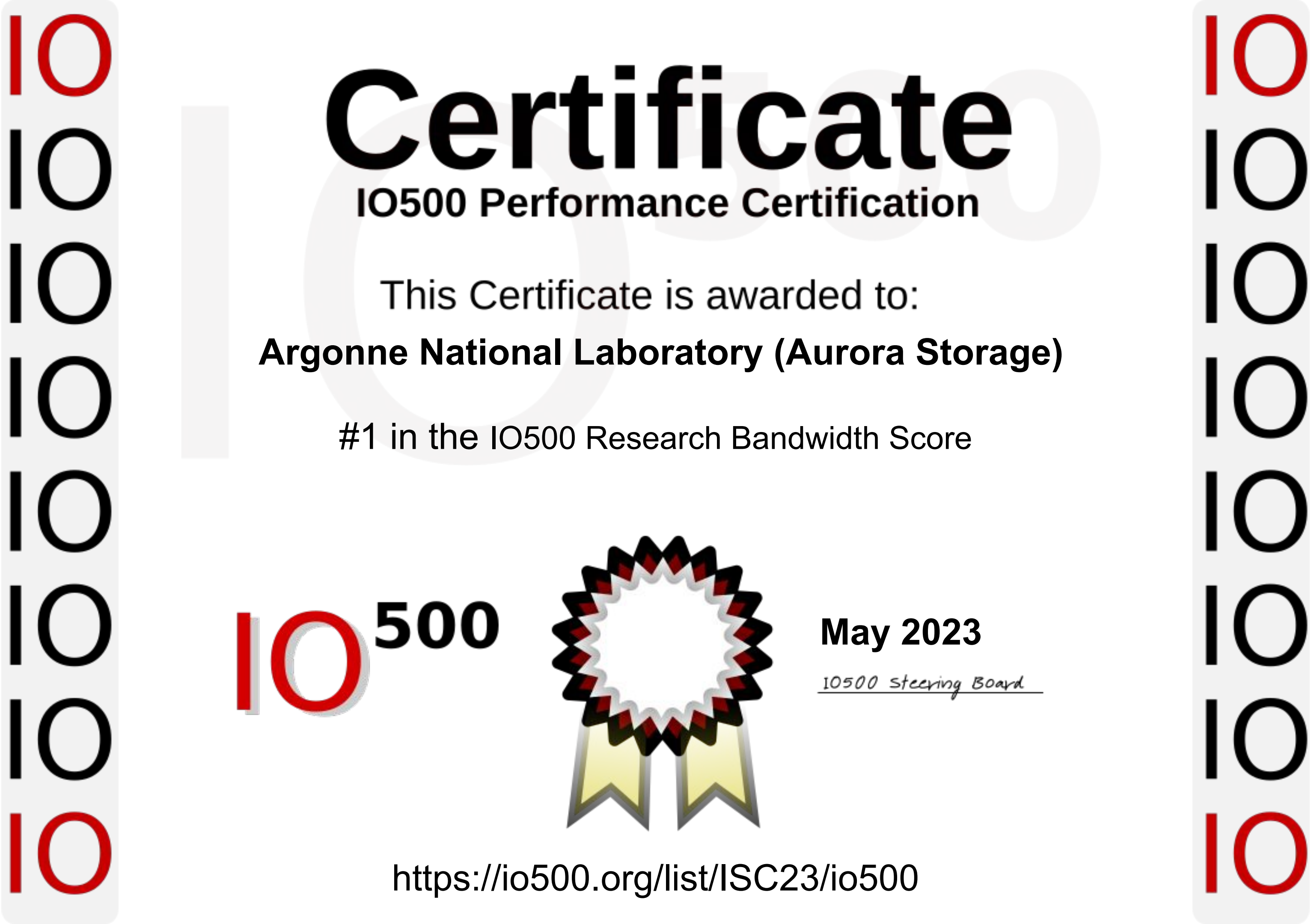
https://io500.org/list/ISC23/production

IO500 Research List - Bandwidth Winner

Sorted by BW



#	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE	BW ↑	MD
						(GIB/S)	(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	



Certificate

IO500 Performance Certification

This Certificate is awarded to:

Argonne National Laboratory (Aurora Storage)

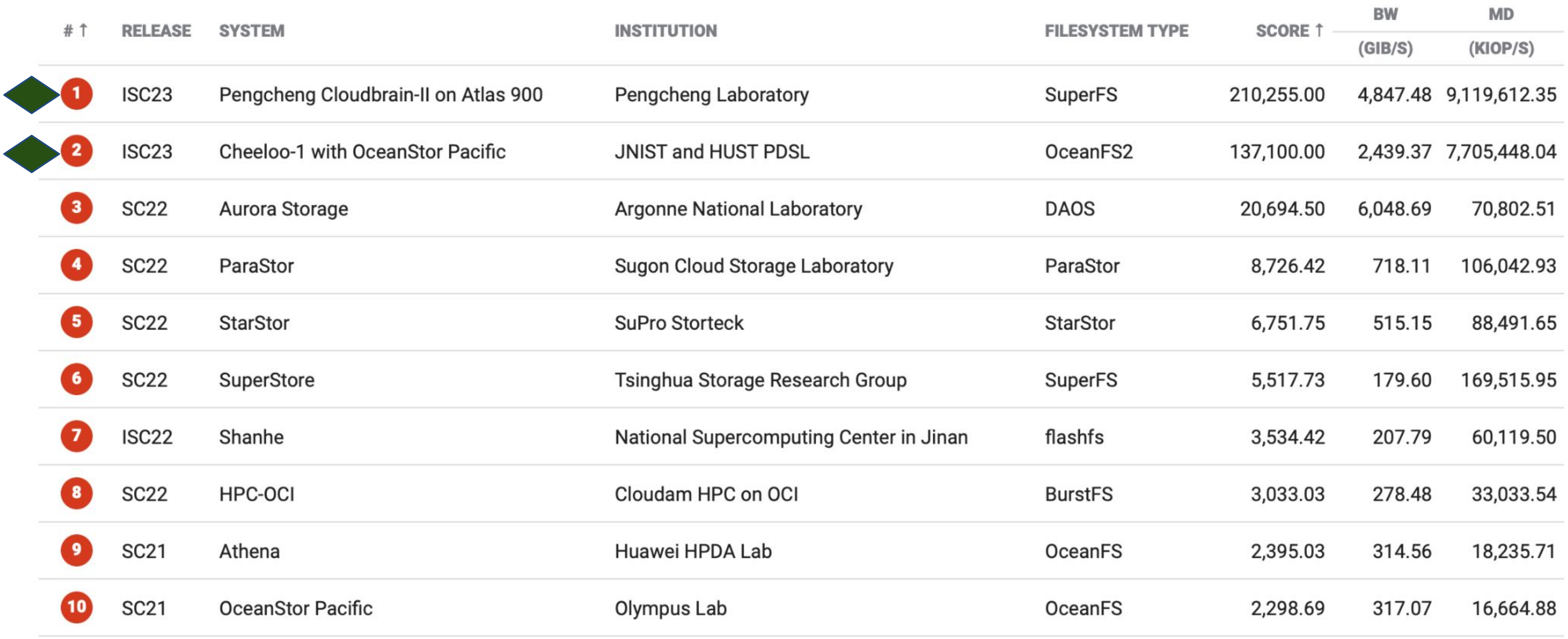
#1 in the IO500 Research Bandwidth Score

May 2023

IO500 Steering Board

https://io500.org/list/ISC23/io500

IO500 Research List - Overall Winner



# ↑	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE ↑	BW	MD
						(GIB/S)	(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

May 2023

IO500 Steering Board

https://io500.org/list/ISC23/io500

List of Awarded Systems in the Ranked Lists

10 Client	**Production**	**Leibniz-Rechenzentrum**	DAOS	**1008.81 score**
10 Client Research	Bandwidth	**JNIST and HUST PDSL**	OceanFS2	2439.37 GiB/s
	Overall	**JNIST and HUST PDSL**	OceanFS2	**137,100.00 score**
IO500 Production	Bandwidth	**EuroHPC-CINECA**	EXA6	807.12 GiB/s
	Overall	**Leibniz-Rechenzentrum**	DAOS	**1386.41 score**
IO500 Research	Bandwidth	**Argonne National Laboratory**	DAOS	6048.69 GiB/s
	Overall	**Pengcheng Laboratory**	SuperFS	**210,255.00 score**

Roadmap

Roadmap for the IO500

- Improve a few usage patterns (random, better find)
- Collect and evaluate results for potential new benchmark phases
  - Not part of benchmark score yet
  - Create proposals to give rationale and details of any potential new phase
    - Proposal must gain community consensus before official inclusion
- Improve `io500.org` submissions page
  - Add more mandatory fields/sections, help text to clarify field usage
  - Please give feedback and be patient in the transition
- Community meeting
  - Skipped a meeting in February due to lack of topics/work on submission system
  - Target August/September 2023 if topics to discuss

SC 23 (Nov 12-17, 2023)

- Call for submission: Sept 22nd
- Testing phase ends: Sept 29th
  - Code freeze, but please test before!
- Submission deadline: Nov 3rd
- List release: BoF date TBD (SC’23 during Nov 12-17)
- Looking forward to many more Production submissions

New IO500 Submission Form

New IO500 submission platform launched!

New Features

- Manage account and submissions
- List all previous submissions
- Make new submissions when calls are open
- Allow users to update metadata of submissions until deadline
- Easier for users to see current status of
- Integrated workflow for submission review and publication
- Mandatory fields
- Reproducibility questionnaire

Many thanks to Jean Luca Bez for development!!!

(With additional thanks to Kaushik Velusamy for their valuable contributions)

Thanks to everyone who submitted for their patience

(it will be worth it)

Soliciting volunteers to help with ongoing maintenance and improvements

New Submission System Status

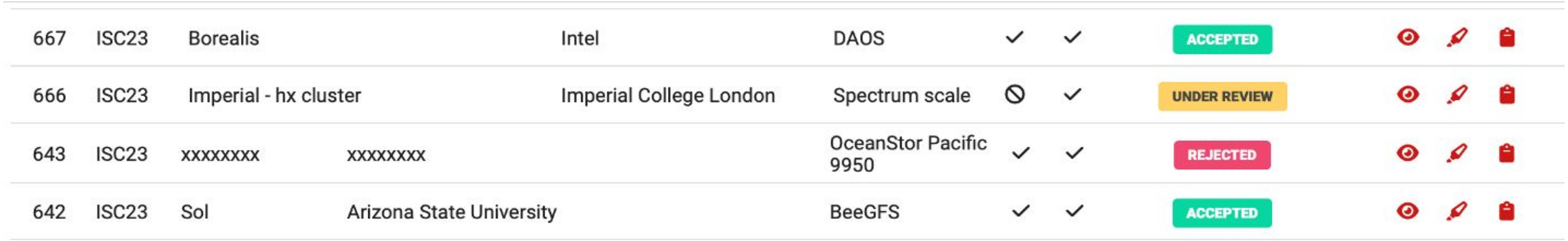
First list release since changing over to new system

- Some issues found with forms by early submitters (e.g. special characters)
- Able to address these problems as they were being reported

Some parts of submission form need further improvement

- Make Research/Test vs. Production submission selection more prominent
- Need to fully import and link historical submission results and data
- Allow JSON submission for Reproducibility Questionnaire
- Storage System mandatory Servers, Storage, Interconnect if no overlapping with compute

Looking forward to a further improved submission process for SC'23



667	ISC23	Borealis	Intel	DAOS	✓	✓	ACCEPTED	👁	✎	📋
666	ISC23	Imperial - hx cluster	Imperial College London	Spectrum scale	⊘	✓	UNDER REVIEW	👁	✎	📋
643	ISC23	xxxxxxxx	xxxxxxxx	OceanStor Pacific 9950	✓	✓	REJECTED	👁	✎	📋
642	ISC23	Sol	Arizona State University	BeeGFS	✓	✓	ACCEPTED	👁	✎	📋

Benchmark Phases and Extended Access Patterns

Benchmark Phases and Extended Access Patterns

- Experimental **`--mode=extended`** run with extra benchmark phases
  - `ior-rnd4k-{read,write}`, `ior-rnd1m-{read,write}`
  - `find-{easy,hard}`, `mdworkbench-{create,bench,delete}`
  - New phases subject to change until final agreement
- Comparison of score between standard / extended modes
  - New phases may change the result of existing phases in rare cases
  - Take only the values of **current** IO500 phases to calculate score
  - Allows to compare new results with historical submissions
- Request that future submission use extended mode
  - Two submissions for ISC22 with extended data, need more feedback
- Need better description for all I/O patterns
  - Motivation, use cases, description of actual IO pattern, …
- Code base is there, please give us feedback anytime

Open Questions About Extended Access Patterns

- Should both 4KB and 1MB patterns be added, or only one (which)?
  - Current IOR implementation needs write phase at same IO size as read
  - `ior-random` IO pattern ensures “dense” files, allows data verify
- Should `ior-random-`**`write`** be counted in the score, or only reads?
  - Relatively few HPC workloads have purely random writes
- Want `find-hard` to be “harder” than just “`find` in `mdtest-hard/` dir”
  - 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 make up `find`?
- Should a directory `mdtest-rename` phase be added?
  - Is this a hierarchical namespace, or flat strings with ‘`/`’ in them?
- Expect runtime would increase by about 30 minutes if all phases added

Reproducibility & List Split

Production and Research List Split

We did it!

Production and Research List Split

- Overall process appeared to go smoothly
- 9 total Production submissions
  - Great start, looking forward to many more for SC23
- Some improvements still to be done
  - Reword “usage” question as “Intended List”
  - Mouseover/help text for system submission fields
  - Tweak questionnaire to clarify fault tolerance, production usage requirements

Reproducibility

- First time that every submission filled out the reproducibility questionnaire
    - It will all be made public after ISC
- Every new ISC23 now has a reproducibility score
    - 16 - Fully Reproducible (all metadata, and software/hardware available to public)
    - 5 - Proprietary (all metadata, but software/hardware unavailable to public)
- Mandatory fields key to making this possible
- Some feedback
    - Make fields less freeform and more standardized
    - Add additional system design questions
    - Upload YAML file
- Next steps
    - Clarify several reproducibility questions based on feedback
    - Upload previous questionnaires to website
- Highlight
    - Huge thanks to Michael Hennecke at Intel for creating a model of how we would like every submission to provide reproducibility information
    - https://github.com/daos-stack/daos-reproducibility/tree/master/io500/isc23/lrz/sng2

Voice of the Community & Open Discussion

Open Floor

- How to collect storage system metadata more easily?
- Can we encourage vendors to support tool and schema development?
- Vote with raised hands
    - random I/O 4KB vs. 1MB, what do people want?
    - random read score only, or read and write score, what do people want?

Collecting Storage System Metadata

- Improved submission schema with templates to simplify collection
  - Supporting storage-system specific schemas
  - Remove uncertainty about the semantics of fields
  - More useful metadata about test system (nodes, storage, network)
- Integrate tools to automatically collect system configuration
  - Support the capturing of accurate system data with each submission
  - Simplify collection of system details for end users
  - Client scripts to capture kernel, filesystem, node, network, and other info
  - Per-filesystem-type script, can be customized to best collect information
  - Seek contributions from users/vendors for scripts for their filesystems
- Explanations with video: https://www.youtube.com/watch?v=R_Fq_ks4hnM