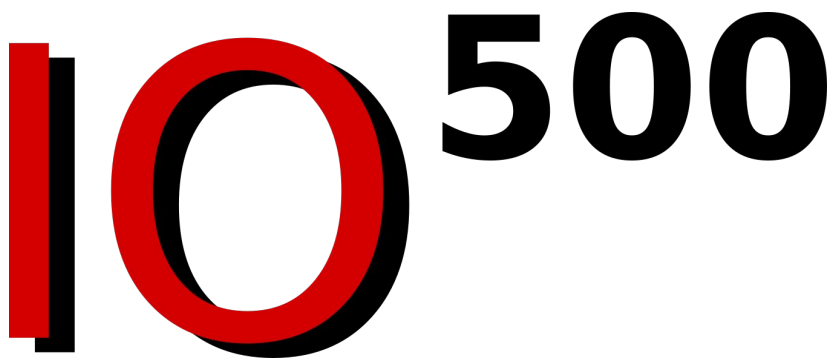


The 10th IO500 and the Virtual Institute of I/O

George Markomanolis, Andreas Dilger, Dean
Hildebrand, Julian M. Kunkel, Jay Lofstead

The logo for the IO500 benchmark. It features the characters 'IO' in a large, bold, red font with a black outline, followed by the number '500' in a smaller, bold, black font.The logo for the Virtual Institute of I/O. It features the characters 'v4io' in a stylized font. The 'v' and '4' are blue, and the 'io' is red. All characters have a black outline and are set against a black rectangular base.

BoF Agenda

1. **Welcome** – George Markomanolis
2. **What's New with IO500** – Andreas Dilger
3. **The New IO500 List Analysis** – Jay Lofstead (remote)
 - **Impact of the list Splitting Proposal**
4. **Award Presentations** – George Markomanolis
5. **Community Presentation** - Frank Gadban
6. **Roadmap** – Julian Kunkel
 - **Benchmark Phases and Extended Access Patterns**
7. **Questions & Discussion Session**

IO500 News

- Versioning of benchmark itself continues to work
 - We check versions in submissions, please use latest `iscYY/scYY` tag
- Exploring usage of new phases in benchmark
 - Open for discussion/modification for future inclusion
 - Optional `--mode=extended` activates experimental phases
 - `ior-rand` (random read/write IOPS - 4KB and/or 1MB)
 - `find-easy`, `find-hard` (many small dirs, single large dir with complex scan)
 - `md-workbench` (concurrent read-write workload)
- Improving storage system metadata schemas with your feedback
 - Simplify accurate system metadata collection for better analysis
- Proposal to split list into separate Production and Research categories
 - Better reflect how systems are being used/deployed, allow better comparisons

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)
- Started integrating 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

IO500 Organization Status

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



Lists and Analysis

10⁵⁰⁰

Growth in Entries and Institutions

IO500 List

ISC22

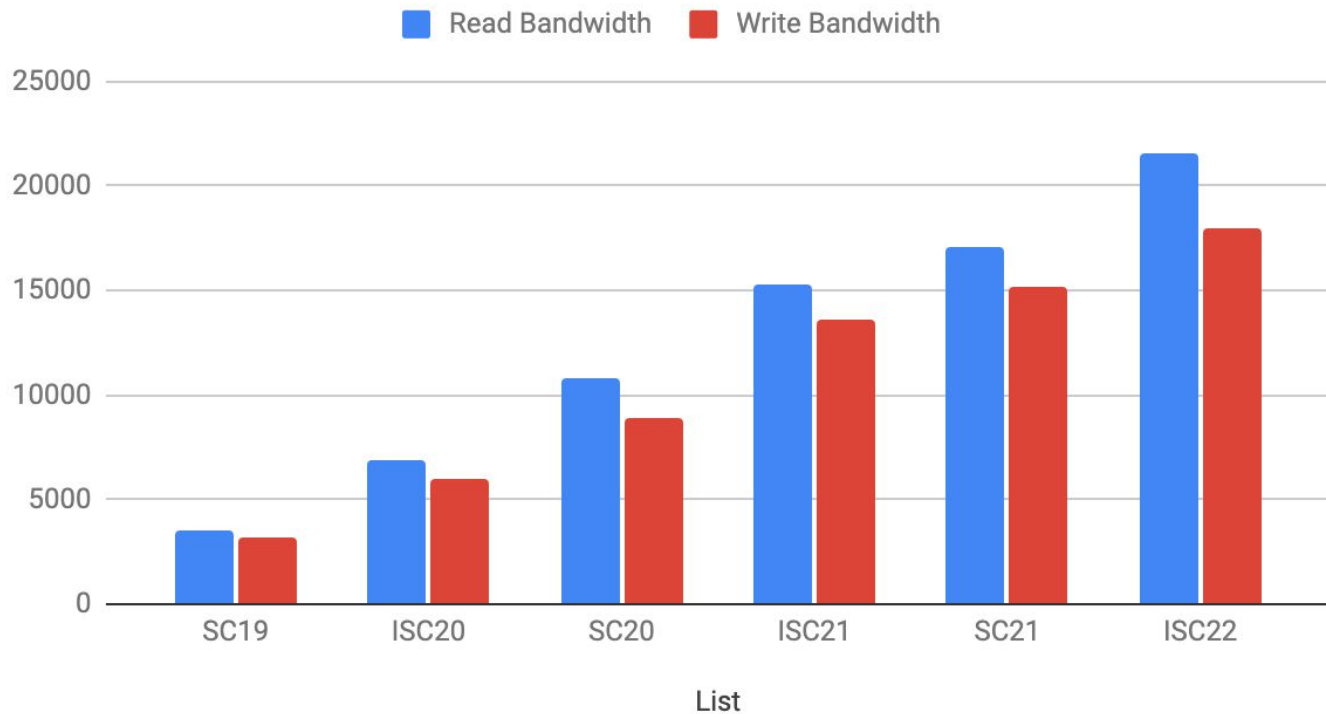
- 15 submissions
 - 7 for IO500
 - 3 for 10-client
 - 5 for both
- 86 list entries
- 67 institutions



Total Bandwidth

IO500 List

Read Bandwidth and Write Bandwidth

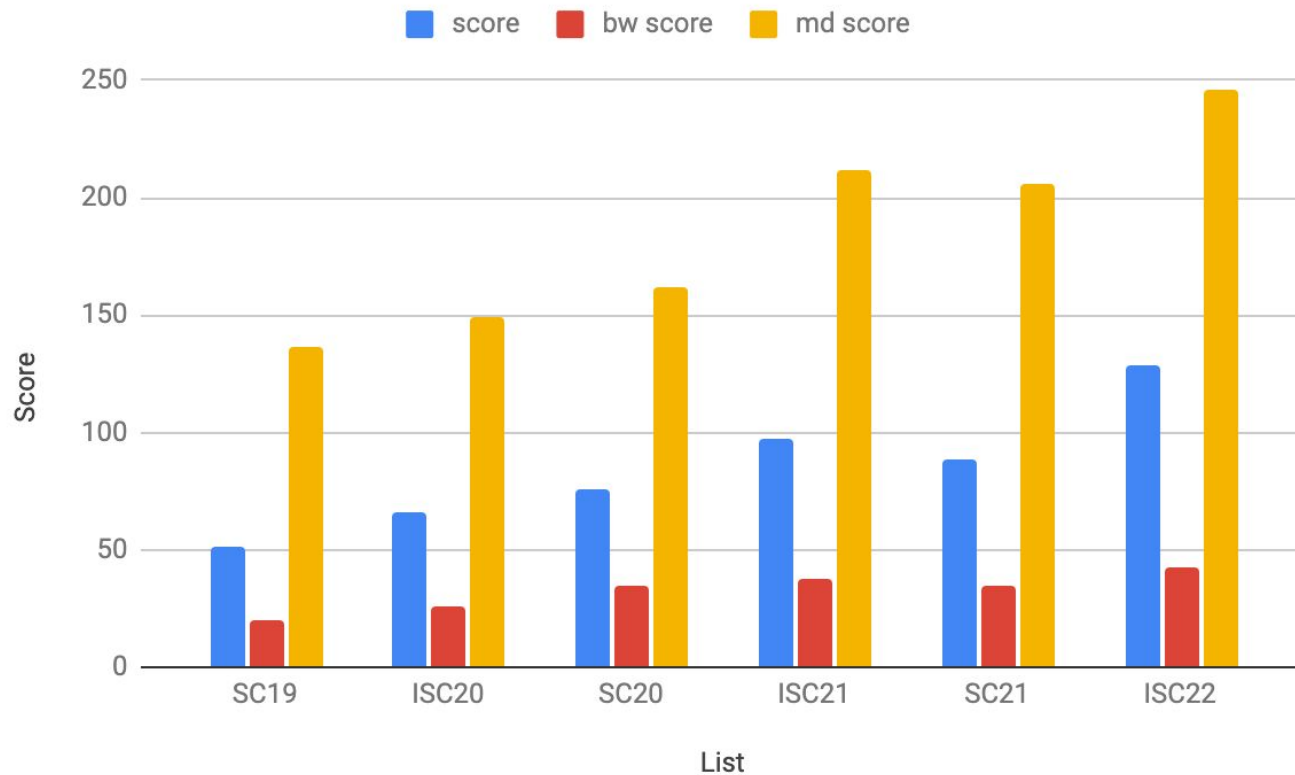


Median Scores

IO500 List

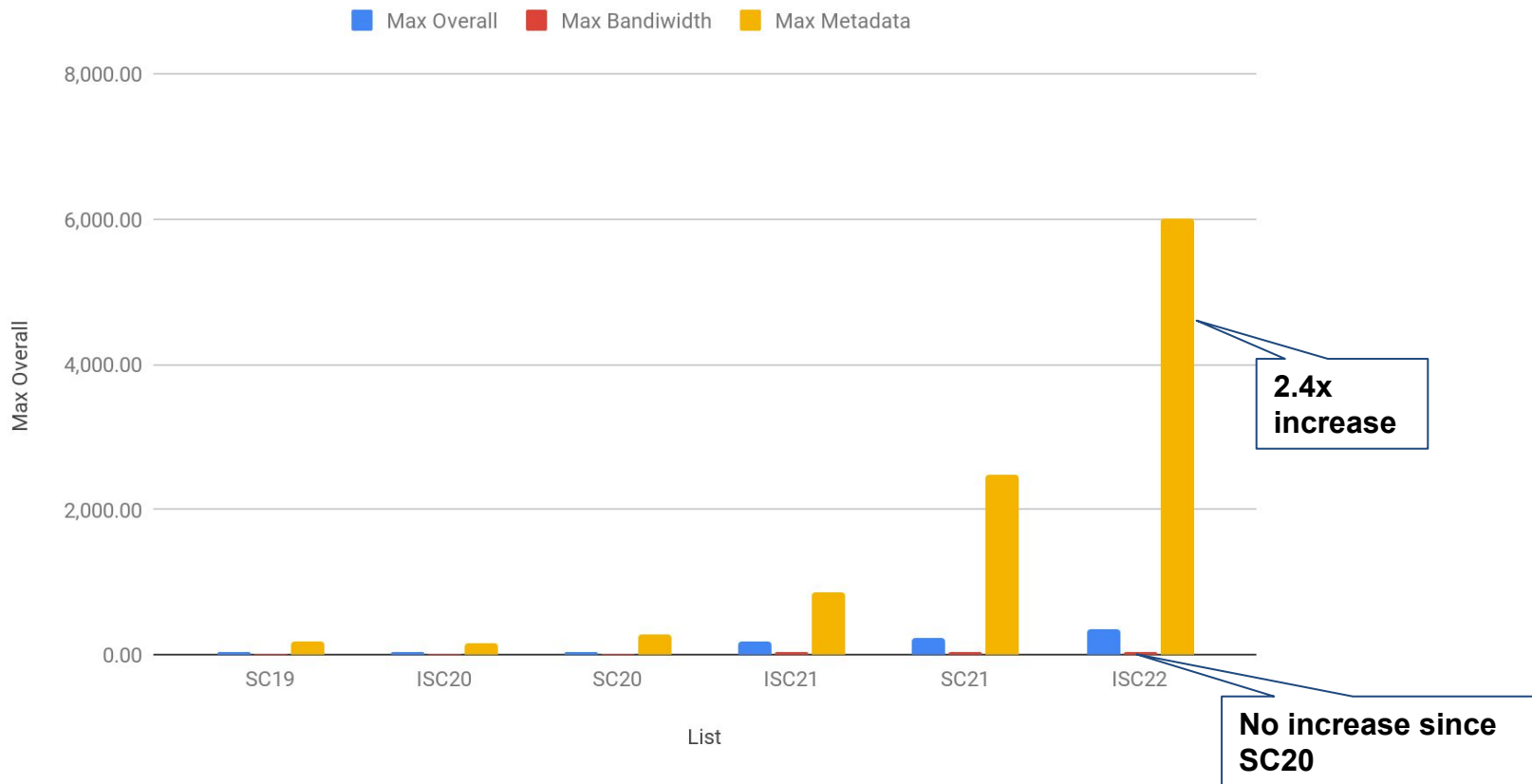
Median scores
reached new highs

(after falling slightly in
SC21)



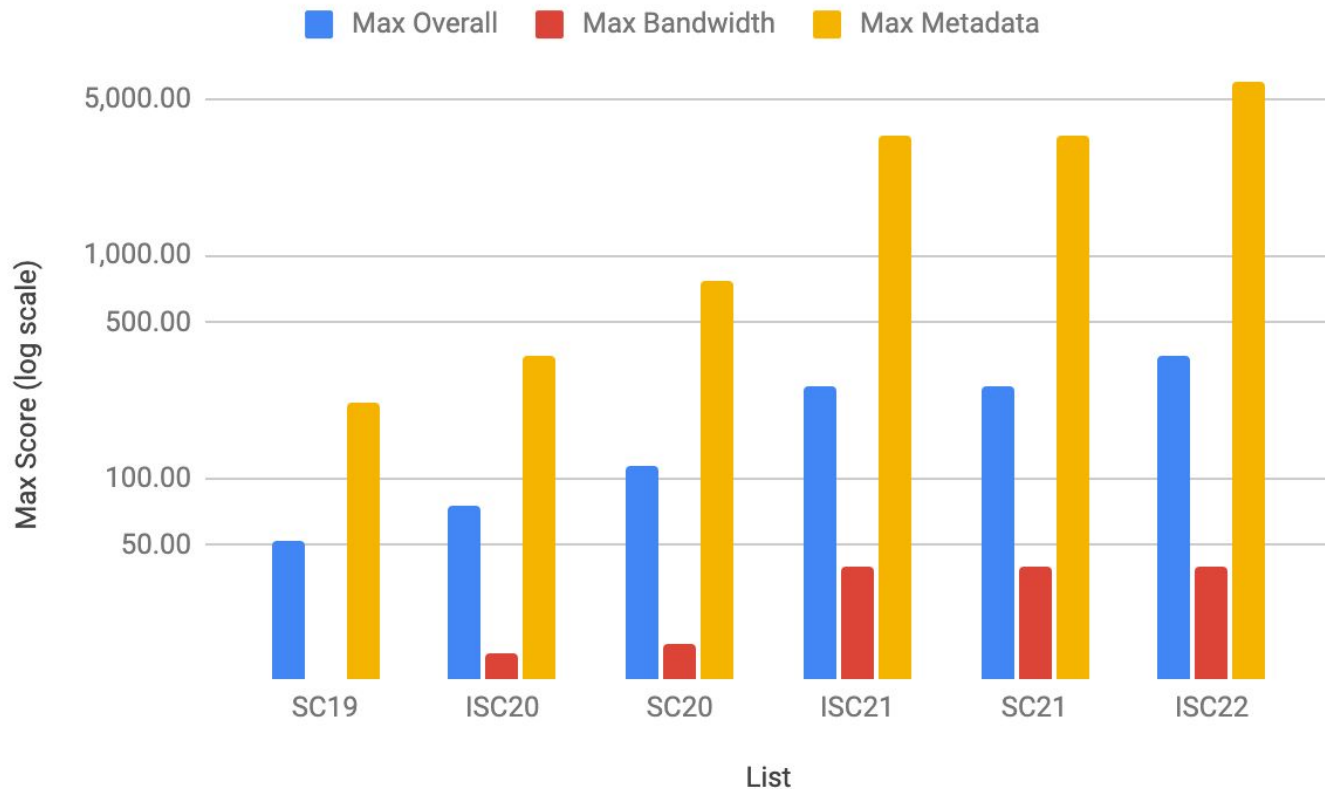
Growth in Max Score per Client

IO500 List



Growth in Max Scores per Client

IO500 - 10-Node Challenge List



Slight growth in metadata

Bandwidth flat since SC20

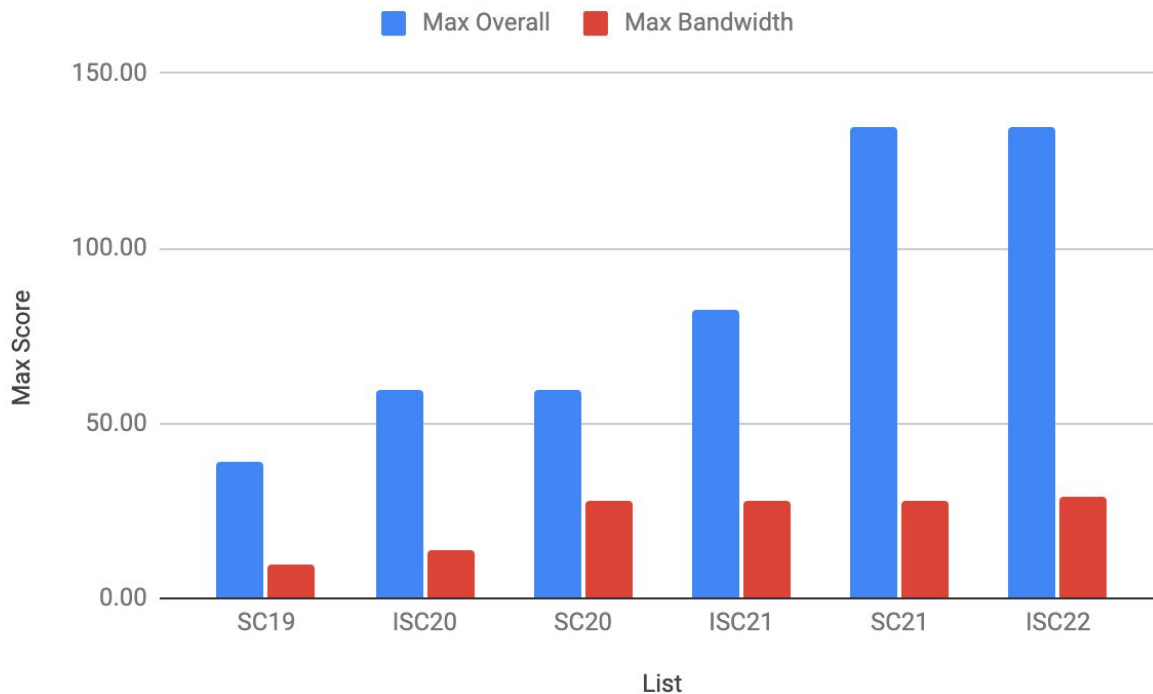
Growth in Max Score per Storage Server

IO500 - List

Per-client scores are growing fast, but...

Per-storage server scores are growing much slower

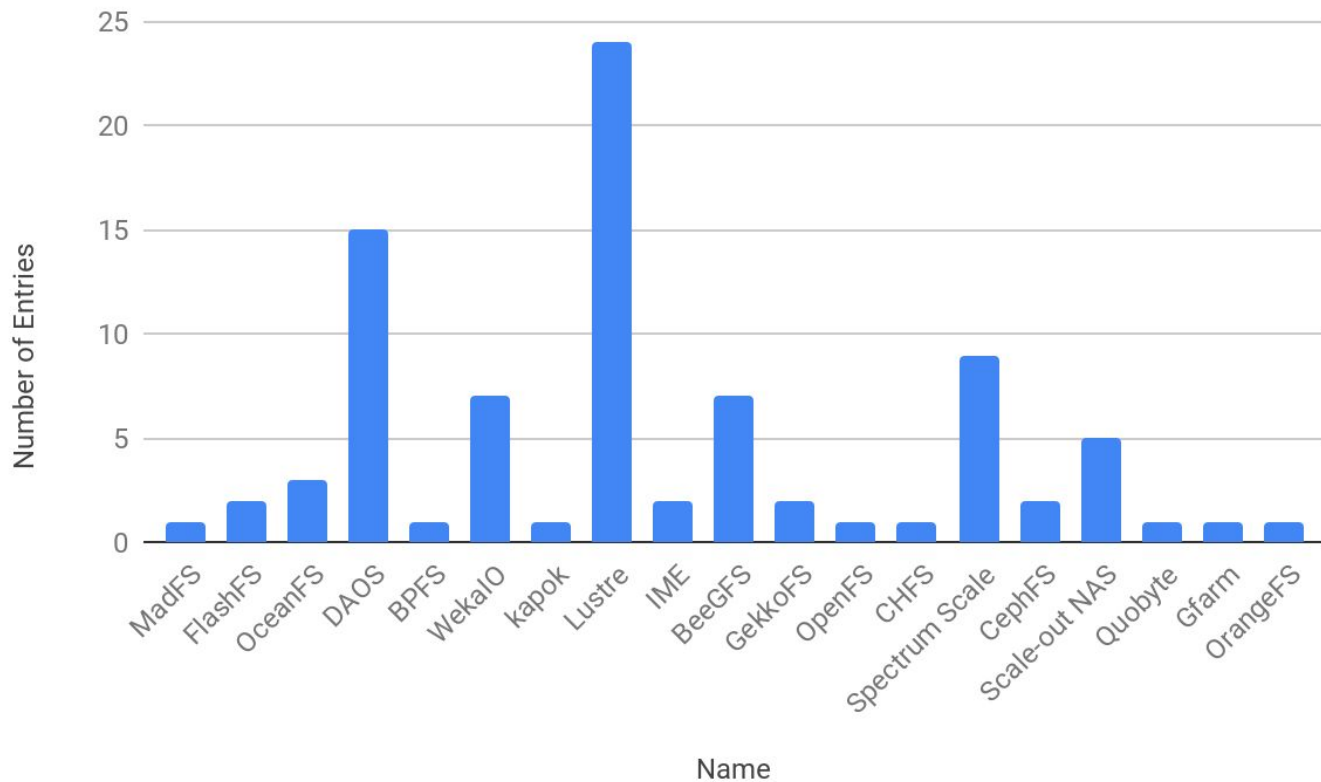
- bandwidth flat for 4 lists
- metadata flat for 2 lists



Note: metadata score per server growth reflected in overall score

Number of File System Entries

IO500 - List



Award Ceremony

10⁵⁰⁰

Six Awards

- Full List
 - Bandwidth
 - Metadata
 - Overall
- 10-Node Challenge List
 - Bandwidth
 - Metadata
 - Overall

10 node challenge - Bandwidth Winner

Sorted by BW

#	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE	BW (GIB/S)	MD (KIOP/S)
1	ISC21	Endeavour	Intel	DAOS		398.77	
2	SC21	OceanStor Pacific	Olympus Lab	OceanFS		317.07	
3	SC21	Athena	Huawei HPDA Lab	OceanFS		314.56	
4	ISC22	SuperMUC-NG Phase2	LRZ	DAOS		209.48	
5	ISC22	Shanhe	National Supercomputing Center in Jinan	flashfs		207.79	
6	ISC21	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng Laboratory	MadFS		193.77	
7	SC20	JUWELS	Forschungszentrum Juelich (FZJ)	IME		178.11	
8	SC20	Frontera	TACC	IME		176.23	
9	ISC20	Wolf	Intel	DAOS		164.77	
10	ISC22	Lenovo-Lenox3	Lenovo	DAOS		115.94	

Certificate

IO500 Performance Certification

This Certificate is awarded to:

Intel (Endeavour)

#1 in the 10 Node Challenge BW Score

IO500



May 2022

IO500 Steering Board

<https://io500.org/list/ISC22/ten>

10-Node Challenge - Metadata Winner

Sorted by MD

#	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE	BW (GIB/S)	MD (KIOP/S)
1	ISC22	Shanhe	National Supercomputing Center in Jinan	flashfs			60,119.50
2	ISC21	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng Laboratory	MadFS			34,777.27
3	SC21	Athena	Huawei HPDA Lab	OceanFS			18,235.71
4	SC21	OceanStor Pacific	Olympus Lab	OceanFS			16,664.88
5	SC21	Kongming	BPFS Lab	BPFS			9,827.09
6	ISC21	Endeavour	Intel	DAOS			8,671.65
7	ISC22	SuperMUC-NG Phase2	LRZ	DAOS			5,109.23
8	ISC21	Lenovo-Lenox	Lenovo	DAOS			3,567.85
9	ISC20	Wolf	Intel	DAOS			3,493.56
10	ISC20	Frontera	TACC	DAOS			3,271.49

Certificate

IO500 Performance Certification

This Certificate is awarded to:
National Supercomputing Center in Jinan (Shanhe)

#1 in the 10 Node Challenge MD Score

IO500



May 2022

IO500 Steering Board

<https://io500.org/list/ISC22/ten>

10-Node Challenge - Winner

Sorted by
score

# ↑	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE ↑	BW (GIB/S)	MD (KIOP/S)
1	ISC22	Shanhe	National Supercomputing Center in Jinan	flashfs	3,534.42	207.79	60,119.50
2	ISC21	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng Laboratory	MadFS	2,595.89	193.77	34,777.27
3	SC21	Athena	Huawei HPDA Lab	OceanFS	2,395.03	314.56	18,235.71
4	SC21	OceanStor Pacific	Olympus Lab	OceanFS	2,298.69	317.07	16,664.88
5	ISC21	Endeavour	Intel	DAOS	1,859.56	398.77	8,671.65
6	ISC22	SuperMUC-NG Phase2	LRZ	DAOS	1,034.55	209.48	5,109.23
7	SC21	Kongming	BPFS Lab	BPFS	972.60	96.26	9,827.09
8	ISC20	Wolf	Intel	DAOS	758.71	164.77	3,493.56
9	ISC21	Lenovo-Lenox	Lenovo	DAOS	612.87	105.28	3,567.85
10	ISC22	Lenovo-Lenox3	Lenovo	DAOS	544.18	115.94	2,554.14

Certificate

IO500 Performance Certification

This Certificate is awarded to:
National Supercomputing Center in Jinan (Shanhe)

#1 in the 10 Node Challenge

IO500



May 2022

IO500 Steering Board

<https://io500.org/list/ISC22/ten>

Full list - Bandwidth Winner

Sorted by BW

#	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE	(GIB/S)	(KIOP/S)
1	ISC21	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng Laboratory	MadFS		3,421.62	
2	SC20	Oakforest-PACS	JCAHPC	IME		697.20	
3	ISC20	NURION	Korea Institute of Science and Technology Information (KISTI)	IME		515.59	
4	ISC21	Endeavour	Intel	DAOS		398.77	
5	ISC20	Wolf	Intel	DAOS		371.67	
6	ISC22	SuperMUC-NG Phase2	LRZ	DAOS		321.75	
7	SC21	OceanStor Pacific	Olympus Lab	OceanFS		317.07	
8	SC21	Athena	Huawei HPDA Lab	OceanFS		314.56	
9	ISC20	BeeGFS on Oracle Cloud	Oracle Cloud Infrastructure	BeeGFS		293.05	
10	ISC22	Cumulus	University of Cambridge	DAOS		283.19	

Certificate

IO500 Performance Certification

This Certificate is awarded to:
Pengcheng Laboratory (Cloudbrain-II)

#1 in the IO500 BW Score

IO500



May 2022

IO500 Steering Board

<https://io500.org/list/ISC22/io500>

Full list - Metadata Winner

Sorted by MD

#	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE	BW (GIB/S)	MD ↑ (KIOP/S)
1	ISC21	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng Laboratory	MadFS			396,872.82
2	ISC22	Shanhe	National Supercomputing Center in Jinan	flashfs			60,119.50
3	SC21		Huawei Cloud	Flashfs			37,034.00
4	SC21	Athena	Huawei HPDA Lab	OceanFS			18,235.71
5	SC21	OceanStor Pacific	Olympus Lab	OceanFS			16,664.88
6	SC21	Kongming	BPFS Lab	BPFS			9,827.09
7	ISC21	Endeavour	Intel	DAOS			8,671.65
8	ISC20	Wolf	Intel	DAOS			8,649.57
9	ISC20	Frontera	TACC	DAOS			7,449.56
10	ISC22	SuperMUC-NG Phase2	LRZ	DAOS			5,844.40

Certificate

IO500 Performance Certification

This Certificate is awarded to:
Pengcheng Laboratory (Cloudbrain-II)

#1 in the IO500 MD Score

IO500



May 2022

IO500 Steering Board

<https://io500.org/list/ISC22/io500>

Full list - Winner

# ↑	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE ↑	BW (GIB/S)	MD (KIOP/S)
1	ISC21	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng Laboratory	MadFS	36,850.40	3,421.62	396,872.82
2	ISC22	Shanhe	National Supercomputing Center in Jinan	flashfs	3,534.42	207.79	60,119.50
3	SC21	Athena	Huawei HPDA Lab	OceanFS	2,395.03	314.56	18,235.71
4	SC21	OceanStor Pacific	Olympus Lab	OceanFS	2,298.69	317.07	16,664.88
5	SC21		Huawei Cloud	Flashfs	2,016.70	109.82	37,034.00
6	ISC21	Endeavour	Intel	DAOS	1,859.56	398.77	8,671.65
7	ISC20	Wolf	Intel	DAOS	1,792.98	371.67	8,649.57
8	ISC22	SuperMUC-NG Phase2	LRZ	DAOS	1,371.30	321.75	5,844.40
9	ISC22	Cumulus	University of Cambridge	DAOS	1,107.17	283.19	4,328.68
10	ISC21	Lenovo-Lenox	Lenovo	DAOS	988.99	176.37	5,545.61

Certificate

IO500 Performance Certification

This Certificate is awarded to:
Pengcheng Laboratory (Cloudbrain-II)

#1 in the IO500

IO500



May 2022

IO500 Steering Board

<https://io500.org/list/ISC22/io500>

List of Awarded Systems in the Ranked Lists

No change of the awarded systems this list

10-Node	Bandwidth	Intel Endeavour	DAOS	398.77	GiB/s
	Metadata	NSC Jinan Shanhe	FlashFS	60119.50	kIOPS
	Overall	NSC Jinan Shanhe	FlashFS	3534.42	score
IO500	Bandwidth	Pengcheng Cloudbrain-II	MadFS	3421.62	GiB/s
	Metadata	Pengcheng Cloudbrain-II	MadFS	396872.82	kIOPS
	Overall	Pengcheng Cloudbrain-II	MadFS	36850.37	score

Reproducibility & List Split Progress

10⁵⁰⁰

IO500 Reproducibility and List Split Progress

Progress

- Finalized both proposals
 - Lists will be called “Production” and “Research”
 - “Production” definition, reproducibility score, all finalized
- Created early version of reproducibility questionnaire for ISC22

Next Steps

- Revamp questionnaire possibly into Google Form (will require Google auth)
- Clarify (and enforce) mandatory fields for reproducibility score in submission
- Split lists and publish all submitted information and reproducibility score
- Build tool to gather full reproducibility scripts
- Build review committee

Reproducibility Stats

- 10 of 15 Submissions completed questionnaire

ISC22 IO500 Submissions “What If...” Production/Research List

Production

SuperMUC-NG-EC - DAOS

Oracle Cloud - WEKA (but need more details)

Lenovo-Lenox3-EC - DAOS

CTPAI - DAOS

Research

Shanhe - FlashFS (No public access, lack H/A, research)

SuperMUC-NG - DAOS (No H/A, unlike -EC config)

UCambridge - DAOS (missing Questionnaire)

Lenovo-Lenox3 - DAOS (No H/A, unlike -EC config)

UCambridge - Lustre (missing Questionnaire)

Wu Dong - OpenFS (Research)

AI400X2 - Lustre - DDN (missing Questionnaire)

Omnibond-Google - OrangeFS (No H/A)

Not Official

Community Presentation

10⁵⁰⁰



ISC

High Performance

TRANSFORMING

THE FUTURE

MAY 29 – JUNE 2, 2022 | HAMBURG, GERMANY

IO500-S3

Frank Gadban¹, Julian Kunkel²

¹University of Hamburg, 20146 Hamburg, Germany

²University of Göttingen / GWDG

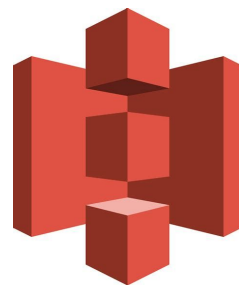
S3 Performance inside HPC - Status Quo



- In the Cloud :
 - Amazon S3 is the de-facto storage API for object-storage.
- In HPC:
 - HPC and Cloud convergence
 - Scientific/Big Data apps run on the same Infrastructure
 - Storage vendors offer S3 API alongside POSIX
 - Which storage is the most suitable ?
 - How to test the performance of S3 inside HPC ?



IBM Cloud



SCALITY



WEKA



HUAWEI



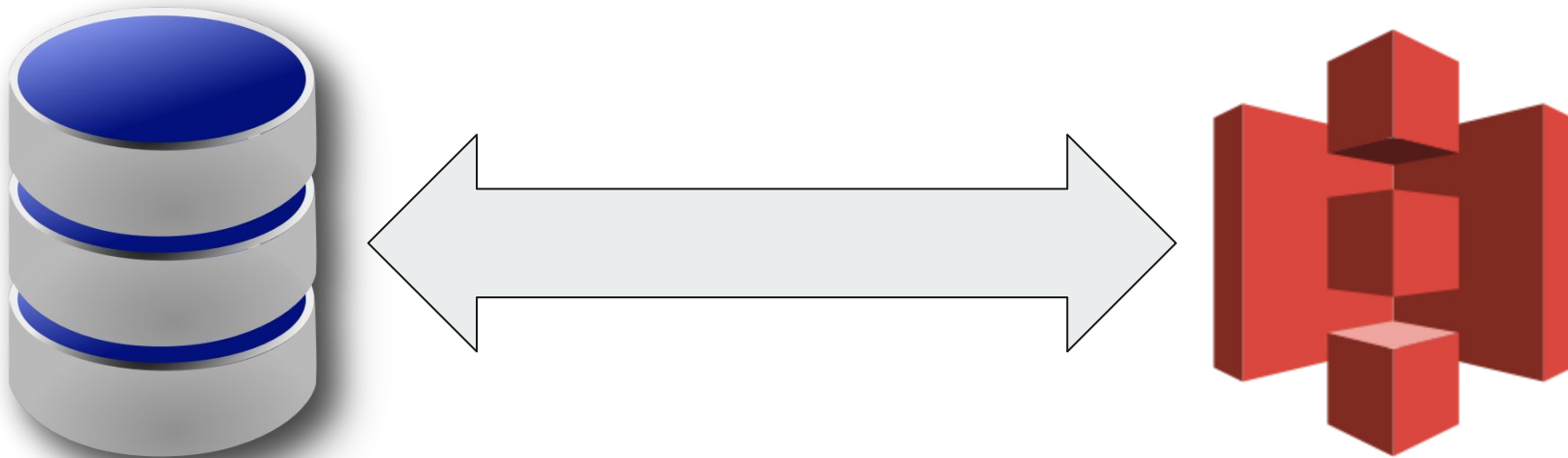
ceph

S3 Performance inside HPC - Status Quo

Performance of the S3 API for HPC I/O applications received too little attention

Most work till now focused on :

- the download performance of Amazon S3
- specific implementations and not a wide range of S3 compatible storage
- the usage of unknown/unpublished tools -> No Open Source
- No lightweight Implementation / C language








IO500-S3














Goal: Enable comparative testing of S3 storage by using the IO500




- Since the setup is non-trivial, submitted a patch - available under contrib/s3




 **IO500** / **io500** Public


 Notifications  Fork 20  Star 45 

 Code  Issues 7  Pull requests 2  Actions  Projects  Wiki  Security  Insights

 main  **io500** / contrib / s3 / 

 **Frankgad** updates prepare-s3  a512138 15 days ago  History

..		
 README.md	updates prepare-s3	15 days ago
 config-s3.ini	moved to contrib folder	21 days ago
 prepare-s3.sh	updates prepare-s3	15 days ago

 **README.md**

io500-s3

The following explains how to use the IO500 to benchmark S3 compatible storage. S3 runs are not fully compliant at the time of writing because find is not yet supported, but this might be useful for testing and comparing different S3 implementations. Some people might find it helpful since it is relatively complicated to set up. (Hopefully, in the long run, IO500 compliant runs would be possible with S3)

- The S3 API specification does not provide any way to search for a file inside buckets.
 - Any provided S3 search functionality is done either on
 - the client-side
 - using an intermediate service to accomplish that
 - Therefore, find can not be implemented
- Using the IOR / libS3, each I/O is stored as an independent object
 - Suits for a best case performance analysis
 - Multi-part might be implemented in the future.
 - The alternative S3 IOR plugin claims to support multi-part (but couldn't get it working)

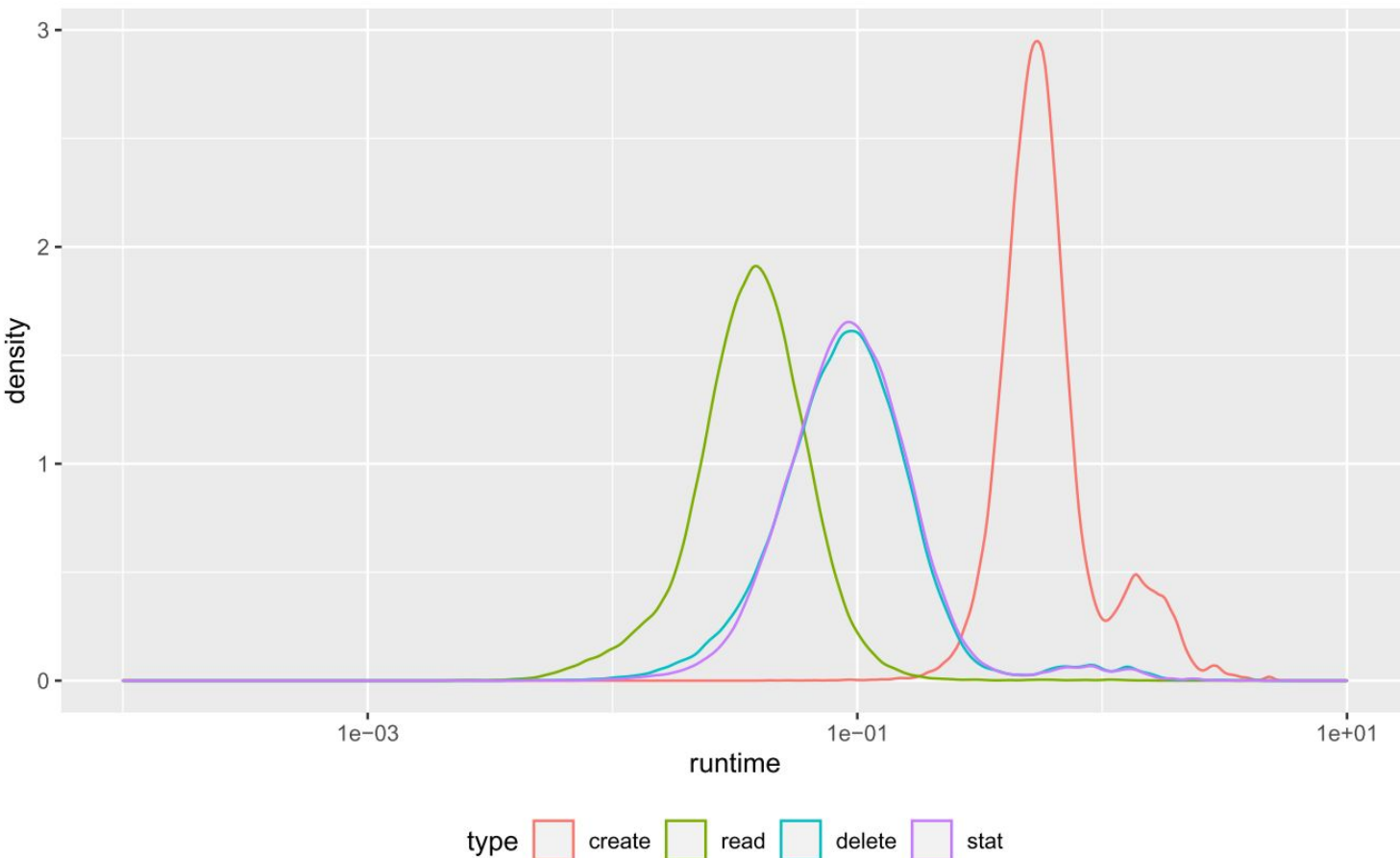
Performance of MinIO Gateway on 4 nodes with 20 PPN

Benchmark	Metric	Unit	Lustre	MinIO		Local-Gw
				Disjoint-Gw	Local-Gw	% of Lustre
IO500	ior-easy-write	GiB/s	18.671	0.153	0.286	1.5%
	mdtest-easy-write	kiOPS	5.892	0.088	0.132	2.2%
	ior-hard-write	GiB/s	0.014	0.003	0.006	45.7%
	mdtest-hard-write	kiOPS	5.071	0.036	0.076	1.5%
	ior-easy-read	GiB/s	11.475	0.693	2.071	18.1%
	mdtest-easy-stat	kiOPS	24.954	1.198	4.092	16.4%
	ior-hard-read	GiB/s	0.452	0.029	0.094	20.7%
	mdtest-hard-stat	kiOPS	18.296	1.281	3.968	21.7%
	mdtest-easy-delete	kiOPS	9.316	0.025	0.023	0.3%
	mdtest-hard-read	kiOPS	6.950	0.449	1.636	23.5%
	mdtest-hard-delete	kiOPS	4.863	0.029	0.025	0.5%

IO500 results comparing S3 Cloud providers

Benchmark/System	Unit	Wasabi	IBM	Google	MinIO-local-gw
Score Bandwidth	MiB/s	0,007	1,642	0,46	12,62
ior-easy-write	MiB/s	2.35	35.00	13.35	46.39
mdtest-easy-write	IOPS	13.04	81.72	21.79	27.96
ior-rnd-write	MiB/s	0.01	0.23	0.07	1.231
mdworkbench-bench	IOPS	5.75	47.23	12.83	15.25
ior-easy-read	MiB/s	1.20	45.37	7.81	73.86
mdtest-easy-stat	IOPS	20.92	145.09	51.10	260.97
ior-hard-read	MiB/s	0.05	5.59	1.38	6.01
mdtest-hard-stat	IOPS	20.74	149.64	49.48	297.62
mdtest-easy-delete	IOPS	10.35	35.02	9.37	81.06
mdtest-hard-read	IOPS	8.54	70.06	18.90	130.36
mdtest-hard-delete	IOPS	10.28	35.25	9.48	94.32

Bonus: Latency Analysis : MinIO local-gw



- Using IO500 extended – md-workbench
- Analysis of the latency of each type of operations.

- # Analyzing the Performance of the S3 Object Storage API for HPC Workloads

[illegible]

Roadmap

10 500

Roadmap for the IO500

- Implementation of the list split
 - Introducing the reproducibility award with **mandatory** questionnaire
- Improvements to system schema for filesystem types
 - **More system metadata fields will be mandatory for better comparisons**
 - Continue to improve with more use and feedback
 - Extending the scripts to automatically collect more system metadata
- Fill in gaps in IO500 to improve usage patterns
 - Collect and evaluate results for new benchmark phases
 - Not officially part of benchmark score yet, still some flexibility to modify
 - Document rationales for existing/new benchmark phases
- New io500.org site for submissions for next list - thanks Jean Luca

New IO500 submission platform

IO500HUB
ACCESS

Goals

- allow users to update metadata of submissions
- integrated workflow for review and publication
- easier submission and results analysis

AUTHENTICATION

EMAIL


PASSWORD


[REGISTER](#) [RESET PASSWORD](#) [LOGIN](#)


IO500HUB
USER ACCESS


[My Submissions](#) [New Submission](#) [Account](#) [Logout](#)

UPLOAD NEW FILES

RESULTS FILE (.TAR.GZ)
 [Browse...](#) No file selected.

JOB SCRIPT
 [Browse...](#) No file selected.

JOB OUTPUT
 [Browse...](#) No file selected.

SYSTEM INFORMATION FILE (.JSON)
 [Browse...](#) No file selected.

[SUBMIT](#)

Roadmap for the IO500

- Prospect feature: Support for GPU Direct
- Community meeting
 - In preparation with call for submission, propose September 7th
- SC 22 Roadmap
 - Call for submission: September ~15th
 - Testing phase ends: October ~15th
 - Code freeze, but please test before!
 - Submission deadline: November 1st
 - List release: November 1Xth (BoF date TBD)

Benchmark Phases and Extended Access Patterns

10500

IO500 Survey Results

- Most users want that the benchmark evolves

- Should test concurrent metadata ops (53%)
- Should split find into easy/hard (38%)
- Should add random read 4k (38%)
- Should add random write 4k (35%)
- Should add random read 1M (36%)
- Should add random write 1M (35%)
- Benchmark should stay as it is (22%)



Part of the `-mode=extended` run

Benchmark Phases and Extended Access Patterns

- Extended mode with extra phases
 - We had 2 submissions for ISC22 with extended data
- Pending issues
 - Comparison of score between standard / extended
 - New phases may change the result of existing phases in rare cases
- We will request all submission for SC22 to use extended mode
 - Take only the values of current IO500 to calculate score
 - Allow to compare results with historical submissions
- The committee will work on specification of all I/O patterns
 - Motivation, use cases, ...,
- Code base is there, please give us feedback anytime

Voice of the Community & Open Discussion

10⁵⁰⁰

Open Floor

- How to collect system metadata more easily?
- Can we encourage vendors to support the tool development and schema development?
- Vote with raised hands -> Survey results, random I/O 4KB vs. 1MB, what do people want?