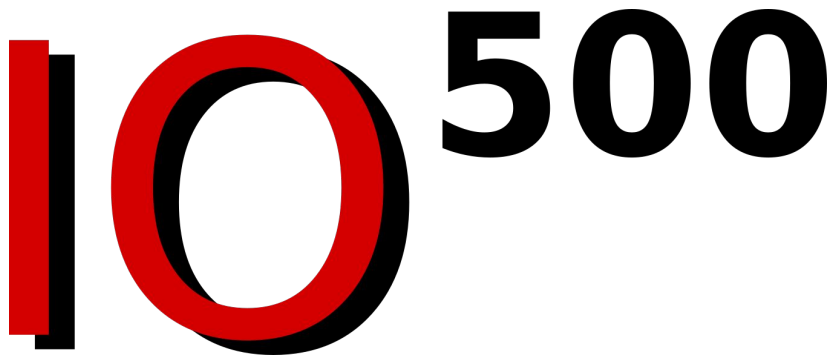


BoF: The IO-500 and the Virtual Institute of I/O

George Markomanolis, Jay Lofstead,
John Bent, Julian M. Kunkel

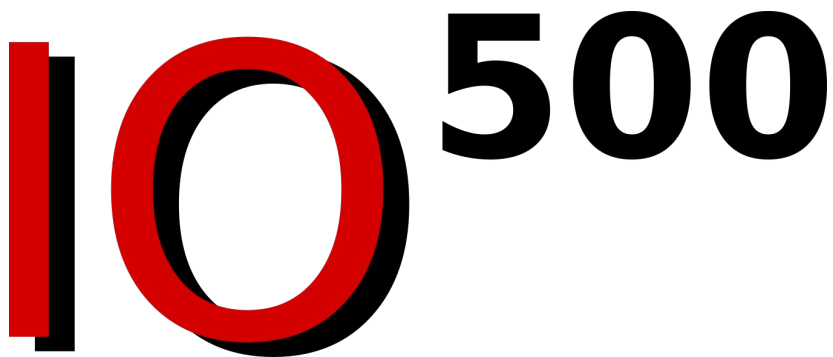
The logo for the IO500 benchmark. It features the letters 'IO' in a large, bold, red font with a black outline. To the right of 'IO' is the number '500' in a bold, black font.The logo for the Virtual Institute of I/O. It features the letters 'vI4IO' in a stylized font. The 'v' and 'I' are blue, the '4' is white with a black outline, and the 'IO' is red with a black outline. The entire logo is set against a black background.

BoF Agenda

1. **The Virtual Institute for IO** (10 min) – Julian Kunkel
2. **What's new with IO-500** (5 min) – George Markomanolis
3. **Community lightning talks** (5 min each)
 - a. Rationalizing Public Clouds HPC Performance – Vinay Gaonkar
 - b. I/O performance variability in practice – Glenn Lockwood
4. **The new IO-500 list** (5 min) – George
 - a. **Award ceremony** (2 min)
5. **Analysis** (5 min) – Julian
6. **Roadmap** (2 min) – Jay
7. **Voice of the community & Open Discussion** (20 min) – Jay Lofstead

IO-500

George Markomanolis, Jay Lofstead,
John Bent, Julian M. Kunkel

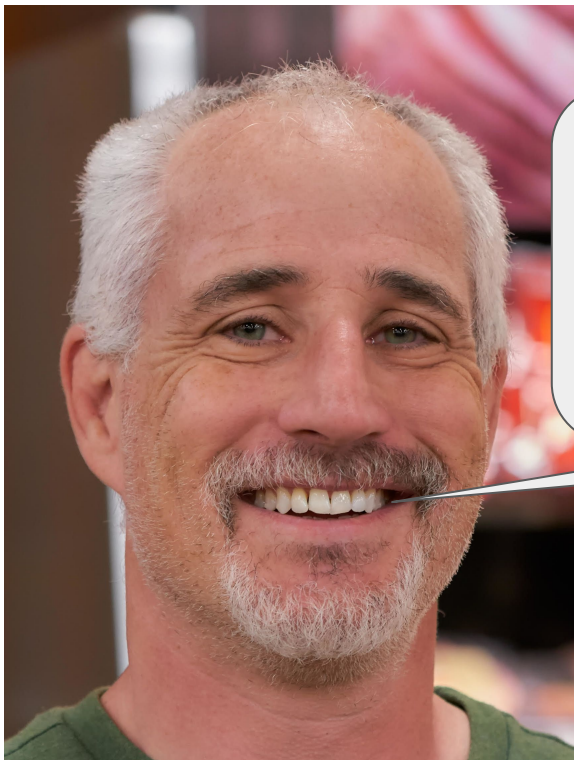
The logo features the letters 'IO' in a large, bold, red font with a thick black outline. To the right of 'IO' is the number '500' in a bold, black, sans-serif font. The entire logo is set against a white background.

IO500

The logo features the letters 'v4lo' in a stylized, blocky font. The 'v' and 'l' are blue, the '4' is white with a black outline, and the 'o' is red with a thick black outline. The letters are set against a white background.

v4lo

Apologies in absentia



- Sorry I'm not with you
- Just started new job; important meeting later today in CA...
- Thanks for all the awesome submissions
- Please don't abuse the other committee members too badly
- See you at SC!

IO-500 What's new?

George S. Markomanolis,
The IO-500 and the Virtual Institute of I/O
Frankfurt, Germany, ISC'19

18 June 2019

ORNL is managed by UT-Battelle, LLC for the US Department of Energy



U.S. DEPARTMENT OF
ENERGY

IO500

- mdtest was patched (an out of cycle list was released):
 - <https://www.vi4io.org/io500/list/19-01/start>
 - <https://www.vi4io.org/io500/list/19-01/10node>
- Creating more standard procedures
 - Rules for running IO-500 (<https://www.vi4io.org/io500/rules/submission>)
 - Vendor engagement (io-500-vendors@vi4io.org)
 - Certificates
- Organization
 - More frequent meetings
 - Clarified rules

IO500

- Procedure to modify IO-500 and/or add new benchmarks (<https://www.vi4io.org/io500/rules/proposals>)
- Deadlines moved earlier (June 10th, 2019)
- Informing submitters about their ranking
- Exhaustive submission checking

Submission Rules

Submission Rules

The following rules should ensure a fair comparison of the IO-500 results between systems and configurations. They serve to reduce mistakes and improve accuracy.

1. The latest version of io500.sh in GitHub must be used and all binaries should be built according to the included build instructions.
2. All required phases must be run and in the same order as they appear in the io500.sh script.
3. Read-after-write semantics: The system must be able to correctly read freshly written data from a different client after the close operation on the writer has been completed.
4. All create phases should run for at least 300 seconds; the stonewall flag can be used as an optional convenience to help
5. For SC19: The stonewall flag must be used and must be set to 300.
6. There can be no edits made to the script beyond changing the allowed variables and adding commands to configure the storage system (e.g. setting striping parameters).
 - For example, there can be no artificial delays added within the script such as calling 'sleep' between phases.
 - No edits are allowed to the utilities/io500_fixed.sh scripts.
 - You may not overwrite any parameters that are set in utilities/io500_fixed.sh.
7. The file names for the mdtest and IOR output files may not be pre-created.
8. You must run the benchmark on a single storage system.
9. You must configure the batch scheduler to allocate the ranks in blocks, e.g. if you are running with five ranks per client node then rank 0-5 must be placed on Node0 and 6-10 on Node1. You should verify the appropriate placement. This ensures the proper shifting in IOR.
10. All data must be written to persistent storage within the measured time for the individual benchmark, e.g. if a file system caches data, it must ensure that data is persistently stored before acknowledging the close.
11. Submitting the results must be done in accordance with the instructions on our submission page.
12. If a tool other than the included pfind is used for the find phase, then it must follow the same input and output behavior as the included pfind.

Please send any requests for changes to these rules or clarifying questions to our mailing list.

Reminder about pfind/sfind/io500.sh

- pfind/sfind are currently provided by the IO500 committee as a convenience
- io500.sh attempts to determine whether a run is invalid
- However, submitters are responsible for checking results for compliance with the rules
 - i.e., just because it seems to run successfully, does not mean it will be accepted as valid

Edit or add functionalities to IO-500

Change Request

The IO-500 aims to be a robust and long-living benchmark. Nevertheless, the community recognizes the need to consider modifications occasional modifications such as including new access patterns, removing deprecated access patterns, or any other modifications deemed necessary by the community. Therefore, we have established a process to add further benchmarks, which works as follows:

1. A member of the community prepares a (up to) 1-page proposal for the new access pattern to include. This should include a motivation, a rough sketch of the access pattern and justification why the pattern is important. This proposal can then be sent to the community mailing list or the steering board. Deadline: 1 month before the next community meeting – at the moment, these are the birds-of-a-feather sessions at ISC or Supercomputing.
2. The steering board will give feedback to the technical quality of the proposal.
3. The member is given the opportunity to present the proposal at the next following community IO-500 meeting.
4. Given there are no technical concerns, the IO-500 benchmark will be modified for the next submission period to allow the execution of a benchmark that represents the pattern as an *optional* benchmarking step. Additionally, the optional field is introduced into subsequent lists and the changes to the benchmark are documented on the webpage.
5. The optional pattern is kept for at least two subsequent IO-500 lists and community meetings. The results and effectiveness of the new pattern are discussed during the community meetings. As a result, it may be removed, remain optional, or may become mandatory.

The committee can be reached at ✉ committee@io500.org.

IO-500

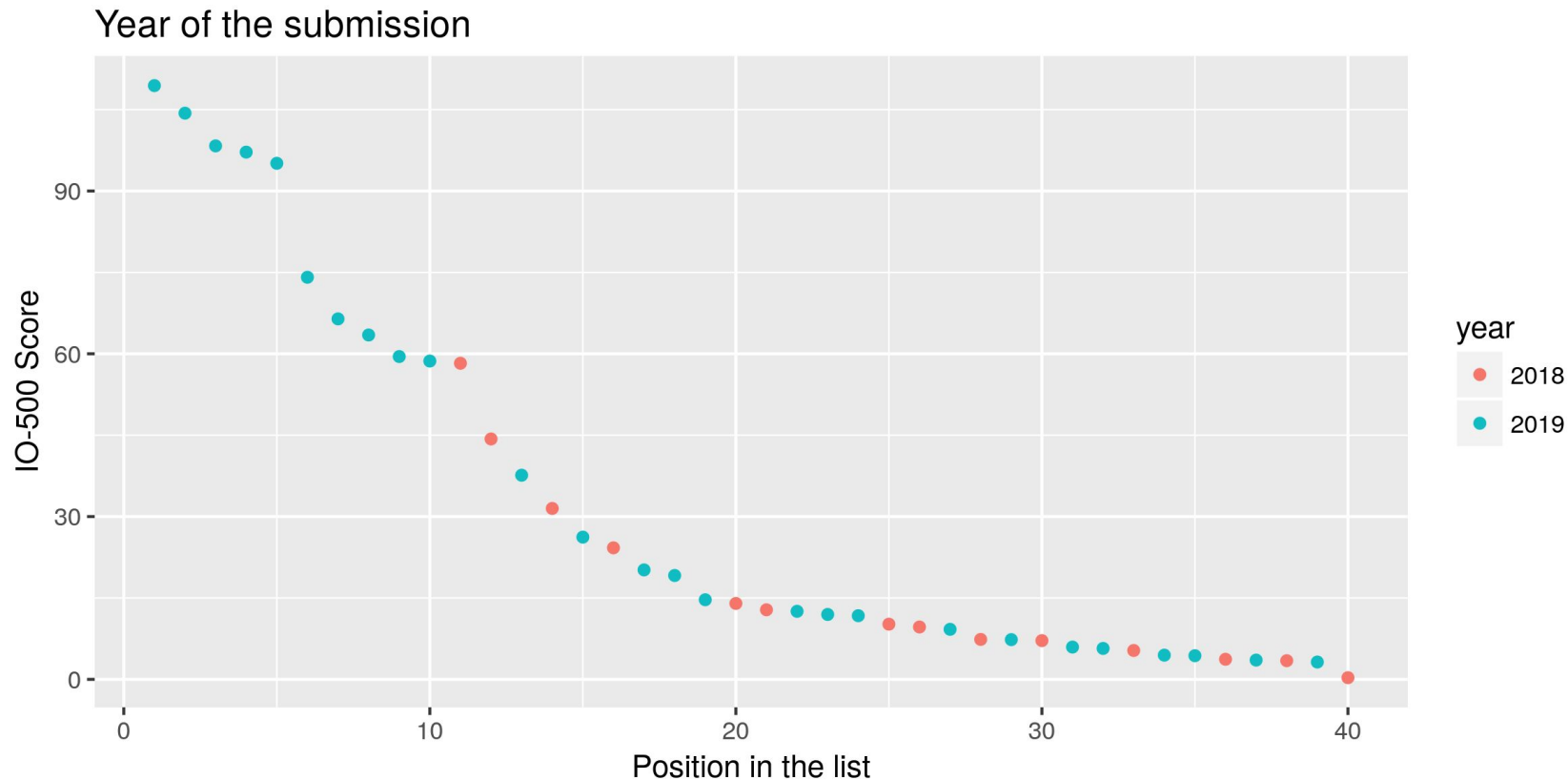
We have more than 100 **total** submissions!

Thank you!

The new IO-500 list and analysis

IO⁵⁰⁰

10 Node Challenge -- Ranked List



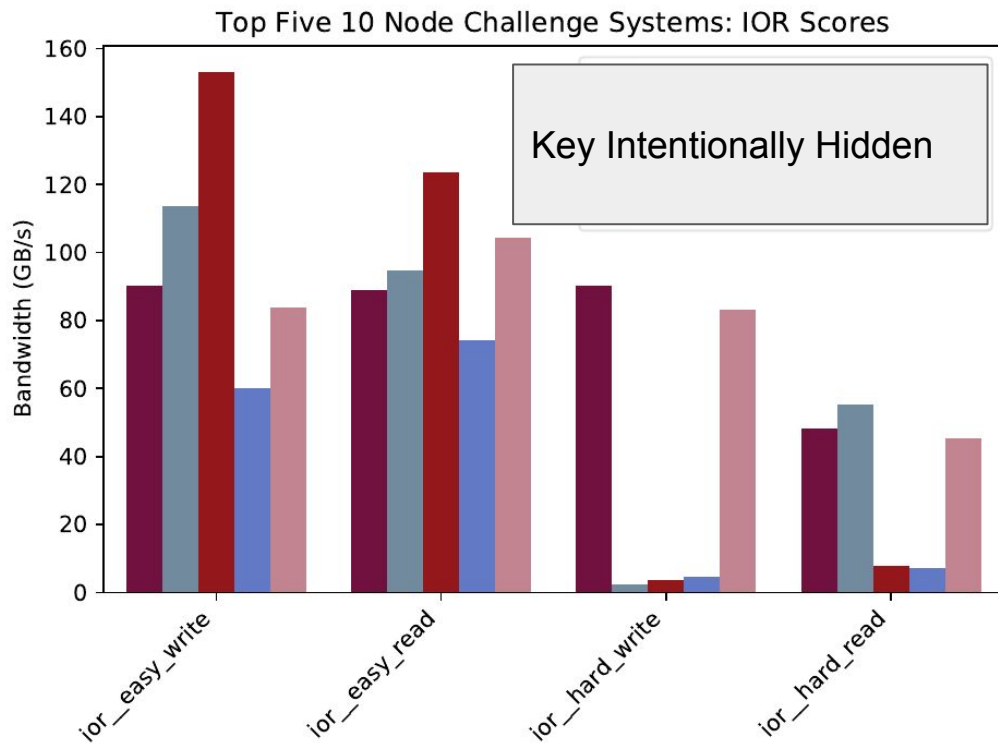
IO-500 10 nodes Bandwidth Winner



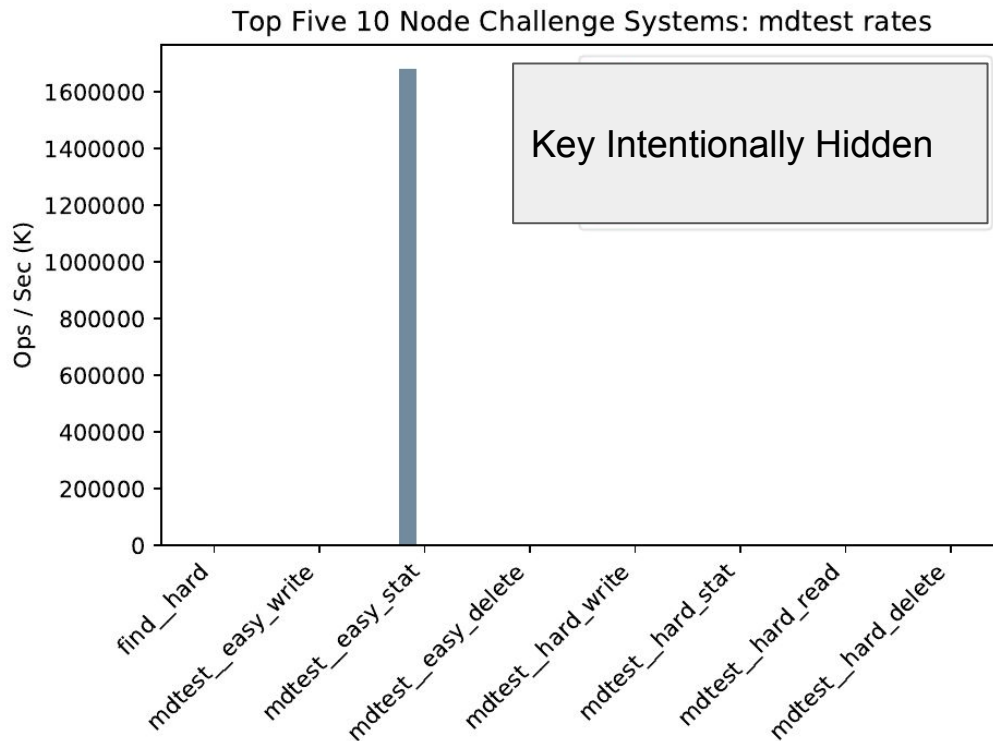
IO-500 10 nodes Metadata Winner



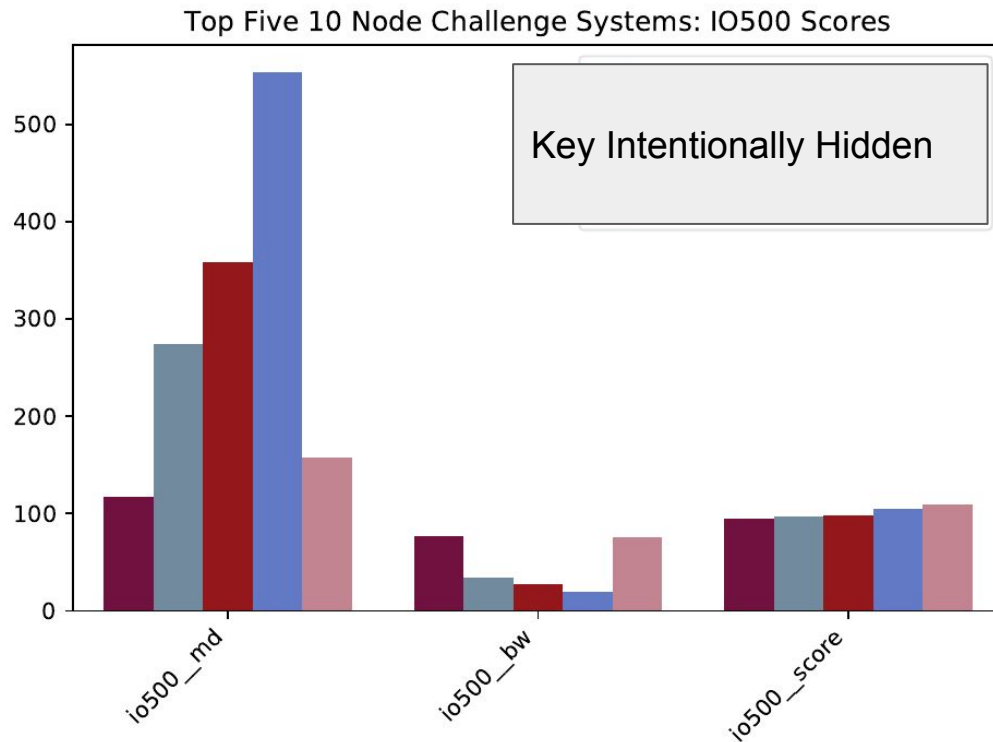
Bandwidth Scores for Top Five in Ten Node Challenge



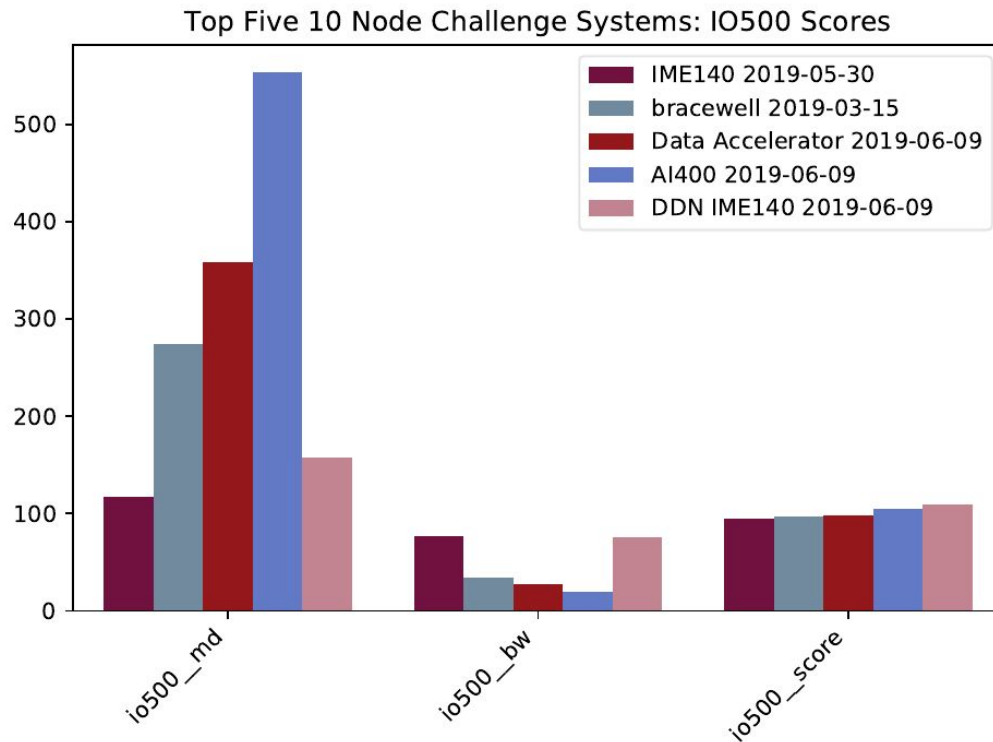
Metadata Scores for Top Five in Ten Node Challenge



Overall Scores for Top Five in Ten Node Challenge



Overall Scores for Top Five in Ten Node Challenge



The Big Reveal!

Congrats to
DDN Colorado!

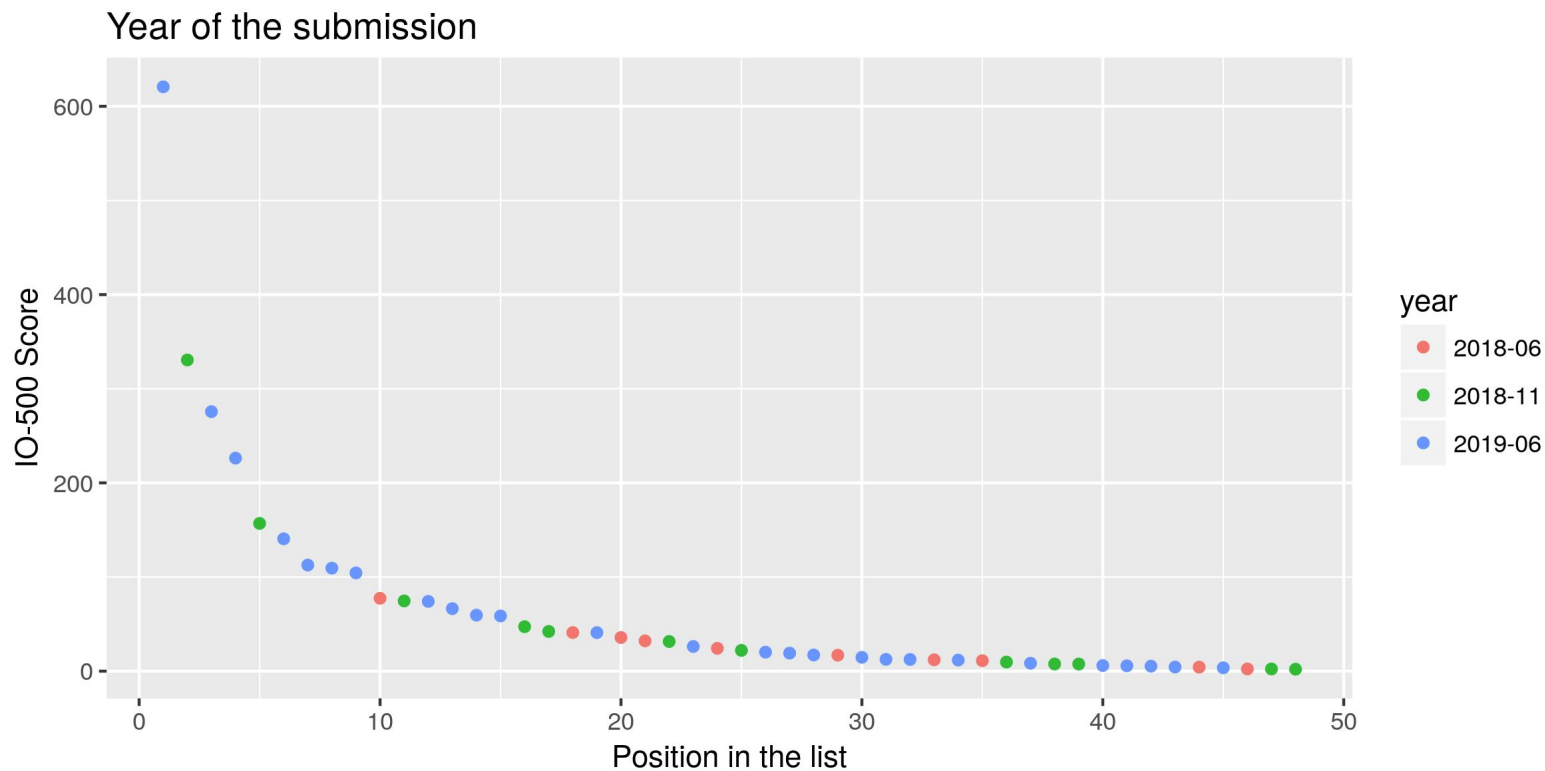
IO-500 10 nodes Winner



10 Node Challenge Ranked List (39 total entries)

#	information							io500		
	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	score	bw	md
									GiB/s	kIOP/s
1	DDN Colorado	DDN IME140	DDN	IME	10	160	zip	109.42	75.79	157.96
2	DDN	AI400	DDN	Lustre	10	160	zip	104.34	19.65	553.98
3	University of Cambridge	Data Accelerator	Dell EMC	Lustre	10	320	zip	98.31	26.94	358.85
4	CSIRO	bracewell	Dell/ThinkParQ	BeeGFS	10	160	zip	97.16	34.43	274.18
5	DDN	IME140	DDN	IME	10	160	zip	95.10	76.89	117.62
6	DDN Japan	AI400	DDN	Lustre	10	160	zip	74.10	12.22	449.28
7	HHMI Janelia Research Campus	Weka	WekaIO		10	3200	zip	66.43	27.74	159.12
8	DDN	400NV	DDN	GRIDScaler	10	30	zip	59.49	13.55	261.21
9	Genomics England	GELous	WekaIO	Wekafs Matrix parallel filesystem	10	1200	zip	58.66	14.83	232.05
10	WekaIO		WekaIO	WekaIO Matrix	10	700	zip	58.25	27.05	125.43

IO-500 Ranked List



IO-500 Bandwidth Winner



Certificate

IO-500 Performance Certification

This Certificate is awarded to:

JCAHPC

to be ranked #1 in the IO-500 BW Score

IO 500



June 2019

IO-500 Steering Board

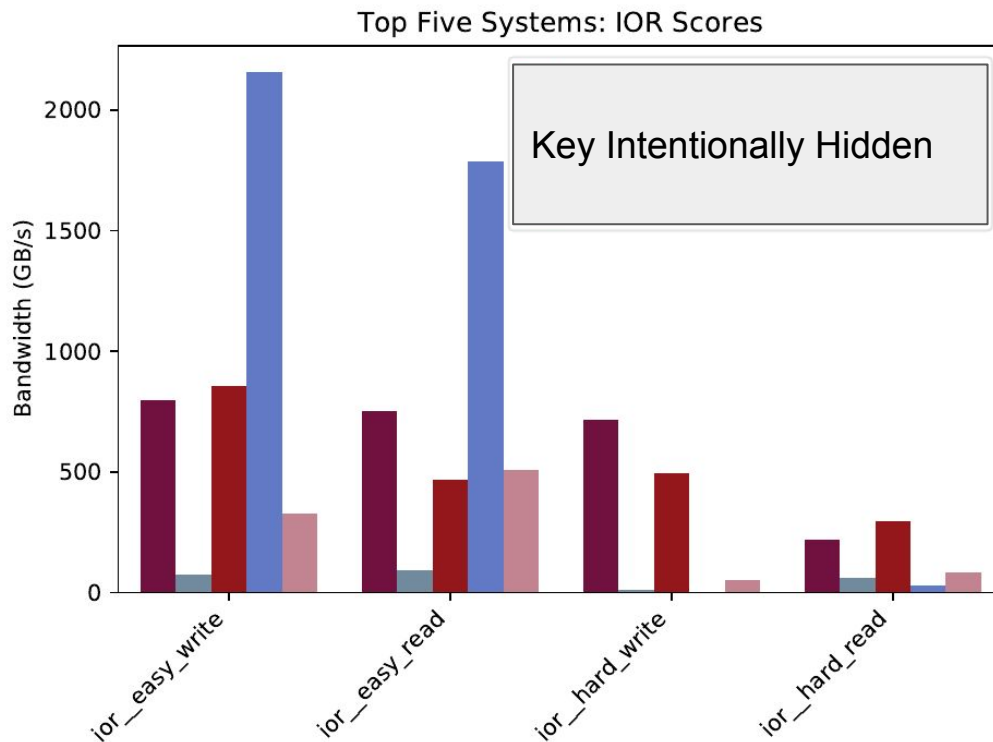
<http://io500.org/list/19-06/>



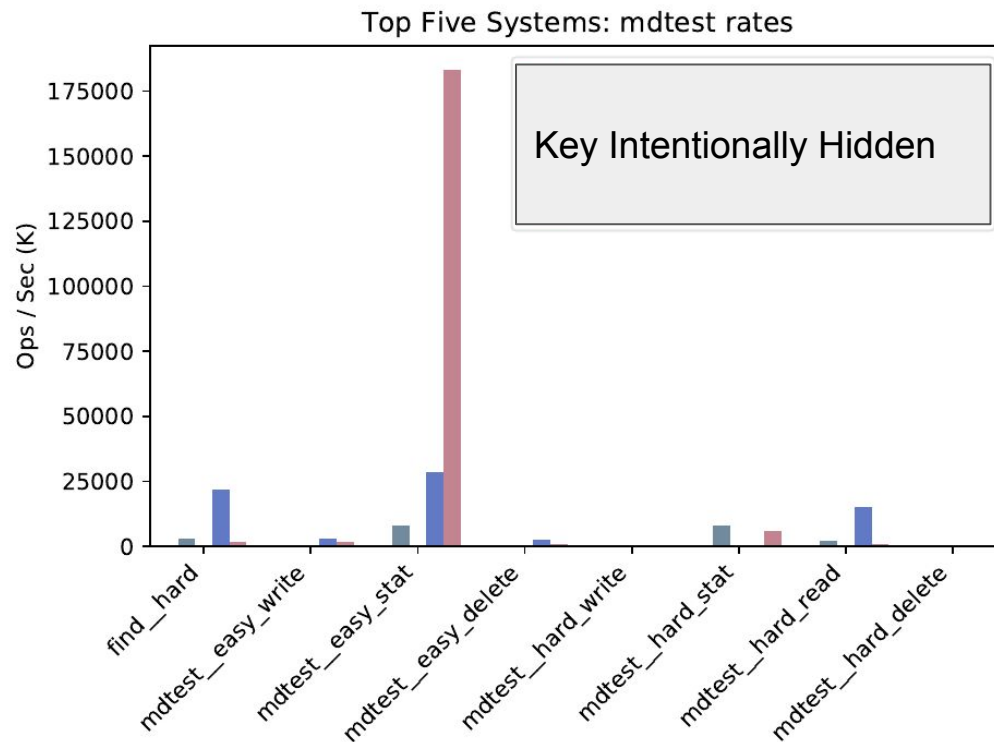
IO-500 Metadata Winner



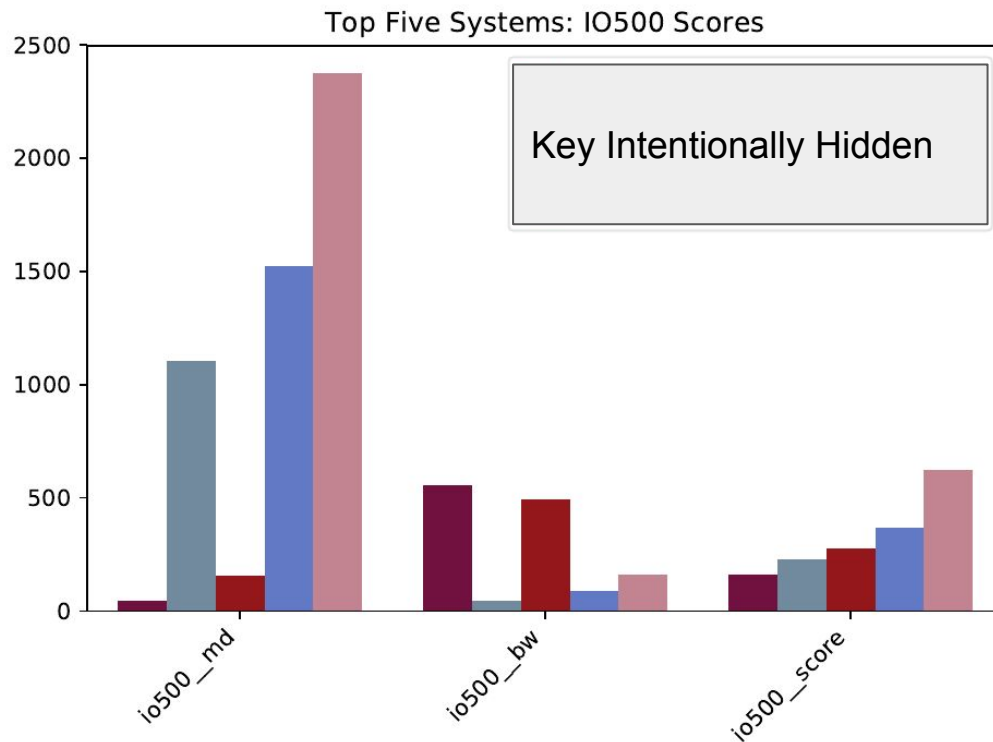
Bandwidth Scores for Top Five Overall



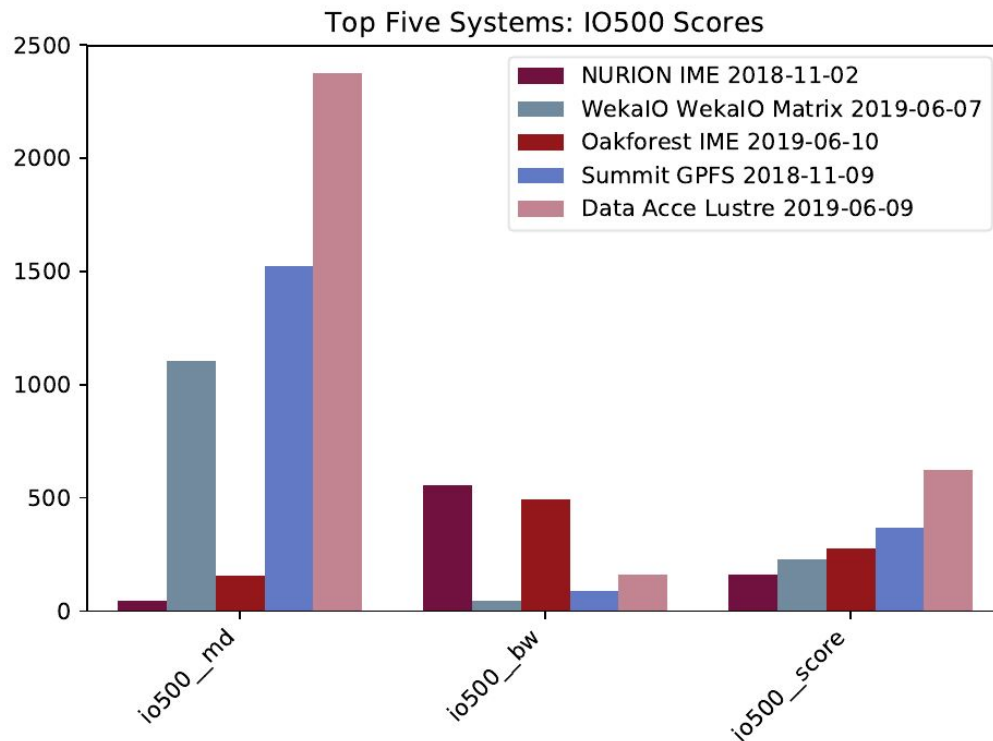
Metadata Scores for Top Five in Ranked List



Overall Scores for Top Five in Ranked List



Overall Scores for Top Five in Ranked List



The Big Reveal!

Congrats to
U Cambridge!

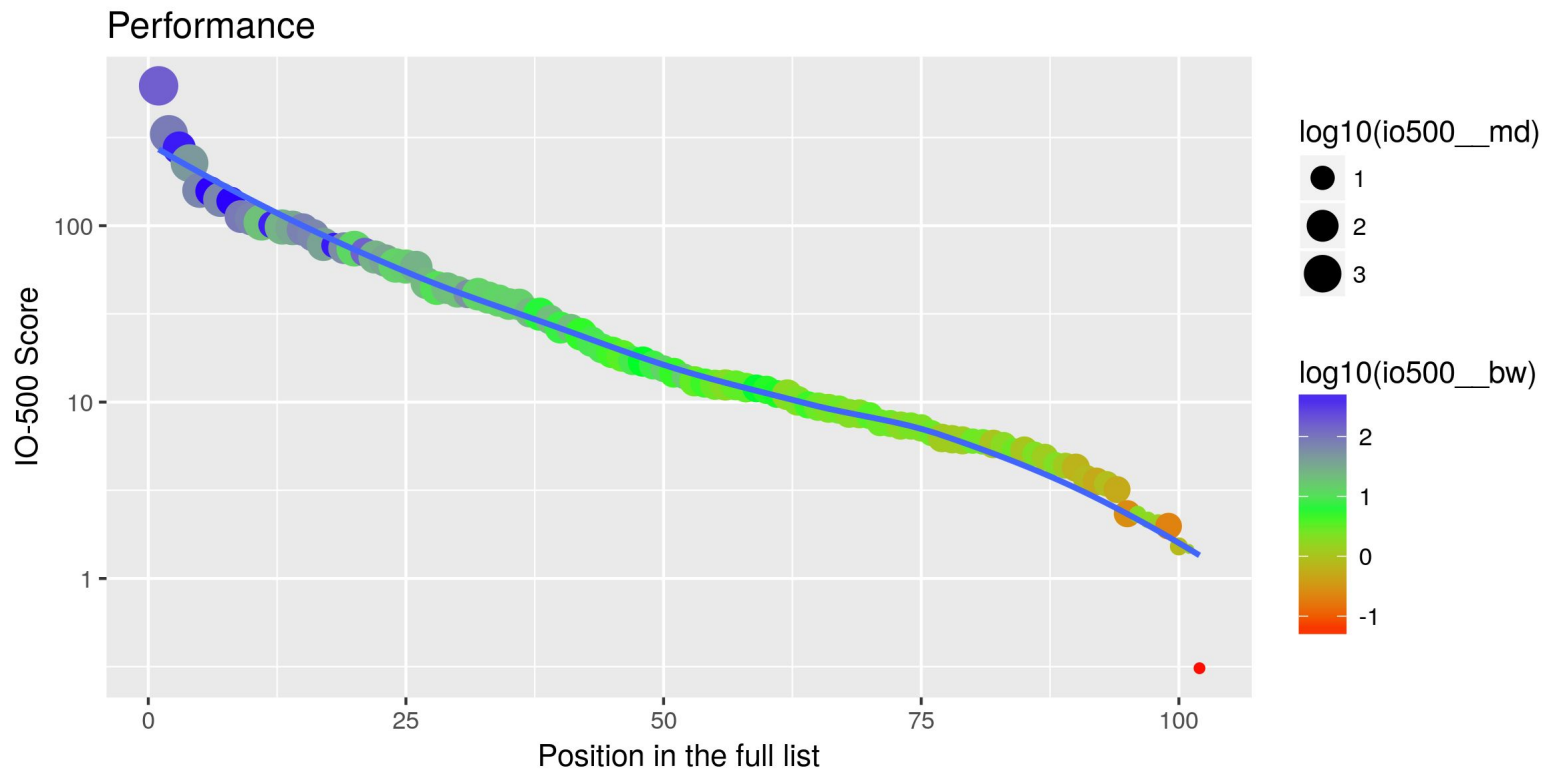
IO-500 Winner



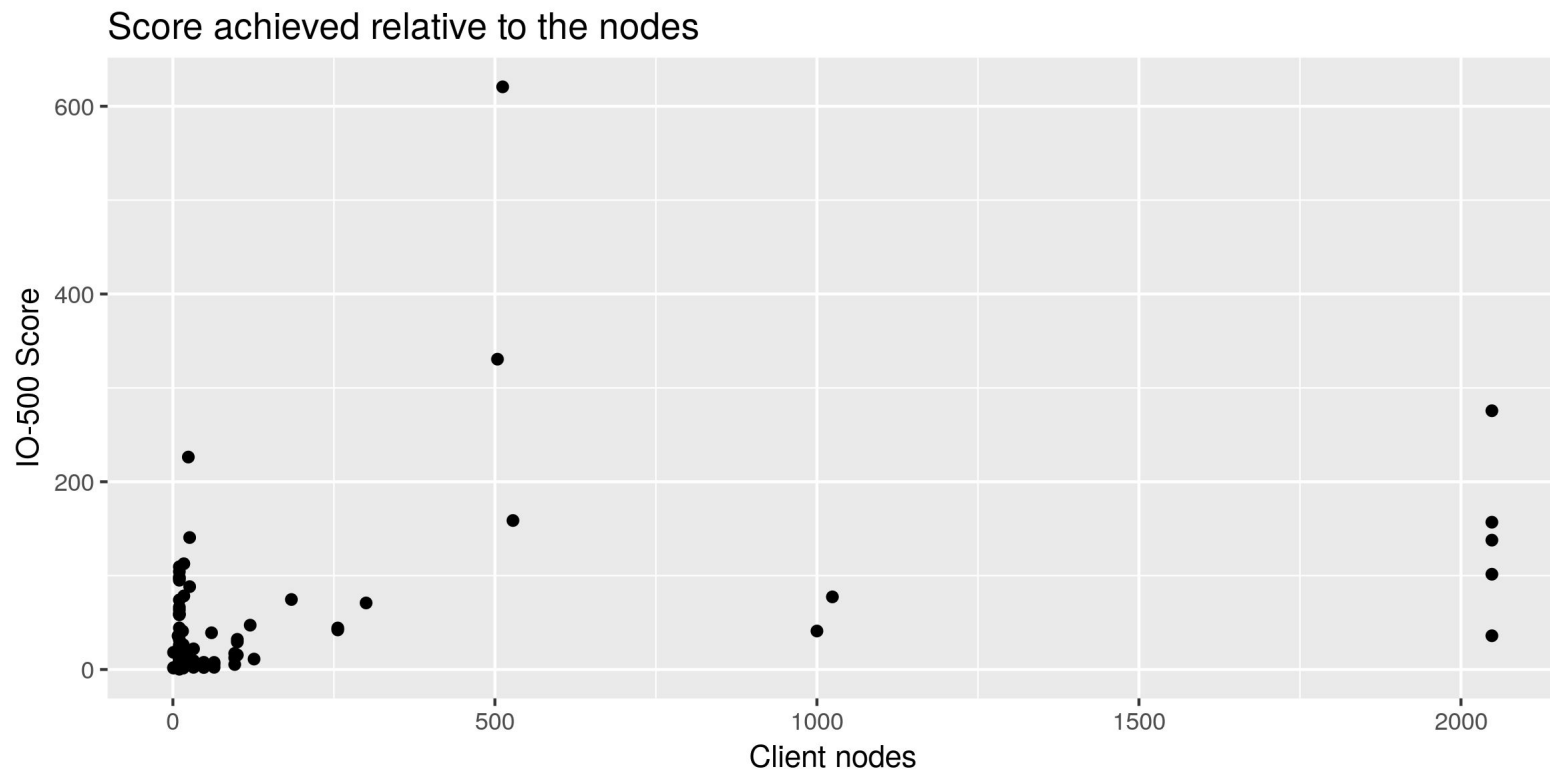
IO-500 Ranked List (48 total entries)

#	information							io500		
	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	score	bw	md
									GiB/s	kIOP/s
1	University of Cambridge	Data Accelerator	Dell EMC	Lustre	512	8192	zip	620.69	162.05	2377.44
2	Oak Ridge National Laboratory	Summit	IBM	Spectrum Scale	504	1008	zip	330.56	88.20	1238.93
3	JCAHPC	Oakforest-PACS	DDN	IME	2048	2048	zip	275.65	492.06	154.41
4	Weka	WekaIO	WekaIO	WekaIO Matrix	24	1704	zip	226.31	46.28	1106.51
5	Korea Institute of Science and Technology Information (KISTI)	NURION	DDN	IME	2048	4096	zip	156.91	554.23	44.43
6	CSIRO	bracewell	Dell/ThinkParQ	BeeGFS	26	260	zip	140.58	69.29	285.21
7	DDN	IME140	DDN	IME	17	272	zip	112.67	90.34	140.52
8	DDN Colorado	DDN IME140	DDN	IME	10	160	zip	109.42	75.79	157.96
9	DDN	AI400	DDN	Lustre	10	160	zip	104.34	19.65	553.98
10	KAUST	ShaheenII	Cray	DataWarp	1024	8192	zip	77.37	496.81	12.05

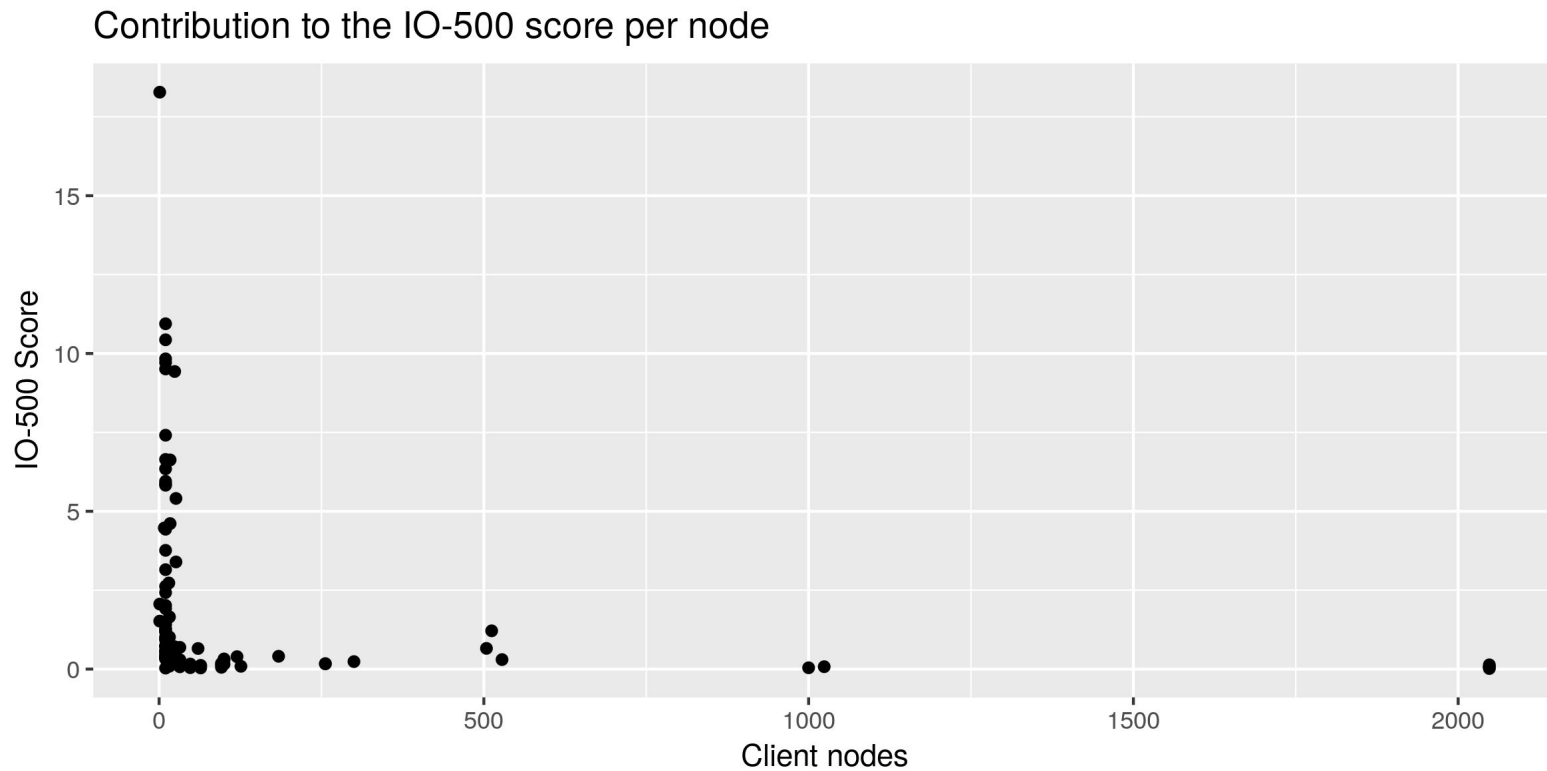
Analysis of the Full List



Analysis of the Full List



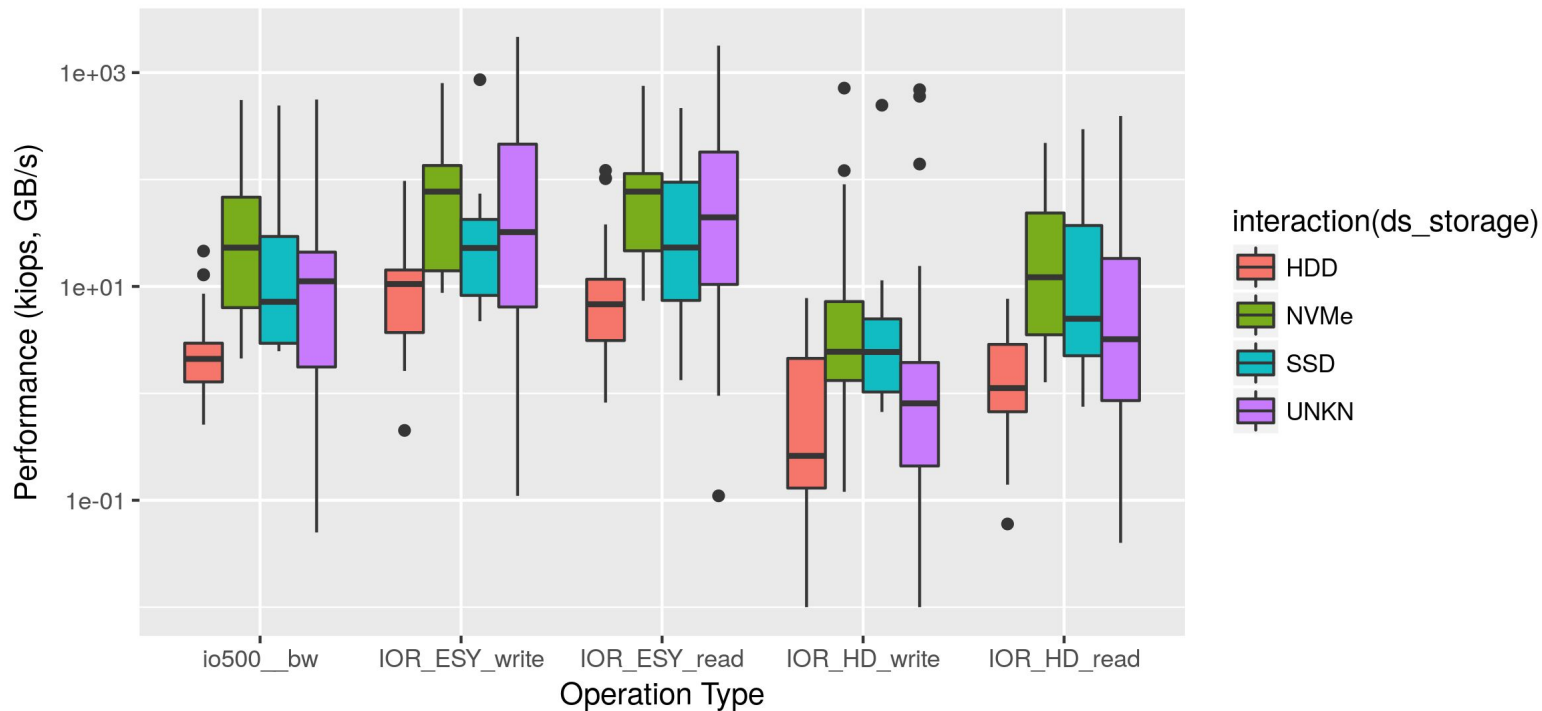
Analysis of the Full List



10⁵⁰⁰

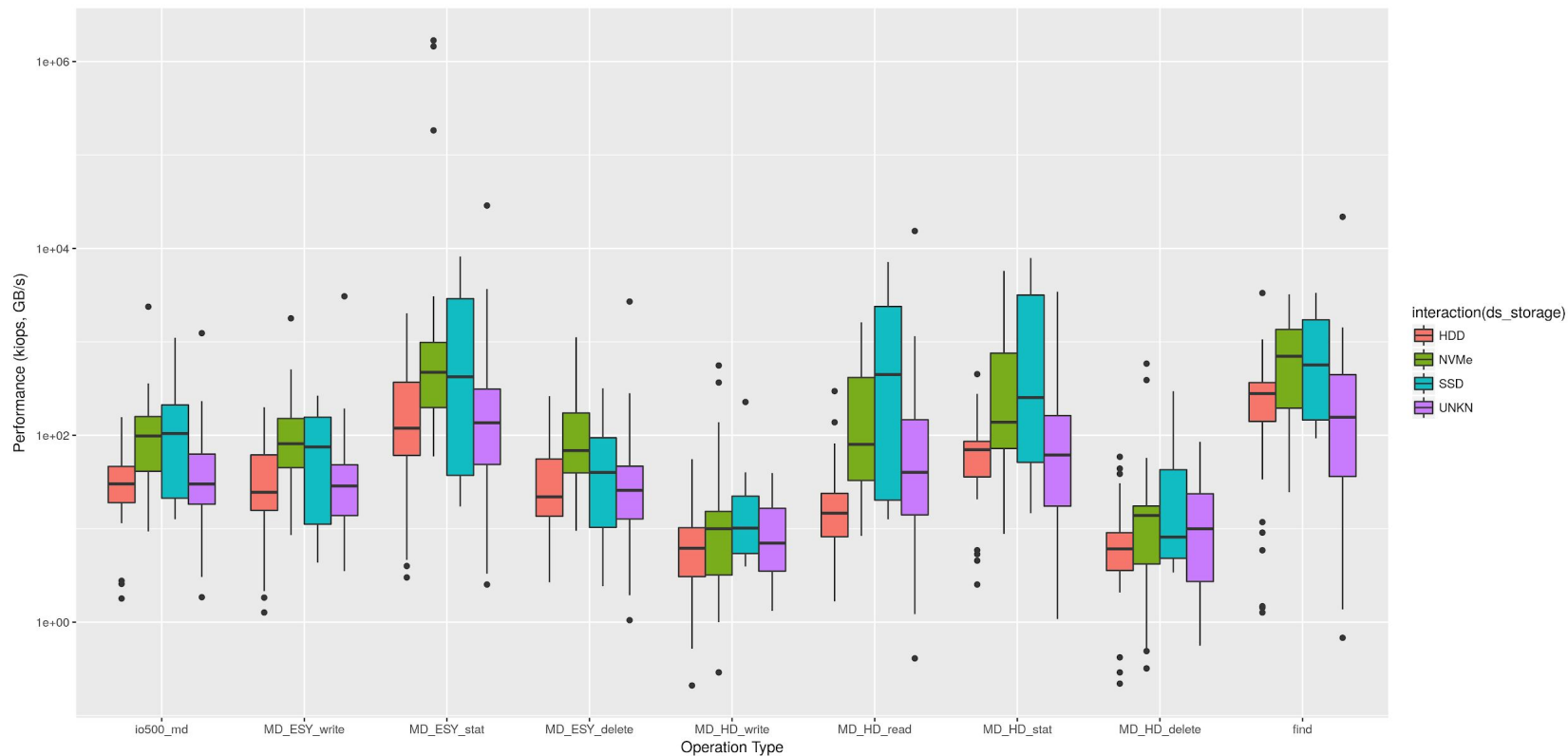
IO Throughput by (DS) Storage Media Type

File system performance per operation

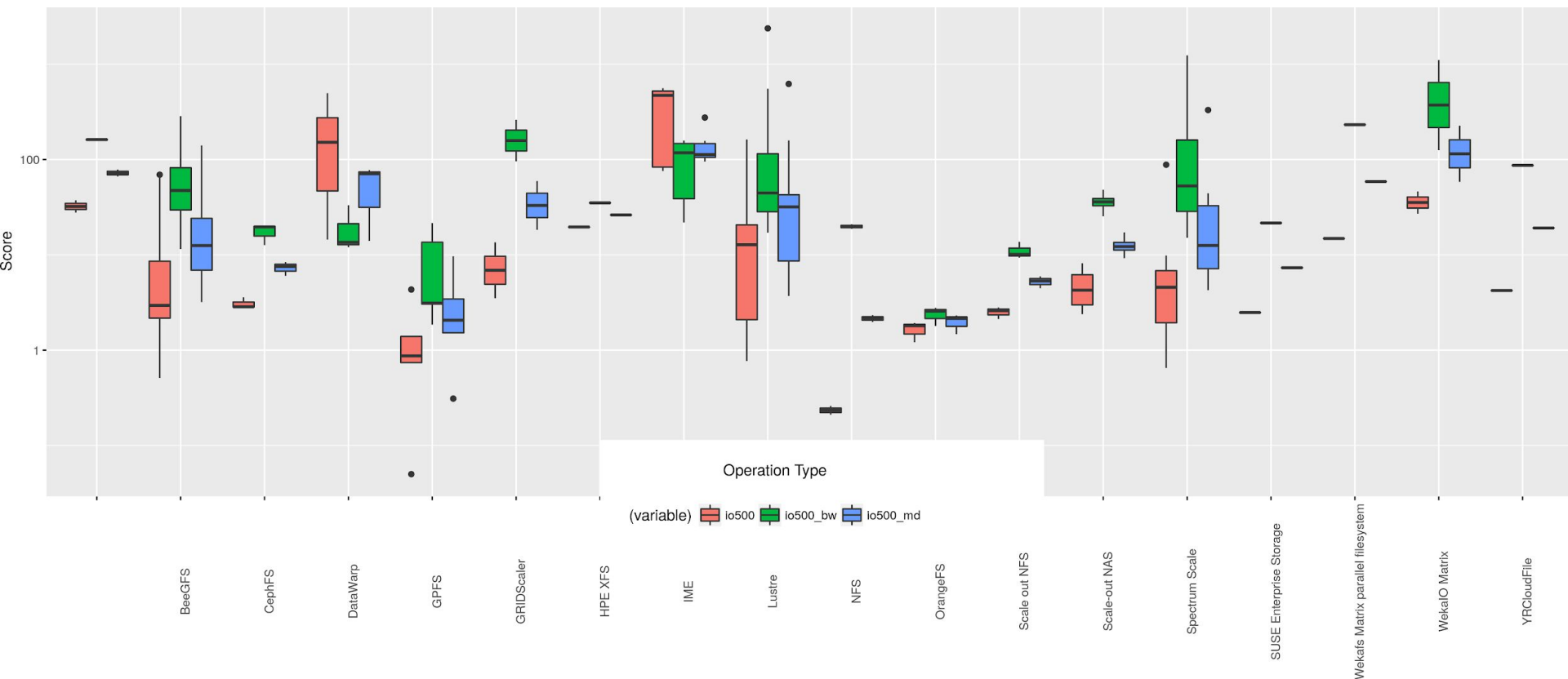


Metadata Performance per DS Storage

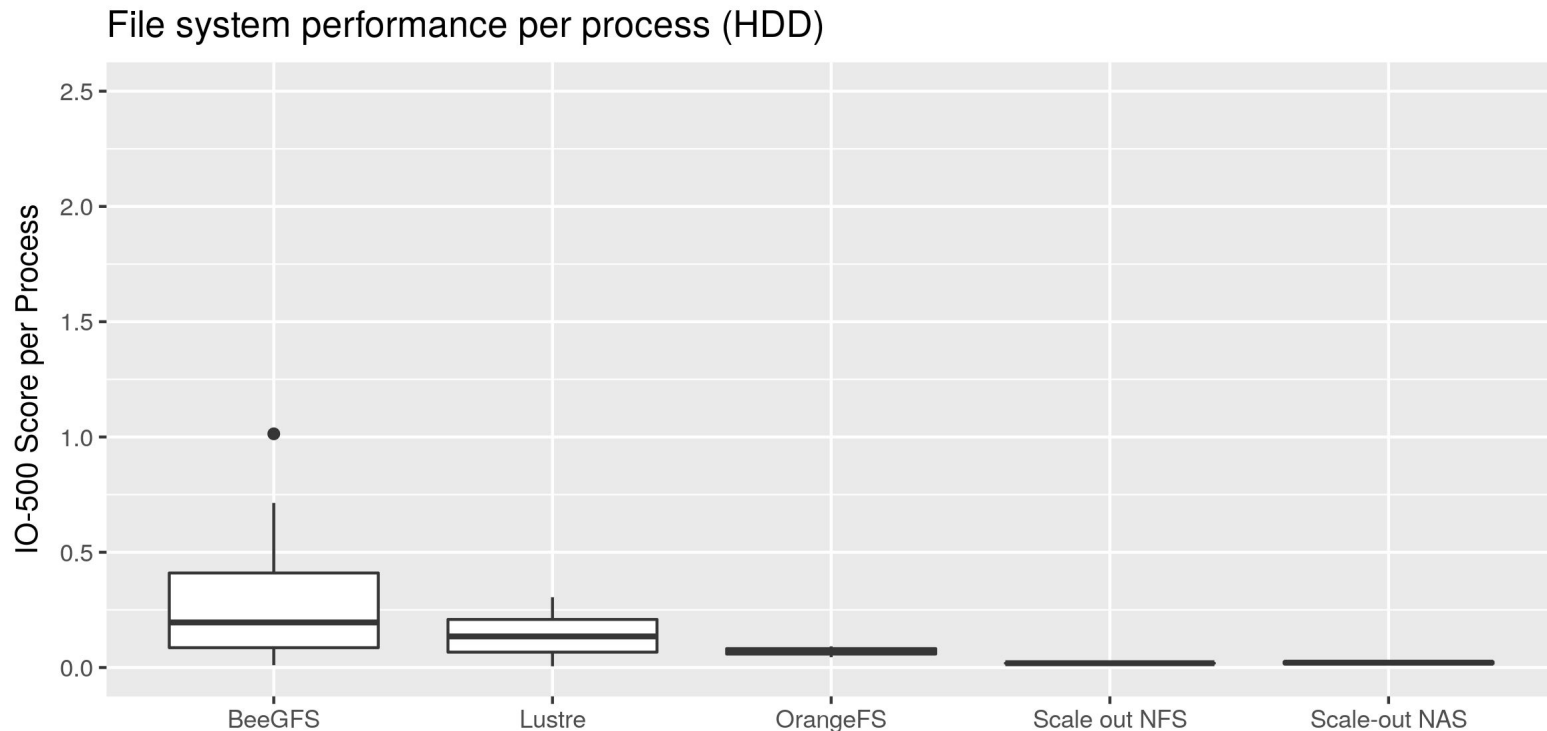
File system performance per operation



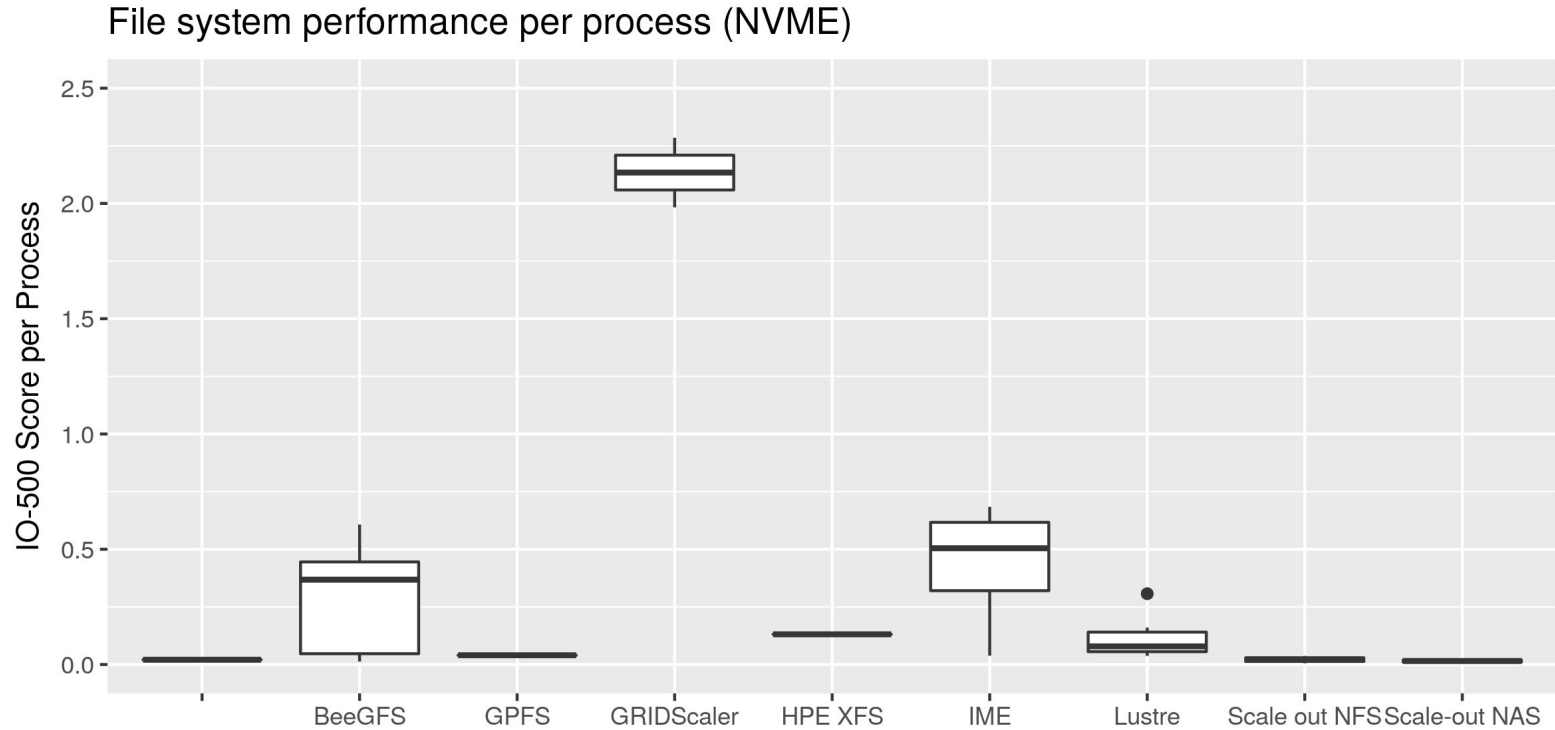
Score By File System



Which FS Yields Best for Storage / Normalized

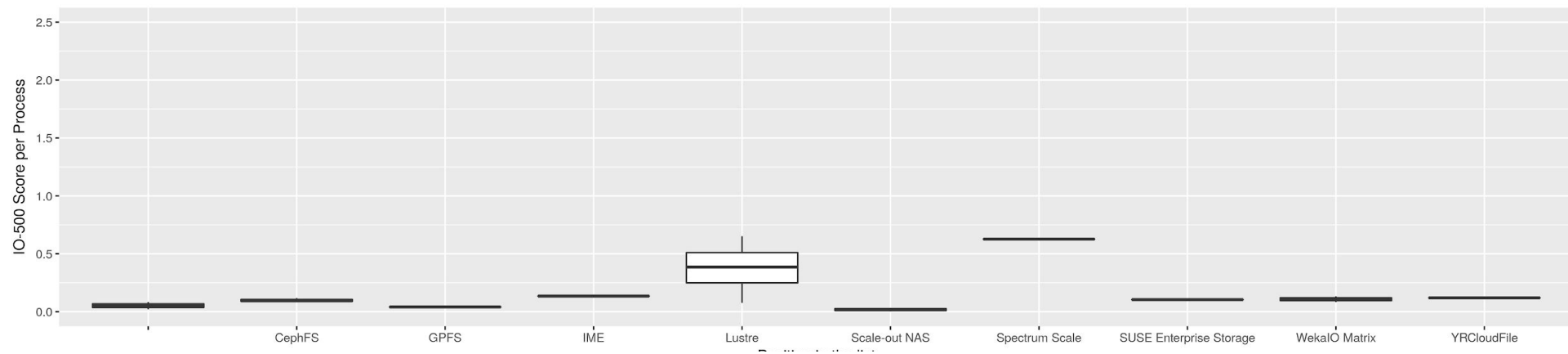


Which FS Yields Best for Storage / Normalized



Which FS Yields Best for Storage / Normalized

File system performance per process (SSD)



Analysis: What if? The “Uber” Storage

Combining the best performance for each benchmark leads to an Uber storage

For the full list: score: 3917

bw: 1020 => current best 560

md: 7674 => current best 2377

For a 10 node system:

score: 680: This would be **faster than #1** in the current list!

bw: 112 => current best 77

md: 1671 => current best 554

Roadmap


10 500

Roadmap for the IO-500 (until Supercomputing)

- Update to the benchmark
 - MDTest: Rank shifting between tests (similar to IOR): a rewrite of MDTest
 - Intent was to eliminate caching effects and it was determined this wasn't working
- Contribute a regression testing platform to IOR and MDTest for validation
- Keeping non-qualified submissions (partial/invalid results) separately
- Filter to generate new sublists, e.g., API=POSIX, storage=NVMe
- Webpage
 - Hardening of submission page; hardware description standardization (e.g., IB = Infiniband)
 - Detailed results for each submission
 - Creating of new sublists (how exactly is part of the discussion)

Evidence for the need for mdtest shifting

Radar Chart

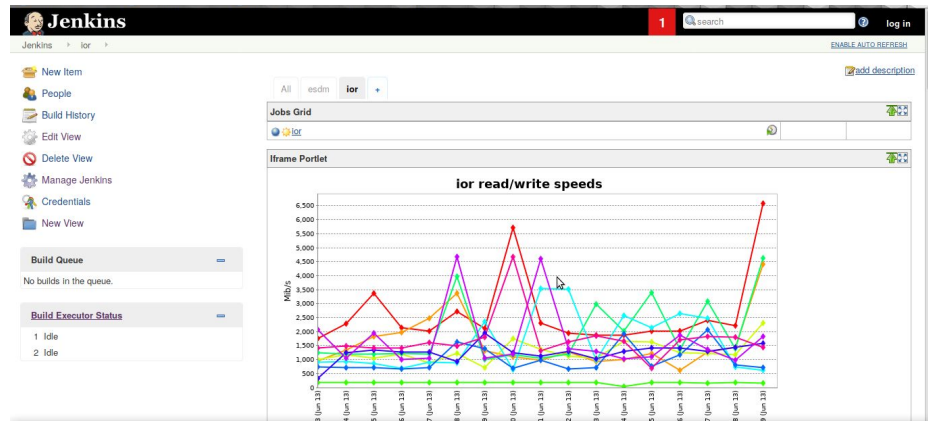
This is the official list from  ISC 2019 with a radar chart and controls to manipulate the ranking. The list shows all results.

IO500

#	information							io500			mdtest
	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	score	bw	md	easy_stat
											kIOP/s
1	CSIRO	bracewell	Dell/ThinkParQ	BeeGFS	10	160	zip	97.16	34.43	274.18	1682210.00
2	CSIRO	bracewell	Dell/ThinkParQ	BeeGFS	26	260	zip	140.58	69.29	285.21	1453520.00
3	University of Cambridge	Data Accelerator	Dell EMC	Lustre	512	8192	zip	620.69	162.05	2377.44	183233.00
4	Oak Ridge National Laboratory	Summit	IBM	Spectrum Scale	504	1008	zip	330.56	88.20	1238.93	28769.40
5	Weka	WekaIO	WekaIO	WekaIO Matrix	24	1704	zip	226.31	46.28	1106.51	8199.12
6	WekaIO	WekaIO	WekaIO		17	935	zip	78.37	37.39	164.26	3738.01
7	Genomics England	GELous	WekaIO	Wekafs Matrix parallel filesystem	10	1200	zip	58.66	14.83	232.05	3681.22

Regression Testing

- For IO500/IOR/MDTest
- Validation: correct behavior
- Performance behavior:
 - Track performance changes over time
- Technical:
 - Using Jenkins (various plugins)
 - Schedule to run tests to determine
 - Some tests will run for minutes!
 - Prototype nearly ready (see figures)
 - Done by Benjamin Hodges
- Anyone willing to run the scripts as well?



Discussion

10 500

Discussion Points

1. Vendor community created.
2. Change process discussion.
3. What to do with old entries as benchmark suite evolves?
4. List reveal process
5. What part of the process is working/not working well for you?
 - a. Possible topics
 - i. Is find working for you?
 - ii. Hard phases too easy?
 - iii. Are we missing anything?
6. Shall the benchmark validate correctness of its results?
 - a. e.g., Is it important to make sure you read what was written?
 - b. reported numbers, client count, etc.
7. How to automatically collect more and better environmentals?