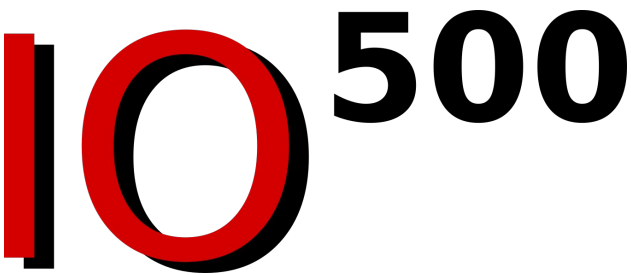
IO500:
The High-Performance Storage Community

Committee

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



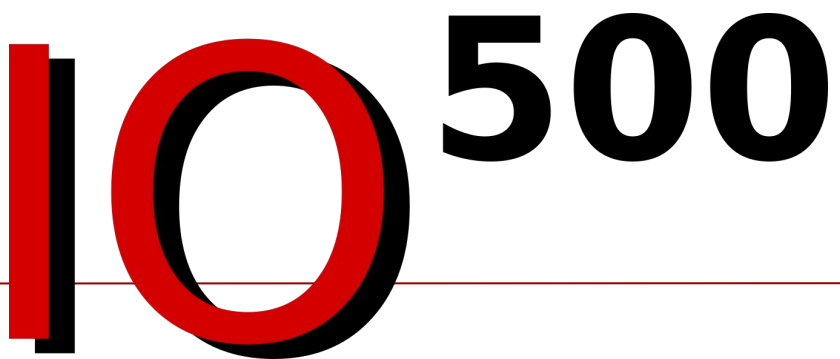
BoF Agenda

1. **Welcome** – Andreas Dilger
2. **The New IO500 List Analysis** – Dean Hildebrand
3. **Award Presentations** – Andreas Dilger
4. **Website Update** - Jean-Luca Bez
5. **Community Presentations**
    - Empowering Lustre Performance Through IO500 - Shuichi Ihara
    - IO500 with GPUDirect and Extended Mode - Hendrik Nolte
6. **Extended Access Patterns** – Andreas Dilger
7. **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.
  - https://io500.org/
  - Issues/submissions 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: rb.gy/sn8esm

IO500 List Analysis



Overall Thoughts

- Production lists submissions had a great boost
- We really appreciate the detailed information in the schema and questionnaire
  - We know it is a lot of work…
  - The questionnaire is really becoming the best source for a quick overview of the submissions
- Really great to see submissions from both top 10 HPC systems as well as ‘regular’ HPC systems
  - Starting to realize IO500’s mission in building a wealth of information on HPC storage systems
- 229 entries across the 4 lists (almost half-way to 500)

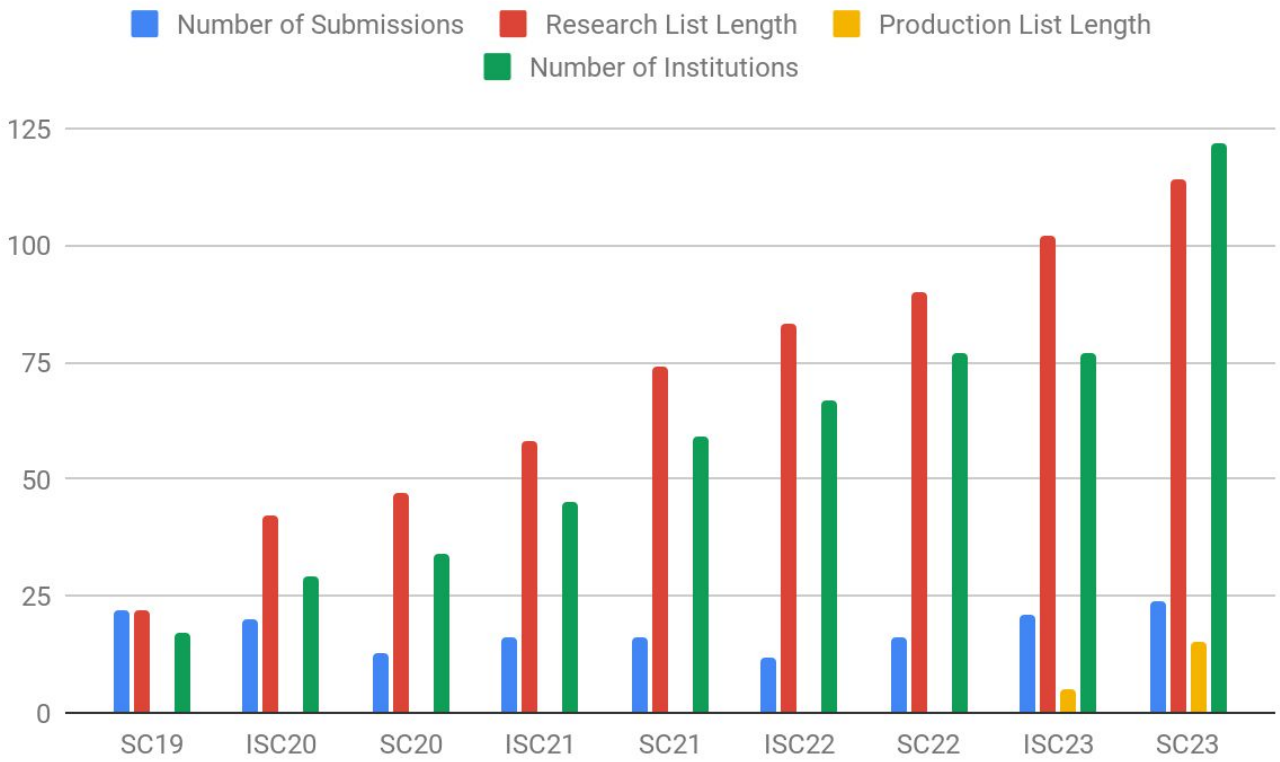
IO500 List - Growth in Entries and Institutions

25 submissions (with overlap)

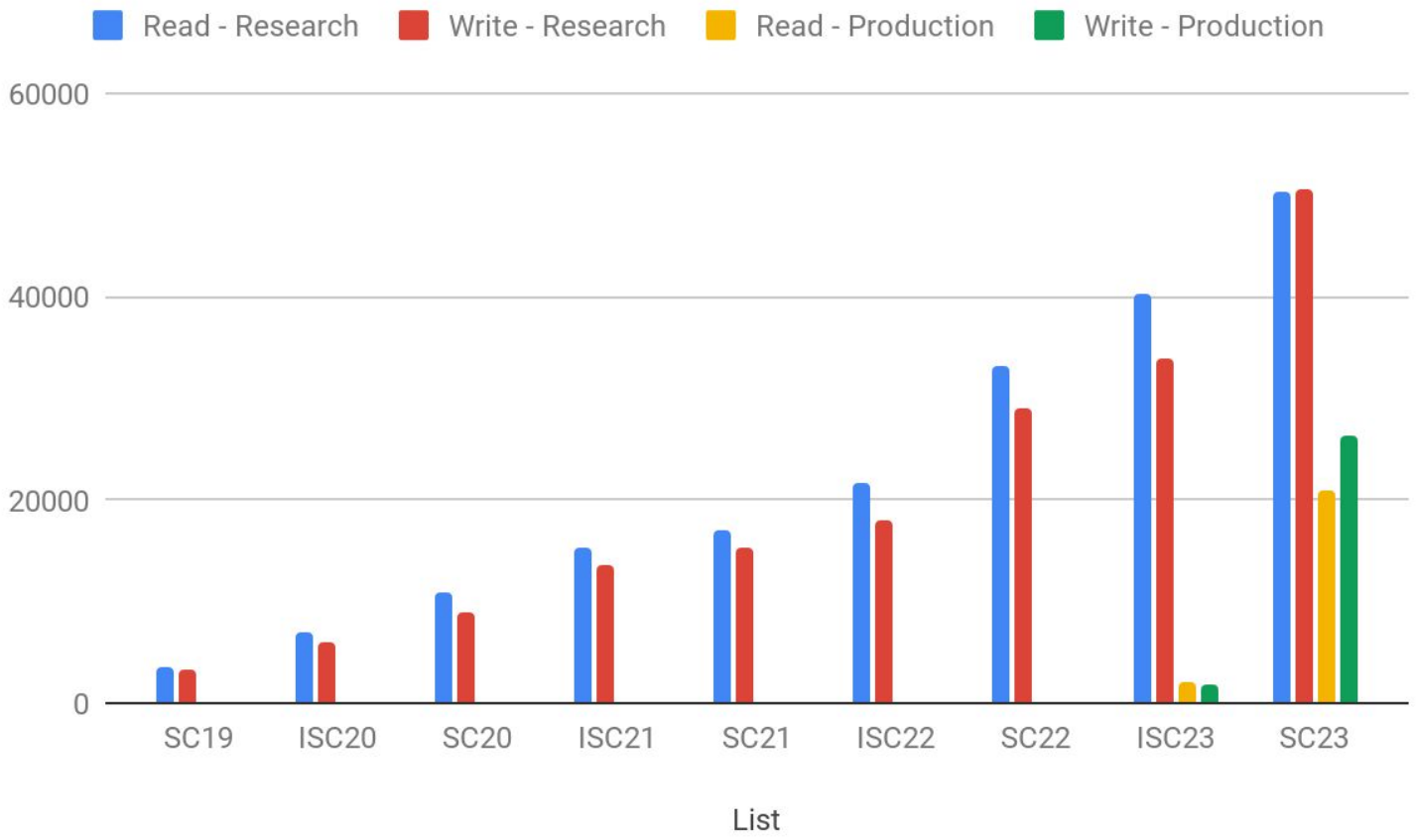
- 11 for 10-Client Research
- 6 for 10-Client Production
- 17 for IO500 Research
- 11 for IO500 Production
- 1 Reject due to lack of persistence

Production: 15
Production-10: 7
Research: 114
Research-10: 101

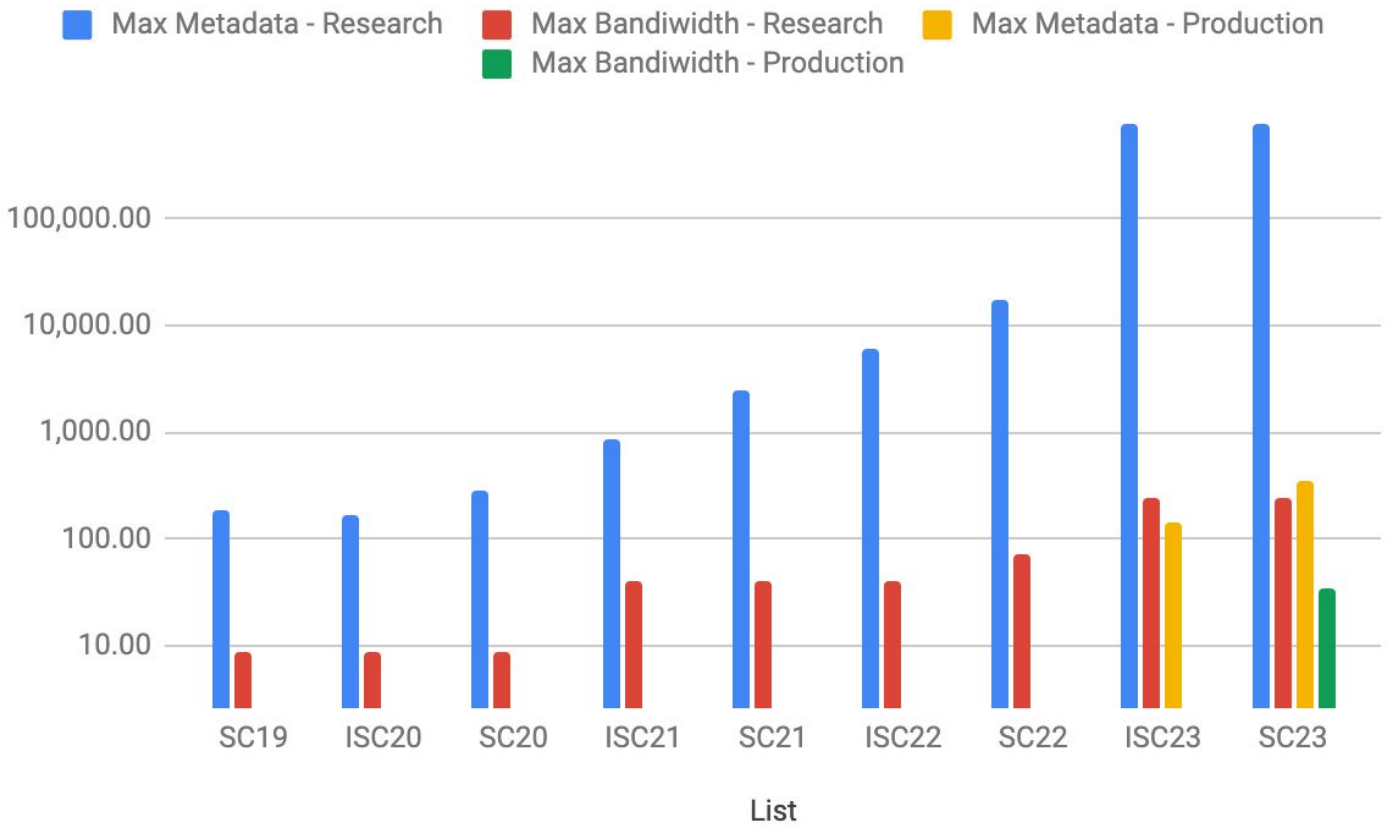
Institutions: 122



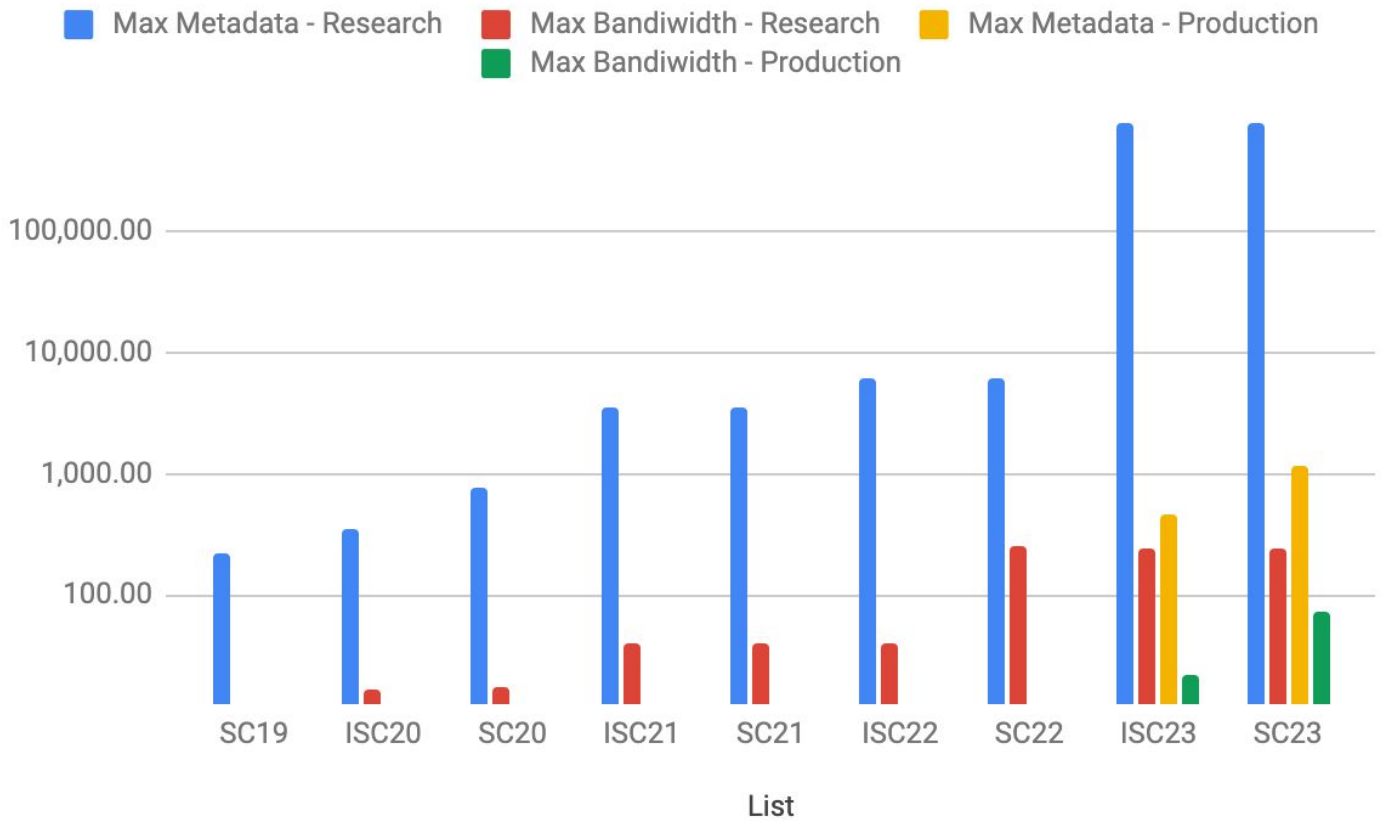
IO500 List - Aggregate List Bandwidth (GB/s)



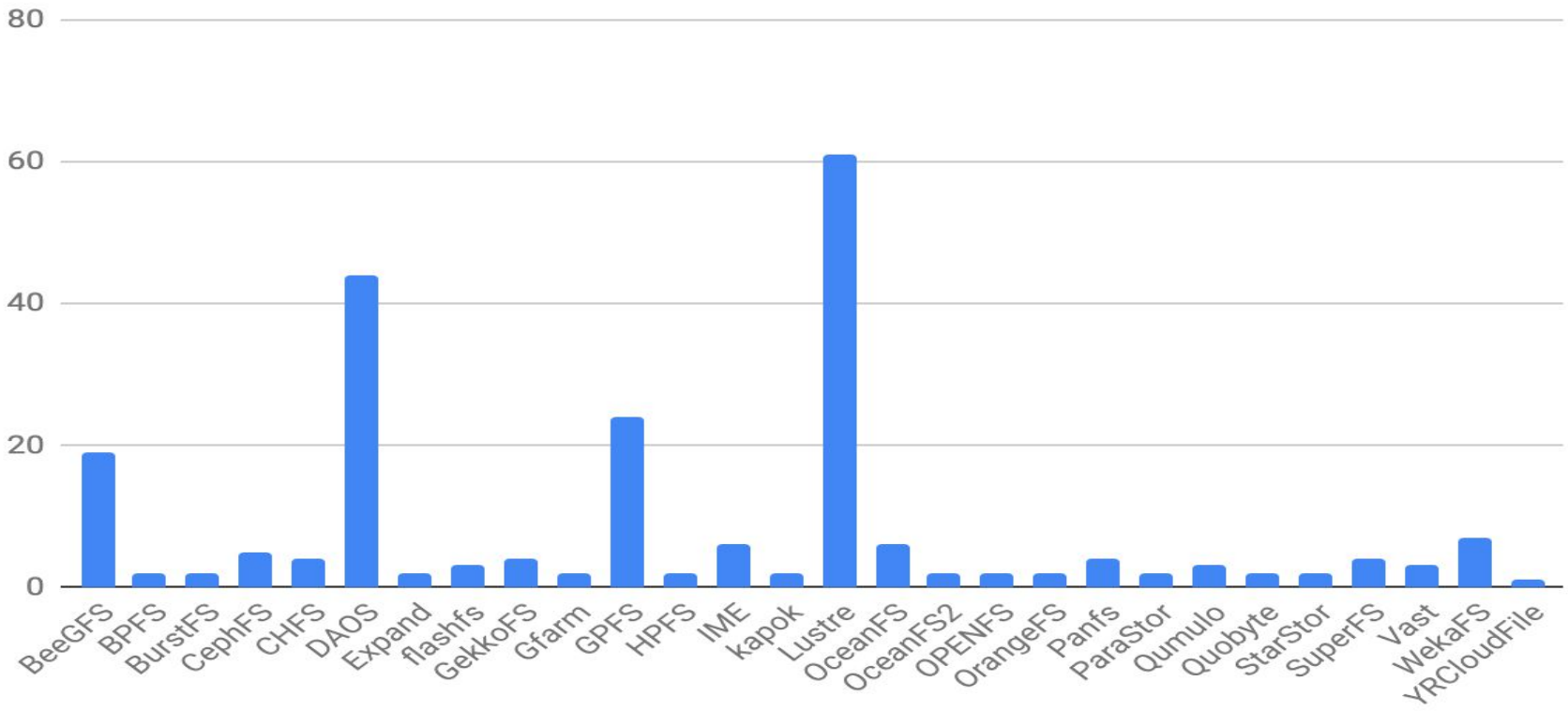
IO500 List - Growth in Max Score per Client



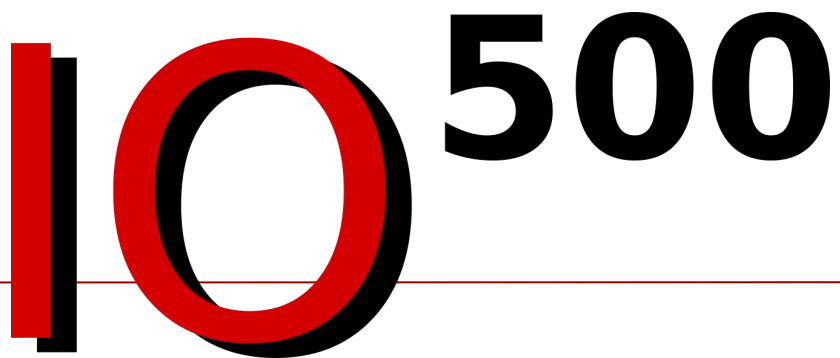
10-Client List - Growth in Max Scores per Client



IO500 List - Number of File System Entries



Award Ceremony



Five Awards

Aggregate awarding of Bandwidth, Metadata, Overall for same winner

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

10 Client Node Production - Bandwidth Winner

Sort by BW

#	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE	BW ↑ (GIB/S)	MD (KIOP/S)
1	SC23	Aurora	Argonne National Laboratory	DAOS		734.50	
2	ISC23	SuperMUC-NG-Phase2-EC-10	LRZ	DAOS		218.38	
3	SC23	Earth Simulator 4	Japan Agency for Marine-Earth Science and Technology	EXAScaler		48.19	
4	SC23	Randi	Center for Research Informatics at University of Chicago	Spectrum Scale		31.05	
5	SC23	Orion	Mississippi State University High Performance Computing Collaboratory	Lustre		6.43	
6	SC23	Orion	Mississippi State University High Performance Computing Collaboratory	Lustre		5.01	
7	SC23	spt-compute1	Eikon Therapeutics	Qumulo Core		2.24	

Indicates new entry on this list

Certificate

IO500 Performance Certification

This Certificate is awarded to:

Argonne National Laboratory
(Aurora DAOS EC)

#1 in the 10 Client Node Production Bandwidth Score

IO500

November 2023

IO500 Steering Board

https://io500.org/list/SC23/ten-production

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	SC23	Earth Simulator 4	Japan Agency for Marine-Earth Science and Technology	EXAScaler	101.88	48.19	215.38
4	SC23	Randi	Center for Research Informatics at University of Chicago	Spectrum Scale	60.88	31.05	119.36
5	SC23	Orion	Mississippi State University High Performance Computing Collaboratory	Lustre	20.83	5.01	86.67
6	SC23	Orion	Mississippi State University High Performance Computing Collaboratory	Lustre	17.57	6.43	48.03
7	SC23	spt-compute1	Eikon Therapeutics	Qumulo Core	5.35	2.24	12.77

Certificate

IO500 Performance Certification

This Certificate is awarded to:

Argonne National Laboratory
(Aurora DAOS EC)

#1 in the 10 Client Node Production Overall Score

IO500

November 2023

IO500 Steering Board

https://io500.org/list/SC23/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	SC23	Aurora	Argonne National Laboratory	DAOS		934.00	
3	SC22	ParaStor	Sugon Cloud Storage Laboratory	ParaStor		718.11	
4	SC22	StarStor	SuPro Storteck	StarStor		515.15	
5	ISC21	Endeavour	Intel	DAOS		398.77	
6	SC21	OceanStor Pacific	Olympus Lab	OceanFS		317.07	
7	SC21	Athena	Huawei HPDA Lab	OceanFS		314.56	
8	ISC23	SuperMUC-NG-Phase2-10	LRZ	DAOS		266.73	
9	ISC23	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng Laboratory	SuperFS		263.97	
10	ISC22	Cumulus	University of Cambridge	DAOS		216.78	

Certificate

IO500 Performance Certification

This Certificate is awarded to:

JNIST and HUST PDSL (Cheeloo-1)
with Huawei OceanStor Pacific

#1 in the 10 Client Node Research Bandwidth Score

IO500

November 2023

IO500 Steering Board

https://io500.org/list/SC23/ten

10 Client Node Research - Overall Winner

# ↑	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE ↑	BW (GIB/S)	MD (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	SC23	Aurora	Argonne National Laboratory	DAOS	3,748.85	934.00	15,046.98
7	ISC22	Shanhe	National Supercomputing Center in Jinan	flashfs	3,534.42	207.79	60,119.50
8	SC21	Athena	Huawei HPDA Lab	OceanFS	2,395.03	314.56	18,235.71
9	SC21	OceanStor Pacific	Olympus Lab	OceanFS	2,298.69	317.07	16,664.88
10	ISC21	Endeavour	Intel	DAOS	1,859.56	398.77	8,671.65

Certificate

IO500 Performance Certification

This Certificate is awarded to:

JNIST and HUST PDSL (Cheeloo-1)
with Huawei OceanStor Pacific

#1 in the 10 Client Node Research Overall Score

IO500

November 2023

IO500 Steering Board

https://io500.org/list/SC23/ten

IO500 Production List - Bandwidth Winner

Sorted by BW

#	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE	BW ↑ (GIB/S)	MD (KIOP/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	
5	SC23	IRIS	Memorial Sloan Kettering Cancer Center	WekaIO		104.79	
6	SC23	Earth Simulator 4	Japan Agency for Marine-Earth Science and Technology	EXAScaler		48.19	
7	ISC23	Imperial - hx cluster	Imperial College London	Spectrum scale		44.63	
8	SC23	Randi	Center for Research Informatics at University of Chicago	Spectrum Scale		31.05	
9	ISC22	CTPAI	China Telecom Research Institute	DAOS		25.29	
10	SC23	Janelia Compute Cluster	Howard Hughes Medical Institute Janelia Research Campus	Vast		11.45	

Certificate

IO500 Performance Certification

This Certificate is awarded to:

Argonne National Laboratory
(Aurora DAOS EC)

#1 in the IO500 Production Bandwidth Score

IO500

November 2023

IO500 Steering Board

https://io500.org/list/SC23/production

IO500 Production List - Overall Winner

# ↑	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
4	ISC23	Leonardo	EuroHPC-CINECA	EXAScaler	648.96	807.12	521.79
5	SC23	IRIS	Memorial Sloan Kettering Cancer Center	WekaIO	308.94	104.79	910.80
6	ISC22	CTPAI	China Telecom Research Institute	DAOS	187.84	25.29	1,395.01
7	ISC23	Imperial - hx cluster	Imperial College London	Spectrum scale	119.56	44.63	320.31
8	SC23	Earth Simulator 4	Japan Agency for Marine-Earth Science and Technology	EXAScaler	101.88	48.19	215.38
9	SC23	Randi	Center for Research Informatics at University of Chicago	Spectrum Scale	60.88	31.05	119.36
10	SC23	Altair	Poznan Supercomputing and Networking Center	Lustre	53.70	8.84	326.39

Certificate

IO500 Performance Certification

This Certificate is awarded to:

Argonne National Laboratory
(Aurora DAOS EC)
#1 in the IO500 Production Overall Score

IO500

November 2023

IO500 Steering Board

https://io500.org/list/SC23/production

IO500 Research List - Bandwidth Winner

Sorted by BW

#	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE	BW ↑ (GIB/S)	MD (KIOP/S)
1	SC23	Aurora	Argonne National Laboratory	DAOS		11,362.27	
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	SC23	SuperMUC-NG-Phase2	LRZ	DAOS		1,054.72	
5	ISC23	Leonardo	EuroHPC-CINECA	EXAScaler		807.12	
6	SC22	ParaStor	Sugon Cloud Storage Laboratory	ParaStor		718.11	
7	SC23	Shaheen III	King Abdullah University of Science and Technology	Lustre		709.52	
8	SC20	Oakforest-PACS	JCAHPC	IME		697.20	
9	ISC20	NURION	Korea Institute of Science and Technology Information (KISTI)	IME		515.59	
10	SC22	StarStor	SuPro Storteck	StarStor		515.15	

Certificate

IO500 Performance Certification

This Certificate is awarded to:

Argonne National Laboratory
(Aurora DAOS)

#1 in the IO500 Research Bandwidth Score

IO500

November 2023

IO500 Steering Board

https://io500.org/list/SC23/io500

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	SC23	Aurora	Argonne National Laboratory	DAOS	43,218.80	11,362.27	164,391.73
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	SC23	SuperMUC-NG-Phase2	LRZ	DAOS	4,585.68	1,054.72	19,937.45
8	ISC22	Shanhe	National Supercomputing Center in Jinan	flashfs	3,534.42	207.79	60,119.50
9	SC22	HPC-OCI	Cloudam HPC on OCI	BurstFS	3,033.03	278.48	33,033.54
10	SC21	Athena	Huawei HPDA Lab	OceanFS	2,395.03	314.56	18,235.71

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

November 2023

IO500 Steering Board

https://io500.org/list/SC23/io500

List of Awarded Systems in the Ranked Lists

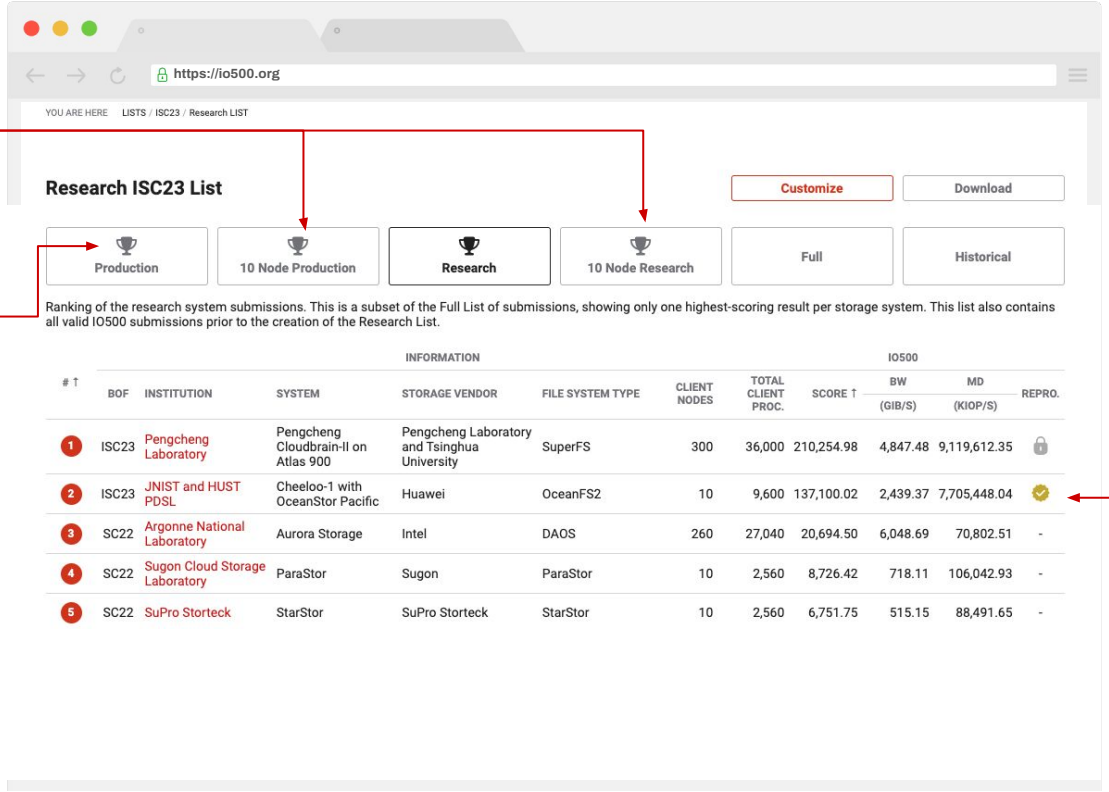
10 Client Production	**Bandwidth**	**Argonne National Laboratory**	DAOS EC	**734.50 GiB/s**
	Metadata	**Argonne National Laboratory**	DAOS EC	**11,336.27 kIOP/s**
	Overall	**Argonne National Laboratory**	DAOS EC	**2,885.57 score**
10 Client Research	Bandwidth	JNIST and HUST PDSL	OceanFS2	2439.37 GiB/s
	Metadata	JNIST and HUST PDSL	OceanFS2	7,705,448.04 kIOP/s
	Overall	JNIST and HUST PDSL	OceanFS2	137,100.00 score
IO500 Production	**Bandwidth**	**Argonne National Laboratory**	DAOS EC	**10,066.09 GiB/s**
	Metadata	**Argonne National Laboratory**	DAOS EC	**102,785.41 kIOP/s**
	Overall	**Argonne National Laboratory**	DAOS EC	**32,165.93 score**
IO500 Research	**Bandwidth**	**Argonne National Laboratory**	DAOS	**11,362.27 GiB/s**
	Metadata	Pengcheng Laboratory	SuperFS	9,119,612.35 kIOP/s
	Overall	Pengcheng Laboratory	SuperFS	210,255.00 score

IO500 Website Updates

IO500 Website - List View

NEW!
Lists are separated by **Research** and **Production**

NEW!
Lists with **Awards** are clearly marked



#		INFORMATION						IO500		
	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE	BW (GIB/S)	MD (KIOP/S)	REPRO.
1	ISC23	Pengcheng Laboratory	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng Laboratory and Tsinghua University	SuperFS	300	36,000	210,254.98	4,847.48	9,119,612.35	
2	ISC23	JNIST and HUST PDSL	Cheeloo-1 with OceanStor Pacific	Huawei	OceanFS2	10	9,600	137,100.02	2,439.37	7,705,448.04	
3	SC22	Argonne National Laboratory	Aurora Storage	Intel	DAOS	260	27,040	20,694.50	6,048.69	70,802.51	-
4	SC22	Sugon Cloud Storage Laboratory	ParaStor	Sugon	ParaStor	10	2,560	8,726.42	718.11	106,042.93	-
5	SC22	SuPro Storteck	StarStor	SuPro Storteck	StarStor	10	2,560	6,751.75	515.15	88,491.65	-

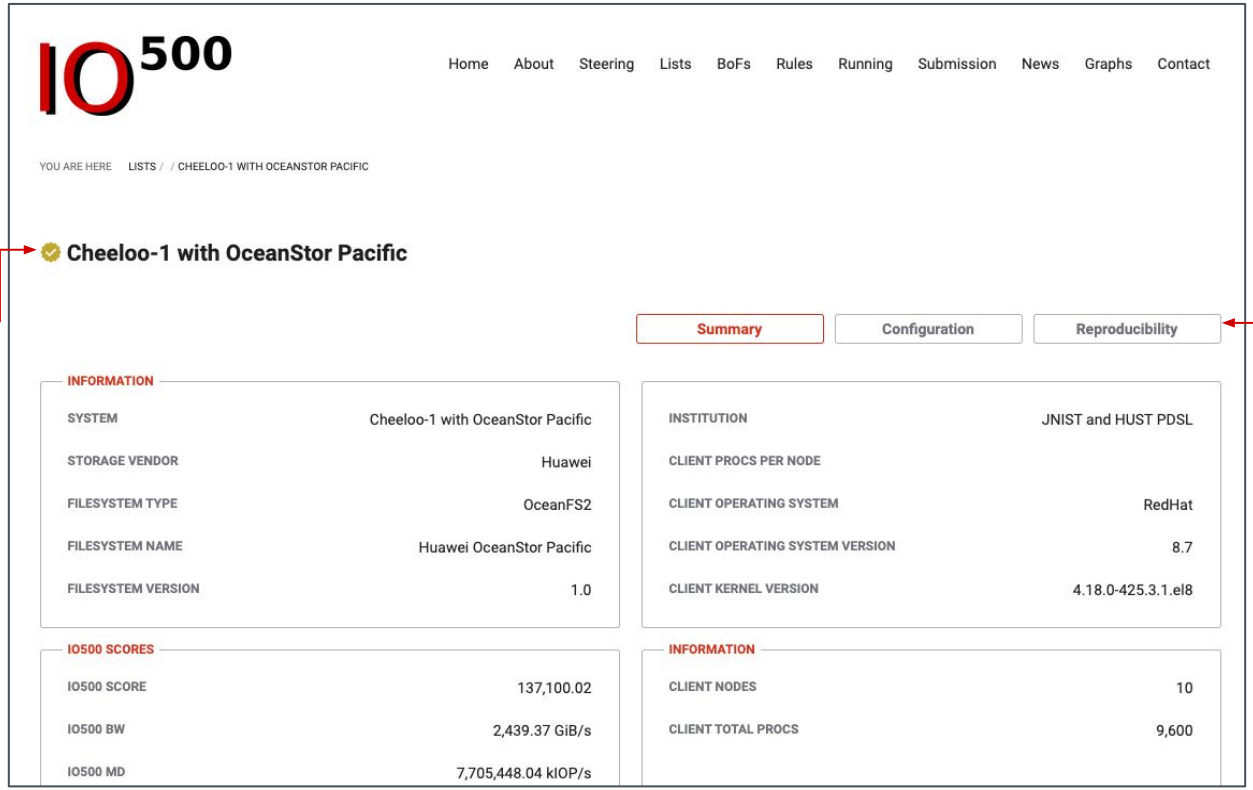
NEW!
Reproducibility is featured on each submission

- Fully Reproducible
- Partially Reproducible
- Proprietary

IO500 Website - Submission View

NEW!

Reproducibility is featured on each submission

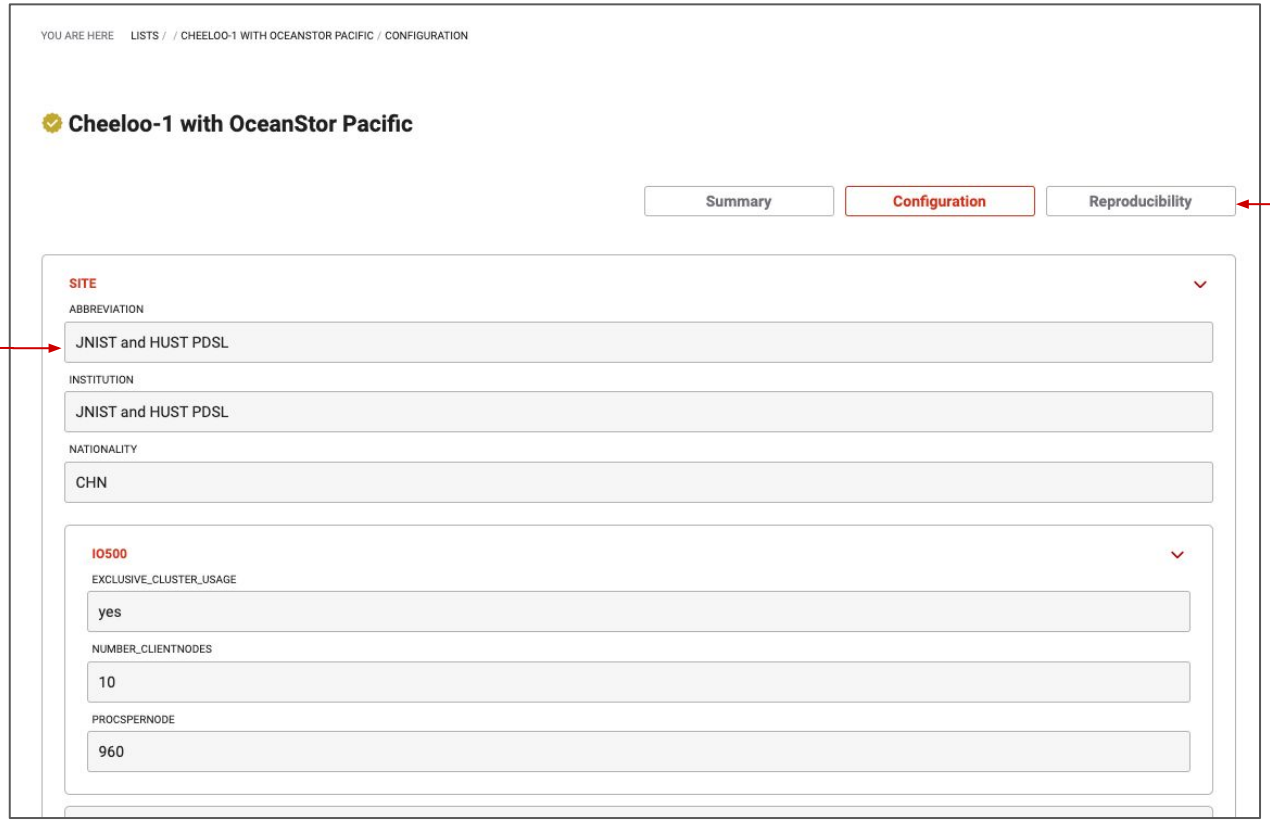


NEW!

Easily **explore** details of each submissions

IO500 Website - Submission Details

NEW!
Visualize all the **details** of each submission



NEW!
Access the **reproducibility** questionnaire

IO500 Submission Management

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

COMING SOON!

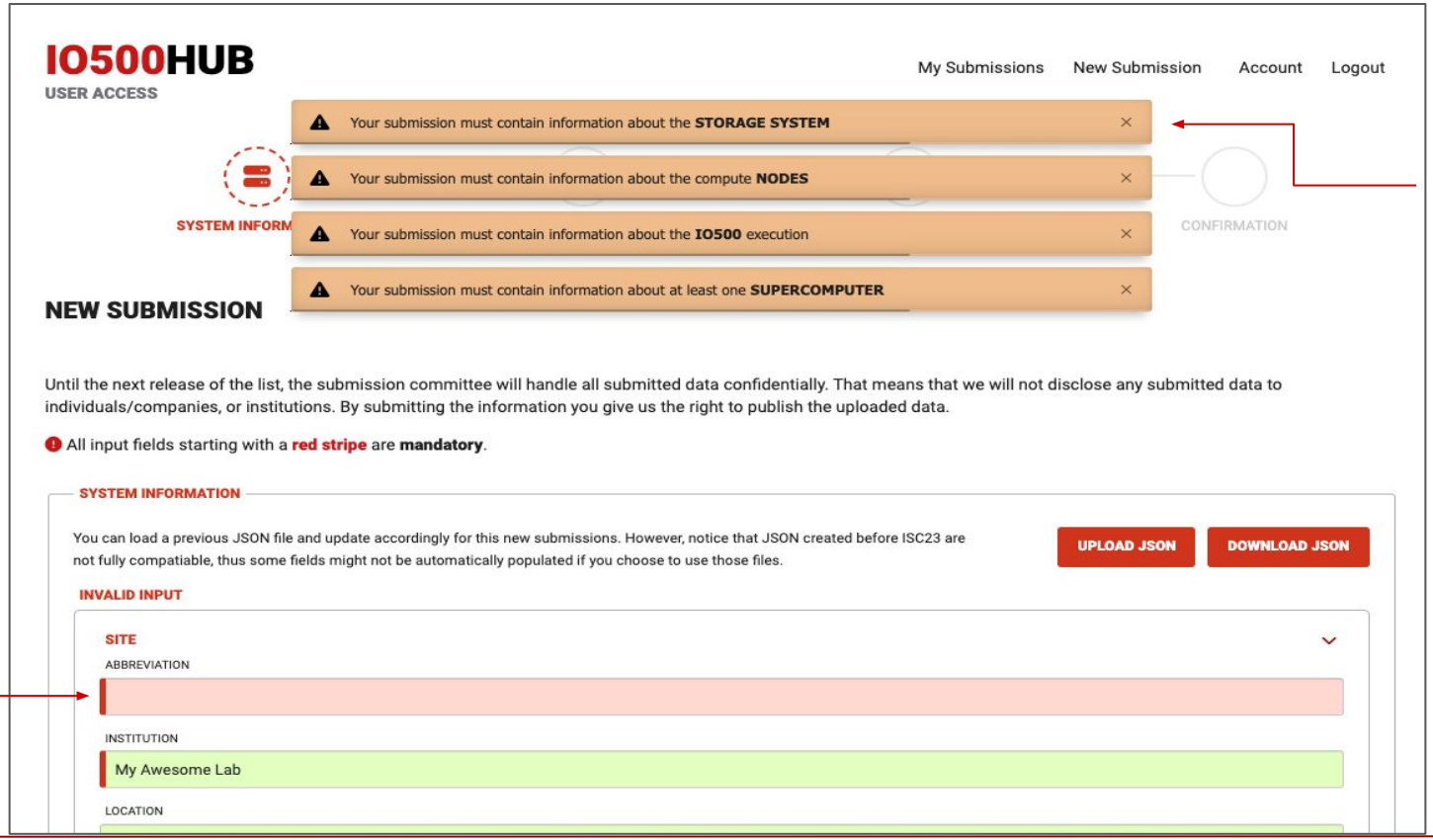
Update metadata and **Reproducibility** Questionnaire of older submissions for inclusion into Production lists

Soliciting volunteers to help with ongoing maintenance and improvements

IO500 Submission Validation

NEW!

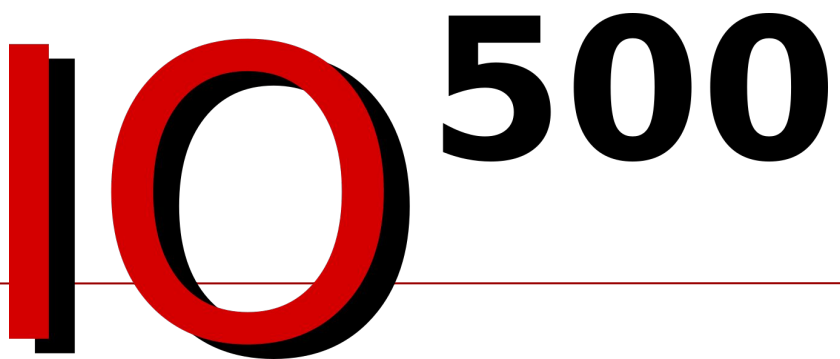
Improved validation of **mandatory** fields



NEW!

Comprehensive **validation** of key information

Community Presentations



Whamcloud

Empowering Lustre Performance Evolution Through IO500

Shuichi Ihara

IO500 Performance History

Whamcloud

Performance improvements go beyond what hardware upgrades can achieve

Hardware Configuration

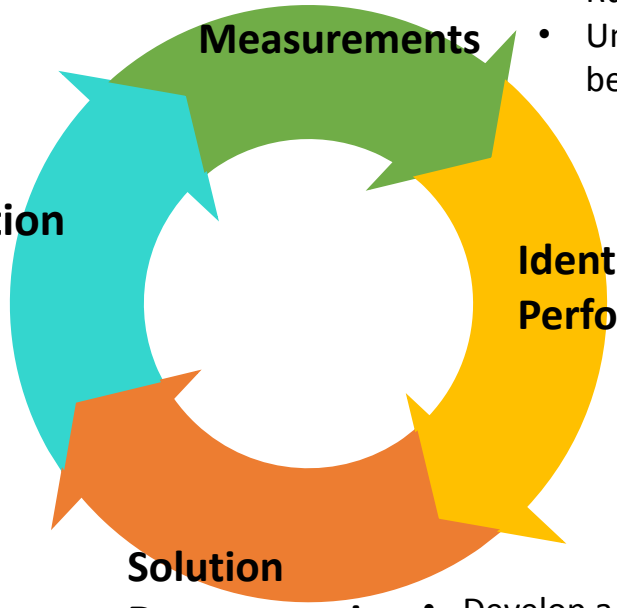
- 4 x Lustre Server VMs
  - 1 MDT, 2 OST
  - 12 x CPU core
  - 142GB RAM
  - 1 x HDR200 InfiniBand
  - 24 x NVMe (shared)
- 10 x Lustre Client
  - 16 x CPU core
  - 96GB RAM
  - 1 x HDR100 InfiniBand

Storage Platform	1x ES400NV	1x ES400NVX	1x ES400NVX2			
	Pre-SC19	SC19	ISC20	ISC22	SC22	ISC23	ISC23/PreSC19
ior-easy-write	25.8	28.62	37.56	55.95	58.07	57.88	2.2x
ior-easy-read	39.9	41.72	45.95	83.86	77.56	79.08	2.0x
ior-hard-write	2.7	2.96	2.77	5.02	5.27	5.38	2.0x
ior-hard-read	8.9	42.19	40.81	39.73	49.36	50.77	5.6x
find	1,735.4	810	1,698.00	6,248.55	12628.78	13,229.11	7.6x
mdtest-easy-write	143.8	152.84	157.22	270.04	312.9	344.70	2.3x
mdtest-easy-stat	455.0	451.97	453.51	740.01	1,278.50	1,276.31	2.8x
mdtest-easy-delete	88.5	132.76	135.09	223.61	272.64	311.16	3.5x
mdtest-hard-write	32.3	79.65	90.47	119.41	157.4	199.36	6.1x
mdtest hard-read	44.9	172.59	169	194.33	238.82	391.09	8.7x
mdtest Hard-stat	20.4	449.93	446.75	514.36	1,214.03	1,105.33	54.1x
mdtest Hard-delete	16.3	75.15	76.94	101.98	122.44	112.58	6.8x
Bandwdith	12.68	19.65	21.02	31.10	32.90	33.43	2.6x
IOPS	91.41	207.6	232.6	368.4	544.2	603.39	6.6x
Score	34.05	63.87	69.93	107.0	133.8	142.03	4.1x

https://io500.org/submissions/view/657

Successful Lustre Performance Improvement cycle

Whamcloud



Performance Measurements

- Run IO500
- Understanding system and Lustre behaviors

Identifying Performance Bottlenecks

- Debug and narrow down the problem
- Identify performance limitations
- Propose solution to address bottleneck and also improve real application workloads

Solution Demonstration

- Develop a prototype of the solution
- Re-run IO500 for performance demonstration
- Collect new results to confirm improvement

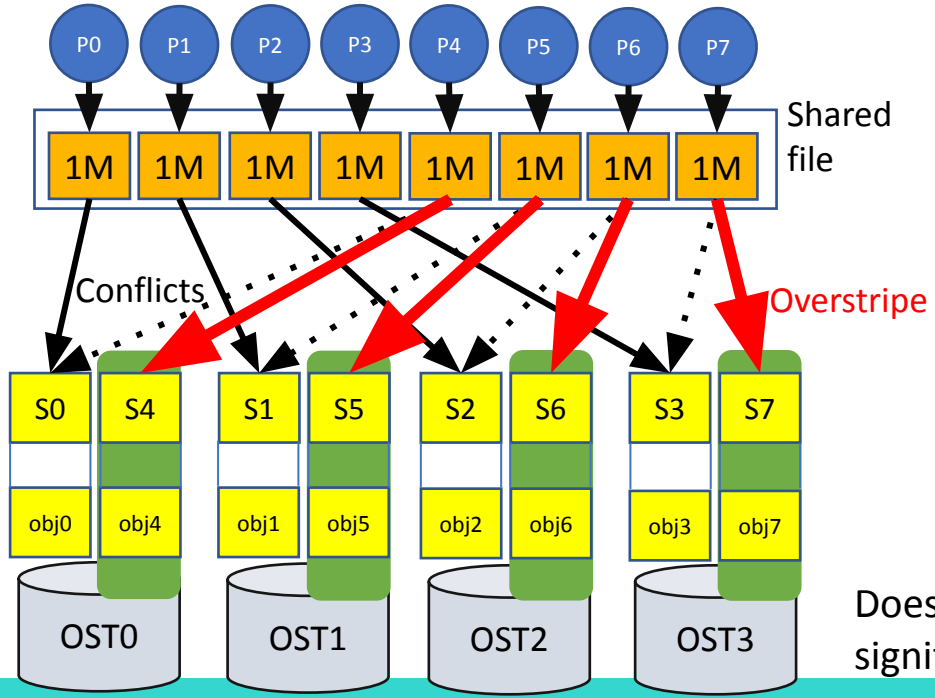
Productization

- Review code and run full regression tests to merge
- Provide feedback to customers and Lustre users

Lustre Overstripe Improves DLM Lock Scalability (Lustre 2.13)

Whamcloud

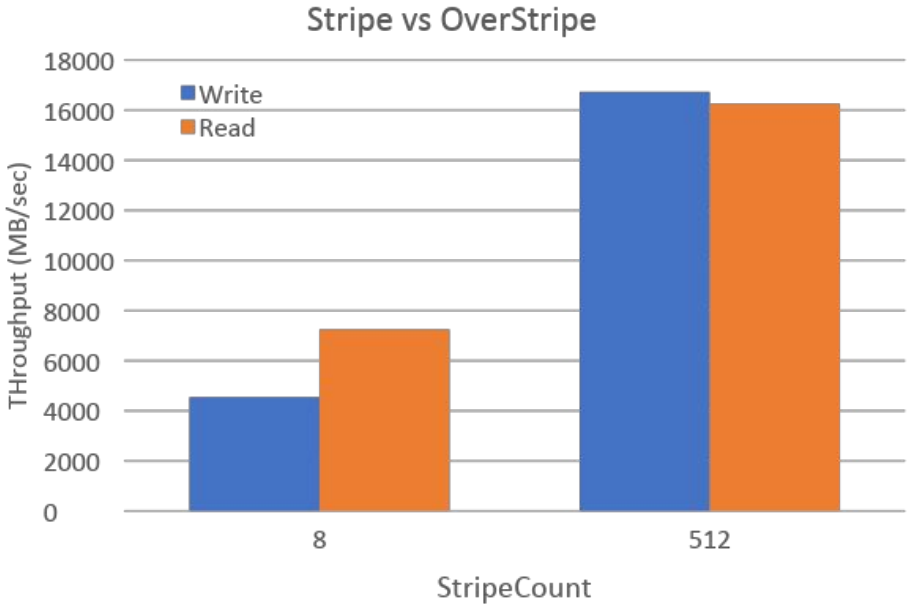
lfs setstripe -c 4 /lustre/file (Lustre Regular Stripe)
lfs setstripe -C 8 /lustre/file (OverStripe)



1MB single shared file

ior -w -r -C -g -i 3 -vv -s 13000 -b 1m -t 1m -a POSIX –e

ES7990(160 x HDD, 2 x OSS, 8 x OST), 32 clients

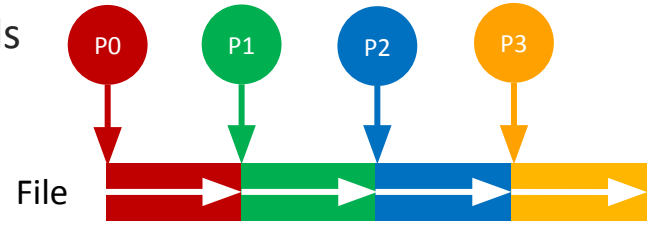


Does not solve `ior-hard-write` entirely, but offers significant performance improvement for single shared file

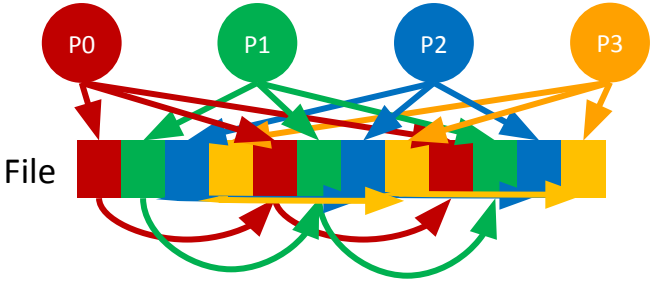
Improvements of Lustre Read Ahead (Lustre-2.14)

Whamcloud

- Accurate detection of I/O patterns
  - Readahead was previously working well for sequential reads
  - Add ”Strided Read” IO pattern for a single shared file
- Change page-granular index to byte-granular offset
  - Support unaligned page (47008-byte in `ior-hard-read`)
  - Avoid many small page RPCs and readahead windows reset
  - Improve readahead cache hit rate

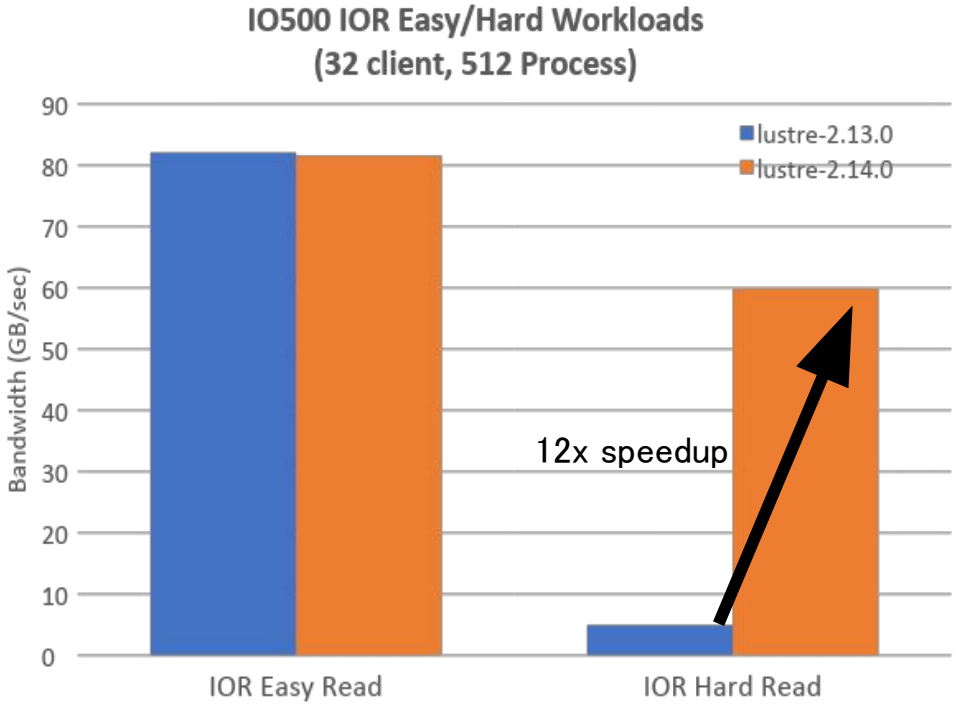


Sequential Read



Stride Read

Performance comparisons of Lustre 2.13 and Lustre 2.14



Readahead stats for `ior-hard-read`

Lustre 2.13

```
# lctl get_param llite.*.read_ahead_stats
llite.exafs-ffff9b96c1349800.read_ahead_stats=
hits                            3340631 samples [pages]
misses                          32901120 samples [pages]
```

Readahed Cache Hit rate: 9%

Lustre 2.14

```
llite.exafs-ffff9b96b8117000.read_ahead_stats=
hits                            33616605 samples [pages]
misses                          4444696 samples [pages]
```

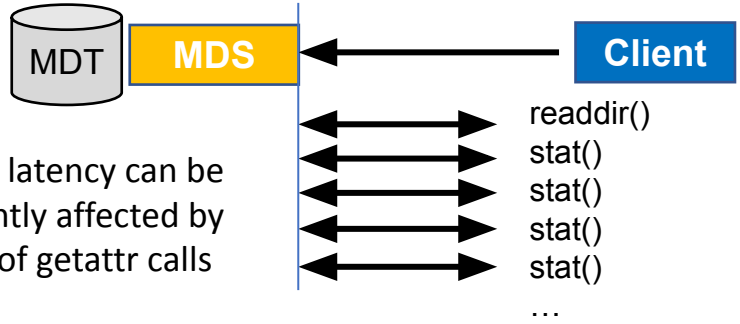
Readahed Cache Hit rate: 88%

Prefetched data by readahead hits expected next read

Lustre Batched RPCs for Statahead (Lustre 2.16)

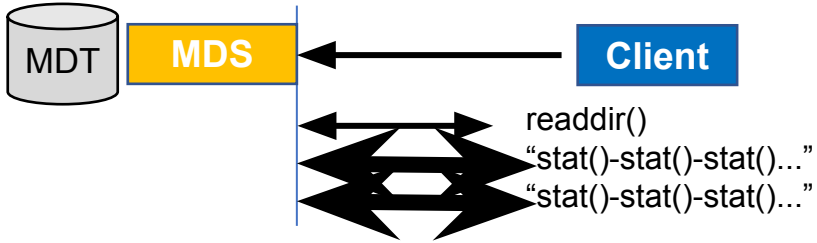
Whamcloud

Traditional statahead

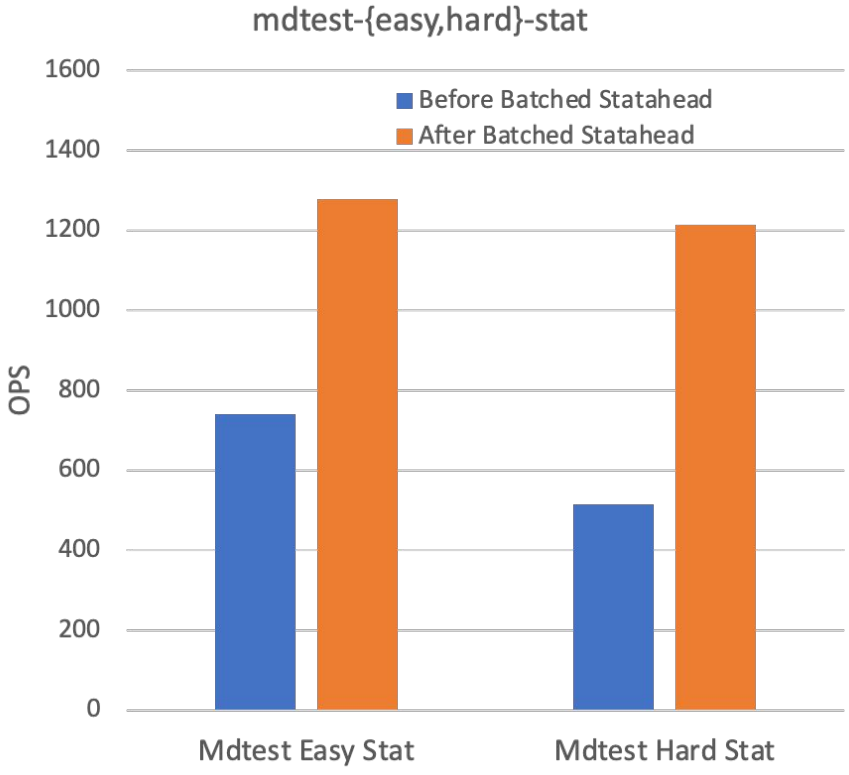


Network latency can be significantly affected by number of getattr calls

Batched statahead



Aggregate multiple getattr RPCs and send them as a batched large request to severs



Additional Lustre Performance Enhancements and Tips

Whamcloud

- Automated MDT directory space/usage balancing (Lustre 2.14/Lustre 2.15)
  - Each unique subdirectory automatically assigned to an MDT, avoid striped directory
- Metadata Overstriping (LU-12273)
  - Similar concept to OST Overstripe, but it allows MDT stripe counts > MDTs
- Developed an external tool for metadata scan/search
  - Use "`lipe_scan`" alternative tool to "`lfs find`", "`find`" and "`pfind`" to scan MDT directly
  - Does not rely on namespace traversal on Lustre clients
  - 7x performance improvements compared to "`pfind`"
- New RHEL Linux kernel for Lustre server and client
  - Upgrading from RHEL7.x to RHEL8.x servers improved metadata performance by 25-30%
  - VFS Parallel Lookup (since Linux 4.7) speeds up `stat()` for shared dir (`mdtest-hard-stat`)

Summary

Whamcloud

- Lustre performance proven on large production HPC systems at many sites
  - IO500 is an important benchmark metric, but it's not the only one
  - In addition to performance, high RAS capability is necessary in large-scale systems
  - IO500 also opened a door for new Lustre performance evolution in HPC, AI, and more
- What's next?
  - Multiple efforts underway to improve unaligned IO (`ior-hard-write`)
    - Direct IO support for unaligned read/write
    - Enable delayed allocation and writeback cache merging in `ldiskfs` backend
  - Cross-file Readahead
    - Expect `mdtest-hard-read` performance boosts
    - It also helps many small file read workload
  - Consider upgrading the Linux kernel for servers (e.g. RHEL 9.x)

Stay tuned!

HPS

Hendrik Nolte, Julian Kunkel

IO500 with GPUDirect and Extended Mode

IO500

- IO500 normally allocates the buffer on CPU only
- IO500 uses currently the timestamp pattern
- Phase-concurrent branch includes options to trigger the benchmarks:
  - allocateBufferDevice
  - gpuDirect
  - The options work as described for the benchmark repositories
- Also includes concurrent phase - runs phases at the same time
  - IOR easy write (20%)
  - RND 1 MB read (40%)
  - MDWorkbench (40%)
- Setting the flags, triggers the options for ALL phases and all benchmarks

Benchmarks

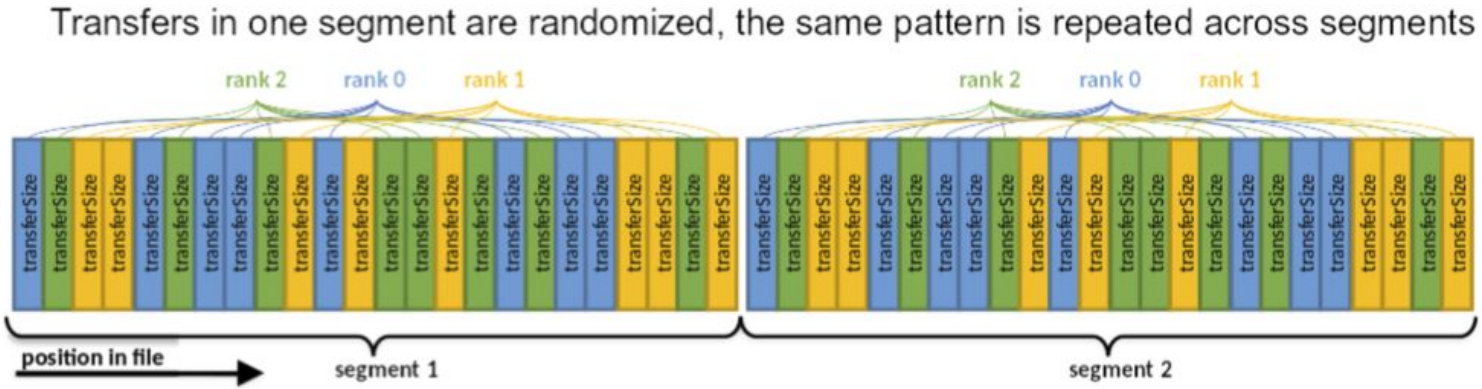
- Core benchmarks IOR/MDTest/MDWorkbench support GPUDirect
  - Normally, IO is done between Client NIC + (host) memory
  - GPUDirect: IO is done between Client NIC + GPU memory - skipping host mem
- We can choose if data buffers and patterns are created/verified on GPU/CPU
- Extra flag: allocateBufferOnGPU=MODE

Mode	Buffer	Creation, Verification (if enabled)	GPUDirect
0	malloc()	CPU	No
1	cudaMallocManaged()	CPU	Optional
2	cudaMallocManaged()	GPU	Optional
3	cudaMalloc()	GPU	Mandatory

- To enable GPUDirect: –gpuDirect
- Requires: POSIX odirect
  GPUDirect supports unaligned blocks (with performance impact)
- Limitations: Verification is supported currently only for timestamp pattern

Mode Extended

- `phase-concurrent` branch also includes the `-mode=extended` option
- introduces new operations like concurrent
  - Default: 20% do write, 40% do reads, 40% do metadata
- And random 4k/1MiB write/reads



Hardware

- IO500 was run on Grete
  - https://www.top500.org/system/180092/
  - CPU: AMD EPYC 7513 32C
  - Interconnect: Infiniband HDR
  - Accelerator: 4xA100 SXM4 80GB
  - Storage: DDN Lustre 130 TiB NVME

Preliminary Results

Task / Mode	0	1	2	2-GPUD	3-GPUD
ior-easy-write [GiB/s]	6.3	7.5	7.0	6.1	5.6
ior-rnd4K-write [GiB/s]	0.2	0.2	0.19	0.02	0.02
mdtest-easy-write [kIOPS]	11.7	11.5	11.5	8.4	8.3
ior-rnd1MB-write [GiB/s]	1.2	0.9	1.2	5.2	3.8
mdworkbench-create	11.0	11.0	11.0	3.4	3.3
find-easy [kIOPS]	2635.5	2219.4	2501.2	2241.7	2433.7
ior-hard-write [GiB/s]	0.7	0.4	0.4	0.1	0.1
mdtest-hard-write [kIOPS]	2.5	2.8	2.8	2.5	2.3
find [kIOPS]	1577.8	1543.6	1473.7	2713.4	2624.0
ior-rnd4K-read [GiB/s]	2.4	0.1	0.2	0.03	0.03
ior-rnd1MB-read [GiB/s]	26.9	2.8	3.3	5.2	4.2
find-hard [kIOPS]	1364.6	1655.5	1398.9	1342.5	1201.2
mdworkbench-bench [kIOPS]	18.6	18.3	3.1	8.4	2.9
concurrent [score]	6.5	6.3	3.6	7.7	4.9
ior-easy-read [GiB/s]	5.8	6.1	3.6	6.2	6.1
mdtest-easy-stat [kIOPS]	28.8	28.7	29.6	207.8	200.0
ior-hard-read [GiB/s]	3.0	2.2	0.24	0.3	0.3
mdtest-hard-stat [kIOPS]	49.5	49.7	46.4	194.0	190.8
mdworkbench-find-delete [kIOPS]	19.9	19.9	20.8	19.9	20.1
mdtest-easy-delete [kIOPS]	21.8	22.3	19.7	22.0	20.0
mdtest-hard-read [kIOPS]	14.4	14.4	4.9	5.0	5.0
mdtest-hard-delete [kIOPS]	5.0	4.9	4.9	5.1	4.7
Score Bandwidth [GiB/s]	2.9	2.5	1.2	1.0	1.0
Score IOPS [kIOPS]	23.8	24.1	20.4	32.6	31.2
ScoreX Bandwidth [GiB/s]	2.4	1.1	0.6	0.6	0.5
ScoreX IOPS [kIOPS]	47.6	48.0	37.1	54.2	48.0

- Number shows the mode
- GPUDirect on/off
- Used a single node on Grete
  - 9 processes
  - 3 GPUs
- Numbers are irrelevant
  → Just want to show it works
  → And how it looks like
- Shows volatility to file count
  slowdown due md create good perf
- Find/Easy hard consistent results, not find.

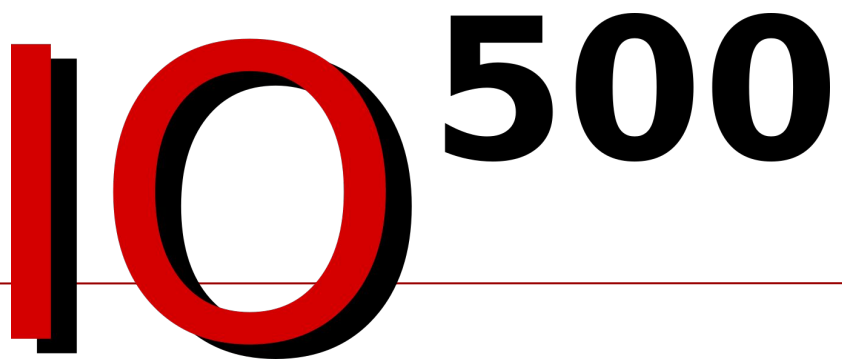
Next Steps

- We need your help!
- Please, test the features on your system
- If there are any issues → open an issue!

How to Get Started

- git clone `https://github.com/IO500/io500`
- git checkout `phase-concurrent`
- set `IOR_HASH=db3c6fb` in `prepare.sh`
- Ensure that you also have CUDA available

Roadmap



Roadmap for the IO500

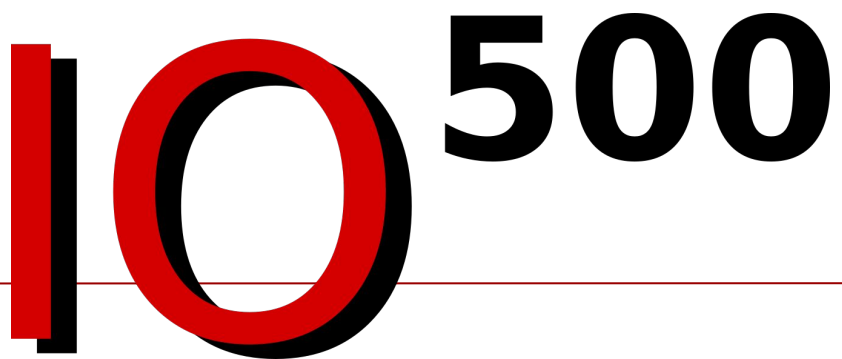
- Create proposals with rationale and details for random I/O and find-hard
  - Hold community meeting when proposals are ready
  - Target February/March 2024 if topics to discuss

- Continued improvements of **`io500.org`** submissions page
  - Add more mandatory fields/sections for storage details
  - Help text to clarify field usage for more accurate input
  - Please give feedback and be patient in the transition

ISC-HPC 2024 (May 12-16, 2024)

- Call for submission: Mar 20th
- Testing phase ends: Mar 27th
  - Code freeze, but please test beforehand!
- Submission deadline: May 3rd
- List release: BoF date TBD (ISC'24 during May 12-17)
- Looking forward to many more Production submissions

Benchmark Phases and Extended Access Patterns



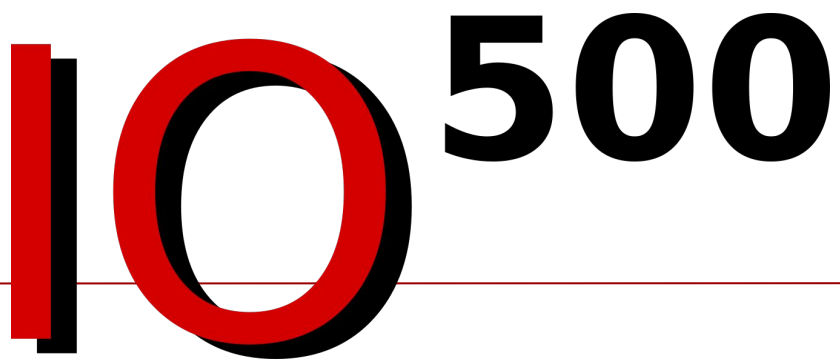
Benchmark Phases and Extended Access Patterns

- Want to add new phases for 4KB random, find hard, rename
  - Need to finalize details of new phases, exact implementation
- Need better description for all I/O patterns
  - Motivation, use cases, description of actual IO pattern, …
- Comparison of score between standard / extended modes
  - New phases may change the result of existing phases in some cases
  - Compute **current** IO500 score based on **current** phases
  - Allows to compare new results with historical submissions
  - Track Extended IO500 score for transition of ranking in future
- Prototype code available, needs further refinement

Open Questions About Extended Access Patterns

- `ior-random-write` pseudo-random pattern to allow data verification
  - May be too small to avoid cache effects for `ior-random-read` pattern
  - Considering using `ior-easy-write` files for random reads to have enough data
  - Could read blocks totally randomly in this case
- Should `ior-random-write` be counted in the score, or only reads?
  - Con: Relatively few HPC workloads have **only** random writes
  - Pro: 4KiB write IOPS often tested/reported, and provide lower bound for storage
- Want `find-hard` to be “harder” than just “`find` in `mdtest-hard/` dir”
  - Output find filename 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 new `find` score?
- Should a directory `mdtest-rename` phase be added?
  - Is this a hierarchical namespace, or a flat namespace with ‘`/`’ in object names?
- Expect runtime would increase by about 20 minutes if all phases added

Voice of the Community & Open Discussion



Open Floor

- Open discussion about proposed new phases
- Feedback on the submission process
- How to collect storage system metadata more easily/accurately?
  - Community volunteers to assist for their favorite storage system?
  - Vendors develop schema/tools for their metadata?
  - How to handle server-side collection?

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