The 8th IO500 and the Virtual Institute of I/O
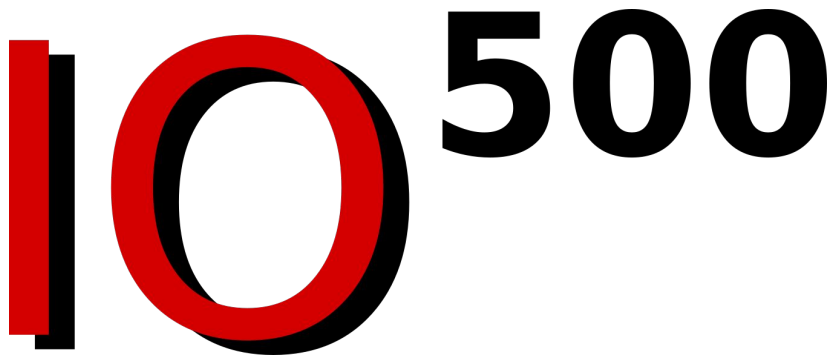
Andreas Dilger, Dean Hildebrand, Julian M. Kunkel, Jay Lofstead, George Markomanolis

BoF Agenda

1. **What's New with IO500** – George Markomanolis
2. **The New IO500 List Analysis** – Dean Hildebrand
3. **8th IO500 Award Presentations** – Andreas Dilger
4. **Roadmap** – Julian Kunkel
5. **Questions & Discussion Session** – Jay Lofstead

Note: Supplementary (extended) presentations available on the BOF webpage

What's New: Webpage

- Improving and fixing bugs, thanks to Jean Luca Bez
  - Thanks to the people who reported them
  - Submit issues via: https://github.com/IO500/webpage/
- Including SCC list and better organize some parts of the website

SC19

IO500
10 Node
Full
Historical
SCC

IO500 Status

- Versioning of benchmark itself continues to work
  - We check the versions from the submissions
  - Please use the correct one isc<YEAR>, sc<YEAR>

- Exploring usage of new phases in benchmark
  - Open for discussion/modification for future inclusion
  - Optional `--mode=extended` activates experimental phases
    - `ior-rand` (small-block random read/write)
    - `find-easy`, `find-hard` (many small dirs, single large dir with complex scan)
    - `md-workbench` (concurrent read-write workload)

- Exploring storage schemas, improving with your feedback

IO500 Organization Status

- A non-profit organization IO500 Foundation
  - o Domain, mailing list, servers, Github belongs to IO500 Foundation
- Updating the new web page:
  - o https://io500.org/
  - o Contribute at https://github.com/IO500/webpage
- Please join our new mailing list:
  - o https://io500.org/contact
- Please join our Slack:
  - o https://io500workspace.slack.com/
  - o Join link: https://rb.gy/sn8esm

Board Changes

- John Bent has stepped down from the board
  - We thank him for co-founding and supporting the IO500

- Dean Hildebrand is replacing John Bent on the board
  - Current Position
    - Technical Director for Storage and HPC
    - Office of the CTO, Google Cloud
  - Previous Position
    - Principal Research Staff Member on Spectrum Scale (GPFS)
    - IBM Research

Ongoing: Specification of the Hardware Schema

- Improved submission schema toward more intuitive, less ambiguous
    - Supporting storage-system specific schemas
    - Remove uncertainty about the semantics of fields

- 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 utility, can be customized to best collect information
- More improvements are planned
    - Need community input to refine schemas and better populate submissions
    - More submission information allows better analysis and comparisons

In-Person and Virtual
Supercomputing Student Cluster Competition 2021

- IO500 is part of the benchmarks in the SCC
- Organization of the SCC
  - The in-person teams will bring their own hardware
  - The remote teams will have access to Microsoft Azure
  - Totally 11 teams
  - If SC21 is virtual-only, all teams will use Azure or another cloud provider
- More information about the teams:
  https://www.studentclustercompetition.us/

Lists and Analysis

IO^{500}

Growth in Length and Institutions
IO500 List
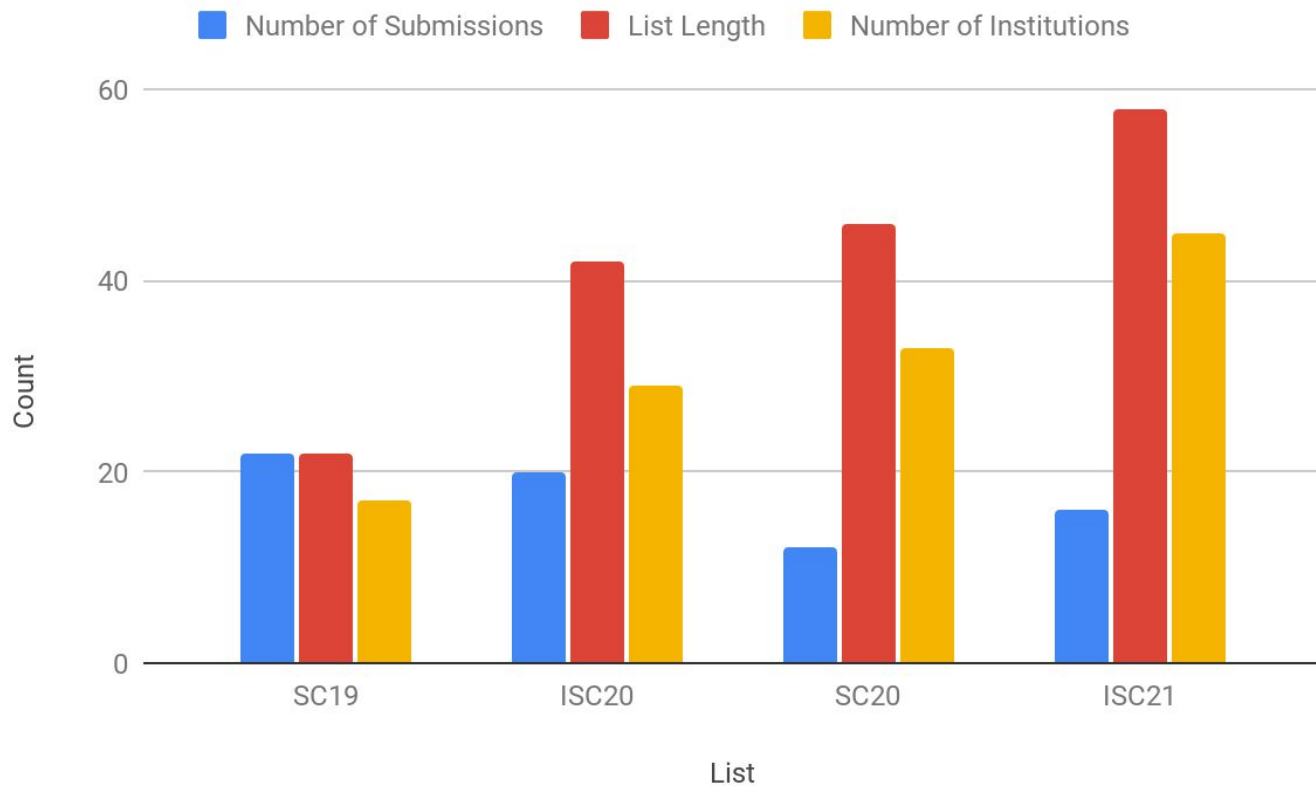
Growth in Length and Institutions
IO500 - 10-Node Challenge List

Total Bandwidth
IO500 List
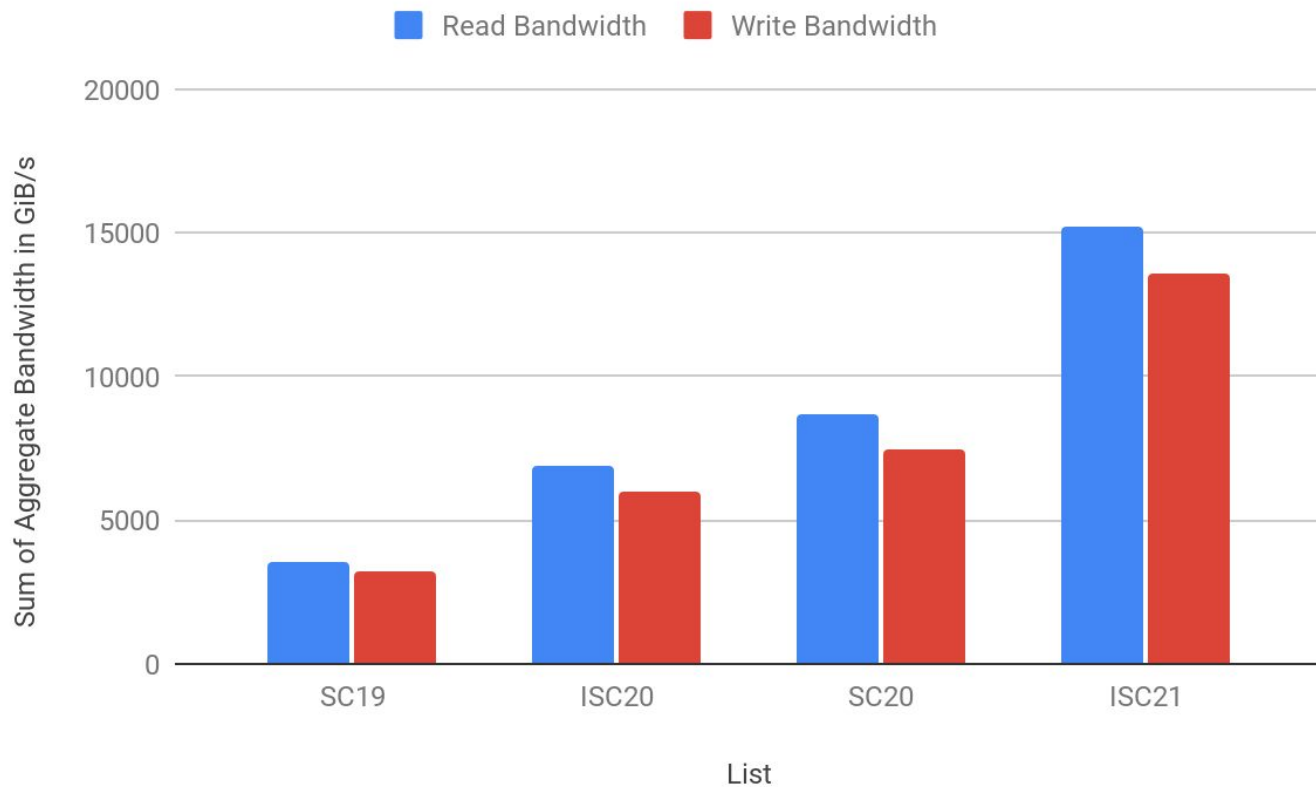
Median Scores
IO500 List

Score growing by
11-32%

Bandwidth Score
growing by 21-27%

Metadata Score
growing by 9-32%
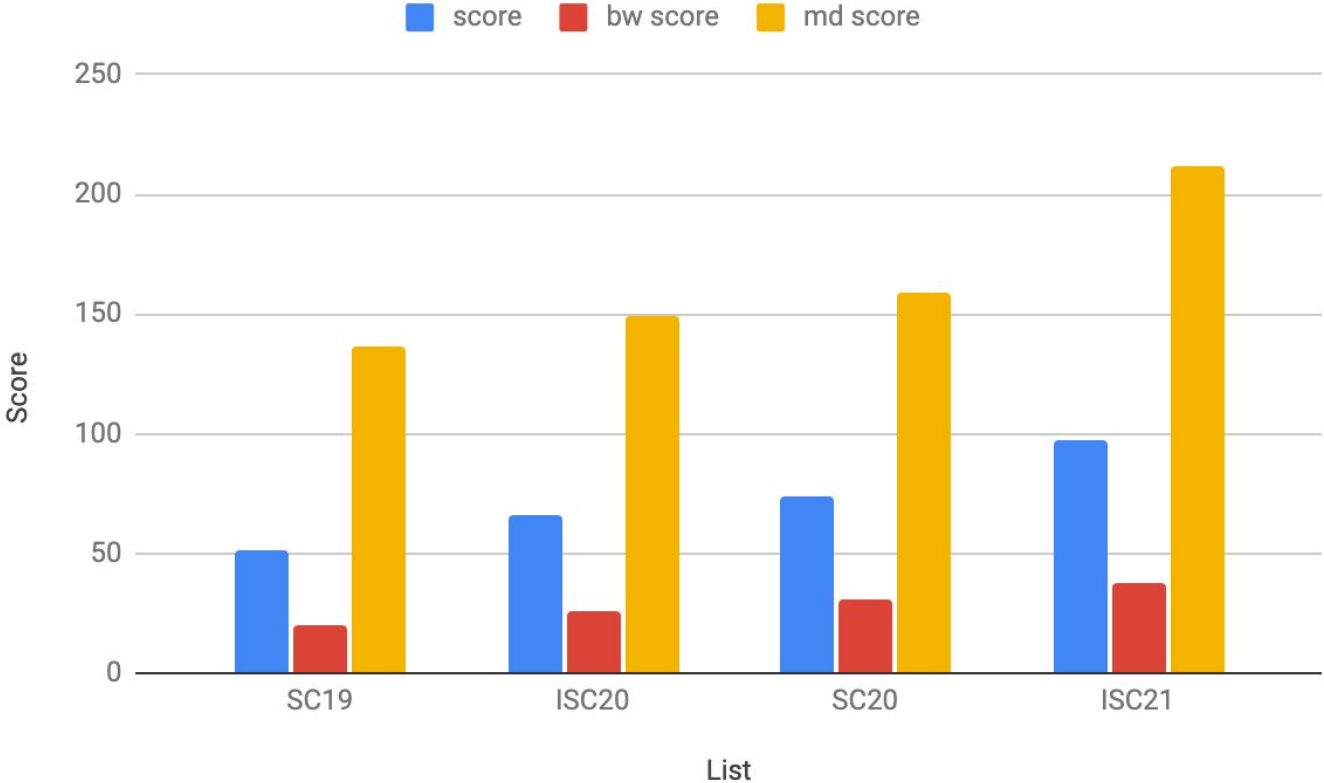
Growth in Max Scores per Client
IO500 List
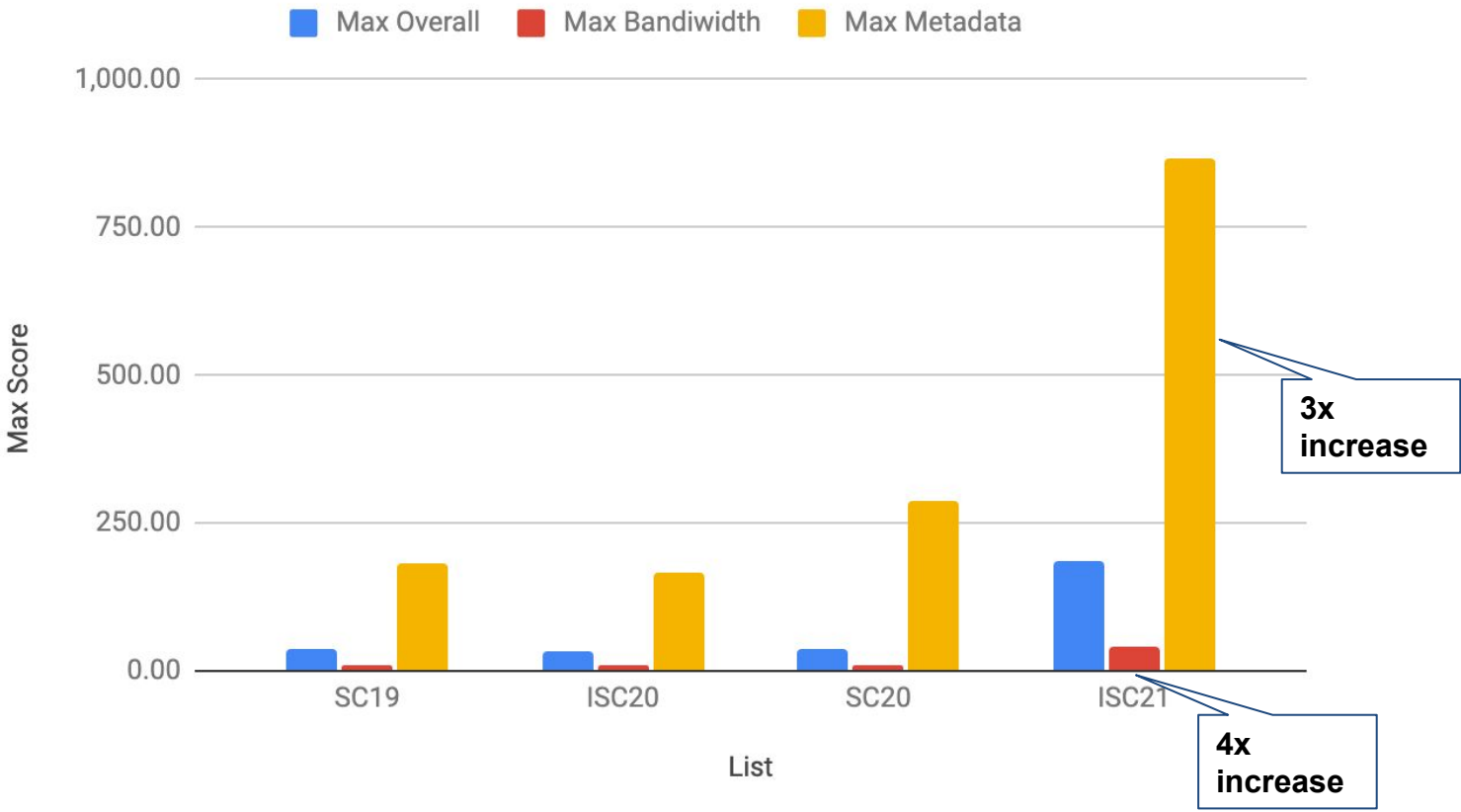
Growth in Max Scores per Client
IO500 - 10-Node Challenge List
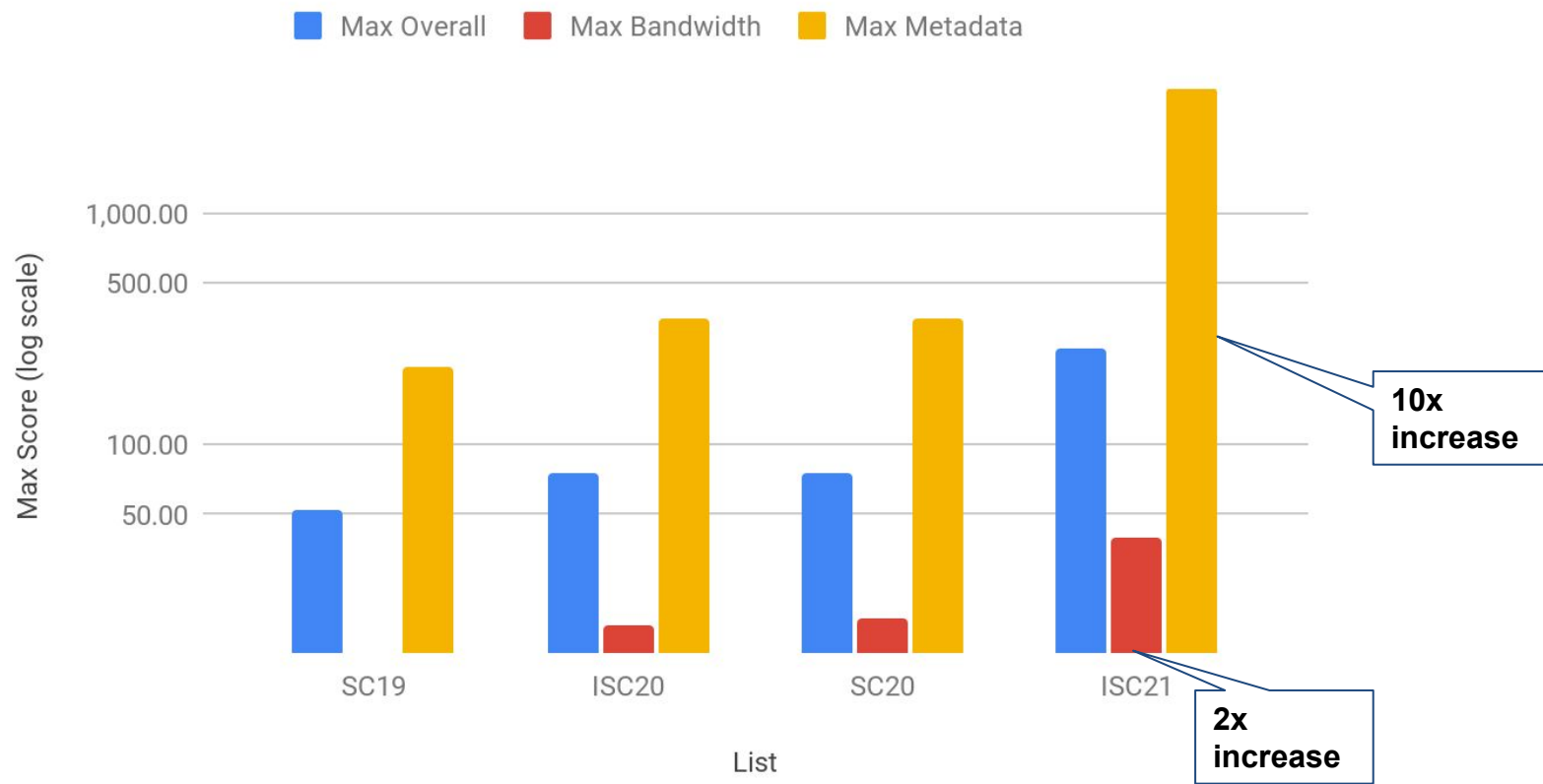
Growth in Max Score per Storage Server
IO500 - List

Per-client scores are growing by orders of magnitude

Per-storage server scores are growing incrementally

Note: Working to clean up metadata server data to provide metadata score per server growth.

Award Ceremony

IO^{500}

Six Awards

- Full List
  - Bandwidth
  - Metadata
  - Overall

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

10 node challenge - Bandwidth Winner

10 Node ISC21 List

IO500   10 Node

Sorted by BW

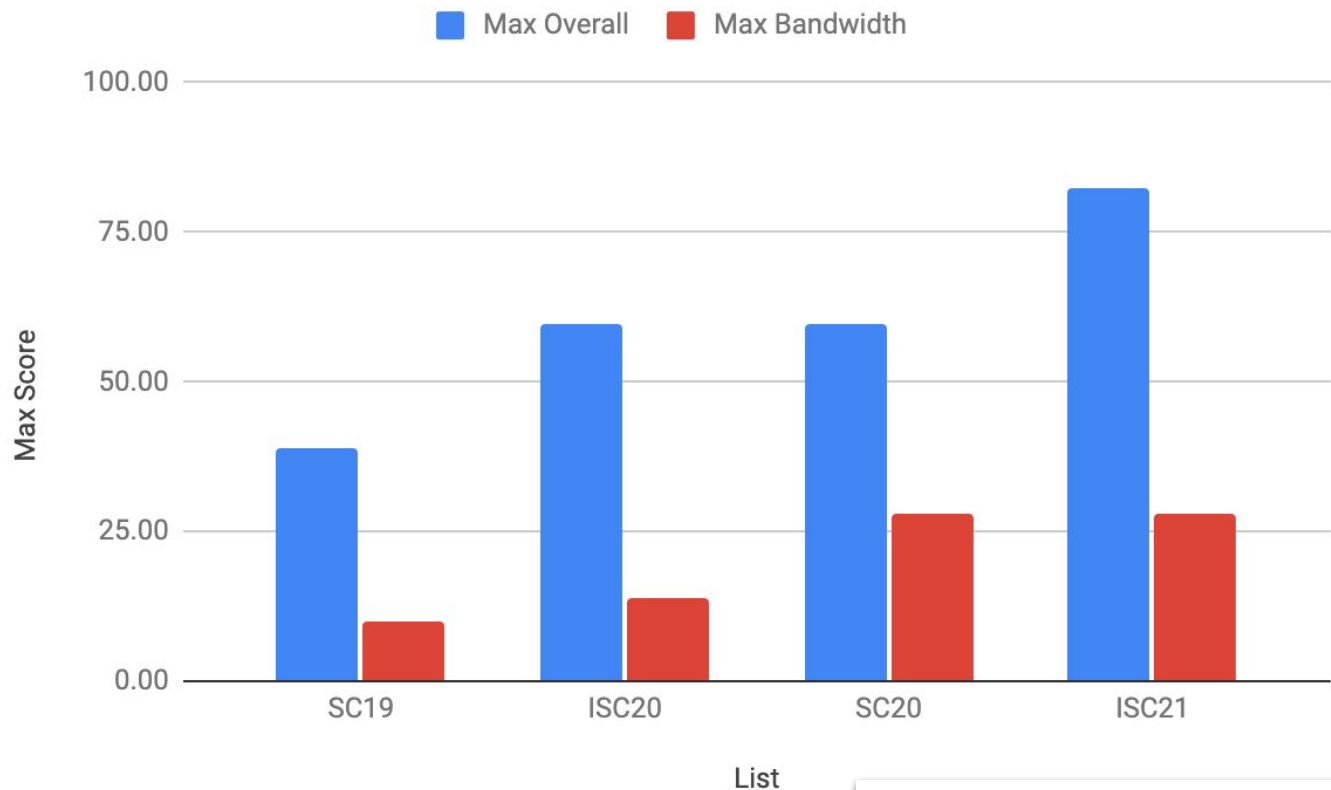| #	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE	BW ↑ (GIB/S)	MD (KIOP/S)
1	ISC21	Intel	Endevour	Intel	DAOS	10	1,440		398.77	
2	ISC21	Pengcheng Laboratory	Pengcheng	Pengcheng	MadFS	10	1,800		193.77	
3	SC20	Forschungszentrum Juelich (FZJ)	JUWELS	HPEDDN	IME	10	400		178.11	
4	SC20	TACC	Frontera	DDN	IME	10	280		176.23	
5	ISC20	Intel	Wolf	Intel	DAOS	10	420		164.77	
6	ISC21	Supermicro		Supermicro	DAOS	10	1,120		112.17	
7	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	10	960		105.28	
8	SC19	National Supercomputing Centre, Singapore	Aspire 1	DDN	IME	10	160		101.75	
9	SC20	DDN	DIME	DDN	IME	10	110		101.60	
10	ISC20	Argonne National Laboratory	Presque	Argonne National Laboratory	DAOS	10	380		95.80	

IO500

Certificate

IO-500 Performance Certification

This Certificate is awarded to:

Intel (Endeavour)
#1 in the 10 Node Challenge BW Score

IO**500**

July 2021

IO-500 Steering Board

https://io500.org/list/isc21/ten

IO**500**

10-Node Challenge - Metadata Winner

10 Node ISC21 List

	IO500	10 Node

Sorted by MD

#		INFORMATION							IO500	
	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE	BW (GIB/S)	MD ↑ (KIOP/S)
1	ISC21	Pengcheng Laboratory	Pengcheng	Pengcheng	MadFS	10	1,800		193.77	34,777.27
2	ISC21	Intel	Endevour	Intel	DAOS	10	1,440		398.77	8,671.65
3	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	10	960		105.28	3,567.85
4	ISC20	Intel	Wolf	Intel	DAOS	10	420		164.77	3,493.56
5	ISC20	TACC	Frontera	Intel	DAOS	10	420		79.16	3,271.49
6	ISC21	National Supercomputer Center in GuangZhou	Venus2	National	kapok	10	480		91.64	2,452.87
7	ISC20	Argonne National Laboratory	Presque	Argonne National Laboratory	DAOS	10	380		95.80	2,026.80
8	ISC21	Supermicro		Supermicro	DAOS	10	1,120		112.17	1,535.63
9	SC20	EPCC	NextGENIO	BSC & JGU	GekkoFS	10	3,800		45.79	1,251.32
10	SC20	Johannes Gutenberg University Mainz	MOGON II	JGU (ADA-FS) & BSC (NEXTGenIO)	GekkoFS	10	240		22.97	1,223.59

IO500

ISC21

21

Certificate

IO-500 Performance Certification

This Certificate is awarded to:

Pengcheng Laboratory (Cloud Brain II)
#1 in the 10 Node Challenge MD Score

IO⁵⁰⁰

July 2021

IO-500 Steering Board

https://io500.org/list/isc21/ten

10-Node Challenge - Winner

10 Node ISC21 List

IO500 | 10 Node

Sorted by score

# ↑	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	IO500 SCORE ↑	IO500 BW (GIB/S)	IO500 MD (KIOP/S)
1	ISC21	Pengcheng Laboratory	Pengcheng	Pengcheng	MadFS	10	1,800	2,595.89	193.77	34,777.27
2	ISC21	Intel	Endevour	Intel	DAOS	10	1,440	1,859.56	398.77	8,671.65
3	ISC20	Intel	Wolf	Intel	DAOS	10	420	758.71	164.77	3,493.56
4	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	10	960	612.87	105.28	3,567.85
5	ISC20	TACC	Frontera	Intel	DAOS	10	420	508.88	79.16	3,271.49
6	ISC21	National Supercomputer Center in GuangZhou	Venus2	National	kapok	10	480	474.10	91.64	2,452.87
7	ISC20	Argonne National Laboratory	Presque	Argonne National Laboratory	DAOS	10	380	440.64	95.80	2,026.80
8	ISC21	Supermicro		Supermicro	DAOS	10	1,120	415.04	112.17	1,535.63
9	SC19	NVIDIA	DGX-2H SuperPOD	DDN	Lustre	10	400	249.50	86.97	715.76
10	SC20	EPCC	NextGENIO	BSC & JGU	GekkoFS	10	3,800	239.37	45.79	1,251.32

IO500

Certificate

IO-500 Performance Certification

This Certificate is awarded to:

Pengcheng Laboratory (Cloud Brain II)
#1 in the 10 Node Challenge

IO**500**

July 2021

IO-500 Steering Board

https://io500.org/list/isc21/ten

IO**500**

Full list - Bandwidth Winner

IO500 ISC21 List

IO500	10 Node

#	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE	BW ↑ (GIB/S)	MD (KIOP/S)
1	ISC21	Pengcheng Laboratory	Pengcheng	Pengcheng	MadFS	512	36,864		3,421.62	
2	SC20	JCAHPC	Oakforest-PACS	DDN	IME	2,048	4,096		697.20	
3	ISC20	Korea Institute of Science and Technology Information (KISTI)	NURION	DDN	IME	2,048	2,048		515.59	
4	ISC21	Intel	Endevour	Intel	DAOS	10	1,440		398.77	
5	ISC20	Intel	Wolf	Intel	DAOS	52	1,664		371.67	
6	ISC20	Oracle Cloud Infrastructure	BeeGFS on Oracle Cloud	Oracle Cloud Infrastructure	BeeGFS	270	3,240		293.05	
7	ISC21	Google Cloud	Google	DDN	Lustre	1,000	5,000		282.78	
8	SC19	National Supercomputing Center in Changsha	Tianhe-2E	National University of Defense Technology	Lustre	480	5,280		209.43	
9	ISC21	University of Florida	HiPerGator	DDN	Lustre	140	4,480		181.37	
10	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	36	3,456		176.37	

IO500

Certificate

IO-500 Performance Certification

This Certificate is awarded to:

Pengcheng Laboratory (Cloud Brain II)
#1 in the IO500 BW Score

IO**500**

July 2021

IO-500 Steering Board

https://io500.org/list/isc21/io500

Full list - Metadata Winner

IO500 ISC21 List

| IO500 | 10 Node |

Sorted by MD

#	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE	IO500 BW (GIB/S)	MD ↑ (KIOP/S)
1	ISC21	Pengcheng Laboratory	Pengcheng	Pengcheng	MadFS	512	36,864		3,421.62	396,872.82
2	ISC21	Intel	Endevour	Intel	DAOS	10	1,440		398.77	8,671.65
3	ISC20	Intel	Wolf	Intel	DAOS	52	1,664		371.67	8,649.57
4	ISC20	TACC	Frontera	Intel	DAOS	60	1,440		78.31	7,449.56
5	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	36	3,456		176.37	5,545.61
6	SC19	WekaIO	WekaIO on AWS	WekaIO	WekaIO Matrix	345	8,625		174.74	5,045.33
7	ISC21	National Supercomputer Center in GuangZhou	Venus2	National	kapok	18	720		104.52	3,195.53
8	ISC20	Argonne National Laboratory	Presque	Argonne National Laboratory	DAOS	16	544		108.19	2,668.57
9	ISC21	Supermicro		Supermicro	DAOS	10	1,120		112.17	1,535.63
10	SC20	EPCC	NextGENIO	BSC & JGU	GekkoFS	10	3,800		45.79	1,251.32

IO500

Certificate

IO-500 Performance Certification

This Certificate is awarded to:

Pengcheng Laboratory (Cloud Brain II)
#1 in the IO500 MD Score

IO⁵⁰⁰   **July 2021**

IO-500 Steering Board

https://io500.org/list/isc21/io500

Full list - Winner

IO500 ISC21 List

| IO500 | | 10 Node | | **Sorted by Score** |

# ↑	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE ↑	IO500 BW (GIB/S)	MD (KIOP/S)
1	ISC21	Pengcheng Laboratory	Pengcheng	Pengcheng	MadFS	512	36,864	36,850.37	3,421.62	396,872.82
2	ISC21	Intel	Endevour	Intel	DAOS	10	1,440	1,859.56	398.77	8,671.65
3	ISC20	Intel	Wolf	Intel	DAOS	52	1,664	1,792.98	371.67	8,649.57
4	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	36	3,456	988.99	176.37	5,545.61
5	SC19	WekaIO	WekaIO on AWS	WekaIO	WekaIO Matrix	345	8,625	938.95	174.74	5,045.33
6	ISC20	TACC	Frontera	Intel	DAOS	60	1,440	763.80	78.31	7,449.56
7	ISC21	National Supercomputer Center in GuangZhou	Venus2	National	kapok	18	720	577.93	104.52	3,195.53
8	ISC21	Google Cloud	Google	DDN	Lustre	1,000	5,000	569.99	282.78	1,148.90
9	ISC20	Argonne National Laboratory	Presque	Argonne National Laboratory	DAOS	16	544	537.31	108.19	2,668.57
10	SC19	National Supercomputing Center in Changsha	Tianhe-2E	National University of Defense Technology	Lustre	480	5,280	453.68	209.43	982.78

IO500

Certificate

IO-500 Performance Certification

This Certificate is awarded to:

Pengcheng Laboratory (Cloud Brain II)
#1 in the IO500

IO**500**   **July 2021**

IO-500 Steering Board

https://io500.org/list/isc21/io500

List of Awarded Systems in the Ranked Lists

10-Node	Bandwidth	Intel Endeavour	DAOS	398.77	GiB/s
	Metadata	Pengcheng Laboratory	MadFS	34777.27	kIOPS
	Overall	Pengcheng Laboratory	MadFS	**2595.89**	**score**
IO500	Bandwidth	Pengcheng Laboratory	MadFS	3421.62	GiB/s
	Metadata	Pengcheng Laboratory	MadFS	396872.82	kIOPS
	Overall	Pengcheng Laboratory	MadFS	**36850.37**	**score**

IO⁵⁰⁰

Roadmap

IO⁵⁰⁰

Roadmap for the IO500

- Continue improving new website
- Continue to improve the system schema
  - Uniformity of presentation and related information
  - We need the community help to define suitable schemas
  - We need submitters to populate information about their systems
  - Extending the scripts to automatically collect system metadata
- Earlier release for the CFS and test version release schedule
- Documentation of rationales for benchmarking phases
- Exploring the integration of new benchmarking phases

IO**500**

Supplementary Presentations

Due to time constraints, additional presentations are on our BoF page:
https://io500.org/pages/bof-isc21

- The Virtual Institute for I/O

  - *Julian Kunkel*

- Community presentation: There is Nothing Mysterious Behind MadFS

  - *Kang Chen (Tsinghua University)*

- Community presentation: An Analysis of the IO500 for Modeling Storage Systems

  - *Luke Logan (Illinois Tech)*

Please watch these presentations and discuss with us on the IO500 Slack after the BoF...

Discussion about new benchmarking phases

- What new benchmarking phases are requested?
- Open discussion
    - Random read/write of 4 KB?
    - Find separation: split into `find-easy`/`find-hard`?
    - Metadata concurrent operations?

IO⁵⁰⁰

Edit or add functionalities to IO500

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.

https://io500.org/rules-benchmark

Issues Regarding Fair Comparisons

- Production vs. benchmark-only system
- Production vs. research file system software
- Non-redundant vs. RAID/Erasure Coded storage
- Vendor submission vs. end-user submission
- Cloud vs. on-prem
- Ephemeral vs. persistent file system
- Storage media technologies

Open Floor