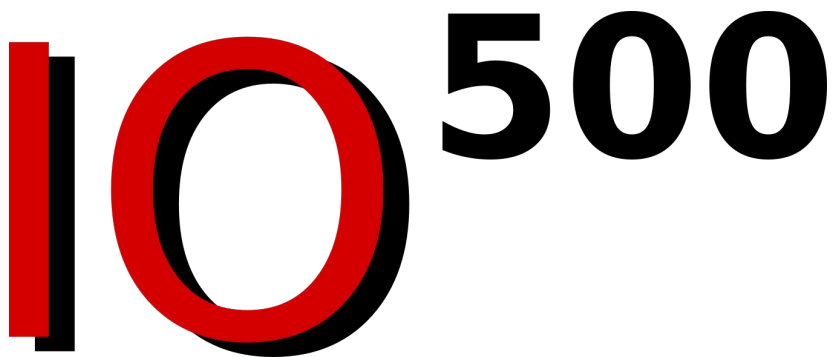


# The 9<sup>th</sup> IO500 and the Virtual Institute of I/O

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

The logo for the IO500 benchmark. It features the letters '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 letters 'v4IO' in a stylized font. The 'v' and '4' are blue, and the 'IO' is red, all with black outlines. The letters are set against a black background.

# BoF Agenda

---

1. **Welcome** – Julian Kunkel
2. **What's New with IO500** – George Markomanolis
3. **The New IO500 List Analysis** – Dean Hildebrand
4. **Award Presentations** – Julian Kunkel
5. **Community Presentation** - Radita Liem
6. **Roadmap** – Andreas Dilger
  - **List Splitting Proposal** – Dean Hildebrand
  - **Benchmark Phases and Extended Access Patterns** – Jay Lofstead
7. **Questions & Discussion Session** – Jay Lofstead

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



## 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
- Explanations with video: [https://www.youtube.com/watch?v=R\\_Fq\\_ks4hnM](https://www.youtube.com/watch?v=R_Fq_ks4hnM)

# In-Person and Virtual SC'21 Student Cluster Competition 2021



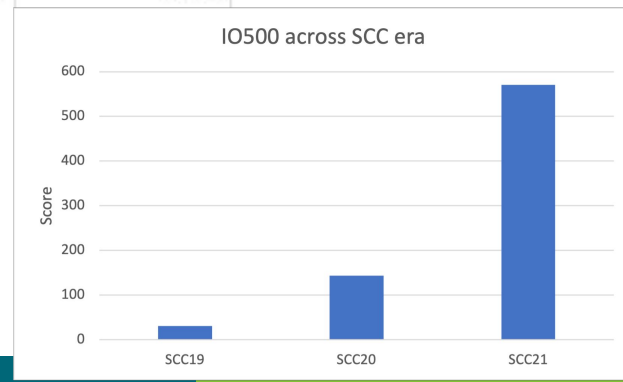
- IO500 is part of the benchmarks in the SCC
- Organization of the SCC
  - The in-person teams will bring their own hardware
  - The remote teams will have access to Microsoft Azure and Oracle cloud
  - For the benchmarks only Azure was used
  - Totally 10 teams (9 teams did submit IO500 results)
  - More information about the teams:

<https://www.studentclustercompetition.us/>

# Results - Student Cluster Competition 2021

#	Team	BW	MD	Score
1	Tsinghua University	46.82	6950.42	570.46
2	ShanghaiTech University	2.79	264.83	27.20
3	University of California San Diego	7.65	47.11	19.00
4	Jinan University	2.23	108.63	15.60
5	Georgia Institute of Technology	2.36	103.19	15.60
6	Massachusetts Green Team	3.32	56.89	13.76
7	Southern University of Science and Technology	1.06	71.25	8.70
8	Peking university	4.05	18.14	8.58
9	Wake Forest University	0.68	9.01	2.48

That's  
Mad(fs)



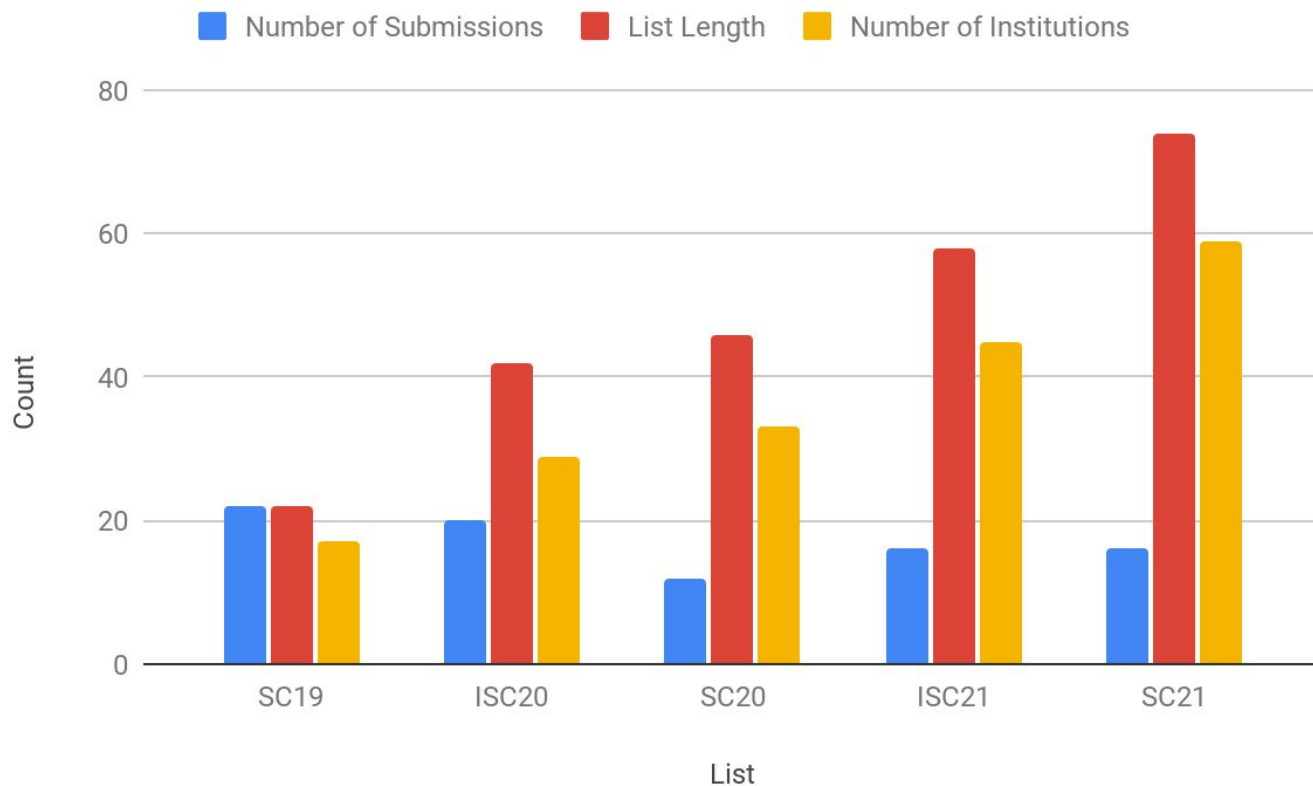
# Lists and Analysis

**10<sup>500</sup>**



# Growth in Length and Institutions

## IO500 List

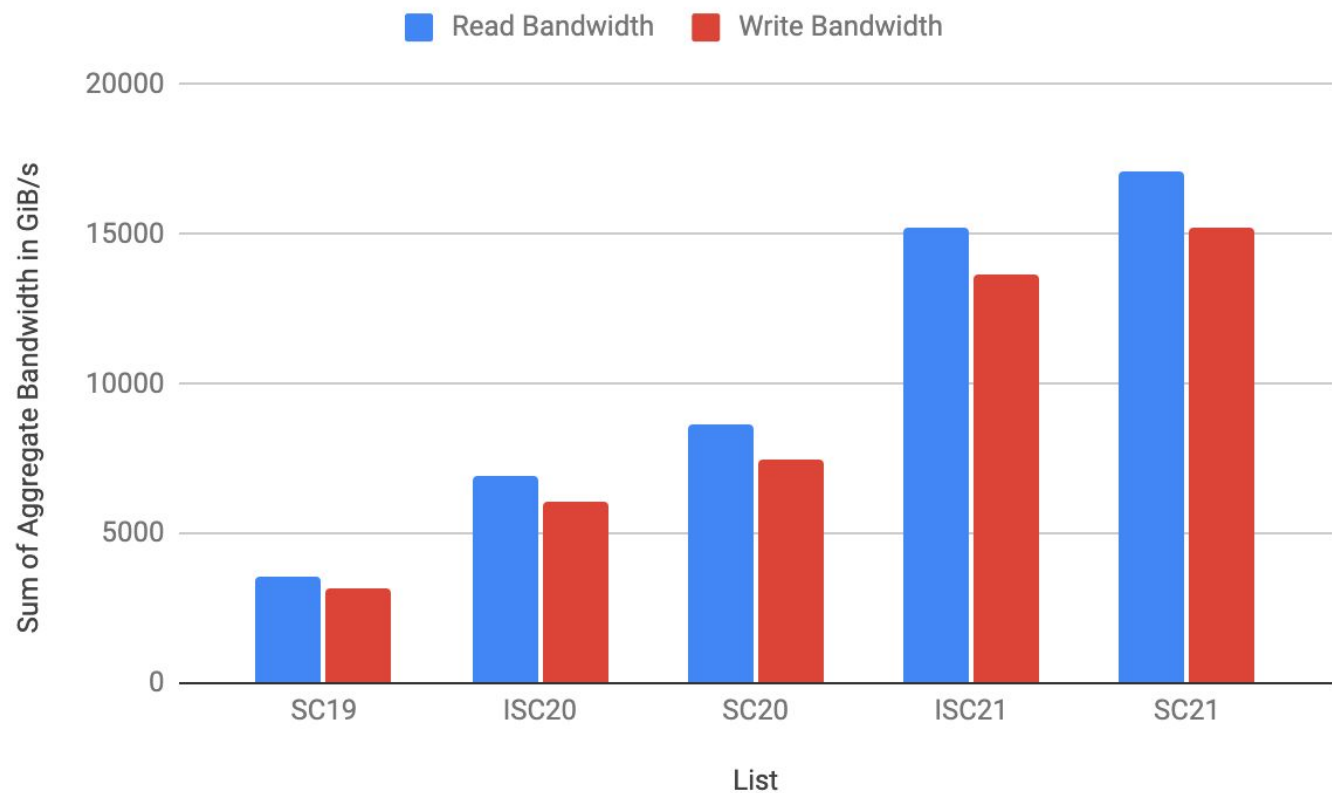


### SC21

- 16 submissions
- 74 list entries
- 59 institutions

# Total Bandwidth

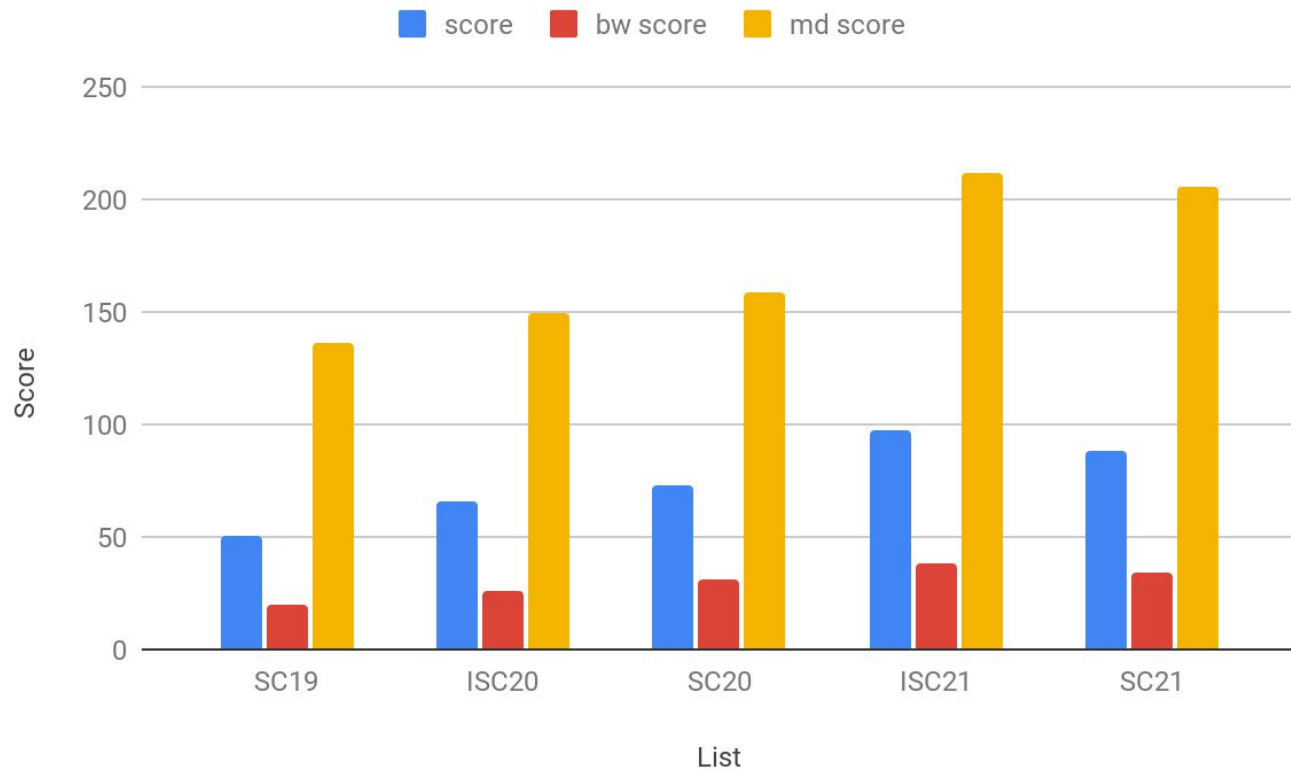
## IO500 List



# Median Scores

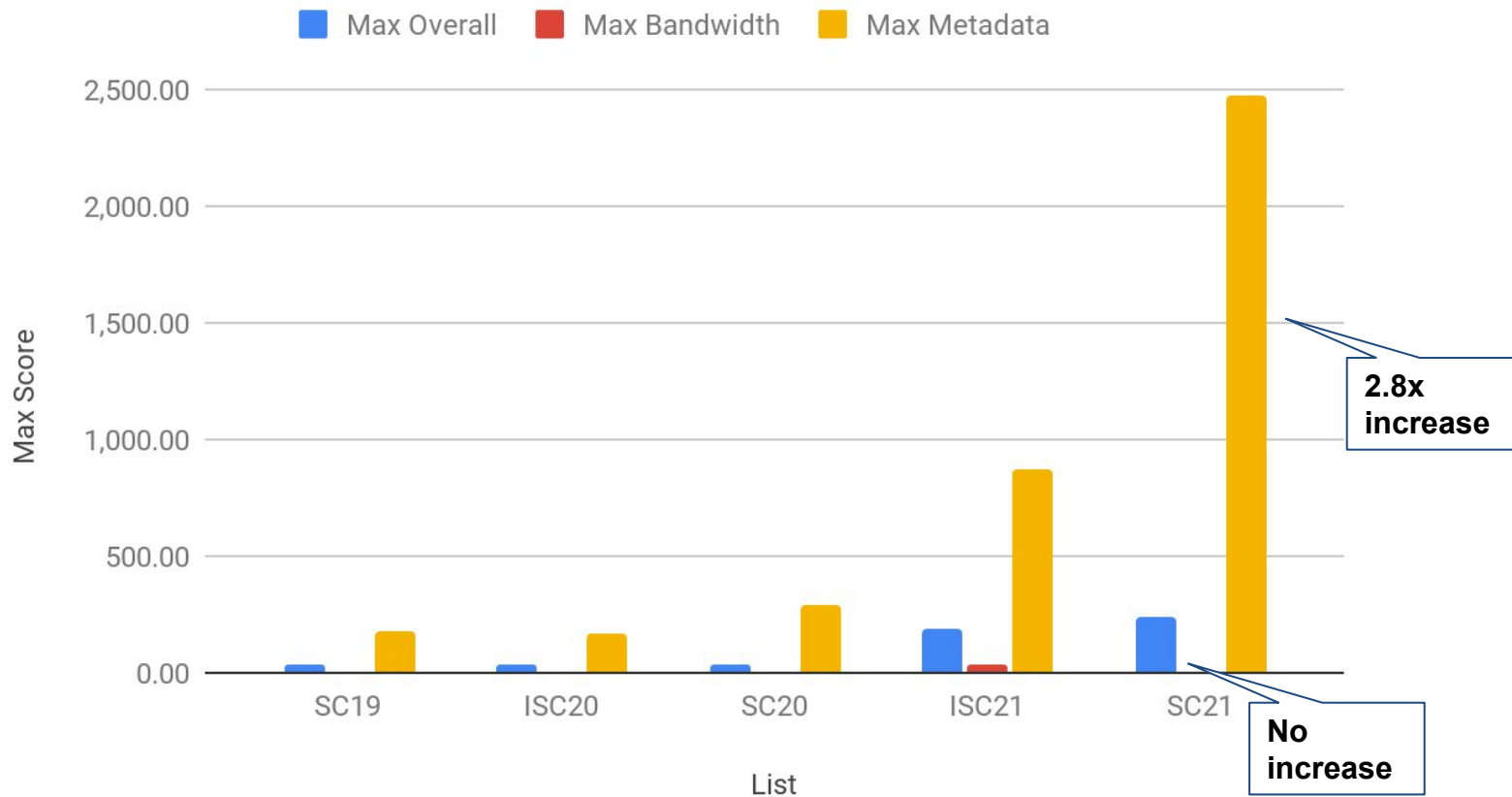
## IO500 List

Slight decreases of  
3-10% for overall,  
bandwidth, and  
metadata scores  
from ISC21 to SC21



# Growth in Max Score per Client

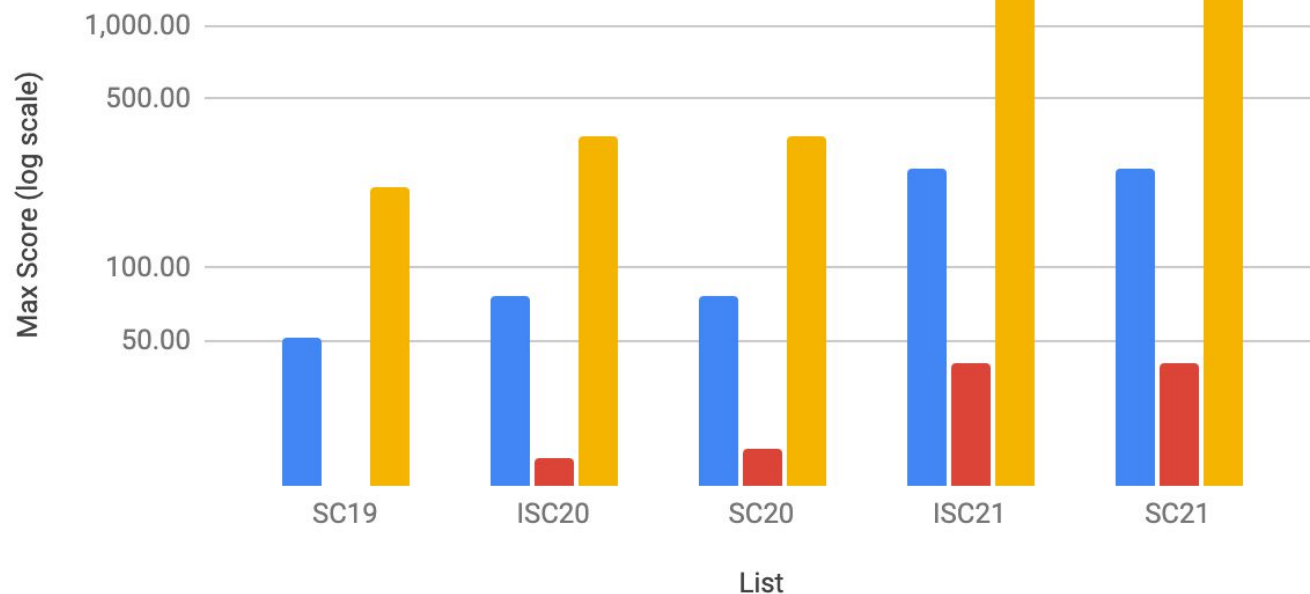
## IO500 List



# Growth in Max Scores per Client

## IO500 - 10-Node Challenge List

Max Overall Max Bandwidth Max Metadata



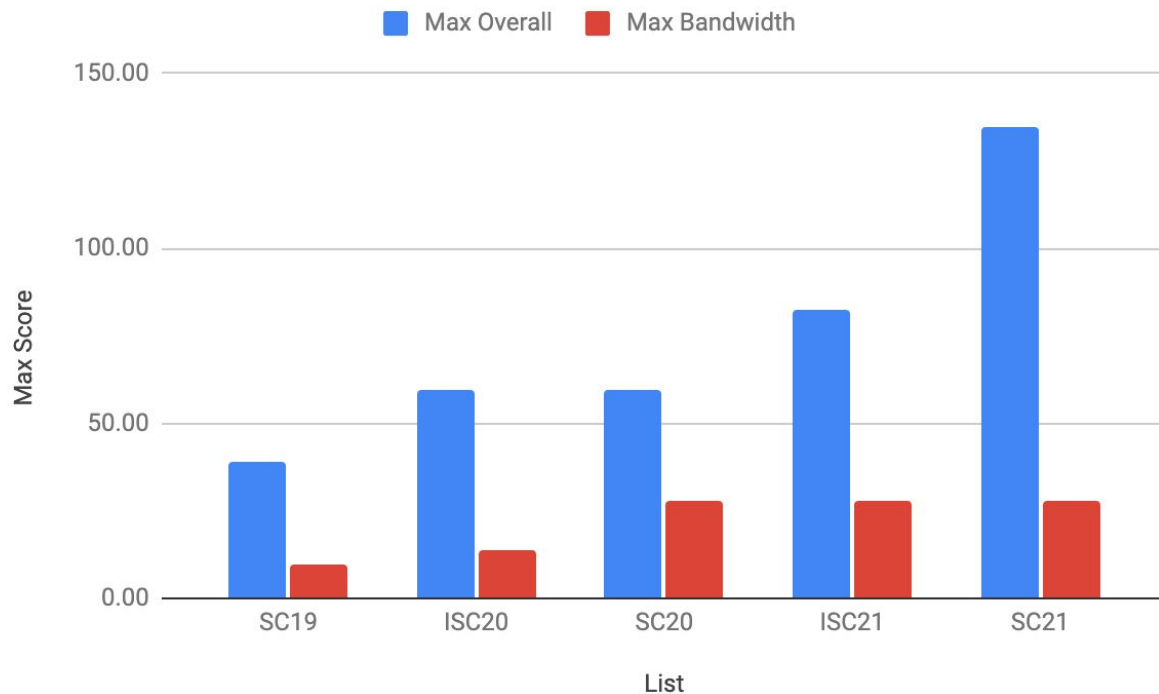
No change from  
ISC21 to SC21

# Growth in Max Score per Storage Server

## IO500 - List

Per-client scores are growing by orders of magnitude

Per-storage server scores are growing slower, with bandwidth flat for 3 lists



Note: metadata score per server growth reflected in overall score

# Award Ceremony

**10<sup>500</sup>**

# Six Awards

---

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



# 10 node challenge - Bandwidth Winner

## 10 Node SC21 List

IO500

10 Node

Sorted by BW

This is the SC21 IO500 10-nodes list

#	INFORMATION							SCORE	IO500	
	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.		BW ↑ (GiB/s)	MD (KIOP/s)
1	ISC21	Intel	Endeavour	Intel	DAOS	10	1,440		398.77	
2	SC21	Olympus Lab	OceanStor Pacific	Huawei	OceanFS	10	1,720		317.07	
3	SC21	Huawei HPDA Lab	Athena	Huawei	OceanFS	10	1,720		314.56	
4	ISC21	Pengcheng Laboratory	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng	MadFS	10	1,800		193.77	
5	SC20	Forschungszentrum Juelich (FZJ)	JUWELS	HPEDDN	IME	10	400		178.11	
6	SC20	TACC	Frontera	DDN	IME	10	280		176.23	
7	ISC20	Intel	Wolf	Intel	DAOS	10	420		164.77	
8	ISC21	Supermicro		Supermicro	DAOS	10	1,120		112.17	
9	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	10	960		105.28	
10	SC21	QCT DevCloud	QCT DevCloud	QCT	DAOS	10	560		102.85	

# Certificate

IO500 Performance Certification

This Certificate is awarded to:

**Intel (Endeavour)**

#1 in the 10 Node Challenge BW Score

**IO**500



**Nov 2021**

*IO500 Steering Board*

<https://io500.org/list/sc21/ten>

IO  
IO  
IO  
IO  
IO  
IO  
IO  
IO  
IO

IO  
IO  
IO  
IO  
IO  
IO  
IO  
IO  
IO

# 10-Node Challenge - Metadata Winner

## 10 Node SC21 List

IO500

10 Node

Sorted by MD

This is the SC21 IO500 10-nodes list

#	INFORMATION							IO500	
	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE	MD ↑ (KIOP/S)
1	ISC21	Pengcheng Laboratory	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng	MadFS	10	1,800		34,777.27
2	SC21	Huawei HPDA Lab	Athena	Huawei	OceanFS	10	1,720		18,235.71
3	SC21	Olympus Lab	OceanStor Pacific	Huawei	OceanFS	10	1,720		16,664.88
4	SC21	BPFS Lab	Kongming		BPFS	10	800		9,827.09
5	ISC21	Intel	Endeavour	Intel	DAOS	10	1,440		8,671.65
6	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	10	960		3,567.85
7	ISC20	Intel	Wolf	Intel	DAOS	10	420		3,493.56
8	ISC20	TACC	Frontera	Intel	DAOS	10	420		3,271.49
9	SC21	NRCTM	ASTRA	NRCTM	DAOS	10	360		2,984.61
10	ISC21	National Supercomputer Center in GuangZhou	Venus2	National Supercomputer Center in GuangZhou	kapok	10	480		2,452.87



# Certificate

IO500 Performance Certification

This Certificate is awarded to:

**Pengcheng Laboratory (Cloudbrain-II)**  
#1 in the 10 Node Challenge MD Score

**IO500**



**Nov 2021**

*IO500 Steering Board*

<https://io500.org/list/sc21/ten>



# 10-Node Challenge - Winner

## 10 Node SC21 List

IO500

10 Node

Sorted by score

This is the SC21 IO500 10-nodes list

# ↑	INFORMATION								IO500	
	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE ↑	BW	MD
									(GIB/S)	(KIOP/S)
1	ISC21	Pengcheng Laboratory	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng	MadFS	10	1,800	2,595.89	193.77	34,777.27
2	SC21	Huawei HPDA Lab	Athena	Huawei	OceanFS	10	1,720	2,395.03	314.56	18,235.71
3	SC21	Olympus Lab	OceanStor Pacific	Huawei	OceanFS	10	1,720	2,298.69	317.07	16,664.88
4	ISC21	Intel	Endeavour	Intel	DAOS	10	1,440	1,859.56	398.77	8,671.65
5	SC21	BPFS Lab	Kongming		BPFS	10	800	972.60	96.26	9,827.09
6	ISC20	Intel	Wolf	Intel	DAOS	10	420	758.71	164.77	3,493.56
7	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	10	960	612.87	105.28	3,567.85
8	SC21	NRCTM	ASTRA	NRCTM	DAOS	10	360	511.02	87.50	2,984.61
9	ISC20	TACC	Frontera	Intel	DAOS	10	420	508.88	79.16	3,271.49
10	ISC21	National Supercomputer Center in GuangZhou	Venus2	National Supercomputer Center in GuangZhou	kapok	10	480	474.10	91.64	2,452.87

IO  
IO  
IO  
IO  
IO  
IO  
IO  
IO  
IO

# Certificate

IO500 Performance Certification

This Certificate is awarded to:

**Pengcheng Laboratory (Cloudbrain-II)**  
#1 in the 10 Node Challenge

IO 500



Nov 2021

*IO500 Steering Board*

<https://io500.org/list/sc21/ten>

IO  
IO  
IO  
IO  
IO  
IO  
IO  
IO  
IO

# Full list - Bandwidth Winner

## IO500 SC21 List

IO500

10 Node

Sorted by BW

This is the SC21 IO500 list

#	INFORMATION							IO500		
	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE	BW ↑	MD
									(GIB/S)	(KIOP/S)
1	ISC21	Pengcheng Laboratory	Pengcheng Cloudbrain-II on Atlas 900	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	Endeavour	Intel	DAOS	10	1,440		398.77	
5	ISC20	Intel	Wolf	Intel	DAOS	52	1,664		371.67	
6	SC21	Olympus Lab	OceanStor Pacific	Huawei	OceanFS	10	1,720		317.07	
7	SC21	Huawei HPDA Lab	Athena	Huawei	OceanFS	10	1,720		314.56	
8	ISC20	Oracle Cloud Infrastructure	BeeGFS on Oracle Cloud	Oracle Cloud Infrastructure	BeeGFS	270	3,240		293.05	
9	ISC21	Google Cloud	Google	DDN	Lustre	1,000	5,000		282.78	
10	SC19	National Supercomputing Center in Changsha	Tianhe-2E	National University of Defense Technology	Lustre	480	5,280		209.43	



# Certificate

IO500 Performance Certification

This Certificate is awarded to:

**Pengcheng Laboratory (Cloudbrain-II)**  
#1 in the IO500 BW Score

**IO 500**



**Nov 2021**

*IO500 Steering Board*

<https://io500.org/list/sc21/io500>





# Full list - Metadata Winner

## IO500 SC21 List

IO500

10 Node

Sorted by MD

This is the SC21 IO500 list

#	INFORMATION							IO500	
	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE	MD ↑ (KIOP/S)
1	ISC21	Pengcheng Laboratory	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng	MadFS	512	36,864		396,872.82
2	SC21	Huawei Cloud		PDSL	Flashfs	15	1,560		37,034.00
3	SC21	Huawei HPDA Lab	Athena	Huawei	OceanFS	10	1,720		18,235.71
4	SC21	Olympus Lab	OceanStor Pacific	Huawei	OceanFS	10	1,720		16,664.88
5	SC21	BPFS Lab	Kongming		BPFS	10	800		9,827.09
6	ISC21	Intel	Endeavour	Intel	DAOS	10	1,440		8,671.65
7	ISC20	Intel	Wolf	Intel	DAOS	52	1,664		8,649.57
8	ISC20	TACC	Frontera	Intel	DAOS	60	1,440		7,449.56
9	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	36	3,456		5,545.61
10	SC19	WekaIO	WekaIO on AWS	WekaIO	WekaIO Matrix	345	8,625		5,045.33



# Certificate

IO500 Performance Certification

This Certificate is awarded to:

**Pengcheng Laboratory (Cloudbrain-II)**  
#1 in the IO500 MD Score

**IO500**



**Nov 2021**

*IO500 Steering Board*

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



# Full list - Winner

## IO500 SC21 List

IO500

10 Node

This is the SC21 IO500 list

# ↑	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 Cloudbrain-II on Atlas 900	Pengcheng	MadFS	512	36,864	36,850.37	3,421.62	396,872.82
2	SC21	Huawei HPDA Lab	Athena	Huawei	OceanFS	10	1,720	2,395.03	314.56	18,235.71
3	SC21	Olympus Lab	OceanStor Pacific	Huawei	OceanFS	10	1,720	2,298.69	317.07	16,664.88
4	SC21	Huawei Cloud		PDSL	Flashfs	15	1,560	2,016.70	109.82	37,034.00
5	ISC21	Intel	Endeavour	Intel	DAOS	10	1,440	1,859.56	398.77	8,671.65
6	ISC20	Intel	Wolf	Intel	DAOS	52	1,664	1,792.98	371.67	8,649.57
7	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	36	3,456	988.99	176.37	5,545.61
8	SC21	BPFS Lab	Kongming		BPFS	10	800	972.60	96.26	9,827.09
9	SC19	WekaIO	WekaIO on AWS	WekaIO	WekaIO Matrix	345	8,625	938.95	174.74	5,045.33
10	ISC20	TACC	Frontera	Intel	DAOS	60	1,440	763.80	78.31	7,449.56



# Certificate

IO500 Performance Certification

This Certificate is awarded to:

**Pengcheng Laboratory (Cloudbrain-II)**  
#1 in the IO500

**IO**500



**Nov 2021**

*IO500 Steering Board*

<https://io500.org/list/sc21/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	Pengcheng Cloudbrain-II	MadFS	34777.27	kIOPS
	<b>Overall</b>	Pengcheng Cloudbrain-II	MadFS	<b>2595.89</b>	<b>score</b>
IO500	Bandwidth	Pengcheng Cloudbrain-II	MadFS	3421.62	GiB/s
	Metadata	Pengcheng Cloudbrain-II	MadFS	396872.82	kIOPS
	<b>Overall</b>	Pengcheng Cloudbrain-II	MadFS	<b>36850.37</b>	<b>score</b>

# Community Presentation

**10<sup>500</sup>**

# Exploring More Ways to Use IO500 Benchmark & List

Radita Liem<sup>\*</sup>, Julian Kunkel<sup>‡</sup>, Jay Lofstead<sup>†</sup>

<sup>\*</sup>Chair for High Performance Computing, IT Center, RWTH Aachen University

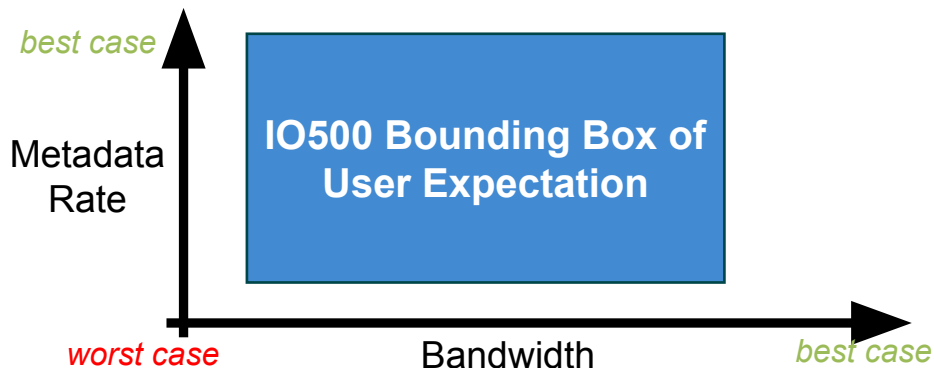
<sup>‡</sup>Göttingen University/GDWG

<sup>†</sup>Sandia National Laboratories



## IO500 Benchmark Usage

- IO500 benchmark's mdtest and IOR scenario can be used to form a bounding box of user expectations<sup>4</sup> as illustrated by the figure below



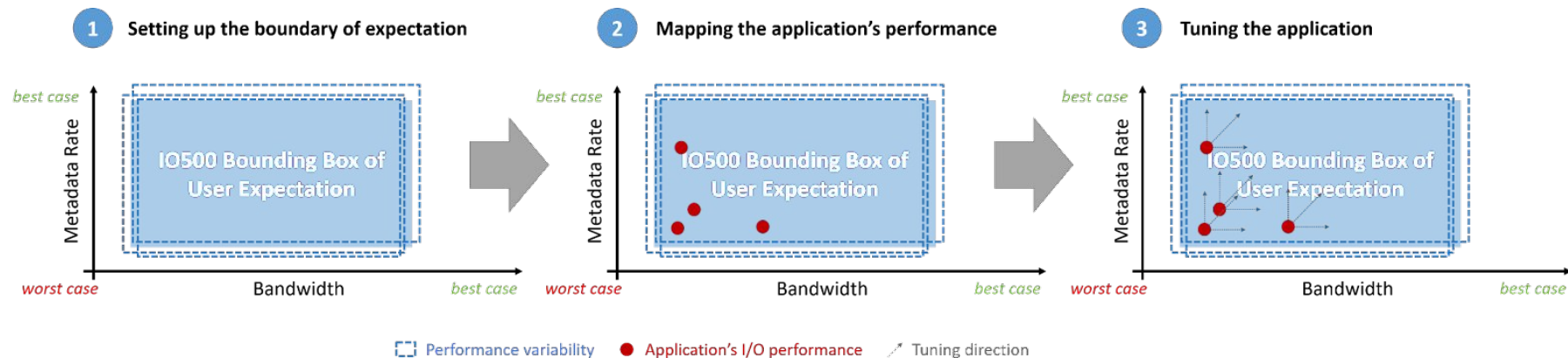
IO<sup>500</sup>

- Worst case scenario** is coming from IOR and mdtest 'hard' scenario
- Best case scenario** is coming from IOR and mdtest 'easy' scenario
- 'Find' is not used in this bounding box model since it is not as controlled as IOR and mdtest and will skew the IO500 numbers

<sup>4</sup> A. Dilger, "IO500 | A storage Benchmark for HPC", 2019. [Online]. Available: [https://wiki.lustre.org/images/9/92/LUG2019-IO500\\_Storage\\_Benchmark\\_for\\_HPC-Dilger.pdf](https://wiki.lustre.org/images/9/92/LUG2019-IO500_Storage_Benchmark_for_HPC-Dilger.pdf). [Accessed: 02-Mar-2021]

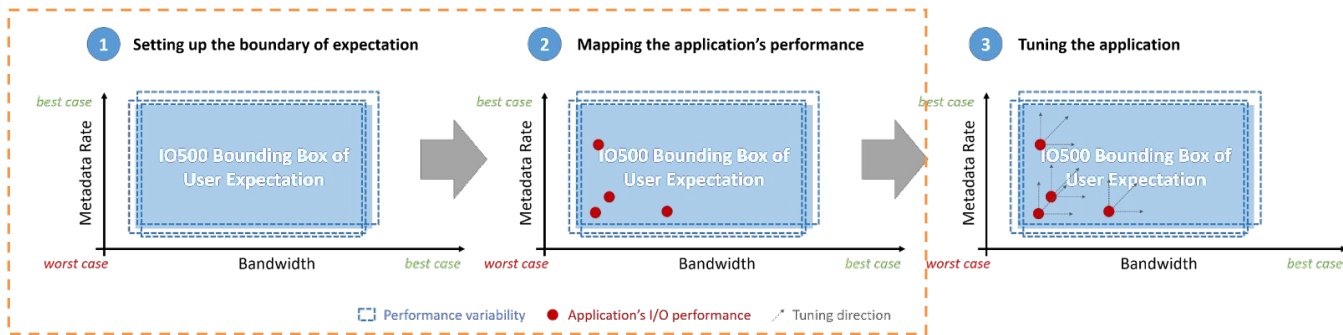


# IO500-based Workflow Proposed



## Proof of Concept Experiment Setup

- The experimentation in this work covers **the first and second step of the workflow**. The second step of the work flow is still in our **exploratory stage** and the third step is for the future work

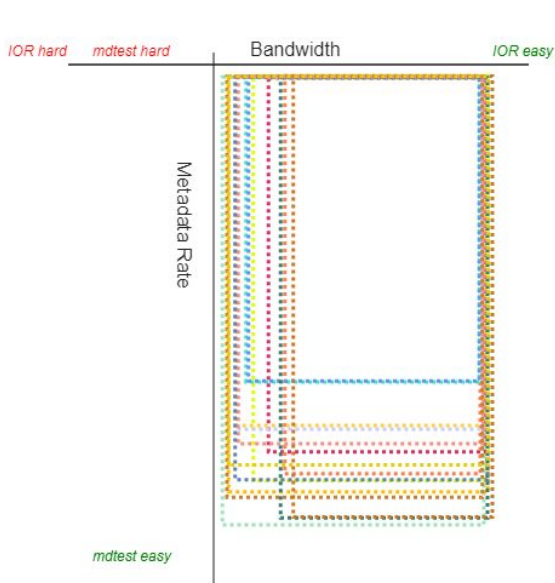


- Experiment environment:**

- CLAIIX-2018 cluster at RWTH Aachen University (48 cores Intel Skylake, 384 GB memory), 40 Gb/s Ethernet.
- 4 nodes BeeGFS Filesystem, each with 480 GB SSD.
- IO500 benchmark - SC20 submission version.
- NAS Parallel Benchmark – BTIO “full” class A,B, and C on 4,9, and 16 processes

## Results: Forming Bounding Box of User Expectation [1]

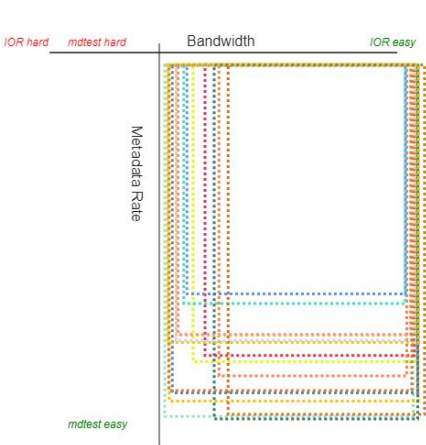
- Bounding box of **POSIX** API, each square represents individual run from the same IO configuration



	IOR hard (GiB/s)	IOR easy (GiB/s)	mdtest hard (KIOPS)	mdtest easy (KIOPS)
	0.85	1.88	10.97	145.17
	1.03	1.89	11.26	123.26
	1.10	1.87	10.59	129.86
	0.87	1.90	10.76	135.27
	0.97	1.90	10.64	131.93
	0.94	1.86	11.06	102.35
	0.95	1.86	10.84	102.06
	0.90	1.89	10.98	116.54
	0.90	1.89	11.16	115.40
	1.08	1.90	11.14	143.09
	0.91	1.88	10.80	120.86
	0.90	1.90	11.05	131.65

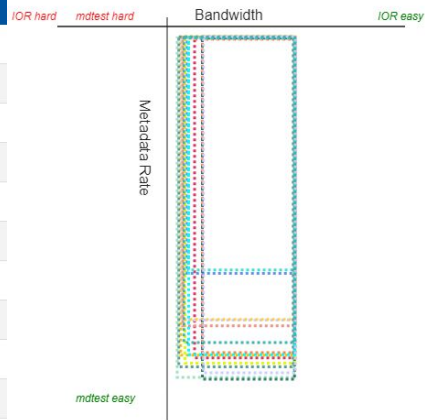
## Results: Forming Bounding Box of User Expectation [2]

- Bounding box of **POSIX** API, read and write show different pattern



	IOR hard (GiB/s)	IOR easy (GiB/s)	mdtest hard (KIOPS)	mdtest easy (KIOPS)
■	0.52	1.75	2.60	76.12
■	0.72	1.76	2.56	63.37
■	0.79	1.72	2.55	67.64
■	0.55	1.78	2.62	72.92
■	0.66	1.77	2.56	64.69
■	0.62	1.71	2.57	52.50
■	0.63	1.71	2.56	50.49
■	0.58	1.77	2.54	60.23
■	0.58	1.76	2.50	59.20
■	0.77	1.77	2.64	76.67

Bounding box from POSIX **write** result



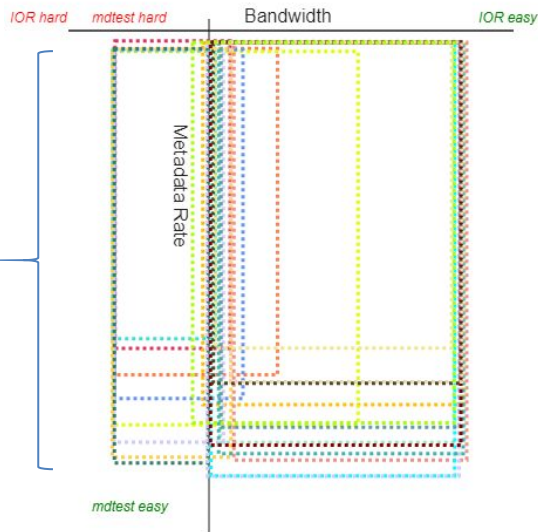
	IOR hard (GiB/s)	IOR easy (GiB/s)	mdtest hard (KIOPS)	mdtest easy (KIOPS)
■	1.37	2.02	24.80	360.31
■	1.47	2.03	26.02	341.16
■	1.51	2.03	23.60	335.88
■	1.39	2.03	24.04	338.53
■	1.42	2.03	23.63	346.27
■	1.44	2.03	24.96	254.26
■	1.43	2.03	24.70	257.68
■	1.40	2.03	25.13	303.84
■	1.41	2.02	25.57	303.20
■	1.51	2.02	25.22	362.23

Bounding box from POSIX **read** result

## Results: Anomalous Bounding Box

- Anomalous result in **MPI-IO** API: IOR 'Easy' score gets lower number than IOR 'hard'
- Broken node is most likely the reason behind these anomalous result

Bounding box skewed  
to the direction of IOR 'hard'

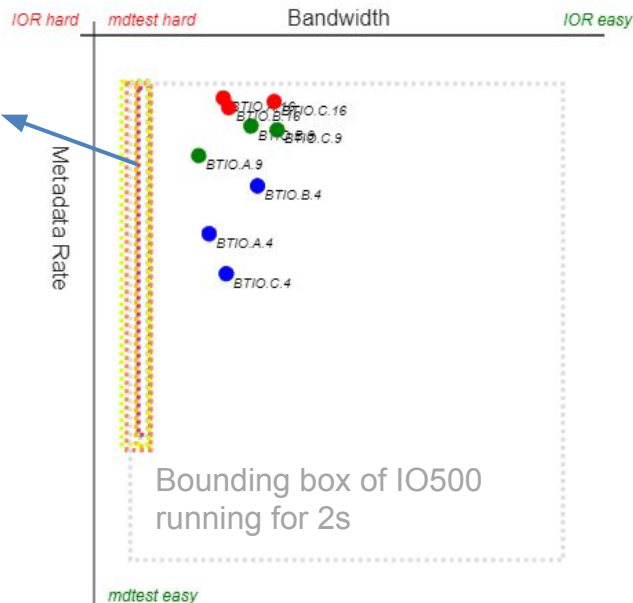


	IOR hard (GiB/s)	IOR easy (GiB/s)	mdtest hard (KIOPS)	mdtest easy (KIOPS)
	0.90	1.89	10.74	135.79
	0.94	0.47	10.50	106.45
	1.14	0.47	12.86	114.73
	0.83	1.89	11.12	124.06
	1.46	0.47	13.78	130.29
	0.88	0.47	13.50	103.47
	0.99	0.47	13.24	122.10
	0.91	0.47	13.04	135.73
	0.95	0.46	13.10	140.41
	0.85	0.47	13.02	142.20
	0.96	1.90	11.00	141.28
	0.86	1.85	10.81	146.24

## Results: Exploration on the I/O Performance Mapping [1]

- Exploration with BTIO shows the application's performance falls within the box for **MPI-IO API with cache effect**

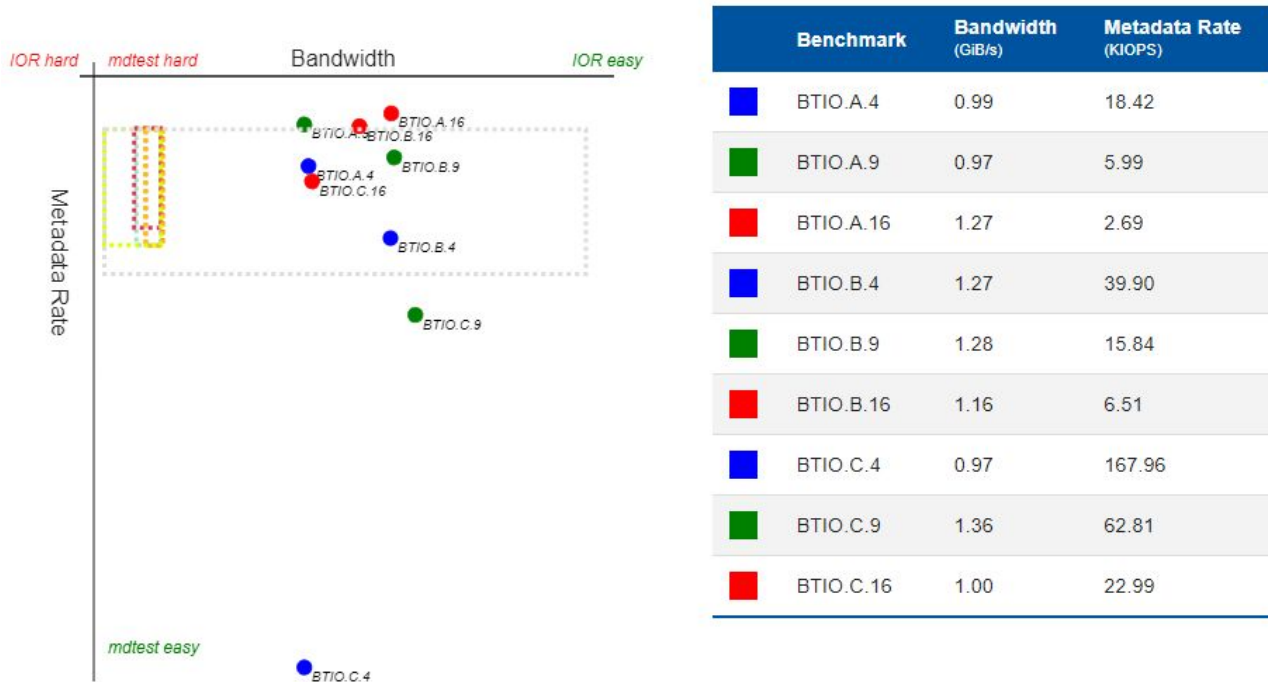
Bounding boxes of IO500 running with default setup (300s)



	Benchmark	Bandwidth (GiB/s)	Metadata Rate (KIOPS)
■	BTIO.A.4	0.68	21.67
■	BTIO.A.9	0.64	14.18
■	BTIO.A.16	0.73	8.66
■	BTIO.B.4	0.85	17.08
■	BTIO.B.9	0.82	11.35
■	BTIO.B.16	0.75	9.58
■	BTIO.C.4	0.74	25.50
■	BTIO.C.9	0.91	11.71
■	BTIO.C.16	0.90	9.00

## Results: Exploration on the I/O Performance Mapping [2]

- However in the **POSIX API**, metadata rate calculation falls outside the cache box



This project is currently displayed in: <https://bit.ly/3BhhAFZ>

## Next Project: Using IO500 List for Modelling & Performance Engineering

- Interesting form of bounding box of user expectation from the top 2 of the list
- IOR hard performs better than the IOR easy in MadFS. Potential improvement?

Tag	IOR hard (GiB/s)	IOR easy (GiB/s)	mdtest hard (KIOPS)	mdtest easy (KIOPS)
ISC21.Pengcheng-Laboratory.MadFS	205.33	182.85	16380.06	26569.90
ISC21.Intel.DAOS	282.97	561.94	8081.06	10934.75
ISC20.Intel.DAOS	139.37	194.80	3038.19	3959.73
ISC21.Lenovo.DAOS	100.11	110.71	2767.18	4457.22
ISC20.TACC.DAOS	74.49	84.13	3023.32	4064.34
ISC21.National-Supercomputer-Center-in-GuangZhou.kapok	90.63	92.65	1831.38	2503.70
ISC20.Argonne-National-Laboratory.DAOS	69.21	132.61	1568.07	2786.25
ISC21.Supermicro.DAOS	92.44	136.12	815.94	3061.71
SC19.NVIDIA.Lustre	27.06	279.59	574.07	919.45
SC20.EPCC.GekkoFS	27.95	75.02	669.01	1824.42



# Thank you!

---

Inquiry and question: **Radita Liem** ([liem@itc.rwth-aachen.de](mailto:liem@itc.rwth-aachen.de))



UNIVERSITY OF GÖTTINGEN  
GERMANY



Sandia  
National  
Laboratories



High  
Performance  
Computing



IT Center

**RWTH**AACHEN  
UNIVERSITY

# Roadmap

**10** 500

# Motivation: IO500 User Survey

---

- 55 survey participants (max one per org)
  - 51% used benchmark, but didn't submit results to list
  - 16% may submit results in future
- What would increase submissions:
  - Requested by end-users (45%)
  - Splitting of the list (35%)
  - Better instructions (33%)
  - Easier installation or usage (31%)
  - Include new/different access patterns (27%)
  - Provision of dedicated system (27%)

# IO500 Survey

---

- What was the key value of IO500?
  - Build database of storage system metrics (69%)
  - Encourage storage vendors/devs to improve (65%)
  - Help storage purchasing decisions (42%)
- What should be criterion for splitting the list?
  - Production vs. non-production (54%)
  - Vendor vs. end-user (45%)
  - POSIX vs. non-POSIX (44%)
  - Type of back-end storage (38%)
  - Shared on-premise vs. cloud (36%)

# IO500 Survey

---

- What defines a production-level storage system?
  - System for production apps (91%)
  - System exists for reasons beyond benchmarking (82%)
  - System for long-term usage (69%)
  - System provides data redundancy (65%)
  - System available to/used by end users (65%)
  - Software/Hardware available to general public/can purchase (61%)
  - Planned to operate for longer than a year (41%)

# IO500 Survey

---

- 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%)
- Reproducibility is very important to the community
  - On scale 1-5, 50% selected 5, 30% selected 4

# Roadmap for the IO500

---

- Proposal for reproducibility and list-splitting
- Fill in gaps in IO500 to improve usage patterns
  - Collect and evaluate results for new benchmark phases
    - Not officially part of benchmark yet, still some flexibility to modify
  - Document rationales for existing/new benchmark phases
- Improvements to system schema for filesystem types
  - Improve system-level data from submitters, uniformity of data collected
  - Some schemas exist, continue to improve with more use and feedback
  - Extending the scripts to automatically collect system metadata
- New io500.org site for submissions for next list - thanks Jean Luca
  - More system metadata fields will be mandatory for better comparisons

# Roadmap for the IO500

---

- ISC 22 Roadmap
  - Call for submission: March
  - Testing phase ends: April ~15th
    - Code freeze, but please test before!
  - Submission deadline: May ~15th
  - ISC Release: May ~29th

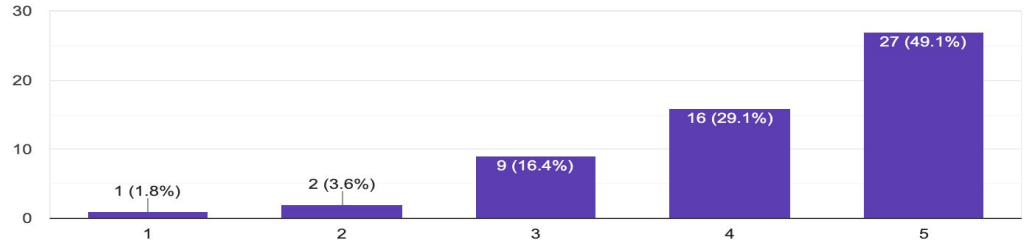


# Reproducibility & List Split

**10<sup>500</sup>**

# IO500 Reproducibility

- 78.2% rated reproducibility as important or very important (4 or 5 out of 5)



- What are we trying to do
  - Improve the transparency (and data quality) of each submission
    - Ensure enough information for someone with enough resources to reproduce IO500 execution
  - Strengthen trust among HPC users in IO500 real-world value
  - Strengthen confidence in the community of a fair playing field for all existing/potential submissions
- What are we NOT trying to do
  - Make every submission fully reproducible
  - Mandate open-source file systems
  - Force access to systems or include legal framework
    - Good faith integrity of participants continues to apply

# IO500 Reproducibility

---

- Rough Proposal Outline

- Improve metadata collection
  - Reduce ambiguity
  - Expand to all core components required to reproduce (e.g., HW, software)
- Provide open access to all custom scripts/tunings
  - e.g., find, file system tunings, setup instructions, etc.
- Mandatory questionnaire
  - e.g., durability type, client/storage config, client API, results integrity steps

- Next Steps

- Define how users share scripts and extra information
  - e.g., via IO500 website, personal github
- Create initial questionnaire as optional for ISC22
- Extend/Improve the schema to cover relevant metadata
- Once ready, make additional information mandatory (hopefully for SC22)

# IO500 Sub Categories

---

- 85.5% believe the list should be split into sub-categories
  - Each list having equal stature
- Top proposal (>55%) to define a “Production” sub-category
  - Many opinions on definition of production
- What defines “Production” could include
  - Require more stringent “Reproducibility” requirements
    - e.g., details of support node/device failure, commercial availability
  - Satisfy XX% of production definition (e.g., data redundancy, real applications, available > 1 year)
  - System actively/will support applications that generate data with business/scientific value
    - Computer science research that read or generate fake data would not meet this criteria
- Next Steps
  - Define “Production” category requirements by ISC22 and preview “production” category impact
    - Leverage mandatory reproducibility initiative
    - Optimally, every submission includes the required data to end on the preview production category
  - Create “Production” category for SC22

# Benchmark Phases and Extended Access Patterns

**10500**

# Benchmark Phases and Extended Access Patterns

---

- Extended mode with extra phases
  - We had two submissions for SC21 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 dual submission for ISC22 to get experience
  - Standard run + extended run with more benchmark
  - 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

**10<sup>500</sup>**

# Supplementary Presentations

---

Due to time constraints, additional presentations are on our BoF page:

<https://io500.org/pages/bof-sc21>

- [The Virtual Institute for I/O](#)
  - *Julian Kunkel*



# Open Floor

---

- Questions from zoom
  - Is the file system a factor or is it the storage system that is more important?
  - How will you reconcile this higher bar for submission with questionnaires/list splitting/reproducibility with the ~33% of survey respondents who said they didn't submit because it was too complicated?
  - Does the utilization of in-storage compute impact the intended goals of the 10-node challenge, since number of servers is unlimited?
  - Is there still a plan to create a vendor advisory list? This was discussed on Slack ...
    - IBM
  - Should we have additional community meetings?
  - 10-node certainly biases the list away from file systems that rely on smart clients
  - as with every ranked list, 10-node is easily gamed
  - Now that it's brought up, it might be a good idea to just drop the find altogether... it is easily the most variable portion of the benchmark
  - Is it possible to setup a repository of the testing configurations that people use as a way to evolve these and speed up the testing for those starting out.