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

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

IO500

vi4io

BoF Agenda

1. **BoF intro + The Virtual Institute for IO** (5 min) – Julian Kunkel
2. **What's new with IO-500** (8 min) – George Markomanolis
3. **Community lightning talks** (5 min each)
   a. **In-node storage and memory-like I/O** — Adrian Jackson (EPCC)
   b. **Demonstrating GPUDirect Storage using the IO500** — CJ Newburn/Sven Oehme
4. **Analysis of the IO-500 data** (12 min) – John Bent
5. **Award ceremony** (5 min) — George Markomanolis, John Bent, Julian Kunkel, Jay Lofstead
6. **Roadmap for the IO-500** (5 min) – Julian Kunkel
7. **Voice of the community & Open Discussion** (15 min) – Jay Lofstead

OAK RIDGE National Laboratory

What's new with IO-500

George S. Markomanolis,

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

Denver, Colorado, SC'19

19 November 2019

U.S. DEPARTMENT OF ENERGY

Outline

- Benchmark improvements
  - MDTest Shift
  - Fix validation with IOR
- Versioning of IO500 via Tagging
  - How IO500 identifies proper versions of IOR/Mdtest
- Webpage: Info-Creator Drop-Down
- Student Cluster Competition 19

MDtest Shift Added; IOR Shift Improved

- It was intentional to implement it since the beginning of IO-500
- Each process handles data that were accessed from another process
- It doesn’t allow local caching
- It can hurt the performance significant
- In most cases, it is more realistic
- About IOR, improved shift how ranks mapped to nodes and IOR detects its mapping, no need to specify in IOR the mapping

Versioning

- IOR/MDTest
  - We were using a tagging version lately
  - Now we use a HASH again inside prepare.sh
- IO500
  - We have a tag of the io-500-dev branch for sc19, …
    - Tag always points to the latest version
    - Changelog shows details of changes
      - SC19-v1 and in Git commit message is the name as well
  - Trying to keep the instructions the same, at least per each list
  - Keep the versions of each IO500 submission through tagging
  - Needs to be improved though
  -

Info-Creator Drop-Down

- https://www.vi4io.org/io500-info-creator/

Metadata server information

Number of nodes

Number of storage devices in each metadata server

Type of the storage media in metadata servers

✓ -- select an option --
SSD
HDD
NVMe
NLSAS
SAS
SAS-SSD
PD-SSD
Other

Volatile memory capacity

Storage interface used by the servers

Network interconnect on the servers

File system software version on the servers

Operating system software version on the servers

Overhead of resilience in %

Data server information

Number of data server nodes

Number of storage devices

Type of the storage media

-- select an option --

Volatile memory capacity

Storage interface

-- select an option --

Network interconnect

✓ -- select an option --
Aries
Slingshot
Omnipath
Infiniband
IB-EDR
IB-HDR
Ethernet
Other

File system software version

Operating system software version

Overhead of resilience in %

Whatever

Comment

Type of the storage media in metadata servers

Other

Please try to use this web page to help us analyze the data

Supercomputing Student Cluster Competition 2019

- IO-500 is part of the Supercomputing Cluster Competition 2019 for extra credits!
- New stonewall rule only for the competitions (30 seconds)
- Drop cache option for single node submissions
- We show that IO is important and should be considered part of such competitions
- New list for such competitions will be announced this week
- For vendors: If IO-500 becomes part of SCC, maybe you would like to provide hardware for a team

The new IO-500 list and analysis

IO500

Reminder about Computing the Scores

- IOR easy
  - Write and read
- IOR hard
  - Write and read
- Mdtest easy
  - Create, stat, delete
- Mdtest hard
  - Create, read, stat, delete
- Namespace search
  - Find across all produced files

Geo mean -> Bandwidth score

Geo mean -> Metadata score

Geo mean -> Overall score

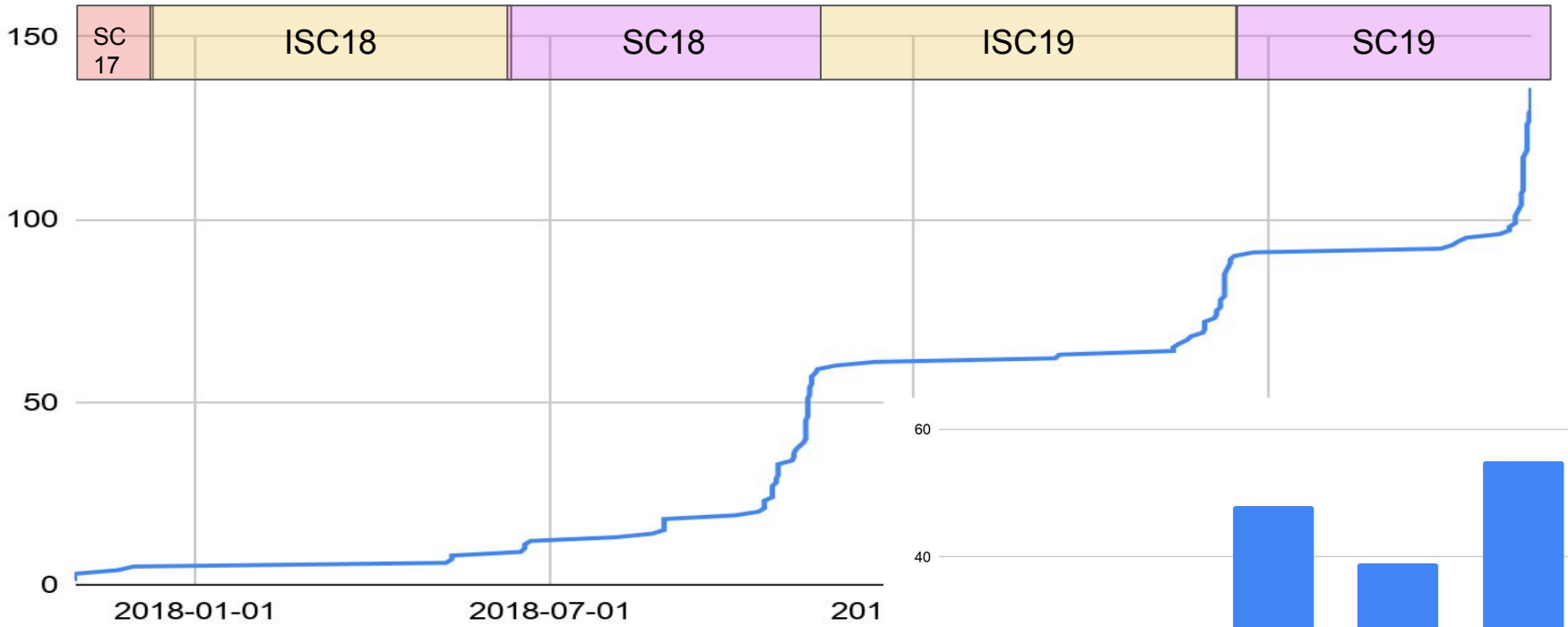
New metrics used for today's analysis:

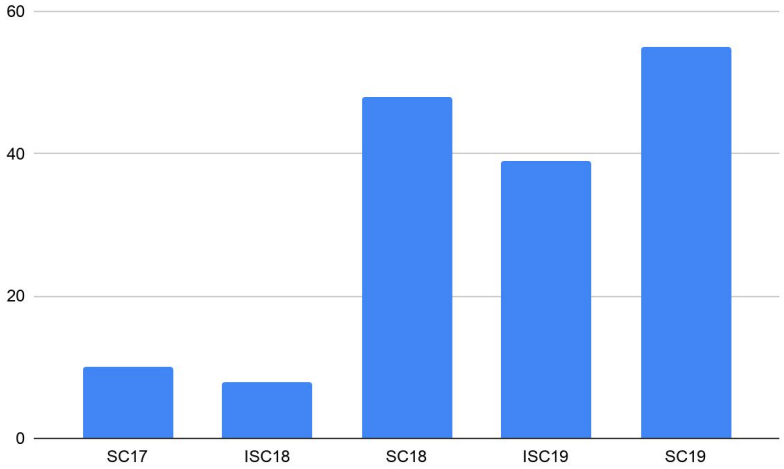
mdt_consume=geo_mean(mdt_easy_stat, mdt_easy_delete, mdt_hard_read, mdt_hard_stat,mdt_hard_delete)

mdt_produce=geo_mean(mdtest_easy_create,mdtest_hard_create)

mdt_ratio = mdt_consume / mdt_produce

Total Submissions by Date



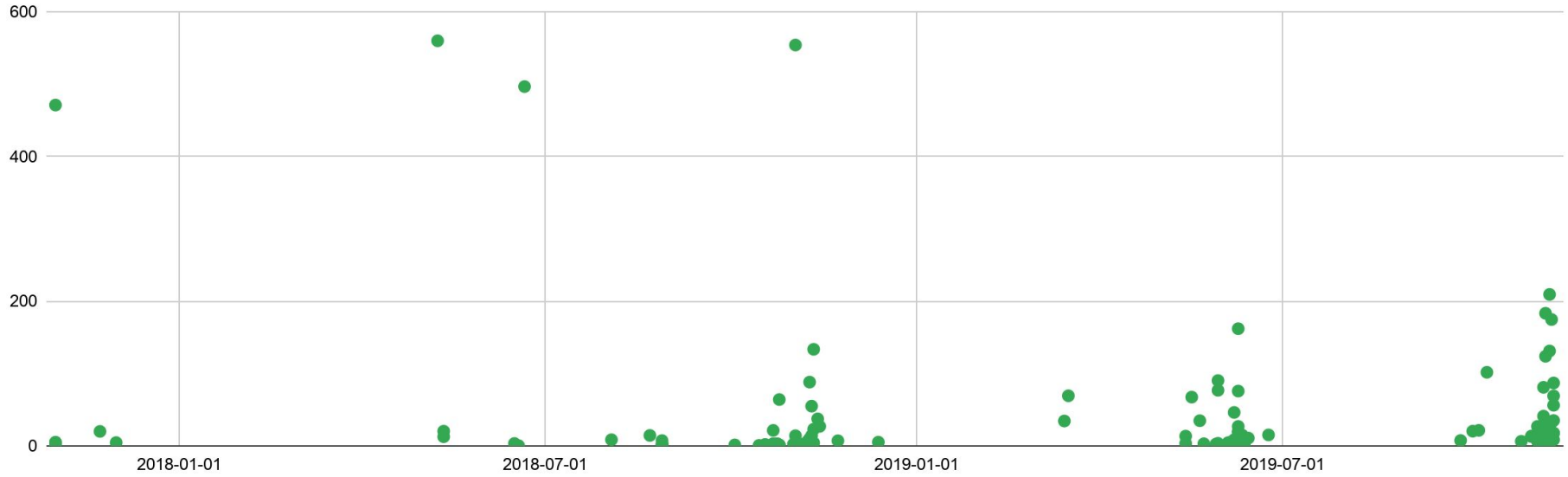


Three new file systems in SC19!

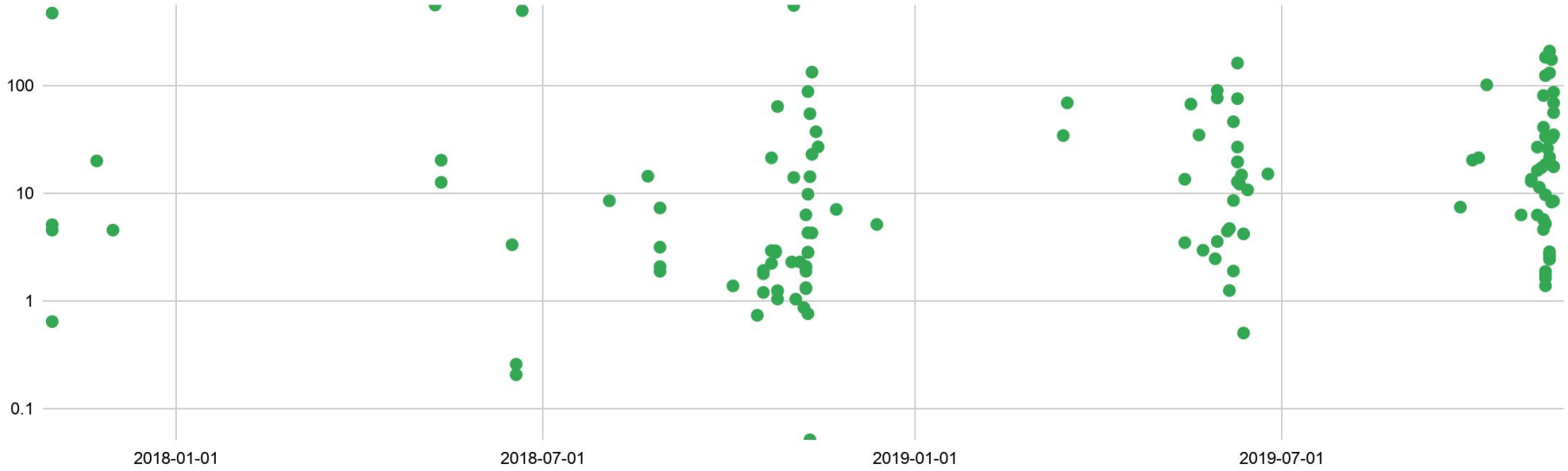
- DAOS
- GekkoFS
- YRCloudFile

Top Bandwidth by List

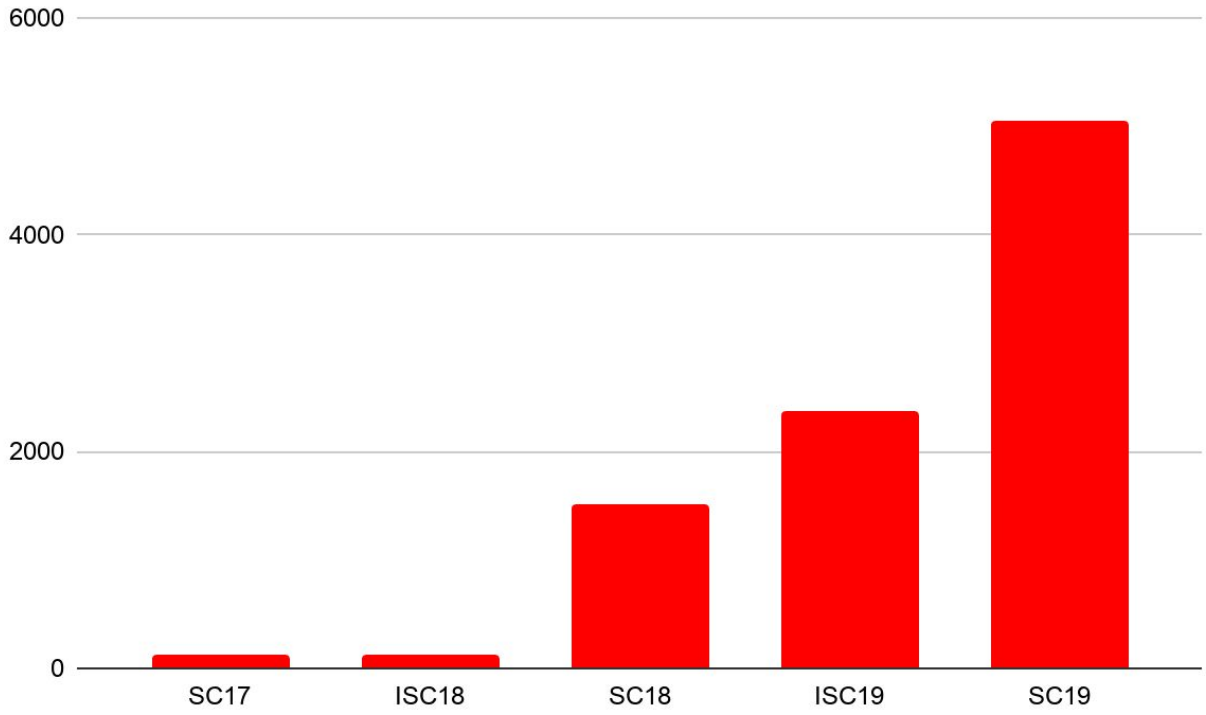
All Bandwidths by Date



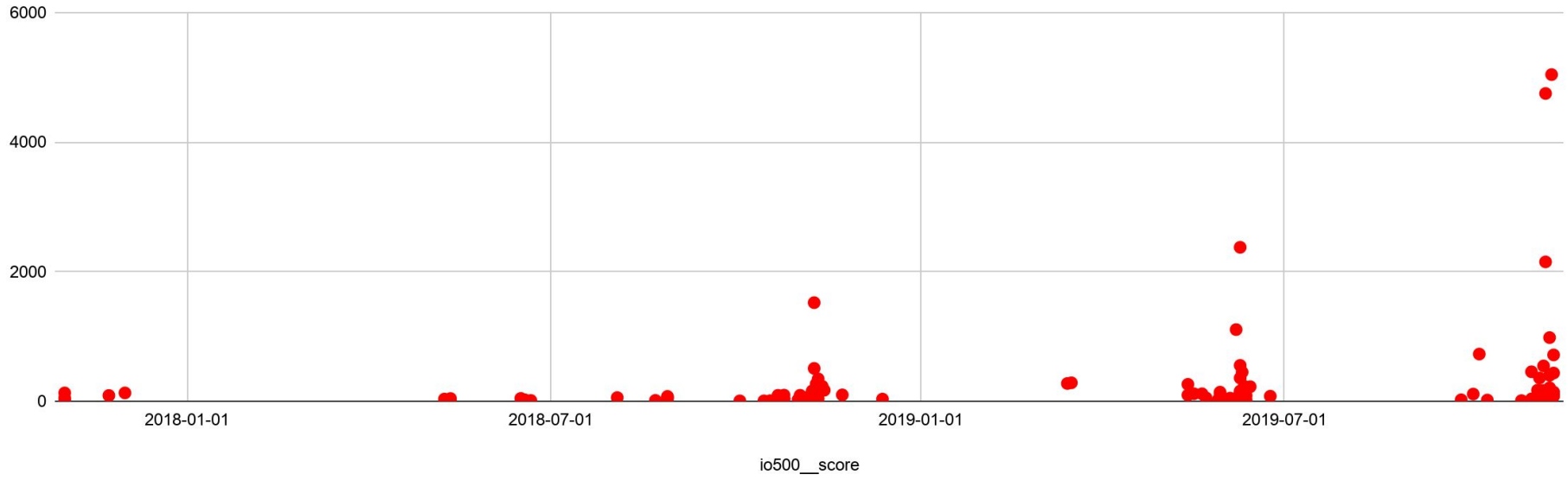
All Bandwidths by Date (log-scale)



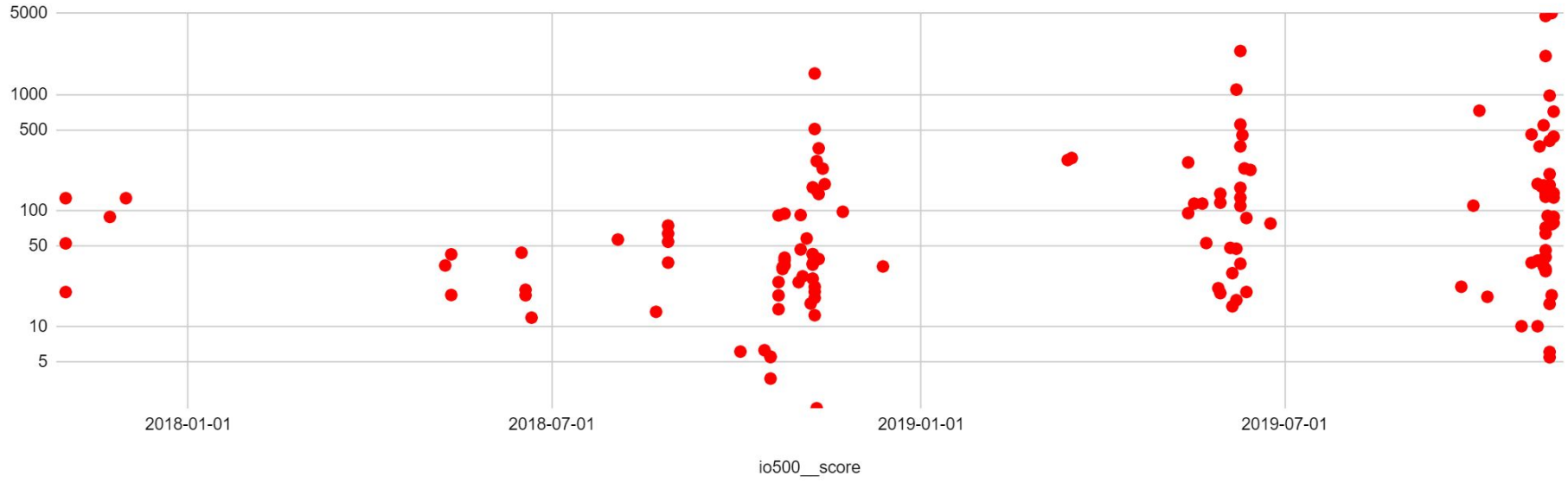
Top Metadata by List



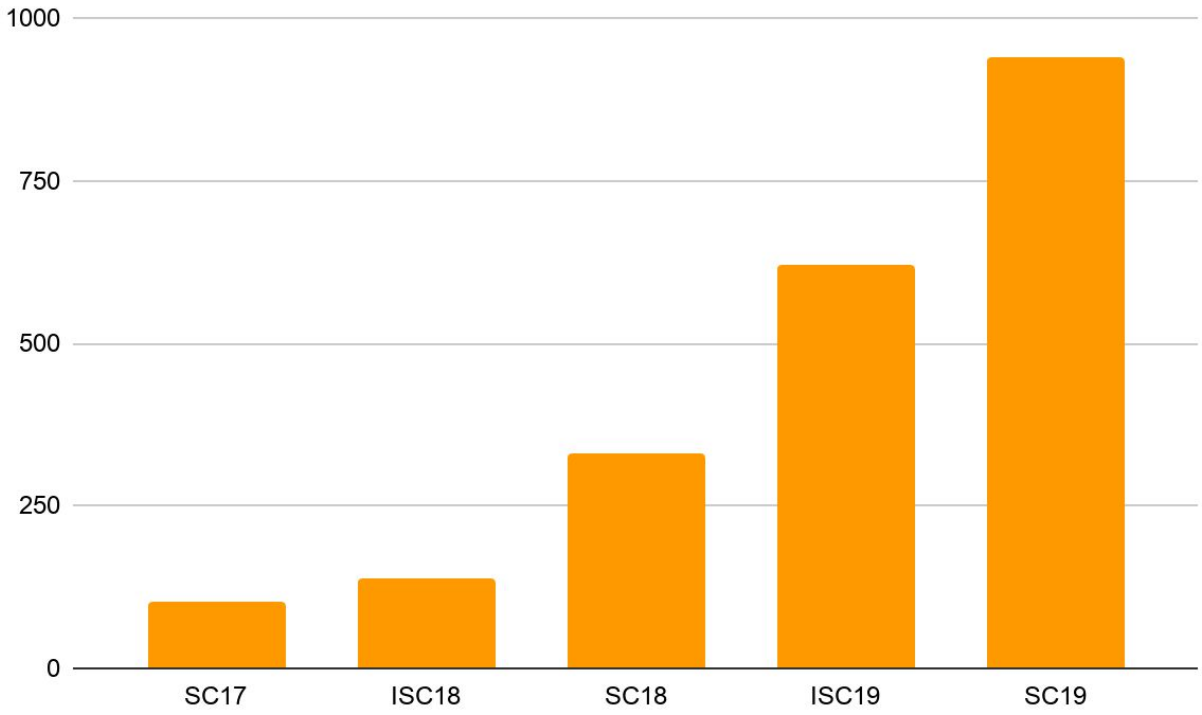
All Metadata by Date



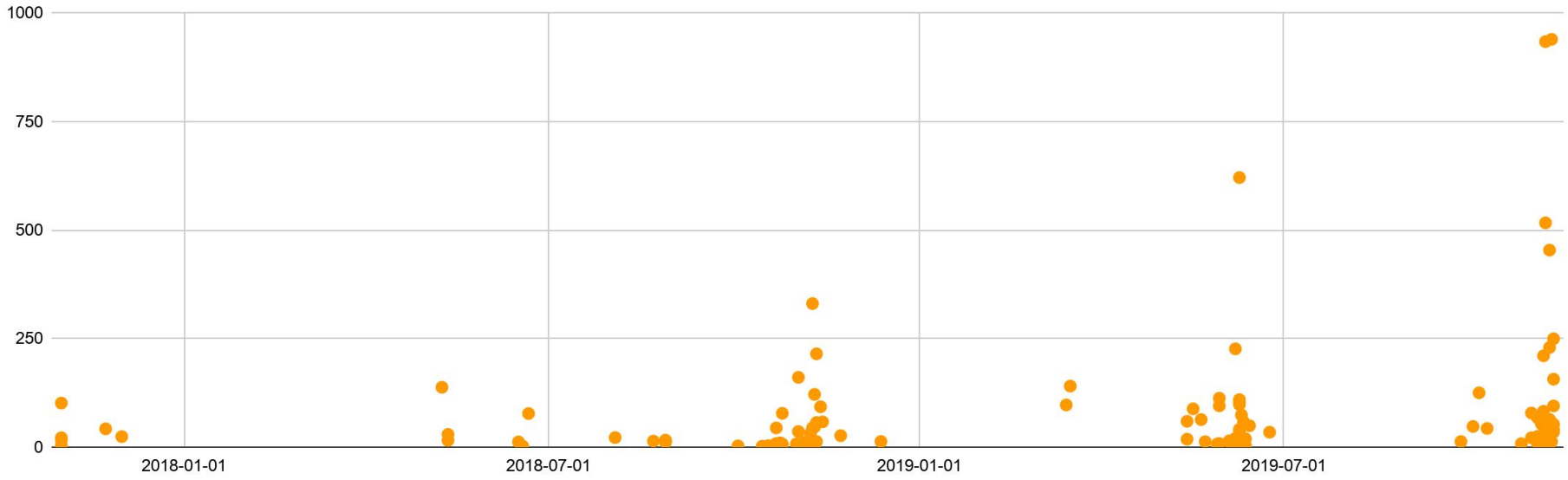
All Metadata by Date (log-scale)



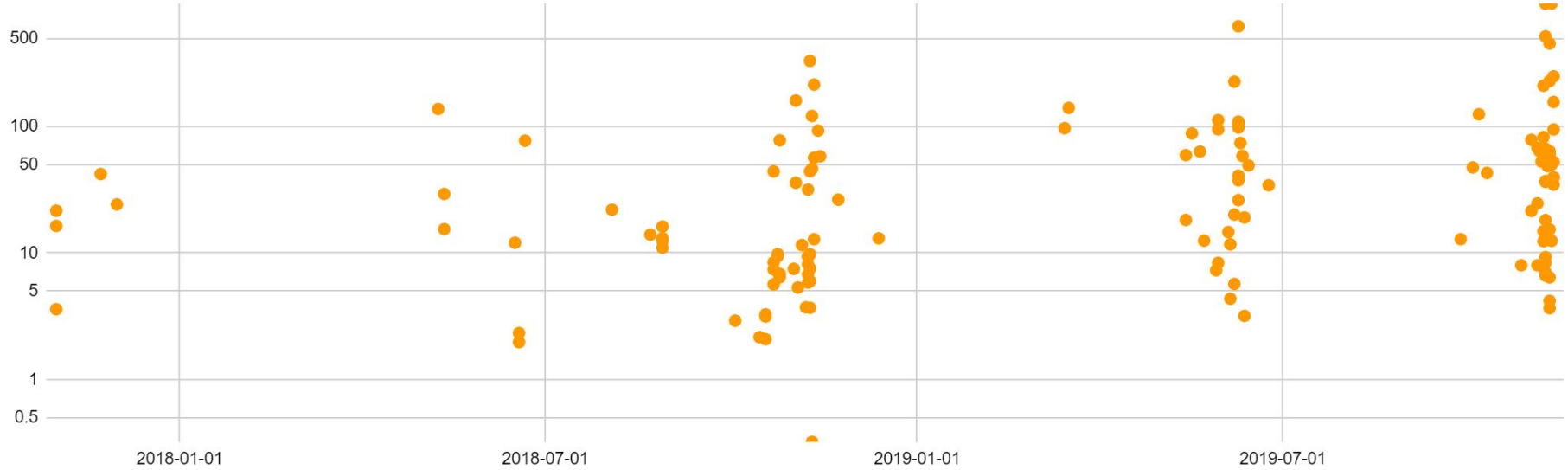
Top Score by List

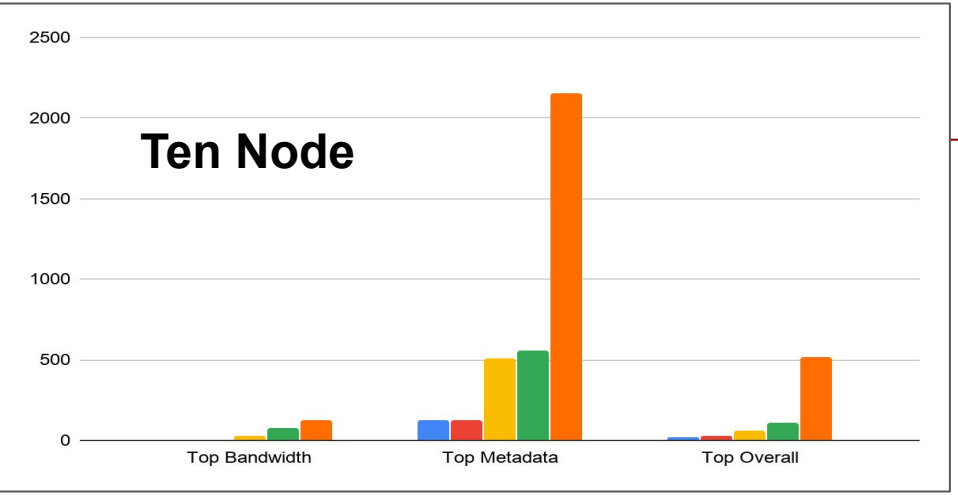


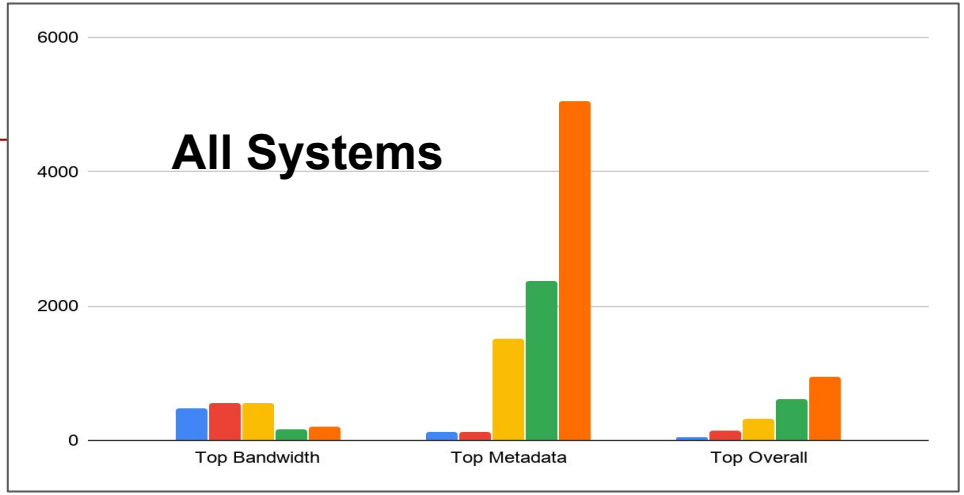
All Scores by Date

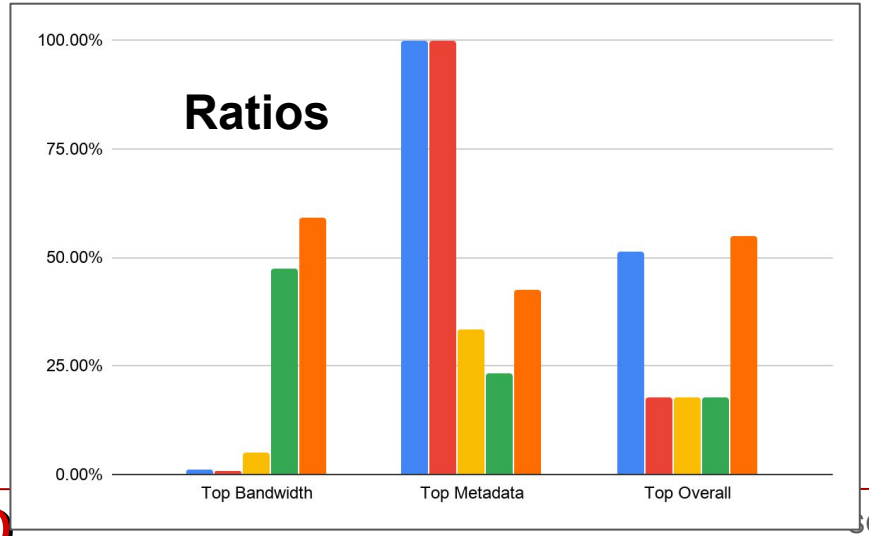


All Scores by Date (log-scale)









Top Scoring Systems by List

- SC17
- ISC18
- SC18
- ISC19
- SC19

What to do with our lists? New Lists or Merge?

- Thus far, it seems like the new rules did not affect people's ability to improve scores
  - Suggests that perhaps we can just merge the new results into the old lists
  - But first let's consider a bit more carefully

- Let's zoom in on mdt_produce, mdt_consume, and mdt_ratio
  - The new rule was designed out of fear that historical mdt_consume rates artificially inflated by cache

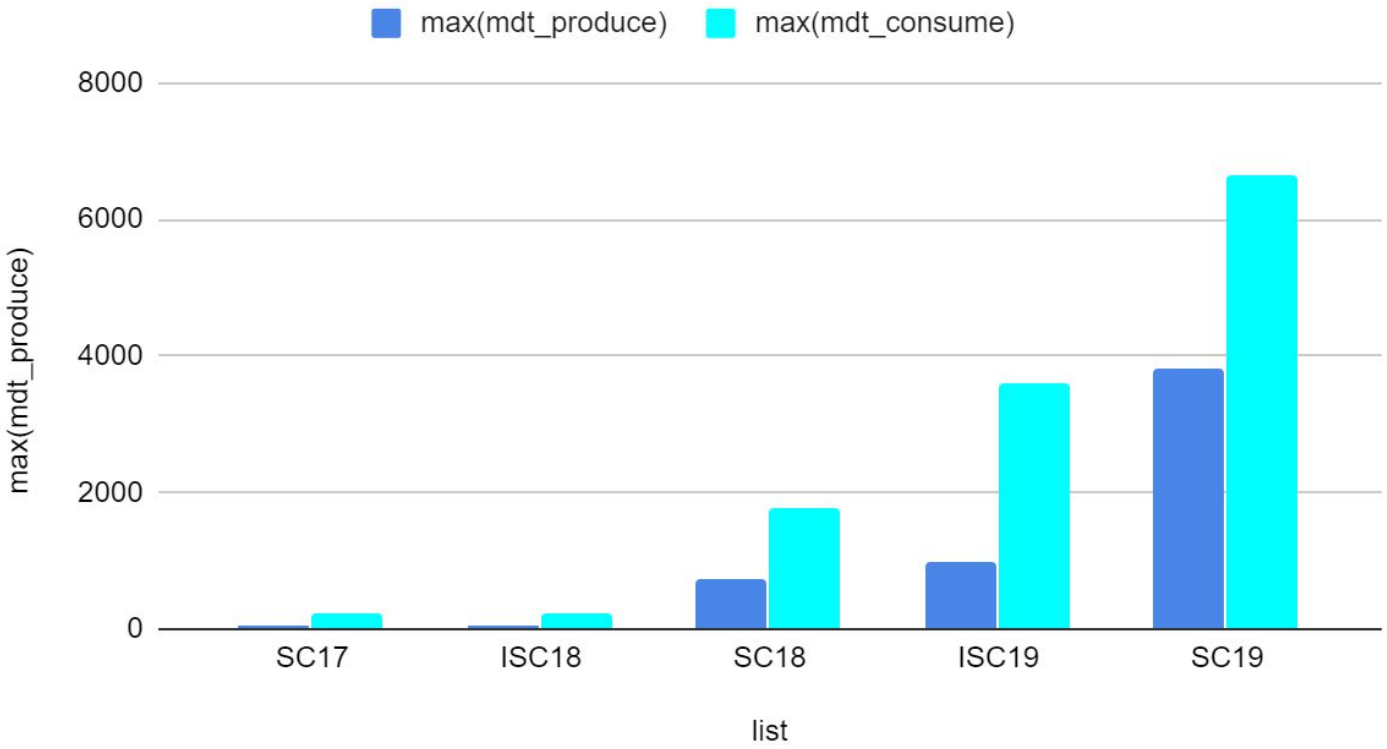
Reminder:

mdt_consume=geo_mean(mdt_easy_stat, mdt_easy_delete, mdt_hard_read, mdt_hard_stat,mdt_hard_delete)

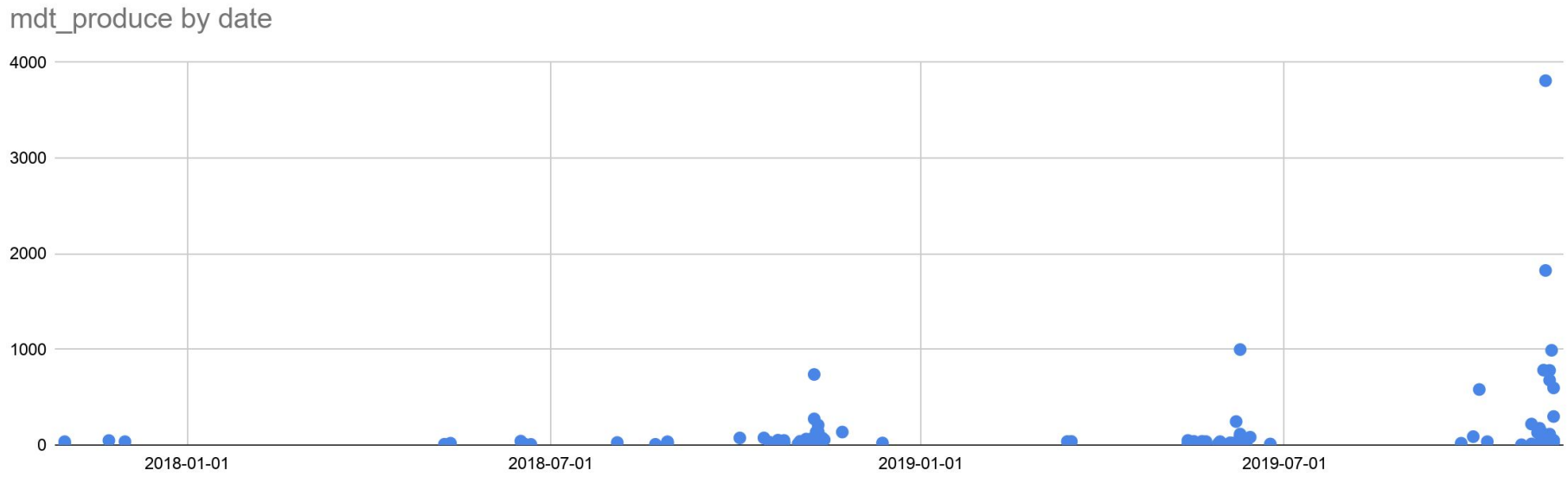
mdt_produce=geo_mean(mdtest_easy_create,mdtest_hard_create)

mdt_ratio = mdt_consume / mdt_produce

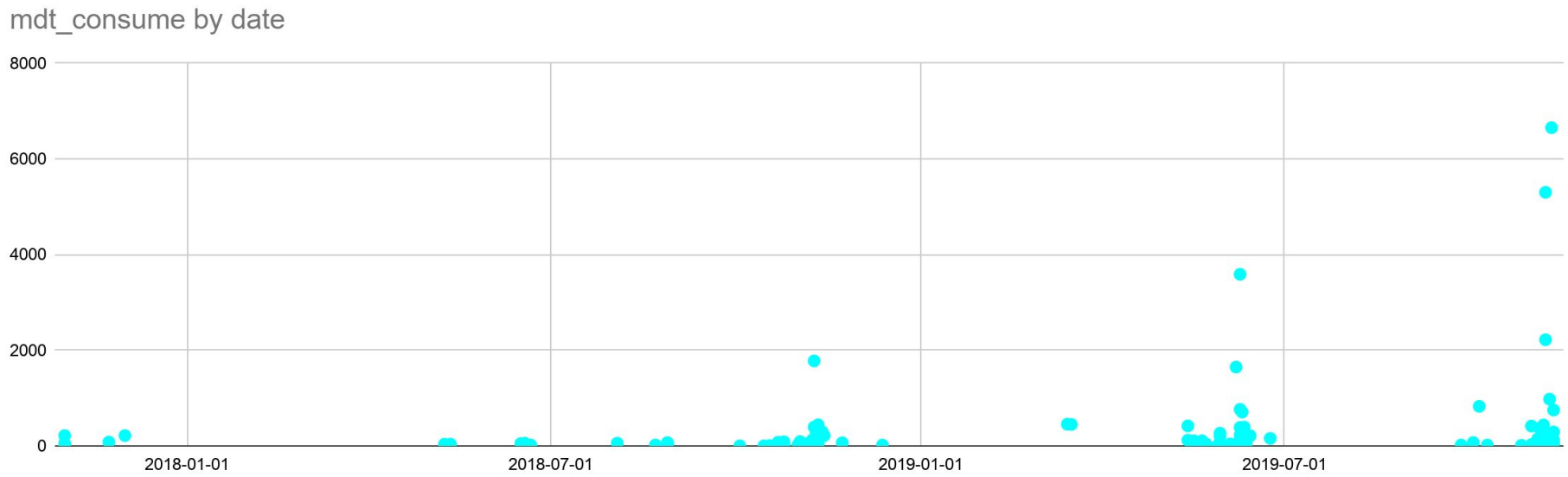
Top mdt_produce and mdt_consume by List



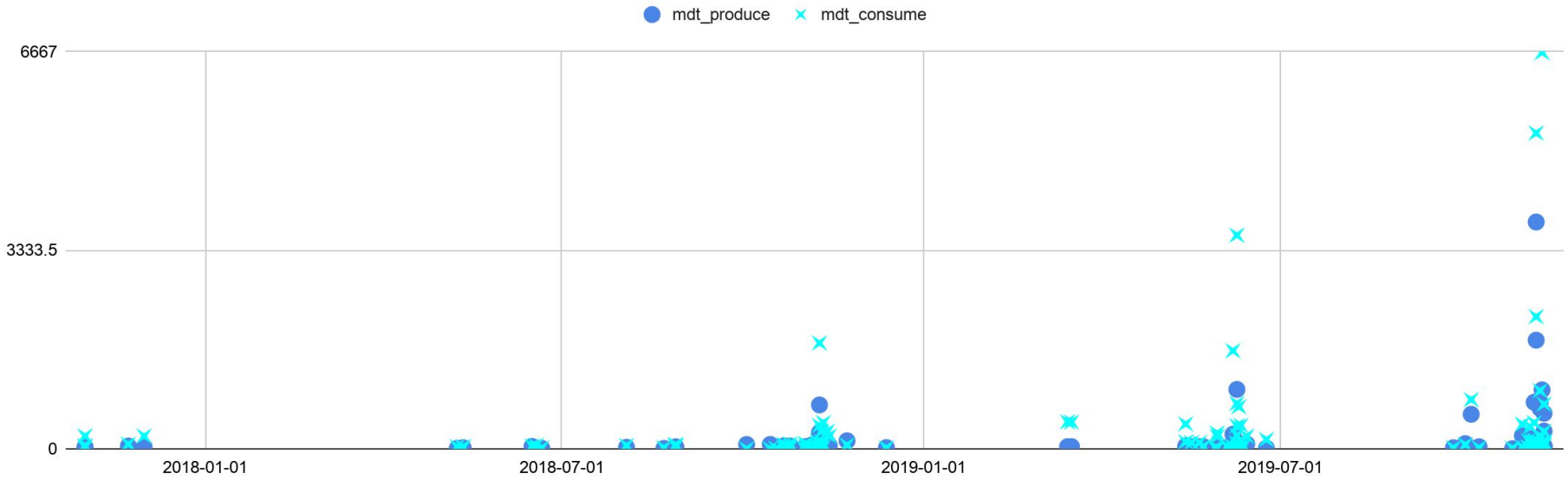
All mdt_produce by Date



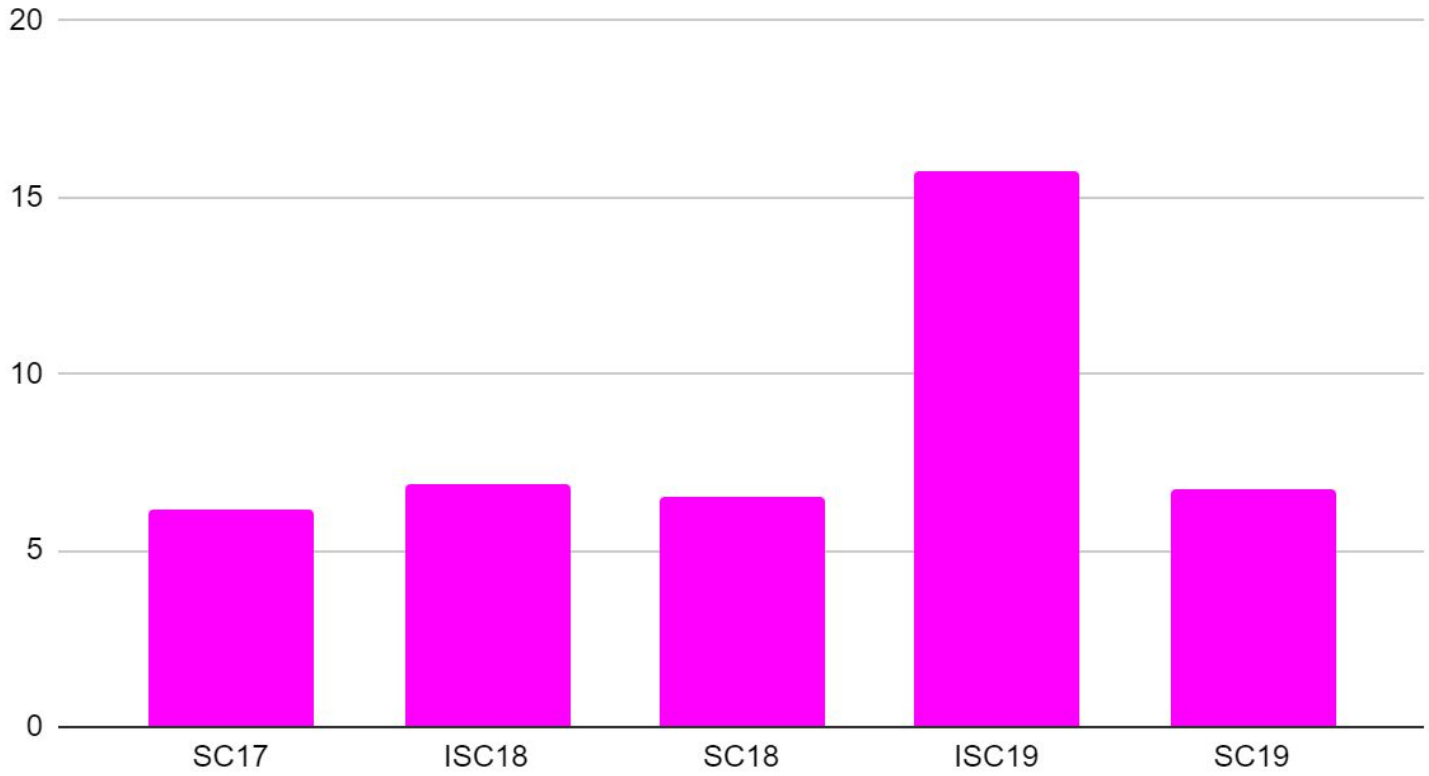
All mdt_consume by Date



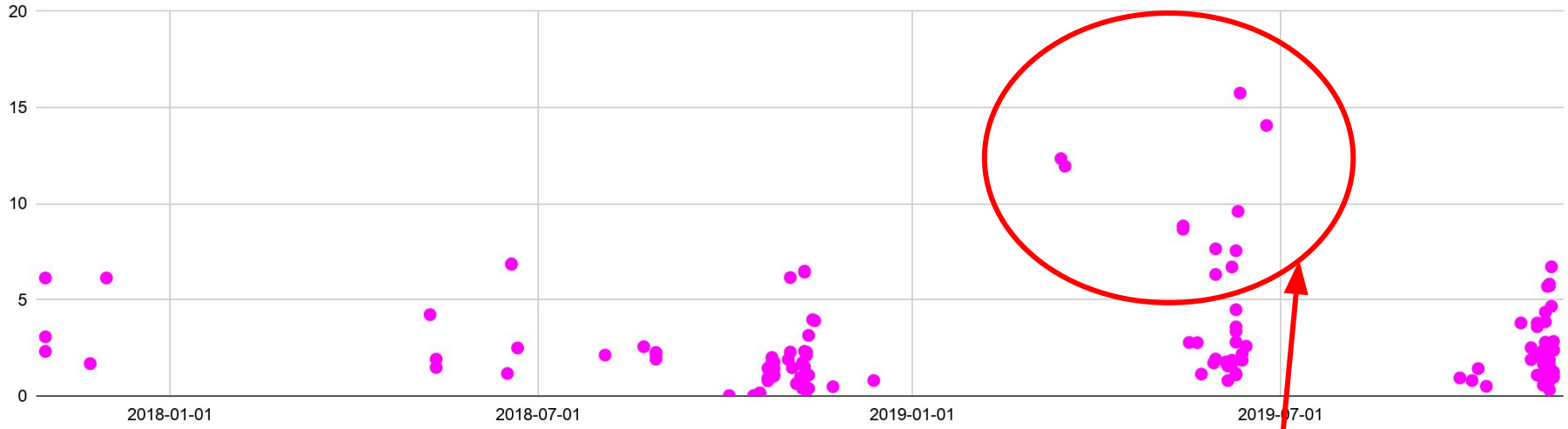
All mdt_produce and mdt_consume by Date



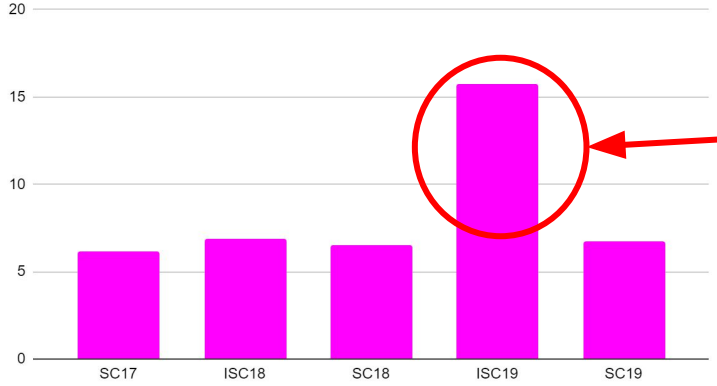
Top mdt_ratio by List



Metadata Consume / Metadata Produce (mdt_ratio) By Date



Top Mdtest Consume / Mdtest Produce (mdt_ratio) by List



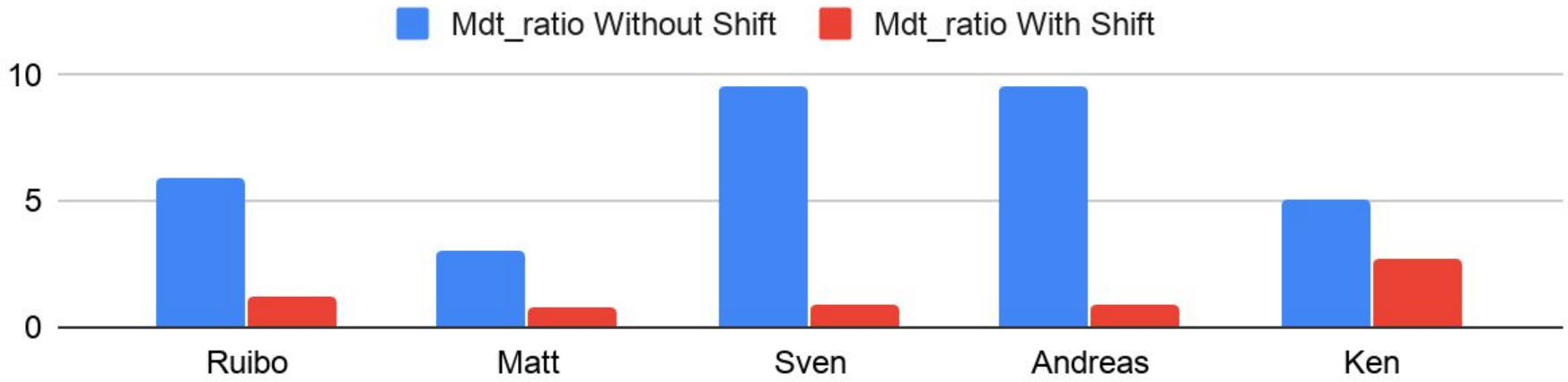
Something definitely changed in ISC19.
Ratio is probably not expected to increase.

Apples:Apples Comparison of Shift Effects

- Five people submitted apples to apples results
  - Ken Carlile, Andreas Dilger, Sven Oehme, Matt Rásó-Barnett, Ruibo Wang
  - Unfortunately four were for Lustre systems, Ken's was Vast however
- Reminder that shift was to avoid client side cache
  - IOR phases and mdtest produce phases and find should not be degraded
  - mdtest stat and read phases could be degraded since client side cache could help these
  - mdtest delete phase might be a bit less likely to be degraded since server must be involved

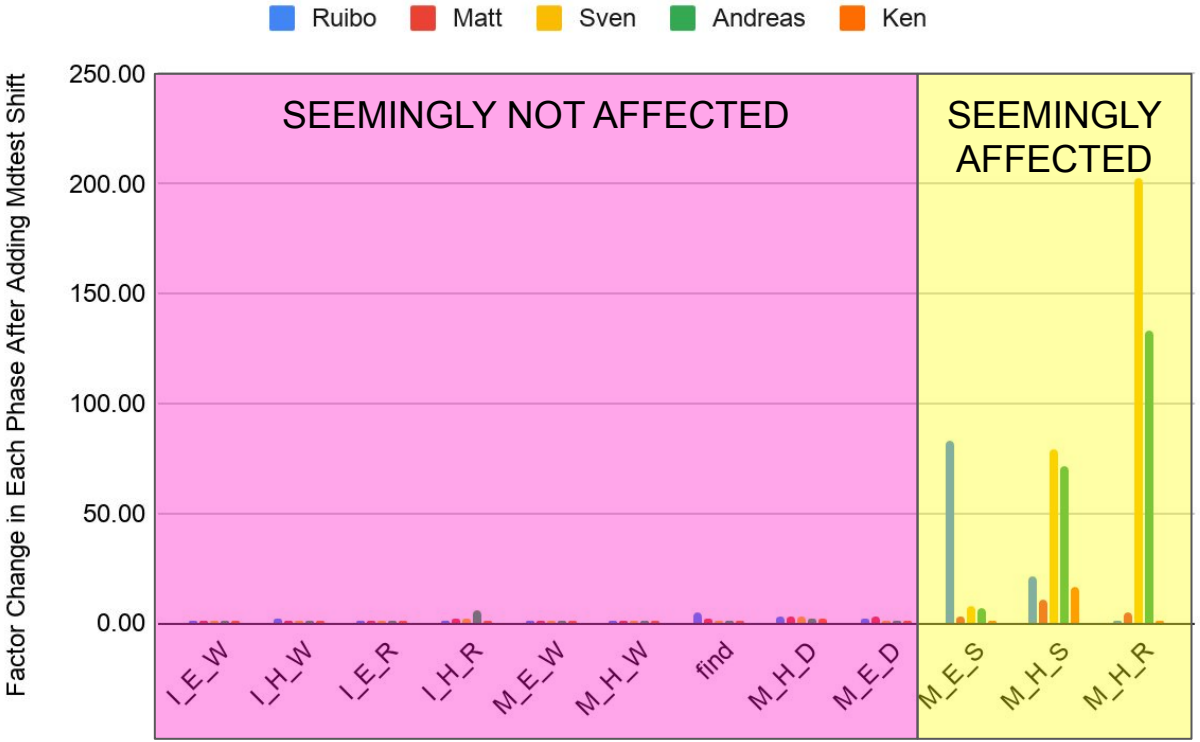
Apples:Apples Comparison of mdtest Shift Effects

- Five people submitted apples to apples results
  - Ken Carlile, Andreas Dilger, Sven Oehme, Matt Rásó-Barnett, Ruibo Wang
- Reminder that shift was added to avoid client side cache
  - IOR phases and mdtest produce phases and find should not be degraded
  - mdtest stat and read phases could be degraded since client side cache could help these
  - mdtest delete phase might be a bit less likely to be degraded since server must be involved
  - If client-side caching had been helping, we would expect to see this in mdt_ratio



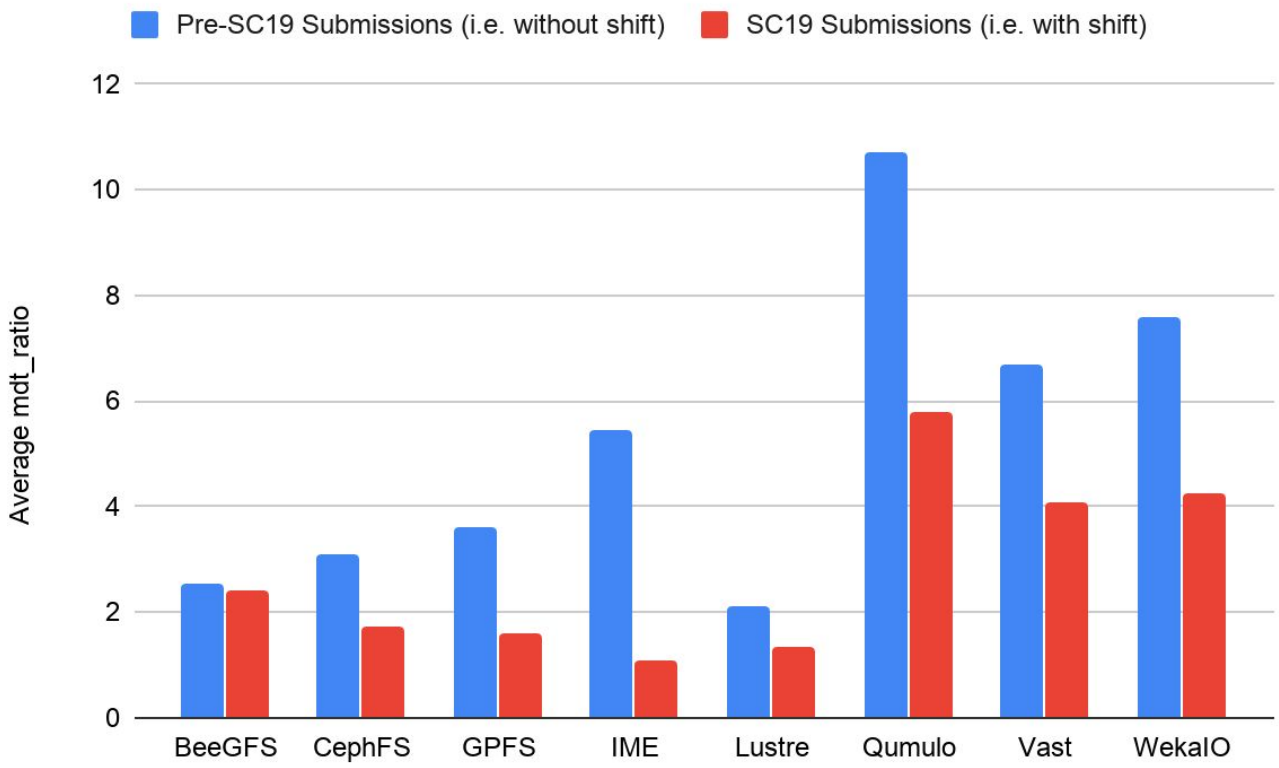
Apples:Apples Results Show Effect of Change

```
I=IOR
M=mdtest
E=easy
H=hard
W=write
R=read
S=stat
D=delete
```



But this was only two file systems! Ken was Vast; others were Lustre.

8 File Systems Had Both Historical and SC19 Results



Drops from blue to red bars suggest that client side caching was benefiting old results.

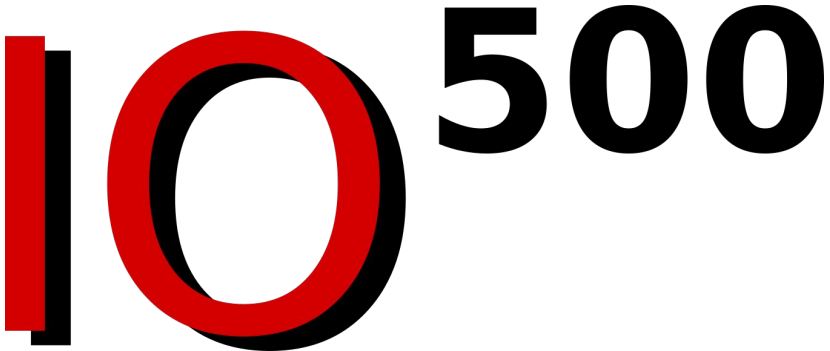
Clearly mdtest Shift Had an Effect - Will Create New Lists

Four Lists Will be Maintained Going Forward

Thanks to community members Ken Carlile, Andreas Dilger, Glenn Lockwood, Sven Oehme, Matt Rásó-Barnett, and Ruibo Wang for offering valuable opinions and data to help with this key decision!

1. Full List
   a. No historical results
   b. All submissions to SC19 and beyond
2. Ranked List
   a. No historical results
   b. Multiple submissions to SC19 and beyond for “system/institution/file system” collapsed into top
3. Ten Node Ranked List
   a. No historical results
   b. Only 10 node submissions to SC19 and beyond will be included
   c. Multiple submissions for “system/institution/file system” collapsed into top submission
4. Historical List
   a. All submissions both historical and new are included

Awards



Six SC19 Awards Will Be Now Given

1. Ten-node
   a. Bandwidth
   b. Metadata
   c. Overall
2. All Systems
   a. Bandwidth
   b. Metadata
   c. Overall

Note that due to the decision about making a new list, only SC19 submissions can compete.

Even though historical bandwidths are fully compatible and were not affected by the mdtest-shift, the committee decided that the move to a new list should be complete. This will minimize any confusion as well as reduce the likelihood that incompatible results are ever inadvertently compared.

10 node challenge - Bandwidth Winner

10 Node Challenge SC19 ONLY

This is the official list from Supercomputing 2019 for the 10 Node Challenge. The list shows the best result for a given combination of system/institution/filesystem qualifying for the 10 Node Challenge.

Sorted by BW

#	information								io500	
	list id	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	bw GiB/s	
1	sc19	Intel	Wolf	Intel	DAOS	10	310	zip	123.89	
2	sc19	National Supercomputing Centre, Singapore	Aspire 1	DDN	IME	10	160	zip	101.75	
3	sc19	NVIDIA	DGX-2H SuperPOD	DDN	Lustre	10	400	zip	86.97	
4	sc19	WekaIO	WekaIO	WekaIO	WekaIO Matrix	10	2610	zip	56.22	
5	sc19	State Key Laboratory of High-end Server & Storage Technology (HSS)	TStor3000	INSPUR	BeeGFS	10	300	zip	41.14	
6	sc19	National Supercomputing Center in Changsha	Tianhe-2E	National University of Defense Technology	Lustre	10	160	zip	35.06	
7	sc19	CSIRO	bracewell	Dell/ThinkParQ	beegfs	10	160	zip	33.77	
8	sc19	Janelia Research Campus, HHMI	weka	WekaIO	wekaio	18	1368	zip	26.22	
9	sc19	University of Cambridge	Data Accelerator	Dell EMC	Lustre	10	320	zip	21.73	
10	sc19	EPCC	NEXTGenIO Prototype	BSC (NEXTGenIO) & JGU (Ada-FS)	Adhoc Filesystem	10	240	zip	21.47	

Certificate

IO-500 Performance Certification

This Certificate is awarded to:

Intel

#1 in the 10 Node BW Score

IO 500

Nov 2019

IO-500 Steering Board

http://io500.org/list/19-11/

10 node challenge - Metadata Winner

10 Node Challenge SC19 ONLY

This is the official list from Supercomputing 2019 for the 10 Node Challenge. The list shows the best result for a given combination of system/institution/filesystem qualifying for the 10 Node Challenge.

Sorted by md

#	information								io500	
	list id	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	bw GiB/s	md kIOP/s
1	sc19	Intel	Wolf	Intel	DAOS	10	310	zip	123.89	2152.46
2	sc19	EPCC	NEXTGenIO Prototype	BSC (NEXTGenIO) & JGU (Ada-FS)	Adhoc Filesystem	10	240	zip	21.47	728.68
3	sc19	NVIDIA	DGX-2H SuperPOD	DDN	Lustre	10	400	zip	86.97	715.76
4	sc19	iFLYTEK	iFLYTEK	Yanrong	YRCloudFile	10	200	zip	13.55	455.18
5	sc19	WekaIO	WekaIO	WekaIO	WekaIO Matrix	10	2610	zip	56.22	435.76
6	sc19	DDN	AI400	DDN	Lustre	10	240	zip	19.65	207.63
7	sc19	University of Cambridge	Data Accelerator	Dell EMC	Lustre	10	320	zip	21.73	167.09
8	sc19	State Key Laboratory of High-end Server & Storage Technology (HSS)	TStor3000	INSPUR	BeeGFS	10	300	zip	41.14	165.71
9	sc19	CSIRO	bracewell	Dell/ThinkParQ	beegfs	10	160	zip	33.77	132.15
10	sc19	Janelia Research Campus, HHMI	weka	WekaIO	wekaio	18	1368	zip	26.22	90.62

Certificate

IO-500 Performance Certification

This Certificate is awarded to:

Intel

#1 in the 10 Node MD Score

IO 500

Nov 2019

IO-500 Steering Board

http://io500.org/list/19-11/

10 node challenge - Winner

10 Node Challenge SC19 ONLY

This is the official list from Supercomputing 2019 for the 10 Node Challenge. The list shows the best result for a given combination of system/institution/filesystem qualifying for the 10 Node Challenge.

Sorted by score

#	information								io500		
	list id	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	score	bw GiB/s	md kIOP/s
1	sc19	Intel	Wolf	Intel	DAOS	10	310	zip	516.41	123.89	2152.46
2	sc19	NVIDIA	DGX-2H SuperPOD	DDN	Lustre	10	400	zip	249.50	86.97	715.76
3	sc19	WekaIO	WekaIO	WekaIO	WekaIO Matrix	10	2610	zip	156.51	56.22	435.76
4	sc19	EPCC	NEXTGenIO Prototype	BSC (NEXTGenIO) & JGU (Ada-FS)	Adhoc Filesystem	10	240	zip	125.08	21.47	728.68
5	sc19	State Key Laboratory of High-end Server & Storage Technology（HSS）	TStor3000	INSPUR	BeeGFS	10	300	zip	82.57	41.14	165.71
6	sc19	iFLYTEK	iFLYTEK	Yanrong	YRCloudFile	10	200	zip	78.54	13.55	455.18
7	sc19	CSIRO	bracewell	Dell/ThinkParQ	beegfs	10	160	zip	66.80	33.77	132.15
8	sc19	DDN	AI400	DDN	Lustre	10	240	zip	63.88	19.65	207.63
9	sc19	University of Cambridge	Data Accelerator	Dell EMC	Lustre	10	320	zip	60.25	21.73	167.09
10	sc19	National Supercomputing Center in Changsha	Tianhe-2E	National University of Defense Technology	Lustre	10	160	zip	52.58	35.06	78.86

Certificate

IO-500 Performance Certification

This Certificate is awarded to:

Intel

#1 in the 10 Node Challenge

IO 500

Nov 2019

IO-500 Steering Board

http://io500.org/list/19-11/

Full list - Bandwidth Winner

Full List

This is the full list from Supercomputing 2019. The list shows all submissions.

IO500

Sorted by BW

#	information								io500	
	list id	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	bw GiB/s	
1	sc19	National Supercomputing Center in Changsha	Tianhe-2E	National University of Defense Technology	Lustre	480	5280	zip	209.43	
2	sc19	Intel	Wolf	Intel	DAOS	26	728	zip	183.36	
3	sc19	WekaIO	WekaIO	WekaIO	WekaIO Matrix	345	8625	zip	174.74	
4	sc19	University of Cambridge	Data Accelerator	Dell EMC	Lustre	128	2048	zip	131.25	
5	sc19	Intel	Wolf	Intel	DAOS	10	310	zip	123.89	
6	sc19	National Supercomputing Centre, Singapore	Aspire 1	DDN	IME	10	160	zip	101.75	
7	sc19	NVIDIA	DGX-2H SuperPOD	DDN	Lustre	10	400	zip	86.97	
8	sc19	CEA	Tera-1000	DDN	Lustre	128	4096	zip	81.01	
9	sc19	CSIRO	bracewell scratch2	Dell/ThinkParQ	beegfs	26	260	zip	69.10	
10	sc19	WekaIO	WekaIO	WekaIO	WekaIO Matrix	10	2610	zip	56.22	

Certificate

IO-500 Performance Certification

This Certificate is awarded to:

National Supercomputing Center in Changsha

#1 in the IO-500 BW Score

IO 500

Nov 2019

IO-500 Steering Board

http://io500.org/list/19-11/

Full list - Metadata Winner

Full List

This is the full list from Supercomputing 2019. The list shows all submissions.

Sorted by md

#	information								io500		
	list id	institution	system	storage vendor	filesystem type	client nodes	client total procs	data		bw GiB/s	md kIOP/s
1	sc19	WekaIO	WekaIO	WekaIO	WekaIO Matrix	345	8625	zip		174.74	5045.33
2	sc19	Intel	Wolf	Intel	DAOS	26	728	zip		183.36	4753.79
3	sc19	Intel	Wolf	Intel	DAOS	10	310	zip		123.89	2152.46
4	sc19	National Supercomputing Center in Changsha	Tianhe-2E	National University of Defense Technology	Lustre	480	5280	zip		209.43	982.78
5	sc19	EPCC	NEXTGenIO Prototype	BSC (NEXTGenIO) & JGU (Ada-FS)	Adhoc Filesystem	10	240	zip		21.47	728.68
6	sc19	NVIDIA	DGX-2H SuperPOD	DDN	Lustre	10	400	zip		86.97	715.76
7	sc19	CEA	Tera-1000	DDN	Lustre	128	4096	zip		81.01	545.74
8	sc19	iFLYTEK	iFLYTEK	Yanrong	YRCloudFile	10	200	zip		13.55	455.18
9	sc19	WekaIO	WekaIO	WekaIO	WekaIO Matrix	10	2610	zip		56.22	435.76
10	sc19	University of Cambridge	Data Accelerator	Dell EMC	Lustre	128	2048	zip		131.25	401.13

Certificate

IO-500 Performance Certification

This Certificate is awarded to:

WekaIO

#1 in the IO-500 MD Score

IO 500

Nov 2019

IO-500 Steering Board

http://io500.org/list/19-11/

Full list - Winner

Full List

This is the full list from Supercomputing 2019. The list shows all submissions.

Sorted by score

#	information								io500		
	list id	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	score	bw	md
										GiB/s	kIOP/s
1	sc19	WekaIO	WekaIO	WekaIO	WekaIO Matrix	345	8625	zip	938.95	174.74	5045.33
2	sc19	Intel	Wolf	Intel	DAOS	26	728	zip	933.64	183.36	4753.79
3	sc19	Intel	Wolf	Intel	DAOS	10	310	zip	516.41	123.89	2152.46
4	sc19	National Supercomputing Center in Changsha	Tianhe-2E	National University of Defense Technology	Lustre	480	5280	zip	453.68	209.43	982.78
5	sc19	NVIDIA	DGX-2H SuperPOD	DDN	Lustre	10	400	zip	249.50	86.97	715.76
6	sc19	University of Cambridge	Data Accelerator	Dell EMC	Lustre	128	2048	zip	229.45	131.25	401.13
7	sc19	CEA	Tera-1000	DDN	Lustre	128	4096	zip	210.26	81.01	545.74
8	sc19	WekaIO	WekaIO	WekaIO	WekaIO Matrix	10	2610	zip	156.51	56.22	435.76
9	sc19	EPCC	NEXTGenIO Prototype	BSC (NEXTGenIO) & JGU (Ada-FS)	Adhoc Filesystem	10	240	zip	125.08	21.47	728.68
10	sc19	CSIRO	bracewell scratch2	Dell/ThinkParQ	beegfs	26	260	zip	94.86	69.10	130.23

Full list - Winner

Full List

This is the full list from Supercomputing 2019. The list shows all submissions.

Sorted by score

#	information								io500		
	list id	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	score	bw GiB/s	md kIOP/s
1	sc19	WekaIO	WekaIO	WekaIO	WekaIO Matrix	345	8625	zip	938.95	174.74	5045.33
2	sc19	Intel	Wolf	Intel	DAOS	26	728	zip	933.64	183.36	4753.79
3	sc19	Intel	Wolf	Intel	DAOS	10	310	zip	516.41	123.89	2152.46
4	sc19	National Supercomputing Center in Changsha	Tianhe-2E	National University of Defense Technology	Lustre	480	5280	zip	453.68	209.43	982.78
5	sc19	NVIDIA	DGX-2H SuperPOD	DDN	Lustre	10	400	zip	249.50	86.97	715.76
6	sc19	University of Cambridge	Data Accelerator	Dell EMC	Lustre	128	2048	zip	229.45	131.25	401.13
7	sc19	CEA	Tera-1000	DDN	Lustre	128	4096	zip	210.26	81.01	545.74
8	sc19	WekaIO	WekaIO	WekaIO	WekaIO Matrix	10	2610	zip	156.51	56.22	435.76
9	sc19	EPCC	NEXTGenIO Prototype	BSC (NEXTGenIO) & JGU (Ada-FS)	Adhoc Filesystem	10	240	zip	125.08	21.47	728.68
10	sc19	CSIRO	bracewell scratch2	Dell/ThinkParQ	beegfs	26	260	zip	94.86	69.10	130.23

0.57% difference

Certificate

IO-500 Performance Certification

This Certificate is awarded to:

WekaIO

#1 in the IO-500

IO 500

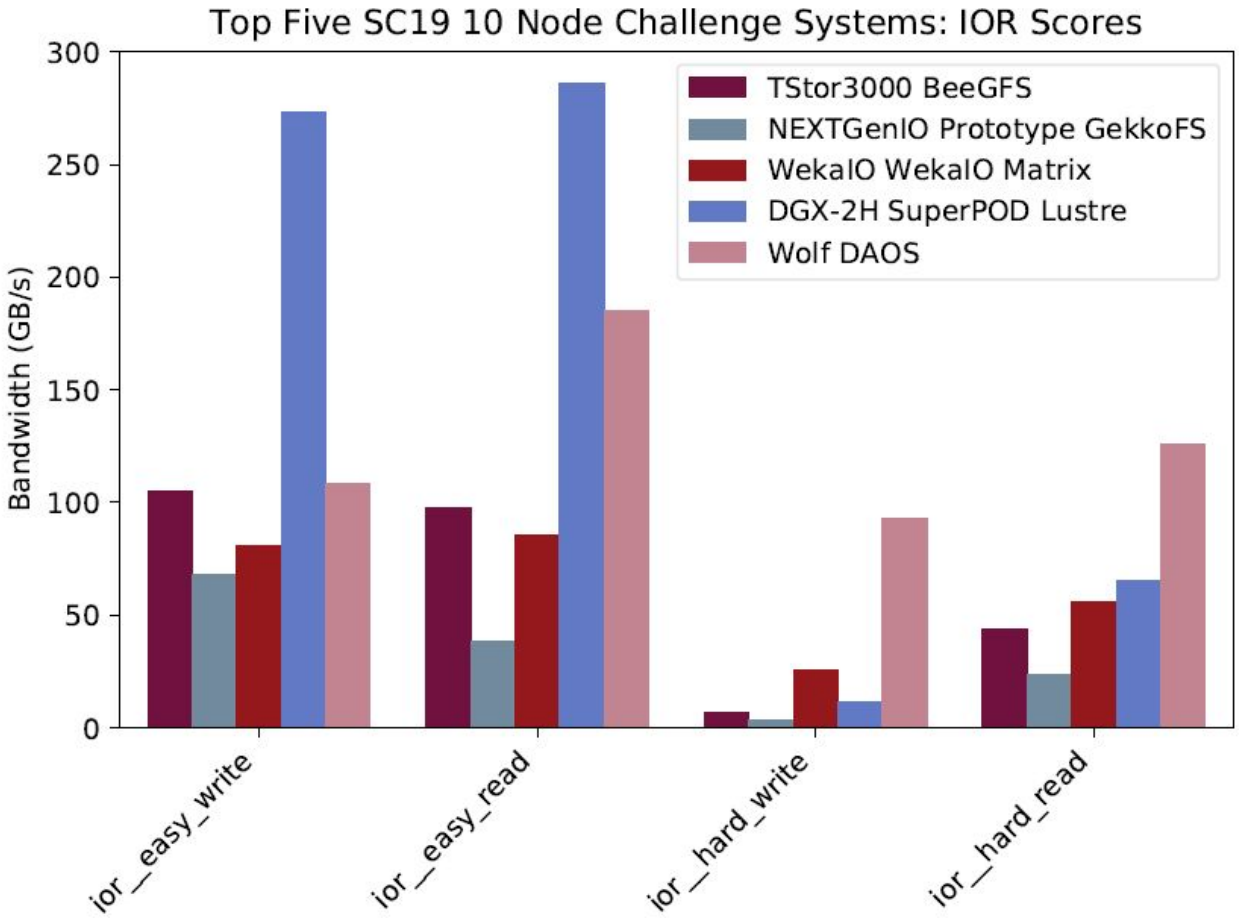
Nov 2019

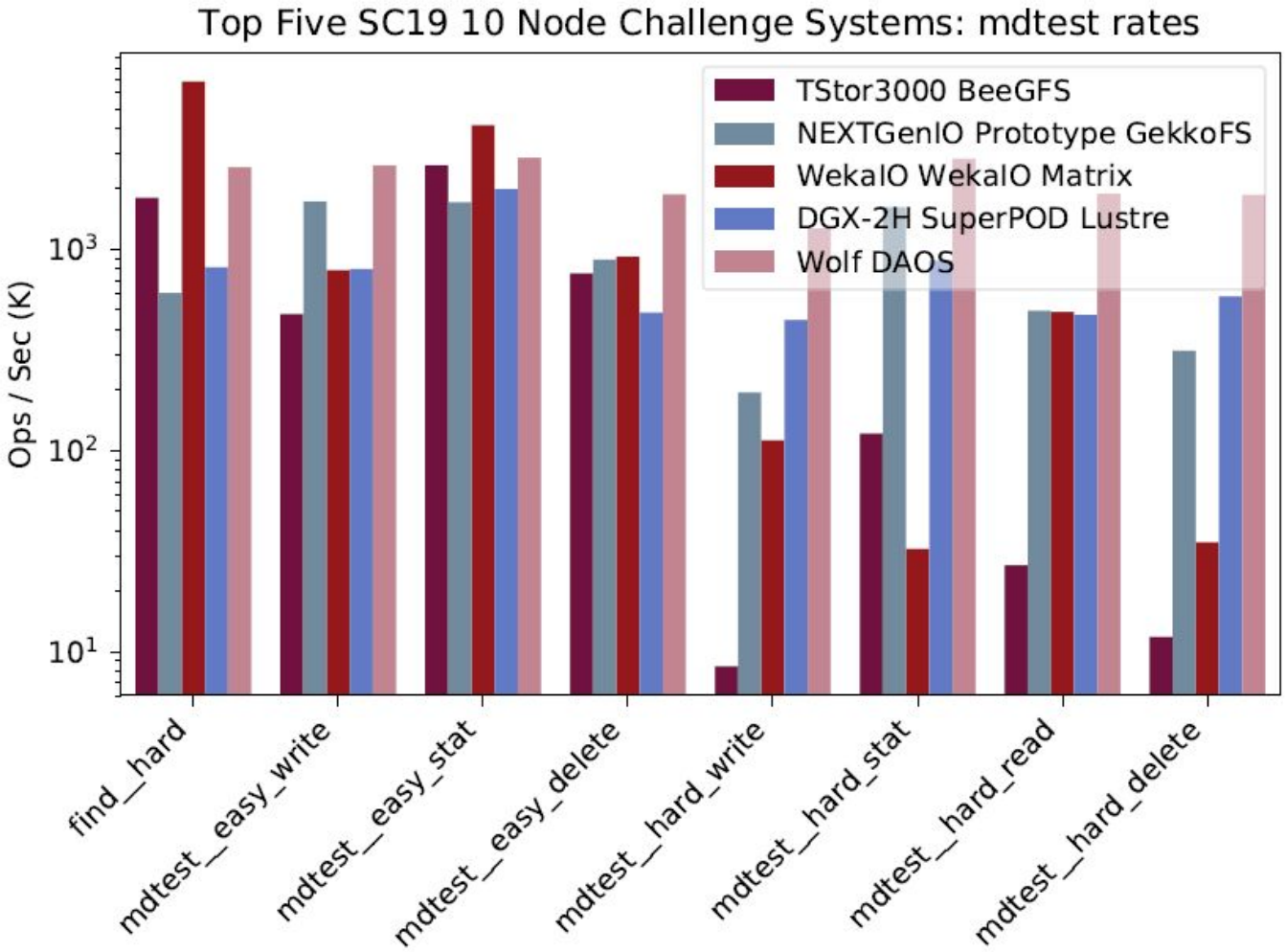
IO-500 Steering Board

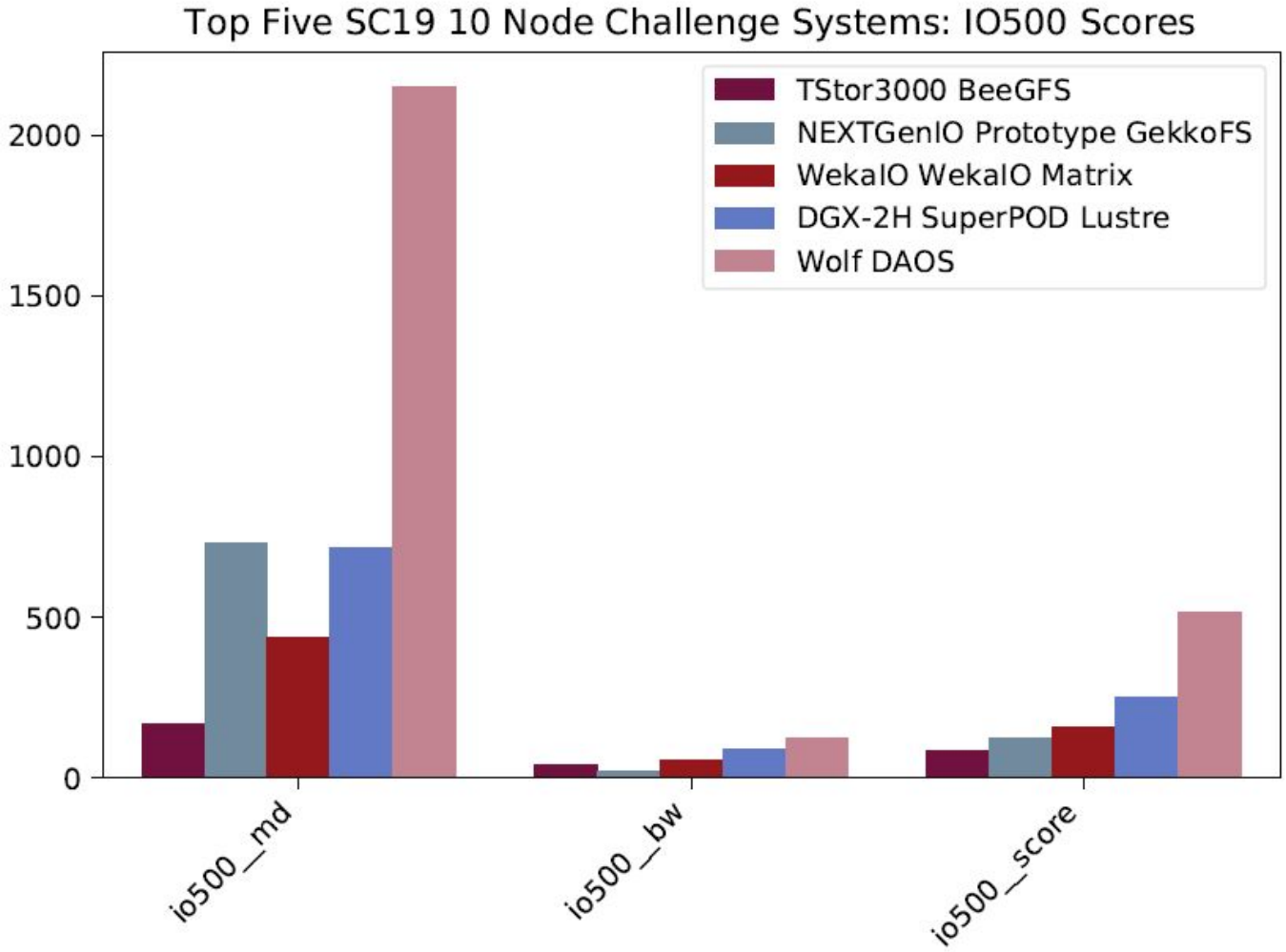
http://io500.org/list/19-11/

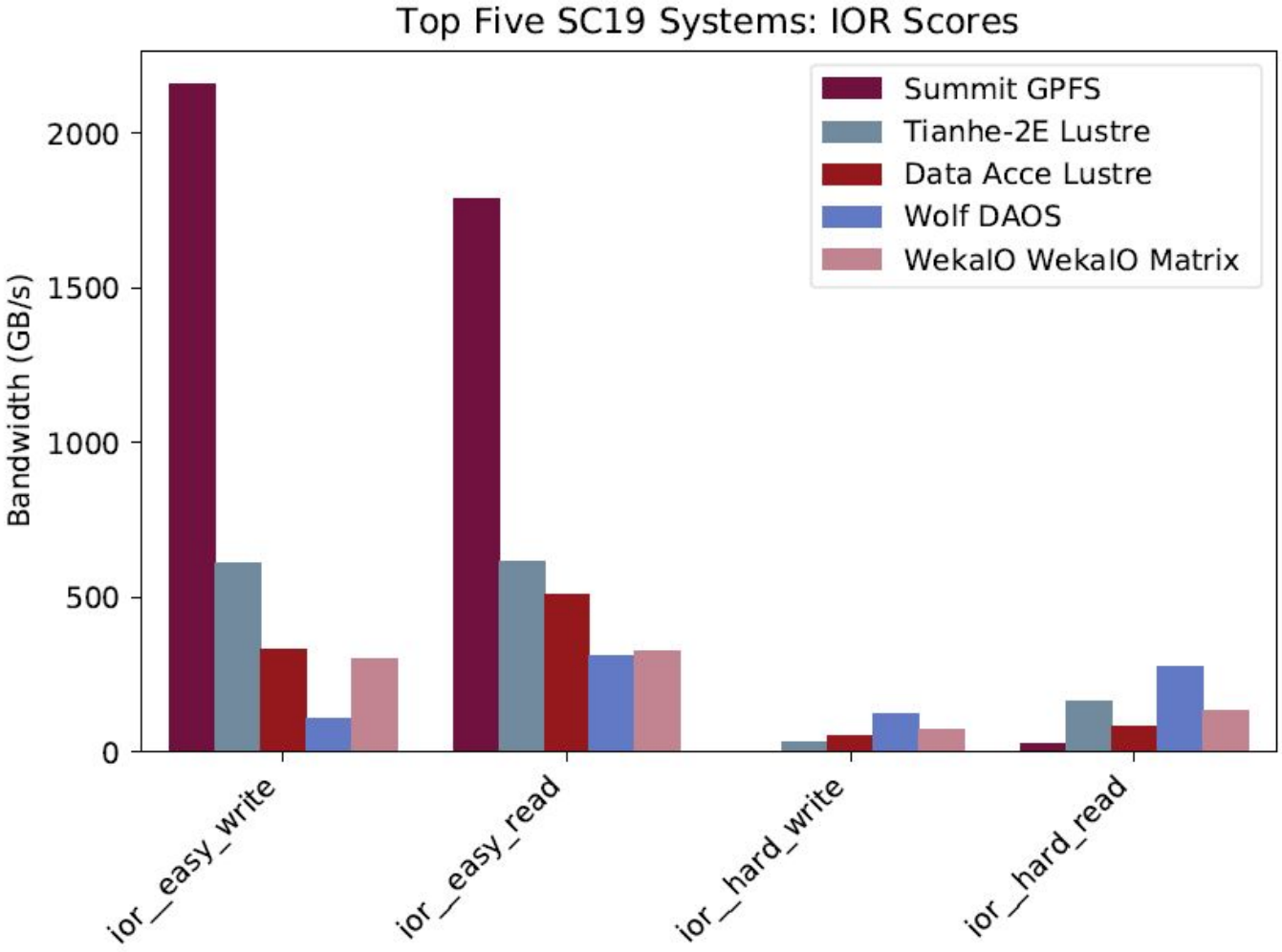
SC19 List of Awarded Systems

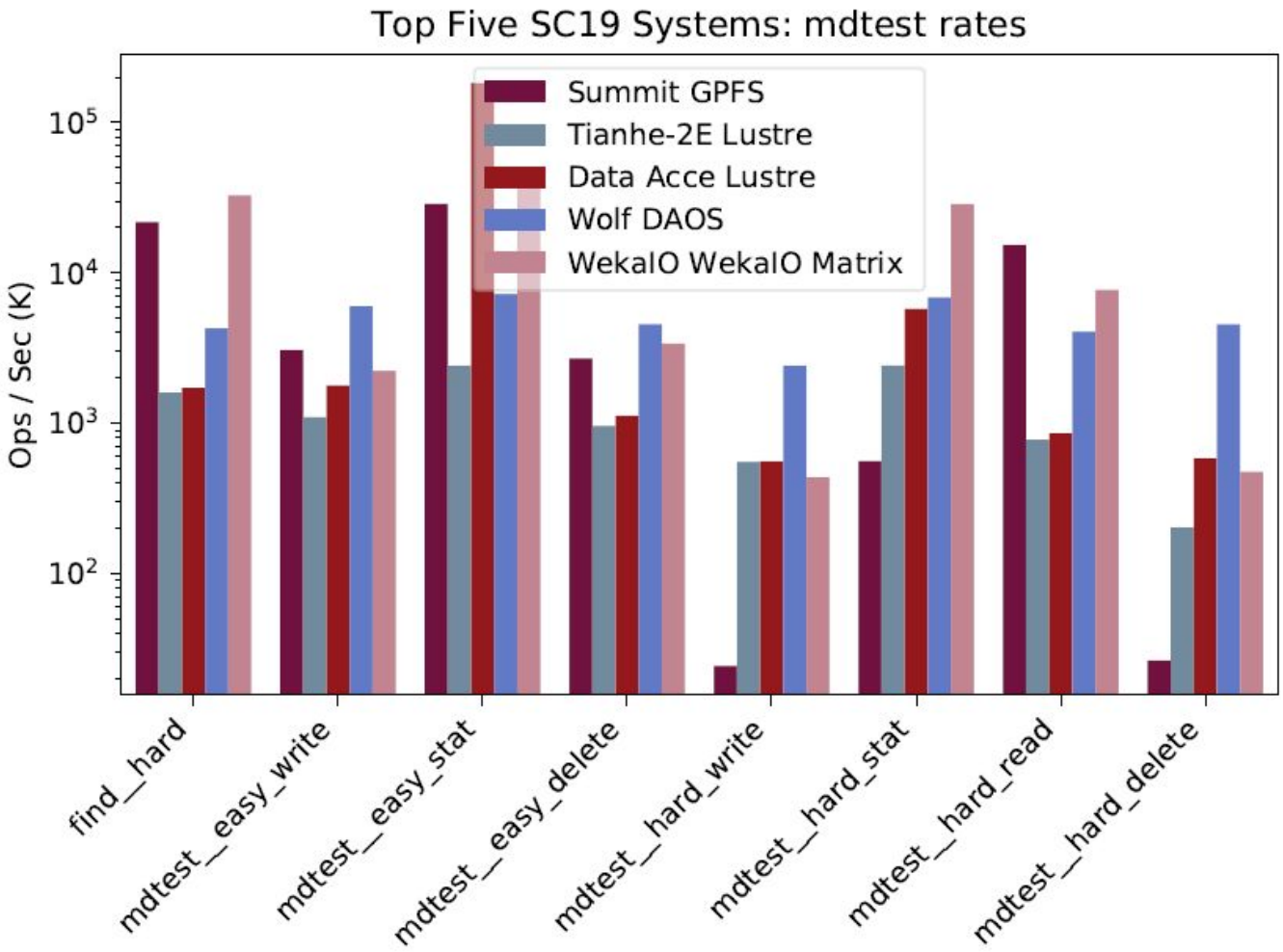
10-Node	Metadata	Intel	DAOS	2152	KIOPS
	Bandwidth	Intel	DAOS	124	GiB/s
	Overall	Intel	DAOS	516	score
All Systems	Metadata	WekaIO	Matrix	5045	KIOPS
	Bandwidth	Tianhe	Lustre	209	GiB/s
	Overall	WekaIO	Matrix	939	score

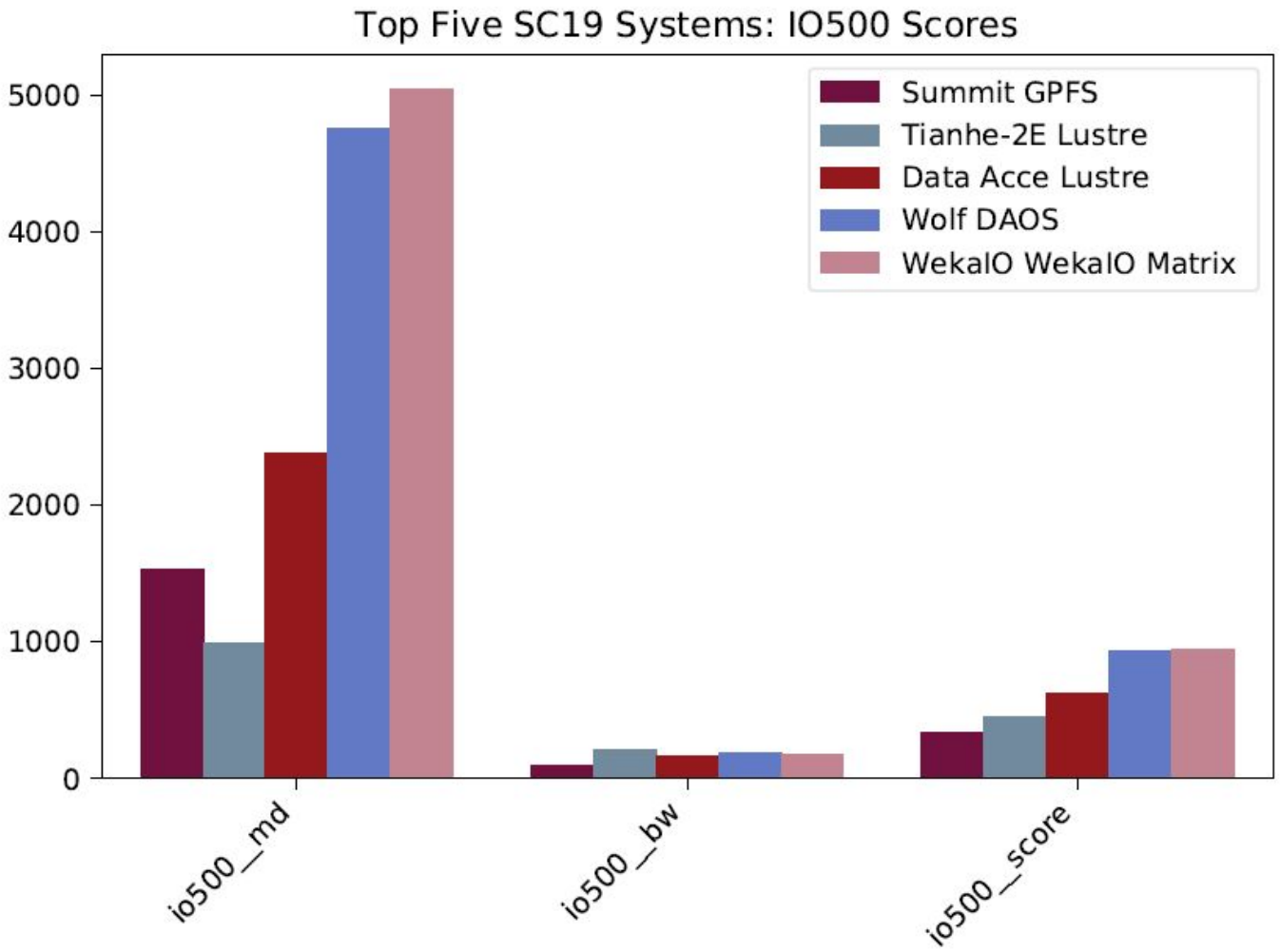

Top Five SC19 10 Node Challenge Systems: IOR Scores
Bandwidth (GB/s)
300
250
200
150
100
50
0
TStor3000 BeeGFS
NEXTGenIO Prototype GekkoFS
WekaIO WekaIO Matrix
DGX-2H SuperPOD Lustre
Wolf DAOS
ior_easy_write
ior_easy_read
ior_hard_write
ior_hard_read


Top Five SC19 10 Node Challenge Systems: mdtest rates
TStor3000 BeeGFS
NEXTGenIO Prototype GekkoFS
WekaIO WekaIO Matrix
DGX-2H SuperPOD Lustre
Wolf DAOS
Ops / Sec (K)
10^3
10^2
10^1
find__hard
mdtest__easy_write
mdtest__easy_stat
mdtest__easy_delete
mdtest__hard_write
mdtest__hard_stat
mdtest__hard_read
mdtest__hard_delete


Top Five SC19 10 Node Challenge Systems: IO500 Scores
TStor3000 BeeGFS
NEXTGenIO Prototype GekkoFS
WekaIO WekaIO Matrix
DGX-2H SuperPOD Lustre
Wolf DAOS
2000
1500
1000
500
0
io500_md
io500_bw
io500_score

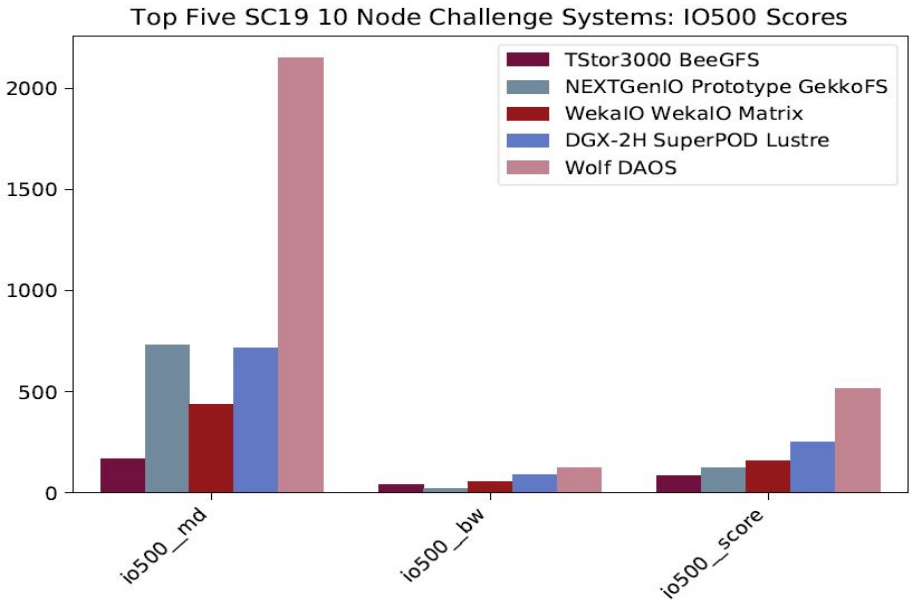

Top Five SC19 Systems: IOR Scores
Summit GPFS
Tianhe-2E Lustre
Data Acce Lustre
Wolf DAOS
WekaIO WekaIO Matrix
Bandwidth (GB/s)
0
500
1000
1500
2000
ior_easy_write
ior_easy_read
ior_hard_write
ior_hard_read

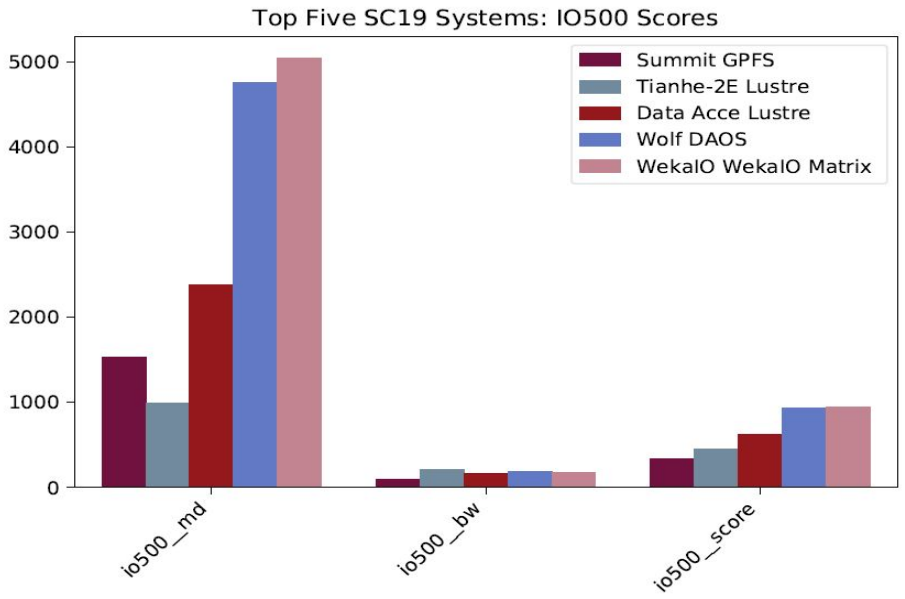

Top Five SC19 Systems: mdtest rates
Summit GPFS
Tianhe-2E Lustre
Data Acce Lustre
Wolf DAOS
WekaIO WekaIO Matrix
Ops / Sec (K)
10^5
10^4
10^3
10^2
find_hard
mdtest_easy_write
mdtest_easy_stat
mdtest_easy_delete
mdtest_hard_write
mdtest_hard_stat
mdtest_hard_read
mdtest_hard_delete


Top Five SC19 Systems: IO500 Scores
5000
4000
3000
2000
1000
0
Summit GPFS
Tianhe-2E Lustre
Data Acce Lustre
Wolf DAOS
WekaIO WekaIO Matrix
io500_md
io500_bw
io500_score

SC19 IO500 Winners and Honorable Mentions

10-Node	Metadata	Intel	DAOS	2152	KIOPS
	Bandwidth	Intel	DAOS	124	GiB/s
	Overall	**Intel**	**DAOS**	**516**	**score**
All Systems	Metadata	WekaIO	Matrix	5045	KIOPS
	Bandwidth	Tianhe	Lustre	209	GiB/s
	Overall	**WekaIO**	**Matrix**	**939**	**score**



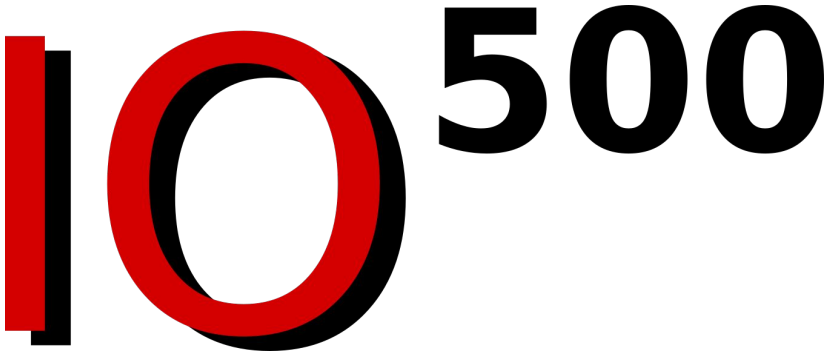


SC Student Cluster Competition Preliminary results*

Ranking	Team	Score	BW	MD	Nodes	DropCache
1	Nanyang	28.28	4.139	193.21	1	yes
2	Purdue	26.0845	9.8	69.4	5	no
3	Shanghai Tech	20.4175	2.61	159.381	1	yes
4	ETH Zurich	19.588	3.23	118.575	1	yes
5	Peking	16.21	2.54	103.2	1	yes
6	FAU	13.82	2.02	94.4	1	yes
7	UIUC	12.69	6.5	24.77	1	yes
8	NTHU	9.4	1.05	84.16	1	yes
9	Tennessee	8.96	1.85	43.38	1	yes
10	Wake Forest	5.44	0.98	30.2	1	yes
11	Warsaw	4.95	1.33	18.38	2	no
12	Washington	4.39	0.248	77.916	1	yes
13	Shanghai Jiao Tong	3.279	1.188	9.05	1	yes
Not valid						
	Tsinghua	30.55	3.08	303.121	1	yes
	NC State	24.64	2.39	253.351	1	no

* Results received less than 18 hours ago and have not been fully validated by the committee yet

Roadmap



Roadmap for the IO-500

- Rewrite the IO-500 into a C-application instead of the script solution
  - Run by using a configuration file and no additional arguments
  - Improved error handling and validation for submitters
  - Produces the same results as the current bash solution
- Using MDTest data validation during mdtest hard read
  - Compares read data with the expectation
- Integration of tools to automatically harvest system configuration
- Rewrite the webpage
  - Move 100% of code into github
  - Mirror at io500.vi4io.org and io500.org

C-Application / Thoughts

- Running the application should be as simple as (e.g. SLURM)

```
#!/bin/bash -e
#SBATCH -p compute2
#SBATCH --nodes=10
module load OpenMPI

mpiexec -np 20 ./io500 final-config.ini
```

- Configuration could be INI or JSON files
  - Providing only options that are allowed to tune; options for additional testing in extra section
- The tool provides a dry-run option showing the exact commands it runs
  - `e.g. mpiexec -np 20 ./io500 --dry-run final-config.ini`
  - Dumps the full INI options that are available and their current values
  - Shows predicted execution behavior:
    - I run mdtest with these arguments, then this then that….
    - The result should be valid or will definitely be invalid based on the options provided

C-Application / Configuration File Draft

```
[find]
external-program = ./bin/mmfind.sh # wrapper returns similar output

[ior-hard]
API = MPIIO # Like when using ior -O <OPTION>=<VALUE>
hintsFileName = my-hints.txt

[ior-hard write] # Some options might be valid for specific sections
posix.odirect = 1

[optional]
ior-random = enable

[debug] # the program will warn if anything is invalid
drop-caches = TRUE
stonewall-time = 10
```

Discussion

10^{500}

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.

http://io500.org/rules/proposals

Open Floor

Issues about Fair Comparisons

- Non-erasure vs erasure systems
- Production system versus benchmark-only system
- Vendor submission versus customer submission
- GA File system versus research file system
- Cloud vs on-prem
- Ephemeral vs persistent file system
- Storage media