BIRDS OF A FEATHER

IO500: The High-Performance Storage Community

Jean Luca Bez – Lawrence Berkeley National Laboratory
Andreas Dilger – The Lustre Collective
Dean Hildebrand – Google
Julian Kunkel – Georg-August-Universität Göttingen/GWDG
Jay Lofstead – Sandia National Laboratories
George Markomanolis – AMD

BoF Agenda

- **Welcome** – Jean Luca
- **Award Presentations** – Jean Luca
- **IO500 List Analysis** – Andreas Dilger
- **Community Talk**
  - “Using IO500 for Storage System Sizing” – Michael Hennecke, HPE
- **Updates** – Dean Hildebrand
  - 4KB easy read phase update
  - Website
  - Proposed community policies
- **Community Discussion** – Jay Lofstead

IO500 Organization Status

- IO500 Foundation is a US non-profit, public charity organization
  - Domain, mailing list, servers, GitHub belongs to IO500 Foundation
- Website contains results with links to details, CFS, BoF slides
  - **io500.org**
  - Contribute fixes at **github.com/IO500/webpage**
- Please join our mailing list for announcements:
  - **io500.org/contact**
- Please join our Slack for discussions:
  - **io500workspace.slack.com**
  - Join link: **rb.gy/sn8esm**

IO500

bit.ly/io500poll

IO500 Award Ceremony

IO500 Award Ceremony

10 Client Node Production List

10 Client Node Production
Bandwidth Winner

IO500

	#	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE	BW ↑ (GIB/S)	MD (KIOP/S)
	1	SC23	Aurora	Argonne National Laboratory	DAOS		734.50	
New	2	SC25	SuperMUC-NG-Phase2-EC-10	LRZ	DAOS		253.98	
New	3	SC25	Maximus-01	Core42	DAOS		247.90	
	4	SC24	CHIE-2	SoftBank Corp	EXAScaler		159.93	
	5	SC24	GEFION	Danish Centre for AI innovation AS	EXAScaler		154.70	
New	6	SC25	CHIE-4	SoftBank Corp	EXAScaler		148.88	
	7	ISC25	HRT	Hudson River Trading	EXAScaler		136.05	
	8	ISC25	SAKURAONE	SAKURA Internet Inc and Prunus Solutions Inc	EXAScaler		133.03	
	9	SC24	HiPerGator AI	University of Florida	EXAScaler		124.89	
New	10	SC25	JoyBuilder AI Development Service Platform	JD Explore Academy	JPFS		117.37	

10 Client Node Production
Metadata Winner

IO500

	#	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE	BW (GIB/S)	MD ↑ (KIOP/S)
	1	SC23	Aurora	Argonne National Laboratory	DAOS		734.50	11,336.27
New	2	SC25	SuperMUC-NG-Phase2-EC-10	LRZ	DAOS		253.98	6,187.21
New	3	SC25	Maximus-01	Core42	DAOS		247.90	1,793.03
	4	ISC24	Lise	Zuse Institute Berlin	DAOS		65.01	1,620.13
New	5	SC25	JoyBuilder AI Development Service Platform	JD Explore Academy	JPFS		117.37	1,400.93
New	6	SC25	HiPerGator	University of Florida	EXAScaler		79.39	944.56
	7	ISC25	HRT	Hudson River Trading	EXAScaler		136.05	890.51
	8	SC24	GEFION	Danish Centre for AI innovation AS	EXAScaler		154.70	637.43
New	9	SC25	CHIE-4	SoftBank Corp	EXAScaler		148.88	617.51
	10	SC24	CHIE-2	SoftBank Corp	EXAScaler		159.93	560.19

10 Client Node Production
Overall Winner

IO500

	# ↑	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
New	2	SC25	SuperMUC-NG-Phase2-EC-10	LRZ	DAOS	1,253.56	253.98	6,187.21
New	3	SC25	Maximus-01	Core42	DAOS	666.70	247.90	1,793.03
New	4	SC25	JoyBuilder AI Development Service Platform	JD Explore Academy	JPFS	405.50	117.37	1,400.93
	5	ISC25	HRT	Hudson River Trading	EXAScaler	348.08	136.05	890.51
	6	ISC24	Lise	Zuse Institute Berlin	DAOS	324.54	65.01	1,620.13
	7	SC24	GEFION	Danish Centre for AI innovation AS	EXAScaler	314.03	154.70	637.43
New	8	SC25	CHIE-4	SoftBank Corp	EXAScaler	303.20	148.88	617.51
	9	SC24	CHIE-2	SoftBank Corp	EXAScaler	299.32	159.93	560.19
New	10	SC25	HiPerGator	University of Florida	EXAScaler	273.84	79.39	944.56

IO500 Award Ceremony
Production List

Production
Bandwidth Winner

IO500

	#	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE	BW ↑ (GIB/S)	MD (KIOP/S)
	1	SC23	Aurora	Argonne National Laboratory	DAOS		10,066.09	
New	2	SC25	SuperMUC-NG-Phase2-EC	LRZ	DAOS		861.22	
	3	ISC23	Leonardo	EuroHPC-CINECA	EXAScaler		807.12	
	4	SC23	Shaheen III	King Abdullah University of Science and Technology	Lustre		709.52	
	5	ISC25	Helma	Erlangen National High Performance Computing Center	Lustre		438.62	
New	6	SC25	CHIE-4	SoftBank Corp	EXAScaler		399.41	
	7	SC24	CHIE-3	SoftBank Corp	EXAScaler		331.66	
	8	ISC25	Miyabi-G	Joint Center for Advanced High Performance Computing	Lustre		319.00	
New	9	SC25	Blue Vela Storage Scale System 6000 - shared	IBM	IBM Storage Scale 5.2.2.1 with ESS 6.2.2.0		265.36	
	10	SC24	IRIS	MSKCC	WekaIO		252.54	

Production
Metadata Winner

IO500

	#	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE	BW (GIB/S)	MD ↑ (KIOP/S)
	1	SC23	Aurora	Argonne National Laboratory	DAOS		10,066.09	102,785.41
New	2	SC25	SuperMUC-NG-Phase2-EC	LRZ	DAOS		861.22	13,982.88
	3	ISC25	SSC-24	Samsung Electronics	WekaIO		248.67	2,749.41
	4	SC24	IRIS	MSKCC	WekaIO		252.54	1,753.69
	5	ISC24	Lise	Zuse Institute Berlin	DAOS		65.01	1,620.13
	6	ISC25	Helma	Erlangen National High Performance Computing Center	Lustre		438.62	1,604.84
	7	ISC22	CTPAI	China Telecom Research Institute	DAOS		25.29	1,395.01
	8	SC23	Shaheen III	King Abdullah University of Science and Technology	Lustre		709.52	895.35
	9	ISC25	HRT	Hudson River Trading	EXAScaler		136.05	890.51
New	10	SC25	CHIE-4	SoftBank Corp	EXAScaler		399.41	762.47

Production
Overall Winner

IO500

	# ↑	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE ↑	BW (GIB/S)	MD (KIOP/S)
	1	SC23	Aurora	Argonne National Laboratory	DAOS	32,165.90		
New	2	SC25	SuperMUC-NG-Phase2-EC	LRZ	DAOS	3,470.22		
	3	ISC25	Helma	Erlangen National High Performance Computing Center	Lustre	838.99		
	4	ISC25	SSC-24	Samsung Electronics	WekaIO	826.86		
	5	SC23	Shaheen III	King Abdullah University of Science and Technology	Lustre	797.04		
	6	SC24	IRIS	MSKCC	WekaIO	665.49		
	7	ISC23	Leonardo	EuroHPC-CINECA	EXAScaler	648.96		
New	8	SC25	CHIE-4	SoftBank Corp	EXAScaler	551.85		
	9	SC24	CHIE-3	SoftBank Corp	EXAScaler	500.20		
	10	ISC25	Miyabi-G	Joint Center for Advanced High Performance Computing	Lustre	391.60		

Production
Overall Winner

IO500

	# ↑	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
New	2	SC25	SuperMUC-NG-Phase2-EC	LRZ	DAOS	3,470.22	861.22	13,982.88
	3	ISC25	Helma	Erlangen National High Performance Computing Center	Lustre	838.99	438.62	1,604.84
	4	ISC25	SSC-24	Samsung Electronics	WekaIO	826.86	248.67	2,749.41
	5	SC23	Shaheen III	King Abdullah University of Science and Technology	Lustre	797.04	709.52	895.35
	6	SC24	IRIS	MSKCC	WekaIO	665.49	252.54	1,753.69
	7	ISC23	Leonardo	EuroHPC-CINECA	EXAScaler	648.96	807.12	521.79
New	8	SC25	CHIE-4	SoftBank Corp	EXAScaler	551.85	399.41	762.47
	9	SC24	CHIE-3	SoftBank Corp	EXAScaler	500.20	331.66	754.41
	10	ISC25	Miyabi-G	Joint Center for Advanced High Performance Computing	Lustre	391.60	319.00	480.72

IO500 Award Ceremony
10 Client Node Research List

10 Client Node Research
Bandwidth Winner

IO500

#	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	
New 6	SC25	SuperMUC-NG-Phase2-10	LRZ	DAOS		323.11	
7	SC21	OceanStor Pacific	Olympus Lab	OceanFS		317.07	
8	SC21	Athena	Huawei HPDA Lab	OceanFS		314.56	
9	ISC23	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng Laboratory	SuperFS		263.97	
10	ISC22	Cumulus	University of Cambridge	DAOS		216.78	

10 Client Node Research
Metadata Winner

IO500

#	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	7,705,448.04
2	ISC23	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng Laboratory	SuperFS		263.97	502,435.85
3	SC22	SuperStore	Tsinghua Storage Research Group	SuperFS		179.60	169,515.95
4	SC22	ParaStor	Sugon Cloud Storage Laboratory	ParaStor		718.11	106,042.93
5	SC22	StarStor	SuPro Storteck	StarStor		515.15	88,491.65
6	ISC22	Shanhe	National Supercomputing Center in Jinan	flashfs		207.79	60,119.50
7	ISC24	Songshan	Institute of Computing Technology Chinese Academy of Sciences and National Supercomputing Center in Zhengzhou	HiStore		43.78	41,580.79
8	SC21	Athena	Huawei HPDA Lab	OceanFS		314.56	18,235.71
9	SC22	HPC-OCI	Cloudam HPC on OCI	BurstFS		95.90	17,224.05
10	SC21	OceanStor Pacific	Olympus Lab	OceanFS		317.07	16,664.88

10 Client Node Research
Overall Winner

IO500

# ↑	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
New 10	SC25	SuperMUC-NG-Phase2-10	LRZ	DAOS	1,997.41	323.11	12,347.70

IO500 Award Ceremony
Research List

Research
Bandwidth Winner

IO500

#	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	SC25	SuperMUC-NG-Phase2	LRZ	DAOS		1,303.25	
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	

New

Research
Metadata Winner

IO500

#	RELEASE	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE	BW (GIB/S)	MD ↑ (KIOP/S)
1	ISC23	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng Laboratory	SuperFS		4,847.48	9,119,612.35
2	ISC23	Cheeloo-1 with OceanStor Pacific	JNIST and HUST PDSL	OceanFS2		2,439.37	7,705,448.04
3	SC22	SuperStore	Tsinghua Storage Research Group	SuperFS		179.60	169,515.95
4	SC23	Aurora	Argonne National Laboratory	DAOS		11,362.27	164,391.73
5	SC22	ParaStor	Sugon Cloud Storage Laboratory	ParaStor		718.11	106,042.93
6	SC22	StarStor	SuPro Storteck	StarStor		515.15	88,491.65
7	ISC22	Shanhe	National Supercomputing Center in Jinan	flashfs		207.79	60,119.50
8	ISC24	Songshan	Institute of Computing Technology Chinese Academy of Sciences and National Supercomputing Center in Zhengzhou	HiStore		43.78	41,580.79
9	SC21		Huawei Cloud	Flashfs		109.82	37,034.00
10	SC22	HPC-OCI	Cloudam HPC on OCI	BurstFS		278.48	33,033.54

Research
Overall Winner

IO500

# ↑	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
New 6	SC25	SuperMUC-NG-Phase2	LRZ	DAOS	6,308.87	1,303.25	30,540.36
7	SC22	SuperStore	Tsinghua Storage Research Group	SuperFS	5,517.73	179.60	169,515.95
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

IO500

bit.ly/io500poll

List of Awarded Systems in the Ranked Lists

10 Client Production	Bandwidth Metadata Overall	**Argonne National Laboratory**	DAOS	734,50 11,336.72 2,885.57	GB/s KIOPS/s score
10 Client Research	Bandwidth Metadata Overall	**JNIST and HUST PDSL**	OceanFS2	2,439.37 7,705,448.04 137,100.00	GB/s KIOPS/s score
Production	Bandwidth Metadata Overall	**Argonne National Laboratory**	DAOS	10,066.09 102,785.41 32,165.90	GB/s KIOPS/s score
Research	Bandwidth Metadata Overall	**Argonne National Laboratory** **Pengcheng Laboratory** **Pengcheng Laboratory**	DAOS SuperFS SuperFS	11,362.27 9,119,612.35 210,255.00	GB/s KIOPS/s score

IO500 List Analysis

SC25 ST. LOUIS hpc ignites.

Growth in Submissions by Year

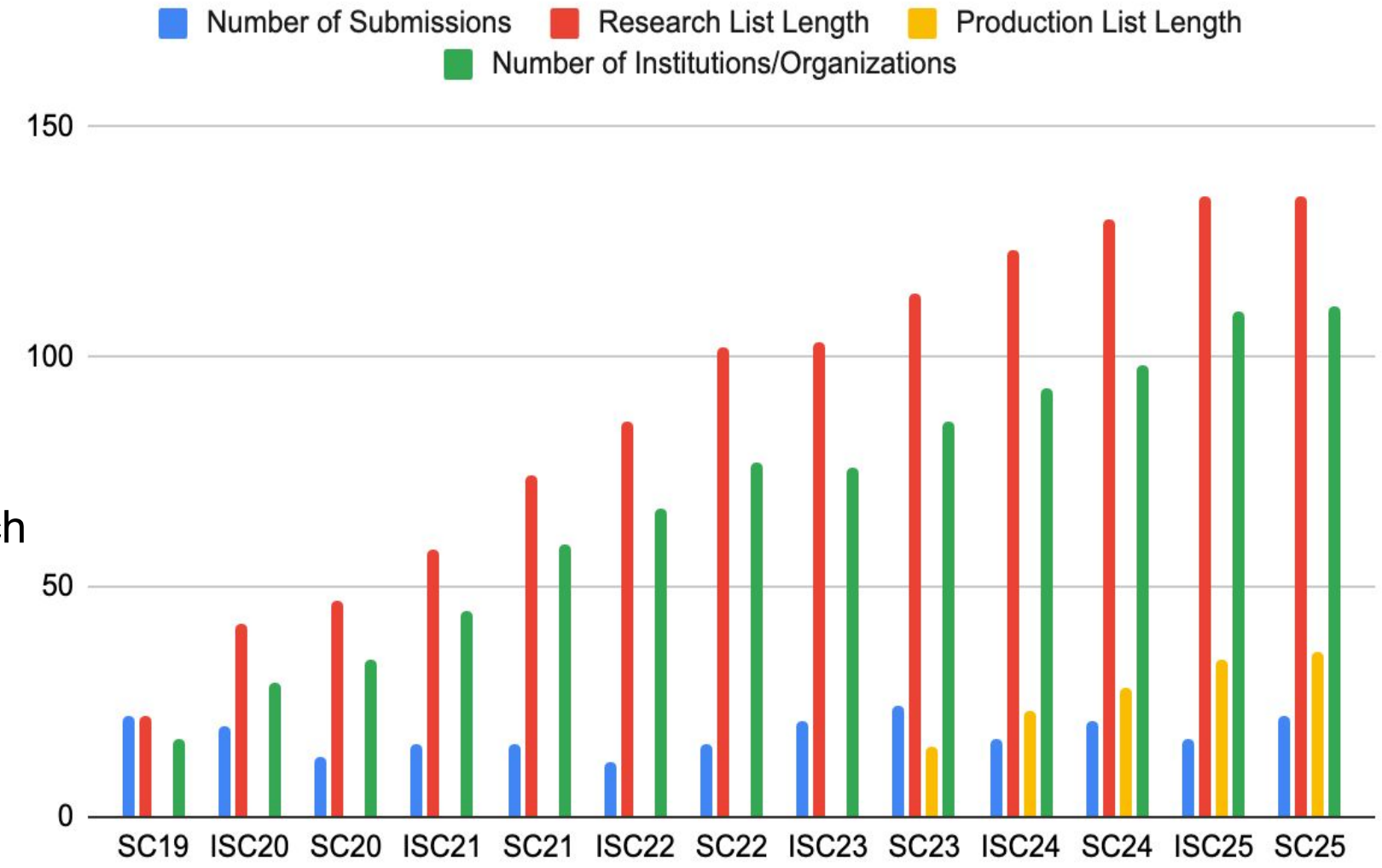
SC25

22 new submissions added

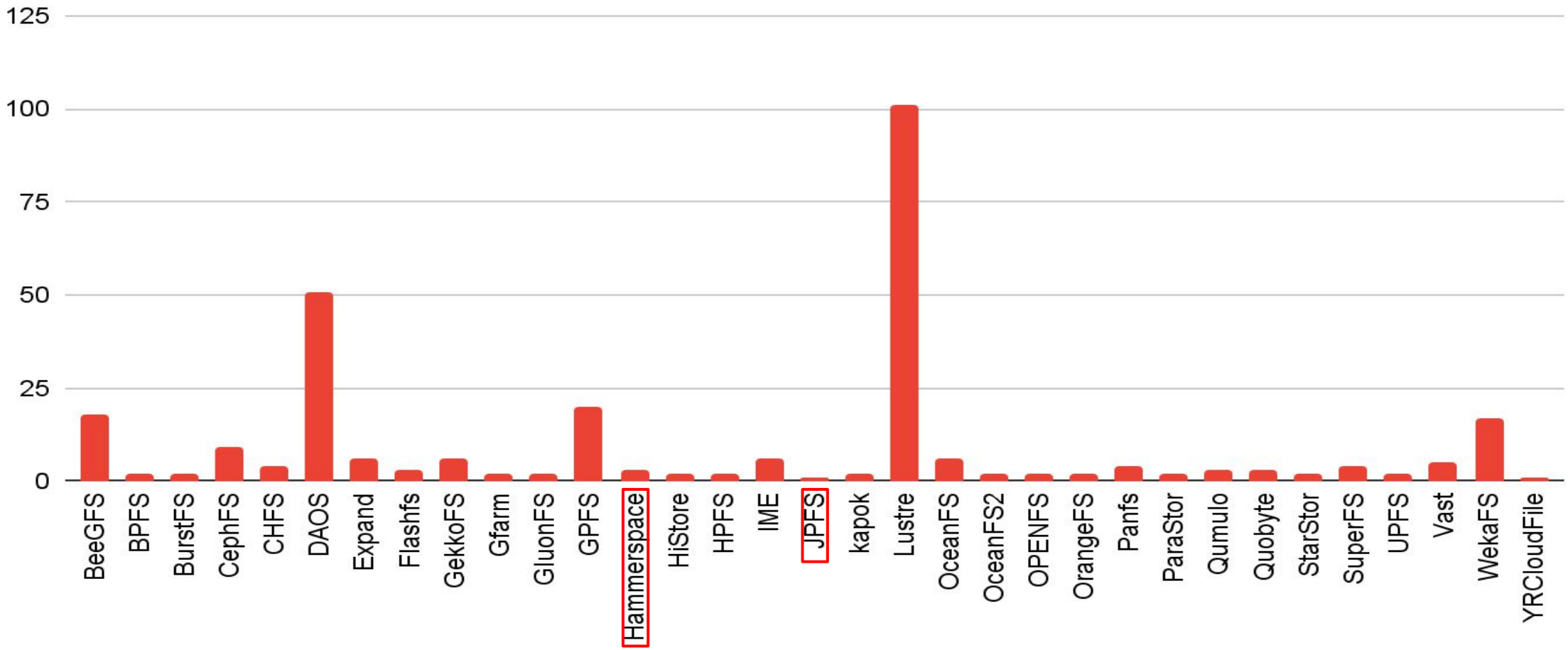
- 3 new on Prod list
- 10 new on 10-Client Prod
- 1 new on Research list
  - 1 was not accepted
- 5 new on 10-Client Research
- 3 new **only** on Full list
  - too few clients to rank

Around 300 list entries

More than 100 institutions

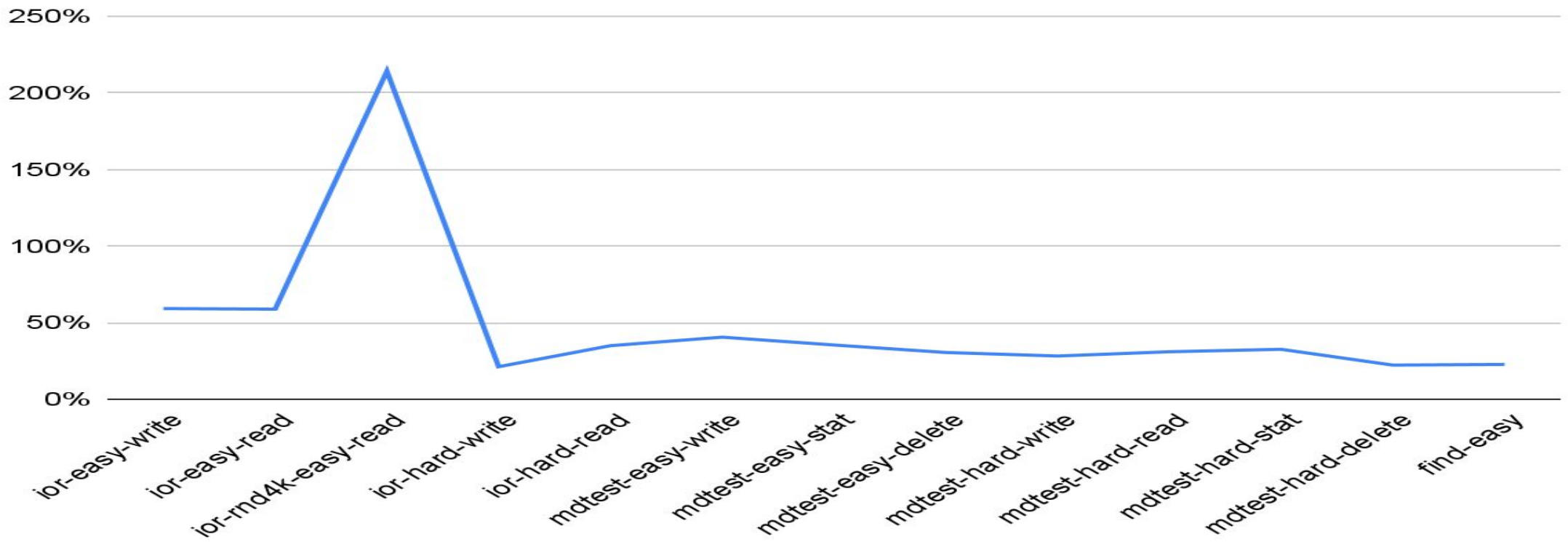


Filesystem Types in Submissions - 33 in Total



2 new storage systems for the first time - pNFS (Hammerspace), JPFS

Growth of Phase Totals SC'25 vs. ISC'25 Submissions



- 3/24 submissions vs. 13/33 have valid **ior-rnd4k-easy-read**
- Aggregate bandwidth (**ior-easy-***) grew almost twice metadata ops rate

Comparison of ior-rnd4k-easy-read vs. Existing Scores

	rnd4k / ioeasy-write	rnd4k / ioeasy-read	rnd4k / iohard-write	rnd4k / iohard-read	rnd4k / mdhard-read	rnd4k / mdhard-write
Avg (mean)	16%	12%	247%	63%	65%	25%
Stddev	21%	14%	**410%**	134%	179%	80%
Geomean	7.4%	6.1%	68%	22%	7.5%	2.3%
Min	0.3%	0.3%	2.2%	2.1%	0.01%	0.01%
Max	73%	45%	**1423%**	551%	951%	337%
Spread	242	157	645	257	**123311**	**41702**

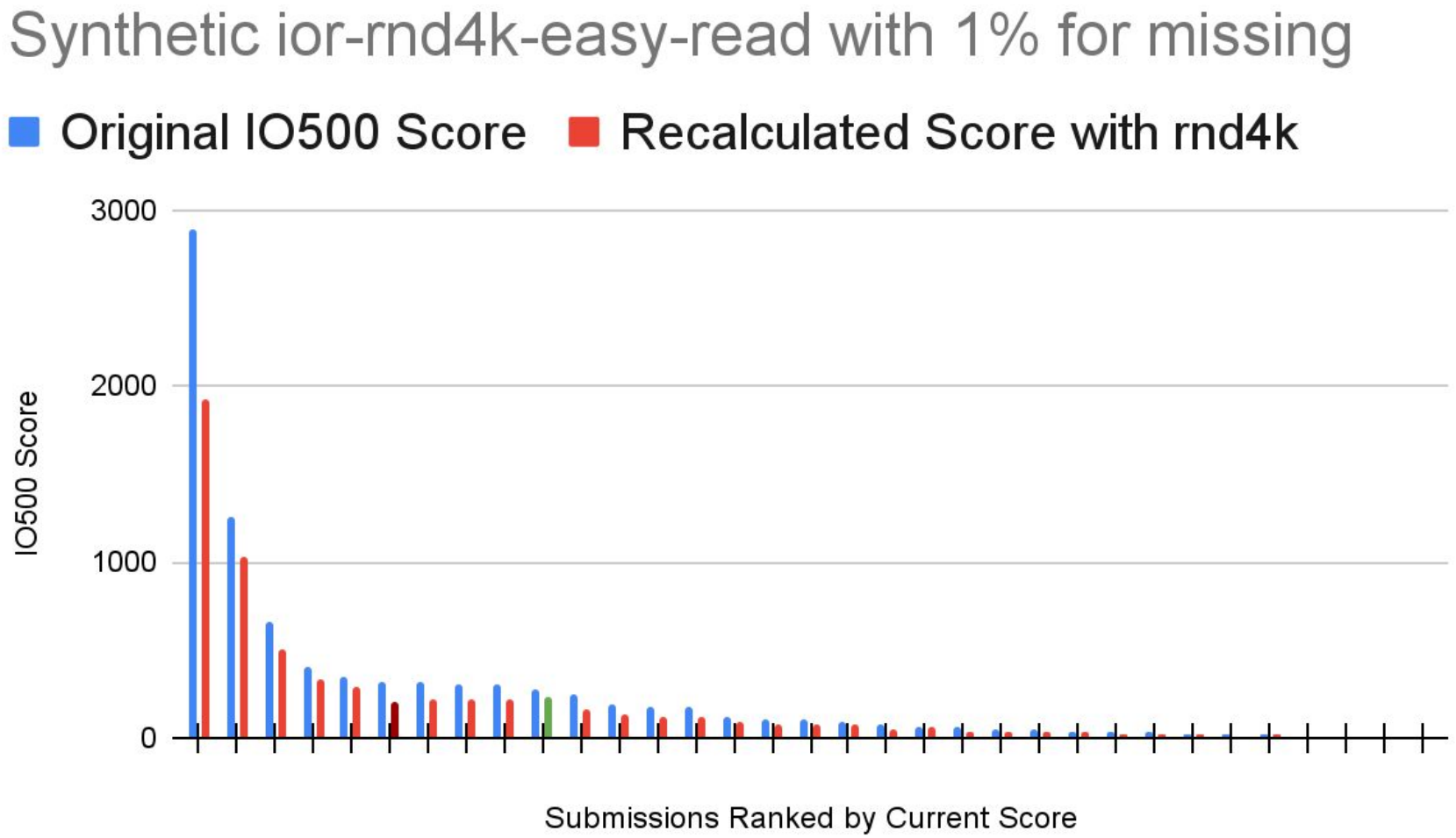
- Looking for correlation between existing phases and **rnd4k**
  - Some submissions had better **rnd4k** than **hard-read**!
  - Strongest correlation seen with **rnd4k/easy-read** (stddev, spread)
- Want to minimize old scores that benefit from no **rnd4k**
  - 90% of submissions have **rnd4k/easy-read** at least 1%

Hypothetical Changes to 10-Client Production Ranking

IO BW	Score	IO BW + rnd4k	Score + rnd4k
734.5	2885.6	**328.5**	1929.7
254.0	1253.6	170.9	1028.3
247.9	666.7	141.3	503.3
117.4	405.5	79.2	333.1
136.1	348.1	92.5	287.1
65.0	324.5	**27.7**	211.7
154.7	314.0	**79.9**	225.6
148.9	303.2	82.0	225.0
159.9	299.3	**88.3**	222.4
79.4	273.8	57.7	233.5
124.9	243.6	**59.6**	168.3
77.4	188.3	**37.7**	131.3
133.0	181.9	**60.7**	122.9
62.6	176.6	**31.9**	126.0
76.6	125.0	**37.5**	87.5

- Top-10 list ranking mostly unchanged
- 4 of top 10 missing **rnd4k** score
- Two entries in SC25 Top-10 would swap
  - Old result penalized due to synthetic 1% score
    - **Average 70% of IO500 score with synthetic rnd4k**
  - New result has strong **rnd4k** score
    - Average 80% of IO500 score with actual **rnd4k**
  - Rest of scores have enough margin to stay
- 10-Client Prod list has 6 results with **rnd4k**
- Production list has 3 results with **rnd4k**

Hypothetical 10-Node Prod Ranking with ior-rnd4k-easy-read



IO500

bit.ly/io500poll

Community Talk
Using IO500 for Storage System Sizing

Michael Hennecke (HPE)

Prologue – Some Trivial Observations (and Opinions)

- To do **sizing**, we need to understand **scaling** behaviour
- For NVMe storage, **PCIe generation** drives bandwidth evolution
  - Per-port **network** speed, per-device **NVMe** bandwidth
- **Absolute** numbers depend on budgets, **per-server** numbers inform about technology
- **Min. Time to Disk Full** starts to prohibit stonewall=300sec runs
  - 3.84 TB divided by 6 GB/s = **640 sec** (getting worse with >10 GB/s write)
  - **PLEASE** do not add more write phases (or reduce the stonewall time)…
- IO500's **SCORE** (absolute, or per-server) is *not* useful for sizing
  - Its equal weighting will likely cause "mis-allocation of capital"
- **Find** phase is *meaningless* (and heavily skews the scores)

Performance Scaling with Client and Server Resources

Client-side variables determining performance:

- Number of network ports for bandwidth, number of CPU cores for IOPS
- Must scale out the number of clients until storage is saturated
- **Total number of client processes** is the correct metric for the “x-axis”

Server-side variables determining performance:

- Number of PCIe lanes – goal is to *balance* NVMe and network BW (here: **2x NDR**)
- **Number** and model of **NVMe drives**
- **Number** of **CPU cores**
  - In DAOS, a **“target”** is a user-level thread running on a physical CPU core (“tgt” in graphs)
  - Different-coloured lines in graphs: # of targets per NVMe disk key to size servers’ CPU model
- Scale out by adding servers – not studied here (graphs below are for a single server)

35

LRZ @ Research List

(SX on 42 servers @ 8 NVMe)

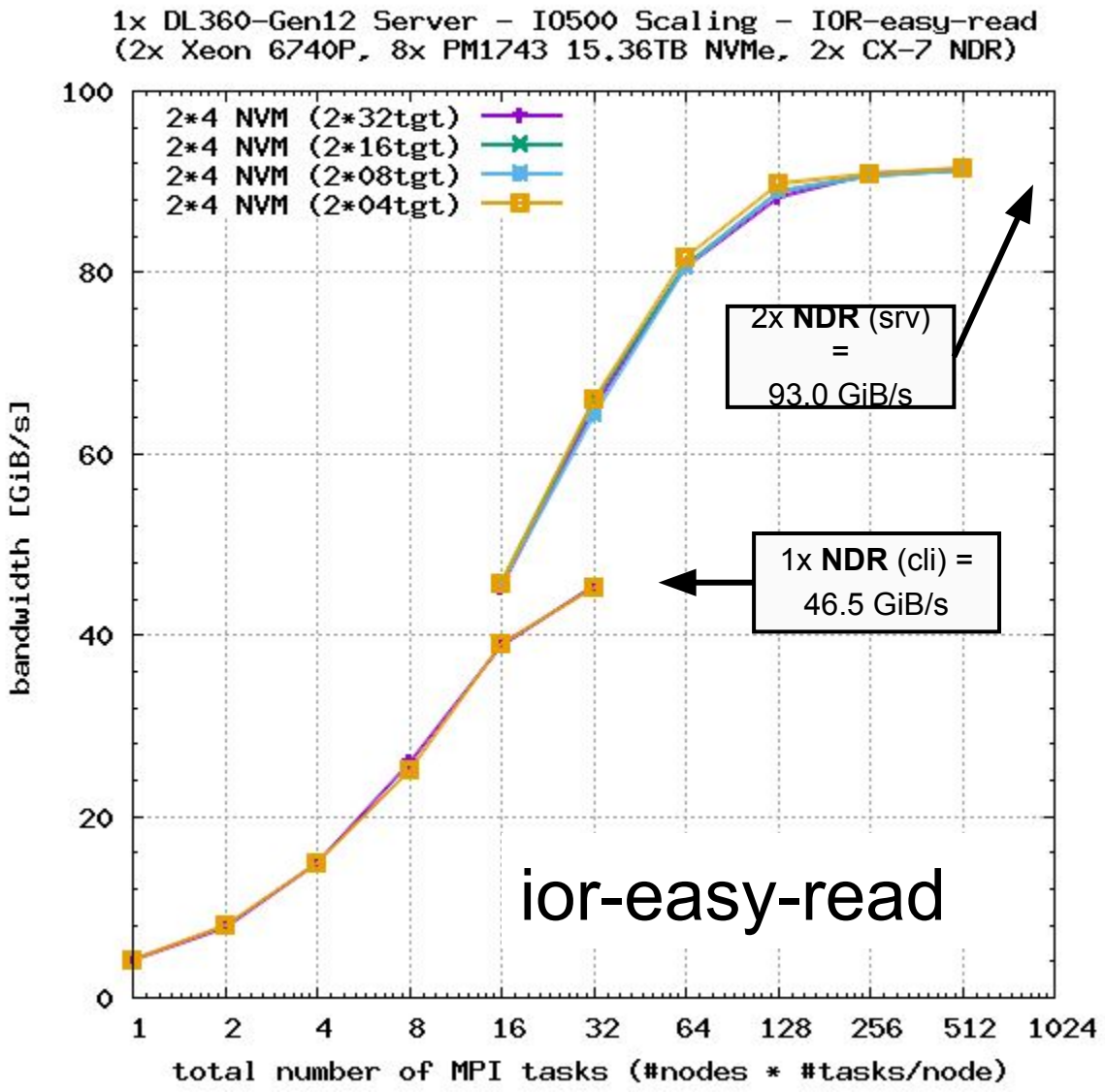
		SC23	SC25
		90*72=6480 tasks	192*112=21504 tasks
[RESULT]	ior-easy-write	1081.065152 GiB/s	1130.084774 GiB/s
[RESULT]	mdtest-easy-write	28285.010669 kIOPS	46279.405319 kIOPS
[RESULT]	ior-hard-write	854.092711 GiB/s	981.241225 GiB/s
[RESULT]	mdtest-hard-write	11326.412741 kIOPS	19020.539326 kIOPS
[RESULT]	find	21144.493586 kIOPS	15026.685891 kIOPS
[RESULT]	ior-easy-read	1854.753978 GiB/s	1785.866162 GiB/s
[RESULT]	mdtest-easy-stat	31709.921027 kIOPS	49871.905565 kIOPS
[RESULT]	ior-hard-read	722.621314 GiB/s	1456.731978 GiB/s
[RESULT]	mdtest-hard-stat	26079.516275 kIOPS	44756.150608 kIOPS
[RESULT]	mdtest-easy-delete	14607.461557 kIOPS	26396.123269 kIOPS
[RESULT]	mdtest-hard-read	19814.883537 kIOPS	35932.392537 kIOPS
[RESULT]	mdtest-hard-delete	15397.518911 kIOPS	27026.142118 kIOPS
[ ]	ior-rnd4K-easy-read	n/a	189.112448 GiB/s
[SCORE]	Bandwidth	1054.723179 GiB/s	1303.253329 GiB/s
	IOPS	19937.454838 kiops	30540.356635 kiops
	TOTAL	4585.683783	6308.868477

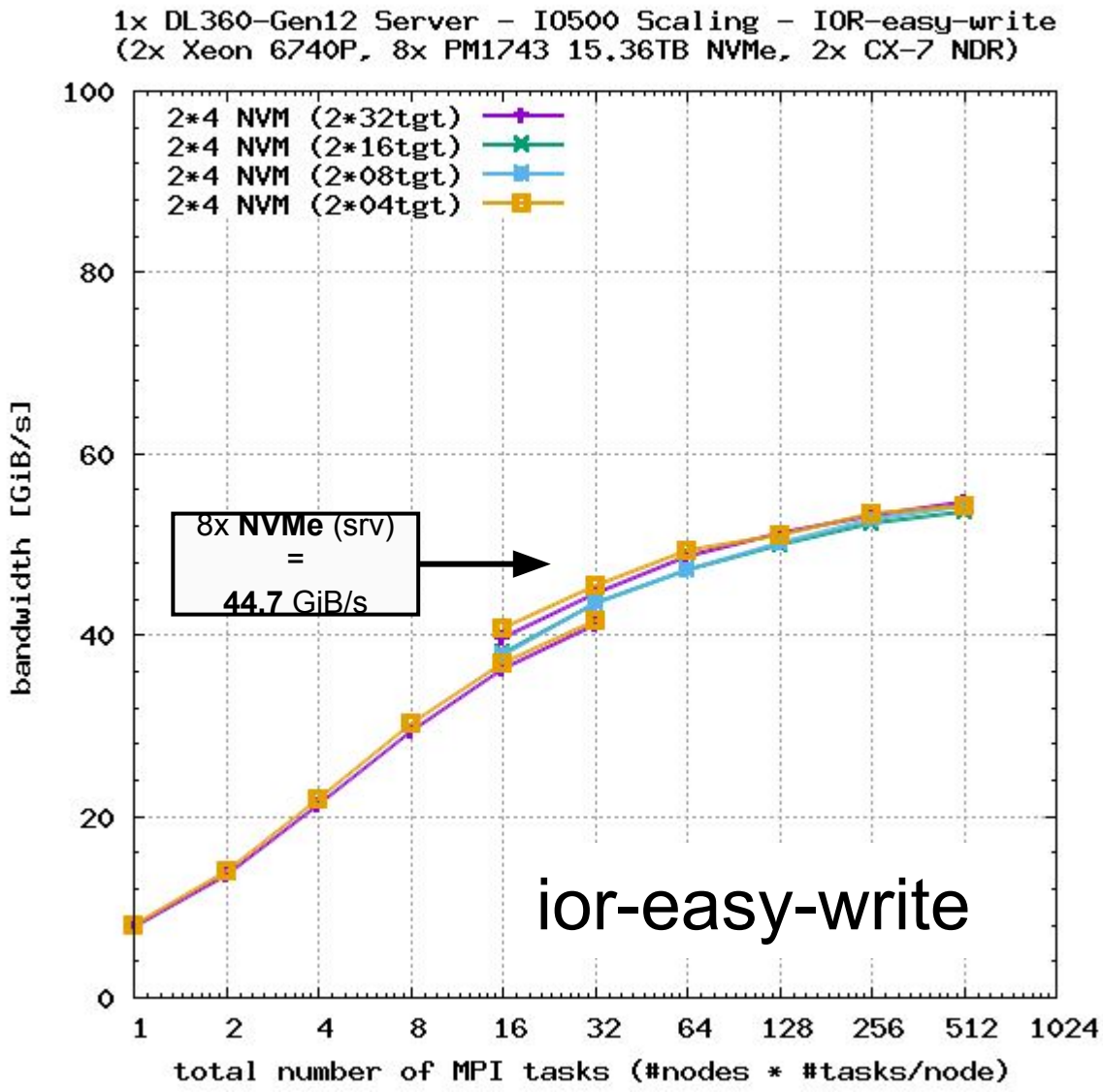
LRZ @ Production List

(EC_16P1GX on 42 servers @ 8 NVMe)

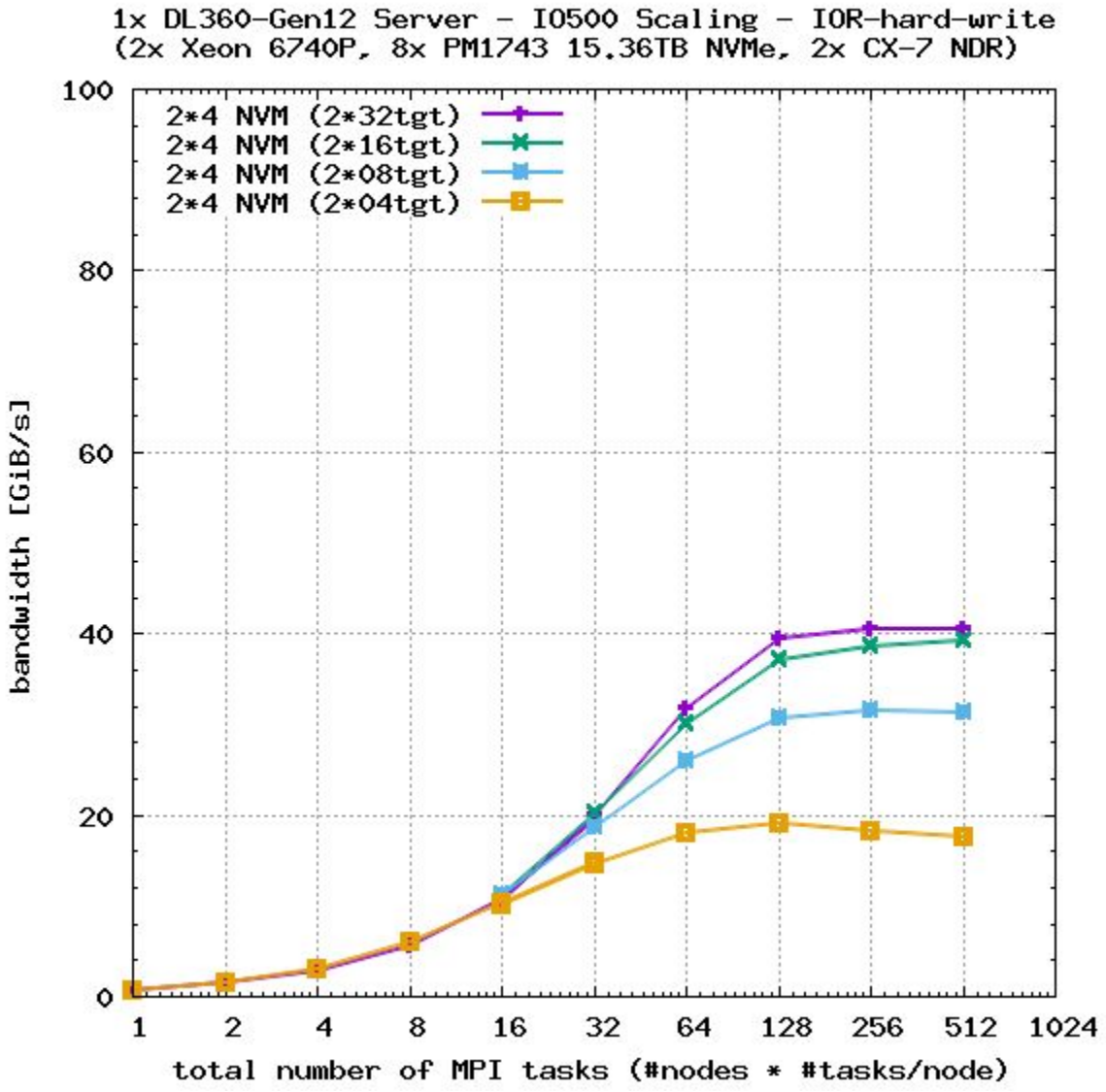
		SC23 90*72=6480 tasks	SC25 192*112=21504 tasks
[RESULT]	ior-easy-write	896.708153 GiB/s	1028.084806 GiB/s
[RESULT]	mdtest-easy-write	6324.788102 kIOPS	11128.689155 kIOPS
[RESULT]	ior-hard-write	252.427284 GiB/s	195.535950 GiB/s
[RESULT]	mdtest-hard-write	2644.926530 kIOPS	4643.273444 kIOPS
[RESULT]	find	12733.442991 kIOPS	10358.107562 kIOPS
[RESULT]	ior-easy-read	1872.091759 GiB/s	1836.458514 GiB/s
[RESULT]	mdtest-easy-stat	29403.338203 kIOPS	46968.114658 kIOPS
[RESULT]	ior-hard-read	718.806938 GiB/s	1490.144590 GiB/s
[RESULT]	mdtest-hard-stat	23242.010086 kIOPS	44099.698465 kIOPS
[RESULT]	mdtest-easy-delete	3442.670418 kIOPS	6473.046718 kIOPS
[RESULT]	mdtest-hard-read	17023.129123 kIOPS	30522.888272 kIOPS
[RESULT]	mdtest-hard-delete	3112.592330 kIOPS	6671.944725 kIOPS
[ ]	ior-rnd4K-easy-read	n/a	218.311055 GiB/s
[SCORE]	Bandwidth	742.902297 GiB/s	861.224294 GiB/s
	IOPS	8472.598104 kiops	13982.884828 kiops
	TOTAL	2508.846865	3470.216148

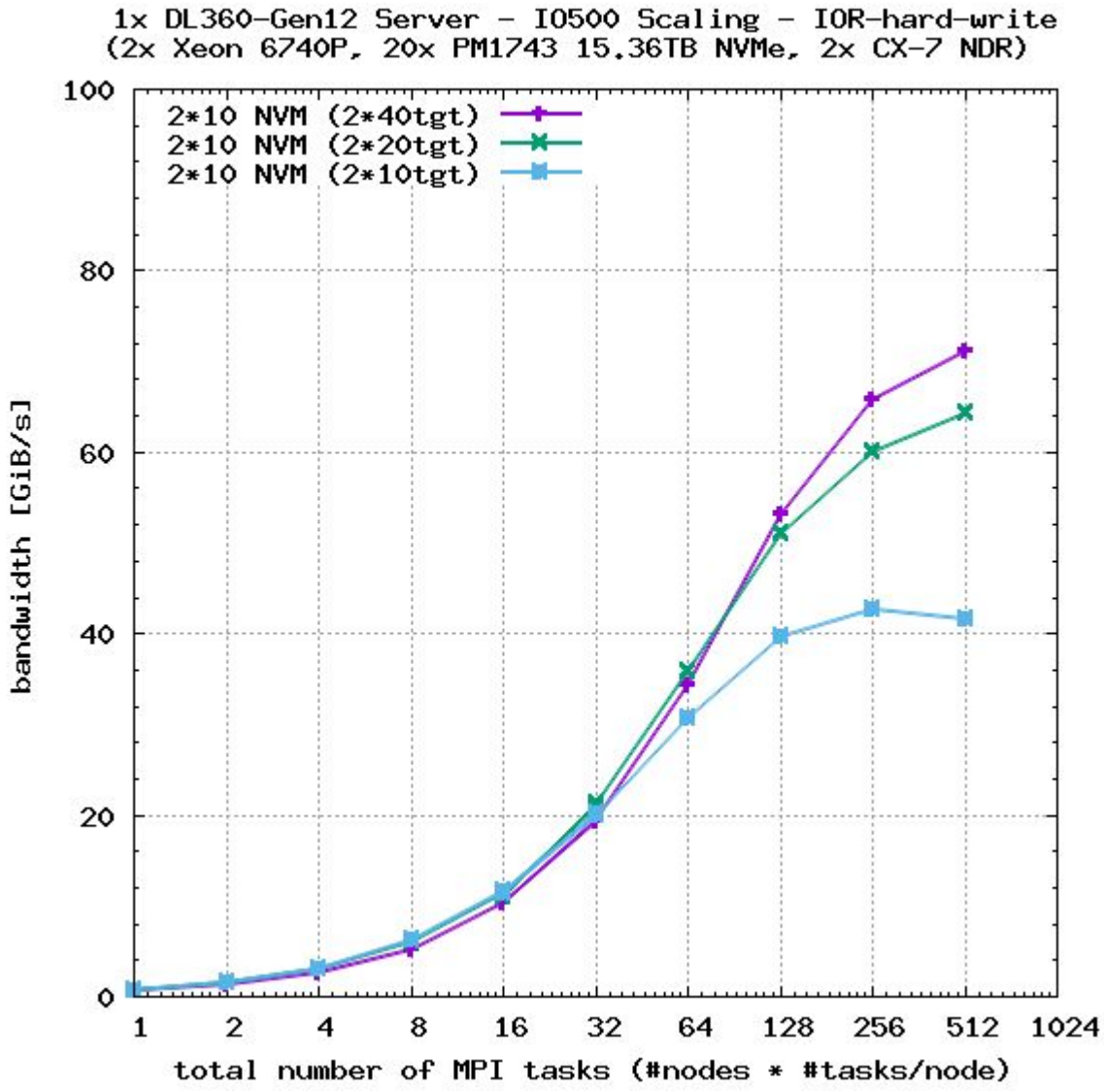
DAOS on DL360-Gen12 (NDR): ior-easy-{read,write} on 8x NVMe



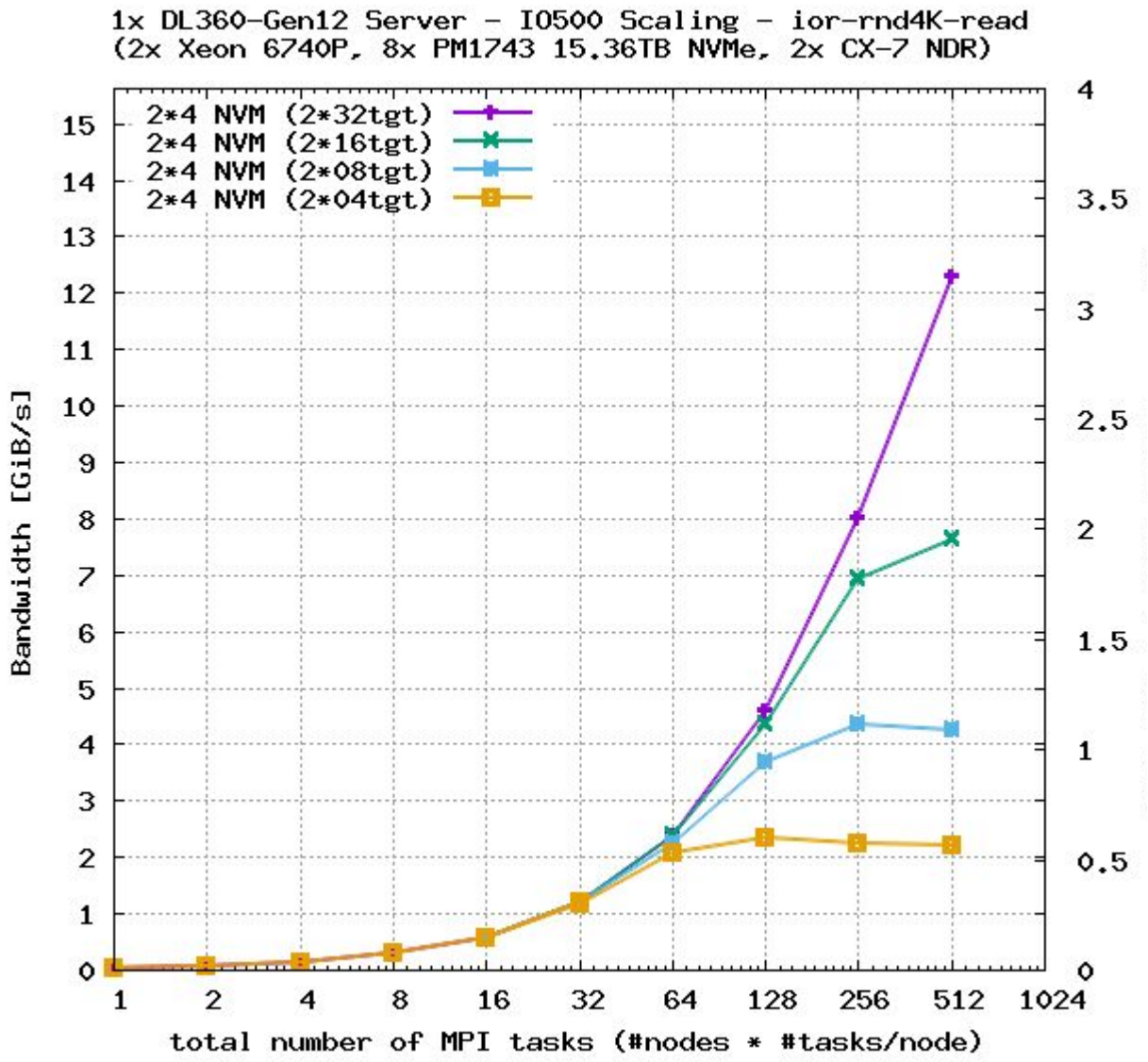


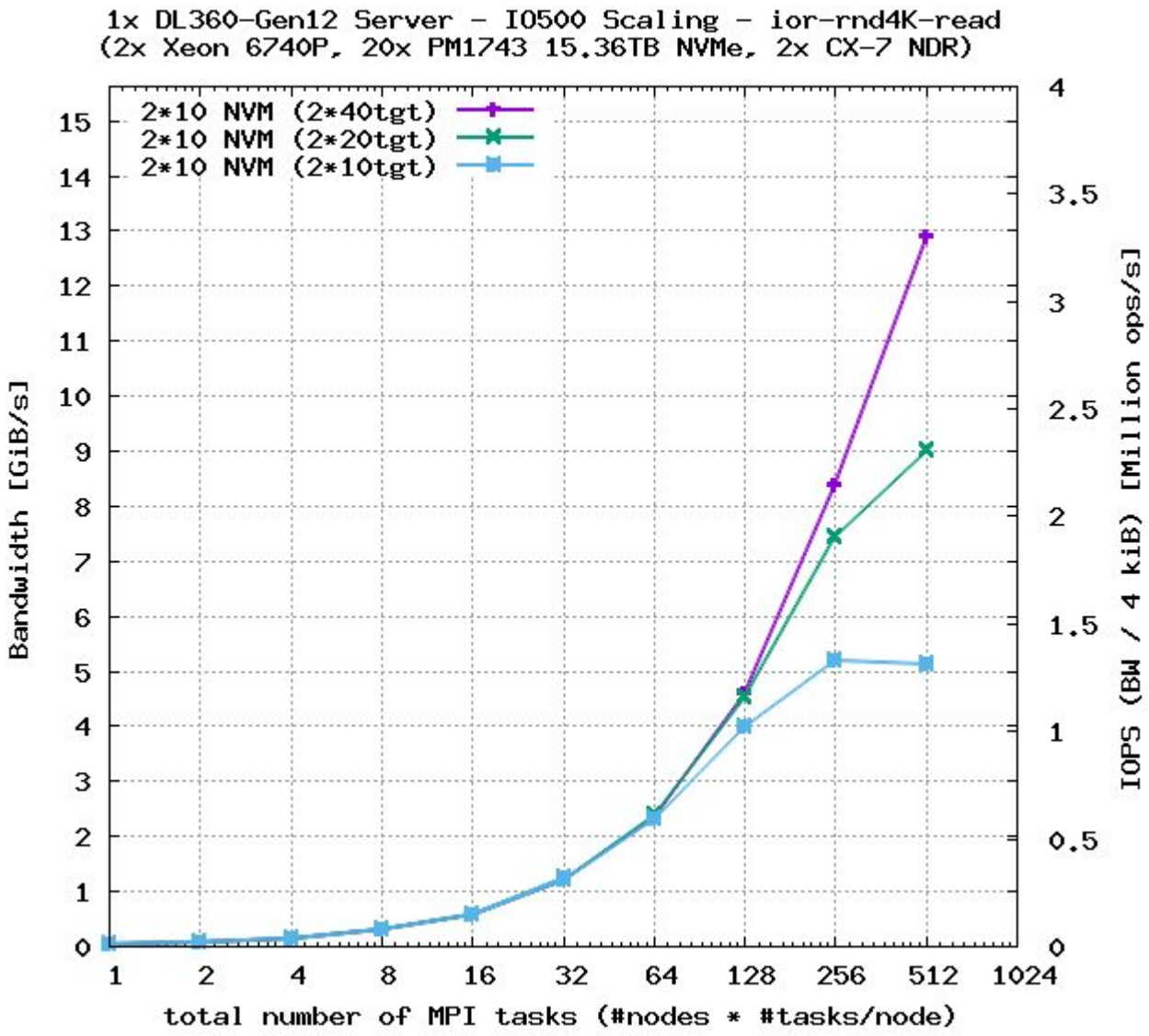
DAOS on DL360-Gen12 (NDR): ior-hard-write on 8x (left) and 20x (right) NVMe





DAOS on DL360-Gen12 (NDR): ior-rnd4K-read on 8x (left) and 20x (right) NVMe





For more information:

- **SC-Asia 2023 paper** : *Understanding DAOS Storage Performance Scalability* https://doi.org/10.1145/3581576.3581577

- **CUG 2025 paper** : *Enhancing RPC on Slingshot for Aurora's DAOS Storage System* https://doi.org/10.1145/3757348.3757350

Thank you !

Updates

SC25 ST. LOUIS hpc ignites.

Random Read Phase Update

- ISC25 was first run with any **ior-rnd4K-easy-read** results
  - Several issues occurred due to IOR bugs
- SC25 is first **full** run with **ior-rnd4K-easy-read** results!
- Workload summary
  - Reuse existing **ior-easy-write** files for input to avoid writing new files
  - Total dataset size is the largest available from previous phases to minimize cache
  - No data verification needed, was done during **ior-easy-read** already
  - Run at end of other phases to avoid conflicting with existing phases/scores
    - Hard stonewall at 300s (with wearout) to limit increase in runtime

Random Read Phase Scoring

- Scoring
  - Not currently utilized to compute final score
  - Reported as bandwidth to allow comparison to other IOR phases
- Next steps to include it into benchmark runs/score
  - Run analysis of **ior-rnd4k-easy-read** score to see how it affects SC25 results
    - Both as bandwidth and as IOPS
  - Our proposals:
    - When 6 of top 10 entries of each list have random-read results, trigger move to new ranking
    - Assign entries without random read scores a value of 1% of **easy-read**

Random Read Poll

Now that we have rnd4k phase, when should we transition to include it in the score?

A. Next list release at ISC'26 (all lists)

B. Next list release at SC'26 (all lists)

C. (recommended) When at least 6 entries in each Top 10 list have valid rnd4k score

D. When all 10 entries in each Top 10 list have valid rnd4k score

bit.ly/io500poll

How do you prefer old submissions that do not have those scores to be handled?

A. Drop them from new ranked list releases (eliminates all submissions before ISC'25)

B. *(recommended) Use 1% of **ior-easy-read** for **rnd4k** (worse than 90% of all submissions)*

C. Use another synthetic **ior-rnd4k-easy-read** score, but at (some) disadvantage vs. actual scores

Website Updates

- Continuing to work on how to simplify:
  - Submission process
  - Improve access to all fields for data analysis of prior submissions
    - e.g., flat schema export
- Work to disambiguate what data is expected in submission fields
- Always looking for volunteers to help!

Proposal: Community Guidelines

- **Community Participation Guidelines**
  - No issues so far in the IO500 community, but want to ensure we are proactive
  - In worst case we could remove offenders from Slack/email/etc.

- **IO500 List and Data Usage Guidelines**
  - Ensure IO500 lists, rankings, submission data, are used in an accurate and fair way
  - Goal is to ensure some teeth on enforcement
  - Start with a request for correction, …, lead to list removal in the worst case

- Watch out for proposals, please help us review

- MLPerf is much further down the path here…

Proposal: New find-hard Phase

- **Spirit of the ‘find’ phase is to represent real-world workloads**
  - Finding files to backup/delete/etc, general user queries
  - Any optimization specific to the benchmark is again the ‘spirit’
- **Current find phase can be circumvented too easily**
  - Offload all searching to the server, precreate indexes to match
  - Benchmark metadata can fit into server RAM, does not show true cold-cache speed
    - In real-world, old files might not be in RAM
- Current thought
  - Run multiple finds with different (varying?) arguments to defeat index?
    - Some searches against **mdtest-easy-create**, some against **mdtest-hard-create**?
    - Random values could penalize some results depending on number of matches
  - Output find results to a file in the storage to allow further analysis
    - Better matches actual usage case (e.g. list of files to be accessed or deleted)

Open Discussion

bit.ly/io500poll

SC25 ST. LOUIS hpc ignites.

IO500

bit.ly/io500poll

BIRDS OF A FEATHER

IO500: The High-Performance Storage Community

Jean Luca Bez – Lawrence Berkeley National Laboratory
Andreas Dilger – The Lustre Collective
Dean Hildebrand – Google
Julian Kunkel – Georg-August-Universität Göttingen/GWDG
Jay Lofstead – Sandia National Laboratories
George Markomanolis – AMD