

## Analyzing Parallel I/O

Data analysis news from GWDG and analysis tasks in the Scalable Storage Competition

Christian Boehme, Sebastian Krey, Julian Kunkel



- GWDG provides HPC services on national level in Germany
  - National supercomputing center within NHR alliance  
NHR-Nord@Göttingen
  - One of four national AI service centers (KISSKI)
  - HPC service center for research organizations DLR, MPG, UGö, ...
  - 2000+ nodes, heterogeneous landscape
- Our focus: Security and availability - performance second
- Central and compute island specific storage

## Storage

- Lustre (Highest performance, cluster specific)
- VastData Storage (HOME, some AI applications, central)
- Ceph (Warm storage, capacity, central)
- BeeGFS (Tier 3)
- Being phased out
  - StorNext (Tier 3 HOME and warm storage)
  - GPFS (Tier 2 HOME)

## Network

- Omni-Path (mainly CPU nodes)
- Infiniband (mainly GPU nodes)
- Ethernet (smaller clusters, Ceph storage)

- **Discovering** application I/O patterns
- **Optimizing** storage system setup
- Fostering understanding of I/O in **teaching and training**

- Monitoring stacks on different systems utilized
  - TIG (Telegraph, Influx, Grafana)
  - Prometheus + Grafana
- Typical I/O metrics monitored
  - Network, e.g.:
    - Bandwidth used per second, Errors
    - TCP connections
  - Storage, e.g. (using Lustre jobstats, BeeGFS userstats or VAST dataflow statistics):
    - Metadata OPs (general, per user or job)
    - read/writes (general, per user or job)
    - read sizes/write sizes (general, per user or job)
    - usage/utilization (general, per user or job)
    - inodes utilization, disk errors
- Data is utilized by admins to identify issues

- **Monitoring stack unification**
  - Context: system unification
  - Goal: move to one monitoring stack to reduce admin effort
- **Extend collected I/O Metrics**
  - e.g. IP connections BeegFS etc.
  - Challenge: impact on system performance to be avoided
- **Extend standard metrics collection with IOFS**
  - provide users with information about application IO behaviour
  - utilization for regression tests from user POV
- **Add automatic analysis methods**

## Goals:

- Increase transparency of I/O behaviour of the systems
- Provide analysis tools to admins/users
- Base for IO optimization/regression testing of the systems

## Capacity cluster (Lustre HDD replacement)

- 45 servers, 2x24 Cores Sapphire Rapids, 512 GB memory, 2x25G ethernet
- $45 \times 24 \times 20T = 21.6P$  HDDs,  $45 \times 4 \times 7.68T = 1.4P$  NVMEs for WAL+DB and metadata pool
- Erasure coding 8+3 for data pools, 3x replication for metadata
- Usage for CephFS and S3

## NVME cluster (Central storage)

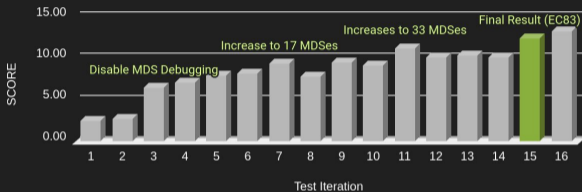
- 8 server, 2x32 Cores Milan, 512GB memory, 100G ethernet
- $8 \times 20 \times 15T = 2.4P$  NVMEs, 3x replication
- Usage for CephFS, S3 and dCache pool

We contracted Clyso to optimize it, demonstrating the need of system optimization

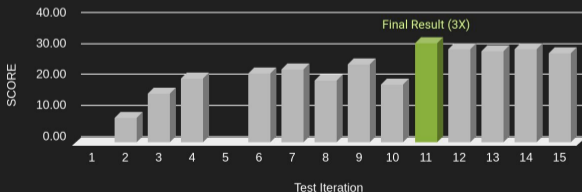
## Case Study: GWDDG

CLISO

### 1080 HDD Ceph Cluster IO500 Score



### 160 NVMe Ceph Cluster IO500 Score





# Ceph IO500 results HDD cluster

CLYSO	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10	Test 11	Test 12	Test 13	Test 14	Test 15	Test 16
Client Nodes	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
MPI Ranks	256	256	256	256	256	256	256	256	256	256	256	512	256	256	256	256
Active MDSes	4	4	4	4	8	9	17	17	17	17	33	33	33	33	33	33
Standby MDSes	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Replication	3	3	3	3	3	3	3	3	3	3	3	3	ECB3	ECB3	ECB3	3X
mdtest-easy Pinning Strategy	N-1 RR	N-1 RR	N-1 RR	RR	RR	N-1 RR	N-1 RR	N-1 RR	N-1 RR	N-1 RR	N-1 RR	N-1 RR	N-1 RR	N-1 RR	N-1 RR	N-1 RR
Meta PGs	128	2048	2048	2048	2048	2048	2048	2048	2048	2048	2048	2048	2048	2048	2048	2048
Data PGs	512	8192	8192	8192	8192	8192	8192	8192	8192	8192	16384	16384	16384	2048	16384	16384
debug_mds	10	10	default (1)	default (1)	default (1)	default (1)	default (1)	default (1)	default (1)	default (1)	default (1)	default (1)	default (1)	default (1)	default (1)	default (1)
mds_bal_interval	default (10)	default (10)	default (10)	default (10)	default (10)	default (10)	default (10)	0	5	5	5	5	5	5	5	5
CPU Turbo	Off	Off	Off	Off	Off	Off	Off	Off	Off	Off	Off	Off	Off	Off	Off	Off
Result Directory	2024.08.04-15.20.30	2024.08.04-19.00.44	2024.08.04-20.36.47	2024.08.05-02.42.10	2024.08.05-04.28.04	2024.08.05-05.45.36	2024.08.05-06.56.02	2024.08.05-06.56.02	2024.08.05-13.12.33	2024.08.05-15.32.15	2024.08.05-16.58.45	2024.08.05-17.58.48	2024.08.09-21.04.57	2024.08.10-03.13.03	2024.08.17-21.41.20	2024.08.17-22.57.05
ior-easy-write (GiB/s)	21.02	23.98	24.13	23.99	24.04	24.04	23.95	24.17	24.10	24.22	24.07	24.07	23.00	23.14	23.21	23.75
mdtest-easy-write (kiOPS)	3.05	2.98	17.36	21.15	32.21	43.16	82.34	79.56	83.07	80.42	156.05	157.90	178.24	173.06	191.39	266.39
timestamp (kiOPS)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ior-hard-write (GiB/s)	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.03	0.04	0.04	0.03	0.03	0.03	0.04
mdtest-hard-write (kiOPS)	2.20	2.14	11.69	9.59	7.89	13.80	7.85	3.69	7.34	7.43	8.15	8.05	10.67	6.74	12.62	12.94
find (kiOPS)	15.89	26.60	117.33	340.63	571.43	340.62	677.20	656.19	583.13	563.73	1365.82	1191.74	1079.99	936.47	1133.54	1546.63
ior-easy-read (GiB/s)	24.26	49.69	47.63	44.40	40.78	40.40	42.24	38.99	43.67	46.46	44.44	38.23	42.82	43.93	49.65	52.98
mdtest-easy-stat (kiOPS)	11.09	11.23	82.86	113.32	132.33	152.96	169.01	207.66	183.78	187.21	195.60	168.91	164.66	177.71	221.15	207.93
ior-hard-read (GiB/s)	0.30	0.20	0.22	0.23	0.24	0.23	0.23	0.23	0.24	0.21	0.23	0.23	0.25	0.20	0.22	0.19
mdtest-hard-stat (kiOPS)	7.29	7.44	36.79	45.01	55.68	74.20	47.01	40.63	43.24	43.69	103.13	51.68	68.12	52.30	100.58	116.40
mdtest-easy-delete (kiOPS)	1.83	1.80	11.16	12.16	12.92	26.66	45.62	43.99	46.17	44.76	67.06	67.13	85.84	91.90	104.92	124.67
mdtest-hard-read (kiOPS)	2.81	2.60	14.09	13.84	25.22	13.74	29.56	6.62	37.85	37.15	61.51	49.43	38.12	51.98	65.75	59.55
mdtest-hard-delete (kiOPS)	0.87	0.91	7.90	5.32	7.73	5.79	7.40	3.93	8.85	7.79	5.04	4.35	6.79	7.46	11.40	10.56
SCORE	2.46	2.65	6.42	7.01	7.93	8.12	9.35	7.79	9.51	9.13	11.23	10.09	10.41	10.06	12.44	13.28

# Ceph IO500 results SSD cluster

CLYSO	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 8 (rep 1)	Test 8 (rep 2)	Test 9	Test 10	Test 11	Test 12	Test 13
OSDs	160	160	160	160	160	160	160	160	160	160	160	159	159	159	159
Client Nodes	14	14	14	20	20	18	18	18	18	18	18	18	18	18	18
MPI Ranks	336	14	224 (wrong pin)	260	240	270	270	270	270	270	270	270	270	270	270
Active MDSes	14	14	14	14	28	28	28	28	28	28	28	28	28	28	28
Standby MDSes	2	2	2	2	4	4	4	4	4	4	4	4	4	4	4
Replication	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
mdtest-easy Pinning Strategy	N-1 RR	N-1 RR	N-1 RR	N-1 RR	N-1 RR	N-1 RR	N-1 RR	N-1 RR	N-1 RR	N-1 RR	N-1 RR	N-1 RR	N-1 RR	N-1 RR	N-1 RR
lor-hard Pinning Strategy	none	none	none	none	none	none	rank 0	rank 0	rank 0	rank 0	rank 0	rank 0	rank 0	rank 0	rank 0
Meta PGs	512	512	512	512	512	512	512	512	512	512	512	512	512	512	512
Data PGs	4096	4096	4096	4096	4096	4096	4096	4096	4096	4096	4096	4096	4096	4096	4096
debug_mds	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mds_bal_interval	default (10)	default (10)	default (10)	default (10)	default (10)	default (10)	default (10)	default (10)	default (10)	default (10)	default (10)	default (10)	4	4	default (10)
mds_bal_sample_interval	default (3)	default (3)	default (3)	default (3)	default (3)	default (3)	default (3)	default (3)	default (3)	default (3)	default (3)	default (3)	2	2	default (3)
mds_bal_replicate_threshold	default (8000)	default (8000)	default (8000)	default (8000)	default (8000)	default (8000)	default (8000)	default (8000)	default (8000)	default (8000)	default (8000)	default (8000)	default (8000)	16000	default (8000)
mds_log_max_segments	default(128)	default(128)	default(128)	default(128)	default(128)	default(128)	default(128)	default(128)	default(128)	default(128)	default(128)	default(128)	default(128)	default(128)	512
CPU Hyperthreading	Off	Off	Off	Off	Off	Off	Off	On	On	On	On	On	On	On	On
mds_cache_memory_limit	4GB GB (partially 64GB)	64GB	64GB	64GB	64GB	64GB	64GB	64GB	64GB	64GB	64GB	64GB	64GB	64GB	64GB
client pagecache	on	on	on	on	on	on	on	on	on	on	Off	Off	Off	Off	Off
client_caps_wanted_delay_*	default (5/60)	default (5/60)	default (5/60)	default (5/60)	default (5/60)	default (5/60)	default (5/60)	default (5/60)	default (5/60)	default (5/60)	default (5/60)	1/1	1/1	1/1	1/1
Result Directory	2024.09.20-03.54.13	2024.09.20-05.43.23	2024.09.20-12.43.51			2024.09.20-19.51.06	2024.09.20-21.00.45	2024.09.24-18.07.38	2024.09.24-19.25.10	2024.09.24-20.56.57	2024.09.25-00.17.25	2024.09.25-01.47.00	2024.09.25-03.02.09	2024.09.25-04.09.50	2024.09.25-05.14.25
lor-easy-write (GiB/s)	17.31	16.47	17.32	17.57	16.52	17.41	17.14	23.53	23.64	23.63	24.87	23.80	24.78	23.84	23.74
mdtest-easy-write (kiOPS)	76.56	43.00	54.99	87.96	156.89	160.99	162.58	173.17	169.30	166.95	164.21	162.92	167.81	167.30	169.20
timstamp (kiOPS)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
lor-hard-write (GiB/s)	0.51	0.40	0.46	0.77	0.30	0.30	0.34	0.53	0.55	0.55	0.56	0.72	0.72	0.72	0.71
mdtest-hard-write (kiOPS)	14.42	17.21	15.26	20.02	12.06	15.29	15.94	16.31	17.78	16.33	21.11	16.43	25.88	20.55	20.92
find (kiOPS)	211.10	113.53	625.14	370.33 (too many open files)	1161.23	1141.51	579.37	858.43	553.85	978.32	710.62	464.27	718.91	761.93	761.93
lor-easy-read (GiB/s)	68.44	19.28	66.37	68.00	71.03	78.96	70.99	80.84	71.94	78.51	78.09	78.04	77.91	78.25	78.25
mdtest-easy-stat (kiOPS)	very slow, canceled	9.85	123.57	117.78		127.21	125.45	114.88	114.29	110.62	118.46	113.24	112.84	118.00	110.38
lor-hard-read (GiB/s)	0.56	2.82	3.42	3.42	3.50	3.36	3.44	3.38	3.38	18.80	17.75	15.03	18.17	14.85	
mdtest-hard-stat (kiOPS)	28.22	87.22	82.98	78.57	81.08	69.36	94.38	55.14	108.64	67.39	118.57	93.85	95.58	95.58	
mdtest-easy-delete (kiOPS)	28.32	15.75	43.64	96.57	89.41	78.57	73.69	74.05	73.43	95.01	72.53	76.49	68.69	68.69	
mdtest-hard-read (kiOPS)	20.21	15.68	66.81	41.27	53.55	8.29	75.81	6.07	78.75	39.13	27.63	28.34	23.85	23.85	
mdtest-hard-delete (kiOPS)	3.44	4.70	6.21	21.98	24.51	7.58	12.58	5.90	13.11	15.14	13.82	12.58	11.65	11.65	
SCORE	7.88	15.76	20.48			22.29	23.56	19.83	25.13	18.80	32.17	30.71	29.45	30.03	28.67

## SSC is:

- Open competition with predefined tasks in regards to optimizing IO
- Teams and individuals
- Active phase for SSC2024/2025 starts Jan 2025
- Planning for 2025/2026 started
- Includes various tasks for **I/O performance analysis**

## Goals:

- Improve participants competencies in I/O behaviour analysis and optimization
- Finding interesting/better task/application specific I/O optimizations
- Encourage individuals to participate in the subject of I/O
- Foster the research by offering a low key entry opportunity for interested persons

- Measure performance for varying configurations
- Address throughput, (max) latency, application startup
- Optimize performance for a fixed benchmark case
- Optimize storage system setup
- Creating performance reports

# Backup slides

Central storage for all compute islands with high availability

- VAST Data (mountable via NFS, home directories, central software installations)
- CephFS (capacity tier), replaces HDD based Lustre and BeeGFS

Compute island specific storage for high performance (migrating to all flash)

- Lustre (DDN ExaScaler or Community Lustre)
- VAST Data via NFS over RDMA (esp. read intensive AI)
- BeeGFS (smaller clusters)

Links from central storage to specific storage - it's project specific

## Lustre:

- DDN 2xES400NVX with 500T SSDs running ExaScaler 6
- DDN 2xES14KX with HDDs (8.5P) and 2xSFA200NV with SSDs (110T) running ExaScaler 5 (will be phased out soon)

## VAST Data

- 3 dBoxes (1 Mavericks 550T, 2 Ceres 330T, total 1.1P)
- 3 cBoxes (1 cBox for 100G Ethernet, 2 cBoxes for Infiniband)

## BeeGFS

- One system with 2 Servers backed by DDN SFA7990 block storage (SAS HDDs (2.1P) and SSDs (100T))
- One system with 3 servers and Dell EMC block storage (450T HDDs and 210T SSDs)
- One system with 2 Supermicro servers (240T HDDs)

- IO500
  - Submission to IO500 Webpage (Research section)
  - Optimize for 10 Client setup
  - Description of the configurations measured and performance improvement made
- MD Workbench - Lowest (maximum latency) for the fixed configuration
- Elbencho
  - arbitrary number of client nodes, 100 KByte files (same file size as DLIO)
  - 10 Client nodes, 100 KByte file vs. 100 KByte access in shared files
  - 1 client node, single 100 MByte block, ensuring cache blows out



- Quantum Espresso
  - Tune a standard use case
  - Sanity check: provide image from the visualization pipeline
- NVIDIA DALI Pipeline - Large file access pipeline with prepared python code
- Conda environment - Startup time of prepared environment, uncached
- Performance analysis (report)
- Secret Task