Lustre usage and compression at DKRZ

Michael Kuhn

Research Group Scientific Computing
Department of Informatics
Universität Hamburg

2016-09-21

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

informatik die zukunft

About us: Scientific Computing

- Analysis of parallel I/O
- I/O & energy tracing tools
- Middleware optimization
- Alternative I/O interfaces
- Data reduction techniques
- Cost & energy efficiency

We are an Intel Parallel Computing Center for Lustre
("Enhanced Adaptive Compression in Lustre")

HLRE3 – Mistral[1]

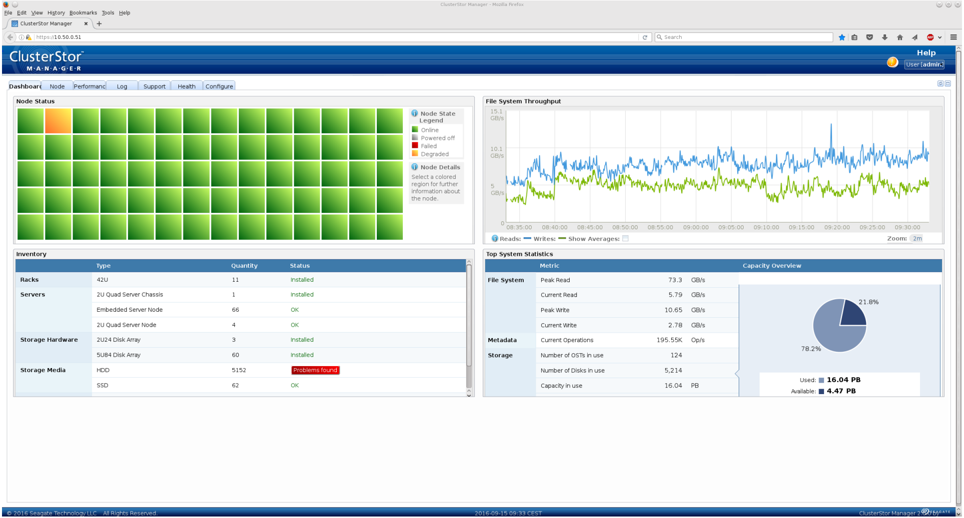
- Went into operation in two phases
  - Spring 2015 and spring 2016
- Currently number 33 on the TOP500
- Approximately 3,000 nodes
  - 1,500 nodes: 2× Intel Xeon E5-2680v3 12C 2.5 GHz (Haswell)
  - 1,600 nodes: 2× Intel Xeon E5-2695V4 18C 2.1 GHz (Broadwell)
- 2.5 PFLOPS (3.14 PFLOPS peak)
- 240 TB RAM
- InfiniBand FDR
  - Fat tree with 2:2:1 blocking

[1] With a lot of information from Carsten Beyer.

HLRE3 – Mistral…

- Lustre with a capacity of 54 PiB
  - Split into two file systems, due to phases
- One of the largest storage systems
  - Storage development is a problem
  - CPU factor 20, storage speed factor 15, storage capacity factor 9.5
- Based on Seagate ClusterStor
  - Scalable Storage Units (SSU) and Expansion Storage Units (ESU)
- Throughput of 450 GB/s
  - 5.9 GB/s per node
  - Single-stream performance: 1 GB/s

HLRE3 – Mistral. . .



HLRE3 – Mistral…

- Phase 1 (CS9000)
  - Lustre 2.5.1 (Seagate)
  - 62 OSSs with 124 OSTs
  - 5 MDSs with DNE
  - Per SSU/ESU: Two trays with 41× 6 TB HDDs each
    - One SSD for parity
  - 80,000 metadata operations per second
- Phase 2 (L300)
  - Lustre 2.5.1 (Seagate)
  - 74 OSSs with 148 OSTs
  - 7 MDSs with DNE
  - Per SSU/ESU: Two trays with 41× 8 TB HDDs each
    - One SSD for parity

HLRE3 – Mistral…

- File system is separated into Home, Work and Scratch
- Home for code, configuration files etc.
  - 24 GB quota per user
  - Backup
- Work for input and output data
  - Project-specific quotas (TBs)
  - No backup
- Scratch for temporary data
  - 15 TB quota per user
  - No backup
  - Data is deleted 14 days after last access

HLRE3 – Mistral…

- Policies are implemented using Robinhood
  - Quota reporting, planned for cleaning up Scratch
- Currently five instances, one per MDS (phase 1)
  - Planned: Two instances for phase 1, three for phase 2
- $2\times$ RAID1 with two SSDs (500 GB each)
  - One for OS (ext4), one for MariaDB (XFS)
- 256 GB RAM, 128 GB dedicated to Robinhood
- Performance is satisfactory
  - Can scan 6,000,000 entries per hour
  - 60,000,000 entries per MDS

HLRE3 – Mistral. . .

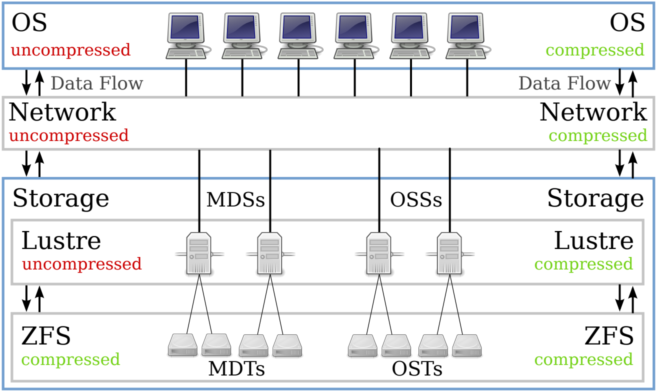
- Tape system with a capacity of 200 PB
  - 15 GB/s throughput
  - No automatic HSM
- System is stable, everything works
  - Failover etc.
- Client upgrade to 2.7 is planned (October)
  - Server upgrade is currently not planned

Workflow

- Climate applications often use CDI/NetCDF/HDF
  - Supports parallel I/O via MPI-IO
- Scientists have application- and domain-specific solutions
  - I/O servers such as XIOS
- Performance is problematic
  - Most applications use serial I/O
  - Data is shipped to master process that performs I/O
  - Simply turning on parallel I/O makes it slower

Gap between computation and storage

- Capacity and performance continue to increase exponentially
  - Different components improve at different speeds
- I/O is becoming an increasingly important problem
  - Data can be produced faster but it becomes harder to store it
- Consequence: Spend more money on storage
  - Results in less available money for computation
  - Or more expensive systems overall
- Storage becomes a considerable portion of the TCO
  - Around 20 % of total costs for DKRZ

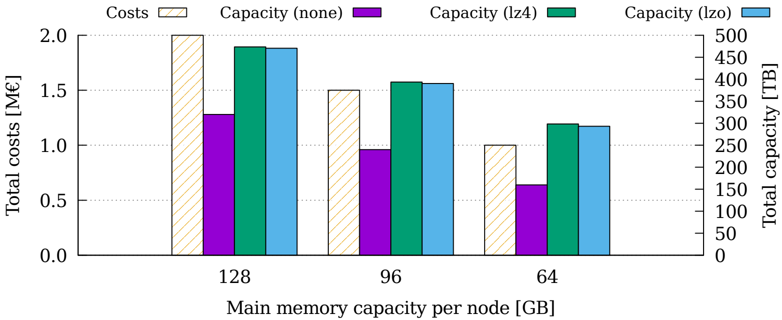


- Left: Compression is only performed on the servers (status quo)
- Right: Compression can be performed on the clients (goal)

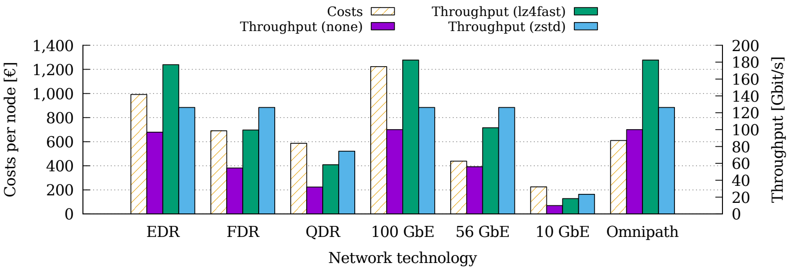
- Investigated compression across the whole I/O stack [1]
  - Main memory, network, storage
  - Both performance and costs
- Compression and HPC usually do not mix well
  - Modern algorithms can provide high performance
- Some interesting results regarding cost efficiency
  - Still have to analyze performance impact in more detail

Algorithm	Compression	Decompression	Ratio
lz4fast	2,945 MB/s	6,460 MB/s	1.825
lz4	1,796 MB/s	5,178 MB/s	1.923
lz4hc	258 MB/s	4,333 MB/s	2.000
lzo	380 MB/s	1,938 MB/s	1.887
xz	26 MB/s	97 MB/s	2.632
zlib	95 MB/s	610 MB/s	2.326
zstd	658 MB/s	2,019 MB/s	2.326

- Measured using lzbench on a climate data set
- lz4 and lz4fast are suspiciously good
  - Additional benchmarks confirm results are realistic
- zstd is also interesting
  - Higher compression ratio with decent performance
- Several good candidates for archival



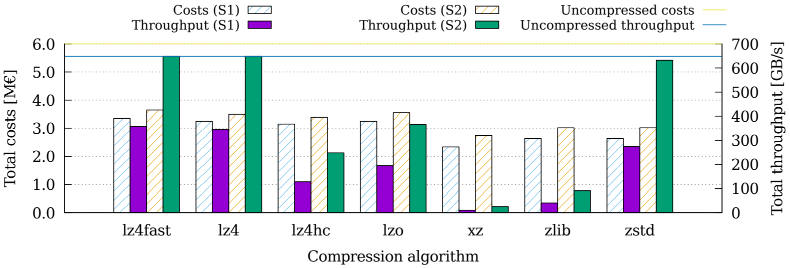
- zram can be used to compress main memory
  - lzo and lz4, multiple compression streams
- Reach a per-node capacity of 128 GB
  - Compress as much as necessary to reach capacity target, leave remaining main memory uncompressed
  - Not possible with 64 GB (leave 4 GB uncompressed)
- Leads to more data that we have to store



- I/O performance not optimal due to network layout
- Per-node throughput could be improved to roughly 100 Gbit/s (lz4fast) or 125 Gbit/s (zstd)
    - zstd limits throughput for networks faster than 54 Gbit/s
- Alternatively, FDR InfiniBand network could be replaced with QDR InfiniBand when using lz4fast, decreasing costs by 15 %

- Assumption: 50 PB of storage with 650 GB/s throughput
  - Costs approximately € 6,000,000
  - Distributed across 60 SSU/ESU pairs
  - Results in 833 TB and 10.8 GB/s per pair
- Costs of € 100,000 per SSU/ESU pair
  - Assume base costs of € 10,000
  - Up to € 90,000 for HDDs
- Additional costs of € 1,500 for compression
  - Each pair currently equipped with two 8-core CPUs
  - Dedicated or faster CPUs for compression

- Scenario 1: Purchase as many fully equipped SSU/ESU pairs as necessary for 50 PB
  - Lower costs: Buy the minimal amount of hardware
  - Decreased throughput: Missing pairs impact performance
- Scenario 2: Purchase as many HDDs as necessary for 50 PB and distribute them across 60 SSU/ESU pairs
  - Slightly higher costs: Base costs for pairs
  - Higher throughput: No pairs are missing



- lz4 and lz4fast do not degrade performance, costs are decreased to roughly € 3,500,000
- zstd decreases throughput by 20 GB/s and costs to € 3,000,000

Conclusion

- DKRZ has one of the largest storage systems
  - Using it efficiently is sometimes problematic
- Storage systems lag behind computation
  - Problem will only get worse over time
  - Compression can help alleviate it
- We are working on compression in Lustre
  - `https://wr.informatik.uni-hamburg.de/research/projects/ipcc-l/start`

[1] Michael Kuhn, Julian Kunkel, and Thomas Ludwig. Data Compression for Climate Data. *Supercomputing Frontiers and Innovations*, pages 75–94, 06 2016.