gStore – the GSI Archive Storage for Experiment Data

H. Göringer, M. Feyerabend, M. Imhof, and S. Sedykh
GSI, Darmstadt, Germany

Overview

gStore is a client/server middleware developed at GSI and tailored according to the requirements of the GSI experiments. gStore provides high performance access 24 hours a day and 7 days a week. For running experiments, highly parallel online writing to dedicated gStore write cache is enabled, including online data copy to lustre for online experiment monitoring and analysis. Due to the design principles:

- reliable long term archive storage,
- full scalability in data capacity and I/O bandwidth,
- high performance access due to intrinsic parallelism,

gStore is well prepared for the challenges of the future FAIR T0 centre. Design principles and functionality are described in detail in GSI reports, talks, and two papers.[1]

Hardware Status

Experiment data are archived in two automatic tape libraries (ATLs), which are also used for backup of user data. The larger ATL has a storage capacity of 8.8 PByte and an I/O bandwidth of 2 GByte/s currently. The smaller ATL (1.3 PByte) is located in the remote BG2 building and contains copies of experiment (raw) and user backup data. This concept prevents from loss of valuable data in case of media damage and enables data recovery even in case of a disaster in the computing centre.

Data are accessed via data movers with read and write disk cache of 220 TByte overall. This large disk buffer hides the tape storage from the users to a big extent. The lustre file system /hera, a cluster file system with ˜ 3 PByte storage capacity, is the online storage for data analysis on Prometheus. The I/O bandwidth between gStore and /hera amounts to 2.5 GByte/s currently.

gStore Enhancements

Data copy to lustre. To utilize the available bandwidth, for retrieve processes from tape to lustre automatical parallelization by gStore has been implemented. The input data are sorted twice, according to their location on tape media and to the storage order on the media. Then files on different tape media are copied in parallel by different processes running on different data movers, and all files are read in optimal order from the corresponding media. Obviously the number of parallel processes is limited by the number of available tape drives, which is up to eight currently. It should be noted that this parallelization cannot be done by the users themselves, because they intentionally need not care about file location on tape and therefore do not have the corresponding information. Due to the high performance connection between gStore and lustre - matching the tape speed of 250 MByte/s - tape files are copied directly to lustre, skipping the otherwise mandatory staging step to gStore read cache. This additionally reduces the copy time considerably.

The parallelization concept works only efficiently if a large number of files is copied with one single (gstore) command, the more files are involved, the better. To specify many files, wildcard characters, a file list, recursive file operations, or any combination of them can be used. Thereby users need not keep track of files already archived in gStore or already staged or retrieved from gStore, respectively, because files already existing are never overwritten, except if explicitly specified otherwise. If some of the specified gStore files are not on tape, but already staged in read cache or still residing on write cache, additional copy processes on each data mover involved are started. Therefore a few tens of copy processes may initiated by one single user command

Removal of limitations in file number. To support as many files as possible in single commands, some limitations in file number have been identified and removed. Up to now with one single command more than 100,000 files have been processed successfully by users.

As an example, in a recursive user archive command 127,473 matching files have been found in lustre. 26,635 files have been rejected, mainly because they were already existing in gStore, or because they were empty or had invalid file names. The remaining 100,838 files, with file sizes from some 10 bytes to ˜ 1 GByte and ˜ 6.5 TByte size in total, have been archived successfully to the write cache of a data mover. With all latencies included, the average data rate amounted to 152 MByte/s.

Outlook

It is planned to implement automatic parallelization also for processes copying from lustre to gStore. Using up to 10 data movers in parallel, the overall copy time would be decreased by an order of magnitude.

Newer lustre versions soon available at GSI enable to implement and test the lustre HSM (Hierarchical Storage Manager) functionality with tape backend in gStore.

According to the road maps of big tape manufacturers, data capacity (now 4 TByte/tape) and I/O speed (now 250 MByte/s) will be doubled in the next years. With additional frames for tape media and tape drives, our ATL data capacity could be enhanced then to ˜ 100 PByte, which is already the order of magnitude needed for FAIR.

References

[1] see `http://www.gsi.de/informationen/wti/it/exp_daten/daten_speicherung_e.html` as starting point for more info

 doi:10.15120/GR-2014-1-IT-15