High Performance Experiment Data Archiving with gStore

H.Goeringer, M.Feyerabend, S.Sedykh

GSI Helmholtzzentrum für Schwerionenforschung – Center for Heavy Ion Research
Planckstr. 1, D-64291 Darmstadt, Germany

E-mail: H.Goeringer@gsi.de

Abstract. GSI in Darmstadt (Germany) is a center for heavy ion research. It hosts an Alice Tier2 center and is the home of the future FAIR facility. The planned data rates of the largest FAIR experiments, CBM and Panda, will be similar to those of the current LHC experiments at Cern. gStore is a hierarchical storage system with unique name space and successfully in operation since more than fifteen years. Its core consists of several tape libraries and currently ~20 data mover nodes connected within a SAN network. The gStore clients transfer data via fast socket connections from/to the disk cache of the data movers (~200 TByte currently). Each data mover has also a high speed connection to the GSI lustre file system (~3 PByte data capacity currently). The overall bandwidth between gStore (disk cache or tape) and lustre amounts to 5 GByte/s and will be duplicated in 2012. In the near future the lustre HSM functionality will be implemented with gStore. Each tape drive is accessible from any data mover, fully transparent to the users. The tapes and libraries are managed by commercial software (IBM Tivoli Storage Manager TSM), whereas the disk cache management and the TSM and user interfaces are provided by GSI software. This provides the flexibility needed to tailor gStore according to the always developing requirements of the GSI and FAIR user communities. For Alice users all gStore data are worldwide accessible via Alice grid software. Data streams from running experiments at GSI (up to 500 MByte/s) are written via sockets from the event builders to gStore write cache for migration to tape. In parallel the data are also copied to lustre for online evaluation and monitoring. As all features related to tapes and libraries are handled by TSM gStore is practically completely hardware independent. Additionally, according to the design principles gStore is fully scalable in data capacity and I/O bandwidth. Therefore we are optimistic to fulfill also the dramatically increased mass storage requirements of the FAIR experiments in 2018, which will be some orders of magnitude higher than those of today.

1. Introduction
The GSI Helmholtzzentrum für Schwerionenforschung [1] in Darmstadt (Germany) is a basic research center for physics with heavy ions. Linear accelerator, synchrotron, and storage ring provide heavy ion beams for experiments mainly in the area of nuclear physics, but also in other research areas such as atomic and plasma physics, biophysics and radiation medicine, material research and accelerator research. GSI is also a member of the Alice collaboration [2] and hosts a large Alice Tier 2 computing center. For the ambitious goals of the 1.2 billion Euro project FAIR [3], the GSI experiment facilities will be enhanced considerably. FAIR is planned to start in 2018.
The gStore mass storage system [4], [5] is a middleware developed at GSI and successfully in operation since more than 15 years. It is based on automatic tape libraries (ATL) for long term archiving, and on data movers (DM) with large read and write disk caches. The caches hide tape operations as far as possible from the users. Tape libraries and data movers are connected via Storage Area Network (SAN). The data are available 24 hours per day and seven days per week for fast and highly parallel access, which is enabled via command client and application programmer's interface (API). All GSI experiment data (0.9 PB in May 2012) are stored in a tape library, which has a maximum data capacity of 8.8 PB actually. The overall capacity of gStore read and write cache currently sums up to 200 TB.

The future challenges for gStore are considerable. For start of FAIR operation in 2018, the FAIR experiments assume an annual data growth of 33 PB. Especially the two main FAIR experiments, CBM and Panda, will have mass storage requirements similar to those of the current LHC experiments at CERN.

gStore cooperates closely with the GSI lustre file system, which has a size of nearly 3 PB and serves as main online mass storage system for offline data analysis. Another important gStore task at GSI is the online data storage from running experiments. Data from data acquisition are transferred online to gStore write cache. Before migration to tape, data are optionally also copied to lustre, where they are immediately available for monitoring and online analysis. The underlying concepts are discussed in more details in the sections 4 and 5 below.

High performance access to the data is provided by the architecture of gStore, which is based on parallel data streams between tapes and data mover cache and between data mover cache and client storage. Besides that, there is strict separation of control and data flow. By design gStore is fully scalable in both, data capacity and I/O bandwidth, and therefore well prepared for the challenges of the future FAIR T0 center. Design principles and functionality of gStore are described in more detail in several GSI reports, talks, and in a paper [4], [5].

2. gStore Software

gStore software consists of two main parts:

1. TSM, the Tivoli Storage Manager, a commercial package provided by IBM, and
2. GSI software, a client/server package, consisting of more than 100,000 lines of C-code and using TCP sockets.

TSM handles the automatic tape libraries and tape drives in gStore. As TSM is world market leader for backup and archive, all relevant storage devices are supported by TSM and therefore also by gStore. As a result gStore works practically independent from the storage devices used.

Besides TSM, all other gStore software has been developed at GSI. It provides the interfaces to the users, utilizes the TSM functionality via TSM API, drives the gStore entry servers and data movers, and manages the read and write caches of gStore. The main tasks of the cache managers are management of the cache meta data, locking and unlocking of the cache files, cache space management, and data mover selection, based on load balancing algorithms.

3. gStore Hardware

Currently there are two IBM 3584-L23 tape libraries in use. The first one has 8 IBM 3592-E07 tape drives, each reading and writing with up to 250 MB/s. The actual tape media capacity is 4 TB uncompressed. The overall library capacity amounts to ~9 PB currently. With user backup data included thereof nearly 1 PB is actually in use.

The second tape library is used for copies of raw experiment data and of some user backup data. It is located in a different building outside of the computing center in order to allow disaster recovery. The backup library has 4 older IBM 3592-E06 tape drives, each reading and writing with up to 160 MB/s. The tape media capacity is still limited to 1 TB uncompressed, and the overall library capacity amounts to ~1.2 PB actually. About 200 TB are currently in use.
For the latest tape drive generation (E07), the current media costs amount to about 40 € per TB. Taking into account additional proportionate library and tape drive costs of about 60%, it is obvious that for reliable long term data archiving there is still no alternative to tape available, which is similarly inexpensive - neither currently nor in the near future. Besides that, due to the low power resources needed - tape storage means really green IT.

Currently gStore has 17 data movers in operation running on Suse Linux. They have disk caches of 3-20 TB, which sums up to overall 200 TB. Each data mover has 4 Gbit SAN connection to the tape libraries, and most of them have 10 Gbit Ethernet connection to the client LAN. The maximum I/O bandwidth between disk cache and tape is 2 GB/s, whereas the maximum I/O bandwidth between disk cache and clients amounts to nearly 5 GB/s, mainly limited by a common 40 Gbit switch.

A summary of the actual gStore hardware status can be found in figure 1.

4. gStore features
All gStore servers are running in 64 bit mode and support 32 bit and 64 bit clients. Where applicable, all gStore functions support recursive file handling, also in combination with wildcard characters ("*" and "?"). Both may replace any character except the (sub)directory delimiter ("/").
4.1. gStore and lustre
Each gStore data mover has a lustre mount point, and data are copied directly between gStore data movers and lustre OSTs. Depending on the speed of the servers selected, up to 500 MB/s have been reached for single file transfers. The overall I/O bandwidth with gStore tape or disk cache amounts to 2 GB/s and 5 GB/s, respectively, see figure 1. Due to the good connectivity to lustre, lustre clients reading from tape are allowed to skip the gStore read cache and copy directly from tape to lustre. This is the only exception – all non-lustre clients can only read data via gStore read cache.

4.2. Automatic process parallelization.
Due to the parallel architecture of gStore, in many cases it happens more or less ‘naturally’ that several file transfers submitted nearly at the same time run in parallel. This is obvious, if, for example, several users read files from read caches of different data movers. It is also true if several users stage files from different gStore archives simultaneously. As each gStore archive has its own tape pool, files from different archives are always located on different tape media, and simultaneous processes may run in parallel.

GStore allows to start large transfers of many files with one single command. Parallelization of such single processes is also necessary, but requires more effort. This can not be done by users, because division into several commands with suitable file subsets is usually too cumbersome and not acceptable. It is even impossible, if tape files are involved, because gStore users have no information about file location on tape media, of course. Therefore parallelization of single processes must be arranged by gStore, which in case of tape files obtains the necessary tape metadata information from the TSM database.

For staging from tape to read cache, automatic parallelization of single processes is already implemented in gStore. The files are sorted according to the tape media where they are located, and according to the optimal retrieve order for each medium. Then different data movers are assigned to different tape media, and the file set is staged with several parallel processes. This not only decreases the staging time, but also enables higher parallelism of file access afterwards. The number of parallel processes is limited by the number of tape media involved and by the number of available tape drives.

If all requested files reside on the same tape medium, parallelization is not possible. In such cases gStore divides the files into an appropriate number of chunks and assigns different data movers to different chunks. Then the files are read sequentially, by one data mover after the other. With increasing media capacity (4 TB for IBM 3592 E07) this feature becomes increasingly important.

Single process parallelization makes no sense if certain types of clients are involved, such as desktops, overloaded group file servers, or badly connected file servers. For transfers between gStore and lustre, however, automatic process parallelization is attractive, since lustre OSTs are powerful clients normally. This project is currently a work in progress.

At GSI there are several hundred lustre OSTs in operation. Therefore the number of parallel processes is only limited by the number of available gStore data movers and the number of available tape drives, if involved. Writing to lustre in parallel will be especially efficient, because not only gStore load balancing, but also lustre load balancing will be utilized. When data are read from lustre, the file distribution on the lustre OSTs and therefore the efficiency of parallelization depends on history.

5. Online data storage from running experiments.
An important gStore use case at GSI is the online data storage from the data acquisition of experiments, which sends constant, continuous data streams to gStore write cache over long time periods up to weeks. For example Hades, the largest GSI experiment, normally uses 16 data streams in parallel. They are equally distributed to dedicated data movers, which are undisturbed by the heavily varying offline activities of other gStore clients.
It is a strong user requirement at GSI that most or all data from data acquisition are as fast as possible available on lustre, for online monitoring and for online analysis. Therefore, data are first stored in gStore write cache, then copied to lustre, and finally migrated to tape, when a preset fill level in write cache file systems is reached. Currently the maximum bandwidth for online storage is limited to 500 MB/s, if all data must be copied to lustre. Without lustre copies, even 1 GB/s is possible.

For online copy to lustre, there are two modes available. In parallel copy mode, each incoming data buffer is written to gStore write cache and to lustre. So each data buffer is immediately available for online analysis. In order to keep the data streams from the experiment and the copy processes to lustre completely independent from each other, also a sequential copy mode has been implemented. Here complete files are copied to lustre after they are closed in write cache. In both copy modes, the user defines which fraction of files is copied to lustre. Optionally new subdirectories can be automatically created in lustre after a specified number of file copies, according to various naming conventions.

6. Conclusions and outlook

gStore is a GSI middleware providing reliable long-term archive storage with high performance access. As tape storage is handled by the commercial TSM package, gStore is able to work with practically all relevant storage hardware. The GSI software based on C-code and TCP sockets runs with minor modifications on practically all operating systems. gStore is designed for highly parallel and high performance access, and it is fully scalable according to data capacity and I/O bandwidth. Due to the in-house development gStore is tailored for the special requirements of the GSI users. It cooperates very closely with the GSI lustre file system, which is used as main online mass storage system for offline data analysis, and provides a comfortable and high performance interface for online data storage from the data acquisition of experiments.

As a long-term future project the implementation of lustre HSM functionality would be very attractive. However, due to a possibly uncertain future of EOFS and lustre GPL, and due to limited manpower it is not yet under investigation currently.

Most important is the preparation of gStore for the FAIR project, which is planned to start in 2018. New technical developments will possibly happen until then, and from today's view the storage situation in 2018 is not yet obvious. The FAIR experiments expect an annual data growth of 33 PB beginning from the start. The current tape library with E07 drives is expandable to 50 PB with a reasonable mixture of media capacity and tape drive upgrade. With the next generation drives E08, which can be expected before the start of FAIR, the library should be expandable to a capacity of more than 100 PB. This can be expected from the capacity increases in the past. Especially the last upgrade from E06 to E07 provided a factor of 4 in data capacity! Possibly more challenging is the needed increase in bandwidth, which amounts to more than a factor of 10.

In the past fifteen years at GSI the scalability of gStore has been proven by accomplishing similar increases in data capacity and in bandwidth as needed for FAIR now. Technical progress helped in the past and will help also in the future. Taking into account the ability of gStore to enable easy adaption for cooperation with external software and middleware, we are optimistic to master also the ambitious challenges of FAIR.

References
[1] GSI Helmholtzzentrum für Schwerionenforschung - Center for Heavy Ion Research, Darmstadt http://www.gsi.de
[2] Alice: A large Ion Collider Experiment, http://aliceinfo.cern.ch/Public/Welcome.html
[3] Facility for Antiproton and Ion Research: http://www.gsi.de/portrait/fair_e.html
[4] gStore: GSI Mass Storage System http://www.gsi.de/informationen/wti/it/exp_daten/daten_speicherung_e.html
[5] Experiences with gStore, a scalable Mass Storage System with Tape Backend Journal of Physics (Conference Proceedings): Conf. Ser. 119 Volume 119(2008) 052018(7pp)
