gStore - a High Performance Experiment Data Archive Storage

H.Göringer, M. Feyerabend, and S. Sedykh

GSI, Darmstadt, Germany

Overview

Archive and storage for experiment data is provided 24 hours a day and 7 days a week by **gStore,** a client/server middleware developed at GSI. Data are archived in automatic tape libraries (ATL), which can be accessed with high performance and in parallel via data movers (DM) with large read and write disk caches.
Design principles and functionality of gStore are described in detail in GSI reports, talks, and two papers [1].

Hardware Status

The eight IBM 3592 tape drives in the ATL located in the computer centre (RZ) have been upgraded from E06 to E07. As a result the I/O speed of each drive has been increased from 160 MByte/s to 250 MByte/s, and up to 4 TByte of uncompressed data can be written to the actual tape media. The current storage capacity for experiment and user backup data has been increased to 8.8 PByte, and the overall I/O bandwidth amounts to 2 GByte/s now. With additional frames for tape media and tape drives, the data capacity could be enhanced to 50 PByte, which is already in the order of magnitude needed for FAIR.
The second ATL located in the remote BG2 building contains copies of experiment (raw) and of user backup data. This concept prevents from loss of valuable data in case of media damage and enables disaster recovery.
Three outdated data movers have been replaced by new ones with larger disk arrays thus increasing the available disk space for read and write cache from 170 TB to nearly 220 TB. A summary of the current hardware resources can be found in table 1.

resource	used	max
storage capacities:		
3584-L23 ATL RZ:	700 TB	8.8 PB
3584-L23 ATL BG2:	240 TB	1.3 PB
overall data mover disk cache	< 90%	0.22 PB
lustre/hera file system	~80%	3.5 PB
overall gStore I/O bandwith:		
DM disk <-> tape (SAN)		2.0 GB/s
DM disk <-> clients (LAN)		5.0 GB/s

Table 1: Hardware Status GSI Storage in December 2012

gStore Enhancements

Access to /hera. Now all data movers have also mount connections with the new lustre file system /hera residing in the testing hall. As /hera is only reachable via Infiniband, data transfers between gStore data movers and /hera file servers take place via so-called LNET routers connecting Ethernet with Infiniband hosts and vice versa. The current I/O bandwidth amounts to 2.5 GByte/s.
Access from Icarus and Prometheus. On the GSI batch farms connected with /lustre (Icarus) and /hera (Prometheus) the gStore clients are made available via cvmfs [2].
New Storage Pool for large data transfers. Large data transfers, e.g. between gStore and lustre/hera, or between tape and disk cache, need the highest performance possible. Therefore they should be made in parallel and on the fastest data movers. However, some of these transfers need read cache and others write cache resources, which were completely separate in gStore in the past. Therefore a new pool has been created for these transfers. It is located on the data movers connected with 10 Gbit and used for both, writing to gStore and reading from gStore.

Data Movements

In 2012, from April to December the online data storage capabilities of gStore were heavily used by up to four experiments running in parallel.
For nearly five weeks, data from Hades event builders, divided into 16 data streams, were written to gStore write cache with an overall average data rate of 100 MByte/s. Additionally, the data were copied automatically from write cache to lustre and migrated to tape afterwards[1] .
The long term stability of gStore was also utilized by the Tasca experiment, which was running for nearly half a year. Nearly all the time data were stored online in gStore with data rates of 10 MByte/s.
Taking into account all data transfers between gStore clients, disk cache, and tape, in 2012 overall 1.37 million files were moved with a data volume of nearly 1.1 PByte.

Outlook

The concept of automatic process parallelization, which is already realized for staging processes from tape to read cache, will also be implemented for data transfer processes between lustre/hera and gStore using the new fast storage pool. Then handling many files in parallel can be done with single gstore commands enabling transfer rates between lustre/hera and gStore with full I/O bandwidth of up to 5 GB/s.

References

[1] see http://www.gsi.de/informationen/wti/it/exp_daten/daten_speicherung_e.html as starting point for more info
[2] http://wiki.gsi.de/cgi-bin/view/Linux/CvmFs

[1] Parts of the data were test data not permanently stored on tape.