EPJ Web of Conferences **214**, 04013 (2019) https://doi.org/10.1051/epjconf/201921404013
CHEP 2018

The Software Defined Online Storage System at the GridKa WLCG Tier-1 Center

Jan Erik Sundermann[1,*], *Jolanta* Bubeliene[1], *Ludmilla* Obholz[1], and *Andreas* Petzold[1]

[1]Karlsruhe Institute of Technology, Steinbuch Centre for Computing

Abstract. The computing center GridKa is serving the ALICE, ATLAS, CMS and LHCb experiments as one of the biggest WLCG Tier-1 centers world wide with compute and storage resources. It is operated by the Steinbuch Centre for Computing at Karlsruhe Institute of Technology in Germany. In April 2017 a new online storage system was put into operation. In its current stage of expansion it offers the HEP experiments a capacity of 34 PB of online storage. The whole storage is partitioned into few large file systems, one for each experiment, using IBM Spectrum Scale as software-defined-storage base layer. The system offers a combined read-write performance of 100 GB/s. It can be scaled transparently both in size and performance allowing to fulfill the growing needs especially of the LHC experiments for online storage in the coming years. In this article we discuss the general architecture of the storage system and present first experiences with the performance of the system in production use.

1 Introduction

The Grid Computing Centre Karlsruhe (GridKa) [1] is a data and computing center for particle and astroparticle physics experiments. It is operated by the Steinbuch Centre for Computing (SCC) at the Karlsruhe Institute of Technology (KIT) in Germany. It was founded in 2002 initially supporting the four particle physics experiments BaBar, D0, CDF and COMPASS. Since 2006 GridKa is providing resources for the Pierre Auger Observatory. In 2008 GridKa started its full production service with 24/7 coverage before the anticipated start of the LHC supporting all four LHC experiments, ALICE, ATLAS, CMS and LHCb, as the German Tier-1 centre. From 2021 onwards, GridKa will serve as raw data center for the BELLE-II experiment. At the moment GridKa is responsible for about 14% of the raw LHC data. It is the largest of the 13 Tier-1 centers in terms of CPU and storage resources pledged to the LHC experiments.

GridKa consists of a high-throughput compute farm, large installations of disk (online) and tape (offline) storage as well as a dedicated network infrastructure. The compute farm consists of 1000 worker nodes with 18000 cores in a classical high-throughput setup. GridKa uses cost-efficient and reliable hardware in terms of power consumption, rack occupancy, and network infrastructure usage. The typical analysis and simulation production workload requires one disk spindle per 10 jobs and 10 Gbit/s Ethernet connections from the worker node server. The GridKa farm has 29000 job slots which were utilized on average by 98%. Jobs are typically single core or multi-core on the same CPU but do not require fast network

*e-mail: jan.sundemann@kit.edu

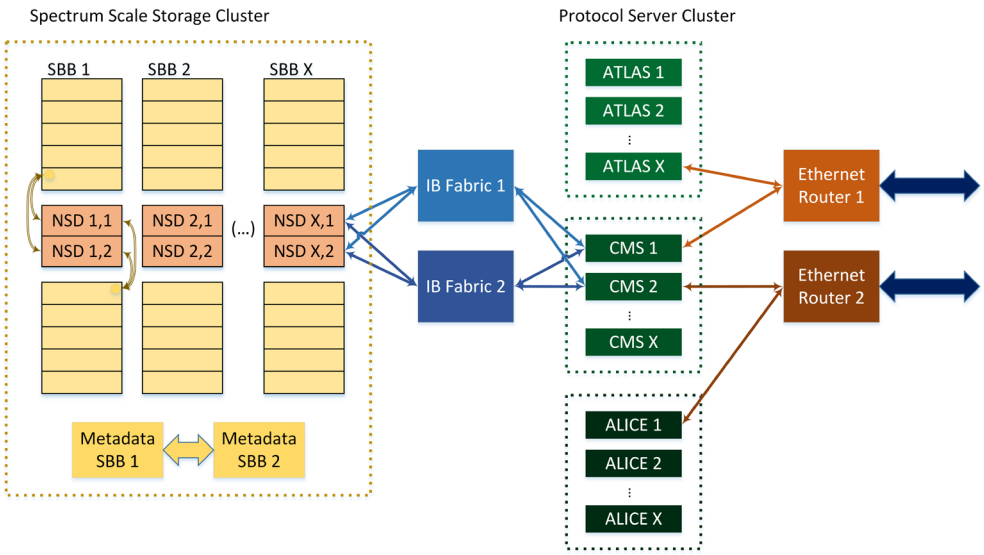


Figure 1: Schematic overview of the GridKa online storage system including a storage cluster with disks (yellow), dedicated protocol clusters per experiment (green) and the network infrastructure consisting of InfiniBand fabrics (blue) and Ethernet switches (brown) to connect to the GridKa farm and remote sites.

interconnects between the worker nodes. During the past 12 months, 24 Million jobs were running in GridKa corresponding to 176 Million CPU hours.

In order to serve as a data distribution hub in WLCG, GridKa is connected with dedicated 100 Gbit/s network connections to CERN, the German research network and private high-energy physics networks. Upgrades of those networks to 200 Gbit/s will be done once required and financially viable.

The tape storage system of GridKa is actively used by the experiments as a distributed backup of LHC data. Data is frequently recalled from tape for the reprocessing of the raw data. Tape operations are initiated from the GridKa disk storage system. At the moment 49 PB of experiment data is stored on tape in two libraries.

2 The GridKa Online Storage System

In 2017 a new disk storage system was put into operation in GridKa. The new storage system is a GxFS storage appliance from NEC [2]. The design of the system allows for a flexible scaling of the storage infrastructure both in size and performance. The system uses IBM Spectrum Scale [3] as software defined storage layer. The storage is partitioned into few very large file systems which enable the operator to manage the storage efficiently and make it possible to optimize for different scenarios, e.g. different experiments or workloads such as tape buffers. The online storage system currently has a usable capacity of 34 PB with a combined maximum read-write performance measured to be 100 GB/s.

Figure 1 shows schematically the structure of the storage systems. The whole setup is divided into different sub-clusters, each of them setup to perform a different task. The storage cluster builds and manages the disk pools and file systems. Data can be written to or read from the file systems through multiple protocol clusters. The protocol clusters only mount the

file systems exported by the storage cluster, but they do not have their own storage attached. GridKa typically operates one protocol cluster per experiment served with disk storage. The division into storage and protocol clusters allows for a clean organizational and administrative separation of the storage on the one hand and the required access protocols on the other hand.

Two redundant Mellanox 56 Gbit/s FDR InfiniBand fabrics provide the necessary network infrastructure to allow for a high performance access with low latency of the protocol servers to the data via remote direct memory access (RDMA). The InfiniBand fabrics are setup with a blocking factor of 2:1. Each fabric is setup in a fat tree topology with 2 spine and 5 leaf switches. Each protocol server is connected via a 40 Gbit/s Ethernet link to one of the two Ethernet switches. The storage system is connected redundantly with 8 100 Gbit/s Ethernet lines to the GridKa network backbone and subsequently to the high-throughput compute farm and remote WLCG sites.

The storage cluster consists of several independent storage building blocks (SBBs) (cf. Fig. 1). Thus, it is possible to scale the storage easily in both size and performance by adding additional SSBs to the storage cluster. Every SBB consists of 10 enclosures (NetApp E5600 / DE6600), which are connected redundantly in sub-blocks of 5 enclosures, each sub-block with 2 raid controllers, to two servers (NEC Express 5800 R120g-2M). Each enclosure holds 60 8 TB or 10 TB nearline SAS hard drives. Utilizing so-called Dynamic Disk Pools (DDPs) [4] always 50 hard disks, distributed evenly over a sub-block, are combined into logical disk pools. Data is written to a DDP in Raid 6 stripes (8+2P). Within each pool a storage capacity equivalent to the capacity of three hard drives is reserved and used as a replacement in case of a hard drive failure. Compared to classical Raid 6 systems, DDPs allow for significantly faster rebuilds in the event of disk failures.

3 Protocol Servers

GridKa uses two different software frameworks to provide protocol access to the data stored in the online storage infrastructure, XRootD [5] and dCache [6]. All file servers are separated from the actual storage and setup in dedicated remote protocol clusters.

We utilize a native XRootD setup for the ALICE experiment. In this setup every XRootD file server is able to serve all files of the ALICE file system (see Fig. 2). The native XRootD setup directly profits from the underlying Spectrum Scale parallel file system and allows to easily scale the provided read and write performance by adding additional file servers to the protocol cluster.

dCache natively provides a large number of different protocols to access the data from the compute cluster or remote sites. The GridKa storage system uses separate dCache instances for ATLAS, CMS, LHCb and Belle II. Each dCache pool lives in sub-directory of a large file system (see Fig. 3). The dCache pools provide no automatic fail over but it is in general easy to deploy a failing pool on a different server in case of a hardware failure.

4 Management and Monitoring

All servers of the GridKa online storage system are provisioned and configured with Foreman and Puppet using the central GridKa infrastructure. The necessary Puppet modules were developed by the GridKa team and allow new servers to automatically join existing Spectrum Scale clusters and reinstalled servers to automatically recover their Spectrum Scale configuration. Since the Spectrum Scale cluster configuration is stored in several copies across many servers, all servers of the storage or protocol cluster are essentially stateless. Therefore, we are able to reinstall any server at any time.

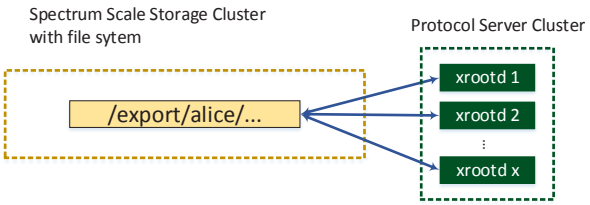


Figure 2: Schematic overview of the XRootD protocol servers. Every protocol server is able to serve every file of the file system.

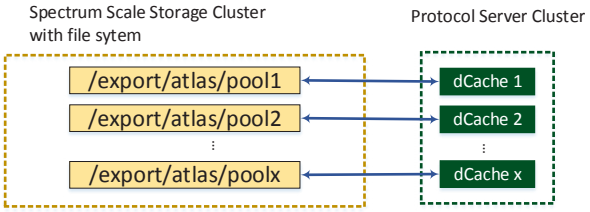


Figure 3: Schematic overview of the dCache protocol servers. Each dCache pool lives in sub-directory of a large file system.

At the moment we are mainly using two software stacks for the monitoring of the GridKa online storage system. We have started to collect log files in an Elasticsearch NoSQL database [7] and enrich them with additional attributes. Log files can be searched and visualized in either a Kibana or a Grafana Dashboard [8]. By collecting log files of all storage and protocol servers in the same database we are able to correlate events in the logs. Figure 4 shows an example of a Kibana dashboard with RDMA errors caused by a faulty InfiniBand cable. Performance metrics are collected using Telegraf and stored in an InfluxDB time series database [9]. Metrics can be aggregated and visualized using Grafana dashboards [10]. Figure 5 shows the aggregated RDMA data read rate for all servers of the storage system in GridKa grouped by experiment.

5 Utilization and Expansion of the Storage System

In 2017 7.5 PB and 4.2 PB of data were read on average per month from the GridKa compute farm and from remote sites, respectively (see figure 6a). In the same time period 1.1 PB and 3.0 PB of data were written on average per month from the GridKa compute farm and from remote sites, respectively (see figure 6b).

In preparation for the provisioning of the 2018/19 storage pledges to the LHC experiments we were able to exercise the upgrade and expansion of the existing storage system. During this expansion, the storage capacity of the cluster was transparently increased from 23 to 34 PB by increasing the number of disk servers from 16 to 22 and the number of protocol servers was increased from 44 to 64. All servers were transparently included in the two InfiniBand fabrics while changing the setup of both fabrics from non-blocking 1:1 to a blocking factor of 1:2.

In the upcoming years, resource increases of 20% per year are envisaged. The available storage capacity of the disk storage system is expected to grow till 2021 to 50 PB.

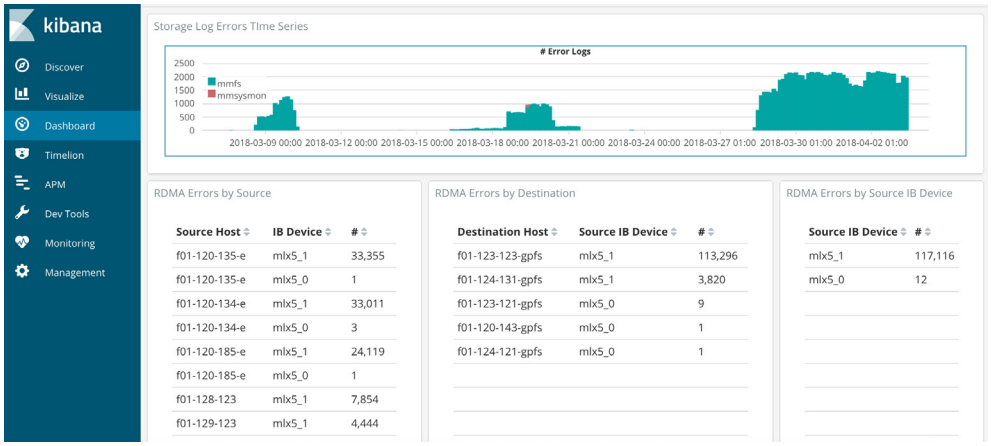


Figure 4: Kibana dashboard showing monitoring information related to RDMA connections between the nodes of the Spectrum Scale clusters.

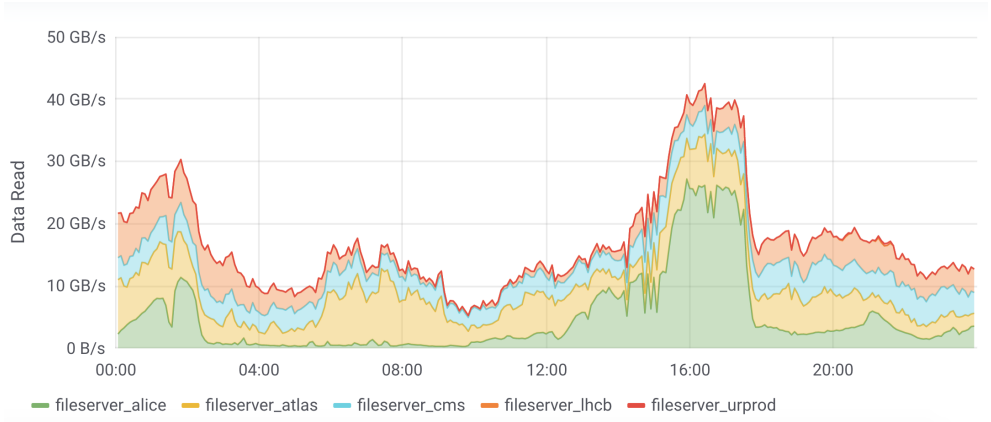
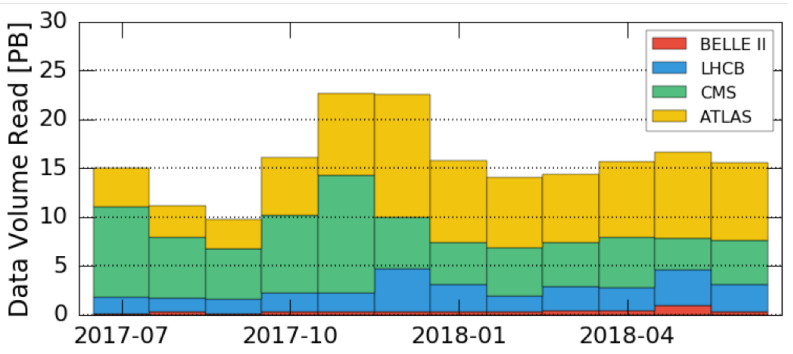


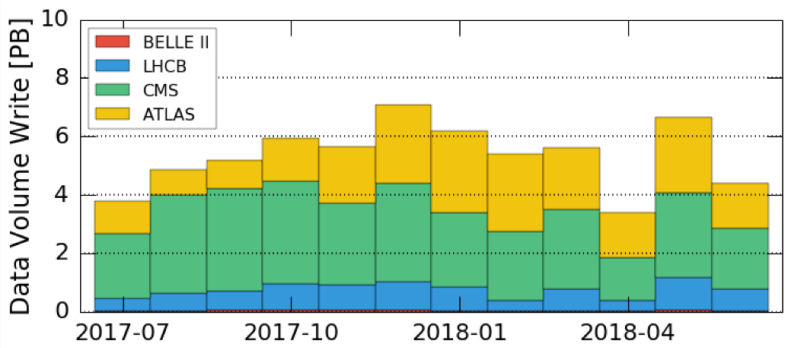
Figure 5: Example for a Grafana dashboard with monitoring information showing the InfiniBand RDMA rates of data being read from the GridKa storage system stacked on top of each other for the experiments ALICE (green), ATLAS (yellow), CMS (blue), LHCb (orange), and others (red).

6 Summary and Outlook

In 2017 a new disk storage system was put successfully into operation in the WLCG Tier-1 center GridKa. The design of the storage infrastructure already proved to be transparently expandable both in size and performance. Meanwhile the GridKa storage team started to investigate technology options for time beyond 2019 where a yearly resource increase of 20% is expected. The storage system is expected to grow till 2021 to a usable capacity of 50 PB. With the current setup we expect to able to scale the system for the requirements towards the HL-LHC.



(a) Data volume read



(b) Data volume write

Figure 6: Data volume read (a) and written (b) to the GridKa online storage system per month and per experiment in 2017.

References

[1] Grid Computing Centre Karlsruhe (GridKa). Available from http://www.gridka.de [accessed 2019-02-14]

[2] NEC GxFS - GPFS Appliance. Available from https://www.nec.com/en/global/solutions/hpc/storage/gxfs.html [accessed 2019-02-14]

[3] IBM Spectrum Scale. Available from https://www.ibm.com/us-en/marketplace/scale-out-file-and-object-storage [accessed 2019-02-14]

[4] NetApp SANtricity Dynamic Disk Pools (DDP). Available from https://www.netapp.com/us/info/what-is-dynamic-disk-pools-technology.aspx [accessed 2019-02-14]

[5] A. Dorigo et al., *XROOTD/TXNetFile: A Highly Scalable Architecture for Data Access in the ROOT Environment*, Proceedings of the 4th WSEAS International Conference on Telecommunications and Informatics (2005)

[6] P. Fuhrmann et al., *dCache, Storage System for the Future*, Euro-Par 2006 Parallel Processing, 1106–1113 (2006)

[7] The ELK stack. Available from https://www.elastic.co/elk-stack [accessed 2019-02-14]

[8] Grafana. Available from https://grafana.com/ [accessed 2019-02-14]

[9] Telegraf/InfluxDB. Available from https://www.influxdata.com/ [accessed 2019-02-14]

[10] SCC/SDM Grafana - WLCG Tier-1 Center. Available from https://grafana-sdm.scc.kit.edu/ [accessed 2019-02-14]