Journal of Physics: Conference Series **664** (2015) 042047 doi:10.1088/1742-6596/664/4/042047

An integrated solution for remote data access

Vladimir Sapunenko[1,6], Domenico D'Urso[2,4], Luca dell'Agnello[1], Vincenzo Vagnoni[3] and Matteo Duranti[4,5]

[1] INFN-CNAF, Viale Berti-Pichat 6/2, 40127 Bologna, Italy
[2] ASDC, Via del Politecnico, 00133 Roma, Italy
[3] INFN Sezione di Bologna, Via Irnerio 46, 40126 Bologna, Italy
[4] INFN Sez. Perugia, Via A. Pascoli, 06123 Perugia, Italy
[5] University of Perugia, Piazza Università 1, 06100 Perugia, Italy

E-mail: vladimir.sapunenko@cnaf.infn.it

Abstract. Data management constitutes one of the major challenges that a geographically-distributed e-Infrastructure has to face, especially when remote data access is involved. We discuss an integrated solution which enables transparent and efficient access to on-line and near-line data through high latency networks. The solution is based on the joint use of the General Parallel File System (GPFS) and of the Tivoli Storage Manager (TSM). Both products, developed by IBM, are well known and extensively used in the HEP computing community. Owing to a new feature introduced in GPFS 3.5, so-called Active File Management (AFM), the definition of a single, geographically-distributed namespace, characterised by automated data flow management between different locations, becomes possible. As a practical example, we present the implementation of AFM-based remote data access between two data centres located in Bologna and Rome, demonstrating the validity of the solution for the use case of the AMS experiment, an astro-particle experiment supported by the INFN CNAF data centre with the large disk space requirements (more than 1.5 PB).

1. Introduction

HEP experiments produce large data sets, ranging from hundreds of terabytes to hundreds of petabytes of data per year. To minimise hardware costs and have major data protection, data is distributed in several replicas over different locations. Centralised Analysis Facilities provide primary data elaboration and make them available for the analysis by a single or a group of scientists. Data from remote locations is not easy to be analysed, especially in an interactive mode, mainly for latency in data access. End-users prefer to have data "close" to their desk, so they usually ask to have local storage large enough to store in all the data that they want to access. Such a solution is unfeasible for petabyte-scale data samples.

A more attractive approach would be to access data in an completely transparent way through a data management system that provides data transfer on demand and eviction of unused files. In this paper we discuss an implementation that enables such kind of data access using Advanced File Management of IBM's General Parallel File System.

[6] Corresponding author

2. Advanced File Management

IBM General Parallel File System™(GPFS) [1] is a scalable high-performance shared-disk clustered file system for AIX®, Linux® and Windows® developed by IBM®. It is an efficient storage management for big data applications. The use of GPFS as the underlying file system allows a global view of files from any client node. It provides parallel access to the data residing on server nodes through the Network Shared Disk interface. In addition to its file system storage capabilities, GPFS provides tools for management and administration of the GPFS cluster and allows for shared access to file systems from remote GPFS clusters.

A multi-cluster configuration allows to connect GPFS clusters within a data centre, across a campus or via reliable WAN links. GPFS version 3.5 introduces a new feature that enables the sharing of data across less reliable WAN links or, when desirable, to have a copy of the data in multiple locations. This new feature is called Active File Management (AFM). AFM allows to create associations between GPFS clusters, to define the location and the flow of file data and to automate the management of the data. AFM implements a single namespace view across sites around the world once again redefining the scope in the terms of a global namespace. Location and flow of file data files between GPFS clusters can be automated. Relationships between GPFS clusters using AFM are defined at the *fileset* level. A *fileset* is a subtree of a file system namespace that in many respects behaves like an independent file system. A *fileset* in a file system can be created as a *cache* that provides a view to a file system in another GPFS cluster called *home*. Data files are moved into a *cache fileset* on demand.

Cache filesets can be read-only or writeable. Cached data is locally read or written. At the time of reading, if data is not in the *cache* and the amount of requested data is more than a *prefetching threshold*, then GPFS automatically creates a copy. When data is written into the *cache*, the write operation completes locally, then GPFS asynchronously pushes the changes back to the *home* location. Multiple *cache filesets* for each *home* data source can be defined. The number of *cache* relationships for each *home* is limited only by the bandwidth available at the *home* location. Placing a quota on the *cache fileset* causes the data to be cleaned (*evicted*) out of the cache automatically, on the basis of the available space. In the absence of quota, a copy of the data file remains in the *cache* until manually evicted or deleted.

2.1. Data Movement

By means of rich cache management features (see table 1) AFM provides seamless data movement between clusters on demand with a persistent scalable POSIX-compliant cache for remote file system even during disconnection. Updates at *cache* sites are pushed back to *home* asynchronously, queuing updates for later execution, while local writes to cache are done in a synchronous way, providing identical performance as in local file system.

2.2. AFM Communication

Communication in AFM is done using NFSv3 protocol (NFSv4 and native GPFS protocols are available in GPFSv.4.1). GPFS has its own NFSv3 client and automatically manages connection, reconnection and recovery in case of failures of the other cluster. In addition, the built-in NFS client has the ability to parallelize data transfers between clusters, even for a single large data file. It can transfer extended attributes and ACL information for the files. When GPFS is used as the *home* there are many optimizations of available WAN bandwidth.

All data transfers between remote GPFS clusters are managed by the so called *gateway*-nodes. The *gateway*-nodes are used as NFS servers and are the only nodes of GPFS clusters exposed to WAN. Transfer of data *home–cache* can happen in parallel within a *gateway* node or across multiple *gateway* nodes.

Table 1. AFM WAN caching features.

Feature	AFM supports
Granularity	*Fileset* (dir sub-tree) logical namespace mapping
Writable cache	Yes (Coalesces writes, other ops)
Policy based pre-fetching	Yes (uses GPFS policy engine rules)
Policy based cache eviction	Yes (uses GPFS policy engine rules)
Disconnected mode operations	Yes (can also expire based on a timeout)
Streaming support	Yes (uses GPFS policy engine rules)
Locking support	No (only local cluster wide locks)
Sparse file support	Yes (can read as sparse files)
Namespace caching	Yes (gets directory structure along with data)
Parallel data transfer	Yes (can use multi nodes)

2.3. Integration with HSM

GPFS extends its Information Lifecycle Management (ILM) functionalities to allow the integration with HSM (Hierarchical Storage System) products like HPSS or TSM [3].

2.4. Configuration example

An example of a 3-site AFM configuration presented in fig. 1. In this example *site2* and *site3* see all of the data from all sites (forming the *Global namespace*), and write to the site's dedicated *fileset* (directory) on the Home (*site1*), which hosts all home directories and backup/HSM areas.

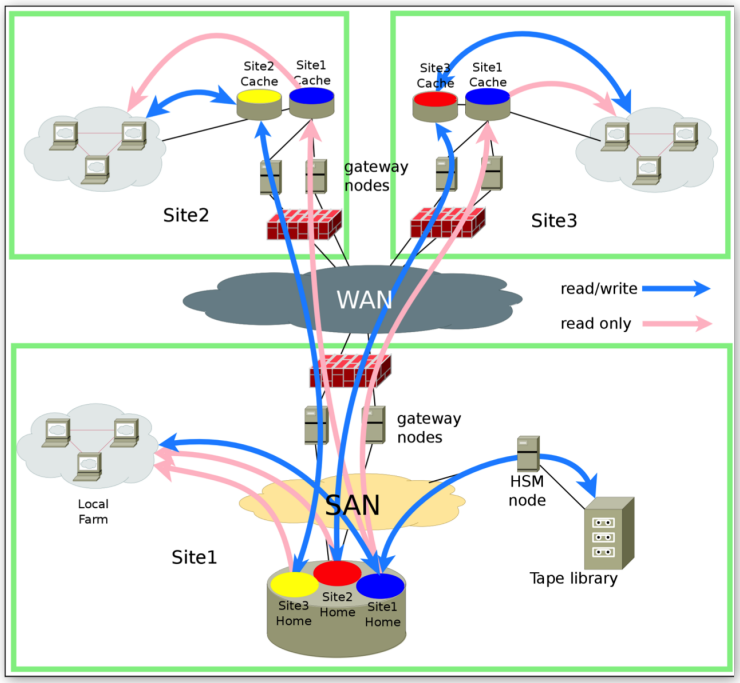

Figure 1. A 3-site distributed storage model with AFM.

3. The CNAF Computing Centre

CNAF is a National Centre of the National Institute of Nuclear Physics (INFN) in Bologna, Italy. It is also devoted to R&D in the field of Information Technologies applied to High Energy Physics (HEP) experiments.

CNAF hosts the Italian World-wide LHC Computing Grid (WLCG [4]) Tier-1 site, the largest Italian computing facility employed in the LHC distributed computing infrastructure. The Centre is also part of the Italian and European Grid Infrastructures (EGI [5]). It also houses computing resources for other particle, astro-particle and physics experiments. The total amount of data stored at CNAF is about 17 PB on-line (on disks) and 18 PB near-line (on tapes).

To provide the highest level of reliability, scalability and performance on the installed mass storage system, a novel solution has been developed and implemented. This solution, called Grid Enabled Mass Storage System (GEMSS) [2], is based on two components:

- a layer between the IBM GPFS (General parallel file system) and the IBM Tivoli Storage Manager HSM;
- StoRM[6], designed to allow direct access to the storage resources by file protocol as well as standard Grid protocols.

The Computing Centre supports many astro-particle experiments, and, in particular, it provides more than 1.5 PB of on-line storage space and 8000 HEPSpecInt of computing power for data processing to the AMS [7] (Alpha Magnetic Spectrometer) experiment. This makes AMS the non-LHC experiment with the largest computing resources at CNAF.

AMS is a large acceptance instrument conceived to search for anti-particles (positrons, anti-protons, anti-deutons) coming from dark matter annihilations, primordial anti-matter (anti-He or light anti nuclei) and to perform accurate measurements in space of the cosmic radiation in the GeV-TeV energy range. AMS has been installed on the International Space Station (ISS) on May 19th 2011 and it is operating continuously since then.

The AMS experiment collects $\sim$ 35 TB of raw data from the ISS each year and produces more than 100 TB per year of reconstructed data, whose format is based on CERN ROOT libraries [8]. In addition to ISS data, more than 200 TB per year of simulated data samples are produced by the AMS Collaboration, crucial for the physics analysis. After 4 years in space AMS collected more than 65 billion events. Each event has an average weight, once reconstructed and written in a ROOT file, of $\sim$ 10 KB. The typical size of an AMS ROOT files is $\sim$ 10 GB.

The AMS computing model [9] is based on the use of a network of computing centres, among them the CNAF, for data processing and Monte Carlo production. Once reconstructed data has been validated, they are copied to the main regional sites and CNAF is one of those. From the repository at CNAF, data has to be accessible from all of the Italian institutions involved in AMS. As in many HEP experiments, due to the amount of data users have to deal with, usually only final histograms or reduced data sets are transferred to local institution sites.

The ASI Science Data Center (ASDC [10]), in Rome, hosts one of the AMS Tier-3s. It has $\sim$ 380 cores and $\sim$ 120 TB of available storage space. As previously noted, performing a complete copy of AMS data set is not a feasible solution, as well as to select a unique subsample of "interesting" events, since the definition of "interesting" depends on the specific analysis each user would like to do. In such a scenario, remote access to data constitutes one of the major challenges.

In 2014 an integrated solution, which enables transparent and efficient access to on-line and near-line data through high latency networks, has been implemented between the ASDC and the CNAF. Owing to AFM, it is possible to define a single, geographically-distributed namespace, characterised by automated data flow management between different locations. This solution was developed in cooperation between CNAF staff and AMS physicists.

Given the limited bandwidth between CNAF and ASDC (see table 2) even the optimisation of all the AFM parameters (such as *Prefetch Threshold*) does not help too much if the ASDC users are requiring to process data not yet cached. The problem has been solved creating a custom mechanism, for the AMS users, to limit the amount of data required from disk, without limiting the potentiality to process any desired particular event from the whole sample. A database (based on ROOT *TTree* objects) with tags of events that have passed certain preselection requirements has been locally created in ASDC. Each data processing job at ASDC queries the preselection database to look for the tags of interesting events, in order to access them (and only them) from a remote file. This gives the users the potential to analyse any desired sub-sample limiting, however, the I/O, and so the usage of the *cache*, only to the interesting events. In this scheme, AFM *Prefetch Threshold* has been tuned to manage 10 GB files (which corresponds to average size of AMS run file) to be accessed randomly. This configuration allows to process the same file remotely paying only a fraction of 15% in execution time compared to that needed to process the same files directly at CNAF. 10 TB of the data storage available at ASDC have been devoted as *cache* for the AFM mechanism.

Table 2. Two site configuration.

Home site location:	CNAF, Bologna
Remote site location:	ASDC, Rome
Distance between sites:	500km
RTT:	23 ms
Bandwidth:	100 Mbps
Home FS size:	1.5 PB
Cache size:	10 TB

4. Conclusions

AFM provides a single namespace with transparent data access via local POSIX calls from remote sites. Configuration of AFM between GPFS clusters is very simple. Parallel prefetch helps when the available bandwidth between sites exceeds the bandwidth of single gateway nodes. When the available bandwidth over WAN is less than the aggregated bandwidth of all gateways, the WAN link can be easily saturated. Many parameters can be tuned on specific use cases, such as the Prefetch Threshold, to specify the amount of a file that should be cached before the whole file is prefetched.

The implemented solution demonstrated the validity of the approach even in the case of limited bandwidth between two data centres. A more performant bandwidth, obviously, would enable more users to perform their analyses without the need to increase the *cache* size.

References

[1] Schmuck F. and Haskin R. 2002, Gpfs: A shared-disk file system for large computing clusters *Proc. of FAST 2002 Conf. on File and Storage Technologies* 2002, http://www.usenix.org/events/fast02/full_papers/schmuck/schmuck.pdf.

[2] P. P. Ricci, D. Bonacorsi, A. Cavalli, L. dell'Agnello, D. Gregori, A. Prosperini, L. Rinaldi, V. Sapunenko and V. Vagnoni J 2012 The Grid Enabled Mass Storage System (GEMSS): the Storage and Data management system used at the INFN Tier1 at CNAF *J. of Physics: Conf. Ser. 396 (2012) 042051*

[3] IBM, Overview - Tivoli Storage Manager Supported Operating Systems, http://www-01.ibm.com/support/docview.wss?uid=swg21243309

[4] WLCG - Worldwide LHC Computing Grid, http://wlcg.web.cern.ch.

[5] EGI - European Grid Infrastructure, http://www.egi.eu.
[6] StoRM - Storage Resource Manager. http://storm.forge.cnaf.infn.it
[7] M. Aguilar et al. 2013, *Physical Review Letters* **110**, 141102.
[8] Brun R and Rademakers F 1996, ROOT - An Object Oriented Data Analysis Framework, *Proc. AIHENP'96 Workshop (Lausanne)*, *Nucl. Inst. & Meth. in Phys. Res.* A **389** 81-86 (1997). See also http://root.cern.ch/.
[9] B. Shan for the AMS Collaboration, "Computing strategy of AMS-02 experiment", this proceeding.
[10] ASI Science Data Center: http://www.asdc.asi.it