An integrated solution for remote data access

V. Sapunenko[1], D. D'Urso[2], V. Vagnoni[3], L. dell'Agnello[1], M. Duranti[4]

[1] INFN CNAF, Viale Berti Pichat 6/2, 40127 Bologna, Italy
[2] INFN and ASDC, via del Politecnico snc, 00133 Roma, Italy
[3] INFN Sezione di Bologna, Via Irnerio 46, 40126 Bologna, Italy
[4] University of Peugia and INFN, via A. Pascoli, 06123 Perugia, Italy

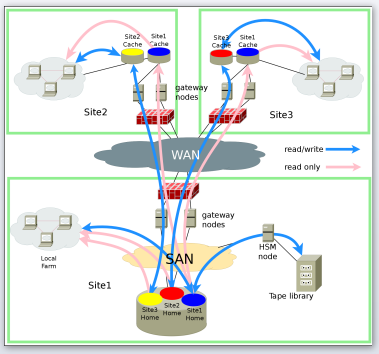
Introduction

- 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.
- This 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 world. Owing to a new feature introduced in GPFS 3.5, so-called Active File Management (AFM), the definition of a single, geographically-distributed namespace, characterized 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.

AFM: Advanced File Management in GPFS

scalable, high-performance, file system caching layer integrated with the GPFS cluster file system

- Enables sharing data across unreliable or high latency networks.
- Location and flow of file data between GPFS clusters can be automated.
- Relationships between GPFS clusters using AFM are defined at the fileset level.
- 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." File data is moved into a cache fileset on demand.
- Transfer home ← → cache can happen in parallel within a gateway node or across multiple gateway nodes.
- 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.



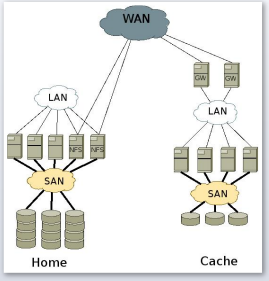
Feature	AFM support
Granularity	Fileset (dir tree) logical namespace mapping
Writable cache	Yes (uses GPFS policy rules select files to replicate). Coalesces writes, other ops
Policy based pre-fetching	Yes (uses GPFS policy engine rules)
Policy based cache eviction	Yes (uses GPFS policy engine rules)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

Data movement

- Communication is done using NFS (v3, v4).
  - Standard protocol can leverage standard WAN accelerators.
- GPFS has its own NFSv3 client.
  - Automatic recovery in case of failures.
  - Parallel data transfers (even for a single file).
  - Extended attributes and ACLs are also transferred.

Use Case: CNAF (Bologna) – ASDC (ASI, Rome) data processing for the AMS experiment

- Home site location: CNAF, Bologna.
- Remote site location: ASDC, Rome.
- Distance between sites: 500km.
- RTT: 23 ms.
- Bandwidth: 100 Mbps.
- Home FS size: 1.1 PB.
- Cache size: 10 TB.



- A DB (based on ROOT TTree objects) with tags of events that have passed certain preselection requirements has been locally created.
- Each data processing job queries the preselection DB to look for the tags of interesting events, in order to access them (and only them) from a remote file.
- AFM Prefetch Threshold has been tuned to manage 10 GB files accessed randomly and sequentially.
- The final configuration allows us to process the same file remotely paying only a fraction of 15% in execution time.

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.

References

1. A. Cavalli et al., "StoRM-GPFS-TSM: a new approach to Hierarchical Storage Management for the LHC experiments", Journal of Physics: Conference Series 219(2010) 072030
2. GPFS. Available Online at http://www-03.ibm.com/systems/clusters/software/gpfs/index.html
3. AFM. Available Online at http://www-01.ibm.com/support/knowledgecenter/SSFKCN_3.5.0/com.ibm.cluster.gpfs.v3r5.gpfs200.doc/bl1adv_afm.htm
4. TSM. Available Online at http://www-01.ibm.com/software/tivoli/products/storage-mgr