WLCG storage, Cloud resources and volatile storage into HTTP/WebDAV-based regional federations

F.Furano, R.Sobie, S.Pardi, A.De Salvo, O.Keeble

The goal of this demonstrator is to evaluate regional federations of stable or volatile storage services that can be seen as a unique read/write multi-tier entity, using HTTP protocols for HEP applications

1 Introduction

The goal of this demonstrator is to create regional federations of stable and volatile storage services that can be viewed as a single entity and accessed using commonly-used, open-source protocols from the applications.
This system will further decouple the compute resources from the storage, and simplify access to data from non-WLCG sites to WLCG Storage Elements ("stable storage").
As the HEP community expands its use of opportunistic (private and commercial) resources, such as IaaS computing clouds, the data federation would facilitate access to input data and simplify the transfer of output data. Further, the data federation opens up the use of non-WLCG storage ("volatile storage") that can be used for short or medium term buffering of input data.

2 Background

Currently there are many WLCG storage systems distributed around the world and there is a view within the WLCG community that fewer large storage sites would reduce operational and manpower costs.
This has encouraged a number of technical explorations on consolidating existing site storage setups into multi-tier storage deployments that have a common primary entry point for data access. Of interest to the project teams, is the Dynafed HTTP/WebDAV-based federation technology that could be a solution to unifying the existing WLCG storage systems.

As we continue to expand our use of opportunistic resources (specifically, clouds), we want to accommodate volatile storage resources. Volatile storage can mean a storage service that can be added or removed quickly, generally under the control of the user but possibly as a result of the requirements of the provider (eg. quota reduction). The volatile storage can be used as a data file cache for the application job either on the worker node or within a larger storage system. An example of a volatile or "opportunistic" storage is the REST-based cloud storage resources provided, for example, by Amazon S3 and Microsoft Azure.

Even in the presence of volatile storage, the data management system of the experiment (e.g. ATLAS) may need to be made aware of file addition/removal/modification in the multi-tier federated storage. One advantage of this kind of consolidation is that the central catalogues may benefit from a reduction in the granularity of the sites whose content they have to index, and that the management of the support could be simplified.

In the next sections we briefly describe the key elements of the demonstrator projects. This includes the Dynafed data federation service, the distributed cloud computing system used by ATLAS and Belle II, and the Belle II data federation.

The demonstrator involves teams from CERN-IT, ATLAS and Belle II.

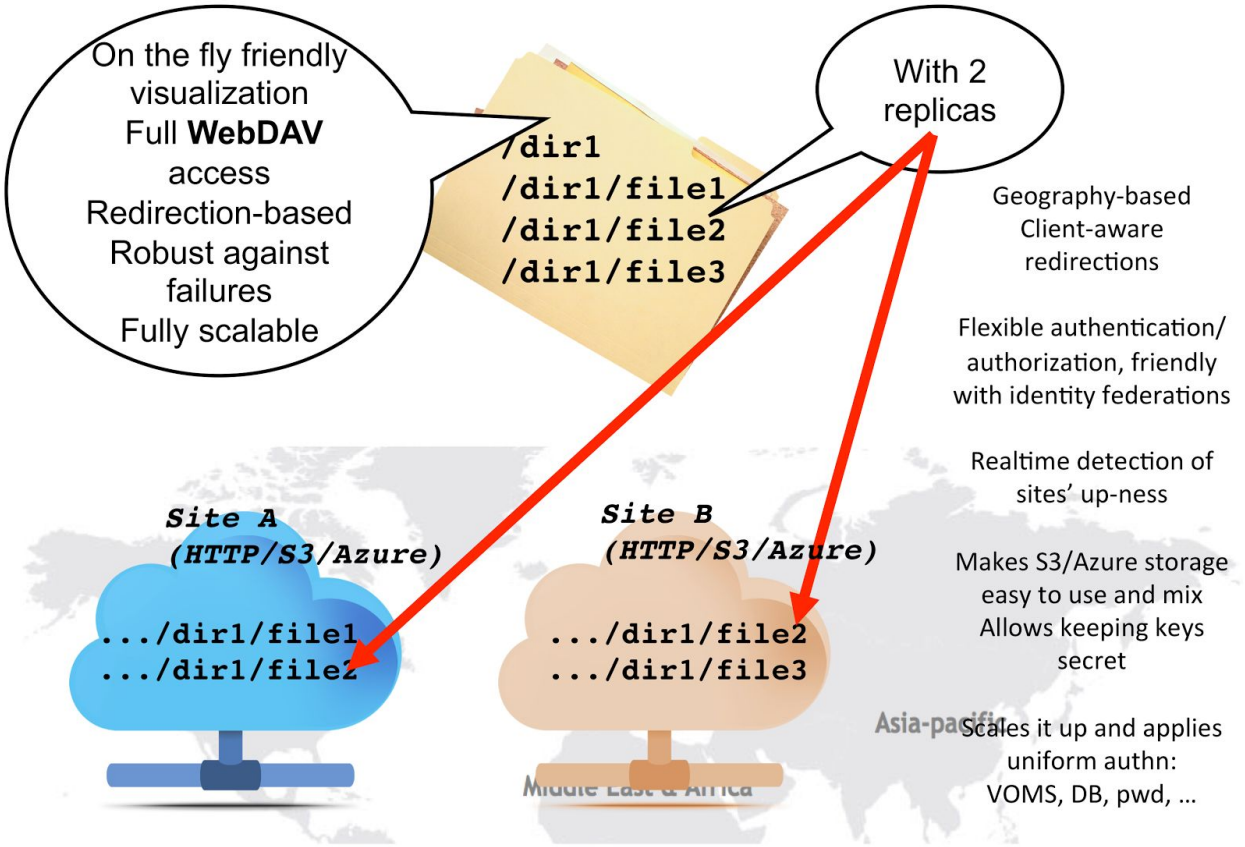
3 Dynafed data federation service

The Dynafed system gives seamless access, via HTTP and WebDAV, to a repository that is presented by merging and caching in memory metadata items that are taken on the fly from a number of eventually remote storage systems or endpoints.
Dynafed natively supports HTTP, WebDAV, S3 and Microsoft Azure as protocol dialects to communicate with the endpoints. In the S3 and Azure case, the system fully supports managing hierarchical content and applying a uniform authorization scheme to any number of buckets from different providers. The authorization scheme allows not to disclose private S3/Azure keys to the clients. The policies can be based on simple rules or on more sophisticated inline scripting.

One of its fundamental features is its' ability to redirect GET/PUT requests to a storage system hosting the requested data that is closest to the requesting client. The focus of the design is on performance, scalability and realtime fault resilience with respect to the status of endpoints.

From the perspective of a normal user, using HTTP and WebDAV clients, they can browse the Dynamic Federation as if it were a large, monolithic filesystem, being redirected to the right host when they ask for a file replica. Dynafed also supports writing. Sysadmins will appreciate the integrated dashboard, providing real time information about the up-ness of endpoints and their history.

The goal of the system is to provide a multi-site, distributed repository as if it were one entity, without the necessity of maintaining an internal catalogue of all the files it contains. Of course external catalogues of what is supposed to be accessible in a federation will work normally, and such a macro-site can be normally indexed by experiments catalogues.



The CERN IT team developed the Dynafed technology in the last few years, and gives support for its usage in the context of workflows and visualizations that use the HTTP protocol. The technology is considered mature and stable and is currently used in production systems.

The ATLAS teams include the ATLAS-ITALY and ATLAS-Canada groups, in close contact with the Rucio development team. An invitation to join will be made to the ATLAS Latin America group.
The Belle II team includes the Italian and Canadian Belle II groups.

The groups have a wide range of complementary experience in data management, cloud computing, storage and networks.

The demonstrator project has its roots into some similar explorations that have been undertaken

independently by the participants. The project will help focus the efforts towards a common solution that will benefit the HEP community.

The previous and ongoing explorations have been:

- The ATLAS Canadian team has evaluated the Dynafed technology for aggregating sites and Cloud storage in the last 3 years, and the start of an official nation-wide project is very likely to happen during 2016.
- A similar work had been performed by ATLAS Italy and presented as a poster at CHEP 2015.
- Belle II computing has an ongoing R&D project aimed at federated, HTTP-based data access.

Among the main technological providers there is DPM team, belonging to CERN IT-ST-AD, providing support for DPM and Dynafed :

http://lcgdm.web.cern.ch/dynafed-dynamic-federation-project

http://lcgdm.web.cern.ch/dpm

4 Canadian experience: Distributed cloud system for ATLAS and Belle II

One of the groups in the project has designed and constructed a system that uses independent dedicated and opportunistic cloud resources for particle physics and astronomy applications.
The system utilizes clouds in Europe and North America for applications from the ATLAS and Belle II experiments. It has been in production operation for over three years typically using 2000-4000 cores and has successfully completed many millions of jobs.
A full description of the system was given at the CHEP 2015 conference (http://heprc.phys.uvic.ca/sites/heprc.phys.uvic.ca/files/Gable-CHEP-2015.pdf) and its use by ATLAS (http://iopscience.iop.org/article/10.1088/1742-6596/664/2/022038) and Belle II (http://iopscience.iop.org/article/10.1088/1742-6596/664/2/022037) were also presented at CHEP 2015.

The system can run simulation (low I/O), high-memory and analysis (high I/O) applications, however, analysis applications are only submitted to clouds with a local WLCG Storage Element (SE). The high-memory jobs are currently limited to the ATLAS Compute Cloud at CERN. Most of the workload for ATLAS and Belle II are simulation jobs. In this case, the output of the jobs is written to a local cache on the node hosting the virtual machine. Once the job is completed, the output is transferred to an SE. At the moment, the output SE is specified in the job parameters but can fail over to another SE if the primary SE is offline.

The distributed cloud uses private and commercial opportunistic resources. These clouds have none of the usual HEP services and storage. To expand the functionality of the system so that it can run both low I/O and high I/O applications, we plan to utilize an HTTP data federation service (such as Dynafed). This will enable the system to write the output files to the optimal SE (one of our first goals is use the federation service for the output of the Belle II simulation jobs). In order to run analysis jobs on opportunistic resources, we would use the data federation service to find the optimal location of the input data, retrieve it to a local cache, process the data and remove it from the cache. Utilizing the volatile storage of the opportunistic clouds could provide a further option for storing and analyzing HEP data.

5 Belle II

In order to deal with the large amount of data it produces, Belle II employs a distributed computing system, spread over more than 20 countries.
In the current version of the Belle II Computing Model, data are distributed over multiple sites, hosting RAW data, reconstructed data (called MDST) or MC data.
Sites use heterogeneous storage elements, which have to provide services for data access, data copy and remote access, in order to support multiple analysis paradigms.
Among Sites, we call Regional Data Center the ones that store the permanent copy of files (produced at example during the MC campaign), and registered in the file catalog based on LFC.
The possibility to use a dynamic federation could help to take advantage from multiple copies of files even if present in ephemeral volumes, or in a cloud storage, with a simplified file catalog management.
Currently there is an R&D activity that involves several sites of the Belle II collaboration, with the goal to demonstrate the possibility to use efficiently storage resources with highly popular protocols as http/webdav and to test the benefit in using dynamic federation system.

6 Italian experience: comparing XROOT, HTTP and HTTP Federation access technologies and evaluating the Dynafed capabilities

In http://iopscience.iop.org/article/10.1088/1742-6596/664/4/042019/pdf the Dynafed technology has been tested in the context of ATLAS, with a PROOF-based analysis that accesses the input datasets through a Dynamic Federation with HTTP/WebDav. The federator server was located at DESY and the SEs were based on DPM.

The federator has given transparent access to the storage resources through the HTTPS/WebDav protocol, presenting a unique name space and redirecting the requests to the

closest available SE, so that HTTP and WebDAV clients could browse the Federation and directly download files using the Rucio syntax.

Moreover, the federator server always was aware of the endpoints' status, that are checked every 10 seconds. If the httpd daemon goes offline, that storage system was not used by the federator. When the HTTP protocol becomes available again, the federator tested it for a few minutes before enabling it again in the Federation.

For the tests a PROOF-based analysis was run on a PROOF cluster that was setup in the ATLAS Frascati Tier-2, with the input dataset stored in two DPM SEs: at Frascati and in the Naples ATLAS Tier-2 and both DPMs were part of the HTTP federation.

Two types of tests were run. In the first one the performance accessing the input dataset were evaluated, comparing the access with XRootD, HTTPS/WebDav directly and HTTPS/WebDav through the federator server and it was shown that the overhead due to the federator is small.

In the second test, the fallback mechanism of the Dynamic Federation was tested, to show that the traffic is redirected to another server in the Federation when the original one is stopped.

During the second test the analysis task running in the Frascati Tier2 was accessing (through the federator) a dataset that was available both in Frascati and Naples DPM SEs. During the tests the network connections of the PROOF cluster nodes were monitored, with a focus on those towards the SEs and the federator. The first PROOF task was launched while both SEs were up and the federator effectively redirected the workers to the Frascati SE, as expected.

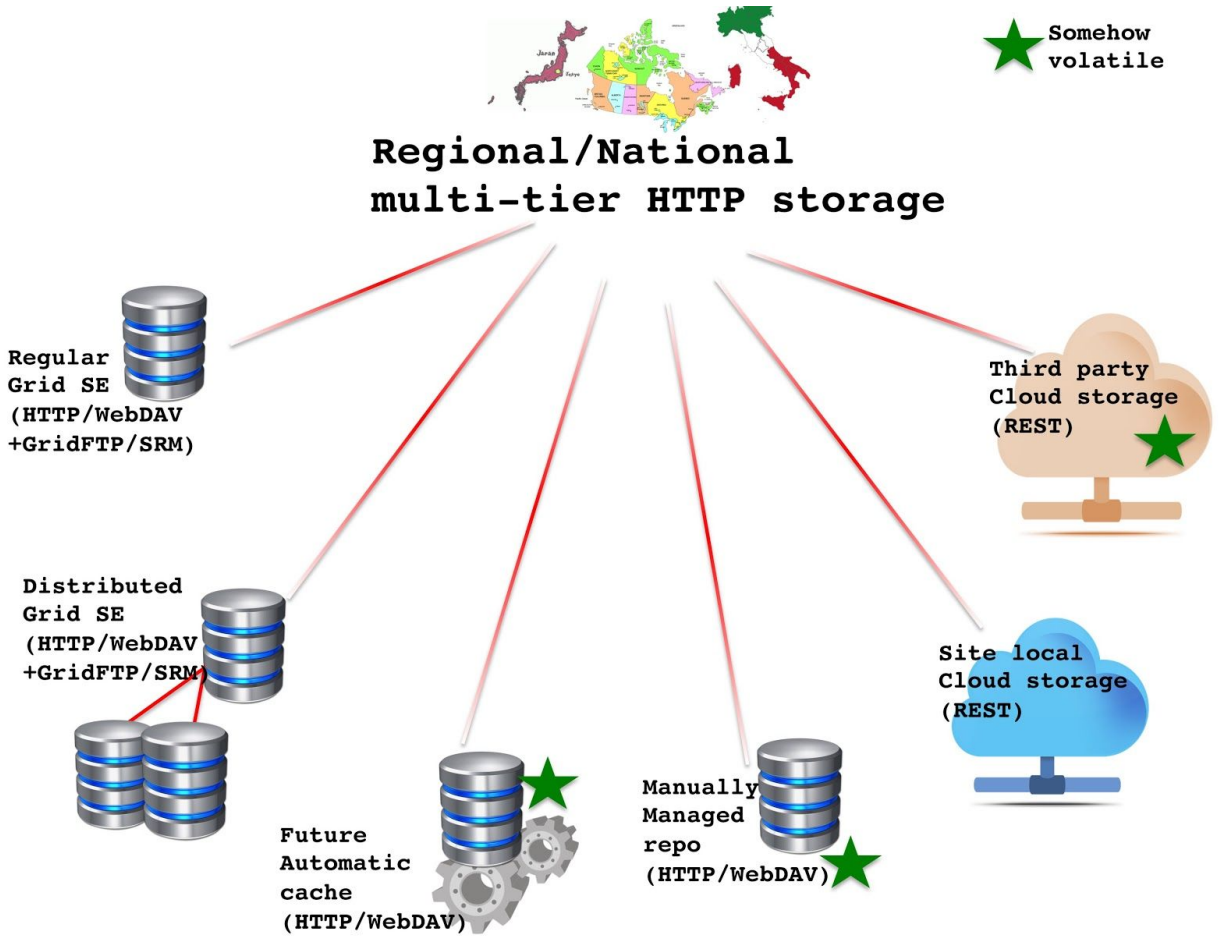
To test the federator responsiveness, the task was launched a second time after bringing down the Frascati SE; the federator in this case correctly redirected the workers to the Naples SE. After two minutes of task execution, we enabled again the Frascati SE and, in about 5 minutes all the workers stopped accessing data from Naples and reconnected to the SE in Frascati, thus proving the validity of the concept; a forced downtime of a storage endpoint was gracefully recognized and the new data connections of the jobs were steered to the closest working site.

7 Goals and conclusion

The goal of this demonstrator is to evaluate regional federations of stable and volatile storage services. These can be seen as a unique multi-tier entity from the point of view of applications reading and writing data into them or indexing parts of their content. The storage services can be, interchangeably:

- Regular WLCG storage elements exposing an HTTP/WebDAV interface (together with the other usual protocols)
- REST-based cloud storage that can be acquired by third parties (e.g. AWS or Azure)

- REST-based cloud storage that can reside in one of the participating sites (e.g. using CephS3)
- “Cache”-type endpoints able to work as data caches that are local to the federation. These objects actually do not exist in production yet. Should they become available, one of the goals of the demonstrator is to evaluate the ease of an eventual integration.



This demonstrator does not take into account the technical possibility of gathering all the mentioned resulting multi-tier storages into a global federation.