A lightweight federation of the Belle II storages through Dynafed

S Pardi\textsuperscript{1} and G Russo\textsuperscript{1,2}
\textsuperscript{1}INFN-Napoli Unit – Italy
\textsuperscript{2}University of Napoli Federico II – Italy
E-mail: spardi@na.infn.it

Abstract. Belle II experiment can take advantage from Federation technologies to simplify data management. The increasing adoption of HTTP and WebDAV protocol by sites, enable to create lightweight solutions to give an aggregate view of multiple distributed storages. In this work we present a study on the possible usage of the software Dynafed developed by CERN for the creation of a dynamic data federation system. We created a first Dynafed server, hosted in the datacentre in Napoli, and connected with about the 60\% of the production storages of Belle II. Then we aggregated all the file systems under a unique HTTP path. Over this infrastructure we performed a set of stress tests, in order to evaluate the impact of federation, the service resilience, and to study the capability of Dynafed to redirect clients properly to the most convenient file replica. The results show good potentialities of the service and suggest further investigation.

1. Introduction

Data Federation is a powerful technology which allows researchers to manage their datasets in a very effective and resilient way, by offering to the final user a transparent access to multiple file replicas geographically distributed. In the last years large High Energy Experiments like CMS and ATLAS have developed ad-hoc Data Federation systems based on xrootd protocol, and have integrated such federation services within the user analysis tools.

However, HTTP and WebDAV have become popular protocols for data analysis applications, with a growing interest from the HEP community. At the same time, Cloud storage with S3 interface has become a largely used protocol as well, thanks to the growing market of public clouds. In this context a team of CERN has developed a lightweight federation service called Dynafed [1], that is able to federate and integrate together HTTP/WebDAV, S3 and other endpoints with a standard interface.

Within the investigation activities of the Belle II experiment [2], a large testbed of Dynafed has been setup, integrating a set of storages currently in production. The goal of this R&D activity is to evaluate performance and opportunities offered by this kind of federation system to support researchers in their day-by-day activities.

The rest of the paper is organized as follows: in Section II we describe how Dynafed works and in which way it could be used within the Belle II collaboration. In section III there are all details related to our testbed, hardware and software configuration. In section IV we introduce our test plan and we show the
results after job probes execution. Finally, we summarize our results, and we give some hints for further investigation in section V.

2. HTTP Data Federation opportunity for Belle II

The Belle II collaboration is investigating the possibility to use open protocols and standard interfaces to their storages, such HTTP and WebDAV. The goal is to simplify the user’s life and to reduce, in the long term, the costs of storage management without sacrificing features and performance.

One of the most interesting tools, that takes advantage from the HTTP interface of storage systems, is the Dynamic Federations system Dynafed developed by CERN. Dynafed is able to aggregate multiple endpoints geographically distributed, presenting to the final users a global storage with a unique virtual name space.

The interesting feature for Belle II use-case is the possibility to manage file replicas in a very easy manner. Indeed Dynafed is able to recognize multiple copies of a single file and represent them as a single metalink. The global file system representation is created on-the-fly in a local cache memory, and does not require any additional services on the site-side (see schema in figure 1).

Moreover with Dynafed it is even possible to aggregate standard HTTP/WebDAV endpoints with Cloud storage resources accessible via s3 or MS Azure interface.

![Diagram of Dynafed Server](image.png)

**Figure 1** – Schema showing how the Dynafed Server aggregates multiple file replicas by creating, on the fly, a metalink as single representation for multiple distributed copies.

3. The Belle II Dynafed implementation

3.1. The Belle II storage infrastructure

A Dynafed server for the Belle II experiment has been setup in the site of PRISMA/ReCaS-Napoli[3][4][5] in order to test features and performance. The federation service integrates a set of 14 storages selected among the SEs in production that offer the HTTP/WebDAV interface. The testbed includes storages of different technologies such as DPM, dCache and STORM, distributed in a wide
geographic area: 9 in Europe, 1 in Asia, 2 in North America and 2 in Australia (figure 2). The federation service covers the 60% of the Belle II storages currently available. The Dynafed server in installed on a virtual machine with 4 cores, 8GB of RAM and 10Gbit connection to the WAN.

![Figure 2 – Geographic distribution of the storages used in our testbed](image)

3.2. Dynafed Views

Taking advantage from the capability of Dynafed to create multiple mount points easily and instantly, we have aggregated the available endpoints in 4 different views:

- **myfed/PerSite/** Shows the file systems of each storage separately (without aggregation)
- **myfed/belle/** Aggregation of all the directory /DATA/belle and /TMP/belle/
- **myfed/site-based-path/** Aggregation of all the root directory of different storages
- **myfed/s3-federation/** Testing area for cloud storage

The final result is a fast-browseable virtual file system (figure 3), which users can access easily with their preferred browser to list and download files, using standard grid certificate and VOMS extensions as authorization and authentication method.

The testbed has been setup without any additional work effort from administrators of sites providing the storage endpoint, and it is immediately accessible by researchers.
4. Performance analysis

Thanks to the aggregation feature provided by Dynafed, we can address a specific file and all the relative replicas with a single url as follow:

https://dynafed01.na.infn.it/myfed/belle/TMP/belle/user/spardi/test

https://mixed_e0001r0009_s00_BGx1.mdst.root

At access time, Dynafed is the able to redirect the client on the most convenient replica, thanks to the geoip-plugin that we enabled in our testbed.

In order to use the meta-link directly in a steering file of the Belle II framework for analysis and simulation (basf2), we recompiled the software using the Libdavix[6] library. In that way we are able to access the Dynafed universal URLs in streaming, within a standard Belle II job.

On the described HTTP/Federated Ecosystem [7], we tested the performance in accessing to data through Dynafed. To do that, we populated all the storage elements of the testbed with a set of identical ROOT files, produced during the 4th belle II MC campaign, each one of 10MB size. Then we created a set of test probes performing the following steps, through a basf2 analysis job:

- Step 1: The probe launches a set of analysis jobs each one accessing the same file remotely via the HTTP protocol, using a specific Storage Element, i.e. addressing the file with the storage URL

- Step 2: The probe launches the same analysis job using the Dynafed URL

We sent the job probes to different Grid Sites, multiple times in order to have better statistics. For each analysis job, we logged the latency between client and the specific storage element, and the time spent by the root library to read the specific file.

The histogram in figure 4 shows the performance of the RootInput function, obtained running the testing probe on a Worker Node in the site of ReCaS-Napoli. The grid job download the same file from 11 different storage elements of the testbed, selected among the most stable at the testing time. The RootInput times are ordered by latency between storage and client. In the last column we have the performance obtained using the Dynafed URL+geoip-plugin.
Results show that using Dynafed, the client is redirected to closer storages, that in this case are in the same geographic areas. In that way users can access data in a convenient way, without knowing the file location, and without the necessity to manage replica catalogues.

![Graph showing time in seconds spent to complete the RootInput routine vs Latency (ms) between client and storage, compared with the performance obtained addressing the file via Dynafed.](image)

**Figure 4.** – Time in seconds spent to complete the RootInput routine vs Latency (ms) between client and storage, compared with the performance obtained addressing the file via Dynafed.

5. Conclusions and future works

A new service of data federation based on Dynafed has been setup for the Belle II collaboration, in the site of ReCaS-Napoli. The testbed includes about the 60% of the storages in production and uses HTTP/WebDAV as main data access protocols.

Over this infrastructure a bunch of tests has been performed, the results demonstrate the possibility to use Dynafed URLs to access transparently a data set from the Belle II standard framework for analysis and simulation basf2. Moreover thanks to the geoip-plugin, users can access a more convenient replica depending on the location of the client.

The next goal is to extend the testbed to all the Belle II storages and perform user analysis jobs to evaluate the full ecosystem.

6. References

[1] Furano F, Brito da Rocha R, Devresse A, Keeble O, Alvarez Ayllon A and Fuhrmann P, “The Dynamic Federations: Federate Storage on the fly using HTTP/WebDAV and DMLite”, 2013 The International Symposium on Grids and Clouds - DOI: 10.13140/RG.2.1.3584.3446

[2] Pardi S et al. “Computing at Belle II” - Belle-II computing Group - Nucl.Part.Phys.Proc. 273-275 (2016) 950-956 - DOI: 10.1016/j.nuclphysbps.2015.09.148

[3] Delprete D, Pardi S and Russo G - “A Grid monitoring model over Network-Aware IaaS Cloud Infrastructure”- Journal: Int. J. of High Performance Computing and Networking, 2013 Vol.7, No.3, pp.195 – 204
[4] Palmieri F, Ficco M, Pardi S, Castiglione A - “A cloud-based architecture for emergency management and first responders localization in smart city environments” - Journal Computers and Electrical Engineering - Volume 56 Issue C, November 2016 Pages 810-830 doi 10.1016/j.compeleceng.2016.02.012

[5] Alfano M et al. “Network Management in Cloud Computing for Public Administration: A Practical Use Case” - Proceedings 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing, UCC 2014 29 January 2014, Article number 7027592, Pages 768-773. DOI: 10.1109/UCC.2014.125

[6] Devresse A, Furano F - “Efficient HTTP Based I/O on Very Large Datasets for High Performance Computing with the Libdavix Library” in Big Data Benchmarks, Performance Optimization, and Emerging Hardware. BPOE 2014. Lecture Notes in Computer Science, vol 8807. Springer, Cham. DOI: 10.1007/978-3-319-13021-7_15

[7] Furano F et al. Towards an HTTP Ecosystem for HEP Data Access, J. Physics: Conf. Series 513, 032034 (2014). HTTP://stacks.iop.org/1742-6596/513/i=3/a=032034

Acknowledgments
Authors wishing to acknowledge assistance of the staff of the ReCaS Data centre, the computing group of Belle II and the site admin of the Belle II production infrastructure.