20th International Conference on Computing in High Energy and Nuclear Physics (CHEP2013) IOP Publishing
Journal of Physics: Conference Series **513** (2014) 032034 doi:10.1088/1742-6596/513/3/032034

Towards an HTTP Ecosystem for HEP Data Access

Fabrizio Furano, Adrien Devresse, Oliver Keeble, Martin Hellmich, Alejandro Álvarez Ayllón

CERN, European Organization for Nuclear Research

E-mail: fabrizio.furano@cern.ch, adrien.devresse@cern.ch, oliver.keeble@cern.ch, martin.hellmich@cern.ch, alejandro.alvarez.ayllon.cern.ch

Abstract. In this contribution we present a vision for the use of the HTTP protocol for data access and data management in the context of HEP. The evolution of the DPM/LFC software stacks towards a modern framework that can be plugged into Apache servers triggered various initiatives that successfully demonstrated the use of HTTP-based protocols for data access, federation and transfer. This includes the evolution of the FTS3 system towards being able to manage third-party transfers using HTTP. Given the flexibility of the methods, the feature set may also include a subset of the SRM functionality that is relevant to disk systems.

The application domain for such an ecosystem of services goes from large scale, Grid-like computing to the data access from laptops, profiting from tools that are shared with the Web community, like browsers, clients libraries and others. Particular focus was put into emphasizing the flexibility of the frameworks, which can interface with a very broad range of components, data stores, catalogues and metadata stores, including the possibility of building high performance dynamic federations of endpoints that build on the fly the feeling of a unique, seamless very efficient system. The overall goal is to leverage standards and standard practices, and use them to provide the higher level functionalities that are needed to fulfil the complex problem of Data Access in HEP. Other points of interest are about harmonizing the possibilities given by the HTTP/WebDAV protocols with existing frameworks like ROOT and already existing Storage Federations based on the XROOTD framework. We also provide quantitative evaluations of the performance that is achievable using HTTP for remote transfer and remote I/O in the context of HEP data. The idea is to contribute the parts that can make possible an ecosystem of services and applications, where the HEP-related features are covered, and the door is open to standard solutions and tools provided by third parties, in the context of the Web and Cloud technologies.

1. Introduction

The support of HTTP/WebDAV, provided by frameworks for scientific data access like DPM [6], dCache [9], STORM [10], GFAL2 [11], FTS3 [12] and foreseen for XROOTD [2], can be seen as a coherent ensemble – an ecosystem – that is based on a single, standard protocol, where the HEP-related features are covered, and the door is open to standard solutions and tools provided by third parties, in the context of the Web and Cloud technologies.

Among our contributions are a fully standard and full-featured client library to ROOT 5 and 6, and the design of a plugin that adds HTTP/WebDav support to the XROOTD framework. At the same time, the Dynamic Federations project gives the possibility of clustering HTTP/WebDAV resources across LANs and WANs in a very efficient and scalable way, thus

allowing one to build large federations of HTTP-based storage sites or file metadata databases.

The application domain goes from large scale, Grid-like computing to data access from laptops, profiting from tools that are shared with the Web community, like browsers, clients libraries and others.
The overall goal is to leverage standards and standard practices, and use them to contribute the components that are needed to fulfil the complex problem of Data Access in HEP while using technologies and toolsets that are attractive also for non-HEP communities, making long term sustainability a closer goal.

2. Enabling scalable, uniform services based on HTTP and WebDAV

The storage elements used in the Grid and for HEP belong to the categories of software systems whose goal is to cluster storage resources and make them available to data processing clients as if they were a unique system. Many other projects have analogous functionalities, in the Scientific community (e.g. DPM, dCache, iRODS, Xrootd) or in the market.

Most of these products work by managing and distributing data across mountpoints that are spread through many servers. Clients that want to access the repository are redirected to the best machine according to various criteria, typically given by some clustering algorithm. There are different clustering techniques using various technologies (p2p-like, database, hashes, ...) and the goal of such systems is always to hide to the users doing analysis and data access the complexity of managing site resources, for repositories that are too big for a single server to be able to sustain the load that the clients would generate.
From the technical point of view, one of the main goals of this kind of systems is to hide to the users doing analysis and data access the complexity linked to the "where is my file problem". The central systems that schedule analysis jobs to be submitted, or the users themselves do not need to deal with details that are internal to the site (e.g. the names of the mount points, which can change); they need to interact with a coherent storage service that is offered and administered by the site.

2.1. DPM and DMLite

One of our most important contributions to the usage of HTTP in High Energy Physics has been the evolution of the Disk Pool Manager (DPM) architecture towards our scalable and flexible framework for designing data management systems, called *dmlite* [4].

The Disk Pool Manager (DPM) is a lightweight solution for grid enabled disk storage management. Operated at more than 200 sites, it has the widest distribution of all grid storage solutions in the WLCG infrastructure. It provides an easy way to manage and configure disk pools, and exposes multiple interfaces for data access (xrootd, NFS, GridFTP and HTTP/WebDAV) and control (SRM).

Some of the most important supported features are:

- Provide HTTP multi-stream transfers for high performance wide area, access matching it with the strict requirement of DPM about coordination of accesses.
- Support for third party copies.
- Support for X.509 authentication with proxy certificates and VOMS extensions
- Support for user credential delegations.
- Support for modern configuration and monitoring solutions based on the industry standards Puppet and Nagios.

2.2. XrdHTTP: Accommodating resources managed in Xrootd clusters
One way we foster the HTTP ecosystem is by contributing the components that fill the gaps between the functionalities that are needed by the World Wide Web browsing and the more advanced functionalities that are needed by intensive data analysis applications.
In this case, our goal has been allowing a storage cluster using the Xrootd framework to join a computing model that needs HTTP access or a Storage Federation based on the HTTP and WebDAV protocols, fully supporting X.509 authentication and the related proxy certificates. For this use case we have designed XrdHTTP.
The xrootd framework has a versatile multiprotocol architecture that allows a developer to use the internal features of the framework to implement data access protocols. One more advantage of this approach is that all the features of a preexisting setup (e.g. monitoring, tapes, etc.) are available to all the loaded protocols automatically.
At this time XrdHTTP is in advanced testing phase. We expect XrdHTTP to be released to the EPEL distribution at the beginning of 2014, following the availability of the version 4 of the Xrootd framework.

2.3. The Dynamic Federations
The goal of the Dynamic Federations system is to federate storage sites and metadata endpoints that expose a suitable data access protocol, into a transparent, high performance storage federation that exposes a unique name space. The architecture can accommodate LFN/PFN algorithmic name translations without the need of catalogues. On the other hand, if catalogues are needed, several of them can be accommodated into the same federation. The idea is to allow applications to access a globally distributed repository, in which sites participate. The applications would be able to efficiently access data that is spread through different sites, by means of a redirection mechanism that is supported by the data access protocol that is used. The focus is on standard protocols for data access, like HTTP and WebDAV, and NFS can be considered as well. The architecture and the components of such a system are anyway detached from the actual protocol that is used.
The system can also accommodate on-the-fly geography-based weighing of clients and replicas.

Another point that is important for our design, is that such a system should be efficient also in the browsing case, e.g. allowing an user to list the content of a directory in a fast and reliable way that does not impact the performance of the whole system.

2.4. Federating Grid storage with third-party farms via HTTP/WebDAV
Given a number of storage endpoints deploying the WebDAV door of the dCache system, and the upcoming versions of DPM [3] [4], we wanted to show that a completely transparent federation of them was possible, using the WebDAV protocol. This use case has been the first one to be demoed by the Dynamic Federations project, and the first two endpoints that were added to a working federation have been a dCache instance at DESY (Germany) and a DPM instance in ASGC (Taipei). The test did what it advertised, i.e. the users could not realize that they were browsing and using a federation of two distant sites. Moreover, the feeling of performance that the system gives is the one of a site that is hosted in the federation's frontend machine, with a fast and smooth interactivity.

This would allow users to browse their files using Internet Explorer without potentially being aware of the location of the items they see, and to run their personal analyses pointing their applications to the unique entry point, using the URLs that they see in the browser.

This use case may also accommodate the use case of the "Cloud storage providers". Technically, we chose to use a Cloud storage service provided by T-Mobile (Germany) through WebDAV, which then became a standard component of the various demos of the Dynamic Federations system. The fact that our Dynamic Federations system applies a dynamic behavior to the problem of federating storage and metadata endpoints also opens the possibility of federating storage endpoints whose content may change at a faster pace with respect to a regular storage element, like distributed file caches.

3. Contributing DAVIX: a full-featured client

Davix is a toolkit that we created for High Performance Remote I/O with HTTP-based protocols in a High Performance Computing environment. It is currently one of our contributions to the EPEL [13] software distribution.
The need for such a component was triggered by the very uneven support for advanced features in the mainstream HTTP clients. To give an example, X.509 authentication, which is fundamental for Grid usage, is supported only partially and by one or two clients, which do not support other fundamental features, like the vectored access.
Davix provides a set of simple command line tools and a high level POSIX-like API for fast optimized data access, abstracting the details of the HTTP protocol and of the interactions with the servers. Davix has been built as a layer on top of the official WebDAV library, called libneon [8] .
Davix supports WebDAV and the S3 protocol for meta-data operations and file manipulation. It also supports supports advanced features like:

- Vectored read operations, that higher the I/O performance for the applications that can use them (e.g. ROOT I/O).
- X509 and VOMS authentication, compatible with the Grid standards.
- Authentication session pools, to reuse previously established authentication sessions.
- Redirection on all the operations and redirection caching.

Other advanced features are foreseen, like multi-source download based on metalinks and chunk caching.

4. HTTP and Third Party Copies

The WebDAV standard defines a COPY method that, applied over a specific resource, allows to copy it into the destination specified with the header "Destination"[14]. The RFC actually allows the destination to be on a remote server.
Note that this specification limits us to a push model, where the source orchestrates the COPY. However, there are needs which are not fully covered by the specification: proxy delegation and performance feedback, which needed some extensions that have been agreed with other storage solutions (dCache).

4.1. Proxy delegation

When a COPY is initially performed over a resource, if a delegation is needed the server will redirect the client to the same resource but appending a header "X-Delegate-To", which points to the endpoint that should be used for delegation. The GFAL2 library, and hence the FTS3 system support this set of extensions. If the client understands this header (i.e. Davix and Davix-based libraries, as gfal2), it will delegate to the aforementioned endpoint and then follow the redirection. The COPY will then proceed.

It is worth mentioning that if there were already a valid delegated proxy, the redirection step won't be needed.

5. Current status and conclusions

In this work we have shortly described the contributions that we have given towards being able to use the HTTP protocol in the context of High Energy Physics distributed data access. Although the Web world is based on HTTP, the set of the features supported by the more widespread components did not include some that are needed for the kind of high performance distributed computing that modern HEP experiments need.
The areas for improvement that we have contributed to are:

- Server side. DPM [3] [4] gives advanced, RFC-compliant HTTP and WebDAV functionalitites, on top of the Grid standards, and together with the other Grid Storage Element implementations [9] [10]
- Compatibility. We are contributing XrdHTTP, a native xrootd plugin that adds HTTP/WebDAV functionalitites to the xrootd [2] [1] framework, keeping its native protocol.
- Clients. We designed DAVIX, a versatile full-features client library on top of the widespread libneon library
- Applications: We contributed a plugin for ROOT I/O that uses DAVIX for data access and data management, thus extending to ROOT the coherent support of all the advanced features through HTTP
- Scalability:We contributed a system that is able to build high performance federations of WebDAV storage endpoints, working through Wide Area Network
- Data replication: FTS3 can use HTTP to replicate data between storage elements that support the COPY method
- We released all these contributions to the EPEL [13] distribution, in addition to the LCG Application Area

Acknowledgment

This work was partially funded by the EMI project under European Commission Grant Agreement INFSO-RI-261611.

References

[1] Scalla/xrootd WAN globalization tools: Where we are Fabrizio Furano and Andrew Hanushevsky 2010 J. Phys.: Conf. Ser. 219 072005 http://iopscience.iop.org/1742-6596/219/7/072005/
[2] The xrootd.org homepage http://www.xrootd.org
[3] DPM: Future Proof Storage Alejandro Alvarez, Alexandre Beche, Fabrizio Furano, Martin Hellmich, Oliver Keeble, Ricardo Rocha CHEP2012
[4] Web enabled data management with DPM & LFC Alejandro Alvarez Ayllon, Alexandre Beche, Fabrizio Furano, Martin Hellmich, Oliver Keeble and Ricardo Brito Da Rocha CHEP2012
[5] Furano F. Data Management in HEP: an approach. The European Physical Journal Plus Volume 126, Number 1 (2011), 12, DOI: 10.1140/epjp/i2011-11012-2
[6] DPM components https://svnweb.cern.ch/trac/lcgdm/wiki/Dpm/Dev/Components
[7] Cristian Traian Cirstea Grid Data Access: Proxy Caches and User Views Eindhoven University of Technology Stan Ackermans Institute / Software Technology ISBN 978-90-444-1067-9
[8] neon: an HTTP and WebDAV client library, with a C interface http://www.webdav.org/neon/
[9] The dCache.org home page http://www.dcache.org
[10] STORM: a storage manager service (SRM) for generic disk based storage system http://storm.forge.cnaf.infn.it/
[11] Grid File Access Library 2.0 official page https://svnweb.cern.ch/trac/lcgutil/wiki/gfal2
[12] File Transfer Service (FTS) 3 https://svnweb.cern.ch/trac/fts3
[13] EPEL - Fedoraproject http://fedoraproject.org/wiki/EPEL
[14] HTTP Extensions for Web Distributed Authoring and Versioning (WebDAV), section 8.8 http://tools.ietf.org/html/rfc2518#section-8.8