

WG on Storage Federations

Oct 2012 - GDB
Status

Fabrizio Furano
furano@cern.ch

- The mandate sounds like:
 - Make the concepts more clear to everybody
 - Make experiments and WLCG sites talk
- 3 months of life, 8-9 meetings (this is the 4th... time passes)
- We started collected the points of view of the experiments, with focus on:
 - What are they trying to accomplish (and who is “they”)
 - The rationale behind
 - Being critical and pragmatic at the same time
- The last episode (yesterday at the pre-GDB) was supposed to move to considering sites. We will reschedule the discussions and change the roadmap accordingly.
- Some bits also mention the “how”, these will be better discussed in the next meetings
- *Thanks again to the participants, as their contribution to the content and to the productive atmosphere so far has been of the highest quality.*

- In the past episodes we put together a snapshot of the recent “storage federations” initiatives
 - Focus on the motivations and on the differences with respect to the current production environments
- The activity and the plans, in particular of ATLAS and CMS are very strong
- From the words of some speakers, this seems to be a shift in the focus. From local to global, from production to user analysis.
 - Related to “lowering the bar” to start doing analysis
- The information in our pages is very up to date and synthetic
 - <https://twiki.cern.ch/twiki/bin/view/LCG/WLCGStorageFederationsWG>
 -

- Everybody agrees on adopting this as a starting point:
 - *A collection of disparate storage resources managed by cooperating but independent administrative domains transparently accessible via a common namespace.*
- In practice, see everything natively as one storage, which works easy and well, minimizing its complexity and the amount of glue.
 - Consequence: Thin clients are preferred
- My impression:
 - All this was also, somehow, in the original ideas of the GRID (evident e.g. in the LFC namespace `/grid/<experiment>/path/file`)
 - The primary difference among the 4 experiments is in the used components and their characteristics, not in the vision
 - The “wave of federations” seems about smoothly evolving in that direction, having learnt something

- Experiments have a main interest in these building blocks:
 - Direct access to data is the main item
 - e.g. Eases personal activity of the users doing analysis and writing papers
 - Coherent file naming with access to everything is the main idea
 - Translation: users (at the client side) do not like being exposed to bits that are private to the site, like their SFNs
 - Being able to use WAN direct access is the ultimate wish
 - ATLAS: *“helping users to make their code perform well through WAN is a key factor”*
 - Give more importance to the chaotic, Web-like user activity
 - Keep the official data processing (jobs, MC, reco, etc.) as it is, if possible enhance
 - Lower the bar for a physicist to run analyses, let him focus less on the data access and data management issues
- Power users seem to have a role in propagating their wishes and their solutions

- A message of the WG is:
 - Keep separated the concept of federation from its applications
 - Federation: *A collection of disparate storage resources managed by cooperating but independent administrative domains transparently accessible via a common namespace.*
 - Applications: What we do with it, e.g. direct data access, self healing, failover, workflows, etc...
 - The difference among the various interpretations of the tech aspects is
 - relatively in the features (what the system provides), as more or less people agree on what such a system should provide (prev. slide)
 - more in the way the data access is performed, i.e. in what the clients doing analysis do to get their job done
 - e.g. which/how many systems they have to contact, how many protocols are involved in a single transaction, how fast it is, etc.

- Fail over for jobs
 - Failover is a feature that is linked to the idea of “protocol that supports redirection”, like Xrootd or HTTP
 - It’s about the client choosing a new destination in the case of an issue accessing a file
 - In the case of dead servers, this has to do with “fault tolerance” rules of a client
 - In the case of files that are not found, this blends with the concept of workflow (go to site A, then B, then to the regional redirector, etc...) as seen from the client’s perspective
 - The destination of a failover can also be a federation of sites, likely able to satisfy the request
 - The workflows “go to the regional redirector” fit here

- Self healing
 - A storage site realizes that it misses a file, then it does automatically something to pull it from *somewhere*. Meanwhile, the requesting client is transparently paused.
 - This *somewhere* can naturally be a federation, because pulling files from it is supposed to be easy and solid
 - Can be the same federation the site belongs to
 - The way it's done is by instrumenting the storage cluster, using hooks of the server-side software framework
- Status
 - CMS leaves this instrumentation of the xrootd servers fully to the good will of the sysadmin
 - ATLAS FAX can accommodate an optional xrootd instrumentation, similar to CMS
 - ALICE has this instrumentation in their SE setup. Deactivated now for manpower and resource reasons (global redirector switched off -> instrumentation does nothing)
 - The various ALICE AFs use this method by default since years, to download files. They are big clusters, they are not federated yet.
 - LHCb has an external framework that tracks troubled files.

- T3 type site, users doing analysis don't want to pre-place any files
 - Often repeated access of same files.
- The “site proxy/cache” recognizes this and keeps copies of the files that are popular (definition of popular should be configurable).
 - So this proxy/cache would serve local users for all data accesses.
 - The cache could decide to:
 - serve a copy already cached
 - retrieve the copy from somewhere else and then serve it
 - redirect the access to somewhere else
- The cache could also participate to a federation, offering its content in a given moment as a source of data to clients sent there by the federation system
- There are examples of similarly inspired things (e.g. PROOF clusters and Xrootd proxies), a challenge is to make this possible also with HTTP/DAV clients

- The old-rooted habit of decorating the path/name of a replica with the name of the site and other tokens is historically difficult to handle
 - Needs a non trivial name translation to be mapped to another site
- in ATLAS there was also historical freedom to mangle the filenames when storing files in sites
- All the exps have their own experiment-oriented metadata catalogue
- LHCb and ATLAS also use the LFC as a replica catalogue, with a subtle and very important difference
 - ATLAS uses the whole SFN from it, as a string
 - LHCb uses only the hostname field and builds the SFN algorithmically at the client app side

- What's happening now is:
 - in ATLAS
 - the FAX federation instruments all the storage elements so that they do this LFN->SFN translation on the fly, contacting the LFC in a synchronous way
 - ATLAS has developed an xrootd plugin that does this
 - the FAX federation does not (want to) expose SFNs or site-private naming conventions to clients
 - Clients and users see exclusively logical file names
 - Clients do not need to translate names, this is done by the servers.
 - A benefit is to decentralize the task. Weak point is that all the servers will depend from the LFC (or similar name translation service)
 - Good translation service --> good service
 - ATLAS: *"when we get rid of the lfc the setup will be streamlined"*... the means of this statement should be better understood
 - through what? Is a worldwide monster file renaming on the way?
 - Changing the name/technology of the DB will not streamline the setup

- What's happening now is (cont.d):
 - CMS uses an algorithmic name translation. This is performed in the xrootd servers with a n2n plugin. No external translating services need to be contacted. Servers export the global file name like in the ATLAS case.
 - LHCb gets the relevant path tokens from their information service and from LFC, then build the path from these. In principle they are compatible with such global name spaces, as the name translation is done by the application that wraps the clients for the various protocols.
 - ALICE configures all the SEs so that they export the same namespace, the older LHCb-like translation was set to identity, being performed by the site's xrootd cluster internally. The filenames are opaque, jobs contact the Alien DB for translating.

- We did not do yet a meeting specific to security
- One requirement was already stated very clearly
 - ATLAS and CMS points of view at the WG are very similar:
 - *Make the data of a storage federation readable to anybody in the collaboration.*
- Both practical and technical reasons behind:
 - Grow fast and smoothly the new system, reducing a deployment effort that is felt as not necessary
 - More performance, to keep up with expectations
 - May have to do with the final users' perception of the framework

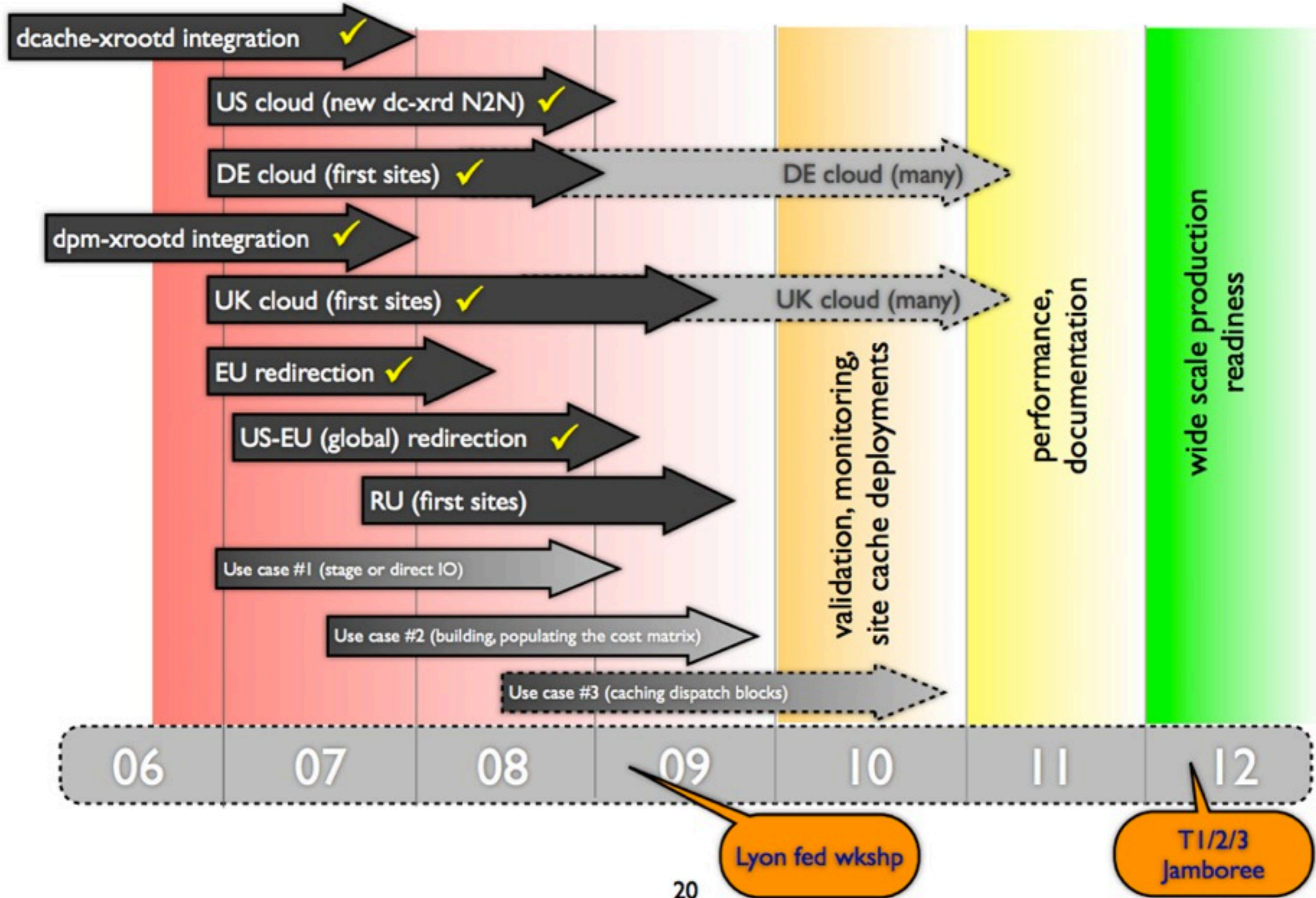
- The focus of the meetings so far has been more on the features and on the characteristics, trying to find answers to “why” and “what”
- At the same time:
 - Open-mindedness about the protocol to use
 - Doug (ATLAS + OSG) raises the attention on the benefits of giving users a standard set of tools
- Some technologies technically can do these federations
 - Xrootd, http/DAV are the technologies that will be available throughout the next years
 - Statement: *The exps and the grid mw should coordinate more to maximise the chances of HTTP being adopted*
- Web browsers trained us in willing “Any Data, Any Time, Anywhere”. Feels natural that the HEP power users try to propose it.

- Apparently, FAX and the CMS AAA federation are progressing very fast in the deployment:
 - AAA: “Any data, Any time, Anywhere”
- This was very evident at the Federated data stores workshop, as it is at our WG
 - <https://indico.in2p3.fr/conferenceDisplay.py?ovw=True&confId=6941>
 - The focus in Lyon was mostly technical, about progressing in the tech development and the deployments of FAX and CMS AAA
 - Monitoring, performance, functionalities
 - The WG is technical and tries to make the concept more known

- CMS status:
 - We have a multi-layered hierarchy.
 - US redirector containing all US T1/T2 sites.
 - EU redirector containing a smaller subset of European sites, including EOS. Sites from Italy, UK, Germany. Finland is working on joining.
 - Doesn't cover all CMS files, but probably has 90% of those relevant to analysis.
 - The target for 2012 is that the majority of sites participate.
 - We will reschedule Adam's talk about CMS UK from yesterday...



ATLAS timeline



- The WG is ongoing, the first milestone was understanding *why*, *what* and the status
- We have put together a synthetic bunch of information in the Indico pages
 - <https://twiki.cern.ch/twiki/bin/view/LCG/WLCGStorageFederationsWG>
 - <https://indico.cern.ch/categoryDisplay.py?categId=4318>
- We will steer towards sites, in order to
 - Better understand what they need
 - Monitor their participation to these new federations
 - Clarify what the WG could become for them, according to the items proposed by Rob (ATLAS)
- Our first goal is to complete a good snapshot, so we will do also some mono thematic meetings, on particularly delicate subjects
 - Security
 - Monitoring

- A tech monography? Xrootd? HTTP/DAV dynamic federations?
- This was the original proposal:
 - *A description of the set of consolidated features with respect to the federations tech*
 - *Where's the config? Central or in the endpoints?*
 - *Discovery, resilience to downtimes*
 - *Indicative steps for their tech to be deployed at a site*
 - *A collaborating one, AND one which does not know anything and will just look at it from time to time*
 - *Bundles available? How the config of a site is handled*
 - *Would be nice to know about:*
 - *Why is it so cool?*
 - *What is the vision about making it really powerful and ubiquitous?*
 - *Are there contacts with other sciences/VOs setting up federations?*
 - *Is there an idea on a future interactive access (tools?)*

Thank you

Questions?