WG on Storage Federations

Sept 2012

Usages and Goals

Summary

Fabrizio Furano

furano@cern.ch

- One meeting to discuss the usages of such federations
    - "fail over" for individual jobs
    - "repair mechanism" of placed data
        - who initiates the repair?
        - how is metadata handled ?
        - pro-active repair?
    - access control and trust mechanism between sites
    - use of catalogues
- Could these be matched among the experiments?
- What should the admin of a remote site do?
- Who will design and maintain each instrumentation?

Tuesday, September 11, 12

What follows is...

- Some sort of summary of our activity so far

- This is what we will report to the community

- Feel free to ask me to add your items now.

Tuesday, September 11, 12

GT

- 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)
- So far we 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

- 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.*

Tuesday, September 11, 12

What is a federation

- 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

Tuesday, September 11, 12

What do exps want

- Experiments now want 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
  - 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

- Power users seem to have a role in propagating their wishes and their solutions

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

30 Aug 2012 - F.Furano - WG on Storage Federations

6

Tuesday, September 11, 12

- **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. user 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 the system should provide
    - 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.

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

30 Aug 2012 - F.Furano - WG on Storage Federations

7

Tuesday, September 11, 12

GT

- 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 choosing a new destination for a client that has 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

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

30 Aug 2012 - F.Furano - WG on Storage Federations

8

Tuesday, September 11, 12

Usages: self-healing

- Self healing
  - A storage site realizes that it misses a file, then it does automatically something to pull it from somewhere
  - 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 software

  - CMS leaves this instrumentation of the xrootd servers fully to the good will of the sysadmin
  - ATLAS FAX proposes an optional xrootd instrumentation....... urgh WHAT?
  - ALICE had it completely bundled for sites, deactivated now for manpower and resource reasons
  - The various ALICE AFs use this method by default since years
  - LHCb has an external framework that tracks troubled files.

Tuesday, September 11, 12

- 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 LFN algorithmically at the client app side

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

30 Aug 2012 - F.Furano - WG on Storage Federations

10

Tuesday, September 11, 12

CERN **IT** Department

- **What's happening now is:**
  - in ATLAS
    - the FAX federation instruments all the storage elements so that they do this LFN->SFN translation in the background, 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 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
      - Good LFC 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? a worldwide monster file renaming?)
      - Changing the name/technology of the DB will not streamline the setup

Tuesday, September 11, 12

CERN
IT
Department

- **What's happening now is (cont.d):**
  - CMS already used an algorithmic name translation at the client side, no external translating services need to be contacted

  - LHCb gets the relevant tokens from their information service and from LFC. In principle they are compatible with such global name spaces, as the name translation is done by the application that wraps the client.

  - 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.

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

30 Aug 2012 - F.Furano - WG on Storage Federations

12

Tuesday, September 11, 12

**GT**

CERN IT Department

- We did not do yet a meeting specific to security

- One requirement was already stated very clearly
  - ATLAS and CMS points of view are very similar:
  - *Make the data (at least the metadata) of a storage federation readable to anybody.*

- 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

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

30 Aug 2012 - F.Furano - WG on Storage Federations

13

Tuesday, September 11, 12

- The focus of the meetings has been more on the features and on the characteristics so far

- 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

- Newer technologies technically can do these federations
  - Xrootd, http/DAV is what we will see in the next years

- Web browsers trained us in willing "Any Data, Any Time, Anywhere". Feels natural that the HEP power users try to propose it.

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

30 Aug 2012 - F.Furano - WG on Storage Federations   14

Tuesday, September 11, 12

- ## Apparently, FAX and the CMS federation are progressing very fast in the deployment:

- ## 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.

Tuesday, September 11, 12

Conclusion (so far)
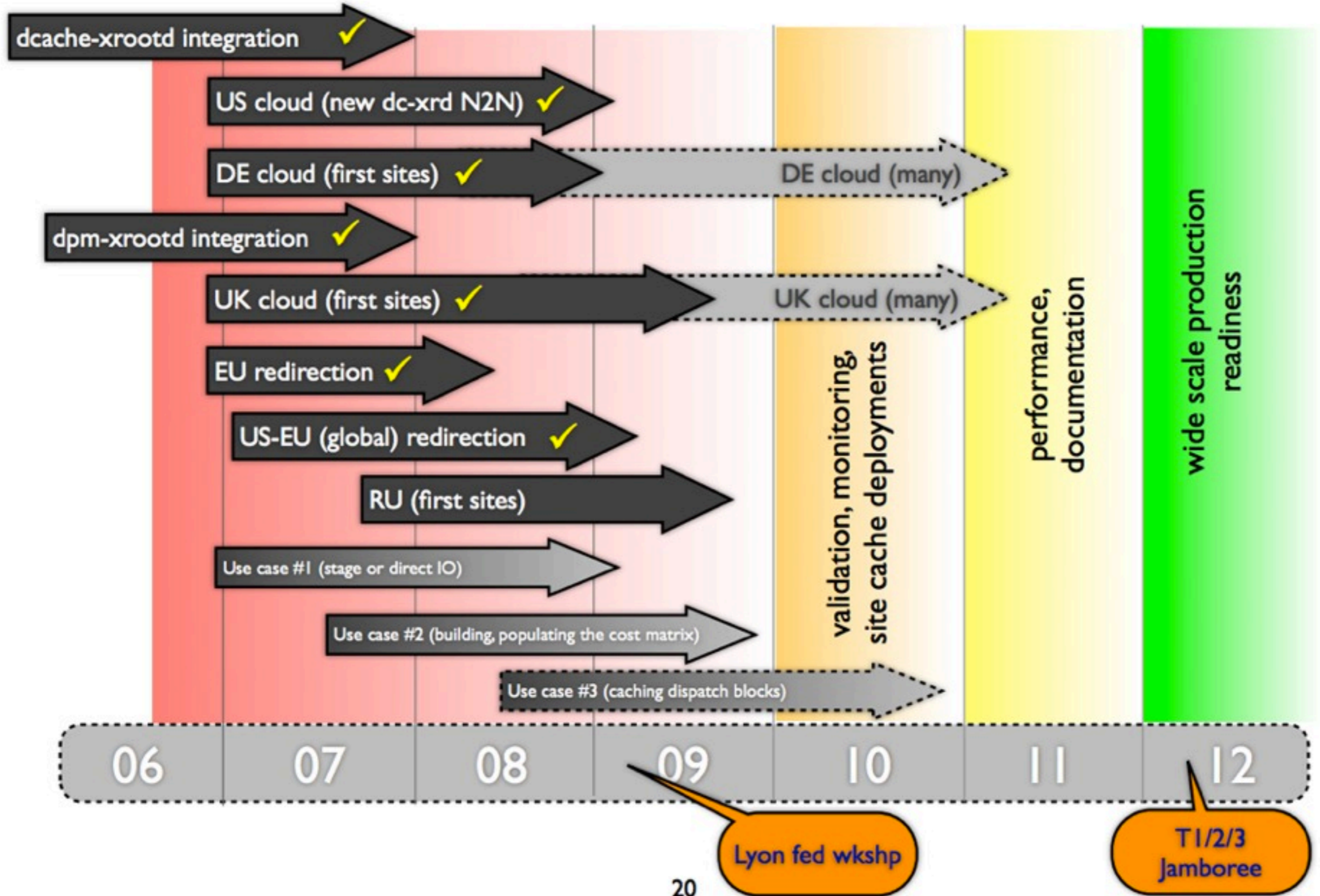
- The WG is ongoing, just finished with the first milestone (understanding *why*, *what* and the status)

- We have put together a synthetic bunch of information in the Indico pages

- These should give a more precise idea

- We will do also some monothematic meetings, on particularily delicate subjects
  - Security
  - Monitoring

Tuesday, September 11, 12

- My proposal is to finish the last point of today:

- *"What shall the admin of a remote [WLCG] site do?"*

- In other words:
  - What do you give and say to CNAF or NIKHEF or BARI or <put your site> to have a well working endpoint
  - Instructions for sysadmins, troubleshooting etc.
  - Firewall config, DNS names, machines with multiple intfcs...

Tuesday, September 11, 12

Thank you

Questions?