DPM in FAX (ATLAS Federation)

Wahid Bhimji
University of Edinburgh

As well as others in the UK, IT and Elsewhere

Outline

- Introductory:
  - What is FAX and the goals (as stated by the project)
  - Some personal perspectives

- DPM FAX deployment status

- Testing / Monitoring / Use-Cases

- Concerns and Benefits

What is FAX?

Description (from the FAX Twiki):

The Federated ATLAS Xrootd (FAX) system is a storage federation aims at bringing Tier1, Tier2 and Tier3 storage together as if it is a giant single storage system, so that users do not have to think of there is the data and how to access the data. A client software like **ROOT** or **xrdcp** will interact with FAX behind the sight and will reach the data whereever it is in the federation.

Goals (from Rob Gardner's talk at Lyon mtg. 2012):

- Common ATLAS namespace across all storage sites, accessible from anywhere;

- Easy to use, homogeneous access to data

- Use as failover for existing systems

- Gain access to more CPUs using WAN direct read access

- Use as caching mechanism at sites to reduce local data management tasks

Other details / oddities of FAX (some of this is my perspective)

- Started in US with pure-xrootd and xrootd-dcache
  - Now worldwide inc. UK; IT ; DE and CERN (EOS)

- Uses topology of "regional" redirectors (see next slide)

- ATLAS federation uses a "Name2Name" LFC lookup

- Now moving from R&D to production
  - But not (quite) there yet IMHO

- There is interest in http(s) federation instead / as well
 - But this is nowhere near as far along.

Regional redirectors
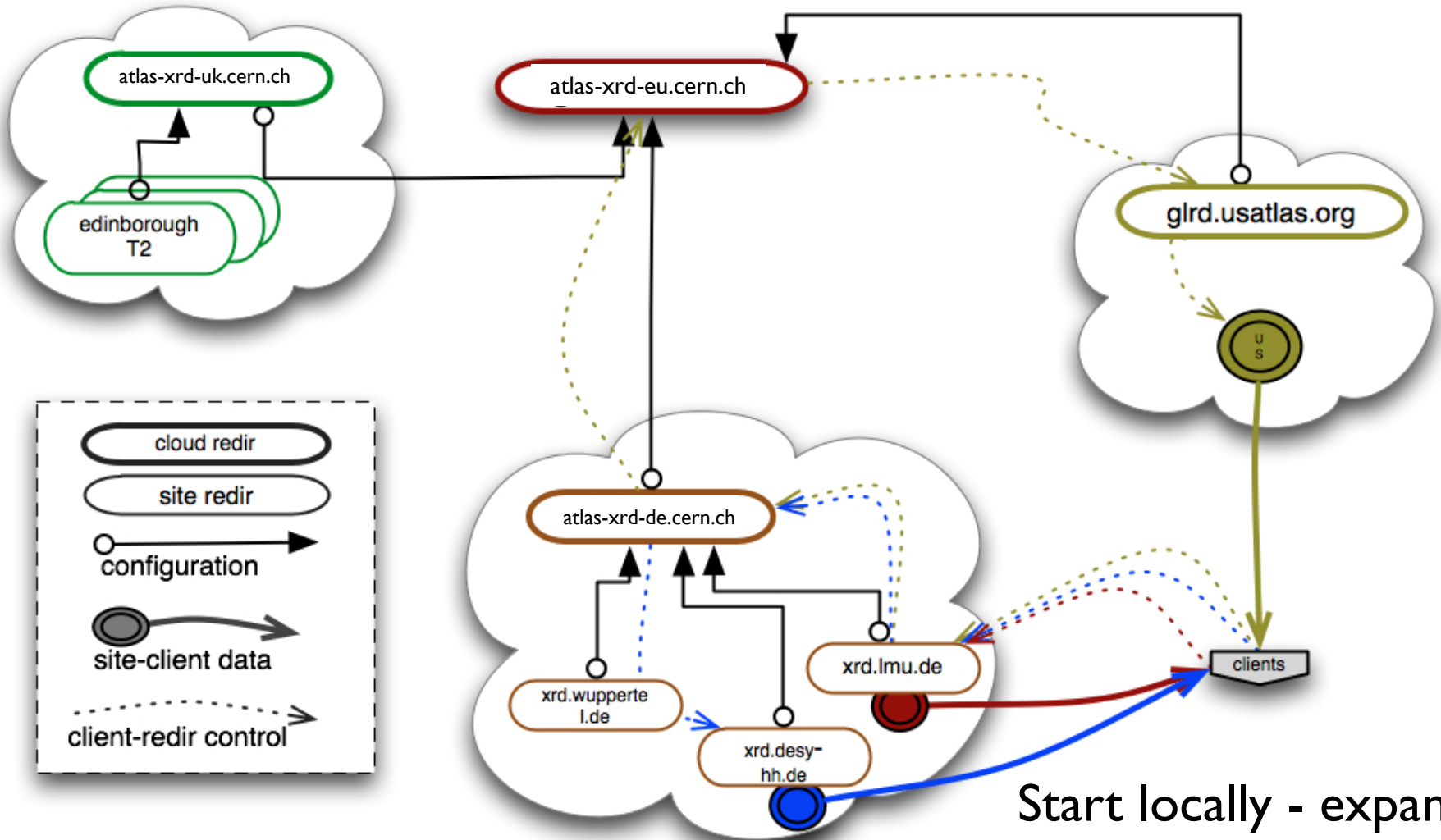
Start locally - expand search as needed

DPM Fax Deployment Status

- Last workshop DPM developers presented a new dpm-xrootd: <u>details</u>

- Initially deployed manually in Scotgrid. A few teething configuration issues, all tweaks now in YAIM (since 1.8.4) and documented at
  - https://svnweb.cern.ch/trac/lcgdm/wiki/Dpm/Xroot/Setup
  - Thanks to David Smith

- **Regional redirectors setup for UK; IT; DE (and EU)**

- Sites working now (see next page): UK: Edinburgh (ECDF); Glasgow; Oxford; DE: Prague; IT: Roma (at least)

- EMI push means many other sites will be able to install soon

  Some Issues.See later

- IT plan to expand to many sites (Frascati,.Napoli++); ASGC (T1) plan to deploy
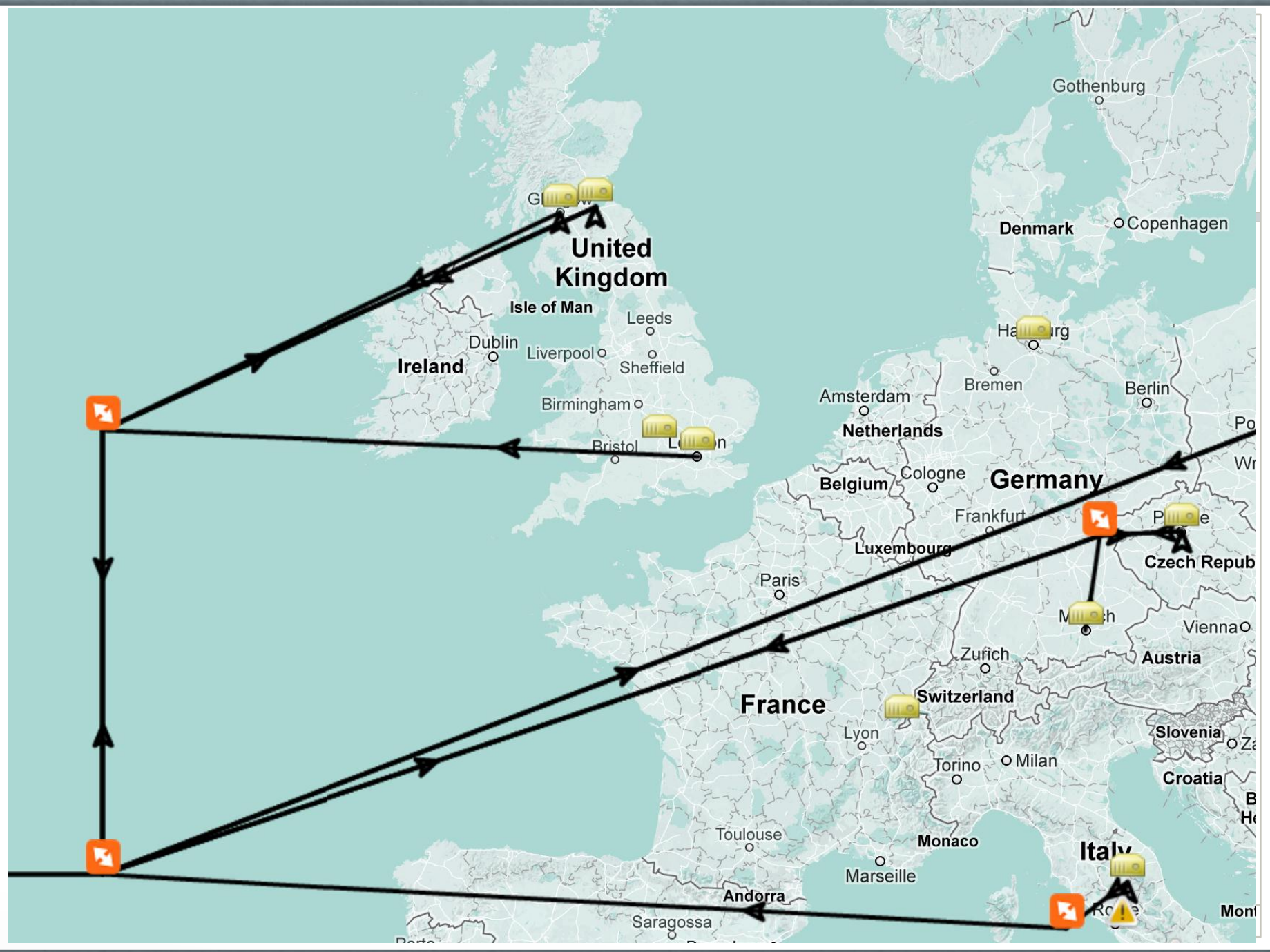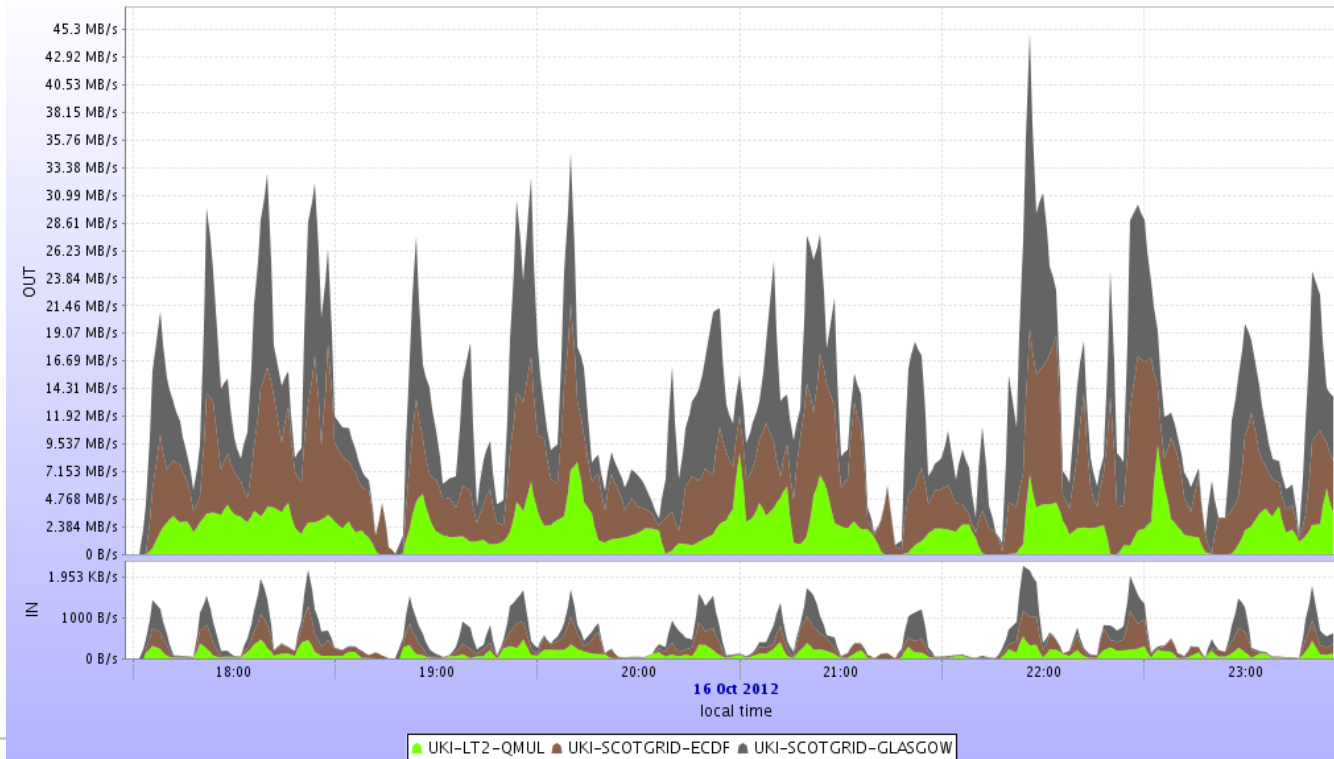
Traffic monitoring

xrootd.monitor all rbuff 32k auth flush 30s window 5s dest files info user io redir atl-prod05.slac.stanford.edu:9930

xrd.report atl-prod05.slac.stanford.edu:9931 every 60s all -buff -poll sync

Aggregated Xrootd traffic

Initial bug OK since updated
to 3.2.5 (in EMI external repo)

See:
http://atl-prod07.slac.stanford.edu:8080/display

Functional Testing

Regular WAN testing in hammercloud and MAP:

http://ivukotic.web.cern.ch/ivukotic/WAN/index.asp (See next page)

http://ivukotic.web.cern.ch/ivukotic/FAX/index.asp (The map)

Traffic monitoring:

http://atl-prod07.slac.stanford.edu:8080/display (previous page)

• Dashboard monitoring (being developed)

http://dashb-atlas-xrootd-transfers.cern.ch/ui/#m.content=(efficiency,errors,successes)&tab=matrix

• Basic service test http://uct3-xrdp.uchicago.edu:8080/rsv/ (next page)

• There will be a SAM test

100% default cache — CPU eff. vs ping [ms]

Legend:
- CERN->AGLT2
- CERN->BNL
- CERN->OU_OCHEP_SWT2
- MWT2->AGLT2
- MWT2->BNL
- MWT2->OU_OCHEP_SWT2
- OU_OCHEP_SWT2->AGLT2
- OU_OCHEP_SWT2->BNL
- OU_OCHEP_SWT2->OU_OCHEP_SWT2
- SLAC->AGLT2
- SLAC->BNL
- SLAC->OU_OCHEP_SWT2
- SWT2_CPB->AGLT2
- SWT2_CPB->BNL
- SWT2_CPB->OU_OCHEP_SWT2

by Ilija Vukotic

Host: srm.glite.ecdf.ed.ac.uk (srm.glite.ecdf.ed.ac.uk)

Metric	Last Executed	Enabled?	Next Run Time	Status
org.usatlas.xrootd.grid-xrdcp-direct	2012-11-29 13:35:00 CST	YES	2012-11-29 13:50:00 CST	OK
org.usatlas.xrootd.grid-xrdcp-fax	2012-11-29 13:35:00 CST	YES	2012-11-29 13:50:00 CST	OK
org.usatlas.xrootd.ping	2012-11-29 13:35:02 CST	YES	2012-11-29 13:50:00 CST	OK

Use Cases – revisiting goals

- Common ATLAS namespace across all storage sites, accessible from anywhere; Easy to use, homogeneous access to data
  - Done – implicit in the setup
  - Keen users being encouraged to try: tutorials etc.

- Use as failover for existing systems
  - Production jobs can now retry from the federation if all local tries fail… works – but not tried on DPM in anger.

- Gain access to more CPUs using WAN direct read access
  - WAN access works – no reason no to use in principle.
  - Timing info from WAN tests ready for brokering – not yet used (AFAIK)

- Use as caching mechanism at sites to reduce local data management tasks
  - Nothing yet has been done on this with DPM (AFAIK).

Stress Testing DPM Xrootd

ANALY_GLASGOW_XROOTD queue
- Stress-tested "local" xrootd access
  - For direct access we saw some server load (same as we do for rfio).
  - David did offer to help – we didn't follow up much
- Trying panda failover tricks
  - Not done yet.
  - Requiring new dq2 tools in cvmfs which requires new python.

- ASGC have done extensive hammerclouds on (non-FAX) dpm-xrootd :
  - Promising results . Using in production now (?)

FAX-DPM Issues encountered

PAST :

- xrootd packaging: would ideally be current in epel – but there have been some problems achieving that
  - Now in EMI externals which is Okish

- Without rbuff 32k in monitoring crashed with initial version
  - Fixed in later xrootd versions

- Getting stuck in LFC lookup:
  - LFC host is an alias and single threaded N2N sometimes trying the "wrong" host
  - Fixed by setenv LFC_CONRETRY=0

PRESENT:

- SL6: xrootd 3.2.5 segfaults on startup so until 3.2.6 available can't use those sites

NONE OF THESE ARE DPM PROBLEMS – IN EACH CASE DAVID FOUND THE FIX. HOWEVER MAY INDICATE A (PRATICAL) ISSUE WITH USING XROOTD ON DIFFERENT STORAGE SYSTEMS.

My (general) concerns

- Even now if users started using this, this could result in a lot of unexpected traffic of files served from DPM sites :
  - The service is not in production – no SAM test; no clear expectations of service etc. Communication with sites currently direct to site admin (not via cloud or ggus). Some network paths are slow.
  - Ideally should be able to configure server (or redirector?) to limit connections / bandwidth. (And to monitor monitoring).

- Multiple VO support: currently separate server instances - sensible?

- (xrootd) Software Documentation.

- http (s) might be preferable standard

- But many site failures are storage related so *if* it solves those then its worth it

Conclusion

Significant development in DPM/FAX integration since last meeting:

- Basically from nothing to something…

- At least 5 T2 sites federated and seemingly working…

- But need to stress test under real use

- And to have some concerns assuaged