Status of the FAX federation

Rob Gardner

Computation and Enrico Fermi Institutes

University of Chicago

ATLAS Tier 1 / Tier 2 / Tier 3 Jamboree

December 10-11, 2012

Computation Institute

THE UNIVERSITY OF CHICAGO

efi.uchicago.edu
ci.uchicago.edu

Recall our goals

- Common ATLAS namespace across all storage sites, accessible from anywhere
- Easy to use, homogeneous access to data
- Identified initial use cases
  - Failover from stage-in problems with local SE
    - Now implemented, in production on several sites
  - Gain access to more CPUs using WAN direct read access
    - Allow brokering to Tier 2s with partial datasets
    - Opportunistic resources without local ATLAS storage
  - Use as caching mechanism at sites to reduce local data management tasks
    - Eliminate cataloging, consistency checking, deletion services
- WAN data access group formed in ATLAS to determine use cases & requirements on infrastructure
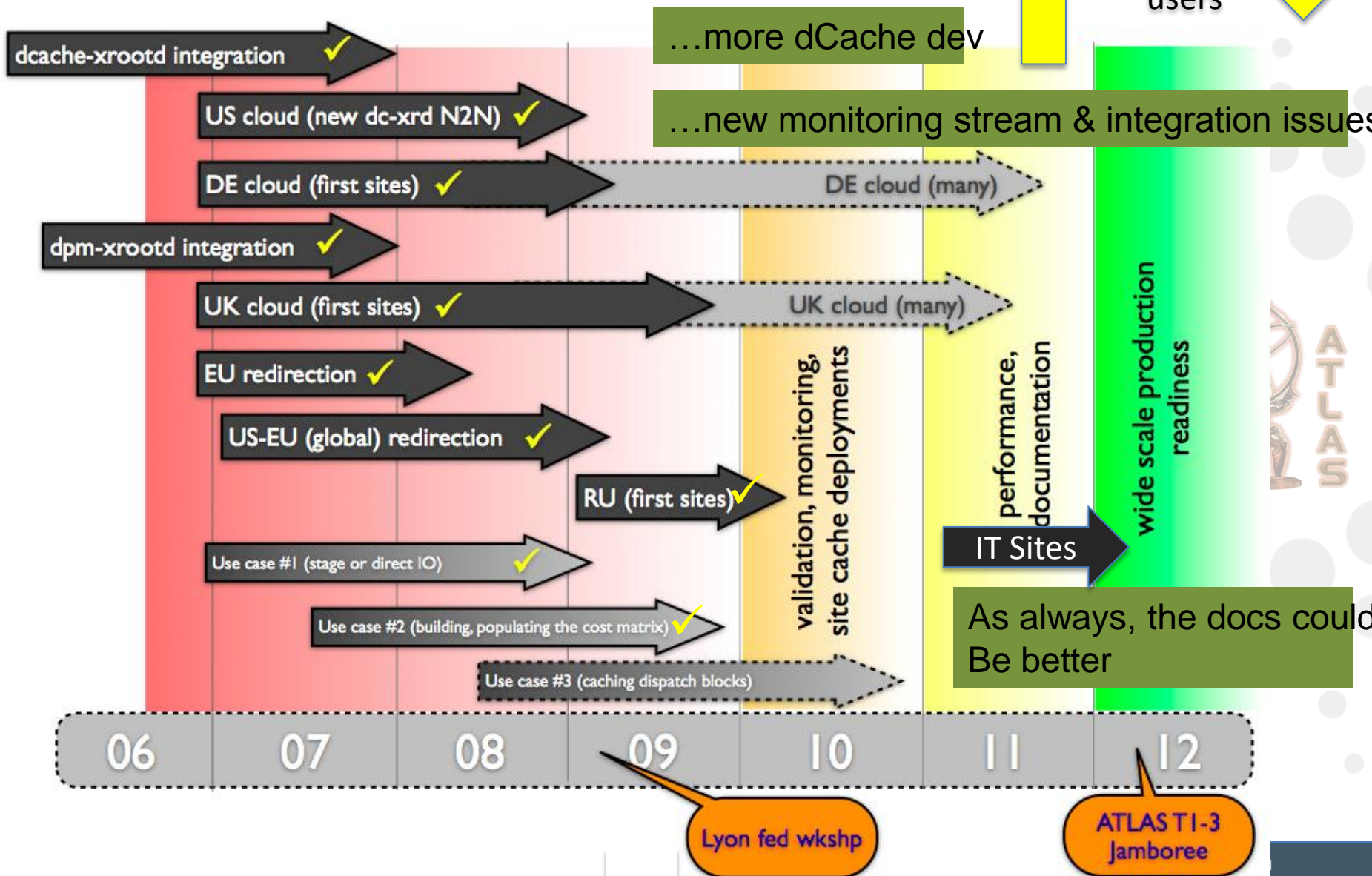
Progress

- Ongoing issues development & operations
- FAX in the pilot.. First use case working now
  - Later → Panda brokering
- Programmatic WAN testing
- FAX as supporting portable, lightweight applications on opportunistic resources (Tier 3, cloud, campus)

Roadmap

EU deployment status

THE UNIVERSITY OF
CHICAGO

efi.uchicago.edu
ci.uchicago.edu

RU Cloud Status

Danila Oleynik

RU cloud status.

- **IHEP**
  - Site in production status.
  - Current xroot version: 3.2.6-1
  - No major problems observed
    - little concern because of the high memory consumption.

```
PID     USER     PR  NI  VIRT   RES   SHR   S  %CPU  %MEM  TIME+      COMMAND
28123   xrootd   15  0   6213m  1.4g  2392  S  3.9   70.8  162:09.14  xrootd
```
  - o Automatic notification about test failures – are very useful. (Thanks to Ilija)

- **JINR**
  - Site in production status.
  - Current xroot version: 3.2.6-1
  - No major problems observed, but memory…

```
PID    USER     PR  NI  VIRT   RES   SHR   S  %CPU  %MEM  TIME+     COMMAND
5331   xrootd   20  0   3604m  1.5g  3204  S  4.0   38.1  2634:20   xrootd
```

THE UNIVERSITY OF CHICAGO | efi.uchicago.edu ci.uchicago.edu

Wahid Bhimji

UK Cloud FAX Status

- Same UK sites in FAX as at s/w week: (Wiki page of status)
  - ECDF; Glasgow; Oxford (T2Ds) (DPM) - Working
  - QMUL (T2D) (Lustre) – Working
  - RAL (T1) (Castor) – Works for CMS; "local" atlas path works
    - "Global" path doesn't –issue contacting LFC?

- Nearly all other T2s in UK are DPM
  - Almost all have upgraded to EMI: DPM 1.8.4 (or later)
  - Allows easy Yaim config of FAX (confirmed this works)
  - Liverpool; Manchester; Lancaster willing to join
    but Xrootd 3.2.5 is broken in SL6
    - David Smith confirmed this is fixed in 3.2.6 but also found a bug for SL5 – confirmed fixed now in devel.
    - Once 3.2.7 available from xrootd will be rebuilt for EPEL/EMI – probably released before Christmas… so these sites should be able to join next year

Wahid Bhimji

## UK Testing and Other Issues (since last meeting)

- All sites occasionally getting stuck in LFC lookup:
  - LFC host address changes, single threaded N2N sometimes (re)trying the "wrong" (cached) address.
  - Fixed by setenv LFC_CONRETRY=0 (now in DPM YAIM config.)

- ANALY_GLASGOW_XROOTD queue
  - "local" access hammerclouds fine (as good as rfio at least)
  - For FAX test waiting to use new dq2-tools in cvmfs
    - Currently a python 2.6 issue

- Real user (!) at QMUL – all his files are local though

- Ready for "dress-rehersal" mid-Jan with up-to 7 T2s and maybe T1

- Also planning WebDav deployment (tracked on same wiki)

THE UNIVERSITY OF CHICAGO

efi.uchicago.edu
ci.uchicago.edu

DE cloud overview

- in DE:
  - 8 dCache sites (1 T1, 7 T2, 1 T3)
  - several T3 w/ Lustre/GPFS/Sonas
- outside DE:
  - DPM sites Prague, Cracow, Innsbruck
  - dCache in CSCS/Manno

12/07/12   Guenter Duckeck, LMU   2

efi.uchicago.edu
ci.uchicago.edu
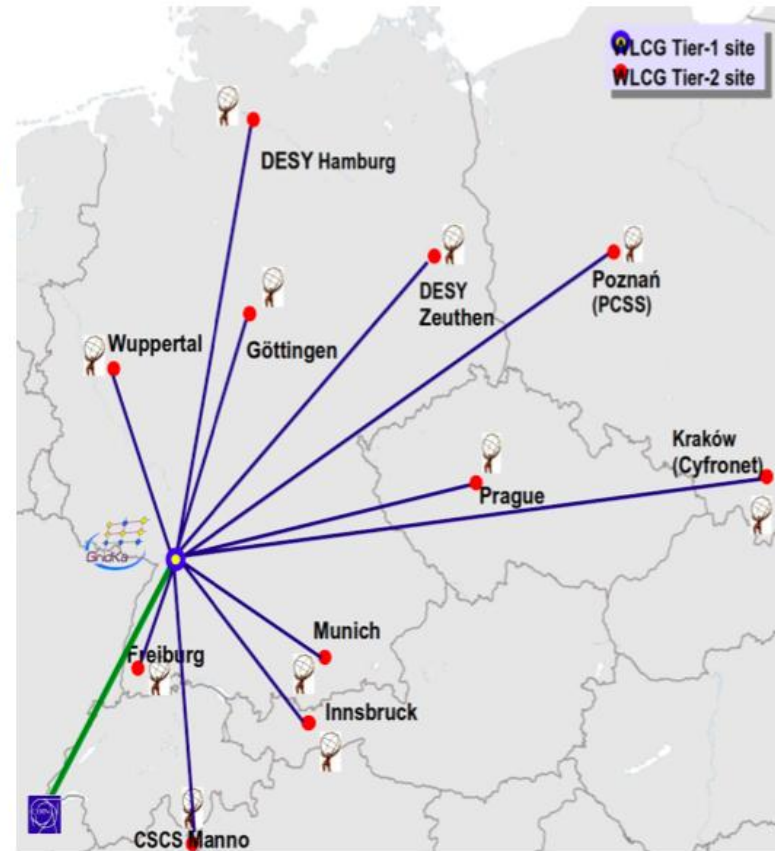
DE FAX status

Sites:

- **LRZ-LMU:**
  - FAX node name (proxy): xrootd-lrz-lmu.grid.lrz.de:1094, tests ok
  - //atlas/dq2/user/HironoriIto/user.HironoriIto.xrootd.lrz-muenchen/user.HironoriIto.xrootd.lrz-muenchen-1M
- **MPPMU:**
  - FAX node name (proxy): xrootd-mppmu.grid.lrz.de:2094 , no tests
  - //atlas/dq2/user/HironoriIto/user.HironoriIto.xrootd.mpg/user.HironoriIto.xrootd.mpg-1M
- **Wuppertal**
  - FAX node name (proxy): up.pleiades.uni-wuppertal.de, local ok, redir fails
  - //atlas/dq2/user/HironoriIto/user.HironoriIto.xrootd.wuppertal/user.HironoriIto.xrootd.wuppertal-1M
  - needs reinstall of uptodate xrootd proxy … in progress
- **Prague**
  - FAX node name (dpm) golias100.farm.particle.cz, tests ok
  - //atlas/dq2/user/HironoriIto/user.HironoriIto.xrootd.prague/user.HironoriIto.xrootd.prague-1M
- **Desy-GridLab:** tests of dCache plugin, → Ilija

DE FAX prospects

- Documented FAX-Proxy setup for dCache in DE cloud Wiki:
    - https://twiki.cern.ch/twiki/bin/view/Atlas/FAXDECloud
- We are waiting for more sites to join
    - we had expected GridKa to join
        - ongoing operational issues with tapes and storage have priority
    - currently very little manpower&resources available at sites.
    - some sites have strict policies how to operate services
        - require dedicated node for xrootd proxy installation, well connected to dcache nodes
            - not straighforward to fulfill
    - Plan B is deployment of dCache plugin for xrootd door (Ilija), needs no extra node
        - requires upgrade to recent dCache version
            - extended downtime & service interruption
            - rather wait until after spring confs in March 2013
- Dress rehearsal January:
    - Currently active sites should be ok, maybe GridKa as well.

IT Cloud Status

- All the DPM (v1.8.4) sites in IT have FAX enabled since the last week of November 2012
  - INFN-FRASCATI
  - INFN-NAPOLI-ATLAS
  - INFN-ROMA1

- All FAX-enabled sites are pointing to the IT redirector at CERN
  - atlas-xrd-it.cern.ch

- Storm sites not ready yet
  - Test facility (XROOTD test machine in front of Storm) requested to CNAF
  - Still to be implemented and tested
  - Major Storm sites in Italy are INFN-T1 and INFN-MILANO-ATLASC

Allessandro de Salvo

THE UNIVERSITY OF CHICAGO

efi.uchicago.edu
ci.uchicago.edu

IT Cloud Status

- ## Installation issues
  - Not so much, the most complex task was to identify the list of ports to be opened on the routers!
  - Update to the new N2N plugin transparent, no evident problem seen in DPM sites

- ## Operational problems
  - Some test dataset missing in Frascati and Napoli, needed for the transfer tests, should we ask manually for a dataset transfer in such cases?

- ## Monitoring
  - All the DPM sites correctly showing in the monitoring
  - …but not trasferring so much data!
  - We would like to have a real test to see if our infrastructure is really working properly

EOS

FAX: Federating EOS

DSS

CERN IT Department

- Issue: (non-trivial) mapping of "global" name to EOS-local, need lookup in LFC - see EOS Federation plans A.Peters (IN2P3 workshop)
- N2N translation done via new service – see LFC plug-in for EOS, E.Sindrilaru, A.Peters
  - Currently will look into a single location only (`/eos/atlas/atlasdatadisk`), prototype/config issue
- Deployed on 2 VMs (atlas-xrd-eos-n2n.cern.ch) – works.

EOS N2N already looks into several locations - match=atlasdatadisk,atlasgroupdisk,atlasscratchdisk

FAX phone conf – CERN redirector status - 1

CERN IT Department CH-1211 Genève 23 Switzerland

THE UNIVERSITY OF CHICAGO

efi.uchicago.edu
ci.uchicago.edu

Other FAX infrastructure at CERN

CERN IT Department

DSS

- Redirectors for EU, UK, DE, FR, RU (and ASIA)
- Common setup:
  - 2x VMs (on different hypervisors), same OS+xrootd versions
  - behind common DNS alias,

```
all.manager meta any atlas-xrd-eu.cern.ch.+ 1098
xrootd.redirect atlas-xrd-eu.cern.ch:1094 ? /atlas
```

- Minor differences:
  - Site access control (to prevent mistakes, != security)
  - User access control (debug)
  - Topology – who subscribes where
- Minimal effort to run
  - Can keep or can give away to "real" regional redirectors
  - No real advantage to have them at CERN, some penalty for faraway sites

CERN IT
Department
CH-1211
Genève 23
Switzerland

4

Development & integration .. operations

On-going work & issues

- SSB and WLCG transfer dashboard with cost matrix decision algorithm
- Xrootd instabilities seen in the UK cloud – perhaps related to N2N blocking at LFC
- FAX extensions to ATLAS information system AGIS
- Need new monitoring f-stream at all sites
- Stand-alone cmsd for dcache sites
- xrootd.org repository & EPEL policy (site guidance, esp. DPM)
- Several dCache specific issues, and many releases under test (1.9.12-22+, 2.2.4,…); f-stream, proper stat response and checksum support from dcache-xrootd doors

- Moved US sites to r.o. LFC
- SLC6 issues – X509 and VOMS attribute checking
- UDP collector service with f-stream format under test
- Functional testing probes & publishing into ActiveMQ and dashboards
- Monitoring will have to be validated at all stages
- FAX-enabled pilot site mover in production at two Tier 2s
- Documentation for users & site admins

Xrootd Issues

Need a xrootd release with fixed f-stream (3.3.0-rc2)

Move all xrootd sites to new version

Test new xrootd monitoring settings

Have all xrootd servers auto-restart and auto report problems (request to Xrootd team)

A standalone version of cmsd to be used in tandem with dCache xrootd doors (request to Xrootd team)

Issue of xrootd repository: Xroot.org/CERN or EPEL (WLCG operations task force)

Ilija Vukotic

Version 2.2.4 in production at AGLT2

Version 1.9.12-23 in production at MWT2

xrootd4j plugin working

N2N plugin working

Monitor plugin working

Have other 1.9.12-X dCache sites need to upgrade to get monitoring

Have dCache properly respond to xrootd stat call (request to dCache team)

Implement kxr_query checksum request (request to dCache)

THE UNIVERSITY OF **CHICAGO**

efi.uchicago.edu
ci.uchicago.edu

Ilija Vukotic

Italian DPM sites sensitive to version of xrdcp used to get file (on-going investigation)

N2N for RAL (Castor) probably nothing fundamental but needs a bit of sustained effort

  Shaun deWitt

  Will setup a call to discuss (Wei, Ilija, Shaun)

FAX infrastructure

German cloud (more from Guenter's slides)	Wuppertal not working
	DESY-HH works on a test node with dCache 2.4. Production on 1.9.12-12

Deployment in France? Where next?

EOS not redirecting properly

　　　　From Jan: *Issue identified (yesterday), new version+ configuration not yet deployed but expect coming week*

Ilija Vukotic

The problem on RHEL6 is fixed in 3.2.6 but the same fix also broke GSI security on RHEL5 (due to openssl). So 3.2.6 can't be deployed on RHEL5.

X509 on SLC6 is broken

VOMS attribute verification

Grid jobs running at some sites (apparently CERN and Glasgow at least) can't access FAX .   Possible problem with limited proxy.

THE UNIVERSITY OF **CHICAGO**

efi.uchicago.edu
ci.uchicago.edu

Monitoring

Test and certify f-stream

Finalize development of dCache Monitor plugin

Have the dashboard results "certified"

Things dashboard needs (proper host/user site name, VO, application)

THE UNIVERSITY OF CHICAGO

Integration
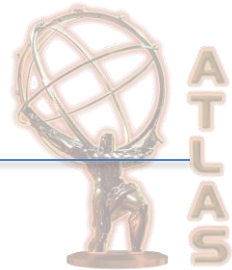
Make RSV publish to ActiveMQ

Move FAX topology viewer to AGIS/AMQ

Add to WAN tests AGIS/AMQ

THE UNIVERSITY OF CHICAGO

efi.uchicago.edu
ci.uchicago.edu

Documentation

Improve on all parts of TWiki

FAX site certification matrix update

Make our own PAT tutorial

Ilija Vukotic

- Growing but could be faster.
  - US ATLAS Hadronic Final State Forum 2012 users

Aggregated Xrootd traffic

efi.uchicago.edu
ci.uchicago.edu

There are many more components as discussed at the Lyon storage federations workshop in September

FAX dashboard – sites transfer matrix

efi.uchicago.edu
ci.uchicago.edu

AGIS developments

- Alessandro, Artem, Danila and Alexey added FAX support to AGIS
  - Database side
  - Web UI
  - Export to JSON
  - Tutorial (on the way)
- Current FAX topology data inserted (Ilija)
- Need to decide on who will take care data are up-to-date.
  - Since it should not change often we could charge one person with doing it
  - But it is rather simple to maybe sites could do it too
- Keep in mind: currently no e-mail field associated to FAX SE or redirector in AGIS
  - Currently FAX topology monitor has handpicked people to send mail to when FAX SE or redirector is down
  - Move to AGIS will mean that people responsible for site will be getting mails

THE UNIVERSITY OF CHICAGO

efi.uchicago.edu
ci.uchicago.edu

WLCG

- We have an on-going set of meetings with two WLCG working groups
  - WLCG Federated Data Storage (F. Furano Chair)
    - Explores issues generally and assesses approach taken by each of the LHC experiments
  - WLCG Xrootd task force (D. Giordano)
    - New group forming to coordinate deployment and operations
    - Will seek to engage Grid infrastructure providing groups (EGI/EMI, OSG) for support
- Both of these will help bring FAX into normal production operations

In the pilot, in Panda

Pilot capabilities – from Paul @ SW week

FAX in the Pilot

- Available since pilot version 54a (August 16)

- Update for user analysis jobs in v 54b (September 26)

- Tested "manually" (i.e. with faked stage-in failures) using dev pilots and test jobs on US test sites; UTA_PAUL_TEST, ANALY_MWT2_TEST

- Now used with user jobs on ANALY_MWT2

- Currently running tests at ANALY_SWT2_CPB and ANALY_AGLT2

- More tests on more sites under way

- Schedconfig activation: allowfax = 'true', faxredirector = '..'

Paul Nillson

FAX Activation Algorithm

- When the pilot encounters a stage-in error, it will retry the transfer (after a 5 min delay).

- If the second stage-in attempt fails as well, switch to FAX

- A special FAX site mover has been developed (can also be used as primary site mover, tested – works fine)

- If FAX fails, report original failure reason, the entire job will be retried

- Requires CVMFS site with an updated DQ2 client (2.3.0)
  - Performs a local root setup, uses xrd-cp from cvmfs, needs new option –p for dq2-list-files

THE UNIVERSITY OF CHICAGO

efi.uchicago.edu
ci.uchicago.edu

Paul Nillson

"Diskless" Use-cases

- FAX mechanism offers some interesting use-cases

- A stage-in problem will trigger the FAX algorithm as mentioned in the previous slide, bypassing the local SE for stage-in

- Planned SE shutdown; FAX as primary copy tool
  - Set schedconfig.copytoolin = "fax" (switch off any direct access)
  - But what about stage-out?
  - Stage-out to alternative SE (from T-2 to T-1) in development
  - Similar mechanism; after repeated stage-out attempts, the pilot switches to attempting stage-out to an alternative SE (belonging to the SE of the corresponding T-1 [within the same cloud as the job] using lcg-cp)
  - Local SE bypassed

Paul Nilsson

- Sites with FAX enabled: ANALY_MWT2, ANALY_SWT2_CPB
- Currently testing FAX at: ANALY_OX, ANALY_GLASGOW_XROOTD, ANALY_CERN_XROOTD, ANALY_LANCS, UKI-LT2-QMUL, ANALY_ECDF and ANALY_LRZ using updated pilot wrapper
- Pilot factory wrappers need to setup python 2.6 + latest DQ2 client (2.3.0)
  - Done for pyfactory wrapper by Peter Love
- Introduced FAX usage info to job metrics
  - If at least one successful FAX transfer, pilot adds "filesWithFAX=N_1 filesWithoutFAX=N_2" to job metrics

Current Pilot problems

Paul Nilsson

- Problems with lcg-cp seen on some sites (ANALY_GLASGOW, ANALY_LANCS, UKI-LT2-QMUL):

  lcg-cp: error while loading shared libraries:

  libcgsi_plugin.so.1: cannot open shared object file: No such file or directory

  - Bad PYTHONPATH
  - lcg-cp works fine e.g. at RAL with the new wrapper; however RAL is not yet part of the federation

- Intermittent transfer problems at ANALY_CERN_XROOTD

  - Reads for a while and then gets interrupted

  [xrootd] Total 1320.18 MB |==========>.........| 52.72 % [1.1 MB/s]

  Last server error 3010 ('Unable to open /atlas/dq2/../AOD.638838._000004.pool.root.1; Permission denied')

- Server error 10000 at ANALY_OX

  - Transfers (all) fail with errors like

  Last server error 10000 ('') Error accessing path/file for root://atlas-xrd-uk.cern.ch//atlas/dq2/mc11_7TeV/AOD/e971_s1310_s1300_r3043_r2993/mc11_7TeV.127627.simple_Gt_G750_T1200_L150_herwigpp_susy.merge.AOD.e971_s1310_s1300_r3043_r2993_tid637585_00/AOD.637585._000001.pool.root.2

efi.uchicago.edu
ci.uchicago.edu

WAN performance

Testing FAX

HammerCloud

ORACLE DB

sites, doors, ports, roles, protocols, paths

SVN
Test code
Sets release
Datasets

WEB site
SSB

Results
ping, copy time, read times

HC based FAX tests
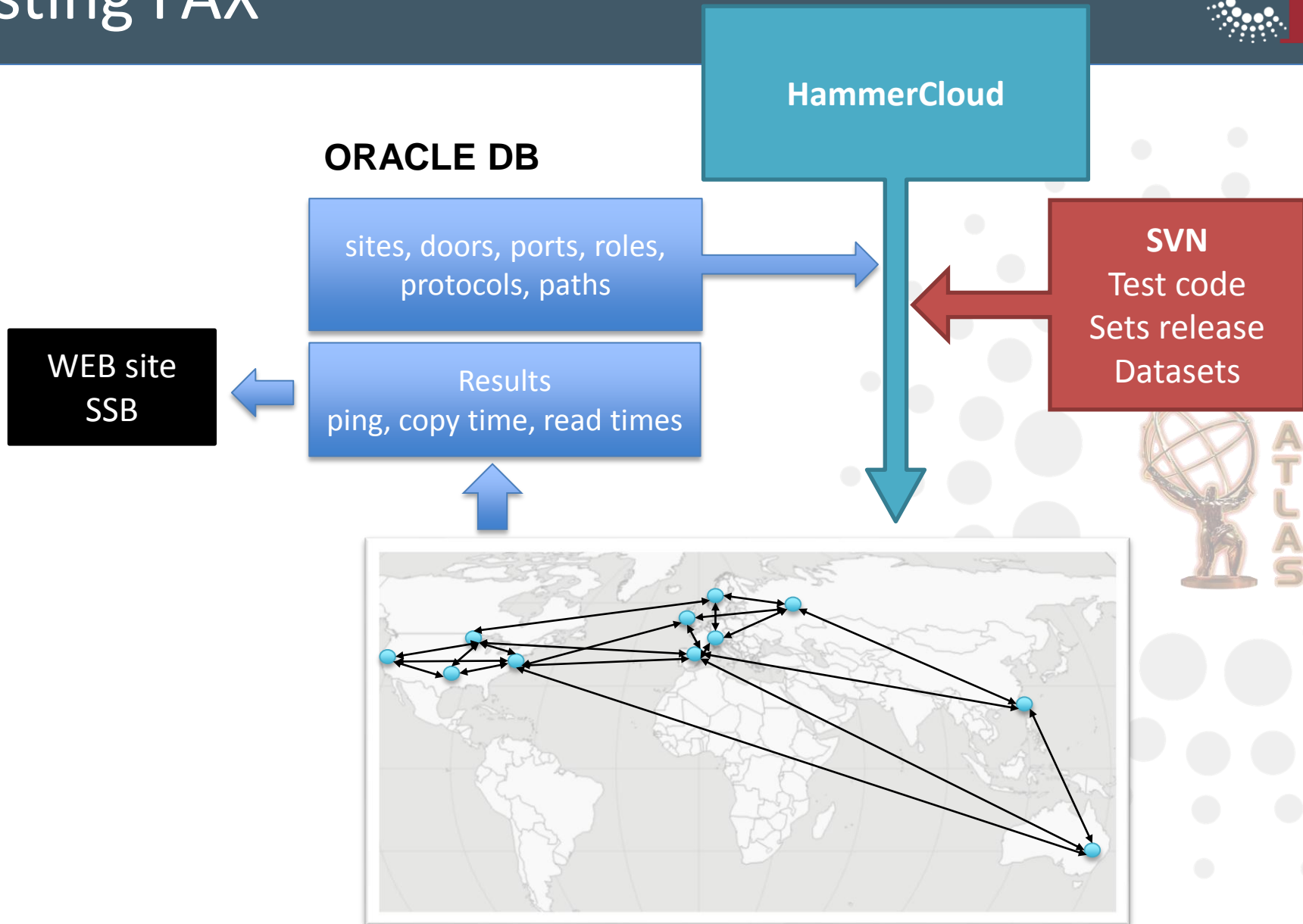
- HC submits 1 job/day to all of the "client" nodes. Client node is the one using the data. It is an ANALY queue
  - All the "server" sites have one and the same dataset. Server sites are the ones delivering data.
- Each job, each half an hour, in parallel:
  - Pings of all of the "server" sites.
  - Copies a file from a site (xrdcp/dccp)
  - Reads the file from a root script
  - Uploads all the results to Oracle DB at CERN
- Result are shown at: http://ivukotic.web.cern.ch/ivukotic/WAN/index.asp
- Results are also given in JSON format to SSB: http://dashb-atlas-ssb.cern.ch/dashboard/request.py/siteview#currentView=Network+Measurements&highlight=false
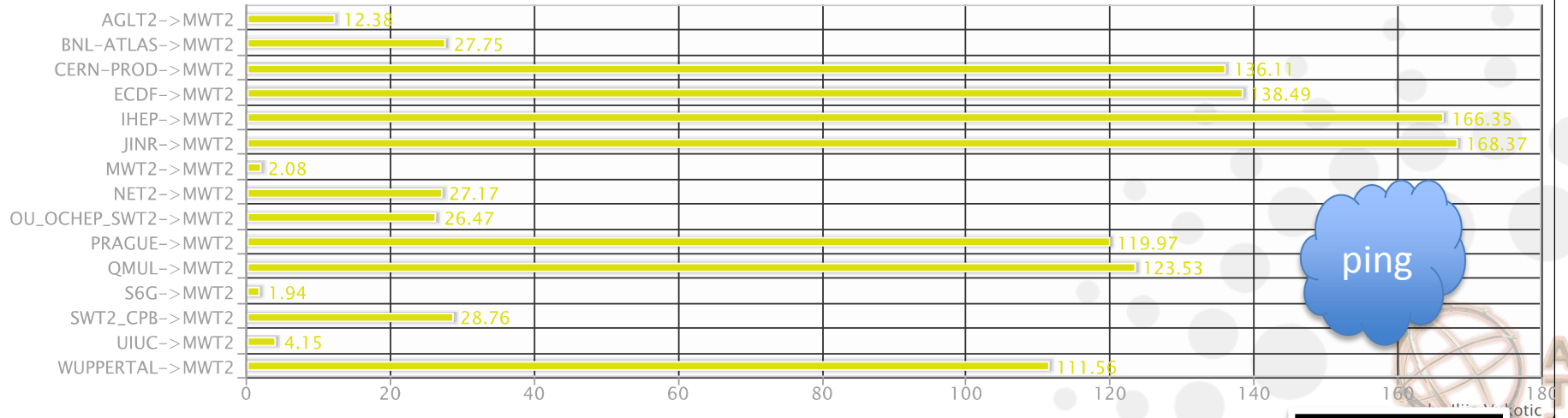
Testing details

- Test file
  - standard ATLAS 760 MB D3PD with 5k branches and 13kevents

- Measurements
  - "direct copy": time (in seconds) for xrdcp to site
  - "read time" time required to read 10% randomly selected consecutive events using default TTreeCache of 30 MB

For jobs at MWT2 (client location)

ping [ms] for 10% default cache

Location	Value
AGLT2->MWT2	12.38
BNL-ATLAS->MWT2	27.75
CERN-PROD->MWT2	136.11
ECDF->MWT2	138.49
IHEP->MWT2	166.35
JINR->MWT2	168.37
MWT2->MWT2	2.08
NET2->MWT2	27.17
OU_OCHEP_SWT2->MWT2	26.47
PRAGUE->MWT2	119.97
QMUL->MWT2	123.53
S6G->MWT2	1.94
SWT2_CPB->MWT2	28.76
UIUC->MWT2	4.15
WUPPERTAL->MWT2	111.56

ping

Read time

read time for 10% default cache

Location	Value
AGLT2->MWT2	90.7
BNL-ATLAS->MWT2	133.15
CERN-PROD->MWT2	216.37
ECDF->MWT2	459.51
IHEP->MWT2	323.84
JINR->MWT2	244.46
MWT2->MWT2	70.74
NET2->MWT2	97.43
OU_OCHEP_SWT2->MWT2	122.25
PRAGUE->MWT2	478.81
QMUL->MWT2	297.73
S6G->MWT2	76.25
SWT2_CPB->MWT2	114.84
UIUC->MWT2	75.71
WUPPERTAL->MWT2	477.86

by Ilija Vukotic

efi.uchicago.edu
ci.uchicago.edu

THE UNIVERSITY OF CHICAGO
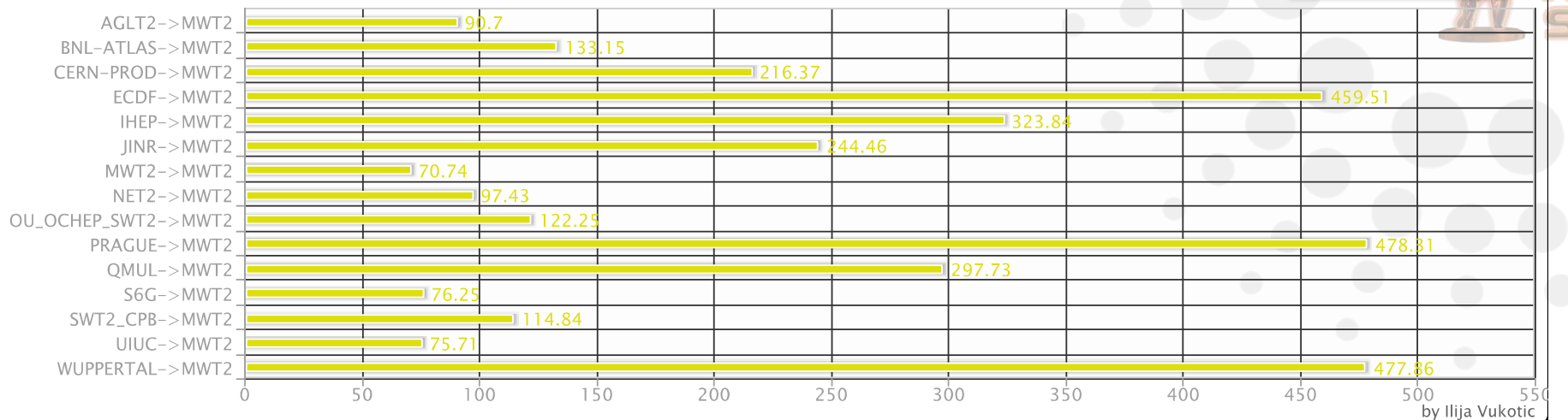
FAX 'Dress Rehearsal' ("FDR")

- Steps towards usability:
  - Define an FDR sufficient to cover most anticipated user workloads, including user docs
  - Define a rehearsal period: ~ week and metrics (and needed monitoring)
  - Poll for site volunteers and an ad-hoc FAX OPS team
  - Execute the FDR; gather monitoring statistics and accounting data
- Propose spend December preparing
  - Identifying five exemplar use cases that can be run by FAX OPS team
  - Preparing a clean set of tutorial-like documents
  - Preplace example datasets
  - Load test redirectors and sites against with examples
  - Solve the main problem of federated access to datasets
- Week of January 21 going live with real users

Conclusions

- FAX usability inches forward – but growing pains due to:
  - Standardizing metrics
  - dCache components
  - Many more sites and SEs
- First pilots bits are in production at a couple sites
- Co-located Tier3 users using FAX doors for LOCALGROUPDISK analysis
- Offers storage access from opportunistic or cloud
- Offers "diskless" use case which would be very attractive to sites for storage admin purposes
- Dress rehearsal being planned for week of January 21

Extra: FAX usability

THE UNIVERSITY OF CHICAGO

How to use it? Part - I

- Datasets should be registered
  - All the grid produced datasets are automatically registered independently if these are part of official production or simply result of a user's job.
  - If files are not registered it is trivial to do so. Very detailed description how to do this is given https://twiki.cern.ch/twiki/bin/viewauth/Atlas/DQ2ClientsHowTo.

- Have your ATLAS grid certificate
  - Make a proxy

```
source /afs/cern.ch/project/gd/LCG-share/current_3.2/etc/profile.d/grid_env.sh
voms-proxy-init -voms atlas
```

- setup DQ2

```
source /afs/cern.ch/atlas/offline/external/GRID/ddm/DQ2Clients/setup.zsh
```

- Make sure your code uses TTreeCache!

CVMFS version

efi.uchicago.edu
ci.uchicago.edu

THE UNIVERSITY OF CHICAGO

CVMFS environment setup

- ## Setup environment

```
export
ATLAS_LOCAL_ROOT_BASE=/cvmfs/atlas.cern.ch/repo/ATLASLocalRootBase
alias setupATLAS='source
${ATLAS_LOCAL_ROOT_BASE}/user/atlasLocalSetup.sh'
export ALRB_localConfigDir=$HOME/localConfig
setupATLAS
localSetupGLite
```

- ## Make a proxy

```
voms-proxy-init -voms atlas
```

- ## setup DQ2

```
localSetupDQ2Client
```

- Check that datasets exist at one of the federated sites

```
dq2-ls -r myDataSetName
```

- Find gLFN's of input datasets
  - Find closest redirector to compute site. List is here: https://twiki.cern.ch/twiki/bin/view/Atlas/FaxRedirectors
  - Do

```
export STORAGEPREFIX=root://closestRedirector:port/
```

  - make a file with the list of all the gLFN's

```
dq2-list-files -p
data12_8TeV.00201556.physics_Muons.recon.DESD_ZMUMU.f437_m716_f437 >
my_list_of_gLFNS.txt
```

How to use it? Part - III

- ## From ROOT

```
TFile *f =
TFile::Open("root://myRedirector:port//atlas/dq2/user/ilijav/HCtest/user.i
lijav.HCtest.1/group.test.hc.NTUP_SMWZ.root");
```

- ## From prun

   Instead of giving --inDS myDataset option, provide it
   with --pfnList my_list_of_gLFNS.txt

- ## copy files locally

```
xrdcp
root://xrddc.mwt2.org:1096//atlas/dq2/user/ilijav/HCtest/user.ilijav.HCtes
t.1/group.test.hc.NTUP_SMWZ.root /tmp/myLocalCopy.root
```