Computation Institute

Data Federation Strategies for ATLAS using XRootD

Ilija Vukotic
On behalf of the ATLAS Collaboration

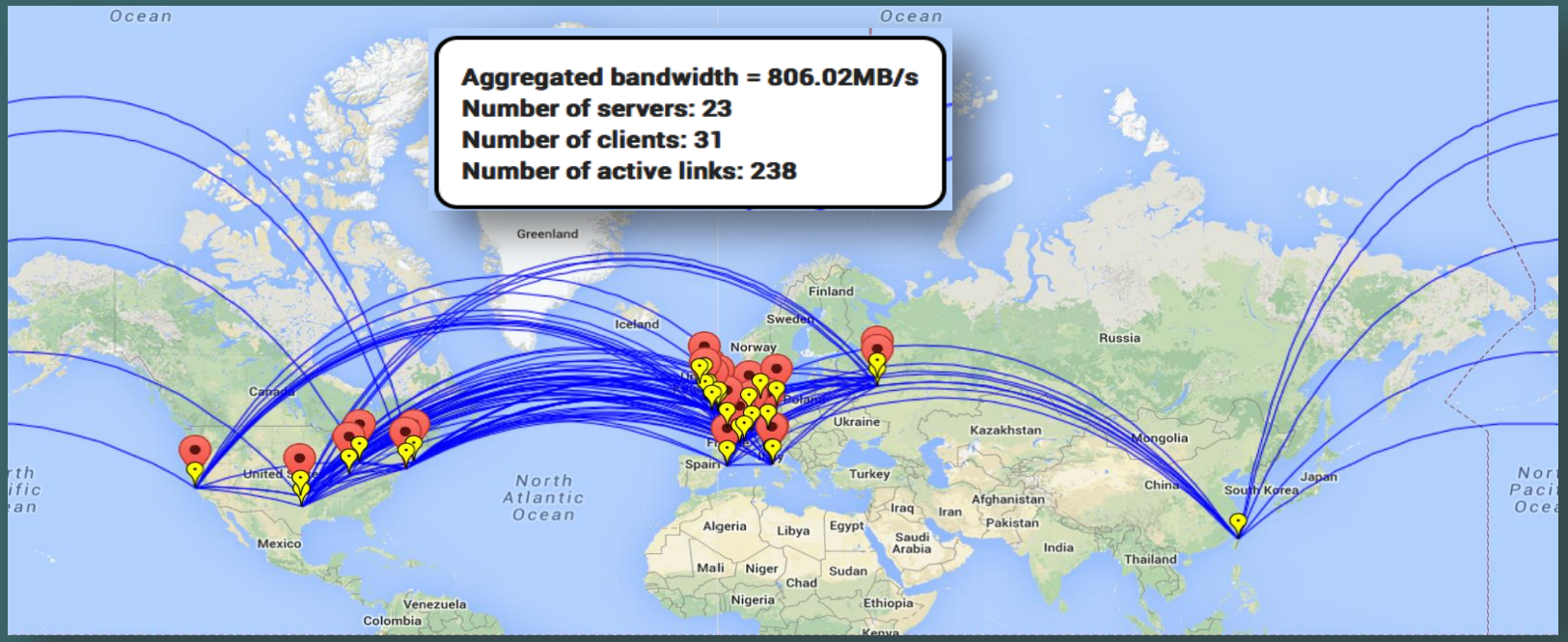
Computation and Enrico Fermi Institutes
University of Chicago

Computing in High Energy and Nuclear Physics (CHEP 2013)
Amsterdam, The Netherlands
October 2013

What is FAX?

FAX (Federated ATLAS Xrootd) is a way to unify direct access to a diversity of storage services used by ATLAS

- Read only access
- Global namespace
- Currently 42 federated sites
- Regions covered: US, DE, UK, ES, and CERN



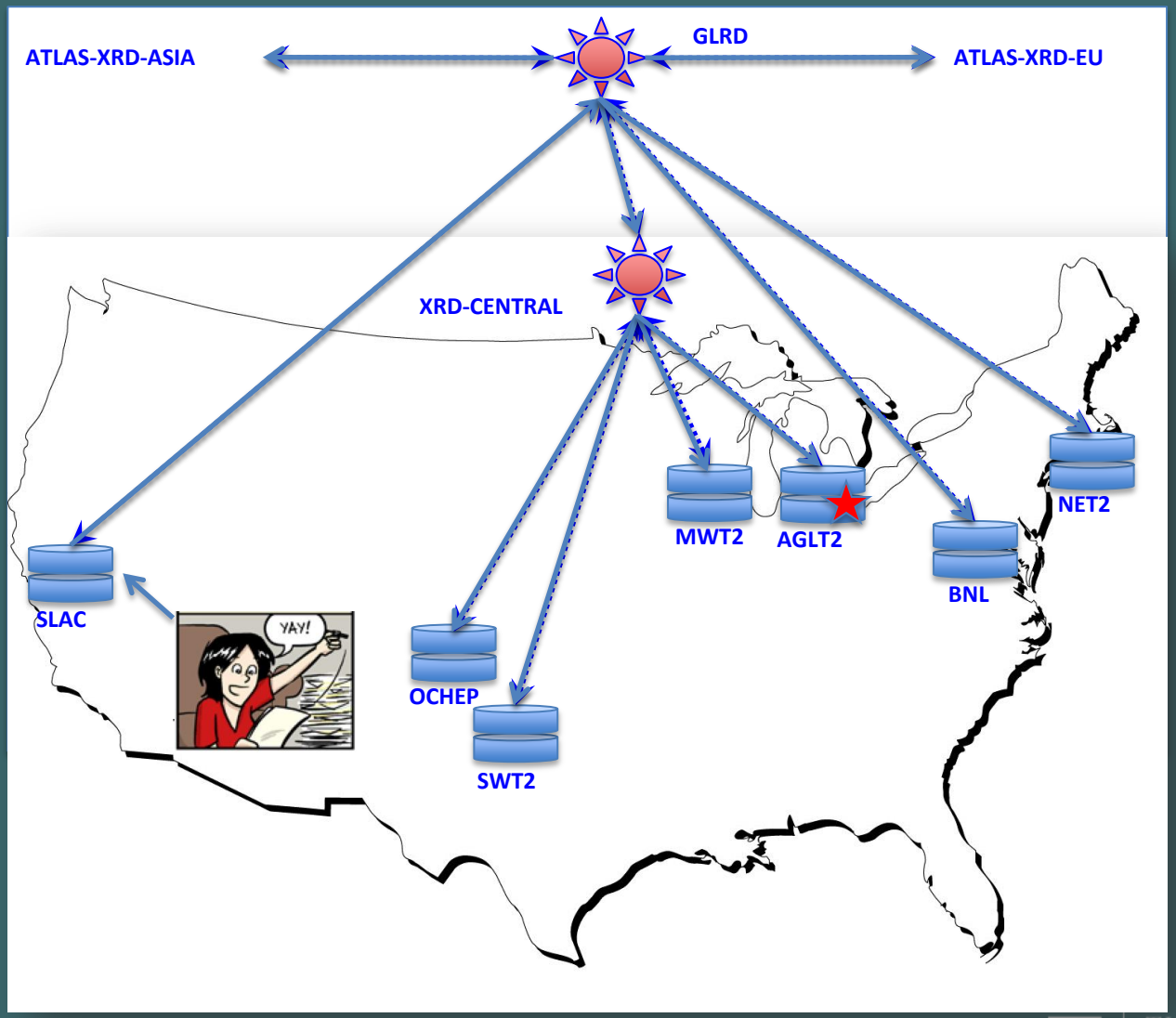
But Not only that!

- Initial use cases
  - Failover from stage-in problems with local storage
  - 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

HOW it works

redirector

endpoint



- Data can be asked for from any endpoint and redirector
- Data are transferred directly from server to user
- Searching for the file is fast, delivering is more important
- Ideally one should use the one with best connection
- Usually that's the closest one

What are the ingredients?

- gLFN – global Logical Filename
- Protocol
- Storage systems at sites
- Redirectors
- Federation description
- Monitoring
- Storage service integration
- Applications
- Documentation

Simplicity for user requires a smart (therefore complex) system.

Global Logical Filenames

- In ATLAS data management based on DQ2
  - Files are organized in DataSets, DataSets in DataContainers, exist in one or more SpaceTokens at one or more StorageElements and are registered in DQ2
  - File catalog database LFC is an Oracle DB and contains mapping of logical file names to their physical path(s)
  - Each endpoint when asked for a file, queries LFC to find exact path of the file before delivering it or responding that file is not there
  - Proved to be large complication: multiple LFCs, authentication issues, scaling, latencies
- Now moving to new DDM system - RUCIO
  - Simple function will derive PFN from gLFN
  - Much faster, more reliable, easier to set up.

Access Protocol: XRootD

- The XROOTD project aims at giving high performance, scalable, fault tolerant access to data repositories of many kinds
- Widely used in High Energy Physics
- Supported by ROOT
- XRootD clients: xrdcp and xrdfs
- File path start with **root://servername:[port]/…**
- Can serve from any posix storage, dCache xrootd plugin implements protocol

A Diversity of Storage Systems

- ATLAS uses 80+ WLCG computing sites organized in Tiers
- Various storage technologies are used: dCache, DPM, Lustre, GPFS, Storm, XRootD and EOS
- Single source of deployment documentation:https://twiki.cern.ch/twiki/bin/view/AtlasComputing/JoiningTheATLASFederation
- To ease support we have experts for different technologies
- To ease communications we have contact persons per national cloud

Storage Endpoints

Continuous Status Monitoring

Site Name	Direct	Upstream redirection	Downstream redirection	X509
UKI-NORTHGRID-LANCS-HEP	OK	OK	OK	On
CERN-PROD	OK	OK	OK	On
BNL-ATLAS	OK	OK	OK	Off
OU_OCHEP_SWT2	OK	OK	OK	On
UKI-SCOTGRID-ECDF	OK	OK	OK	On
UKI-NORTHGRID-LIV-HEP	OK	OK	OK	On
GRIF-LAL	OK	OK	OK	On
IN2P3-CPPM	OK	OK	OK	On
SWT2_CPB	OK	OK	OK	On
JINR-LCG2	OK	OK	OK	On
DESY-HH	OK	OK	OK	Off
MPPMU	OK	OK	OK	On
WUPPERTALPROD	OK	OK	OK	Off
IN2P3-LPSC	OK	OK	OK	On
RAL-LCG2	OK	OK	OK	On
INFN-FRASCATI	OK	OK	OK	On
DESY-ZN	OK	OK	OK	Off
AGLT2	OK	OK	OK	On
LRZ-LMU	OK	OK	OK	On
INFN-NAPOLI-ATLAS	OK	OK	OK	On
UKI-SCOTGRID-GLASGOW	OK	OK	OK	On
UKI-LT2-QMUL	OK	OK	OK	On
BU_ATLAS_TIER2	OK	OK	OK	Off
IFAE	OK	OK	OK	On
UKI-SOUTHGRID-CAM-HEP	OK	OK	OK	On
UKI-SOUTHGRID-OX-HEP	OK	OK	OK	On
IN2P3-LAPP	OK	OK	OK	On
TAIWAN-LCG2	OK	OK	OK	On
FZK-LCG2	OK	OK	OK	On
INFN-T1	OK	OK	OK	On
WT2	OK	OK	OK	On
RU-PROTVINO-IHEP	OK	OK	OK	On
GOEGRID	OK	OK	OK	Off
INFN-ROMA1	OK	OK	OK	On
PIC	OK	OK	OK	On
MWT2	OK	OK	OK	On
UNI-FREIBURG	OK	OK	OK	Off
GRIF-LPNHE	OK	OK	OK	On
UKI-NORTHGRID-MAN-HEP	noDirect	NoUpstreamRedirection	NoFirstLevelRedirection	Off
PRAGUELCG2	noDirect	NoUpstreamRedirection	NoFirstLevelRedirection	Off
GRIF-IRFU	OK	OK	NoFirstLevelRedirection	On

Coverage

T0	T1	T2D
1/1	6/12	34/44

41 sites

72% done

All of USA, UK, DE, IT, RU, ES

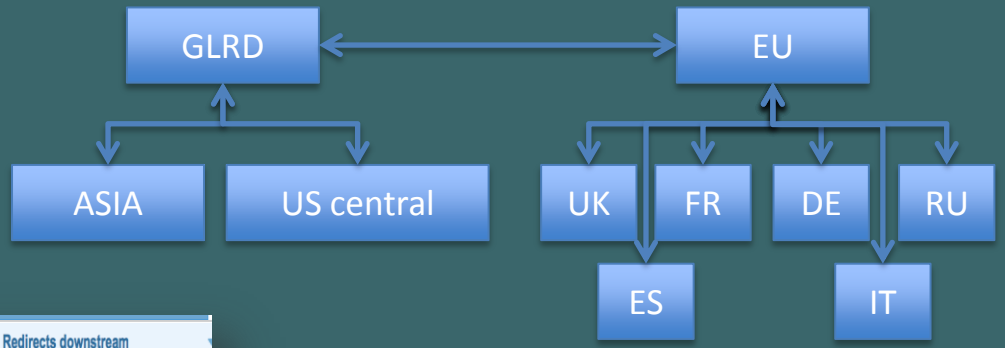
Most of FR

Recently joined: Taiwan, PIC, IFAE

Next: AU, CA, NL

Redirector Network

- Lightweight and easily managed



Site Name	Redirects upstream	Redirects downstream
XROOTD-ATLAS-XRD-FR	OK	OK
XROOTD_GLRD	OK	OK
XROOTD-ATLAS-XRD-UK	OK	OK
XROOTD-XRD-CENTRAL	OK	OK
XROOTD-ATLAS-XRD-IT	OK	OK
XROOTD-ATLAS-XRD-DE	OK	OK
XROOTD-ATLAS-XRD-ASIA-TW	OK	OK
XROOTD_ATLAS-XRD-EU	OK	OK
XROOTD-ATLAS-XRD-ES	OK	OK
XROOTD-ATLAS-XRD-RU	OK	OK

Describing the federation

- Both individual users and production/testing systems need to know access points, redirectors and their current status
- FAX topology recorded in the [ATLAS Grid Info System]

ATLAS Grid Information System

ATLASSite | DDMEndpoint | PANDA Queue | Service | DDM Groups | **Services**

Show 500 entries | First | Previous | 1 | Next | Last

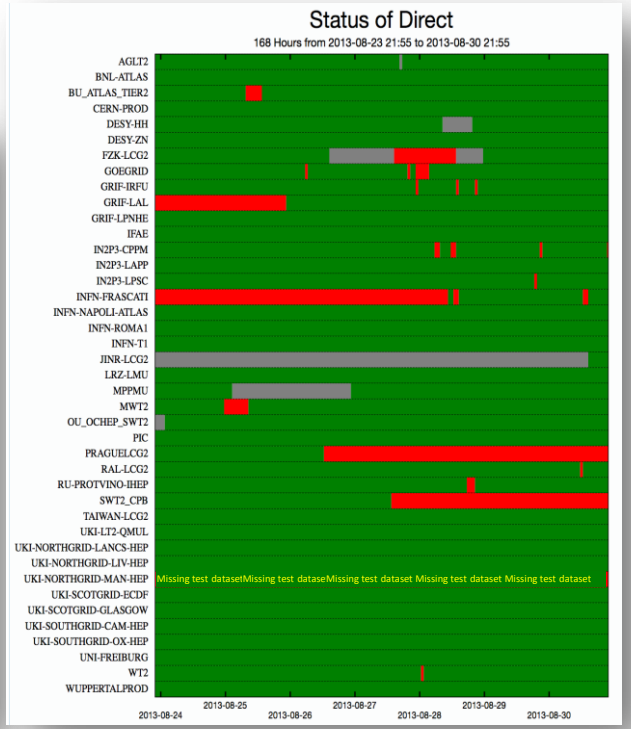
give me url of this page

Name | Service Type | Site | Endpoint | State | Status | Nodes

-- Redirector/XROOTD | ACTIVE

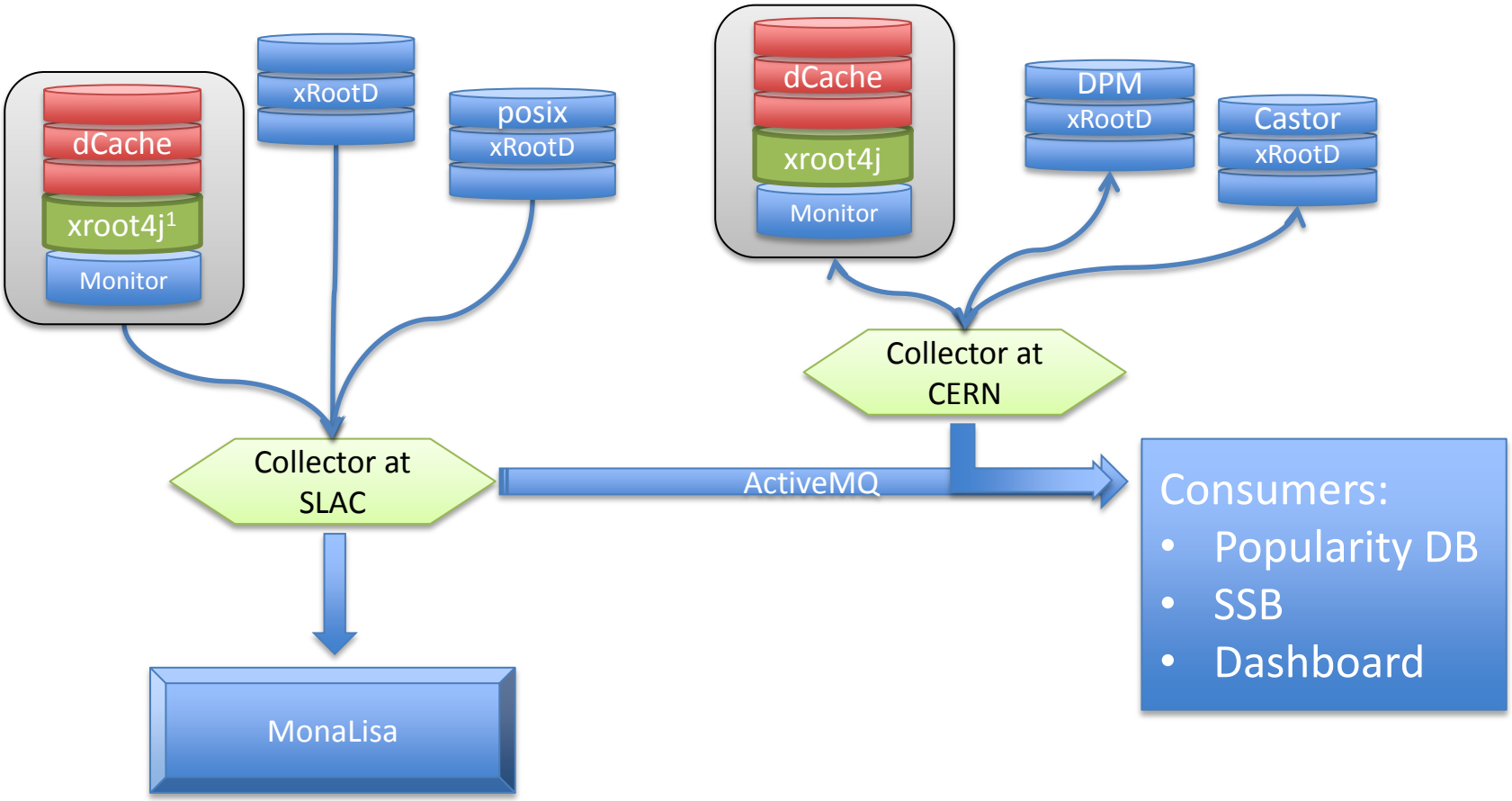
Name	Service Type	Site	Endpoint	State
XROOTD-atlas-xrd-asia-tw	Redirector/XROOTD	Taiwan-LCG2	atlas-xrd-asia.grid.sinica.edu.tw	ACTIVE
XROOTD-atlas-xrd-de	Redirector/XROOTD	CERN-PROD	atlas-xrd-de.cern.ch:1094	ACTIVE
XROOTD-atlas-xrd-es	Redirector/XROOTD	CERN-PROD	atlas-xrd-es.cern.ch:1094	ACTIVE
XROOTD-atlas-xrd-fr	Redirector/XROOTD	CERN-PROD	atlas-xrd-fr.cern.ch:1094	ACTIVE
XROOTD-atlas-xrd-it	Redirector/XROOTD	CERN-PROD	atlas-xrd-it.cern.ch:1094	ACTIVE
XROOTD-atlas-xrd-ru	Redirector/XROOTD	CERN-PROD	atlas-xrd-ru.cern.ch:1094	ACTIVE
XROOTD-atlas-xrd-uk	Redirector/XROOTD	CERN-PROD	atlas-xrd-uk.cern.ch:1094	ACTIVE
XROOTD-xrd-central	Redirector/XROOTD	MWT2	xrd-central.usatlasfacility.org:1094	ACTIVE
XROOTD_atlas-xrd-eu	Redirector/XROOTD	CERN-PROD	atlas-xrd-eu.cern.ch:1094	ACTIVE
XROOTD_glrd	Redirector/XROOTD	BNL-ATLAS	glrd.usatlas.org:1094	ACTIVE

Showing 1 to 10 of 10 entries



- Current status (and its history) are kept in [Site Status Board] (SSB)

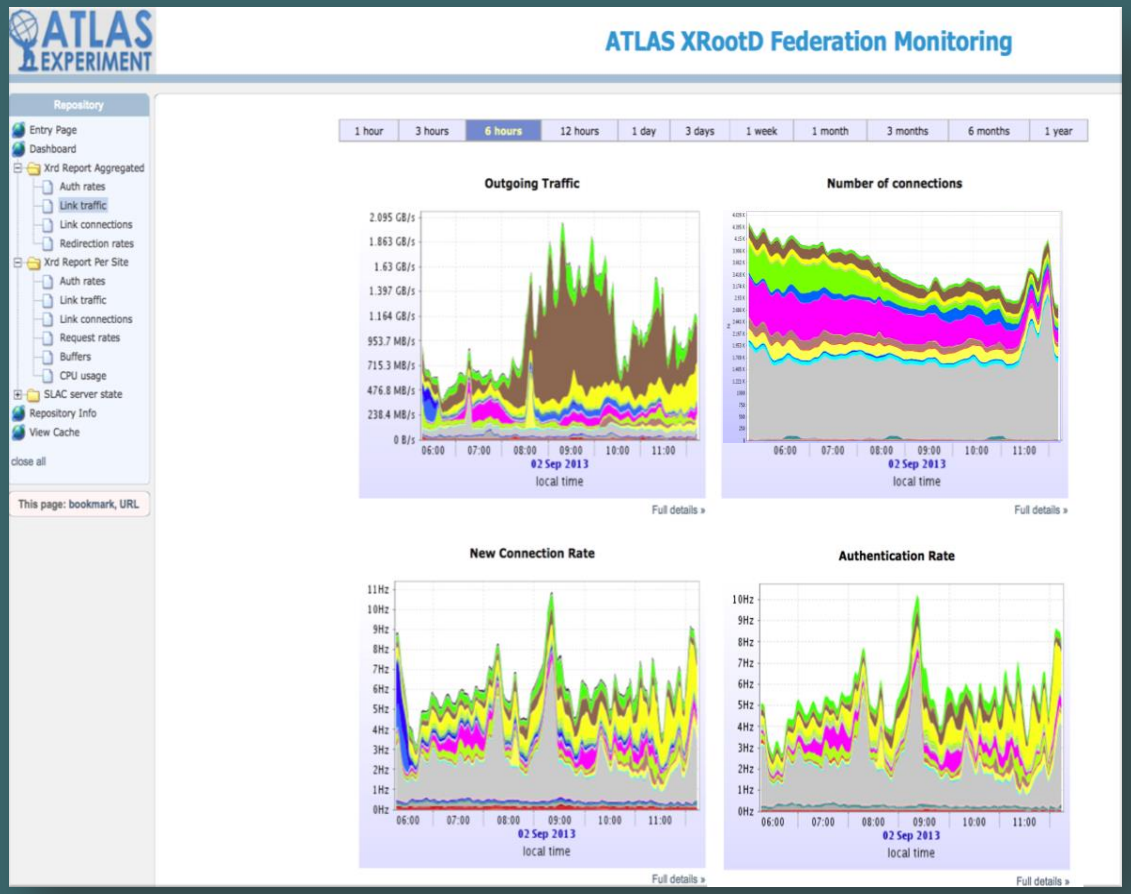
Monitoring the infrastructure



[1] dCache XRootD door

“Summary” Monitoring Stream

- Lightweight
- XML
  - throughputs
  - connections
  - redirections
  - per site
  - per server
- Results stored in postgresql
- Shown in MonaLisa



“Detailed” Monitoring Stream

- Real time → actionable!
- File info
  - Path
  - Size
  - Open
  - Close
  - Reads/writes,…
- User info
  - Hostname
  - DN
  - Process ID
- Server info
- Application Info

Can see each currently open file

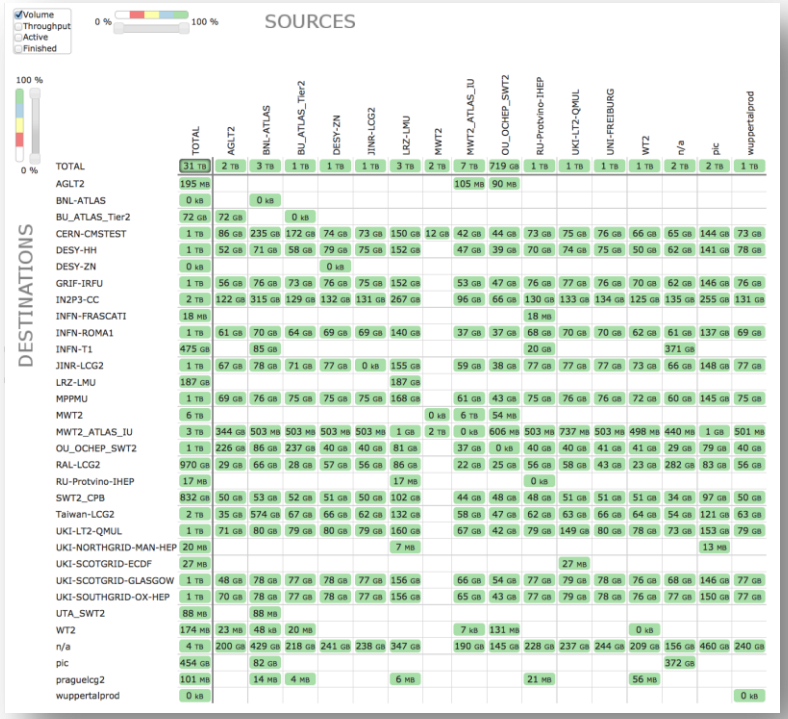
Usually >5k files

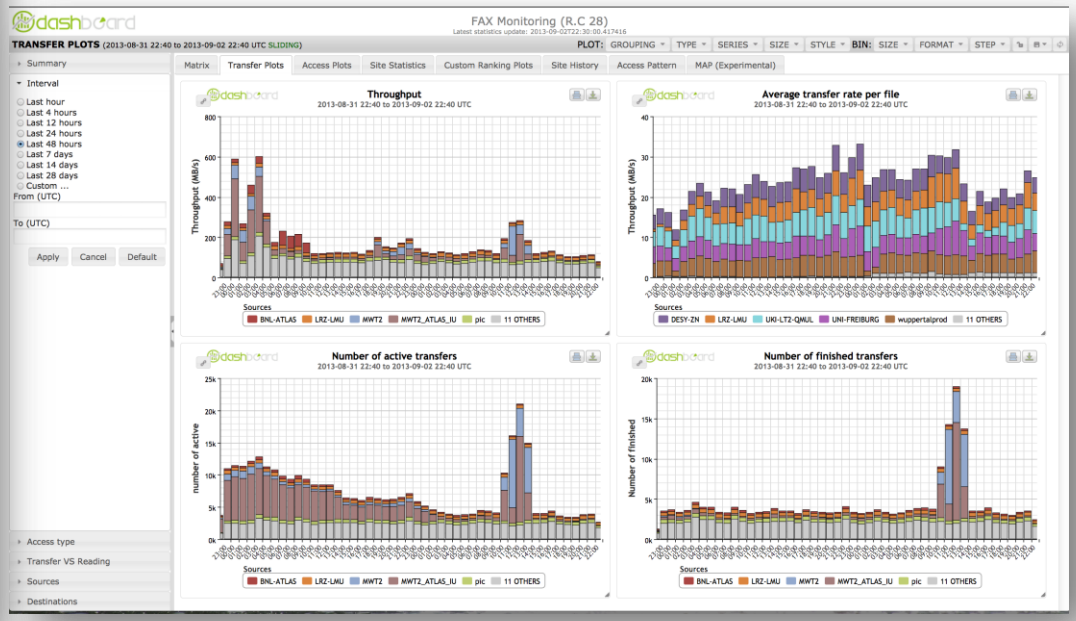
http://atl-prod05.slac.stanford.edu:4242/

File V	User	Server Domain	Client Domain	Open Ago	Update Ago	Read [MB]	Read [%]	Rate [MB/s] Λ
/atlas/dq2/mc12_8TeV/NTUP_JETMET /e1126_s1469_s1470_r3542_r3549_p...	kkrizka	uchicago.edu	uchicago.edu	08:01:11	08:00:59	9.608	0.375	0.801
/atlas/dq2/mc12_8TeV/NTUP_JETMET /e1126_s1469_s1470_r3542_r3549_p...	uct3	uchicago.edu	mwt2.org	13:24:12	12:04:10	1322.356	56.040	0.275
/atlas/dq2/user/flegger/AGLT2/user.flegger.AGLT2.data12_8TeV.002...	usatlas1	aglt2.org	bu.edu	00:02:25	00:01:25	199.356	6.837	3.323
/atlas/dq2/user/flegger/AGLT2/user.flegger.AGLT2.data12_8TeV.002...	usatlas1	aglt2.org	bu.edu	00:31:18	00:01:18	2664.860	75.331	1.480
/atlas/dq2/user/HironoriIto/user.HironoriIto.xrootd.aglt2/user.H...	atlaspil	aglt2.org	grid.sinica.edu.tw	00:01:05	00:00:58	125.000	100.000	18.750
/atlas/dq2/user/HironoriIto/user.HironoriIto.xrootd.aglt2/user.H...	atlaspil	aglt2.org	grid.sinica.edu.tw	00:00:45	00:00:38	125.000	100.000	18.750
/atlas/dq2/user/HironoriIto/user.HironoriIto.xrootd.aglt2/user.H...	atlpilot	aglt2.org	in2p3.fr	00:02:11	00:01:11	125.000	100.000	2.083
/atlas/dq2/user/HironoriIto/user.HironoriIto.xrootd.aglt2/user.H...	atlpil21	aglt2.org	gina.sara.nl	00:03:20	00:01:42	125.000	100.000	1.282
/atlas/dq2/user/HironoriIto/user.HironoriIto.xrootd.aglt2/user.H...	atlpilot	aglt2.org	datagrid.cea.fr	00:02:57	00:00:52	125.000	100.000	1.000
/atlas/dq2/user/HironoriIto/user.HironoriIto.xrootd.aglt2/user.H...	patlas92	aglt2.org	gridka.de	00:03:29	00:00:18	125.000	100.000	0.653
/atlas/dq2/user/HironoriIto/user.HironoriIto.xrootd.aglt2/user.H...	atlplt10	aglt2.org	cern.ch	00:14:10	00:01:32	125.000	100.000	0.165
/atlas/dq2/user/HironoriIto/user.HironoriIto.xrootd.aglt2/user.H...	atlplt10	aglt2.org	cern.ch	00:14:33	00:00:51	125.000	100.000	0.152

Detailed Monitoring: WLCG dashboard

- Collects and keeps all the information
- Will make it easy to slice & dice data
- This is what shifters will look at when in production





Job Failover using FAX

PanDA (**distributed production and analysis system for ATLAS**) sends a job only to a site having all the input data. In case that an input file can not be obtained after 2 tries, the file will be obtained through FAX in case it exists anywhere else.

There are in average 2.8 copies of files from recent reprocessing in FAX so there is a large change of success.

Ultimate success would be for a site to for example update its storage without first draining it of jobs.

- Small number of jobs failing
- But these failures cost the most in terms of user's turn-around time

FAX usage in Job Brokering

One can broker jobs to sites that don't have all or part of input data and use FAX to read them directly. Beneficial in cases where

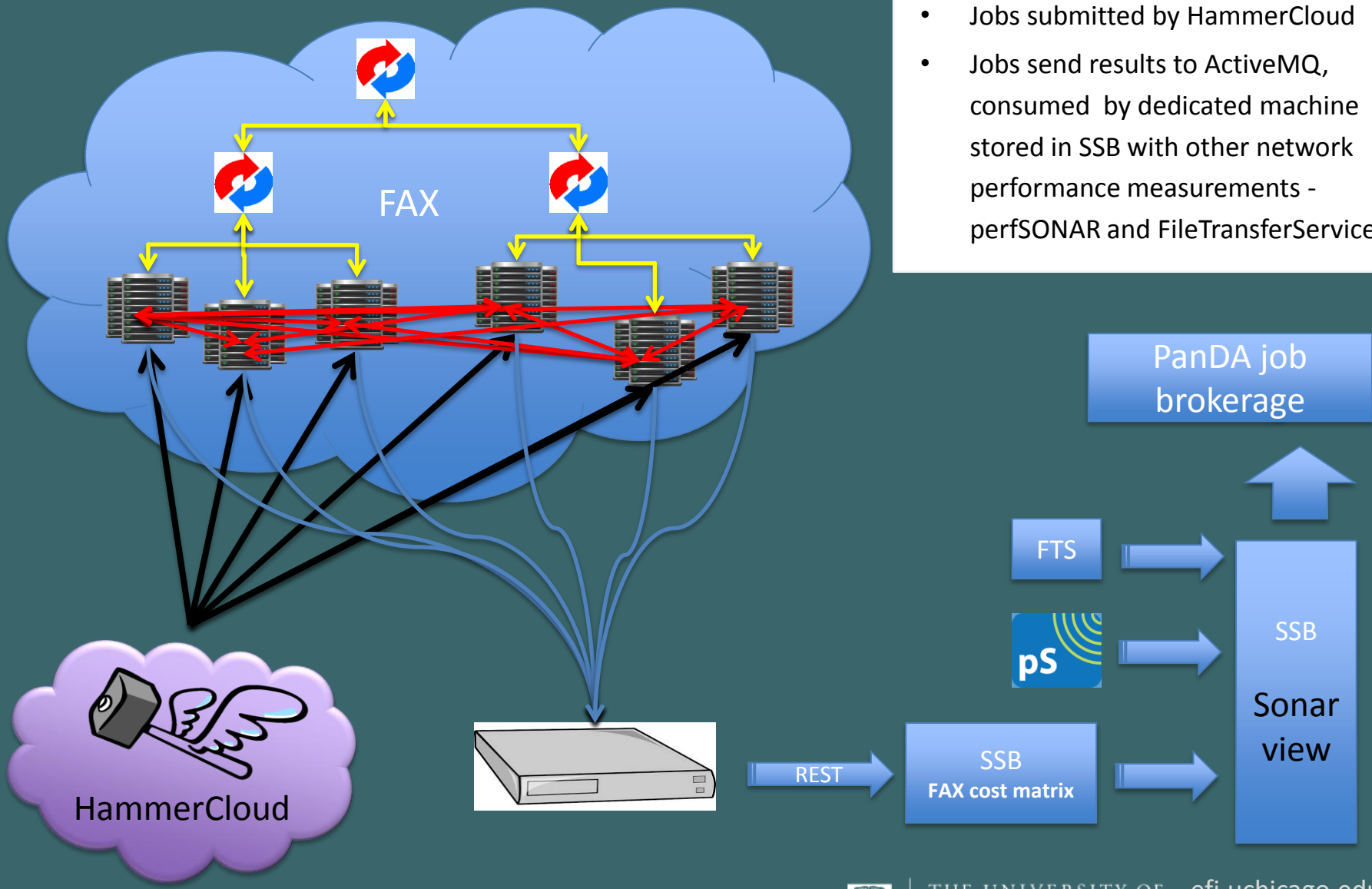
- A site has very often full queues as some specific data exist only there
- A site has free CPUs, good connectivity but not enough storage/data
- One can use external sources of CPU cycles (OSG, Amazon/Google cloud based queues,…)

For this approach to be efficient, system has to "know" expected available bandwidth between a queue and all of the FAX endpoints

- We continuously measure this data and store it in **Cost Matrix**.
- We expect brokering to FAX to be functional in a few months.

Cost matrix

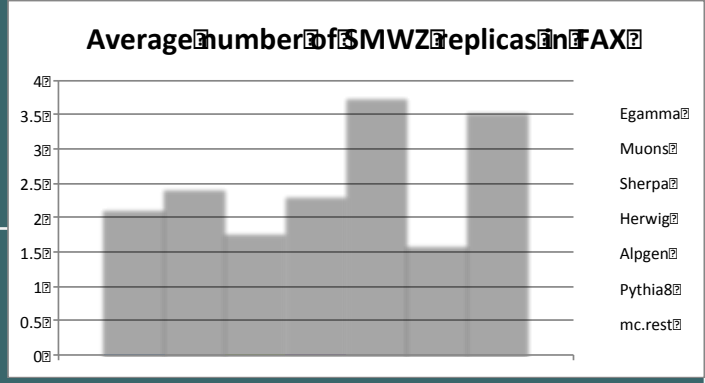
- Measures transfer rates (memory-to-memory) between 42 ANALY queues and each FAX endpoint
- Jobs submitted by HammerCloud
- Jobs send results to ActiveMQ, consumed by dedicated machine stored in SSB with other network performance measurements - perfSONAR and FileTransferService



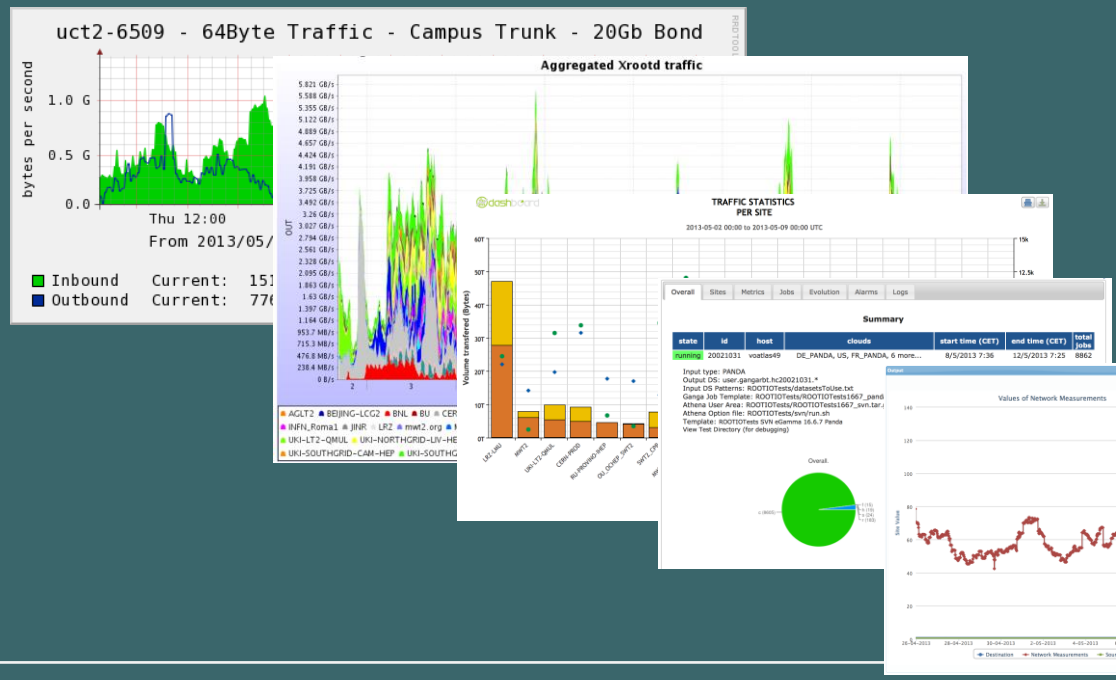
FAX-based applications

- With dedicated or on-demand CPU, one can provide *"MapReduce-like"* web service applications
- Ongoing development
  - Event Picking Service
    - Plucks small chunks of data from large numbers of data sets
  - Skim and Slim Service
    - To help users go from multi-terabyte scale to multi-gigabyte scale
    - A common grid use case
    - A lot of optimizations possible if provided as a service
    - Need to find balance between making it general and useful enough while limiting resource usage

Performance



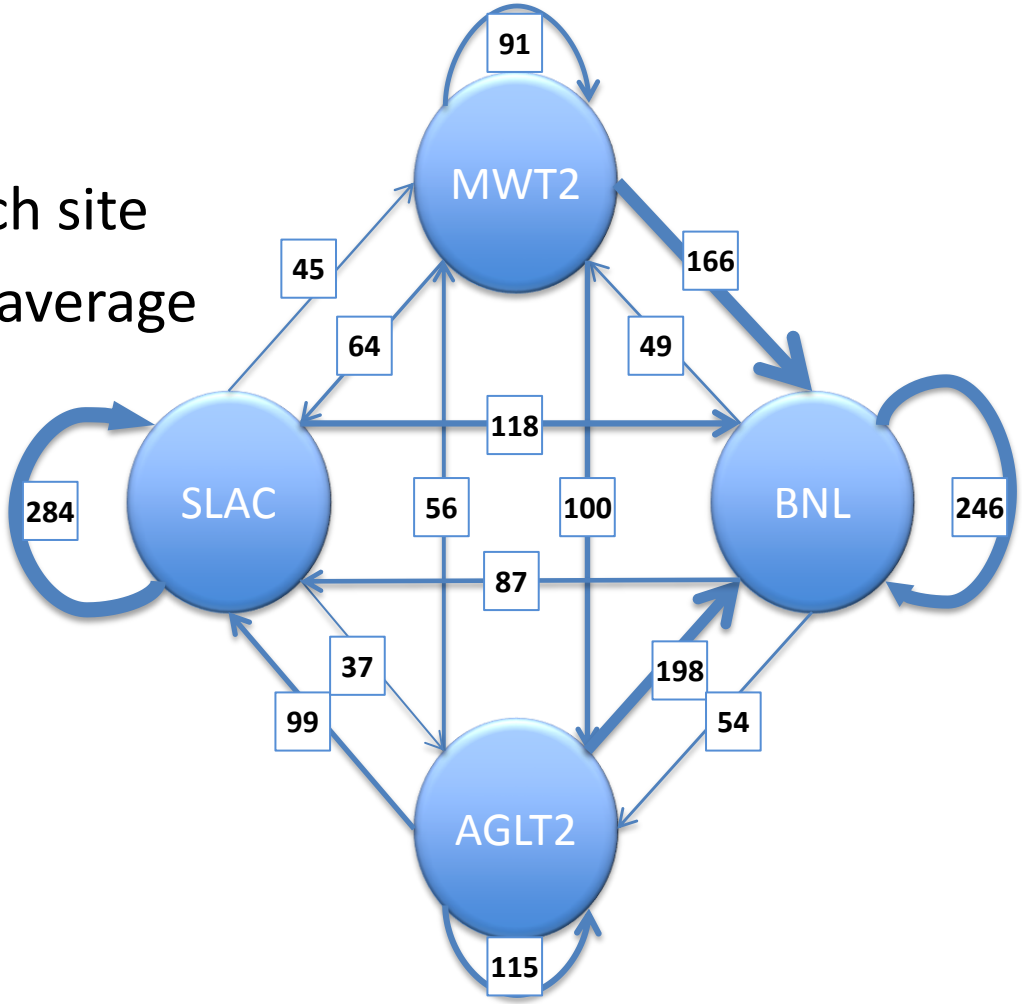
- Metrics
  - Data Coverage **>98%, >2.8 replicas**
  - Number of users **production, few tens of individual users**
  - Percentage of successful jobs **>99% from last tests**
  - Total amount of data delivered **~2PB/week**
  - **Bandwidth usage**
- Source
  - Ganglia plots
  - MonaLisa
  - FAX Dashboard
  - HC tests
  - CostMatrix tests
  - Special tests using dedicated resources



Performance

- HammerCloud based test
- Running real analysis code
- 100 concurrent jobs from each site
- Arrow width proportional to average event rate

WAN performance can be as good as LAN



Future

- Increase coverage, add redundancy, increase total bandwidth
  - Add the rest of sites
- Increase performance, reduce bandwidth needs
  - Caching
  - Cost matrix – smart FAX
  - Smart network - Bandwidth requests, QOS assurance
- Improve adoption rate
  - Presenting, teaching, preaching
  - New services
- Improve satisfaction
  - FAX tuning
  - Application tuning
  - New services

Conclusions

- Federation is functional
- First use cases implemented
- Usage is increasing
- Powerful tool that has to be used wisely
- Took more effort than initially expected
  - A lot of people (in different time zones) involved
  - Had to develop things multiple times for different storage technologies
  - Had to integrate it in our frameworks
  - Establishing effective monitoring as complex as establishing federation

Thank you!
Questions?