Computation Institute

ATLAS Experiment Status
Run2 Plans
Federation Requirements

Ilija Vukotic

XRootD Workshop @ UCSD
San Diego
27 January, 2015

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

Part 1 – overview of changes

Run2 brings new challenges:

- 1kHz trigger rate
- 40 events pile-up
- flat resources

But the long shutdown gave us an opportunity for fundamental changes. And we used it:

- New models and policies
- New DDM system: Rucio
- New distributed production framework: ProdSys2
- Opportunistic resources – clouds, HPC
- New data format: xAOD
- New analysis model
- Large changes in reconstruction codes – 3x speed up!

Run-1 model, briefly

- Strict hierarchical model (Monarc):
  - Clouds: T1 + T2s (+ T3s)
  - No direct transfers between foreign T2s
  - Relaxed towards the end of Run-1 (Multi-cloud production – T2s can process jobs of many clouds)
- Production organization:
  - Tasks assigned to T1s
  - T1 is the aggregation point for the output datasets of the tasks
  - T2 PRODDISK used for input/output transfers from/to T1
- T2 disk space:
  - distribute the final data to be used by analysis
  - store secondary replicas of precious datasets

Planning for Run-2 model - facts

- Network globally improved
  - Much higher bandwidth (an order of magnitude increase)
  - Most of the links between ATLAS sites provide sufficient throughput : full mesh for transfers can be used
- Many Tier-2 sites provide the Tier-1 level stability of computing, storage and WAN
  - Many in LHCONE or other high-throughput networks
  - Tape resource is the only difference between Tier-1s and large Tier-2s, as far as the usability for ATLAS is concerned
- CPU only (opportunistic) centers are fully integrated in ATLAS
  - Some run all kind of tasks, including data reprocessing
  - Have good connectivity to geographically close Storage Elements

CPU and Storage organization

- Breaking the barrier between the Storage Element and Computing Element:
  - Remote I/O, job overflow, remote fail-over of input or output file staging → storage not strictly bound to the site computing resource
  - Tier-1, Tier-2, Tier-3 storage classification does not make much sense anymore
- ATLAS Storage pool:
  - TAPE
  - STABLE disk storage – T1 + reliable T2 (former T2Ds). For storing custodial, primary data
  - UNSTABLE disk storage – less reliable T2s. For secondary data (for analysis)
  - VOLATILE disk storage – unreliable T2s, T3s, opportunistic storage. LOCALGROUPDISK SEs, Rucio cache storage, Tier3 Analysis
- Disk Space:
  - Lifetime-based Storage Model
  - Disk - Tape residency almost 100% algorithmically managed

Job optimizations

- Production / Analysis
  - Run-1: 75% / 25% (slots occupancy ~ cputime usage)
  - Run-2: 90% / 10% (no estimate yet)
    - Bulk of analysis (Derivation) moving to (group) production
    - Remaining analysis will be shorter and I/O intensive
- Reduce the merging
  - Avoid it if possible (simulation, reconstruction)
  - Local merging – merge on the site, where the files to be merged are
- Jobs will produce bigger outputs
  - Good for tape storage
  - Bigger files transferred – good for efficient transfers (but less files to transfer)
- Massive multicore for ~80% of production
  - All G4 simulation and all digi+reco
  - Effective drop in running jobs from 200k to 60k (20k 8-core + 40k single-core)
- JEDI dynamic resizing – tune the jobs to 6-12h
  - Avoid failures and cpu losses for very long jobs
- Automatic healing:
  - Split jobs too long
  - Increase memory requirements for out-of-memory failing jobs

Specializing the sites for workloads

- not all the sites are equal
- not all the job types run equally well on all the sites
- some sites are slow for analysis but they are good for data reprocessing
- some sites are very big but cannot run 100% of heavy I/O jobs
- differentiation was already used during Run-1 by limiting the job types through the fairshare (AGIS settings)
  - e.g. evgensimul=60%,all=40%
- not all the jobs are EQUALLY important:
  - Some tasks have short deadline
  - Some large activities have close deadline (physics conferences)
- FUTURE:
- Dynamic specialization:
  - I/O expensive jobs will be automatically throttled by the central system based on recent history – keeping track of data transferred to site and reduce the heavy job assignment
  - Migration from fixed bamboo queues to per task/job heaviness estimates
- Forced specialization:
  - ADC will specialize sites for certain activities, if the site provides custom resources (more memory per cpu, GPU availability ...)

Rucio

Run 2 Data Management model

In production
Ironing kinks
Adding functionality

- File level granularity
- Multiple ownerships (user/group/activity)
- Policy based replication for space/network optimization
- Features:
  - Plug in based architecture supporting multiple protocols (SRM/gridFTP/xrootd/HTTP...)
  - Unified dataset/file catalog, supports metadata

Lifetime-based Storage Model

- Use updated version of current model to calculate CPU usage and dataset volumes
- Keep track of creation date of each dataset
- Remove each dataset from disk and tape when it has exceeded its (type-specific) Lifetime.
- Adjust total request for disk and tape to be sufficient to accommodate the requested lifetimes.
- Assign an experience-driven fraction of the disk+ tape storage to disk.
- Model is now being refined
- Will be used to inform the ATLAS resource request for 2016 in February 2015

ProdSys2

- DEfT – task request and task definition
  - new web interface to requests, review, submission
  - templates for easy task definitions.
  - chaining of steps, requests – series of chains
  - post-production interface
- JEDI – dynamic job definition and task execution
  - integrated with PanDA
  - engine for user analysis tasks
  - New features:
    - dynamic job definition
    - lost file recovery
    - network-aware brokerage
    - log file merging
    - output merging
    - support for event service
- PanDA – covered later today by Kaushik De
- BigPanDA - brand new monitor

Adding functionality
Validating workflows

In production since August
Tuning parameters

In production since August

New data format: xAOD

- "Dual-use" xAOD replaces separate ATHENA-readable and Root-readable formats.
- Will be covered in details by Doug Benjamin

In use
Some tuning still possible

New Analysis Model

- More efficient
- More user-friendly
- Derivation Framework (trains with carriages provided by groups) replaces incoherent "group production"
- Analysis Framework, supporting standardized use of performance group recommendations (jet energy scale etc.)

In validation

In use
Some tools in validation

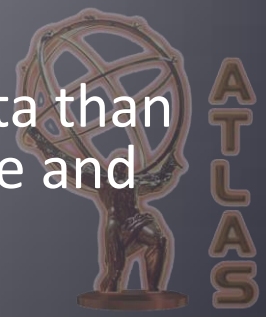
Part 1 – Conclusion

- ATLAS is making excellent progress towards readiness for Run 2.
- New production and data management system provides many possibilities for further improvements and dynamic optimizations
- Many of the changes can be implemented before the Run-2 starts
- A lot of things need tuning
- Even during the Run-2 we can afford to bring drastic improvements to our distributed system
- The production STABILITY will be the FIRST PRIORITY during data taking

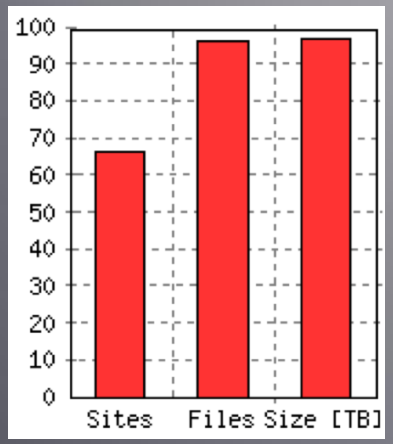
- Unknown: What will groups and physicists really do?

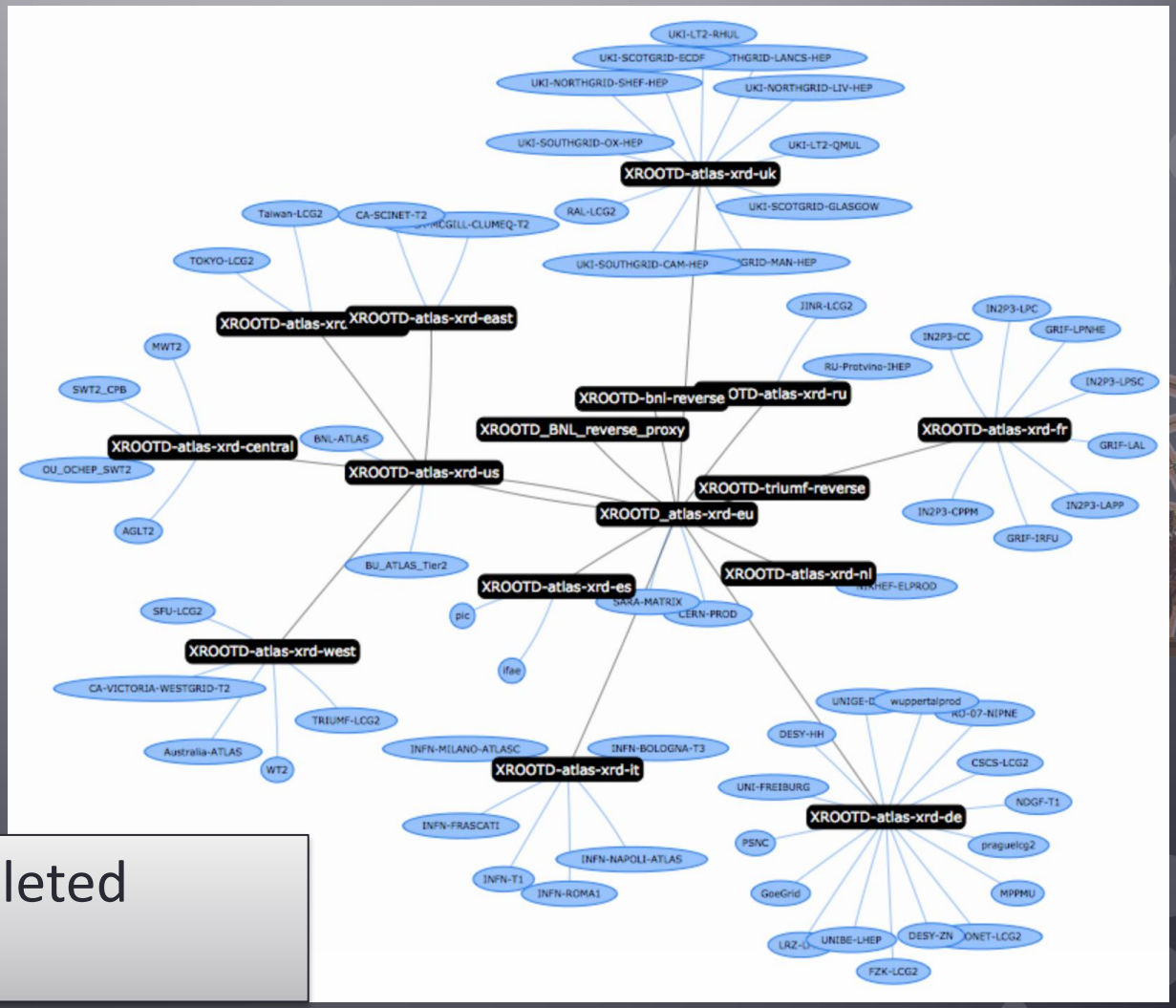
Part 2 – Federation Role and scale in Run2

- Three main roles
  - Enable remote IO jobs
    - Increase job turnover (shorter wait in queues, faster startup)
    - Increase wall time utilization of CE (no wait for input data)
    - Lower number of replicas needed
    - Better use available bandwidth by streaming only what is actually needed
  - Enabling users to easily and efficiently access much more data than they could possibly have locally. Make diskless Tier3s possible and practical.
  - Add redundancy to the existing data delivery mechanisms.
- An ideal scale would be the one where we use all the available bandwidth to make all the CPU's busy and have a minimal number of rarely accessed files/datasets.
  - Currently system is much simple than that.



Part 3 – FAX currently





Deployment completed
Coverage >96%

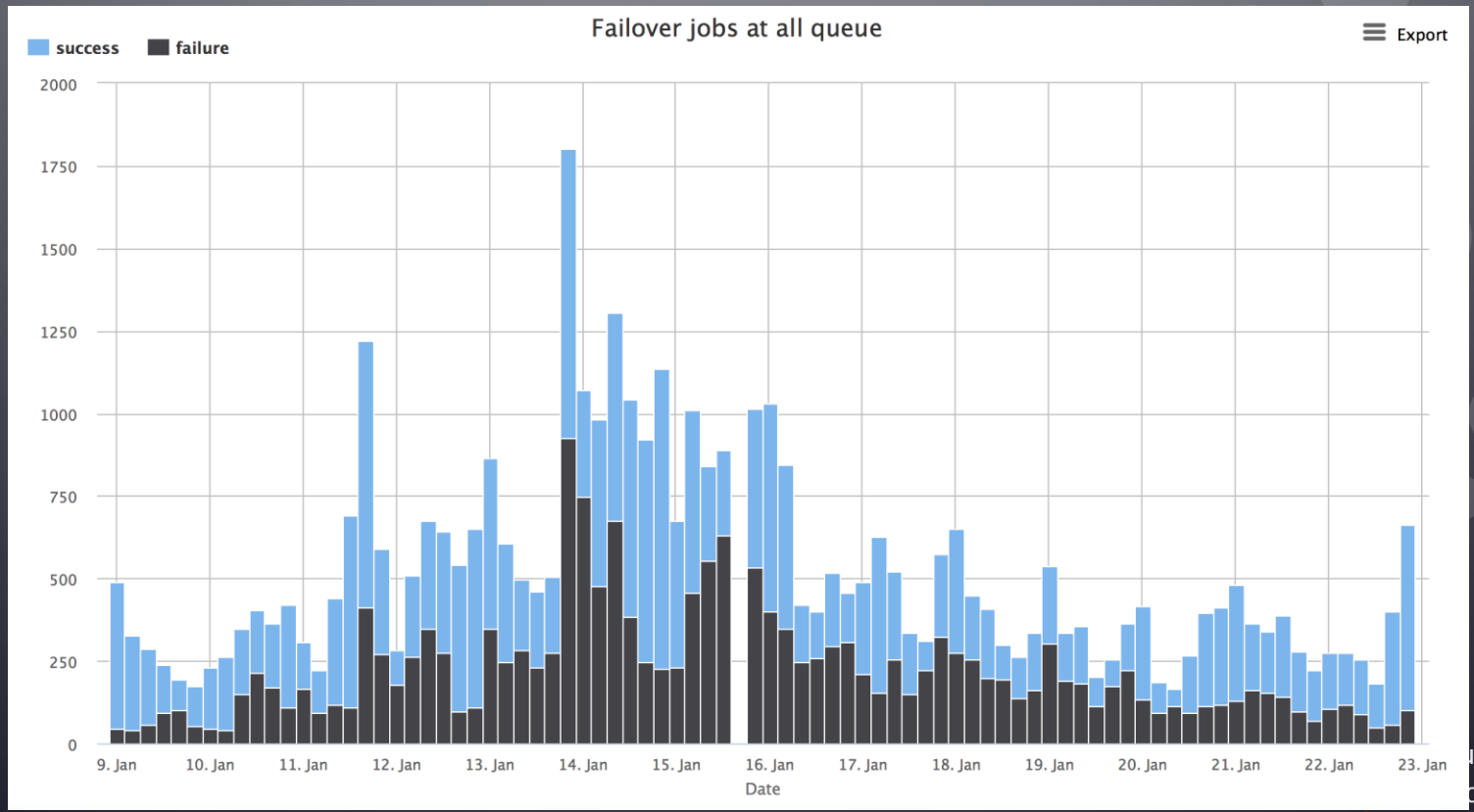
FAX - stability

- Most sites running stably
- Glitches do happen but are fixed usually in few hours
- No signs of stability issues caused by load

Site Name	Direct	Upstream redirection	Downstream redirection	ATLAS readonly
UKI-NORTHGRID-LANCS-HEP	OK	OK	OK	On
pic	OK	OK	OK	On
UKI-SCOTGRID-ECDF	OK	OK	NoFirstLevelRedirection	On
UKI-NORTHGRID-LIV-HEP	OK	OK	OK	On
wuppertalprod	OK	OK	OK	Off
CERN-PROD	OK	OK	OK	On
BNL-ATLAS	OK	OK	OK	On
OU_OCHEP_SWT2	OK	OK	OK	On
IN2P3-LPC	OK	OK	OK	On
CYFRONET-LCG2	OK	OK	OK	On
SFU-LCG2	OK	OK	OK	On
MWT2	OK	OK	OK	On
INFN-ROMA1	OK	OK	OK	On
INFN-BOLOGNA-T3	OK	OK	OK	On
TOKYO-LCG2	OK	OK	OK	On
RU-Protvino-IHEP	OK	OK	OK	On
INFN-MILANO-ATLASC	OK	OK	OK	Off
IN2P3-CPPM	OK	OK	OK	On
SWT2_CPB	OK	OK	NoFirstLevelRedirection	On
CA-SCINET-T2	OK	OK	OK	On
NIKHEF-ELPROD	OK	OK	NoFirstLevelRedirection	On
DESY-HH	OK	OK	OK	On
MPPMU	OK	OK	OK	On
IN2P3-LPSC	OK	OK	OK	On
RAL-LCG2	OK	OK	NoFirstLevelRedirection	On
NDGF-T1	OK	OK	OK	Off
IN2P3-CC	OK	OK	OK	On
INFN-FRASCATI	OK	OK	OK	On
DESY-ZN	OK	OK	OK	On
AGLT2	OK	OK	OK	On
UNIGE-DPNC	OK	OK	OK	On
LRZ-LMU	OK	OK	OK	On
CA-MCGILL-CLUMEQ-T2	OK	OK	OK	On
praguelcg2	OK	OK	OK	On
INFN-NAPOLI-ATLAS	OK	OK	NoFirstLevelRedirection	On
PSNC	OK	OK	OK	On
Taiwan-LCG2	OK	OK	OK	On
UKI-SCOTGRID-GLASGOW	OK	OK	OK	On
UKI-LT2-QMUL	OK	OK	NoFirstLevelRedirection	Off
UNIBE-LHEP	OK	OK	NoFirstLevelRedirection	On
ifae	OK	OK	OK	On
UKI-NORTHGRID-MAN-HEP	OK	OK	OK	On
RO-07-NIPNE	OK	OK	OK	On
UKI-SOUTHGRID-CAM-HEP	OK	OK	OK	On
UKI-SOUTHGRID-OX-HEP	OK	OK	OK	On
UKI-LT2-RHUL	OK	OK	OK	On
IN2P3-LAPP	OK	OK	OK	On
BU_ATLAS_Tier2	OK	OK	OK	On
FZK-LCG2	OK	OK	OK	On
TRIUMF-LCG2	OK	OK	OK	On
INFN-T1	OK	OK	OK	Off
WT2	OK	OK	OK	On
UKI-NORTHGRID-SHEF-HEP	OK	OK	OK	On
JINR-LCG2	noDirect	NoUpstreamRedirection	NoFirstLevelRedirection	Off
SARA-MATRIX	noDirect	NoUpstreamRedirection	NoFirstLevelRedirection	Off
CSCS-LCG2	noDirect	NoUpstreamRedirection	NoFirstLevelRedirection	Off
GoeGrid	noDirect	NoUpstreamRedirection	NoFirstLevelRedirection	Off
UNI-FREIBURG	noDirect	NoUpstreamRedirection	NoFirstLevelRedirection	Off
GRIF-IRFU	offline	offline	offline	offline

FAX usage

- Physicists are slowly starting to use it
  - Mainly thanks to our Offline software tutorial sessions.
  - Only anecdotal evidence – due to both privacy and monitoring issues
- Failovers – jobs that could not access the data from local storage, try to get them elsewhere
  - Few thousands jobs / day. Saves roughly half of them.



uchicago.edu
chicago.edu

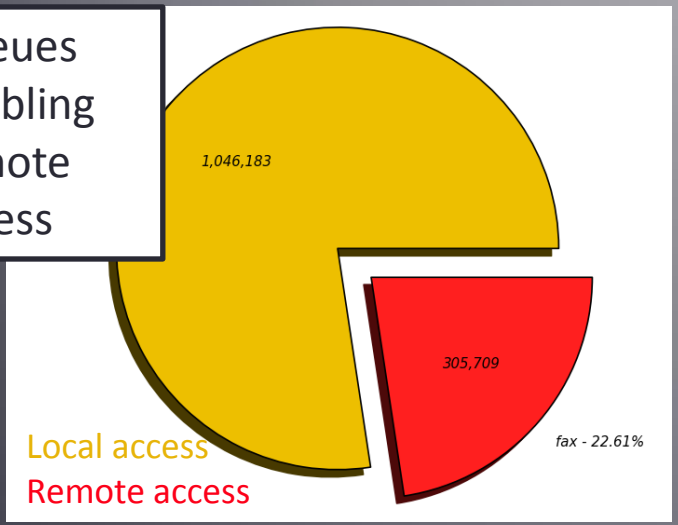
FAX usage

- Overflows – our term for jobs brokered to a site that does not have the input data, sent with explicit instruction to use FAX to get them.
- Goals for beginning of Run2:
  - Handle 5-10% of all the Analysis jobs
  - Have job efficiency roughly the same as jobs locally accessing data
  - Have CPU efficiency at least 50% of the locally run jobs
- Decision to overflow is made by JEDI based on:
  - Where is the data
  - How busy is the destination site
  - What kind of data rate job can expect between source and destination
- In operation since August.
- Initially only enabled in USA. Now includes 4 sites in Europe.
- Started with a suboptimal system, big improvements still to come.

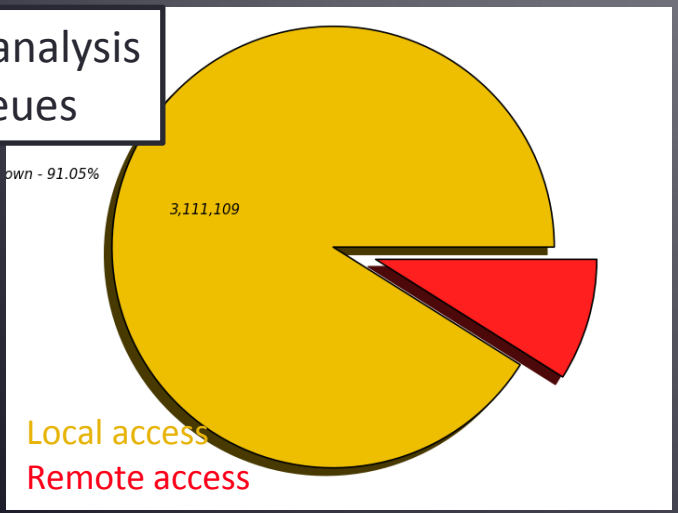
Job rates and efficiencies

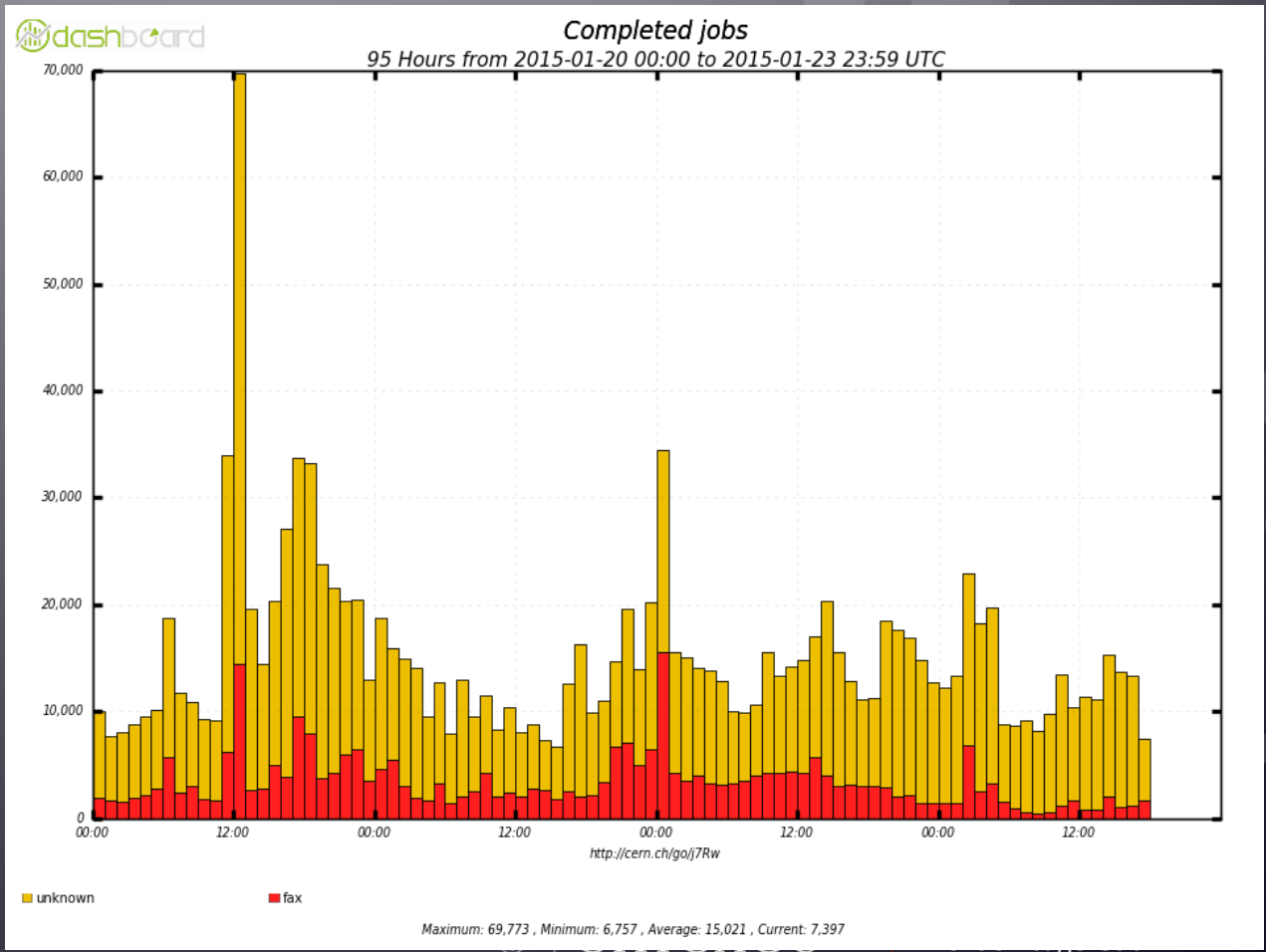
Analysis jobs per hour		
Local access		Remote access
All queues	Queues enabling overflow	
33k	11k	3.2k (9/22%)

Queues enabling remote access



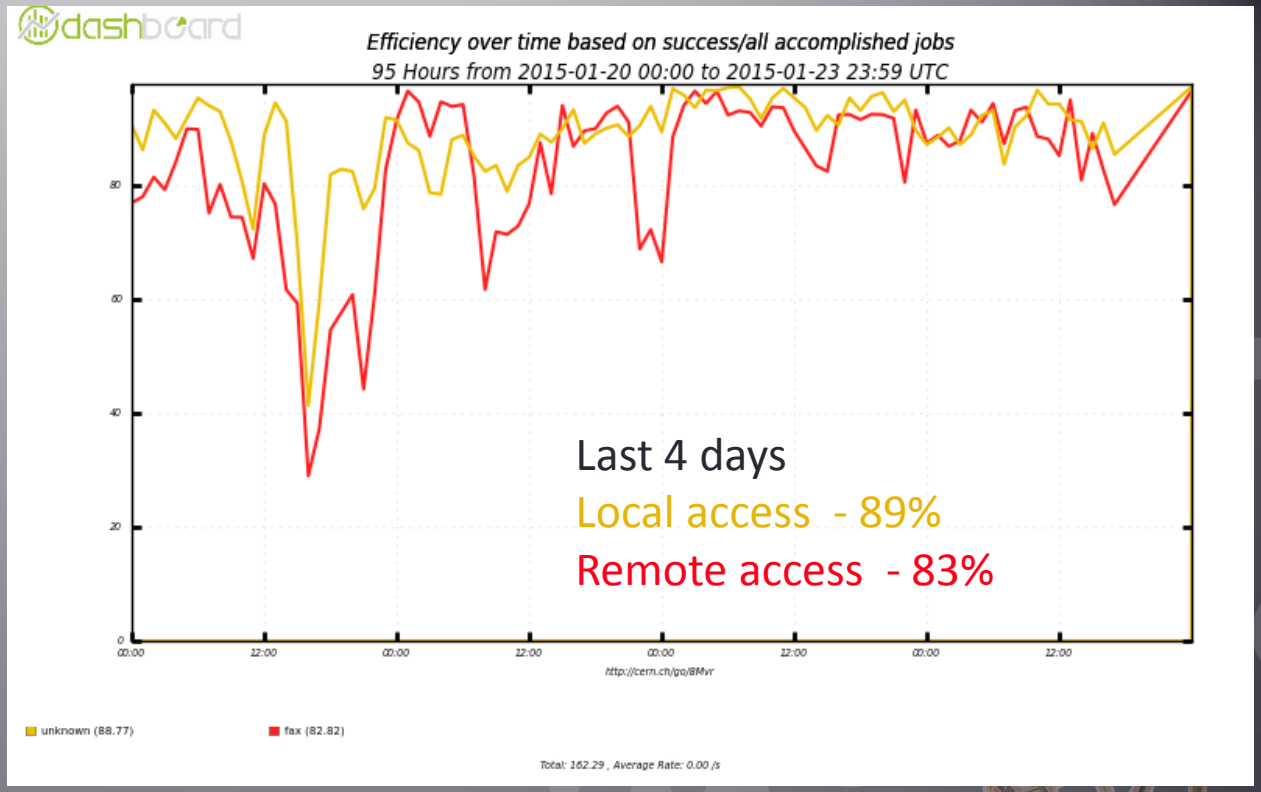
All analysis Queues

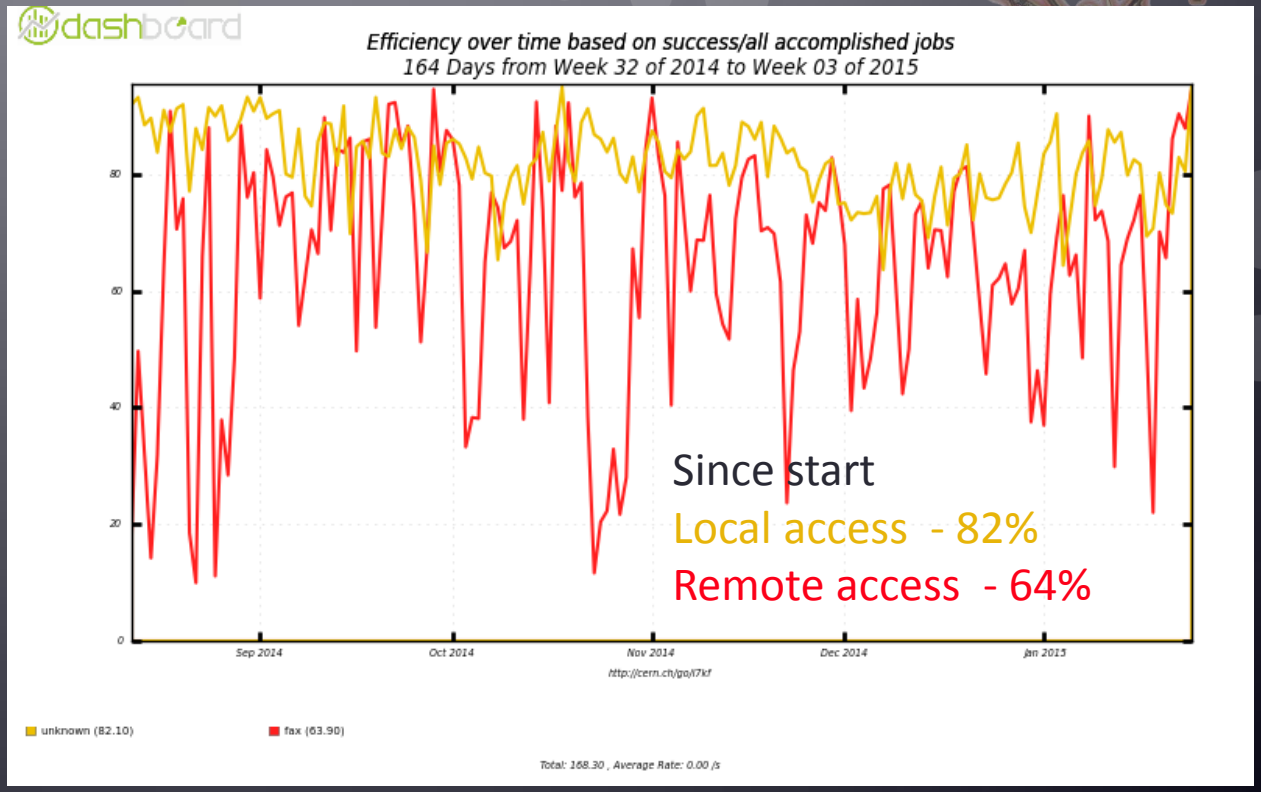




Job efficiency

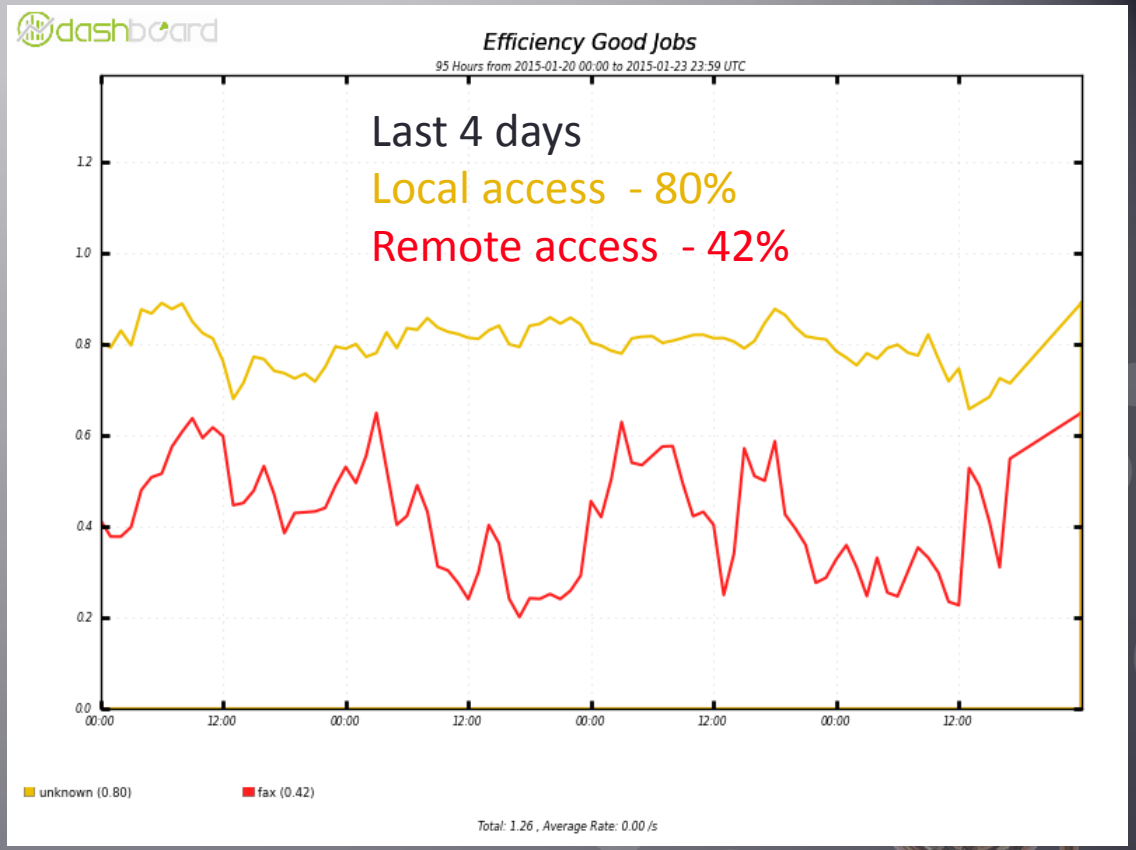
- We were and still are debugging the system and that causes lower job efficiency
- We will start blacklisting sites accepting/delivering overflow jobs when their FAX endpoint fails the tests.
- We are confident that error rate will be at the level of jobs accessing data locally.

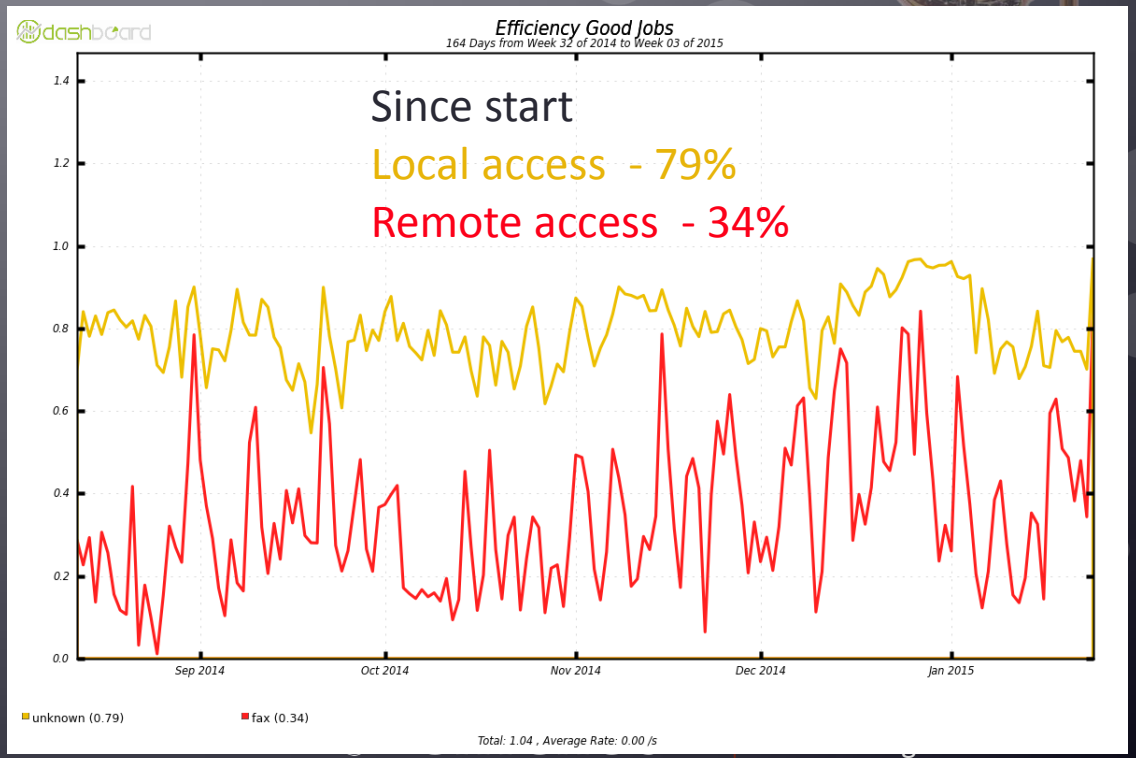




CPU efficiency

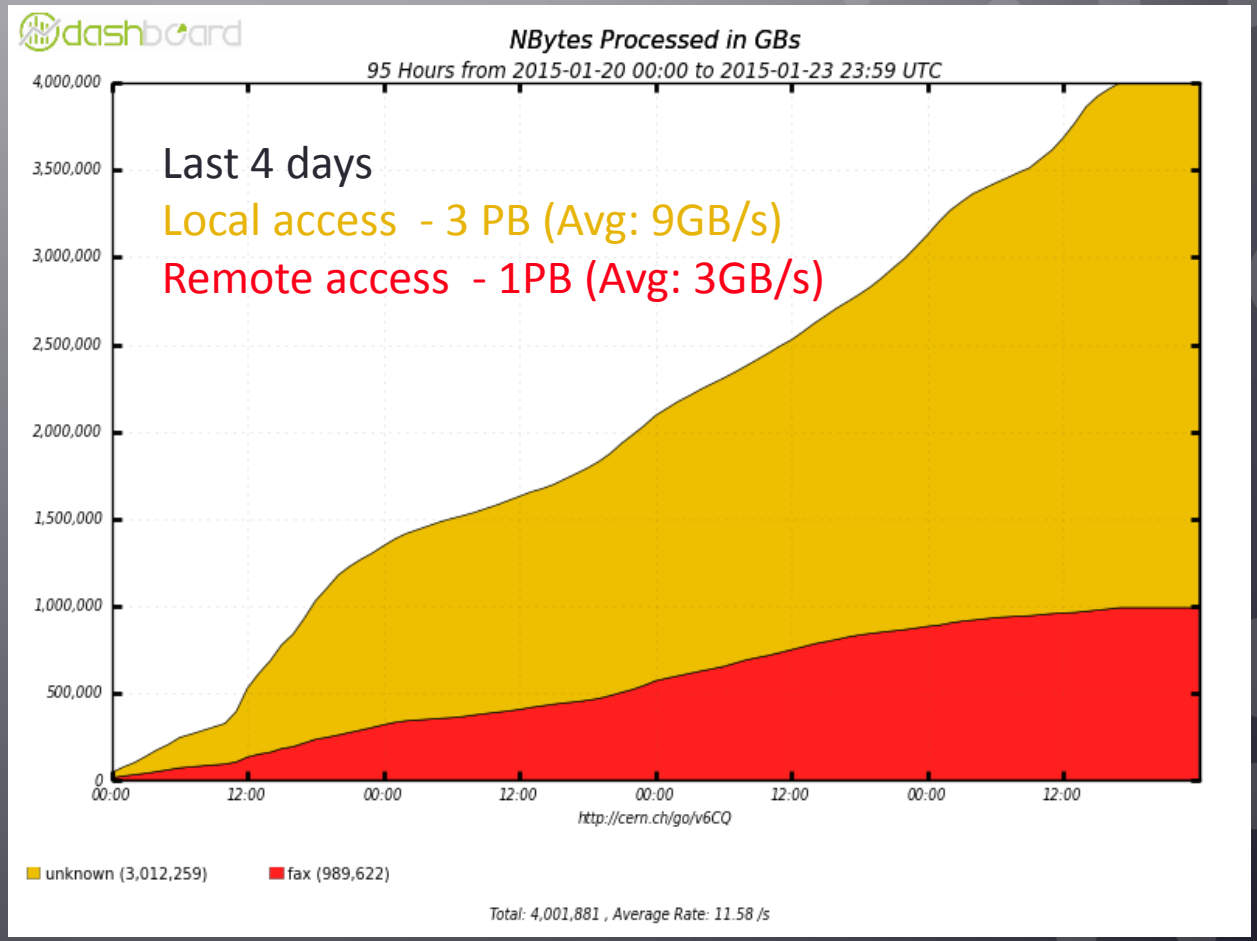
- Completely depends on the job mix.
- People still not consistently using TTreeCache
- Version of ROOT auto-enabling TTC still not in wide use





FAX - scaling

- We are confident it can scale to all the ATLAS queues
- A few important optimizations still in queue:
  - Access data from the optimal place
  - Reducing number of space tokens where N2N looks for the file

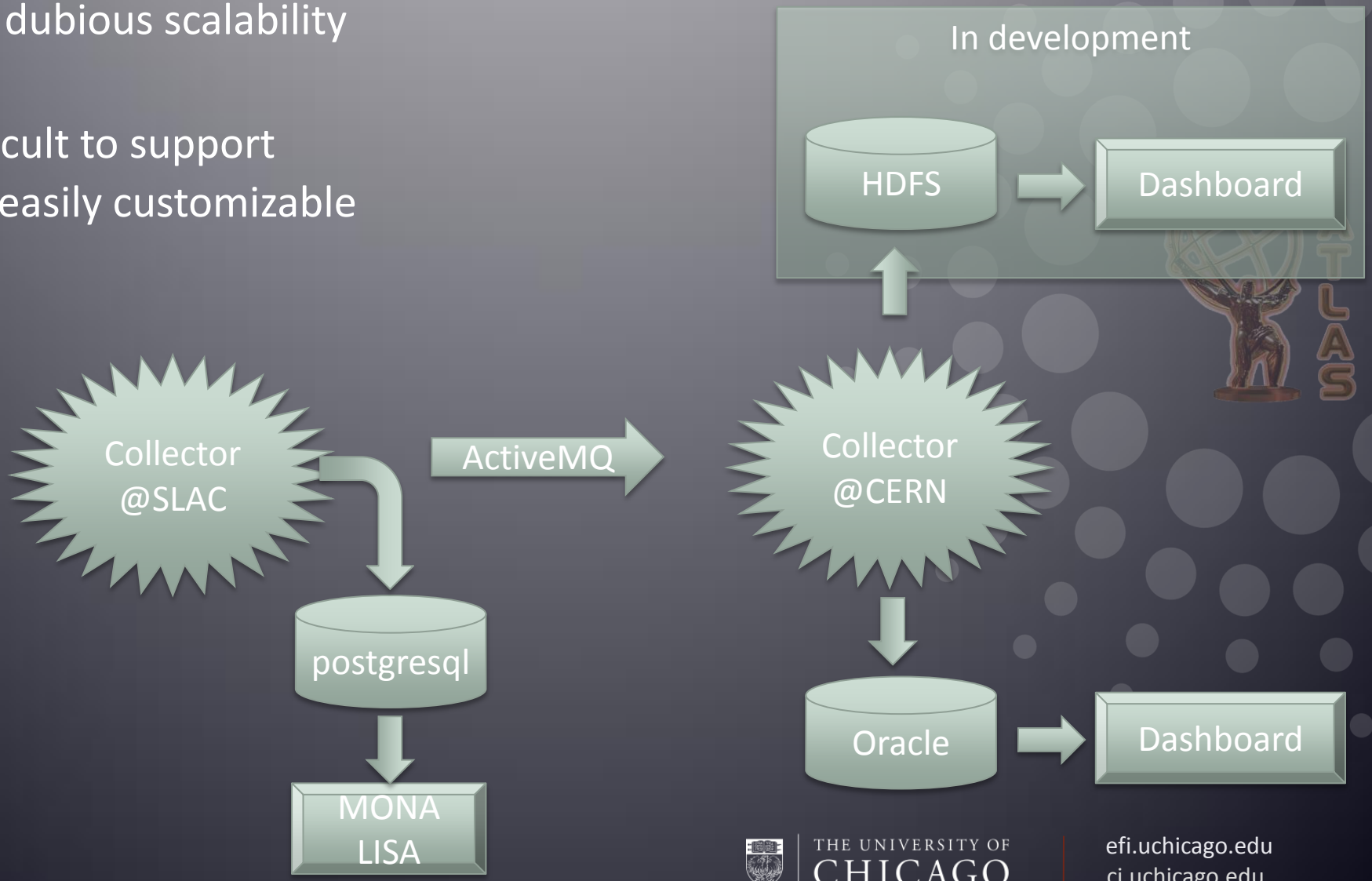


Monitoring

- Network has to be managed the same way we manage CPU and storage usage.
- And for that we need accounting.
- That's the part that failed us the most
  - Sites don't really cooperate in enabling it
    - Even when obliged by law
    - We did not make it easy / clear how to do it
  - We don't trust currently shown ML info
    - Instances where ganglia shows 3GB/s and ML 100MB/s
    - Huge differences between summary and detailed monitoring plots.

Current Monitoring Chain

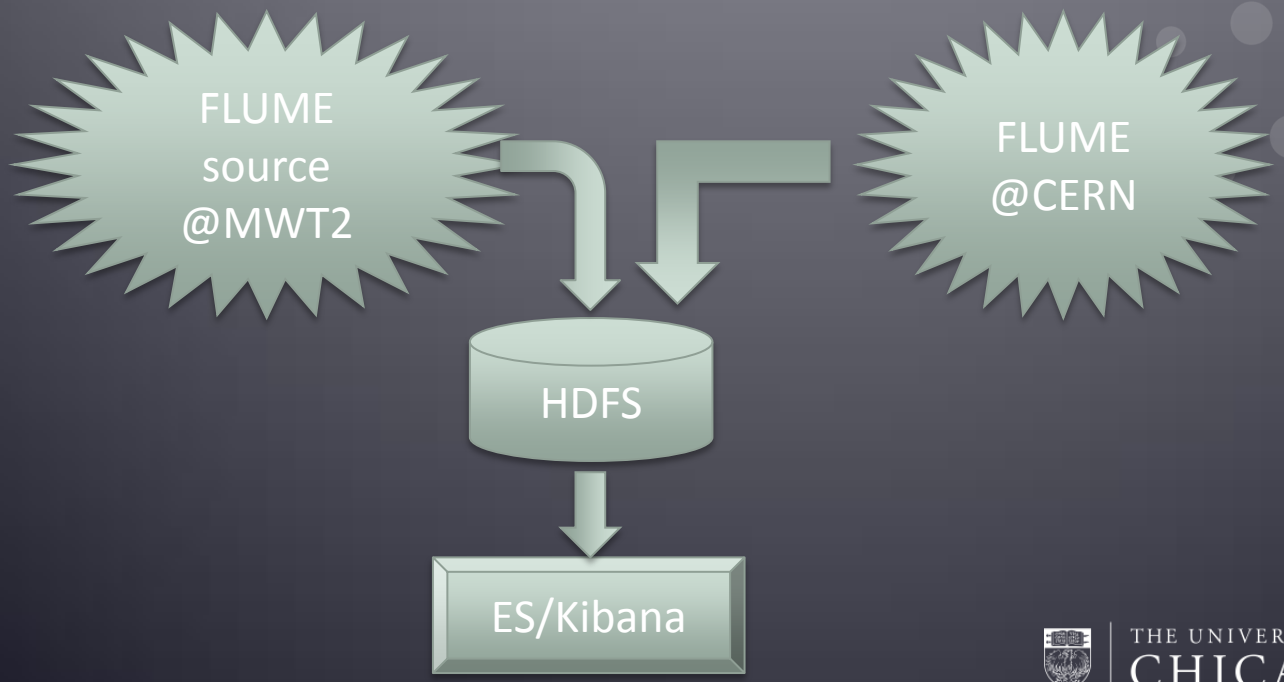
- Collector at SLAC
  - a complicated custom made construction
  - quite difficult to support
  - of a dubious scalability
- ML
  - Difficult to support
  - not easily customizable



Proposal for a new Monitoring Chain

- Collectors
  - FLUME source(s)
    - directly accept UDP messages
  - FLUME sink
    - Aggregates and outputs straight into HDFS
- Cleaning, re-summing
  - PIG running each 10 min.
  - Writing back into HDFS
- Dashboard
  - ElasticSearch indexes data in HDFS
  - Kibana to visualize

Scalable
Industry standard tools
Very customizable dashboard
Easy to do detailed analytics



Federation requirements

- Stability
- Stability
- Stability
- Monitoring
- User friendliness

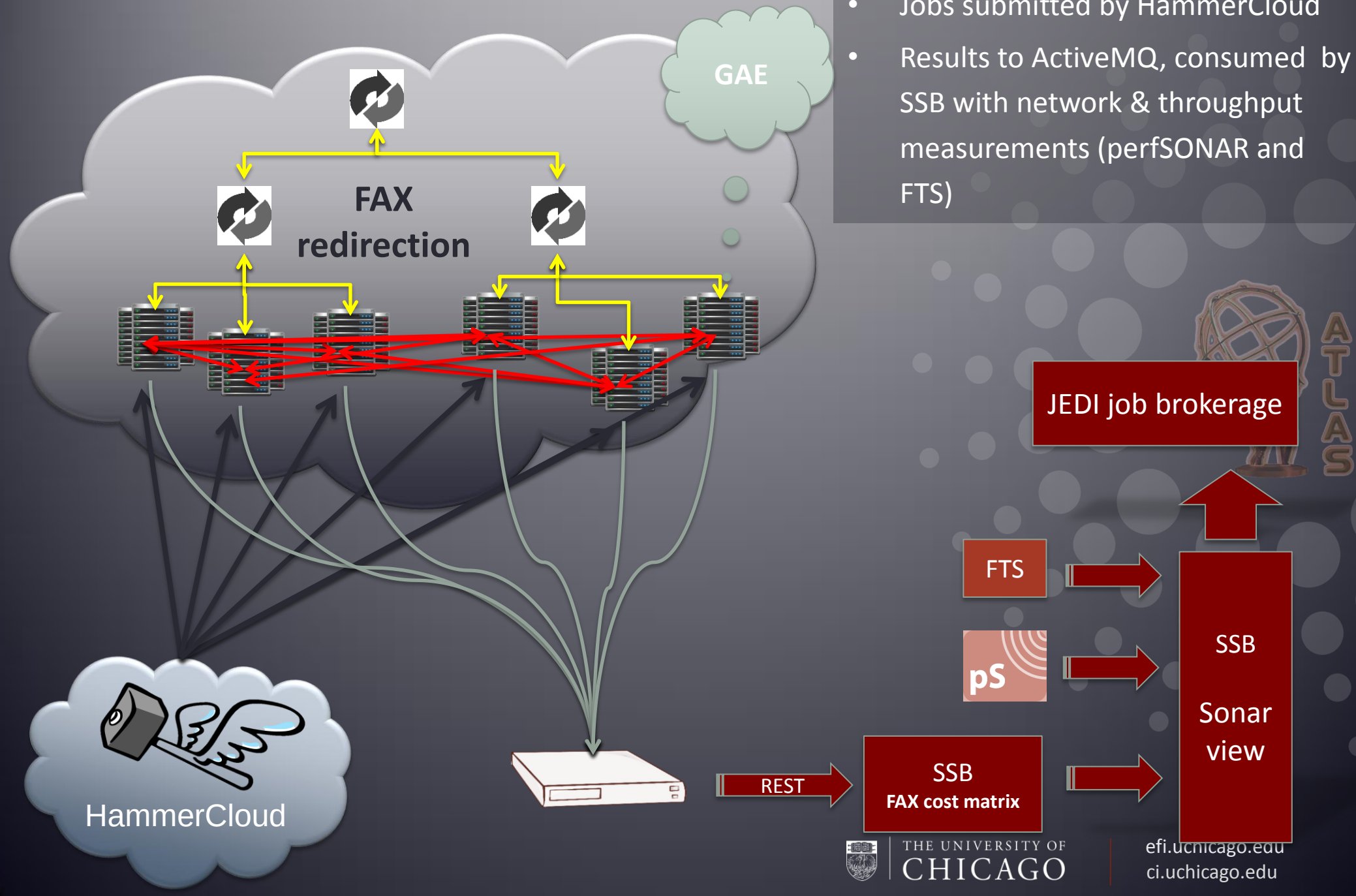
Reserve

Actions on solving protocol zoo

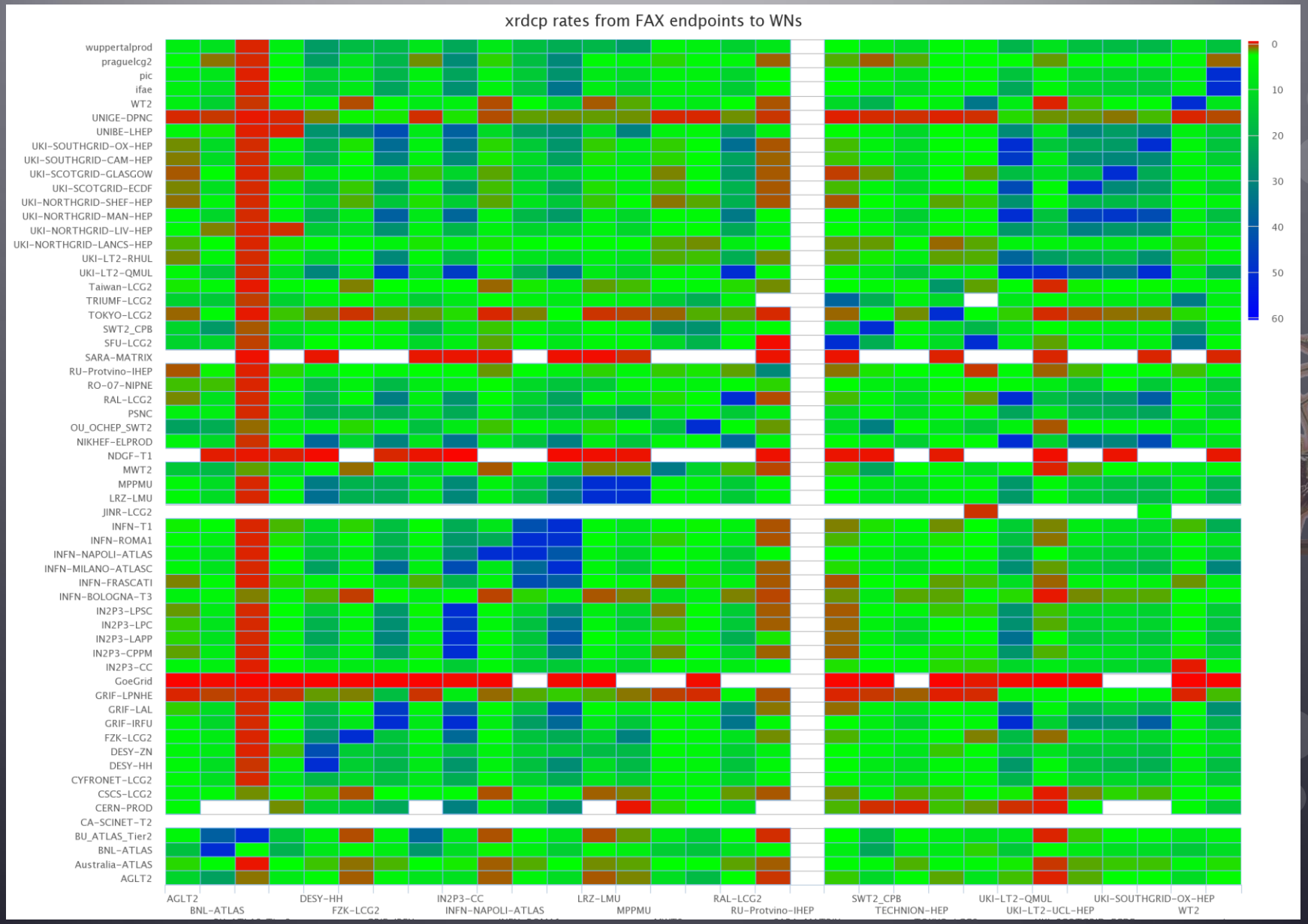
1. Short Term (?) : Eliminate the need of space tokens (rely on paths) for accounting.
2. Short Term: Move to gridFTP only for 3rd party transfer (requires 1)
3. Short Term: Move to xrootd/http for uploads and downloads (requires 1)
4. Short/Medium Term: commission http/WebDAV to production quality for deletions
5. Medium Term: Decommission SRMfrom non tape sites (requires all the above)
6. Short Term: Move to xrootd (or file) all the directIO
7. Short Term: Decommission other directIO protocols (requires 6)
8. Medium/Long Term: Commission xrootd and WebDAV for 3rd party transfers
9. Medium/Long Term: Decommission gridFTP (requires 8)
10. Long Term: Consolidate Davix and evaluate Davix and xrootd for directIO
11. Long Term: Keep both webDAV and xrootd or decommission one

FAX cost matrix

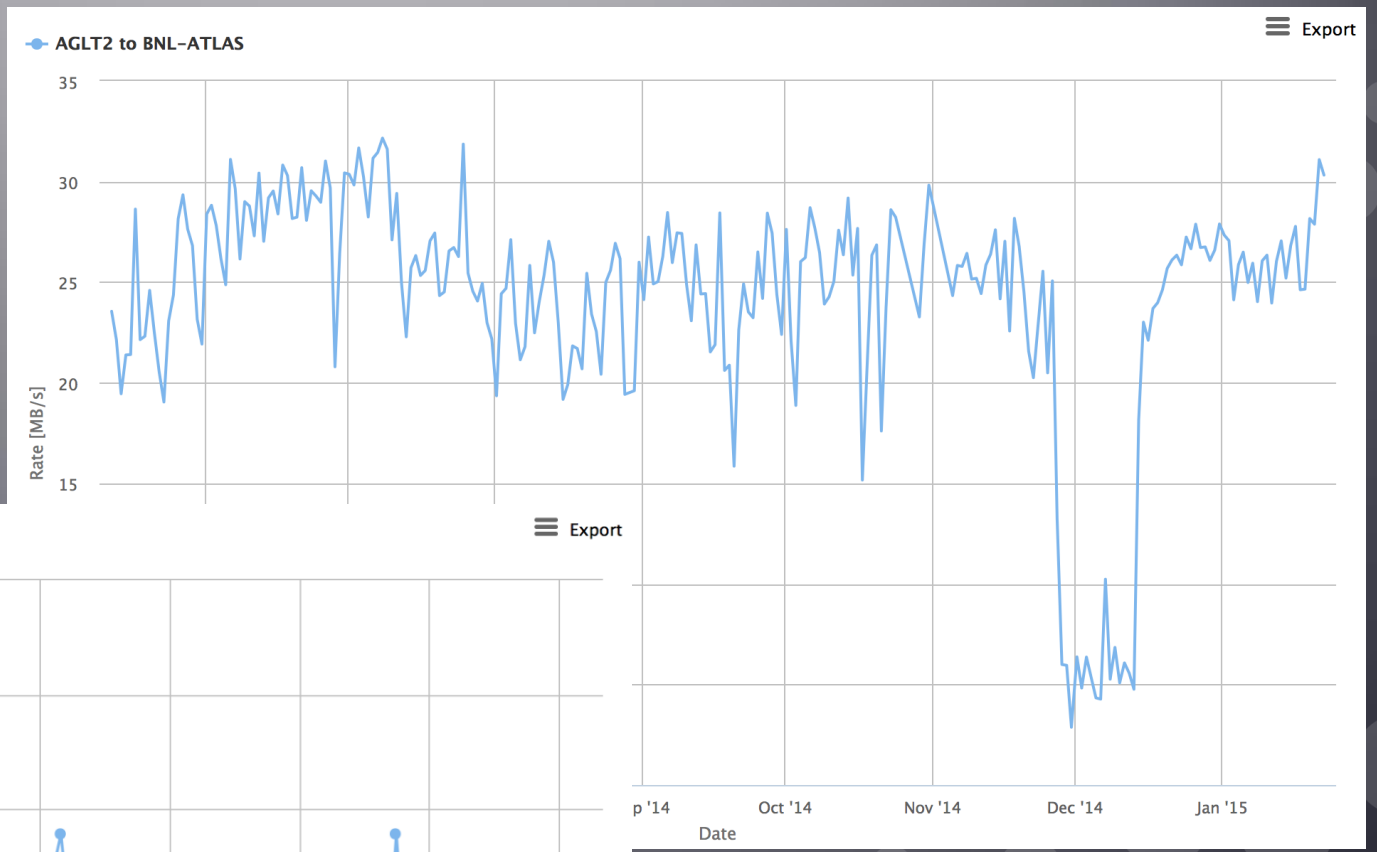
- Data collected between 20 ANALY queues (compute sites) and 58 FAX endpoints
- Jobs submitted by HammerCloud
- Results to ActiveMQ, consumed by SSB with network & throughput measurements (perfSONAR and FTS)

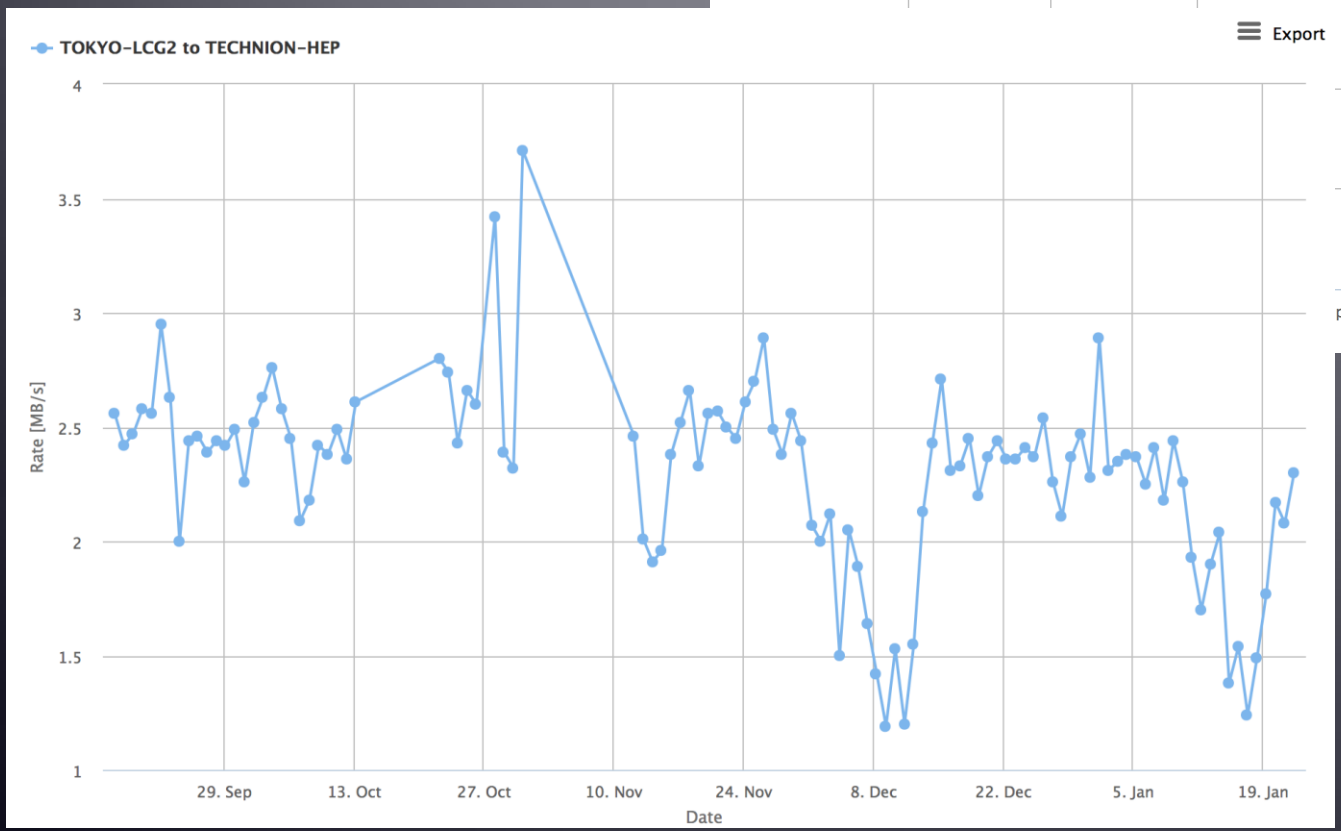


Cost matrix - results



Cost matrix - results





GAE based dashboard
http://waniotest.appspot.com/