Computation Institute

Federating ATLAS storage using XrootD (FAX)

Rob Gardner
on behalf of the atlas-adc-federated-xrootd working group

Computation and Enrico Fermi Institutes
University of Chicago

OSG All Hands Meeting
March 11, 2013

Data federation goals

- Create a common ATLAS namespace across all storage sites, accessible from anywhere
- Make easy to use, homogeneous access to data
- Identified initial use cases
  - Failover from stage-in problems with local storage
    - 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

Implications for Production & Analysis

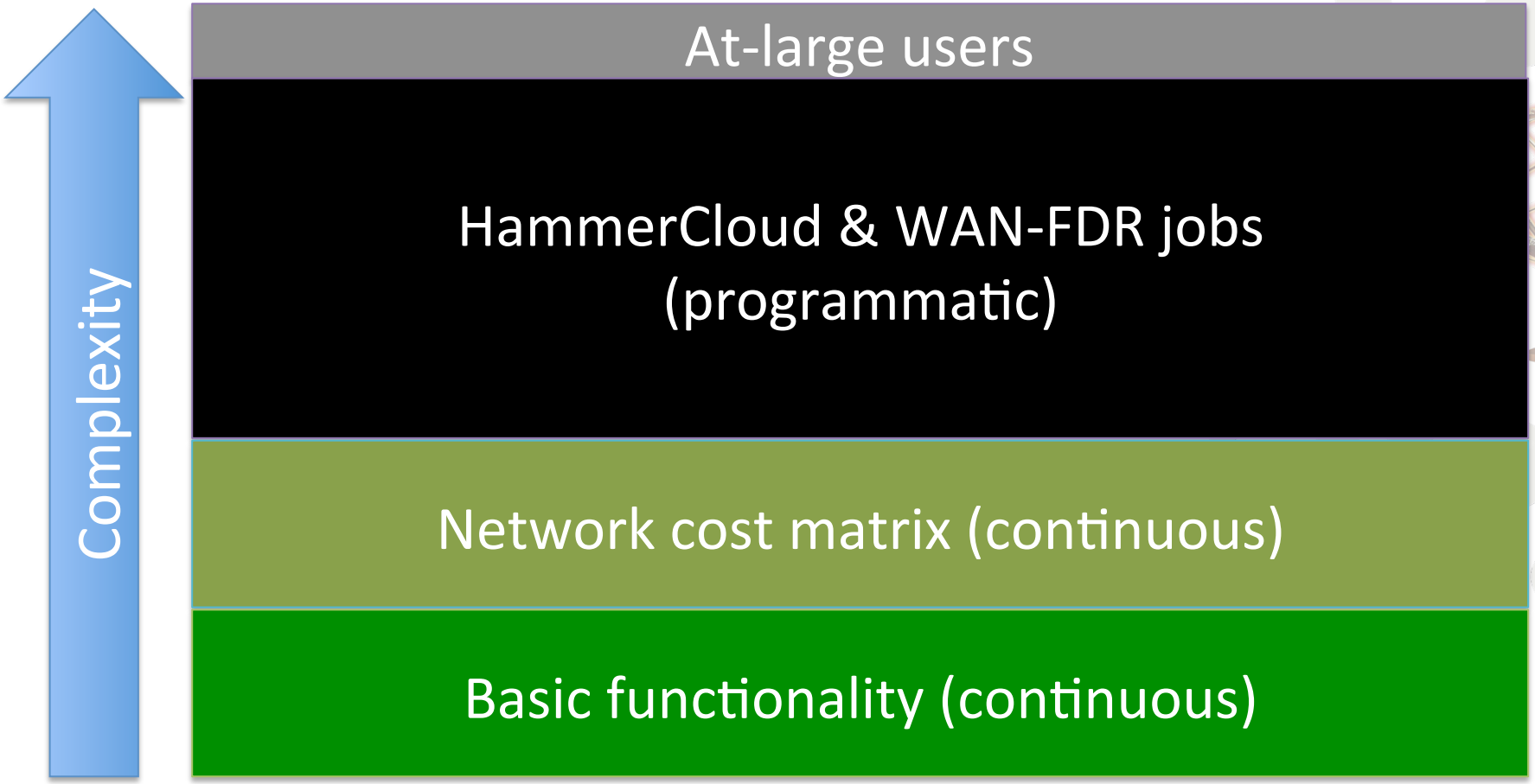
- Behind the scenes in the Panda + Pilot systems:
  - Recover from stage-in to local disk failures
  - This is in production at a few sites
- Development coming to allow advanced brokering which includes network performance
  - Would mean jobs no longer require dataset to be complete at a site
  - Access "diskless" compute sites
- Ability to use non-WLCG resources
  - "Off-grid" analysis clusters
  - Opportunistic resources
  - Cloud resources

Site Metrics

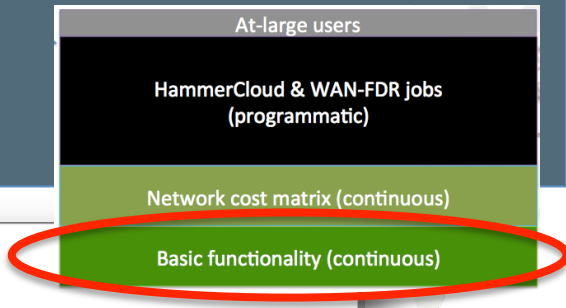
- "Connectivity" – copy and read test matrices
  - Snapshots per site as sever
- HC runs with modest job numbers
  - Stage-in & direct read
  - Local, nearby, far-away
- HC metrics
  - Simple job efficiency
  - Wallclock, # files, CPU %, event rate,
- Load tests
  - For well functioning sites only
  - Graduated tests 50, 100, 200 jobs vs. various # files
  - Will notify the site and/or list when these are launched

Testing elements

- Starting week of January 21, we've been following a bottoms-up approach which builds stability in lower layers
- Will review progress in each of the layers in this interim report



Probes, integrated with AGIS



ATLAS Federated Xrootd Status - 2013-02-25 21:03:58

Frequently Asked Questions

Host: atl-prod09.slac.stanford.edu (atl-prod09.slac.stanford.edu)

Metric	Last Executed	Enabled?	Next Run Time	Status
org.usatlas.xrootd.grid-xrdcp-direct	2013-02-25 20:50:02 CST	YES	2013-02-25 21:05:00 CST	OK
org.usatlas.xrootd.grid-xrdcp-fax	2013-02-25 20:50:02 CST	YES	2013-02-25 21:05:00 CST	OK

Host: atlas-cm4.bu.edu (atlas-cm4.bu.edu)

Metric	Last Executed	Enabled?	Next Run Time	Status
org.usatlas.xrootd.grid-xrdcp-direct	2013-02-25 20:50:04 CST	YES	2013-02-25 21:05:00 CST	OK
org.usatlas.xrootd.grid-xrdcp-fax	2013-02-25 20:50:04 CST	YES	2013-02-25 21:05:00 CST	OK

Host: atlas-xrd-eos-n2n.cern.ch (atlas-xrd-eos-n2n.cern.ch)

Metric	Last Executed	Enabled?	Next Run Time	Status
org.usatlas.xrootd.grid-xrdcp-direct	2013-02-25 20:50:04 CST	YES	2013-02-25 21:05:00 CST	OK
org.usatlas.xrootd.grid-xrdcp-fax	2013-02-25 20:50:04 CST	YES	2013-02-25 21:05:00 CST	OK

Host: atlasse.lnf.infn.it (atlasse.lnf.infn.it)

Metric	Last Executed	Enabled?	Next Run Time	Status
org.usatlas.xrootd.grid-xrdcp-direct	2013-02-25 20:50:04 CST	YES	2013-02-25 21:05:00 CST	OK
org.usatlas.xrootd.grid-xrdcp-fax	2013-02-25 20:50:04 CST	YES	2013-02-25 21:05:00 CST	OK

Host: bohr3226.tier2.hep.manchester.ac.uk (bohr3226.tier2.hep.manchester.ac.uk)

Metric	Last Executed	Enabled?	Next Run Time	Status
org.usatlas.xrootd.grid-xrdcp-direct	2013-02-25 20:50:05 CST	YES	2013-02-25 21:05:00 CST	OK
org.usatlas.xrootd.grid-xrdcp-fax	2013-02-25 20:50:02 CST	YES	2013-02-25 21:05:00 CST	CRITICAL

Host: dcache-door-atlas18.desy.de (dcache-door-atlas18.desy.de)

Metric	Last Executed	Enabled?	Next Run Time	Status
org.usatlas.xrootd.grid-xrdcp-direct	2013-02-25 20:50:03 CST	YES	2013-02-25 21:05:00 CST	OK
org.usatlas.xrootd.grid-xrdcp-fax	2013-02-25 20:50:03 CST	YES	2013-02-25 21:05:00 CST	OK

Host: dcdoor11.usatlas.bnl.gov (dcdoor11.usatlas.bnl.gov)

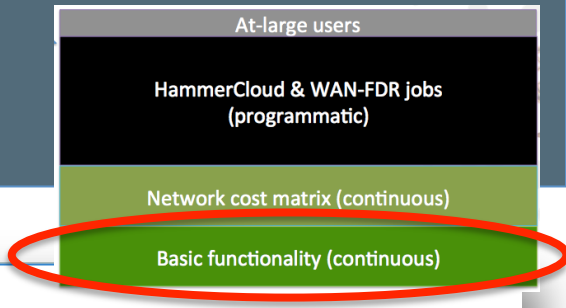
Metric	Last Executed	Enabled?	Next Run Time	Status
org.usatlas.xrootd.grid-xrdcp-direct	2013-02-25 20:50:02 CST	YES	2013-02-25 21:05:00 CST	OK
org.usatlas.xrootd.grid-xrdcp-fax	2013-02-25 20:50:04 CST	YES	2013-02-25 21:05:00 CST	OK

Host: fal-pygrid-30.lancs.ac.uk (fal-pygrid-30.lancs.ac.uk)

Metric	Last Executed	Enabled?
org.usatlas.xrootd.grid-xrdcp-direct	2013-02-25 20:50:04 CST	YES
org.usatlas.xrootd.grid-xrdcp-fax	2013-02-25 20:50:03 CST	YES

Direct xrdcp copy of test files
Copy using regional redirector

Basic redirection functionality



- Direct access from clients to sites
- Redirection to non-local data ("upstream")
- Redirection from central redirectors to the site ("downstream")

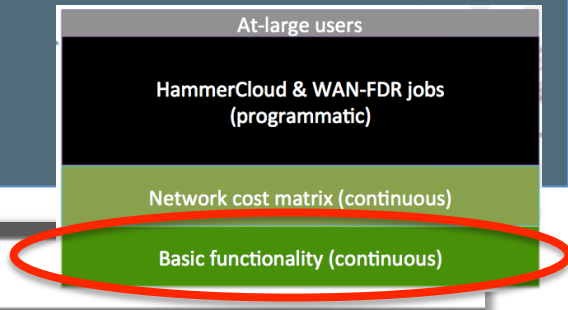
dashboard | Index | Expanded Table

Show 200 entries | Copy | Print | Save | view: FAXMON

Site Name	FAXMON_1	FAXMON_2	FAXMON_3
AGLT2	OK	OK	OK
BNL-ATLAS	OK	OK	OK
BU_ATLAS_TIER2	OK	OK	OK
CERN-PROD	OK	OK	OK
DESY-HH	OK	OK	OK
INFN-FRASCATI	OK	OK	OK
INFN-NAPOLI-ATLAS	OK	OK	OK
INFN-ROMA1	OK	OK	OK
JINR-LCG2	OK	OK	OK
LRZ-LMU	OK	OK	OK
MPPMU	OK	OK	OK
MWT2	OK	OK	OK
OU_OCHEP_SWT2	OK	NoUpstreamRedirection	OK
PRAGUELCG2	OK	OK	OK
RAL-LCG2	OK	OK	OK
RU-PROTVINO-IHEP	OK	OK	OK
SWT2_CPB	OK	OK	OK
UKI-LT2-QMUL	OK	OK	OK
UKI-NORTHGRID-LANCS-HEP	OK	OK	NoFirstLevelRedirection
UKI-NORTHGRID-LIV-HEP	OK	OK	OK
UKI-SCOTGRID-ECDF	OK	OK	OK
UKI-SCOTGRID-GLASGOW	OK	OK	OK
UKI-SOUTHGRID-OX-HEP	OK	OK	OK
WT2	OK	OK	OK
WUPPERTALPROD	noDirect	NoUpstreamRedirection	NoFirstLevelRedirection

Uses a host at CERN which runs set of probes against sites

Redirectors - regional and global



atlas-agis.cern.ch/agis/service/table_view/

Information System

Service monitor

DDMEndpoint | PANDA Queue | Service | DDM Groups

entries First Previous 1 Next Last

Name | Service Type | Site | Related ATLAS Sites | Endpoint | state | status

Redirector

	Service Type	Site	Endpoint
-xrd-de	Redirector/XROOTD	CERN-PROD	atlas-xrd-de.cern.ch:1094
-xrd-it	Redirector/XROOTD	CERN-PROD	atlas-xrd-it.cern.ch:1094
-xrd-ru	Redirector/XROOTD	CERN-PROD	atlas-xrd-ru.cern.ch:1094
-xrd-uk	Redirector/XROOTD	CERN-PROD	atlas-xrd-uk.cern.ch:1094
central	Redirector/XROOTD	MWT2	xrd-central.usatlasfacility.org:1
-xrd-eu	Redirector/XROOTD	CERN-PROD	atlas-xrd-eu.cern.ch:1094

of 6 entries

XRDFEDATLAS

availability: (more)

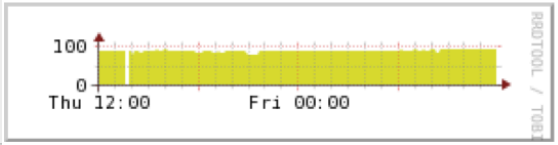
percentage: 88%

status: affected

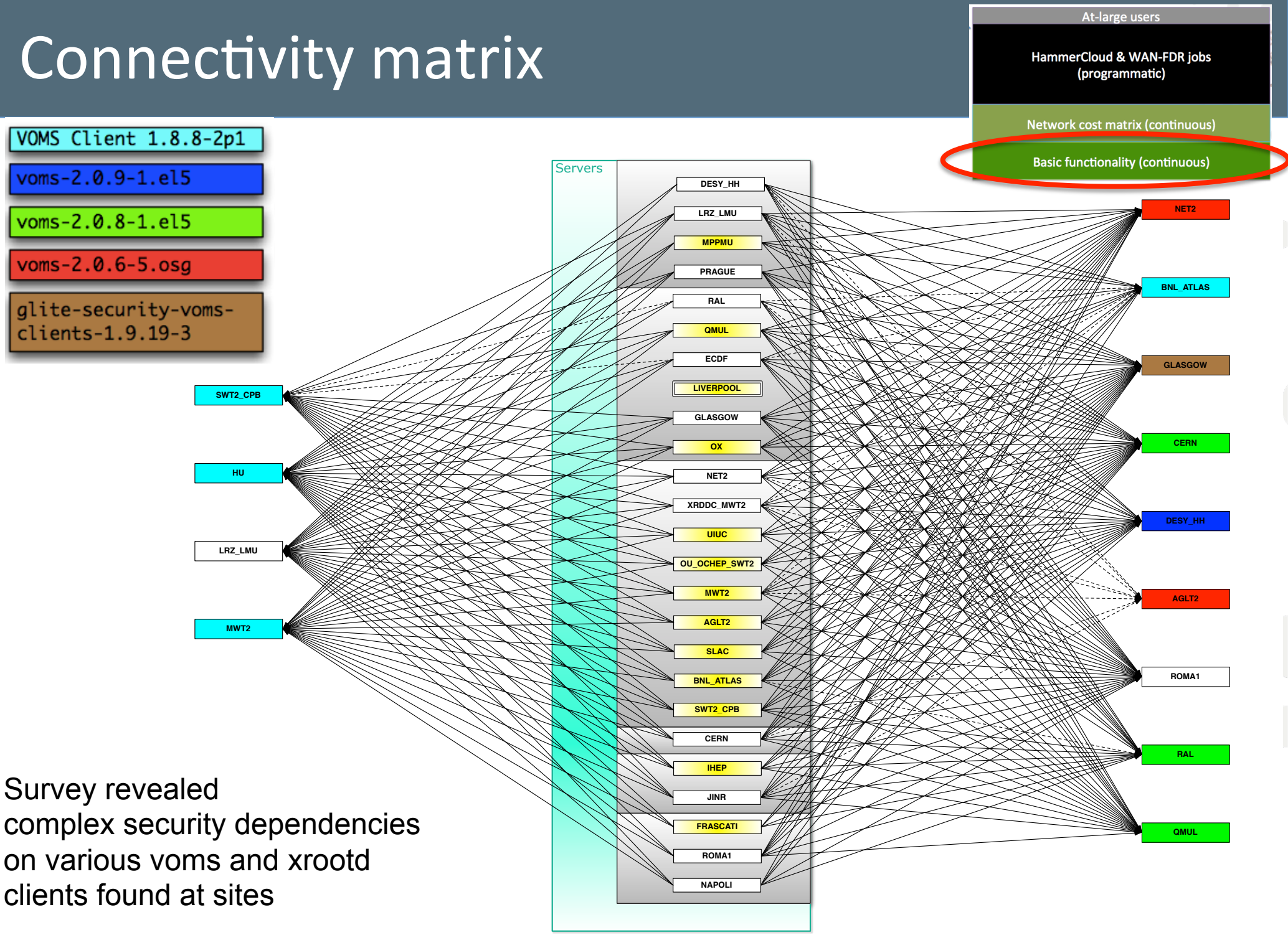
this service consists of:

- XRDFED_ATLAS-EU
- XRDFED_ATLAS-EOS-N2N
- XRDFED_ATLAS-DE
- XRDFED_ATLAS-IT
- XRDFED_ATLAS-ES
- XRDFED_ATLAS-FR
- XRDFED_ATLAS-ASIA
- XRDFED_ATLAS-RU
- XRDFED_ATLAS-UK

availability in the last 24 hours (more):

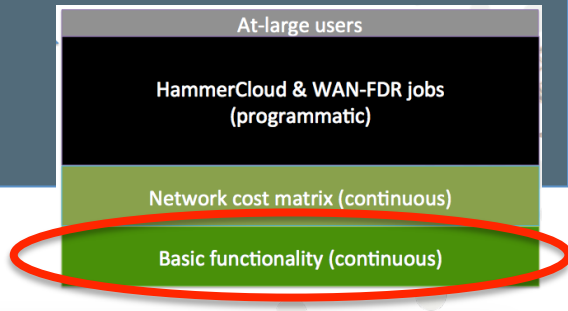


Connectivity matrix



Survey revealed complex security dependencies on various voms and xrootd clients found at sites

Data federated (1)



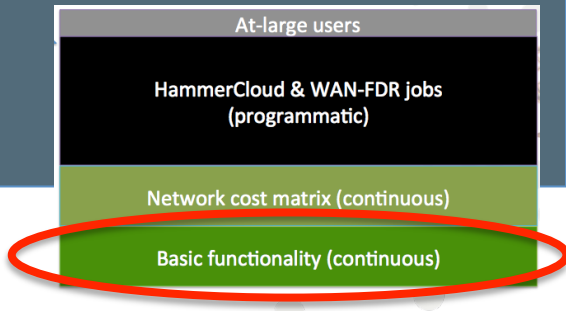
Top 100 sites used by ATLAS (bold=FAX accessible)

Values for 2013-02-20

Sites	No. of Dataset Replicas	No. of Physical Files	No. of Physical Bytes(in TBs)
CERN-PROD	**530796**	**47518062**	**32420.099** *
BNL-ATLAS	**844552**	**61226021**	**15519.543** *
IN2P3-CC	231311	28094169	7901.322
FZK-LCG2	**182141**	**28722971**	**7026.415**
RAL-LCG2	**173389**	**22913769**	**6942.899**
SARA-MATRIX	126548	19907328	5242.638
TRIUMF-LCG2	124095	14451310	5208.316
INFN-T1	124404	13198843	4288.349
NDGF-T1	121453	14446185	3834.052
TAIWAN-LCG2	75104	10971537	3624.539
PIC	88313	10478026	3347.56
AGLT2	**195943**	**8504294**	**2653.895**
MWT2	**138562**	**10568581**	**2345.397**
WT2	**146915**	**7413448**	**2007.233**
DESY-HH	**69208**	**4087641**	**1983.852**
UKI-LT2-QMUL	**36077**	**2181772**	**1397.756**
TOKYO-LCG2	80375	4559482	1338.497
DESY-ZN	41199	2662858	1210.058
PRAGUELCG2	**24237**	**1477373**	**1065.529**
BU_ATLAS_TIER2	**57261**	**2245432**	**1040.986**
IFIC-LCG2	30123	1741242	1008.129
SWT2_CPB	**92554**	**3185476**	**991.167**
NIKHEF-ELPROD	29547	1139314	983.224
INFN-NAPOLI-ATLAS	**38131**	**1668688**	**924.815**
UKI-SCOTGRID-GLASGOW	**27603**	**1346190**	**919.348**
UNI-FREIBURG	24430	1228722	865.058
INFN-ROMA1	**67192**	**2329638**	**841.959**
UKI-NORTHGRID-LANCS-HEP	**19693**	**1270750**	**825.473**
GRIF-IRFU	**21820**	**950667**	**824.421**

* Includes tape, which we do not federate

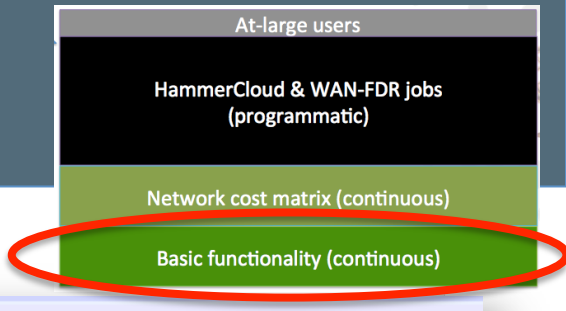
Data federated (2)



Top 100 sites used by ATLAS (bold=FAX accessible)

GRIF-IRFU	**21820**	**950667**	**824.421**
INFN-MILANO-ATLASC	37308	1363614	777.098
GOEGRID	40849	1574264	766.624
WUPPERTALPROD	**30922**	**1566320**	**764.091**
LRZ-LMU	**45780**	**1786397**	**745.085**
UKI-NORTHGRID-MAN-HEP	**24592**	**1061597**	**696.129**
MPPMU	**24407**	**2657478**	**684.402**
GRIF-LPNHE	**23191**	**1309095**	**626.494**
CA-MCGILL-CLUMEQ-T2	14968	935888	616.643
CA-VICTORIA-WESTGRID-T2	20682	870217	586.662
IFAE	28452	975357	555.763
CA-SCINET-T2	17299	732496	552.452
AUSTRALIA-ATLAS	67522	1385186	547.474
CSCS-LCG2	13409	671856	534.262
UKI-SOUTHGRID-OX-HEP	**29579**	**1275013**	**520.352**
UKI-LT2-RHUL	18400	712343	503.246
IN2P3-LAPP	**30061**	**1016122**	**497.957**
IN2P3-LPC	10695	483930	475.307
UAM-LCG2	20545	750274	465.196
SFU-LCG2	22157	982551	463.28
IN2P3-CPPM	16373	1179026	443.637
JINR-LCG2	**14730**	**584517**	**424.687**
IN2P3-LPSC	41437	840272	384.559
GRIF-LAL	**18663**	**597276**	**371.101**
RU-PROTVINO-IHEP	**19201**	**900094**	**359.037**
UKI-NORTHGRID-LIV-HEP	**35665**	**1054536**	**346.976**
CYFRONET-LCG2	8748	592935	311.13
INFN-FRASCATI	**9093**	**424917**	**308.104**
UKI-SCOTGRID-ECDF	**8952**	**583241**	**293.909**

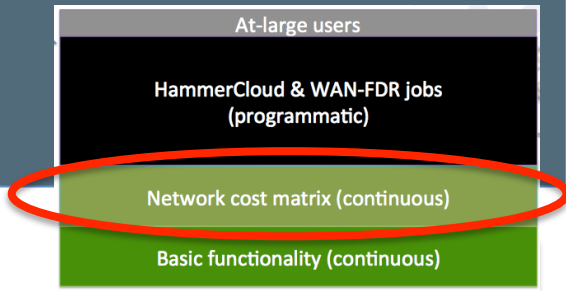
Data federated (3)



Top 100 sites used by ATLAS (bold=FAX accessible)

TR-10-ULAKBIM	9788	504053	278.02
RO-07-NIPNE	6942	466744	277.784
BEIJING-LCG2	5275	276359	265.556
RRC-KI	6244	292816	258.819
UKI-SOUTHGRID-RALPP	8099	484516	256.433
UKI-SOUTHGRID-CAM-HEP	**6546**	**222005**	**206.996**
OU_OCHEP_SWT2	**9643**	**243745**	**196.336**
UKI-NORTHGRID-SHEF-HEP	7274	375315	196.08
WEIZMANN-LCG2	7391	350577	170.41
UKI-SOUTHGRID-BHAM-HEP	6042	240539	165.214
TECHNION-HEP	7557	627099	147.7
RO-02-NIPNE	9895	290792	141.984
IL-TAU-HEP	19804	443696	122.509
LIP-COIMBRA	11518	382359	104.828
UKI-LT2-UCL-HEP	7760	201189	102.026
HEPHY-UIBK	4030	135794	99.388
RU-PNPI	2600	109158	82.695
PSNC	8156	128760	50.304
NCG-INGRID-PT	3515	107992	47.107
CA-ALBERTA-WESTGRID-T2	5039	146383	43.163
SE-SNIC-T2	6327	816522	35.296
LIP-LISBON	222	33450	6.04
RO-14-ITIM	915	20577	5.613
RO-16-UAIC	918	19665	4.3
UTA_SWT2	792	11470	3.407
ITEP	1200	30197	3.227
RU-MOSCOW-FIAN-LCG2	1098	10629	3.039
FMPHI-UNIBA	163	4524	2.992
UKI-LT2-IC-HEP	1009	10202	1.726
UKI-LT2-BRUNEL	903	5630	0.641
UKI-SCOTGRID-DURHAM	776	6180	0.56
RRC-KI-T1	710	3171	0.496
IEPSAS-KOSICE	165	1048	0.267

Cost matrix measurements



Requires sites to install an XRootD server

Redirectors for each region ("cloud")

Currently **32** sites, including EOS

Redirection network touches **six** clouds (DE, FR, IT, RU, UK, US) plus CERN

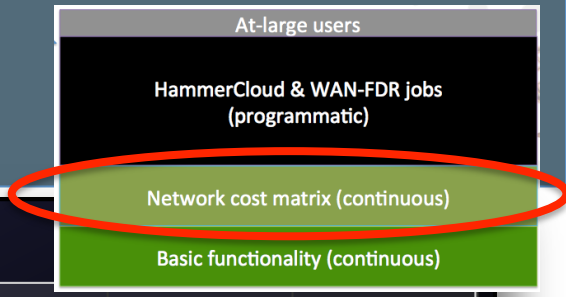
Redirectors ready for ES and Asia regions

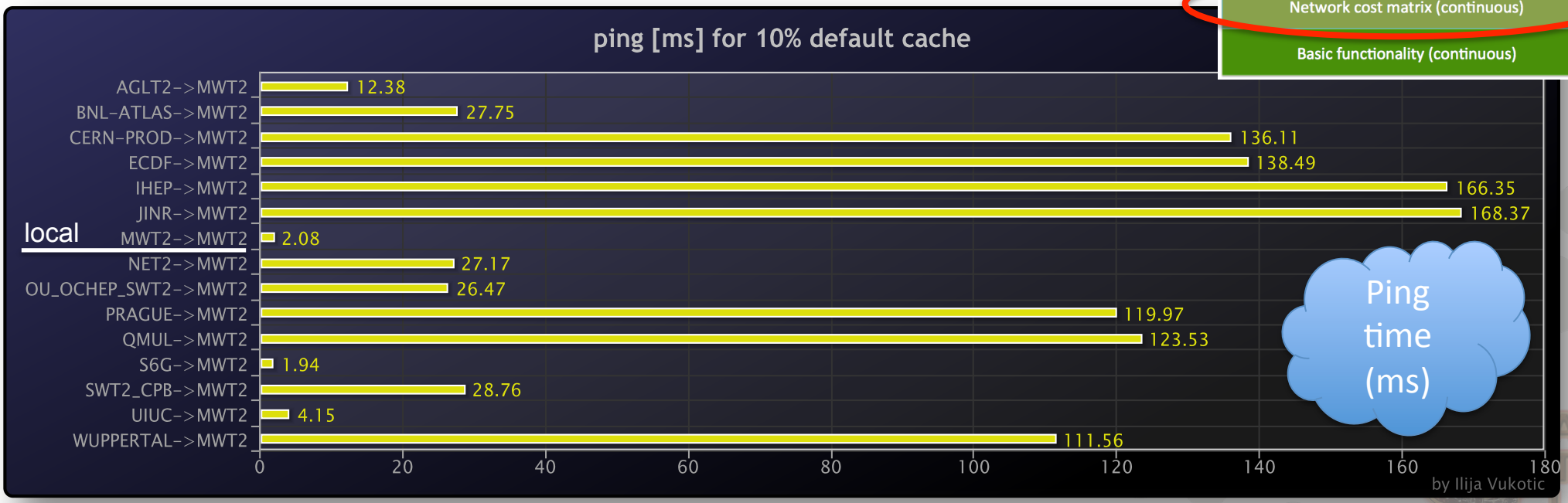
Show 200 entries | Copy | Print | Save | view: Network

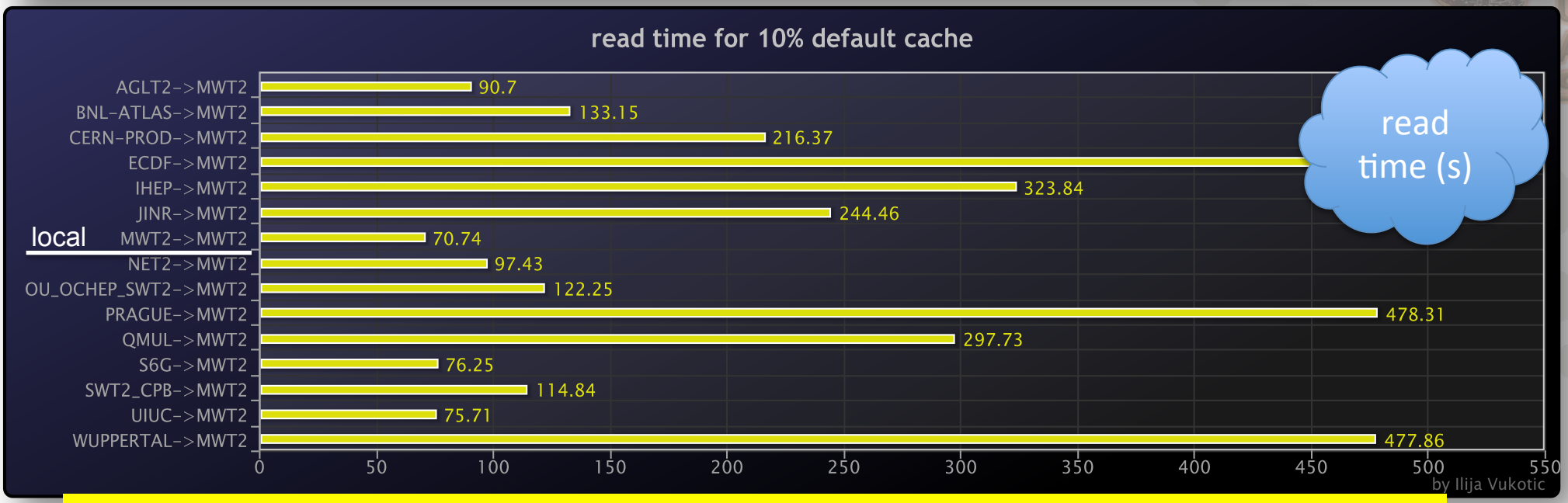
Site Name	Site Info: Source	Site Info: Destination	Network Measurements
ANALY_MWT2_to_ANALY_CERN_XROOTD	MWT2	CERN-PROD	2.074
ANALY_SWT2_CPB_to_ANALY_CERN_XROOTD	SWT2_CPB	CERN-PROD	2.079
ANALY_OU_OCHEP_SWT2_to_ANALY_CERN_XROOTD	OU_OCHEP_SWT2	CERN-PROD	2.083
ANALY_CERN_XROOTD_to_ANALY_AGLT2	CERN-PROD	AGLT2	6.143
ANALY_MWT2_to_ANALY_NET2	MWT2	BU_ATLAS_Tier2	6.6
ANALY_CERN_XROOTD_to_ANALY_NET2	CERN-PROD	BU_ATLAS_Tier2	6.794
ANALY_AGLT2_to_ANALY_CERN_XROOTD	AGLT2	CERN-PROD	7.769
ANALY_NET2_to_ANALY_NET2	BU_ATLAS_Tier2	BU_ATLAS_Tier2	7.982
ANALY_AGLT2_to_ANALY_NET2	AGLT2	BU_ATLAS_Tier2	8
ANALY_SWT2_CPB_to_ANALY_NET2	SWT2_CPB	BU_ATLAS_Tier2	8
ANALY_OU_OCHEP_SWT2_to_ANALY_NET2	OU_OCHEP_SWT2	BU_ATLAS_Tier2	8.25
ANALY_CERN_XROOTD_to_ANALY_CERN_XROOTD	CERN-PROD	CERN-PROD	10.513
ANALY_OU_OCHEP_SWT2_to_ANALY_OU_OCHEP_SWT2	OU_OCHEP_SWT2	OU_OCHEP_SWT2	13.631
ANALY_SWT2_CPB_to_ANALY_SWT2_CPB	SWT2_CPB	SWT2_CPB	14.296
ANALY_QMUL_to_ANALY_CERN_XROOTD	UKI-LT2-QMUL	CERN-PROD	34.761
ANALY_ILLINOISHEP_to_ANALY_AGLT2	IllinoisHEP	AGLT2	35.042
ANALY_ILLINOISHEP_to_ANALY_MWT2	IllinoisHEP	MWT2	35.046
ANALY_ILLINOISHEP_to_ANALY_SWT2_CPB	IllinoisHEP	SWT2_CPB	35.122
ANALY_wuppertalprod_to_ANALY_CERN_XROOTD	wuppertalprod	CERN-PROD	35.195
ANALY_QMUL_to_ANALY_NET2	UKI-LT2-QMUL	BU_ATLAS_Tier2	35.403
ANALY_ILLINOISHEP_to_ANALY_NET2	IllinoisHEP	BU_ATLAS_Tier2	35.545
ANALY_QMUL_to_ANALY_AGLT2	UKI-LT2-QMUL	AGLT2	36
ANALY_wuppertalprod_to_ANALY_AGLT2	wuppertalprod	AGLT2	36.068
ANALY_wuppertalprod_to_ANALY_SWT2_CPB	wuppertalprod	SWT2_CPB	36.102
ANALY_wuppertalprod_to_ANALY_MWT2	wuppertalprod	MWT2	36.139
ANALY_ILLINOISHEP_to_ANALY_CERN_XROOTD	IllinoisHEP	CERN-PROD	36.2
ANALY_wuppertalprod_to_ANALY_NET2	wuppertalprod	BU_ATLAS_Tier2	36.38

Cost-of-access: (pairwise network links, storage load, etc.)

Comparing local to wide area performance

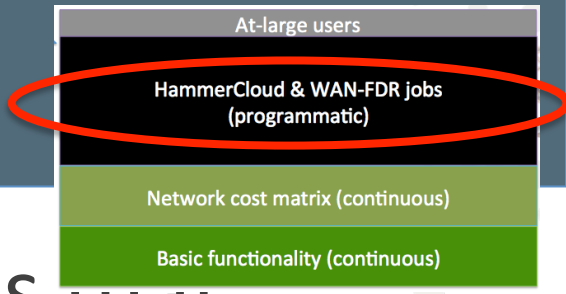






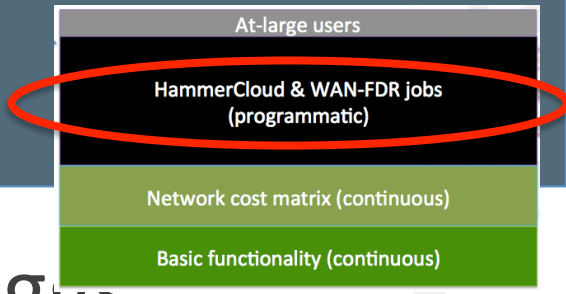
Each site can check its connectivity and IO performance for copy and direct read

Programmatic Hammer Cloud tests



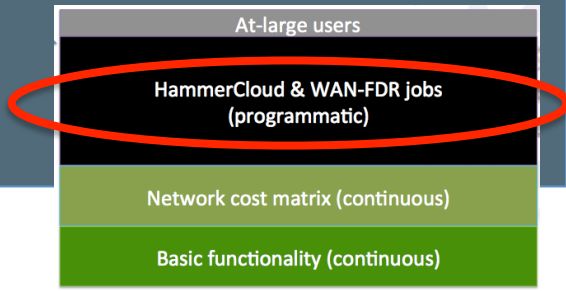
- Defined a set of Hammer Cloud tests that probe the infrastructure and which will collect measures of various data access patterns
- Setup by Johannes and Federica using Higgs → WW, and a SUSY D3PD analysis
  - 17.2.2 (Root 5.30) HWW analysis code which analyzes NTUP SMWZ
  - 17.6.0 (Root 5.34) HWW analysis code which analyzes NTUP SMWZ
  - 17.5.0 (Root 5.32) SUSY analysis code which analyzes NTUP SUSYSKIM (p1328, p1329)

Hammer Cloud testing



- Pre-placed, site-unique SUSY and Higgs datasets at all sites (see coverage next slide)
- Realistic, typical analysis templates for SUSY D3PD maker and Higgs analysis
- New pilot equipped for **stage-in** or **direct access** with XrootD
- Choose ANALY queue, and redirector
- Submission runs for (both modes):
  - Phase 1: Local performance
  - Phase 2: Nearby performance (e.g. within a cloud)
  - Phase 3: Far-away performance

Test datasets



Each of these datasets gets copied to a version with site-specific names in order to so as to automatically test redirection access and to provide a benchmark comparison

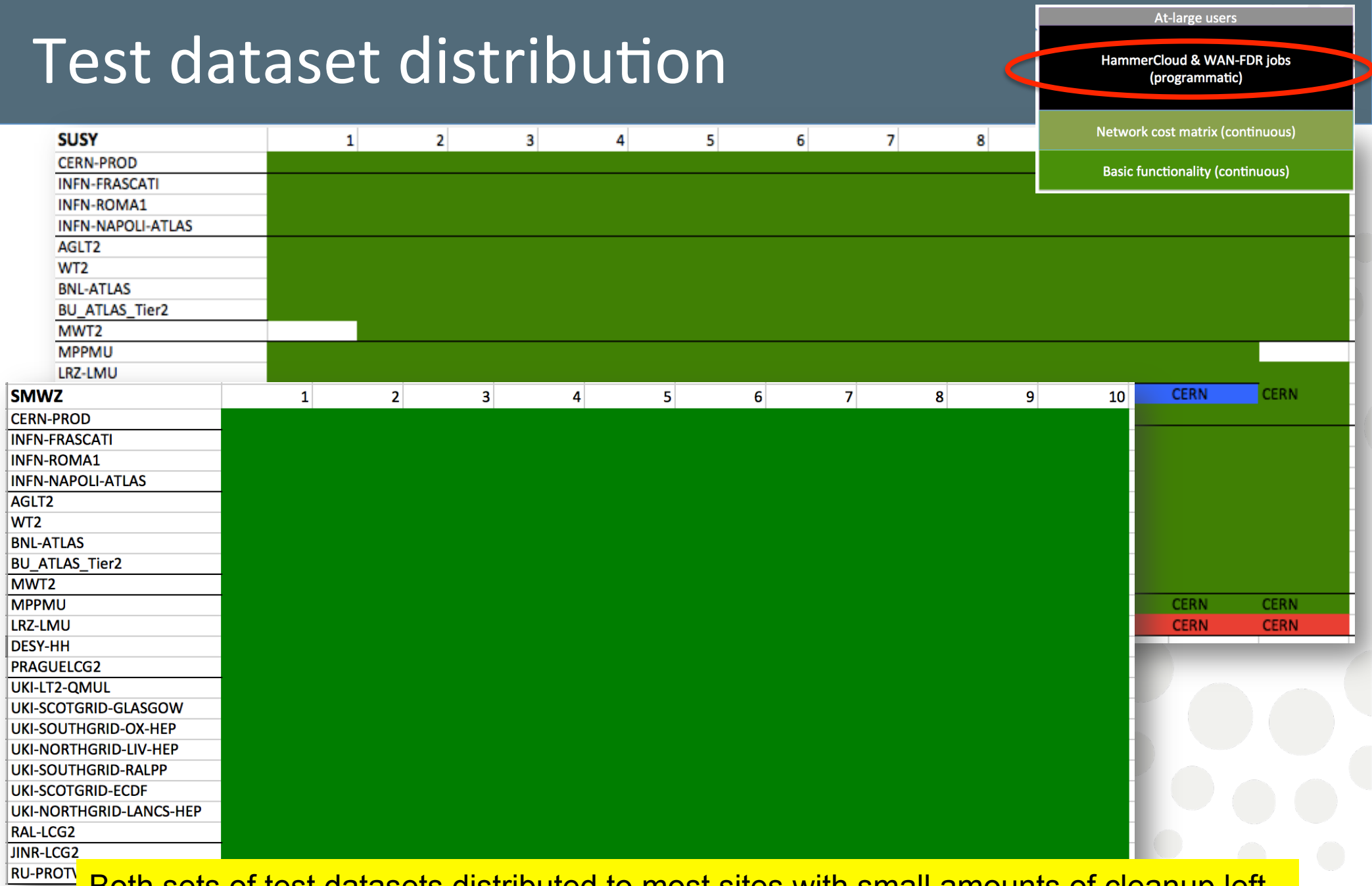
SUSY

data12_8TeV.00203195.physics_JetTauEtmiss.merge.NTUP_SUSYSKIM.r4065_p1278_p1328_p1329_tid01163314_00
data12_8TeV.00203934.physics_JetTauEtmiss.merge.NTUP_SUSYSKIM.r4065_p1278_p1328_p1329_tid01163289_00
data12_8TeV.00209074.physics_JetTauEtmiss.merge.NTUP_SUSYSKIM.r4065_p1278_p1328_p1329_tid01106330_00
data12_8TeV.00209084.physics_JetTauEtmiss.merge.NTUP_SUSYSKIM.r4065_p1278_p1328_p1329_tid01106329_00
data12_8TeV.00209109.physics_JetTauEtmiss.merge.NTUP_SUSYSKIM.r4065_p1278_p1328_p1329_tid01106328_00
data12_8TeV.00209161.physics_JetTauEtmiss.merge.NTUP_SUSYSKIM.r4065_p1278_p1328_p1329_tid01106327_00
data12_8TeV.00209183.physics_JetTauEtmiss.merge.NTUP_SUSYSKIM.r4065_p1278_p1328_p1329_tid01106326_00
data12_8TeV.00209265.physics_JetTauEtmiss.merge.NTUP_SUSYSKIM.r4065_p1278_p1328_p1329_tid01106323_00
data12_8TeV.00209269.physics_JetTauEtmiss.merge.NTUP_SUSYSKIM.r4065_p1278_p1328_p1329_tid01106322_00
data12_8TeV.00209550.physics_JetTauEtmiss.merge.NTUP_SUSYSKIM.r4065_p1278_p1328_p1329_tid01106319_00
data12_8TeV.00209628.physics_JetTauEtmiss.merge.NTUP_SUSYSKIM.r4065_p1278_p1328_p1329_tid01106316_00
data12_8TeV.00209629.physics_JetTauEtmiss.merge.NTUP_SUSYSKIM.r4065_p1278_p1328_p1329_tid01106315_00

SMWZ

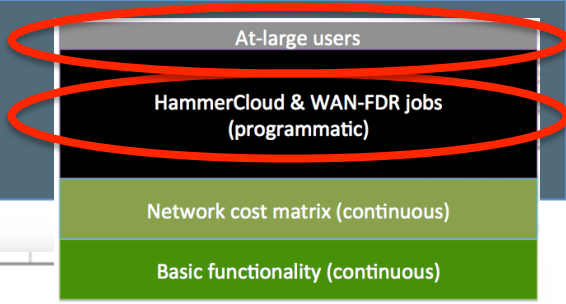
data12_8TeV.00211697.physics_Muons.merge.NTUP_SMWZ.f479_m1228_p1067_p1141_tid00987986_00
data12_8TeV.00211620.physics_Muons.merge.NTUP_SMWZ.f479_m1228_p1067_p1141_tid00986521_00
data12_8TeV.00211522.physics_Muons.merge.NTUP_SMWZ.f479_m1228_p1067_p1141_tid00986520_00
data12_8TeV.00212172.physics_Muons.merge.NTUP_SMWZ.f479_m1228_p1067_p1141_tid01007411_00
data12_8TeV.00212144.physics_Muons.merge.NTUP_SMWZ.f479_m1228_p1067_p1141_tid00999023_00
data12_8TeV.00211937.physics_Muons.merge.NTUP_SMWZ.f479_m1228_p1067_p1141_tid00994157_00
data12_8TeV.00212000.physics_Muons.merge.NTUP_SMWZ.f479_m1228_p1067_p1141_tid00994158_00
data12_8TeV.00212199.physics_Muons.merge.NTUP_SMWZ.f479_m1228_p1067_p1141_tid01007410_00
data12_8TeV.00211772.physics_Muons.merge.NTUP_SMWZ.f479_m1228_p1067_p1141_tid00990030_00
data12_8TeV.00211787.physics_Muons.merge.NTUP_SMWZ.f479_m1228_p1067_p1141_tid00990029_00

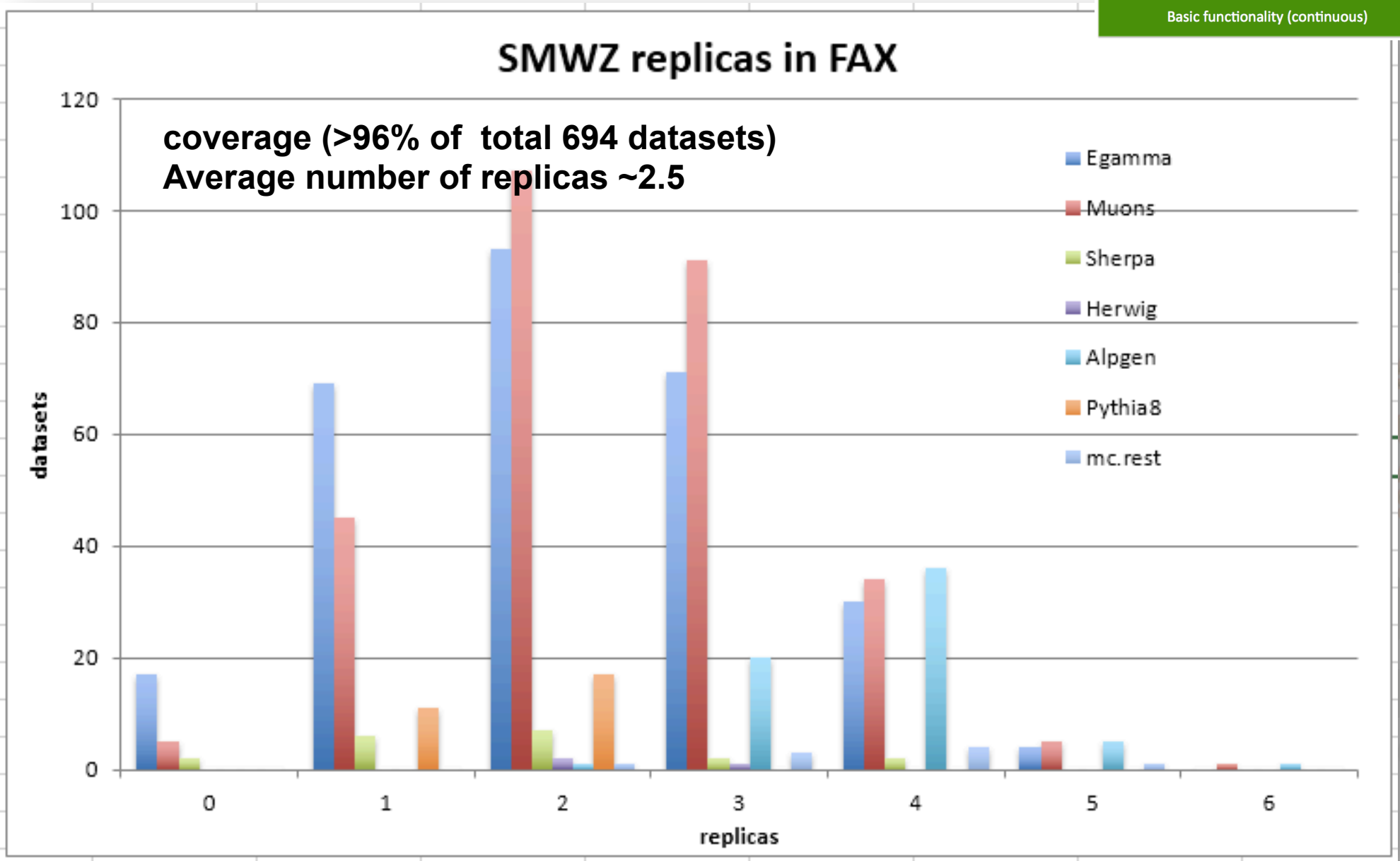
Test dataset distribution



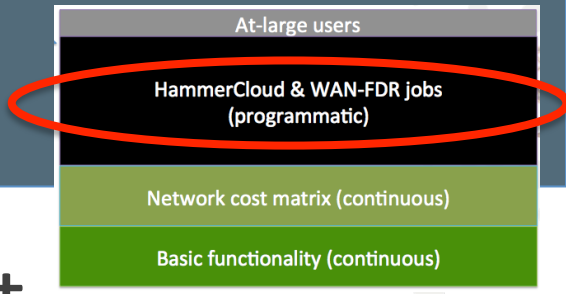
Both sets of test datasets distributed to most sites with small amounts of cleanup left. These datasets will be used to gather reference benchmarks for the various access configuration

Full SMWZ DATA+MC





Queue configurations



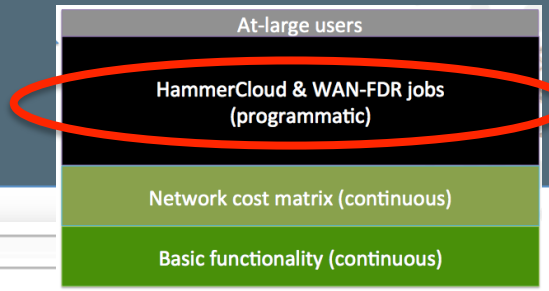
- This turns out to be the hardest part
- Providing federated XRootD access exposes the full extent of heterogeneity of sites, in terms of schedconfig queue parameters
- Each site's "copysetup" parameters seems to differ, and specific parameter settings need to be tried in the Hammer Cloud job submission scripts using –overwriteQueuedata
- Amazingly, in spite of this there are a good fraction of FAX-functional sites

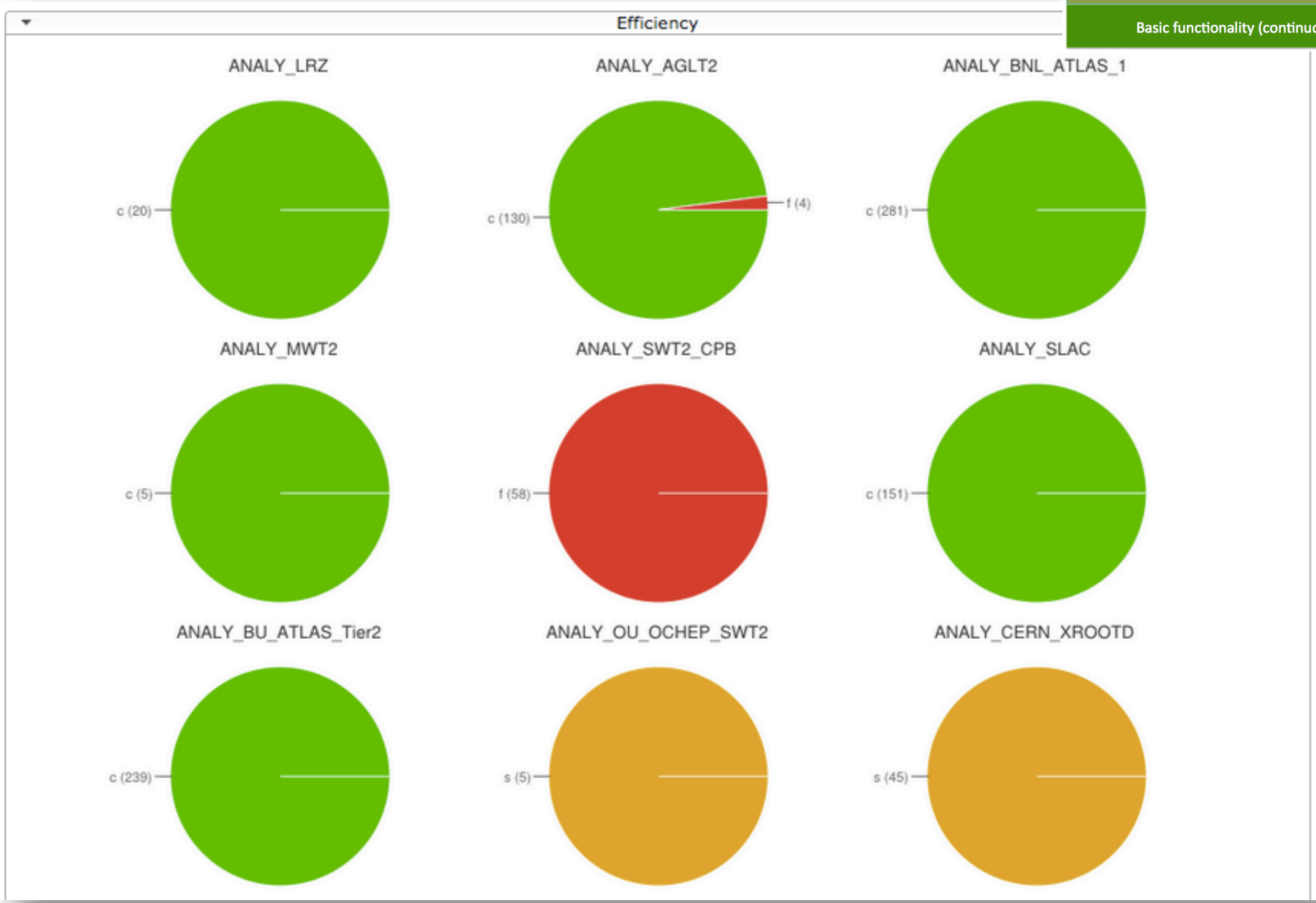
First phase of HC tests: local access

- HC run
  - http://hammercloud.cern.ch/hc/app/atlas/test/20018041/
  - HWW code with regular SMWZ input, FAX directIO, production version pilots
  - This is for **access to local data, but via direct access xrootd**
- Results:
  - 26 sites in the test
  - 16 sites with job successes
  - 3 sites where no job started/finished during test
  - (CERN, ROMA1, OU_OCHEP_SWT2)
  - 1 site does not have input data (GLASGOW)
  - 1 site blacklisted (FZU)
  - 1 site used xrdcp instead of directIO (BNL)
  - 4 sites with 100% failures (EDCF, IHEP, JINR, LANCS)
  - 4 sites with job successes and failures
  - (FRASCATI, NAPOLI, LRZ, RAL)
  - LRZ experienced again xrootd crashes
  - SLAC jobs finally succeed

Johannes, 3 weeks ago

HC efficiencies for selected sites





First phase of HC tests: local access

- HC run
  - http://hammercloud.cern.ch/hc/app/atlas/test/20018258/
  - HWW code with regular SMWZ input, FAX directIO, production version pilots
  - This is for **access to local data, but xrdcp to scratch**
- Results:
  - 28 sites in the test
  - 17 sites with job successes
  - 12 sites with actual xrdcp job successes
  - 7 sites used directIO

    AGLT2, LRZ, MPPMU, MWT2, SLAC, SWT2_CPB, WUPPERTAL
  - 3 sites with all job failures

    IHEP, JINR, SWT2_CPB
  - 3 sites with no jobs started during test

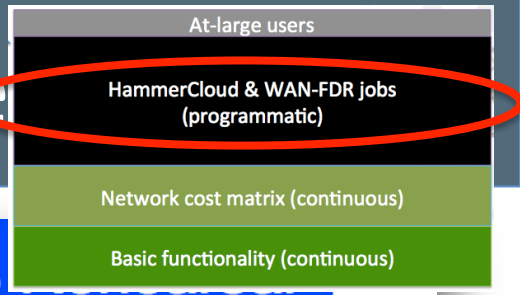
    ECDF, CAM, CERN
  - 1 site with black-listed ANALY queue

    OU_OCHEP_SWT2
  - 2 sites with no input data

    LANCS, GRID-LAL

Johannes, 2 weeks ago

Systematic FDR load tests in progre



ATLAS EXPERIMENT

FAX - Full Dress

To run and get here results of your own tests please contact Ilija Vukotic. ivukotic at standard cern adddress.

Refresh

Show 25 entries

Search:

	TaskID	name	server	client	created	type	files	value	jobs	finished	started	timeout
	84	IT test with real data	INFN-ROMA1	ANALY_CERN_XROOTD	13-01-23 16:41	FAXcopy	10	2798.8	10	10	10	3600
							10	484.7	10	10	10	3600
							10	1013.5	10	4	10	3600
	88	UK test with real data	UKI-SOUTHGRID-OX-HEP	ANALY_OX	13-01-23 22:34	FAXcopy	10	307.9	10	10	10	3600
	90	UK test with real data	UKI-SOUTHGRID-OX-HEP	ANALY_RAL	13-01-23 22:36	FAXcopy	10	4338.8	10	10	10	5000
	92	UK test with real data	UKI-SOUTHGRID-OX-HEP	ANALY_CERN_XROOTD	13-01-24 00:08	FAXcopy						
	93	UK test with real data	UKI-SOUTHGRID-OX-HEP	ANALY_DESY-HH	13-01-24 00:08	FAXcopy						
	94	UK test with real data	UKI-SOUTHGRID-OX-HEP	ANALY_BNL_ATLAS_1	13-01-24 01:14	FAXcopy						
	101	IT test with real data	INFN-ROMA1	ANALY_NET2	13-01-23 22:24	FAXcopy						
	102	IT test with real data	INFN-ROMA1	ANALY_INFN-NAPOLI	13-01-23 22:28	FAXcopy						
	103	UK test with real data	UKI-SOUTHGRID-OX-HEP	ANALY_GLASGOW	13-01-23 22:55	FAXcopy						
	105	UK test with real data	UKI-SOUTHGRID-OX-HEP	ANALY_OX	13-01-24 00:03	read10pc						
	107	UK test with real data	UKI-SOUTHGRID-OX-HEP	ANALY_SWT2_CPB	13-01-24 01:19	FAXcopy						
	108	UK test with real data	UKI-SOUTHGRID-OX-HEP	ANALY_OU_OCHEP_SWT2	13-01-24 01:19	FAXcopy						
	109	IT test with real data	INFN-ROMA1	ANALY_CERN_XROOTD	13-01-24 07:22	read10pc						
	110	IT test with real data	INFN-ROMA1	ANALY_INFN-ROMA1	13-01-24 20:36	read10pc						
	111	IT test with real data	INFN-ROMA1	ANALY_INFN-NAPOLI	13-01-24 20:37	read10pc						
	112	IT test with real data	INFN-ROMA1	ANALY_LRZ	13-01-24 20:37	read10pc	1	null	10	0	0	7200
	121	IT test with	INFN-ROMA1	ANALY_LRZ	13-01-24	FAXcopy	10	null	10	0	3	3600

Adapted WAN framework for specific FDR load tests

Choose analysis queue & FAX server sites, #jobs, #files

Choose access type: copy files or direct ROOT access (10% events, 30 MB client cache)

Record timings in Oracle @ CERN

Systematic FDR load tests in progress

At-large users

HammerCloud & WAN-FDR jobs (programmatic)

Network cost matrix (continuous)

Basic functionality (continuous)

ATLAS EXPERIMENT

FAX - Full Dress

To run and get here results of your own tests please contact Ilija Vukotic. ivukotic at standard cern

Individual job lists + links back to Panda logs

Drill down

Show 25 entries Search:

TaskID	name	server	client	created	type	files	value	jobs	finished	started	timeout
84	IT test with real data	INFN-ROMA1	ANALY_CERN_XROOTD	13-01-23 16:41	FAXcopy	1	2798.8	10	10	10	3600
[illegible]	IT test with real data	INFN-ROMA1	ANALY_INFN-ROMA1	13-01-23 16:41	FAXcopy	1	484.7	10	10	10	3600
[illegible]	[illegible] test with [illegible]	UKI-SOUTHGRID-OX-HEP	ANALY_QMUL	13-01-23 19:25	FAXcopy	1	1013.5	10	4	10	3600
88	[illegible]	UKI-SOUTHGRID-OX-HEP	ANALY_OX	13-01-23 22:34	FAXcopy	[illegible]	[illegible]	10	10	10	3600
90	UK [illegible] real da[illegible]	UKI-SOUTHGRID-[illegible]	ANALY_RAL	13-01-23 22:36	FAXcopy	[illegible]	[illegible]338.8	10	10	10	5000
92	UK test wi[illegible] real data	[illegible]THGRID-	ANALY_CERN_XROOTD	13-01-24 [illegible]	FAXcopy	10	3132.6	10	10	10	5000
93	UK test with real data										
94	UK test with real data										
101	IT test with real data										
102	IT test with real data										
103	UK test with real data										
105	UK test with real data										
107	UK test with real data										
108	UK test with real data										
109	IT test with real data										
110	IT test with real data										
111	IT test with real data										
112	IT test with real data	INFN-ROMA1	ANALY_LRZ	13-01-24 20:37	read10pc	1	null	10	0	0	7200
121	IT test with	INFN-ROMA1	ANALY_LRZ	13-01-24	FAXcopy	10	null	10	0	3	3600

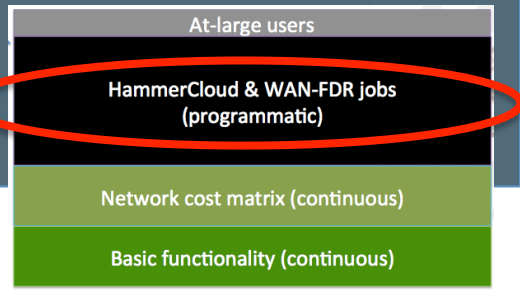
Show 25 entries Search:

TaskID	name	server	client	created	type	files	value	jobs	finished	started	timeout
84	IT test with real data	INFN-ROMA1	ANALY_CERN_XROOTD	13-01-23 16:41	FAXcopy	10	2798.756	10	10	10	3600

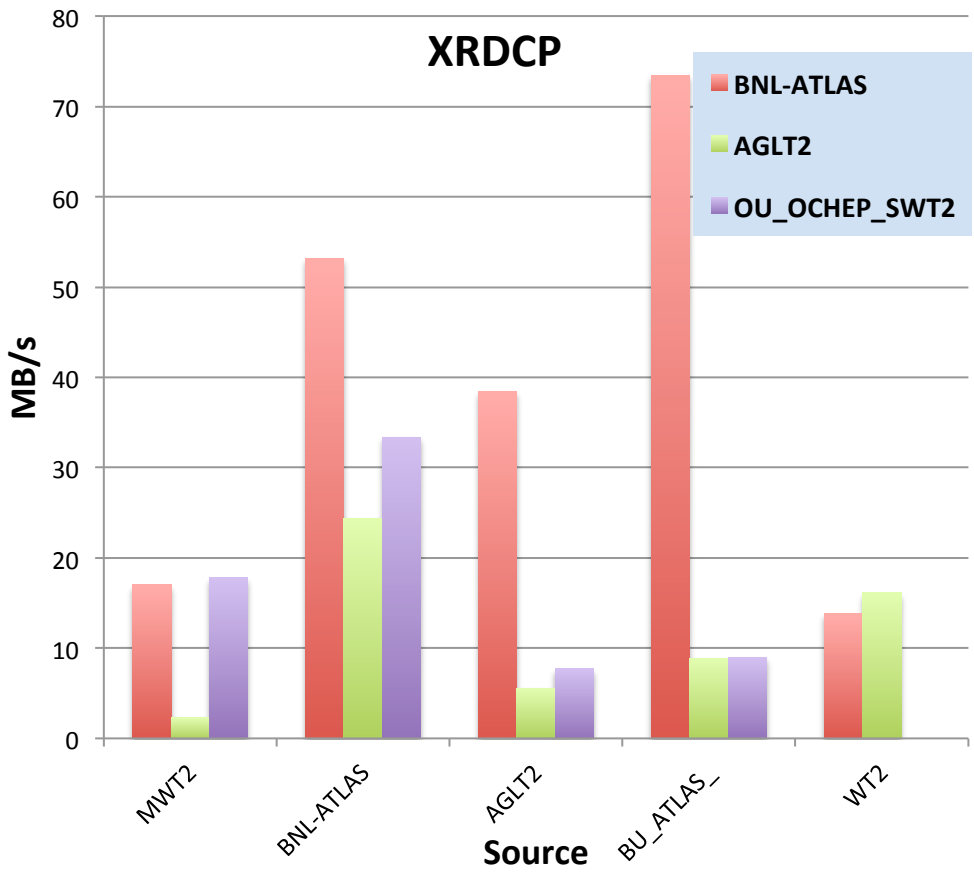
Result arrived	Comment	Value	PandaID
13-01-24 00:02	0 copies failed. 8 were OK. 8.59022138389 MB/s	3563.47	1711941711
13-01-23 23:00	OK. 18.9125960385 MB/s	1993.23	1711911755
13-01-24 00:04	OK. 12.1001014776 MB/s	3055.89	1711952316
13-01-23 23:19	OK. 13.4641255845 MB/s	2815.95	1711915611
13-01-23 23:31	OK. 13.6356259807 MB/s	2616.4	1711925637
13-01-23 23:52	0 copies failed. 6 were OK. 6.9805032048 MB/s	3301.43	1711929535
13-01-23 23:55	OK. 10.6919201596 MB/s	3512.31	1711935163
13-01-23 23:36	0 copies failed. 5 were OK. 5.17533673019 MB/s	3400.96	1711917082
13-01-23 23:41	0 copies failed. 7 were OK. 7.78473594517 MB/s	3335.52	1711921565
13-01-24 00:09	0 copies failed. 8 were OK. 8.29854905735 MB/s	3584.12	1711947841

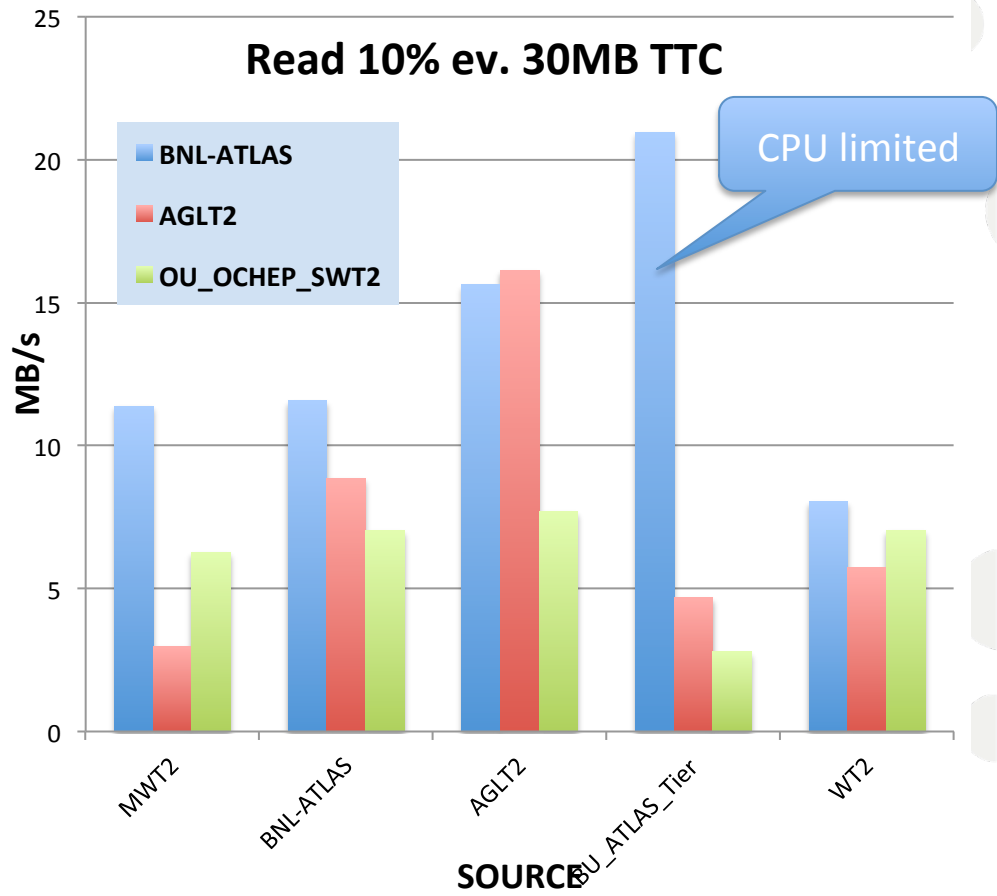
Current Call

Systematic FDR load tests in progress



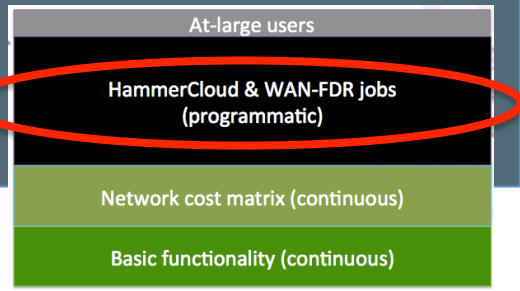
US cloud results.
10 jobs * 10 SMWZ files ~ 50GB



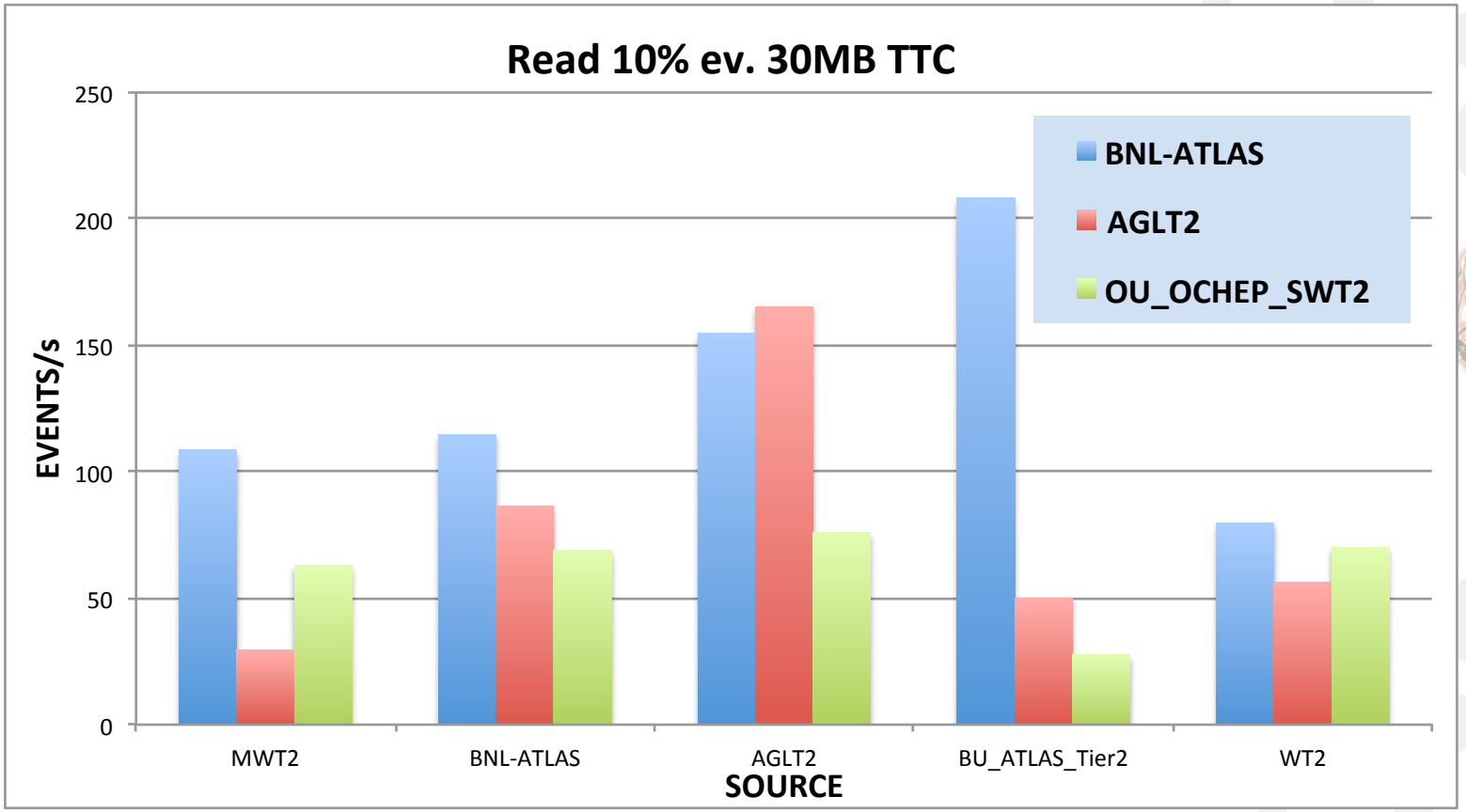


Factors affecting spreads: pair-wise network latency, throughput, storage "business"

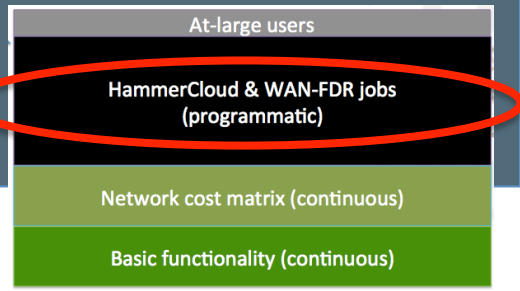
Systematic FDR load tests in progress



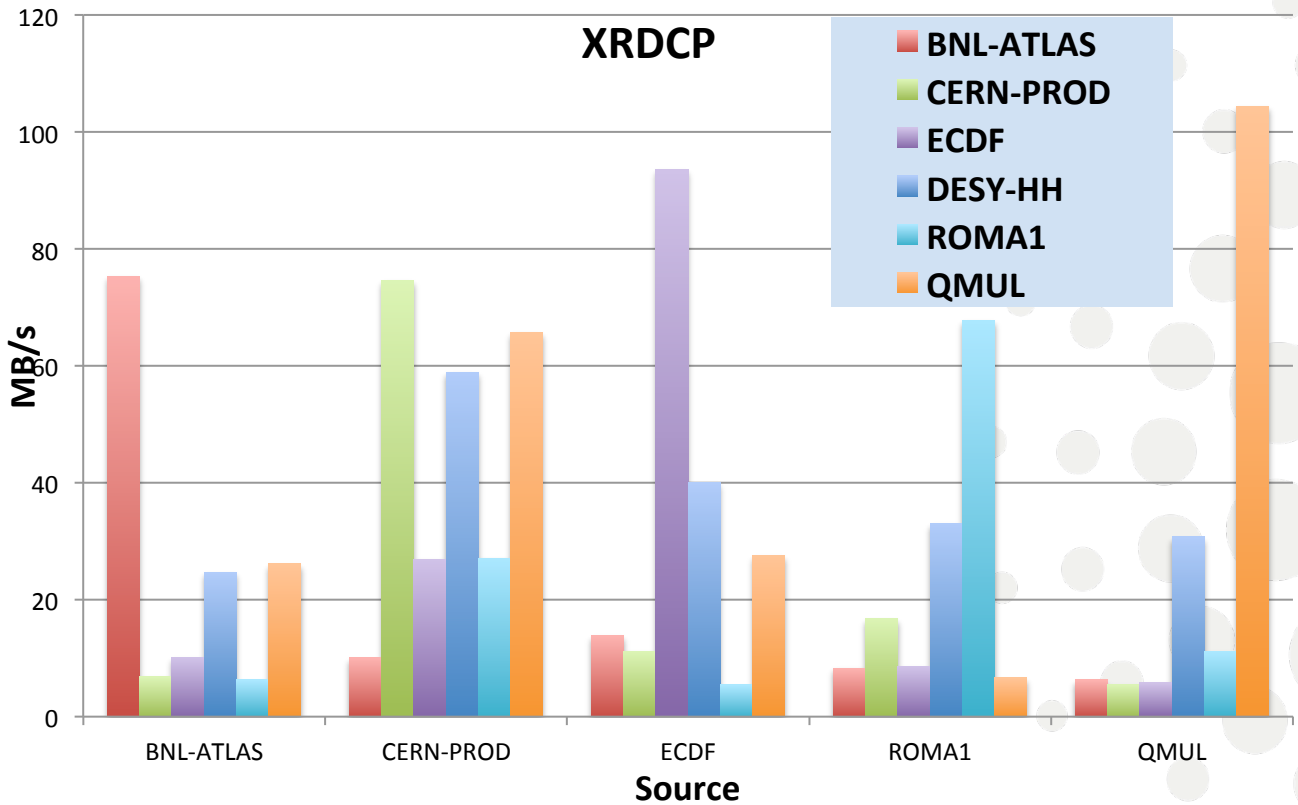
US cloud results



Systematic FDR load tests in progress

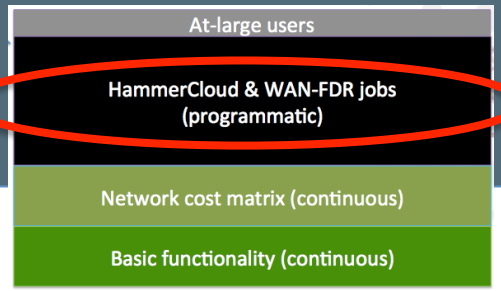


EU cloud results

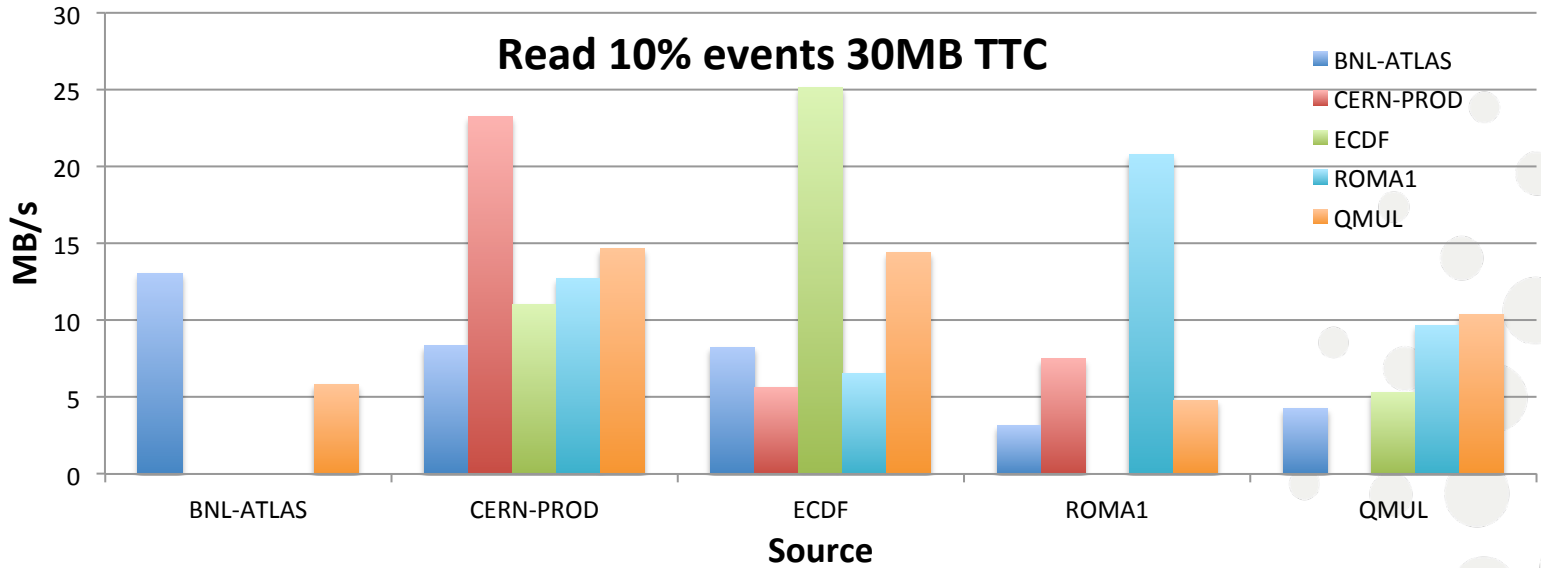


MB/s		destination					
		BNL-ATLAS	CERN-PROD	ECDF	DESY-HH	ROMA1	QMUL
source	BNL-ATLAS	75.27	6.83	10.14	24.68	6.28	26.16
	CERN-PROD	10.08	74.63	26.89	58.88	27.1	65.68
	ECDF	13.87	11.22	93.57	39.99	5.44	27.59
	ROMA1	8.31	16.81	8.61	33.1	67.78	6.65
	QMUL	6.3	5.49	5.82	30.76	11.15	104.39

Systematic FDR load tests in progress



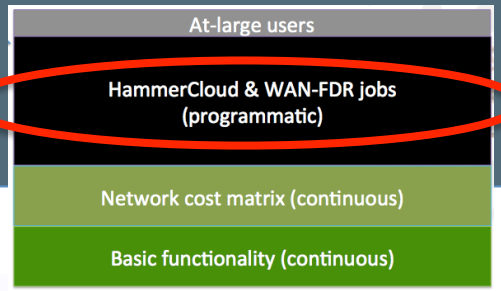
EU cloud results

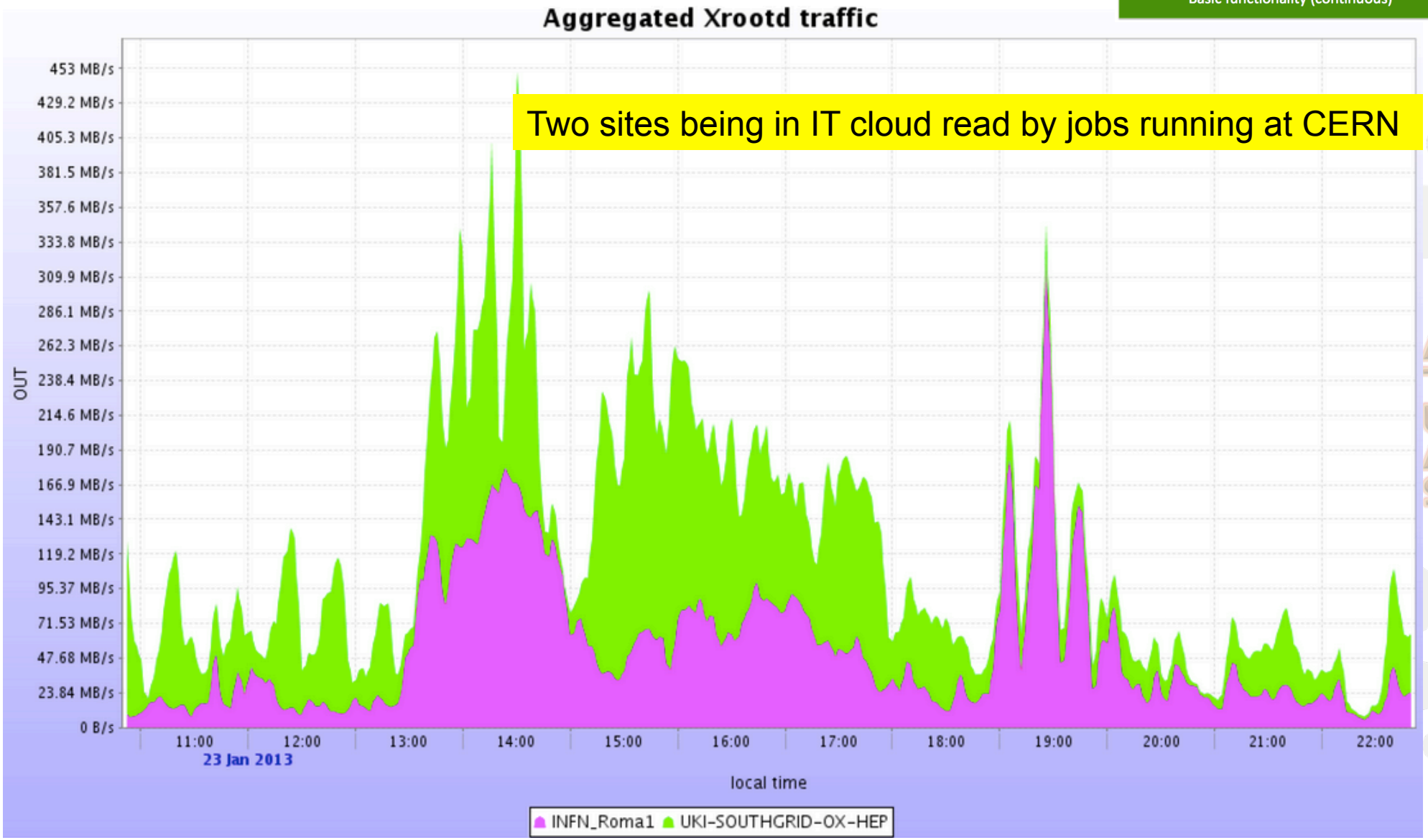


	events/s	destination: BNL-ATLAS	CERN-PROD	ECDF	ROMA1	QMUL
source	BNL-ATLAS	126.76				57.26
	CERN-PROD	82.68	232.52	108.46	123.52	145.96
	ECDF	80.68	56.06	252.39	62.83	145.18
	ROMA1	32	73.66		197.01	49.72
	QMUL	41.34		52.2	99.43	105.46

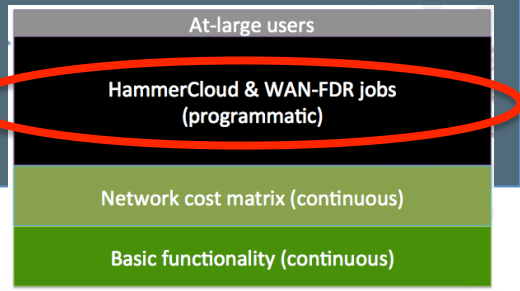
	MB/s	destination: BNL-ATLAS	CERN-PROD	ECDF	ROMA1	QMUL
source	BNL-ATLAS	13.07				5.84
	CERN-PROD	8.36	23.26	11.02	12.71	14.68
	ECDF	8.23	5.64	25.14	6.52	14.42
	ROMA1	3.15	7.49		20.77	4.79
	QMUL	4.26		5.33	9.65	10.38

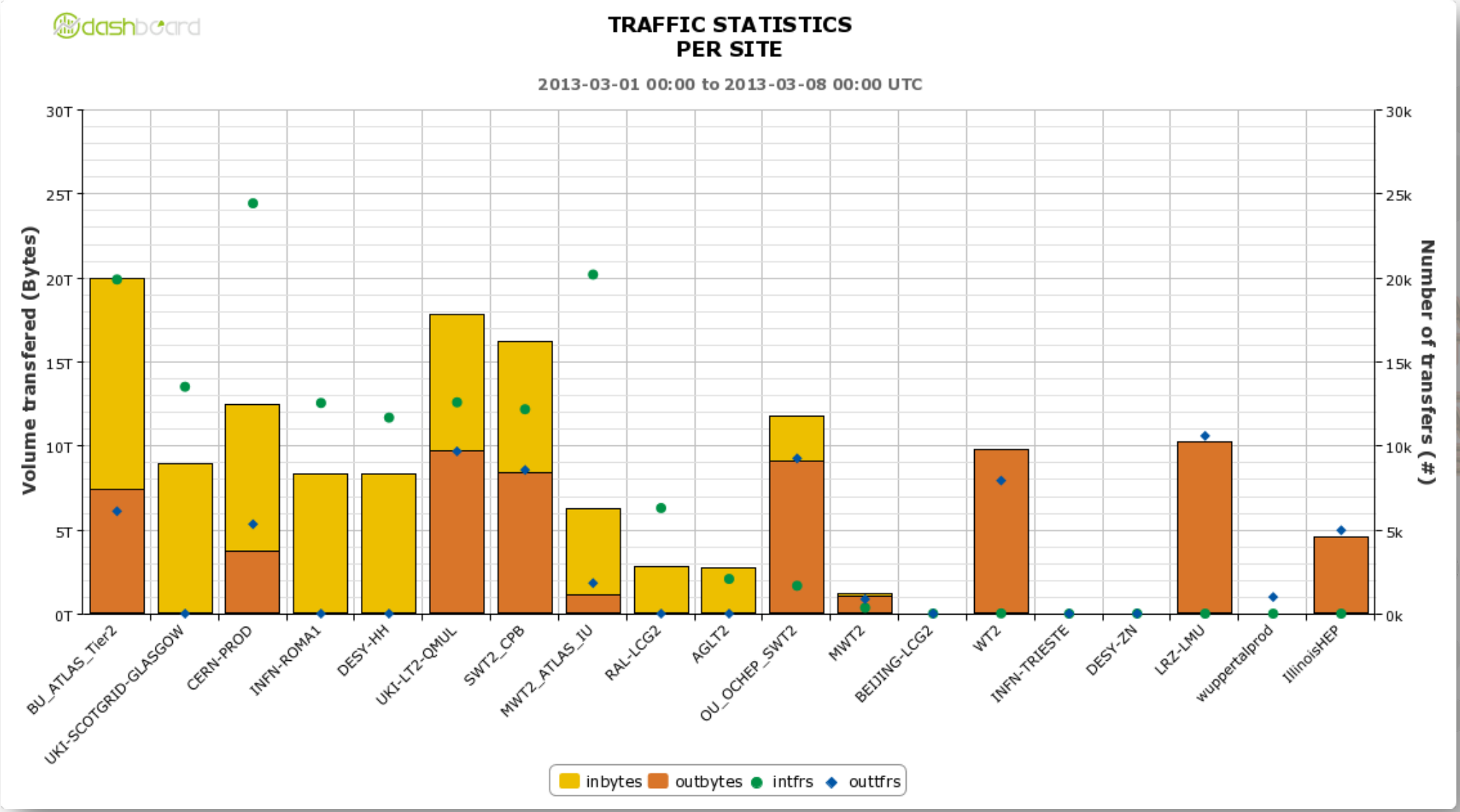
Controlled site “load” testing





Federated traffic seen in the WLCG dashboard





FAX WLCG dashboard

At-large users

HammerCloud & WAN-FDR jobs (programmatic)

Network cost matrix (continuous)

Basic functionality (continuous)

dashb-atlas-xrootd-transfers.cern.ch/ui/#

dashboard

FAX Monitoring (R.C 23)
Latest statistics update: 2013-03-10T17:10:00.031133

TRANSFER PLOTS (2013-03-10 13:10 to 2013-03-10 17:10 UTC SLIDING) PLOT: GROUPING TYPE SERIES SIZE STYLE BIN: SIZE FORMAT STEP

Summary

Common options

Interval
Last 4 hours

Access type
Remote access

Transfer VS Reading
Reading
Copy

Client - Server options

Servers:

Clients:
Grouping:SITE

Users:

File:

Source - Destination options

Sources
Grouping:SITE

Destinations
Grouping:SITE

Centres
Grouping:SITE

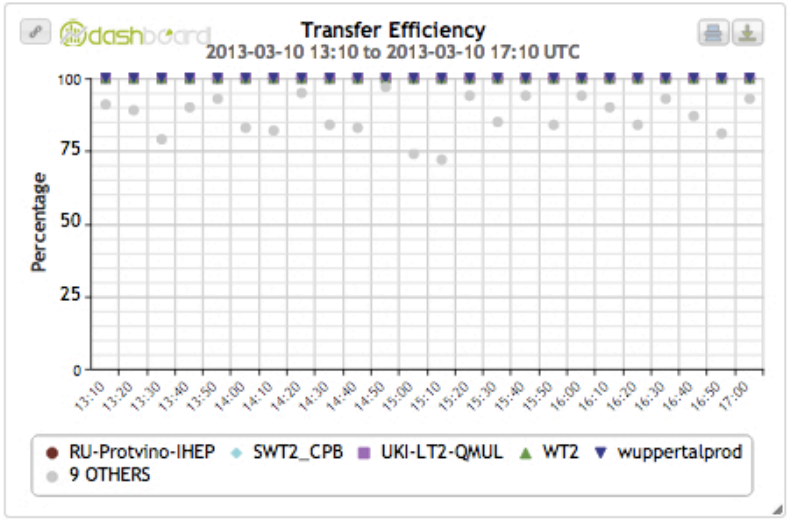
Interval

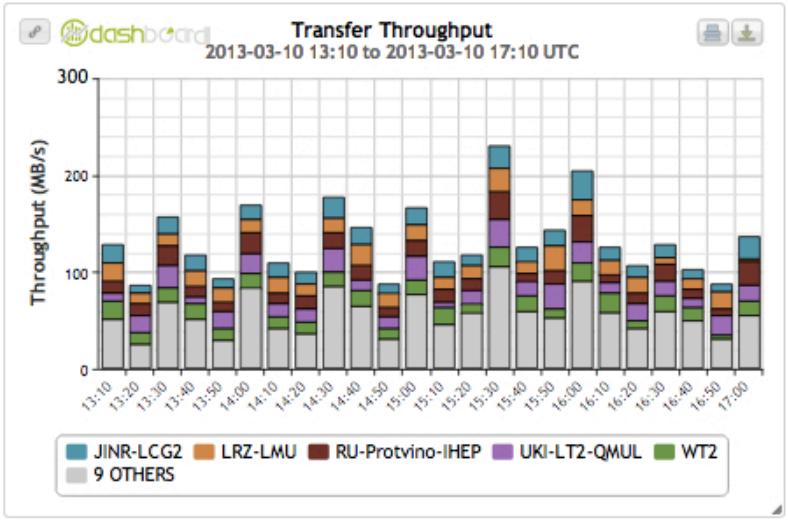
Access type

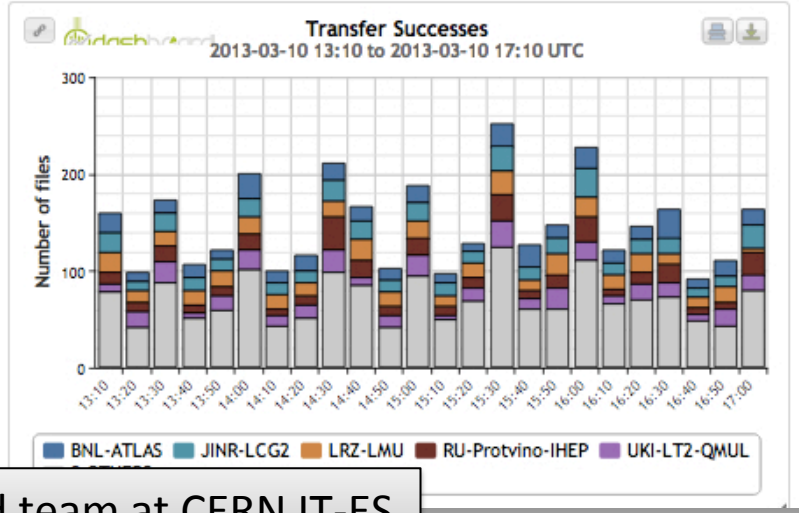
Transfer VS Reading

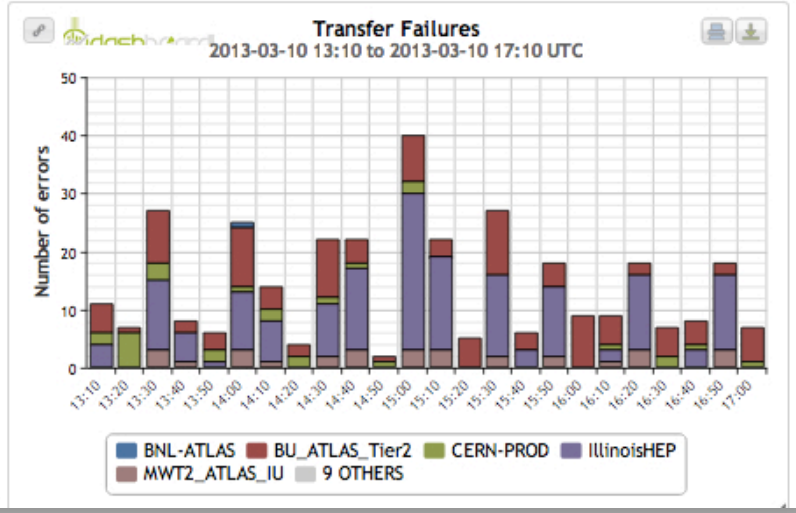
Sources

Matrix | Transfer Plots | Access Plots | Site Statistics | Custom Ranking Plots | Site History | Server Access Pattern



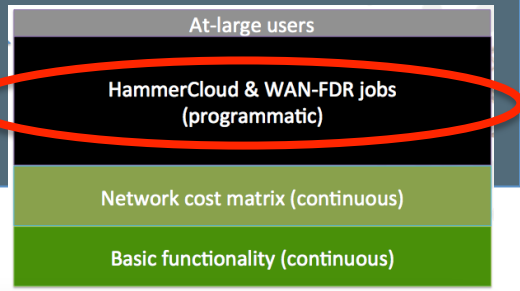


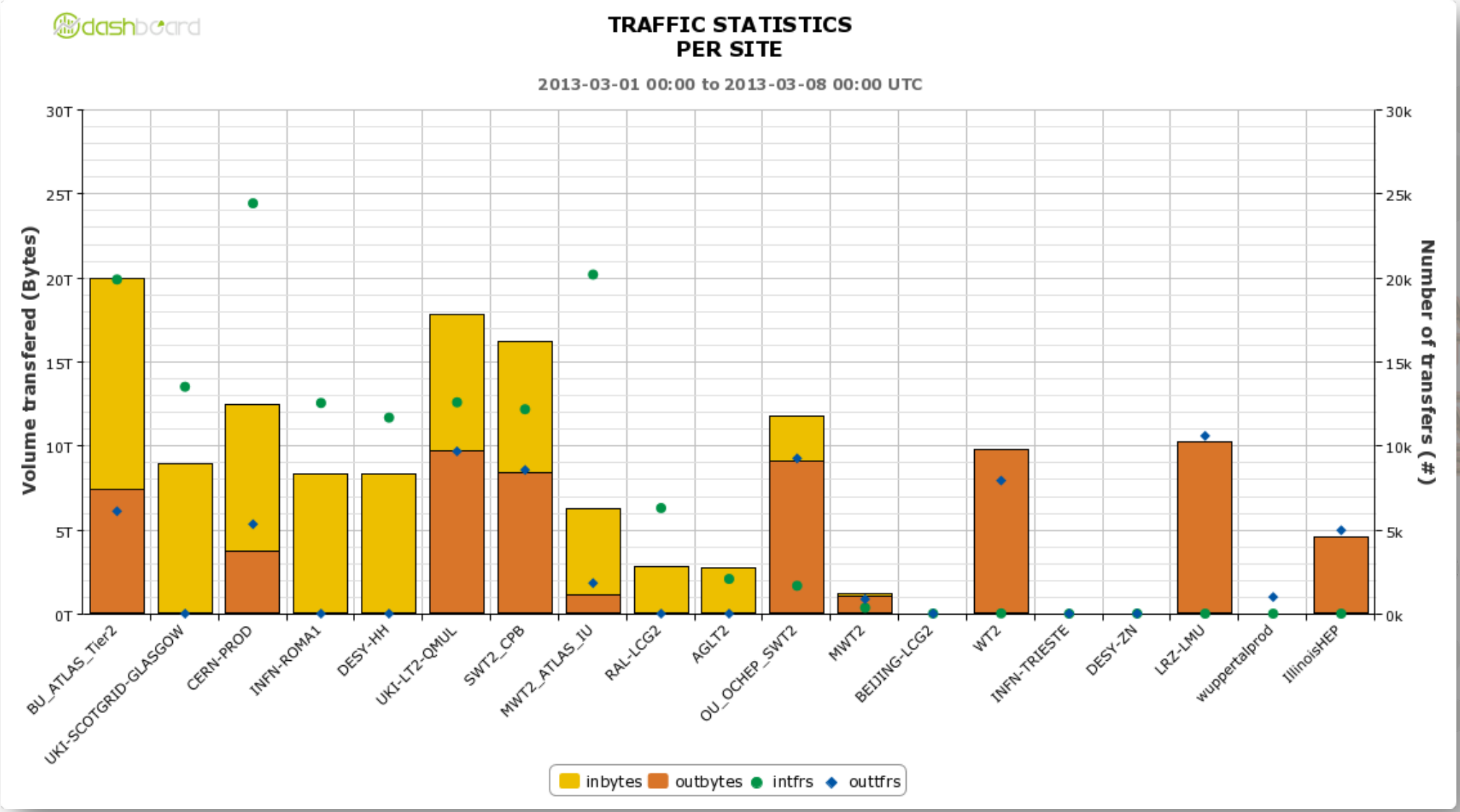




Thanks to dashboard team at CERN IT-ES

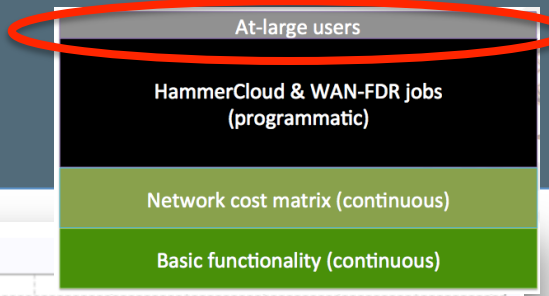
Federated traffic seen in the WLCG dashboard

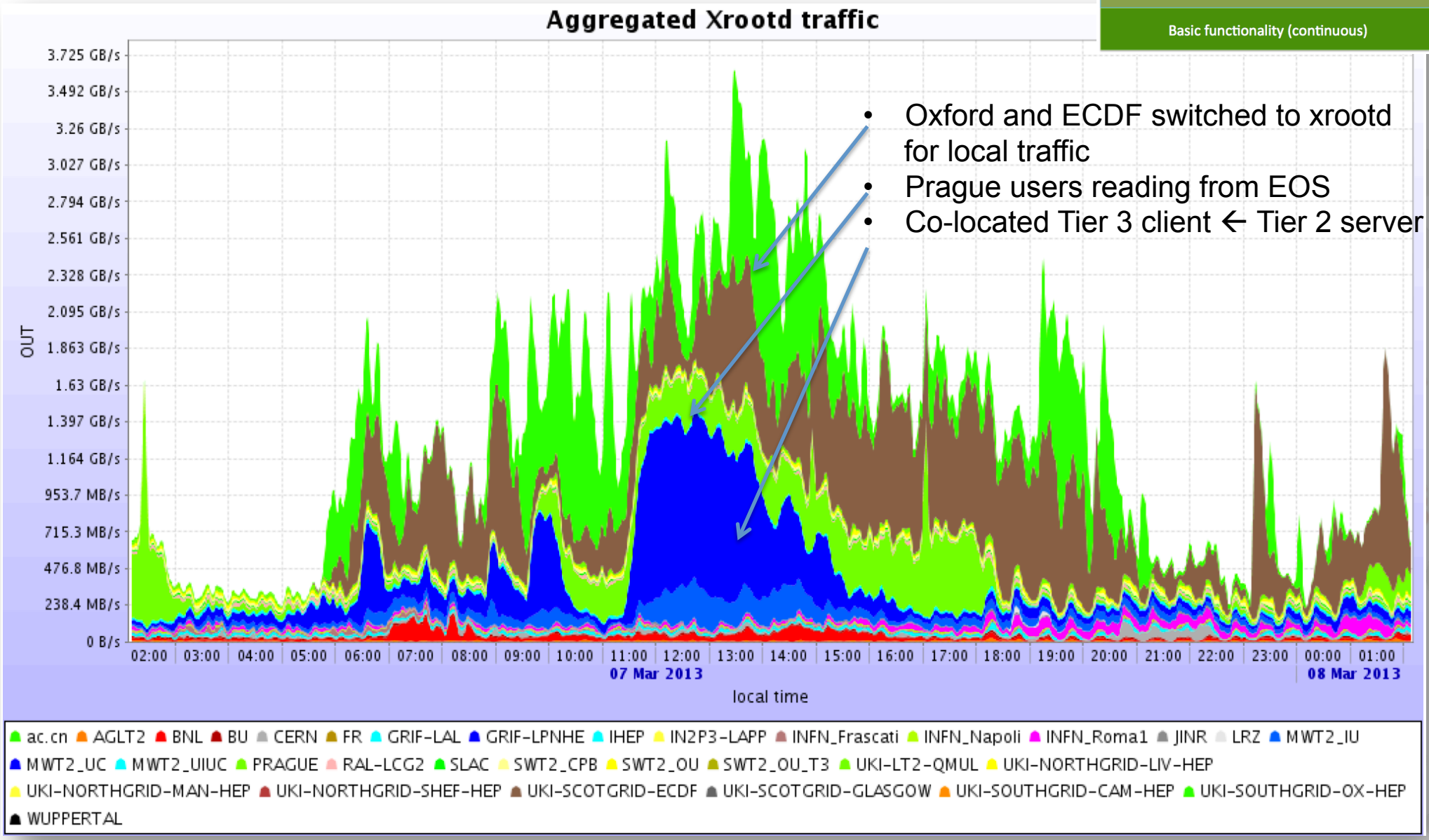




Federation traffic

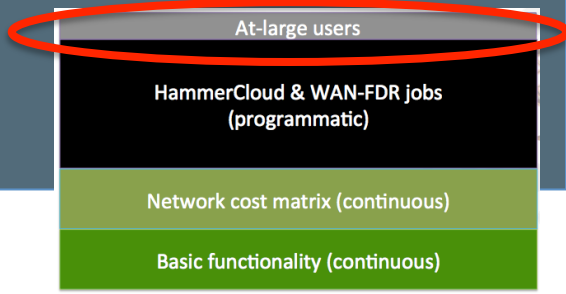
Modest levels now
will grow when in production





- Oxford and ECDF switched to xrootd for local traffic
- Prague users reading from EOS
- Co-located Tier 3 client ← Tier 2 server

Studies from Shuwei Ye at BNL



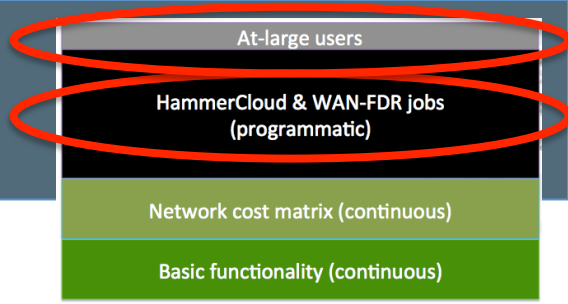
Comparing wall and CPU times for access from Tier3 to datasets at BNL, NET2 and RAL (only BNL results shown)

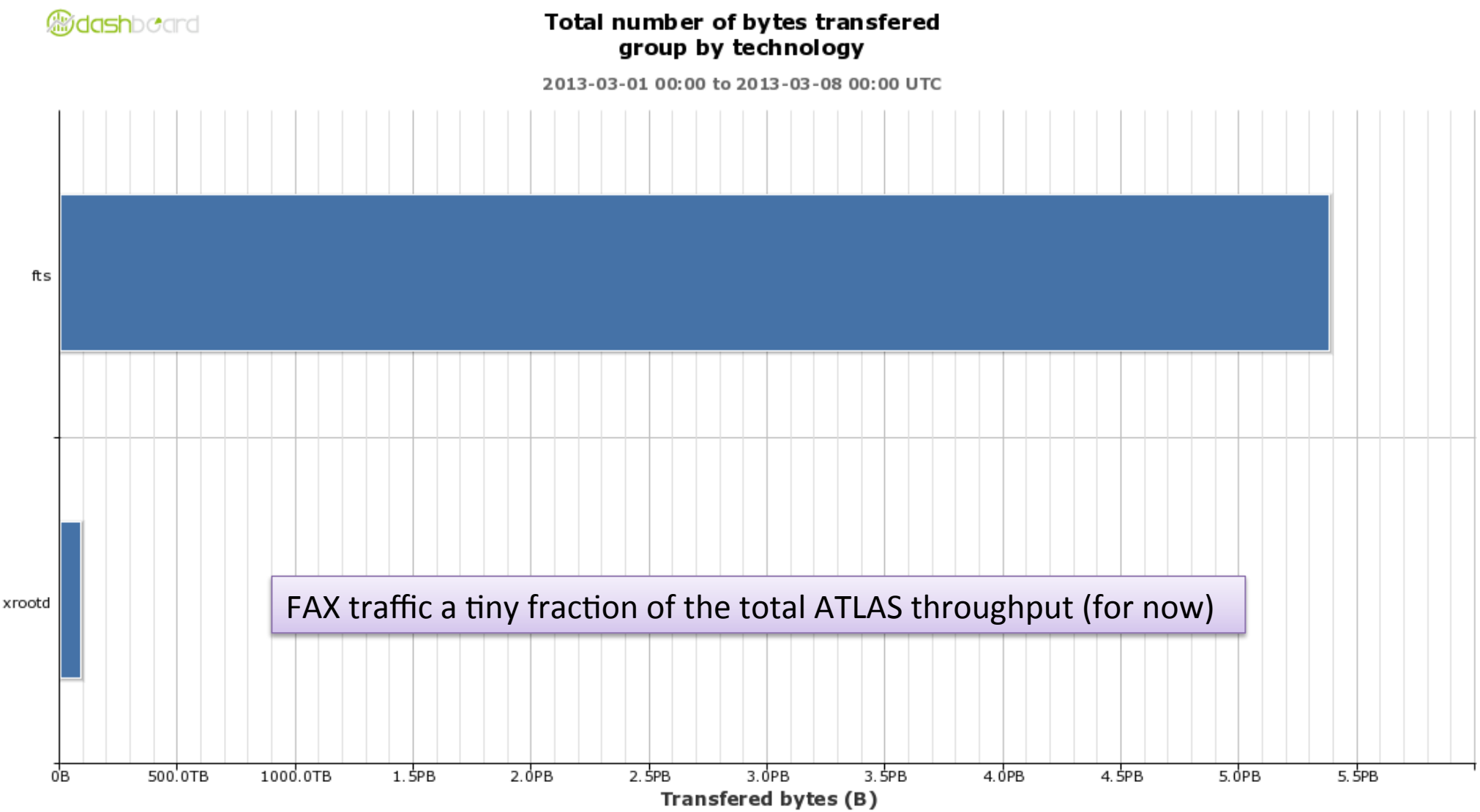
10 TOPMU files at BNL
==========
TotalEvents=244499

Redirector	BNL	NET2	RAL	NFS
chain->MakeSelector(run1)	Wall 0:19	Wall 1:44	Wall 7:24	Wall 0:12
chain->MakeSelector(run2)	Wall 0:15	Wall 1:40	Wall 7:47	
chain->Process(run1)	Wall 1:07	Wall 1:58	Wall 1:42	Wall 0:51
	CPU 26s	CPU 25s	CPU 25s	CPU 21s
Rate(evts/s, run1)	3645	2073	2386	4834
chain->Process(run2)	Wall 1:09	Wall 1:57	Wall 1:56	
	CPU 26s	CPU 25s	CPU 26s	
	3557	2098	2107	

Concludes nearby redirector reduces time to process (validates ATLAS redirection model)
Usual performance hit for “long reach” datasets over slow networks (to RAL)
More systematic studies to come.

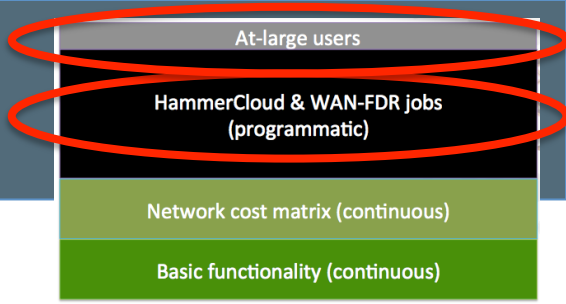
ATLAS throughputs (from US)

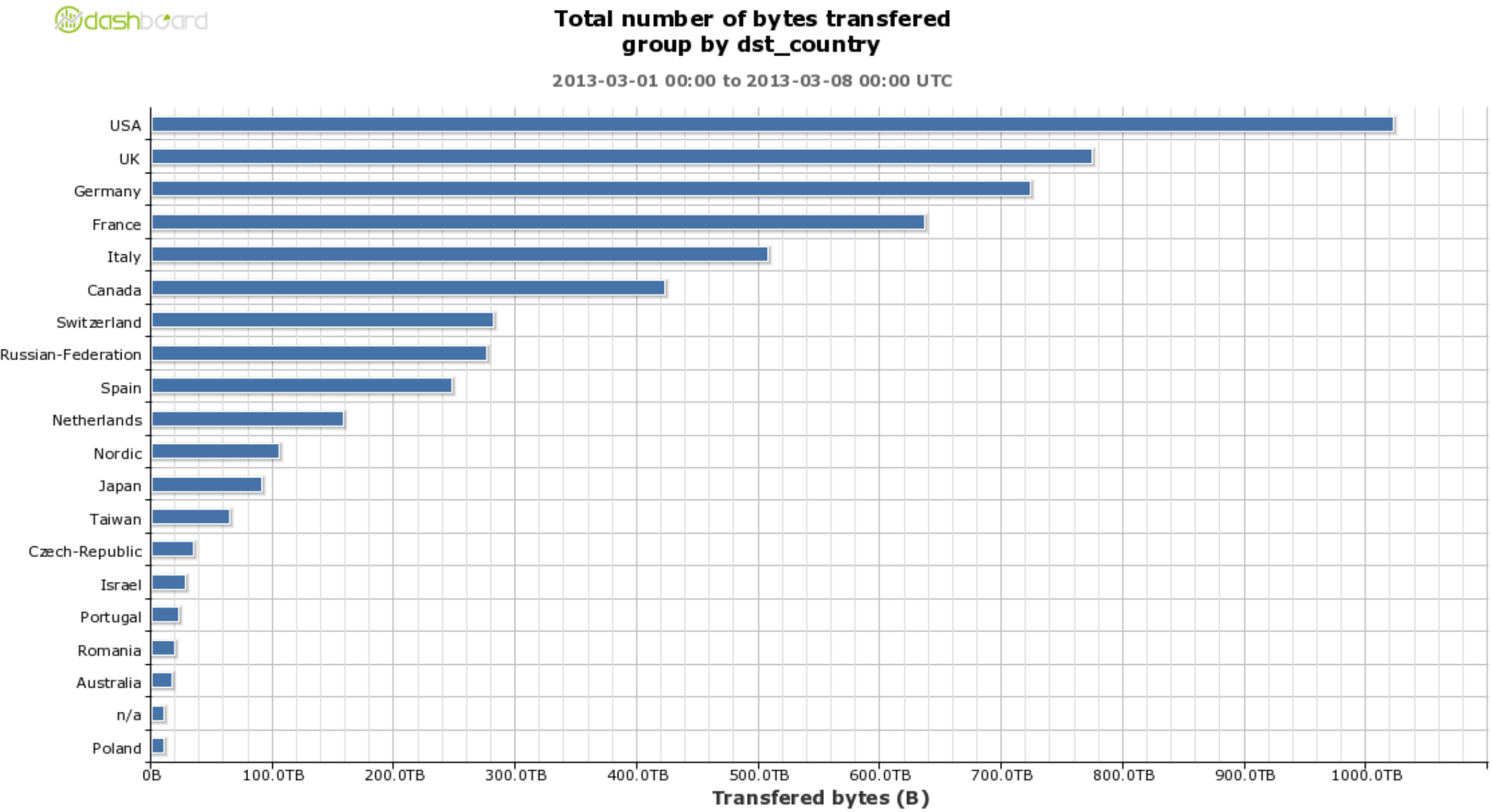




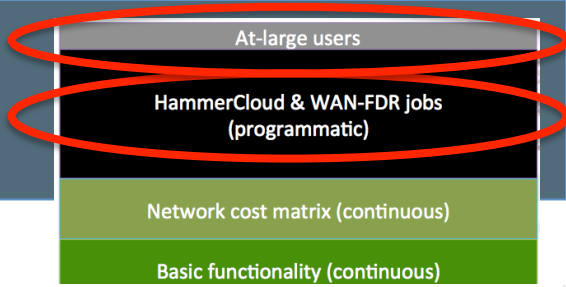
FAX traffic a tiny fraction of the total ATLAS throughput (for now)

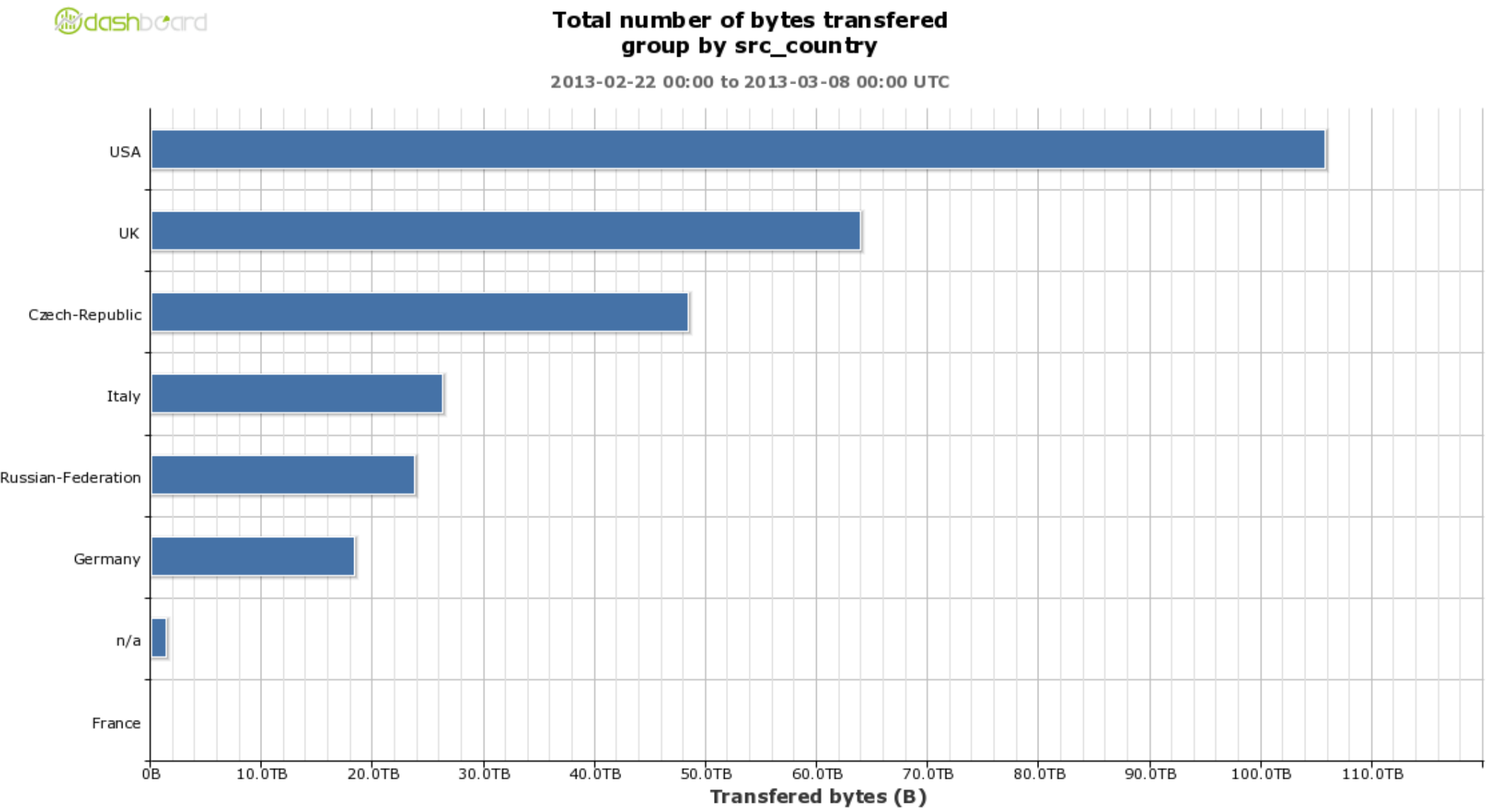
By destination (FTS + FAX)



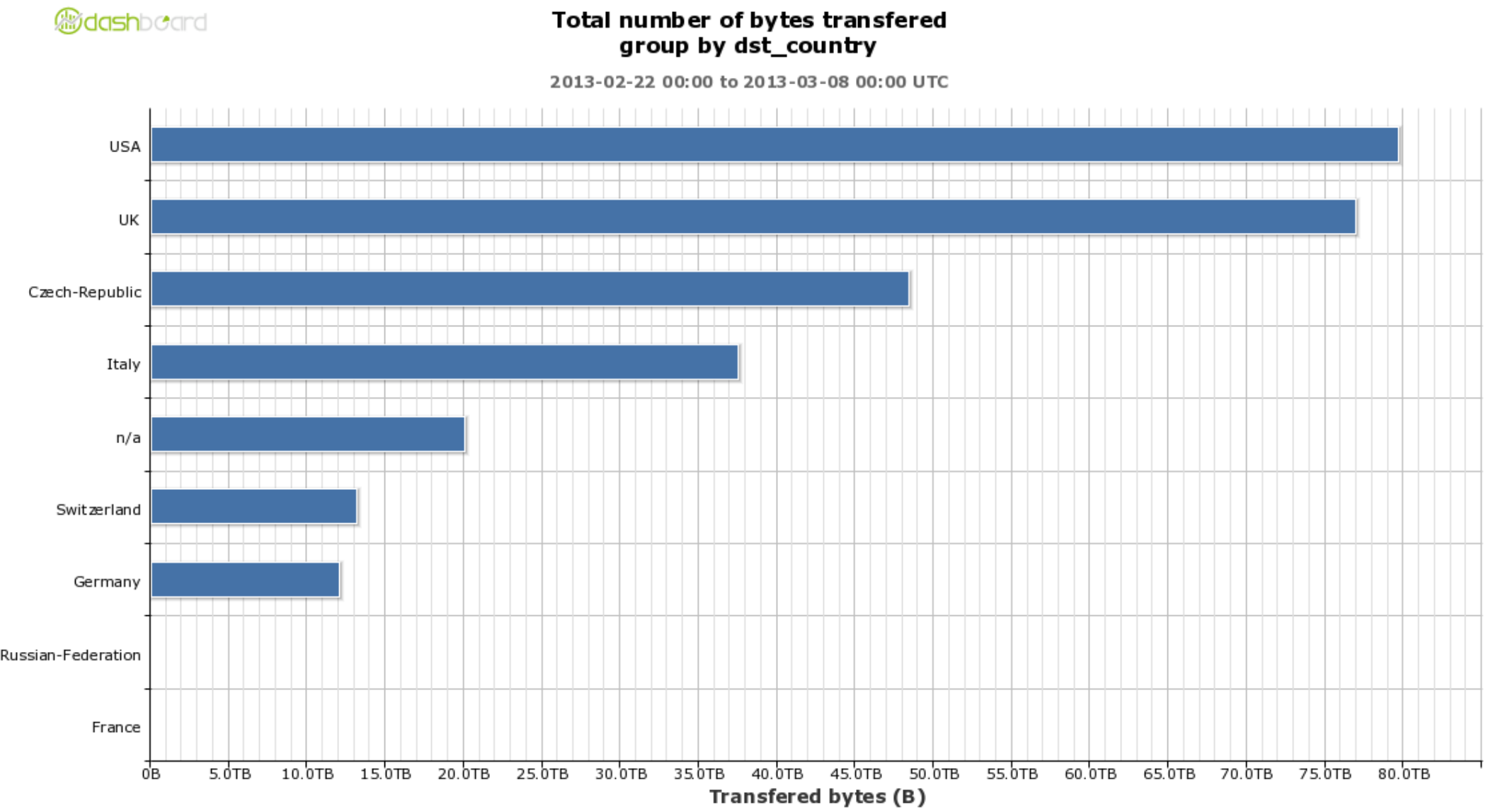


FAX by source cloud

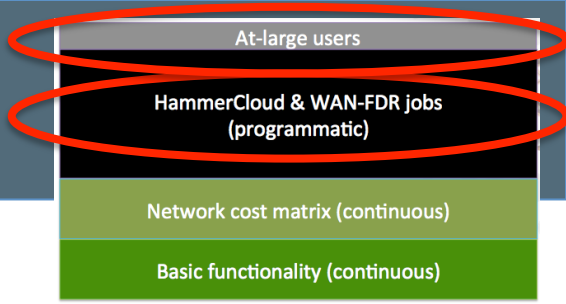


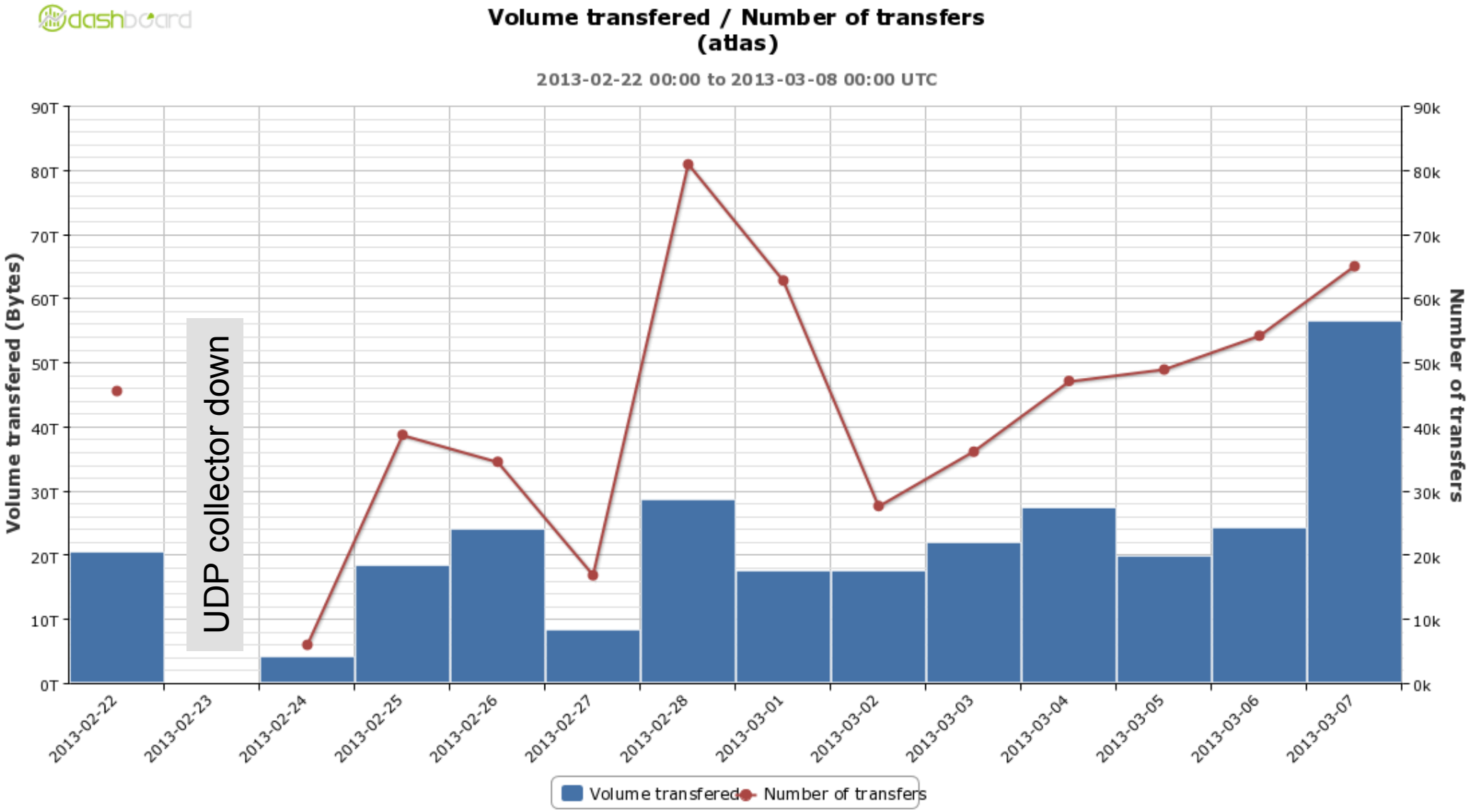


FAX by destination cloud



Daily FAX transfer





Conclusions

- The FDR has been a good exercise in exposing a number of site & system integration issues
  - Site specific client differences → limited proxy check not always working
  - Non-uniform copysetup parameters in schedconfig for sites
  - Lack of fault checking in the rungen script for read failures
  - Tweaks necessary to brokering to allow sending jobs to sites missing datasets
- In spite of this, much progress:
  - New functionality in the pilot to handle global paths without using dq2-client & forcing python 2.6 compatibility at all sites
  - First phase of programmatic HC stress testing nearing completion (local site access)
  - Some FAX accesses from Tier 3s
  - Test datasets in place
- Next steps
  - Programmatic HC stress tests for regional data access (Phase 2)
  - Address remaining integration issues above & continue to validate sites
  - Recruit, acquire feedback from early-adopting users
  - Outsource monitoring services where possible to WLCG, including central UDP collectors, availability probes, etc.
  - Global and Rucio namespace mapping, dev. of new N2N module
  - Set a timeframe for an ATLAS requirement of federating xrootd services at sites

Thanks

- A hearty thanks goes out to all the members of the atlas-adc-federated-xrootd group, especially site admins and providers of redirection & monitoring infrastructure
- Special thanks to Johannes and Federica for preparing HC FAX analysis stress test templates and detailed reporting on test results
- Simone & Hiro for test dataset distribution & Simone for getting involved in HC testing
- Paul, John, Jose for pilot and wrapper changes
- Ilija for FDR testing framework and site-by-site testing
- Wei for doggedly tracking down xrootd security issues & other site problems & Andy for getting ATLAS' required features into xrootd releases