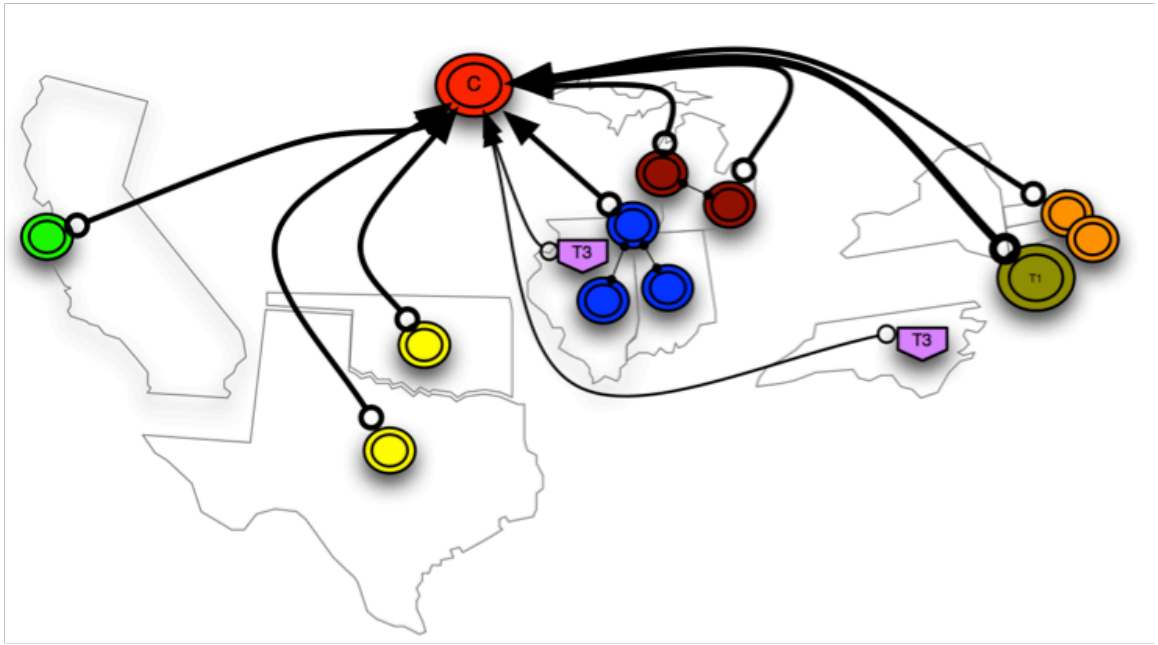
Federating ATLAS Data Stores using Xrootd

Rob Gardner
Lyon Federated Data Meeting
Sep 13, 2012

R&D Activity to Production

BNL Tier 1
AGLT2 (Tier 2)
MWT2 (Tier 2)
SWT2 (Tier 2)
SLAC (Tier 2)
ANL (Tier 3)
BNL (Tier 3)
Chicago (Tier 3)
Duke(Tier 3)
OU (Tier 3)
SLAC (Tier 3)
UTA (Tier 3)
NET (Tier 2)

- 2011 R&D project FAX (Federating ATLAS data stores using Xrootd) was deployed over US Tier 1, Tier 2s and some Tier3s
- Feasibility testing monitoring, site integrations
- In June 2012 extended effort to European sites as an ATLAS-wide project



Federation goals and usage in ATLAS

- Common ATLAS namespace across all storage sites, accessible from anywhere
- Easy to use, homogeneous access to data
- Use as failover for existing systems
- Gain access to more CPUs using WAN direct read access
- Use as caching mechanism at sites to reduce local data management tasks
- WAN data access group formed in ATLAS (see Torre Wenaus' talk later) to determine use cases & requirements on infrastructure

Prerequisites to federation

- Concept of a global name or easy way for analysis users to name input files
- Means to integrate to site's backend storage
  - Name-to-name mapping between LFC name and physical storage path for dCache, Xrootd, GPFS/Lustre(Posix), EOS (in progress)
  - Redirect client to data or proxy server
- System of Xrootd 'redirectors' to open files based network proximity (later cost functions)
- Analysis code that intelligently caches (eg. TTreeCache)

Name translation in ATLAS

- An ATLAS global namespace convention based on unique dataset and filenames has been developed
- While the LFN (and GUID) is unique, the physical path at sites is not
- Use of space tokens as an organizational data management tool means files can live in more than one directory
  - Prior to FTS 2.8 and re-writes, DQ2 suffixes appended to subsequent transfer attempts, preventing deterministic lookups
- Therefore an LFC query is required, at least right now
- Rucio (DQ2 replacement) offers a chance to revisit this

Recent Progress (1)

- dcache-xrootd doors now functional
  - new N2N working
  - recipe for adding Tier 1 sites and others using dCache
    - requires a "helper" xrootd+cmsd pair & billing database publisher for monitoring data
    - New version will have have monitoring plugins for pool services to publish data at source, similar to Xrootd
- Deployment of read-only LFC for unauthenticated LFC lookup (data access still uses GSI security requiring ATLAS voms attribute)

Recent Progress (2)

- Development of general ATLAS monitoring requirements & work with dcache and xrootd teams to standardize (will come out in new xrootd and dcache releases - cf. A.Hanushevsky's talk)
- WLCG dashboard, detailed monitoring services discussions and prototype work (cf. Julia Andreeva, Domenico Giordano talks)
- First European sites federated - from the UK and DE clouds
- Redirectors setup at CERN, good support from
- FAX-enabled pilot available and tested at two sites, implementing "missing file" use-case

Validating new sites, towards production federation readiness over many regions

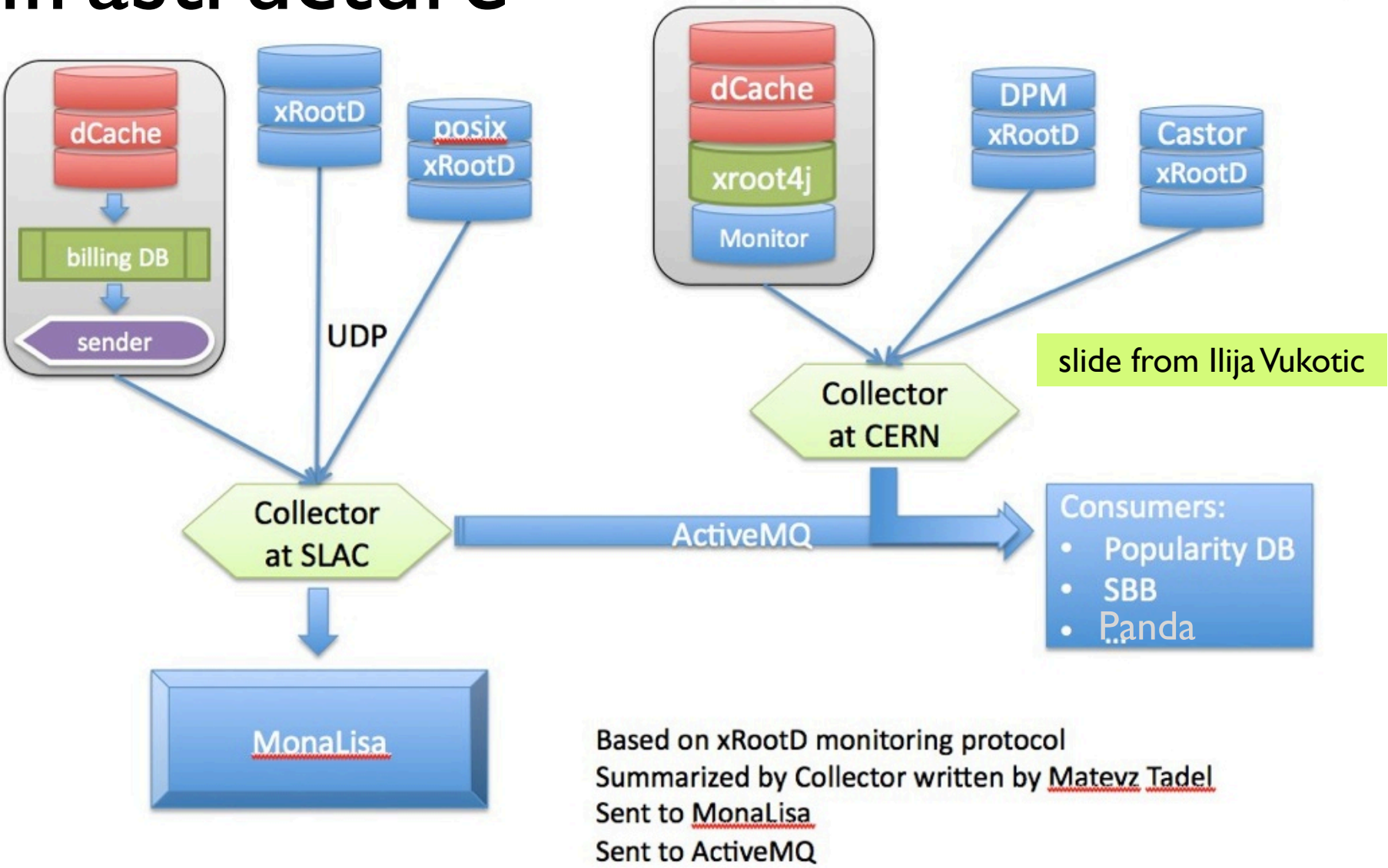
Site certification for federation

Notation: completed | work is in progress | to do | problem

Site	Cloud	SE type-(door)	Regional RD	Federated	X509	GlobalN2N	FAX status mon	UDP Collector	Redir Cloud	Redir Gobal	Failover	Analy Test
BNL	US	dcache	glrd.usatlas.org	completed	completed	completed	completed	completed	completed	completed	work is in progress	work is in progress
AGLT2	US	dcache-xrootd	xrd-central to glrd.usatlas.org	completed	completed	completed	completed	completed	completed	completed	work is in progress	work is in progress
MWT2.org	US	dcache-xrootd	xrd-central to glrd.usatlas.org	completed	completed	completed	completed	completed	completed	completed	work is in progress	work is in progress
MWT2_UC,IU	US	xrootd	xrd-central to glrd.usatlas.org	completed	completed	completed	completed	completed	completed	completed	work is in progress	work is in progress
NET2	US	GPFS	glrd.usatlas.org	completed	completed	completed	completed	completed	completed	completed	work is in progress	work is in progress
SWT2 (UTA)	US	xrootd	glrd.usatlas.org	completed	completed	completed	completed	completed	completed	completed	work is in progress	work is in progress
SWT2 (OU)	US	Lustre	xrd-central to glrd.usatlas.org	completed	completed	completed	completed	completed	completed	completed	work is in progress	work is in progress
SLAC	US	xrootd	glrd.usatlas.org	completed	completed	completed	completed	completed	completed	completed	work is in progress	work is in progress
Wuppertal	DE	dcache-xrootd	atlas-xrd-de.cern.ch	completed	completed	completed	work is in progress	completed	completed	work is in progress	work is in progress	work is in progress
LRZ-LMU	DE	dcache-xrootd via xrootd proxy	atlas-xrd-de.cern.ch	completed	work is in progress	work is in progress	work is in progress	completed	completed	work is in progress	work is in progress	work is in progress
Edinburgh	UK	DPM	atlas-xrd-uk.cern.ch	completed	completed	completed	work is in progress	completed	completed	work is in progress	work is in progress	work is in progress
Glasgow	UK	DPM	atlas-xrd-uk.cern.ch	completed	completed	completed	work is in progress	completed	completed	work is in progress	work is in progress	work is in progress
Oxford	UK	DPM	atlas-xrd-uk.cern.ch	completed	completed	work is in progress	work is in progress	work is in progress	work is in progress	work is in progress	work is in progress	work is in progress
QMUL	UK	Storm/Lustre	atlas-xrd-uk.cern.ch	work is in progress	work is in progress	work is in progress	work is in progress	work is in progress	work is in progress	work is in progress	work is in progress	work is in progress
EOS	EU	EOS	atlas-xrd-eu.cern.ch	completed	completed	work is in progress	work is in progress	work is in progress	work is in progress	work is in progress	work is in progress	work is in progress
Dubna	RU		atlas-xrd-ru.cern.ch	work is in progress	work is in progress	work is in progress	work is in progress	work is in progress	work is in progress	work is in progress	work is in progress	work is in progress

Monitoring infrastructure

Large extra development
define and implement a new monitoring standard for both xRootD and dCache.



slide from Ilija Vukotic

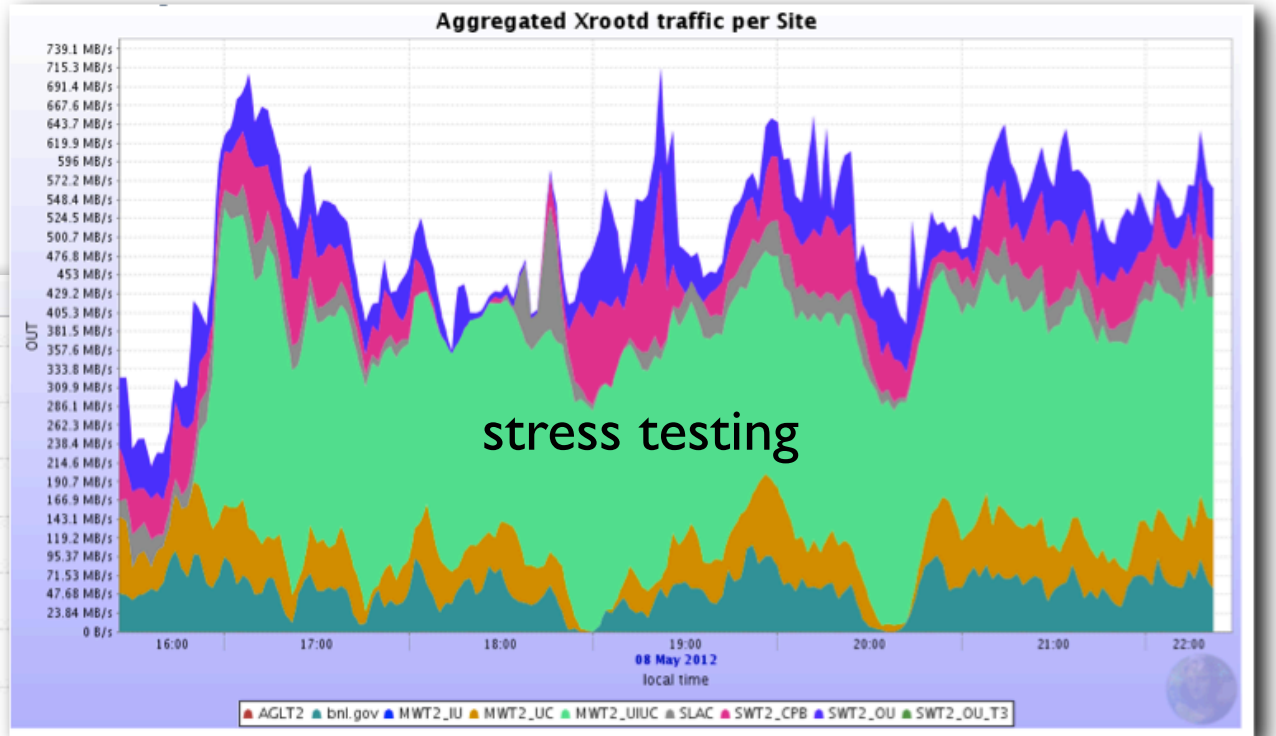
Based on xRootD monitoring protocol
Summarized by Collector written by Matevz Tadel
Sent to MonaLisa
Sent to ActiveMQ

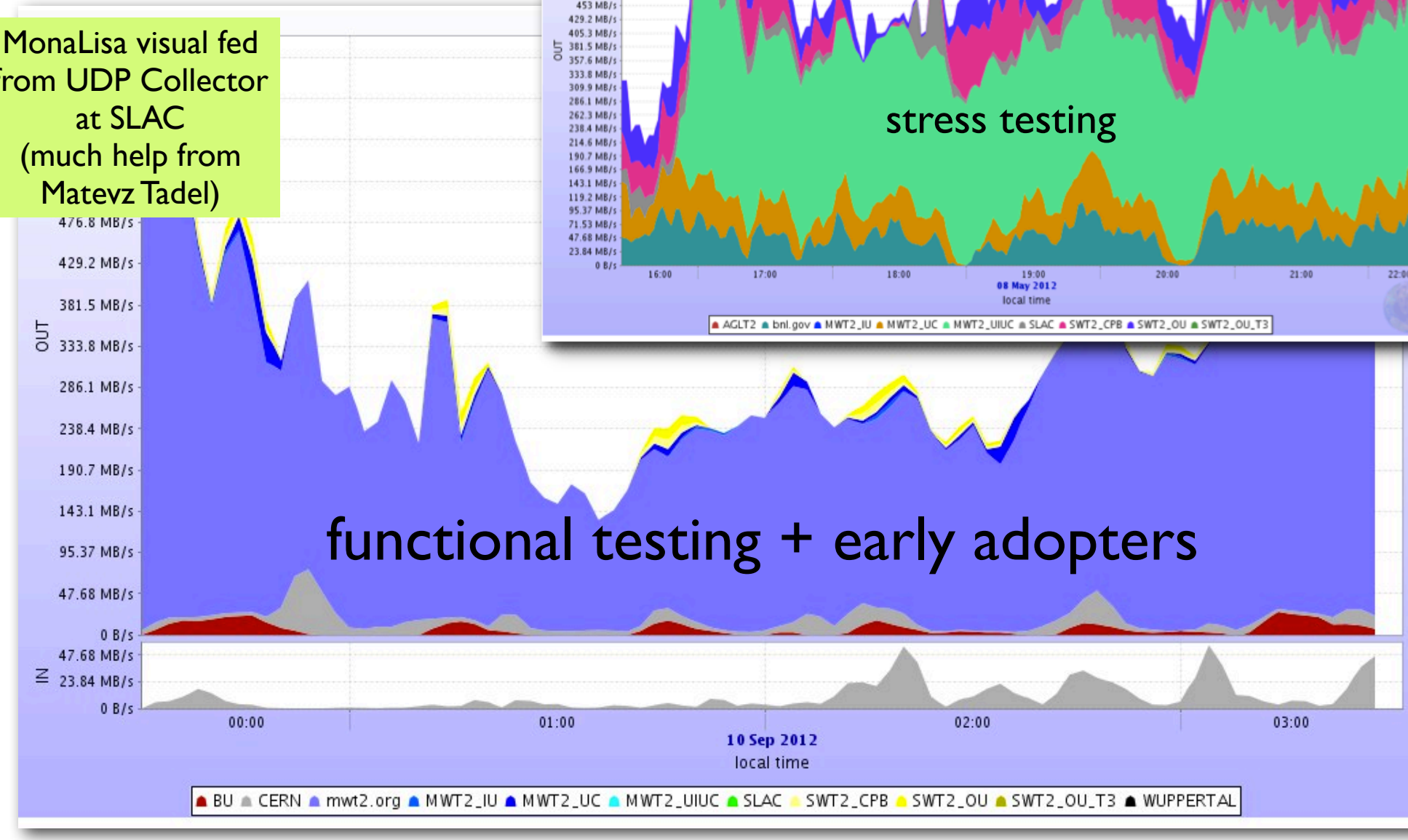
Collaborative work w/WLCG, AAA, Xrootd; cf Julia Andreeva et. al.
https://twiki.cern.ch/twiki/bin/view/LCG/WLCGTransferMonitoring

More in later talks!

Summarized monitoring

MonaLisa visual fed from UDP Collector at SLAC (much help from Matevz Tadel)





Xrootd DATA POPULARITY

CERN IT Experiment Support

PROOF OF CONCEPT ON ATLAS EOS

Home | xrd monitor | Plots | Tables

FAX REAL-TIME MONITORING

Provide the real time snapshot of the last 10 mins of collected Xrootd monitoring data.
UTC Time is reported.

publishing detailed Xrootd stream from SLAC to ActiveMQ at CERN

Show 10 entries

starttime	endtime	server domain	client domain	read bytes	write bytes	file size	filename
2012-09-11 15:24:06	2012-09-11 15:25:50	slac.stanford.edu	uchicago.edu	797152257	0	797152257	/atlas/dq2/user/ilijav/HCtest/user.ilijav.HCtest.1/group.test.hc.NTUP_SMWZ.root
2012-09-11 15:25:17	2012-09-11 15:25:24	atlas-swt2.org	atlas-swt2.org	797152257	0	797152257	/atlas/dq2/user/ilijav/HCtest/user.ilijav.HCtest.1/group.test.hc.NTUP_SMWZ.root
2012-09-11 15:24:14	2012-09-11 15:25:09	atlas-swt2.org	uchicago.edu	797152257	0	797152257	/atlas/dq2/user/ilijav/HCtest/user.ilijav.HCtest.1/group.test.hc.NTUP_SMWZ.root
2012-09-11 15:24:12	2012-09-11 15:25:07	bu.edu	uchicago.edu	0	0	797152257	/atlas/dq2/user/ilijav/HCtest/user.ilijav.HCtest.1/group.test.hc.NTUP_SMWZ.root
2012-09-11 15:24:14	2012-09-11 15:25:04	uchicago.edu	mwt2.org	0	0	797152257	/atlas/dq2/user/ilijav/HCtest/user.ilijav.HCtest.1/group.test.hc.NTUP_SMWZ.root
2012-09-11 15:23:28	2012-09-11 15:24:38	uchicago.edu	ochep.ou.edu	0	0	797152257	/atlas/dq2/user/ilijav/HCtest/user.ilijav.HCtest.1/group.test.hc.NTUP_SMWZ.root
2012-09-11 15:22:39	2012-09-11 15:24:13	slac.stanford.edu	uchicago.edu	797152257	0	797152257	/atlas/dq2/user/ilijav/HCtest/user.ilijav.HCtest.1/group.test.hc.NTUP_SMWZ.root
2012-09-11 15:23:27	2012-09-11 15:24:07	bu.edu	ochep.ou.edu	0	0	797152257	/atlas/dq2/user/ilijav/HCtest/user.ilijav.HCtest.1/group.test.hc.NTUP_SMWZ.root
2012-09-11 15:23:24	2012-09-11 15:23:58	slac.stanford.edu	ochep.ou.edu	797152257	0	797152257	/atlas/dq2/user/ilijav/HCtest/user.ilijav.HCtest.1/group.test.hc.NTUP_SMWZ.root
2012-09-11 15:23:29	2012-09-11 15:23:54	atlas-swt2.org	ochep.ou.edu	797152257	0	797152257	/atlas/dq2/user/ilijav/HCtest/user.ilijav.HCtest.1/group.test.hc.NTUP_SMWZ.root

Showing 1 to 10 of 76 entries

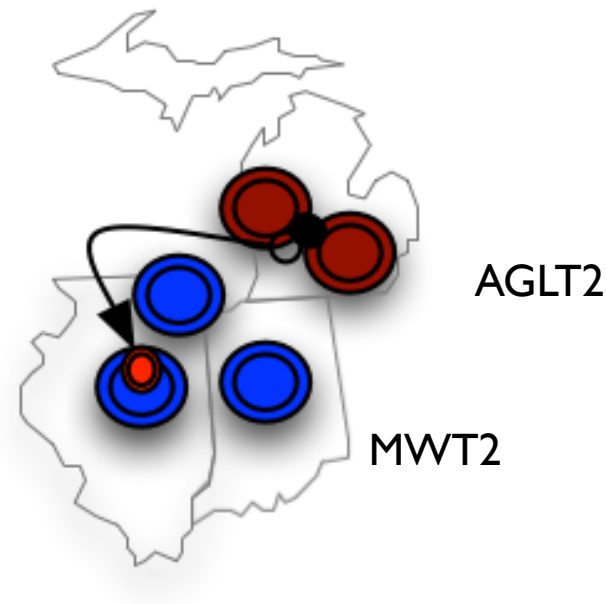
Get source JSON || Resource URL: /eosatl/xrdrealmonfax?

cf. Domenico Giordano's talk

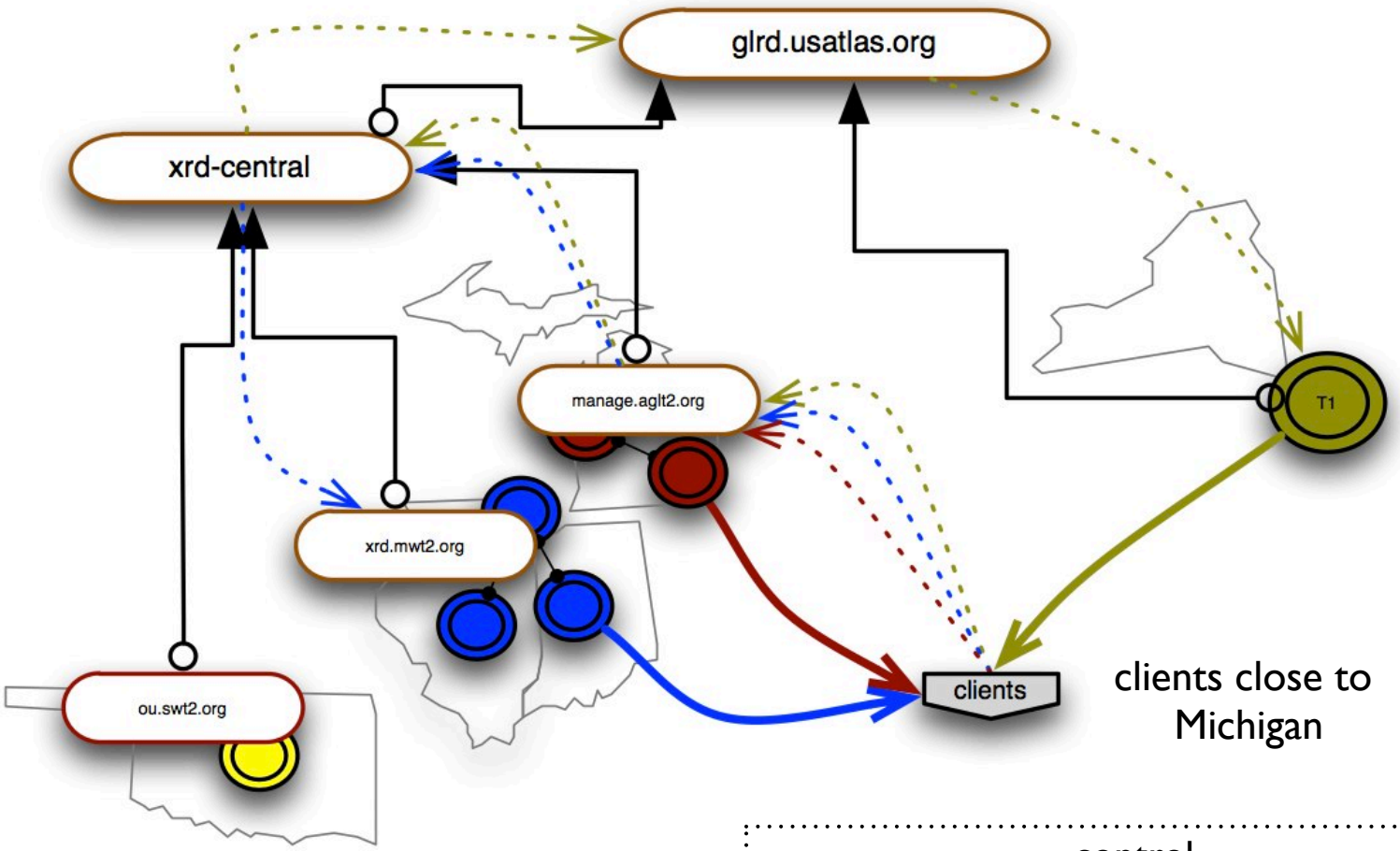
Capabilities: regional redirection

`root://xrd-central.usatlasfacility.org/global-name`

- MWT2 and AGLT2 internally are multi-site federations using dCache (5 total)
- Combined storage ~ 5 PB and about 10K job slots
- ANALY_AGLT2 & ANALY_MWT2 within 7 ms RTT
- Added SWT2 Oklahoma (< 10ms)
- Not found here? Fails over to US cloud level redirector glrd.usatlas.org



Regional redirection



clients close to Michigan

case 2:
client redirected to file on site in region

control
data

Regional WAN analysis: jobs running on one site reading data from itself and two others

PandaID, Owner, Working group	Job	Status	Created	Time to start	Duration	Ended/ Modified	Cloud/Site, Type
1594538009 Robert W. Gardner Jr.	jobsetID=2511 runGen-00-00-02	finished	2012-09-06 07:46	2 days, 1:27:14	0:07:16	09-08 09:20	US/US.ANALY_OU_OCHEP_SWT2, analysis-run
	Out: user.rwg.SMExample_OU_OCHEP_SWT2_xrd-central.001/						

Associated build job: 11141357

Run jobs in this job set: 11141358

Job ran on Oklahoma analysis site
read data from two nearby federations

Panda monitor

Job 1594538009 details

3 files for job 1594538009:

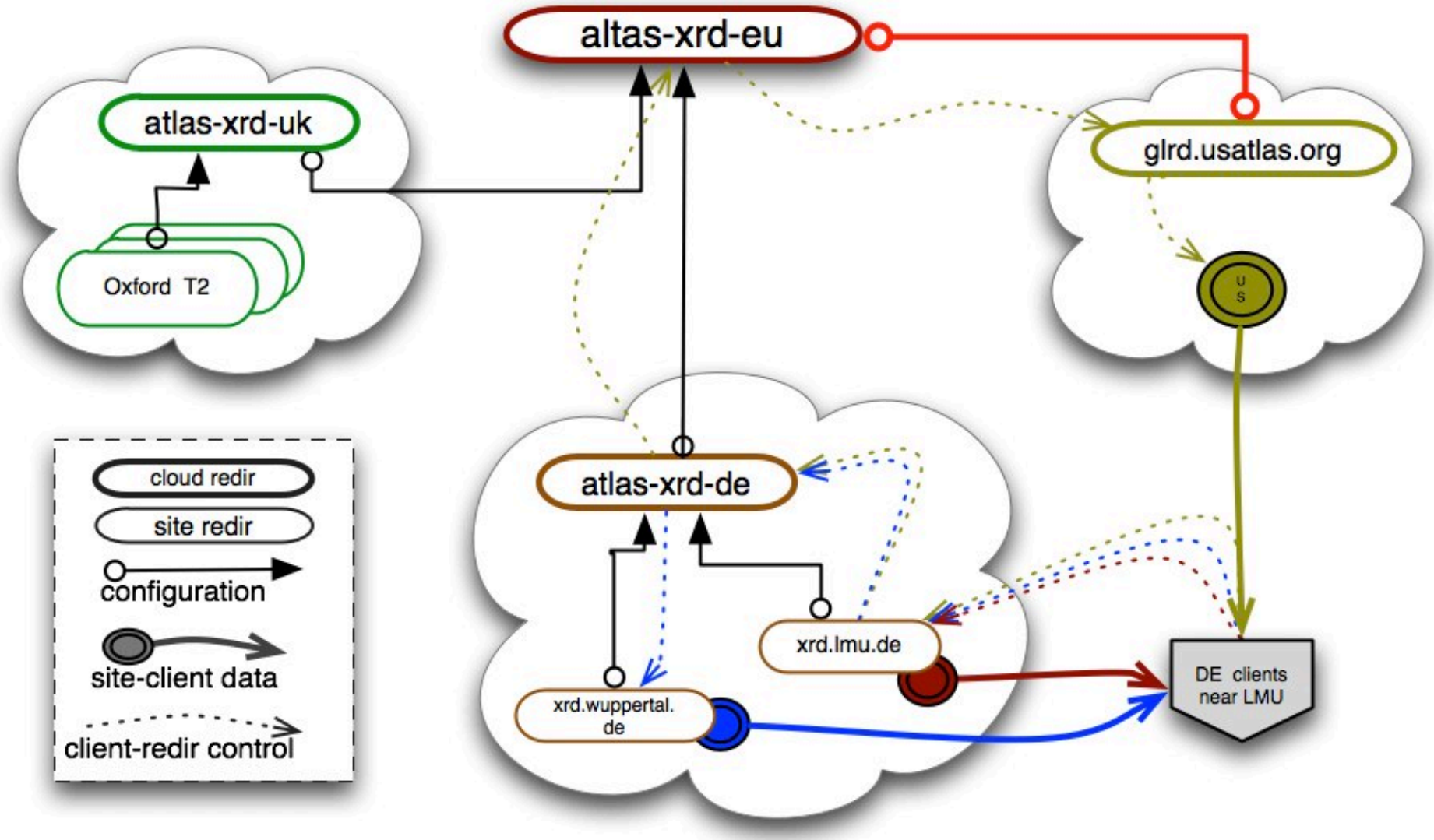
Filename	Type	Status	Dataset
user.rwg.0906074600.297152.lib._002511.lib.tgz guid=0dd239c9-309a-4df4-ad35-2f3552f8ae1b	input	ready	user.rwg.0906074600.297152.lib._002511
user.rwg.002511._1594538009.log.tgz guid=e42d4df2-c6d0-499a-b731-4479938d7351 Space token OU_OCHEP_SWT2_USERDISK	log	ready	user.rwg.SMExample_OU_OCHEP_SWT2_xrd-central.001/ (destination block: user.rwg.SMExample_OU_OCHEP_SWT2_xrd-central.001.1209060246
user.rwg.002511._00025.SMWZExampleXYZ.log.tgz guid=eac1209e-747e-4efe-ab1d-20c7d05bf664 Space token OU_OCHEP_SWT2_USERDISK	output	ready	user.rwg.SMExample_OU_OCHEP_SWT2_xrd-central.001/ (destination block: user.rwg.SMExample_OU_OCHEP_SWT2_xrd-central.001.1209060246

```
.458670 :
xample
connect: 128.135.158.237:1094  cmd: SMWZd3pdExample
connect: 128.135.158.190:1094  cmd: SMWZd3pdExample
connect: 192.41.230.187:1094  cmd: SMWZd3pdExample
connect: 192.41.230.21:1094  cmd: SMWZd3pdExample
connect: 192.41.236.55:23197  cmd: SMWZd3pdExample
connect: 128.135.158.180:1094  cmd: SMWZd3pdExample
2012-09-08 09:14:14.486698 : INFO  connect: 192.41.236.63:23982  cmd: SMWZd3pdExample
2012  : INFO  connect: 128.135.158.252:1094  cmd: SMWZd3pdExample
2012  : INFO  connect: 128.135.158.186:1094  cmd: SMWZd3pdExample
2012  : INFO  connect: 192.41.236.60:24377  cmd: SMWZd3pdExample
2012  : INFO  connect: 192.41.230.23:20398  cmd: SMWZd3pdExample
2012  : INFO  connect: 128.135.158.221:1094  cmd: SMWZd3pdExample
2012  : INFO  connect: 128.135.158.189:1094  cmd: SMWZd3pdExample
2012  : INFO  connect: 192.41.236.60:23501  cmd: SMWZd3pdExample
```

root://xrd-central.usatlasfacility.org//filename

read data from:
MWT2 (128.135.x.x)
AGLT2 (192.41.x.x)
OU (129.15.x.x)

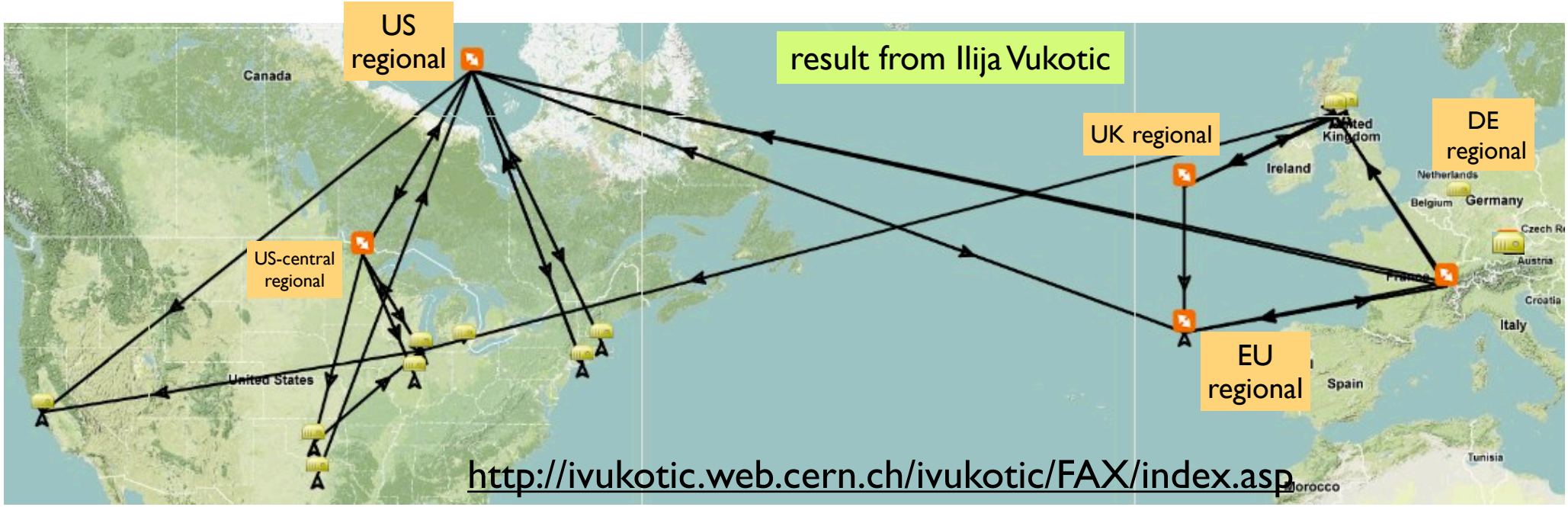
Four levels of redirection: site-cloud-zone-global



Start locally - expand search as needed

Topology validation

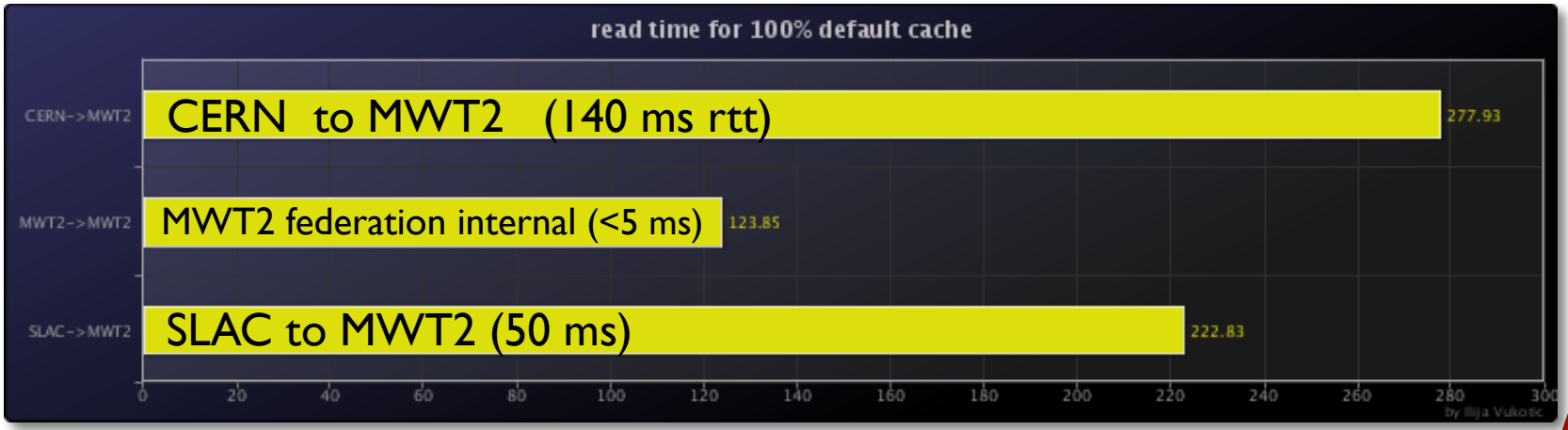
- Launch jobs to every site, test reading of site-specific files at every other site
- Parse client logs to infer resulting redirection

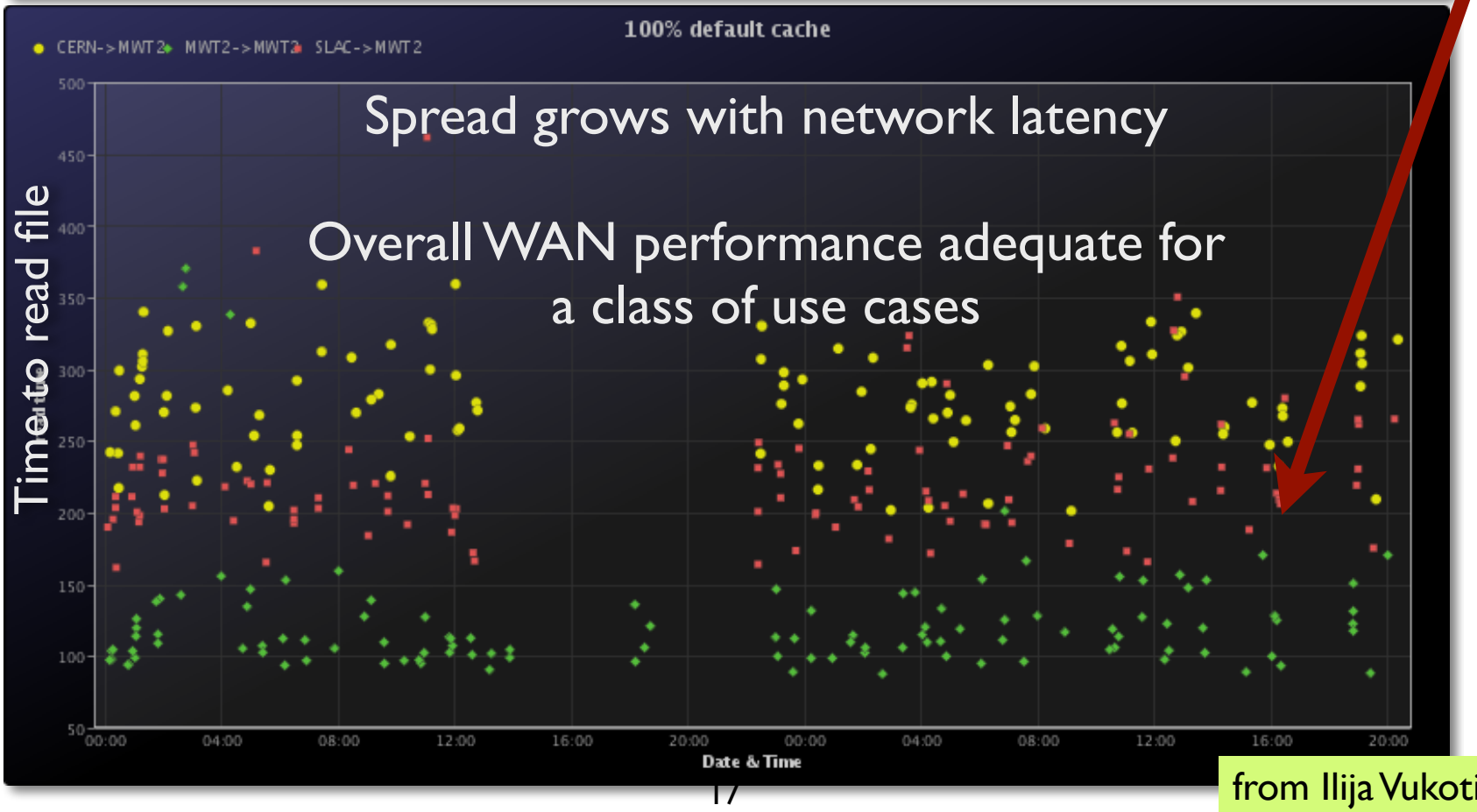


XRD redirector

federating site

WAN Read Tests (basis for "cost matrix")

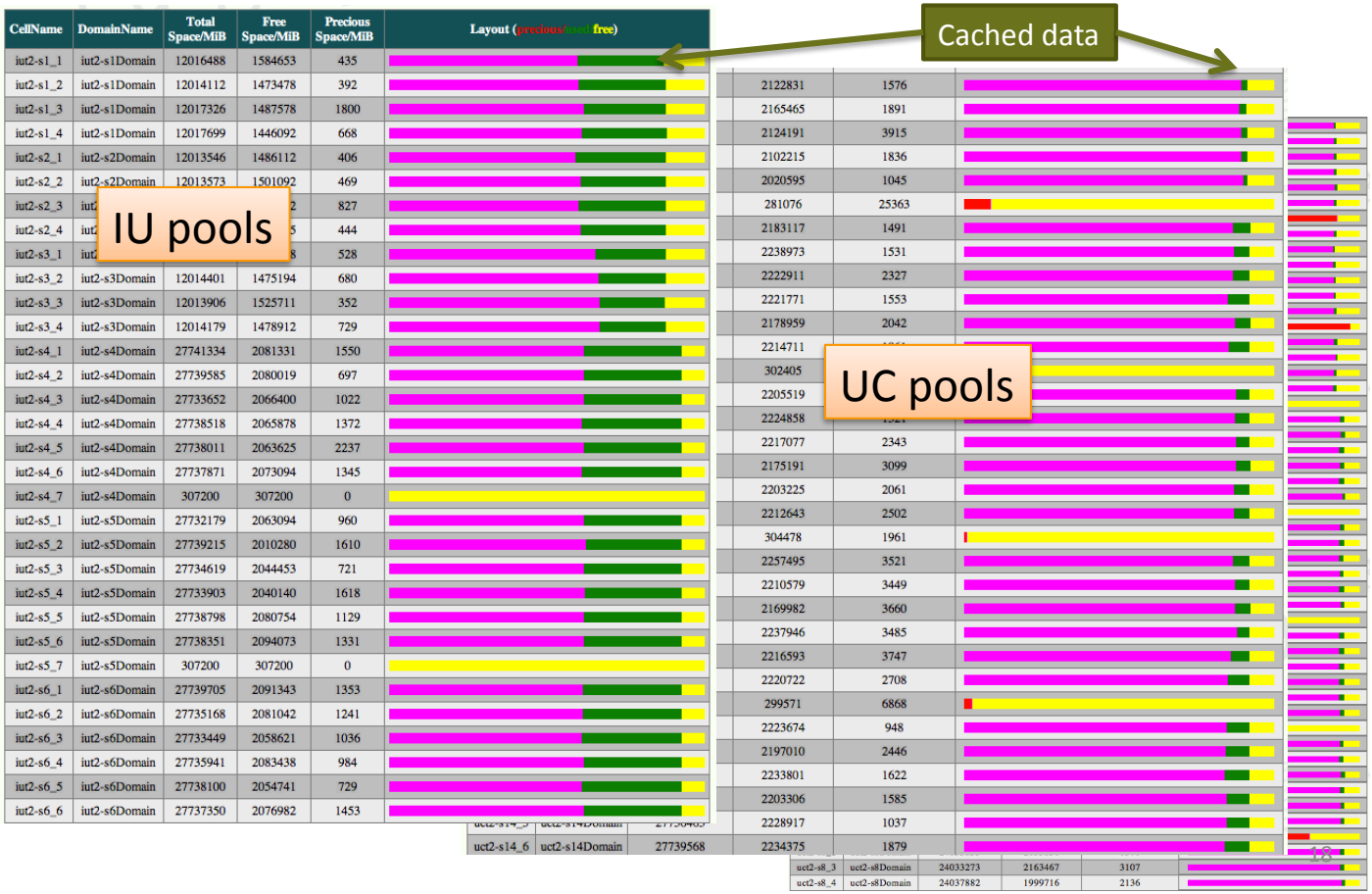




from Ilija Vukotic

WAN direct access versus caching

- MWT2 federation is split between three sites separated by 5 ms (storage & cpu); storage at two of the sites.
- Recently configured dCache's 'locality' caching

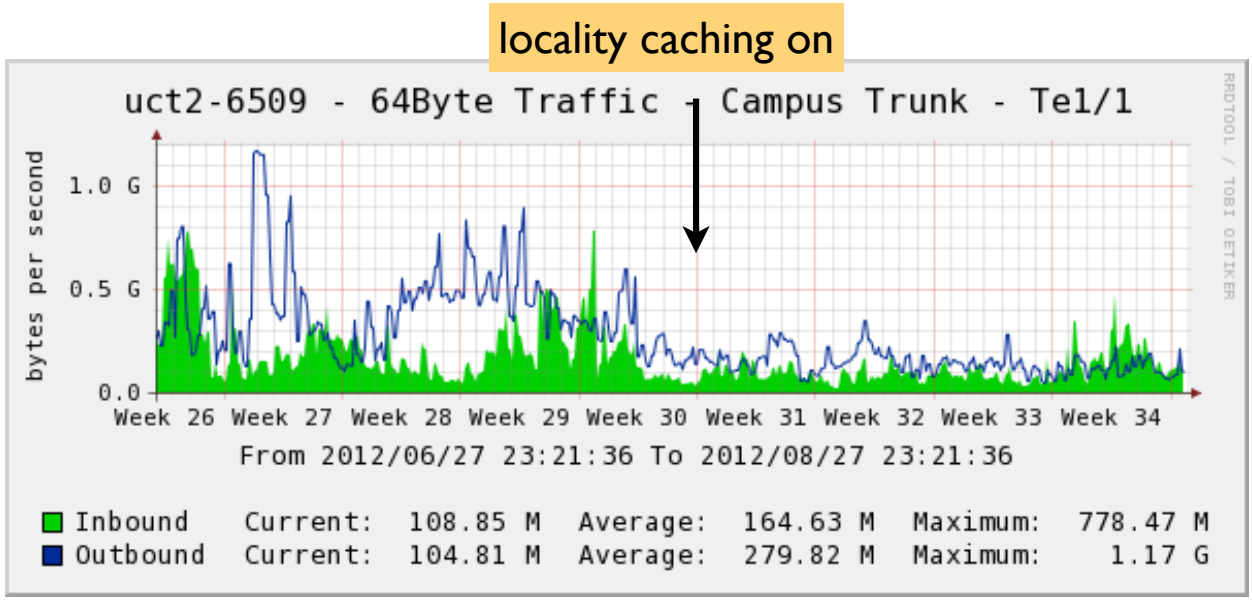


CellName	DomainName	Total Space/MiB	Free Space/MiB	Precious Space/MiB	Layout (precious/cached free)
iut2-s1_1	iut2-s1Domain	12016488	1584653	435	
iut2-s1_2	iut2-s1Domain	12014112	1473478	392	
iut2-s1_3	iut2-s1Domain	12017326	1487578	1800	
iut2-s1_4	iut2-s1Domain	12017699	1446092	668	
iut2-s2_1	iut2-s2Domain	12013546	1486112	406	
iut2-s2_2	iut2-s2Domain	12013573	1501092	469	
iut2-s2_3	iut2		2	827	
iut2-s2_4	iut2		5	444	
iut2-s3_1	iut2		8	528	
iut2-s3_2	iut2-s3Domain	12014401	1475194	680	
iut2-s3_3	iut2-s3Domain	12013906	1525711	352	
iut2-s3_4	iut2-s3Domain	12014179	1478912	729	
iut2-s4_1	iut2-s4Domain	27741334	2081331	1550	
iut2-s4_2	iut2-s4Domain	27739585	2080019	697	
iut2-s4_3	iut2-s4Domain	27733652	2066400	1022	
iut2-s4_4	iut2-s4Domain	27738518	2065878	1372	
iut2-s4_5	iut2-s4Domain	27738011	2063625	2237	
iut2-s4_6	iut2-s4Domain	27737871	2073094	1345	
iut2-s4_7	iut2-s4Domain	307200	307200	0	
iut2-s5_1	iut2-s5Domain	27732179	2063094	960	
iut2-s5_2	iut2-s5Domain	27739215	2010280	1610	
iut2-s5_3	iut2-s5Domain	27734619	2044453	721	
iut2-s5_4	iut2-s5Domain	27733903	2040140	1618	
iut2-s5_5	iut2-s5Domain	27738798	2080754	1129	
iut2-s5_6	iut2-s5Domain	27738351	2094073	1331	
iut2-s5_7	iut2-s5Domain	307200	307200	0	
iut2-s6_1	iut2-s6Domain	27739705	2091343	1353	
iut2-s6_2	iut2-s6Domain	27735168	2081042	1241	
iut2-s6_3	iut2-s6Domain	27733449	2058621	1036	
iut2-s6_4	iut2-s6Domain	27735941	2083438	984	
iut2-s6_5	iut2-s6Domain	27738100	2054741	729	
iut2-s6_6	iut2-s6Domain	27737350	2076982	1453	
uct2-s14_6	uct2-s14Domain	27739568	2234375	1879	
uct2-s8_3	uct2-s8Domain	24033273	2163467	3107	
uct2-s8_4	uct2-s8Domain	24037882	1999716	2136	

2122831	1576
2165465	1891
2124191	3915
2102215	1836
2020595	1045
281076	25363
2183117	1491
2238973	1531
2222911	2327
2221771	1553
2178959	2042
2214711	
302405	
2205519	
2224858	
2217077	2343
2175191	3099
2203225	2061
2212643	2502
304478	1961
2257495	3521
2210579	3449
2169982	3660
2237946	3485
2216593	3747
2220722	2708
299571	6868
2223674	948
2197010	2446
2233801	1622
2203306	1585
2228917	1037

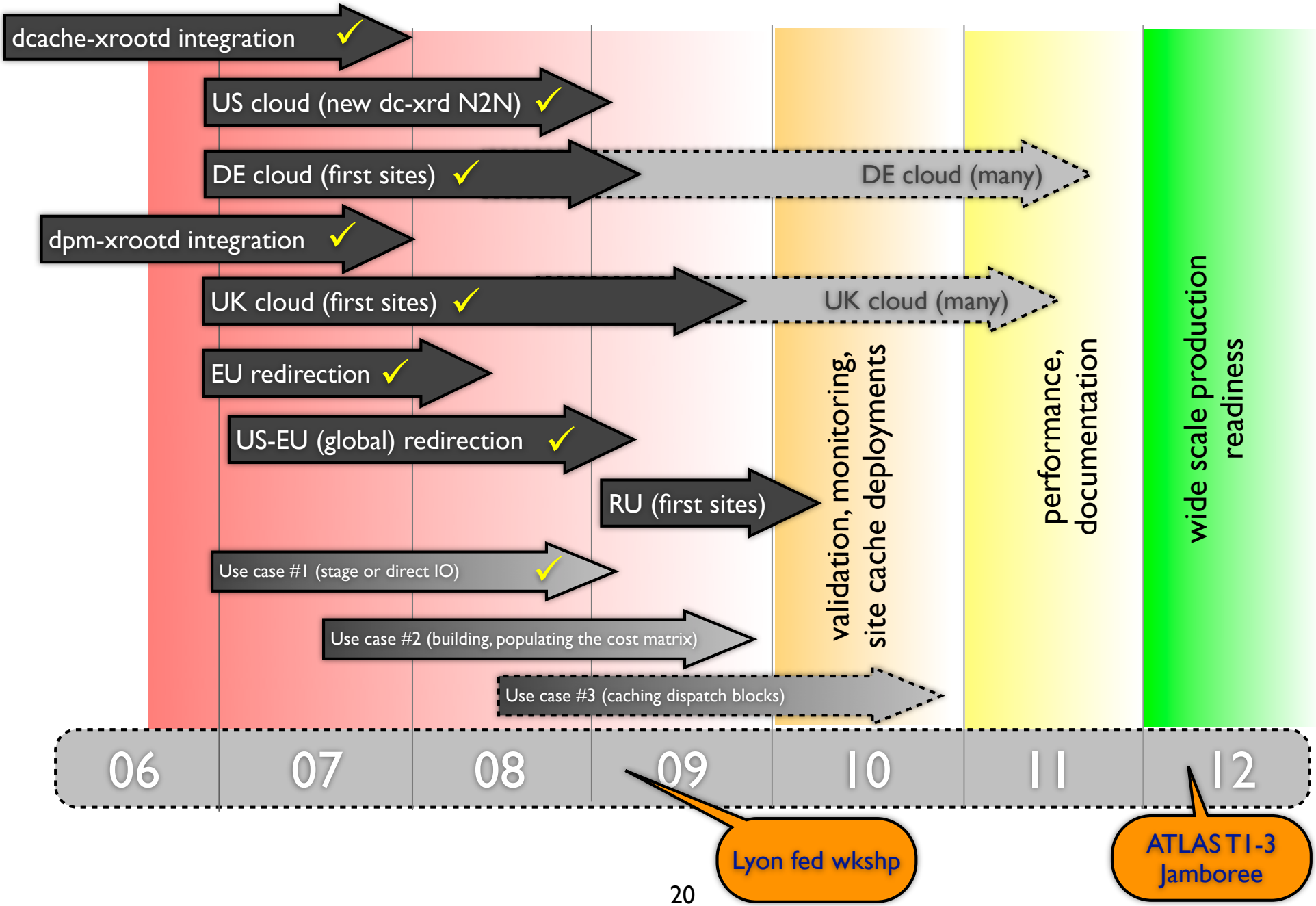
WAN direct access versus caching

- Expect WAN traffic reduced provided there is re-use and non-trivial fraction of file read
- Evidence of reduced WAN traffic
- Suggests equipping sites with federated caches, alongside managed storage



WAN IO

ATLAS timeline



Summary

- Federated xrootd infrastructure for ATLAS now an ATLAS-wide project - presence in three clouds plus EOS
- Regional to global topology testing (up & down the trees)
- Accessible from Panda analysis jobs, either direct access or stage-in from pilot
- In parallel much work in ATLAS to optimize analysis code for wide area reads
- While **not yet in production** ATLAS is on a good course to provide new capabilities to users using federating Xrootd technologies