Status of Federated Xrootd in ATLAS

R. Gardner

19-Mar-12

FAX Working Group

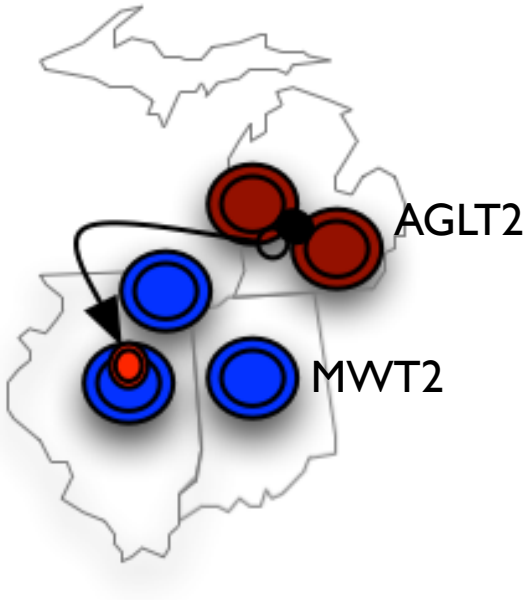
- Led by Wei Yang, R.Gardner (US facilities integration)
- Doug Benjamin
- Andy Hanushevsky
- Hiro Ito
- Patrick McGuigan
- Shawn McKee
- Ofer Rind
- Horst Severini
- Sarah Williams
- Meet bi-weekly
- Next workshop coming up April 11-12, Chicago

Motivations for federation

- Provide transparent **read** access to remote data from any compute server
- Reduce local storage and data management requirements for Tier3 clusters or from within Cloud resources
- More efficienct utilization of storage & cpu resources

Example regional case

- The 5 sites in ANALY_AGLT2 & ANALY_MWT2 are all within 7 ms RTT
- ANALY_MWT2 is already an (internally) federated 3-site wide area queue
- Combined storage ~ 4.4 PB
- A regional redirector would allow sharing of datastets between AGLT2 and MWT2

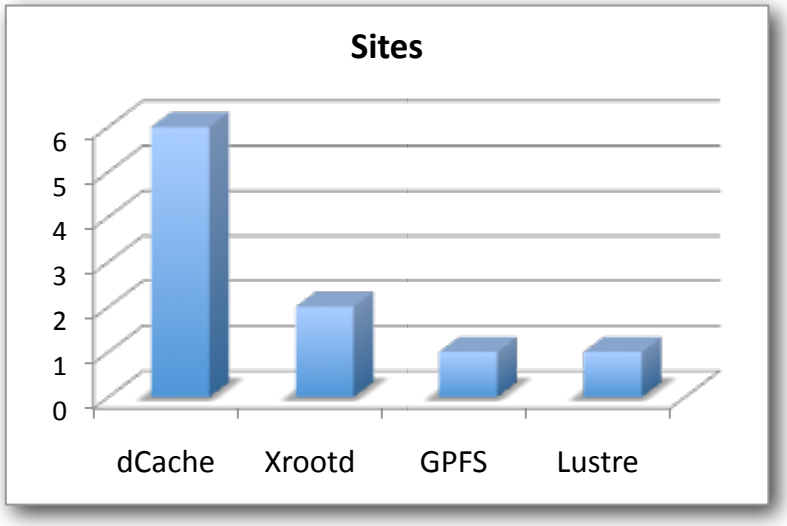


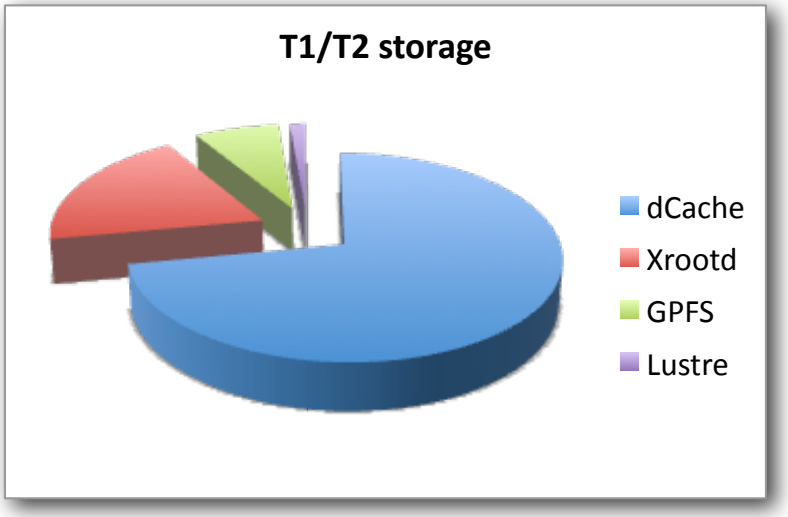
Federating data stores in the US Cloud

- Storage resources at T1 and 5 Tier2 centers (10 sites) curently total 17.7 PB disk
- Three Tier2 centers are mulit-site and share distributed storage resources across WAN (AGLT2, MWT2, NET2)
- O(20) analysis T3g sites: some would federate as sources, others would use the federation as clients

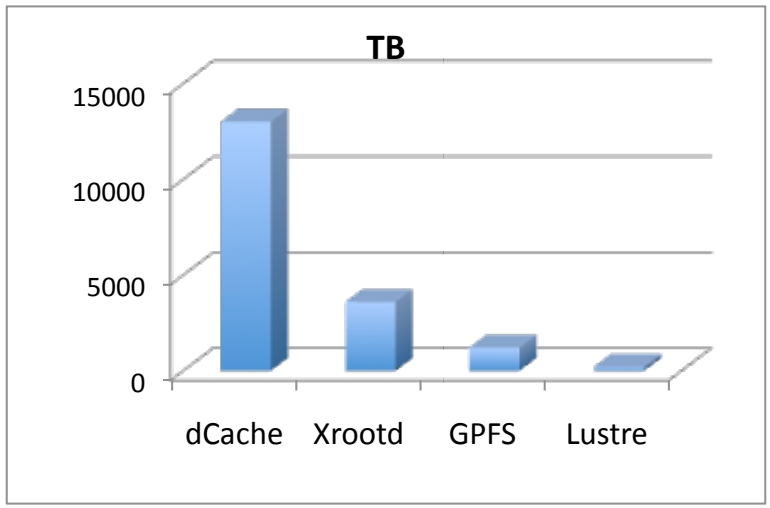
backend profile in US

(T1, T2 sites)



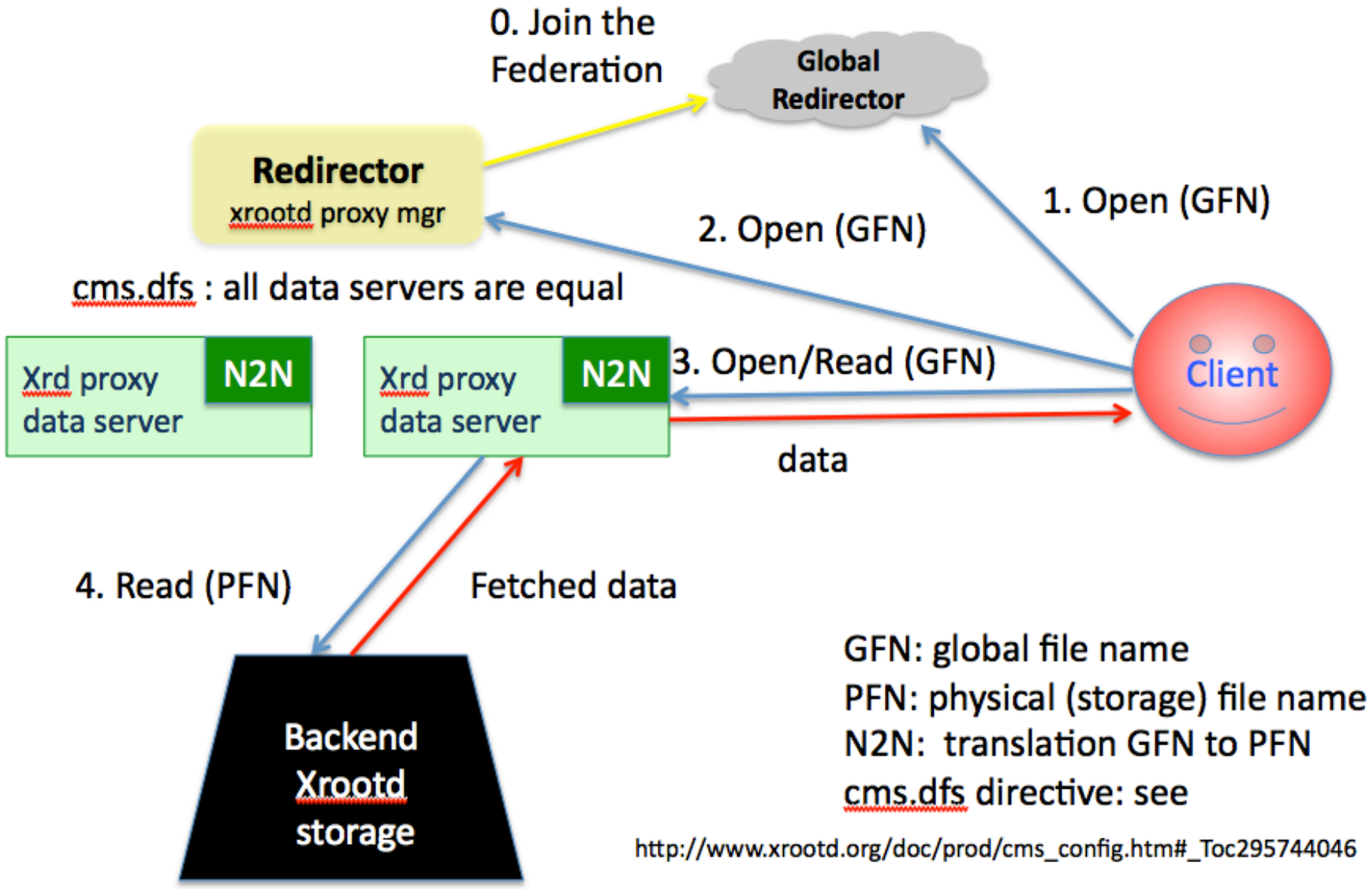


- ~13 PB in dCache
- ~3.5 PB in Xrootd
- By April:
  - 2.2 PB at each T2
  - 8.1 PB at T1
  - ~ 17.7 PB total



Export Xrootd storage via Xrootd Proxy cluster

SLAC, SWT2

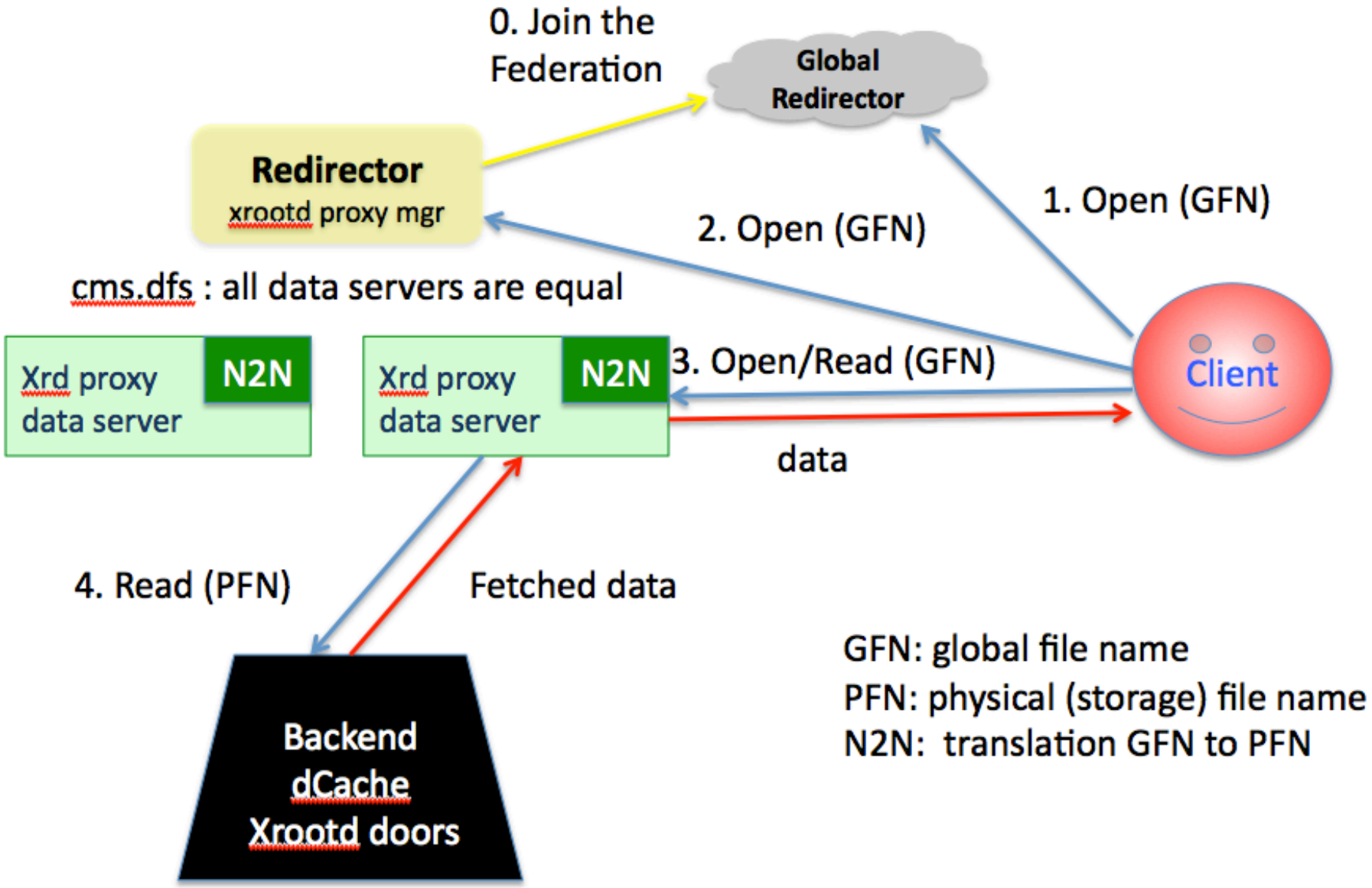


GFN: global file name
PFN: physical (storage) file name
N2N: translation GFN to PFN
cms.dfs directive: see

http://www.xrootd.org/doc/prod/cms_config.htm#_Toc295744046

Export dCache Xrootd doors via Xrootd Proxy cluster, 1

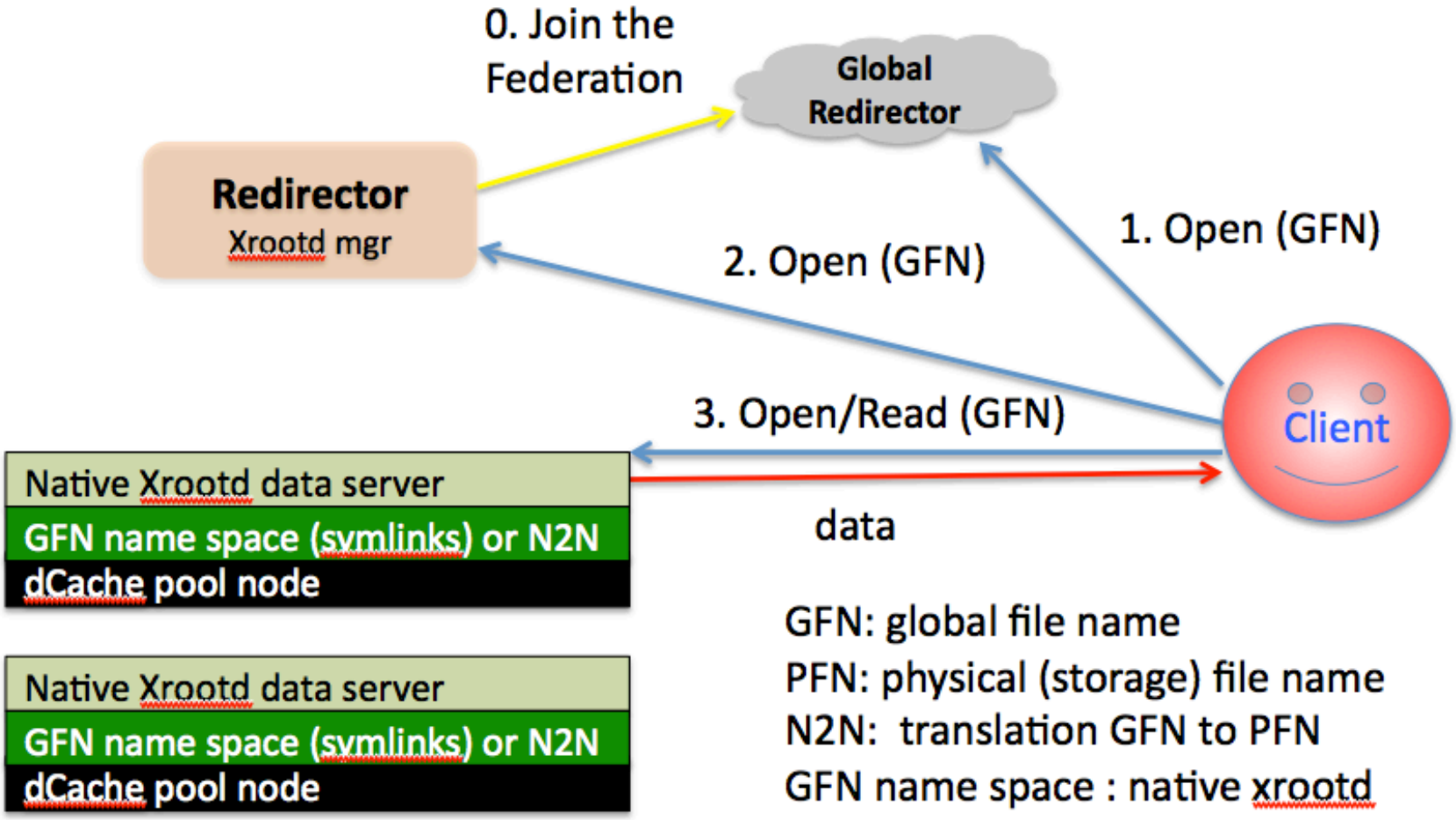
BNL, AGLT2



GFN: global file name
PFN: physical (storage) file name
N2N: translation GFN to PFN

Overlapping Xrootd cluster on top of dCache

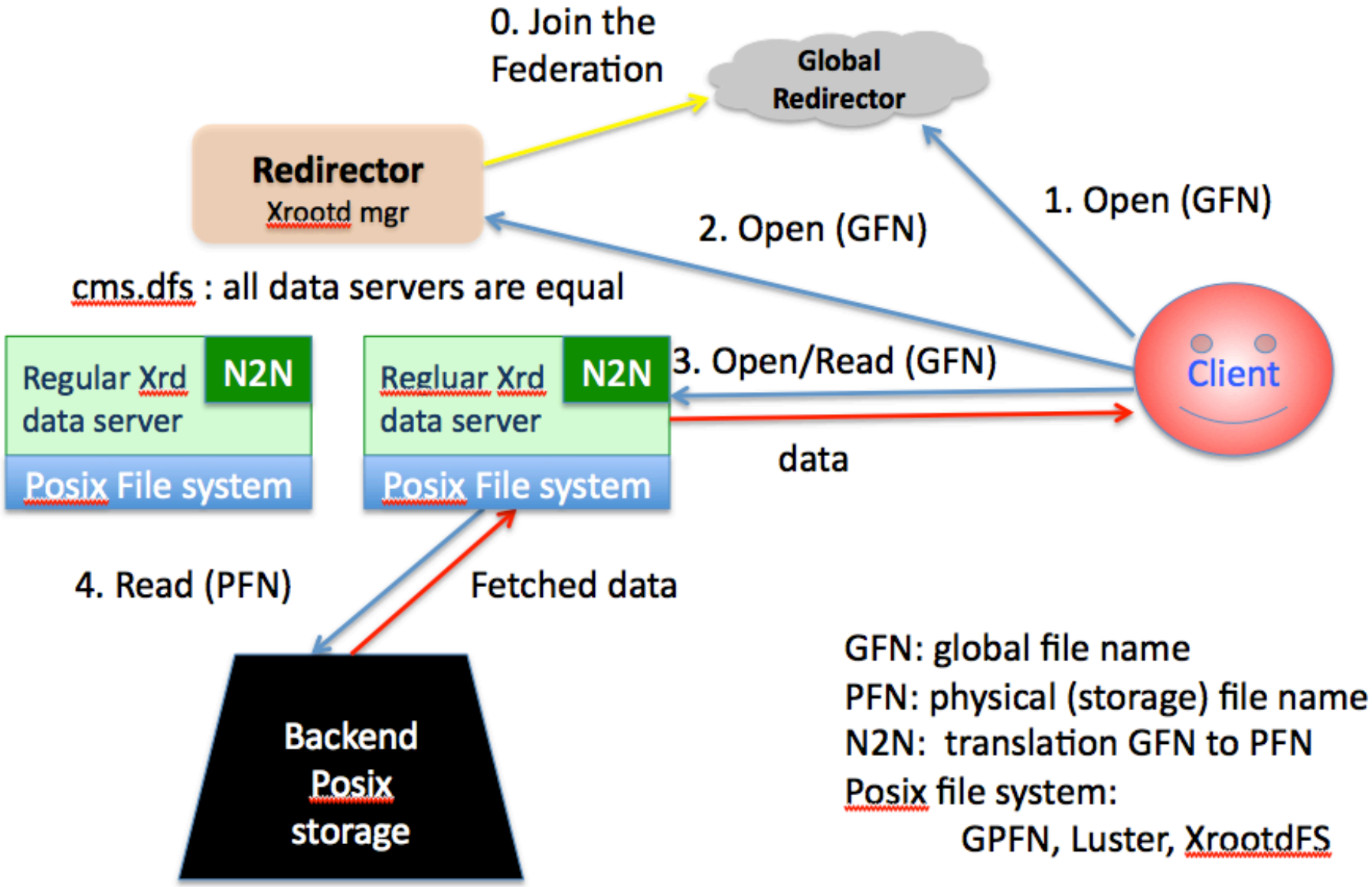
MWT2



GFN: global file name
PFN: physical (storage) file name
N2N: translation GFN to PFN
GFN name space : native xrootd
name space layout according to GFN, with symlinks pointing to actual file on the same node
(some CMS sites do this)

Export Posix storage via regular Xrootd cluster

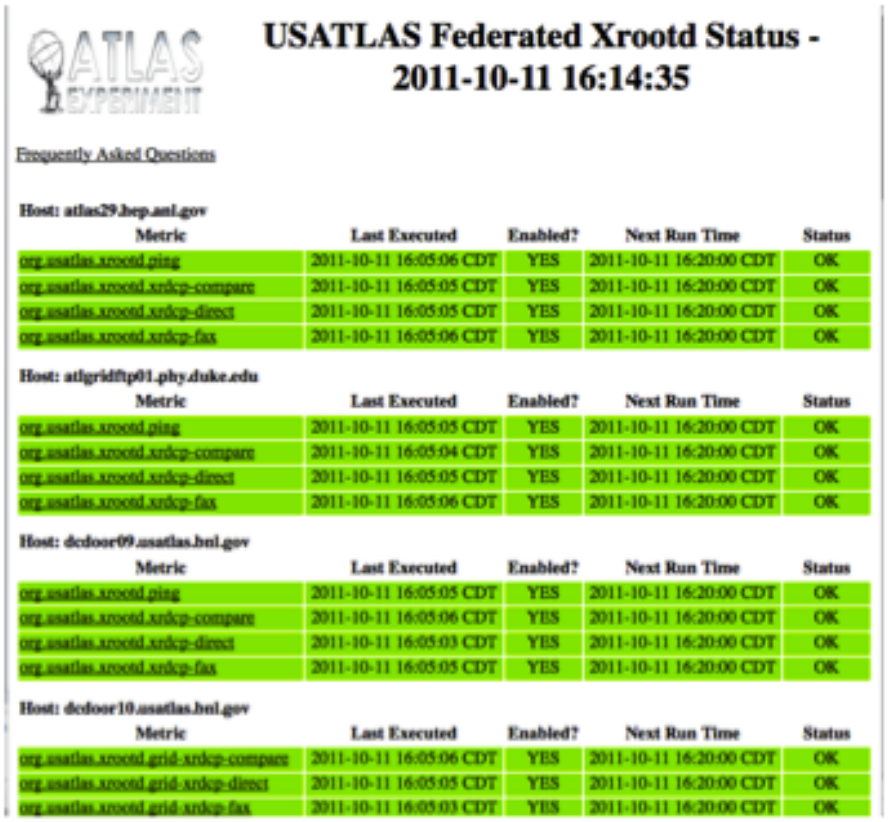
SWT2_OU, BU



GFN: global file name
PFN: physical (storage) file name
N2N: translation GFN to PFN
Posix file system:
GPFN, Luster, XrootdFS

Deployment Status

- http://uct3-xrdp.uchicago.edu:8080/rsv/
- based on OSG monitoring framework
- Probes sites every 15 minutes
- Tests direct transfers and via global redirector
- Also does simple ping and file comparision checks



USATLAS Federated Xrootd Status - 2011-10-11 16:14:35

Frequently Asked Questions

Host: atlas29.hep.anl.gov

Metric	Last Executed	Enabled?	Next Run Time	Status
org.usatlas.xrootd.ping	2011-10-11 16:05:06 CDT	YES	2011-10-11 16:20:00 CDT	OK
org.usatlas.xrootd.xrdcp-compare	2011-10-11 16:05:05 CDT	YES	2011-10-11 16:20:00 CDT	OK
org.usatlas.xrootd.xrdcp-direct	2011-10-11 16:05:05 CDT	YES	2011-10-11 16:20:00 CDT	OK
org.usatlas.xrootd.xrdcp-fax	2011-10-11 16:05:06 CDT	YES	2011-10-11 16:20:00 CDT	OK

Host: atlgridftp01.phy.duke.edu

Metric	Last Executed	Enabled?	Next Run Time	Status
org.usatlas.xrootd.ping	2011-10-11 16:05:05 CDT	YES	2011-10-11 16:20:00 CDT	OK
org.usatlas.xrootd.xrdcp-compare	2011-10-11 16:05:04 CDT	YES	2011-10-11 16:20:00 CDT	OK
org.usatlas.xrootd.xrdcp-direct	2011-10-11 16:05:05 CDT	YES	2011-10-11 16:20:00 CDT	OK
org.usatlas.xrootd.xrdcp-fax	2011-10-11 16:05:06 CDT	YES	2011-10-11 16:20:00 CDT	OK

Host: dcdoor09.usatlas.bnl.gov

Metric	Last Executed	Enabled?	Next Run Time	Status
org.usatlas.xrootd.ping	2011-10-11 16:05:05 CDT	YES	2011-10-11 16:20:00 CDT	OK
org.usatlas.xrootd.xrdcp-compare	2011-10-11 16:05:06 CDT	YES	2011-10-11 16:20:00 CDT	OK
org.usatlas.xrootd.xrdcp-direct	2011-10-11 16:05:03 CDT	YES	2011-10-11 16:20:00 CDT	OK
org.usatlas.xrootd.xrdcp-fax	2011-10-11 16:05:05 CDT	YES	2011-10-11 16:20:00 CDT	OK

Host: dcdoor10.usatlas.bnl.gov

Metric	Last Executed	Enabled?	Next Run Time	Status
org.usatlas.xrootd.grid-xrdcp-compare	2011-10-11 16:05:06 CDT	YES	2011-10-11 16:20:00 CDT	OK
org.usatlas.xrootd.grid-xrdcp-direct	2011-10-11 16:05:05 CDT	YES	2011-10-11 16:20:00 CDT	OK
org.usatlas.xrootd.grid-xrdcp-fax	2011-10-11 16:05:03 CDT	YES	2011-10-11 16:20:00 CDT	OK

7

adopting CMS-like monitor

- Courtesy Matevz Tadel UCSD (thank you!)
- Detailed xrootd monitoring information sent to collector
- Tracks files (global names) in use, when opened, server, client, MB read
- Provides IO visibility in federation

OpenAgo	ServerDomain	ClientDomain	User	Read [MB]	UpdateAgo
02:38:26	slac.stanford.edu	23.40.189		347.835	00:00:05
00:04:05	slac.stanford.edu	uchicago.edu		11.003	00:02:54
00:04:05	slac.stanford.edu	uchicago.edu		11.008	00:02:53
00:02:53	slac.stanford.edu	uchicago.edu		11.126	00:01:48
00:02:53	slac.stanford.edu	uchicago.edu		11.020	00:01:51
00:01:49	slac.stanford.edu	uchicago.edu		10.982	00:00:46
00:01:49	slac.stanford.edu	uchicago.edu		11.125	00:00:48
00:01:49	slac.stanford.edu	uchicago.edu		11.125	00:00:49
00:01:49	slac.stanford.edu	uchicago.edu		11.125	00:00:48
					00:00:-6
					00:00:-7
					00:00:-7
					00:00:-4

File
/atlas/xrootd/atlasdatadisk/data11_7TeV/AOD/r260_p659_tid493619_00/AOD.493619._000001.pool.root.1
/atlas/dq2/mc10_7TeV/NTUP_SMWZ/e773_s933_5_WW2lep.merge.NTUP_SMWZ.e773_s933_s946_r2302_r2300_p591_tid408566_00/NTUP_SMWZ.408566._000001.root.1
/atlas/dq2/mc10_7TeV/NTUP_SMWZ/e773_s933_5_WW2lep.merge.NTUP_SMWZ.e773_s933_s946_r2302_r2300_p591_tid408566_00/NTUP_SMWZ.408566._000001.root.1
/atlas/dq2/mc10_7TeV/NTUP_SMWZ/e773_s933_5_WW2lep.merge.NTUP_SMWZ.e773_s933_s946_r2302_r2300_p591_tid408566_00/NTUP_SMWZ.408566._000002.root.1
/atlas/dq2/mc10_7TeV/NTUP_SMWZ/e773_s933_5_WW2lep.merge.NTUP_SMWZ.e773_s933_s946_r2302_r2300_p591_tid408566_00/NTUP_SMWZ.408566._000002.root.1
/atlas/dq2/mc10_7TeV/NTUP_SMWZ/e773_s933_5_WW2lep.merge.NTUP_SMWZ.e773_s933_s946_r2302_r2300_p591_tid408566_00/NTUP_SMWZ.408566._000003.root.1
/atlas/dq2/mc10_7TeV/NTUP_SMWZ/e773_s933_5_WW2lep.merge.NTUP_SMWZ.e773_s933_s946_r2302_r2300_p591_tid408566_00/NTUP_SMWZ.408566._000003.root.1
/atlas/dq2/mc10_7TeV/NTUP_SMWZ/e773_s933_5_WW2lep.merge.NTUP_SMWZ.e773_s933_s946_r2302_r2300_p591_tid408566_00/NTUP_SMWZ.408566._000003.root.1
/atlas/dq2/mc10_7TeV/NTUP_SMWZ/e773_s933_5_WW2lep.merge.NTUP_SMWZ.e773_s933_s946_r2302_r2300_p591_tid408566_00/NTUP_SMWZ.408566._000003.root.1
/atlas/dq2/mc10_7TeV/NTUP_SMWZ/e773_s933_5_WW2lep.merge.NTUP_SMWZ.e773_s933_s946_r2302_r2300_p591_tid408566_00/NTUP_SMWZ.408566._000004.root.1
/atlas/dq2/mc10_7TeV/NTUP_SMWZ/e773_s933_5_WW2lep.merge.NTUP_SMWZ.e773_s933_s946_r2302_r2300_p591_tid408566_00/NTUP_SMWZ.408566._000004.root.1
/atlas/dq2/mc10_7TeV/NTUP_SMWZ/e773_s933_5_WW2lep.merge.NTUP_SMWZ.e773_s933_s946_r2302_r2300_p591_tid408566_00/NTUP_SMWZ.408566._000004.root.1
/atlas/dq2/mc10_7TeV/NTUP_SMWZ/e773_s933_5_WW2lep.merge.NTUP_SMWZ.e773_s933_s946_r2302_r2300_p591_tid408566_00/NTUP_SMWZ.408566._000004.root.1

TTreeCache WAN tests (1)

- Standard model analysis over ntuple datasets by D. Benjamin show good results local versus remote (Argonne to BNL)
- Systematically measure walltime efficiency for reads between sites & determine optimal TTreeCache options

WAN
CPUTIME=1168
WALLTIME=2744
eff = 43%

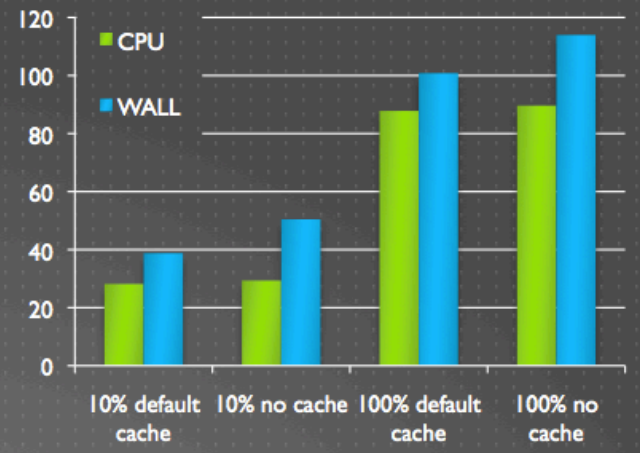
Local xrootd
CPUTIME=1150
WALLTIME=1442
eff=80%

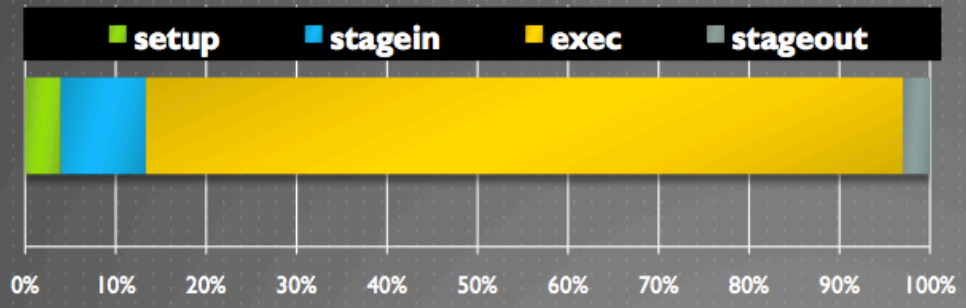
TTreeCache studies: read vs analysis

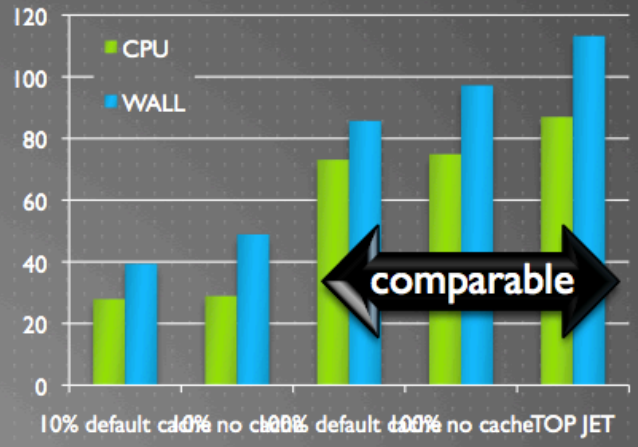
local IO testing ...

RESULT – EFFICIENCY

- Average results over all the sites during last month using 17.0.4 (ROOT 5.28)
- 77% Event loop CPU efficiency
- Total job CPU efficiency 41%







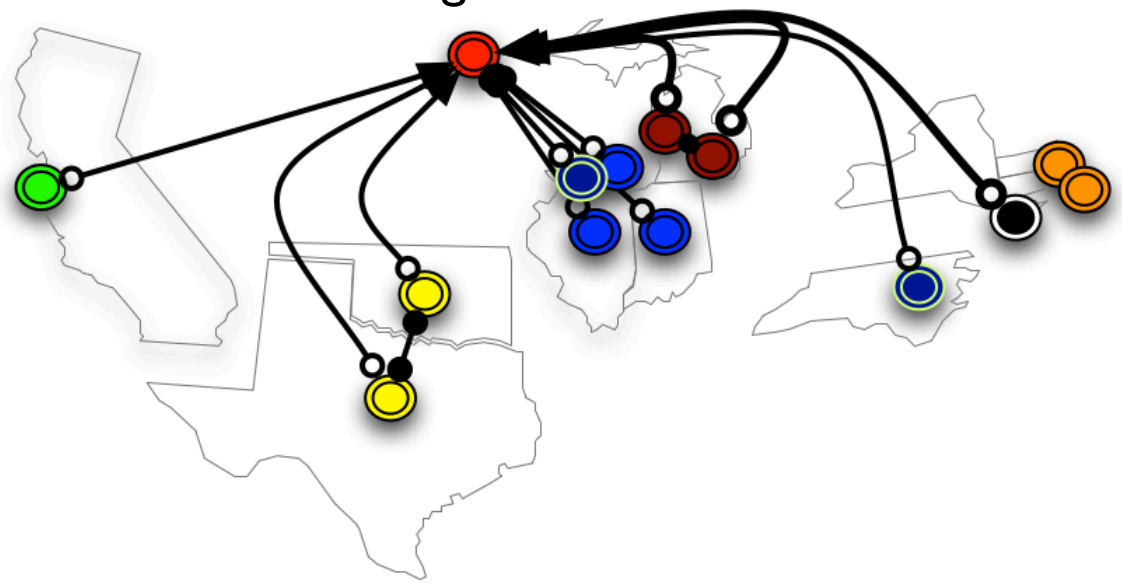
Realistic analysis

Latencies in the federation

..to remote IO testing

ep.anl.gov	4.0 ms
001.phy.duke.edu	25.2 ms
9.usatlas.bnl.gov	24.6 ms
0.usatlas.bnl.gov	35-42 ms
s-swt2.org	25-36 ms
las.org	35-42 ms
Host: griddev01.slac.stanford.edu	52 ms
Host: griddev02.slac.stanford.edu	52 ms
Host: gw01.tier3-atlas.uta.edu	22 ms
Host: itb3.uchicago.edu	< 1ms
Host: manage.aglt2.org	6.7 ms
Host: osgx3.hep.uiuc.edu	3.38 ms
Host: ouhep03.nhn.ou.edu	19.2 ms
Host: tier2-03.ochep.ou.edu	19.5 ms
Host: uct2-grid5.uchicago.edu	< 1ms
Host: uct3-xrdp.uchicago.edu	< 1ms
Host: lxplus.cern.ch	140 ms

latencies seen from a root client in Chicago



TTreeCache WAN tests (2)

- Investigate efficiency varying %events read and TTreeCache size
- Steady improvement with buffer size
- With large enough buffers 80% to ~50% wall time efficiency

Client: *.uchicago.edu				
	% events read (30MB buffer)			100 MB buffer
Server	10%	50%	100%	100%
SLAC	WALLTIME=35.8	WALLTIME=74.5	WALLTIME=105.9	WALLTIME=76.0
	CPUTIME=11.9	CPUTIME=25.12	CPUTIME=41.57	CPUTIME=41.78
BNL	WALLTIME=28.2	WALLTIME=61.6	WALLTIME=87.8	WALLTIME=62.3
	CPUTIME=12.01	CPUTIME=25.27	CPUTIME=45.66	CPUTIME=41.69
SWT2-UTA	WALLTIME=28.1	WALLTIME=40.9	WALLTIME=66.78	WALLTIME=56.4
	CPUTIME=12.06	CPUTIME=22.6	CPUTIME=41.69	CPUTIME=41.78
AGLT2	WALLTIME=25.4	WALLTIME=45.0	WALLTIME=58.5	WALLTIME=49.5
	CPUTIME=11.9	CPUTIME=25.3	CPUTIME=44	CPUTIME=41.65
MWT2	WALLTIME=18.8	WALLTIME=29.4	WALLTIME=48.6	WALLTIME=46.2
	CPUTIME=11.93	CPUTIME=25.2	CPUTIME=44	CPUTIME=42.11

Summary of direct access federation testing

Mode	Tested	Type	relative performance
• T3 access to other T3s via global name		local script	Limited testing so far - T3gs tested
• T3 access to T2 via global name		local script	Good
• T3 access to T1 via global name		local script	Good
• T2 access to itself via global name		local script, Panda	Excellent
• T2 access to T1 via global name		local script, Panda	Good
• T2 access to other T2 via global name		Panda	Varies
• T2 access to T3 in the federation		local script, Panda(!)	Can be good

- Many access modes are possible, and with appropriate TTreeCache settings performance can be good enough to compliment local access

Other work

- Xrootd 3.1.1 now deployed at most US sites
- X509 VOMS mapping now available - implementing on sites
- dq2-client 1.0 supports the global file name
- git repo for sharing configurations
- Improvements to N2N methods working well
  - requires LFC lookup so potentiall an issue with consolidation, to be tested
- Focus has been on direct reading over WAN - but in future we know stage-in/caching will be important

R&D to production?

- Subject the current set of sites to regular testing at signficant analysis job scale (in HammerCloud)
- Provide redirector of highly performing data sources
- More experience with TTreeCache settings with well defined examples for users
- Explore augmenting current ANALY workflow to use FAX when problems with local SE or missing files (Athena or lsm, eg.)
  - or to expand number of queues available to users (no local input dataset requirement)
- Other regions in ATLAS are interested in trying out federation

Conclusions

- An R&D federated xrootd has been deployed over production storage resources over a large region
- Some Tier3's are using T1 and T2 redirectors directly in production (BNL and UC)
- Wide adoption would be indicated so long as WAN performance is decent enough