

Using Xrootd to Federate Regional Storage

Presentation by Brian Bockelman
Representing work done by many others:

Lothar Bauerdick, Doug Benjamin, Ken Bloom, Brian Bockelman, Dan Bradley, Sridhara Dasu, Michael Ernst, Rob Gardner, Andy Hanushevsky, Hironori Ito, David Lesny, Patrick McGuigan, Shawn McKee, Ofer Rind, Horst Severini, Igor Sfiligoi, Matevz Tadel, Ilija Vukotic, Sarah Williams, Frank Würthwein, Avi Yagil, Wei Yang

Introductions

Scale Up, Scale Down

- The LHC experiments have thoroughly demonstrated the ability to *scale up*.
- Combined, we run around 100k cores a day, delivering petabytes to the local cluster, and about a petabyte over the WAN.
- What's questionable is our ability to *scale down* to the needs of a single physicist.

The Woe of One Event

- If I want to read a single event, how do I get the data? Options:

- **Run a grid job on that event:** best case, 15 minutes (create the job, submit it, have it run, fetch results). Worst case, hours of queue time.

Events per second when jobs are running can be impressive - but the overhead kills things when looking at a single event!

Download the file: First you have to find it and setup the tools. If you're lucky, only 5 minutes to download.

Direct Remote Access is Key

- We turned to the Xrootd project to provide remote, direct access to data stored at sites.
- Mature project for remote-I/O.
- Client almost always integrated into ROOT.
- Has the security mechanisms WLCG needs.
- Time to open event interactively is limited to network latency.

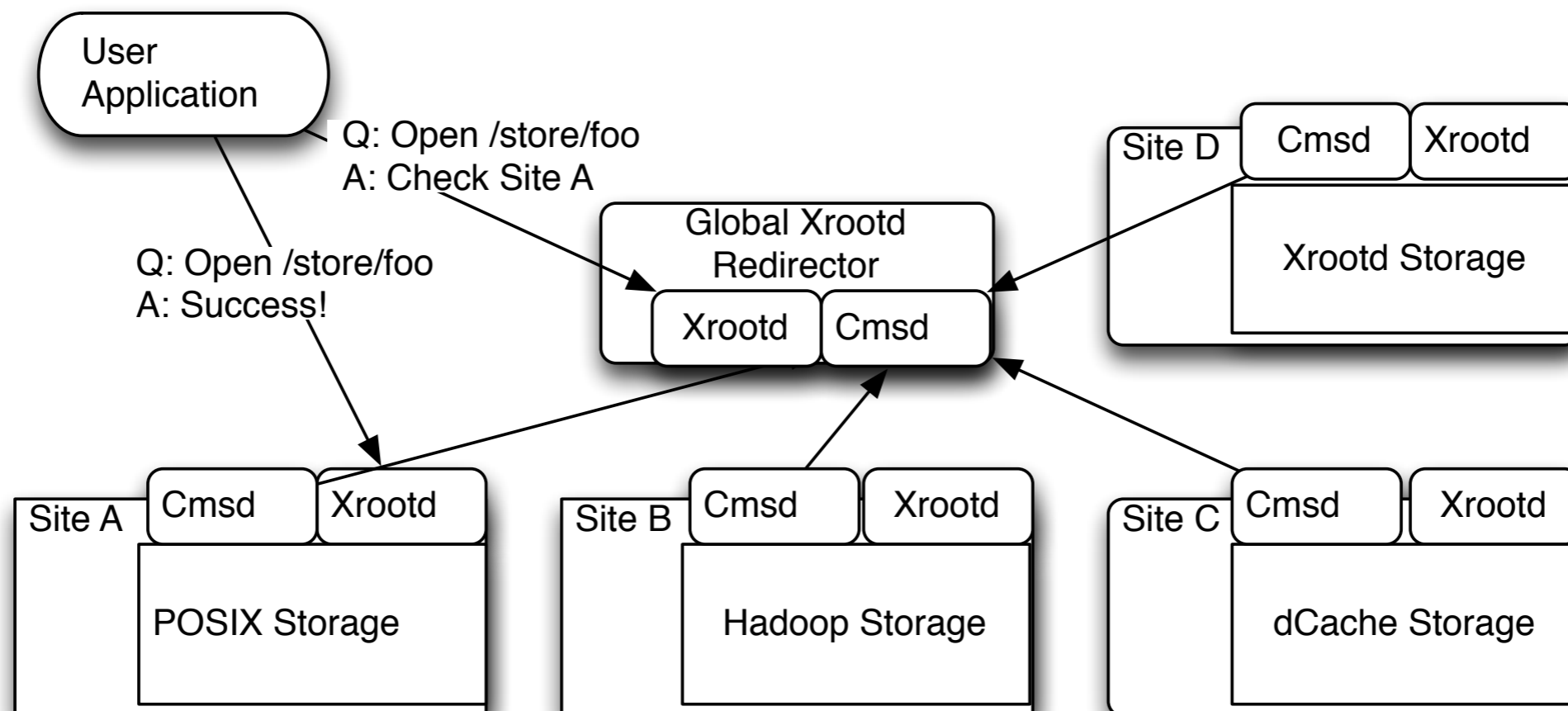
Introducing Federations

- Remote access gives us data for *one* site. We need a federation to access all sites.
- Definition of a **federated storage system***:
 - A collection of disparate storage resources managed by cooperating but independent administrative domains transparently accessible via a common namespace.

* From the Lyon workshop on Federated Data Stores: <http://indico.in2p3.fr/conferenceProgram.py?confId=5527>

Federating Xrootd

- The simplest kind of federation is illustrated below:



Federation overlays on top of existing storage

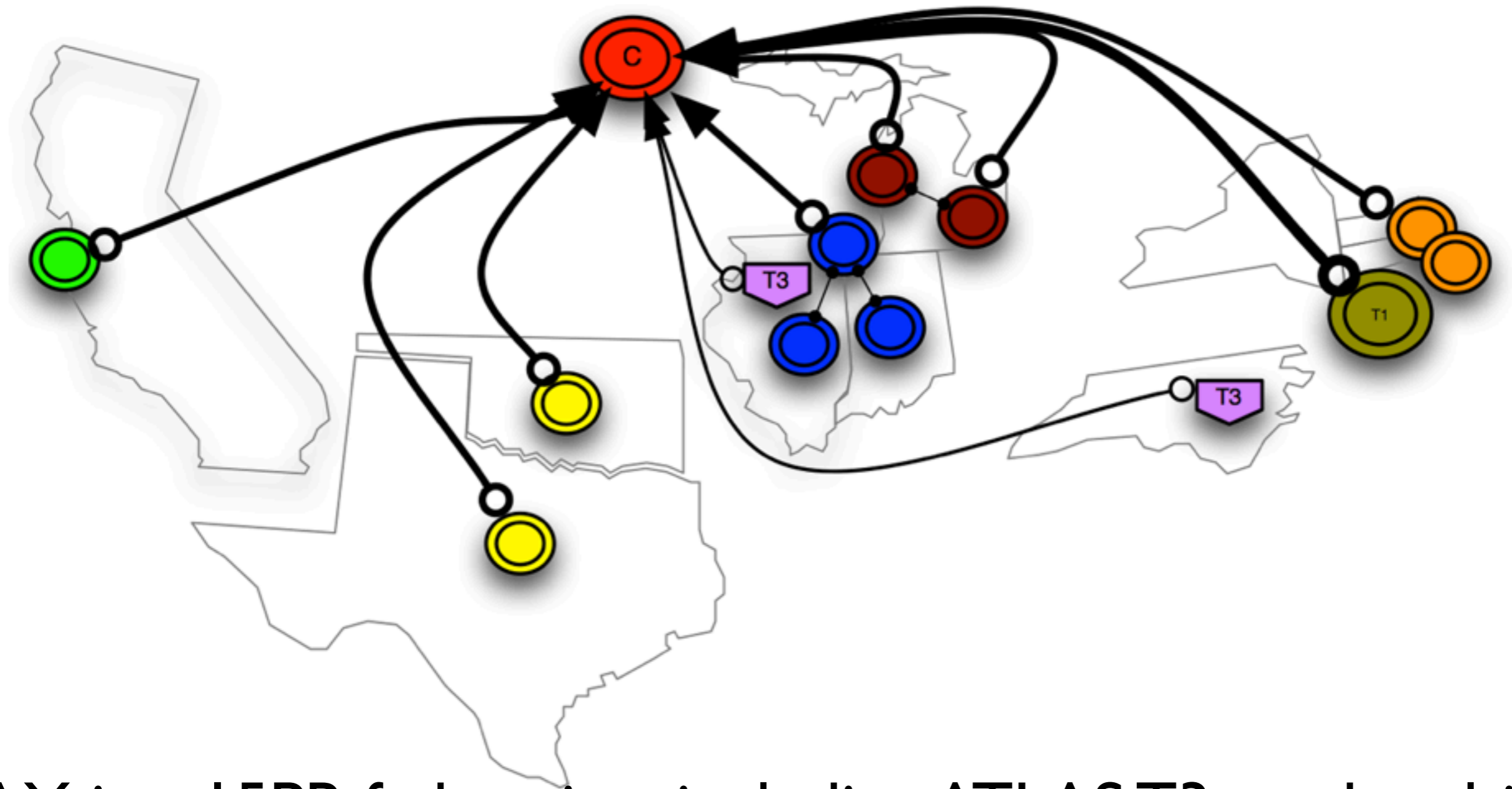
Federations, in practice

- The federation approach has been used by ALICE for many years; used ALIEN, not Xrootd to federate.
- USCMS started federating T2s in 2010; grew to all sites in 2011.
 - Project is named “Any Data, Any Time, Anywhere” or AAA.
- USATLAS started in 2011 and quickly grew to all sites.
 - Project named “Federated Atlas Xrootd”, or FAX.
- Equivalent projects in EU are being worked on.

AAA Deployment

- Currently, redirector at xrootd.unl.edu.
- Includes the FNAL T1 (dCache) and 8 T2s (5 HDFS, 1 dCache, 1 Lustre, 1 L-Store).
- During April, our monitoring recorded:
 - Over 300 unique users,
 - 900K file transfers
 - 300TB moved.

FAX Deployment



FAX is a 15PB federation, including ATLAS T3s and multiple layers of hierarchy.

Technical Concerns

Site Integration

- Most sites integrate via installing a plugin specific to their storage system.
- This causes Xrootd to be a *proxy* to the outer world.
- Anything with a C or POSIX interface can be integrated. E.g., HDFS or Lustre.
- A native Xrootd can do *direct integration* with the redirector.
- dCache has a few options besides proxying:
 - *Native implementation* of the Xrootd protocol, plus a standalone cmsd server for integration.
 - The *overlay* approach has a SLAC Xrootd server running on each dCache pool, reading out of dCache's data directories.

Merging Namespaces

- To provide a uniform namespace, each site must export the *global* filename, not the *local* filename. This is achieved through a Xrootd plugin.
- For CMS, this mapping is achieved through a list of mapping rules and regular expressions.
- For ATLAS, this requires a MySQL database lookup. The higher overhead is partly ameliorated by aggressive caching.
- ATLAS namespace is currently evolving to something simpler.

Authorization

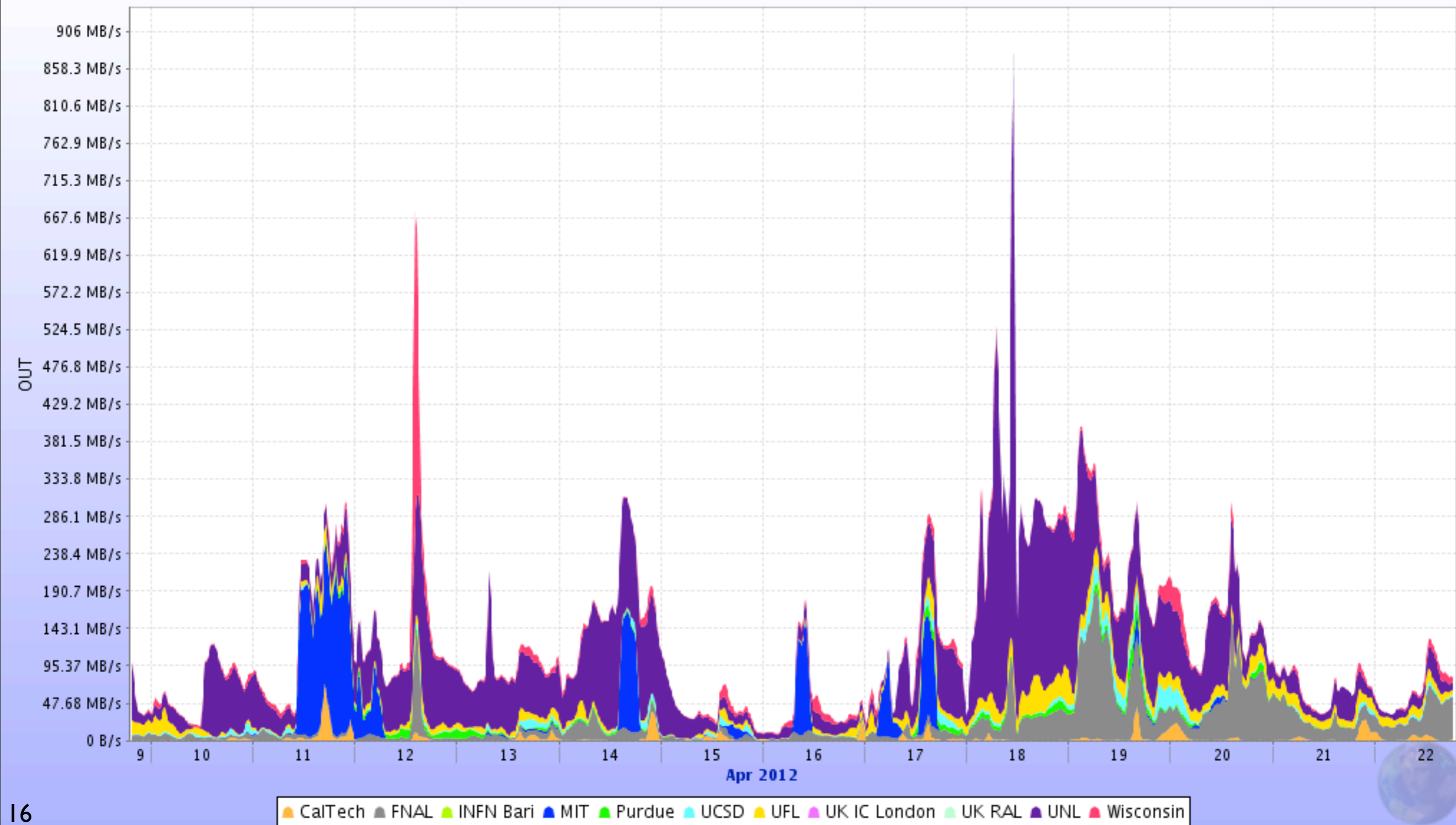
- While Xrootd has many security protocols, like SRM or GridFTP, GSI-based authentication is used in AAA/FAX.
- Xrootd has a plugin for mapping the GSI credentials to a username:
 - FAX uses a simple map file based on VOMS attributes.
 - AAA passes the DN and VOMS attributes to the site mapping service (GUMS).
- Once mapped, a separate file containing authorizations for each user determines file access.

Monitoring

- See poster 233, “Xrootd Monitoring for the CMS experiment,” in track 3.
- We have two streams of monitoring from Xrootd:
 - **Summary:** gives high-level statistics (# connections, failures, MB/s) per-server.
 - **Detailed:** a highly compressed trace of the user’s activities. Works at the per-open-file level.
- In addition, external server health checks - Nagios and RSV.

Monitoring - AAA

Aggregated Xrootd traffic





USATLAS Federated Xrootd Status - 2012-05-08 20:07:

Frequently Asked Questions

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

Metric	Last Executed	Enabled?	Next Run Time	
org.usatlas.xrootd.grid-xrdcp-compare	2012-05-08 19:50:03 CDT	YES	2012-05-08 20:05:00 CDT	
org.usatlas.xrootd.grid-xrdcp-direct	2012-05-08 20:05:01 CDT	YES	2012-05-08 20:20:00 CDT	
org.usatlas.xrootd.grid-xrdcp-fax	2012-05-08 20:05:00 CDT	YES	2012-05-08 20:20:00 CDT	
org.usatlas.xrootd.ping	2012-05-08 20:05:02 CDT	YES	2012-05-08 20:20:00 CDT	

Host: atlas29.hep.anl.gov (atlas29.hep.anl.gov)

Metric	Last Executed	Enabled?	Next Run Time	
org.usatlas.xrootd.ping	2012-05-08 20:05:02 CDT	YES	2012-05-08 20:20:00 CDT	
org.usatlas.xrootd.xrdcp-compare	2012-05-08 20:05:01 CDT	YES	2012-05-08 20:20:00 CDT	
org.usatlas.xrootd.xrdcp-direct	2012-05-08 20:05:02 CDT	YES	2012-05-08 20:20:00 CDT	
org.usatlas.xrootd.xrdcp-fax	2012-05-08 20:05:01 CDT	YES	2012-05-08 20:20:00 CDT	

Host: atlgridftp01.phy.duke.edu (atlgridftp01.phy.duke.edu)

Metric	Last Executed	Enabled?	Next Run Time	
org.usatlas.xrootd.ping	2012-05-08 20:05:01 CDT	YES	2012-05-08 20:20:00 CDT	
org.usatlas.xrootd.xrdcp-compare	2012-05-08 20:05:00 CDT	YES	2012-05-08 20:20:00 CDT	
org.usatlas.xrootd.xrdcp-direct	2012-05-08 20:05:02 CDT	YES	2012-05-08 20:20:00 CDT	
org.usatlas.xrootd.xrdcp-fax	2012-05-08 20:05:01 CDT	YES	2012-05-08 20:20:00 CDT	

Host: dcdoor09.usatlas.bnl.gov (dcdoor09.usatlas.bnl.gov)

Metric	Last Executed	Enabled?	Next Run Time	
org.usatlas.xrootd.ping	2012-05-08 20:05:02 CDT	YES	2012-05-08 20:20:00 CDT	

Use Cases

- Our initial goals tended toward interactive use, but we have found several other compelling use cases for a federation.

Fallback

- If a grid job fails to open a file, don't fail the application. Instead, have it try again, reading from the redirector.
- Loss of efficiency, but the job doesn't crash.
- Great for “breaking in” new federations.
- The job was going to fail anyway; even an unreliable federation is better than nothing.

Storage Healing

- Fallback prevents jobs from failing. Taking it one step further, we can use the federation as a source to re-download the broken file.
- Still an R&D topic. Issues:
 - How to handle failure “bursts” (all files in your storage fail for 10 minutes).
 - How do you determine whether the file *should* be there?

Overflow

- See poster 232, “Controlled overflowing of data-intensive jobs from oversubscribed sites,” in track 3.
- If jobs are queued for a long time, we can purposely send the job to the “wrong” site if
 - Fallback is enabled at the destination.
 - One data source is in the federation.
- This forced fallback allows us to “backfill” otherwise-idle CMS analysis CPUs, work around non-optimal data distribution.

File Caching

- Remote I/O is expensive in terms of WAN bandwidth, especially if the file is re-read many times.
- Instead of asking the redirector for the file, ask a local Xrootd install.
- On the file open failure, the local Xrootd will stage the file to a local disk. On next access, file will be local.
- Using this at the BNL T3, and under investigation at Duke.
- Work needed to make it more fault tolerant. Cache management - which files to evict - is a deep, difficult topic.

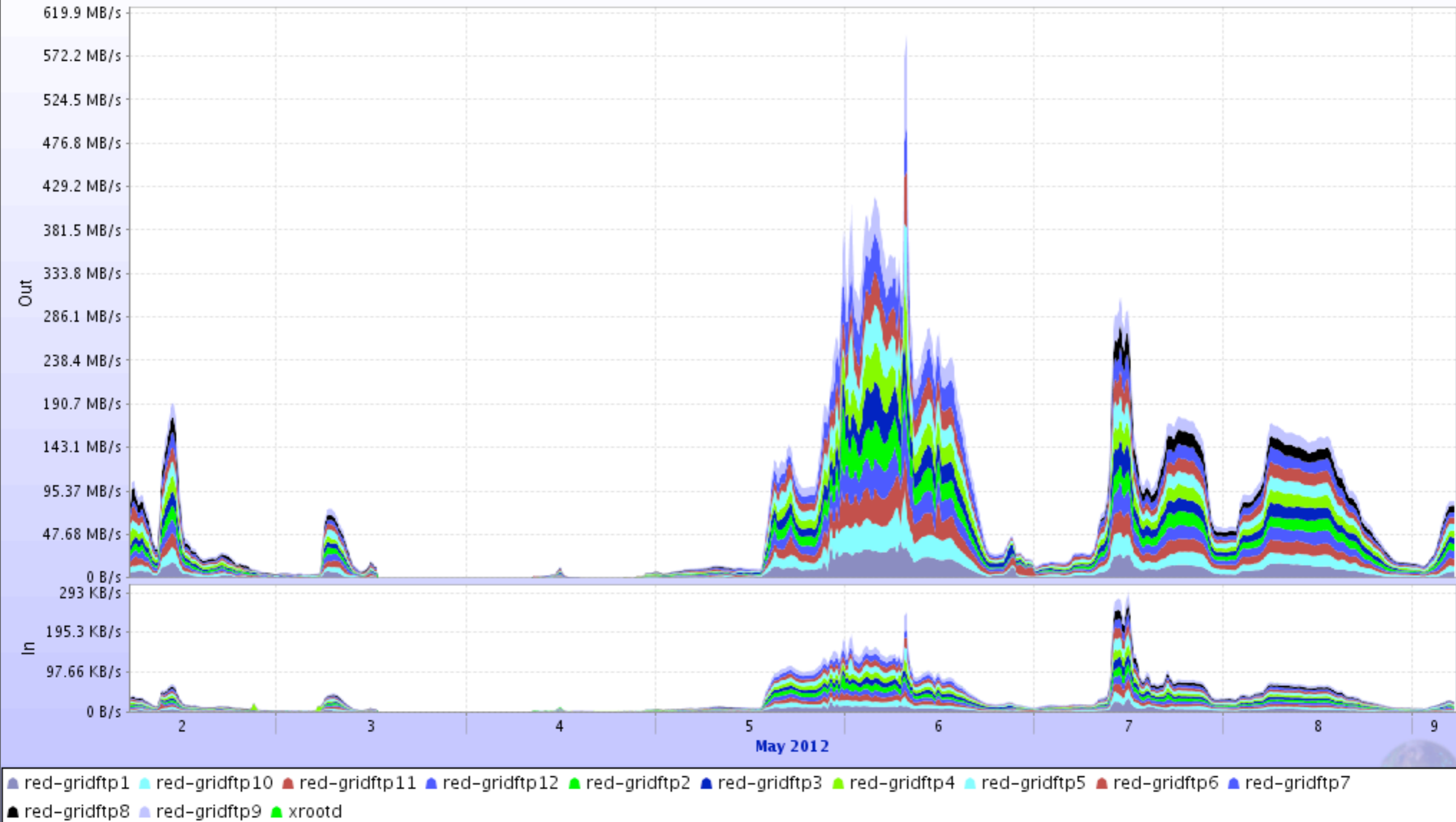
Parting Remarks

- We've been doing this for several years now, and Xrootd-based federations are starting to enter large-scale, day-to-day use for CMS. ATLAS is quickly hardening FAX.
- We've found the **major issues are** not making the federation deliver data, but **covering the failure cases** - esp. within the client.
- The ability to do wide-area, direct Xrootd access is dependent on the application code's ability to handle large network latency.
- Effort by the experiment is a prerequisite to many of these use cases work.

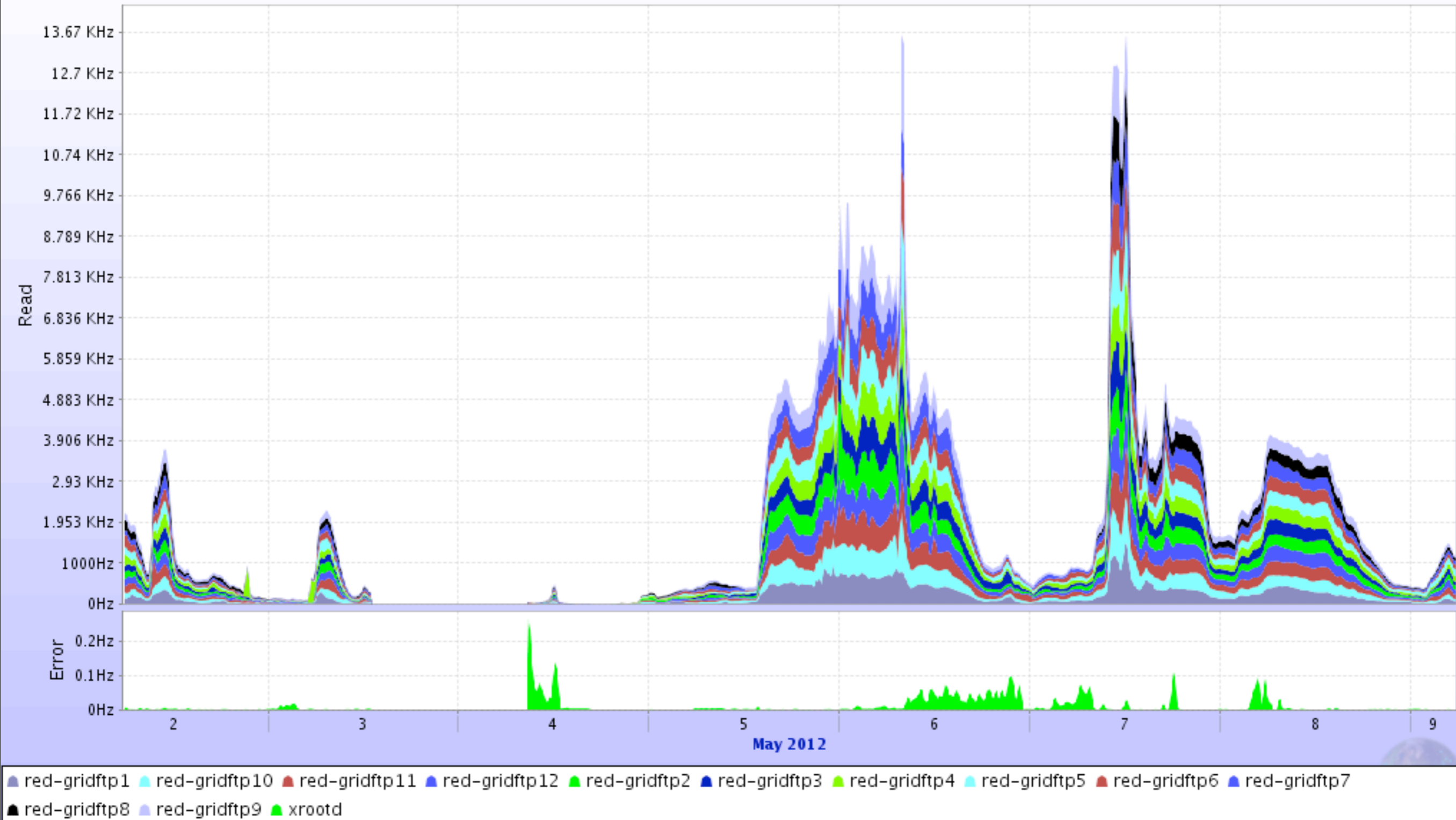
Questions?

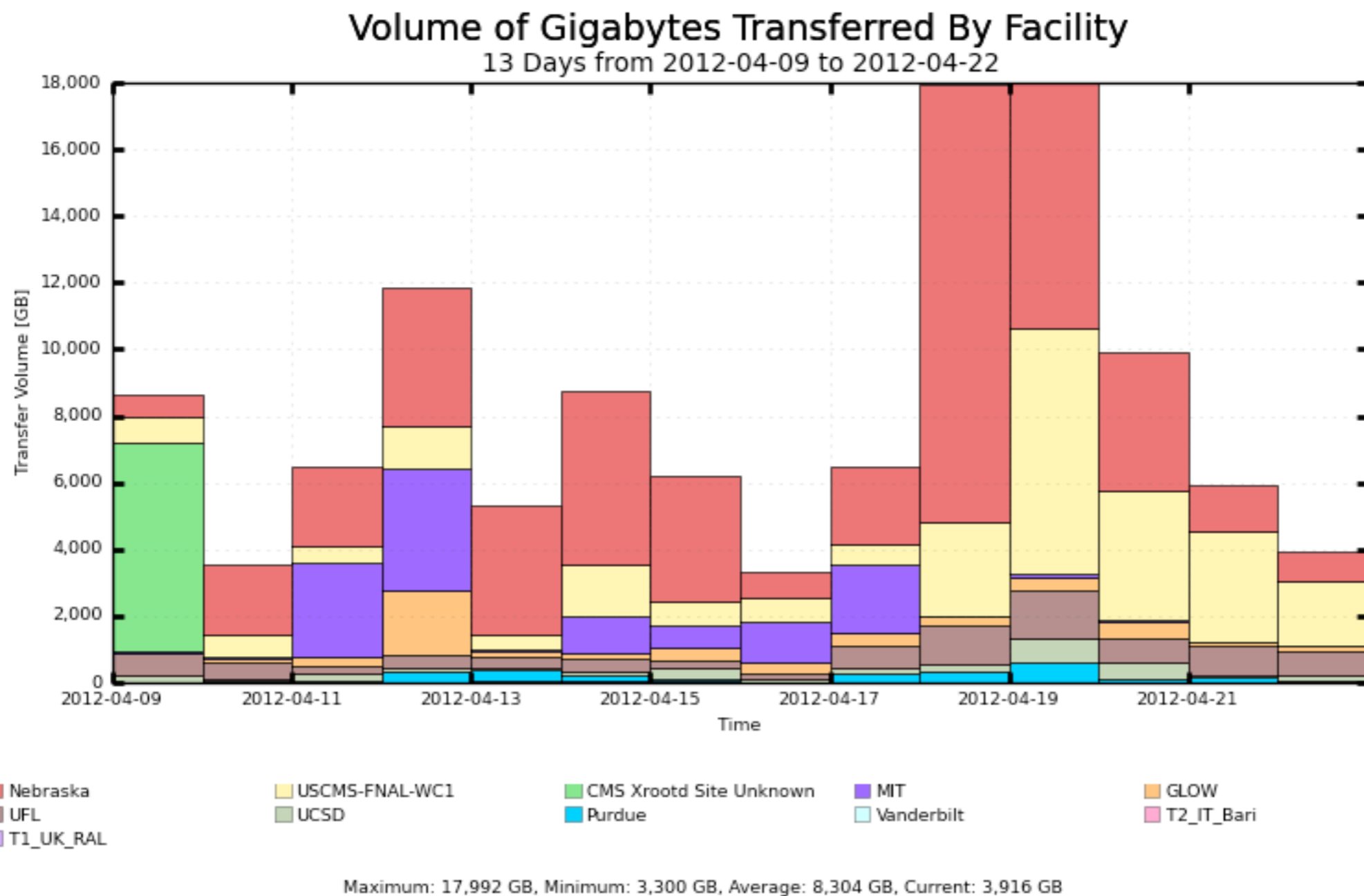
Backup Slides

XrdReport for link traffic on UNL



XrdReport operation rates on UNL





File ^	User Hash	Server Domain	Client Domain	Open Ago	Update Ago	Read [MB] ^	Read [%]	Rate
/store/mc/Fall11	6DB2B1B7	hep.wisc.edu	unl.edu	25:37:11	21:22:11	2728.245	68.827	0.030
/store/mc/Summer11	3C12BE84	fnal.gov	unl.edu	01:06:03	00:05:23	2523.107	62.609	0.637
/store/data/Run2012A	3D8C2B76	fnal.gov	t2.ucsd.edu	04:30:43	00:08:38	2498.118	54.639	0.154
/store/data/Run2012A	3D8C2B76	fnal.gov	t2.ucsd.edu	03:15:18	00:01:08	2474.463	57.433	0.211
/store/data/Run2012A	3D8C2B76	fnal.gov	t2.ucsd.edu	04:00:37	00:01:42	2442.054	57.521	0.169
/store/mc/Summer11	3C12BE84	fnal.gov	unl.edu	00:51:58	00:00:58	2429.944	64.505	0.779
/store/mc/Summer11	3C12BE84	fnal.gov	t2.ucsd.edu	02:14:36	00:04:21	2420.094	59.497	0.300
/store/data/Run2011A	3C12BE84	ihepa.ufl.edu	unl.edu	01:26:01	00:04:41	2413.841	63.327	0.468
/store/data/Run2012A	3D8C2B76	hep.wisc.edu	t2.ucsd.edu	04:00:47	00:10:22	2411.461	52.835	0.167
/store/mc/Summer11	3C12BE84	unl.edu	t2.ucsd.edu	01:31:55	00:00:50	2366.587	64.929	0.429
/store/data/Run2011A	3C12BE84	ihepa.ufl.edu	unl.edu	03:52:41	00:01:56	2365.060	59.156	0.169
/store/data/Run2012A	3AC086B	fnal.gov	t2.ucsd.edu	01:17:18	00:00:18	2320.691	62.855	0.500
/store/data/Run2012A	3D8C2B76	hep.wisc.edu	t2.ucsd.edu	04:41:24	00:08:09	2274.699	48.299	0.135
/store/data/Run2012A	3D8C2B76	hep.wisc.edu	t2.ucsd.edu	04:01:48	00:00:23	2212.339	59.802	0.152
/store/mc/Summer11	3C12BE84	hep.wisc.edu	ultralight.org	02:02:28	00:16:48	2170.151	54.353	0.295
/store/data/Run2012A	3D8C2B76	fnal.gov	t2.ucsd.edu	04:27:46	00:05:21	2164.812	53.054	0.135
/store/data/Run2011A	3C12BE84	ihepa.ufl.edu	unl.edu	01:20:25	00:00:30	2117.791	53.866	0.439
/store/data/Run2011A	3C12BE84	ihepa.ufl.edu	unl.edu	03:52:41	00:11:16	2101.871	52.162	0.151
/store/data/Run2012A	3D8C2B76	fnal.gov	t2.ucsd.edu	04:20:26	00:00:56	1991.785	61.253	0.127
/store/data/Run2012A	3AC086B	ihepa.ufl.edu	t2.ucsd.edu	08:41:15	00:00:05	1976.205	62.562	0.063
/store/mc/Summer11	3C12BE84	unl.edu	ultralight.org	01:17:14	00:10:34	1953.870	53.664	0.422
/store/mc/Fall11	B83B3C94	hep.wisc.edu	ultralight.org	11:32:01	02:49:41	1944.444	52.812	0.047
/store/data/Run2012A	3D8C2B76	hep.wisc.edu	t2.ucsd.edu	03:17:33	00:02:03	1919.433	64.035	0.162
/store/data/Run2011A	3C12BE84	ihepa.ufl.edu	unl.edu	03:38:01	00:00:01	1908.636	50.158	0.146

Sample Daily Report

=====

Xrootd Summary for 2012-05-09 | 59.92 TB | 37% increase

=====

Source Site	Volume GB	# of Transfers	Yesterday Diff	One Week Diff	

GLOW	1,080	1,223	55%	1908%	
GLOW_Internal	46,053	38,388	65%	32%	
MIT	4,950	11,460	103%	95800%	
Nebraska	4,195	9,118	-43%	65%	
Purdue	415	2,168	279%	374%	
T2_IT_Bari	0	6	Unknown	-96%	
UCSD	551	4,245	-65%	927%	
UFL	1,138	2,066	945%	1250%	
USCMS-FNAL-WC1	1,521	2,222	-57%	-32%	
Vanderbilt	14	746	400%	598%	
