Data on OSG

Frank Würthwein
OSG Executive Director
Professor of Physics
UCSD/SDSC

The purpose of this presentation is to give the Council a summary of where we are at with supporting data on OSG

High Level Messages

- First line of defense is HTCondor file transfer
- If that's not sufficient for needed scale:
  - Pull/put data to/from job via gridftp and/or xrdcp
    - We offer data hosting in some cases.
  - Use cashing if same input data is reused often.
    - We can support "reasonable" privacy of data, but not HIPAA or FISMA.
- If data movement needs to be managed across multiple global locations, independent of jobs:
  - We helped Xenon1T to adopt Rucio, the ATLAS data management solutions.
  - We expect to help others evaluate this as possible solution in the future.
    - First potential customers for an eval are LSST and LIGO.

The services provided by OSG in support of data on OSG vary, depending on the size of the needs of the communities we deal with.

Benchmarking HTCondor Filetransfer

Initiated by GlueX and Jefferson Lab.

Wanted to know if a single submit host at JLab can support glueX operations needs.

Concern was primarily the IO in and out of the system.

OSG did the test on our system.

Then provided instructions for deployment at JLab.

Then repeated test on their system, and helped debug until expected performance was achieved.

GlueX Requirements

| Parameter	GlueX Spec	OSG Test
Running jobs	20,000	4,000
Output Size	10-100 MB	250 MB
Input Size	1-10 MB	1-10 MB
Job Runtime	8h - 9h	0.5 h

GlueX specs translated into 55.5MB/sec and ~1Hz transaction rate

$$O(n, l, s) = \frac{nJobs * size}{length} = \frac{20000 * 90}{9 * 3600} \approx 55.5 \frac{\text{MB}}{\text{sec}}$$

We tested x10 larger IO and x3 more transactions per second.

Benchmarking Result

- Smooth operations at the scale tested.

- Lessons Learned:

  - Stay away from exceeding significantly more than half the 10Gbps network bandwidth on the submit host.

  - Be careful with TCPIP settings to avoids latencies of schedd communications with far away worker nodes.

10 Gbit interface limit
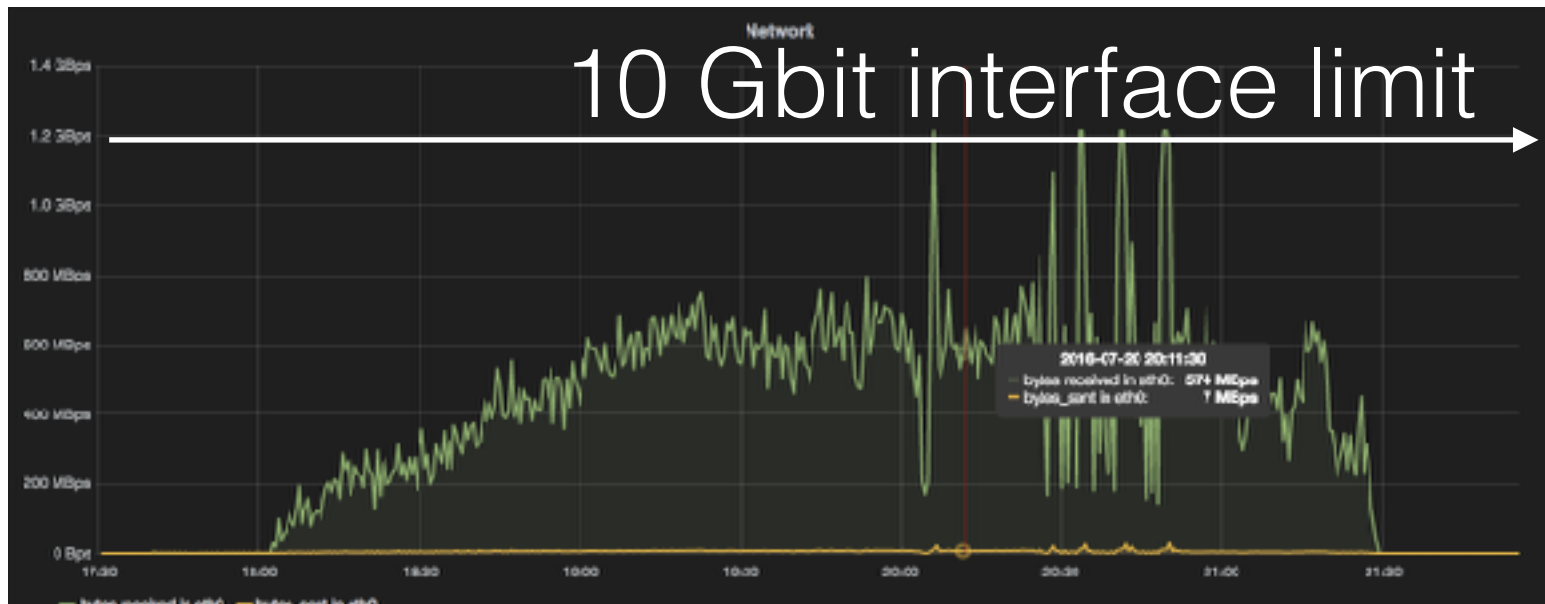
Put and Get at 100Gbps

OSG offers installation instructions for deploying a cluster of Gridftp or xrdcp hosts, each of which is 10Gbps connected, and seen by the clients as a single service, using Linux Virtual Server.

This is the OSG strategy pursued for replacing SRM.

It's also what LIGO used for its first gravitational wave detection work on OSG.

Aside on Reducing Complexity

- We are working to reduce the complexity we support for the LHC in order to sustain it with less effort in the future.

- E.g. SRM:

  - In OSG 3.2 there were 4 SRM clients

  - In OSG 3.4 there are none.

- E.g. X509:

  - We are working on eliminating the need for X509 from OSG.

    - More on that later.

Caching via StashCache

- OSG now operates its own Data Federation.
  - We support federations inside ours that have privacy from each other.
  - We support people to build their own.
  - The advantage of living inside OSG is that you have access to the StashCache deployed infrastructure.
  - If you roll your own, you are on your own.

OSG Data Federation

Applications connect to regional cache transparently.
Regional cache asks redirector for location of file.
Redirector redirects to relevant origin.
File gets cached in regional cache.

Caches at:
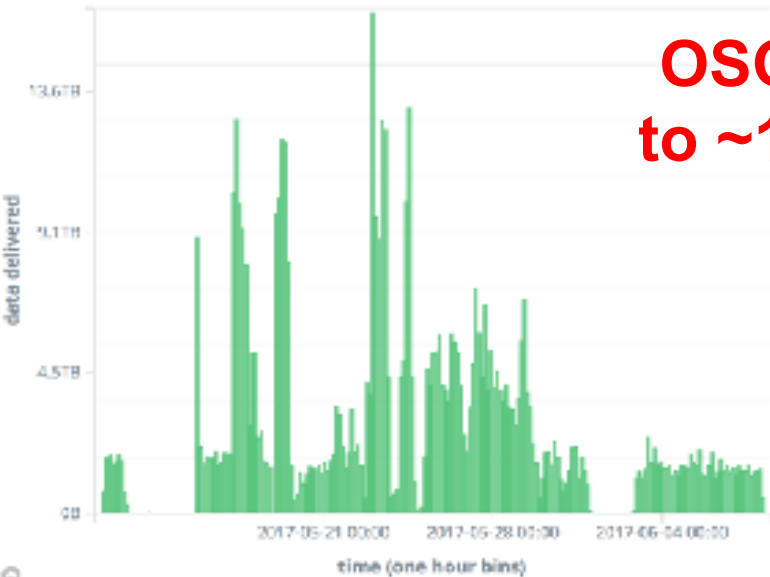**BNL, FZU, UNL,**
Syracuse, UChicago,
UCSD/SDSC, UIUC.
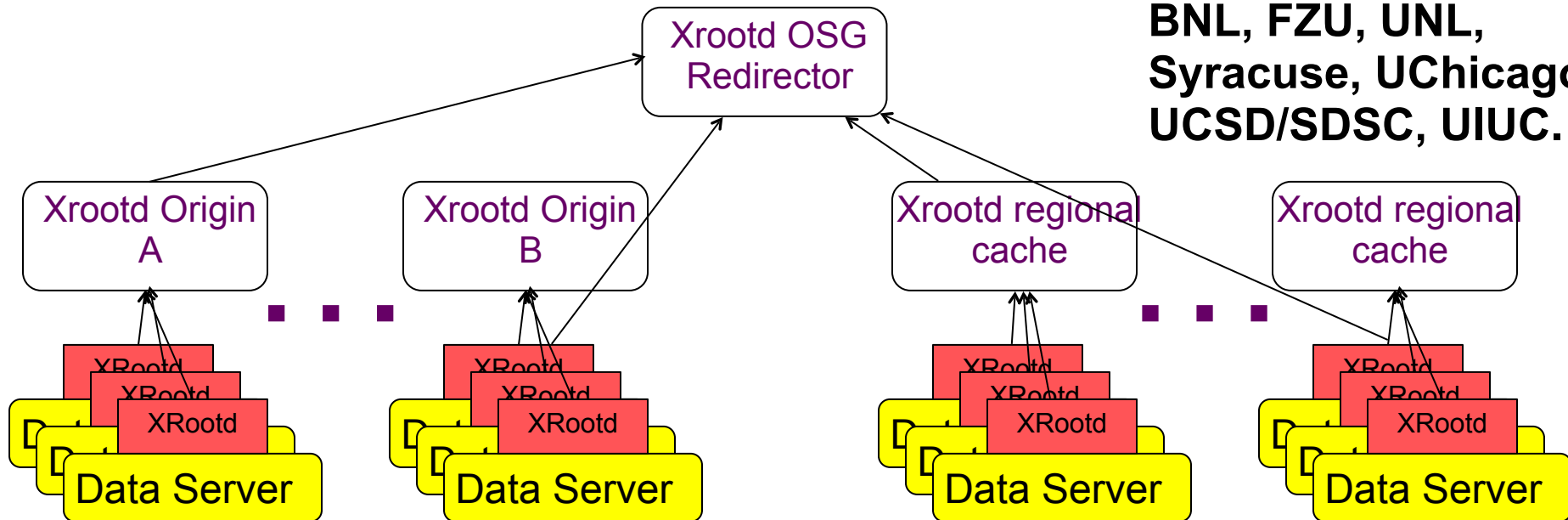
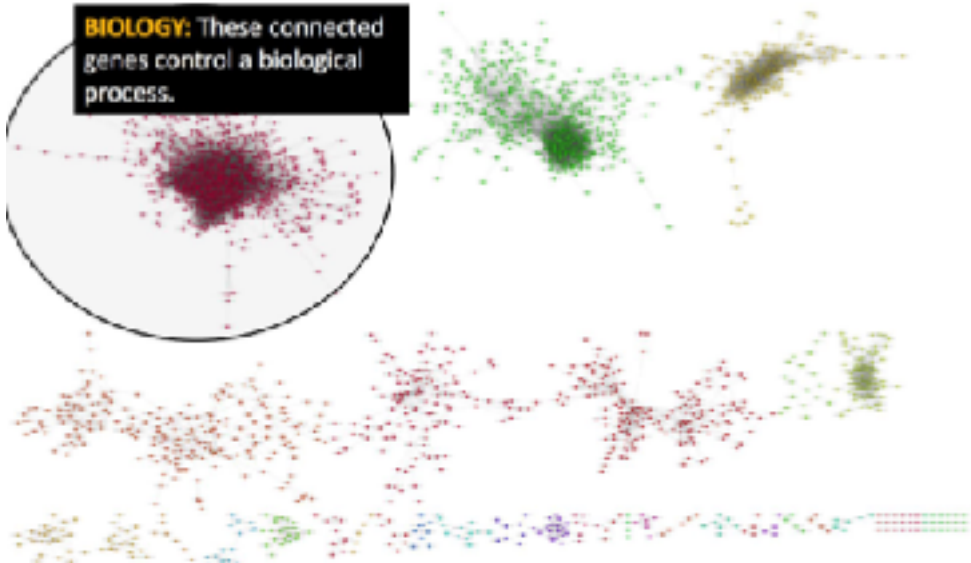One Data Origin per Community

Multiple caches across US

This is a technology transfer from LHC with some OSG value added.

Communities using StashCache

- OSG-Connect
  - See next slide for examples.
- LIGO
- Nova
- And some expression of interest:
  - Xenon1T, expects future use in front of Comet@SDSC, and potentially elsewhere.
  - GlueX, initial interest, not yet concrete.

Big Data beyond Big Science

OSG caching infrastructure used at up to ~10TB/hour for meta- or exo-genetics

Motivation

- In a pervious project, we have isolated exosomes from the commercial cow's milk and assessed the bovine molecules inside the exosomes.

- Moreover, we also found many unmapped reads are from microbial species.

- Thus, we designed a follow-up study to understand the origin of microbial sequences in bovine milk exosomes.
  - **Metagenomics analysis**

Jang Shu · University of Nebraska Lincoln

COMPLEX GENETIC SYSTEMS WOVEN INTO GRAPHS

BIOLOGY: These connected genes control a biological process.

William Poehlman, Alex Feltus · Clemson University

Stephen Ficklin, Washington State University

StashCP Dashboard Info
Last 3 Months
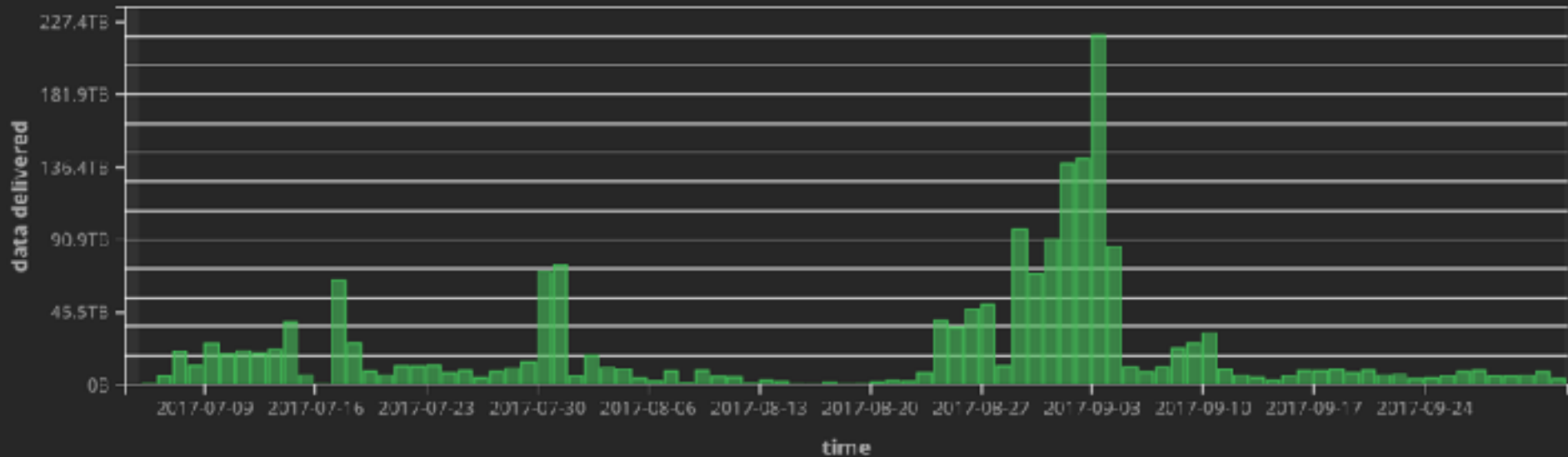
Dashboards Hosted at Kibana Instance at MWT2

Sum of link_out
- 195.6KB – 1.1TB
- 1.1TB – 2.2TB
- 2.2TB – 3.3TB
- 3.3TB – 4.3TB
- 4.3TB – 5.4TB

Details on data in/out, connections, errors, timeouts, retries, … for each cache are monitored.

Rucio and its use in Xenon1T

- Xenon1T needed something to manage its transfers between the experiment DAQ in Italy, and various disk locations in EU, Israel, and the US.

Data Flow (now)

Legend:
- ☐ Disk space
- ☐ Rucio Storage Element
- ☐ Tape

Raw Data
- — Rucio transfers (simultaneous)
- — Tape upload (download)
- — Ruciax upload (download)

Processed Data
- — Cax
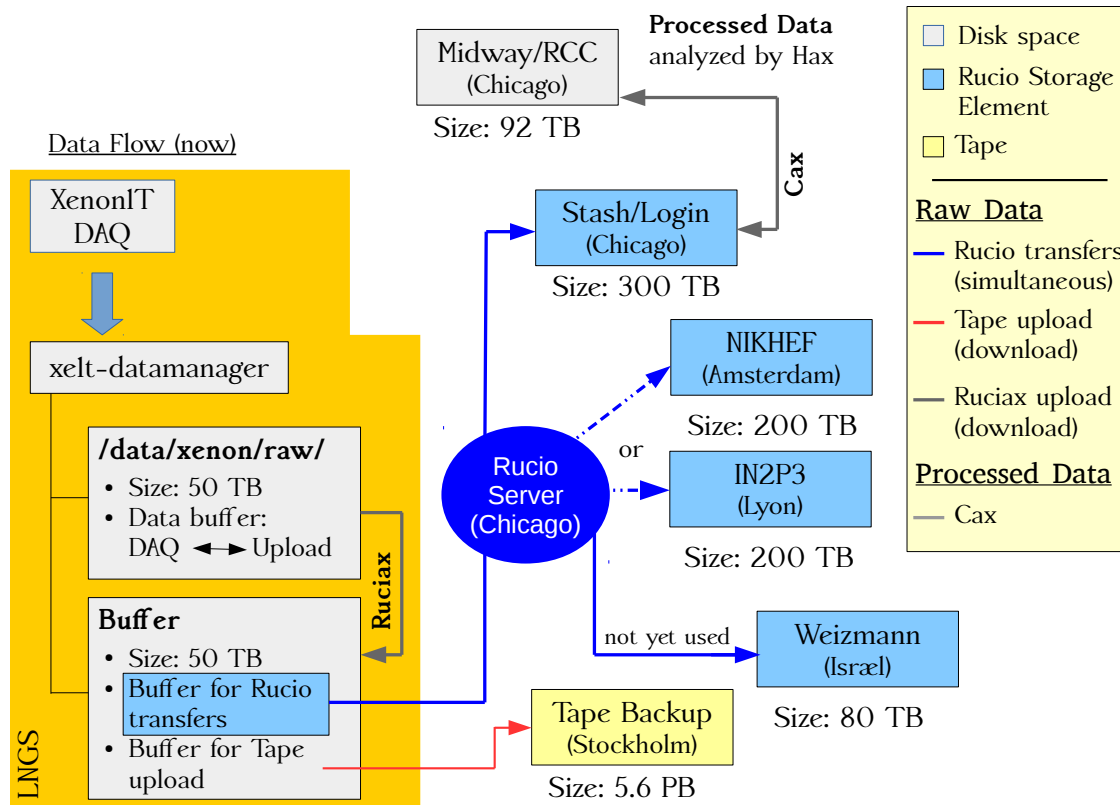
- Xenon1T adopted Rucio for this after joint evaluation with OSG.
- Since then, LSST and LIGO expressed an interest in a similar evaluation.
- Next steps:
  - Two pager to define metrics for eval project with LSST.
  - OSG Blueprint to better understand technical concept underlying Rucio.

A Future without X509

We already have eliminated X509 for user job submission in OSG.

The two remaining use cases are:

Pilots being authenticated at CEs

Users staging out data to Storage Endpoints from jobs.

Problem Statement

- At present, Storage Endpoints authenticate users
  - X509 certificate is delegated to the job for that job to stage out data to a Storage Endpoint from the worker node.
- In the future, we want Storage Tokens that define capability rather than personhood.
  - You are allowed to store data in your directory at OSG-VOs storage endpoint(s).
- Working with NSF funded SciToken project to accomplish this.
  - https://scitokens.org

Initial "Demo"

- Initial Demo showed:
  - OSG-Connect HTCondor submit host transparently generates SciToken
    - Users are oblivious of the existence of such Tokens.
  - User jobs put files from worker node into user-owned directory at Stash Endpoint using HTTPS protocol.
    - Stash is the Origin of StashCache, implemented as Xrootd server => This implies that data staged out can be used for subsequent processing via StashCache.
- Based on OAuth2 framework
  - Same as when you authorize 3rd party website to use your Facebook/Google/DropBox login.

Status of SciToken

- Initial hackathon led to initial demo, and thus understanding of viability of basic concept.

- Draft Design write up for technical director evaluation, and broader discussion exists.

  - Needs a bit more work before it's ready for sharing.

Summary & Conclusion

- OSG made a lot of progress in supporting data on OSG broadly for anybody.

- We build on technologies that have broad community support and/or are NSF funded projects.

- We reduced, and will continue to reduce the complexity of the software stack required to use data on OSG.

- There's some i's to dot and t's to cross, but decent functionality now exists, and the geek gap between Big Science and the rest of scientific endeavors has shrunk, and continues shrinking.