

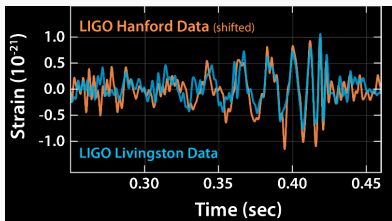
# Bulk Data Management For LIGO

---

J. A. Clark

January 16, 2018

LIGO: Laser Interferometer Gravitational Wave Observatory: Large-scale physics experiment / observatory for astrophysical gravitational waves

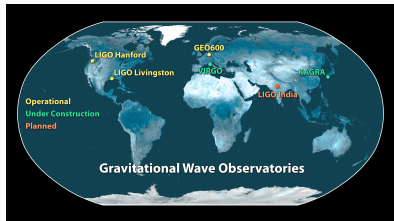


**Figure 1:** GW150914: first GWs observed [1]

Principal data products (Each instrument: 25 MBytes/s):

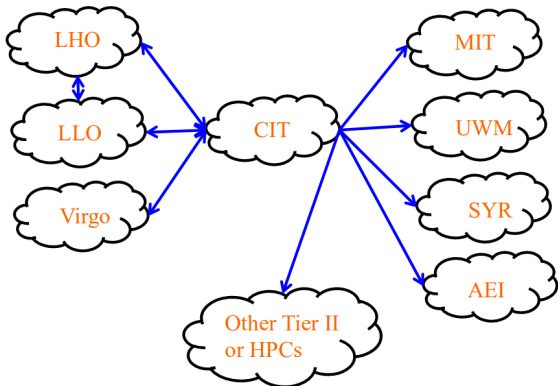
- Strain  $h(t)$ : 1 digitized time series / detector<sup>1</sup>
- $N$  auxillary channels

<sup>1</sup>LIGO,GEO: 16384 Hz, Virgo: 32768 Hz



**Figure 2:** Global network of (operational) detectors: 2x LIGO, 1x Virgo, 1x GEO600

# Bulk Data Management & LIGO



Data distribution:

1. Low-latency service: data @ CIT in  $\sim 10$ s
2. Bulk data management: complete datasets for massively parallel computing

## Data Access: Consumer Perspective

LIGO Data Grid (LDG): Collection of Tier-1, 2 compute sites operated by LIGO personnel.

Users on the LDG

- Users log in to an individual site, jobs submitted there run there [**HTCondor (majority)**, **Pegasus**].
- Data & software accessible via NFS
- Jobs run at the site using data hosted at that site

Most familiar mode of operation for the majority of analyses.

~Recent deployment to Open Science Grid (OSG)

- Dedicated OSG-submit nodes at LDG sites
- Software accessed via /cvmfs, increasingly singularity
- HTCondor-based workflows require local data storage

# Bulk Data Management & LIGO Data Sets

LIGO/Virgo data acquired over engineering (ER) & observing (O) runs.

E.g.,:

- O1: Sept 2015 – Jan 2016
- O2: Nov 2016 – Aug 2017

Data is contained in “frame” files:

- in-house specification for time-series data
- frame types: raw, RDS,  $h(t)$  & calibration version (C00, C01, ...)

Data sets defined by:

1. observational period (e.g., O1, ER10, O2, ...)
2. type

Note: bulk data management is done for official, published data only<sup>2</sup>

---

<sup>2</sup>conceivable the scope of “official” may broaden

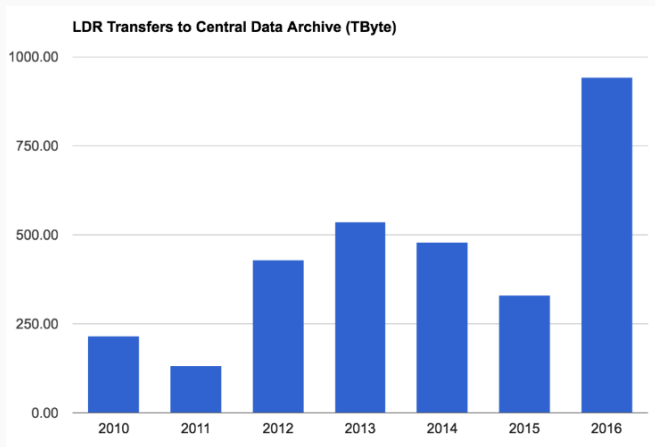
# Data Sets

Data Set	File count	Total Size
H1 O2 h(t) C01	5506	2.41 TB
H1 O2 h(t)	5795	2.38 TB
H1 O2 raw	367547	600.11 TB
L1 O2 h(t) C01	5496	2.52 TB
L1 O2 h(t)	5774	2.55 TB
L1 O2 raw	367386	518.84 TB
V1 O2 h(t)	10961	546.62 GB
H1 O1 h(t) C02	2705	784.26 GB
L1 O1 h(t) C02	3113	638.55 GB
H1 O1 h(t) C01	2684	1.17 TB
L1 O1 h(t) C01	3178	1.14 TB
H1 O1 h(t)	2746	1.16 TB
L1 O1 h(t)	2611	1.18 TB
H1 O1 rds	41547	25.85 TB
L1 O1 rds	41554	24.53 TB
H1 O1 raw	166160	126.09 TB
L1 O1 raw	166188	116.12 TB
L1 O2 h(t)	5774	2.55 TB
H1 O2 h(t)	5795	2.38 TB

# Data Distribution



# Data Flow Statistics



- First aLIGO run had 294 Tb of data distributed over 418,657 files
- Reduced data replicated to CIT, LHO, LLO, AEI, UWM, SU
- Subsets replicated to OSG sites (GATech, Nebraska) and to Virgo.



# Current Practice(s)

~ 2 current approaches to data management:

## 1. LDG sites

- Local storage
- Distribution managed by LIGO Data Replicator (LDR)
- Manual workflow partitioning

## 2. OSG sites

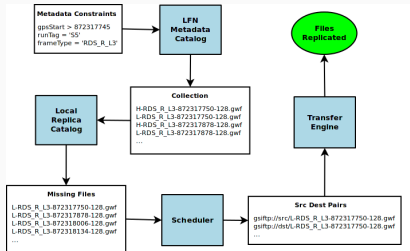
- Mix of local, non-local storage
- Distribution managed by LDR, globus, CVMFS
- Generally automated workflow partitioning

The LIGO Data Replicator (LDR) [2]:

- a system for replicating bulk data sets or file sets from one site to another.
- a system for data or file discovery by people and their compute jobs across multiple sites or grid.
- a collection of grid middleware sewn together using the Python scripting language as glue.

# LDR Workflow/Architecture

1. Constraints on logical filename (LFN) metadata are compared to a catalog to determine the list of LFNs for replication
2. A local replica catalog (LRC) reports which LFNs exist locally. Intersection of the target list and LRC yields the list of files which need to be replicated.
3. A scheduler locates physical filenames (PFNs)/URLs available from remote sites for the LFNs. Scheduler output is a list of source-destination pairs for input to a transfer engine or service.
4. A transfer engine acts on the source-destination pair and replicates the missing files or LFNs. The LRC is updated after replication.



**Figure 3:** LDR workflow [2]: LFNs for replication are identified via metadata & only files absent from the local catalog are transferred.

Data distribution via:

1. Completely centralized: use HTCondor file transfer
2. Centralized with stage-in: use LDR to pre-stage data to an OSG site with storage (e.g., Nebraska), input data downloaded (GridFTP or XRootD) at job startup
3. Distributed CVMFS storage: use LDR to pre-stage data to an OSG site with storage (e.g., Nebraska), use StashCache to cache data locally.

Site	Secure CVMFS	LDR	Other
Comet	✓		
Fermilab	✓		
Georgia Tech	✓	✓	
Caltech/Louisiana	✓	✓	
Nebraska	✓	✓	
NIKHEF		✓	
Omaha	✓		
Michigan	✓		
Polish VIRGO		✓	
Syracuse	✓	✓	
UCSD	✓		
Wisconsin			
TACC			✓

**Table 1: Availability of data at processing sites**

**Figure 4:** Availability of data at processing sites [3]

# LDR & The Future?

- Good: stable
  1. LDR: extremely mature, no further development expected
  2. Long, successful history on LDG
- Bad: changing data ecosystem
  1. Relies on GridFTP
  2. In-house software: limited support
  3. Multiple distribution paths: LDR, /cvmfs, out-of-band globus transfers (e.g., TACC)

Clear need to explore other technologies, potentially unify all bulk data management





Partnering with OSG for data management:

1. ~~Archival: Moving non-reproducible data from an instrument (such as a detector) to a long-term storage (archival) site.~~
2. Replica-based ✓: Management of storage where the data (contents and lifetime) is managed by the experiment. [**LDG sites**]
3. Cache-based ✓: Data located at a storage service whose lifetime is decided by the storage service itself. [**LIGO & OSG**]
4. Compute-Workflow-based ✓: Data is moved to/from the job sandbox as part of the compute workflow. [**LIGO & OSG**]

See [workshop guidance notes](#)

# Summary

- LDR & LDG: mature technology / stable clusters with LIGO-control
- Diversifying resource ecosystem: LDR += OSG
- New personnel associated with data management (me!)
- Multiple data management models in use:
  1. Traditional: LDR + LDG
  2. Centralized OSG: ship data at submission
  3. Pre-staged: LDR → OSG site with local control
  4. Federated: CVMFS
- LIGO already using OSG:
  - Pegasus workflows (pycbc) already play nicely with OSG
  - Many non-Pegasus workflows: limited to sites with LIGO data
- **interested in ways to streamline data management**

-  B. P. Abbott *et al.* [Virgo and LIGO Scientific Collaborations], ,  
“Observation of Gravitational Waves from a Binary Black Hole  
Merger”, Phys. Rev. Lett. **116**, 061102 (2016), arXiv:1602.0383
-  “LDR Administrator Manual”,  
<https://wiki.ligo.org/DASWG/LIGODataReplicatorAdminManual>
-  “Data Access for LIGO on the OSG”, D. Weitzel, B. Bockelman, D.  
A. Brown, P. Couvares, F. Wrthwein, E. F. Hernandez,  
[arXiv:1705.06202](https://arxiv.org/abs/1705.06202)
-  “Rucio The next generation of large scale distributed system for  
ATLAS Data Management”, V Garonne et al 2014 J. Phys.: Conf.  
Ser. 513 042021



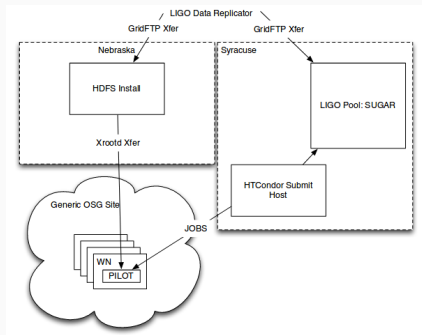
# Supplementary Material

---

# Centralized OSG Implementation (staged)

O1 implementation of data distribution for pycbc on Sugar & OSG

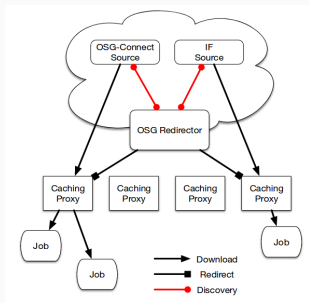
- LDR stages frames to HDFS at Nebraska
- Jobs submitted from HTCondor Submit Host at Syracuse
- OSG jobs see Nebraska as a large shared filesystem
- At job startup: input data downloaded from Nebraska via GridFTP / XRootD
- Simple but scales poorly
- Only amenable to Pegasus workflows



**Figure 5:** Centralized storage deployment for pycbc on the OSG in O1 [3]

## Distributed OSG Implementation (2)

“StashCache provides data caches near compute site . . . uses a distributed network filesystem (based on XRootD proxy caching)”

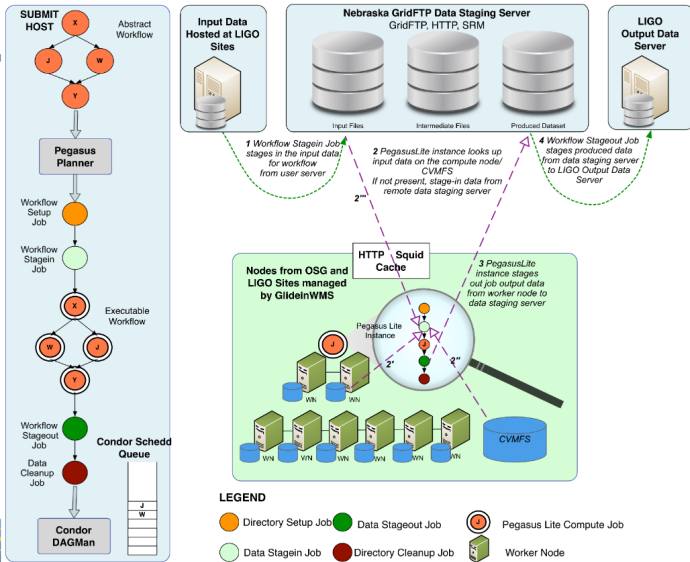


**Figure 6:** StashCache architecture (read from bottom): Jobs request data from caching proxies. Caching proxies retrieve data from multiple sources via XRootD [3]

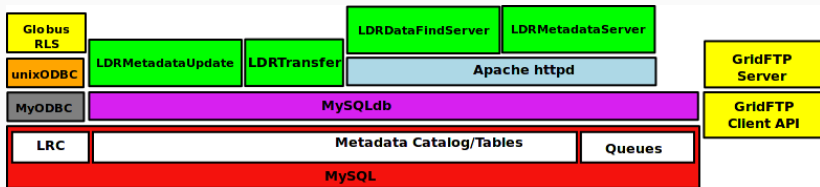
1. LIGO data is LDR'd to Nebraska and published to CVMFS: [ligo.osgstorage.org](http://ligo.osgstorage.org)
2. Redirector determines appropriate data source
3. If data is non-local, caching nodes query OSG XRootD redirector
4. OSG jobs pull data from geographically near caching proxies

# Centralized OSG Implementation (3)

## Data Flow for LIGO Pegasus Workflows in OSG



# LDR Components



**Globus RLS:** Each site runs a Globus Replica Location Service (RLS) instance as part of LDR. RLS provides a local replica catalog (LRC) and a replica location index (RLI). The RLI allows mappings from LFNs to remote LRCs so that to locate a particular LFN for replication one need not query every remote LRC. LDR administrators may sometimes use client tools to directly query the RLS server when trouble shooting.

**GridFTP server:** The data files residing on a file system are exposed via a GridFTP? server. The GridFTP? server enables multiple concurrent file transfers with each one using multiple data streams and tunable TCP windows for high performance data transfers. Striped server configurations are also possible for systems with data on high performance file systems (not NFS) available to multiple nodes.