Data Management for IceCube

Patrick Meade

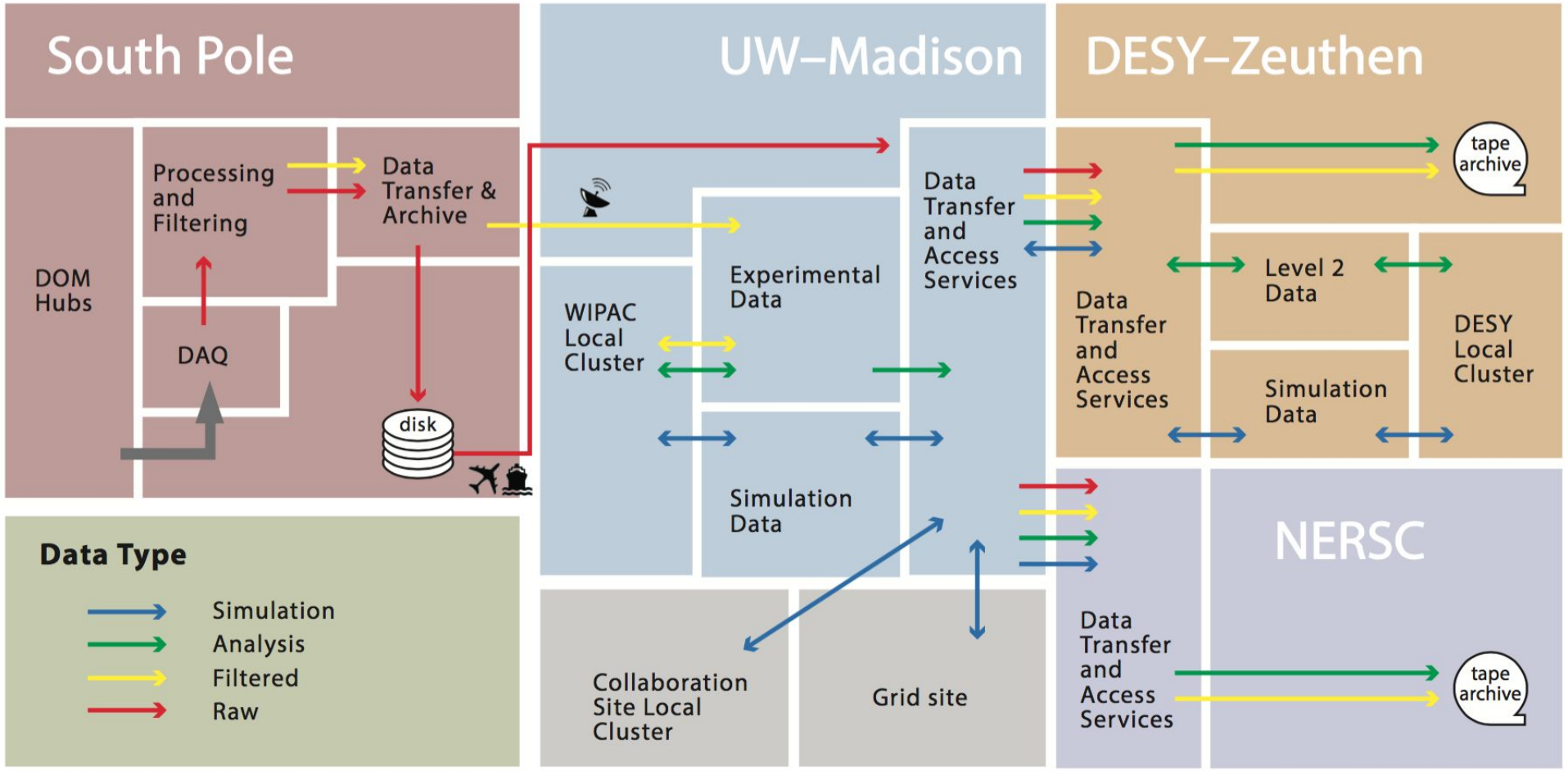
16 January 2018

IceCube Laboratory at Night

Project Background

- Neutrino Observatory at Geographic South Pole
- ~5400 Digital Optical Modules (DOMs) in a cubic kilometer of ice
- Faint traces of light become signal to the IceCube Laboratory (ICL)
- South Pole Systems (SPS) data center does processing and filtering
- RAW: 1 TB/day to Archival Disk Filtered: 100 GB/day to Satellite Transfer

IceCube Data Flow



Data Management: JADE

- JADE: Java Archival and Data Exchange
- SPS JADE (South Pole JADE): Consumption, packaging, and transmission at the South Pole Systems (SPS).
- JADE North: Indexing and warehousing satellite data in the Data Warehouse at UW-Madison.
- JADE Long Term Archive (JADE LTA): Indexing and large (>500+ GB) archival to collaboration partners at DESY and NERSC.

Data Management: File Catalog

- File Catalog: Simple REST API over MongoDB
- Flexible schema: Fixed top level fields, variable sub-objects
- JADE populates transfer and archival information
- Simulation and Production (IceProd) data pipelines also populate
- Query via web interface is in development now

Question 1

Of the broad categories of data management workflows outlined above, which are you interested in partnering with the OSG on?

- Replica-based
- Compute-Workflow-based

Our South Pole data flow is specialized due to limited network connectivity. More opportunities exist in the north with replication to DESY/NERSC, and making data available for analysis at grid sites.

Question 2A

Using the framework / terminology outlined above, can you describe your existing data management workflows?

A. Are you looking to implement new workflows or replace the existing one?

We have existing workflows, so we'd be looking to integrate and/or replace them. Right now our replication to DESY and NERSC are tracked in a JADE database and we also look to populate the File Catalog. Perhaps Rucio could support or enable that at some level.

Question 2B

Using the framework / terminology outlined above, can you describe your existing data management workflows?

> B. If you are looking to replace systems, what are the perceived weaknesses of the current system (or other motivation)?

We have the ability to create archives for replication. The process to make the replicas at DESY and NERSC is operator driven. Also, the tools to track replica status are bare bones, and somewhat plagued by data bugs.

Question 2C

Using the framework / terminology outlined above, can you describe your existing data management workflows?

> C. At what granularity do you manage data (individual files, set of files/datasets or sub-files/objects)? What are example replica management policies that you will have?

Data are collected at the file level and managed by “bundles” (tar/zip archive) for replication. Policy is driven by type. RAW is large and goes only to NERSC. Filtered (and L2/L3/L4) go to DESY and NERSC. Data should be retained until replication is confirmed.

Question 2D

Using the framework / terminology outlined above, can you describe your existing data management workflows?

> D. What are the requirements on responsiveness of the system? Are there any soft or hard deadlines when data replicas must be created?

Replication from the pole is ~24 hours, but this governed by network connectivity, not software. Replication to collaborators at DESY and NERSC is not deadline driven. We do want it to happen, but no fixed deadline (soft or hard) exists.

Question 2E

Do you use a workflow management system ? If yes, which one ? for which use cases will it need to interact with data management services ?

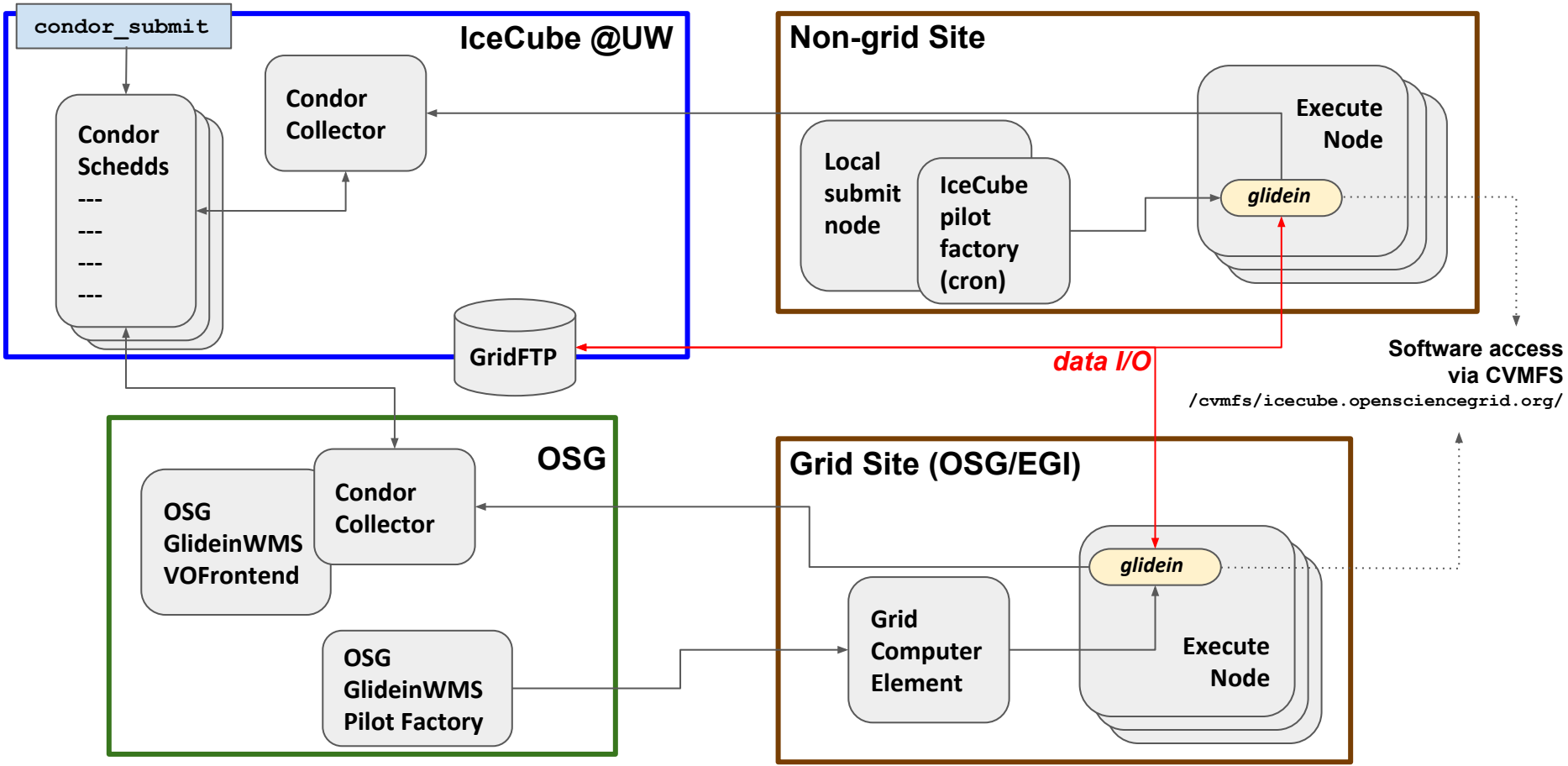
Users: HTCondor submit to a central pool. Use DAGman for workflows. Data I/O is “trivial” and handled by the Icecube data analysis framework:

Gridftp input from UW-Madison → data processing → Gridftp output to UW-Madison

Simulation Production: in-house developed framework IceProd that submits to HTCondor.

- Bookkeeping of simulation datasets (configuration settings, file locations …)
- So far, only two storage endpoints per dataset: intermediate and final files.
- For the future, we would like to extend this to handle multiple storage endpoints
  - If a job fails to store an intermediate file in the default SE, try another one. Subsequent jobs would need to know the location of that file.
  - Sites like TITAN might need data pre-stage into a local SE before/after processing.

IceCube Workflow



Question 2F

Do you use a metadata service (e.g., schema, data curation and characterization) ? If yes, do you foresee some interactions with data management services ? which ones ?

So far, we have not used a metadata service. Different s/w like JADE or IceProd store file metadata in their internal DBs. We are now starting a project to develop a central metadata catalog that both JADE and IceProd will use.

JADE indexes an NSF provided data format DIF. Some of our current work involves integration with our File Catalog metadata service. We foresee JADE North populating the catalog with satellite data, and JADE LTA updating the File Catalog with respect to replicas that exist at DESY and NERSC.

Question 3

For any potential new system, what are your order-magnitude-accuracy estimates for:

a. Number of files managed?
   We currently have ~500 million files. Aim for a system that can handle at least ~1 billion files.

b. Volume of data managed?
   Currently ~7 PB on disk at UW-Madison, ~1PB disk / 500TB tape at DESY, ~3 PB tape at NERSC
   Yearly growth: ~+1PB disk at UW-Madison, ~+500TB disk ~+200TB tape at DESY, ~+700TB tape at NERSC

c. Number of files transferred per day? Deleted per day?
   ~1 million reads ~300k writes per day via gridftp. We do not track deletes for now, so we do not have data.

d. Volume transferred per day?
   Grid ←→ UW: ~50-100 TB/day read and ~10-50 TB/day written
   UW → NERSC/DESY: up to ~50TB/day

e. Number of storage services managed? Types of storage (disk or tape)? Locations ?
   Just three: UW (disk), NERSC (tape), DESY (disk/tape)

f. Number of users/size of the collaboration ?
   ~400 users

Question 4

Is data management only done for “official” files or is the VO expected to manage user-produced data as well?

A. If user data is to be managed, how is access and ownership handled?

Data management covers data that is produced in the detector, or goes through our simulation and/or production workflows. A recent focus is capturing metadata for reproducibility. Access and ownership are currently handled with POSIX permissions in the Data Warehouse.

Our data management targets only “official” data for now. Final data products findable and accessible by the whole collaboration.

Question 5

If you have done any evaluation or study of Rucio, can you summarize your experience or progress?

A. For non-HEP experiments: Rucio was developed to meet the needs of a HEP experiment. Given any pre-existing knowledge of HEP computing you may have, do you foresee any requirements for your VO that might not be addressed by Rucio?

We have not done an evaluation of Rucio, but we hoping to learn where Rucio might fit with our data management needs during this conference.

Rucio Goals

IceCube would like to discover what Rucio can offer to our existing data management systems.

Our environment is generally open and receptive to new tools that are practical and stable.

https://github.com/rucio/rucio