CI CoE PILOT

Funded by the National Science Foundation Grant #1842042

Data Life Cycle (DLC) for Large Facilities

Anirban Mandal
Renaissance Computing Institute (RENCI), UNC – Chapel Hill
anirban@renci.org

CICoE Pilot: Anirban Mandal, Laura Christopherson, Erik Scott, Ilya Baldin, Paul Ruth (RENCI)
NEON: Philip Harvey, Steve Jacobs, Tom Gulbransen (NEON Large Facility, Boulder)
IceCube: Benedikt Riedel (Wisconsin IceCube Particle Astrophysics Center)

2019 NSF Workshop on Connecting Large Facilities and Cyberinfrastructure, Alexandria, VA, Sep 16, 2019

USC Viterbi School of Engineering *Information Sciences Institute*

UNIVERSITY OF NOTRE DAME

THE UNIVERSITY OF UTAH

CENTER FOR APPLIED CYBERSECURITY RESEARCH

renci

Data Life Cycle: Goals

- Understand and document the cyberinfrastructure (CI) best practices and solutions for data life cycle (DLC) for Large Facilities (LFs).
- Can a generalized DLC abstraction help us understand the diverse CI landscape for LFs ?
    - Can it be ***ONE way to learn/catalog the CI functionalities*** at each stage of data operation for LFs ?
    - What *services are offered* by each DLC stage ?
    - What *CI architectural elements* support each DLC stage ?
- Study the end-to-end life cycle for data as it traverses different CI entities inside a LF and then catalog the underlying services/tools/platforms for the life cycle stages.

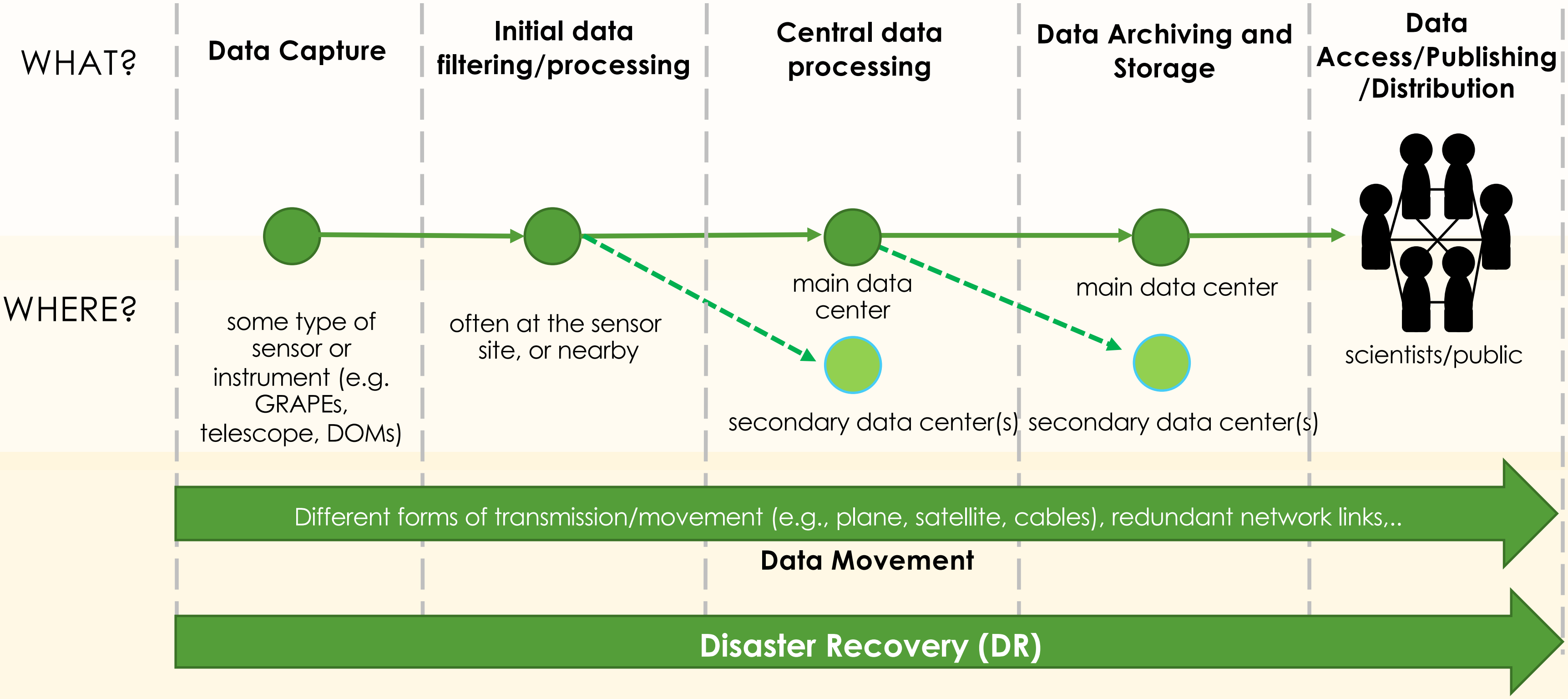
Data Life Cycle: Approach

- Develop initial DLC abstraction based on input from
  - engagement with NEON and IceCube LFs as part of the CICoE Pilot project,
  - the CI architectures submitted as part of the 2017 NSF LF CI workshop report, and
  - publicly available information about a collection of LFs.
- Develop DLC taxonomy and Disaster Recovery (DR) planning models
  - a taxonomy of CI services, architectures, and functionalities that support the different DLC stages for a collection of LFs.
  - effective process guides for DR planning for LFs in the context of DLC stages.
- **We are looking for feedback to refine our DLC abstraction and to extend or modify it as we learn about other LFs' CI and get valuable input from the LF CI community.**

Overview

- Data Life Cycle (DLC) abstraction for Large Facilities (LF)
- DLC case study - IceCube Large Facility (LF)
- Toward a taxonomy for DLC for LFs
- How can LFs utilize the DLC model ? – a Disaster Recovery (DR) example
- DR planning case study – IceCube LF
- Lessons learned from initial engagements

Data Life Cycle for LFs



WHAT?	Data Capture	Initial data filtering/processing	Central data processing	Data Archiving and Storage	Data Access/Publishing /Distribution
WHERE?	some type of sensor or instrument (e.g. GRAPEs, telescope, DOMs)	often at the sensor site, or nearby	main data center secondary data center(s)	main data center secondary data center(s)	scientists/public

Different forms of transmission/movement (e.g., plane, satellite, cables), redundant network links,..

Data Movement

Disaster Recovery (DR)

Data Life Cycle case study – IceCube Large Facility

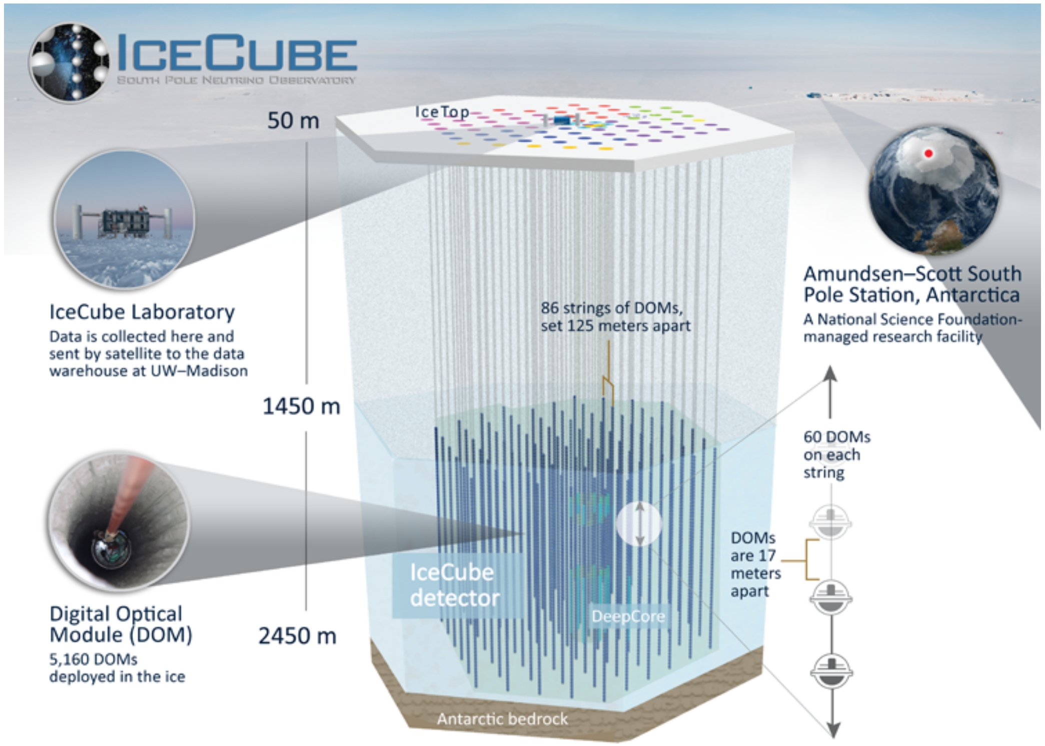
IceCube materials courtesy: Dr. Benedikt Riedel, Wisconsin IceCube Particle Astrophysics Center

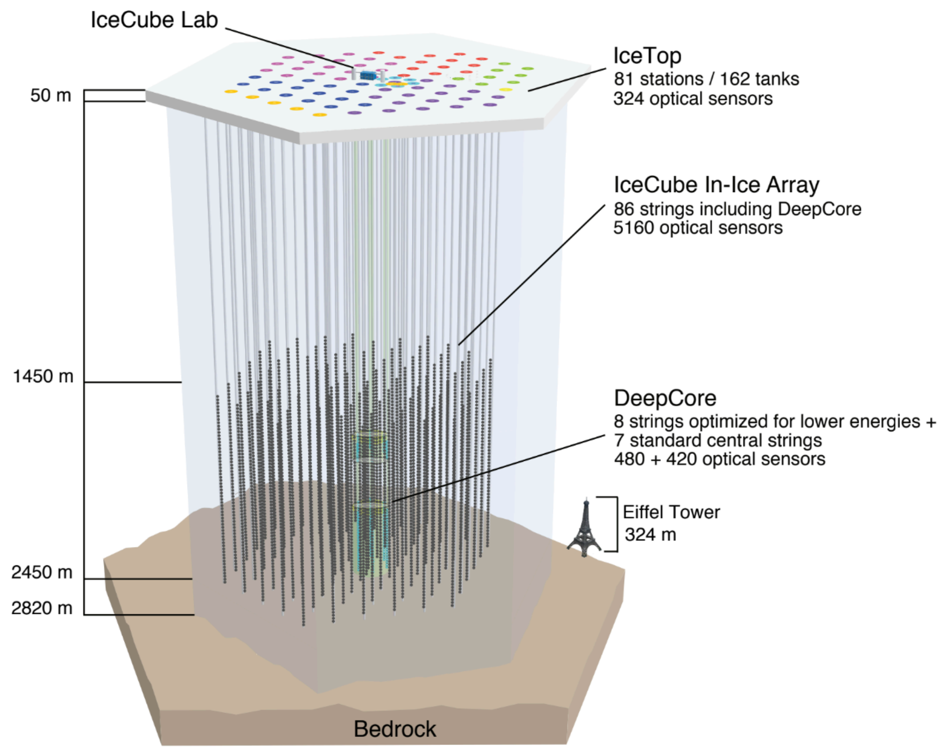
IceCube: About the Data

- Data represents
  - Hits
  - Events – time period of interest with fixed read out window
  - metadata and secondary streams (e.g., time calibration, monitoring)
- 4 types of data
  - *PFRAW* - full data set originating at the South Pole (~3TB/day)
  - *PFFILT Level 1* - ~100 GB/day of PFRAW that is filtered and send to UW-Madison
  - *Level 2* - Level 1 data that has directional reconstructions and is "science ready"
  - *Level 3* - Level 2 data that has been reduced, with extra reconstructions applied, by a particular science working group

CI CoE PILOT

IceCube Data Life Cycle: Data Capture


ICECUBE
SOUTH POLE NEUTRINO OBSERVATORY
50 m
IceTop
1450 m
2450 m
IceCube Laboratory
Data is collected here and sent by satellite to the data warehouse at UW–Madison
86 strings of DOMs, set 125 meters apart
Amundsen–Scott South Pole Station, Antarctica
A National Science Foundation-managed research facility
60 DOMs on each string
DOMs are 17 meters apart
IceCube detector
DeepCore
Digital Optical Module (DOM)
5,160 DOMs deployed in the ice
Antarctic bedrock



USC Viterbi School of Engineering *Information Sciences Institute*

UNIVERSITY OF NOTRE DAME

THE UNIVERSITY OF UTAH

CENTER FOR APPLIED CYBERSECURITY RESEARCH

renci

IceCube Data Life Cycle: Initial Processing/Filtering

Initial processing, filtering: South Pole

- Data is received by DOMHubs in the IceCube Lab (surface of South Pole)
- ~500 core filtering cluster; ~100 machines for detector readout
- Hits are output as events. Internal PnF system selects events based on their usefulness for a particular analysis. It also creates event metadata and reduces data volume before it is transmitted away from the South Pole.
- **Alert production** is an important process that happens in this stage of the DLC.

IceCube Data Life Cycle: Central Processing

Central processing: UW-Madison processes what is sent from the South Pole to a "science ready level" up to level 3.

- UW-Madison: 7600 core, 400 GPU cluster, ~10 PB storage. PFFILT → L2 and L3.
- Additional downstream processing happens using a mix of resources: DESY, OSG, IceCube Grid (campus clusters, contributed resources, etc.), XSEDE allocations, DOE resources (e.g. NERSC).
- Increased demand for GPU resources.
- PyGlidein + HTCondor based distributed computing middleware.
- Exploring cloud resources for CPU, GPU, ML.

IceCube Data Life Cycle: Data Movement

1. Hits at DOMs → DOMHubs → Data Acquisition System (DAQ) → Events (PFRAW)
2. Sent to Processing and Filtering System (PnF) - PFRAW made ready for analyses
3. Sent to South Pole Station JADE for archival storage to disk (PFRAW and PFFILT/Level 1)
4. JADE transmits via satellite to UW-Madison (PFFILT)
5. PFFILT sent to DESY and PFRAW sent to NERSC for additional tape backups

In addition, Alerts are sent out using GCN (Gamma Ray Coordination Network - operated by NASA) or Astronomical telegrams along with initial estimate of PFRAW data sample via satellite link to UW

- Limited bandwidth of ~125 GB/day from South Pole to UW; 3TB/day raw data is filtered down to ~80GB/day and transmitted via satellite from South Pole Station to UW
- Once a year, raw data from the South Pole is sent via plane, boat in disks to UW-Madison
- UW connected to SciDMZ through Starlight-ESNet for connection to DOE facilities
- Leverages GridFTP for data transfers from UW-Madison to DESY/NERSC/OSG

IceCube Data Life Cycle: Data Storage/Archiving

JADE (archival system) exists in ~3 locations

- South Pole JADE - writes 2 copies to disk (3 TB/day)
- JADE North (UW) - warehouses the data to disk (~200 TB/yr)
- JADE Long Term Archive (LTA) in DESY – keeps replicas of Level 1 and 2 data

NERSC archives PFRAW (the raw data) in tape archives.

IceCube Data Life Cycle: Access/Publishing/Distribution

Dissemination of Alerts

- Alerts happen at the South Pole during Level 1 processing.
- Alerting systems detect events and then an immediate alert is sent out using GCN (Gamma Ray Coordination Network - operated by NASA) or Astronomical telegrams along with initial estimate/small portion of PFRAW data sample via satellite link to UW.
- When a full PFFILT (Level 1) data set is available at UW later, a refinement of the first alert is sent.

IceCube Data Life Cycle: Access/Publishing/Distribution

3 forms of **data access** for other types of data:

- Be a member of the IceCube collaboration
- Be an "associate member" – one applies for use of the data for a particular purpose but is not required to fulfill collaboration obligations
- Public web portal

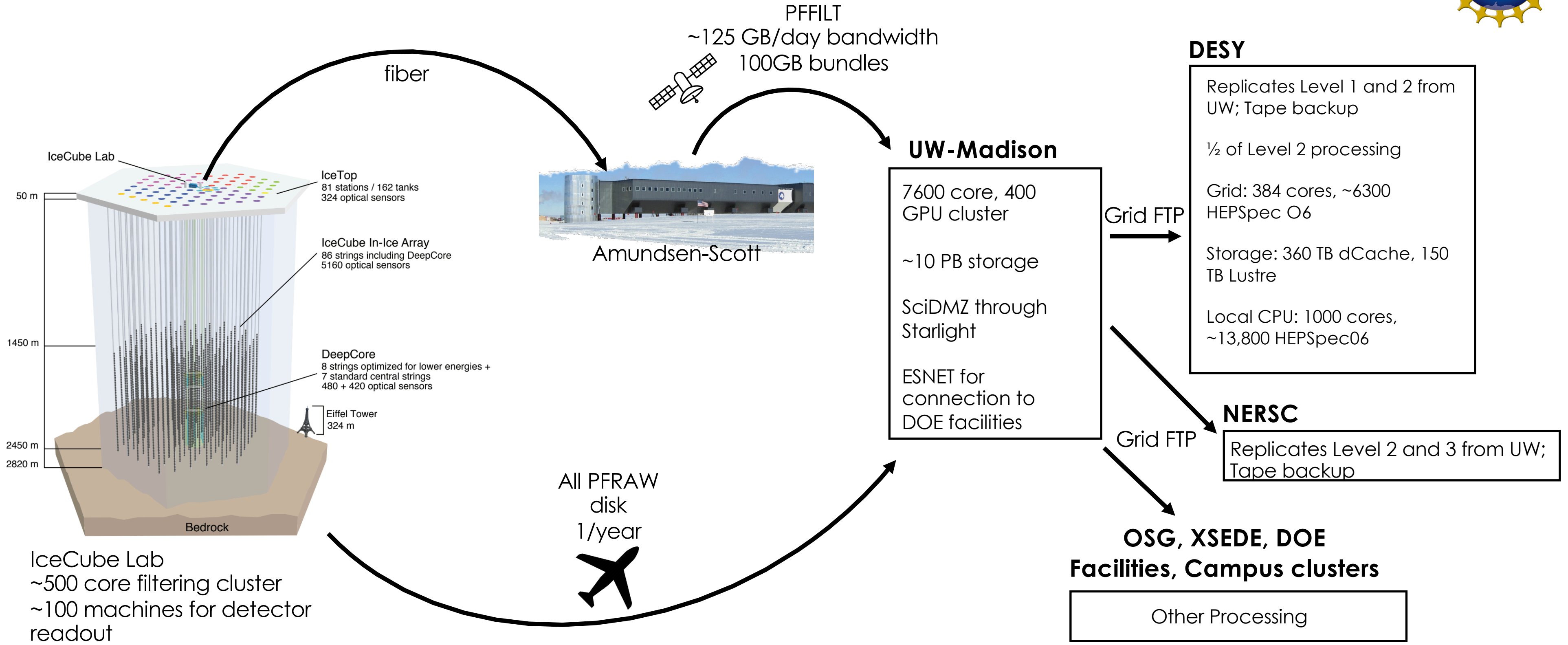
Planned enhancements for data organization, management, access, and data catalog

- Xrootd-based solution, Ceph/www

Data is released to members and associates. When the data has been analyzed and those analyses published, it becomes available for release to others.

CI CoE PILOT

IceCube DLC Logical Architecture



fiber

Amundsen-Scott

PFFILT
~125 GB/day bandwidth
100GB bundles

All PFRAW
disk
1/year

IceCube Lab
~500 core filtering cluster
~100 machines for detector readout
2 copies of all data archived

UW-Madison

7600 core, 400 GPU cluster

~10 PB storage

SciDMZ through Starlight

ESNET for connection to DOE facilities

Grid FTP

DESY

Replicates Level 1 and 2 from UW; Tape backup

½ of Level 2 processing

Grid: 384 cores, ~6300 HEPSpec O6

Storage: 360 TB dCache, 150 TB Lustre

Local CPU: 1000 cores, ~13,800 HEPSpec06

Grid FTP

NERSC

Replicates Level 2 and 3 from UW; Tape backup

OSG, XSEDE, DOE Facilities, Campus clusters

Other Processing

IceCube materials courtesy: Dr. Benedikt Riedel, Wisconsin IceCube Particle Astrophysics Center

USC Viterbi School of Engineering Information Sciences Institute

University of Notre Dame

The University of Utah

Center for Applied Cybersecurity Research

renci

Toward a Taxonomy for Data Life Cycle for LFs

- Can we document the current best practices for CI supporting each DLC stage for LFs ?
- Can that help uncover the shared CI challenges across DLC stages for LFs ?
- Can we infer the fundamental CI principles used to operate LFs' CI in the long-term ?

NEON materials courtesy: Tom Gulbransen, Battelle
OOI materials courtesy: Dr. Ivan Rodero, Rutgers Discovery Informatics Institute
IceCube materials courtesy: Dr. Benedikt Riedel, Wisconsin IceCube Particle Astrophysics Center

CI CoE PILOT

Data Life Cycle Taxonomy: Central Processing: Sample LFs (Distributed Data Capture)

NEON	• Iron Mountain Data Center (DEN-1) in Denver with planned 2nd data center in Wyoming. • Data in Oracle PDR (OS data), Elastic Cloud Storage (IS data), Common Object Storage (AOP data). • Computations on VMWare resource cluster with metadata/data pulled from Dell ECS and SC9000 SAN; 476 total cores, 8.5 TB RAM (total). • Runs transitions pipelines with Apache Airflow, Pachyderm for IS/OS/AOP data, produces L0 – L4 data.
OOI	• Rutgers Data Center with 2nd data center near Portland, OR connected with redundant I2 100G links. • 126 dual socket Xeon nodes; VMWare environment; Cassandra cluster; Palo Alto Firewalls, Raytheon uFrame framework operating on a ~2PB SAN (gold copy) and ~500TB NAS (user access). • Performs quality control, construct alerts/alarms, create calibration info, format data, and generate metadata; Performs synchronous/asynchronous processing based on user requests.
SAGE / IRIS	• Central Data center, aka. IRIS DMC in Seattle, WA + Auxiliary Data Center (ADC) in LLNL, with 10 Gb/s LAN backbone at DMC/ADC and between DMC and ADC. • VMWare on Dell, Forcepoint Firewalls, A10 Load Balancers, with data on Hitachi and NetApp RAID. • Currently also evaluating use of services on Amazon AWS and XSEDE. • Produces L0, L1 and L2 data for observations coming from about 30 types of sensors, worldwide.
GAGE / UNAVCO	• Primary data center is in Boulder, CO with an onsite failover; Other analysis centers at NMT, CWU, MIT. • VMWare on Dell; SAN storage (Oracle, Infotrend). • Investigating use of XSEDE, and deploying services in the cloud through Earthcube GeoSciCloud. • At Boulder, L0 to L1, quality controlled, and archived; At NMT, CWU: L1 to L2a; At MIT: L2a to L2b

USC Viterbi School of Engineering Information Sciences Institute

University of Notre Dame

The University of Utah

Center for Applied Cybersecurity Research

renci

Data Life Cycle Taxonomy: Data Storage/Archiving: Sample LFs (Distributed Data Capture)

NEON

- Dell EMC SC9000 SAN: Utility storage for VMs, databases, and NAS.
- Dell Elastic Cloud Storage (ECS): DEN1 Primary (2PB); DEN2 Replica (2PB); DEN3 Development (200TB).
- Syncs/Backups: DEN1 Primary ECS synced with Amazon AWS Glacier Deep Storage offsite; Unitrends Backup Server (160 TB) in DEN-1 with replica in Boulder HQ; Veeam and Backup Server in DEN-1 with replica in Boulder HQ.

OOI

- Rutgers Data Center - Cassandra cluster (~50TB) for raw, SAN (~2PB) for formatted/"gold", NAS for user access (~500TB); Tape archive (~18PB) - SAN, Cassandra, and VMWare are backed up to tape.
- Portland Data Center has identical storage setup (no Tape); each partner keeps a copy of the data.
- Overall capacity: 25PB.

SAGE / IRIS

- Both DMC and ADC have copies of data and identical storage setup.
- Large volume Hitachi RAID with RAID contents indexed in PostgreSql DBMS for better performance.
- High performance NetApp RAID system for ingest of real-time data and PostgreSql DB transactions.
- Internal/external access to storage transitioning to web-services based access.

GAGE / UNAVCO

- Primary Data Center at Boulder, CO: SAN storage (Oracle and Infortrend) used for Long-term storage archive + Metadata DB + FTP Data server with onsite mirror/backup/failover.
- Offsite storage backup (IRIS tape and Amazon AWS); Offsite Metadata DB failover (FRII Colocation) and Offsite FTP Data server failover (FRII colocation).

Data Life Cycle Taxonomy: Central Processing: Sample LFs (Centralized i.e. 1-2 Site Data Capture)

IceCube	• UW-Madison: 7600 core CPU, 400 GPU cluster, ~10 PB storage. L1 → L2 data. • Additional processing done on a mix of resources: DESY, OSG, IceCube Grid (campus clusters, contributed resources, etc.), XSEDE allocations, DOE resources (e.g. NERSC). • PyGlidein + HTCondor based distributed computing middleware. • Exploring cloud resources for CPU, GPU, ML. Increased demand for GPU resources.
LSST	• NCSA: Computing ~18,000 cores: Nightly/Alert production, 50% of Data Release production, Calibration production, moving object • CC-IN2P3, France: other 50% of Data Release production
LIGO	• Parallel workflows executed on HTC resources: LIGO Scientific Collaboration clusters; Data on Oracle (HSM, ZFS), HDFS • External shared resources integrated via the Open Science Grid (e.g., Virgo, XSEDE, universities). • HTCondor for job scheduling, DAGMan and Pegasus WMS for workflow management. • BOINC infrastructure to manage search for continuous wave signals via Einstein@Home. • Shibboletch, Grouper, InCommon, and CILogon for identity/access management; Kerberos, LDAP, GSI; Docker/Singularity/Shifter; • CVMFS, StashCache/Xrootd, Globus GridFTP, and in-house tools for managing distributed data.

Data Life Cycle Taxonomy: Initial Processing/Filtering: Sample LFs (Centralized i.e. 1-2 Site Data Capture)

IceCube	• Data received by DOMHubs at IceCube Lab: ~100 nodes detector readout; ~500 core filtering cluster; • Hits output as events. PnF system selects events based on usefulness for a particular analysis. It also creates event metadata and reduces data volume (PFRAW to PFFILT/L1). • Alert production is an important process that happens in this stage of the DLC.
LSST	• Base Facility, La Serena Chile: real time alert generation, initial detector cross-talk correction, metadata creation. • Computing capacity: ~ 2400-3000 cores.
GEMINI	• Base Facilities in La Serena, Chile and Hilo, Hawaii: Substantial computing; Has a virtual machine cluster and physical server farm; Real time cross/off-site replication of data;

Similar characteristics for **LIGO, NRAO, DKIST/NSO**

Data Life Cycle Taxonomy: Data Movement: Sample LFs (Mobile Data Capture)

Academic Research Vessels + R2R	• Data is delivered via "sneakernet" to R2R project. • Working on standardized real time delivery of data from ships with HiSeasNet satellite communications. • R2R moves and archives one copy of raw data to NOAA and another copy to Amazon Glacier.
Regional Class Research Vessels	• Sensors → RPi Sensor network interface (Analog/Serial observations → UDP with XML payload) • RPi Sensor network → Isolated data distribution network (10G Fiber & CAT7 backbone) • Data distribution network → Shipside datastore (Data aggregation, parsing creates files from UDP) • Shipside datastore → Shoreside datastore via satellite communications (also used in other direction)
JOIDES Resolution / IODP	• VSAT (very small aperture terminal) for ship to shore satellite connectivity using a dedicated asynchronous WAN circuit, 2Mbps down to the ship, and 1 Mbps up; Data downloaded at TAMU. • Data is also moved to the NCEI facility in Boulder, CO for archiving.

Data Life Cycle Taxonomy: Data Movement: Sample LFs (Centralized i.e. 1-2 Site Data Capture)

IceCube	• IceCube Lab → South Pole Station JADE for archival storage to disk for L0 and L1: Dedicated fiber link. • South Pole JADE → UW-Madison for L1: satellite link with limited bandwidth of ~125 GB/day. • UW-Madison → DESY/OSG/Other Compute resources for L1 and L2 using GridFTP. • UW-Madison → NERSC for tape backup of L0 using GridFTP leveraging SciDMZ through StarLight-ESNet. • Once a year, raw data from the South Pole is sent via plane and disks to UW-Madison.
LSST	• Telescope at Summit in Cerro Pachon, Chile → Base Facility in La Serena, Chile: 600 Gb/s dedicated circuit. • Base Facility → NCSA, Illinois: 2 x 100 Gbps redundant network with dedicated circuits. • NCSA → CC-IN2P3, France
GEMINI	• Redundant core network (I2 + REUNA) between GEMINI North and GEMINI South connected through VPN tunnels.

How can Large Facilities utilize the DLC abstraction ?

- Disaster Recovery (DR) Planning example

Disaster Recovery (DR) Effective Processes

- **Cross-cutting finding:** Although some DR strategies exist across some stages of the DLC for some LFs, DR hasn't been taken into account to the fullest extent it warrants when designing the CI architecture for LFs.
- There is a need for some careful consideration of **requirement analysis and planning for DR** as an effective process to be followed **before and after** a possible disaster.
- Developing an effective processes guide for planning for Disaster Recovery for LFs
  - **DR Planning Phase template** that Large Facilities can follow for planning for Disaster Recovery.
  - Based on federal guidance for developing an *Information System Contingency Plan (ISCP)* after doing a thorough *Business Impact Assessment (BIA)* – **NIST 800-34r1** (https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-34r1.pdf)

Disaster Recovery Effective Processes: ISCP

- **Information System Contingency Planning (ISCP) major steps**
  - **Overview:** Top-level view of CI DR needs of the LF and serves as a summary and context.
  - **Conduct Business Impact Analysis (BIA)**
    - **Mission, data life cycle, and recovery criticality:** CI for the entire data life cycle is identified and impact of disruption to those systems is determined along with ***outage impacts*** and ***estimated downtimes***.
    - **Resource requirements:** Thorough evaluation of resources required to restore systems and processes supporting the data life cycle and related interdependencies.
    - **Recovery priorities for system resources:** Link system resources to critical mission and business processes for LFs and establish priority levels for sequencing recovery activities and resources.
  - **Create Contingency Strategies**
    - (a) Backup and Recovery; (b) Backup methods and Offsite Storage; (c)Alternate Sites; (d) Equipment Replacement; (e) Cost Considerations; (f) Roles and Responsibilities;)g) Plan testing, training and exercises

We created example DR Planning Phase templates for NEON and IceCube and engaged with them to validate and refine the template. Example templates on CI CoE Pilot website. Plan to engage with other LFs.

DR Effective Processes: BIA: Mission, DLC, Criticality

- **Mission and data life cycle example (IceCube)**

Mission/Business Process	Description
Collect observational data for scientific inquiry	Operate an array of photodetectors buried in a cubic kilometer of ice at the south pole
Ingest data from sensors for further use	Operate networks at the south pole, satellite and physical data transport to UW-Madison, and bulk internet transfer to DESY-ZN and to NERSC (National Energy Research Scientific Computing Center). Notification of neutrino events to worldwide observatories, both radio and optical.
Process data for QA, events, and production of private and public datasets	Operate clusters of computers to analyze data for interesting events, QA, and calibration. Data reduction for low bandwidth satellite link.
Archive data for future use	Operate a curated repository of collected and processed data redundantly copied across three institutions.
Disseminate data	Provide **alerts** for interesting astrophysical events, i.e. during L1 analysis, if an event of interest is detected, an alert is sent out immediately. Operate an access portal to support querying the collection and accessing the stored and/or computed results. Provide well documented access methods and ensure that access methods can be added into the far future as needed.

DR Effective Processes: BIA: Mission, DLC, Criticality

- **Recovery Criticality – Outage Impacts example (IceCube)**

Outage Impacts

Impact category: {Severe, Moderate, Minimal}

Data Life Cycle Stage	**Impact Category**			
	Mission Impact	**Science Return**	**Cost**	**Impact**
Collect observational data for scientific inquiry	Severe	Severe	Severe	Severe
Ingest data from sensors for further use	Severe	Moderate	Moderate	Severe
Process Data for Multimessenger GCN Alerts	Severe	Severe	Moderate	Severe
Process data for QA, and production of private and public datasets	Moderate	Moderate	Minimal	Moderate
Archive data for future use	Severe	Severe	Severe	Severe
Disseminate data - Alerts	Severe	Severe	Minimal (short term)	Severe
Disseminate data – Level 2+ processed products	Moderate	Minimal	Minimal	Moderate

DR Effective Processes: BIA: Mission, DLC, Criticality

- **Recovery Criticality – Estimated Downtimes example (IceCube)**

Estimated Downtime

MTD: Maximum Tolerable Downtime
RTO: Recovery Time Objective
RPO: Recovery Point Objective

Data Lifecycle Stage	MTD	RTO	RPO
Collect sensor data for scientific inquiry – operation of the actual sensors	1 hour	1 hour	N/A
Ingest data from sensors for further use	Optical DOMs to IC Lab: 2 days; IC Lab to South Pole Lab: 6 months	1 day from DOM; 1 month from lab to station	1 day
Process data for QA, events, and production of private and public datasets	Events: 1 hour; All others: 1 month	Events: 30 minutes; Others: 4 days	Equal to recovery time
Archive sensor data for future use	6 months w/ no loss	1 hr	1 day
Disseminate data - alerts	6 hours	3 hours	None
Dissemination – Level 2+	1 year	1 month	1 day

DR Effective Processes: BIA: Resource Requirements

- **Resource requirements example (IceCube)**
  - For each step in the Data Lifecycle and the identified level of recovery criticality, describe resources used for day-to-day operation and the resources needed for recovery in the event of an outage.

System Resource/Component	Platform/OS/Version (as applicable)	Description
Collect observational data for scientific inquiry	DOMs (custom)	Sensors buried in the ice. Also power, IceCube Laboratory Facilities, staff
Ingest data from sensors for further use	Data communication to DOMs, network to South Pole Station, rest of world	Communications from DOMs, network to South Pole station, satellite and sneakernet to rest of world. JADE software for archive and distribution worldwide.
Process data for QA and for computed results	Server clusters, other CPU/GPU resources, HTCondor compute middleware, data distribution software	Datacenters at IceCube Lab and South Pole Station with special challenges. Data centers at UW-Madison, DESY, and NERSC. Distributed computing on OSG, XSEDE, DOE resources.
Archive sensor data for future use	JADE	Long term archive and archive management
Disseminate data	Web servers, data access/distribution middleware	Public-facing Portal, other data distribution methods (xrootd, ceph etc.)

DR Effective Processes: BIA: Recovery Priorities

- **Recovery priorities example (IceCube)**

Priority	**System Resource/Component**	**Recovery Time Objective**
1	***Collect environmental data for scientific inquiry***	1 hour
4	***Ingest data from sensors for further use***	Optical DOMs to IC Lab: 1 day; IC Lab to South Pole Lab: 2 months
5	***Process data for QA and for intermediate results***	Events: 30 minutes Others: 4 days
3	***Archive sensor data for future use***	1 hour
6	***Disseminate data – Level 2+***	1 month
2	***Disseminate Alerts***	3 hours

CI CoE PILOT

Data Life Cycle and DR Planning: Lessons Learned

- DLC can be **ONE** way to **learn, reason and catalog the CI functionalities** at each stage of data operation for LFs.
- DLC abstraction helps reasoning about
  - What ***services are offered*** by each DLC stage ?
  - What ***CI architectural elements support*** each DLC stage ?
- DLC taxonomy helps us uncover ***commonalities and differences in CI architectural choices***, and to potentially ***discover shared CI challenges*** across DLC for LFs.
- ***Heterogeneity*** *of data processing* – the set of processes handling the data differs according to the type of data; Heterogeneity of tools and CI stacks.

Data Life Cycle and DR Planning: Lessons Learned

- DR planning template was a good way to start thinking about disaster recovery – a ***framework to document and prioritize CI architecture and operations***.
- DR planning template can be used by LFs to ***quantitatively justify CI elements*** for construction or future enhancements.
- ***Effective communication*** *between LF CI professionals and CICoE Pilot team* was key to thorough understanding; Documentations and reports are necessary but not sufficient.

Funded by the National Science Foundation Grant #1842042

Thank you !!

Questions ?

Please send feedback to

cicoe-pilot@isi.edu or anirban@renci.org