NOAA’s NCDC iRODS Implementation

iRODS User’s Meeting

June 18, 2014

Alan Hall

NOAA Enterprise Systems

Comprehensive Large Array Stewardship System (CLASS)

Enterprise Systems

Support Services

Enterprise Ground Segments

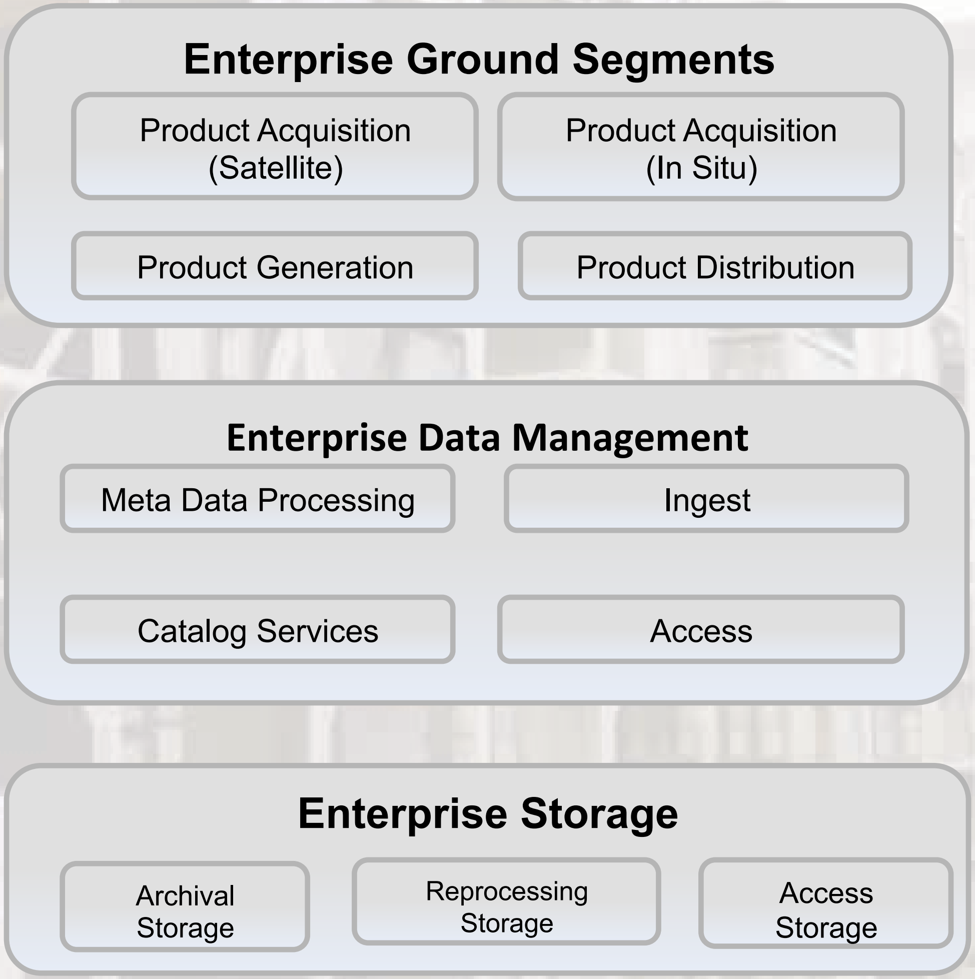
- Retrieve and distribute Satellite data and products
- Retrieve ground base measurements
- Retrieve Model outputs
- Retrieve derived products
- Distribute and/or Submit to Archive

Enterprise Data Management

- Extract meta data during ingest
- Stewardship of metadata
- Document metadata changes
- Provide catalog services
- Enable access and dissemination
- Search and Discovery

Enterprise Storage

- Long term preservation
- Guaranteed delivery and integrity
- Data Migration services
- Cloud Storage



Enterprise Ground Segments

Product Acquisition (Satellite)

Product Acquisition (In Situ)

Product Generation

Product Distribution

Enterprise Data Management

Meta Data Processing

Ingest

Catalog Services

Access

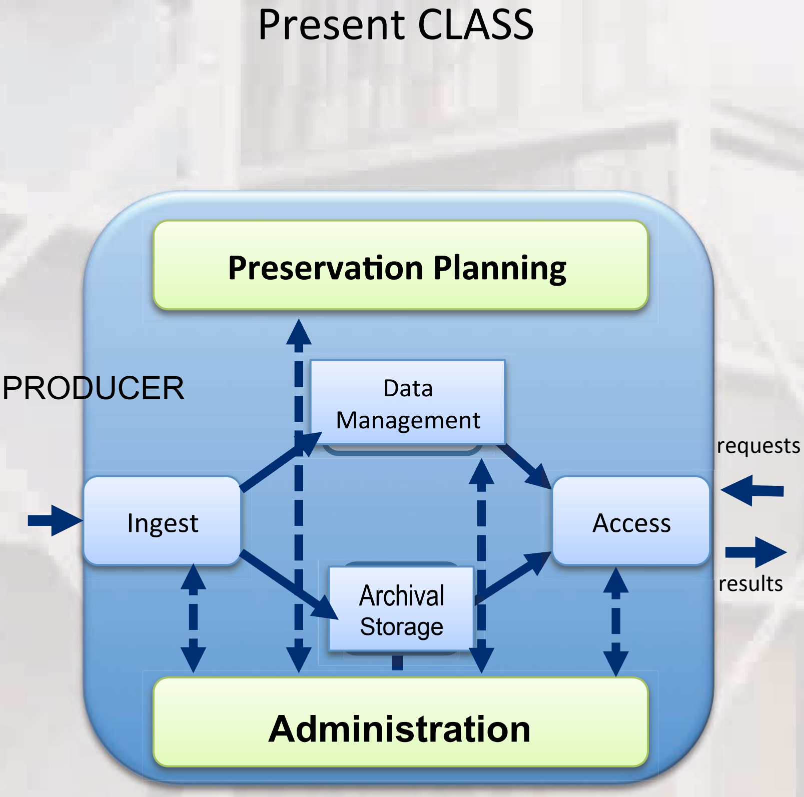
Enterprise Storage

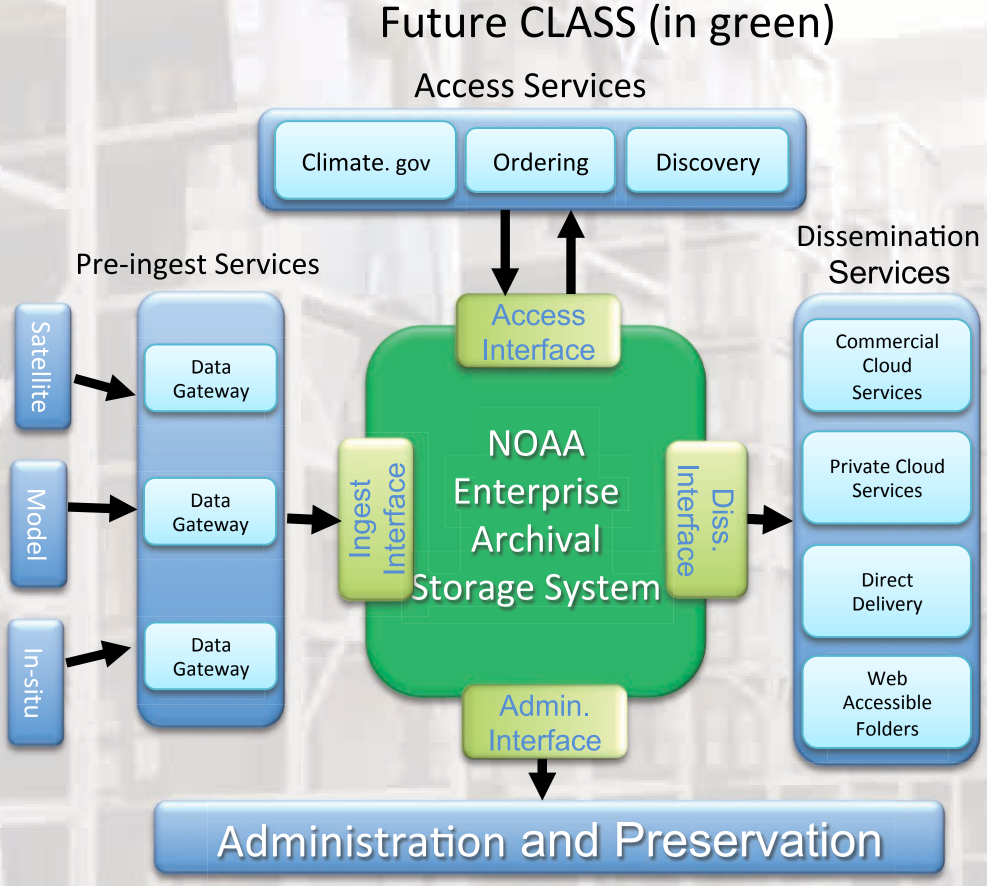
Archival Storage

Reprocessing Storage

Access Storage

CLASS Evolution to an Enterprise Archival System (EAS)





CLASS provides components of the OAIS-RM as services and interfaces.
Additional systems and services implemented by the data centers.

OAIS-RM

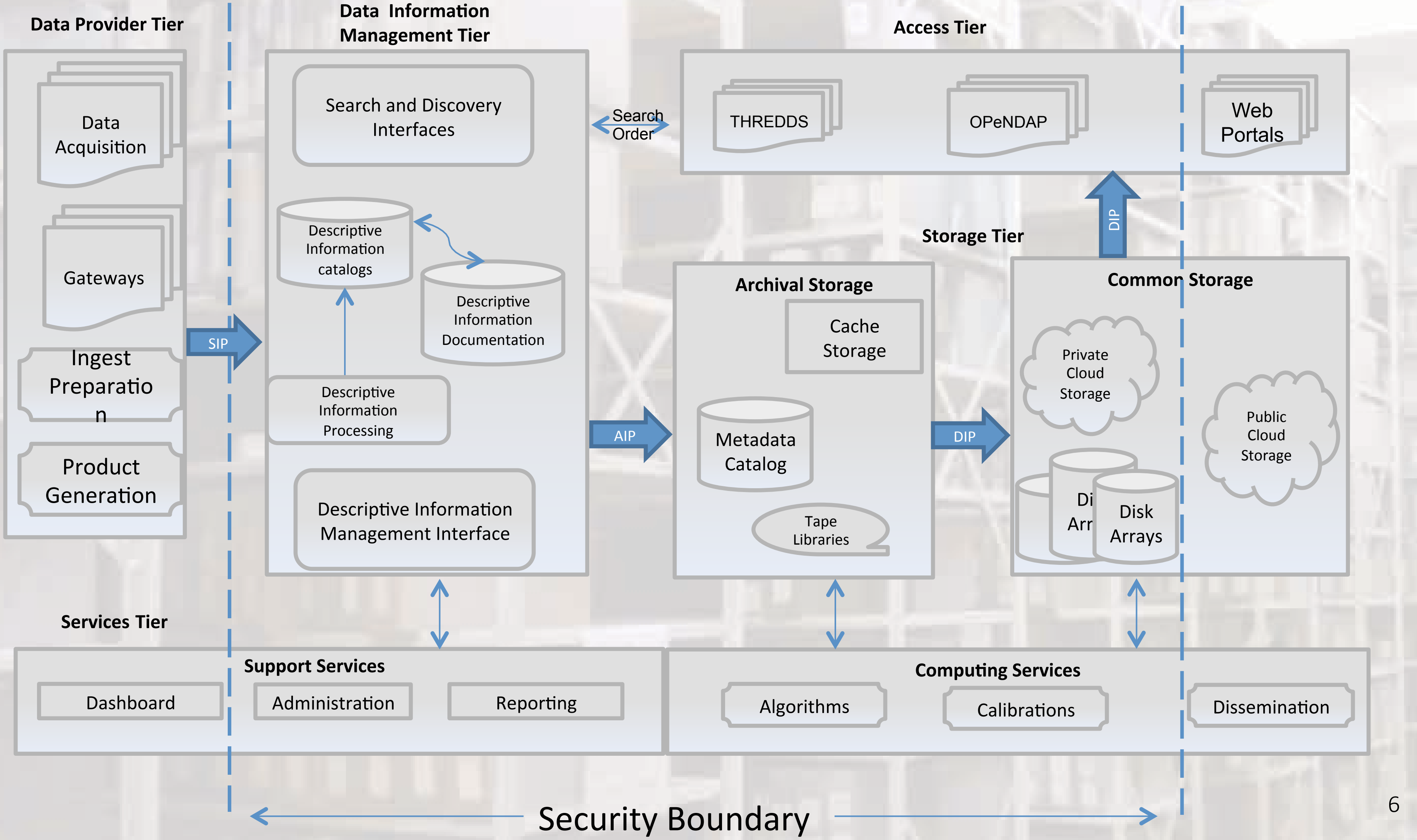
Open Archival Information System – Reference Model (OAIS-RM):
An Archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community.

Submission Information Package (**SIP**): An Information Package that is delivered by the Producer to the OAIS for use in the construction or update of one or more AIPs and/or the associated Descriptive Information.

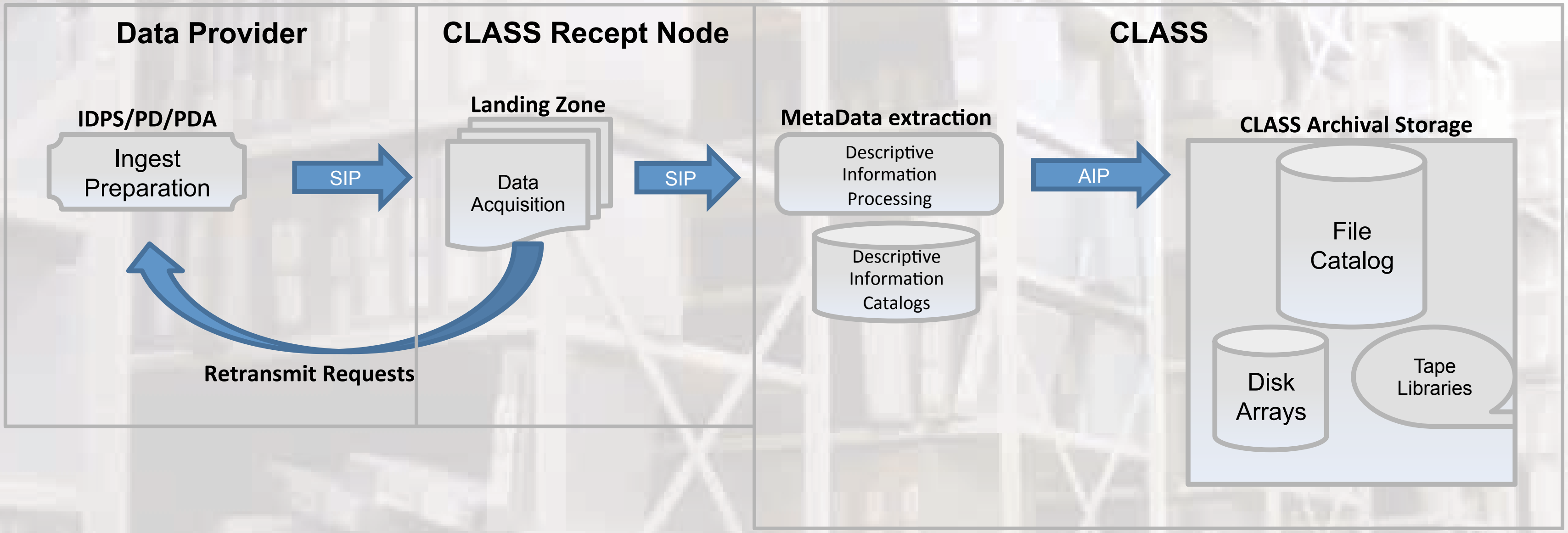
Archival Information Package (**AIP**): An Information Package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an OAIS.

Dissemination Information Package (**DIP**): An Information Package, derived from one or more AIPs, and sent by Archives to the Consumer in response to a request to the OAIS.

NOAA Enterprise Archive Tiered Architecture Components



CLASS Ingest Workflow

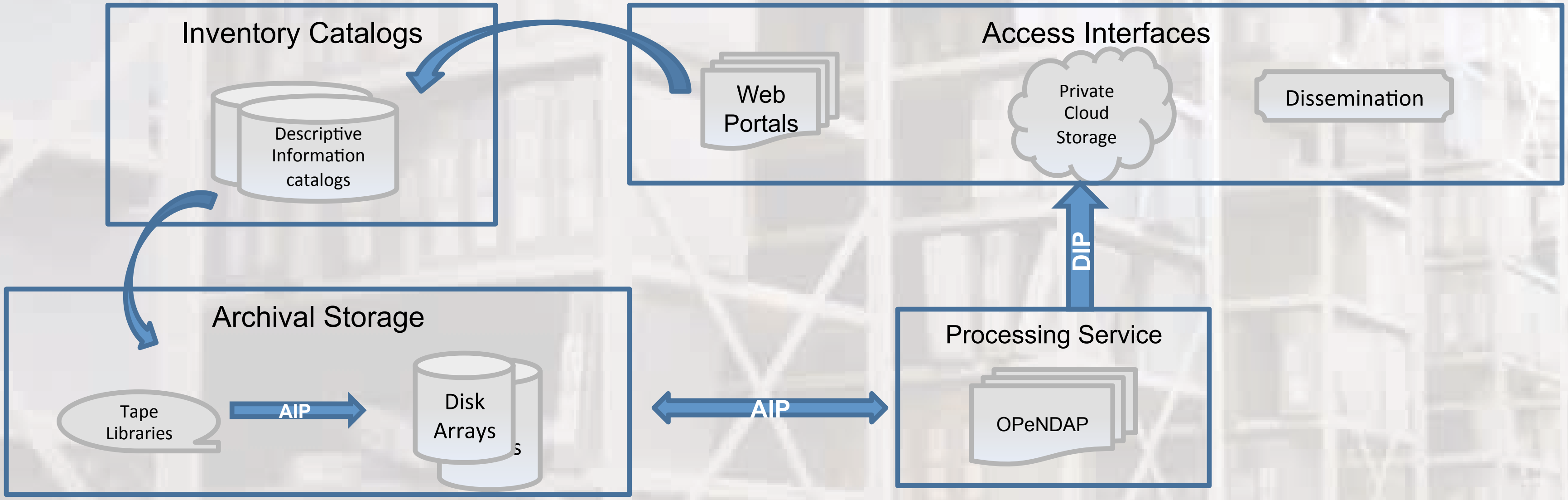


CLASS data arrives in structured packages at predictable periodicity

- CLASS Providers are well defined and documented (ICD)
- CLASS data predictable and well structured
- Standardized formats based on programs
- Submission packages generally match archive packages
- No transformations, repackaging, or other processing

OAIS Reference Model:
SIP – Submission Information Package
AIP – Archive Information Package
DIP – Dissemination Information Package

CLASS Order Workflow

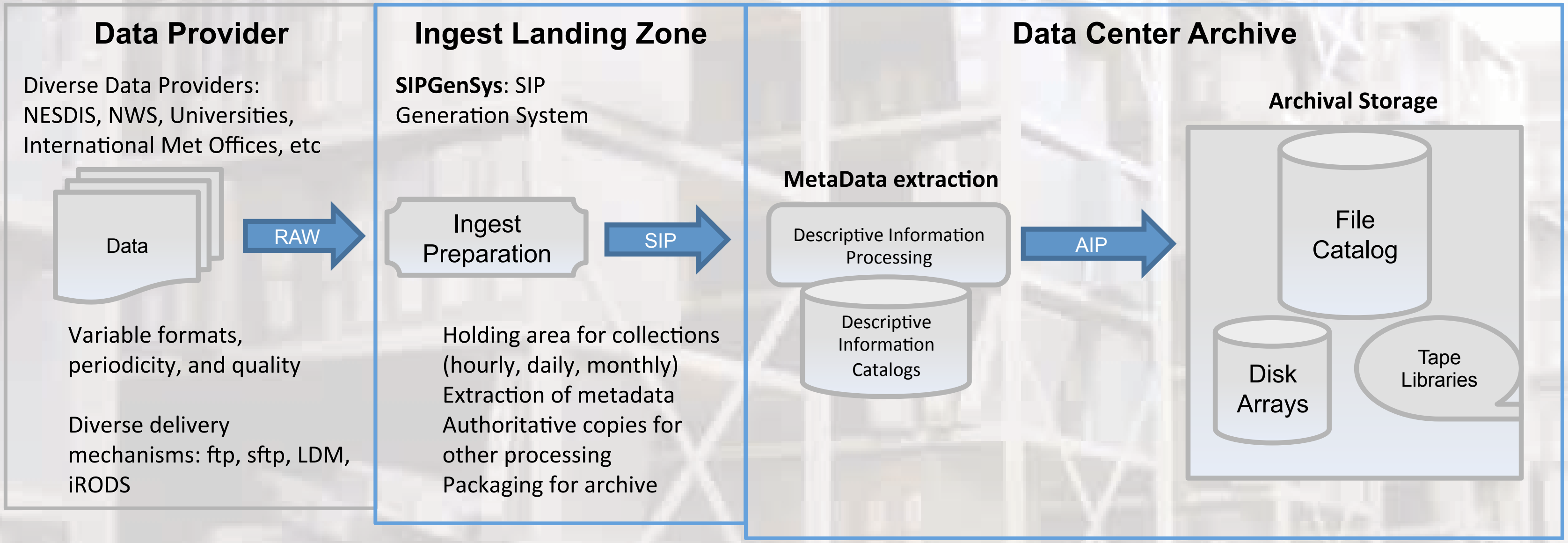


Order workflow

1) User Queries catalogs and places a order
2) iRODS accepts order and stages data to disk arrays
3) iRODS calls OPenDAP subsetting routines
4) iRODS copis data to Access storage for dissemination.

OAIS Reference Model:
SIP – Submission Information Package
AIP – Archive Information Package
DIP – Dissemination Information Package

Data Center Ingest/Archive Workflow



Data Provider

Diverse Data Providers: NESDIS, NWS, Universities, International Met Offices, etc

Variable formats, periodicity, and quality

Diverse delivery mechanisms: ftp, sftp, LDM, iRODS

Ingest Landing Zone

SIPGenSys: SIP Generation System

Holding area for collections (hourly, daily, monthly)
Extraction of metadata
Authoritative copies for other processing
Packaging for archive

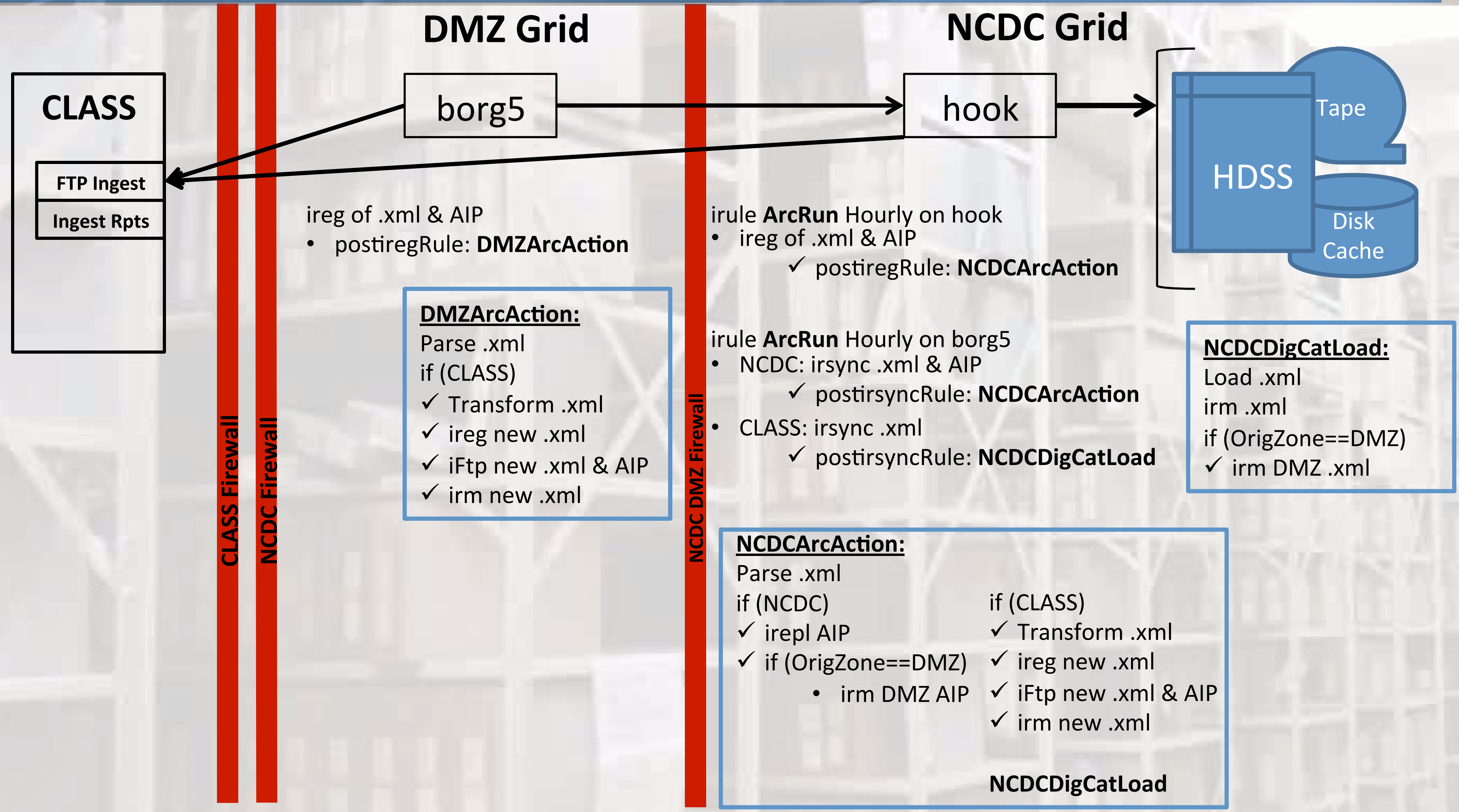
Data Center Archive

iRODS Workflow:

- Time-based script is executed on remote servers in the DMZ via iRODS
- **Event Driven:**
  - ✓ SIP is registered on creation
  - ✓ SIP is irsync'd to NCDC grid
  - ✓ SIP → AIP is irepl'd to NCDC HPSS MassStore
  - ✓ SIP is removed from DMZ grid
  - ✓ AIP is itrim'd after 24 hours

OAIS Reference Model:
SIP – Submission Information Package
AIP – Archive Information Package
DIP – Dissemination Information Package

iRODS Event Driven Workflow
A Prototype



DMZ Grid

ireg of .xml & AIP

- postiregRule: **DMZArcAction**

DMZArcAction:

Parse .xml

if (CLASS)

- ✓ Transform .xml
- ✓ ireg new .xml
- ✓ iFtp new .xml & AIP
- ✓ irm new .xml

NCDC Grid

irule **ArcRun** Hourly on hook

- ireg of .xml & AIP
  - ✓ postiregRule: **NCDCArcAction**

irule **ArcRun** Hourly on borg5

- NCDC: irsync .xml & AIP
  - ✓ postirsyncRule: **NCDCArcAction**
- CLASS: irsync .xml
  - ✓ postirsyncRule: **NCDCDigCatLoad**

NCDCDigCatLoad:

Load .xml

irm .xml

if (OrigZone==DMZ)

- ✓ irm DMZ .xml

NCDCArcAction:

Parse .xml

if (NCDC)

- ✓ irepl AIP
- ✓ if (OrigZone==DMZ)
  - irm DMZ AIP

if (CLASS)

- ✓ Transform .xml
- ✓ ireg new .xml
- ✓ iFtp new .xml & AIP
- ✓ irm new .xml

NCDCDigCatLoad

Python Script

iRODS Even Driven Workflow

Workflow:

Still time-based archive

- ~10M+ SIPs are received each day
    - Most SIPs are removed after AIP created
    - Not managed in iRODS (seemed overkill)
- ~25K+ AIPs are archived each day
    - iRODS will manage the AIPs, these are delivered to customers
- ~1TB+/day

Replaces crons on local servers:

- Workflows are coordinated between grids
- Control workflow from within firewall

Simplifies workflow:

- Event-driven execution and Remote-rule execution
- Reduces coordination of crons across many servers
    - time-based execution is very cumbersome to coordinate