Migrating NASA Archives to Disk: Challenges and Opportunities

NASA Langley Research Center
Chris Harris
June 2, 2015

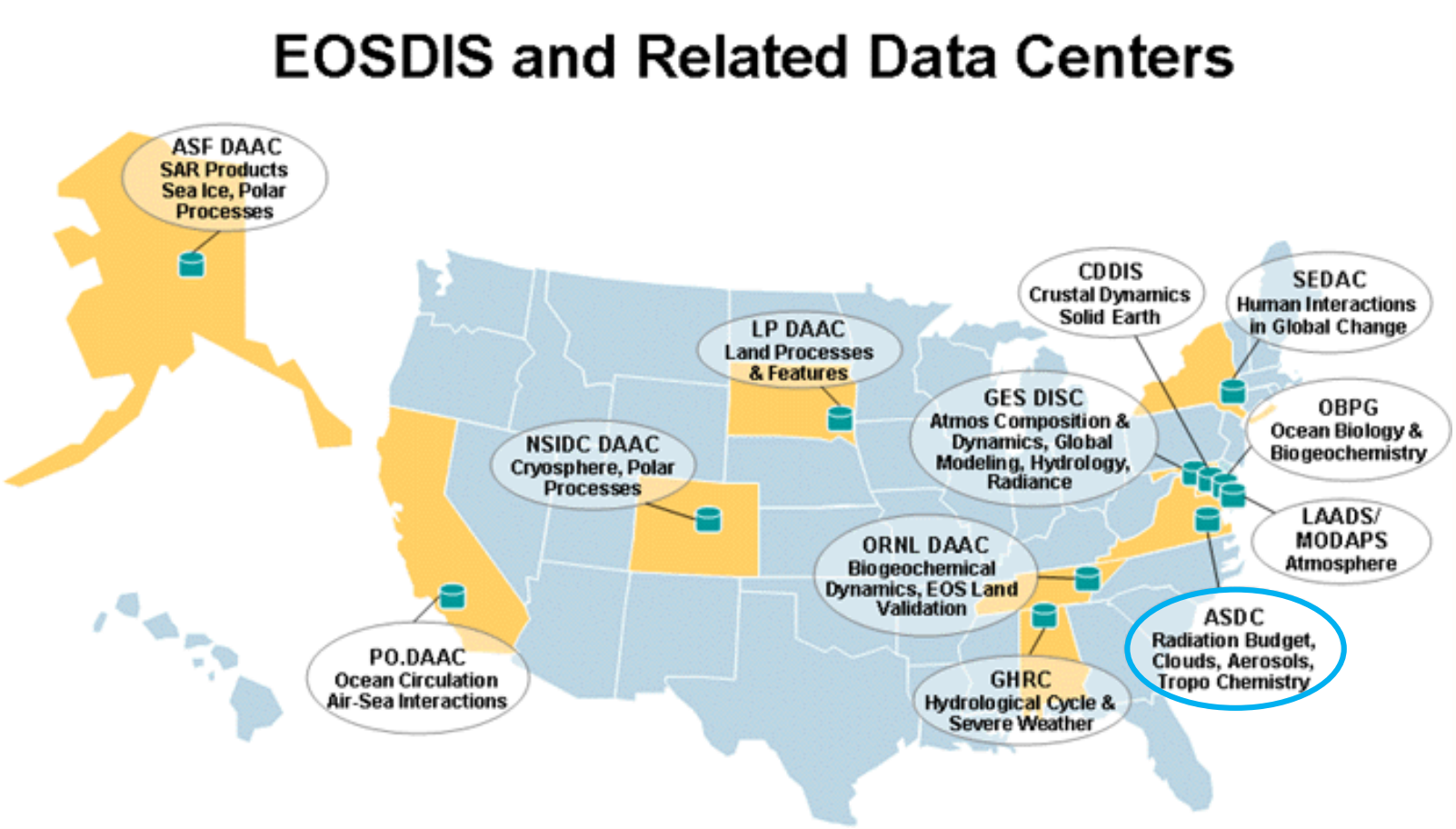
MSST 2015

Topics

- ASDC – Who we are? What we do?
- Evolution of storage technologies
- Why we need an online archive? (opportunities)
- Why we still need tape?
- Challenges
- What lies ahead?
- Summary

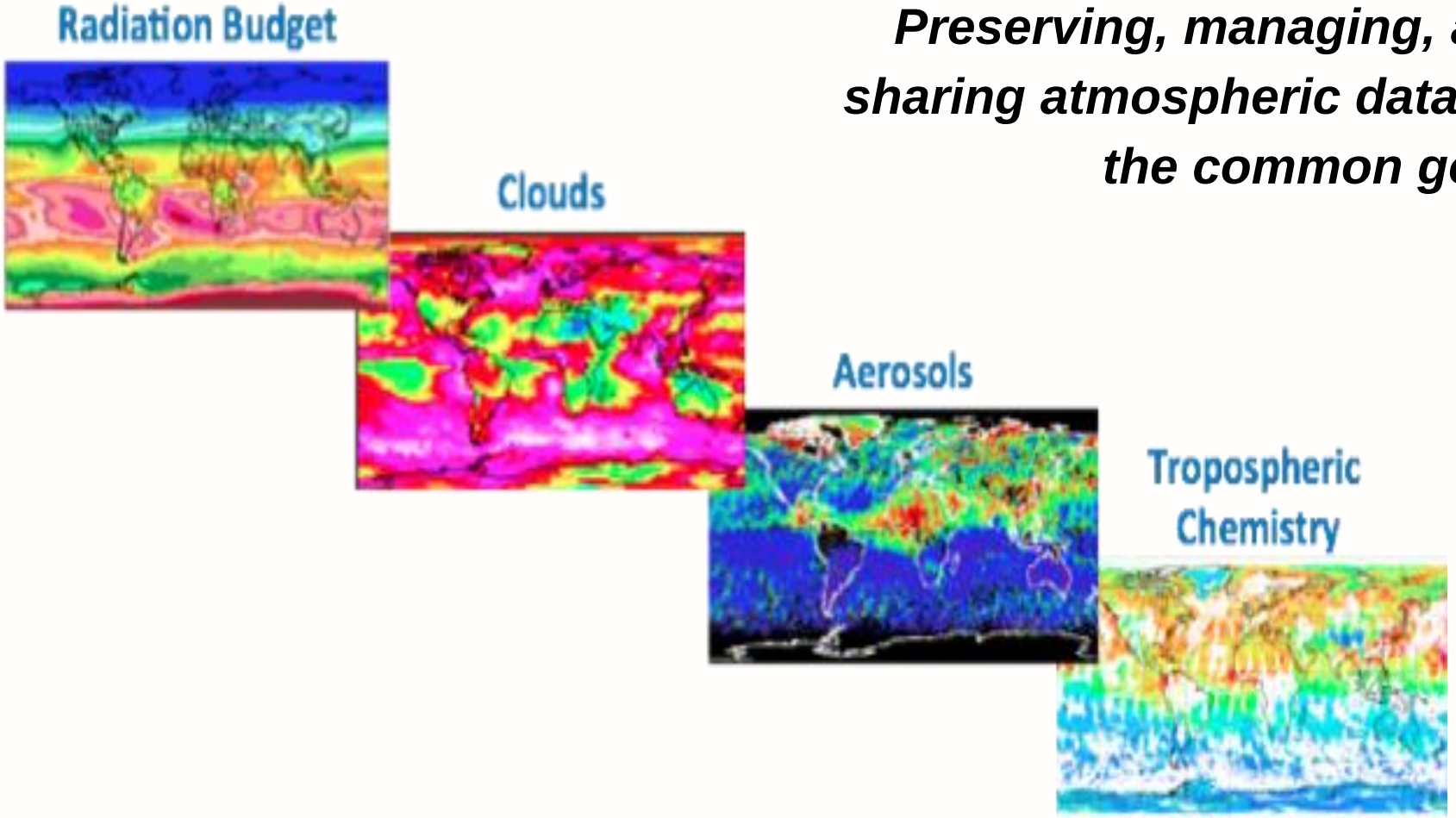
ASDC Under ESDIS

The Atmospheric Science Data Center (ASDC) is chartered under the Earth Science Data and Information System (ESDIS) Project at NASA GSFC.



NASA Langley Research Center's Atmospheric Science Data Center (ASDC)

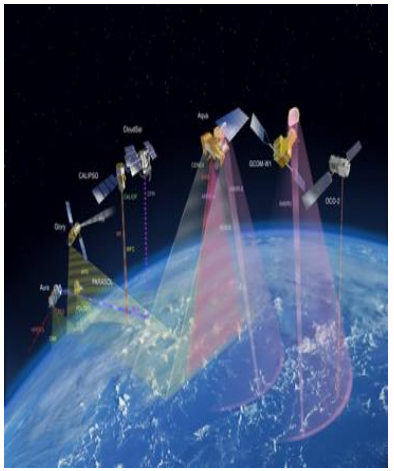
Preserving, managing, and sharing atmospheric data for the common good



Primary Functions

Receive (*Ingest*) data to archive and support science driven requirements

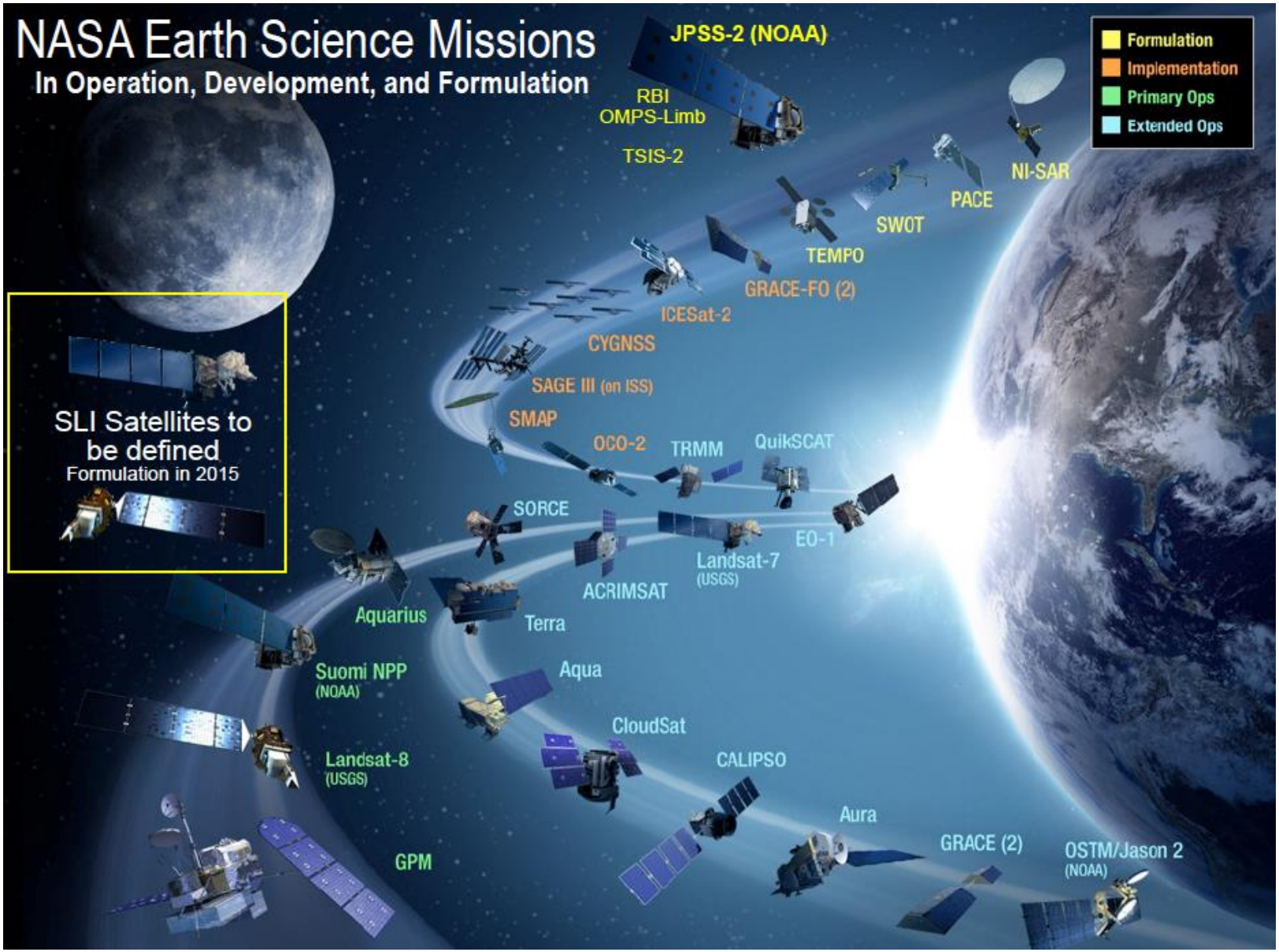
Archive data to ensure long term preservation, integrity, provenance, and proper use



Process data in various environments to create higher-level data products for science community

Distribute data to as many scientific communities as possible in as many formats and through as many mechanisms as possible

Provide customer support and outreach to the science community to support science teams and facilitate use of data products and associated technologies by current and emerging users


NASA Earth Science Missions
In Operation, Development, and Formulation
Formulation
Implementation
Primary Ops
Extended Ops
JPSS-2 (NOAA)
RBI
OMPS-Limb
TSIS-2
NI-SAR
PACE
SWOT
TEMPO
GRACE-FO (2)
ICESat-2
CYGNSS
SAGE III (on ISS)
SMAP
OCO-2
TRMM
QuikSCAT
SLI Satellites to be defined
Formulation in 2015
SORCE
EO-1
Landsat-7
(USGS)
ACRIMSAT
Aquarius
Terra
Suomi NPP
(NOAA)
Aqua
CloudSat
Landsat-8
(USGS)
CALIPSO
Aura
GRACE (2)
OSTM/Jason 2
(NOAA)
GPM

Featured Projects

Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO)

Clouds and the Earth's Radiant Energy System (CERES)

International Satellite Cloud Climatology Project (ISCCP)

Measurements Of Pollution In The Troposphere (MOPITT)

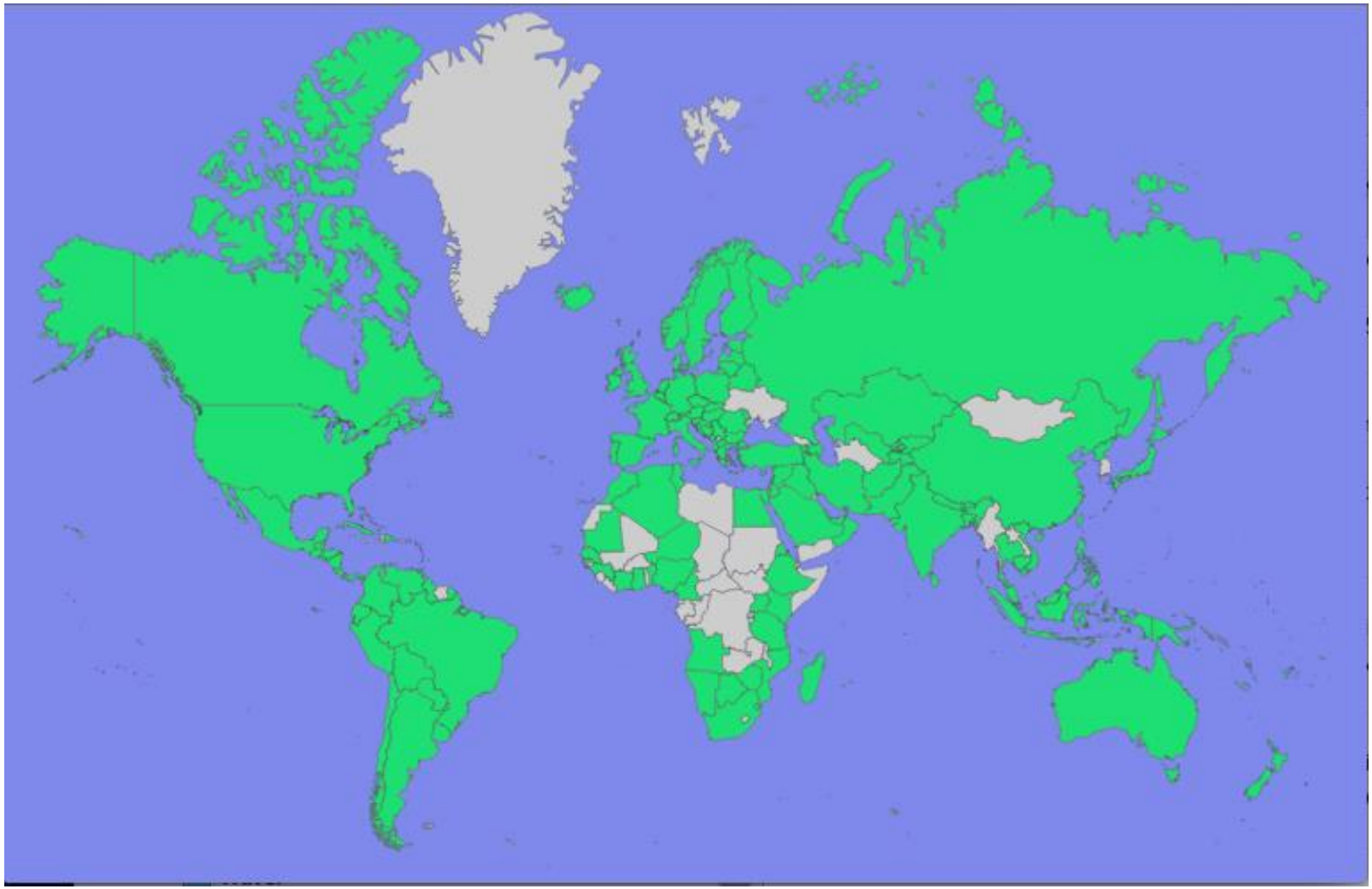
Multi-angle Imaging SpectroRadiometer (MISR)

GEWEX Surface Radiation Budget (GEWEX/SRB)

Tropospheric Emission Spectrometer (TES)

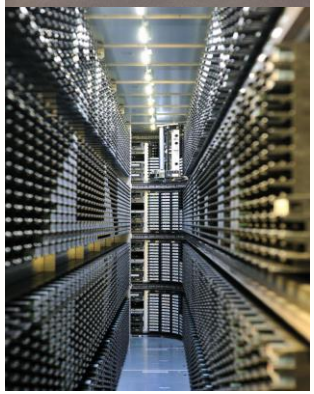
ASDC Data Subscribers

Over 165,000 subscribers in 158 Countries



ASDC Archive - Technology Evolution

1997
- 1 PB
- 9940B



2010
- 4 PB
- LTO-4

2013
- 1 PB
- 3 TB HDDs

2015
- 5.5 PB
- 4 TB HDDs

Use of Evolving Technologies

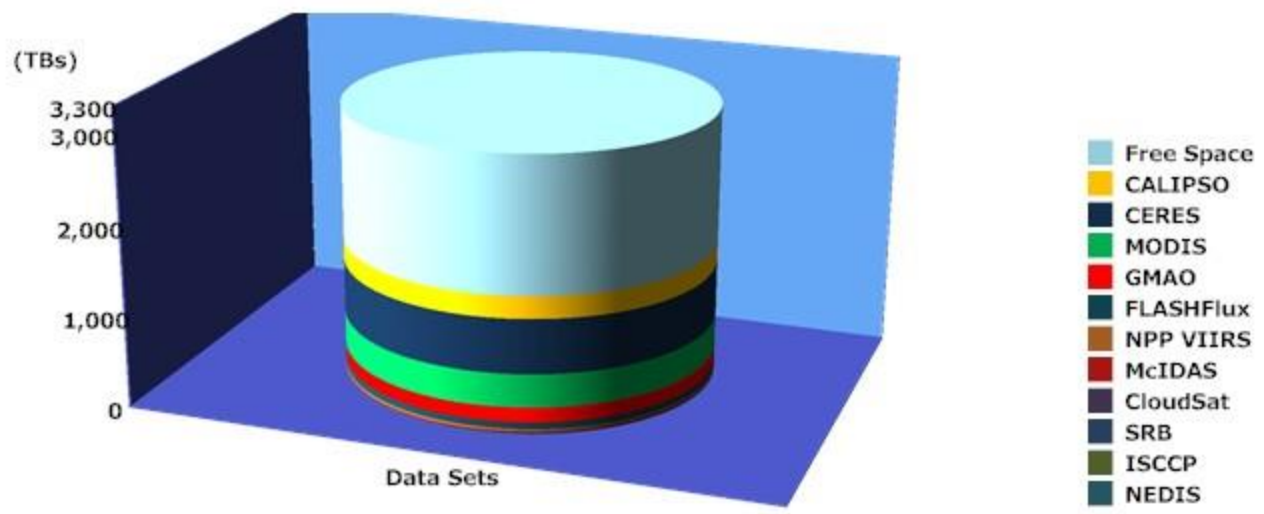
- As an active archive – must provide support production and ingest of new data products and distribution of data holdings to science communities
- ASDC strives to utilize mature technologies to meet data stewardship and increasing storage requirements
  - Current size: ~3.5 PB and over 100 million files
  - Growth Rate: 1.5 TB – 3 TB/day; 70K – 100K files/day
- Using IBM GPFS for multi-petabyte online disk archive
- Using Quantum StorNext as HSM with tape libraries for deep archive and disaster recovery copies

WHY WE NEED An Online ARCHIVE? (Opportunities)

- Can support long term climate studies requiring analysis of data sets collected over the past 20 years
- Can support the reprocessing of the entire mission of data sets to produce new editions of higher quality products with improved algorithms supplied by scientists
- Can more readily support distribution of hundreds of terabytes of data to other scientific research facilities
  - 25 Terabytes sent to University of Wisconsin-Madison over a 4 day period in October 2014 using 16 concurrent FTP streams
  - ~ 250 TB of ASDC data required at NCSA Blue Waters to support research; Using GridFTP; NASA and NCSA currently working to optimize data flow; WAN network upgrades underway; Additional data requirements could reach 1 PB

Data Products Online (DPO) disk cache provides unprecedented local access to many climate data sets

Important to have the Right Data at the Right Place at the Right Time



DPO (3.3 PB): Data Currently Stored = 1.7 PB
(~60 million files)

- BIG data access to ASDC data production services (no more staging data from tapes)
- Local scientists can run BIG data analytics across years of climate data records
- Validation of revised algorithms more comprehensive using BIGGER test data sets
- BIG data available to users desktop systems over NASA Langley campus network
- BIG data readily available for distribution to external customers

WHY WE STILL NEED TAPE?

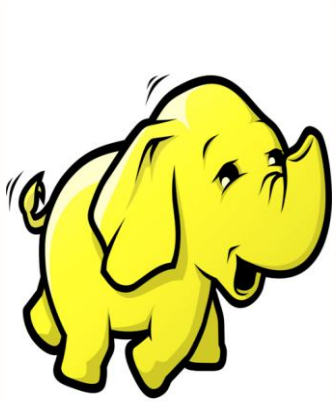
- ASDC must maintain a off-site disaster recovery copy of data holdings. Tapes are shipped to Iron Mountain.
- Use of tapes for backup copies is a more cost effective option than remote disk systems
- Our experience suggests that failure of disk systems will likely require recovery from backup copies at some point in the future
- Moving to LTO-6 later this year
- Investigating use cases for Linear Tape File System (LTFS)

CHALLENGES

- Maintaining data integrity throughout the data life cycle (data migrations to new disk and tape technologies)
- Keeping pace with increasing data archiving and distribution requirements. More data is coming from instruments on satellite missions scheduled through 2022
- Accomplishing large data recovery operations of hundreds of terabytes in weeks instead of months
- Working with scientists on relocating less likely to be used data sets to near-line or offline media (slower access)
- Working with scientists to purge obsolete data sets
- Supplying unpredictable large volumes of data sets to external customers

WHAT LIES AHEAD?

- ASDC now a member of iRODS (Integrated Rule-Oriented Data System) Consortium. Working with iRODS team on use cases applicable for more effective management of ASDC data holdings.
- Cloud Computing & Hadoop experiments conducted and private cloud pilot project underway
- Determination of the right balance between data sets held online versus near-line versus offline
- Assessments of emerging technologies
- Collaboration with partners (NCCS, NCSA, etc.)



SUMMARY

Opportunities

- Support BIG Data Analytics
- Support BIG Data Product Generation Campaigns
- Support BIG Data Access/Distribution

Challenges

- Managing BIG Data Growth at ASDC
- Distribution of BIG Data to external customers
- Deploying the best fit technologies to insure ready access and long term stewardship of ASDC data holdings

For more information or support:

ASDC Website: https://eosweb.larc.nasa.gov/

Email: support-asdc@earthdata.nasa.gov/

Phone: 757-864-8656

Located at Langley Research Center (LaRC)

Hampton, VA

Chris Harris, Systems Operations Lead

c.j.harris@nasa.gov

757-864-8590