National Aeronautics and Space Administration

EARTHDATA

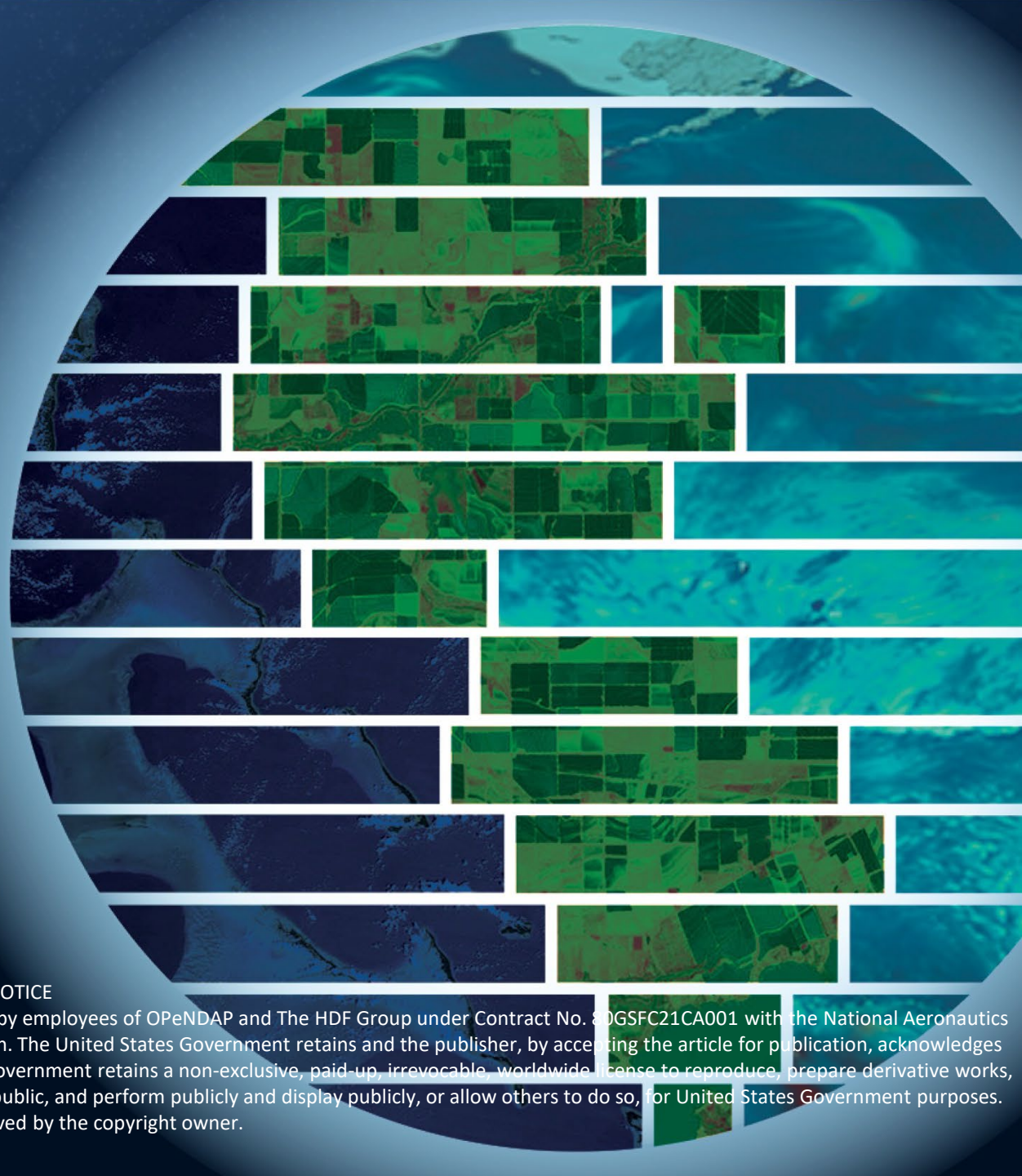
Support the Access of NASA HDF Data in the Cloud via OPeNDAP - Work Update

07/23/2025

Kent Yang[1], James Gallagher[2], Joe Lee[1] and Aleksandar Jelenak[1]

[1]NASA EED-3/HDF Group

[2]NASA EED-3/OPeNDAP

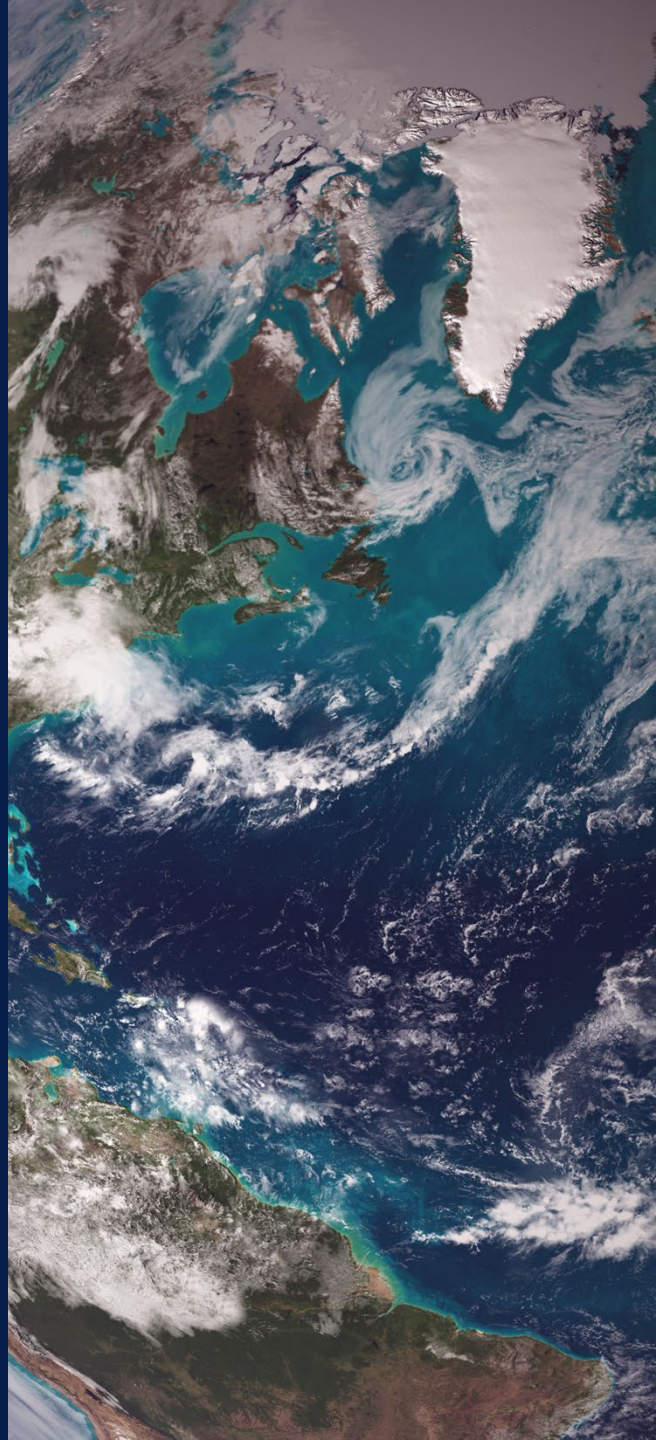


Agenda/Table of Contents

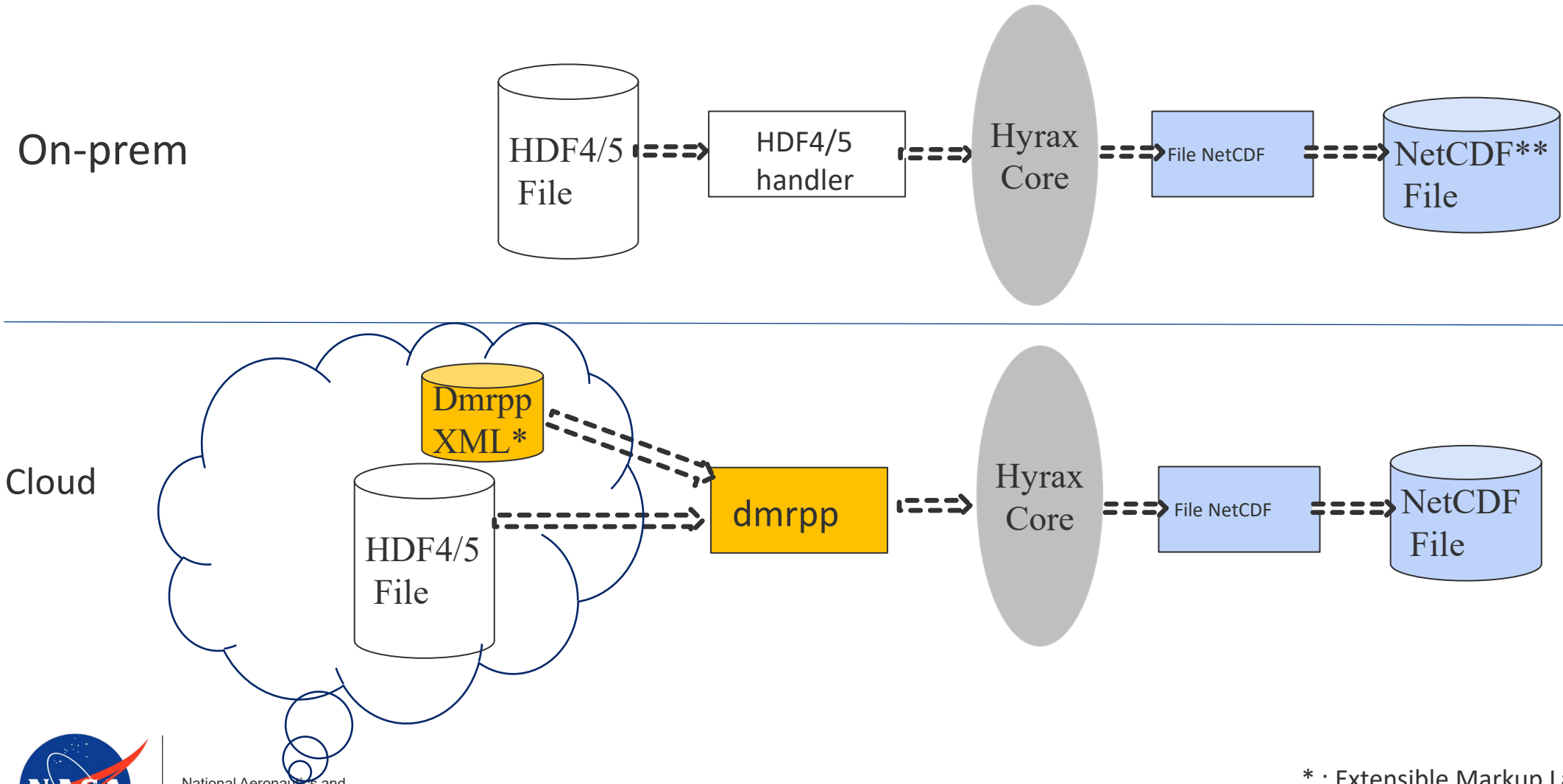
- Accessing HDF* Data in the Cloud via dmrpp**
- Support Updates
- Performance Updates and Results

*: Hierarchical Data Format

**: Dataset Metadata Response Plus Plus



HYRAX Web Service Chain

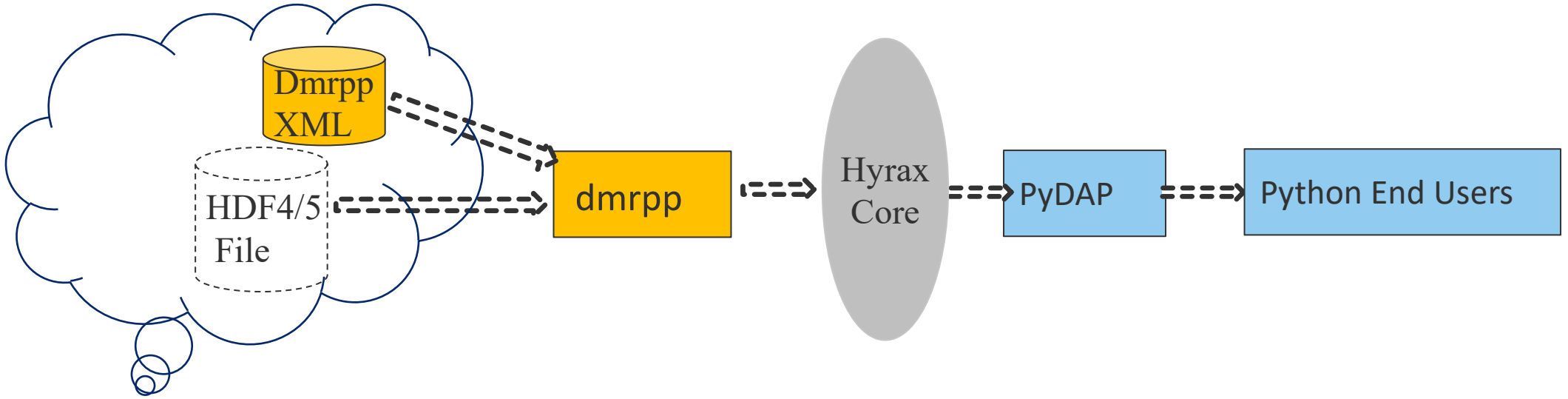


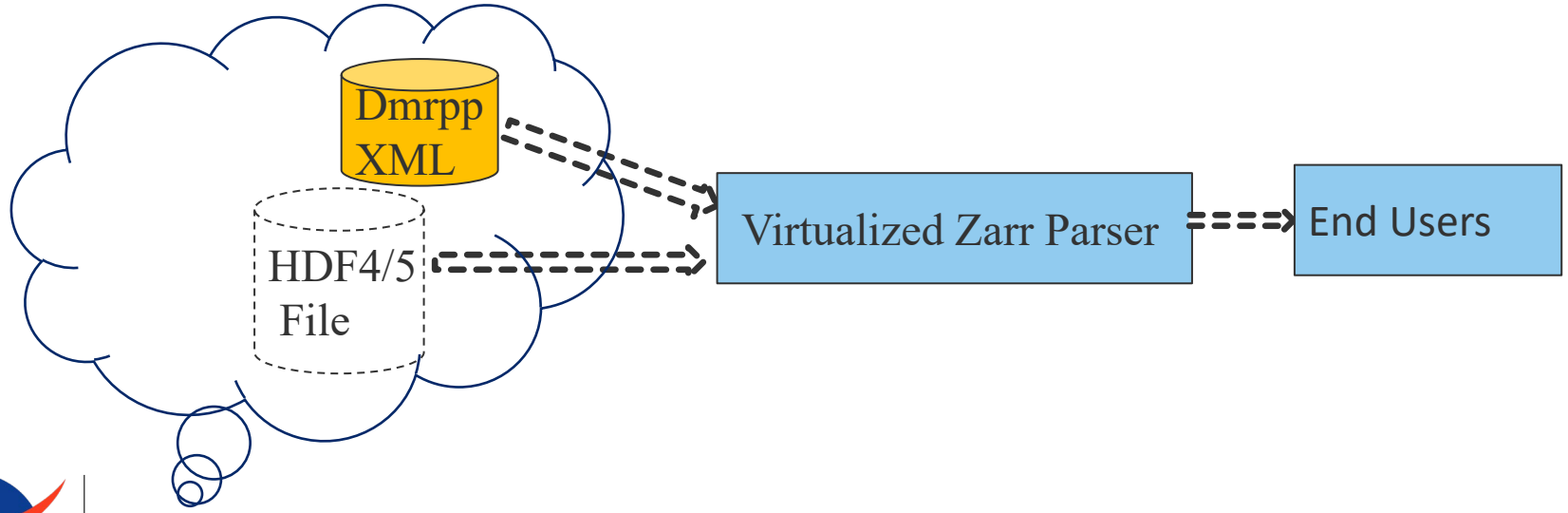
National Aeronautics and Space Administration

* : Extensible Markup Language

** : Network Common Data Form

Other Ways to Access HDF via dmrpp





Key Component to Access HDF Files in the Cloud

- dmrpp XML file
- To generate dmrpp files
  - The HDF4/5 libraries developed and maintained by the HDF Group are required.
  - Hyrax's HDF4/5 handlers developed by the HDF Group are required.
    - The handlers are components of the Hyrax software packages, the HDF Group is the main maintainer of the handlers.
- The HDF Group implemented the dmrpp generation program for HDF4 files.
  - Compression for the contiguous storage and linked blocks are supported.
- The HDF Group implemented a python wrapper to generate dmrpp files for both HDF4 and HDF5 files.
- The HDF Group is the major contributor that supports the special data handling in the dmrpp file(See the next slide).

National Aeronautics and
Space Administration

Special Data Stored in a dmrpp File

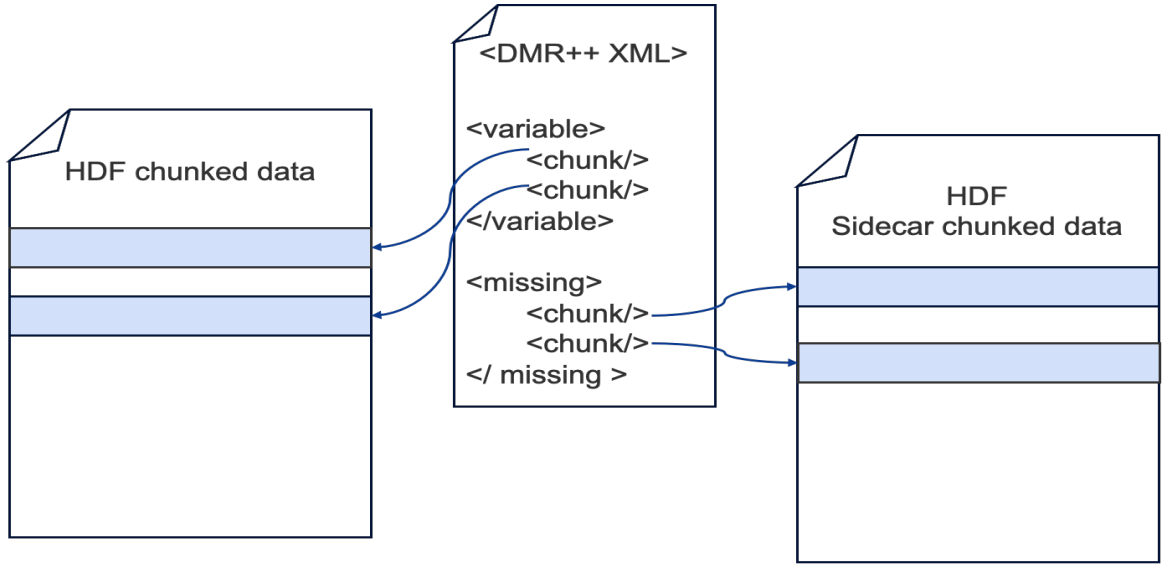
- The data's offset and length information cannot be retrieved through HDF API*s
- Examples:
  - HDF5 variables in compact storage
    - Data values are stored in the object header, no offset/length are available.
  - HDF5 variable-length variables
    - Data values are stored in the global heap, no offset/length are available.
  - HDF-EOS** geolocation information
    - Grid: Latitude and longitude may need to be calculated
    - Swath: Latitude and longitude may be stored in another file
- Question:
  - How to store them in a dmrpp file?

National Aeronautics and Space Administration

*: Application Programming Interface
**: Earth Observing System

Special Data Stored in a dmrpp File -Solution

- The smaller size or highly compressed data
  - Stored in the dmrpp file with base64 encoding and deflate compression
- The bigger size data such as HDF-EOS geolocation data
  - Data is stored in a separate sidecar file
  - The offset and length of the data in the sidecar file are stored in the dmrpp file with an additional link that points to the sidecar file



Support NASA HDF Products with Hyrax in the Cloud

- The HDF Group takes quick actions on the requests or bugs reported from NASA Distributed Active Archive Centers(DAACs)
  - Bugs reported by DAACs are usually fixed within a week
- The HDF Group also fixes bugs or add features proactively
  - By testing newer versions of NASA HDF sample files with Hyrax
  - By actively engaging in the discussions with the Hyrax team lead and other team members
- The HDF Group also addresses a few performance issues discovered by testing NASA HDF sample files with Hyrax
- Before working on the above tasks, we discussed with the Hyrax team lead and sometimes with other team members. All the above tasks are approved/guided by the Hyrax team lead.

National Aeronautics and
Space Administration

Newly Added Support By the HDF Group

- Complete the support of using dmrpp to access NASA HDF4/HDF-EOS2 files
- Support the variable length integer and float data in HDF5 handler and dmrpp
  - Discovered in the TROPOMI* data
- Support the enum datatype in HDF5 handler, dmrpp and Fileout netCDF
  - Discovered in NISAR** simulation data
- Support the HDF5 compound datatype data that contains variable-length and fixed size string in HDF5 handler, dmrpp and Fileout netCDF
  - Discovered in the TROPOMI and GEDI*** data

National Aeronautics and
Space Administration

*: TROPOspheric Monitoring Instrument
**: NASA-ISRO SAR Mission
***: Global Ecosystem Dynamics Investigation

Performance Improvement Work Update

- Significantly reduce the time to generate the dmrpp file for HDF4 files that have many chunks
  - Example: For an AIRS* level 1B file, the generation time is reduced from minutes to seconds.
- Add the buffer chunk feature in the dmrpp module to reduce the number of times to fetch the data from S3 to the Hyrax EC2 server
  - This is an enhancement to the current super chunk feature in the dmrpp module.
  - Suit for the case when the chunks in a variable are not adjacent to each other
    - Preliminary useful for accessing the HDF4 files
- Significantly reduce the time to write a large size string array in the fileout netCDF module
  - Example: For a SMAP** level 3 file, the time to write a large size string is reduced from hours to less than 1 minute.

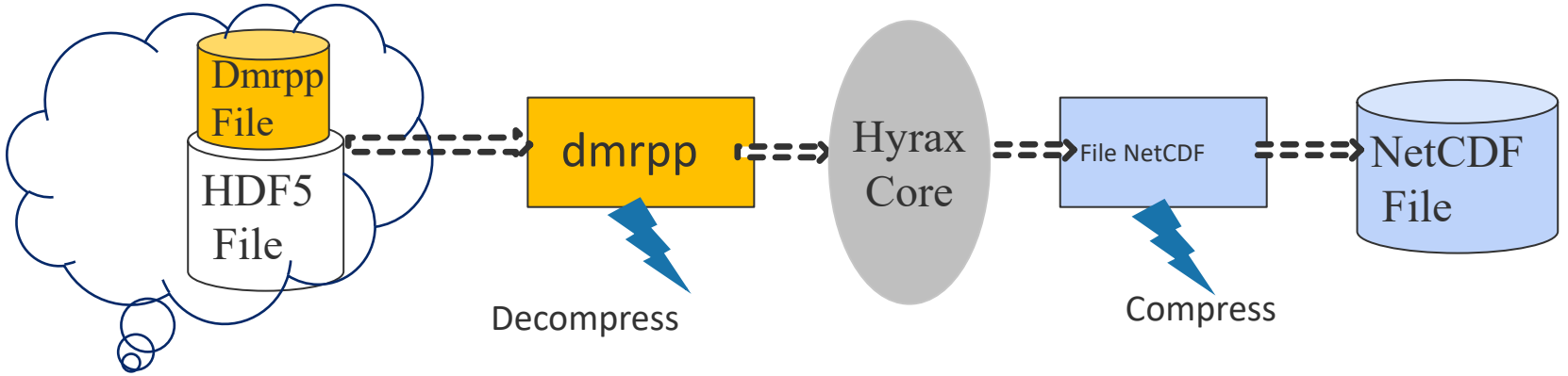
National Aeronautics and Space Administration

*: Atmospheric Infrared Sounder
**: Soil Moisture Active Passive

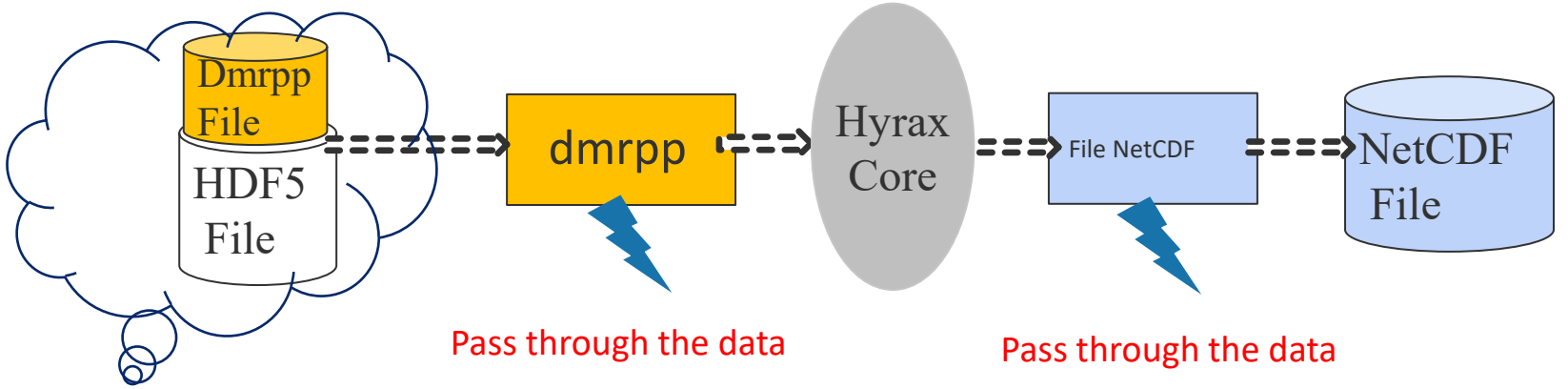
Performance Improvement Work Update 2

- Direct IO* review and evaluation
  - Concept review

General Approach



Approach with Direct IO



National Aeronautics and Space Administration

* Input Output

Direct IO Performance Results

- Evaluate more products that have Hyrax service in the earth data cloud
- Use the Hyrax testing server hosted by the Hyrax team
- Continue observing the ~10 times faster service response for many products

Product Sample File	File Size (GB)	Response Time with Direct IO (Seconds)	Response Time without Direct IO (Seconds)	Speed-up by using Direct IO
pre_SWOT*	1.2	11	123	11 X
MERRA2**	1.2	13	110	8 X
VIIRS***	0.2	4.5	49	10 X
SMAP	1.8	44	203	4.5 X

National Aeronautics and Space Administration

*: Surface Water and Ocean Topography
**: Modern-Era Retrospective analysis for Research and Applications
***: Visible Infrared Imaging Radiometer Suite

Direct IO Performance Results

- We also observe some products that benefit less from direct IO optimization.
  - It takes longer response time if a variable contains many small chunks.
    - Example: SMAP's 4X speed-up compared with 10X speed-up of equivalent size files
  - Currently Hyrax doesn't support the direct IO optimization if a variable contains chunks that are just filled with the fill values.
    - Example: This happens to one of GHRSST* testing file.
  - Currently Hyrax doesn't support the direct IO optimization if a variable is an array of string.
    - Example: One SWOT level 2 product contains string variables.
  - Direct IO only applies to the compressed variables.
    - Example: One big-size IMPACT** product doesn't have compressed variables.

National Aeronautics and Space Administration

*: Group for High Resolution Sea Surface Temperature
**: Interagency Implementation and Advanced Concepts Team

Facts for the Direct IO Feature

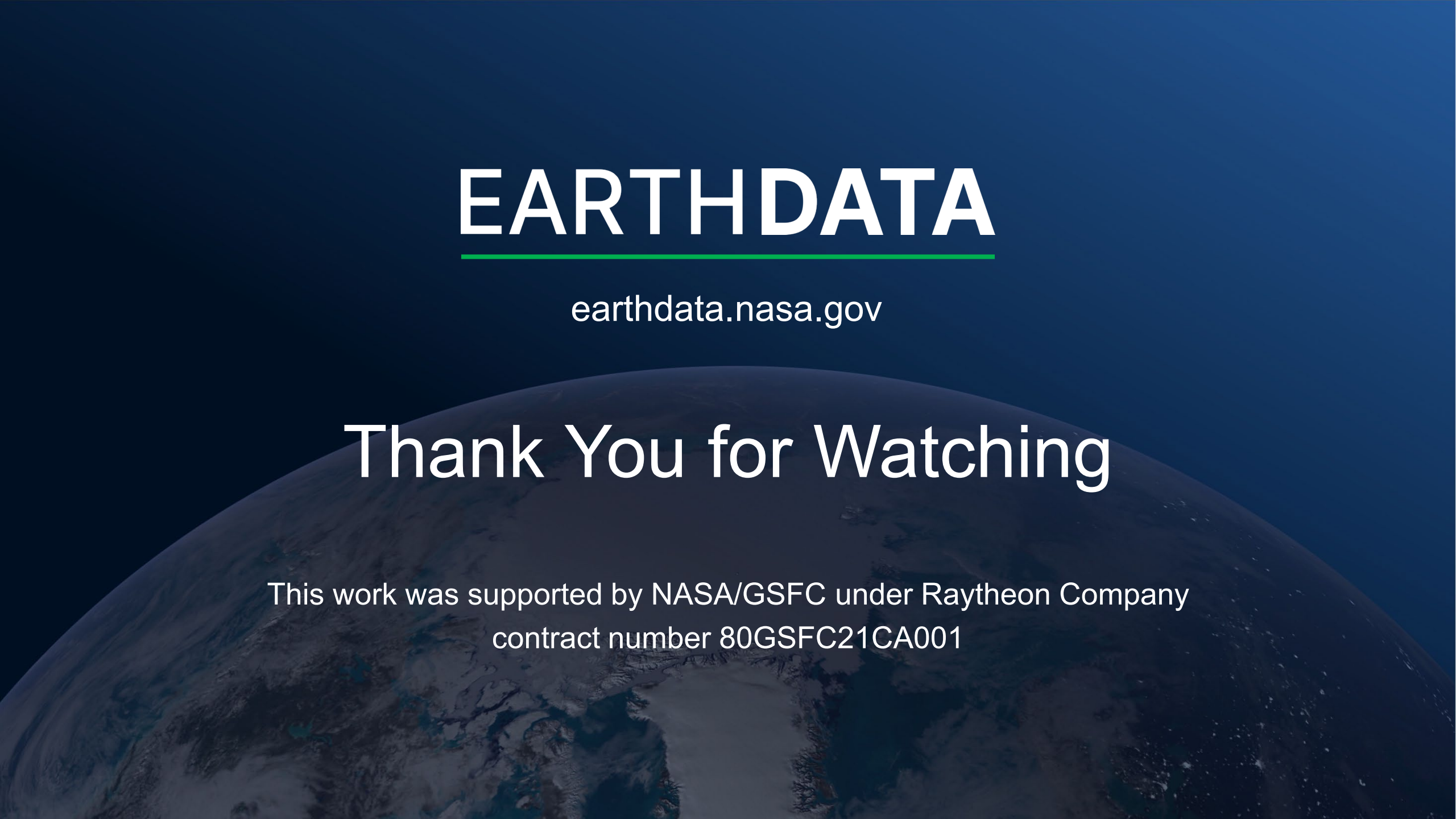
- Hyrax will use direct IO automatically for those cases when end users request to obtain the whole array of the selected variable(s) in integer or float datatypes and those variable(s) are compressed.
- Direct IO doesn't work for the variable that contain "no-data" chunks.
- This process is entirely transparent to the end users.
- Direct IO doesn't work for some old dmrpp files if they don't contain the key information needed for using the Direct IO feature. These dmrpp files need to be regenerated to take advantage of the Direct IO feature.

National Aeronautics and
Space Administration

Direct IO Future Work

- Will evaluate if we can support
  - Variables that have chunks that are filled with fillvalues
  - Good variable subset cases

National Aeronautics and
Space Administration



This work was supported by NASA/GSFC under Raytheon Company
contract number 80GSFC21CA001