Open Source Science for ESO Mission Processing Study

Identify a system architecture that meets the ESO mission processing objectives, supports open science, enables system efficiencies, and promotes earth-system science.

Workshop #1
October 19-20, 2021

NISAR Mission Data Processing System Perspective

Hook Hua¹
¹Jet Propulsion Laboratory, California Institute of Technology

Copyright 2021, by the California Institute of Technology. ALL RIGHTS RESERVED. United States Government Sponsorship acknowledged. Any commercial use must be negotiated with the Office of Technology Transfer at the California Institute of Technology.
Data Processing System Architecture

Science Data System (SDS) - JPL
- Receive / ingest data
  - Instrument and spacecraft data
  - Externally provided data
  - Standard Data Products
- Manage Data Processing
  - Level 0 S-Band Processing
  - Level 0/1 L-Band Processing
  - Level 2 Interferometric Processing
  - Level 2 Biomass Processing
  - L3 for soil moisture and Cal/Val
  - Bulk Reprocessing Campaigns
  - Urgent Response Processing
  - Instrument Calibration Tools
  - On-demand processing
    - Generate / Manage Data Products
      - Standard Data Products
      - Ancillary Data Sets
      - Auxiliary Data Sets
      - Internal & Temporary Data Sets
    - Deliver data products
      - ASF, ISRO
      - Bulk Reprocessing Campaigns
    - Urgent Response Processing
- Generate / Manage Data Products
  - Standard Data Products
  - Ancillary Data Sets
  - Auxiliary Data Sets
  - Internal & Temporary Data Sets
- Deliver data products
  - ASF, ISRO
  - Bulk Reprocessing Campaigns

Open science-like platform for algorithm development and analysis test bed environment for ADT, Cal/Val, Science Team
- Data accountability, Ins reports, system performance, data product latency...

Ground Stations
- Instr raw data
- Ancil (HK, etc), POE

JPL GDS
- Aux data

External Data Providers
- Ancil (HK, etc), POE

ISRO SDS
- S-SAR data
- L-SAR data

General Users
- All products
- L & S SAR data

ASF
- All products
- L & S SAR data

NISAR Proj Users
- (SDS, Instr, ADT, CalVal, UR users ...)
- Cal Params, etc

Ground Stations
- Instr raw data
- Ancil (HK, etc), POE

JPL GDS
- Aux data

External Data Providers
- Ancil (HK, etc), POE

ISRO SDS
- S-SAR data
- L-SAR data

General Users
- All products
- L & S SAR data

ASF
- All products
- L & S SAR data

NISAR Proj Users
- (SDS, Instr, ADT, CalVal, UR users ...)
- Cal Params, etc
Component View

Science Data System

- Algorithm Development Environment
- Processing Services
  - Analysis Processing
  - Data Product Generation
- Data Services
  - Data Store
  - Data Products (including ARDs in COG and Zarr)
- Algorithm Developers, Science Teams, Product Validation
- On-demand requests
- registers
- Data access
- uses & generates
- External Systems and Archives
  - ingest
  - delivery
Implementation Plan

Project Schedule (from pre-MCR to end of Science phase)

Road to MS-CDR
- ADT L5 requirements Peer Review
- PGE L5 Requirements Peer Review (internal Team)
- L0-L2 PGE Design Peer Review
- Cyber Security Peer Review
- Design Benchmark tests between NEN, GDS and SDS
- PCM subsystem LPR
- ADT Algorithm Peer Review
- GRFN review/status
- PPQA Process products Review
- GPU Peer Review
- AWS Cost Review using Resource Utilization Tool
- SDS Pre-CDR Peer Review
- MS CDR

Inherited software capabilities
- Open source ISCE used by ADT
- Open source HySDS used by SDS

This is an older schedule from 2019 that contained items of interest from before phase C/D.
Supporting Earth System Science

- Does your MDPS use/share any data, algorithms, etc from other ESO projects to support Earth System Science?
  - NISAR SDS artifacts (data, algorithms, IT security plans, processes, procedures, workforce and expertise) are shared across Integrated Earth Mission Systems (IEMS)
    - Shared with SWOT, SMAP, OPERA, MAAP
  - NISAR standard products will be used by SNWG/OPERA for displacement and water extent products to the applications communities

- What are your barriers to enabling collaboration to support Earth System Science within your ESO MDPS? (e.g., firewalls, access, schedule, developments costs, etc)
  - Current policies on Authentication and Authorization are barriers to opening up of systems for access and collaboration. e.g. access by public users (US and Foreign Nationals) working with Project
  - Current clearance process is not equipped to deal with modern complex web tools that are used to generate new datasets with a rapid turnaround time
  - Creating any new web tools or an online presence for NASA-funded systems is hampered by the current NASA moratorium on new website domains.

- What are the opportunities for improved support of Earth System Science?
  - Opening up NASA data and platforms to NISAR science users
  - Opening up NISAR project algorithms, data, workflows, etc to open science community to foster transparency and reproducibility by other systems and platforms
  - Systematic processing of standard products for future missions may be costly. Opportunities for more on-demand processing.
Supporting Open Science

SMD defines open science as a collaborative culture enabled by technology that empowers the open sharing of data, information, and knowledge within the scientific community and the wider public to accelerate scientific research and understanding.

- What does this definition of Open Science mean in the context of data processing systems?
  - Algorithm development is done in **collaborative open science platforms** per project scope
  - Data products during development, Cal/Val, product validation, and on-demand requests are open and accessible
  - Algorithms, workflows, ancillary/auxiliary data are open
  - Data products are **reproducible outside canonical SDS**

- What do you feel are the most beneficial opportunities for improvement in the MDPS to support Open Science?
  - Updates to Flight **project scope, requirements (and therefore funding)** to support open science
  - Updates to **cybersecurity policies and procedures** to improve open science procedures (e.g. accessibility)

- What are the barriers to supporting open science?
  - Current authentication and authorization policies required for **open access**
  - Development of algorithms in **open science context** requires updates to mission requirements

- What components (Data system, PGEs, algorithms, data) of your system will be developed in the open (open source from the outset)?
  - For NISAR SDS, the algorithms and core processing system are open sourced and **developed in the open**
  - NISAR SDS is **contributing** to the open source software
  - SNWG/OPERA will be **inherit**ing from this open source software as a starting point
Other

- What are your **pain points** in support of this mission?
  - Current NASA policies and cybersecurity restrictions are providing hurdles to open access to members of science teams.
  - To support science team, algorithm development team (ADT), and Cal/Val needs, setting up an accessible on-demand algorithm development and processing environment allowing users to develop, deploy, test, and generate on-demand data products.

- What does **system efficiency** mean in the context of an MDPS? (cost, data storage, processing time, etc.)
  - Processing of L0 to L2 data products are done in most **compute and storage** efficient manner.
    - Algorithm implementation optimally matches with deployed compute sizes. (e.g. single core, 36-core, GPU, etc)
    - Intermediate and standard products generated are **cloud-optimized**
  - Additional **economies of scale** can be achieved if there is a multi-mission processing capability
    - **Reuse** of processing, common pre-processing, more cost-efficient cloud costs models, sharing of ops workforce
  - Harmonized approach of **algorithm development-deployment-execution** use in standard product production, algorithm development, and analysis for Cal/Val and product validation.
    - Allows algorithm developers to harness the large-scale compute powers of MDPS

- Is there anything else you’d like to share that you feel would be helpful in our study?
  - **Cost constraints** and estimated peak throughput were drivers for early architecture design
    - Desirement for High Availability (HA) design vs requirements
  - Need for **common processing platform** for ESO-era missions.
    - Helps to address barriers mentioned above
    - Minimizes need to have MDPS formulation early on
  - Evaluation of **systematic processing vs on-demand processing** of standard products