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

High-End Computing Program Perspective
Tsengdar Lee

High-End Computing Program Scope

High-end computing program sustains the most advanced cyberinfrastructure which provides computational resource to NASA funded projects that include large scale computing tasks.

- Large data and computing infrastructure designed for most demanding computational workloads
- Specialized computing systems focusing on parallel computing tasks that cannot be done in commodity based computing infrastructure
- Mostly supporting modeling and simulation (MODSIM)
- Allowing data processing and analytics at peta- to Exa- scale
- Routinely refreshing data and computing infrastructure
- Funding research and development efforts to maintain the capability at leading edge

HEC’s implementation of Open Science and Lessons learned

- HEC already embraces Open Science.
- HEC invested in open source software development in previous CMAC solicitations.
- HEC developed a reference architecture when soliciting projects.
- HEC required software being developed under Apache 2.0 in CMAC calls.
- HEC encountered resistance from NASA centers and PIs
  - Policies at a NASA center had been a significant barrier
  - Algorithms, software and systems were not documented to facilitate sharing
  - Competition among NASA centers had prevented true sharing

Vision for Mission Data Processing

- HEC is currently not part of the Mission Data Processing System (MDPS).
- HEC supported data re-processing for a number of missions.
  - Open science is essential so that data processing algorithms may be revised/modified for data reprocessing purposes
  - Carefully curated long-term climate data records are critical for advancing Earth system science
  - Large computing infrastructure maximize the throughput for data reprocessing and greatly reduced the time to production
  - Open science will enable reuse of algorithms, workflow and shared knowledge
- Vision: HEC will become part of MDPS specifically for data reprocessing and modeling

Prioritized Guidance for the System Architecture Study

1. The System Architecture (SA) shall adhere to cybersecurity policies
2. The SA shall allow collaboration and support open science
3. The SA shall support SMD's procurements (processes and policies)
4. The SA shall be [hyper]scaleable
    - Shifting a strategy is a scale project,. Without scale, we have hobby.
    - Without hyperscaling, projects will die prematurally without reaching the critical mass for self sustaining.
5. The SA shall include both on premises and in-the-cloud capabilities
6. The algorithms, software, tools, and data shall be well documented and registered so that they are searchable.

An incomplete wish list:

- Reusable
- Reconfigurable
- Searchable
- Scalable
- Modularized (microservices), serverless, and server hardware agnostic
- Registration (App store) for services

Back up

Jupyter Notebook Workflow Analyzer

- Each Jupyter Notebook represents a scientific workflow
- Reverse engineer a notebook and extract its comprised function calls and dependencies among them
- From notebook code, automatically generate its flow chart showing its comprising external calls (e.g., APIs or services), package calls (library calls), and internal calls (user defined function calls)

- External Function
- Internal Function
- Package Function

Goal: Identify reusable functions to become potential APIs -> to store into App Store

CMDA Summer School Topic1 Goal

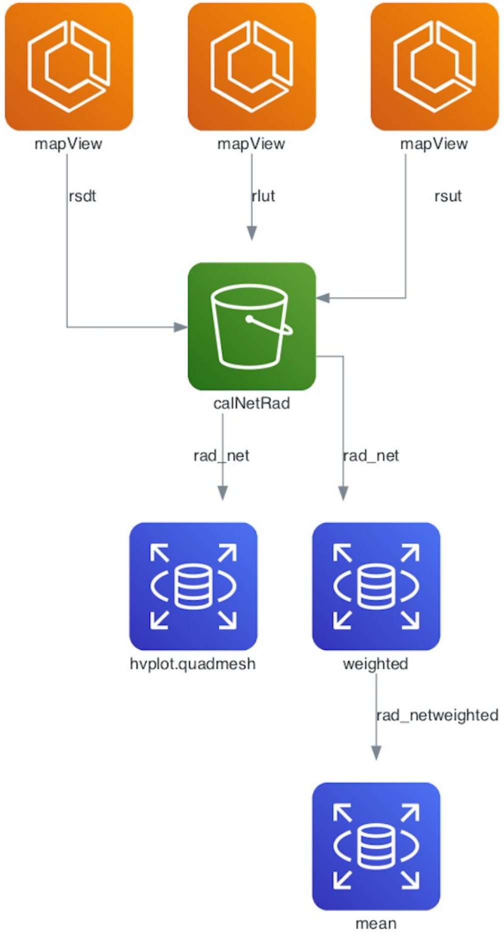
Topic 1: Where is the global warming?

Question 1: Calculate the global net radiative flux imbalance at TOA

Net radiative flux at TOA is calculated from:

$$\Delta F = F_{SW}^{\downarrow} - F_{SW}^{\uparrow} - F_{LW}^{\uparrow}$$

Where $F_{SW}^{\downarrow} - F_{SW}^{\uparrow}$ is the net incoming shortwave radiation and $F_{LW}^{\uparrow}$ is the outgoing longwave radiation at TOA respectively. First, let's load each of these from the CERES satellite instrument data:



CMDA Summer School Topic 3 goal

Topic 3: Variability of Clouds and Precipitation

Question 1: What are the spatial distributions of clouds, cloud radiative effects and precipitation?

The spatial distributions of clouds can be obtained using the MODIS total cloud fraction dataset.

Topic 3: It studies the spatial distributions of clouds, cloud radiative effects, and precipitation using NASA MODIS, CERES, and GPM datasets.

Generated Flow Chart