General Data Challenges (Environmental and Climate Modeling)

Dali Wang

Climate Change Science Institute

Environmental Science Division

U.S. DEPARTMENT OF ENERGY

Climate Change Science Institute AT OAK RIDGE NATIONAL LABORATORY

OAK RIDGE NATIONAL LABORATORY

MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Can you define your workflow from the data and I/O perspective? How do you classify your workflow data (I.e. experimental, observational, and simulation)?

- Simulation centric workflow (Community Earth System Models (CESM or ACME) and Parallel Reactive Flow and Transport Model, etc.)
- Analysis centric workflow (Diagnostic Analysis, Visualization, Backup and Archive, etc.)

What is the rate of data generated? Is it metadata or object I/O heavy? Is it checkpoint/restart heavy? Any other particulars you would like to share about the characteristics of your data?

- The rate of data generations in the term of simulation time (annually, monthly, daily, 6 hours, and even half hourly).
- Many domain specific languages, tools, and data processing utilities have been developed around the world.
- The checkpoint and restart is usually setup after every simulation year, depends on machines status and research requirements

What is the ratio of data to be placed on home, work, and archival storage areas during the life cycle of your allocation?

- We heavily use shared project space (such as those in cli017, cli106, etc.) share our simulation data.
- We normally generate 10 TB every quarter. Very small amount of code and data have been moved into home directory.
- Archival storage is constantly used during the simulation (with an easy turn-on option using CESM build-in scripts).

How often and what percent of data is accessed from the archival storage?

- Very often, small amount of simulation data (5-10 %) need to be accessed from the archival storage.
- I guess this number will become bigger during the ACME project, since we branched off from NCAR.

How often large amounts of archival storage data need to be moved across the WAN during the lifetime of your project? How will you share your data with other researchers in your domain?

- Sharing data across WAN is very limited between DOE labs/facilities.
- However, we shared large amount of data between OIC (ESD) cluster and Titan. We shared the data via the project-shared space on both Titan and OIC cluster.

What data tools/libraries you need to improve the efficiency of your workflow management, data sharing and access, and long-term data preservation? Which tools/libraries do you use currently at OLCF for data manipulating and handling?

- Traditionally, we use NETCDF/HDF5, or Parallel NETCDF etc. for data management and processing.
- We are developing new workflow management system through ACME project.
- Interactive and efficient tools for data analysis and manipulation and handling (such as ParCAT, etc.) are much needed.