Automating the Workflow of Light Source Scientific Applications

Eun-Sung Jung and Rajkumar Kettimuthu
Argonne National Laboratory
Email: {esjung,kettimut}@mcs.anl.gov

***Abstract*—The light source scientific applications deals with large-scale data, which is projected to increase rapidly in next a few years. The current workflow of such applications suffers from non-automated procedures, which eventually result in a prolonged time-to-solution. We propose that the adoption of multiclouds will help automate the workflow of such applications.**

I. Light source facility and related scientific applications

At Argonne National Laboratory, we have the Advanced Photon Source (APS) facility, one of most advanced light source facilities in the nation. There are multiple X-ray beamlines at the facility and each beamline, which has different capability for diverse scientific applications, can capture the images of subjects.

In general, the workflows of such applications follow the similar steps as in Figure 2. At first, data is generated by X-ray sensors and moved to a local HPC cluster through a LAN for basic image processing such as 3D image reconstruction. The data then is moved to the users institutions for post analysis for their own purposes. If the experimental results are different from what is sought at the experimental design time, repetitive experimental would be conducted with different parameters of experiments.

As of now, 3.3 TB of data is generated per day. The data is distributed by USB or hard drive. In future, the data is projected to increase around up to 250 TB/day.

II. Future workflow of light source applications

The workflow of light source scientific applications needs to be automated through transferring data over WANs directly to users institution. This helps users identify the experimental errors in the early stage, which is the major delay in the overall time-to-solution, and they can adjust experimental parameters on the fly to get the accurate results. Recently, a near-real-time analysis of a few terabytes of APS data at a remote compute cluster at Pacific Northwest National Laboratory was done.

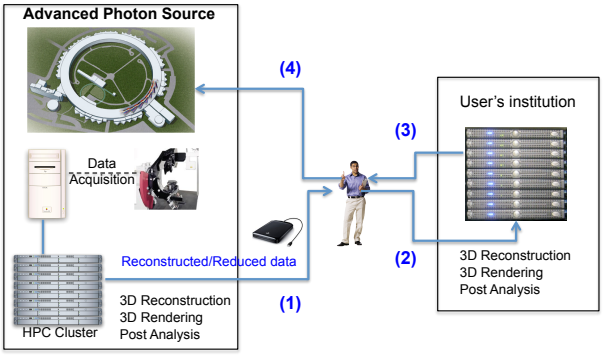

Fig. 1. Automated workflow of Advanced Photon Source (APS) scientific applications

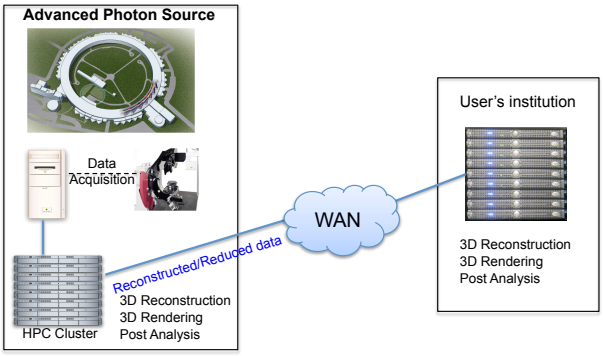

Fig. 2. Automated workflow of Advanced Photon Source (APS) scientific applications

We can further detail the requirements from both of the users' and facility's perspective. First, users want to do the following:

- **Centralized control over their workflows**: Currently users do not know whether the experiments at the APS are going well or not until the experiments get done. Even after data is given to users, they spend time on physical disk ingest into their institution. These situations get worse if multiple experiments are going on and additional experiments are required due to unsatisfactory results of preceding experiments.
- **Cost-effective use of their available resources**: Users may have allocated resources at multiple sites, but have difficulties in utilizing all the computing resources be-

cause each resource has different computing environment, which deters users from using all the resources.
- **Reduced time-to-solution**: Eventually, users want to get the reduce time-to-solution, most part of which depends on experiments.

Second, administrators of a facility want to do the following:

- **Up-to-date local HPC resources**: The APS should keep its local HPC resources up-to-date as the data grows.
- **Provision of user tailored post analysis**: The APS perform basic computations on the generated data before handing it over to users because each user has its own post analysis program, which is one of obstacles toward the full automation of the workflow.
- **Fast distribution of generated data and improved utilization of facilities**: Human intervention is required for distribution of experimental data and unnecessary experiments due to inability of confirming experimental data in near real-time lower the utilization of the APS facility.

III. Challenges and opportunities

To satisfy the requirements, we need to address challenges as follows.

- APS does not have enough IT management organization.
- Upgrading their local HPC resources as the data grows is costly.
- Users suffer delays in processing their data due to human intervention.
- Users' programs for post analysis cannot run on any other platforms.

We believe that the multi-cloud platform is the viable solution to address the challenges. The possible approaches can be as follows.

- **Outsourcing local HPC resources to clouds**: Considering insufficient budget for local HPC resources, cloud is a good resource to oursource their function.
- **Utilizing multiclouds for users' post analysis**: If we can provide a platform for multiclouds where users can run their post-analysis program (e.g., Docker), users can get benefits such as more reduced time-to-solution and no lock-in to one computing resource.
- **Providing a central dashboard**: The ability to check the progress of a workflow and experimental results quickly will help users make smart decisions on subsequent experiments.