Implementing Petascale Workflow development into XSEDE Science Gateways

Yuan Liu
Joint Institute for Computational Sciences
Univ. of Tennessee - Oak Ridge National Laboratory
Bldg 5100, Rm 231, ORNL
Oak Ridge, TN 37831
yliu89@vols.utk.edu

Bhanu Rekepalli
Joint Institute for Computational Sciences
Univ. of Tennessee - Oak Ridge National Laboratory
Bldg 5100, Rm 231, ORNL
Oak Ridge, TN 37831
bhanu@vols.utk.edu

ABSTRACT

Next generation knowledge discovery in life sciences research urgently needs compute-intensive and memory-intensive bioinformatics analysis to keep up with exponential data growth. We developed parallelized versions of some of the most widely used life sciences applications such as BLAST, MUSCLE, HMMER, and DOCK6 in order to efficiently utilize large-scale XSEDE (Extreme Science and Engineering Discovery Environment) HPC (High-Performance Computing) resources. We have designed large scale genomics data analysis platform with web-enabled portal known as the PoPLAR (Portal for Petascale Lifescience Applications and Research) science gateway to enable the researcher to effectively and systematically analyze large scale data using the afore mentioned applications on HPC resources without the necessity to be a technical expert. This has directly impacted the science achieved by collaborators who previously could not run these tools on supercomputers. However, many research tasks require the execution of several loosely-coupled processes. Researchers/users currently need to submit data and job requests for each individual process, manually obtain intermediate outputs, and resubmit it for sequential procedures for downstream analysis. The manual intervention is not only time consuming for the user, but also reduces the overall throughput of the computational jobs. Therefore, we developed several high-performance intermediate bioinformatics processing tools to analyze tera-bytes of output data and connect parallel applications in PoPLAR in order to enable automated workflow features for users, which will be discussed in this poster. We will demonstrate the development of large-scale scientific workflows using the parallel tools on PoPLAR science gateway with successful real-world workflows as an example, for large-scale discovery of novel protein domain models used by biologists at the University of Tennessee, Knoxville, for protein classification and annotation by researchers at the Seattle Children's Research Institute, and for metagenomics and metatranscriptomics analysis by researchers at the University of South Carolina who currently use XSEDE resources. We will display the process and efforts needed to integrate these workflows into the PoPLAR science gateway. Finally, we will also discuss the challenges faced in integration of Globus tools for large scale data movement between users and science gateway.

Keywords

Scientific workflow, science gateway, bioinformatics, genomics, HPC