ECSS Workflows – An Overview and Update

Mats Rynge, Marlon Pierce and Suresh Marru
Extreme Science and Engineering Discovery Environment

XSEDE
Extreme Science and Engineering Discovery Environment

The XSEDE Workflow Community Applications Team's charter is to assist researchers to use scientific workflow technologies on XSEDE to solve challenging scientific problems involving parameter sweeps, multiple applications combined in dependency chains, tightly coupled applications, and similar execution patterns that require multiple applications and multiple XSEDE resources. The workflow team accomplishes its mission through the use of third party workflow software in collaboration with the workflow developers, service providers and XSEDE Extended Collaborative Support Services.

Introduction

Do you have a scientific workflow? Do you use XSEDE in any of the following ways?

- Large scale parameter sweep problems.
- Computations that need to run on more than one XSEDE site.
- Computations that combine multiple scientific applications in novel ways.
- Computations with demanding pre- and post-processing steps that require XSEDE all by themselves.
- Computations requiring non-trivial data movement between XSEDE resources.

The Workflow Community Applications Team is available to help. We provide expertise and assistance with scientific workflow software tools that make challenging workflow problems more efficiently executed, easier to manage, and easier to reproduce.

A keystone in our approach is to use existing tools. Even though the team is made up of developers of several workflow systems, the team is workflow agnostic. The goal is not to develop workflow systems, but to help users use the existing ones. Users are guided when it comes to selecting the appropriate workflow system for their problem, and then the team helps with the integration with the user's existing codes, and XSEDE resources.

Outreach and Community

In order to make the workflows effort sustainable, it is not enough to help individual users, but what is needed is a to build a community. This is similar to other successful XSEDE efforts such as Gateways and Campus Champions. To build the community, general discussion and information mechanisms, such as mailing lists and XSEDE wiki space, have been set up.

At the SC'13 conference, a workflows BOF was hosted to highlight some of the existing workflows on XSEDE, and to bring people in the XSEDE community and the workflow community together.

This summer, we will host a seminar series introducing different workflow systems. The talks given so far are:

- *Swift: implicitly parallel scripting for XSEDE researchers*
- *Building Scalable Applications using Makeflow and Work Queue*
- *MoSGrid - Workflows for Molecular Simulations in a Science Gateway*

Community Mailing List: workflows@xsede.org This is a community mailing list that is open to anyone interested in scientific workflows. To subscribe, email majordomo@xsede.org with "subscribe workflows" in the body of the message.

Example Engagement: USATLAS and LIGO

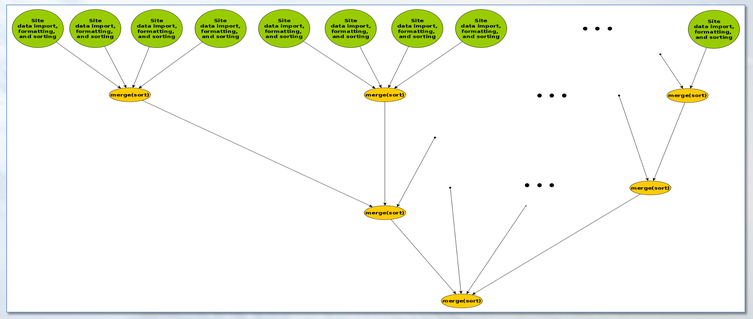
The ECSS Workflows team can provide support porting existing workflows to the XSEDE infrastructure. Current examples of this are USATLAS and LIGO, both large projects with their own infrastructures. ECSS Workflows consultants are helping evaluate what modifications will have to be made to the workflows and possibly the target resources in order to make these workflows run efficiently on XSEDE.

Example Engagement: BioKepler

BioKepler is a module distributed on top of the core Kepler scientific workflow system to design and execute workflows using a set of bioinformatics tools using distributed execution patterns. This ECSS workflows effort is fairly new, and the goal is to develop NGS microbiome data analysis workflows that can run on XSEDE resources computers, and make the workflows available to researchers that have access to XSEDE resources.

Example Engagement: RSQSIM

RSQSIM is an earthquake and fault systems dynamics code which was initially implemented in R, but later moved to C. The code was set up to generate data files fore each grid point, with the result being that a run could potentially generated millions of files. In parallel to improving the output format, the ECSS workflows team was asked to provide a solution to merge the files into a single output files. The solution include, for each grid point, running a tool for post processing which generates a data file sorted by timestamp. As the goal for the final data file is to be sorted by timestamp, all the files from the first level is just merged in order. The files are merged in small groups to minimize the memory usage, and levels are added to the workflow until only the final data file remains.



Example Engagement: SoyKB and iPlant Collaborative

SoyKB, in collaboration with iPlant and the ECSS workflows team, implemented a workflow to process soybean genome sequence data. The workflow is based on community tools such as GATK, BWA, and Picard. Interesting characteristics of the workflow include a close integration with the iPlant DataStore, which is used to store inputs and outputs of the workflow, and mapping of fine-grained tasks with varying core/memory requirements to TACC Stampede using MPI based cluster jobs. SoyKB has a detailed poster at this XSEDE'14 poster session.