SC'13 BOF Report - Science and Scientific Workflows: Putting Workflows to Work

The purpose of this session was for computational scientists, scientific workflow software researchers, cyberinfrastructure architects, and cyberinfrastructure operations leaders to share ideas on how to best use cyberinfrastructure such as the NSFs XSEDE to tackle the most challenging and scientifically valuable distributed computing problems. The emphasis was on discussing open, unsolved scientific problems that will benefit from using multiple XSEDE, Open Science Grid, campus, and international cyberinfrastructure resources in coordinated fashion.

The format of the BOF consisted of a series of short "lightning" talks on workflow topics, followed by a discussion session dubbed "Birds on the Hot Seat", or BOTHS. The BOF organizers invited all presenters, with an open call for participation issued to several forums. During the session, a Google Docs document was editable by anyone in the audience who wanted to contribute to the notes. Notes are available from https://docs.google.com/document/d/1o4sTWO1HrrVYrhj2c5uRUvjIu8cE55Ds1HxC62KRImw/edit. Links to all talk abstracts, presentations, and notes are available from the BOF website, https://sites.google.com/site/sc13workflowbof/home.

Lightning Talks

Lightning talk agenda, with links to the abstracts and slides:

- Marlon Pierce, Mats Rynge, "BOF Intro and Getting Help from the XSEDE Workflow Community Applications Team"
- Scott Callaghan, Southern California Earthquake Center, "Experiences Running Seismic Hazard Workflows"
- Justin Wozniak, Argonne National Laboratory, "Swift: Solutions to Workflow Challenges on Grids and Supercomputers"
- Sandra Gesing, University of Notre Dame, "Experiences with Computational Workflows at Notre Dame's Center for Research Computing"
- Yifeng Cui and Heming Xu, San Diego Supercomputer Center, "Efficient Aftershock Sequence Generation Using Workflow-Based Algorithms"
- Sangmi Pallickara, Colorado State University, "Supporting Scalable Server-side Data Preprocessing in Distributed Data Storage Systems"
- David E. Bernholdt, Wael R. Elwasif, and Samantha S. Foley, Oak Ridge National Laboratory, "The IPS Framework: A Flexible Approach to Loosely-Coupled Multiphysics Simulations"

The second half of the BOF was Birds on the Hot Seat (BOTHS): Probing questions on the state of scientific workflow software, its integration with cyberinfrastructure, and its use to solve challenging scientific computing problems.

Questions for the Lightning Talks

Scott Callaghan, Southern California Earthquake Center, "Experiences Running Seismic Hazard Workflows"

- Do we have this problems on the custom workflows or in the Pegasus workflows? Answer: custom workflows had more issues. More streamlined with Pegasus.
- Why aren't sub-workflows the solution to the "group workflows"? Issues include large scope, for example just the number of files. The SCEC workflows are already hierarchical workflows.
- How much workflow reuse do you see? The challenges are not only from the workflow tool, but the whole hardware/software stack. Using workflows have been an overall positive experience.

Justin Wozniak, Argonne National Laboratory, "Swift: Solutions to Workflow Challenges on Grids and Supercomputers"

- Moving to a new machine leads to new problems. How do you handle that in Swift? We are doing a lot of testing. With Swift/T we are trying to make it easier for user to test themselves by copy and pasting into templates.
- What language is Swift written in? Originally it was Java, but Swift/T is C.
- Do you have to install? Yes, the runtime must be compiled for the local MPI library.

Sandra Gesing, University of Notre Dame, "Experiences with Computational Workflows at Notre Dame's Center for Research Computing"

- Don't make me think! The number one challenge is usability.
- You work a lot with EU projects. How would you characterize the state of the Cyberinfrastructure in US and Europe? US seems to be more connected. Europe is a little bit more disconnected with many different smaller grids. US likes to reinvent the wheel. New user interfaces are created, but existing workflow engines are used.

Yifeng Cui and Heming Xu, San Diego Supercomputer Center, "Efficient Aftershock Sequence Generation Using Workflow-Based Algorithms" (presented by Mats Rynge)

- Can you describe how you got involved with the project and how the support works? As users can request for XSEDE Allocation, they can request for support by consultants. As ECSS consultant we help setup the workflow and help the users in getting started on XSEDE using workflows.

David E. Bernholdt, Wael R. Elwasif, and Samantha S. Foley, Oak Ridge National Laboratory, "The IPS Framework: A Flexible Approach to Loosely-Coupled Multiphysics

Simulations"

- What is the availabilty? It is on SourceForge.
- Curious about the Python wrappers - are they parameterized? It is just a driver which interacts with the other parts of the system (such as the plasma state system).

Sangmi Pallickara, Colorado State University, "Supporting Scalable Server-side Data Preprocessing in Distributed Data Storage Systems"

- Is Galileo a software layer? It is a meta file system layered on top of commodity hardware.

BOTHS Questions

What is a scientific workflow?

- Are the other types of workflows, other than scientific? Business workflows. BPEL comes to mind, the primary difference being that business workflows have manual (human) steps. This might be something we might want to bring into scientific workflows, for example error handling and progress.
- Not all science is computation but has a workflow in the process of scientists work.
- Instrument samples, experiment execution, connectivity between scientific tasks can benefit. We should not restrict ourselves to large-scale computations. There are tasks that are related with some decision making in the middle of processes. Any tools in this space?

If workflows == reproducibility, why aren't we all using them all the time?

- Do users know they need reproducibility? Scalability is more common selling point.
- Barrier of entry.
- Going down the wrong path, hacks on top of hacks, makes it hard to switch to a workflow system later.
- Still hard to share workflows? Trust, but also technical.
- Workflows are not the only way to achieve reproducibility. Scripts for example.
- When is a script a workflow? Maybe the difference is superficial and abstract. You stick to what you know.
- There is difference. It is not only the tasks, but fault tolerance, error recovery, optimization.
- But can those things be implemented as packages for scripts?

Are there examples of successful workflow technologies not for/from scientists that we can learn from?

- See the business workflow comments above
- Rendering pipelines. Visualization pipelines are technologies we could learn from.
- How would we find out what for example industry is doing? How would we do that

research? Industry wants to purchase code/support. Industry have higher requirements, for example GUIs to define the workflows.

- In industry, workflows might use more seldom in the process, which means it has to be super easy because the users would not remember how to do it otherwise.