Workflows At NERSC

New User Training
June 16, 2020

Bill Arndt
Data Science Engagement Group
warndt@lbl.gov

Agenda

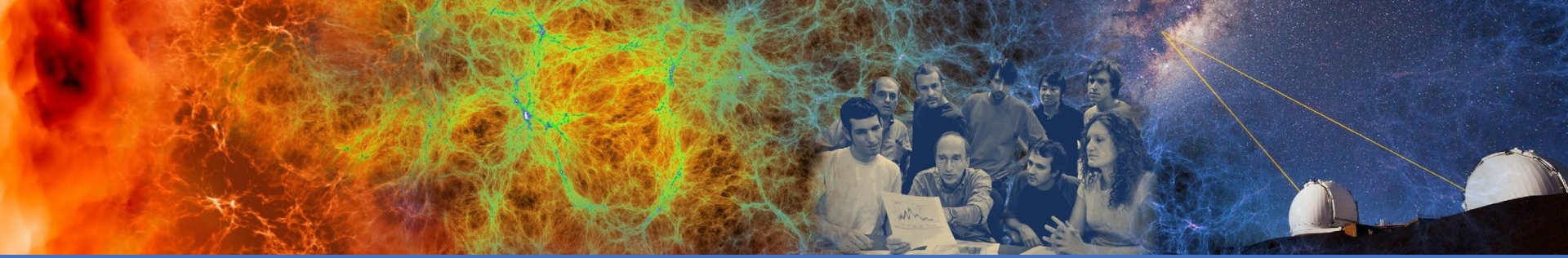
- What are Workflows?
- Workflow Resources at NERSC
- Best Practices

What Are Workflows?

- A workflow is a problem best solved by inserting automation between user action and the interfaces* used to get computation and data resources**.
  - *Interfaces like: Slurm commands, shell on a login node, HPSS, Globus, Data Warp
  - **Resources like: Cori compute nodes, storage, network bandwidth and data transfer, identity management
- *Workflow Management Tools* (WMT) are the software systems that perform that automation.

Workflow General Examples

- “I need to run my application thousands of times.”
- “My data needs several stages of processing with different applications running in an ordered sequence.”
- Application has a 2% chance of crashing and needing rerun
- Rerun this application every month



Workflow Resources at NERSC

What NERSC Is Doing To Support Workflows

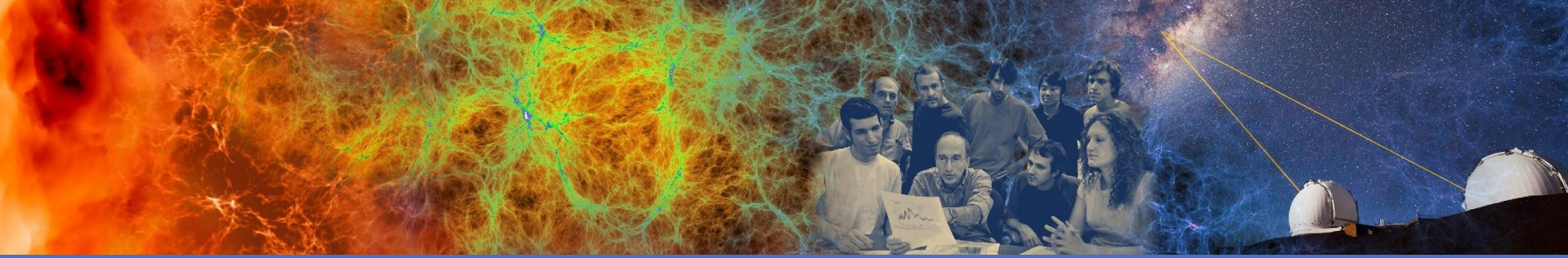
- Specialized infrastructure, software, and support
- Workflows Working Group
  - Formed September 2019 - Laurie Stephey (DAS), Bjoern Enders (DSEG), Bill Arndt (DSEG)
  - Thorough evaluation of many WMT ongoing
  - Documentation and guidance refresh
  - Outreach to users, facilities, tool developers, and infrastructure providers

WMT Documentation and Guidance

- https://docs.nersc.gov/jobs/workflow-tools/
  - A work in progress; expanding and refining as our tool evaluation continues
  - Detailed information, examples, pitfalls, and suggestions regarding specific tools and use cases
- We *want* to get tickets about workflow management tools
  - Builds our experience and knowledge of what users need
  - Shares our experience

Cori Workflow Nodes

- Cori has two service nodes specifically reserved for WMTs
  - Same environment as login nodes
  - Access is limited to approved users
  - Heavy compute not allowed
  - The preferred place for crontabs
  - Uptime same as Cori login nodes, prepare accordingly
- Gain access by submitting a request to NERSC support
  - Be prepared to describe your WMT and its resource footprint
  - Provide a list of users who need access to set up and maintain the WMT



Best Practices

GNU Parallel Is Better Than Shared QOS

```
elvis@cori07:~> seq 1 5 | parallel -j 2 'echo \
> "Hello world {}!"; sleep 10; date'
Hello world 1!
Thu Jun 11 00:21:00 PDT 2020
Hello world 2!
Thu Jun 11 00:21:00 PDT 2020
Hello world 3!
Thu Jun 11 00:21:10 PDT 2020
Hello world 4!
Thu Jun 11 00:21:10 PDT 2020
Hello world 5!
Thu Jun 11 00:21:20 PDT 2020
elvis@cori07:~>
```

- Packed jobs have massively reduced total queue wait
  - Can also pack single-node tasks into multiple node jobs
- No risk of Slurm overload
- Run combinations of tasks in parallel and sequence
- Easy input substitution
  - If you need it, *much* more power is available
- Superior to task arrays, too
- See documentation

Burst Buffer and Data Intensive Computing

- The Burst Buffer has excellent I/O operations capacity
  - Necessary to scale an I/O intensive Data/HTC workload to hundreds of compute nodes or beyond
  - Up to hundreds of metadata server on Burst Buffer vs. two for Cori scratch
- See Cori Burst Buffer documentation

Data Centric Workflow Management Tools

- “I have many different applications and data types chained together in a network of dependencies.”
- *Plenty* of options. Snakemake and Parsl are two examples, among *many*
  - See documentation
- Pitfalls:
  - Many expect cloud availability and can’t understand queue waiting
  - Often lack job packing
  - Naive Slurm integration can use too many requests
  - Risks with networked filesystems

Thank You and Welcome to NERSC!