

Increasing Portability of Scientific Workflows with Linking

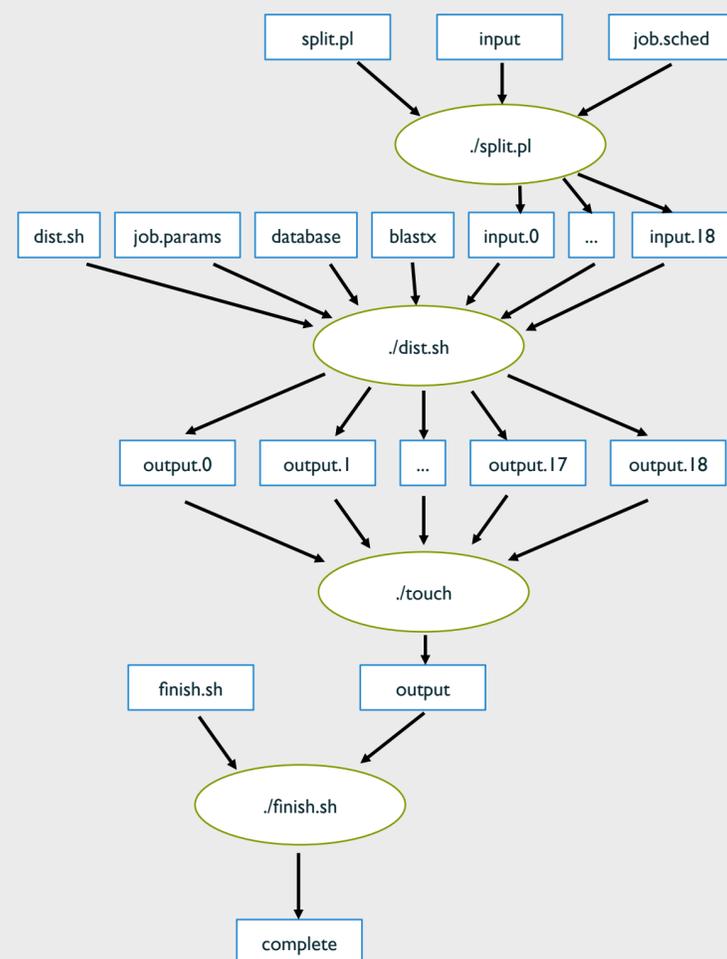
Casey Robinson and Douglas Thain
University of Notre Dame



Abstract

Dependency management remains a major challenge for all forms of software. A program implemented in a given environment typically has many implicit dependencies on programs, libraries, and other objects present within that environment, so that is virtually guaranteed to break when moved to another environment. Workflows particularly suffer from dependency management problems, precisely because they tie together multiple independent programs into a coherent whole. To address this problem, we propose applying the old idea of a linker into the new context of workflow systems. We have implemented a linker for the Makeflow workflow system, and extended the concept to apply recursively to executables and scripted languages within the workflow. We evaluate the system by applying it to a selection of bioinformatics workflows including BLAST^[3], BWA^[4], and SHRiMP^[5], enabling them to be moved across multiple computation environments.

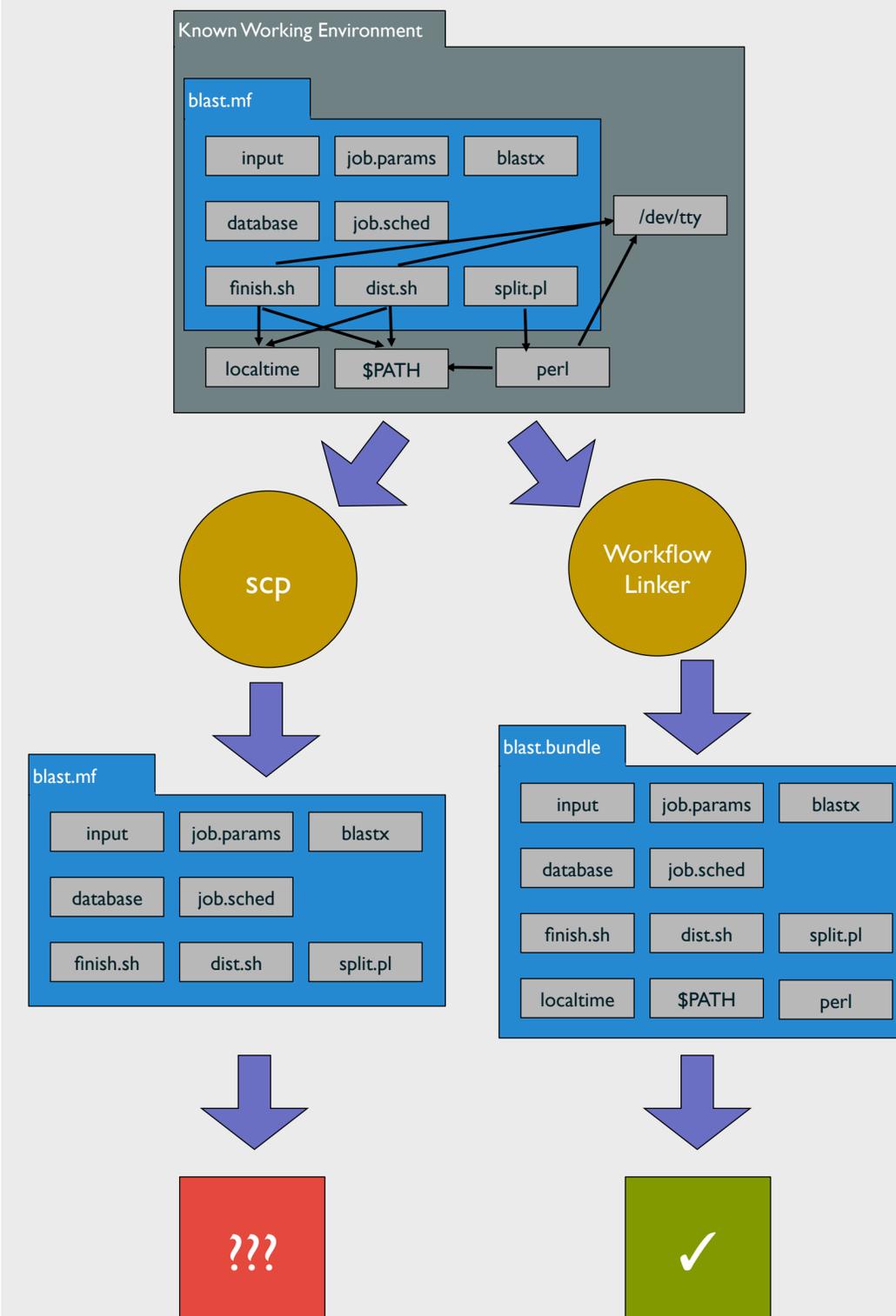
BLAST^[3] Workflow



Additional Examples

BWA, Biocompute^[6], Matlab, SHRiMP, SSAHA

Solution



Challenges

- ▶ Which files should be collected?
- ▶ Minimizing data transfer

- ▶ Requirements for scripting languages

Where to get it?

<http://www.nd.edu/~ccl>

The Cooperative Computing Lab

Participate in the Tutorial at CCGrid 2013 on May 13 2013 at Delft, Netherlands!

About the CCL: We design software that enables our collaborators to easily harness large scale distributed systems such as clusters, clouds, and grids. We perform fundamental computer science research in that enables new discoveries through computing in fields such as physics, chemistry, bioinformatics, biometrics, and data mining.

Community Highlight: Scientists searching for the Higgs boson have profited from Parro's new support for the CernVM Filesystem (CVMFs), a network filesystem tailored to providing world-wide access to software installations. By using Parrot, CVMFS, and additional components integrated by the AnyData, Anytime, Anywhere project, physicists working in the Compact Muon Solenoid experiment have been able to create a uniform computing environment across the Open Science Grid. Instead of maintaining large software installations at each participating institution, Parrot is used to provide access to a single highly-available CVMFS installation of the software from which files are downloaded as needed and aggressively cached for efficiency. A pilot project at the University of Wisconsin has demonstrated the feasibility of this approach by exporting excess compute jobs to run in the Open Science Grid, opportunistically harnessing 370,000 CPU-hours across 15 sites with seamless access to 400 gigabytes of software in the Wisconsin CVMFS repository. -Dan Bradley, University of Wisconsin and the Open Science Grid

Recent News:

- Tutorial on Building Scalable Scientific Applications at XSEDE13
- Tutorial on Makeflow and Work Queue at CCGrid 2013
- Elastic Apps Paper at CCGrid 2013
- Genome Assembly Paper in IEEE TPDS
- CCTools 3.7.0 Released!
- CCTools 3.6.2 Released!
- CCTools 3.6.1 Released!

Research:

- Papers
- Projects
- People
- Jobs
- HEU

Software:

- Download
- Manuals
- Makeflow
- Work Queue
- Parrot
- Chimp
- SAND
- AWE

Community:

- Highlights
- Annual Meeting
- Training
- Getting Help
- Mailing List
- For Developers

Operations:

- Visual Status
- Condor Pool
- History Cluster
- Biocompute
- BXGrid
- Condor Log Analyzer
- Internal

or
<https://github.com/cooperative-computing-lab/cctools>

References

1. Guo, P., Engler, D. CDE: Using System Call Interposition to Automatically Create Portable Software Packages USENIX ATC '11
 2. Zhao, J., et. al. Why Workflows Break - Understanding and Combating Decay in Taverna Workflows E-Science 2012 IEEE 8th International Conference
 3. Basic Local Alignment Search Tool <http://blast.ncbi.nlm.nih.gov/>
 4. Burrows-Wheeler Aligner <http://bio-bwa.sourceforge.net/>
 5. Short Read Mapping Package <http://compbio.cs.toronto.edu/shrimp/>
- Biocompute <http://biocompute.cse.nd.edu>

Acknowledgement

This work was supported in part by the Department of Energy and the National Science Foundation via grants OCI-1148330 and CBET-0941565