Increasing Portability of Scientific Workflows with Linking

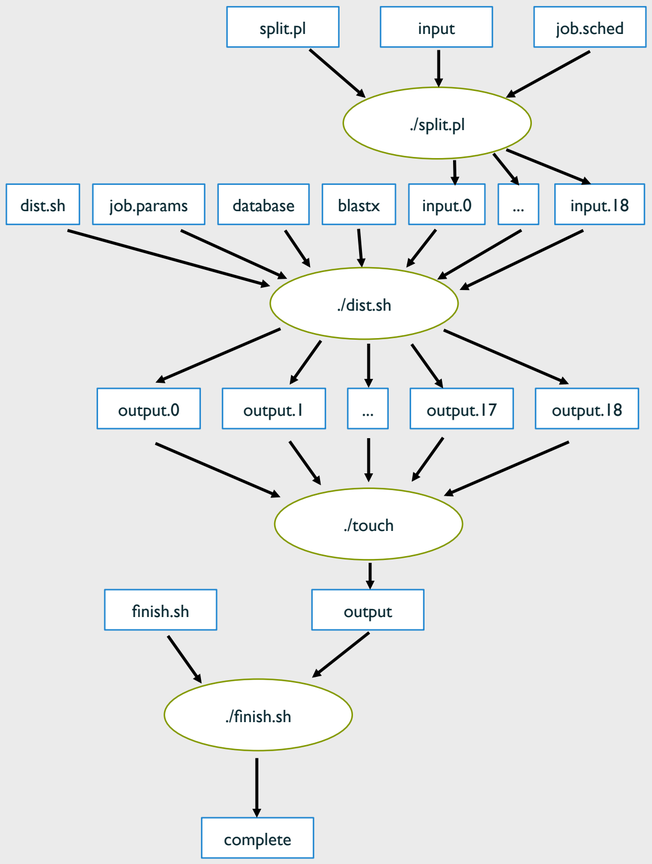
Casey Robinson and Douglas Thain
University of Notre Dame

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.

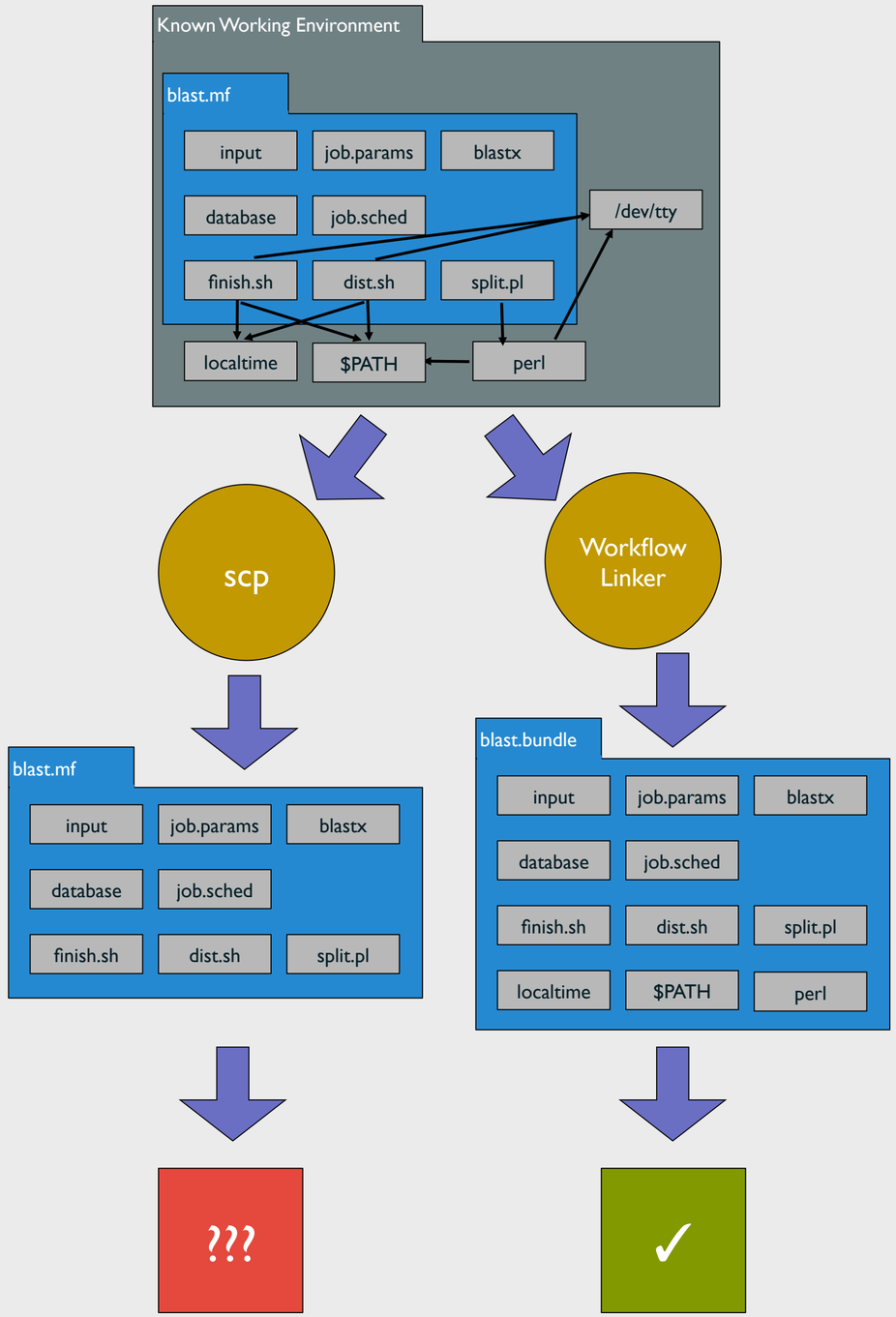
BLAST[3] Workflow



Additional Examples

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

Solution

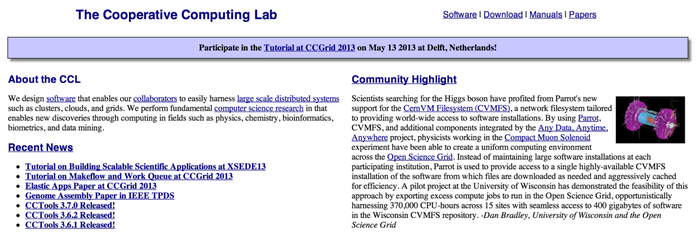


Challenges

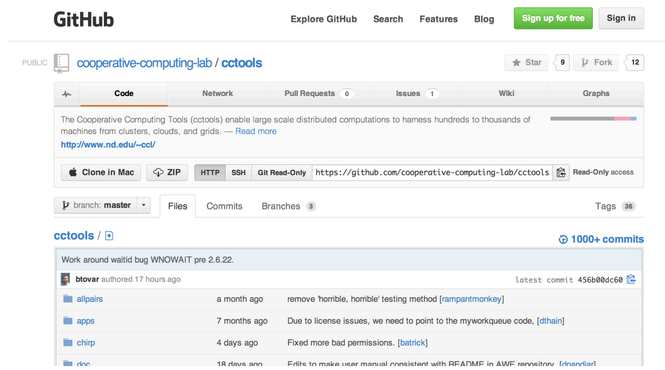
- Which files should be collected?
- Minimizing data transfer
- Requirements for scripting languages

Where to get it?

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



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