Umbrella: A Portable Environment Creator for Reproducible Computing on Clusters, Clouds, and Grids
Haiyan Meng and Douglas Thain
Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA
{hmeng|dthain}@nd.edu

1 Abstract
Environment configuration is a significant challenge in large scale computing. An application that runs correctly on one carefully-prepared machine may fail completely on another machine, creating wasted effort and serious concerns about long-term reproducibility. Virtual machines and system containers provide a partial solution to this problem, in that they allow for the accurate reconstruction of an entire computing environment. However, when used directly, they have the dual problems of significant overhead and a lack of portability. To avoid this problem, we present Umbrella, a tool for specifying and materializing comprehensive execution environments from the hardware all the way up to software and data. A user simply invokes Umbrella with the desired task, and Umbrella determines the minimum mechanism necessary to run the task - direct execution, a system container, a local virtual machine, or submission to a cloud or grid environment. We present the overall design of Umbrella and demonstrate its use to precisely execute a high energy physics application across many platforms using combinations of chroot, Docker, Parrot, Condor, and Amazon EC2.

2 Motivation

3 Architecture of Umbrella

4 Specification
Figure 2 shows the specification for a CMS physics application. Umbrella allows a user to specify a dependency in two ways: unique identifier (one referent) and attribute description (a class of referents). The only except is the environ section, which has a fixed syntax: <env name>:<env value>.

4.1 Hardware
*hardware*: ['platform': "x86_64", 'cpu cores': '1', 'memory': '1 GB', 'disk': '4 GB'],

4.2 Kernel
*kernel*: ['type': 'linux', 'release': '>=2.6.32'],

4.3 Software
*software*: ['name': 'Red Hat', 'version': '6.5', 'id': '62165b27c8fcd0b97907a']

4.4 Data
*data*: ['name': 'LHE_v12-th_samples-2381', 'resources': ['cvmfs://cvmfs.fnal.gov/cern.ch/']

5 Evaluation of Matching Degree
Umbrella deploys the minimum virtualization technology necessary to achieve the desired environment. (S1) If the host machine is fully compatible, the task is run directly. (S2) If the OS is compatible but some additional software or data are needed, Parrot is used to deliver the files. (S3) If only the kernel is compatible, Docker is used to deliver the operating system. (S4) If the kernel is not compatible, a virtual machine is created.

6 Local Cache
To minimize the execution environment construction time, each software dependency should be pre-built and configured. To improve the portability of archived software, software dependencies should conform to common-used internal organizations.

7 Evaluation

8 Conclusion
Cooperative Computing Lab: ccl.cse.nd.edu
Data and Software Preservation for Open Science: www.daspos.org