Workflow Virtualization for Data Intensive Computation (WVDIC)

Sreekanth Pothanis

Scientific Workflows and Data Grids

- Scientific workflows
  - Manage complex scientific applications
  - Integrate compute and data sources
  - Generate large amounts of data
    - Cactus simulations

- Data grids
  - Provide long term storage
  - Enable collaboration and sharing
  - Provide context for recovery

Integration with Data Grids

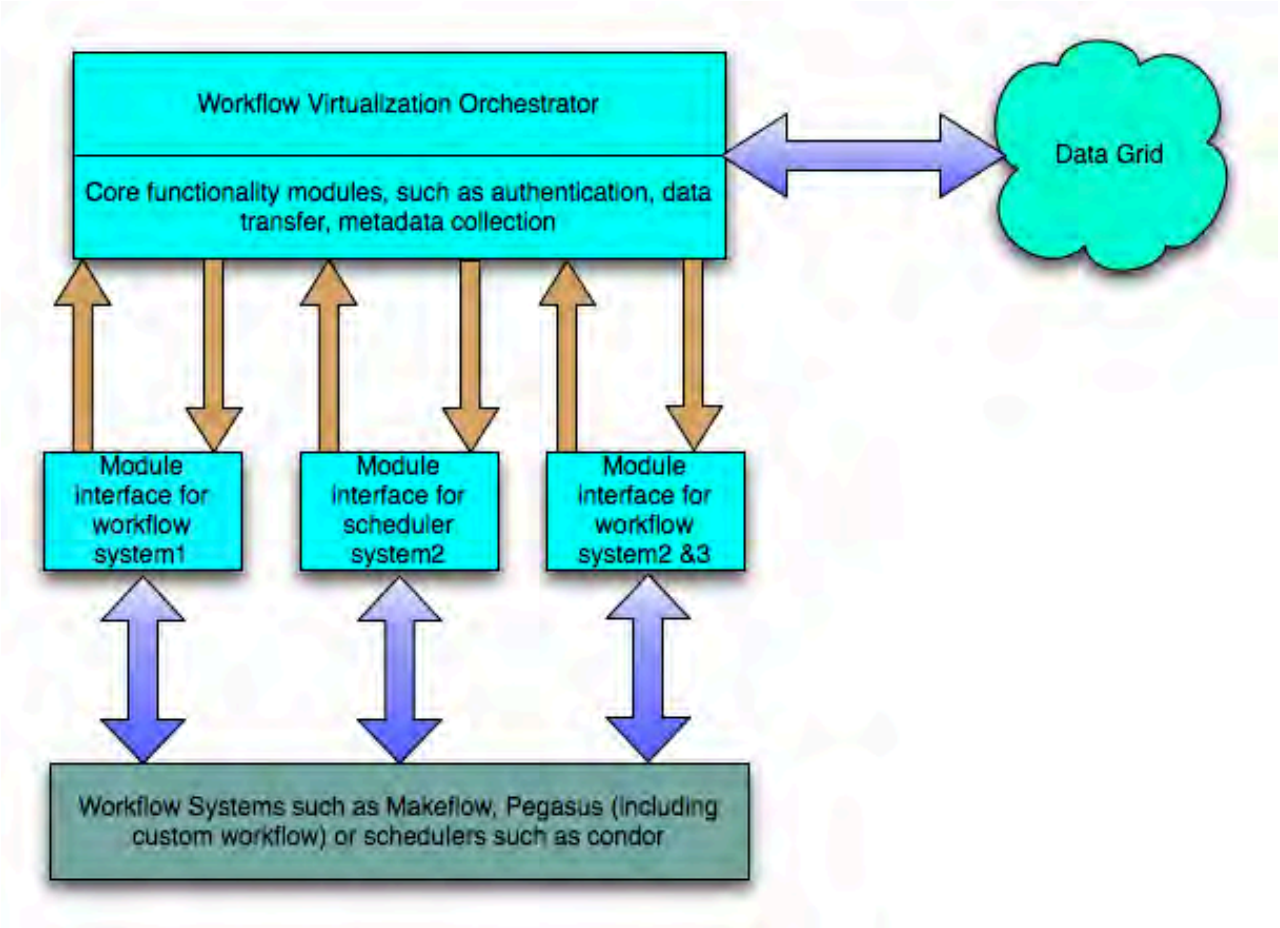
- Automates execution of workflows
- Allows staging and post processing
- Enables automation of archival of produced data sets
- Simplifies environment set-up

Workflow Virtualization

- Management of the properties
- Manage interactions with each workflow system for input and output of files.
- Provides higher control
- Enables execution of complex workflows spanning multiple different workflow systems
- External to the environment that actually runs the workflow
  - Increases generality

Workflow Virtualization Server (WVS)

- Stand alone and modular
- External to any workflow



WVS: Authentication and Context Handling

- Handled at two levels
  - Grid level to perform grid transactions
  - OS level to execute workflows
- Data grid context
  - Provides information about data grid
    - User privileges, quotas
- Workflow context
  - Generated during the execution
    - List of output files, destination, metadata

WVS: Staging, Execution and post Processing

- Sets up the working environment before initiating the interfacing module
- Decreases execution time by pipelining where possible
- Executed by invoking appropriate modules
  - Modularity allows high level of customization
  - Provides higher control
- Handles custom post processing scenarios

Integration with iRODS

- Implemented through micro-services and rules
  - Client interface
- Client design and configuration
  - Configuration file and rules

```
WORKFLOW=MAKEFLOW
CONFIG=/tempZone/home/wfuser/test.makeflow
INPUT=/tempZone/home/wfuser/capitol.jpg
INPUT=/tempZone/home/wfuser/local.jpg
INPUT=/tempZone/home/wfuser/meta.jpg
DEST=/tempZone/home/wfuser/test_dest/
METADATA=NAME1=VAL1
METADATA=NAME2=VAL2
```

Integration with iRODS

- Server Configuration
  - Authentication
  - Data Transfer
  - Metadata
  - Module execution
- Interacts with iRODS server as an admin

```
[MAKEFLOW] path=/usr/local/cctools/redhat5/
bin/makeflow
args= -T condor
[MAKEFLOW]
[MAKEFLOW1]
path=/usr/local/Makeflow/bin/makeflow
args= -p 9876
[MAKEFLOW1]
#[KEPLER]
#path=path to kepler
#args=-t –P
#[KEPLER]
[PEGASUS] path=/usr/local/Pegasus/Pegasus-
plan
path_to_sites.xml = /usr/local/Pegasus/sites.xml
path_to_rc.data /usr/local/Pegasus/rc.data
path_to_tc.data = /usr/local/Pegasus/tc.data
[PEGASUS]
```

Conclusion – WVDIC

- Automates execution of workflow
- Orchestrates at sub-workflow levels across multiple workflow systems
- Provides a generic solution
  - Implemented with iRODS, Makeflow, Pegasus

Thank you