HTCondor, Docker, iRODS and HPC for every scientist with the CyVerse Discovery Environment

Ian McEwen – Software Engineer
University of Arizona
mian@cyverse.org
Discovery Environment

Main, high-level goals:
- Make storing and analyzing data easy (help scientists focus on science, not IT)
- Use prior work, serve as glue between others' excellent work:

iRODS  
HTCondor  
docker  
The AGAVE Platform  
GWT  
PostgreSQL  
RabbitMQ™  
Grouper™  
elastic

(and more)
**iRODS**

Integrated Rule-oriented Database System

- Data storage & organization for large amounts of data
- Known and used by scientists already
- Where mostly-all the data comes from & goes to in our usage

- Gateway into High-Performance Computing capabilities through HTTP APIs
- Used for jobs which need HPC capabilities, but through same interfaces for end users within the DE

**docker**

- Software container system, registry, etc.
- Known and used by scientists already
- Used for encapsulation & reproducibility of execution, distribution of tools within cluster, ease of use
A brief tour of the interfaces around jobs
App interface
(for final users who just want to do some science)
App editor interface (for intermediate users or internal people, creating interfaces around tools)
Workflow editor (a bit rudimentary though...)
Tool interface

(for software authors or internal people, who build images that get sent to internal Docker registry)
Now let's talk about just the HTCondor part, behind the scenes
(maybe I should focus this a bit)
Future (potential) Work & Ideas

- Time limitation (e.g., 48h per job, then request more)
- Other computational limits (number of processes (no forkbombs!), memory, CPU, network)
- Interactive jobs (iPython etc., or even just for exploration)
- More complex (non-linear, more flexible) workflows
- Public APIs and command-line tools (expand into more technical users)
- “connect your own compute” – spin up machines with appropriate software, put keys/other credentials into DE and have your own jobs go to non-limited, no-waiting compute you control

(question time!)

https://de.iplantcollaborative.org/
(try it out yourself)