Galaxy and Condor:
Challenges with harnessing widely-distributed resources

Condor Project
Computer Sciences Department
University of Wisconsin-Madison

Condor

- Batch scheduler
  - Similar to PBS, SGE
  - Manages a cluster of machines that can run Galaxy tasks
- Well-suited to widely-distributed systems and sharing resources between groups

James Thomson Lab

- Stem cell research
- http://discovery.wisc.edu/home/morgridge/research/regenerative-biology/
- Use Galaxy
- Condor cluster
  - 72 cpus
- We wrote Galaxy module to run tasks using Condor

Additional Machines

- Dozen Condor clusters at UW
  - 17,000 cpus
- Open Science Grid
  - Collaboration of 100 academic institutions
  - 80,000 cpus
- Amazon EC2 and similar
  - How much do you want to spend?

Problem

- Access to data
  - No shared file system
  - Condor can transfer files
    - Full list of files not easily available
    - Tasks arguments and input need rewriting
- Applications
  - Probably not installed

First Solution

- Write custom wrappers
- Time-consuming
  - Only suitable for most-used tools
- Not easily re-usable by other Galaxy users

New Solution

- Parrot
  - Developed at UW-Madison and Notre Dame
  - http://nd.edu/~ccl/software/parrot/
- Transparently intercept all disk I/O
  - Perform I/O on Galaxy machine
  - Use http cache for common input files
  - Reduce I/O for sparse file access