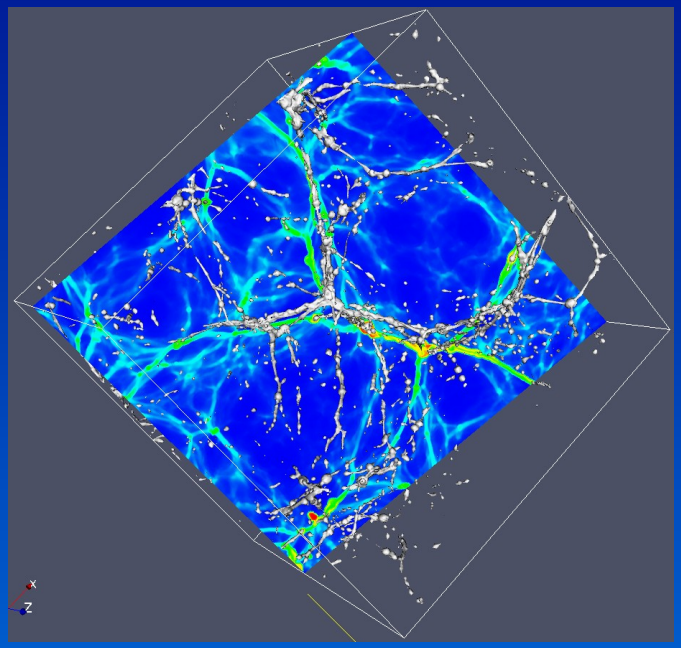
# Galaxy Cluster Simulations at NCSA using FLASH and Teuthis

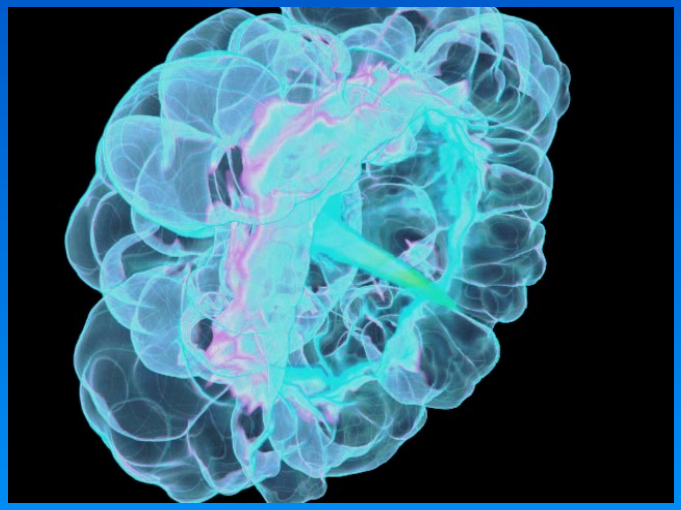
Paul Ricker
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

SC06
November 14, 2006

# Astrophysical simulations

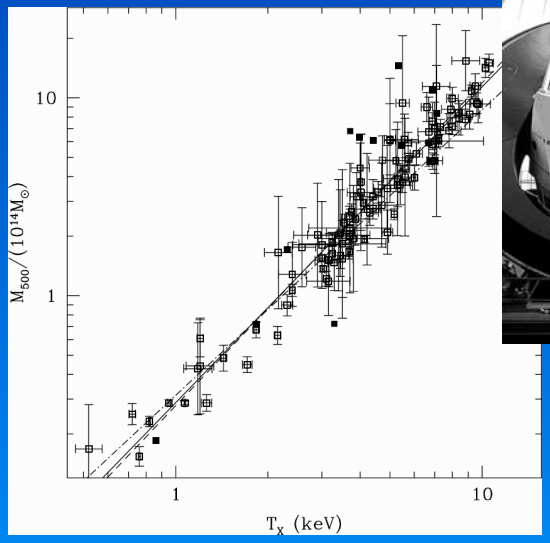
- Typical problems
  - Cosmological structure formation
  - Supernova explosions
  - Planetary disks
- Algorithms
  - Partial differential equations
  - Feature extraction
  - Statistical analysis
  - Ray casting
- Data types
  - Large checkpoint files (100 GB+)
  - Data subsets for plotting
  - Object/feature catalogs
  - Images and movies

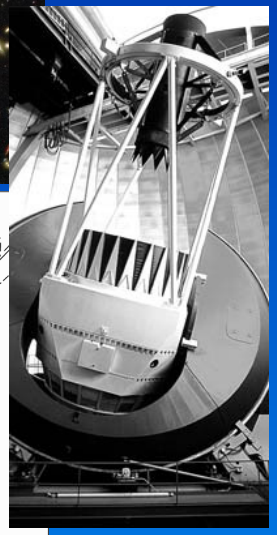




# Galaxy clusters

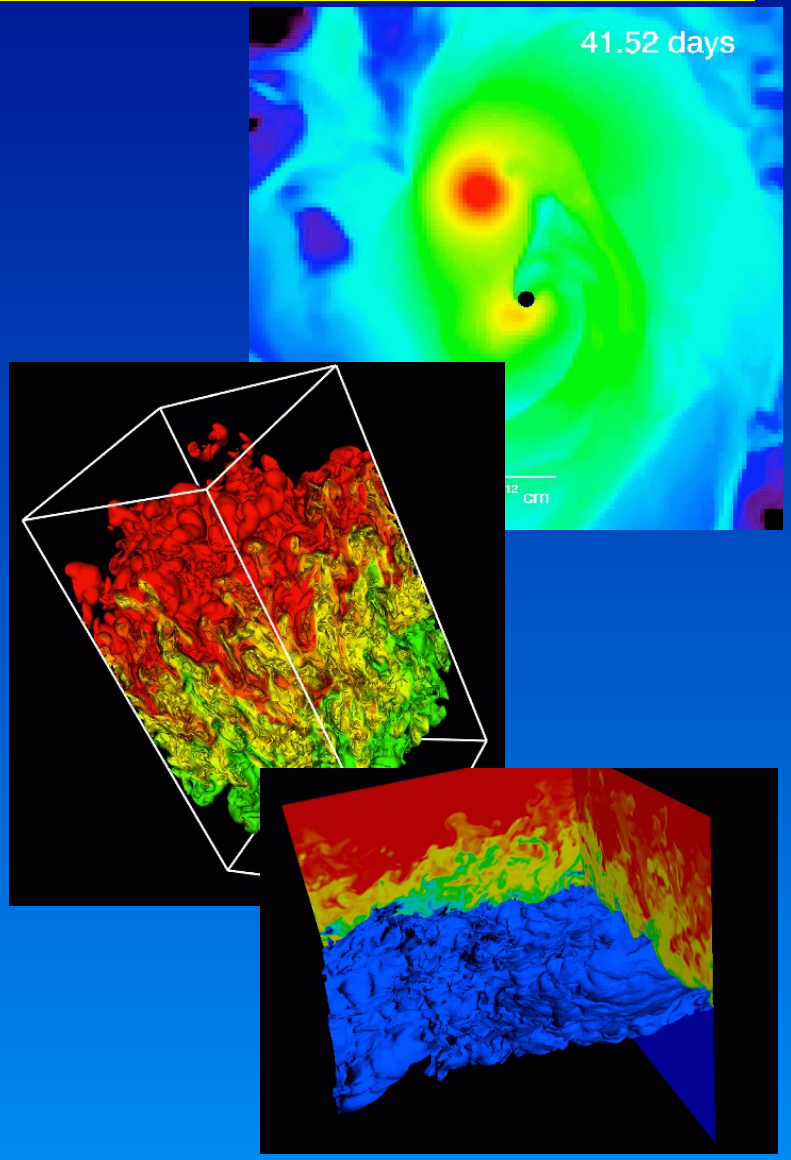
- Largest gravitationally bound objects in the Universe
  - ~ few to > 1000 galaxies
  - ~ $10^{13\text{-}15}$ solar masses
  - ~ few x $10^6$ light-years across
- Simulations
  - Form through gravitational instability
  - Create many to study statistics
  - Simulated observations to compare with cluster surveys



# What is FLASH?

- Astrophysical simulation code developed under DOE ASCI program at University of Chicago
  - Adaptive mesh refinement (AMR)
  - Particles and gravity
  - Nuclear reactions
- Community code
  - Free: http://flash.uchicago.edu
  - Modular framework (560,000 lines)
  - MPI parallel
  - Validated against experiments
  - 200+ users



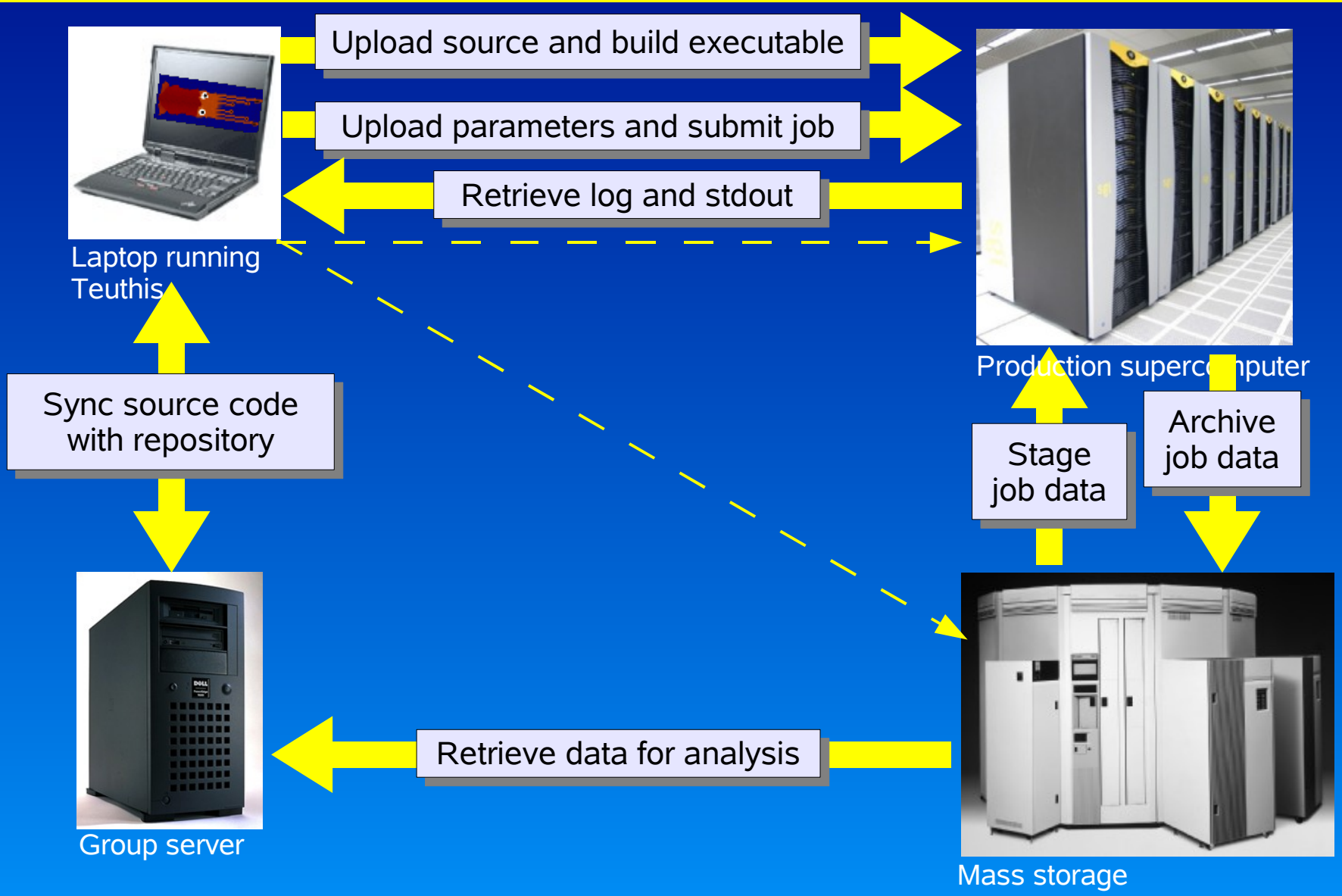
# What is Teuthis?

- A control panel
  - Remotely configure and build applications
  - Submit and track remote jobs
  - Painlessly create parameter studies and restart jobs
- A data manager
  - Stage and archive data
  - Keep track of where datasets are stored
- A notebook
  - Organize job metadata by purpose and disposition
- An aid to collaboration
  - Share notebook files with collaborators

# Design philosophy

- Small is beautiful
  - Small footprint – run on Tungsten someday?
  - Minimal prerequisites – avoid dependency hell
- Exploit others' expertise
  - Use external tools when possible…
  - … but only for “extra” functionality
- Think locally
  - Single point of interaction with one's work
  - Local metadata store complete and authoritative
- Don't get tied down
  - Separate GUI from backend
  - Everything open-source and cross-platform

# Running simulations with Teuthis



# Objects manipulated by Teuthis

## Workflow hierarchy

**Project**
Galaxy cluster scaling relations

**Experiment**
How does the level of galaxy feedback affect the mass-temperature relation?

**Run**
Run with 10x fiducial energy input

**Job**
Job 123456 on 128 processors for 18 hours

**Job**
Restart job 789012 on 256 processors for 18 hours

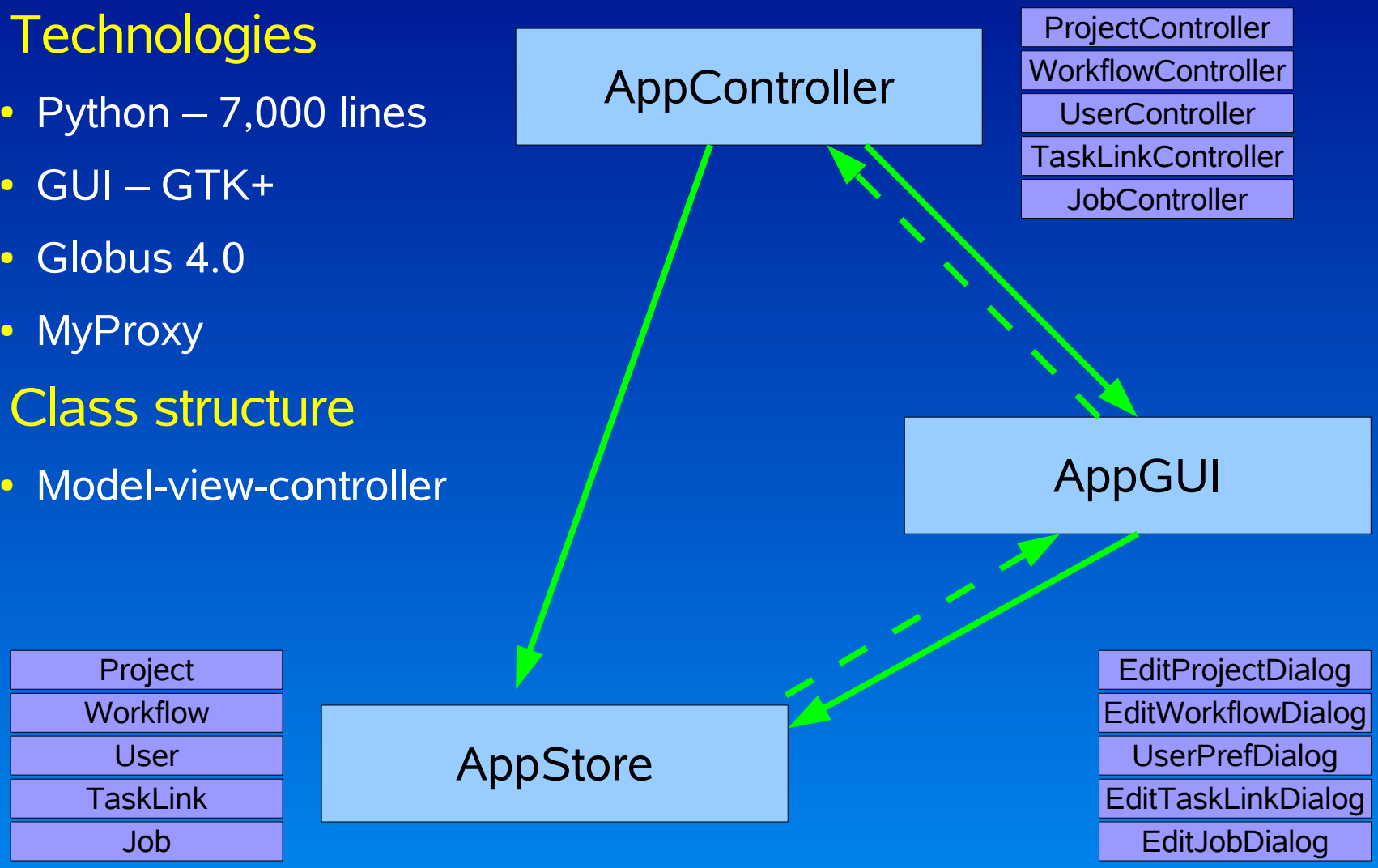
## Resources

**Application**
- Accepts text parameter file
- Executes noninteractively
- May need to be compiled
- Produces log file, screen output, data files

**Machine**
- Login host
- Access method
- Queuing system
- Paths and commands

# Under the hood

- Technologies
  - Python – 7,000 lines
  - GUI – GTK+
  - Globus 4.0
  - MyProxy
- Class structure
  - Model-view-controller



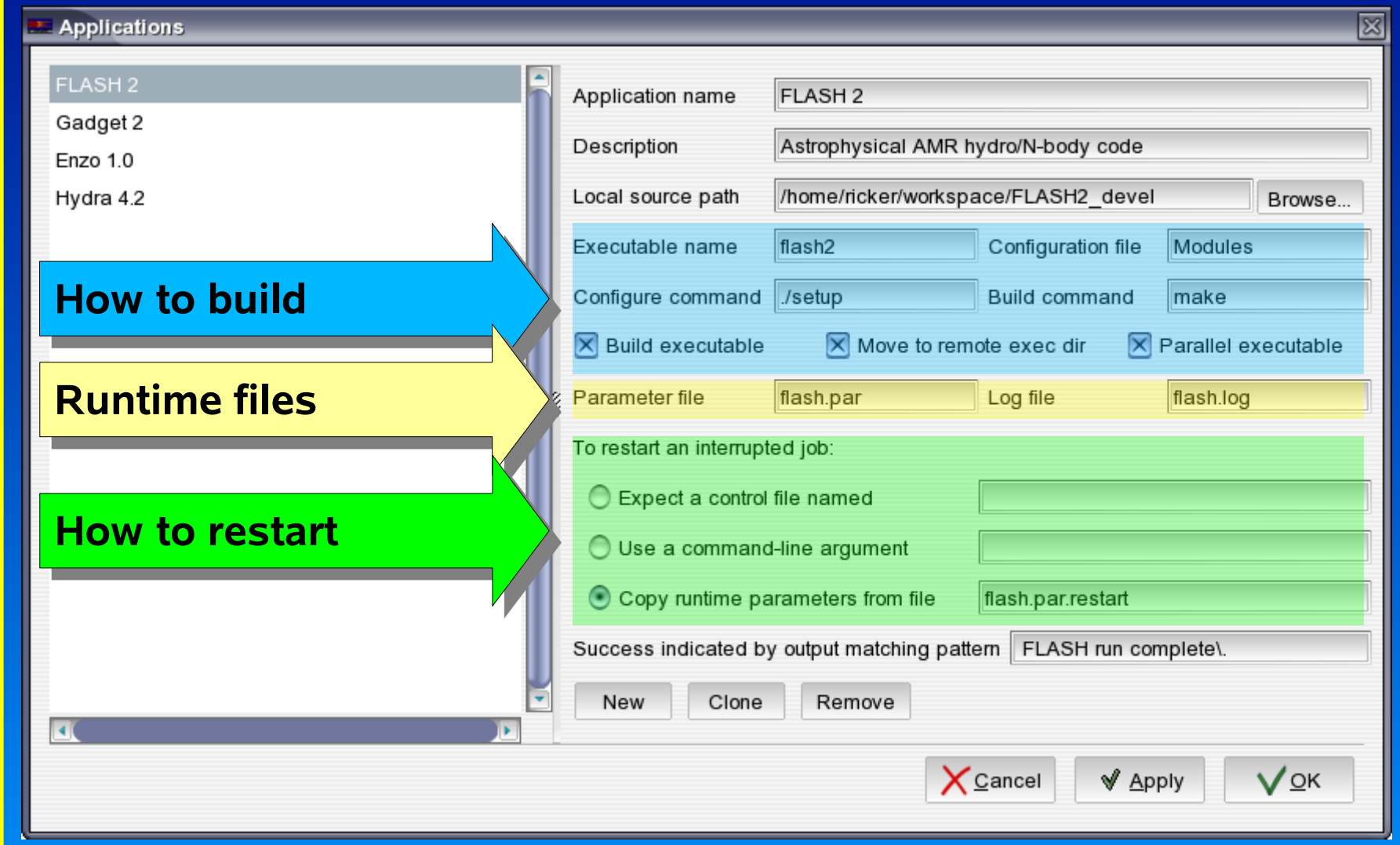
# Project view

Simulation Manager 1.0

File View Settings Help

| Name | Description | Status | Date last modified |
|---|---|---|---|
| **FLASH testing** | **Testing Simulation Manager using FLASH** | | **Tue Oct 4 17:1** |
| **Basic Sedov test (local)** | **Test of local jobs** | | **Tue Oct 4 19:14:5** |
| Run A | | | Tue Oct 4 19:15:0 |
| **Basic Sedov test (cobalt)** | **Test of jobs on a machine with PBS queuing and using** | | **Wed Oct 5 01:43:** |
| Run A | Single run using default parameters to test job submission | Complete | Wed Oct 5 01:17: |
| *Job A0004* | *Original* | *28909 [21:26 10/04/2005] 1 CPU/00:10 (Complete) Successful completion* | *Wed Oct 5 01:50:* |
| *Job A0004* | *Restart of 28909* | *28910 [21:27 10/04/2005] 1 CPU/00:10 (Complete) Successful completion* | *Wed Oct 5 01:16:* |
| **Sedov scaling test (cobalt)** | **Test with varying number of processors** | | **Wed Oct 5 01:43:** |
| Run A1 | | Complete | Wed Oct 5 02:07: |
| *Job A10001* | *Original* | *28924 [01:44 10/05/2005] 1 CPU/00:10 (Complete) Successful completion* | *Wed Oct 5 02:05:* |
| Run A2 | | Complete | Wed Oct 5 02:07: |
| *Job A20001* | *Original* | *28925 [01:45 10/05/2005] 2 CPUs/00:10 (Complete) Successful completion* | *Wed Oct 5 02:05:* |
| Run A4 | | Complete | Wed Oct 5 02:07: |
| *Job A40001* | *Original* | *28926 [01:45 10/05/2005] 4 CPUs/00:10 (Complete) Successful completion* | *Wed Oct 5 02:05:* |
| Run A8 | | Complete | Wed Oct 5 02:07: |
| *Job A80001* | *Original* | *28927 [01:45 10/05/2005] 8 CPUs/00:10 (Complete) Successful completion* | *Wed Oct 5 02:05:* |
| **Sedov test with varying parameter (cobalt)** | **Test of jobs with a single varying parameter (lrefine_mi** | | **Wed Oct 5 02:09:** |
| Run A | | Complete | Wed Oct 5 02:08: |
| *Job A0001* | *Original* | *28928 [01:52 10/05/2005] 1 CPU/00:10 (Complete) Successful completion* | *Wed Oct 5 01:58:* |
| *Job A0001 Copy* | *Original* | *29231 [15:01 10/05/2005] 1 CPU/00:10 (Complete) No data* | *Wed Oct 5 15:01:* |
| Run B | | Complete | Wed Oct 5 02:08: |
| *Job B0001* | *Original* | *28929 [01:52 10/05/2005] 1 CPU/00:10 (Complete) Successful completion* | *Wed Oct 5 01:58:* |
| Run C | | Complete | Wed Oct 5 02:08: |
| *Job C0001* | *Original* | *28930 [01:53 10/05/2005] 1 CPU/00:10 (Complete) Successful completion* | *Wed Oct 5 01:58:* |
| Run D | | Complete | Wed Oct 5 02:08: |
| *Job D0001* | *Original* | *28931 [01:53 10/05/2005] 1 CPU/00:20 (Complete) Successful completion* | *Wed Oct 5 01:59:* |
| Run E | | In progress | Wed Oct 5 02:08: |
| *Job E0001* | *Original* | *28932 [01:53 10/05/2005] 1 CPU/00:20 (Complete) Successful completion;* | *Wed Oct 5 02:02:* |
| Run F | | In progress | Wed Oct 5 02:09: |
| *Job F0001* | *Original* | *28933 [01:53 10/05/2005] 1 CPU/00:30 (Complete) Exceeded MAXBLOCKS* | *Wed Oct 5 01:55:* |

# Configuring applications



# Tested applications

| Application | Exec name and args | Config file | Config command | Build command | Move executable | Parallel | Parameter file | Log file | Auto restart method | Notes |
|---|---|---|---|---|---|---|---|---|---|---|
| FLASH 2.x | flash2 | Modules | ./setup | gmake | optional | yes | flash.par | flash.log | Copy from flash.par.restart; or command line argument "-chk_file" followed by manual addition of checkpoint file name | flash.par.restart not available in standard distribution; need patch<br>Need to set up site directory for remote site<br>Leave log_file parameter unset |
| Gadget 2 | Gadget2 gadget.param | Makefile | N/A | gmake | optional | yes | gadget.param | info.txt | Command line argument "1" | Upload custom makefile as your configuration file<br>Use "." for OutputDir parameter<br>Leave InfoFile parameter unset |
| Enzo 1.0.1 | enzo.exe EnzoParms | N/A | ./configure –bindir=XX | cd amr_mpi/src; gmake mach-YY; gmake; gmake install | yes | yes | EnzoParms | OutputLevelInformation | Command line argument "-r"; must manually add name of last restart file | XX = absolute path to build directory<br>Need to set up Make.mach.YY file for remote machine; place in config directory |
| Hydra 4.2 | hydra | makeflags | N/A | make clean; make | yes | no | prun.dat | pr0001.log | None; manually edit prun.dat | May need to create a new src/system.YY file for remote machine YY<br>Modify src/dumpdata.F, src/readdata.F, and src/gravsubs.F to read/write to ./ rather than data/<br>To change array sizes, edit include/psize.inc on local machine and sync source<br>Upload custom makeflags file as your configuration file; set RUNDIR to ".."<br>Use 0001 as run number in prun.dat |

# Configuring machines

Machines

cobalt
tungsten
tungsten (ssh-agent)

**Basic information**

Machine name: cobalt
Description: SGI Altix at NCSA
Type: Compute | Online data

**Access**

Host: cobalt.ncsa.uiuc.edu
User ID
gsissh+uberftp
Realm
X.509 certificate subject: /C=US/O=National Center
Auth init command: grid-proxy-init -cert %K -key %K -valid %
Remote exec command: gsissh -o BatchMode=yes -o GSSAPIAut
File upload command: gsiscp -B -r -p -q "%f" %u@%h:"%r"
File download command: gsiscp -B -r -p -q %u@%h:"%r" "%f"
Third-party transfer cmd: uberftp %i "lopen %h; put %f %r"

**Job submission**

Job template
Browse...
pbs
/node (MB): 1000
R Queues standard A R
ker/build/test
ker/exec
/users/ricker
Remote Teuthis dir: /u/ac/ricker/teuthis
OS type: Unix

**Actions**

New | Clone | Remove

Cancel | Apply | OK

**Access methods**

local
ssh
ssh-agent
Kerberos
GSI

**Queuing methods**

Unix (no queue)
PBS
LSF
LoadLeveler
User-specified

# Experiment dialog

1 Configure and build application

2 Set up input and archive data

3 Set up execution host

4 Choose parameters to fix/vary

5 Generate runs

Experiment properties

**Information**

Experiment name: Sedov test with file transfer

Description: Sedov test on a remote machine with two third-party file transfers.

**Data**

Src machine: tungsten

Src files: /u/ac/ricker/input/la128.tar
/u/ac/ricker/input/lcdms64.tar

**Application**

Application: FLASH 2.4 | Upload source

Configuration file: Browse... | Upload file

Remote build dir: /u/ac/ricker/build/test/FLASH_2.4

Config command: ./setup sedov -auto | Do it | Output

Build command: make; mv flash2 /u/ac/ricker/exec | Do it | Output

Executable to use: /u/ac/ricker/exec/flash2

**Execution**

Exec machine: cobalt

Queue: normal | Account:

# of CPUs (range): 1 | Tiling: 1 | Mem/node (MB): 1000

**Actions**

Generate runs | Clone experiment

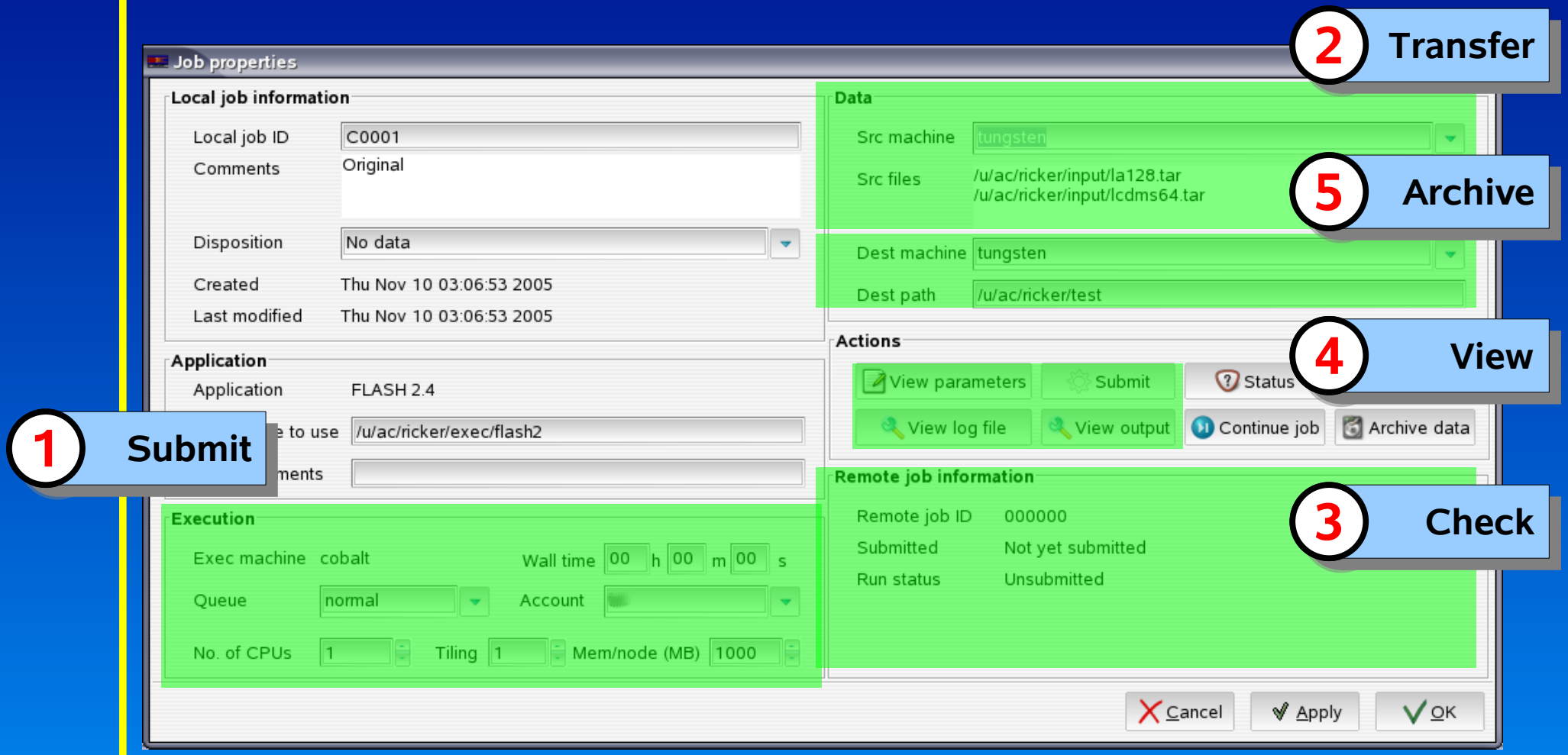
**Parameters**

Template: /home/ricker/flash.par | Browse... | Edit

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Parameter 1 | lrefine_max | Range | 1-6 | Parameter 5 | | Range | |
| Parameter 2 | | Range | | Parameter 6 | | Range | |
| Parameter 3 | | Range | | Parameter 7 | | Range | |
| Parameter 4 | | Range | | Parameter 8 | | Range | |

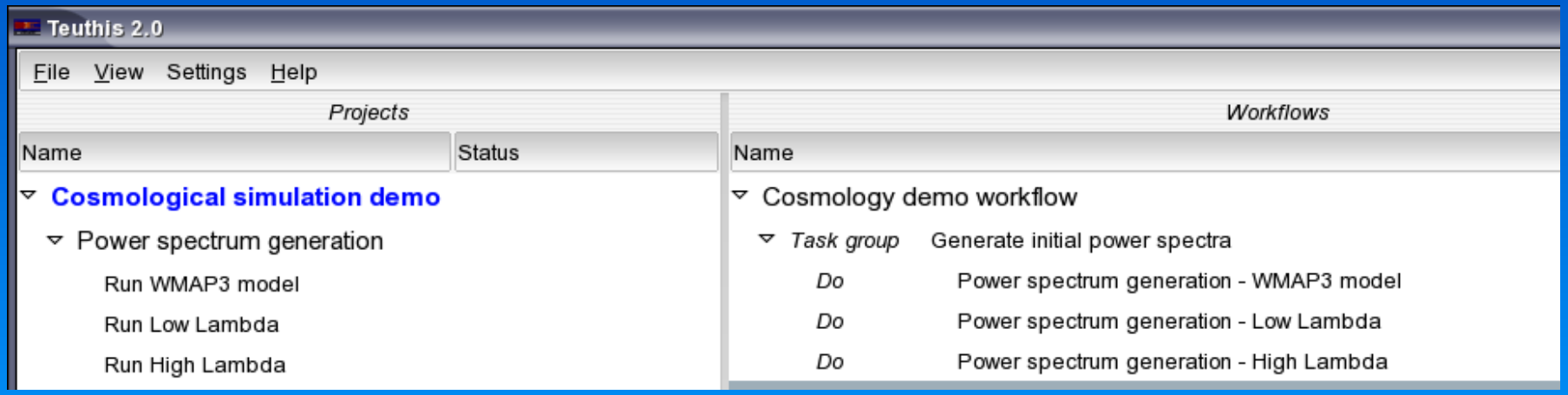
Cancel | Apply | OK

# Job dialog


Job properties
Local job information
Local job ID
C0001
Comments
Original
Disposition
No data
Created
Thu Nov 10 03:06:53 2005
Last modified
Thu Nov 10 03:06:53 2005
Application
Application
FLASH 2.4
to use
/u/ac/ricker/exec/flash2
ments
Execution
Exec machine
cobalt
Wall time
00 h 00 m 00 s
Queue
normal
Account
No. of CPUs
1
Tiling
1
Mem/node (MB)
1000
Data
Src machine
tungsten
Src files
/u/ac/ricker/input/la128.tar
/u/ac/ricker/input/lcdms64.tar
Dest machine
tungsten
Dest path
/u/ac/ricker/test
Actions
View parameters
Submit
Status
View log file
View output
Continue job
Archive data
Remote job information
Remote job ID
000000
Submitted
Not yet submitted
Run status
Unsubmitted
Cancel
Apply
OK
1 Submit
2 Transfer
3 Check
4 View
5 Archive

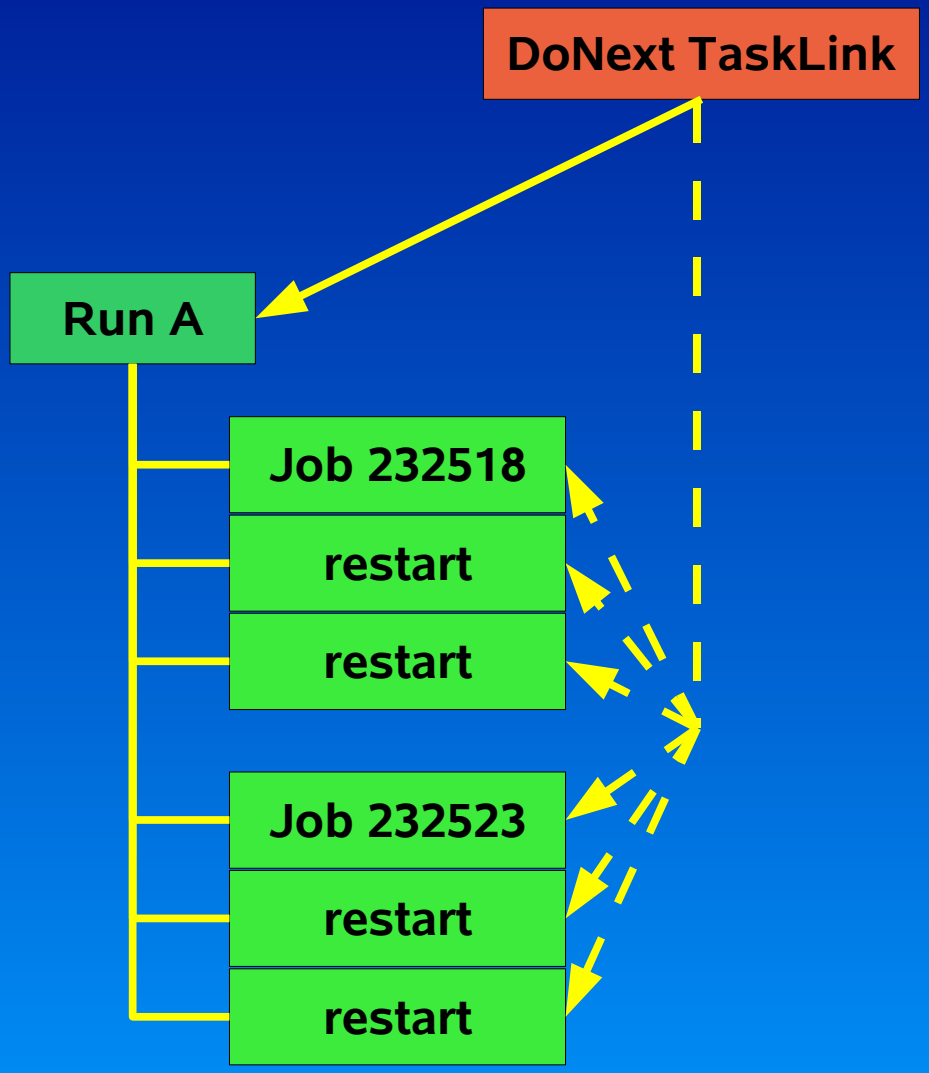
# New features in Teuthis 2.0

- Data
  - Background file transfers with retry
- Job submission and monitoring
  - GRAM job submission
  - Workflow management
- User interface
  - Refactoring
  - Usability improvements



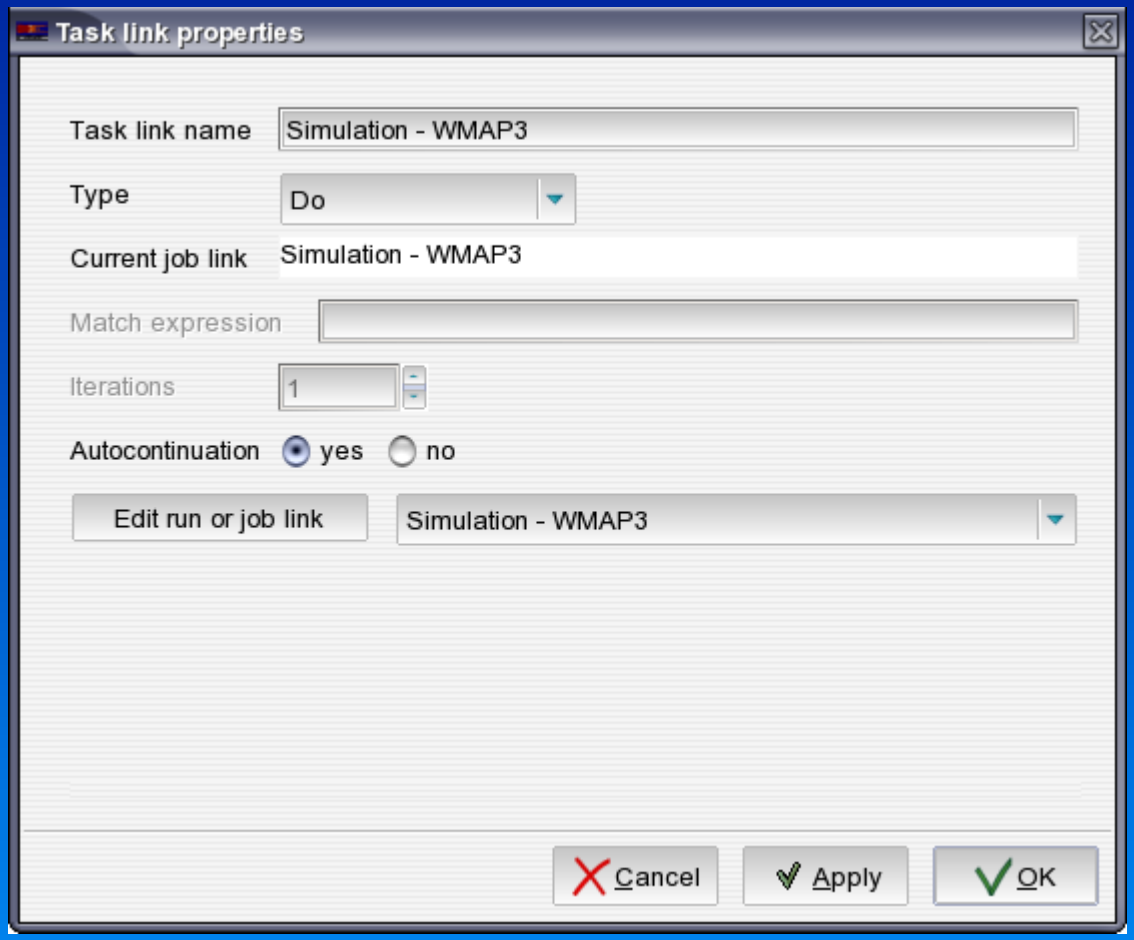
# Teuthis workflow management

- Task links
  - Atomic workflow unit
  - Linked to a particular run
  - Invocation triggers new job: stage in – exec – stage out
  - Automatic job continuation
  - Pattern matching conditionals
- Types
  - Data source – Static
  - Unconditional – DoNext, DoTogether
  - Conditional – DoIf, WhileDo
  - Grouping – Task Groups



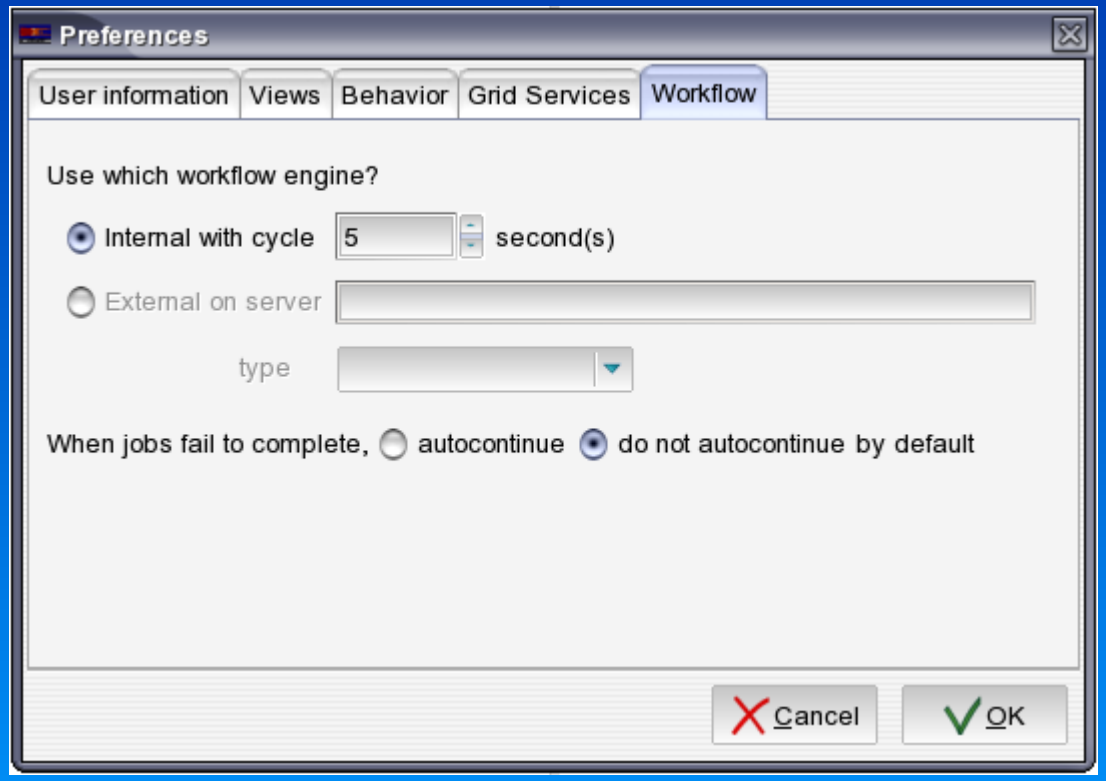
# Creating and editing task links

- Creating and editing task links
  - Add from popup menus
  - Drag runs/jobs from Project Pane



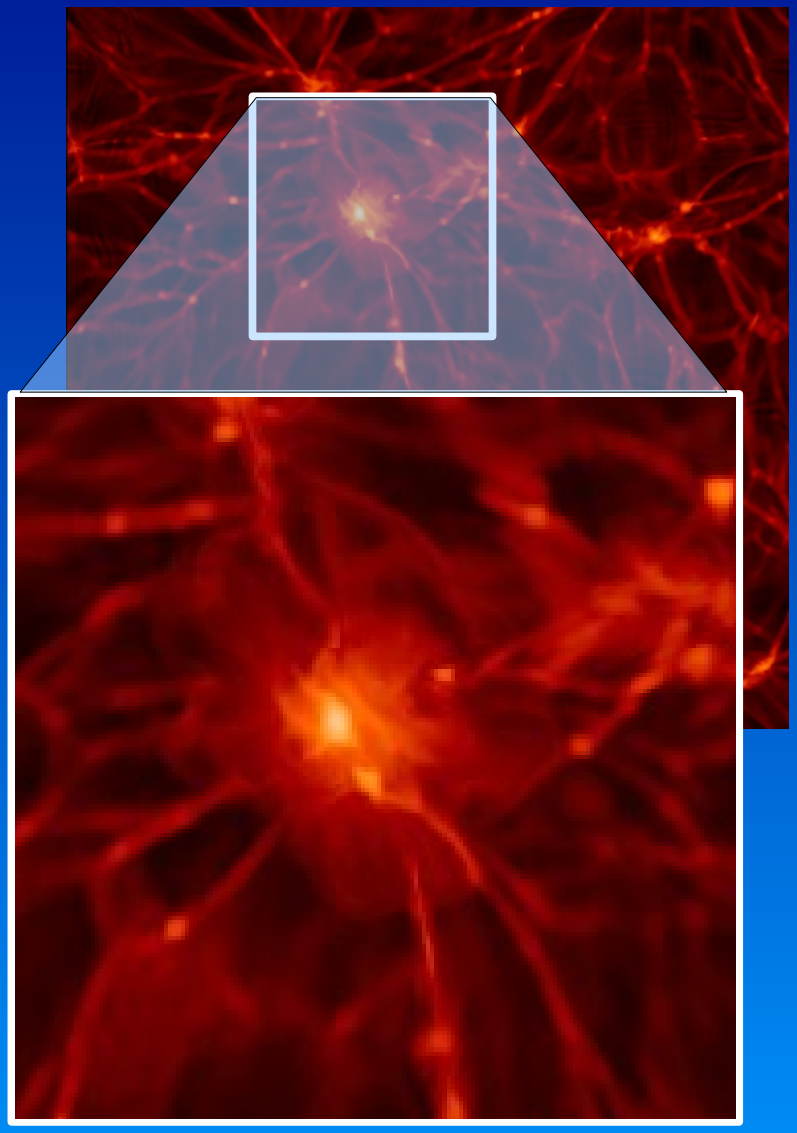
# Teuthis workflow management

- Workflow execution choices
  - Internal engine: use polling to test job status and advance at preset interval
  - External engine:  hand script off, watch for messages
  - Pause/halt/resume



# An example workflow

- Galaxy cluster formation
  - Gravity amplifies random initial perturbations
  - Very large dynamic range
- Resimulation approach
  - Large, low-resolution box to locate clusters
  - Resample initial volume containing matter that ends up in cluster
  - Resimulation this volume with surroundings at low resolution to provide boundary conditions



# Example workflow #1 – simple

Workstation

Supercomputer

Mass storage

Generate initial conditions

Run simulation

Find halos at present day

Find halo mass function

Archive initial conditions

Archive simulation outputs

Archive halo catalog

Stage in

Stage out

# Example workflow #1 – Teuthis implementation

Teuthis 2.0

File View Settings Help

*Projects*

| Name | Status |
|---|---|
| **Cosmological simulation demo** | |
| Power spectrum generation | |
| Run WMAP3 model | |
| *Job WMA0002* | *16399 [02:41 11/14/2006] 1 CPU/00:00 (Complete) No data* |
| Run Low Lambda | |
| Run High Lambda | |
| Initial conditions generation | |
| Run A | |
| *Job A0002* | *000000 [02:41 11/14/2006] 1 CPU/00:00 (Complete) No data* |
| Run B | |
| *Job B0001* | *16431 [02:42 11/14/2006] 1 CPU/00:00 (Complete) No data* |
| Run C | |
| Run D | |
| Simulation | |
| Run WMAP3 | |
| *Job WMA0002* | *16454 [02:42 11/14/2006] 1 CPU/00:00 (Complete) No data* |
| *Job WMA0003* | *000000 [Not yet submitted] 1 CPU/00:00 (Unsubmitted) No data* |
| Run Low Lambda | |
| Run High Lambda | |
| Halo catalog construction | |
| Run WMAP3 - realization A | |
| Run WMAP3 - realization B | |
| Run WMAP3 - realization C | |
| Run WMAP3 - realization D | |
| Plot halos | |
| Run WMAP3 - realization A | |
| Run WMAP3 - realization B | |
| Run WMAP3 - realization C | |
| Run WMAP3 - realization D | |
| Data movement | |
| Run Copy power spectrum data for | |
| Run Copy power spectrum data for | |
| Run Copy power spectrum data for | |
| Run Copy initial conditions for realiz | |
| Run Copy initial conditions for realiz | |
| Run Copy initial conditions for realiz | |

*Workflows*

| Name | | Status |
|---|---|---|
| Cosmology demo workflow | | Executing |
| *Task group* | Generate initial power spectra | Complete |
| *Do* | Power spectrum generation - WMAP3 model | Complete |
| *Task group* | Generate initial conditions | Complete |
| *Do* | Data movement - Copy power spectrum data for WMAP3 | Complete |
| *Do* | Initial conditions generation - A | Complete |
| *Do* | Initial conditions generation - B | Complete |
| *Task group* | Run simulations | In Group |
| *Do* | Data movement - Copy initial conditions for realization A | Complete |
| *Do* | Simulation - WMAP3 | Complete |
| *Do* | Data movement - Copy z=0 output for realization A | Complete |
| *Do* | Data movement - Copy initial conditions for realization B | Complete |
| *Do* | Simulation - WMAP3 | |
| *Do* | Data movement - Copy z=0 output for realization B | |
| *Task group* | Convert z=0 outputs to PMHDF output format | |
| *Do* | Data format conversion - WMAP3 - realization A | |
| *Do* | Data format conversion - WMAP3 - realization B | |
| *Task group* | Construct halo catalogs | |
| *Do* | Halo catalog construction - WMAP3 - realization A | |
| *Do* | Halo catalog construction - WMAP3 - realization B | |
| *Task group* | Make halo plots | |
| *Do* | Plot halos - WMAP3 - realization A | |
| *Do* | Plot halos - WMAP3 - realization B | |

# Example workflow #2 - resimulation


Workstation


Supercomputer


Mass storage

Generate level 0 initial conditions

Run low resolution large box

Find halos at present day

Choose sample to resimulate

Archive initial conditions

Archive simulation outputs

Archive halo catalog

Stage in

Stage out

# Example workflow #2 - resimulation


Workstation


Supercomputer


Mass storage

Do together (one for each cluster in sample)

Locate halo particles in initial conditions → Generate level 1 initial conditions → Run high resolution box → Halo finder → Simulated observation

Generate level 1 initial conditions → Archive initial conditions

Run high resolution box → Archive simulation outputs

Halo finder → Archive halo catalogs

Simulated observation → Archive simulated observations

# Future plans

- More sophisticated experiment designs
  - Latin hypercube, random samples
- More sophisticated workflows
  - DoTogether with multiple execution threads
  - DoTogether with linked threads (e.g. data mover)
- Integration with other workflow engines
  - Export workflow script and submit
- More complete data tracking
  - Store lists of files transferred with types, checksums, sizes
  - Should be able to store other data besides logs/stdout
- Integration with observational data management tools
  - Portal version

# Getting Teuthis

1.0 release available at

http://mazama.ncsa.uiuc.edu/projects/teuthis