Computation Institute

UC3: A Framework for Cooperative Computing at the University of Chicago

UC3

uchicago

Marco Mambelli (marco@hep.uchicago.edu), Rob Gardner

Computation and Enrico Fermi Institutes

UC Computing Cooperative

- A shared Campus distributed high throughput computing infrastructure (DHTC)
- Inspired by need to promote resource sharing and "mutual opportunity" at the Campus level, with eye towards integration with national-scale resources such as the Open Science Grid
- Framework concept that leverages present and future investments from the funding agencies and the University

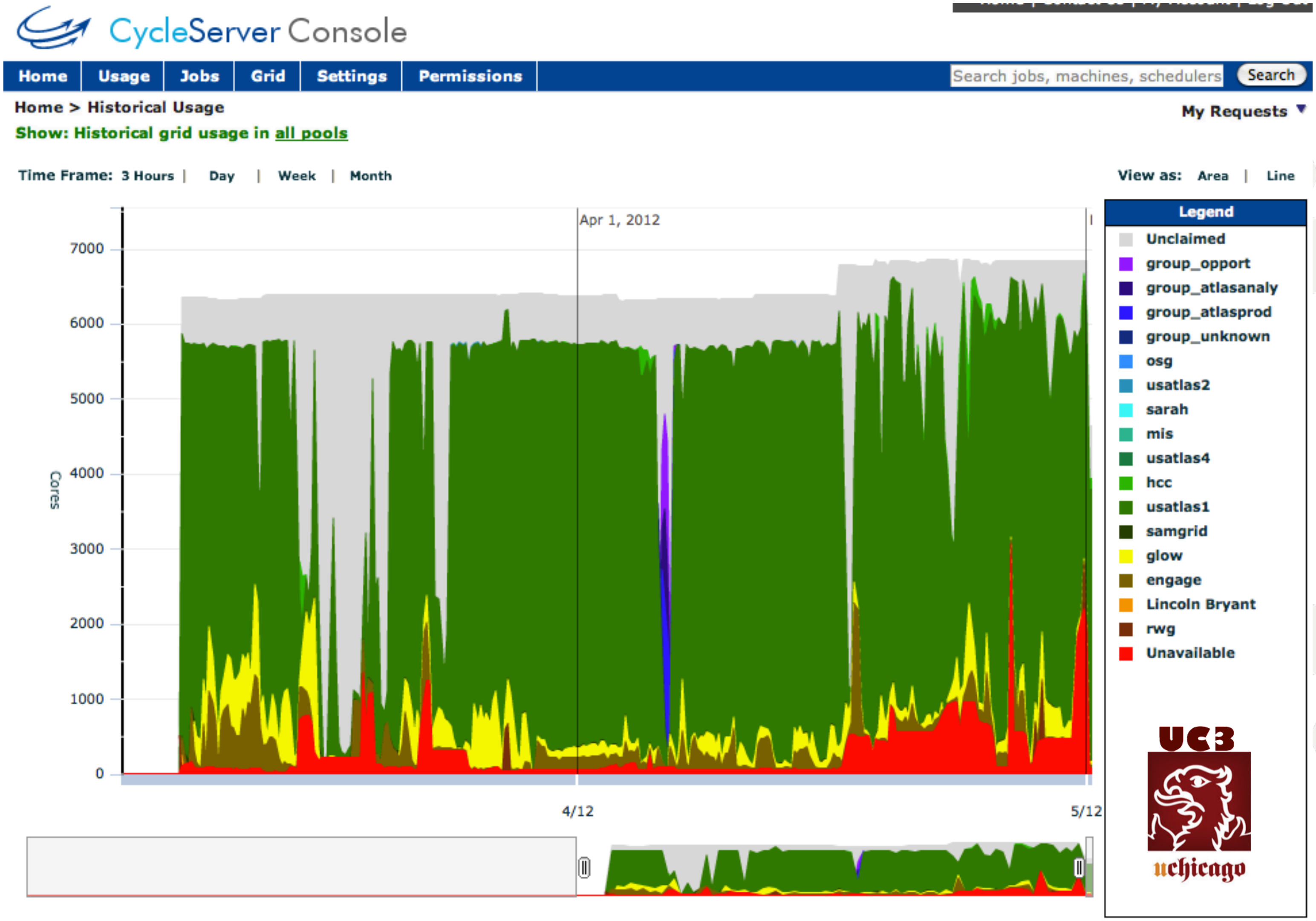
Collaboration and Acknowledgements

- Enrico Fermi Institute in the Physical Sciences Division
  - ATLAS Collaboration (HEP)
  - South Pole Telescope Collaboration
- Departments of Radiology and Radiation Oncology (SIRAF project)
- Computation Institute at UC (OSG, Swift)
- Center for Robust Decision Making on Climate and Energy Policy group at UC (CI, Economics)
- UC Information Technology Services
- UC Research Computing Center

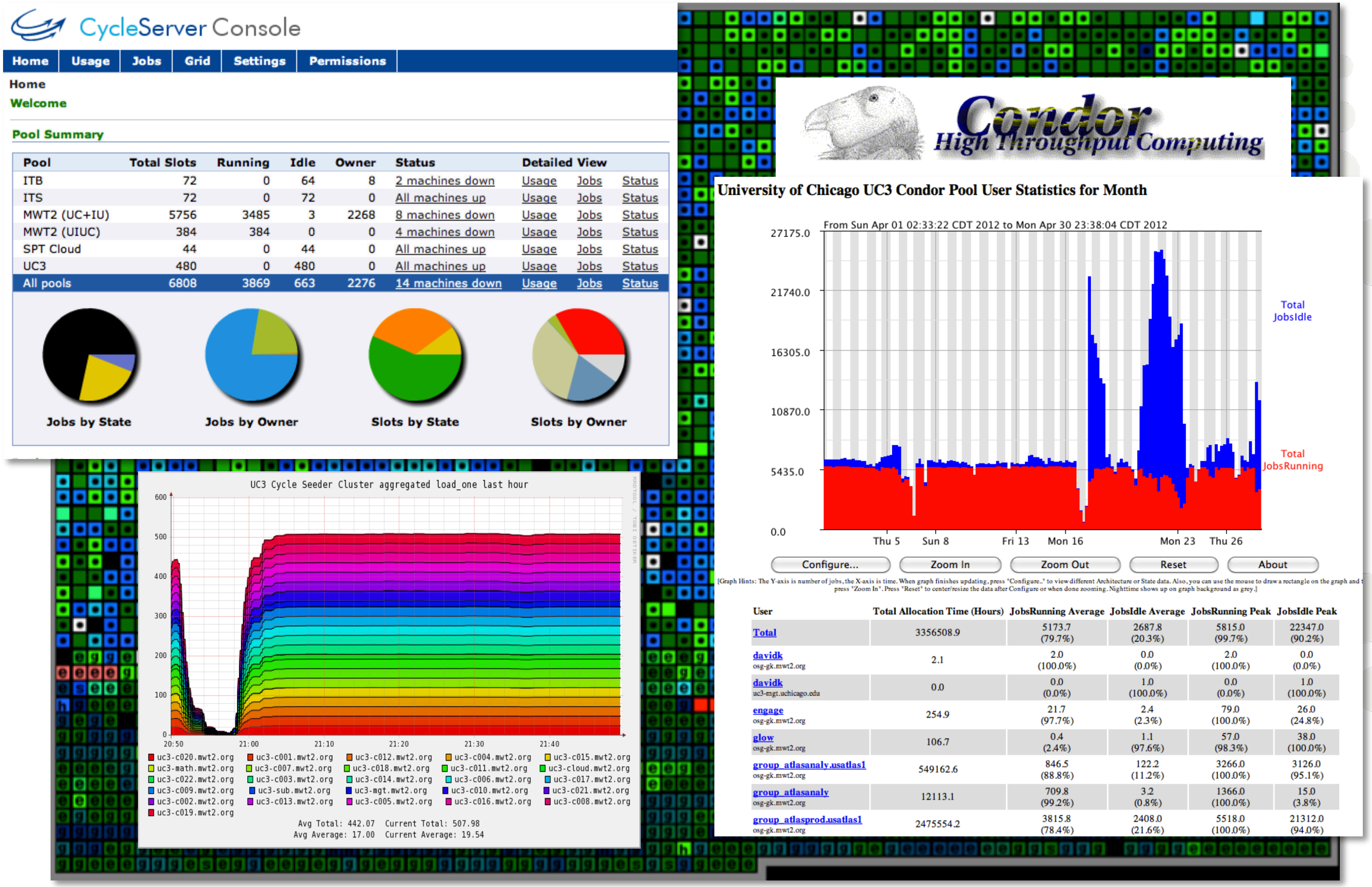
Building UC3 - principles

- UC3 focus is solely on DHTC
- UC3 participating resource owners control their assets and local policies & contribute to the collective infrastructure as possible
- UC3 will have or use a baseline set of services for job management, data access, monitoring and identity management
- Community-oriented with a focus on connecting computational resources and scientists
- Grass-roots driven by U Chicago faculty from various University divisions and institutes
- UC3 has in its scope connecting to resources off-campus (regional resources, Open Science Grid, …) driven by community demand

Install Fest – March 2012



Monitoring & Console – using tools out there!



CycleServer Console

Home

Welcome

Pool Summary

Pool	Total Slots	Running	Idle	Owner	Status	Detailed View		
ITB	72	0	64	8	2 machines down	Usage	Jobs	Status
ITS	72	0	72	0	All machines up	Usage	Jobs	Status
MWT2 (UC+IU)	5756	3485	3	2268	8 machines down	Usage	Jobs	Status
MWT2 (UIUC)	384	384	0	0	4 machines down	Usage	Jobs	Status
SPT Cloud	44	0	44	0	All machines up	Usage	Jobs	Status
UC3	480	0	480	0	All machines up	Usage	Jobs	Status
All pools	6808	3869	663	2276	14 machines down	Usage	Jobs	Status

University of Chicago UC3 Condor Pool User Statistics for Month

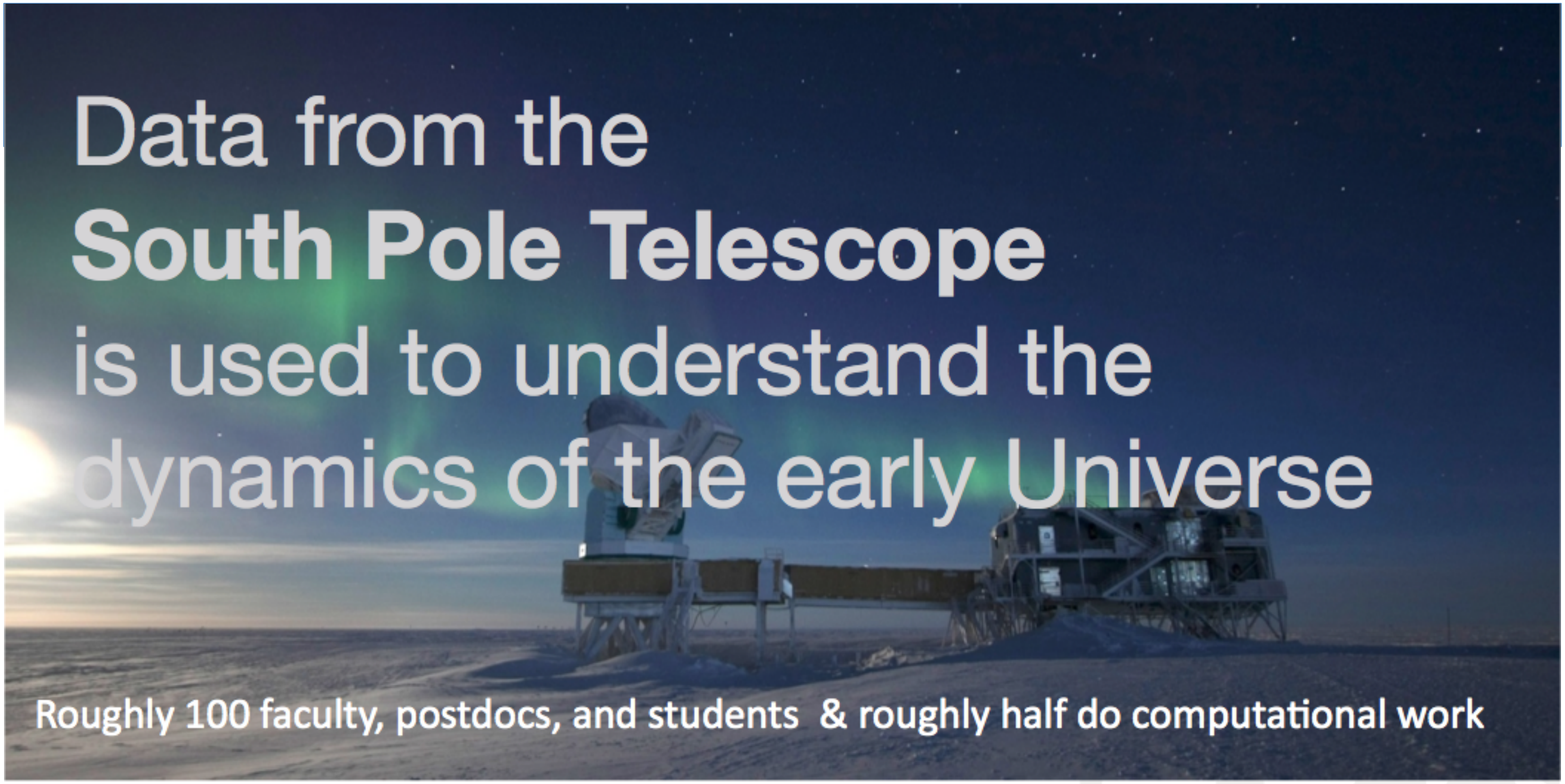
From Sun Apr 01 02:33:22 CDT 2012 to Mon Apr 30 23:38:04 CDT 2012

[Graph Hints: The Y-axis is number of jobs, the X-axis is time. When graph finishes updating, press "Configure.." to view different Architecture or State data. Also, you can use the mouse to draw a rectangle on the graph and press "Zoom In". Press "Reset" to center/resize the data after Configure or when done zooming. Nighttime shows up on graph background as grey.]

User	Total Allocation Time (Hours)	JobsRunning Average	JobsIdle Average	JobsRunning Peak	JobsIdle Peak
Total	3356508.9	5173.7 (79.7%)	2687.8 (20.3%)	5815.0 (99.7%)	22347.0 (90.2%)
davidk osg-gk.mwt2.org	2.1	2.0 (100.0%)	0.0 (0.0%)	2.0 (100.0%)	0.0 (0.0%)
davidk uc3-mgt.uchicago.edu	0.0	0.0 (0.0%)	1.0 (100.0%)	0.0 (0.0%)	1.0 (100.0%)
engage osg-gk.mwt2.org	254.9	21.7 (97.7%)	2.4 (2.3%)	79.0 (100.0%)	26.0 (24.8%)
glow osg-gk.mwt2.org	106.7	0.4 (2.4%)	1.1 (97.6%)	57.0 (98.3%)	38.0 (100.0%)
group_atlasanaly.usatlas1 osg-gk.mwt2.org	549162.6	846.5 (88.8%)	122.2 (11.2%)	3266.0 (100.0%)	3126.0 (95.1%)
group_atlasanaly osg-gk.mwt2.org	12113.1	709.8 (99.2%)	3.2 (0.8%)	1366.0 (100.0%)	15.0 (3.8%)
group_atlasprod.usatlas1 osg-gk.mwt2.org	2475554.2	3815.8 (78.4%)	2408.0 (21.6%)	5518.0 (100.0%)	21312.0 (94.0%)

UC3 Cycle Seeder Cluster aggregated load_one last hour

Avg Total: 442.07 Current Total: 507.98
Avg Average: 17.00 Current Average: 19.54



Data from the **South Pole Telescope** is used to understand the dynamics of the early Universe

Roughly 100 faculty, postdocs, and students & roughly half do computational work

Core SPT Institutions:

- Case Western Reserve University
- Harvard-Smithsonian Astrophysical Observatory
- Ludwig-Maximilians Universität / University of Illinois
- McGill University
- University of California, Berkeley
- University of California, Davis
- University of Chicago
- University of Colorado at Boulder

Other Participating Institutions:

- California Institute of Technology
- University of Michigan
- Yale University
- University of Arizona

The raw time streams per detector are recorded at 100 Hz, with 960 detectors, 60 Gigabytes per day via satellite to Chicago for processing

South Pole Telescope Collaboration

- Low-level processing on raw data and conversion to intermediate-level data products (IDL based)
- Simulated observations of "fake skies" (**main UC3 workload**)
  - Theoretical power spectrum fourier-transformed into a 2D real-space map. "Observe" using the actual telescope pointing information, make maps of what we would have seen given the fake sky and observing pattern. We then push this fake observation through the full pipeline to calculate systematic and statistical errors
- Exploring large-dimensional likelihood spaces with Markov Chain Monte Carlo methods
  - dark energy equation of state, the sum of the neutrino masses, the normalization of the matter power spectrum, the effective number of neutrinos in the early universe, the "tilt" of the primordial matter power spectrum
- Relation to UC3
  - Offload high throughput tasks from SPT core resources onto UC3 during interactive sessions

Radiological Imaging with SIRAF

- SIRAF – Scientific Image Reconstruction and Analysis Facility
  - Users of the UC Comprehensive Cancer Center
  - Medical physicists from departments of Radiology and Radiation Oncology
- Projects supported by SIRAF
  - Real-time computer aided diagnosis for diagnostic mammography
  - CAD for Lung Cancer Screening Using Computed Tomography
  - CAD for Breast Tomosynthesis
  - MR Imaging of Breast and Prostate with High Spectral and Spatial Resolution
  - Targeted Imaging in Helical Cone-Beam CT
  - Development and Evaluation of Receiver Operator Characteristic Software
  - Multi-modality CAD in Breast Imaging
  - Real-time CAD for Diagnosis of Lung Nodules

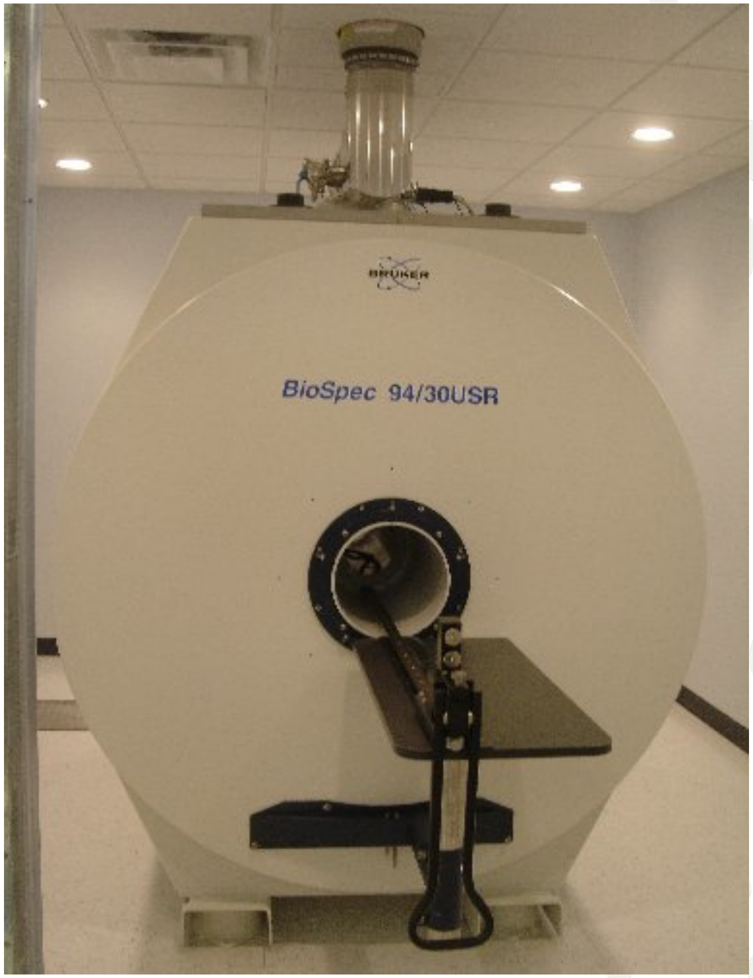
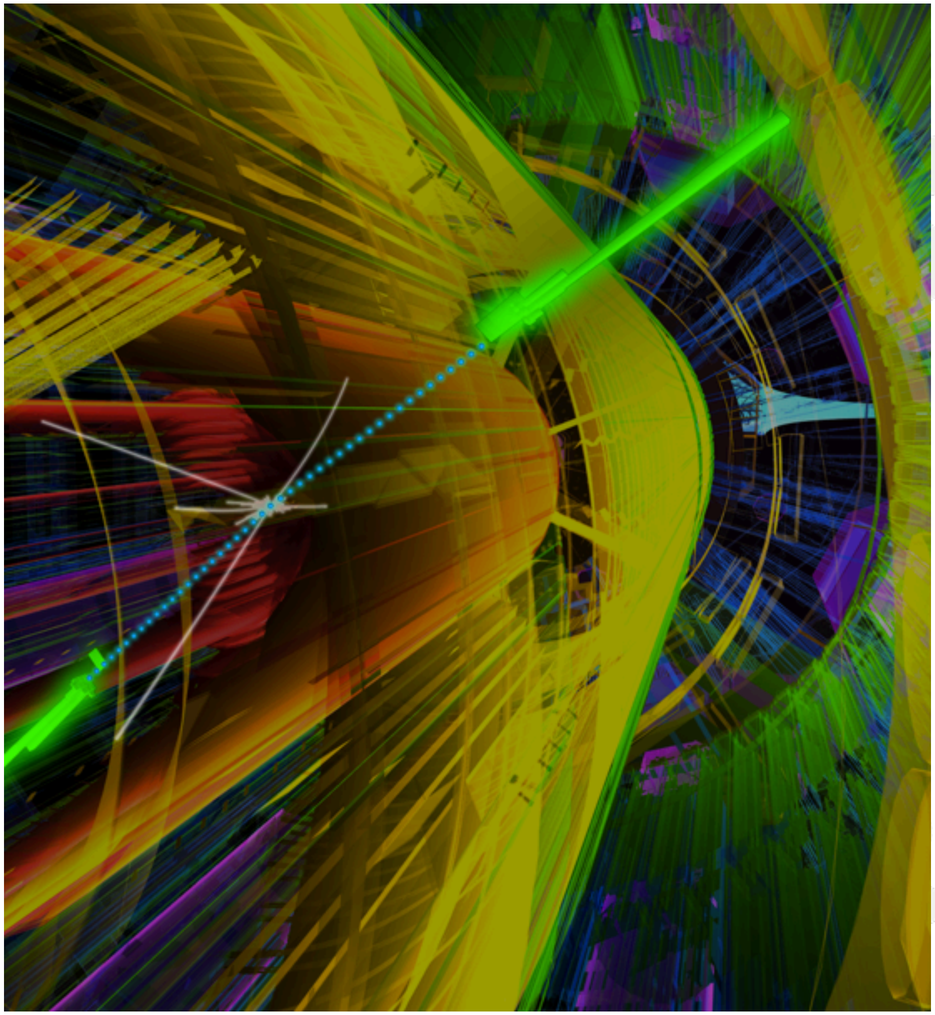


Image Reconstruction & Analysis

- Classes of problems addressed by SIRAF
  - Image processing of data from MRI, PET, xray, CT, SPECT, and tomosynthesis scanners to reconstruct 2D & 3D images
  - Image Analysis - Given a reconstructed image, derive relevant parameters of medical/biological interest. Similar to data mining in other domains. Many algorithms highly parallel or **high throughput - neural network training, genetic algorithms, Monte Carlos**
  - Biomaterial physics - radiation transport through living tissue/bone/organs. Used by RadOnc for treatment planning, new less damaging methods, etc. **Many methods are high throughput - Monte Carlos, ray tracing.**
  - Medical Visualization - presentation of reconstructed 2-D/3-D images often with annotations/enhancements derived from analysis. Usually interactive
- Relation to UC3
  - It is difficult to schedule both interactive and batch computations on the same cluster. By partnering with other UC3 members, we can offload many of the long running batch computations to the campus grid and maintain better interactive performance on more nodes during work hours, then make unused resources available during off-hours to other UC3 members.
  - SIRAF will upgrade to GPUs in Q4 2012 and will make available to UC3 campus grid users for development

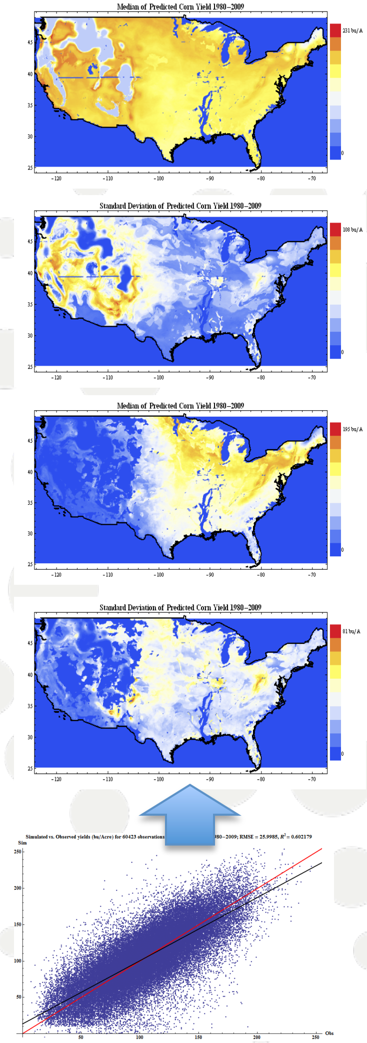
ATLAS at LHC

- Investigations at forefront particle physics at the CERN Large Hadron Collider
- UC3 partnering with both Tier 2 and Tier 3 data centers
- Provide flocking to unused ATLAS resources
- Allow flocking of ATLAS to spare UC3 cycles
- Facilitated with CERN Virtual File System for release directories, and federated Xrootd for storage access (→ minimal UC3 system modifications for a large class of jobs)



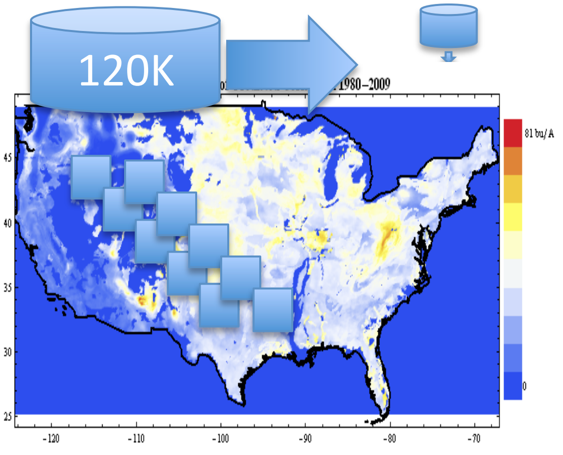
Economics and land use models

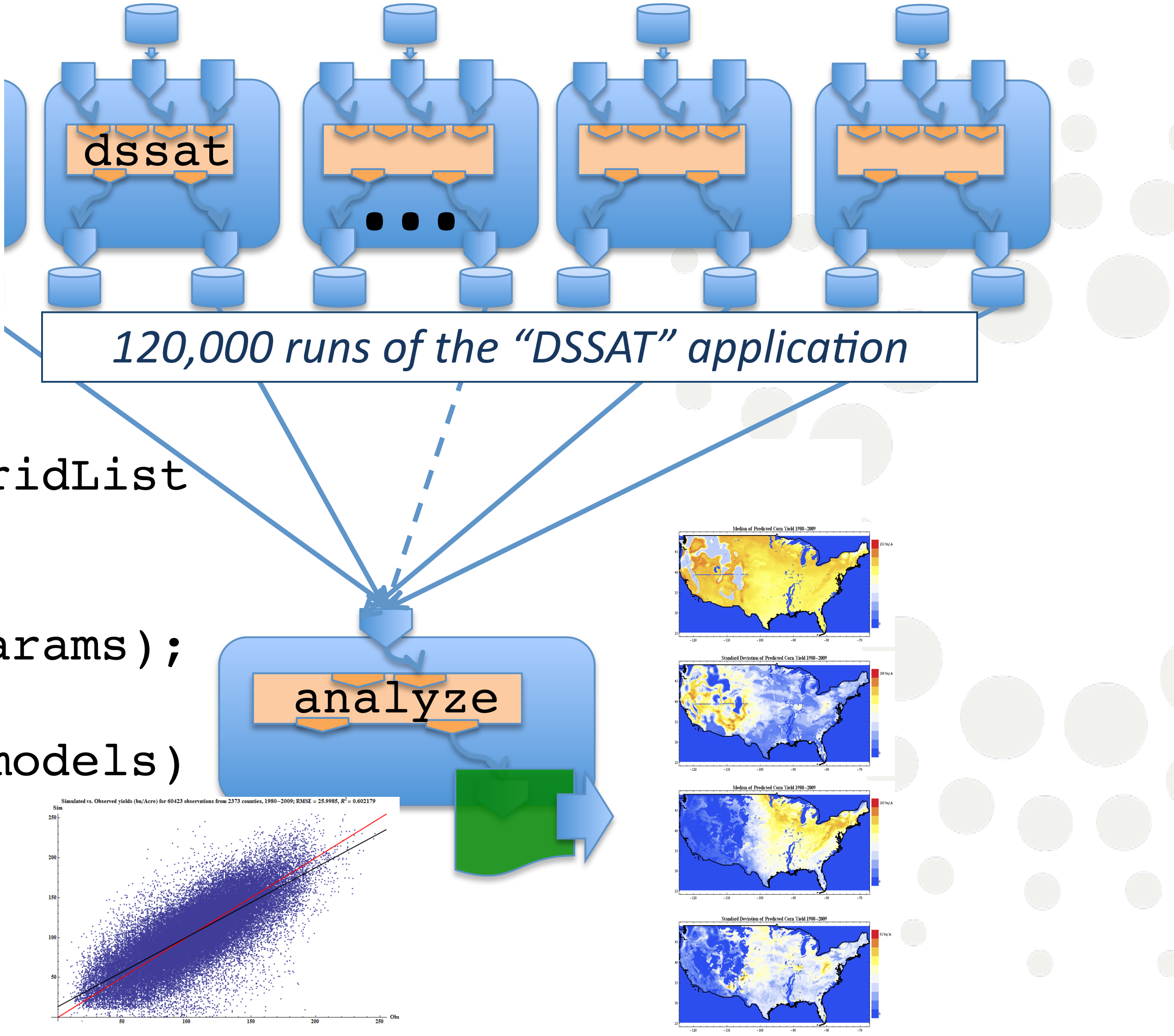
- The **CIM-EARTH** and **RCDEP** projects develop a large-scale integrated modeling frameworks climate and energy policy (Foster, Elliott)
- Open Science Grid and UC3 are being used to study land use, land cover, and the impacts of climate change on agriculture and the global food supply.
- Using a **DSSAT 4.0** ("Decision Support System for Agrotechnology Transfer") crop systems model, a parallel simulation framework was implemented using **Swift**. Benchmarks of this framework have been performed on a prototype simulation campaign, measuring yield and climate impact for a single crop (maize) across the conterminous USA with daily weather data and climate model output spanning 120 years (1981-2100) and 16 different configurations of local management (fertilizer and irrigation) and cultivar choice.
- Preliminary results of parallel DSSAT run using Swift have been presented in an NSF/advisory board meeting of the CIM-EARTH project. At right, top 2 maps: Preliminary results of parallel DSSAT: maize yields across the USA with intensive nitrogen application and full irrigation; bottom 2 maps show results with no irrigation. Each model run is ~120,000 DSSAT invocations.


Median of Predicted Corn Yield 1980–2009
231 bu/A
0
Standard Deviation of Predicted Corn Yield 1980–2009
108 bu/A
0
Median of Predicted Corn Yield 1980–2009
195 bu/A
0
Standard Deviation of Predicted Corn Yield 1980–2009
81 bu/A
0
Simulated vs. Observed yields (bu/Acre) for 60423 observations 980–2009; RMSE = 25.9985, R² = 0.602179
Sim
Obs
250
200
150
100
50

DSSAT models of corn yield.
Courtesy J. Elliott and K. Maheshwari

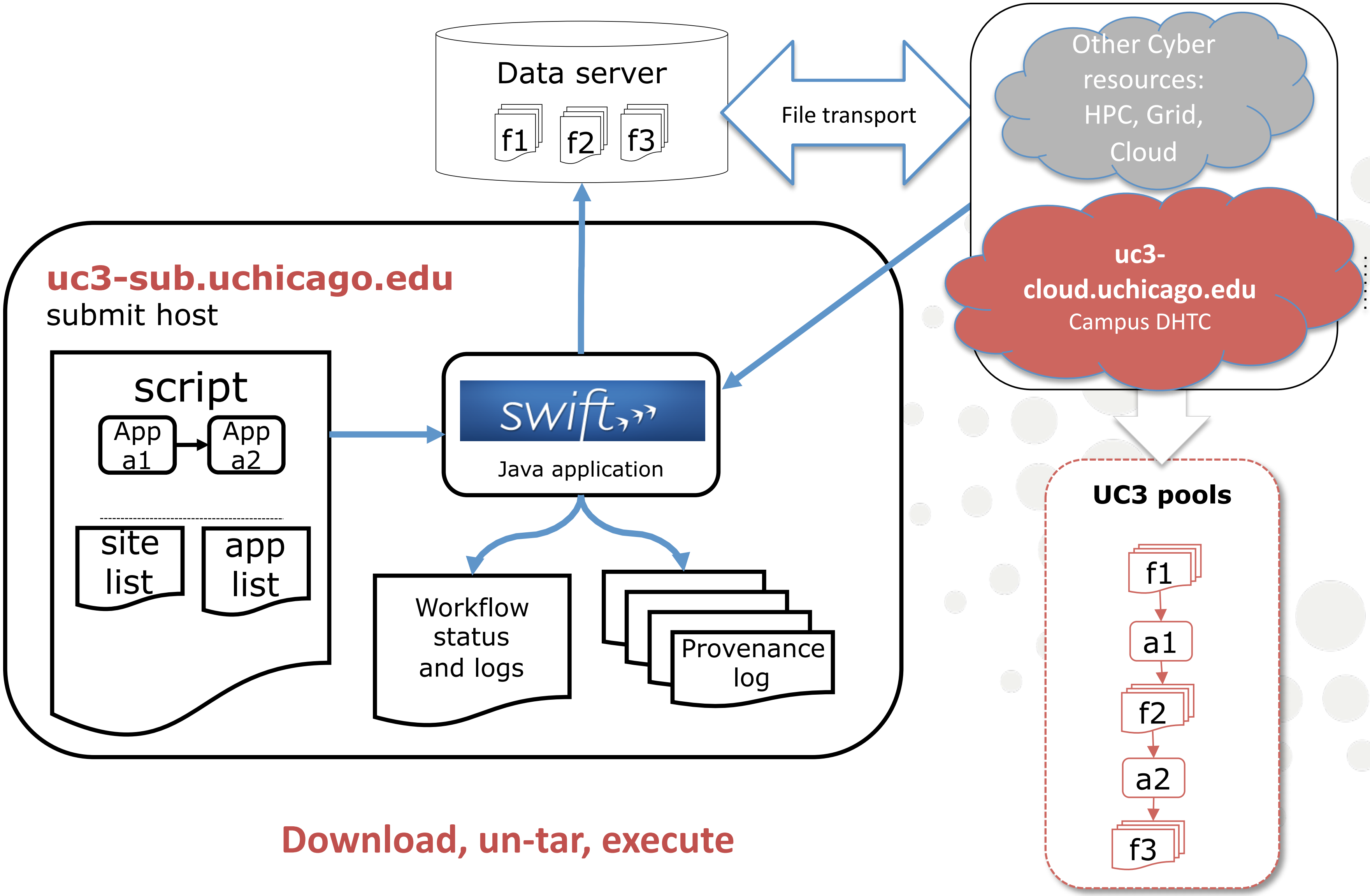
Large scale parallelization with simple loops




dssat
. . .
120,000 runs of the “DSSAT” application
analyze
Simulated vs. Observed yields (bu/Acre) for 60423 observations from 2373 counties, 1980–2009; RMSE = 25.9985, R2 = 0.602179
Sim
Obs
250
200
150
100
50
Median of Predicted Corn Yield 1980–2009
Standard Deviation of Predicted Corn Yield 1980–2009

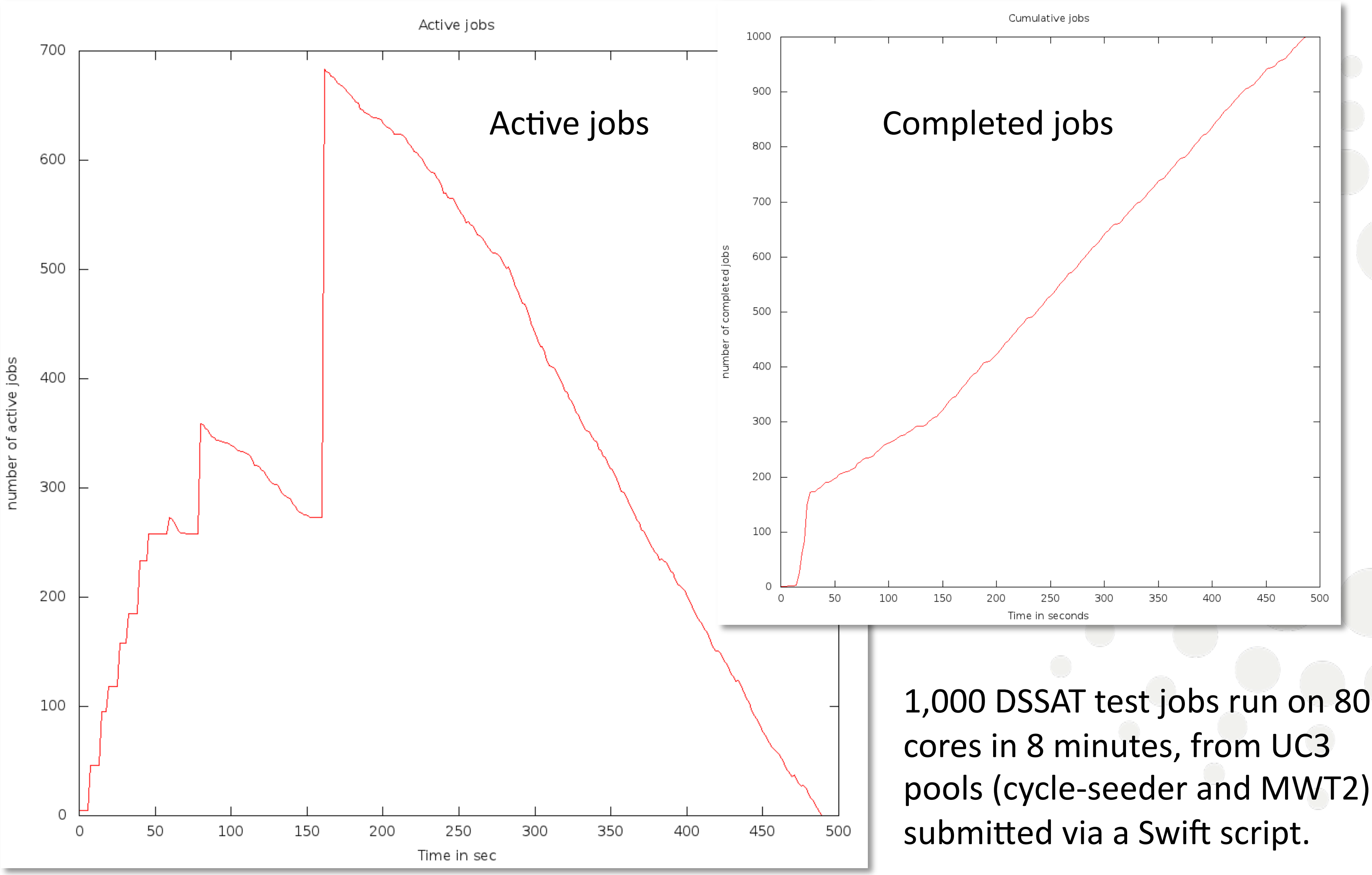
```
foreach cell in gridList
{
   models[cell] =
     dssat(cell,params);
}
result = analyze(models)
```

Implemented as Swift scripts on UC3



Download, un-tar, execute

Test run of DSSAT on UC3

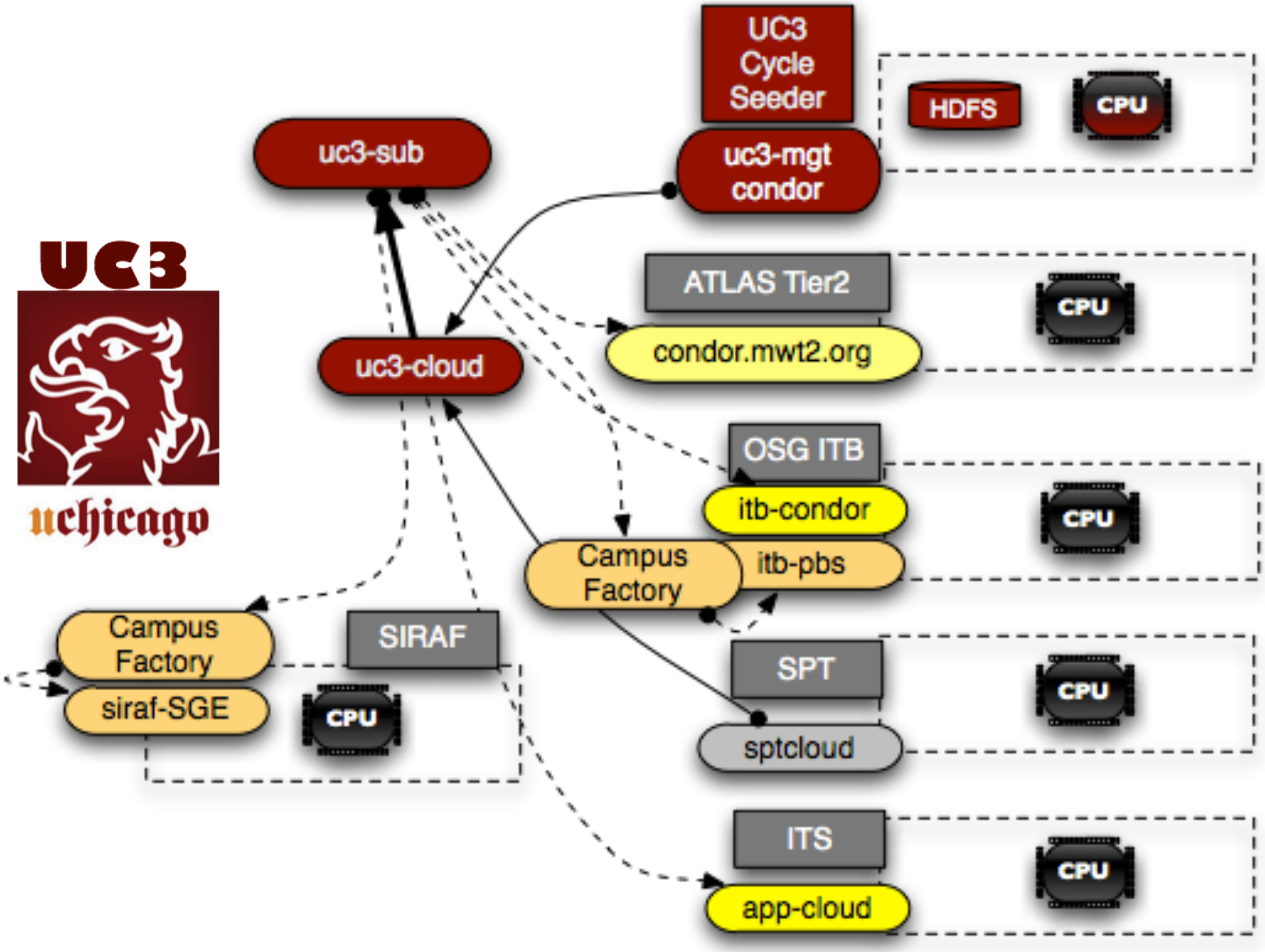


Active jobs

Completed jobs

1,000 DSSAT test jobs run on 800 cores in 8 minutes, from UC3 pools (cycle-seeder and MWT2) submitted via a Swift script.

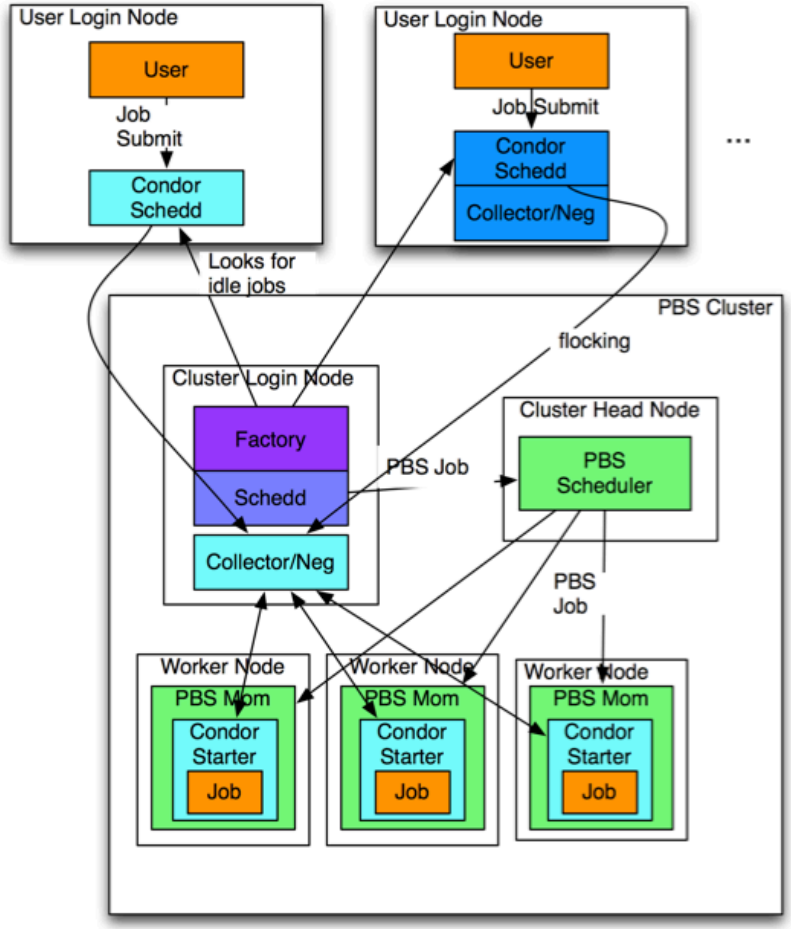
Current UC3 implementation



Lost in Space
Knob

Campus Factory to reach non-Condor pools

- Developed by OSG
- Used in UC3 for SIRAF cluster and OSG Integration testbed cluster
- Components and use:
  - Collector, Negotiator
  - Local scheduling to PBS/SGE via BLAHP
  - Condor glidein (Startd) starting as PBS jobs and reporting to the CF head node
  - UC3 flocking to the CF head node



Campus Factory issues

- Condor, BLAHP and CF configurations are separate
- Adapt to the local cluster
  - May have to run on non-standard port (other Condor)
  - Use CCB (if Firewall/NAT is present)
- Control that the file movement works as expected:
  - Set the shared directories in BLAHP configuration
  - See if LRM staging works correctly
  - Use the latest version of Condor (latest BLAHP)
- BLAHP fixes for SGE (bug submitted)
  - Not propagating the Cell (sge_cell) but using “default”

Running in a group account

- On some clusters the preference is to group account to simplify management
- This is done with Condor SLOT_USER
  - Single user for all slots
  - Multiple slot users (uc3usr[1..32])
    - More complex setup
    - Safer (isolation)

```
# Dedicated account per slot
SLOT1_USER = uc3
SLOT2_USER = uc3
SLOT3_USER = uc3
SLOT4_USER = uc3
SLOT5_USER = uc3
SLOT6_USER = uc3
…
SLOT21_USER = uc3
SLOT22_USER = uc3
SLOT23_USER = uc3
SLOT24_USER = uc3
SLOT25_USER = uc3
SLOT26_USER = uc3
SLOT27_USER = uc3
SLOT28_USER = uc3
SLOT29_USER = uc3
SLOT30_USER = uc3
SLOT31_USER = uc3
SLOT32_USER = uc3
```

Other technical issues

- Firewalls – even inter-campus – options:
  - Use CCB
  - Use shared port
  - Add the host to the ALLOW_WRITE list (if not standard port or with SOAP expression)
- GSI Authentication as first option
  - Ran into an issue where Condor doesn't failover as expected for clusters with multiple authentication systems

Special applications

- Mathematica
  - Installed a license manager
  - Installed and advertised on some nodes
  - Available to Condor jobs

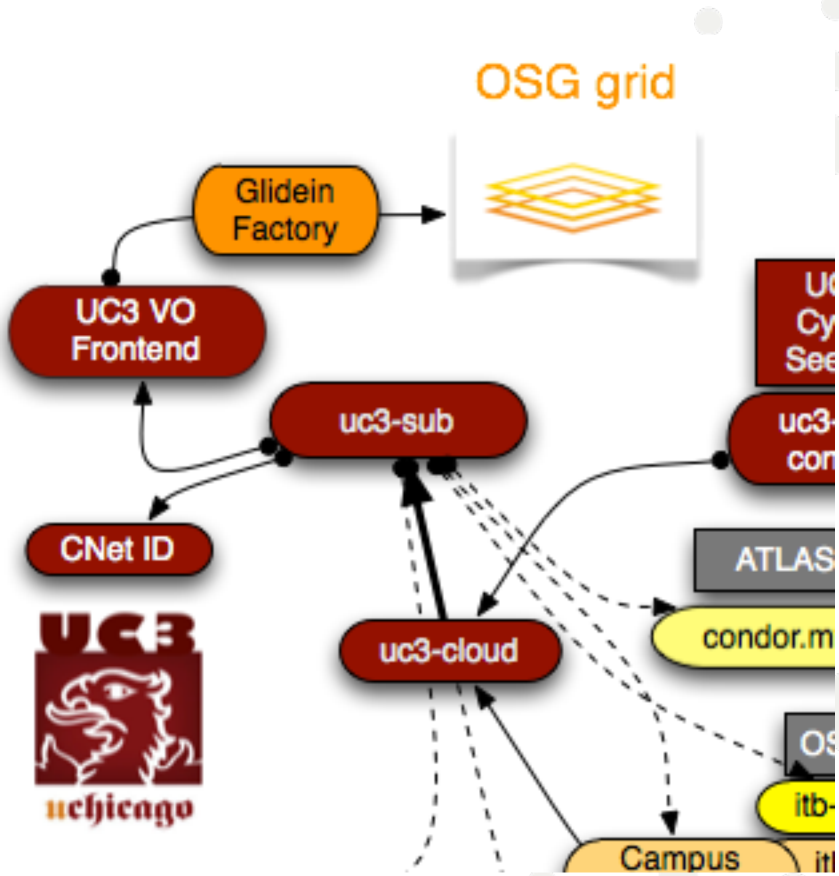
    ```
    requirements = (HAS_MATHEMATICA =?= True)
    ```
- Follow the example of other Condor pools for Matlab and R

Future work

- Job routing
  - Provide a recipe for selective user mapping using Condor mapfile
  - Evaluate/use Condor SSH submission to PBS/SGE/Condor (BOSCO)
  - Compare Condor flocking vs Condor-C vs Condor to Condor via BOSCO (BLAHP/SSH) vs rcondor (J.Dost talk)
- Identity management
  - Integration with University's LDAP system
- Data and software access
  - Flexible access to UC3 HDFS data staging via Parrot and Xrootd
  - Utilize solutions for ease of software delivery to target compute sites (e.g. CVMFS)

Future work, cont

- Off campus opportunistic overflow
  - UC3 collective VO established in OSG
  - Submission to remote sites on OSG via GlideinWMS
  - Explore InCommon for seamless local-to-grid ID management



Thank you!

https://wiki.uchicago.edu/display/uc3/UC3+Home

UC3

uchicago