Lessons Learned Developing the Computational Chemistry Grid Cyberinfrastructure

Rion Dooley, **Kent Milfeld, Chona Guiang,**

Sudhakar Pamidighantam, Gabrielle Allen

Presentation Outline

- Computational Chemistry Grid (CCG) Overview
- Why GridChem?
- Technological challenges
- Laignappe

CCG Overview

- Computational Chemistry Grid (CCG)
- 3-year NSF-funded project
- 5 sites:

NCSA

TACC

CCS
Center for Computational Sciences

OSC

CCG Overview

- CCG Resources

System (Site)	Procs Avail	Total CPU Hours/Year
HP Intel Cluster (OSC)	**12**	**100,000**
Intel Cluster (OSC)	**96**	**840,000**
Intel Cluster (UKy)	**96**	**840,000**
HP Integrity Superdome	**33**	**290,000**
Intel Cluster (NCSA)	**64**	**560,000**
Intel Cluster (NCSA)	**384**	**N/A**
Intel Cluster (NCSA)	**2688**	**N/A**
SGI Origin2000 (NCSA)	**128**	**1,000,000**
Intel Cluster (LSU)	**1024**	**1,000,000**
IBM Power4 (TACC)	**16**	**40,000**
Total	**7808**	**> 4,670,000**

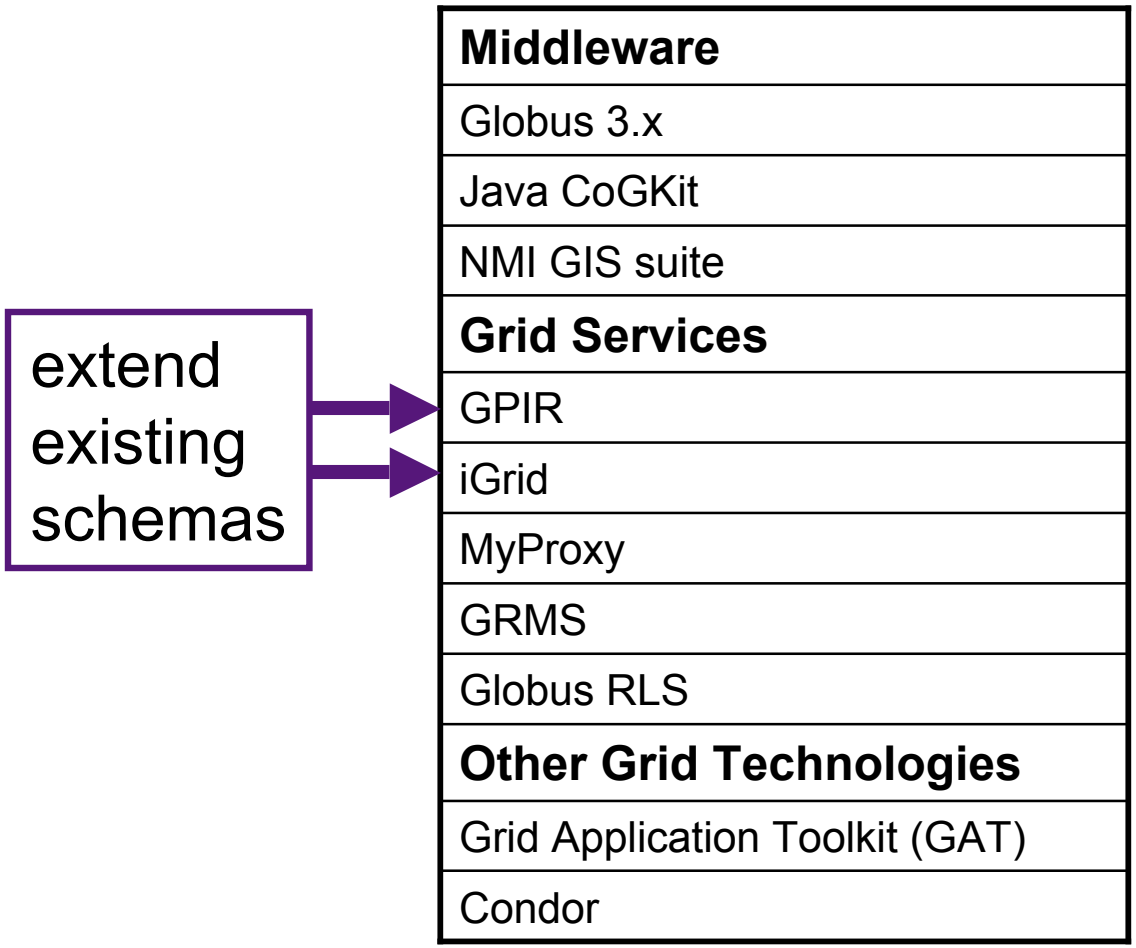
CCG Overview

- Goal is to develop “cyberinfrastructure” for computational chemistry community
  - Middleware
  - Grid services
  - End-user applications
  - Client application, **GridChem**
- NMI funded, which means

  integration not *development*

CCG Overview

- Leverage and extend existing technology



GridChem

- "The goal of GridChem is to create a powerful and useful tool for the computational chemistry community that allows users to easily submit, monitor, and manage their jobs using a large set of existing computational chemistry applications on a broad set of resources."

  --http://www.gridchem.org

GridChem

- Lightweight client relies on CCG infrastructure for grid functionality
- Currently supports Gaussian03, GAMESS, and NWchem
- User driven features:
  - molecular editor
  - GUI for input file generation
  - output file parsers
  - Grid file browser
  - multiple authentication methods**
- First release April 2005 (https://www.gridchem.org/)

Tech Challenges

- Architectural Design
- Security
- Accounting
- Information Provisioning
- Resource Management

Tech Challenges

- Architectural Design
  - "Production v. Principles"
  - NSF timeline required production in first 6 months
  - Grid architecture looks like it will take ~18 months to deploy, test, evaluate, and adjust
  - Rolling out architecture in phases

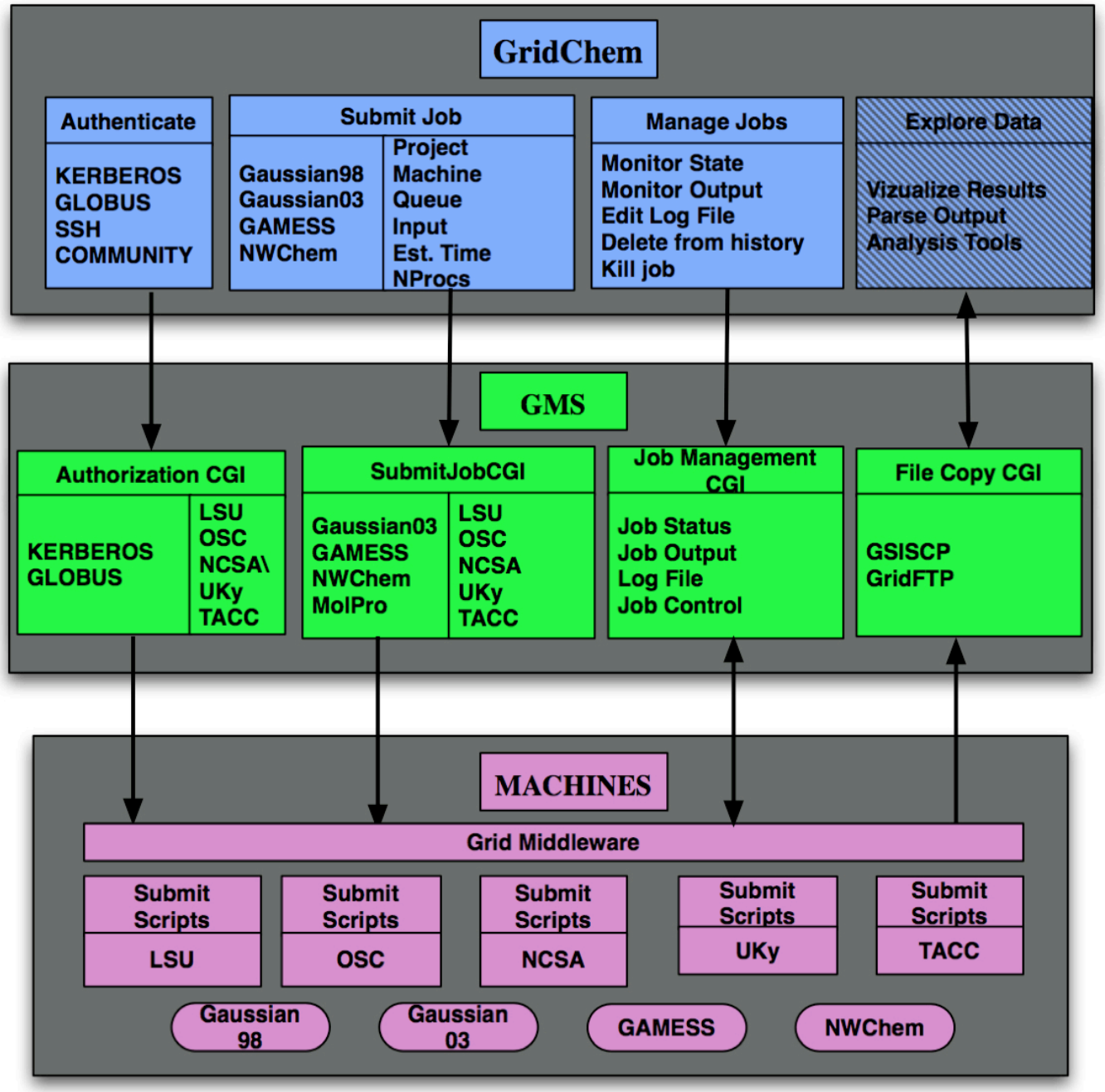


Figure 1: Current CCG Architecture

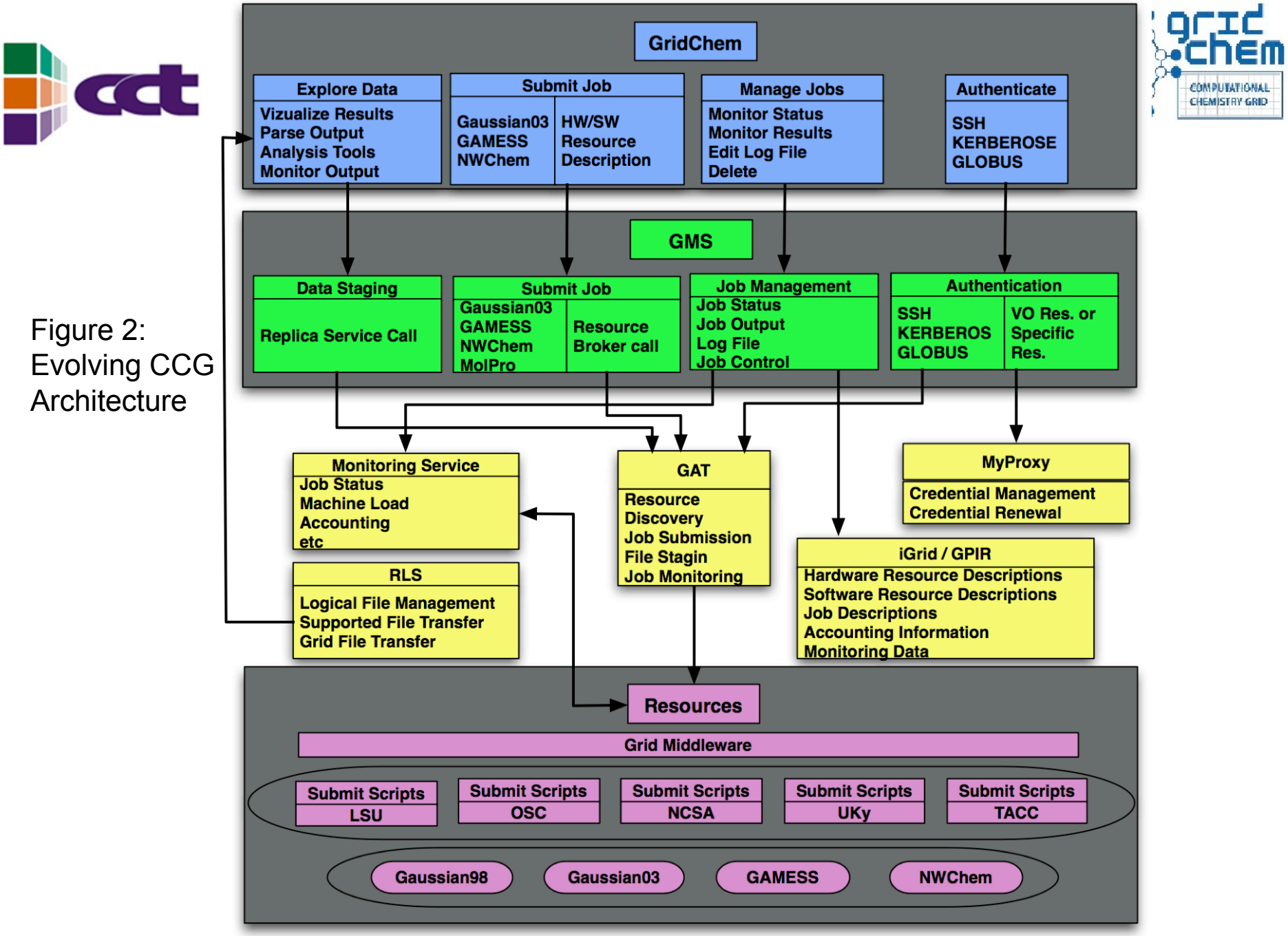


Figure 2: Evolving CCG Architecture

Tech Challenges

- Security
  - Tradeoff between standard grid mechanisms and wide community acceptance
    - Support GSI, Kerberos, SSH
    - Single sign-on
    - Use MyProxy to get user's credentials
    - Support for credential creation coming soon!!

Tech Challenges

- Security (cont.)
  - Need to incorporate the concept of a "community user"
  - Intimately tied to accounting
  - Certificates must be manually handled through our own middleware
  - Lack of credential renewal and management would be great additions to a grid API like GAT, CoGKit, or SAGA

Tech Challenges

- Accounting
  - Supporting general grid users and our "community user" creates havoc!!
  - How to track user jobs submitted outside of GridChem?
    - Dependent on advanced information provisioning
    - Real performance considerations
    - Necessity or nicety?
  - Require user registration
  - Only track GridChem usage

Tech Challenges

- Accounting (cont.)

  Q: From what perspective is the data meaningful?

  A: User perspective

  Q: From what perspective is the information published?

  A: Resource perspective

Tech Challenges

- Accounting (cont.)
  - Need API flexible enough to support these different perspectives.
  - GAT is very close to doing this.
  - Current grid API's are very good at tracking jobs submitted through their API

Tech Challenges

- Information Provisioning
  - Information is our most important piece of the puzzle
  - Existing providers are very good at pulling static and queue information on selected resources
  - Existing schemas are very good for providing resource-centric views of data

Tech Challenges

- Information Provisioning (cont.)
  - Existing API's don't support direct information discovery
  - Existing schemas missing support for information that we need to model a user's VO
  - Existing providers are missing information (job, queue, history, software, user) on most platforms

Tech Challenges

- Information Provisioning (cont.)
  - Wrote our own providers (JAMMS) (http://www.gridchem.org/consult/jobmon/omon.php)
  - Work with GPIR and iGrid people to extend their schemas to meet our needs

Tech Challenges

- Resource Management
  - Not solving the general case!!
    - Know the applications
    - Fixed set of resources
    - Fresh information on all resources
    - Have history on the users and a finite set of job classes
  - Don't need a meta scheduler, but willing to use one if it *makes life easier.*

Tech Challenges

- Resource Management (cont.)
  - GAT API to submit jobs using
    - GRMS
    - Condor
    - GRAM
    - etc
  - Relying on underlying monitoring systems to provide user-centric data

  Q: How do I track a user's jobs across their VO when they submit jobs externally to the API?

  Q: Where do I get job-specific information on a jobs submitted externally to the API?

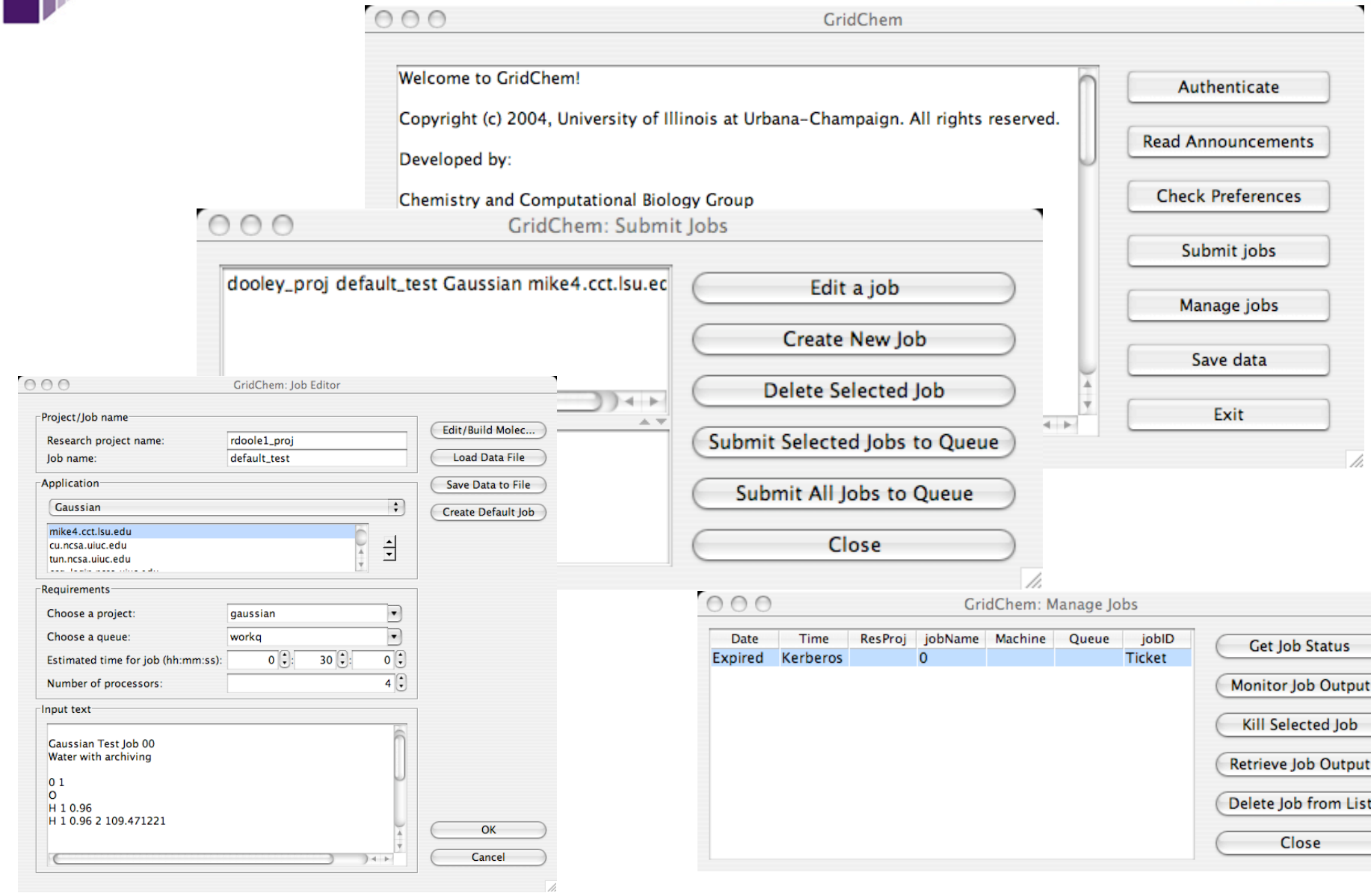
Summary

- We are in production not maintenance; Infrastructure development is ongoing.
- Rolling out grid architecture components (accounting, monitoring, submission, file management, etc.) over next 8 months
- Next GridChem release July 11, 2005

Links

- GridChem website: http://www.gridchem.org
- GAT: http://www.gridlab.org/WorkPackages/wp-1/
- GPIR: http://www.gridport.net/
- CogKit: www.cogkit.org
- Triana: http://www.trianacode.org/
- GRMS: http://www.gridlab.org/WorkPackages/wp-9/
- Condor: http://www.cs.wisc.edu/condor/

GridChem


GridChem
Welcome to GridChem!
Copyright (c) 2004, University of Illinois at Urbana-Champaign. All rights reserved.
Developed by:
Chemistry and Computational Biology Group
Authenticate
Read Announcements
Check Preferences
Submit jobs
Manage jobs
Save data
Exit
GridChem: Submit Jobs
dooley_proj default_test Gaussian mike4.cct.lsu.ec
Edit a job
Create New Job
Delete Selected Job
Submit Selected Jobs to Queue
Submit All Jobs to Queue
Close
GridChem: Job Editor
Project/Job name
Research project name:
rdoole1_proj
Job name:
default_test
Application
Gaussian
mike4.cct.lsu.edu
cu.ncsa.uiuc.edu
tun.ncsa.uiuc.edu
Requirements
Choose a project:
gaussian
Choose a queue:
workq
Estimated time for job (hh:mm:ss):
0
30
0
Number of processors:
4
Input text
Gaussian Test Job 00
Water with archiving
0 1
O
H 1 0.96
H 1 0.96 2 109.471221
Edit/Build Molec...
Load Data File
Save Data to File
Create Default Job
OK
Cancel
GridChem: Manage Jobs
Date
Time
ResProj
jobName
Machine
Queue
jobID
Expired
Kerberos
0
Ticket
Get Job Status
Monitor Job Output
Kill Selected Job
Retrieve Job Output
Delete Job from List
Close