Cyberinfrastructure & Grid Computing in New York State

Russ Miller

Cyberinfrastructure Lab, SUNY-Buffalo

Hauptman-Woodward Med Res Inst

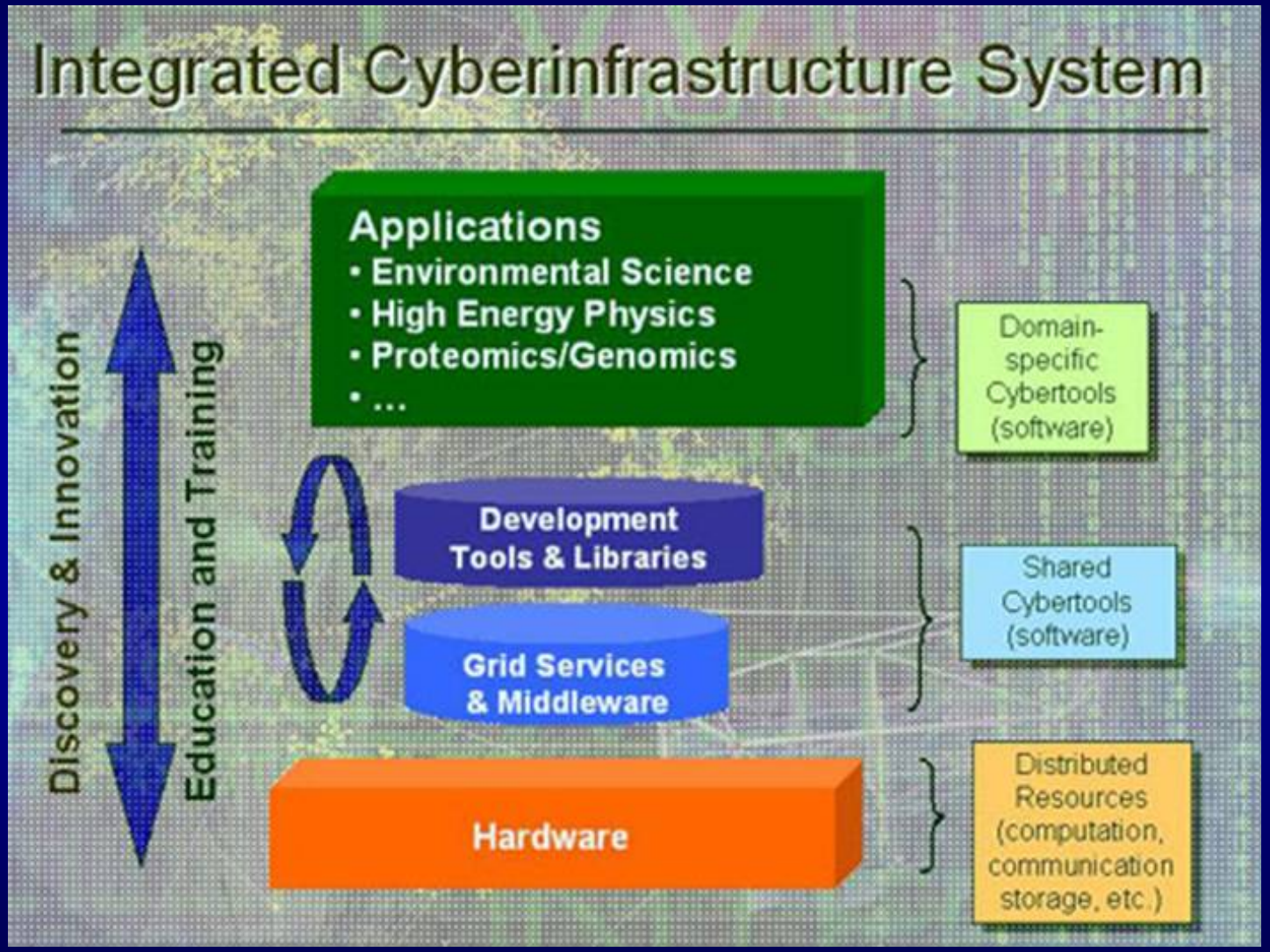
NSF, NIH, DOE, NIMA, NYS, HP

www.cse.buffalo.edu/faculty/miller/CI/

Cyberinfrastructure

- **Generic: transparent and ubiquitous application of technologies central to contemporary engineering and science**
- **NSF: "comprehensive phenomenon that involves creation, dissemination, preservation, and application of knowledge"**
- **Foster & Kesselman: "a domain-independent computational infrastructure designed to support science."**
- **NSF Cyberinfrastructure (OCI)**
  - **HPC Hardware and Software**
  - **Data Collections**
  - **Science Gateways/Virtual Organizations**
  - **Support of Next Generation Observing Systems**

NSF Integrated Cyberinfrastructure



NSF Director Arden L. Bement: "leadership in cyberinfrastructure may determine America's continued ability to innovate – and thus our ability to compete successfully in the global arena."

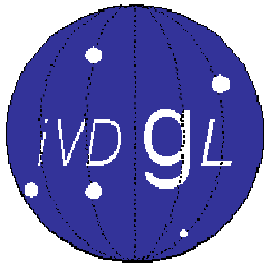
Cyberinfrastructure in Academia

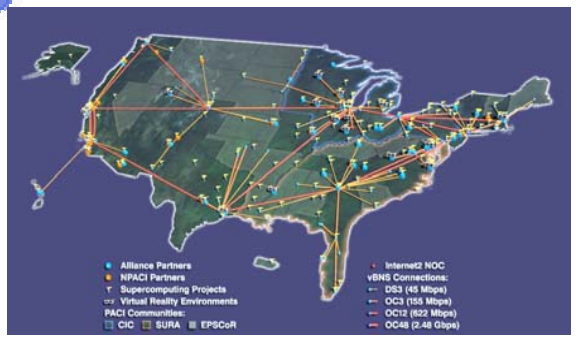
- **Empower students to compete in knowledge-based economy**
- **Embrace digital data-driven society**
- **Accelerate discovery and comprehension**
- **Enhance virtual organizations**
- **Provide increased education, outreach, and training**
- **Enhance and expand relationships between academia and the corporate world**

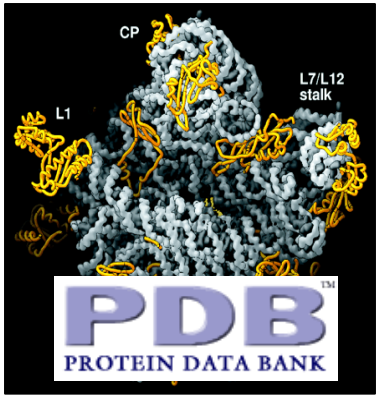
Critical Academic Initiatives

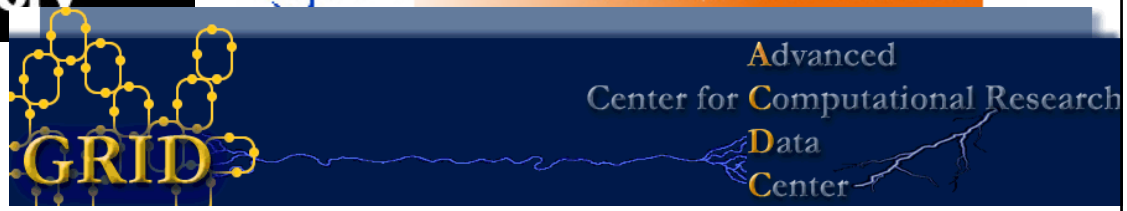
- **Create links between enabling technologists and disciplinary users**
- **Improve efficiency of knowledge-driven applications in myriad disciplines**
  - **New Techniques**
  - **New Algorithms**
  - **New Interactions (people & systems)**
- **Support HPC infrastructure, research, and applications**
- **Deliver high-end cyberinfrastructure to enable efficient**
  - **Collection of data**
  - **Management/Organization of data**
  - **Distribution of data**
  - **Analysis of data**
  - **Visualization of data**

Grid Computing

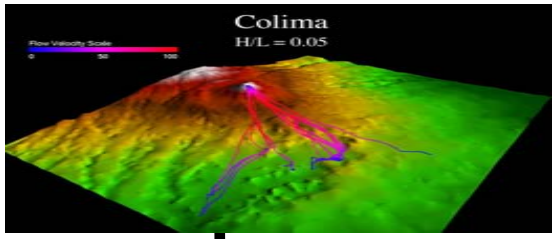




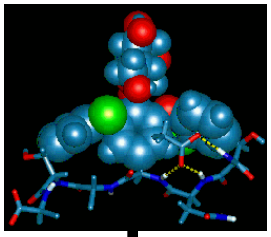




Grid Computing Overview



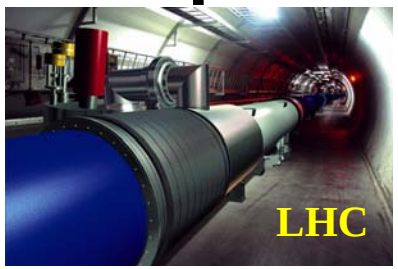
Data Acquisition

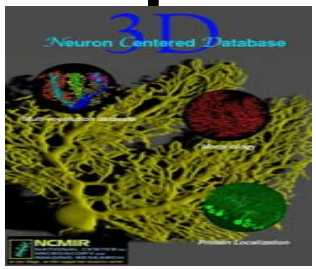


Advanced Visualization

Analysis

Imaging Instruments





Large-Scale Databases

- Coordinate Computing Resources, People, Instruments in Dynamic Geographically-Distributed Multi-Institutional Environment
- Treat Computing Resources like Commodities
  - Compute cycles, data storage, instruments
  - Human communication environments
- No Central Control; No Trust

Major Grid Initiatives

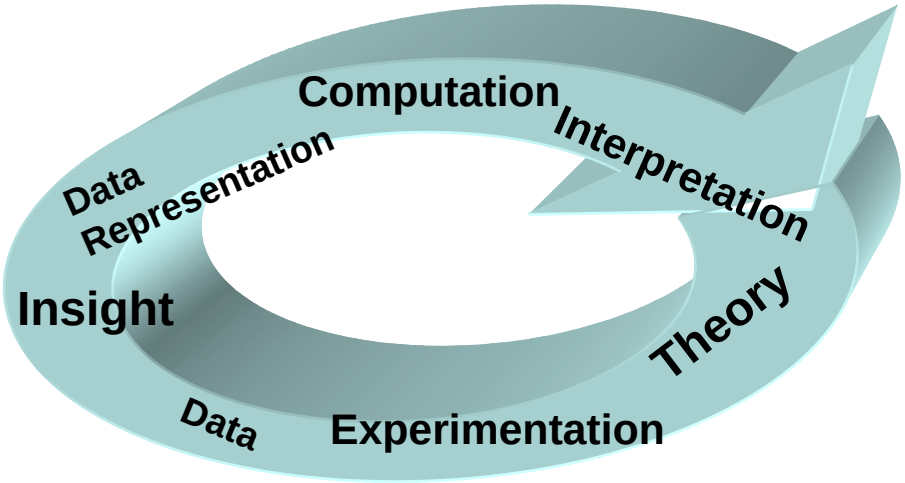
- TeraGrid (NSF)
  - Integrates High-End Resources
  - High-Performance (Dedicated) Networks
  - 9 Sites; 250TF & 30PB
  - 100+ Databases Available
- OSG (DOE, NSF)
  - High-Throughput Distributed Facility
  - Open & Heterogeneous
  - Biology, Computer Science, Astrophysics, LHC
  - 57 Compute Sites; 11 Storage Sites;
  - 10K CPUS; 6PB
- EGEE: Enabling Grids for E-SciencE (European Commision)
  - Initial Focus on CERN (5PB of Data/Year)
    - High-Energy Physics and Life Sciences
  - Expanded Focus Includes Virtually All Scientific Domains
  - 200 Institutions; 40 Countries
  - 20K+ CPUs; 5PB; 25,000 jobs per day!

A Depiction of Cyber-enabled Discover (NSF)

*How will **CDI** (**C**yber-Enabled **D**iscover and **I**nnovation) change the way we will think about and practice science?*

"The Old Way: Classical Science"

Where, for the most part, classical science is practiced in a constrained domain.

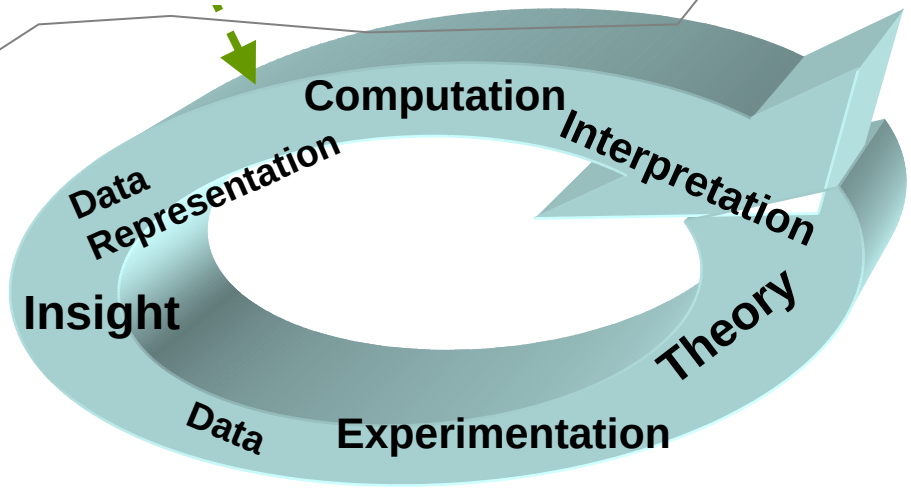


A Depiction of Cyber-enabled Discover (NSF)

Example:
Discovering causality of a phenomenon, such as a disease or weather pattern, by mining massive Data stores.

"The New Way: Cyber-Enabled Science"

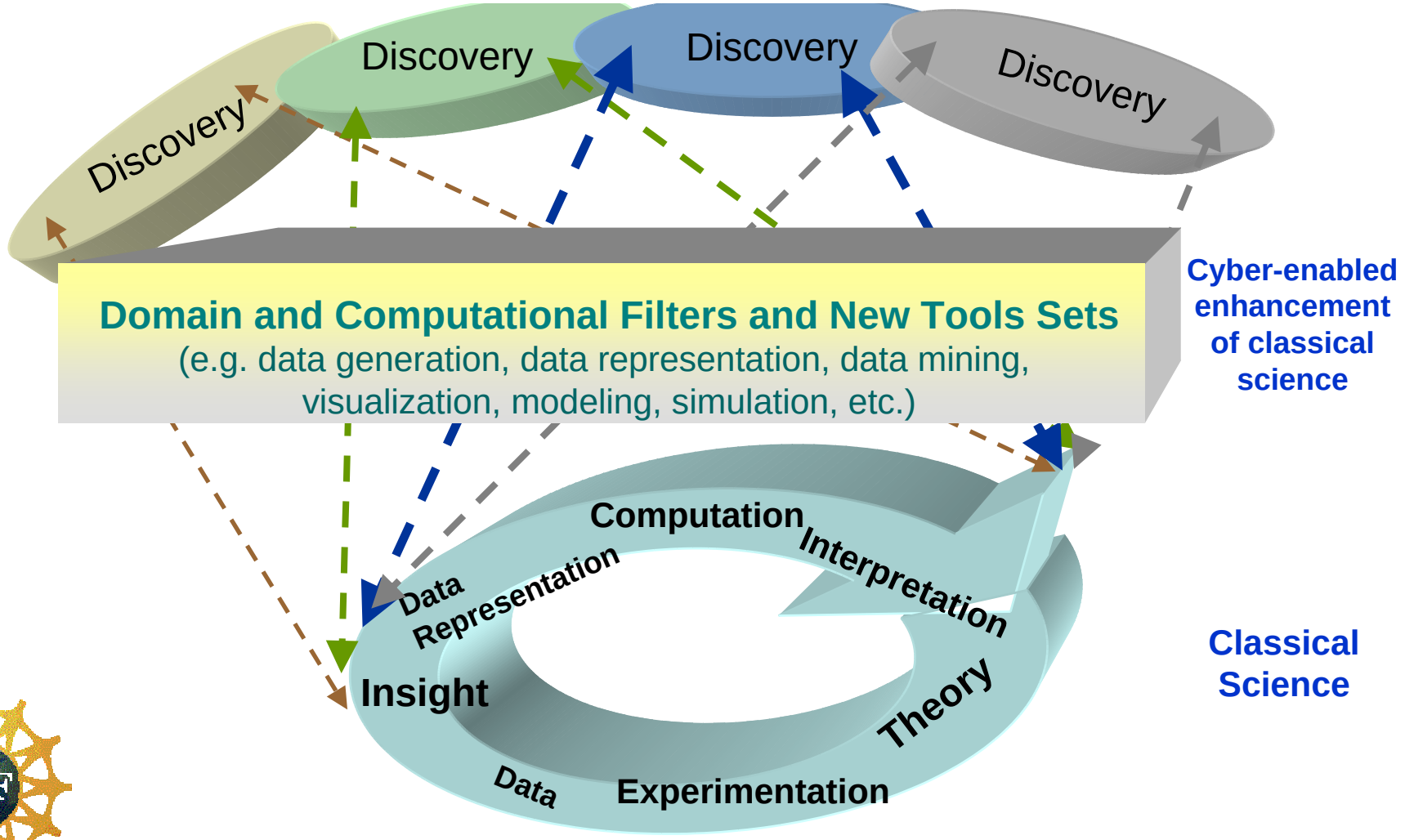
For biologists, meteorologists, computational scientists, and others, a "cyber-layer" between their problem space and the way they traditionally engaged in science and may provide new sources of data, new ways of visualizing the data, new tools for extracting meaning from the data, new ways to collaborate with other scientists, and other enabling mechanisms and computational tools for discovery in new areas of science.



Classical Science

A Depiction of Cyber-enabled Discover (NSF)

Ubiquitous, cyber-enabled discovery across many fields.

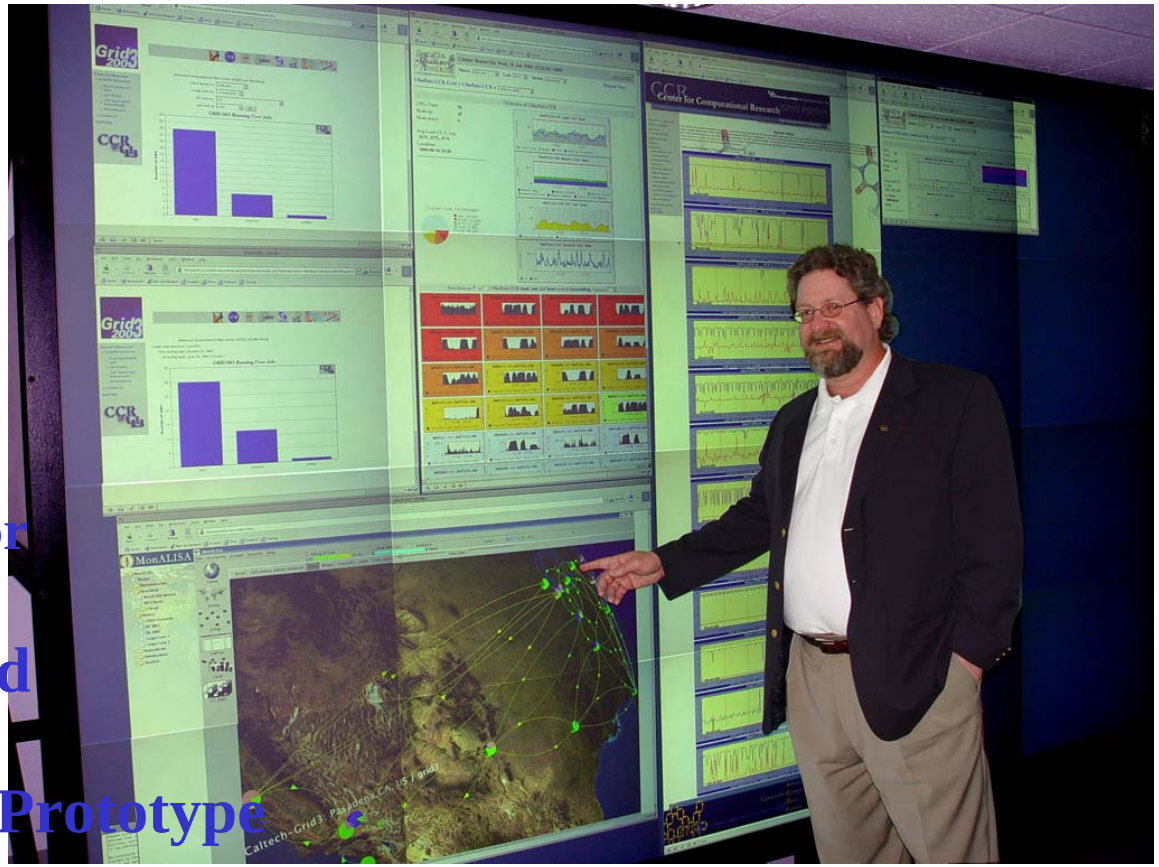


Miller's Cyberinfrastructure Lab

- CI sits at core of modern simulation & modeling
- CI allows for new methods of investigation to address previously unsolvable problems
- Focus on development of
  - Algorithms
  - Portals
  - Interfaces
  - Middleware
- Free end-users to do disciplinary work
- Funding (2001-pres): NSF ITR, NSF CRI, NSF MRI, NYS, Fed

CI Lab Collaborations

- High-Performance Networking Infrastructure
- Grid3+ Collaboration
- iVDGL Member
  - Only External Member
- Open Science Grid
  - GRASE VO
- NYSGrid.org
  - NYS CI Initiative
  - Fndg Executive Director
  - Various WGs
- Grid-Lite: Campus Grid
  - HP Labs Collaboration
- Innovative Laboratory Prototype
  - Dell Collaboration

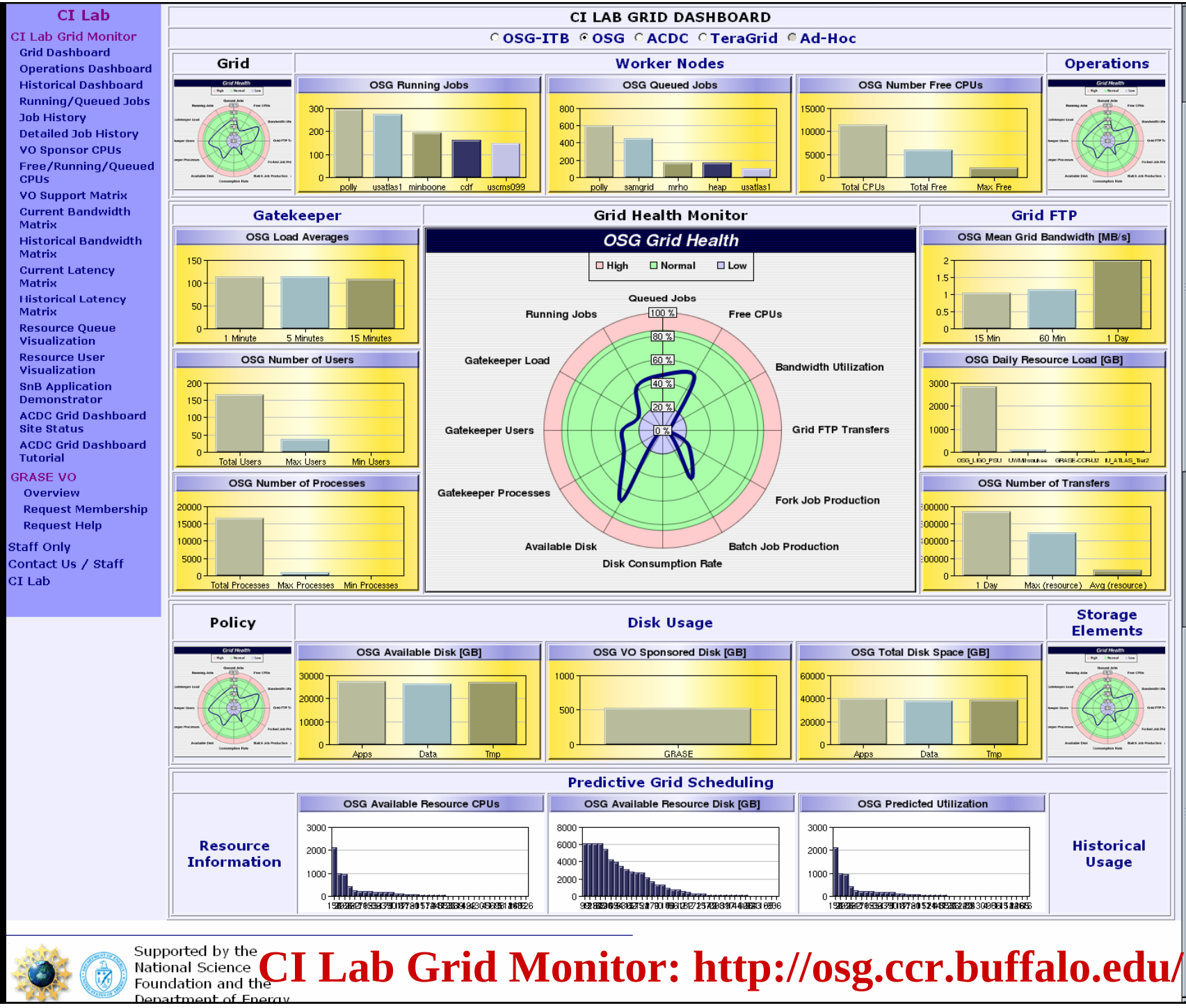


Evolution of CI Lab Projects

- **ACDC-Grid**
  - **Experimental Grid: Globus & Condor**
  - **Integrate Data & Compute, Monitor, Portal, Node Swapping, Predictive Scheduling/Resource Management**
  - **GRASE VO: Structural Biology, Groundwater Modeling, Earthquake Eng, Comp Chemistry, GIS/BioHazards**
  - **Buffalo, Buffalo State, Canisius, Hauptman-Woodward**
- **WNY Grid**
  - **Heterogeneous System: Hardware, Networking, Utilization**
  - **Buffalo, Geneseo, Hauptman-Woodward, Niagara**
- **NYS Grid**
  - **Extension to Hardened Production-Level System State-Wide**
  - **Albany, Binghamton, Buffalo, Geneseo, Canisius, Columbia, HWI, Niagara, [Cornell, NYU, RIT, Rochester, Syracuse, Marist], {Stony Brook, RPI, Iona}**
  - **VOs: Engage, GADU, GRASE, Nanohub, SDSS, USATLAS, USCMS**

NYS Grid Resources

- Albany: 8 Dual-Processor Xeon Nodes
- Binghamton: 15 Dual-Processor Xeon Nodes
- Buffalo: 1050 Dual-Processor Xeon Nodes
- Cornell: 30 Dual-Processor Xeon Nodes
- Geneseo State: Sun/AMD with 128 Compute Cores
- Hauptman-Woodward Institute: 50 Dual-Core G5 Nodes
- Marist: 9 P4 Nodes
- Niagara University: 64 Dual-Processor Xeon Nodes
- NYU: 58 Dual-Processor PowerPC Nodes
- RIT: 4 Dual-Processor Xeon Nodes
- Syracuse: 8 Dual-Processor Xeon Nodes


CI Lab
CI Lab Grid Monitor
Grid Dashboard
Operations Dashboard
Historical Dashboard
Running/Queued Jobs
Job History
Detailed Job History
VO Sponsor CPUs
Free/Running/Queued CPUs
VO Support Matrix
Current Bandwidth Matrix
Historical Bandwidth Matrix
Current Latency Matrix
Historical Latency Matrix
Resource Queue Visualization
Resource User Visualization
SnB Application Demonstrator
ACDC Grid Dashboard Site Status
ACDC Grid Dashboard Tutorial
GRASE VO
Overview
Request Membership
Request Help
Staff Only
Contact Us / Staff
CI Lab
CI LAB GRID DASHBOARD
OSG-ITB OSG ACDC TeraGrid Ad-Hoc
Grid
Worker Nodes
Operations
OSG Running Jobs
polly usatlas1 minboone cdf uscms099
OSG Queued Jobs
polly samgrid mrho heap usatlas1
OSG Number Free CPUs
Total CPUs Total Free Max Free
Gatekeeper
OSG Load Averages
1 Minute 5 Minutes 15 Minutes
OSG Number of Users
Total Users Max Users Min Users
OSG Number of Processes
Total Processes Max Processes Min Processes
Grid Health Monitor
OSG Grid Health
High Normal Low
Queued Jobs
Running Jobs
Free CPUs
Gatekeeper Load
Bandwidth Utilization
Gatekeeper Users
Grid FTP Transfers
Gatekeeper Processes
Fork Job Production
Available Disk
Batch Job Production
Disk Consumption Rate
Grid FTP
OSG Mean Grid Bandwidth [MB/s]
15 Min 60 Min 1 Day
OSG Daily Resource Load [GB]
OSG Number of Transfers
1 Day Max (resource) Avg (resource)
Policy
Disk Usage
Storage Elements
OSG Available Disk [GB]
Apps Data Tmp
OSG VO Sponsored Disk [GB]
GRASE
OSG Total Disk Space [GB]
Apps Data Tmp
Predictive Grid Scheduling
Resource Information
OSG Available Resource CPUs
OSG Available Resource Disk [GB]
OSG Predicted Utilization
Historical Usage
Supported by the National Science Foundation and the Department of Energy
CI Lab Grid Monitor: http://osg.ccr.buffalo.edu/

Predictive Scheduler

- **Build profiles based on statistical analysis of logs of past jobs**
  - **Per User/Group**
  - **Per Resource**
- **Use these profiles to predict runtimes of new jobs**
- **Make use of these predictions to determine**
  - **Resources to be utilized**
  - **Availability of Backfill**

ACDC-Grid Dynamic Resource Allocation at SC03 with Grid3

- **Small number (40) of CPUs were dedicated at night**
- **An additional 400 CPUs were dynamically allocated during the day**
- **No human intervention was required**
- **Grid applications were able to utilize the resources and surpassed the Grid3 goals**

ACDC-Grid Data Grid Functionality

- **Basic file management functions are accessible via a platform-independent web interface.**
- **User-friendly menus/interface.**
- **File Upload/Download to/from the Data Grid Portal.**
- **Simple Web-based file editor.**
- **Efficient search utility.**
- **Logical display of files (user/ group/ public).**
- **Ability to logically display files based on metadata (file name, size, modification date, etc.)**

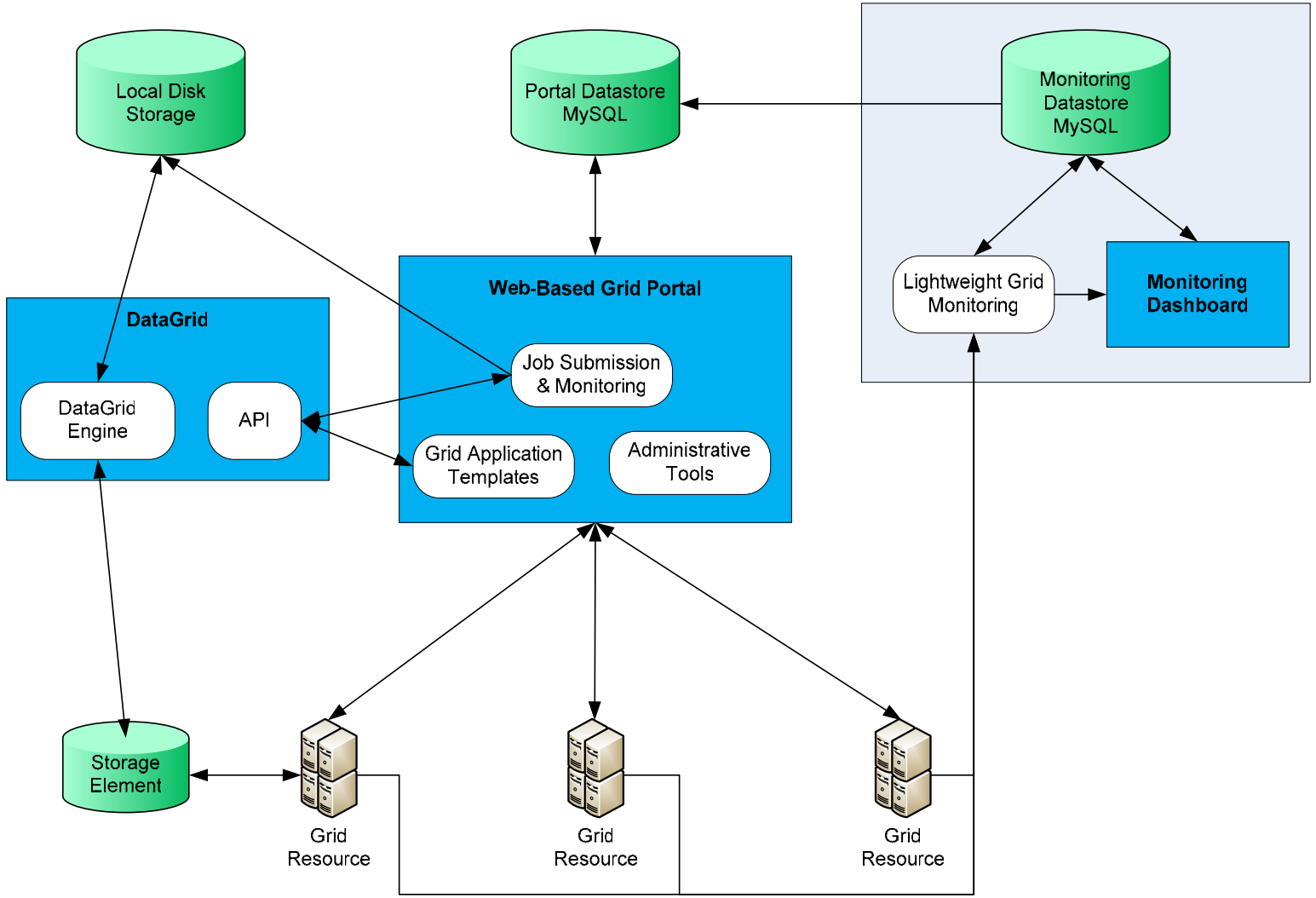
Grid-Enabling Application Templates (GATs)

- **Structural Biology**
  - ***SnB* and *BnP* for Molecular Structure Determination/Phasing**
- **Groundwater Modeling**
  - ***Ostrich*: Optimization and Parameter Estimation Tool**
  - ***POMGL*: Princeton Ocean Model Great Lakes for Hydrodynamic Circulation**
  - ***Split*: Modeling Groundwater Flow with Analytic Element Method**
- **Earthquake Engineering**
  - ***EADR*: Evolutionary Aseismic Design and Retrofit; Passive Energy Dissipation System for Designing Earthquake Resilient Structures**
- **Computational Chemistry**
  - ***Q-Chem*: Quantum Chemistry Package**
- **Geographic Information Systems & BioHazards**
  - ***Titan*: Computational Modeling of Hazardous Geophysical Mass Flows**

Grid Enabled *Shake-and-Bake* (Molecular Structure Determination)

- **Required Layered Grid Services**
  - **Grid-enabled Application Layer**
    - ***Shake–and–Bake* application**
    - **Apache web server**
    - **MySQL database**
  - **High-level Service Layer**
    - **Globus, NWS, PHP, Fortran, and C**
  - **Core Service Layer**
    - **Metacomputing Directory Service, Globus Security Interface, GRAM, GASS**
  - **Local Service Layer**
    - **Condor, MPI, PBS, Maui, WINNT, IRIX, Solaris, RedHat Linux**

NYS Grid Portal



https://grid.ccr.buffalo.edu/

Mail Home Netscape

New Tab | CCR Grid Computing Services:

Cyberinfrastructure Laboratory

Grid Portal

Dr. Russ Miller
UB Distinguished Professor of Computer Science & Engineering

Welcome to the Cyberinfrastructure Laboratory Grid Portal

The **Cyberinfrastructure Laboratory**, in conjunction with the **Center for Computational Research**, has created an integrated Data and Computational Grid. This site is devoted to a Grid Portal that provides access to applications that can be run on a variety of grids. A related site contains a **Grid Monitoring System** designed by the Cyberinfrastructure Laboratory.

Applications may be run on the Cyberinfrastructure Laboratory's **ACDC Grid**, **Western New York Grid**, and **New York State Grid**, which includes computational and data storage systems from dozens of institutions throughout the State of New York.

The applications available to the users cover a variety of disciplines, including Bioinformatics, Computational Chemistry, Crystallography and Medical Imaging, to name a few.

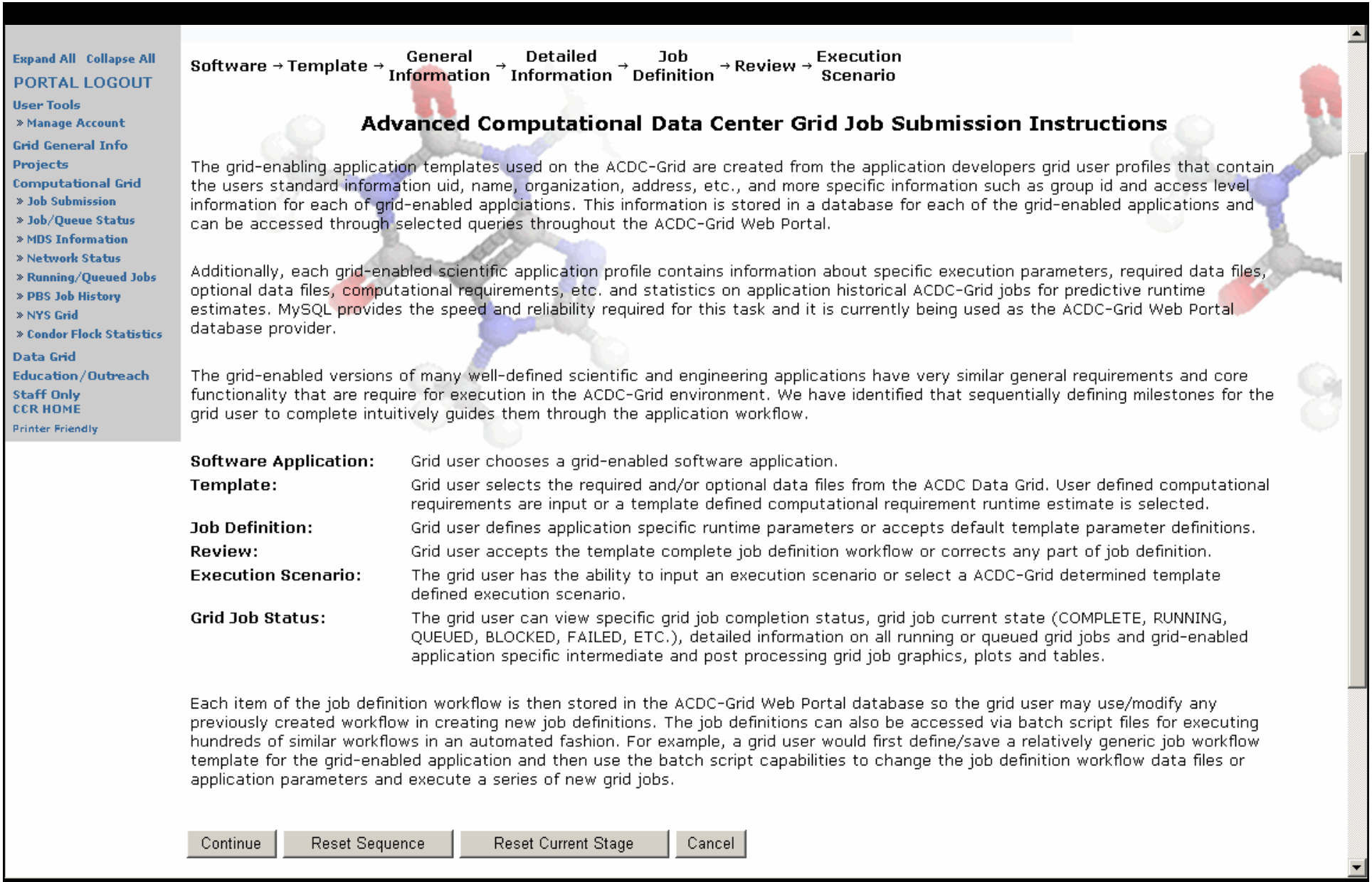
The grids developed by the CI Lab support teaching and research activities, as well as providing infrastructure that includes high-end data, computing, imaging, grid-enabled software, all of which relies on the New York State Research Network (**NYSERNet**).

This work is funded by the National Science Foundation (ITR, MRI, CRI), three program projects from The National Institutes of Health, and the Department of Energy.

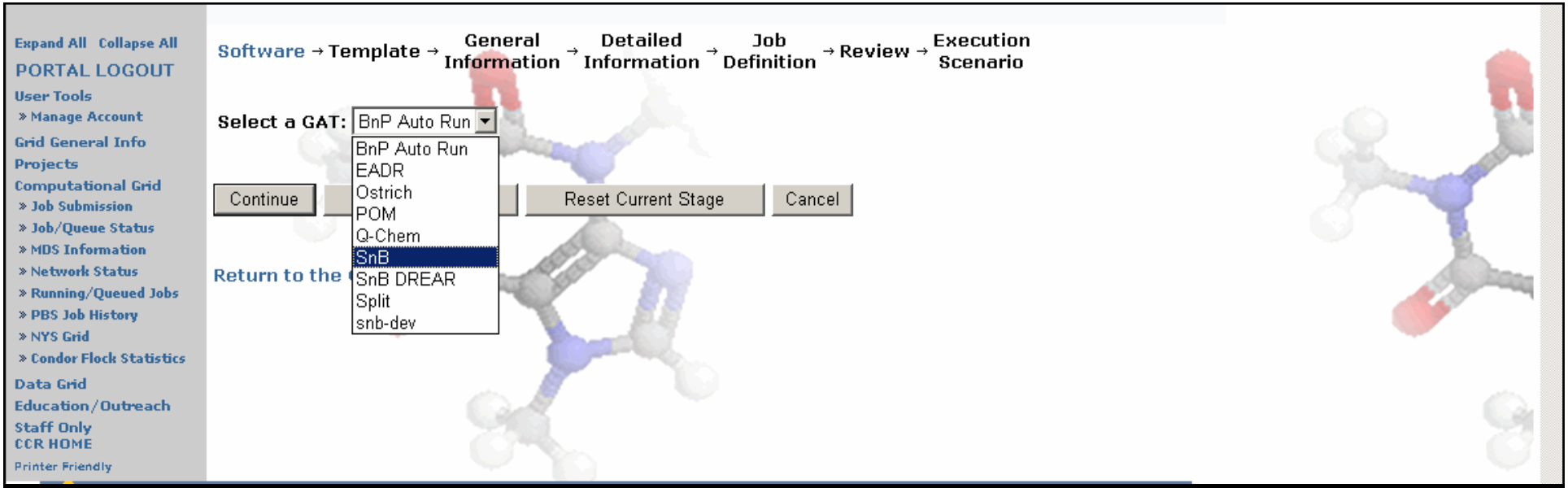
BnP

Software : BnP
Field : Protein crystal structure determination

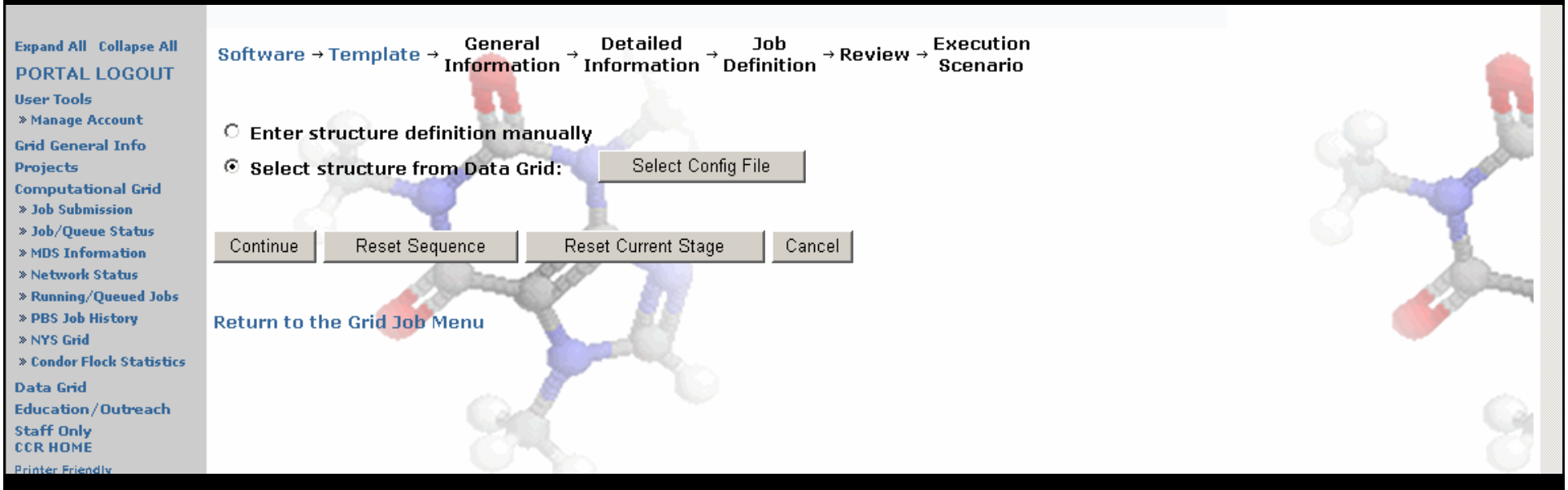
Startup Screen for CI Lab Grid Job Submission



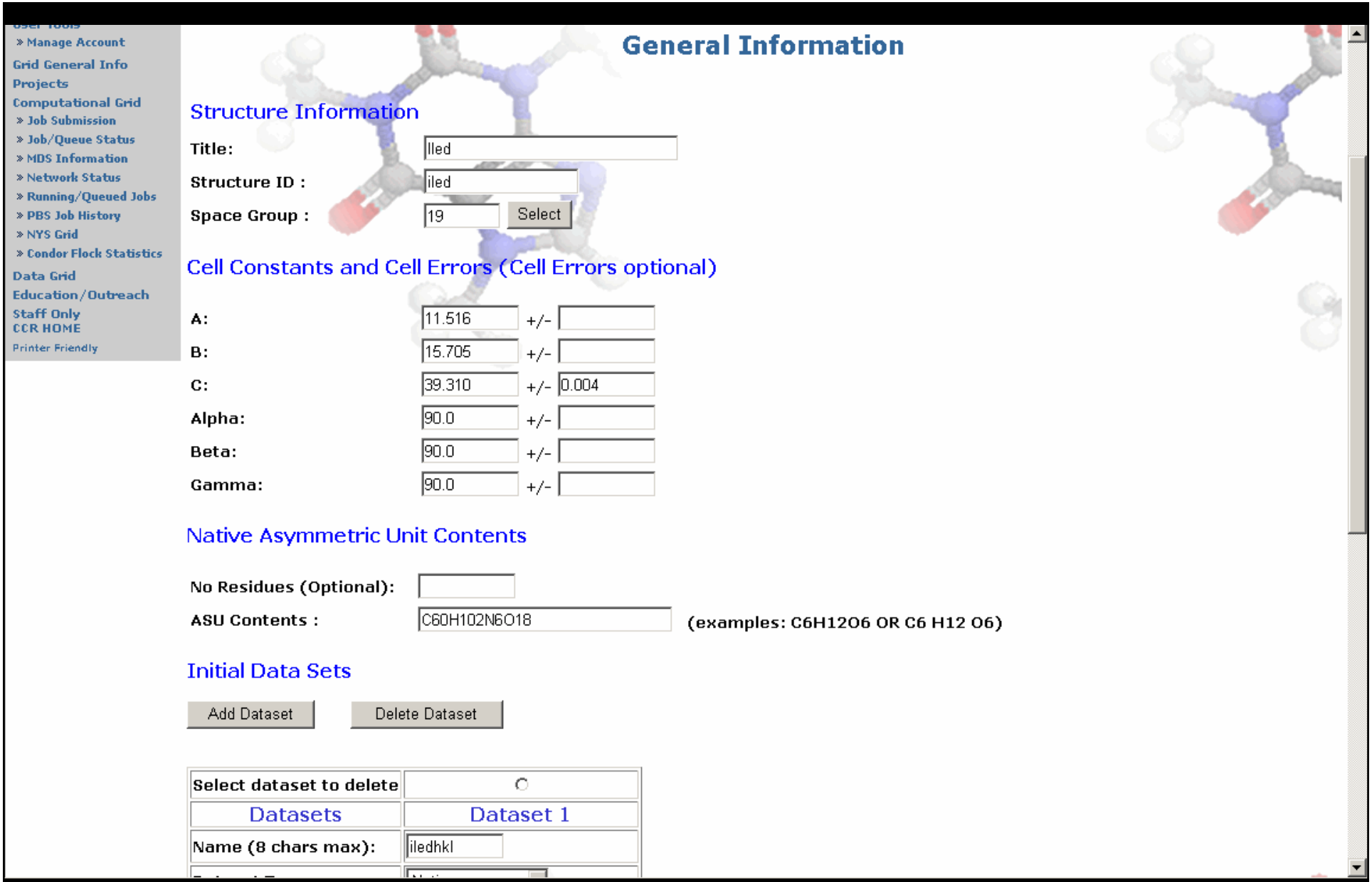
Instructions and Description for Running a Job on ACDC-Grid



Software Package Selection



Full Structure / Substructure Template Selection

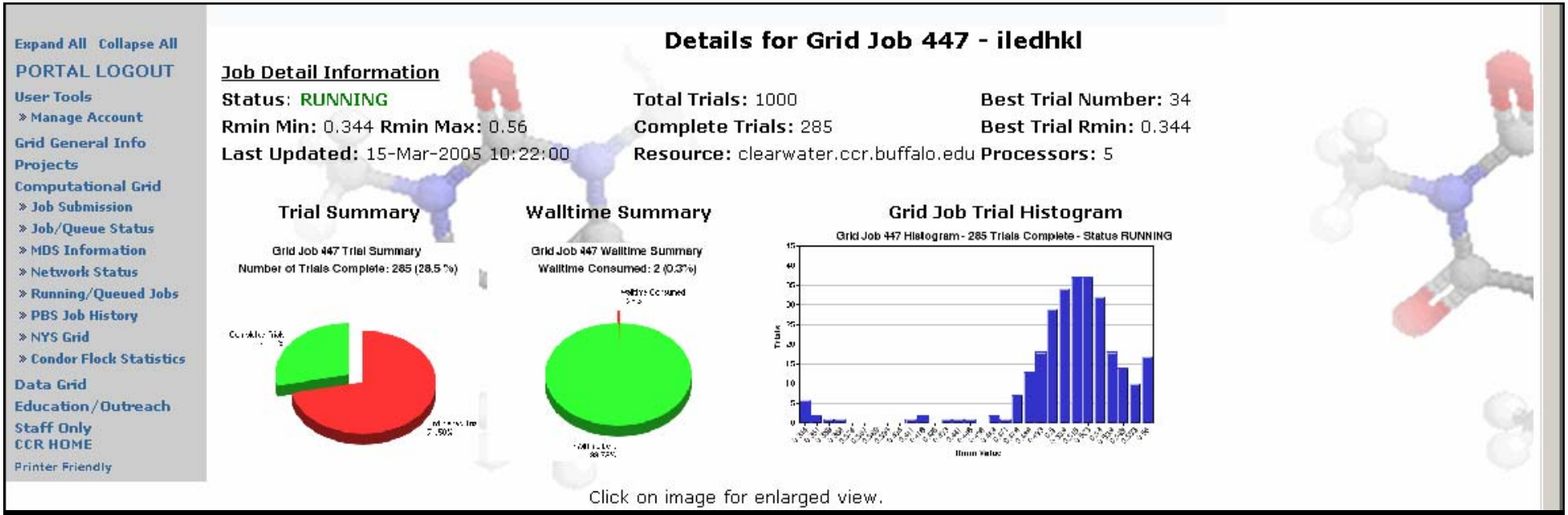


Default Parameters Based on Template

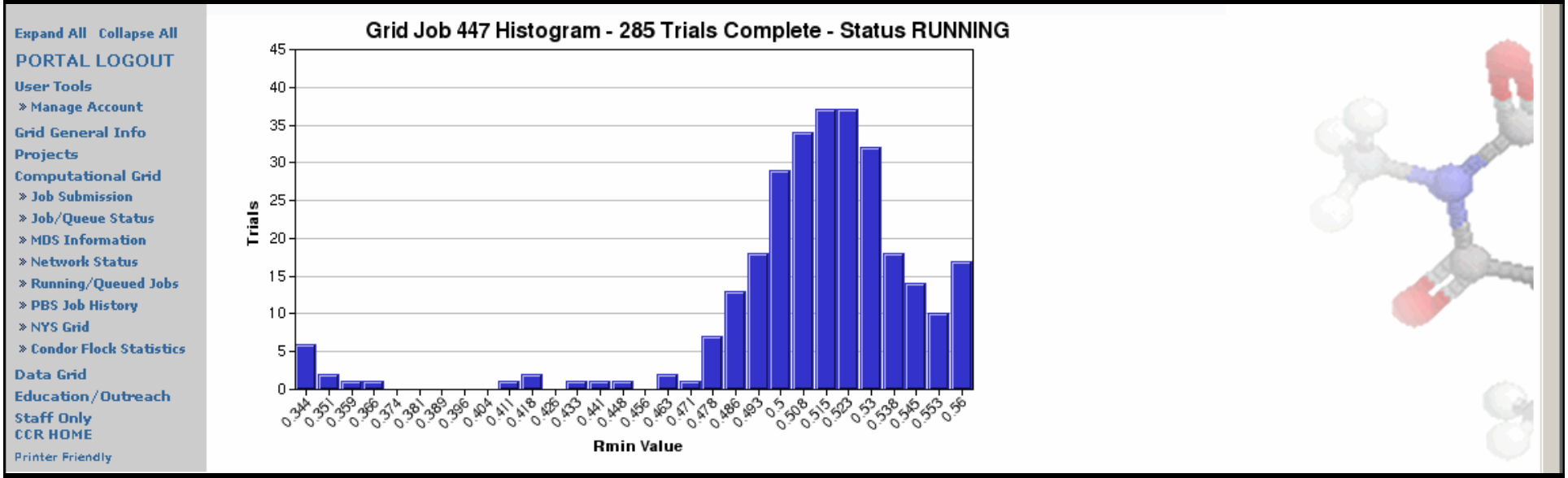
SnB Job Review

Grid Job ID:	447
Selected resource:	clearwater.ccr.buffalo.edu
Number of processors:	5
Wallclock time requested:	720
Number of triplet invariant to use:	8400
Start Phases From:	Random Atoms
Random seed (prime):	11909
Number of trials:	1000
Starting Trial:	1
Input Phase File:	Unused
Input Atom File:	Unused
Keep complete (every trial) peak file? :	Yes
Number of Shake-and-bake cycles:	20
Keep complete (every cycle) trace file? :	No
Terminate trials failing the R-Ratio test? :	No
R-Ratio cutoff:	Unused
Phase Refinement Method:	Parameter Shift(Fast)
Number of passes through phase set:	3
Phase shift:	90.0
Number of shifts:	2
Number of peaks to select:	84
Minimum interpeak distance:	3
Minimum distance between symmetry-related peaks:	3.0
Number of special position peaks to keep:	0
Fourier grid size:	0.31
Perform extra cycles with more peaks? :	No
Number of extra cycles:	Unused
Number of peaks:	Unused
Trials for E-Fourier filtering (fourier refinement)? :	None
Number of cycles:	Unused
Number of peaks:	Unused
Minimum |E|:	Unused

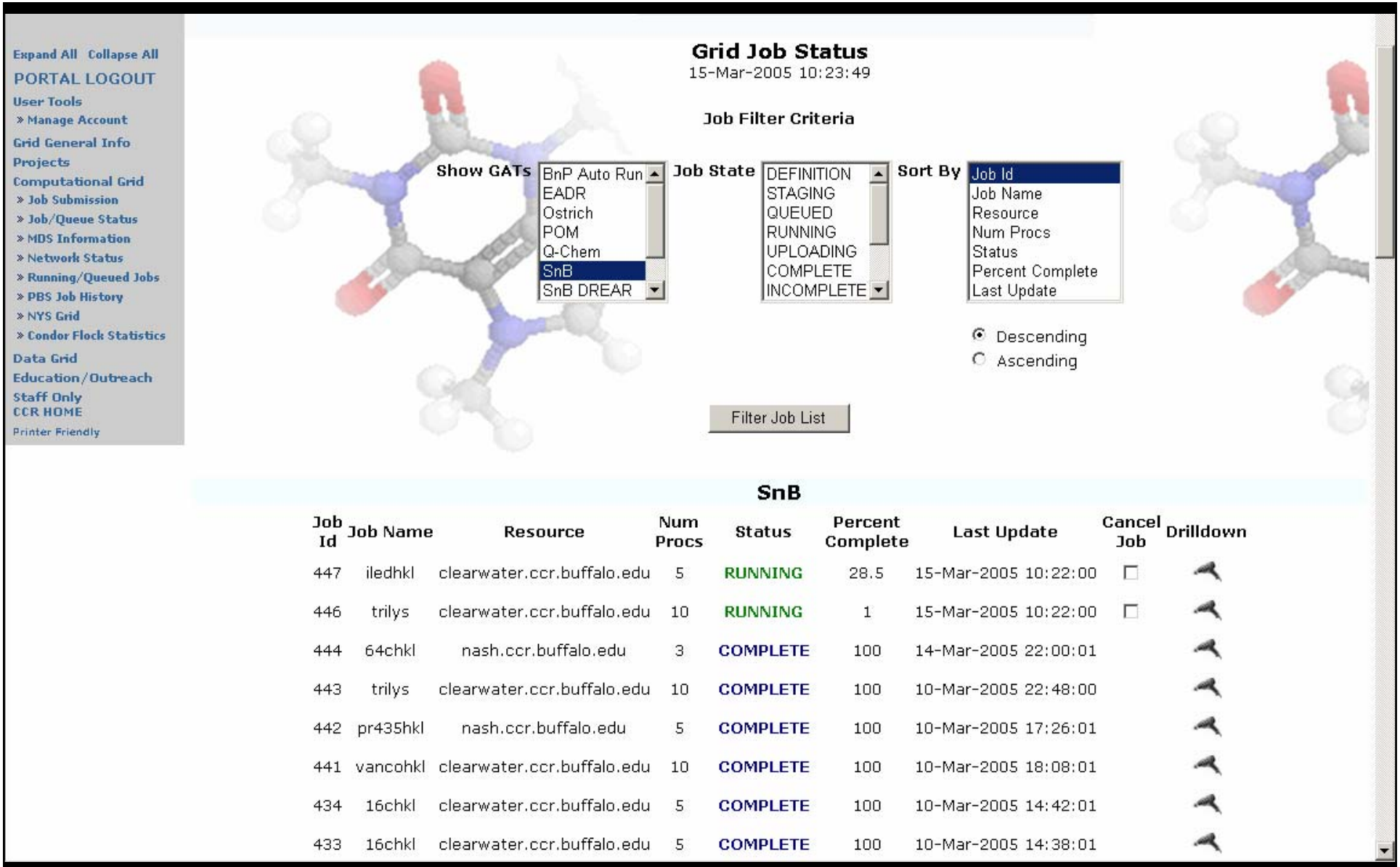
SnB Review (Grid job ID: 447)



Graphical Representation of Intermediate Job Status

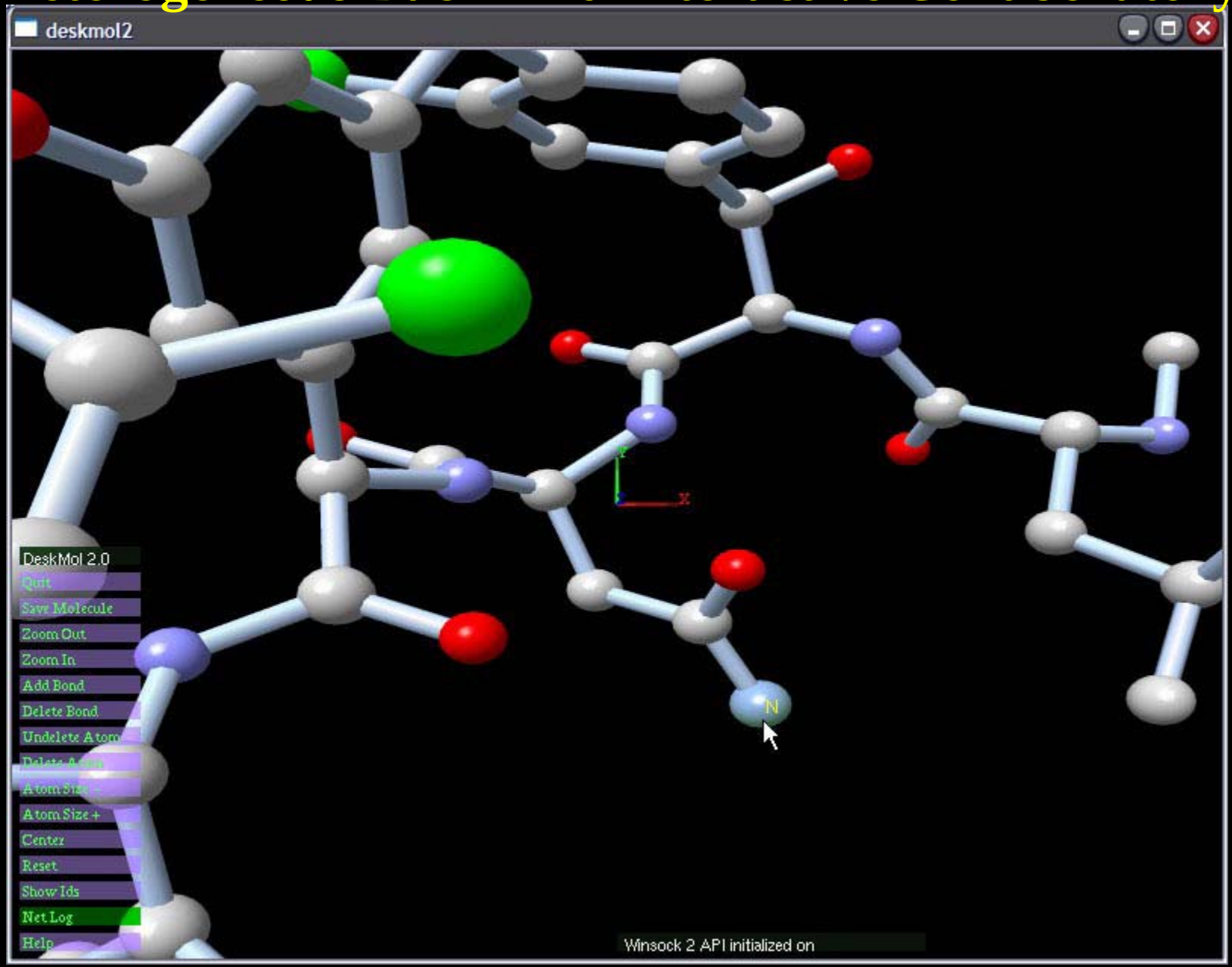


Histogram of Completed Trial Structures

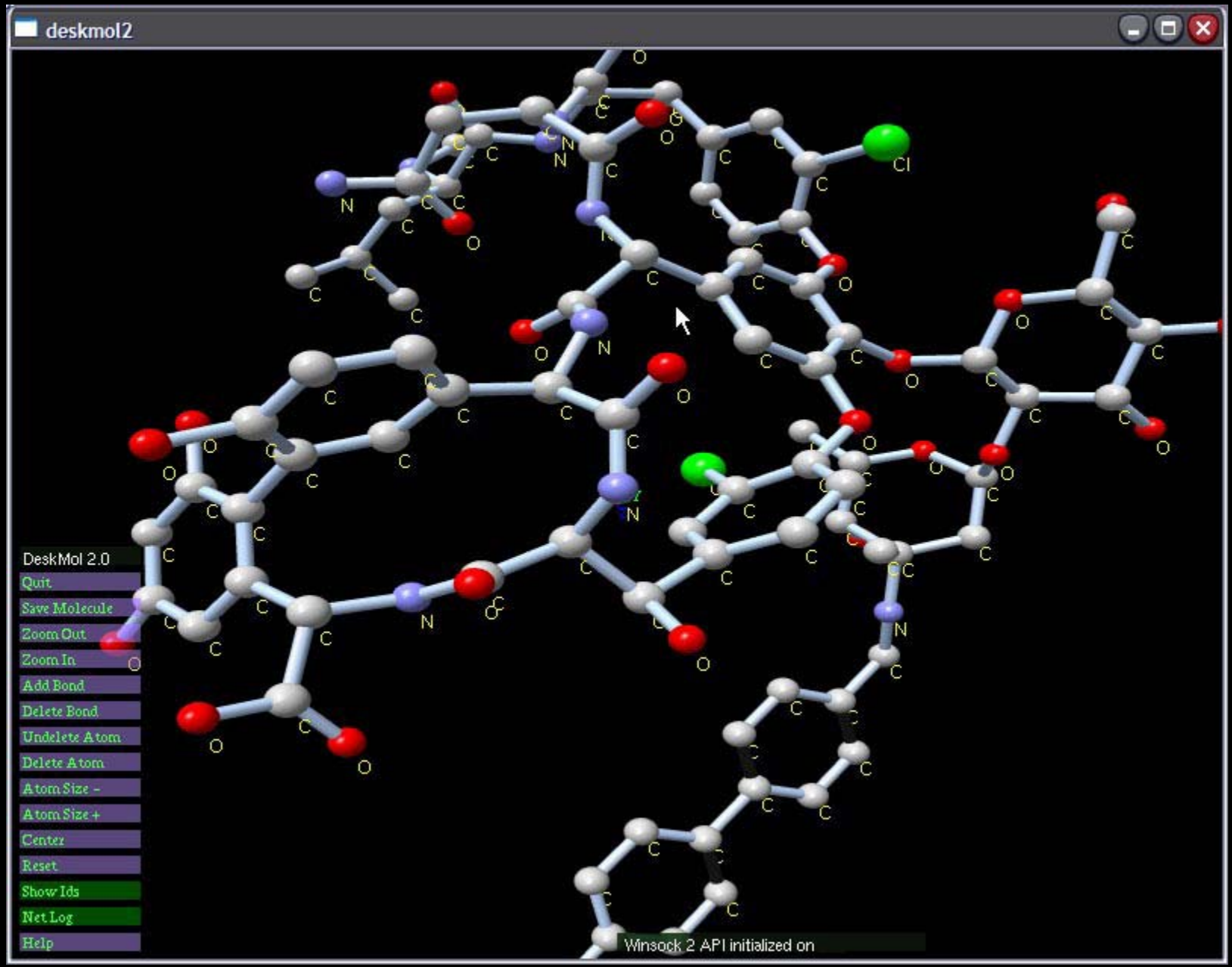


Status of Jobs

Heterogeneous Back-End Interactive Collaboratory



User starts up – default image of structure.



Molecule scaled, rotated, and labeled.

Binghamton University

- Grid Computing Research Laboratory
- Drs. Kenneth Chiu, Madhu Govindaraju, and Michael Lewis.
- Techniques for Web and grid service performance optimization
- Component frameworks for grids
- Instruments and sensors for grid environments
- Adaptive information dissemination protocols across grid overlays
- Emulation framework for grid computation on multi-core processors
- Secure grid data transfer
- www.grid.cs.binghamton.edu/

NYSGrid.org

- **Grass-Roots Cyberinfrastructure Initiative in NYS.**
- **Open to academic and research institutions.**
- **Mission Stmt: To create and advance collaborative technological infrastructure that supports and enhances the research and educational missions of institutions in NYS.**
- **Enable Research, Scholarship, and Economic Development in NYS.**
- **To date, no utilization on any Grid through their VO.**
- **www.nysgrid.org**
- **Recently made part of NYSERNet.**

Acknowledgments

- Mark Green
- Cathy Ruby
- Amin Ghadersohi
- Naimesh Shah
- Steve Gallo
- Jason Rappleye
- Jon Bednasz
- Sam Guercio
- Martins Innus
- Cynthia Cornelius

- George DeTitta
- Herb Hauptman
- Charles Weeks
- Steve Potter

- Alan Rabideau
- Igor Janckovic
- Michael Sheridan
- Abani Patra
- Matt Jones

- NSF ITR
- NSF CRI
- NSF MRI
- NYS
- CCR


www.cse.buffalo.edu/faculty/miller
CCR
CCR