Molecular Structure Determination and the ACDC Computational and Data Grid

Russ Miller

Center for Computational Research

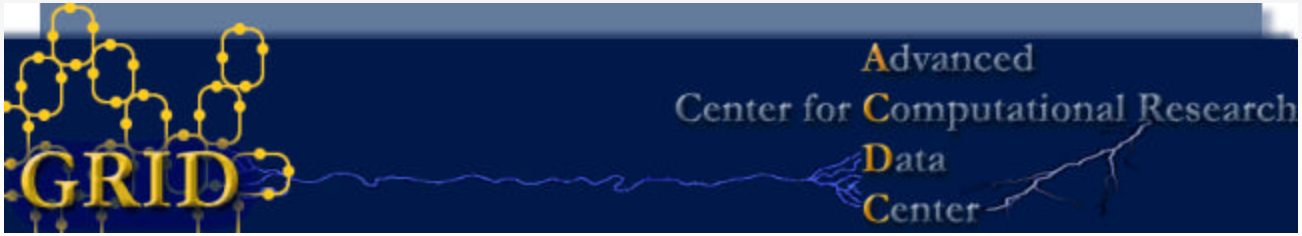
Computer Science & Engineering
SUNY-Buffalo

Hauptman-Woodward Medical Inst

NSF, NIH, DOE
NIMA, NYS, HP

Grid3 2003

SnB



University at Buffalo
The State University of New York

Research Activities

- **Theory/Algorithms**
  - Fundamental Problems, Data Movement, Computational Geometry, Image Analysis
  - Mesh, Pyramid, Hypercube, PRAM, Reconfigurable Mesh, CGM
- **Experimentation**
  - Distributed- and Shared-Memory Machines
  - Computational Geometry, NP-Hard Approximation Algorithms, Image Analysis
- **Applications**
  - Molecular Structure Determination
- **Systems**
  - Grid Computing

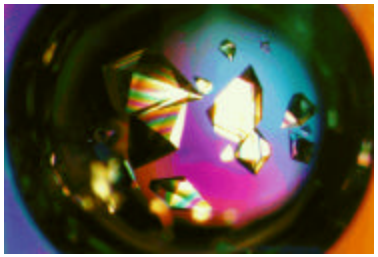
"Science is a Team Sport"

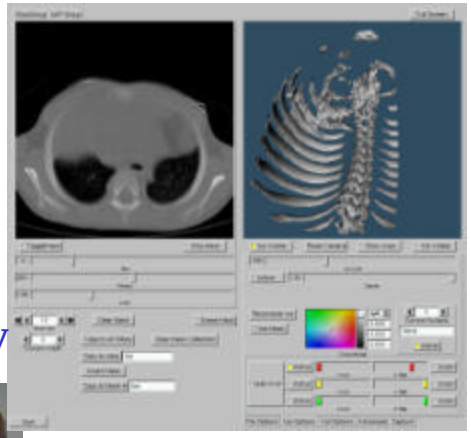
Academia in the 21st Century

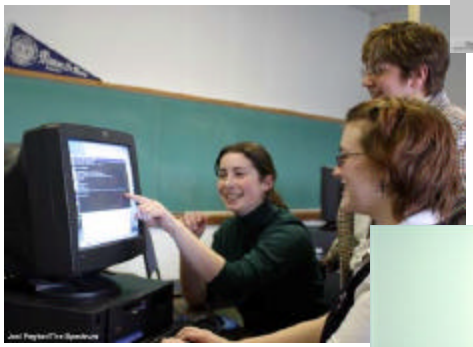
- Embrace digital data-driven society
- Empower students to compete in knowledge-based economy
- Support research, scholarship, education, and outreach
- Support HPC infrastructure, research, and applications
- Deliver *high-end cyberinfrastructure* to enable efficient
  - Collection of data
  - Management/Organization of data
  - Distribution of data
  - Analysis of data
  - Visualization of data

Center for Computational Research 1998-2005 Snapshot

- **High-End Computing, Storage, Networking, and Visualization**
  - **~140 Research Groups in 37 Depts**
    - **Physical Sciences**
    - **Life Sciences**
    - **Engineering**
    - **Scientific Visualization, Medical Imaging, Virtual Reality**
  - **13 Local Companies**
  - **10 Local Institutions**
- **External Funding: $300M+**
- **Total Leveraged WNY: $500M+**
- **Deliverables**
  - **1100+ Publications**
  - **Software, Media, Algorithms, Consulting, Training, CPU Cycles…**







Major Compute/Storage Resources

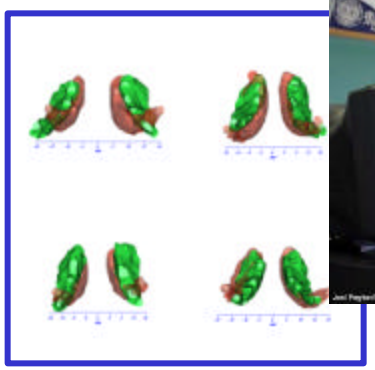
- Dell Linux Cluster (10TF peak)
  - 1600 Xeon EM64T Processors (3.2 GHz)
  - 2 TB RAM; 65 TB Disk
  - Myrinet / Force10
  - 30 TB EMC SAN
- Dell Linux Cluster (2.9TF peak)
  - 600 P4 Processors (2.4 GHz)
  - 600 GB RAM; 40 TB Disk; Myrinet
- Dell Linux Cluster (6TF peak)
  - 4036 Processors (PIII 1.2 GHz)
  - 2TB RAM; 160TB Disk; 16TB SAN
- IBM BladeCenter Cluster (3TF peak)
  - 532 P4 Processors (2.8 GHz)
  - 5TB SAN
- SGI Intel Linux Cluster (0.1TF peak)
  - 150 PIII Processors (1 GHz)
  - Myrinet
- SGI Altix3700 (0.4TF peak)
  - 64 Processors (1.3GHz ITF2)
  - 256 GB RAM
  - 2.5 TB Disk
- Apex Bioinformatics System
  - Sun V880 (3), Sun 6800
  - Sun 280R (2)
  - Intel PIIIs
  - Sun 3960: 7 TB Disk Storage
- HP/Compaq SAN
  - 75 TB Disk; 190 TB Tape
  - 64 Alpha Processors (400 MHz)
  - 32 GB RAM; 400 GB Disk

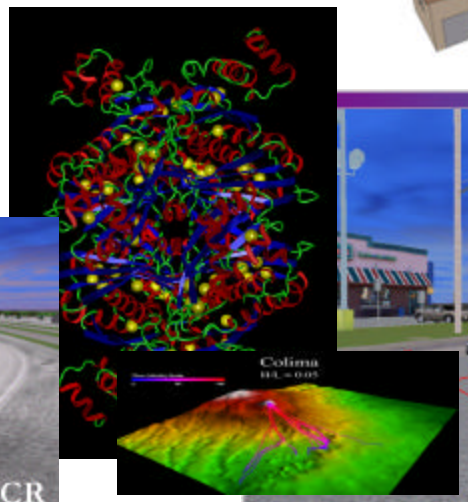
CCR Visualization Resources

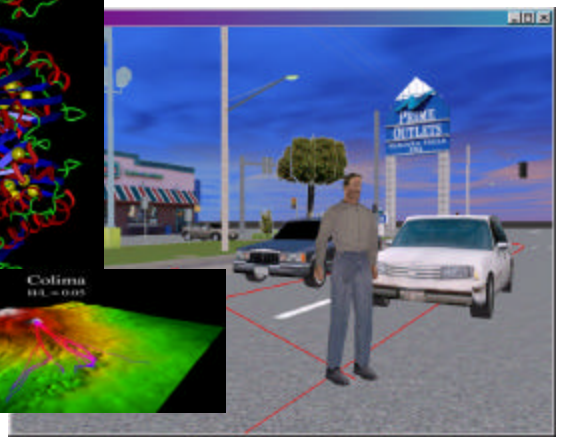
- Fakespace ImmersaDesk R2
  - Portable 3D Device
  - Onyx2: 6 R10000 @ 250MHz
  - 2 IR2 Pipes; 3 64MB texture memory mgrs.
- Tiled-Display Wall
  - 20 NEC projectors: 15.7M pixels
  - Screen is 11’ ´ 7’
  - Dell PCs with Myrinet2000
- Access Grid Nodes (2)
  - Group-to-Group Communication
  - Commodity components
- SGI Reality Center 3300W
  - Dual Barco’s on 8’ ´ 4’ screen
  - Onyx300: 10 R14000 @ 500MHz
  - 2 IR4 Pipes; 1 GB texture mem per pipe

CCR Research & Projects

- Ground Water Modeling
- Computational Fluid Dynamics
- Molecular Structure Determination
- Protein Folding
- Digital Signal Processing
- Grid Computing
- Computational Chemistry
- Bioinformatics
- Real-time Simulations and Urban Visualization
- Accident Reconstruction
- Risk Mitigation (GIS)
- Medical Visualization
- High School Workshops
- Virtual Reality

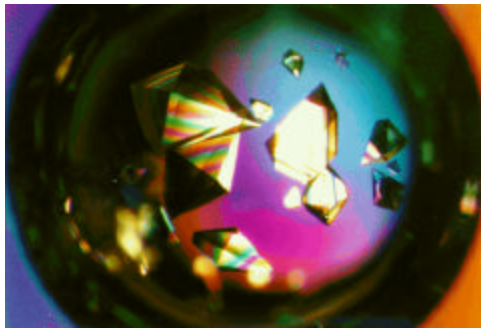


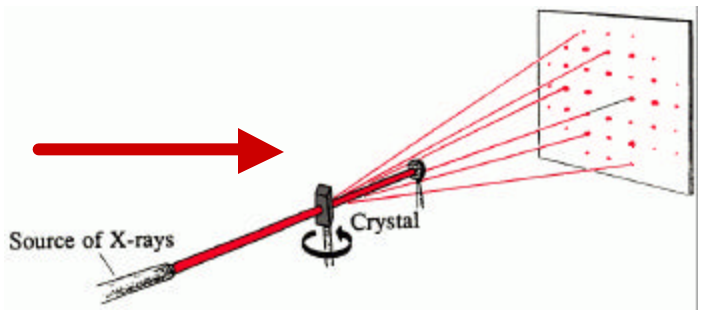


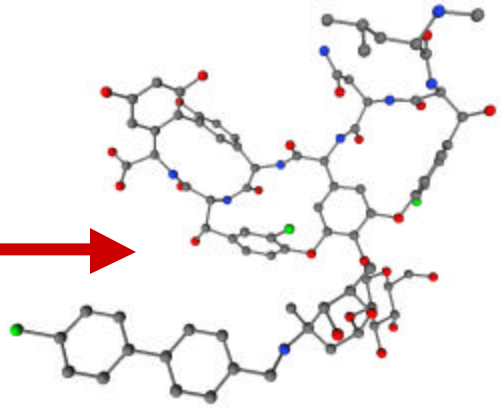


Molecular Structure Determination via *Shake-and-Bake*

- ***SnB* Software by UB/HWI**
  - IEEE "Top Algorithms of the Century"
- **Worldwide Utilization**
- **Critical Step**
  - Rational Drug Design
  - Structural Biology
  - Systems Biology
- **Vancomycin**
  - "Antibiotic of Last Resort"
- **Current Efforts**
  - Grid
  - Collaboratory
  - Intelligent Learning

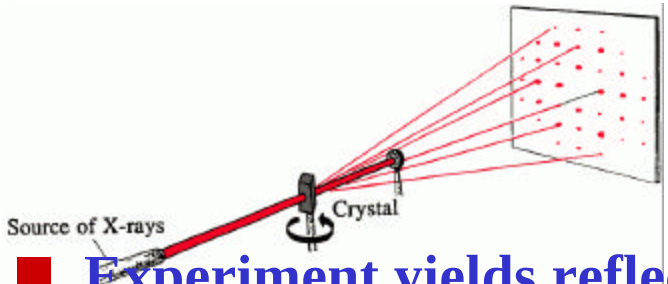




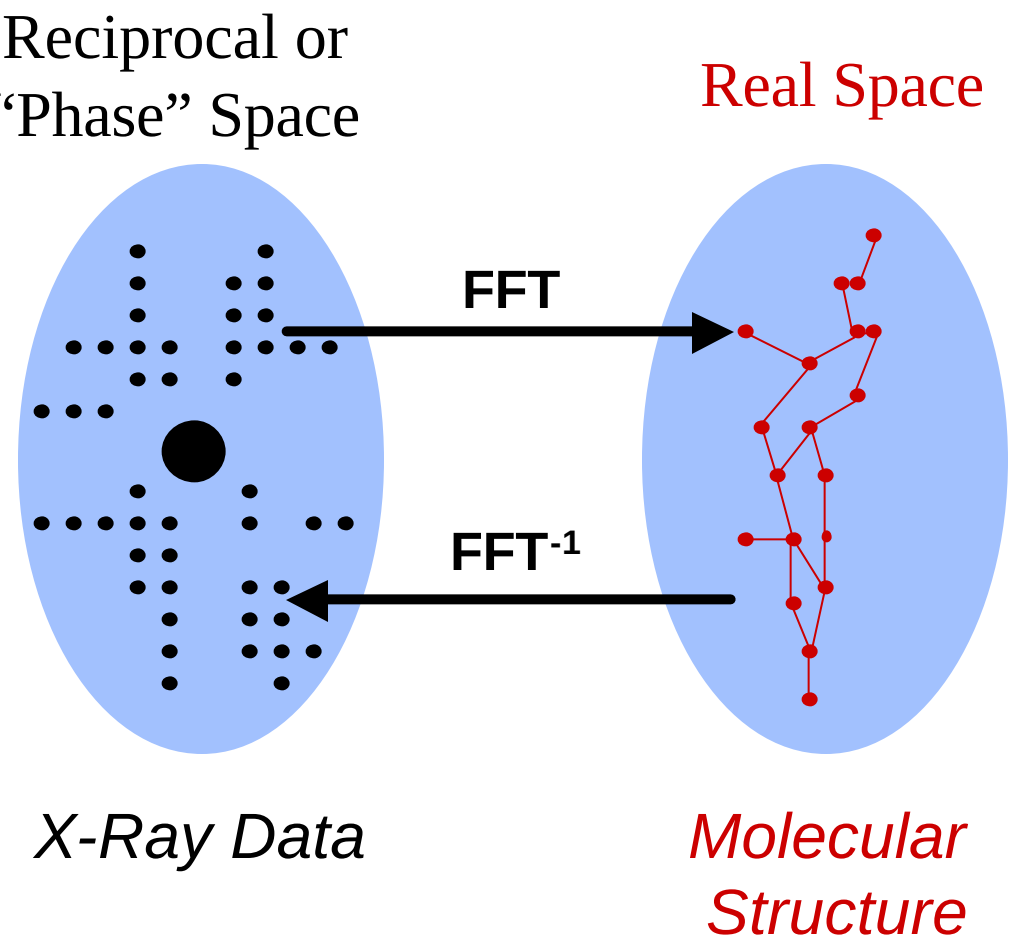


1. Isolate a single crystal 2. Perform the X-Ray diffraction experiment 3. Determine the crystal structure

X-Ray Data & Corresponding Molecular Structure



- **Experiment yields reflections and associated intensities.**
- **Underlying atomic arrangement is related to the reflections by a 3-D Fourier transform.**
- ***Phase angles are lost in experiment.***
- ***Phase Problem*: Determine the set of phases corresponding to the reflections.**



Overview of Direct Methods

- **Probability theory gives information about certain linear combinations of phases.**
  - **In particular, the triples $f_H + f_K + f_{-H-K} = 0$ with high probability.**
- **Probabilistic estimates are expressed in terms of normalized structure factor magnitudes (|E|).**
- **Optimization methods are used to extract the values of individual phases.**
- **A multiple trial approach is used during the optimization process.**
- **A suitable figure-of-merit is used to determine the trials that represent solutions.**

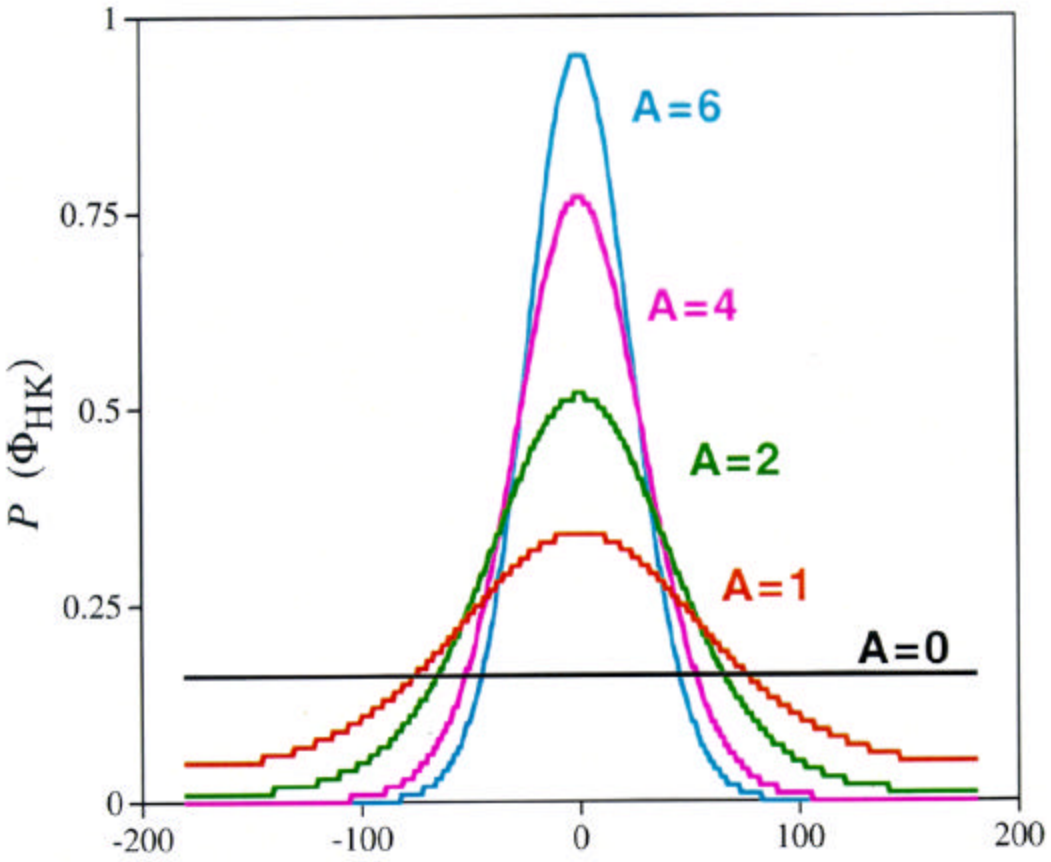
Normalized Structure-Factor Magnitudes: $|E_H|$

$$E_H = |E_H| \exp(i f_H)$$

$$|E_H| = \frac{|F_H|}{\left\langle |F_H|^2 \right\rangle^{1/2}} = \frac{k\left\langle \exp[-B_{iso}(\sin q)^2 / l^2 \right\rangle^{-1} |F_H|_{meas}}{\left(e_H \sum_{j=1}^{N} f_j^{\,2} \right)^{1/2}}$$

- á|E|ñ **constant for concentric resolution shells.**
- á|E|ñ **constant regardless of reflection class (**e_H **correction factor).**
- **The *renormalization* condition,** á|E|²ñ**=1 is always imposed**

Cochran Distribution



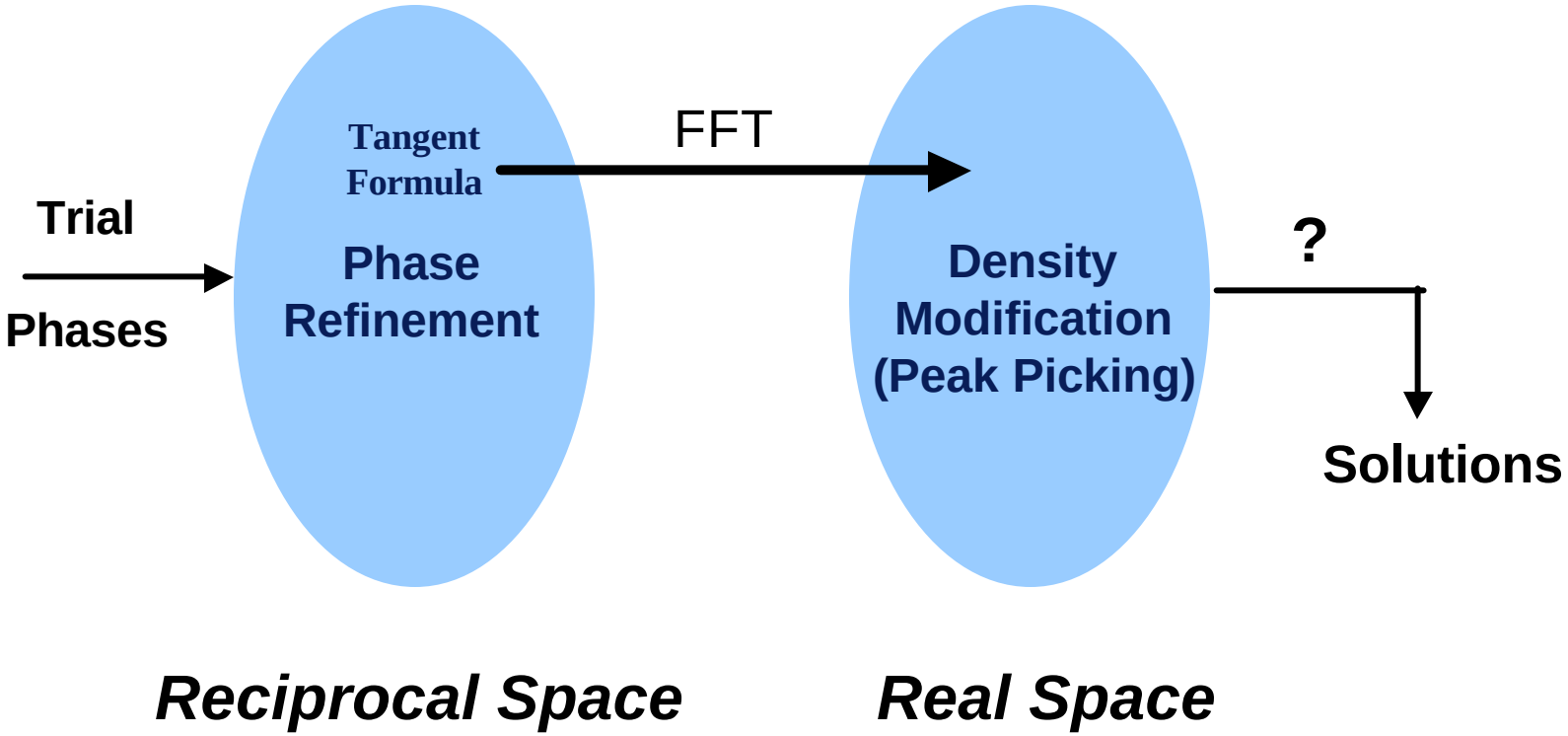
$F_{HK} = f_H + f_K + f_{-H-K}$

- *N*=non-H atoms in unit cell
- Each triplet of phases or structure invariant, F_{HK}, has an associated parameter

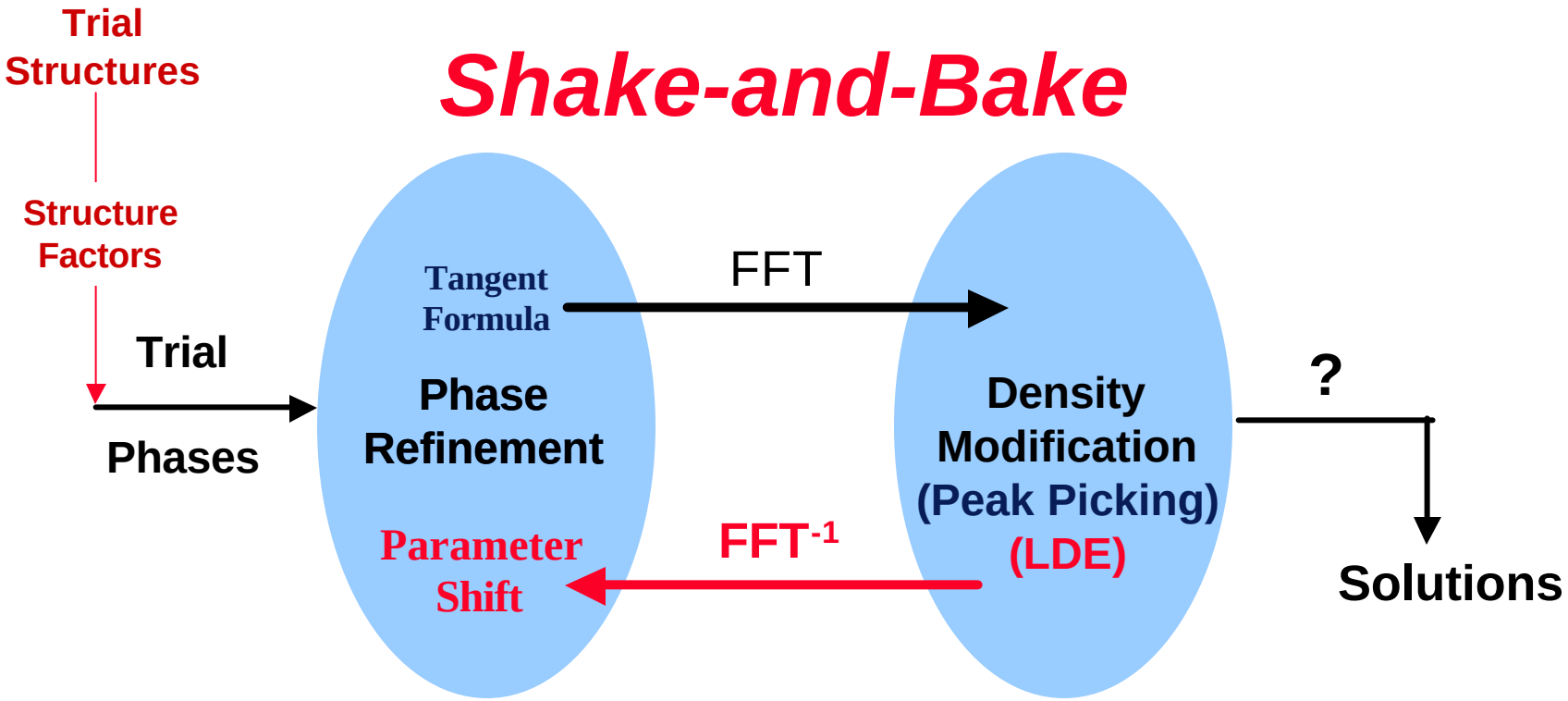
$$A_{HK} = 2|E_H E_K E_{-H-K}|/N^{1/2}$$

- A_{HK} is large if
  - $|E_H|$, $|E_K|$, $|E_{-H-K}|$ are large
  - *N* is small
- If A_{HK} is large, $F_{HK} \approx 0$

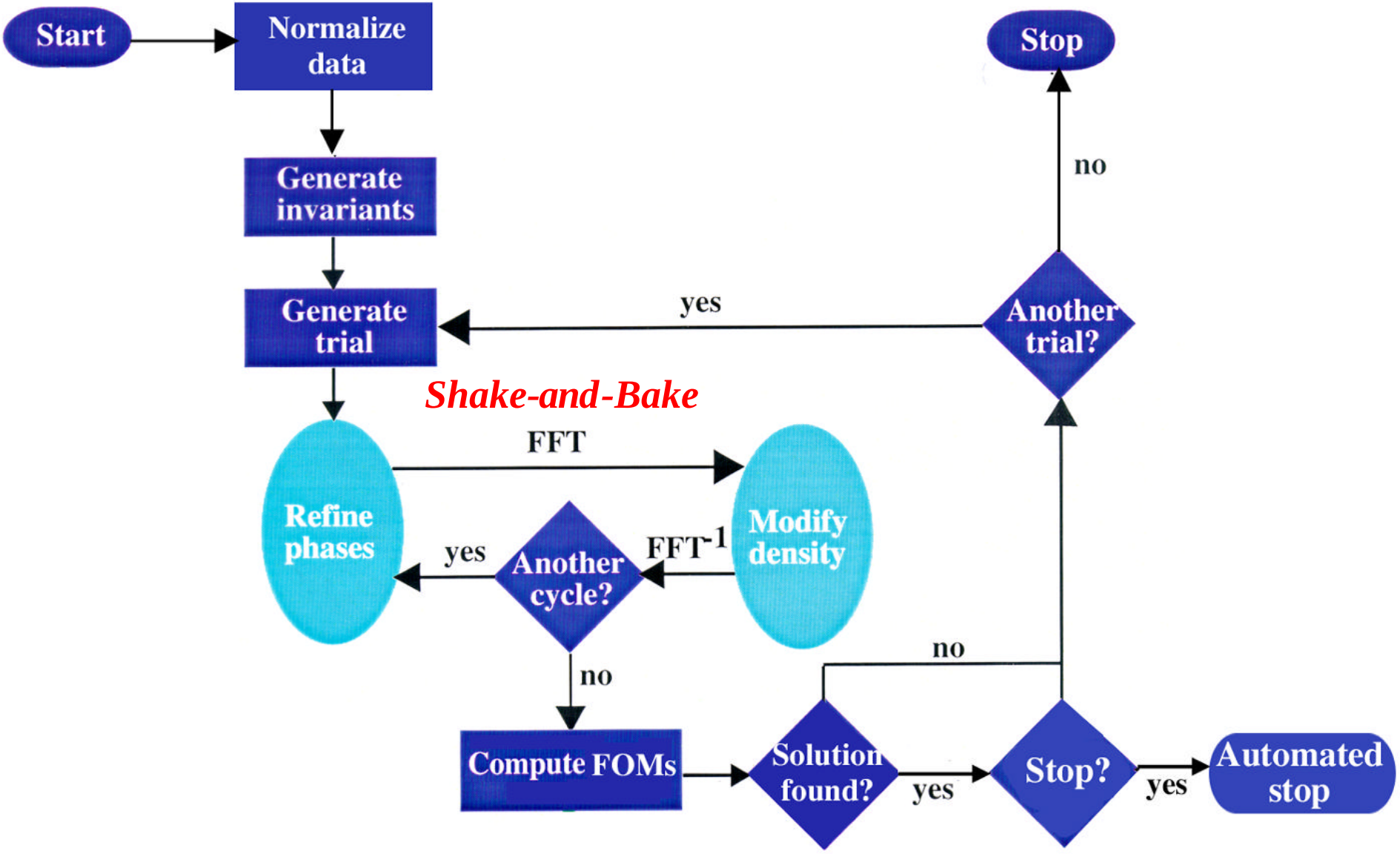
Conventional Direct Methods



Shake-and-Bake Method: Dual-Space Refinement



A Direct Methods Flowchart



Generate Triplet Invariants

Reflections

Rank	h	k	l	|E|
1	0	3	4	4.65
2	0	7	30	3.67
3	5	1	1	3.67
4	8	8	5	3.26
5	6	0	1	3.15
⋮	⋮	⋮	⋮	⋮
10n=840	7	0	3	1.33
841	2	4	30	1.33

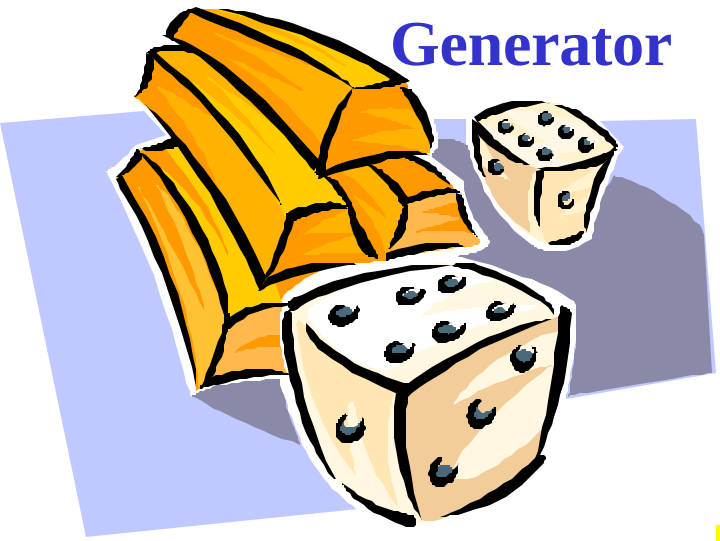
Triplets

Rank	H	K	-H-K	A
1	1	4	45	3.90
2	1	3	165	3.52
3	3	5	17	3.37
4	1	3	289	3.16
5	1	28	40	3.09
⋮	⋮	⋮	⋮	⋮
100n=840	19	259	734	0.71
8401	142	179	283	0.71

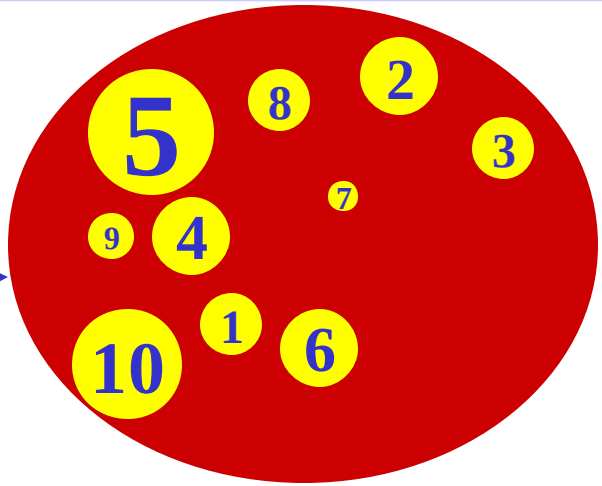
$n = 84$ unique atoms

Getting Started: Random Atoms

Random Number Generator



n = 10 atoms
(30 coordinates)



Structure Factor Calculation

ϕ_1 ϕ_2
ϕ_3 ϕ_4
ϕ_5 ϕ_6
ϕ_7 ϕ_8
ϕ_9 ϕ_{10}

Useful Relationships for Multiple Trial Phasing

Tangent Formula

$$\tan f_H = \frac{-\sum_K |E_K E_{-H-K}| \sin(f_K + f_{-H-K})}{\sum_K |E_K E_{-H-K}| \cos(f_K + f_{-H-K})}$$

Parameter Shift Optimization

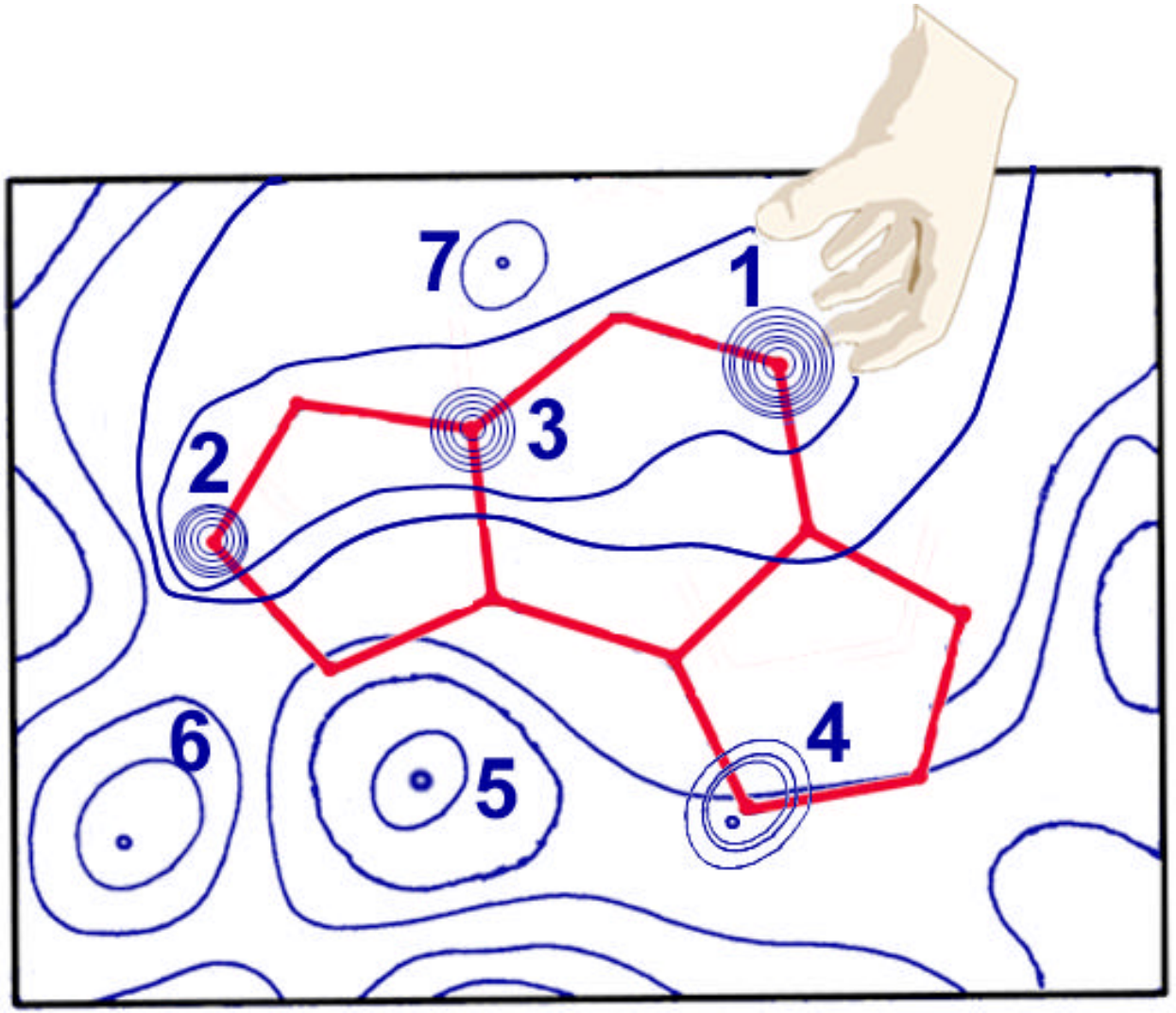
$$R(f) = \frac{1}{\sum_{H,K} W_{HK}} \sum_{H,K} W_{HK} \left(\cos \Phi_{HK} - \frac{I_1(W_{HK})}{I_0(W_{HK})} \right)^2$$

where $|E_H| \propto |F_H|$ normalized in resolution shells

Invariants : $\Phi_{HK} = f_H + f_K + f_{-H-K} \approx 0$

Weights : $W_{HK} = A_{HK} = 2N^{-1/2} |E_H E_K E_{-H-K}|$

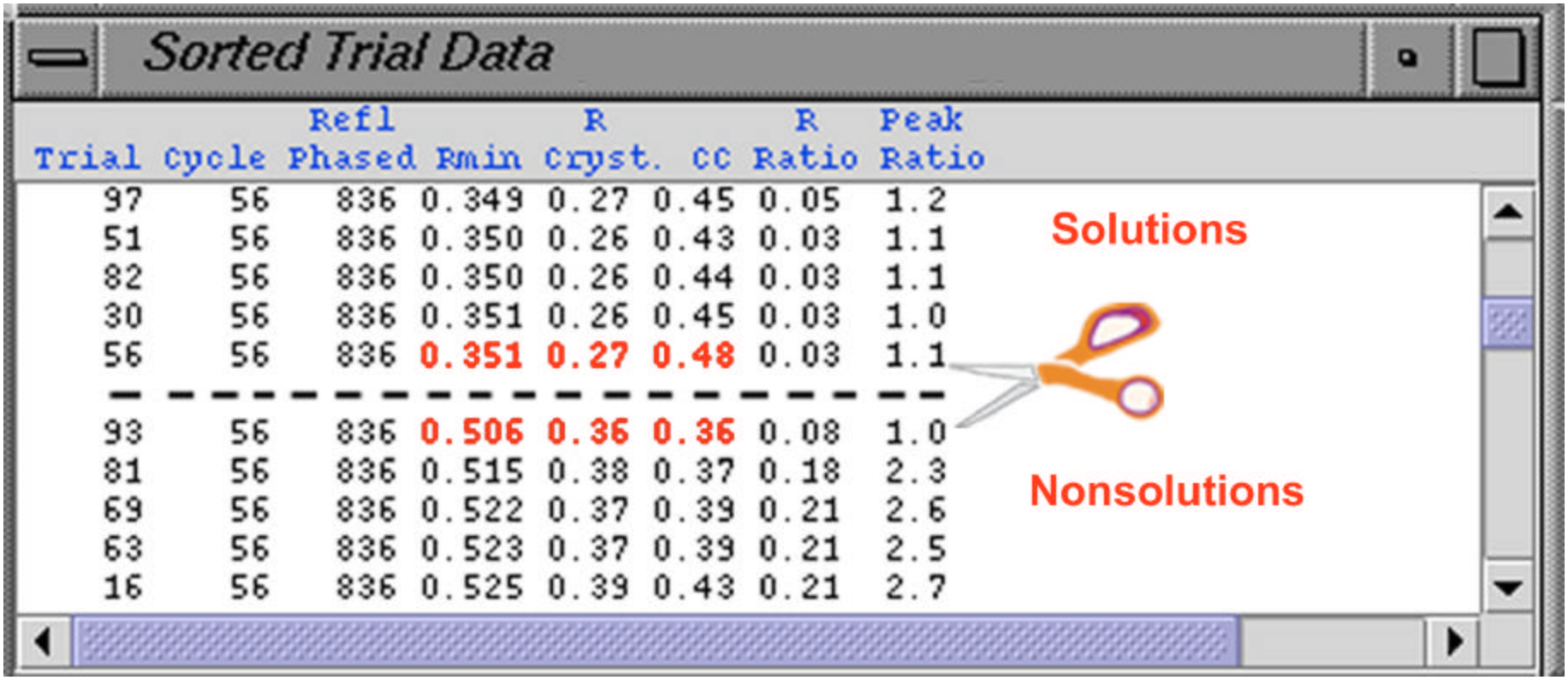
Peak Picking



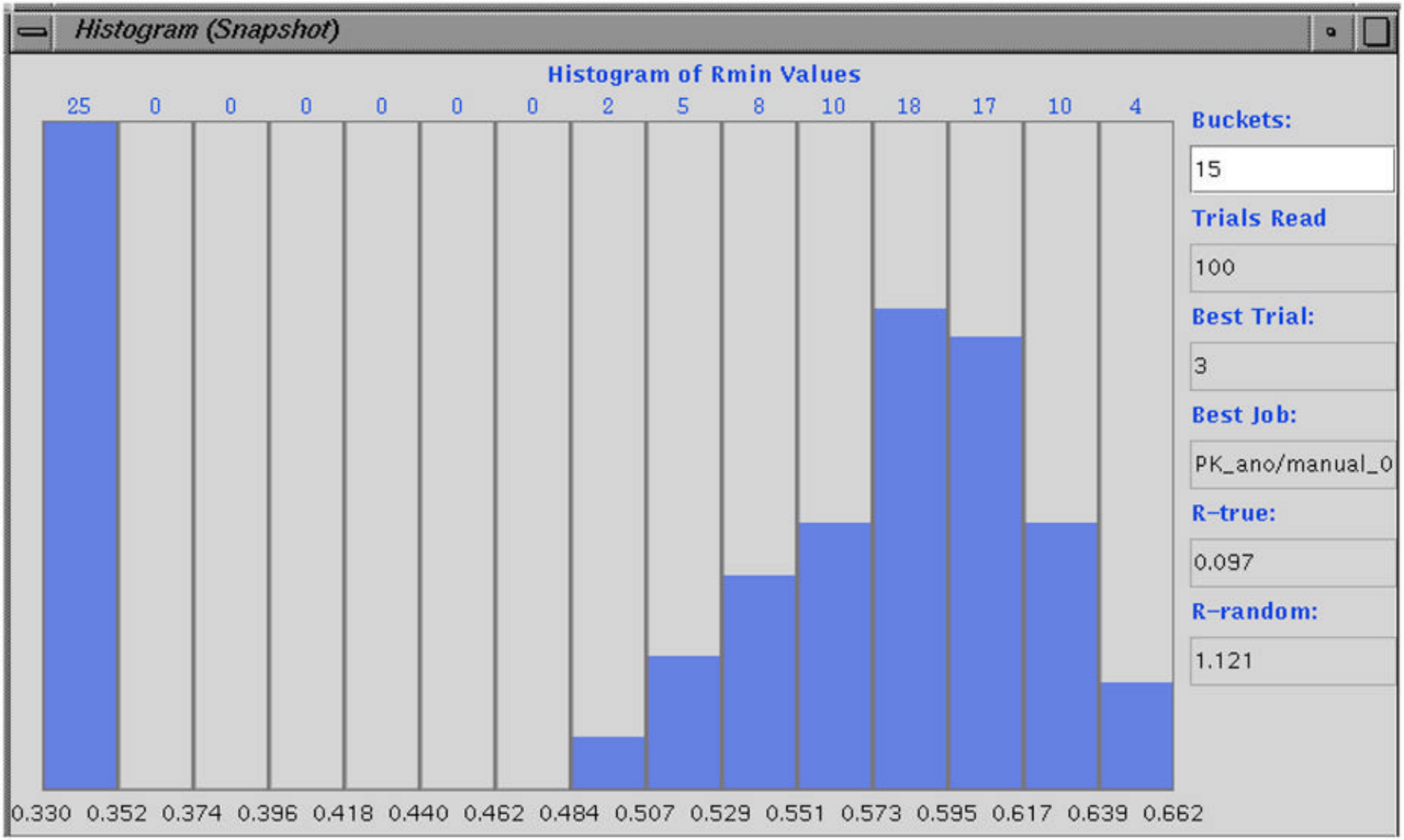
Default *SnB* Parameters (given *n* atoms)

Parameter	Full Structures	Substructures
Phases	10*n*	30*n*
Triplet Invariants	100*n*	300*n*
Cycles		
n<100	*n*/2	2*n*
n>100	*n*	2*n*
Peaks		
n<100	*n*	*n*
n>100	0.8*n*	0.8*n*

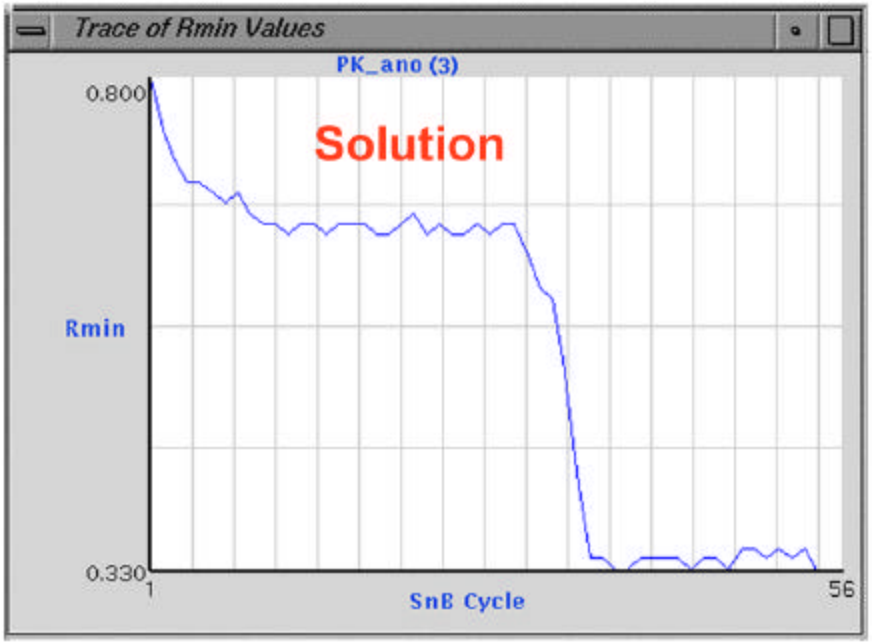
Sorted Trials

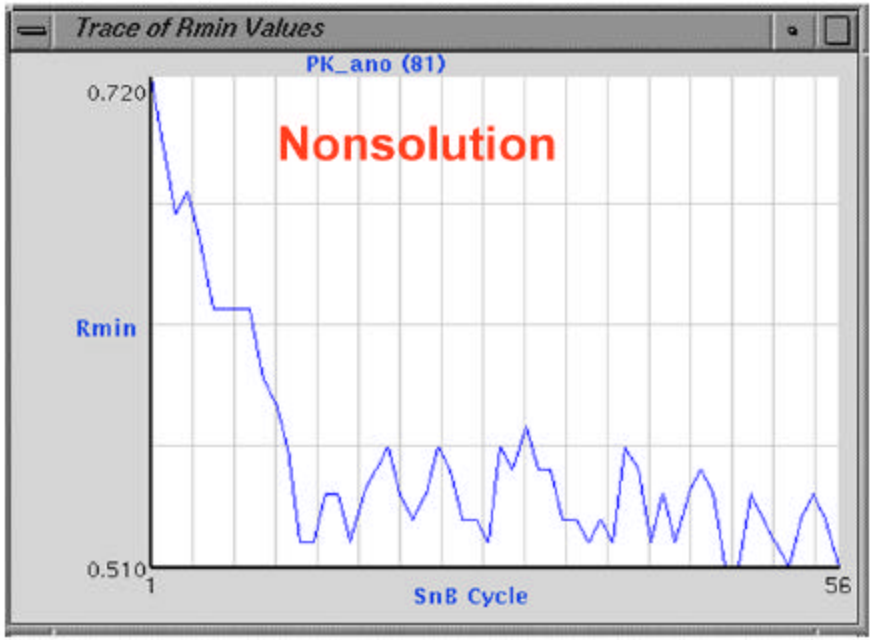


Ph8755: *SnB* Histogram

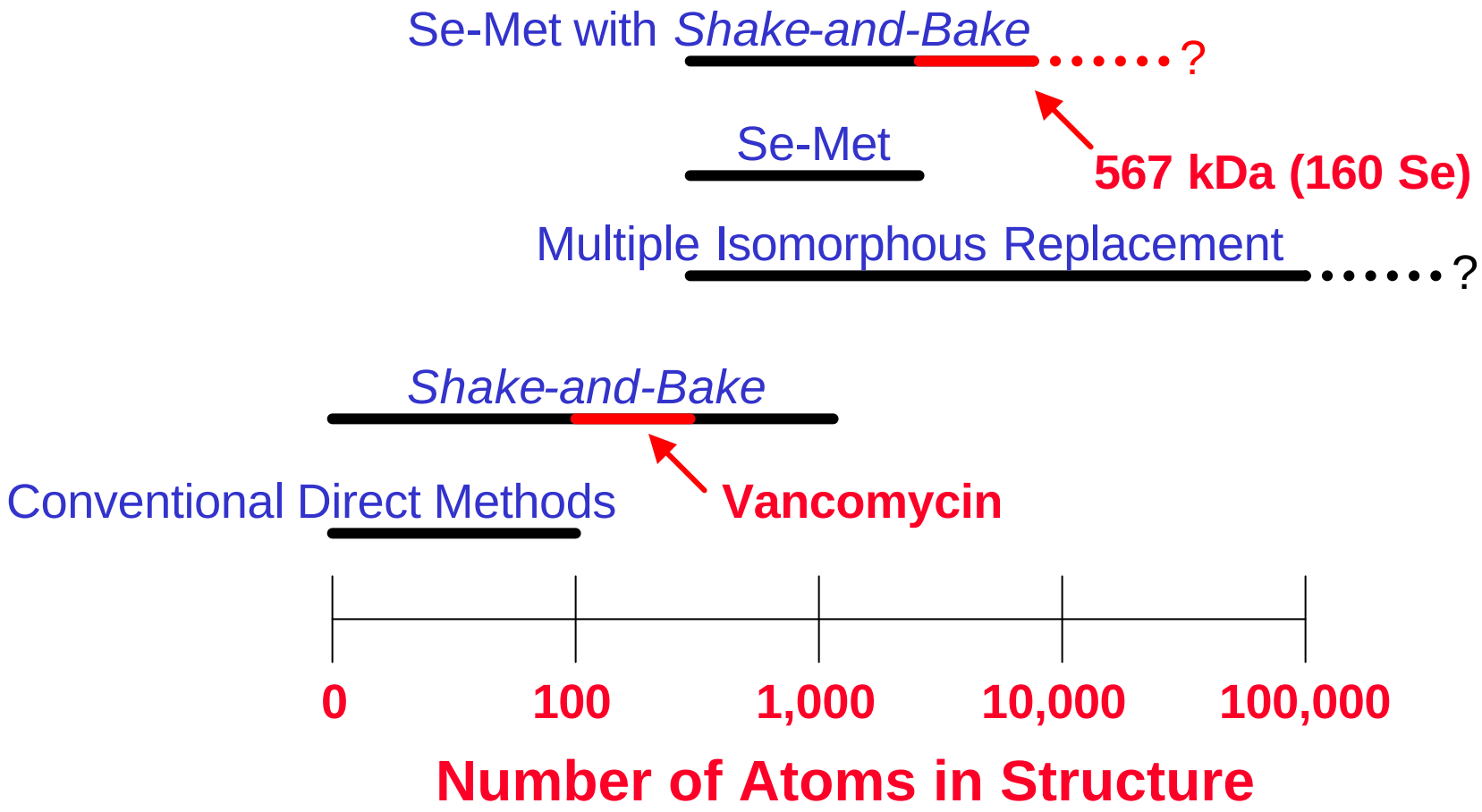


Minimal Function Traces

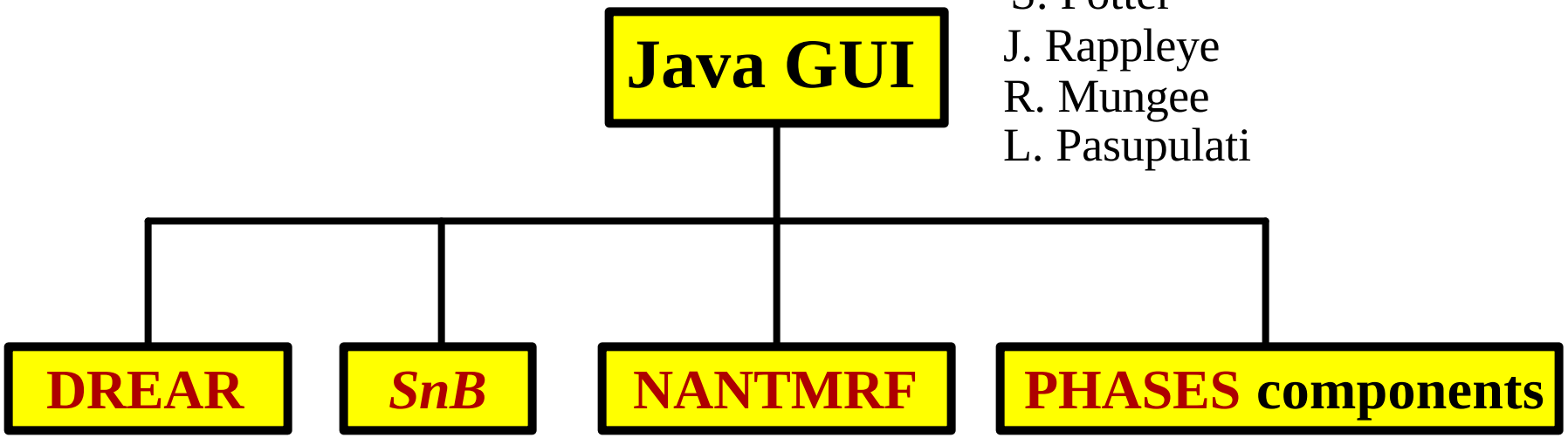




Phasing and Structure Size



BnP: The Buffalo 'n Pittsburgh Interface



BnP Overview

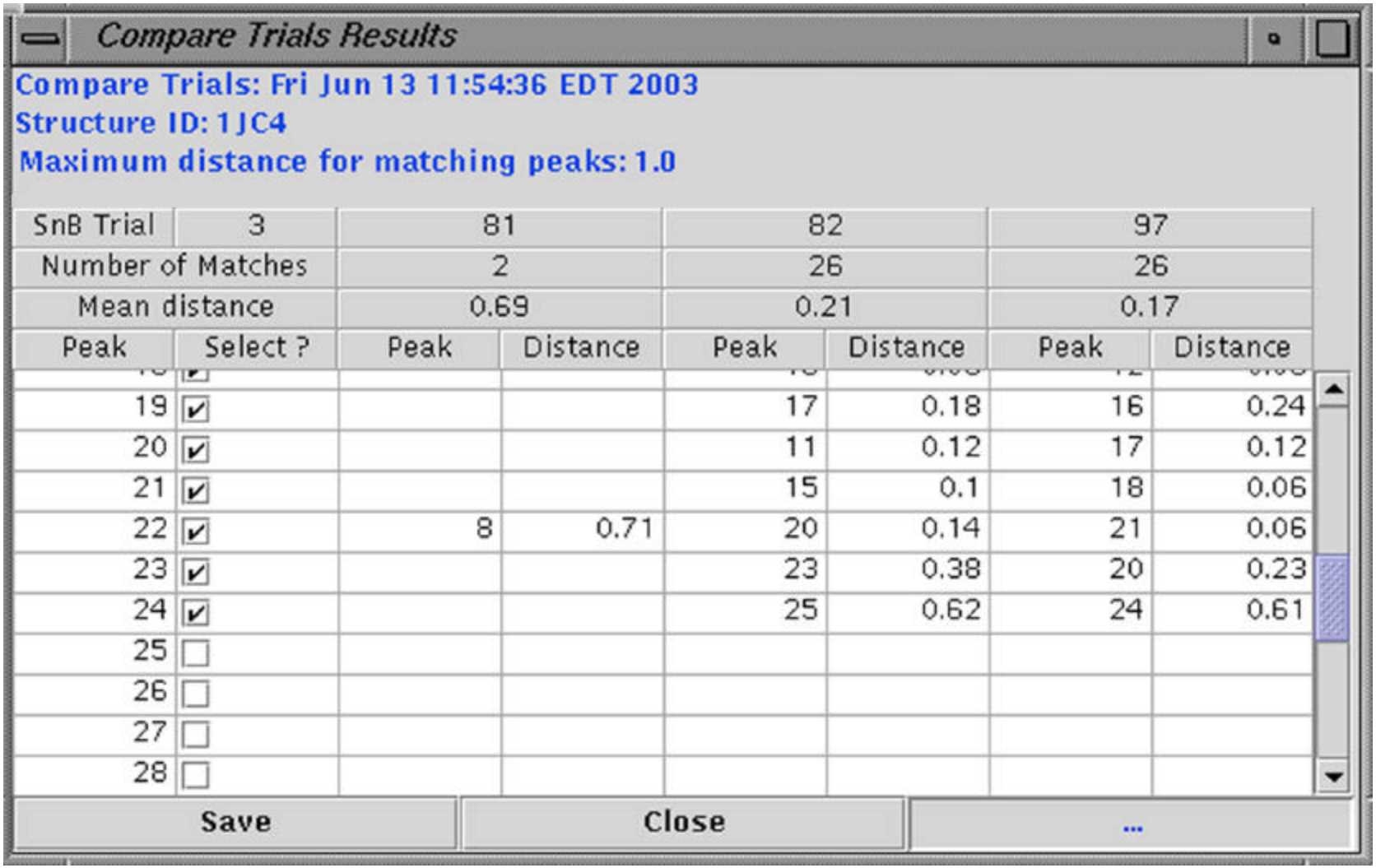
Task	Manual Mode (Workflow)	Auto Mode
Substructure Determination	FOM Histogram and Trace	FOM Deviation from Mean
Site Validation	Trial Comparison	Occupancy Refinement
Enantiomorph Determination	Map Inspection	s(protein)/ s (solvent)
Substructure Refinement (Optional)	← SAME →	
Solvent Flattening	← SAME →	

SeMet Test Data: Auto Mode Results

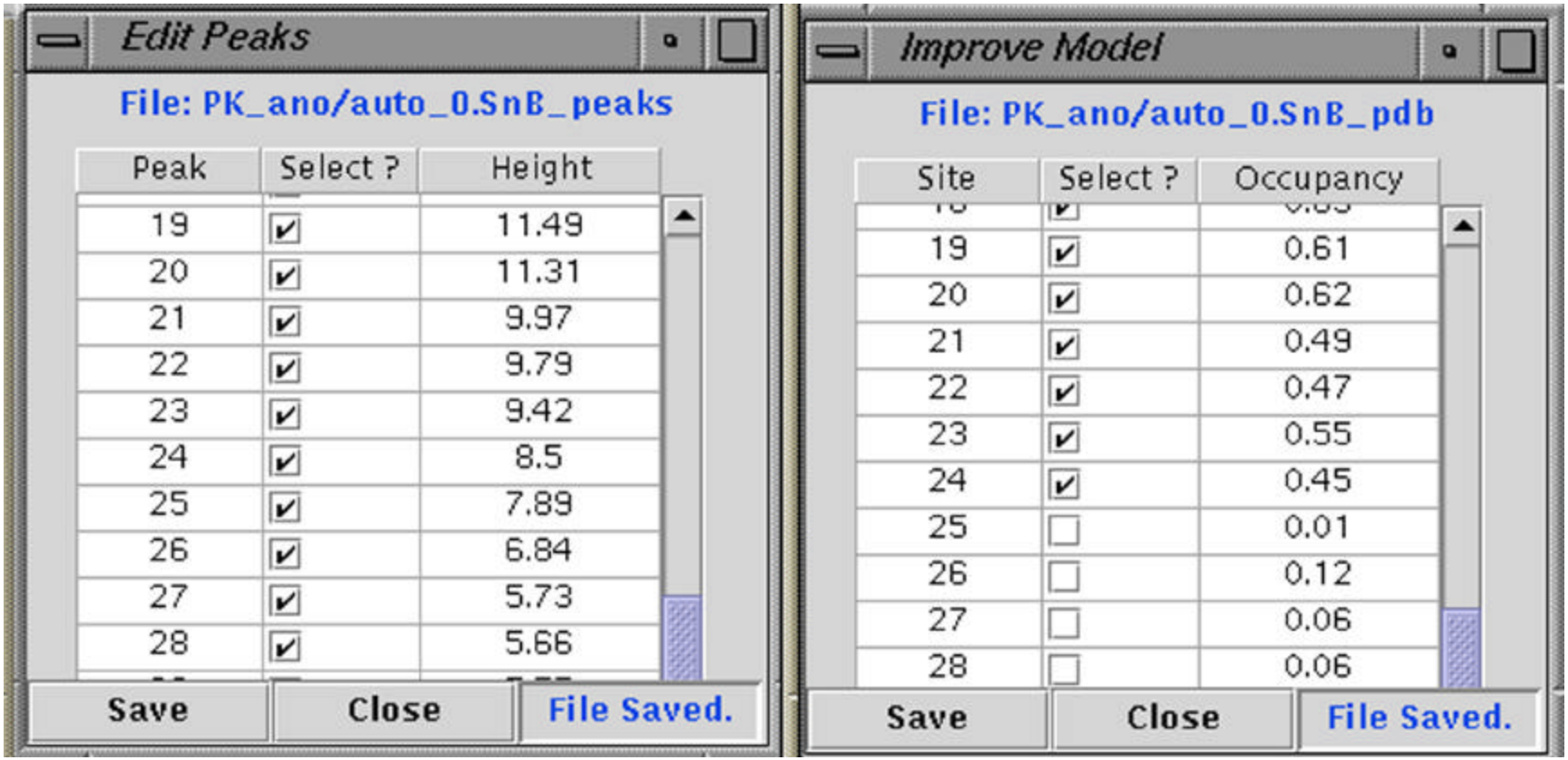
PDB Code	No. Sites	No. Trials	Time (min) Apple Power Mac G5	PDB Code	No. Sites	No. Trials	Time (min) Apple Power Mac G5
1QC2	4	6	<1	1CLI	28	43	1
1BX4	7	17	<1	1A7A	30	195	9
1CB0	8	119	<1	1L8A	40	111	2
1T5H	10	6	<1	1E3M	45	28	2
1GSO	13	111	<1	1HI8	50	28	2
2JXH	14	106	<1	1GKP	54	578	102
2TPS	15	107	<1*	1DQ8	60	119	7
1DBT	19	61	<1	1E2Y	60	19	<1
1JEN	22	6	<1	1M32	66	111	5
1JC4	24	6	<1	1EQ2	70	8	1

* Solution not recognized automatically

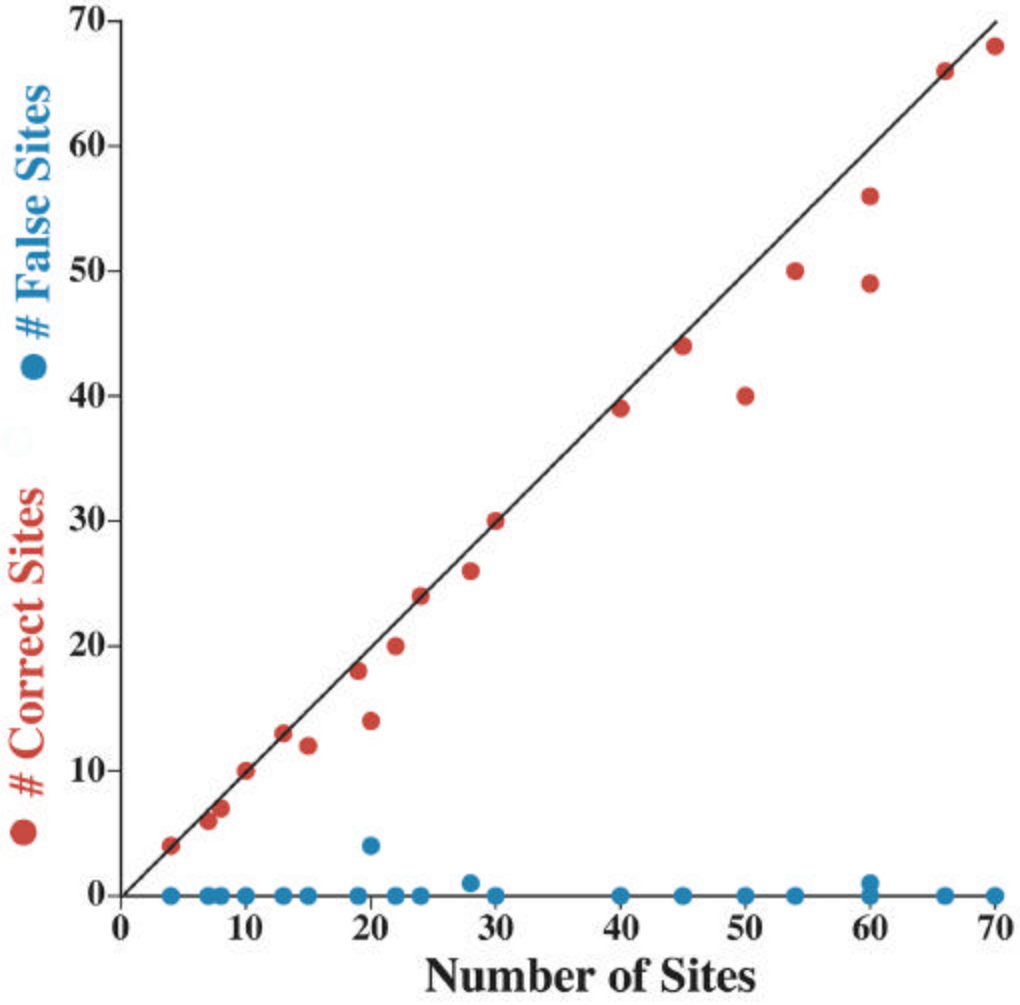
Manual Site Validation: Trial Comparison



Automated Site Validation: Occupancy Refinement

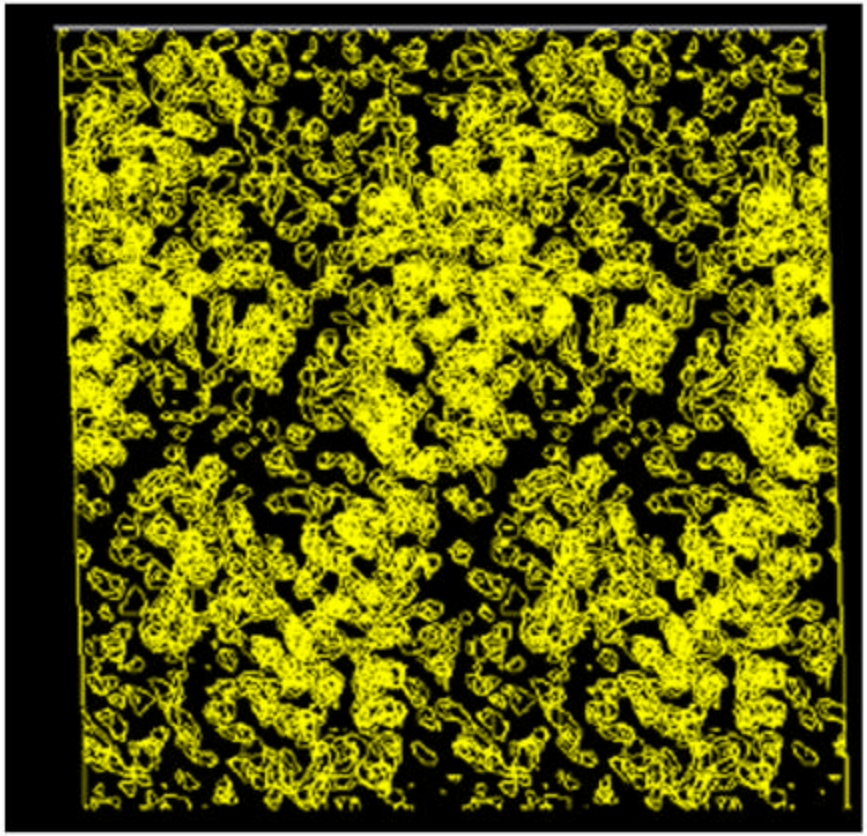


SeMet Test Data: Site Validation

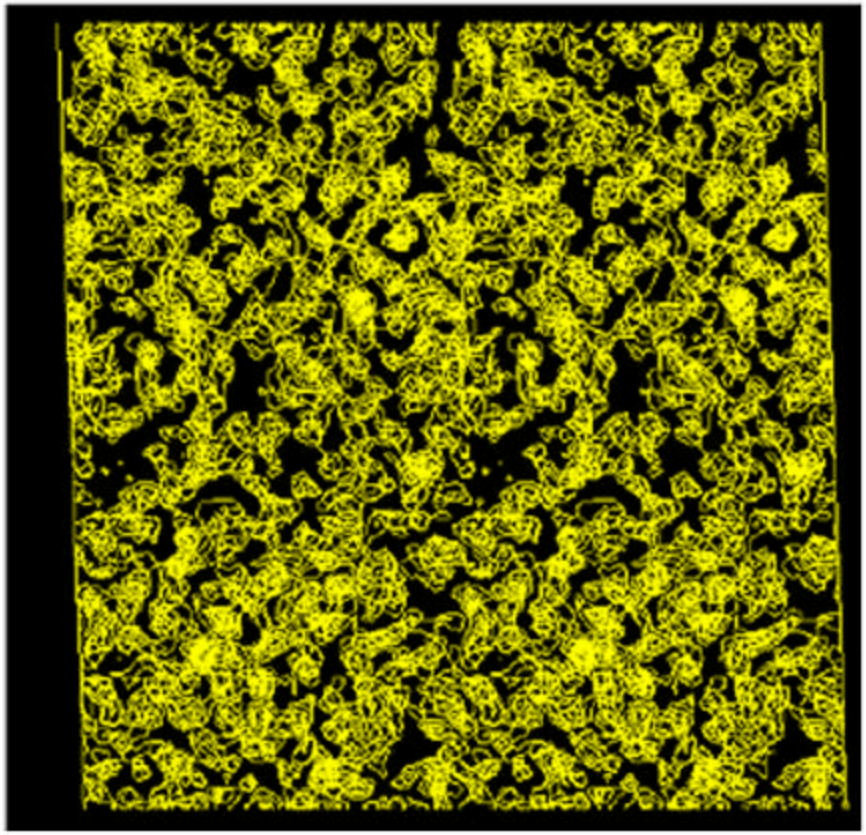


Manual Enantiomorph Determination: Map Inspection

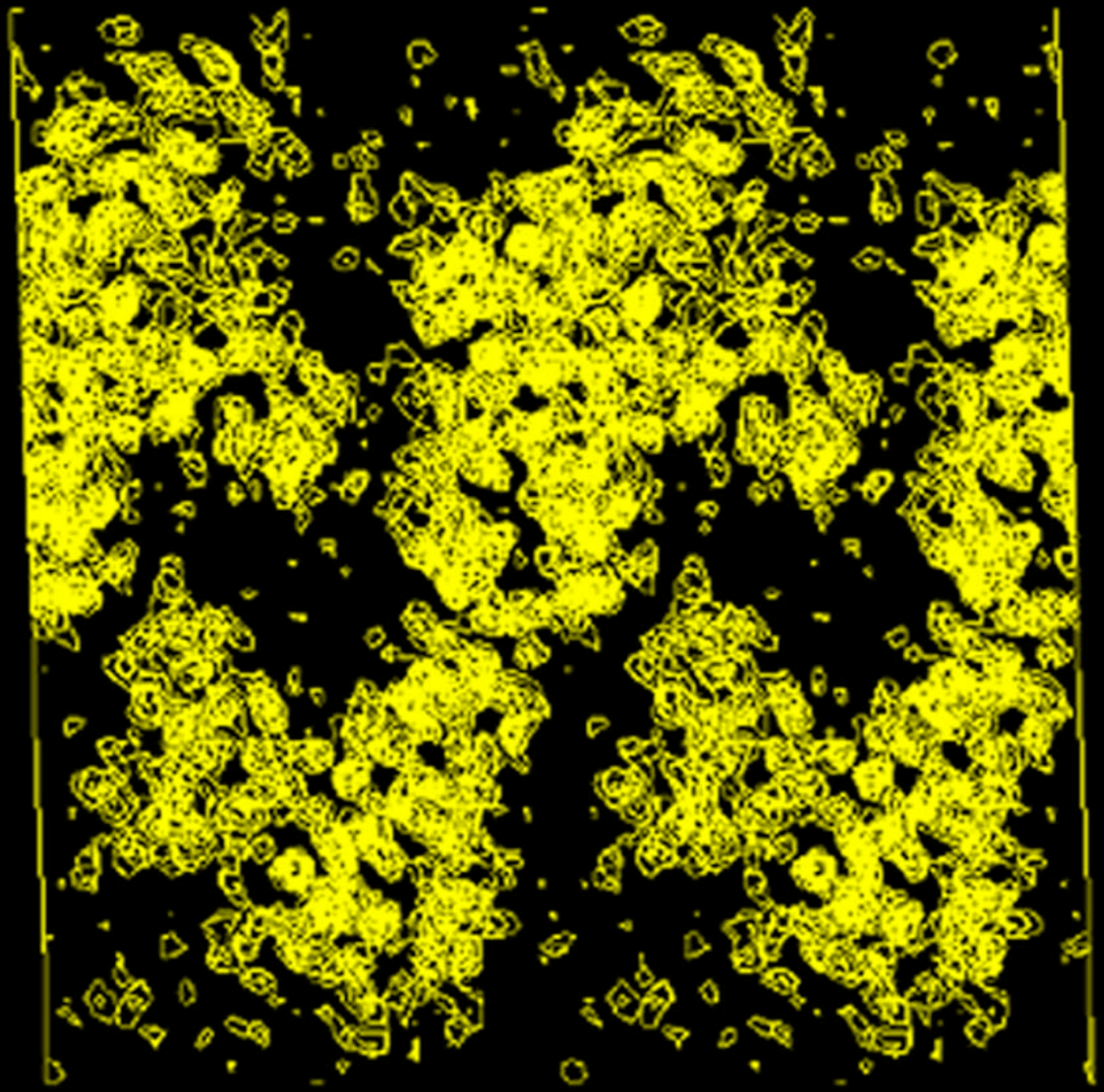
Original Hand



Alternate Hand



Solvent Flattening



Shake-and-Bake Applications: Structure Size and Data Resolution

- **Basic Data (Full Structure)**
  - ~750 unique non-H atoms (equal)
  - ~2000 such atoms including 8 Fe's
  - 1.1-1.2Å data (equal atom)
  - 1.3-1.4Å data (unequal atoms, sometimes)
- **SAS or SIR Difference Data (substructures)**
  - 160 Se (567 kDa / ASU)
  - 3-4Å data
  - 5Å truncated data have also worked

Grid Computing

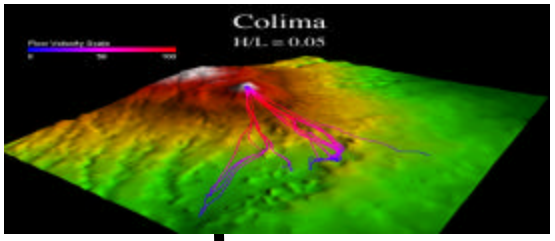
DISCOM

SinRG

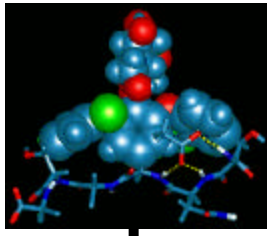
APGrid

IPG ...

Grid Computing Overview



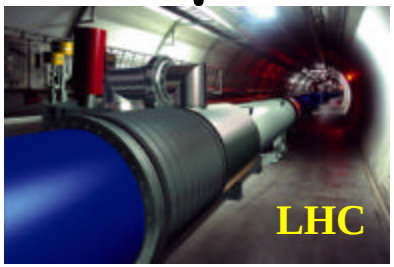
Data Acquisition

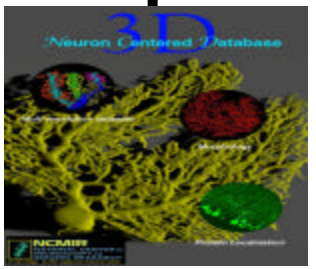


Advanced Visualization

Analysis

Imaging Instruments





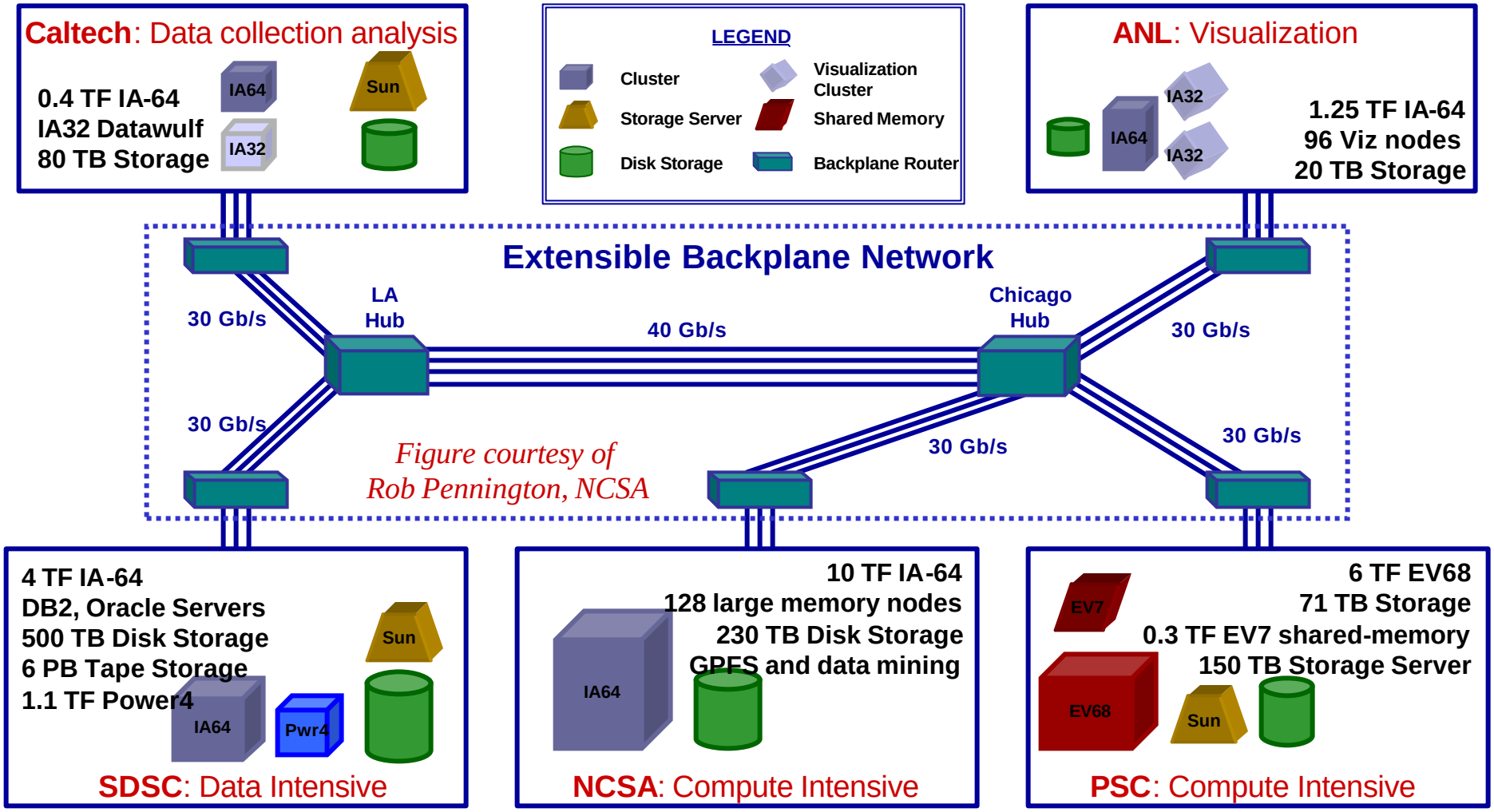
Large-Scale Databases

- 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

Factors Enabling the Grid

- **Internet is Infrastructure**
  - **Increased network bandwidth and advanced services**
- **Advances in Storage Capacity**
  - **Terabyte costs less than $5,000**
- **Internet-Aware Instruments**
- **Increased Availability of Compute Resources**
  - **Clusters, supercomputers, storage, visualization devices**
- **Advances in Application Concepts**
  - **Computational science: simulation and modeling**
  - **Collaborative environments ® large and varied teams**
- **Grids Today**
  - **Moving towards production; Focus on middleware**

NSF Extensible TeraGrid Facility



Computational Grids & Electric Power Grids

- **Similarities/Goals of CG and EPG**
  - **Ubiquitous**
  - **Consumer is comfortable with lack of knowledge of details**
- **Differences Between CG and EPG**
  - **Wider spectrum of performance & services**
  - **Access governed by more complicated issues**
    - **Security**
    - **Performance**
    - **Socio-political factors**

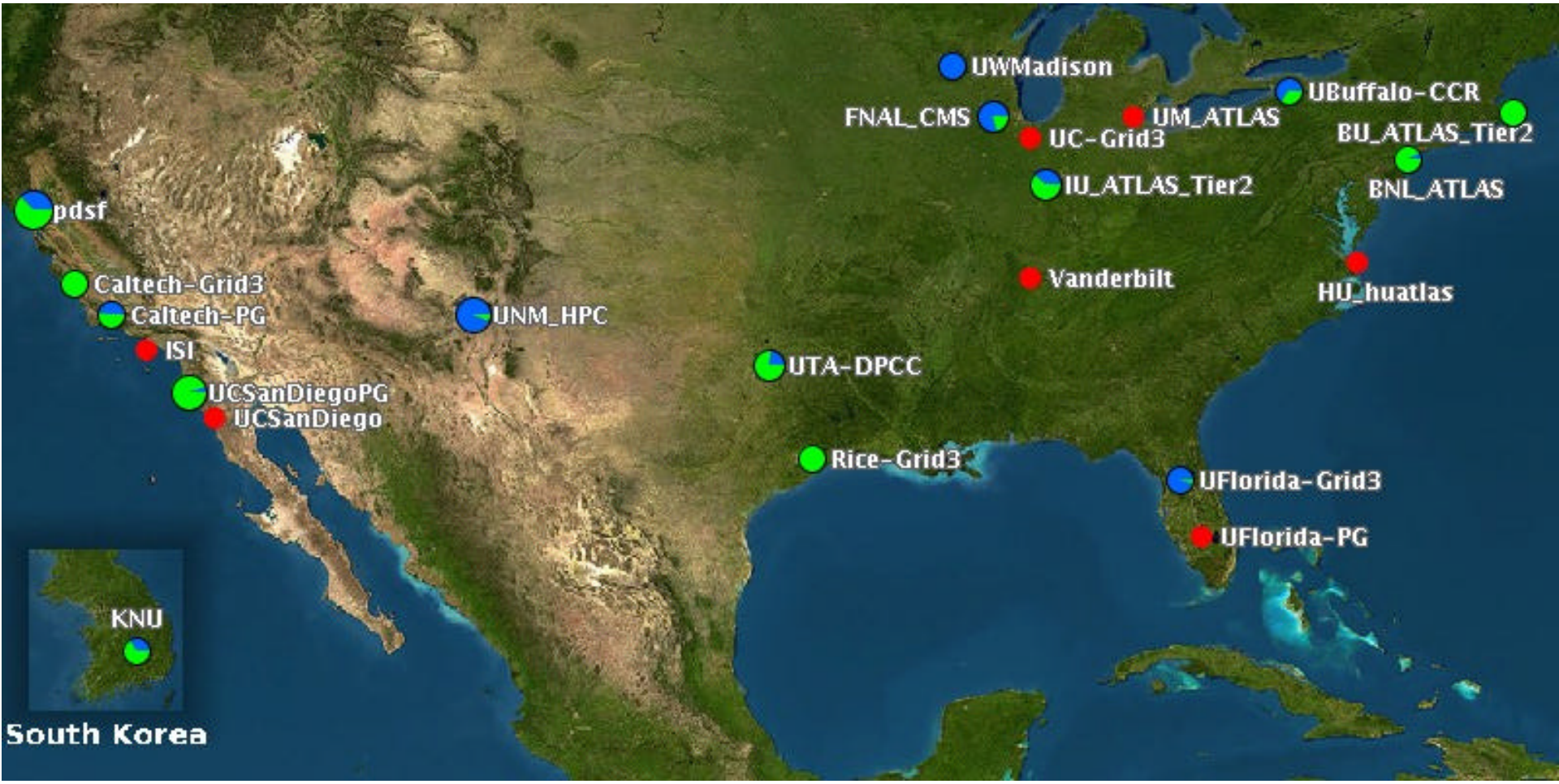
ACDC-Grid Cyber-Infrastructure

- **Integrated Data Grid**
  - **Automated Data File Migration based on profiling users.**
- **Lightweight Grid Monitor (Dashboard)**
- **Predictive Scheduler**
  - **Define quality of service estimates of job completion, by better estimating job runtimes by profiling users.**
- **Dynamic Resource Allocation**
  - **Develop automated procedures for dynamic computational resource allocation.**
- **High-Performance Grid-Enabled Data Repositories**
  - **Develop automated procedures for dynamic data repository creation and deletion.**

ACDC-Grid Collaborations

- High-Performance Networking Infrastructure
- WNY Grid Initiative
- Grid3+ Collaboration
- iVDGL Member
  - Only External Member
- Open Science Grid Member
  - Organizational Committee
  - Blueprint Committee
  - Security Working Group
  - Data Working Group
- Grid-Based Visualization
  - SGI Collaboration
- Grid-Lite: Campus Grid
  - HP Labs Collaboration
- Innovative Laboratory Prototype
  - Dell Collaboration

Grid3 Snapshot of Sites



UBuffalo-CCR Virtual Organization

Grid Resources for Advanced Science and Engineering (GRASE)

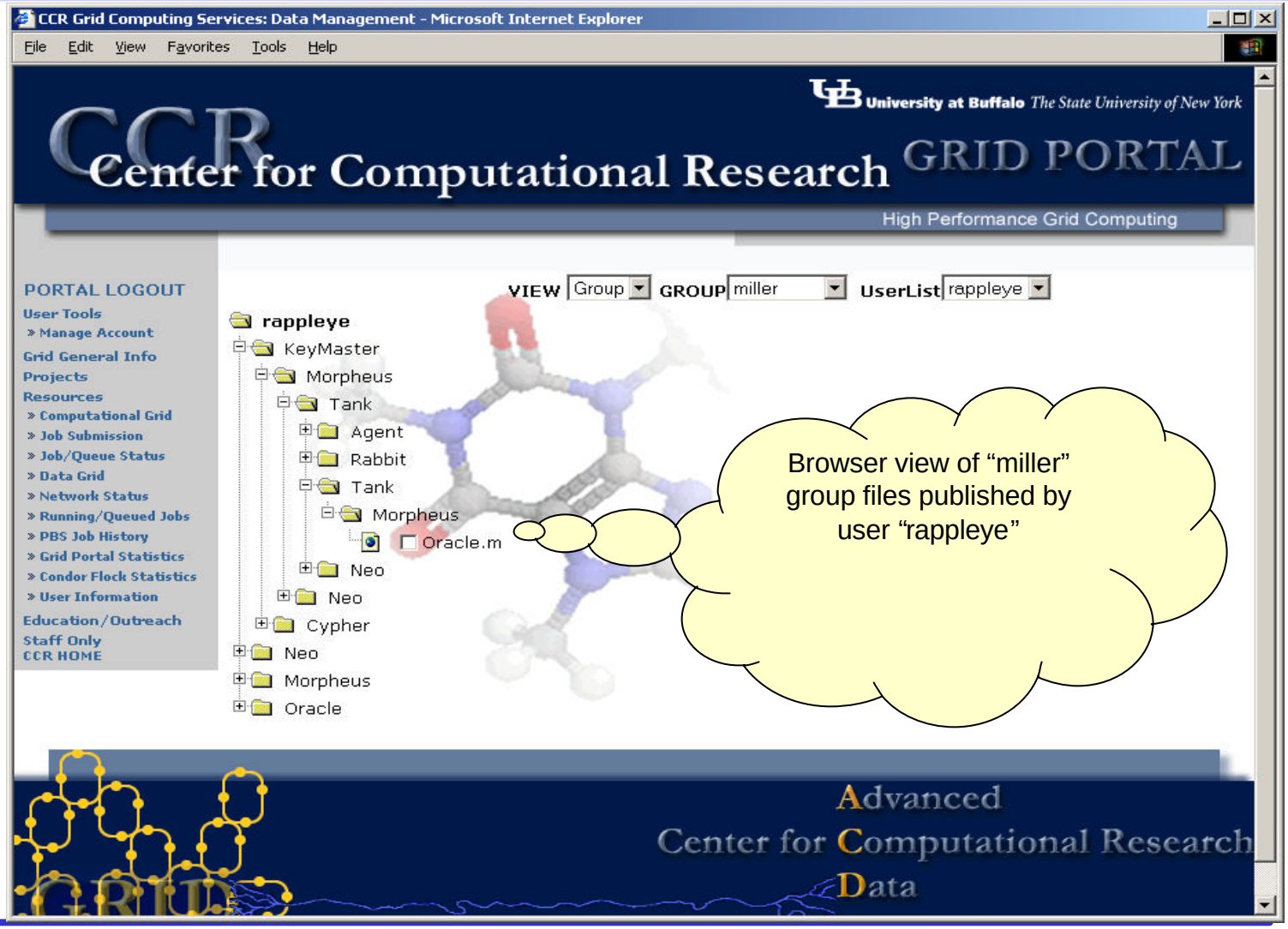
Data Grid

- **Motivation:**
  - **Large data collections are emerging as important community resources.**
  - **Data Grids complement Computational Grids.**
- **Definition:** ***A data grid is a network of distributed storage resources, including archival systems, caches, and databases, which are linked logically to create a sense of global persistence.***
- **Goal: Design and implement transparent management of data distributed across heterogeneous resources.**

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.)

ACDC-Grid Data Grid

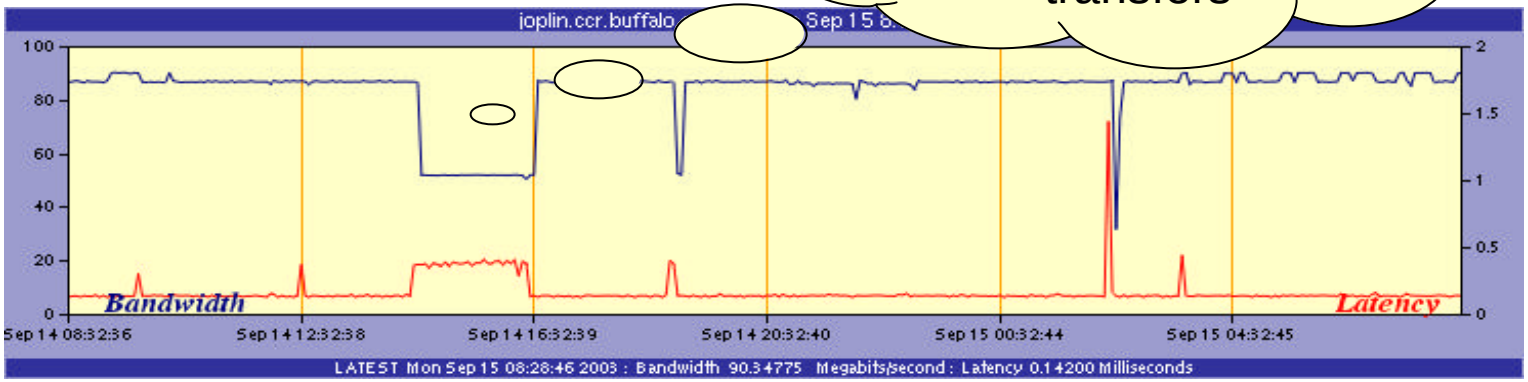


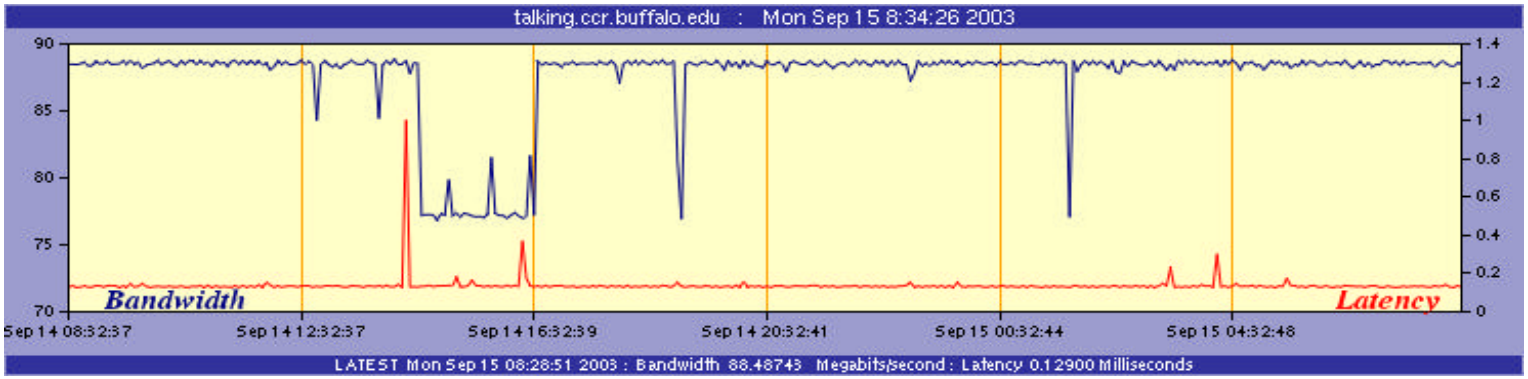
ACDC-Grid
Data Grid File Migration

- **Migration Algorithm dependent on**
  - **User access time**
  - **Network capacity at time of migration**
  - **User profile**
  - **User disk quotas on various resources**

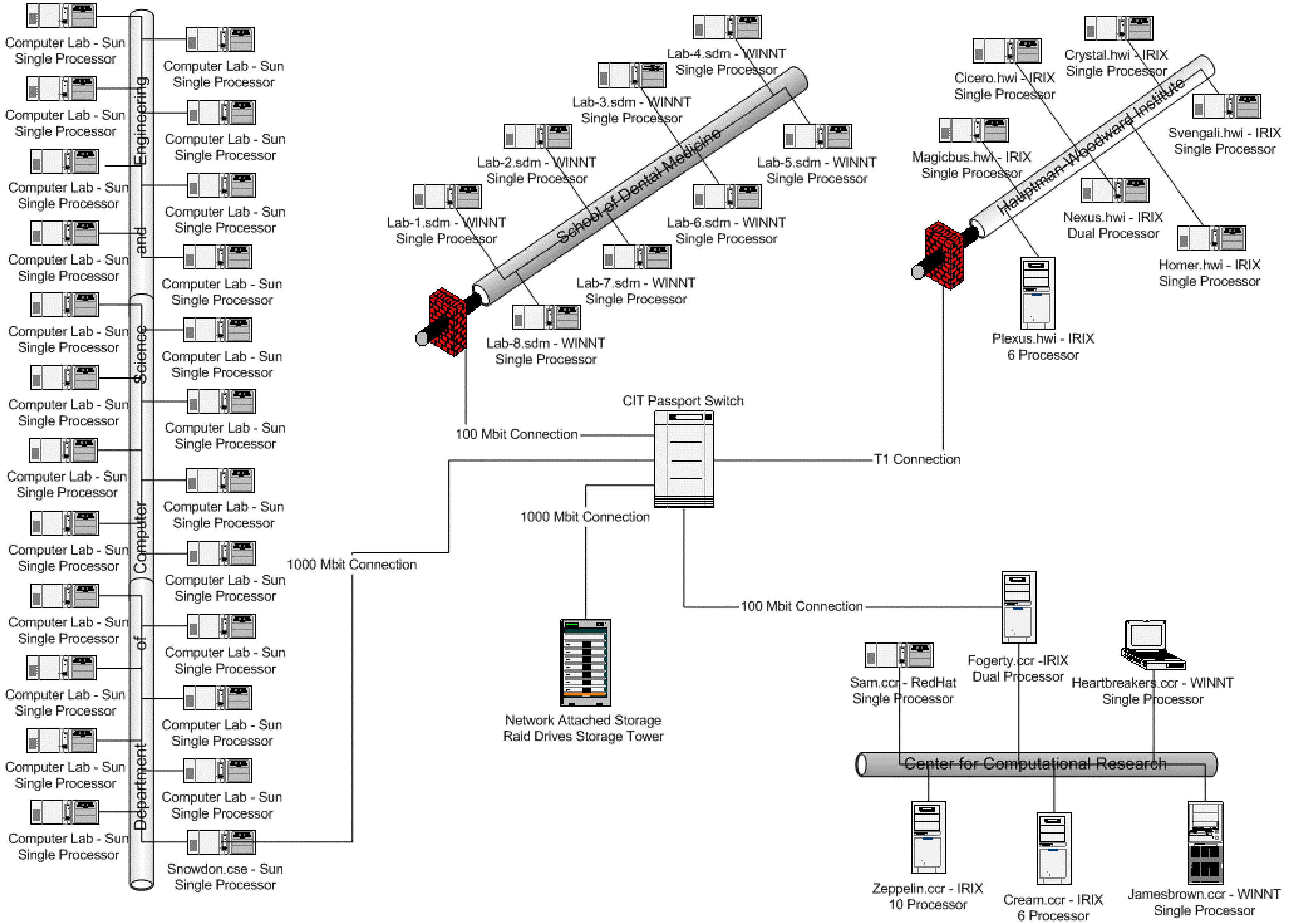
Data Grid Resource Info

Both platforms have reduced bandwidth available for additional transfers

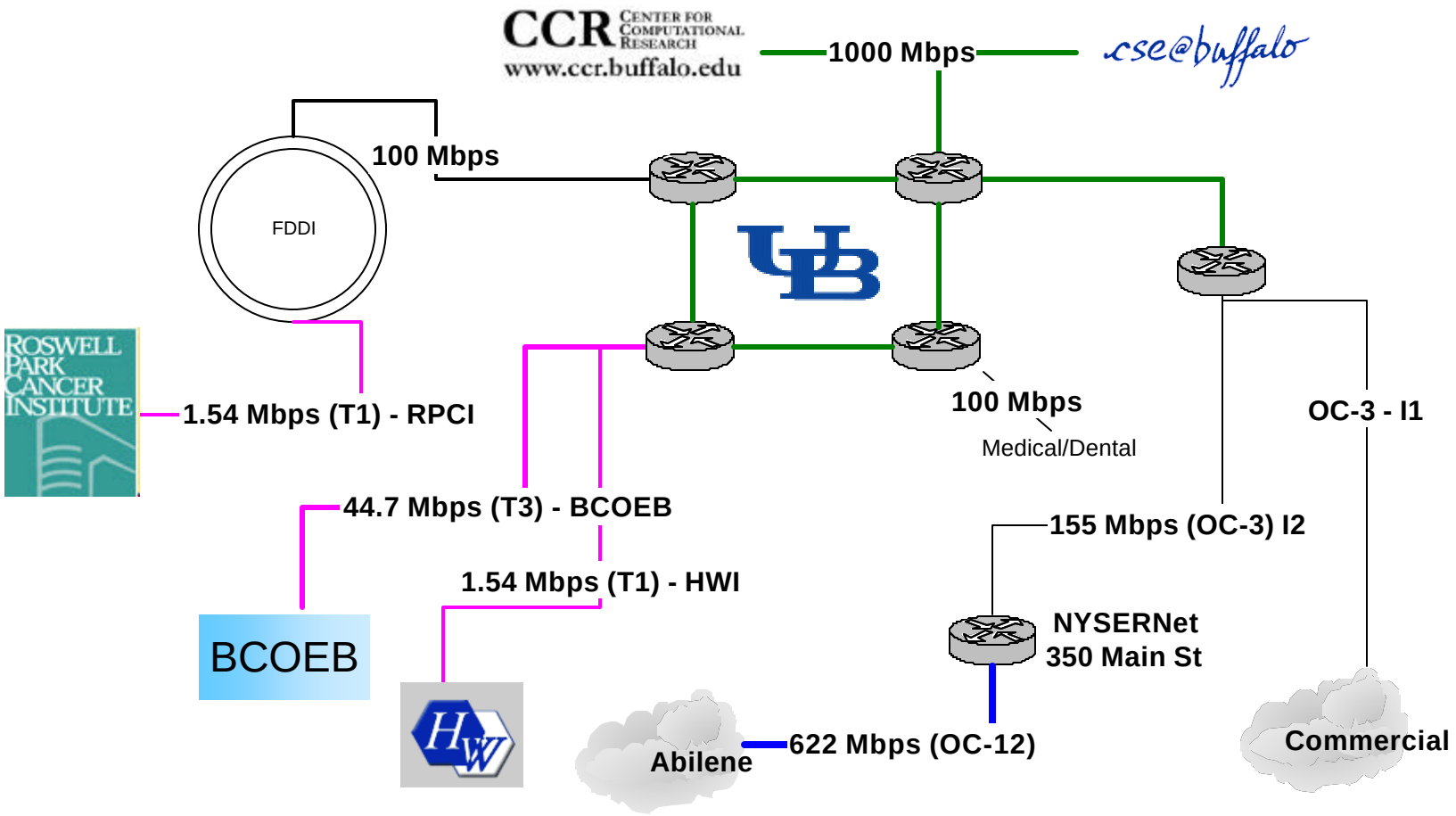




Initial ACDC Campus Grid



Network Connections



ACDC Data Grid Overview
(Grid-Available Data Repositories)

182 GB Storage

Joplin: Compute Cluster
300 Dual Processor
2.4 GHz Intel Xeon
RedHat Linux 7.3
38.7 TB Scratch Space

Nash: Compute Cluster
75 Dual Processor
1 GHz Pentium III
RedHat Linux 7.3
1.8 TB Scratch Space

100 GB Storage

70 GB Storage

Mama: Compute Cluster
9 Dual Processor
1 GHz Pentium III
RedHat Linux 7.3
315 GB Scratch Space

ACDC: Grid Portal
4 Processor Dell 6650
1.6 GHz Intel Xeon
RedHat Linux 9.0
66 GB Scratch Space

56 GB Storage

100 GB Storage

Young: Compute Cluster
16 Dual Sun Blades
47 Sun Ultra5
Solaris 8
770 GB Scratch Space

136 GB Storage

Crosby: Compute Cluster
SGI Origin 3800
64 - 400 MHz IP35
IRIX 6.5.14m
360 GB Scratch Space

CSE Multi-Store
40 TB

Storage Area Network
75 TB

Network Attached Storage
1.2 TB

Fogerty: Condor Flock Master
1 Dual Processor
250 MHz IP30
IRIX 6.5

CCR
19 IRIX, RedHat, & WINNT Processors

T1 Connection

Computer Science & Engineering
25 Single Processor Sun Ultra5s

School of Dental Medicine
9 Single Processor Dell P4 Desktops

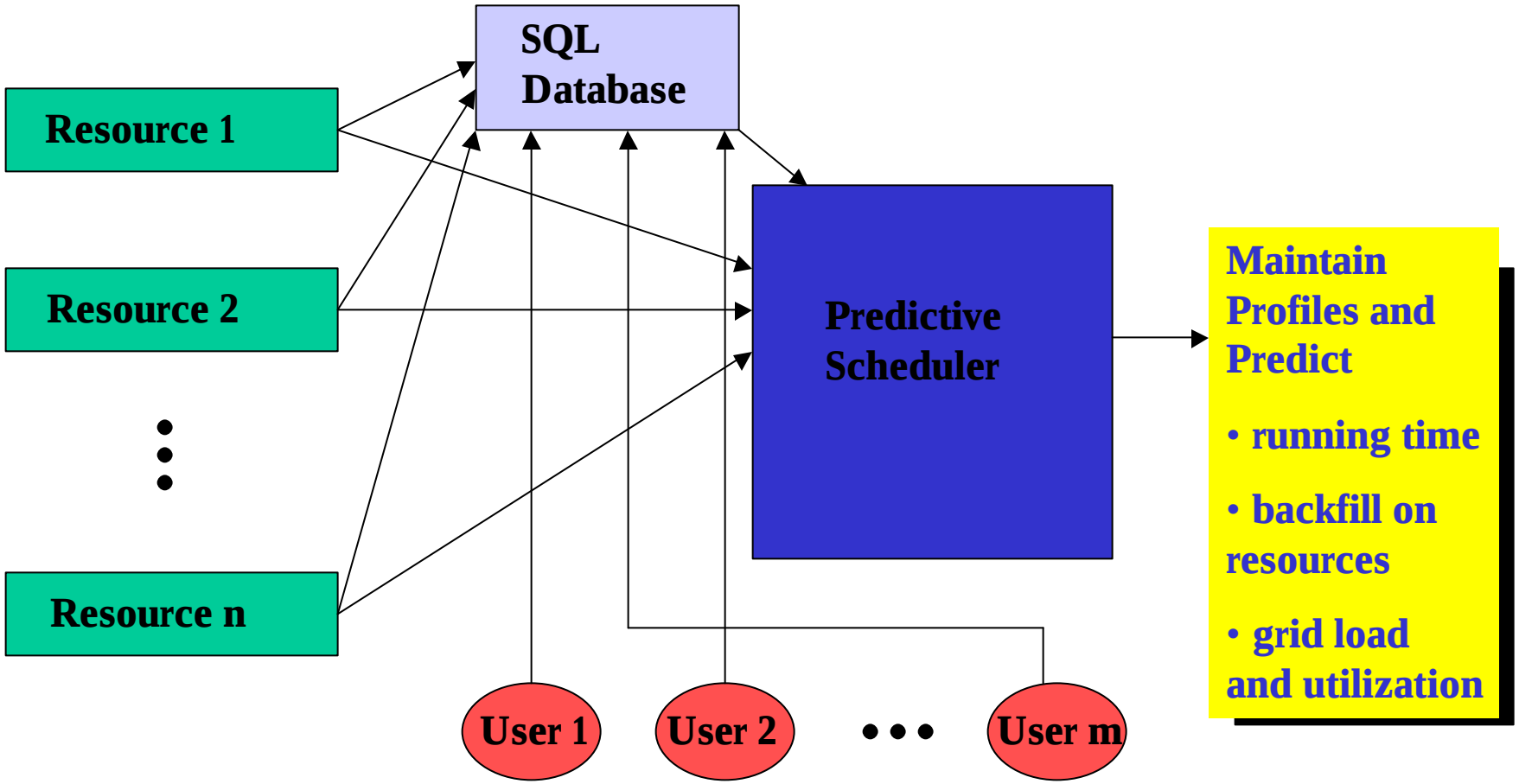
Hauptman-Woodward Institute
13 Various SGI IRIX Processors

Note: Network connections are 100 Mbps unless otherwise noted.

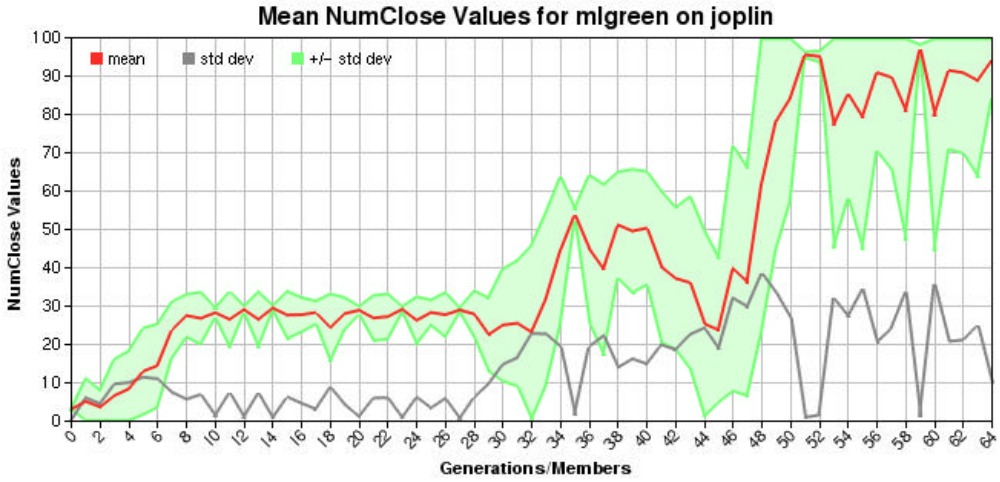
Predictive Scheduler

- 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

System Diagram

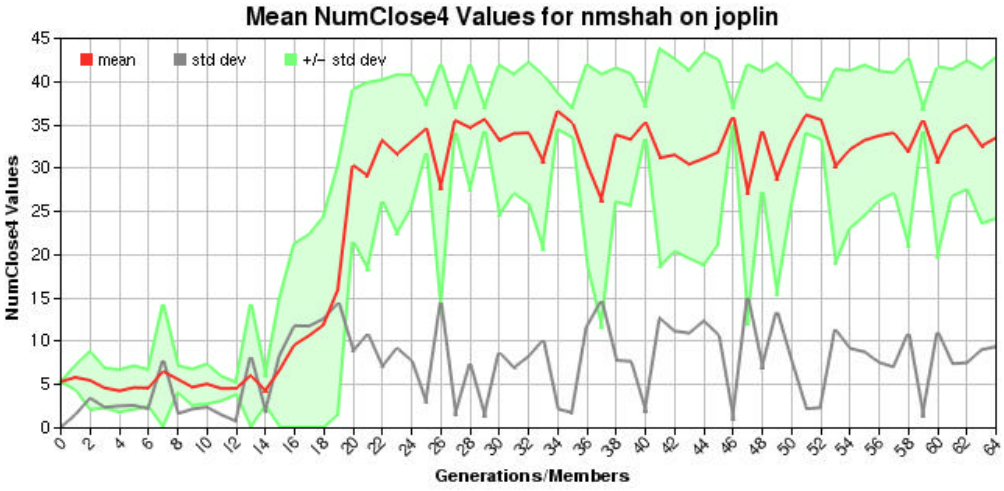


Preliminary GA results



Percent of estimates within 5% of actual values

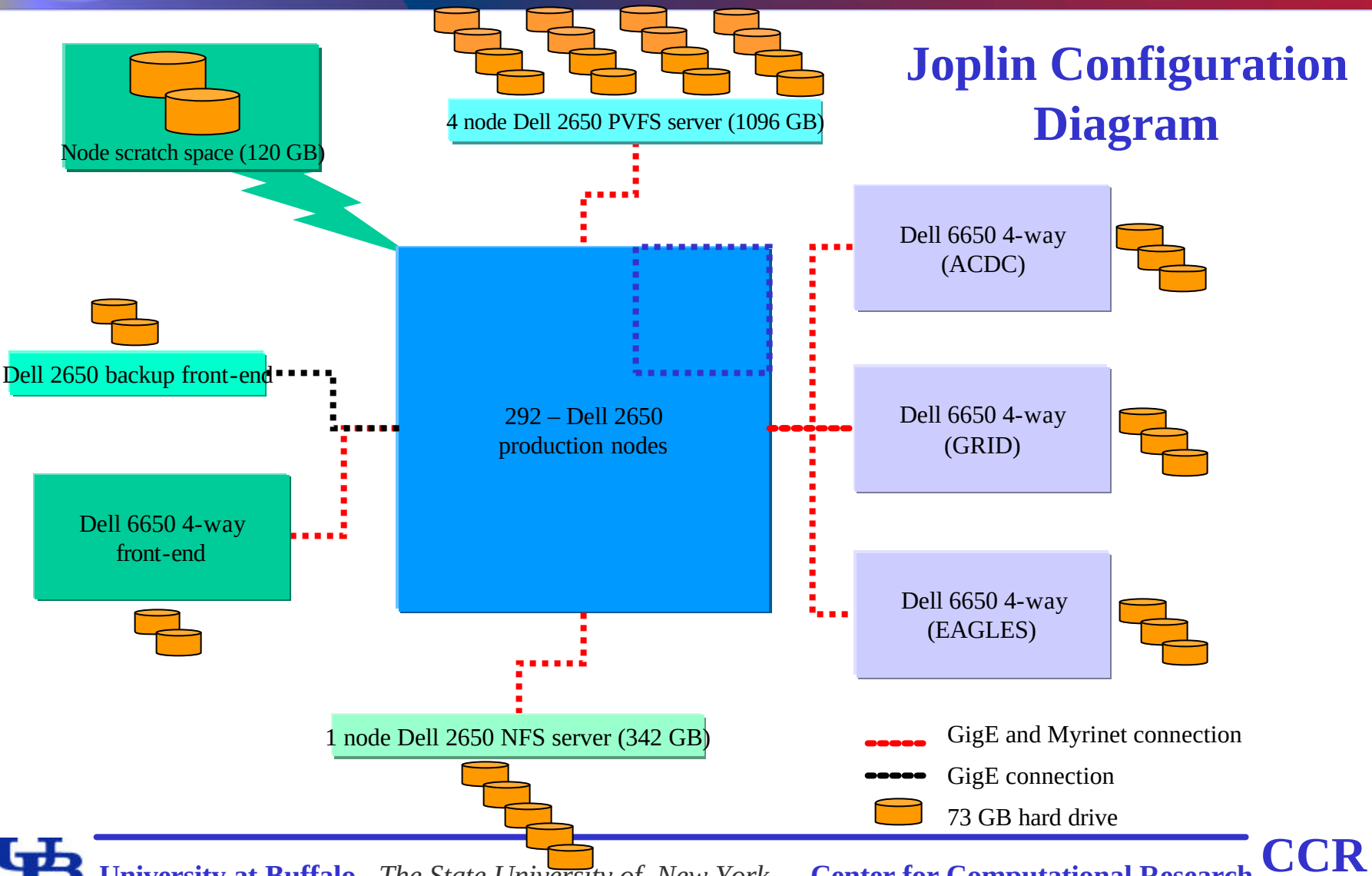
Percent of estimates within 20% of actual values



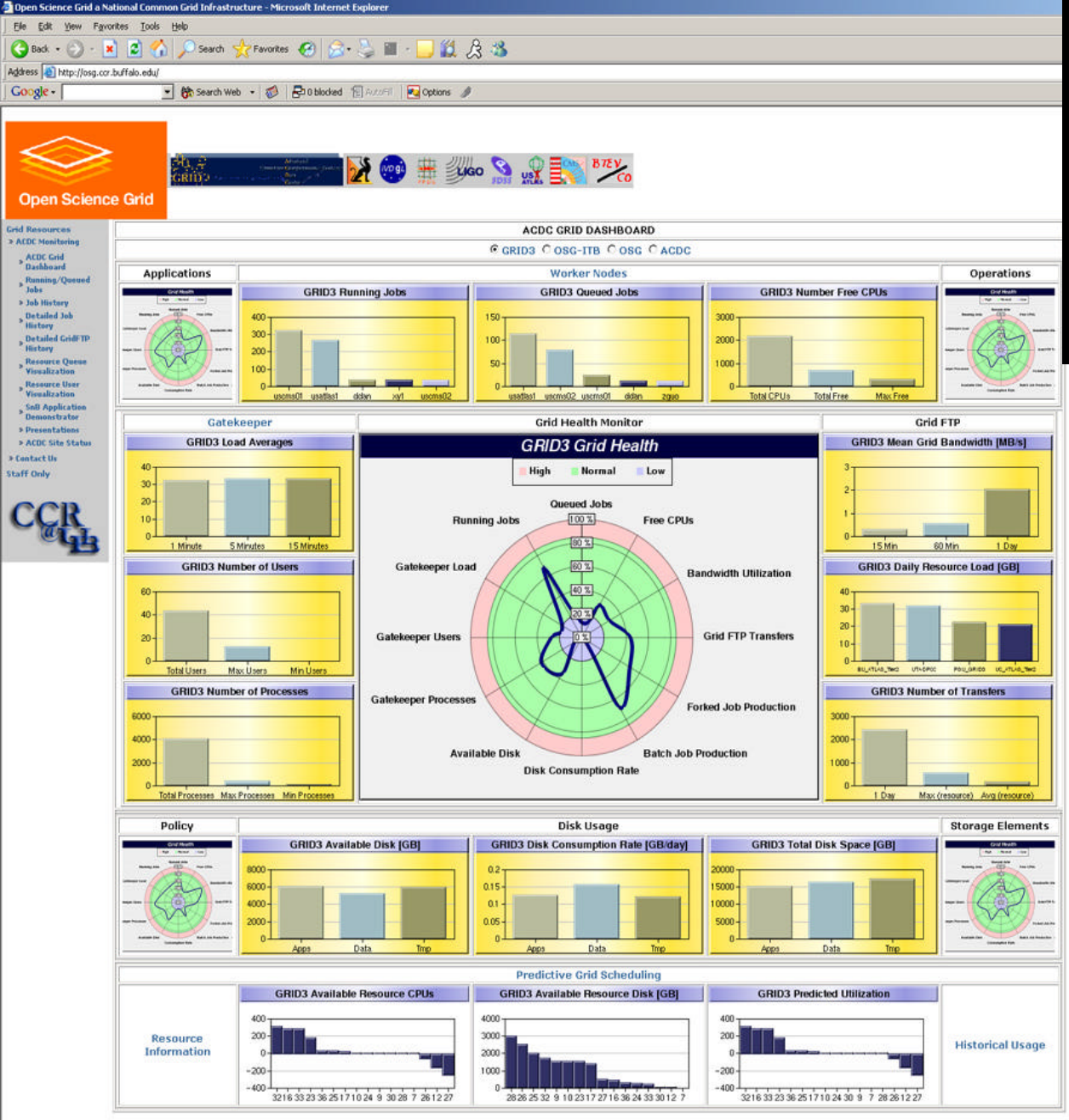
ACDC-Grid Dynamic Resource Allocation at SC03 with Grid3

- 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 Dynamic Resource Allocation

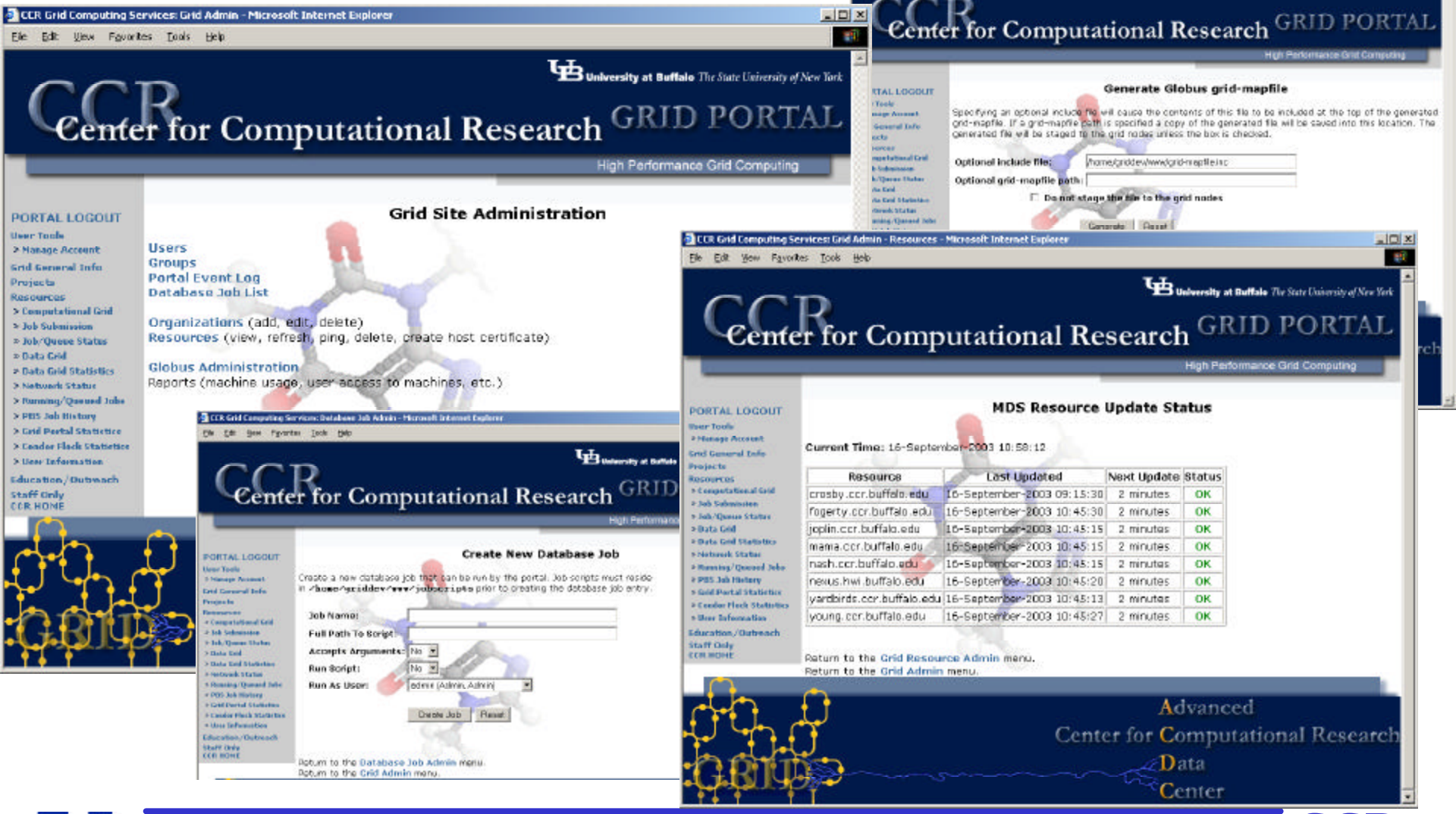


ACDC-Grid Monitoring: The ACDC-Grid DASHBOARD



http://osg.ccr.buffalo.edu

ACDC-Grid Administration



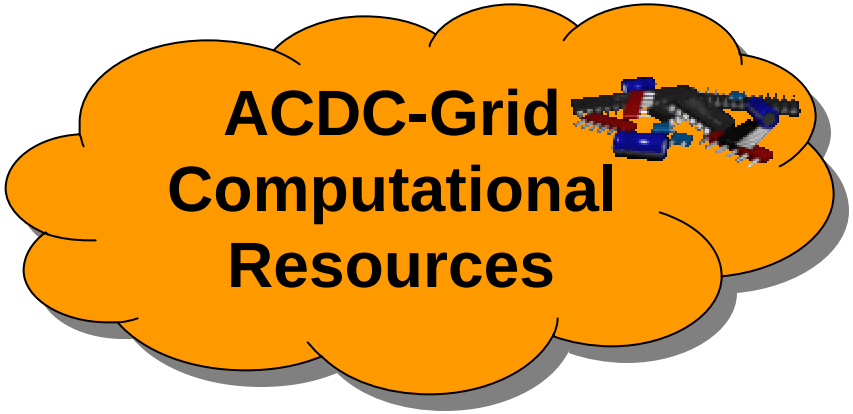
Grid-Enabling Application Templates (GATs)

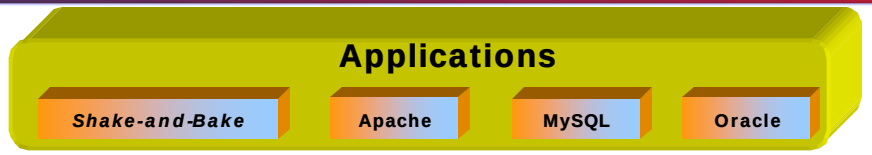
- **Structural Biology**
  - ***SnB* and *BnP* for Molecular Structure Determination/Phasing**
- **Groundwater Modeling**
  - ***Ostrich*: Optimization and Parameter Estimation Tool**
  - ***POMGL*: Princeton Ocean Model Great Lakes for Hydrodynamic Circulation**
  - ***Split*: Modeling Groundwater Flow 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**

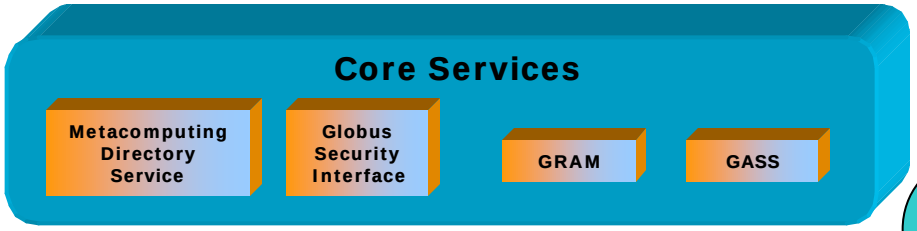
Grid Enabled *SnB*

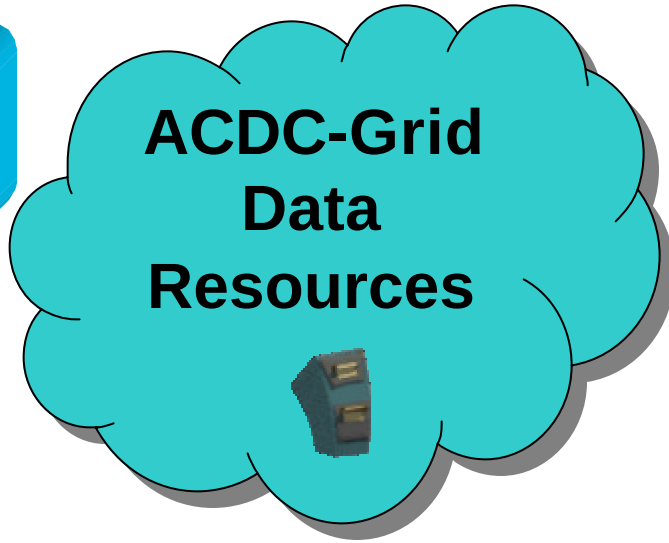
- **Problem Statement**
  - **Use all available resources for determining a single structure**
- **Grid Enabling Criteria**
  - **Run on heterogeneous set of resources**
  - **Store results in *SnB* database**
  - **Mine database (and automagically deploy new jobs) to improve parameter settings**
- **Runtime Parameters Transparent to User**
  - **Assembling Necessary Files**
  - **Number of Processors**
  - **Trials per Processor**
  - **Appropriate Queue and Running Times**

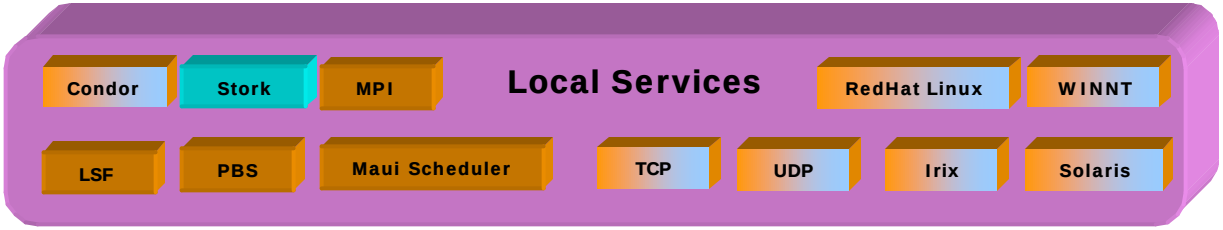
Grid Services and Applications







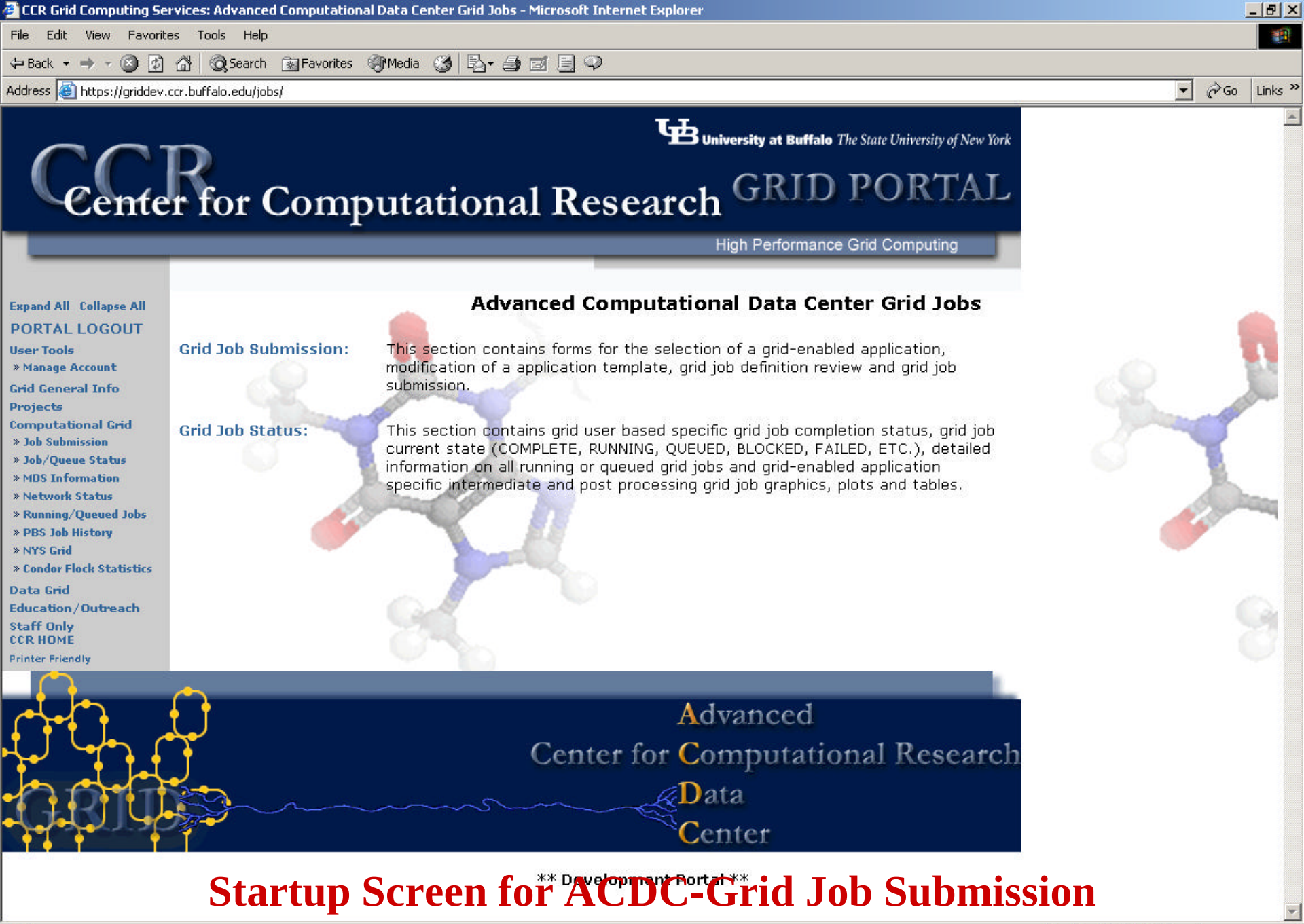




Adapted from Ian Foster and Carl Kesselman

Middleware

- **Grid (Computational and Data)**
  - **Globus Toolkit 2.2.4 ® direct upgrade WSRF**
  - **Condor 6.6.0**
  - **Network Weather Service 2.6**
  - **Apache2 HTTP Server**
  - **PHP 4.3.0**
  - **MySQL 3.23**
  - **phpMyAdmin 2.5.1**
- **Collaboratory**
  - **OpenGL (LibDMS, DevIL, GLUT)**
  - **Windows, IRIX, Mac OS X, Linux**
  - **CAVE, Desktop**



Startup Screen for ACDC-Grid Job Submission

Software → Template → General Information → Detailed Information → Job Definition → Review → Execution Scenario

Advanced Computational Data Center Grid Job Submission Instructions

The grid-enabling application templates used on the ACDC-Grid are created from the application developers grid user profiles that contain the users standard information uid, name, organization, address, etc., and more specific information such as group id and access level information for each of grid-enabled applciations. This information is stored in a database for each of the grid-enabled applications and can be accessed through selected queries throughout the ACDC-Grid Web Portal.

Additionally, each grid-enabled scientific application profile contains information about specific execution parameters, required data files, optional data files, computational requirements, etc. and statistics on application historical ACDC-Grid jobs for predictive runtime estimates. MySQL provides the speed and reliability required for this task and it is currently being used as the ACDC-Grid Web Portal database provider.

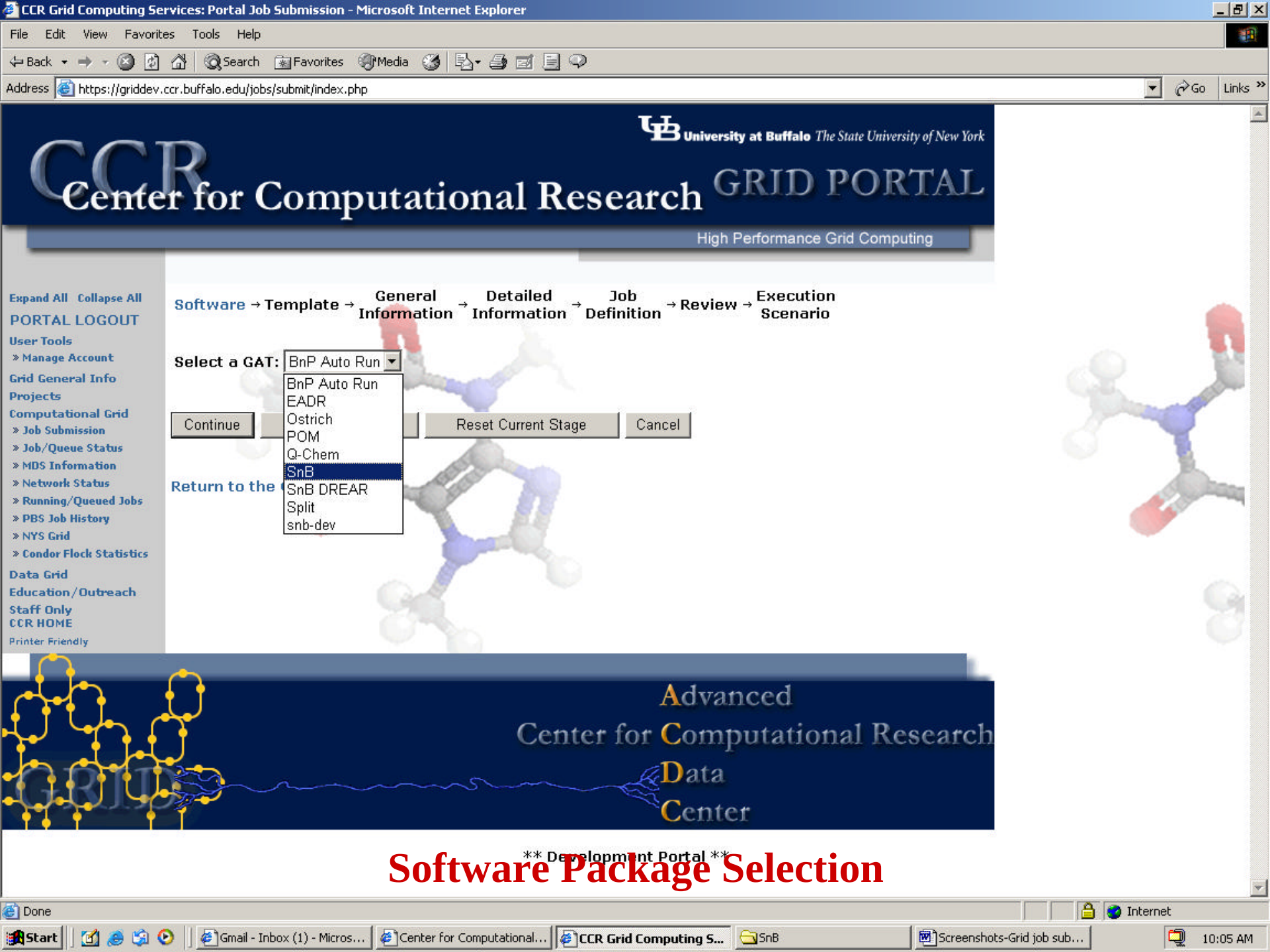
The grid-enabled versions of many well-defined scientific and engineering applications have very similar general requirements and core functionality that are require for execution in the ACDC-Grid environment. We have identified that sequentially defining milestones for the grid user to complete intuitively guides them through the application workflow.

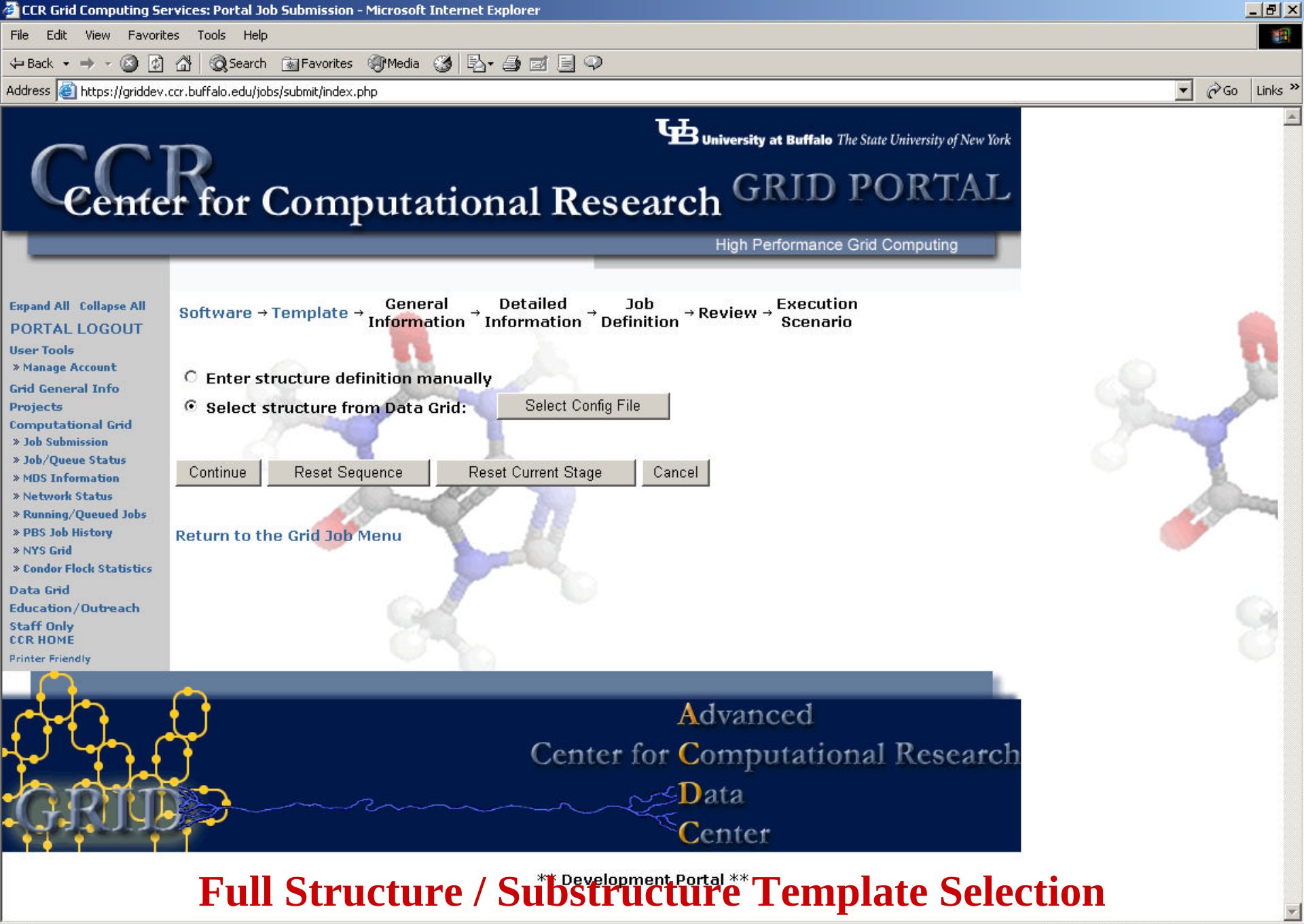
Software Application:	Grid user chooses a grid-enabled software application.
Template:	Grid user selects the required and/or optional data files from the ACDC Data Grid. User defined computational requirements are input or a template defined computational requirement runtime estimate is selected.
Job Definition:	Grid user defines application specific runtime parameters or accepts default template parameter definitions.
Review:	Grid user accepts the template complete job definition workflow or corrects any part of job definition.
Execution Scenario:	The grid user has the ability to input an execution scenario or select a ACDC-Grid determined template defined execution scenario.
Grid Job Status:	The grid user can view specific grid job completion status, grid job current state (COMPLETE, RUNNING, QUEUED, BLOCKED, FAILED, ETC.), detailed information on all running or queued grid jobs and grid-enabled application specific intermediate and post processing grid job graphics, plots and tables.

Each item of the job definition workflow is then stored in the ACDC-Grid Web Portal database so the grid user may use/modify any previously created workflow in creating new job definitions. The job definitions can also be accessed via batch script files for executing hundreds of similar workflows in an automated fashion. For example, a grid user would first define/save a relatively generic job workflow template for the grid-enabled application and then use the batch script capabilities to change the job definition workflow data files or application parameters and execute a series of new grid jobs.

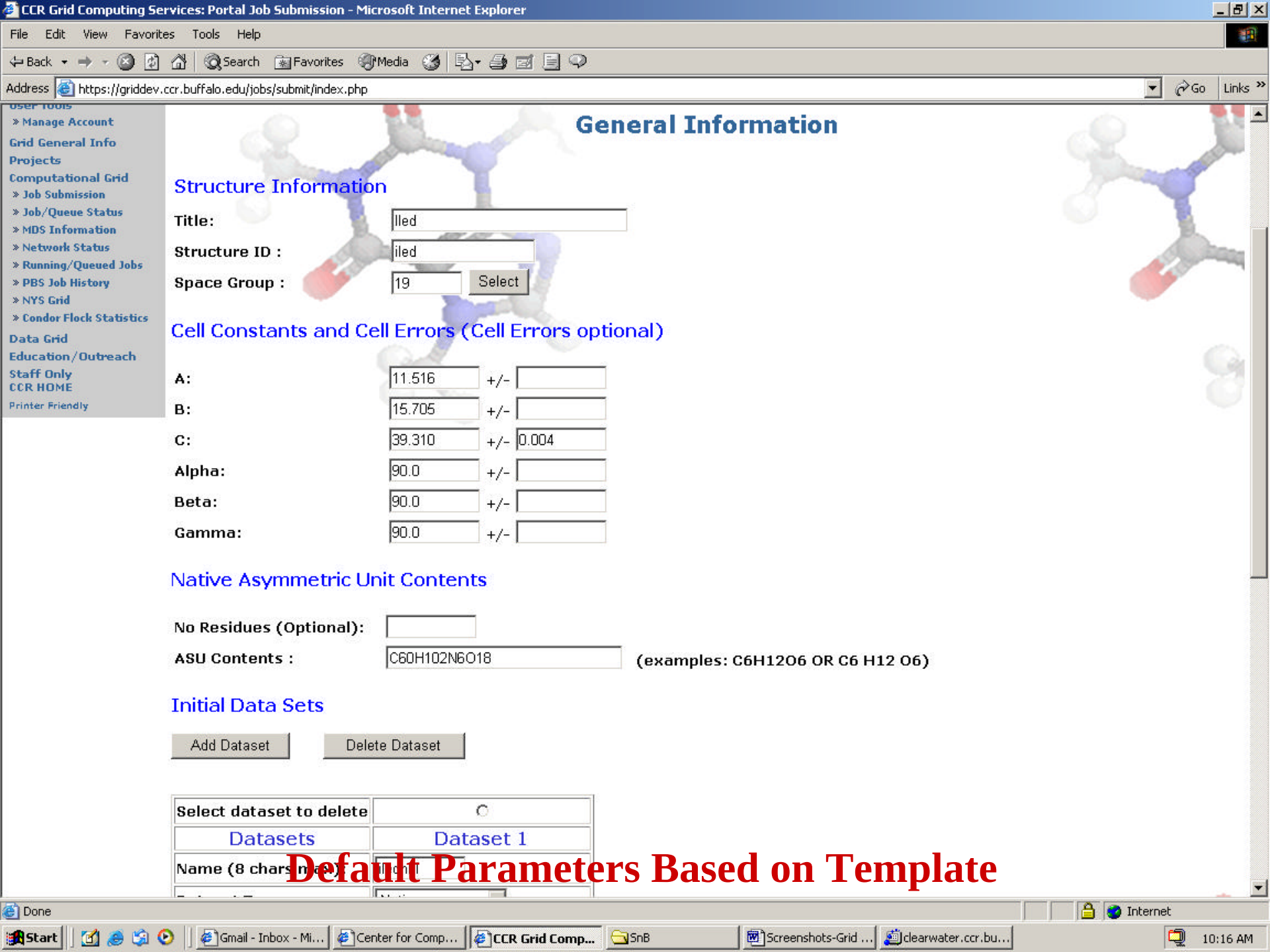
Continue Reset Sequence Reset Current Stage Cancel

Instructions and Description for Running a Job on ACDC-Grid


CCR Grid Computing Services: Portal Job Submission - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address https://griddev.ccr.buffalo.edu/jobs/submit/index.php
Go Links
University at Buffalo The State University of New York
CCR
Center for Computational Research
GRID PORTAL
High Performance Grid Computing
Expand All Collapse All
PORTAL LOGOUT
User Tools
» Manage Account
Grid General Info
Projects
Computational Grid
» Job Submission
» Job/Queue Status
» MDS Information
» Network Status
» Running/Queued Jobs
» PBS Job History
» NYS Grid
» Condor Flock Statistics
Data Grid
Education/Outreach
Staff Only
CCR HOME
Printer Friendly
Software → Template → General Information → Detailed Information → Job Definition → Review → Execution Scenario
Select a GAT: BnP Auto Run
BnP Auto Run
EADR
Ostrich
POM
Q-Chem
SnB
SnB DREAR
Split
snb-dev
Continue
Reset Current Stage
Cancel
Return to the
Advanced
Center for Computational Research
Data
Center
** Development Portal **
Software Package Selection
Done
Internet
Start
Gmail - Inbox (1) - Micros...
Center for Computational...
CCR Grid Computing S...
SnB
Screenshots-Grid job sub...
10:05 AM



Full Structure / Substructure Template Selection


CCR Grid Computing Services: Portal Job Submission - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address https://griddev.ccr.buffalo.edu/jobs/submit/index.php
Go Links
User Tools
» Manage Account
Grid General Info
Projects
Computational Grid
» Job Submission
» Job/Queue Status
» MDS Information
» Network Status
» Running/Queued Jobs
» PBS Job History
» NYS Grid
» Condor Flock Statistics
Data Grid
Education/Outreach
Staff Only
CCR HOME
Printer Friendly
General Information
Structure Information
Title: lled
Structure ID : iled
Space Group : 19 Select
Cell Constants and Cell Errors (Cell Errors optional)
A: 11.516 +/-
B: 15.705 +/-
C: 39.310 +/- 0.004
Alpha: 90.0 +/-
Beta: 90.0 +/-
Gamma: 90.0 +/-
Native Asymmetric Unit Contents
No Residues (Optional):
ASU Contents : C60H102N6O18 (examples: C6H12O6 OR C6 H12 O6)
Initial Data Sets
Add Dataset
Delete Dataset
Select dataset to delete
Datasets
Dataset 1
Name (8 chars max)
Default Parameters Based on Template
Done
Internet
Start
Gmail - Inbox - Mi...
Center for Comp...
CCR Grid Comp...
SnB
Screenshots-Grid ...
clearwater.ccr.bu...
10:16 AM

CCR Grid Computing Services: Portal Job Submission - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address https://griddev.ccr.buffalo.edu/jobs/submit/index.php Go Links

Initial Data Sets

Add Dataset | Delete Dataset

Select dataset to delete	
Datasets	Dataset 1
Name (8 chars max):	iledhkl
Dataset Type:	Native
File Name (*.hkl) :	Browse
File Type:	F, Sig(F)
Wavelength:	1.5418
Max. Resolution:	0.94
Anomalous Dispersion:	Not Measured
Heavy Element Type:	
Nat. Element Replaced:	
No. Expected Sites:	
F Prime (f'):	
F Double Prime (f''):	

Continue | Reset Sequence | Reset Current Stage | Cancel

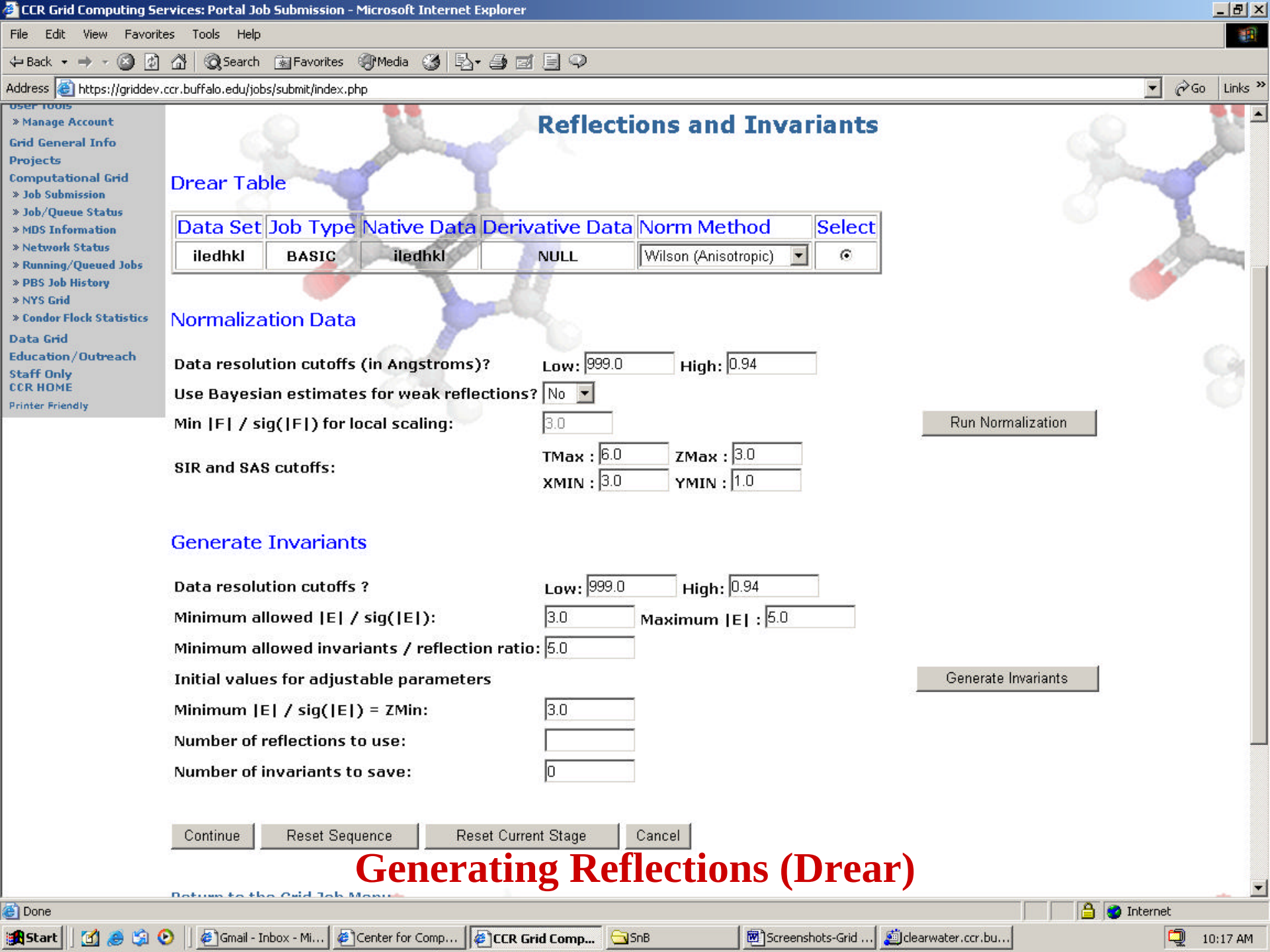
Return to the Grid Job Menu

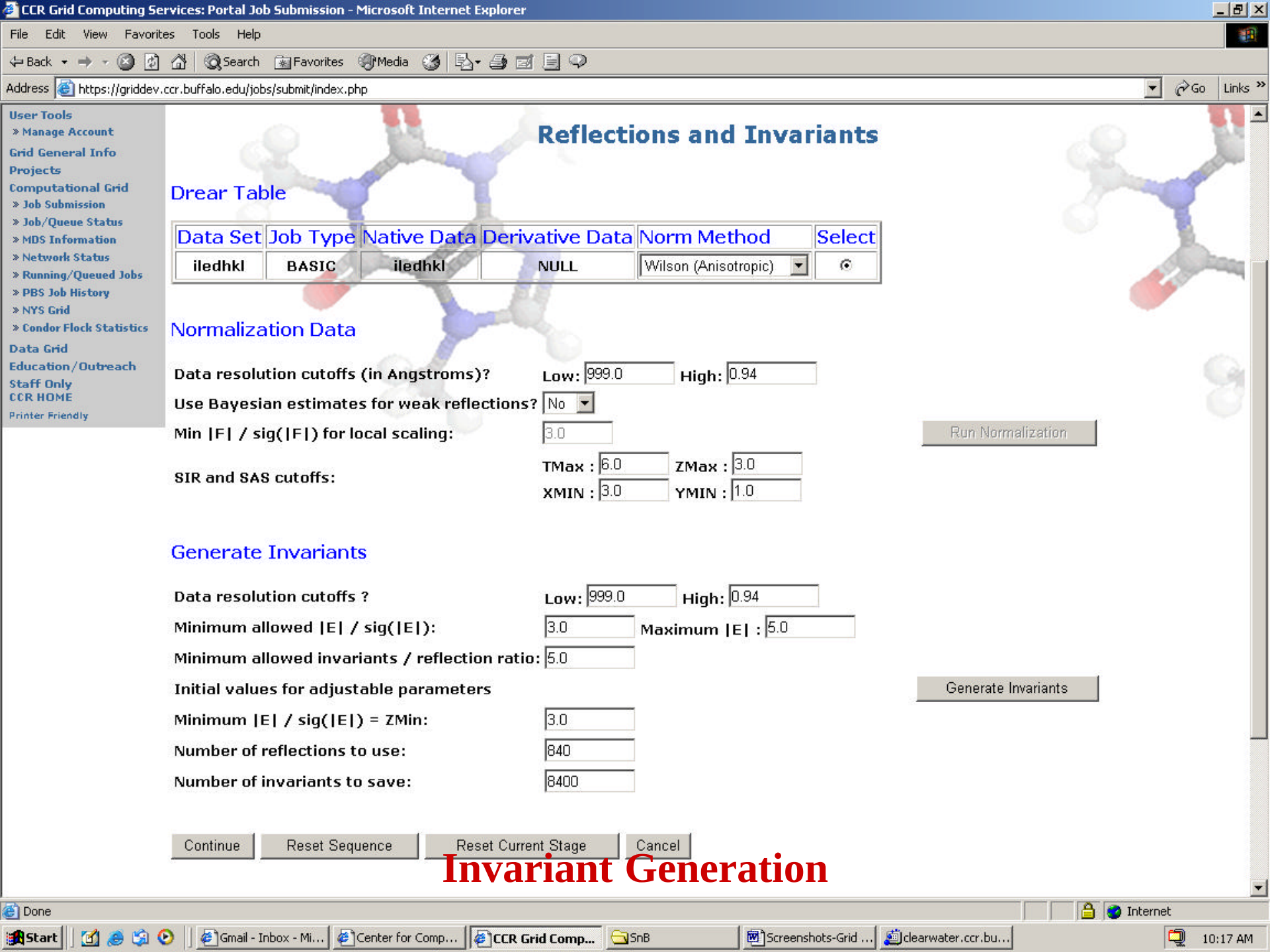
Default Parameters (cont'd)

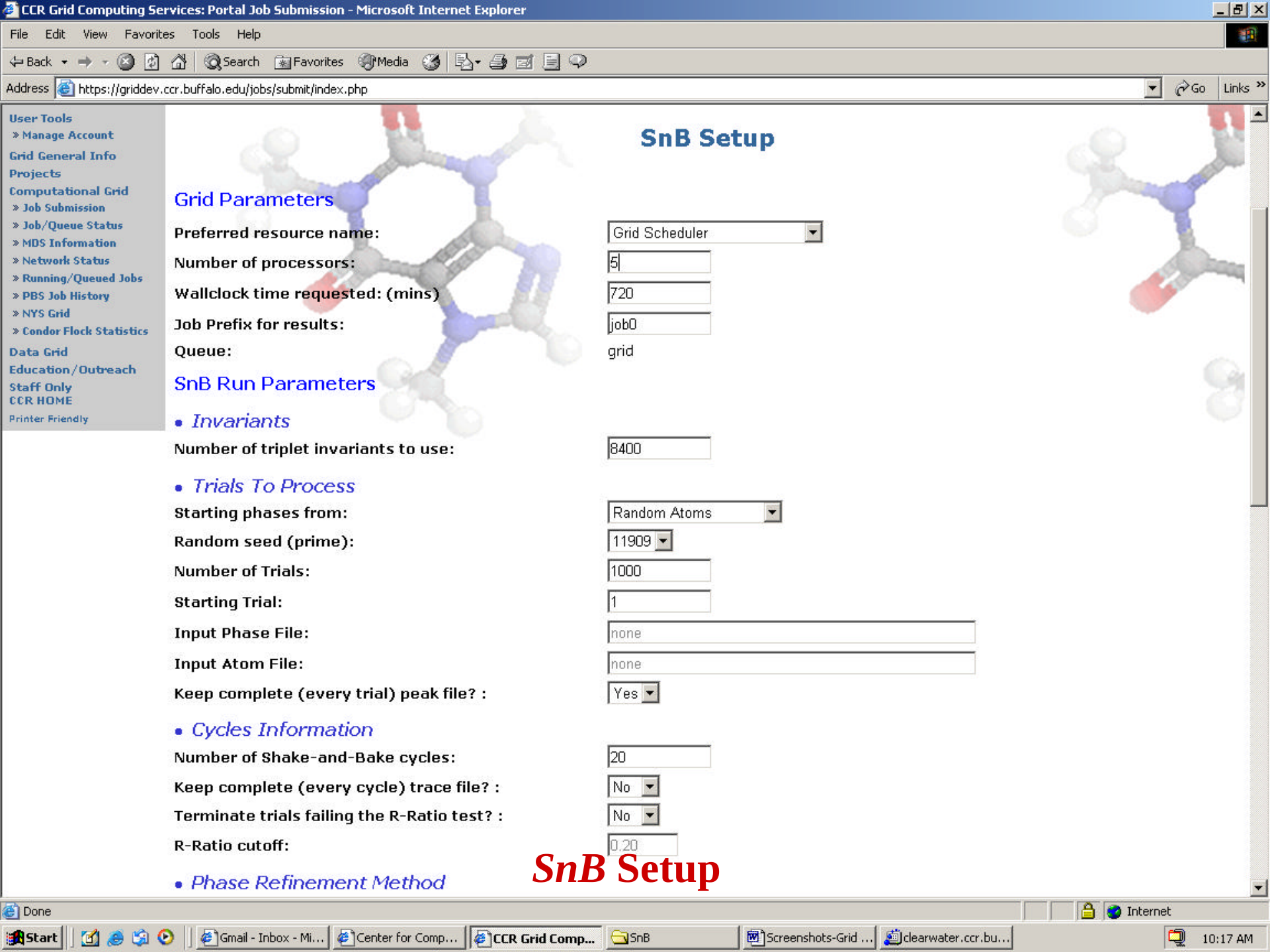
Advanced Center for Computational Research

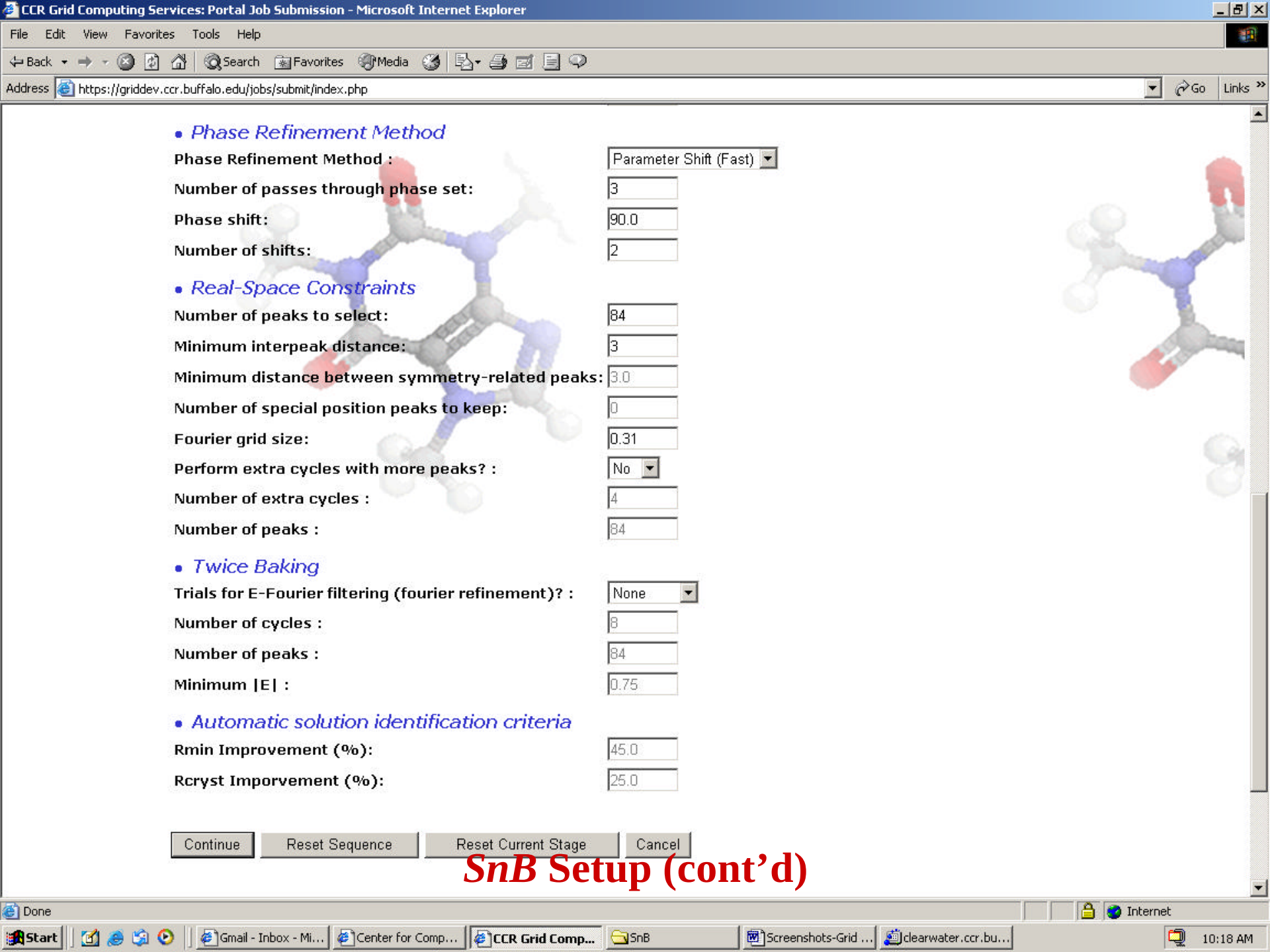
Done Internet

Start | Gmail - Inbox - Mi... | Center for Comp... | CCR Grid Comp... | SnB | Screenshots-Grid ... | clearwater.ccr.bu... | 10:16 AM


CCR Grid Computing Services: Portal Job Submission - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address https://griddev.ccr.buffalo.edu/jobs/submit/index.php
Go Links
» Manage Account
Grid General Info
Projects
Computational Grid
» Job Submission
» Job/Queue Status
» MDS Information
» Network Status
» Running/Queued Jobs
» PBS Job History
» NYS Grid
» Condor Flock Statistics
Data Grid
Education/Outreach
Staff Only
CCR HOME
Printer Friendly
Reflections and Invariants
Drear Table
Data Set
Job Type
Native Data
Derivative Data
Norm Method
Select
iledhkl
BASIC
iledhkl
NULL
Wilson (Anisotropic)
Normalization Data
Data resolution cutoffs (in Angstroms)?
Low: 999.0
High: 0.94
Use Bayesian estimates for weak reflections?
No
Min |F| / sig(|F|) for local scaling:
3.0
Run Normalization
SIR and SAS cutoffs:
TMax : 6.0
ZMax : 3.0
XMIN : 3.0
YMIN : 1.0
Generate Invariants
Data resolution cutoffs ?
Low: 999.0
High: 0.94
Minimum allowed |E| / sig(|E|):
3.0
Maximum |E| : 5.0
Minimum allowed invariants / reflection ratio:
5.0
Initial values for adjustable parameters
Generate Invariants
Minimum |E| / sig(|E|) = ZMin:
3.0
Number of reflections to use:
Number of invariants to save:
0
Continue
Reset Sequence
Reset Current Stage
Cancel
Generating Reflections (Drear)
Done
Internet
Start
Gmail - Inbox - Mi...
Center for Comp...
CCR Grid Comp...
SnB
Screenshots-Grid ...
clearwater.ccr.bu...
10:17 AM


CCR Grid Computing Services: Portal Job Submission - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address https://griddev.ccr.buffalo.edu/jobs/submit/index.php
Go Links
User Tools
» Manage Account
Grid General Info
Projects
Computational Grid
» Job Submission
» Job/Queue Status
» MDS Information
» Network Status
» Running/Queued Jobs
» PBS Job History
» NYS Grid
» Condor Flock Statistics
Data Grid
Education/Outreach
Staff Only
CCR HOME
Printer Friendly
Reflections and Invariants
Drear Table
Data Set
Job Type
Native Data
Derivative Data
Norm Method
Select
iledhkl
BASIC
iledhkl
NULL
Wilson (Anisotropic)
Normalization Data
Data resolution cutoffs (in Angstroms)?
Low: 999.0
High: 0.94
Use Bayesian estimates for weak reflections?
No
Min |F| / sig(|F|) for local scaling:
3.0
Run Normalization
SIR and SAS cutoffs:
TMax : 6.0
ZMax : 3.0
XMIN : 3.0
YMIN : 1.0
Generate Invariants
Data resolution cutoffs ?
Low: 999.0
High: 0.94
Minimum allowed |E| / sig(|E|):
3.0
Maximum |E| : 5.0
Minimum allowed invariants / reflection ratio:
5.0
Initial values for adjustable parameters
Generate Invariants
Minimum |E| / sig(|E|) = ZMin:
3.0
Number of reflections to use:
840
Number of invariants to save:
8400
Continue
Reset Sequence
Reset Current Stage
Cancel
Invariant Generation
Done
Internet
Start
Gmail - Inbox - Mi...
Center for Comp...
CCR Grid Comp...
SnB
Screenshots-Grid ...
clearwater.ccr.bu...
10:17 AM


CCR Grid Computing Services: Portal Job Submission - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address https://griddev.ccr.buffalo.edu/jobs/submit/index.php
Go Links
User Tools
» Manage Account
Grid General Info
Projects
Computational Grid
» Job Submission
» Job/Queue Status
» MDS Information
» Network Status
» Running/Queued Jobs
» PBS Job History
» NYS Grid
» Condor Flock Statistics
Data Grid
Education/Outreach
Staff Only
CCR HOME
Printer Friendly
SnB Setup
Grid Parameters
Preferred resource name: Grid Scheduler
Number of processors: 5
Wallclock time requested: (mins) 720
Job Prefix for results: job0
Queue: grid
SnB Run Parameters
• Invariants
Number of triplet invariants to use: 8400
• Trials To Process
Starting phases from: Random Atoms
Random seed (prime): 11909
Number of Trials: 1000
Starting Trial: 1
Input Phase File: none
Input Atom File: none
Keep complete (every trial) peak file? : Yes
• Cycles Information
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: 0.20
• Phase Refinement Method
SnB Setup
Done
Internet
Start
Gmail - Inbox - Mi...
Center for Comp...
CCR Grid Comp...
SnB
Screenshots-Grid ...
clearwater.ccr.bu...
10:17 AM



SnB Setup (cont'd)

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)

Details for Grid Job 447 - iledhkl

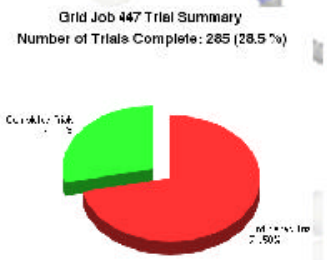
Job Detail Information

Status: RUNNING
Rmin Min: 0.344 **Rmin Max:** 0.56
Last Updated: 15-Mar-2005 10:22:00

Total Trials: 1000
Complete Trials: 285
Resource: clearwater.ccr.buffalo.edu **Processors:** 5

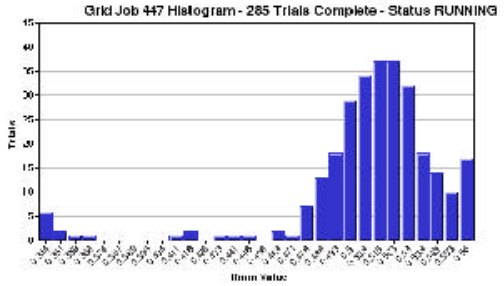
Best Trial Number: 34
Best Trial Rmin: 0.344

Trial Summary



Walltime Summary

Grid Job Trial Histogram



Click on image for enlarged view.

Refresh Graphical Job Status
Return to the Grid Job Menu

Graphical Representation of Intermediate Job Status

CCR Grid Computing Services: Grid Job Status Detail - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address https://griddev.ccr.buffalo.edu/jobs/submit/snb/display_chart.php?id=447&gat=snb&chart=rmin_histogram

Go Links

University at Buffalo The State University of New York

CCR Center for Computational Research GRID PORTAL

High Performance Grid Computing

Expand All Collapse All

PORTAL LOGOUT

User Tools
» Manage Account

Grid General Info

Projects

Computational Grid
» Job Submission
» Job/Queue Status
» MDS Information
» Network Status
» Running/Queued Jobs
» PBS Job History
» NYS Grid
» Condor Flock Statistics

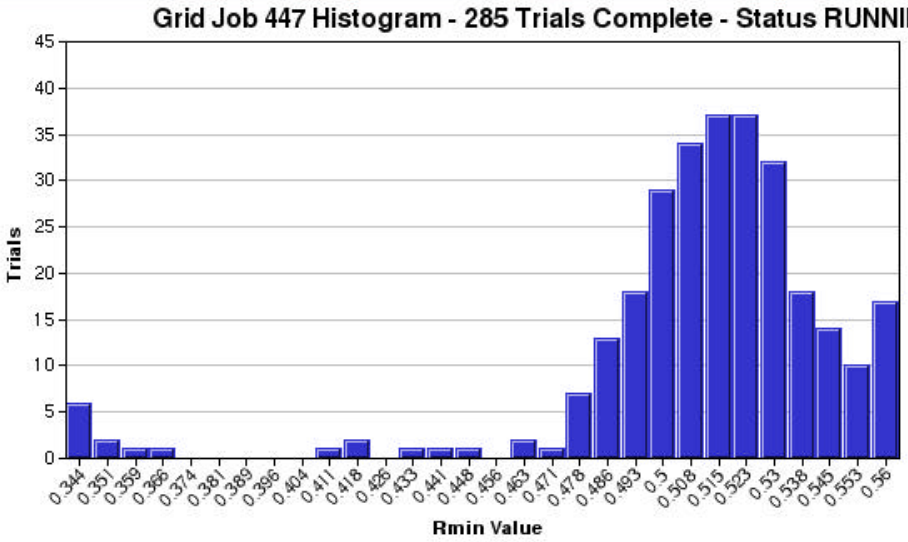
Data Grid

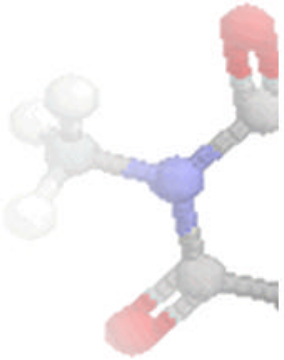
Education/Outreach

Staff Only

CCR HOME

Printer Friendly





Return to Job Details

Return to the Grid Job Status List

Return to the Grid Job Menu

Histogram of Completed Trial Structures

Advanced Center for Computational Research

Done

Internet

Start | Gmail - Inbox - Mi... | Center for Comp... | CCR Grid Comp... | SnB | Screenshots-Grid ... | clearwater.ccr.bu... | 10:22 AM

Grid Job Status

15-Mar-2005 10:23:49

Job Filter Criteria

Show GATs: BnP Auto Run, EADR, Ostrich, POM, Q-Chem, SnB, SnB DREAR

Job State: DEFINITION, STAGING, QUEUED, RUNNING, UPLOADING, COMPLETE, INCOMPLETE

Sort By: Job Id, Job Name, Resource, Num Procs, Status, Percent Complete, Last Update

(•) Descending
( ) Ascending

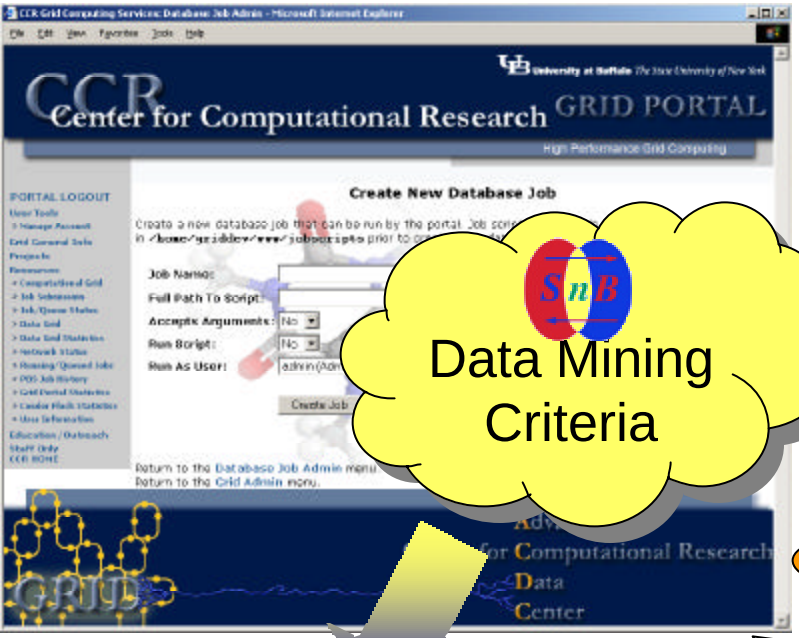
Filter Job List

SnB

Job Id	Job Name	Resource	Num Procs	Status	Percent Complete	Last Update	Cancel Job	Drilldown
447	iledhkl	clearwater.ccr.buffalo.edu	5	RUNNING	28.5	15-Mar-2005 10:22:00	☐	
446	trilys	clearwater.ccr.buffalo.edu	10	RUNNING	1	15-Mar-2005 10:22:00	☐	
444	64chkl	nash.ccr.buffalo.edu	3	COMPLETE	100	14-Mar-2005 22:00:01		
443	trilys	clearwater.ccr.buffalo.edu	10	COMPLETE	100	10-Mar-2005 22:48:00		
442	pr435hkl	nash.ccr.buffalo.edu	5	COMPLETE	100	10-Mar-2005 17:26:01		
441	vancohkl	clearwater.ccr.buffalo.edu	10	COMPLETE	100	10-Mar-2005 18:08:01		
434	16chkl	clearwater.ccr.buffalo.edu	5	COMPLETE	100	10-Mar-2005 14:42:01		
433	16chkl	clearwater.ccr.buffalo.edu	3	COMPLETE	100	10-Mar-2005 14:38:01		

Status of Jobs

SnB Grid Enabled Data Mining



Data Mining Criteria

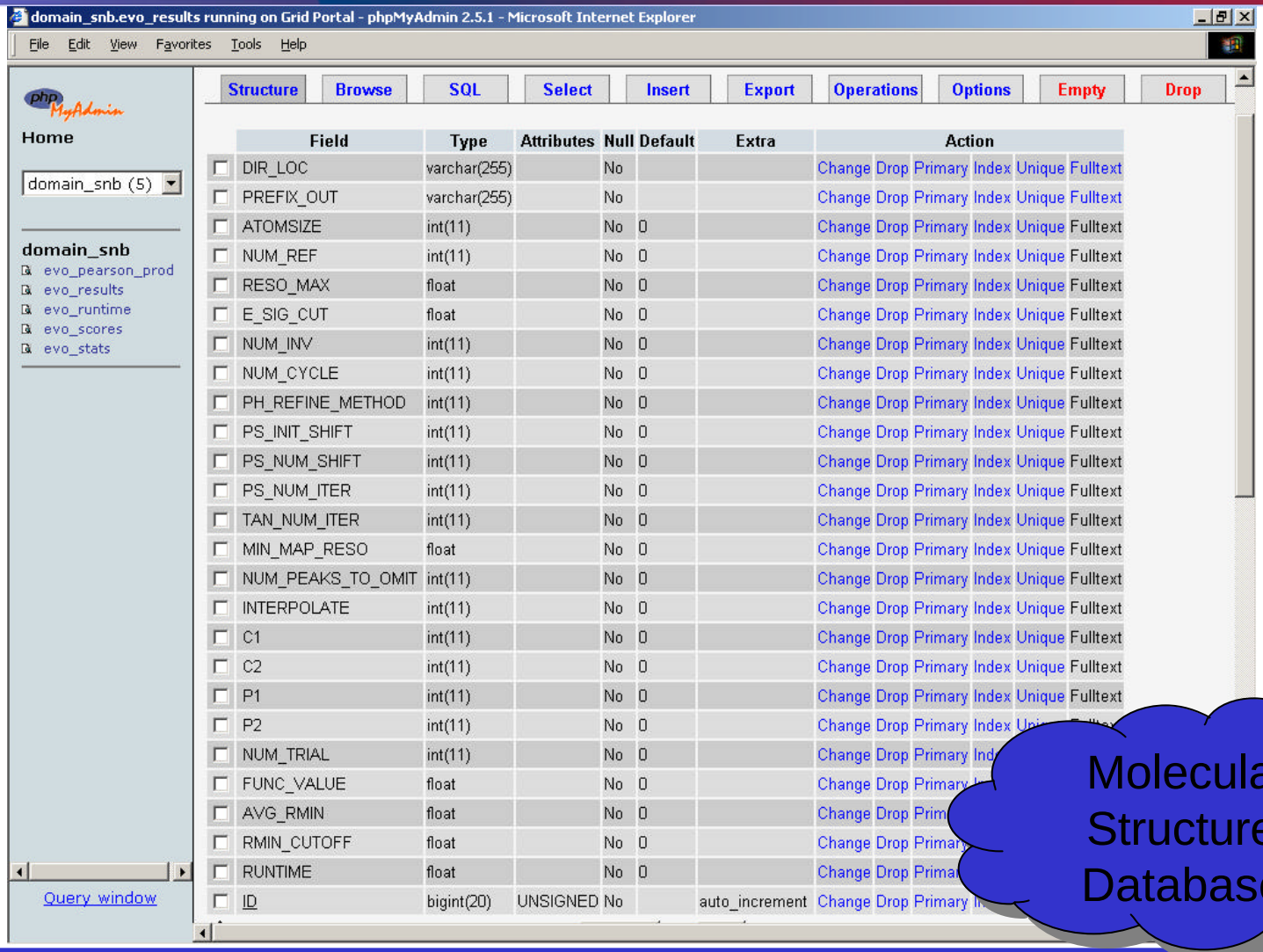
ACDC-Grid Data Grid

ACDC-Grid Computational Resources

Grid Portal Workflow Job Manager

Molecular Structure Database

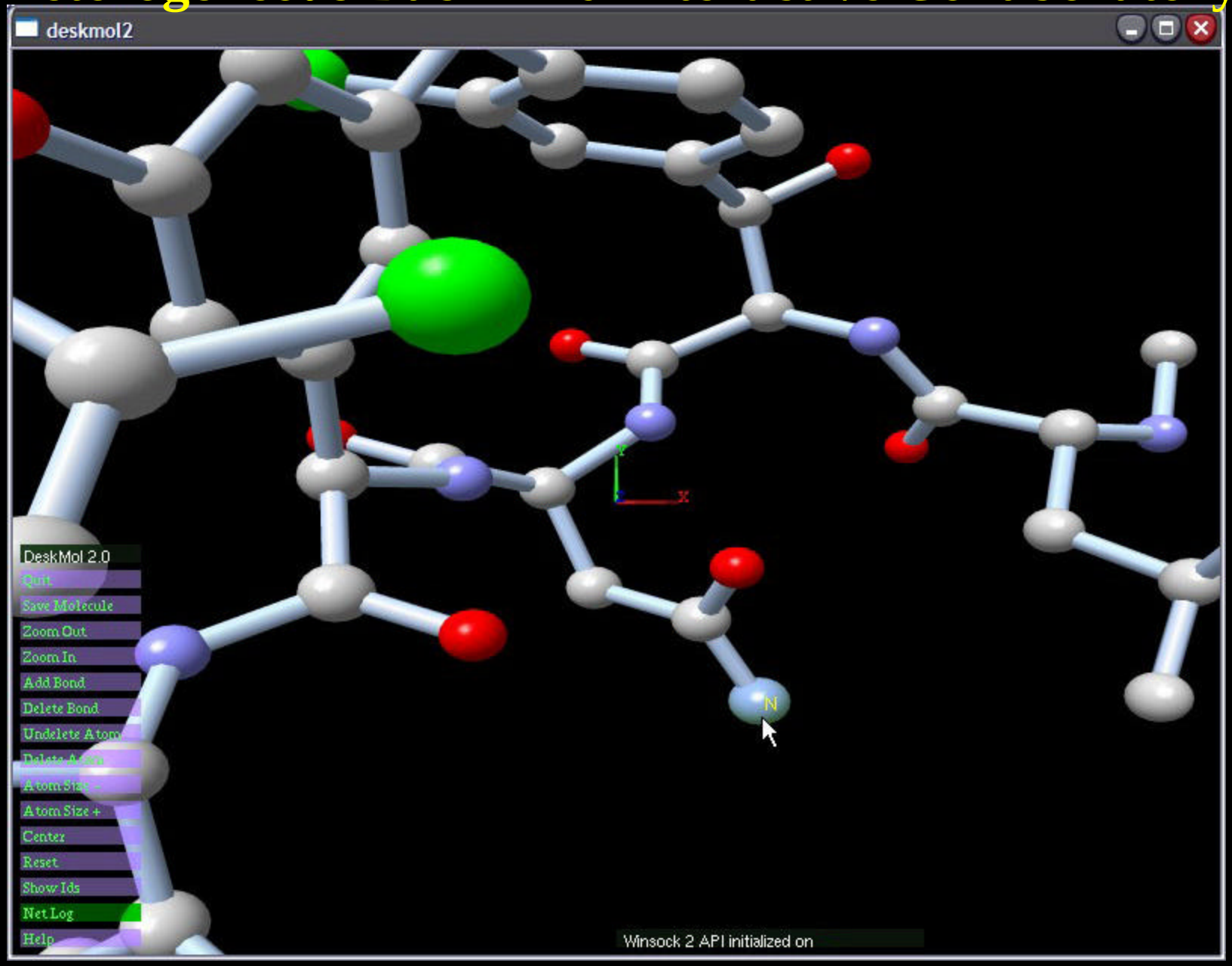
SnB Molecular Structure Database



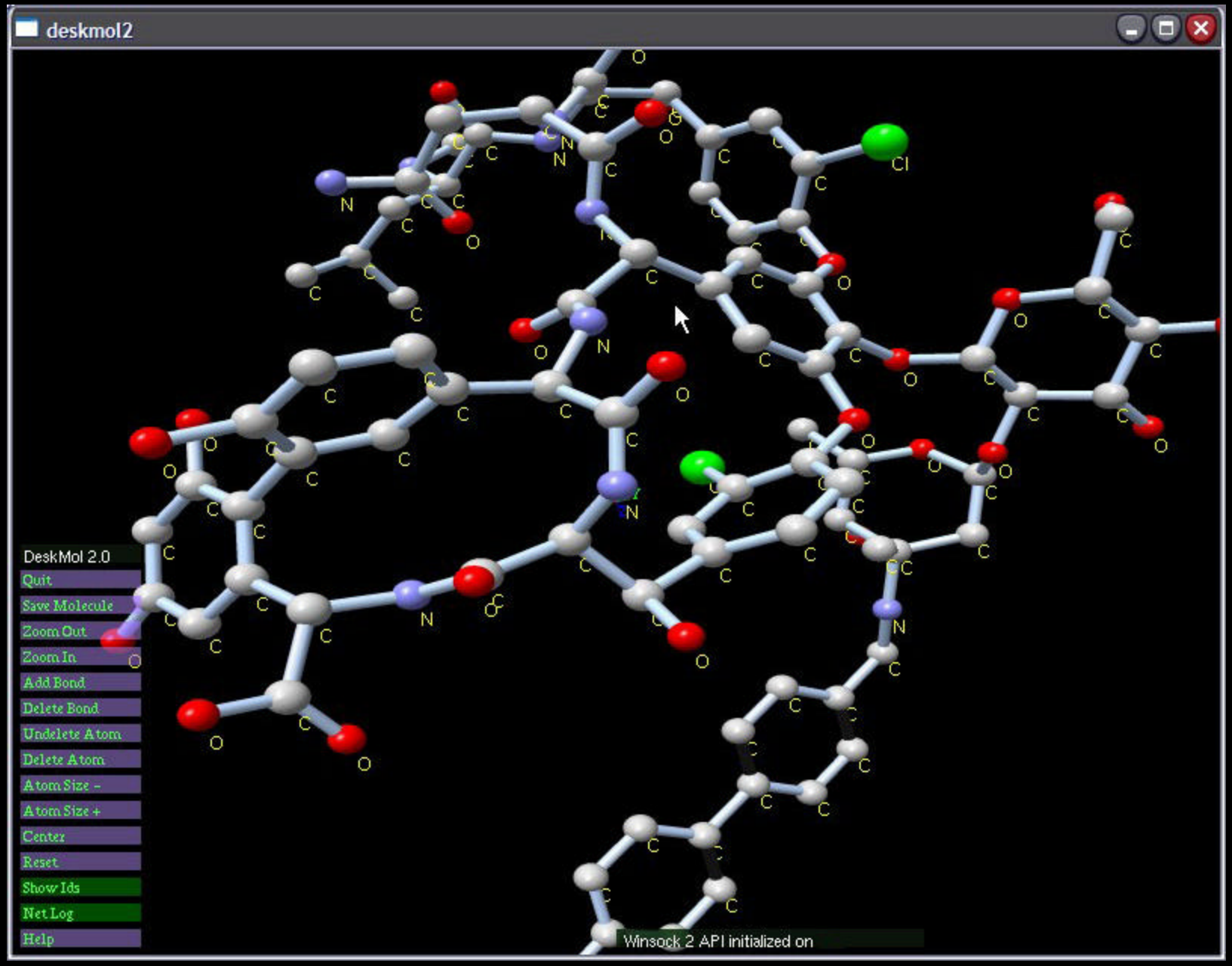
SnB Data Mining Results (Preliminary)

- Database over 22 atomic-resolution structures
- There are 5 different base data sets per structure
- Genetic Algorithm optimization on these 110 datasets
- GA results provide order(s) of magnitude improvement in cost-effectiveness
- Current focus on automation of algorithm for large-scale (international) grid deployment
- Decisions continue on which parameters to refine

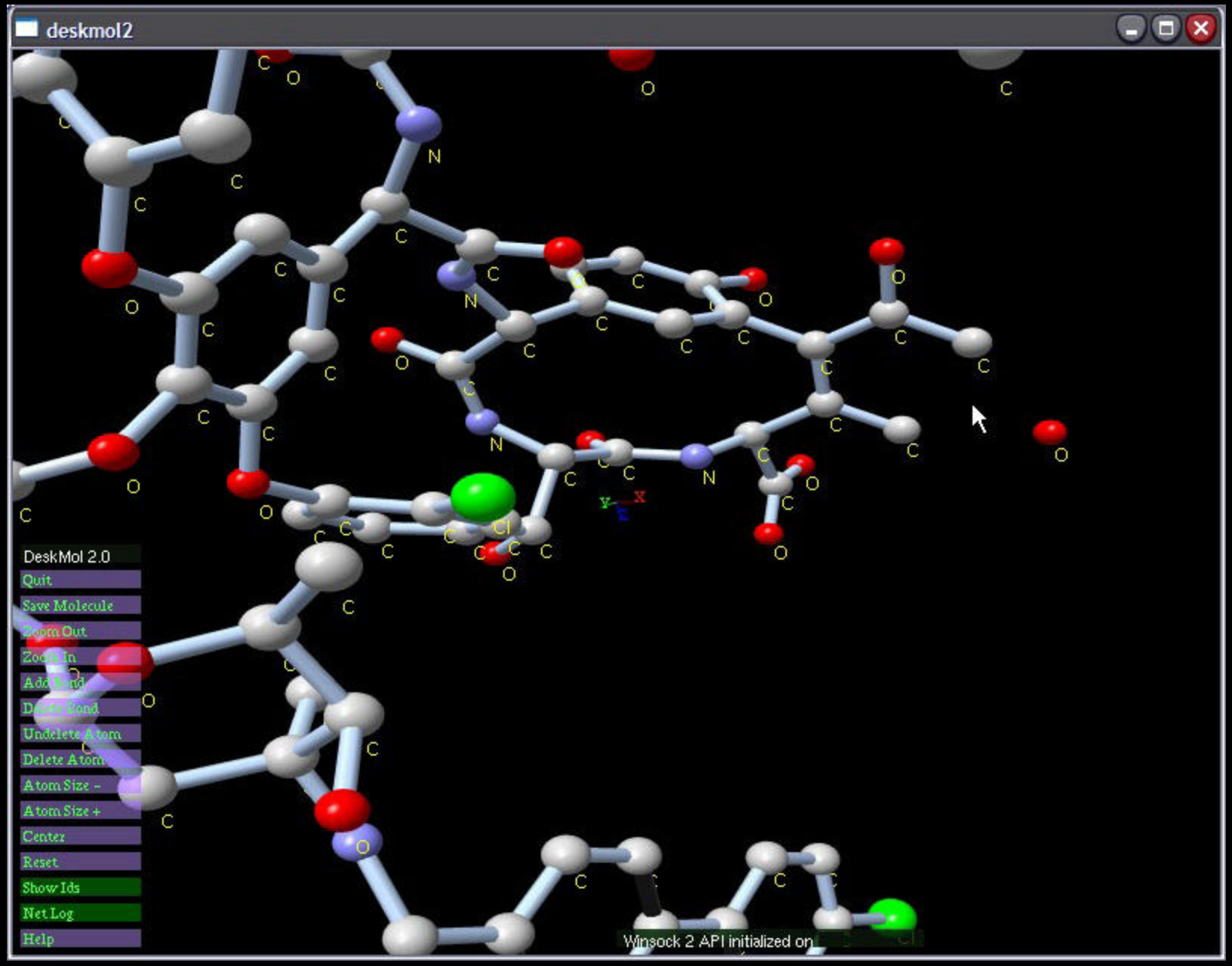
Heterogeneous Back-End Interactive Collaboratory



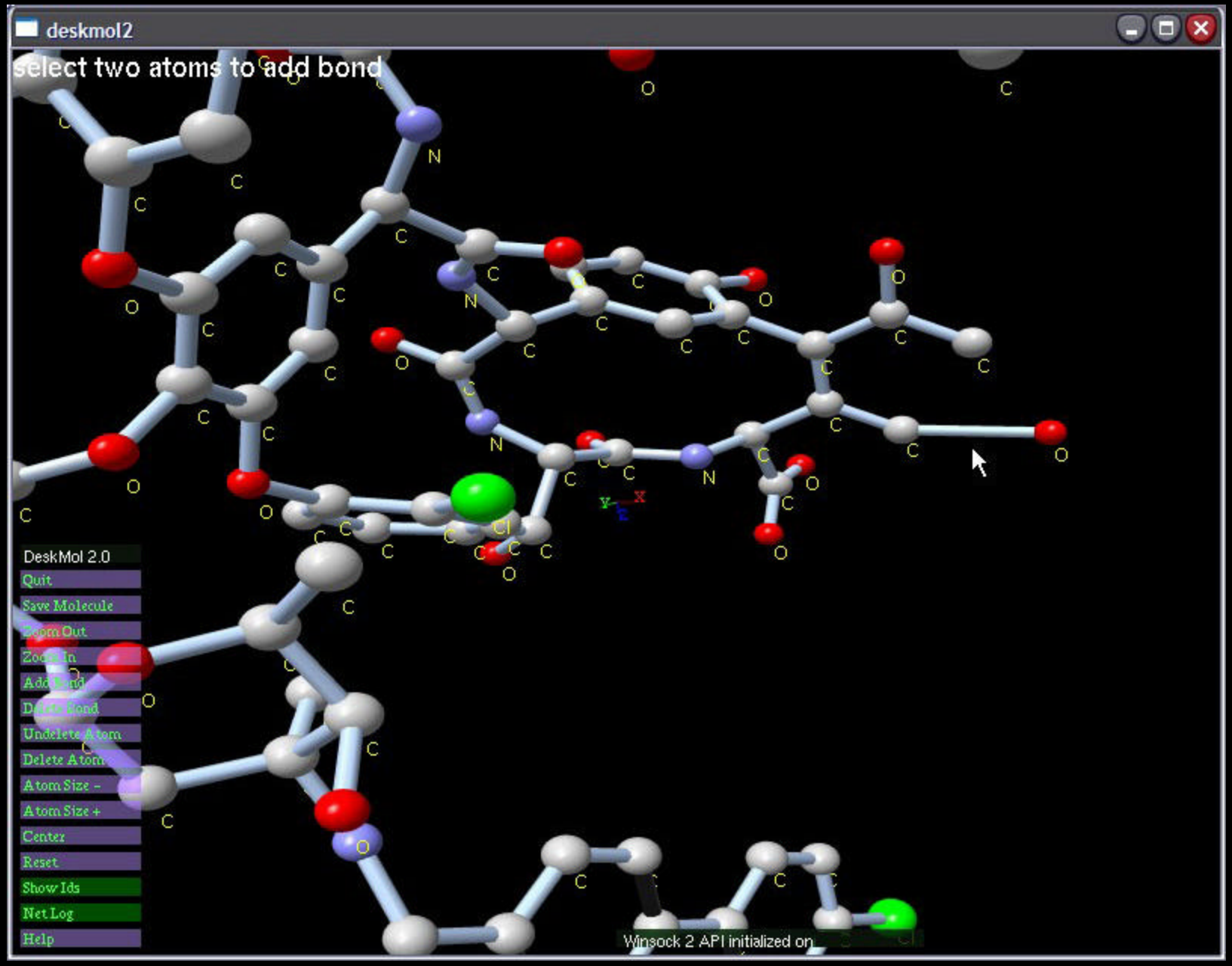
User starts up – default image of structure.



Molecule scaled, rotated, and labeled.



Remove Carbon Atoms (and Links)



User Adds Bond Between Atoms

Outreach

- **HS Summer Workshops in Computational Science**
  - **Chemistry, Visualization, Bioinformatics**
  - **10-14 HS Students Participate Each Summer for 2 weeks**
  - **Project-Based Program**

Outreach

- **Pilot HS Program in Computational Science**
  - **Year long extracurricular activity at Mount St. Mary's, City Honors, and Orchard Park HS**
  - **Produce next generation scientists and engineers**
  - **Students learn Perl, SQL, Bioinformatics**
  - **$50,000 startup funding from Verizon, PC's from HP**

Community Service

- Managed numerous baseball teams (LGYB, MMYB)
- Coached numerous basketball teams
  - House League (AYB)
  - PAL
  - Local Tournaments
- President of Amherst Youth Basketball (6 yrs.)
  - 350 boys/girls Þ ~1000 boys/girls
  - Web based
- Co-President/Coach WNY Warriors (AAU)
  - 6 Travel Teams
  - Numerous DI players
- Inducted into Amherst (NY) Avenue of Athletes
- Board Member of infoTechNiagara,……

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
- Bill Furey
- Steve Potter

- NSF, NIH, NYS, NIMA, NTA, Oishei, Wendt, DOE


www.ccr.buffalo.edu
CCR