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

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 PIIs
  - 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
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.
Probability theory gives information about certain linear combinations of phases.

- In particular, the triples $\phi_H + \phi_K + \phi_{-H-K} = 0$ 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: $|E_H|$  

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

$$|E_H| = \frac{|F_H|}{\sqrt{\langle |F_H|^2 \rangle}} = \frac{k \langle \exp[-B_{iso}(\sin\theta)^2/\lambda^2] \rangle^{-1} |F_H|_{meas}}{\left( \varepsilon_H \sum_{j=1}^{N} f_j^2 \right)^{1/2}}$$

- $\langle |E| \rangle$ constant for concentric resolution shells.  
- $\langle |E| \rangle$ constant regardless of reflection class ($\varepsilon_H$ correction factor).  
- The renormalization condition, $\langle |E|^2 \rangle = 1$ is always imposed
• $N$ = non-H atoms in unit cell

• Each triplet of phases or structure invariant, $\Phi_{HK}$, 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, $\Phi_{HK} \approx 0$
Conventional Direct Methods

Trial Phases → Phase Refinement

Tangent Formula → FFT → Density Modification (Peak Picking) → Solutions

Reciprocal Space → Real Space

Conventional Direct Methods
Shake-and-Bake Method: Dual-Space Refinement

Trial Structures → Structure Factors → Trial Phases → Phase Refinement → FFT → Density Modification (Peak Picking) (LDE) → Solutions

Reciprocal Space
“Shake”

Real Space
“Bake”

Parameter Shift → FFT⁻¹ → Tangent Formula

Trial Structures

University at Buffalo The State University of New York Center for Computational Research
A Direct Methods Flowchart

Start

Normalize data

Generate invariants

Generate trial

Refine phases

Shake-and-Bake

Another cycle?

FFT

Modify density

Solution found?

Compute FOMs

Another trial?

Stop

FFT^-1

Automated stop
### 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 |

#### Triplets

<table>
<thead>
<tr>
<th>Rank</th>
<th>H</th>
<th>K</th>
<th>-H-K</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
<td>45</td>
<td>3.90</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
<td>165</td>
<td>3.52</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>5</td>
<td>17</td>
<td>3.37</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>3</td>
<td>289</td>
<td>3.16</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>28</td>
<td>40</td>
<td>3.09</td>
</tr>
<tr>
<td>100n=840</td>
<td>19</td>
<td>259</td>
<td>734</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Total $n = 84$ unique atoms
Getting Started: Random Atoms

Random Number Generator

$n = 10$ atoms
(30 coordinates)

Structure Factor Calculation

$\phi_1 \phi_2 \phi_3 \phi_4 \phi_5 \phi_6 \phi_7 \phi_8 \phi_9 \phi_{10}$
Tangent Formula

\[ \tan \phi_H = \frac{-\sum_K |E_K E_{-H-K}| \sin(\phi_K + \phi_{-H-K})}{\sum_K |E_K E_{-H-K}| \cos(\phi_K + \phi_{-H-K})} \]

Parameter Shift Optimization

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

where \( |E_H| \propto |F_H| \) normalized in resolution shells

Invariants: \( \Phi_{HK} = \phi_H + \phi_K + \phi_{-H-K} \approx 0 \)

Weights: \( W_{HK} = A_{HK} = 2N^{-1/2} |E_H E_K E_{-H-K}| \)
Peak Picking
## Default SnB Parameters (given n atoms)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Full Structures</th>
<th>Substructures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phases</td>
<td>(10n)</td>
<td>(30n)</td>
</tr>
<tr>
<td>Triplet Invariants</td>
<td>(100n)</td>
<td>(300n)</td>
</tr>
<tr>
<td>Cycles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n&lt;100)</td>
<td>(n/2)</td>
<td>(2n)</td>
</tr>
<tr>
<td>(n&gt;100)</td>
<td>(n)</td>
<td>(2n)</td>
</tr>
<tr>
<td>Peaks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n&lt;100)</td>
<td>(n)</td>
<td>(n)</td>
</tr>
<tr>
<td>(n&gt;100)</td>
<td>(0.8n)</td>
<td>(0.8n)</td>
</tr>
</tbody>
</table>
## Sorted Trial Data

<table>
<thead>
<tr>
<th>Trial</th>
<th>Cycle</th>
<th>Phased</th>
<th>Rmin</th>
<th>Cryst. CC Ratio</th>
<th>Ratio</th>
<th>Peak</th>
</tr>
</thead>
<tbody>
<tr>
<td>97</td>
<td>56</td>
<td>836</td>
<td>0.349</td>
<td>0.27 0.45</td>
<td>0.05 1.2</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>56</td>
<td>836</td>
<td>0.350</td>
<td>0.26 0.43</td>
<td>0.03 1.1</td>
<td></td>
</tr>
<tr>
<td>82</td>
<td>56</td>
<td>836</td>
<td>0.350</td>
<td>0.26 0.44</td>
<td>0.03 1.1</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>56</td>
<td>836</td>
<td>0.351</td>
<td>0.26 0.45</td>
<td>0.03 1.0</td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>56</td>
<td>836</td>
<td><strong>0.351</strong></td>
<td><strong>0.27 0.48</strong></td>
<td>0.03 1.1</td>
<td></td>
</tr>
<tr>
<td>93</td>
<td>56</td>
<td>836</td>
<td>0.506</td>
<td>0.36 0.36</td>
<td>0.08 1.0</td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>56</td>
<td>836</td>
<td>0.515</td>
<td>0.38 0.37</td>
<td>0.18 2.3</td>
<td></td>
</tr>
<tr>
<td>69</td>
<td>56</td>
<td>836</td>
<td>0.522</td>
<td>0.37 0.39</td>
<td>0.21 2.6</td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>56</td>
<td>836</td>
<td>0.523</td>
<td>0.37 0.39</td>
<td>0.21 2.5</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>56</td>
<td>836</td>
<td>0.525</td>
<td>0.39 0.43</td>
<td>0.21 2.7</td>
<td></td>
</tr>
</tbody>
</table>

**Solutions**

**Nonsolutions**
Ph8755: SnB Histogram

Histogram of Rmin Values

Buckets:
15

Trials Read:
100

Best Trial:
3

Best Job:
PK_ano/manual_0

R-true:
0.097

R-random:
1.121
Minimal Function Traces

- **Solution**
  - Trace of $R_{\text{min}}$ Values
  - Initial value: 0.800
  - Final value: 0.330
  - Non-linear decrease

- **Nonsolution**
  - Trace of $R_{\text{min}}$ Values
  - Initial value: 0.720
  - Final value: 0.510
  - Fluctuations and non-linear trend
Phasing and Structure Size

Se-Met with *Shake-and-Bake* ......?

Se-Met

Multiple Isomorphous Replacement ......?

567 kDa (160 Se)

*Shake-and-Bake*

Conventional Direct Methods

Vancomycin

Number of Atoms in Structure

0 100 1,000 10,000 100,000
BnP: The Buffalo ‘n Pittsburgh Interface

Java GUI

DREAR
R. Blessing
Normalization

SnB
C. Weeks
R. Miller
H. Xu
Substructure solution

NANTMRF
G. D. Smith
Substructure comparison

PHASES components
W. Furey
Substructure refinement, protein phasing, solvent flattening, preparation for map viewing

S. Potter
J. Rappleye
R. Mungée
L. Pasupulati
<table>
<thead>
<tr>
<th>Task</th>
<th>Manual Mode (Workflow)</th>
<th>Auto Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substructure Determination</td>
<td>FOM Histogram and Trace</td>
<td>FOM Deviation from Mean</td>
</tr>
<tr>
<td>Site Validation</td>
<td>Trial Comparison</td>
<td>Occupancy Refinement</td>
</tr>
<tr>
<td>Enantiomorph Determination</td>
<td>Map Inspection</td>
<td>$\sigma$(protein)/$\sigma$(solvent)</td>
</tr>
<tr>
<td>Substructure Refinement (Optional)</td>
<td>SAME</td>
<td>SAME</td>
</tr>
<tr>
<td>Solvent Flattening</td>
<td>SAME</td>
<td>SAME</td>
</tr>
</tbody>
</table>
SeMet Test Data: Auto Mode Results

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

* Solution not recognized automatically
Manual Site Validation: Trial Comparison

### Compare Trials Results

**Compare Trials:** Fri Jun 13 11:54:36 EDT 2003  
**Structure ID:** 1JC4  
**Maximum distance for matching peaks:** 1.0

<table>
<thead>
<tr>
<th>SnB Trial</th>
<th>3</th>
<th>81</th>
<th>82</th>
<th>97</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Matches</td>
<td>2</td>
<td>26</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>Mean distance</td>
<td>0.69</td>
<td>0.21</td>
<td>0.17</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Peak</th>
<th>Select?</th>
<th>Peak</th>
<th>Distance</th>
<th>Peak</th>
<th>Distance</th>
<th>Peak</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>19</td>
<td>17</td>
<td>0.18</td>
<td>16</td>
<td>0.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>11</td>
<td>0.12</td>
<td>17</td>
<td>0.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>21</td>
<td>15</td>
<td>0.1</td>
<td>18</td>
<td>0.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>22</td>
<td>8</td>
<td>0.71</td>
<td>20</td>
<td>0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>23</td>
<td>23</td>
<td>0.38</td>
<td>21</td>
<td>0.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>24</td>
<td>25</td>
<td>0.62</td>
<td>24</td>
<td>0.61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td></td>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td></td>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td></td>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Save**  
**Close**  

---

University at Buffalo  The State University of New York  Center for Computational Research
Automated Site Validation: Occupancy Refinement

File: PK_ano/auto_0.SnB_peaks

<table>
<thead>
<tr>
<th>Peak</th>
<th>Select?</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>✓</td>
<td>11.49</td>
</tr>
<tr>
<td>20</td>
<td>✓</td>
<td>11.31</td>
</tr>
<tr>
<td>21</td>
<td>✓</td>
<td>9.97</td>
</tr>
<tr>
<td>22</td>
<td>✓</td>
<td>9.79</td>
</tr>
<tr>
<td>23</td>
<td>✓</td>
<td>9.42</td>
</tr>
<tr>
<td>24</td>
<td>✓</td>
<td>8.5</td>
</tr>
<tr>
<td>25</td>
<td>✓</td>
<td>7.89</td>
</tr>
<tr>
<td>26</td>
<td>✓</td>
<td>6.84</td>
</tr>
<tr>
<td>27</td>
<td>✓</td>
<td>5.73</td>
</tr>
<tr>
<td>28</td>
<td>✓</td>
<td>5.66</td>
</tr>
</tbody>
</table>

File: PK_ano/auto_0.SnB_pdb

<table>
<thead>
<tr>
<th>Site</th>
<th>Select?</th>
<th>Occupancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>✓</td>
<td>0.61</td>
</tr>
<tr>
<td>20</td>
<td>✓</td>
<td>0.62</td>
</tr>
<tr>
<td>21</td>
<td>✓</td>
<td>0.49</td>
</tr>
<tr>
<td>22</td>
<td>✓</td>
<td>0.47</td>
</tr>
<tr>
<td>23</td>
<td>✓</td>
<td>0.55</td>
</tr>
<tr>
<td>24</td>
<td>✓</td>
<td>0.45</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>26</td>
<td></td>
<td>0.12</td>
</tr>
<tr>
<td>27</td>
<td></td>
<td>0.06</td>
</tr>
<tr>
<td>28</td>
<td></td>
<td>0.06</td>
</tr>
</tbody>
</table>
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

Various logos and images related to grid computing and related projects are displayed:

- NSF Grid
- NEESgrid
- Data Grid Global
- GriPhys
- European Grid Forum
- TERAGRID
- NSF PAMI
- DISCOM
- SinRG
- APGrid
- PDB
- United States virtual observatory
- APAN (Asia-Pacific Advanced Network)

The image also includes the logos of the University at Buffalo, The State University of New York, and the Center for Computational Research (CCR).
Grid Computing Overview

- 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
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)
**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

Browser view of “miller” group files published by user “rappleye”
Migration Algorithm dependent on

- User access time
- Network capacity at time of migration
- User profile
- User disk quotas on various resources
Both platforms have reduced bandwidth available for additional transfers.
Network Connections
ACDC Data Grid Overview
(Grid-Available Data Repositories)

**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

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

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

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

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

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

**CSE Multi-Store**: 40 TB

**Storage Area Network**: 75 TB

**Network Attached Storage**: 1.2 TB

**182 GB Storage**

**56 GB Storage**

**100 GB Storage**

**136 GB Storage**

**70 GB Storage**

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

SQL Database

Resource 1
Resource 2
Resource n

User 1 User 2 User m

Predictive Scheduler

Maintain Profiles and Predict
- running time
- backfill on resources
- grid load and utilization
Preliminary GA results

Percent of estimates within 5% of actual values

Percent of estimates within 20% of actual values
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

Joplin Configuration Diagram

- Node scratch space (120 GB)
- 4 node Dell 2650 PVFS server (1096 GB)
- Dell 6650 4-way (ACDC)
- Dell 6650 4-way (GRID)
- Dell 6650 4-way (EAGLES)

Dell 2650 backup front-end
Dell 6650 4-way front-end
1 node Dell 2650 NFS server (342 GB)

GigE and Myrinet connection
GigE connection
73 GB hard drive
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**
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

ACDC-Grid
Computational Resources

Core Services
- Metacomputing Directory Service
- Globus Security Interface
- GRAM
- GASS

High-level Services and Tools
- Globus Toolkit
- MPI
- MPI-IO
- C, C++, Fortran, PHP
- globusrun

Applications
- Shake-and-Bake
- Apache
- MySQL
- Oracle

ACDC-Grid
Data Resources

Local Services
- Condor
- Stork
- MPI
- LSF
- PBS
- Maui Scheduler
- TCP
- UDP
- Irix
- Solaris
- RedHat Linux
- WINNT

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
Advanced Computational Data Center Grid Jobs

Grid Job Submission:
This section contains forms for the selection of a grid-enabled application, modification of an application template, grid job definition review and grid job submission.

Grid Job Status:
This section contains grid user-based 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.
Instructions and Description for Running a Job on ACDC-Grid

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 applications. 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 the 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 required 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.
Select a GAT: BnP Auto Run, EADR, Catfish, POM, Q-Chem, SnB, SnB DREAM, Sipit, snb-dev.
Full Structure / Substructure Template Selection
Default Parameters Based on Template
### Default Parameters (cont’d)

#### Initial Data Sets

- **Add Dataset**
- **Delete Dataset**

<table>
<thead>
<tr>
<th>Select dataset to delete</th>
<th>Dataset 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Datasets</strong></td>
<td><strong>Dataset 1</strong></td>
</tr>
<tr>
<td>Name (<em>8 chars max</em>)</td>
<td>Icedhkl</td>
</tr>
<tr>
<td>Dataset Type</td>
<td>Native</td>
</tr>
<tr>
<td>File Name (*.hkl)</td>
<td></td>
</tr>
<tr>
<td>File Type</td>
<td>F, Sig(F)</td>
</tr>
<tr>
<td>Wavelength</td>
<td>1.5418</td>
</tr>
<tr>
<td>Max. Resolution</td>
<td>0.94</td>
</tr>
<tr>
<td>Anomalous Dispersion</td>
<td>Not Measured</td>
</tr>
<tr>
<td>Heavy Element Type</td>
<td></td>
</tr>
<tr>
<td>Nat. Element Replaced</td>
<td></td>
</tr>
<tr>
<td>No. Expected Sites</td>
<td></td>
</tr>
<tr>
<td>F Primes (F')</td>
<td></td>
</tr>
<tr>
<td>F Double Primes (F'')</td>
<td></td>
</tr>
</tbody>
</table>

- **Continue**
- **Reset Sequence**
- **Reset Current Stage**
- **Cancel**
Generating Reflections (Drear)

Reflections and Invariants

Drear Table

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Job Type</th>
<th>Native Data</th>
<th>Derivative Data</th>
<th>Norm Method</th>
<th>Select</th>
</tr>
</thead>
<tbody>
<tr>
<td>iledhkl</td>
<td>BASIC</td>
<td>NULL</td>
<td>NULL</td>
<td>Wilson (Anisotropic)</td>
<td></td>
</tr>
</tbody>
</table>

Normalization Data

Data resolution cutoffs (in Angstroms)?
Low: 0.990, High: 0.94
Use Bayesian estimates for weak reflections? No
Min |F| / ||(|F|) for local scaling: 1.0
SIR and SAS cutoffs:
TMax: 6.0, ZMax: 3.0
XMIN: 3.0, YMn: 1.0

Generate Invariants

Data resolution cutoffs?
Low: 0.990, High: 0.94
Minimum allowed |E| / ||(|E|): 3.0, Maximum |E|: 5.0
Minimum allowed invariants / reflection ratio: 5.0
Initial values for adjustable parameters
Minimum |E| / ||(|E|) = ZMin: 3.0
Number of reflections to use: 0
Number of invariants to save: 0

Continue, Reset Sequence, Reset Current Stage, Cancel
Invariant Generation

Reflections and Invariants

**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: 3.990, High: 0.94
- **Use Bayesian estimates for weak reflections?** No
- **Min |F| / |sig|(|F|)**: 3.0
- **SIR and SAS cutoffs**:
  - Tmax: 6.0, ZMax: 3.0
  - XMin: 3.0, YMin: 1.0

**Generate Invariants**

- **Data resolution cutoffs?**
  - Low: 3.990, High: 0.94
- **Minimum allowed |E| / |sig|(|E|)**: 3.0
- **Minimum allowed invariants / reflection ratio**: 6.0
- **Initial values for adjustable parameters**
  - Minimum |E| / |sig|(|E|) = ZMin: 3.0
  - Number of reflections to use: 5400
  - Number of invariants to save: 5400

---

**Invariant Generation**
SnB Setup

Grid Parameters
- Preferred resource name: Grid Scheduler
- Number of processors: 8
- Wallclock time requested: (mins) 720
- Job Prefix for results: job00
- Queue: grid

SnB Run Parameters
- Invariants
  - Number of triplet invariants to use: 8400

- Trials To Process
  - Starting phases from: Random Atoms
  - Random seed (primo): 119209
  - Number of Trials: 1000
  - Starting Trial: 1
  - Input Phase File: none
  - Input Atom File: none

- 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
**Phase Refinement Method**

- Phase Refinement Method:
- Parameter Shift (Fast)
- Number of passes through phase set: 3
- Phase shift: 0.0
- Number of shifts: 2

**Real-Space Constraints**

- 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: 0
- Number of peaks: 84

**Twice Baking**

- Trials for E-Fourier filtering (fourier refinement)?: None
- Number of cycles: 3
- Number of peaks: 84
- Minimum |E|: 0.75

**Automatic solution identification criteria**

- Rmin Improvement (%): 95.0
- Rcryst Improvement (%): 25.0

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

**Minimum |E|:** Unused

---

**SnB Review (Grid job ID: 447)**
Graphical Representation of Intermediate Job Status
Histogram of Completed Trial Structures
Status of Jobs

<table>
<thead>
<tr>
<th>Job Id</th>
<th>Job Name</th>
<th>Resource</th>
<th>Num Procs</th>
<th>Status</th>
<th>Percent Complete</th>
<th>Last Update</th>
<th>Cancel Job</th>
<th>Drilldown</th>
</tr>
</thead>
<tbody>
<tr>
<td>447</td>
<td>iledhkl</td>
<td>clearwater.ccr.buffalo.edu</td>
<td>5</td>
<td>RUNNING</td>
<td>20.5</td>
<td>15-Mar-2005 10:22:00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>446</td>
<td>triyys</td>
<td>clearwater.ccr.buffalo.edu</td>
<td>10</td>
<td>RUNNING</td>
<td>1</td>
<td>15-Mar-2005 10:22:00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>444</td>
<td>64chkl</td>
<td>nash.ccr.buffalo.edu</td>
<td>3</td>
<td>COMPLETE</td>
<td>100</td>
<td>14-Mar-2005 22:00:01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>443</td>
<td>triyys</td>
<td>clearwater.ccr.buffalo.edu</td>
<td>10</td>
<td>COMPLETE</td>
<td>100</td>
<td>10-Mar-2005 22:48:00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>442</td>
<td>pr43hkl</td>
<td>nash.ccr.buffalo.edu</td>
<td>5</td>
<td>COMPLETE</td>
<td>100</td>
<td>10-Mar-2005 17:26:01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>441</td>
<td>vancohkl</td>
<td>clearwater.ccr.buffalo.edu</td>
<td>10</td>
<td>COMPLETE</td>
<td>100</td>
<td>10-Mar-2005 18:08:01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>434</td>
<td>15ehkl</td>
<td>clearwater.ccr.buffalo.edu</td>
<td>5</td>
<td>COMPLETE</td>
<td>100</td>
<td>10-Mar-2005 14:42:01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>433</td>
<td>15ehkl</td>
<td>clearwater.ccr.buffalo.edu</td>
<td>8</td>
<td>COMPLETE</td>
<td>100</td>
<td>10-Mar-2005 14:39:01</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SnB Grid Enabled Data Mining
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
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
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 ⇒ ~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