Molecular Structure Determination on a Computational & Data Grid

Mark Green & Russ Miller
Center for Computational Research, SUNY-Buffalo
Computer Science & Engineering, SUNY-Buffalo
Hauptman-Woodward Medical Inst
Biomedical Advances

- PSA Test (screen for Prostate Cancer)
- Avonex: Interferon Treatment for Multiple Sclerosis
- Artificial Blood
- Nicorette Gum
- Fetal Viability Test
- Implantable Pacemaker
- Edible Vaccine for Hepatitis C
- Timed-Release Insulin Therapy
- Anti-Arrythmia Therapy
  - Tarantula venom
- Direct Methods Structure Determination
  - Listed on “Top Ten Algorithms of the 20th Century”
  - Vancomycin
  - Gramacidin A
- High Throughput Crystallization Method: Patented
- NIH National Genomics Center: Northeast Consortium
- Howard Hughes Medical Institute: Center for Genomics & Proteomics
Bioinformatics in Buffalo
A $360M Initiative

- New York State: $121M
- Federal Appropriations: $13M
- Corporate: $146
- Foundation: $15M
- Grants & Contracts: $64M
Lead Institutions
- University at Buffalo (UB)
- Hauptman-Woodward Medical Research Inst.
- Roswell Park Cancer Institute

Corporate Partners
- Amersham Pharmacia, Beckman Coulter, Bristol Myers Squibb, General Electric, Human Genome Sciences, Immco, Invitrogen, Pfizer Pharmaceutical, Wyeth Lederle, Zeptometrix
- Dell, HP, SGI, Stryker, Sun
- AT&T, Sloan Foundation
- InforMax, Q-Chem, 3M, Veridian
- BioPharma Ireland, Confederation of Indian Industries
High-Performance Computing and High-End Visualization
- 110 Research Groups in 27 Depts
- 13 Local Companies
- 10 Local Institutions

External Funding
- $111M External Funding
  - $13.5M as lead
  - $97.5M in support
- $41.8M Vendor Donations
- $360M Bioinformatics Initiative

Deliverables
- 350+ Publications
- Software, Media, Algorithms, Consulting, Training, CPU Cycles…
Major CCR Resources
(12TF & 80TB)

- Dell Linux Cluster: #22 → #25 → #38
  - 600 P4 Processors (2.4 GHz)
  - 600 GB RAM; 40 TB Disk; Myrinet
- Dell Linux Cluster: #187 → #368 → off
  - 4036 Processors (PIII 1.2 GHz)
  - 2TB RAM; 160TB Disk; 16TB SN
  - Restricted Use (Skolnick)
- IBM BladeCenter Cluster
  - 532 P4 Processors (2.8 GHz)
  - 5TB SAN
- Apex Bioinformatics System
  - Sun V880 (3), 6800, 280R (2), PIIIs
  - Sun 3960: 7 TB Disk Storage
- HP/Compaq SAN
  - 75 TB Disk; 190 TB Tape SGI Origin3800
  - 64 Alpha Processors (400 MHz)
  - 32 GB RAM; 400 GB Disk
- IBM RS/6000 SP: 78 Processor
- Sun Cluster: 80 Processors
- SGI Intel Linux Cluster
  - 150 PIII Processors (1 GHz)
  - Myrinet
**CCR Visualization Resources**

- **Fakespace ImmersaDesk R2**
  - Portable 3D Device

- **Tiled-Display Wall**
  - 20 NEC projectors: 15.7M pixels
  - Screen is 11’×7’
  - Dell PCs with Myrinet2000

- **Access Grid Node**
  - Group-to-Group Communication
  - Commodity Components

- **SGI Reality Center 3300W**
  - Dual Barco’s on 8’×4’ screen
Advanced CCR Data Center (ACDC)
Computational Grid Overview

**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

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

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

**Expanding**
- RedHat, IRIX, Solaris, WINNT, etc

**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.
ACDC Data Grid Overview

**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

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

**CSE Multi-Store**
- 2 TB

**Storage Area Network**
- 75 TB

**Network Attached Storage**
- 480 GB

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

**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.
WNY Grid Highlights

- Heterogeneous Computational & Data Grid
- Currently in Beta with *Shake-and-Bake*
- WNY Release in 2H04
- Bottom-Up General Purpose Implementation
  - Ease-of-Use User Tools
  - Administrative Tools
- Back-End Intelligence
  - Backfill Operations
  - Prediction and Analysis of Resources to Run Jobs
    (Compute Nodes + Requisite Data)
Objective: Provide a 3-D mapping of the atoms in a crystal.

Procedure:
1. Isolate a single crystal.
2. Perform the X-Ray diffraction experiment.
3. Determine molecular structure that agrees with diffraction data.
Underlying atomic arrangement is related to the reflections by a 3-D Fourier transform.

- Phases lost during the crystallographic experiment.
- **Phase Problem:** Determine phases of the reflections.
Shake-and-Bake Method: Dual-Space Refinement

**Shake-and-Bake**

- **Trial Structures**
- **Structure Factors**

**Trial Phases**

**Phase Refinement**
- Tangent Formula
- Parameter Shift

**Density Modification** (Peak Picking) (LDE)

**Real Space**
- **“Bake”**
- **Solutions**

**Reciprocal Space**
- **“Shake”**
- DeTitta, Hauptman, Miller, Weeks

**Equations**
- Trial Structures
- Phase Refinement
- FFT
- FFT$^{-1}$
- Parameter Shift
- Tangent Formula
- Density Modification (Peak Picking) (LDE)
- Solutions

**Method Overview**

- Shake-and-Bake Method: Dual-Space Refinement
- Trial Structures
- Structure Factors
- Trial Phases
- Phase Refinement
- Density Modification (Peak Picking) (LDE)
- Real Space
- “Bake”
- Solutions

**References**
- DeTitta, Hauptman, Miller, Weeks
- Trial Structures
- Phase Refinement
- Density Modification (Peak Picking) (LDE)
- Real Space
- “Bake”
- Solutions
Ph8755: SnB Histogram

Atoms: 74
Space Group: P1
Phases: 740
Triples: 7,400

Trials: 100
Cycles: 40
Rmin range: 0.243 - 0.429
Phasing and Structure Size

Number of Atoms in Structure

0  100  1,000  10,000  100,000

Se-Met

Se-Met with Shake-and-Bake

Multiple Isomorphous Replacement

190kDa

Shake-and-Bake

Conventional Direct Methods

Vancomycin
Grid-Based SnB Objectives

- Install Grid-Enabled Version of SnB
- Job Submission and Monitoring over Internet
- SnB Output Stored in Database
- SnB Output Mined through Internet-Based Integrated Querying Tool

- Serve as Template for Chem-Grid & Bio-Grid
- Experience with Globus and Related Tools
Grid Services and Applications

ACDC-Grid
Computational Resources

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

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

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

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

ACDC-Grid Data Resources

Adapted from Ian Foster and Carl Kesselman
ACDC-Grid Portal Login

Grid Portal login screen
Data Grid Capabilities

Browser view of “miller” group files published by user “rappleye”
Grid enabled jobs can be monitored using the Grid Portal web interface dynamically.

- Charts are based on:
  - total CPU hours, or
  - total jobs, or
  - total runtime.

- Usage data for:
  - running jobs, or
  - queued jobs.

- Individual or all resources.

- Grouped by:
  - group, or
  - user, or
  - queue.
ACDC-Grid Portal User Management

Administrator based

user based
Administrator grants a user access to ACDC-Grid

- resources,
- software, and
- web pages.
Data Grid Resource Info

The ACDC Grid network bandwidth and latency are monitored by the Network Weather Service (NWS) developed and maintained by NPPC, Grid Computing Lab (UGC), CrAI Lab (SDRC), Globus (ANL/ISTI), Legion (UVa), Condor (Wisconsin), NCSA, Ninf, and Martin Quinson (ENS Lyon) Collaborators. The Grid Portal periodically monitors and dynamically forecasts the network performance of all grid computational resources with a 240 second frequency. The ACDC Data Grid extensively uses the network bandwidth and latency sensor data for optimizing the location and distribution of data required by ACDC Grid Portal computational jobs. Please refer to the Data Grid section of this site for an overview of new data mining techniques and algorithm enhancements that are being used to minimize the ACDC grid network traffic. Currently, the system sensor data can be viewed at any of the selected historical periods for all grid computational resources.
Both platforms have reduced bandwidth available for additional transfers.
File age, access time, and resource id denote:

- the amount of time since a file was accessed,
- when the file was accessed, and
- where the file currently resides respectively.
Development Requirements

- 7 – Person months for Grid Services Coordinator
  - Including Grid and Database conceptual design and implementation
- 5 – Person months for Grid Services Programmer
  - Web portal programming
- 5 – Person months for System Administrator
  - Globus, NWS, MDS, etc. installations
- 3 – Person months for Database Administrator
  - Grid Portal Database implementation

Minimum Maintenance Requirements

- 1 – Grid Services Coordinator
  - 100% level of effort
- 1 – Grid Services Programmer
  - 100% level of effort
- 1 – System Administrator
  - 50% level of effort
- 1 – Database Administrator
  - 10% level of effort
Acknowledgments

- Steve Gallo
- Jason Rappleye
- Jeff Tilson
- Martins Innus
- Cynthia Cornelius
- National Science Foundation, National Institutes of Health, Oishei Foundation, Wendt Foundation, Sloan Foundation, Verizon, NYS
- Gov Pataki, Congressman Reynolds, Senator Clinton, Senator Schumer, Congressman Quinn
- George DeTitta
- Herb Hauptman
- Charles Weeks
- Steve Potter