Enabling Collaborative Science Through Grid Technology

Russ Miller

Director, Center for Computational Research
UB Distinguished Professor, Computer Science & Engineering
Senior Research Scientist, Hauptman-Woodward Medical Inst

“Top 10 Worldwide Supercomputing Center”
- www.gapcon.com

University at Buffalo
The State University of New York
Outline

- Bioinformatics in Buffalo
- Supercomputing in Buffalo
- Grid Computing
- Grid Computing in Buffalo
  - Shake-and-Bake: Computational Crystallography
  - ECCE: Computational Chemistry
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 $290M Initiative

- UB Center for Advanced Bioengineering & Biomedical Technologies
  - $1M/yr NYS

- Center Disease Modeling & Therapy Discovery
  - UB, HWI, RPCI, Kaleida
  - $15.3M NYS
  - Software, device development, and drug therapies

- Buffalo Center of Excellence in Bioinformatics
  - UB, HWI, RPCI
  - $61M NYS
  - $10M Federal Government
  - $151 Corporate Funding

- UB Faculty Funding: $64M
Partnerships

- **Lead Partners:** SUNY-Buffalo, Hauptman-Woodward Medical Research Institute, Roswell Park Cancer Institute

- **Corporate Partners:** Amersham Pharmacia, AT&T, Beckman Coulter, BioPharma Ireland, Bristol Myers Squibb, Confederation of Indian Industries, Dell, General Electric, Human Genome Sciences, HP, Immco, InforMax, Invitrogen, Pfizer Pharmaceutical, Q-Chem, Sloan Foundation, SGI, Stryker, Sun, 3M, Veridian, Wyeth Lederle, Zeptometrix
Experimental Facilities I

- Molecular Targeting Laboratory
  - Screen 30-50K compounds every 3 months
  - Apply compound to cell (different genes treated w fluor markers)
  - Rapidly identify effect on specific gene expression pathways

- Gene Expression Laboratory
  - High-throughput microarray and gene chip
  - Discover new genes, their functions, and pathways

- Proteomics and Molecular Kinetics Lab
  - Identify molecular targets found in Gene Expression Lab

- Disease Modeling Laboratory
  - In vivo testing (flies, mice, baboons,…)
  - Gene targeting and genetic mapping facilities
Experimental Facilities II

- **Bioengineering Support Laboratory**
  - Capabilities in photonics and nano-tech research
  - E.g., handheld devices to test for diseases

- **Protein Scale-Up and Purification**

- **High-Throughput Robotic Combinatorial Chemistry/Parallel Synthetic Chemistry Capabilities**
  - Drugs created robotically; Tested for interaction with target protein
  - Rapid identification of a large number of potential drugs

- **Public Health and Molecular Pathology**
  - Tissue repositories; disease gene maps; medical informatics

- **High-Throughput Search Process for Structural Biology**
  - Tests 1536 “chemical cocktails” to determine effective parameters for crystallization
Personnel
- Hired Jeff Skolnick as Director (7/02)
  - Brought 13 additional staff to Buffalo
  - Authorized to hire 10 additional research groups
- Hired Norma Nowak as co-Director (4/03)
  - Authorized to hire 10 additional research groups
- Additional members TBD

External Funding ($0)
- Applications submitted

Deliverables
- Six (6) scientific papers

Resources
- Building
- 6TF → 10TF Compute Cluster
Center for Computational Research

- High-Performance Computing and High-End Visualization
  - 110 Research Groups in 27 Depts
  - 25 Companies and Institutions

- Sample Areas
  - Urban Visualization and Simulation
  - Computational Chemistry
  - Ground Water Modeling
  - Geophysical Mass Flows
  - Networked Multimedia
  - Medical Imaging

- Training
  - Workshops; Courses
  - Degree Programs
Personnel

- 18 State-Supported Staff
- 2 Grant-Supported Staff

External Funding

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

Deliverables

- 350+ Publications
- Software, Media, Algorithms, Consulting, Training, CPU Cycles, etc.
Computational Resources (9TF)

- **Dell Linux Cluster - #22 on top500**
  - 600 P4 Processors (2.4 GHz)
  - 600 GB RAM; 40 TB Disk; Myrinet

- **Dell Linux Cluster - #187 on top500**
  - 4036 Processors (PIII 1.2 GHz)
  - 2TB RAM; 160TB Disk; 16TB SN

- **SGI Origin3800**
  - 64 Processors (400 MHz)
  - 32 GB RAM; 400 GB Disk

- **IBM RS/6000 SP**
  - 78 Processors
  - 26 GB RAM; 640 GB Disk

- **Sun Microsystems Cluster**
  - 48 Sun Ultra 5s (333MHz)
  - 16 Dual Sunblades (750MHz)
  - 30 GB RAM, Myrinet

- **SGI Intel Linux Cluster**
  - 150 PIII Processors (1 GHz)
  - 75 GB RAM, 2.5 TB Disk Storage

- **Apex Bioinformatics System**
  - Sun V880 (3), 6800, 280R (2), PIIIs
  - Sun 3960: 7 TB Disk Storage

- **HP/Compaq SAN**
  - 25 TB Disk; 250 TB Tape

UBCOEB System
Sample Computational Research

- **Computational Chemistry** (King, Kofke, Coppens, Furlani, Tilson, Lund, Swihart, Ruckenstein, Garvey)
  - Algorithm development & simulations
- **Groundwater Flow Modeling** (Rabideau, Jankovic, Becker, Flewelling)
  - Predict contaminant flow in groundwater & possible migration into streams and lakes
- **Geophysical Mass Flows** (Patra, Sheridan, Pitman, Bursik, Jones, Winer)
  - Study of geophysical mass flows for risk assessment of lava flows and mudslides
- **Bioinformatics** (Zhou, Miller, Hu, Szyperski – NIH Consortium, HWI)
  - Protein Folding: computer simulations to understand the 3D structure of proteins
  - Structural Biology; Pharmacology
- **Computational Fluid Dynamics** (Madnia, DesJardin, Lordi, Taulbee)
  - Modeling turbulent flows and combustion to improve design of chemical reactors, turbine engines, and airplanes
- **Physics** (Jones, Sen)
  - Many-body phenomena in condensed matter physics
- **Chemical Reactions** (Mountziarlis)
- **Molecular Simulation** (Errington)
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

- **VREX VR-4200 Stereo Imaging Projector**
  - Portable projector works with PC
Sample Visualization Areas

- **Computational Science** (Patra, Sheridan, Becker, Flewelling, Baker, Miller, Pitman)
  - Simulation and modeling

- **Urban Visualization and Simulation** (CCR)
  - Public projects involving urban planning

- **Medical Imaging** (Hoffmann, Bakshi, Glick, Miletich, Baker)
  - Tools for pre-operative planning; predictive disease analysis

- **Geographic Information Systems** (CCR, Bisantz, Llinas, Kesavadas, Green)
  - Parallel data sourcing software

- **Historical Reenactments** (Paley, Kesavadas, More)
  - Faithful representations of previously existing scenarios

- **Multimedia Presentations** (Anstey, Pape)
  - Networked, interactive, 3D activities
3D Medical Visualization App

- Collaboration with Children’s Hospital
  - Leading miniature access surgery center
- Application reads data output from a CT Scan
- Visualize multiple surfaces and volumes
- Export images, movies or CAD representation of model
Multiple Sclerosis Project

- Collaboration with Buffalo Neuroimaging Analysis Center (BNAC)
  - Developers of Avonex, drug of choice for treatment of MS
- MS Project examines patients and compares scans to healthy volunteers
Multiple Sclerosis Project

- Compare caudate nuclei between MS patients and healthy controls
- Looking for size as well as structure changes
  - Localized deformities
  - Spacing between halves
- Able to see correlation between disease progression and physical structure changes
Grid Computing 2003

European GRID Forum

TERAGRID

NSF PACI

DISCOM
SinRG
APGrid

PDB
PROTEIN DATA BANK

USA virtual observatory

APAN

Asia-Pacific Advanced Network
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

Thanks to Mark Ellisman
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
Growth of Data and Load vs. Moore’s Law

1990

2000

2010

Combinatorial Chemistry

ESTs

Human Genome

Pharmacogenomics

Metabolic Pathways

Computational Load

Genome Data

Moore’s Law

Courtesy of Rick Stevens
A Short History of the Grid

- Grand Challenge Problems (1980s)
  - NSF and DOE initiatives
  - “Science is a team sport”
  - Initiate multi-resource projects involving computation, instruments, visualization, data

- Evolution of Related Communities
  - Parallel computation
    - Address resource limitations
  - Networking
    - Gigabit testbed program
      - Investigate potential testbed network architectures
      - Explore usefulness for end-users

CASA Gigabit Testbed (1990s)
Globus model focuses on providing key Grid services:
- Resource access and management
- Grid FTP
- Information Service
- Security services:
  - Authentication
  - Authorization
  - Policy
  - Delegation
- Network reservation, monitoring, control

The Grid as a Layered Set of Services:
- Applications
- Core Services
- High-level Services and Tools
- Local Services
Extensible TeraGrid Facility (ETF)

Caltech: Data collection analysis
- 0.4 TF IA-64
- IA32 Datawulf
- 80 TB Storage

SDSC: Data Intensive
- 4 TF IA-64
- DB2, Oracle Servers
- 500 TB Disk Storage
- 6 PB Tape Storage
- 1.1 TF Power4

LEGEND
- Clusters
- Storage Servers
- Shared Memory
- Disk Storage
- Backplane Routers

30 Gb/s
LA Hub
40 Gb/s
Chicago Hub
30 Gb/s
30 Gb/s

4 TF IA-64
DB2, Oracle Servers
500 TB Disk Storage
6 PB Tape Storage
1.1 TF Power4

ANL: Visualization
- 1.25 TF IA-64
- 96 Viz nodes
- 20 TB Storage

NCSA: Compute Intensive
- 10 TF IA-64
- 128 large memory nodes
- 230 TB Disk Storage
- GPES and data mining

PSC: Compute Intensive
- 6 TF EV68
- 71 TB Storage
- 0.3 TF EV7 shared-memory
- 150 TB Storage Server

Figure courtesy of Rob Pennington, NCSA
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
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)

Reciprocal Space “Shake”

Real Space “Bake”

Solutions
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
Proof of Concept

- Combine CCR’s Heterogeneous Compute Platforms into a Grid
  - Client/Server Configurations
  - Rapid Prototype 4Q02 (not Globus)

- Develop a user interface to monitor system
  - Dynamic HTML Grid Interface

- Key Features for Proof of Concept
  - Load Balancing
  - Fault Tolerance
  - Result and Grid Statistics
Dynamic HTML Grid Status

Grid Server Information
- Date/Completion Time
- Parallel Run Time/Serial Run Time/Speedup
- Trial Result Rate (Trial/Minute)

Shows Configured Platform Information Dynamically
- Platform – Type/Name/Picture
- Status – Idle/Working/Offline
- Resources – Nodes/Total Process/Available Process/Running Process

Shows Job Status Dynamically
- Trails – Total Number/Amount Processed
- Platform Server State – Block Queue/Float/Race
- Result Figure of Merit Histogram
# Grid Server Console

(Vancomycin)

## Center for Computational Research

University at Buffalo  The State University of New York

### Grid Server Console

<table>
<thead>
<tr>
<th>CONSOLE</th>
<th>GRID SERVER</th>
<th>PLATFORM SERVER STATUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPLETE</td>
<td>WORKING</td>
<td>IDLE</td>
</tr>
<tr>
<td>100% FLOAT</td>
<td>Tue Nov 5 23:29:10 2092</td>
<td>RACING</td>
</tr>
<tr>
<td>PARALLEL RUN TIME:</td>
<td>108.18 minute</td>
<td>100%</td>
</tr>
<tr>
<td>TRIAL/MINUTE:</td>
<td>265.82</td>
<td>90%</td>
</tr>
<tr>
<td>COMPLETION TIME:</td>
<td>0.00 / 188.10 minute</td>
<td>75%</td>
</tr>
<tr>
<td>SERIAL RUN TIME:</td>
<td>114785.88 minute</td>
<td>54%</td>
</tr>
<tr>
<td>SPEEDUP:</td>
<td>609.81</td>
<td>45%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>READY</th>
<th>READY</th>
<th>READY</th>
<th>OFFLINE</th>
<th>READY</th>
<th>OFFLINE</th>
<th>READY</th>
<th>READY</th>
<th>READY</th>
<th>OFFLINE</th>
<th>READY</th>
<th>READY</th>
<th>READY</th>
<th>READY</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>41699</td>
<td>42008</td>
<td>1558</td>
<td>0</td>
<td>41974</td>
<td>41144</td>
<td>41244</td>
</tr>
<tr>
<td>50000</td>
<td>50000</td>
<td>1557</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>41973</td>
<td>42501</td>
<td>41143</td>
<td>0</td>
<td>42007</td>
<td>41243</td>
<td>41606</td>
</tr>
<tr>
<td>50000</td>
<td>50000 / 50060</td>
<td>544 / 1557</td>
<td>0 / 0</td>
<td>0 / 0</td>
<td>0 / 0</td>
<td>232 / 365</td>
<td>296 / 494</td>
<td>48645 / 39586</td>
<td>0 / 0</td>
<td>25 / 34</td>
<td>99 / 108</td>
<td>169 / 365</td>
<td></td>
</tr>
</tbody>
</table>

### Node Information

- Nodes: 320
- Process Available: 649
- Running: 645

### Job Status

- SGI INTEL/ALPHA: 1
- SGI INTEL/ALPHA: 1
- DNA RNA: 1
- ORIGIN: 0
- BRIGGS (SOLAR POWERED): 0
- SUN BLADE/ULTRA: 0
- IBM SP2 PWR2/PWR3: 0
- DELL XEON: 0
- IBM 340: 0
- IBM 44F: 0
- SGI OCTANE: 0
- SGI ONYXX: 0

### Shake-N-Bake

- NASH/MOONGLOWS: 1
- NASH/MOONGLOWS: 1
- DNA RNA: 1
- CROSBY: 1
- BRIG: 1
- YOUNG: 1
- STILLS: 1
- JOPLIN: 1
- MAMA PAPAS: 0
- COASTERS: 0
- THEODOORS: 0
- CREAM: 0

---

University at Buffalo  The State University of New York  Center for Computational Research
Grid Portal
- Access control lists, security groups
- User attributes, history, proxies
- Managed through MySQL database
- Distributed data grid

Globus
- Vers 2.2.4 installed and in production
- Metacomputing Directory Services (MDS) stored in MySQL
  - Eliminates need for LDAP

Condor and Condor-G
- Used for resource management and grid job submissions
Red queue color indicates that there are currently running or queued jobs.
ECCE “Grid” at CCR

- Computational Chemistry
  - Relativistic effects/Heavy elements
  - Algorithm development
  - Theoretical physical chemistry

- Structural/Systems Biology
  - Protein structure
  - Enzyme catalysis

- Chemical Engineering
  - Condensed phases/Mixed phase predictions
  - Catalysis

- Geology, Pharmacology, Medical School

- Import Scientific Information
  - Application independent input
  - ECCE automatically formats for target application (Gaussian98, NWChem)

- Computing at CCR
  - 881 available CPUs (>2.5TFlops)
    - (Xeon, P3, Power3, R12K)
  - Uniform access to all platforms via ECCE “job launcher”

- Chemical Analysis
  - Full complement of visual tools for understanding data/publication quality graphics
“Genomics is powering the new biology, but Computing is in the driver’s seat.”

BioGrids
BioGrids provide scalable computing so that biologists can focus on biology.

- EUROGRID BioGRID
- Asia Pacific BioGRID
- NC BioGrid
- Bioinformatics Research Network
- Osaka University Biogrid
- Indiana University BioArchive BioGrid
Contact Information

miller@buffalo.edu
www.ccr.buffalo.edu
Acknowledgments

Mark Green  
Steve Gallo  
Jason Rappleye  
Jeff Tilson  
Martins Innus

Betty Capaldi  
Bruce Holm  
Janet Penksa

George DeTitta  
Herb Hauptman  
Charles Weeks  
Steve Potter

Rohit Bakshi  
Philip Glick