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

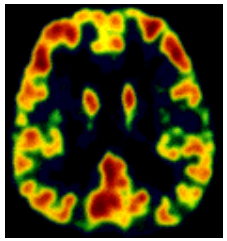
The State University of New York

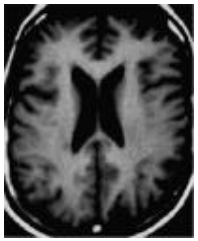
Outline

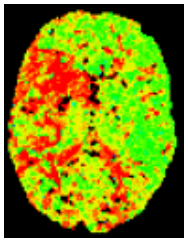
- Bioinformatics in Buffalo
- Supercomputing in Buffalo
- Grid Computing
- Computational Crystallography
- Buffalo Direct Methods Crystallography
- Crystallographic Grid Computing in Buffalo

WNY Biomedical Advances

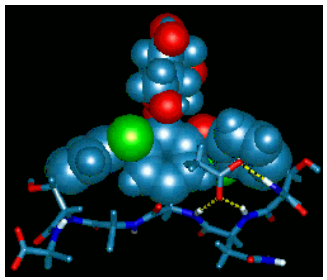
- PSA Test (screen for Prostate Cancer)
- Avonex: Interferon Treatment for Multiple Sclerosis
- Artificial Blood
- Nicorette Gum
- Fetal Viability Test
- Implantable Pacemaker
- Edible Vaccine for Hepatitis C
- Timed-Release Insulin Therapy
- Anti-Arrythmia Therapy
  - Tarantula venom







- 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
  - Med Tech for Product Dev & Commer.
- 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

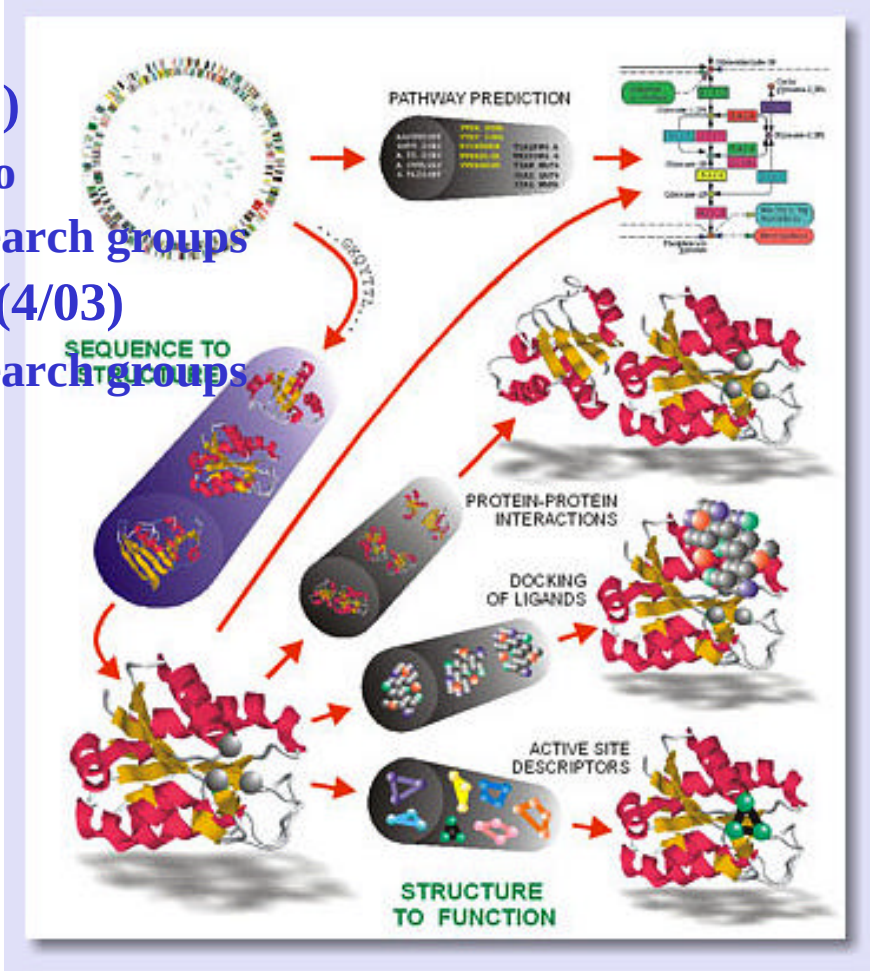
- **Molecular Targeting Laboratory**
  - Screen 30-50K compounds every 3 months
  - Apply compound to cell (different genes treated w fluor markers)
  - Rapidly identify effect on specific gene expression pathways
- **Gene Expression Laboratory**
  - High-throughput microarray and gene chip
  - Discover 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**

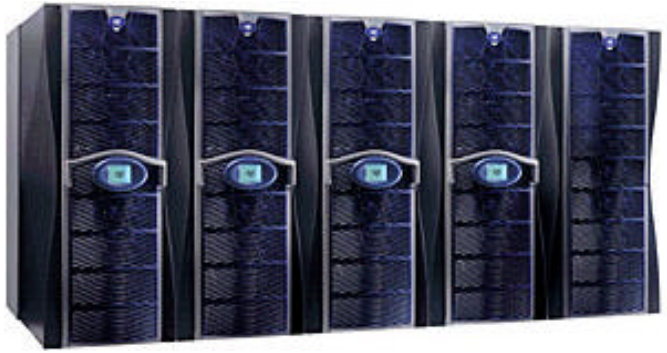
UBCOEB 2002-03 Snapshot

- **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



CCR 1999-2003 Snapshot

- **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

- Dell Linux Cluster - #22 on top500
  - 600 P4 Processors (2.4 GHz)
  - 600 GB RAM; 40 TB Disk
- Dell Linux Cluster - #187 on top500
  - 4036 Processors (PIII 1.2 GHz)
  - 2TB RAM; 160TB Disk; 16TB SN

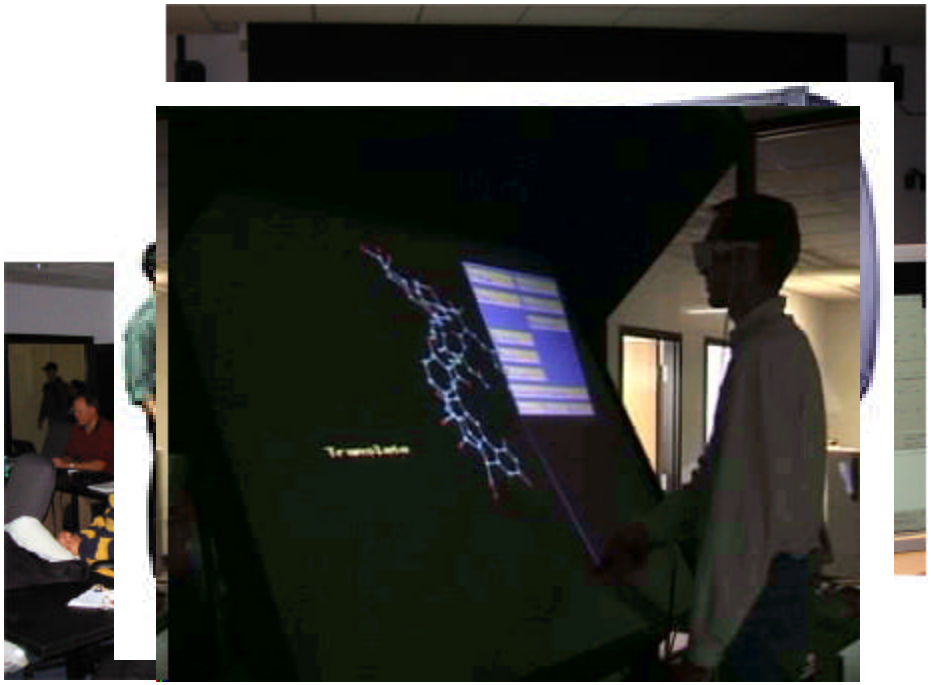
- SGI Origin3800
  - 64 Processors (400 MHz)
  - 32 GB RAM; 400 GB Disk
- IBM RS/6000 SP
  - 78 Processors
  - 26 GB RAM; 640 GB Disk
- Sun Microsystems Cluster
  - 48 Sun Ultra 5s (333MHz)
  - 16 Dual Sunblades (750MHz)
  - 30 GB RAM, Myrinet
- SGI Intel Linux Cluster
  - 150 PIII Processors (1 GHz)
  - 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

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** (Mountziaris)
- **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

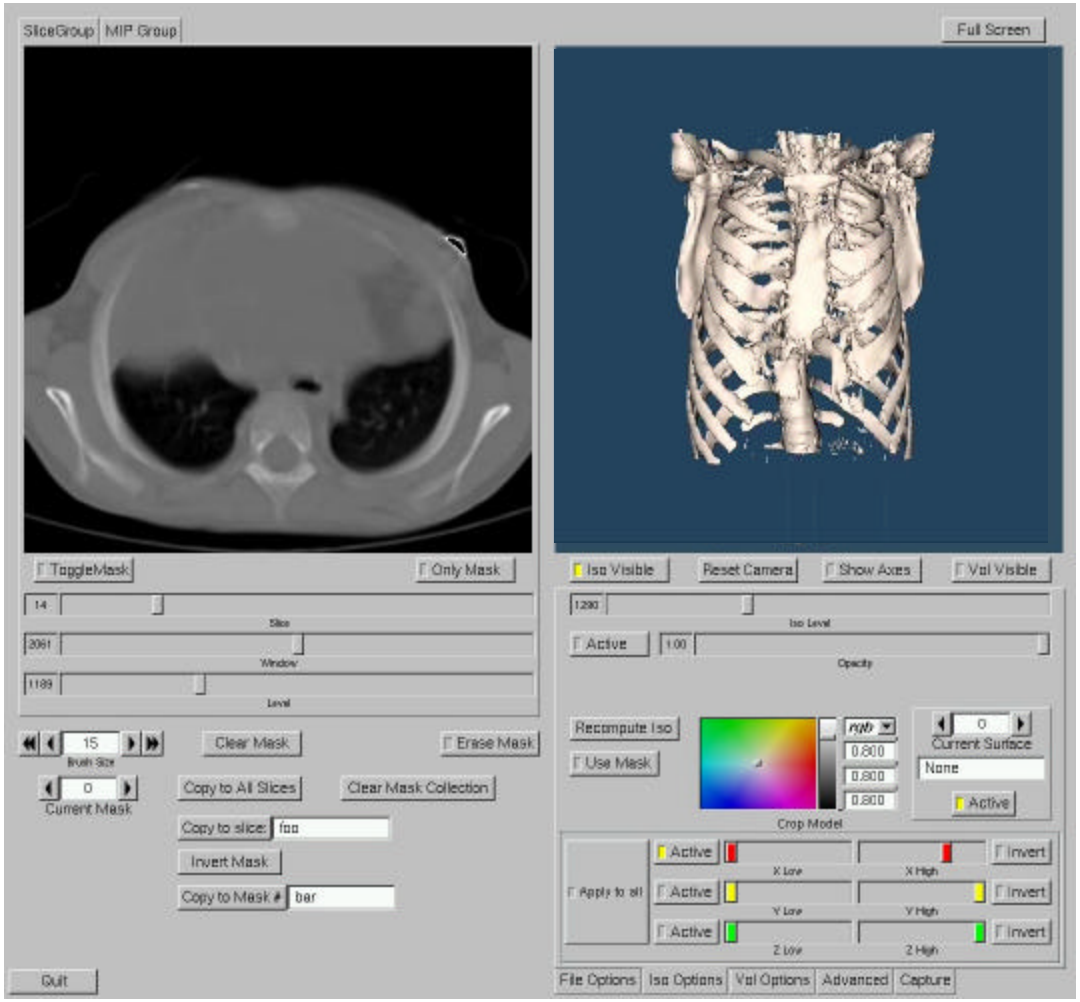


Sample Visualization Areas

- **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**

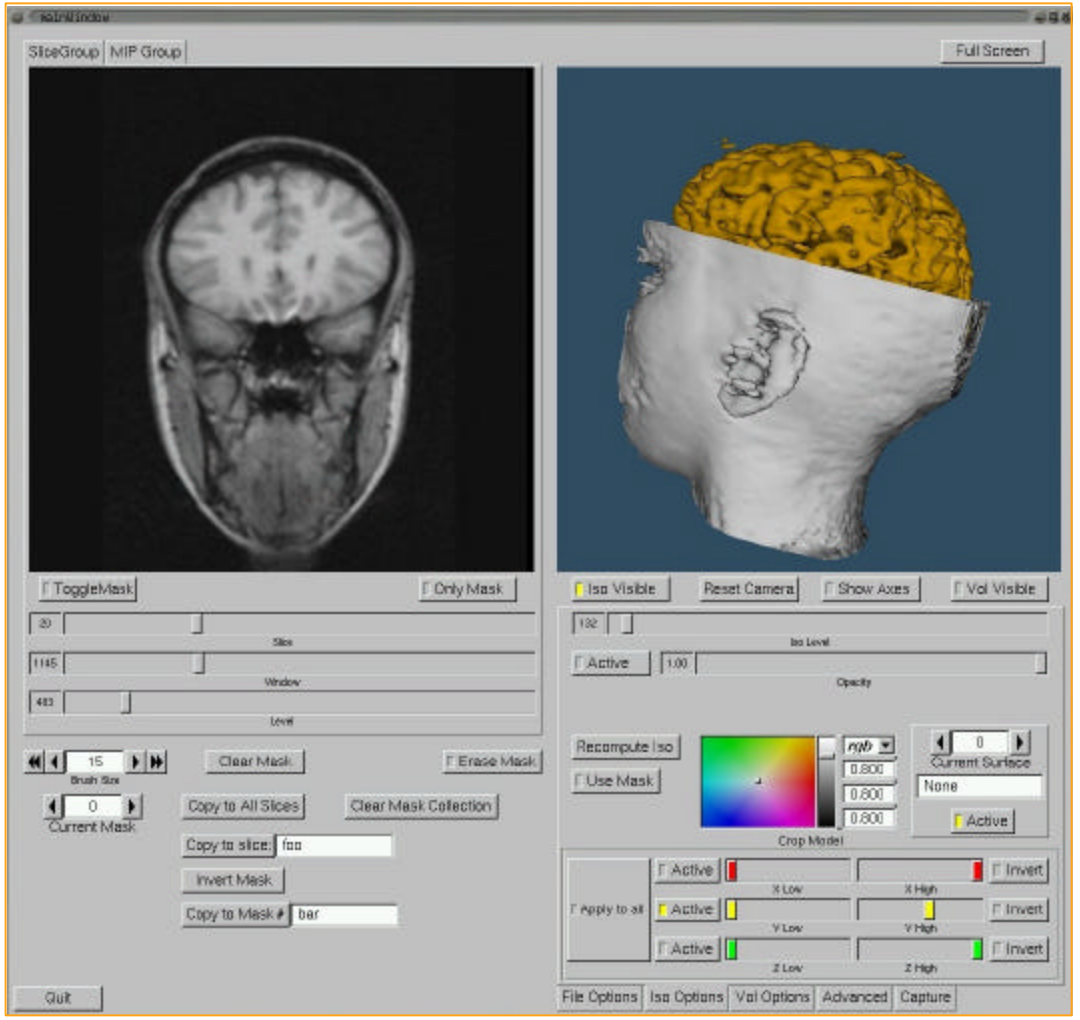
3D Medical Visualization App

- 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



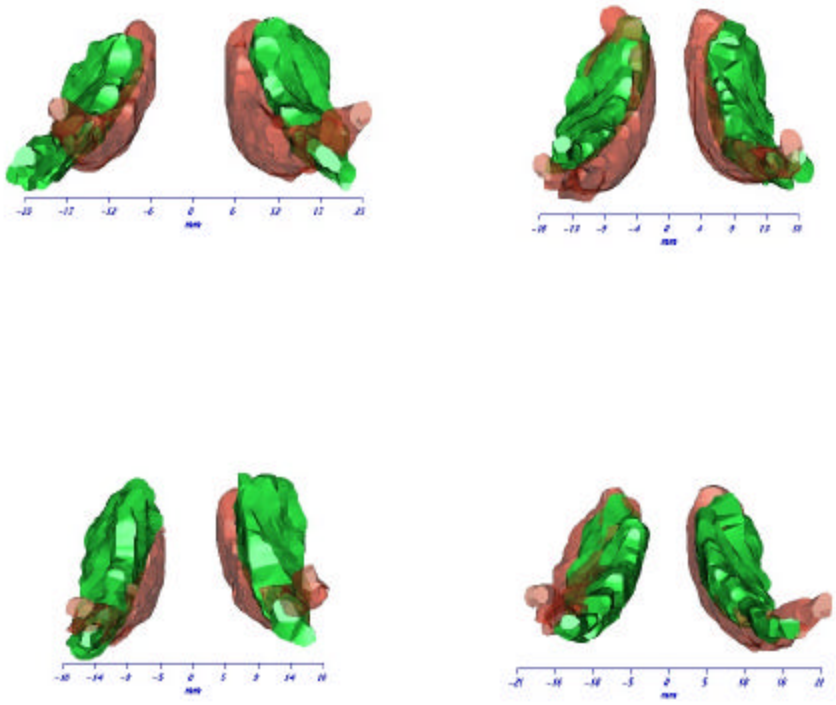
Multiple Sclerosis Project

- **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

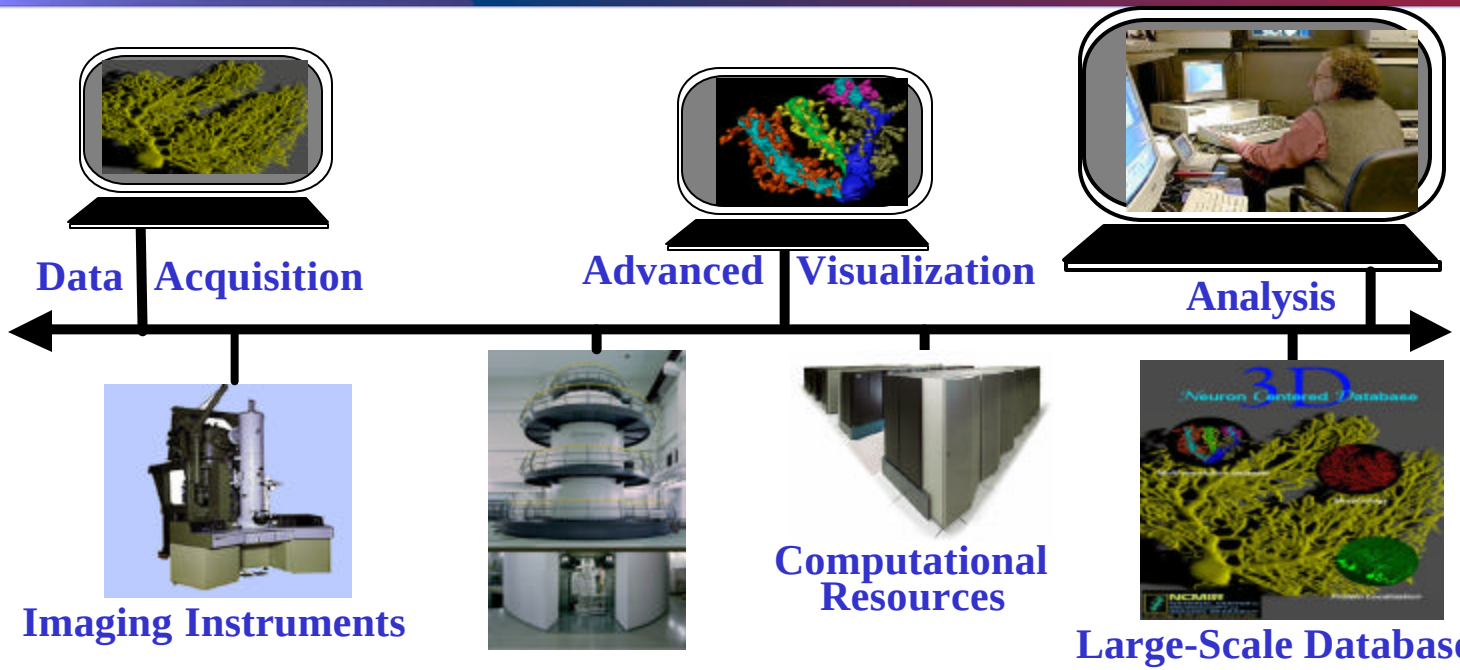
DISCOM

SinRG

APGrid

IPG …

Grid Computing Overview



Thanks to Mark Ellisman

- 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

Computational Grids & Electric Power Grids

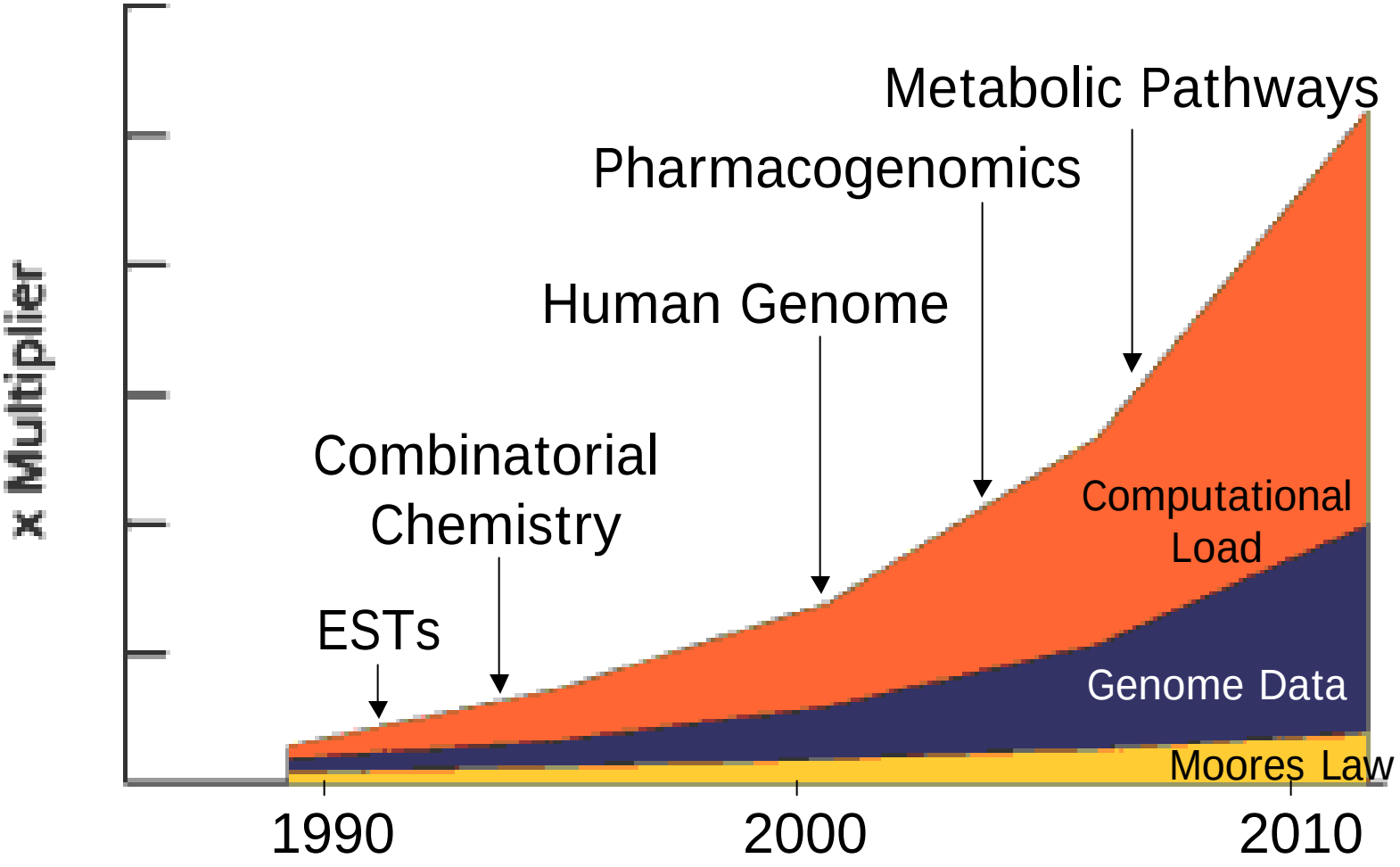
- **Similarities/Goals of CG and EPG**
  - **Ubiquitous**
  - **Consumer is comfortable with lack of knowledge of details**
- **Differences Between CG and EPG**
  - **Wider spectrum of performance & services**
  - **Access governed by more complicated issues**
    - **Security**
    - **Performance**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

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

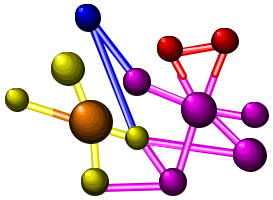
Growth of Data and Load vs. Moore's Law

Courtesy of Rick Stevens



Biomedical Data:
High Complexity and Large Scale

Courtesy of Rick Stevens



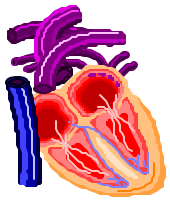
millions

Protein-Protein Interactions
- metabolism
- pathways
- receptor-ligand

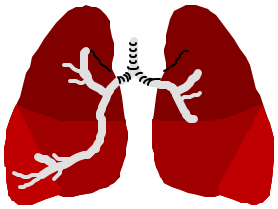
4º structure

billions

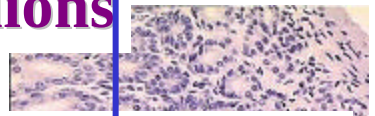
Polymorphism and Variants
- genetic variants
- individual patients
- epidemiology



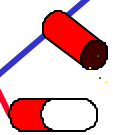
Physiology
- Cellular biology
- Biochemistry
- Neurobiology
- Endocrinology
- etc

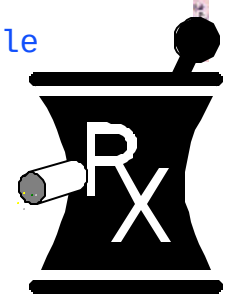


millions



ESTs
Expression patterns
Large-scale screens





Hundred thousands

Proteins
- sequence

2º structure

3º structure

MPMILGYWDIRGLAHAIRLLLEYTDSSYEEKKYT.

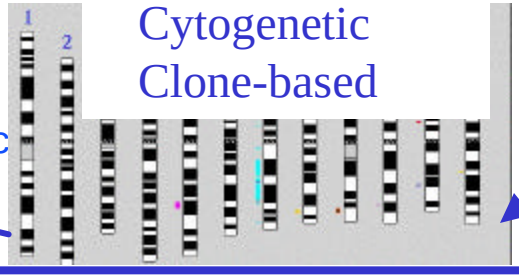
billions

DNA sequences
- alignments

...atcgaattccaggcgtcacattctcaattcc

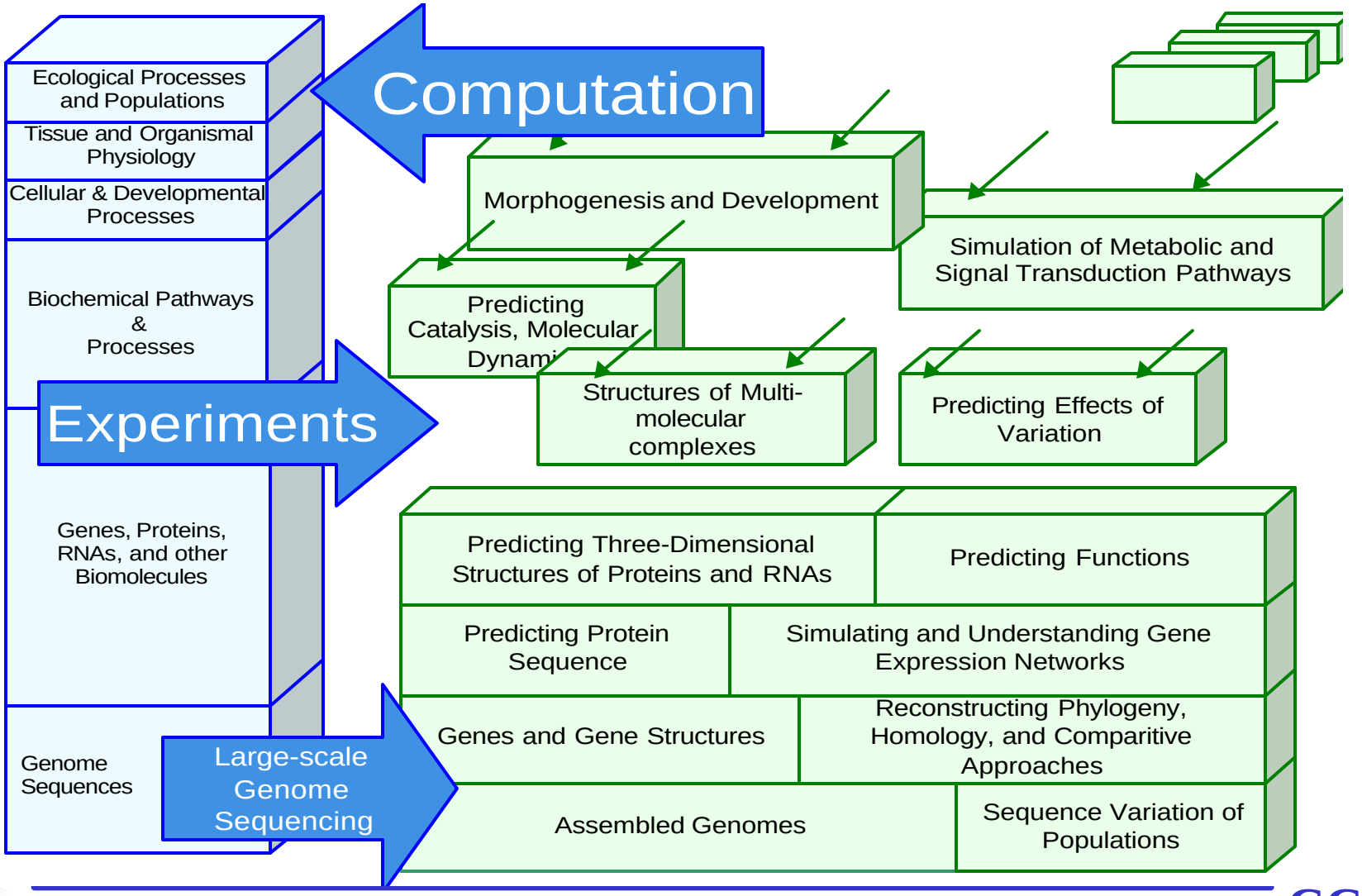
millions

Genetics and Maps
- Linkage
- Cytogenetic
- Clone-based



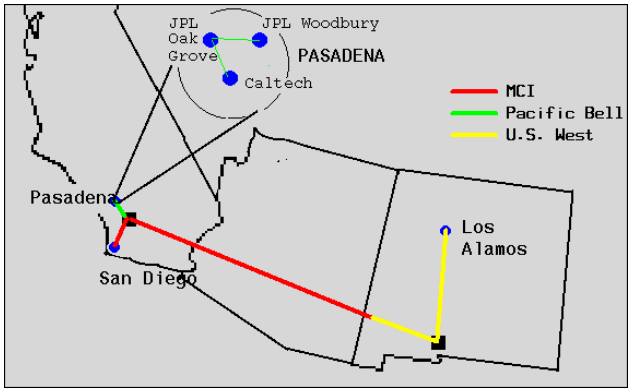
Computational Motivation

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

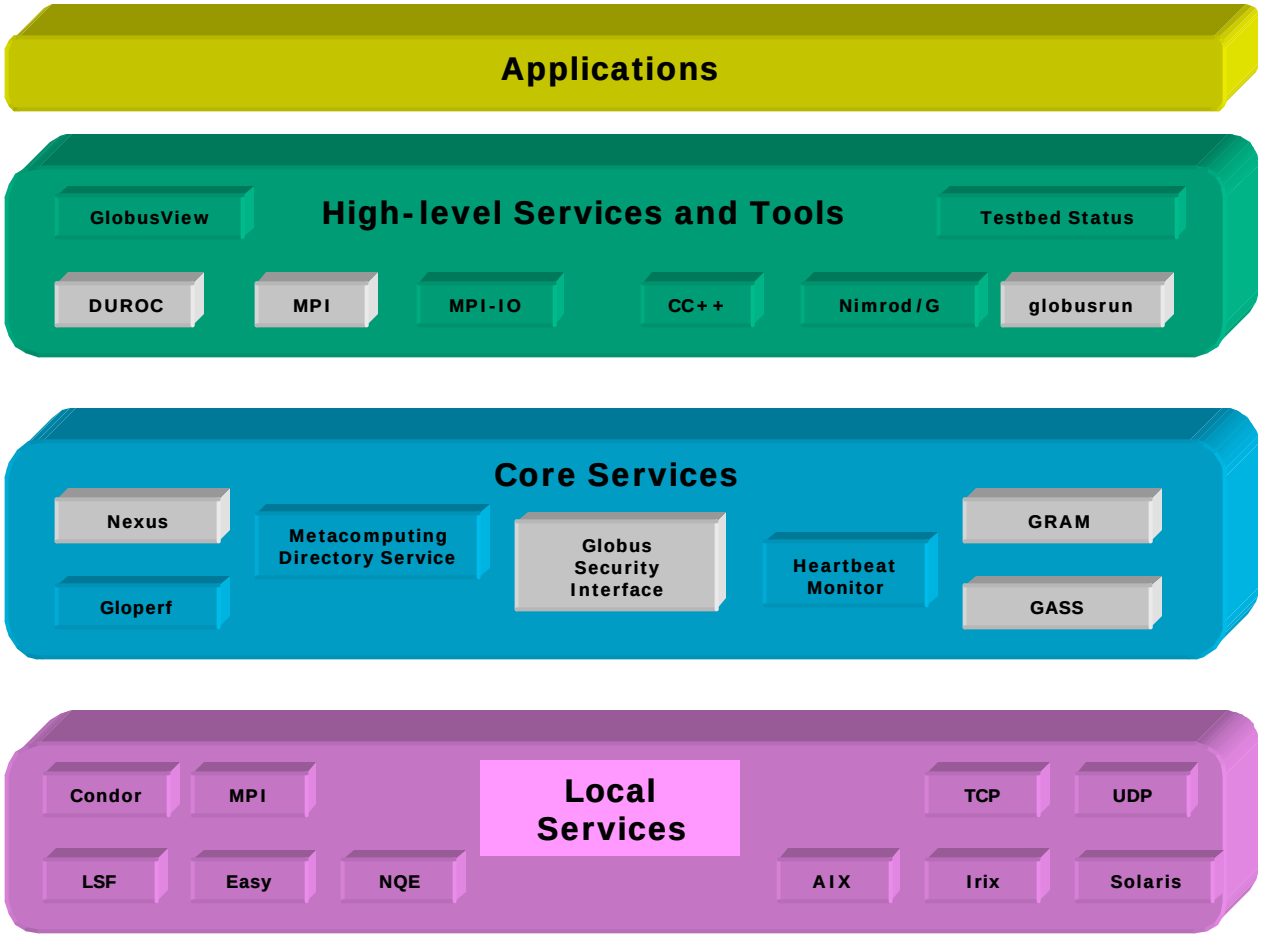


CASA Gigabit Testbed (1990s)

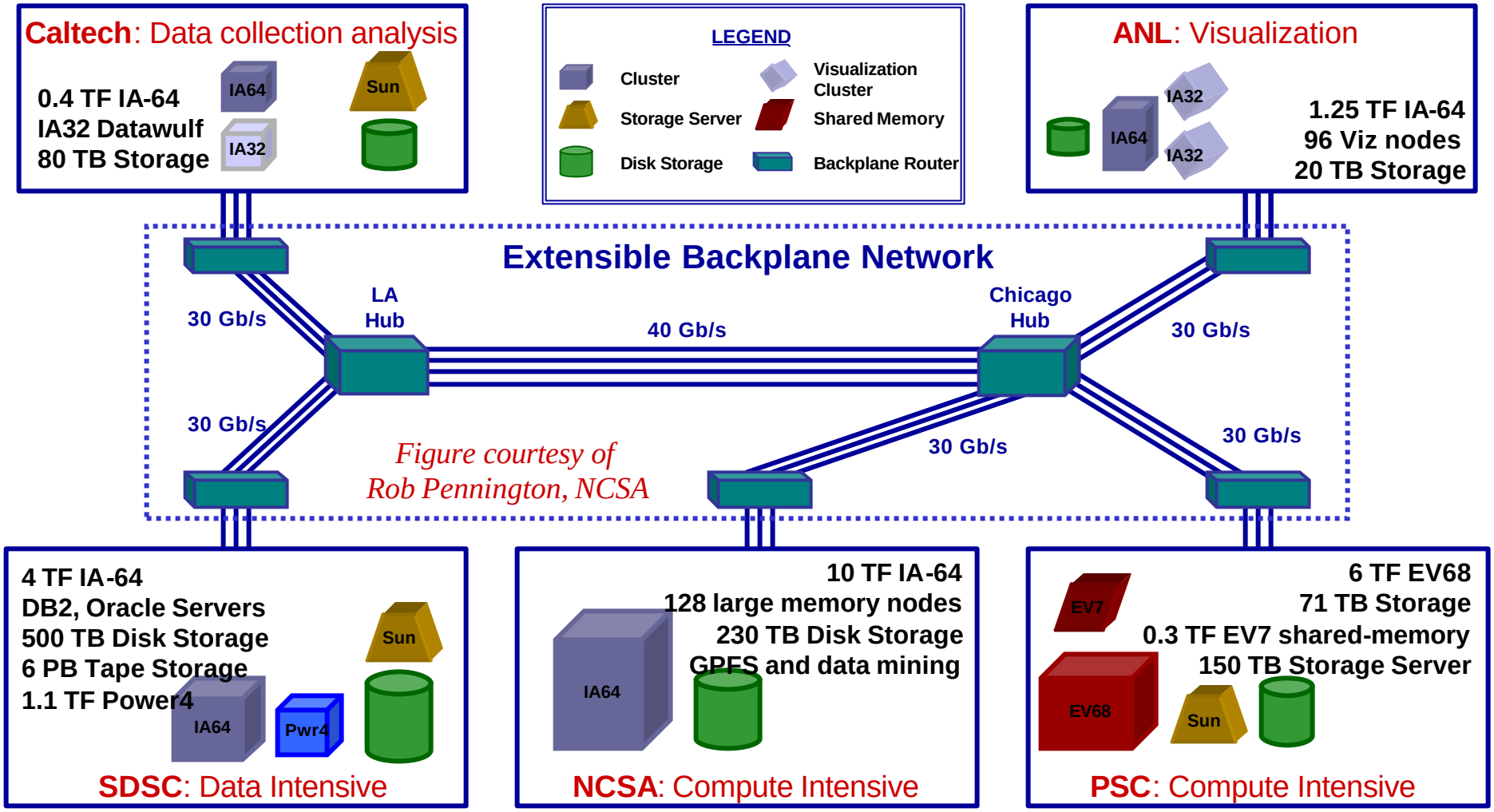
The Globus Project (Ian Foster and Carl Kesselman)

- 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

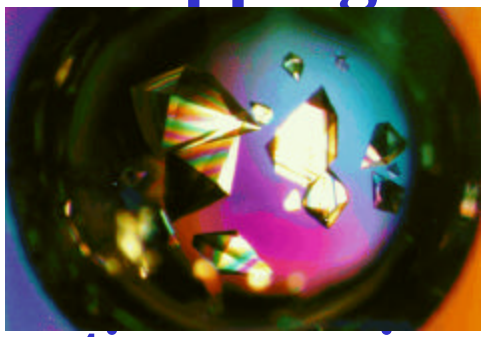


Extensible TeraGrid Facility (ETF)

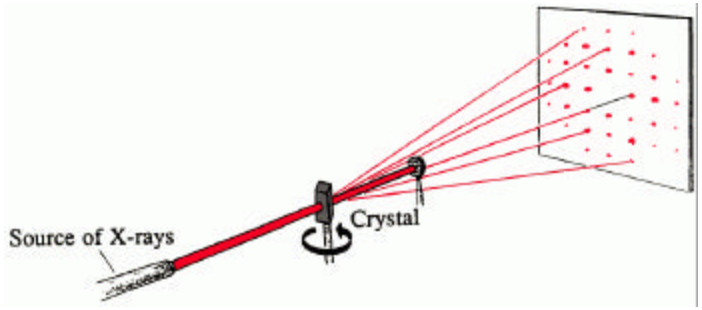


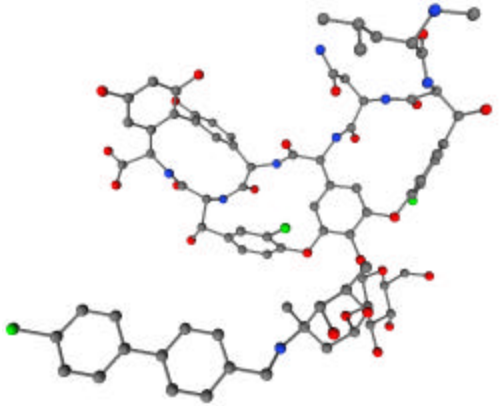
X-Ray Crystallography

- **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.**



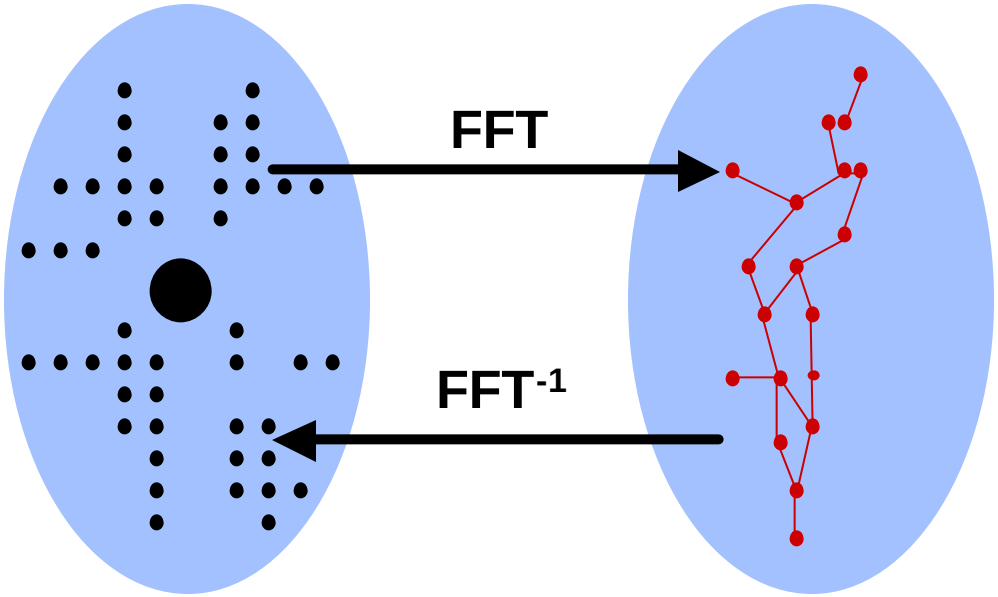


  3. **Determine molecular structure that agrees with diffration data.**

X-Ray Data & Corresponding Molecular Structure

Underlying atomic arrangement is related to the reflections by a 3-D Fourier transform.

Reciprocal or "Phase" Space

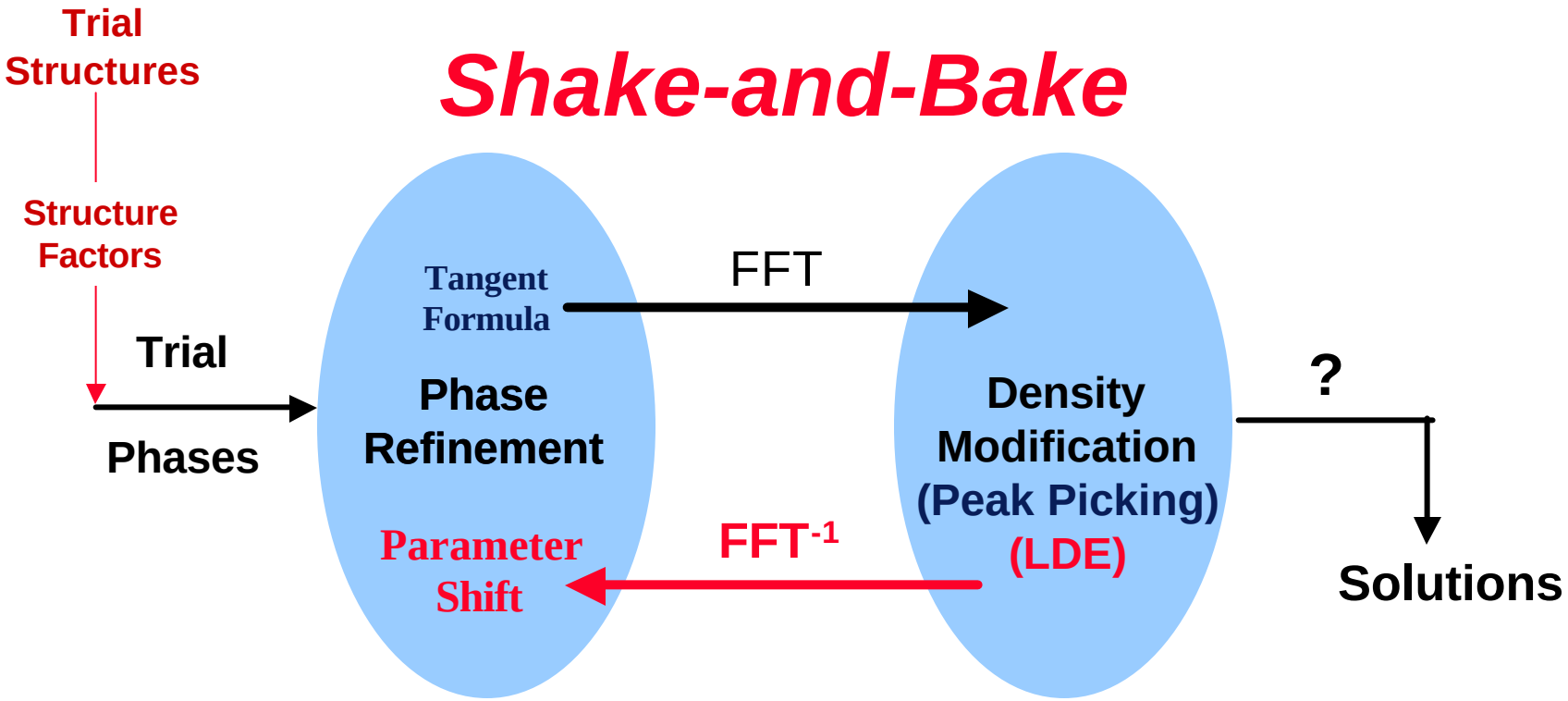


Real Space

X-Ray Data *Molecular Structure*

- **Phases lost during the crystallographic experiment.**
- ***Phase Problem*: Determine phases of the reflections.**

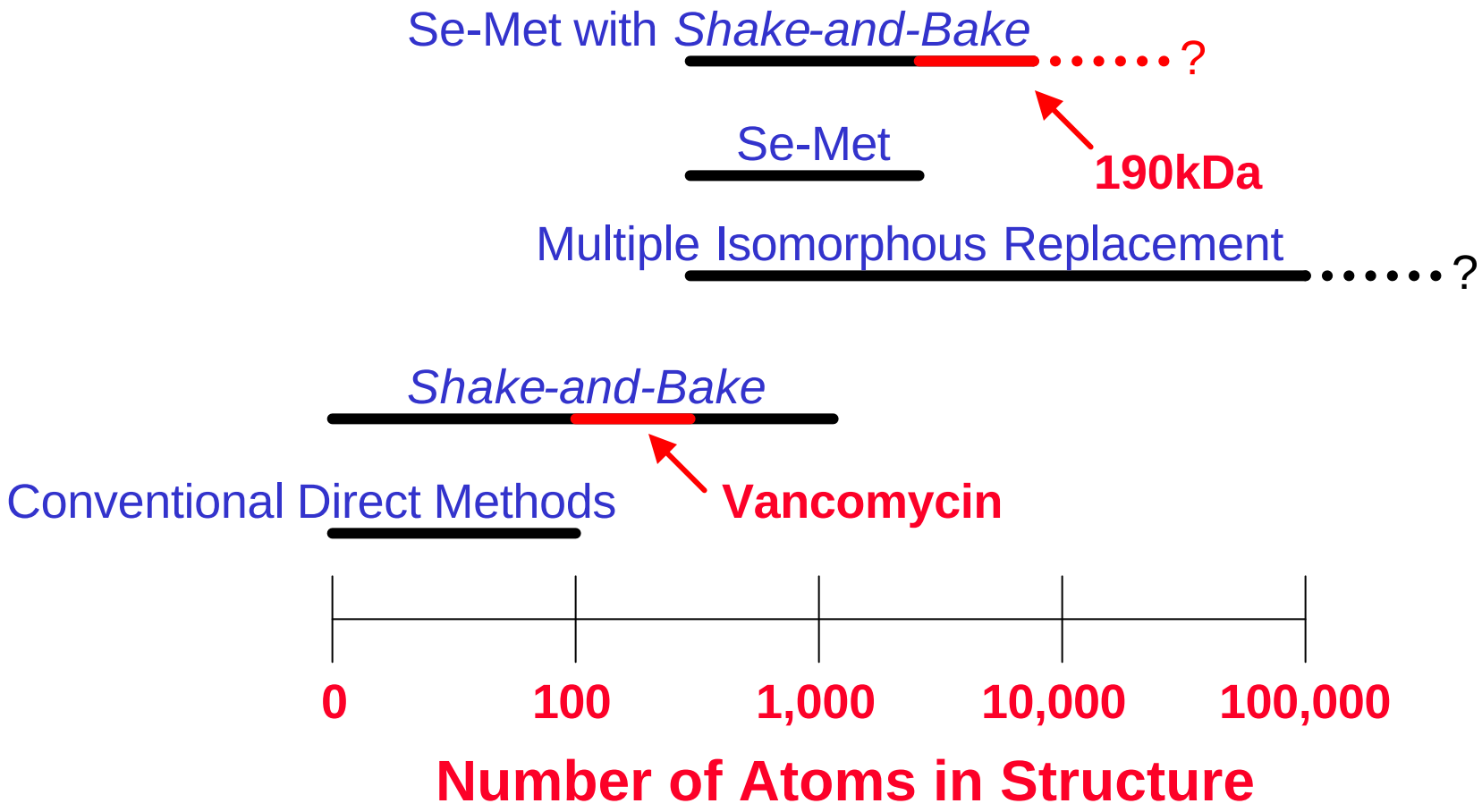
Shake-and-Bake Method: Dual-Space Refinement



Reciprocal Space
"Shake"

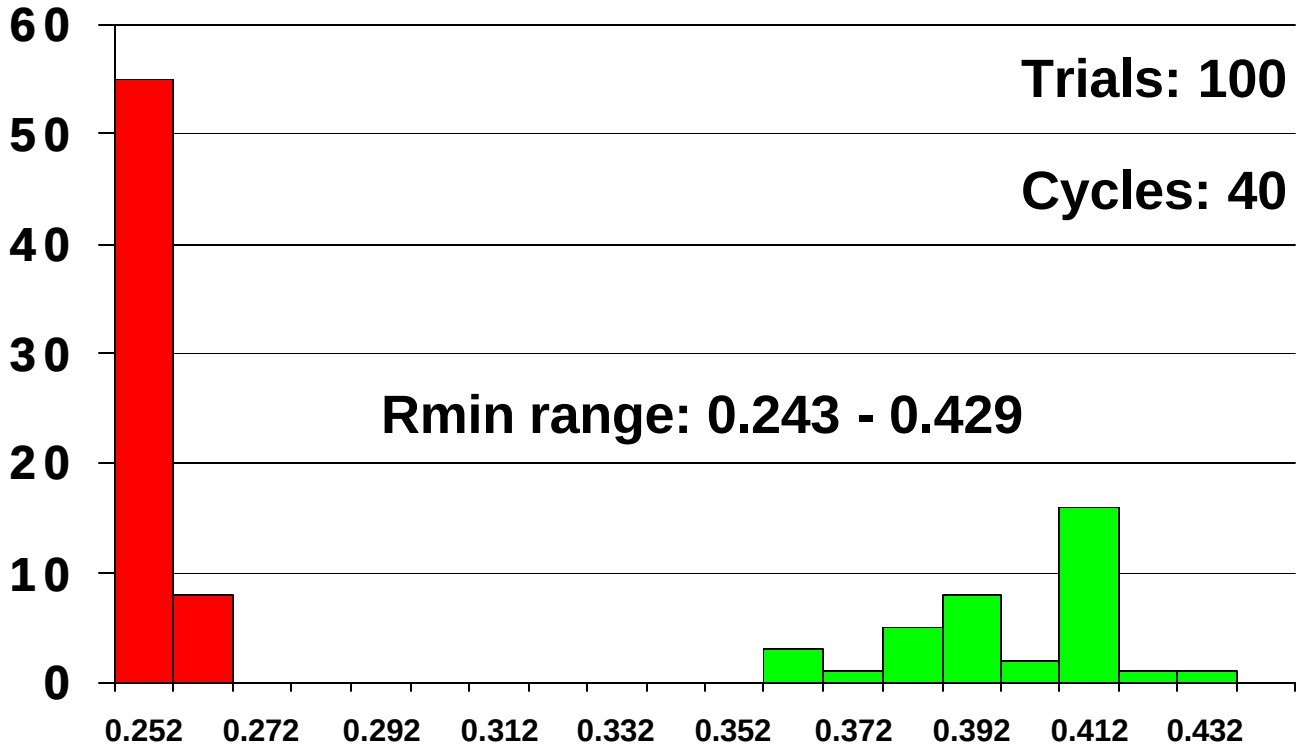
Real Space
"Bake"

Phasing and Structure Size



Ph8755: *SnB* Histogram

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



Grid-Based *SnB* Objectives

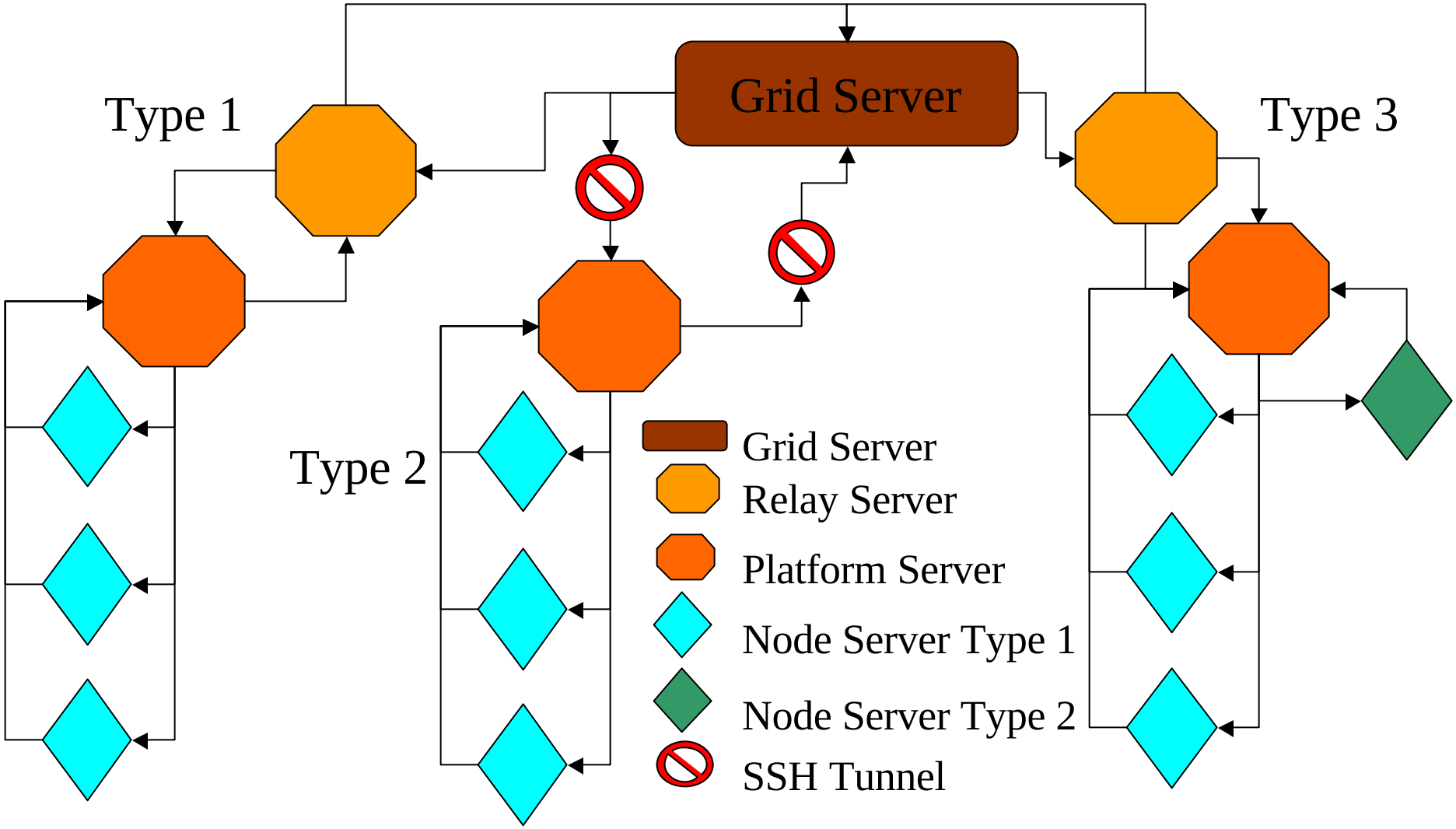
- **Install Grid-Enabled Version of *SnB***
- **Job Submission and Monitoring over Internet**
- ***SnB* Output Stored in Database**
- ***SnB* Output Mined through Internet-Based Integrated Querying Tool**

- **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**

Client/Server Configuration



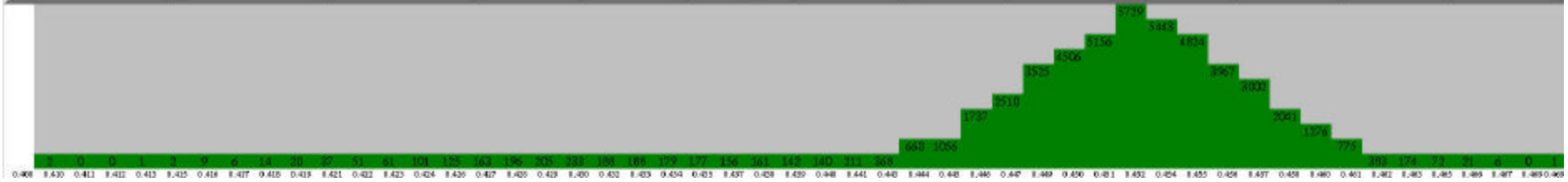
Internet Grid Console

- **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)

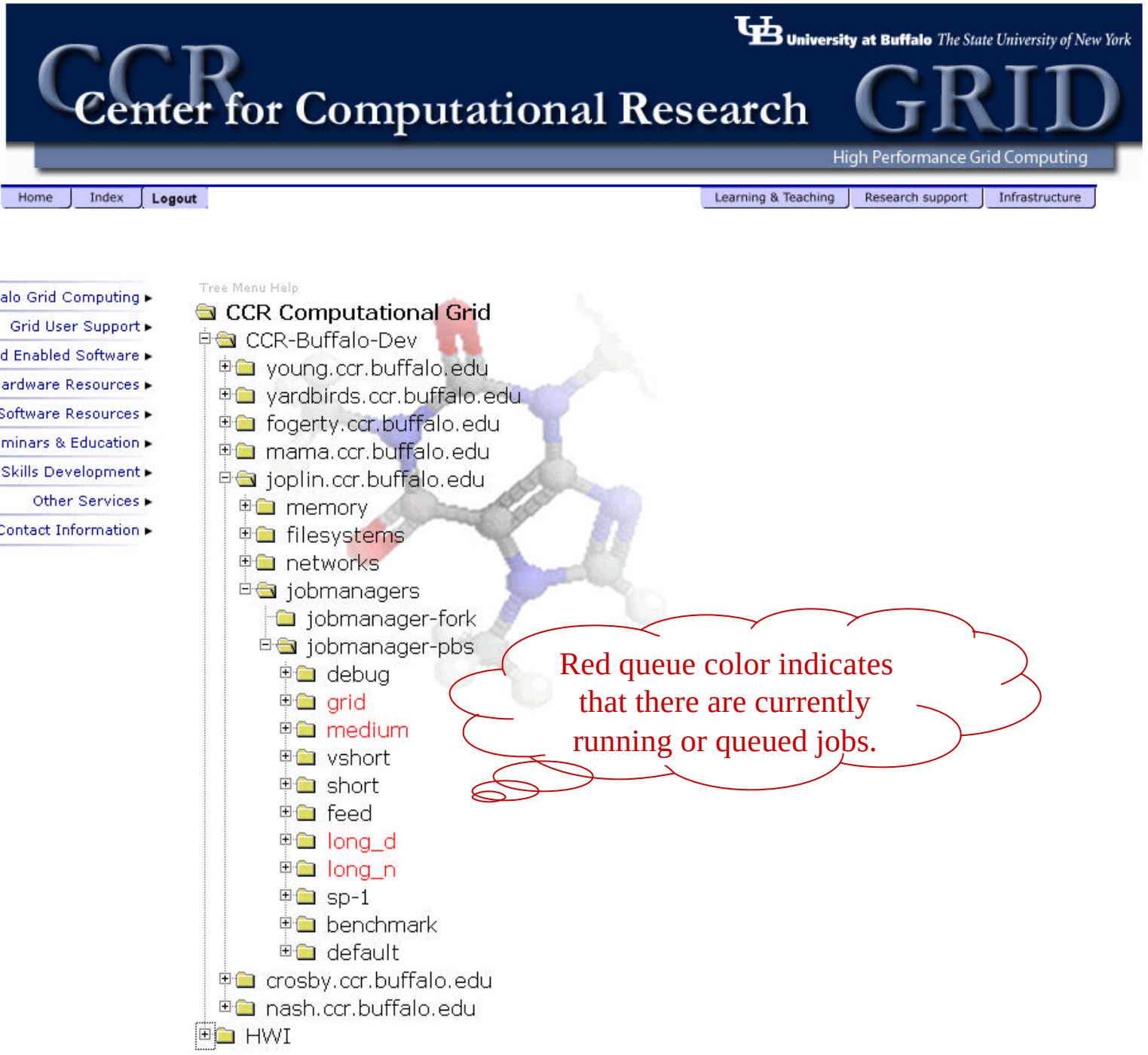
CENTER FOR COMPUTATIONAL RESEARCH
University at Buffalo *The State University of New York*

CONSOLE	GRID SERVER	PLATFORM SERVER STATUS										
COMPLETE	WORKING	IDLE	IDLE	IDLE	IDLE	IDLE	IDLE	IDLE	IDLE	IDLE	IDLE	IDLE
100 % FLOAT	Tue Nov 5 23:29:10 2002							RACING				
	PARALLEL RUN TIME: 188.10 minute							100 %				
											90 %	
	TRIAL/MINUTE: 265.82											
										73 %		
	COMPLETION TIME: 0.00 / 188.10 minute					63 %						
							59 %					
	SERIAL RUN TIME: 114705.08 minute											46 %
		34 %										
	SPEEDUP: 609.81											
READY	READY	READY	OFFLINE	READY	OFFLINE	READY	READY	READY	OFFLINE	READY	READY	READY
1	1	1	0	0	0	41609	42008	1558	0	41974	41144	41244
50000	50000	1557	0	0	0	41973	42501	41143	0	42007	41243	41608
50000	50000 / 50000	544 / 1557	0 / 0	0 / 0	0 / 0	232 / 365	296 / 494	48645 / 39586	0 / 0	25 / 34	90 / 100	168 / 365
Nodes	320	11	0	1	0	4	2	299	0	1	1	1
Process	649	24	0	4	0	6	8	598	0	1	2	6
Available	4	0	0	0	0	0	0	0	0	0	0	0
Running	645	24	0	4	0	6	8	598	0	1	2	6
JOB STATUS	SGI INTEL/ALPHA	SGI INTEL/ALPHA	DNA RNA DELL	SGI 3800 ORIGIN	BRIQS (SOLAR POWERED)	SUN BLADE/ULTRA	IBM SP2 PWR2/PWR3	DELL XEON	IBM 340	IBM 44P	SGI OCTANE	SGI ONYX2
SHAKE-N-BAKE	NASH/MOONGLOWS	NASH/MOONGLOWS	DNA RNA	CROSBY	BRIQ	YOUNG	STILLS	JOPLIN	MAMA PAPAS	COASTERS	THEDOORS	CREAM



Status Report

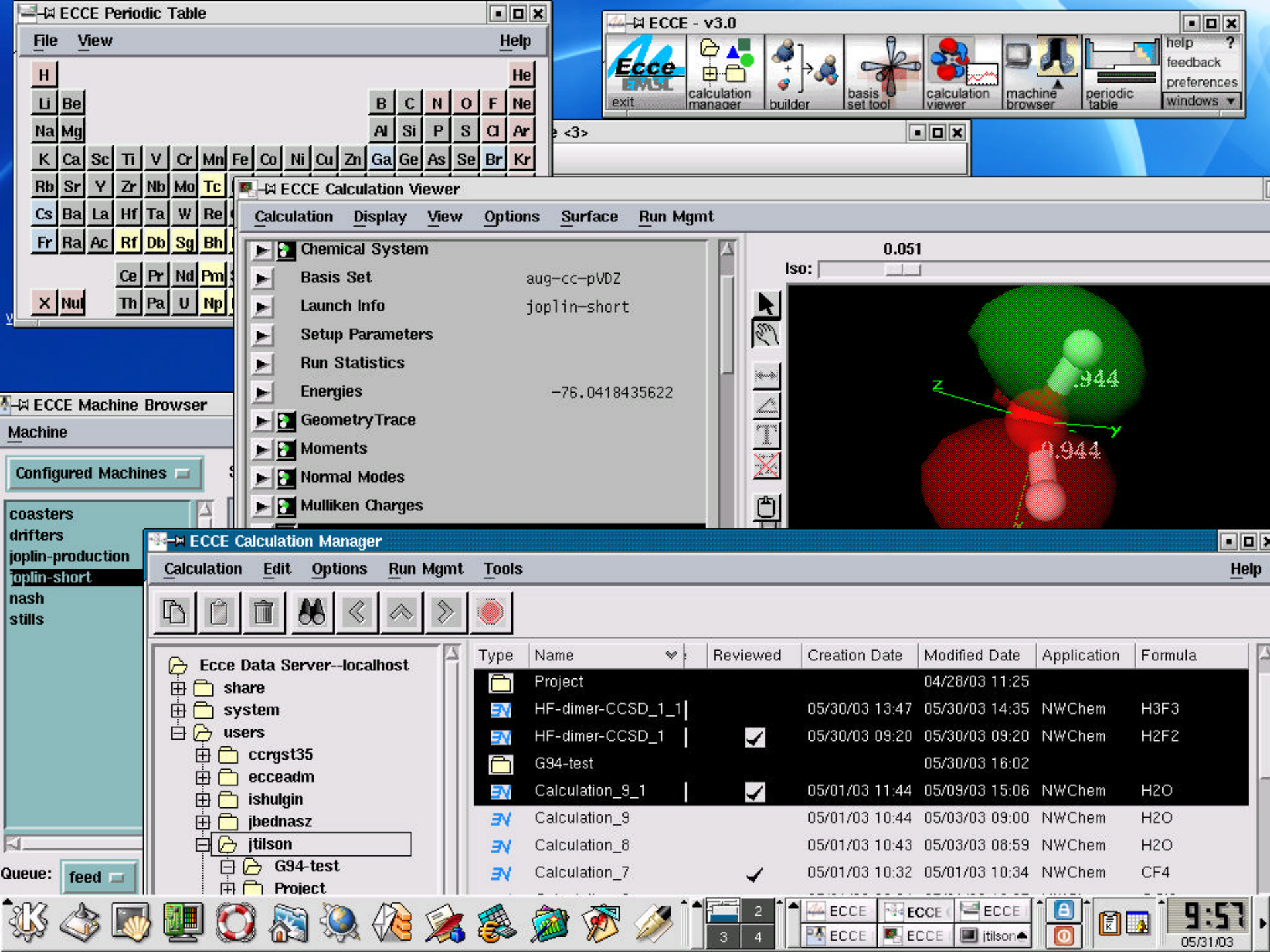
- **Grid Portal**
  - **Access control lists, security groups**
  - **User attributes, history, proxies**
  - **Managed through MySQL database**
- **Globus**
  - **Vers 2.2.4 installed and in production**
  - **Metacomputing Directory Services (MDS) stored in MySQL**
    - **Eliminates need for LDAP**
- **Condor and Condor-G**
  - **Used for resource management and grid job submissions**


University at Buffalo The State University of New York
CCR Center for Computational Research GRID
High Performance Grid Computing
Home
Index
Logout
Learning & Teaching
Research support
Infrastructure
Buffalo Grid Computing
Grid User Support
Grid Enabled Software
Hardware Resources
Software Resources
Seminars & Education
Skills Development
Other Services
Contact Information
Tree Menu Help
CCR Computational Grid
CCR-Buffalo-Dev
young.ccr.buffalo.edu
yardbirds.ccr.buffalo.edu
fogerty.ccr.buffalo.edu
mama.ccr.buffalo.edu
joplin.ccr.buffalo.edu
memory
filesystems
networks
jobmanagers
jobmanager-fork
jobmanager-pbs
debug
grid
medium
vshort
short
feed
long_d
long_n
sp-1
benchmark
default
crosby.ccr.buffalo.edu
nash.ccr.buffalo.edu
HWI
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


ECCE Periodic Table
File View Help
ECCE - v3.0
exit
calculation manager
builder
basis set tool
calculation viewer
machine browser
periodic table
help
feedback
preferences
windows
ECCE Calculation Viewer
Calculation Display View Options Surface Run Mgmt
Chemical System
Basis Set aug-cc-pVDZ
Launch Info joplin-short
Setup Parameters
Run Statistics
Energies -76.0418435622
GeometryTrace
Moments
Normal Modes
Mulliken Charges
0.051
Iso:
ECCE Machine Browser
Machine
Configured Machines
coasters
drifters
joplin-production
joplin-short
nash
stills
Queue: feed
ECCE Calculation Manager
Calculation Edit Options Run Mgmt Tools Help
Ecce Data Server--localhost
share
system
users
ccrgst35
ecceadm
ishulgin
jbednasz
jtilson
G94-test
Project
Type Name Reviewed Creation Date Modified Date Application Formula
Project 04/28/03 11:25
HF-dimer-CCSD_1_1 05/30/03 13:47 05/30/03 14:35 NWChem H3F3
HF-dimer-CCSD_1 05/30/03 09:20 05/30/03 09:20 NWChem H2F2
G94-test 05/30/03 16:02
Calculation_9_1 05/01/03 11:44 05/09/03 15:06 NWChem H2O
Calculation_9 05/01/03 10:44 05/03/03 09:00 NWChem H2O
Calculation_8 05/01/03 10:43 05/03/03 08:59 NWChem H2O
Calculation_7 05/01/03 10:32 05/01/03 10:34 NWChem CF4
9:57
05/31/03

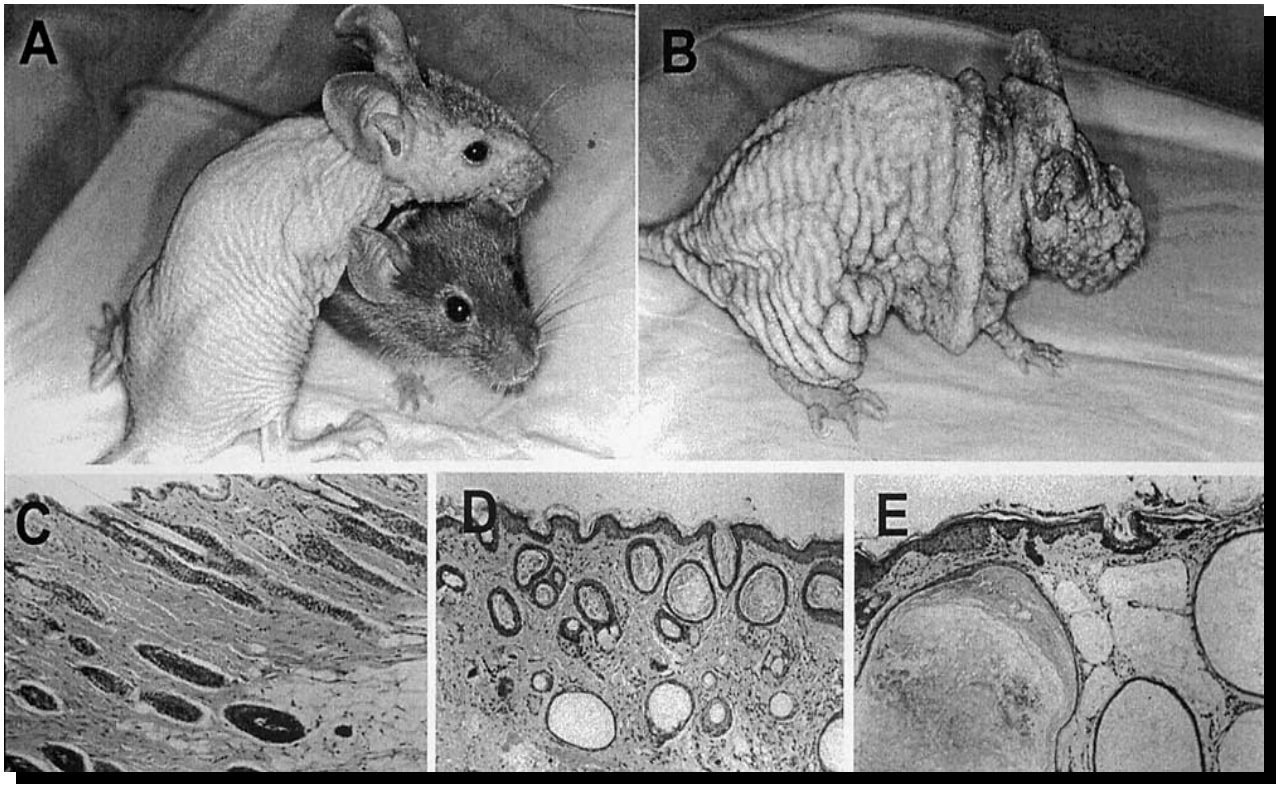
"Genomics is powering the new biology, but Computing is in the driver's seat."

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