

The Center for Computational Research: Grid, Visualization, and BioMedical Computing

Russ Miller

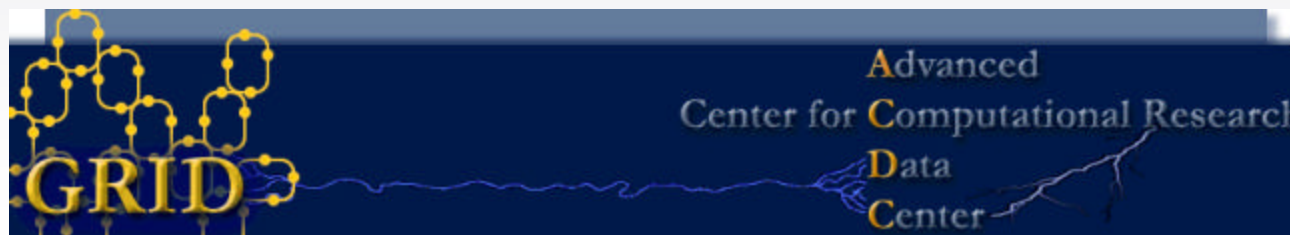
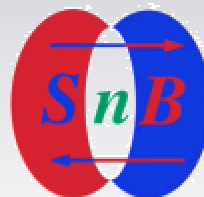
Center for Computational Research

Computer Science & Engineering

SUNY-Buffalo

Hauptman-Woodward Medical Inst

NSF, NIH, DOE
NIMA, NYS, HP



University at Buffalo

The State University of New York

Center for Computational Research 1999-2004 Snapshot

■ High-Performance Computing and High-End Visualization

- ❑ 110 Research Groups in 27 Depts
- ❑ 13 Local Companies
- ❑ 10 Local Institutions

■ External Funding

- ❑ \$116M External Funding
 - \$16M as lead
 - \$100M in support
- ❑ \$43M Vendor Donations
- ❑ Total Leveraged: \$0.5B

■ Deliverables

- ❑ 400+ Publications
- ❑ Software, Media, Algorithms, Consulting, Training, CPU Cycles...



Major Compute Resources

- **Dell Linux Cluster: #22® #25® #38® #95**
 - ❑ 600 P4 Processors (2.4 GHz)
 - ❑ 600 GB RAM; 40 TB Disk; Myrinet
- **SGI Origin3700 (Altix)**
 - ❑ 64 Processors (1.3GHz ITF2)
 - ❑ 256 GB RAM
 - ❑ 2.5 TB Disk
- **SGI Origin3800**
 - ❑ 64 Processors (400 MHz)
 - ❑ 32 GB RAM; 400 GB Disk
- **Apex Bioinformatics System**
 - ❑ Sun V880 (3), Sun 6800
 - ❑ Sun 280R (2)
 - ❑ Intel PIIIs
 - ❑ Sun 3960: 7 TB Disk Storage
- **HP/Compaq SAN**
 - ❑ 75 TB Disk
 - ❑ 190 TB Tape
 - ❑ 64 Alpha Processors (400 MHz)
 - ❑ 32 GB RAM; 400 GB Disk
- **Dell Linux Cluster: #187® #368® off**
 - ❑ 4036 Processors (PIII 1.2 GHz)
 - ❑ 2TB RAM; 160TB Disk; 16TB SAN
- **IBM BladeCenter Cluster: #106**
 - ❑ 532 P4 Processors (2.8 GHz)
 - ❑ 5TB SAN
- **IBM RS/6000 SP: 78 Processors**
- **Sun Cluster: 80 Processors**
- **SGI Intel Linux Cluster**
 - ❑ 150 PIII Processors (1 GHz)
 - ❑ Myrinet

CCR's BioACE System

■ BioACE Computing Environment

- ❑ SunFire 6800 (12P, 24GB), 2 SunFire V880's (16P, 32GB)
- ❑ 16 RLX Pentium 3 Server Blades
- ❑ 104 GB of RAM; 7 TB of disk storage

■ Software

❑ Genomics Packages

- GCG, Vector NTI, Vector Xpression, Vector PathBlazer

❑ Sequence Analysis

- EMBOSS, ClustalW, MUMmer,

❑ Database Search

- Blast, PSI Blast, HMMER

❑ Gene Expression

- Cluster/Xcluster, TreeView, J-Express

❑ Statistics Packages

- R & Bioconductor

CCR Visualization Resources

■ Fakespace ImmersaDesk R2

- Portable 3D Device

■ Tiled-Display Wall

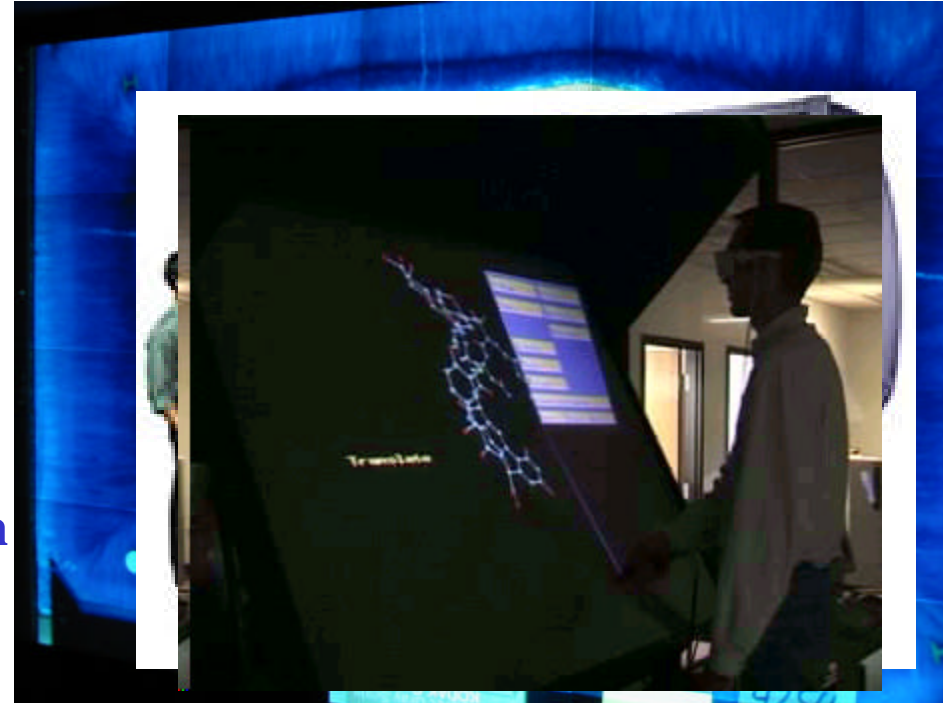
- 20 NEC projectors: 15.7M pixels
- Screen is 11' ´ 7'
- Dell PCs with Myrinet2000

■ Access Grid Nodes (2)

- Group-to-Group Communication
- Commodity components

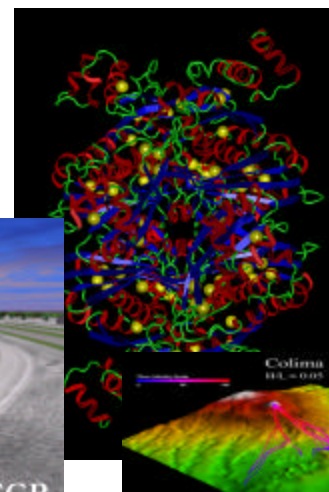
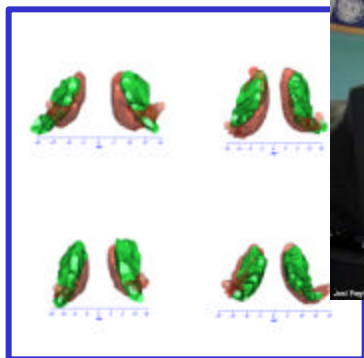
■ SGI Reality Center 3300W

- Dual Barco's on 8' ´ 4' screen



CCR Research & Projects

- Ground Water Modeling
- Computational Fluid Dynamics
- Molecular Structure Determination via *Shake-and-Bake*
- 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

- ❑ “Top Algorithms of the Century”

■ Worldwide Utilization

■ Critical Step

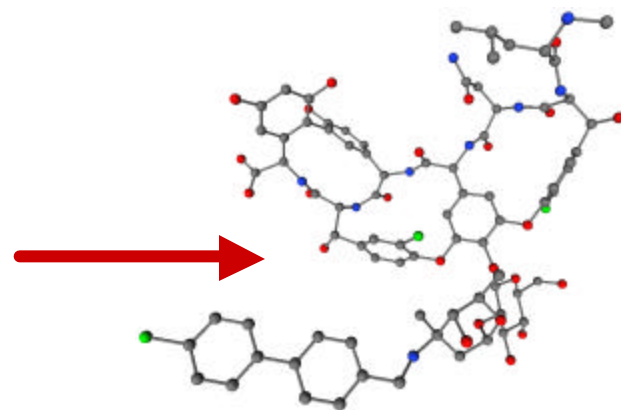
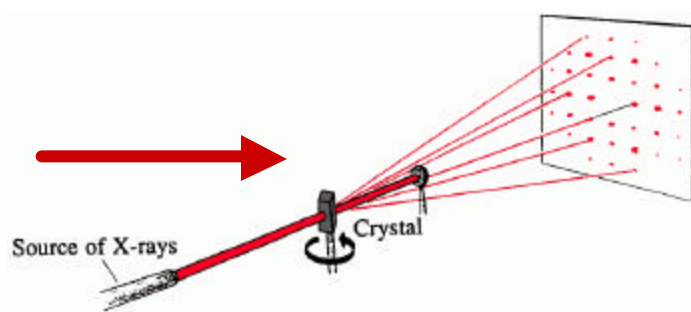
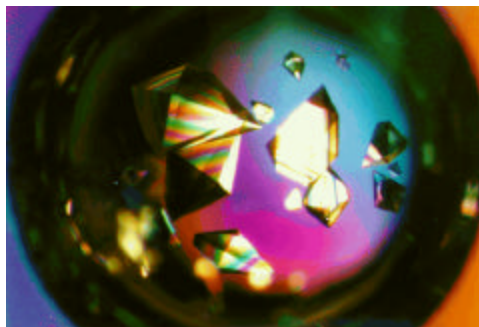
- ❑ Rational Drug Design
- ❑ Structural Biology
- ❑ Systems Biology

■ Vancomycin

- ❑ “Antibiotic of Last Resort”

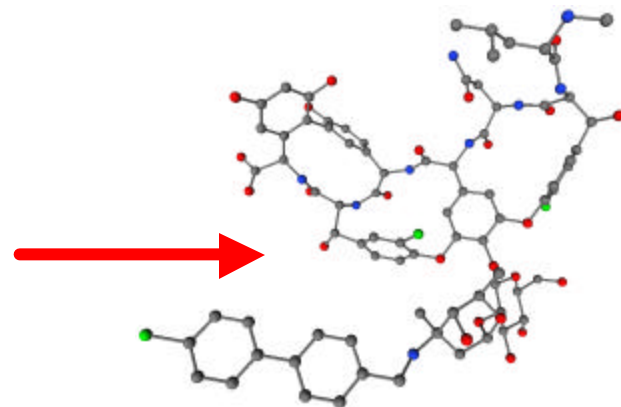
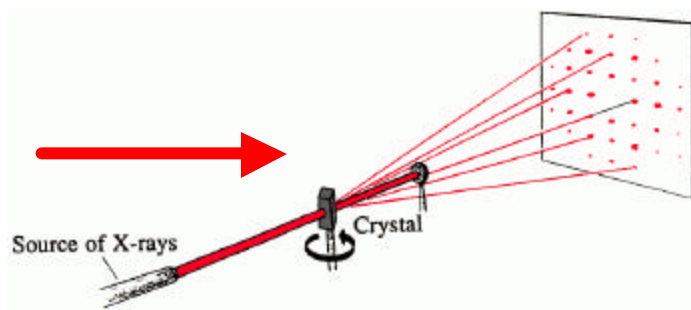
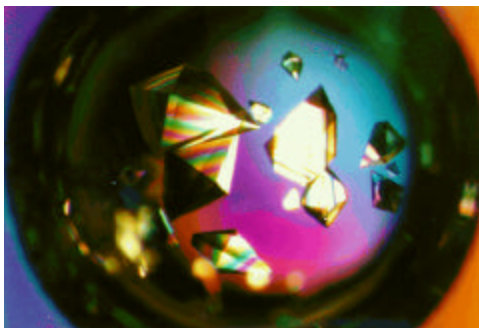
■ Current Efforts

- ❑ Grid
- ❑ Collaboratory
- ❑ Intelligent Learning



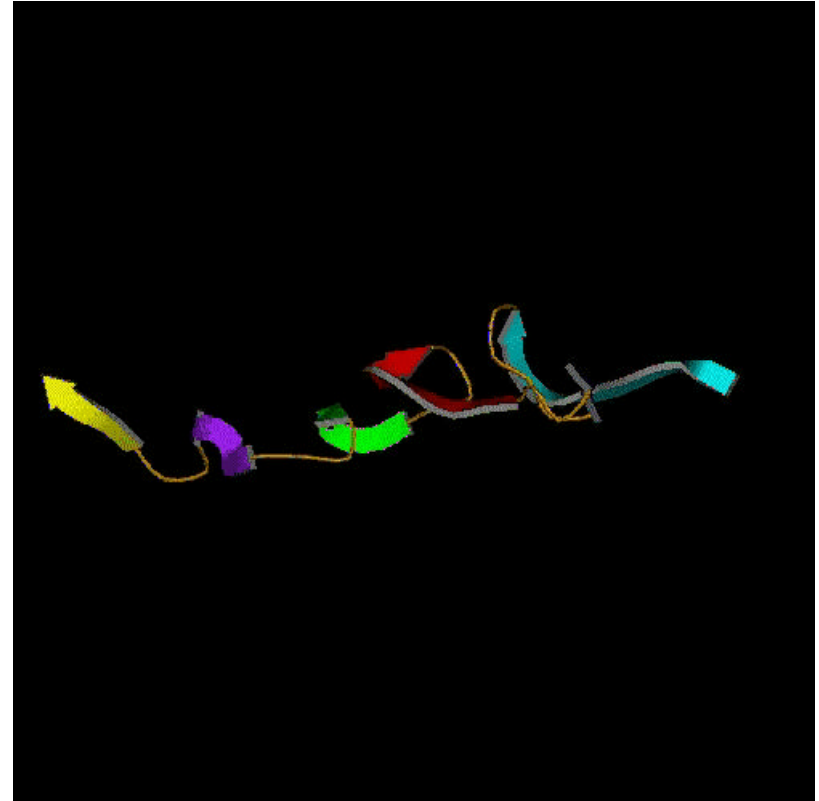
X-Ray Crystallography

- **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 diffraction data.



Protein Folding

- Ability of protein to perform biological function is attributed to 3D structure
- Protein folding problem
 - Predict 3D structure from amino-acid sequence
- Solving the folding problem impacts drug design
- Research underway at UB on the development of models to improve accuracy and efficiency of 3D prediction
- 4000 processor Dell P3 cluster dedicated solely to protein folding problem



Computational Chemistry

■ UB Software Development in Quantum Chemistry

- ❑ **Q-Chem** – development of parallel algorithms and combined QM/MM methods for large molecular systems
- ❑ **ADF** – development of algorithms to calculate magnetic and optical properties of molecules

■ Used to determine

- ❑ Molecular Structure
- ❑ Electronic Spectra
- ❑ Chemical Reactivity

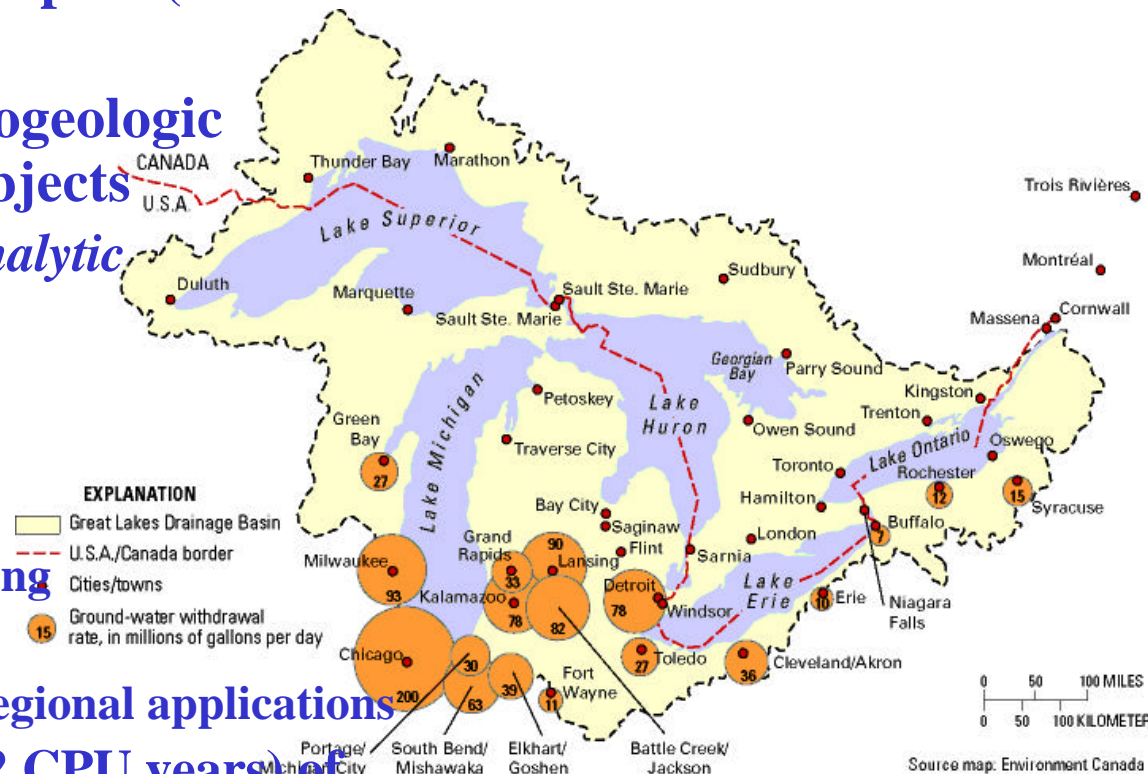
■ Applications

- ❑ Pharmaceutical Drug Design
- ❑ Industrial Catalysis
- ❑ Materials Science
- ❑ Nanotechnology
- ❑ Solution Phase Chemistry
- ❑ Chemical Kinetics



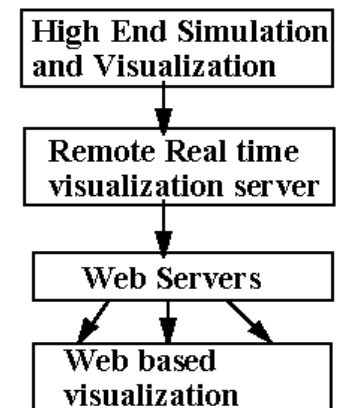
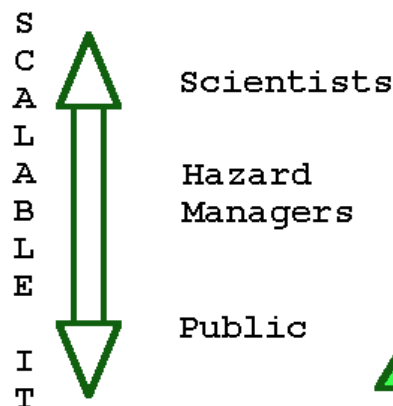
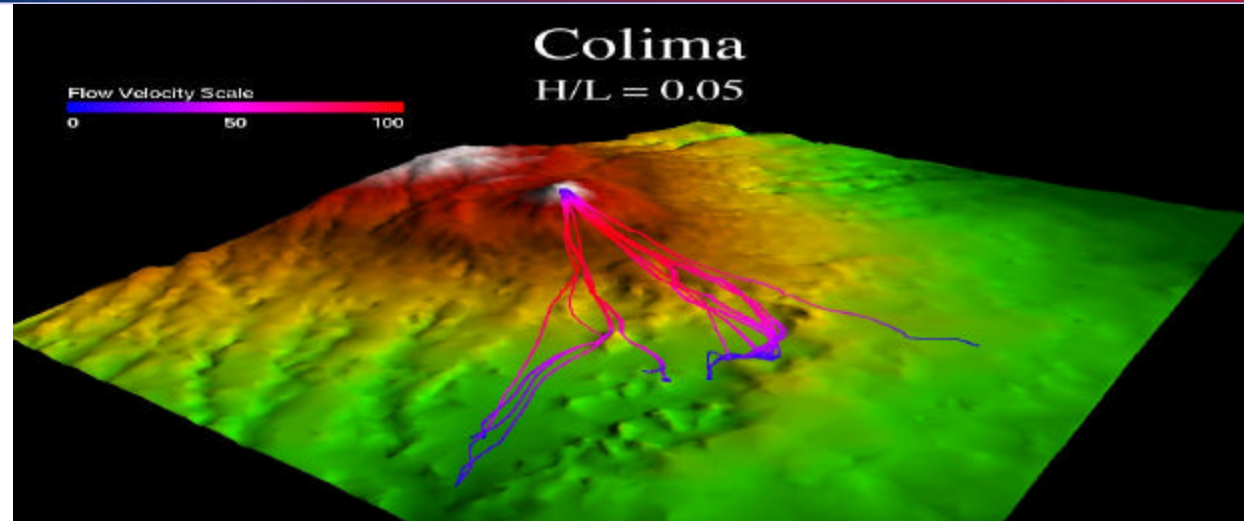
Groundwater Flow Modeling

- Regional-scale modeling of groundwater flow and contaminant transport (Great Lakes Region)
- Ability to include all hydrogeologic features as independent objects
- Current work is based on *Analytic Element Method*
- Key features:
 - High precision
 - Highly parallel
 - Object-oriented programming
 - Intelligent user interface
 - GIS facilitates large-scale regional applications
- Utilized 10,661 CPU days (32 CPU years) of computing in past year on CCR's commodity clusters



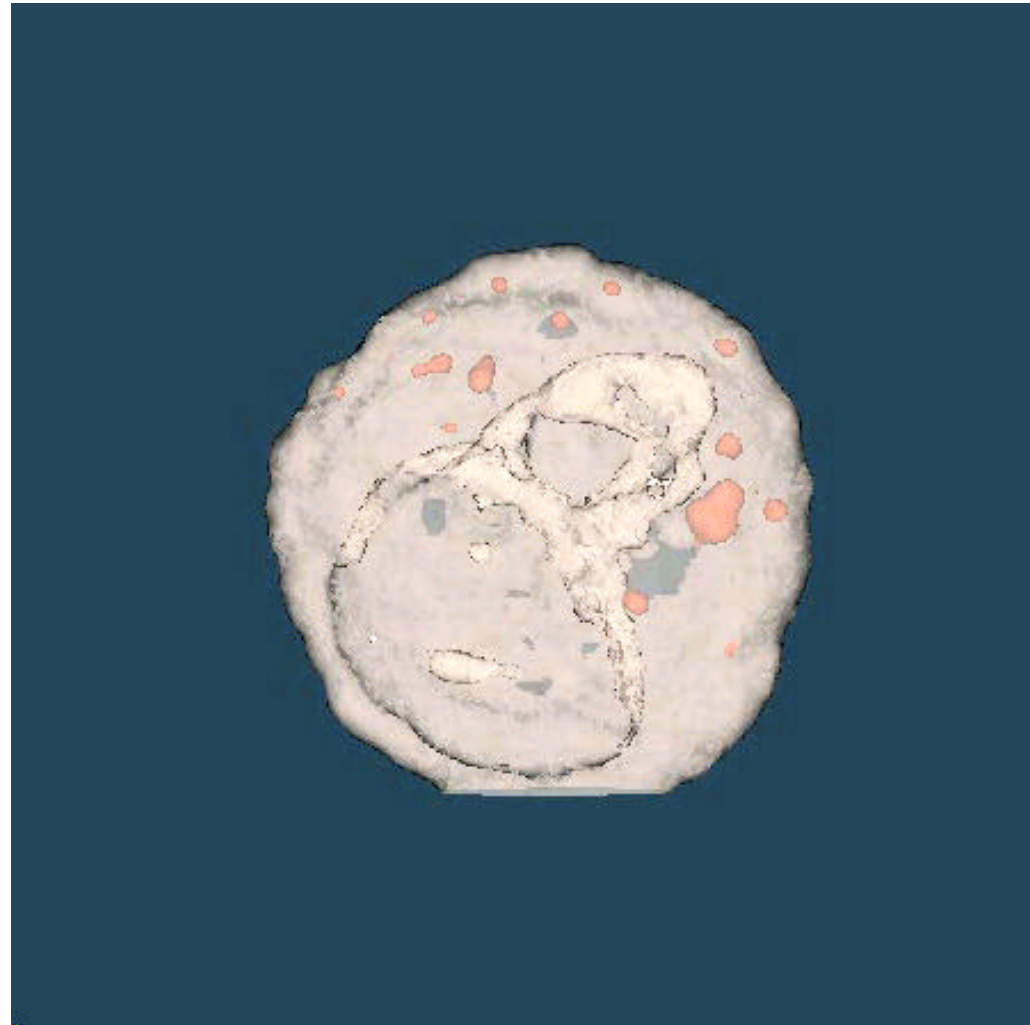
Geophysical Mass Flow Modeling

- Modeling of Volcanic Flows, Mud flows (flash flooding), and Avalanches
- Integrate information from several sources
 - Simulation results
 - Remote sensing
 - GIS data
- Develop realistic 3D models of mass flows
- Present information at appropriate level



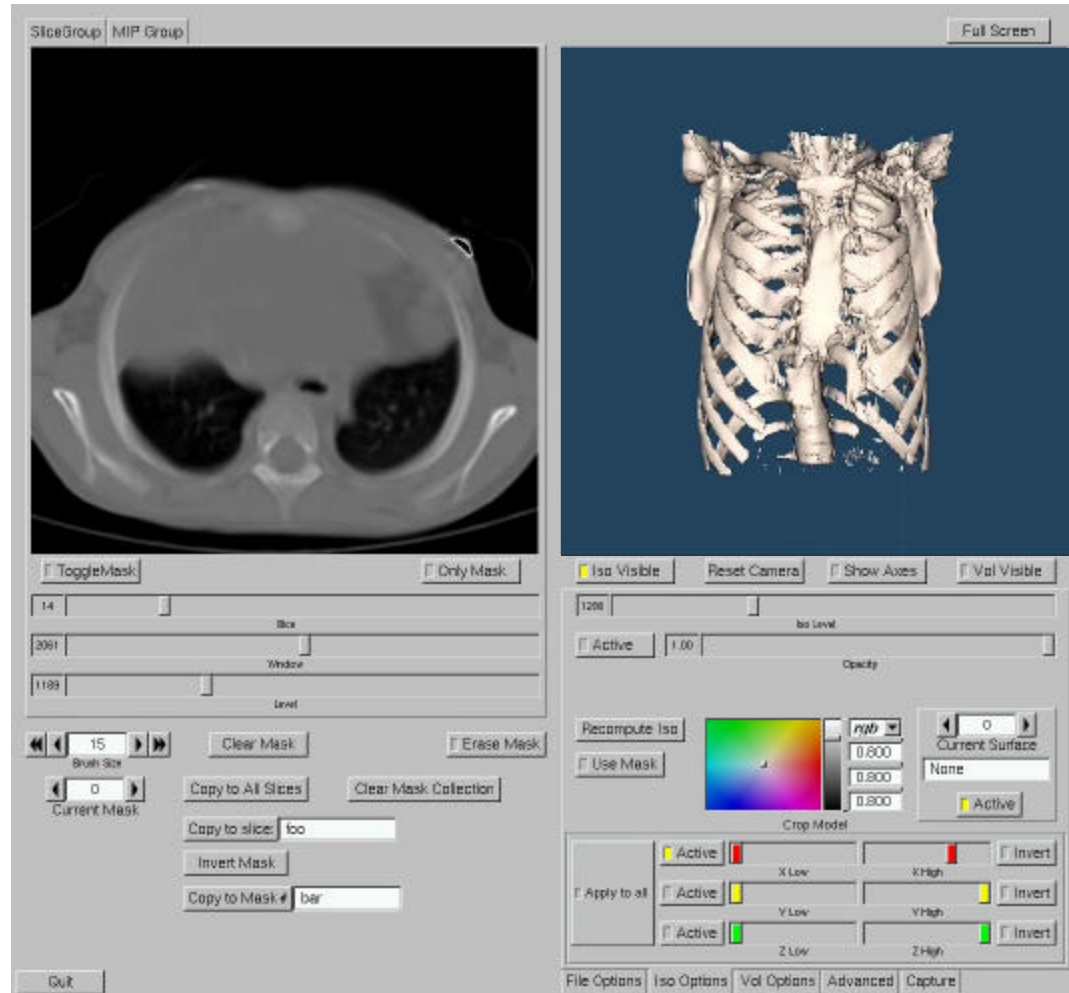
Confocal Microscopy

- **3D Reconstruction of an Oral Epithelial Cell**
- **Translucent White Surface Represents the Cell Membrane**
- **Reddish Surface Represents Groups of Bacteria**



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

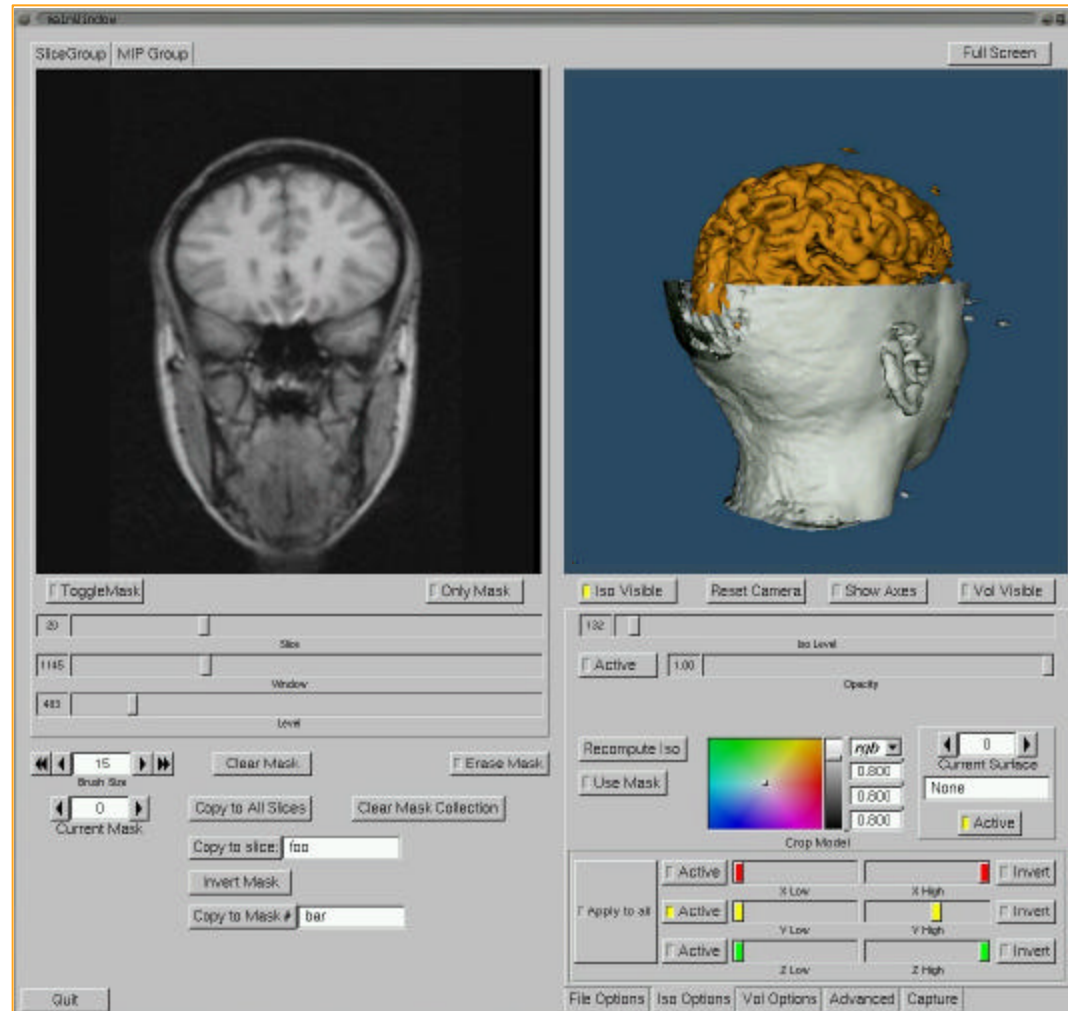


Multiple Sclerosis Project

- Collaboration with Buffalo Neuroimaging Analysis Center (BNAC)

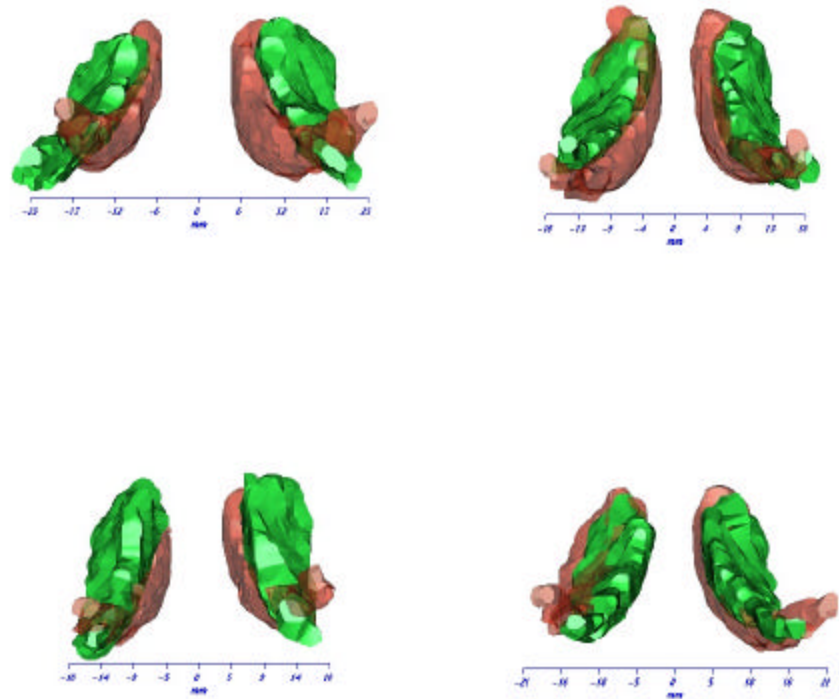
- Developers of Avonex, drug of choice for treatment of MS

- MS Project examines patients and compares scans to healthy volunteers



Multiple Sclerosis Project

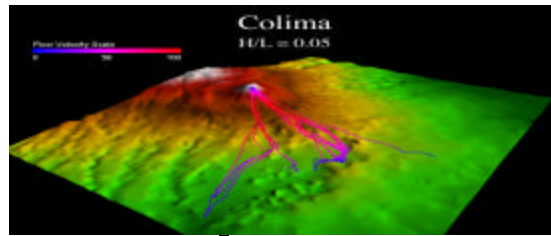
- Compare caudate nuclei between MS patients and healthy controls
- Looking for size as well as structure changes
 - Localized deformities
 - Spacing between halves
- Able to see correlation between disease progression and physical structure changes



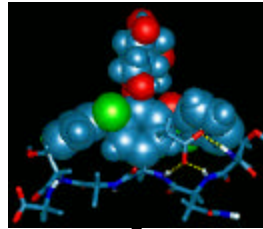
Grid Computing



Grid Computing Overview



Data Acquisition



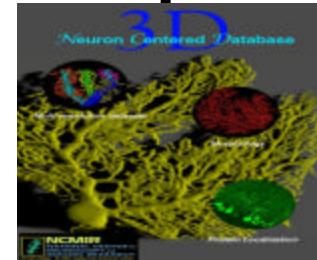
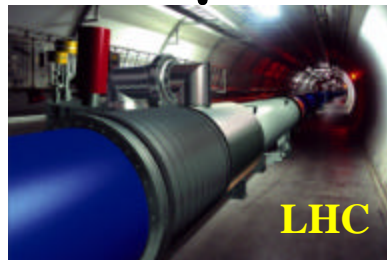
Advanced Visualization



Analysis



Imaging Instruments



Large-Scale Databases

- Coordinate Computing Resources, People, Instruments in Dynamic Geographically-Distributed Multi-Institutional Environment
- Treat Computing Resources like Commodities
 - ❑ Compute cycles, data storage, instruments
 - ❑ Human communication environments
- No Central Control; No Trust

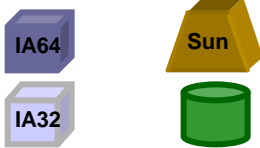
Factors Enabling the Grid

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

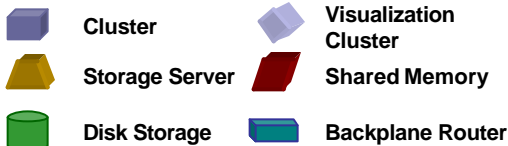
NSF Extensible TeraGrid Facility

Caltech: Data collection analysis

0.4 TF IA-64
IA32 Datawulf
80 TB Storage

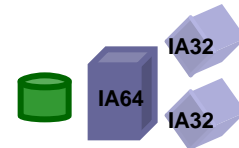


LEGEND

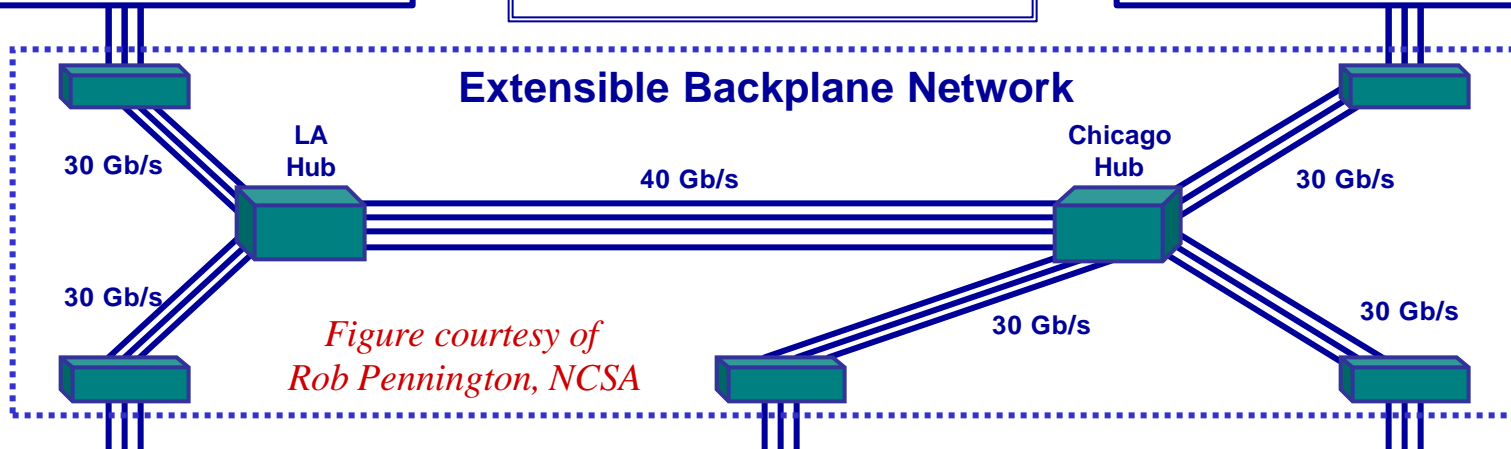


ANL: Visualization

1.25 TF IA-64
96 Viz nodes
20 TB Storage

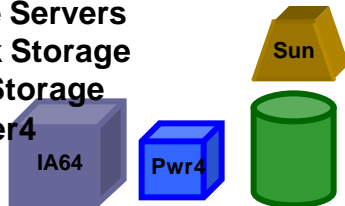


Extensible Backplane Network



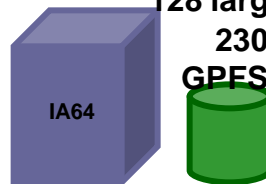
*Figure courtesy of
Rob Pennington, NCSA*

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



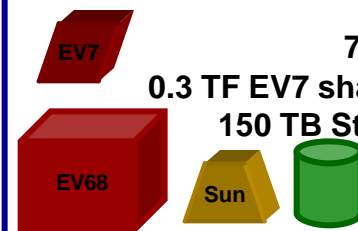
SDSC: Data Intensive

10 TF IA-64
128 large memory nodes
230 TB Disk Storage
GPFS and data mining



NCSA: Compute Intensive

6 TF EV68
71 TB Storage
0.3 TF EV7 shared-memory
150 TB Storage Server



PSC: Compute Intensive

Advanced Computational Data Center

ACDC: 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

CCR

19 IRIX, RedHat, &
WINNT Processors

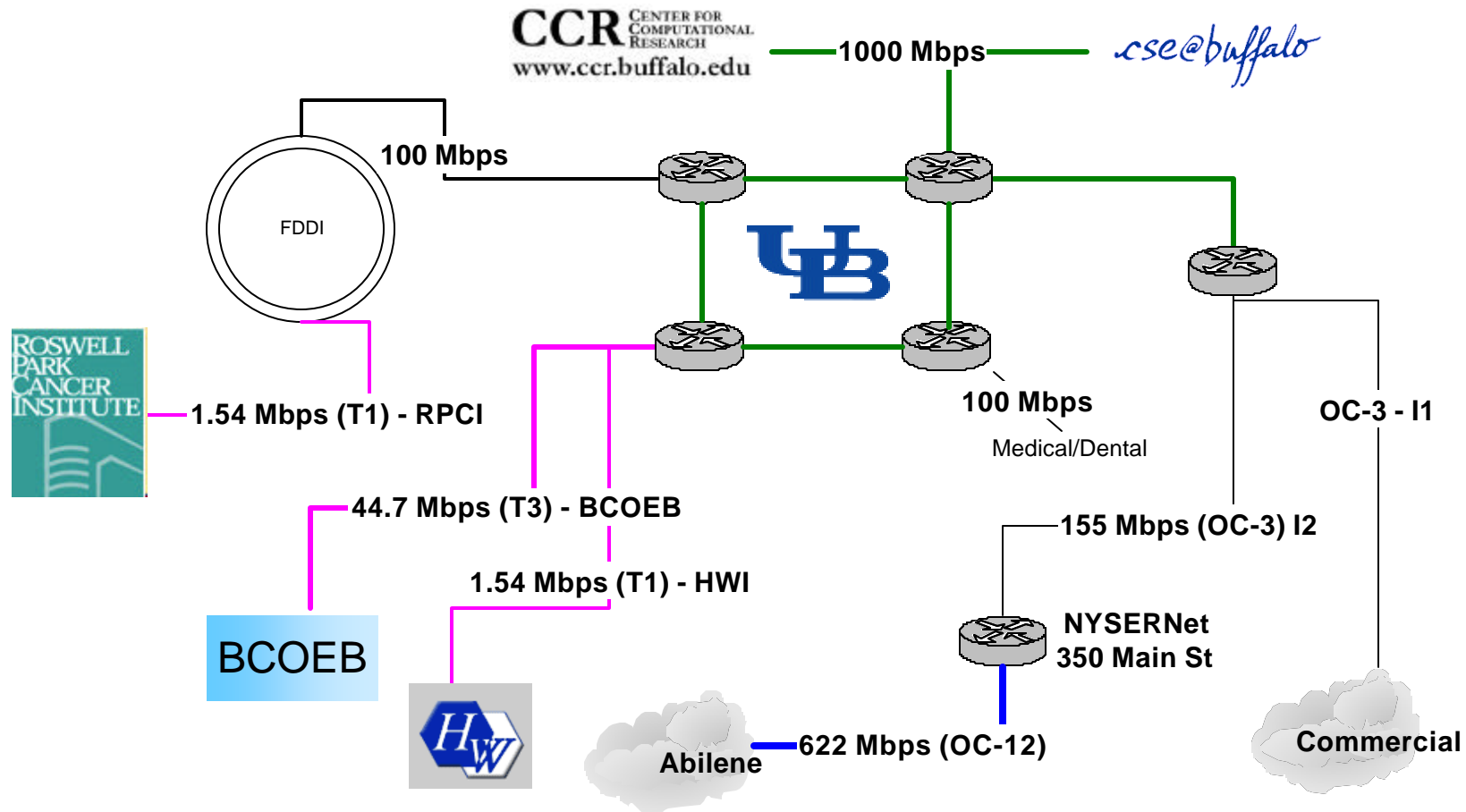
Computer Science & Engineering
25 Single Processor Sun Ultra5s

School of Dental Medicine
9 Single Processor Dell P4 Desktops

Hauptman-Woodward Institute
13 Various SGI IRIX Processors

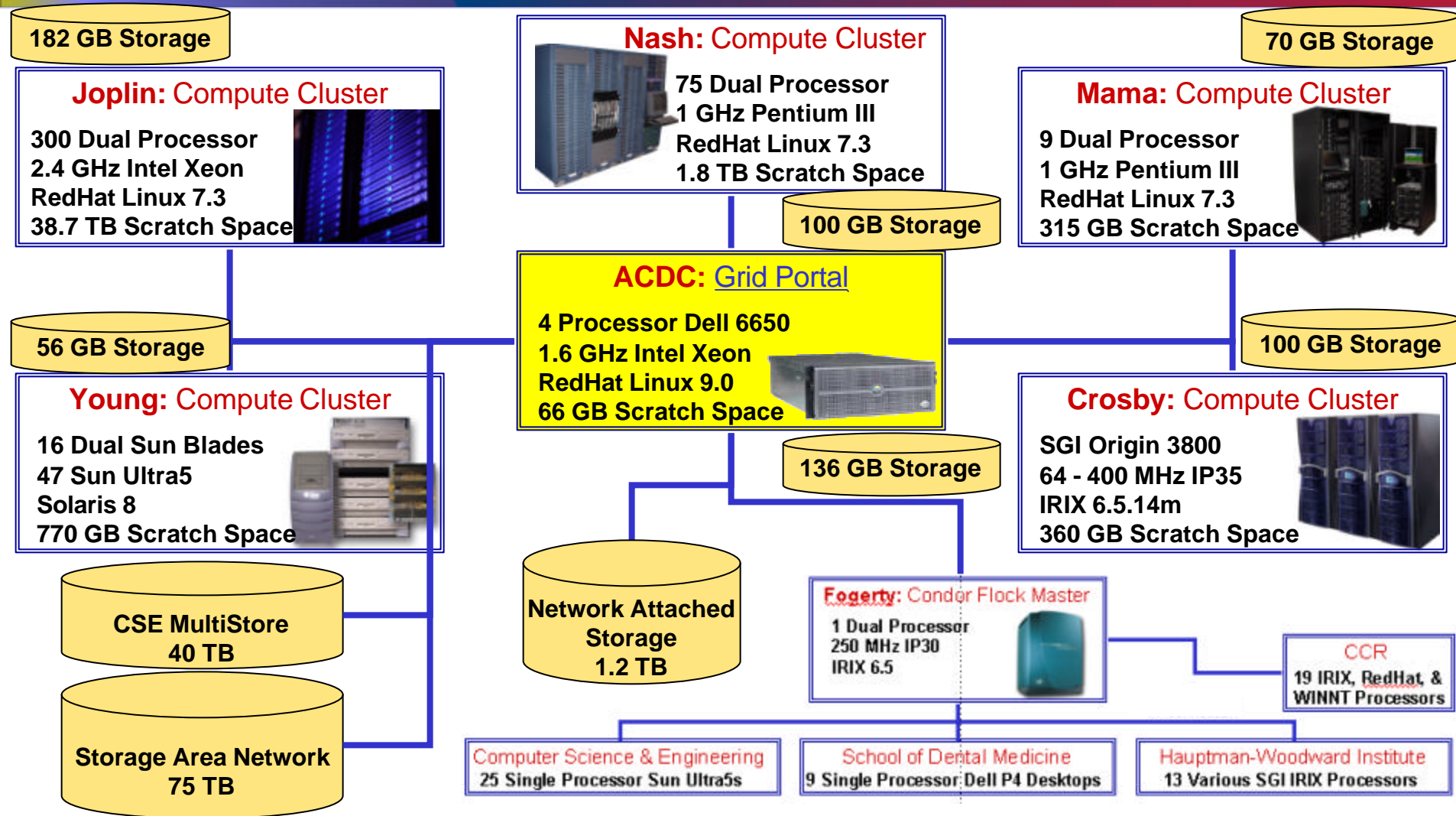


Network Connections



ACDC Data Grid Overview

(Grid-Available Data Repositories)



ACDC-Grid

CCR Grid Computing Services - Microsoft Internet Explorer

CCR University at Buffalo The State University of New York

Center for Computational Research GRID PORTAL

High Performance Grid Computing

Welcome to Grid Computing Services

University at Buffalo Center for Computational Research is currently forming the first Western New York computational grid. The computational grid consist of many supercomputers located at the Center and several other networked supercomputers throughout the Western New York region. These resources will be shared by many researchers from several departments working on a diverse suite of problems including Bioinformatics, Computational Chemistry, and Medical Imaging to name a few.

We also provide grid computing support for the University's Center for Computational Research learning, teaching and research activities plus the infrastructure for both high performance computing and grid enabled software.

Get your "Grid Computing Guide"?

PORTAL LOGIN

- Grid General Info
- Manage Account
- Grid General Info
- Projects
- Resources
- Computational Grid
- Job Submission
- Job Queue Status
- Data Grid
- Network Status
- Running/Queued Jobs
- PBS Job History
- Grid Portal Statistics
- Grid Portal Statistics
- User Info
- Education/Outreach
- Staff Only
- CCR HOME

CCR Grid Computing Services Data Management - Microsoft Internet Explorer

CCR University at Buffalo The State University of New York

Center for Computational Research GRID PORTAL

High Performance Grid Computing

PORTAL LOGOUT

VIEW Group GROUP miller UserList rappleys

reppleye

- KeyMaster
- Morpheus
 - Tank
 - Agent
 - Rabbit
 - Tank
 - Morpheus
 - Oracle.m
 - Nao

Browser view of "miller" group files published by user

CCR Grid Computing Services Grid Admin - Microsoft Internet Explorer

CCR University at Buffalo The State University of New York

Center for Computational Research GRID PORTAL

High Performance Grid Computing

PORTAL LOGOUT

View statistics for: disk_space

Data based on: group

from starting date: January 1 2000

to ending date: September 13 2003 inclusive

for: Grid Portal resources OK

Baagrid Historical Group Disk Space Usage

| Group | Disk Space Usage (KB) |
|----------|-----------------------|
| miller | ~10,500,000 |
| griddev | ~10,000,000 |
| ccrstaff | ~1,500,000 |
| mlgreen | ~1,500,000 |

CCR Grid Computing Services Grid Admin - Microsoft Internet Explorer

CCR University at Buffalo The State University of New York

Center for Computational Research GRID PORTAL

High Performance Grid Computing

PORTAL LOGOUT

View statistics for: disk_space

Data based on: user

from starting date: January 1 2000

to ending date: September 13 2003 inclusive

for: Grid Portal resources OK

| File num | File ID | Filename | Dir ID | Resource ID | Owner | Groupname | Type |
|----------|---------|---------------|--------|-------------|---------|-----------|------|
| 1 | 56033 | Cypher.txt | 52831 | 10 | mlgreen | griddev | txt |
| 2 | 56034 | Cypher.sh | 52858 | 10 | mlgreen | griddev | sh |
| 3 | 56035 | Oracle.asc | 52958 | 10 | mlgreen | griddev | asc |
| 4 | 56036 | Cypher.sh | 52634 | 10 | mlgreen | miller | sh |
| 5 | 56037 | Rabbit.dat | 52830 | 10 | mlgreen | ccrstaff | dat |
| 6 | 56038 | Agent.exe | 53064 | 10 | mlgreen | griddev | exe |
| 7 | 56039 | Dozer.sh | 52852 | 10 | mlgreen | griddev | sh |
| 8 | 56040 | Nao.asc | 52187 | 10 | mlgreen | mlgreen | asc |
| 9 | 56041 | Agent.mpg | 52833 | 10 | mlgreen | mlgreen | mpg |
| 10 | 56042 | Tank.txt | 52188 | 10 | mlgreen | mlgreen | txt |
| 11 | 56043 | Smith.xls | 52258 | 10 | mlgreen | ccrstaff | xls |
| 12 | 56044 | KeyMaster.csh | 52186 | 10 | mlgreen | miller | csh |
| 13 | 56045 | Oracle.csh | 52632 | 10 | mlgreen | griddev | csh |
| 14 | 56046 | Dozer.xls | 52808 | 10 | mlgreen | mlgreen | xls |
| 15 | 56047 | Cypher.exe | 52204 | 10 | mlgreen | griddev | exe |
| 16 | 56048 | Rabbit.ppt | 52861 | 10 | mlgreen | miller | ppt |
| 17 | 56049 | Nao.dat | 52217 | 10 | mlgreen | ccrstaff | dat |
| 18 | 56050 | Cypher.asc | 53086 | 10 | mlgreen | griddev | asc |

ACDC-Grid Administration

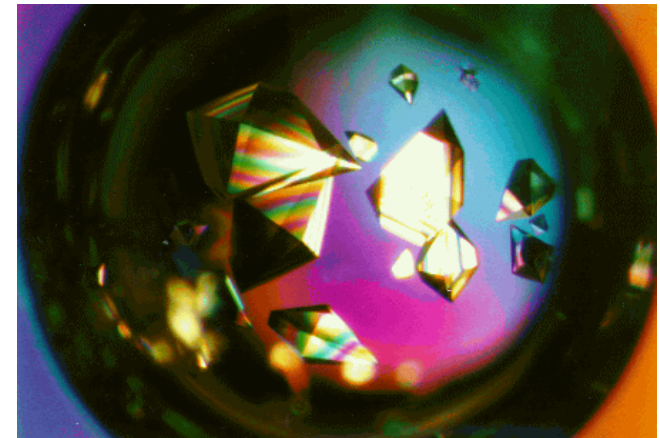
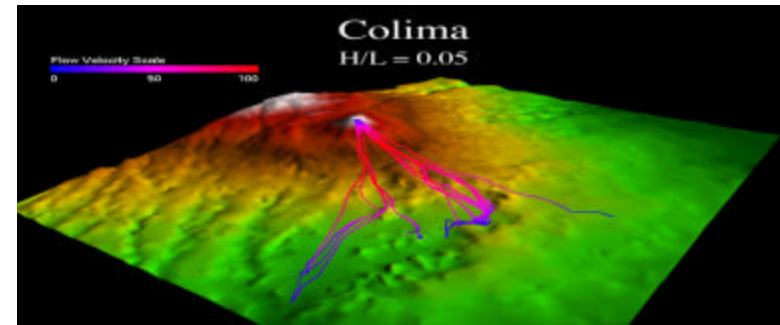
The collage displays four different pages from the CCR Grid Portal, all accessed via Microsoft Internet Explorer. The portal's header consistently shows the University at Buffalo logo and the text 'Center for Computational Research GRID PORTAL High Performance Grid Computing'.

- Top Left:** The 'Grid Site Administration' page. It features a 'PORTAL LOGOUT' sidebar with links like 'Manage Account', 'Grid General Info', and 'Project Resources'. The main content area lists 'Users', 'Groups', 'Portal Event Log', and 'Database Job List'. Below these are sections for 'Organizations (add, edit, delete)', 'Resources (view, refresh, ping, delete, create host certificate)', and 'Globus Administration Reports (machine usage, user access to machines, etc.)'.
- Top Right:** The 'Generate Globus grid-mapfile' page. It includes a 'PORTAL LOGOUT' sidebar and a main section with instructions on specifying an optional include file and a grid-mapfile path. There are input fields for these values and a checkbox for 'Do not stage the file to the grid nodes'.
- Bottom Left:** The 'Create New Database Job' page. It has a 'PORTAL LOGOUT' sidebar and a main section with a form to create a new database job. The form includes fields for 'Job Name', 'Full Path To Script', 'Accepts Arguments' (a dropdown menu), 'Run Script' (a dropdown menu), and 'Run As User' (a dropdown menu). There are 'Create Job' and 'Cancel' buttons at the bottom.
- Bottom Right:** The 'MDS Resource Update Status' page. It features a 'PORTAL LOGOUT' sidebar and a main section with a table showing the current time and a list of resources with their last update times and next update times. Below the table are links to return to the 'Grid Resource Admin menu' and the 'Grid Admin menu'.



Grid-Enabling Application Templates

- Structural Biology
- Earthquake Engineering
- Pollution Abatement
- Geographic Information Systems & BioHazards



ACDC-Grid

Cyber-Infrastructure

■ Predictive Scheduler

- ❑ Define quality of service estimates of job completion, by better estimating job runtimes by profiling users.

■ Data Grid

- ❑ Automated Data File Migration based on profiling users.

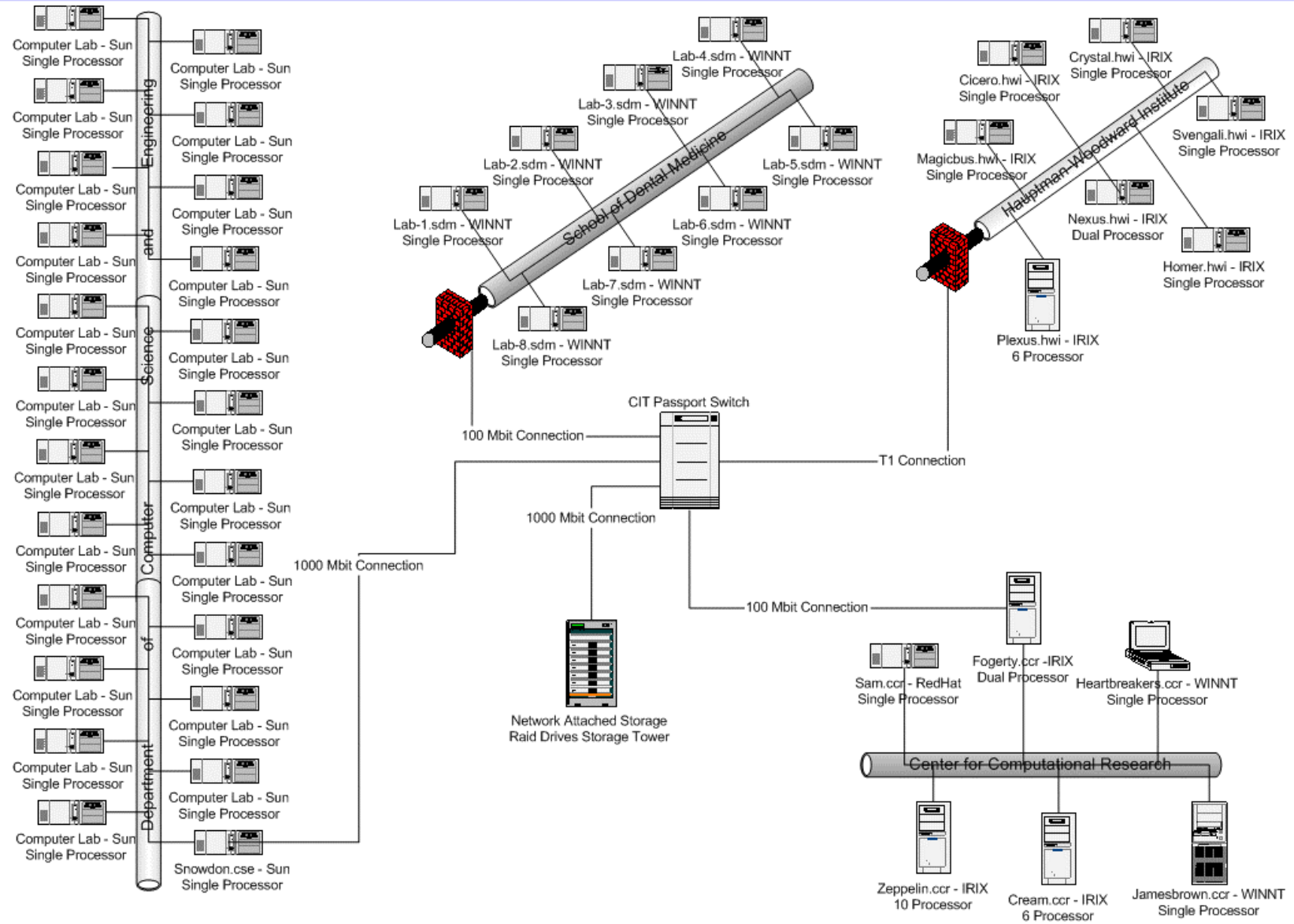
■ High-performance Grid-enabled Data Repositories

- ❑ Develop automated procedures for dynamic data repository creation and deletion.

■ Dynamic Resource Allocation

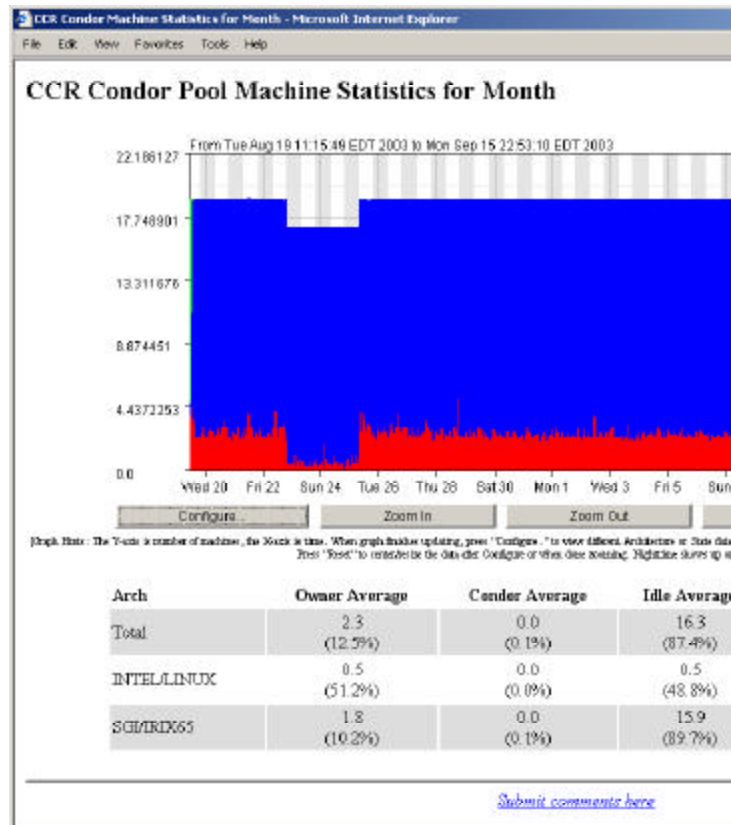
- ❑ Develop automated procedures for dynamic computational resource allocation.

Initial ACDC Campus Grid



ACDC-Grid Portal Condor Flock

■ CondorView integrated into ACDC-Grid Portal



CCR Grid Computing Services: Grid Admin - Microsoft Internet Explorer

CCR Center for Computational Research GRID PORTAL

High Performance Grid Computing

PORTAL LOGOUT

User Tools

- » Manage Account
- Grid General Info
- Projects
- Resources
 - » Computational Grid
 - » Job Submission
 - » Job/Queue Status
 - » Data Grid
 - » Data Grid Statistics
 - » Network Status
 - » Running/Queued Jobs
 - » PBS Job History
 - » Grid Portal Statistics
 - » Condor Flock Statistics
 - » User Information
- Education/Outreach
- Staff Only
- CCR HOME

Condor
High Throughput Computing

Condor Pool Statistics for CCR

Pool Resource (Machine) Statistics

- For the past hour
- For the past day
- For the past week
- For the past month
- For the month of [Jan] [Feb] [Mar] [Apr] [May] [Jun] [Jul] [Aug] [Sep] [Oct] [Nov] [Dec]

Pool User (Job) Statistics

- For the past hour
- For the past day
- For the past week
- For the past month
- For the month of [Jan] [Feb] [Mar] [Apr] [May] [Jun] [Jul] [Aug] [Sep] [Oct] [Nov] [Dec]

[Submit comments here](#)

Advanced
Center for Computational Research
Data
Center

GRID

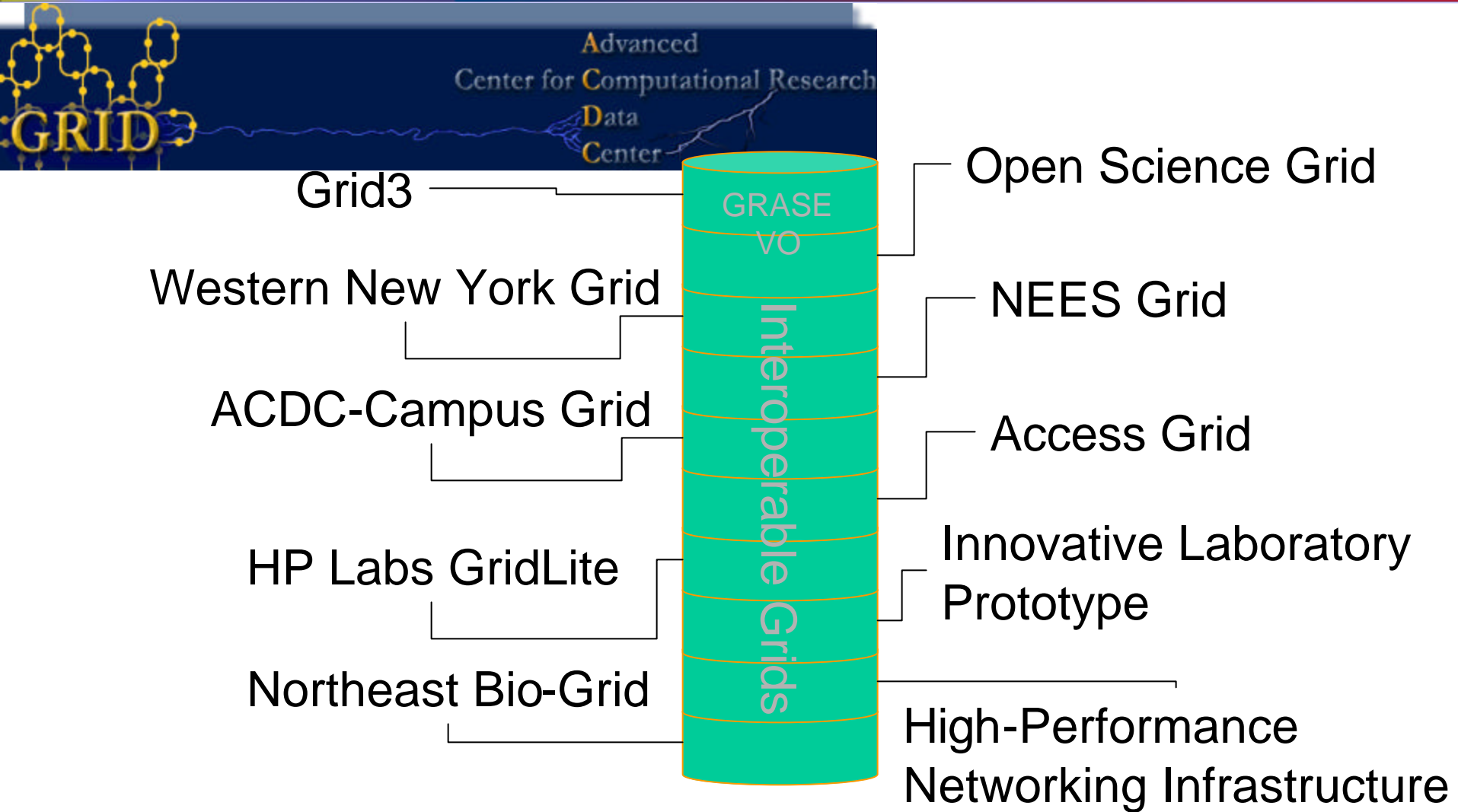


ACDC-Grid Collaborations

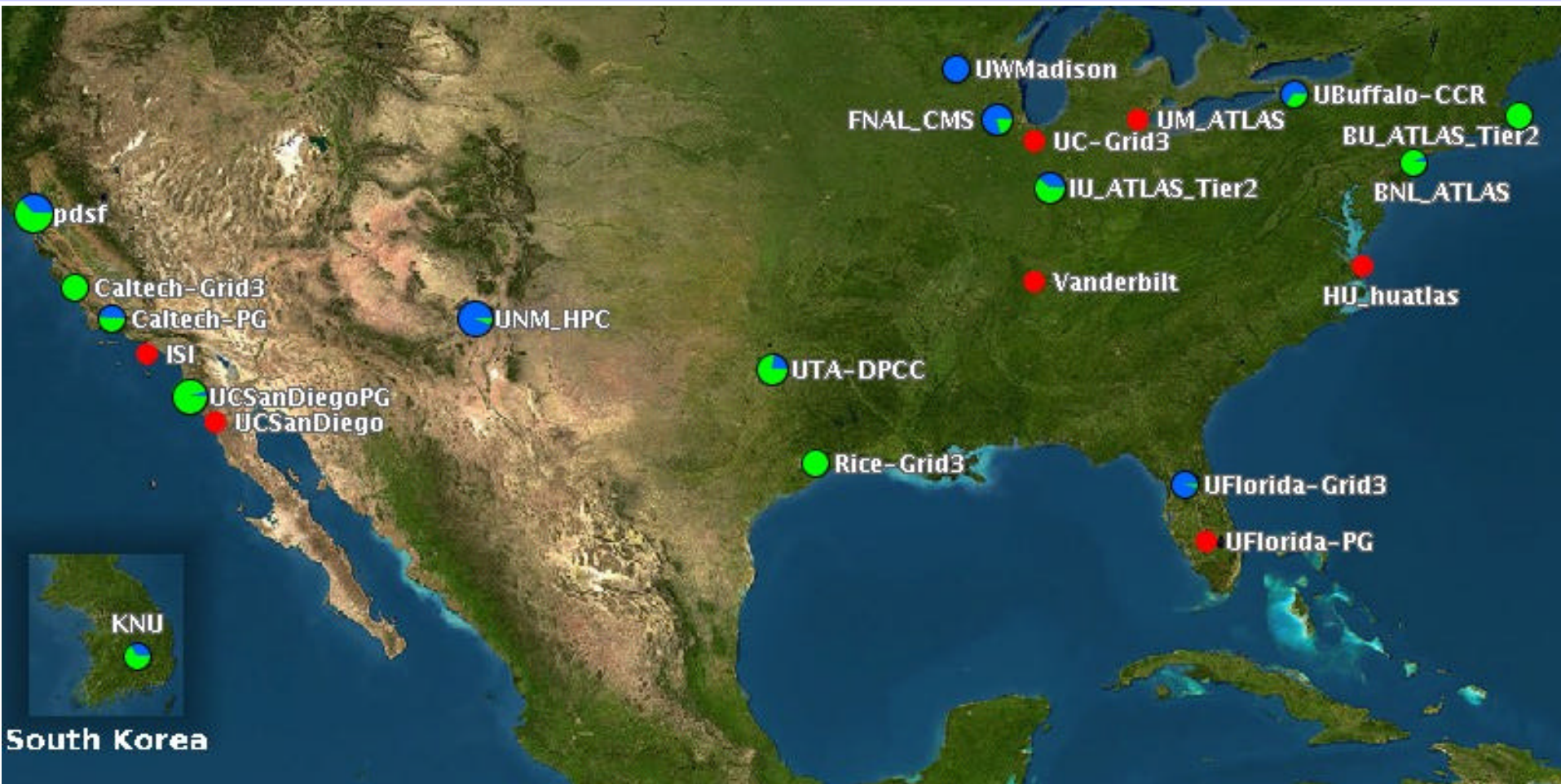
- Grid3+ Collaboration / iVDGL Member
- Open Science Grid Founding Participant
 - Monitoring & Information Services, co-chair
 - Security, Tech Working Group Participant
- WNY Grid Initiative
- Grid-Based Visualization
 - SGI Collaboration
- Grid-Lite
 - HP Labs Collaboration
- Innovative Laboratory Prototype
 - Dell Collaboration
- NE Bio-Grid
 - IBM Research Collaboration
 - MIT, Harvard



ACDC-Grid Collaborations



Grid3 Snapshot of Sites



UBuffalo-CCR Virtual Organization

Grid Resources for Advanced Science and Engineering (GRASE)



University at Buffalo

The State University of New York

Center for Computational Research

CCR

Northeast Structural Genomics Consortium

■ Consortium

- ❑ UB, Rutgers, Columbia, Cornell, PNNL, Yale, UToronto, Robert Wood Johnson Medical Center, Hauptman-Woodward Medical Research Center

■ Mission

- ❑ Develop integrated technologies for high-throughput (htp) protein production and 3D structure determination
- ❑ The goal is to determine 500 new protein structures over 5 years
- ❑ Combination of strong parallel efforts in both X-ray crystallography and solution-state NMR spectroscopy
- ❑ UB Professor Thomas Szyperski awarded Scientific American's Top 50 Scientists in 2003 for novel work in high-throughput structure determination with NMR

Western New York Health Information Project

Goals:

- Build a secure community-wide healthcare database
- Develop an electronic patient medical record that “follows the patient”
- Provide care providers with real-time patient information wherever they are
- Provide a tool to aid agencies in community safety, epidemiology, resource allocation, and bioterrorism response
- Improve the overall quality of healthcare while reducing costs

Selected Participants:

- University at Buffalo (CCR, School of Informatics, School of Medicine, Health Science Library)
- Buffalo Academy of Medicine
- Erie County DoH
- New York State DoH
- WNY HealtheNet
- Involvement from Kaleida Health, ECMC, Catholic Health System, Independent Health, HealthNow, and Univera Healthcare



Outreach

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



Outreach

■ Pilot HS Program in Computational Science

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



THE BUFFALO NEWS

EDUCATION

RONALD C. COLLEMA/BUFFALO NEWS

University at Buffalo undergraduate David Walsh works with Jacklyn Shure, right, to demonstrate the "Next Generation Scientists" program. At left is Shannon O'Rourke.

An early look at bioinformatics

By EMMA D. SAPING
New Northtown Bureau

For most of Darcy Brown's educational career, science classes have been instructive but somewhat abstract. They've been steeped in theories and ideas that she left behind in the classroom.

But that's not the case anymore for the senior at Mount St. Mary Academy. The world of science has come alive and is practical.

She's in her second year of a University at Buffalo Center for Computational Research bioinformatics program geared to high school students. And when she studies DNA in biology class, she can bring that lesson to life by writing a DNA program.

The innovative and rigorous pilot program, called "Next Generation Scientists: Training for Students and Teachers," merges life sciences and

computational science. It is being taught at Mount St. Mary, Orchard Park High School and City Honors School. About two dozen students are enrolled in the program; they work on smaller versions of the computers used at the research center.

Brown and the three other students in the program demonstrated and spoke about the program Thursday at Mount St. Mary. Awarding was official from UB and Vermont, which funded the program with a \$50,000 grant.

"When you take science in school, it's mostly not practical," Brown said. "Bioinformatics has shown me how to apply science in real life. It has really opened doors for me."

L. Bruce Péterson, associate dean for research and sponsored programs at UB, said the program aims to bring bioinformatics to high schools by developing a curriculum and training

teachers. It will expand into other schools in opening years.

The students met with a couple of selected teachers in their schools who also are receiving training and three UB undergraduate students.

Senior Courtney Kiosowski, who plans to major in mechanical engineering at Clarkson University, said bioinformatics has prepared her for her field of study. She said it's "going to give me a stronger background in engineering."

Because the students use all graduating, Brown said they are trying to recruit students for the program.

"Bioinformatics is really a different experience," she said. "You think of computers and computer programs and the way they are, and now you know the work that goes behind them."

emma.d.saping@buffnews.com

