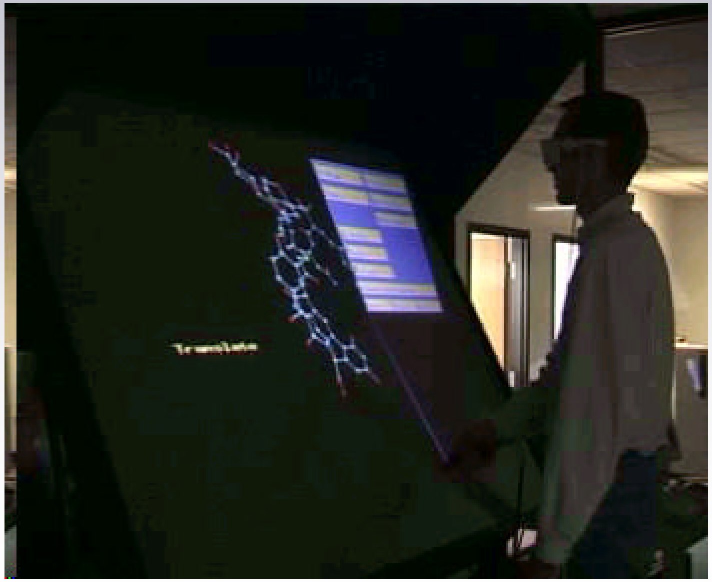
Molecular Structure Determination on a Computational & Data Grid

Mark Green & Russ Miller

Center for Computational Research, SUNY-Buffalo

Computer Science & Engineering, SUNY-Buffalo

Hauptman-Woodward Medical Inst

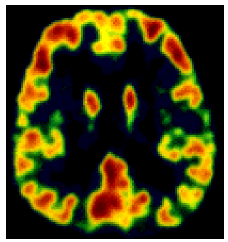


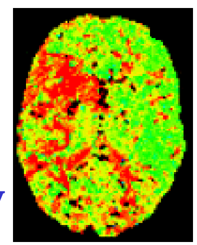
NSF ITR
ACI-02-04918

University at Buffalo
The State University of New York

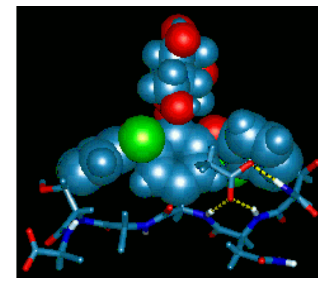
Biomedical Advances

- 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



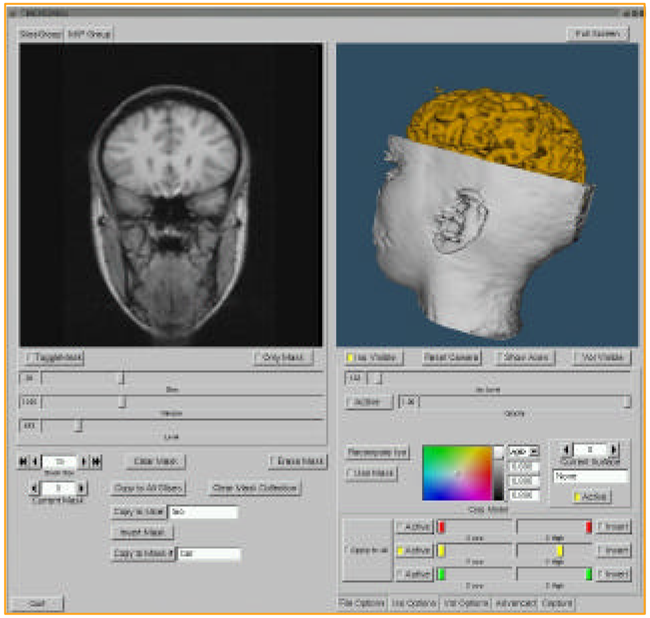


- Direct Methods Structure Determination
  - Listed on “Top Ten Algorithms of the 20th Century”
  - Vancomycin
  - Gramacidin A
- High Throughput Crystallization Method: Patented
- NIH National Genomics Center: Northeast Consortium
- Howard Hughes Medical Institute: Center for Genomics & Proteomics



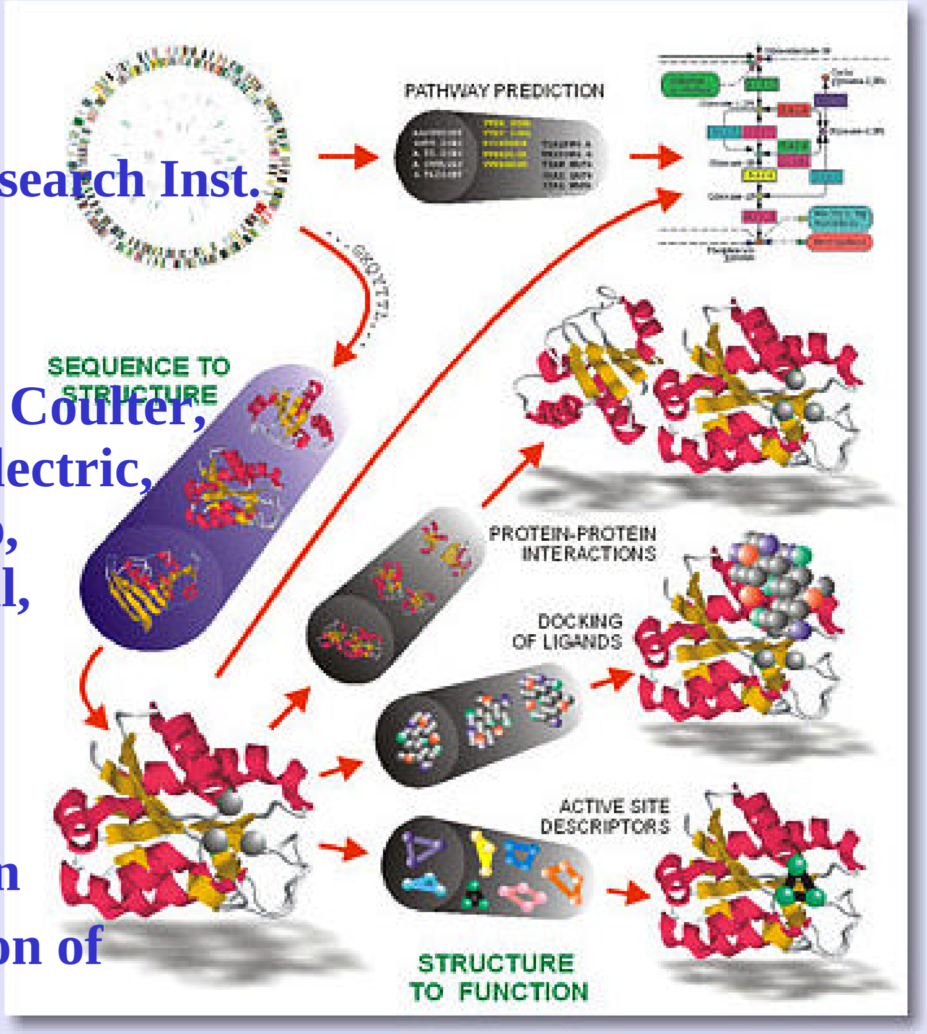
Bioinformatics in Buffalo
A $360M Initiative

- **New York State: $121M**
- **Federal Appropriations: $13M**
- **Corporate: $146**
- **Foundation: $15M**
- **Grants & Contracts: $64M**



Bioinformatics Partners

- **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



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**
  - $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…

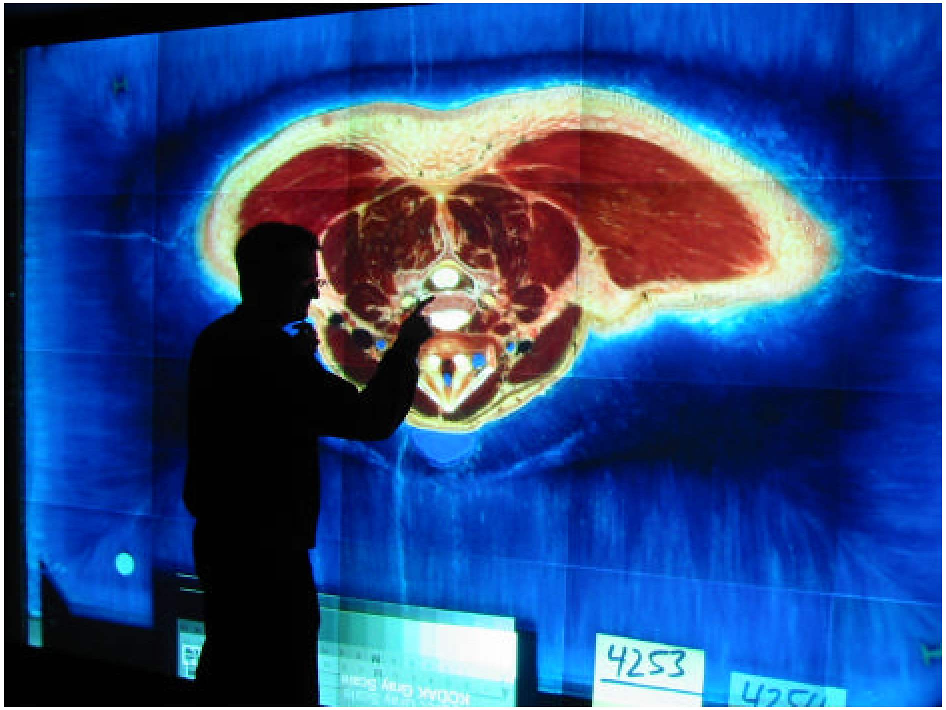
Major CCR Resources (12TF & 80TB)

- 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)

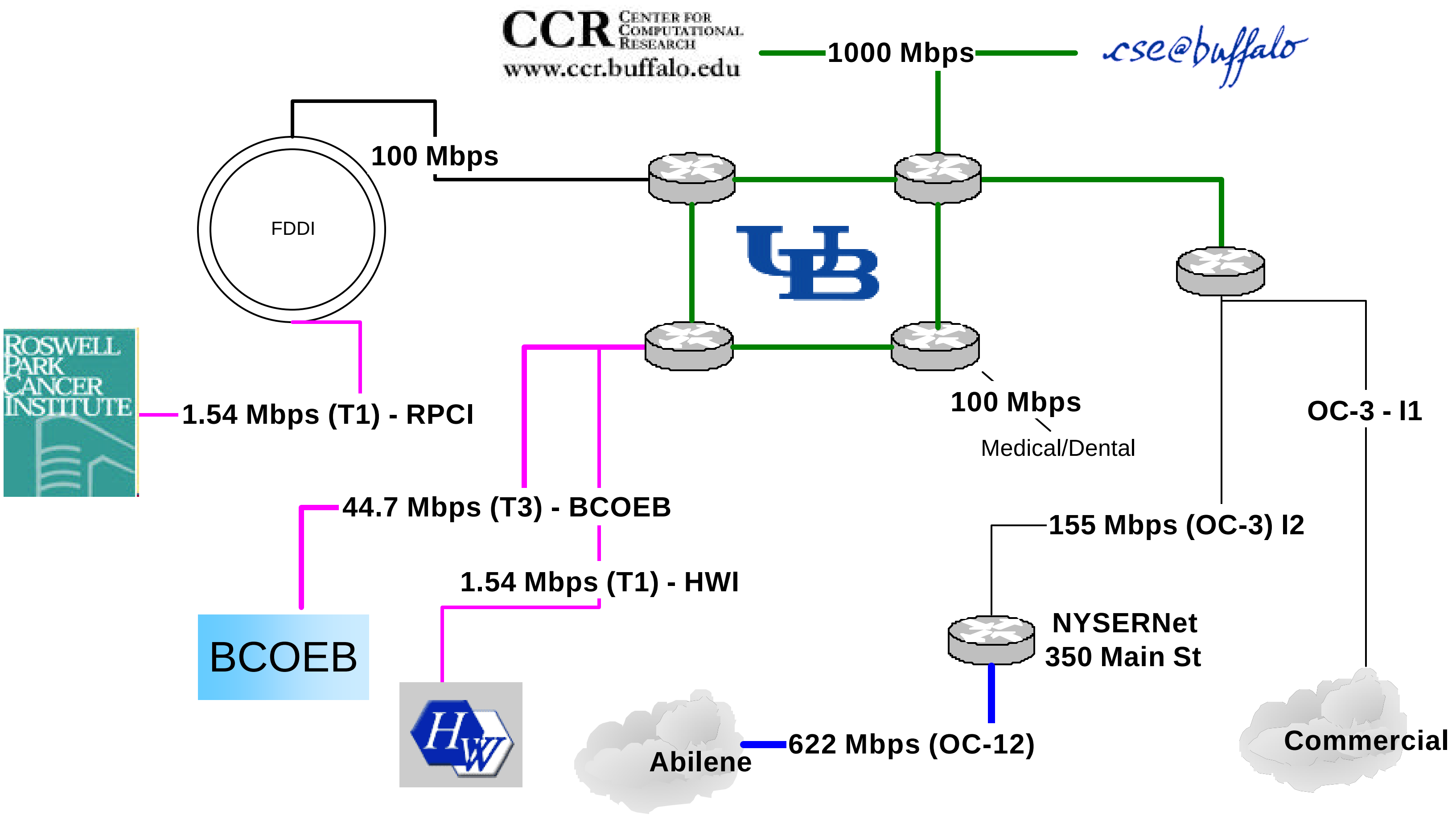
- IBM BladeCenter Cluster
  - 532 P4 Processors (2.8 GHz)
  - 5TB SAN
- Apex Bioinformatics System
  - Sun V880 (3), 6800, 280R (2), PIIIs
  - Sun 3960: 7 TB Disk Storage
- HP/Compaq SAN
  - 75 TB Disk; 190 TB Tape SGI Origin3800
  - 64 Alpha Processors (400 MHz)
  - 32 GB RAM; 400 GB Disk
- IBM RS/6000 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**



Network Connections

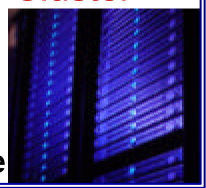


Advanced CCR Data Center (ACDC) Computational Grid Overview

Nash: Compute Cluster

75 Dual Processor
1 GHz Pentium III
RedHat Linux 7.3
1.8 TB Scratch Space

Joplin: Compute Cluster



300 Dual Processor
2.4 GHz Intel Xeon
RedHat Linux 7.3
38.7 TB Scratch Space

Mama: Compute Cluster

9 Dual Processor
1 GHz Pentium III
RedHat Linux 7.3
315 GB Scratch Space

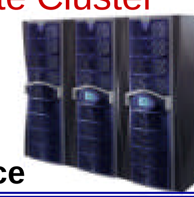
ACDC: Grid Portal

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

Expanding

RedHat, IRIX, Solaris, WINNT, etc

Fogerty: Condor Flock Master

1 Dual Processor
250 MHz IP30
IRIX 6.5

CCR

19 IRIX, RedHat, & WINNT Processors

T1 Connection

Computer Science & Engineering

25 Single Processor Sun Ultra5s

School of Dental Medicine

9 Single Processor Dell P4 Desktops

Hauptman-Woodward Institute

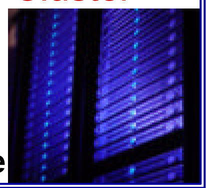
13 Various SGI IRIX Processors

Note: Network connections are 100 Mbps unless otherwise noted.

ACDC Data Grid Overview

182 GB Storage

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

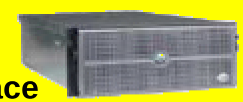
100 GB Storage

70 GB Storage

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

56 GB Storage

Young: Compute Cluster

16 Dual Sun Blades
47 Sun Ultra5
Solaris 8
770 GB Scratch Space

100 GB Storage

Crosby: Compute Cluster

SGI Origin 3800
64 - 400 MHz IP35
IRIX 6.5.14m
360 GB Scratch Space

136 GB Storage

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

CCR
19 IRIX, RedHat, & WINNT Processors

T1 Connection

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

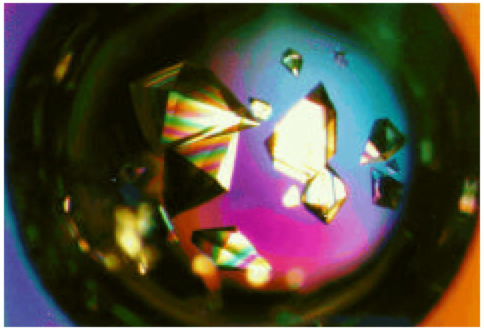
Note: Network connections are 100 Mbps unless otherwise noted.

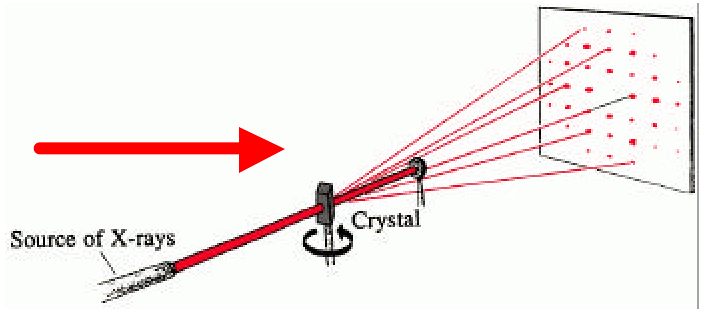
WNY Grid Highlights

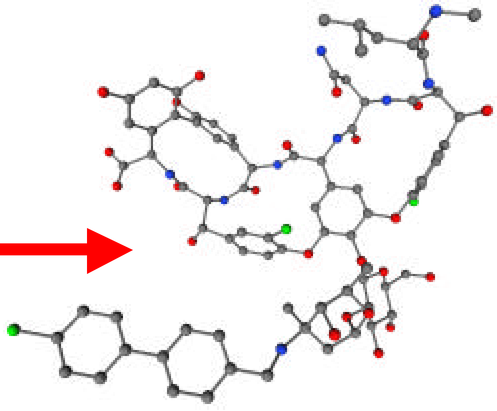
- **Heterogeneous Computational & Data Grid**
- **Currently in Beta with *Shake-and-Bake***
- **WNY Release in 2H04**
- **Bottom-Up General Purpose Implementation**
  - **Ease-of-Use User Tools**
  - **Administrative Tools**
- **Back-End Intelligence**
  - **Backfill Operations**
  - **Prediction and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

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.**

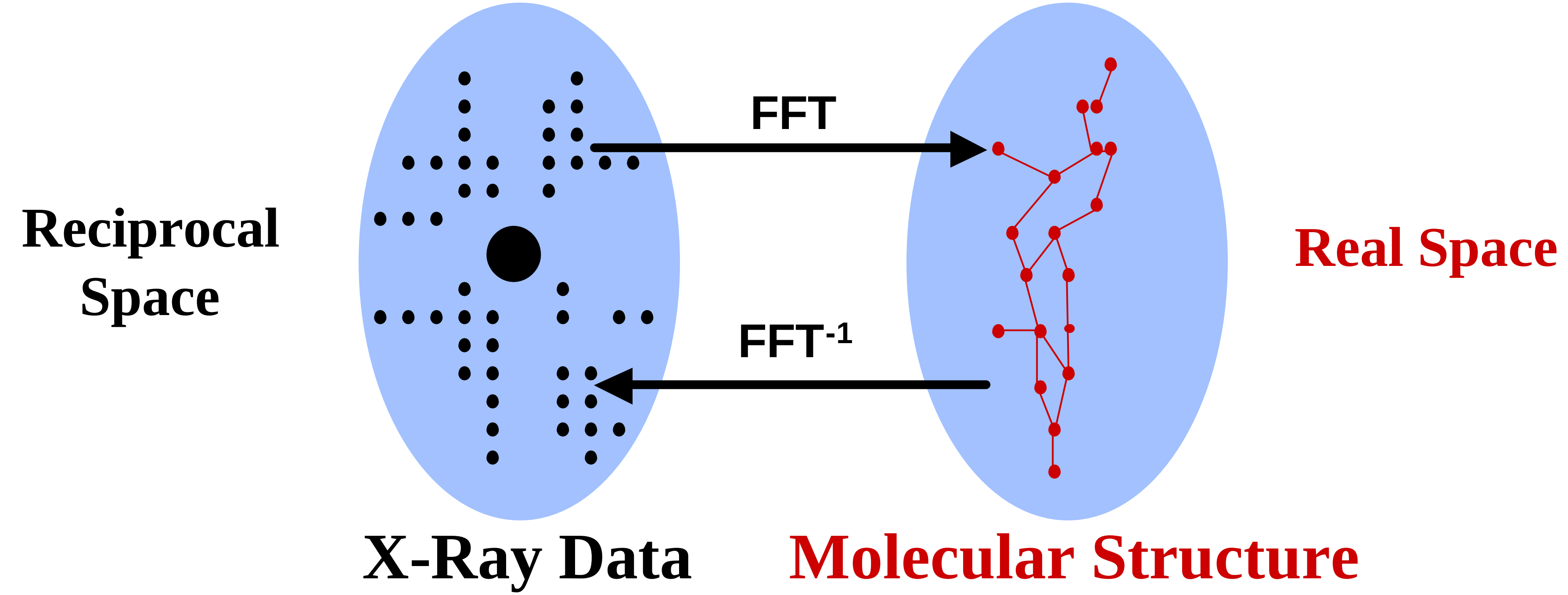






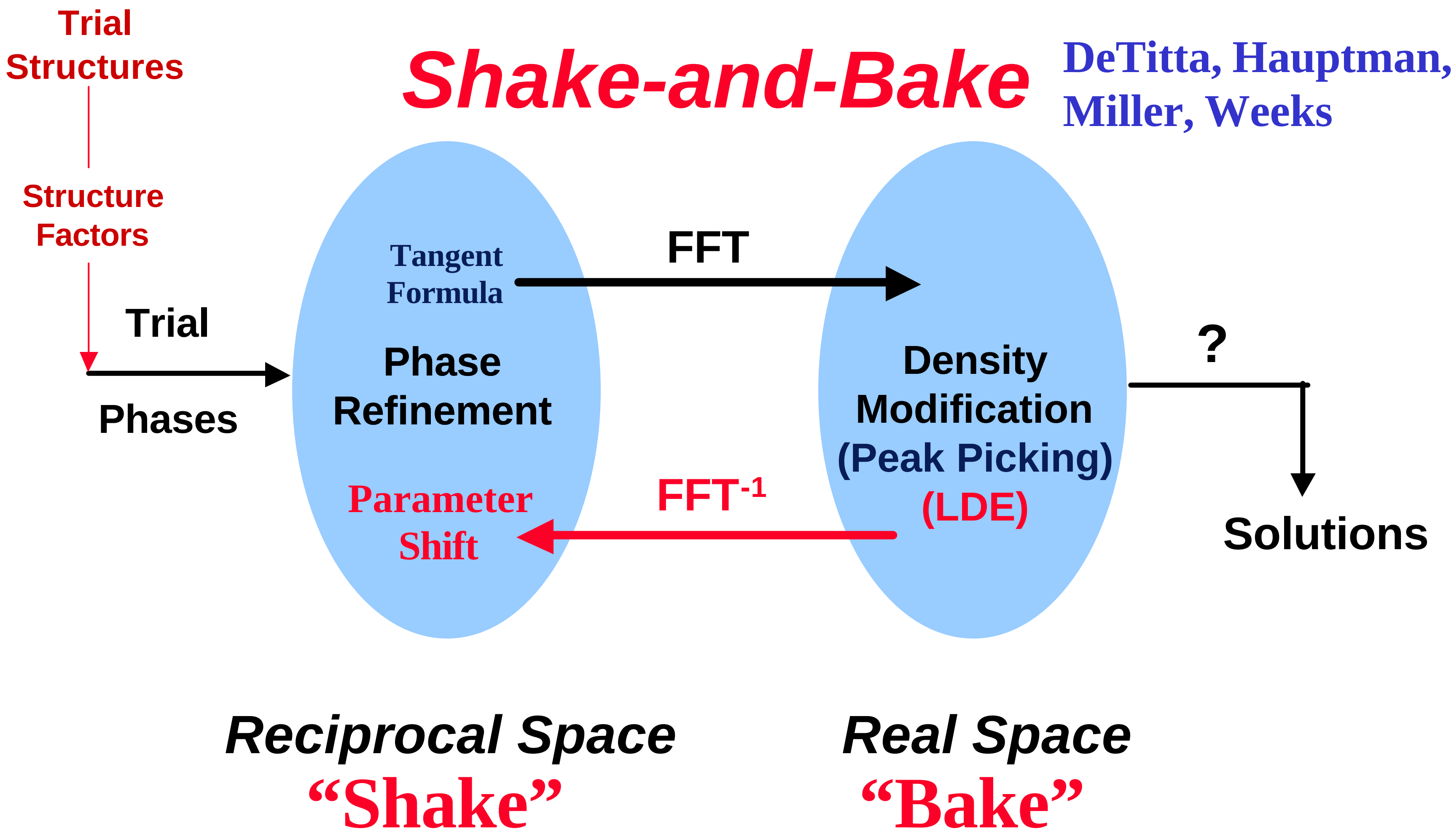
X-Ray Data & Corresponding Molecular Structure

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



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

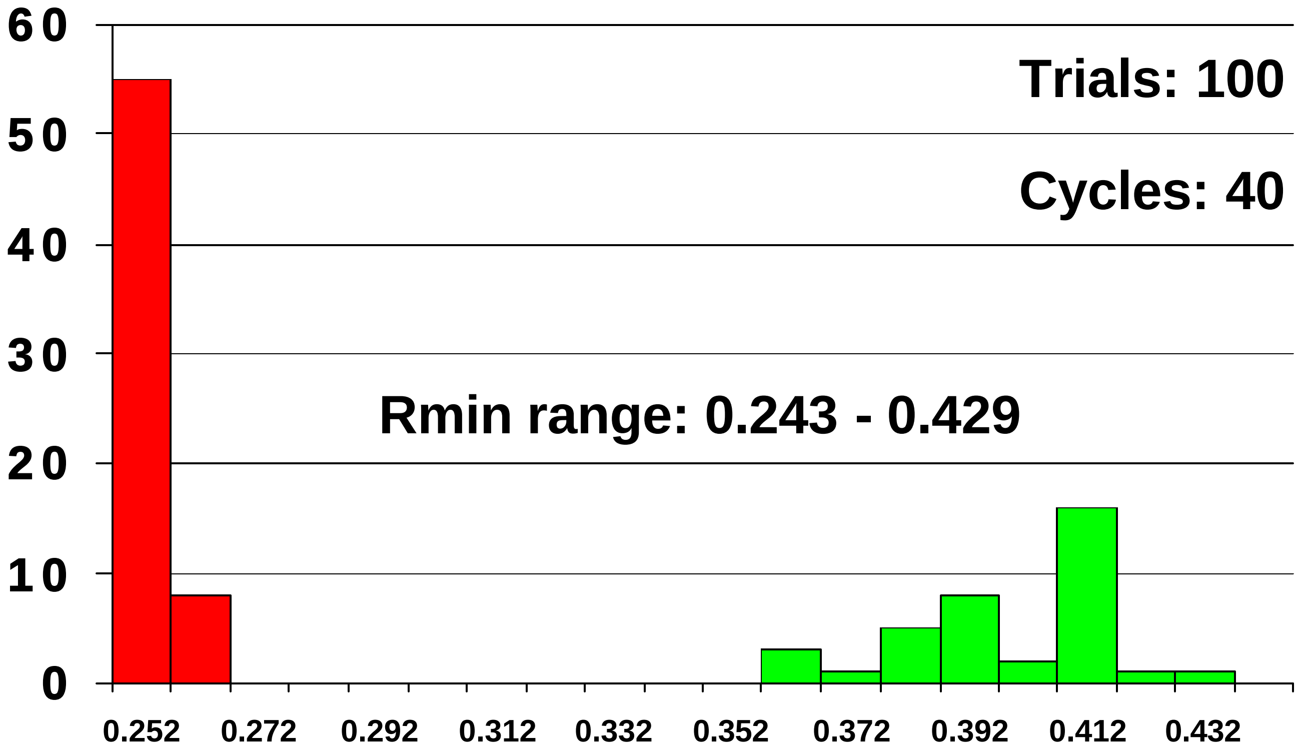
Shake-and-Bake Method: Dual-Space Refinement



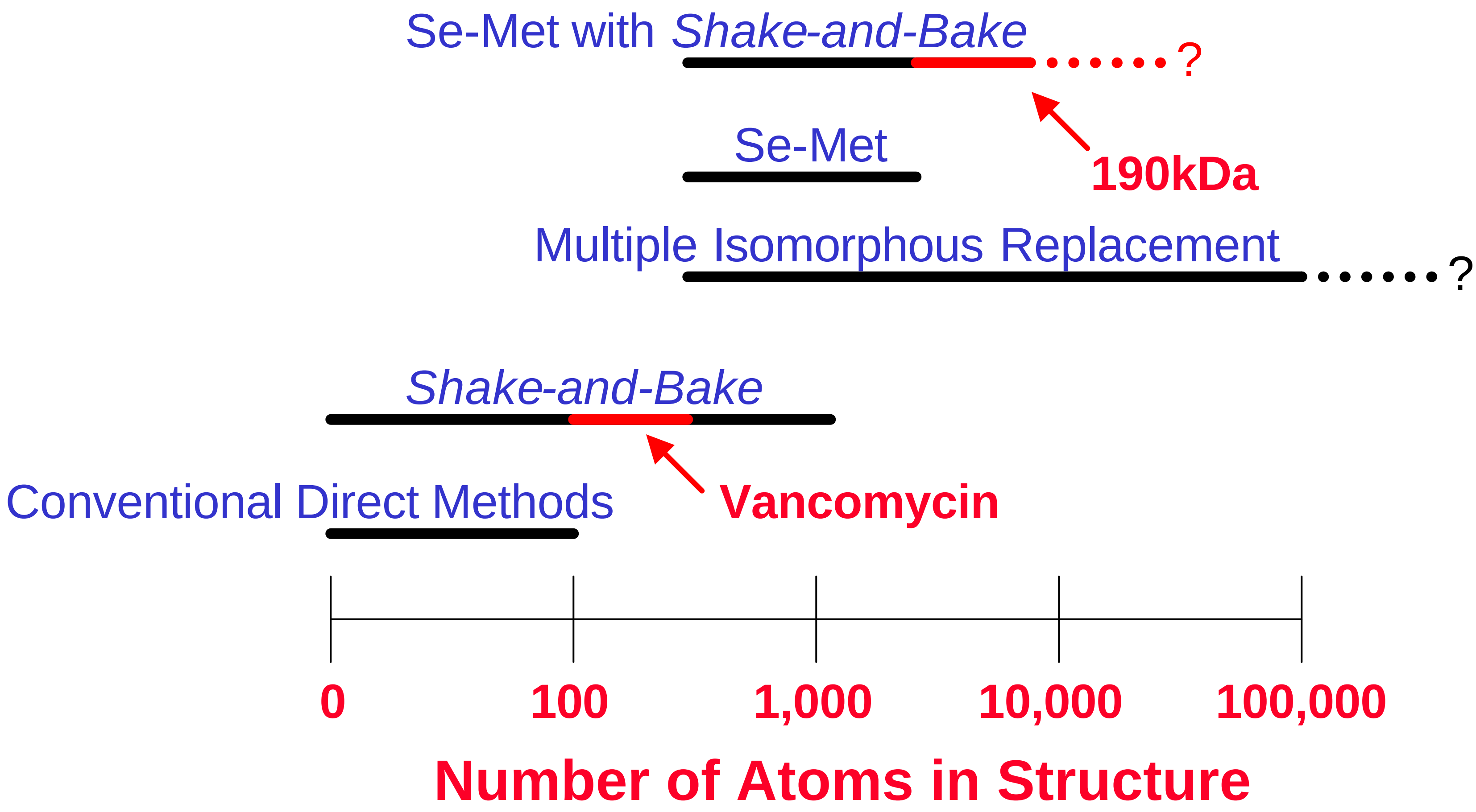
Ph8755: *SnB* Histogram

Atoms: 74
Space Group: P1

Phases: 740
Triples: 7,400



Phasing and Structure Size

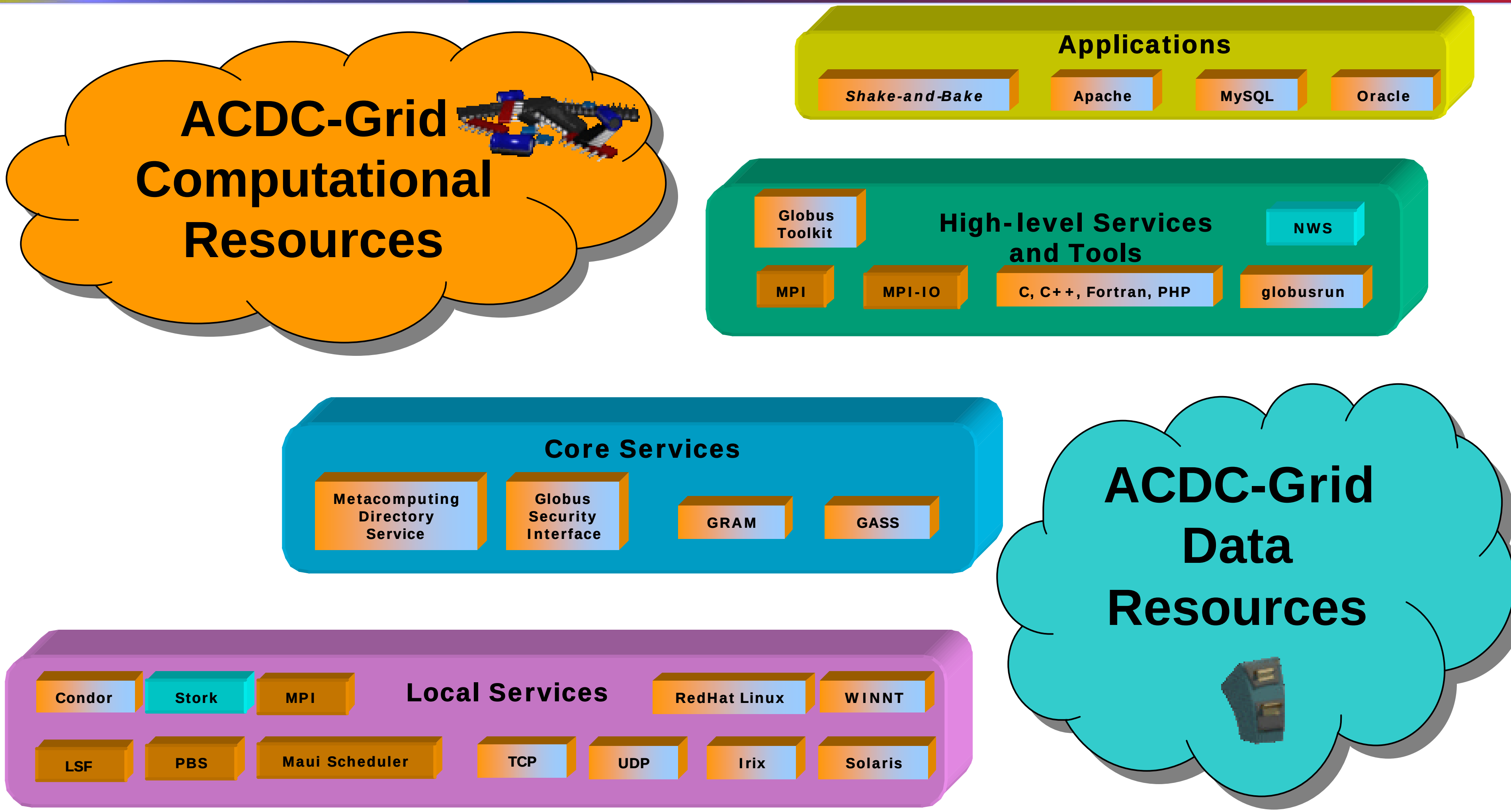


Grid-Based *SnB* Objectives

- 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

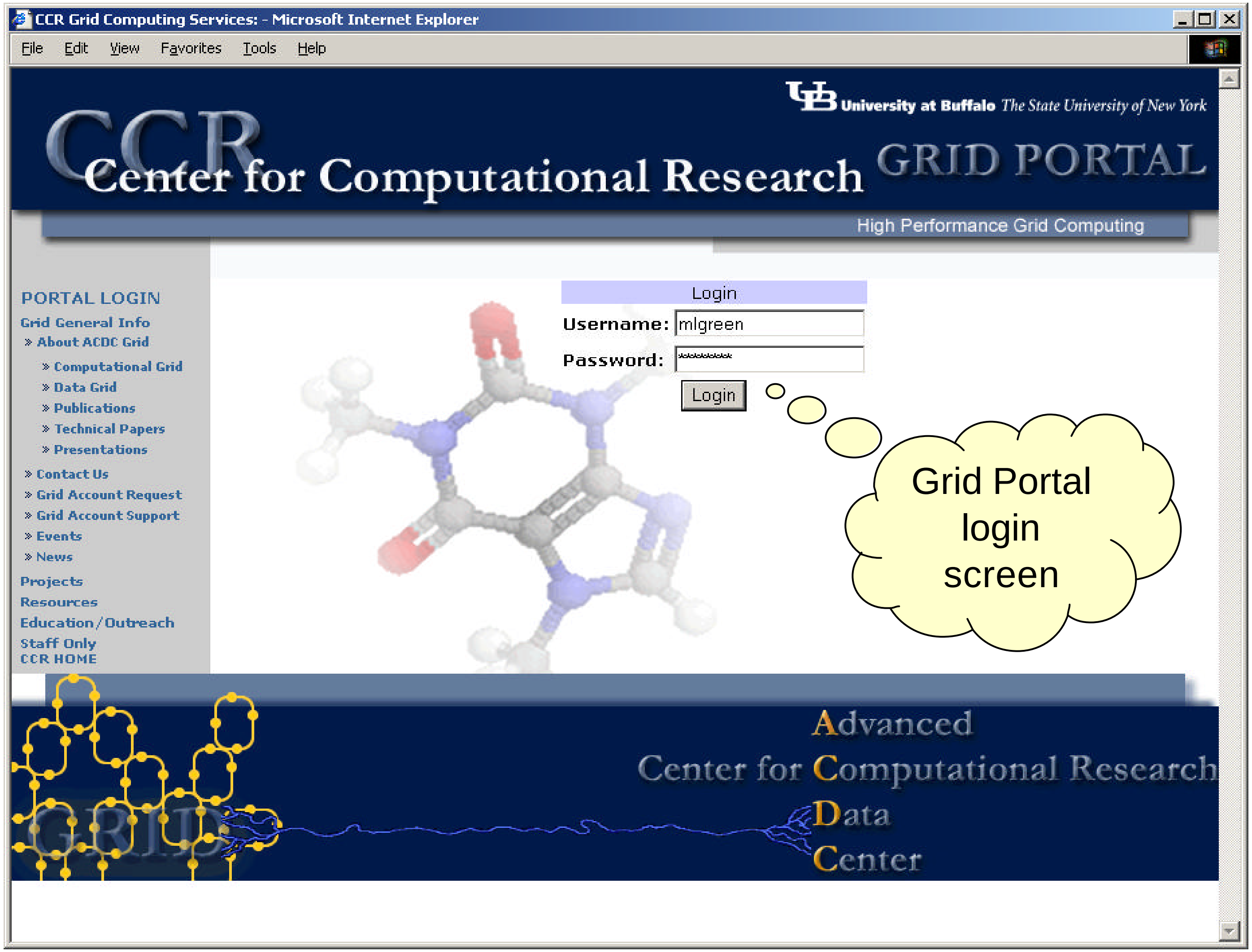
- Serve as Template for Chem-Grid & Bio-Grid
- Experience with Globus and Related Tools

Grid Services and Applications

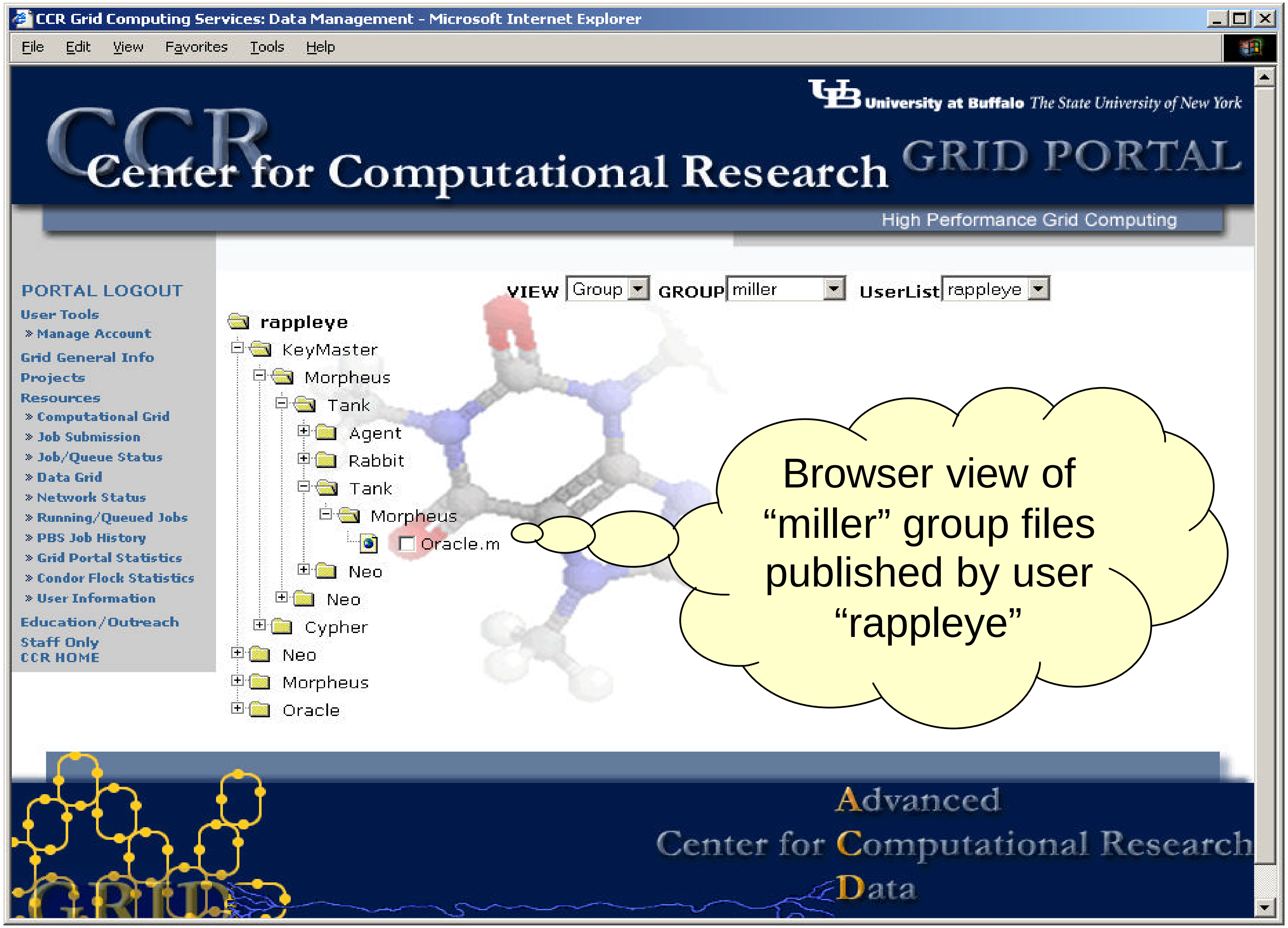


Adapted from Ian Foster and Carl Kesselman

ACDC-Grid Portal Login

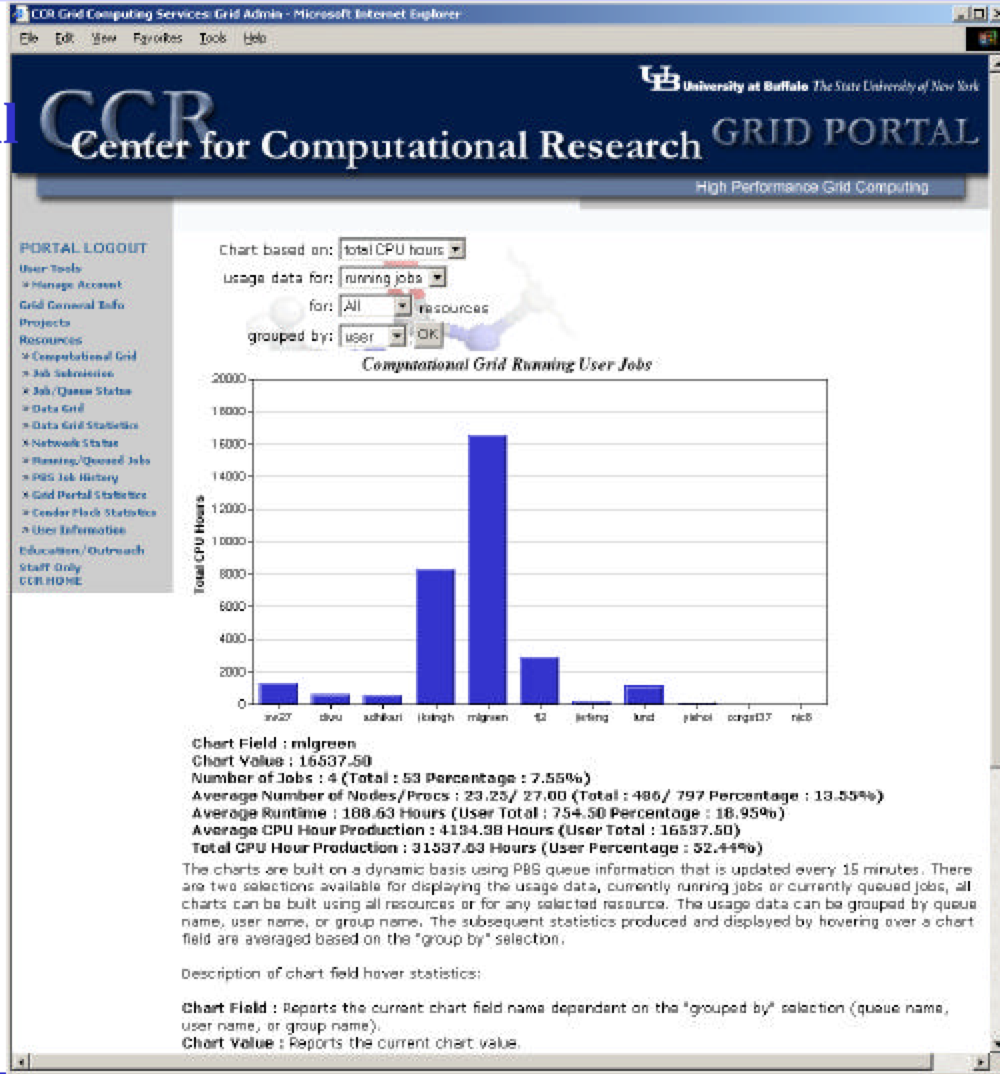


Data Grid Capabilities

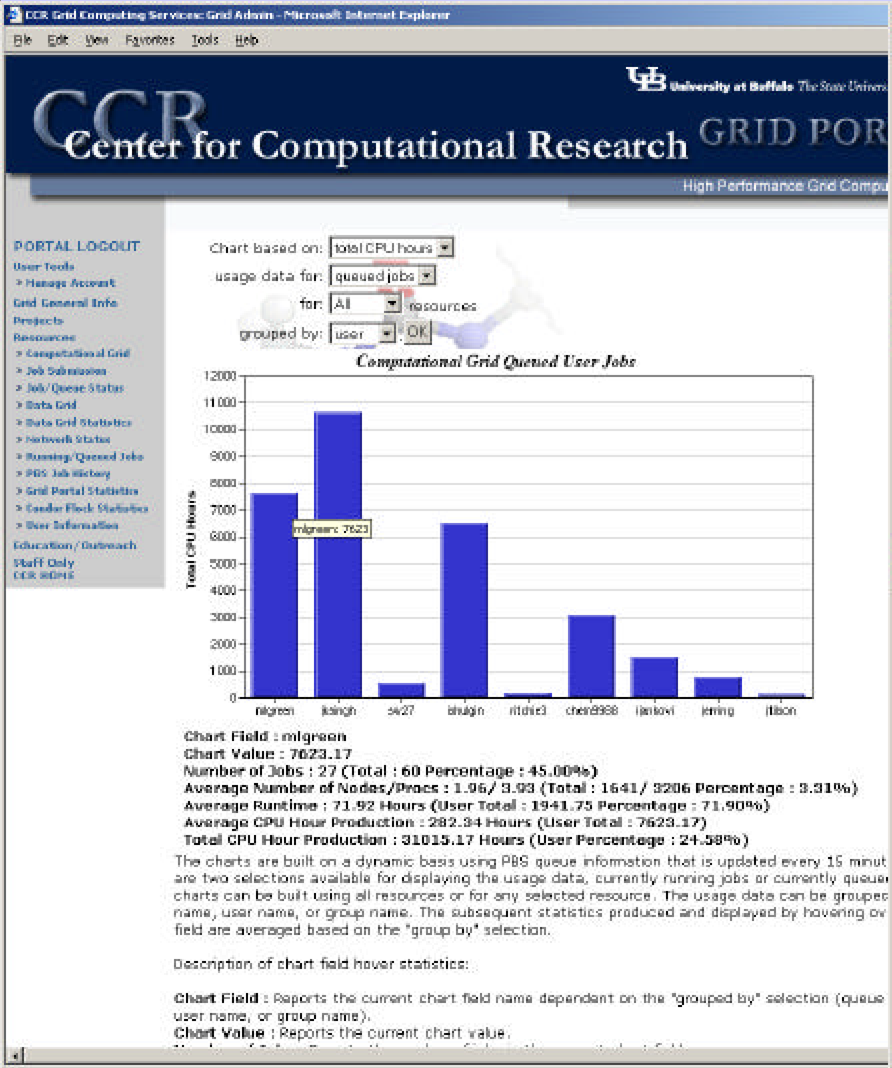


Grid Portal Job Status

- 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.



Grid Portal Job Status



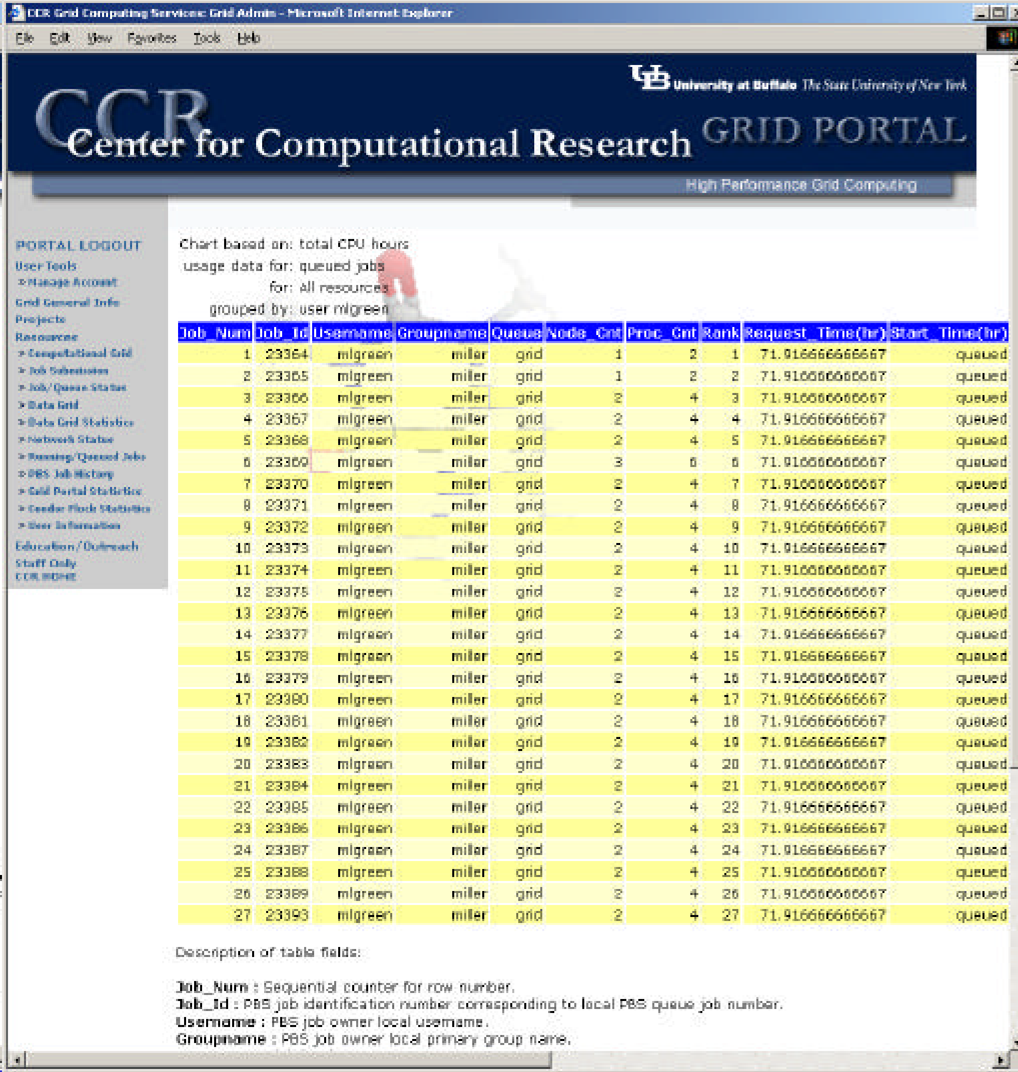


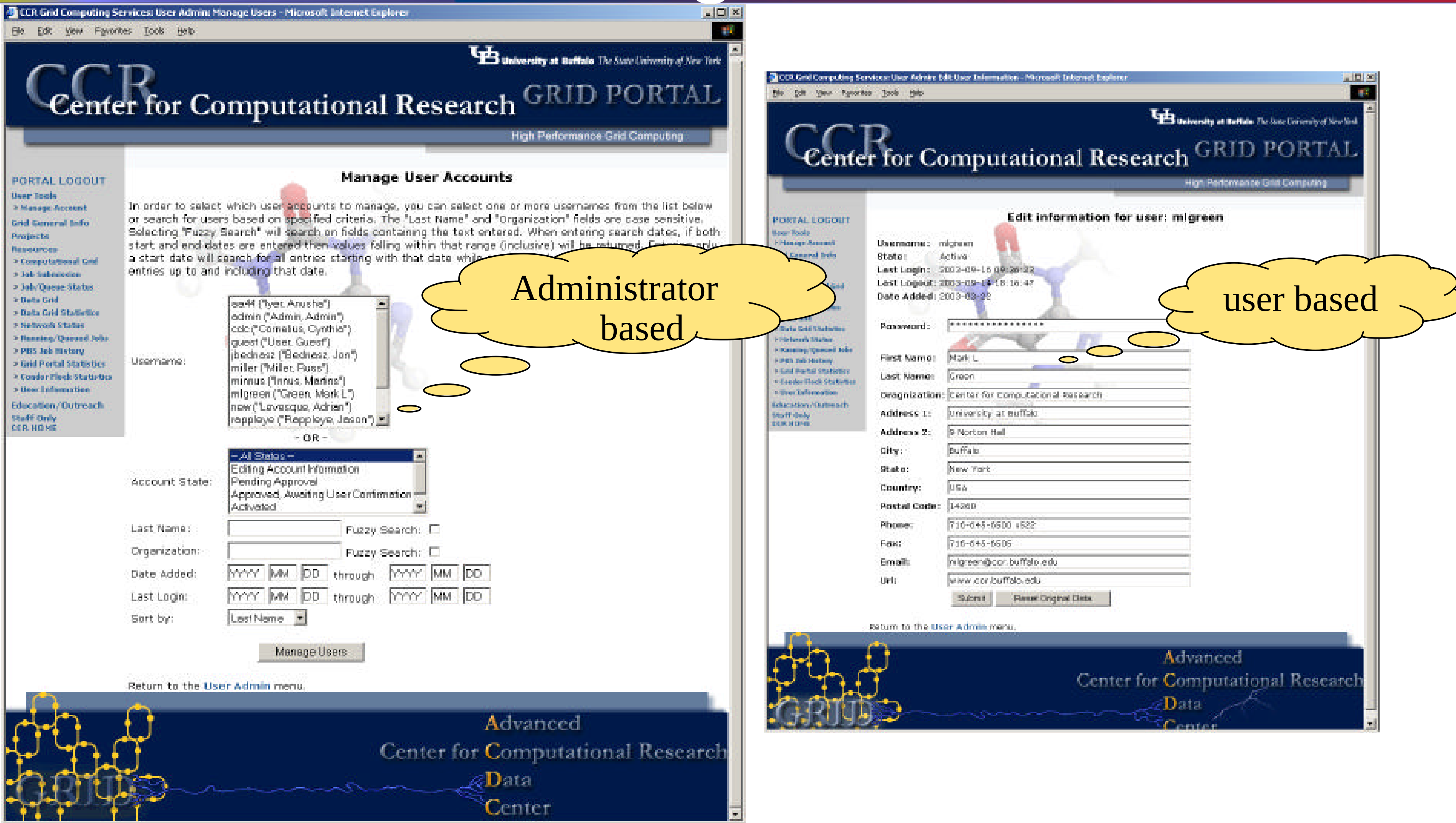
Chart based on: total CPU hours
usage data for: queued jobs
for: All resources
grouped by: user mlgreen

Job_Num	Job_Id	Username	Groupname	Queue	Node_Cnt	Proc_Cnt	Rank	Request_Time(hr)	Start_Time(hr)
1	23364	mlgreen	miler	grid	1	2	1	71.91666666667	queued
2	23365	mlgreen	miler	grid	1	2	2	71.91666666667	queued
3	23366	mlgreen	miler	grid	2	4	3	71.91666666667	queued
4	23367	mlgreen	miler	grid	2	4	4	71.91666666667	queued
5	23368	mlgreen	miler	grid	2	4	5	71.91666666667	queued
6	23369	mlgreen	miler	grid	3	6	6	71.91666666667	queued
7	23370	mlgreen	miler	grid	2	4	7	71.91666666667	queued
8	23371	mlgreen	miler	grid	2	4	8	71.91666666667	queued
9	23372	mlgreen	miler	grid	2	4	9	71.91666666667	queued
10	23373	mlgreen	miler	grid	2	4	10	71.91666666667	queued
11	23374	mlgreen	miler	grid	2	4	11	71.91666666667	queued
12	23375	mlgreen	miler	grid	2	4	12	71.91666666667	queued
13	23376	mlgreen	miler	grid	2	4	13	71.91666666667	queued
14	23377	mlgreen	miler	grid	2	4	14	71.91666666667	queued
15	23378	mlgreen	miler	grid	2	4	15	71.91666666667	queued
16	23379	mlgreen	miler	grid	2	4	16	71.91666666667	queued
17	23380	mlgreen	miler	grid	2	4	17	71.91666666667	queued
18	23381	mlgreen	miler	grid	2	4	18	71.91666666667	queued
19	23382	mlgreen	miler	grid	2	4	19	71.91666666667	queued
20	23383	mlgreen	miler	grid	2	4	20	71.91666666667	queued
21	23384	mlgreen	miler	grid	2	4	21	71.91666666667	queued
22	23385	mlgreen	miler	grid	2	4	22	71.91666666667	queued
23	23386	mlgreen	miler	grid	2	4	23	71.91666666667	queued
24	23387	mlgreen	miler	grid	2	4	24	71.91666666667	queued
25	23388	mlgreen	miler	grid	2	4	25	71.91666666667	queued
26	23389	mlgreen	miler	grid	2	4	26	71.91666666667	queued
27	23393	mlgreen	miler	grid	2	4	27	71.91666666667	queued

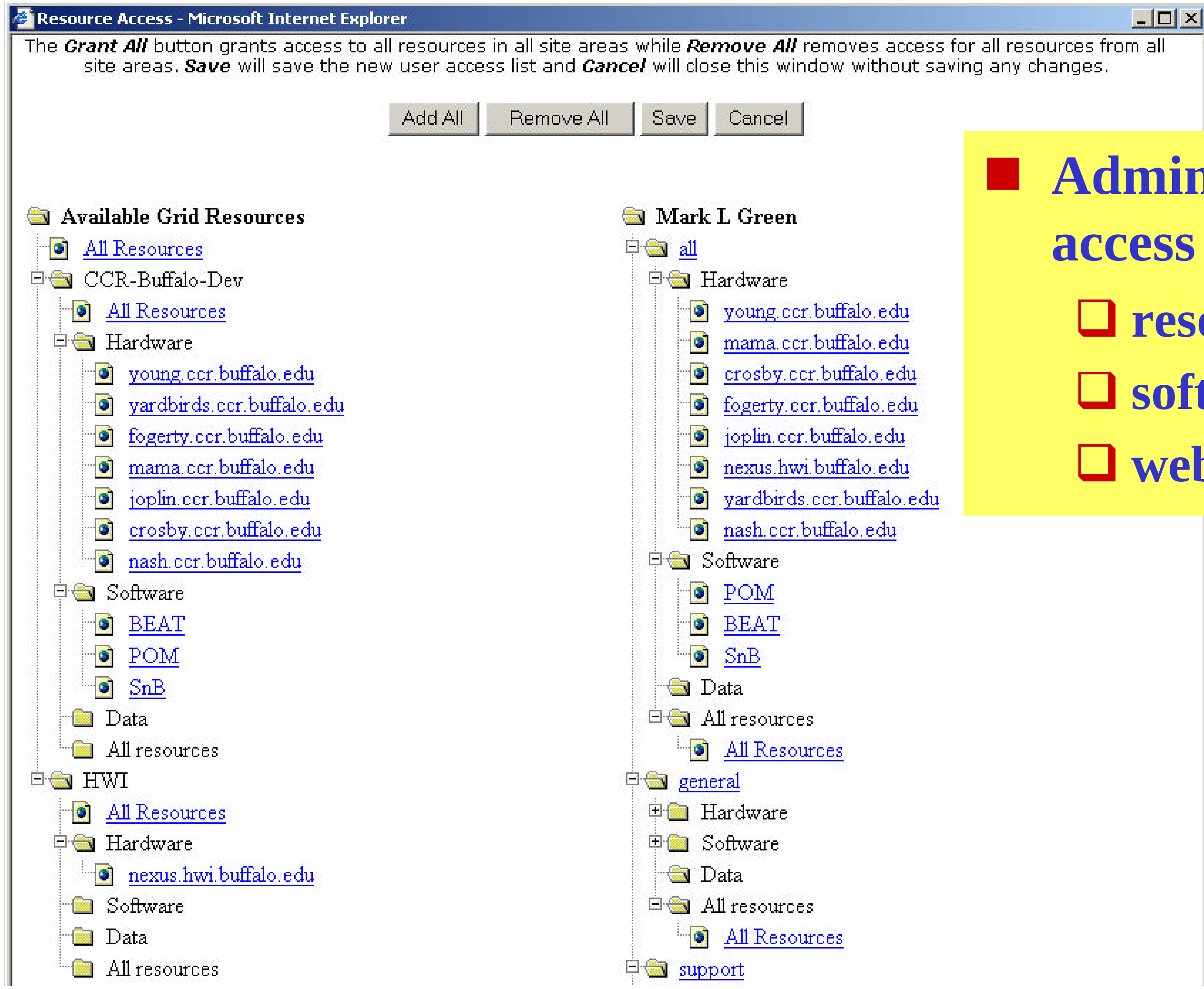
Description of table fields:

Job_Num : Sequential counter for row number.
Job_Id : PBS job identification number corresponding to local PBS queue job number.
Username : PBS job owner local username.
Groupname : PBS job owner local primary group name.

ACDC-Grid Portal User Management

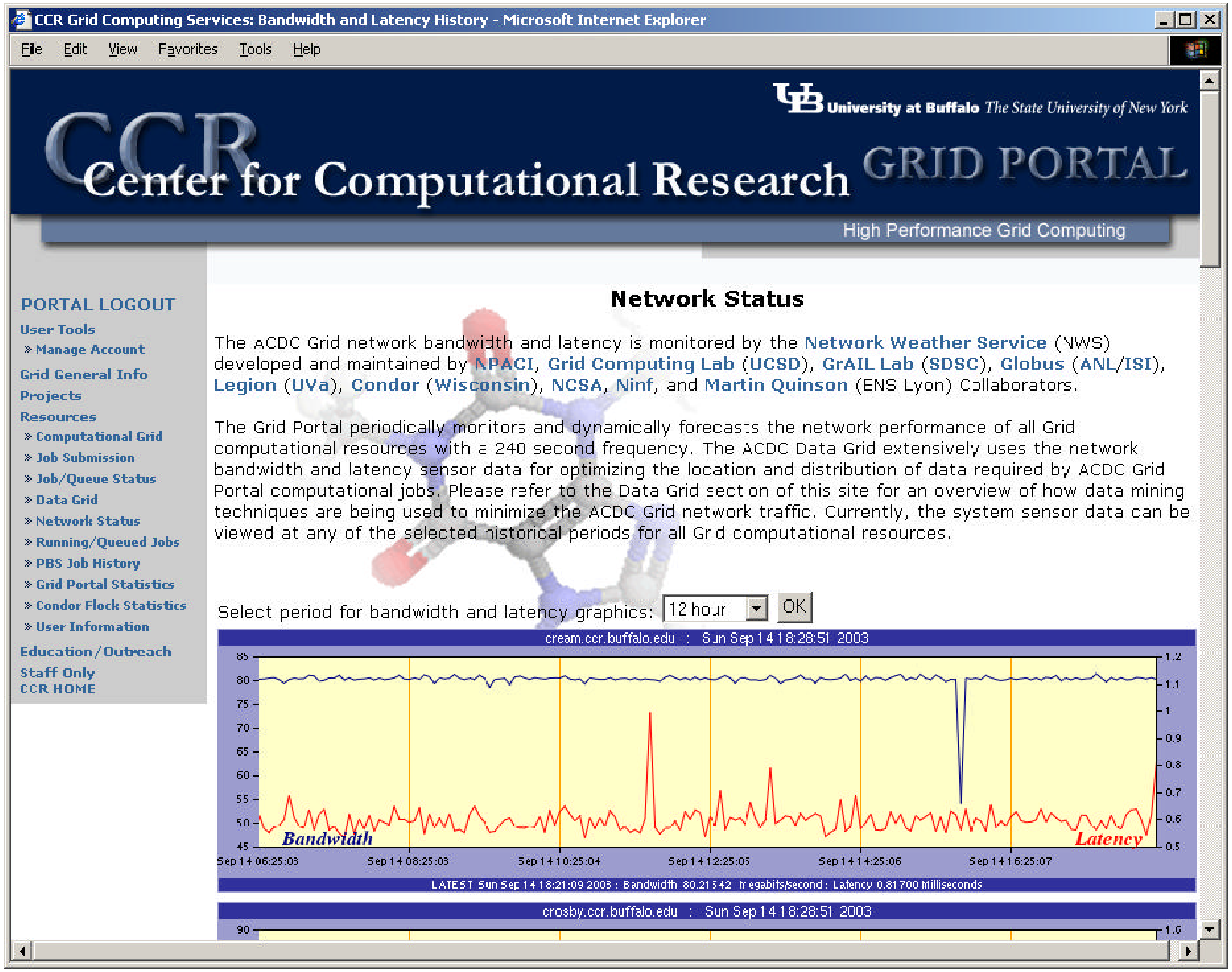


ACDC-Grid Portal Resource Management



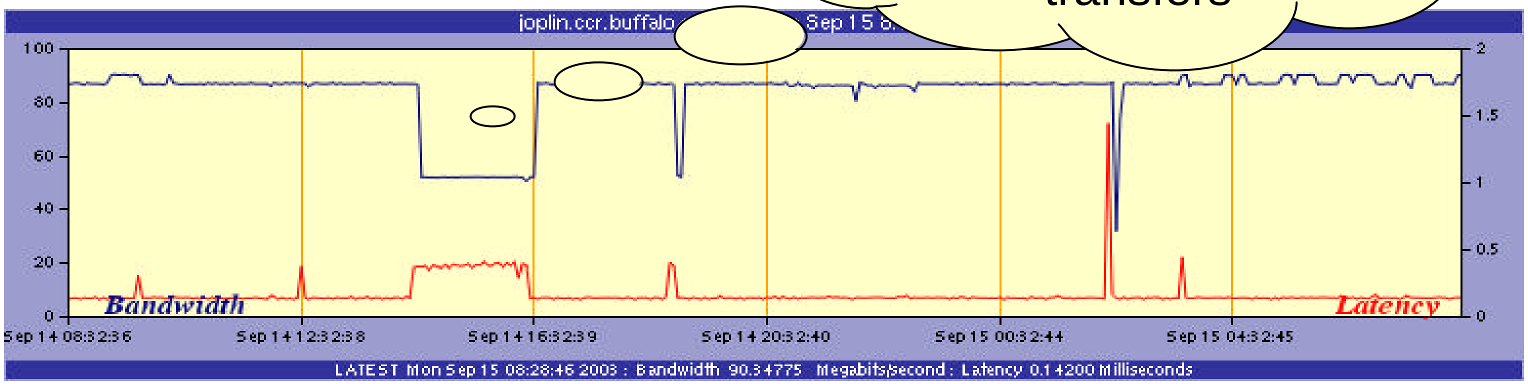
- **Administrator grants a user access to ACDC-Grid**
  - **resources,**
  - **software, and**
  - **web pages.**

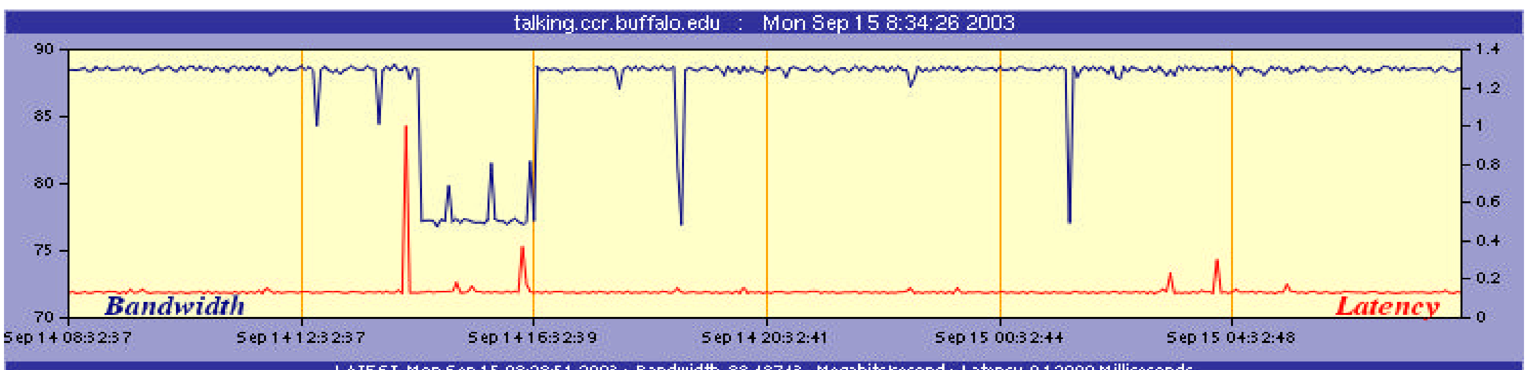
Data Grid Resource Info



Data Grid Resource Info

Both platforms have reduced bandwidth available for additional transfers





Data Grid File Age

Database *data_grid* - Table *file_management* running on *Grid Portal*

Structure | Browse | SQL | Select | Insert | Export | Operations | Options | Empty | Drop

Showing rows 0 - 29 (13000 total, Query took 0.0022 sec)

SQL-query : [Edit] [Explain SQL] [Create PHP Code]

```
SELECT `File_ID` , resource_id, filename, dir_id, access_time, file_age
FROM `file_management`
WHERE 1  LIMIT 0 , 30
```

Show : 30 row(s) starting from record # 30 > >> Page number: 1

in horizontal mode and repeat headers after 100 cells

		File_ID Unique File ID Value	resource_id	filename	dir_id	access_time	file_age
Edit	Delete	62844	10	Dozer.csh	54030	2003-08-16 11:24:52	925832
Edit	Delete	57120	10	Tank.mpg	53304	2003-04-01 01:45:23	12837001
Edit	Delete	57121	10	Neo.ksh	53505	2003-06-08 06:41:05	6904459
Edit	Delete	57122	10	Trinity.m	53499	2003-07-30 11:34:54	2394030
Edit	Delete	57123	10	Rabbit.ksh	53541	2003-07-13 01:43:34	3898310
Edit	Delete	57124	10	Neo.ksh	53407	2003-06-22 06:19:03	5696181
Edit	Delete	57049	10	Agent.ppt	53928	2003-02-24 12:15:39	15909585
Edit	Delete	61710	10	Neo.txt	52724	2003-07-26 09:44:48	2746236
Edit	Delete	61711	10	Morpheus.sh	52710	2003-07-31 07:03:43	2367101
Edit	Delete	61712	10	Morpheus.ppt	52761	2003-08-26 08:04:38	117046
Edit	Delete	61713	10	Tank.jpg	52929	2003-06-26 09:59:37	5337347
Edit	Delete	61714	10	Rabbit.dat	52624	2003-08-26 05:57:43	124661
Edit	Delete	61715	10	KeyMaster.mpg	52770	2003-06-17 04:16:44	6178720

- **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.

ACDC-Grid Development/Maintenance

- **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
- George DeTitta
- Herb Hauptman
- Charles Weeks
- Steve Potter
- 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


CCR
www.ccr.buffalo.edu