

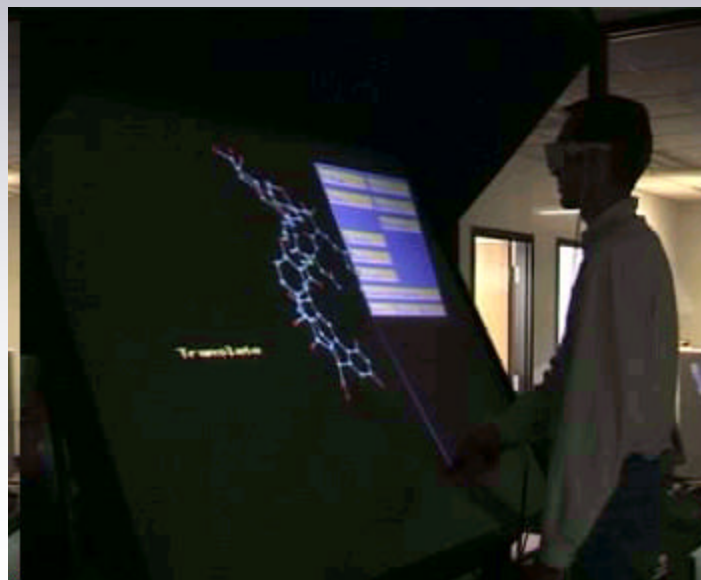
Molecular Structure Determination on a Computational & Data Grid

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**

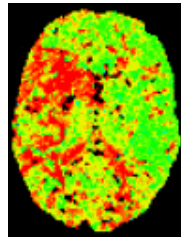
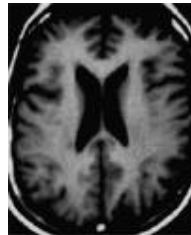
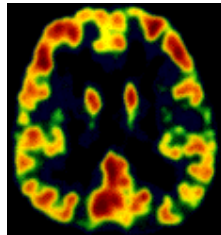
■ **Implantable Pacemaker**

■ **Edible Vaccine for Hepatitis C**

■ **Timed-Release Insulin Therapy**

■ **Anti-Arrhythmia Therapy**

□ **Tarantula venom**



■ **Direct Methods Structure Determination**

□ **Listed on “Top Ten Algorithms of the 20th Century”**

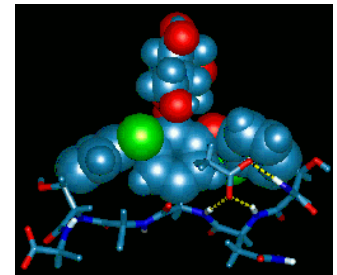
□ **Vancomycin**

□ **Gramacidin A**

■ **High Throughput Crystallization Method: Patented**

■ **NIH National Genomics Center: Northeast Consortium**

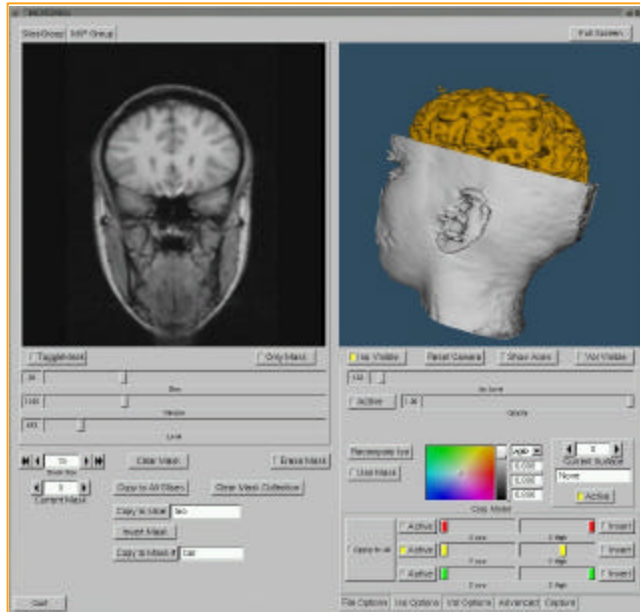
■ **Howard Hughes Medical Institute: Center for Genomics & Proteomics**



Bioinformatics in Buffalo

A \$360M Initiative

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



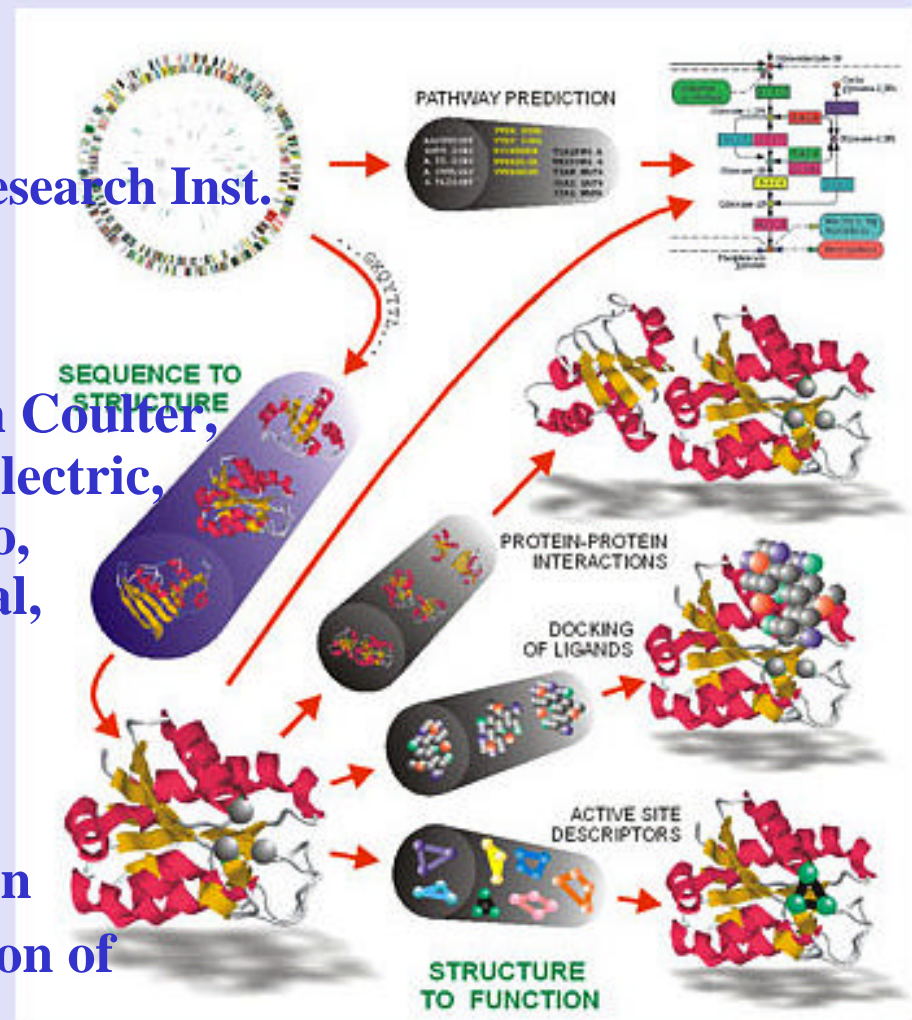
Bioinformatics Partners

■ 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, Zeptomatrix
- ❑ 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

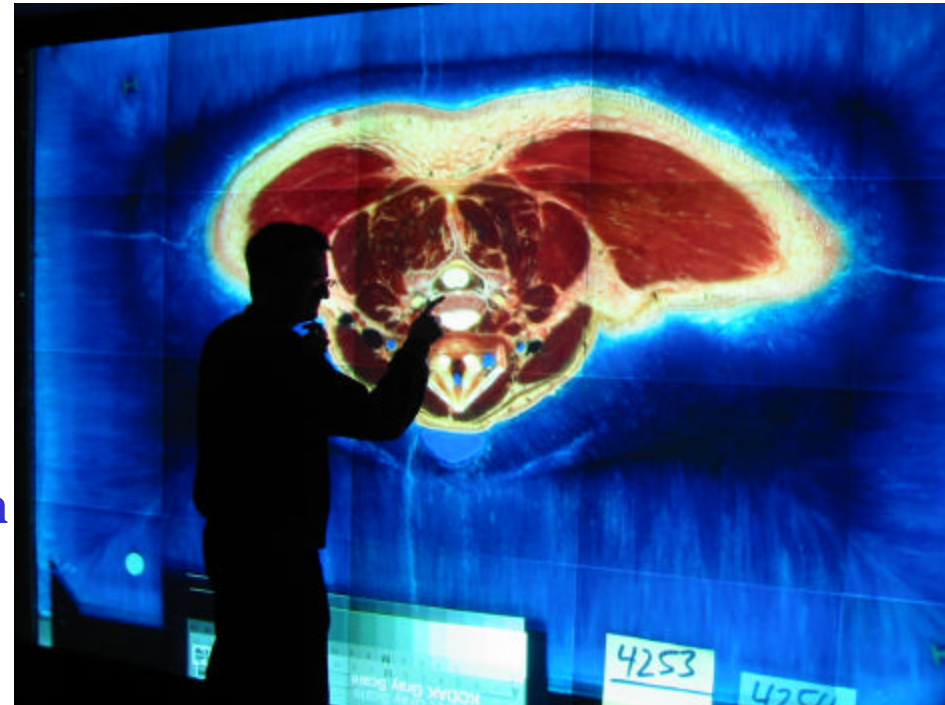
- ❑ 20 NEC projectors: 15.7M pixels
- ❑ Screen is 11' ´ 7'
- ❑ Dell PCs with Myrinet2000

■ Access Grid Node

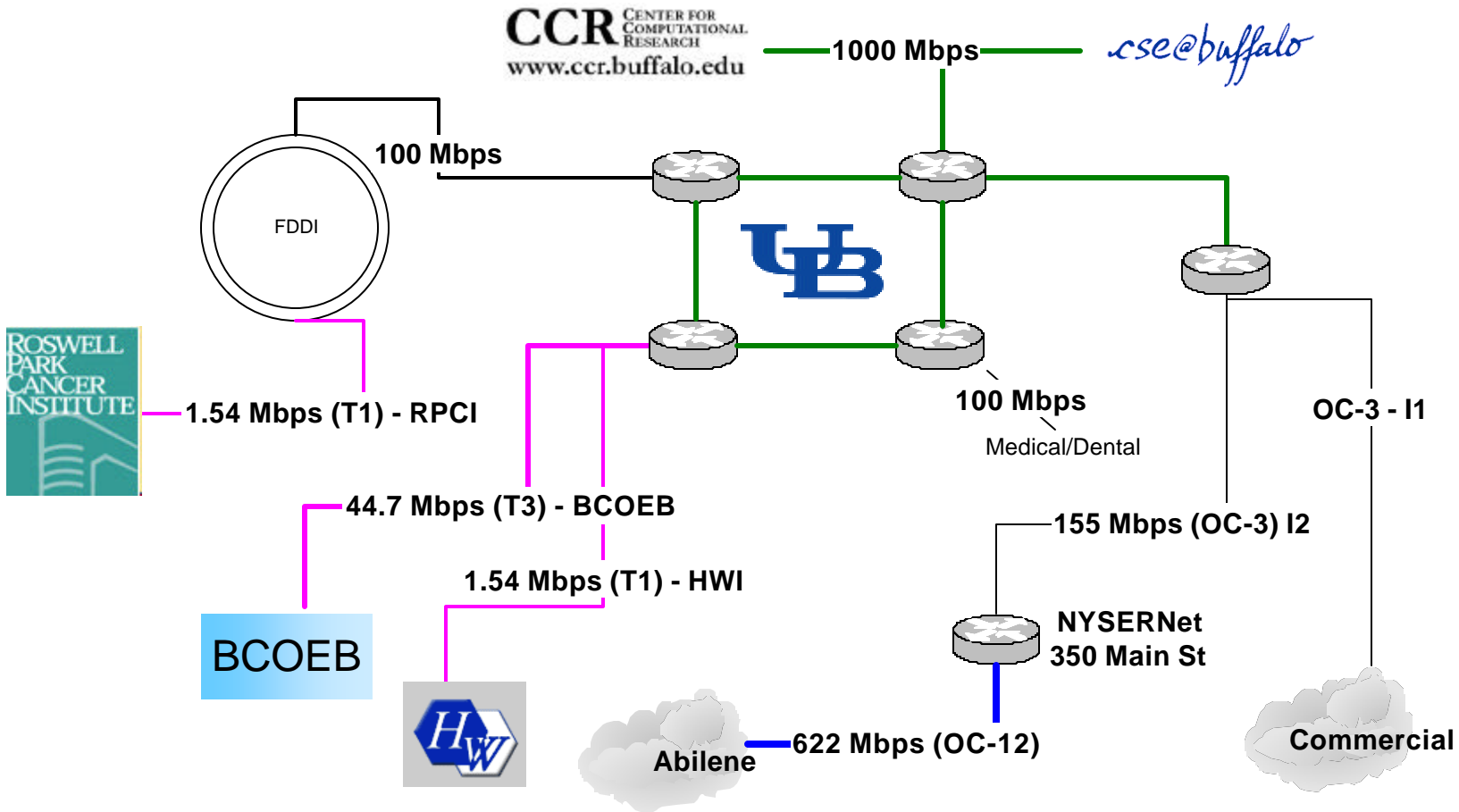
- ❑ Group-to-Group Communication
- ❑ Commodity components

■ SGI Reality Center 3300W

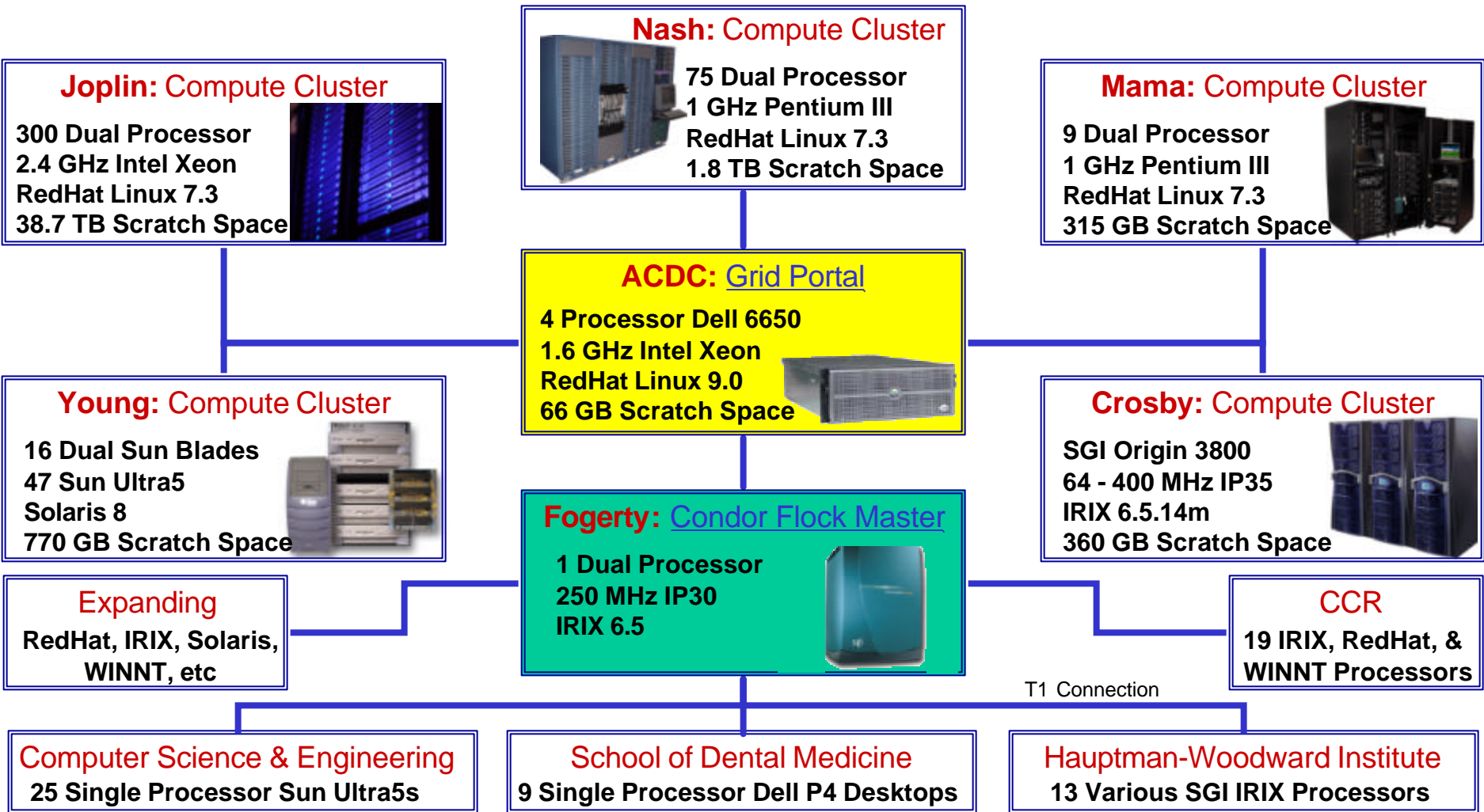
- ❑ Dual Barco's on 8' ´ 4' screen



Network Connections

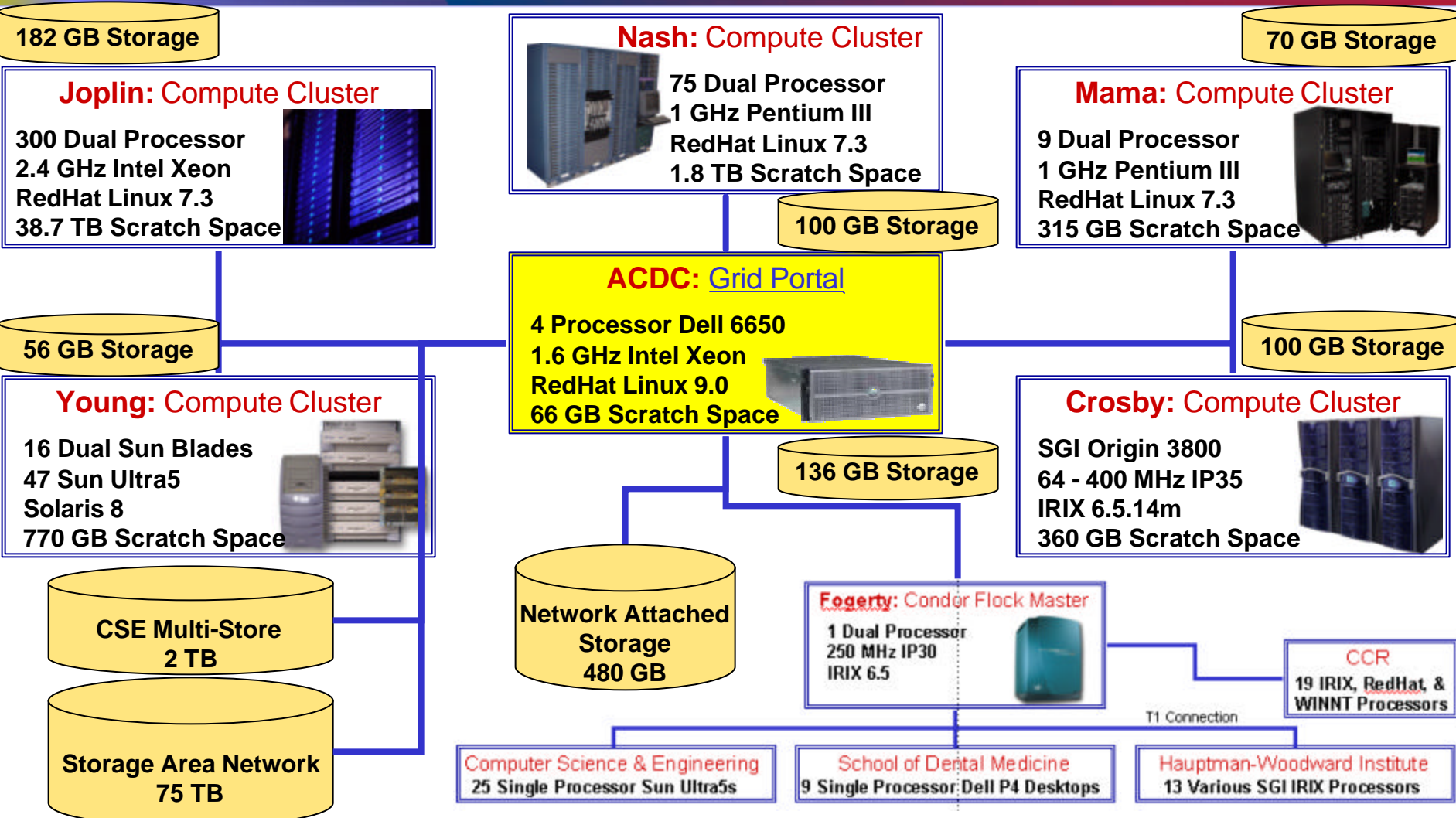


Advanced CCR Data Center (ACDC) Computational Grid Overview



Note: Network connections are 100 Mbps unless otherwise noted.

ACDC Data Grid Overview



Note: Network connections are 100 Mbps unless otherwise noted.

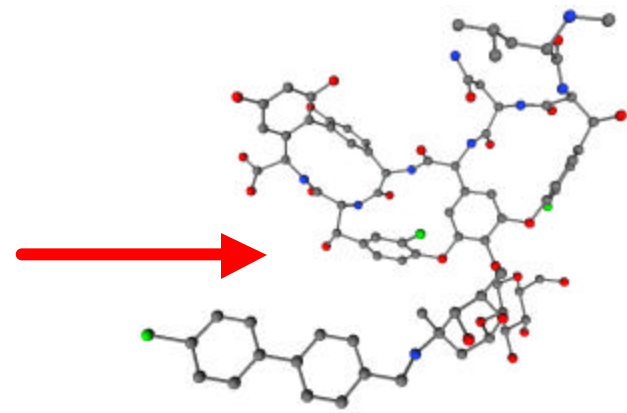
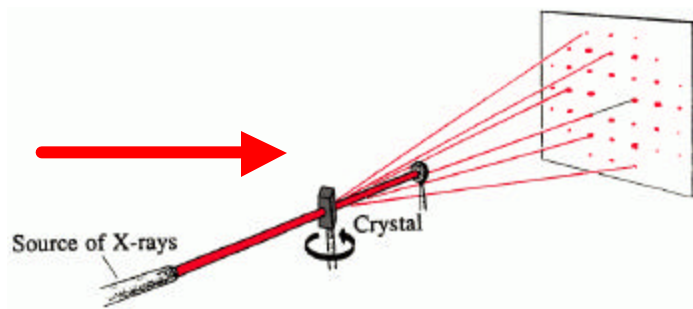
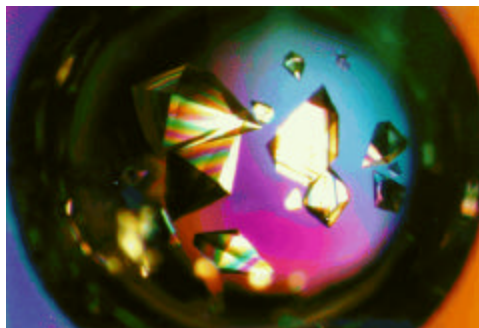


WNY Grid Highlights

- **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)

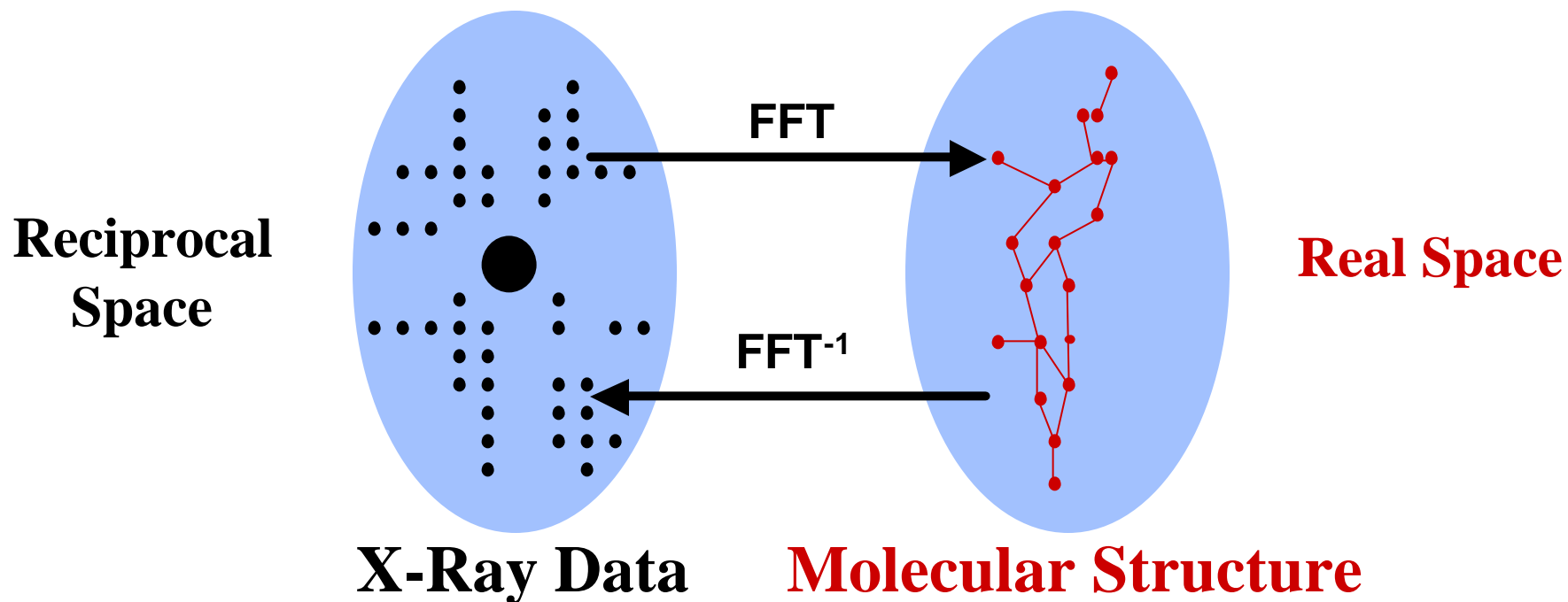
X-Ray Crystallography

- **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 diffraction data.**



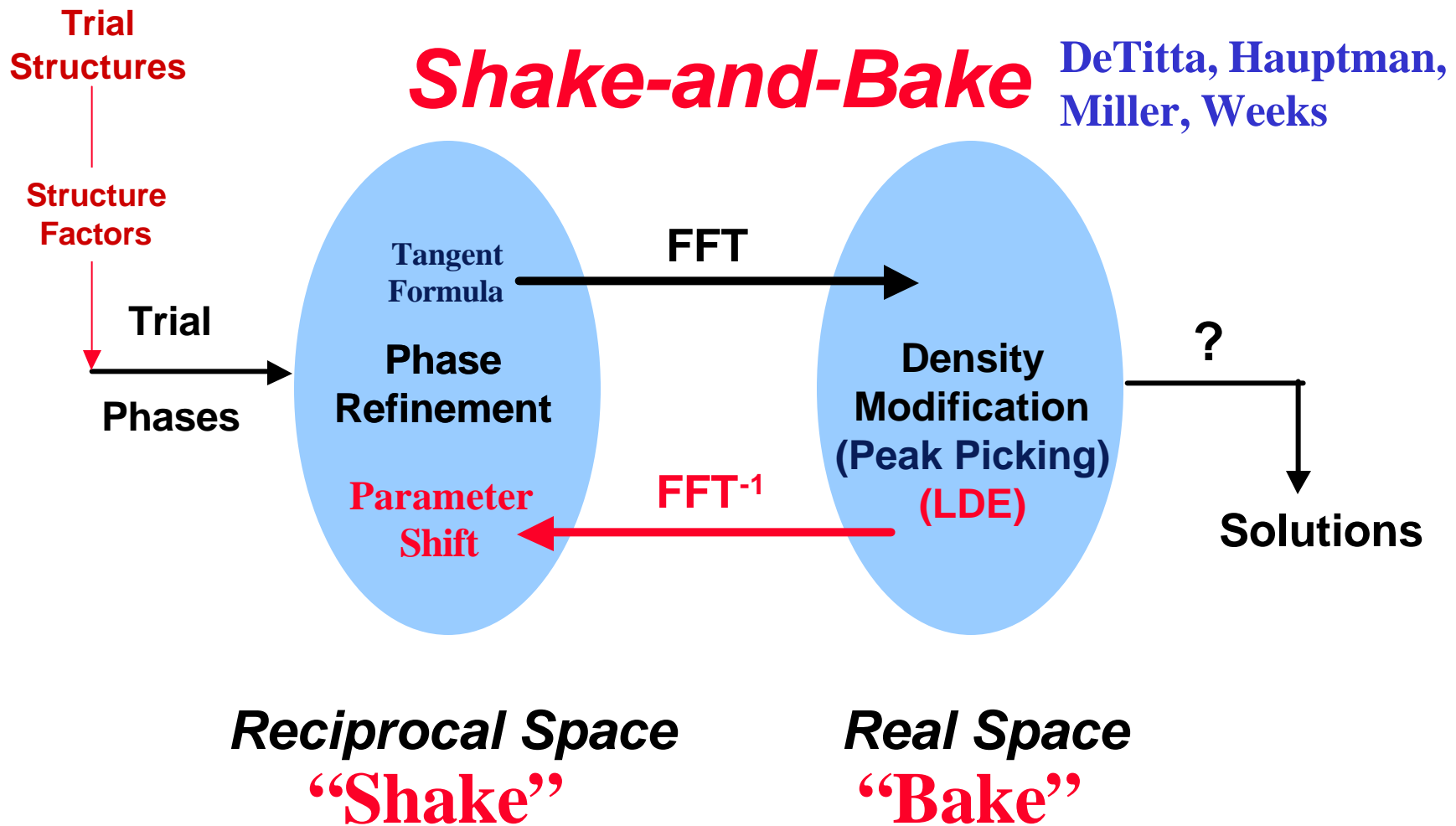
X-Ray Data & Corresponding Molecular Structure

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.

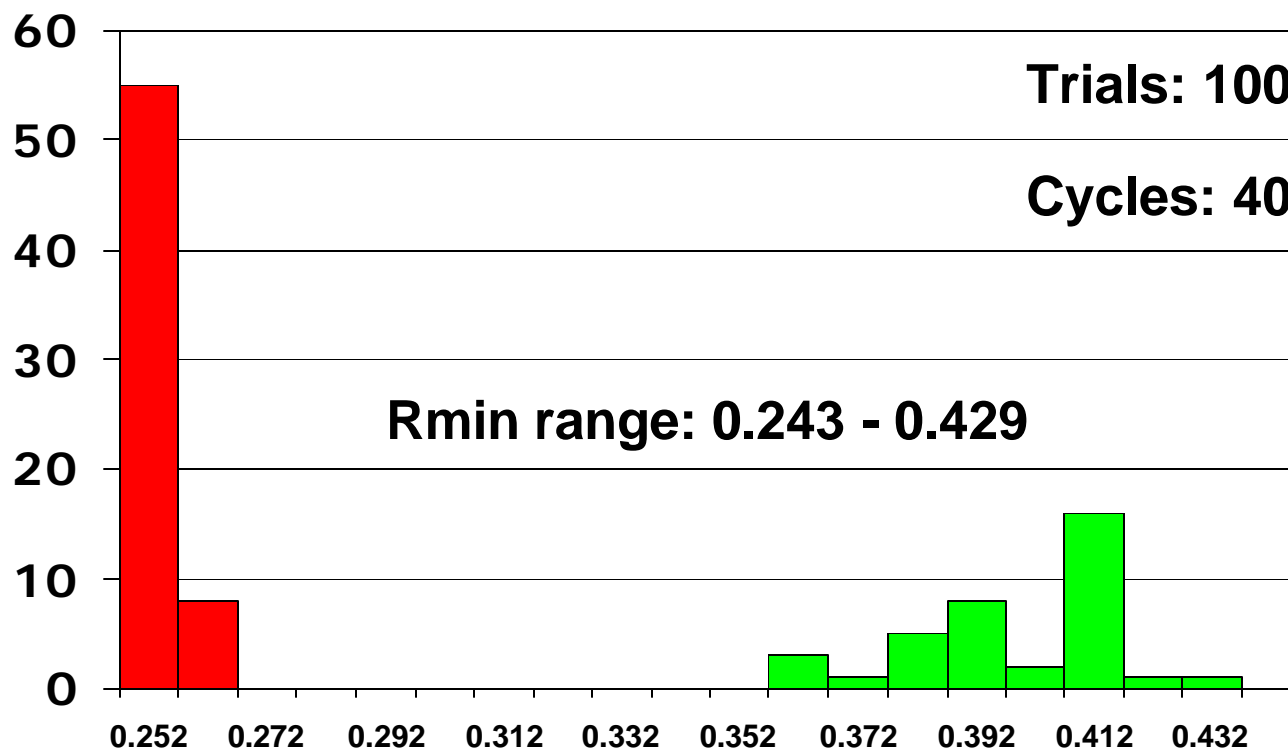
Shake-and-Bake Method: Dual-Space Refinement



Ph8755: *SnB* Histogram

Atoms: 74
Space Group: P1

Phases: 740
Triples: 7,400



Phasing and Structure Size

Se-Met with *Shake-and-Bake*?

Se-Met

190kDa

Multiple Isomorphous Replacement?

Shake-and-Bake

Conventional Direct Methods

Vancomycin



Number of Atoms in Structure

Grid-Based *SnB* Objectives

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

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

Grid Services and Applications

**ACDC-Grid
Computational
Resources**



Applications

Shake-and-Bake

Apache

MySQL

Oracle

High-level Services and Tools

Globus
Toolkit

NWS

MPI

MPI-IO

C, C++, Fortran, PHP

globusrun

Core Services

Metacomputing
Directory
Service

Globus
Security
Interface

GRAM

GASS

Local Services

Condor

Stork

MPI

RedHat Linux

WINNT

LSF

PBS

Maui Scheduler

TCP

UDP

Irix

Solaris

**ACDC-Grid
Data
Resources**



Adapted from Ian Foster and Carl Kesselman



ACDC-Grid Portal Login

CCR Grid Computing Services: - Microsoft Internet Explorer

File Edit View Favorites Tools Help

UB University at Buffalo The State University of New York

CCR Center for Computational Research GRID PORTAL

High Performance Grid Computing

PORTAL LOGIN

- Grid General Info
 - » About ACDC Grid
 - » Computational Grid
 - » Data Grid
 - » Publications
 - » Technical Papers
 - » Presentations
 - » Contact Us
 - » Grid Account Request
 - » Grid Account Support
 - » Events
 - » News
- Projects
- Resources
- Education/Outreach
- Staff Only
- CCR HOME

Login

Username:

Password:

Grid Portal login screen

Advanced Center for Computational Research Data Center

GRID

Data Grid Capabilities

The screenshot shows a web browser window titled "CCR Grid Computing Services: Data Management - Microsoft Internet Explorer". The page header includes the University at Buffalo logo and the text "Center for Computational Research GRID PORTAL High Performance Grid Computing".

On the left side, there is a "PORTAL LOGOUT" menu with the following items:

- User Tools
 - » 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
 - » Condor Flock Statistics
 - » User Information
- Education/Outreach
- Staff Only
- CCR HOME

The main content area displays a file tree for the user "rappleye". At the top, there are controls: "VIEW" set to "Group", "GROUP" set to "miller", and "UserList" set to "rappleye". The file tree shows the following structure:

- Folder: rappleye
 - Folder: KeyMaster
 - Folder: Morpheus
 - Folder: Tank
 - Folder: Agent
 - Folder: Rabbit
 - Folder: Tank
 - Folder: Morpheus
 - File: Oracle.m
 - Folder: Neo
 - Folder: Cypher
 - Folder: Neo
 - Folder: Morpheus
 - Folder: Oracle

A yellow callout bubble points to the "Oracle.m" file with the text: "Browser view of 'miller' group files published by user 'rappleye'".

At the bottom of the page, there is a decorative footer with the text: "Advanced Center for Computational Research Data".

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.

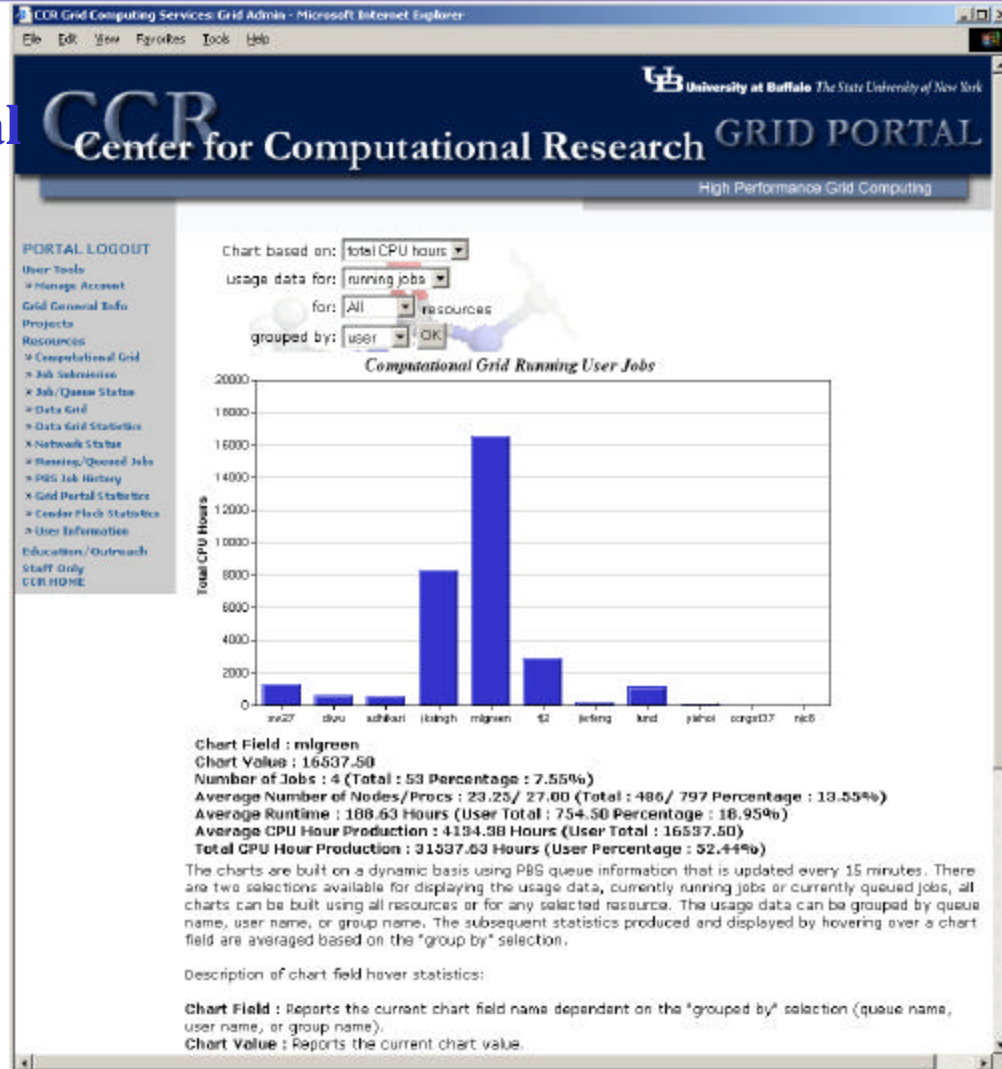
□ Usage data for:

- running jobs, or
- queued jobs.

□ Individual or all resources.

□ Grouped by:

- group, or
- user, or
- queue.



Grid Portal Job Status

CCR Grid Computing Services: Grid Admin - Microsoft Internet Explorer

File Edit View Favorites Tools Help

CCR University at Buffalo The State University of New York
Center for Computational Research GRID PORTAL
High Performance Grid Computing

PORTAL LOGOUT
User Tools
Manage Account
Grid General Info
Projects
Resources
Computational Grid
Job Submissions
Job/Queue Status
Data Grid
Data Grid Statistics
Network Status
Running/Queued Jobs
PBS Job History
Grid Portal Statistics
Center Fleck Statistics
User Information
Education/Outreach
Staff Only
CCR 80246

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

Computational Grid Queued User Jobs

User	Total CPU Hours
mlgreen	7623.17
ksingh	10800.00
sa27	500.00
dsingh	6500.00
rtidici	200.00
chen938	3000.00
janjoti	1500.00
janmg	800.00
libon	100.00

Chart Field : mlgreen
Chart Value : 7623.17
Number of Jobs : 27 (Total : 60 Percentage : 45.00%)
Average Number of Nodes/Procs : 1.96 / 3.93 (Total : 1641 / 3206 Percentage : 9.31%)
Average Runtime : 71.92 Hours (User Total : 1941.75 Percentage : 71.90%)
Average CPU Hour Production : 282.34 Hours (User Total : 7623.17)
Total CPU Hour Production : 31015.17 Hours (User Percentage : 24.58%)

The charts are built on a dynamic basis using PBS queue information that is updated every 15 minutes are two selections available for displaying the usage data, currently running jobs or currently queued charts can be built using all resources or for any selected resource. The usage data can be grouped name, user name, or group name. The subsequent statistics produced and displayed by hovering over field are averaged based on the "group by" selection.

Description of chart field hover statistics:

Chart Field : Reports the current chart field name dependant on the "grouped by" selection (queue user name, or group name).
Chart Value : Reports the current chart value.

CCR Grid Computing Services: Grid Admin - Microsoft Internet Explorer

File Edit View Favorites Tools Help

CCR University at Buffalo The State University of New York
Center for Computational Research GRID PORTAL
High Performance Grid Computing

PORTAL LOGOUT
User Tools
Manage Account
Grid General Info
Projects
Resources
Computational Grid
Job Submissions
Job/Queue Status
Data Grid
Data Grid Statistics
Network Status
Running/Queued Jobs
PBS Job History
Grid Portal Statistics
Center Fleck Statistics
User Information
Education/Outreach
Staff Only
CCR 80246

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	Nodes	Cnt	Proc	Cnt	Rank	Request_Time(hr)	Start_Time(hr)
1	23364	mlgreen	miller	grid	1	2	1	71.916666666667	queued		
2	23365	mlgreen	miller	grid	1	2	2	71.916666666667	queued		
3	23366	mlgreen	miller	grid	2	4	3	71.916666666667	queued		
4	23367	mlgreen	miller	grid	2	4	4	71.916666666667	queued		
5	23368	mlgreen	miller	grid	2	4	5	71.916666666667	queued		
6	23369	mlgreen	miller	grid	3	6	6	71.916666666667	queued		
7	23370	mlgreen	miller	grid	2	4	7	71.916666666667	queued		
8	23371	mlgreen	miller	grid	2	4	8	71.916666666667	queued		
9	23372	mlgreen	miller	grid	2	4	9	71.916666666667	queued		
10	23373	mlgreen	miller	grid	2	4	10	71.916666666667	queued		
11	23374	mlgreen	miller	grid	2	4	11	71.916666666667	queued		
12	23375	mlgreen	miller	grid	2	4	12	71.916666666667	queued		
13	23376	mlgreen	miller	grid	2	4	13	71.916666666667	queued		
14	23377	mlgreen	miller	grid	2	4	14	71.916666666667	queued		
15	23378	mlgreen	miller	grid	2	4	15	71.916666666667	queued		
16	23379	mlgreen	miller	grid	2	4	16	71.916666666667	queued		
17	23380	mlgreen	miller	grid	2	4	17	71.916666666667	queued		
18	23381	mlgreen	miller	grid	2	4	18	71.916666666667	queued		
19	23382	mlgreen	miller	grid	2	4	19	71.916666666667	queued		
20	23383	mlgreen	miller	grid	2	4	20	71.916666666667	queued		
21	23384	mlgreen	miller	grid	2	4	21	71.916666666667	queued		
22	23385	mlgreen	miller	grid	2	4	22	71.916666666667	queued		
23	23386	mlgreen	miller	grid	2	4	23	71.916666666667	queued		
24	23387	mlgreen	miller	grid	2	4	24	71.916666666667	queued		
25	23388	mlgreen	miller	grid	2	4	25	71.916666666667	queued		
26	23389	mlgreen	miller	grid	2	4	26	71.916666666667	queued		
27	23393	mlgreen	miller	grid	2	4	27	71.916666666667	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

CCR Grid Computing Services: User Admin: Manage Users - 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 Queue Status

Job Submission

Data Grid

Data Grid Statistics

Network Status

Running/Queued Jobs

PBS Job History

Grid Portal Statistics

Consider Med. Statistics

New Information

Education / Outreach

Staff Only

CCR HD MC

Manage User Accounts

In order to select which user accounts to manage, you can select one or more usernames from the list below or search for users based on specified criteria. The "Last Name" and "Organization" fields are case sensitive. Selecting "Fuzzy Search" will search on fields containing the text entered. When entering search dates, if both start and end dates are entered then values falling within that range (inclusive) will be returned. Entering only a start date will search for all entries starting with that date while entries up to and including that date.

Username: Fuzzy Search

Account State: Fuzzy Search

Last Name: Fuzzy Search

Organization: Fuzzy Search

Date Added: / / through / /

Last Login: / / through / /

Sort by:

Return to the User Admin menu.

Advanced Center for Computational Research Data Center

Administrator based

CCR Grid Computing Services: User Admin: Edit User Information - 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 Queue Status

Job Submission

Data Grid

Data Grid Statistics

Network Status

Running/Queued Jobs

PBS Job History

Grid Portal Statistics

Consider Med. Statistics

New Information

Education / Outreach

Staff Only

CCR HD MC

Edit information for user: mlgreen

Username: mlgreen

State: Active

Last Login: 2003-09-15 00:20:22

Last Logout: 2003-09-15 15:10:47

Date Added: 2003-02-22

Password: [masked]

First Name: Mark L.

Last Name: Green

Organization: Center for Computational Research

Address 1: University at Buffalo

Address 2: 0 Norton Hall

City: Buffalo

State: New York

Country: USA

Postal Code: 14260

Phone: 716-645-6500 x522

Fax: 716-645-6505

Email: mlgreen@ccr.buffalo.edu

Url: www.ccr.buffalo.edu

Return to the User Admin menu.

Advanced Center for Computational Research Data Center

user based

ACDC-Grid Portal Resource Management

Resource Access - Microsoft Internet Explorer

The **Grant All** button grants access to all resources in all site areas while **Remove All** removes access for all resources from all site areas. **Save** will save the new user access list and **Cancel** will close this window without saving any changes.

Available Grid Resources

- [All Resources](#)
- CCR-Buffalo-Dev
 - [All Resources](#)
 - Hardware
 - [young.ccr.buffalo.edu](#)
 - [yardbirds.ccr.buffalo.edu](#)
 - [fogerty.ccr.buffalo.edu](#)
 - [mama.ccr.buffalo.edu](#)
 - [joplin.ccr.buffalo.edu](#)
 - [crosby.ccr.buffalo.edu](#)
 - [nash.ccr.buffalo.edu](#)
 - Software
 - [BEAT](#)
 - [POM](#)
 - [SnB](#)
 - Data
 - All resources
- HWI
 - [All Resources](#)
 - Hardware
 - [nexus.hwi.buffalo.edu](#)
 - Software
 - Data
 - All resources

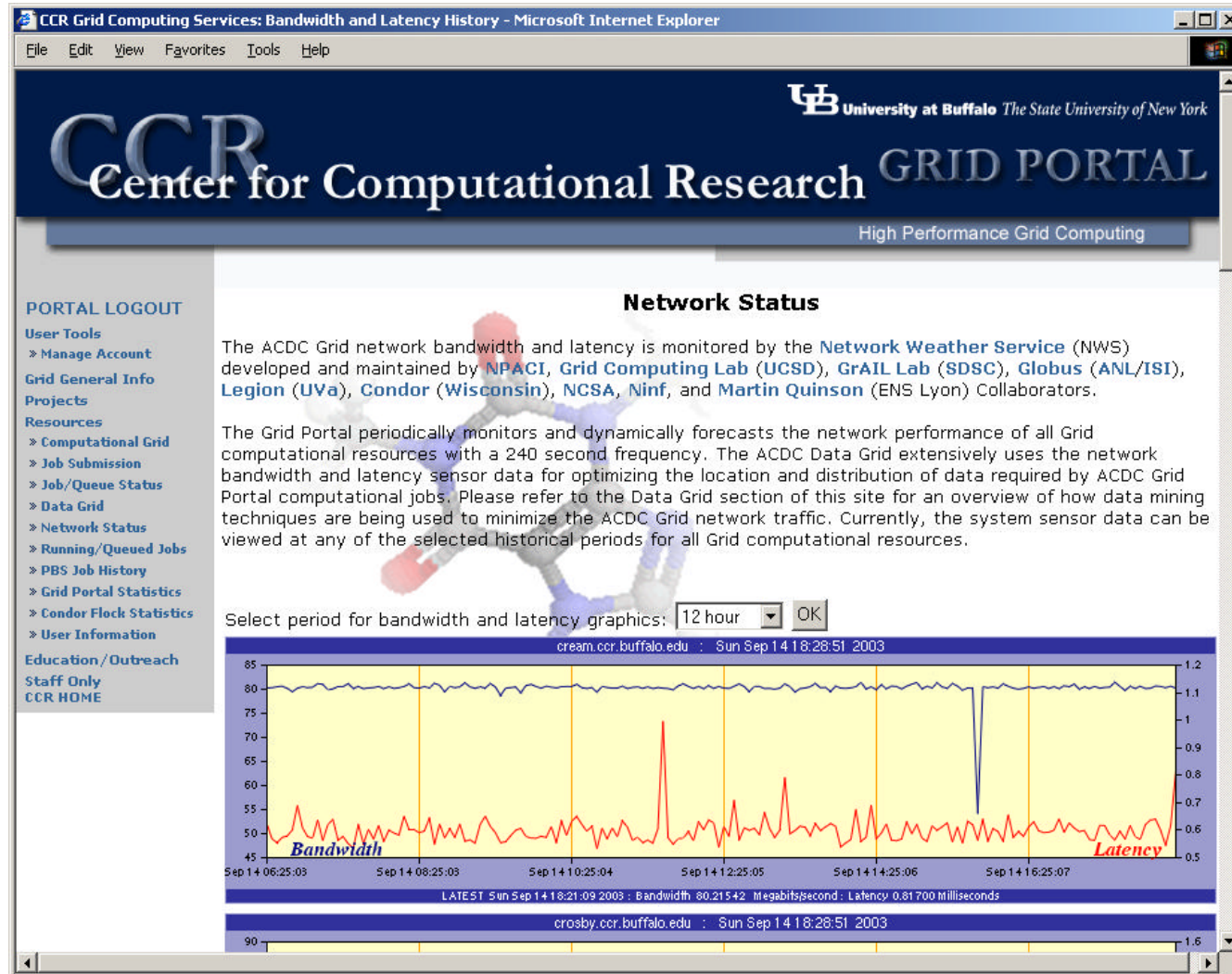
Mark I. Green

- [all](#)
 - Hardware
 - [young.ccr.buffalo.edu](#)
 - [mama.ccr.buffalo.edu](#)
 - [crosby.ccr.buffalo.edu](#)
 - [fogerty.ccr.buffalo.edu](#)
 - [joplin.ccr.buffalo.edu](#)
 - [nexus.hwi.buffalo.edu](#)
 - [yardbirds.ccr.buffalo.edu](#)
 - [nash.ccr.buffalo.edu](#)
 - Software
 - [POM](#)
 - [BEAT](#)
 - [SnB](#)
 - Data
 - All resources
 - [All Resources](#)
 - [general](#)
 - Hardware
 - Software
 - Data
 - All resources
 - [All Resources](#)
 - [support](#)

■ Administrator grants a user access to ACDC-Grid

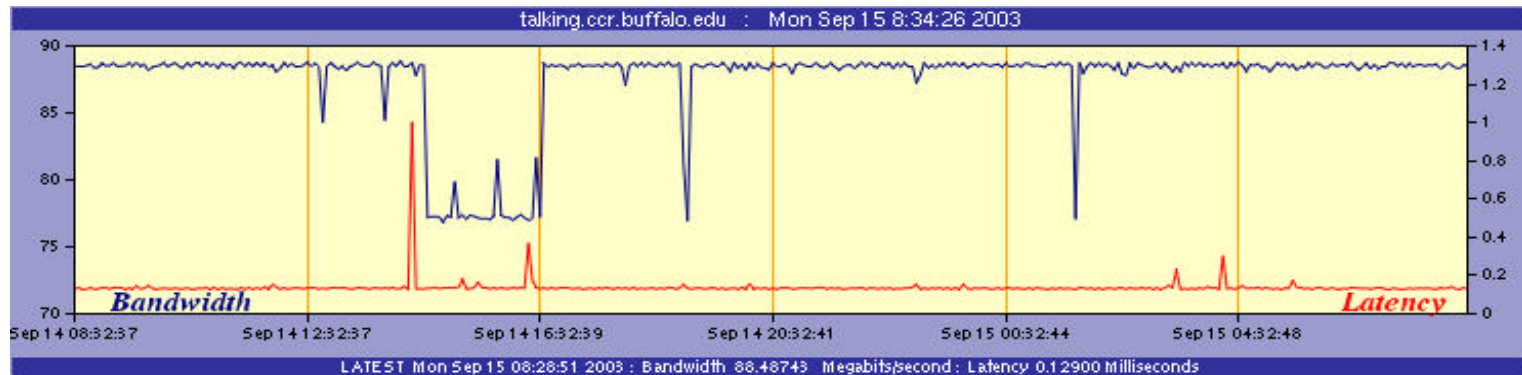
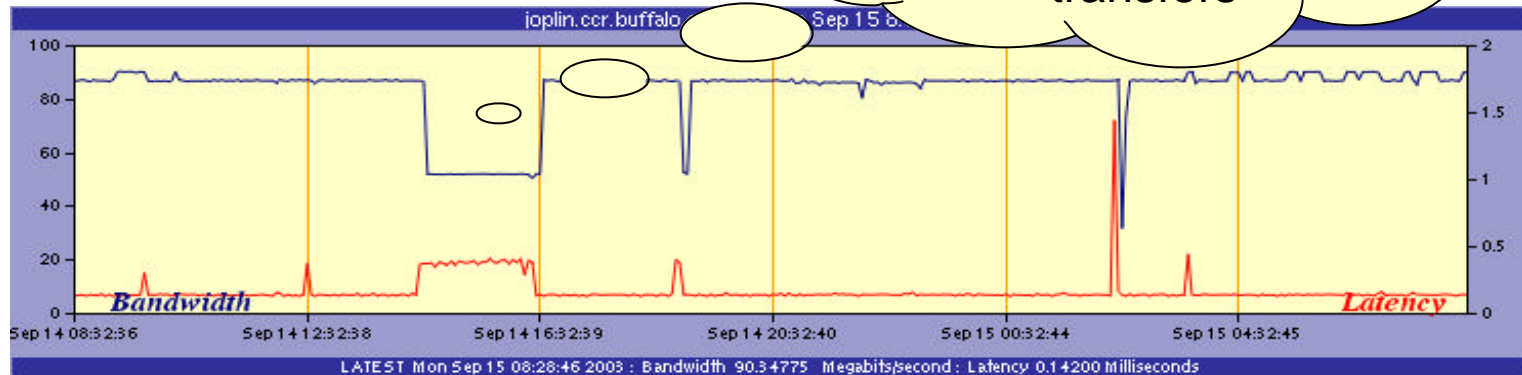
- resources,
- software, and
- web pages.

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
in horizontal mode and repeat headers after 100 cells Page number: 1

		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
- National Science Foundation, National Institutes of Health, Oishei Foundation, Wendt Foundation, Sloan Foundation, Verizon, NYS
- Gov Pataki, Congressman Reynolds, Senator Clinton, Senator Schumer, Congressman Quinn
- George DeTitta
- Herb Hauptman
- Charles Weeks
- Steve Potter



University at Buffalo
Department of

www.ccr.buffalo.edu