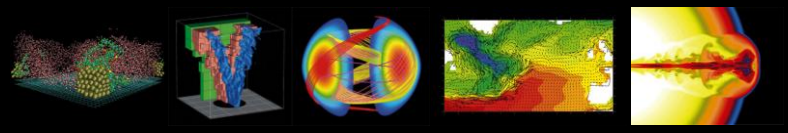
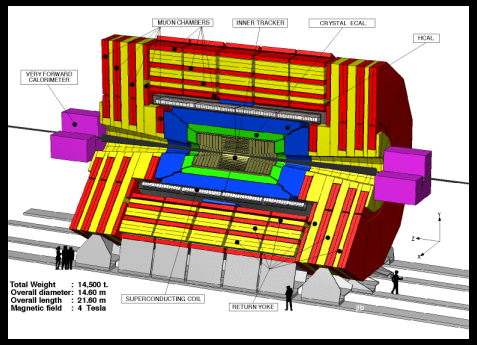
# Research and Development of CyberInfrastructure to Support End Science

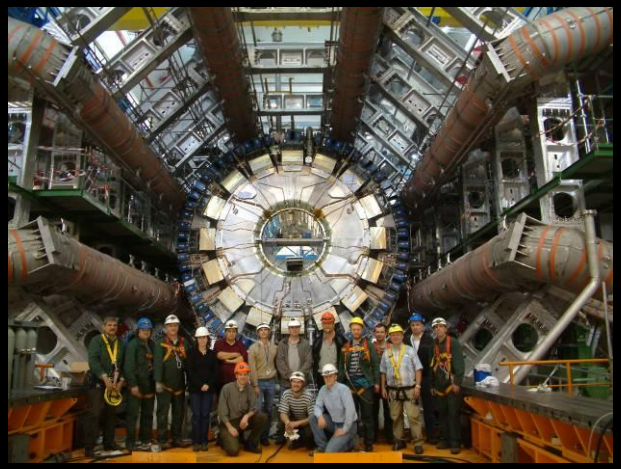
**Deb Agarwal**
**Distributed Systems Department**

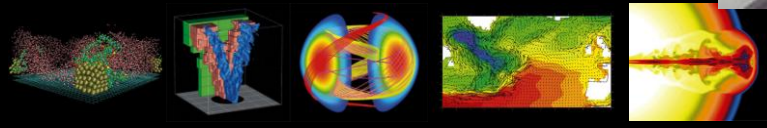
**Lawrence Berkeley National Laboratory**



# Distributed Large-Scale Science





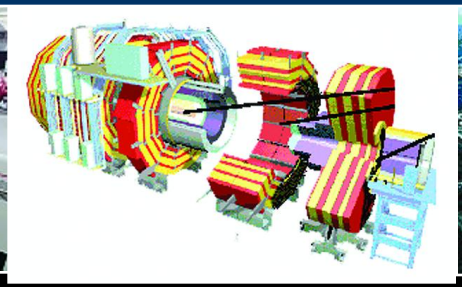
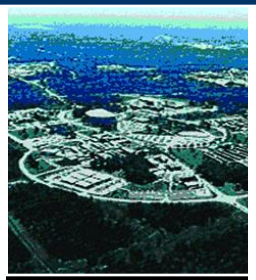
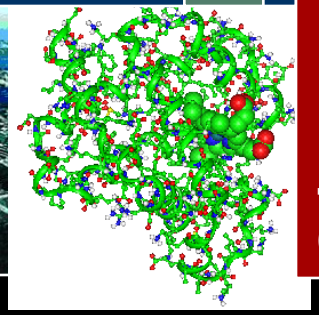


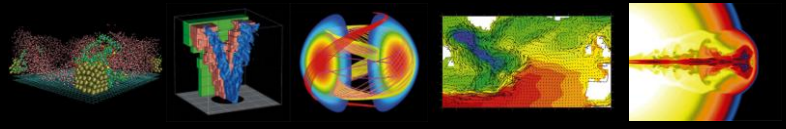
# Robust Science Support Framework - Grid

**Web Services, Portals, Collaboration Tools, Problem Solving Environments**

- Authentication and Authorization
- Resource Discovery
- Secure Communication
- Event Services And Monitoring
- Data Transfer
- Scheduling
- Data Curation
- Compute Services
- Application Servers
- Asynchrony Support
- Virtual Organization

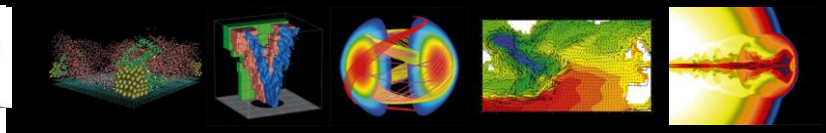
**Cybersecurity Protections**



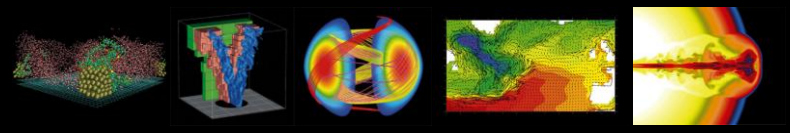
# DSD Expertise

- Grid middleware
- Cybersecurity mechanisms
- Application interfaces and tools
- Collaborative interaction tools
- End-2-end monitoring
- Network transport protocols and overlays
- Real-time systems
- Software engineering

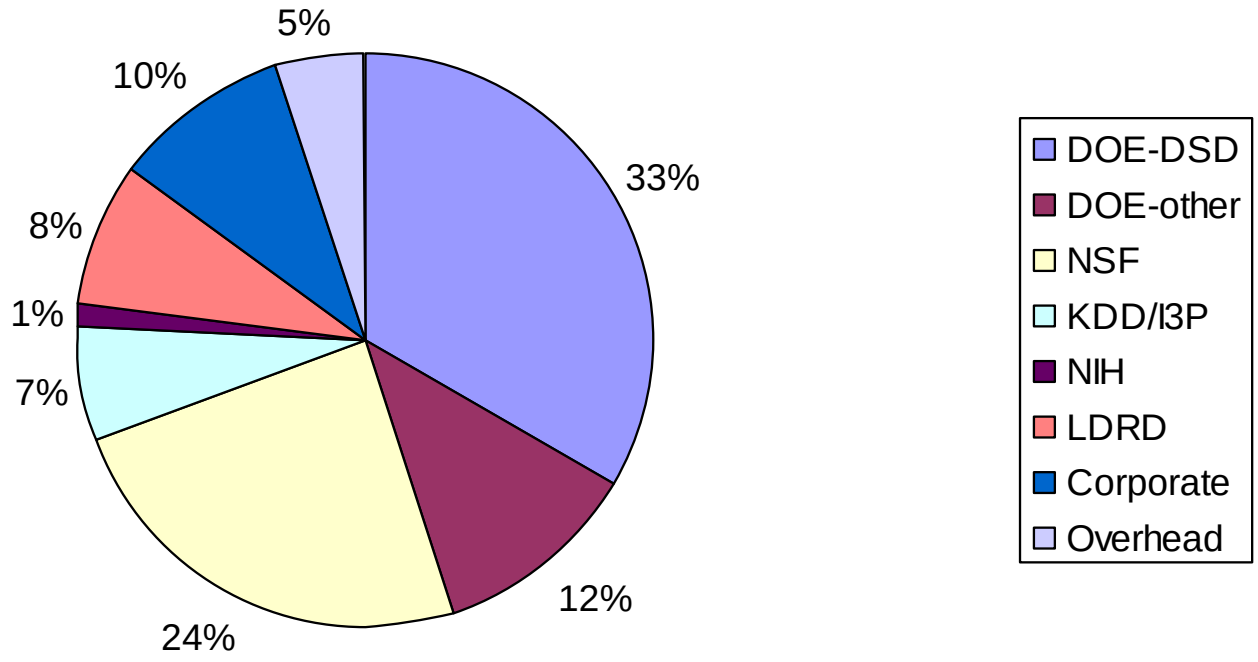


# Mix of Projects and Activities

- Research and development of new software-based capabilities and paradigms
  - Workflow
  - Security
  - Collaboration
  - Intrusion detection
  - …
- Software development for a specific science group
  - IceCube
  - Ameriflux
  - Fusion
  - …
- Large multi-agency collaborations
  - Grid middleware
  - Cybersecurity
  - …
- Community Service



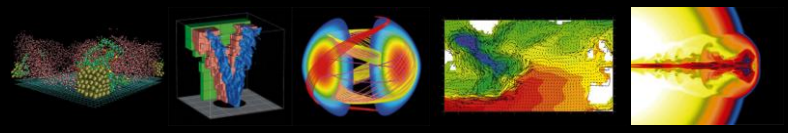
# Mix of Funding Sources




~$4 million/year

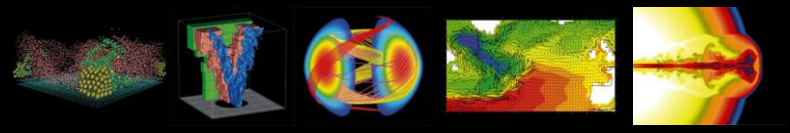
# Research and Development of New Software-based Capabilities and Paradigms

A few examples



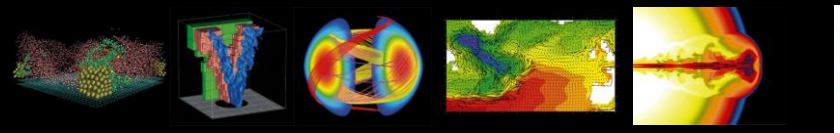
# Bro Intrusion Detection System

- Passively monitors network traffic for suspicious behavior or policy violations: incoming/outgoing/internal
  - Customizable for individual site policy
- In conjunction with blocking routers, Bro acts as a dynamic and intelligent firewall
  - Bro controls the blocking router, dynamically blocking hosts when an intrusion attempt is detected or alerts upon suspicious activity
- Very different approach from “Signature-based” IDS’s such as Snort
- Bro does full protocol analysis of most common application level protocols:
  - HTTP, FTP, telnet, rlogin, rsh, RPC, DCE/RPC, DNS, Windows Domain Service, SMTP, IRC, POP3, NTP, ARP, ICMP, SSH, SSL, TFTP, Gnutella, etc.



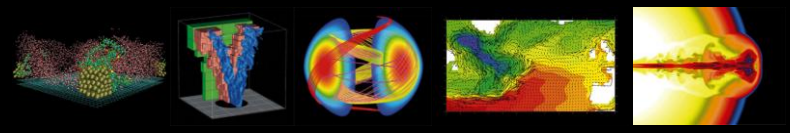
# DHS Enterprise Traces Project

- Several anonymized public datasets of Internet backbone and site border traffic exist
- But *NO* public datasets of internal traffic from a large site have ever been released, until now!
- Bro's protocol analyzers were used to characterize and anonymize the data.

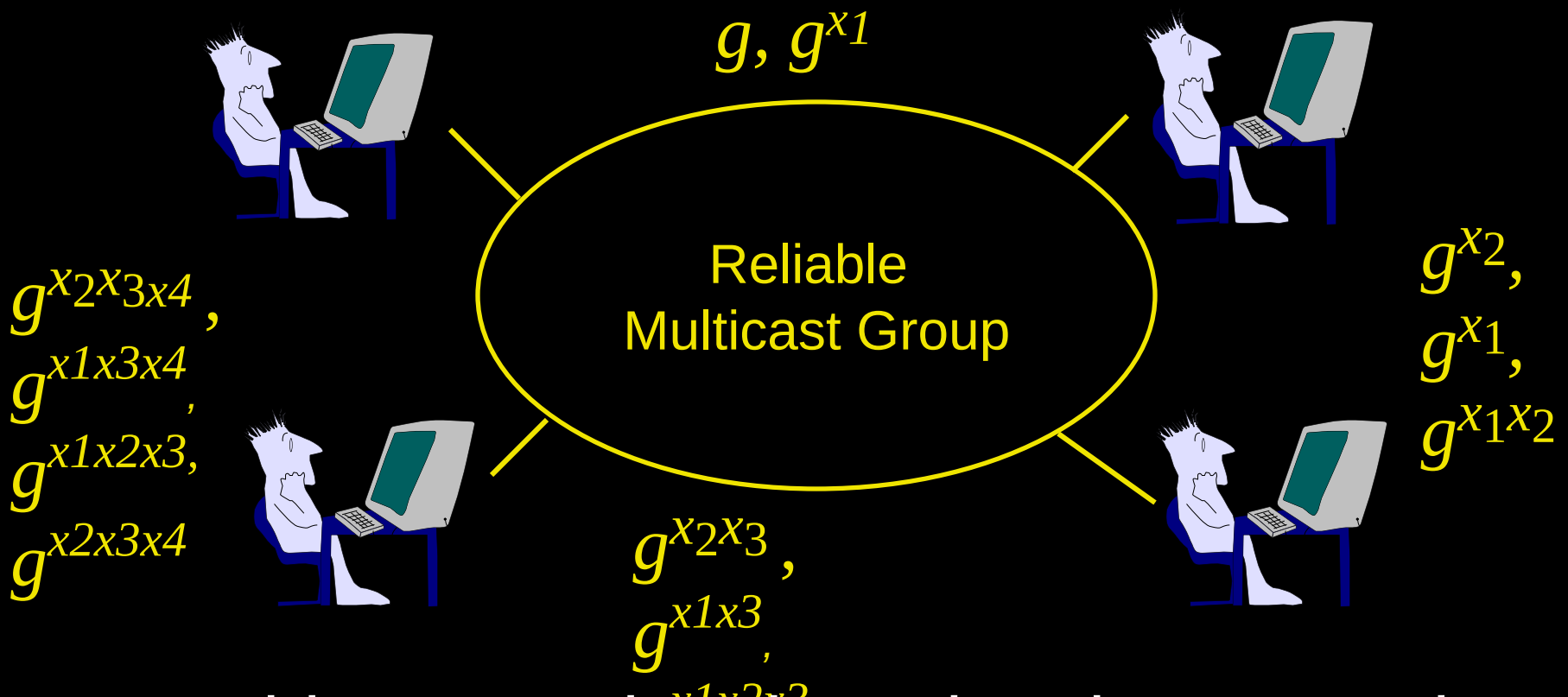


# Python-based Grid Middleware

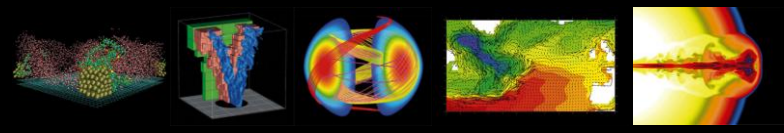
- Grid Services implementation and application interfaces in Python
  - Standalone implementation in python
    - Both client and server side
  - High-level services
  - Application-specific services and infrastructure
- Distributed as part of Globus releases



# Secure and Reliable Group Communication Protocols

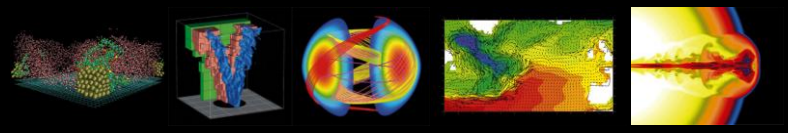



Provides an authenticated and encrypted reliable group channel (similar to SSH/TCP)



# Software development for a specific science group

a few examples



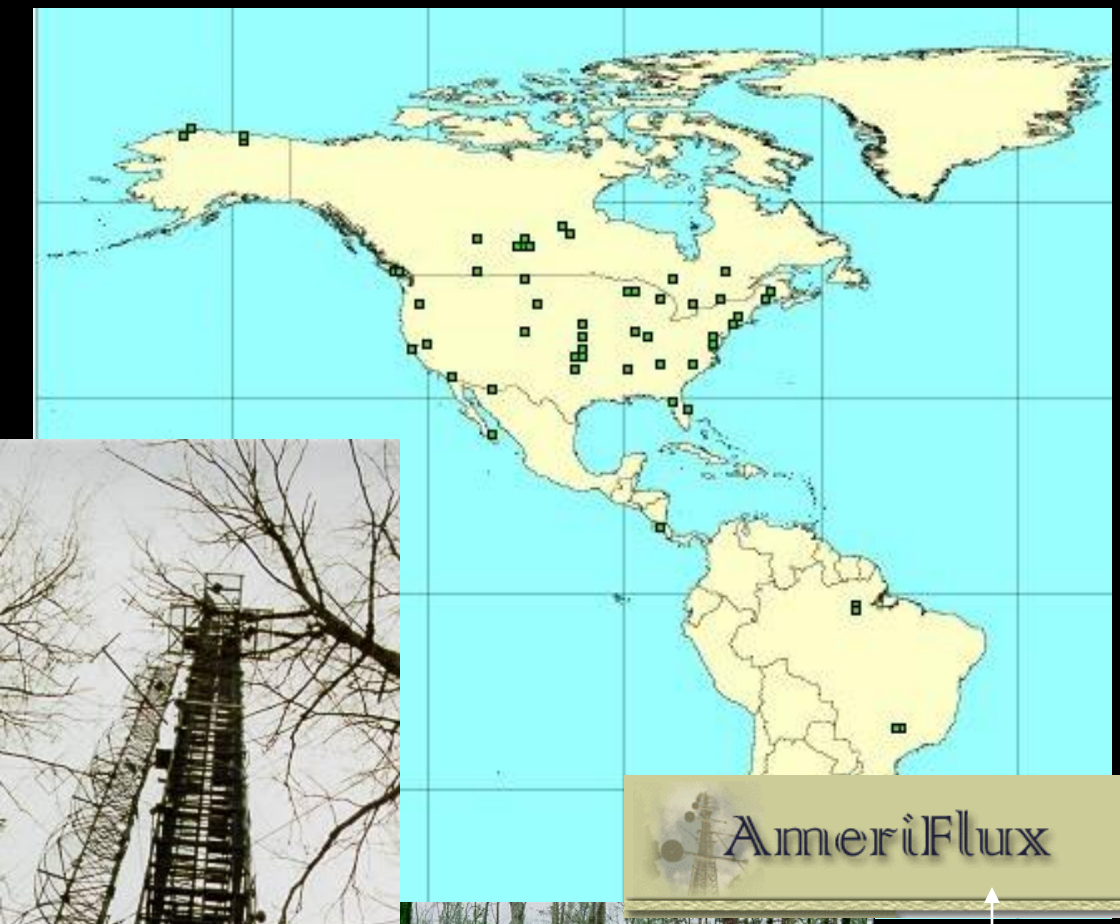
# Berkeley Water Center (BWC): A Water Center of Excellence

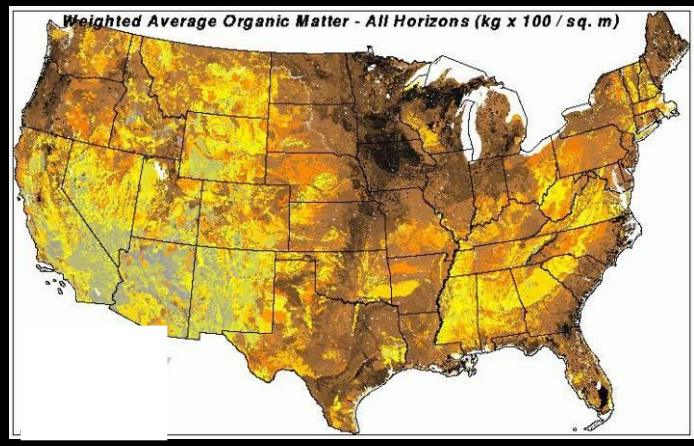
- Is developing a new mode of for doing business at Berkeley by developing a **seamless integration of UCB and LBNL expertise**;

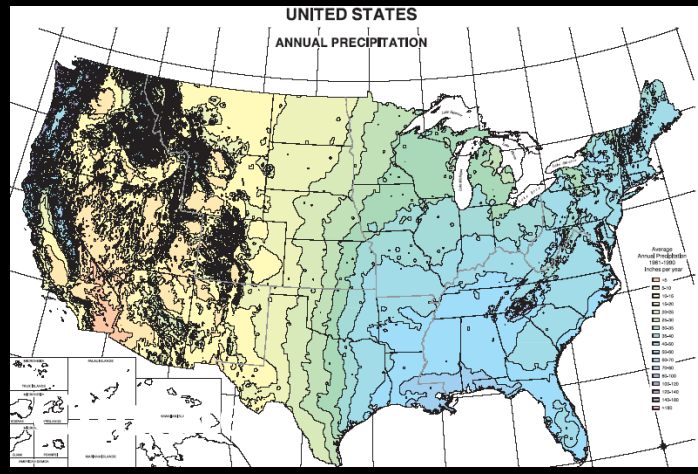
*BWC involves faculty from 3 UCB Colleges and scientists from 3 LBNL Divisions*

- Conducts interdisciplinary investigations that are coordinated through **research thrust areas**;
- Accelerates thrust area results into **applications**;
- Develops **collaborations** between Berkeley water researchers and other expert groups;
- Creates strong, **mutually beneficial partnerships** between Berkeley and other academic, governmental, and private sector institutions;

MALCOLM PIRNIE Independent Environmental Engineers, Scientists and Consultants
VTT
CITRIS Center for Information Technology Research in the Interest of Society
TEKES
northgate
UFZ CENTRE FOR ENVIRONMENTAL RESEARCH LEIPZIG - HALLE IN THE HELMHOLTZ ASSOCIATION
LFR LEVINE•FRICKE LFR Levine•Fricke provides innovative environmental management and consulting services... for today and tomorrow
USGS science for a changing world
BECHTEL

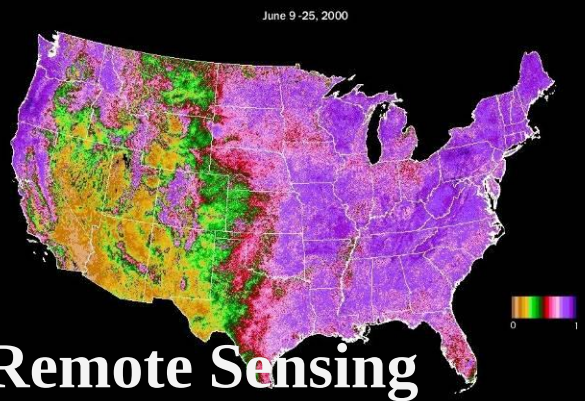





Observatory datasets

Spatially continuous datasets




# Examples of Carbon-Climate Datasets

## Remote Sensing

## Distributed Data Sets

AmeriFlux

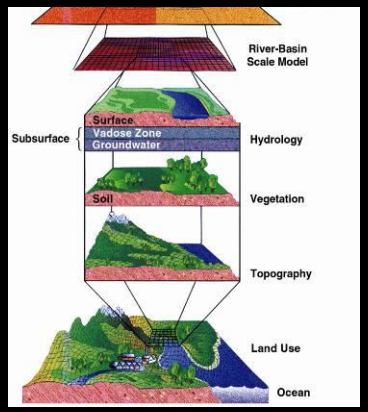
USGS science for a changing world

CIMIS

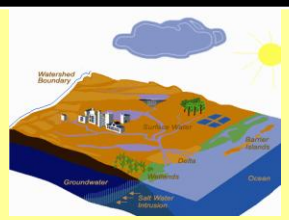
NAS

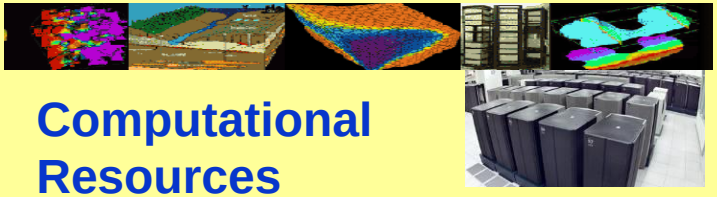
DEPARTMENT OF WATER RESOURCES

## Data Harvesting and Transformations




## Data Cleaning, Models, Analysis Tools



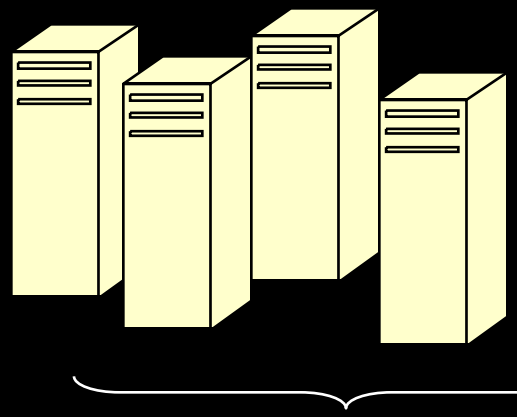



## Computational Resources

## Science Portal

*Building Water Cyberinfrastructure to Connect Data, Resources, and People*

# Carbon-Climate Analysis Environment



**Data Portals**:
Host Ameriflux Climate Data, Statsgo Soils Data, MODIS products

**Tools:**
Statistical
Graphical

Microsoft

Web Service Interface to Data and Tools

Web-based Workbench access

**Choose Ameriflux Area/Transect, Time Range, Data Type**

LAI
Temp
Fpar
Veg Index
Surf Refl
NPP
Albedo

**Import other Datasets**

Climate
Statsgo
MODIS

**Design Workflow**

Data harvest Sites 1-16
Gap Fill, A technique
Gap Fill, B technique
Statistical & graphical analysis
Version control
Canoak Model Site 1
Canoak Model Site 9
Network display LAI
Statistical & Graphical analysis

**Ecology Toolbox**

Data Cleaning Tools

Knowledge Generation Tools

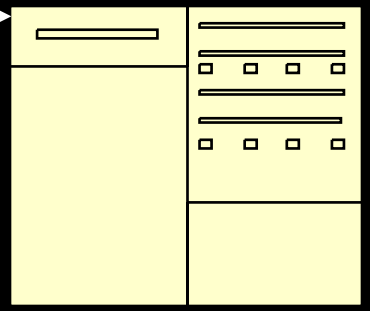
Data Mining and Analysis Tools

Modeling Tools

Visualization Tools

**Carbon-Climate Workbench**

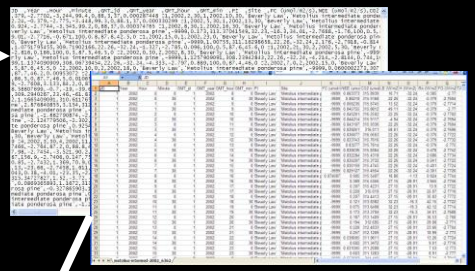
**Compute Resources**



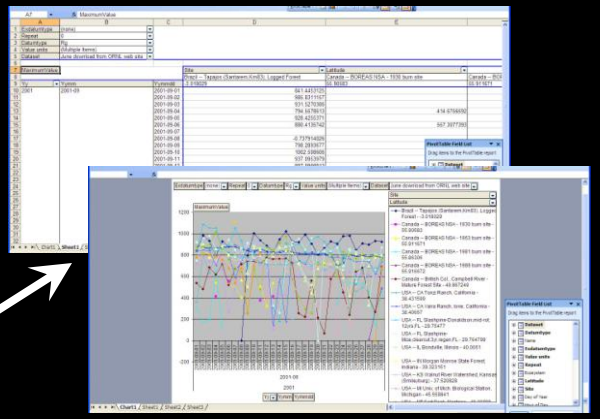
# Data Pipeline

ORNL Ameriflux Site

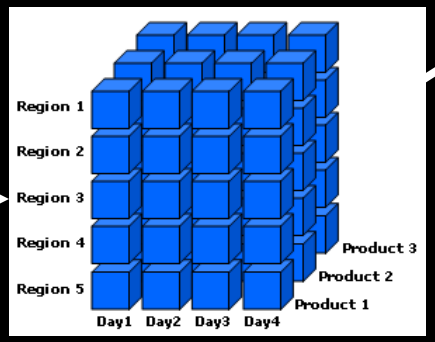
CSV Files



Excel Pivot Table and Chart

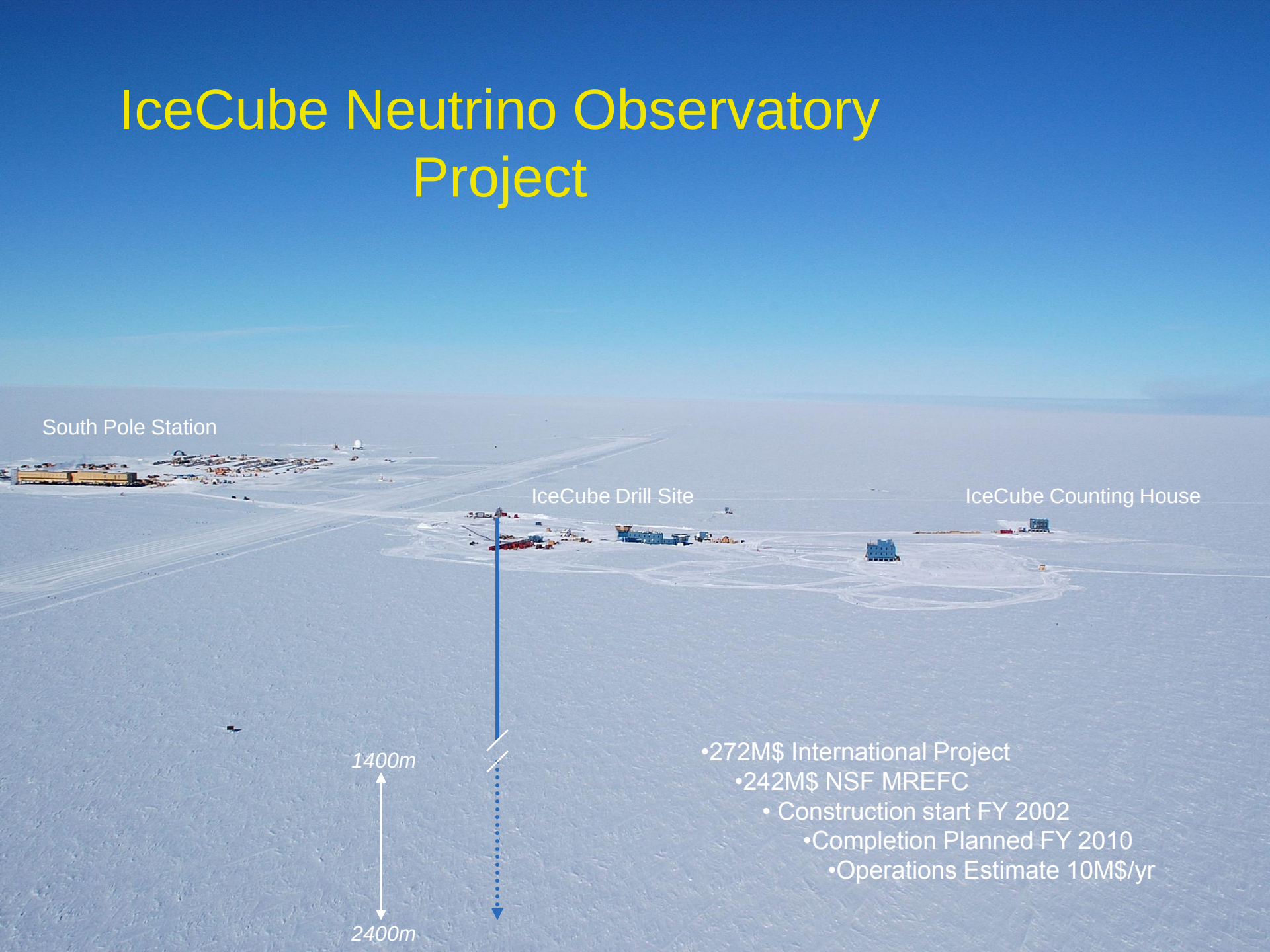


BWC SQL Server Database



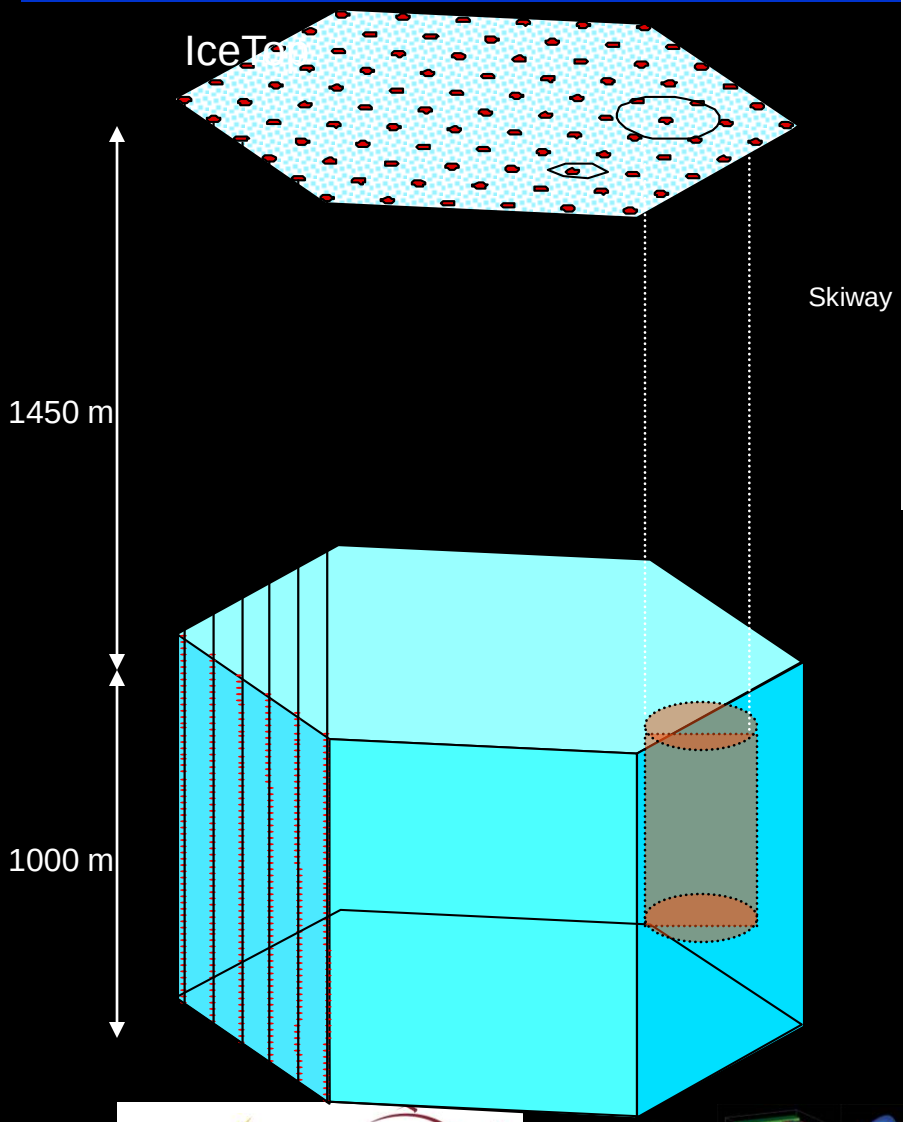

Data Cube

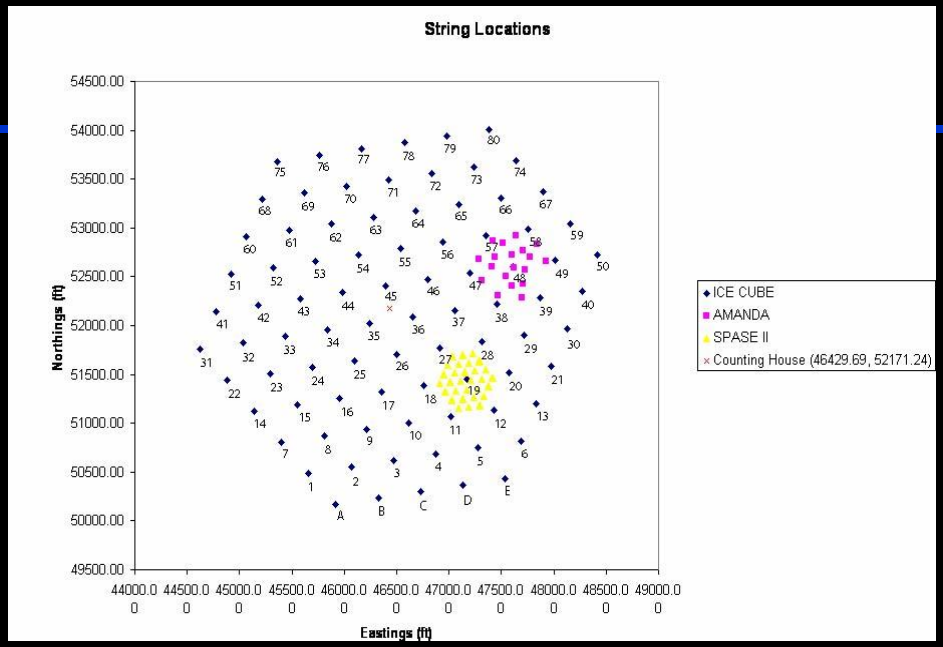
# IceCube Neutrino Observatory Project



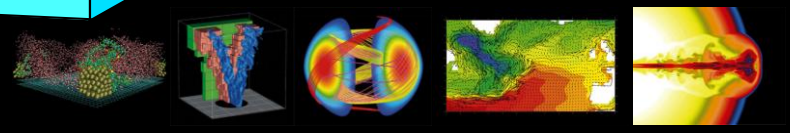

- 272M$ International Project
  - 242M$ NSF MREFC
    - Construction start FY 2002
      - Completion Planned FY 2010
        - Operations Estimate 10M$/yr

# IceCube

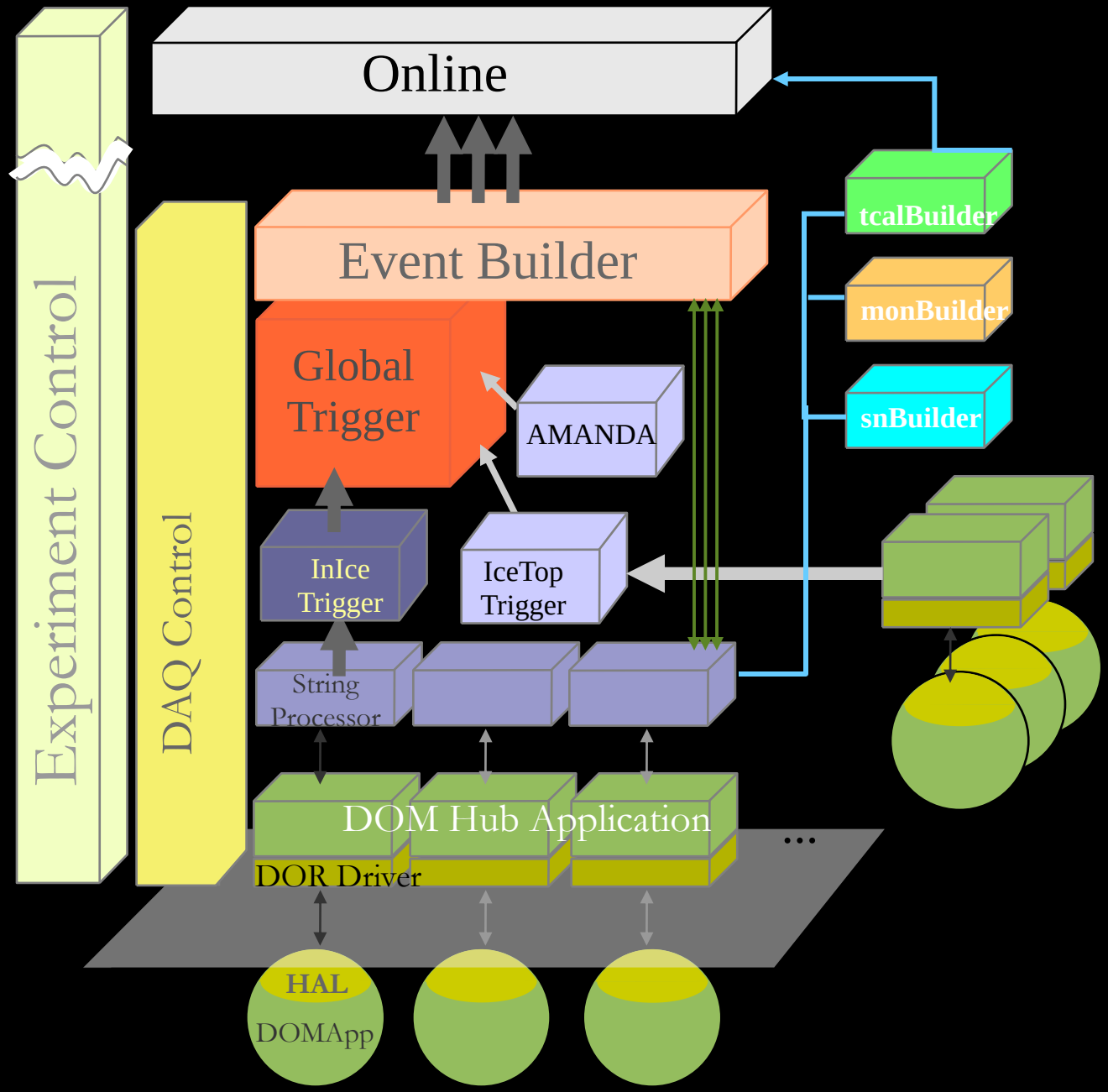



- 70 Strings w/60 Digital Optical Modules per String
- 140 Surface Tanks
- 4200 In-ice DOMs
- ~1 km$^3$ Instrumented Volume
- Depth: 1450 m to 2450 m

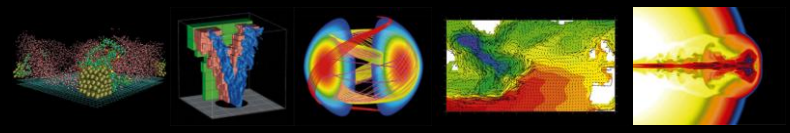


# IceCube DAQ Architecture



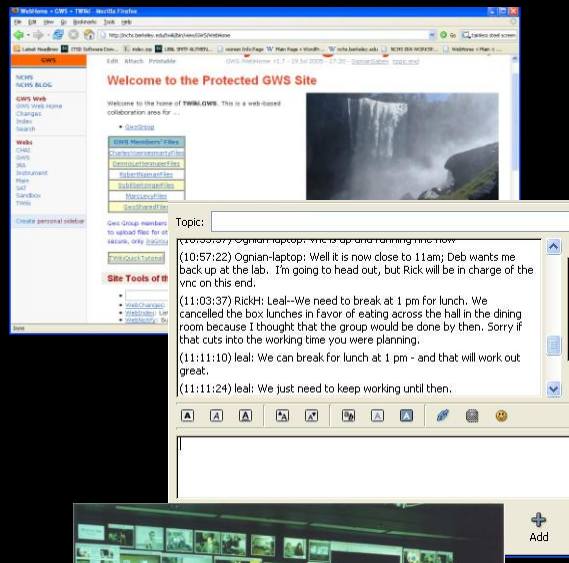
# LBNL IceCube Responsibilities

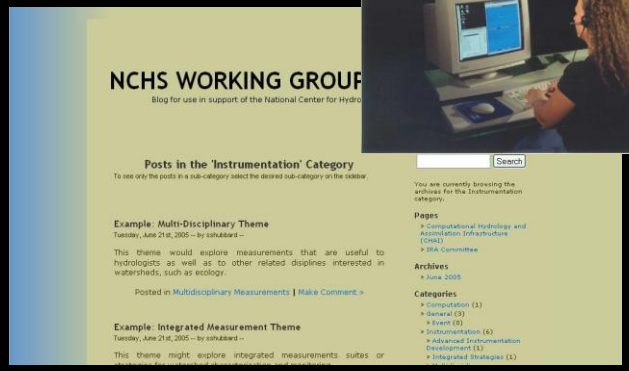
- Overall responsibility for DAQ software system design and implementation.
  - 9 LBNL staff comp. scientists and software engineers over a period of 4-5 yrs.
  - Total of 15 software FTE's across multiple institutions at peak.
- Example specific responsibilities
  - Daily DAQ operations at pole
  - Software release management and testing prior to deployment at pole
  - primary IceCube software code and tool repository (glacier.lbl.gov)
- NSF-funded project. We report to project office (U.W.) and NSF.

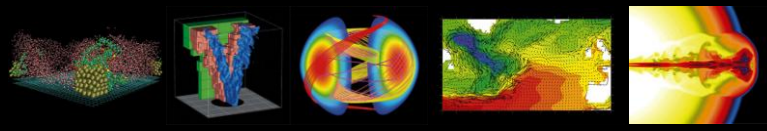


# Collaborative Tools

- Messaging and chat development
- Wiki – supporting the CUAHSI Science Advisory Team
- Access Grid chat library
- Blog – supporting Berkeley Water Center
- Secure group communication protocols

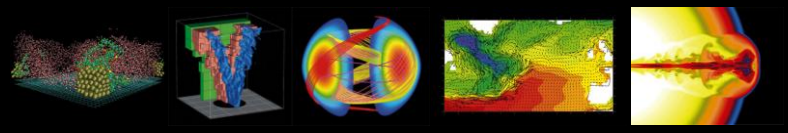




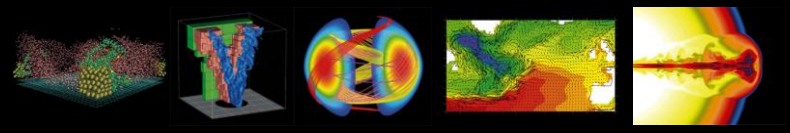
# Large Multi-site and Often Multi-agency Collaborations

A few examples



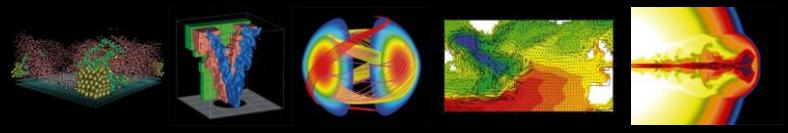
# Security for Open Science – DOE Funding Expected FY07

- Institutions – ANL, ESnet, LBNL, NCSA, NERSC, PNNL, Univ. Delaware, Univ. Virginia, Univ. Wisconsin
- Auditing and forensics
  - Services to enable sites, communities, and application scientists to determine precisely *who* did *what, where* and *when*.
- Dynamic ports in firewalls
  - Services to open and close ports dynamically for applications while enforcing site policy.
- Identity management
  - Services to manage identity and access control across sites and collaborations, and to allow for rapid response to security incidents.
- Secure middleware
  - Services to proactively find and fix software vulnerabilities.



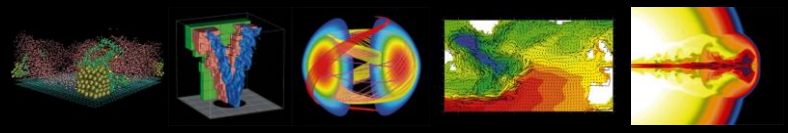
# Open Science Grid – DOE and NSF Funding Expected FY07

- Institutions – ANL, BNL, CalTech, Columbia, Cornell, FNAL, Harvard, Imperial College, Indiana Univ., ISI, LBNL, RCI, SLAC, TACC, Univ. Buffalo, Univ. California SD, Univ. Chicago, Univ. Florida, Univ. Iowa, and Univ. Wisconsin
- Goal – A national distributed computational facility



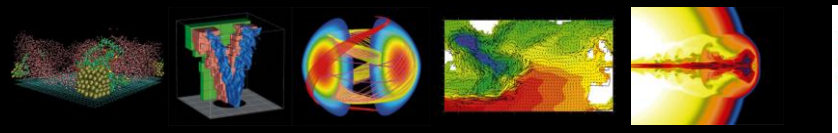
# Center for Enabling Distributed Petascale Science – DOE Funding Expected FY07

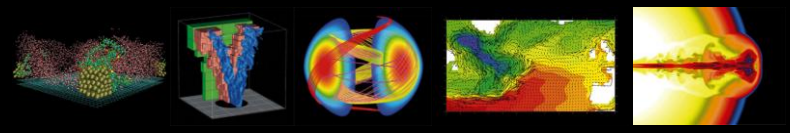
- Institutions – ANL, FNAL, ISI, LBNL, and Univ. Wisconsin
- Data placement – rapid and dependable
- Scalable science services – reliable and high performance processing
- Troubleshooting – effective and timely feedback



# Summary

- Modern science operates in a distributed global environment
- Our goal is to enable scientists addressing complex and large-scale computing and data analysis problems
- We develop, deploy, and support robust software components to enable science in a distributed environment



# Community Service Examples

- Open Science Grid
  - Deputy executive director
  - Operations activity co-chair
- Global Grid Forum
  - Served as chair and members of the Nominating Committee
  - Active member of several working groups and co-chaired several
  - Helped author many GGF published documents.
  - Co-presenter of workshops
- EU DataGrid
  - Architecture task force member
- EU DataTAG / US iVDGL
  - Co-chair of GLUE schema group
- Review panels
  - Members of many NSF proposal review panels
  - Members of the US-LHC Software and Computing review panel
- SC
  - Members of the workshop, tutorials and SC Global committees
  - Many tutorials
- Conferences
  - Co-chair and program committee for many conferences
- DOE Grids Policy Management Authority member