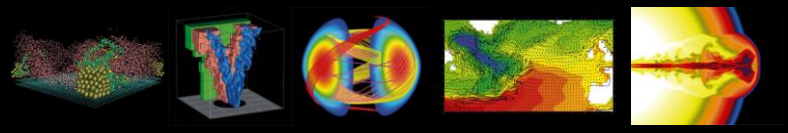
Research and Development of CyberInfrastructure to Support Science

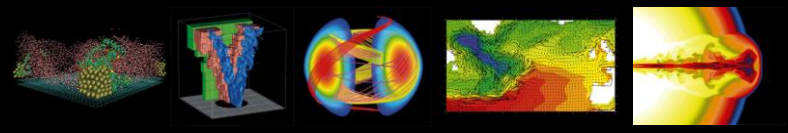
Deb Agarwal
Distributed Systems Department

Lawrence Berkeley National Laboratory

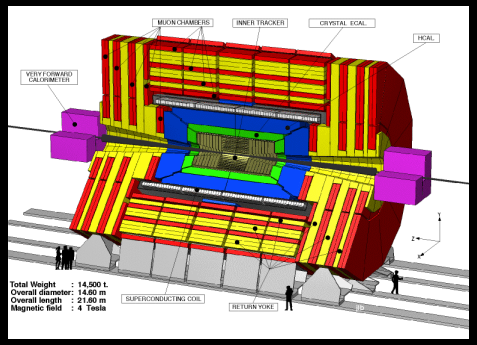


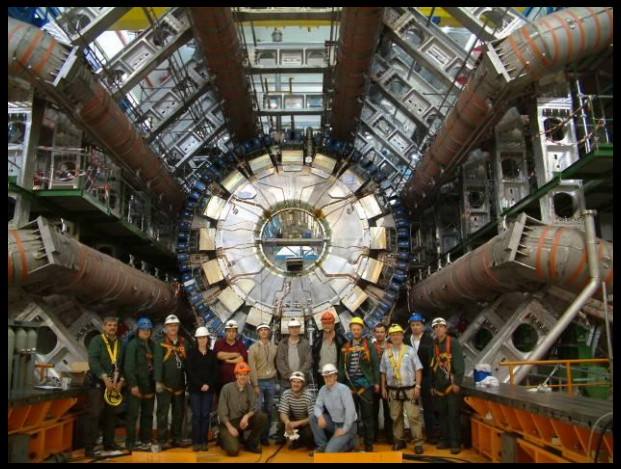
DSD Organization

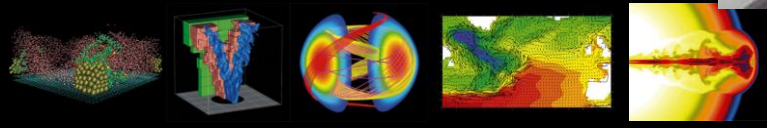
- Collaborative Computing Technologies (Brian Tierney)
- Distributed Secure Grid Infrastructures (Mary Thompson)
- Advanced Application Interface Technologies (Keith Jackson)
- Distributed Scientific Research Tools (Chuck McParland)
- Cybersecurity and Network Research (Deb Agarwal)



Distributed Large-Scale Science





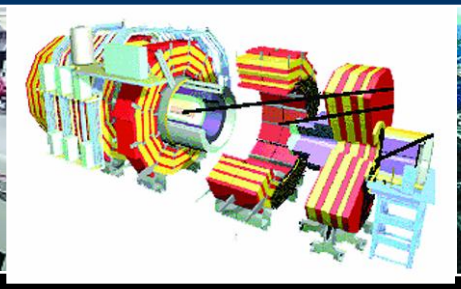
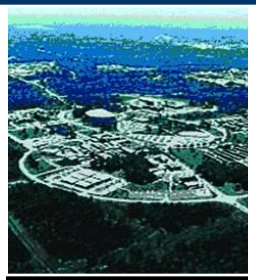
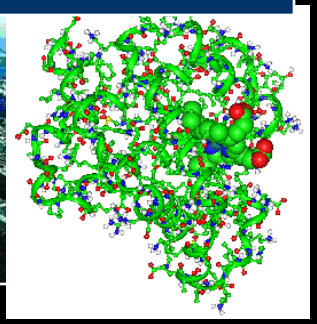


Robust Science Support Framework - Grid

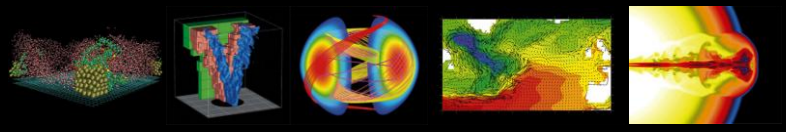
Cybersecurity Protections

Web Services, Portals, Collaboration Tools, Problem Solving Environments

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

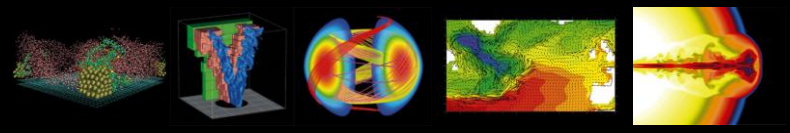
NERSC · Berkeley Lab



Office of Science, U.S. Department of Energy · ESnet Energy Sciences Network

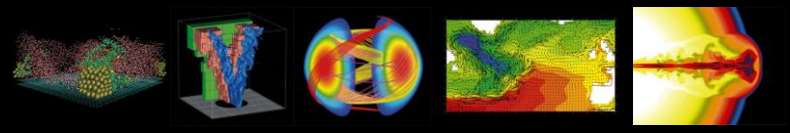
DSD Expertise

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



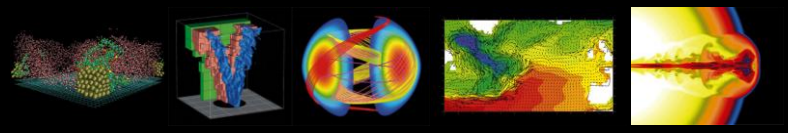
Mix of Projects and Activities

- Research and development of new software-based capabilities and paradigms
  - Authentication and Authorization -- Workflow
  - Network Protocols -- Cybersecurity
  - Collaboration Tools -- Troubleshooting
  - Intrusion detection
  - Grid middleware
- Software development for a specific science or CS group
  - IceCube -- PERC
  - Ameriflux/Berkeley Water Center -- OSCARS
  - Fusion -- Cybersecurity
  - Structural biology
- Contribution to non-DSD Computing Sciences projects and infrastructure
  - PERC
  - OSCARS
  - Ecoinformatics
  - LBL Cybersecurity and networking



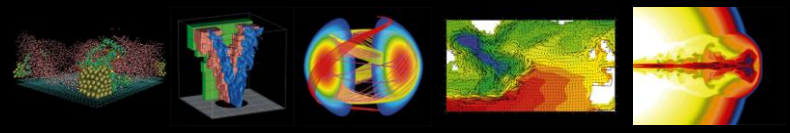
Research and Development of New Software-based Capabilities and Paradigms

A few examples



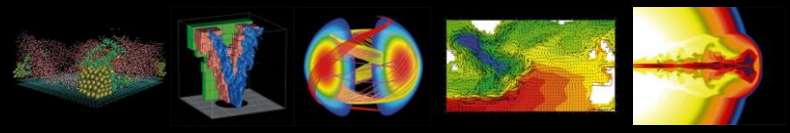
Cryptographic Foundations for Next Generation Distributed Systems - LDRD

- Commercial products for one-time password-authentication do not satisfy the needs of large-scale, data intensive computing applications
- Provide a complete, strong authentication solution for large-scale scientific experiments and the Laboratory
- Leverage open source projects to reduce the time-to-solution
- Scientific contributions



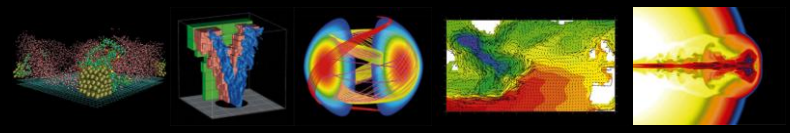
Software Application Infrastructure for Efficiently Managing Large-Scale Computational Biology Experiments - LDRD

- Advanced computational infrastructure that significantly enhances and automates computational biology investigations
- Approach:
  - Develop and establish parameter mesh refinement as a valid technique for parameter sweeps
  - Allow computational biology experiments to run more efficiently and reliably, enabling scale-up to current problem sizes.
  - Link the lab's distributed computing expertise to the lab's genome sequence data capabilities to develop improved information and knowledge tools and reduced time to solution
  - Through close coordination with the Brenner lab, demonstrate the capabilities of this approach on real problems and provide a proof-of-concept that can be leveraged into future funding from NIH and DOE.



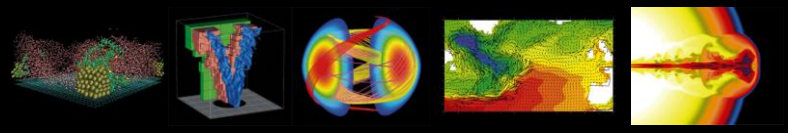
On-demand Overlays for Scientific Applications - LDRD

- Research infrastructure-integrated overlay networks as a means of fully exploiting the high speed and redundant capabilities of the physical network for bulk data transfer and data dissemination needs of scientific applications.
- Two major research threads
  - Investigate methods of instantiating virtual overlays on the physical ESnet network
  - Research and develop overlay network algorithms tuned to application requirements and underlying network characteristics



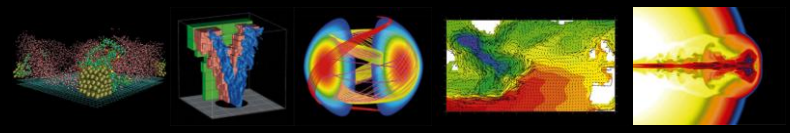
KDD - Visual Workflow Tools to Support Collaborative Hypothesis Generation and Evidence Chaining

- Application of workflow tools to data mining
- Enable hypothesis generation
- Track and record queries that lead to a conclusion
- Enable complex queries across databases



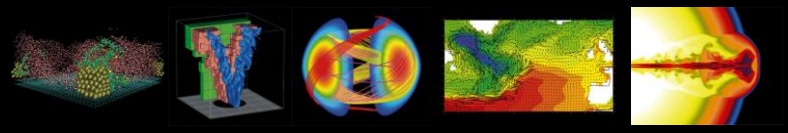
Security for Open Science – DOE Funding Expected FY07

- 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.



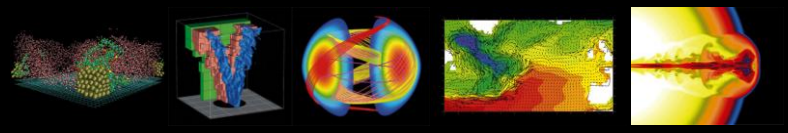
Center for Enabling Distributed Petascale Science – DOE Funding Expected FY07

- 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



Software development for a specific science or CS group

a few examples



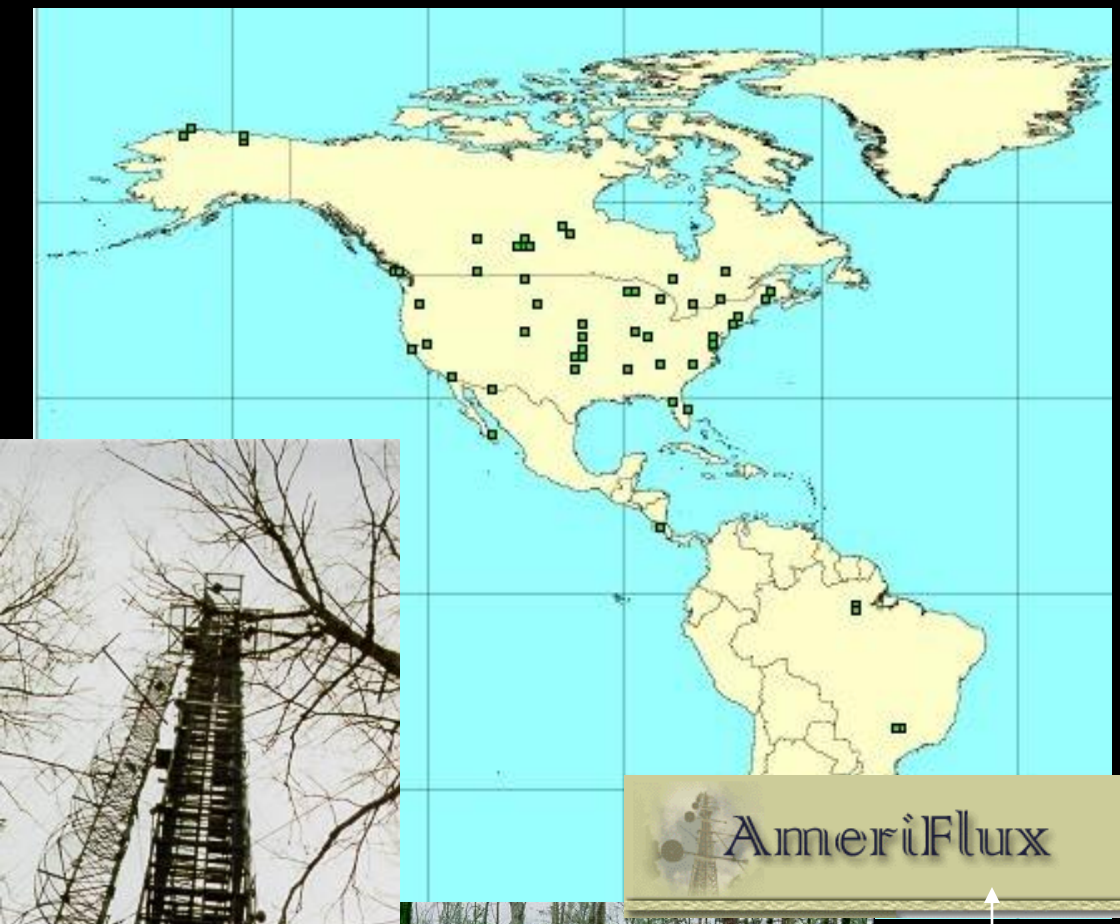
Berkeley Water Center (BWC): A Water Center of Excellence

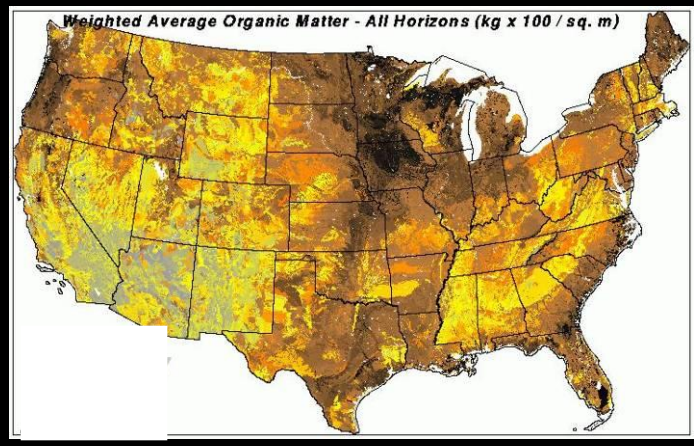
- Is developing a new mode of for doing business at Berkeley by developing a **seamless integration of UCB and LBNL expertise**;

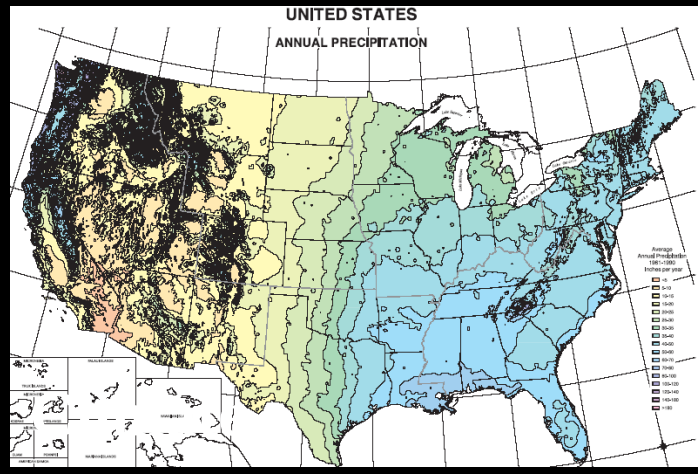
BWC involves faculty from 3 UCB Colleges and scientists from 3 LBNL Divisions

- 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

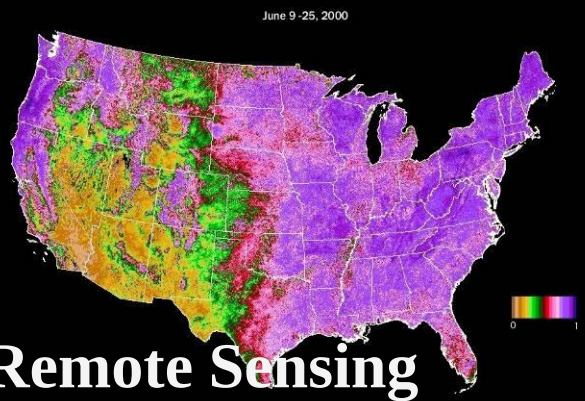






Observatory datasets

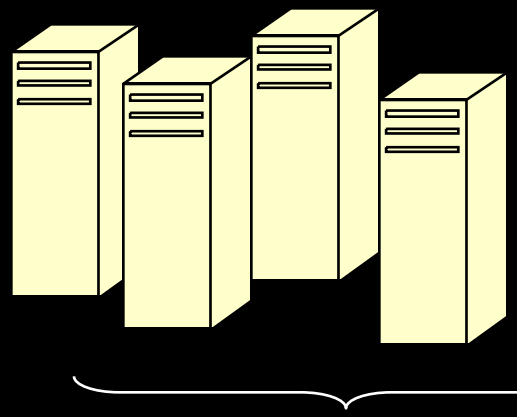
Spatially continuous datasets



Examples of Carbon-Climate Datasets

Remote Sensing

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

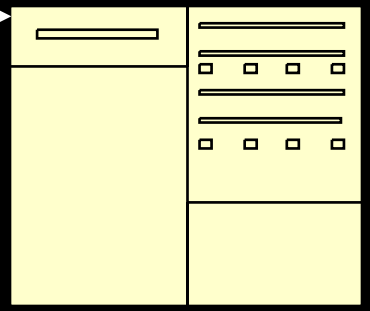
Data Mining and Analysis Tools

Modeling Tools

Visualization Tools

Carbon-Climate Workbench

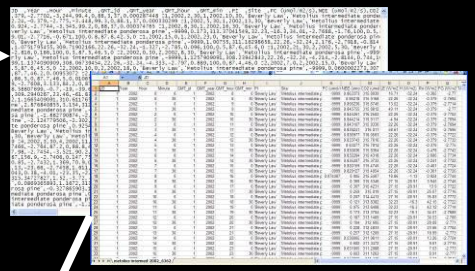
Compute Resources



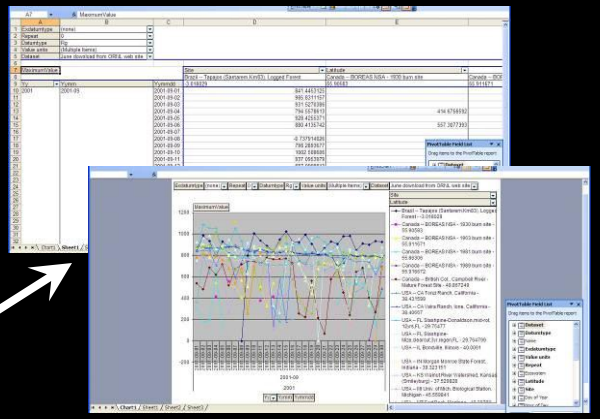
Data Pipeline

ORNL Ameriflux Site

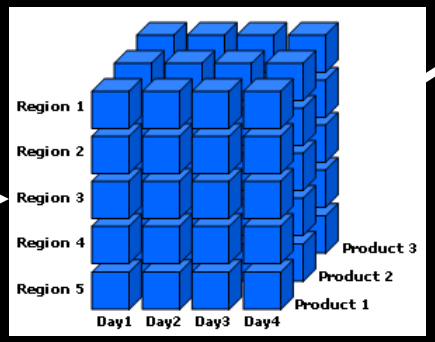
CSV Files



Excel Pivot Table and Chart

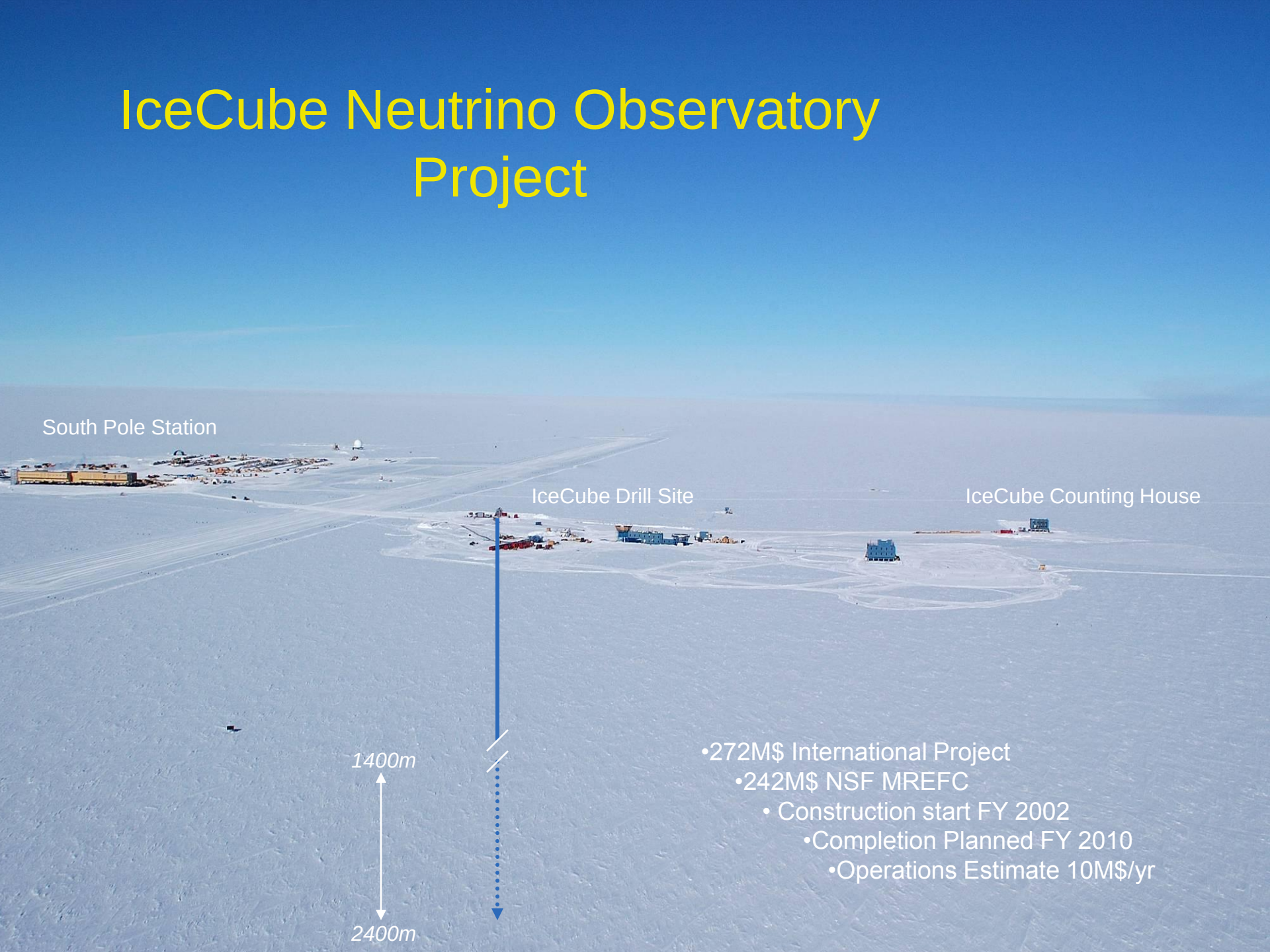


BWC SQL Server Database



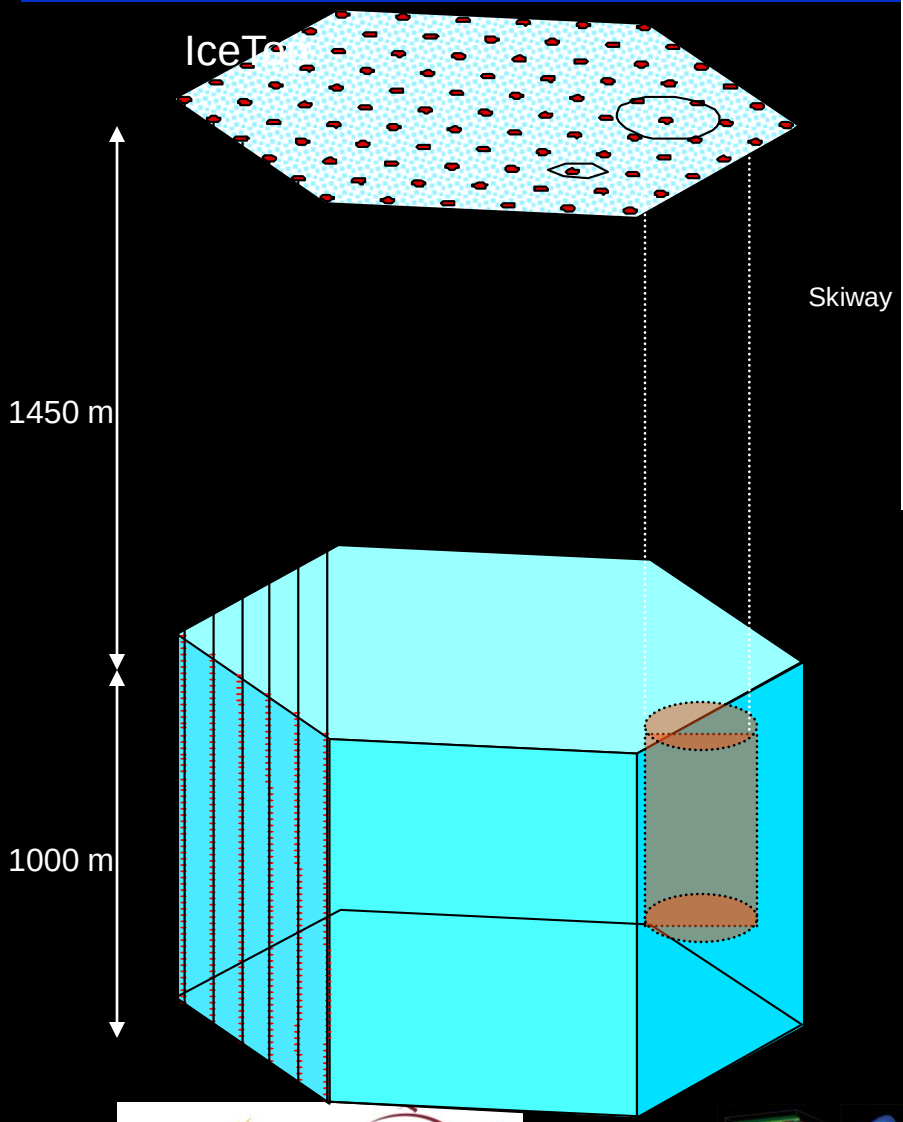
Data Cube

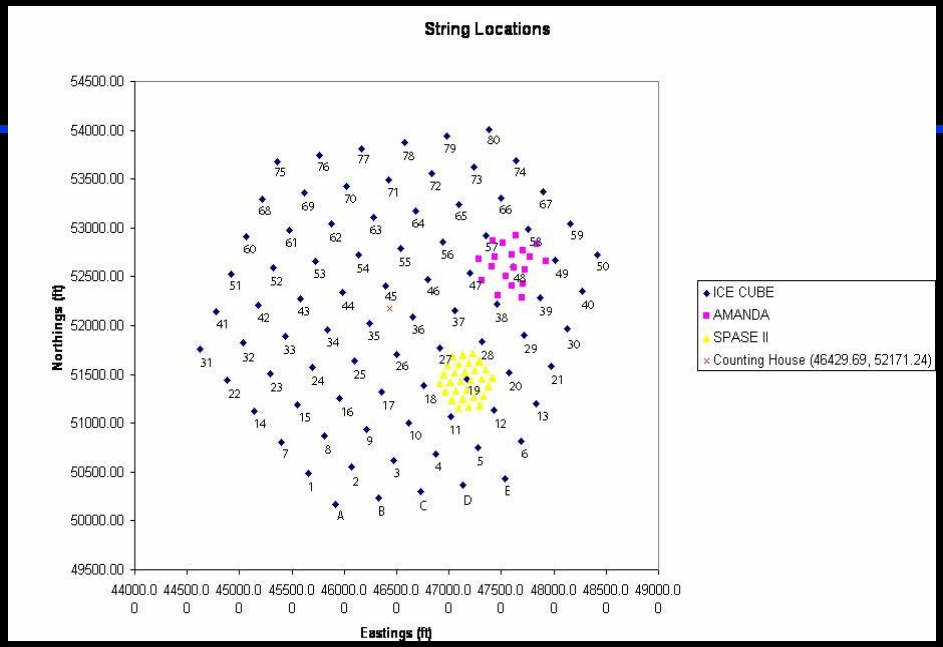
IceCube Neutrino Observatory Project



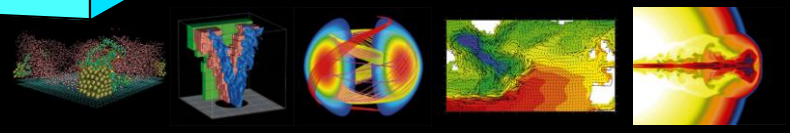
- 272M$ International Project
- 242M$ NSF MREFC
- Construction start FY 2002
- Completion Planned FY 2010
- Operations Estimate 10M$/yr

IceCube

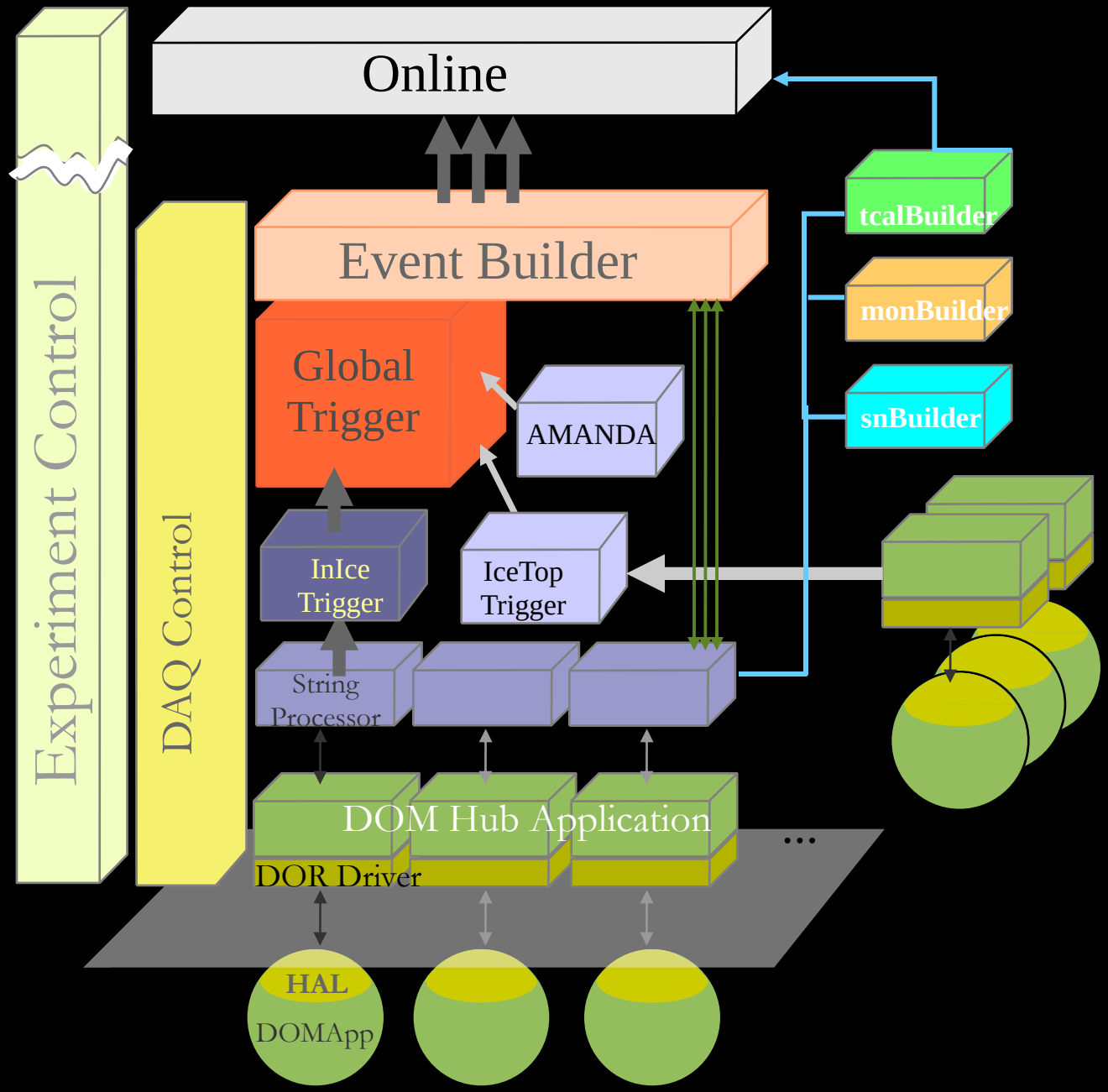




- 70 Strings w/60 Digital Optical Modules per String
- 140 Surface Tanks
- 4200 In-ice DOMs
- ~1 km^3 Instrumented Volume
- Depth: 1450 m to 2450 m

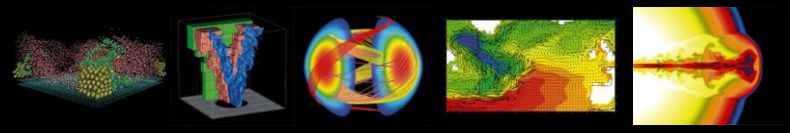


IceCube DAQ Architecture



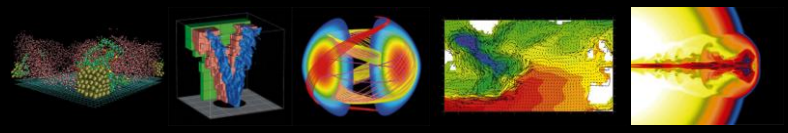
LBNL IceCube Responsibilities

- 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.



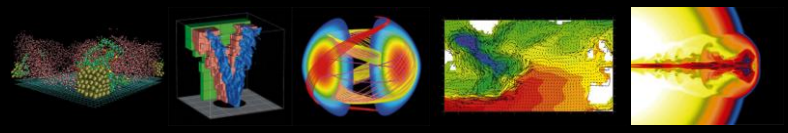
Open Science Grid – DOE and NSF Funding Expected FY07

- 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

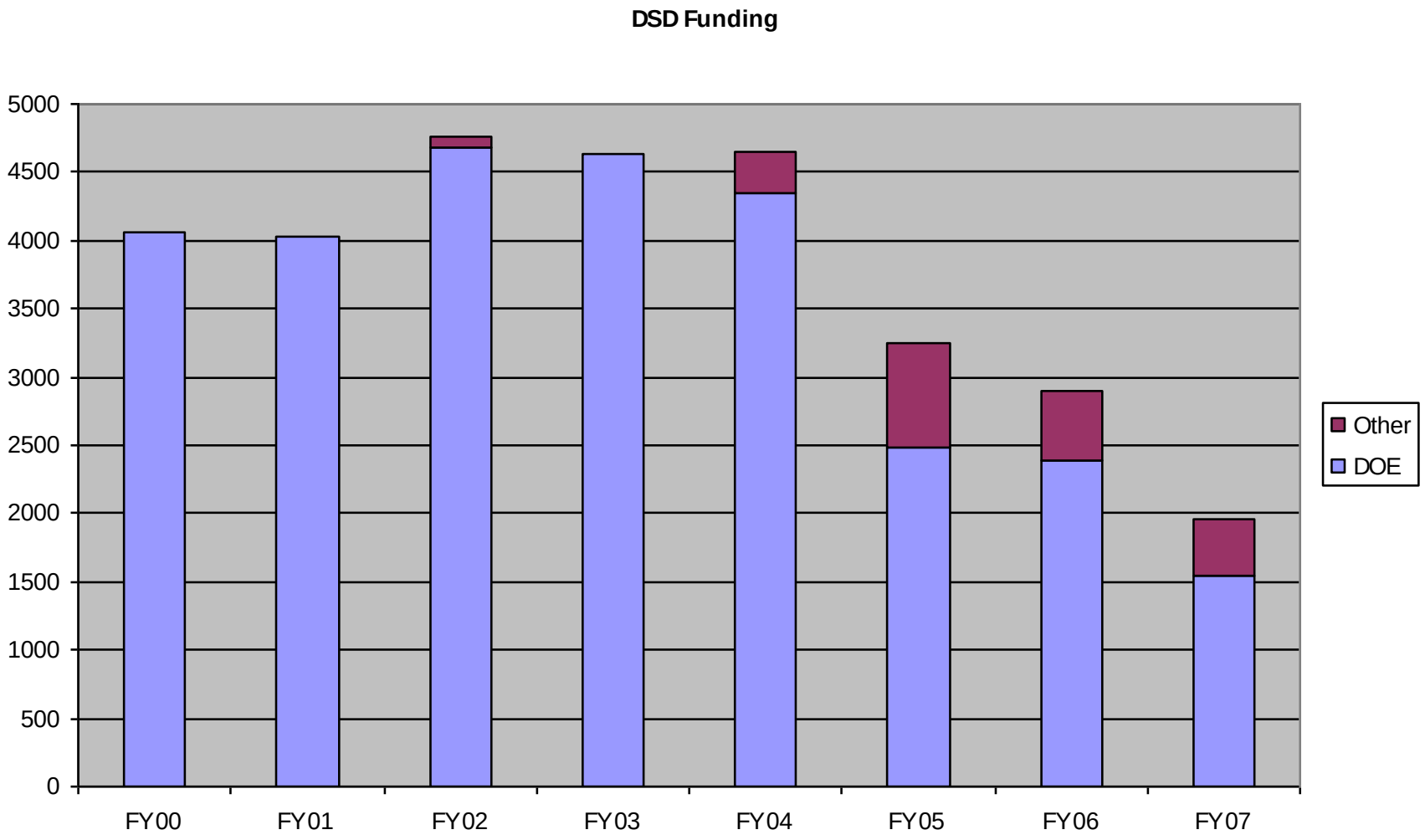


Support of Other CS and IT Departments and Divisions

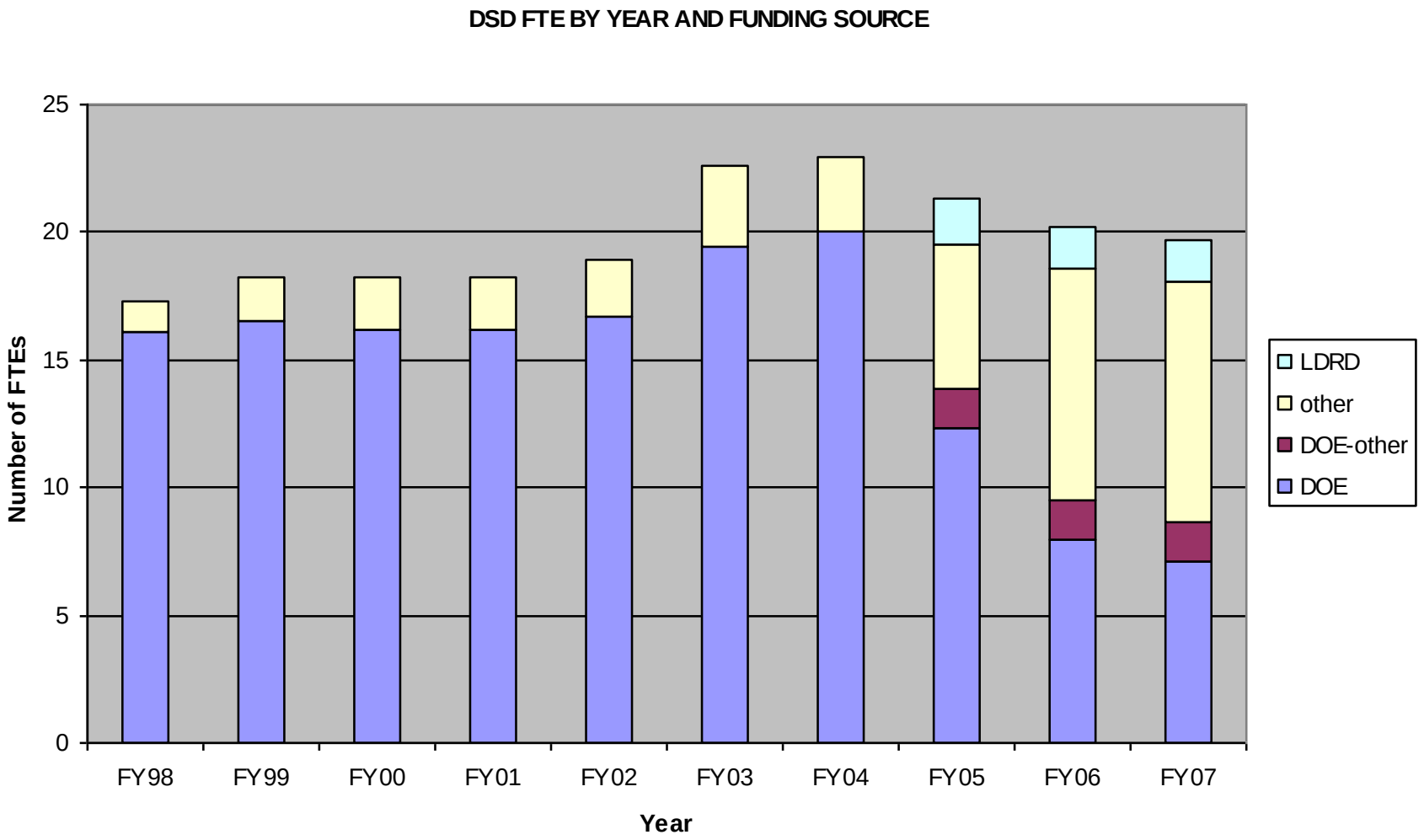
- ESNet - OSCARS (~1.5 FTE)
- HPCRD – PERC (~0.5 FTE)
- Cybersecurity operations (~1 FTE)
- HPCRD – Ontologies (~0.5 FTE)



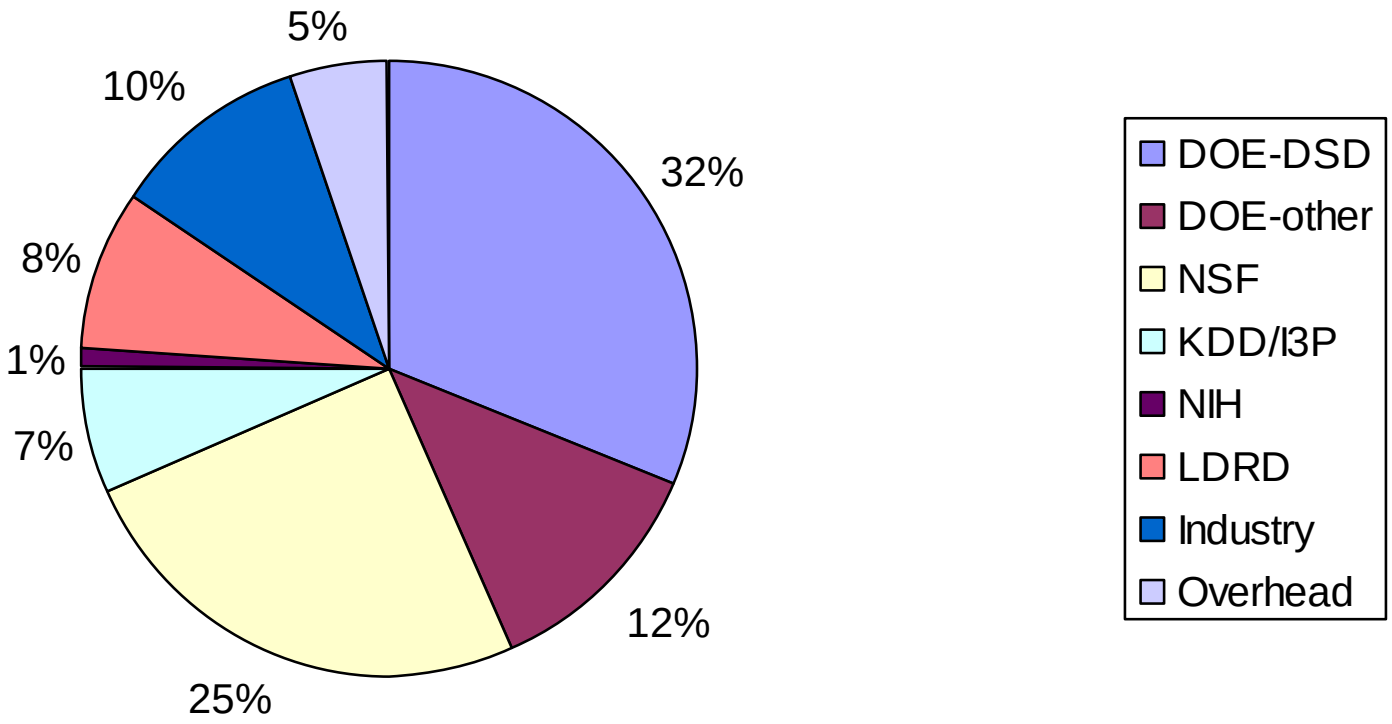
Historical DSD Department Funding (only the money in DSD accounts)



Historical DSD Department FTE

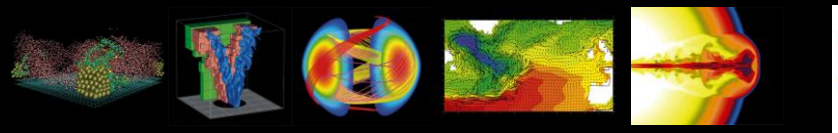


DSD FY07 Mix of Funding Sources



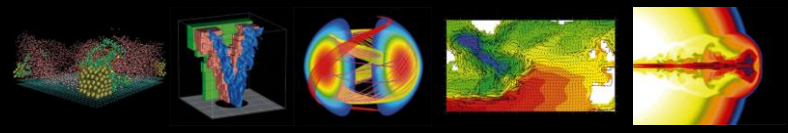
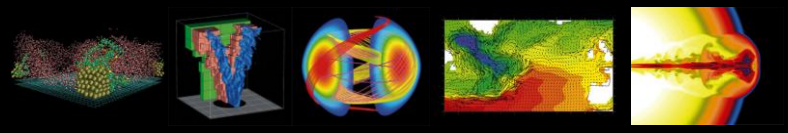
DSD in a Nutshell

- Modern science operates in a distributed global environment
- Our goal is to help 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



DSD Challenges for the Next Five Years

- Get back to the point where we have a moderate number of focused research areas
- Maintain a strong team of career personnel without “base program” funding
- Increase research funding to achieve a balanced mix of research effort, matrixed personnel, and applied effort
- Increase our DOE funding level to closer to ~75%
- Staff development and re-invigoration



The End