

Cyberinfrastructure

9th Team Meeting
Ft Collins, CO
August, 2010

John Helly



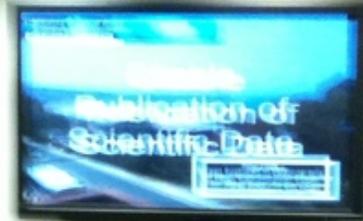
CIWG Objectives



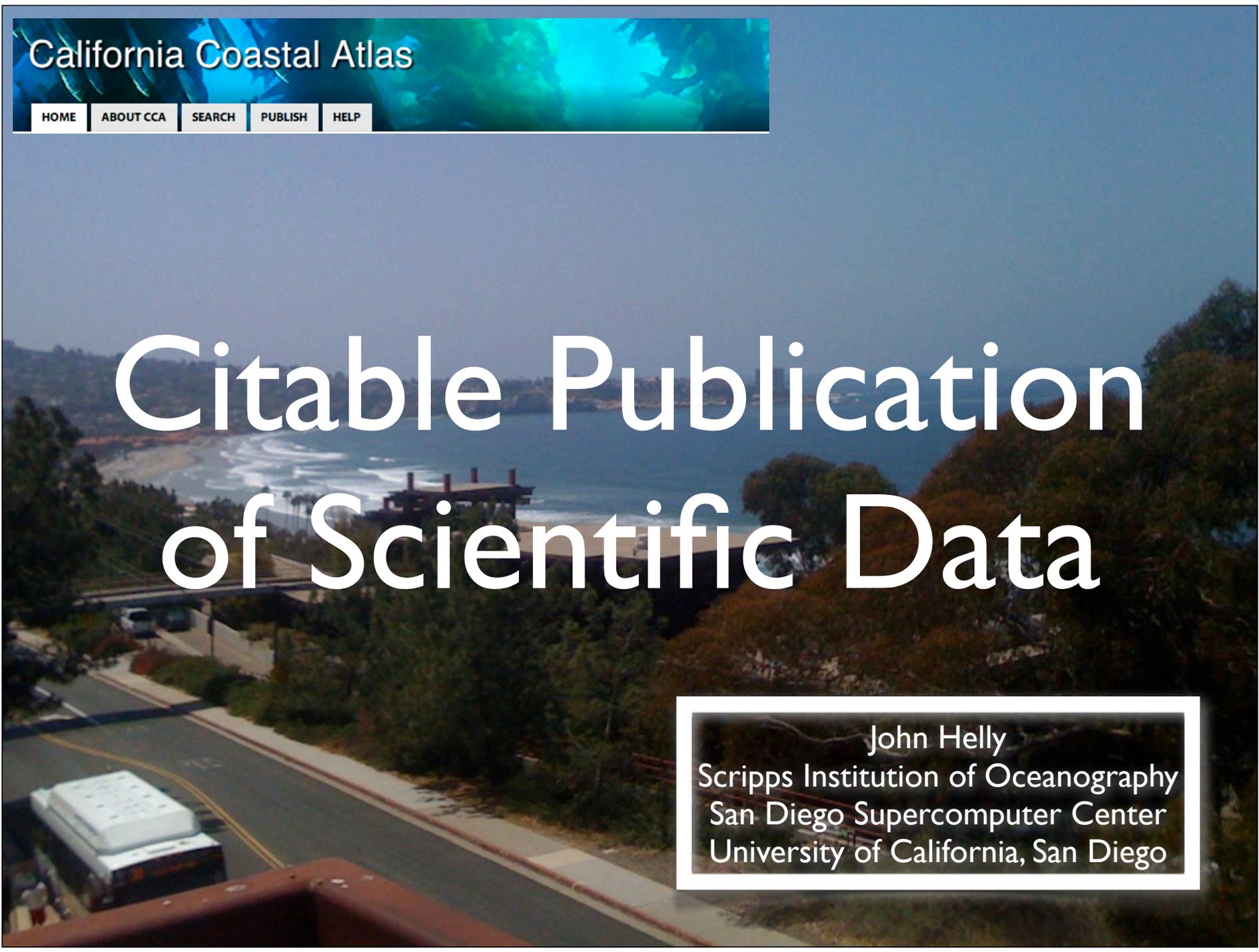
- Make efficient use of computing and data resources
- acquire resources
- coordinate resource utilization
- collaborate to leverage joint efforts
- Provide technology look-ahead
- Validate goals and provide advice and consent to Executive Committee



Optiputer@SDSC



April 2010
Talk to OSTP @ National Archives
Citable Publication of Scientific Data



Citable Publication of Scientific Data

John Helly
Scripps Institution of Oceanography
San Diego Supercomputer Center
University of California, San Diego

R: Analysis and Plotting

Click to **LOOK INSIDE!**

Use **AI**

Hadley Wickham

ggplot2

GSP's Guide to netCDF and R

$[Y|X, \theta][X|\theta][\theta]$

UCAR | NCAR | **Statistics** | Software | Data | Publications | Projects

CISL | IMAGE | Statistics | Contact Us | Visit Us | People Search

GSP's guide to netCDF format data and the 'R' package 'ncdf'.

netCDF is a common, self-describing, portable binary format for geophysical data. GSP made an executive decision earlier this year (i.e. Tim and Doug talked after lunch) to use this format as much as possible when creating or manipulating data sets. For the statistical readership we should note that there are contributed packages for R that allow for the efficient reading and writing of netCDF files and part of the intent of this web page is to provide some simple [examples](#) to get users started. Some advantages of this format are:

- The netCDF libraries to create and access files for many (all?) architectures are free and available through UNIDATA.
- A netCDF file not only contains the "data" but also a description of the variables, the creation history, and any other important attributes about the data set.
- A netCDF file is a reasonably compact **portable** binary format. I.e. You can make one on a 'supercomputer' and read it on a PC.
- The netCDF interface can extract parts of a large data file without having to read the entire record. Since many netCDF files are 0~ Gigabytes, this is important.
- netCDF is a standard format not only for geophysical observational data but also for numerical model output, such as the NCAR community climate system model.
- A contributed package in R, **ncdf** is free, readily available, and simplifies the interface to the otherwise gory low-level routines available in the Unidata library.

A brief description of netCDF

With regards to netCDF, a little philosophy goes a long way. A netCDF file is intended to provide all the information needed to interpret the data, as well as the data itself. The information about the length of each dimension, how many dimensions, the units of the quantity, etc. are all contained in a netCDF file. They are intended to be **self-describing**. The netCDF format is flexible enough to allow for a tremendous variety of incarnations. Some netCDF files contain the data for (a set of) radiosonde instruments, some represent the output of a General Circulation Model (GCM), or a set of observations taken at a weather station, for example. All three cases will be explored in the following discussion.

The netCDF file can be broken down into logical parts. To that end, lets take a look at the **header** of a very simple netCDF file.

```
netCDF example {
```

The R Project for Statistical Computing

http://www.r-project.org/

The R Project for Statistical Computing

PCA 5 vars
principle1 = data.ccr = ccr

Clustering 4 groups

Factor 1 [41%]

Factor 3 [19%]

Getting Started:

- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).
- If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

News:

RGL - 3D Real-Time Visualization Device System for R

http://rgl.neoscientists.org/download.shtml

RGL 3D Real-Time Visualization Device System for R

News
About
Download
Docs
Gallery
User
Developer

Get it from **CRAN**

rgl package details at <http://cran.r-project.org/src/contrib/Descriptions/rgl.html>

Get it from our local **Archive**

Checkout the **latest revision**

See **Developer** section for details.

1. *NSF Architecture Planning Proposal
 1. Projected experiments, requirements
 2. Participants, node partners, workshops/locations
 3. Data transport, post-processing, visualization
2. *New Teragrid Computing Award
 1. Hugh Morrison's Experiments and Plans
 2. Community account for MMF runs
3. *Subversion MMF Development Repository (Mark B.)
4. Update on current CMMAP CI Architecture (Roadmap)
 1. Digital Library (transition to iRODS)
 2. *New iRODS web-browser and parallel data transfer service
 1. 'How to store your data in the CMMAP capacity at SDSC and get it back on your desktop whenever you want it.'
 3. *Data transpose project on Dash (SDSC Flash Memory Machine)
5. *New MMF (SP-CAM) Community Account interface through CMMAP Digital Library using GFDL FRE system (V. Balaji Group)

Proposal to the National Science Foundation for

**Community Infrastructure Planning:
Cyber-Infrastructure for the Cloud-Climate Community**

Prepared in response to
CISE Computing Research Infrastructure (CRI) Program Solicitation 08-570

Principal Investigator:

David A. Randall

Department of Atmospheric Science
Colorado State University
Fort Collins, Colorado 80523

tel: (970) 491-8474

email: randall@atmos.colostate.edu

Co-Investigators:

John Helly

San Diego Supercomputer Center and Scripps Institution of Oceanography
University of California, San Diego
10100 Hopkins Drive
La Jolla, CA 92093-0505

Kelley Wittmeyer

Department of Atmospheric Science
Colorado State University
Fort Collins, Colorado 80523

Michelle Strout and Sangmi Pallickara

Department of Computer Science
Colorado State University
Fort Collins, Colorado 80523

NSF Planning Proposal

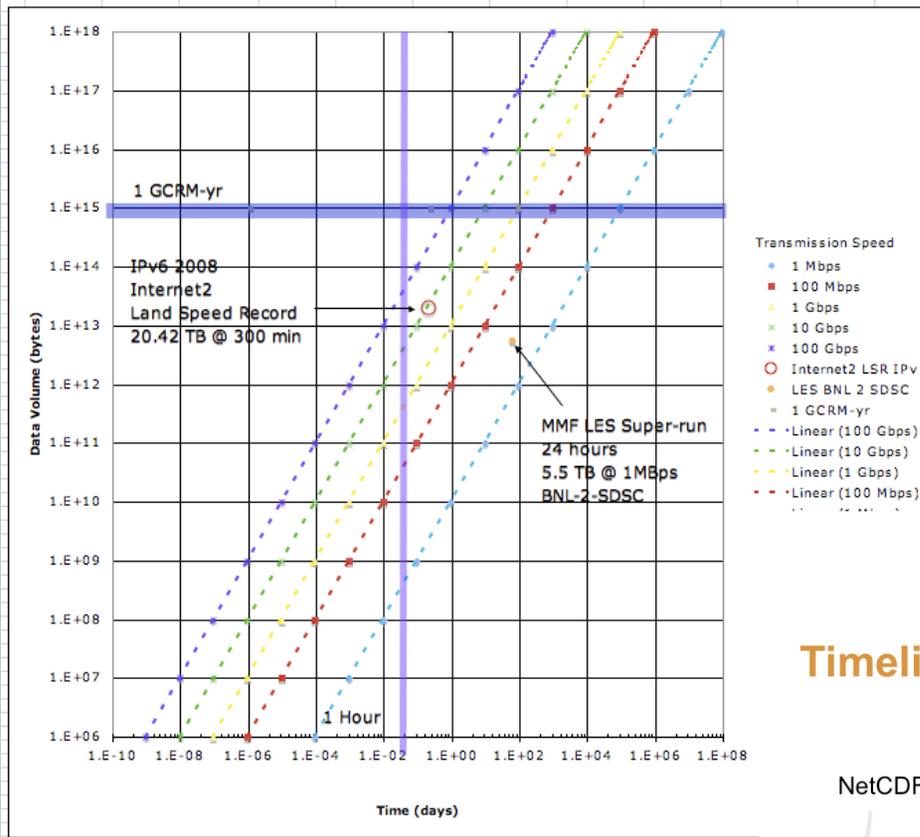
GCRM Data Problem

The GCRMs and Giga-LES models are conceptually complex, but in addition they pose problems that are technical, practical, and fiscal, rather than conceptual in nature. This is where the need for new infrastructure arises. Our proposed infrastructure project relates to data management, analysis, and visualization:

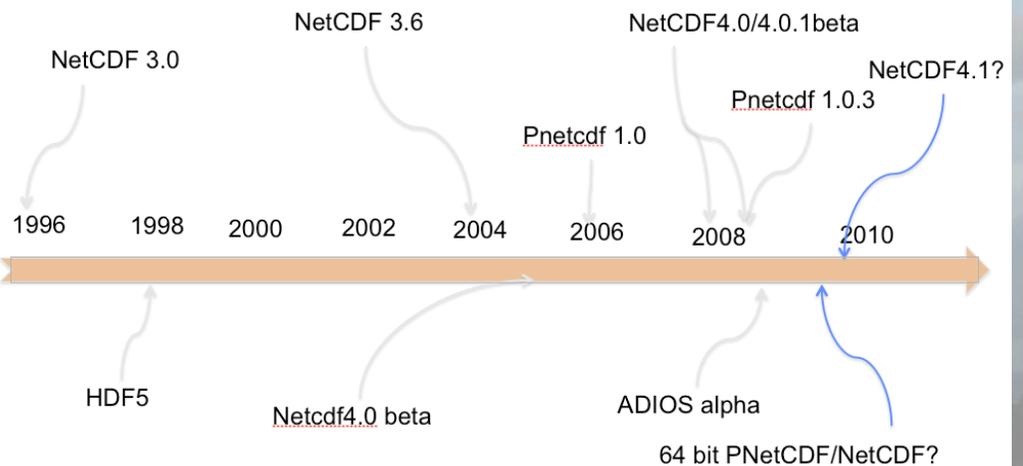
- GCRMs produce terabytes to petabytes of model output. The data is created at supercomputer centers. It must be archived, curated, and made available to users at remote sites.
- Many difficult choices must be made; for example, choosing which fields to output, and what subsetting spatial and temporal resolutions to save, are complex.
- Routinely saving global model output with high temporal resolution is not practical.
- A possible strategy is to save regional model output (for one or more selected regions) with high temporal resolution, and full spatial resolution, and global model output with lower temporal resolution and perhaps even reduced spatial resolution.
- Extraction of useful information from GCRM output is complicated by the sheer volume of data produced, the wide range of scales represented, and the diverse phenomena included. New methods are needed for comparison of model output with a variety of observations, including satellite data.
- New methods are needed for the efficient and effective visualization of GCRM results. The range of scales is so large that “zooming” capabilities are essential. New approaches are needed to visualize and analyze the time evolution of complex three-dimensional structures (such as large rotating convective clouds) that are associated with multiple interacting fields, including vector fields.

In short, *the very large models used in cloud-climate studies must be supported by a suitably designed infrastructure for data management, analysis, and visualization.* These needs are community wide and should be addressed in a coordinated fashion that serves the community as a whole.

Internet Data Transfer Capacity



Timeline



Parallel Input/Output Technology Progress

4. Planning process

Our goal is to submit the CI proposal in the summer of 2011.

Planning will begin at CMMAP's Team Meeting in August 2010, the same week that this planning proposal will be submitted to NSF. We will also take advantage of CMMAP's Team Meeting in January 2011, in Berkeley, California, where we systematically collect ideas from the CMMAP team, and also engage the local talent at NERSC, including the developers of ViSIT. The planning activity will be a major component of the January 2011 meeting.

The tasks to be completed before the submission of the CI proposal include:

- Scoping the hardware and software systems, including the determination of expected life-cycle cost to operate over its useful lifetime.
- Developing a plan for maintaining and enhancing the proposed infrastructure up to and beyond the sunset of the planned CI grant.
- Outreach to solicit input from the CMMAP community and the national HPCC centers on the design of the hardware and software components and external and internal interfaces of the proposed infrastructure.
- Soliciting input from the CMMAP community on the design of the education and outreach activities that will be associated with the proposed infrastructure.
- Developing a site plan for the hardware, in cooperation with the CSU administration, including the university's Facilities office.

13

- Developing an operations plan, which must take into account the education and outreach activities. Initiation of a planning process to determine hardware and software configuration and a likely data loading model to size the system against

5. Management Plan

- Many difficult choices must be made; for example, choosing which fields to output, and what subsetted spatial and temporal resolutions to save, are complex.
- CMMAP standard?

Computing Resources



New Allocations

[Home](#)
[My TeraGrid](#)
[Resources](#)
[User Forums](#)
[Documentation](#)
[Training](#)
[Consulting](#)
[Allocations](#)

[Allocations/Usage](#)
[Accounts](#)
[Profile](#)
[Tickets](#)
[Registered DNs](#)
[Change Portal Password](#)
[Add/Remove User](#)
[Community Account](#)
[SSH Terminal](#)
[Citation Info](#)

Allocations/Usage

Projects

[Show Inactive Projects](#) | [Show Expired Allocations](#)

Modeling Global Climate Variability with the Multi-scale Modeling Framework New parameterizations of Cloud Micro-physics and Developing Community Accounts
 Portal for Running the MMF

[Show Project Details](#)

Allocations*

Start Date	End Date	Resource	SUs Remaining	SUs Awarded	My Usage (SU)	% Remaining	Alloc. Type	State
2010-07-01	2011-06-30	abe-queenbee-steele-lonestar.teragrid	2,307,000	2,307,000	0.0	100 %	new	active
Show Users on abe-queenbee-steele-lonestar.teragrid								
2010-07-01	2011-06-30	NCSA Tape	5	5	0.0	100 %	new	active
Show Users on NCSA Tape								

Data Transposition Development for Exa-scale Data in Memory

[Show Project Details](#)

Allocations*

Start Date	End Date	Resource	SUs Remaining	SUs Awarded	My Usage (SU)	% Remaining	Alloc. Type	State
2010-05-04	2010-11-07	dash.sdsc.teragrid	29,970	30,000	30.0	100 %	supplement	active
Show Users on dash.sdsc.teragrid								
2009-11-07	2010-11-07	Spur	30,000	30,000	0.0	100 %	new	active
Show Users on Spur								

Regionalization of Anthropogenic Climate Change Simulations

[Show Project Details](#)

Allocations*

Start Date	End Date	Resource	SUs Remaining	SUs Awarded	My Usage (SU)	% Remaining	Alloc. Type	State
2009-04-01	2010-09-30	Ranger	43,186	3,950,000	0.0	1 %	new	active
2009-04-01	2010-09-30	Spur	397	500	103.0	79 %	new	active

Modeling Global Climate Variability with the Multi-scale Modeling Framework: The Boundary-layer Cloud Problem

[Show Project Details](#)

Allocations*

Start Date	End Date	Resource	SUs Remaining	SUs Awarded	My Usage (SU)	% Remaining	Alloc. Type	State
2009-04-01	2010-09-30	Steele	103,900	950,000	0.0	11 %	new	active
Show Users on Steele								

Subversion Repository

The screenshot displays a Mac OS X desktop with three Subversion client windows open. The top menu bar includes 'Grab', 'File', 'Edit', 'Capture', 'Window', and 'Help'. The system tray shows various icons and the date 'Thu 7:12'.

Repositories Window: Lists several repositories with their names and URLs. The 'CMMAP' repository is highlighted.

Name	Url
Neptune	/Users/hellyj/Active/svn_repos
Geospatial	file:///Volumes/Geospatial001/svn_repos
CESM V1.0	https://svn-ccsm-release.cgd.ucar.edu/model_versions/cesm1_0/
CMMAP	https://svn.sdsc.edu/repo/cmmap

Working Copies Window: Lists working copies with their names and paths. The 'DLF' working copy is highlighted.

Name	Path
DLF	/Users/hellyj/Active/svn_working/DLF
Geospatial	/Users/hellyj/svn_work/IcebergIII

CMMAP Window: Shows the Subversion client interface for the CMMAP repository. The URL is `https://svn.sdsc.edu/repo/cmmap/`. It features a search log table and a file browser view.

Search Log Table:

Rev #	Date	Author	Log message
52	08/02/10 15:23:15	u8753	JNR 02 Aug 2010 4page version install file
51	08/02/10 15:22:35	u8753	JNR Aug 02 4 page version
50	07/29/10 19:27:25	hellyj	First import
49	07/29/10 19:20:39	hellyj	Deleting DrupalModules
48	07/29/10 18:47:13	hellyj	Modules developed to integrate modeling servi
47	07/28/10 15:07:12	mbranson	Additions and fixes to allow the SPCAM to out
46	06/02/10 15:28:20	mbranson	Uncomment out the previously commented out li

File Browser View: Shows the directory structure of the repository at Revision 52. The 'trunk' directory is selected.

```
root
├── CAM
├── DrupalModules
├── SAM
├── SPCAM
├── branches
├── tags
├── trunk
├── form_spcam3
├── querym
└── querymv2
```

Leveraging National & Partner Resources

	Organization	Resource	2007	2008	2009	2010
Data Allocations	San Diego Supercomputer Center (SDSC)	Disk	15 Terabytes	15 Terabytes	30 Terabytes	45 Terabytes
		BlueGene			30,000 SUs*	
		Triton				30,000 SUs
Computing Allocations	Teragrid (multi-institution)	SDSC DataStar (IBM SP4)	600,000 SUs	1,200,000 SUs		
		Grid Roaming			600,000 SUs	2,703,000 SUs
		LSU Steele			900,000 SUs	
	Lawrence Berkeley National Laboratory (LBNL)	National Energy Research Scientific Computing Center (NERSC)			700,000 SUs	
	Oak Ridge National Laboratory (ORNL)	Cray XT			2,000,000 hrs	3,000,000 hrs
	National Center for Atmospheric research (NCAR)	Bluelce IBM Power5			500,000 SUs	
	IBM Watson Research Center	BGW - eServer Blue Gene Solution			TBD	
	Stonybrook				TBD	

CMMAP Architecture Roadmap

