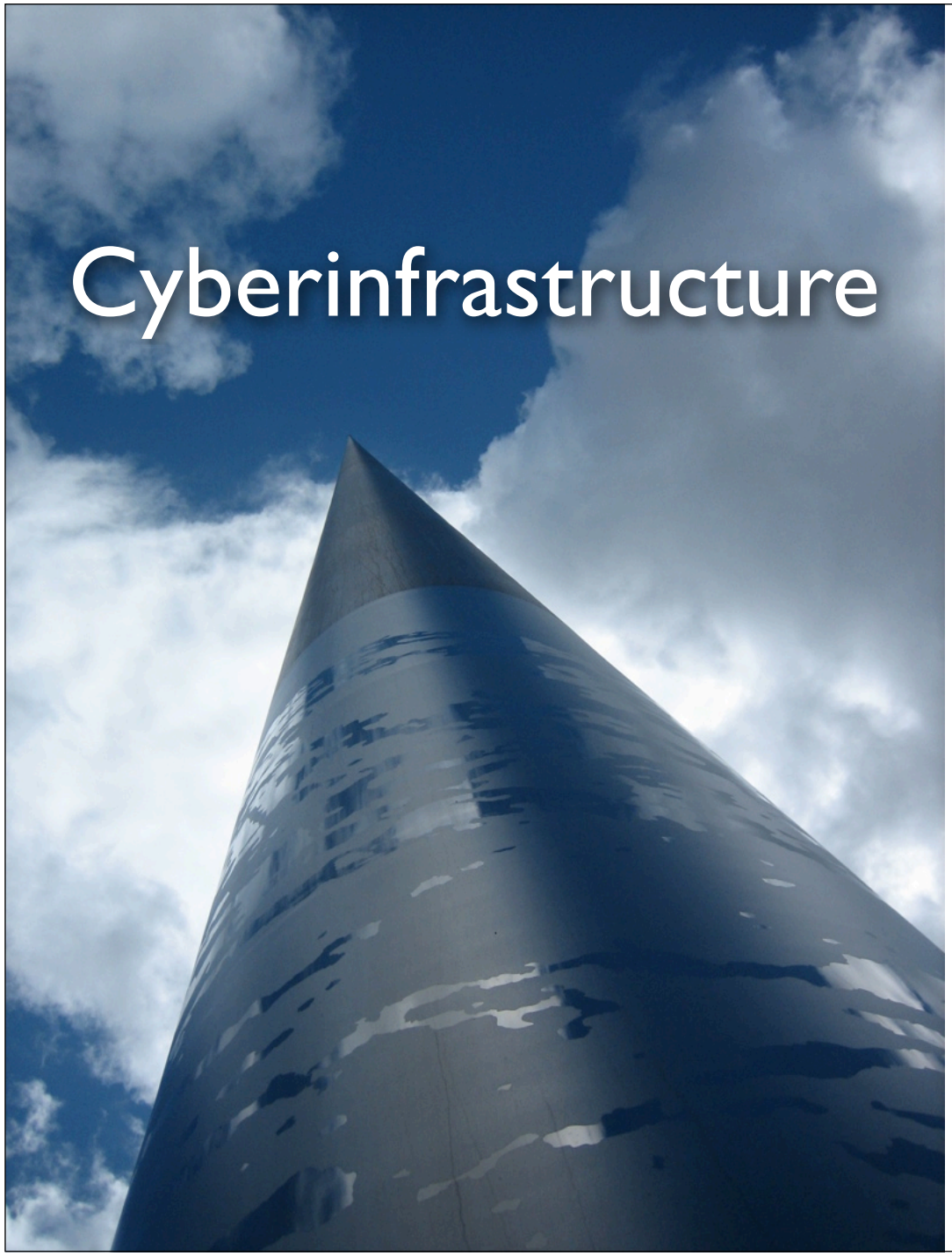


Cyberinfrastructure

9th Team Meeting
Ft Collins, CO
August, 2010

John Helly

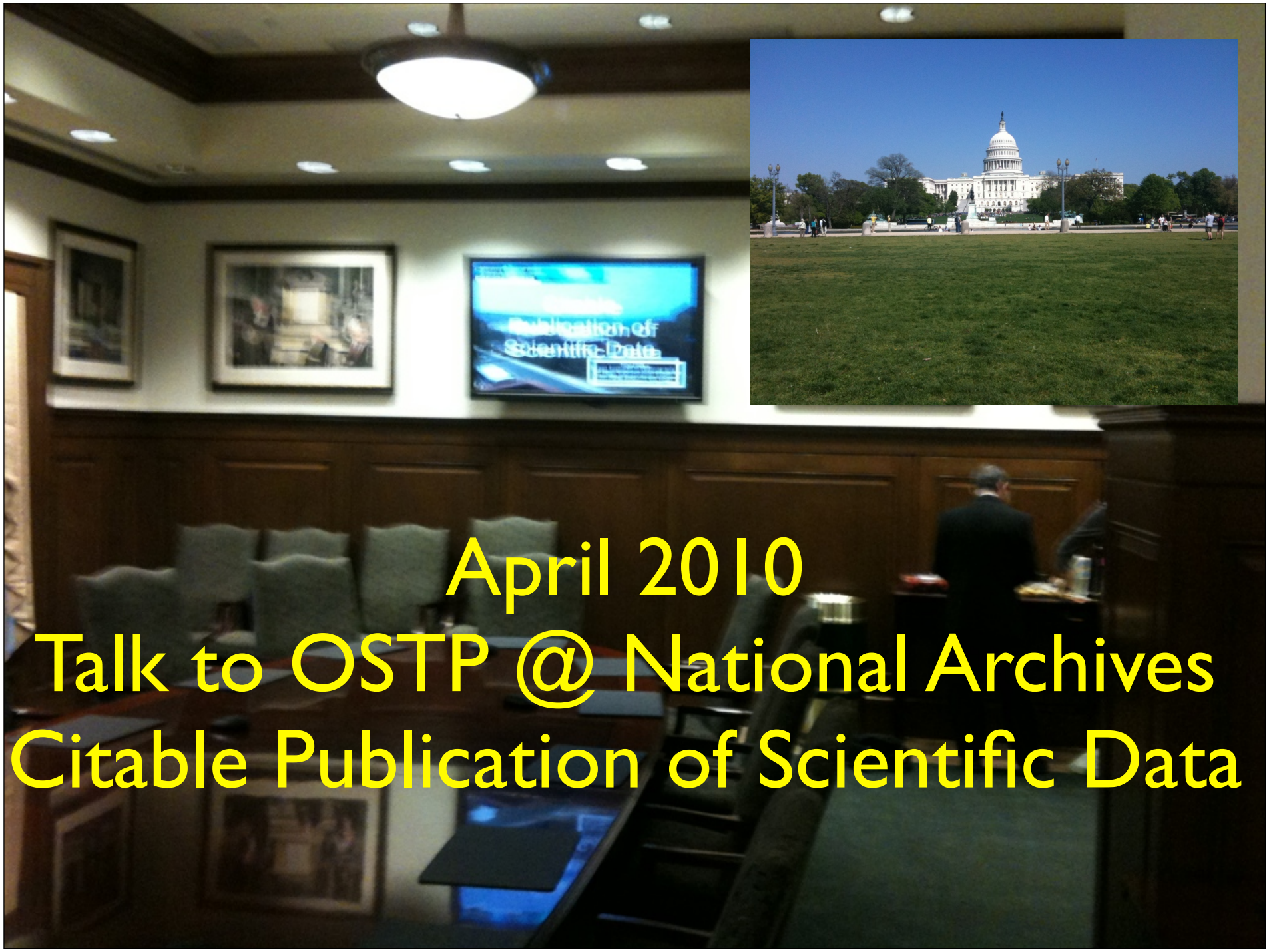
CMMAP

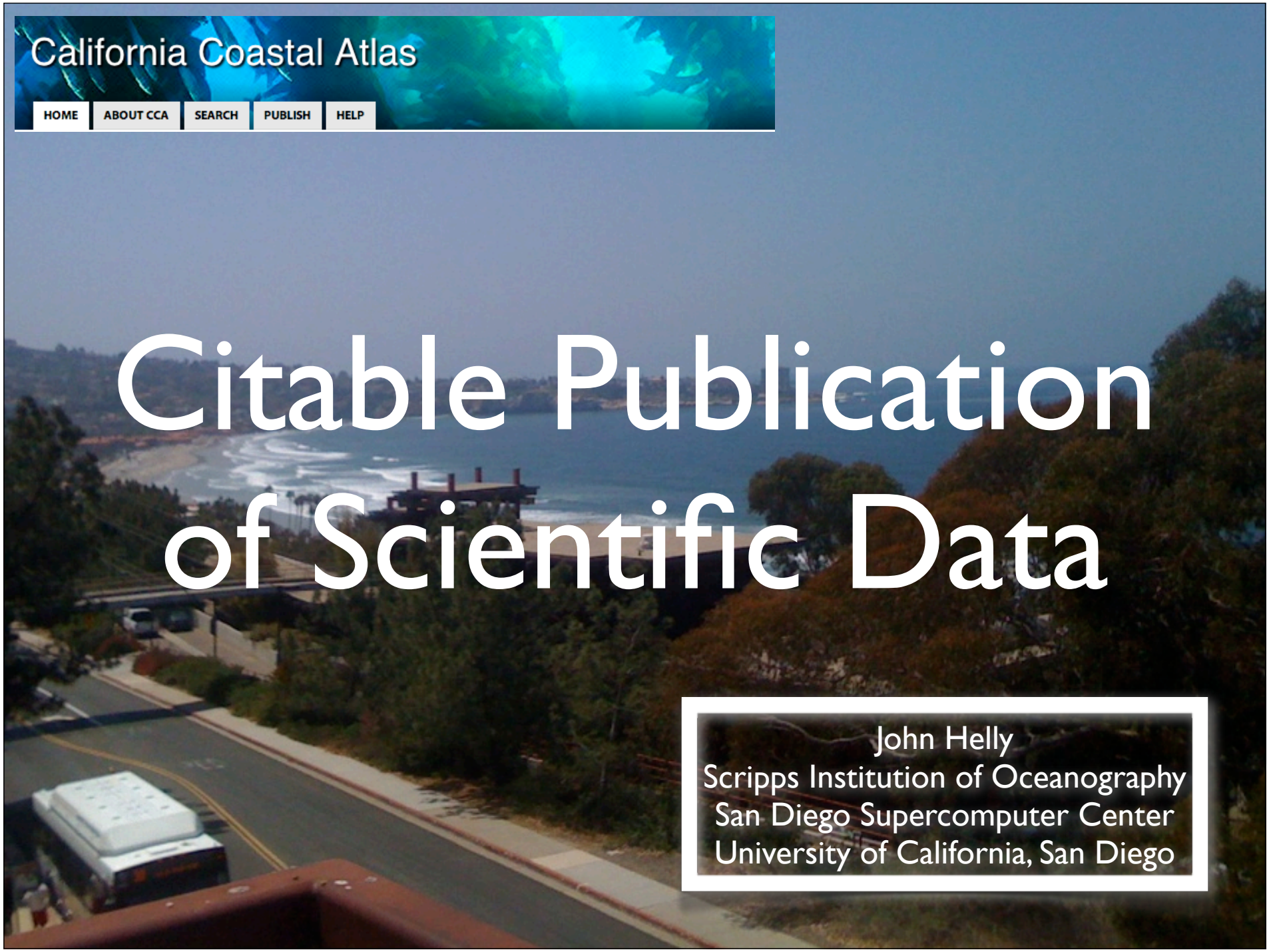
Reach for the sky

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

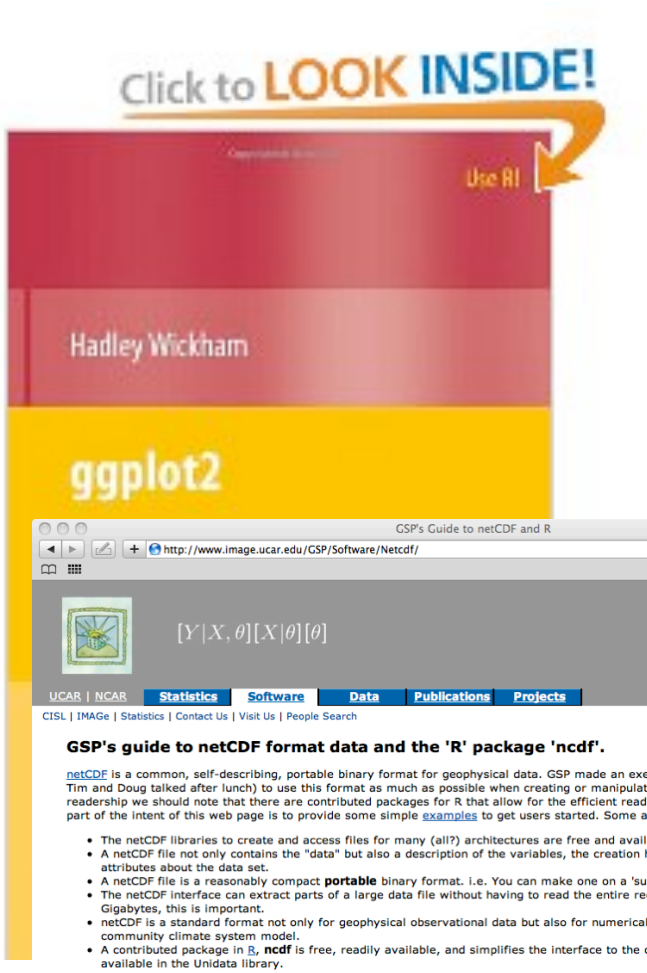

April 2010
Talk to OSTP @ National Archives
Citable Publication of Scientific Data

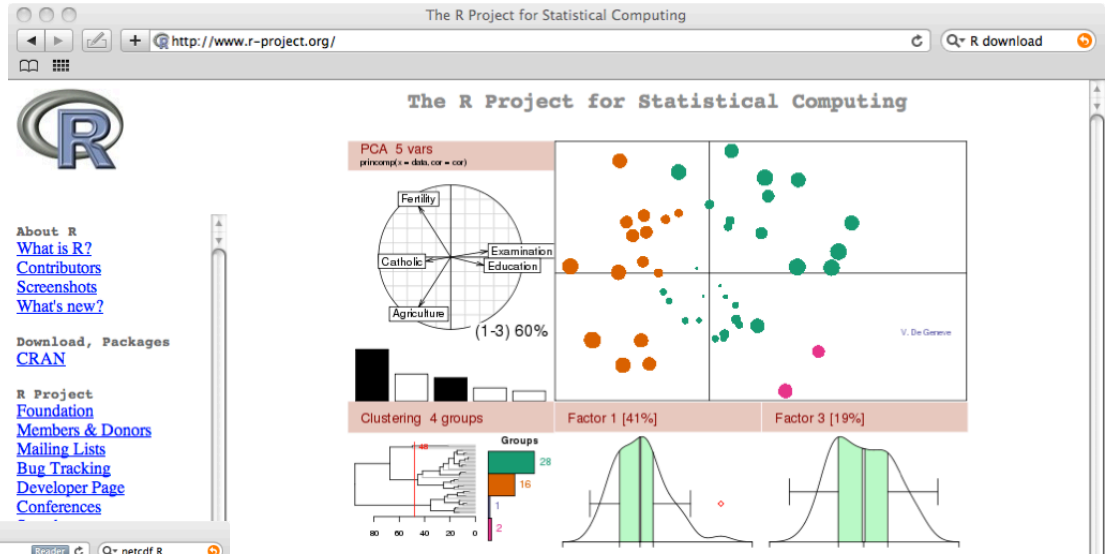


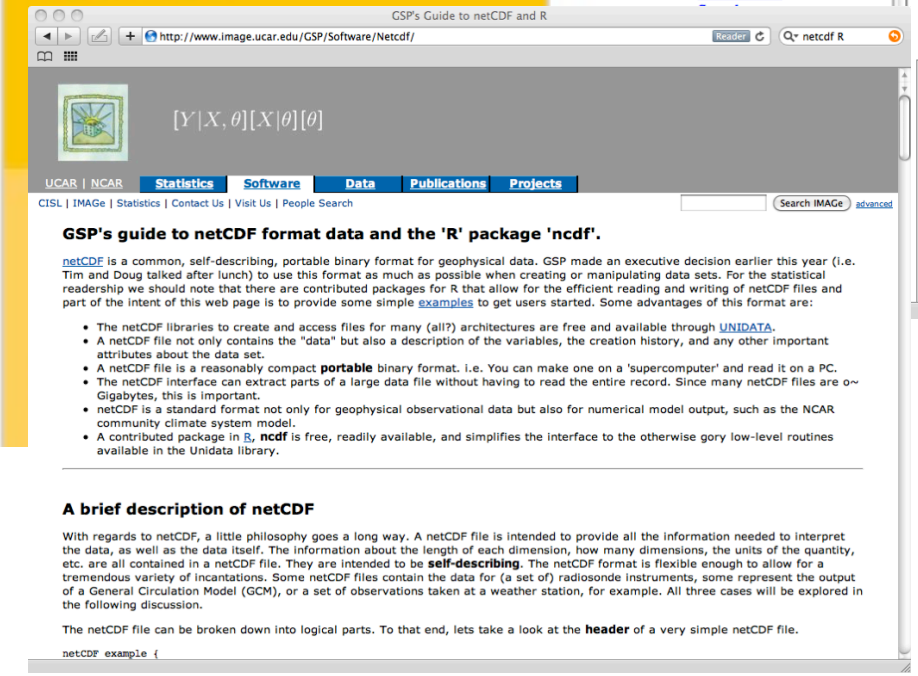
Citable Publication of Scientific Data

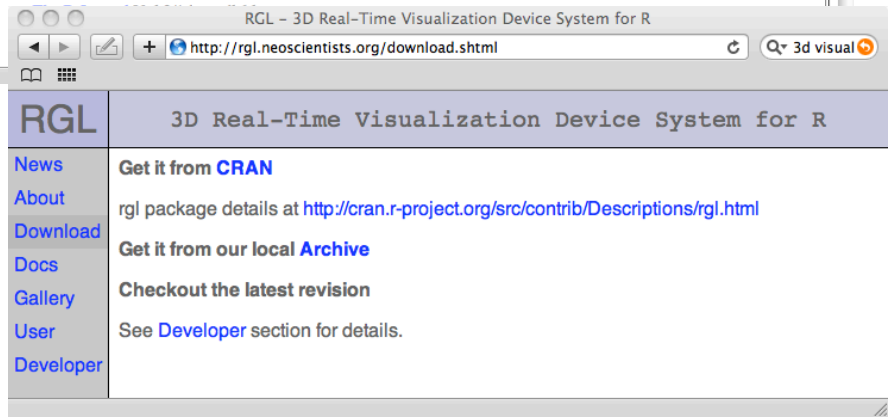
John Helly
Scripps Institution of Oceanography
San Diego Supercomputer Center
University of California, San Diego

R: Analysis and Plotting









CIWG Agenda

* News item

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)

CI-P-1_asSubmitted.pages (Word Processing)

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

92% 8,942 Words Page 1 of 26

NSF Planning Proposal

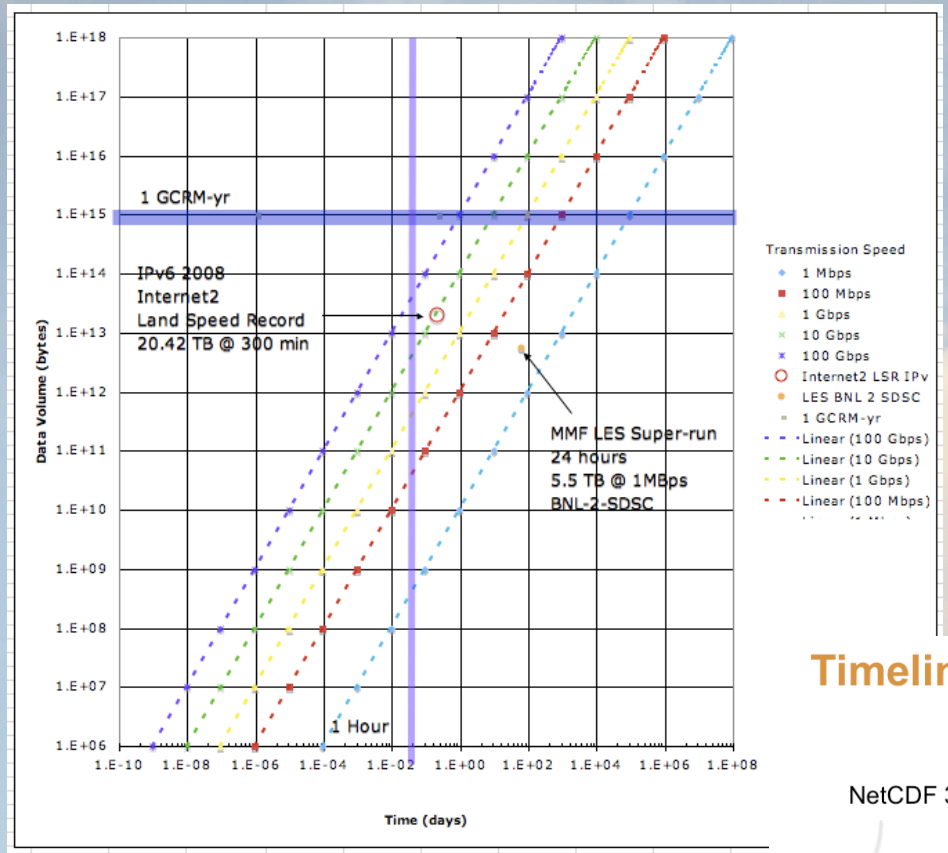
GCRM Data Problem

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:

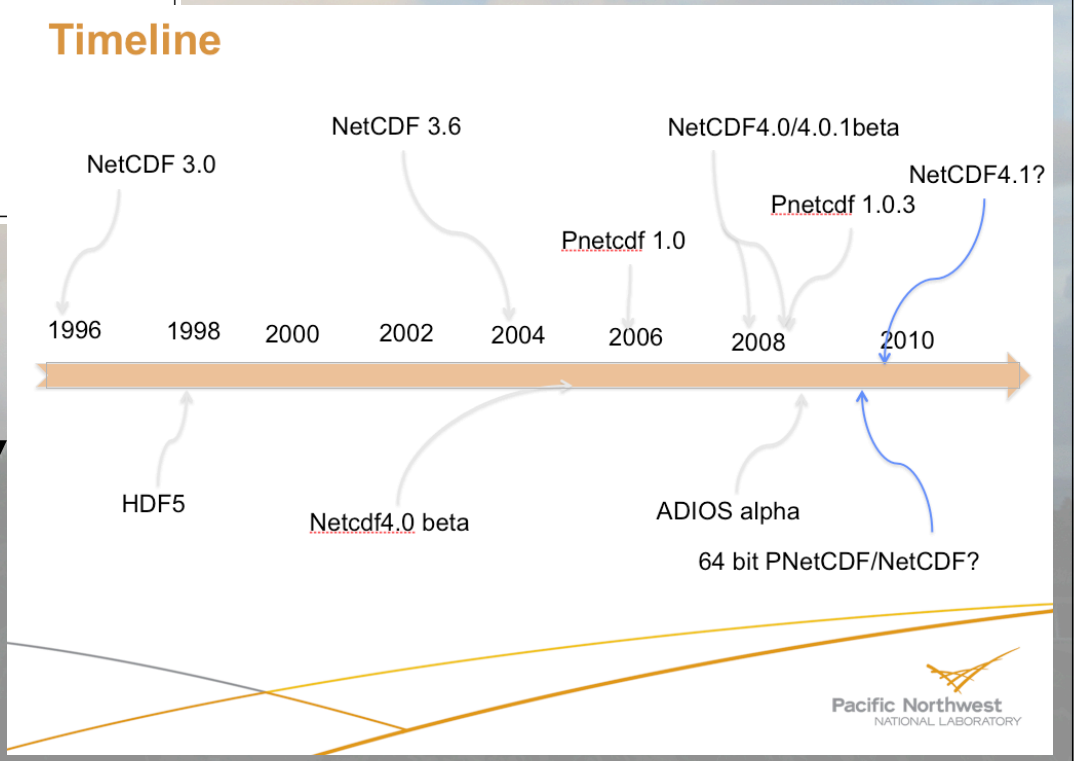
- GCRMs produce terabytes to petabytes of model output. The data is created at supercomputer centers. It must be archived, curated, and made available to users at remote sites.
  - Many difficult choices must be made; for example, choosing which fields to output, and what subsetted spatial and temporal 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



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.

- 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 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

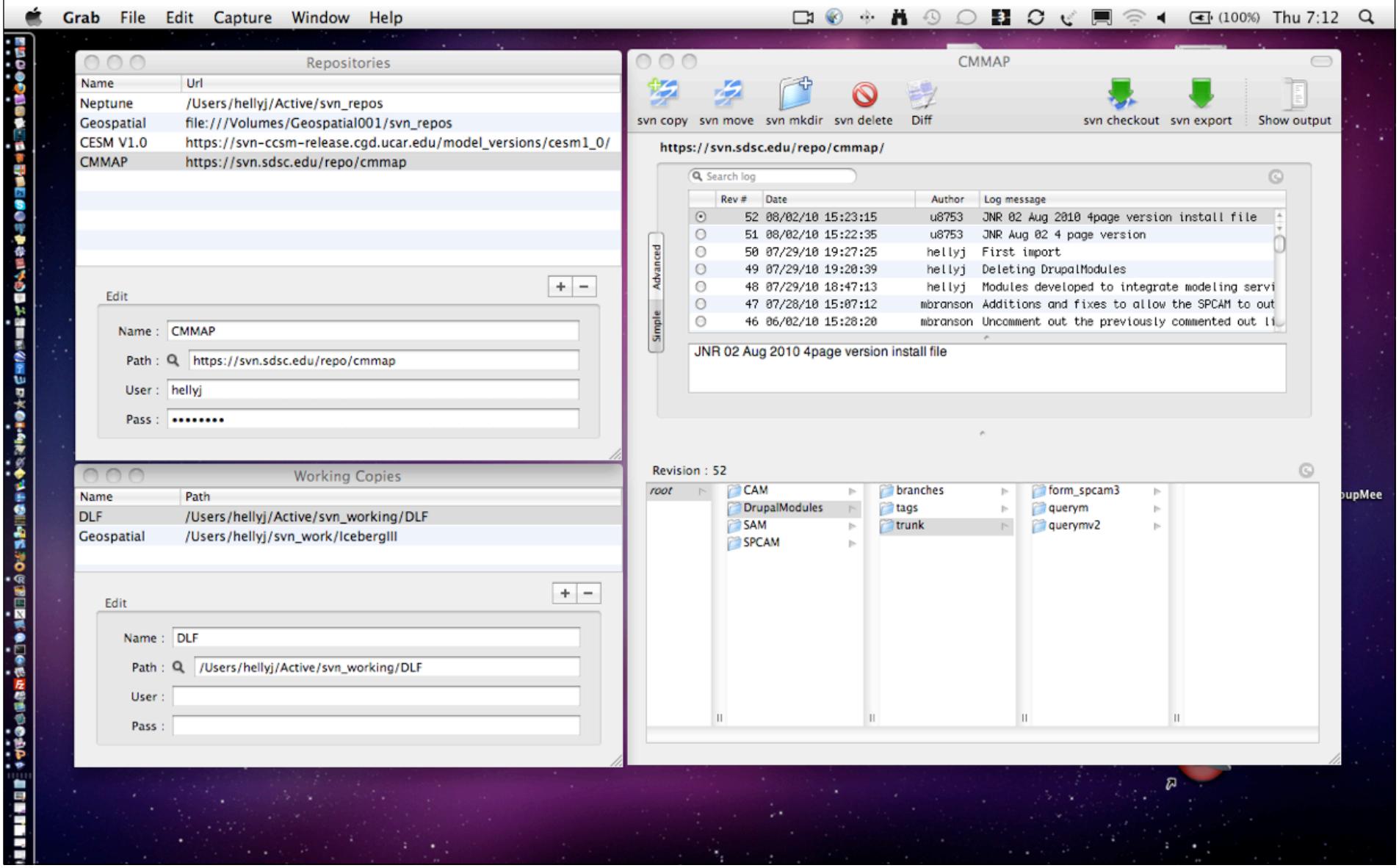
Modeling Global Climate Variability with the Multi-scale Modeling Framework: The Boundary-layer Cloud Problem

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

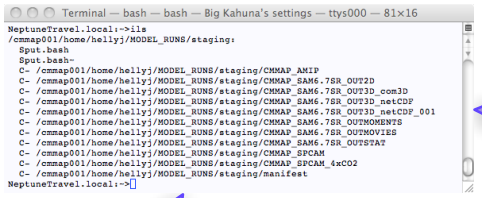


Leveraging National & Partner Resources

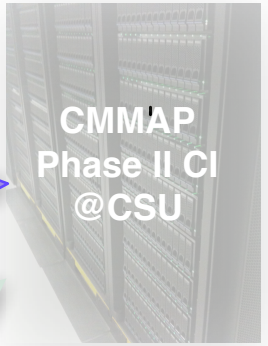
	Organization	Resource	2007	2008	2009	2010
Data Allocations	San Diego Supercomputer Center (SDSC)	Disk	15 Terabytes	15 Terabytes	30 Terabytes	45 Terabytes
		BlueGene			30,000 SUs*	
		Triton				30,000 SUs
Computing Allocations	Teragrid (multi-institution)	SDSC DataStar (IBM SP4)	600,000 SUs	1,200,000 SUs		
		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)	BlueIce IBM Power5			500,000 SUs	
	IBM Watson Research Center	BGW - eServer Blue Gene Solution			TBD	
	Stonybrook				TBD	

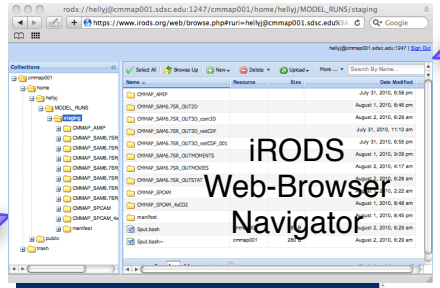
CMMAP Architecture Roadmap

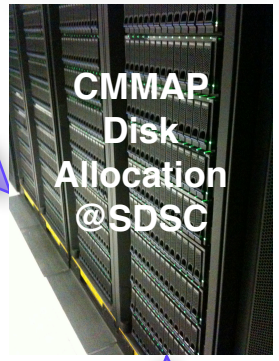
iRODS Command-line Client: Scriptable



iRODS Federation Mechanism

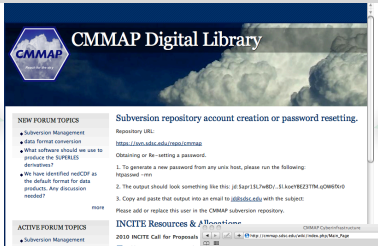


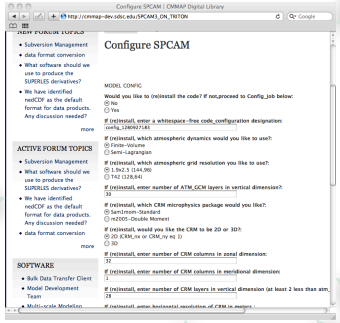


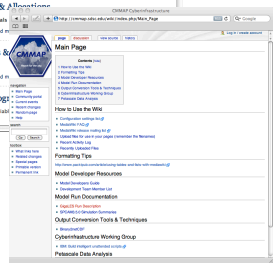


Drupal: programmability

- iquery
- MMF Workbench







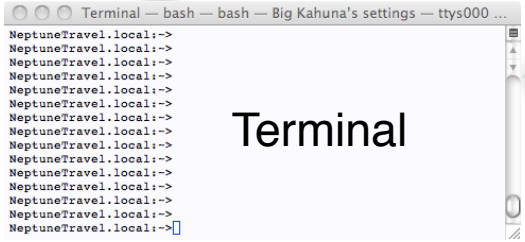
MediaWik: Documentationi

FRE
Workflow Control
Collaboration w/GFDL

others

Community Account

Steele@Teragrid



Terminal

Computing

Data