CMMAP CIWG Breakout

Summer Meeting 2012

John Helly

San Diego Supercomputer Center & Scripps Institution of Oceanography
University of California, San Diego
La Jolla, California 92093

hellyj@ucsd.edu

August 7, 2012

Table of Contents

Table of Contents

Existing CMMAP Resources at SDSC

- Disk allocation (45 TB) used to host digital library and storage for CMMAP researchers via iRODS
- CMMAP Digital Library
  - CMMAP platform for access to data, source code, and XSEDE Community Gateway for CMMAP
  - A new project is starting within the XSEDE 'umbrella' to improve Gateway support. We will participate in that.
- Subversion code repository
  - Currently used for community versions of SPCAM and MACM
  - What is the status of the version in CCSM and how do changes in the svn versions correlate (or not)?
- Disk allocation (45 TB) used to host digital library and storage for CMMAP researchers via iRODS

Digital Library Holdings

Digital Library Holdings

```
NeptuneStonesteps.local:~/work/SLR002/src/bin>ils
/cmmap001/home/cmmap/model_runs:
  INSTALL.txt
  C- /cmmap001/home/cmmap/model_runs/CMMAP_AMIP
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01
  C- /cmmap001/home/cmmap/model_runs/CMMAP_SAM6.7SR_OUT2D
  C- /cmmap001/home/cmmap/model_runs/CMMAP_SAM6.7SR_OUT3D_com3D
  C- /cmmap001/home/cmmap/model_runs/CMMAP_SAM6.7SR_OUT3D_netCDF
  C- /cmmap001/home/cmmap/model_runs/CMMAP_SAM6.7SR_OUT3D_netCDF_001
  C- /cmmap001/home/cmmap/model_runs/CMMAP_SAM6.7SR_OUTMOMENTS
  C- /cmmap001/home/cmmap/model_runs/CMMAP_SAM6.7SR_OUTMOVIES
  C- /cmmap001/home/cmmap/model_runs/CMMAP_SAM6.7SR_OUTSTAT
  C- /cmmap001/home/cmmap/model_runs/CMMAP_SPCAM
  C- /cmmap001/home/cmmap/model_runs/CMMAP_SPCAM_4xCO2
  C- /cmmap001/home/cmmap/model_runs/manifest
NeptuneStonesteps.local:~/work/SLR002/src/bin>
```

```
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/SENS_MW0_wbg
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/SENS_MW1_wbg
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/SENS_MW2_wbg
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/SENS_MW3_wbg
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/SENSvn_MW0
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/SENSvn_MW1
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/SENSvn_MW10
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/SENSvn_MW11
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/SENSvn_MW12
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/SENSvn_MW13
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/SENSvn_MW2
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/SENSvn_MW3
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/SENSvn_MW4
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/SENSvn_MW5
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/SENSvn_MW6
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/SENSvn_MW7
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/SENSvn_MW8
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/SENSvn_MW9
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/SENSvn_MWr4
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/sppt
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/z_MP32_2MW0_ac
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/z_MP32_2MW0_wb
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/z_MP32_2MW11
  C- /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01/z_MP32_2MW12
NeptuneStonesteps.local:~/work/SLR002/src/bin>
```

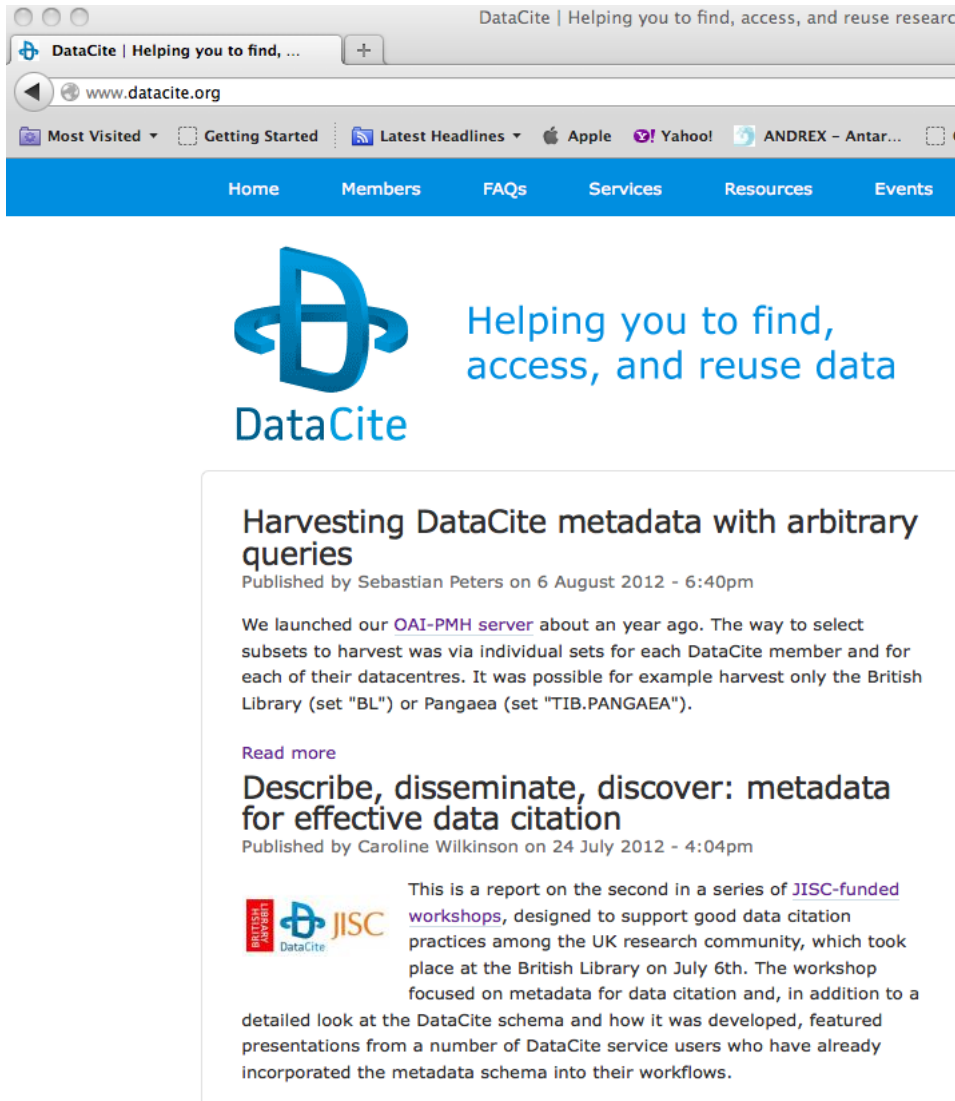
Thanks to Jack Ritchie and Hugh Morrison

Configuration Management Challenges

Model Configuration	PI	Status of Data
SP-CCSM3.0	Rachel Cristina Stan	?
MACM EXP0, EXP1	Hugh Morrison Minghuai Wang	SDSC
SPCAM	Mike Pritchard	UW (coming to SDSC)
GigaLES I	Marat Kharoutdinov	Some at SDSC, some at UTAH, some at NCAR, some missing
GigaLES-SIB3	?	?

Data Publication System for CMMAP as Integral Part of Digital LIbrary

- UC system is supporting DataCite for the assignment of DOIs
- CMMAP should be publishing datasets and source code using DOIs
- We should join DataCite as an authorized issuer of DOIs
- CIWG should recommend to the EXCOMM to authorize becoming a member of DataCite



Data Publication System for CMMAP as Integral Part of Digital LIbrary

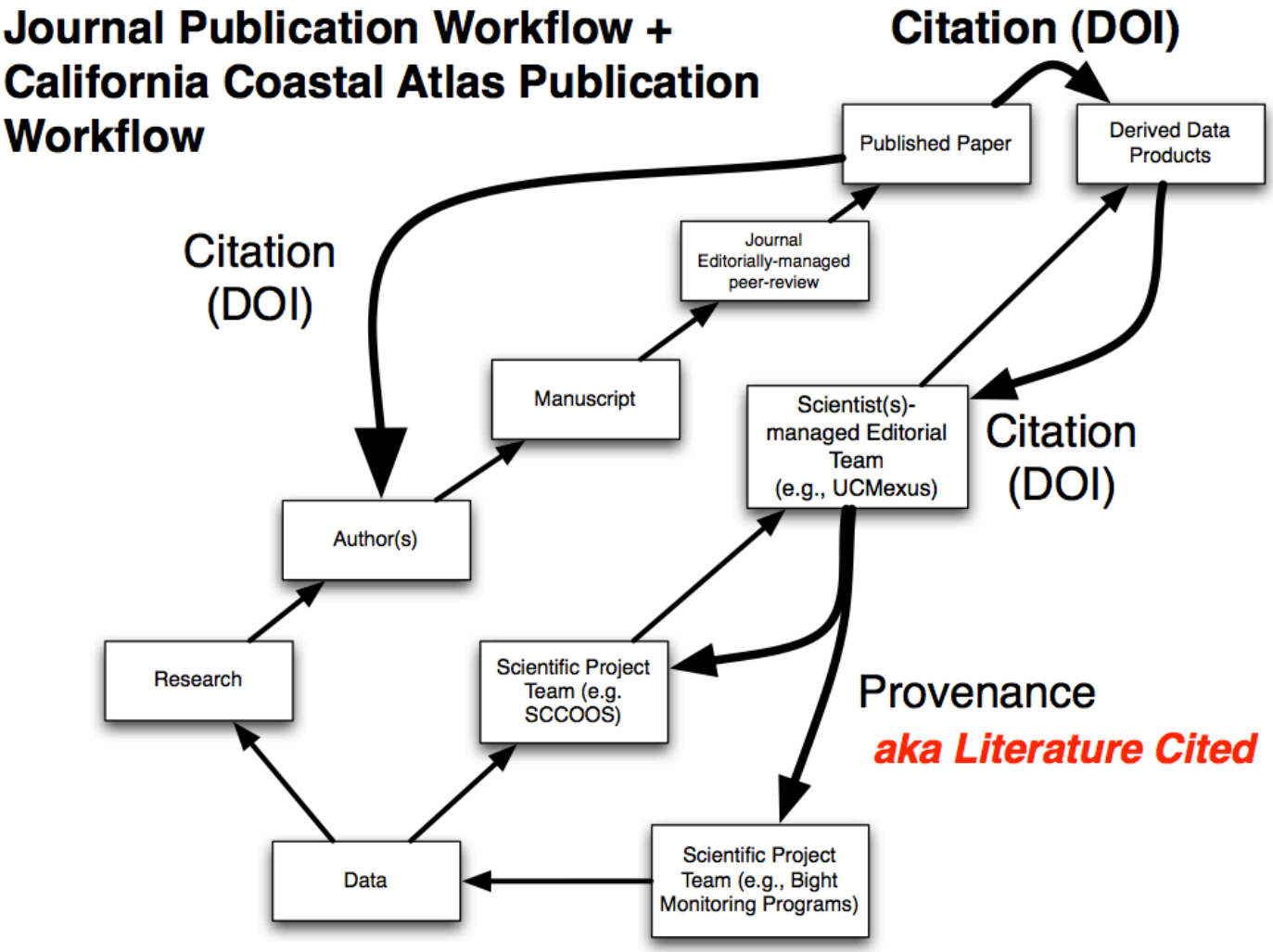
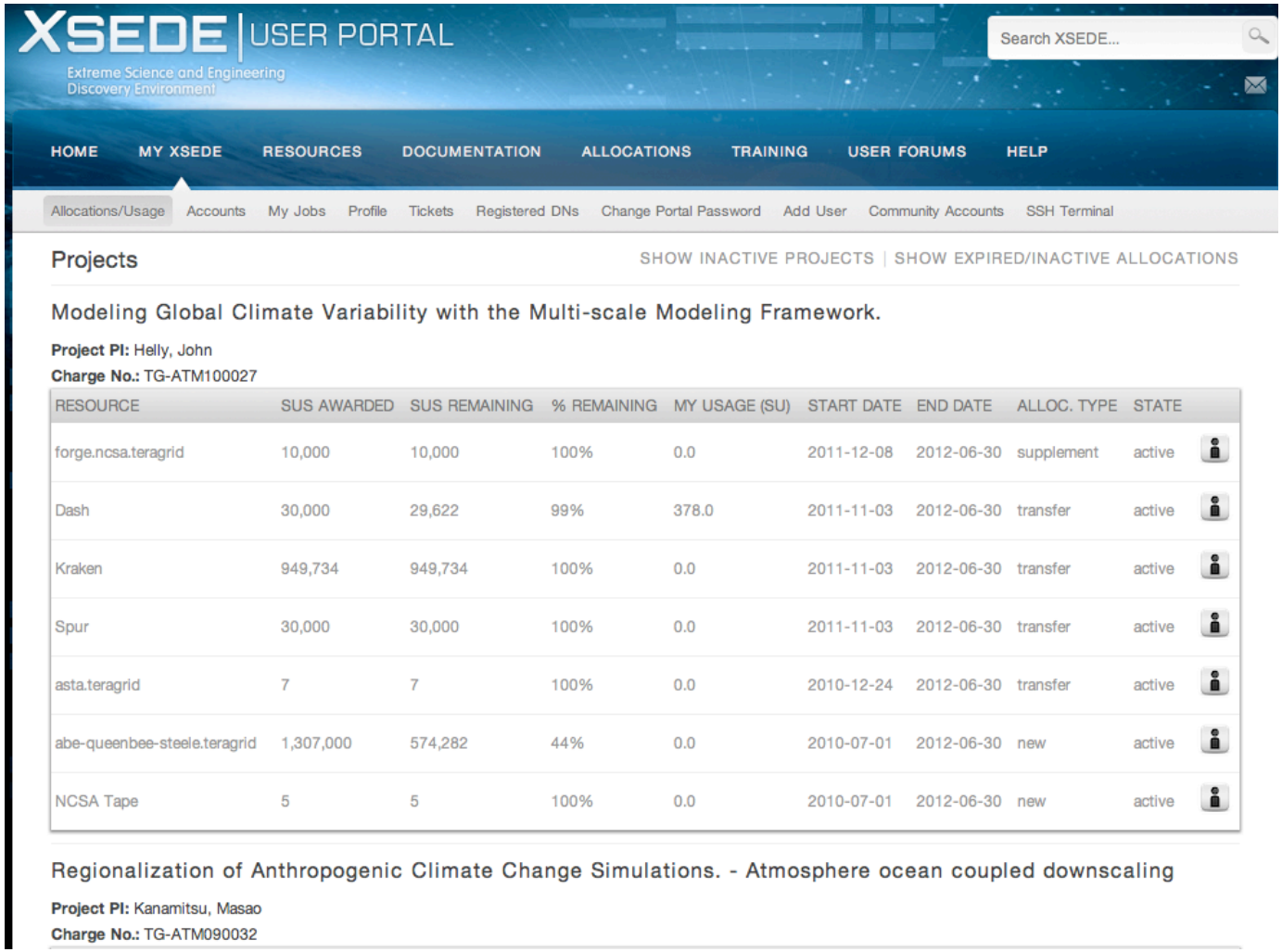


Table of Contents

2M SUs Consumed on Kraken



XSEDE | USER PORTAL
Extreme Science and Engineering Discovery Environment

Search XSEDE...

HOME MY XSEDE RESOURCES DOCUMENTATION ALLOCATIONS TRAINING USER FORUMS HELP

Allocations/Usage Accounts My Jobs Profile Tickets Registered DNs Change Portal Password Add User Community Accounts SSH Terminal

Projects

SHOW INACTIVE PROJECTS | SHOW EXPIRED/INACTIVE ALLOCATIONS

Modeling Global Climate Variability with the Multi-scale Modeling Framework.

Project PI: Helly, John
Charge No.: TG-ATM100027

RESOURCE	SUS AWARDED	SUS REMAINING	% REMAINING	MY USAGE (SU)	START DATE	END DATE	ALLOC. TYPE	STATE
forge.ncsa.teragrid	10,000	10,000	100%	0.0	2011-12-08	2012-06-30	supplement	active
Dash	30,000	29,622	99%	378.0	2011-11-03	2012-06-30	transfer	active
Kraken	949,734	949,734	100%	0.0	2011-11-03	2012-06-30	transfer	active
Spur	30,000	30,000	100%	0.0	2011-11-03	2012-06-30	transfer	active
asta.teragrid	7	7	100%	0.0	2010-12-24	2012-06-30	transfer	active
abe-queenbee-steele.teragrid	1,307,000	574,282	44%	0.0	2010-07-01	2012-06-30	new	active
NCSA Tape	5	5	100%	0.0	2010-07-01	2012-06-30	new	active

Regionalization of Anthropogenic Climate Change Simulations. - Atmosphere ocean coupled downscaling

Project PI: Kanamitsu, Masao
Charge No.: TG-ATM090032

Latest Results: 2M SUs Consumed on Kraken (cont.)

- CMMAP MMF portal was moved from Steele (Purdue) to Kraken (ORNL) last year for use by Hugh Morrison for his microphysics development
- Initially the intention was to use the 'standard' MMF but late breaking results suggested that switching to the PNNL/MACM version would be more desirable. We made this decision at the Winter meeting in Ft. Lauderdale
- Hugh was able to complete some preliminary runs but the results suggested that longer runs did not make sense until he could analyze what was going on with the preliminary results.
- Of course the switch to MACM set back the CMMAP portal development that was predicated on the use of the 'standard' MMF. However, Jack Ritchie made heroic efforts to switch to the MACM.

2M SUs Consumed on Kraken (cont.)

- Mike Pritchard and Gabe Kooperman stepped up and made use of the allocation for some of Gabe's preliminary dissertation work and Mike's continuing investigation of CONUS diurnal variability.
- These results are being analyzed and the runs will be added to the CMMAP Digital Library.
- Don Dazlich was also able to use some SUs to do some benchmarking of SAM coupled with the land-surface model (SIB3, gigales-sib3) from kraken, a Cray XT5 in support of a planned proposal for September to perform new Giga-LES runs.

Table of Contents

Status of Portal Development

Status of Portal Development

- Output from the runs /archive/science/cmmap-staging/source/CMMAP_SPCAM_MACM_EXP0
- iRODS repository /cmmap001/home/cmmap/model_runs/CMMAP_MACM_EXP01
- FRE (GFDL workflow management tools) partially integrated

Table of Contents

Due Dates

Research Allocations

Research allocation requests are reviewed at quarterly XRAC meetings.

Research allocations may be requested for any computation resource and require a formal request document and CVs (for PI/Co-PIs). Research allocations are typically appropriate as follow-ons to Startups; but a PI need not request a Startup prior to submitting a Research request. Requests may be submitted the quarter before they are needed during designated submission windows. The XSEDE Resource Allocations Committee (XRAC) meets quarterly to review requests based on merit and award available SUs.

OPEN SUBMISSIONS	CLOSE SUBMISSIONS	ALLOCATIONS BEGIN
Dec 15	Jan 15	April 1
Mar 15	Apr 15	Jul 1
Jun 15	Jul 15	Oct 1
Sep 15	Oct 15	Jan 1

XSEDE: What Experiments and What Resources Do We Want?

1. New GigaLES Run with SAM and SIB3 (Kraken)
   1. How many runs?
   2. How much data storage?
   3. What data format(s)?
2. CMMAP Model Intercomparison?
3. CMMAP Community Account Portal for SPCAM5 (Kraken)
4. Visualization and Data Analysis (Gordon)
5. Data Transfer via DYNES and possibly BISON

First GigaLES Specifications: Idealized GATE Simulation of Convection over Tropical Atlantic) from Marat Kharoutdinov

- Based on average forcing and sounding from GATE Phase III observations (30 August - 19 September 1074, Tropical Atlantic)
- Forcing: SST, horizontal advective tendencies of s and q; mean wind nudged to observed; radiative heating prescribed; surface fluxes - interactive.
- Domain: 2048 x 2048 x 256 grid points, or 205 x 205 x 27 km^3 (horizontal grid spacing 100m)
- Vertical grid spacing: (50m below 1km, 50-100m @1-5km, 100m @5-18km, 100-300m above)
- Time step: 2 sec, duration: 1 day
- Initialization: random small-amplitude noise in temperature near the surface; run done over 6 days wall-clock time on 2048 processors of IBM BlueGene BG/L of NYCCS

Table of Contents

Parallel I/O Update

Pagoda: Parallel Analysis of Geoscience Data

Overview

Pagoda is both an API for data-parallel analysis of geodesic climate data as well as the set of data-parallel processing tools based on this API. The API and the tools were designed first to support geodesic semi-structured NetCDF data, however they are generic enough to work with regularly gridded data as well. The command line tools are designed to be similar to the NetCDF Operators.

Pagoda is currently in an early release stage. While we intend to mimic the NCO tools, some functionality is not yet available. In particular, the command line tools do not yet process missing values. Metadata editing operations are not implemented since they are not inherently parallel operations.

Pagoda has been tested on Linux workstations and Cray XT4 (franklin at NERSC).

- Download has the latest release
- Build Documentation has information on building pagoda
- User's Guide has information on running the command-line tools

Currently Supported Command-line Tools

- pgra record averages
- pgea ensemble averages
- pgbo binary operations
- pgsub data subsetting
- pgflint file iterpolation

News

- 05/11/2011: Presented at GO-ESSP. Here are the slides (PDF,2.1M) on the GO-ESSP web page. Also available on our presentations page.
- 04/11/2011: Version 0.6 released

Parallel I/O Update

NetCDF 4.2.1 Release Notes

The 4.2.1 relelase for the netCDF C libraries contains new features, bug fixes, performance improvements, and internal refactoring.

- Ported static and shared libraries (DLL's) for both 32- and 64-bit Windows, including support for DAP remote access, with netCDF-3 and netCDF-4/HDF5 support enabled. The environment for this build is MSYS/MinGW/MinGW64, but the resulting DLLs may be used with Visual Studio. [NCF-112] [NCF-54] [NCF-57] [NCF-65]
- Implemented diskless files for all netCDF formats. For nc_create(), diskless operation performs all operations in memory and then optionally persists the results to a file on close. For nc_open(), but only for netcdf classic files, diskless operation caches the file in-memory, performs all operations on the memory resident version and then writes all changes back to the original file on close. [NCF-110] [NCF-109] [NCF-5]
- Added MMAP support. If diskless file support is enabled, then it is possible to enable implementation of diskless files using the operating system's MMAP facility (if available). The enabling flag is "--enable-mmap". This is most useful when using nc_open() and when only parts of files, a single variable say, need to be read.
- Changed the file protections for NC_DISKLESS created files to 0666. [NCF-182]
- Added a specific NC_MMAP mode flag to netcdf.h to modify behavior of NC_DISKLESS.
- Added configure flag for --disable-diskless.
- Added nccopy command-line options to exploit diskless files, resulting in large speedups for some operations, for example converting unlimited dimension to fixed size or rechunking files for faster access. Upgraded doxygen and man-page documentation for ncdump and nccopy utilities, including new -w option for diskless nccopy, with an example.
- Modified Makefile to allow for concurrent builds and to support builds outside the source tree, e.g. 'mkdir build; cd build; SOURCE-DIR/configure' where SOURCE-DIR is the top-level source directory.
- Fixed some netCDF-4 bugs with handling strings in non-netCDF-4 HDF5 files. [NCF-150]
- Fixed bug using nccopy to compress with shuffling that doesn't compress output variables unless they were already compressed in the input file. [NCF-162]
- Fixed bug in 64-bit offset files with large records, when last record variable requires more than 2**32 bytes per record. [NCF-164]
- Fix bug in which passing a NULL path to nc_open() causes failure. [NCF-173]
- Fixed ncgen bugs in parsing and handling opaque data.

Parallel I/O Update

Parallel netCDF: A High Performance API for NetCDF File Access

Overview

Parallel netCDF (PnetCDF) is a library providing high-performance I/O while still maintaining file-format compatibility with Unidata's NetCDF.

NetCDF gives scientific programmers a space-efficient and portable means for storing data. However, it does so in a serial manner, making it difficult to achieve high I/O performance. By making some small changes to the API specified by NetCDF, we can use MPI-IO and its collective operations.

- Download has the latest release and development links as well as information about svn access.
- Documentation: a QuickTutorial, plus papers, presentations, articles, and other resources
- Benchmarking: tools and suggestions for evaluating pnetcdf performance

News

- In the 1.3.0 release, the unsigned and 64-bit integer data types are supported for CDF-5 format. The unsigned data types include NC_UBYTE, NC_USHORT, NC_UINT, and NC_UINT64. The 64-bit integer data types are NC_INT64 and NC_UINT64.
- New APIs for supporting more data types are added. For C, they are ncmpi_(i)put/(i)get_var*_ushort/uint/longlong/ulonglong. For Fortran, they are nfmpi_(i)put/(i)get_var*_int8.
- A new set of "buffered"-put APIs is supported in 1.3.0 release. The nonblocking iput/iget APIs require the contents of user buffers not to be changed until the wait call completed. The bput APIs use a user attached buffer to make a copy of request data, so the user buffer is free to change once the bput call returns.
- The special character set, "special2", and multi-byte UTF-8 encoded characters introduced in the CDF-2 file format for variable, dimension, and attribute name strings are now supported.
- A set of example programs and QuickTutorial are now available.

- Nonblocking I/O is redesigned in the 1.2.0 release. It defers the I/O requests until "wait" call, so small requests can be aggregated into large ones for better performance.
- Two new hints, `nc_header_align_size` and `nc_var_align_size`, are added. The former allows pre-allocation of a larger header size to accommodate new header data in case new variables or attributed are added later. The latter aligns the starting file offsets of non-record variables. Refer to VariableAlignment for a more detailed description.
- Data consistency control has been revised. A more strict consistency can be enforced by using `NC_SHARE` mode at the file open/create time. In this mode, the file header is synchronized to the fil if its contents have changed. Such file synchronization of calling `MPI_File_sync()` happens in many places, including `ncmpi_enddef()`, `ncmpi_redef()`, all APIs that change global or variable attributes, dimensions, and number of records.
- As calling `MPI_File_sync()` is very expensive on many file systems, users can choose more relaxed data consistency, i.e. by not using `NC_SHARE`. In this case, file header is synchronized among processes in memories. No `MPI_File_sync()` will be called if header contents have changed. `MPI_File_sync()` will only be called when switching data mode, i.e `ncmpi_begin_indep_data()` and `ncmpi_end_indep_data()`.

A note about Large File Support

As of Pnetcdf-0.9.2, we ship with support for CDF-2 formated data. With this format, even 32 bit platforms can create netcdf datasets (files) greater than 2GB in size. See the file README.large_files in the source tree for more information. CDF-2 also allows more special characters in the name strings of defined dimension, variables, and attributes.

The maintainers of the serial NetCDF library added support for the CDF-2 format in netcdf-3.6.0. The support was based largely on work from Greg Sjaardema.

The CD-5 file format specification: supports unsigned and 64-bit integer data types and variables with more than 2^{32} array elements.

The CDF (or CDF-1) file format specification has been in use through netCDF library version 3.5.1.

File and Variable Limits

Both PnetCDF and NetCDF share limitations on file and variable sizes. More information can be found on the FileLimits page.

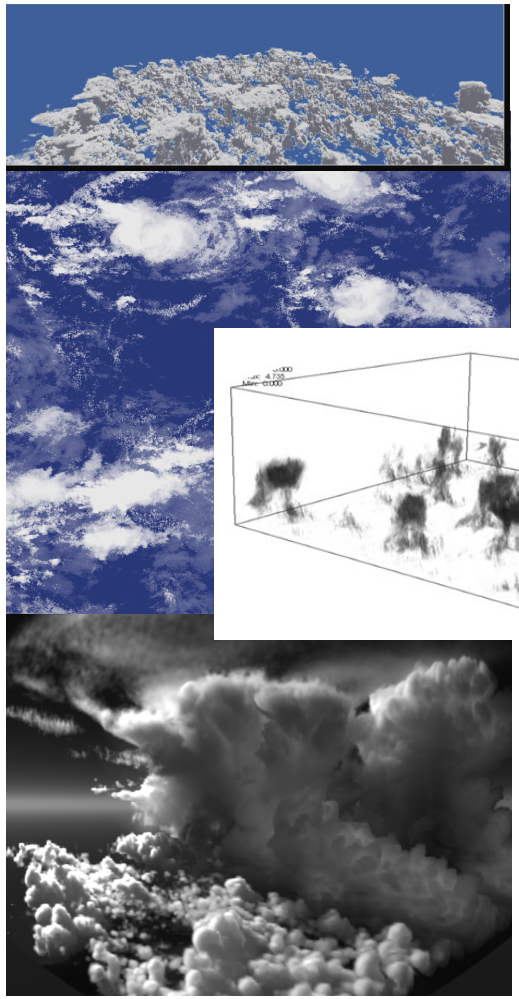
Required Software

PnetCDF requires an MPI implementation with MPI-IO support. Most MPI libraries have this nowadays. A parallel file system would also go a long way towards achieving highest performance.

Table of Contents

Visualization and Data Analysis

- VisIT testing has been done on Gordon@SDSC to evaluate how it handles netCDF files and how to control the rendering engine
- Marat Kharoutdinov has a ray-tracing (ray-casting?) algorithm
- Steve Krueger and his student Ian have some nice results with SHDOM (radiative transfer code)



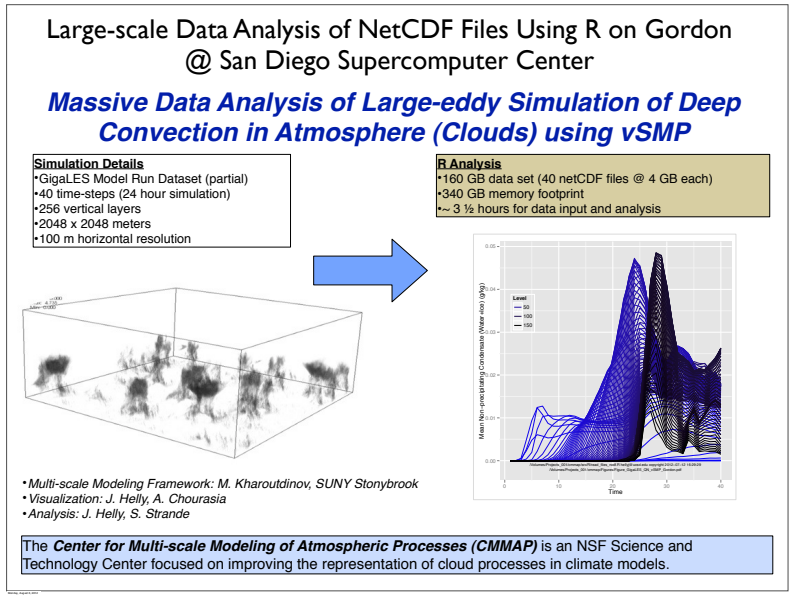
Marat Kharoutdinov

VisIT
John Helly
Amit Chourasia

SHDOM
Ian Glenn
Steve Krueger

Visualization and Data Analysis (cont.)

- This has been combined with NCO-based 'stacking' procedures to combine 40 netCDF files into 4GB single time-step files for a 160 GB dataset
- Same data files used for viz were then analyzed with R using Gordon



Gordon Specifications

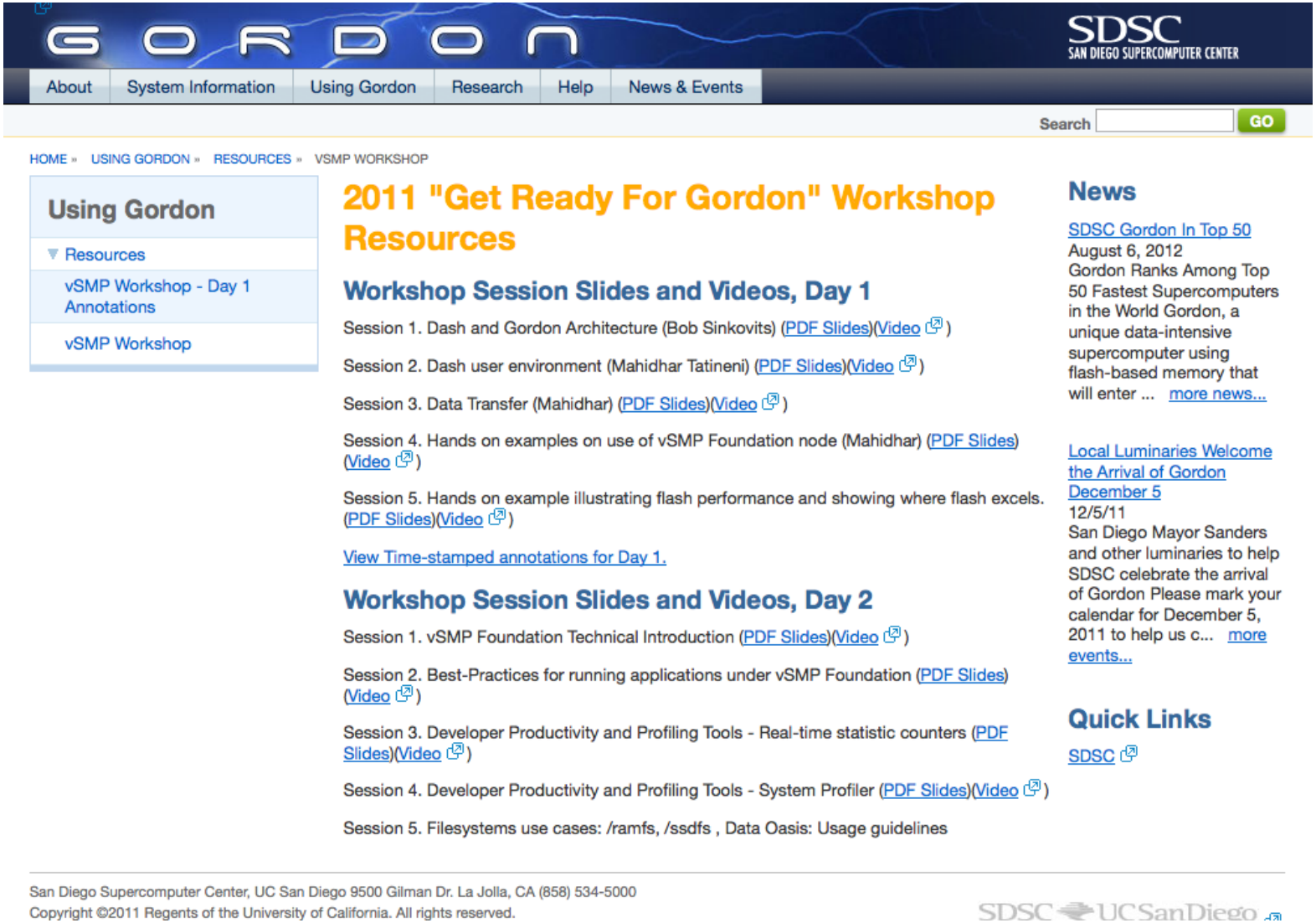
System Component	Configuration
Intel EM64T Xeon E5 Compute Nodes	
Sockets	2
Cores	16
Clock speed	2.6 GHz
Flop speed	333 Gflop/s
Memory capacity	64 GB
Memory bandwidth	85 GB/s
STREAM Triad bandwidth	GB/s

I/O Nodes	
Sockets	2
Cores	12
Clock speed	2.67 GHz
Memory capacity	48 GB
Memory bandwidth	64 GB/s
Flash memory	4.8 TB

Full System	
Total compute nodes	1024
Total compute cores	16384
Peak performance	341 Tflop/s
Total memory	64 TB
Total memory bandwidth	87 TB/s
Total flash memory	300 TB

QDR InfiniBand Interconnect	
Topology	3D Torus
Link bandwidth	8 GB/s (bidirectional)
Peak bisection bandwidth	TB/s (bidirectional)
MPI latency	1.3 µs
DISK I/O Subsystem	
File Systems	NFS, Lustre
Storage capacity (usable)	150 TB: Dec 2010 2 PB: June 2011 4 PB: July 2012
I/O bandwidth	GB/s

vSMP on Gordon



GORDON

SDSC SAN DIEGO SUPERCOMPUTER CENTER

About | System Information | Using Gordon | Research | Help | News & Events

Search GO

HOME » USING GORDON » RESOURCES » VSMP WORKSHOP

Using Gordon

- Resources
  - vSMP Workshop - Day 1 Annotations
  - vSMP Workshop

2011 "Get Ready For Gordon" Workshop Resources

Workshop Session Slides and Videos, Day 1

Session 1. Dash and Gordon Architecture (Bob Sinkovits) (PDF Slides)(Video)

Session 2. Dash user environment (Mahidhar Tatineni) (PDF Slides)(Video)

Session 3. Data Transfer (Mahidhar) (PDF Slides)(Video)

Session 4. Hands on examples on use of vSMP Foundation node (Mahidhar) (PDF Slides) (Video)

Session 5. Hands on example illustrating flash performance and showing where flash excels. (PDF Slides)(Video)

View Time-stamped annotations for Day 1.

Workshop Session Slides and Videos, Day 2

Session 1. vSMP Foundation Technical Introduction (PDF Slides)(Video)

Session 2. Best-Practices for running applications under vSMP Foundation (PDF Slides) (Video)

Session 3. Developer Productivity and Profiling Tools - Real-time statistic counters (PDF Slides)(Video)

Session 4. Developer Productivity and Profiling Tools - System Profiler (PDF Slides)(Video)

Session 5. Filesystems use cases: /ramfs, /ssdfs , Data Oasis: Usage guidelines

News

SDSC Gordon In Top 50
August 6, 2012
Gordon Ranks Among Top 50 Fastest Supercomputers in the World Gordon, a unique data-intensive supercomputer using flash-based memory that will enter ... more news...

Local Luminaries Welcome the Arrival of Gordon December 5
12/5/11
San Diego Mayor Sanders and other luminaries to help SDSC celebrate the arrival of Gordon Please mark your calendar for December 5, 2011 to help us c... more events...

Quick Links

SDSC

San Diego Supercomputer Center, UC San Diego 9500 Gilman Dr. La Jolla, CA (858) 534-5000
Copyright ©2011 Regents of the University of California. All rights reserved.

SDSC UCSanDiego

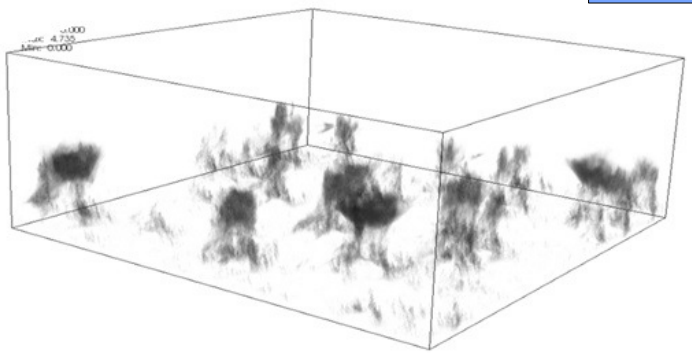
Visualization and Data Analysis (cont.)

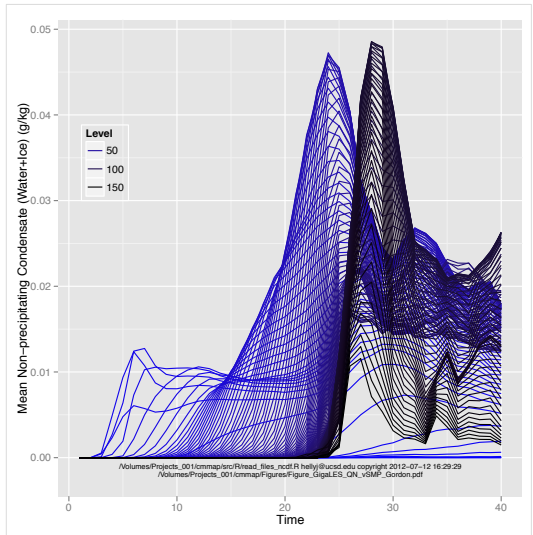
Large-scale Data Analysis of NetCDF Files Using R on Gordon @ San Diego Supercomputer Center

Massive Data Analysis of Large-eddy Simulation of Deep Convection in Atmosphere (Clouds) using vSMP

Simulation Details
- GigaLES Model Run Dataset (partial)
- 40 time-steps (24 hour simulation)
- 256 vertical layers
- 2048 x 2048 meters
- 100 m horizontal resolution

R Analysis
- 160 GB data set (40 netCDF files @ 4 GB each)
- 340 GB memory footprint
- ~ 3 ½ hours for data input and analysis





- *Multi-scale Modeling Framework: M. Kharoutdinov, SUNY Stonybrook*
- *Visualization: J. Helly, A. Chourasia*
- *Analysis: J. Helly, S. Strande*

The ***Center for Multi-scale Modeling of Atmospheric Processes (CMMAP)*** is an NSF Science and Technology Center focused on improving the representation of cloud processes in climate models.

Table of Contents

DYNES and BISON Data Transportation Rates

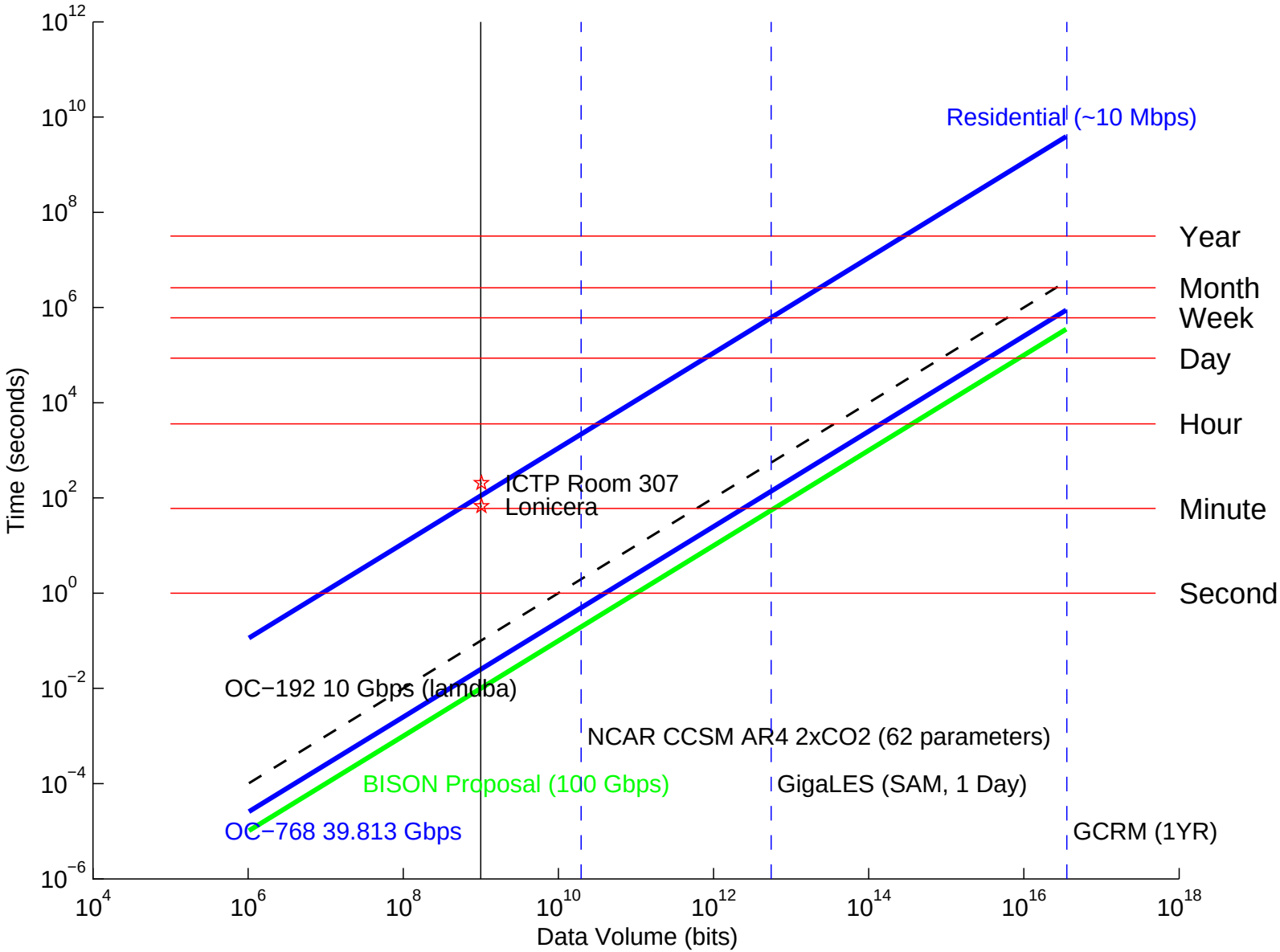


Table of Contents

GPU processing on Keeneland

- Ross Heikes provided GCRM kernel code for testing
- Some account problems delayed testing so this will proceed next quarter

Table of Contents

Discussion