DOE's Climate and Earth System Modeling Town Hall:

Climate Model Analysis and Visualization Efforts for Next Generation Needs

Dean N. Williams

williams13@llnl.gov

U.S. DEPARTMENT OF ENERGY | Office of Science

DOE Climate Town Hall – Data and Visualization Infrastructure
Monday, December 5, 2011

Drivers and Directions

Overview

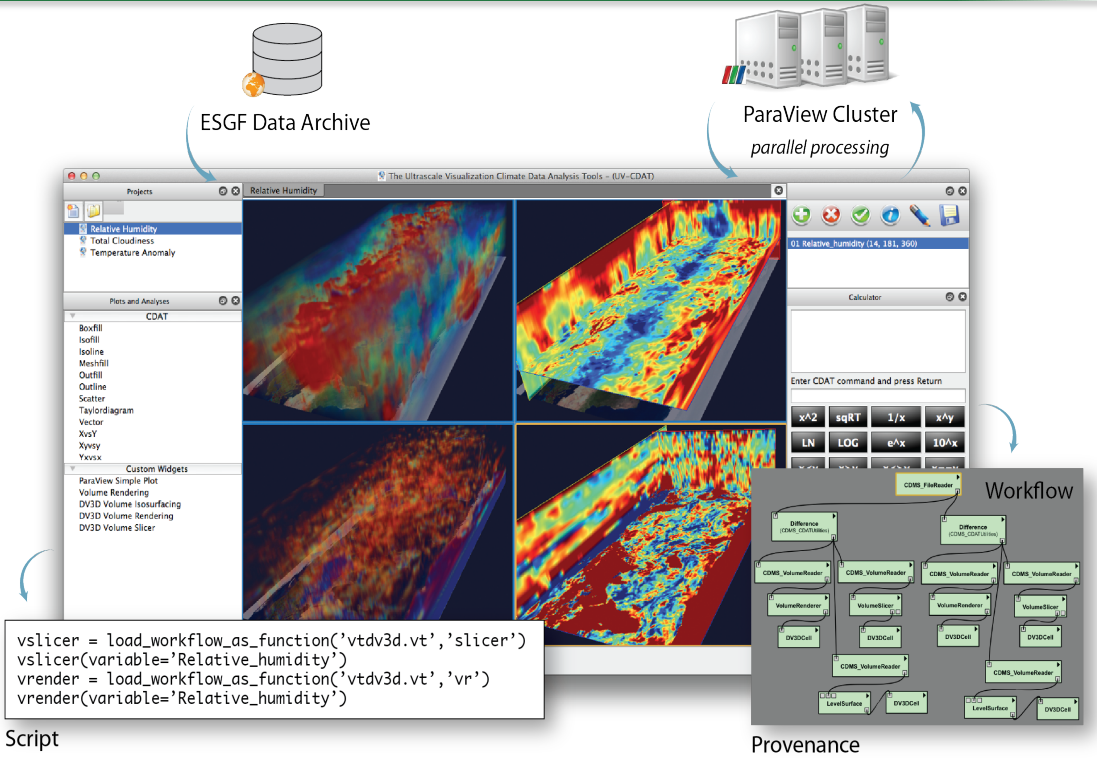
- **Exponential expansion of model and observational data and scientific needs**
  - Existing solutions do not scale to new data volumes
  - Provenance must be documented and retained
  - Stewardship and security
  - Discoverability and accessibility
- **New and evolving technologies**
  - Data Informatics Computational Platforms (Data Engines)
    - GPU and Hybrid architectures with fast networks and large memory
    - Cloud and On-demand computing
  - Expanded network bandwidth and reach
  - Data science and data analytics is a rapidly expanding computational science sub-discipline
- **New paradigm**
  - Federations of data centers
  - Distribution of workload between clients and "data engines"
  - Automation and capture of workflows for reproducibility and efficiency
- **Current projects**
  - Make current tools and approaches more efficient by incorporating new computational and information technologies
  - Research to build next generation capabilities
  - Examples follow

Ultra-scale Visualization Climate Data Analysis Tools

Objective: Integrates several existing, widely used open-source data analysis and visualization packages into seamless environment

- CDAT – Climate data analysis/viz
- VTK - Visualization Toolkit
- R – Statistical analysis
- VisTrails – Workflow Provenance
- VisIt, ParaView – 3D Visualization

- Local and remote visualization and data access
- Comparative visualization and statistical analyses
- Robust tools for regridding, reprojection, and aggregation
- Support for unstructured grids and non-gridded observational data, including geospatial formats often used for observational data sets
- Workflow analysis and provenance management



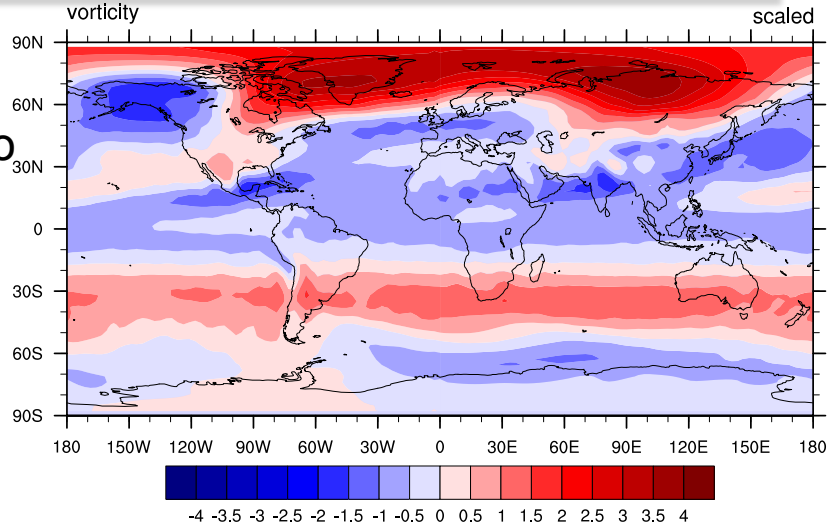
Recent Accomplishment

- ParaView successfully demonstrated the scalability of a new spatio-teomporal pipeline by processing 1/2 TeraByte of data composed of 365 time steps of 1/10 degree POP ocean model in under 2 minutes.

Contact Dean N. Williams (williams13@llnl.gov) for more information or see http://uv-cdat.org/wiki

parVis

Parallel Analysis Tools and New Visualization Techniques for Ultra-Large Climate Data Sets



Objective: Speed up the production of standard 2D plots of climate model output and allow application to ultra-large data sets and native grids

- Speed up current diagnostics (e.g the CESM-CAM atmospheric model diagnostics) with **task parallelism**
- Create a **data-parallel version** of the NCAR Command Language (NCL) analysis and visualization package.
- Build a new library: ParCAL – Parallel Climate Analysis Library.
- Use existing software technology (MOAB, PnetCDF, Intrepid).
- ParNCL (built with ParCAL) will allow users to run their NCL scripts unaltered.
- Explore news ways of doing 3D visualization of climate data

Also anticipate **future hardware landscape** for climate analysis

- Introduce compression within future NetCDF to cope with relatively small disk sizes.
- Building MapReduce-based climate analysis tools for cloud-based platforms.

Recent Accomplishments:

- 3x to 4x speedup of CESM-CAM diagnostics. Released to community.
- Data parallel time averaging, vorticity (see above) and divergence calculations implemented.
- ParNCL interface to ParCAL working with simple scripts.
- Developed 2x – 3x lossless compression for smooth climate data

Contact Rob Jacob (jacob@mcs.anl.gov) for more information or see http://trac.mcs.anl.gov/projects/parvis

U.S. DEPARTMENT OF ENERGY | Office of Science | Argonne NATIONAL LABORATORY | Computation Institute | NCAR | Pacific Northwest NATIONAL LABORATORY | Sandia National Laboratories | UCDAVIS UNIVERSITY OF CALIFORNIA

Visual Data Exploration and Analysis of Ultra-large Climate Data

Goal *– Develop and apply advanced parallel visualization and analysis software, such as Visit and R for the climate community.*

Approach

- Multidisciplinary team composed of computer and climate scientists.
- Use real climate science problems to drive algorithmic R&D, software development, tool use and deployment.
- Deliver production-quality software infrastructure to the climate science community.

Recent CAM5 0.25° runs can produce upwards of 100TB of model output for 20 simulated years.

Recent Accomplishments

- Cyclone tracking: finds/tracks cyclones on 100TB dataset in 2hrs on 7K processors, compared to 583 days if run on a desktop-class platform.
- Atmospheric river detection/analysis: demonstrated scalability to 3400 CPUs on Advanced Microwave Sounding Radiometer satellite data, 10K CPUs on fvCAM output.
- Spatial Extreme Value Analysis: characterize extremes in precipitation using statistical analysis. New tool uses R, run in parallel on new VisIt+R platform.
- Software infrastructure and engineering:
  - New techniques to parallelize computations over time (along with space).
  - ***Close coordination with UV-CDAT team to deliver new s/w.***

Contact Wes Bethel (ewbethel@lbl.gov) for more information

Earth System Grid Federation (ESGF)

The ESGF consortium's mission is to provide climate researchers worldwide with a system of federated science gateways to access data, information, models, analysis tools, and computational capabilities required to evaluate ultrascale data sets. Its goals are to make data more useful to climate researchers by developing collaborative technology that enhances data usability, and provide a universal and secure Web-based data access portal for broad-based multi-model data collections.

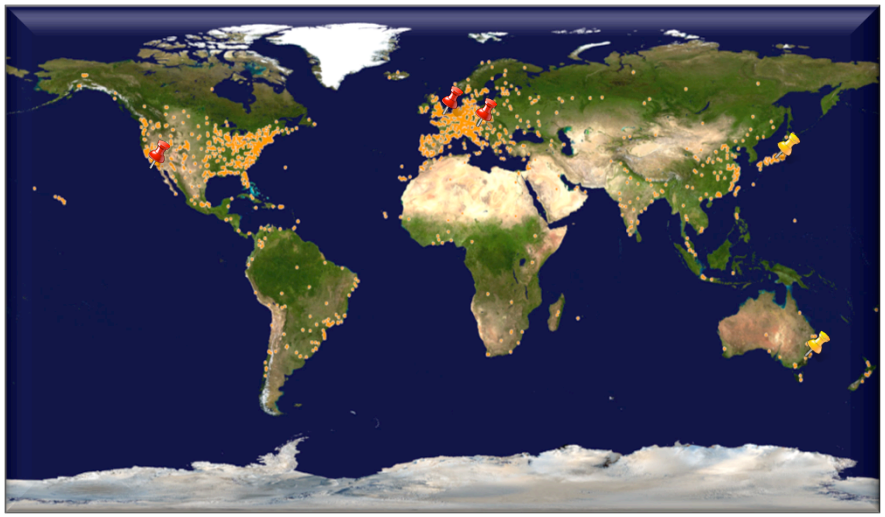
Objectives

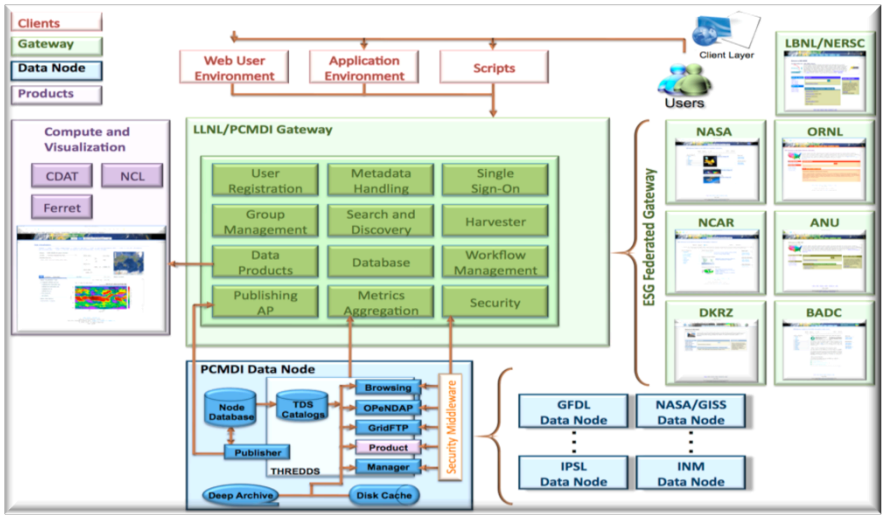
- Meet the specific needs of national and international climate projects for distributed databases, data access, and data movement.
- Provide a wide range of climate data-analysis tools and diagnostic methods to international climate centers and U.S. government agencies.
- **Integrate in a collaborative problem-solving environment highly publicized climate data sets using distributed storage management, remote high-performance units, high-bandwidth wide-area networks, and user desktop platforms.**

Impact

- Massive data archives (PB moving to XB)
- Multiple data centers worldwide
  - **Existing IT infrastructure and separate security domains**
- Heterogeneous data sources (models, observations, reanalysis)
- Multiple physical realms (atmosphere, ocean, land, sea ice)
- Multiple data, metadata formats and conventions
- Multiple scales (global, regional and local)
- Cyber security
- Multiple audiences (scientists, policy makers, students, educators)

Deployment of Climate Data Management and Access



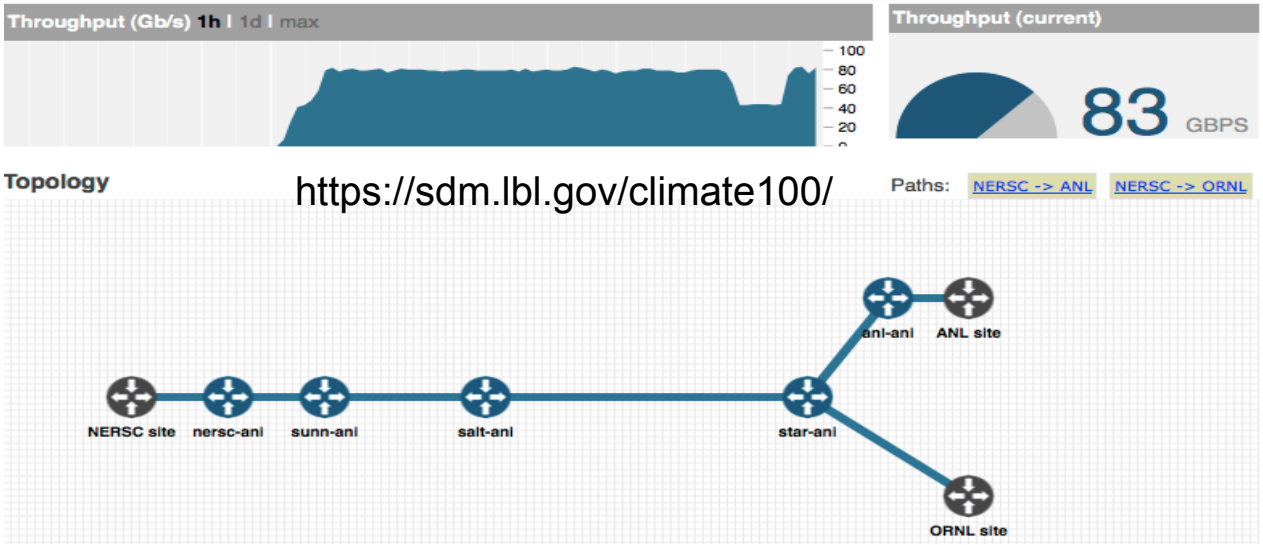


Contact Dean N. Williams (williams13@llnl.gov) for more information or see http://esgf.org/wiki

U.S. DEPARTMENT OF ENERGY | Office of Science

British Atmospheric Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL

Climate100: Scaling Climate Applications to 100Gbps Network



Network bandwidth is increasing!

Problems

- Total size of data is increasing.
- There are many files, relatively small files, in climate data sets.
- It requires efficient methods to fully utilize the underlying network infrastructure with limited resources.

Current State

- The 100Gbps network is in the testing phase.
- Expected to be in production by the end of 2012.

Recent Accomplishment

- Entire IPCC AR4 CMIP-3 (~35TB) is moved under an hour from LBNL/NERSC to ANL and to ORNL over 100Gbps network link.
  - Achieved 83Gbps on average over TCP connections.

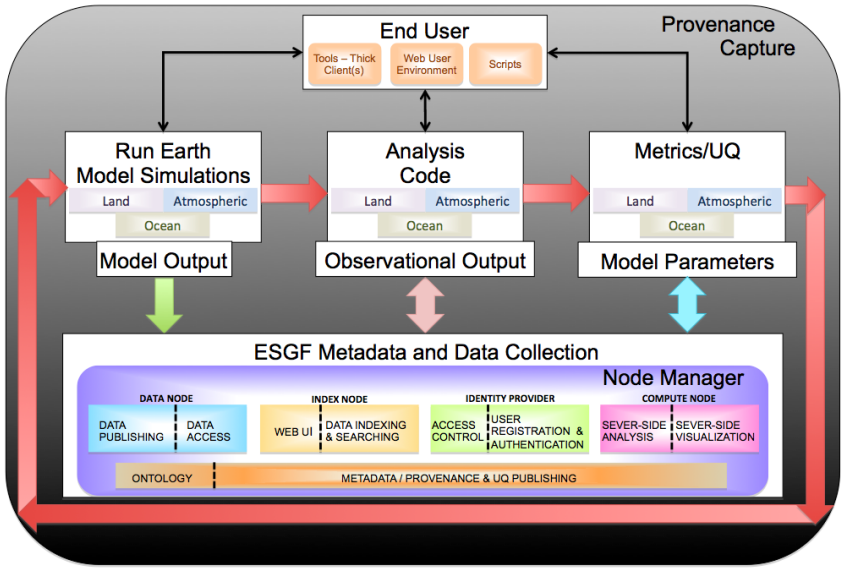
Contact Alexander Sim (Asim@lbl.gov) for more information or see https://sdm.lbl.gov/climate100/ and http://www.es.net/

Climate Science for a Sustainable Energy Future (CSSEF)

Overview

The Climate Science for a Sustainable Energy Future (CSSEF) project [CSSEF 2011] is constructing a test-bed infrastructure that will integrate the existing Earth System Grid Federation (ESGF) Peer-To-Peer (P2P) data-management system [ESGF 2011] with requisite model simulation and analysis codes coupled with methods for capturing metrics and calculating uncertainty quantification (UQ). These components will be linked together with an iterative workflow that will interface with end users via thick client, web client, and script management capabilities.

CSSEF Architecture



High-level conceptual view of the CSSEF test bed architecture and workflow. Working for all model components (atmosphere, land, and ocean), provenance capture is pervasive throughout the test bed, capturing and saving all user actions. The red arrow lines show the baseline ensemble loop in which Earth model simulations are conducted using variety of model input parameters generated by Metrics and UQ ensemble drivers. At any stage, data can be collected and stored in the ESGF distributed archive. The black arrow lines show access to the test bed is attained through desktop clients, web browsers, or scripts.

Future

Contact David Bader (bader2@llnl.gov) and Dean N. Williams (williams13@llnl.gov) for more information

Other AGU Session

Global Environmental Change – *Climate Modeling: Innovative Application of Observations for Diagnosing CMIP5 and IPCC Simulations: Quantifying Model Processes and Uncertainties I Posters*

Earth and Space Science Informatics – *Challenges in Analysis and Visualization of Large Earth Science Data Sets I*

Next week, presentations can be found at the following URLs:
http://esgf.org/wiki and http://uv-cdat.org/wiki

Discussion questions: 5 year timeframe

- Model Development -> Data Transfer -> Validation -> Analysis, Intercomparison
  - Where are the bottlenecks?
  - What software is lacking (e.g. for handing high or variable resolution data)?
  - What analysis tools are needed?

- In what ways could DOE's support of model intercomparison activities better serve
  - Model development
  - Model evaluation
  - Impacts and vulnerability studies
  - …..
- Of DOE's climate modeling software development efforts, which are most critical?
  - High-speed data sharing and transfer
  - Enhanced visualization and analysis tools
  - Distributed analysis performed remotely across multiple data archives
  - Automated recording of analysis procedures (provenance, reproducibility)
  - Framework development for Earth System Models
  - …..