The Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT): A Vision for Large-Scale Climate Data

?

Hank Childs (LBNL) and Charles Doutriaux (LLNL)

September 22, 2011

On behalf of the UV-CDAT and Data Explorer Teams

10-05 team responsible for developing and deploying UV-CDAT

10-05 team responsible for developing and deploying large data climate analysis

Lawrence Livermore National Laboratory

Los Alamos NATIONAL LABORATORY EST. 1943

Kitware

OAK RIDGE National Laboratory

SCI www.sci.utah.edu

NYU·poly

BERKELEY LAB

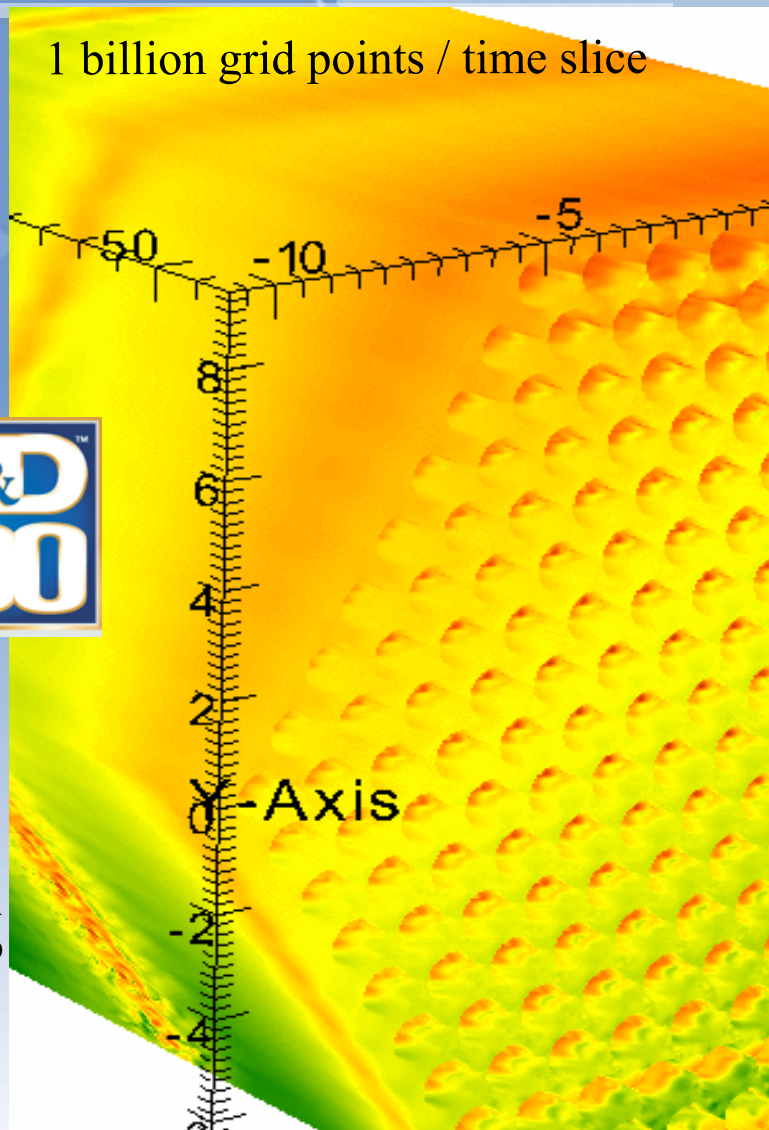
NASA

UV-CDAT

- **UV-CDAT = Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT)**
- **Goal: robust tool, capable of doing powerful analysis on large climate data sets.**
- **Collaboration between two 10-05 teams**
- **One tool that incorporates many packages…**
  - **… including “VisIt”** ← Focal point of Data Explorer team

VisIt is an open source, richly featured, turn-key application for large data.

- **Used by:**
  - **Visualization experts**
  - **Simulation code developers**
  - **Simulation code consumers**
- **Popular**
  - **R&D 100 award in 2005**
  - **Used on many of the Top500**
  - **>>>100K downloads**
- **Developed by:**
  - **NNSA, SciDAC, NEAMS, NSF, and more**



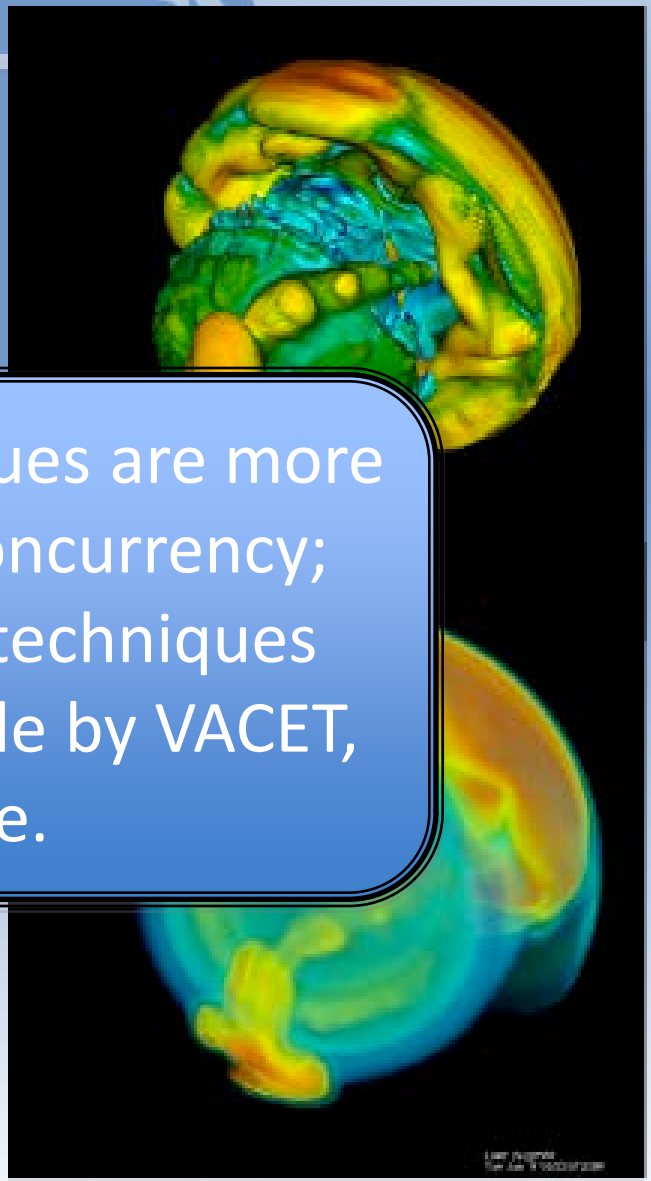
217 pin reactor cooling simulation
Run on ¼ of Argonne BG/P
Image credit: Paul Fischer, ANL

VisIt recently demonstrated good performance at unprecedented scale.

- **Weak scaling study: ~62.5M cells/core**

Machine	Model	Problem Size	#cores
Franklin			
Dawn			
JaguarPF			
Juno			
Purple			
Ranger			

VisIt's data processing techniques are more than scalability at massive concurrency; we are leveraging a suite of techniques developed over the last decade by VACET, the NNSA, and more.



Two trillion cell data set, rendered in VisIt by David Pugmire on ORNL Jaguar

UV CDAT

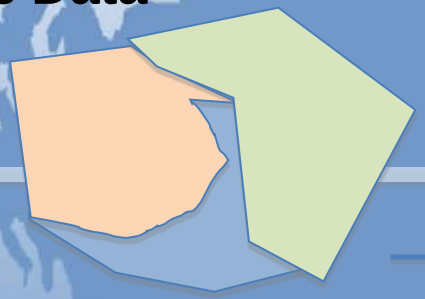
Production visualization tools use "pure parallelism" to process data.

This is a good approach for high resolution meshes
… but climate data is different

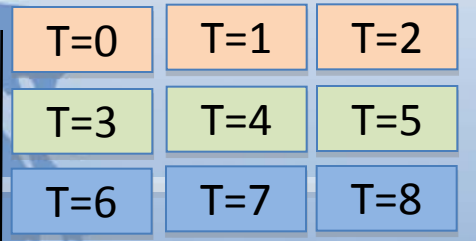
Parallelizing Processing Over Time Slices: Improving Performance For Climate Data

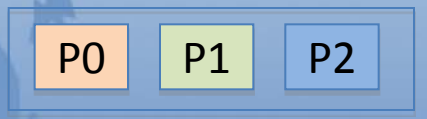
Objective

VisIt's parallel processing techniques were designed for single time slices of very high resolution meshes. We must adapt this approach for the lower resolution and high temporal frequency characteristic of climate data.



VisIt parallelizing over a high resolution spatial mesh





VisIt parallelizing temporally over a low resolution mesh

Progress & Results

- We have modified VisIt's underlying infrastructure to have a "parallelize over time" processing mode.
  - We implemented a simple algorithm ("maximum value over time") as a proof of concept

Concurrency	Average Time	Speedup
1	480.0s	-
2	223.0s	2.1X
4	120.0s	4.0X
8	62.0s	7.8X
16	29.0s	16.5X
32	15.5s	31.0X
64	7.8s	61.5X
128	4.2s	114X

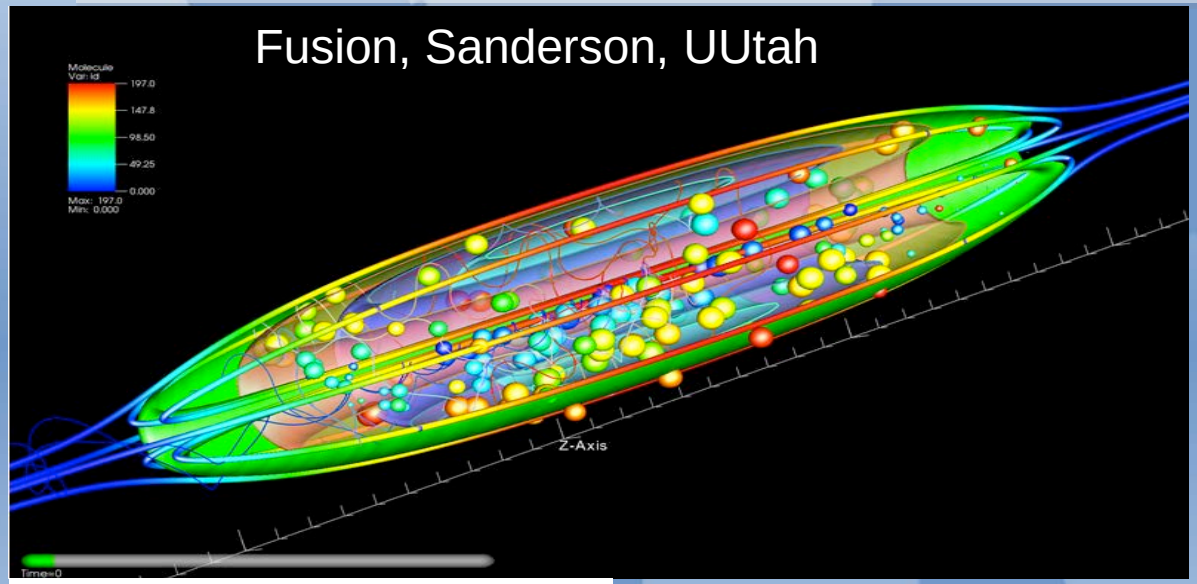
Scaling on 2130 time slices of NetCDF climate data (source: Wehner)

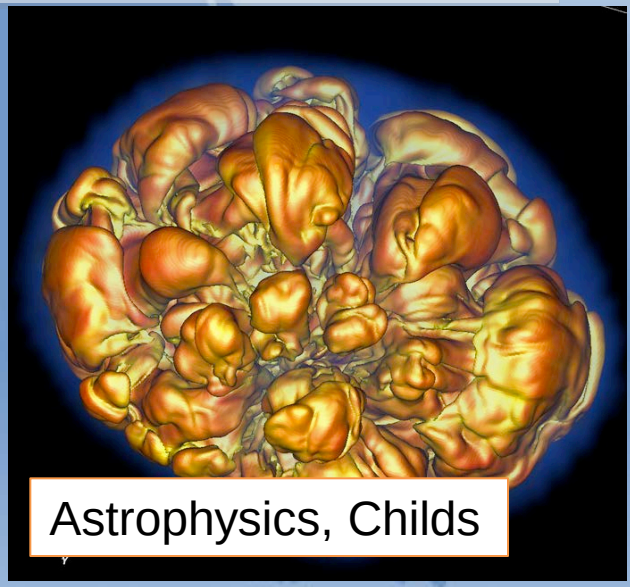
Impact

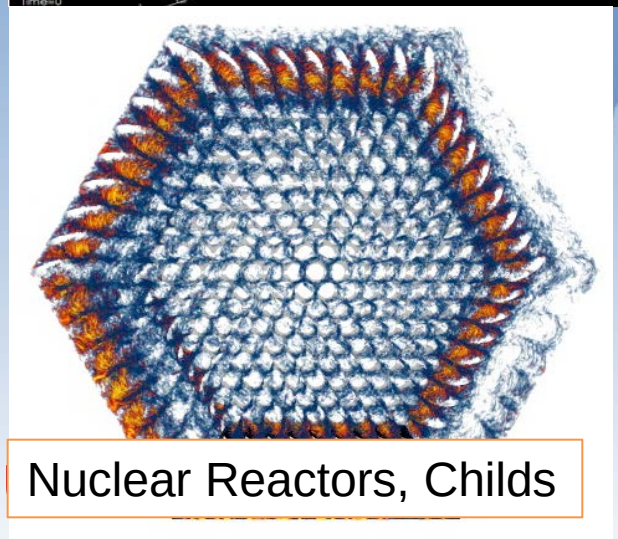
Climate scientists with access to parallel resources for their data will be able to process their data significantly faster through parallelization. This software investment will enable other algorithms developed by the project to also be accelerated through parallelization.

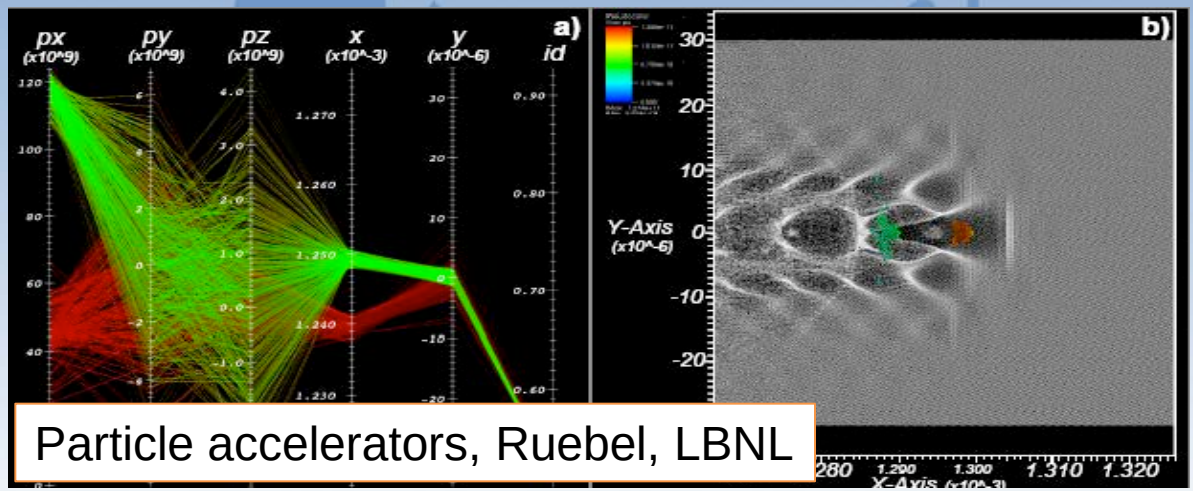
UV-CDAT

VisIt is used to look at simulated and experimental data from many areas.









General-purpose tools vs application-specific tools

General-purpose tools:

- Are developed by large teams, leading to:
  - Robustness
  - Efficient algorithms
  - Rich set of features
- But:
  - They aren't streamlined for a given appl
  (lots of bu
  - They don'
  applicatio
  methods

Application-specific tools:

- Are made specifically to solve your problem:
  - Streamlined user interface
  - Application-specific analysis
- But, they have smaller user and developer communities,

ust

cient algorithms

set of features

So which do users want?

Amazing developments over the last decade…

- Very useful packages now available that make quick tool development possible:
  - Python, R, VTK, Qt
- But this doesn't solve the large data issue.
  - … but tools now are available that do that as well:
    - VisIt & ParaView
- Great idea: put all these products together into one tool. VisTrails
- Users get:
  - The robustness, richness, and efficiency of large development efforts
  - Streamlined user interface & climate-specific analysis (via CDAT)

This tool is UV-CDAT.

UV-CDAT

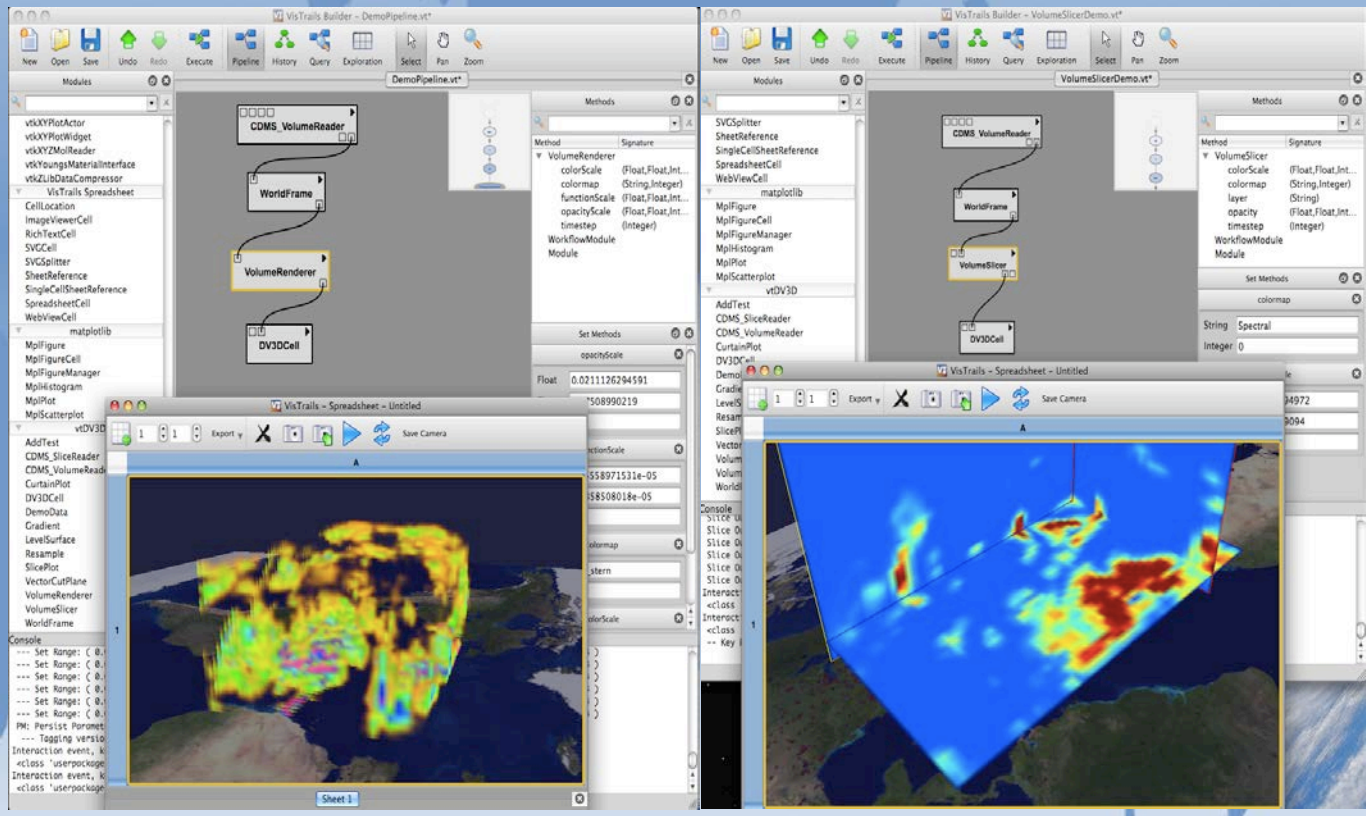
CDAT

- Designed for climate science data, CDAT was first released in 1997
- Based on the object-oriented Python computer language
- Added Packages that are useful to the climate community and other geophysical sciences
  - Climate Data Management System (CDMS)
  - NumPy / Masked Array / Metadata
  - Visualization (VCS, IaGraphics, Xmgrace, Matplotlib, VTK, Visus, etc.)
  - Graphical User Interface (VCDAT)
  - XML representation (CDML/NcML) for data sets
- One environment from start to finish
- Integrated with other packages (i.e., LAS, OPeNDAP, ESG, etc.)
- Community Software (BSD open source license)
- URL: http://www-pcmdi.llnl.gov/software-portal (CDAT Plone site)

UV-CDAT

ParaView: vtDV3D, Scientific Data Analysis and Visualization

- High level analysis and visualization modules for UV-CDAT workflows.
- Encapsulate complex data visualization and processing operations.
  - At a level of complexity appropriate for scientists.
- Interactive configuration enabling data exploration (with provenance).

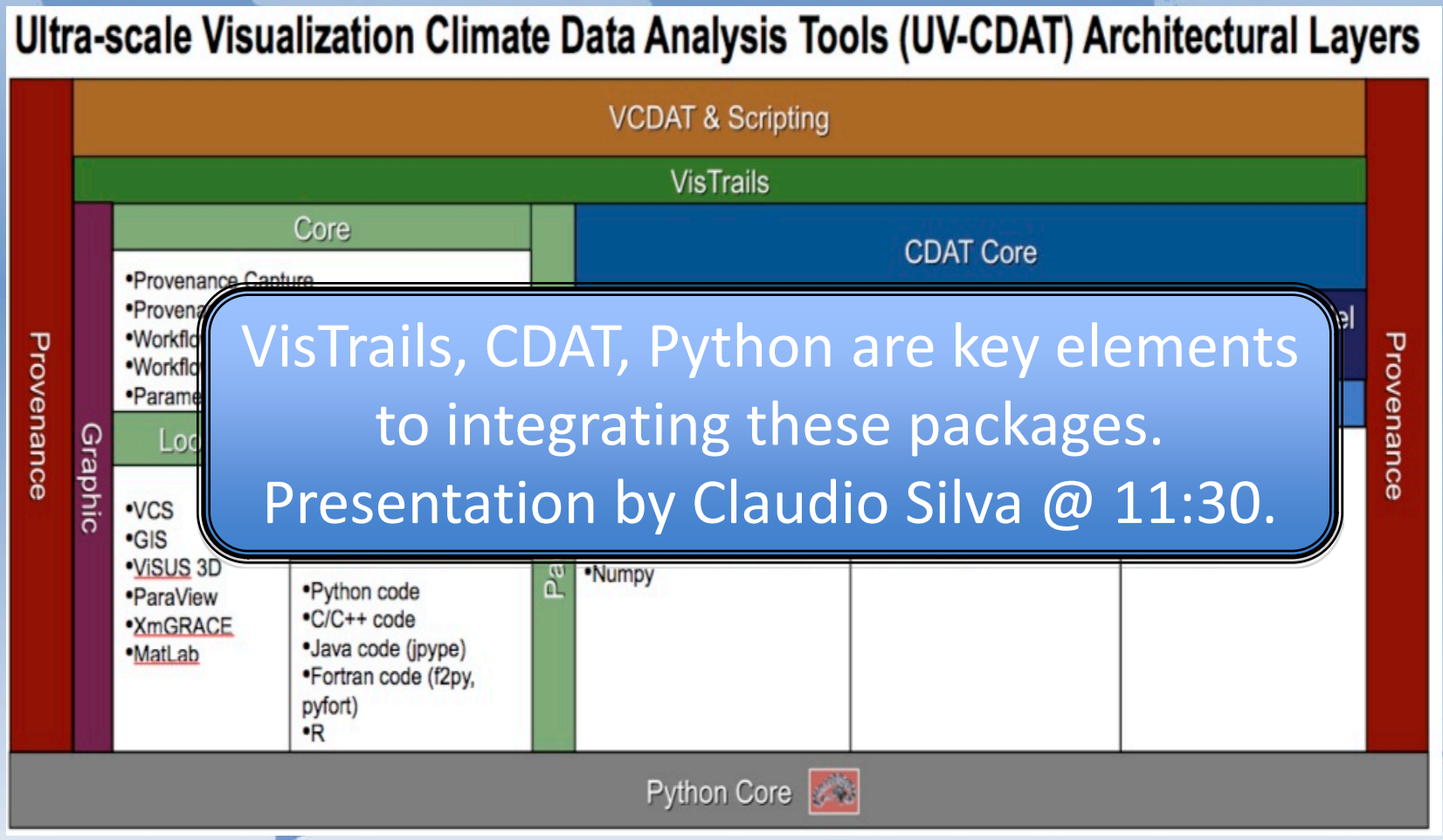


VisTrails

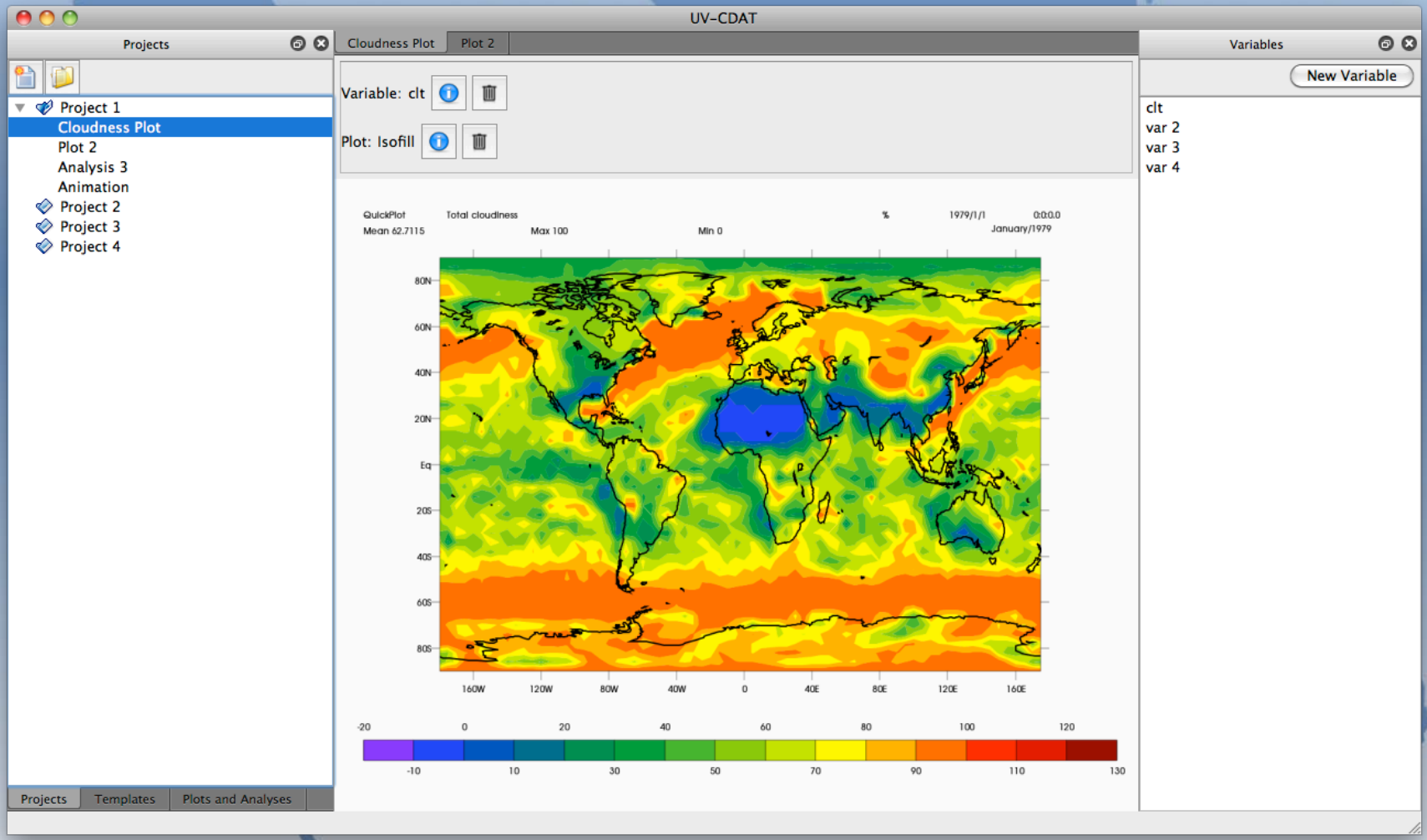
VisTrails is an open-source scientific workflow and provenance management system developed at the University of Utah that provides support for data exploration and visualization. Whereas workflows have been traditionally used to automate repetitive tasks, for applications that are exploratory in nature, such as simulations, data analysis and visualization, very little is repeated---change is the norm. As an engineer or scientist generates and evaluates hypotheses about data under study, a series of different, albeit related, workflows are created while a workflow is adjusted in an interactive process. VisTrails was designed to manage these rapidly-evolving workflows.

A key distinguishing feature of VisTrails is a comprehensive provenance infrastructure that maintains detailed history information about the steps followed and data derived in the course of an exploratory task: VisTrails maintains provenance of data products, of the workflows that derive these products and their executions. This information is persisted as XML files or in a relational database, and it allows users to navigate workflow versions in an intuitive way, to undo changes but not lose any results, to visually compare different workflows and their results, and to examine the actions that led to a result. It also enables a series operations and user interfaces that simplify workflow design and use, including the ability to create and refine workflows by analogy and to query workflows by example.

UV-CDAT Architecture Layers



Integrated UV-CDAT GUI: Project, Plot, and Variables View



Lots of work to do for Data Explorer effort…

- Lots of software engineering to get pieces working together:
  - Integration of VisIt and R
  - Improved integration of VisIt and VisTrails
  - Connect VisIt to UV-CDAT
  - (Upgrade to VTK 5.6)
  - Incorporate analysis work described by Wes (cyclone detection, atmospheric rivers)

UV-CDAT

Summary

- UV-CDAT: tool being developed by two 10-05 teams.
  - Goal: robust tool, capable of doing powerful analysis on large climate data sets.
  - Its development is significantly leveraged by many existing packages.

- The Data Explorer team is focusing on deploying VisIt and R as part of UV-CDAT.
  - VisIt is excellent with large data and has been adapted to work with climate data.
  - Our SWE effort is now to integrate with UV-CDAT
  - We also are collaborating on cutting edge analysis techniques with climate scientists (Bethel talk)

UV-CDAT

Example climate visualizations with VisIt