A UT/ORNL PARTNERSHIP
NATIONAL INSTITUTE FOR COMPUTATIONAL SCIENCES
NICS

# RDAV
# Center for Remote Data Analysis and Visualization

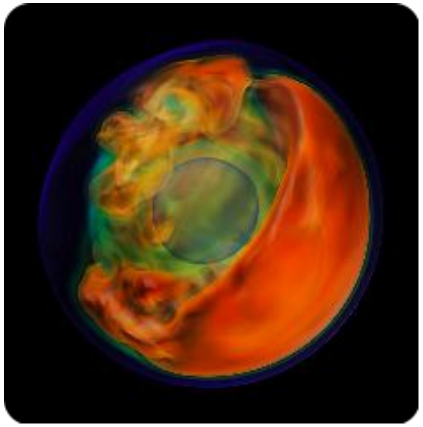
Sean Ahern

in collaboration with

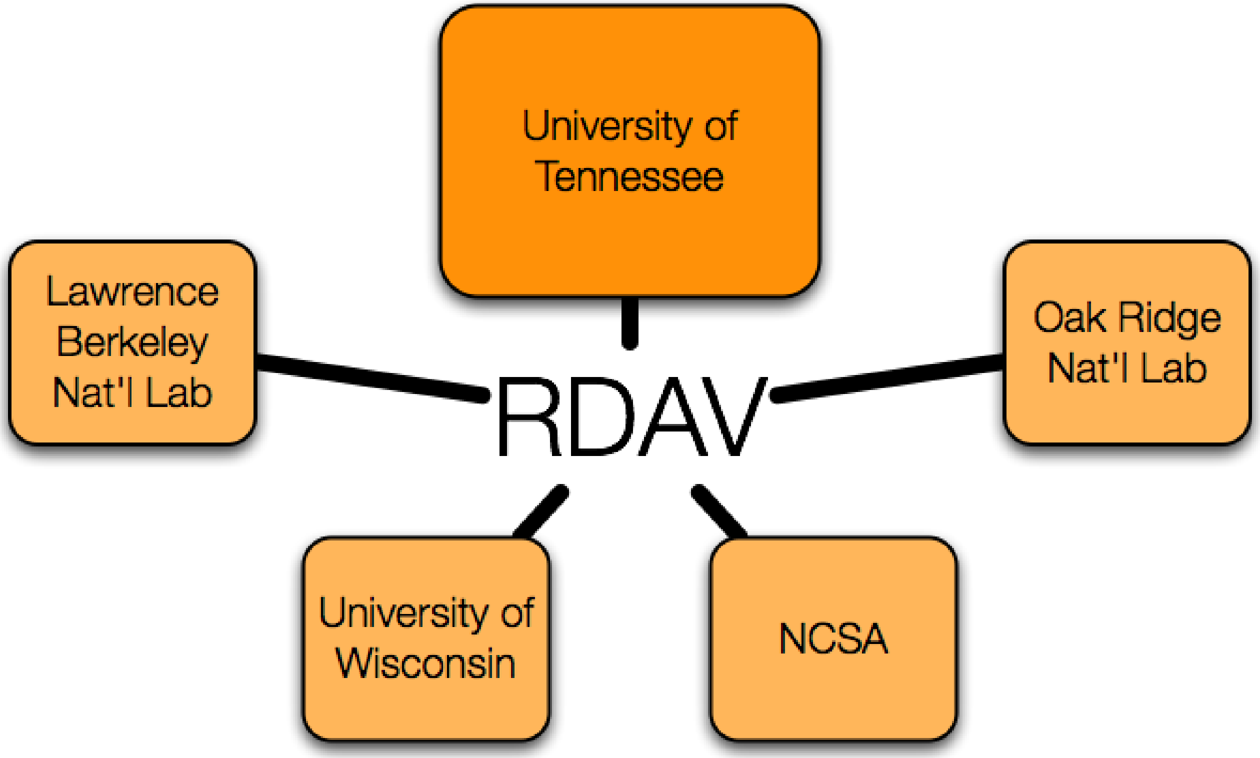
Jian Huang – University of Tennessee, Wes Bethel – LBNL, Scott Klasky – ORNL, Dave Semeraro – NCSA, George Ostrouchov – ORNL, Miron Livny – University of Wisconsin

# Providing analysis services for XSEDE users

- RDAV: The Center for Remote Data Analysis and Visualization
- Provide remote and shared resources for the purpose of exploring/analyzing/visualizing large scale data
- Provide the ability to easily take advantage of remote and shared computing/data storage infrastructure
- Provide unique architecture for data analysis and visualization
- Leverage large amount of existing experience in deploying similar capabilities
- Allocated through XRAC

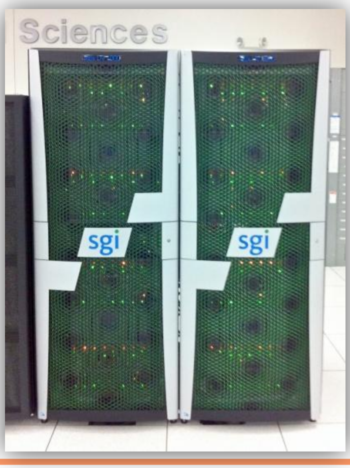


# RDAV is a partnership between 5 institutions



# Diverse use cases dictate SMP architecture

- Many HPC users can use distributed memory analysis:
  - Data parallel, space/time parallel
- However, many general and statistical analysis algorithms favor large shared memory
  - Document clustering/searching
  - Generalized graph structures
  - Bioinformatics, genomics



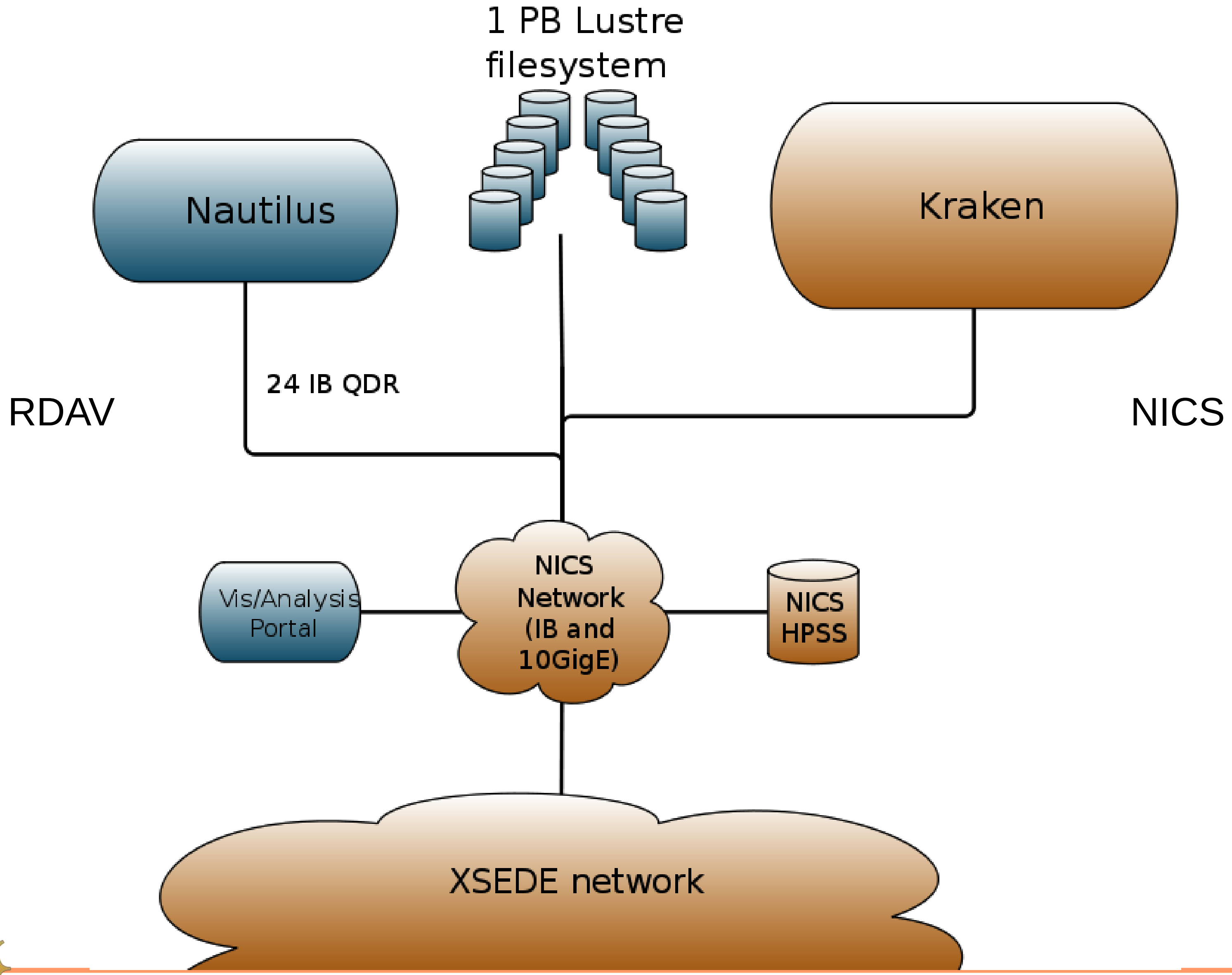
- Large shared memory is the only reasonable way to address all of these needs
- University of Tennessee partnered with SGI to site an Altix UV1000 machine
  - Large memory single-system image through NUMA
  - A “better” cluster architecture, accelerating distributed memory MPI
- NSF XSEDE resource dedicated to analysis and visualization
  - 1024 cores (Intel Nehalem EX)
  - 4 TB Global Shared Memory
  - 8 NVIDIA Fermi Tesla GPUs

# Systems housed at National Institute for Computational Sciences (NICS)

- NICS is a collaboration between UT and ORNL
- Major partner in the $121M XSEDE effort.
- Awarded the NSF Track 2B
- Home of Kraken

THE UNIVERSITY of TENNESSEE
OAK RIDGE National Laboratory
NSF
NATIONAL INSTITUTE FOR COMPUTATIONAL SCIENCES
5100

# How Nautilus fits into NICS

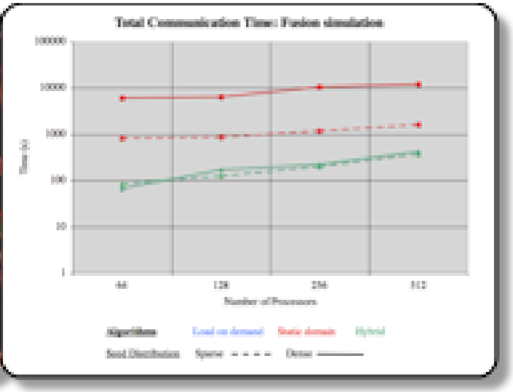


# We provide a range of software services

- Analysis applications to be dictated by user needs and technology needed to solve user problems. “Whatever it takes!”
- Remote visualization and image generation
  - Provide interactive and batch image generation tools (gnuplot, ImageMagick, etc.)
  - Remote parallel visualization (VisIt, ParaView, yt, etc.)
  - Tools for custom app development
- Data analysis and statistical analysis
  - Octave, Parallel R, Matlab, etc.
- Workflow systems
  - DAGMan and Kepler systems automate actions on behalf of users
  - Use is increasing and many users wish to explore
- Portal system
  - Builds upon standard Liferay platform
  - Provides SAS services for analysis and visualization

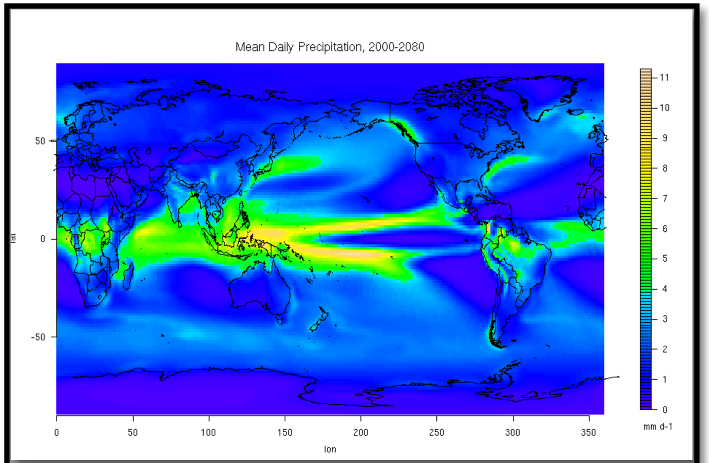
# We provide scalable visualization tools

- Scalable tools: VisIt, ParaView, yt
  - VisIt great at exploiting large memory and large core count systems
  - ParaView great at scalable rendering
  - yt works with data from many astrophysics codes, such as Enzo
- Serial tools (gnuplot, ImageMagick)

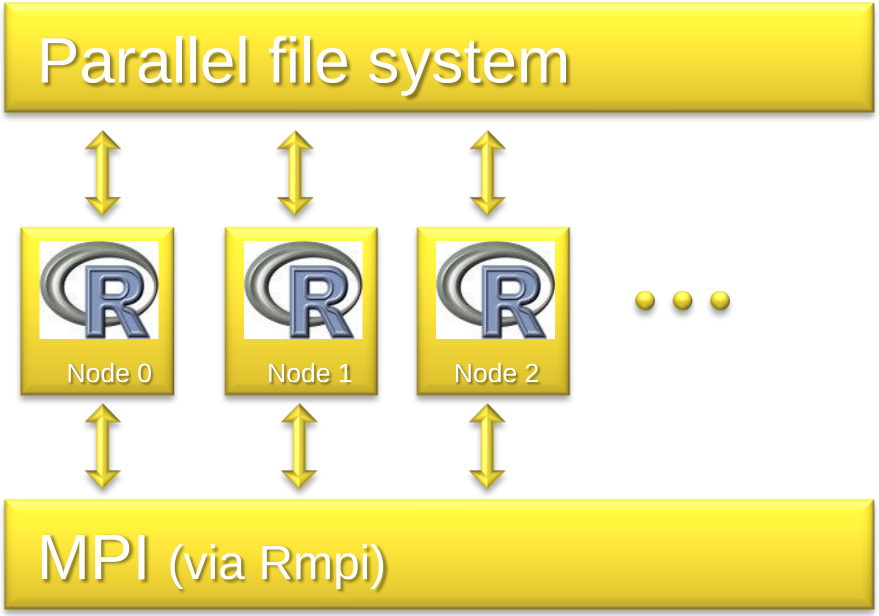


- Remote "vis" delivery software
  - Some high-end packages already have a client-server architecture, so those can be used in "remote vis mode" today with no special third-party software
  - Some packages have GUIs and displays that we'll want to provide remote access to. We can do so using either protocol accelerators (NX, VirtualGL) or remote desktop software (NX, VNC). These two approaches have some overlap in capability and functionality

# We provide scalable statistical analysis tools



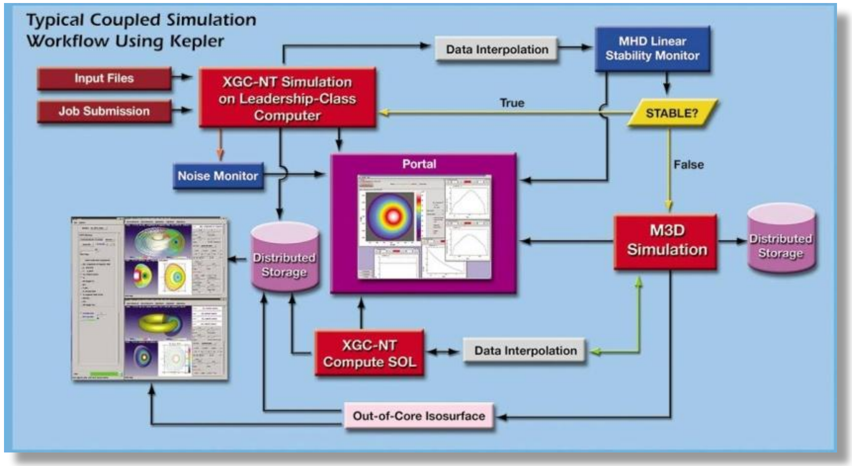
Core 0 produces graphics files



- Data parallel R leverages success of scalable visualization
- Widely used open source statistical analysis
- Recently used to provide interactive analysis of 100 GB climate dataset

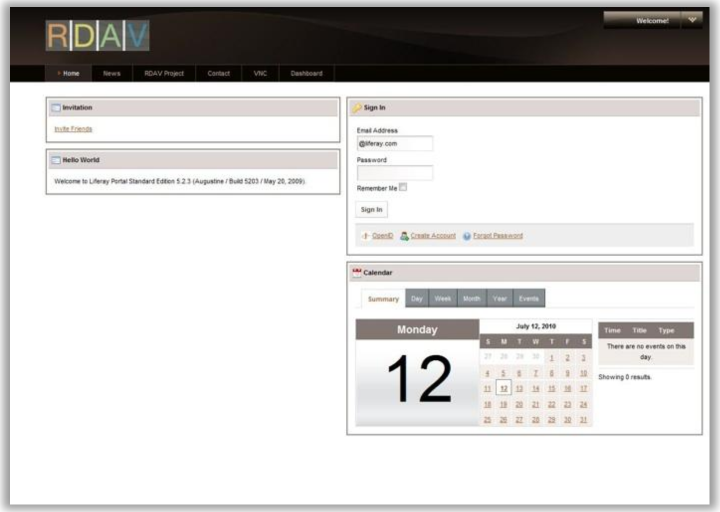
# We provide scientific workflow management tools

- Help in the construction and automation of scientific problem-solving processes that include executable sequences of components and data movement
- Scientific workflow systems often need to provide for load balancing, parallelism, and complex data flow patterns between servers on distributed networks
  - Aiming to solve complex scientific data integration, analysis, management, visualization tasks
  - Error checking and retry
  - Maximizes compute resources / human time
  - “Launch and forget!”
- Deployed through DAGMan or Kepler



# RDAV will provide portals for science gateways

- Portals give developers a chance to create collaborative environments that allow researchers or any community to share knowledge, analyze data, and solve problems
  - Common area to combine users, resources, services
    - Portlets for chat, e-mail, forums, group invites by e-mail, etc.
    - Plug-ins to provide rich services for research communities
    - Custom portlets to access backend services
- Accessible through any browser
- Using the Liferay framework, adopted by XSEDE as a whole



Welcome to Liferay Portal Standard Edition 5.2.3 (Augustine / Build 5203 / May 20, 2009).

Email Address: @liferay.com

Password:

Remember Me □

Monday

12

July 12, 2010

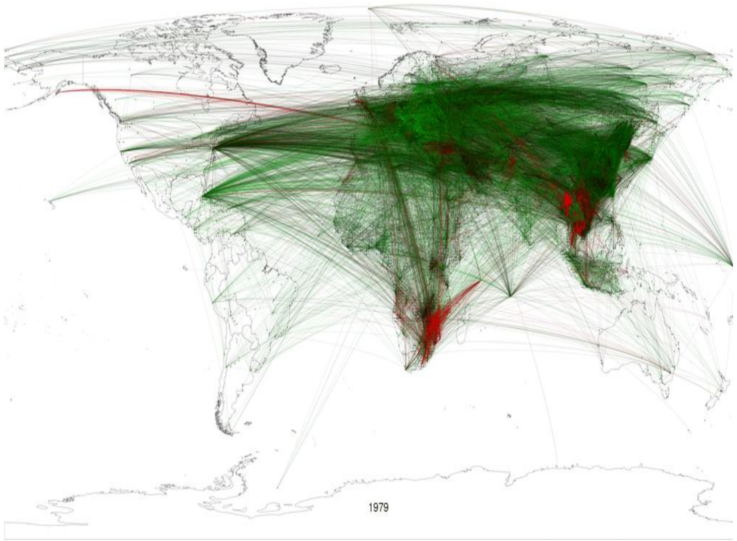
| S | M | T | W | T | F | S |
|---|---|---|---|---|---|---|
| 27 | 28 | 29 | 30 | 1 | 2 | 3 |
| 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 11 | 12 | 13 | 14 | 15 | 16 | 17 |
| 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| 25 | 26 | 27 | 28 | 29 | 30 | 31 |

| Time | Title | Type |
|---|---|---|
| There are no events on this day. | | |

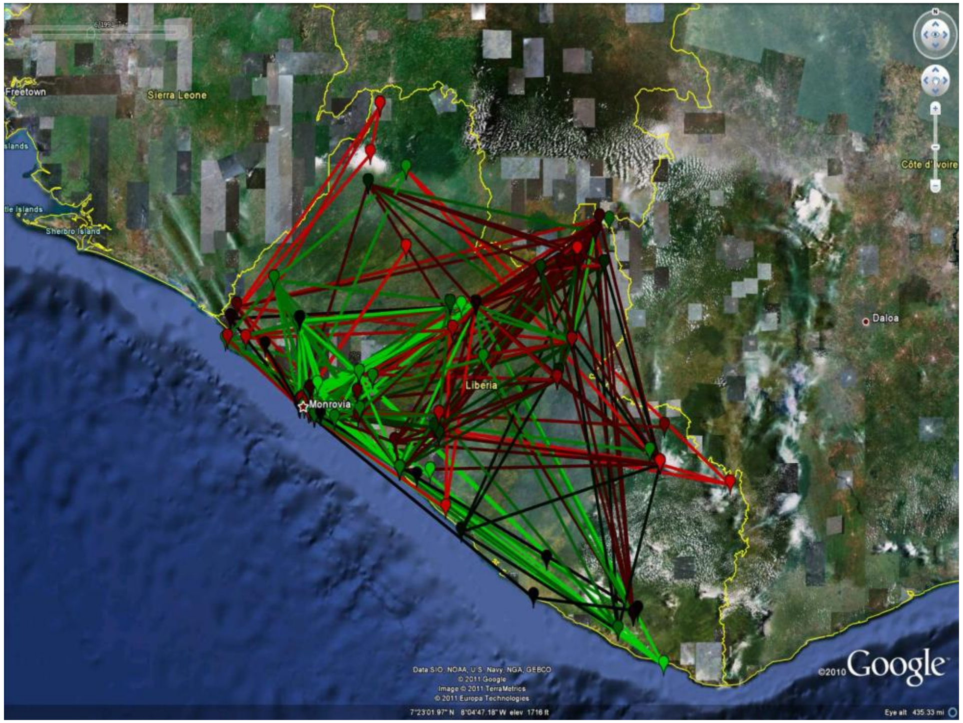
Showing 0 results.

# Work with users

**Kalev Leetaru** of the University of Illinois creates a knowledge network based on information taken from databases of newspaper articles and other publicly available information to understand relationships between geographic regions. These analyses have identified patterns that describe the “Arab spring” of 2011 and that can be used to understand future events.



Cooperation and competition between cities, based on new articles from 1979



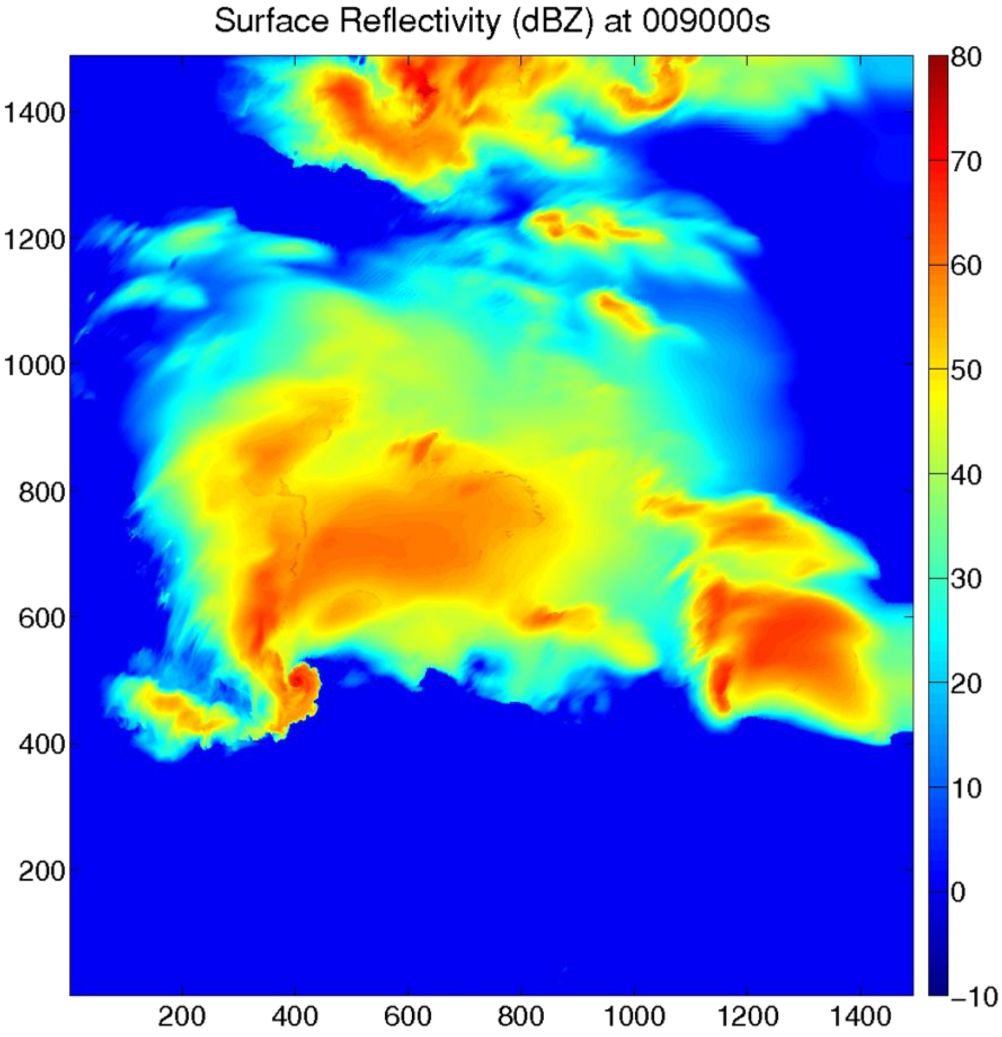
Connections between cities in Pakistan

# Work with users

**Amy McGovern** of the University of Oklahoma uses Kraken to produce high resolution simulations of severe thunderstorms.

These simulations yield roughly 1 TB of data each.

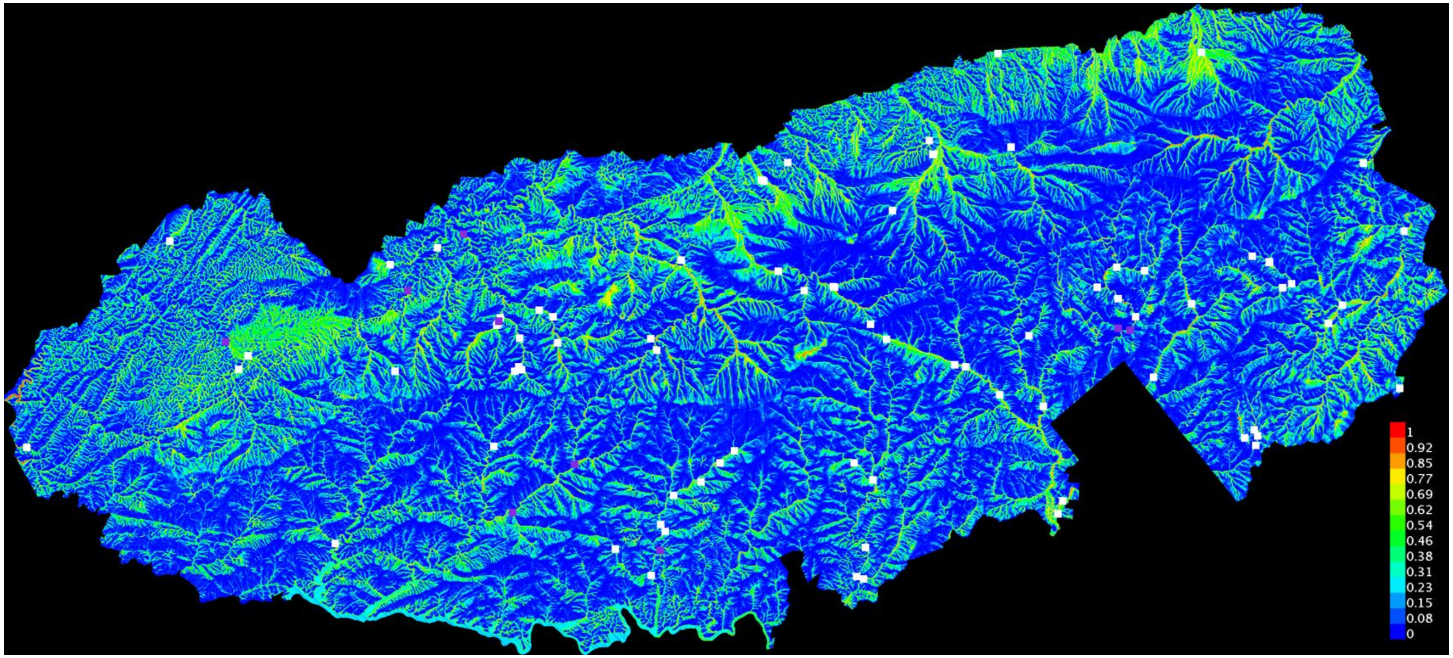
In order to understand the results of the simulations, the researchers use Nautilus to carry out data mining to identify patterns that lead to the formation of tornadoes. They then use MATLAB on Nautilus to visualize the results.



Formation of a supercell thunderstorm

# Work with users

**Lou Gross and Will Godsoe** of the National Institute for Mathematical and Biological Synthesis (NIMBioS) are studying statistical ecology. They performed comparative analyses of over 6000 grids of output from statistical analyses of biodiversity data from the Great Smoky Mountains National Park. RDAV developed tools to allow them to perform parameter sweeps with legacy serial code and then to use the VisIt visualization tool to visualize their results.



Probability of presence of the Greenbrier Springfly. White squares are recorded observation points, and purple squares are sample points that were left out for crossvalidation. This plot uses ten layers of environment data in the prediction.

# Education, outreach, and training activities

- We taught a visualization class at the Petascale Programming Environments and Tools classes in July 2011
- We presented tutorials on the SGI Altix UV 1000 and on OpenMP at the TeraGrid ’11 conference in July 2011
- We gave a hands-on visualization workshop at the NSF/PRACE summer school in Lake Tahoe in August 2011
- We taught a Fall 2011 seminar on data analysis and visualization to freshman at the University of Tennessee

# Contact

**Sean Ahern**
Director, Center for Remote Data Analysis and Visualization
ahern@utk.edu