BERKELEY LAB
LAWRENCE BERKELEY NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

Large-Scale Scientific Data Management and Visualization

Juan Meza

Department Head and Senior Scientist
High Performance Computing Research
Lawrence Berkeley National Laboratory

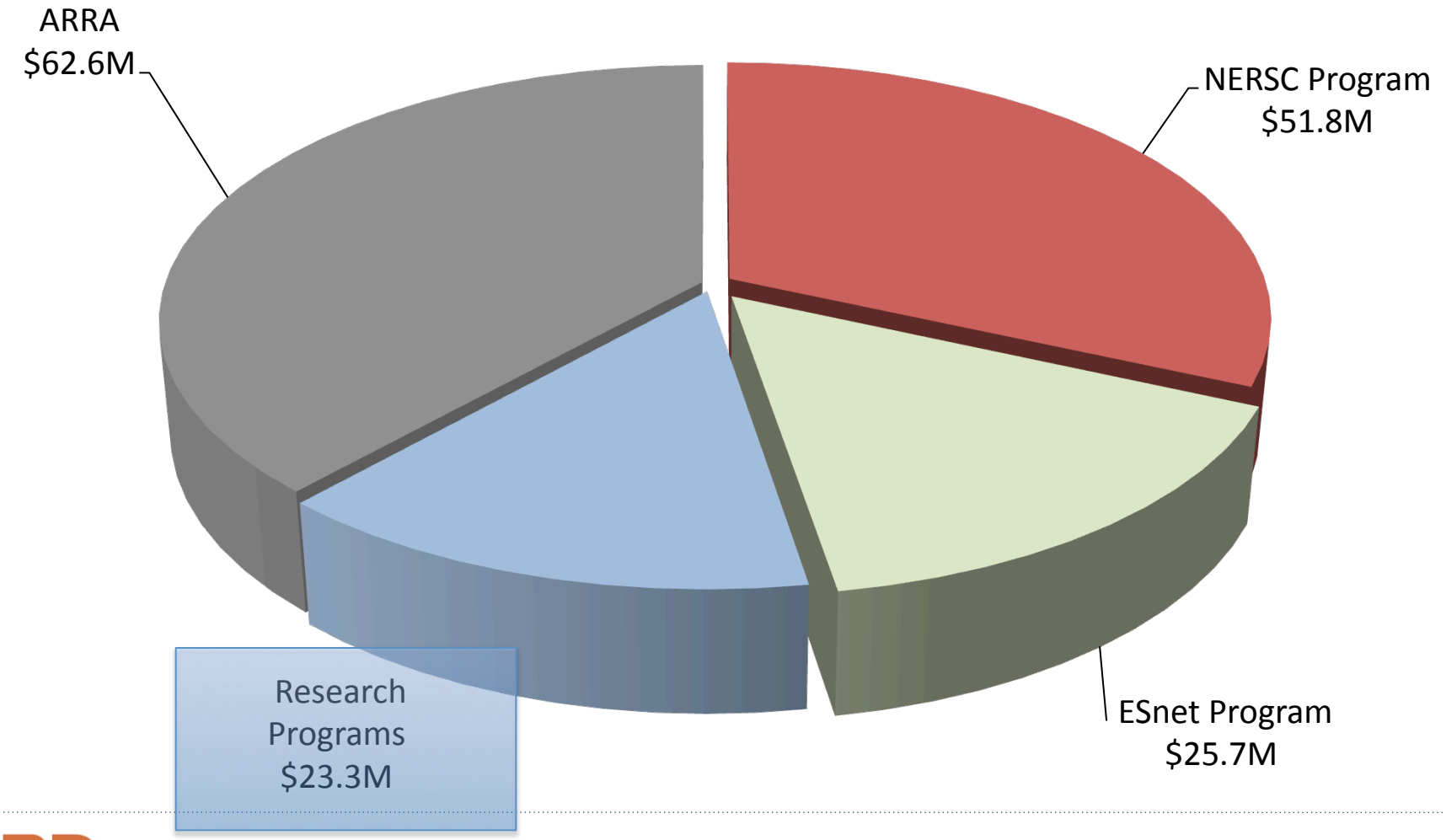
NETL, Morgantown, WV, September 22, 2009

Computing Sciences at LBNL Mission

- **Deliver world class facilities, NERSC and ESnet, supporting the DOE Office of Science computational mission**
- **Conduct world-leading research in applied mathematics and computer science in support of DOE science mission**
- **Build and maintain an outstanding computational science and engineering (CSE) research effort in close collaboration with other divisions at the lab and the UC campuses**

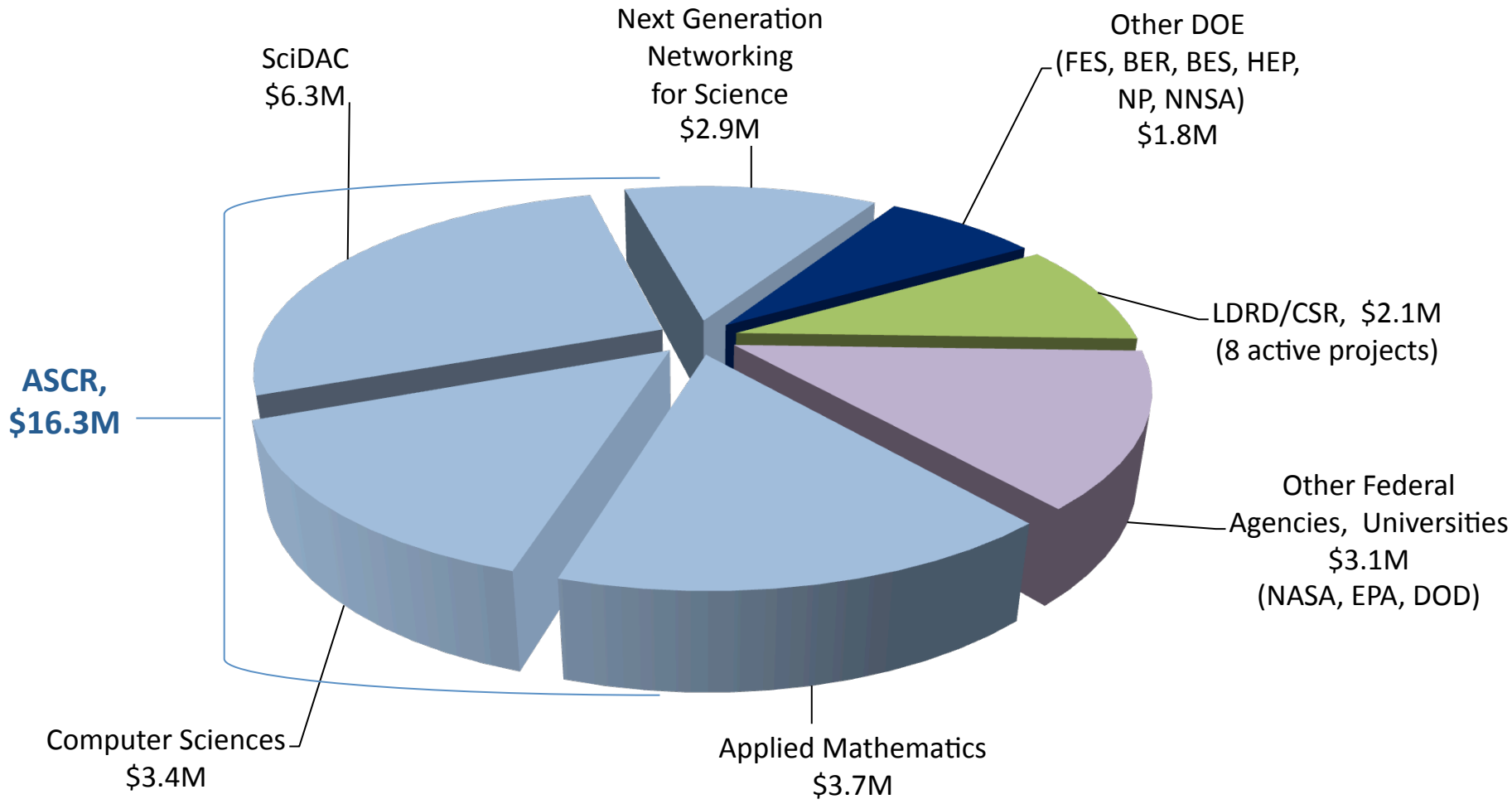
Computing Sciences Program

$163.4M Annual Budget for FY09 as of 8/1/09



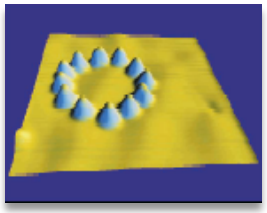
Research Portfolio

$23.3M in FY09



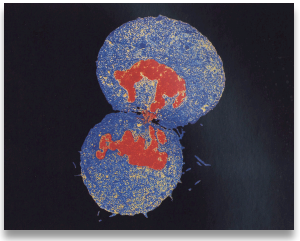
Computational Science Mission

energy technology

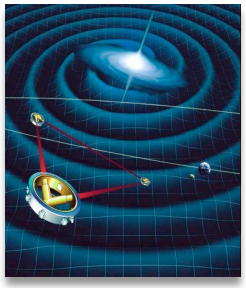


nano systems

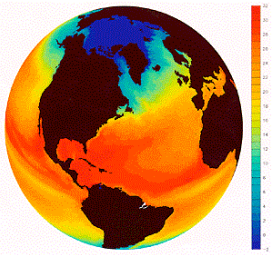
biological systems



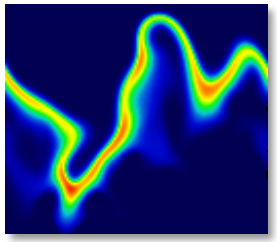
The Computational Research Division is engaged in computational science collaborations, creating tools and techniques that will enable computational modeling and simulation, and lead to new understanding in areas such as



astrophysics simulation



global climate



combustion processes

Finding vulnerabilities in a complex system

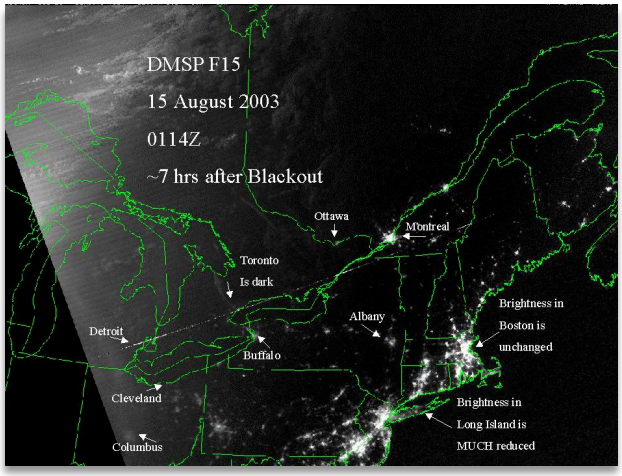
2003 NE US Blackout demonstrated vulnerability of power grid.

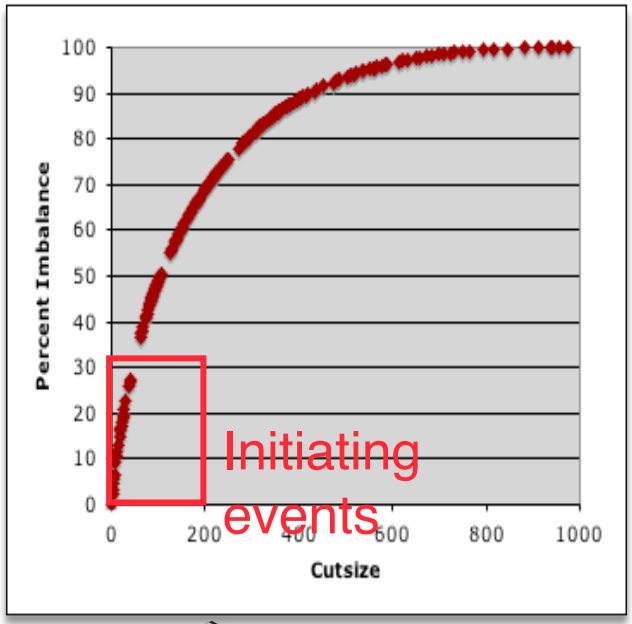
Applied for an LDRD in 2004 and obtained funding from 2005 – 2007 to investigate the application of combinatorial optimization and no-linear optimization for large scale problem.

Can now analyze vulnerabilities of large systems in minutes.

Solutions with small cut size can be used to detect initiating events and groups of vulnerabilities.

Received $875K DOE funding in 2009 (PI Meza).



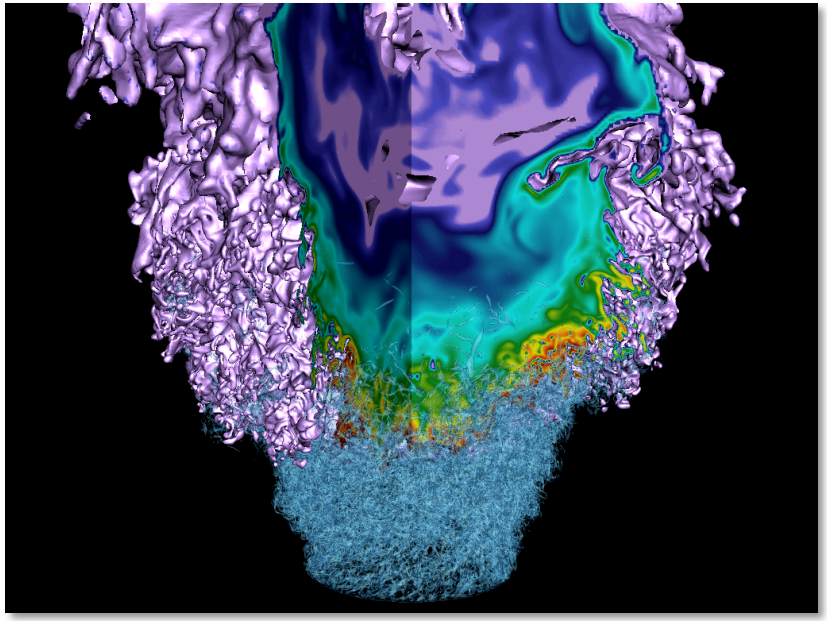


Transformative Mathematics and Computer Science

Leverage expertise in applied mathematics, computational methods and algorithms and apply them to science and engineering problems throughout DOE

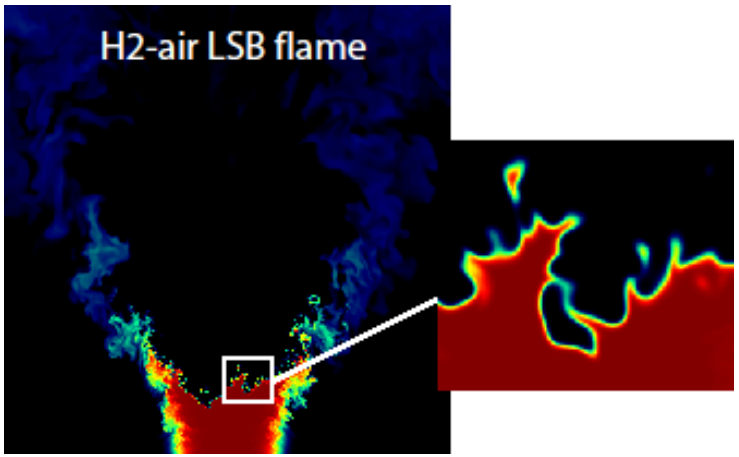
Potential to transform research in:

- **Extreme climate prediction**
- **Energy efficiency**
- **Design new photovoltaic materials**
- **Addressing tokamak plasma instabilities**
- **Carbon sequestration strategies**
- **Nuclear reactor safety**
- **Improving efficiency, reliability of nation's power grid**

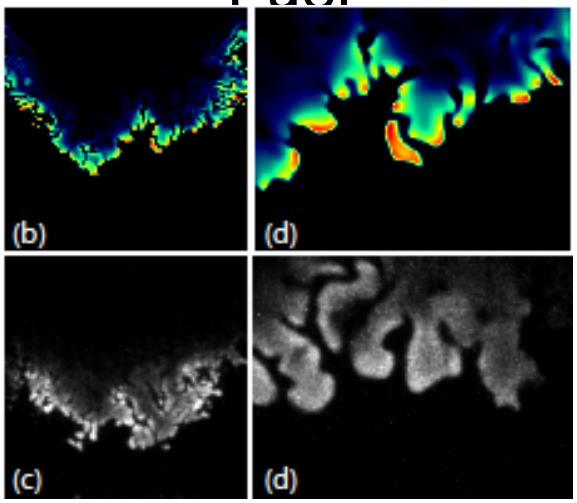


Simulation of lean premixed hydrogen flames stabilized on a low-swirl burner

- Low Mach number formulation exploits mathematical structure of the problem
  - Advanced numerical methodology, including projection methodology, adaptive mesh refinement, and parallel implementation using BoxLib
  - Detailed chemistry and transport
  - No explicit models for turbulence or turbulence / chemistry interaction
  - 25 cm x 25 cm x 25 cm
- Combined methodology enables simulation at effective resolution of 8B cells (2048^3)
- Simulation captures cellular structure of thermodiffusively unstable lean hydrogen flames and provides insight into experimental diagnostics



Fuel

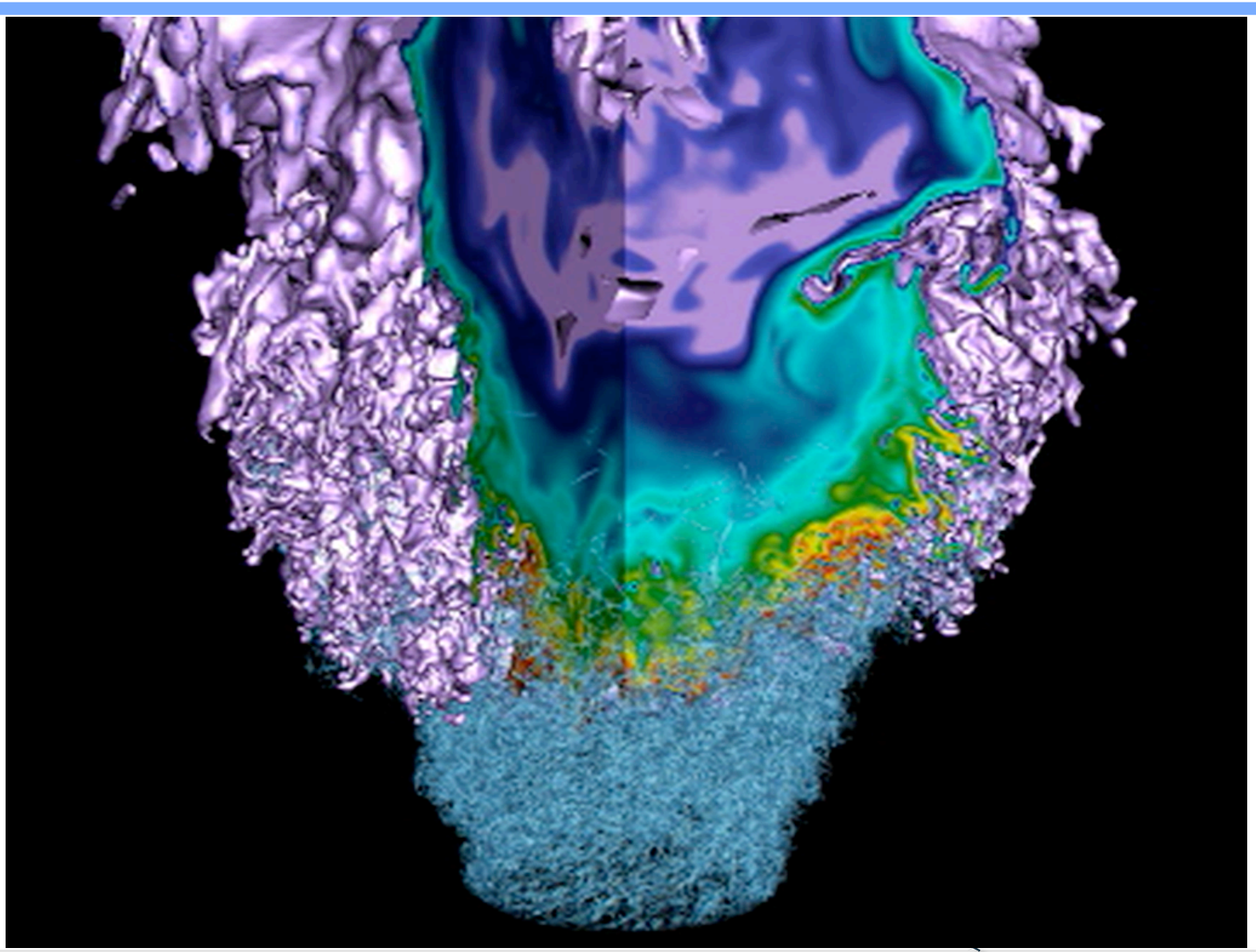


Experimental comparison of OH Radical

PI: J. Bell, LBNL

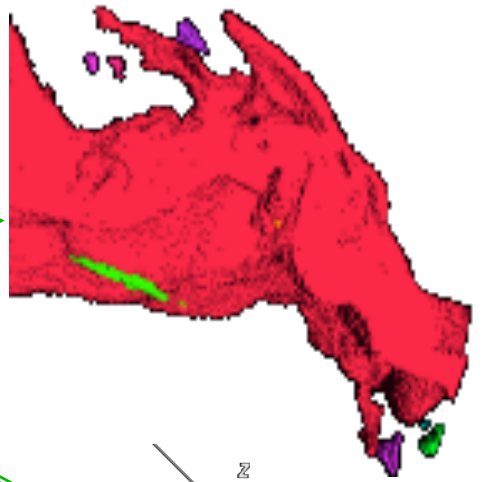
Simulations performed at NERSC under an INCITE grant

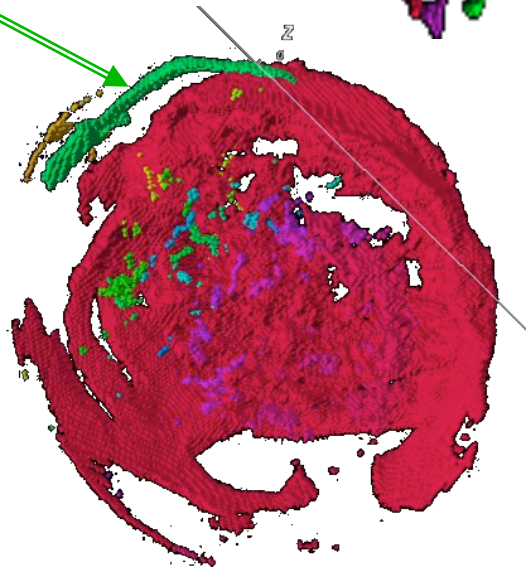
Simultaneous rendering of OH and Vorticity



Data Intensive Science Examples

- Find the HEP collision events with the most distinct signature of Quark Gluon Plasma
- Find the ignition kernels in a combustion simulation
- Track a layer of exploding supernova





These are not typical database searches:

- **Large high-dimensional** data sets (1000 time steps X 1000 X 1000 X 1000 cells X 100 variables)
- No modification of individual records during queries, i.e., **append-only data**
- Complex questions: 500 < Temp < 1000 && CH3 > 10^{-4} && …
- Large answers (hit thousands or millions of records)
- Seek collective features such as regions of interest, histograms, etc.

Challenges and Goals

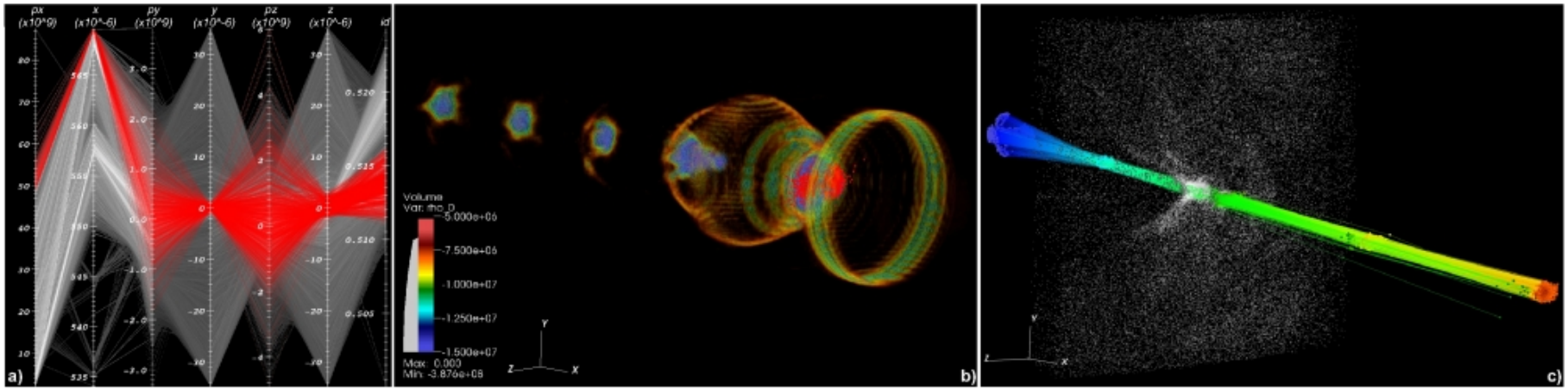
- Why is managing scientific data important to scientists?
  - Sheer volume and increasing complexity of data being collected are already interfering with the scientific investigation process
  - Managing the data by scientists greatly wastes scientists' time
  - Data collection, storage, transfer, and archival often conflict with effectively using computational resources
  - Effectively managing, and analyzing this data and associated metadata requires a comprehensive, end-to-end approach from the initial data acquisition to the final analysis

Enable scientists to most effectively discover new knowledge by removing data management bottlenecks and enabling effective data analysis

FastBit: accelerating analysis of large datasets

- Most data analysis algorithm cannot handle a whole dataset
  - Therefore, most data analysis tasks are performed on a subset of the data
  - Need fast indexing for real-time analysis
- FastBit is an extremely efficient compressed bitmap indexing technology
  - Can search billion data values in seconds
  - FastBit improves the search speed by 10–100X over best known indexing methods
  - Uses LBNL patented compression techniques
- FastBit indexes are modest in size compared to well-known database indexes
  - On average about 1/3 of data volume compared to 3-4 times in common indexes (e.g. B-trees)

Query-Driven Visualization

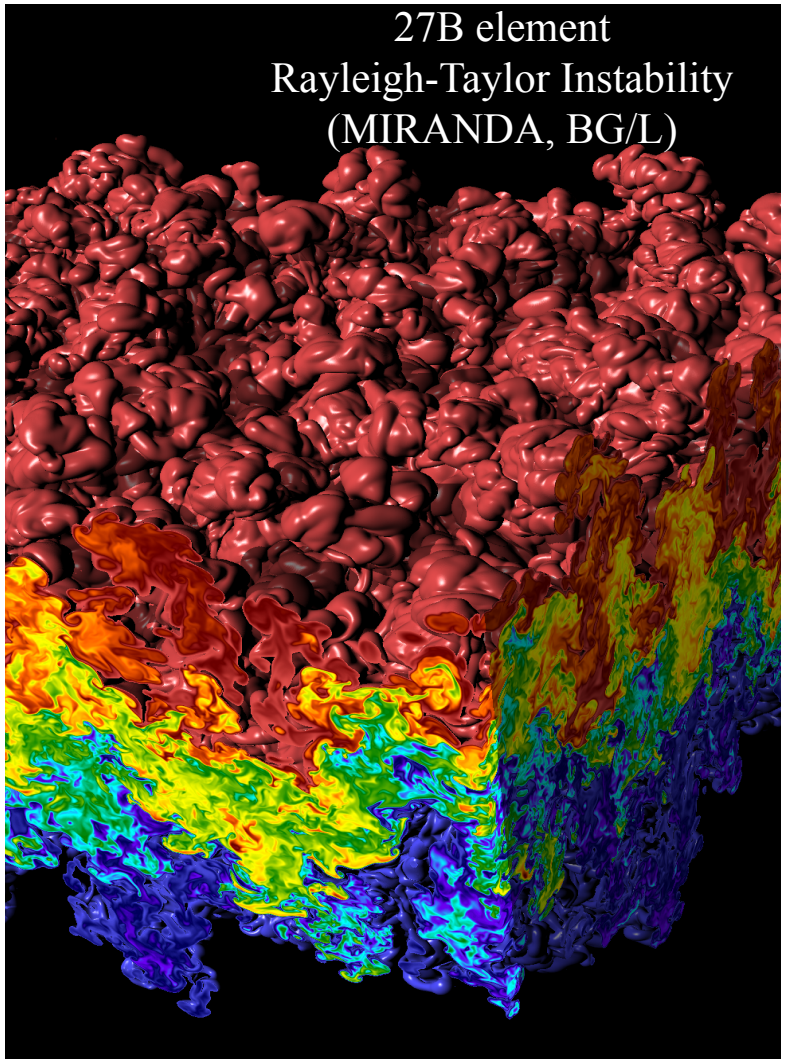


- Collaboration between SDM and VACET
  - Use FastBit indexes to efficiently select the most interesting data for visualization
- Laser wakefield accelerator simulation
  - VORPAL produces 2D and 3D simulations of particles in laser wakefield
  - Finding and tracking particles with large momentum is key to design the accelerator
  - Brute-force algorithm is quadratic (taking 5 minutes on 0.5 mil particles), FastBit time is linear in the number of results - takes 0.3 s,
  - 1000 X speedup

Contact: John Wu, Wes Bethel, LBNL

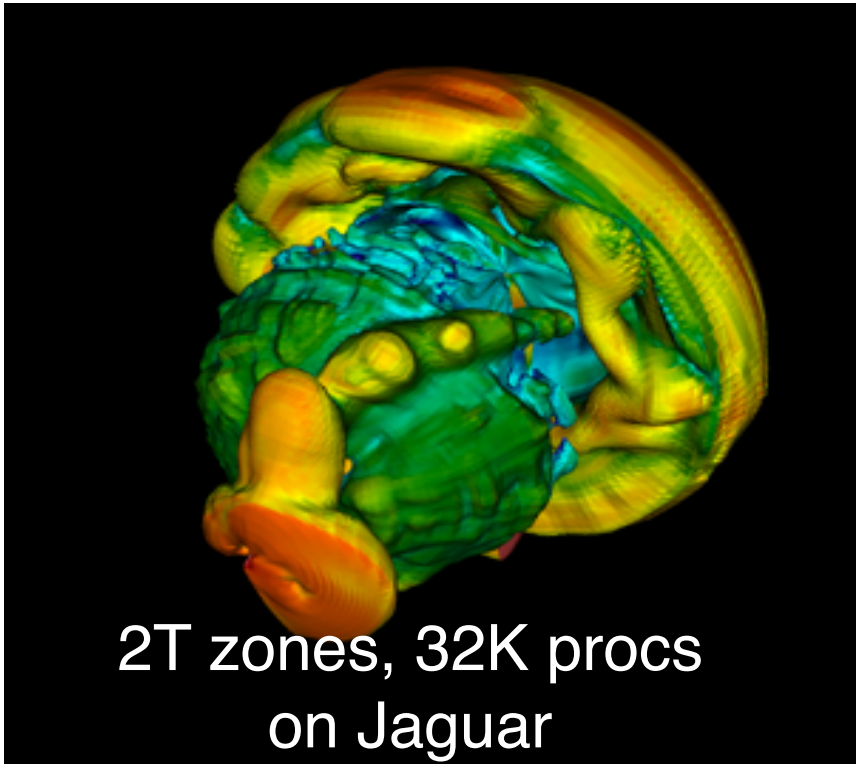
VisIt: a full-featured application for large data

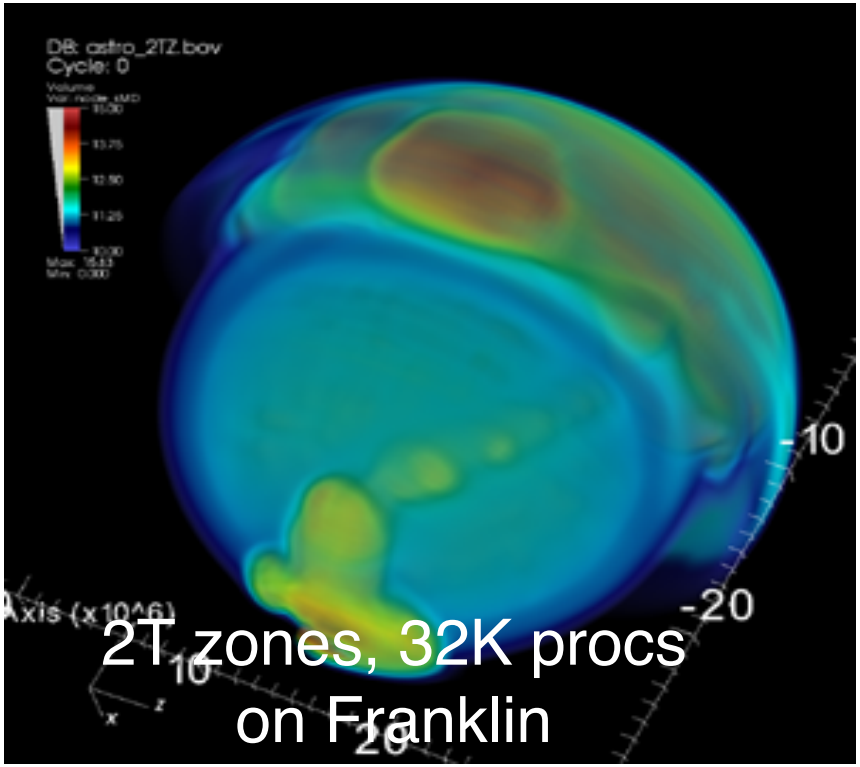
- VisIt is an open source, end user visualization and analysis tool for simulated and experimental data
  - Used by: physicists, engineers, code developers, vis experts
  - >100K downloads on web
- R&D 100 award in 2005
- Used “heavily to exclusively” on 8 of world’s top 12 supercomputers



Trillion cell experiment methodology:

- Two common visualization techniques:
  - Isocontouring, volume rendering





Summary

- SDM and Vis groups have developed data management and analysis tools through large multidisciplinary, multi-institution efforts
- High performance
  - Specialized Indexing technologies
  - Parallel analysis tools
  - Remote visualization of large distributed data sets
- Usability and effectiveness
  - Provide real-time monitoring, repeated analysis, code coupling
  - Production-quality visualization software
- Enabling data understanding
  - Analysis pipeline framework for combining multiple techniques
  - Parallel statistics tools
  - Use of indexing in query-based visualization

Questions?