



Argonne  
NATIONAL  
LABORATORY

*... for a brighter future*



U.S. Department  
of Energy

UChicago ►  
Argonne<sub>LLC</sub>



Office of  
Science

U.S. DEPARTMENT OF ENERGY

A U.S. Department of Energy laboratory  
managed by UChicago Argonne, LLC

# Visualization and Analysis at Extreme Scale

*Michael E. Papka*

*Argonne National Laboratory & The University of Chicago*

# Overview

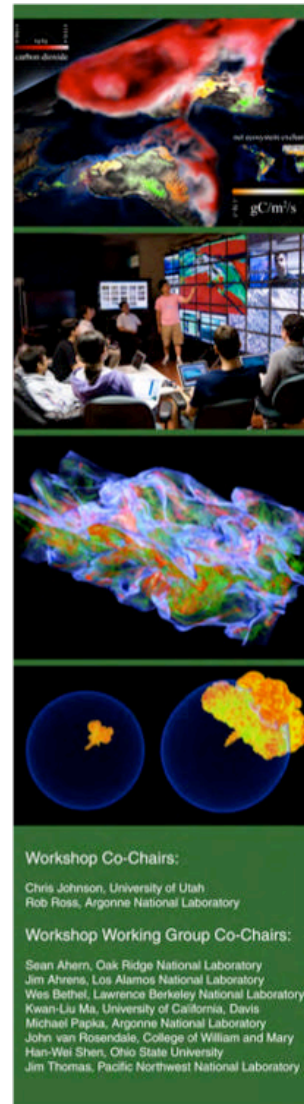
- DOE visualization report
- Extreme scale
- Where are we today
- Connecting to the visualization report

# Visualization and Knowledge Discovery

- Researchers from national labs, universities, industry
- Report on fundamental research in visualization and analysis necessary to enable discovery for computational science applications at extreme scale
- Four focus areas
  - Fundamental algorithms
  - Complexity of scientific datasets
  - Advanced architectures and systems
  - Knowledge-enabling visualization and analysis

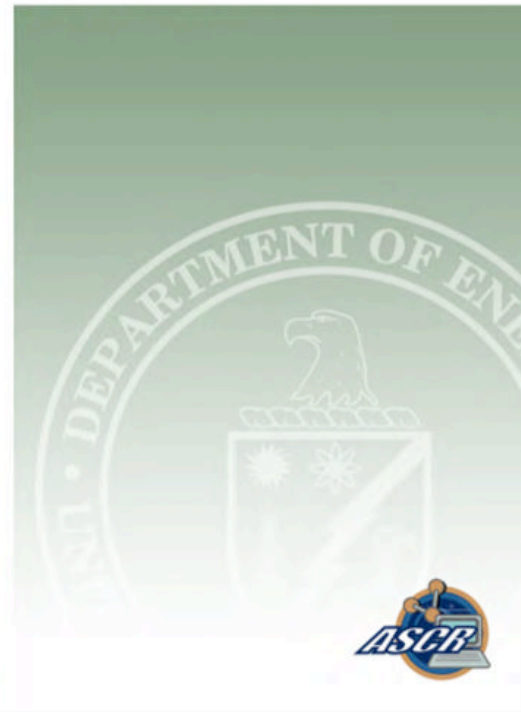
# Visualization and Knowledge Discovery

- Interaction and Collaboration
- Pervasive Parallelism and Multiscale Analysis
- Feature Detection and Tracking
- Multifield and Multimodel Data Understanding
- Distance Visualization
- In Situ Processing
- Time-Varying Datasets
- Visual Analysis, Quantification, and Representation of Uncertainty and Error
- End-to-End Integration



## Visualization and Knowledge Discovery:

Report from the DOE/ASCR  
Workshop on Visual Analysis and Data  
Exploration at Extreme Scale  
October 2007



# What is Extreme Scale?

- Operational Definition of Extreme Scale
  - From here to just beyond the current horizon
- Tera Scale
  - 1K - 100K processes, 1TB - 10TB data, 10K files
  - Multi-billion atom simulation
  - 3D simulation of weakly compressible turbulent flow

# What is Extreme Scale?

## ■ Peta Scale

- 100K - 10M processes, 10TB - 1PB data, 100K - 1M files
- 100M neuron network of multi-compartment cells
- Climate models at many times today's resolution

## ■ Exa Scale

- 10M - 100M processes?, 1PB - 100PB data?, ...
- Storm impacting modeling
- Sea and land ice into climate models
- Ecosystems and climate coupling

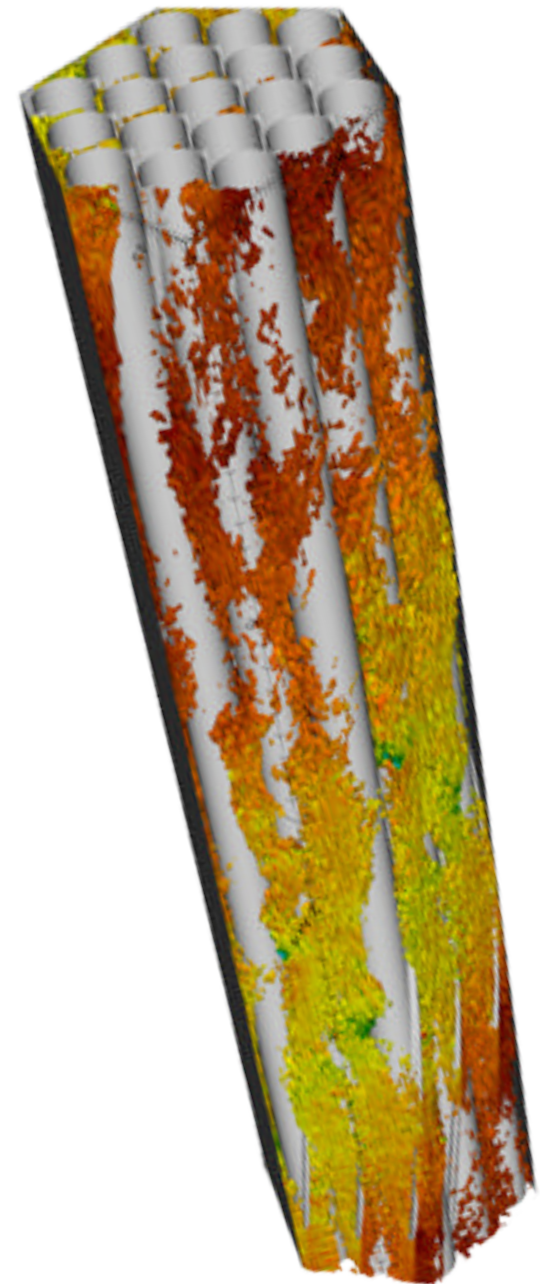
# Nuclear Reactor Simulation

## ■ Preliminary studies

- 4.5 million elements
- 7 variables per element
- 20K timesteps
- Total data produced 2.5TB

## ■ Science runs

- 3 – 4 @ 120 million elements
- Several runs at  $\frac{1}{2}$  and  $\frac{1}{4}$  resolution
- 90K timesteps
- Total data produced 900TB – 1.2PB



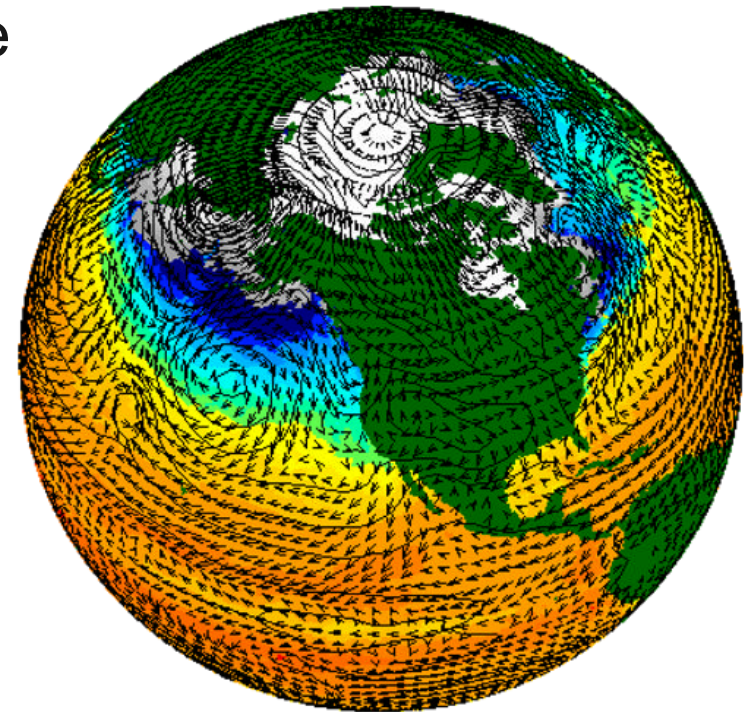
# Climate Modeling

## ■ Preliminary studies

- 50-100 @ 3 million grid points (1M atmosphere, 2M ocean)
- 100 variables per grid point (30 vectors, 70 scalars)
- Simulating 5 - 10 years of climate
- Total data produced 30 -124TB

## ■ Science runs

- 50 @ 6 million grid points
- Simulating 100 years of climate
- Total data produced 1.2PB



# Astrophysics

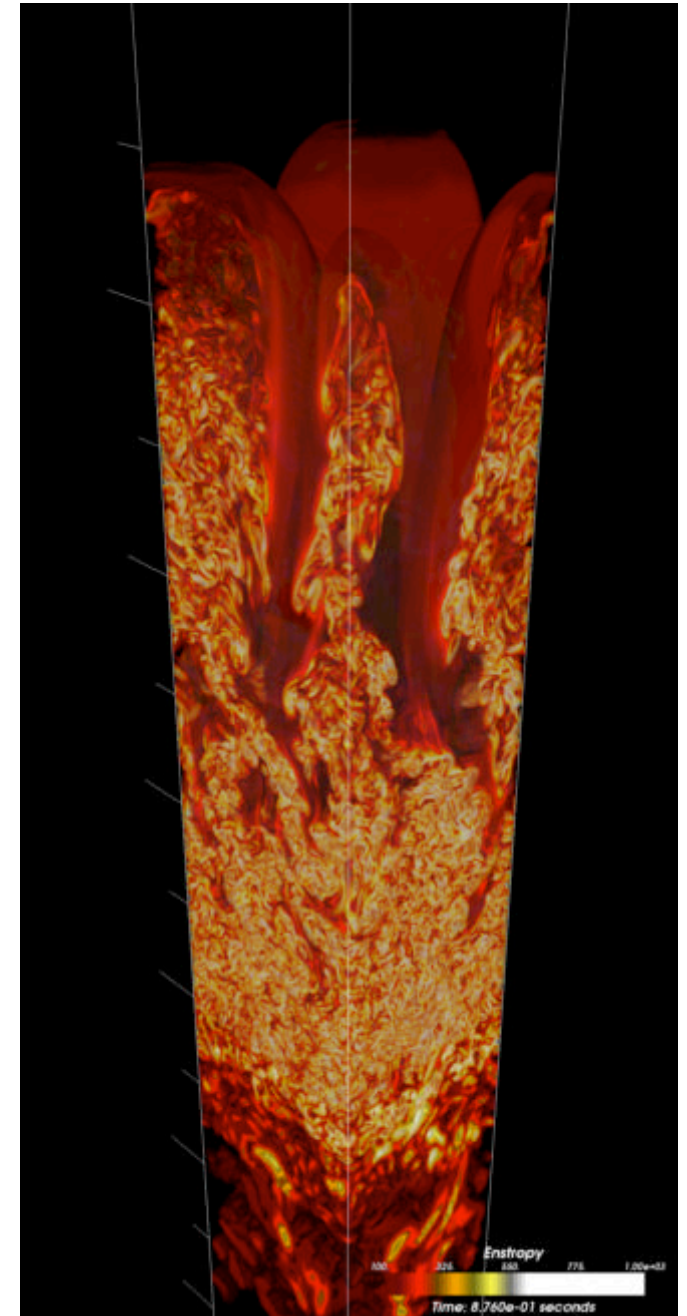
## ■ Preliminary studies

- ~80 @ 67M grid points
- ~5 @ 536M grid points
- 6 variables (1 vector, 3 scalars)
- ~1800 timesteps
- Total data produced 78TB

## ■ Science run\*

- 4.3B grid points
- 6 variables (1 vector, 3 scalars)
- ~1800 timesteps
- Total data produced 48TB

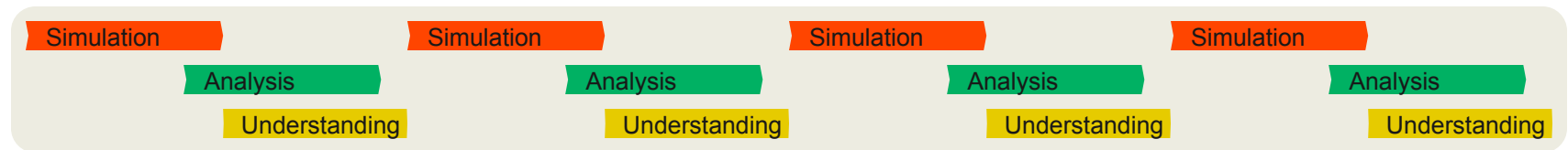
\* 3-5 times bigger allocation is needed



# Pre-Extreme Science

## ■ Typical scenario

- 3-4 *science* runs a year
- 3-4 months of analysis and development

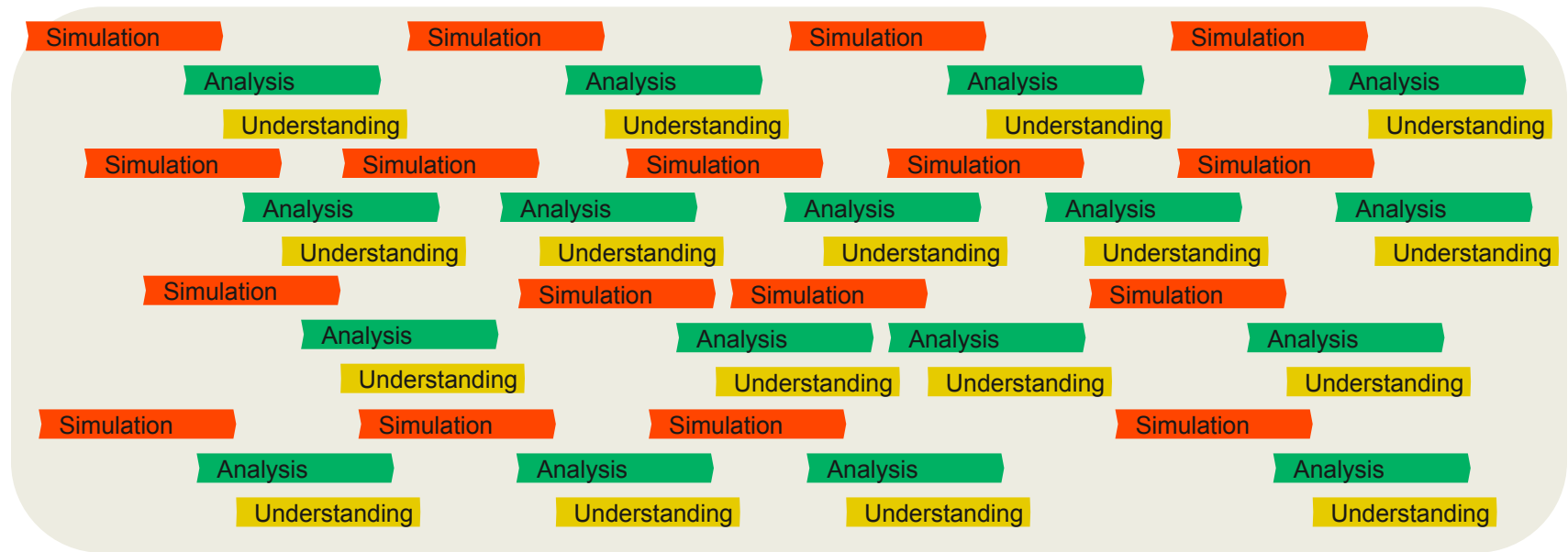


One year, 3-4 runs

# Extreme Science

## ■ Typical scenario

- 10s to 100s of *science* runs a year
- Days to weeks of analysis and development



One year, 10s – 100s of runs

# Feeding the Engine

- Our ability to generate data experimentally and by simulation far outstrips our ability to understand it
- We can overcome this by:
  - Applying visualization and analysis earlier in the process, during simulation
  - Developing more efficient data representations
  - Improved user interfaces so scientists can adapt their runs based on early analysis

# Interaction and Collaboration

## Big Data

- Hard to interact with in real time
- Hard to share

## Workflow

- Collaboration more important
- Interactive science

# Pervasive Parallelism and Multiscale Analysis

## Big Data

- Parallelism needed to process
- Exploit hardware trends
- Global and local views

## Workflow

- Overviews to find areas of interest
- Detailed views for analysis

# Feature Detection and Tracking

## Big Data

- Finding key components
- Objects through time

## Workflow

- Identification of areas of interest
- Classification
- Events

# Multifield and Multimodel Data Understanding

## Big Data

- Data parsing/filtering
- Use of multiple sensory channels

## Workflow

- Help speedup understanding

# Distance Visualization

## Big Data

- The challenge of relatively small pipes

## Workflow

- Usability

# In Situ Processing

## Big Data

- Processing data while in memory

## Workflow

- Starts analysis earlier

# Time-Varying Datasets

## Big Data

- Compression opportunities
- Model generation

## Workflow

- ?

# Visual Analysis, Quantification, and Representation of Uncertainty and Error

## Big Data

- Verification
- Steering

## Workflow

- Efficiency
- Guidance

# End-to-End Integration

## Big Data

- Needed to meld all the pieces mentioned before

## Workflow

- Less complexity

# Analysis Is Becoming A Bottleneck

- Extreme scale is driving to critical point
- Need for a shift in priorities
- Decoupling simulation from visualization and analysis, has created a additional resource footprint that is becoming too large to tolerate
- Need to consider strategies through hardware configuration and software methods to optimize the coupling of simulation, visualization and analysis

# Conclusions

- Report states: “A wealth of information creates a poverty of attention and a need to allocate it efficiently.” from Herbert Simon
- We can allocate our limited attention more efficiently by:
  - Effectively targeting our analyses
  - Exploring results earlier and at more stages of the process, to gain understanding earlier, and incrementally refocusing our analyses

# Conclusions

- Push more of computational science into infrastructure, so scientists can concentrate on Science.
- Scientists today are far too involved in the computation aspect of “computational science”;
  - build tools that expose useful interfaces to them,
  - hide the underlying details (where data is stored, how/ where jobs are submitted),
- Consider this maxim from Alan Kay, in application to scientists’ approach to computational science:  
“Simple things should be simple; Complex things should be possible”

# Acknowledgments

- This work supported in part by DOE, NSF, and NIH
- The visualization group at Argonne/UChicago – Brad Gallager, Mark Hereld, Randy Hudson, Joe Insley, John Norris, Eric Olson, Tom Peterka, Rob Ross, Thomas Uram
- Application groups– Bob Fisher, Rob Jacob, Andrew Siegel
- Additional input – Pete Beckman, Rusty Lusk, Rick Stevens