LA-UR-11-11032

Approved for public release; distribution is unlimited.

Title: VISUALIZATION AND DATA ANALYSIS IN THE EXTREME SCALE ERA

Author(s): James Ahrens

Intended for: SCIDAC 2011- July 12 - DENVER, COLORADO

VISUALIZATION AND DATA ANALYSIS IN THE EXTREME SCALE ERA

James Ahrens
Los Alamos National Laboratory

Jonathan Woodring, John Patchett, Li-Ta Lo, Chris Sewell, Susan Mniszewski, Patricia Fasel, Joshua Wu, Christopher Brislawn, Christopher Mitchell, Sean Williams, Dave DeMarle, Berk Geveci, William Daughton, Katrin Heitmann, Salman Habib, Mat Maltrud, Phil Jones, Daniel Livescu

SCIDAC 2011- July 12 - DENVER, COLORADO

Introduction

- What are the challenges in the extreme scale supercomputing era for visualization and data analysis?
  - Challenge #1 – changing supercomputing architectures
    - Solution: New processes, algorithms, foundations
  - Challenge #2 – massive data
    - Solution: New quantifiable data reduction techniques
  - Challenge #3 – massive compute enables new physics
    - Solution: Custom visualization and data analysis approaches

Supercomputing Architectural Challenges for Data Analysis and Visualization

Mega	Giga	Tera	Peta	Exa
10^6	10^9	10^{12}	10^{15}	10^{18}
Displays	Networks & Storage bandwidths	Operations per second	Operations per second	Operations per second

Introduction

- Structure of this presentation
  - Review our state of the art
  - Discuss challenges #1 and #2
  - Present research work on specific solutions applied to scientific applications

State of the art foundational concepts

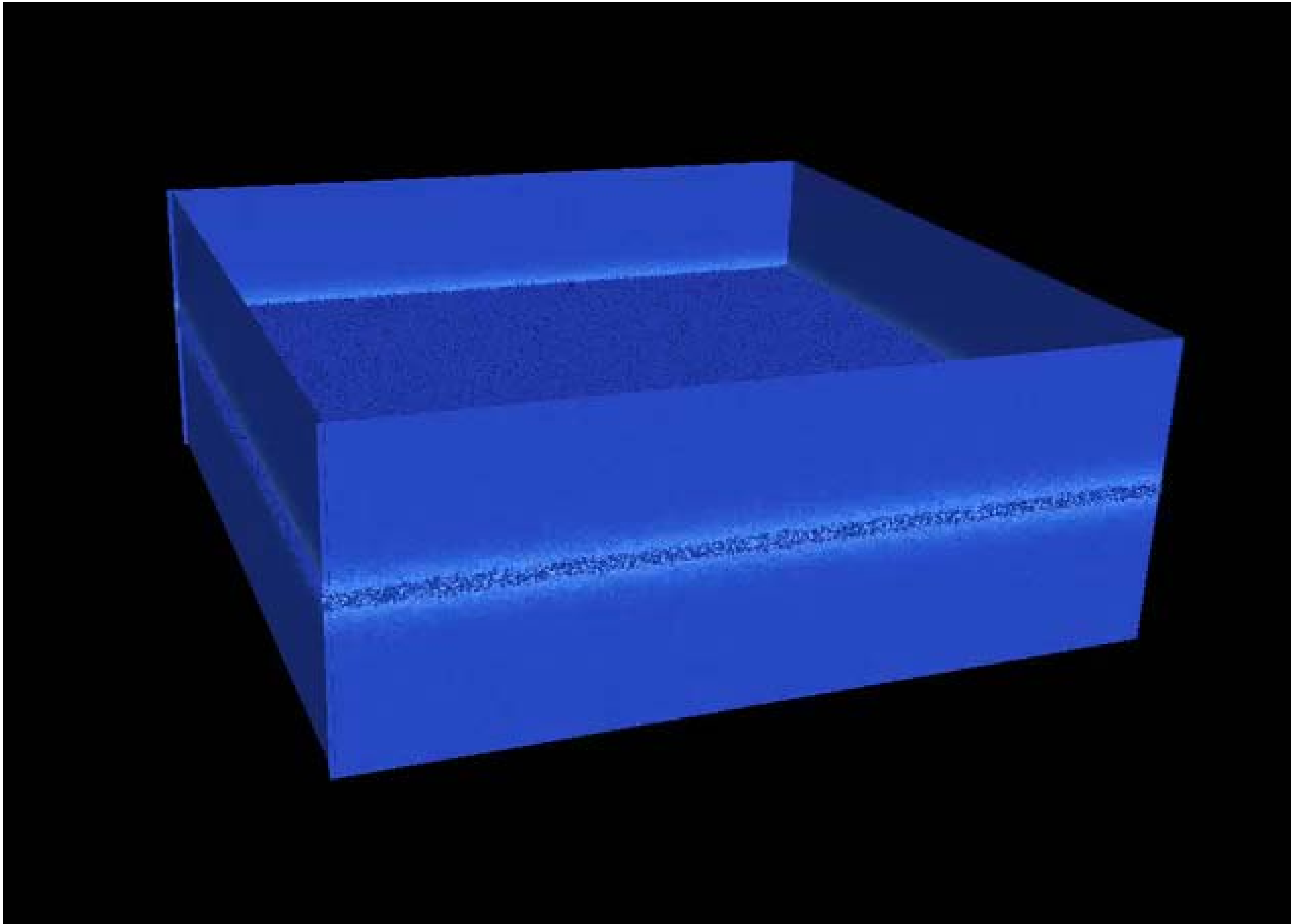
1) Open-source
2) Portability to most architectures
3) Full-featured toolkit of visualization and analysis operations
4) Data parallelism
5) Multi-resolution

In vtk, ParaView, Visit

- Streaming data model
  - the incremental independent processing of data
    - Enables out-of-core processing, parallelism and multi-resolution
    - Supports culling and prioritization

VPIC Plasma Simulation State of the Art Example

- Magnetic reconnection is a basic plasma process involving the rapid conversion of magnetic field energy into various forms of plasma kinetic energy, including high-speed flows, thermal heating, and highly energetic particles.
- Simulation runs on Roadrunner, Kraken and Jaguar
  - Computing massive grid sizes - 8096x8096x448
- Saving data for later post-processing using supercomputing platform or attached visualization cluster
  - Striding and subsetting data to explore and understand their data
- The VPIC team considers interactive visualization critical to the success of their project

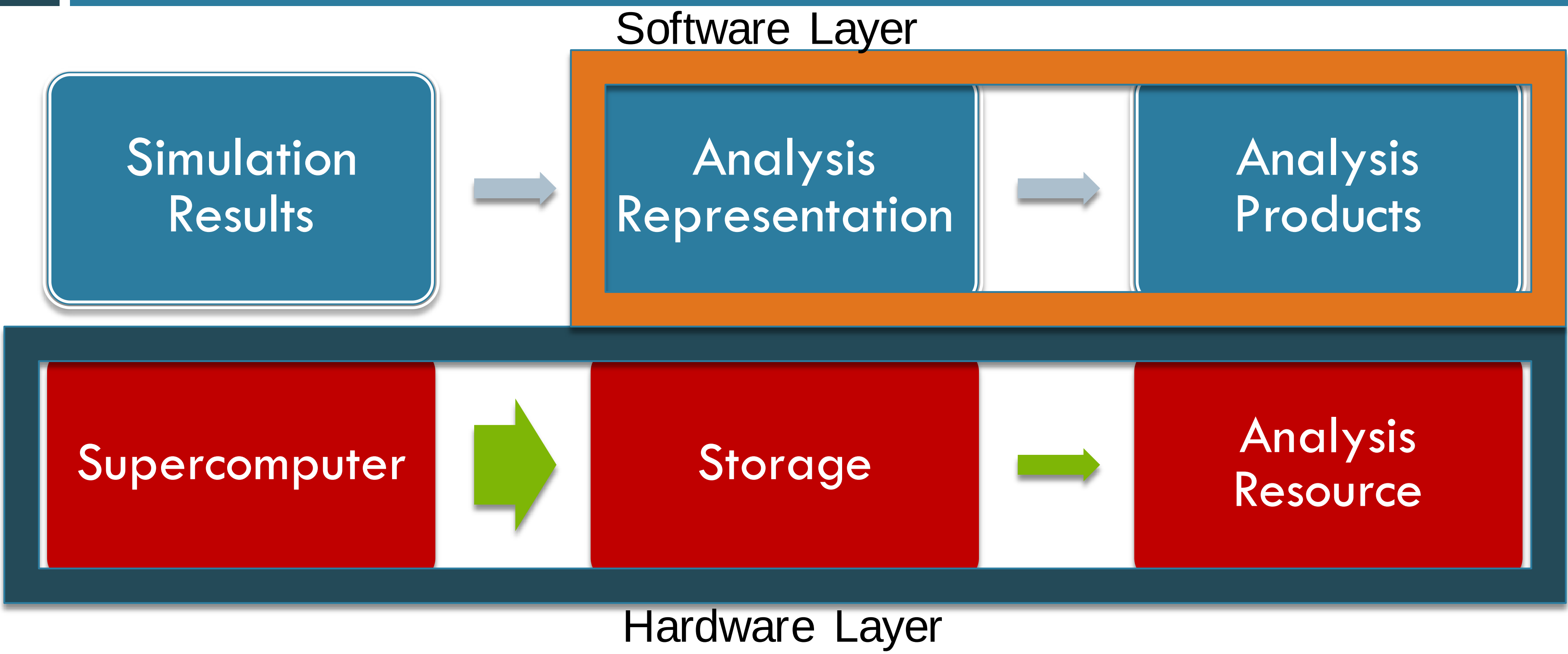


The central electron current sheet shown using an isosurface of the current density

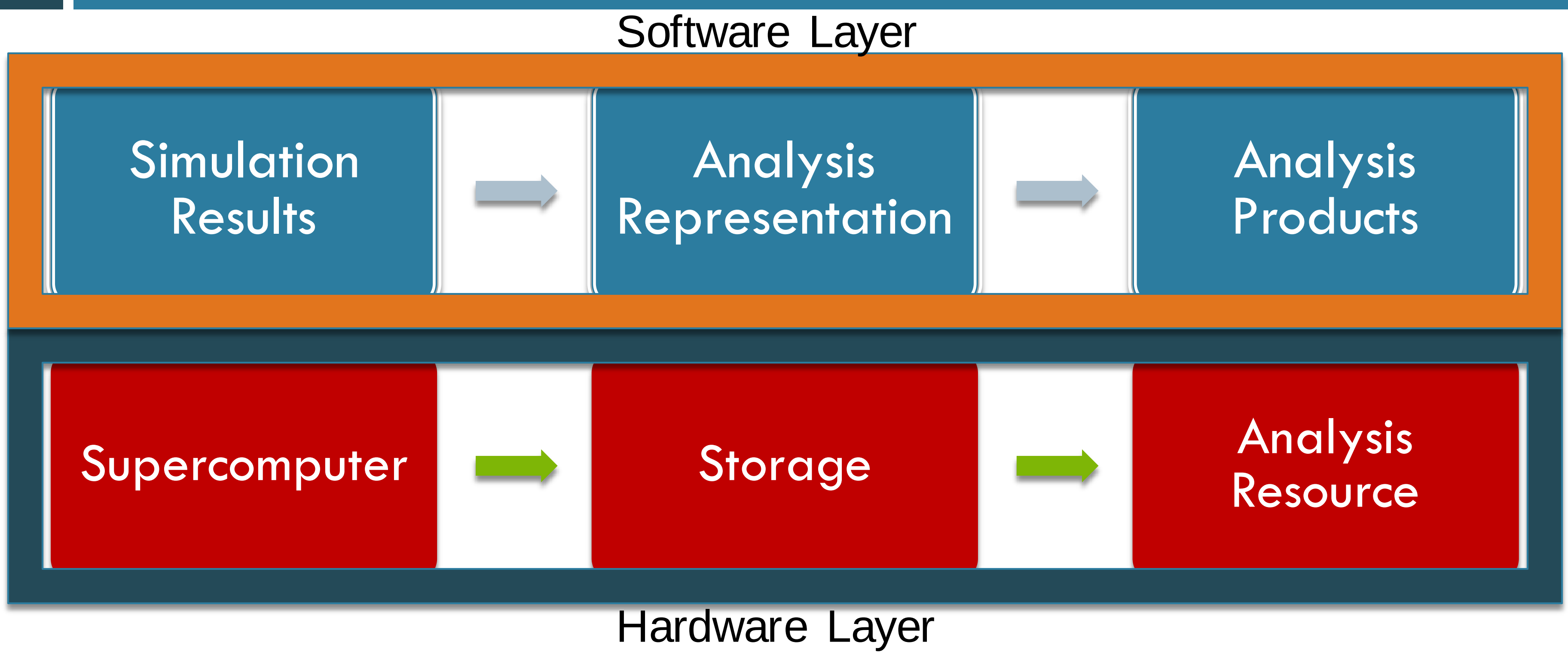
Challenge #1: Changing supercomputing architectures

- The rate of performance improvement of rotating storage is not keeping pace with compute
  - Provisioning additional disks is a possible mitigation strategy
  - However, power, cost and reliability issues will become a significant issue
- In addition, data movement is proportional to power costs
  - Must reduce data in-situ while simulation is running
- A new integrated in-situ and post-processing visualization and data analysis approach is needed

Current Analysis Workflow

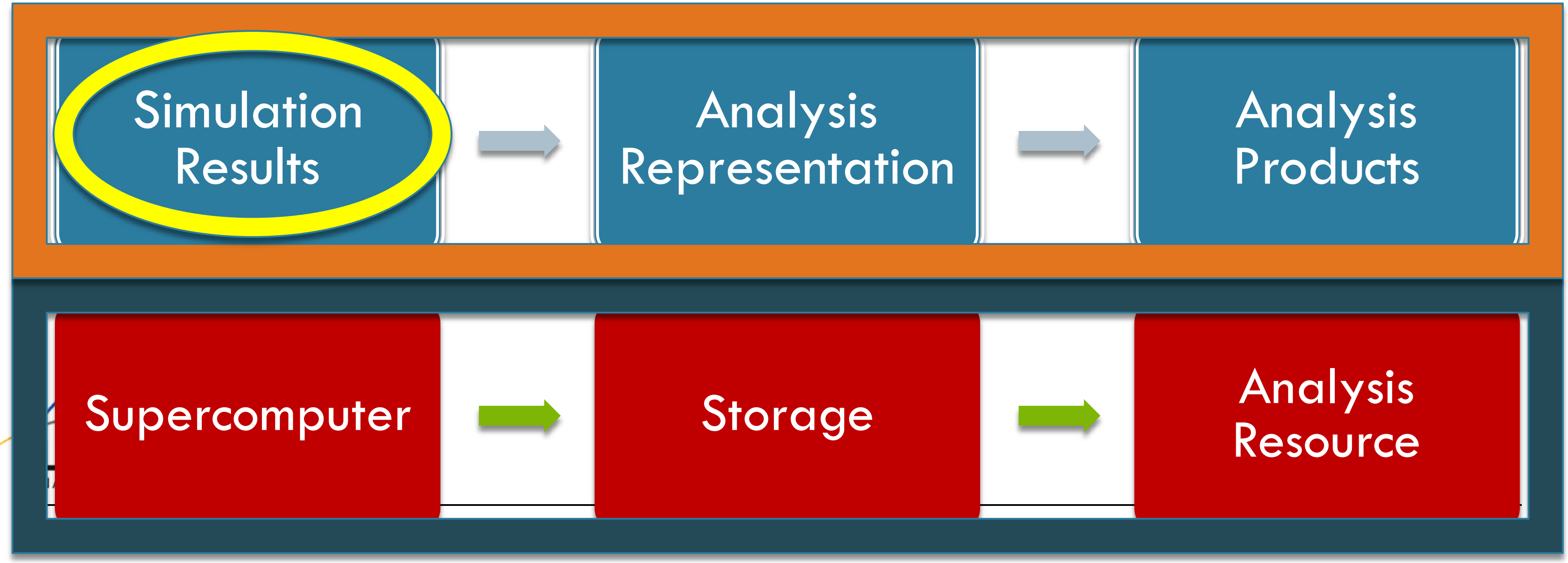


Evolving the Analysis Workflow

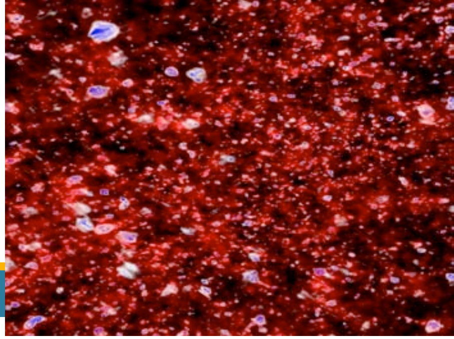


Challenge #2: Massive Data

- Extreme scale simulation results must be distilled with quantifiable data reduction techniques
  - Feature extraction, Statistical sampling, Compression, Multi-resolution



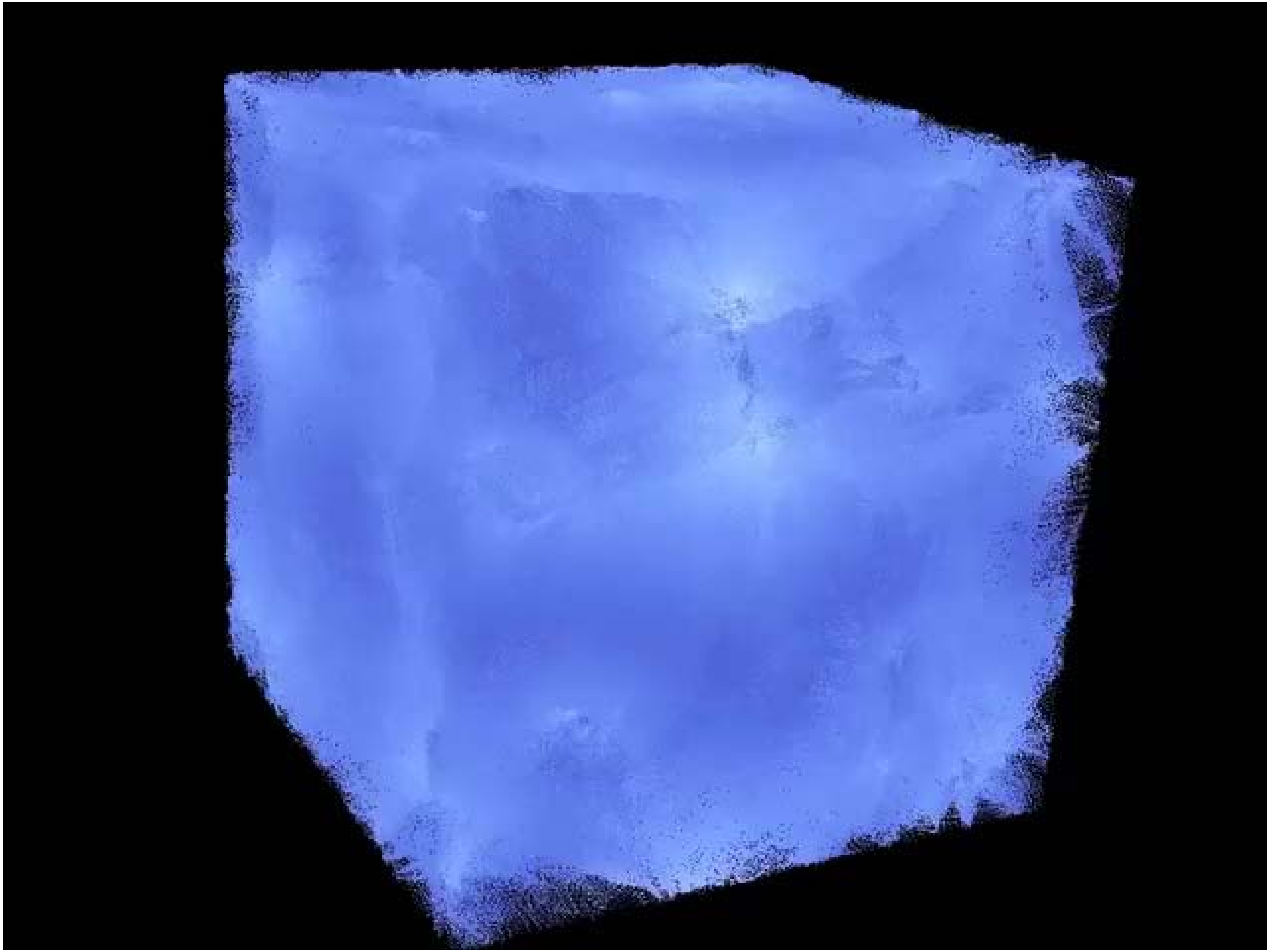
Example from Cosmological Science



- The data sizes for the simulations are exceedingly large
  - A 4000^3 (16 billion) particle run is approximately 2.3 TB per time step
- Simulation storage is optimized for fast checkpoint restart writes, assuming only 10%-20% of simulation time is used
  - Therefore there is a limit on how much data can be saved
    - Decision to save halos and halo properties
    - ~ 2 orders of magnitude data reduction

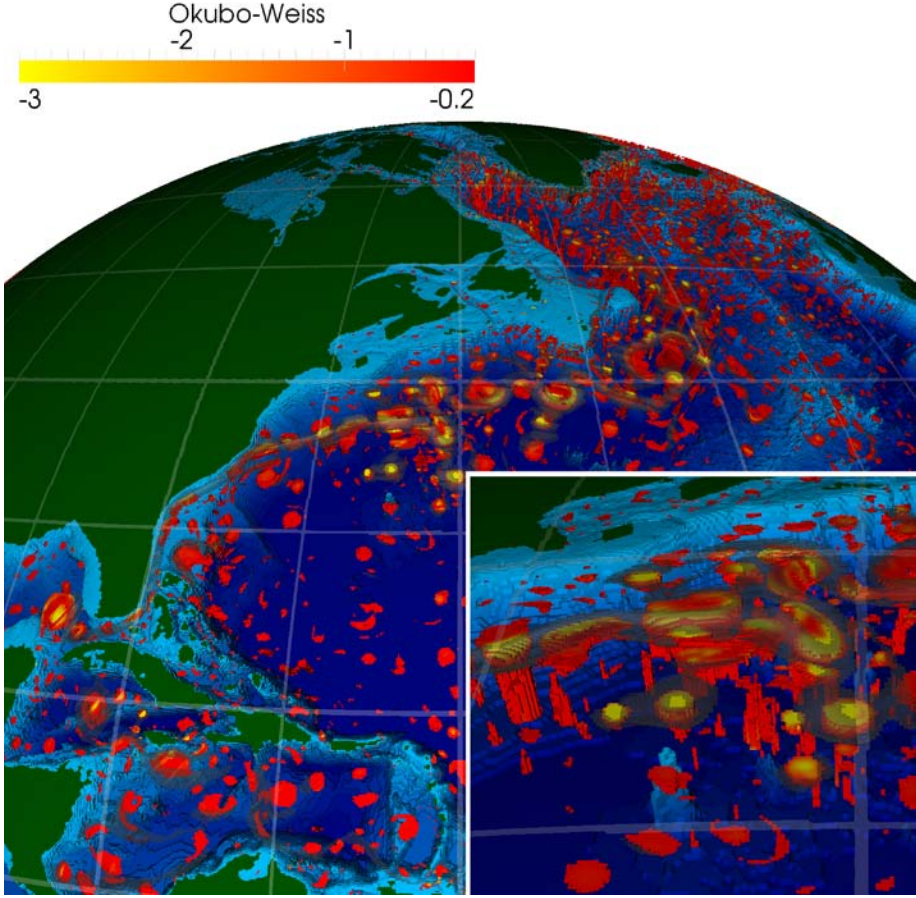
Solution to Massive Data Challenge: Feature Extraction

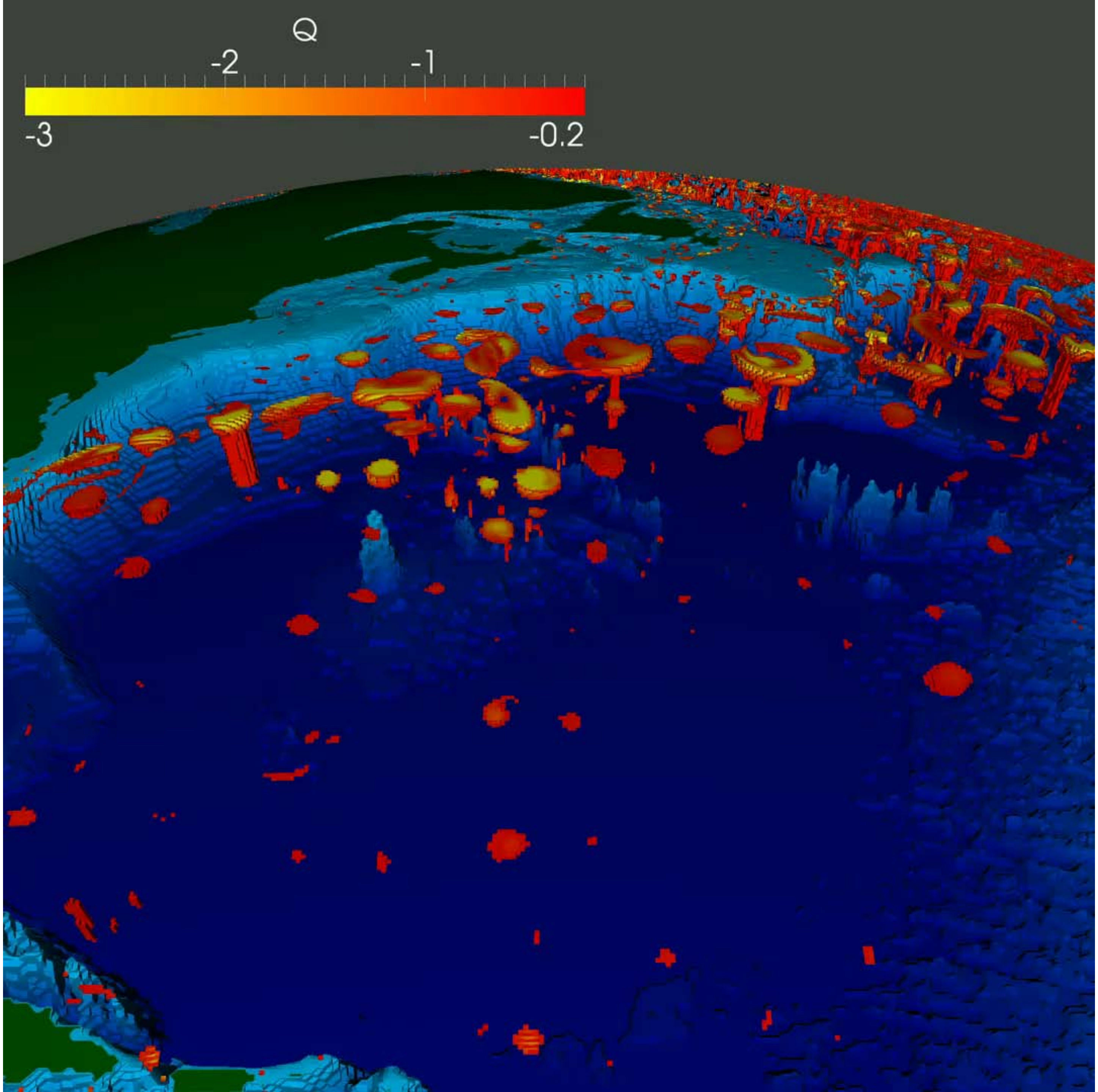
- Science specific techniques need to be created and generalized
  - Cosmology
    - Friend of friends halo
      - 3D connected component for particle data
        - Linking length
      - Implementation
        - spatial kd tree
        - similar to merge sort
  - Materials
    - Reusing halo finder for atomistic queries
- Techniques needs to run in parallel on the supercomputing platform



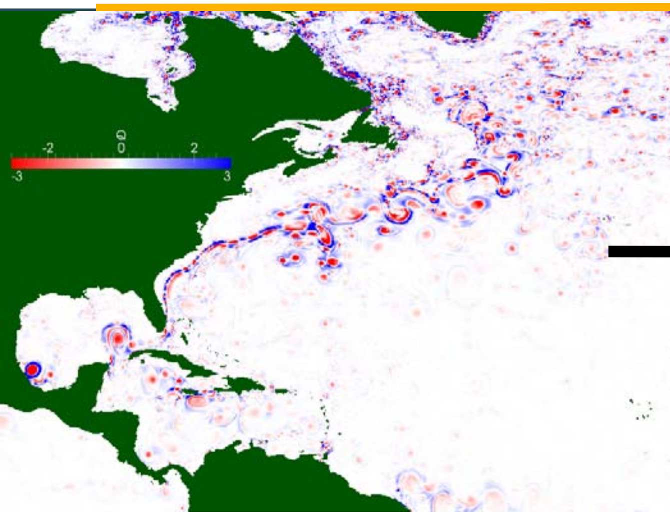
Case Study from Climate Science

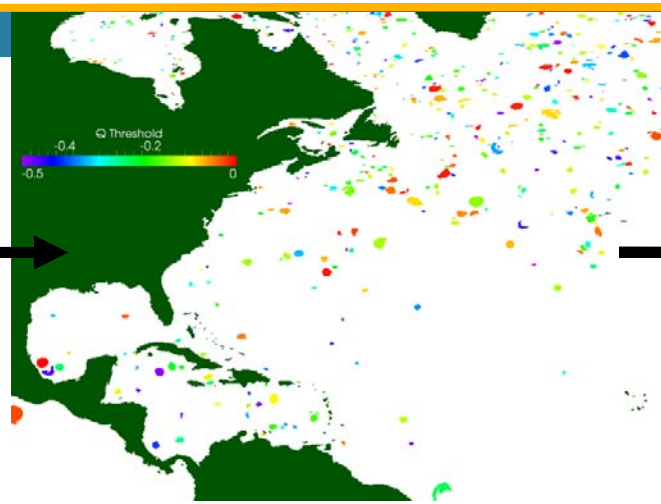
- Mesoscale eddies are large, long-lived vortices
- Eddies transport heat, salt, and nutrients
- This impacts the global energy budget
- But, the impact is not well-understood




Q
-2
-1
-3
-0.2

Eddy Feature Extraction Reduces Data Size





Longitude	Latitude	Depth	Radius	vx
-179.77	-66.8128	35.7585	1.78412	
-179.867	-63.8611	96.9241	0.977205	
-179.508	-59.4139	5.00622	5.49904	
-179.743	-51.7519	5.00622	2.70576	
-179.756	-49.0265	5.00622	2.25676	
-179.9	-46.891	82.9241	1.26157	
-179.531	-46.341	96.9241	4.44244	
-179.857	-42.7571	25.2834	1.49271	
-179.847	-38.4564	579.307	2.1851	

- 2 * 1.4 GB per time step * 350 time steps = 980 GB
- 5000 eddies per time step * 6 floats * 350 time steps = 30,000 floats = 120 KB

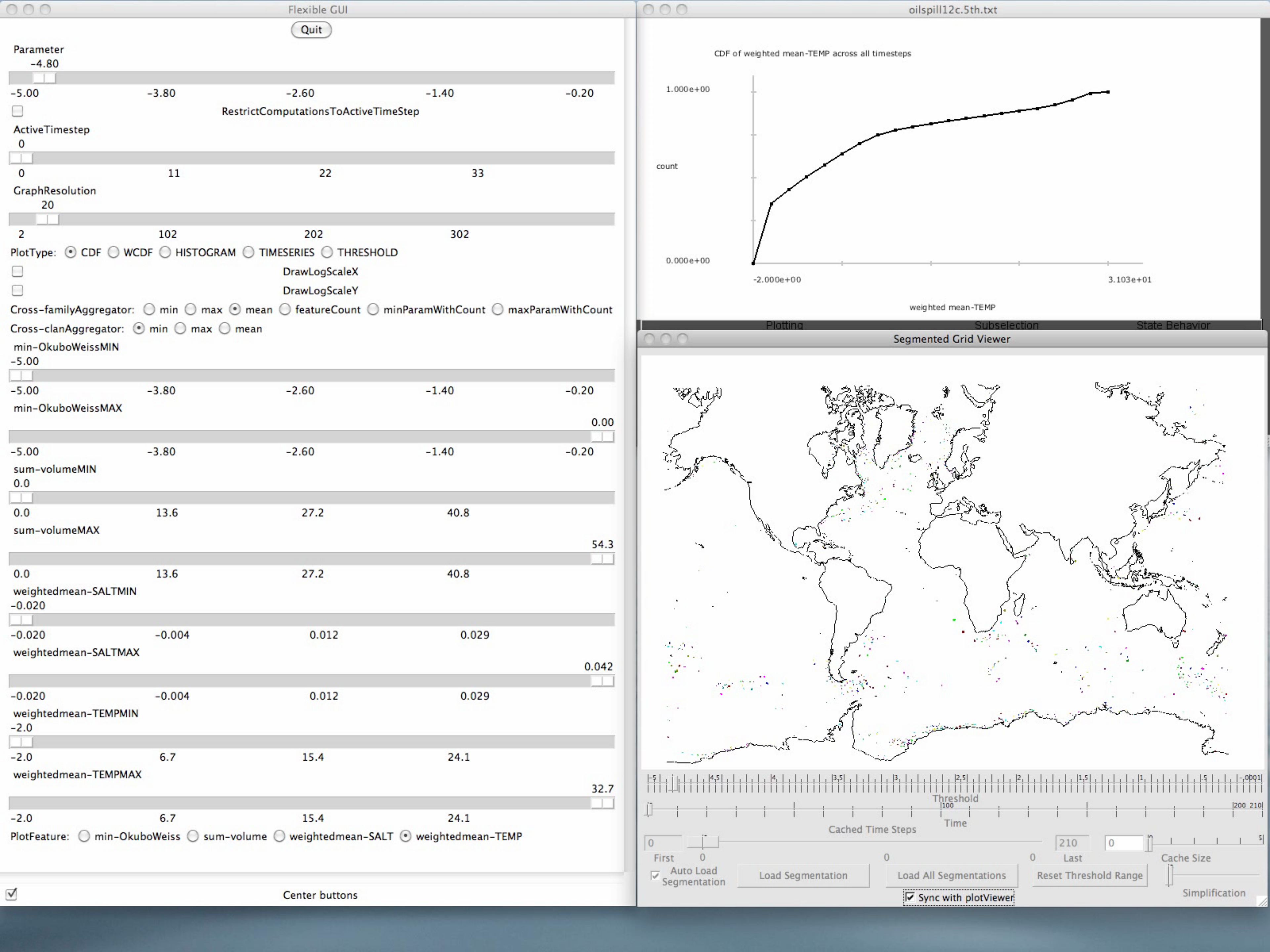
Interactive Analysis and Visualization of Ocean Eddies

Peer-Timo Bremer, Janine Benette,
Sean Williams, Cameron Christensen,
James Ahrens, Valerio Pascucci

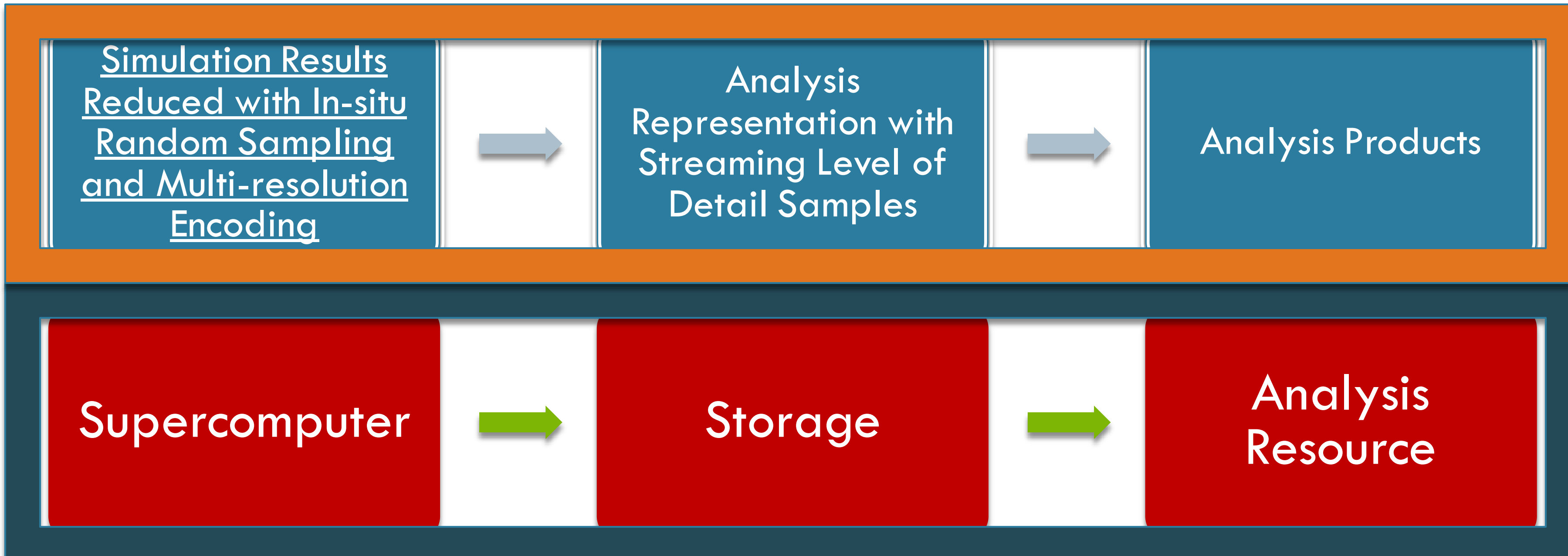
Ultrascale Visualization
Climate Data Analysis Toolkit

Visualization and Analytics Center of
Enabling Technology

This work has been has been performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344

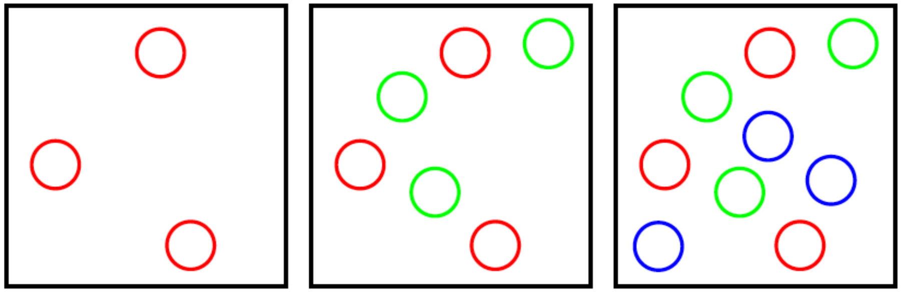

Flexible GUI
Quit
Parameter
-4.80
-5.00 -3.80 -2.60 -1.40 -0.20
RestrictComputationsToActiveTimeStep
ActiveTimestep
0
0 11 22 33
GraphResolution
20
2 102 202 302
PlotType: CDF WCDF HISTOGRAM TIMESERIES THRESHOLD
DrawLogScaleX
DrawLogScaleY
Cross-familyAggregator: min max mean featureCount minParamWithCount maxParamWithCount
Cross-clanAggregator: min max mean
min-OkuboWeissMIN
-5.00
-5.00 -3.80 -2.60 -1.40 -0.20
min-OkuboWeissMAX
0.00
-5.00 -3.80 -2.60 -1.40 -0.20
sum-volumeMIN
0.0
0.0 13.6 27.2 40.8
sum-volumeMAX
54.3
0.0 13.6 27.2 40.8
weightedmean-SALTMIN
-0.020
-0.020 -0.004 0.012 0.029
weightedmean-SALTMAX
0.042
-0.020 -0.004 0.012 0.029
weightedmean-TEMPMIN
-2.0
-2.0 6.7 15.4 24.1
weightedmean-TEMPMAX
32.7
-2.0 6.7 15.4 24.1
PlotFeature: min-OkuboWeiss sum-volume weightedmean-SALT weightedmean-TEMP
Center buttons
oilspill12c.5th.txt
CDF of weighted mean-TEMP across all timesteps
1.000e+00
count
0.000e+00
-2.000e+00
3.103e+01
weighted mean-TEMP
Plotting
Subselection
State Behavior
Segmented Grid Viewer
Threshold
Time
Cached Time Steps
0
210
First
Last
Cache Size
Auto Load Segmentation
Load Segmentation
Load All Segmentations
Reset Threshold Range
Sync with plotViewer
Simplification

Evolving the Analysis Workflow with Random Sampling and LOD Encoding

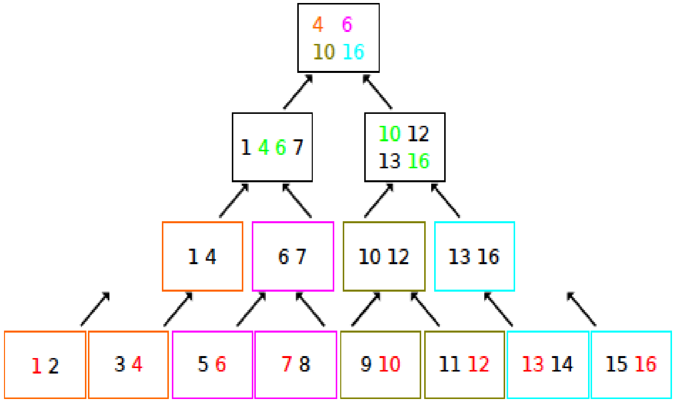


Solution to Massive Data Challenge: Use In Situ Statistical Multi-resolution Sampling to Store Simulation Data

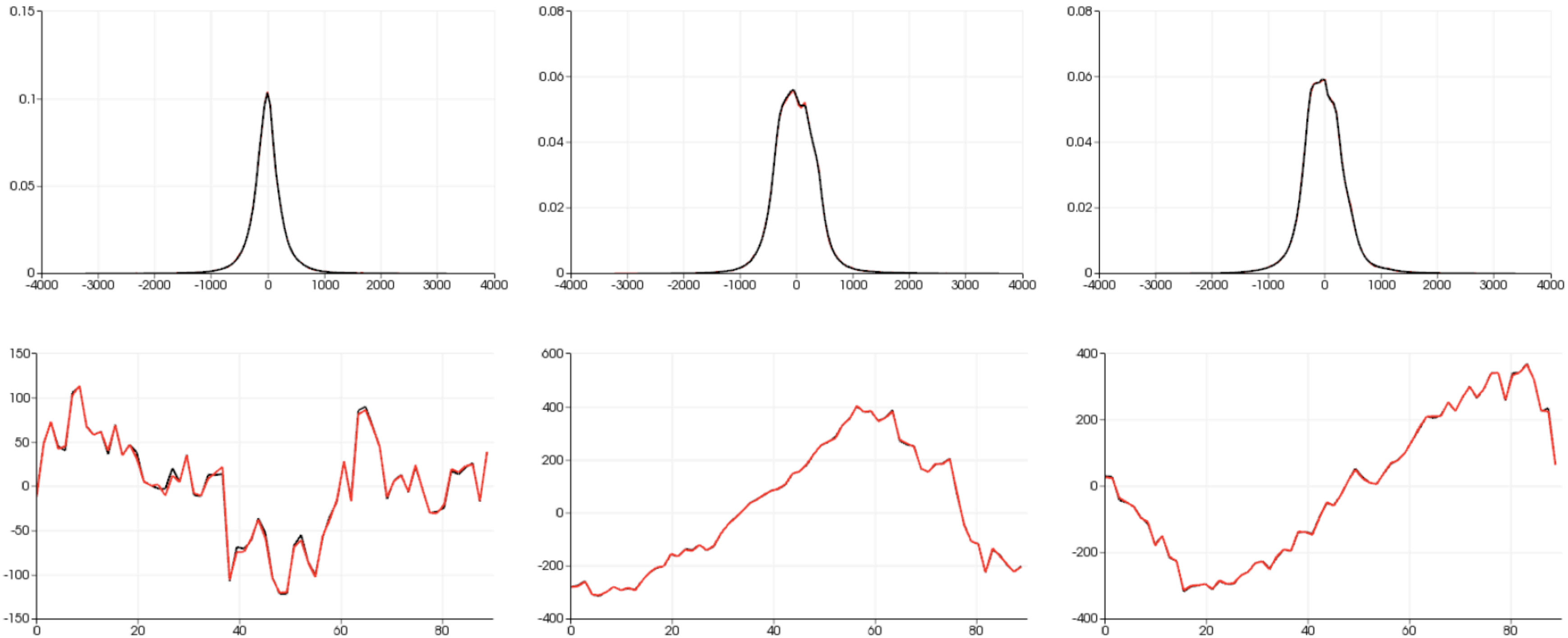
- Random sampling provides a data representation that is unbiased for statistical estimators, e.g., mean and others
- Since the sampling algorithm is done in situ, we are able to measure the local differences between sample data and full resolution data
- (Simulation Data – Sampled Representation) provides an accuracy metric



An abstract depiction of LOD particle data under increasing resolution with visual continuity. The particles in the lower resolution data are always present in the higher resolution data.

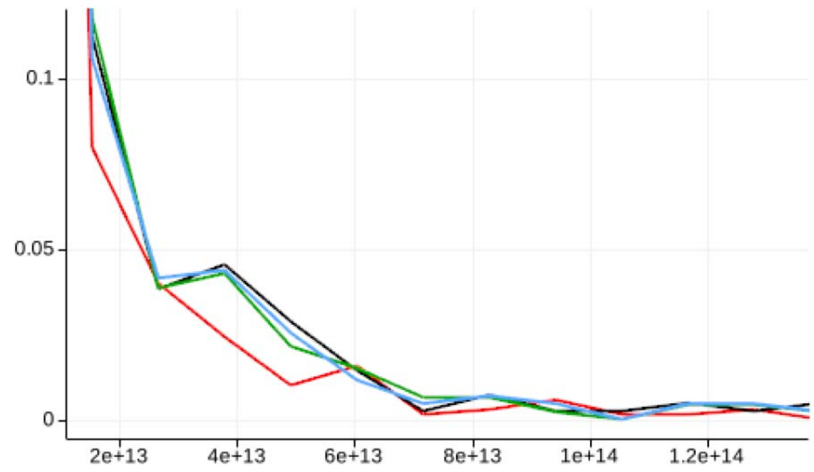


Empirically Comparing a 0.19% Sample compared to Full Resolution MC^3 Data



Red is 0.19% sample data, black is original simulation data. Both curves exist in all graphs, but the curve occlusion is reversed on top graphs compared to bottom graphs.

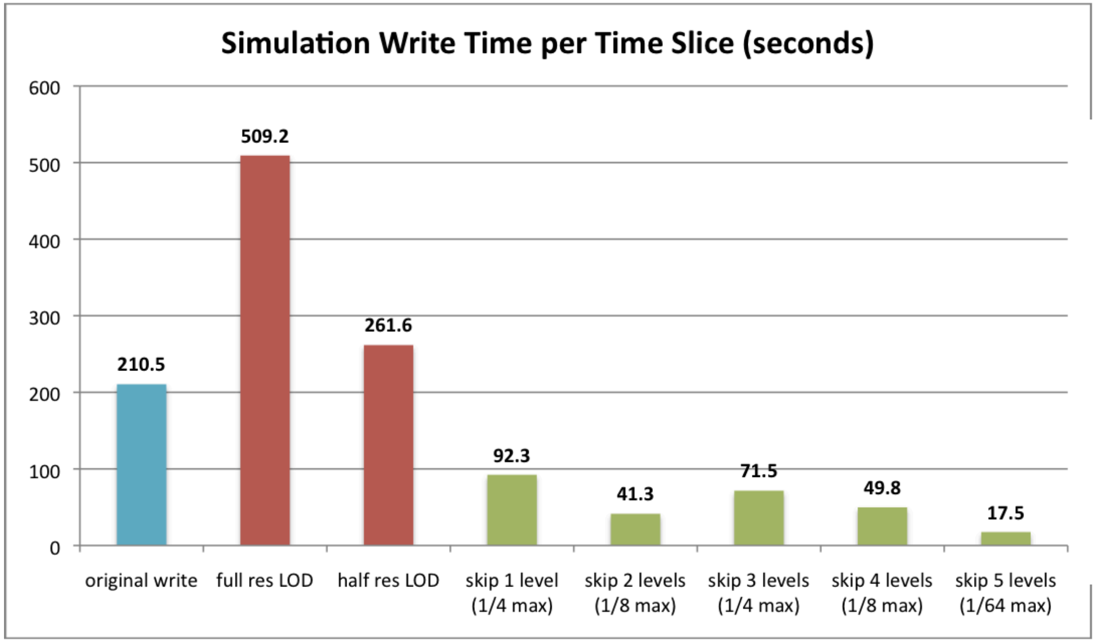
Effect of Sampling on Friend of Friends Algorithm

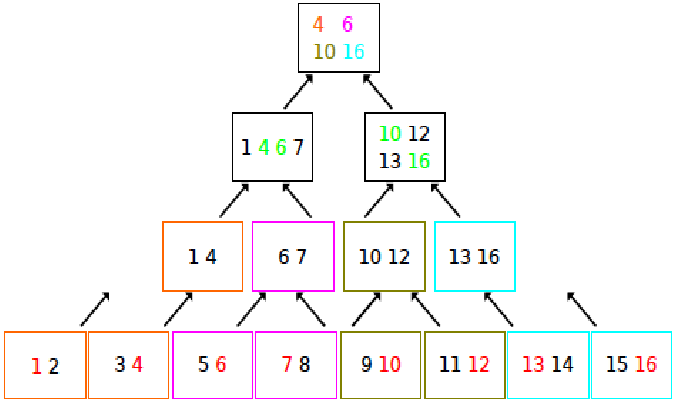


- The halo mass function for different sample sizes of 256^3 particles. The black curve is the original data. The red, green, and blue curves are 0.19%, 1.6%, and 12.5% samples, respectively.

512-way Simulation I/O Time Savings per Time Step for 2048^3 particles (8 billion)

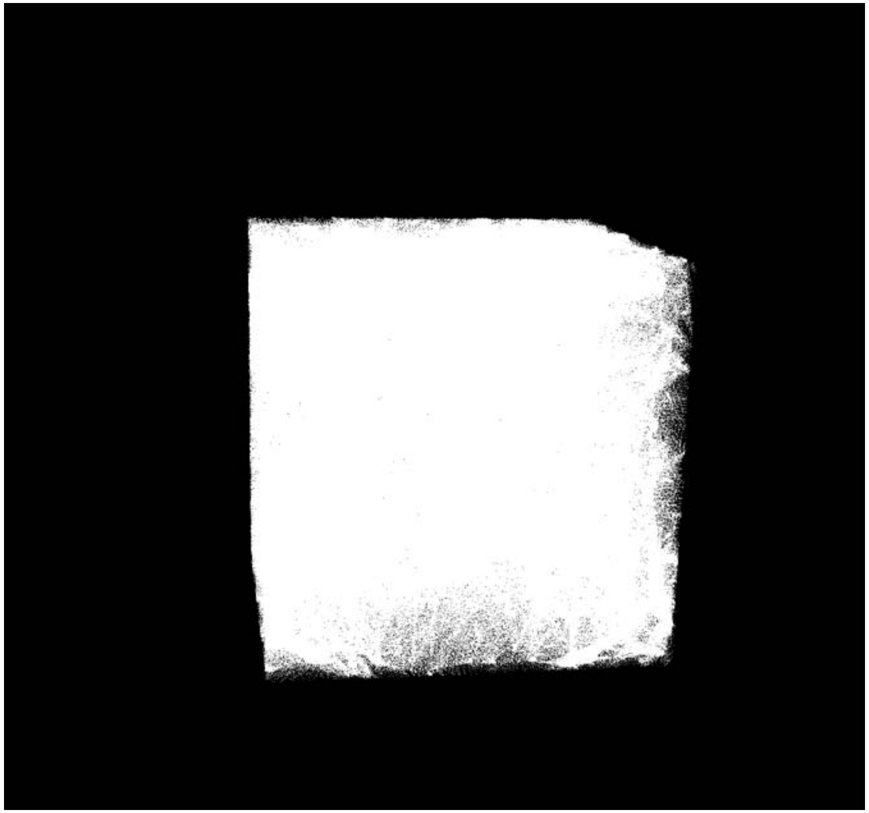
Storing less data through sampling significantly reduces the amount of time spent in I/O



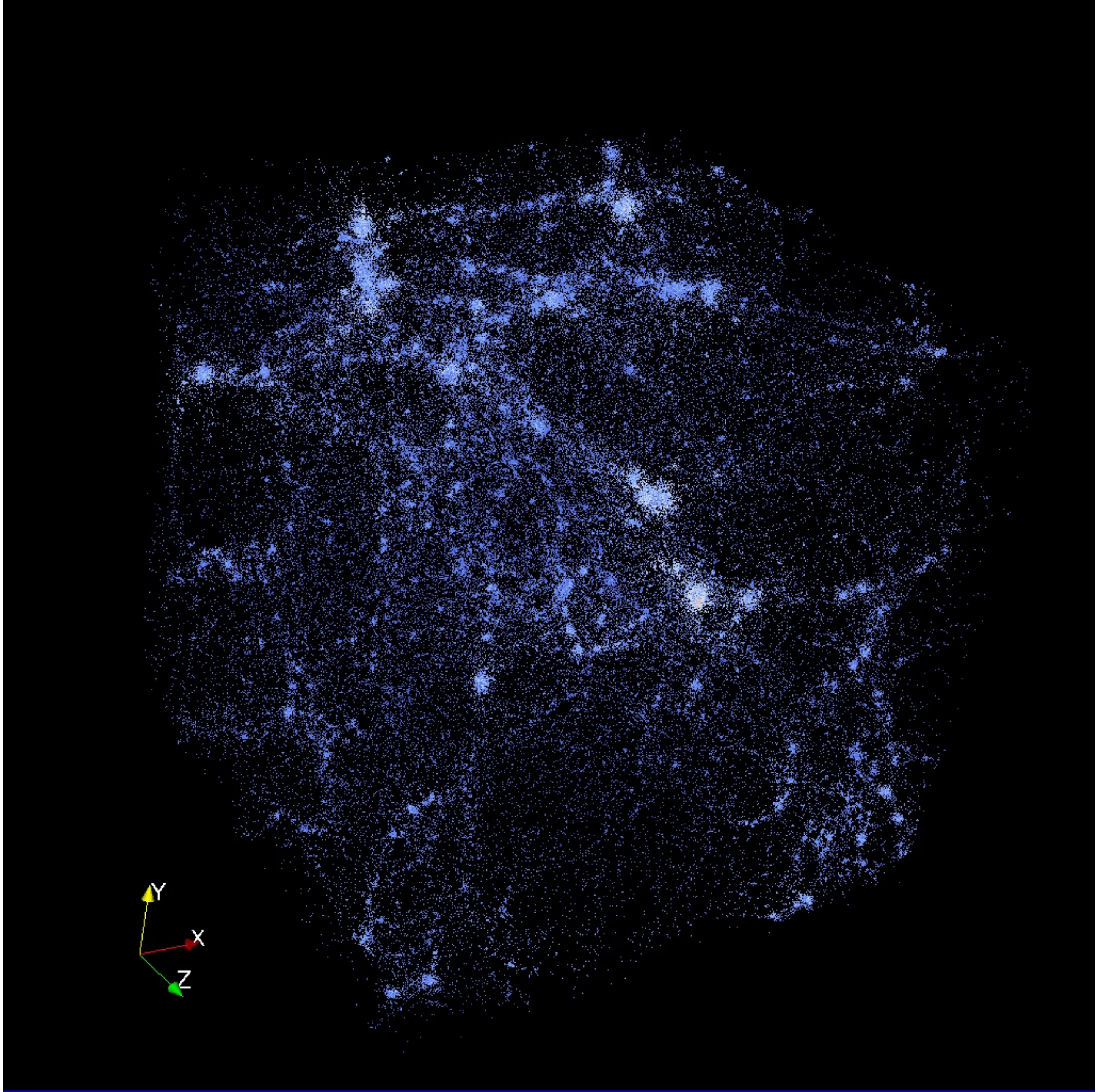


Visual Downsampling

- Large-scale visualization tools (ParaView, VisIt, Ensight, etc.) have been effective, but render everything – a lack of display bandwidth compared to data sizes
- For the MC^3 data there is too much occlusion and clutter to see anything



MC^3 visualization in ParaView

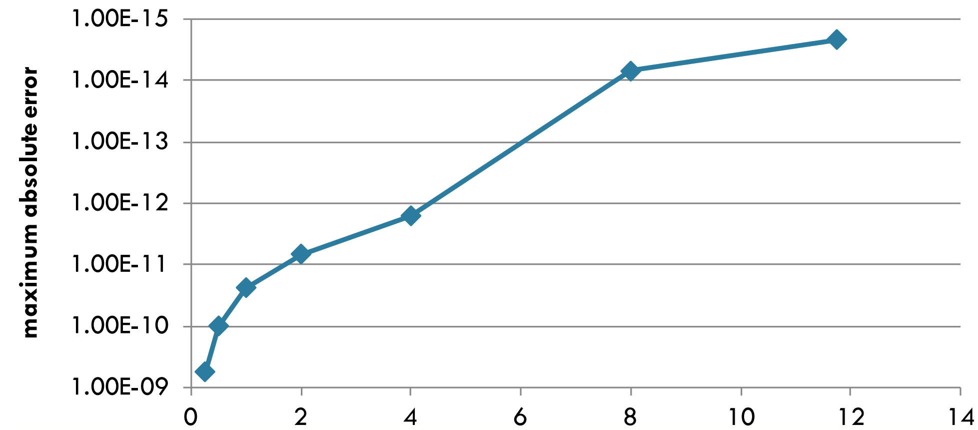

Y
X
Z

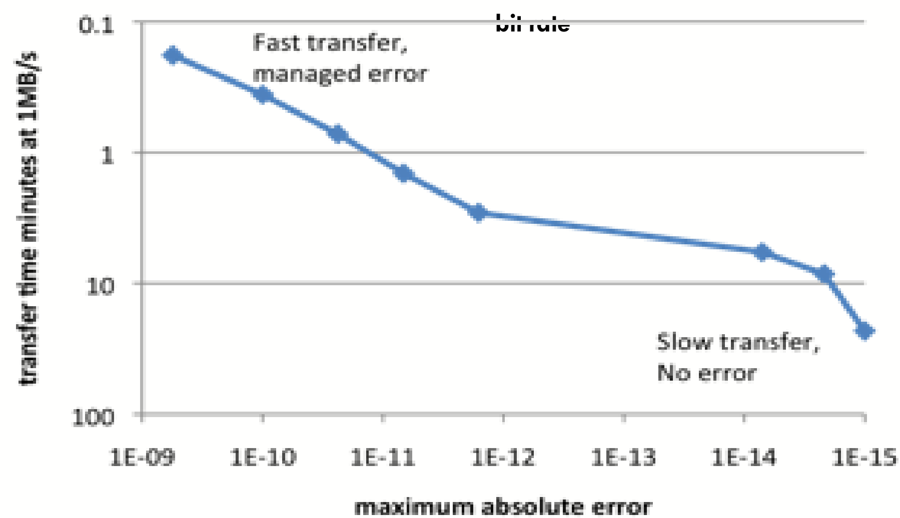
Solution to Massive Data Challenge: Data Compression with Quantified Accuracy

- In visualization and image processing, data compression and the resulting error has been measured as average difference
  - concerned with reducing visual quality differences
- Compression directly in-situ on simulation data as a data reduction mechanism
  - our research focus is to quantify the maximum/L-infinity norm (rather than average/L2 norm) data quality for scientific analysis
  - Provide a solution that automatically compresses simulation data with visualization and analysis accuracy guarantees
- (Simulation Data – Compressed Representation) provides an accuracy metric

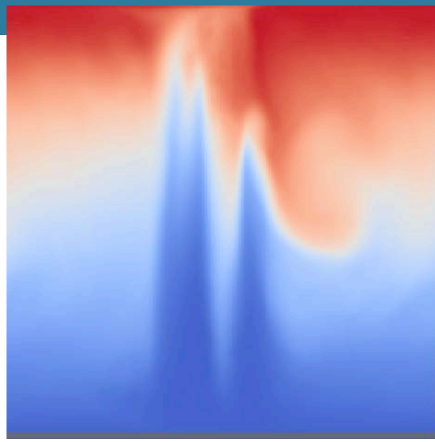
Quantify the Maximum Error (L-infinity norm) so the Scientist Knows the Data Precision

- We measure the maximum point error so there is a guarantee that the data are accurate to x decimal places
- The user can trade read I/O time vs. data accuracy in a quantifiable manner

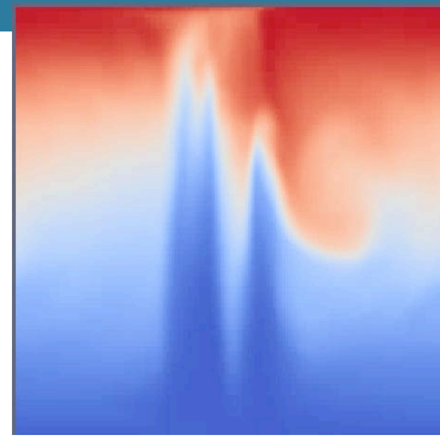




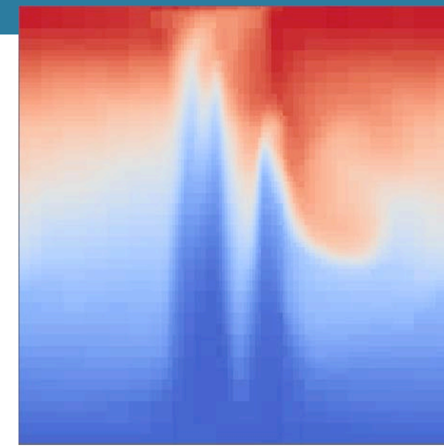
Zoomed Portion of the Local Point Error Difference in Compressed Simulation Data



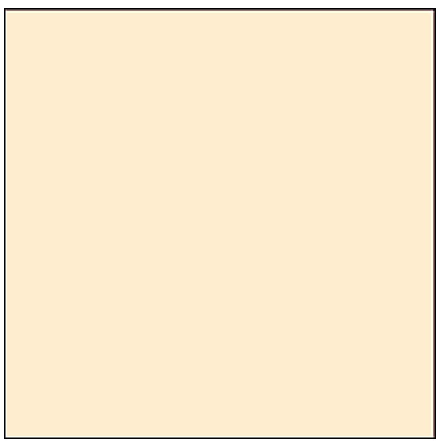
SNR = 130.3 dB



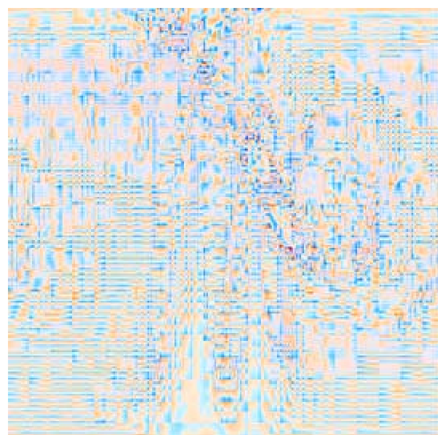
SNR = 48.9 dB



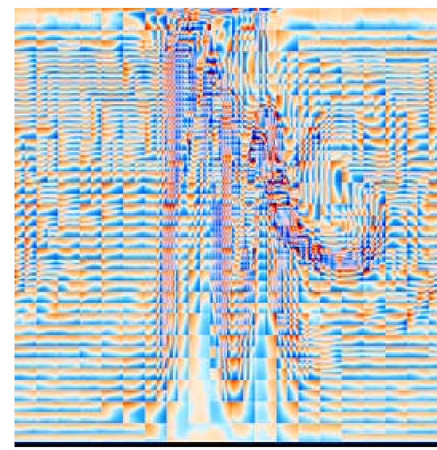
SNR = 41.9 dB



SNR = 130.3 dB



SNR = 48.9 dB

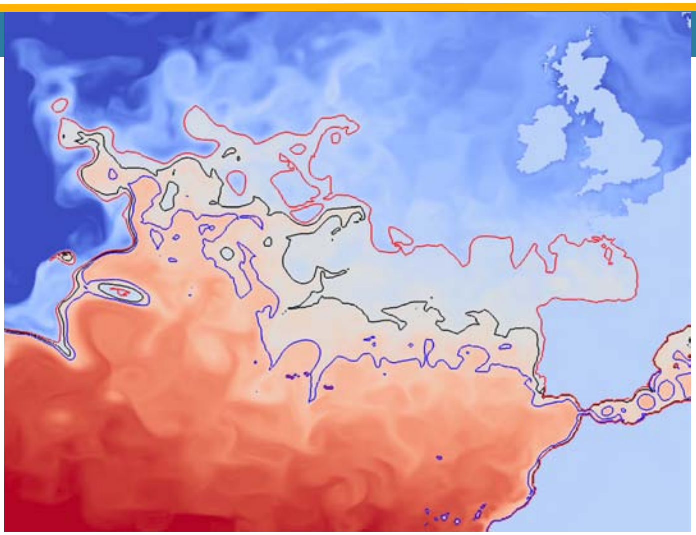


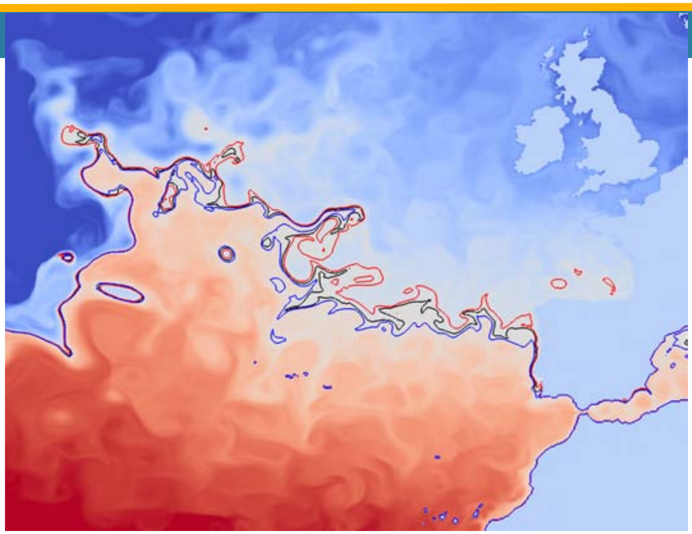
SNR = 41.9 dB

Isovalues on Compressed Simulation Data with Bounding Error - (32 bits, 3200x2400x42, 1.4 GB)

0.25 bits

10.8 MB



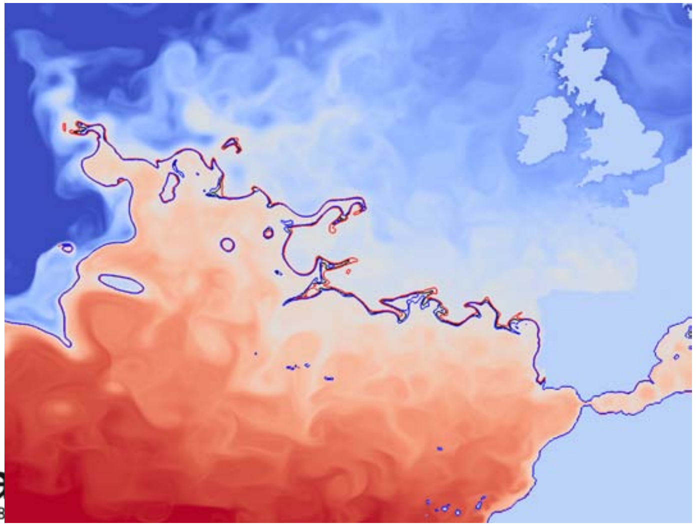


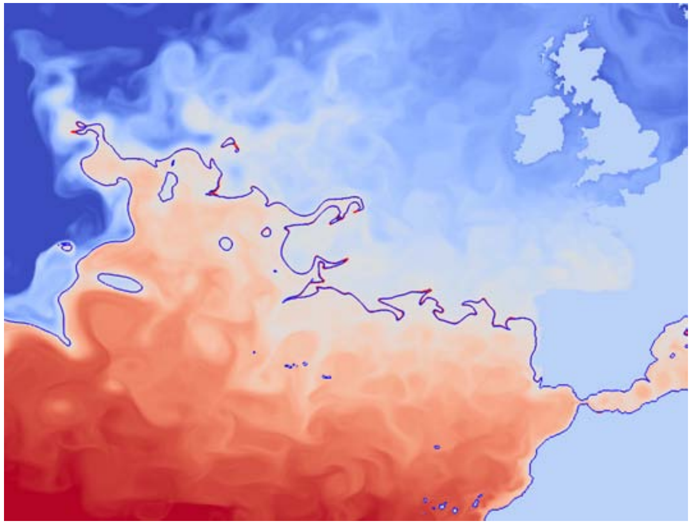
0.5 bits

21.6 MB

1.0 bits

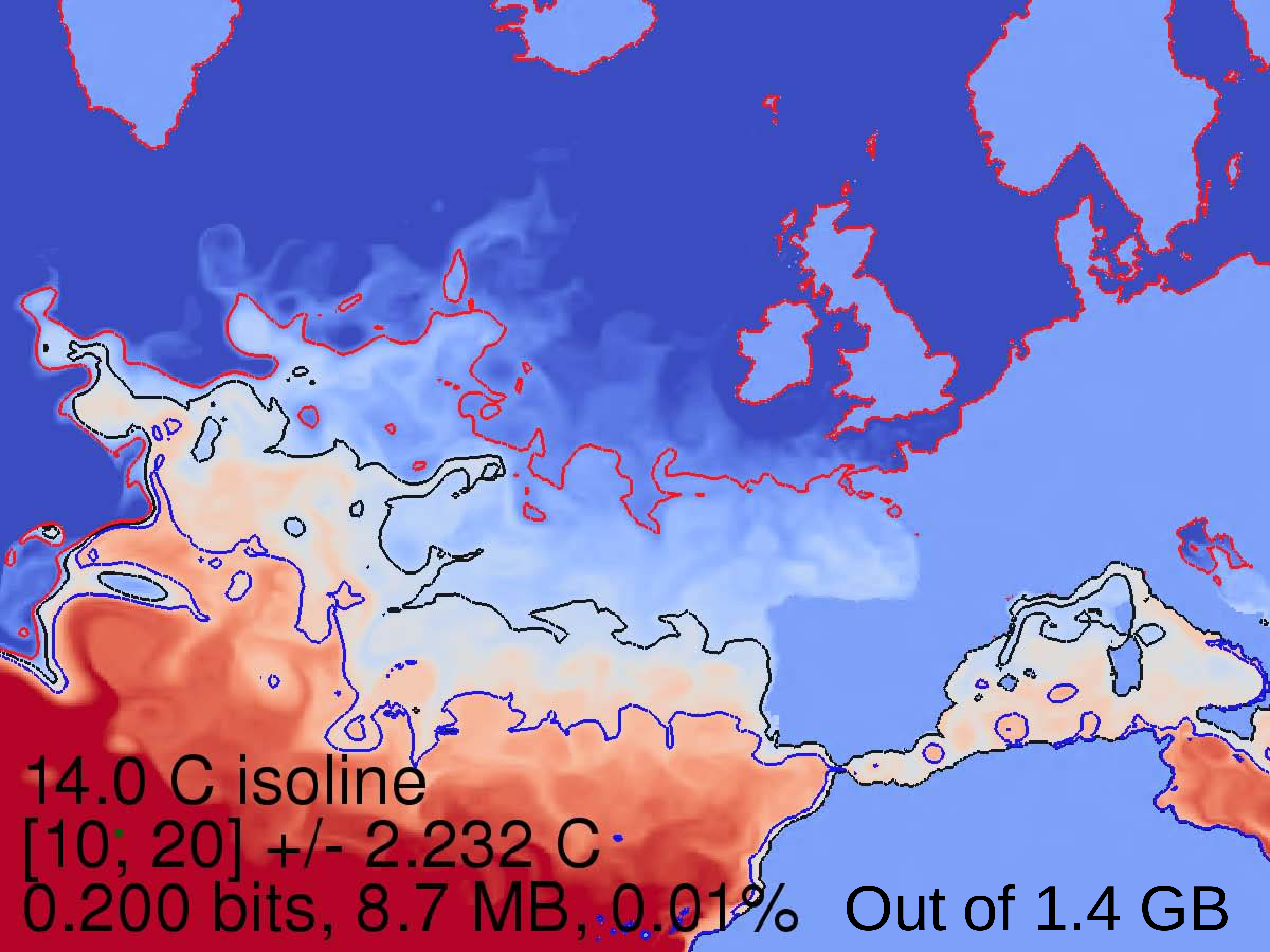
43.3 MB





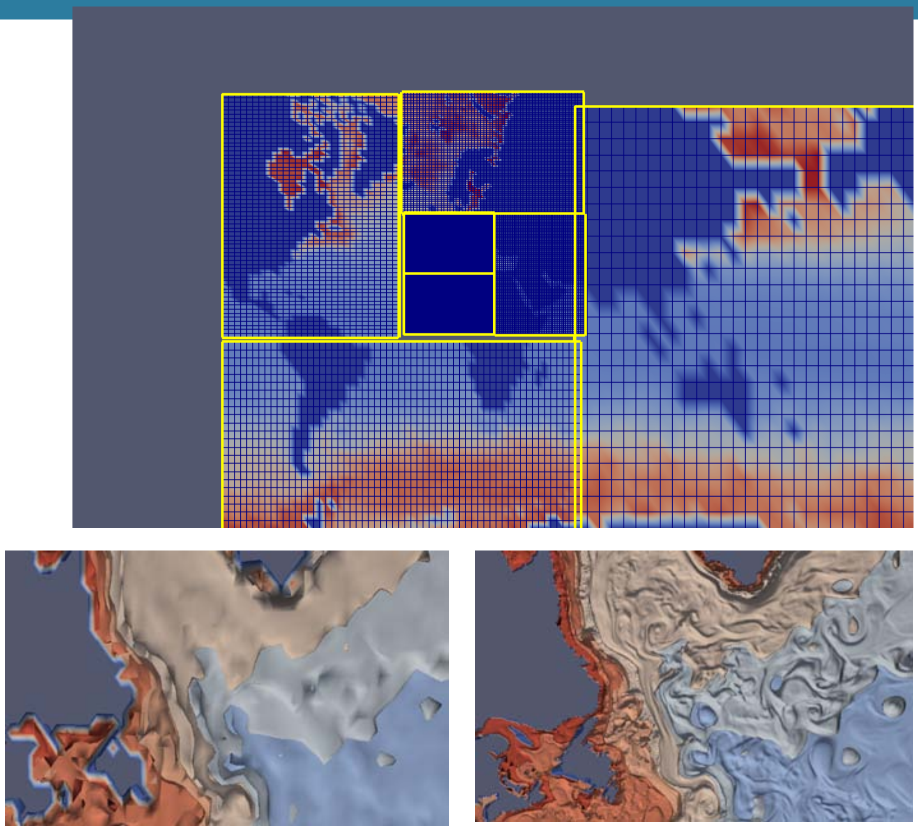
2.0 bits

86.5 MB

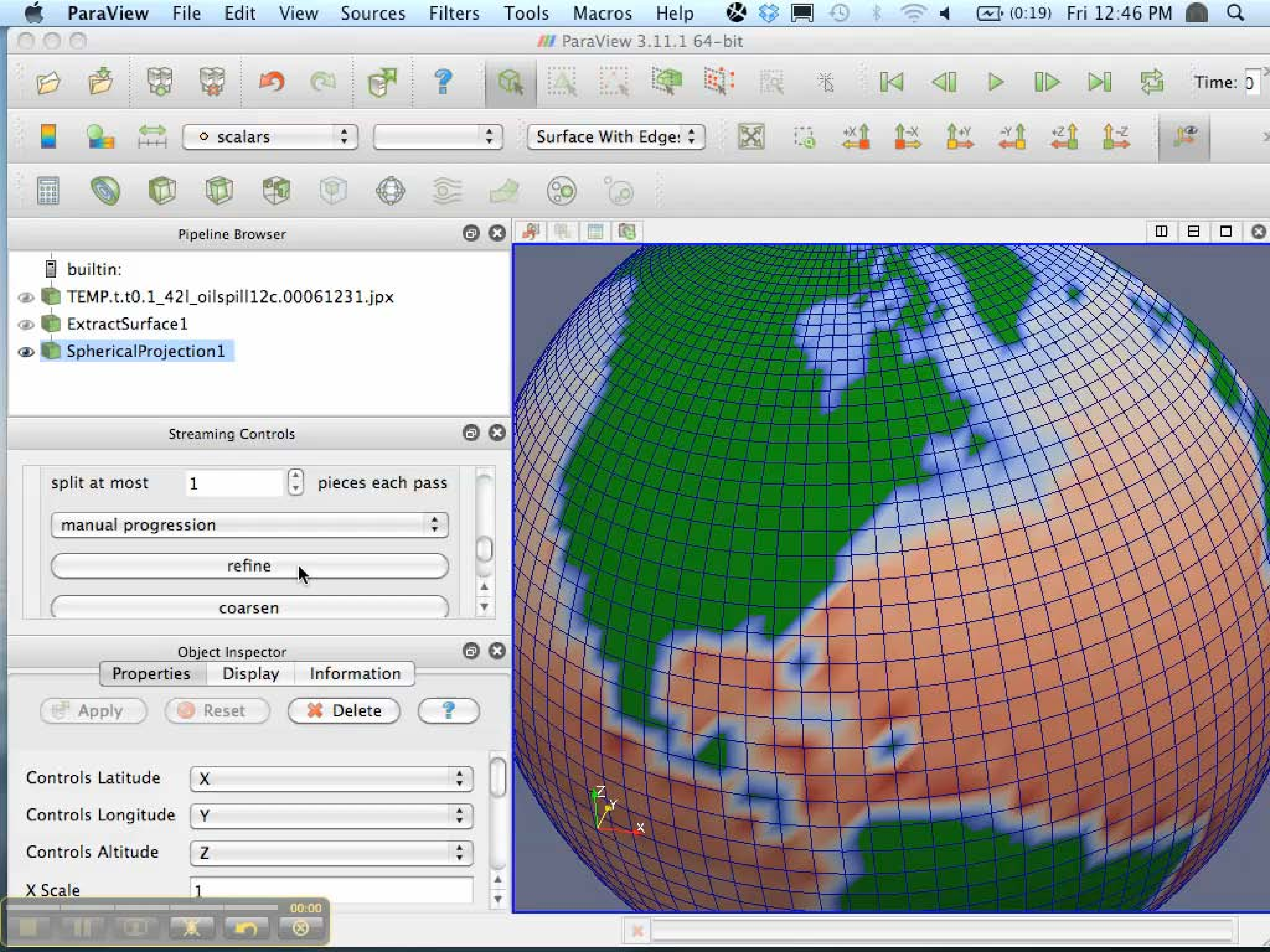

14.0 C isoline
[10; 20] +/- 2.232 C
0.200 bits, 8.7 MB, 0.01%
Out of 1.4 GB

Multi-resolution Compression and Streaming in ParaView

- A multi-resolution representation of simulation data is create using spatial compression or sampling
- View in a multi-resolution visualization and analysis tool
- Mat Maltrud, Climate Scientist, LANL: "This new distance visualization technology will increase our productivity by significantly reducing the amount of time spent in transferring and analyzing our remote data."



Images from multi-resolution streaming ParaView


ParaView File Edit View Sources Filters Tools Macros Help
(0:19) Fri 12:46 PM
ParaView 3.11.1 64-bit
Time: 0
scalars
Surface With Edge:
Pipeline Browser
builtin:
TEMP.t.t0.1_42l_oilspill12c.00061231.jpx
ExtractSurface1
SphericalProjection1
Streaming Controls
split at most
1
pieces each pass
manual progression
refine
coarsen
Object Inspector
Properties
Display
Information
Apply
Reset
Delete
Controls Latitude
X
Controls Longitude
Y
Controls Altitude
Z
X Scale
1
00:00
Z
Y
X

Data reduction summary

Algorithm	Reduction
Data parallelism	Handle large datasets Make reduction possible
Multi-resolution	Make focused exploration possible
Visualization and analysis operators (isosurface)	A dimension reduction
Statistical sampling	1-2 orders of magnitude
Compression	1 order of magnitude
Feature extraction	2 orders of magnitude

Challenge: Changing supercomputing architectures

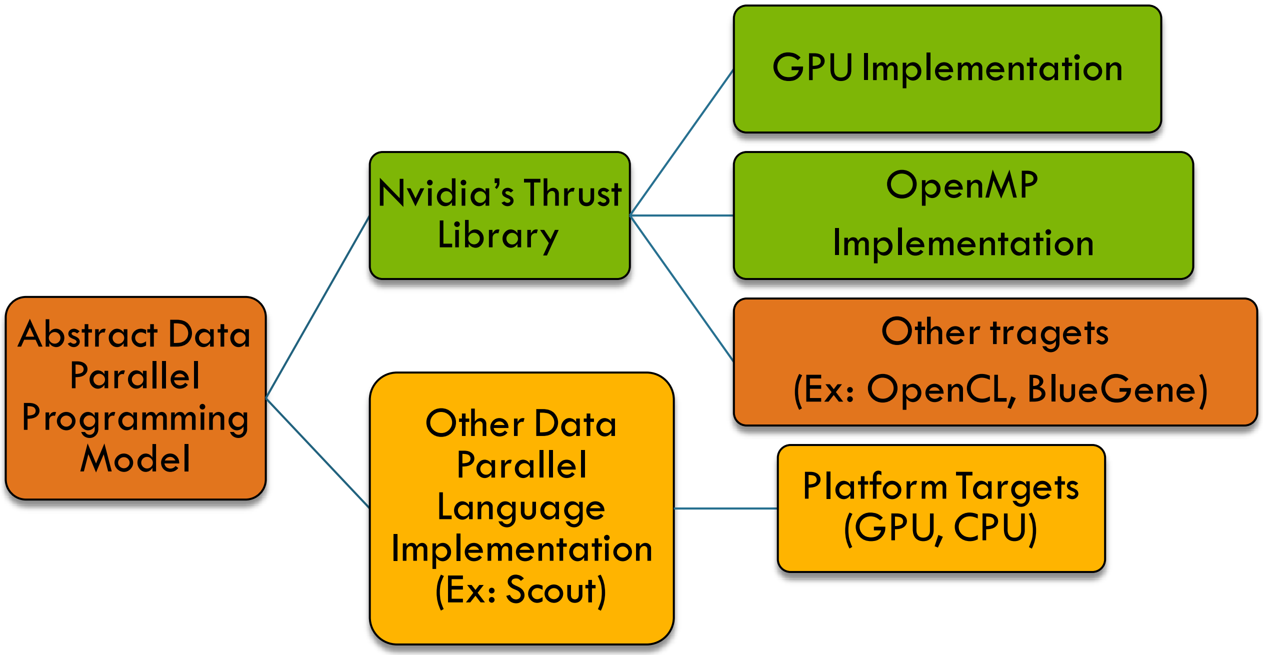
- Solution: new visualization and analysis algorithms, implementations and infrastructures
  - Limited programming resources
  - Many emerging architectures
- How do I best allocate my programming resources?
  - To move field forward we need reusable code base
    - Otherwise spend most resources rewriting

Start with an Isosurface Algorithm

- Important visualization technique
  - Each cell in the data set is examined and isosurface is generated by interpolation
- Highly parallel since each cell can be processed independently
- Numerous research on hardware specific acceleration of the algorithm
  - Vector and SIMD machine
  - GPU with GLSL or CUDA
    - Very low-level programming model, no abstraction used
    - Non-portable across different hardware

Solution: Explore Using Data Parallel Programming Model

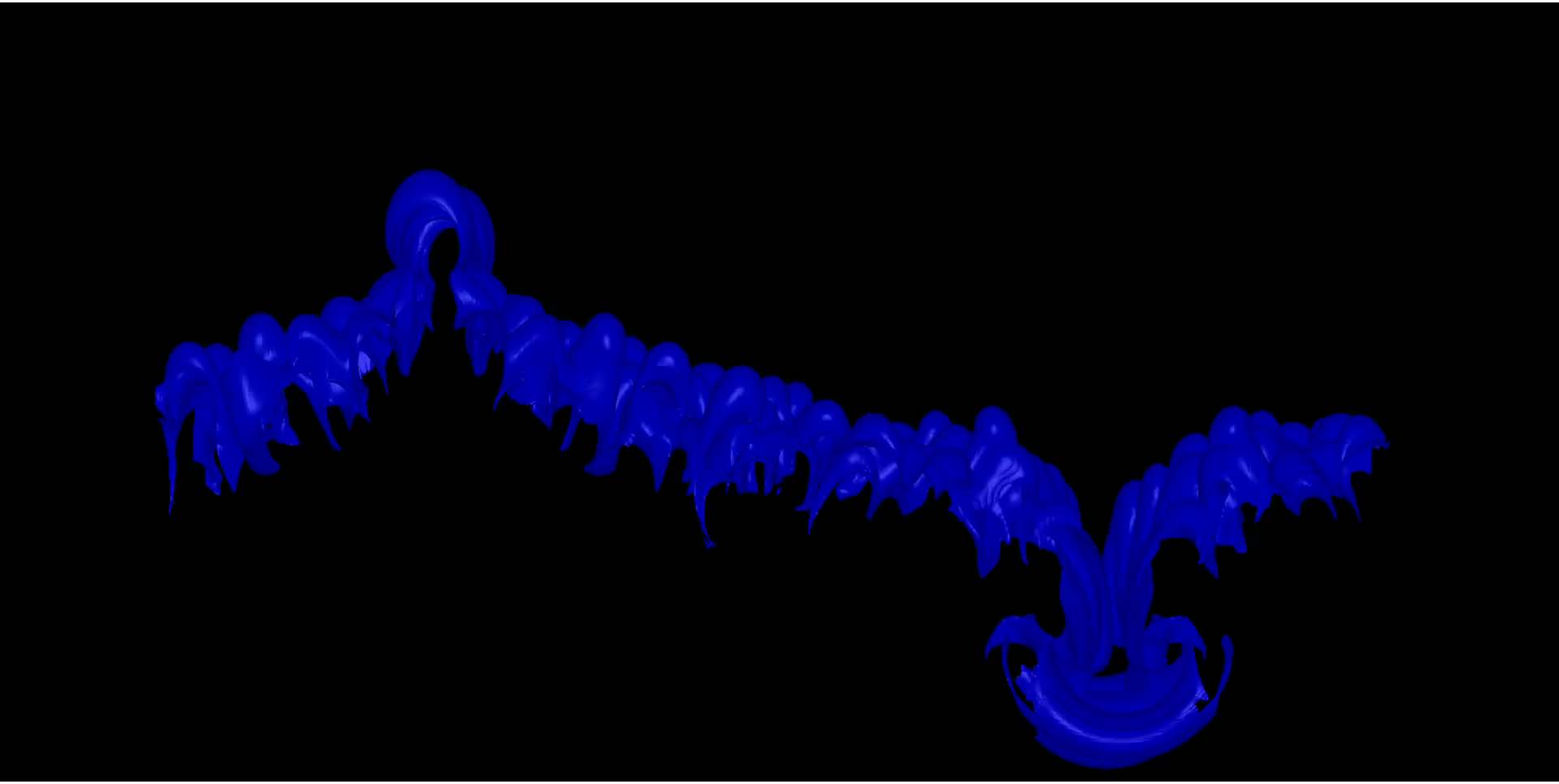
- All operations run on each data element
  - Mathematical, Reductions, Prefix sums, Sorting, Gather/Scatter



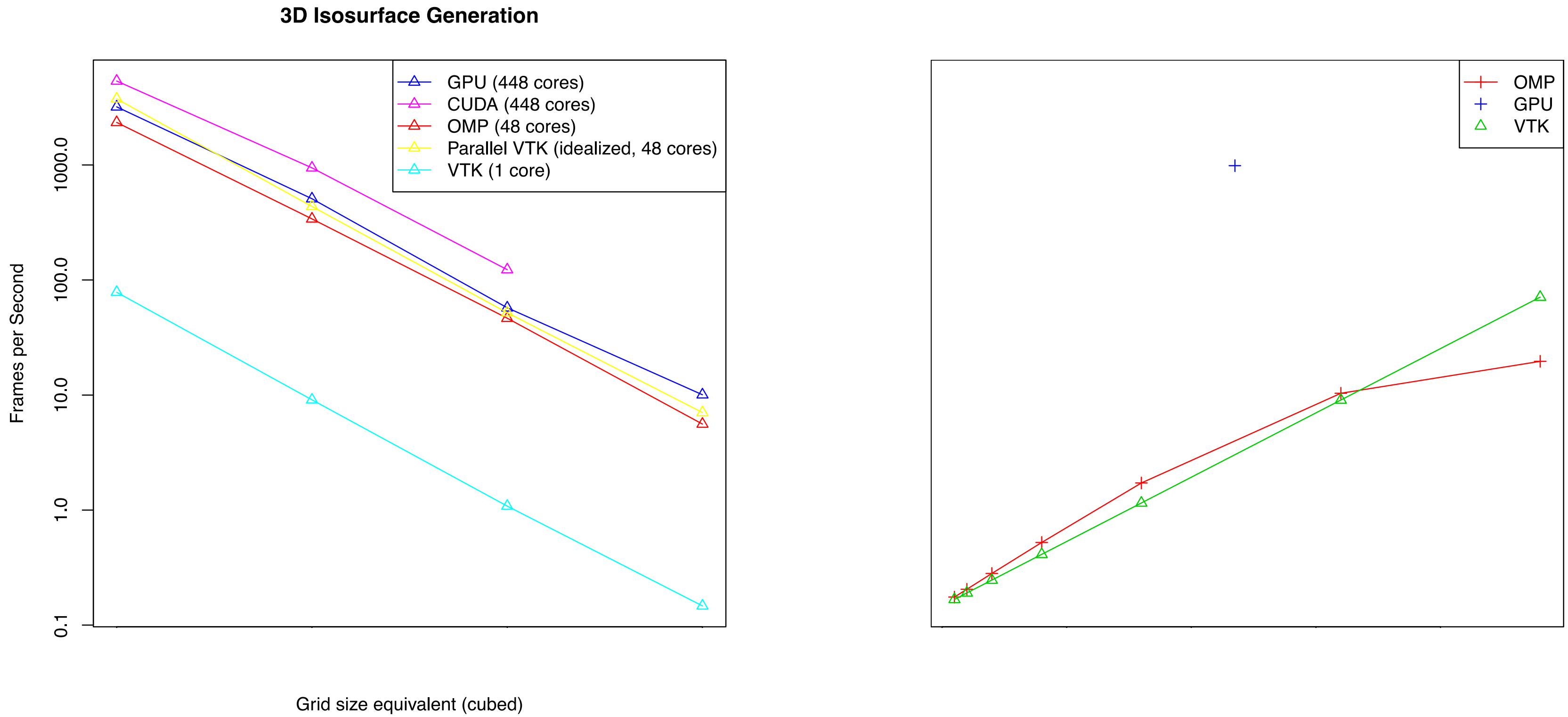
The Thrust Library

- A C++ Template Library for Data-Parallel programming with STL like syntax
- Data can reside on the host or "device" side.

```
#include <thrust/host_vector.h>
#include <thrust/device_vector.h>
#include <thrust/sort.h>

int main(void)
{
    // generate 16M random numbers on the host
    thrust::host_vector<int> h_vec(1 << 24);
    thrust::generate(h_vec.begin(), h_vec.end(), rand);

    // transfer data to the device
    thrust::device_vector<int> d_vec = h_vec;

    // sort data on the device
    thrust::sort(d_vec.begin(), d_vec.end());

    // transfer data back to host
    thrust::copy(d_vec.begin(), d_vec.end(), h_vec.begin());
```



Portable Performance for Isosurface Generation



For more information
Publications (http://viz.lanl.gov)

- **2011**
  - "Revisiting wavelet compression for fast quantified data visualization and analysis", J. Woodring, S. Mniszewski, C. Brislawn and J. Ahrens, in publication, IEEE Large Data Analysis and Visualization Symposium, 2011.
  - Sean J. Williams, Matthew W. Hecht, Mark R. Petersen, Richard Strelitz, Mathew E. Maltrud, James P. Ahrens, Mario Hlawitschka, and Bernd Hamann. "Visualization and Analysis of Eddies in a Global Ocean Simulation". EuroVis 2011, May 31–June 3, Bergen, Norway.
  - Jonathan L. Woodring, James P. Ahrens, Jeannette A. Figg, Joanne R. Wendelberger, and Katrin Heitmann. "In-situ Sampling of a Large-Scale Particle Simulation for Interactive Visualization and Analysis". EuroVis 2011, May 31–June 3, Bergen, Norway.
- **2010**
  - Jonathan Woodring, Katrin Heitmann, James Ahrens, Patricia Fasel, Chung-Hsing Hsu, Salman Habib, and Adrian Pope. "Analyzing and Visualizing Cosmological Simulations with ParaView". Astrophysical Journal Supplements, Oct 2010.

Acknowledgements

- DOE ASC program
- DOE ASCR base program
  - Remote Visualization for Extreme Scale Simulations
- DOE BER Climate Visualization program
  - UV-CDAT project
- DOE LDRD

End