The Impact of Data Complexity on Scientific Visualization Methods

Ken Joy
Director, Institute for Data Analysis and Visualization
Computer Science Department
University of California, Davis
VACET

- SciDAC II Visualization and Analytics Center for Emerging Technologies
  - Lawrence Berkeley National Laboratory
  - Lawrence Livermore National Laboratory
  - Oak Ridge National Laboratory
  - University of Utah
  - UC Davis
What am I going to talk about?

- Data size:
  - O(3/4 PB) raw field data
  - Particle data on Jag

- Data complexity:
  - Data is multi-variant
  - Turbulence chaotic
    - Wide range of
    - High intermittent
      - Time-varying
        - Organized columns
        - Non-locality in
          - Temporal correlated

- GenASiS data get large quickly as we move from moments of the distribution function to the distribution function itself
- Otherwise, data sizes increase modestly
  - The exascale machine will be a "strong scaling" platform (relative dearth of memory)

DoE Exascale Visualization Workshop, Houston Texas, February 22-23, 2011

(Courtesy, Sean Ahern)
### The Opportunity

**Data Volumes**

<table>
<thead>
<tr>
<th>Code</th>
<th># Variables</th>
<th>Resolution</th>
<th># Dumps</th>
<th>Total Volume</th>
<th>Runtime</th>
<th>Machine</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHIMERA 1.0</td>
<td>~ 200</td>
<td>576x96x192</td>
<td>3000</td>
<td>~50 TB</td>
<td>~ 3 Months</td>
<td>1 PF</td>
</tr>
<tr>
<td>CHIMERA 2.0</td>
<td>~ 350</td>
<td>576x96x192</td>
<td>3000</td>
<td>~90 TB</td>
<td>~ 3 Months</td>
<td>20 PF</td>
</tr>
<tr>
<td>GenASiS</td>
<td>~5000</td>
<td>512x512x512</td>
<td>3000</td>
<td>~30 PB</td>
<td>?</td>
<td>1 EF</td>
</tr>
</tbody>
</table>

- GenASiS data get large quickly as we move from moments of the distribution function to the distribution function itself.
- Otherwise, data sizes increase modestly.
  - The exascale machine will be a “strong scaling” platform (relative dearth of memory).

(Courtesy, Bronson Messer, ORNL)
The Opportunity

Data Volumes

<table>
<thead>
<tr>
<th>Code</th>
<th># Variables</th>
<th>Resolution</th>
<th># Dumps</th>
<th>Total Volume</th>
<th>Runtime</th>
<th>Machine</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHIMERA 1.0</td>
<td>~ 200</td>
<td>576X96X192</td>
<td>3000</td>
<td>~50 TB</td>
<td>~ 3 Months</td>
<td>1 PF</td>
</tr>
<tr>
<td>CHIMERA 2.0</td>
<td>~ 350</td>
<td>(expanded nuclear</td>
<td>576X96X192</td>
<td>3000</td>
<td>~90 TB</td>
<td>~ 3 Months</td>
</tr>
<tr>
<td></td>
<td>(expanded nuclear</td>
<td>network</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>α to 150-species)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GenASiS</td>
<td>~5000</td>
<td>512X512X512</td>
<td>3000</td>
<td>~30 PB</td>
<td>?</td>
<td>1 EF</td>
</tr>
</tbody>
</table>

- GenASiS data get large quickly as we move from moments of the distribution function to the distribution function itself.
- Otherwise, data sizes increase modestly
  - The exascale machine will be a “strong scaling” platform (relative dearth of memory)
### The Opportunity

<table>
<thead>
<tr>
<th>Code</th>
<th># Variables</th>
<th>Resolution</th>
<th># Dumps</th>
<th>Total Volume</th>
<th>Runtime</th>
<th>Machine</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHIMERA 1.0</td>
<td>~ 200</td>
<td>576X96X192</td>
<td>3000</td>
<td>~50 TB</td>
<td>~ 3 Months</td>
<td>1 PF</td>
</tr>
<tr>
<td>CHIMERA 2.0</td>
<td>~ 350</td>
<td>576X96X192</td>
<td>3000</td>
<td>~90 TB</td>
<td>~ 3 Months</td>
<td>20 PF</td>
</tr>
<tr>
<td>GenASiS</td>
<td>~5000</td>
<td>512X512X512</td>
<td>3000</td>
<td>~30 PB</td>
<td>?</td>
<td>1 EF</td>
</tr>
</tbody>
</table>

- GenASiS data get large quickly as we move from moments of the distribution function to the distribution function itself.
- Otherwise, data sizes increase modestly.
  - The exascale machine will be a “strong scaling” platform (relative dearth of memory).

(Courtesy, Bronson Messer, ORNL)
Challenges

• The number of variables studied in large-scale simulations will continue to increase.
  
  • ...interactions/correlations between the variables explodes

• So -- there are “scale” difficulties upcoming in visualization that are based on the complexity of our data
First

• “Fields” are becoming more complex...
  • Vector fields
  • Tensor fields
  • Function Fields
  • Material Fractions
  • Uncertainty
  • Ensembles
  • Distribution Fields
  • Others...
Vector Fields

- In the past five years, we have finally found good ways to represent flow.
- Based on “integral curve methods”

Streamlines

Streamsurfaces
Vector Fields

- In the past five years, we have finally found good ways to represent flow.
  - Based on “integral curve methods”
Vector Fields

- In the past five years, we have finally found good ways to represent flow.
  - Based on “integral curve methods”
Unsteady vs. Steady

(Courtesy, Christoph Garth, UC Davis)
Streak Surfaces

(Courtesy, Christoph Garth, UC Davis)
Streak Surfaces

(Courtesy, Christoph Garth, UC Davis)
Time Surfaces

(Courtesy, Hari Krishnan, Christoph Garth, UC Davis)
Particle movement on the surface is indicated by stripes with animated highlights.

(Courtesy, Mattias Hummel, Christoph Garth)
Lagrangian Methods

Finite-time Lyapunov Methods
Parallel Methods for Visualizing Flow

- Flow visualization methods do not parallelize well.

- We have several results...

- Working on “-surfaces”

Interesting to note: FTLE -will- scale
The “Fish Tank”


(Courtesy, Hank Childs, Christoph Garth)
The “Fish Tank”
The “Fish Tank”
Field Type: Volume Fractions

(.65,.35)  (.1,.9)  (0,1)

(1,0)  (.8,.2)  (.12,.88)

(1,0)  (1,0)  (.64,.36)
Embedded Boundary/Material Interfaces

Based on Adaptive Interface Methods

Data Courtesy of APDEC

(John Anderson et al., UC Davis)
Embedded Boundary/Material Interfaces

Cross section of a “swirler” (Courtesy of APDEC). The swirler is represented by volume fractions in a uniform grid, and boundaries are generated by active contour methods.
Material Interface Methods

Classic approximations

Higher-order Approximations

(Juri Prilepov, UC Davis)

(Juri Prilepov, UC Davis)
Multi-Material Works Now

Three-Material Example

(Iuri Prilepov, UC Davis)
Reproduces Fine Detail Better

Iuri Prilepov, UC Davis

Elongated Bubble

Three-Material Example
At each vertex of a mesh, one is given a “function”.

Air-pollution data from the San Joaquin Valley, CA. Each vertex has an associated function [particle size by number of particles]. This frame is from a large-scale 24-hour simulation of the air quality in the valley.

(John Anderson, UC Davis)
Function Fields
Function Fields
More Function Fields

• Time-Varying Vector Fields
Multi-Temporal Visualization

This shows output from a Raleigh-Taylor simulation, where the color indicates the time at which mixing has occurred in the simulation.
Multi-temporal visualization

This multi-temporal image depicts summary statistic information gathered from 18 time steps from the Argon Bubble dataset. Yellow regions indicate areas where high gas density predominantly resides over the course of the Argon Bubble simulation.

This multi-temporal image depicts summary statistic information gathered from queries processed over 48 time steps from a Hurricane dataset. In this image, yellow regions indicate where low pressure predominantly exists across the 48 timesteps.
Distribution Fields

- One has a distribution given at each data point

  Studying 25 Rayleigh-Taylor Instability calculations (all at 10us)
  Two “knobs”: turbulent viscosity coefficient, buoyancy coefficient
  Five values for each knob, 25 pairs total

  Average Speed over all 25
  Max Speed over all 25
  Min Speed over all 25
  Biggest difference over all 25

(Hank Childs)
Much more work to do!

- **Function Fields, Distribution Fields**
  - Only early results

- **Ensembles**
  - Query-driven methods

- **Vector Fields (and other field types)**
  - Parallel Implementation
How do you solve the “multi-field” Problem

• Query-driven Data Exploration

  • The data is so large, and so complex, that we may as well treat it as a data base.

  • We can “mine” previous work on data bases

  • We can address uncertainty through, “Fuzzy Data Bases”, and “Uncertain Queries”
Data Complexity will become a primary constraint in the analysis and visualization of scientific simulations in the immediate future.

We have made substantial progress over the past four+ years. But much more to do.
Thank to all the help.

- Contributors
  - Christoph Garth (UC Davis, Kaiserslautern), Harald Obermaier (UC Davis), Hank Childs (LBNL, UC Davis), Sean Ahern (ORNL), Valerio Pascucci (Utah), Peer-Timo Bremer (LLNL, Utah), Kathleen Bonnell (LLNL), John Anderson (Makai Ocean Engineering), Luke Gosink (PNNL, UC Davis), Iuri Prileov (UC Davis), Eduard Deines (UC Davis), Hans Hagen (Kaiserslautern), Mattias Hummel (Kaiserslautern), Hari Krishnan (UC Davis, LBNL), Wes Bethel (LBNL), Chris Johnson (Utah), Kevin Bessemer (UC Davis), George Ostrachov (ORNL), Mark Miller (LLNL), David Camp (LBNL, UC Davis), Alan Sanderson (Utah), Xavier Tricoche (Purdue), Sean Williams (LANL, UC Davis), Sohail Shafi (UC Davis), Rob Sisneros (ORNL).

Things I don’t have time to talk about!

Monday, November 28, 11