Knowledge Discovery panel
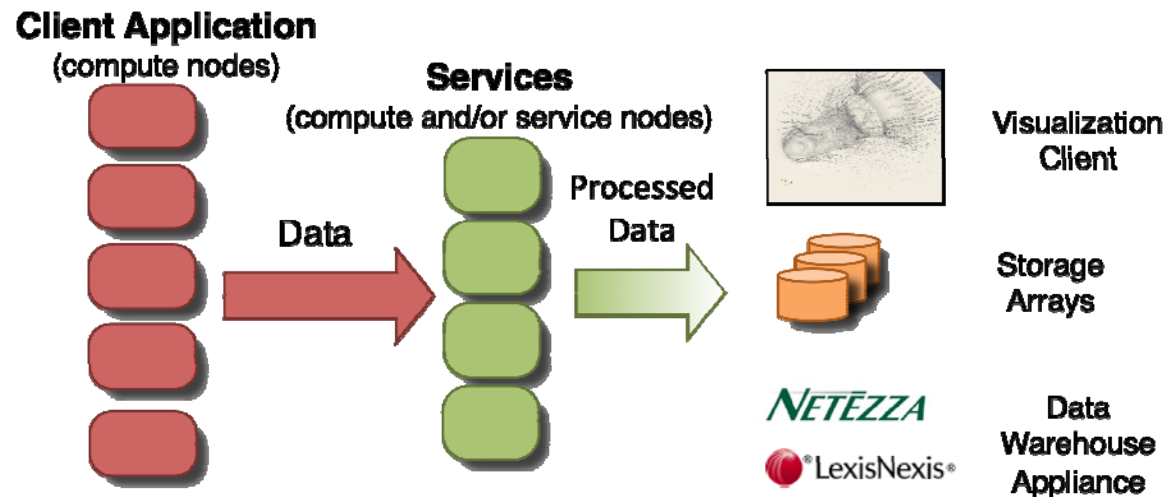
At the crossing point between data analysis and simulation

Ron Oldfield
Sandia National Laboratories
March 11th, 2010

Nachhaltige Begegnungen

What are the challenges for bridging data analysis and simulation in your field?

- File systems are too slow and they are often used for transient data

- Do your analysis with your simulation ... avoid I/O

Client Application
(compute nodes)

Services
(compute and/or service nodes)

Data

Processed Data

Visualization Client

Storage Arrays

NETEZZA

LexisNexis

Data Warehouse Appliance

Network Scalable Service Interface (Nessie)

- Developed for the Lightweight File Systems Project
- Framework for HPC client/server development
- Designed for scalable data movement
- RPC-like API (client and server stubs)

Nessie
NEtwork-Scalable Service InterfacE

I/O Services Example
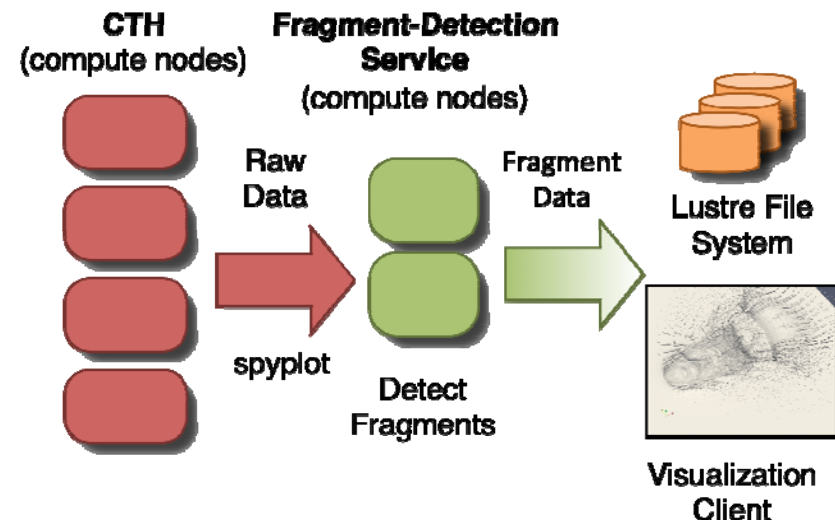## CTH Fragment Detection

Motivation

- Fragment detection requires data from every time step (I/O intensive)
- Detection process takes 30% of time-step calculation (scaling issues)
- Integrating detection software with CTH is intrusive on developer

CTH fragment detection service

- Extra compute nodes provide in-line processing (overlap fragment detection with time step calculation)
- Only output fragments to storage (reduce I/O)
- Non-intrusive
  - Looks like normal I/O (spyplot interface)
  - Can be configured out-of-band

Status

- Developing client/server stubs for spyplot
- Developing Paraview proxy service

Fragment detection service provides on-the-fly data analysis with no modifications to CTH.

Contact: Ron Oldfield raoldfi@sandia.gov

- Not really sure… we're trying all sorts of things though

SICAIDA

Storage-Intensive Computing Architectures for In-situ Data Analysis

Research Questions
- Can we leverage data warehouse appliances (e.g., Netezza) for ASC analysis
- Can we utilize data parallel languages for scientific calculations?

Accomplishments and Activities
- Volume calculation study (ongoing)
  - What is total volume of a gas mixture in a region?
  - Easy in C++, not bad in MapReduce, painful in SQL
- Platforms study (ongoing)
  - Survey of different appliances/languages
  - Ported k-Nearest Neighbors (kNN) to SQL, Map/Reduce

Contact: Craig Ulmer cdulmer@sandia.gov
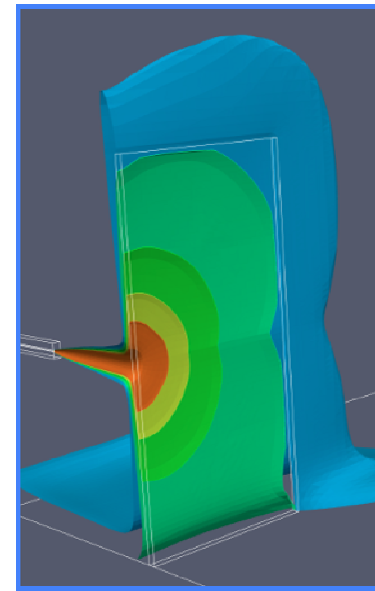
Other Questions

- **How would you define the "P" of HPC in your community: Performance, Productivity, Portability, Pain, ...?**

    - Performance... and pain (that's our niche)

- **What would your wish list from the HPC community consist of?**

    - If I wished for it and it came true, then I wouldn't be able to experience the agony of trying to do it.
    - Here's what I'm looking forward to:
        - Fun with resilience (all sorts of issues there)
        - Fun with architectures and programming models
          Shooting at the moving targets
        - Process movement instead of data movement
          Locality seems more important today than ever before... woo hoo!

Other Questions

- **In your field, do you expect the data wealth and complexity to influence simulation models and methods?**

    - Does system capacity/bandwidth influence data wealth?
    - If you had unlimited storage, would you dump more data?