Sandia National Laboratories

SAND2011-8691P

National Nuclear Security Administration

Examples of In Transit Visualization

Petascale Data Analytics: Challenges and Opportunities

November 14, 2011

Kenneth Moreland (SNL), Ron Oldfield (SNL), Pat Marion (Kitware), Sebastien Jourdain (Kitware), Norbert Podhorszki (ORNL), Venkatram Vishwanath (ANL), Nathan Fabian (SNL), Ciprian Docan (Rutgers), Manish Parashar (Rutgers), Mark Hereld (ANL), Michael E. Papka (ANL), Scott Klasky (ORNL)

SAND 2011-XXXXP

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

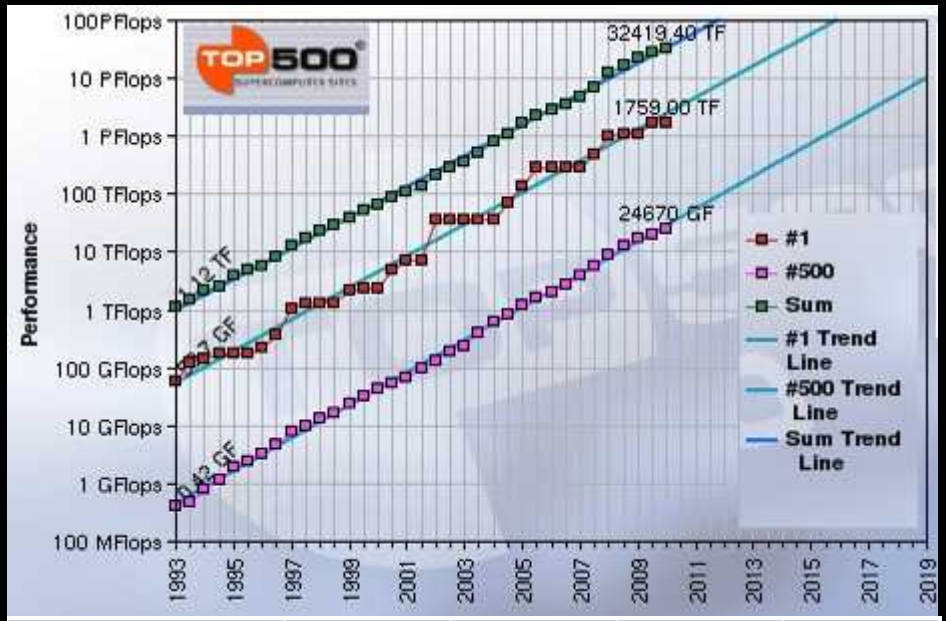
Collaborators

- **Sandia National Laboratories**
  - Kenneth Moreland
  - Nathan Fabian
  - David Thompson
  - Ron Oldfield
- **Los Alamos National Laboratories**
  - James Ahrens
  - Jonathan Woodring
- **Oak Ridge National Laboratory**
  - Scott Klasky
  - Norbert Podhorszki
- **Argonne National Laboratory**
  - Venkatram Vishwanath
  - Mark Hereld
  - Michael E. Papka
- **Kitware, Inc.**
  - Berk Geveci
  - Utkarsh Ayachit
  - Andrew C. Bauer
  - Pat Marion
  - Sebastien Jourdain
  - David DeMarle
- **University of Colorado at Boulder**
  - Michel Rasquin
  - Kenneth E. Jansen
- **Rutgers University**
  - Ciprian Docan
  - Manish Parashar

Extreme scale computing

Sandia National Laboratories

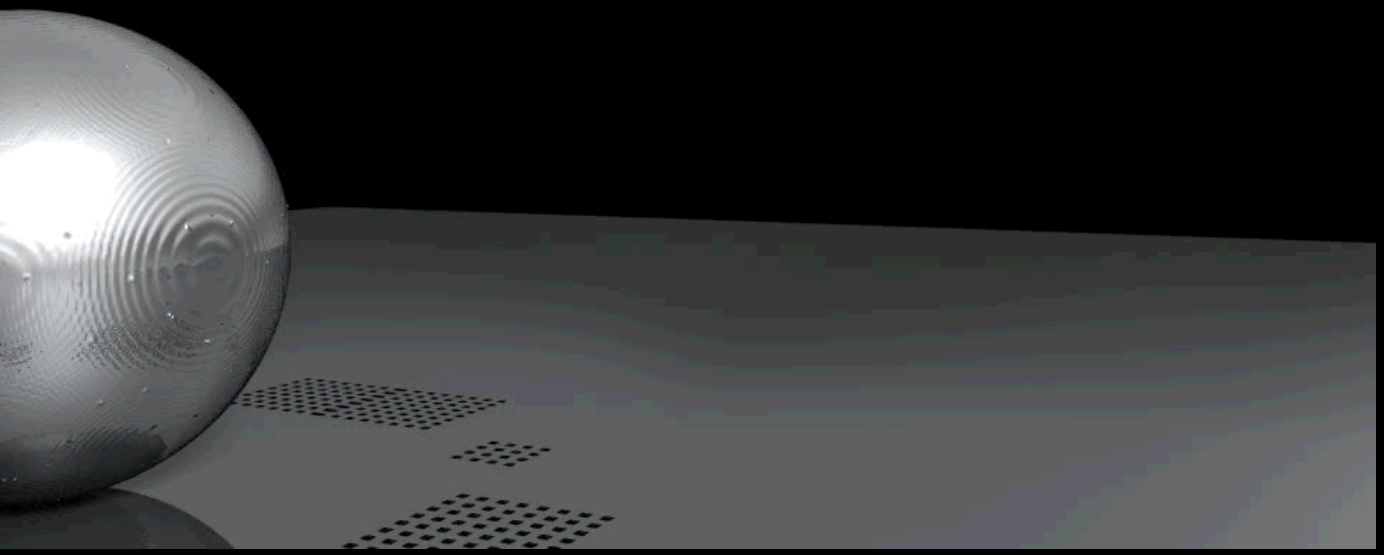
- Trends
  - More FLOPS
  - More concurrency
  - Comparatively less storage, I/O bandwidth
- ASCI purple (49 TB/140 GB/s) – JaguarPF (300 TB/200 GB/s)
  - Most people get < 5 GB/sec at scale



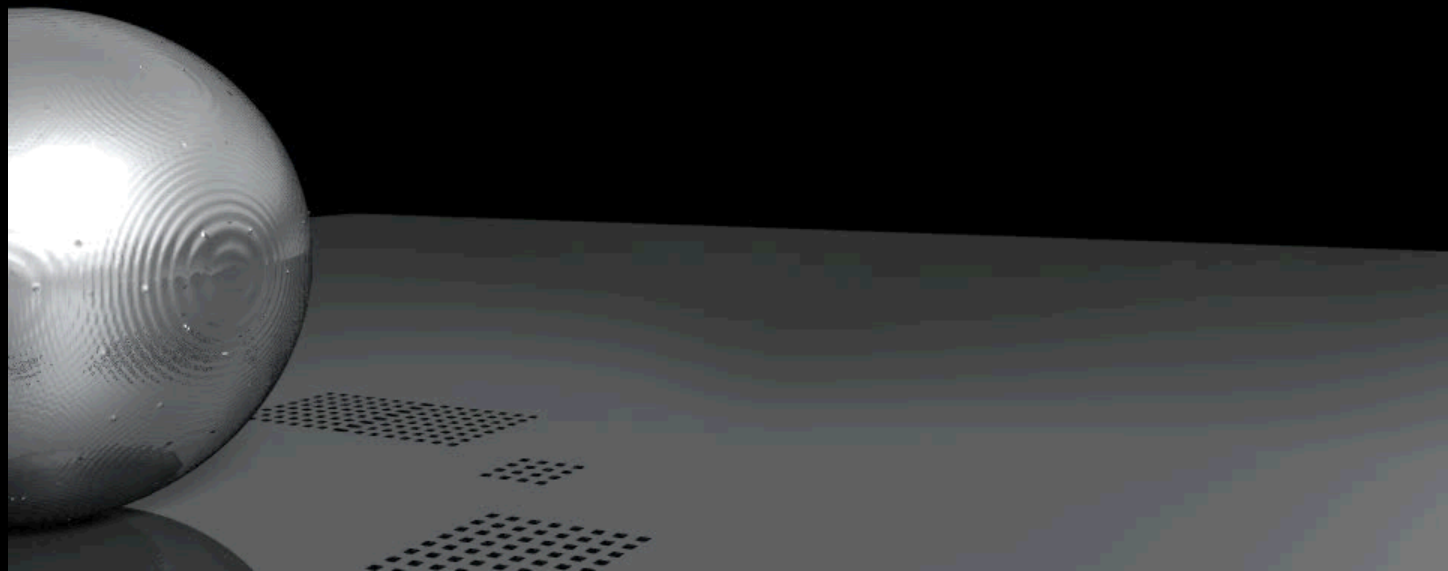
Systems	2009	2011	2015	2018
System Peak Flops/s	2 Peta	20 Peta	100-200 Peta	1 Exa
System Memory	0.3 PB	1 PB	5 PB	10 PB
Node Performance	125 GF	200 GF	400 GF	1-10 TF
Node Memory BW	25 GB/s	40 GB/s	100 GB/s	200-400 GB/s
Node Concurrency	12	32	O(100)	O(1000)
Interconnect BW	1.5 GB/s	10 GB/s	25 GB/s	50 GB/s
System Size (Nodes)	18,700	100,000	500,000	O(Million)
Total Concurrency	225,000	3 Million	50 Million	O(Billion)
Storage	15 PB	30 PB	150 PB	300 PB
I/O	0.2 TB/s	2 TB/s	10 TB/s	20 TB/s
MTTI	Days	Days	Days	O(1Day)
Power	6 MW	~10 MW	~10 MW	~20 MW

ILLUSTRATION: A. TOVEY

From J. Dongarra, "Impact of Architecture and Technology for Extreme Scale on Software and Algorithm Design," Cross-cutting Technologies for Computing at the Exascale, February 2-5, 2010.

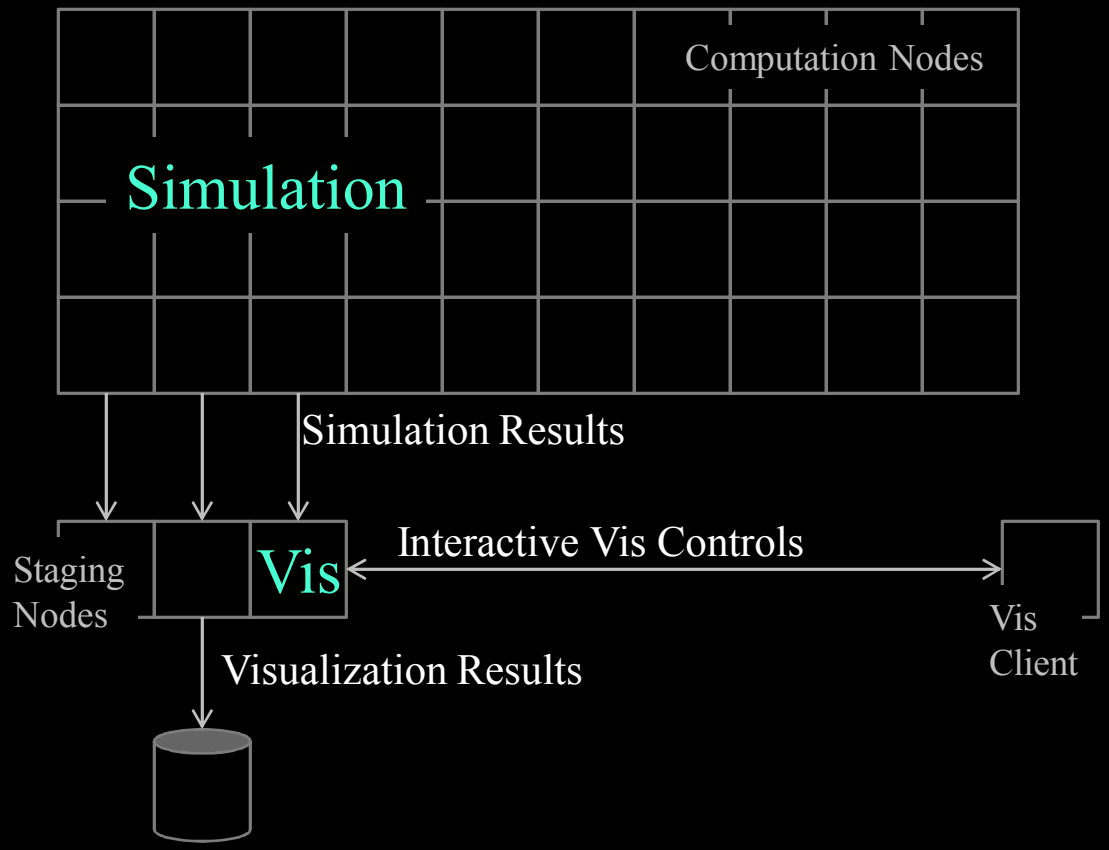


1 mesh per 100 steps
~1% write time

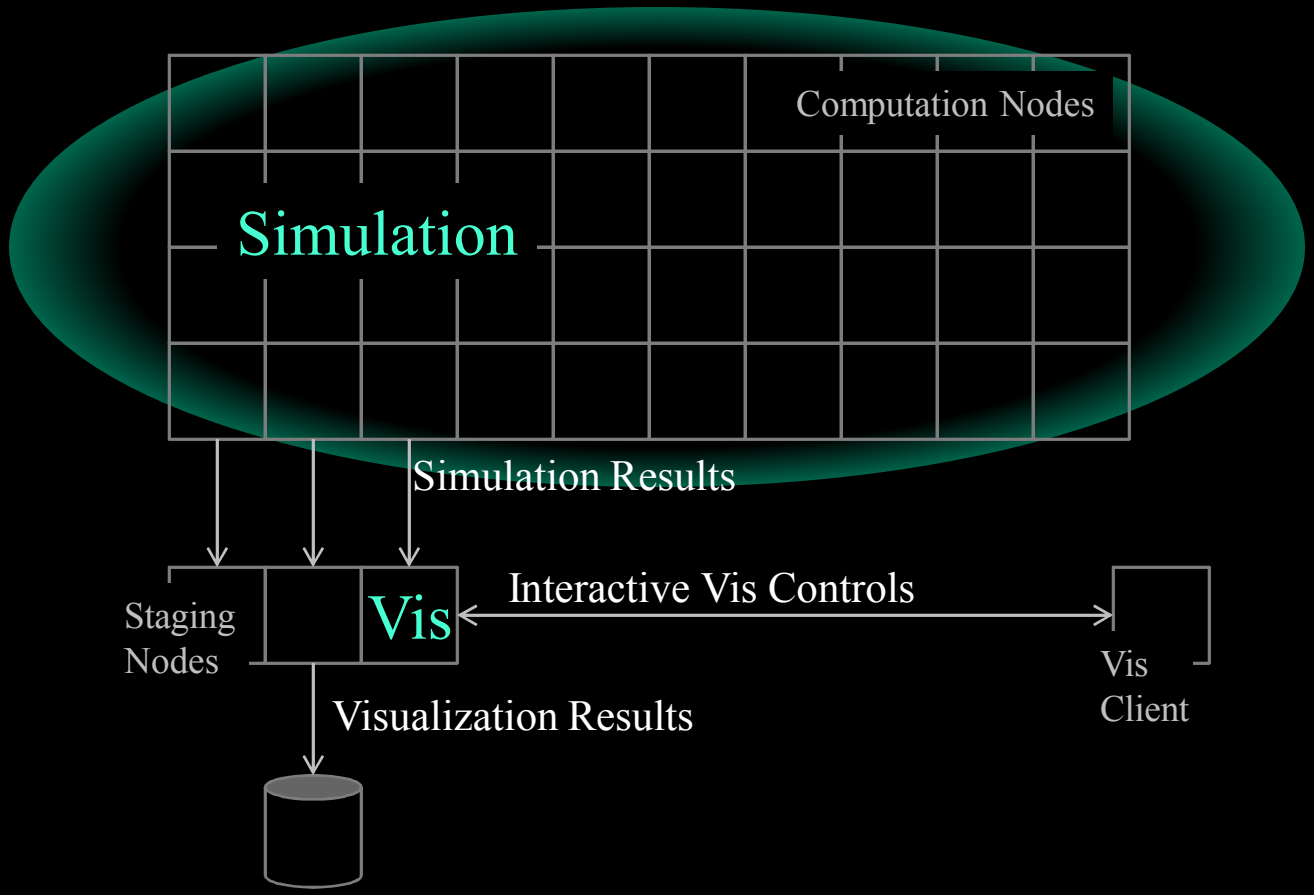


1 image per 1 step
~0.1% write time

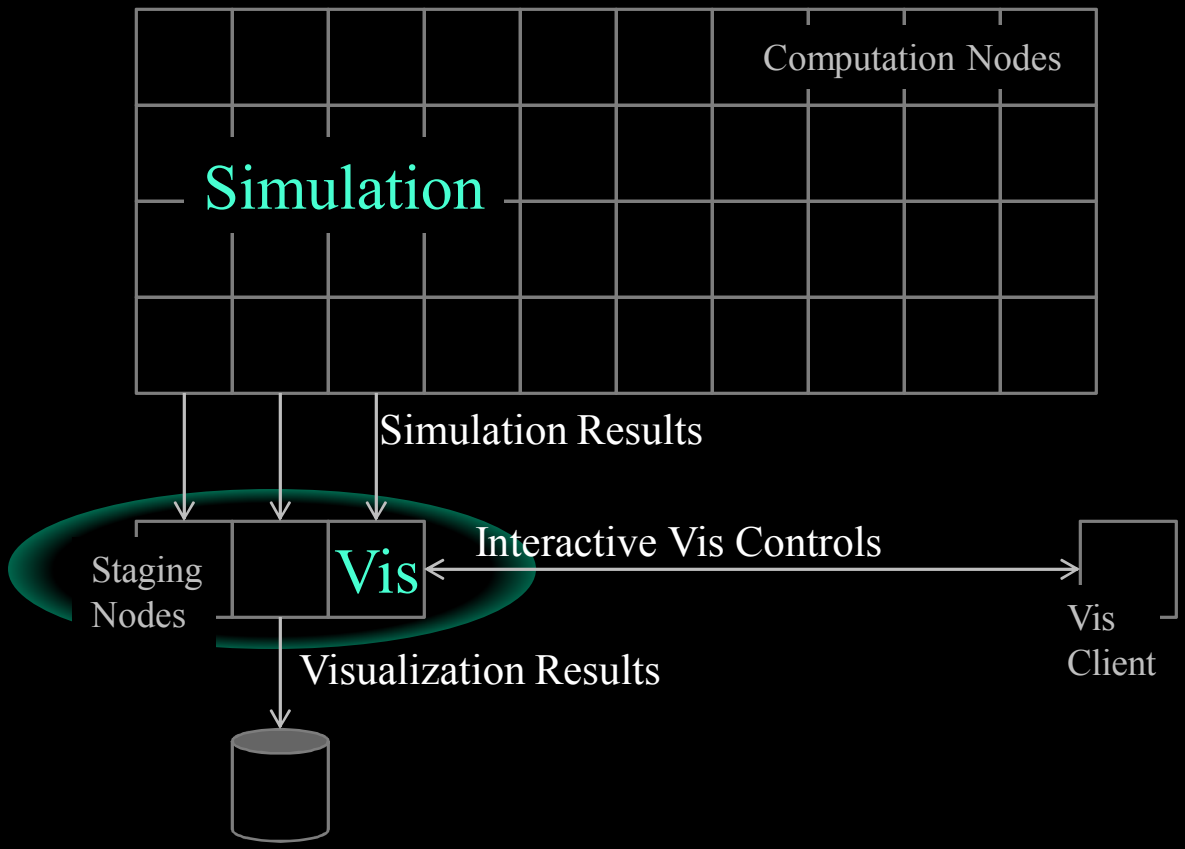
	Capability	Coupling	Footprint	Transfer	Interactive
ParaView In Transit	**High**	**Loose**	~5% Extra Nodes	Hi Speed Transfer	**Yes**



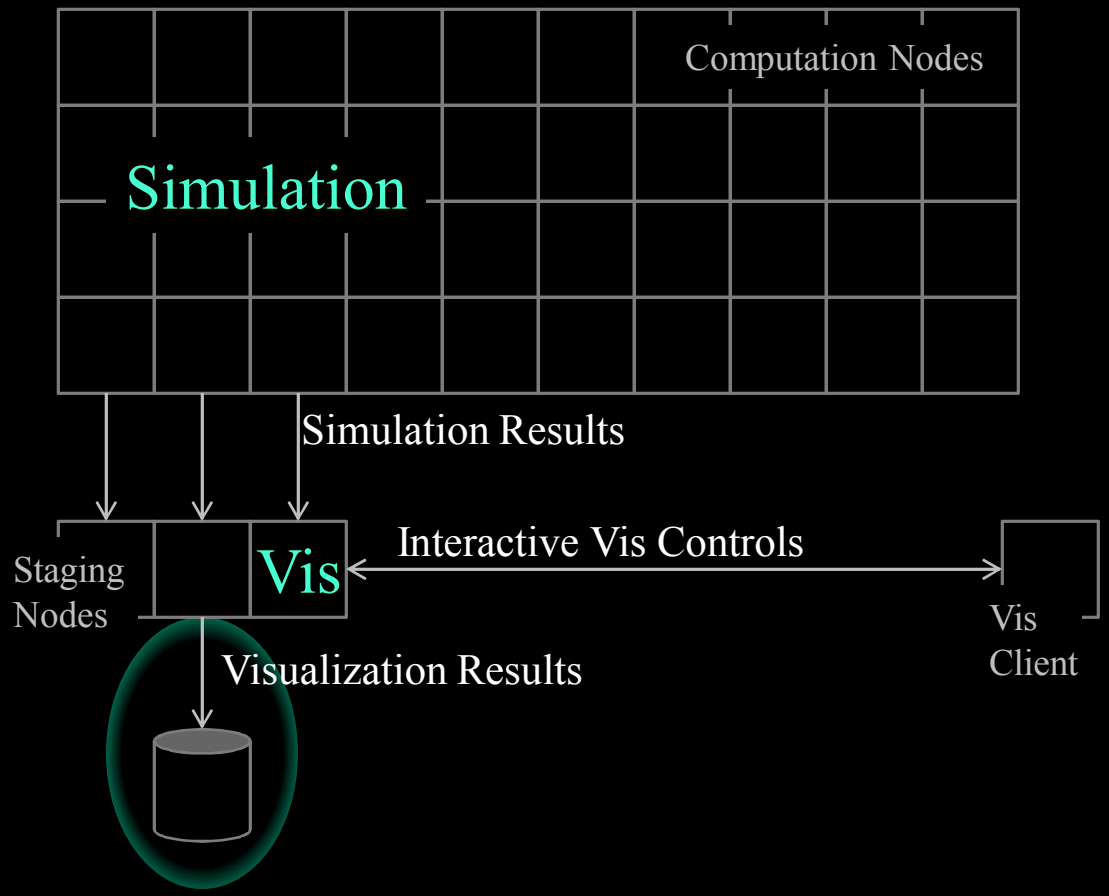
	Capability	Coupling	Footprint	Transfer	Interactive
ParaView In Transit	**High**	**Loose**	~5% Extra Nodes	Hi Speed Transfer	**Yes**



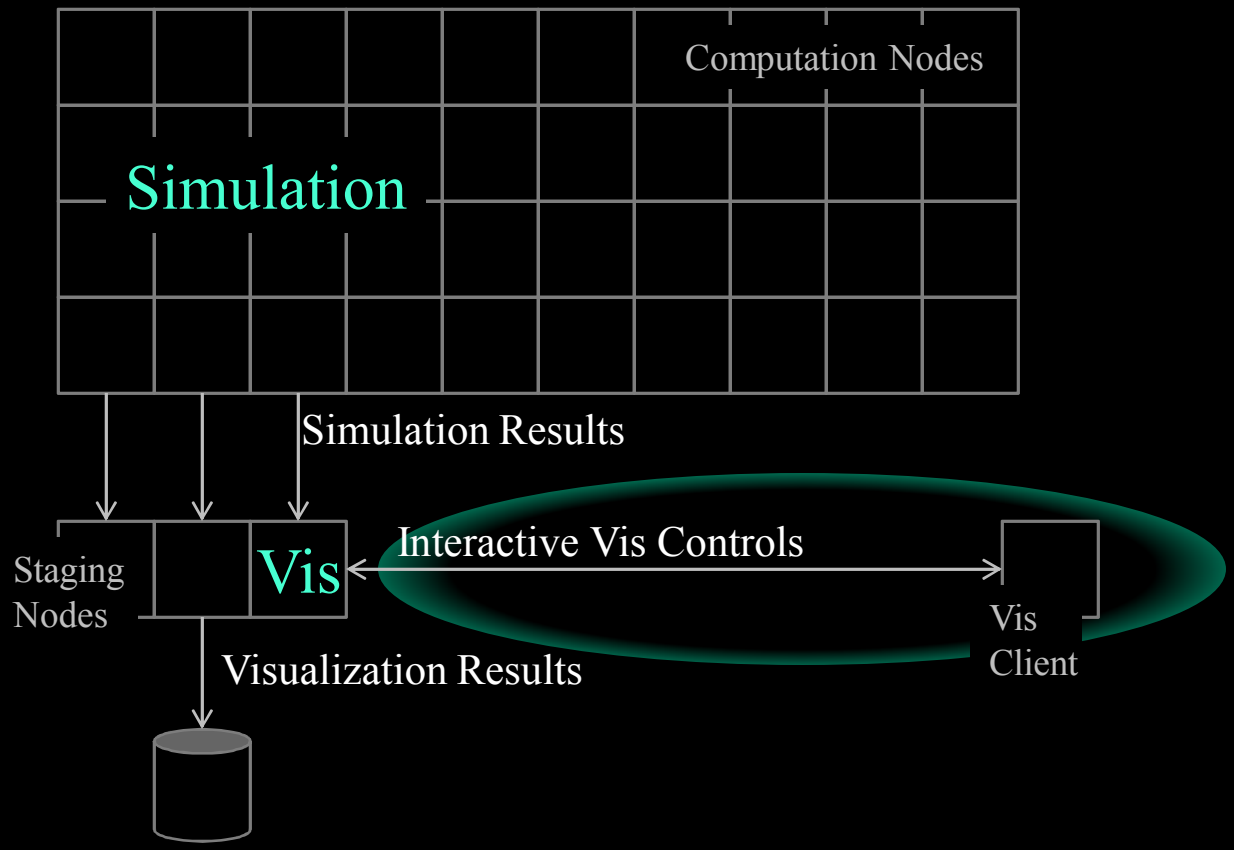
	Capability	Coupling	Footprint	Transfer	Interactive
ParaView In Transit	**High**	**Loose**	~5% Extra Nodes	Hi Speed Transfer	**Yes**

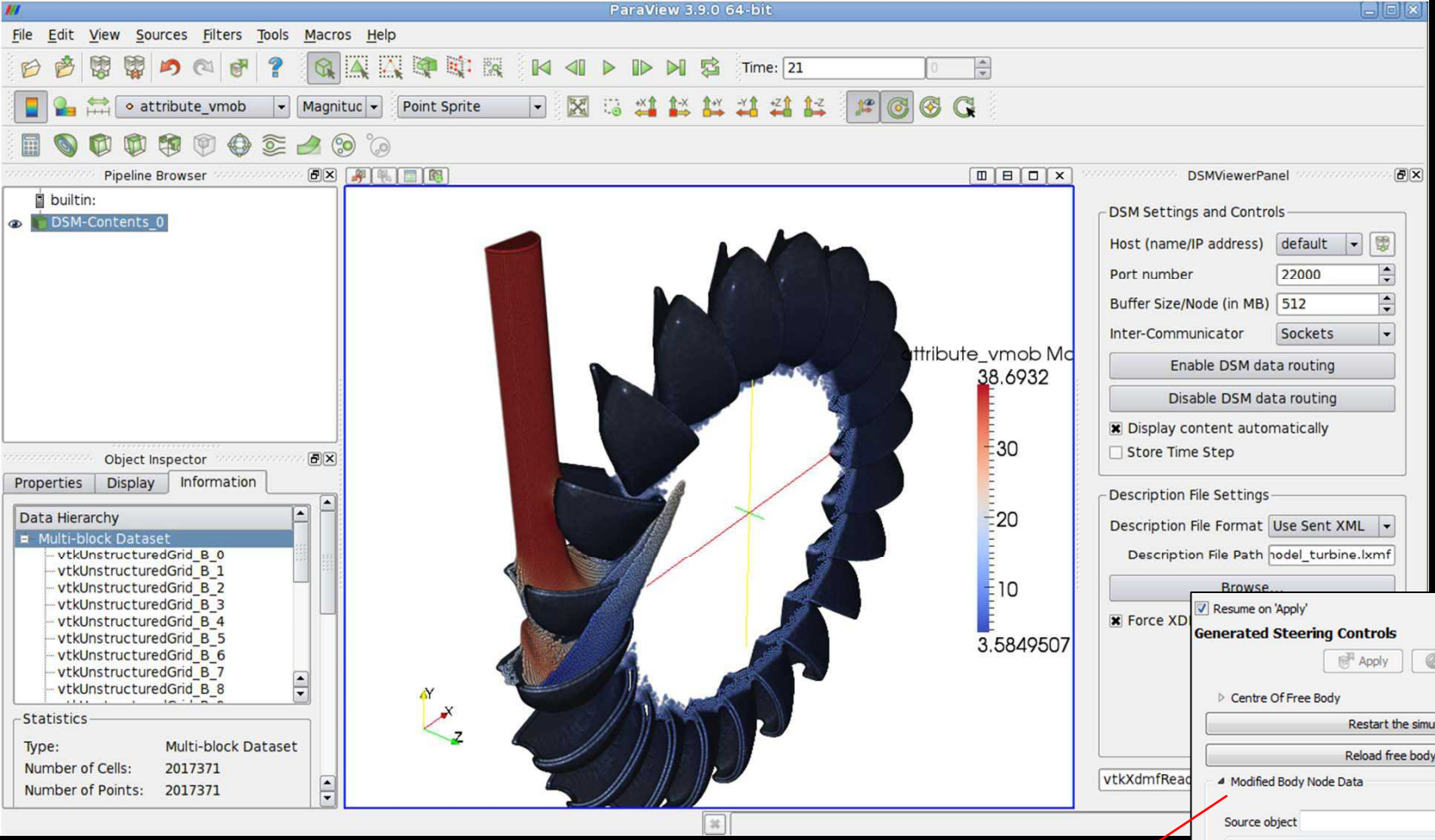


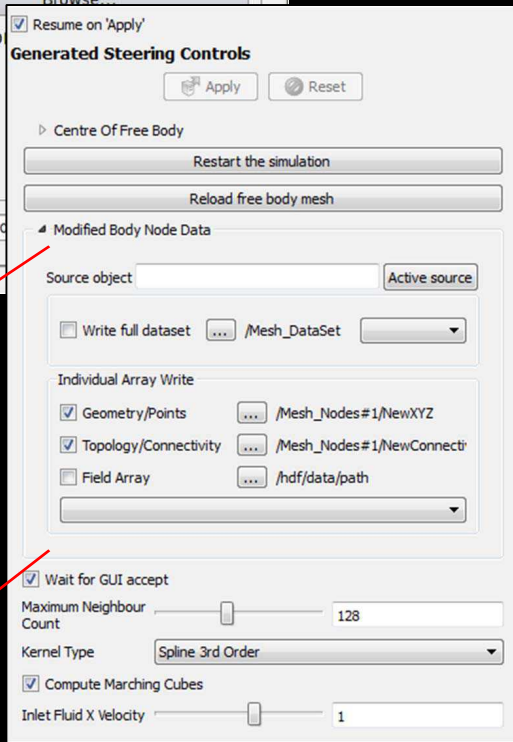
	Capability	Coupling	Footprint	Transfer	Interactive
ParaView In Transit	**High**	**Loose**	~5% Extra Nodes	Hi Speed Transfer	**Yes**



	Capability	Coupling	Footprint	Transfer	Interactive
ParaView In Transit	**High**	**Loose**	~5% Extra Nodes	Hi Speed Transfer	**Yes**








```
<DataExportProperty name="ModifiedBodyNodes"
                    command="SetSteeringArray"
                    label="Modified Body Node Data">
  <DataExportDomain name="data_export"
                    full_path="/Mesh_DataSet"
                    geometry_path="/Mesh_Nodes#1/NewXYZ"
                    topology_path="/Mesh_Nodes#1/NewCo..."
                    command_property="ReloadFreeBodyMesh" />
</DataExportProperty>
```

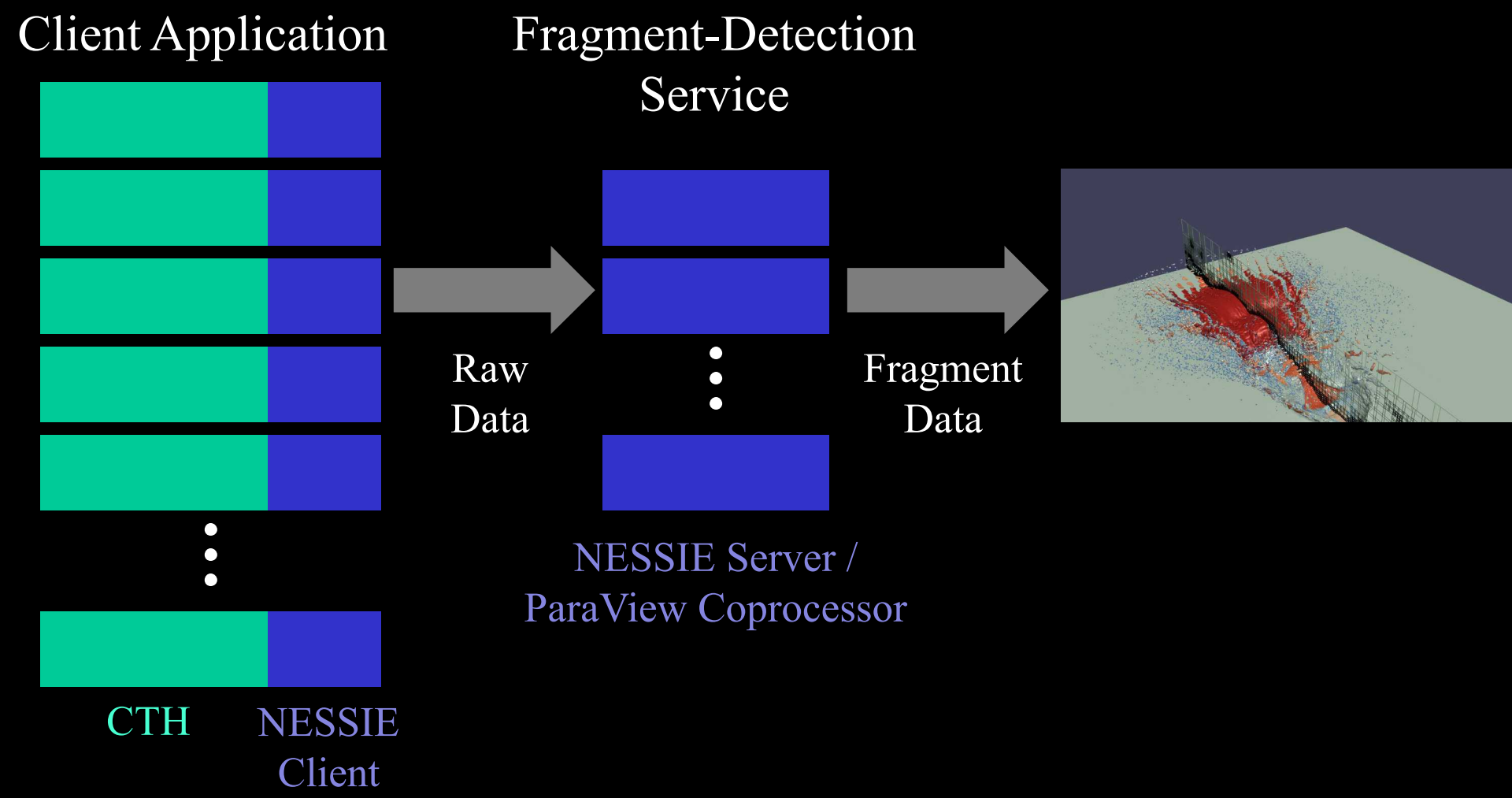
Biddiscombe, et al, "Parallel Computational Steering and Analysis for HPC Applications using a ParaView Interface and the HDF5 DSM Virtual File Driver."

Network Scalable Service Interface (NESSIE)

Sandia National Laboratories

- Provides a remote-procedure call (RPC) abstraction.
- Relies on client/server stubs to marshal/unmarshal procedure call parameters.
  - Also supports data messages for raw data transfers.
- Server uses one-sided API for native remote direct-memory (RDMA).
- Used to implement:
  - File system services
  - HPC database access proxies
  - PnetCDF data staging
- Our work leverages PnetCDF data staging for in transit analysis of CTH.

NESSIE-CTH-ParaView Integration

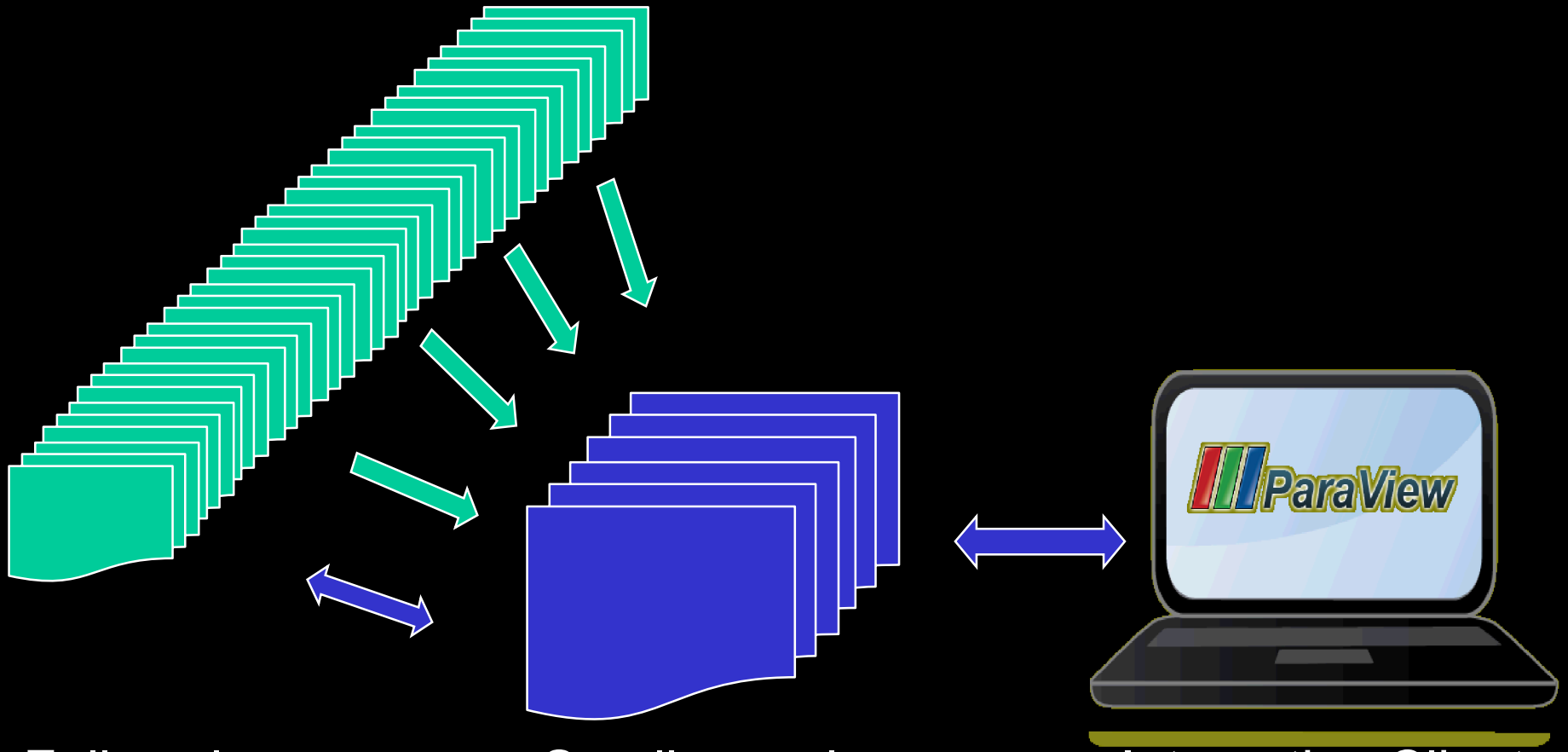


GLEAN

- Flexible/extensible framework for simulation-time data analysis and I/O acceleration.
- Adjusts to analysis and system characteristics.
- Client/server architecture.
  - Data streams from running solver (client) to dedicated staging nodes (server).
- Integrated in ParaView via VTK reader/writer components.
  - ParaView coprocessing (in situ) library (integrated with server) uses VTK GLEAN writer (“client,” compute nodes).
  - ParaView pvserver (interactive vis service) uses VTK GLEAN reader (“server,” staging nodes).
  - Behaves like file read/write from ParaView perspective.

ParaView In Situ + Covis + Interactive Vis with GLEAN

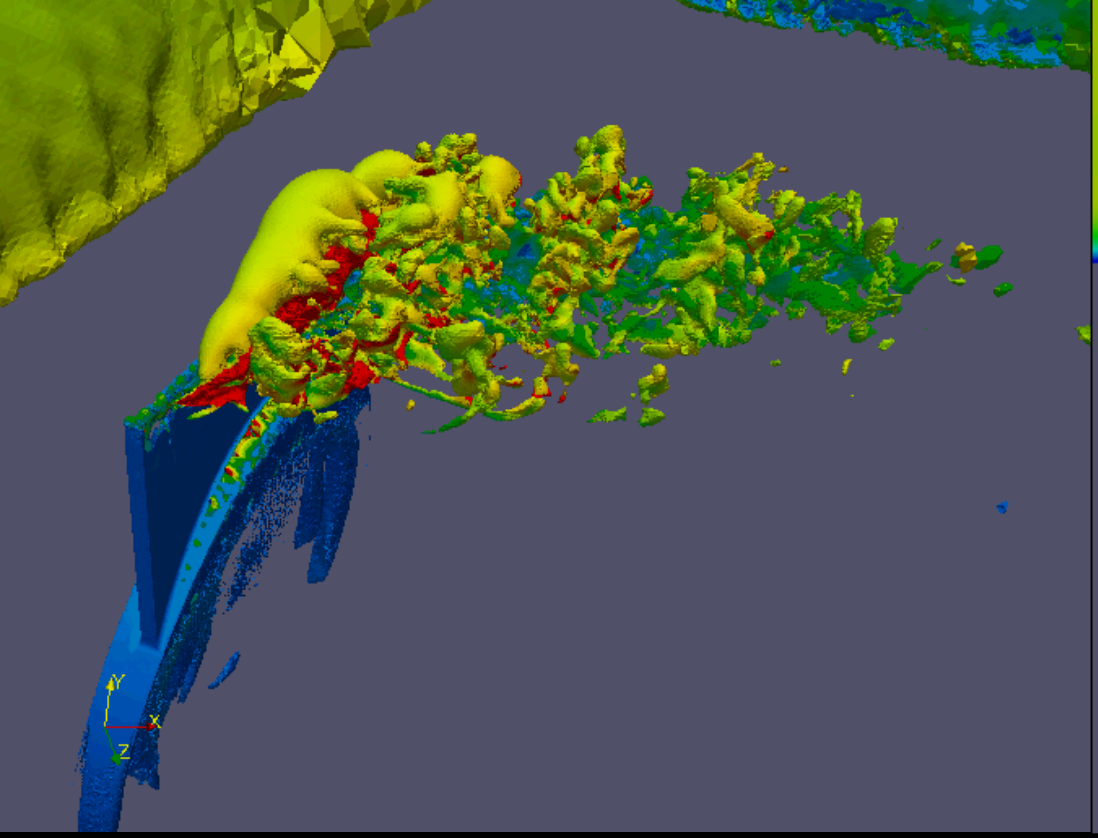
Sandia National Laboratories

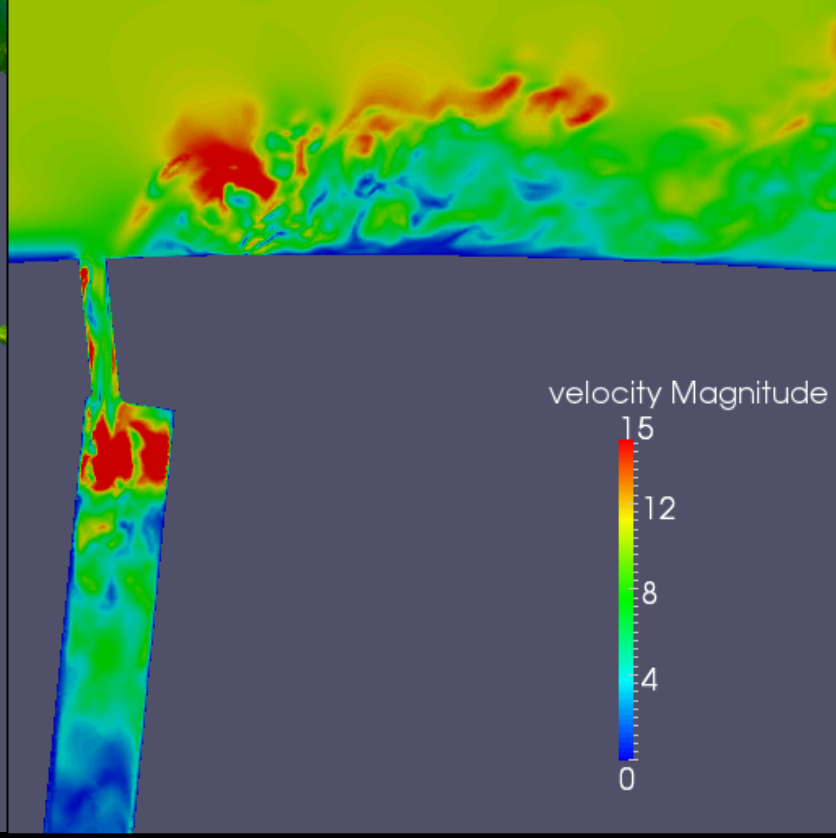


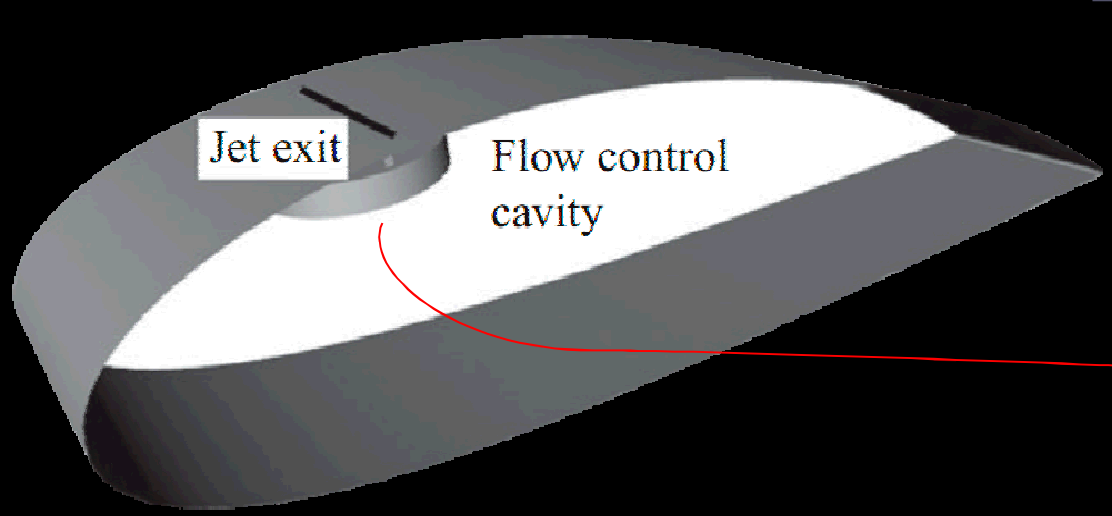
Full scale Simulation w/ In situ

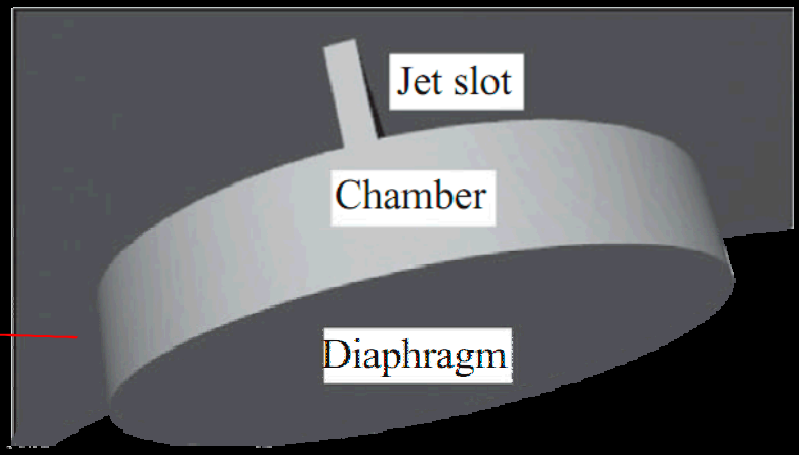
Smaller scale Covisualization

Interactive Client



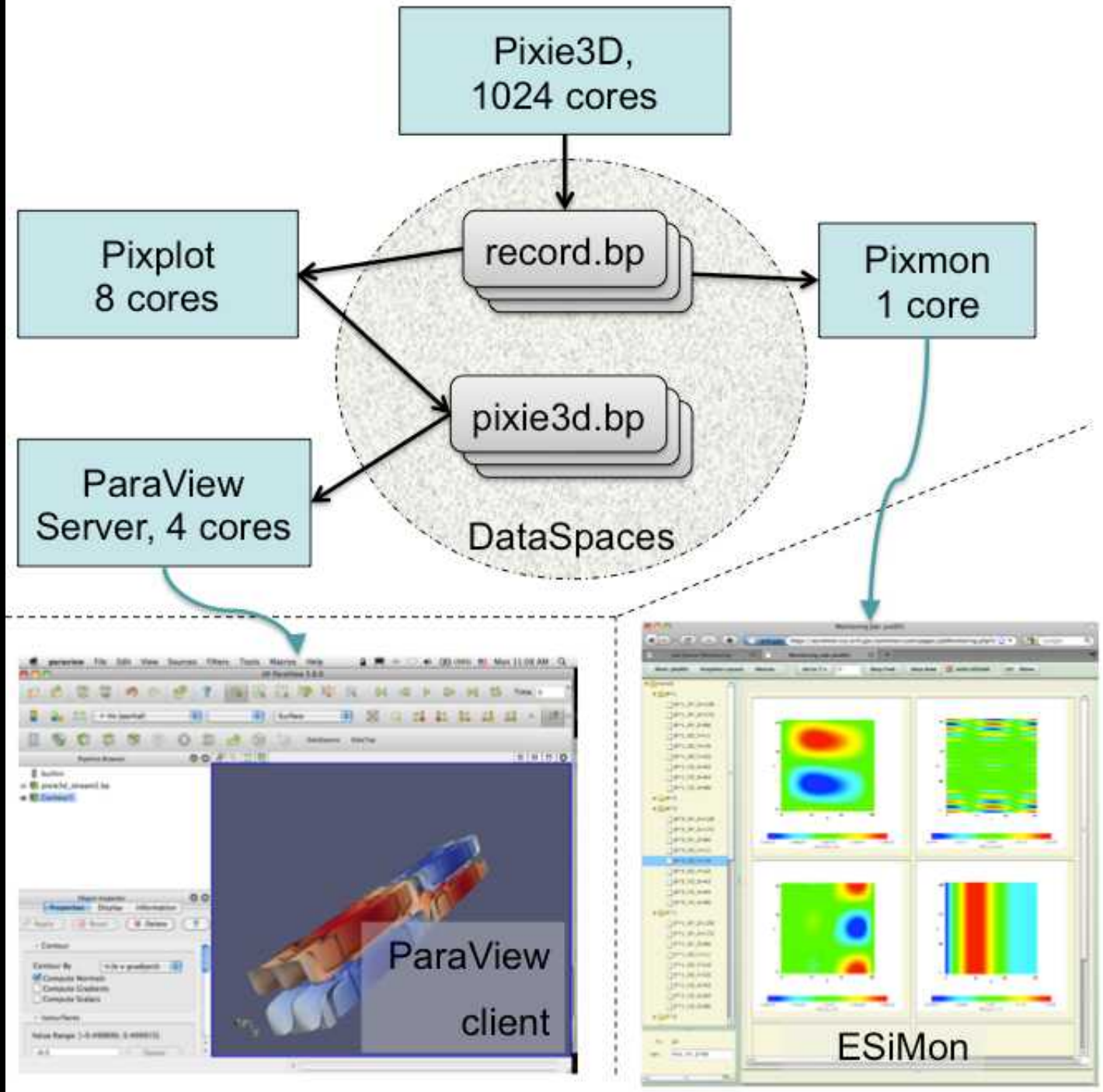

velocity Magnitude
15
12
8
4
0


Jet exit
Flow control
cavity


Jet slot
Chamber
Diaphragm

Adaptable I/O System (ADIOS)

- Separates I/O API from I/O implementation.
  - Abstracts I/O transport, destination, format.
  - Can be changed at runtime through XML file.
- Uses DataSpaces for virtual semantically-specialized shared space.
  - Location/distribution agnostic.
  - Accessed with asynchronous operations.
  - Exists as separate application – provides fault isolation.
    - Coupled codes do not bring each other down.
  - Manages multiple versions of dataset (e.g. timesteps).
    - Autonomous buffering and eviction.
- Using ADIOS for files automatically enables in transit.
  - But, change read semantics for unavailable timesteps.


Pixie3D,
1024 cores
Pixplot
8 cores
record.bp
Pixmon
1 core
pixie3d.bp
ParaView
Server, 4 cores
DataSpaces
ParaView
client
ESiMon

Conclusion

- Potential advantages of in transit visualization.
  - About everything in situ has going for it.
  - Loose coupling between simulation and visualization.
  - Fault isolation.
  - Multiple event loops / asynchronous operation.
  - Federated environment.
- Should stop thinking of I/O in terms of “files.”
  - I/O libraries should (will) progress to more general.
  - Application should stop considering data sources/sinks as static collections.
  - Applications should *not* go low level (POSIX, MPI/IO).

Shameless Promotion of Other Presentations

Sandia National Laboratories

Sun	1:30	TCC 205	ParaView In Situ Tutorial
Tue	10:00	ASC Booth	In Situ Visualization and Analysis
Tue	11-1	ASC Booth	ParaView Demonstrations
Tue	5:15	Poster Reception	Co-visualization of full data and in situ data extracts from unstructured grid CFD at 160k cores
Tue	5:30	TCC LL1	BoF: Next Generation I/O and visualization requirements for in situ processing
Thu	10-12	ASC Booth	ParaView Demonstrations
Thu	10:30	ASC Booth	ASC Tri-Lab Visualization on the Cielo Platform

Questions?