SSDBM PANEL
ON-LINE ANALYTICS

CHAIR: SCOTT KLASKY (ORNL)

PANELISTS: WILL SCHROEDER (KITWARE), DAVE PEARAH (HDF GROUP), MATTHEW WOLF (ORNL), TOM PETERKA (ANL) AND SHINJAE YOO (BNL)

THE PANELIST

DAVID PEARAH

- David is CEO and Chairman of Board of Directors of The HDF Group.

MATTHEW WOLF

- Matthew is the deputy group lead, and senior scientist in the Scientific Data Group at ORNL. His research targets high performance, scalable applications, particularly focused on I/O and adaptive event middlewares.

SHINJAE YOO

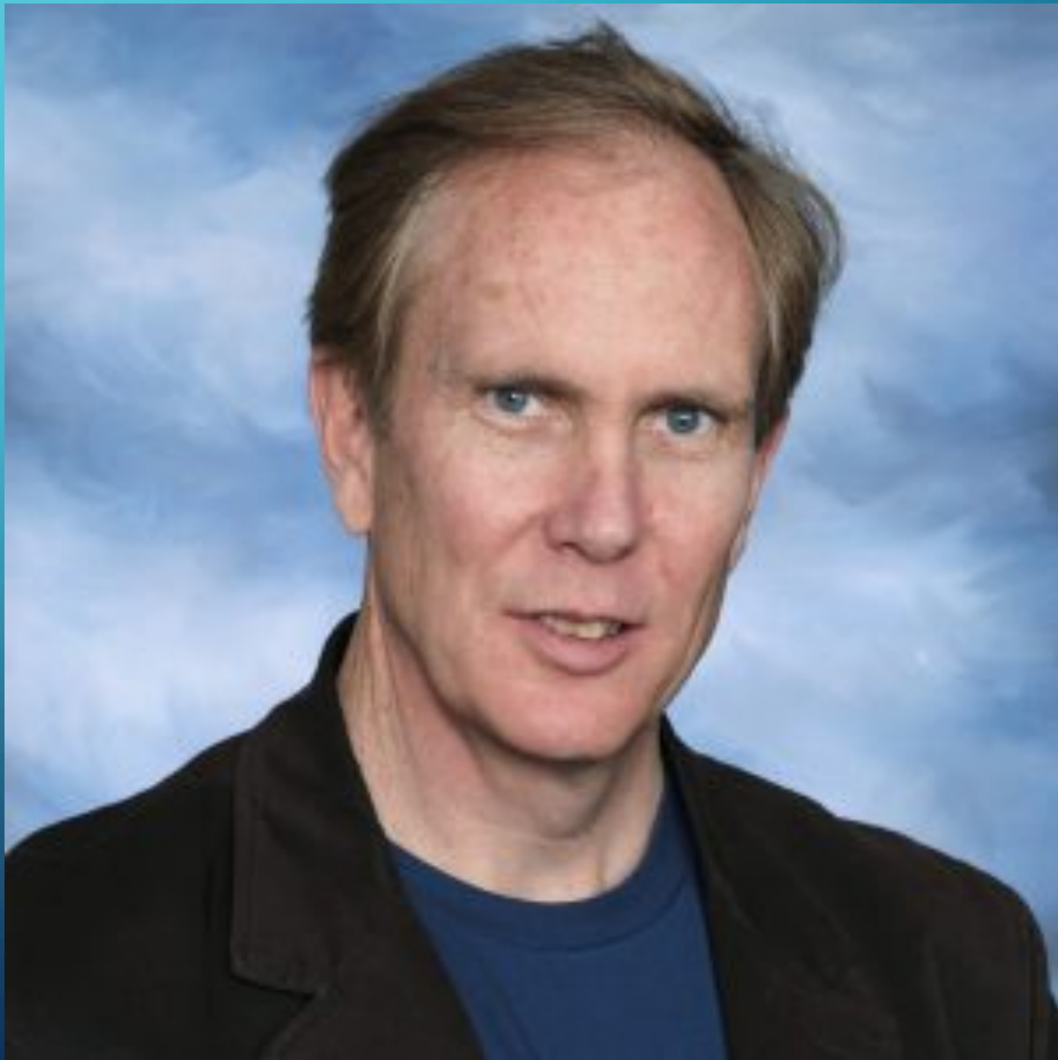
- Shinjae is computational scientist @ BNL and his research interest includes large scale scientific data mining and learning.

TOM PETERKA

- Tom is a computer scientist at Argonne National Laboratory, fellow at the Computation Institute of the University of Chicago, adjunct assistant professor at the University of Illinois at Chicago, and fellow at the Northwestern Argonne Institute for Science and Engineering. His research interests are in large-scale parallelism for in situ analysis of scientific data.

WILL SCHROEDER

- Will Schroeder is a founder and former CEO/President of Kitware. In his current role as Opportunity Catalyst, he works on advanced computing algorithms and software platforms such as VTK; also engaging in collaborator outreach, mentoring, and technology visioning.



INSTRUCTIONS: DISCUSS ON-LINE ANALYTICS IN TERMS OF

- Which services are most important?
- Usability
- Reliability
- Decision support
- Infrastructure
- Incorporation of Visualization service
- Services for analysis

SSDBM: On-line Analytics Panel

June 27, 2017

Dave Pearah, CEO

The HDF Group

Who is the HDF Group?

HDF Group has developed open source solutions for Big Data challenges for nearly 30 years

Small company (~ 40 employees) with focus on High Performance Computing and Scientific Data

Offices in Champaign, IL + Boulder, CO

Our flagship platform – HDF5 – is at the heart of our open source ecosystem.

Tens of thousands use HDF5 every day, as well as build their own solutions (~~600~~ ~~700~~ 800+ projects on Github)

What does the HDF Group do?

Products

- HDF5 Community (Open Source) + Enterprise (Coming Soon)
- Connectors: ODBC + Cloud (Beta)
- Add-Ons: e.g. compression + encryption

Support

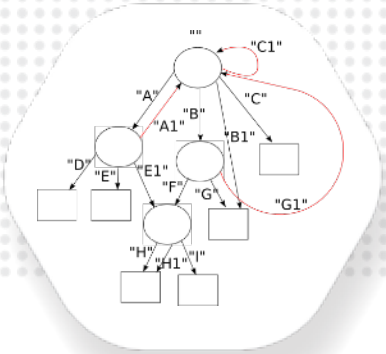
- HDF Support Packages (Basic + Pro + Premier)
- Support for h5py + PyTables + pandas (NEW)
- Training

Consulting

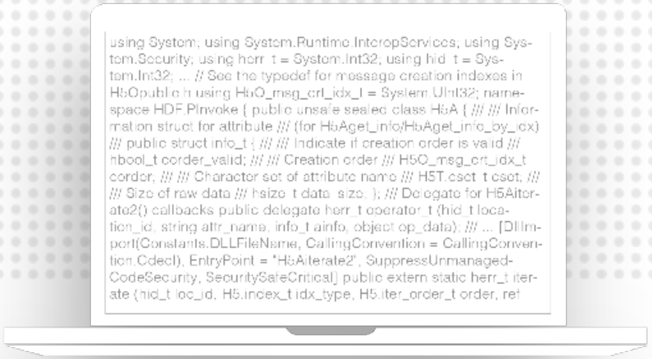
- HDF: new functionality + performance tuning for specific platforms
- General HPC software engineering with scientific expertise
- Metadata science and expert services

The HDF5 Platform

Marriage of data model + I/O software + binary container

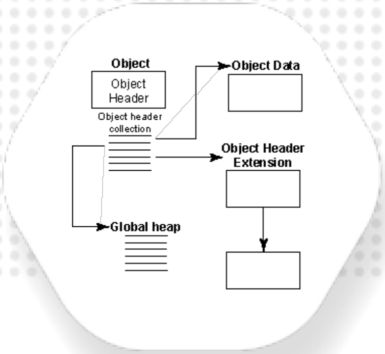


HDF5 abstract data model



```
using System; using System.Runtime.InteropServices; using Sys-
tem.Security; using herr_t = System.Int32; using hid_t = Sys-
tem.Int32; ... // See the typedef for message creation indexes in
H5Opublic.h using H5O_msg_crt_idx_t = System.UInt32; name-
space HDF.PInvoke { public unsafe sealed class H5A { /// /// Infor-
mation struct for attribute /// (for H5Aget_info/H5Aget_info_by_idx)
/// public struct info_t { /// /// Indicate if creation order is valid ///
hbool_t corder_valid; /// /// Creation order /// H5O_msg_crt_idx_t
corder; /// /// Character set of attribute name /// H5T.cset_t cset; ///
/// Size of raw data /// hsize_t data_size; }; /// Delegate for H5Aiter-
ate2() callbacks public delegate herr_t operator_t (hid_t loca-
tion_id, string attr_name, info_t ainfo, object op_data); /// ... [DllIm-
port(Constants.DLLFileName, CallingConvention = CallingConven-
tion.Cdecl), EntryPoint = "H5Aiterate2", SuppressUnmanaged-
CodeSecurity, SecuritySafeCritical] public extern static herr_t iter-
ate (hid_t loc_id, H5.index_t idx_type, H5.iter_order_t order, ref
```

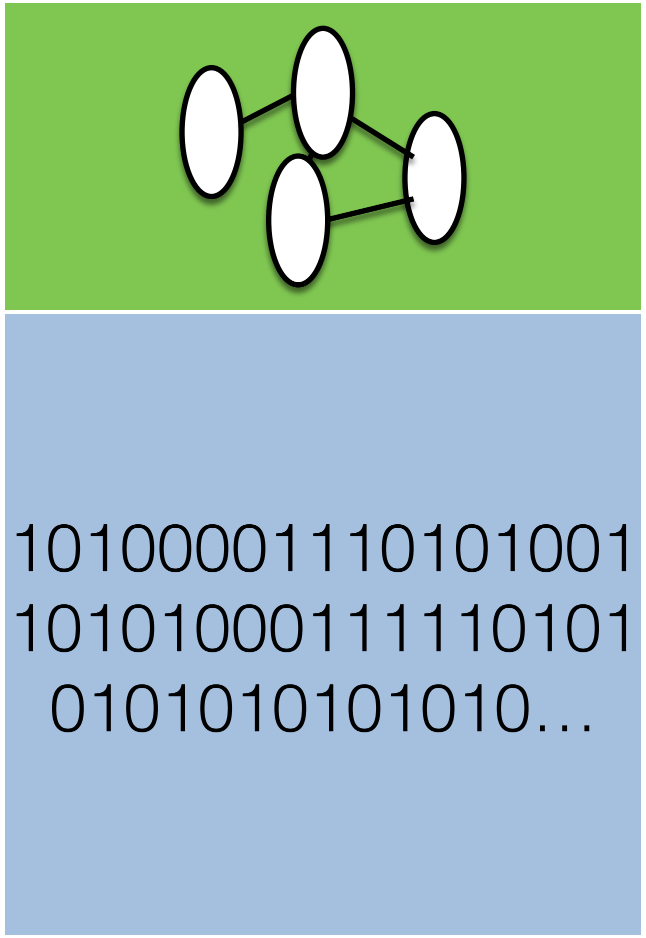
HDF5 library



HDF5 file format

HDF5 “File”

NOT an accurate depiction
of the physical
layout!



10100001110101001
10101000111110101
0101010101010…

Metadata (“data about data”, documentation)

- Structure + organization
- Types and encodings
- Names
- Array shapes
- Conventions
- Annotations

Data (payload, variable, measurement, “raw”)

- Problem-sized (small to large)
- Binary
- Optionally
  - compressed
  - encrypted
  - checksum’d
  - “filtered”

What isn’t HDF5?

Shrink-Wrapped Service

HDF5 is an SDK for developers to embed into their own solutions

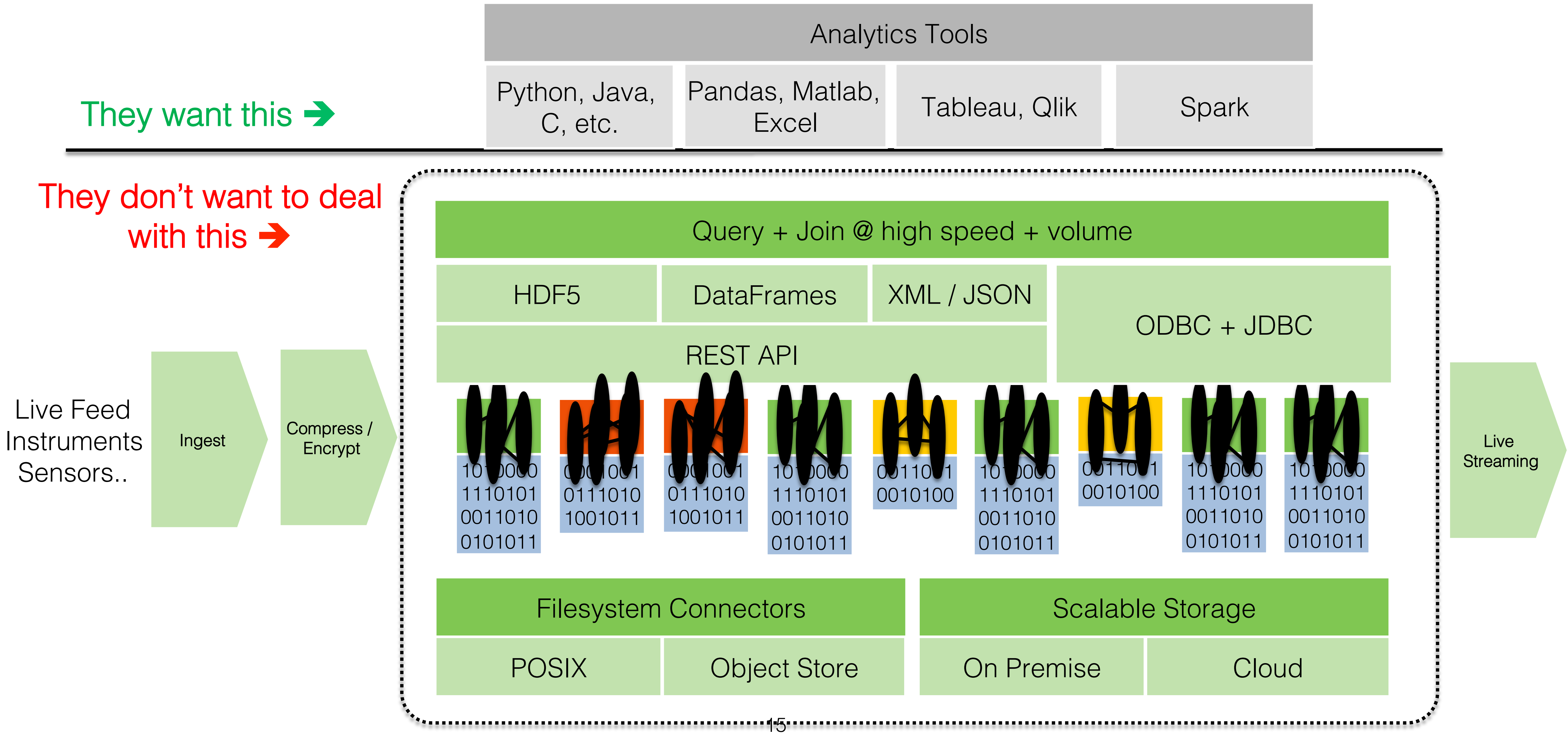
Database

Example: no scalability beyond single file (particularly without parallel file system)

Analytics Tool

Example: run STAC M3 benchmark for backtesting financial market data

Scientific users want analytics, not low-level tools



They want this ➔

They don't want to deal with this ➔

Analytics Tools

Python, Java, C, etc.

Pandas, Matlab, Excel

Tableau, Qlik

Spark

Query + Join @ high speed + volume

HDF5

DataFrames

XML / JSON

ODBC + JDBC

REST API

Filesystem Connectors

Scalable Storage

POSIX

Object Store

On Premise

Cloud

Live Feed Instruments Sensors..

Ingest

Compress / Encrypt

Live Streaming

Typical Feedback from Scientific Users

- **Piles of Files**
  - “OK, I have a stupid-large repository of these HDF files… now what?”
- **Use Case #1: finding remote data then grabbing it for local analytics**
  - “Right now, I can barely keep up with indexing the metadata”
- **Use Case #2: forget two steps ➔ just run the analytics!**
  - “I want to explore data in real-time, and ideally I would like to avoid writing any code or as little as possible”

Scientific Computing is Different

- **Communities optimized for different things**
  - Scientific community: value **performance** (run-time) over productivity (code-time)
  - Commercial community: value **productivity** over performance
- **In commercial world, you want to use an existing framework or tool because you would never build yourself**
  - High-level tools: pandas, Matlab, Spark
  - BI: Tableau, Qlik, etc.
  - Excel

Open Questions

- What is need for higher-level tools? Are scientists willing to adopt?
- If there is need, should these be unique to HPC or port tools (e.g. Spark) from conventional computing community
- Is scientific community willing to invest in creating not only prototypes but sustainable software (and companies to support this software)

Online Analytics: Opening a New Ecosystem

Matthew Wolf

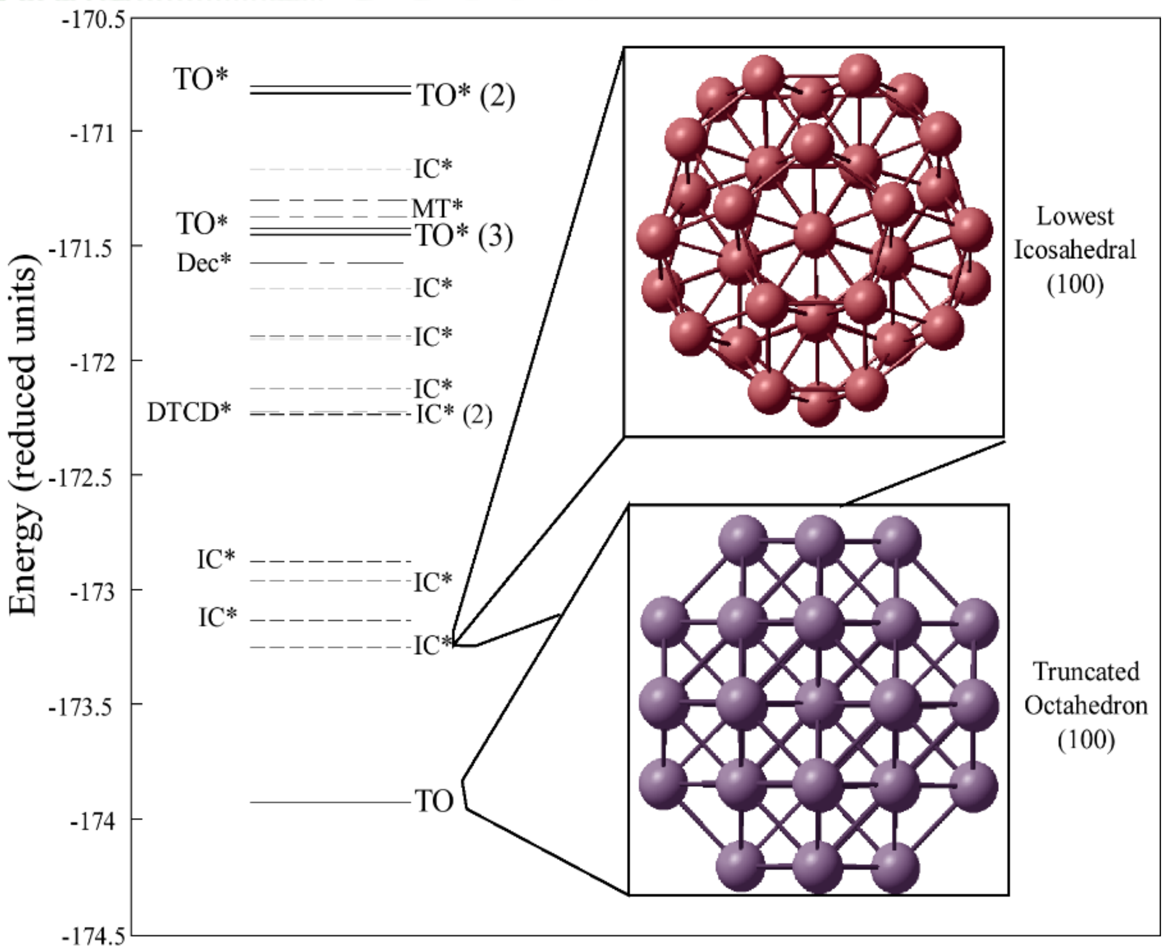
Scientific Data Group

Oak Ridge National Laboratory

Online vs In Situ

- In Situ has gained quite a bit of attention recently.
  - "In Place" is a pretty limiting statement, though.
- Users don't think about where there data is in Hadoop or Spark… why should high performance analytics be different?
  - High performance scientific data needs its own platform.
  - Forcing the user API to manage placement, concurrency, memory access, etc. so that data can stay "in place" is very limiting
- Online Analytics often is phrased as a performance improvement over post-hoc, but that's missing the point
  - It's also an opportunity for a significant change in functionality.

Personal Example: Understanding Nanocrystals



- The application I wanted to write wasn’t an all-in-one simulation from one initial condition moving forward in time.

- Physics aside, this is a challenging problem for data management
  - Each item you generate has complex (geometric) features that you want to categorize, as well as metadata properties you want to evaluate (energy, etc.)
  - The algorithm I wanted needed a function where I could do this evaluation **online, at runtime**, so that you could reject/modify a sample if it was redundant with what you already had.

- There are a whole set of “what-if” or “only-if” scientific data analytics problems like this that are prohibitive in current HPC programming models.

Online Runtimes

- Great idea… but it only works if you can actually move the data around online in an intelligent way.
  - Remember: may be large, parallel data sets (TBs per record, on 100k+ nodes)
  - Leave it in situ if that's most efficient right now, or move it if it's not.
  - Have to know where to move it if you want performance – this is where tools like Spark tend to fall down for scientific data.

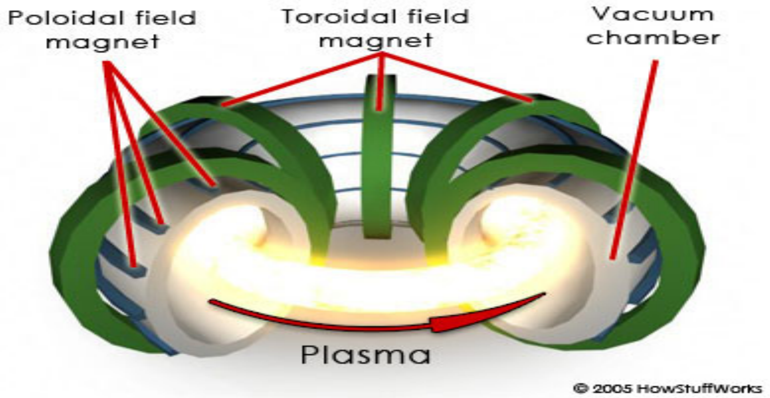
- So how do you make these functionality changes?
  - Scientists choose to investigate problems that they have the tools for, and if there wasn't anything there till your widget came along, you won't have any examples to work with.
  - Solution: Leverage something the scientist already needs to do in a new way.

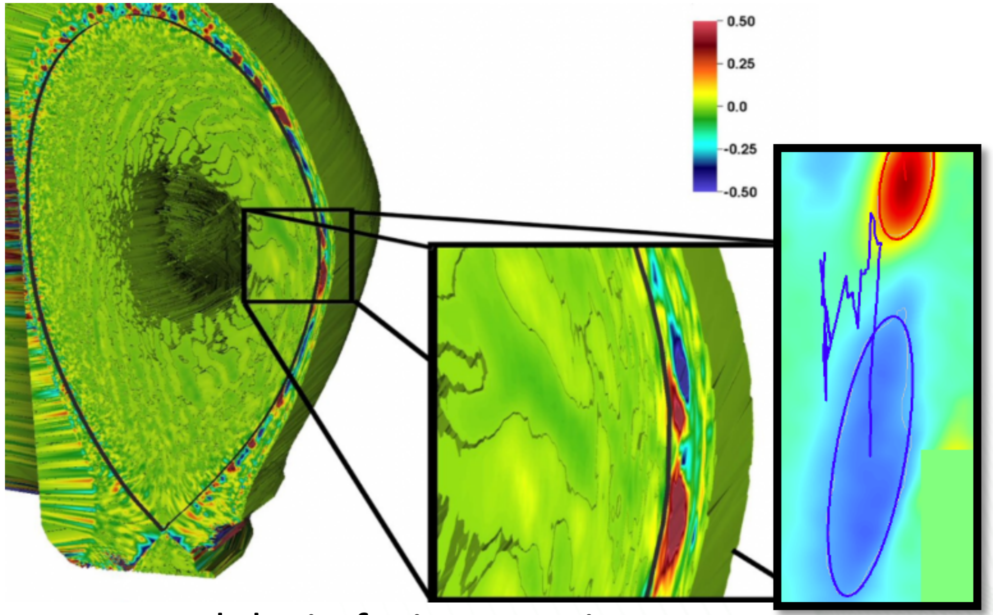
Adaptable I/O System (as a non-storage service provider)

- This is something we're demonstrating now in the ADIOS ecosystem.
  - Memory-to-memory coupling with DataSpaces, FlexPath, and others.
  - Use the exact same executable that you'd use for post-hoc processing; just change the service provider.
  - These services can be complex and include compression, transformation, indexing, etc.
  - The executables can be C, Fortran, Python, Matlab, R, etc.
- So "what-if" scenarios that were played out over months of processing by post-docs can be moved into online analytics.

Feature Extraction: Near Real Time Detection of Blobs

- Fusion Plasma blobs
  - Lead to the loss of energy from tokamak plasmas
  - Could damage multi-billion tokamak
- In experimental data sources, these blobs are important to find in real time.
  - In simulation, finding them in real time allows you to start asking “what-if”, feature-driven questions.
- ADIOS: Distributed online processing
  - Make more processing power available
  - Allow more scientists to participate in the data analysis operations and monitor the experiment remotely
  - Enable scientists to share knowledge and processes



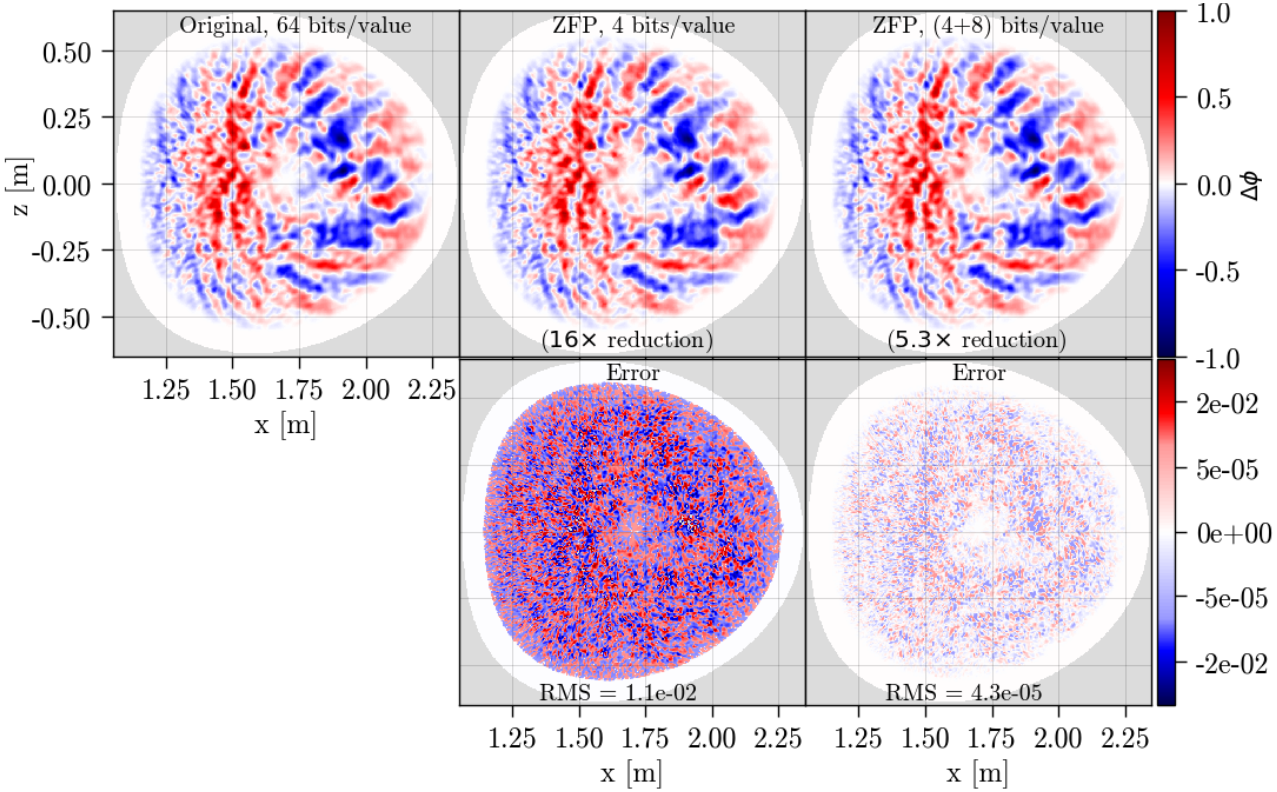


Blobs in fusion reaction
(Source: EPSI project)

Blob trajectory

Composition of Services: Using Lossy Compression

- Sometimes the features you want don't need the full precision of data as generated.
- Composing transformation services in ADIOS gives the user an opportunity to compress and segment data to meet their needs
  - The bitwise segmentation in ZFP generates a multi-resolution indexing data structure that can be exploited for analysis.
- Performance optimization (compression) opens up new analytics capabilities



Progressive refinement with plasma physics data. The top panels show how the electric potential in a simulated tokamak deviates from background. The data has been saved at different precision-levels in each column, compressing with ZFP. The bottom row shows the errors incurred for that precision-level. Low precision can be saved to faster storage (SSD) with additional accuracy added from larger, slower storage (parallel file system) if enough bandwidth is available.

S. Klasky, E. Suchyta, M. Ainsworth, Q. Liu, et al., "Exacution: Enhancing Scientific Data management for Exascale," in ICDCS'17, Atlanta, GA, 2017.

ASCR PI: Scott Klasky (ORNL)

Analysis on the Wire (AoW)

Shinjae Yoo

70 YEARS OF DISCOVERY

A CENTURY OF SERVICE

U.S. DEPARTMENT OF ENERGY

BROOKHAVEN NATIONAL LABORATORY

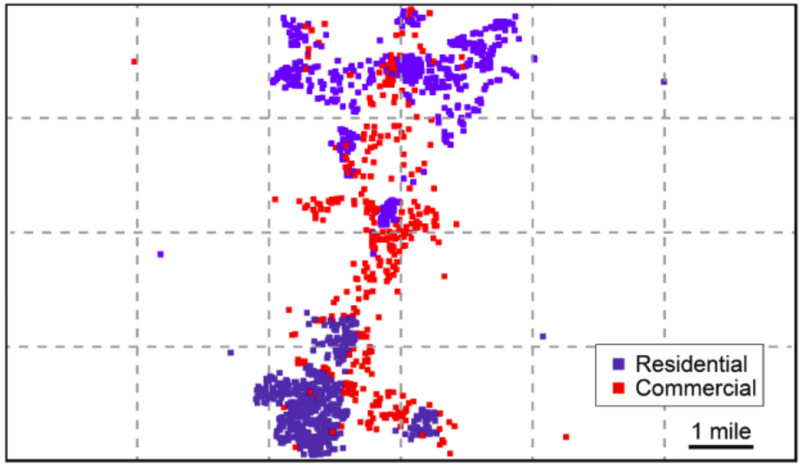
UNCLASSIFIED

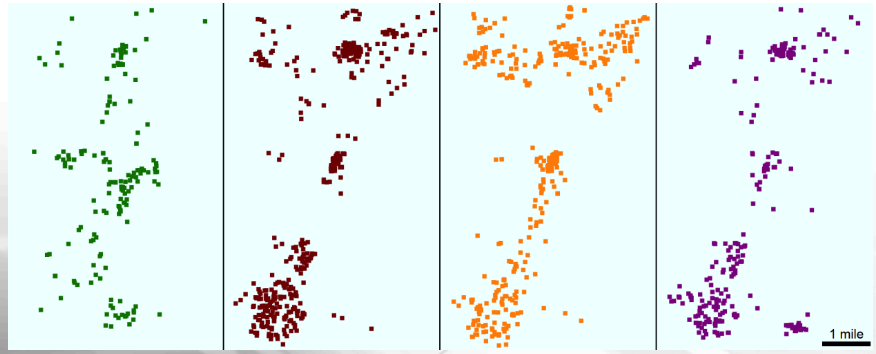
Distributed Sensor Networks (DSNs)

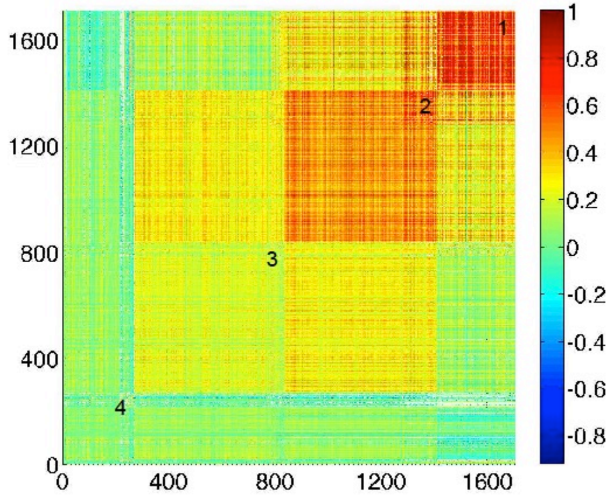
- Services: Spatio-temporal data analytics or IoT
- Questions
  - Do we have to move all such data into data center to analyze?
  - Can we reduce the data volume in advance?
  - Can we analyze the data while in transit?
  - How can we robustly provide analysis services on faulty network?
  - What kinds of infrastructure can be applied to DSNs data analytics?
  - How can we effectively communicate with analyst on such stream of data?
  - What are potential services for DSNs?

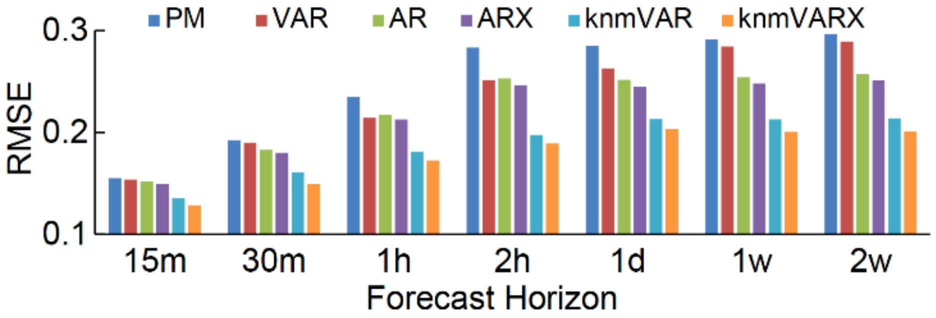
Spatio-Temporal Load Forecasting

- Smart Meter load data used to predict future grid load
- Northeast US utility company
- 1,708 load profiles from residential and commercial customers
- 15 minute intervals, covering a span of 4 months

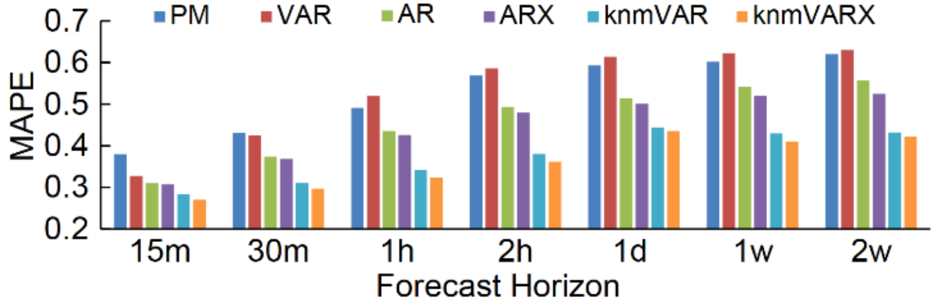








(a) RMSE



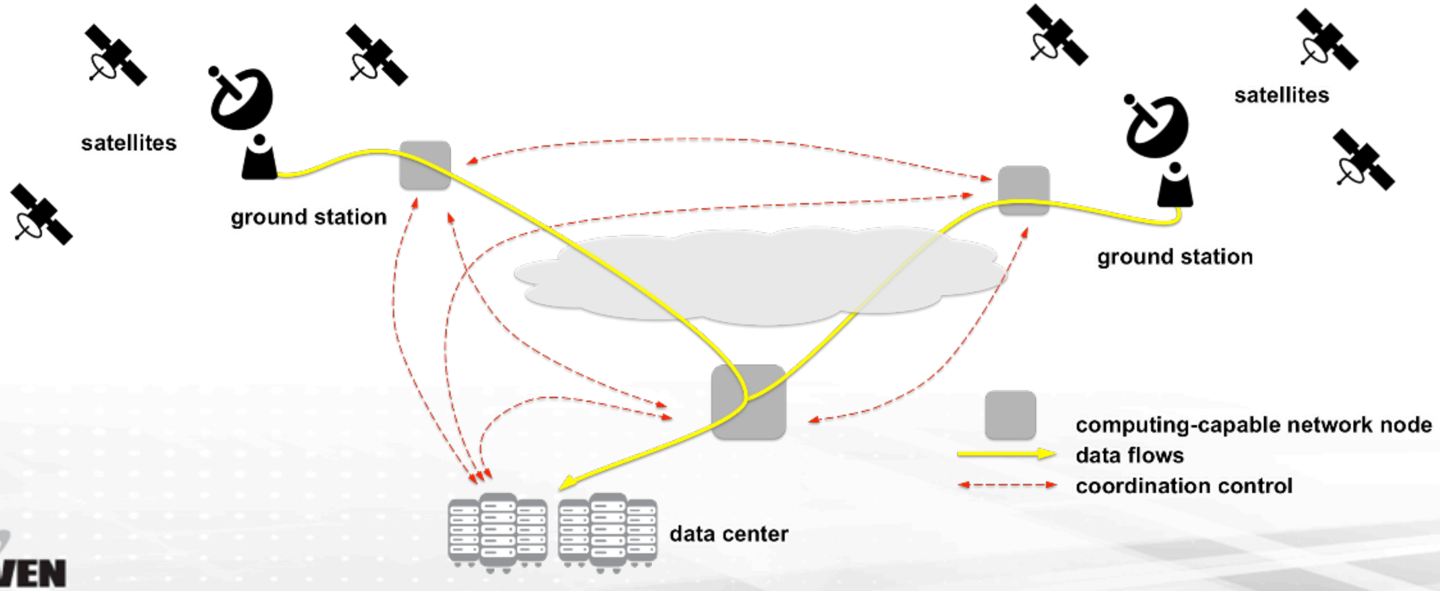
(b) MAPE

UNCLASSIFIED

Streaming Data Analysis on the Wire

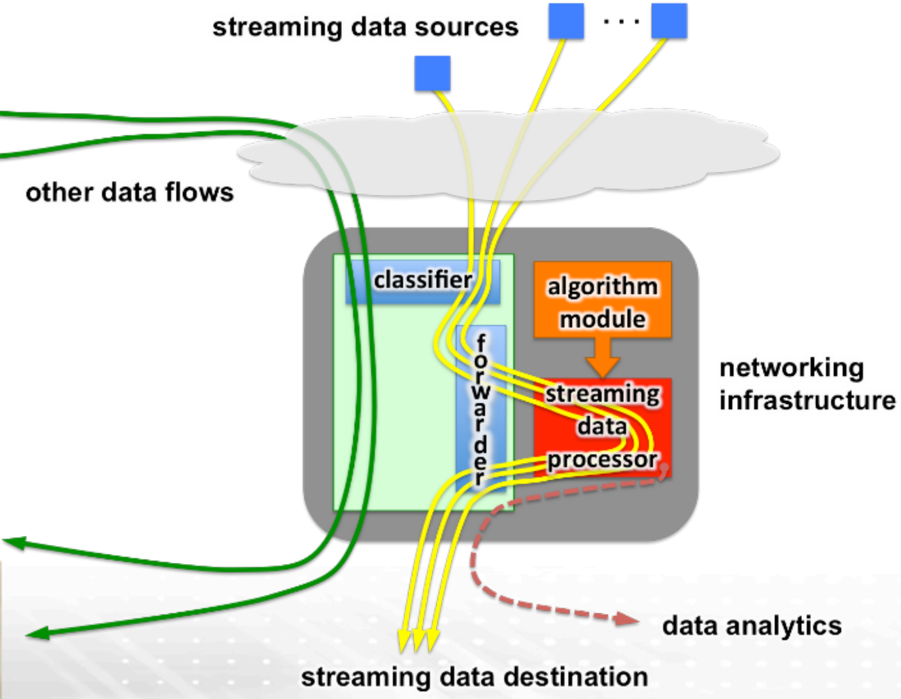
"Analysis on the Wire", a framework that can selectively and transparently perform generic computations on data while in transit in the network fabric. Multiple potential benefits including:

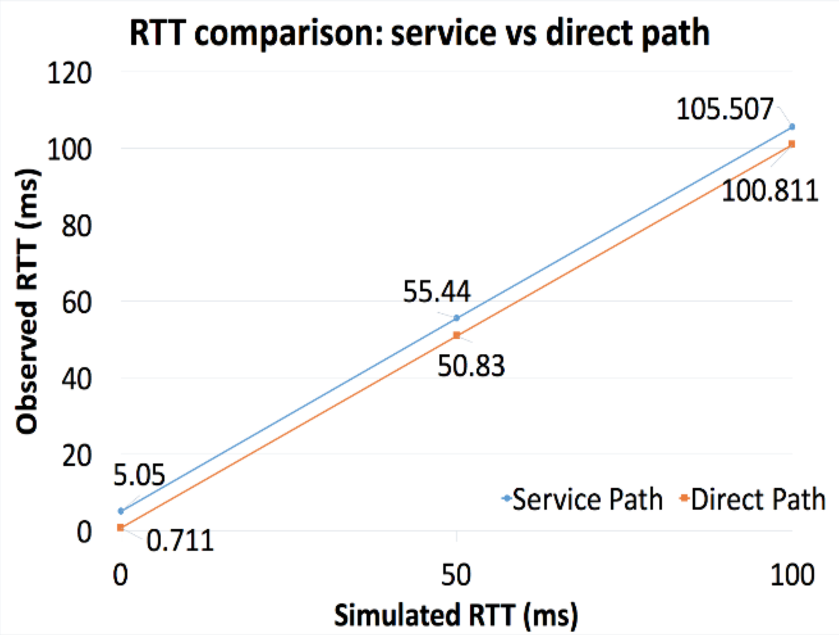
- Process streaming data (e.g., imagery, sensors) for early decision-making and reduced downstream bandwidth requirements
- Extract data analytics, perform generic computations, use distributed computing capabilities
- Examples: Forecasting, deep learning, pattern recognition (e.g., cyber security, automation)
- Data preprocessing: tag, triage, filter, bin all while data is in-transit
- Cost effective solution (non-specialized commodity hardware, scalable to increased bandwidth)

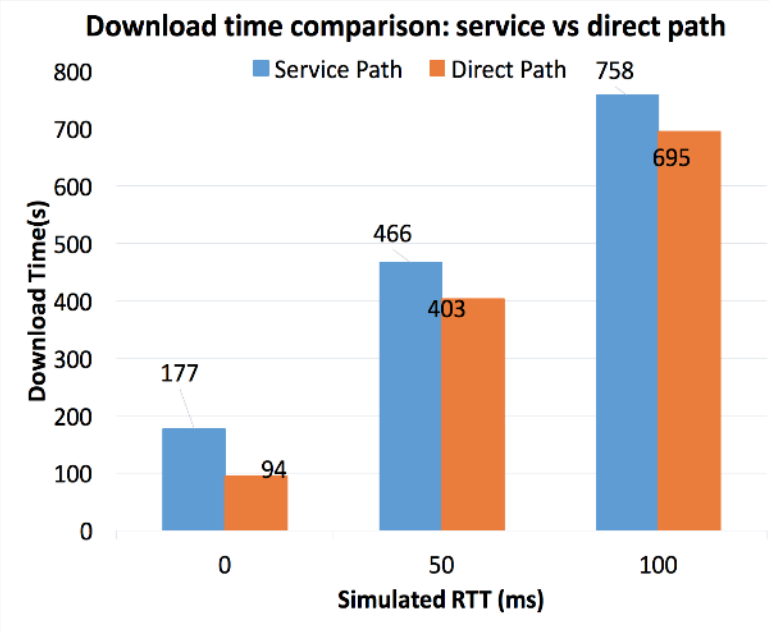


Exploring the SDN

- Tested mechanism with pings, simple file transfers
- Introduced RTT delay to simulate real traffic
- Worst case (2-way) overhead measured ~5ms



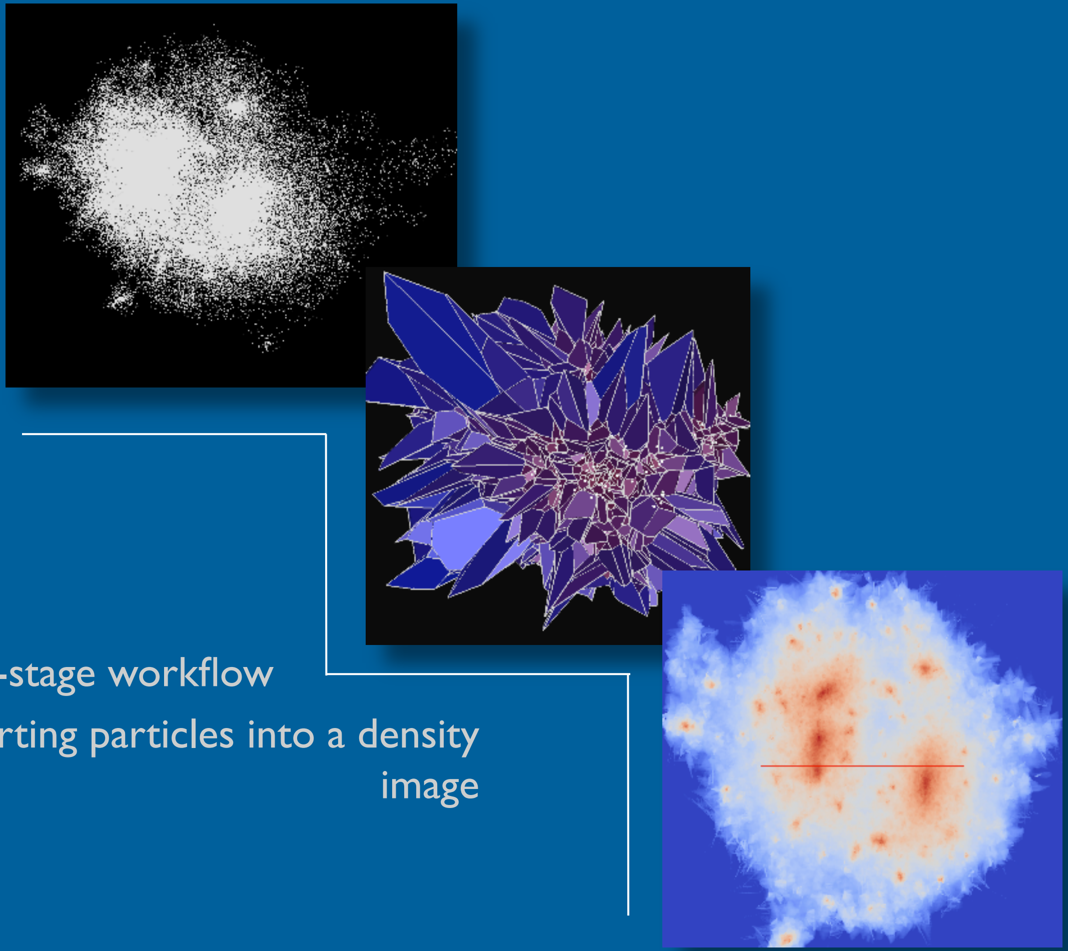




Online Analytics:
Think Globally, Act Locally

“Data movement, rather than computational processing, will be the constrained resource at exascale.”

– Dongarra et al. 2011



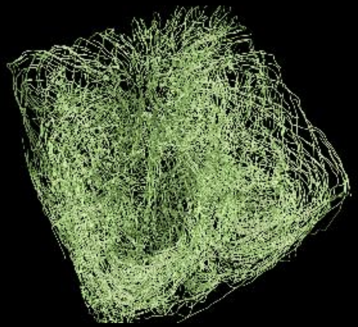
Three-stage workflow converting particles into a density image

SSDM Panel on Online Analytics
June 27, 2017

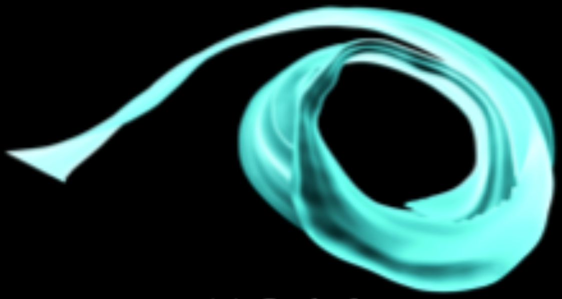
Tom Peterka
tpeterka@mcs.anl.gov
Mathematics and Computer Science Division

Services

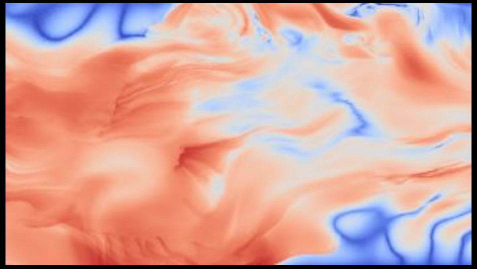
- Definition of a service
  - A data transformation
- Which services should be onlined?
  - Those that can
    - Be streamed
    - Be automated
  - Those that make sense
    - Reduce data movement
    - Provide immediate information
    - Enable later information



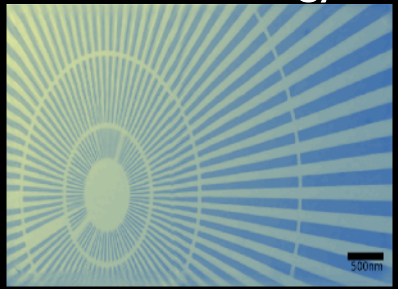
Streamlines in nuclear engineering



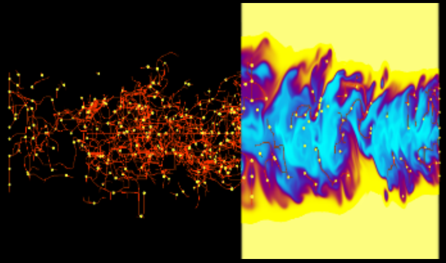
Stream surfaces in meteorology



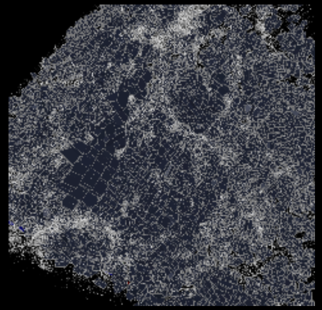
FTLE in climate modeling



Phase reconstruction in X-ray microscopy



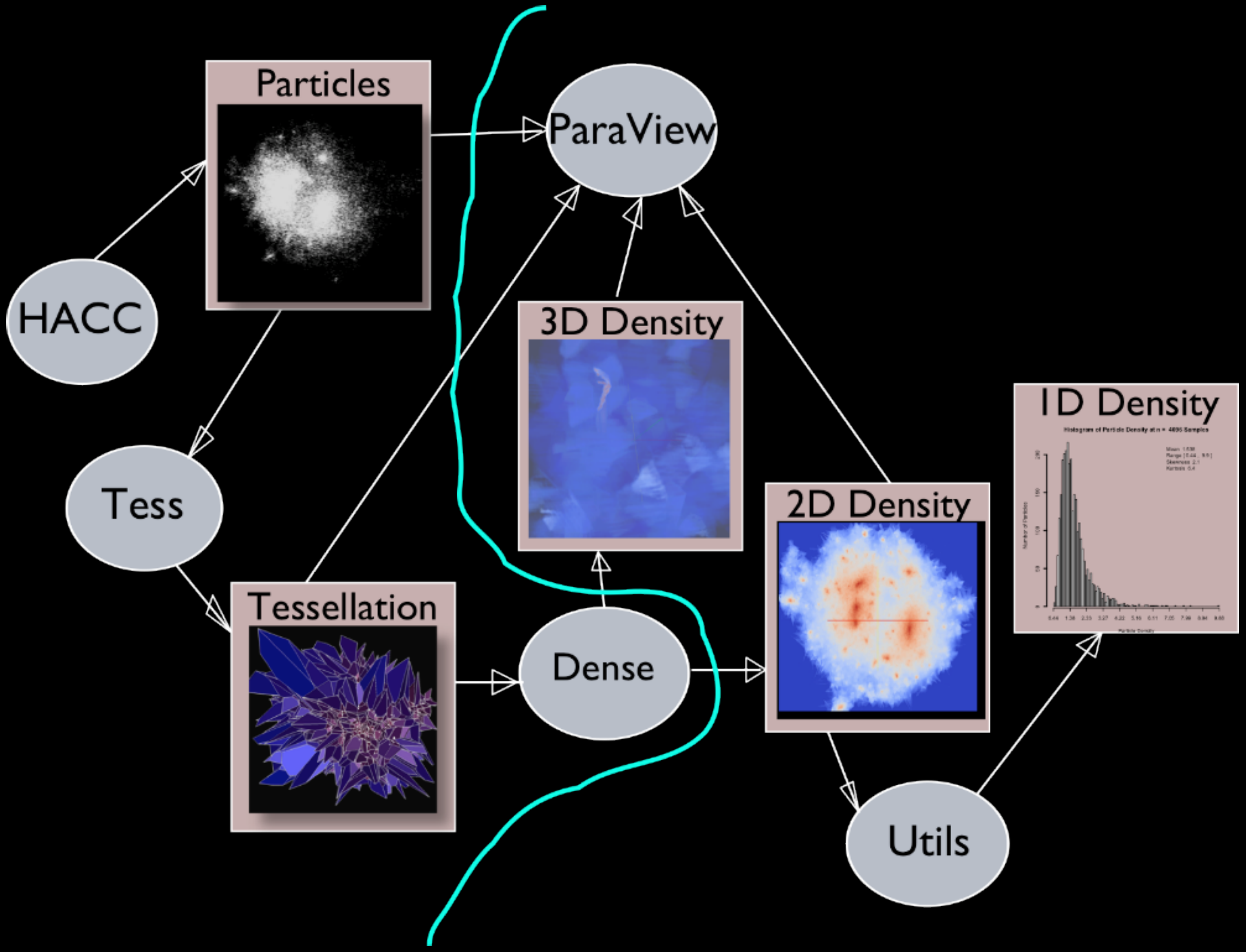
Morse-Smale complex in combustion



Voronoi, Delaunay tessellation in cosmology

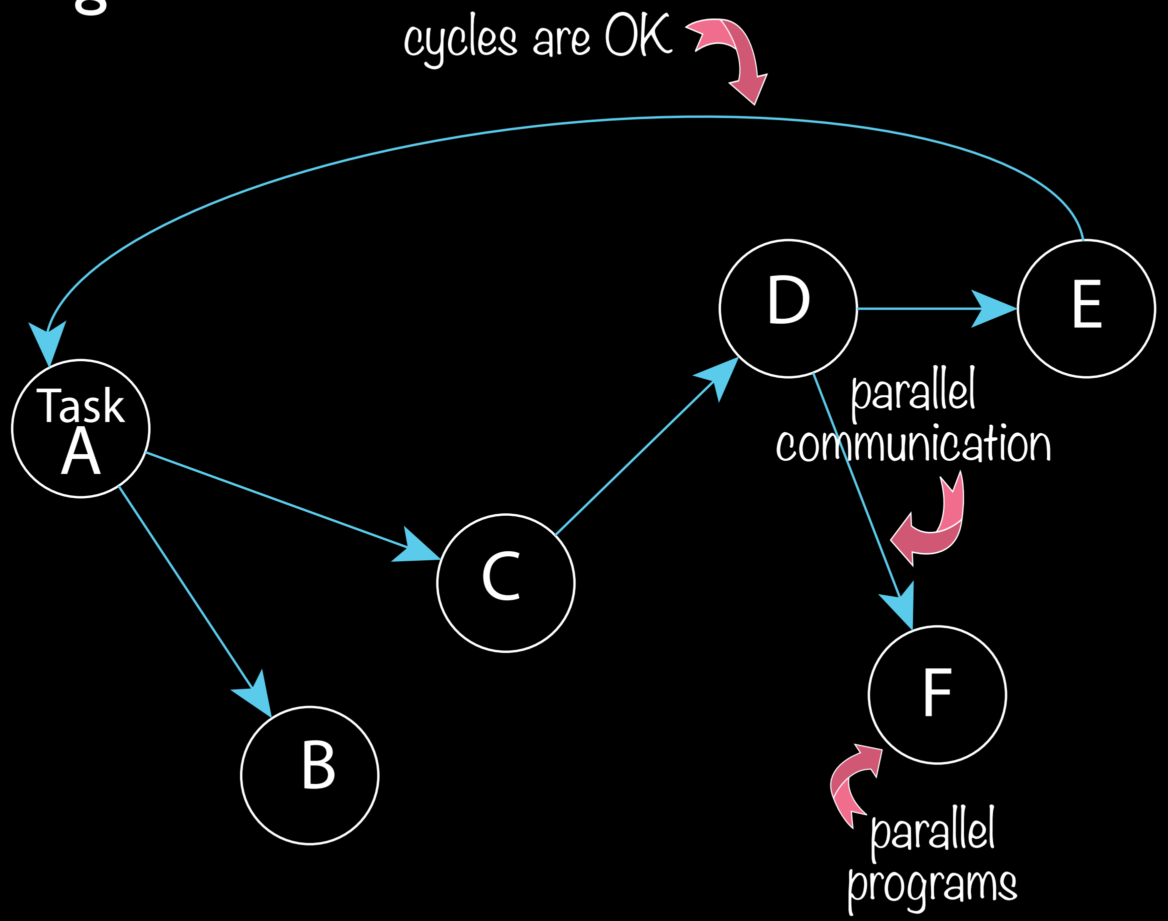
Simple In Situ Workflow Example
Analysis of Cosmology Simulations

- Just one small part of the complete cosmology workflow
- Converts dark matter particles to an unstructured mesh
- Converts an unstructured mesh to a regular grid
- Computes statistics over the grid and visualizes the results



Intertask Programming Model: Workflow

- A directed graph of tasks and communication between them
- Graph nodes are the tasks
- Graph links are the communication

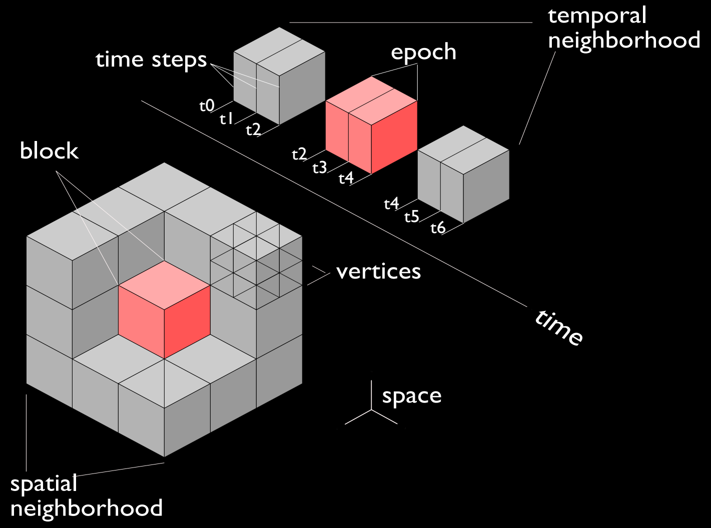


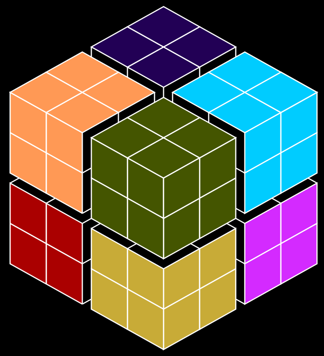
Footnotes

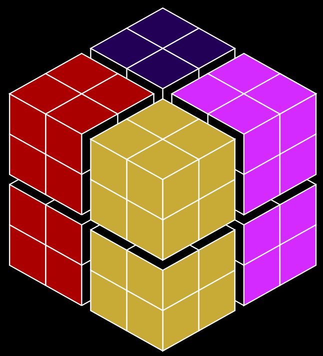
- Notice the graph does not have to be acyclic (digraph, not DAG)
- Think of “large tasks” (programs), not “small tasks” (threads)
- Nodes and links are parallel (parallel programs and parallel communication)

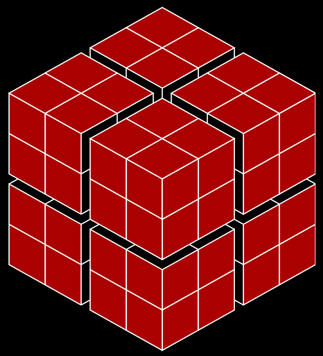
Intratask Programming Model: Block-Parallelism

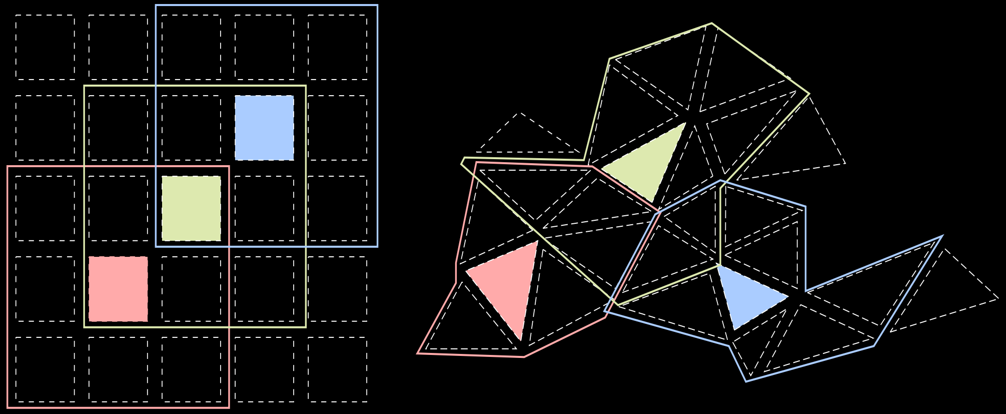
1. Separate analysis ops from data ops
2. Group data items into blocks
3. Assign blocks to processes
4. Group blocks into neighborhoods
5. Handle time
6. Communicate between blocks in reusable design patterns
7. Read data and write results



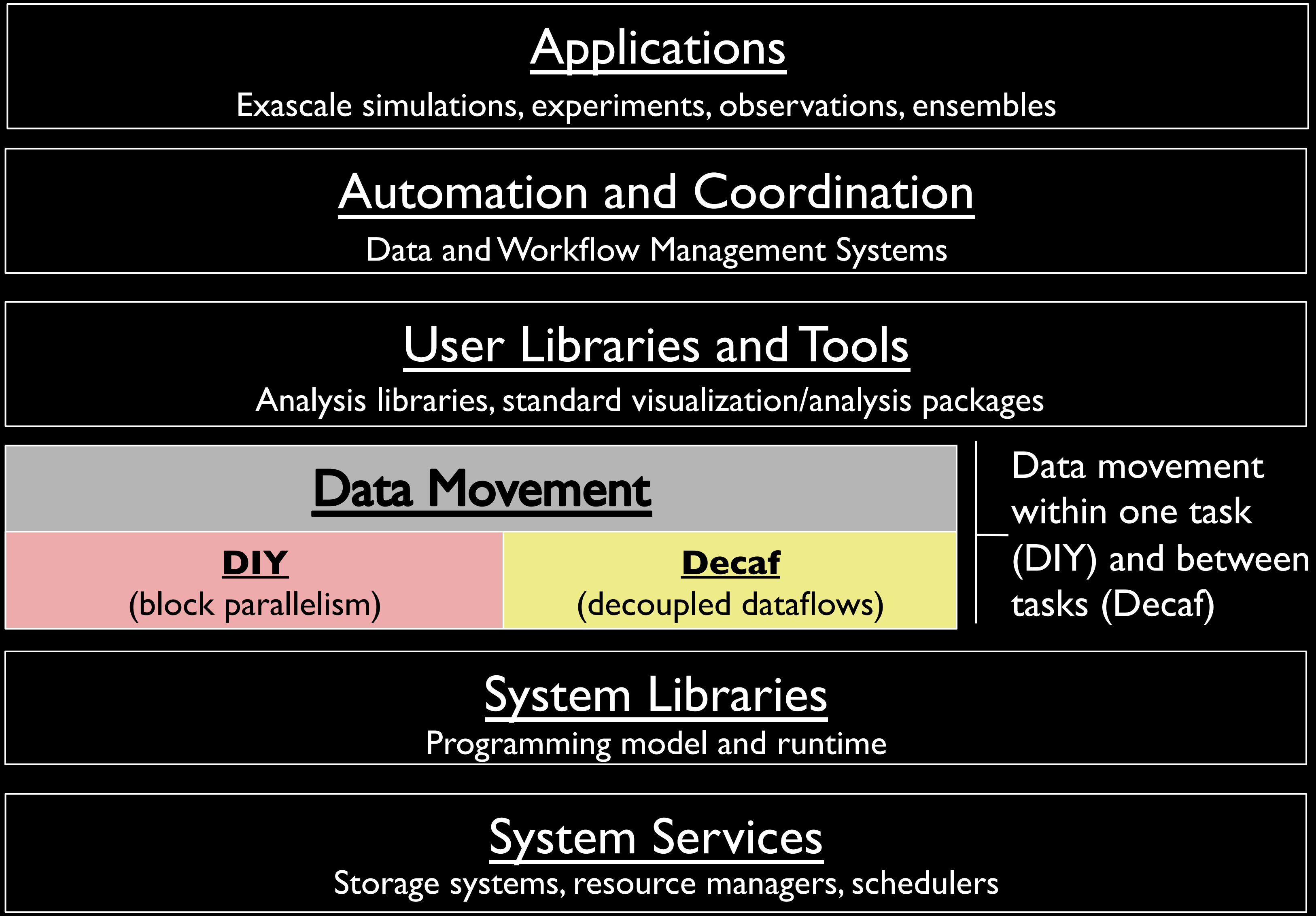

8 processes


4 processes


1 process


Two examples of 3 out of a total of 25 neighborhoods

Software: Exascale Data Analytics Software Stack



Applications

Exascale simulations, experiments, observations, ensembles

Automation and Coordination

Data and Workflow Management Systems

User Libraries and Tools

Analysis libraries, standard visualization/analysis packages

Data Movement

DIY
(block parallelism)

Decaf
(decoupled dataflows)

Data movement within one task (DIY) and between tasks (Decaf)

System Libraries

Programming model and runtime

System Services

Storage systems, resource managers, schedulers

Usability

- Q: Why don’t domain scientists use online analytics?
  - Well, they do, actually. They often embed analytics as function calls directly in codes.

- Q: Why don’t domain scientists use generic middleware for online analytics?
  - Resource cost
  - Reliability
  - Learning curve and usability even later
  - Perceived value
  - Support

- Q: What can we, computer scientists, do to change that?
  - (Why) should we?
  - How?

Acknowledgments

Facilities

Argonne Leadership Computing Facility (ALCF)
Oak Ridge National Center for Computational Sciences (NCCS)
National Energy Research Scientific Computing Center (NERSC)

Funding

DOE SDMAV Exascale Initiative
DOE SciDAC SDAV Institute

People

Franck Cappello (ANL), Matthieu Dreher (ANL), Jay Lofstead (SNL),
Patrick Widener (SNL), Dmitriy Morozov (LBNL)

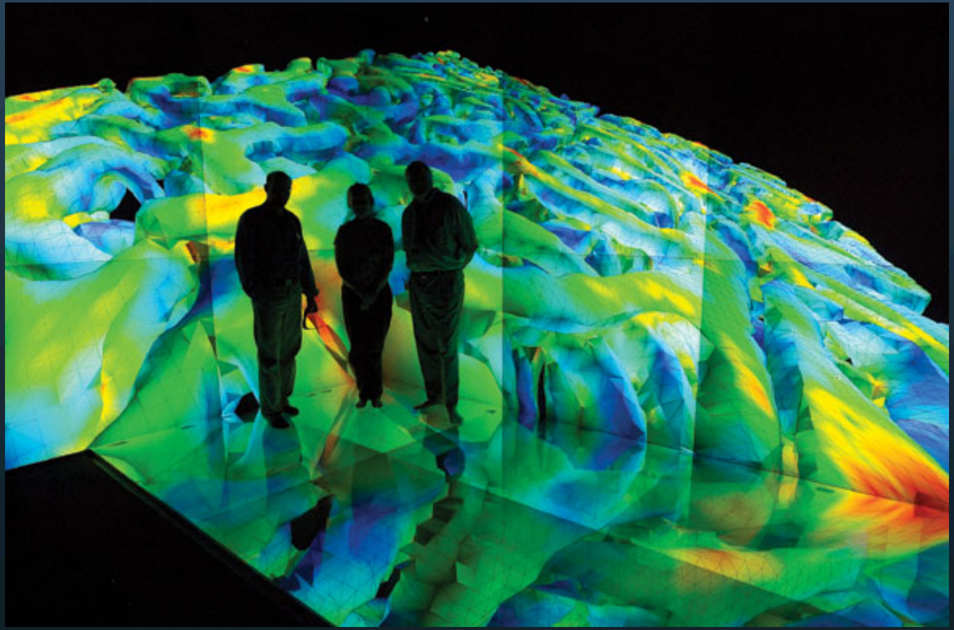
Tom Peterka
tpeterka@mcs.anl.gov
Mathematics and Computer Science Division

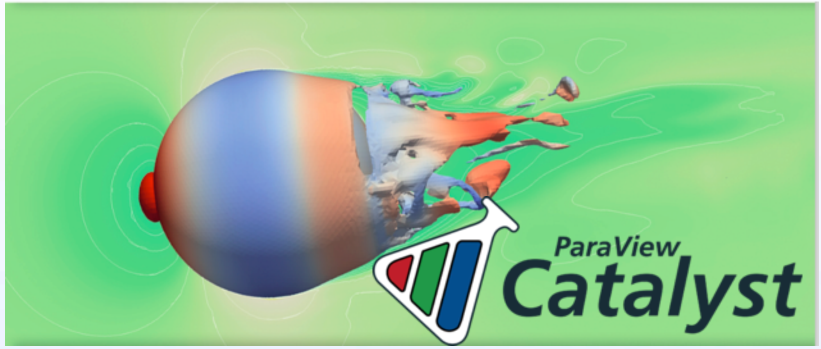
Kitware

On-Line Analytics:
In situ

SSDBM June 2017

Will Schroeder
Kitware Inc

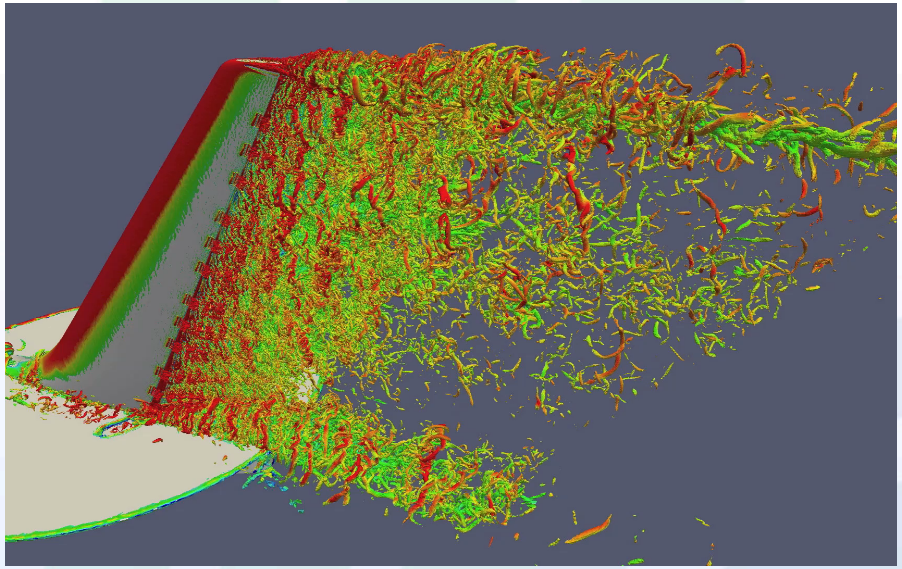




LibSim

ParaView Catalyst

- *In situ* data analysis and visualization
- Challenges
  - There is an explosion of scientific data
  - “Data analysis infrastructure and storage system bandwidth have not scaled proportionately to processing power”
- “Move” the analysis package NOT the data

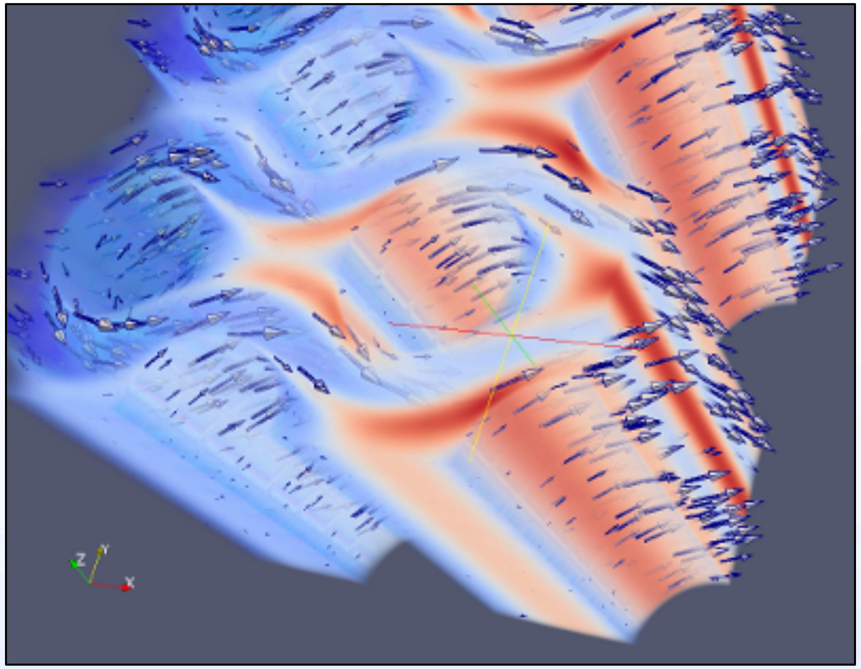


256K MPI ranks on BG/Q Mira at Argonne

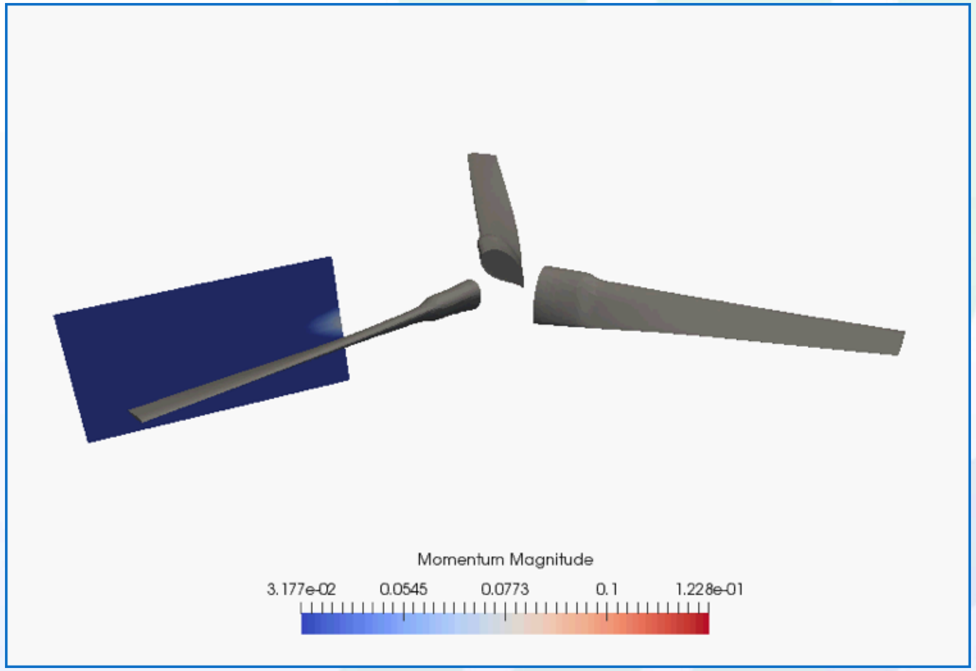
Best HPC Visualization Product or Technology

HPC wire

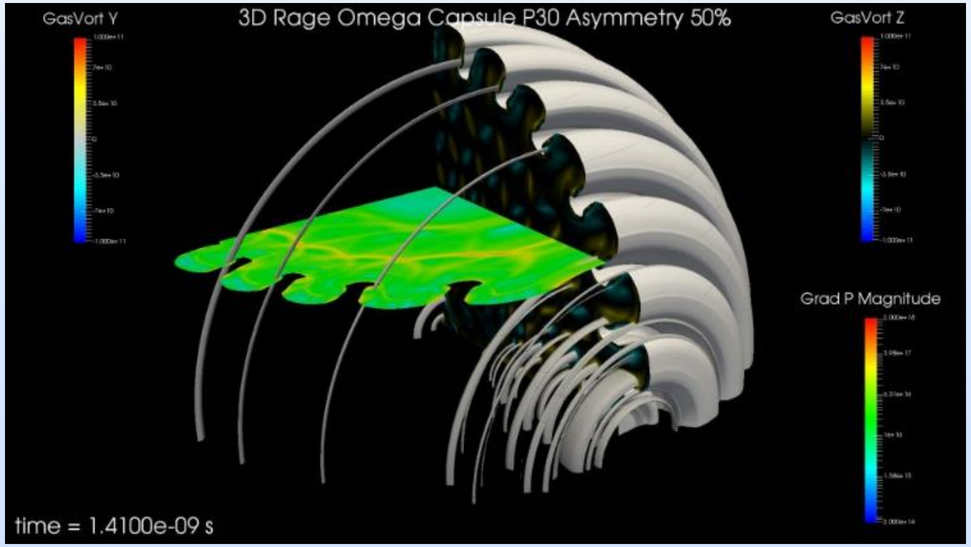
Readers’ Choice Awards 2015



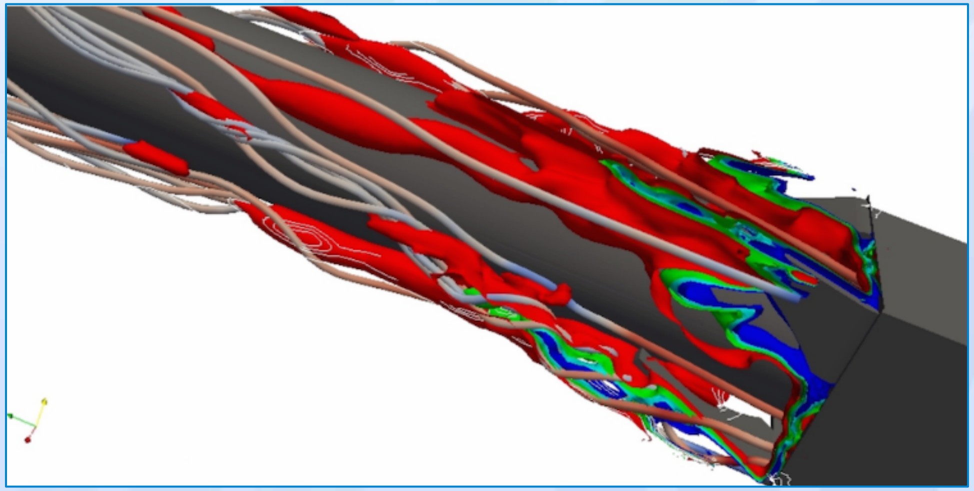
Code Saturne (EDF)



HPCMP CREATE-AV Helios (Army)

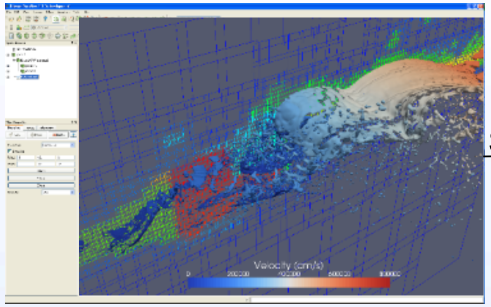


RAGE (LANL)

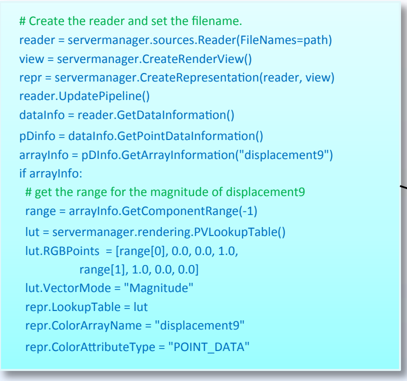


Hydra-TH (LANL)

In Situ Workflow



Script Export



```
# Create the reader and set the filename.
reader = servermanager.sources.Reader(FileNames=path)
view = servermanager.CreateRenderView()
repr = servermanager.CreateRepresentation(reader, view)
reader.UpdatePipeline()
dataInfo = reader.GetDataInformation()
pDinfo = dataInfo.GetPointDataInformation()
arrayInfo = pDInfo.GetArrayInformation("displacement9")
if arrayInfo:
  # get the range for the magnitude of displacement9
  range = arrayInfo.GetComponentRange(-1)
  lut = servermanager.rendering.PVLookupTable()
  lut.RGBPoints  = [range[0], 0.0, 0.0, 1.0,
                range[1], 1.0, 0.0, 0.0]
  lut.VectorMode = "Magnitude"
  repr.LookupTable = lut
  repr.ColorArrayName = "displacement9"
  repr.ColorAttributeType = "POINT_DATA"
```

Augmented script in input deck.

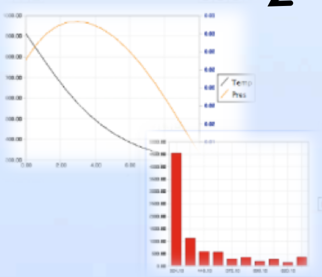
Simulation

ParaView Catalyst

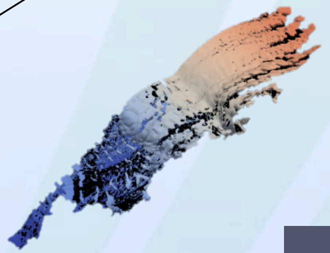
Output Processed Data

256K MPI ranks on BG/Q Mira at Argonne

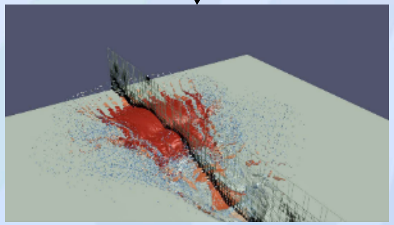
Statistics



Series Data

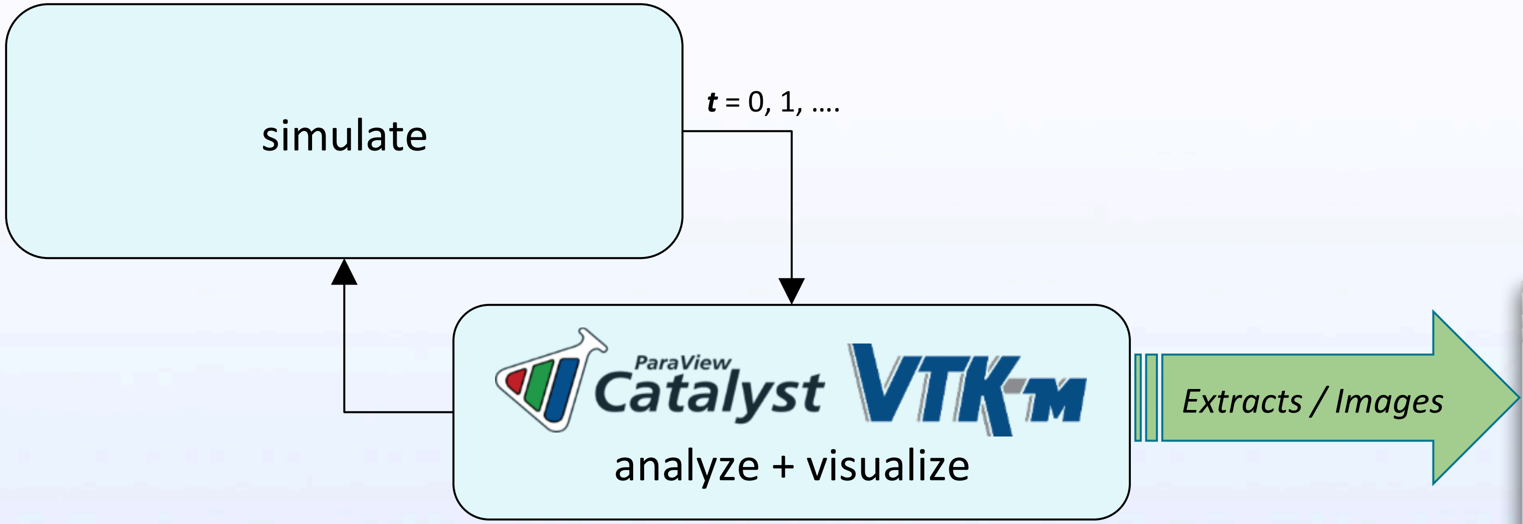


Polygonal Output with Field Data



Rendered Images

ParaView ‘Live’

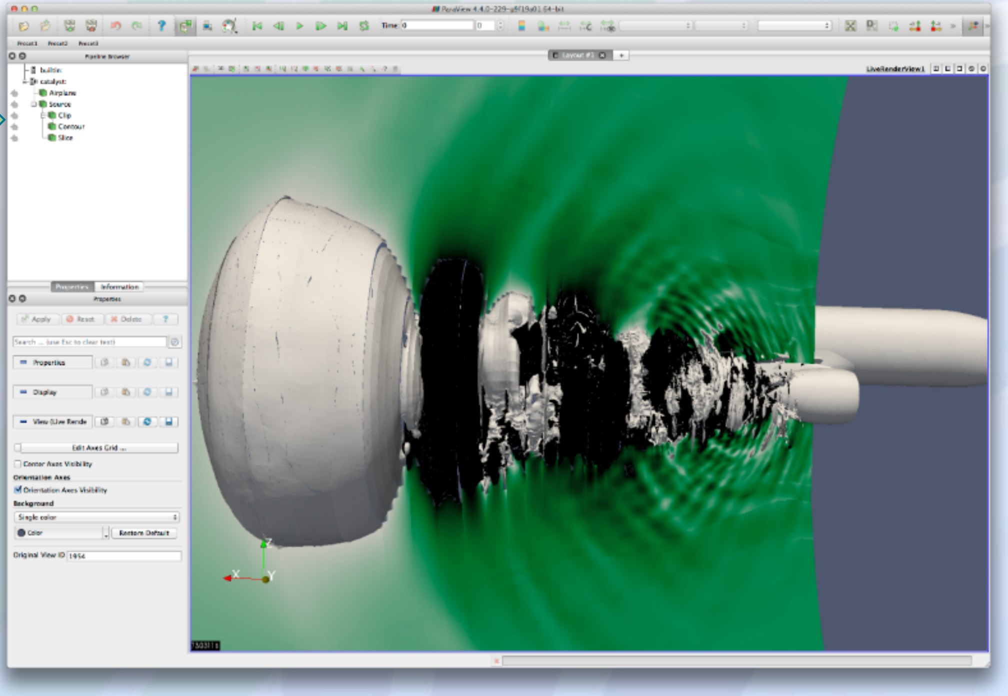


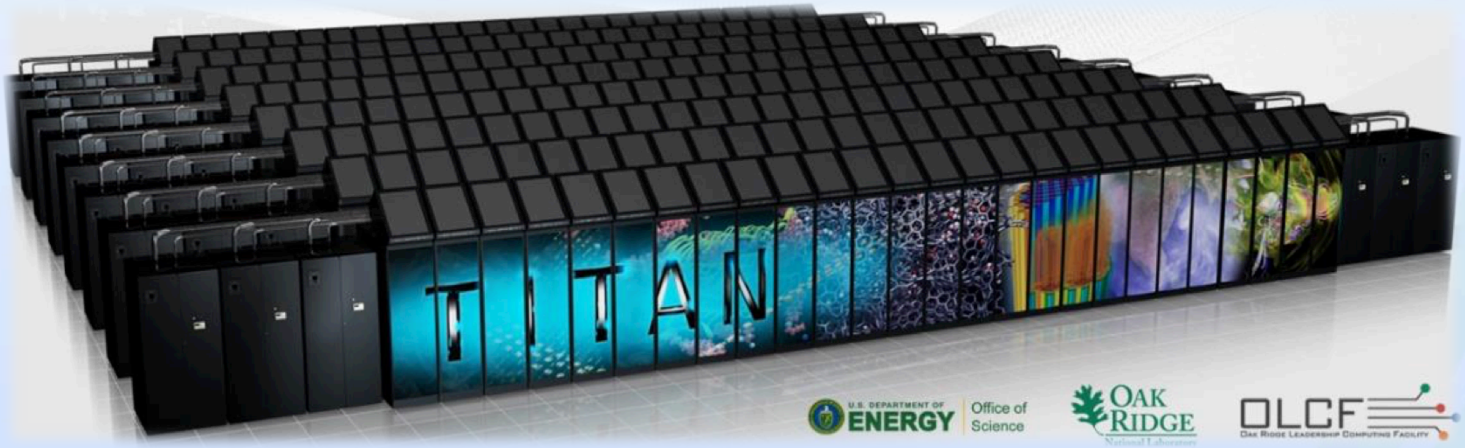
simulate

t = 0, 1, ….

analyze + visualize

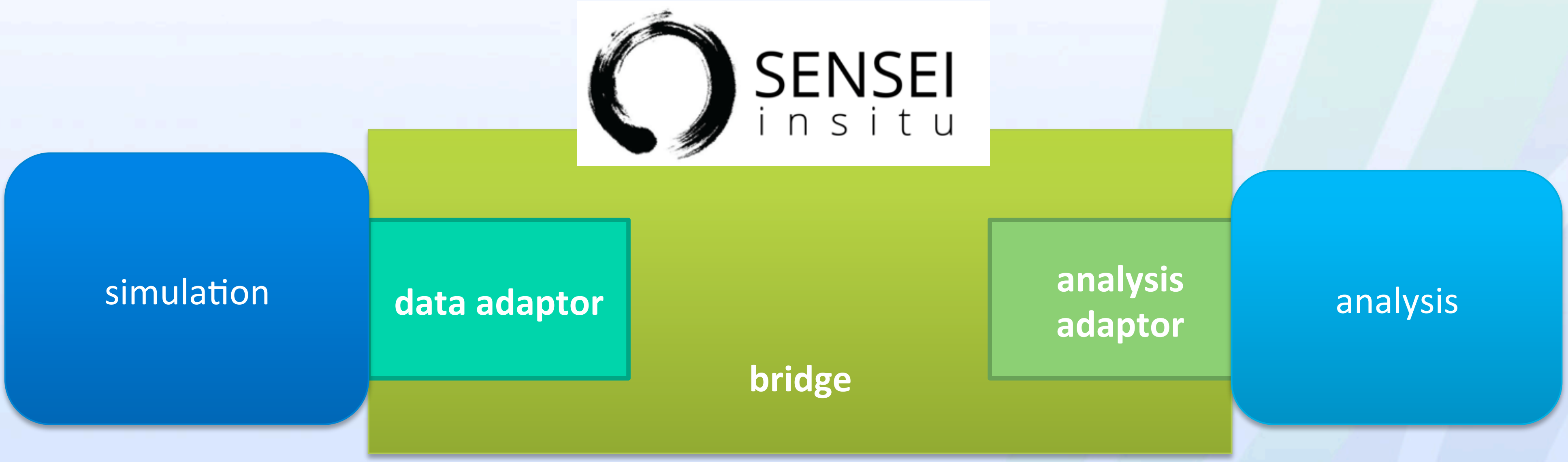
Extracts / Images



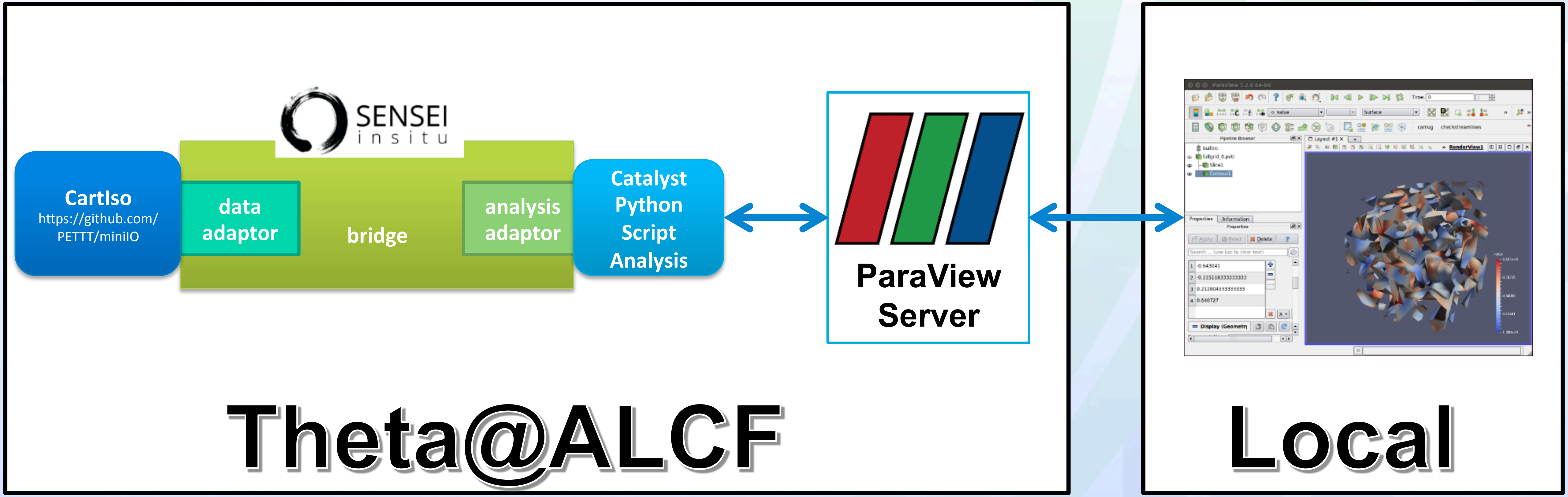


TITAN
U.S. DEPARTMENT OF ENERGY
Office of Science
OAK RIDGE
National Laboratory
OLCF
Oak Ridge Leadership Computing Facility

SENSEI: API: Components



Catalyst Live through Python script

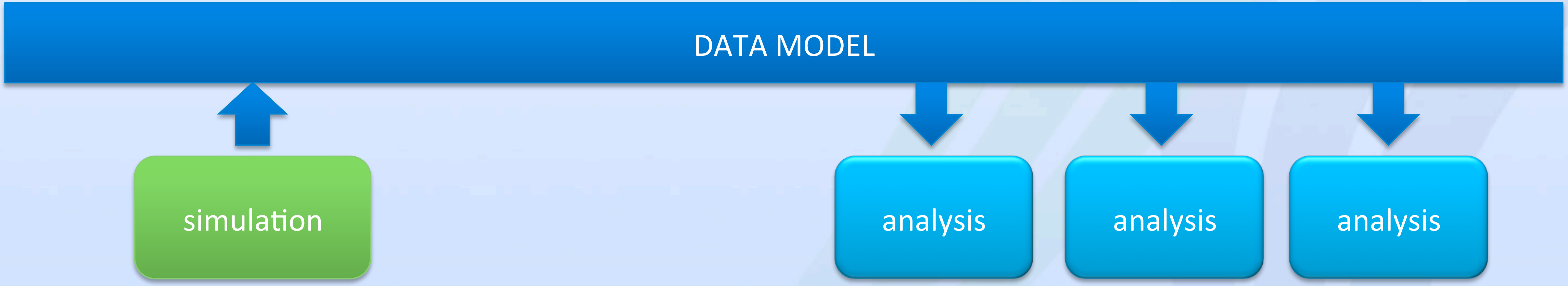


1 -0.643041
2 -0.215118333333333
3 0.212804333333333
4 0.640727

Data Model: VTK

- Used by ParaView/Catalyst and VisIt/Libsim
- Supports common scientific dataset types
- On going independent efforts to evolve for exascale
- Supports using simulation memory directly (zero-copy) for multiple memory layouts

http://www.vtk.org/



DATA MODEL

simulation

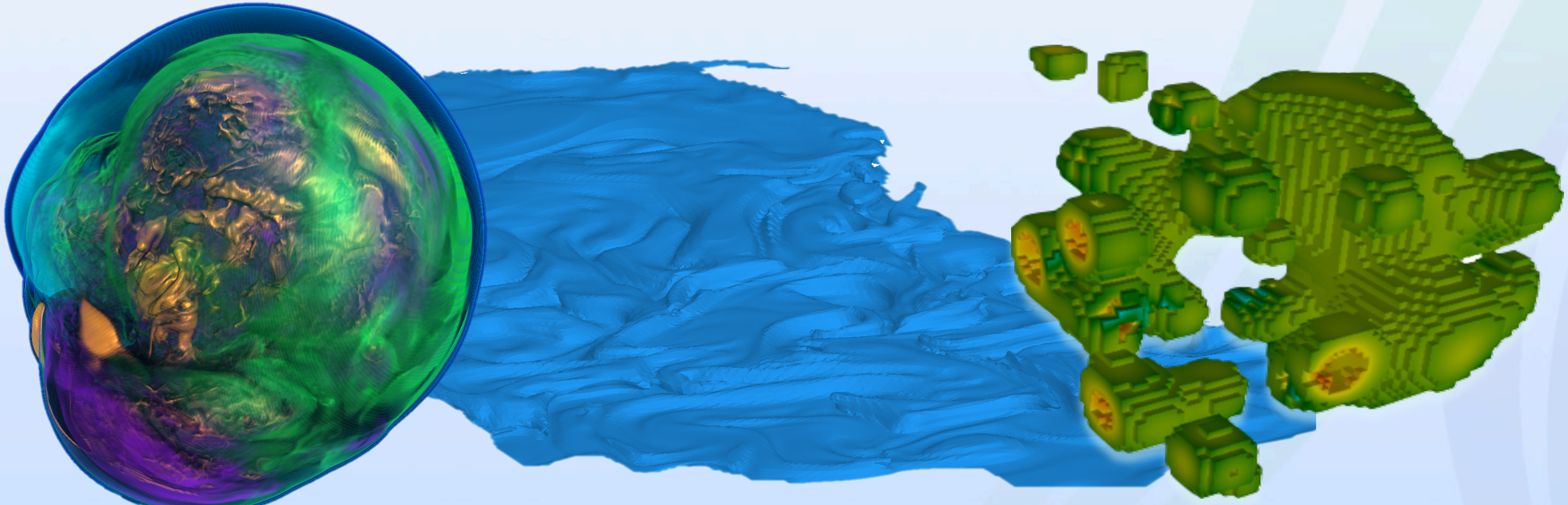
analysis

analysis

analysis

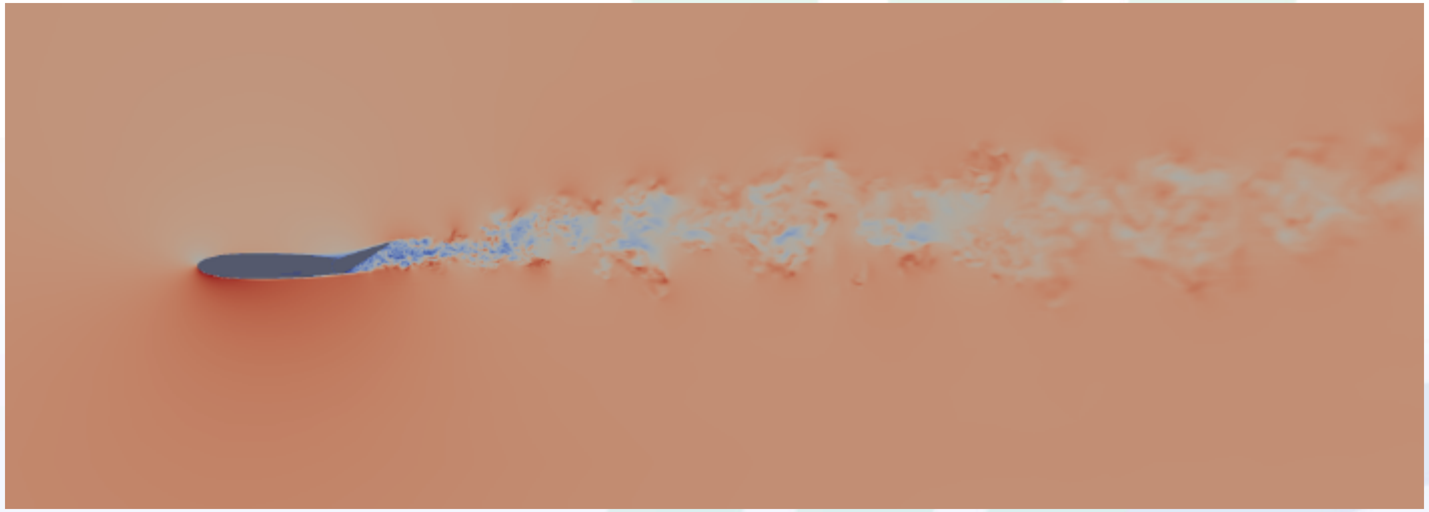
VTK-m

- A single place for the analysis and visualization community to collaborate, contribute, and leverage massively threaded algorithms
- Make it easier for simulation codes to take advantage of these parallel analysis and visualization and algorithms on all next-generation hardware
- Data parallel primitives provide an abstraction layer between the hardware's architecture and the high-level algorithm

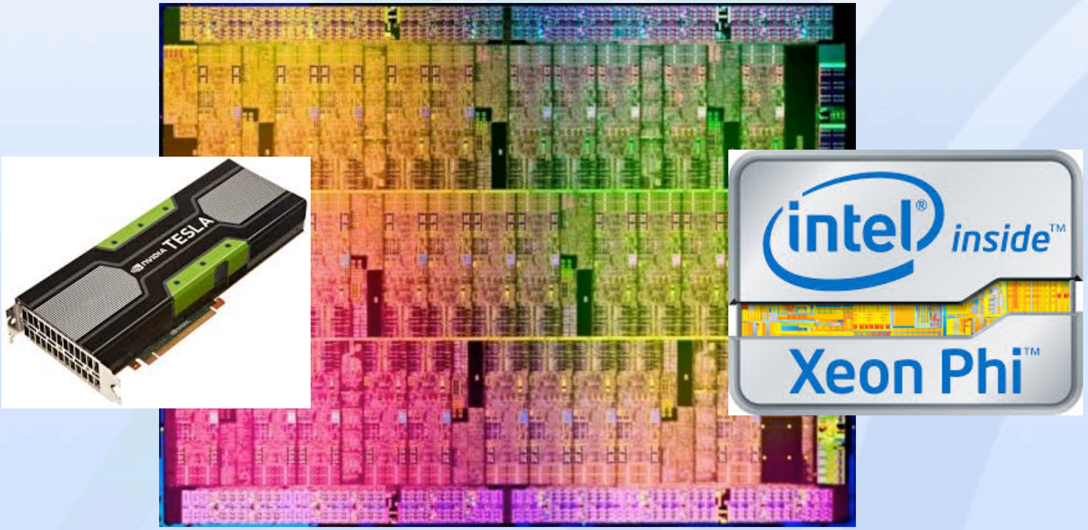


Scalability

- Distributed / MPI
- Shared Memory (TBB / OpenMP / SMPTools)
- VTK-m
  - Many / multi-core
  - TBB / OpenMP / CUDA



6.3 billion-cell unstructured grid using 32,768 nodes on Mira, for a total of 1,048,576 Message Passing Interface (MPI) processes on >500,000 cores (Catalyst / SENSEI / PHASTA). In situ process



Many / Multi-core Architectures

In Situ: What information to extract?

- Emerging, important data features during simulation

- Best visualization “viewpoint” with which to convey information
  - Static
  - Dynamic

- Smart extraction algorithms
  - Useful metrics
  - Feature recognition
  - Machine learning
  - ?

In Situ: Form is the extraction

- Produce minimal information that can be later:
  - Reconstructed
  - Interacted with

- Are classic visualization features best?
  - Other forms (compressed data subsets)
  - Basis or parameter sets that can be applied to preprocessed data
    - Evolution of geometry
    - Deltas from low-res baseline

Visualization Algorithms

- Effectively support simulation process
  - Fast, interactive exploration
  - Low memory footprint; minimal resource consumption
  - Temporal visualization
  - Hierarchy of tools that tradeoff speed for fidelity