Near real-time processing of scientific data

WoNDP: 3rd workshop on near-data processing
Waikiki Hawaii
December 5 2015

S. Klasky	ORNL, GT U. Tenn., NCSU
H. Abbasi, J. Choi, Q. Liu, N. Podhorszki	ORNL
M. Ainsworth	Brown, ORNL
C. S. Chang, M. Churchill	PPPL
J. Wu	LBNL

klasky@ornl.gov

Outline

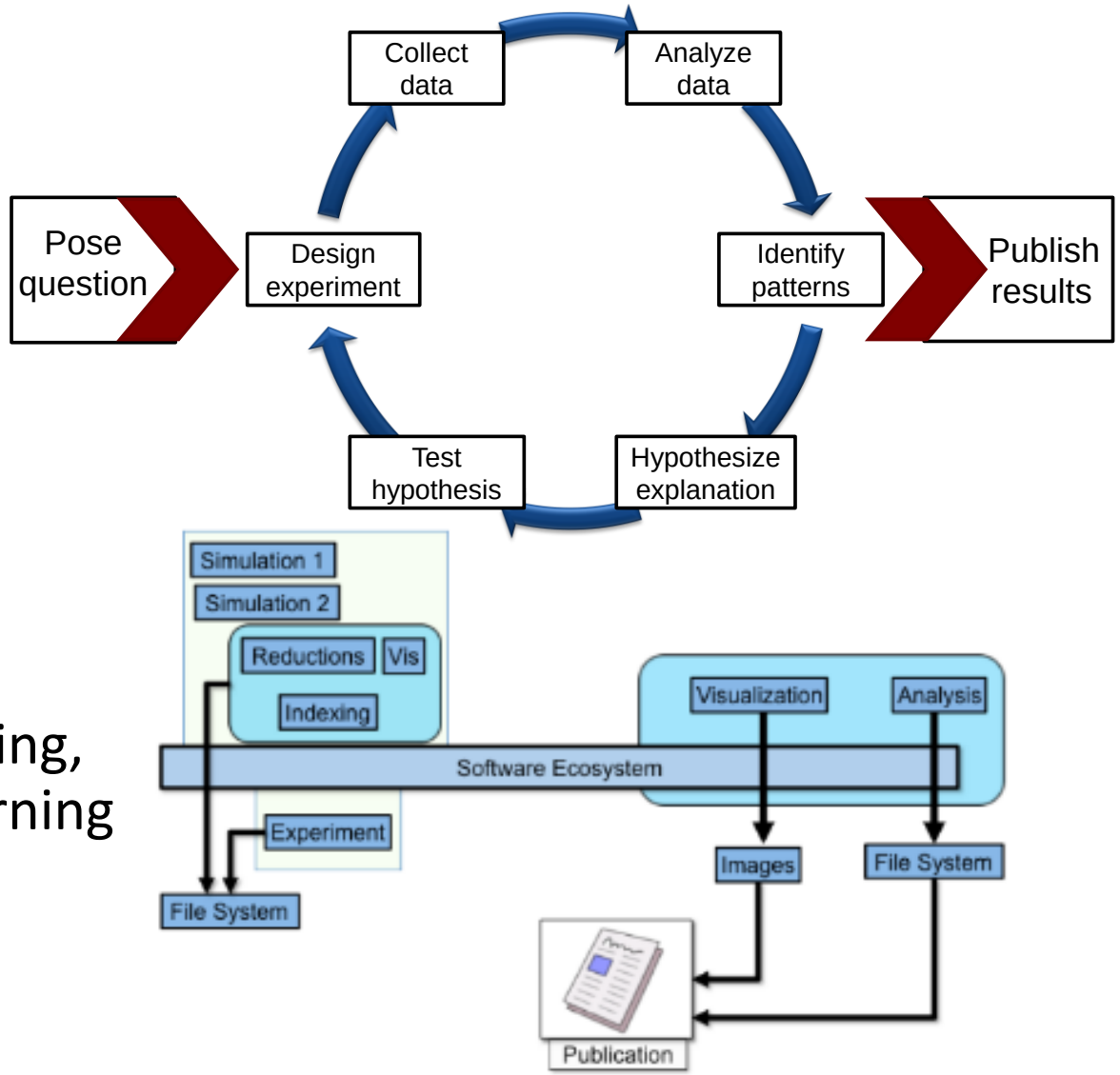
- Motivation
- Our contributions
- Examples
- SIRIUS: Next generation Multi-tier storage and I/O system
- Closing thoughts

Disclaimer: I do NOT pay much attention to hardware….

klasky@ornl.gov

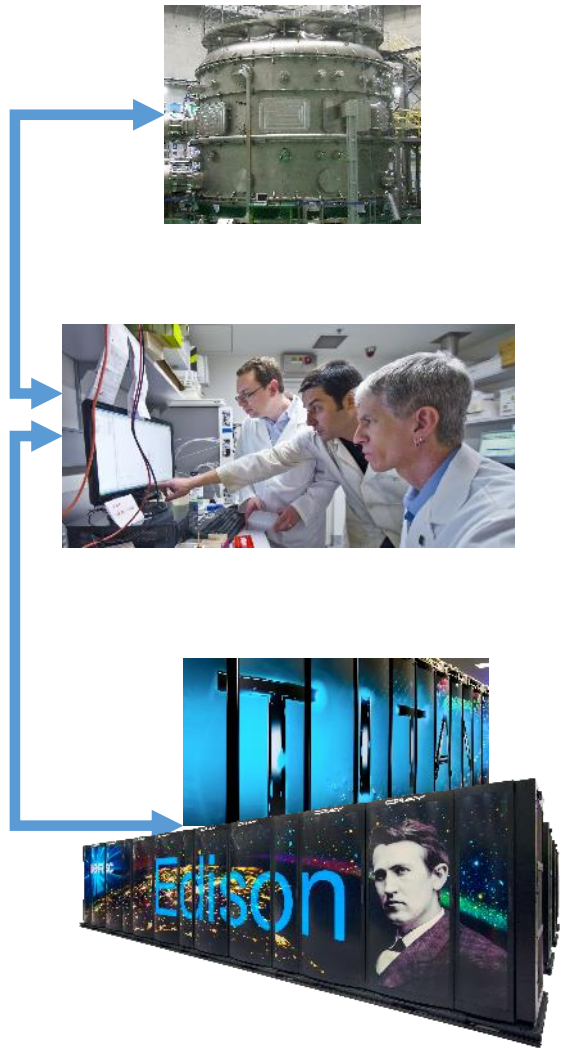
Where do we spend our time in science

- Goals
  - Accelerate this process
  - Make the process **predictable**
  - Make the process **adaptable**
  - Make the process **scalable** as the complexity increases
  - Make the software **easy-to-use**
- Observation
  - Most of the time is spent in managing, moving, storing, retrieving , and turning the science data into **knowledge**



Vision: Enable Rapid Collaborative Decision Making

- Vision: Enable distributed, collaborative, real-time decisions
  - Workflows including both experiments and simulations
  - Reduce cost, improve utilization of expensive experimental devices
- Metrics of Success:
  - Reduction of time to make a “good” decision, across the entire scientific process
  - Adoption of technology by “important users”



How to Enable Rapid Decision Making

- Effective data management
  - Easily express data accesses: high-level data model instead of offsets into files
  - Transparent accesses to remote data
  - Convenient querying operations
- Effective workflow management
  - Tight integration of workflow components to reduce latency
  - Make the best uses of known resources
- Reduce the time to solution
  - Streaming data accesses, avoid waiting for all data before analysis could start
  - Only access the necessary data records (selective data accesses)
  - Keep the data in memory as much as possible

klasky@ornl.gov

Example: Fusion Experiments

- Complex DOE experiments, such as a fusion reactor, contain numerous diagnostics that need Near-Real-Time analysis for feedback to the experiment
  - For guiding the experiment
  - For faster and better understanding of the data
- Current techniques to write, read, transfer, and analyze "files" require a long time to produce an answer
  - Long delay due to slow disks involved to store and retrieve files
  - Slow start up of many workflow execution engines

Big Data in Fusion Science: ITER example

- **Volume**: Initially 90 TB per day, 18 PB per year, maturing to 2.2 PB per day, 440 PB per year
- **Value**: All data are taken from expensive instruments for valuable reasons.
- **Velocity**: Peak 50 GB/s, with near real-time analysis needs
- **Variety**: ~100 different types of instruments and sensors, numbering in the thousands, producing interdependent data in various formats
- **Veracity**: The quality of the data can vary greatly depending upon the instruments and sensors.

The pre-ITER superconducting fusion experiments outside of US will also produce increasingly bigger data (KSTAR, EAST, Wendelstein 7-X, and JT60-SU later).

klasky@ornl.gov

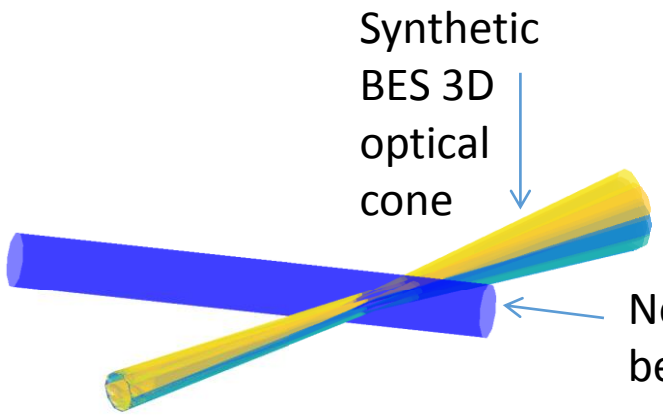
Validation Laboratories

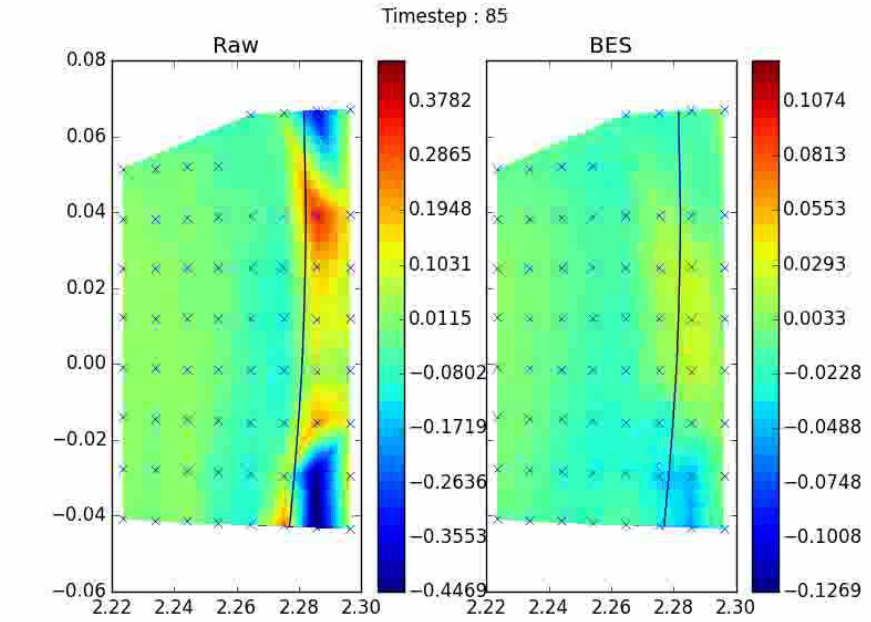
- **Goal**
  - Create a framework which can allow scientists to fuse experimental and computational data to aid in the validation process, transitioning this from an Art to a Science
- **Research Challenges**
  - Encapsulate sufficient semantic information in a workflow language to allow global optimizations to be performed
  - Scheduling across heterogeneous resources (memory, cores, systems, networks)
  - Fusion of data in an automated workflow
  - Ensemble comparison using comparative analytics
  - Extract relationships of experimental and simulation data
- **Metrics of Success**
  - Increasing the number of users who manually validate their data to automated workflows, decreasing their time in “baby-sitting”
  - Accuracy of data mining for discovery of correlations to aid validation

klasky@ornl.gov

Synthetic Diagnostics

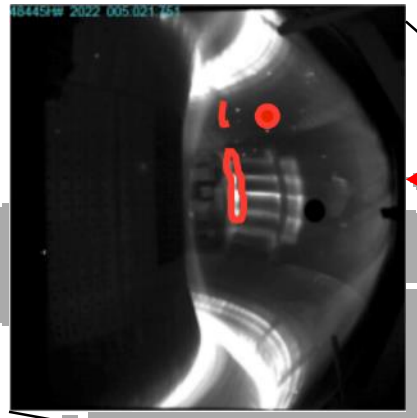
- Enables direct comparison of simulation results to experiment
- Example of beam emission spectroscopy (BES) using XGC1 simulation data





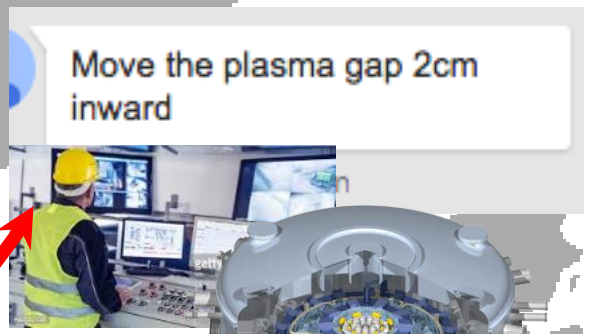
klasky@ornl.gov

Example ITER workflow: Anomaly detection, analysis, and feedback



Hot spots detected & analyzed!!

Remote scientist

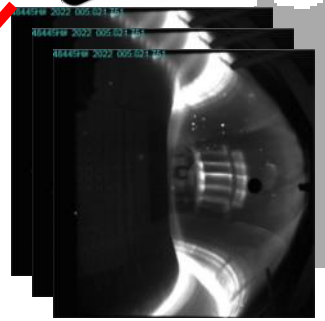


near-real-time Feedback

(Resource management) offload compute depending on availability

Image processing, anomaly detection, and near-real-time prediction at remote compute center

ICEE WAN data transfer



~600MB/s

klasky@ornl.gov

Outline

- Motivation
- Our contributions
- Examples
- SIRIUS: Next generation Multi-tier storage and I/O system
- Closing thoughts

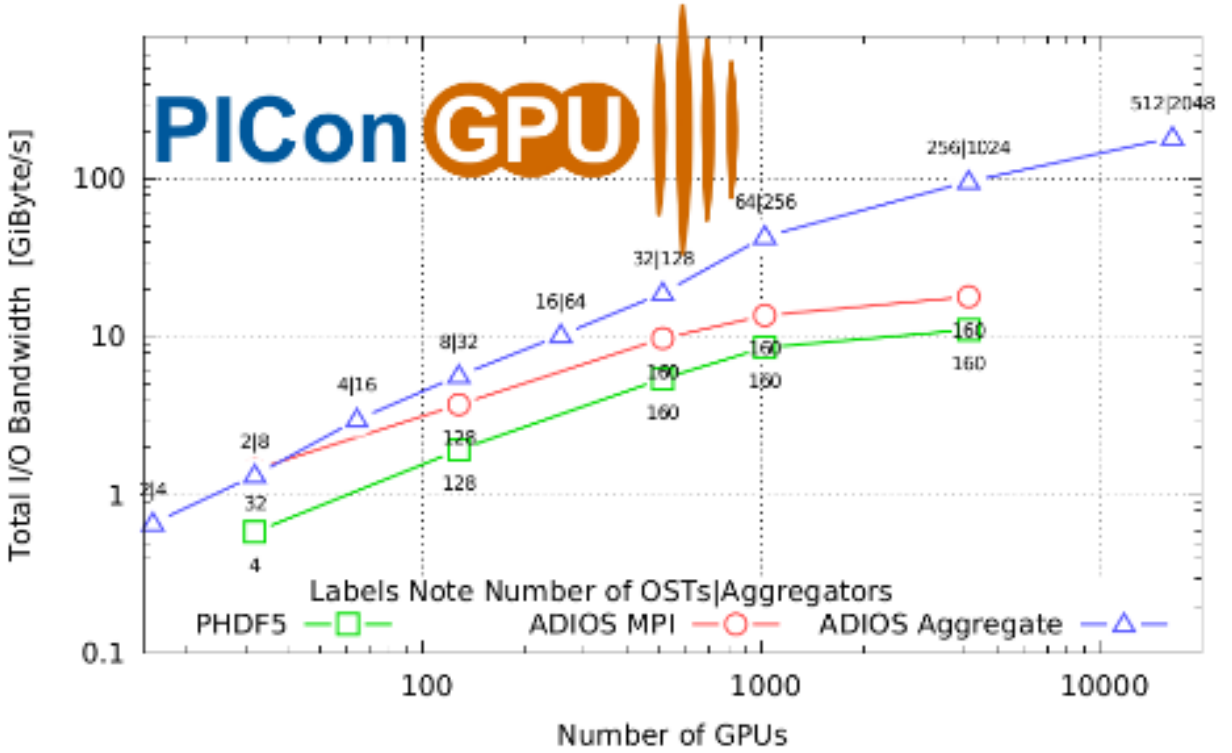
klasky@ornl.gov

Our Approach

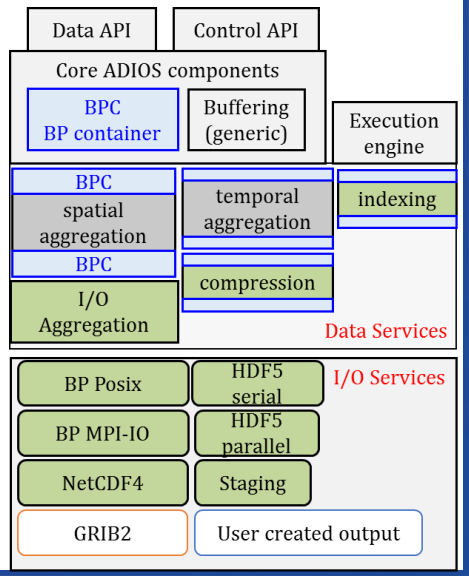
- Create an I/O abstraction layer for
  - Writing data quickly on exa, peta, tera, giga scale resources transparently
  - Streaming data on these resources, and across the world
- Place different parts of a workflow at different locations
  - Move work to data whenever possible
- Research new techniques for quickly indexing data to reduce the amount of information moved in the experimental workflow
  - Prioritize data
- Create new techniques to identify important features, which turn the workflow into a data-driven streaming workflow

ADIOS

- An I/O abstraction framework
- Provides portable, fast, scalable, easy-to-use, metadata rich output
- Choose the I/O method at runtime
- Abstracts the API from the method
- Need to provide solutions for “90% of the applications”



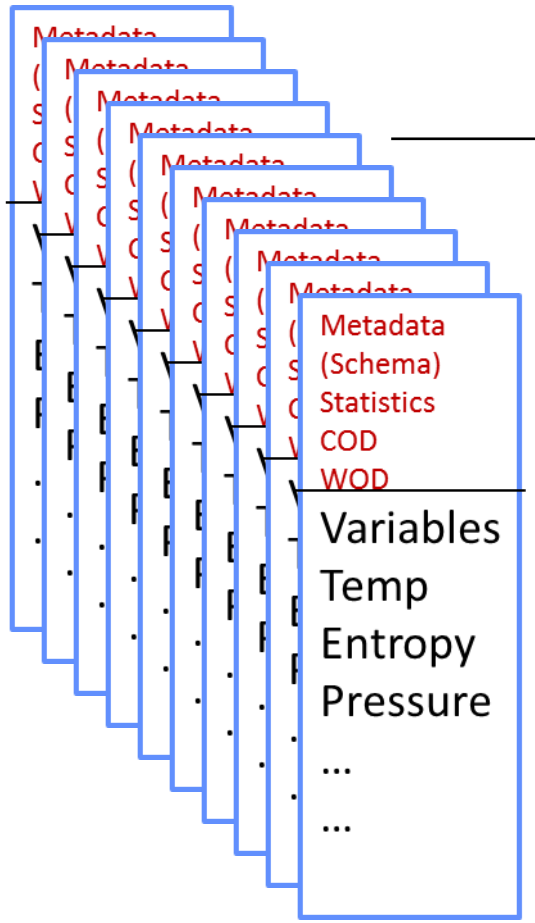
- Astrophysics
- Climate
- Combustion
- CFD
- Environmental Science
- Fusion
- Geoscience
- Materials Science
- Medical:
- Pathology
- Neutron Science
- Nuclear Science
- Quantum Turbulence
- Relativity
- Seismology
- Sub-surface modeling
- Weather

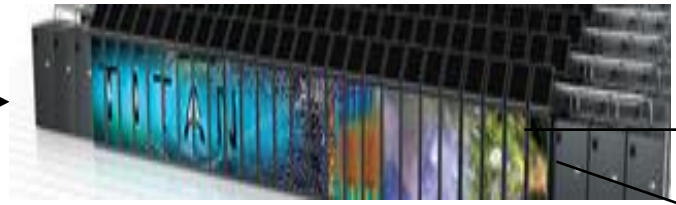


http://www.nccs.gov/user-support/center-projects/adios/

klasky@ornl.gov

The ADIOS-BP Stream/File format

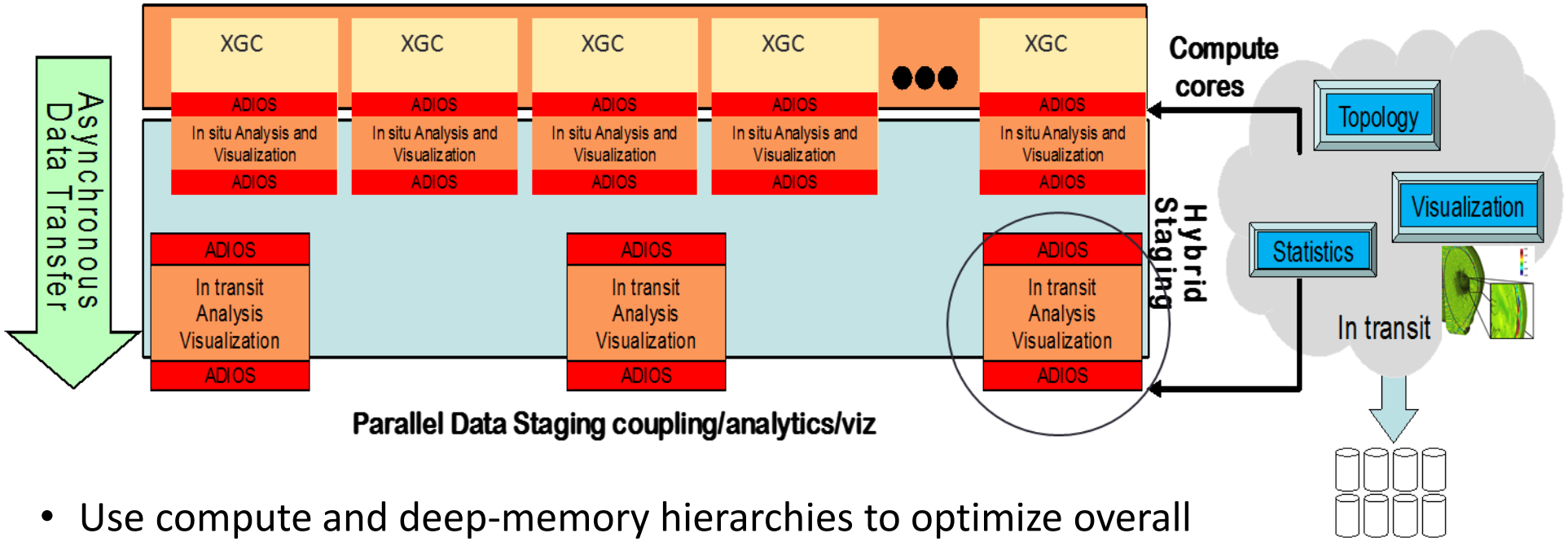




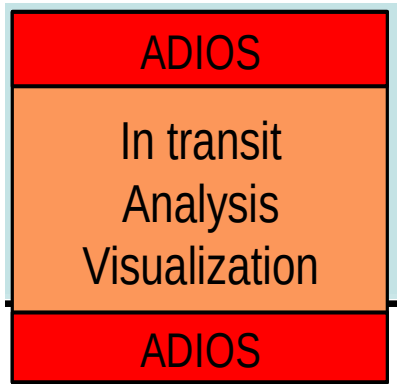
- All data chunks are from a single producer
  - MPI process, Single diagnostic
- Ability to create a separate metadata file when "sub-files" are generated
- Allows variables to be individually compressed
- Has a schema to introspect the information
- Has workflows embedded into the data streams
- Format is for "data-in-motion" and "data-at-rest"

Ensemble of chunks = file

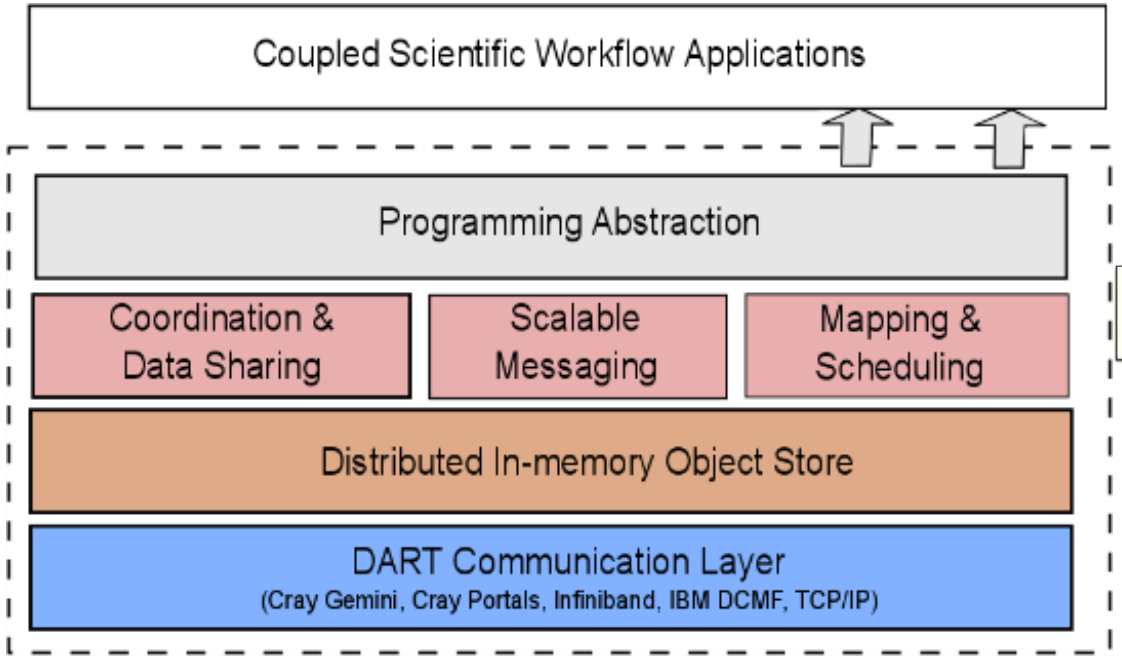
Hybrid Staging: Flexibility in processing



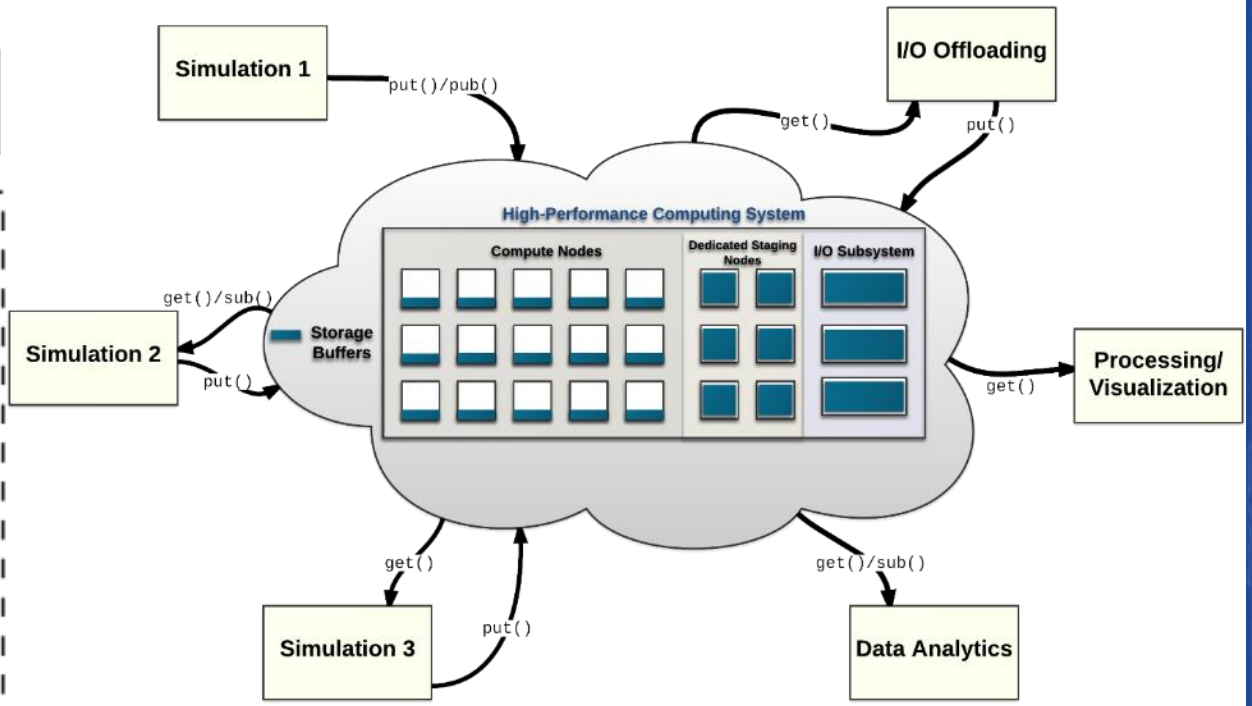
- Use compute and deep-memory hierarchies to optimize overall workflow for power vs. performance tradeoffs
- Abstract complex/deep memory hierarchy access
- Placement of analysis and visualization tasks in a complex system
- Impact of network data movement compared to memory movement



DataSpaces – Rutgers



The DataSpaces Abstraction

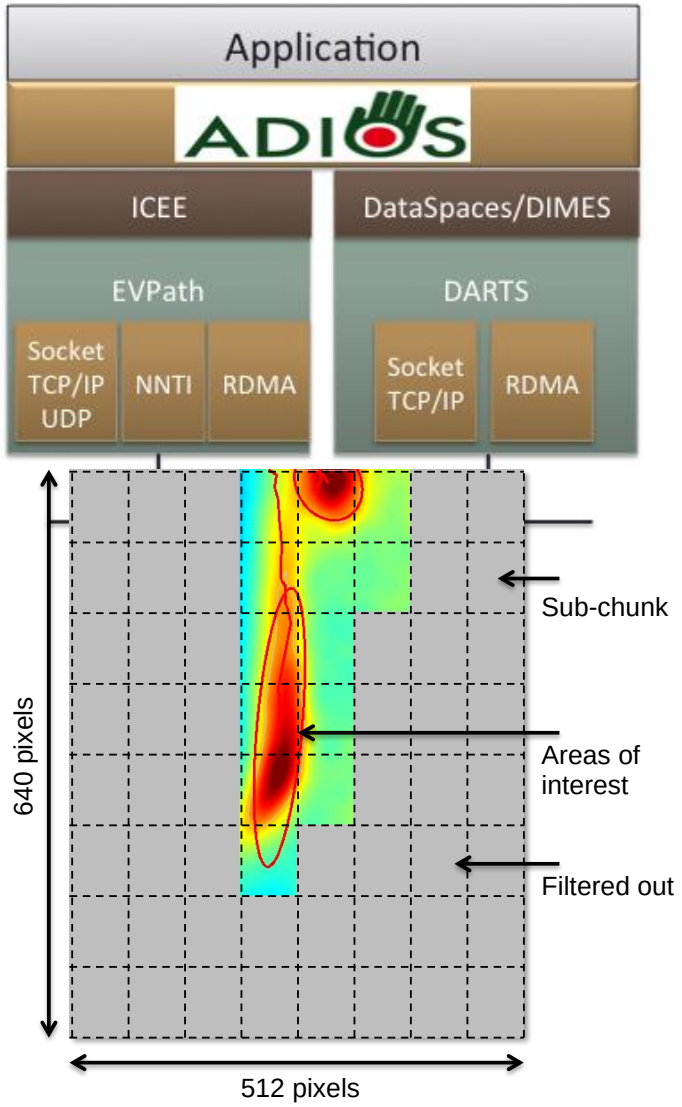


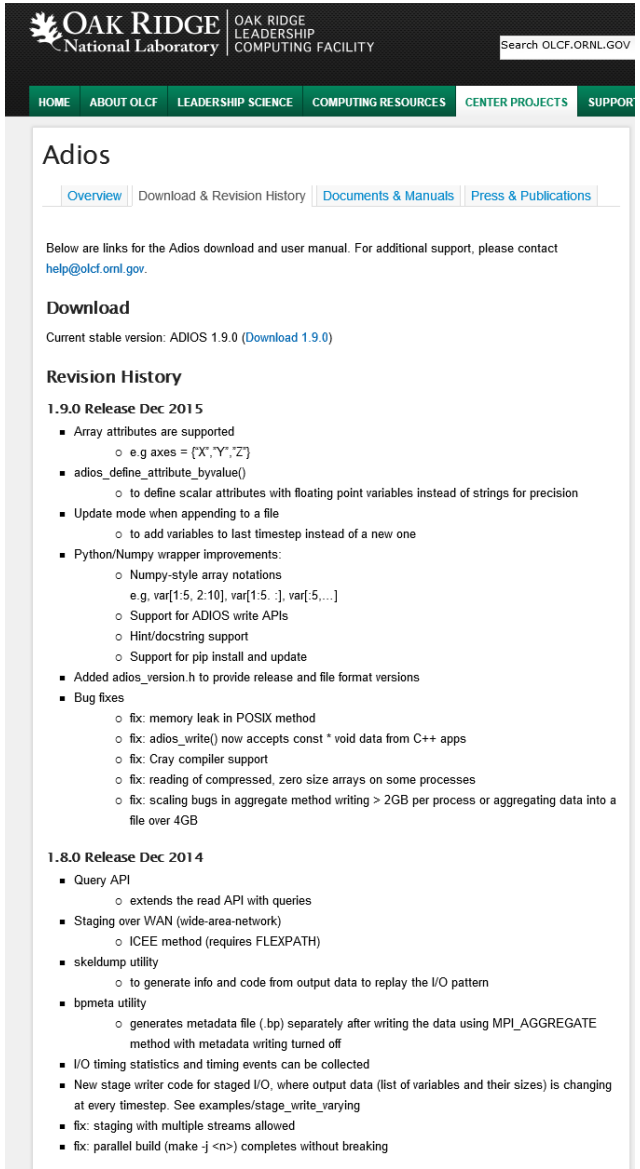
- Virtual shared-space programming abstraction
  - Simple API for coordination, interaction and messaging
- Distributed, associative, in-memory object store
  - Online data indexing, flexible querying
- Adaptive cross-layer runtime management
  - Hybrid in-situ/in-transit execution
- Efficient, high-throughput/low-latency asynchronous

klasky@ornl.gov

Data Movement methods

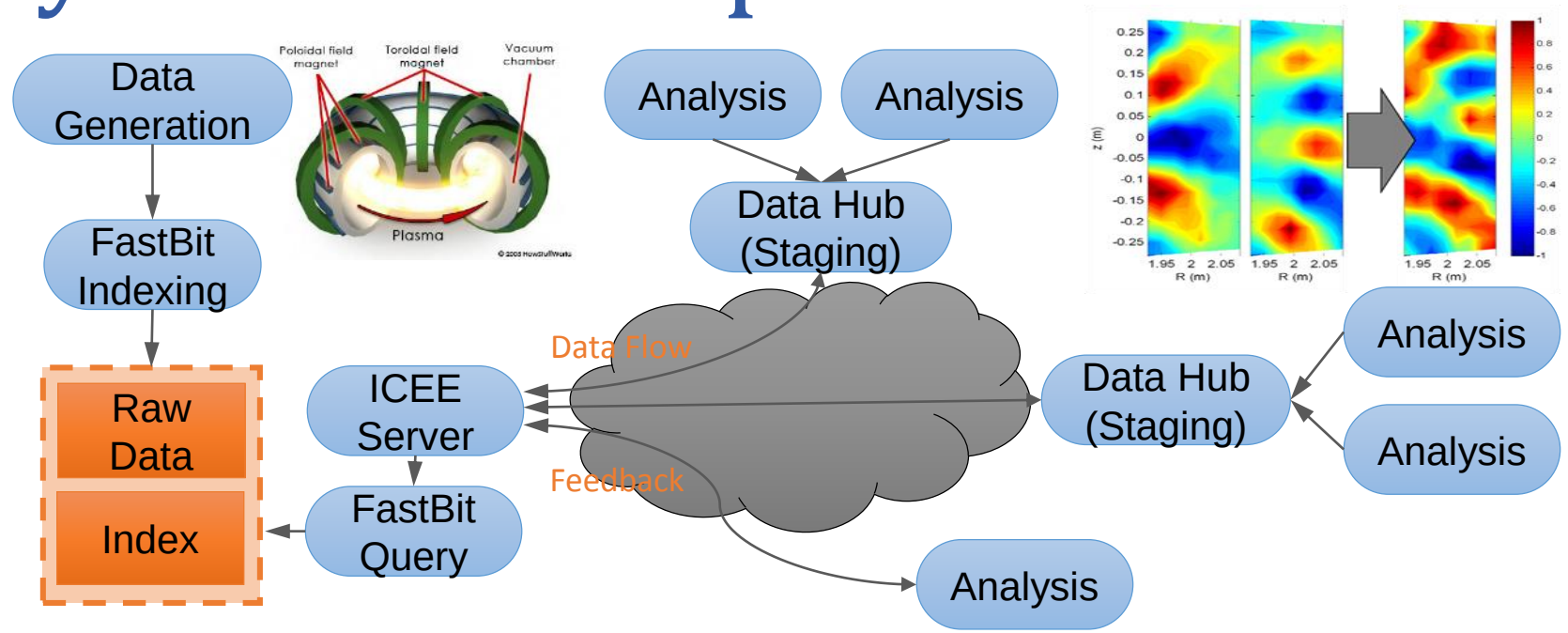
- ICEE
  - Using EVPath package (GATech)
  - Support uniform network interface for TCP/IP and RDMA
  - Easy to build an overlay network
- Dataspaces (with sockets)
  - Developed by Rutgers
  - Support TCP/IP and RDMA
- Select only areas of interest and send (e.g., blobs)
- Reduce payload on average by about 5X





klasky@ornl.gov

ICEE System Development With ADIOS



- Features
  - ADIOS provides an overlay network to share data and give feedbacks
  - Stream data processing – supports stream-based IO to process pulse data
  - In transit processing – provides remote memory-to-memory mapping between data source (data generator) and client (data consumer)
  - Indexing and querying with FastBit technology

klasky@ornl.gov

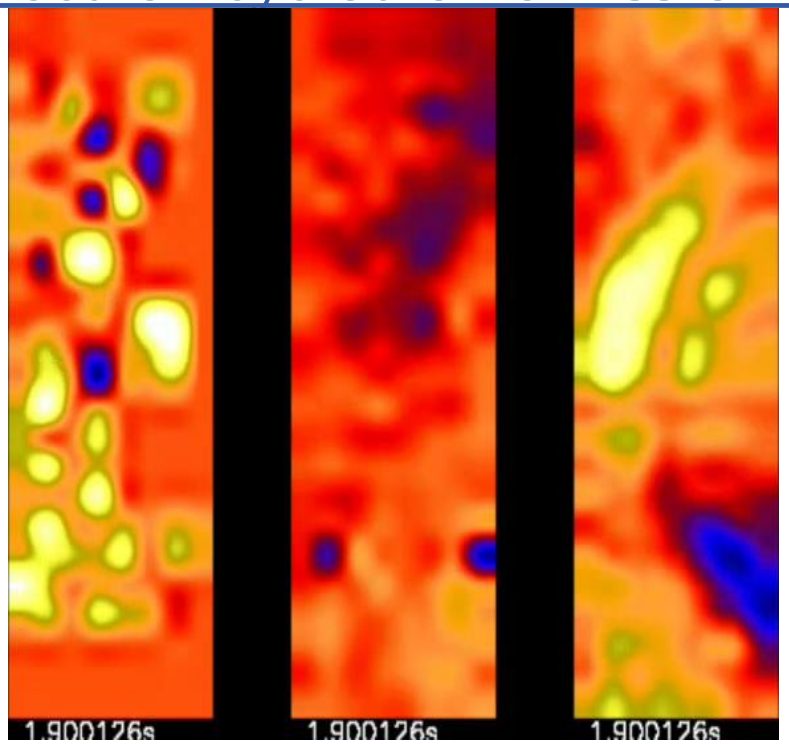
Outline

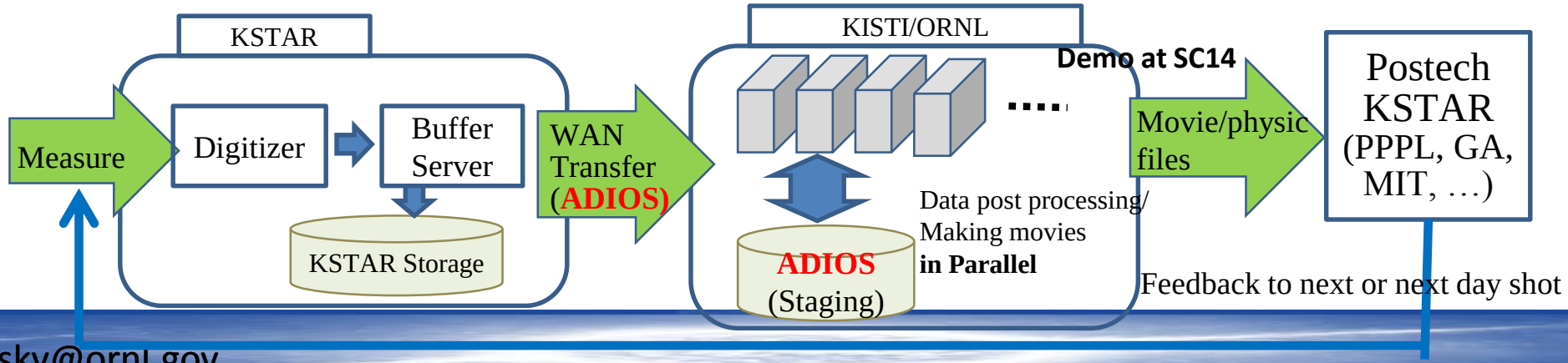
- Motivation
- Our contributions
- Examples
- SIRIUS: Next generation Multi-tier storage and I/O system
- Closing thoughts

ICEE, Enabling International Collaborations

Example: KSTAR ECEI Sample Workflow (Electron cyclotron emission)

- **Objective**: To enable remote scientists to study ECE-Image movies of blobby turbulence and instabilities between experimental shots in near real-time.
- **Input**: Raw ECEi voltage data (~550MB/s, over 300 seconds in the future) + Metadata (experimental setting)
- **Requirement**: Data transfer, processing, and feedback **within <15min** (inter-shot time)
- **Implementation**: distributed data processing with ADIOS ICEE method

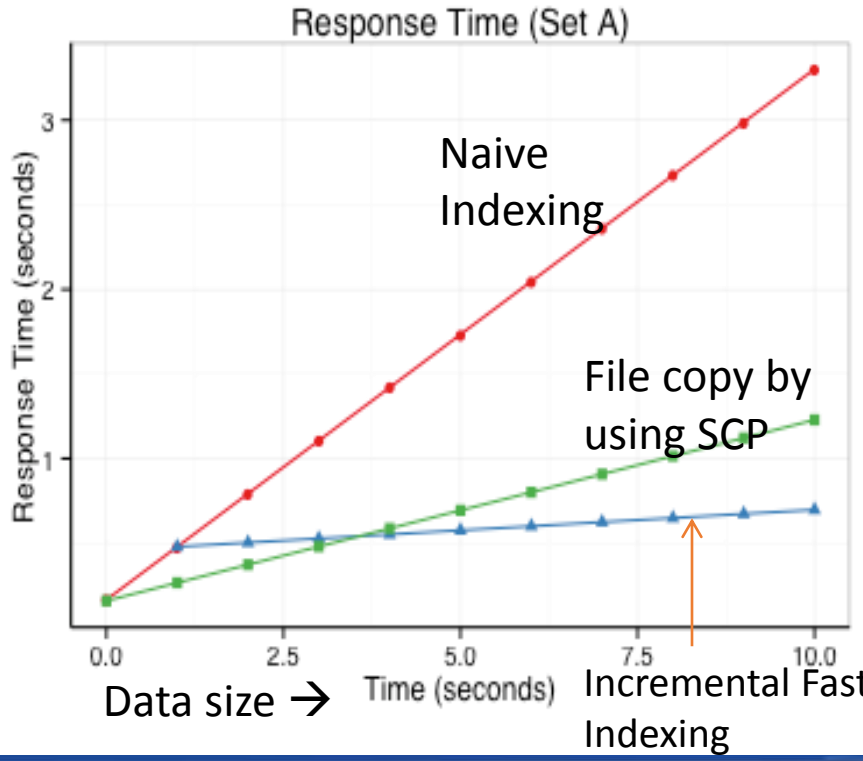


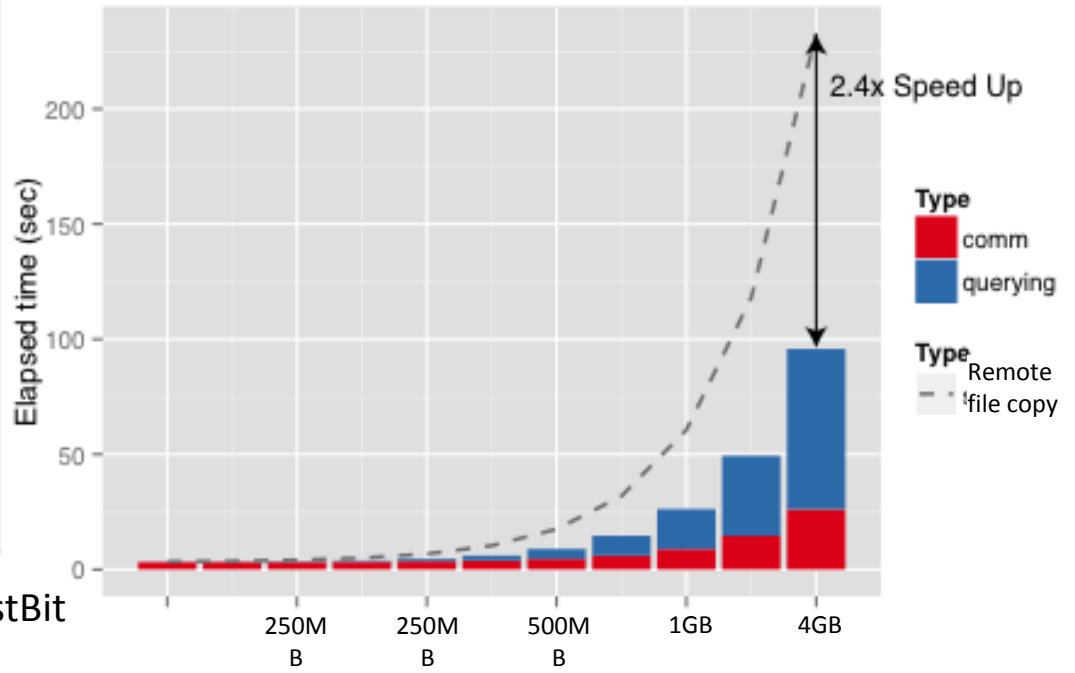


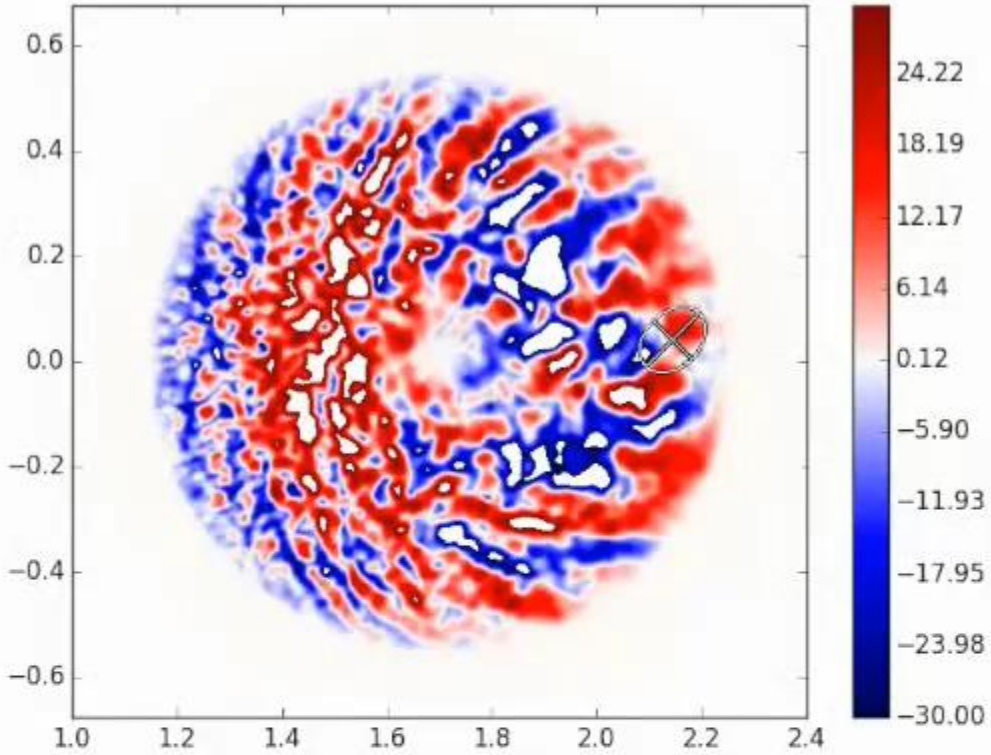
klasky@ornl.gov

Index-and-Query Reduces Execution Time

- Remote file copy VS. index-and-query
  - Measured between LBL and ORNL to simulate KSTAR-LBL-ORNL connection
  - Indexed by FastBit. Observed a linear performance (i.e., indexing cost increased by data size) ➔ Expensive indexing cost
  - However, once we have index built, index-and-query can be a better choice over remote file copy







R1 R2 Angle X Y REV TRACKING
0.0717066031626 0.0517881022841 45.0 2.14366197183 0.0422535211268 1.0 0

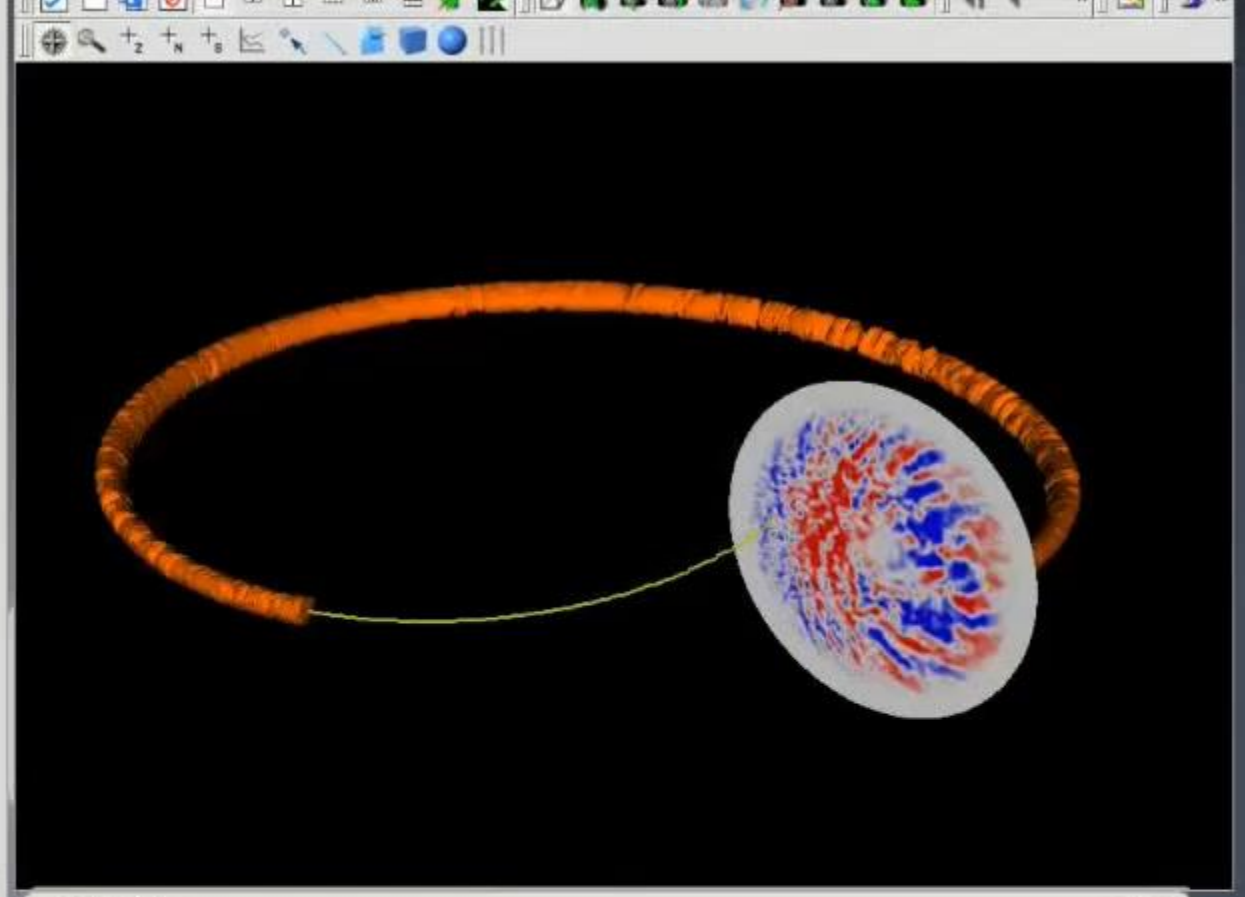
MultiPick

Accept Multi Select

Clear Selection

Number of Revolutions

ID Tracking: OFF



Sith-ADIOS

File Edit View Search Terminal Tabs Help

Sith-ADIOS × | sith-Visit × | sith × | focus × | rhea-visit-build ×

```
>>> Sleeping ... 5 (seconds)
>>> Reading ... totalf_itg_tiny/xgc.3d.08010.bp
>>> Writing ... xgc.3d.bp
>>> Written 84,764,672 bytes, Elapsed 0.066 seconds (Throughput: 1,216.485 MB/sec)
>>> Reading ... xgc-bbox.bp
>>> Reading ... totalf_itg_tiny/restart_dir/xgc.restart.08010.bp
>>> Reading ... totalf_itg_tiny/restart_dir/xgc.restart.08010.bp.info
>>> Read particle efficient ...
>>> Reading particles elapsed 2.191 seconds
>>> Writing ... xgc.particle.bp
>>> Written 7,251,168 bytes, Elapsed 0.050 seconds (Throughput: 139.048 MB/sec)
>>> Sleeping ... 5 (seconds)
>>> Reading ... totalf_itg_tiny/xgc.3d.08020.bp
>>> Writing ... xgc.3d.bp
>>> Written 84,764,672 bytes, Elapsed 0.049 seconds (Throughput: 1,637.854 MB/sec)
>>> Reading ... xgc-bbox.bp
>>> Waiting a newer version: xgc-bbox.bp
>>> Waiting a newer version: xgc-bbox.bp
>>> Reading ... totalf_itg_tiny/restart_dir/xgc.restart.08020.bp
>>> Reading ... totalf_itg_tiny/restart_dir/xgc.restart.08020.bp.info
```

Outline

- Motivation
- Our contributions
- Examples
- SIRIUS: Next generation Multi-tier storage and I/O system
- Closing thoughts

klasky@ornl.gov

Compute-Data Gap

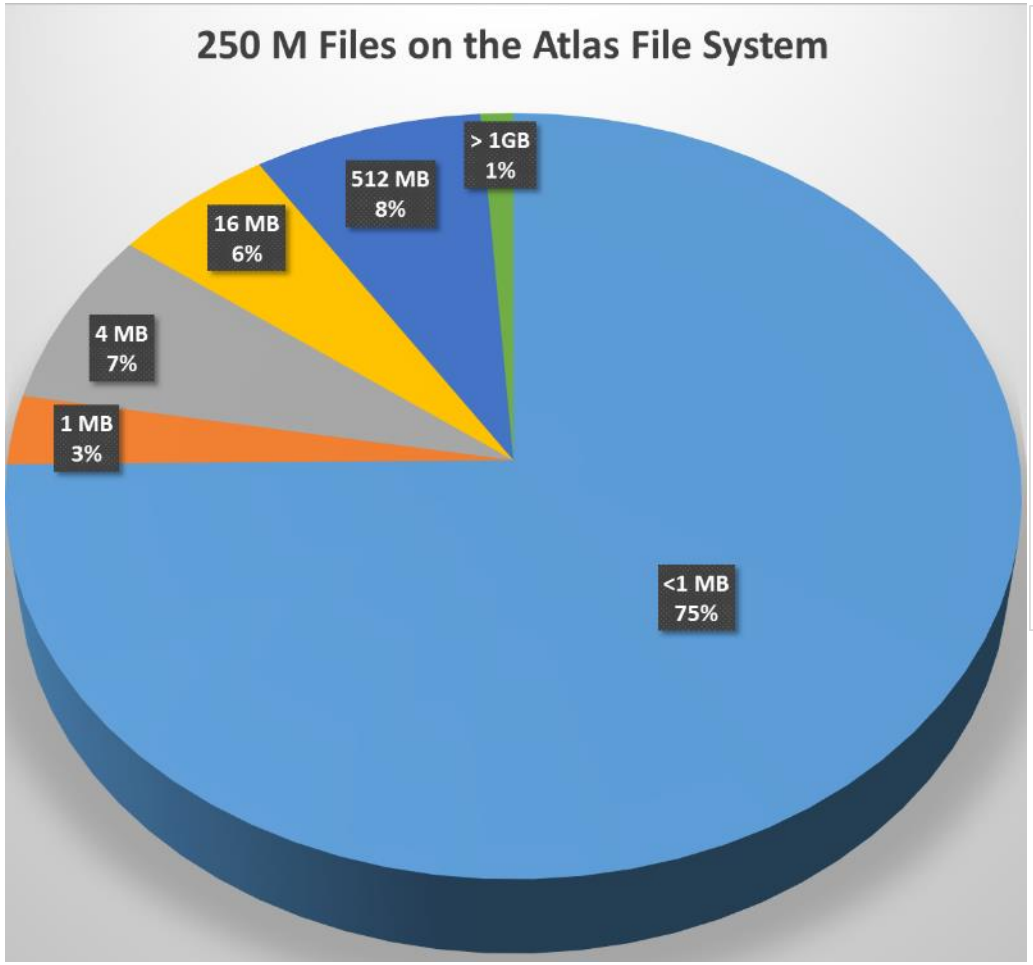
- Data storage and management will be limiting factor for exascale and beyond
- New research into utilizing computation to fill in data gap
- Scientific data can be modelled and refactored
  - Exploit structure to optimize data storage and processing
  - Split data into blocks with varying precision
  - Remember how data was originally created to regenerate on demand

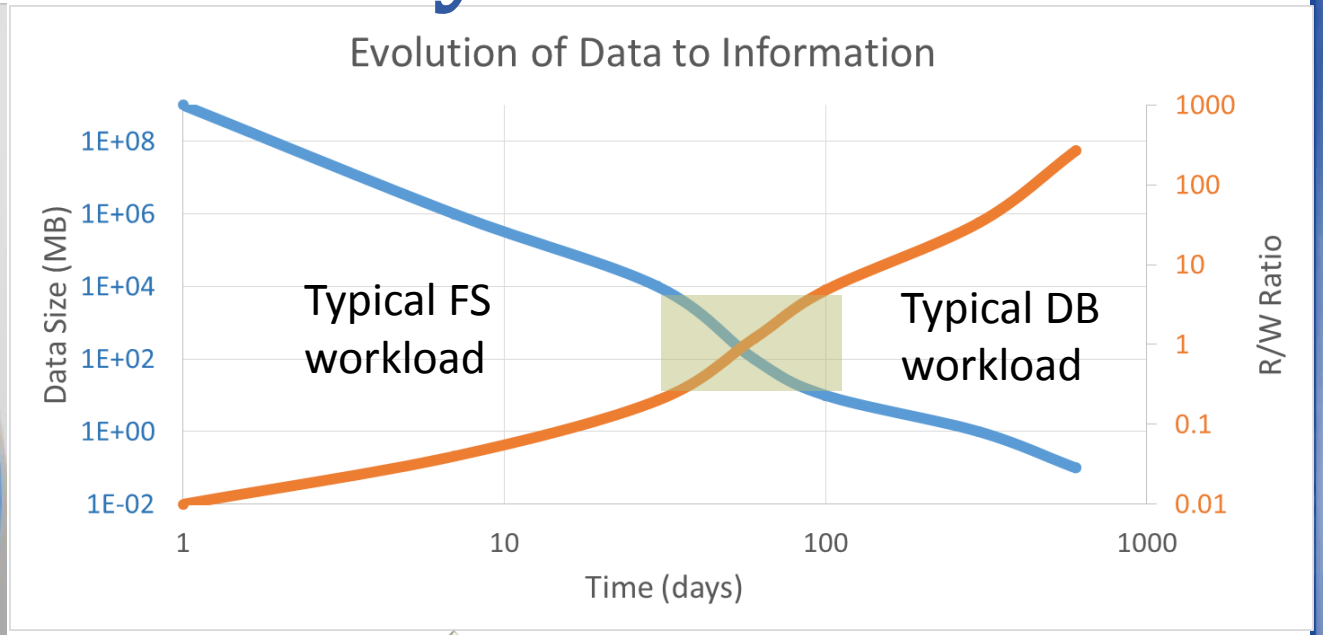
Next Generation doe computing

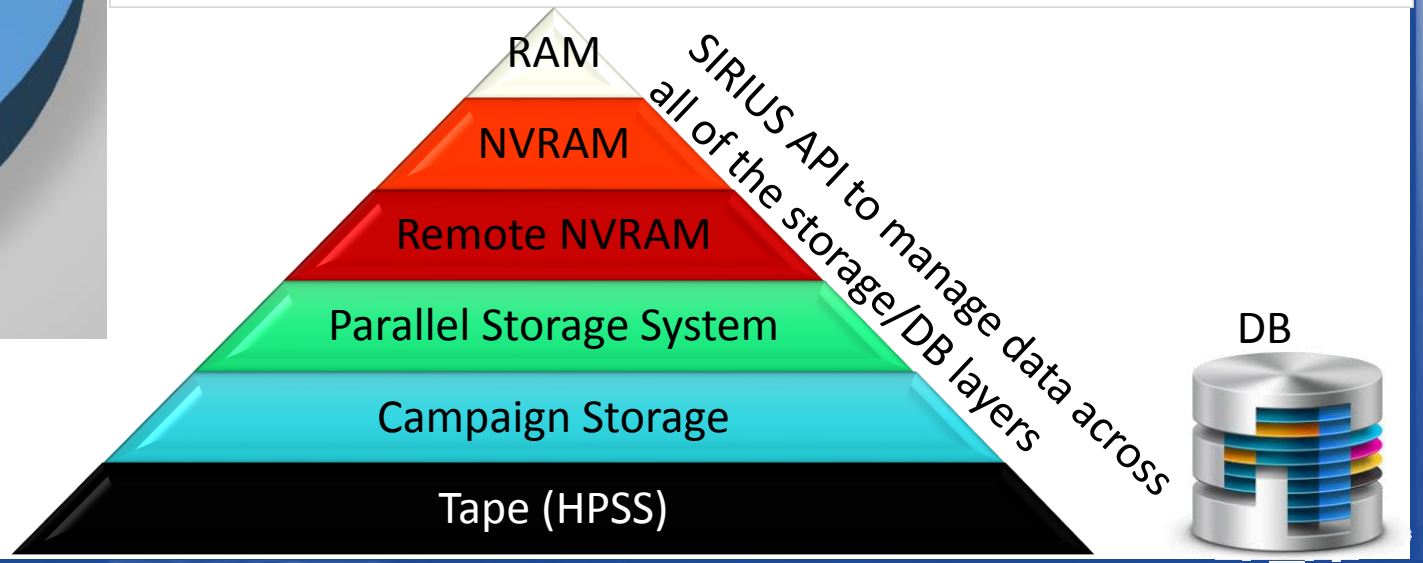
System attributes	NERSC Now	OLCF Now	ALCF Now	NERSC Upgrade	OLCF Upgrade	ALCF Upgrades	
Name Planned Installation	Edison	TITAN	MIRA	Cori 2016	Summit 2017-2018	Theta 2016	Aurora 2018-2019
System peak (PF)	2.6	27	10	> 30	150	>8.5	180
Peak Power (MW)	2	9	4.8	< 3.7	10	1.7	13
Total system memory	357 TB	710TB	768TB	~1 PB DDR4 + High Bandwidth Memory (HBM)+1.5PB persistent memory	> 1.74 PB DDR4 + HBM + 2.8 PB persistent memory	>480 TB DDR4 + High Bandwidth Memory (HBM)	> 7 PB High Bandwidth On-Package Memory Local Memory and Persistent Memory
Node performance (TF)	0.460	1.452	0.204	> 3	> 40	> 3	> 17 times Mira
Node processors	Intel Ivy Bridge	AMD Opteron Nvidia Kepler	64-bit PowerPC A2	Intel Knights Landing many core CPUs Intel Haswell CPU in data partition	Multiple IBM Power9 CPUs & multiple Nvidia Voltas GPUS	Intel Knights Landing Xeon Phi many core CPUs	Knights Hill Xeon Phi many core CPUs
System size (nodes)	5,600 nodes	18,688 nodes	49,152	9,300 nodes 1,900 nodes in data partition	~3,500 nodes	>2,500 nodes	>50,000 nodes
System Interconnect	Aries	Gemini	5D Torus	Aries	Dual Rail EDR-IB	Aries	2nd Generation Intel Omni-Path Architecture
File System	7.6 PB 168 GB/s, Lustre®	32 PB 1 TB/s, Lustre®	26 PB 300 GB/s GPFS™	28 PB 744 GB/s Lustre®	120 PB 1 TB/s GPFS™	10PB, 210 GB/s Lustre initial	150 PB 1 TB/s Lustre®

klasky@ornl.gov

Abstractions across File System to DB







klasky@ornl.gov

AUDITOR: New techniques for "Data Intensive Science"

AUDITOR: An additional "simulation" whose purpose is to monitor the fine scale simulation and initiate appropriate actions when anomalies are detected

Examples

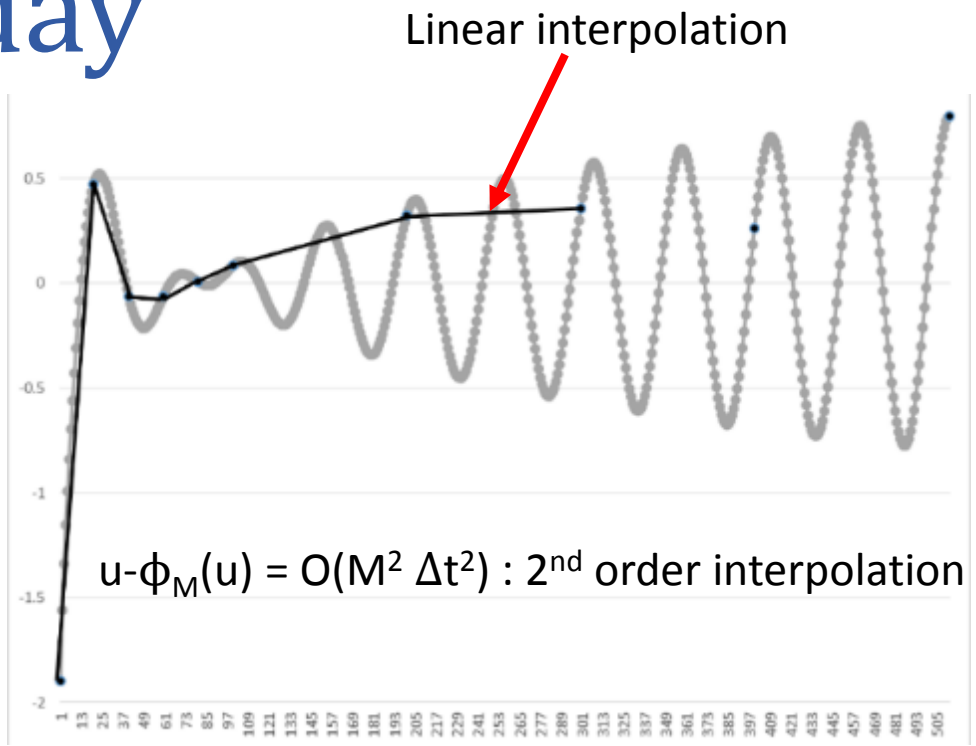
- Trigger a: checkpoint, roll-back, local change in a function, …
- Not confined to stability issues because it will always reset
- Can allow data regeneration cheaply

Basic quantities in Information Theory

- Data stream S and for $x \in S$ $\quad$ let $P_r(X=x) = p_x \in [0,1]$
- Shannon Information Content: $\quad h(x) = - \log_2 p_x$
- Entropy $\quad H(S) = - \Sigma\, p_x \log_2 p_x$
- **Noisy/random data has HIGH ENTROPY**

klasky@ornl.gov

Current practices of today

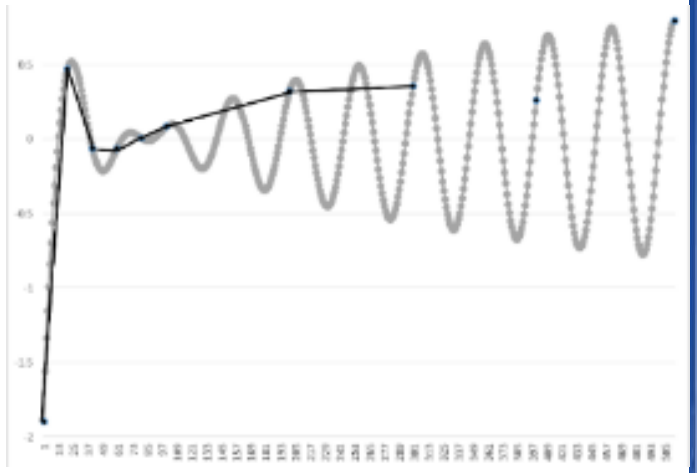


- Want to write data every n^{th} timestep
  - Because of the Storage and I/O requirements users are forced to writing less
- Common practice is to write data at every m^{th} timestep, stride = M
- If the users reconstruct their data, u(t), at the n^{th} timestep, they need to interpolate between the neighboring timesteps
  - $\Phi_M(u)$ = interpolant on coarser grid (stride M), reduce storage my 1/M
- Assume (C=constant depending on the complexity of the data)
  - Original storage cost = 32*N bits (floats)
  - New storage cost = 32*N/M bits + $\{ 23 - \log_2 (C\ M^2\ \Delta t^2)\}N$
  - Ratio = $(1/M - 1/16\ \log_2 M) - 1/16\ \log_2 \Delta t$ + constant

Cost to store ϕ_M +
Cost to store mantissa of u- $\phi_M(u)$

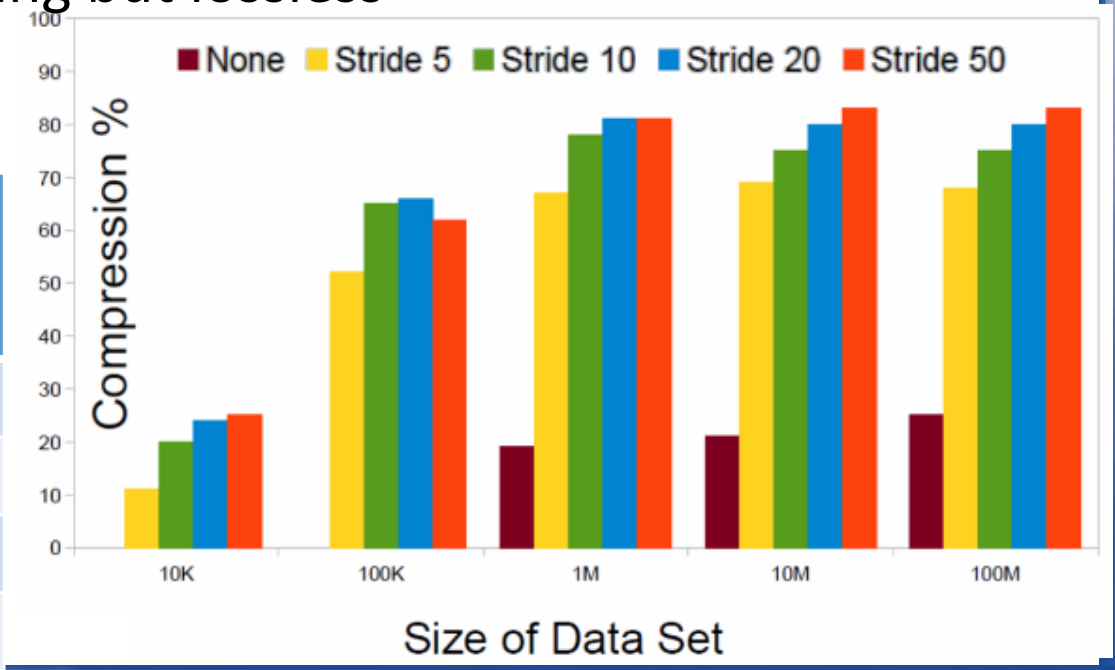
klasky@ornl.gov

Compression with an interpolation auditor



- Linear interpolation (LA) is the auditor
- If we look at 10MB output, with a stride of 5
  - Total output = 50MB for 5 steps
  - 10 MB, if we output 1 step, 43MB “typical lossless compression”, 18MB, using linear auditing but lossless
- Investigating adaptive techniques

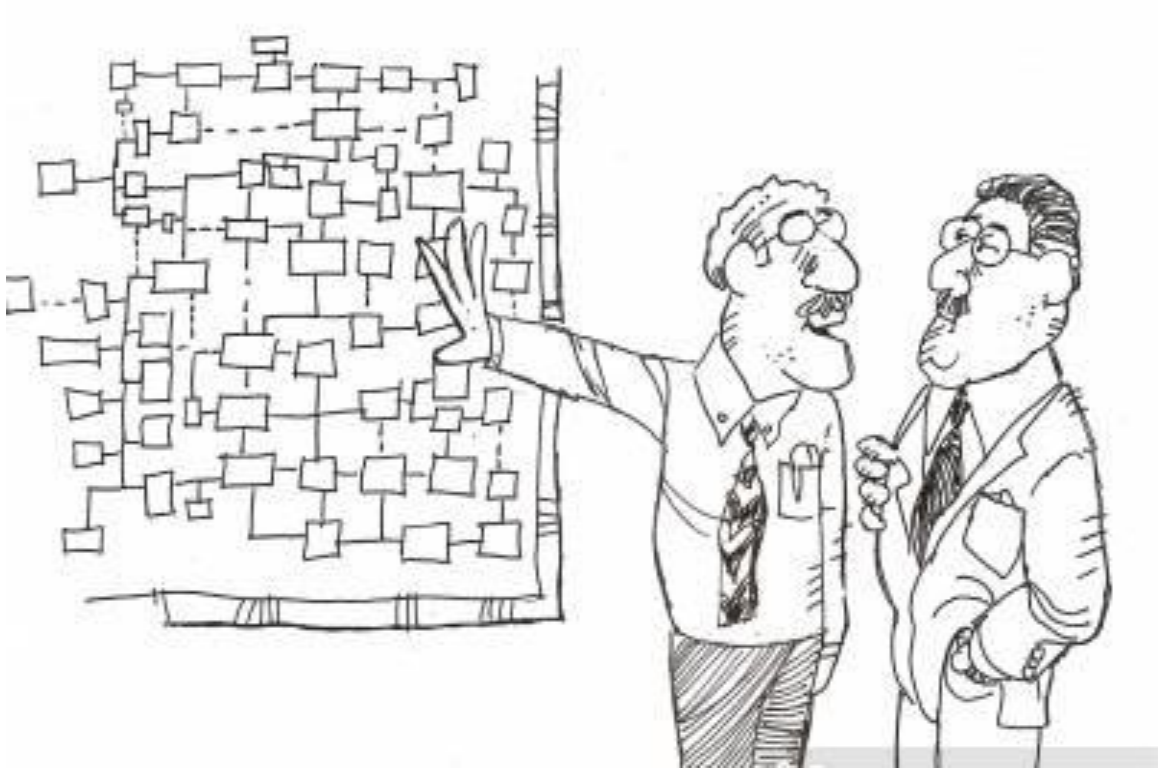
Stride	1 step (MB)	lossless compression (MB)	Linear Audit (MB)	Total Data in 50 steps, typical compression	**Total data in 50 steps in LA**
5	10	43	18	430	**180**
10	10	85	25	850	**125**
20	10	170	40	1700	**100**
50	10	**425**	100	4250	**100**



klasky@ornl.gov

Outline

- Motivation
- Our contributions
- Examples
- SIRIUS: Next generation Multi-tier storage and I/O system
- Closing thoughts



Lessons Learned

- Velocity
  - Critical to quickly build an index which can be done in a timely fashion
- Veracity
  - Understand the trade-offs for accuracy (of the query) vs. accuracy of the results vs. performance (time to solution).
- Volume
  - Reduce the volume of data being moved and processed over the WAN (size vs. accuracy)
- Variety
  - Enable multiple streams of data to be analyzed together
- Value
  - Provide the freedom for scientists to access and analyze their data interactively

Next steps

- Zetascale computing will usher a new age of computing
- Knowledge discovery in the validation process will become the overarching theme of scientific computing
  - Design of computation
  - Design of experiments
- Data Movement is the costly factor (you know this)
- Too many cores ….
  - Tradeoff between more cores and specialized accelerators
- Move from the “Big Data” age to the knowledge discover age
  - Move, process, only what’s necessary

Questions

klasky@ornl.gov