The data deluge: Overcoming the barriers to extreme scale science

Scott Klasky

SSDBM - Copenhagen 2022

Ana Gainaru, Qian Gong, Norbert Podhorszki, Dave Pugmire

ORNL

Adjunct Faculty in Computer Science: Georgia Tech

Thanks to the many of my collaborators

- Hasan Abbasi
- Mark Ainsworth
- Chuck Atkins
- Vicente Bolea
- Michael Bussmann
- CS. Chang
- Jieyang Chen
- Hank Childs
- Jong Choi
- Michael Churchill
- Philip Davis
- Ciprian Docan
- Greg Eisenhauer
- Stephane Ethier
- Ana Gainaru
- Dmitry Ganyushin
- Kai Germaschewski
- Berk Geveci
- William Godoy
- Qian Gong
- Junmin Gu
- Kevin Huck
- Axel Huebl
- Chen Jin
- Mark Kim
- Brad King
- James Kress
- S.H. Ku
- Ralph Kube
- Tahsin Kurc
- Xin Liang
- Zhihong Lin
- Qing Liu
- Jay Lofstead
- Jeremy Logan
- Kshitij Mehta
- Ken Moreland
- Todd Munson
- Manish Parashar
- Franz Pöschel
- Dave Pugmire
- Anand Rangarajan
- Sanjay Ranka
- Caitlin Ross
- Nagiza Samatova
- Karsten Schwan
- Ari Shoshani
- Eric Suchyta
- Fred Suter
- Keichi Takahashi
- William Tang
- Roselyne Tchoua
- Nick Thompson
- Seiji Tsutsumi
- Ozan Tugluk
- Lipeng Wan
- Ruonan Wang
- Ben Whitney
- Matthew Wolf
- Kesheng Wu
- Bing Xie
- Fan Zhang
- Fang Zheng

klasky@ornl.gov, scottk007@gmail.com

We are looking for motivated scientists and software engineers

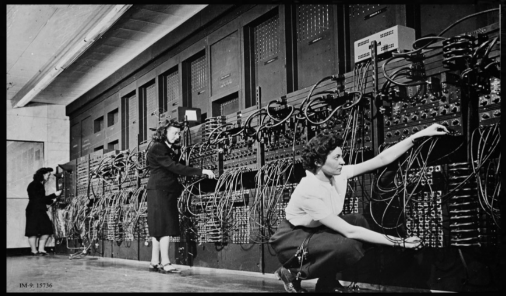
- Postdoctoral Research Associate - Computer Science/Applied Mathematics
  - https://jobs.ornl.gov/job-invite/8517/
- Scientific Software Engineer
  - https://jobs.ornl.gov/job/Oak-Ridge-Scientific-Software-Engineer-TN-37831/883774400/
- Computer Scientist
  - https://jobs.ornl.gov/job/Oak-Ridge-Computer-Scientist-TN-37830/883927400/

klasky@ornl.gov, scottk007@gmail.com

The data deluge

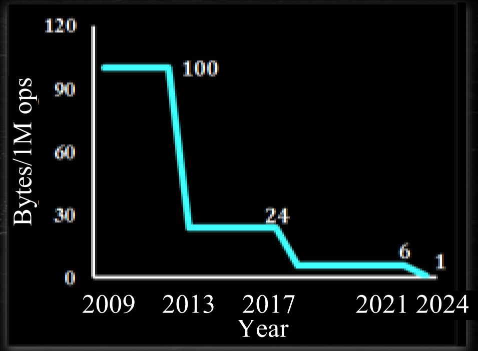
- Computing capability has increased by a factor of 10^{15} over the past 70 years
- Applications push the boundaries of data: e.g., Radio Astronomy
  - In 2022 the Vera C. Rubin Observatory in Chile will collect 20 terabytes/night as part of the Legacy Survey of Space and Time (LSST)
  - In 2028 the Square Kilometre Array, will generate 100 times that amount
- Filesystem/network bandwidth falls behind CPU/memory: Fewer bytes/operation
- Our goal is to create the proper abstractions & frameworks to cope with this deluge

1943 , 5K Flops, Electronic Numerical Integrator And Computer – no filesystem



2022 , 1.6 Ex Flops, Frontier, LUSTRE filesystem

Filesystems continue to fall behind computing power



OAK RIDGE
National Laboratory

klasky@ornl.gov, scottk007@gmail.com

Outline

R&D
Analytics
Visualization

Understand simulations, experiments and observations

Understanding new technologies

R&D in Scientific Data Management

R&D Staging & Coupling

R&D Data Refactoring

Application Impact:
High Energy Physics, Accelerator Physics, Cancer, Smart Medicine
, Climate , Combustion, Hypersonics, Rocket Science, General Relativity, Molecular Dynamics, Chemistry, Radio Astronomy, Fusion , Seismology, Wind Turbine

OAK RIDGE
National Laboratory

Supercomputing changes in the last 30 years

1988: Cray Y-MP:0.0000027 Pflops: Vector Processors, SSD for storage (13.6 GB/s)

1996:Cray T3E: 0.001 PFlops, massively parallel

1998: 0.0025 Pflops,ASCI Blue Mountain: shared memory across all procs

2002: Earth Simulator 0.040 Tflops, 50MW, vector procs

2009: Cray XT5: 2.5 Pflops: Multi-core, LUSTRE storage system

2013: Cray Titan: 27 Pflops , NVIDIA GPUs

2018: IBM Summit: 200 Pflops , NVIDIA GPUs

2022: Cray Frontier: 2000 Pflops , AMD GPUs, Burst Buffer Storage 10TB/s, long term at 4.6 TB/s

Observations:

- Ratio of Storage/Flops keeps getting worse
- I/O variability gets worse as systems scale

- New applications are running complex workflows with AI + HPC applications
- Experimental/Observational data is outpacing compute & storage

OAK RIDGE National Laboratory Foster, I., Klasky, S., et al., (2017, August). Computing just what you need: Online data analysis and reduction at extreme scales. In European conference on parallel processing (pp. 3-19). Springer, Cham.

klasky@ornl.gov, scottk007@gmail.com

Frontier: First Exascale computer on the top500 list

- System URL: https://www.olcf.ornl.gov/frontier/
- Manufacturer: HPE
- Cores: 8,730,112
- Processor: AMD 3rd Gen EPYC 64C 2GHz
- Interconnect: Slingshot-11
- Installation Year: 2021
- 9,472 nodes, 4 GPUs/node

Performance

- Linpack Performance (Rmax) 1,102.00 PFlop/s
- Theoretical Peak (Rpeak) 1,685.65 PFlop/s
- Nmax 24,440,832
- Power: 21,100.00 kW
- Operating System: HPE Cray OS

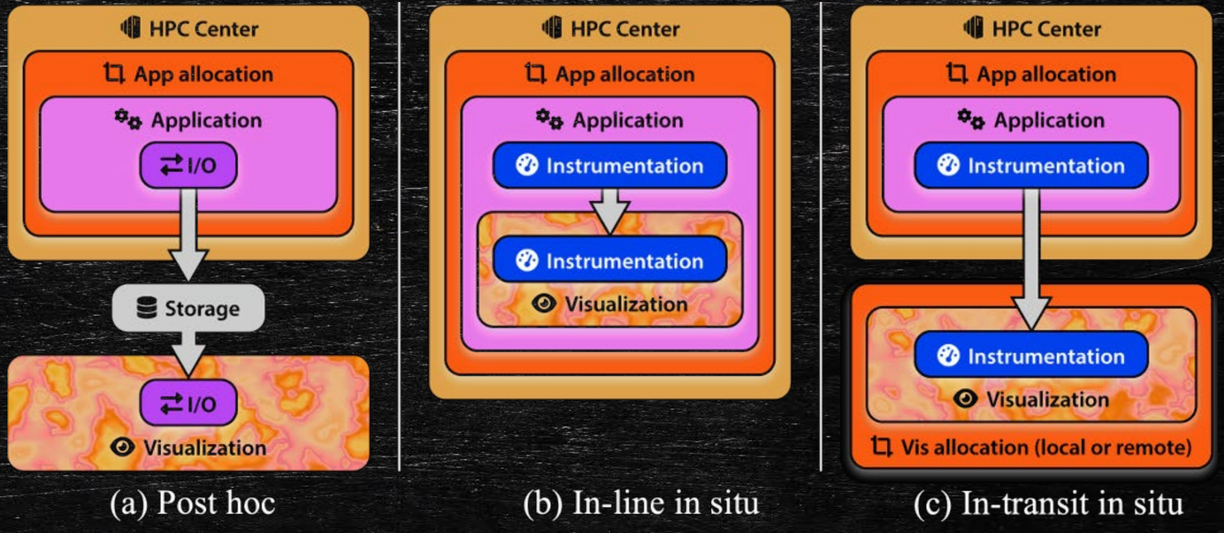
Storage

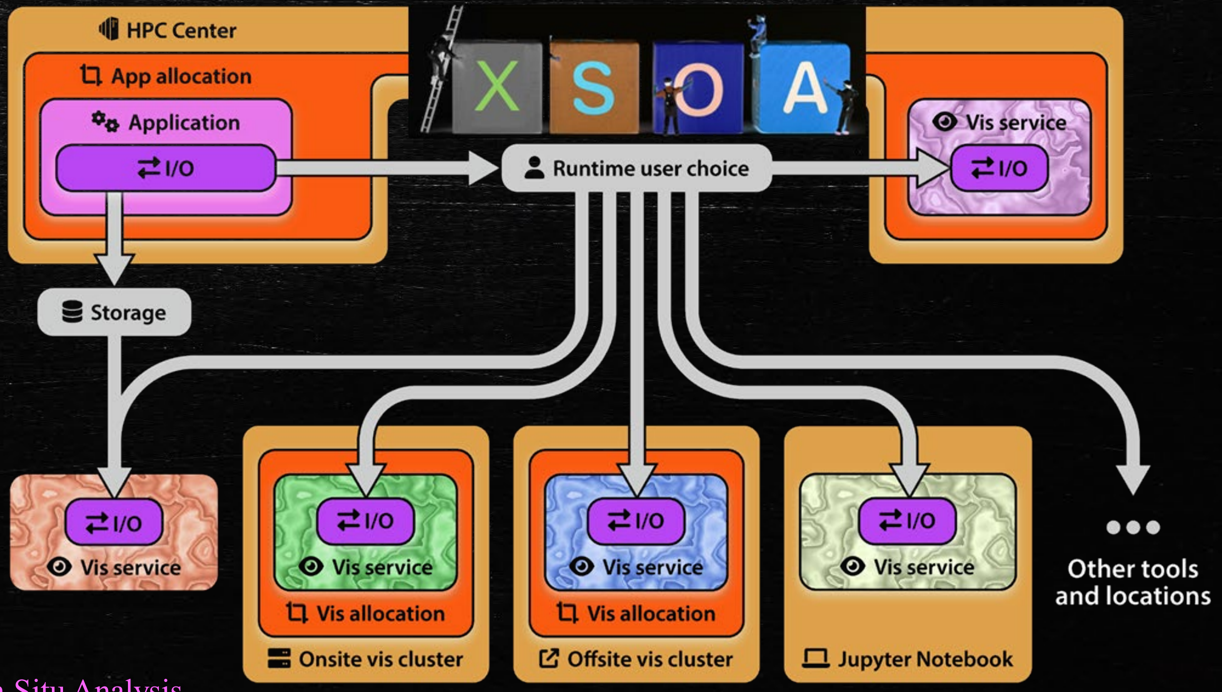
- 37 PB of node local NVMe, 716 PB of center-wide storage

klasky@ornl.gov, scottk007@gmail.com

Our vision: creating a pub/sub system for high performance SDM

- Currently applications are typically programmed to
  - Write/read to storage
  - Perform in-line visualization or in-transit
- Our vision is to allows applications to publish and subscribe to data
  - With no modifications, any code can tap into the I/O system
  - Data can stream in a “refactored” manner allowing the “most important” information to be prioritized in the streams
  - Data written and read to storage will be highly optimized on HPC resources and queriable





OAK RIDGE National Laboratory D. Pugmire, J. Huang, S. Klasky, and K. Moreland: The Need for Pervasive In Situ Analysis and Visualization (P-ISAV), WOIV’22.

klasky@ornl.gov, scottk007@gmail.com

eXtreme Scale Service Oriented Architecture (XSOA)

- Philosophy based on Service-Oriented Architecture
  - Deal with system/application complexity, rapidly changing requirements, evolving target platforms, and diverse teams
- Applications constructed by assembling services based on a universal view of their functionality using an API
- Implementations can be changed and assembled easily
- Manage complexity while maintaining performance, scalability
  - Complexity from the problem (complex physics) and the codes
  - Complexity of underlying disruptive infrastructure
  - Complexity from coordination across codes and research teams
  - Complexity of the end-to-end workflows

S. Klasky et al., "A View from ORNL: Scientific Data Research Opportunities in the Big Data Age," 2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS), 2018, pp. 1357-1368, doi: 10.1109/ICDCS.2018.00136

klasky@ornl.gov, scottk007@gmail.com

Think about Storage, I/O, analysis, visualization in a new way

- Design abstractions to work with data at rest and in motion
- Understand the differences between data and information
- Create new mathematical frameworks to create a hierarchy of information from the data
  - Similar to Adaptive Mesh Refinement, Multigrid techniques for Partial Differential Equations
  - Similar to how we deal with images, movies, but have well defined error bounds
- Create new analytics and visualization to take advantage of thew new abstractions and frameworks
- Record provenance to aid in the data lifecycle
- Use automation to aid in the process

klasky@ornl.gov, scottk007@gmail.com

Moving towards new physics

Homoclinic tangle is produced by microturbulence and breaks the last confinement surface in tokamak, and was first discovered while running a coupled workflow using ADIOS-2

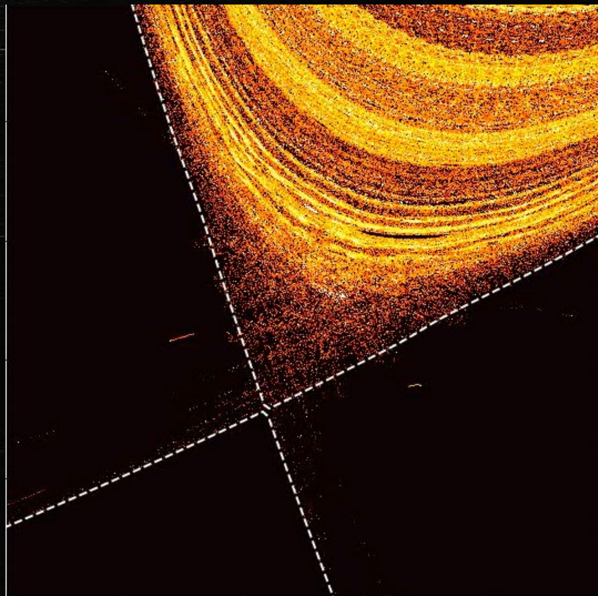
Science

- Tokamak plasma is designed to have the last magnetic confinement surface, called separatrix surface possessing a hyperbolic fixed point: X-point
- It has been long known that homoclinic tangle can exist when there is 3D perturbation δB in the magnetic field [Poincare, 1881], disturbing or destroying the last confinement surface if δB is large enough.

Discovery

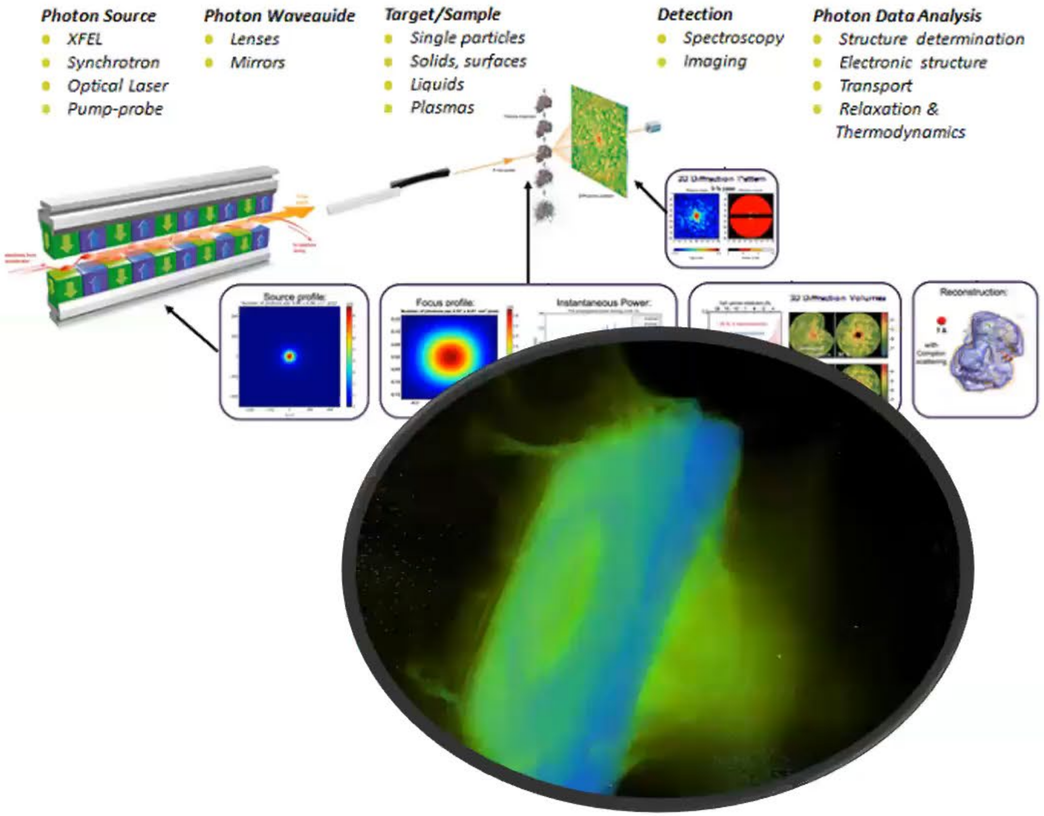
- Gyrokinetic simulation in XGC discovers that the intrinsic electromagnetic microturbulence in a stationary operation condition generates homoclinic tangle and destroys the last confinement surface around the X-point
- This discovery opens up new research topics:
  - A new escape route for confined plasma to open region
  - Non-local physics interactions between edge pedestal and divertor plasmas
  - Spreading the divertor heat-load footprint in fusion reactors, such as ITER

Fluctuating homoclinic tangles in full-current ITER edge, predicted by XGC



In-memory coupling and online analysis on Top 10 HPC systems

Coupling several codes for a full digital twin of HIBEF experiments



- Use openPMD API with ADIOS2
- Add interactive, in-memory analysis
- Enable interactive simulation steering

Run full digital twin of planned experiment on Top 10 HPC

Outline

R&D
Analytics
Visualization

Understand
simulations,
experiments and
observations

Understanding
new technologies

R&D in Scientific Data
Management

R&D Staging
& Coupling

R&D Data
Refactoring

Application Impact:
Celeritas, DeepdriveMD, E3SM, GE, GENE, GEM, GTC, JAXA, KSTAR, LAMMPS, NNESH, PIConGPU, S3D, SKA, SPECFEM3D_GLOBE, TAE, XGC, WarpX,WRF

ADIOS: high-performance publisher/subscriber I/O framework

Vision

- Create an easy-to-use, high performance I/O abstraction to allow for on-line/off-line memory/file data subscription service
- Create a sustainable solution to work with multi-tier storage and memory systems

Research Details

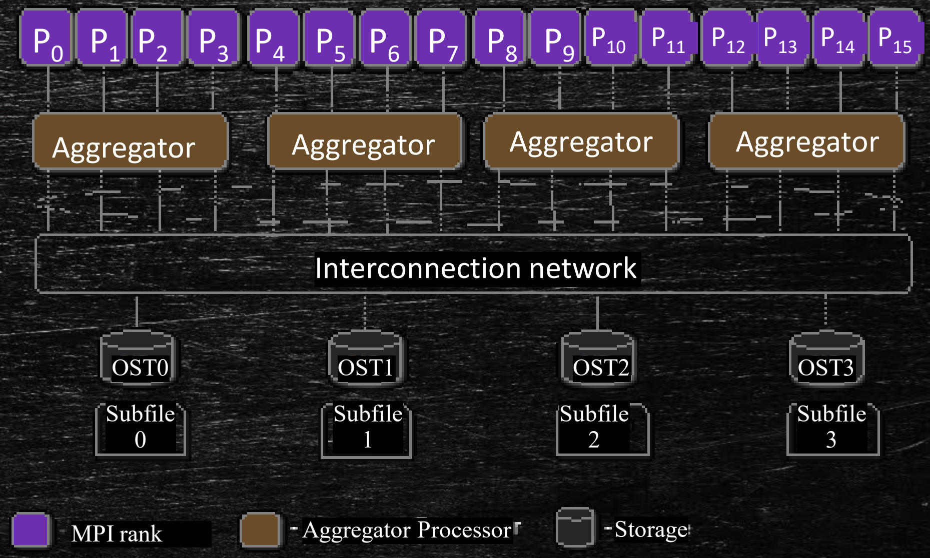
- Declarative, publish/subscribe API is separated from the I/O strategy
- Multiple implementations (engines) provide functionality and performance
- Rigorous testing ensures portability
- Data reduction techniques are incorporated to decrease storage cost
- https://github.com/ornladios/ADIOS2

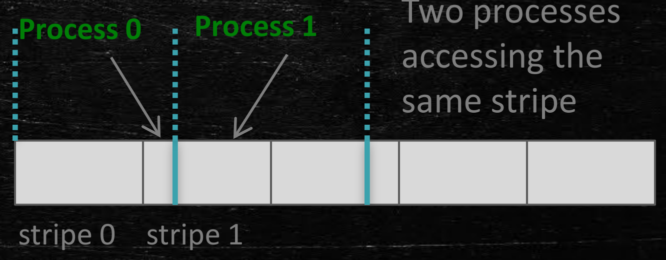
OAK RIDGE National Laboratory Godoy, W. F., Klasky, S. ,et al. (2020). ADIOS-2: The adaptable input output system. a framework for high-performance data management. SoftwareX, 12, 100561.

klasky@ornl.gov, scottk007@gmail.com

Optimizations for a parallel file system

- Avoid latency (of small writes): Buffer data for large bursts – use a type of self-describing log file format
- Avoid accessing a file system target from many processes at once
  - Aggregate to a small number of actual writers:
  - Avoid lock contention
  - Striping correctly & writing to subfiles
- Avoid global communication
- Topology-aware data movement that takes advantage of topology
  - Find the closest I/O node to each writer
  - Minimize data movement across racks/mid-planes





Application	Nodes/GPUs	Data Size/step	I/O speed
SPECFEM3D	3200/19200	250 TB	~2 TB/sec
GTC	512/3072	2.6 TB	~2 TB/sec
XGC	512/3072	64 TB	1.2 TB/sec
LAMMPS	512/3072	457 GB	1 TB/sec

OAK RIDGE National Laboratory

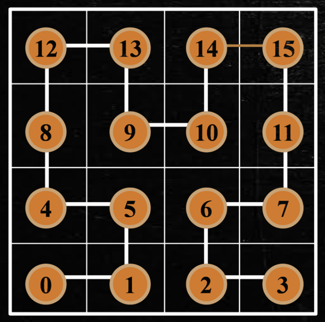
Liu, Q., Klasky, S., et al. "Hello ADIOS: the challenges and lessons of developing leadership class I/O frameworks." Concurrency and Computation: Practice and Experience 26.7 (2014): 1453-1473.

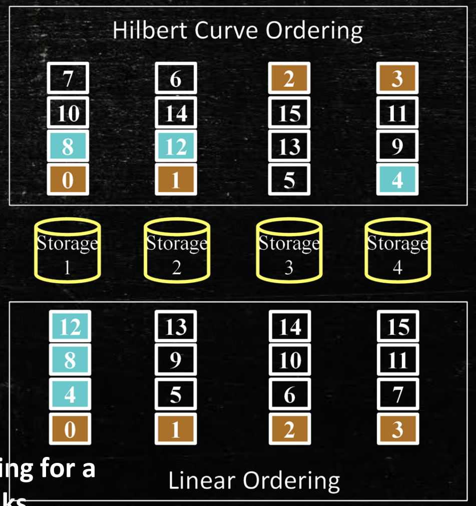
klasky@ornl.gov, scottk007@gmail.com

Space filling curve reordering for optimizing reading performance

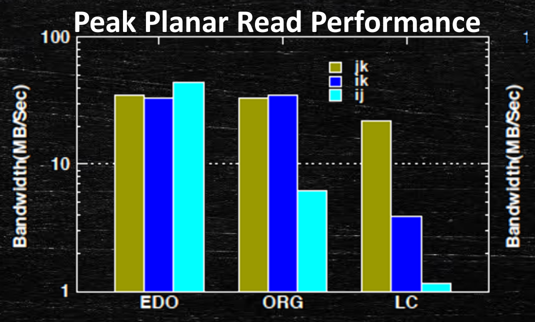
- Linear placement of data leads to hotspot on storage nodes
  - Can't leverage aggregated bandwidth, poor scalability
- Distribute data chunks on storage targets along the Hilbert curve ordering
  - Does not change the data organization within each chunk
  - Achieving near-optimal concurrency for any access pattern

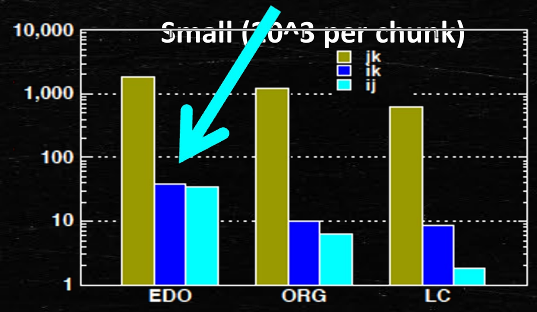
First Place, ACM Student Research Competition Grand Finals 2012



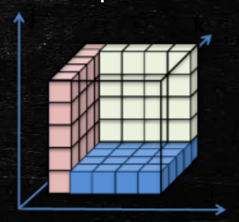


Hilbert Curve Reordering for a 2D Array with 16 chunks





Extra-large (250^3 per chunk)



- Consistently good and balanced read performance
- Up to **37X** speedup on Jaguar for S3D

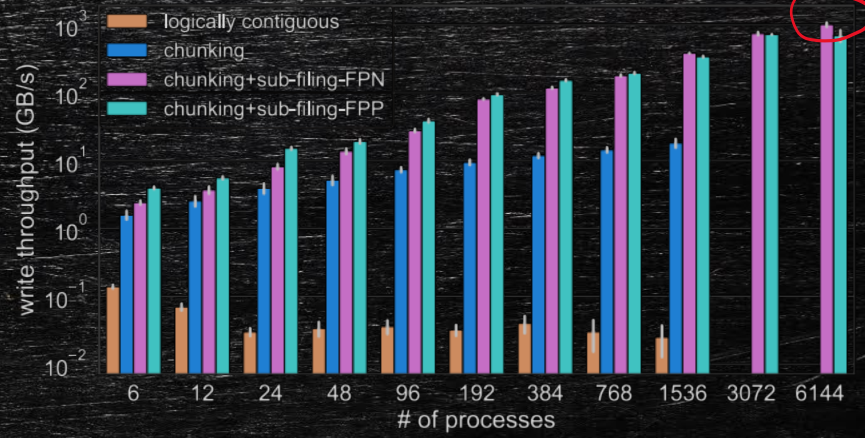
Tian, Y., Klasky, S., et al.(2011, September). EDO: improving read performance for scientific applications through elastic data organization. In 2011 IEEE International Conference on Cluster Computing (pp. 93-102).

klasky@ornl.gov, scottk007@gmail.com

Case study with the WarpX code

- How do we balance the write vs. read cost of a large scale HPC application such as WarpX?
  - Typical choices are to write to a logically contiguous file or chunk versions of this (e.g., HDF5) or to write separate chunks in many files (e.g., ADIOS-2) using one file per process (FPP) or one file per node (FPN)
- The challenge can be in optimizing reading
  - What is the most optimal organization?

WarpX write performance on Summit: weak scaling



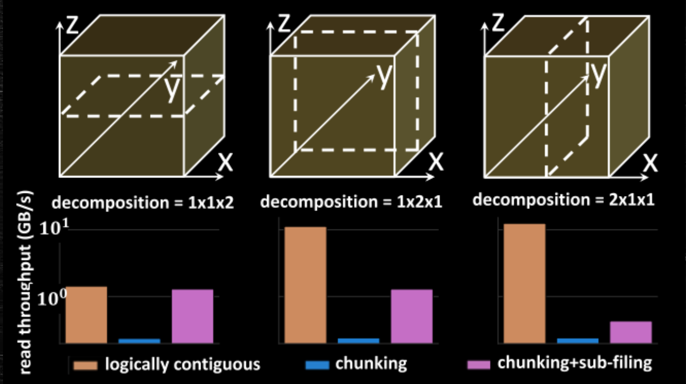


Fig. 5. Impact of decomposition schemes when reading.

OAK RIDGE National Laboratory

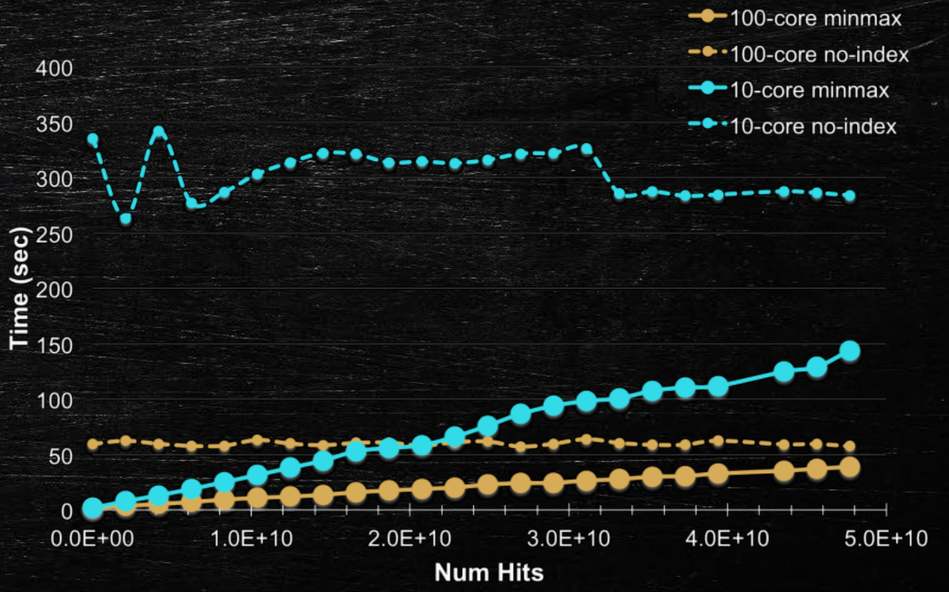
Wan, L., Huebl, A., Gu, J., Poeschel, F., Gainaru, A., Wang, R., ... & Klasky, S. (2021). Improving I/O performance for exascale applications through online data layout reorganization. IEEE Transactions on Parallel and Distributed Systems, 33(4), 878-890.

klasky@ornl.gov, scottk007@gmail.com

Querying Large Scientific Data Sets

Metadata location, attributes, min, max, … Variables Scalars, Arrays	Metadata location, attributes, min, max, … Variables Scalars, Arrays	Metadata location, attributes, min, max, … Variables Scalars, Arrays	Metadata location, attributes, min, max, … Variables Scalars, Arrays
Metadata location, attributes, min, max, … Variables Scalars, Arrays	Metadata location, attributes, min, max, … Variables Scalars, Arrays	Metadata location, attributes, min, max, … Variables Scalars, Arrays	Metadata location, attributes, min, max, … Variables Scalars, Arrays

- ADIOS writes metadata for each variable on each “chunk” of data
  - Currently it contains min/max, but it can include variance, mean, …
- Queries can contain 3 parts
  1. The selection box to limit the points considered
  2. The query conditions - in a form of query predicates connected with AND/OR operators
  3. The query output



- The chunks which satisfy the queries are returned to ADIOS and then the resultant data is given to the application.
  - This allows analysis tools to quickly query the data with no additional cost for storing indices
- Research is in optimizing the merge/split of chunks for Write/Read performance

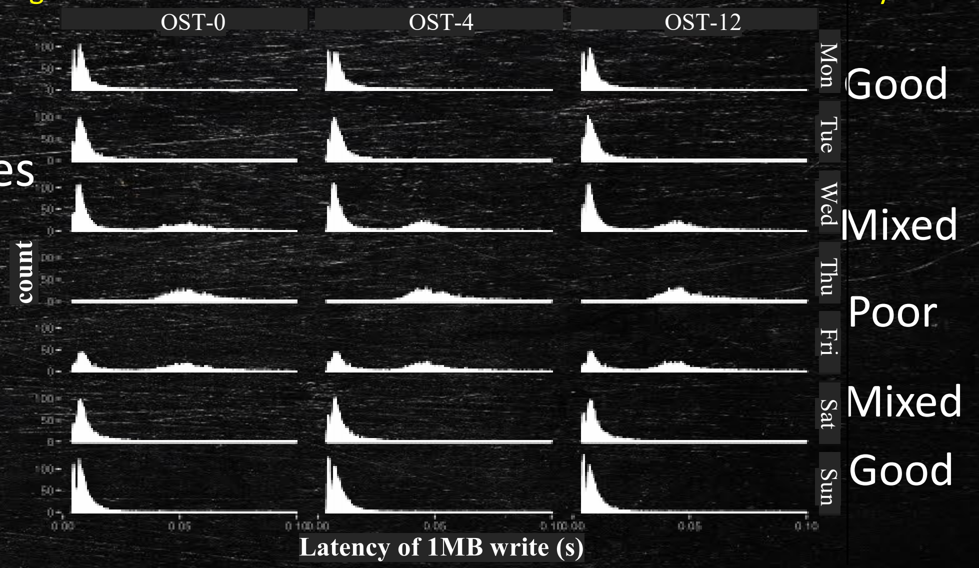
Gu, J., Klasky, S., Podhorszki, N., Qiang, J., & Wu, K. (2018, March). Querying large scientific data sets with adaptable IO system ADIOS. In Asian Conference on Supercomputing Frontiers (pp. 51-69). Springer, Cham.
klasky@ornl.gov, scottk007@gmail.com

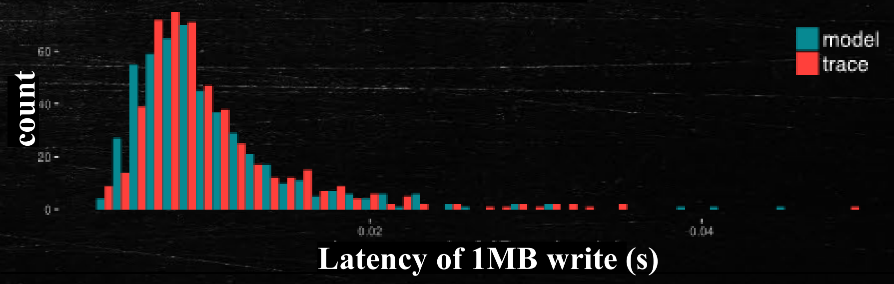
Managing I/O variability for applications on HPC systems (Titan)

- Scientific Achievement
  - Created a hidden Markov model of the I/O performance
  - Based on observed properties of the distribution of I/O latencies
  - Used to characterize and predict the I/O performance of applications
- Significance and Impact
  - Being integrated into ADIOS so that ADIOS can leverage these results to guide the data placement
- Research Details
  - Conduct I/O tests and collect time-dependent I/O traces for seven consecutive days
  - Build a hidden Markov model based on the statistical properties observed in the I/O traces
  - Validate the model by comparing the distribution of the predicted
    I/O latencies against the distribution of the real latency

- Xie, B., Klasky, S., et al. Characterizing output bottlenecks in a supercomputer. In SC'12: (pp. 1-11). Best student paper nominee.

Histograms of latencies of 1MB writes to different OSTs on different days



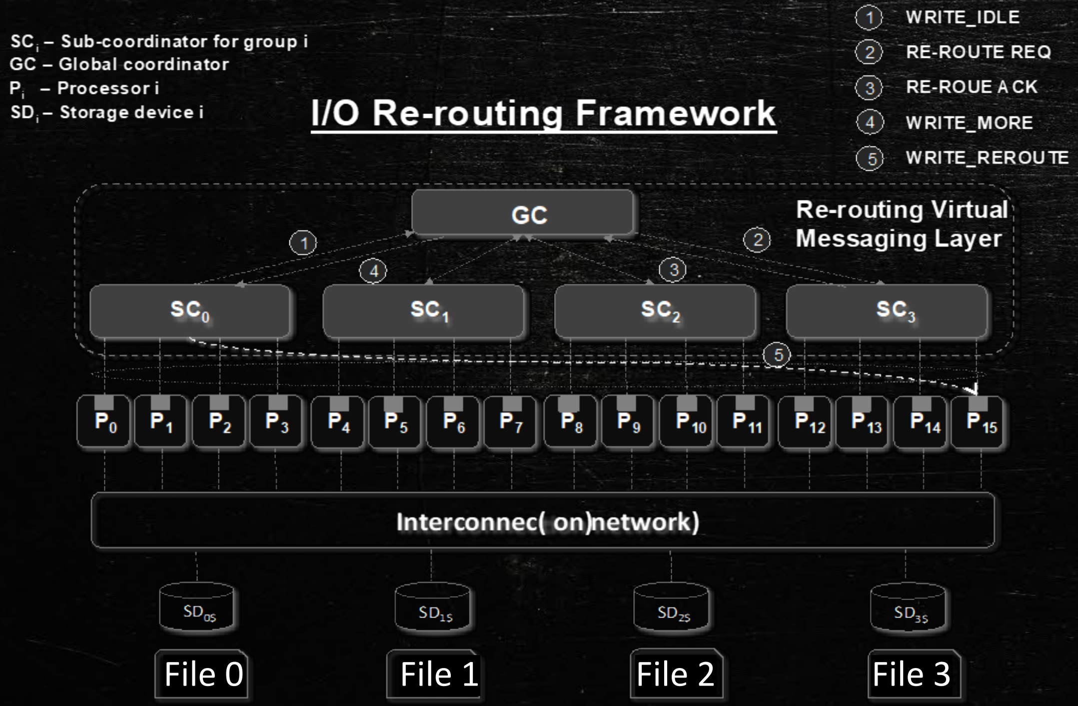


Wan, L., Klasky, S., et al. Comprehensive measurement and analysis of the user-perceived I/O performance in a production leadership-class storage system. In 2017 IEEE 37th (ICDCS) (pp. 1022-1031). Best paper finalist.

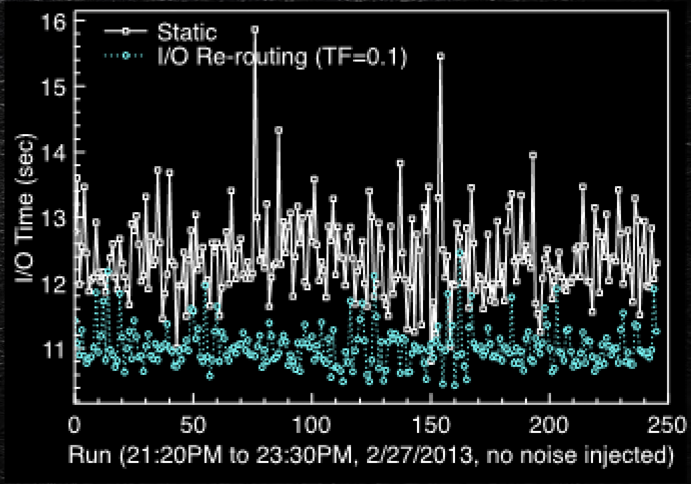
OAK RIDGE National Laboratory
klasky@ornl.gov, scottk007@gmail.com

I/O Variability

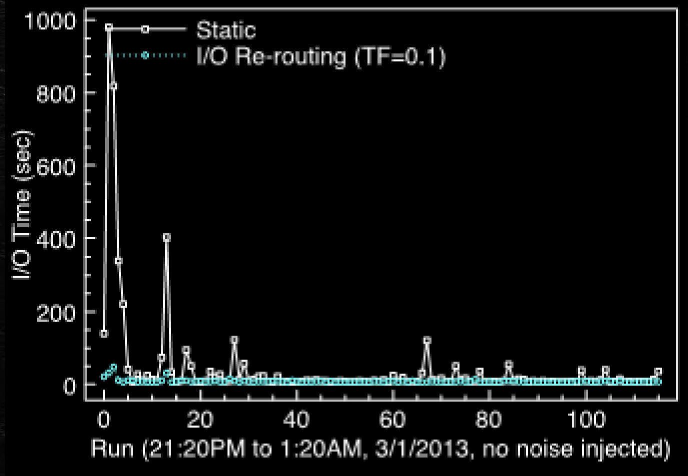
- Developed a runtime system that use short messages to distribute storage state and direct I/O re-routing
- The system consider both write and read performance, by limit the degree of re-routing, and is scalable using a hierarchical scheduler



Write Performance



Titan



Hopper

OAK RIDGE National Laboratory

Liu, Q., Klasky, S., et al. Runtime {I/O} {Re-Routing}+ Throttling on {HPC} Storage. In 5th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage 13).

klasky@ornl.gov, scottk007@gmail.com

Outline

R&D
Analytics
Visualization

Understand
simulations,
experiments and
observations

Understanding
new technologies

R&D in Scientific Data
Management

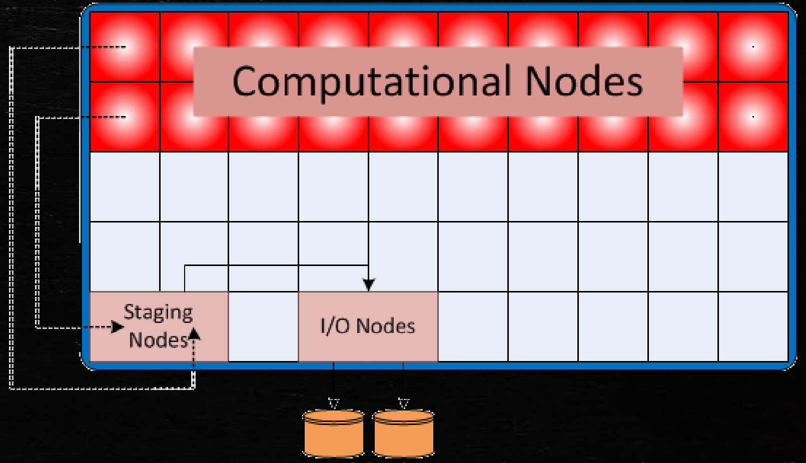
R&D Staging
& Coupling

R&D Data
Refactoring

Application Impact:
Celeritas, DeepdriveMD, E3SM, GE, GENE, GEM, GTC, JAXA, KSTAR, LAMMPS, NNESH, PIConGPU, S3D, SKA, SPECFEM3D_GLOBE, TAE, XGC, WarpX,WRF

Introduction to staging

- Simplistic approach to staging
  - Decouple application performance from storage performance (burst buffer)
- Built on past work with threaded buffered I/O
  - Buffered asynchronous data movement with a single memory copy for networks which support RDMA, TCP, RUDP, or MPI
  - Application blocks for a very short time to copy data to outbound buffer
  - Data is moved asynchronously using server-directed remote reads
- Exploits network hardware for fast data transfer to remote memory
- Value added
  - Allows scientists to use a data-in-transit technique to write, reduce, analyze data
  - Can be used to couple multiple codes together
  - Can be used for asynchronous I/O



Abbasi, H., Klasky, S., et al., (2010). Datastager: scalable data staging services for petascale applications. Cluster Computing, 13(3), 277-290.
klasky@ornl.gov, scottk007@gmail.com

Staging Options

Transfer mechanisms

- File based
- Network based on the same resource
  - MPI communication
  - RDMA (libfabric, UCX)
  - MPI (one sided, two sided)
  - TCP/ RUDP
- Memory references
- WAN data transfer
  - Files – GridFTP, scp, …
  - Streams – TCP, RUDP, RoCE

Placement options

- Same core
- Different cores/same node
- Different nodes
- Different resource (LAN)
- Different resource (WAN)
- Hybrid (mixture of options)

Scheduling options

- Fully synchronous
- Fully asynchronous
- Hybrid

Refactoring options

- Prioritize which data gets moved first

Storage options

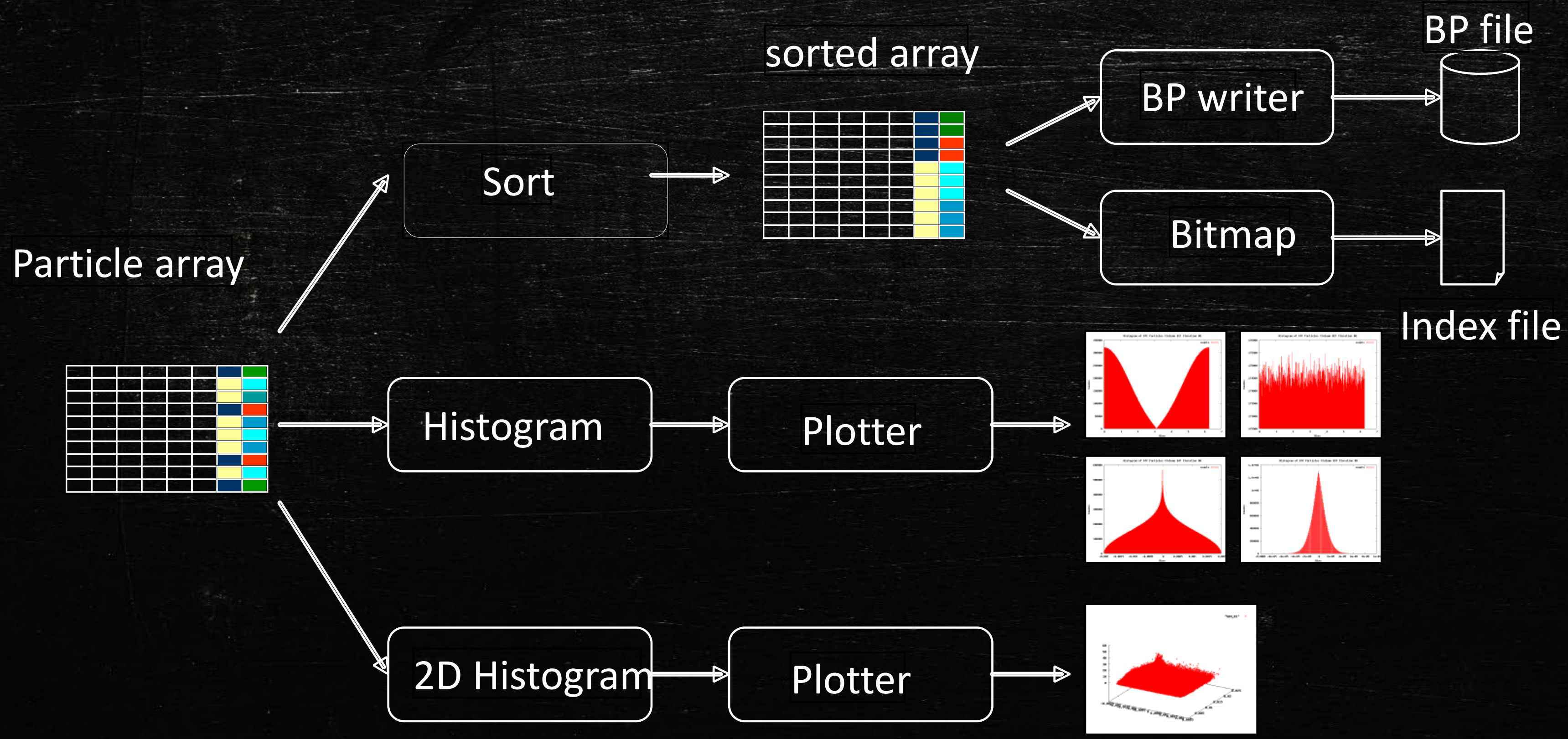
- HDF5
- ADIOS-BP5, …

Gainaru, A., Klasky, S., et al., 2022, Understanding the Impact of Data Staging for Coupled Scientific Workflows , IEEE transactions on parallel and distributed systems. 2022.

klasky@ornl.gov, scottk007@gmail.com

I/O pipelines in staging

- Use the staging nodes and create a workflow in the staging nodes
- Mitigate performance impact of I/O of the GTC code by using asynchronous data movement
- Improve total simulation time by 2.7% , but we also improved the reading performance + analyzed the data + visualized the data



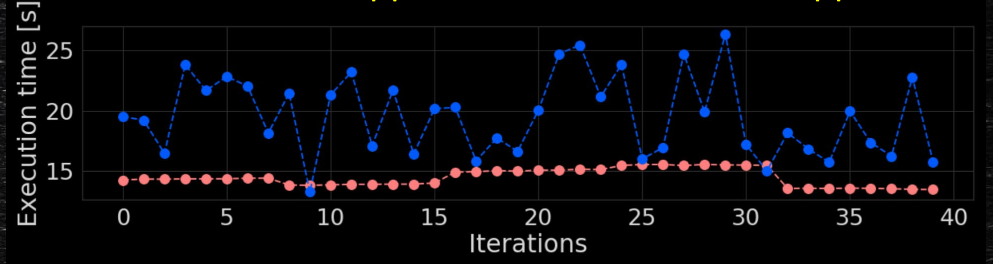
OAK RIDGE National Laboratory F. Zheng, S. Klasky, et al., "PreDatA - Preparatory Data Analytics on Peta-Scale Machines", IPDPS 2010.

klasky@ornl.gov, scottk007@gmail.com

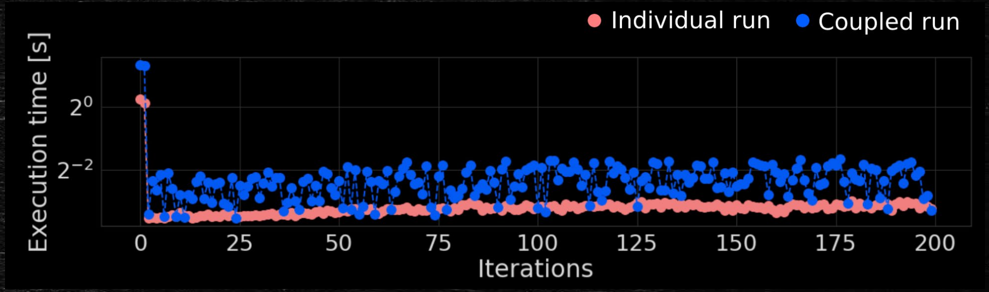
The movement for SDM for the convergence of HPC with AI

- Traditional simulations focused on scaling a single app
  - As they move to digital twins, they have been evolving into complex workflows which can mix simulations with AI
- The graphs show the cumulative execution times of a histopathology analysis pipeline that classifies image patches in WSIs to characterize tumor regions and lymphocyte distributions:
  - (Top) when multiple concurrent instances of the models are running in parallel, interference at the storage level causes a delay in the execution of each application shifting the distribution to the right
  - (Bottom) when streaming data directly to the consumer the I/O inference is decreased by 20%
- The results for analyzing 7,000 WSIs with the workflow using 500 concurrent instances on Summit is presented in the bottom: as we see, we can reduce congestion by streaming the data

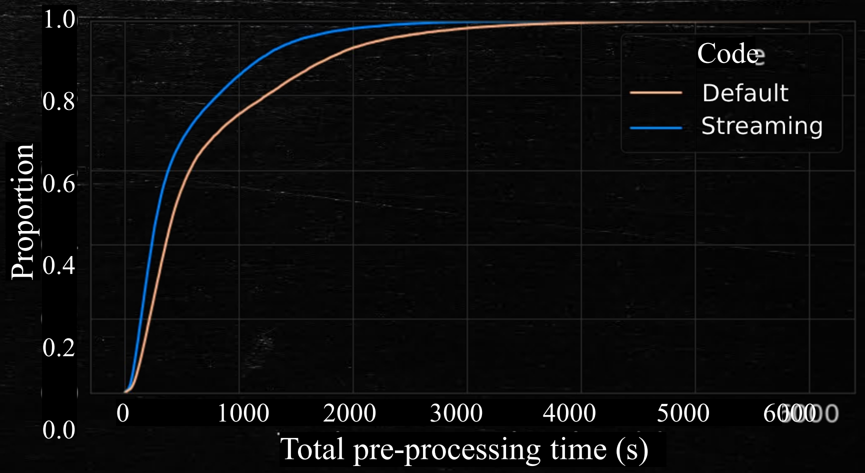
The effects of the DL application run on the HPC application



The effects of the HPC application run on the AI application



Use streaming to improve the performance by 20%

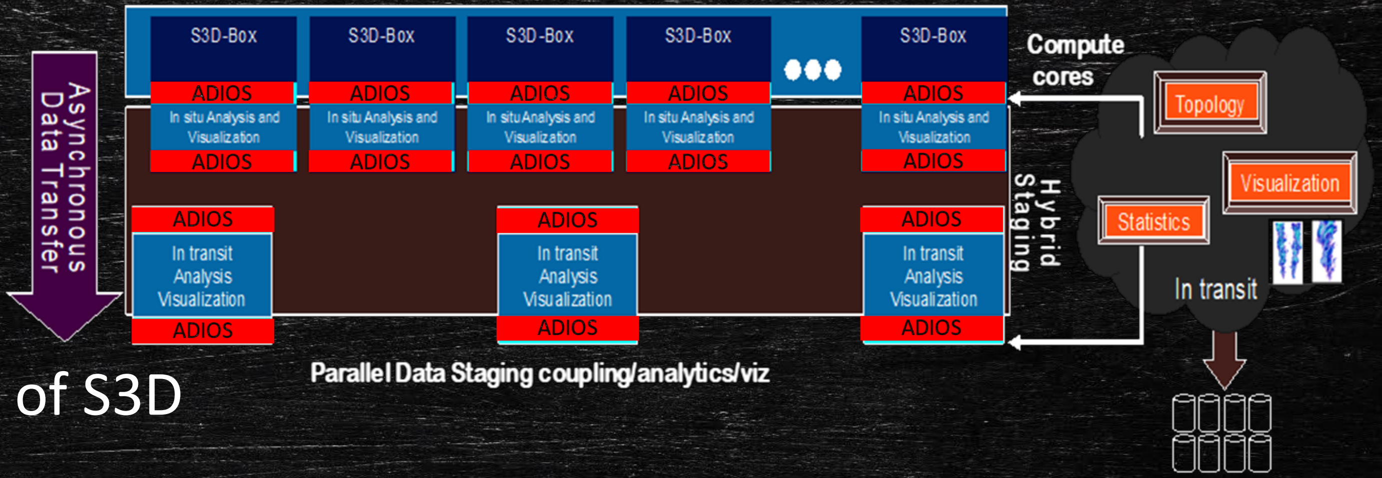


Gainaru, A., Klasky, S., et al., Framework for Automating the I/O of Deep Learning Methods in Biomedical Imaging Applications, submitted to IEE/ACM Transactions on computational biology and bioinformatics, 2022.

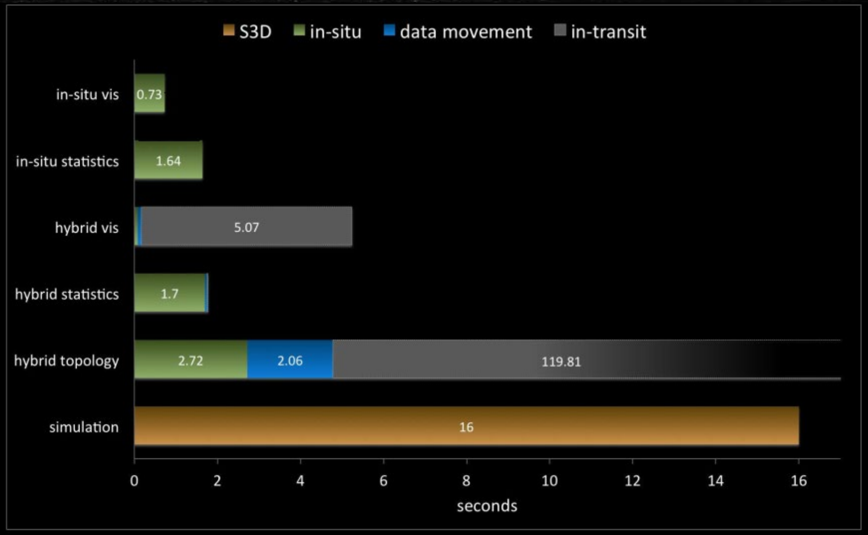
klasky@ornl.gov, scottk007@gmail.com

Using hybrid staging to enable extreme-scale scientific analysis

- Can we break algorithms into two parts
  - Embarrassingly parallel – do this inline
  - Communication heavy – do this asynchronously on separate staging nodes



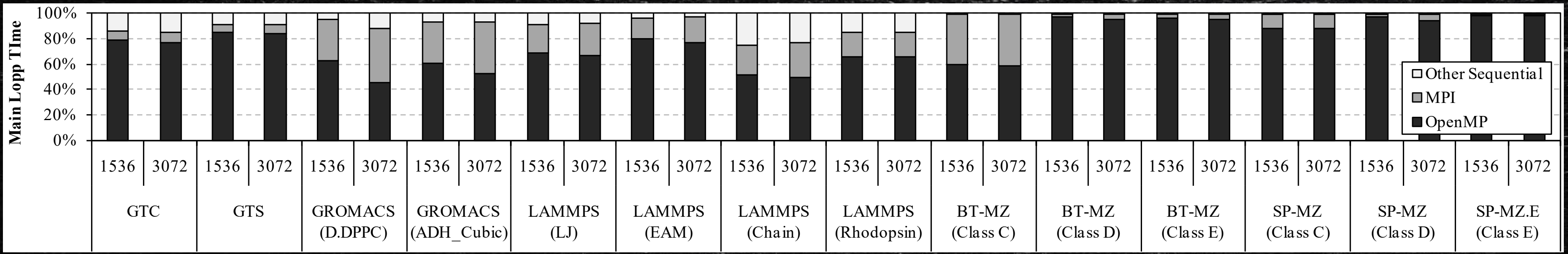
- Workflow to enable topological analysis of S3D data from large scale simulations
  - Visualization – in situ volume rendering
  - Topological feature extraction – uses a merge tree approach
- Main findings
  - Topological feature extraction uses the hybrid approach
  - Statistics is best with inline processing
  - Volume rendering - works well in all cases



OAK RIDGE National Laboratory Bennett, Klasky, S., et al. (2012). Combining in-situ and in-transit processing to enable extreme-scale scientific analysis. In SC'12: (pp. 1-9).

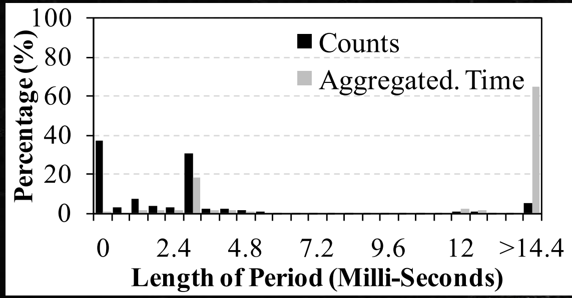
klasky@ornl.gov, scottk007@gmail.com

Data Staging considering Idle CPU Resources Un-used by simulations



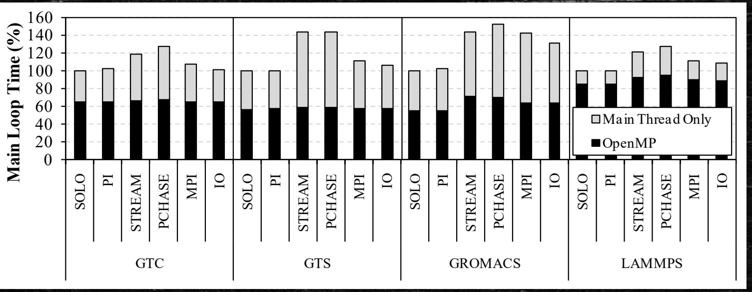
Challenge I

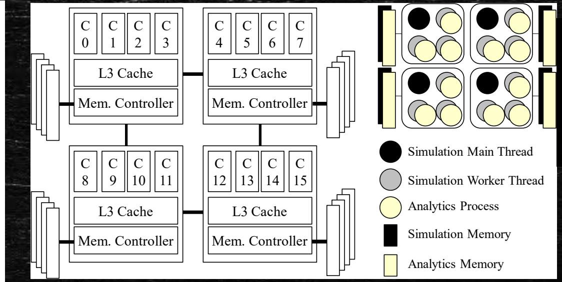
-Select suitable idle periods to amortize scheduling costs

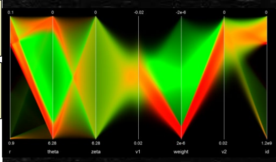


Challenge II

-Interference between simulation and analytics due to contention on memory hierarchy

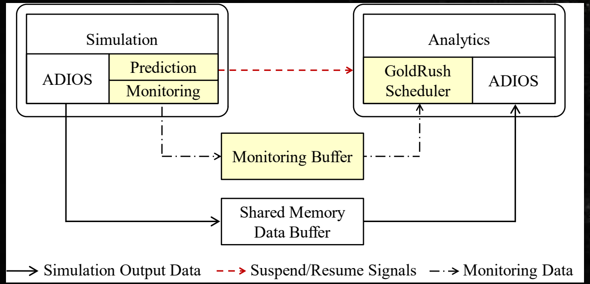






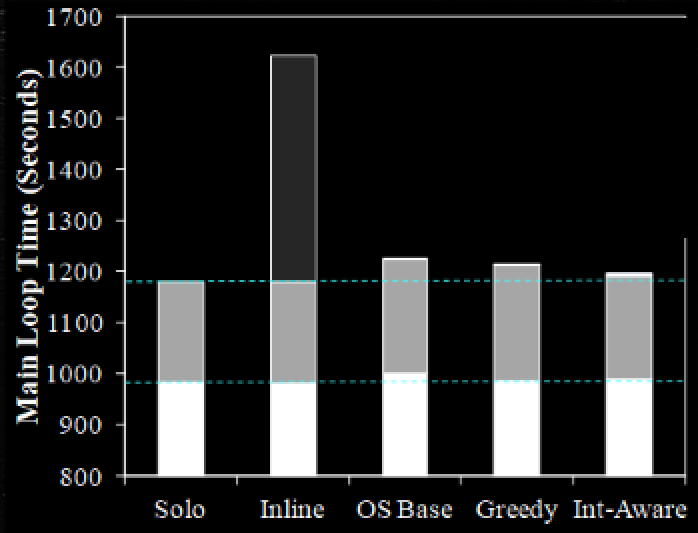
Harvest Idle Resource for In-Situ Analytics

- Dynamically predict idle resource availability
- Reduce interference with execution throttling



GTS simulation with parallel coordinate visualization

- Improve time to solution and resource efficiency
- Reduce off-node data movements
- Scale to up to 12288 cores on Hopper Cray XE6

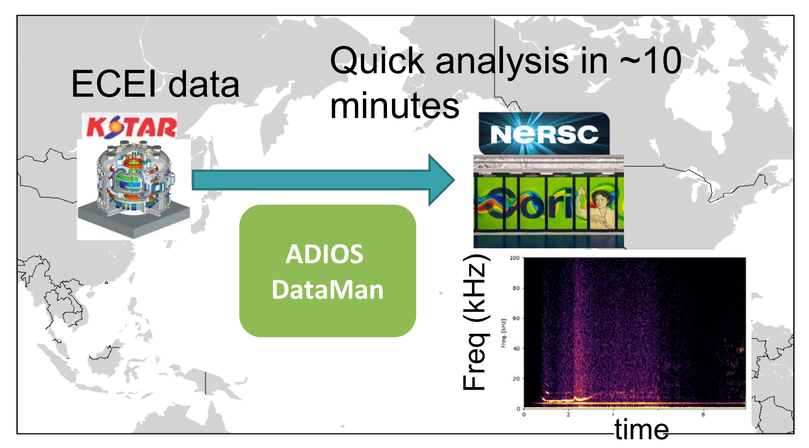


Zheng, F., Klasky, S., et al. (2013). Goldrush: Resource efficient in situ scientific data analytics using fine-grained interference aware execution. In SC'13 (pp. 1-12).

Using staging to establish capability for near-real time networked analysis of fusion experimental data (KSTAR)

Research and develop a streaming workflow framework, to enable near-real-time streaming analysis of KSTAR data on a US HPC

- Allow the framework to adopt ML/AI algorithms to enable adaptive near-real-time analysis on large data streams
- Created a framework to enable US fusion researchers to have broader and faster access to the KSTAR data, enabling
  - Faster analysis of data
  - Faster and autonomous utilization of ML/AI algorithms for incoming data
  - More informed steering of experiment
- Accomplishments
  - Created end-to-end Python framework, streams data using ADIOS over WAN (at rates > 4 Gbps), asynchronously processes on multiple workers with MPImulti-threading
  - Applied to KSTAR streaming data to NERSC Cori.
  - Reduces time for analysis from 12 hours to 10 minutes

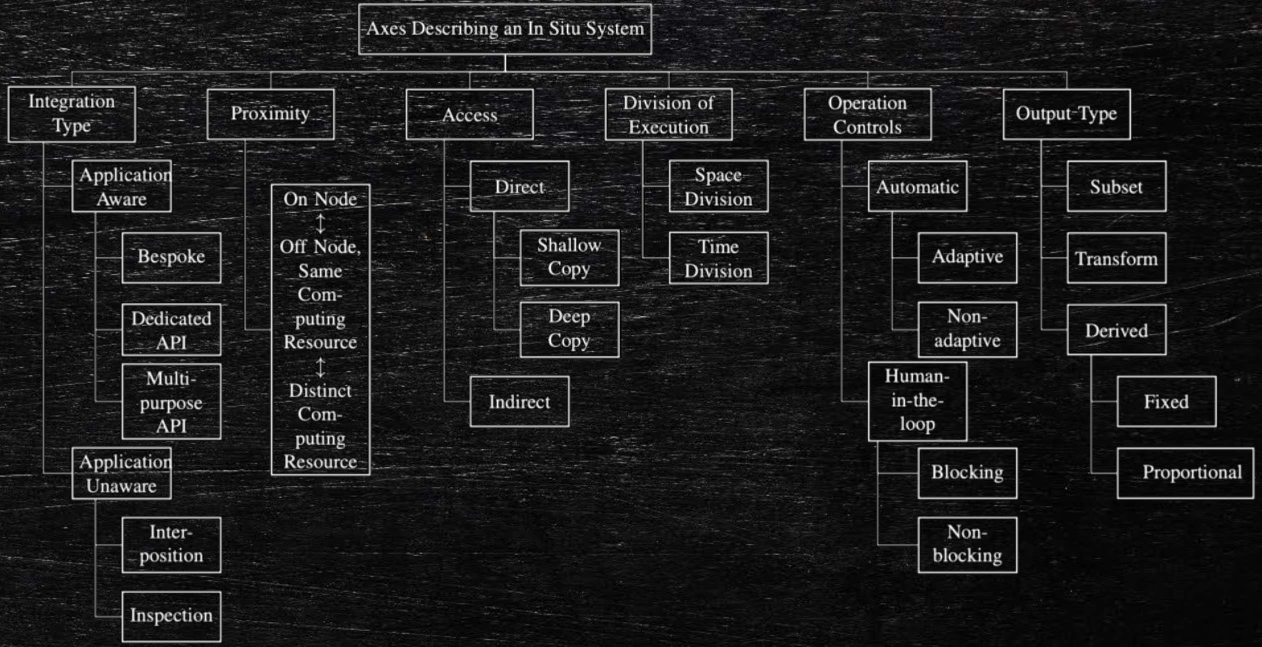


Churchill, Klasky, S., et al.(2021). A Framework for International Collaboration on ITER Using Large-Scale Data Transfer to Enable Near-Real-Time Analysis. Fusion Science and Technology, 77(2), 98-108.

klasky@ornl.gov, scottk007@gmail.com

In situ visualization

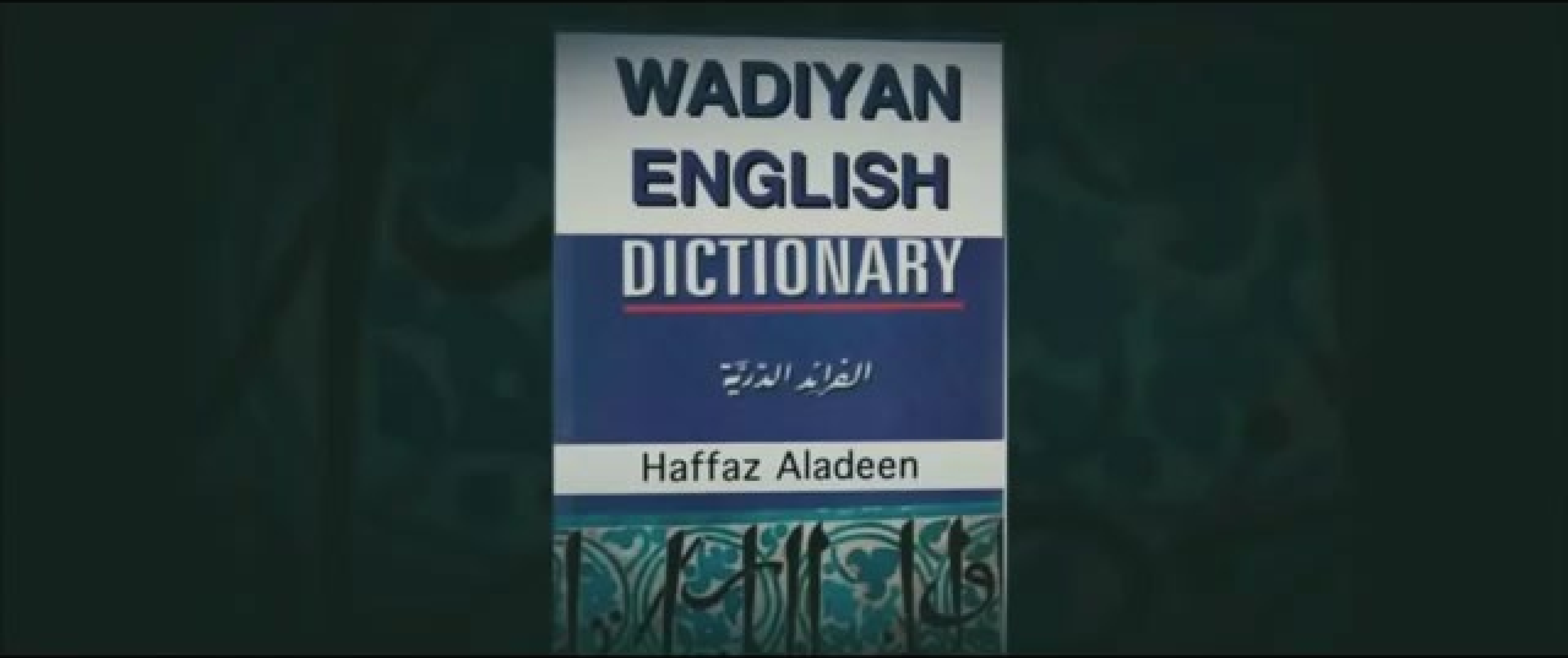
- Why in situ?
  - Imbalance between compute and I/O
  - Real time analysis and visualization
  - Probe a running simulation / experiment
- Post hoc: hard, but the 'how to do it' is straightforward
- In situ: hard **and** the 'how to do it' is harder
- Why ??
  - No 'file' to open
  - In situ is not just one thing
  - Block or crash the simulation
  - Unbounded costs



Childs, Hank, et al. "A terminology for in situ visualization and analysis systems." The International Journal of High Performance Computing Applications 34.6 (2020): 676-691.

klasky@ornl.gov, scottk007@gmail.com

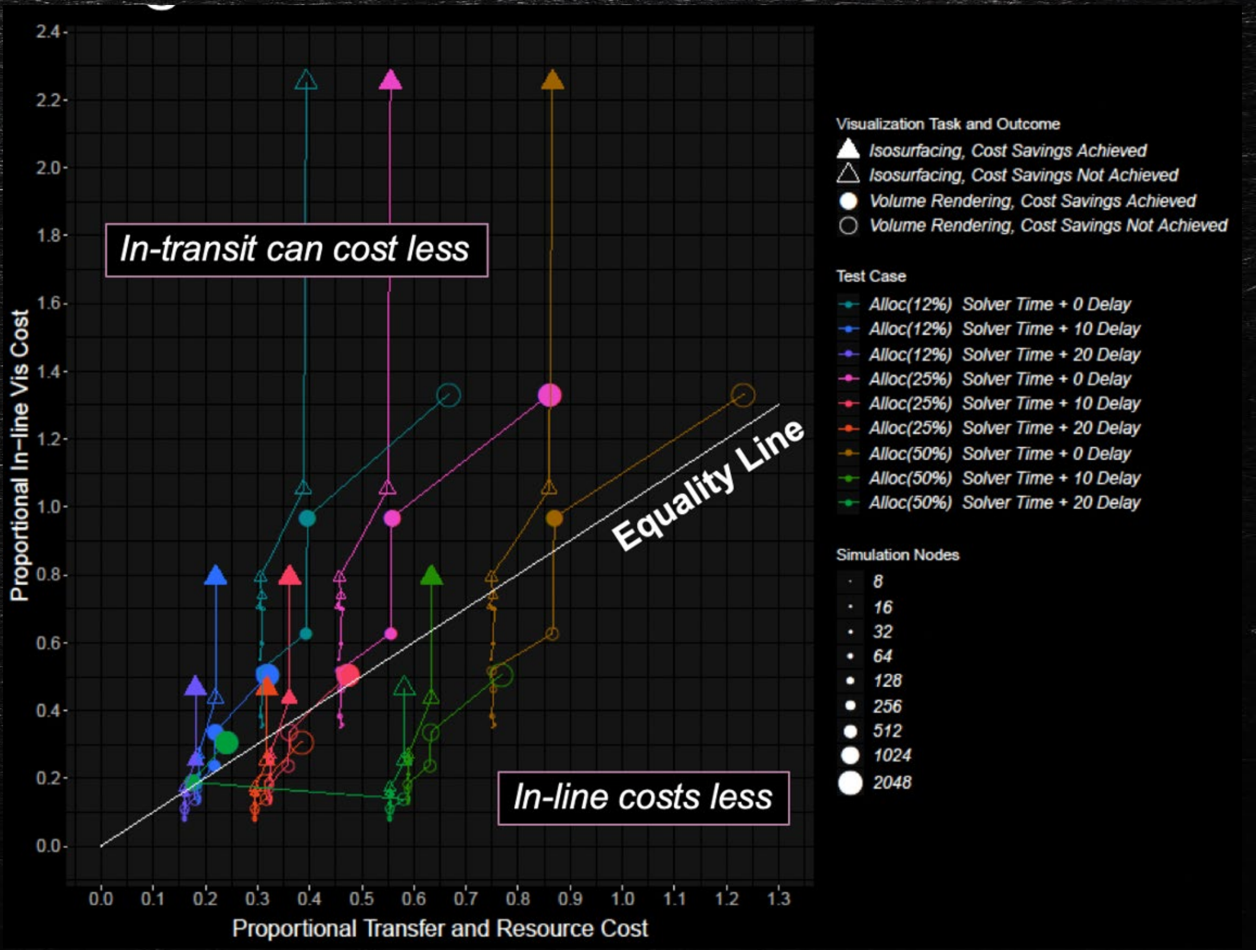
But it’s all in situ



[The Dictator (2012 film) - Wikipedia]

Comparing the efficiency of in situ visualization paradigms at scale

- The use of in-transit visualization requires more resources than in-line visualization
  - Overall costs for each paradigm will vary based on use cases
  - Relative efficiencies and costs of different use cases and in situ visualization paradigms was not well understood
- We identify and demonstrate the use cases where in-transit techniques are both faster and more cost efficient than in-line techniques
  - In-transit techniques are more cost efficient than in-line for communication heavy algorithms at large-scale
- We created a cost model for in situ visualization that shows the performance of each technique at a given scale
- Study conducted on proxy applications running at OLCF
  - Used VTK-m for visualization and ADIOS for in-transit data movement



Kress, J., Klasky, S., et al., (2019). Comparing the efficiency of in situ visualization paradigms at scale. In International Conference on High Performance Computing (pp. 99-117).

klasky@ornl.gov, scottk007@gmail.com

Outline

R&D
Analytics
Visualization

Understand simulations, experiments and observations

Understanding new technologies

R&D in Scientific Data Management

R&D Staging & Coupling

R&D Data Refactoring

Application Impact:
Celeritas, DeepdriveMD, E3SM, GE, GENE, GEM, GTC, JAXA, KSTAR, LAMMPS, NNESH, PIConGPU, S3D, SKA, SPECFEM3D_GLOBE, TAE, XGC, WarpX,WRF

DOE/ASCR Data Reduction

- Bill Spotz (ASCR)
  - Klasky, Najm, Thayer
- Main findings
- Data volumes and velocities from next generation experiments, observations and simulations require new R&D in data reduction

4 Priority Research Directions

1. Trust: Accuracy and Performance
2. Progressive Streaming
3. Feature Preserving
4. Platform Portability

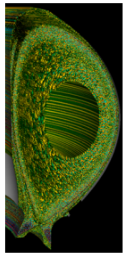
Data Reduction for Science: Brochure from the Advanced Scientific Computing Research Workshop

Scott Klasky, Oak Ridge National Laboratory
Jana Thayer, SLAC National Accelerator Laboratory
Habib Najm, Sandia National Laboratories

Publication date: April 15, 2021
Web DOI: 10.2171/1770192
DOE Office of Science Technical Contact: William Spotz (William.Spotz@science.doe.gov)

Introduction

The reduction of streaming and voluminous data sets while maintaining accurate representations of quantities of interest (QoIs) is a critical capability across the Office of Science (SC). SC-supported experiments, observations, and simulations produce data at volumes and velocities that are already overwhelming network, storage, and compute capabilities and their projected growth will greatly exacerbate this imbalance. The Advanced Scientific Computing Research program office held a virtual workshop in January 2021, bringing together 155 participants and 41 observers across experimental, observational, and computational application areas and research thrust areas in compression, reduced representations, experiment-specific triggers, filtering, and feature extraction/QoIs to identify priority research directions (PRD) leading to enhanced capabilities in data reduction. This workshop examined many scientific drivers, such as radio astronomy, fusion, combustion, climate, light sources, nuclear physics, and genomics, which are in desperate need for new Research & Development (R&D) in data reduction, because they currently risk ad hoc decisions that can limit the amount of knowledge gathered from SC facilities.



New workflows are beginning to emerge to both manage data and fully exploit the incredibly rich information produced by SC facilities. These data reduction workflows employ triggering, filtering, sampling, compression, reduced order modeling and feature detection. The workflows extend from observational/experimental devices to networks to remote and local storage to desktop and leadership computing facilities and require optimization across a diverse range of hardware.

In order for application scientists to trust data reduction methodologies, reduction techniques should be usable and adoptable by communities through best practices, benchmarks, data sharing, resource sharing, and through the development of tools that enable scientists to navigate these resources. The workshop focused on new R&D capabilities which can allow scientists to quantify the uncertainties in QoIs, along with preserving features to a specified tolerance. Furthermore, progressive techniques for streaming data need to be developed to enable scientists to make tradeoffs between the uncertainty, speed, and resource utilization. Since these workflows typically run on all types of

Home Agenda Presentations Contacts
Registration is Closed
Already Registered?

Data Reduction for Science Workshop

Sponsored by the U.S. Department of Energy,
Office of Advanced Scientific Computing Research

January 25, 26, and 28, 2021

The workshop will be held using a virtual format.

Scientific observations, experiments, and simulations are producing data at a rate beyond our capacity to store, analyze, stream, and archive. This data almost always contains redundancies and trivialities that hide the important information of interest to scientists. Of necessity, many research groups have already begun reducing the size of their data sets via techniques such as compression, reduced representations, experiment-specific triggers, filtering, and feature extraction. These efforts should be expanded to include mathematical rigor to ensure that quantities of interest are conserved, to be offered as services from scientific user facilities, to be integrated into scientific workflows, and to be implemented in a manner that inspires trust that the desired information is preserved. The purpose of this workshop is to:

- Bring together disparate communities of practice in the data reduction space to foster collaboration and improved understanding of the various techniques
- Outline requirements for data reduction techniques from domain scientists
- Highlight the relevant state-of-the-art in computer science and mathematics
- Identify priority research directions leading to enhanced capabilities in data reduction

Add to Calendar

U.S. DEPARTMENT OF ENERGY | Office of Science
ORAU
OAK RIDGE INSTITUTE FOR SCIENCE AND EDUCATION
Managed by ORAU for DOE

ORAU/ORISE Privacy Policy | Contract Acknowledgement

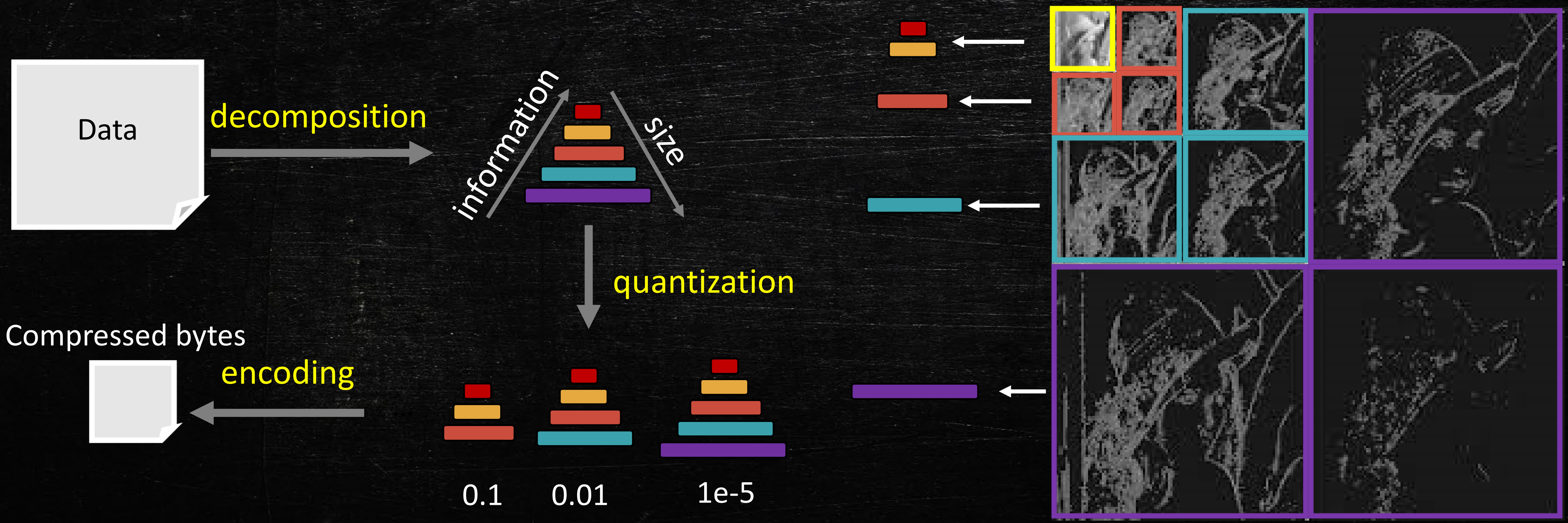
Support for Internet Explorer 11 will be ending soon. For the best experience using this site, we recommend using Google Chrome, Mozilla FireFox, or Microsoft Edge as your desktop browser.

Klasky, S., Thayer, J., & Najm, H. (2021). Data Reduction for Science: Brochure from the Advanced Scientific Computing Research Workshop..ORNL; SLAC; SNL-CA, SNL-NM.

MGARD - Multi-Grid Adaptive Reduction of Data

MGARD is a transform-based compressor

- multi-resolution, multi-precision



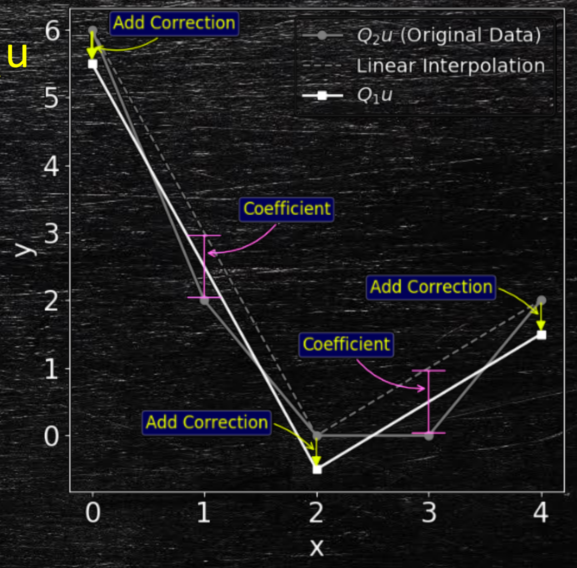
Similar to filtering Fourier coefficients, JPEG, wavelet methods

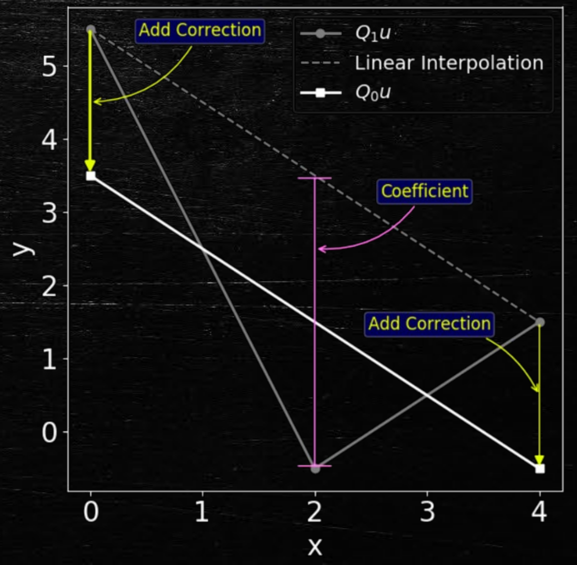
Ainsworth, M., Klasky, S., et al., (2018). Multilevel techniques for compression and reduction of scientific data—the univariate case. Computing and Visualization in Science, 19(5), 65-76.
klasky@ornl.gov, scottk007@gmail.com

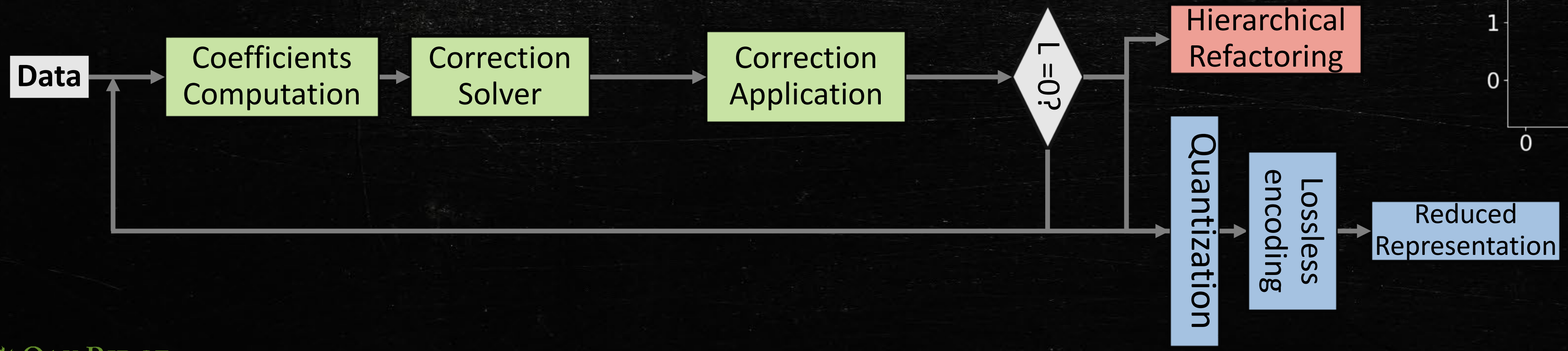
Decomposition Algorithm

$Q_\ell u \rightarrow (I - \Pi_{\ell-1}) Q_\ell u$ and $Q_{\ell-1} u$

- Coefficient computation
  - L^2 projection, linear interpolation and subtraction
- Correction solver
  - another L^2 projection from fine grid to coarse grid
- Correction application
  - Add correction to the nodal values on the coarse grid
- Recomposition is the inverse of decomposition.







klasky@ornl.gov, scottk007@gmail.com

MGARD Theory: Multilevel Decomposition

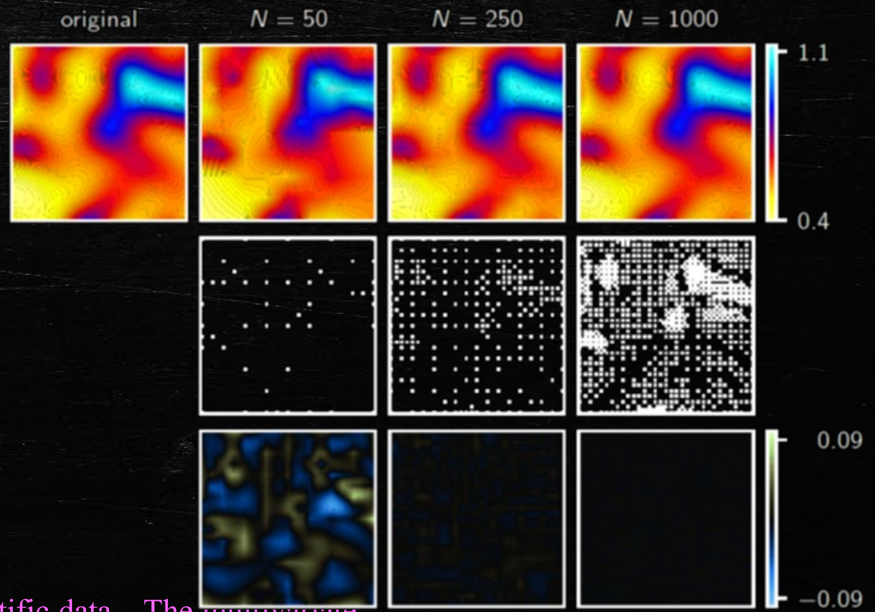
- Goal: transform u into a presentation that's amenable to compression

Define multilevel coefficients u_mc by

$$\texttt{u\_mc[x]} = (I - \Pi_{\ell-1})\, \mathcal{Q}_\ell\, u(x) \quad \text{if} \quad x \in N_\ell^*$$

where

- $\mathcal{Q}_\ell$ is the L^2 projection onto V_ℓ, the space of continuous piecewise linear functions on (intermediate) mesh P_ℓ ,
- Π_ℓ is the linear interpolation onto $V_{\ell-1}$ the space of continuous piecewise linears on $\mathcal{P}_{\ell-1}$
  - This of this as a restriction operator in MultiGrid
- $\mathcal{N}_\ell^*$ is the set of 'new' nodes in $\mathcal{P}_\ell$
- X is the spatial location



Ainsworth, M., Klasky, S., et al., 2019. Multilevel techniques for compression and reduction of scientific data---The multivariate case. SIAM Journal on Scientific Computing, 41(2), pp.A1278-A1303.

klasky@ornl.gov, scottk007@gmail.com

Control of Errors in Raw Data

Theorem: Let u, $\tilde{u} \in V_l$ with multilevel coefficients `u_mc`, `ũ_mc`. Let vol($\mathcal{P}_\ell$) be the size of the $\mathcal{P}_\ell$ mesh elements. If

$$\sum_{\ell=0}^{L} 2^{2s\ell}\mathrm{vol}(P_\ell) \sum_{X \epsilon N_\ell^*} |u_{mc}[x] - \tilde{u}_{mc}[x]|^2 \leq \tau^2$$

then $\| u - \tilde{u} \|_s \leq \tau$.

To apply the theorem, take *u* to be the original function

- Compute the multilevel coefficients `u_mc` of *u*
- Generate `ũ_mc` so that the inequality holds
- Recompose to obtain a reduced function $\tilde{u}$ respecting the error tolerance

klasky@ornl.gov, scottk007@gmail.com

MGARD - Multi-Grid Adaptive Reduction of Data

MGARD controls the compression errors in quantities of interest

- If we know how we'll use the reduced data, we can more aggressively compress

Let $\mathcal{Q}: V_L \rightarrow \mathbb{R}$ be a function of analyzer (e.g., the average over some piece of the domain)

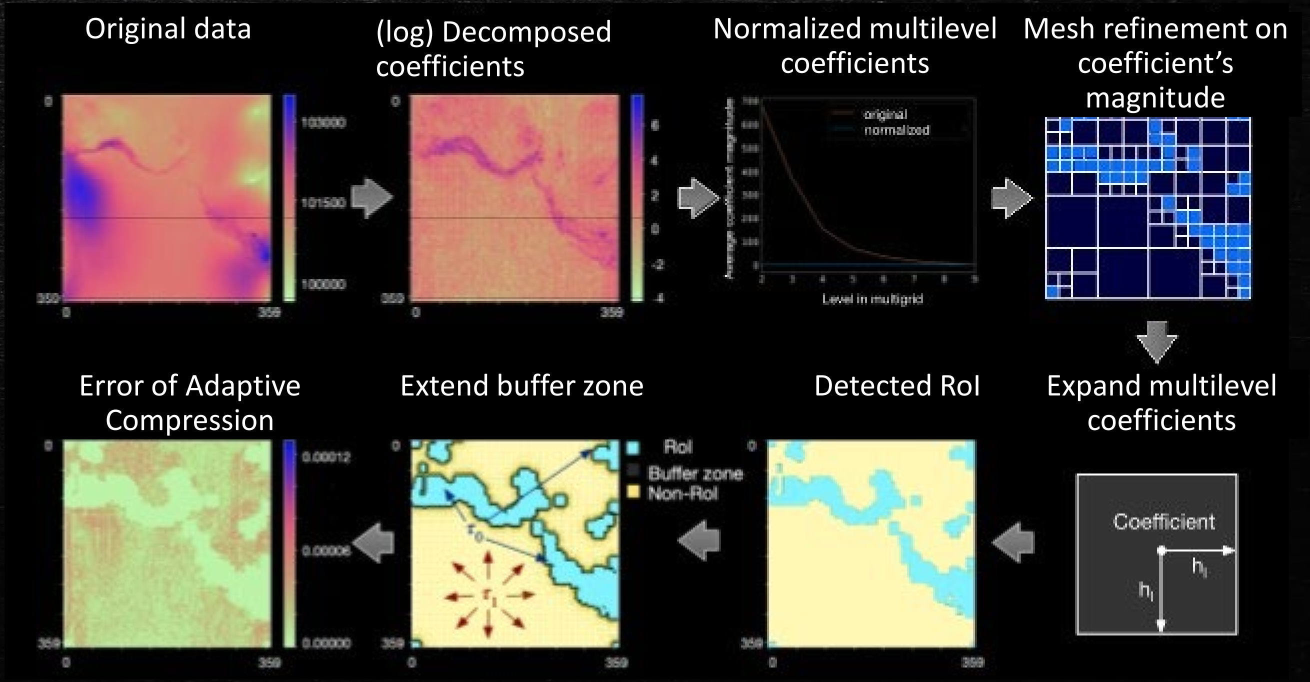
$$|\mathcal{Q}(u) - \mathcal{Q}(\tilde{u})| = |\mathcal{Q}(u - \tilde{u})| \leq \| \mathcal{Q} \|_s \| u - \tilde{u} \|_{-s} \leq \tau$$

- Here $\| \mathcal{Q} \|_s$ is the operator norm of $\mathcal{Q}$
- The choice of s (i.e., norm) affects how u is compressed
- Depending on the QoI, we will choose different values of s

Ainsworth, M., Klasky, S., et al. (2018). Multilevel techniques for compression and reduction of scientific data—the univariate case. Computing and Visualization in Science, 19(5), 65-76.
klasky@ornl.gov, scottk007@gmail.com

Region-adaptive compression for reduction of climate data

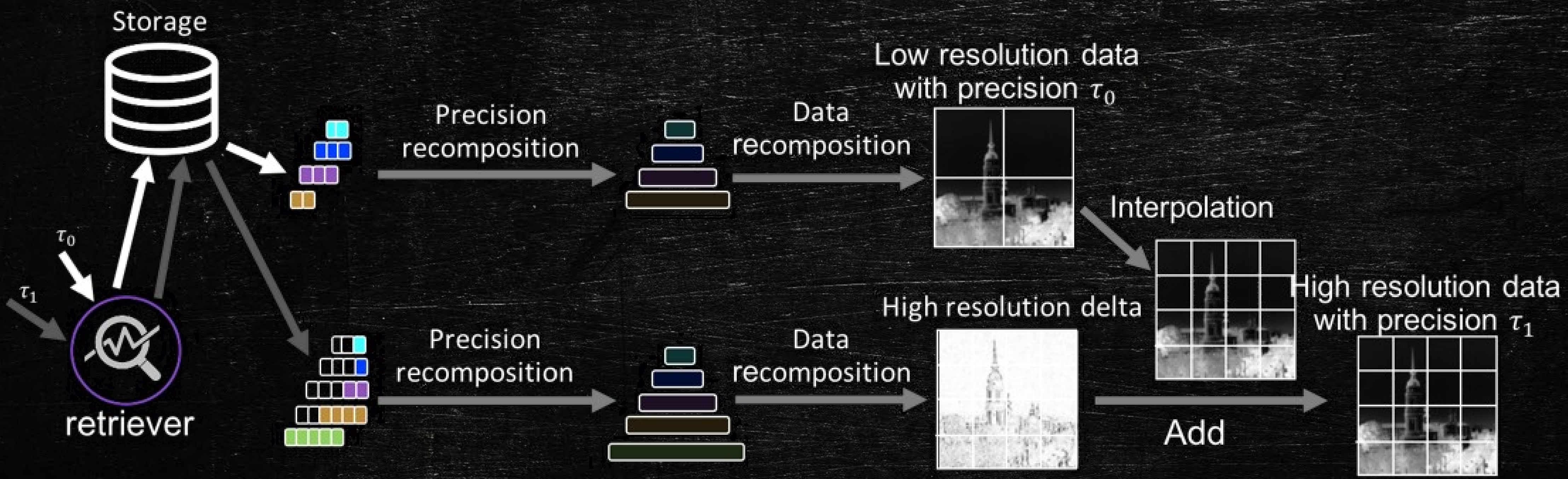
- Motivation:
  - It's common in scientific datasets that only a small portion of space are of interest
  - Store region-of-interest (RoI) with more bits and compress the rest with lower precision, so that larger compression ratios can be achieved while task-interested information are preserved
- Key Techniques
- MGARD decomposition
- Critical region detection by applying mesh refinement on the decomposed coefficients
- Region-wise error control through multi-level extended buffer zone
- Mask-free, multi-error bounded data compression and reconstruction



Gong, Q., Klasky, S., et al., 2022, Region-adaptive, Error-controlled Scientific Data Compression using Multilevel Decomposition, SSDBM022.

klasky@ornl.gov, scottk007@gmail.com

Progressive data reconstruction



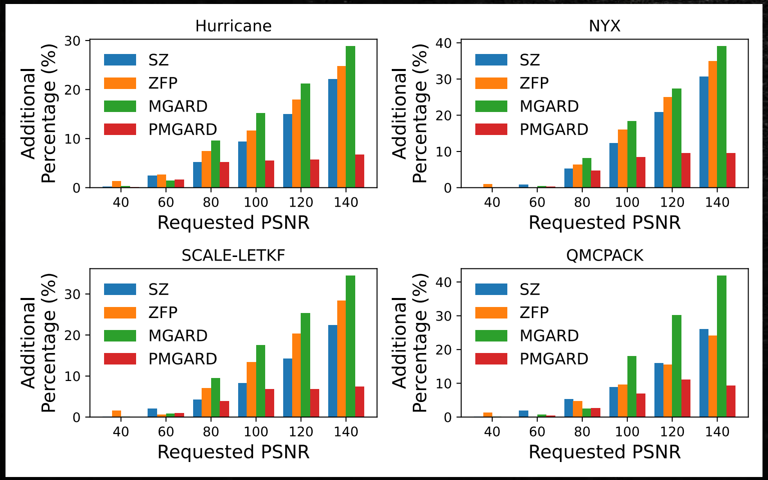
Progressively fetch the data based on requested precision (i.e., tolerance τ_i)

- Reduces data movement for requested precision
- Allow incremental data reconstruction, from low precision to high precision
- Asynchronous data streaming and data analyzing

Results of progressive retrieval on SDRB datasets

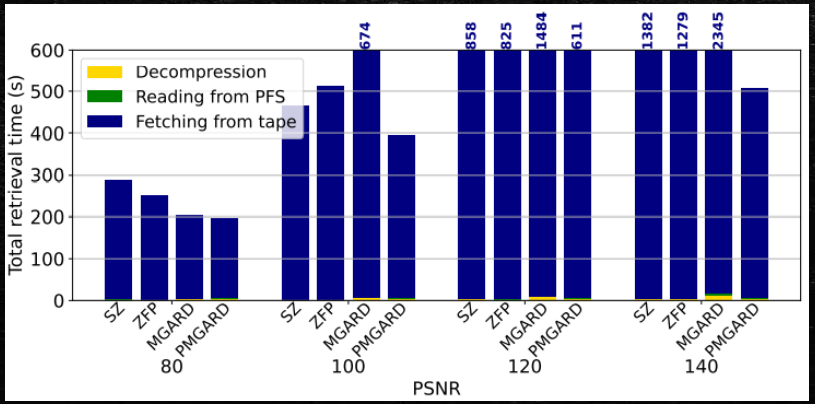
Reduction in Retrieved Data Size

- Comparing to SZ, ZFP, and non-progressive MGARD: additional retrieval percentage under given PSNR when data of previous precision are available



Reduction Retrieval and Recomposing

- Comparing to SZ, ZFP, and non-progressive MGARD: total retrieval time when data is transferred from High Performance Storage System and progressively reconstructed using 1024 cores on Summit



Reduction in Analysis Time

- Comparing to SZ, ZFP, and non-progressive MGARD: speed of iso-surface analysis when target PSNR is 60

Model	%	Resolution	Analysis time (s)	Analysis error
SZ	2.19%	512^3	60.83	2.25%
ZFP	1.78%	512^3	59.99	5.89%
MGARD	10.62%	257^3	10.99	5.08%
PMGARD	0.81	257^3	11.16	5.67%

Liang, X, Klasky, S., et al. (2021). Error-controlled, progressive, and adaptable retrieval of scientific data with multilevel decomposition. In SC'21 (pp. 1-13).

Moving to the future: Control of quantities of interest

Theorem

Let $V_0 \subset \ldots \subset V_L$ be space of continuous piecewise multilinear functions defined on uniform tensor product grids on a domain $\Omega \subset \mathbb{R}^d$. Let $\mathcal{Q}$ be a *bounded linear functional* on V_L. Let $u \in V_L$ with multilevel coefficients u_mc. Let ũ_mc be a set of multilevel coefficients and let $\tilde{u} \in V_L$ be the corresponding function. Then the loss in quantity of interest is bounded by:

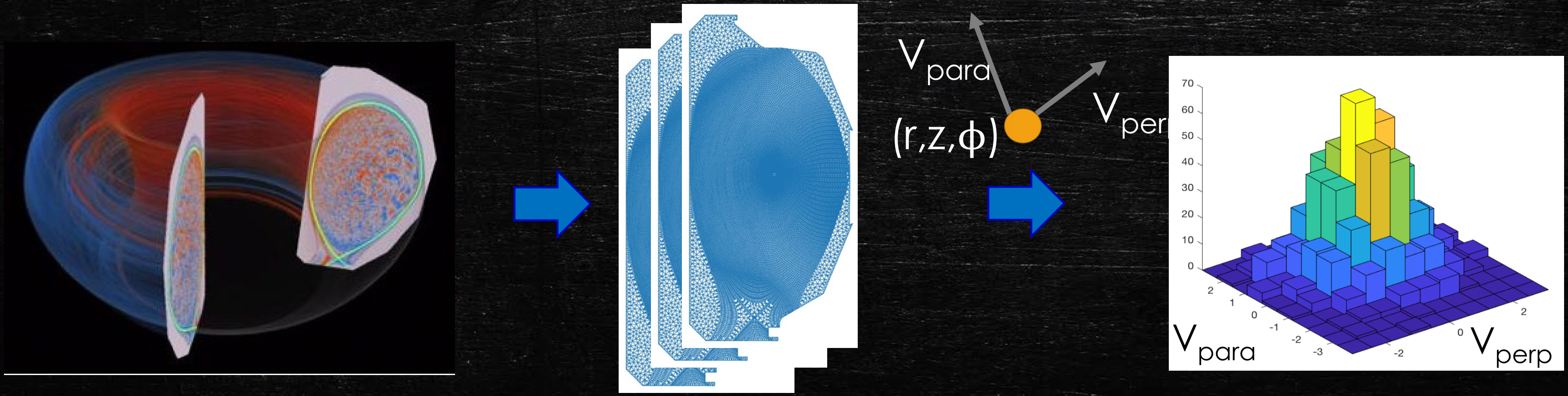
$$|\mathcal{Q}(u) - \mathcal{Q}(\tilde{u})| \leq \Upsilon_s(\mathcal{Q}) \left(\sum_{\ell=0}^{L} 2^{2s\ell} \operatorname{vol}(\mathcal{P}_\ell) \sum_{x \in \mathcal{N}_\ell^*} |\texttt{u\_mc}[x] - \tilde{\texttt{u}}\texttt{\_mc}[x]|^2 \right)^{1/2}$$

where $\Upsilon_s(\mathcal{Q})$ is the operator norm of $\mathcal{Q}$ and can be mathematically derived

Ainsworth, M., Klasky, S., et al. (2019). Multilevel techniques for compression and reduction of scientific data-quantitative control of accuracy in derived quantities. SIAM Journal on Scientific Computing, 41(4), A2146-A2171.
klasky@ornl.gov, scottk007@gmail.com

XGC Fusion Code

- A full-f gyrokinetic particle-in-cell (PIC) code which specializes in simulating kinetic transport in edge tokamak
- The code solves for a 5-dimensional ($\{r, z, \phi\}, \{v_x, v_y\}$) particle distribution function f defined on unstructured-meshed radial-poloidal (RZ) planes



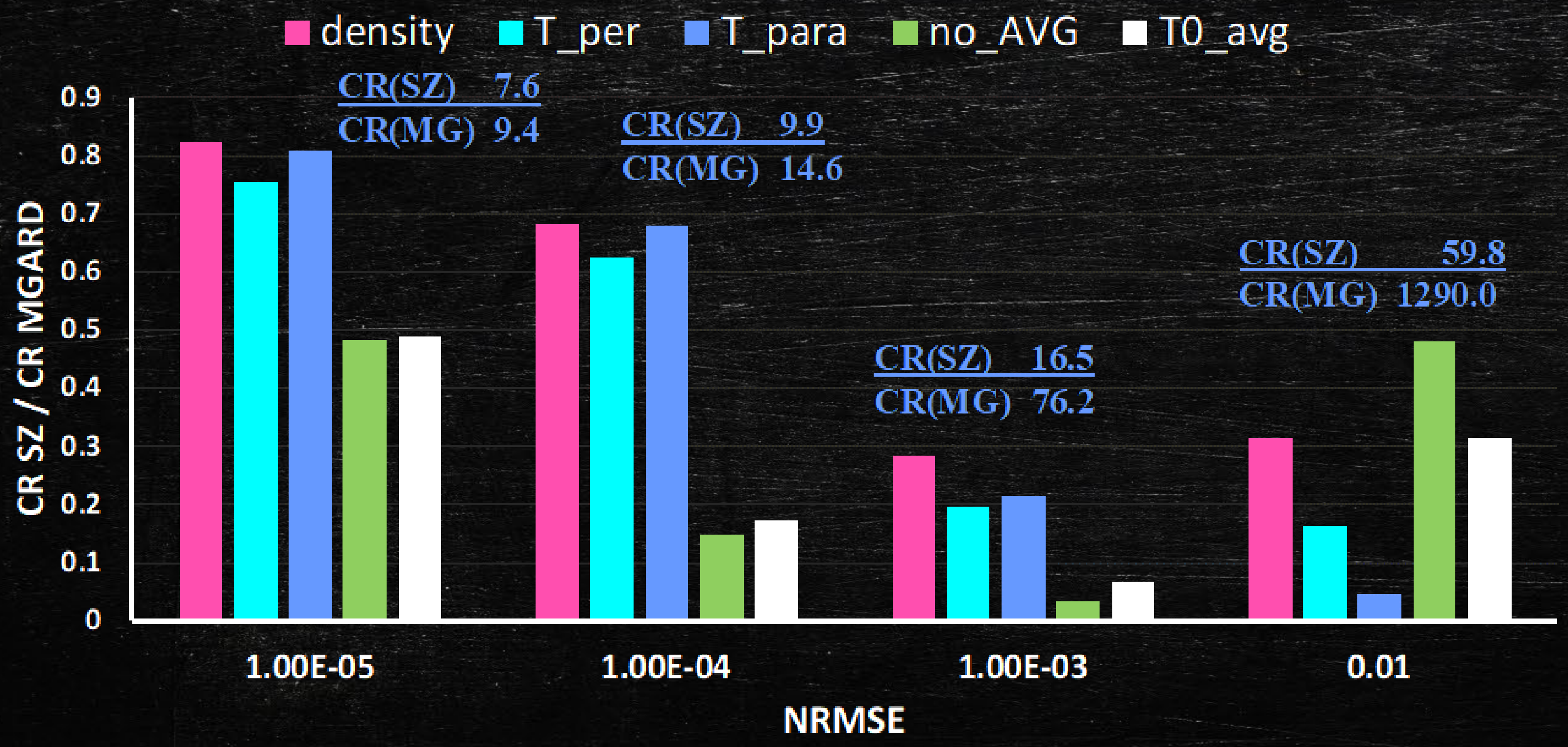
- A simulation modeling ITER-scale problems will typically contain trillions of particles and can each day produce over 200 PB of data

Ku, S., Klasky, S., et al. "Gyrokinetic particle simulation of neoclassical transport in the pedestal/scrape-off region of a tokamak plasma." Journal of Physics: Conference Series. Vol. 46. No. 1. IOP Publishing, 2006.

klasky@ornl.gov, scottk007@gmail.com

Comparing MGARD against state-of-the-art compressor: SZ

Ratio of compression ratios: SZ / MGARD



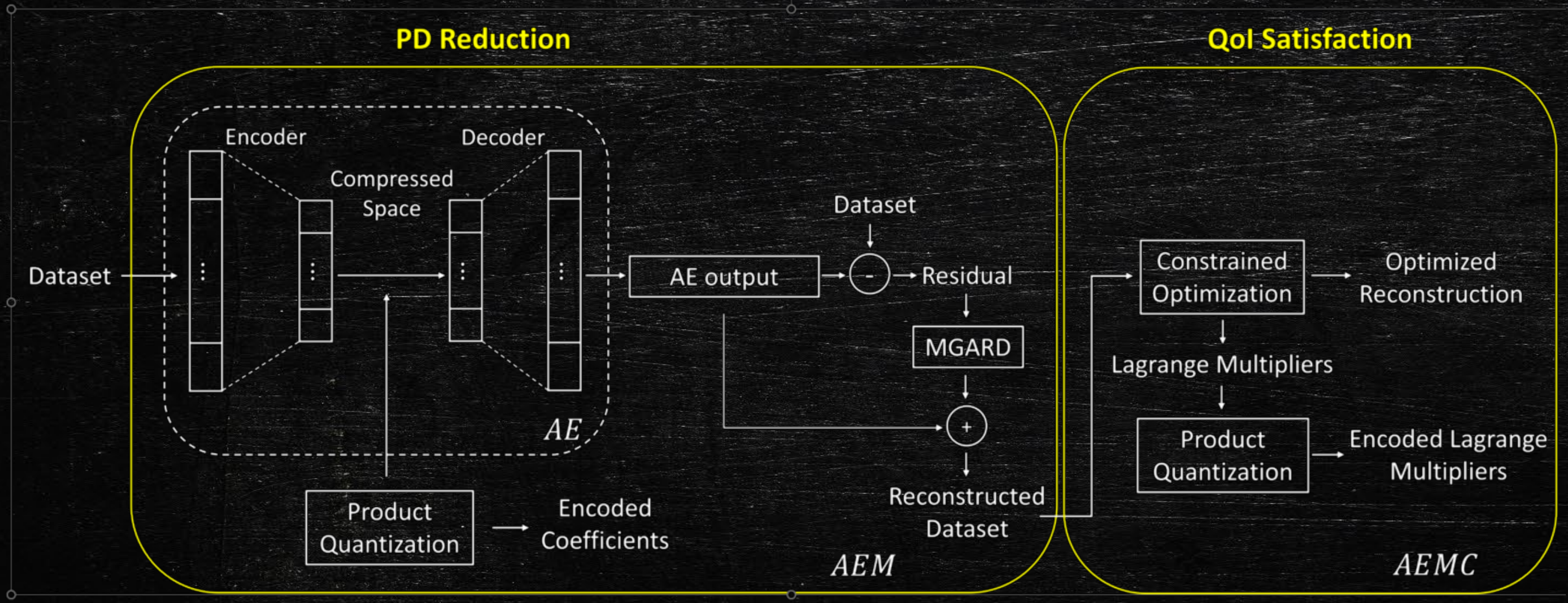
MGARD shows more advantage on low-frequency QoIs and in situations when the requested error bounds are loose

OAK RIDGE National Laboratory Gong, Q., Klasky, S., et al., (2021). Maintaining Trust in Reduction: Preserving the Accuracy of Quantities of Interest for Lossy Compression. In Smoky Mountains Computational Sciences and Engineering Conference (pp. 22-39). Springer, Cham.

klasky@ornl.gov, scottk007@gmail.com

Compression Framework - AEMC

AEMC: combination of PD reduction (AEM) and constraint satisfaction (C).



- **Auto-Encoder (AE):** an artificial neural network that has an encoder and decoder for compression.
- **Product Quantization (PQ):** decompose high dimensional vector into the Cartesian product of subspace and then quantize the subspace vectors separately.
- **MGARD:** an error-bounded lossy compression technique that guarantees PD reconstruction error

J. Lee, S. Ranka, A. Rangarajan, et al. Error-bounded Learned Scientific Data Compression with Preservation of Derived Quantities, accepted to Artificial Intelligence Data Engineering in Engineering Applications, 2022.

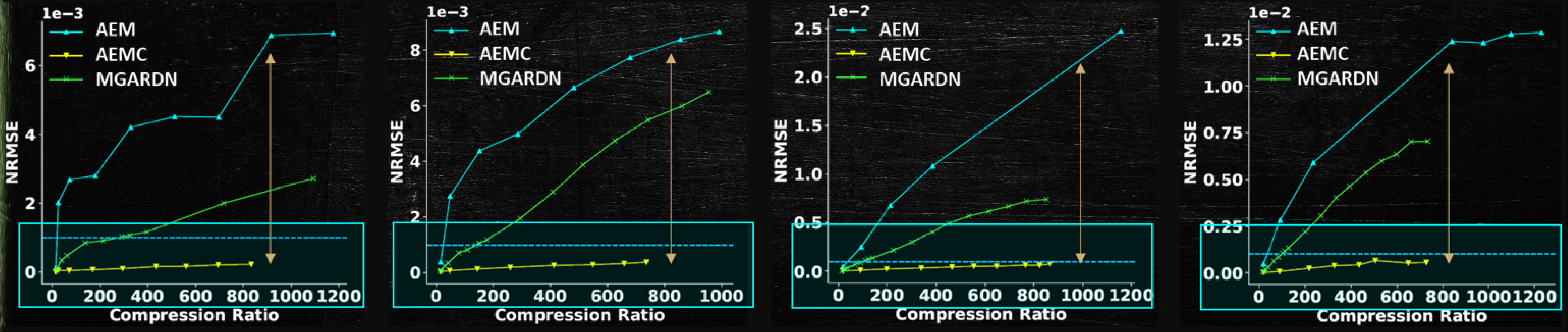
QoI Satisfaction for XGC – cont.

IMPROVEMENT of QoI by Constraint Satisfaction

- Works for various XGC timesteps.
- The dash line indicates the requirement of XGC scientists.

Averaged QoI

MGARDN: MGARD non-uniform (s=-1)



Timestep: 100 | Timestep: 400 | Timestep: 700 | Timestep: 1000

QoI errors are reduced significantly

Outline

R&D
Analytics
Visualization

Understand simulations, experiments and observations

Understanding new technologies

R&D in Scientific Data Management

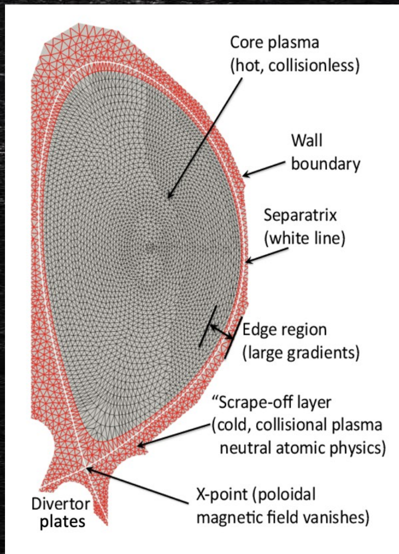
R&D Staging & Coupling

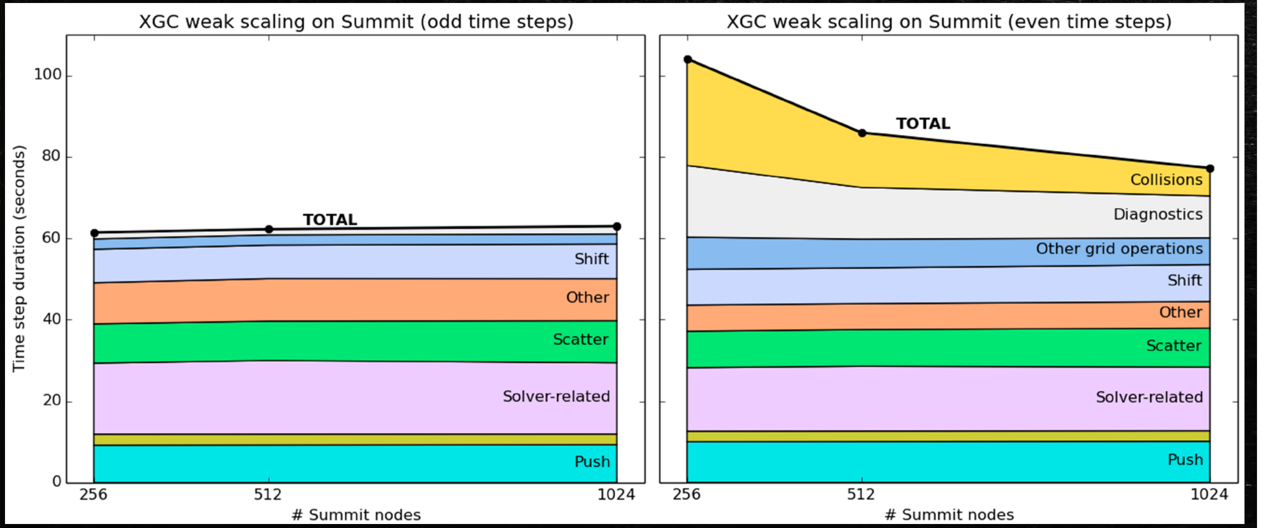
R&D Data Refactoring

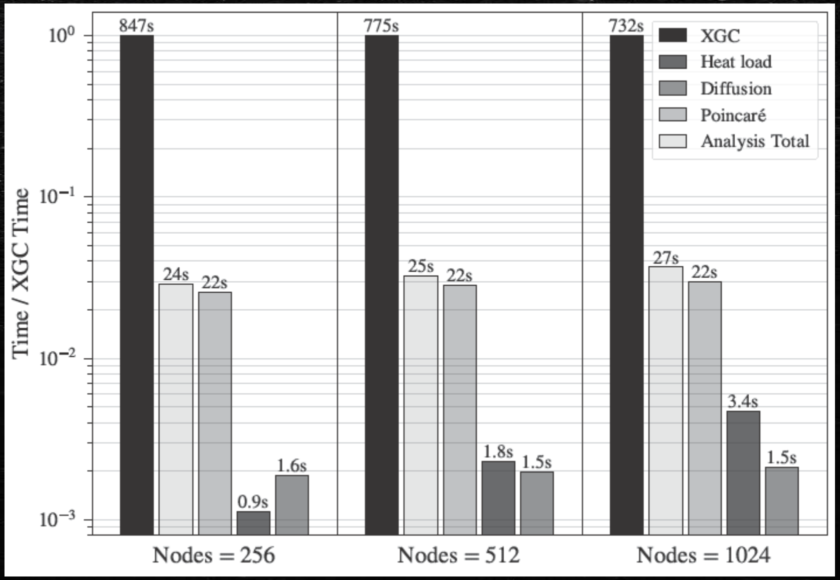
Application Impact:
Celeritas, DeepdriveMD, E3SM, GE, GENE, GEM, GTC, JAXA, KSTAR, LAMMPS, NNESH, PIConGPU, S3D, SKA, SPECFEM3D_GLOBE, TAE, XGC, WarpX,WRF

Hybrid Analysis of Fusion Data for Online Understanding of Complex Science on Extreme Scale Computers

- We examine a complex workflow using XGC on Summit, with three in situ analysis for new scientific discovery
- We execute XGC along with three analysis routines
  - Poincare Puncture Plot
  - Head Load calculation
  - Diffusion Calculation







OAK RIDGE National Laboratory Suchyta, E., Klasky, S. , et al., Hybrid Analysis of Fusion Data for online understanding of complex science on extreme scale computers, Cluster 2023.

klasky@ornl.gov, scottk007@gmail.com

ICRAR

International Centre for Radio Astronomy Research

Using ADIOS2 to enable SKA-scale processing

Andreas Wicenec

On behalf of ICRAR DIA team and the Gordon Bell prize finalist team

Curtin University

THE UNIVERSITY OF WESTERN AUSTRALIA

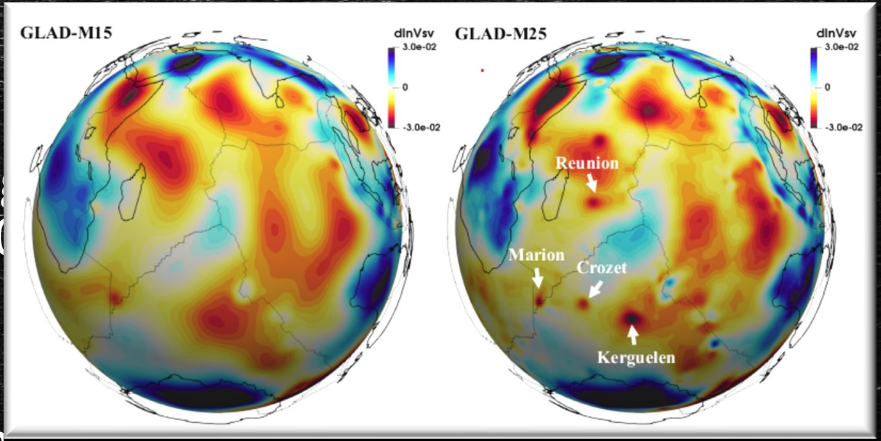
Government of Western Australia
Department of the Premier and Cabinet
Office of Science

Wang, R., Klasky, S., et al. (2020). Processing full-scale square kilometre array data on the summit supercomputer. In SC20: International Conference for High Performance Computing, Networking, Storage and Analysis (pp. 1-12). IEEE. Gordon Bell nom.

klasky@ornl.gov, scottk007@gmail.com

Seismic Tomography Workflow (PBs of data/run) [2.2 TB/s]

PI: Jeroen Tromp, Princeton



Map views at 250 km depth of vertically polarized shear wave speed perturbations in GLAD-M15 (2017) and GLAD-M25 (2020) in the Indian Ocean. New features have emerged in GLAD-M25, such as the Reunion, Marion, Kerguelen, Maldives, Seychelles, Cocos and Crozet hotspots.

Scientific Achievement

- Most detailed **3-D model of Earth**'s interior showing the entire from the surface to the core–mantle boundary, a depth of 1,800

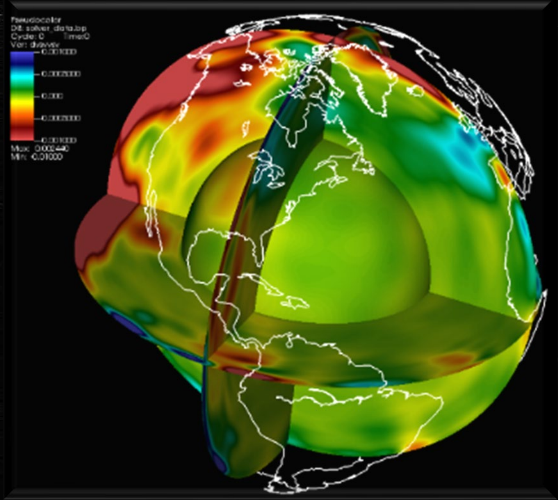
Significance and Impact

- Updated (transversely isotropic) global seismic model GLAD-M2 used to simulate how seismic waves travel through the Earth. T processing are challenging even for leadership computer
- **7.5 PB** of data is produced in a single workflow step
  - which is fully processed later in another step

Improvement by appending steps

- 3200 nodes ensemble run, 19200 GPUs
- 50 tasks at once
- 5.2 TB per task in 133 steps
- 260 TB total per 50 tasks
- 7.5 PB per 1500 tasks (total run)

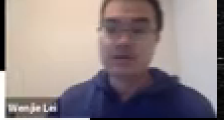
50 tasks, 133 steps, 3200 nodes	Time
No I/O	94s
BP3, one file per step	235s
BP4 one dataset per job 133x reduction in # of files	156s



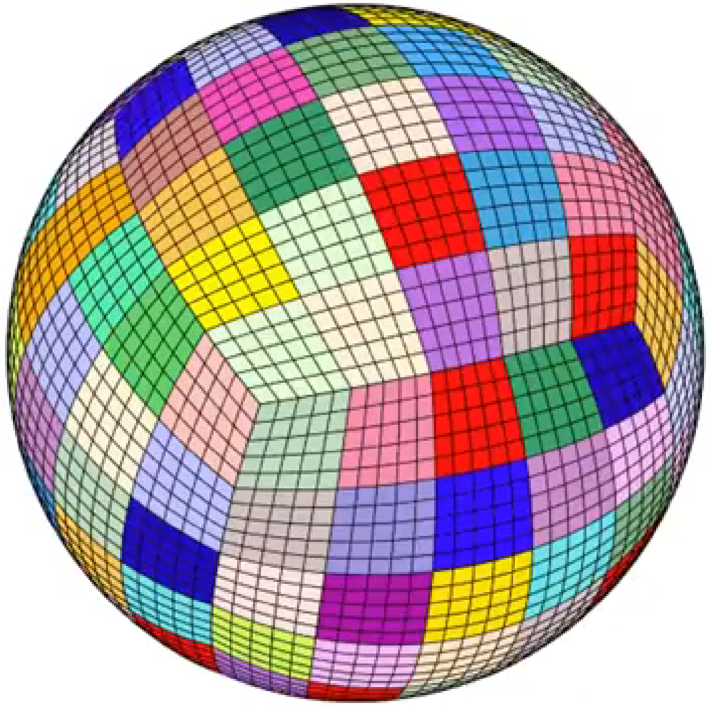
Oak Ridge National Laboratory

Lei, Wenjie, et al. "Global adjoint tomography—model GLAD-M25." Geophysical Journal International 223.1 (2020): 1-21.

klasky@ornl.gov, scottk007@gmail.com



Global Adjoint Tomography and Inversion Workflow



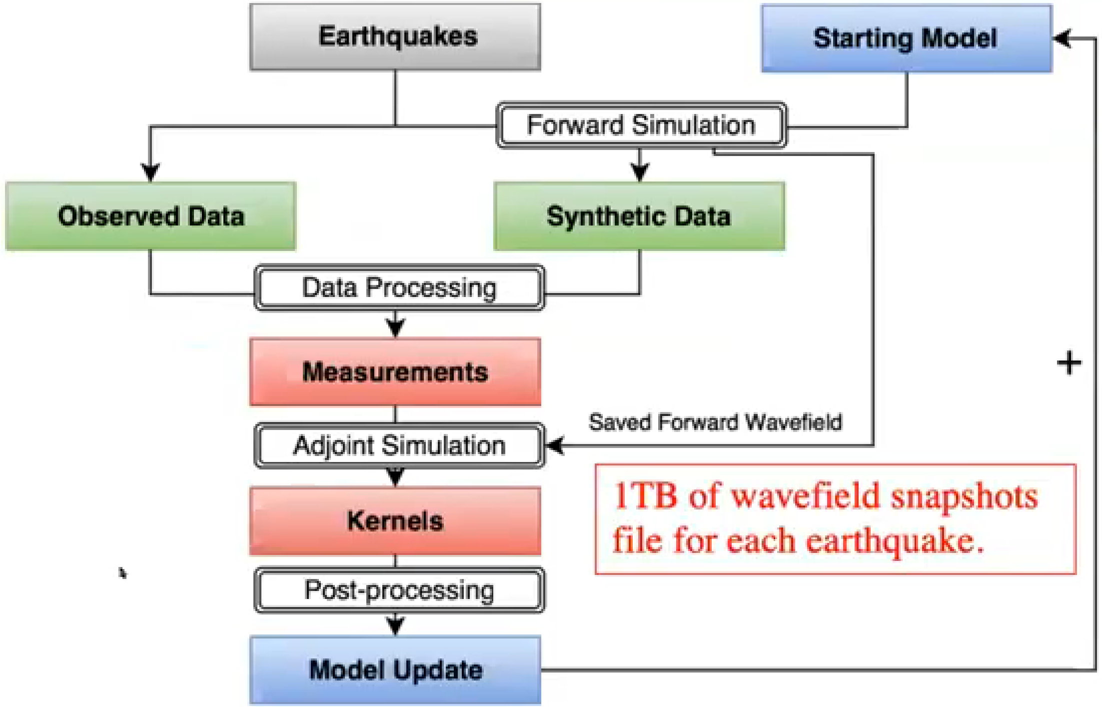
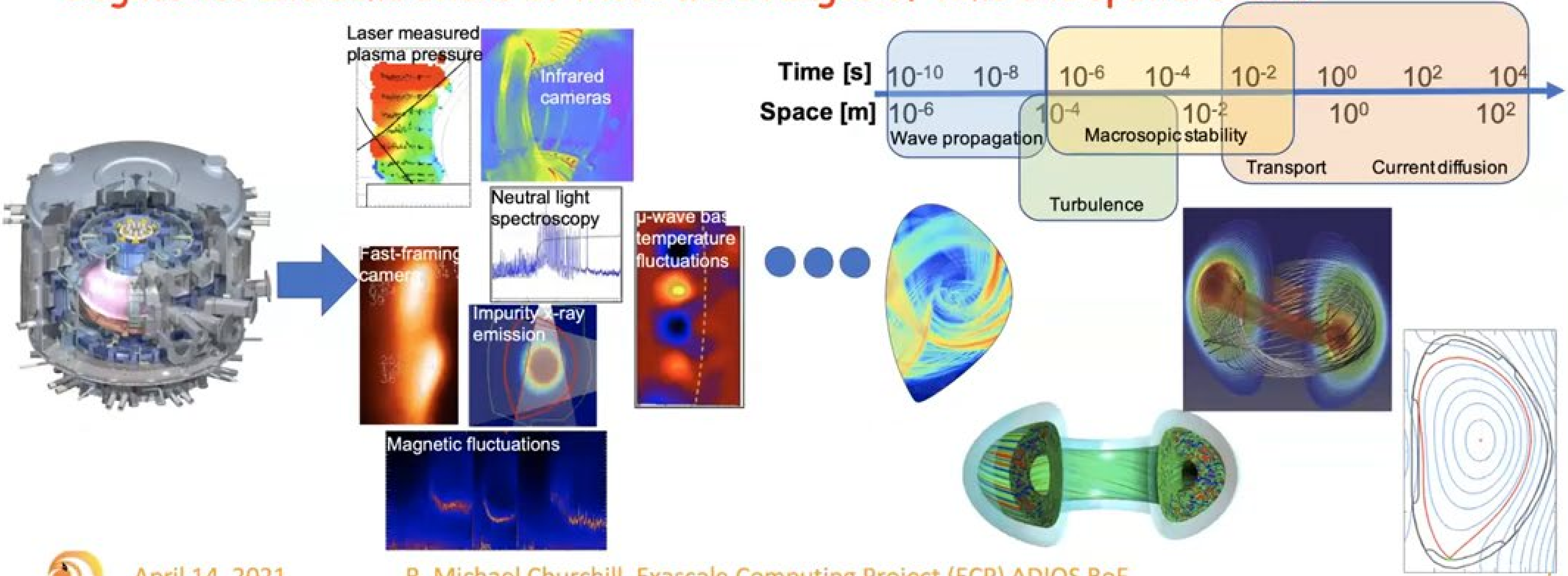


Figure 1. Demo figure of the spectral-finite-element (SEM) mesh of the globe. This figure shows the the Earth is partitioned into finite element. During the forward simulation, we need to save the snapshots of wavefield at certain checkpoints. Degree of Freedom: **~10^10 (10 billion)**

Figure 2: The workflow of adjoint tomography. In the forward simulation, the wavefield snapshots are saved on each mesh points. Those wavefield data will then be read back during the adjoint simulation for wavefield reconstruction. Given our current resolution and simulation length, each earthquake will generate ~1TB of wavefield snapshots file.

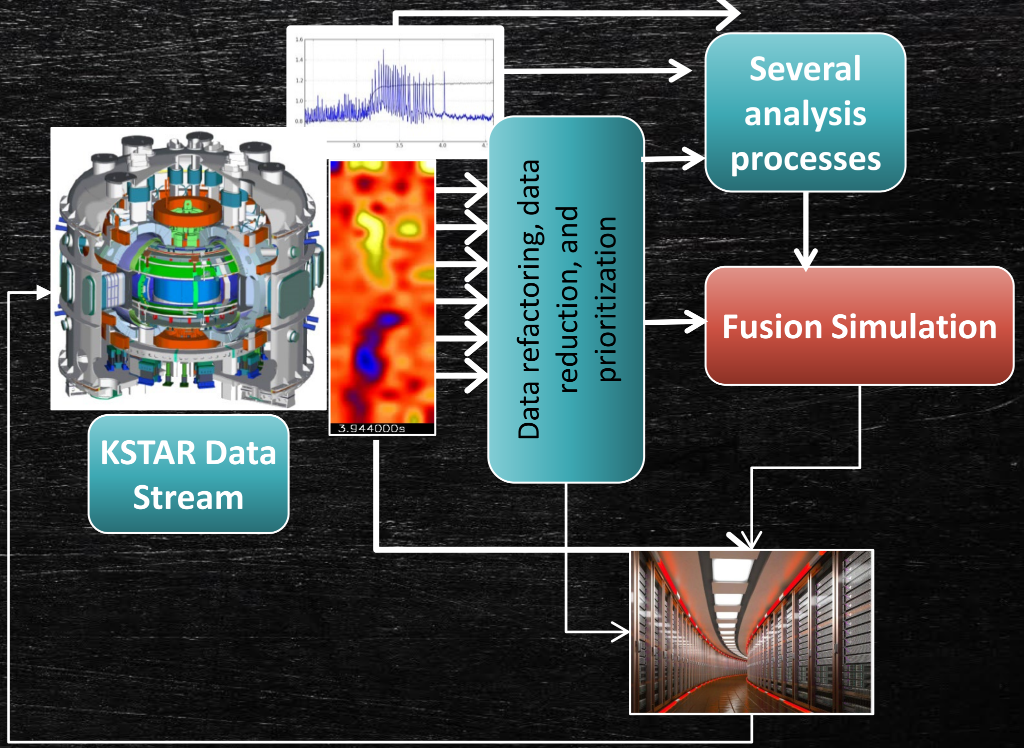
Fusion plasmas have a range of phenomena, manifest over multiple time and spatial scales

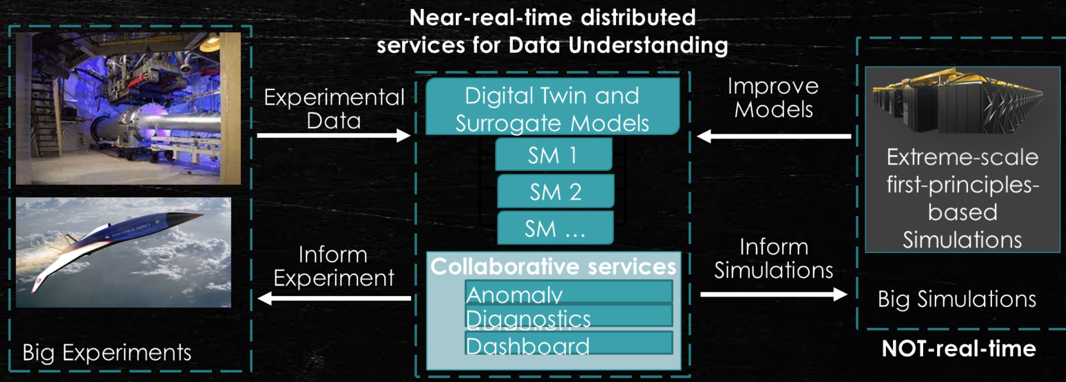
Understanding fusion plasmas requires extracting information from multiple diagnostics and simulations at these wide ranges of time and spatial scales



Future R&D

- Edge to HPC integration for NRT command & control
- Data Management technologies for the convergence of HPC with AI/ML
- Prioritization of refactored data for use in streaming environments
- Accelerate analytics from refactored data
- Further extensions of our theory to bound non-linear QoIs and work with complex unstructured meshes





UMD NNESH

Oak Ridge National Laboratory

klasky@ornl.gov, scottk007@gmail.com

Questions