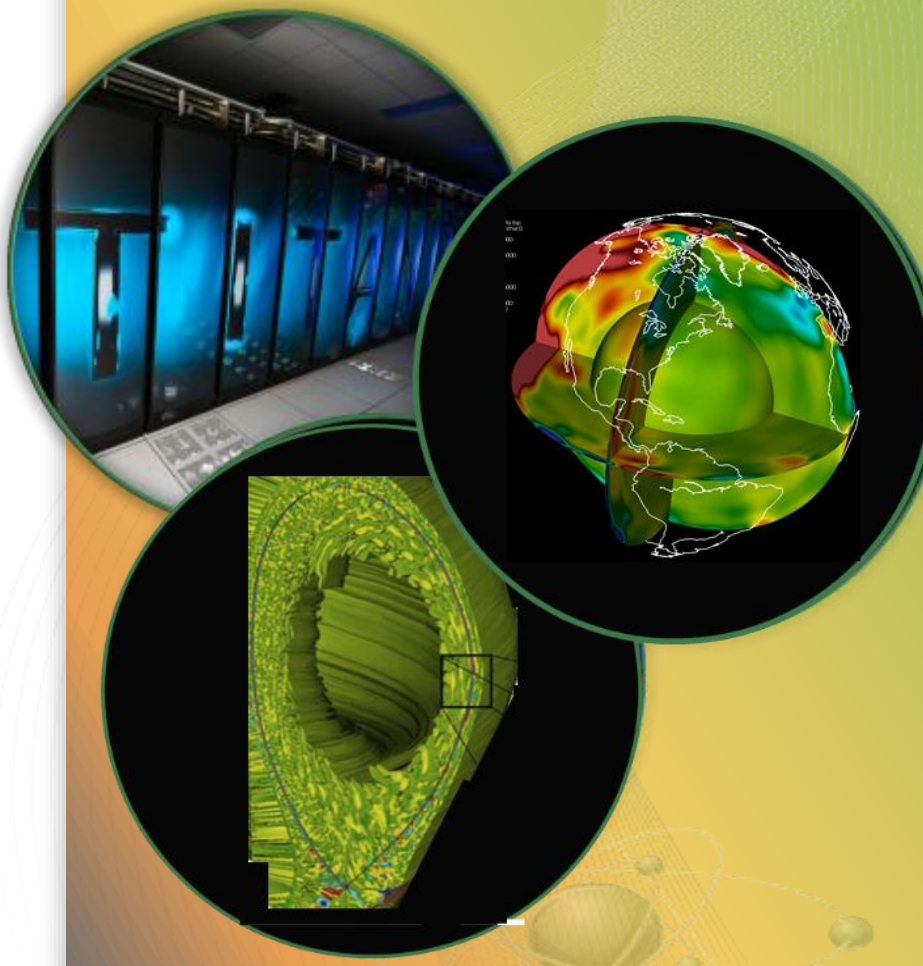
ADIOS
Matthew Wolf, Scott Klasky: ORNL
ASCAC
9/26/2017

- Mark Ainsworth, *Brown
- Jong Choi
- Dongliang Chu
- William Godoy
- Mark Kim
- Scott Klasky, Group Lead
- James Kress
- Tahsin Kurc, *SBU
- Qing Liu, *NJIT
- Jeremy Logan
- Kshitij Mehta
- George Ostrouchov
- Norbert Podhorszki, SDM task lead
- Dave Pugmire, Visual Analytics task lead
- Eric Suchyta
- Lipeng Wan
- Ruonan Wang
- Matthew Wolf, Deputy Group Lead
- Rochelle Womble, Admin

Kitware, ParaTools, PPPL, Sandia, LBNL, ANL, BNL, Oregon, Rutgers, Georgia Tech, ++



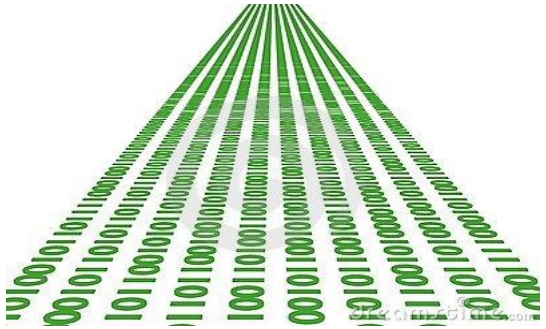
Acknowledgement

- ADIOS has been fully funded by the DOE through
  - ASCR-Research, which we kindly acknowledge the strong interaction from Dr. Lucy Nowell, and secondly from Mr. Richard Carlson who funded the work to allow us to included WAN data staging
  - ASCR-SciDAC, which we kindly acknowledge Drs. R. Laviolette, C. Susut, and many of the wonderful interaction of the ASCR Staff, and especially Drs. Chang, Lin, Tang
  - ASCR-OLCF,ORNL, especially Mr. B. Bland, and Dr. J. Wells, Dr. M. Shankar, Dr. A. Maccabe
  - DOE-ECP, from the ADIOS-ECP project (PI=Klasky), CODAR co-design project (PI=Foster), and Fusion WDM (PI=Bhattacharjee)

- Download
  - https://github.com/ornladios/ADIOS2/
  - https://github.com/ornladios/ADIOS
  - https://www.olcf.ornl.gov/center-projects/adios/

Data reproducibility through self-describing data

- Raw data (byte streams) have little to no use without appropriate description
- Recovering actual information costs scale with the size of the data.
- Self-describing data: Raw data (expensive access) + Metadata (quick access, embedded).



Field	Value
Size	100
Min	0.00002352422
Max	1203213123.4542
Dimensions	{2,50}
Type	float
Name	“pressure”
Units	“KPa”

Question:
What extra annotations do we add to the output to make it more valuable, without “greatly” effecting performance ?

- **Cost:** metadata overhead. **Benefit:** make fast decisions based on metadata.
- Modern software self-describing components:
  - C++ STL containers, Python numpy arrays, Self-describing file formats
- Goal: Write high performance self-describing data streams for in situ + post processing of data

The compute-data gap is a major challenge for HPC

Filesystem/network bandwidth falls behind CPU/memory: Fewer bytes/operation

- **Filesystem/network bandwidth falls behind CPU/memory; most apps become I/O bound**

⇒ Must write many fewer bytes/ops

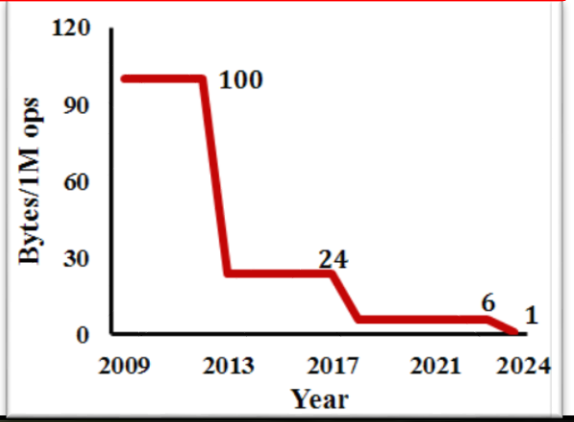
⇒ Online analysis and reduction

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®

Massive increase in computing power and concurrency

⇒ SPMD → MPMD

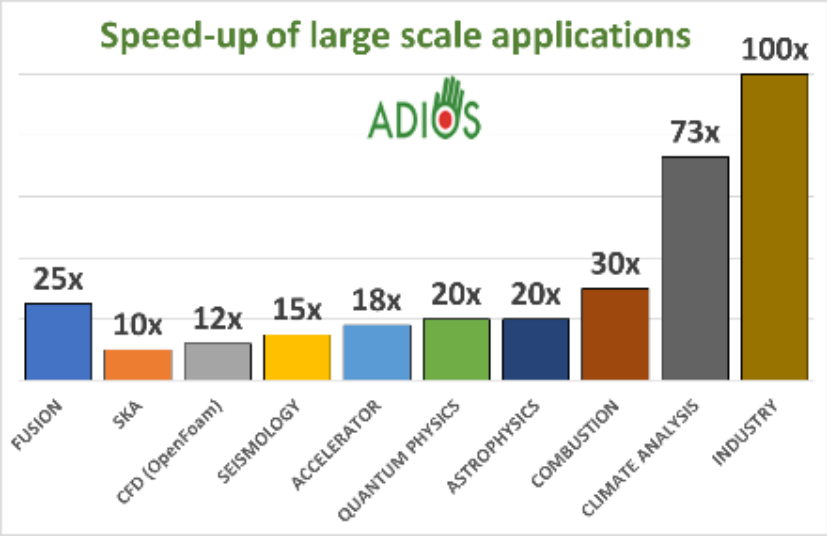
⇒ Increasing need for online coupling



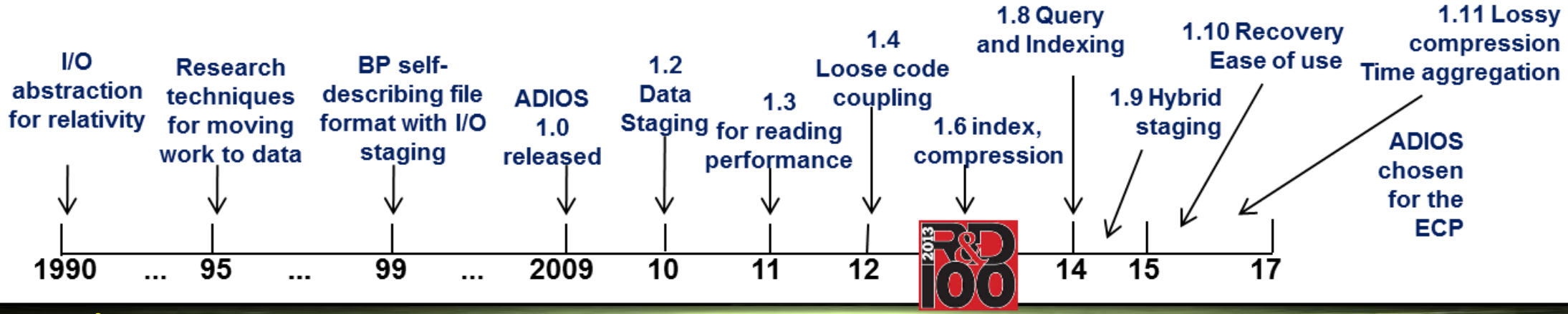
klasky@ornl.gov

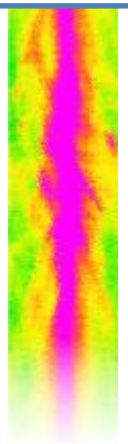
I/O Framework for Data Intensive Science

- Problem
  - Input/output (I/O) performance is severely bottlenecked on HPC systems and experimental instruments because of hardware limitations
    - Applicable to both checkpoint restart and data analysis and visualization
- Solution
  - ADIOS is a DOE I/O framework developed for sustainable I/O on Leadership Class Facilities (LCF) supercomputers
  - ADIOS supports in situ processing of data, which means processing data during the run of a simulation
  - Burst Buffers, a simplified version of data staging, were created as part of ADIOS and are becoming the de-facto standard for exascale I/O
- Impact
  - Over 10X I/O improvement from previous I/O methods on LCF apps
  - Now used by more than 30 LCF application areas, totaling over 1B hours on the LCFs,
  - Outside DOE: Used in Industrial Engineering, Oil Exploration, Computational Fluid Dynamics
  - ADIOS won R&D 100 Award in 2013



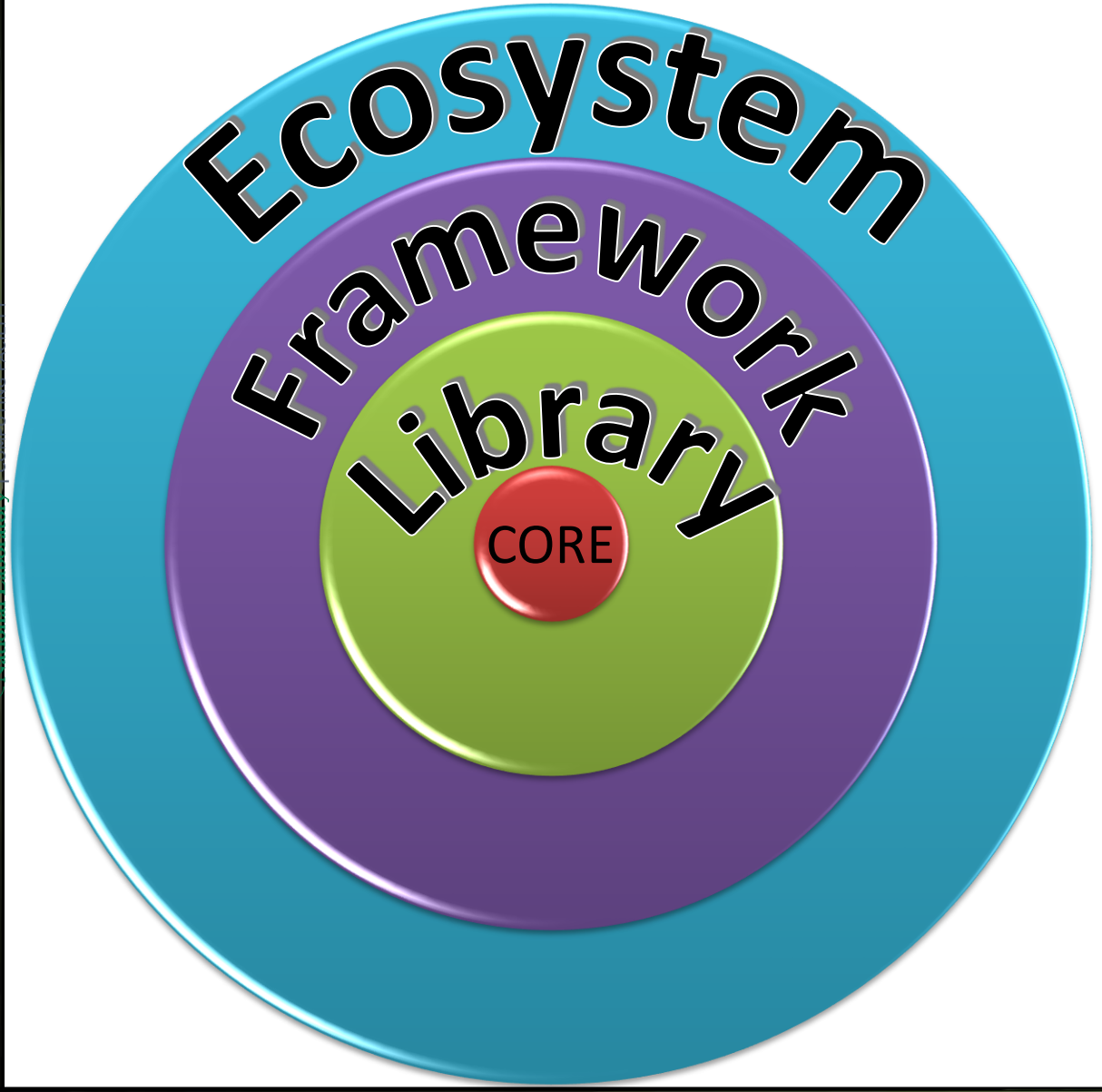
With ADIOS we saw a 20-fold increase in I/O performance compared to our best previous solution which enabled us to study our laser-driven particle **accelerator** from the single-particle level to the full system; which is a game changer for our team", M. Bussmann (PIConGPU)





klasky@ornl.gov

ADIOS



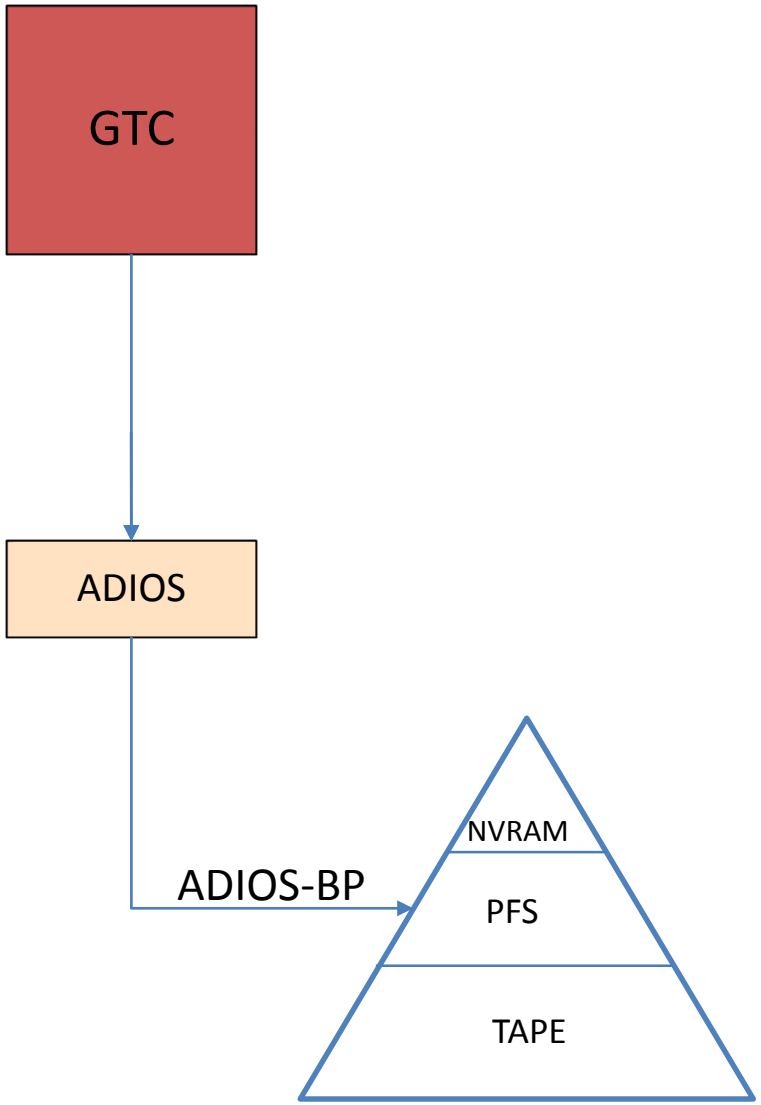
- ADIOS core – provides the basic infrastructure for scientific data
  - BP stream format, Memory Buffering, Data Movement strategies
- ADIOS library - allow "best practice" from external components
  - Engines, Transformations, Indexing, Transports
- ADIOS Framework – allow scientific libraries to be used inside ADIOS
  - Staging libraries, reduction libraries, Indexing libraries, I/O libraries
- ADIOS ecosystem – Allow applications to interact with ADIOS codes/data
  - Analysis- Visualization services, Performance services, Living Miniapps

Application Example

- Application wants to write data at scale

CORE

High Performance, self-describing data output to the file system



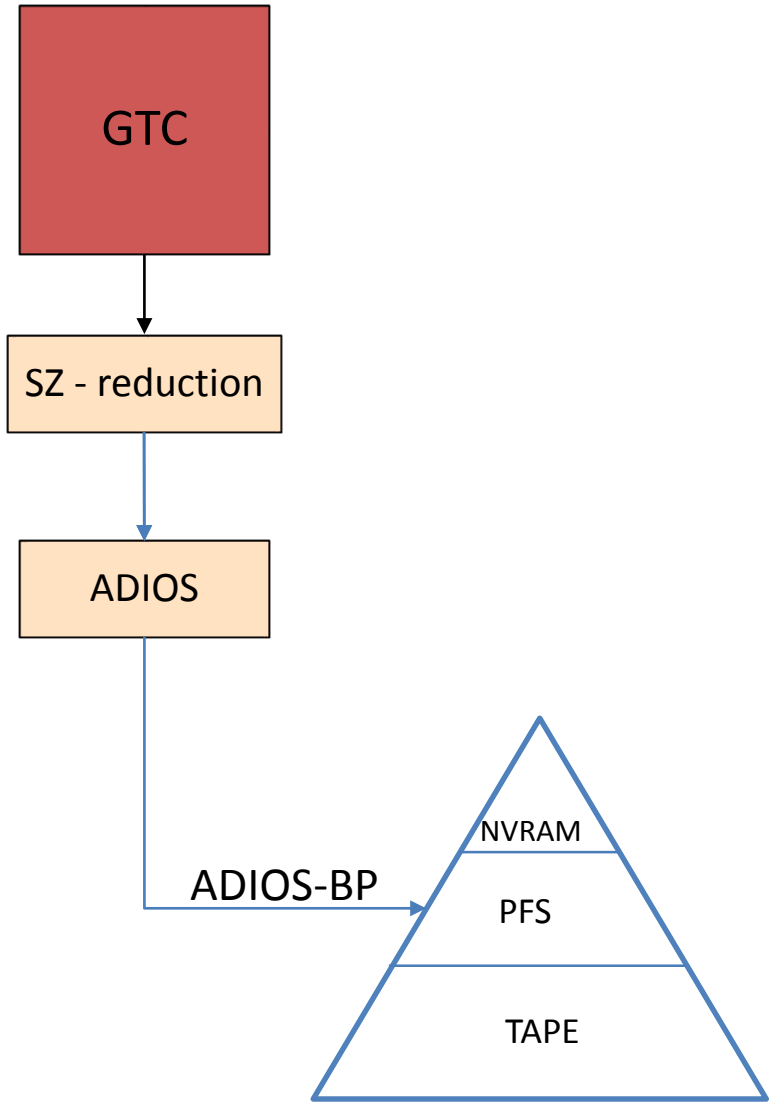
Application Example

- Application wants to write data at scale

CORE High Performance, self-describing data output to the file system

- Applications need to reduce - compress data, index data

Library Encapsulate “best” practices in ADIOS



Application Example

- Application wants to write data at scale

CORE

High Performance, self-describing data output to the file system

- Applications need to reduce - compress data, index data

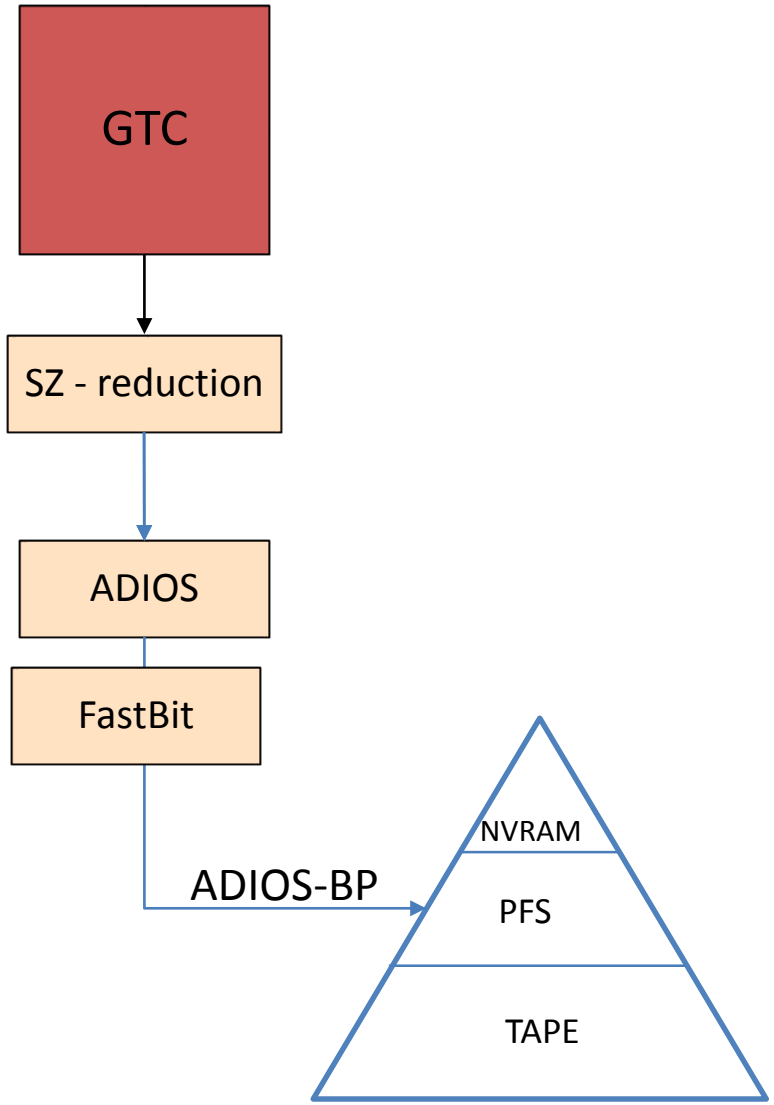
Library

Encapsulate “best” practices in ADIOS

- Applications may want to use “other” compression, indexing technologies

Framework

Allow libraries to be added into ADIOS



Application Example

- Application wants to write data at scale

  CORE High Performance, self-describing data output to the file system

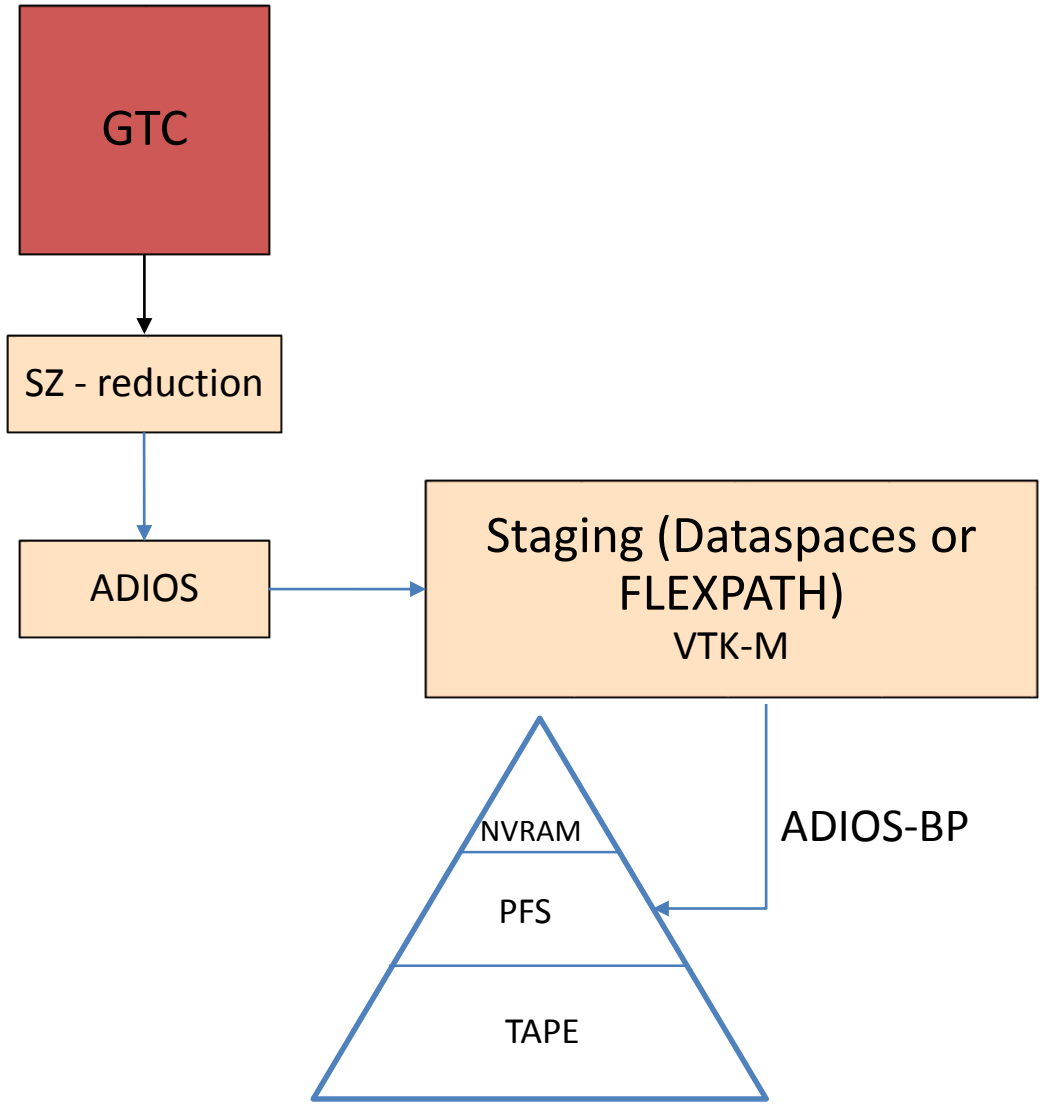
- Applications need to reduce - compress data, index data

  Library Encapsulate "best" practices in ADIOS

- Applications may want to use "other" compression, indexing technologies

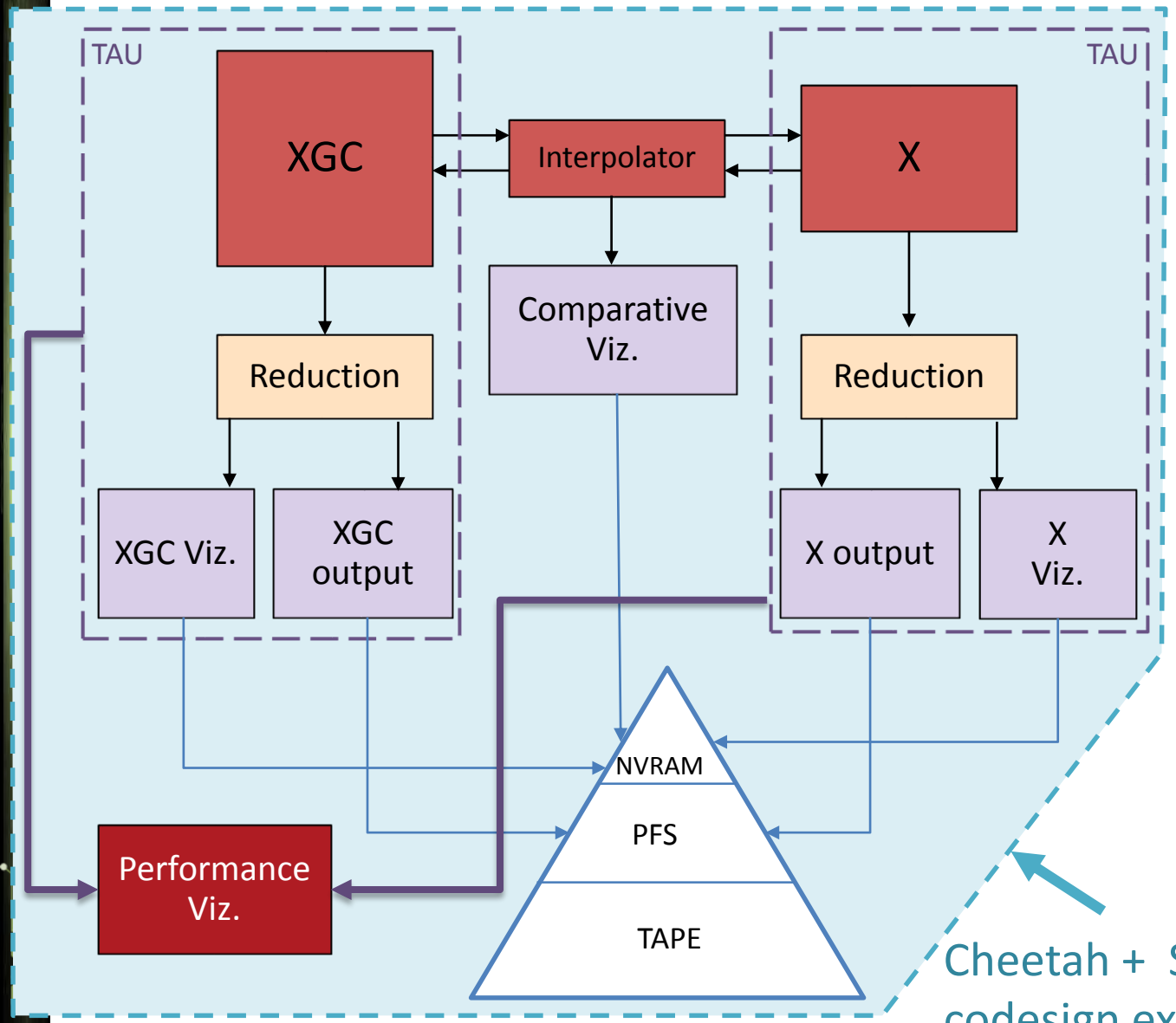
  Framework Allow libraries to be added into ADIOS

- Now we need to analyze and visualize the data (in situ) Ecosystem



klasky@ornl.gov

Driving Collaborative Co-Design with Applications

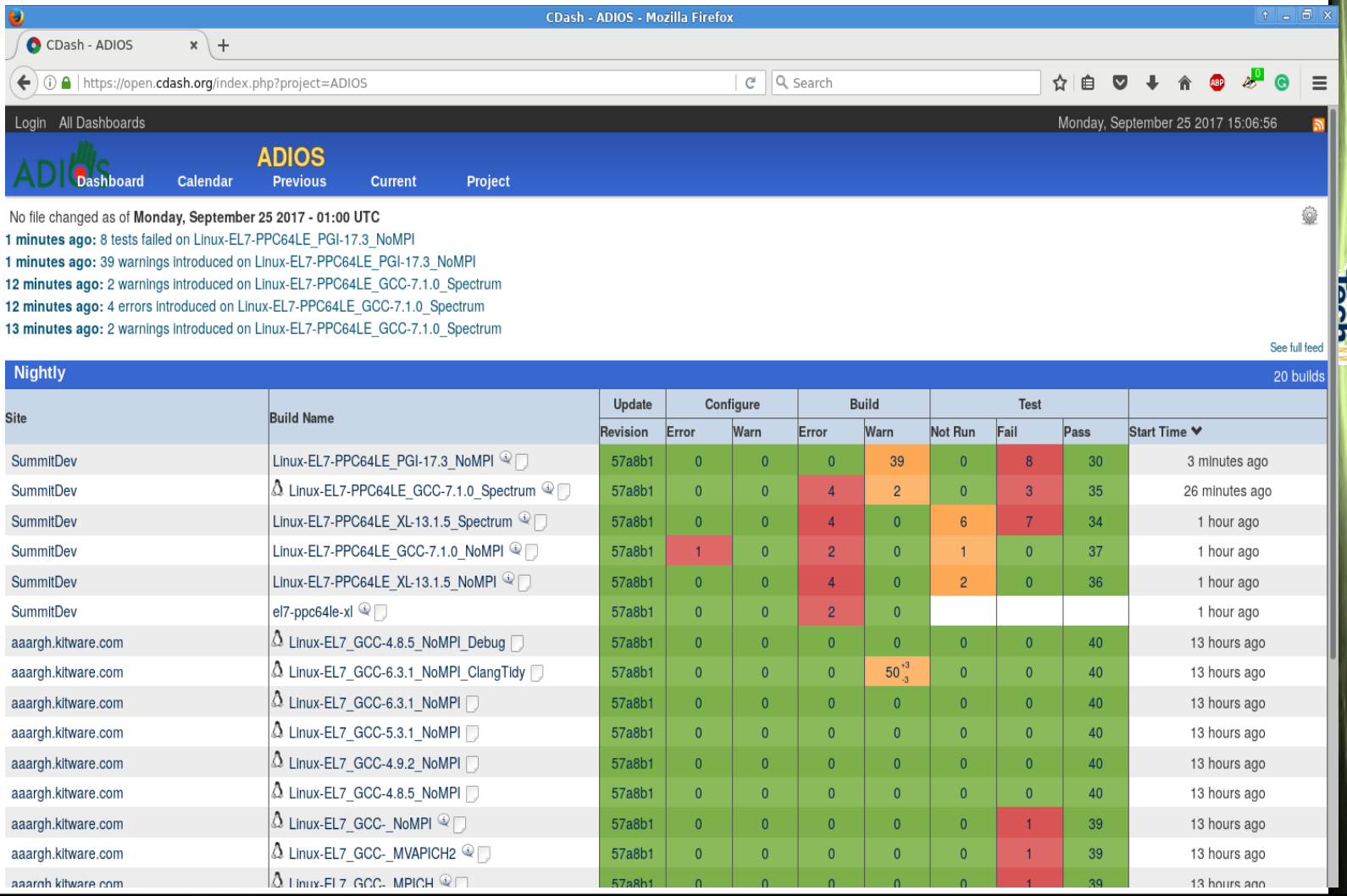


Cheetah + Savanna drive codesign experiments

- <u>Showcase using numerous ECP and other technologies</u>:
  - ADIOS staging (DataSpaces) for coupling
  - Sirius (ADIOS + Ceph) for storage
  - ZFP, SZ, Dogstar for reduction
  - VTK-M services for visualization
  - TAU for instrumenting the code
  - Cheetah + Savanna to test the different configurations (same node, different node, hybrid-combination) to determine where to place the different services
  - Flexpath for staged-write from XGC to storage
  - Ceph + ADIOS to manage storage hierarchy
  - Swift for workflow automation

Engineering ADIOS for Sustainability

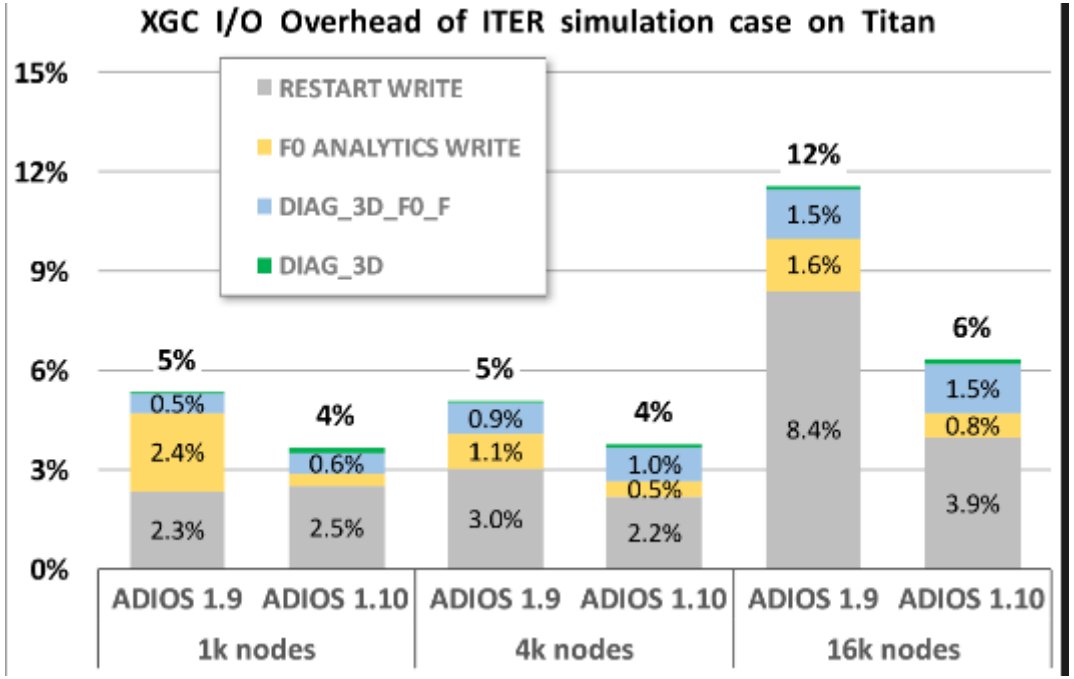
- On-going effort to take what we've learned and build a better stack to support community engagement
- Re-engineering of ADIOS (ADIOS2) from the framework to the inside
  - Make the engagement at the tool/framework level as easy as possible.
  - Build the high performance core out to serve that.
- Uses community practices
  - Continuous integration
  - Github & C++
  - <u>Test-driven development based on applications</u>

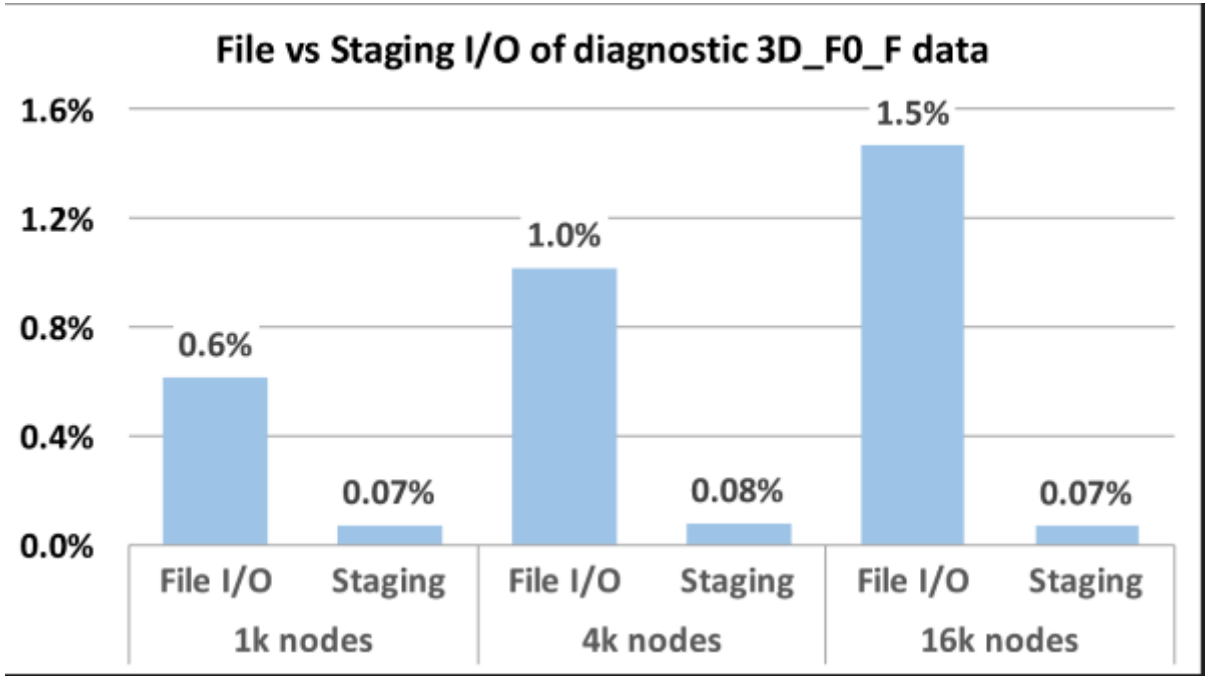


Site	Build Name	Update	Configure		Build		Test			
		Revision	Error	Warn	Error	Warn	Not Run	Fail	Pass	Start Time
SummitDev	Linux-EL7-PPC64LE_PGI-17.3_NoMPI	57a8b1	0	0	0	39	0	8	30	3 minutes ago
SummitDev	Linux-EL7-PPC64LE_GCC-7.1.0_Spectrum	57a8b1	0	0	4	2	0	3	35	26 minutes ago
SummitDev	Linux-EL7-PPC64LE_XL-13.1.5_Spectrum	57a8b1	0	0	4	0	6	7	34	1 hour ago
SummitDev	Linux-EL7-PPC64LE_GCC-7.1.0_NoMPI	57a8b1	1	0	2	0	1	0	37	1 hour ago
SummitDev	Linux-EL7-PPC64LE_XL-13.1.5_NoMPI	57a8b1	0	0	4	0	2	0	36	1 hour ago
SummitDev	el7-ppc64le-xl	57a8b1	0	0	2	0				1 hour ago
aaargh.kitware.com	Linux-EL7_GCC-4.8.5_NoMPI_Debug	57a8b1	0	0	0	0	0	0	40	13 hours ago
aaargh.kitware.com	Linux-EL7_GCC-6.3.1_NoMPI_ClangTidy	57a8b1	0	0	0	50^{+3}_{-3}	0	0	40	13 hours ago
aaargh.kitware.com	Linux-EL7_GCC-6.3.1_NoMPI	57a8b1	0	0	0	0	0	0	40	13 hours ago
aaargh.kitware.com	Linux-EL7_GCC-5.3.1_NoMPI	57a8b1	0	0	0	0	0	0	40	13 hours ago
aaargh.kitware.com	Linux-EL7_GCC-4.9.2_NoMPI	57a8b1	0	0	0	0	0	0	40	13 hours ago
aaargh.kitware.com	Linux-EL7_GCC-4.8.5_NoMPI	57a8b1	0	0	0	0	0	0	40	13 hours ago
aaargh.kitware.com	Linux-EL7_GCC-_NoMPI	57a8b1	0	0	0	0	0	1	39	13 hours ago
aaargh.kitware.com	Linux-EL7_GCC-_MVAPICH2	57a8b1	0	0	0	0	0	1	39	13 hours ago
aaargh.kitware.com	Linux-EL7_GCC-_MPICH	57a8b1	0	0	0	0	0	1	39	13 hours ago

klasky@ornl.gov

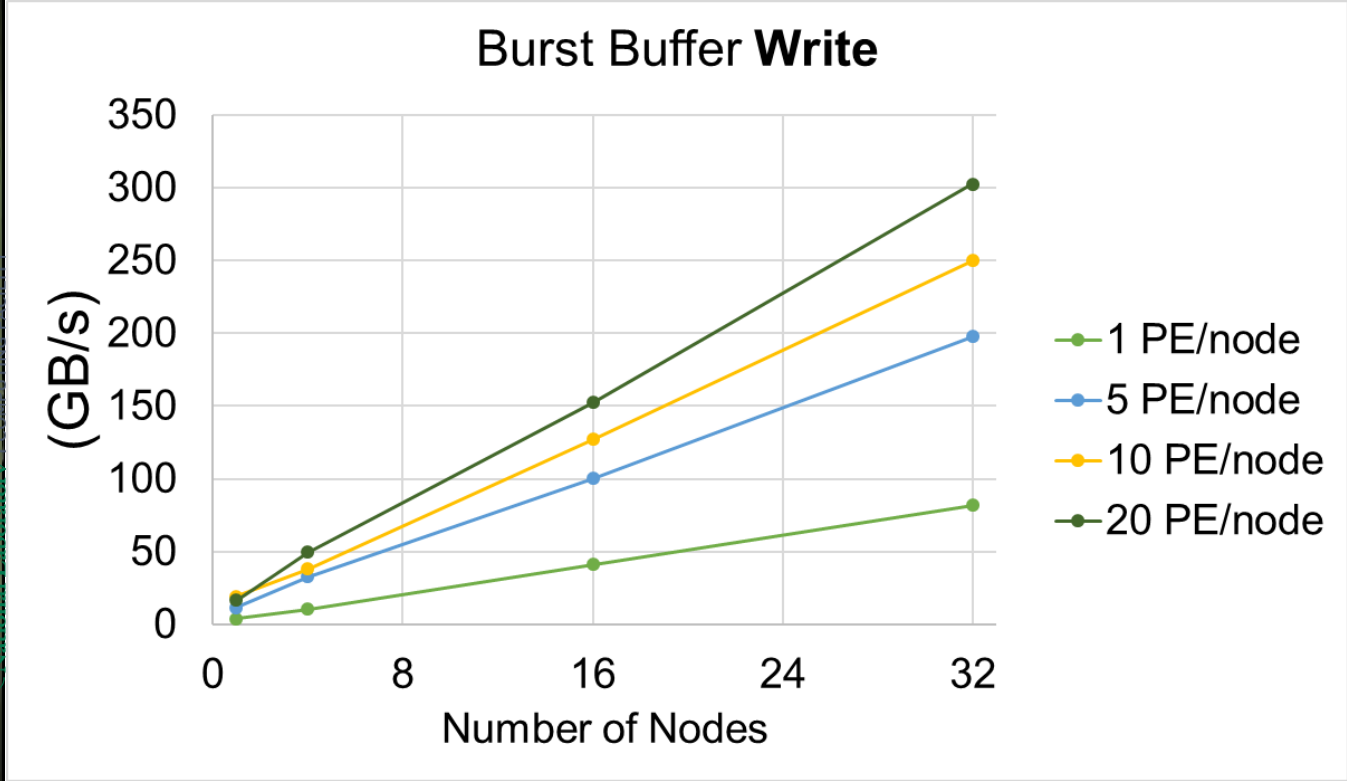
Overheads with XGC

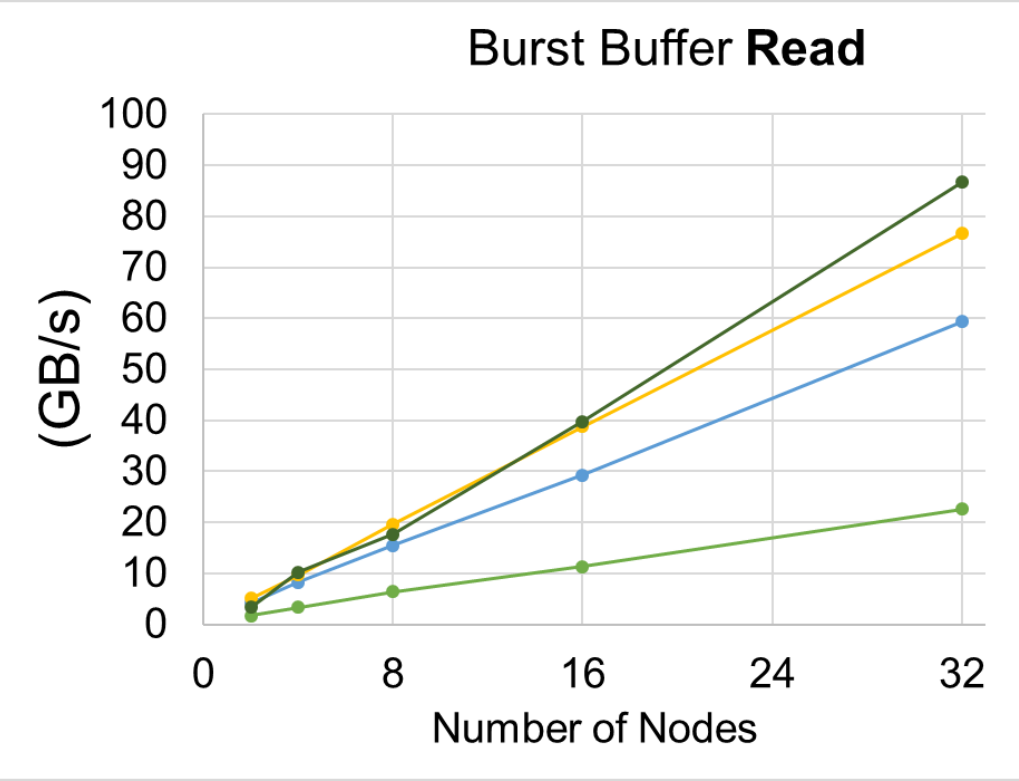




- 1 PB of total output per 24 hours, but really wanted 10 PB/24 hours

ADIOS on Summitdev Burst Buffer

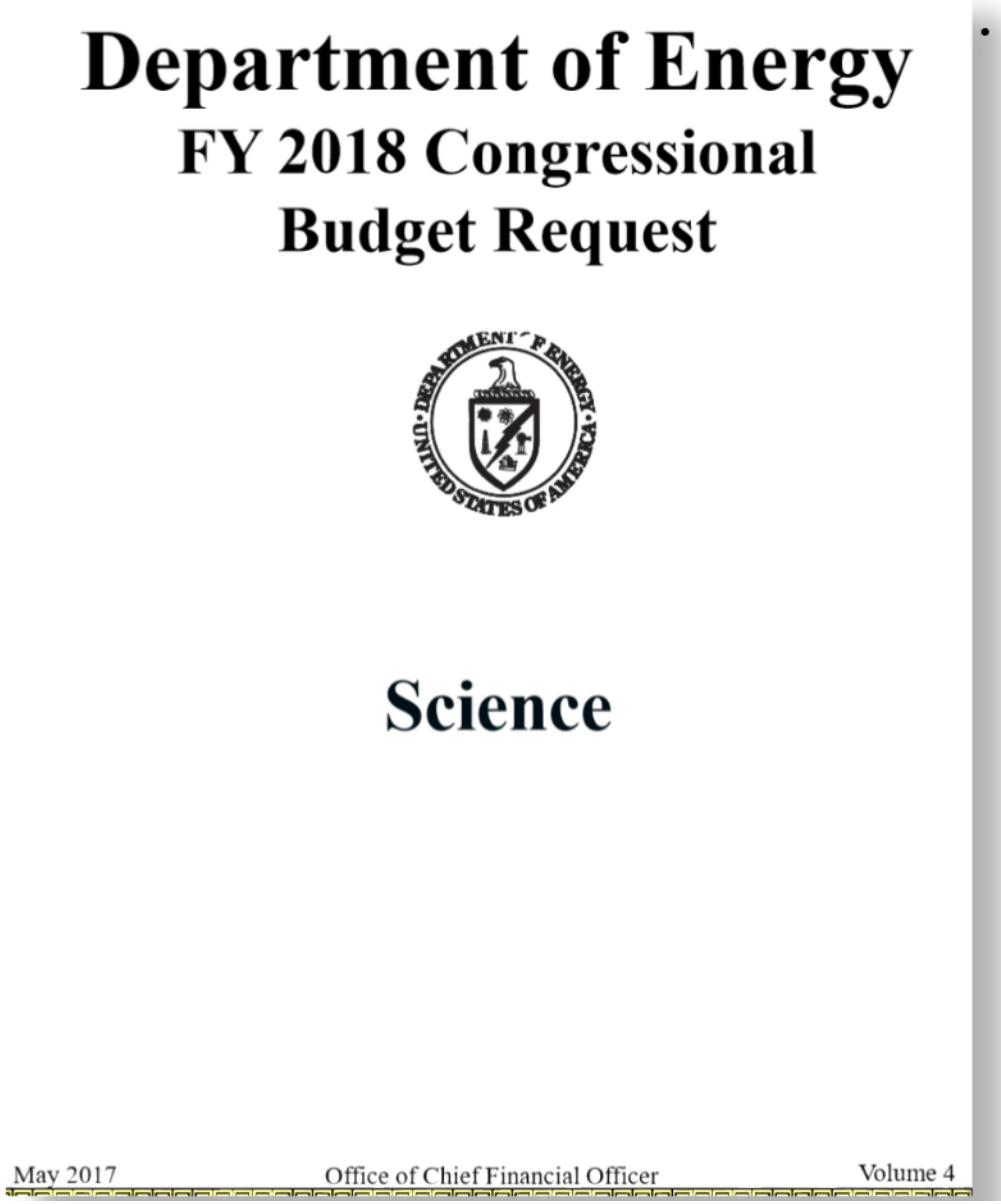




- 1 GB per PE written/read – in 20 PE/node case that means 20GB/node
- caveat: Plenty of free RAM available on each node

Further impact: OpenFOAM CFD simulations

Yi Wang, Karl Meredith – FM Global
S. Klasky, N. Podhorszki - ORNL



- Reducing the Damage Caused by Industrial Fires. Warehouse fires are the leading cause of commercial property damage, responsible for 40% of all industry property loss at a cost of approximately $188 million per year. Understanding how fires spread has the potential to save both business owners and insurance companies hundreds of millions of dollars. However, some of the most destructive fires – those that take place in mega-warehouses with ceilings up to 100 ft. high and a footprint in excess of 100,000 sq. ft.– are among the most difficult to study because they cannot be replicated in a test facility. To solve this problem, one of the world's largest commercial and industrial insurance companies partnered with the Oak Ridge Leadership Computing Facility to adapt an open-source fluid dynamics code to include the complex processes that occur during an industrial fire, including soot formation and sprinkler spray dynamics. After running their high resolution FireFOAM code on the Oak Ridge Leadership Computing Facility's Titan machine to learn how to stack storage boxes on pallets to impede the spread of horizontal flames, **the team incorporated the SciDAC-developed Adaptable I/O System (ADIOS) into FireFOAM to improve its efficiency in moving data on and off the supercomputer**. The new and improved code is now being used to simulate other commodities stored in warehouses, starting with large paper rolls. Both the results and the code are shared publicly to promote the improvement of fire protection standards across industry.