RAPIDS **Resource and Application Productivity through computation, Information, and Data Science**

Application Engagement & Community Outreach
Tiger Teams
Liaisons
Outreach
Data Understanding
Platform Readiness
Scientific Data Management

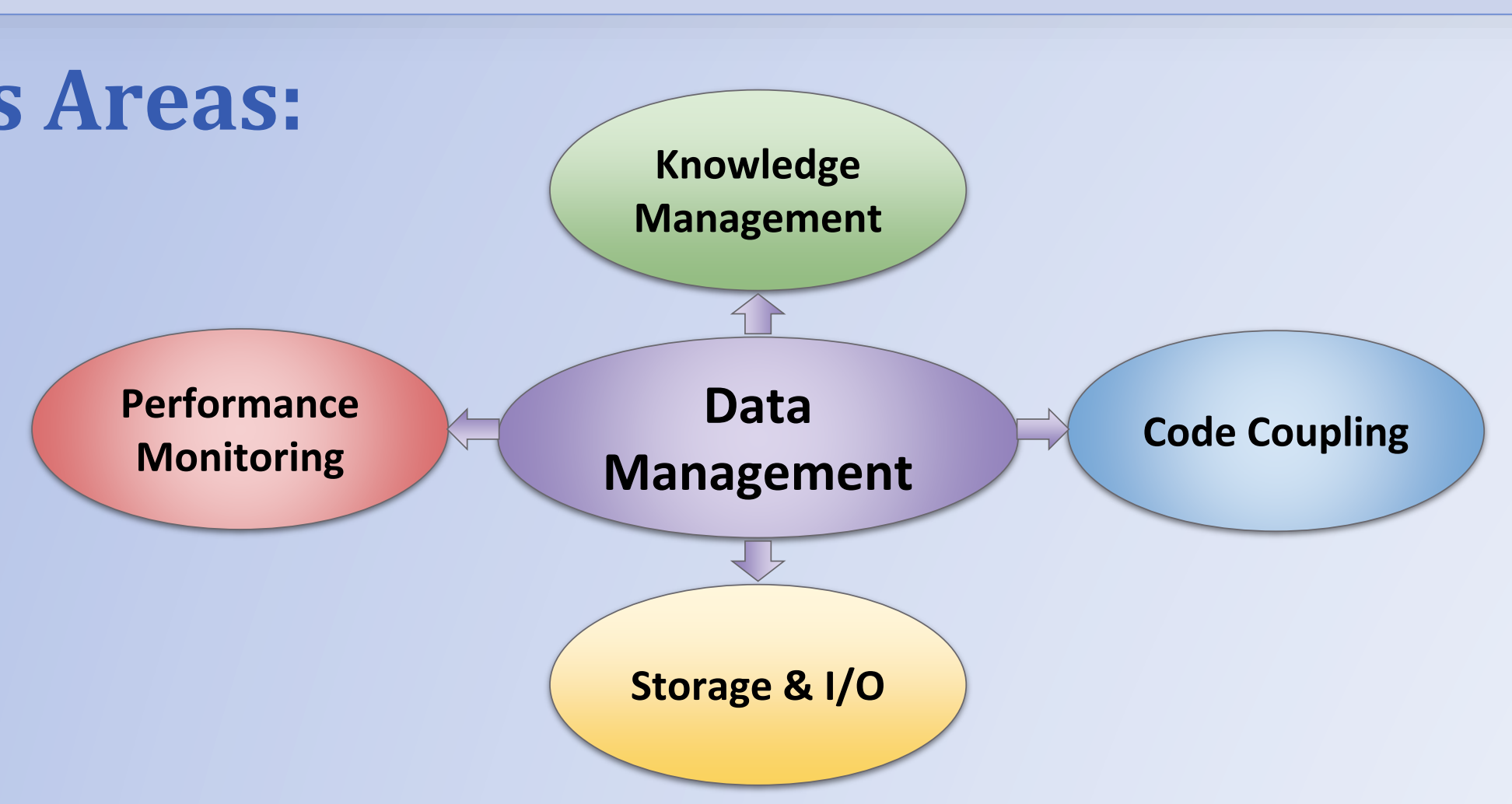
Scientific Data Management: I/O libraries, coupling, knowledge management

S. Klasky, J. Wu, N. Fortner, R. Latham, W. Liao, K. Mehta, N. Podhorszki, K. Huck, A. Sim, P. Davis, M. Parashar, B. Geveci

Motivation

- **Challenges for I/O on for SciDAC applications running on HPC resources**
  - File system/network bandwidth is not keeping up with computing power
  - Data is difficult to organize and find from in a scientific campaign
- **Approach**
  - Optimize the I/O layer for SciDAC applications (Storage & I/O)
  - Monitor and save the performance data (Performance Monitoring)
  - Organize all of the data during a campaign and allow efficient queries (Knowledge Management)
  - Allow synchronous and asynchronous (Code Coupling)

Focus Areas:



- **Storage and I/O**
  - Fast I/O to read/write to memory & storage layers
  - Self Describing I/O formats
- **Performance Monitoring**
  - Capture and understand I/O performance
- **Knowledge Management**
  - Provenance Capture
  - Manage and query data across a campaign
- **Code coupling (couple sims to)**
  - analysis
  - visualization
  - simulations
  - experiments

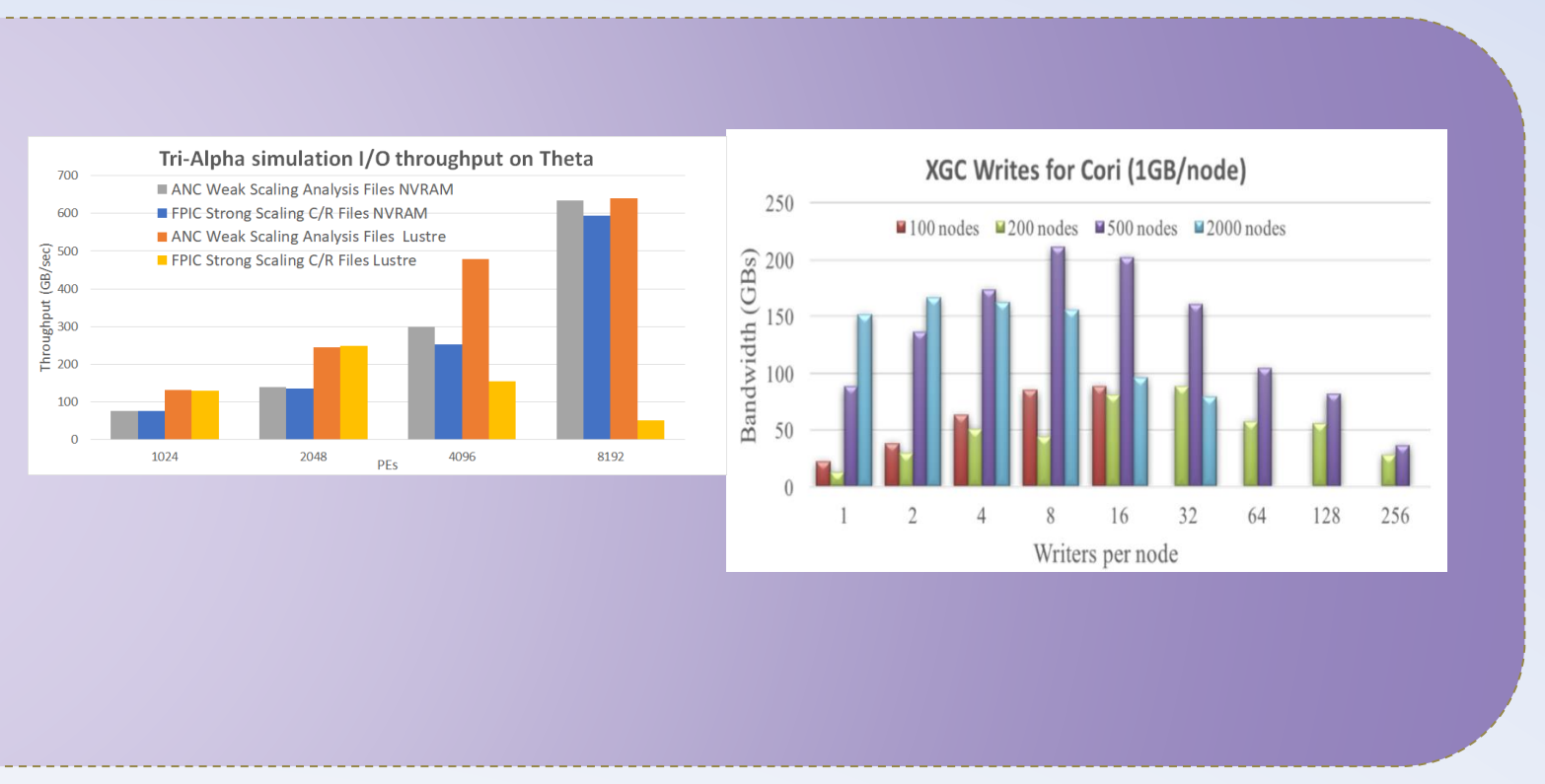
Application and Facility Highlights

- **Accelerating Fusion Scientific Discovery through advanced code coupling.**
  - The fusion community created a new tight-coupling scheme to enable the strong coupling of two large-scale kinetic turbulence codes.
  - The coupling was accomplished via ADIOS and Dataspaces, coupling the fusion codes along with using TAU, VTK-M viz services from RAPIDS
  - Julien Dominski, Seung-Hoe Ku, Choong-Seock Chang, Jong Choi, Eric Suchyta, Scott Parker, Scott Klasky, Amitava Bhattacharjee, A tight-coupling scheme sharing minimum information across a spatial interface between gyrokinetic turbulence codes, arXiv preprint arXiv:1806.05251, 2018/6/13; accepted to the physics of plasmas, (2018).
- **Simulation on Titan Helps Scientists Fine-Tune Laser Interactions to Advance Cancer Treatments**
  - Today scientists are focusing on leveraging heavy particle to target cancers, since these particles can reach deep tumors and reduce the amount of radiation to healthy tissues.
  - Scientist need the ability to model compact particle accelerators on HPC resources to understand how to build this device for medical facilities.
  - Using the Titan supercomputer at Oak Ridge National Lab (ORNL), a team led by Michael Bussmann (HZDR) performed a simulation of a novel laser target that describes the physics behind the acceleration using the PIConGPU code. The code produced 4 PB of data and was accurately able to reproduce the physical experiment, and the RAPIDS contributions were in reducing the I/O time from 20 minutes/time step for output to 30 seconds.
  - Schellhammer, Sonja M., Sebastian Gantz, Armin Lühr, Bradley M. Oborn, Michael Bussmann, and Aswin L. Hoffmann. "Experimental verification of magnetic field induced beam deflection and Bragg peak displacement for MR-integrated proton therapy." Medical physics (2018).
- **ADIOS is helping the GTC code manage large-scale data and perform online analysis and visualization**
  - Recent RAPIDS advancements in the ADIOS framework, using SST, are being integrated into the GTC code, to allow for the efficient coupling of visualization services into GTC.
- **Petascale I/O in Seismic Tomography Workflow**
  - The most detailed 3D model of Earth's interior showing the entire globe from the surface to the core–mantle boundary, a depth of 1,800 miles was run with the SPECFEM3D_GLOBE code on the titan supercomputer.
  - The first global seismic model where no approximations were used to simulate how seismic waves travel through the Earth. The data sizes required for processing are challenging even for leadership computer facilities.
  - RAPIDS researchers were able to reduce their I/O time by over 20X allowing over 1 PB of data to be written/read in a 12 hour simulation on all of Titan.
  - https://www.olcf.ornl.gov/2017/03/28/a-seismic-mapping-milestone

Software Products

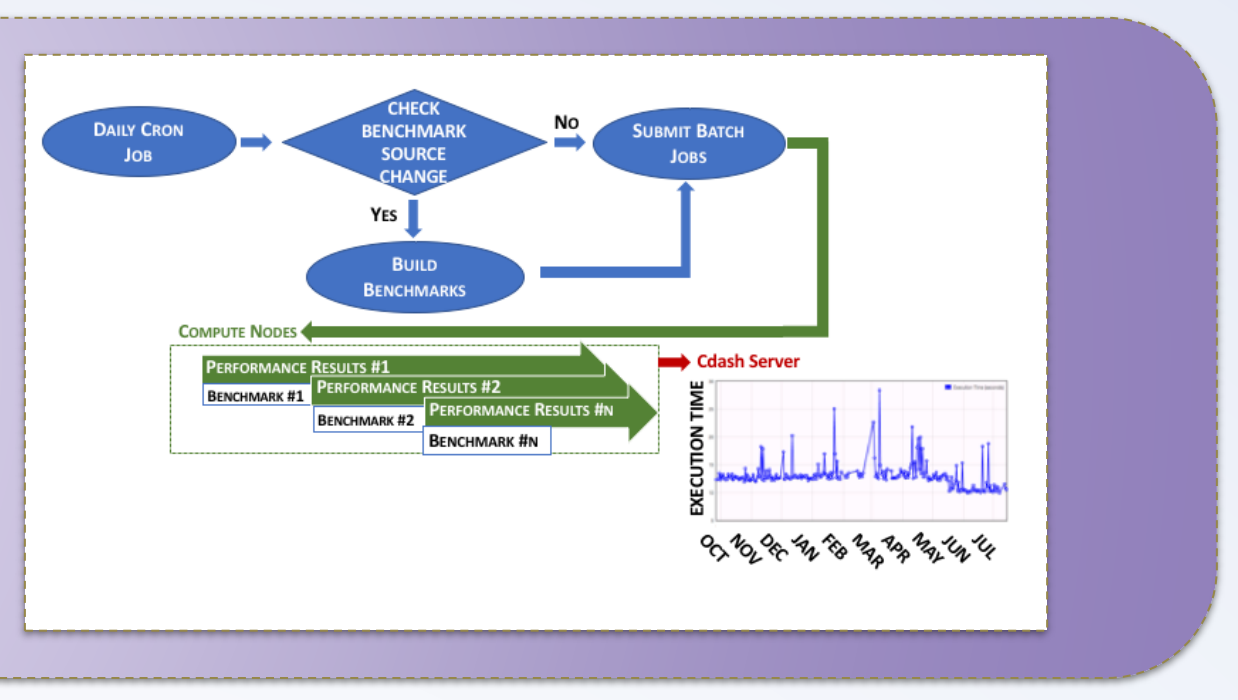
ADIOS

- A community I/O framework which acts as the "glue" for codes to communicate with one another or to storage
- Incorporates the "state of the art" I/O techniques for self describing data for C/R, analysis, visualization and in situ data movement between codes
- **RAPIDS work**
- Develop new I/O drivers to coordinate data movement in the storage layers
- Optimize I/O drivers for patterns from our Partner applications and libraries
- https://github.com/ornladios/ADIOS2



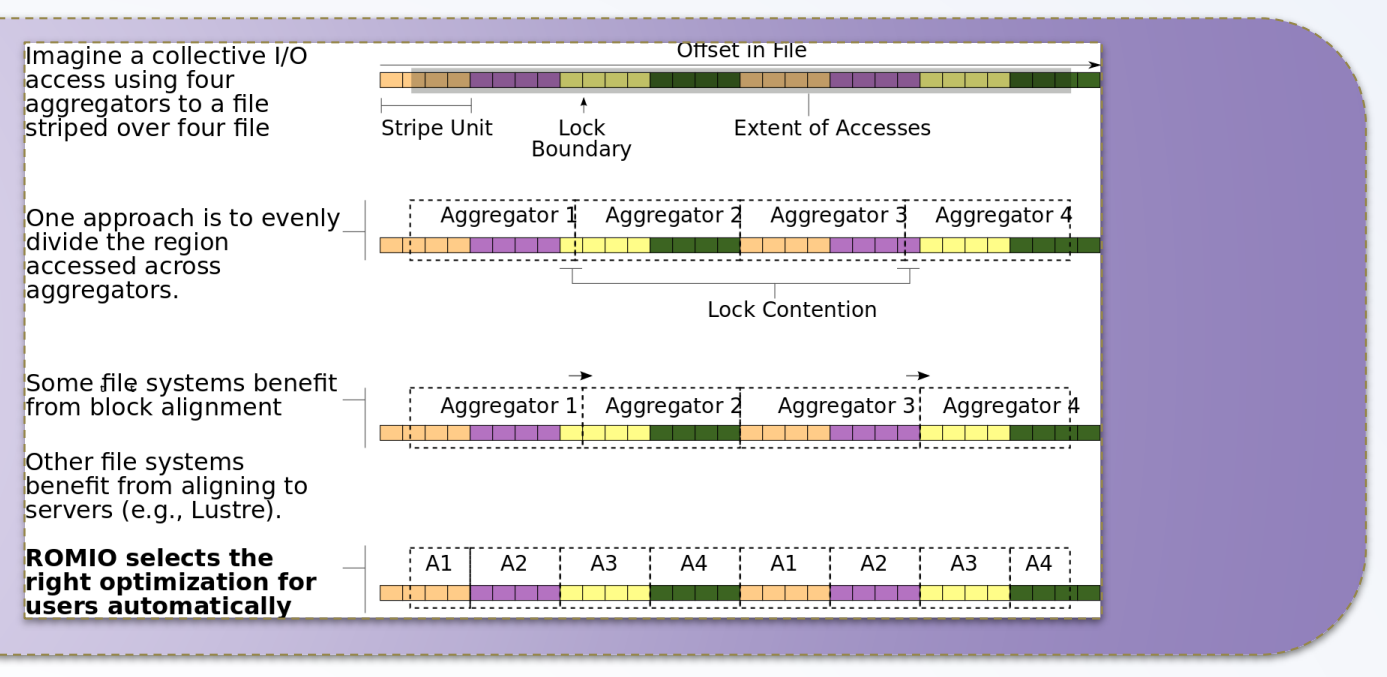
HDF5

- **HDF5** is a is a data model, parallel I/O library, and file format for storing and managing data, hdfgroup.org
- HDF5 is Flexible, self-describing, portable, high performance
- **RAPIDS work**
- Education and training of next-generation researchers on using HDF5 in HPC
- Development of I/O kernel benchmarks exercising common DoE software/hardware configuration
- Targeted metadata performance improvements based on results of I/O benchmarks



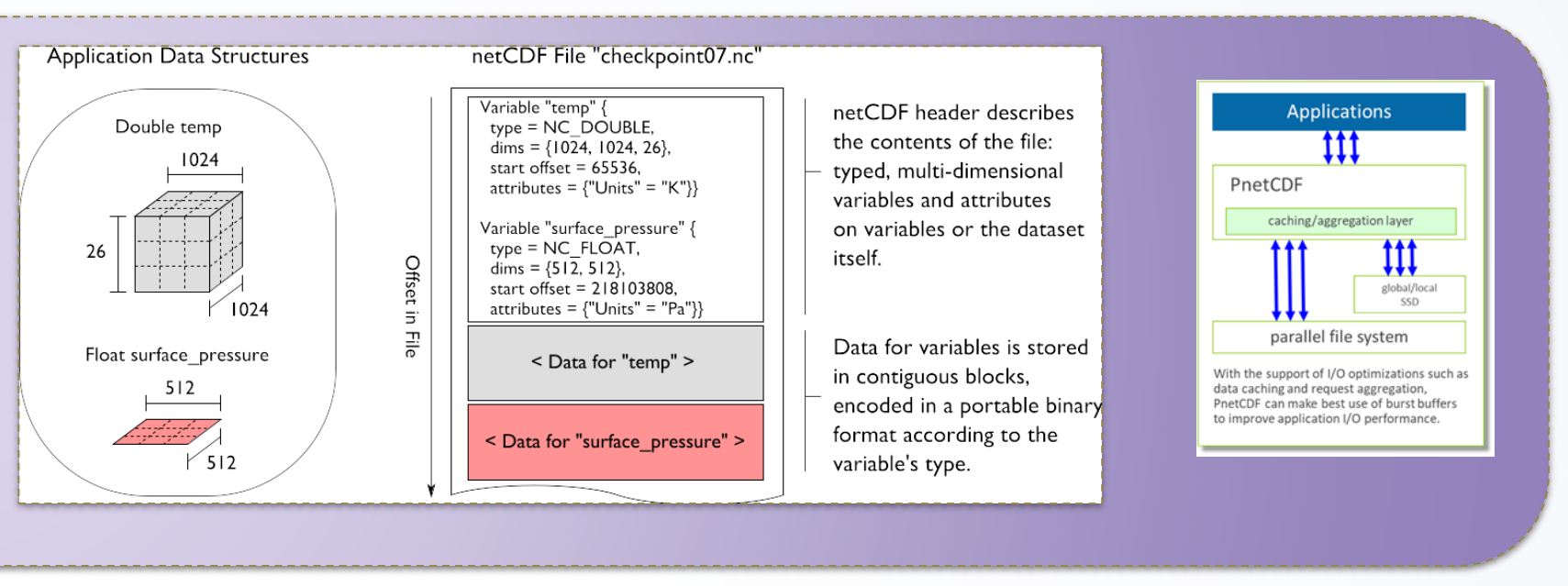
ROMIO

- ROMIO provides a widely-deployed implementation of the I/O routines in MPI
- Incorporated into many vendor MPI implementations (Intel MPI, HPE, Cray)
- Also available as part of MPICH (https://github.com/pmodels/mpich/)
- **RAPIDS work**
- Working with industry partners to incorporate improvements in LUSTRE


Imagine a collective I/O access using four aggregators to a file striped over four file
Offset in File
Stripe Unit
Lock Boundary
Extent of Accesses
One approach is to evenly divide the region accessed across aggregators.
Aggregator 1
Aggregator 2
Aggregator 3
Aggregator 4
Lock Contention
Some file systems benefit from block alignment
Other file systems benefit from aligning to servers (e.g., Lustre).
ROMIO selects the right optimization for users automatically
A1
A2
A3
A4

PnetCDF

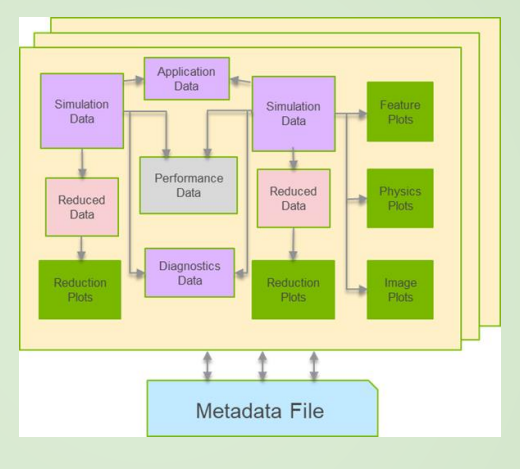
- Parallel I/O library for accessing NetCDF files on HPC systems
- v1.10.0 released on 2018-07-02, contains a new burst buffering feature and progressive performance tuning
- http://cucis.ece.northwestern.edu/projects/PnetCDF/download.html


Application Data Structures
netCDF File "checkpoint07.nc"
Double temp
1024
26
1024
Float surface_pressure
512
512
Offset in File
< Data for "temp" >
< Data for "surface_pressure" >
netCDF header describes the contents of the file: typed, multi-dimensional variables and attributes on variables or the dataset itself.
Data for variables is stored in contiguous blocks, encoded in a portable binary format according to the variable's type.
Applications
PnetCDF
caching/aggregation layer
global/local SSD
parallel file system
With the support of I/O optimizations such as data caching and request aggregation, PnetCDF can make best use of burst buffers to improve application I/O performance.

```
Variable "temp" {
  type = NC_DOUBLE,
  dims = {1024, 1024, 26},
  start offset = 65536,
  attributes = {"Units" = "K"}}
Variable "surface_pressure" {
  type = NC_FLOAT,
  dims = {512, 512},
  start offset = 218103808,
  attributes = {"Units" = "Pa"}}
```

FastBit +

- Organize and quickly find records across files generated & used during a scientific campaign
- **Research and Development** for novel techniques to maintain the relationships of input and output from source code, workflows, images, input data, and output data
- **Algorithmic research**: develop new indexing techniques



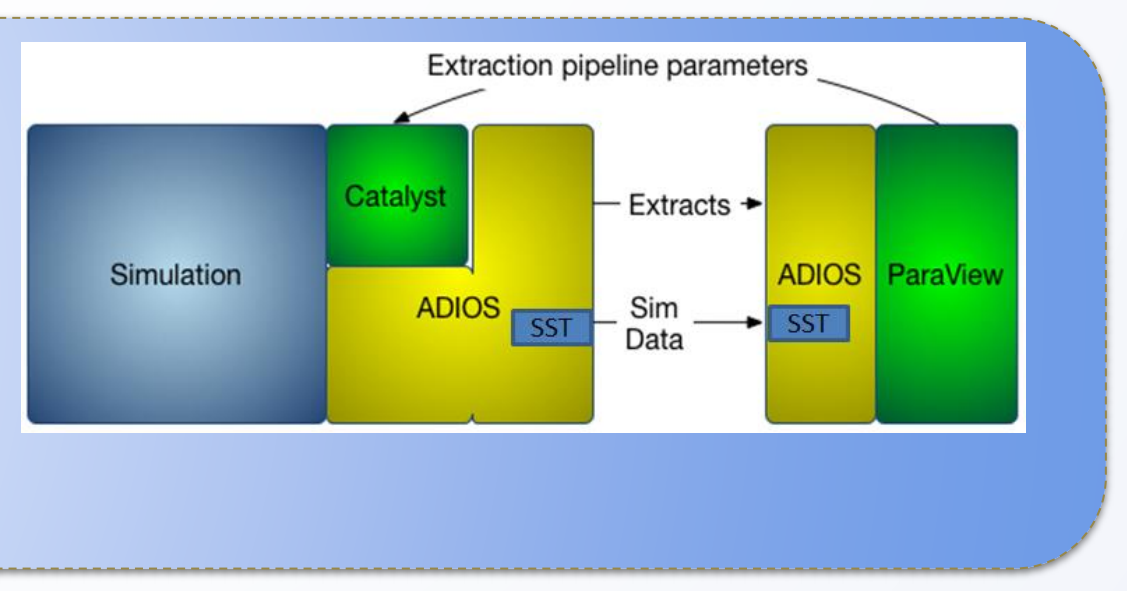
Scan
Query+Read
Query
Time (sec)
# Hits

RAPIDS Work

- Utilizing FastBit to create a novel knowledge management tool for managing data across all of the files of a scientific campaign including the metadata for fast indexing, and efficient queries
- Query across source code, input data, workflow traces, output data

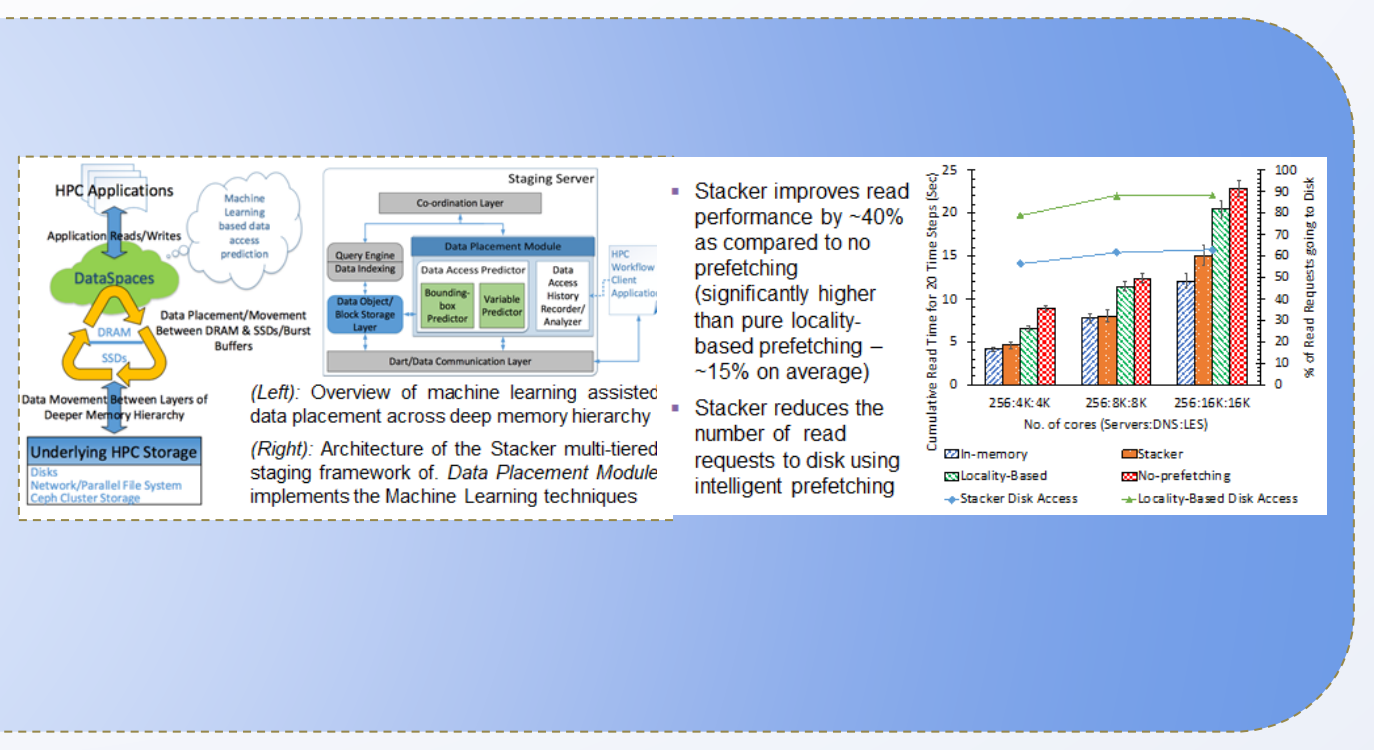
ParaView Catalyst

- Catalyst is a state of the art in situ framework for data analysis and visualization. It is based on VTK and ParaView.
- Synchronous in situ data analysis and extraction through Catalyst.
- Asynchronous data transfer and interactive analysis / visualization through ParaView.
- Interactive management of in situ analysis parameters.



Dataspaces (SST2)

- Dataspaces is an In-memory storage distributed across set of cores/nodes, using RAM, NVRAM/Burst Buffers
- In-staging data processing, querying, sharing, and exchange
- Virtual shared-space programming abstraction
- Provides an efficient, high-throughput/low-latency asynchronous data transport
- **RAPIDS work**
- Make it easier to use and run with DS/SST2
- Create a series of examples/benchmarks to make it easier to work with

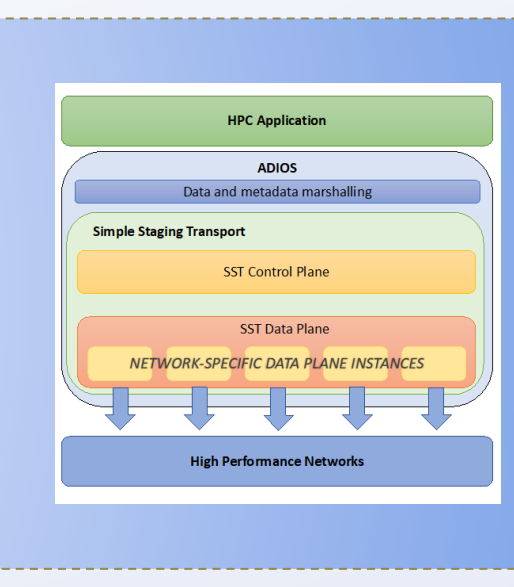

HPC Applications
Application Reads/Writes
DataSpaces
Machine Learning based data access prediction
Staging Server
Co-ordination Layer
Data Placement Module
Data Placement/Movement Between DRAM & SSDs/Burst Buffers
Data Movement Between Layers of Deeper Memory Hierarchy
Underlying HPC Storage
No. of cores (Servers:DNS:LES)
In-memory
Stacker
Locality-Based
No-prefetching
Stacker Disk Access
Locality-Based Disk Access

(*Left*): Overview of machine learning assisted data placement across deep memory hierarchy (*Right*): Architecture of the Stacker multi-tiered staging framework of. *Data Placement Module* implements the Machine Learning techniques

- Stacker improves read performance by ~40% as compared to no prefetching (significantly higher than pure locality-based prefetching – ~15% on average)
- Stacker reduces the number of read requests to disk using intelligent prefetching

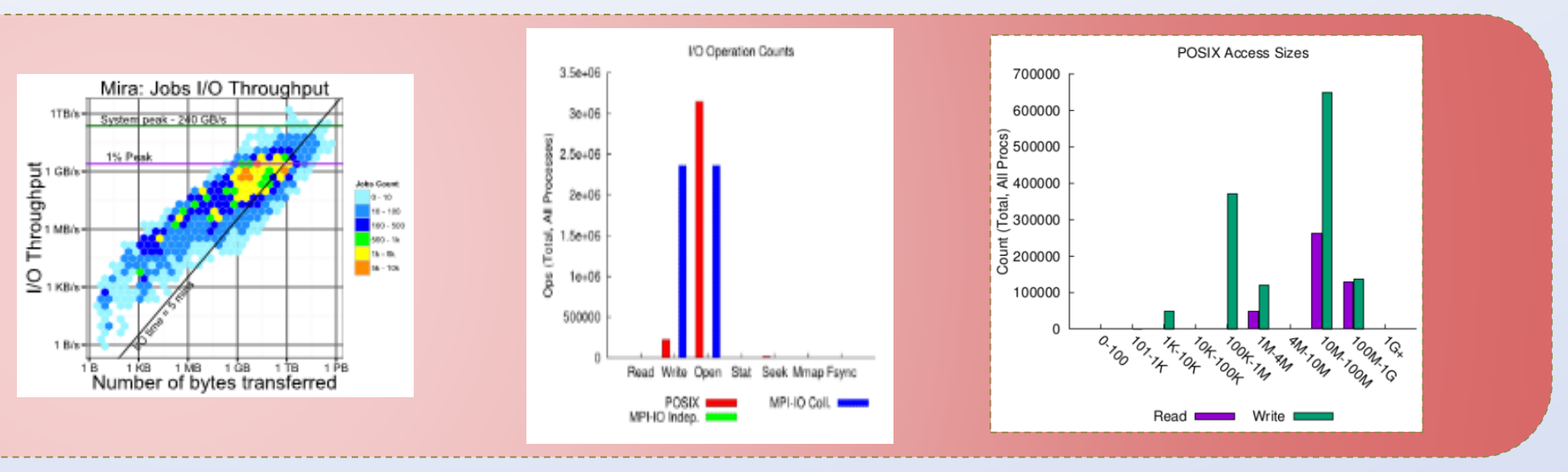
Simple Staging Transport (SST)

- Direct connection between data producers and one or more data consumers
- Communication between separate MPI cohorts
- Multiple transport mechanisms: RDMA, TCP sockets, Shared memory / NVRAM
- Being integrated with Catalyst, and python, VTK-M, XGC1-HPIC, GTC-Python, Visit, Paraview


HPC Application
ADIOS
Data and metadata marshalling
Simple Staging Transport
SST Control Plane
SST Data Plane
NETWORK-SPECIFIC DATA PLANE INSTANCES
High Performance Networks

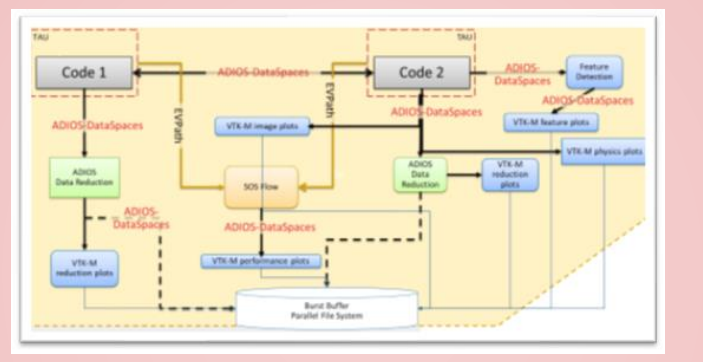
Darshan

- Collects lightweight, always-on I/O characterization for both individual applications and system-wide aggregate
- reports on POSIX, MPI-IO, STDIO
- Modular architecture makes adding new interfaces straightforward
- http://www.mcs.anl.gov/research/projects/darshan/



TAU Performance System®

- Profile, trace, and sampling measurement, analysis, and visualization toolkit
- POSIX, ADIOS, HDF5, NetCDF support
- http://tau.uoregon.edu
- Contact: Kevin Huck (khuck@cs.uoregon.edu)



- **SOSFlow:** (Scalable Observation System for Scientific Workflows) provides a flexible, scalable, and programmable framework for observation, introspection, feedback, and control of HPC applications, output in ADIOS-BP
- Always-on performance capture integrated into the DM tools

Application Engagement

- **SciDAC: HBPS**
  - Optimize I/O on HPC resources utilized in the project for HBPS codes (XGC1, XGCA, HPIC, GENE, …)
  - HPIC coupled through ADIOS for enabling interactive asynchronous data exploration
  - Creation of an XGC1 simulation dashboard to monitor simulation progress
  - On line coupling of HBPS codes with analysis and data reduction services
- **SciDAC: ISEP**
  - ADIOS and visualization services integrated into the GTC simulation
- **SciDAC: HEP Data Analytics on HPC**
  - Working with the pnetCDF team on optimizing the end to end I/O performance
- **SciDAC: HACC (Hardware Accelerated Cosmology Code)**
  - Implementing HDF5 I/O for checkpointing, restarts and analysis
  - Evaluating HDF5 performance compared to POSIX I/O and MPI I/O
- **SciDAC, LCF: E3SM**
  - pnetcdf working to integrate into the PIO/2 code.
  - ADIOS working to integrate into the PIO/2 code
- **SciDAC: SCREAM-2**
  - Kinetic Orbit Runaway electrons Code (KORC) I/O optimizations with ADIOS
  - Code coupling
- **LCF: SPECFEM3D_GLOBE**
  - Scaling of ADIOS with the simulation and incorporation of ADIOS into online workflow
- **LCF: S3D**
  - Integration of Dataspaces to their code
- **LCF: IMPACT**
  - Integration of FastBit and ADIOS to their code
- **LCF: Tri Alpha (ANC, FPIC)**
  - Integrate ADIOS and SST and Visit into their I/O pipelines for high performance I/O on Theta and Summit
- **LCF: OpenFOAM**
  - Performance optimizations using ADIOS
- **LCF: PicOnGPU**
  - Integration of Visualization Services with SST into the PiconGPU code

Outreach and publications

Tutorials

- ATPESC 2018: Data and I/O - Rob Latham, Phil Carns, Quincey Koziol, Jialin Lu
- Climate workshop: Software ecosystem for scientific exascale, N. Podhorszki
- SC 2018: High Performance I/O Frameworks 101 - Klasky, Liu, Parashar, Podhorszki, Pugmire, Wu, Wolf, Atkins
- SC 2018: Parallel I/O in Practice - R Ross, R Latham, B Welch, G Lockwood
- Wuxi 2018: Providing a framework for Self Describing Data, S. Klasky
- ORNL Software Expo 18, ADIOS 2 Tutorial, N. Podhorszki
- Riken 2018, ADIOS 2 Tutorial, S. Klasky
- IXPUG Software-Defined Visualization Workshop: Short ADIOS tutorial, N. Podhorszki
- NASA Langley Workshop 2018: Introduction to HDF5 - M.S. Breitenfeld, E. Pourmal
- Costa Rica Institute of Technology 2018: HDF5 Tutorial - M.S. Breitenfeld

Publications

- Subedi, Davis, Duan, Klasky, Kolla, Parashar , *Stacker: An Autonomic Data Movement Engine for Extreme-Scale Data Staging-based In-Situ Workflows* -(accepted for SC 2018)
- L. Wan, M. Wolf, F. Wang, J. Y. Choi, G. Ostrouchov, S. Klasky, Analysis and Modeling of the End-to-End I/O Performance on OLCF's Titan Supercomputer in High Performance Computing and Communications; IEEE 15th International Conference on Smart City; IEEE 3rd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 2017 IEEE 19th International Conference on, IEEE, pp. 1–9, nominated for best paper.
- T. Lu, Q. Liu, X. He, H. Luo, E. Suchyta, J. Choi, N. Podhorszki, S. Klasky, M. Wolf, T. Liu, et al.. Understanding and Modeling Lossy Compression Schemes on HPC Scientific Data, IPDPS 2018, nominated for best paper.
- J. Gu, S. Klasky, N. Podhorszki, J. Qiang, K. Wu, Querying Large Scientific Data Sets with Adaptable IO System ADIOS in Asian Conference on Supercomputing Frontiers, Springer, Cham, pp. 51–69, best paper award.
- Wang, D., Luo, X., Yuan, F., & Podhorszki, N. (2017). A Data Analysis Framework for Earth System Simulation within an In-Situ Infrastructure. Journal of Computer and Communications, 5(14), 76.
- Bozdag, E., Pugmire, D., Lefebvre, M. P., Hill, J., Komatitsch, D., Peter, D. B., ... & Tromp, J. (2017, December). Visualising Earth's Mantle based on Global Adjoint Tomography. In AGU Fall Meeting Abstracts.
- Lefebvre, M., Chen, Y., Lei, W., Luet, D., Ruan, Y., Bozdag, E., ... & Podhorszki, N. (2017). 13 Data and Workflow Management for Exascale Global Adjoint Tomography. Exascale Scientific Applications: Scalability and Performance Portability, 279.
- Dominski, J., Ku, S. H., Chang, C. S., Choi, J., Suchyta, E., Parker, S., ... & Bhattacharjee, A. (2018). A tight-coupling scheme sharing minimum information across a spatial interface between gyrokinetic turbulence codes. arXiv preprint arXiv:1806.05251. (accepted)