



J. WU, A. SIM -- LBNL
S. KLASKY, J. CHOI -- ORNL
G. S. CHANG -- PPPL



PROJECT GOALS/ RESEARCH CHALLENGES

- Create a framework to allow researchers to conduct distributed analyses on extreme scale data efficiently and easily
- Increase the data handling capability of collaborative workflow systems by leveraging ADIOS, and FastQuery to provide selective data accesses
- Enable large international projects to make near real-time collaborative decisions
- Allow workflows to be modified dynamically for evolving user requirements



ULTIMATE GOALS

- Increase the productivity of diverse teams of scientist by creating an efficient collaborative data management system which can
 - Componentize the I/O abstraction for data: chunks, streams, files
 - Record and capture data provenance for all stages of the collaborative scientific workflow
- Create a paradigm for KSTAR scientists to collaborate during and between shots for analytics to make real-time decisions
- Silly thought “Exascale computing has similar data challenges, so leverage each other to build common re-examine research tools and technologies for **ICEE**.”

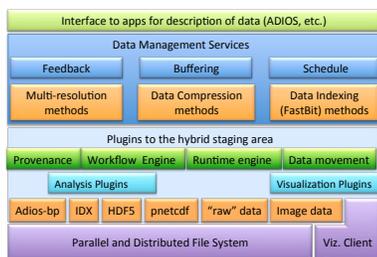
ICEE

WHY NOW?

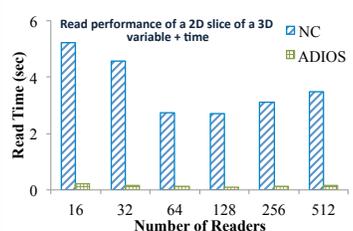
- Big Data is taking the world by storm!
- KSTAR has similar characteristics
 - Volume :
 - Velocity
 - Variety
- Square Kilometer Array has similar aspects (explains my recent trip to Perth)
- Shots grow from 1s to 500s
 - Move from files to streams, from after shot to during shot analysis
- Must understand how we can accelerate workflows
 - Dynamic workflows, minimize data movement, maximize productivity
 - Can write equations to optimize collective services
- Good news is that KSTAR is a solvable problem for the 3Vs, and Fusion scientist have collaborated for years.

ICEE

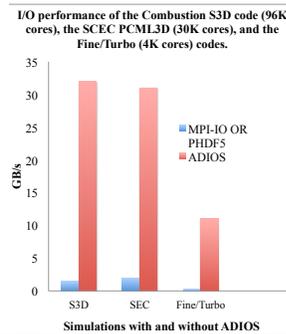
CREATE AN I/O ABSTRACTION



- An I/O abstraction framework
- Provides portable, fast, scalable, easy-to-use, metadata rich output
- Change I/O method on-the-fly
- Abstracts the API from the method <http://www.nccs.gov/user-support/center-projects/adios/>

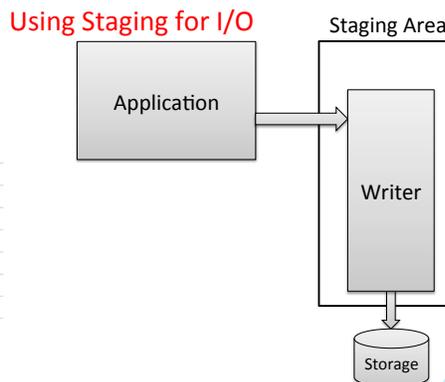
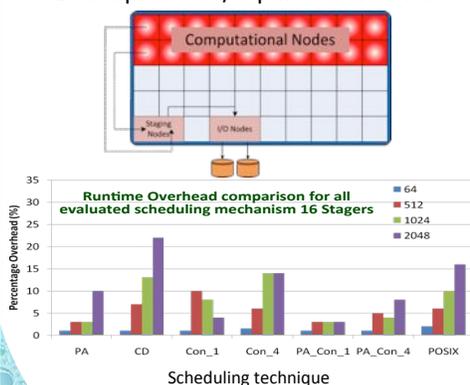


- Provides typical 10X performance improvement for synchronous I/O



CREATE A STAGING AREA WITH THE SAME ABSTRACTION

- Initial development as a research effort to minimize I/O overhead
- Draws from past work on threaded I/O
- Decouples the I/O performance from the File System



Hasan Abbasi, Matthew Wolf, Greg Eisenhauer, Scott Klasky, Karsten Schwan, Fang Zheng: DataStager: scalable data staging services for petascale applications. Cluster Computing 13(3): 277-290 (2010)
 Ciprian Doan, Manish Parashar, Scott Klasky: DataSpaces: an interaction and coordination framework for coupled simulation workflows. Cluster Computing 15(2): 163-181 (2012)



CREATE SELF-DESCRIBING FORMAT FOR DATA CHUNKS/ STREAMS

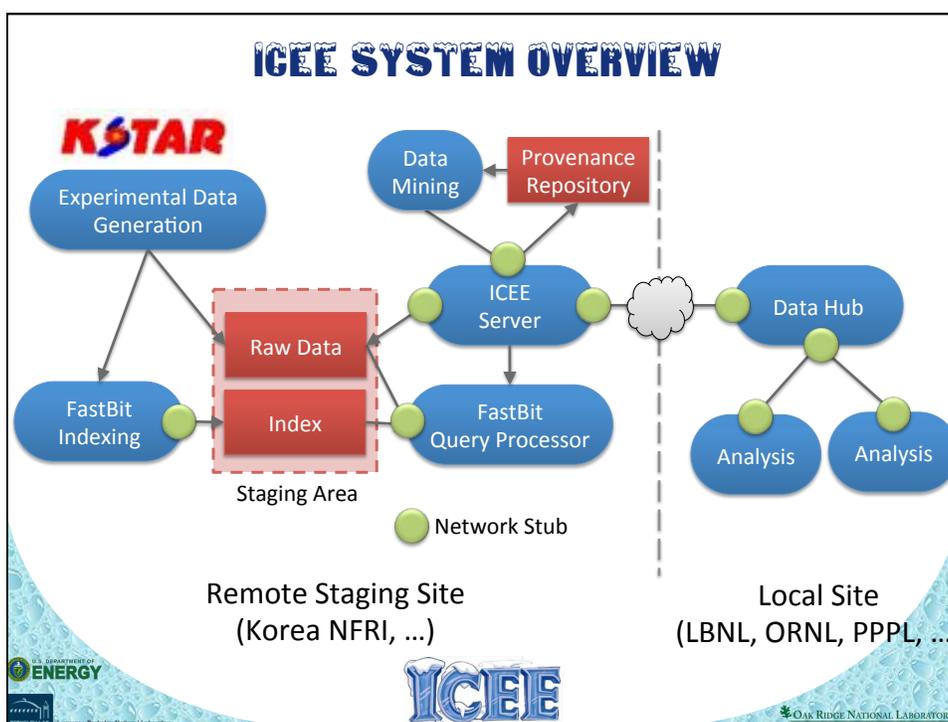
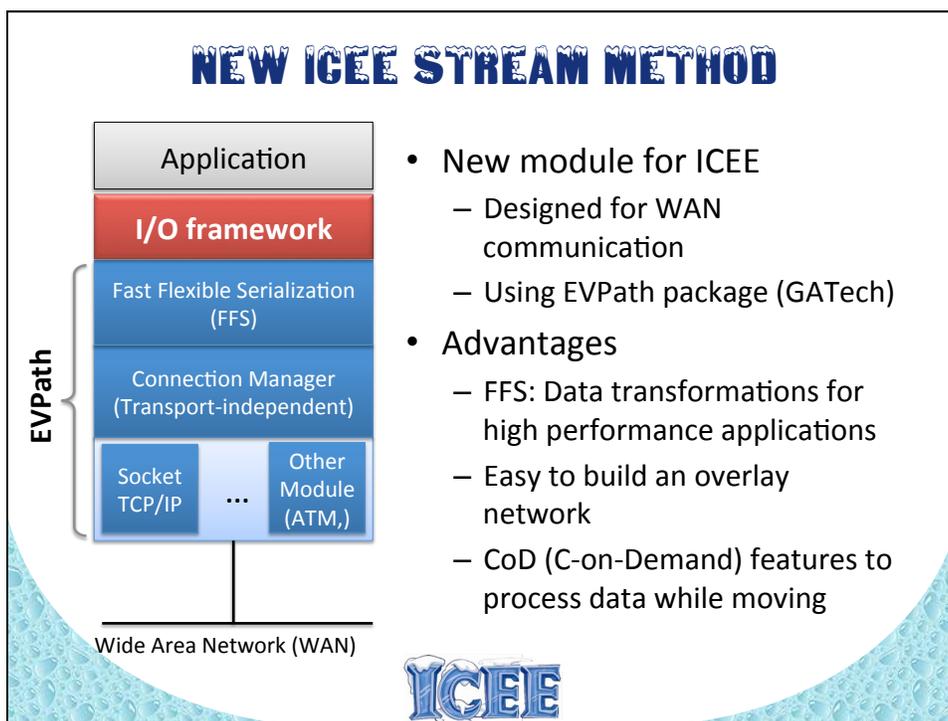
- All data chunks are from a single producer
 - MPI process
 - Single diagnostic
- Ability to create a separate metadata file when “sub-files” are generated
- Allow code to be integrated to streams
- Allows variables to be individually compressed
- Format is for “data-in-motion” and “data-at-rest”

ICEE

DATA IN FILE VS. DATA STREAM

- Research issues in stream-based data process
 - Prioritization of data streams for “individual” consumers
 - Prioritization of data streams for “collaborative” consumers
- Move work to data or data to work?

ICEE

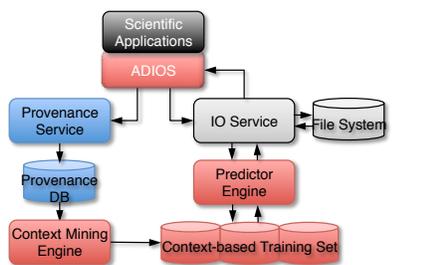


KNOWLEDGE DISCOVERY IN COLLABORATORIES

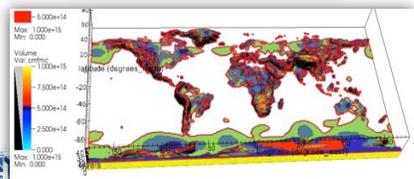
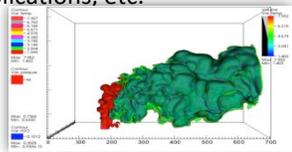
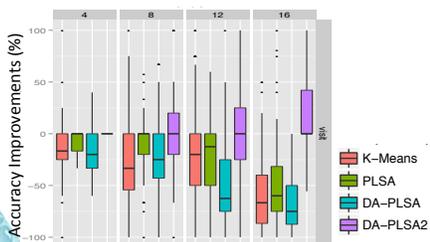
- Collecting provenance data
 - Data lineage (tracking inputs and outputs, metadata, data locations, ...)
 - Data access patterns
- Mining hidden knowledge to support:
 - Parameter sensitivity study (UQ, ...)
 - Access prediction and data prefetching to reduce latency
 - Information for performance tuning
 - Extract common knowledge
- Such activity should be performed transparently and efficiently → Transparent plug-in approach



MINING PROVENANCE DATA OF I/O



- Provenance module for middleware
- Store and index user data access activities
- Transparently integrated with scientific and visualization applications (e.g., VisIt)
- Mining provenance data for IO prefetching, auto-tuning applications, etc.



J. Choi, H. Abbasi, D. Pugmire, S. Klasky, J. Qiu, G. Fox, "Mining Hidden Mixture Context With ADIOS-P To Improve Predictive Pre-fetcher Accuracy", escience 2012



STREAM-BASED DYNAMIC WORKFLOW

- Integrating I/O system with workflow concept
 - To support seamless data processing
 - To provide end-to-end data access pipeline
- Various strategies
 - Management: centralized, hierarchical, distributed, ...
 - Execution models: static, dynamic, just-in-time, ...
 - Workflow definition: scripts, language-based definition, embedded, ...

ICEE

COLLABORATIVE DATA ANALYSIS

- EVPath-based WAN communication
 - Memory-to-memory coupling through WAN or between institutions
 - Tested between ORNL and LBNL
 - Integrate with RDMA/Globus “if necessary”
- CoD (C-on-Demand)
 - Dynamic code generation
 - Investigating data processing with CoD on transmitting nodes (relay nodes, routers, network H/W, etc.)
- Still need to address the “difficult” challenges
 - Where are the “workflows” stored and communicated?
 - How do we allow for Adaptive workflows?
 - How do we minimize the data movement?

ICEE

DIVING INTO SOME DETAIL

- How to pinpoint interesting data records
- How to find interesting features

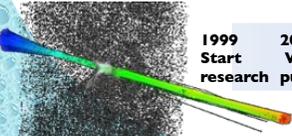




PINPOINTING THE INTERESTING DATA: FASTBIT INDEX



- **Problem:** given a large data collection, quickly find records satisfying user-specified conditions
 - Example: in billions of high-energy collision events, find a few thousand based on energy level, number of particles and so on
- **Solutions**
 - **Algorithmic research:** developed new indexing techniques, achieved 10-100 fold speedup compared with existing methods
 - **Efficient software:** available open source, received a R&D 100 Award
- **Enabled science**
 - **Laser Wakefield Particle Accelerator:** FastBit acts as an efficient back-end for identifying and tracking particles (lower left figure)
 - **Combustion:** FastBit identifies ignition kernels based on user specified conditions and tracks evolution of the regions
- **Testimonial** “FastBit is at least 10x, in many situations 100x, faster than current commercial database technologies” – *Senior Software Engineer, Yahoo!*



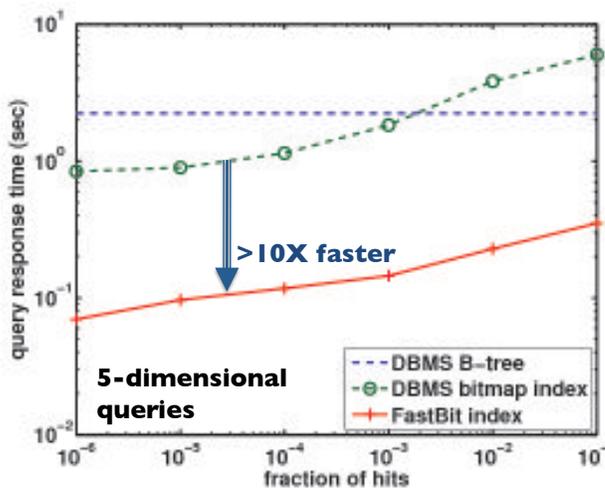
1999	2001	2004	2006	2007	2008
Start	WAH	WAH	- Query Driven Vis	- FastBit released	- R&D100 Award
research	published	patented	- Published theory	- BioSolveIT begin use	- Yahoo! begin use



[\[Wu, et al. 2009\]](#)

SPEED: >10X FASTER THAN COMMON DBMS

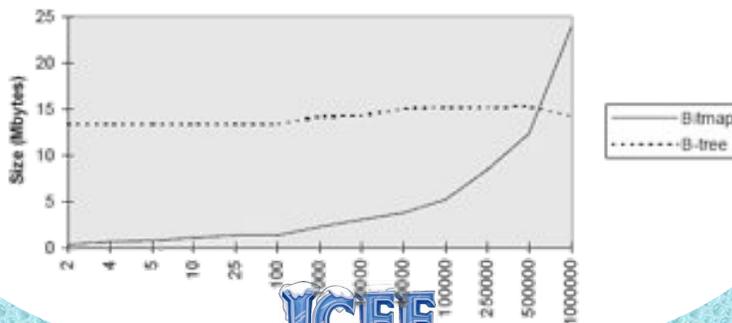
- Queries 5 out of 12 most popular variables from STAR (2.2 million records)
- Average attribute cardinality (distinct values): 222,000
- FastBit uses WAH compression
- DBMS uses BBC compression
- FastBit >10X faster than DBMS
- FastBit indexes are 30% of raw data sizes



[Wu, Otoo and Shoshani 2002]

SIZE: CAN BE CONTROLLED

- The figure below compares the sizes of a bitmap index against a B-tree index (for 1,000,000 records)
 - Bitmap index could be much smaller than the widely used B-tree index
- General strategies for controlling index sizes
 - Reduce precision (binning), reduce granularity (indexing blocks), complex bitmap encoding

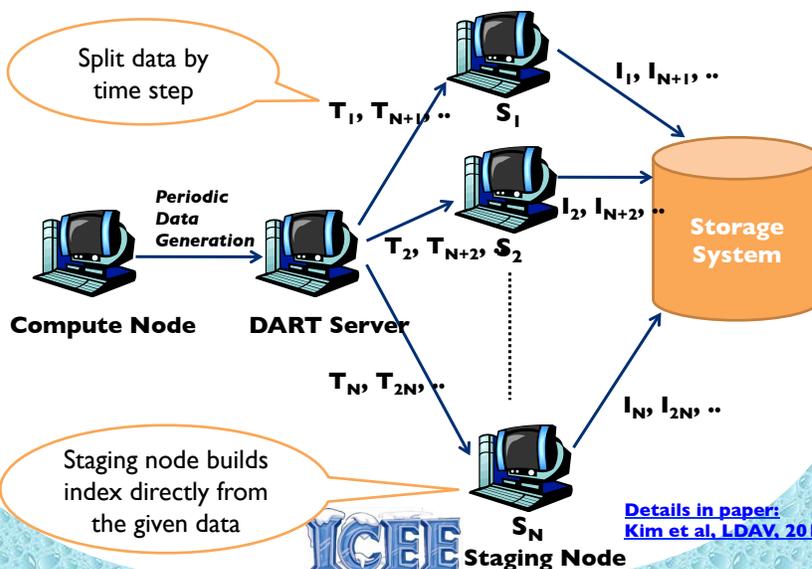


WHY IS NOT EVERYONE RUSHING TO USE FASTBIT?

- **Challenge:** using indexes can significantly speed up data accesses, but building indexes takes a lot of time
- **Observation:** building index requires reading all raw data, which can be very time consuming
- **Solution:** avoid reading data from disk by building indexes in transit

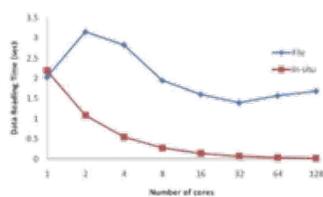
ICEE

SCHEMATICS OF IN TRANSIT INDEXING

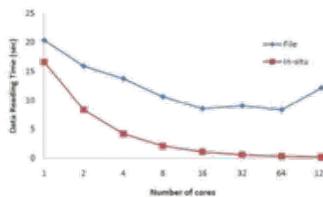


ICEE

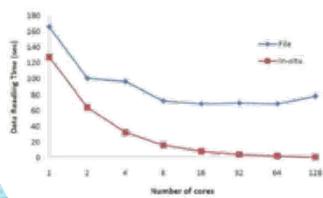
IN TRANSIT INDEXING REDUCES READ TIME



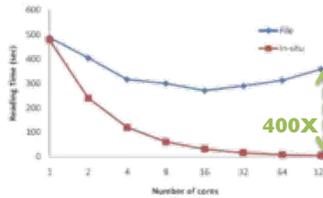
(a) Small



(b) Medium



(c) Large



(d) Large2

Franklin @ LBNL

- ~10000 nodes
- 8 cores
- 8 GB memory
- Lustre file system

Data Sizes

- Small: 3.6GB
- Medium: 27GB
- Large: 208GB
- Large2: 173GB

- Getting data from other processors could be 400x faster than getting data from disk



KNOWLEDGE DISCOVERY & DATA MINING PROBLEMS

- Discovering and mining knowledge from the workflow provenance data
 - Accumulated and collected provenance information
 - Study the viability of predicting data access requirements for the entire set of users.
 - Various clustering or classification algorithms (e.g. model-based clustering, high-dimensional data clustering, pairwise data clustering, support vector machine, etc.) can be used
- Data access pattern recognition
 - Increasing data pre-caching chances to reduce data acquisition latency
 - Keeping up with a quickly evolving set of user requirements
 - The data access patterns can be used to select indexing options for the data and to minimize the data movement in workflow orchestration.



VALIDATING WORKFLOWS

- Validating workflows for changing conditions based on machine learning
 - Knowledge discovery in the collaborator workflow system can validate real-time or near-line workflow conditions based on the previous cases.
 - Plan to study adaptive CBR for the dynamic workflow validation model to validate the changing conditions, and provide the workflow system with learning capabilities.

ICEE

MOBILE ACCESS TO REMOTE WORKFLOW MONITORING

- To explore tablet mobile computing support in collaborative environment, plan to study the following
 - Remote monitoring and control of the workflow and data analysis
 - Mobile distributed access to the monitoring information for the workflow and data analysis, through tablets with iOS or Android OS
 - Tablet access to monitor and update the workflow
 - Tablet access to the results of the workflow
 - Tablet access to analysis results and science discovery
 - Exploration of mobile tablet computing in science
 - Exploring mobile tablet computing as a computing resource for data analysis
 - Exploring mobile tablet computing as a collaborative tool

ICEE

SUMMARY

- ICEE is investigating ways to optimize the real-time decision making process for the international KSTAR experiment
 - New method in I/O abstraction for WAN data movement with efficient data I/O on LCFS and small clusters
 - Research state of the art indexing and queries into I/O pipelines
 - Research the inclusion of COD + workflows embedded into data streams
 - SC-13 paper in preparation
- Many challenges remain
 - Optimizing indexing time – data movement time
 - Optimize queries across ensembles of workflows
 - RDMA over the WAN from KSTAR to LBNL/ORNL/PPPL
 - Optimizations of ensembles of workflows
 - Creation of new services for the larger CS/physics community

ICEE