International Collaboration Framework for Extreme Scale Experiments (ICEE)

Annual Project Report – 1st year

October 1, 2012

Project period of
September 1, 2012 through August 31, 2013

Investigators

John Wu[1], CS Chang[2], Scott Klasky[3], Alex Sim[1], Jong Youl Choi[3]

Project Website: **http://sdm.lbl.gov/icee**

TABLE OF CONTENTS

[1] Lawrence Berkeley National Laboratory

[2] Princeton Plasma Physics Laboratory

[3] Oak Ridge National Laboratory

1 Summary

Large-scale scientific exploration in high-energy physics, fusion, climate and so on are based on international collaborations. As these collaborations produce more and more data at faster rates, the existing workflow management systems are hard pressed to keep pace. A necessary solution is to process, analyze, summarize and reduce the data before it reaches the relatively slow disk storage system. This approach is generally known as in-transit processing or in-flight analysis. The ICEE project aims to introduce this in-transit analysis capability into a collaborative workflow system by leveraging the in-transit capability of ADIOS and selective data access capability of FastBit. Additionally, ICEE seeks to provide an effective data flow management for distributed workflows and enable large international projects to make near real-time collaborative decisions.

As scientific teams tackle increasingly complex problems, many data analysis workflows have to dynamically adjust to the experimental conditions and sensor outputs; thus, workflows also need to be modified dynamically following the evolving user requirements. The ICEE framework will not only allow users to modify parameters of a workflow, but also dynamically modify its processing elements and alter its structure. Additionally, we plan to incorporate data mining features to provide feedback and recommendations while the user is constructing or modifying a workflow. Overall, the ICEE framework will allow researchers to conduct distributed analyses on extreme scale data efficiently and easily. It will enable collaborative decisions in near real-time for geographically distributed teams, reduce the turn-around time on large instruments, and improve scientific productivity.

In the first year of this project, we have developed a prototype system for integrating ADIOS and FastBit and demonstrate its functionality by deploying it between LBNL and ORNL. The current prototype is based on a static workflow for analyzing Soft X-Ray data from the Korean KSTAR project. At the same time, we have started working on the monitoring software based on iOS devices. The project team has made a trip to KSTAR to create a detailed collaboration plan for the coming years. Additionally, we are working with ESnet engineers to diagnose network connections between KSTAR and ESnet. These activities are setting up the groundwork for more advanced wide-area workflow system development planned for the next two years. In the remainder of this report, we provide a brief outline of the key technical activities in the past year.

2 Wide-Area Workflow Prototype

2.1 Overall architecture

Figure 1 shows the current design of the overall ICEE framework. Figure 2 shows the software stack of the current implementation. A technical paper describing the system is currently under review. Next, we provide a brief outline of the key software components.

2.2 Key components

Our system consists of three main components: data acquisition, data server called ICEE server, and remote clients connecting through wide-area networks. We provide programing interfaces (APIs) for clients to query data hosted by an ICEE server (See Figure 1).

Data Acquisition: This is an interface between data sensors, which generate raw experimental data, and ICEE server. It manages the raw data in memory in order to be accessible for ICEE server.

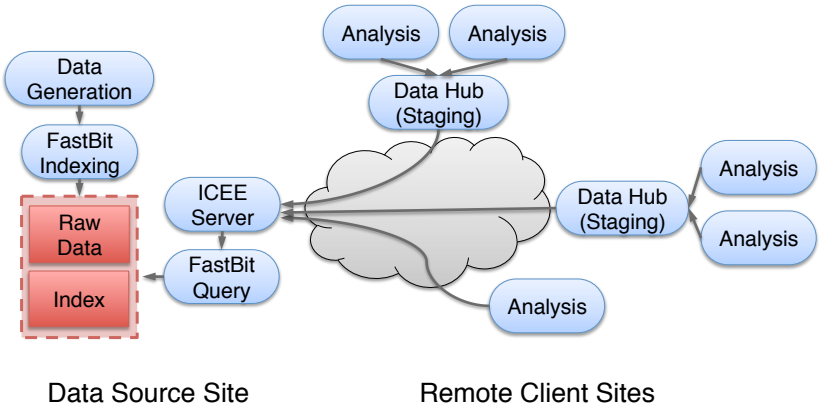

Figure 1. Overall architecture of ICEE prototype system.

The user data can be either pushed or pulled into the ICEE system. Data may be pushed to ICEE for common activities, such as indexing of commonly used variables, which requires a significant amount of computer time. It is important to prepare the indexes before they are needed. We anticipate that the system will automatically build commonly used indexes as soon as the data is available, while a user may specifically request indexes on infrequently used variables.

ICEE server: ICEE server is in charge of providing data to the remote users. It supports common data related activities, such as indexing and processing queries.

In order to minimize disk I/O overheads, operations are performed in memory as much as possible; i.e., the index and the raw data will be residing mainly in memory and provide as a data stream without accessing files. Data in memory is separated by logical time steps (or called as shots in fusion experiments) but shares similar structures between time steps. This is a natural format of many scientific data that are generated or collected over time. Our initial prototype has only a single server, but we plan to expand the number of servers to provide more computing power for the widely distributed user community.

Remote clients: Clients will be distributed over wide- area networks and will access remote data by connecting to an ICEE server directly or via data hub, which will be delegated by multiple clients. Data hub can act as an aggregator for local clients or as a parallel data analysis by providing in-memory staging services.

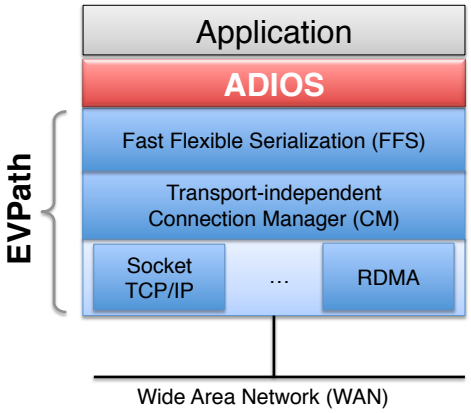

Figure 2. The software stack of the current ICEE prototype

Clients can request data with queries to access a portion of data selectively. Server will exploit index data to extract a portion of satisfactory data set.

All communications between clients and servers are a form of Remote Procedure Calls (RPCs), provided by a set of APIs we developed. APIs are based on ADIOS and EVPath, which will be discussed in detail next.

We develop a new connection module (or plug-in) for ICEE system to help the ICEE server and the remote clients to be connected easily and to exchange data. Our module is integrated with ADIOS and thus shares the same ADIOS APIs (or same interfaces) but supports wide-area data

exchanges under the hood. All operations are transparent to users.

2.3 Performance measurements

We have measured the performance of our prototype system using a static wide-area data analysis scenario. We employed two clusters running in geographically separated locations: Sith, located in Oak Ridge National Lab (ORNL), and Carver, located in Lawrence Berkeley National Laboratory (LBL). These two locations are connected by 10Gbps/100Gbps backbones operated by Energy Sciences Network (ESnet). Sith is a Linux cluster comprised with 40 compute nodes. Each node contains four AMD Opteron Processors (2.3GHz with 8 cores), and 64 GB of memory.

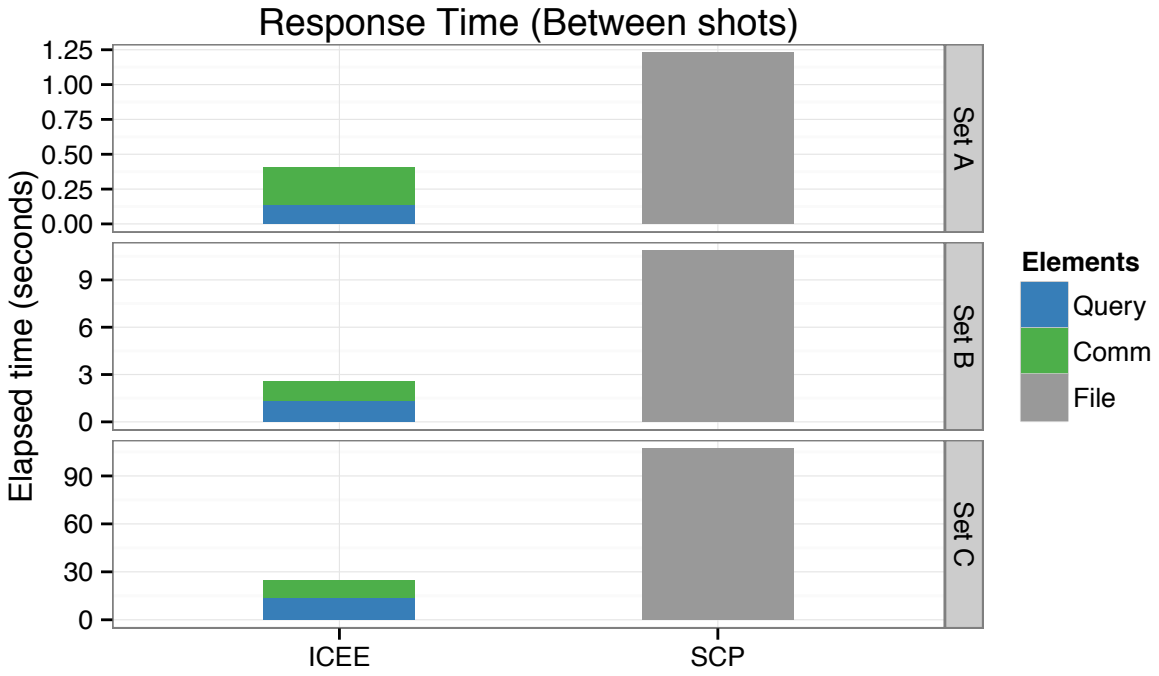

Figure 3. Estimated response times between shot periods with an assumption of having 10% query response results. Sets A-C were prepared from real-world KSTAR data with different workloads; small, medium, and large outputs respectively.

We used a real-world experimental data set measured from KSTAR tokamak, called Soft X-ray Array (SXR). SXR is a time-series data measured by a camera array with 64 channels. Each channel can generate measurement data (as integer numbers) per 2 microseconds. As of current technology, a fusion reaction can last about 10 seconds per shot and a SXR data set contains 5,000,000 integer numbers per channel. KSTAR tokamak is expected to increase the reaction time up to about 300 seconds in future.

Figure 3. shows the estimation of response times for our three different workload data sets; small, medium, and large, denoted by Set A, Set B, and Set C respectively. With an assumption of 10% query response result, we can achieve significant performance gains by using FastBit, compared with just sending data (SCP copy) as files. Especially, with Set C, we achieved 4.2x speedup.

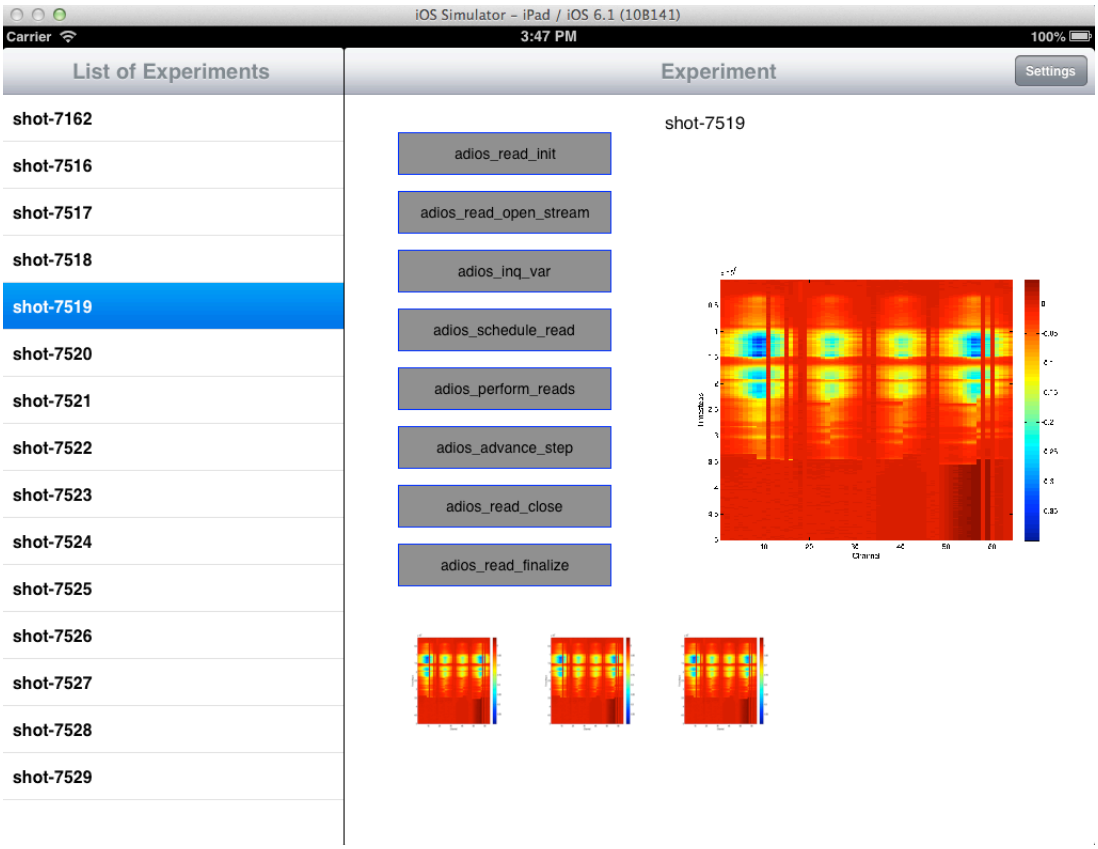

Figure 4. A screen of the iOS monitoring software

3 Monitoring Software

A front-end for monitoring the distributed workflow is under development at LBNL. Figure 4 is a screen shot of the current prototype system based on iOS running on an Apple iPad. In addition to displaying output images from the analysis framework, it also offers limited functionality for manipulating the images. Current prototype makes use of the output files from the remote workflows. Work is underway to more directly tap into the remote workflow by using the ADIOS API and EVPATH API.

4 KSTAR Collaboration

KSTAR (Korean Superconducting Tokamak Advanced Research) device is an advanced fusion device located at National Fusion Research Institute (NFRI) of Korea to study near-steady state aspects of magnetic confinement fusion as in ITER. Large-scale data will be produced from the KSTAR experiment. It has a large number of international participants that could significantly benefit from the real-time accesses to the diagnostic data collected during the operation of the device. ICEE (International Collaboration Framework for Extreme-scale Experiments) is a project motivated by this need for large-scale data collaboration. The key solution strategy of ICEE is to analyze, summarize and reduce the data before it reaches the relatively slow disk storage system. A workshop was held at KSTAR for the software developers to report

progress, review the system design, and plan for shared tasks among the participants of the ICEE project from both US and Korea. ICEE project had been underway for nearly nine months prior to this workshop.

KSTAR has plans to provide data and computing resources to support the ICEE project. Figure 5 is the schematics of how ICEE could fit into the KSTAR data analysis ecosystem presented by Dr. S. I. Lee at the workshop.

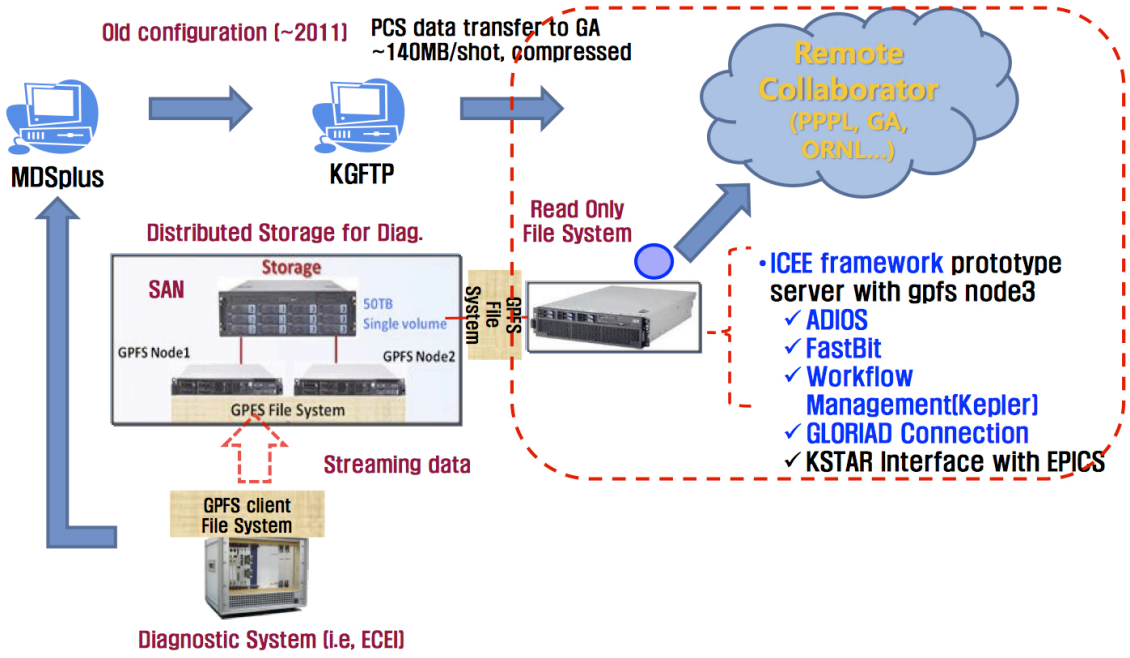

Figure 5: A plan for integration, presented by Dr. S. I. Lee of KSTAR

With the demonstrated commitment from KSTAR, the investigators of ICEE project plan to work on the following items.

1. Work with the network engineers at NFRI, NISN and ESnet to exercise the network connection from KSTAR to LBNL. Making sure of a good network connectivity can benefit all the collaborators of KSTAR and is an essential first step for the ICEE framework to work efficiently.
2. Demonstrate the stream processing of data in the wide-area network with ADIOS on KSTAR's ECEI data. As an intermediate step, Dr. Chang has arranged data to be temporarily stored at NISN (Korean National Institute of Supercomputing and Networking) so that the initial testing could be done from NISN to LBNL or ORNL.
3. Build an analysis workflow for ECEI data analysis. Use the same analysis workflow for both measurement data and modeling data. The actual measurement data consists of 8x24 16-bit values per time step. The device is capable of collecting data at a high frequency of 500KHz. To provide a robust feature identification using this data is potentially very challenging because of the low resolution and high frequency. We also plan to work with model output, which could be of higher resolution and easier for the feature identification algorithms to deal with.

5 Plans for Year 2

- Prototype of a repository for workflows that allows dynamic additions and modifications
- Static workflow scheduling and execution over a data stream
- A mechanism for associating attributes and provenance to the data stream, and recording provenance
- Identify common selection criteria and common features that could be indexed
- Explore options and trade-offs of the in-transit index generation
- Investigate authentication strategies for data integrity
- Work with the KSTAR data management team to characterize their current diagnostics data
- Use data placement in conjunction with static workflow scheduling to optimize the usage of network resources
- Stream the specialization and the differentiation, based on user requirements.
- Initial prototype of a stream synchronization controller based on collaborative groups
- Connect the monitoring tool with distributed workflow through ADIOS API