Title of Pre-application: Science-driven Data Management for multi-tiered storage
Principal Investigator: Scott Klasky, Group Leader, ORNL, 854-241-9980, klasky@ornl.gov
Laboratory Announcement Number: LAB 15-1338, Funding/year: $1.25M
Co-PIs and Key/Senior Personnel

Hasan Abbasi, ORNL	Mark Ainsworth, ORNL JFA, Brown University
Matthew Curry, SNL	Qing Liu, ORNL
Jay Lofstead, SNL	Kimmy Mu, ORNL
Carlos Malzahn, U. Cal. Santa Cruz	Manish Parashar, Rutgers University, ORNL JFA
Feiyi Wang, ORNL	Lee Ward, SNL

Objectives: Exascale scientific discovery will be severely bottlenecked without sufficient new research into managing and storing the large amounts of data that will be produced during the simulation, and analyzed for months afterwards. In this project we will demonstrate novel techniques to facilitate efficient mapping of data objects, even partitioning individual variables, from the user space onto multiple storage tiers, and enable application-guided data reductions/transformations to address capacity and bandwidth bottlenecks. Our goal in this project is to address the associated I/O and storage challenges in the context of current and emerging storage landscapes, and expedite insights into mission critical scientific processes; which is associated with **theme two** of the FOA. To that end, we will build on the capabilities offered by ADIOS and DataSpaces at ORNL that provide I/O abstractions and services, the Sirocco peer-to-peer file system at Sandia and the object storage and annotation expertise of UC Santa Cruz, to explore application and multi-tier storage aware data management solutions. This project brings together a team with strong expertise in I/O middleware (ORNL, Rutgers), file system (SNL, UCSC) and storage (UCSC), and connects and coordinates these key storage components in a seamless fashion.

Our objective here is to reduce the time to knowledge, not just for a single application, but for the entire workload in a multi-user environment, where the storage is shared among users. We achieve this goal by allowing selectable data quality, by trading its accuracy and error in order to meet the time or resource constraint. We will explore beyond checkpoint/restart I/O, and will address the challenges posed by key data access patterns in the knowledge gathering process. Ultimately, we will take the knowledge from the storage system to provide vital feedback to the middleware so that the best possible decisions can be autonomically made between the user intentions and the available system resources. We will test our prototypes on current and future DOE system with many of today's applications, including the s XGC1, GTC, QMCPack, and SpecFM3D simulations.

Our solutions will provide new functionalities and APIs for 1) specifying, at the application level, data annotations that enable the quantification of the relative importance and utility of data objects and enable partitioning across the storage hierarchy; 2) specifying selectable performance/quality/cost tradeoffs from both the application and system perspectives; 3) evaluating these tradeoffs at runtime during data placement and movement, and executing the resultant policies in an autonomic system using models, heuristics and continuous learning; and 4) leveraging techniques such as application-aware data compression, and I/O prioritization to enforce these policies.

Key Technical Approach: Our approach is based on the insight that storage system awareness of application data requirements provides a powerful tool for guiding key I/O challenges such as data access, data representation in storage, and data placement within the hierarchy. Towards this goal, we will allow both the developers of applications and users running these applications, to plug-in their knowledge about their data, represented not as bytes but as motifs, allowing the I/O and storage system to understand their intentions, the relationships between the data objects, as well as access and transport patterns. For example, based on insights into how the data is used, we will offer a coarse grained, quick overview of a data set with progressively more detail in areas

of interest such as those with features, and less detail in areas without. These progressively more detailed data views require more storage space and time to retrieve if left unmanaged. By defining mechanisms to provide this knowledge and incorporating its awareness into the middleware and storage, we will be able to selectively store data at different levels, and allow a user to select not only what data to retrieve but also specify an acceptable timeframe and accuracy. The storage system layer may also dynamically adjust the level of data that will be retrieved for an individual application, to maximize the overall system efficiency and provide fairness among applications. The proposed work will address the following technical areas:
First, the storage system will offer the ability to directly store different portions of a single variable to different levels of the hierarchy and allow each version to be processed differently. For example, full data can be queued to tape while a highly compressed version intended for high-level analytical views can be stored in non-volatile memory (NVM) within the storage hierarchy. The key challenge would be defining and maintaining the metadata connecting different object quality and utility levels. The storage system will support the metadata and plug-ins required to support this approach. **Second**, new storage access APIs will be developed that incorporates the notion of time and data quality. The storage system, through the knowledge of both the data quality stored in different storage hierarchy tiers and the approximate time to retrieve it from the different tiers, can manage the tradeoff between error bound and the time to retrieve the data. This storage system support will be executed and managed through our middleware, insulating the user as much as possible from the new APIs. The storage system will also support plug-ins, for example, to potentially decompress or expand data to the original size. Both lossless and lossy with error bounds style storage is assumed. The time factor must take into account how long this operation takes for servicing the user request. **Third**, the storage system will offer annotations within the metadata and support for both predictive and reactive data placement and migration. While past access patterns may not indicate future access because the simulation details may have changed, we are focused on scalability where subsequent runs are larger in scale as the simulation prepares for a capability run. By learning from the output and access patterns during this run sequence, we can accurately decide how to place and organize data for the critical capability runs. The reactive mechanisms will consider the space and performance characteristics, as well as the requested data fidelity, to determine where and how to pull and store data sets. The challenge here is to utilize this metadata to manage the placement and access of data object based on the defined constraints as well as storage system state. **Fourth**, to effectively manage data storage capacities, the storage system's capability for storing multiple, potentially different data quality versions of the same data will be leveraged to guide data eviction and migration. The main challenge of this approach is to be aware of intentional placement decisions made to optimize future access while ensuring proper system operation by not exhausting any tier inappropriately. If a single application is allocated on the entire machine, completely consuming resources on the machine is expected. For smaller runs, these allocations must be balanced to support the diverse workload. Our research efforts will be heavily focused on the need to, in a coordinated manner, adapt data and metadata retention policies according to the dynamic resource balancing that will need to take place between the application, OS/R, and hardware.

The success of this project will provide insights into how to build extreme scale autonomic middleware and storage layers that can interact effectively with each other, and bring the user in the loop through user-provided hints, allowing user and system knowledge to be incorporated into the SSIO software layers. Storage resources will be managed against the needs of individual simulations as well as what's happening on the entire system. The impact of such a solution will be a significant reduction in the time to knowledge, and the overall increase in the effectiveness and utility of exascale simulations.

List of Conflicts

The following is a consolidated list of collaborators co-editors and advisors for all the key personnel. Anyone not on this list but employed or affiliated with one of the submitting institutions ORNL, SNL, U.C.S.C., and Rutgers, including our affiliated institutions (N. C. State, Georgia Tech, U. Tenn. Knoxville) should also be considered as a conflict.

Advisor-Student conflicts

Allendes, A., Universidad Tecnica Federico, Chile
Andriamaro, G., Strathclyde University, UK
Ma, X., Abertay University, UK
Mihai, L., Cardiff University, UK
Rankin, R., Rice University, US
Wajid, H., COMSATSIT, Pakistan
Craig, A., European Commission Brussels
Grunwald, D., UCSC
Lysecky, R., University of Arizona
Oden, J., University of Texas at Austin
Schwan, K., Georgia Institute of Technology
Wu, F., University of California

Co-Author conflicts

Achlioptas, D., UCSC
Ah Nam, H., Los Alamos National Laboratory
Ames, S., LLNL
Ananthakrishnan, R., University of Chicago
Arpaci-Dusseau, A., UW Madison
Arpaci-Dusseau, R., UW Madison
Atchley, S., Oak Ridge National Laboratory
Barrenechea, G., Strathclyde University, UK
Barreto, R., University of Chicago
Barrett, B., Amazon
Barrett, R., Sandia National Laboratories
Bent, J., EMC Corporation
Bhagwan, V., Yahoo!
Booth, M., Oracle
Brandt, J., Sandia National Laboratories
Brandt, S., UCSC
Brightwell, R., Sandia National Laboratories
Brim, M., Oak Ridge National Laboratory
Buck, J., Context Relevant
Caldwell, B., Oak Ridge National Laboratory
Carns, P., Argonne National Laboratory
Carothers, C., Rensselaer Polytechnic Institute
Chen, J., Sandia National Laboratories
Cope, J., DDN
Crichton, D., NASA
Cummings, J., California Institute of Technology
Davydov, O., Giessen University, Germany
Denvil, S., Institut Pierre Simon Laplace
Dillow, D., Google
Docan, C., Rutgers
Doutriaux, C., Lawrence Livermore National Laboratory
Drach, R., Lawrence Livermore National Laboratory
Ethier, S., Princenton Univeristy
Ezell, M., Oak Ridge National Laboratory
Fabian, N., Sandia National Laboratories
Ferreira, K., Sandia National Laboratories
Fiore, S., Euro-Mediterranean Center on Climate Change
Fuller, D., Redhat
Gavrilovska, A., Georgia Institute of Technology
Gemmill, J., Clemson University
Gentile, A., Sandia National Laboratories
Gibson, G., Carnegie Mellon University
Gokhale, M., LLNL
Gonzalez, E., German Climate Computing Center (DKRZ), Germany
Grandison, T., Proficiency Labs
Grider, G., LANL
Grout, R., National Renewal Energy Laboratory
Gunasekaran, R., Oak Ridge National Laboratory
Gupta, S., Oak Ridge National Laboratory
Guzman, J., Brown University
Harms, K., Argonne National Laboratory
Harney, J., Oak Ridge National Laboratory
Harris, J., Clemson University

He, J., UW Madison
Hemmert, S., Sandia National Laboratories
Hill, J., Oak Ridge National Laboratory
Ioannidou, K., TidalScale
Ionkov, L., LANL
Jean-Baptiste, G., Florida International University
Jhala, A., UCSC
Jimenez, I., University of California Santa Cruz
Jong, C., ORNL
Karniadakis, G., Brown University
Kato, S., Nagoya University
Kelly, S., Sandia National Laboratories
Kershaw, P., STFC R. A. Laboratory, UK
Kim, Y., Oak Ridge National Laboratory
Kimpe, D., Argonne National Laboratory
Klasky, S., Oak Ridge National Laboratory
Klundt, R., Sandia National Laboratories
Koziol, Q., HDF Group
Ku, S., New York University
Laros III, J., Sandia National Laboratories
Leung, V., Sandia National Laboratories
Levenhagen, M., Sandia National Laboratories
Leverman, D., Oak Ridge National Laboratory
Levin, G., Google
Liu, N., ANL
Liu, Q., Oak Ridge National Laboratory
Lofstead, G., Sandia National Laboratories
Lofstead, J., SNL
Logan, J., Oak Ridge National Laboratory
M. Bell, G., Lawrence Livermore National Laboratory
Manzanares, A., HGST
Mateas, M., UCSC
Mattmann, C., Jet Propulsion Laboratory
Maxwell, D., Oak Ridge National Laboratory
McCormick, P., LANL
Miller, R., Oak Ridge National Laboratory
Moreland, K., Sandia National Laboratories
Morgan, M., Institut Pierre Simon Laplace
Narasimha, R., Texas A&M
Nassi, I., TidalScale
Nelson, M., Redhat
Obraczka, K., UCSC
Oral, S., Oak Ridge National Laboratory
Parashar, M., Rutgers
Pascoe, S., STFC R. A. Laboratory, UK
Pedretti, K., Sandia National Laboratories
Pobre, Z., Goddard Space Flight Center
Podhorzski, N., Oak Ridge National Laboratory
Polte, M., WibiData
Polyzotis, N., Google
Pye, I., Cloud Helix
R. Resnick, D., Sandia National Laboratories
Reiss, C., University of California Berkeley
Ricci, R., University of Utah
Rodrigues, A., Sandia National Laboratories
Rogers, J., Oak Ridge National Laboratory
Ross, R., Argonne National Laboratory
S. Bland, A., Oak Ridge National Laboratory
Sayas, F., University of Delaware
Schweitzer, R., Pacific Marine Environmental Laboratory
Shen, H., Clemson University
Shewmaker, A., LANL
Shipman, G., Los Alamos National Laboratory
Shoshani, A., Lawrence Berkeley National Laboratory
Sim, A., Lawrence Berkeley National Laboratory
Simmons, J., Oak Ridge National Laboratory
Sjaardema, G., Sandia National Laboratories
Skjellum, A., University of Alabama
Skourtis, D., VMWare
Sun, W., Google
Sun, X., Illinois Institute of Technology
Sun, Z., Data Direct Networks
Tchoua, R., University of Chicago
Thompson, D., Sandia National Laboratories
Tian, Y., Oak Ridge National Laboratory
Tiwari, D., Oak Ridge National Laboratory
Torres, A., Linden Lab
Tucker, T., Sandia National Laboratories
Van Dyke, J., Sandia National Laboratories
Vaughan, C., Sandia National Laboratories
Vergara Larrea, V., Oak Ridge National Laboratory
Wang, Y., Auburn University
Ward, L., Sandia National Laboratories
Weil, S., Redhat

Wheeler, K., Micron Technologies
Whitehead, J., UCSC
Widener, P., SNL
Williams, D., LLNL
Wingate McClelland, M., Xyratex
Yu, W., Auburn University
Zheng, F., IBM
Long, D., University of California, Santa Cruz

Collaborator conflicts

Agrawal, A., NWU
Ahrens, J., LLNL
Berk, G., Kitware
Bethel, W., LBL
Bremer, P., LLNL
Brooks, R., Clemson University
Cao, C., University of Tennessee
Cappello, F., Argonne National Laboratory
Cinquini, L., NOAA
CS, C., PPPL
Danielson, G., Sandia National Laboratories
Dayal, J., Georgia Institute of Technology
DeBardeleben, N., LANL
Hanrahan, P., Stanford
Hansen, C., Utah
Hesthaven, J., EPFL
Hjelm, N., LANL
Humphrey, M., U. Va.
Johnson, C., Utah
Joy, K., UCD
Kahn, A., Utah
Kolla, H., SNL
Kordenbrock, T., Hewlett-Packard
Kritz, A., Lehigh
Kroeger, T., Sandia National Laboratories
Kurc, T., Stony Brook University
Lang, M., Los Alamos National Laboratory
Lange, J., University of Pittsburgh
Liao, W. K., NWU
Ludaescher, B., UCD
Ma, K. L. , UCD
Maltzahn, C., University of California Santa Cruz
Meredith, J., ORNL
Miller, E., University of California-Santa Cruz
Miller, N., Argonne National Laboratory
Mohror, K., LANL
Moody, A., LANL
Moser, R., TX.
Nevins, W., LLNL
Oldfield, R., Sandia National Laboratories
Panda, D. K., OSU
Papka, M., ANL
Parker, S., U. Col.
Pascucci, V., University of Utah
Peterka, T., Argonne National Laboratory
Pordes, R., Fermilab
Quinlan, D., LLNL
Rotem, D., LBNL
Saltz, J., Stonybrook SUNY
Sanderson, A., Utah
Seidel, E., Illinios
Settlemyer, B., LANL
Shen, H. W., OSU
Shephard, M., RPI
Silvia, C., NYU
Sugiyama, L., MIT
Taflove, A., NWU
Tang, W., PPPL
Taufer, M., U. Dell
Thapaliya, S., University of Alabama, Birmingham
Tynan, G., UCSD
V. Bangalore, P., University of Albama, Birmingham
Wolf, M., Georgia Tech
Wolpert, D., LANL
Woodring, J., LANL
Wu, K. J., LBNL