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 Collaborators

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 collaborator

Abdelwahed, S., MSU
Adams, M., Columbia
Adams, M., LBNL
Agarawal, A., NWU
Agrawal, G., Ohio State
Ahrens, J., LANL
Aiken, A., Stanford
Akoglu, A., U. Arizona
Al-Kiswany, S., UBC
Allen, G., Illinois
Altintas, I., UCSD
Ames, S., LLNL
Ananthakrishnan, R., U. Chicago
Arpaci-Dusseau, A., UW
Arpaci-Dusseau, R., UW
Avery, P., Florida
Baden, S., UCSD
Bader, D., G. Tech
Bai, Z., U. C. Davis
Balaji, P., ANL
Bangalore, P., U. of Alabama at Birmingham
Banicescu, I., MSU
Bardeen, M., Fermi
Baretto, R., U Chicago
Barrett, B., Amazon
Bell, J., LBNL
Bent, J., EMC
Bell, G., LLNL
Berman, F., RPI
Berriman, G., Caltech
Bethel, W., LBNL
Bhagwan, V., Yahoo!
Boker, S., Virginia
Bradley, S., BNL
Brandt, J., Sandia
Bremer, P. T., LLNL
Brenner, P., Notre Dame
Bridges, P., UNM
Brightwell, R., Sanda
Brugger, E., LLNL
Buck, J., Context Relevant
Callaghan, S., USC
Cappello, F., ANL
Carns, P., ANL
Carothers, C., RPI
Catalyurek, U, OSU
Caton, S., Karlsruhe
Chandy, J., UConn
Chang, C. S., PPPL
Chapin, S., Syracuse
Charles, D., LLNL
Chatterjee, S., ND
Chen, J., Sandia
Childs, H., U. Oregon
Chou, J., N. Tsinghua
Choudhary, A., NWU
Cinquini, L., NOAA
Cole, M., Edinburgh
Cope, J., DDN
Cormier-Michel, E., LBNL
Costa, L., UBC
Crichton, D., NASA
DeBardeleben, N., LANL
Deelman, E., ISI/USC
Denis, J., NCAR
Denvil, S., Institut Pierre Simon Laplace
Dillow, D., Google
Doutriaux, C., LLNL
Drach, R., LLNL
Espinosa, A., U. Chicago
Ethier, S., PPPL
Feibush, E., PPPL
Feng, W., Va. Tech
Figueiredo, R., U. FL
Fiore, S., Euro-Mediterranean Center
Fortes, J., U. FL.
Foster, I., ANL
Fox, G., Indiana
Gadhela, L., UFRJ
Geddes, C., LBNL
Gemmill, J., Clemson U.
Gentile, A., Sandia
Geveci, B., Kitware
Gibson, G., CMU
Gil, Y., ISI/USC
Gokhale, M., LLNL
Gonzalez, E., DKRZ
Grandison, T., Proficiency
Greenwald, M., MIT
Grider, G., LANL
Grimmer, B., IIT
Gropp, W., Illinois
Gu, J., LBNL
Hagen, H., U. Kaiserlautern
Hall, M., U. Utah
Hamann, B., U. C. Davis
Hanrahan, P., Stanford
Hansen, C., U. Utah
Hariri, S., U. Arizona
Harms, K., ANL
Harris, J., Clemson U.
Harrison, R., BNL
He, J., UW Madison
Hemmert, S., Sandia
Hesthaven, J., EPFL
Hill, M., U. W.
Hjelm, N., LANL
Hocky, G., Columbia
Hollingsworth, J., UMD
Hong, N. C., Edinburgh
Humphrey, M., U. VA.
Ioannidou, K., TidalScale
Ionkov, L., LANL
Izaguirre, J., ND
Jaja, J., UMD
Jean-Baptiste, G., FIU
Jian, T., BNL
Jiang, T., BNL
Jin, S., SBU
Johnson, J., IIT
Johnson, C., U. Utah
Johnson, Chris, Utah
Joy, K., U. C. Davis
Kahn, A., U. Utah
Karakoyunlu, C, U. Conn
Kato, S., Nagoya U.
Katramatos, D., BNL
Kaufmann, Arie, SBU
Keleher, P., UMD
Kelly, S., Sandia
Kershaw, P., STFC Rutherford
Kim, J., Texas A&M
Kimpe, D., ANL
Klundt, R., Sandia
Kordenbrock, T., HP
Koziol, Q., HDF Group
Krieder, S., IIT
Kritz, A., Lehigh
Ku, S. H., PPPL
Kurc, Tahsin, Stonybrook
Lang, M., LANL
Lange, J., U. of Pittsburgh
Laros, J., Sandia
Latham, R., ANL
Lauret, J., BNL
Lee, W., PPPL
Leung, V., Sandia
Levenhagen, M., Sandia
Levin, G., Google
Liao, W. K., NWU
Lin, Z., U.C. Irvine
Liu, N., ANL
Liu, X., Microsoft
Luckow, A., BMW
Ludaescher, B., UIUC.
Ma, K. L., U.C. Davis
Ma, X., QCRI
Madduri, K., Penn. State
Mann, R., Edinburgh
Manzanares, A., HGST
Mattmann, C., JPL
Mattoso, M., UFRJ
Nelson, M., Redhat
McClelland, M., Xyratex
McCormick, P., LANL
McKee, S., U. Mich.
Messmer, P., TechX
Meyer, L., UFRJ

Miller, N., ANL
Mohror, K., LLNL
Moody, A., LLNL
Mookherjee, M., Cornell
Morgan, M., P. Simon Lap.
Moser, R., U. Texas Austin
Nam, H., LANL
Nassi, I., TidalScale
Natarajan, V., LBNL
Neale, M., VT
Nevins, W., LLNL
Olabarriaga, S., AMC
Otoo, E., U. Witwatersrand
Panda, D.K., Ohio State
Papka, M., ANL
Parker, S., U. CO
Parker, Scott, U. Col.
Pascoe, S., STFC Rutherford
Pascucci, V., U. Utah
Peterka, T., ANL
Pobre, Z., Goddard Center
Polyzotis, N., Google
Pordes, R., Fermi
Prahbat, , LBNL
Prasanna, V., USC
Pye, I., Cloud Helix
Qiao, C., SUNY
Qin, H., SBU
Quigg, E., Fermi
Quinlan, D., LLNL
Raicu, I., IIT
Rana, O., Cardiff
Ricci, R., U. of Utah
Ripeanu, M., UBC
Robertazzi, T., SBU
Rodrigues, A., Sandia
Roskies, R., PSC
Ross, R., ANL
Rotem, D., LBNL
Ruchti, R., Notre Dame
Sadayappan, P, Ohio State
Saltz, Joel, Stony brook
Samatova, N., NCSU
Sanderson, A., Utah
Seidel, E., Illinois
Settlemyer, B., LANL
Shalf, J., LBNL
Sharma, S., Microsoft
Shen, H., Clemson U.
Shen, H. W., OSU
Shephard, M., RPI
Shewmaker, A., LANL
Shi, L., SBU
Shipman, G., LANL
Shoshani, A., LBNL
Shroff, K., BNL
Silvia, C., NYU
Sim, A., LBNL
Simmhan, Y., Indian Institute of Science
Sinha, R., Microsoft
Skjellum, A., Auburn U.
Skourtis, D., VMWare
Srinivasan, S., U. TX
Striegel, A., Notre Dame
Sugiyama, L., MIT
Sun, X., IIT
Sun, Z., DDN
Sun, W., Google
Sussman, A., UMD
Suziki, K., U. Chicago
Taflove, A., Northwestern
Tang, W., PPPL
Tang, William, PPPL
Taufer, M., U. Dell.
Thain, D., Notre Dame
Thapaliya, S., UAB
Torres, A., Linden Lab
Tseng, C.W., UMD
Tucker, T., Sandia
Tynan, G., UCSD
Vairavanathan, E., UBC
Venkata, P. A., LSU
Vishwanath, V., ANL
Wang, W., PPPL
Wang, Y., Auburn
Wang, L. W., LBNL
Weil, S., Red Hat
Weissman, J., U. Minn
Wheeler, K., Micron
Wheeler, M., U. TX Austin
Williams, D., LLNL
Winslett, M., UIUC
Wolpert, D., LANL
Woodring, J., LANL
Worley, P., ORNL
Wu, F., U. Cal.
Wu, J., LBNL
X., Liu, Microsoft
Xu, W., BNL
Yoo, S., BNL
Yu, W., Florida
Yu, D., BNL
Yu, H., U. Neb.
Yu., W., Auburn
Yun, Z., LSU
Zhang, A., SUNY
Zhang, Z., U. Chicago
Zhao, Y., U. Chicago
Zheng, F., IBM