Title of Pre-application: Hierarchal Extreme Scale Knowledge Management
Principal Investigator: Scott Klasky, Group Leader, ORNL, 854-241-9980, klasky@ornl.gov
Funding Opportunity Announcement Number: DE-FOA-0001338
List of all co-PIs and Key/Senior Personnel
Hasan Abbasi, Oak Ridge National Laboratory
Mark Ainsworth, Oak Ridge National Laboratory JFA, Brown University
Matthew Curry, Sandia National Laboratory
Qing Liu, Oak Ridge National Laboratory
Jay Lofstead, Sandia National Laboratory
Kimmy Mu, Oak Ridge National Laboratory
Carlos Malzahn, U. Cal. Santa Cruz
Manish Parashar, Rutgers University, Oak Ridge National Laboratory
Sudharshan Vazhkudai, Oak Ridge National Laboratory
Lee Ward, Sandia National Laboratory

Objectives: Exascale scientific discovery will be bottlenecked without sufficient new research into managing and storing the large data which will be produced during the simulation, and analzyed for months after the simulations. Our goal is to investigate a unique approach to addressing this challenge in a multi-tier storage platform, expediting insights into mission critical scientific processes. We propose a research program that aims to explore a hiearchical organization and storage infstraturure which will facilitiate how data can be written to and read from a complex heterogenous storage hierarchy efficiently. We will apply a technique which will re-organize and possibly reduce the total amount of information based on user intentions along with direct feedback from the storage and I/O system to allow data to be optimized not just for output, but also for many of the complex reading patterns. Since data analysis and visualization will be performed in both an in situ and post processing environment, we will continue to use our I/O and storage abstractions to help enable applications to take full advantage of the hardware and softare environments. Because data volumes are strictly outstripping available bandwidths, compression will be required. Rather than focusing solely on lossless compression, we will incorporate lossy compression with the ability to annotate data allowing optimized algorithm selection based on the numerical methods employed in the source simulation. We plan to epxlore this tradeoff of storage capacity and bandwidth for computation to determine how to offer the user options that can meet their needs with desireable performance characteristics.

In order to explore the many research challenges in this domain, we will build upon the success of our middleware system, ADIOS, our multi-tier storage system, Sirocco, and our distributed storage system, Ceph. We will also leverage data movement and staging technologies, Data Spaces and the Common Communication Interface (CCI), as part of a software stack used successfully on current leadership class machines. Our experience with this stack will guide us in understanding the new directions for research in storage systems and I/O for exascale and beyond systems.

Our objective here is to reduce the time to knowledge, an end-to-end metric relevant to the scientific discovery process. Beyond the traditional high volume I/O pattern of checkpoint/restart, we will address the challenges posed by additional data access patterns in the knowledge gathering process. With a deeper insight into the scientific process we will encode and utilize errors and accuracy as an optimization parameter. For example,

For example, all scientific simulations contain approximations, and all measurements from obervations and experiments also contain errors. These errors can help guide how information can be presented to users, by allowing us to ask for information within a given accuracy bound, and also allow us to offer a mode for recomputation vs. data storage and reterival.

: please list other things from UCSC, OLCF.

Key Technical Approach:

Our approach will allow users to "plug-in" techniques to classify data, not as bytes but as motifs, where we can understand relationships between objects (into data models) and relationships of data in variables. This will allow us to adapt data from a variable into the various storage tiers. Rather than focus on simple data compression, we will incoroarte a multitude of techniques to reduce data on the faster storage tiers, and keep the less reduced information on the lower tiers.

We wish to support additional data access modes as well. First, within a given accuracy bounds, offer a mode for recomputation from data stored in a "fast" tier to reduce data latencies. Second, exploratory analysis operations frequently entail an overall data set view followed by targeted data exploration based on identified features. We will offer support for a configurable data access mode to support these sorts of analysis accesses. An "overview" access mode that gives a quick, approximate within error bounds, data view that can guide feature selection offering rapid coarse-grained data exploration without requiring loading the complete, detailed data set from storage. Based on the granularity requested, the accuracy and size of the data returned can be adjusted. At the most extreme setting, the original data can be retrieved at the time cost of moving the potentially huge data quantity. Third, to ensure available storage for subsequent operations, we will offer automatic data migration based on user annotations for required data lifetimes using monitoring and learning techniques. Unlike existing approaches, this will be tempered both by the user annotations and through learned access patterns. While past access patterns may not indicate future access because the simulation run purpose may have changed, we are focused on scalability where runs are subsequently larger as the simulation prepares for a capability run. By learning from the output and access patterns during this run sequence, we can accurate decide how to place and organize data for the critical capability runs. Fourth, we anticipiate storing mutliple data copies, each compressed in different ways according to the underlying media, some of these copies will disappear based on storage pressures, but data persistence will be maintained according to user specifications. Assuming a relatively low latency cache layer before a tape system, we can offer exporatory data access reserving pulling data from tape to just the data required. This will save scientists time and make data stored on tape usable without long delays.

Our research efforts will be heavily focused on the need to, in a coordinated manner, adapt data and metadata retention policies 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 middleware and storage layers which can interact well with each other, and take user-provided hints. Today data is reduced by application scientist who have limited information on what the storage layer can provide. They often make compromises based on this limited knowledge and either tune their output for writing or reading performance. This data then gets moved to other locations, and much of the tuning is lost when the data is read back during their post processing. Furthermore, there is a limited set of operations which users will be able to stage to other staging nodes for real-tiime-reduction and visualization.

We will identify, through concrete application evaluation, the requirements for highly usable and scalable middleware and storage layers