SDM Center Technologies for Accelerating Scientific Discoveries

PI: Arie Shoshani[2], Co-PIs: Ilkay Altintas[8], Alok Choudhary[5], Terence Critchlow[7], Chandrika Kamath[3], Bertram Ludäscher[9], Jarek Nieplocha[7], Steve Parker[10], Rob Ross[1], Nagiza Samatova[6], Mladen Vouk[4]

[1]Argonne National Laboratory, [2]Lawrence Berkeley National Laboratory, [3]Lawrence Livermore National Laboratory, [4]North Carolina State University,[5]Northwestern University,[6]Oak Ridge National Laboratory, [7]Pacific Northwest National Laboratory, [8]San Diego Supercomputer Center, [9]University of California, Davis, [10]University of Utah

http://sdmcenter.lbl.gov

Abstract

With the increasing volume and complexity of data produced by ultra-scale simulations and high-throughput experiments, understanding the science is largely hampered by the lack of comprehensive, end-to-end data management solutions ranging from initial data acquisition to final analysis and visualization. The SciDAC-1 Scientific Data Management (SDM) Center succeeded in bringing an initial set of advanced data management technologies to DOE application scientists in astrophysics, climate, fusion, and biology. Equally important, it established collaborations with these scientists to better understand their science as well as their forthcoming data management and data analytics challenges. Our future focus is on improving the SDM framework to address the needs of ultra-scale science during SciDAC-2. Specifically, we are enhancing and extending our existing tools to allow for more interactivity and fault tolerance when managing scientists' workflows, for better parallelism and feature extraction capabilities in their data analytics operations, and for greater efficiency and functionality in users' interactions with local parallel file systems, active storage, and access to remote storage. These improvements are necessary for the scalability and complexity challenges presented by hardware and applications at ultra scale, and are complemented by continued efforts to work with application scientists in various domains.

The Three-Layer Organization of the SDM Center

Managing scientific data has been identified as one of the most important emerging needs by the scientific community because of the sheer volume and increasing complexity of data being collected. Effectively generating, managing, and analyzing this information requires a comprehensive, end-to-end approach to data management that encompasses all of the stages from the initial data acquisition to the final analysis of the data. Fortunately, the data management problems encountered by most scientific domains are common enough to be addressed through shared technology solutions. Based on the community input, we have identified three significant requirements. First, more efficient access to storage systems is needed. In particular, parallel file system improvements are needed to write and read large volumes of data without slowing a simulation, analysis, or visualization engine. To facilitate subsequent access, it is necessary to keep track of the location of the datasets, effectively manage storage resources, and provide secure and efficient data movement. These processes are complicated by the fact that scientific data are structured differently for specific application domains, and are stored in specialized file formats. Second, scientists require technologies to facilitate better understanding of their data, in particular the ability to effectively perform complex data analysis and searches over large data sets. Specialized feature discovery and statistical analysis techniques are needed before the data can be understood or visualized. To facilitate efficient access it is necessary to efficiently query and select subsets of the data. Finally, generating the data, collecting and storing the results, data post-processing, and analysis of results is a tedious, fragmented process. Tools for automation of this process in a robust, tractable, and recoverable fashion are required to enhance scientific exploration.

As part of our evolutionary technology development and deployment process (from research through prototypes to deployment and infrastructure) we have organized our activities in three layers that abstract the end-to-end data flow described above. We labeled the layers as Storage Efficient Access (SEA), Data Mining and Analytics (DMA), and Scientific Process Automation (SPA). The SEA layer is immediately on top of hardware, operating systems, file systems, and mass storage systems, and provides parallel data access technology and transparent access to archival storage. The DMA layer, which builds on the functionality of the SEA layer, consists of indexing, feature selection, and parallel statistical analysis technology. The SPA layer, which is on top of the DMA layer, provides the ability to compose scientific workflows from the components in the DMA layer as well as application specific modules. Figure 1 shows this organization and the components developed by the center and applied to various scientific applications.

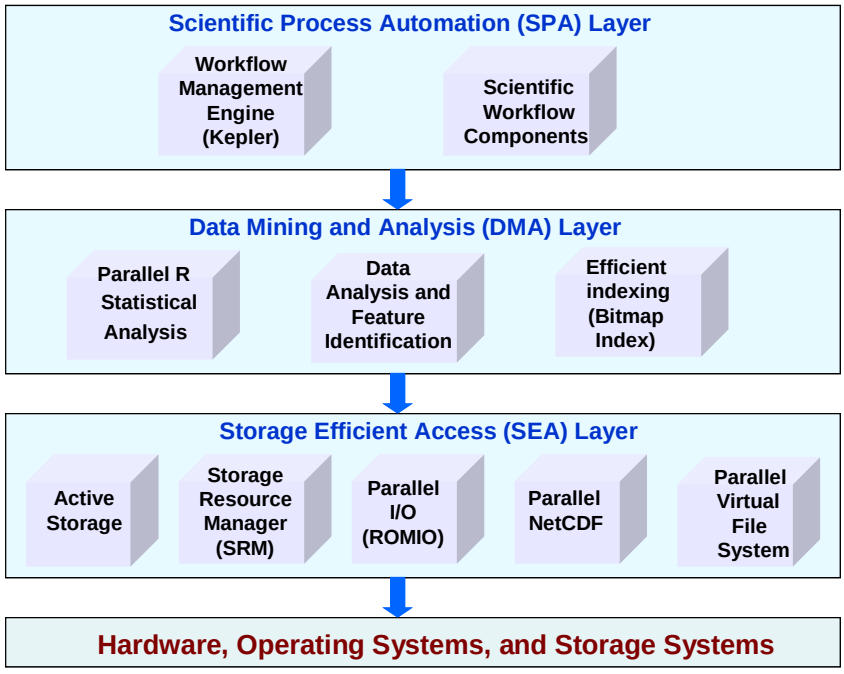

Figure 1: The three-layer organization of technologies in the SDM Center

Descriptions of Technologies

In this section we describe briefly the SDM Center technologies, and include some examples of their application in various application projects. We proceed with technologies from the top layer to the bottom layer.

1. The Kepler Scientific Workflow System

A practical bottleneck for more effective use of available computational and data resources is often the design of resource access and use of processes, and the corresponding execution environments, i.e., in the scientific workflow environment of end user scientists. The goal of the Kepler system is to provide solutions and products for effective and efficient modeling, design and execution of scientific workflows. Kepler is a multi-site open source effort, co-founded by the SDM center, to extend the Ptolemy system (from UC Berkeley) and create an integrated scientific workflow infrastructure. We have also started to incorporate data, process, system and workflow provenance and run-time tracking and monitoring. We have worked closely with application scientists to design, implement, and deploy workflows that address their real-world needs. In particular, we have active users on the SciDAC Terascale Supernova Initiative (TSI) team and an LLNL Biotechnology project.

More recently, we have applied Kepler technology to the Center for Plasma Edge Simulation (CPES) fusion project. CPES is a Fusion Simulation Project whose aim is to develop a novel integrated plasma edge simulation framework. Figure 2, shows a Kepler workflow developed by Norbert Podhorszki (UC Davis) and Scott Klasky (ORNL) to automate coupling XGC-0 and M3D codes. The processing loop within the workflow transfers data regularly from the machine that runs XGC-0 to another machine for equilibrium and linear stability computations. If the linear stability test fails, the workflow system stops the XGC-0 code, a job is submitted to perform nonlinear parallel M3D-MPP computation. Based on the results, XGC-0 can be restarted using the updated data from the nonlinear computation. The restarts will be automatically handled Kepler in the future. This offers a great advantage of avoiding wasting compute resources as soon as they are found to be

unstable. In addition, the workflow system eliminates the need for human monitoring and coordination of task submissions. Another part of the workflow generates intermediate images and status report on the progress of the workflow process. Kepler supports task- and pipeline-parallel execution required for this workflow.

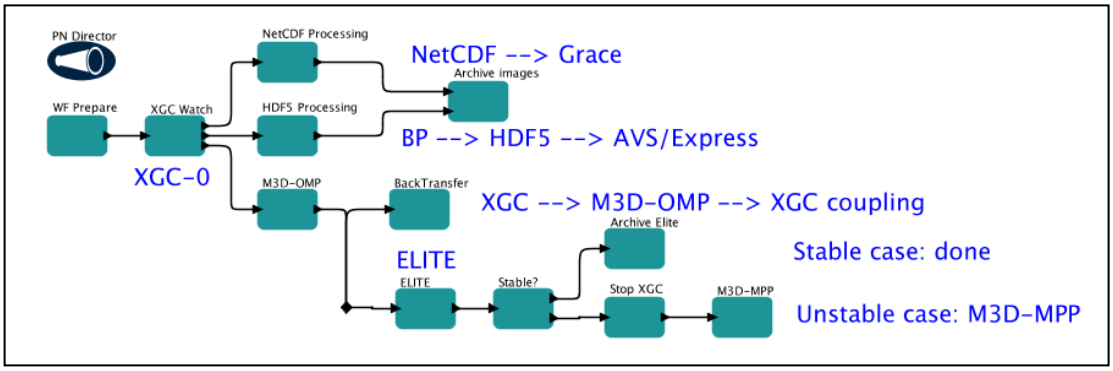

Figure 2. Code coupling workflow steers computations automatically when computations are found unstable, eliminating waste of computational resources

2. Feature Extraction and Tracking

The SDM center is developing scalable algorithms for the interactive exploration of large, complex, multi-dimensional scientific data. By applying and extending ideas from data mining, image and video processing, statistics, and pattern recognition, we are developing at LLNL a new generation of computational tools and techniques that are being used to improve the way in which scientists extract useful information from data. These tools are being applied to problems in a variety of application areas, including separation of signals in climate data from simulations, the identification of key features in sensor data from the D-III-D Tokamak, and the classification and characterization of orbits in Poincaré plots in Fusion data. Recently, these technologies are being applied to identifying the movement of "blobs" in images form fusion experiments. A blob is a coherent structure in the image that carries heat and energy from the center of the torus to the wall. Figure 3 shows bright blobs extracted from experimental images from the National Spherical Torus Experiment (NSTX). The blobs are high energy regions. If they hit the torus wall that confines the plasma, it can vaporize. The figure shows movement of the blobs over time. A key challenge to the analysis is the lack of a precise definition for these structures. Top row is the original image after removing camera noise. Second row is after removal of ambient or background intensity, which is approximated by the median of the sequence. In the third row, we use image processing techniques to identify and track the blobs over time. The goal is validate and refine the theory of plasma turbulence.

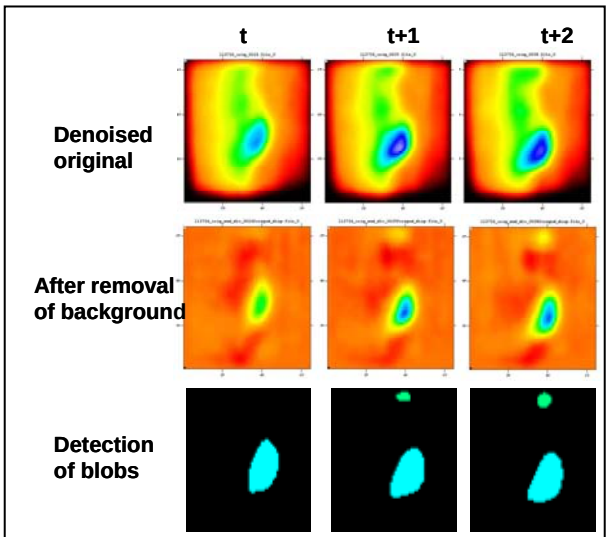

Figure 3: tracking of "blobs" in Fusion images

3. Parallel Statistical Analysis

Present data analysis tools such as Matlab, IDL, and R, even though highly advanced in providing various statistical analysis capabilities, are not apt to handle large data-sets. Most of the researchers' time is spent on addressing data preparation and management needs of their analyses. Parallel R is an open source parallel statistical analysis package developed by the SDM center at ORNL, that lets scientists employ a wide range of

statistical analysis routines on high performance shared and distributed memory architectures without having to deal with the intricacies of parallelizing these routines. Parallel R lets scientists employ a wide range of statistical analysis routines on high performance architectures without having to deal with the intricacies of parallelizing these routines. Through Parallel R the user can distribute data and carry out the required parallel computation but maintain the same look-and-feel interface of the R system. Two major levels of parallelism are supported: data parallelism (k-means clustering, Principal Component Analysis, Hierarchical Clustering, Distance matrix, Histogram) and task parallelism (Likelihood Maximization, Bootstrap and Jackknife Re-sampling, Markov Chain Monte Carlo, Animations). Figure 4 shows an example of the minimal changes to the analysis code to support task and data parallelism.

pR script

```
A ← matrix (1:10000, 100,100)
library (pR)
PE (
S ← sla.svd(A)          Data parallel
b ← list ()
for (k in 1:dim (A) [ 1 ] ) {
   b [ k ] ← sum ( A [ k, ] )
}        Embarrassingly parallel
m ← mean ( A )
d ← sum ( A )            Task parallel
)
```

Figure 4: Example of task and data parallel script

4. Scientific Data Indexing

As the volume of data grows, there is an urgent need for efficient searching and filtering of large-scale scientific multivariate datasets with hundreds of searchable attributes. FastBit is an extremely efficient bitmap indexing technology, developed at LBNL that uses a CPU-friendly bitmap compression technique called the Word-Aligned Hybrid (WAH) code. Unlike other bitmap indexes that assume low cardinality of possible data values, FastBit is particularly useful for scientific data, since it is designed for high-cardinality numeric data. FastBit performs 12 times faster than any known compressed bitmap index in answering range queries. Because of its speed, Fastbit facilitates real-time analysis of data, searching over billions of data values in seconds. FastBit has been applied to several application domains, including finding flame fronts in combustion data, searching for rare events from billion of high energy physics collision events, and more recently to facilitate query-based visualization. Figure 5 shows a 3D histogram over the IP address space and time, to identify malicious network traffic attacks. The query-driven visualization reveals consecutive regions that represent coordinated attacks. It was obtained in real time by using FastBit.

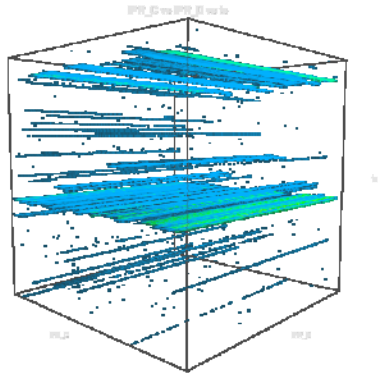

Figure 5: identifying network attacks using FasBit indexing

5. Advanced I/O Infrastructure

Today's scientific applications demand that high performance I/O be part of their operating environment. These applications access datasets of many gigabytes (GB) or terabytes, checkpoint frequently, and create large volumes of visualization data. Such applications are hamstrung by bottlenecks anywhere in the I/O path, including the storage hardware, file system, low-level I/O middleware, and application level interface. Just above the I/O hardware in a high-performance machine sits software known as the parallel file system. At ANL, as part of the SDM center activity, a parallel file system, called PVFS, was developed to address these needs. PVFS can provide multiple GB/second parallel access rates, and is freely available. Above the parallel file system is software designed to aid applications in more efficiently accessing the parallel file system. Implementations of the MPI-IO interface are arguably the best example of this type of software. MPI-IO provides optimizations that help map complex data movement into efficient parallel file system operations. Our ROMIO MPI-IO interface implementation is freely distributed and is the most popular MPI-IO implementation for both clusters and a wide variety of vendor platforms. MPI-IO is a powerful but low-level interface that operates in terms of basic types, such as floating point numbers, stored at offsets in a file. However, some scientific applications desire more structured formats that map more closely to the structures applications use, such as multidimensional datasets. NetCDF is a widely used API and portable file format that is popular in the climate simulation and data fusion communities. As part of the work in the SDM center, a parallel version of NetCDF (pNetCDF) was developed by NWU and ANL. It provides a new interface for accessing NetCDF data sets in parallel. This new parallel API closely mimics the original API, but is designed with scalability in mind

and is implemented on top of MPI-IO. Performance evaluations using micro-benchmarks as well as application I/O kernels have shown major scalability improvements over previous efforts. Upcoming systems, such as shown schematically in Figure 6, will incorporate hundreds of thousands of compute processors along with a collection of support nodes. Using POSIX and MPI-IO interfaces, I/O operations are forwarded through a set of I/O nodes to storage targets. On the Argonne Leadership Computing Facility, the PVFS parallel file system, supported under the SDM Center, will provide a system-wide storage space for applications storing hundreds of terabytes of data.

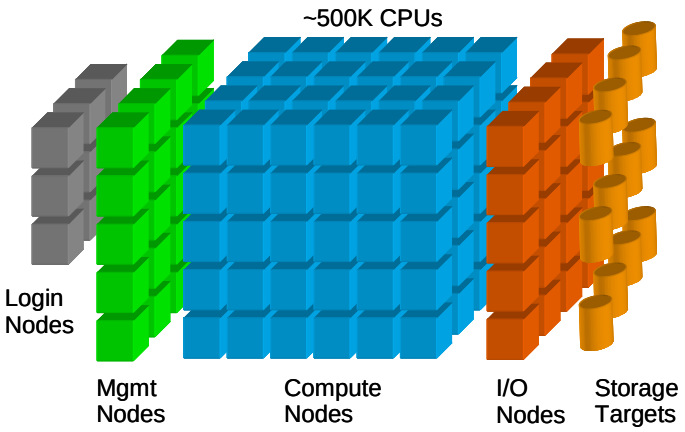

Figure 6: Allocation of processors on large scale machines to support various support functions, including I/O

6. Active Storage

Despite recent advancements in storage technologies for many data intensive applications, analysis of data remains a serious bottleneck. In traditional cluster systems, I/O-intensive tasks must be performed in the compute nodes. This produces a high volume of network traffic. One option for data analysis is to leverage resources not on the client side, but on the storage side referred to as Active Storage. The original research efforts on active storage were based on a premise that modern storage architectures might include usable processing resources at the storage controller or disk; unfortunately, commodity storage has not yet reached this point. However, parallel file systems offer a similar opportunity. Because the servers used in parallel file systems often include commodity processors similar to the ones used in compute nodes, many Giga-op/s of aggregate processing power are often available in the parallel file system. Our goal, in the Active Storage project at PNNL, is to leverage these resources for data processing. Scientific applications that rely on out-of-core computation are likely candidates for application of this technique, because their data is already being moved through the file system. The Active Storage approach allows moving computations involving data stored in a parallel file system from the compute nodes to the storage nodes. Benefits of Active Storage include: low network traffic, local I/O operations, and better overall performance. The SDM center has implemented Active Storage on Lustre and PVFS parallel file systems. We plan to pursue deployment of Active Storage in biology or climate application.

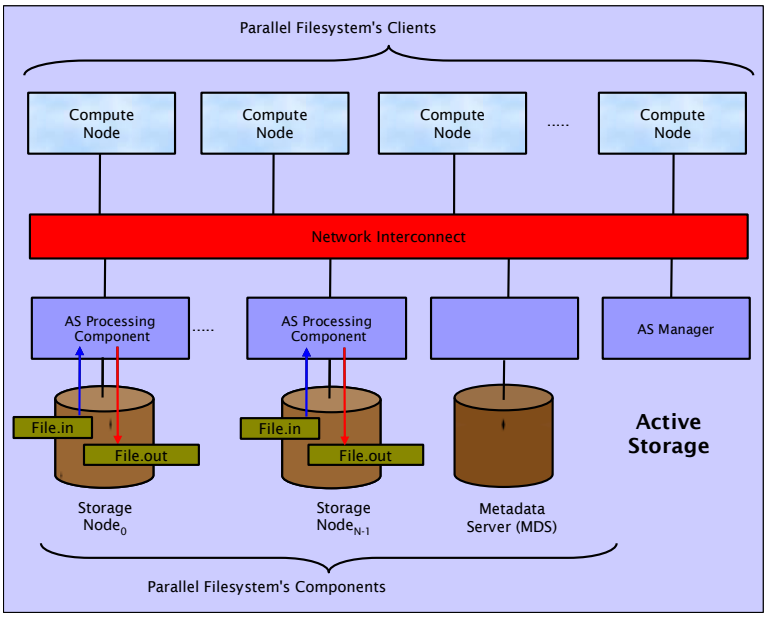

Figure 7: The Active Storage architecture for parallel file systems

Publications

Numerous publications on the technologies described here are available in the SDM center web site (http://sdmcenter.lbl.gov) under “publications”.