The Scientific Data Management Center: Available Technologies and Highlights

A. Shoshani, I. Altintas, J. Chen, G. Chin, A. Choudhary, D. Crawl, T. Critchlow, K. Gao, B. Grimm, H. Iyer, C. Kamath, A. Khan, S. Klasky, S. Koehler, S. Lang, R. Latham, J. W. Li, W. Liao, J. Ligon, Q. Liu, B. Ludaescher, P. Mouallem, M. Nagappan, N. Podhorszki, R. Ross, D. Rotem, N. Samatova, C. Silva, A. Sim, R. Tchoua, R. Thakur, M. Vouk, K. Wu, W. Yu

Managing scientific data has been identified by the scientific community as one of the most important emerging needs 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 the stages from the initial data acquisition to the final analysis of the data. Based on community input, we have identified three significant requirements. First, more efficient access to storage systems is needed. In particular, parallel file system and I/O system improvements are needed to write and read large volumes of data without slowing a simulation. Second, scientists require technologies to facilitate better understanding of their data, in particular the ability to effectively perform complex data analysis and searches over extremely large data sets. Furthermore, exploratory analysis requires techniques for efficiently selecting subsets of the data. Third, generating the data, collecting and storing the results, keeping track of data provenance, data postprocessing, 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.

SDM center technology layers

Based on these needs, the SDM center has developed and deployed a collection of technologies that are organized as three technology layers. We labeled the layers (from bottom to top) Storage Efficient Access (SEA), Data Mining and Analysis (DMA), and Scientific Process Automation (SPA). Figure 1 shows these layers and the technologies deployed in each layer.

In this paper we describe the key technologies developed and deployed by various scientific applications and now available as open source software. We also include selected highlights that had significant impact on the applications. Because of space limitation, no references are provided in this document; however, all papers can be found at https://sdm.lbl.gov/, under "publications."

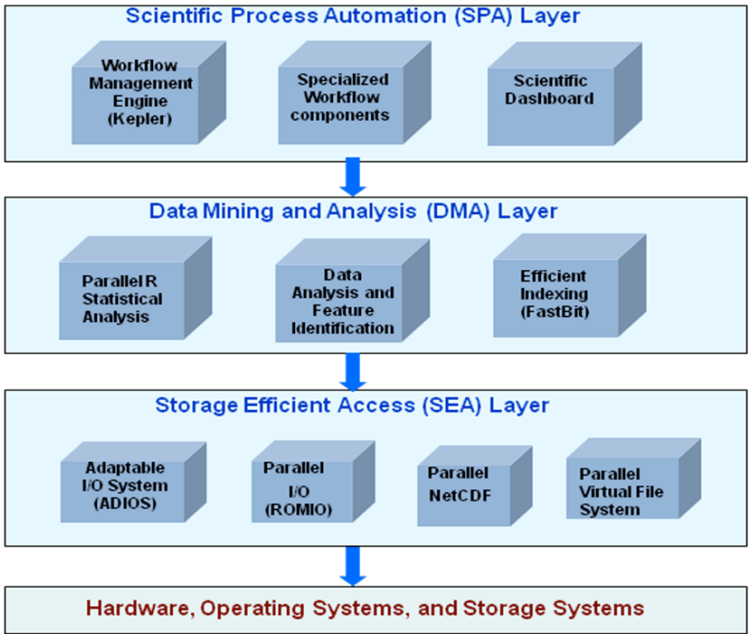


Figure 1: Layers and technologies developed by SDM center.

Highlight: Three books published by members of the SDM center. Members of the SDM center edited and contributed chapters to the book titled *Scientific Data Management: Challenges, Existing Technology, and Deployment*, published in 2009. A second book, titled *Scientific Data Mining: A Practical Perspective*, was authored by a member of the SDM center and published in 2009. A third textbook, titled *Practical Graph Mining with R,* was written entirely by students under the guidance of members of the SDM center; it is scheduled to be published by the end of 2011.

Storage Efficient Access (SEA) Technologies and Highlights

Technology: Parallel Virtual File System (PVFS)

PVFS brings state-of-the-art parallel I/O concepts to production parallel systems. It is designed to scale to petabytes of storage and provide access rates at hundreds of gigabytes per second. Its main features are: performance, reliability, optimized MPI-IO support, hardware independence, and painless deployment. For details and downloads, see http://www.pvfs.org/.

Technology: ROMIO: A high-performance, portable MPI-IO implementation

ROMIO is a high-performance, portable implementation of MPI-IO, the I/O chapter in MPI-2. Version 1.2.5.1 of ROMIO is freely available. For details and downloads, see http://www.mcs.anl.gov/romio.

Highlight: Keeping pace with architectural change in I/O libraries

System architectures have rapidly changed over the past five years, and will continue to do so as we move into the exascale era. While much attention is paid to the performance of storage and file systems as new systems are deployed, the success of these storage and file systems is dependent on additional software—I/O libraries—that bridge between applications and these lower software layers. The SDM center is active in MPI-IO development, working with vendors and the science community to ensure that MPI-IO implementations are efficient and scale on new architectures. This work includes the development and integration of new collective I/O algorithms into the ROMIO MPI-IO implementation (Figure 2), as well as new algorithms to reduce the time to open files and improvements to limit the memory resource requirements of MPI-IO implementations for specific workloads. Equally important, our work on ROMIO provides the infrastructure for other research and development teams to investigate new approaches to HPC I/O, facilitating broader attention to these challenges. Overall the MPI-IO activities in the SDM center have resulted in scalable MPI-IO performance on a wide variety of DOE platforms.

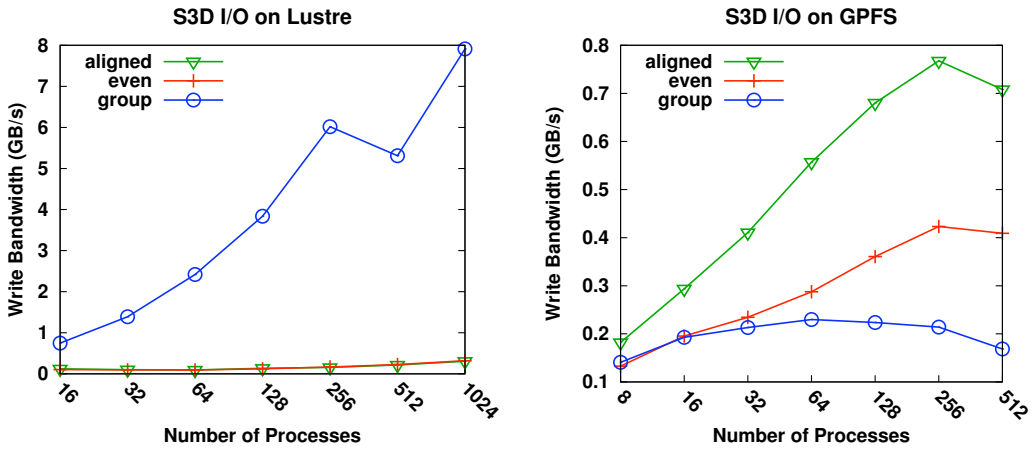


Figure 2: Collective I/O algorithms has shown the importance of matching collective I/O algorithms to underlying file system. On the left, the group cyclic algorithm provides highest performance for Lustre file systems, while on the right, simply aligning accesses with block boundaries provides the highest performance on GPFS file systems.

Technology: Parallel-NetCDF

Parallel-NetCDF is a library providing high-performance I/O while still maintaining file-format compatibility with Unidata's NetCDF. NetCDF gives scientific programmers a space-efficient and portable means for storing data. However, it does so in a serial manner, making it difficult to achieve high

I/O performance. By making some small changes to the API specified by NetCDF, we can use MPI-IO and its collective operations. For details and downloads, see http://www.mcs.anl.gov/parallel-netcdf.

Highlight: Improving I/O performance for climate simulation applications using Parallel-netCDF

Parallel-NetCDF (PnetCDF)—designed, built, and supported by SDM center members—is now used in several production codes. Recently, it has been successfully used by the large-scale Global Cloud Resolve Model (GCRM) and the Community Climate System Model (CCSM). The new PnetCDF file format, called "CDF-5," now supports an array size larger than 4 GB, which allows storage of variables of effectively unlimited size in the netCDF format. Recently, an optimization was developed to enable data aggregation for multiple, small-sized requests that can better utilize I/O bandwidth on modern parallel computers. A significant performance improvement is observed over the current best I/O method. Figure 3 shows up to 140% improvement in terms of write bandwidth.

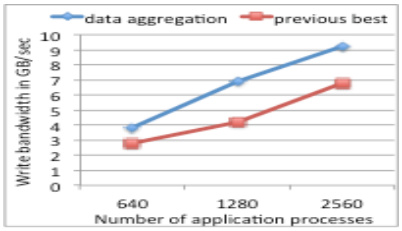


Figure 3: Evaluation of GCRM I/O performance with Parallel-NetCDF data aggregation.

Technology: The Adaptable IO System (ADIOS)

ADIOS provides a simple, flexible way for scientists to describe the data in their code that may need to be written, read, or processed outside the running simulation. By providing an external to the code XML file describing the various elements, their types, and how one wishes to process them this run, the routines in the host code (either Fortran or C) can transparently change how they process the data. For details and downloads, see http://www.olcf.ornl.gov/center-projects/adios.

Highlight: Speeding I/O with ADIOS on Cray XT, InfiniBand clusters, and IBM Blue Gene/P

ADIOS has been incorporated into many scientific simulations and has been used to greatly speed both writing and reading of large-scale simulation data. New methods available in ADIOS allow application scientist to plug in visualization services, such as ParaView and VisIt, and run these services in the staging area, a set of nodes reserved for I/O pipelines. ADIOS supports many file formats, including ADIOS-BP (binary-packed), NetCDF4, and HDF5, along with many other custom formats. ADIOS addresses concerns for exascale I/O by focusing on a framework that can couple service together in an asynchronous fashion, using a staging mechanism. By allowing plug-ins such as services that generate images and deliver them to the SDM dashboard, scientists can closely monitor simulations without perturbing their simulation. ADIOS has been integrated into various codes including Chimera, GTC, GTS, XGC-1, and S3D codes and has achieved up to 31 GB/s on a 40 GB/s file system. More recently, ADIOS has shown a writing speedup of the S3D and SCEC PMCL3D by over 8X from the original I/O; see Figure 4.

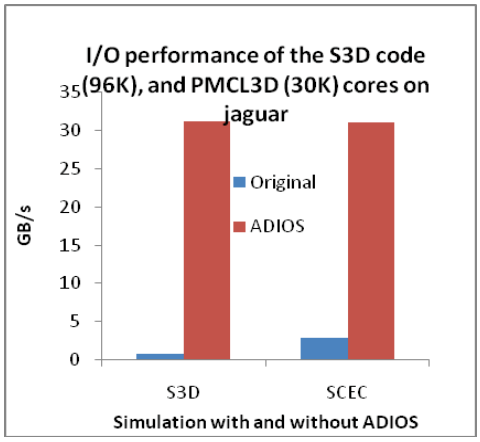


Figure 4: I/O speedup with ADIOS.

Data Mining and Analysis (DMA) Technologies and Highlights

Technology: FastBit indexing

FastBit is a very efficient indexing technology for accelerating database queries on massive datasets. FastBit has been proven to be theoretically optimal; it performs 50–100 times faster than any known indexing method based on its use of our patented compression method. It can search over multivariable, scientific data where attributes have high cardinality (number of possible values). It received an R&D 100

award in 2008. FastBit is available under an open source license. For details and downloads, see http://sdm.lbl.gov/fastbit.

Highlight: Gaining 1000-fold search speedup with FastBit in two applications

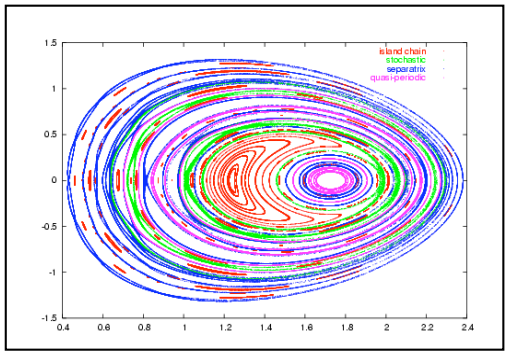
A huge (1000-fold) speedup of particle search for the Laser Wakefield Particle Accelerator project was attained with FastBit. We used FastBit to speed up the operations of searching and tracking particles in Laser Wakefield Particle Accelerator (LWPA) project (joint with VACET center). By replacing an existing IDL-based analysis program with a FastBit-based program, we observed a speedup of 3 orders of magnitude (from 300 seconds to 0.3 seconds) in the first test run. Another 1000-fold speedup was achieved with specialized FastBit structures for a fusion code, the Gyrokinetic Transport Code. By taking advantage of the regularity of the magnetic field-line following mesh in the magnetic coordinate system used for fusion simulation, and by using the compact data structures from FastBit software, we are able to find special regions of interest much faster.

Technology: Sapphire - Scientific data-mining software

Sapphire provides scalable algorithms for the interactive exploration of large, complex, multidimensional scientific data. By applying and extending ideas from data mining, image and video processing, statistics, and pattern recognition, a new generation of computational tools and techniques is being used to improve the way in which scientists extract useful information from data. Sapphire received an R&D 100 award in 2006. For details and downloads, see http://www.llnl.gov/casc/sapphire.

Highlight: Automatically classifying orbits in Poincaré plots

We developed software to extract representative features from the x-y coordinates of points in orbits of a Poincaré plot and used these features in a patented decision tree algorithm to obtain a classification accuracy of 96%. This automated technique replaces a tedious and error-prone manual classification of the orbits. Figure 5 shows an example of the automatic classification of orbits in Poincaré plots deployed for use by fusion scientists.



Highlight: Detecting coherent structure in GSEP simulations

We have been analyzing the fluid data from the GSEP fusion SciDAC center to detect and track the coherent structures in different variables. Our analysis indicated the unexpected presence of negative structures in the ion heat flux variable in the collisionless trapped electron mode simulations; further analysis confirmed that these were due to physics and not noise. Our on-going work to understand these structures has also indicated the presence of small, paired structures with high positive and negative values in the electron heat flux variable; these are under further investigation.

Technology: Parallel R (pR)

Parallel R (pR) is middleware for an easy-to-use, almost-zero-overhead plug-in of parallel analysis functions written in compiled languages into the widely used open source R statistical environment. For example, in contrast to RScaLAPACK, which is designed specifically for ScaLAPACK, the pR architecture has been abstracted to allow parallel third-party analysis functions to integrate with the R environment without requiring major modifications to either pR or the external third-party libraries. For details and downloads, see http://www.r-project.org/.

Highlight: Providing superlinear scaling with pR

The pR middleware delivered superlinear scalability in terms of the number of processors and improved the performance of the state-of-the-art technology by an average factor of 37. Its RScaLAPACK library is distributed as an RPM package across different Linux distributions and in more than 30 countries world-wide through the R's CRAN distribution site. Parallel R forms a server-side analysis engine, with a select set of analysis routines in the Dashboard web application. The initial set of routines was identified based on their frequent use by climate and fusions communities. We also extended the capability of RScaLAPACK library to support an Open MPI back-end in response to multiple users' requests, eased RScaLAPACK's installation via improved autoconf, provided processor grid manipulation routines, and provided both static and dynamic MPI library support.

Scientific Process Automation (SPA) Technologies and Highlights

Technology: Kepler – a scientific workflow system

The Kepler scientific workflow system is developed by a cross-project collaboration including the SciDAC SDM center to serve scientists from different disciplines. Since its initiation in 2003, over twenty diverse projects encompassing multiple disciplines have used Kepler (adding up to more than 40,000 downloads) to manage, process and analyze scientific data. Inherited from Ptolemy II, Kepler adopts the actor-oriented modeling paradigm for scientific workflows. The workflows that are designed through Kepler's graphical user interface can then be executed through the same user interface or in batch mode from other applications. In addition, Kepler provides a provenance framework that keeps a record of chain of custody for data and process products within a workflow design and execution. For details and downloads, see http://kepler-project.org/. More details on features of scientific workflow capabilities developed by the SDM center are provided in a companion paper in this issue titled "Working with Workflows: Highlights from 5 years Building Scientific Workflows."

Technology: Electronic Simulation Monitoring Dashboard (eSiMon)

The eSiMon dashboard combines fast visualization with Web access, the ability to compare images from multiple simulation runs, and the ability to display movies composed from multiple images produced by the workflow system. Using provenance recording, users can download/transfer the original dataset by selecting an image, or start an analytics job (like Matlab) for the dataset. Vector graphics allows for interactive analysis in the web browser. For details and downloads, see http://www.olcf.ornl.gov/center-projects/esimmon/.

Highlight: Eliminating unnecessary computations with an integrated framework for real-time monitoring of large-scale simulations.

The SDM center has developed an integrated framework, currently being used in production runs by fusion and combustion scientists. The technologies provided by the center include ADIOS, the Kepler workflow system, a dashboard, provenance tracking and recording, and parallel analysis capabilities. This integrated system has been used to perform simulation monitoring in real time, as well as complex code-coupling tasks. Monitoring includes dynamic generation of graphs and images posted on the dashboard (see Figure 6).

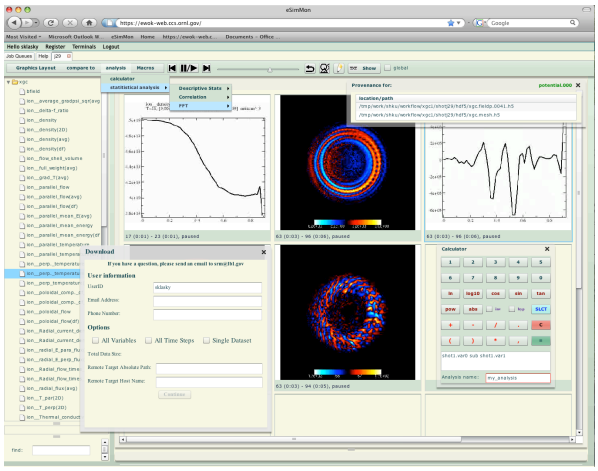


Figure 6: Results from a fusion simulation shown on the eSiMon dashboard.