SDM Center FY 2009 Annual Report

http://sdmcenter.lbl.gov

Table of Contents

SDM Center FY 2009 Annual Report

http://sdmcenter.lbl.gov

Introduction

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 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 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, analysis, or visualization engine. 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 extremely large data sets. Specialized feature discovery and statistical analysis techniques are needed before the data can be understood or visualized. Furthermore, interactive analysis requires techniques for efficiently selecting subsets of the data. Finally, generating the data, collecting and storing the results, keeping track of data provenance, 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.

Our approach is to employ an evolutionary development and deployment process: from research through prototypes to deployment and infrastructure. Accordingly, we have organized our activities in three layers that abstract the end-to-end data flow described above. We labeled the layers (from bottom to top):

- Storage Efficient Access (SEA)
- Data Mining and Analysis (DMA)
- 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 identification, 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.

Over the last year we have focused on enhancing the existing SDM tools in response to application scientists' requirements. These advances were possible since we had close interactions with many scientific application users, and their feedback to the efficacy and usefulness of our tools was essential. The table in Appendix 2 summarizes these interactions in a table format. This report only covers progress in the 2009 fiscal year. Comprehensive reports on progress in previous years are available at the SDM center web site.

Organization of the report

This report consists of the following sections, organized according to the three layers, as follows.

The **Storage Efficient Access** (SEA) area includes the following activities: (1) file system benchmarking and application I/O behavior; (2) parallel I/O infrastructure evolution; (3) application interfaces and data models; and (4) next-generation I/O software technologies.

The **Data Mining and Analysis** (DMA) area includes the following activities: (1) high-performance statistical computing; (2) efficient searching and filtering in data-intensive scientific applications; and (3) Feature extraction and tracking in scientific applications.

The **Scientific Process Automation** (SPA) area includes the following activities: (1) workflow development; (2) provenance collection; (3) generic actors; and (4) workflow fault tolerance.

In addition to the three sections covering progress in the three focus areas above, we include an additional section that describes our efforts in providing an **Framework for Integrated SDM Technologies for Applications** (referred to as FIESTA) that uses multiple SDM center technologies for a specific SciDAC Fusion application project, called CPES (Center for Plasma Edge Simulation). The technologies used are from all three areas, and include workflow, analysis, I/O speed up, data movement technologies, and visual data analysis. FIESTA was designed in collaboration with application users to provide them with sophisticated and powerful capabilities accessed through intuitive web interfaces. While FIESTA was designed in response to the CPES project, it was developed as a general framework that can be used in other application domains. We are currently engaging Combustion and Astrophysics scientists who expressed interest in using this framework as well. This is described in section 3.6.

Details of progress in each of the three areas, SEA, DMA, and SPA, as well as EIF, follow. This is followed by a **publications** and references section, and outreach, **tutorials**, and invited talks in ***Appendix 1***. The SDM center has have developed numerous **collaborations** with application projects and other centers and institutes. This is summarized in a color-coded table form in ***appendix 2***, as well as a summary description of the collaborative tasks. But, first we describe selected highlights that are discussed in these sections in more detail.

Selected Highlights of Achievements

- **Parallel NetCDF successfully used in production.** 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 NCAR Community Atmosphere Model (CAM). According to Yu-heng Tseng, Department of Atmospheric Sciences National Taiwan University, "Parallel NetCDF indeed solves a big problem on the large scale computing." A new parallel netCDF file format to support larger than 4 GB array size was also developed and is being tested extensively. The new format, called the "CDF-5" format, allows storage of variables of effectively unlimited size in the netCDF format. See details in Section 1.3.
- **400% I/O improvement achieved for collective I/O patterms on the Lustre parallel file system.** We enhanced I/O efficiency for the Lustre file system by as much as 400% by introducing a novel technique called partitioned collective IO (ParColl). ParColl partitions parallel processes into subgroups, each carrying out smaller, yet aggregated IO operations. This technique is important in part because it does not require a change in file format. See details in Section 1.2.
- **New high performance driver for Lustre developed and integrated into popular packages.** We have developed a Lustre driver for MPI-IO that enables higher performance by better tuning access to avoid performance pitfalls with Lustre file system on the Cray machines. This driver has been integrated into MVAPICH version 1.0, which is a popular MPI implementation for InfiniBand clusters, and MPICH2 version 1.0.7, which serves as the basis of most vendor implementations.

Through these distributions, this technology will enhance I/O performance for a variety of applications using MPI-IO directly or through such high-level I/O libraries as HDF5 and PnetCDF. See details in Section 1.2.

- **FastBit indexing technology received R&D 100 award.** 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 multi-variable, scientific data where attributes have high cardinality (number of possible values). These unique characteristics made it useful in a variety of scientific applications. Our implementation was packaged in 2007 and released under an open-source license, and has attracted a lot of interest in multiple scientific applications, as well as new areas, such as network traffic data analysis, and query-based visualization. It received the prestigious R&D 100 award in 2008. See details in Section 2.2.

- **1000 times speedup of particle search for the Laser Wakefield Particle Accelerator project 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 Visualization group). By replacing an existing IDL based analysis program with a FastBit based program, we observed a three orders of magnitude speedup (from 300 seconds to 0.3 seconds) in the first test run. In another application, FastBit open source technology was used (without involving the FastBit developers) in a software called TrixX-BMI. It has been reported that FastBit enabled screening libraries of ligands 12 times faster than the state of art screening tools. See details in Section 2.2.

- **An accurate tool for classification of orbits in Poincare plots developed and deployed.** Software for the automatic classification of orbits in Poincare plots was developed and deployed for use by PPPL and other fusion scientists, solving a long-standing problem for scientists in this domain. It takes as input the coordinates of the points in an orbit and assigns to it one of four class labels based on the shape traced out by the points. Key challenges to solving this problem were the extraction of robust features representing the orbits and the creation of a high-quality training set. The cross-validation error rate using ensembles of decision trees is less than 4% and the code "works quite well" in the words of a physicist using it, who has recommended it to colleagues. This software would replace tedious manual labeling of the orbits, which is often error-prone and subjective. See details in section 2.3.

- **Parallel R (pR) for high performance statistical computing delivered super-linear scaling.** We have developed parallel R (pR) middleware for an easy-to-use almost-zero-overhead plug-in of parallel analysis functions written in compiled languages into a widely used open source R statistical environment. pR delivered a super-linear 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 usage by climate and fusions communities. pR is discussed further in Section 2.1.

- **ProRata enabled systems biology studies in various DOE energy and environment applications.** We have brought to production our open source ProRata robust statistical software with the GUI for high-throughput quantitative shotgun proteomics. ProRata has been downloaded more than 1,000 times and has been used by the DOE Bioenergy centers and Genomics:GTL projects to predict the composition of the cellulosomal complex for biomass degradation and to perform genome-scale functional annotation of ethanol-producing bacteria, to infer metabolic aromatic compound degradation pathways by hydrogen-producing bacteria, and to understand the composition of complex microbial communities from environmentally-hazardous sites. ProRata is discussed in Section 2.3.

- **The Kepler developers hosted and collaborated with the ITER team.** The SDM Center collaborated with the ITER European Integrated Tokamak Modelling project team at the Institute of Fusion Research, France. This group has selected Kepler, the workflow system developed by the SDM center, for their workflow development, and visited twice to coordinate their work with the Kepler development team. During the visits by ITER project team members in 2007 and 2008, we shared our developments of various components of interest to the ITER teams. Kepler development is discussed in Section 3.1.
- **Automatic provenance capture framework developed.** Run-time provenance capture scripts and automatic data-base feed have been developed to be used with the Kepler workflow system. This includes system provenance on the setup of simulation programs, the workflow provenance, and data provenance that captures the history of each data file. The data provenance is now used to find files of interest and move them to the user machine directly from the dashboard, which is used to monitor simulations, and support remote analysis of simulation data. See details in Section 3.4 and Section 4.
- **Integrated framework for real-time monitoring of large-scale simulations eliminated unnecessary computations.** The SDM center has developed an integrated framework, currently being used in production runs by the Center for Plasma Fusion Edge Simulation (CPES) scientists. The technologies provided by the center include the Kepler workflow system, a dashboard, provenance tracking and recording, parallel analysis capabilities, and SRM-based data movement. The Dashboard now has fast visualization with Web access, and other features, including the ability to compare images from multiple time-steps (shots), and display movies composed from multiple images by the workflow system. 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. In a recent review of the CPES project, all the reviewers gave the project the highest possible grade of "excellent." See details in Section 4.
- **Two books to be published by members of the SDM center.** Members of the SDM center edited and contributed chapters to an upcoming book, entitled "Scientific Data Management: Challenges, Existing Technology, and Deployment" [SR09]. In six out of thirteen chapters, the lead authors are members of the SDM center, and additional members contributed to the content of these chapters. The second book, entitled "Scientific Data Mining: A Practical Perspective," was also authored by a member of the SDM center [Kam09].
- **High productivity in the SDM center.** During SciDAC-2 (October 2006 – present) members of the SDM center published **61** papers (see publication list), organized and presented **16** tutorials, invited talks, or invited sessions.

1. Storage Efficient Access (SEA)

The core I/O functionality on today's high-performance computing (HPC) systems consists of a collection of I/O software that provides a convenient and efficient interface to the available I/O hardware. The projects in this layer focus on this core I/O functionality, and they have two complementary goals. First, we develop and support a collection of highly-scalable and freely available I/O software components that are used in production applications by scientists, and we actively engage the community to help application scientists better understand how to use these tools. Second, through our interactions with the community we identify specific deficiencies in functionality, performance, and usability that we then work to address. Successful improvements are subsequently integrated into production releases, ensuring that these benefits are made widely available.

Overall, our work can be placed in four categories, discussed in the following sections:

- File system benchmarking and application I/O behavior

- Parallel I/O infrastructure evolution
- Application interfaces and data models
- Next-generation I/O software technologies

1.1 File System Benchmarking and Application I/O Behavior

The high peak rates of HPC I/O systems simply do not translate into adequate sustained performance for computational science applications. The root cause of this performance gap is the mismatch between the requirements of the system's applications and the capabilities of I/O hardware and software. Systematic evaluation of both I/O system capabilities and application requirements provides much needed insight into the efficient use of existing systems and help guide the design of next-generation I/O systems.

The objective of this work is to study file system characteristics that have significant impacts to the parallel I/O operations and evaluate the relative performance of the file systems available to important SciDAC applications on DOE compute platforms. The performance, functionality, and scalability of MPI-IO, parallel netCDF, and HDF5 are critical for many applications

Progress to Date

We have studied the Lustre and GPFS parallel file systems with respective to their file locking protocols. Both file systems adopt the extent-based locking mechanism, but implement a very different distributed lock management. Therefore, our study concludes that the high-level parallel I/O library, such as MPI-IO, must take consideration of the file locking behavior of underlying file system. We have developed several file domain partitioning methods in MPI-IO library to improve collective I/O performance. The methods include file stripe boundary alignment, static-cyclic stripe assignment, and group-cyclic stripe assignment [LC08]. By evaluating these methods using FLASH and S3D I/O kernels on Jaguar at ORNL, Franklin at NERSC, and the TeraGrid IBM IA-64 cluster at NCSA, we found that the best file domain partitioning method for GPFS is the stripe boundary alignment and the best for Lustre is the group-cyclic stripe assignment. We observed up to 30 times write bandwidth improvement on Jaguar for Lustre and up to 4 times improvement for GPFS on the IBM IA-64 cluster. Such performance enhancement is achieved without changing the source of the applications. The stripe boundary alignment method has been incorporated in the release of MPICH version 2-1.0.8 as a general collective I/O optimization [LC08, CCS+09].

Because of its high throughput, low CPU utilization, and direct data placement, RDMA (Remote Direct Memory Access) has been adopted for transport in a number of storage protocols, such as NFS and iSCSI. We have performed a performance evaluation of RDMA-based NFS and iSCSI on Wide-Area Network (WAN). We have shown that these protocols, though benefit from RDMA on Local Area Network (LAN) and on WAN of short distance, are faced with a number of challenges to achieve good performance on long distance WAN. We have revealed that this is because of (a) the low performance of RDMA reads on WAN, (b) the small 4KB chunks used in NFS over RDMA, and (c) the lack of RDMA capability in handling noncontiguous data [YRW+08].

In conjunction with the Argonne Leadership Computing Facility (ALCF) we have recently completed a study of the I/O system of the 557 TFlop IBM Blue Gene/P at Argonne [LCL+09]. In this work we detail the performance of individual components that make up the system and then examine how the performance of various components impacts overall bandwidth for a variety of access patterns, including patterns exhibited by SciDAC and INCITE applications. This is to our knowledge the most comprehensive study of I/O performance at this scale.

We have placed additional emphasis on understanding the behavior of analysis applications at large scale in the last year. Working with T. Peterka, H-W. Shen, and K-L. Ma of the UltraVis Institute, we have

examined the performance of volume rendering of scientific datasets at very large scale [PYR+09, PRS+09, PRY+08]. Along the way we developed a new compositing algorithm for use at large scale on systems with multi-ported networks [PGR+09], beating the traditional binary swap algorithm.

1.2 Parallel I/O Infrastructure Evolution

Multiple parallel file system options are now available, and most HPC systems now include a rudimentary I/O software stack. However, the performance of the I/O stack on many systems is much lower than possible given the hardware available. As HPC systems scale and application complexity increases, extracting the highest possible performance from the I/O hardware is critical to the overall effectiveness of the system. The objective of this work is to improve the state of parallel I/O support for HPC. The Parallel Virtual File System (PVFS) and ROMIO MPI-IO implementations are in wide use and provide key parallel I/O functionality. This work builds on these two components by enhancing them in order to ensure these capabilities continue to be available as systems continue to scale. In addition to improvements to these tools, special attention is paid to Cray systems using the Lustre parallel file system.

Progress to Date

Many advances in the **PVFS parallel file system** [CLR+00] project were facilitated by SDM support. Three PVFS releases were made between 11/2006 and 1/2007, including many bug fixes, Myricom MX and Portals communication drivers, and a new file distribution mechanism. Functionality was also added to PVFS to allow control of layout of files, facilitating research being performed in active storage at PNNL. This functionality was also rolled into a release.

Through collaboration with the Argonne Leadership Computing Facility (ALCF) we have ensured that the I/O system on the Blue Gene/P system will meet performance and reliability goals. This includes aiding in the specification of the storage hardware, porting and deployment of PVFS at large scale, and working with IBM to solve a significant functionality problem in early versions of their MPI-IO software for the system. We implemented a lock-free driver for the Blue Gene that enables PVFS use, improved the scalability of some metadata operations, and integrated IBM's changes back into the ROMIO source tree.

We incorporated successful research efforts into the production **ROMIO MPI-IO library** [TGL99], including Lustre-specific improvements, file domain, and strided I/O optimizations. We have also developed additional test cases to exercise new code paths in the IBM Blue Gene/P MPI-IO implementation resulting from their work to fix limitations due to a 32-bit pointer on the system. We continue to integrate patches and fix bugs in ROMIO as they are uncovered on the IBM Blue Gene/P system.

Parallel NFS (pNFS) is an emergent open standard for parallelizing data transfer over a variety of I/O protocols. Prototypes of pNFS are actively being developed by industry and academia to examine its viability and possible enhancements. We have designed **Lustre-based parallel NFS** (lpNFS) as an enabling technology for transparent pNFS accesses to an opaque Lustre file system. We optimize the data flow paths in lpNFS by using two techniques: (a) fast memory coping of small messages, and (b) page sharing for zero-copy bulk data transfer. Our initial performance evaluation shows that the performance of lpNFS is comparable to that of original Lustre. Our results have shown that lpNFS is a promising approach to combining the benefits of pNFS and Lustre, and it exposes the underlying capabilities of Lustre file systems while transparently supporting pNFS clients [YDV09].

1.3 Application Interfaces and Data Models

In order to make applications more nimble with respect to their I/O behavior, more effort must be spent on the applications and the interfaces that they use to interact with the I/O system. The objective of this work is to improve the usability and observed I/O throughput for applications using parallel I/O by enhancements to or replacements for popular application interfaces to parallel I/O resources. This task was added in response to a perceived need for improved performance at this layer, in part due to our previous work with the FLASH I/O benchmark. Because of their popularity in the scientific community we have focused on the NetCDF and HDF5 interfaces, and in particular on a parallel interface to NetCDF files.

Progress to Date

Significant work has gone into making the **Parallel netCDF (PnetCDF)** [LLC+03] software ready for production. We moved to using SVN and Trac to manage the Parallel netCDF source tree, facilitating greater community involvement. We also improved the package to better operate on Blue Gene/L, Blue Gene/P, and SiCortex systems. Support for increasingly large datasets has become a critical issue for PnetCDF. The original UCAR netCDF format supports variables up to 2 GBytes (due to 32-bit limitations) without special work, but our users are beginning to surpass this limit. Our first pass was to fix internal overflows (inherited from the original serial netCDF code), and now we have developed an extension (that uses 64-bits for sizes), the "CDF-5" format, that allows us to store variables of effectively unlimited size. We are synchronizing these changes with the serial netCDF team so that serial tools can interoperate.

We have been actively supporting new users of PnetCDF, including A. Koontz (PNNL) in her work on global cloud resolving models and J. Edwards (UCAR) and others in their move of the CCSM climate code to use PnetCDF. We have given presentations at UCAR, to the GARPA meeting group, and to CCSM developers visiting Argonne to help these groups better understand both parallel I/O systems in general and the advantages of PnetCDF specifically.

We have completed implementations of **Disk Resident Multidimensional Extendible Array (DRXTA)** functions in C to create, read, write and manipulate out-of-core arrays stored in Unix file systems [OR07]. Support for accessing Parallel Extendible Array files in Global Array applications has been implemented as well, providing an alternative to Disk Resident Arrays. Array chunks are cached in and out of memory using the cache pool implementation from BerkeleyDB., and we have experimented with alternatives to LRU for cache replacement in this system. We have also implemented skip-list access methods for chunked dense arrays, also using the BerkeleyDB cache pool module underneath. This is a more common method for organizing this type of data, and using this implementation we performed a performance comparison of DRXTA storage of dense arrays with skip-list indexing of dense array chunks. Currently we are evaluating the DRXTA approach as an alternative storage organization for use in storing chunked data within the HDF5 library.

As part of our efforts to provide a versatile I/O middleware for scientific applications, we have developed in cooperation with ORNL, Georgia Tech, and others, a framework that isolates application codes from the specifics of the I/O layer, thus permitting various optimization techniques to be applied without changing the application code. By capturing the desired I/O properties (e.g. type of data, parallelism, format) in an XML description outside the application, experimenting with alternatives is made easier. This **Adaptable I/O System (ADIOS)** provides an easy-to-use programming interface, which can be as simple as Fortran file I/O statements. Abstracting the I/O metadata information and data structures from the source code into an external XML file can reduce the code pollution and create the connection between high-level APIs and the underlying I/O implementation details, as well as other technical descriptions, such as buffering and scheduling. By separating the detailed I/O implementation from the APIs, ADIOS also allows users to only change the declaration of the transport methods in the XML file without any source code modification.

We have created matlab and parallel visit ADIOS BP readers. We have added new methods to greatly speed up the writing of large and small files on the Cray XT5 at ORNL. We worked with LBL to get Chombo integrated with ADIOS. Preliminary results show a speedup over the originally optimized hdf5 data output. We have created new APIs to read in ADIOS-bp data. This works for global arrays from n-dimensional domain decompositions, and allows for serial and parallel access to variables in BP files. We have worked with the following codes for ADIOS output: GTC, GTS, XGC-0, XGC-1, GEM, Pixie3D, Gysela5D, and Chombo. We are also actively working with the Cactus team and a few other codes in other areas.

Parallel netCDF (PnetCDF) [LLC+03] software has been augmented with several new features, aiming for leveraging the productivity and performance. We have developed the new "CDF-5" file format to allow a netCDF file to store array variables of size more than 2 billion elements. The new file format and library use 64-bit integer data type to store all the metadata. This eliminates the limitation inherited from the earlier PnetCDF that uses only 32-bit integers to present all integer data. We also added a feature for detecting metadata inconsistency across all application processes to help PnetCDF programmers quickly find inconsistent metadata inputs. Another feature is a new set of APIs that allow accessing multiple netCDF variables in a single function call [GLC+09]. This feature is particular useful for applications accessing a group of small sized variables at the same time frame. With this new functionality, I/O requests can be aggregated and enhance the I/O performance. Based on the same idea, we also revised the PnetCDF asynchronous functions so such aggregation can be achieved transparently from the users [GLC+09]

We have developed a subfiling utility for PnetCDF. As the number of processes running a single application job increases beyond thousands, it is becoming inefficient and ineffective to having all processes accessing to a shared file. The subfiling utility can systematically divide a netCDF file into a group of subfiles to reduce the number of processes sharing a file [GLN+09]. Although the netCDF arrays are physically partitioned on different files, they still appear as single ones to the program users. In addition, arrays are partitioned along the most significant dimension so they are still stored in the canonical order in files. Data extraction and combination from the group of subfiles are carried out by the subfiling utility, and are transparent to users. Compared with the single shared-file approach, the subfiling strategy provides better scalability beyond thousands of processes. Compared with the one-file-per-process I/O approach, the subfiling keeps the number of files generated by large-scale applications in a reasonable and manageable range [GLN+09].

A prototype of data analysis capabilities in PnetCDF has been implemented. We have incorporated a parallel K-means data clustering functionality as a new data analysis API. We are in the process of developing more data analysis utilities to provide users with range query, statistical and other content-based data analysis functions on netCDF datasets.

An application can take advantage of multicore processors only if it is able to exploit its functional parallelism within the limited data parallelism. HPC parallel I/O libraries such as MPI-IO do not yet exploit multi-core and many-core processors for data input/output. It would be desirable if the surpluses in computing power can be transformed into gains in I/O performance, for example by compressing and consolidating data, and reducing the network bandwidth consumption and storage footprint. We have introduced an **opportunistic data compression** scheme at the MPI-IO level such that all scientific applications on top of this popular interface can take advantage of it. Data segments are indexed and organized in a hierarchical data format to allow for the global deduplication. We are evaluating the initial prototype to show its benefits using synthetic I/O datasets at a varying degree of compression ratio.

1.4 Next-Generation I/O Software Technologies

Some challenges for future I/O systems call for the development of altogether new software technologies. One focus for software development is on the creation of a collaborative file caching system for use in

HPC environments. Such a system would take advantage of small portions of memory on a collection of machines to generate a cache of sufficient size to enable aggregation and reorganization of I/O operations from HPC applications. A second focus of our work is in improving the analysis capabilities of HPC systems. A promising technology for improving analysis is **active storage**, which provides the ability to perform data processing on the storage nodes of modern file systems. We will discuss our successes in both of these areas in this section.

I/O forwarding is another technology that we believe will be a critical component of successful I/O systems on future machines, and a simple implementation is already present on the IBM Blue Gene series of machines. In conjunction with the I/O Forwarding Software Layer project, we have incorporated support for the ZoidFS I/O forwarding API and improvements to the MPI-IO implementation into ROMIO.

Progress to Date

An **I/O delegate and caching system** (IODC) has been implemented [NLC08]. It is a software layer in MPI-IO where certain tasks, such as file caching, consistency control, and collective I/O optimization are delegated to a small set of compute nodes, collectively termed as I/O Delegate (IOD) nodes. IODC system is implemented in ROMIO where it intercepts all the system I/O requests initiated by ROMIO and redirects them to IOD nodes. It uses MPI dynamic process management functions to allocate additional MPI processes for the IODC system so that it can run side-by-side with the application processes. In our design, only IOD nodes can access underlying parallel file system. The tasks performed at the IOD nodes include data caching, aggregation for small requests, cache page migration, and I/O access alignment. By exploiting processing computational power and memory space at IOD nodes for performing data caching, and aggregation, we achieve considerably high percentage of I/O bandwidth improvement. We conducted our experiments using the FLASH and S3D I/O kernels on GPFS and Lustre. Testing and development were performed on several parallel machines: Tungsten running Lustre file system and the IBM cluster running GPFS file system at NCSA. Our experiments show that with 3% of compute nodes allocated as I/O delegates, I/O bandwidth improvements range from 40% to 250%. When the number of I/O delegates is 10% of the application nodes, we observed up to 500% improvement [CLG+09, NLC08, IBC+08].

1.5 Outreach

We take outreach very seriously. We have presented 3 tutorials on topics related to storage and parallel I/O in the last year, including two full-day tutorials at the SC conference series. We actively participate in DOE Exascale workshops and other application-oriented meetings to help educate the community on I/O best practices, and we help organize the HEC FSIO meeting, helping guide research into file systems and I/O for high-end computing [GNB+09].

We have worked closely with Garth Gibson of Carnegie Mellon University and the SciDAC Petascale Data Storage Institute to help in using PVFS as the foundation for class projects in parallel file systems. So far PVFS has been used in two courses as the basis for work in distributed directory storage in high-performance file systems and in alternative data organizations on local storage. Both efforts have resulted in student publications [PGL+07, PST+08]. Likewise, we have been working with J. Wang and his students at the University of Central Florida on approaches to data organization under PVFS [GWR08], and we recently hosted an IIT student for the summer and presented a talk on PVFS to students and faculty at IIT.

We have recently led the development of two chapters on parallel I/O for books on data management and analysis as well [RCG+09, RCM09].

2. Scientific Data Mining and Analysis (DMA)

The Data Mining and Analysis (DMA) layer provides the data-understanding technologies necessary for efficient and effective analytics of complex scientific data. This is accomplished through the development and deployment of the three core technologies:

- High performance parallel statistical computing
- Efficient searching and filtering in data-intensive scientific applications
- Feature extraction and tracking for scientific applications

2.1 High performance parallel statistical computing

The research cycle of science applications includes more than just designing an experiment (e.g. simulation model) and collecting data (e.g. simulation output). After, or while, the results are generated, scientists perform data analysis to discover, build, test, or annotate a new view of scientific reality (scientific discovery). These analytical predictions feed back into the design of new experiments.

Fundamental differences in how data analysis is performed exist as data increases in its size and complexity. This places some unique computation, memory, data, and knowledge management requirements on data analysis environments for massive scientific data sets. These requirements motivate our development of the advanced framework that focuses on the three steps of the iterative knowledge discovery cycle:

1. How will an end-user define a data analysis pipeline over massive data sets?
2. How will such analysis pipelines be efficiently executed?
3. How will analytical predictions derived from these analysis pipelines be annotated by the scientific community at large?

Progress to Date

To date, the framework that aims to enrich and optimize the knowledge discovery cycle has three major components, which parallel the three core steps of the knowledge discovery cycle:

Defining Data Analysis Pipelines through Web-Enabled R

The emerging trend is for geographically distributed multidisciplinary research groups collaborate to solve data analysis problems. This creates a crucial need for Web-enabled collaborative data analysis environments capable of managing the increasing scale of data. Most traditional data analysis environments, such as R, are stand-alone applications that run on a researcher's desktop or laptop.

In FY09, a general architecture is designed and a prototype is developed for Web-enabled analytics of large-scale data in a collaborative environment, using R as its back-end analytical engine. A systematic evaluation and comparison with the current Web-enabled R projects demonstrated the superior performance across a multitude of metrics derived from a set of the identified requirements [BSK09].

Executing Data Analysis Pipelines with pR

Once the data analysis pipeline is defined, the next problem for data intensive statistical computing is how to execute these analysis scripts in the most efficient manner. The difficulty is that many of the current data analysis routines are written in non-parallel interpretable scripting languages such as R, IDL, or MATLAB and are not scalable to massive data sets. The emerging approaches, such as those being developed by our team, aim to provide parallel solutions. In the case of R, several of our software packages including our RScaLAPACK and pR are maturing [PKW+09.a, SPK+09, SAC+07, SBG+06, SCR+06, POS07].

While promising, they assume that parallel implementations of data analysis routines either exist or are possible. However, there is a growing gap between the rate at which data analysis functions are developed and the rate at which they are parallelized. Moreover, the majority of domain scientists that develop domain-specific analysis routines are typically not well versed in parallel computing and write serial data processing and analysis routines in compiled languages such as C/C++/Fortran. By doing so, they face a dilemma: on the one hand, they achieve improved performance by choosing the path of custom built compiled analysis codes. On the other hand, they lose access to the rich set of statistical analysis routines provided by scripting environments such as R or IDL.

Although scripting languages like R provide `hooks' for calling external compiled functions in C/C++/Fortran, the burden on the user is quite high in terms of familiarity with R internals, which limits the effectiveness of these hooks. In addition, it is often the case that the overhead introduced by inter-language translation mechanisms is quite high and, in some cases, dominates the computational cost of executing analysis routines in a compiled language. As a result, these domain-specific analysis routines, while efficient and highly valuable for the community at large, rarely become an integral part of statistically rich and community-shared environments such as R.

In FY09, we further extended pR, a light-weight, easy-to-use plug-in that bridges compiled serial and parallel analysis routines into the R scripting environment for efficient execution of data analysis tasks. It provides a C/C++/Fortran open-source API that exposes third-party serial or parallel codes within the R statistical package, provides automatic parallelization of certain data-parallel operations, and minimizes R overhead. Users can write analysis routines in any supported language and call them from R, therefore bridging the costly divide between current robust analysis tools. The ways to convert data types between R and a compiled languages (e.g., C/C++) are explored, and a templated approach that handles casting between systems is designed and implemented. To the best of our knowledge, pR is the first such system for R. pR often achieves superior performance speedups through parallelization and minimizing overhead compared to existing approaches.

Annotating and Curating Knowledge via BioDEAL

In many scientific domains, especially those of Biomedicine and Earth Science, the predictions derived from analytical pipelines are recorded in distributed public databases that are shared by their respective communities. It is often the case that these predictions (e.g., protein functions or anomalies in carbon pools) are imprecise or incomplete due to the noisy and uncertain nature of the data from which they are derived, as well as the incomplete physical models used to generate such data. As part of an iterative discovery cycle, these predictions may get tested, validated or improved (e.g., identified causes for anomalies or experimentally confirmed protein functions). The results of this downstream research process often get published but rarely are populated into the public databases as supporting evidence. There is a growing need for streamlining such downstream discovered evidence into the public databases to improve their analytical predictions. Currently, no capability of this sort exists.

In FY09, we designed and prototyped BioDEAL, a community **Bio**logical **D**ata-**E**vidence-**A**nnotation **L**inkage system that can introduce a feedback loop into the database-publication cycle to allow biologists

to make connections between data-driven biological concepts and publications, and vice versa. By subscribing to the services provided by BioDEAL, an end-user, at a minimum, can annotate the facts reported in literature, associate these facts with ontological concepts, and share these literature annotations with other researchers in a social network. For example, while reading the paper by Lovley et. al, a genome annotation expert may decide to link the omcB (GSU2887) gene with electron carrier activity (GO:0009055) in the Gene Ontology (GO) and add a comment on experimental validation of its predicted function as the Fe(III)-reductase. BioDEAL will record these annotations in a structured (XML) format so that other databases, such as GenBank, may parse this information and potentially update its ``Related Articles in PubMed'' field for this gene's web page with this PubMed ID. Although a number of databases and frameworks can benefit from and/or enhance the functionality of BioDEAL to the best of our knowledge, BioDEAL is the first system that enables such a feedback loop into the database-publication cycle. Although the proposed framework is tailored to the Biomedical community, its underlying architecture is quite generic in nature and is easily adaptable to other domains.

Future Plans

- Extend the capability of RScaLAPACK: support for openMPI back-end in response to multiple users' requests, ease the installation via improved autoconf, provide processor grid manipulation routines, provide both static and dynamic MPI library support.
- Extend pR to provide automatic parallelization for certain classes of analysis functions, such as those based on apply() family in R. Release the library.
- Start exploring the support of analytics functions inside of the ADIOS I/O library in collaboration with Scott Klasky.
- Explore the ways to incorporate fusion science analytical pipelines such as front tracking (not described here) into the Dahsboard.
- Build and release a library of parallel graph analysis routines using pR, pRapply, RScaLAPACK frameworks.

2.2. Efficient searching and filtering in data-intensive applications

In many scientific applications, to gain insight from massive amounts of data, the scientists have to locate a relative small number of data records that hold the key information. For example, in a study of turbulent combustions, these special data records might be called ignition kernels. In a study of laser-wakefield particle accelerators, these special data records might be called particle bunches. Sifting through mountains of data to locate these "interesting" data records is a significant challenge. To meet this challenge, we have been developing and expanding an indexing software package called FastBit.
FastBit is based on a database indexing technique called bitmap index. This type of index is well-suited for searching scientific data where the data records usually remain unchanged after they are created. Taking advantage of this "read-only" (or "read-mostly") nature, FastBit indexes are designed to answer queries fast by sacrifice some efficiency in updating the indexes. This allows us to package the indexing data structures tightly, reduce the I/O requirement when answering a query and improve query response time. Additionally, FastBit is designed to work with user data in their existing formats, instead of demanding the user data to be transform into a particular format or loaded into a database management system. This flexibility makes it possible for users to accelerate their search operations with a minimal amount of change to their existing data analysis framework.

Progress to Date

In FY09, we have extended the FastBit software in a number of different fronts. We have significantly extended the histograming feature in support of the analysis of simulation of Laser Wakefield Particle Accelerators (LWPA), implemented new algorithms for devising multi-level indexes, and added support

for arithmetic expressions to compute values in the output of the query results. Additionally, we have added support for sorted data and binary objects (more commonly known as BLOBs). In addition, we are also working with application scientists to extend FastBit software to work on queries that require addition information outside of the data table. This particular task is generally known as finding regions of interest can be implemented as database joins, however, by taking advantage of the structure present in the underlying data, we are able to find regions of interest nearly as fast as finding the data records in the regions of interest. In other word, we can perform this special type of database join faster than the filtering step, where the typical database join algorithm would take much longer than the filtering step. Instead of providing technical details on these topics, we next describe two applications and explain how FastBit improves these applications.

Analyzing Laser Wakefield Particle Accelerator Data. Laser Wakefield Particle Accelerator (LWPA) has the potential of replacing current generation of kilometer sized particle accelerators with new ones that can fit on table tops. One crucial piece of the puzzle for this technology breakthrough is the precise control of the plasma density. To understand this process, physicists have developed extensive simulation capabilities. In analyzing the simulation output, the physicists need to be able to select the appropriate particles and track those particles through time. FastBit is able to provide several orders of magnitude reduction in the turn-around time of this analysis process.

To help the users select the "interesting" particles, our visualization collaborators have settled on a parallel coordinate display. This special parallel coordinate display uses a series of two-dimensional histograms instead of accessing the actual raw data values. In this process, we provide a set of efficient conditional histogram functions to reduce the time needed to generate the parallel coordinate display. By using FastBit indexes, we are able to identify the records needed for the histogram computation quickly and efficiently. This significantly reduces the amount of time used to read the data values.
The parallel coordinate display helps the user to decide what particles might be of interest. This "interest" is specified as a set of ranges on various variables. FastBit selects the particles satisfying these range conditions and outputs the particle identifiers to be used in tracking these particles through time. This selection process is part of the core functionality of FastBit and can be carried out very efficiently with the indexes.

The most time consuming part of the data analysis is the step of tracking particles through time. Using FastBit, we are able to conduct the tracking operations as a set of searches on particle identifiers. Because a long list of particle identifiers is involved, we have developed mechanism to bypass certain common operations such as feeding the selection criteria as strings. Because the particle identifiers are often in sorted order, we have also developed functions to take advantage of the ordered data. In time measurements, we have observed that using FastBit can be orders of magnitude faster than not using any indexes.

Finding Regions of Interest. Commonly, the records in a data table are considered as independent of each other. This is true in the case of particles in a LWPA simulation, but for image data and data with an underlying mesh, the connectivity among the data records is crucial to most data analysis tasks. Conceptually, the connectivity can be stored in a separate data table and the connectivity information can be reintegrated into the records through database joins. However, database joins are typically very expensive and very slow. We would like to avoid it by taking advantage of the structure present in the connectivity. For example, we have demonstrated in the past that by extending bitmap indexes we can find regions of interest on data from regular meshes in less time than the algorithms that only find the boundaries of these regions of interest. In the past year, we have been exploring the possibility of extending that work to data from irregular meshes. Our first attempt is to work with data from a toroidal mesh used for fusion simulations.

To efficiently identify regions of interest on toroidal meshes, we recognize that the data from the simulation program is order in a way that is more "natural" in the magnetic coordinate system than in the Cartesian coordinate system in the real space. More specifically, the computations are performed mostly in the magnetic coordinates and the mesh points are ordered in a more "regular" way in the magnetic coordinates. By taking advantage of the structure in the magnetic coordinates, we are able to provide a much more efficient connectivity definition and more compact data structure for storing the points in the regions of interest. In addition, we have developed an implicit union-find data structure that can record connectivity information among the points efficiently. All these together allow us to find regions of interests hundreds of times faster than the most competitive methods in literature.

Supporting FastBit users. Since the release of the FastBit software in 2007, we have spent considerable amount of time interacting with a growing user community. In the past year, we have seen a number of publications that make extensive use of FastBit. Here we briefly mention the key results from two such publications. In a paper by Jochen Schlosser and Matthias Rarey from University of Hamburg in the Journal of Chemical Information and Modeling, the authors demonstrated that FastBit can be used efficiently to support molecular docking. Their tests show that the new molecular docking software is 100-200 times faster than the existing software packages. Thiago Luís Lopes Siqueira and colleagues from Brazil have published a number of articles on using FastBit in geographical data warehouses. They have demonstrated that the new system with FastBit can outperform the existing systems by a factor of 10 to 20 times. We are very much encouraged by such creative and successful use cases, and would endeavor to continue our support of such users.

Plans for the Future

- Transfer the region finding work into the user's hand. This would require us to update the software to better interact with the evolving output data files used by the simulation code, and to provide a reasonable user interface to their data analysis framework.
- Work with collaborators in visualization to expand the capability of the parallel histogram based data analysis system. The current set of functions are based on HDF5, most of them can potentially be used for other applications that output HDF5 files as well.
- Continue to support our users and expand the functionality of FastBit to meet their needs.

2.3 Feature extraction and tracking for scientific applications

As the data from scientific simulations, observations, and experiments approach the petascale and beyond, we need to extract features to fully realize the benefits of our advanced computational and data collecting abilities. This area focuses on the development and application of analysis techniques to data from scientific simulations, observations, and experiments. We use techniques from several disciplines, including image and video processing, machine learning, statistics, and pattern recognition, to find useful information in massive, complex data sets [Kam06, Kam08b]. Our goal is two-fold – to use data mining techniques to understand scientific phenomena and, as appropriate, to deploy our solution for use by application scientists. We have worked with a number of application projects, initiated by the SDM center or at the request of the domain scientists.

Progress to Date

The problems we focused on were driven by applications scientists. Each problem presented different challenges, and required different techniques. The challenge was not only to discover the combination of techniques that addressed the problem at hand, but also to discover new approaches for previously unsolved problems. This could only be achieved by working closely with the application scientists, understanding their problems, providing solutions, and iterating the process. We had great success in addressing several problem classes as described below.

Poincaré plots are an important tool for understanding data in Fusion science applications. A Poincaré plot is composed of orbits, each of which consists of several points created when a field line, tracked around a toroid, intersects a poloidal plane. The shape of the orbit depends on the starting point of the field line. Our task is to assign an orbit to one of four classes – island chain, quasiperiodic, separatrix, and stochastic. This is currently done visually, a process which is tedious, error-prone, and subjective. Our early work, based on fitting second order polynomials to the points, appeared promising, but was sensitive to the choice of parameters, did not easily extend to stochastic orbits, and did not lend itself to a simple extraction of rules for classification. Another approach, called KAM, proposed in the context of dynamical systems was suggested by our collaborators, Neil Pomphrey and Don Monticello (PPPL). This used graph-based features to represent the orbit, which was classified using heuristic rules. Our experiences with KAM indicated that it was not suitable for the characteristics of our data from simulations, which were not only noisy, but also had very thin lobes in the separatrix and island orbits. However, these experiences indicated a solution more suitable for our data. Through extensive experimentation, we **developed a system for classification of Poincaré plots**. This approach extracted robust features that were scale, rotation, and translation invariant and then used these features in a decision tree classifier. We addressed several major challenges, such as: i) improving the quality of the training data by varying the class labels with the number of points; ii) applying techniques from spatial statistics to identify locally stochastic orbits; iii) incorporating appropriate scaling to handle orbits with thin lobes; iv) exploiting the alignment of peaks and valleys to capture local variation along the orbit; and v) using wavelet analysis to represent the multi-scale structure of orbits. After several iterations we obtained an error rate of ~4% using one of our patented algorithms for ensembles of decision trees. To the best of our knowledge, this is the first time an accurate, automated solution has been developed for this problem. Josh Breslau (PPPL) found the code worked quite well and is distributing it to M3D users. As the problem is of broad interest in the fusion community, it is also being deployed for use by others through the workflows being developed by the SPA team. By replacing a tedious and error-prone visual classification, this code is enabling fusion scientists to evaluate their simulations quickly. A journal paper summarizing the approach is in progress.

A second project involves **blob tracking in experimental images**. We are working with Fusion physicists Stewart Zweben and Ricardo Maqueda at PPPL to analyze high-speed, high-resolution images of plasma from NSTX to understand edge turbulence. This involves the segmentation, characterization, and tracking of coherent structures, known as blobs, in the image sequences. There are several challenges: the images are noisy; the theory behind edge turbulence is poorly understood and cannot influence what the scientists expect to see in the data; and the images in a sequence are varied, with both bright and faint blobs, as well as bright blobs with extended faint tails. It is non-trivial to come up with an algorithm which, with a single set of parameters, will perform well across all images in a sequence. In earlier work, using sample images, we investigated several algorithms to de-noise the data, remove the ambient intensity, and identify the blobs [LK07]. In FY09, we considered the more promising of these algorithms and identified those which could handle the variation across sequences of images, working correctly when a sequence had images which were either mainly noise, or composed of faint blobs, or composed of both bright and faint blobs. We also addressed the issue of visual display of these images and created scripts which enabled processing a sequence of images.

A third project is a collaboration with the GSEP SciDAC (Zhihong Lin, PI). The analysis goal is to **identify coherent structures in GSEP simulation fluid and particle data and to understand the non-linear interactions** between the two. This is difficult as: (i) there is no definition of coherent structures; (ii) the fluid data are on a twisted toroidal mesh while the particle data are unstructured, making existing algorithms inapplicable; and (iii) the data are currently in terabytes, with petabytes expected in the future. In FY09, we obtained new data with additional variables for each grid point, used exploratory analysis to identify correlations between the variables, and implemented an initial algorithm to identify the coherent structures in the fluid data in a poloidal plane. These preliminary results are being discussed with Zhihong Lin and Yong Xiao to determine if the approach should be applied to additional time steps.

A fourth project, started in early 2009, is a collaboration with the WindSENSE project funded through EERE. This resulted from the CRNARE workshop organized by DOE in 2007 to identify opportunities in renewable energy, with a focus on EERE missions and Office of Science capabilities. Since renewables, such as wind, tend to be intermittent, it is difficult to schedule them on the power grid while maintaining its reliability. Ramp events, where the wind power increases or decreases by a large amount in a short time, are becoming a major problem as wind resources form an increasing percentage of the total generation. We are using **feature selection to identify weather conditions associated with wind ramp events.** This work is being done using data from Southern California Edison (SCE) and Bonneville Power Administration (BPA). In FY09, we had extensive discussions with SCE, CaISO, and BPA forecasters and schedulers, identified different ways of defining ramp events, and obtained statistics on 2007-2008 data from BPA. Our early results show that these statistics can provide insights into the generation from the wind farms in the Columbia Basin. For example, we found that negative ramp events rarely occur in the morning, while positive ramp events occur more frequently in the afternoon. Further, ramp events tend to occur more frequently in some months than others.

Plans for the Future

In the near future, we will:

- Complete the journal paper on classification of orbits; identify and extract "X points" in the separatrix orbits and deploy the code for use by fusion scientists.
- Complete the identification of the blobs in NSTX images for the shorter sequences, extract relevant features, track the blobs by exploiting the overlapping between frames [GK08], and extend the approach to the longer sequences of 7000-8000 frames.
- Investigate further the current approach to segmentation of coherent structures in fluid data from GSEP, extend it to work on all time steps, and extract characteristics of the structures.
- Leverage our SciDAC 1 work in edge-harmonic oscillations to identify weather conditions associated with wind ramp events.

This work will be done in close collaboration with the domain scientists. In addition, we will continue to explore opportunities for data analysis in additional Office of Science applications.

2.4 Training and Outreach

Our training and outreach span a wide spectrum of activities. We presented our research findings at various national and international conferences, including invited talks at Supercomputing 2008 conference [SHB+09]. We summarized some of our findings as book chapters in the "Scientific Data Management" book, edited by A. Shoshani and D. Rotem [SR09, KWK+09, OW09]. We organize international conferences, such as the SIAM Data Mining conference, which, along with our editorial responsibilities [GKK08, Kam08a], allow us to influence the broader scientific and technical communities in our areas of expertise.

We actively participate in a series of DOE Exascale workshops [HZS07], including a recent 2009-series such as ``Data Analysis, Management and Visualization in Fusion Energy Sciences at Extreme Scale." We co-organized the DOE/NSF workshop on "Mathematics for Analysis of Petascale Data" [KCC+08], the DOE OBER/OASCR workshop on "Genomics GTL Knowledgebase" [GFS09]. In the training arena, we contributed data analysis modules to the Supercomputing 2007 tutorial.

Some of our technologies, such as pR, are being taught as part of the undergraduate- and graduate-level curriculum on Automated Learning at NCSU, Computer Science Department; with the parallel data

mining codes developed by the students using pR through their course projects. FastBit is attracting a growing user community with an active discussion mailing list and a number of enthusiastic contributors from around the world. Many research efforts presented in this report are the result of the four PhD and one MS students' theses. Our work on FastBit, for example, is sparking various research efforts around the country. During 2007, right after the public release of FastBit, it contributed significantly to two PhD theses from UC Berkeley and UIUC [SMW+08, RSW+07]. We are also aware of research efforts deriving from FastBit technologies or utilizing the software from other universities and private companies.

We work closely with a number of DOE projects that has resulted in a number of joint publications or software usages including: (a) SciDAC Ultra-scale Visualization Institute (PI: K-L Ma) [SJH+07, SBH+08]; (b) SciDAC VACET Center (PIs: C. Johnson and W. Bethel) [OPW+08]; (c) SciDAC CPES Center (PI: C.S. Chang) [CKD+09]; (d) DOE Bioenergy Centers and GTL projects [RPH+09, POL+08, KPP+09, YHL+09, BPV+0, BGB+09]; (e) SciDAC CEMM (PI: S. Jardin).

3. Scientific Process Automation (SPA)

Effectively generating, managing, and analyzing scientific data requires a comprehensive, end-to-end approach that encompasses all stages from the initial data acquisition to the final analysis of the data. As part of the SPA thrust area, we are developing a suite of tools and frameworks that integrate into a robust and auditable system for automation of scientific processes to enhance and speed up scientific discovery [CRI09]. Our technologies provide run-time management of the workflow processes, provenance collection, and analysis and display of results. This has led to the deployment of production workflows that allow scientists to a) monitor, in near-real-time, complex tasks such as the execution of large simulation codes, and b) facilitate complex analyses of the process metadata and of the simulation results. This has resulted in significant savings in scientists' time, in more efficient use of resources, and in a more cost-effective scientific discovery process overall.

Workflow technologies have a long history in the database and information systems communities [GHS95]. Similarly, the scientific community has developed a number of problem-solving environments, most of them as integrated solutions [HRG+00]. Component-based solution support systems are also proliferating [CL02, CCA06]. Scientific workflows merge advances in all these areas to automate support for sophisticated scientific problem-solving [LAB+06, LG05, DOE04, ABB+03, BVP00, VS97, SV96]. We use the term scientific workflow as a blanket term describing a series of structured activities and computations (called workflow components or actors) that arise in scientific problem-solving as part of the discovery process. This description includes the actions performed (by actors), the decisions made (control-flow), and the underlying coordination, such as data transfers (dataflow) and scheduling, required to execute the workflow. In its simplest case, a workflow is a linear sequence of tasks, each one implemented by an actor. An example of a scientific workflow is: transfer a configuration file to a large cluster, run a simulation passing this file as an input parameter, transfer the results of the simulation to a secondary system (e.g. a smaller cluster), select a known variable, and generate a movie showing how this variable evolves over time. Scientific workflows can exhibit and exploit data-, task-, and pipeline-parallelism. In science and engineering, process tasks and computations often are large-scale, complex, and structured with intricate dependencies [DOE04, DBN+96, EBV95, Elm66].

We have not only developed a considerable body of software, but we have transferred our technology to a number of ongoing science projects, published numerous papers, and conducted several tutorials and workshops. The challenge is to provide adequate tools and support for four categories of user levels:

- Level 1: Scientist (uses workflows, uses parameterized templates and web-based **wizards** to adapt existing workflows for own use)

- Level 2: Advanced Users (writes more complex new workflows using existing tools, including Kepler-level graphical user interface)
- Level 3: Workflow, template, and actor developer, very advanced workflows
- Level 4: Workflow framework and engine developer

Level 3 and 4 tools and environments are well understood and we have a comprehensive suite of tool sand educational materials for those two levels. We have completed studying level 1 and 2 interface and interaction needs, and we are in the process of developing a comprehensive suite of wizards and educational tools for those to level. In the past year, our contributions to advancing the state-of-the-art in scientific workflows have focused on the following areas. Progress in each of these areas is described in subsequent sections.

- **Workflow development.** The development of a deeper understanding of scientific workflows "in the wild" and of the requirements for support tools that allow easy construction of complex scientific workflows;
- **Generic workflow components, patterns and templates.** The development of generic actors, patterns and templates (i.e. workflow components and processes) which can be broadly applied to scientific problems;
- **Dashboard development.** The development of a one-stop-shopping workflow monitoring and analytics dashboard;
- **Provenance collection and analysis**. The design of a flexible provenance collection and analysis infrastructure within the workflow environment; and
- **Workflow reliability and fault tolerance.** The improvement of the reliability and fault-tolerance of workflow environments.

3.1 Workflow development

The original base-line contribution of the SPA team has been to co-found the Kepler project – an open source workflow support environment [http://www.kepler-project.org]. Kepler is now a widely accepted scientific workflow development and execution environment that powers a number of research and production projects all over the world. The SPA researchers and engineers continue to regularly contribute to Kepler. We are constantly working with the Kepler Core team [e.g., AJB+04, LAB+06, GBA+07] to enhance Kepler at all levels including the user interface, documentation, and tutorials. Our work has led to a significant reduction in the effort required to generate real workflows [ABC+06, SAC+07]. We also actively partner with science teams to transfer technology to their projects and develop and deploy their workflows. This provides real-life case-studies which are then used to enhance Kepler requirements, to identify Kepler enhancements, generic functionalities, and canonical generic workflow solutions, and to improve user interfaces.

In general we distinguish workflow for three stages of scientific computational processes: a) preparation workflows – those used to prepare data and environments for simulations that will run on supercomputers; b) run-time simulation workflows – those that manage launching of the jobs and run-time monitoring, data collection, and steering, and c) post-processing workflows – those that facilitate output management, viewing, post-run analytics and knowledge creation, archiving, and other post-processing activities. Of course, we also have end-to-end workflows, those that encompass preparation, launching, monitoring and post-processing.

Progress to Date

As active participants, and founding members, in the Kepler research community we **contribute to the development of the Kepler environment**. Specifically, we participate in Kepler and Ptolemy workshops that have produced significant enhancements in the underlying workflow environment, and we have

identified and implemented new requirements for scientific workflows, fixed bugs, and improved the overall software development environment, as well as execution-time interfaces and data collection practices [e.g., BMR+08, CA08, NV08, SAC+07, VAB+07]. Some of the specific tools and additions to Kepler are discussed in sections that follow (e.g., provenance recorder, generic actors, ADIOS).

S3D "preparation" Workflows

When executing large-scale simulations on expensive supercomputers, it is critical that simulation runs be "prepared" and do not run into avoidable problems, thus wasting costly compute cycles. For example, simulation code developers often check in revisions to their Fortran codes, that may e.g. "break the build" (fail to compile), create runtime exceptions, or produce unexpected results, possibly caused by bugs or other undesired effects of the new code revisions. Currently, scientists execute "test runs" in a time consuming, error-prone process to avoid an even costlier detection of these failures when running the actual, large-scale simulation. A typical manual process requires the scientist to perform a number of steps (often repeatedly), for example:

(a) checking out of the current version of the simulation code (e.g. from a SVN repository);
(b) building, i.e., compiling the code (with "make" or similar);
(c) running the executables on various test datasets (requires job submission on a remote cluster);
(d) inspecting the test results to look for any "abnormal" results.

This process also requires access to different remote machines (e.g., to access the code repository, build the code, submit jobs, monitor job execution, retrieve or inspect results, etc).

To facilitate these laborious steps, we have developed several different "preparation workflows" for S3D code. The purpose of these workflows is to prevent costly large-scale runs of error-prone S3D code on ORNL computers. Combustion scientists at SANDIA have used the preparation workflows. We have had positive feedbacks for the use of these workflows. The first version of the preparation workflow includes different nested fragments that implements the different parts of the S3D process, (i.e. (a), (b), (c) and (d) as above). Combustion scientists have tested this version. We have developed an extended second version based on the first deliverable and the feedback from scientists. The extended version links the results to Dashboard, making their review easier for the scientists. In addition to this feature, we have added archiving and fault-tolerance capabilities specific to S3D process. The current workflow includes several hundred actors, dependencies and challenging design choices. We took advantage from this experience in developing the scientific workflow patterns that are discussed in following sections.

3.2 Generic workflow components and templates

Many workflows contain sections with very similar functionality but subtle differences in how that functionality is obtained. For example, transferring a file between machines, submitting a job to a batch processing system, monitoring the execution of a running job, and remotely executing a command are found in almost all of the scientific workflows we have developed. However, the actual implementation of these capabilities varies dramatically depending on features such as the specific machine configurations (e.g., which batch processor is used), the security requirements (e.g. ssh or rsh, certificates or one-time-passwords), and the workflow requirements (e.g. failover options, fault tolerance requirements, validation options). Tools that provide the instantiation of a case-specific workflow or workflow component from more general templates or components would substantially improve the ability to reuse existing solutions, leading to greater productivity when developing workflows. The SPA team has had prior experience developing patterns, templates and generic processes [e.g., DKV97, YGN09], and we believe that it is possible to develop a basic set of patterns and templates for construction of certain types of scientific workflows to make the workflow construction easier and less time consuming through elimination of unnecessary rework that occurs when a workflow solution is developed from scratch.

Progress to Date

Despite KEPLER being employed for the implementation of most SDM processes, the design and development of scientific workflow models are still not well understood by scientists and workflow developers in general. Skeptical arguments of process designers take on, among others, the understanding of dataflow-based programming, the modeling of data and control dependencies using the same constructs and the lack of overall design paradigms. Current KEPLER and scientific workflow literature do not support these issues well; specifically, there are very few works that help beginner process designers, in the earlier stages of modeling, to elaborate their models. Workflow designers have to follow intuitive time-consuming and error-prone techniques to develop their models. Typically, they browse and test existing implementations one after another until they are satisfied they have the corresponding fragments that can be modified for the target process or, more likely, they end up writing external scripts for the parts that they cannot model using the available primitives. Pattern oriented model development is a well-known and widely accepted initiative in workflow and traditional application development [e.g., GHJ+95]. Basically, a pattern is the abstraction from a concrete form, which keeps recurring in specific non-arbitrary contexts. Workflow patterns help to study "the suitability of a particular process language or workflow system for a particular project, assessing relative strengths and weaknesses of various approaches to process specification, implementing certain requirements in a particular process-aware information system, and as a basis for language and tool development".

We have developed an initial set of 15 understandable patterns and different versions for each pattern from the previous SDM workflows in order to capitalize the costly development experiences of previous years [YGN09]. Currently, we work on an application platform that will allow non-specialist scientists to create their workflows using those patterns without dealing with low-level implementation details of KEPLER.

We have also continued to develop and deploy generic actors, which provide general capabilities across underlying representations. These generic actors allow the development and deployment of workflows that effectively run on a variety of platforms – dramatically simplifying the use of workflows to support science. We have developed and deployed a generic remote execution actor that can effectively execute a command on a remote machine, and a job submission actor that is capable of submitting a batch request to the job schedulers on the DOE LCF machines (including LoadLeveler and Condor). We are currently working on integrating grid certificate capabilities into these actors.

3.3 Dashboard development

The emergence of leadership class computing is creating a tsunami of data from petascale simulations. Results are typically analyzed by dozens of scientists. In order for the scientist to digest the vast amount of data being produced from the simulations and auxiliary programs, it is critical to automate the effort to manage, analyze, visualize, and share this data. One aspect of automation is to provide an easy-to-use web-based mechanism to monitor the progress of simulations, and view and compare the results generated with the use of the workflow system. A second aspect is to leverage the collective knowledge and experiences of the scientists collaborating on a project through a scientific social network. This can be achieved through a combination of parallel back-end services, provenance capturing and analysis, and an easy-to-use front-end tool. The SDM Center Dashboard is one such tool [e.g., CRI09, BCK+07, BKM09].

Progress to Date

Development of the eSimMon dashboard is ongoing in cooperation with the group at ORNL [e.g., BCK+07, BKM09]. The dashboard machine monitoring component, which displays supercomputer status and job scheduler information, was initially migrated from AJAX to an Adobe Flex prototype implementation with a large amount of duplicate, complex and redundant code. For example, about ten

new functions were needed to add a new machine to the dashboard. Dashboard components have been refactored with the emphasis on improving stability, usability and developer productivity. The component implementation has now been streamlined using object-oriented software design practices so that **one line of code** is all that is needed to add a new machine. Future work on the dashboard will include adding parameterized configuration, as well as web services to retrieve and display useful information, such as machine status.

To address demand for more interactive visualization tools on the dashboard, the prototype of a 3D visualization application was recently developed in collaboration with Valerio Pascucci's VACET group for use with the eSimMon dashboard. The application displays a set of isosurfaces and provides a contour tree, which captures the topological structure of a scalar field and can be used as an efficient indexing key to quickly access and query features in the displayed isosurfaces. The user interface was implemented in Adobe Flex and the visualization images were compiled into interactive Flash videos. The S3D combustion simulation was used as a case study and the prototype was deployed on the eSimMon dashboard. Future work on this application includes expanding S3D support and adding support for XGC simulations, integration into Kepler workflows, allowing users to access high resolution visualization images from the dashboard, and improving the usability and responsiveness of the interface.

The ORNL eSimMon dashboard is relatively tightly coupled to the ORNL environment. It can only be used there (through the network), and a number of items in it are "hardwired" to that environment. To enable portability and transfer of the dashboard technology to other DOE users, NC State has developed two versions of "portable" dashboards. One is a portable ORNL dashboard that can be taken to places where there is no network – for example, airplanes, by installing it on a laptop and by downloading some of the relevant data to the local database [MOU08]. The other is a "distributable" dashboard – a one-click installation package that can be installed in any LAMP environment and with appropriate configuration used at sites away from ORNL and even with workflow engines other than Kepler [NAG09].

3.4 Provenance collection and analysis

In scientific applications, effectively managing data provenance is extremely important [e.g. CFS+05] Provenance is at the heart of almost all functionalities, capabilities, and productivity improvements offered by workflow solutions. We distinguish system, workflow, process and data provenance cagtegories. In the first category we collect data about meta-data related to preparation, run-time and post-processing environments – things like which compiler was used to make the run-time simulation code, possibly pre-processing testing activities, run-time machine characteristics, etc. Workflow provenances is concerned with workflow preparation, history, templating, and so on, while process provenance offers information about the order and success of the processes executed on both workflow control plane – Kepler, and at sites where remote resource are being used. Data provenances is often considered most important, in our case a lot of it comes from the Kepler provenance recorder, but other sources are used as well. In the case of Kepler, data provenance can be thought of as the complete processing history (transformations, types, etc.) of a data product, for example, actor identification and invocation parameters (or application codes launched by those actors), properties such as the time, location, and userid of invocation, relevant environment and configuration parameters. This information needs to be persistent and permanently associated with a data product so that its provenance is readily available. It also needs to be searchable, so that data with certain provenance can be easily identified – for example, if a bug is identified and corrected, the provenance can help identify which runs should be repeated., or at run-time or post-processing phases, we need quick access to certain information.

Progress to Date

During the last year, we have worked on improving provenance collection and management functionalities of the provenance system we have developed [e.g., BCK+07]. For example, we have

investigated the issue of security and privacy of provenance data sharing [NM08[, and we have been investigating integration of provenance information that may come from different sources [CFS+08]. By consolidating provenance information or a variety of applications, we can provide a uniform environment for querying, sharing, and re-using provenance in large-scale, collaborative settings. In this context, an important source of provenance information are different application, system and other logs. For example, logs can be used to understand issues that may arise, collect information that otherwise would not be available, and in general enrich the meta-data about all provenance aspects. One issue with that is the efficiency of extraction of that information from logs which can have millions of records. We developed a suite of very fast algorithms based on prefix arrays [NVW+08]. In order to transfer our technology to DOE research groups and practitioners, we have also developed two variants of portable dashboards for viewing and using provenance information [NAG09]. This is further discussed in the dashboard section. In addition, we worked on algorithms for fast information retrieval from provenance streams [MBK+09], and we have, working on model-oriented data retrieval [ABM+09].

One of the challenges is fast retrieval of detailed information about files that can be produced in very large numbers during different workflow stages, such as visualization related files. The Kepler provenance framework collects all or part of the raw information flowing through the workflow graph. This information then needs to be further parsed to extract meta-data of interest. This can be done through add-on tools and algorithms. We have developed a suite of algorithms that enable dashboard users to quickly find locations of very specific files (e.g., the history of a image, including its location) [MBK+09]. These algorithms leverage Kepler provenance recorder streams and through backtracking identify required information.

Another open issue is that existing approaches for representing the provenance of scientific workflow runs largely ignore computation models that work over structured data, including XML. Unlike models based on transformation semantics, these computation models often employ update semantics, in which only a portion of an incoming XML stream is modified by each workflow step. Applying conventional provenance approaches to such models results in provenance information that is either too coarse (e.g., stating that one version of an XML document depends entirely on a prior version) or potentially incorrect (e.g., stating that each element of an XML document depends on every element in a prior version). We developed a generic provenance model that naturally represents workflow runs involving processes that work over nested data collections and that employ update semantics. We showed how hybrid queries can be expressed against our model using high-level query constructs and implemented efficiently over relational provenance storage schemes [ABM+09].

Provenance also plays an important role in implementation of fault-tolerance strategies in our workflows, as well as in development of design and implementation patterns and templates. This is discussed in following sections.

3.5 Workflow reliability and fault tolerance

One of problems associated with scientific workflows is the fault-tolerant workflow modeling. There are two basic forms of run-time fault-tolerance: forward-recovery (e.g. failure masking and redundancy based failover), and backward-recovery (e.g. check-pointing) [LB90, MV06, DKV97, VOU05, VAB+07]. Exception handling is a very traditional way of managing run-time problems [MV96, VOU05]. It is also used in the workflow-oriented environments [HA96, CCPP99]. Exception handling can involve forward-recovery, backward-recovery, or graceful termination. More recently the web services community has recognized the need for some form of standardized fault-tolerance in the service provisioning through replication [SPP+06]. An important component is collection of sufficient amount of meta-data (provenance information) about the workflow, processes, data, and environments to enable fault-tolerance actions.

At this stage of our work, it is the provenance information that is being collected through our provenance recorder that is at the heart of the fault and failure management mechanisms we are implementing. It has the capability of providing meta-data needed to detect and locate workflow run-time issues that can be handled at the Kepler control layer.

Progress to Date

We have implemented a Kepler-based fault-tolerance framework that leverages provenance information it collects, and provides options for forward recovery as well as backward recovery of the workflows at control plane [MV06, VAB+07, CA08]. We are in the process of extending the model beyond the workflow control plane, so that we can catch and process problem signals from environments that Kepler has no direct control over. The key concept in that context is operational profiles – the frequency of usage of different Kepler and other operations, and their relationship to run-time failures. Operational profiles are an essential part of software reliability engineering. Typically they are created from the software requirements, and through customer reviews. Creation of operational profiles often is laborious and requires human intervention. Our approach builds an operational profile based on the actual usage from execution logs. The difficulty in using execution logs is that the amount of data to be analyzed is extremely large (more than a million records per day in many applications). Our solution constructs operational profiles by identifying all the possible clustered sequences of events (patterns) that exist in the logs. This is done very efficiently using suffix arrays data structure [NWV09, NVW+08].

Our current solution includes a contingency actor for execution of alternatives, and intelligent workflow resume option [ACM+09]. Implementation is in the process of being evaluated with different production workflows.

In contrast to typical fault-tolerant designs in business processes that are based on undoing the transactions (e.g. payment is not received then cancel booking), the recovery strategies of scientific processes can be based on re-executing the failed work items such as invoking failed services [MV06, VAB+07] or re-executing same steps. This approach is more reasonable in the context of long running, repetitive, and multi-instance scientific applications in which individual failures don't affect the overall execution of the processes. However, the main challenge of the re-execution based fault tolerance recovery is the integration of appropriate workflow fragments that re-invoke necessary services to recover the failures without making redundant computing. In designing a fault tolerant workflow, there are two possible ways: (i) Workflow designer specifies his/her fault tolerant workflow that includes the necessary fragments that will be re-executed in case of failures (via our contingency actor). In this case, the workflow management system should be able to verify the model provided by workflow designer to ensure the fault tolerance is configured correctly; (ii) Workflow management system provides the *rescue* fragments without the intervention of the designer deriving them from the principal workflow model w.r.t. error types, design choices and error recovery policies. In both cases, the technical challenges are very similar. They require analysis of workflow models, actor dependencies and, the provenance and execution traces of workflow data.

In relation with the scientific workflow patterns work (next subsection), we aim to develop an infrastructure extension for KEPLER that will allow workflow designers to automate the definition of rescue fragments. Thus, with respect to the predefined error types and error tokens, the rescue fragments can be automatically defined without making redundant designs. We have developed a set of recovery strategies appropriate for SDM workflows. The implementation of these strategies is tightly coupled with the operation of generic and non-generic actors (i.e. error tokens, error types that they provide).

3.6 Patterns, Templates and Generic Actors

Scientific workflows are a set of actions performed in a given sequence in scientific problem solving. The workflows of interest to the DOE Scientific Data Management Center (SDMC) also deal with huge amounts of data, and one of the issues that arises is data-movement and whether computations should go to the data or vice versa. Currently, construction of complex workflows using available workflow capturing and development tools, languages and engines is very much an art. On the other hand, Level 1 and Level 2 users prefer using workflows, and if they have to construct them to do so in the simplest possible manner. They, understandably, prefer to focus on their work, rather than the intricasies of the underlying information technology. We believe that it is possible to develop a basic set of patterns and templates for construction of certain types of scientific workflows to make the workflow construction easier and less time consuming through elimination of unnecessary rework that occurs when a workflow solution is developed from scratch.

In software engineering, a design pattern [e.g., GHJ+05, DKV97] is a reusable abstracted solution for, or an approach to solution of, different variations of frequently occurring problems. A template, on the other hand, is a more specific [e.g., BBC+94]. It focuses on explicit details of the solution, explicit parameterization, and even explicit codes – template can be composed of multiple patterns. In the case of small scale design patterns, templates and design patterns often can be used interchangeably. Patterns make life simpler by providing faster solutions that are known to work thus minimizing rework by allowing us to reduce a problem to a known solution. Design pattern in general, can be described as a canonical solution to a specific problem which then needs to be refined, modified and customized to suit the problem. The customization of a design pattern depends on the input parameters and the specifications of the problem.

Progress to Date

A number of basic detailed workflow patterns have been identified in [YGN09]. An example of a more complex pattern that could be mapped to manage part of the Kepler dynamic forward-error correction fault-tolerance behaviors is given in [DKV97]. Additional check-pointing and "watchdog" [6] patterns would complete the picture. Similarly, provenance provisioning pattern can be constructed based on the earlier work of the SPA group [ACC+07, NV08]. The base-line for creation of most of the SDMC scientific workflow patterns is the sequence – prepare simulation, run simulation and monitor it, move data to analytics engine, analyze and view data and results. Other patterns may be needed.

Our goal is to create a library of patterns and templates involving appropriate collection of generic actors that may be used as stand-alone actors, or may have a bigger template set behind it (for example, a template that lists and splits the data files). Our first proposed step is to use information from different workflows we currently work with (specifically XGC, GEM and pixie3D to start with), and to extract similarities and formulate solution pattern(s). In this context, PNNL has been developing so called generic actors. They abstract actor providing specific capability with implementation details hidden (for example file transfer without concerning the user, unless the user really wants to know, which protocol is used for that). They may encapsulate family of actors with similar function and signature, and there is usually a general set of parameters that relate to actor's function (e.g., source and destination for generic transport actor). Such actors also have selection mechanisms to activate a concrete function or protocol, and they may encapsulate simple or complex functions

3.7 Framework for Integrated SDM Technologies for Applications (FIESTA)

SDM center computer scientists actively collaborate with fusion scientist from the Center for Plasma Edge Simulations (CPES). The main theme of this collaborative effort is to provide enabling technologies for complex coupled simulations running from petascale computers. The technologies being developed by the SDM center for CPES are fully applicable to other projects as well, and in fact, the success of many of

our techniques has led to adoption by some of the biggest data-producing codes in the DOE (*e.g.*, (*e.g.*,CHIMERA for astrophysics simulations and S3D for combustion simulations). Furthermore, there is already a lot of interest from other FSP projects in the technologies developed within CPES, as well as collaboration with ITER simulation software developers in France.

Progress to Date

This year we have worked on a number of important elements of our end-to-end solution. This includes

- Integrating GSI certificates into the dashboard. Now users will have the capability to submit workflow analysis jobs on our analysis cluster through the dashboard.
- Vector graphics for the dashboard. This allows users the capability to change the plots for interactive analysis for x-y plots.
- New functionality in vector graphics to allow for many windows to be dragged onto the canvas, and we have a zoom which zooms all of the windows for faster scientific discovery.
- A new plotting utility for use for the workflow and dashboard which can read in netcdf, hdf5, and adios-bp files. We have shown that for the adios-bp files, they are approximately 2x faster than the other file formats for xgc-1 files. The backbone of the plotting package is vtk and xmgrace. This now replaces avs/express. The plotter routine will eventually be merged into the portable dashboard package so that outside users will be able to produce publication quality plots even outside of the workflow environment.
- Many internal changes in the dashboard that has been done in conjunction with the Utah team.
- Integration of 3D visualization into the dashboard. This work is ongoing and is done in conjunction with Rutgers University and Utah. In both cases users can go to the dashboard for semi-interactive 3D visualization.
- Integration of Matlab jobs with the dashboard. This allows us to upload jobs, and launch matlab jobs. This work in on-going.
- Change in the architecture to allow for no-xml files. This allows the dashboard to be truly portable and allows us to perform different actions depending on the rank of the data.
- New workflows for the GTC team, S3D team, GTS team, GEM team.
- New coupling workflow being created for XGC0-GEM coupling. This is ongoing work.
- A workflow which works with ADIOS with DataSpaces for memory to memory coupling and memory to file to file coupling.
- Wis ESMF team to make initial contact and to provide some support for their workflow.
- Work with N. Samatova to provide R support in the dashboard.

Another aspect of an end-to-end solution is attention to the environment in which workflows may operate. Currently those environments include preparation clusters, supercomputers for simulation runs, and post-processing analytics clusters. A relatively novel paradigm is that of cloud computing. In fact, in the last month DOE has announced that it intends to pursue so called science clouds [FEL09]. We have anticipated these developments, and in order to understand this environment as well as understand its impact on scientific workflows (initially probably in the pre- and post-processing workflow domain) we have been tracking cloud computing technology for several years now [e.g., VOU08, VOU09, DVS+09]. We envision situations where transfer of the data produced during a workflow managed simulation run would be too expensive or too impractical to move to an analytics environment too distant from the supercomputer. Instead the simulation facility would have a component of an analytics cloud on site. The analytics “personality” of the scientists – full analytics and visualization stack or “image”: operating

system, middleware and analytics and other algorithms and software - would be moved to the data via that cloud. After the analysis is over, and the results of the analysis saved, the visiting "image" would be wiped out and that of another scientist or scientific team loaded and the process would repeat.

3.8 Dissemination and Outreach

Two key activities at SDMC SPA are dissemination of the results of its work through publications, presentations and participation in different forums, and transfer of its technology through outreach activities such as tutorials and direct work with user groups. All publications and principal event (such as All Hands Meetings) meetings are on-line, along with other SDMC information (

Over the last year we have published or prepared for publication over 30 papers, reports and tutorials (see Publications subsection). We have also participated in numerous conferences, meetings, panels and other events.

- Participation and report on Scientific Grand Challenges in Fusion Energy Sciences and the Role of Computing at the Extreme (Washington DC, 18-Mar-2009, http://extremecomputing.labworks.org/fusion/index.stm
- Participation in SDMC review (Washington DC, 21-Apr-2009)
- Tutorial at SC08 (Austin, TX, 16-Nov-2008, https://kepler-project.org/developers/events/full-day-kepler-tutorial, tutorial #S06 - http://sc08.supercomputing.org/?pg=tutorials.html)
- List of major conferences we presented at, and participated in, is the publication list.
- Fault-Tolerance Meeting (San Diego, CA,
    - Document: http://groups.google.com/group/spa-dev/web/SPA-FaultTolerance-V6.docx
- Templates Meeting (University of Utah, Salt Lake City,
    - Templates paper (http://www.cs.ucdavis.edu/research/tech-reports/2009/CSE-2009-3.pdf)
    - Templates summary slides (https://indico.lbl.gov/materialDisplay.py?contribId=22&materialId=slides&confId=0

SPA Software

- Distributable Dashboard (http://sdm7.csc.ncsu.edu/download_dashboard/)
- Portable ORNL Dashboard (https://ewok-web2.ccs.ornl.gov/portable/)
- Contributions to KEPLER (https://kepler-project.org/)

FY 2009 Publications by SDM center (61)

[ABM+09] M. Anand, S. Bowers, T. McPhillips, and B. Ludaescher, Exploring Scientific Workflow Provenance Using Hybrid Queries over Nested Data and Lineage Graphs, In 21st Intl. Conf. on Scientific and Statistical Database Management (SSDBM), New Orleans, 2009. in Lecture Notes In Computer Science; Vol. 5566, Proceedings of the 21st International Conference on Scientific and Statistical Database Management, New Orleans, LA, USA , Pages: 237 – 254.

[ACM+09] Ilkay Altintas, Daniel Crawl, Pierre Mouallem, Ustin Yildiz, Mladen Vouk, SPA Fault-Tolerance Architecture Design, internal document (http://groups.google.com/group/spa-dev/web/SPA-FaultTolerance-V6.docx).

[AVK+08] Ilkay Altintas, Mladen Vouk, Scott Klasky, Norbert Podhorszki, Daniel Crawl, "Introduction to scientific workflow management," Tutorial S06 presented at Supercomputing 2008, Nov 16, 2008, Austin, Texas.

[AWE+09] Abbasi, H., Wolf, M., Eisenhauer, G., Klasky, S., Schwan, K., and Zheng, F. 2009. DataStager: scalable data staging services for petascale applications. In Proceedings of the 18th ACM international Symposium on High Performance Distributed Computing (Garching, Germany, June 11 - 13, 2009). HPDC '09. ACM, New York, NY, 39-48.

[BAD+09] D Batchelor, G Abla, E D'Azevedo, G Bateman, D E Bernholdt, L Berry, P Bonoli, R Bramley, J Breslau, M Chance, J Chen, M Choi, W Elwasif, S Foley, G Fu, R Harvey, E Jaeger, S Jardin, T Jenkins, D Keyes, S Klasky, S Kruger, L Ku, V Lynch, D McCune, J Ramos, D Schissel, D Schnack and J Wright, Advances in simulation of wave interactions with extended MHD phenomena, IOP conference series SciDAC, 2009.

[BKM09] Roselyne Barreto, Scott Klasky, Norbert Podhorszki, Pierre Mouallem, Mladen Vouk, Collaboration Portal for Petascale Simulations," CTS 2009, pp. 384-393.

[BMR+08] Shawn Bowers, Timothy McPhillips, Sean Riddle, Manish Anand, Bertram Ludaescher, Kepler/pPOD: Scientific Workflow and Provenance Support for Assembling the Tree of Life, in the proceedings of the International Provenance and Annotation Workshop (IPAW) June 17-18, 2008, Salt Lake City, published in "Provenance and Annotation of Data and Processes, editors J. Freire, D. Koop and L. Moreau, Springer-Verlag, Berlin Heidelberg, 2008, 70-77.

[BSK09] P. Breimyer, N.F. Samatova, G. Kora, Web-Enabled R for Large-Scale Collaborative Data Mining: A Survey, The International Conference on Information and Knowledge Engineering (IKE), 2009.

[CCS+09] Jacqueline Chen, Alok Choudhary, Bronis de Supinski, Matthew DeVries, Evatt Hawkes, Scott Klasky, Wei-Keng Liao, Kwan-Liu Ma, John Mellor-Crummy, Norbert Podhorski, Ramanan Sankaran, Sameer Shende, Chun Sang Yoo. Terascale Direct Numerical Simulations of Turbulent Combustion Using S3D. In the Journal of Computational Science & Discovery, Volume 2, Number 015001, 2009.

[CKD+09] C S Chang, S Ku, P Diamond, M Adams, R Barreto, Y Chen, J Cummings, E D'Azevedo, G Dif-Pradalier, S Ethier, L Greengard, T S Hahm, F Hinton, D Keyes, S Klasky, Z Lin, J Lofstead, G Park, S Parker, N Podhorszki, K Schwan, A Shoshani, D Silver, M Wolf, P Worley, H Weitzner, E Yoon and D Zorin, Whole-volume integrated gyrokinetic simulation of plasma turbulence in realistic diverted-tokamak geometry", IOP conference series SciDAC, 2009.

[CKP+10] Cummings, Klasky, Podhorszki, Barreto, Lofstead, Schwan, Docan, Parashar, Sim, Shoshani, "EFFIS: and End-to-end Framework for Fusion Integrated Simulation", submitted to PDP 2010, http://www.pdp2010.org/.

[CLG+09] Alok Choudhary, Wei-keng Liao, Kui Gao, Arifa Nisar, Robert Ross, Rajeev Thakur, and Robert Latham. Scalable I/O and Analytics. In the Journal of Physics: Conference Series, Volume 180, No. 012048 (10pp), August 2009. (Proceedings of SciDAC conference, 14-18 June 2009, San Diego, California, USA).

[CPP+088] J. Cummings, A. Pankin, N. Podhorszki, G. Park, S. Ku, R. Barreto, S. Klasky, C. S. Chang, H. Strauss, L. Sugiyama, P. Snyder, D. Pearlstein, B. Ludaescher, G. Bateman, A. Kritz, and the CPES Team, Plasma Edge Kinetic-MHD Modeling in Tokamaks Using Kepler Workflow for Code Coupling, Data Management and Visualization, Communications in Computational Physics, 4(3), September 2008.

[CRI09] Terence Critchlow, Scientific Process Automation Improves Data Interaction Workflow infrastructure automates time-intensive manual processes, Scientific Computing, Rockaway NJ 07866, 2009, http://www.scientificcomputing.com/article-hpc-Scientific-Process-Automation-Improves-Data-Interaction-082809.aspx#

[CSD+09] J. Chen, Choudhary A, De Supinski B, DeVries M, Hawkes E, Klasky S, Liao W, Ma K, Mellor-Crummey J, Podhorszki N, Sankaran R, Shende S and Yoo C. Terascale direct numerical simulations of turbulent combustion using S3D. Computational Science and Discovery, 2 015001 (31pp), Jan 2009.

[DVS+09] Patrick Dreher, Mladen A. Vouk, Eric Sills, Sam Averitt, "Cost Effective Cloud Computing Using VCL," to appear in Trends in HPC and Grids", 2009.

[GK08] A. Gezahegne and C. Kamath, "Tracking non-rigid structures in computer simulations," IEEE International Conference on Image Processing, San Diego, October 2008, pp. 1548-1551.

[GLC+09] Kui Gao, Wei-keng Liao, Alok Choudhary, Robert Ross, and Robert Latham. Combining I/O Operations for Multiple Array Variables in Parallel NetCDF. In the Proceedings of the Workshop on Interfaces and Architectures for Scientific Data Storage, held in conjunction with the the IEEE Cluster Conference, New Orleans, Louisiana, September 2009.

[GLN+09] Kui Gao, Wei-keng Liao, Arifa Nisar, Alok Choudhary, Robert Ross, and Robert Latham. Using Subfiling to Improve Programming Flexibility and Performance of Parallel Shared-file I/O. In the Proceedings of the International Conference on Parallel Processing, Vienna, Austria, September 2009.

[GNB+09] G. Grider, J. Nunez, J. Bent, S. Poole, R. Ross, and E. Felix. Coordinating government funding of file system and I/O research through the high end computing university research activity. In SIGOPS Operating Systems Review, January 2009.

[GWR08] Peng Gu, Jun Wang, and Robert Ross. Bridging the gap between parallelfile systems and local file systems: A case study with PVFS. In 37th International Conference on Parallel Processing, pages 554–561, September 2008.

[IBC+08] Florin Isaila, Francisco Javier Garcia Blas, Jesus Carretero, Wei-keng Liao, and Alok Choudhary. AHPIOS: An MPI-based Ad-hoc Parallel I/O System. In the Proceedings of 14th Intl Conference on Parallel and Distributed Systems, Melbourne, Victoria, Australia, December 2008.

[Kam08a] C. Kamath, "Application-driven data analysis", Editorial, Statistical Analysis and Data Mining, Vol 1, Issue 5, April 2009.

[Kam09] C. Kamath, Scientific Data Mining: A Practical Perspective, SIAM, Philadelphia, PA, May 2009.

[KWK+09] C. Kamath, N. Wale, G. Karypis, G. Pandey, V. Kumar, K. Rajan, N. F. Samatova, P. Breimyer, G. Kora, C. Pan,S.Yoginath, "Scientific Data Analysis", book chapter in "Scientific Data Management", A. Shoshani and D. Rotem, editors, CRC Press/Taylor and Francis books, 2009, to appear.

[LAB+09] B. Ludäscher, I. Altintas, S. Bowers, J. Cummings, T. Critchlow, E. Deelman, D. D. Roure, J. Freire, C. Goble, M. Jones, S. Klasky, T. McPhillips, N. Podhorszki, C. Silva, I. Taylor, and M. Vouk. In A. Shoshani and D. Rotem, editors Scientific Process Automation and Workflow Management, Scientific Data Management: Challenges, Existing Technology, and Deployment, Computational Science Series, chapter 13. Chapman & Hall/CRC, 2009.

[LBM+09] B. Ludaescher, S. Bowers, and T. McPhillips. M. T. Özsu and L. Liu, editors, Scientific Workflows, Encyclopedia of Database Systems. Springer, 2009

[LC08] Wei-keng Liao, and Alok Choudhary. Dynamically Adapting File Domain Partitioning Methods for Collective I/O Based on Underlying Parallel File System Locking Protocols. In the Proceedings of International Conference for High Performance Computing, Networking, Storage and Analysis, Austin, Texas, November 2008.

[LCL+09] Samuel Lang, Philip Carns, Robert Latham, Robert Ross, Kevin Harms, and William Allcock. I/O performance challenges at leadership scale. In Proceedings of Supercomputing, November 2009.

[LWM+09] B. Ludaescher, M. Weske, T. McPhillips, S. Bowers, Scientific Workflows: Business as Usual?, In 7th Intl. Conf. on Business Process Management (BPM), Ulm, Germany, 2009.

[LXH+09] Z. Lin, Y Xiao, I Holod, W Zhang, W Deng, S Klasky, J Lofstead, C Kamath and N Wichmann, Advanced simulation of electron heat transport in fusion plasmas, IOP conference series SciDAC, 2009.

[LXH+09] Z. Lin, Y. Xiao, I. Holod, W. Zhang, W. Deng, S. Klasky, J. Lofstead, C. Kamath, and N. Wichmann, "Advanced simulation of electron heat transport in fusion plasmas", SciDAC 2009, Journal of Physics Conference Series.

[MBK+09] Pierre Mouallem, Roselyne Barreto, Scott Klasky, Norbert Podhorszki, Mladen Vouk, Tracking Files using the Kepler Provenance Framework, In 21st Intl. Conf. on Scientific and Statistical Database Management (SSDBM), New Orleans, 2009. in Lecture Notes In Computer Science; Vol. 5566, Proceedings of the 21st International Conference on Scientific and Statistical Database Management, New Orleans, LA, USA, pp. 273-282

[MOU08] Pierre Mouallem, Portable ORNL Dashboard, https://ewok-web2.ccs.ornl.gov/portable/

[NAG09] Mei Nagappan, Distributable Dashboard, http://sdm7.csc.ncsu.edu/download_dashboard/

[NLC08] Arifa Nisar, Wei-keng Liao, and Alok Choudhary. Scaling Parallel I/O Performance through I/O Delegate and Caching System. In the Proceedings of International Conference for High Performance Computing, Networking, Storage and Analysis, Austin, Texas, November 2008.

[NV08] Meiyappan Nagappan, and Mladen Vouk, "A Privacy Policy Model for Sharing of Provenance Information in a Query Based System," presented at the International Provenance and Annotation Workshop (IPAW) June 17-18, 2008, Salt Lake City, published in "Provenance and Annotation of Data and Processes, editors J. Freire, D. Koop and L. Moreau, Springer-Verlag, Berlin Heidelberg, 2008, pp. 62-69.

[NVW+08] M. Nagappan, Vouk, M.A., Wu, K., Sim, A., Shoshani, A.. "Efficient Operational Profiling of Systems using Suffix Arrays on Execution Logs." Student Paper in the 19th International Symposium on Software Reliability Engineering, 10-14 Nov, 2008, Redmond, WA, pp. 313-314.

[NWV09] Meiyappan Nagappan, Kesheng Wu, and Mladen Vouk, Efficiently Extracting Operational Profiles from Execution Logs using Suffix Arrays, In proceedings of the 20th International Symposium on Software Reliability Engineering, November 16-19 2009, Mysuru, India (to appear).

[PGK+08.a] Paul Breimyer, Nathan Green, Vinay Kumar, Nagiza F. Samatova, BioDEAL: Biological Data-Evidence-Annotation Linkage System; Proceedings of the Conference on Bioinformatics and Biomedicine (BIBM), 2008.

[PGK+08.b] Paul Breimyer, Nathan Green, Vinay Kumar, Nagiza F. Samatova, BioDEAL: Biological Data-Evidence-Annotation Linkage System; Journal of BMC Medical Informatics and Decision Making, 2009 (invited).

[PGR+09] Thomas Peterka, David Goodell, Robert Ross, Han-Wei Shen, and Rajeev Thakur. A configurable algorithm for parallel image-compositing applications. In Proceedings of Supercomputing, November 2009.

[PKH+09] P. Breimyer, G. Kora, W. Hendrix, N.F. Samatova, pR: Lightweight, Easy-to-Use Middleware to Plugin Parallel Analytical Computing with R; The International Conference on Information and Knowledge Engineering (IKE), 2009.

[PKL+09] N. Podhorszki, S. Klasky, Q. Liu, C. Docan, M. Parashar, H. Abbasi, J. Lofstead, K. Schwan, M. Wolf, F. Zheng, J. Cummings, "Plasma fusion code coupling using scalable I/O services and scientific workflows", accepted to works09, http://www.isi.edu/works09/.

[PLB+09] M. Polte, J. Lofstead, J. Bent, G. Gibson, S. Klasky, Q. Liu, M. Parashar, K. Schwan, M. Wolf, "… And eat it too: High read performance in write-optimized HPC I/O middlware file formats", submited to PDSW 2009.

[PRS+09] T. Peterka, R. Ross, H-W. Shen, K-L. Ma, W. Kendall, and H. Yu. Parallel visualization on leadership computing resources. In SciDAC 2009, Journal of Physics: Conference Series, San Diego, CA, July 2009.

[PRY+08] T. Peterka, R. Ross, H. Yu, K. Ma, W. Kendall, and J. Huang. Assessing and improving large-scale parallel volume rendering on the IBM Blue Gene/P. In Proceedings of Supercomputing 2008 Ultrascale Visualization Workshop, Austin, TX, November 2008.

[PYR+09] Tom Peterka, Hongfeng Yu, Robert Ross, Kwan-Liu Ma, and Rob Latham. End-to-end study of parallel volume rendering on the IBM Blue Gene/P. In Proc. ICPP 09, Vienna, Austria, September 2009.

[RCG+09] Robert Ross, Alok Choudhary, Garth Gibson, and Wei-Keng Liao. Parallel data storage and access. In Arie Shoshani and Doron Rotem, editors, Scientfic Data Management: Challenges, Technology, and Deployment. Chapman & Hall/CRC, 2009 (expected).

[RCM09] Robert Ross, Philip Carns, and David Metheney. Parallel file systems. In Yupo Chan, John Talburt, and Terry Talley, editors, Data Engineering: Mining, Information and Intelligence. Springer, October 2009.

[SAH+09] Henry E. Schaffer, Samuel F. Averitt, Marc I. Hoit, Aaron Peeler, Eric D. Sills, and Mladen A. Vouk, "NCSU's Virtual Computing Lab: A Cloud Computing Solution," IEEE Computer, July 2009, pp. 94-97. (http://www2.computer.org/portal/c/document_library/get_file?uuid=b150980d-b8dc-445c-950c-25c0549b7982&groupId=889101

[STK+09] E. Santos, J. Tierny, A, Khan, B. Grimm, L. Lins, J. Freire, V. Pascucci, C. T. Silva, S. Klasky, R. Barreto, N. Podhorszki, "Enabling Advanced Visualization Tools in a Web-Based Simulation Monitoring System", IEEE International Conference on e-Science, 2009:

[VOU08] Mladen Vouk, "Cloud Computing – Issues, Research and Implementations," Journal of Computing and Information Technology, Vol 16 (4), 2008, pp 235-246.

[VOU09] M. Vouk, "Cloud Computing and the Next Generation Data Centers," Invited Presentation at the CECIIS 2008 - Central European Conference on Information and Intelligent Systems, Varazdin, Croatia, September 24, 2008 – September 26, 2008.

[VRA+09] M. A. Vouk, A. Rindos, S. F. Averitt, J. Bass, M. Bugaev, A. Kurth, A. Peeler, H. E. Schaffer, E. D. Sills, S. Stein, J. Thompson, and M. Valenzisi, "Using VCL technology to implement distributed reconfigurable data centers and computational services for educational institutions," IBM Journal of Research and Development, Volume 53, No 4, June, 2009 (http://www.research.ibm.com/journal/rd53-4.html)

[VSA+09] M. Vouk, E. Sills, S. Averitt, A. Peeler, "Using VCL to Power Clouds", OGF25/EGEE User Forum Workshop “From Grids to Clouds – a workshop for Grid users facing the Cloud,” Catania, Italy, March 2-6, 2009.

[YDV09] Yu,W., Drokin, O., Vetter, J.S. Design, Implementation, and Evaluation of Transparent pNFS on Lustre, 23rd IEEE International Parallel and Distributed Processing Symposium (IPDPS'09)

[YGN09a1] U. Yildiz, A. Guabtni, and A. H. H. Ngu, Towards scientific workflow patterns," in Proceedings of the 4th Workshop on Workflows in Support of Large-Scale Science, In conjunction with Super Computing, SC (USA), ACM Press, 2009.

[YGN09b2] U. Yildiz, A. Guabtni, and A. H. H. Ngu, Business versus scientific workflows: A comparative study," in Proceedings of the IEEE Third International Workshop on Scientific Workflows, SWF (In conjunction with 7th IEEE International Conference on Web Services (ICWS 2009)), (USA), IEEE Computer Society Press, 2009.

[YGN09c3] U. Yildiz, A. Guabtni, and A. H. H. Ngu, Business versus Scientific Workflow: A Comparative Study, 2009. Research Report, Department of Computer Science, University of California, Davis, CSE-2009-3, http://www.cs.ucdavis.edu/research/tech-reports/2009/CSE-2009-3.pdf.

[YRW+08] Yu,W., Rao,N.S.V., Wyckoff,P.,, Vetter,J.S. (2008). Performance of RDMA-capable Storage Protocols on Wide-Area Network, Peta-Byte Storage Workshop 2008 (PDSW08).

References

[GHS95] D. Georgakopoulos, M. Hornick, and A. Sheth, "An Overview of Workflow Management: From Process Modeling to Workflow Automation Infrastructure,"Distributed and Parallel Databases, Vol. 3(2), April 1995.

[HRG+00] Elias N. Houstis, John R. Rice, Efstratios Gallopoulos, Randall Bramley (editors), Enabling Technologies for Computational Science Frameworks, Middleware and Environments, Kluwer-Academic Publishers, Hardbound, ISBN 0-7923-7809-1, 2000.

[LAB+06] Scientific Workflow Management and the Kepler System, B. Ludaescher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. A. Lee, J. Tao, and Y. Zhao, Concurrency and Computation: Practice & Experience, 2006.

[ABB+03] Altintas I., S. Bhagwanani, D. Buttler, S. Chandra, Z. Cheng, M. Coleman, T. Critchlow, A. Gupta, W. Han, L. Liu, B. Ludaescher, C. Pu, R. Moore, A. Shoshani, M. Vouk, "A Modeling and Execution Environment for Distributed Scientific Workflows," Proc. 15th IEEE International Conference on Scientific and Statisical Database Management (SSDBM 2003).

[ABC+06] Ilkay Altintas, Oscar Barney, Zhengang Cheng, Terence Critchlow, Bertram Ludaescher, Steve Parker, Arie Shoshani and Mladen Vouk, "Accelerating the scientific exploration process with scientific workflows," sciDAC 2006, Journal of Physics: Conference Series 46 (2006), 468-478, doi:10.1088/1742-6596/46/1/065

[ACC+07] I. Altintas, G. Chin, D. Crawl, T. Critchlow, D. Koop, J. Ligon, B. Ludaescher, P. Mouallem, Nagappan, N. Podhorszki, C Silva, M. Vouk, "Provenance in Kepler-based Scientific Workflow Systems,” Poster # 41, at Microsoft eScience Workshop Friday Center, University of North Carolina, Chapell Hill, NC, October 13 - 15, 2007, pp. 82

[BBC+94] R. Barrett, M. Berry, T. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine and H. Vorst, "Templates - for the Solution of Linear Systems: Building Blocks for Iterative Methods". Siam, 1994

[BCK+07] Roselyne Barreto, Terence Critchlow, Ayla Khan, Scott Klasky,, Leena Kora, Jeffrey Ligon, Pierre Mouallem, Meiyappan Nagappan, Norbert Podhorszki, Mladen Vouk, “Managing and

Monitoring Scientific Workflows through Dashboards, Poster # 93, at Microsoft eScience Workshop Friday Center, University of North Carolina, Chapell Hill, NC, October 13 - 15, 2007, pp. 108

[BVP00] Balay R.I, Vouk M.A., Perros H., "Performance of Network-Based Problem-Solving Environments," Chapter 18, in Enabling Technologies for Computational Science Frameworks, Middleware and Environments, editors Elias N. Houstis, John R. Rice, Efstratios Gallopoulos, Randall Bramley, Hardbound, ISBN 0-7923-7809-1, 2000

[CA08] Daniel Crawl and Ilkay Altintas, A Provenance-Based Fault Tolerance Mechanism for Scientific Workflows, in the proceedings of the International Provenance and Annotation Workshop (IPAW) June 17-18, 2008, Salt Lake City, published in "Provenance and Annotation of Data and Processes, editors J. Freire, D. Koop and L. Moreau, Springer-Verlag, Berlin Heidelberg, 2008, pp. 152-159

[CCA09] http://www.cca-forum.org/, accessed October 2009

[CFS+08] S. P. Callahan, J. Freire, C. E. Scheidegger, C. Silva, and Huy T. Vo, Towards Provenance-Enabling ParaView, in the proceedings of the International Provenance and Annotation Workshop (IPAW) June 17-18, 2008, Salt Lake City, published in "Provenance and Annotation of Data and Processes, editors J. Freire, D. Koop and L. Moreau, Springer-Verlag, Berlin Heidelberg, 2008, pp. 120-127

[CL02] I. Crnkovic and M. Larsson (editors), Building Reliable Component-Based Software Systems, Artech House Publishers, ISBN 1-58053-327-2, 2002, http://www.idt.mdh.se/cbse-book/

[DBN+96] R.L. Dennis, D.W. Byun, J.H. Novak, K.J. Galluppi, C.C. Coats, and M.A. Vouk, "The Next Generation of Integrated Air Quality Modeling: EPA's Models-3," Atmospheric Environment, accepted, in print, expected 1996.

[DKV97] Daniels, K. Kim and M. A. Vouk, "The Reliable Hybrid Pattern - A Generalized Software Fault Tolerance Design Pattern," The fourth Pattern Languages of Programming Conference (PLoP'97), Monticello, IL, September, Washington University Technical Report, Pages 97-34, 1997

[DOE04] R. Mount et al., Department of Energy, Office of Science report, "Data Management Challenge". Nov 2004, http://www.er.doe.gov/ascr/Final-report-v26.pdf

[DOU02] B. Douglass, "Real-Time Design Patterns: Robust Scalable Architecture for Real Time Systems". Addison-Wesley, 2002.

[EBV95] Elmaghraby S.E., Baxter E.I., and Vouk M.A., "An Approach to the Modeling and Analysis of Software Production Processes," Intl. Trans. Operational Res., Vol. 2(1), pp. 117-135, 1995.

[Elm66] Elmaghraby S.E., "On generalized activity networks," J. Ind. Eng., Vol. 17, 621-631, 1966

[FEL09] Michael Feldman, DOE Labs to Build Science Clouds, HPC Wire, October 13, 2009, http://www.hpcwire.com/features/DOE-Labs-to-Build-Science-Clouds-64189872.html

[GBA+07] Antoon Goderis, Christopher Brooks, Ilkay Altintas, Edward A. Lee and Carole Goble. Composing Different Models of Computation in Kepler and Ptolemy II, Proc. of the 2nd Int. Workshop on Workflow Systems in e-Science (WSES 07) in conjunction with the Int. Conference on Computational Science (ICCS) 2007, Beijing, China, May 27-30, 2007

[GHJ+95] Gamma, E., Helm, R., Johnson, R. and Vlissides, J. "Design Patterns: Elements of Reusable Object Oriented Software". Addison-Wesley, 1995.

[Kam06] C. Kamath, "Mining Science Data", SciDAC 2006, Scientific Discovery through Advanced Computing, Journal of Physics Conference Series, Volume 46, 2006, pp. 500-504.

[Kam08a] C. Kamath, "Application-driven data analysis", Editorial, Statistical Analysis and Data Mining, Vol 1, Issue 5, 2008.

[Kam08b] C. Kamath, "Sapphire: Experiences in Scientific Data Mining,", SciDAC 2008, Journal of Physics Conference Series 125, 012094, July 2008.

[LB90] J.C. Laprie, and C. Beounes, “Definition and Analysis of Hardware- and Software-Fault-Tolerant Architectures”, IEEE Computer Society Press, Volume 23, Issue 7, Pages: 39 – 51, July 1990.

[LG05] B. Ludaescher and C. A. Goble, editors. ACM SIGMOD Record, Special Section on Scientific Workflows, volume 34(3), September 2005.

[LK07] N. S. Love and C. Kamath, "Image analysis for the identification of coherent structures in plasma," Applications of Digital Image Processing, XXX, SPIE Conference 6696, San Diego, August 2007.]

[MV06] Mouallem, P. and Vouk.M., “Fault Tolerance and Reliability in Scientific Workflows, “ Proc. of ETFS 2006 – International Workshop on Engineering of Fault-Tolerant Software, Luxembourg, Luxembourg, 12-13 June 2006, pp.27-41

[MV96] D.F. McAllister, and M.A. Vouk, "Software Fault-Tolerance Engineering," Chapter 14 in Handbook of Software Reliability Engineering, McGraw Hill, pp. 567-614, January 1996

[SAC+07] Arie Shoshani, Ilkay Altintas, Alok Choudhary, Terence Critchlow, Chandrika Kamath, Bertram Ludäscher, Jarek Nieplocha, Steve Parker, Rob Ross, Nagiza Samatova, Mladen Vouk , :SDM Center Technologies for Accelerating Scientific Discoveries," SciDac 2007 Proceedings Dec 2007, Journal of Physics, Conference Series, Vol. 78, paper #012068, 5 pages.

[SAC+07a] Arie Shoshani, Ilkay Altintas, Alok Choudhary, Terence Critchlow, Chandrika Kamath, Bertram Ludaescher, Jarek Nieplocha, Steve Parker, Rob Ross, Nagiza Samatova, Mladen Vouk, Scientific Data Management: Essential Technology for Accelerating Scientific Discoveries, CTWatch Quarterly, Volume 3, Number 4, November 2007.

[SPP+06] J. Salas, F. Perez, M. Patia-Martinez, R. Jiminez-Peris, “WS-Replication: A Framework for Highly Available Web Services, Proceedings of the 15th International World Wide Web Conference, Edinburgh, Scotland, May 2006, pp. 357 – 366

[SV96] Singh M.P., Vouk M.A., “Scientific workflows: scientific computing meets transactional workflows,” Proceedings of the NSF Workshop on Workflow and Process Automation in Information Systems: State-of-the-Art and Future Directions, Univ. Georgia, Athens, GA, USA; 1996, pp.SUPL28-34.

[VAB+07] M. A. Vouk, I. Altintas R. Barreto, J. Blondin, Z.Cheng, T. Critchlow, A. Khan, S. Klasky, J. Ligon, B. Ludaescher, P. A. Mouallem, S. Parker, N. Podhorszki, A. Shoshani, C. Silva, ” Automation of Network-Based Scientific Workflows,” Proc. of the IFIP WoCo 9 on Grid-based Problem Solving Environemnts: Implications for Development and Deployment of Numerical Software, IFIP WG 2.5 on Numerical Software, Prescott, AZ, 2006, printed in IFIP, Vol 239, “Grid-Based Problem Solving Environments, eds. Gaffney PW and Pool JCT (Boston: Springer), pp. 35-61, 2007

[VOU05] Mladen A Vouk, "Software Reliability Engineering of Numerical Systems," Chapter 13, in Accuracy and Reliability in Scientific Computing, Editor: Bo Einarsson, ISBN 0-89871-584-9, SIAM, 2005, pp 265-300.

[VS97] Vouk M.A., and M.P. Singh, "Quality of Service and Scientific Workflows," in The Quality of Numerical Software: Assessment and Enhancements, editor: R. Boisvert, Chapman & Hall, pp.77-89 , 1997.

Appendix 1: Tutorials, training, outreach, invited presentations

Tutorial: Robert Latham, Robert Ross, Marc Unangst, and Brent Welch. Parallel i/o in practice. SC 2009, Portland, OR, November 2009 (upcoming).

Tutorial: William Gropp, Ewing Lusk, Robert Ross, and Rajeev Thakur. Advanced MPI. SC2009, Portland, OR, November 2009 (upcoming).

Invited talk: R. Ross, "Extreme Scale I/O Systems," IEEE Nuclear Science Symposium Data-Intensive Workshop, Orlando, FL, October 2009.

Invited talk: Robert Ross, "The Scientific Data Management Center", High-End Computing File Systems and I/O Conference, Arlington, VA, August 2009.

Invited talk: Robert Ross, "Parallel I/O in Practice," CScADS Workshop on Leadership-class Machines, Petascale Applications, and Performance Strategies, Tahoe City, CA, July 2009.

Tutorial: Robert Latham and Robert Ross. Parallel I/O in practice. SciDAC Tutorials Day, San Diego, CA, June 2009.

Invited talk: Alok Choudhary, Invited Panelist on the Panel on "Power Management Challenges" at the International Conference on Parallel Processing, Vienna, Austria, September 2009.

Invited talk: Alok Choudhary, "Scalable I/O and Analytics", invited talk in the SciDAC conference, San Diego, CA, June 2009.

Invited talk: Sam Lang. PVFS: A file system for high performance computing, 2009. Invited talk presented at Illinois Institute of Technology.

Tutorial: W. Gropp, E. Lusk, R. Ross, R. Thakur, "Advanced MPI," SC2008, Austin, TX, November 2008.

Tutorial: R. Latham, R. Ross, M. Unangst, and B. Welch, "Parallel I/O in Practice," SC2008, Austin, TX, November 2008.

Invited talk: C. Kamath, "WindSENSE: Enabling control room operators to improve management of wind resources", BPA Wind Ramp Event Meeting, Sept 30-Oct 1, 2009

Invited talk: C. Kamath, "Experience in Scientific Data Mining," Keynote presentation, NREL Informatics workshop, July 16, 2009

Tutorial: Ilkay Altintas, Mladen Vouk, Scott Klasky, Norbert Podhorszki, Daniel Crawl, "Introduction to scientific workflow management," Tutorial S06 presented at Supercomputing 2008, Nov 16, 2008, Austin, Texas.

Talks: presented at All Hands Meeting July 2008 (Myrtle Beach, SC, 10-Jul-2008, https://sdm.lbl.gov/twiki/bin/view/SDMCenter/Meetings/2008-10-07

Talks: presented at All Hands Meeting October 2009 (UC Davis, CA, 12-Oct-2009, https://indico.lbl.gov/conferenceTimeTable.py?confId=0)

Appendix 2:
Collaboration with Application Projects and other Centers and Institutes

Application Domains	Workflow (Kepler)	Metadata and Provenance	Data Movement	Indexing (FastBit)	Parallel I/O (pNetCDF, etc.)	Parallel Statistics (pR, …)	Feature Extraction	ADIOS
Fusion (Chang, Ku, CPES)	monitoring, code-couple	Dashboard	DataMover-Lite	Toroidal meshes		pR	Blob tracking	Accelerate I/O
Fusion (Breslau, Pomphrey, PPPL)							Poincare plots	
Fusion (Nevins, LLNL)			DataMover-Lite			parallelize GVK	poincare plots	
Combustion (Chen, Sandia)	monitoring, code-couple	dashboard		flame front	Global Access	pMatlab	tranient events	Accelerate I/O
Combustion (Bell, LBNL)			DataMover					
Climate Modeling (Drake)	monitoring				pNetCDF	pMatlab		
Climate-CCSM (Murphy, NCAR)	provenance recording							
Climate-Cloud (Randall, Colorado St.)					pNetCDF			
Accelarator-LWF(Geddes, …, LBNL)				visual analytics				
Accelerator Science (Ryne)					MPIO			
Astrophysics (Mezzacappa, ORNL)	monitoring, code-couple	dashboard						Accelerate I/O
Biology (Banfield, Buchanan, ORNL)	ScalaBlast					ProRata		
High Energy Physics (Lauret, BNL)			SRM, DataMover	event finding				
Groundwater (Schuchardt, PNNL)	parameter studies							

In use In progress

Details of activities described in applications vs. technologies table

Below we summarized the activities in the cells of the scientific applications vs. SDM center technologies shown in the table above. The summary is organized by technologies, and for each technology the tasks for each application project are described.

Workflow Technology tasks

- ***Application: Fusion (CPES)***

  Code Coupling: A first workflow demonstrating the coupling of XGC-0 and M3D has been developed. Next steps: Scientists want to couple other codes (e.g. ELITE and NIMROD) for improved accuracy and performance reasons. The goal is to use these workflows in production.

  Monitoring and Archiving: We have developed a workflow which on the fly (i) moves simulation output data to a secondary (remote) resource, (ii) processes (converts) data, (iii) creates images from the data, and (iv) archives the results.

- ***Application: Groundwater Modeling***

  Five different workflows associated with multi-scale simulations of subsurface biogeochemical processes have been identified as potential candidates to be represented and modeled as computational workflows. Of these five, the first workflow under implementation is one focused on a continuum simulation of flow and transport for two non-reacting tracers.

- ***Application: Combustion***

  We have worked with the S3D team to understand the basic workflow requirements. We have started to work with them to run their netcdf output and run this through a series of services which have been constructed through our alliance with the CPES SciDAC project. This includes: splitting netcdf files with infinite time dimensions to one time dimension, joining the files on another system, creating grace and png files from the data, and finally running an avs/express offscreen rendering server to produce 2d colormap images. In order to avoid costly large-scale runs with untested code, we have developed a "preparation workflow" that automatically checks out the latest simulation code from a repository, builds ("makes") the application, and runs it with a number of test cases. We are planning to extend the workflow to handle other domains (e.g. Fusion) in the future.

- ***Application: Biology***

  We are investigating using a simple web interface to define and execute multiple ScalaBLAST (large-scale sequence homology comparisons) workflows and summarize the resulting data set, providing computational biologists a novel and efficient data management capability.

Metadata and Provenance tasks

- ***Application: Combustion***

  We have discussed a need for a simplified tool for day-to-day tracking, analysis, and graphing of simulations that is integrated with the workflow and simulation tracking systems, and are currently exploring the possibility of extending our web-based data management and query tools for use with this project. This will leverage similar work that we are doing with CPES but will require adaptation to the grids used by the combustion simulations, and may require additional analysis tools to be integrated.

- ***Application: Astrophysics (TSI) and Fusion (CPES)***

Provenance: Considerable progress has been made on unification of the provenance approaches. General classification is in place (process, data, workflow and system) and we are working on the general solution. Details are at http://www.vistrails.org/index.php/SDM_Provenance. Some astrophysics and fusion specific data schemas are also in place. SDM center main contact: Mladen Vouk.

- ***Applications: Astrophysics (TSI) and Fusion (CPES) and combustion***

  Dashboard: Dashboard activities are progressing very fast. We have a prototype for CPES that is quite sophisticated. The group has regular teleconferences related to tasks and design. The architecture is now solidifying around a data-base centered repository with remote feeds and real-time updates of the job progress and states.

- **Application: Climate**

  We are working with scientists from NCAR to help develop Kepler-driven workflows and provenance for the Community Climate System Model (CCSM). CCSM belongs to an elite category of computer-based simulation models known as general-circulation models. The automatic capturing of provenance is an essential part of this activity, especially as the number and volume of simulations is expected to significantly grow.

Data Movement and Storage tasks

- ***Application: High Energy Physics***

  Storage Resource Managers (SRMs) have been used for several years by High Energy Physics projects. In cooperation with the Open Science Grid (OSG), we continue to support the STAR project in its use of our SRM. This includes its use for large scale robust data movement activity, as well as its use for dynamic data analysis tasks.

- ***Application: Fusion (CPES)***

  We have used SRM-Lite for this project as well in two different ways. The first is for a user to pull files into their workstation of laptop. For this purpose SRM-Lite has a GUI that shows progress of the transfer. The second way is for SRM-Lite to be used by a Kepler actor – future work.

- ***Collaboration: with Open Science Grid***

  LBNL has developed a test-suite for SRMs used extensively by OSG to test the compatibility and adherence to the SRM specification of several SRM implementations in the US and Europe. This work is funded by the OSG.

- ***Collaboration: with Earth System Grid***

  LBNL has been providing SRM software as well as SRM-Lite for several years now. The latest version provided is a new implementation, called the Berkeley Storage Manager (BeStMan). This work continues to evolve. This work is funded by the ESG.

- ***Application: Combustion***

  We have built a new workflow for migrating an archive from one mass storage to another. This workflow enhances earlier work with concurrent transfers over the network. It was successfully used to migrate a 10TB INCITE archive from NERSC to ORNL within 11 days. The data migration workflow has mechanisms to deal with failures, i.e., allows the user to continue the migration even after some intermediate steps have failed (e.g., due to network problems).

Indexing Technology tasks

- ***Application: Combustion***

  We had previously applied Fastbit to develop software for flame front identification, region growing, and region tracking. We also developed a simple GUI application for displaying and tracking

features in 2D combustion data. The application has moved on to 3D simulations and requires more sophisticated visualization.

- ***Application: Fusion (CPES)***

  The goal is to use Fastbit technology for searching over data in toroidal meshes. We have identified the problem and the algorithms that could potentially address the problem. We are implementing the algorithms to study the actual performance characteristics.

- ***Application: High-Energy Physics***

  In order to have the broadest impact in this community, we plan to integrate FastBit with the popular ROOT framework. The STAR software team is willing to help with ROOT expertise and manpower for testing. Work scheduled to start by June, 2007.

- ***Collaboration: with the Visualization Center (VACET)***

  We have integrated and deployed FastBit software for visualization applications. This includes extending HDFpart with FastBit indexing as well as real-time analysis of Laser Wakefield Particle Accelerator simulation data.

- ***Collaboration: with UltraScale Visualization Institute (USVI)***

  We have developed a special version of FastBit for indexing data from toroidal meshes. This code also generates regions that are selected by conditions on the variables. In collaboration with USVI, we are working on using this software for real-time explorations of toroidal data from Fusion simulations.

Parallel I/O Technologies tasks

- ***Application: Climate (CCSM)***

  The Community Climate System Model (CCSM) groups are interested in using PnetCDF as a mechanism for improving I/O performance for their large-scale simulations. We are routinely participating in concalls with NCAR and others, and PnetCDF is now an output format for the POP ocean code. Main application contact: John Drake.

- ***Application: Climate (GCRM)***

  The Global Cloud Resolving Model (GCRM) group at PNNL uses netCDF as a storage format. We have been working with their developers to use PnetCDF for better scalability on large systems. Main application contact: Bruce Palmer

- ***Application: Combustion (Jackie Chen)***

  This group is interested in improving overall I/O performance. NWU obtained an I/O kernel and developed approaches to store simulation data in a canonical format that eliminates most post-processing prior to analysis.

- ***Application: Materials (QBOX)***

  This group is interested in improving I/O performance for the QBOX code on the IBM BlueGene systems. We have performed initial experiments at ANL to better understand their I/O patterns. Main contact: Guilia Galli.

- ***Collaboration: with Petascale Data Storage Institute (PDSI)***

  We are interacting with the PDSI to further specify and prototype POSIX I/O extensions for High End Computing (HEC). We have had numerous meetings and email discussions on this topic.

- ***Collaboration: with UltraScale Visualization Institute (USVI)***

  We are discussing I/O concerns with participants in the USVI. We hope to apply parallel I/O techniques in visualization codes that will be used to view petascale simulation data.

- ***Collaboration: with Universal Nuclear Energy Density Functional (UNEDF)***

We are discussing I/O concerns with participants in the UNEDF. Our goal is to devise I/O approaches to make full utilization of leadership-class machines.

Feature Extraction tasks

- ***Application: Combustion (TSTC)***

  The goal is to develop robust techniques for quantitative identification and tracking of transient events in combustion simulation data. The purpose is to understand the process of ignition, extinction, and re-ignition.

- ***Application: Fusion (CPES)***

  The goal is to characterize and track the blobs in high-resolution, ultra-high-speed images from the gas-puff diagnostic on the NSTX. The purpose is to contribute to the success of devices such as ITER by improving the understanding of the coherent structures and validating or invalidating theories.

  Application: Fusion (RF)

  The goal is the classification and characterization of Poincare plots for simulation and experimental data. The purpose is to use the simulations to drive the experiments and use the experiments to validate the simulations. The package developed achieves high accuracy classification, and is now being used.

- ***Application: Fusion (GPS)***

  The goal is tracking of blobs in a high-dimensional particle simulations. The techniques have been developed and collaboration with the scientists is continuing.

- Application: Fusion (GSEP)

  The goal is to identify coherent structures in fluid and particle data and understand their interactions. The purpose is to gain insights into the effects of energetic particles on the performance of burning plasmas. [To be colored as "in progress"]

- Application: Renewable energy

  The goals are to understand the effects of increased wind energy on the power grid and improve the forecast of energy generated by wind farms through the identification of sensors important to the forecast as well as wind-driven anomalous events on the system. [To be colored as "in progress"]

High Performance Statistical Analysis tasks

- ***Application: Combustion***

  We initiated a dialog on providing parallel Matlab interface to her S3D library. Jackie handed over to us the parts of her Fortran90 library that deals with I/O and would like to get a plug-in of this library into parallel Matlab environment so that the subsequent analysis and visualization capabilities of parallel Matlab could be utilized. She has assigned her PhD student (David Lignel) to help us in this task. Application contact: Jackie Chen.

- ***Application: Fusion (CPES) and Collaboration with the Visualization Center***

  This task is in collaboration with George Ostrouchov and Sean Ahern. The goal is to parallelize their data analysis routines written in R using our parallel R platform. They need to handle data consisting of billion of particles and sequential R is limited for this task.

- ***Application: Climate***

  The spherical harmonic transform is a critical computational kernel of the dynamics portion of spectral atmospheric weather and climate codes. John and his team currently develop and use Matlab library for computing spherical harmonic transforms to solve simple partial differential equations on

the sphere. We identified a strategy on how to parallelize this library for them so that it could be applied to more realistic problem sizes using parallel Matlab. Application contact: John Drake

- ***Application: Climate***

  Assessment of global climate change impacts requires increasingly finer spatial and temporal resolutions from existing Earth Systems Modeling predictions. Given a fine resolution observational data and a course grain resolution simulation data, statistical downscaling could be applied to learn statistical relationships that link large-scale simulation results with fine grain regional observations. We develop a parallel Matlab library to support that. The library includes a number of components that are routinely used by climate community such as EOF, CCA, MLR, filtering routines. Application contacts: John Drake and George Ostrouchov

- ***Application: Nanoscience (DOE CNMS Center***

  This group is simulating an electron beam induced deposition process using Matlab library. We are providing parallelism to this simulation framework using parallel Matlab to bring the required efficiency. Applicatiuon contact: Philip Rack.

- ***Application: Biology (DOE Genomics:GTL projects)***

  We provide quantitative proteomics capabilities with ProRata. Application of our technologies to a number of problems in GTL community has been demonstrated. Specifically, in collaboration with B. Hettich and Carol Harwood, we reconstructed aromatic compound degradation pathways in a hydrogen producing bacteria using ProRata. Joint paper is under review. Also, we applied ProRata to quantifying the abundance of microbial communities in several DOE contaminated sites. Joint paper is being written and interesting hypotheses are generated about the presence of virus in the community that significantly changed the structure of the communities in one of the two sites. Application contact: R. Hettich, J. Banfield, C. Harwood, M. Buchanan.

- ***Collaboration: with UltraScale Visualization Institute (USVI)***

  We are discussing heterogeneous information analysis and visualization issues with participants in the USVI (Kwan-Lu Ma and Juan Huang). The primary application area is biology. We discuss issues of uncertainty representation in biological networks. We also worked with them on interactive remote visualization. The multi-cache framework with adaptive adjustment of cache parameters using statistical analysis and parameter optimization techniques has been developed and jointly published with Dr. J. Huang.

Active Storage tasks

- ***Application: base program biology project***

  This task is in collaboration with Chris Oehmen and project Scalablast which is a part of DoE base program project named "Data Intensive Computing for Complex Biological Systems" led by TP Straatsma. We identified an opportunity for postprocessing of large data files generated with Scalablast. This would be in support of their work with the Joint Genome Institute. This new task is still in definition stage.

- ***Application: climate SciDAC project***

  This task is in collaboration with Karen Schurchard who runs a SAP for the Dave Randell (Colorado) climate scidac project - Design and Testing of a Global Cloud-Resolving Model. The goal is to be able to compute statistics on the data generated from the simulations that need to be computed in response to queries coming from remote users. We still do not have the actual data and might end up working with simulated data first. This is because the SAP and the problem are new.

- ***Other activities: integration of Active Storage with Lustre***

With recent progress on getting Active Storage running with Lustre 1.6 and a new more flexible implementation approach we have been developing (user rather than kernel space), we should be in a position to actually start working with these apps in the next 2 months.