SDM center: Evolving Management Plan
May, 2007

This Management Plan for the SDM center reflects the current state of tasks performed by the members of the center in support of scientific applications and collaboration with other centers and institutes. This activity is also summarized in a matrix form enclosed at the end of this document. The plan is organized by technologies, and for each technology the tasks for each application project are described.

Workflow Technology tasks

- ***Application: Fusion (CPES) - UCD***
  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. SDM center main contacts: Bertram Ludaescher & Norbert Podhorszki.
  Monitoring and Archivng: 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. SDM center main contact: Scott Klasky.
- ***Application: Groundwater Modeling - PNNL***
  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 acontinuum simulation of flow and transport for two non-reacting tracers. SDM center main contact: George Chin.
- ***Application: Combustion (Jackie Chen) - ORNL***
  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 inifinite 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. On going work will be to push this system into the mainstream S3D runs during summer 2007. SDM center main contact: Scott Klasky.
- ***Application: Biology - PNNL***
  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. SDM center main contact: Terence Critchlow.

Metadata and Provenance tasks

- ***Application: Spallation Neutron Source - LBNL***
  The task is to develop a web-based data entry tool for samples used by the SNS. Contact people are Shelly Ren and Steve Miller. Over a period of several months, we used the Data Entry and Browsing Tool (DEB) to provide such web interfaces, based on schema definition provided by Shelly. The interfaces are now under evaluation by scientists. They may be used only as a prototype for special purpose interfaces that will be developed by the SNS team. SDM center main contact: Arie Shoshani.
- ***Application: Combustion (Jackie Chen) Utah***
  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. SDM center main contact: Steve Parker.

- ***Application: Astrophysics (TSI) and Fusion (CPES) – NCSU, ORNL, Utah***
  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) – NCSU, ORNL, Davis***
  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. SDM center main contact: Mladen Vouk.

Data Movement and Storage tasks

- ***Application: Combustion (Bell) - LBNL***
  The application needs to generate data on multiple sites (e.g. Jaguar at ORNL), tar, archive at the remote site and/or move to LBNL (e.g. on DaVinci), archive to HPSS at LBNL. The difficulty is with site that have OTP firewalls, such as ORNL. Based on this requirement a module was developed, called SRM-Lite which as be invoked at the remote site, and which can provide robust multi-file movement capability. SRM-Lite is currently being tested. SDM center main contact: Alex Sim.
- ***Application: Fusion (CPES) - LBNL***
  There are plans to 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. SDM center main contact: Arie Shoshani.
- ***Application: High Energy Physics - LBNL***
  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. SDM center main contact: Arie Shoshani.
- ***Collaboration: with Open Science Grid - LBNL***
  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. SDM center main contact: Alex Sim.
- ***Collaboration: with Earth System Grid - LBNL***
  LBNL has been providing SRM software as well as SRM-Lite for several years now. This work continues to evolve. This work is funded by the OSG. SDM center main contact: Alex Sim.
- ***Application: Combustion (Jackie Chen) - UCD***
  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). SDM center main contact: Norbert Podhorszki.

Indexing Technology tasks – LBNL

SDM center main contact: John Wu

- ***Application: Combustion (Jackie Chen)***
  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. We are exploring collaboration with Valerio Pascucci of imbedding the Fastbit technology into visualization tools for this application domain.

- ***Application: Fusion (CPES)***
  This task is in collaboration with Scott Klasky. 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***
  There is an agreement to deploy FastBit software for visualization applications. The Vis center will also provide expertise to help with analysis of AMR data.

Parallel I/O Technologies tasks – ANL and NWU

SDM center main contacts: Rob Ross and Alok Choudhari

- ***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: Combustion (Jackie Chen)***
  This group is interested in improving overall I/O performance. NWU has obtained an I/O kernel and is experimenting with approaches to storing simulation data in a canonical format that eliminates most postprocessing 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.
- ***Application: Cloud Modeling***
  This group is interested in using PnetCDF in their applications to improve I/O performance. We had a concall with the group, and they have subsequently begun experimenting with PnetCDF.
- ***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.

Feature Extraction tasks – LLNL

SDM center main contact: Chandrika Kamath.

- ***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.

- ***Application: Fusion (GPS)***
  The goal is tracking of blobs in a high-dimensional particle simulations.

High Performance Statistical Analysis task (ORNL)

SDM center main contact: Nagiza Samatova.

- ***Application: Combustion (Jackie Chen)***
  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: 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 (John Drake)***
  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: Climate (John Drake and George Ostrouchov)***
  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: Nanoscience (DOE CNMS Center) (Philip Rack)***
  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.
- ***Application: Biology (DOE Genomics:GTL projects) (R. Hettich, J. Banfield, C. Harwood, M. Buchanan)***
  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.
- ***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 (PNNL)

SDM center main contact: Jarek Nieplocha

- ***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.

Below is a table identifying SDM center technologies to be applied to different application projects. The information is summarized in a form of a colored matrix. The cells of the matrix have labels that indicate the problem or technology being applied, and a color coding as follows: Red - currently in progress, Orange - problem identified, Yellow - interest expressed. In this document only the Red and Orange cells are described in the bullets above. The matrix is intended as a quick summary of current and planned activities.

				SDM center Technologies				
Application Domains	Workflow Technology (Kepler)	Metadata and Provenance	Data Movement and Storage	Indexing (FastBit)	Parallel I/O (pNetCDF, etc.)	Parallel Statistics (pR, …)	Feature Extraction	Active Storage
Climate Modeling (Drake)	workflow				pNetCDF	pMatlab		
Astrophysics (Blondin)	data movement	dashboard						
Combustion (Jackie Chen)	data movement	distributed analysis	DataMover-Lite	flame front	Global Access	pMatlab	tranient events	
Combustion (Bell)			DataMover-Lite		MPIO-SRM-client			
Fusion (PPPL)							poincare plots blob tracking	
Fusion (CPES)	data-move, code-couple	Dashboard	DataMover-Lite	Toroidal meshes			structure tracking	
Materials - QBOX (Galli)					XML			
High Energy Physics	Lattice-QCD		SRM, DataMover	event finding				
Groundwater Modeling	identified 4-5 workflows							
Accelarator Science (Ryne)					MPIO-SRM			
SNS	workflow	Data Entry tool (DEB)						
Biology	scalaBlast					ProRata		
Climate Cloud modeling (Randall)			DataMover-Lite		pNetCDF			
Data-to-Model Coversion (Kotamathi)								
Biology (H2)								
Fusion (SWIM) (bachelor)				Future interest	Future interest			

Subsurface Modeling (Lichtner)					Future interest	Over AMR		
Flow with strong shocks (Lele)						conditional statistics		
Fusion (extended MHD) (Jardin)								
Centers & institutions								
Open Science Grid			SRM-tester					
Earth System Grid			SRM and DML					
Petascale Storage Institute (Gibson)					Posix-IO			
Vis Institute (Ma)				query-based vis				
Vis Center (Bethel)				query-based vis				
Legend		currently in progress						
		problem identified						
		interest expressed						