From DATA to

The SciDAC-funded Scientific Data Management (SDM) Center is creating novel tools for managing and mining extensive datasets, enabling exciting new possibilities for science and discovery. The three-layer management system developed by the center facilitates workflow, analyses, and data access in innovative ways.

"The three interrelated layers of the SDM Center play equally vital roles in accomplishing the primary goal—enabling scientists to use data effectively to advance science."

Dr. Arie Shoshani
Principal Investigator, SDM Center

Introduction

Experiments and computer simulations involved in today's cutting-edge research generate extremely large datasets, often many orders of magnitude greater than the datasets handled by previous generations of scientists. In addition to being extremely large, modern datasets are also immensely complex and can represent systems spanning multiple scales of length, time, and temperature. For example, the Solenoidal Tracker at the Relativistic Heavy Ion Collider (STAR; RHIC) experiment explores nuclear matter under extreme conditions (sidebar "STAR Experiment," p35), and can collect seventy million pixels of information one hundred times per second. In modeling atmospheric parameters, climate simulations tracking long-term variations may require up to eleven terabytes of storage for a one hundred year run. Similarly immense and complex datasets are being produced in numerous other research disciplines, such as biology, materials science, astrophysics, and fusion. Such extraordinarily rich data present researchers with unprecedented possibilities for scientific advancement.

Along with the empowering possibilities for science arrives a new set of problems: storing, managing, and mining very large quantities of data. Although the cost of hardware needed to store and move large amounts of data is continually decreasing, the time and effort required to access, process, analyze, and visualize data is rapidly increasing. How can scientists utilize these impressive amounts of data efficiently to maximize scientific advancement without being constrained by the details of data management? The SciDAC-funded Scientific Data Management (SDM) Center addresses this challenge with a number of innovative solutions for processing and managing the vast datasets generated by complex systems.

The SDM Center is one of the SciDAC Integrated Software Infrastructure Centers (ISIC) that provides the computer science and mathematical technologies critical for efficient use of computational systems by applications-oriented projects. By facilitating close relationships between computer scientists, mathematicians, and applications scientists, the SDM Center has achieved remarkable success in improving scientific productivity. Formally located at Lawrence Berkeley National Laboratory (LBNL), this cyber facility is a national collaborative effort consisting of scientists and area leaders positioned at major institutions across the country.

Activities at the SDM Center are organized into three layers (figure 1). Principal Investigator of the SDM Center, Dr. Arie Shoshani of LBNL asserts the three interrelated layers play equally vital roles in accomplishing the primary goal—enabling scientists to use data effectively to advance science. The top layer functions to help scientists orchestrate and automate the execution of data management tasks. The next layer of organization supplies components for efficiently searching, analyzing, and identifying useful features in data. The capabilities of these first two layers rely on practical storage and parallel access innovations at the third layer, the area of infrastructure closest to the hardware.

This three-layered structure emerged during the first phase of the SciDAC program. Close contact between computer scientists and applications scientists is one of the strong points of SciDAC. The SDM team is currently collaborating with scientists to develop systems that enable efficient and effective use of their data resources. While working with applications experts, SDM researchers recognized that scientists focus on science issues rather than on finding ways to manage the data. Dr. Shoshani emphasizes "scientists prefer to concentrate on their science."

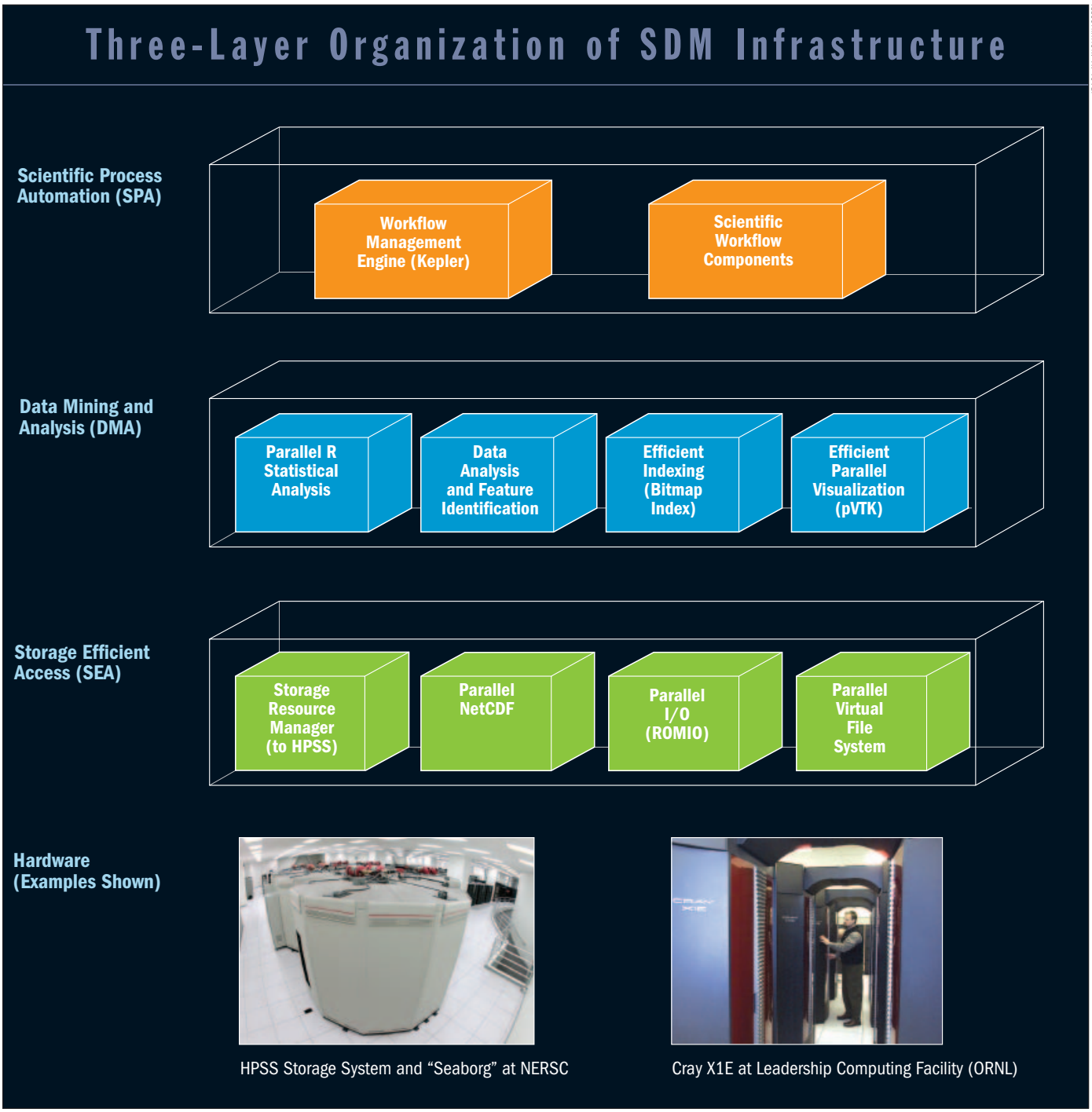


ILLUSTRATION: A. TOVEY

Figure 1. *To most effectively handle vast quantities of scientific data, the SDM Center organizes activities into a three-layered infrastructure. The top layer, Scientific Process Automation (SPA), directly assists users by planning and executing practical workflows. An intermediate layer, Data Mining and Analysis (DMA), provides innovative tools for searching, analyzing, and presenting data. The base layer, Storage Efficient Access (SEA), improves the efficiency of other programs by solving problems inherent to parallel storage and access of large datasets.*

"The goal is to remove the tedium of computational and data management tasks from scientists and put the burden back on computers."

Dr. Terence Critchlow
LLNL

Scientific Process Automation—Channeling the Flow

In the first phase of the SciDAC program, SDM teams explored a variety of strategies for moving information efficiently. It soon became apparent that managing the flow of information among tools, a slow and laborious task, was a major barrier to productivity. The Scientific Process Automation (SPA) layer of the SDM structure now concentrates explicitly on such workflow management issues.

Working on projects in various disciplines helps SDM researchers understand what scientists need from a workflow scheme. "The goal is to remove the tedium of computational and data management tasks from scientists and put the burden back on computers, where it belongs, thereby freeing scientists to focus on science rather than computational process," says the area leader for SPA activities in the SDM Center, Dr. Terence Critchlow of Lawrence Livermore National Laboratory (LLNL). In an early example of developing scientific workflows, the SDM Center worked with Dr. Matthew Coleman, a biologist at LLNL. SDM researchers constructed a workflow process (see figure 2 and sidebar "Mining Biology Databases," p31) to automate the tedious process of extracting and compiling information from various databases. As Dr. Coleman points out, "If I were to do this by hand, it would take an extraordinary amount of time." Instead, he now has more time to focus on the scientific value of the data.

Many SPA activities utilize Kepler, a scientific-workflow system that coordinates the execution of various software components. At times, SDM researchers generate these components, but the system is flexible enough to accommodate external applications, such as specialized programs developed by project scientists. Kepler allows scientists to manipulate the workflow through a graphical user interface.

An ideal workflow system would allow scientists to combine the components at will, with the system coordinating all transfers of information among the components. Achieving such end-to-end capability requires clearly defined and controlled interactions among components. Building on the earlier Ptolemy system from the University of California, the Kepler system uses "actors" to manage components. This approach provides users with a uniform interface for directing workflow, and is an important step towards achieving an entirely modular system of interchangeable components. A robust workflow management system also requires the ability to monitor ongoing calculations and document the provenance, a detailed history of every bit of data. As Dr. Critchlow explains, "knowing where things came from, and being able to find something that you did two years ago is a very challenging and important problem," especially for the sort of flexible, reconfigurable workflow demanded by oft-refined analyses.

Dr. Shoshani understands that setting up workflows for cutting-edge computational tools entails a great deal of expertise. According to Dr. Shoshani, "What you hope is that you work with one group of scientists, and they will show it to their friends." In practice the most successful projects involve assigning someone from the SDM Center to work with scientists on developing the workflow. "It's sufficiently complex that you need to have expertise to set it up properly," Shoshani says. But when such commitment is made, "it saves people a tremendous amount of time." Consequently, the SPA team consists of institutions that perform workflow development at the San Diego Super Computer Center and University of Utah, as well as institutions that interact with application scientists at North Carolina State University (astrophysics), University of California–Davis (fusion), and LLNL (biology and astrophysics).

Although Kepler was developed initially for biological applications, it is a nonspecific system currently used by other application domains, including astrophysics and fusion. Such applications require coordination of code execution and data movement. For example, in fusion plasma simulation research, a single simulation program capable of covering the entire toroidal volume of a reactor does not exist. Thus, two different codes are often used. A magnetohydrodynamic (MHD) code uses a fluid-like description of ion motion to efficiently simulate the behavior at points deep within the plasma. For tracking individual particles at the periphery of the plasma, a gyrokinetic code is utilized. The suitability of individual code usage depends on the dynamic properties of the solutions. Kepler's automated workflow execution system can manage the shifting between the two different simulation engines. A large volume of data describing the plasma state must be exchanged from one program to another, all during the active simulation process. Coordinating this type of data transfer would be extremely time consuming and error-prone if it were executed manually.

Data Mining and Analysis—Gleaning Meaning

Analyzing the data generated by high-throughput tools is a critical component of modern scientific exploration. Important information must be extracted from vast quantities of raw data. However, human beings can only deal with a tiny fraction of the data generated by modern simulation and experimental methods.

Mining Biology Databases

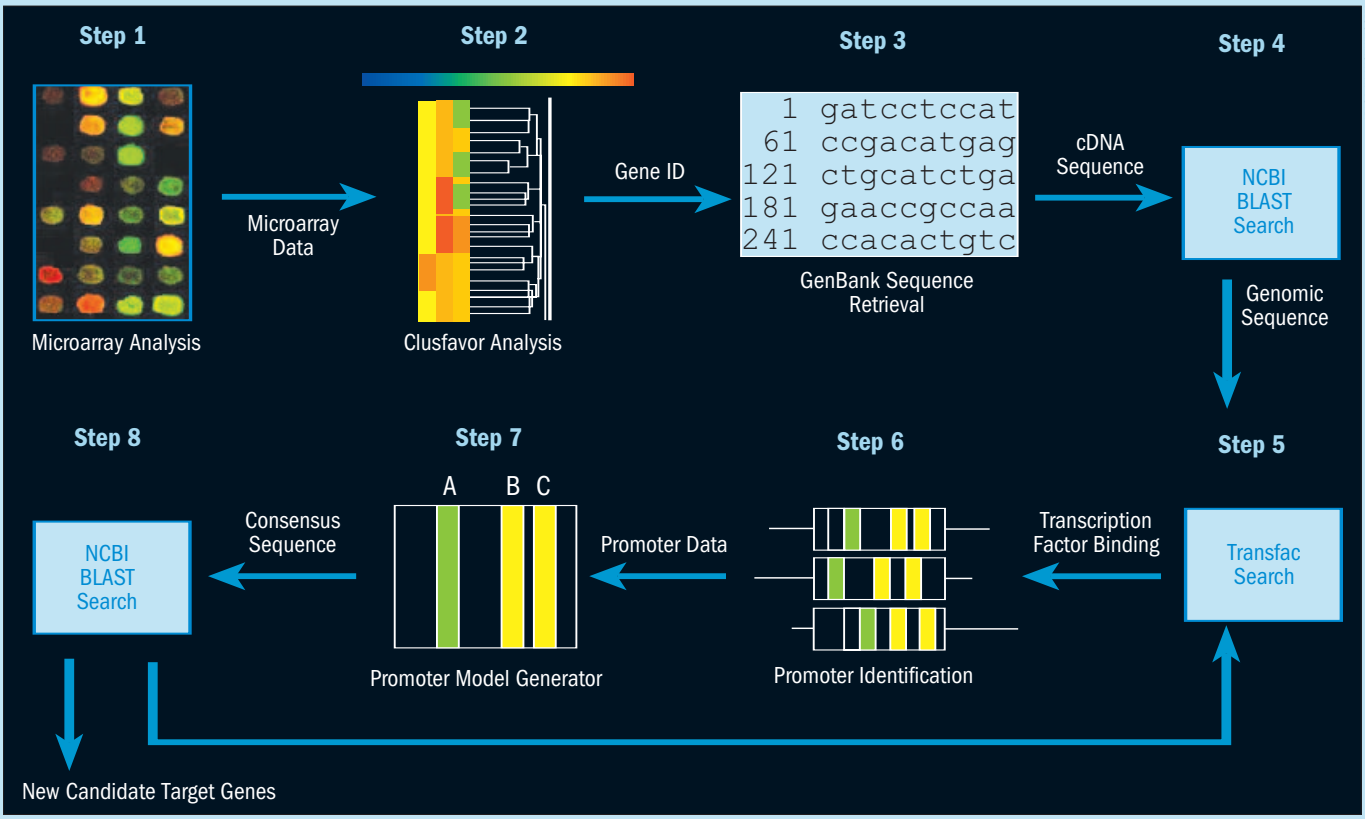


Figure 2. *This diagram depicts the flow of steps involved in using gene sequences to identify shared regulatory regions. Microarrays (step 1) containing the genes of interest are used to measure differential gene expression in cells exposed to ionizing radiation. Gene expression changes are then grouped using computational tools, and the generated clusters are then parsed to identify interesting groups of genes (step 2). Each gene in the cluster is then linked with a sequence database such as GenBank (step 3) to identify its specific cDNA sequences. The cDNA sequence is then compared against genomic sequence databases using a BLAST search (step 4) to identify the predicted start of transcription. To characterize the basal promoter regulatory elements the sequences are then analyzed using multiple databases (step 5). Once found, regulatory profiles are then compared across each gene in the cluster (step 6) to delineate common regulatory elements that can be used to generate a novel promoter consensus sequence. This consensus sequence becomes a model (step 7) to be used to search multiple sequence databases to relate other candidate genes relevant to the study. The Kepler workflow management system now manages this process. Many of the steps involve performing remote, web-based operations.*

Helping scientists manage workflow is a growing emphasis of the SDM program. For example, Dr. Matt Coleman of LLNL teamed with SDM researchers to manipulate data from various remote biology databases. Dataset size is small relative to advanced simulation data, but the process exemplifies how automating data management workflow allows scientists to focus on science.

Dr. Coleman's research explores relationships between DNA damage, low doses of ionizing radiation, and associated health effects. Is there a threshold radiation dose below which health risks are negligible? To explore this controversial question, microarrays are used to monitor tens of thousands of genes simultaneously. Microarrays indicate how strongly each gene is expressed, or transcribed into RNA that will be translated into protein. Monitoring expression for two cell lines at fourteen radiation doses yields a "massive amount of data," says Dr. Coleman.

As expression data are collected for various levels of radiation, genes are clustered with other genes showing similar expression patterns. Co-regulation is indicated when transcription levels for a group of genes rise or fall together. This implies that expression may be governed by common transcription factors, molecules that control expression by binding to DNA regions near the gene in question. These binding regions, called promoter sequences, are often short, recognizable sequences of DNA base pairs. Many promoter sequences are known and available from Internet databases. Some sequences from databases can aid in identifying new promoter sequences, but others are repetitive and so common that including them in the analysis could overwhelm the informative signal with unwanted noise.

Databases use different interfaces, so accessing and combining information can be tedious and time consuming. SDM scientists worked with Dr. Coleman to devise a workflow application for automating the many steps of this process (figure 2). Tasks that once required hours or days can now be completed in minutes, allowing the biologists to spend their saved time on the science of characterizing new promoter regions.

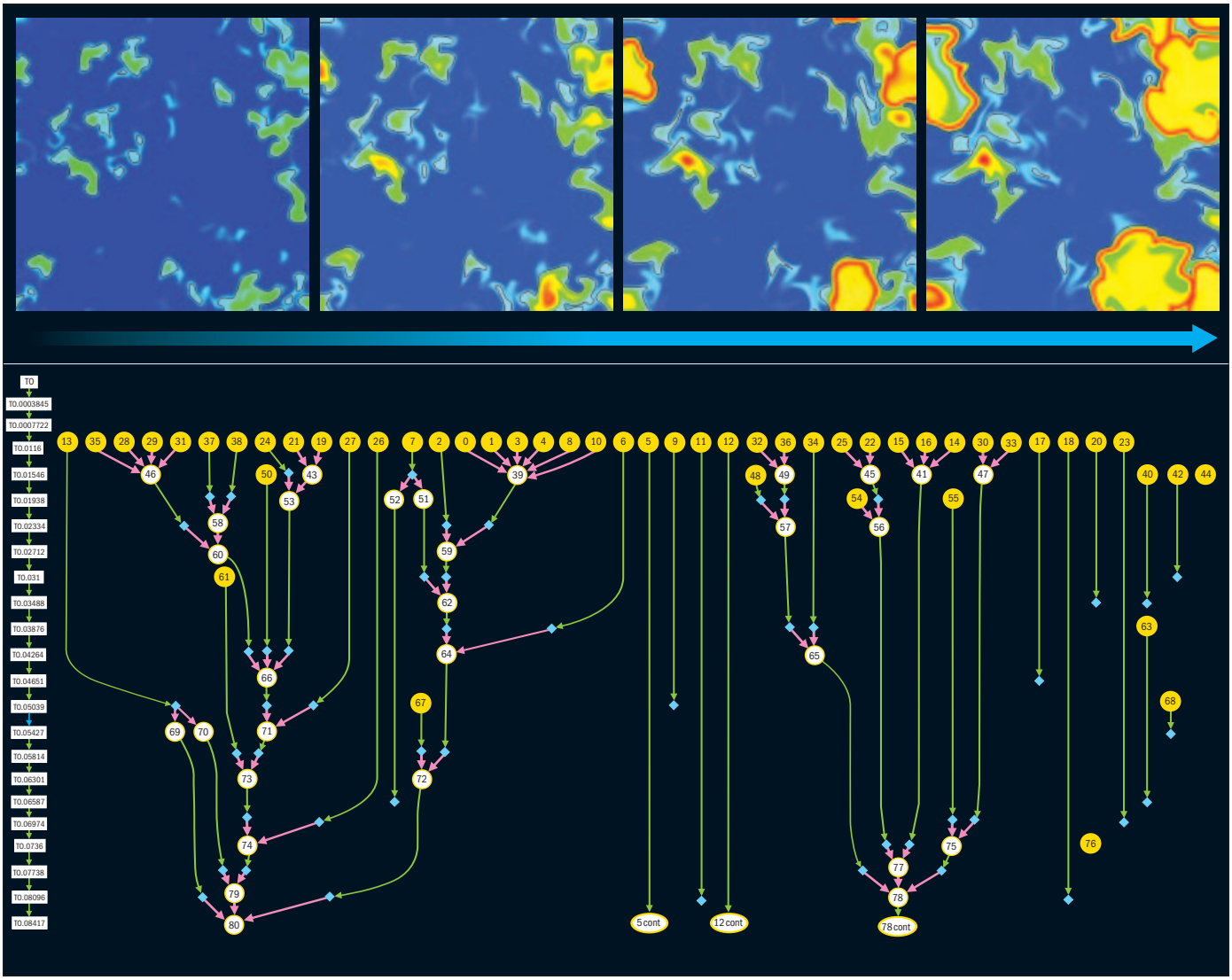


ILLUSTRATION: A. TOVEY

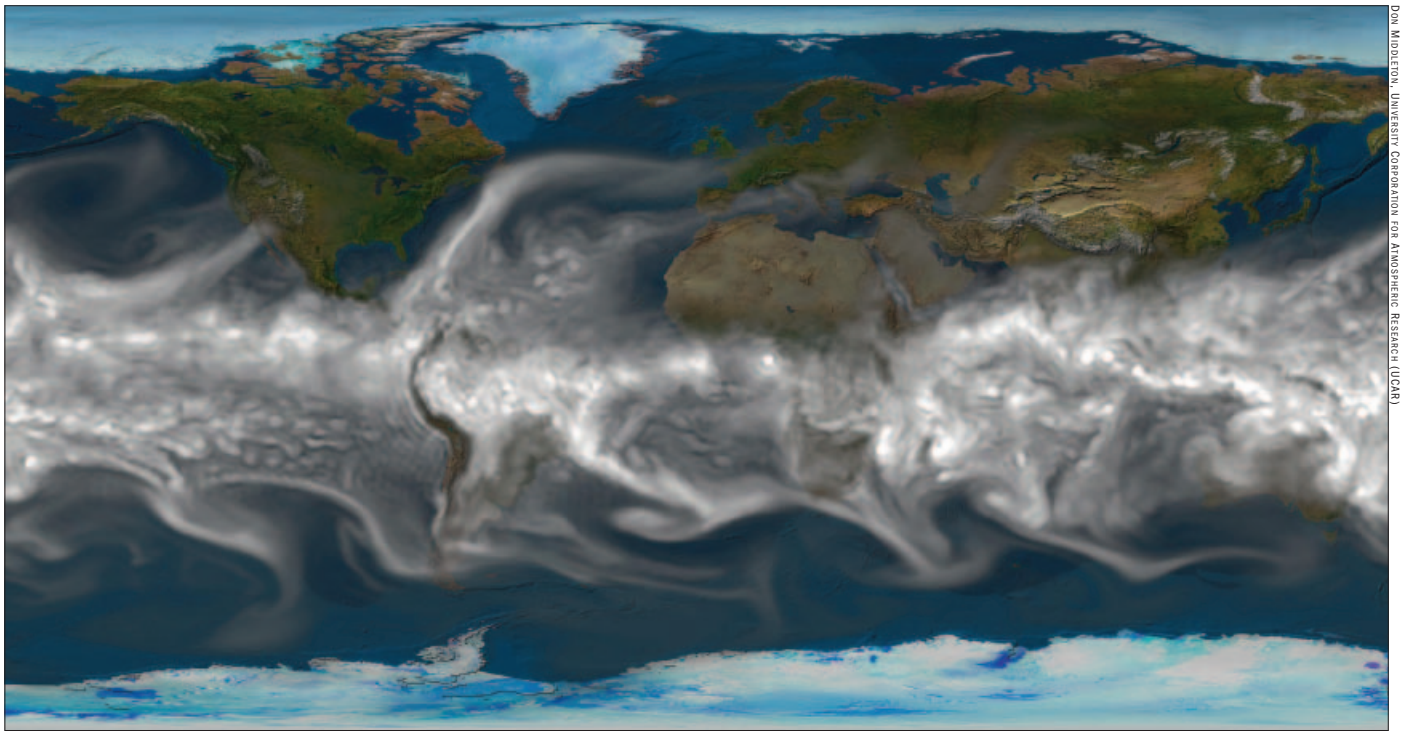
Figure 3. *A combustion simulation is performed as a series of time steps that indicate the progression of flame fronts. Each frame at the top of the figure is a single time step, progressing in time from left to right, as indicated by the arrow. Each flame front starts from a "kernel" and can combine with other fronts. FastBit is used to identify various regions efficiently and find overlaps over time steps. At the bottom of the figure is a graphical representation of the progression of flame fronts over time. Here, time progresses from the top to the bottom of the diagram. Each flame front is assigned a number that is retained over time, and shown with vertical lines. Flame fronts that combine are shown as new nodes, and assigned a new number. This graphical representation is a very compact way of visually representing the progression of a combustion simulation.*

Activities in the intermediate SDM layer, Data Mining and Analysis (DMA), help scientists find meaningful subsets of data for display and further analysis. By applying analysis techniques to datasets, DMA researchers are able to enhance the interpretation of data, the design of successive experiments, and comparisons between simulation and experimental datasets. As an additional benefit, DMA activities often reduce the amount of data scientists are required to handle. DMA tools can further increase efficiency by monitoring simulations and conducting preliminary analyses on partial datasets while computations are still running. This allows scientists to adjust parameters or abort unproductive calculations.

Searching and Indexing

Searching results for regions that satisfy particular criteria recurs frequently in computational science. In a combustion simulation, for example, a flame will continue to propagate if the temperature and reactant concentrations are sufficiently high. Combustion researchers look for regions of neighboring points that meet those conditions. Defining the combustion front, the boundary separating regions of reactants and products, plays an important role in combustion modeling. By characterizing the position, shape, and evolution of fronts, researchers access information that is simpler and more powerful than large amounts of raw data.

The FastBit indexing algorithm, developed by SDM researchers at LBNL, allows efficient



Don Middleton, University Corporation for Atmospheric Research (UCAR)

Figure 4. *Provided by Don Middleton, National Center for Atmospheric Research (NCAR)—one of the PIs of the Earth System Grid (ESG), a SciDAC-supported project—this data visualization provides a dramatic example of the capabilities needed to analyze such large datasets. ESG, which uses SDM technology, provides wide-area access to large-scale climate simulation data from multiple storage systems at various sites.*

searches of large databases for fronts and similar features. Somewhat like an Internet search engine, the algorithm works by cataloguing an existing set of data and creating an index that quickly identifies spatial points with specified attributes (sidebar "FastBit," p34). Originally developed to catalog collision events in high-energy physics experiments, the FastBit software has since been applied to a variety of problems, including visualization of astrophysics simulations, analyses of DNA sequence data, and systems for detecting network intrusion.

The SDM team has collaborated with combustion researchers, Dr. Jacqueline Chen and Wendy Doyle at Sandia National Laboratories (SNL), and Dr. Tarek Echekki at North Carolina State University. This group worked to analyze high-fidelity simulation data from turbulent autoignition processes in which the initial temperature and reactants are non-uniformly distributed. These simulations provide key underlying insights required to develop predictive models used to design novel engine concepts, including homogeneous charge-compression ignition (HCCI). In a combustion simulation, concentrated regions of chemical activity known as ignition kernels play an important role in the modeling of autoignition and subsequent flame propagation. This research aims to exploit the spontaneous combustion of a turbulently mixed, compressed fuel-air mixture and promises the efficiency of diesel engines without the associated polluting particles and emissions. The researchers used FastBit to identify and track combustion fronts and regions of autoignition (figure 3, upper aspect), thus testing their modeling of the process. They learned that regions of interest are found directly from the index. Once the ignition kernels and subsequent flame fronts are identified, they can be tracked over time and represented as a graph showing a temporal progression (figure 3, lower aspect). Combustion researchers ultimately hope to use these findings to design engines that improve the exploitation of HCCI.

Feature Identification

Indexing can quickly identify regions in physical space that have specific properties. Researchers are often looking for phenomena that can only be recognized as complex spatial or temporal patterns, and in many cases, patterns corresponding to a particular behavior are not known in advance. When the volume of data increases, performing efficient pattern analysis is a daunting challenge. For example, large-scale climate simulations (such as those illustrated in figure 4) require increasingly finer meshes in order to achieve accurate and detailed results. Current simulations that use a resolution of 280 km per dimension of the mesh generate about 0.75 terabytes for a 100-year simulation run. A high-resolution dataset will have a 70 km resolution

Often scientists cannot recognize characteristic features beforehand, so a visual representation of the entire dataset is needed to identify potentially important patterns.

FastBit: Indexing for Fast Searches

In data mining and analyses, the process of quickly isolating important information from much larger pools of data is critical. FastBit is a software package capable of extremely fast searches of large databases. During a series of head-to-head trials, FastBit considerably outperformed a leading database management system. Searching a dataset composed of 250,000 email messages, FastBit handled queries between ten and one thousand times faster than the popular commercial software.

Bitmaps are sequences of bits, basic yes/no units of information represented by 1 or 0, a computationally practical representation. A bitmap index is a set of bit sequences that represent information about certain indexed attributes. Because FastBit uses bitmap indices, user queries can be addressed by bitwise logical operations, which computer hardware systems generally handle quite efficiently. However, scientific applications often involve indices containing information about a large number of bitmaps, and such bitmap indices demand impractical storage requirements. Schemes that compress index files can reduce space requirements, but compression can also slow down search methods. To maximize FastBit performance, researchers had to optimize this tradeoff between storage space and speed. Using the Word-Aligned Hybrid (WAH) compression method, FastBit achieves this functional balance. Bitmap indices compressed by the WAH scheme are a little larger than indices compressed by other methods, but WAH-compressed indices can be queried much faster because they can be searched without being fully decompressed.

SDM researchers at LBNL developed both the FastBit software package and the WAH compression scheme it employs. A number of SciDAC-supported projects, including the STAR experiment (sidebar, p35) and combustion research (figure 3, p32), have benefited from the impressive power of FastBit.

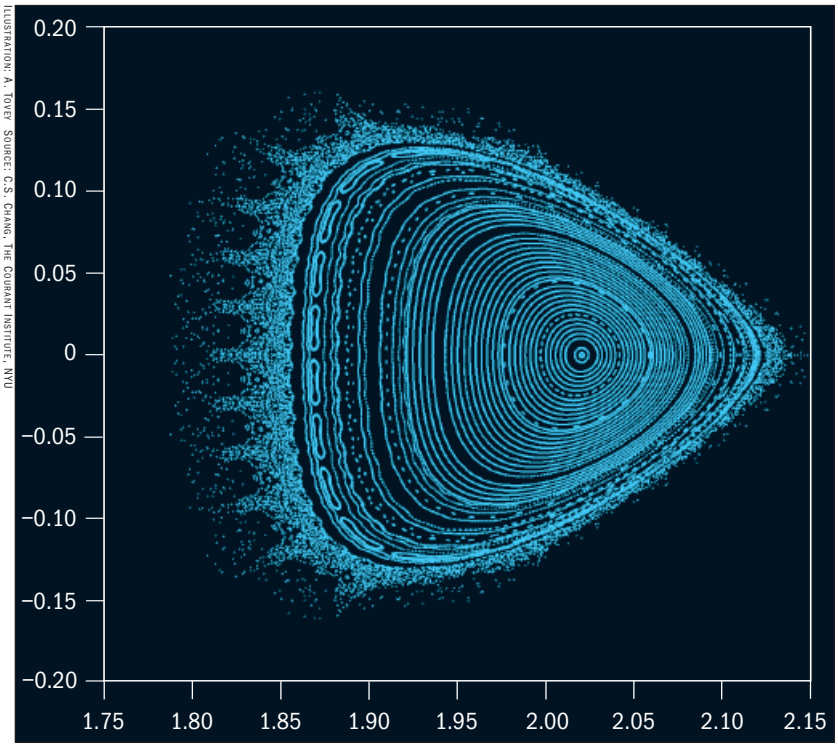


ILLUSTRATION: A. TOVEY SOURCE: C.S. CHANG, THE COURANT INSTITUTE, NYU

Figure 5. *The plot of orbits in a cross-section of a fusion experiment shows different types of orbits, including both circle-like "quasi-periodic orbits" and "island orbits." Note the deformation in some of the circular orbits, indicating the presence of other island orbits nearby.*

"We don't require users to know anything about parallel processes."

DR. NAGIZA SAMATOVA
ORNL

and will require 11 terabytes for a 100-year run. The ability to analyze such data volumes requires not only efficient indexing for the selection of desired data subsets, but also parallel methods for data analysis.

Early in the SciDAC program, researchers looking at decades of climate data wanted to separate the effects of volcanoes and El Niño from their simulations. SDM Center members from LLNL used a combination of principal components analysis (PCA) and independent components analysis (ICA) to identify the characteristic pattern of global temperature changes associated with each of these driving forces. Having identified a spatial fingerprint for each of these features, they could compare the results from different simulations in a more meaningful manner.

A more recent example of feature identification involves orbits of particles around fusion plasma (figure 5). For example, some orbits can transect a particular cross-section of the reactor anywhere in a continuous circle-like shape, while other orbits, called islands, are confined to crossing in only a portion of the circle. The positions in which one type of orbit changes to the other are important for plasma confinement. Because both experiments and simulations report individual crossing points rather than a continuous curve, identifying the orbit topology is challenging. LLNL scientists have developed analysis techniques to classify the orbits automatically.

Large-Scale Statistical Analysis

Often scientists cannot recognize characteristic features beforehand, so a visual representation of the entire dataset is needed to identify potentially important patterns. By incorporating visualization of ongoing simulations into the workflow, researchers can prevent wasted time and effort by isolating problems with starting conditions or system evolution before long computer runs are completed.

Creating an image of data can involve complex processing and analysis, and using various combinations of data management methods can take unexpected turns. Dr. Nagiza Samatova of Oak

STAR Experiment: Finding One Event Out of a Million

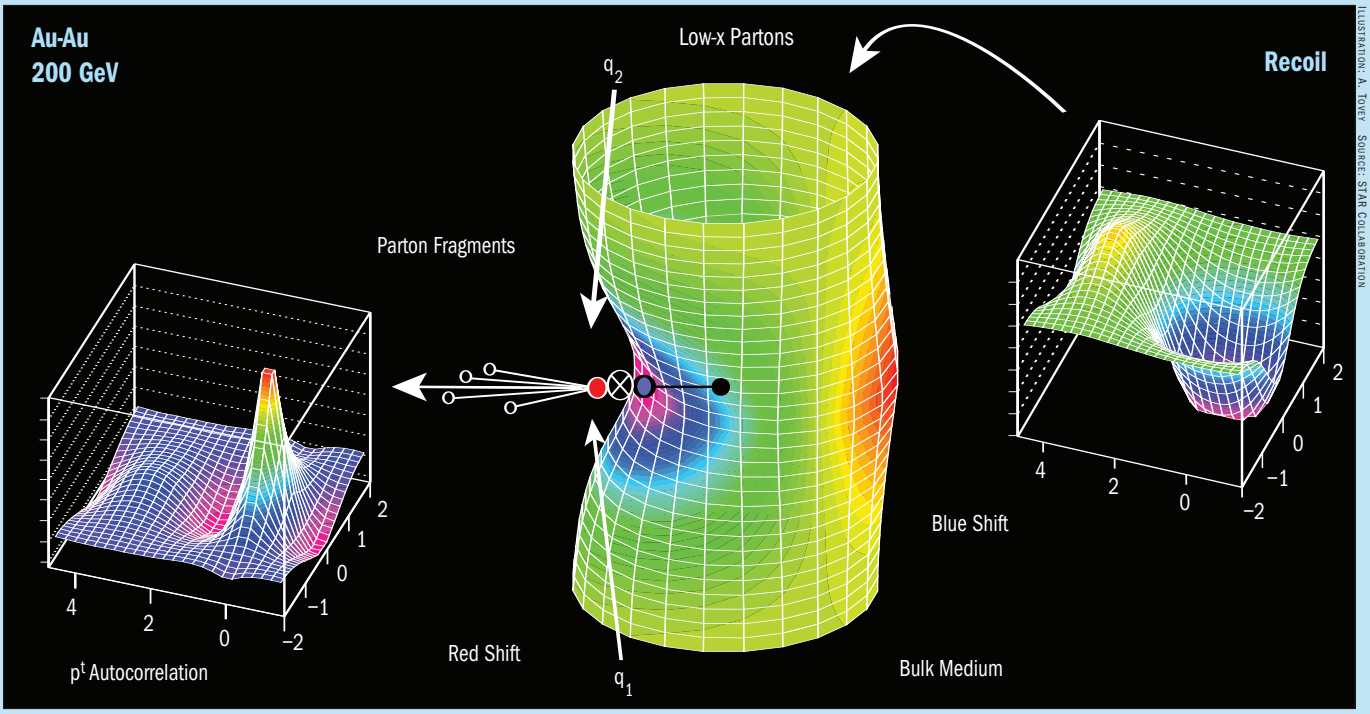


ILLUSTRATION: A. TOVEY SOURCE: STAR COLLABORATION

Figure 6. *Using sophisticated computer analysis to combine the results of hundreds of millions of collision events, physicists constructed this detailed picture of the angular correlations of quarks and gluons interacting with the colored medium produced in RHIC heavy ion collisions. Novel event-wise correlation techniques were utilized in this process.*

The Solenoidal Tracker at Relativistic Heavy Ion Collider (STAR; RHIC) experiment is a collaboration of over five hundred scientists in thirteen countries. Located at Brookhaven National Laboratory (BNL), RHIC smashes together large nuclei, like those of gold atoms, to recreate the hot and dense conditions hypothesized to exist when the universe was less than one second old. For every event, the STAR records the responses of detectors arranged around the collision point. By comparing the responses, physicists infer the type and motion of particles released from the collision to better understand possible states of deconfined quarks and gluons.

In order to efficiently mine the hundreds of terabytes of data generated by the STAR detector, physicists have worked with the Scientific Data Management (SDM) Center to develop a system called Grid Collector. The Grid Collector employs two critical techniques for mining data effectively—Storage Resource Managers (SRM) and FastBit indexing software (sidebar "FastBit," p34). These innovations allow STAR researchers easy access to any subset of data they need.

The FastBit indexing software accelerates searches of all types of databases, including those from massive simulations. Physicists can now rapidly identify events that meet certain criteria, such as particles within a specific range of energies. Rather than extracting every data point and eliminating the large irrelevant fraction, FastBit speeds the task by focusing the analysis on events exhibiting specific properties. For example, one analysis found that only eighty events out of hundreds of millions were relevant for inclusion in more detailed analyses.

Now SDM technology performs data management tasks while researchers concentrate on the exciting science of the RHIC program. With the SDM system, STAR scientists extract data of interest within fifteen minutes. Before the powerful SciDAC-developed tools, accomplishing the same task consumed weeks. "Such achievement is to me a true incarnation of the power of collaborative work between physicists and computer scientists," says Dr. Jerome Lauret, architect of STAR's distributed computing model. "When the best of both worlds meet, the intersection is a tool such as the Grid Collector and from there the implications (and applications) are endless."

"We are performing high-energy nuclear physics experiments of unprecedented complexity and precision which generate enormous data volumes having a very high richness factor. These experiments, which require a level of data mining and resource management previously unimaginable, have already pointed to the existence of a new state of matter which probably last existed near the beginning of the universe more than fifteen billion years ago. Understanding the properties of this new matter in detail will require yet an additional order of magnitude greater capability, so that far larger data sets can be searched for extremely rare probes. This simply will not be possible without continued breakthroughs such as those now coming on the market from ongoing SciDAC developments."

TIM HALLMAN, STAR EXPERIMENT SPOKESPERSON

Scaling Computational Science Parallel Input/Output

Generating, storing, managing, and discovering knowledge from petascale-level data is a daunting task. At least three measures are very important in this context, namely performance, productivity, and portability (P^3). Given that storage and manipulation of data involves disk systems, which are slower by several orders of magnitude as compared to processors, obtaining reasonable performance requires the use of large-scale optimized parallel input/output (I/O) using scalable software. Consequently, the productivity of computational tasks is tremendously reduced if large-scale storage systems that can support efficient storage and retrieval of data are not available. Moreover, since many scientists share data on different machines using different applications, the portability of data and data formats becomes critical for discovery and collaboration. However, scientists tend to use high-level formats for describing, storing, and manipulating data, and leave it to tool developers to provide capabilities that satisfy the P^3 measures. For example, in the climate and weather modeling applications, network Common Data Format (netCDF) is a standard format that describes multidimensional data in a form that contains meta-data in self-describing format. Achieving the P^3 measures have presented challenging requirements for software and tool developers for supporting petascale datasets.

Members of the SDM Center at Northwestern University, in collaboration with Argonne National Laboratory (ANL), have developed software architecture and its implementation, using several layers, that attempts to satisfy all three measures. Figure 7 describes this architecture. The highest layer (Parallel-netCDF) has semantic information associated with it, which aims to capture the intent of a user. For example, the user may have represented a three-dimensional structure with a specific distribution of data, each element of which may be a collection of variables such as temperature, pressure, density, and humidity. In our design, relationships and semantic information are preserved at the middleware level to enable optimizations while preserving portability on different systems. In order to obtain the best optimizations, the lowest level of software, the parallel file system, uses the information transferred through the layers for physical access optimizations.

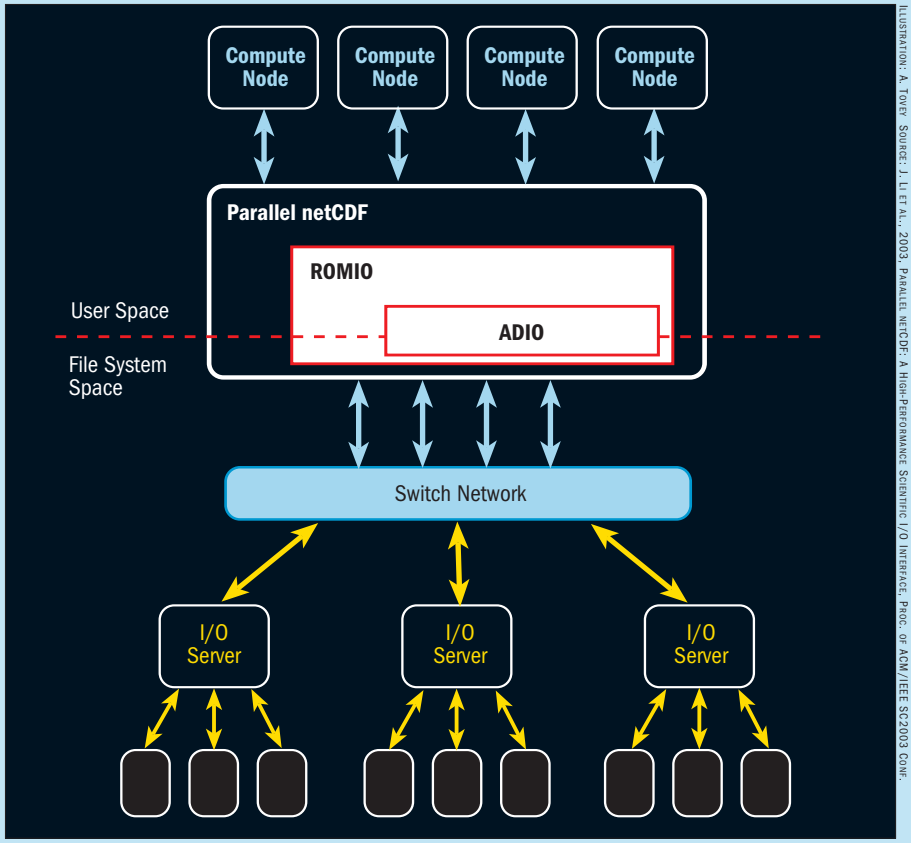


ILLUSTRATION: A. TOVEY. SOURCE: J. LI ET AL., 2003, PARALLEL NETCDF: A HIGH-PERFORMANCE SCIENTIFIC I/O INTERFACE, PROC. OF ACM/IEEE SC2003 CONF.

Figure 7. *The multi-layer design for the support of Parallel-netCDF libraries. The libraries are built on top of an MPI-IO implementation called ROMIO, which is in turn built on top of an Abstract Device Interface for I/O (ADIO) system, used to access a parallel storage system.*

The SDM Center has developed a library, called PnetCDF, where all three layers of software fulfill the principles described above. The previous implementations had to serialize I/O accesses for maintaining data consistency, which resulted in poor performance. Our implementation used the semantic information, collective I/O optimizations, and parallel accesses to significantly improve and scale I/O performance while maintaining consistency. Scientists in several fields including climate, weather, and astrophysics use the PnetCDF library. For example, in the FLASH astrophysics application, used at the University of Chicago and ANL, the PnetCDF library software improved performance by an order of magnitude above what was previously achieved for parallel I/O.

"The applications want to see an interface that looks very simple. The challenging part for us is to hide all those components."

DR. ROB ROSS
ANL

Ridge National Laboratory (ORNL) is area leader of DMA activity at the SDM Center. She stresses that "early on, the idea was to integrate the analytic system with visualization." Soon, however, "we realized that the major bottleneck was the ability to analyze very large data sets." In response, Samatova's team created a version of the popular, open-source statistics package "R" that is optimized for use in parallel computing environments. Although so-called "embarrassingly parallel" problems can be run trivially on parallel machines, typical problems require greater optimization efforts, testing which parts of the calculation depend on other parts, and unrolling repetitive loops in the calculation so the steps can be performed in parallel.

Because each applications group possesses extensive legacy code, the DMA team has pro-

vided "hooks" that allow different client programs to use their "parallel-R" routines for performing these checks and scheduling the calculations. The team has already produced a version for the MATLAB environment used in climate research, and they hope to produce one for the IDL environment that some fusion researchers prefer. They also applied tools for analyzing gigabytes of mass-spectroscopy data to identify proteins that contribute to bacterial hydrogen production. In each case, the scientists invoke the parallel version using the same commands they already know. "We don't require users to know anything about parallel processes," Dr. Samatova says.

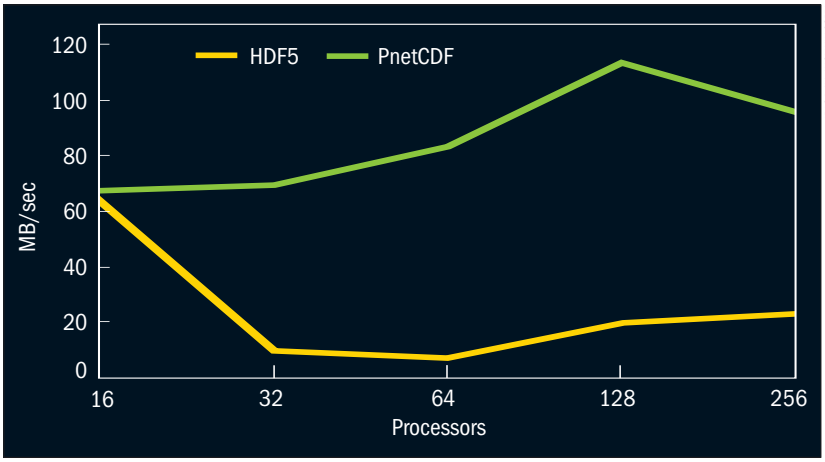


ILLUSTRATION: A. TOVEY

Figure 8. *Algorithms that explicitly address the fact that resources are distributed can speed the transfer of data dramatically. In this example, the rate of data transfer using the Hierarchical Data Format (HDF5) decreases when a particular problem is divided among more processors. In contrast the parallel version of netCDF developed by SDM researchers improves due to the low-overhead nature of PnetCDF and its tight coupling to MPI-IO (sidebar, p36).*

Storage Efficient Access—Under the Hood

Orchestration of data transfers and speedy analyses depend on efficient systems for storage, access, and movement of data among modules. Improving these tasks is the goal of the layer at the base of the SDM structure, Storage Efficient Access (SEA), the realm closest to the hardware. The innovations of SDM researchers at this layer are often invisible to the end user, but they are critical to exploiting the power that supercomputers bring to science. A main function of SEA is retrieving and saving data in parallel from the many sources where data are generated or stored. Doing so efficiently involves completing this task without burdening the programs using the data with the details of the SEA operation. The SEA area leader Dr. Rob Ross, of Argonne National Laboratory (ANL), states "the applications want to see an interface that looks very simple." Moreover, he adds that "the challenging part for us is to hide all those components" that perform sophisticated parallel access while still storing and retrieving data quickly.

Although the detailed implementations of the SEA are complex, the payoff is immense. One example is Parallel-netCDF (network Common Data Format; PnetCDF), a sophisticated data access tool developed at Northwestern University and ANL (see sidebar "Scaling Computational Science Parallel Input/Output," p36). In operations involving many processors, Parallel-netCDF reads and writes data more than ten times faster than a traditional system (figure 8).

Other tools developed by SDM researchers provide transparent, uniform access to files that may be distributed across many local and remote storage systems. This "Storage Resource Manager" (SRM) technology has been used extensively for accessing mass storage systems, such as those that hold distributed climate simulation data in the Earth System Grid SciDAC project, and distributed experimental data in the STAR High Energy Nuclear Physics Experiment (see sidebar "STAR Experiment," p35). Relieving applications scientists of these burdens allows them to be more productive at actual science.

These SEA programs shield other applications from the details of how and where data are stored. Yet, some programs will function better when the storage system is provided with some information about the stored data. The storage layer can utilize information to provide a uniformly structured way of accessing data, thereby increasing the efficiency of the programs using the data. Because scientific domains use different file formats—such as netCDF or Hierarchical Data Format 5 (HDF5)—the SDM Center will continue to improve parallel storage and access technology to comply with the broad range of file structures used by the scientific community.

"We knew something was there but without being able to interact with the data we could not really explore the physics."

Dr. John Blondin
North Carolina State University

Summary

Much of the day-to-day work of SDM researchers involves painstaking optimization of data-handling programs and the interactions among these individual applications. When these efforts are successful, they are largely invisible to users. From workflow management to data analysis and efficient storage access, close cooperation among the scientists improving the computer systems and those studying scientific phenomena enables both parties to better understand and respond to each other's needs. Such cooperation is one of the great strengths of the SciDAC program, in general.

Today, specialized research centers and the scientists who utilize them are often separated by geographic distance, presenting an inherent difficulty with data management. Dr. John Blondin, an astrophysicist at North Carolina State University, explains that in the early days of the SciDAC program development, geographic distance hindered

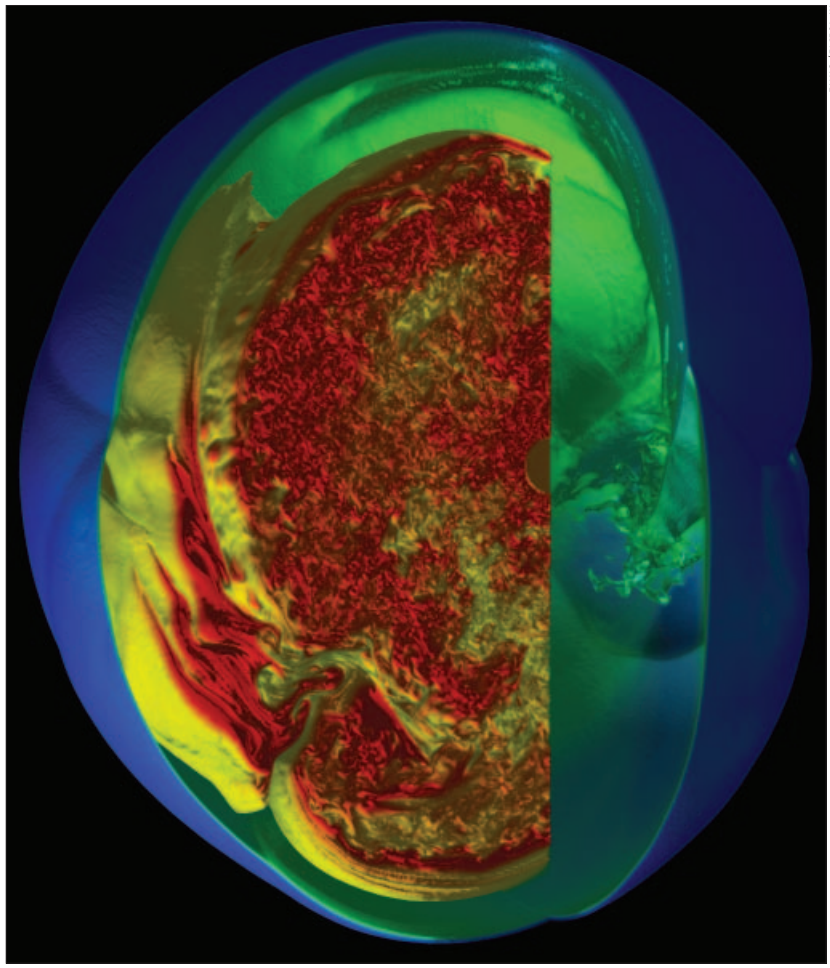
the effectiveness of some calculations. Dr. Blondin and his colleagues on the East Coast were conducting exciting simulations of supernova collapse remotely at NERSC in California ("Modeling the First Instants of a Star's Death," *SciDAC Review*, Spring 2006, p26). As a result, critical scientific information was hidden within the tremendous datasets stored on machines across the continent. "It was literally trying to do science in the dark," he recalled. New software from the SDM Center facilitated efficient transfer and analysis of the data, helping Dr. Blondin's team realize that angular momentum indicates asymmetry as a critical component for modeling supernova collapse (figure 9). "We knew something was there but without being able to interact with the data we could not really explore the physics." As soon as they got the software to move the data, Blondin says, the researchers could "really start understanding what was going on."

The researchers at the SDM Center continue to work closely with applications scientists to identify current and future data management needs and ease the burdens of large-scale data handling. Figure 10 shows some of the exciting science with which the SDM Center is connected. SDM scientists hope to expand participation in scientific discovery to other frontier areas of science.

Writer: Don Monroe, Ph.D.

Further Reading

SDM Center
http://sdmcenter.lbl.gov/



R. Toedte, ORNL

Figure 9. *Astrophysics researchers have recently discovered that modeling the asymmetry of supernova collapse is critical to understanding the process. Full three-dimensional modeling, made possible by SDM Center technology, facilitated this advancement in knowledge.*

Dr. Arie Shoshani on Future Plans

The technology developed by the SDM Center is of a general nature and can be applied to multiple application areas. Building on our early successes, we plan to improve the SDM framework to address the requirements of petascale science. To provide end-to-end support for the data management needs of scientists, we will incorporate appropriate technologies into the scientific workflow framework. Innovations, including both externally- and SDM Center-developed technology, such as new visualization tools and fast wide-area data transfer methods, will be incorporated into the SDM infrastructure. Specific advances are planned for each of the three areas.

In the SEA area we plan to extend the efficiency of parallel-IO and MPI-IO through advances in the areas of collaborative caching, efficient coordination, extended attributes, and collective input/output (I/O). We will also work to develop a common "bridge" API that many applications can use to abstract away the details of specific underlying storage (such as HDF5 and PnetCDF) and allow for eventual migration to new formats. Additionally, we intend to provide transparent access to archived and remote data using Storage Resource Managers (SRM) in collaboration with the SRM Center.

In the DMA area we plan to enhance and parallelize the specialized indexing technology, FastBit, to support important new data structures, such as Adaptive Mesh Refinement (AMR) structures. We also want to expand and apply the feature extraction software to new applications areas (notably combustion and fusion) and package the software for use in workflows. We will also develop the "Parallel Scientific Data Analysis Library" technology to allow various data analysis modules (such as R, MATLAB, and IDL) to be more easily parallelized and integrated into a generic data analysis framework.

In the SPA area we plan to provide systematic support for monitoring and debugging of large-scale workflow processes. We aim to enhance the Kepler system to provide fault tolerance and graceful handling of failures. Also, we hope to improve scientists' ability to interact with workflows.

Through such improvements the SDM Center will keep up with the continuously evolving and expanding data management needs of scientists in various fields of research.

Dr. Arie Shoshani
Principal Investigator, SDM Center

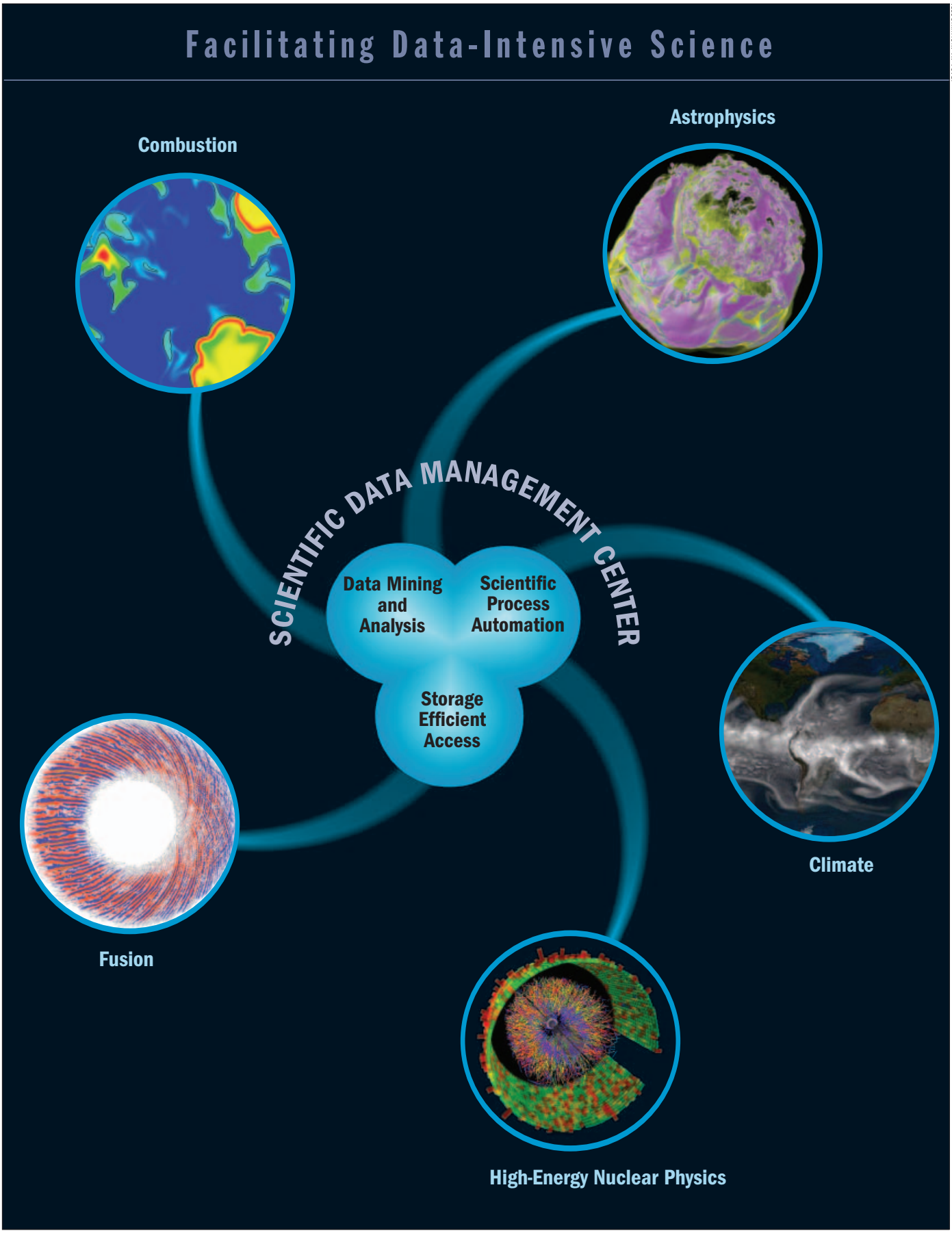


Figure 10. *Combustion, astrophysics, climate, high-energy physics, and fusion are some of the many research applications that have benefited from the creative technologies offered by the SDM Center.*