Identifying Major Scientific Challenges

in the

Mathematical and Physical Sciences

and Their CyberInfrastructure Needs

A workshop funded by the National Science Foundation
Held on April 21, 2004

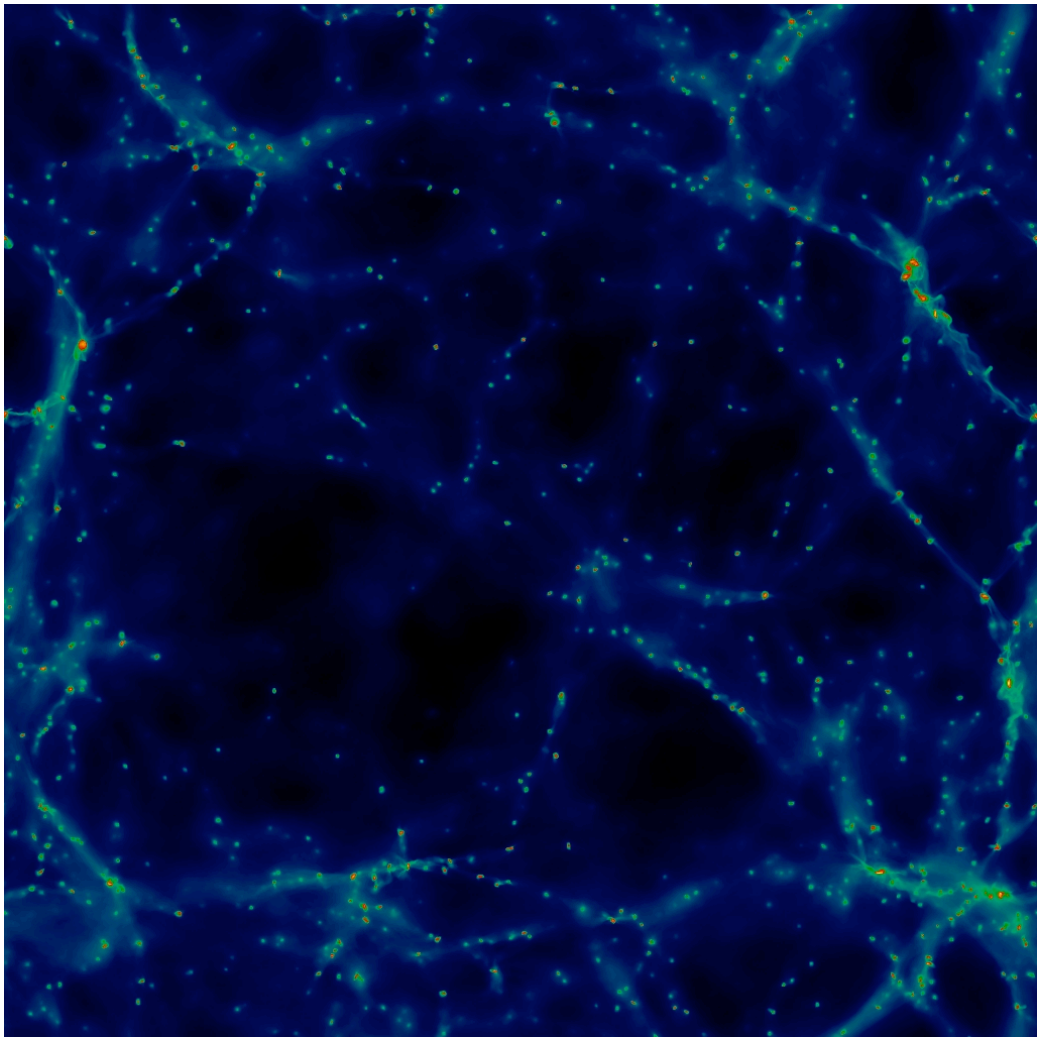


Table of Contents

Executive Summary

Cyberscience Needs and Drivers in Mathematical and Physical Sciences

As the technology available to researchers in the mathematical and physical sciences become more advanced, the opportunities for doing science also evolve, requiring development of the new computational tools that will enable the frontiers of science discoveries. An important role for the members of the scientific community is to identify needs for cyberscience, defined as the science that cannot be done without the advanced capability of cyberinfrastructure. Advances in information technology-enabled systems, tools and services are revolutionizing the practice of science and engineering research and education. Today, simulation and modeling can be as important to discovery and innovation as theory and experimentation. Advances in sensor technology and the availability of affordable mass data storage devices are making possible the collection, creation and federation of large, complex datasets, and pervasive networking technology has made possible broad access to them. Grids are providing scientists and engineers, both professional and amateur alike, with access to a wide range of valuable, heterogeneous computing and information resources as well as sophisticated research instrumentation. Information technology has also permitted the creation of collaboration-rich digital knowledge environments, where researchers and educators work together on projects of common interest.

A Cyberscience Workshop sponsored by the Mathematical and Physical Sciences (MPS) Directorate of the National Science Foundation (NSF) was held April 21, 2004. Defining the role of the MPS Directorate for supporting these opportunities to push the horizons of research was the goal of this workshop.

Main Recommendations

The workshop covered a tremendous amount of material in one day. However, these recommendations reflect the discussion of many people over a very short period of time and must be considered guidelines for establishing more detailed recommendations to come in the future. Also note that many conclusions specific to the breakout groups are not captured in these broad recommendations.

1. Tools for cyberscience should be supported, must be shown to enable accountable science research and sit on the cyberinfrastructure being funded by NSF. The needs of each division will vary across MPS, so guidelines must be developed to allow for these variations. These guidelines must clearly define consistent and reliable policies and programs, and describe science-driven activity, differentiating from that which CISE would naturally fund (or delineate joint projects). [Note that not all cyberscience is large scale. As such, an SBIR-like process could be developed for support of such development, with successful Phase 1 projects leading to further support as Phase 2 projects] ;

2. Divisions within the directorate should determine what they are currently spending on cyberscience and share this information with the MPS Directorate and MPSAC;

3. The MPS budget should be assessed with respect to cyberscience awards, and

reallocated with specific verbiage to cyberscience and supplements to proposals with cyberscience tool components. As much as possible, collaborative efforts with CISE should be encouraged as well as interagency collaborations;

4. Coordination of cyberscience as seen by MPS (and other directorates) and cyberinfrastructure as seen by CISE should be addressed "up front." A cross-cutting office, such as an Office of Shared Cyberscience and Cyberinfrastructure, could oversee these efforts, and its location within the Foundation should be evaluated. Such an organization, recently established in CISE as the "Division of Shared Cyberinfrastructure" would serve all NSF Directorates. For the needs of MPS, however, a "Cyber working group" with members from MPS community and representation from the CISE community should be immediately established to begin to develop the detailed guidelines of how MPS operates in this construct based on its enabling science. This group should communicate its findings often to the MPSAC, to the CISEAC and to a larger, overarching NSF-wide team.

5. The goals of the cyberscience programs should be clearly presented. Metrics and assessment guidelines must be developed to assure accountability and to assess effectiveness of the program

6. MPS with advice from the MPSAC should develop a means to communicate the opportunities and advances in cyberscience and cyberinfrastructure. One mechanism could be to add an obvious cyberscience component to its web page

Mathematical and Physical Sciences Workshop on Cyberscience

The purpose of the workshop was to identify needs of the larger community of mathematics and physical sciences for supporting and enabling cyberscience. A number of independent, disciplinary groups have recently held similar workshops to address discipline or mission specific needs in cyberscience. For example, see:

The report on computational chemistry conducted by the National Research Council available at http://books.nap.edu/openbook/030908721X/html/index.html

Blue Ribbon, Atkins Report January 2003 http://www.cise.nsf.gov/sci/reports/toc.cfm or www.communitytechnology.org/nsf_ci_report/report.pdf

DOE Workshop on Theory and Modeling in Nanosciece May 2002 is found at http://www.sc.doe.gov/bes/reports//files/TMN_rpt.pdf

Recent report on *Statistics: Challenges and Opportunities for he 21st Century* with broader implications for cyberscience in mathematics is at http://www.amstat.org/news/nsf4Aug04.pdf

Report on *Computation as a Tool for Discovery in Physics* from a Workshop on Sept 2001 http://www.nsf.gov/pubs/2002/nsf02176/start.htm

The report of a workshop conducted by the Department of Energy held June 24-25, 2003 entitled *A Science Based Case for Large-Scale Simulation* available at http://www.pnl.gov/scales/

The uniqueness of this workshop was the bringing together of the diverse members and disciplines of the MPS community to define scientific problems that would be extremely difficult or impossible to accomplish without advancing capabilities of Cyberinfrastructure. Unlike previous committees and reports on this matter (discussed above), this group was cross-cutting, having representation from all divisions in the MPS Directorate. The group was to identify the major computational and infrastructure challenges that must be overcome in order to enable MPS researchers to solve the pressing scientific problems of the next decade and beyond. These cyberscience opportunities will ultimately drive MPS investments in cyberinfrastructure and computational needs across the directorate. This proposal draws freely from reports of the Mathematics and Physical Sciences Advisory Committee (MPSAC), and from other recent reports on the topic of cyberinfrastructure.

Collaborations

Exploiting this emerging opportunity requires several forms of strategic planning, collaboration, and investment. The development of the global networking tools for the mathematical and physical sciences needs to be done through collaboration between the Computer Information Sciences and Engineering (CISE) and MPS Directorates. For example, GRID development at

the present time is being driven by the needs of the Large Hadron Collider (LHC) collaboration, with a collaboration involving Physics division in MPS (PHY), CISE, the Department of Energy (DOE), and the European Union. This collaboration is well along in organizing and needs the resources to succeed. Similar examples show scientific disciplines across MPS are likewise involved in conceptualizing the national and international infrastructures and supporting tools. Some computational science requires sophisticated use of distributed resources, with the necessary hardware, software, security, dedicated IT staff, etc. Use of distributed facilities on all scales requires the infrastructure to enable scientists to become directly involved in the operation and use of facilities distributed across the country and beyond. These partnerships need to be more strongly coupled with the needs of MPS and maintained as tools that push the frontiers as MPS moves forward in planning for cyberscience investments.

Organization and Content

Representatives from all of the divisions in MPS contributed content to the meeting. The workshop consisted of a morning session with six invited speakers drawn from the MPS disciplines, each presenting "science drivers" as illustrated by their own research. The morning speakers were Dan Reed of North Carolina State University, Larry Smarr of the University of California at San Diego, Alex Szalay of Johns Hopkins University, Brent Fultz of the California Institute of Technology, Vijay Pande of Stanford University, and David Keyes of Columbia University. The afternoon consisted of several breakout sessions organized around common themes consisting of needs associated with 1) algorithms and software; 2) software infrastructure; 3) hardware and facilities; 4) network infrastructure; and 5) data management and infrastructure. The evening session reviewed the day and sought commonalities across the MPS disciplines. and the workshop began in the morning with six invited talks.

Keeping in mind the purpose of this meeting, Dr. Reed emphasized that "the purpose of computing is insight, not numbers (Hamming)" that we must focus on computing for science versus computing as science. Dr. Szalay described how discoveries are made at the edges and boundaries of science. These boundaries are pushed forward by the tools that we have available: the utility of computer networks grow as the number of possible connections, internet and grid tools are converging, and optimization of searching is needed (software and algorithms). Dr. Keyes postulated that simulation can produce more than "insight" and had quoted J.S. Langer as "The computer literally is providing a new window through which we can observe the natural world in exquisite detail." He also noted that Dr, Ray Orbach, Director of Science at the Department of Energy, claimed, based on a simulation, that the ITER design of plasma reactor would be capable of achieving fusion. Dr. Fultz, in discussing large experimental facilities (neutron scattering as an example), presented the frontiers and needs in data collection, software systems, data handling and reduction required for these high thru-put experiments and visualized the need for direct comparison of real time data to simulations of the detector in real time, real-time visualization with "smart experiments" and data archiving with meaningful meta data. Dr. Pande presented work on protein folding and the coupling of theory, simulations, and experiment, again asserting that the computer models must provide insight and not just reproduce experiments. In his model of multi-node systems, computations are done via new algorithms designed to efficiently use large-scale distributed computing on public machines much like "*SETI@Home,*" leading to calculations on the 100TFLOP scale, far beyond what would be

possible with the fastest supercomputers.

Explanations, Comments and Cautions

Because of differing computational needs and complexity implications across disciplines, a common and important theme emphasized that one size does not fit all—across MPS, different divisions and groups within divisions will have different needs in order to access cutting edge science in their fields. The key rate-limiting step is not necessarily access to central processing unit time, but to people and development. For example, the development of algorithms is one of the areas requiring development, and this need may require MPS investment in people at the interface of science and computer science. Much of the most frontier science will arise from collaborations between those who know the important physical science problems and those who understand in detail the fundamental computational science limitations and opportunities. NSF should invest in the people in that area between computer science and science, and accelerate that capacity in a manner that enhances developments in both. The workshop participants generally agreed that long-term support for people in *infrastructure* should not be an MPS role.

As Cyberscience evolves, the model of the supercomputer center could become one of a broader variety of facilities to tackle large or complex computational problems. It was noted that assistant professors are running their own, ever increasing number of Beowulf clusters. These needs must be met by whatever means is most efficient and cost effective for the entire community. The NSF Small Business Innovative Research (SBIR) Phase I/Phase II model is one that could be explored to address these disparate computational needs when the scales are sufficiently small. As well, this interface between science needs and CS development needs to be cultivated and developments made at this junction between the directorates.

Considerable discussion ensued regarding the balance between the needs for standards versus "letting a thousand flowers bloom." Cyberscience rests on cyberinfrastructure but should have no other restrictions. On the other hand, duplication of efforts is inefficient and fiscally wasteful. MPS will likely not dictate a standard for all divisions; rather, foster debate and encourage dissemination of different ideas, test codes, and converge on a standard. This is a community-driven, bottom-up process. NSF as a whole needs to determine the best mechanisms to facilitate communication across different communities to avoid "reinventing the wheel".

Cyberinfrastructure developments can have a tremendous impact in education. For example, access to modern instrumentation and virtual experiments could be accessible in real time from any school in the country. Tools for visualization and simulation are being developed, and MPS is encouraged to both assess and develop these activities internally as well as to examine these outreach tools developed by other NSF directorates. The socialization and human aspects of the widespread use and acceptance of cyberinfrastructure as well as the international access and funding questions are also noted as areas in which NSF should contribute.

Finally, the science component of cyberscience--that is fundamental developments in mathematics and physical science—should maintain its base within MPS and should not be treated as computer science research. In a parallel fashion, other Directorates will have components that reside in and are funded internally. Science at the interface that pushes the

development of computational power and algorithms (and thus becomes fundamental CS) and that of larger scale with natural ties to CISE could reside in the office of Shared Cyberinfrastructure. The CISE budget is currently thought of by that organization as that of a computer center plus a computer science department, and these are two very different organizations. Learning the needs of many MPS scientists will generate good computer science research, which will in turn, enhance the science that MPS can produce. These are truly synergistic activities.

Each of the following chapters reflects the thoughts, considerations and collective wisdom of each of the breakout groups. No attempt has been made to modify or homogenize these independent reports.

1. Models, Algorithms, and Software

T. Dunning, Jr. (co-chair), R. Larter (co-chair), G. Allen, A. Chan, M. Ciment, R. Fedkiw, J. Fish, A, Frank, M. Gunzburger, G. Hummer, L. Lehner, H. Levine, D. Lockhart, M. Norman, M. Novotny, V. Pande, T. Russell, W. Symes, W. Tang and S. Weidman

Questions, Principles, and Recommendations

The focus of the breakout session on "Algorithms and Software" was on scientific applications software and the algorithms found in that software. Both generic algorithms (*e.g.*, algorithms for matrix diagonalization or solution of a set of linear equations) as well as application-specific algorithms (*e.g.*, algorithms for construction of a Hamiltonian matrix or the calculation of special types of integrals) are found in scientific applications software. Generic algorithms and libraries were the focus of the breakout session on "Software Infrastructure." This session focused on scientific applications software and the application-specific algorithms found therein.

Before beginning its deliberations the breakout group decided that it should include the development of *computational models* in its discussions. Computational models, algorithms, and software are inextricably linked—the model defines the equations that will be solved, the algorithms for solving those equations define the software that will be developed, and, closing the loop, the software developed using those algorithms allows the model to be validated or refined.

QUESTIONS

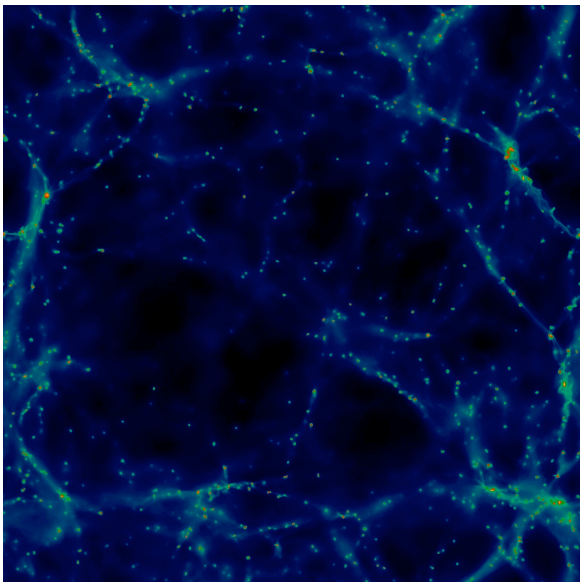
The panel found it difficult to determine what new investments were needed in "Models, Algorithms, and Software" because it did not have sufficient information on the current investments in this area from the Mathematical and Physical Sciences Directorate. In addition, it did not have information on the proposed cyberinfrastructure investments in the Computer and Information Science and Engineering Directorate. Because of the interlocking nature of the investments in MPS and CISE, it is critical that the plans for both directorates be well integrated and that each directorate understands what the other directorate will provide.

The lack of information on the current investments in cyberscience in MPS and on the plans for the development of the underlying cyberinfrastructure in CISE led to the following questions:

- What does MPS currently invest in the development of models, algorithms, and software? What does MPS currently invest in the development of algorithms and of scientific applications software for advanced computer architectures? Are these investments made with the understanding that sustained support is needed for those algorithms and applications software found to be useful by the scientific community? What funding model will be used to support these long-term investments?
- What does CISE currently invest in cyberinfrastructure? What elements of the cyberinfrastructure will CISE be responsible for in the future? Will any elements of the cyberinfrastructure be jointly funded by MPS and CISE? Are these investments made with

the understanding that sustained support is needed for the cyberinfrastructure to be useful to the scientific community? What funding model will be used to support these long-term investments?

– How will MPS and CISE ensure that the elements of the cyberinfrastructure acquired or developed in CISE are useful to MPS researchers and educators? The plan for this should be shared with this (and other) affected groups.



Despite the lack of answers to these questions, the breakout group decided that it was possible to provide MPS with its thoughts on the basic principles that should be used to guide the investments in cyberscience as well as make a number of basic funding recommendations.

Galaxy formation and large scale structure. Early galaxies form in filamentary chains in this 1 billion cell hydrodynamic simulation done at the NSF's San Diego Supercomputer Center. Future cyberinfrastructure will permit resolving the galaxies' internal structure and detailed comparisons with observations of high redshift galaxies. [M. L. Norman, UCSD]

BASIC PRINCIPLES

Investments in cyberscience must be viewed as long term, sustained investments that will raise the level of research and education across *all* fields of science supported by MPS. The Panel felt strongly that the investments in "Models, Algorithms, and Software" should be guided by the following basic principles:

– *Cyberscience is at different levels of maturity in the various fields of science, and these differences must drive the cyberscience investment strategies in MPS.*

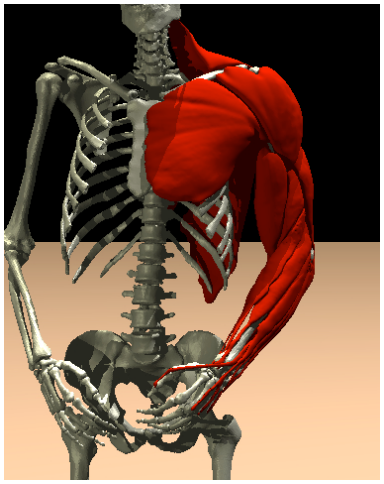
In molecular science the basic mathematical equation that describes the phenomena of interest is known—the Schrödinger equation. On the other hand, in the biology of complex systems, the basic models describing the phenomena of interest are still under active development; these models, when completed and validated, will define the mathematical equations to be solved and the software to be developed. In many disciplines where the model is, in principle, well established (*e.g.*, astrophysics, electromagnetism, fluid mechanics, general relativity, molecular and materials science, and plasma physics), the complexity of the actual implementation require significant advances both at the algorithmic, software, and hardware level to achieve an accurate description of the phenomena of interest. Cyberscience investments will differ depending on the field but are critical to advancing all fields.

– *Investments should be made in the entire knowledge chain in cyberscience: theory and computational models, algorithms, computational frameworks, and scientific applications software as well as in applied mathematics and statistics*.

  Cyberscience is still a rapidly developing field and additional, sustained funding in all of these areas will contribute to its advancement. In some cases, progress will be determined by theoretical or mathematical advances, *e.g.*, the development of new multiscale or multiphysics methods or new functionals for density functional theory. In other cases progress will be determined by the development of new algorithms and computational science software that embody these theoretical advances or make efficient use of new, more powerful computing systems. Some of the research efforts will involve single principle investigators, others will involve collaborations of a few scientists, and still others will involve large, multidiscplinary groups. The fraction of any new funding invested in each of the four areas noted above should be guided by the needs of the various scientific fields and subfields.

– *Funding for computational models, algorithms, and applications software should be at a level commensurate with the strategic role that it plays in cyberscience*.

  It is critical that the investments in the software infrastructure for cyberscience be in balance with those in the hardware infrastructure. Otherwise, the scientific advances promised by the hardware investments will not be realized. In determining the level of support for the cyberscience software infrastructure, MPS should consider that costs in this area are dominated by funding for personnel rather than funding for equipment. The group also noted that, since computational hardware continues to evolve, it is essential that investments in computational methods, algorithms, and software be a *sustained* effort with continuous support.



One new frontier is the simulation of humans. The image on the left shows a biomechanically accurate reconstruction of an NIH data set depicting tetrahedral meshes for muscles and triangle meshes for bones. [Ronald Fedkiw, Stanford University]

BASIC RECOMMENDATIONS

The development of scientific applications software is an intellectual challenge comparable to any that is faced in scientific research. This is especially true for those large software systems that are being developed to serve a broader community of scientists. In fact, these codes are often developed by a community of scientists and referred to as “community codes” (examples include CCSM for climate modeling, ENZO for astrophysical modeling, and NWChem for molecular modeling). These software systems are the “research instruments” of computational science, turning “raw computing cycles” into science advancements (many also support educational activities).

The Panel urges MPS to show leadership in fostering and supporting the development of computational models, algorithms, and applications software systems for the broader scientific community. At a time when computation is playing an increasingly important role in scientific research and education, a sustained effort to develop the software infrastructure for realizing the goals of cyberscience must be aggressively supported. This will require leadership from MPS for many scientists, especially experimental scientists, do not yet appreciate the importance of such investments (although they would enthusiastically support the development of a new experimental instrument for advancing science). Even some in computational science, so long forced to make approximations of unverified reliability, do not appreciate what can be learned from benchmark studies enabled by the new generation of computing systems.

The panel recommends that:

- *MPS invest in the sustained development, deployment, and support of computational modeling software of high value to the broader scientific community*.

  Support for these activities must be structured to have a duration consistent with their nature—the standard three year award period is usually insufficient to bring projects of this type to fruition. In addition, once the software is in use in the community, sustained support is needed to maintain the software, adapt it to new computing systems, and integrate computational models and algorithms developed in other research efforts into it. The NIH Biomedical Technology Resources program can be used as a model for funding this latter activity. The NIH BTR program funds the professional staff (programmers, software engineers, application specialists) needed to carry out these tasks.

- *MPS develop program that fosters collaborations between applied mathematicians/statisticians and physical scientists that will serve as a venue for mathematical scientists to fully participate in cyberscience research and education*.

  A number of the fundamental limitations in cyberscience at present are, at their base, problems in mathematics or statistics. The combined efforts of applied mathematicians/statisticians and physical scientists are often required to solve such problems.

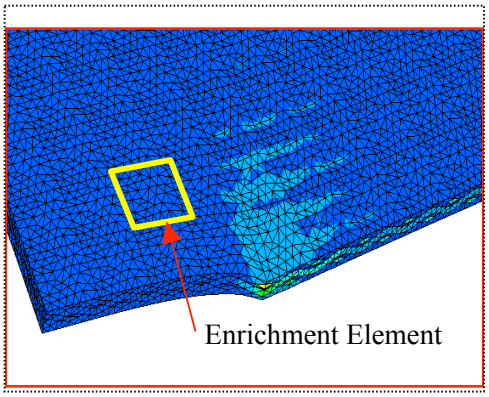
- *MPS ensure that cyberscience is properly represented in the solicitations for any new Science & Technology Centers*.

  There are a number of opportunities in cyberscience that would benefit from the long term, multidisciplinary approach taken in NSF's S&T Centers.

- *MPS fund an Institute for Cyberscience with the mission to advance cyberscience within the areas supported by MPS and to provide a bridge between these areas as well as with the research and educational activities supported in CISE*.

  There are many opportunities for synergy between the various disciplines in MPS and between MPS and CISE. Establishment of an Institute for Cyberscience would help optimize the return on cyberscience funding in the various disciplines by expediting the transfer of knowledge acquired in a particular discipline to others for which it is relevant. Additionally it would spur closer communication between applied mathematicians, computer scientists, and computational scientists.

This figure illustrates application of the Multiscale Enrichment method based on the Partition of Unity (MEPU) idea. MEPU is being developed for enriching the coarse scale continuum descriptions with fine scale features and the quasi-continuum formulations with relevant atomistic data to allow for simulations across broad ranges of length scales and shows considerable improvement over classical homogenization theory and quasi-continuum formulations. [Jacob Fish , RPI]



2. SOFTWARE INFRASTRUCTURE

Linda Petzold, chair; Daryl Hess, co-chair; R. Friesner; R. LeVeque; R. Martin, R. Nichol; W. Pulleyblank, K. Rajan; G. Sanders, L. Smarr; A. Szalay; W. Tang, T. Windus

SCIENCE DRIVERS

Materials Research

The search for new materials, whether through experiment or simulation has been a slow and arduous task, punctuated by infrequent and often unexpected discoveries. Each of these findings prompts a flurry of studies to better understand the underlying science governing the behavior of these materials. The few systematic efforts that have been made to analyze trends in data as a basis for predictions have in large part been inconclusive, not least of which is due to the lack of large amounts of organized data and even more importantly the challenge of sifting through them in a timely and efficient manner. When combined with a huge combinatorial space of chemistries as defined by even a small portion of the periodic table, it is clearly seen that searching for new materials with a tailored modulus is a prohibitive task. Hence the search for new materials for new applications is limited to educated guesses. Data that does exist is often limited to small regions of compositional space. Experimental data is dispersed in the literature and computationally derived data is limited to a few systems for which reliable quantum mechanical information exists for calculations. Even with recent advances in high speed computing, there are limits to how many new alloys can be calculated. Hence this poses both a challenge and opportunity. The challenge is to both build and analyze extremely large disparate databases and large-scale computation. There is a need for multiscale computational methods and software.

Plasma Science

Since plasmas comprise over 99% of the visible universe, understanding turbulent transport behavior in magnetized plasmas is a major scientific challenge of interest in laboratory as well as natural environments. In particular, the efficiency of magnetic confinement of a hot plasma is dictated by the degree to which it is possible to control or eliminate the rapid losses caused by microscopic turbulence driven by pressure gradients. Gaining the physics insights into the associated plasma turbulence and transport is needed for developing techniques to improve confinement in toroidal plasmas. This is recognized as a grand challenge for analytic theories and experimental investigations due to the nonlinear particle and wave dynamics, kinetic resonances, and complex geometries. Consequently, direct numerical simulation, such as nonlinear gyrokinetic particle-in-cell (PIC) simulation, is playing an increasingly important role in turbulence and transport studies.

Barriers include the management of terabytes of data, computer power, visualization of massive amounts of data, facilitation of multidisciplinary collaboration, and multiscale algorithms in computational fluid dynamics.

Cosmology

Issues in cosmology include what is the energy content of the universe? What are the atoms doing? How are galaxies built? How are stars built? There are massive data sets, which require cross-correlation. Simulation is used to build thousands of virtual universes, to quantify the error

of measurements. As the measurements become capable of greater and greater resolution, more and more simulations are required to obtain a smaller statistical error.

Biology
The spectrum of applications here includes genomics, regulatory networks, signal transduction cascades, protein folding, fluid transport, and thermodynamics. Barriers include getting the data, dealing with low signal to noise ratios and false positives in the data, developing models, integrating heterogeneous data, finding the right level of abstraction, simulation at multiple scales, interoperability of models, communication of models driving SBML-type modeling languages, and dealing with data when some of it is proprietary.

Relativity
Issues in relativity include digging through the noise of the gravity wave detector to look for the signal. General relativity simulations are used to first derive the possible shapes of the signal. Thus the discovery potential depends on computational templates that must precede the experiment. Barriers are data mining, statistical analysis, and the low signal to noise ratio.

Computational Chemistry
One of the challenges for molecular chemical sciences is the maintaining chemical accuracy in bond making and bond making processes while increasing the size and complexity of the systems under consideration and while increasing the length scale on which important phenomena occur. There are many examples of this type of science, but one good example is examining electron transfer in biological systems. Usually, these large systems are examined using classical approaches. However, electron transfer cannot be modeled in this manner and quantum mechanical approaches are needed. This requires multiscale mathematics, which carefully examines boundary conditions as well as the physics of the problem. Another part of this problem, however, is the use of data from one scale (such as molecular level simulations) to other scales (such as the development of kinetic models). There are many data and command representations as well as software packages at each of these levels and very few of them are "coordinated". Data sources are also diverse in their location and their use of storage technology. All of these make the development of informatics-based approaches to new scientific discoveries a difficult and slow process. Semantic web capabilities and ease of use are samples of issues that must be addressed in this context.

Another issue is that there are many people developing different theories and algorithms that are advancing the science. This is a positive process in that new methods become available to solve science problems. However, it leads to complications in comparing the results of these methods as well as the "sharing" of these algorithms between codes (different languages and parallel models for example). Often, chemists have to have a diverse set of software codes available to them to accomplish their science. Being able to share algorithms between software codes would help in this process. Also, using new mathematical software/algorithms as well as enabling multiscale science will require software to work together that was not necessarily designed to work together. Enabling this reuse of software is necessary to take advantage of advances in all fields. At a minimum, being able to have the explicit and implicit data that are the inputs and outputs to these algorithms so that the results can be carefully compared and used is necessary.

ELEMENTS OF THE SOFTWARE INFRASTRUCTURE

Data Management

It is clear that data management will be a major issue. The data management problem comes in several forms: managing terabytes of data, and managing less data but which is distributed in many files. Challenges include archiving the data, annotating, retrieving, distributed archiving and annotation, interoperability, searching, querying, protecting the data, streaming data, proprietary data, and the integration of data management with knowledge management. A data-mining infrastructure featuring both interchangeable data structures and interchangeable data mining algorithms is desired. Before proceeding on this front, however, NSF should take a careful look at the commercial data management options to determine which of these problems have already been solved in the commercial market, and what are the unique requirements of science.

Software ease of use and interoperability is another barrier. The applications require increasingly complex and integrated software. Scientists would like to be able to update/improve a specialized part with new knowledge or new algorithms, and be assured that the system will still work. This is not the case with current software. Mathematics is needed to address the stability and control of the proposed interoperable computational software structure.

Algorithms and Computational Software

Algorithm research and development is needed. Algorithms have been at least as fast and as powerful as hardware in following Moore's Law, AND you get to multiply the speedups of algorithms by the speedups due to hardware. Figure 1 shows the performance improvements for a core computational problem over a period of 30 years for algorithms and for hardware. We note that the dates for the algorithms are the approximate dates when the algorithms went into production in large-scale DOE applications.

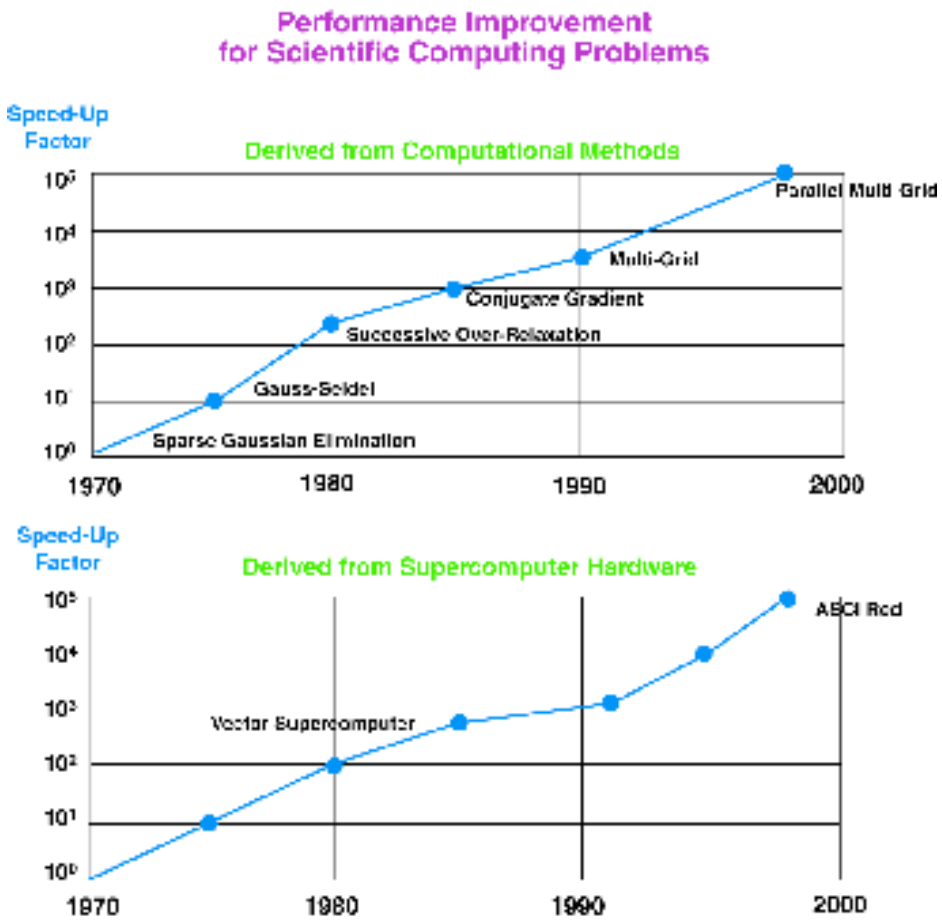


FIGURE 1 (software infrastructure)

Which algorithms are needed for the next generation of scientific problems? They should address the statistical challenges of massive dimensions, and the heterogeneous multiscale nature of the science driver problems.

Infrastructure Issues

Databases of software should be made available to practitioners. Automatic assistance needs to be provided to help locate the most appropriate algorithm for a given problem and data. Problems such as how do you communicate the meta-assumptions of your data need to be resolved. We need to be able to assure quality whenever possible, but not set standards prematurely or too strictly as to inhibit development of new tools and technologies.

RECOMMENDATIONS

Software Infrastructure Support

To develop, manage, leverage and maintain the software infrastructure needed for the next generation of science, sustained infrastructure and facility support will be required.

Science needs to be the driver for software investments. At the same time, software and algorithms which have proved successful at a research-group level for scientific applications should be extended and made available to the wider scientific community. The latter is not a

trivial problem; often the manpower and time required to take a code from the level where it is useful within the research group which generated it to where it can be applied easily and without assistance to a more general class of problems is greater than the effort to produce it in the first place. To deal with both of these issues: that science should be the driver and to identify those software projects that will have the best potential for impact on science; and at the same time to provide a funding mechanism for taking that software which has been identified for its potential of high-impact and chances for success and leveraging its use over a wider range of problems, we propose the following funding structure, illustrated in Figure 2. At the top layer are multidisciplinary, science-driven research projects focused on the development of algorithms and software. These types of projects have often been successful in fostering multidisciplinary collaborations. But where they have resulted in powerful new algorithms or software, this infrastructure has rarely migrated out of the research group where it was generated.. These top-layer projects will be called the 'Phase I' projects. Phase I is the 'proof of principle' stage for the algorithms and software that are generated in these projects. It would be known from the start of Phase I that those projects that generate potentially high-impact software which could benefit the scientific community by development of more generic, user-friendly 'production-level' codes would be candidates for a Phase II project. Thus Phase I projects should be motivated to produce not only state of the art, high-impact science, but also to be the seeds of state of the art, high-impact software. Phase II projects should involve a core of researchers from selected Phase I projects, in collaboration with significant computer science expertise, with the aim of extending and leveraging the software infrastructure, and working with appropriate personnel at a Center to bring the software to a state where it is maintainable, widely accessible, able to make use of appropriate high-performance architectures, etc. Phase II projects should be subject to full and intensive panel review, with funding criteria to include 'proof of principle' for the algorithms and software based on Phase I results, the potential impact of leveraging the software infrastructure, and the capability of the team and viability of the proposed plan to complete and release the proposed software within the given time frame. We emphasize that a Center (or Centers) is required because software maintenance and consulting is labor-intensive, more cost-effective and reliable if centralized, and can easily overwhelm the researchers who originally developed the software (and who are not funded to do maintenance).

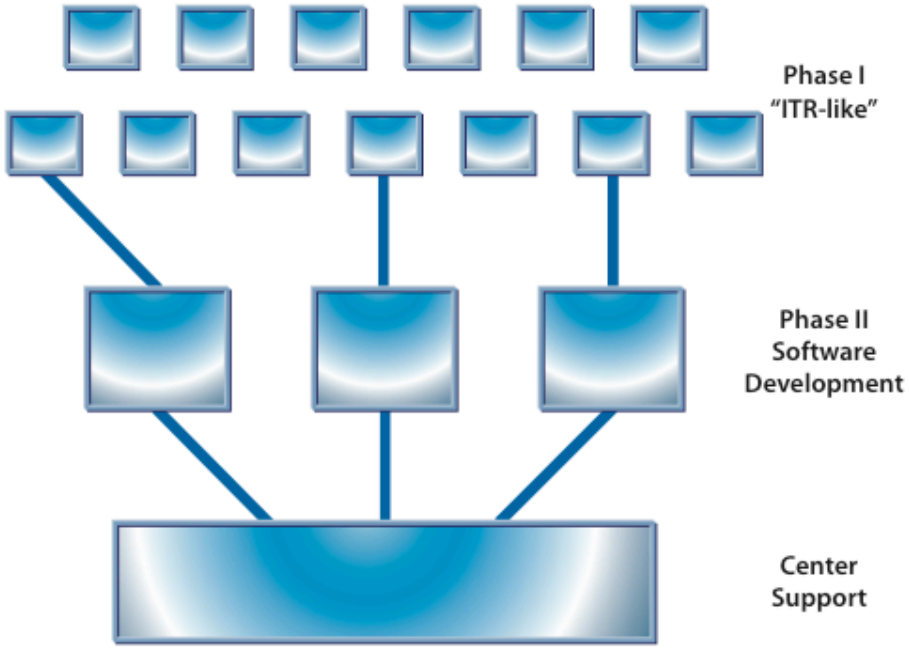


FIGURE 2 (software infrastructure)

Software begins with algorithms. As discussed earlier, algorithms have been at least as powerful as hardware in terms of their ability to accelerate not only computations, but the development of science itself. Yet the investment in algorithms has been miniscule in comparison to the investment in hardware. We suggest that NSF increase the ratio of ($$ spent on algorithms) to ($$ spent on hardware) substantially.

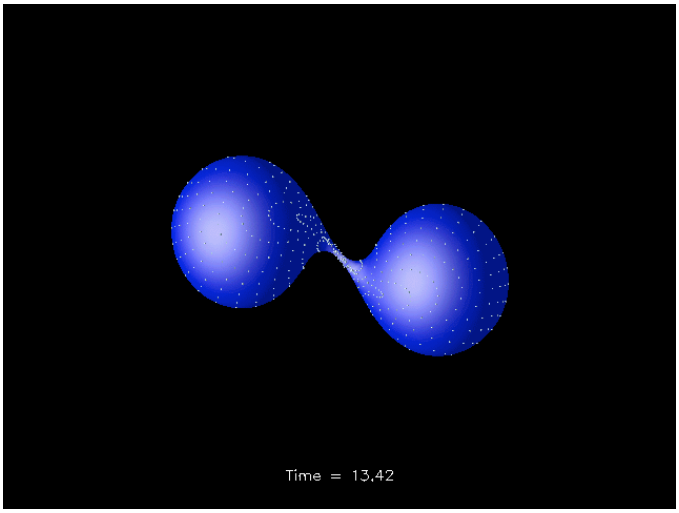
Cultural Issues

There are cultural issues that must be addressed in any large-scale program to undertake multidisciplinary research and software infrastructure development. How can we reward such activity, which is often high-impact, so that more researchers will be in a position to do it? NSF has played a big role and has been very successful in fostering multidisciplinary research collaborations. There is more that needs to be done, particularly in the area of software infrastructure development.

The obvious step is to fund such work. It needs to be emphasized that software infrastructure development generally requires substantially longer than the standard NSF grant length of three years to develop. Grant lengths of AT LEAST 5 years for software development are recommended, and of 10 years for Centers. There is a big startup cost and time for software development. The impact will only be realized if NSF recognizes from the start that software development requires both manpower and time. For very large software projects, it may be advisable to get experts involved in the design, to enhance the possibilities for collaboration and to ensure that users will feel comfortable enough to use it.

A problem for software-oriented proposals is that they often are not considered a priority with review panels. To emphasize the importance of infrastructure development, NSF might consider adding to the list of proposal evaluation criteria, which is currently intellectual merit and broader impacts, the area of tool development. However it is accomplished, if one of the goals of a proposal solicitation is the development of tools, this needs to be emphasized in such a way that reviewers are keenly aware of it.

Multidisciplinary research, as well as software development, has generally been considered to be a tenure risk for young faculty. To encourage academic departments to recognize these activities, which often have a high impact on the acceleration of science but are not well-regarded in disciplinary science, we suggest that NSF consider adding a Career Award for Multidisciplinary Science. As it currently stands, as much as NSF would not like to admit it, Career Awards are a major factor in tenure decisions at Research I Universities. A young faculty member is, under the current system, more likely to receive one of these awards if he/she is doing disciplinary science, as a result of the disciplinary composition of the NSF Career Award Panels.



The image shows the event horizon just after the merger of a binary black hole system. The initial data was set up so the system was inside the ISCO and consequently the black holes merged before an orbit was completed. The white dots on the surface show the location of individual horizon generators. The effect of the angular momentum of the system can be clearly seen in the asymmetry of the horizon. In order to push this science forward, support is needed for certain tools such as: 1. adaptive mesh refinement, 2. multi coordinate patch techniques that will allow us to have boundaries that better conform to the physical system, 3. better visualization tools that can handle to complex data structures and 4. better resource management to automatically employ available executables and keep track of the output. [Peter Diener, LSU, This work was done in collaboration between the AEI and LSU numerical relativity groups and used the Cactus Computational Toolkit.]

3. Hardware and Facilities

Bill McCurdy: chair, Barry Schneider, co-chair, E. Goldfield, R. Pennington, B. Brandt, B. Schütter, R. Roskies, T. Martinez

The "Hardware Breakout Group" of the NSF Cyberscience and Cyberinfrastructure Workshop addressed itself to a single question, "What hardware resources and investments does the MPS scientific community need in its cyberinfrastructure over the next 5 to 10 years?" It is of course clear that the entire range of computing hardware is important to the MPS funded community. However, specific issues were identified concerning the two upper parts of the spectrum of computing needs – the *high-end* and the *midrange* – that are not currently being addressed by the Foundation. Of particular urgency in that respect is an examination of the support of midrange computing resources for MPS scientists.

HIGH-END COMPUTING RESOURCES

The high-end computing resources include supercomputers such as those currently at the NSF funded computing facilities and their possible upgrades, as well as other potential unique resources.

MPS has a serious stake in existence and effectiveness of computing facilities at the high-end. We note that the current facilities funded by the NSF remain over subscribed by a factor of 2 or 3, so that many investigators and collaborations must make do with less computing than they can actually use productively. In addition to the areas within the MPS portfolio with already well-established requirements for these resources, there is an emerging and increasing need for high-end computing in several new areas, among which are nanoscience and the physical biosciences.

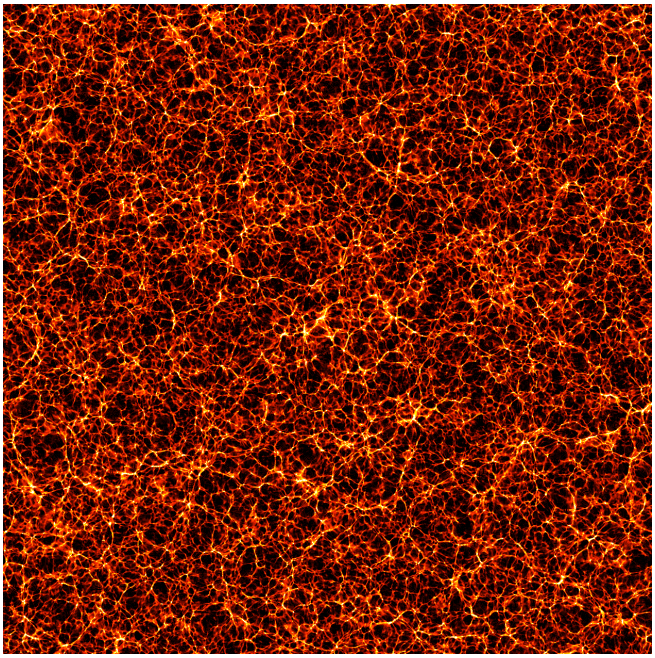
Because of the diversity of disciplines in MPS, we believe that a diversity of supercomputer architectures remains necessary. However it should not come at the expense of large-scale implementations of at least a small number of them that can be used for production scientific computing. Many MPS researchers rely on these facilities, but they can only make good use of them if they can depend on adequate allocations of computer time so that they can plan and execute projects of large scale.

We also noted that, within the MPS portfolio, Quantum Chromodynamics is one of the very largest users of general facilities at the high end of capability, although the resources available in other countries are much greater than in the U.S. Special purpose machines are also of interest, and the QCD community has traditionally made investments in them, but we believe that it is still not clear which of the current competing strategies -- special purpose, general purpose or hybrid -- will be optimal.

MID-RANGE COMPUTING RESOURCES

MPS has an array of important opportunities in the arena of midrange computing. This is the range sometimes thought of as "capacity" computing that extends from the power of clusters containing a few tens of processors, to larger resources that can now approach a TeraFlop of peak power. This is the class of resources most generally used by smaller communities, e.g., 1-5 principal investigators. It is cost-effective and growing in importance. The first point we must emphasize is that this class of computing is of critical to MPS, and we can do that with a few simple observations:

1. Scientific opportunities, like computation in nanoscience problems, design of materials, etc. are being missed because of insufficient resources at the midscale. The point is not that one group needs to run one large simulation on such a resource. It is that many groups need to run very many calculations to make progress.
2. Rapid turn around and sufficient resources encourage more exploratory, higher risk computational research all across in the MPS portfolio.
3. Midrange parallel computing is the resource upon which training of students to do high-end computing depends.
4. Scientific software development for the high-end depends on the availability of midrange parallel computational resources.
5. Some classes of research are best done on midrange facilities. Dedicated midrange facilities make possible real-time, on-demand interactive steering and analysis of progressing simulations. Also, ensembles of simulations can frequently be done most cheaply and effectively with midrange computing resources.



Using PSC's Terascale Computing System, Jeremiah Ostriker and Paul Bode of Princeton, carried out the largest simulation of the universe to date. Their two-billion particle computation used 420 processors for five days. The simulation tracked dark matter and dark energy in a cold dark matter universe from 379,000 years after the Big Bang to the present. This thin-slice snapshot (about one megaparsec thick by 1,500 megaparsecs on each side) shows the filamental structure of dark-matter clusters. (Brightness corresponds to density.) However, important issues that deserve consideration at NSF arise in any discussion of midrange computing, as they did during the workshop. [Ralph Roskies, Pittsburgh Supercomputing Center]

The breakout group discussed the potential impact of dramatically increased funding for midrange computing resources. To pick just one example, we believe that an argument can be made that a large-scale investment of tens of millions of dollars per year in midrange computing

to support nanoscale science could play a key role in determining whether "nanoscience" ultimately will become viable "nanotechnology". Similar arguments can be made for biotechnological applications of the physical biosciences. However, there seems to be no active debate about whether large computing initiatives targeted in this way, rather than initiatives targeted at large facilities, would be a better way to support computational research in the MPS portfolio. It is the responsibility of MPS to stimulate that debate, and act on its consequences.

We also noted that mechanisms do in fact exist within MPS for funding smaller clusters that cost less (generally much less) than $300,000. While more support of computing at that level is certainly desirable and necessary, as we discussed above, no mechanism exists, as far as we know, within the sciences at NSF for funding computing resources costing in the range of $2,000,000. The latter scale represents capacity computing in the class of one or two TeraFlops. Terascale capacity computing is becoming increasingly important to MPS researchers, because it now represents the upper level of parallel computing performance that is routinely practical in many fields. However, the support and maintenance issues associated with such hardware are difficult to address except on an institutional scale. The gap between what is too large for a single research group or department, but too small to fit the mission of a supercomputer center has not been addressed, and should be of concern to MPS.

SPECULATIVE AND EMERGING HARDWARE OPTIONS

There are new opportunities for leveraging commodity hardware that should not be ignored, and there are research groups willing to make the effort as part of their scientific programs. Examples include the exploitation of gaming consoles, graphics processors, etc. In many cases such platforms benefit from an economy of scale and strong isomorphism between the mathematics of science and that of visualization or gaming, like vector processing.

Finally, it must be emphasized as well that MPS has the opportunity to impact the long-range development of hardware with its investments in: quantum computing and information, molecular computing, nanoscale electronics, spintronics, and biological computing.

A CONCLUDING POINT OF CONCENSUS

In the discussions at the workshop, a single point came up repeatedly concerning the topic of hardware, and that was the issue of the consistency and reliability of policies and programs at the NSF in this area. The MPS hardware strategy, like the rest of its cyber infrastructure strategy, must be a long-term, persistent program driven by the science requirements.

4. DATA MANAGEMENT AND ANALYSIS

Participants: S. Keller-McNulty (chair), H. Kaper (co-chair), M. Aronson, B. Allen, P. Avery, C. Breneman, T. Chan, E. Fontes, B. Fultz, J. Green, J. Kettenring, A. Szalay, C. Shoemaker, R. Wilhelmson

Prior to the workshop, the Data Management and Analysis group generated a series of questions, listed below, to frame their discussion. At the workshop we only had time to scratch the surface of this comprehensive set of questions. This report captures that discussion and our workshop recommendations.

PRE-WORKSHOP MOTIVATIONAL QUESTIONS

- What are some of the science drivers in the Mathematical and Physical Sciences (MPS) that generate massive amounts of data, now or in the near future?
- What are some of the major databases of the disciplines? Where are they? How accessible are they? If access is not satisfactory, what should be done to improve it? What should MPS do to improve it?
- What are the characteristics of these databases? Are they maintained in some orderly way? How are they managed? How satisfied are their users? What have we learned?
- What mechanisms are available to display and query results? Do these mechanisms meet the current needs of the disciplines? What about future needs? If improvement is needed, or if bottlenecks can be identified, what should MPS do to improve these mechanisms?
- What are the data analysis issues? What are the data management issues?
- Are existing statistical methods sufficient to extract the information that we are looking for, particularly in the presence of diverse and heterogeneous information sources?
- What is the state of the art when we have many degrees of freedom but (relatively) few measurements? How do we analyze the structure of high-dimensional sparse data sets?
- Do we know how to quantify uncertainty in inferences from data sets including, massive data sets, massive data streams, or high dimensional sparse datasets?
- What software is available? Is it discipline-specific, or generic?
- Is the current proposal/award mechanism responsive to the needs of the MPS community in the areas of data analysis and data management?
- Is the current reward system responsive the needs to develop cyberscience and cyberinfrastructure?

The Data Management and Analysis group found it impossible to separate cyberscience and cyberinfrastructure. The two areas complement each other, they are mutually dependent, and they achieve advances only through integration. In that sense, cyberscience and cyberinfrastructure are the two strands of a double-stranded helix. Nonetheless, the group strongly recommends that an integrated program of cyberscience and cyberinfrastructure be driven by the science. Disconnected research efforts in either area will not lead to the desired end result: advancing the frontiers of scientific knowledge through computing.

The group found it equally difficult to separate data management and data analysis. Our recommendations below reflect these integrated views.

The morning presentations gave wonderful background material on some of the science drivers in MPS for cyberinfrastructure and cyberscience and their need for data management and analysis. It is clear that massive amounts of data are being generated in the various disciplines, that the analysis of these data requires ever more powerful tools, and that there is a need for an evolving analysis methodology. It is also clear that there is no coordination for capturing and archiving the ***entire*** process of scientific discovery. Thus began the discussion of our group, resulting in the following sets of recommendations.

BROAD RECOMMENDATIONS

- Given its mandate to advance the frontiers of knowledge in the mathematical and physical sciences, its history of supporting computation in the various disciplines, and its commitment to promote interdisciplinary scientific efforts, MPS is the directorate most qualified to take the lead in promoting and coordinating cyberscience across the foundation.
- The development of the necessary infrastructure for cyberscience requires cross-agency collaboration, as well as the engagement of the commercial sector. The NSF needs to share the responsibility for developing cyberscience and cyberinfrastructure with other federal science agencies (DOE, NIH, DARPA, NASA).
- The scientific discovery and its corresponding process must be captured as an exciting and living enterprise. The development of "data museums" or "obsolete science archives" should be avoided.
- The cost of a sustained effort to develop and deploy a cyberinfrastructure that is useful for science should not be underestimated. Serious consideration should be given to the development of a business plan that realistically characterizes these costs and the expected return on investment (to the scientific community), complete with a realistic timeline. It should also be understood that most progress in science is evolutionary; revolutionary scientific advances may be glamorous, but they don't occur on a predictable basis. Consequently, it should be acknowledged that investments in and development of cyberinfrastructure are strategic decisions. Measures of success should be realistic and clearly defined; the necessity for long-term maintenance of the infrastructure should be acknowledged, and the responsibility for long-term maintenance determined. Guidance for what will not be developed and funded should likewise be part of the plan, and not an accident.
- Fields that are traditionally thought of as individual-investigator driven ("small" science) are no longer small. Collectively, they generate substantial amounts of data that could benefit from systematic organization and integration, requiring the development of participatory systems with central repositories, accessibility, preservation, and long-term archiving.
- Engagement of the software engineering community in cyberscience is critical. Interdisciplinary teams need to include application scientists **and** software engineers. To be successful, a healthier professional recognition needs to be developed for the

contributions from software engineering. Interaction with the commercial world could be beneficial here.

The term "data management" is too narrow to describe the cyberinfrastucture needs in the context of cyberscience: it is not simply "data" that needs to be managed. The scientific discoveries, as well as the scientific process---including the integration of theory, experiments, and computation---is what needs to be captured by the cyberinfrastructure. The following data management recommendations apply to the entire scientific discovery cycle.

Data Management Recommendations:

- Develop a general-purpose open-source distributed RAID file system with redundancy, put together by teams (application scientists and computer scientists).
- Develop authentication systems to guarantee data quality and integrity.
- Draw on the experience of the digital-library initiative to understand the issues of long-term integrity of data, data preservation, and data curation. The relevant domain science must be engaged in developing this understanding.
- Establish a common framework for the development of metadata, with emphasis on automation. Such a framework will require standards and an enforcement mechanism.
- Engage the computer science community in database research.

The term "data analysis" is similarly too narrow to represent what is needed by cyberscience. The following data analysis recommendations apply to all types of information that is generated through the scientific discovery process.

Data Analysis Recommendations:

- Integrate statistical thinking in cyberscience from the beginning, with attention to issues of data quality, sampling, analysis plans, and analysis tools. Develop an understanding of when various statistical analyses and displays are "good enough," versus the need to deeper methodological development.
- Support efforts to make data analyses useful for the various scientific communities. One size does not fit all.
- Support the development of research for the creation of dynamic data management and analysis tools and standards. This must include the integration of computation and data (e.g., scheduling issues).
- Develop data visualization techniques for both sparse and dense data. Data visualization for the purposes of integration of information appears to be in its infancy, versus the current state of scientific visualization.
- Encourage the development of cyber systems that facilitate data visualization. Current systems for data visualization, for example for the display of genomic structures, are primitive. The problem is hard, because of the dimensionality and the complexity of the structures to be represented.
- Develop pattern recognition tools for high-throughput data streams (e.g., proteomics).

- Develop image analysis tools for complex surfaces (e.g., brain mapping is currently done by hand!).
- Support the development of geometrical and topological methods for analysis of high-dimensional complex data sets (e.g., astronomy, LIGO).
- Promote policies that guarantee open access and integration of data and information; avoid the tendency to make data and information proprietary.

Discussions about the mechanisms for moving cyberinfrastructure forward in the areas of data management and analysis were far ranging, covering topics from education to NSF programs.

MECHANISMS FOR THE DEVELOPMENT AND DEPLOYMENT OF CYBERINFRASTRUCTURE

The following sketch represents a crude hierarchy for the capture and deployment of science, depicting the role of cyberinfrastructure:

> Research → Tech transfer, standards → Cyberinfrastructure for the common good, suite of software → Commercial infrastructure

The point of this representation is to recognize that cyberinfrastructure should not be developed in a vacuum, it must be driven by science and it must recognize that there already exists a huge commercial infrastructure.

The final discussion focused on what NSF could do to promote cyberinfrascture. Two themes emerged.

- Education: There is a significant need for the education and training of a computationally literate workforce in areas of cyberinfrastructure related to data management and analysis. This need exists for "big science," as well as for "small science." To meet this need, NSF should foster the development of a computational science curriculum. It should also promote the enhancement of cyberinfrastructure for remote learning and collaboration-at-a-distance beyond videoconferencing.
- Collaboration across disciplines: Computational science requires the joint expertise of application scientists, computational scientists, computer scientists, and software engineers. NSF has recognized the benefits of these interdisciplinary research efforts, for example through its investments in the ITR priority area. By actively participating in these efforts, MPS has managed to enrich its research portfolio, especially at the frontiers of science. As a result, the Directorate is well positioned to take the lead and promote cyberscience through collaborations across the disciplines.

Finally, it was stressed that a vibrant cyberinfrastructure to support data management and analysis is expensive, both in money and in people resources. It is not too early for NSF to think of cyberinfrastructure as an integral part of its overall support for scientific research.

HGK, 6-14-04

5. NETWORK INFRASTRUCTURE

Participants: Sandip Tiwari, chair; Nigel Sharp, co-chair, Kenny Lipkowitz; Glenn Martyna; Vicky White; Alan Whitney

SCIENCE DRIVERS AND THEIR NEEDS

One way of defining the desired outcome from this workshop is to attempt to address two needs, or perhaps a single need phrased two ways, which affect the community's ability to perform cutting-edge science:

- If we had this capability, then we could do that science

or

- If we want to do this science, then we need that capability

Our small but lively group still managed to represent a wide range of scales of cyber-science and of networking needs, and discussed various ways to approach these problems, concentrating more on generalities than on specific science problems in each MPS discipline.

Science drivers discussed included the increase in wide spread, even international, collaborations, such as the Large Hadron Collider, which will need advances in connectivity, data storage, and interoperability standards, to achieve its scientific promise (and find the Higgs boson!). At the other end of the scale, there is a clear need to link scientists together to make a human network out of the hardware network, capable of expanding and expediting science by a smart linkage of knowledge resources. The specific example given was in molecular design, where it is necessary not only to know what others may have accomplished, but also to inform as many people as possible that you have a molecule that they might be able to use. A similar problem of *ab initio* design was mentioned, where the tools are needed to make predictions and to spread the results and the solutions far and wide. In astronomy, very long baseline interferometry has an interesting technical challenge, in that the data rate between telescopes controls the sensitivity, so you can make what is effectively a larger telescope from the same antennae by increasing network capacity.

The science drivers coalesced around three main needs from a network, which are the ability to handle data as they are generated, the ability to exploit the data, and the ability to know what exists.

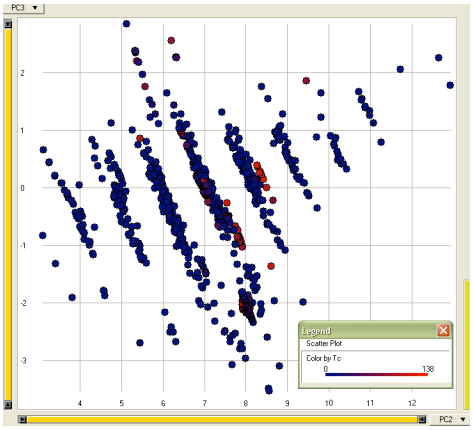
This last requirement was identified as a relatively unusual one, but one of crucial importance in making a network useful. Without prejudicing the implementation, this might involve intelligent agents that look for resources and, as one person pointed out, can also find gaps where the "owner" of the agent might have an application of relevance. The questions are, firstly, how to find out who might be interested in a technique you've developed, and secondly, the reciprocal, how to find techniques from other scientists applicable to your problem. An ideal would be some sort of alert system, capable of recognizing where techniques in one area might be useful in another, and then bringing this connection to everyone's attention. It has been said that the best library is the minds of our colleagues, because they include filters and experience: the current

main problem of, e.g., a web-based search is that of sifting the vast amount of irrelevant material returned. For a network to be usable, there must be some way to incorporate "smarts" into search capabilities.

The problems related to handling data are already being well addressed. For exploiting data, smart visualization algorithms tied to an understanding of how people absorb information are of increasing importance, but are not really network related. Areas for possible investment to try to develop new ideas include: smarter data caching at intermediate sites, with consistency and compliance checks; smart processing built into databases to return only selected subsets of distilled data and thus reduce bandwidth requirements; and compression schemes that convey the essential information with user confidence, possibly including under-used channels of perception such as audio.

We felt that the network infrastructure should encompass not just the hardware, but the software that makes it all function, such as standard network protocols, advanced caching and routing algorithms, and standardized access control (along with identification and verification controls). Much of the work to meet this need is already happening under the aegis of distributed and grid computing.

EXAMPLE: A CYBER DATABASE FOR MATERIALS SCIENCE: AN INTERACTIVE TOOL FOR DATA SHARING AND ASSESSMENT



Databases are traditionally viewed as "static" documents which are used in a "search and retrieval" mode. With the use of IT tools, materials scientists can contribute to the growth of databases and compare their input to other contributors. We are building collaborative databases which are integrated into data mining tools which permit a statistically based visualization of the data "growth" process. The first such database to be posted onto the web portal deals with high temperature superconductors. The figure shows how patterns of behavior have evolved relating systematic changes in chemistry to superconducting transition temperatures. The compounds marked in red with the highest temperatures have been found to be confined to certain valence structures. [Krishna Rajan, Combinatorial Sciences and Materials Informatics Collaboratory ,NSF International Materials Institute, RPI]

Although there will always be a need to "push the envelope" for a few high-end users, because of needs which cannot be met "off the shelf", the general network infrastructure should not be a part of the NSF's research portfolio. Once it is off the research and development track and into commercial production, it is simply a resource to be purchased, like any other. That said, it is a recommendation that the NSF should be involved in creating a coherent US network infrastructure for science, preferably as an interagency partnership, so that once the fundamentals have been created, the running of reliable production services can be handed off to somebody else, rather like the Internet 2 organization. This would need to emphasize the "last mile"

problem, to ensure the focus was on connecting scientists and not purely on connecting centers or institutions. It must allow connectivity across and between national laboratories, universities, community colleges, K-12, non-traditional education systems, and commercial organizations. Great research depends on a reliable and effective infrastructure.

Any created cyberinfrastructure must be a close collaboration of scientists and implementers. However, a major worry is that the research aspects of networking will overwhelm the service aspects. There should therefore be a clear distinction made, so that there is a science net which is service and needs based, and not subject to interruption, development, and alteration for computer science research purposes.

There must also be recognition that network software is as important as hardware in making the network usable, especially when trying to maximize the scientific returns from any provided connectivity.

There was some interest in requesting that the NSF work out how to meet the needs, once they are identified, by taking the initiative to create an appropriate structure, but there is always a problem with NSF, as an agency that responds to proposals, taking that sort of initiative. Nevertheless, the participants felt it should be a suggestion, if not a full recommendation.

In the short term, the NSF is urged to consider how investing in currently available dark fiber, perhaps partnering with Internet 2, might assist in the creation of a cyberscience net.

There were also worries about how to measure effectiveness, should the NSF set out a cyberinfrastructure program advertised as most enhancing science. It would be very helpful to have metrics and assessment methods defined in advance of any program.

The network is the enabler, without which access to all other capabilities is denied.

List of Participants

Bruce Allen	University of Wisconsin, Madison
Gabrielle Allen	Louisiana State University
Meigan Aronson	University of Michigan
Paul Avery	University of Florida
Shenda Baker	Harvey Mudd College
Bruce Brandt	Florida State University
Curt Breneman	RPI
Tony Chan	University of California, Los Angeles
Thom Dunning	University of Tennessee
Sallie Keller-McNulty	Los Alamos National Laboratory
Jon Kettenring	Telcordia
Ronald Fedkiw	Stanford University
Ernie Fontes	Cornell University
Richard Friesner	Columbia University
Brent T. Fultz	California Institute of Technology
Adam Frank	University of Rochester
Jim Green	NASA
Evi Goldfield	Wayne State University
Max Gunzburger	Florida State University
David Keyes	Columbia University
Luis Lehner	Louisiana State University
Herb Levine	University of California, San Diego
Randall Leveque	University of Washington
Kenny Lipkowitz	North Dakota State University
Todd Martinez	University of Illinois, Urbana-Champaign
Richard Martin	University of Illinois, Urbana-Champaign
Bill McCurdy	Livermore Berkeley Laboratory
Glenn Martyna	IBM
Mark Novotny	Mississippi State University

Bob Nichol	Carnegie Mellon University
Mike Norman	University of California, San Diego
Harvey Newman	California Institute of Technology
Vijay Pande	Stanford University
Linda Petzold	University of California, Santa Barbara
Bill Pulleyblank	IBM
Krishna Rajan	Renseleer Polytechnical Institute
Dan Reed	University of North Carolina
Ralph Roskies	Pittsburgh Supercomputer Center
Gary Sanders	California Institute of Technology
Heinz-Bernd Schuttler	University of Georgia
Christine Shoemaker	Cornell University
Larry Smarr	University of California, San Diego
William Symes	Rice University
Alex Szalay	Johns Hopkins University
Sandip Tiwari	Cornell University
Vicky White	white@fnal.gov
Theresa Windus	Pacific Northwestern Laboratory
Alan Whitney	MIT, Haystack Observatory

Acknowledgements

The MPSAC working group of Shenda Baker, Bill Pulleyblank and Gary Sanders would like to gratefully acknowledge Morris Aizenman for his tireless organization and encouragement of this group. We would also like to thank Michael Turner, Judith Sunley, Barry Schneider, Raima Larter, Hans Kaper, Celeste Rohlfing, Daryl Hess, Nigel Sharp, and Jeanne Pemberton for their enthusiastic support for as well as their intellectual contributions to the structure and formation of this workshop and report.