6.0 Budget Recommendations

6.1 Scope of the Program

Achieving the vision of the Advanced Cyberinfrastructure Program (ACP) will require coordinated NSF support of a broader set of activities and facilities than the agency has historically supported. In addition, existing activities (e.g. providing access to high-end computers, enduring data archives, and middleware software development) will need substantially higher funding levels. NSF's role is not limited to financial backing — it is also critical that NSF provide an effective organizational structure to coordinate the ACP, establish operational and user support centers, provide leadership for the nation (including other research funding agencies), and coordinate with similar international activities. This requires not a one-time or short-term initiative, but rather the Panel advocates a material modification to the direction and priorities for the Foundation through a program of sustained long-term funding. In this section, we provide our best estimates of the level and allocation of funding needed near the beginning of the ACP, although we expect this estimate to be modified over time as needs and priorities change.

As described in Section 2, information technology tools and resources should not only support high-end numerical simulations and network connectivity (the major emphases in the past), but also digital libraries, instruments for data acquisition, massive archives of observational data, community application frameworks, and collaboration tools for routine use by researchers. Research communities and disciplines should be able to prototype, refine, develop, and deploy community-specific distributed applications. Robust software (both cyberinfrastructure and application) must be developed, maintained, upgraded, distributed, supported, and in some cases (as in distributed middleware, data curation, and scientific computing) professionally operated. To make these tools and resources accessible across a wide range of academic institutions, we must create a "grid" that provides convenient access to distributed resources, both in the United States and internationally.

Two very important principles that the Panel would like to maintain are:

- The high-end scientific computational resources available to the United States academic research community should be second to none.

- NSF, in collaboration with other appropriate mission agencies, should take lead responsibility for creating and maintaining the crucial data repositories necessary for contemporary, data driven science. The definition of "crucial" will come from the research communities.

The resulting cyberinfrastructure will be much more comprehensive in function and scope, and will be utilized by many more researchers than past NSF infrastructure programs (with the possible exception of the Internet). To gain maximum benefit, it is crucial that NSF support not only the development, provisioning, and operation of cyberinfrastructure and applications, but also their use in the daily conduct of science and engineering research. While support of domain science and engineering research per se is outside the scope of the ACP, successful use in the conduct of this research does require adequate professional staff to provide advice, assistance, and technical support, and these services are within the scope of the ACP.

To achieve the greatest benefits and broadest use, and also to work against Balkanization that inhibits interdisciplinary collaboration, commonality must be captured across disciplines, solutions for common issues identified and solved, and interoperability facilitated through standardization and the choice of common technical solutions. This is another important budgetary priority for the ACP.

The charge to this Panel included the request to “recommend an implementation plan to enact any changes anticipated in the recommendations for new areas of emphasis.” In the following, the budget requirements of this broad spectrum of activities are estimated. In the course of describing these needs, we supply additional recommendations and detail an implementation plan.

6.2 Budget Summary

A high-level summary of the budget is given in the following table. Later subsections describe each of these activities in greater detail.

Estimated annual budget	Millions of $ per year	
	Subcategories	**Total**
Fundamental and applied research to advance cyberinfrastructure		**$60**
Research into applications of information technology to advance science and engineering research		**$100**
Acquisition and development of cyberinfrastructure and applications		**$200**
Provisioning and operations of cyberinfrastructure and applications		**$660**
Computational centers	$375	
Data repositories	$185	
Digital libraries	$30	
Networking and connections	$60	
Application service centers	$10	
Total		**$1020**

These amounts are meant to be in addition to the current NSF investments in these areas, with the exception of the $375 million per year for "computational centers," which does include the current level of funding of approximately $75M/year. The funding described here augments, leverages, and creates incentives for exploiting commonality in the cyberinfrastructure investments already underway in the various NSF directorates. These funding recommendations are for NSF programs only, and presume that other federal agencies and institutions will continue to invest in related research and development. The ACP would increase its funding level as the program is defined and implemented. We estimate a credible ramp up to $545M/year of additional funding over two years and to the full $1020M funding in three years.

6.3 Discussion of Budget Categories

This section provides additional information and justification for the budget estimates. Our primary methodology was to estimate, in each category, how many individual projects and centers meet the goals of the program, and what average level of funding would be appropriate for each in order to reach a desirable critical mass. These "per project/center" costs are estimates and are average annual budgets, not upper bounds, and we would expect a range of actual expenditures around this average.

When budgetary components such as centers, research activities, equipment, and data repositories are described separately, they are not necessarily meant to be freestanding entities. They are elements of one overarching integrated, multidisciplinary, systemic program. It would be appropriate to co-locate and put some of these under a common management umbrella, thus benefiting from increased economies of scale and aiding overall coordination. For example, disciplinary-based data coordination projects may be affiliated with one of the data repositories, and large-scale operational centers may house substantial software development and deployment projects. In addition, we sometimes describe projects, and these may or may not be organizationally located within centers.

While we use the number of centers as one element of a budgetary estimate, the Panel generally provides a range rather than advocating a single hard number. Details at this level should be based on substantial analysis and community input, taking into account a number of factors, including existing resources, economies of scale and scope, availability of appropriate sites and institutions, and the willingness and ability of the community to establish and manage such activities. The actual outcome may reasonably differ materially from our recommendations.

The Panel does, however, feel strongly about several points:

- The existing centers (the leading-edge sites for the Alliance and NPACI plus PSC, and perhaps NCAR) have already accumulated significant expertise and experience relative to the ACP, and, subject to appropriate reviews, are likely to be among the initial sites;

- The supporting systems (data storage, high-performance computers, networks etc.) made available to United States academic researchers should be second to none, and

- There should be sufficient capability (scientific application performance, memory size, I/O speed, etc.) and available job time on such systems to support dozens of qualified groups conducting high-quality and high-impact research utilizing these systems.

Each of the major budget categories will now be discussed further.

Research to advance cyberinfrastructure - As discussed in Section 4, cyberinfrastructure is a system incorporating many processing, storage, and communication technologies, as well as large amounts of software. It encompasses the many roles discussed in the Section 2, principal ones being the sharing of common resources, functions, and expertise among institutions and disciplines, as well as lowering the barriers to entry for the development, provisioning, operations and use of new applications.

From a budgetary perspective, there are significant challenges and opportunities that demand research. Significant advances are required in human-computer interaction, database systems, software engineering, networks, parallel computing, advanced architectures, security, reliability, interoperability and many other areas. While many present and future technologies can be acquired commercially and must be intelligently leveraged, these often do not meet the specialized needs of science and engineering research. Because these needs are often high-end and stretch available technologies, there is a significant opportunity to leverage the ACP to advance information technology itself, one of the important missions of NSF. Both the Internet and supercomputing architectures are historical illustrations of this process of turning the needs of academic researchers into valuable new technologies while simultaneously empowering the research community.

The cyberinfrastructure also raises numerous social issues, for example, those related to security, privacy, intellectual property, and use of information technology in support of research communities in collaborative work across distance, organizations, and disciplines, and associated new modes of scholarly communication. Research into these issues will also pay numerous dividends, both within the NSF community and in the nation as a whole.

Thus, the ACP requires significant basic research activities that address both the technical and social challenges as well as opportunities that surround the construction, management, and use of the nation's evolving cyberinfrastructure. The ACP must also evaluate the outcomes and support the evolution of the cyberinfrastructure to meet ever expanding needs.

Although a portion of these funds should support individual investigators exploring ground-breaking new activities, we also envision a number of larger multi-investigator projects that explore many technical and social issues and mixtures of the two, and involve substantial prototyping, testbeds, and experimentation. Each larger project needs substantial funding, averaging about $2 million annually. Past examples of this type of project include the Titanium[47] Compiler Project at UC-Berkeley, the Storage Resource Broker[48] project at SDSC, the DataCutter[49] project at Ohio State, and the Network Weather Service[50] Project at UC-Santa Barbara. This is also in line with large projects in the ITR program, which we view as successful in bringing together interdisciplinary teams addressing similar issues and we hope will continue as part of the base budget. We estimate conservatively that 30 to 40 projects would be needed to cover the breadth of research issues related to the proposed infrastructure, from making it usable to making it secure.

In our budget estimate we assume 30 projects, for a total of $60 million annually, spent largely on researchers, equipment, and supporting professional staff. Some appropriate and evolving level of these funds could be allocated to individual investigator grants keeping in mind that the CISE base budget will also support many such grants.

Research into the application of information technology to domain science and engineering research - The goal of the ACP is to revolutionize scientific and engineering research through the innovative application of the information technologies. While cyberinfrastructure is an important enabler for this to happen, the ACP also requires researchers within the domain-specific science and engineering research communities to collaborate with computer and information scientists and mathematicians and social scientists in identifying opportunities, refining these ideas through experiments and trials, and ultimately moving these application ideas into production, broad deployment and use. It also requires research into generic applications that span disciplines, and identification of common threads across applications that can be captured within the cyberinfrastructure.

This type of investigation allows research communities to take advantage of the new information technologies and infrastructure, as well as support development of new methods and facilities to tackle research challenges previously out of reach. This research will involve long-term efforts in the science and engineering disciplines, computer and information science, the social sciences, and mathematics. We envision discipline scientists partnering with colleagues from other fields who can contribute to devising technical approaches to advance knowledge in new ways.

Once opportunities have been identified, they should be prototyped and introduced to real users, who will provide feedback to guide refinements and improvements. Ultimately, success will be measured by turning these applications into production software that is broadly adopted and used, and, importantly, many associated new processes and methodologies for the conduct of science and engineering research.

This activity should include a mixture of individual investigator and larger-scale grants or cooperative agreements. Turning successful prototypes into production, and the development of prototypes themselves, may call for partnerships with operational centers which offer expertise in software engineering, especially as these applications are turned over to production. On the other hand, one goal in advancing the cyberinfrastructure is to make it easier to develop and support new applications directly within application groups and disciplines. The distribution of grant sizes and types will likely vary by discipline. Successful models include the Grand Challenge awards of the mid 1990s and the application-oriented ITR grants of recent years. The large number of worthy but unfunded ITR proposals in recent years is a strong indicator of latent interest.

Our budget estimate is based on 50 grants at an average annual funding of $2 million, but also with considerable variation in grant size depending on discipline and the problem being tackled. Experience with application-oriented large ITR grants (roughly $2M-$3M/year for up to five years) has shown that some complex applications require substantially more funding. Some of these grants are large because of their interdisciplinary and inter-institutional character and the substantial needs for facilities, prototyping and experimentation, and supporting professional staff (software engineers, system administrators, user support, etc.).

Acquisition and development of cyberinfrastructure and applications - As the ACP evolves, increasing levels of support will be required for the development of production software, coupled with the licensing of commercial software components and the integration of the various custom and commercial components. Successful cyberinfrastructure and applications, as they move out of the prototype and experimentation stage, will require initial product creation, ongoing maintenance, upgrade, distribution, and user support. Where possible, any cyberinfrastructure and application software that is developed within this ACP should be subsequently commercialized, resulting in (hopefully) lower commercial licensing fees.

Cyberinfrastructure to support the myriad scientific and engineering applications will comprise many software tools, system software components, and other software building blocks. Examples of system software include grid middleware, parallelizing compilers for a variety of machine architectures, scalable parallel file systems and distributed databases, and sophisticated schedulers. Where appropriate these components will be commercially licensed, and NSF will purchase a "site" license on behalf of the community of NSF researchers.

An important activity will be an ongoing effort to identify the appropriate mix of commercial custom-developed software in accordance with an overall architectural plan, and then to acquire or develop and integrate these components. The outcome should be a single unified software distribution that users can download and install. Alternatively, centers will provision and operate this cyberinfrastructure and applications and offer them as services invoked over the network.

The NSF Middleware Initiative is exemplary of the type of program required to create and support the software aspects of cyberinfrastructure. While only a fraction of prototypes will require conversion to production status, the development costs of achieving the levels of stability and usability suitable for the larger community will require a development cost at least an order of magnitude greater than a prototype. The recurring costs of maintenance, upgrade, and user support will also be substantial. An active program to commercialize successful cyberinfrastructure and applications (especially the generic variety) will help to contain these costs.

These software development efforts would be supported wherever the expertise in computational science and software engineering is located, not just in large academic centers, and possibly in the commercial sector. Selection of the software development and maintenance groups should be based on expertise and experience, proposed plans and methodology, and anticipated costs. We estimate initially 20 such projects with an average annual budget of $5 million each.

We propose the creation and support of "cyberinfrastructure software centers" dedicated to developing the more difficult and sophisticated system and infrastructure software. These centers must have a scale necessary to attack the significant challenges of developing standards and production software for grids, programming tools, and data access and analysis, to name a few examples. Each center might employ on the order of 50 full-time-equivalent staff who would engage in professional software engineering, with a funding level in the $10M/year range. Ten such centers funded at this level would be a good starting point, with each center attacking one area, such as grid computing, compilers and runtime systems, visualization, program development environments, global scalable and parallel file systems, human computer interfaces, highly scalable operating systems, system management software, and so forth.

Provisioning and operations of cyberinfrastructure and applications - Whether software is acquired or developed, once it's integrated into a single distribution there are many operational issues to be addressed. Software to be downloaded and installed locally will need to be maintained (possibly) on multiple platforms, made available for download (including issues of authentication and access control), and supported through helpdesk facilities. Where services are provided over the network, the appropriate equipment and software must be acquired, integrated, installed, and operated, and again user support and helpdesk functions must be provided. While capital expenditures for facilities will be necessary, the bulk of the costs are recurring salaries for professional staff, including software engineers, system administrators and operators, and user support personnel. We anticipate that most of these activities will be conducted in centers funded under cooperative agreements with NSF.These needs can be broken down into several categories discussed in the following subsections.

High-end general-purpose centers

One class of centers will provide high-end computing resources, similar to the leading-edge sites of the current PACI program. These will feature some or all of the facilities currently found at such centers, including computers, large data archives, sophisticated visualization systems, collaboration services, licensed application packages, software libraries, digital libraries, very high-speed connections to a national research network backbone, and a cadre of skilled support personnel helping users take advantage of the facilities. Since the technologies deployed in these centers will be cutting-edge, the support staff also may have to develop software to provide missing functionality in the environment and to integrate the various resources and services.

Since progress in many science and engineering disciplines is paced by the capacity and peak performance of the available systems, as well as by the allocation and scheduling policies, there is need for both higher peak performance and higher capacity than currently available in the PACI program. The Panel strongly recommends the following principle: The United States academic research community should have access to the most powerful computers that can be built and operated in production mode at any point in time, rather than an order of magnitude less powerful, as has often been the case in the last decade.

The most powerful scientific computer in the world today is Japan's Earth Simulator System[20], with a peak speed of 40 teraflops (10^{12} floating point operations per second), built at a cost of around $400 million. DOE's Lawrence Livermore National Laboratory[51] (LLNL) has a 12 teraflops machine and its Los Alamos National Laboratory[52] is in the process of installing a 30 teraflops system. In FY 2004 (perhaps the first year of the ACP) at least one of the DOE laboratories is expected to install a system in the 60 – 100 teraflops range. All these systems have been justified and are being used by a relatively small number of applications projects.

The Panel believes it is important that NSF make comparable systems available to the United States academic community, but, due to the large size and diversity of this community, such systems must support a much wider range of applications. If the U. S. academic community is to be competitive internationally in large-scale simulation, these considerations suggest systems in the 60 teraflops range in FY 2004, thereafter tracking the state of the art. In addition, at least a dozen individual American universities have acquired or are installing systems with peak speeds of over one teraflops. In order to enable new applications, national resources should be more powerful than those at individual universities by at least one to two orders of magnitude.

In terms of capacity, there should be a sufficient number of such systems that individual projects (with appropriate justification) can be granted the resource units to run many jobs per year that use a large fraction (at least 25%) of peak performance for tens or hundreds of hours. Such jobs usually access or produce vast amounts of data that need to be stored, visualized, and interacted with; hence, the entire environment needs to be balanced and scaled according to peak processing speeds. A typical balanced configuration meeting this criterion would have:

- At least 1 Byte of memory per FLOP/s.
- Memory Bandwidth (Byte/s/FLOP/s) ≥ 1.
- Internal Network Aggregate Link Bandwidth (Bytes/s/FLOP/s) ≥ 0.2.
- Internal Network Bi-Section Bandwidth (Bytes/s/FLOP/s) ≥ 0.1.
- System Sustained Productive Disk I/O Bandwidth (Byte/s/FLOP/s) ≥ 0.001.
- System High Speed External Network Interfaces (bit/s/FLOP/s) ≥ 0.00125.
- The internal network that connects the nodes with latency in the 1-2 microseconds or less, user memory to user memory.
- Globally addressable disks with at least 20 times the capacity of main memory.

Using those ratios, a 60 teraflops system with a balanced configuration would have 60 TB of memory and 1.2 PB of globally addressable disk space. Current estimates are that in FY 2004 such a system will cost on the order of $180 million. In FY 2007, $180 million might suffice to purchase a balanced system with a peak speed of 100 to 150 teraflops.

The panel recommends that about five such centers be supported; the two leading-edge sites of the PACI program plus the Pittsburgh Supercomputer Center should be considered as three of these centers, following appropriate review. While there are substantial economies of scale in operating large computers – a modestly larger staff can support a much larger computer or several systems – there are other considerations in the number of centers. Each center tends to develop affinity with different disciplines or strengths in different aspects of information technology. Centers are training grounds for computational scientists and engineers, who then migrate to (more likely nearby) research institutions. A greater diversity of centers encourages novel approaches and new ideas. The primary measure of effectiveness of such centers is user satisfaction, and competition among a larger number of centers leads to greater satisfaction. On the other hand and as mentioned earlier, the number of centers is secondary and should in the end be based on additional analysis and community input.

There should be no shortage of institutions interested in creating and operating large-scale centers; over a dozen universities already operate substantial centers and have participated in previous competitions. For the purpose of the budget estimate we assume five centers, each with an annual budget of $75 million, for a combined annual budget of about $375 million ($300 million more than the current level). This is larger than the current centers primarily because we advocate higher-peak-performance and capacity computers and ancillary systems than at present. On the order of $50 million annually would be devoted to these equipment procurements, assuming that a major new system will be acquired by each center every three to four years. Most of the rest of the budget would be for recurring personnel costs; development, integration, maintenance, and upgrade of software; as well as provisioning and operations of cyberinfrastructure and user support.

In staging the operational portion of the ACP, in FY2004 and FY2005 (after appropriate review) the existing centers might acquire upgraded facilities and related infrastructure. (Spreading the ramp-up over two fiscal years will provide more choices and may increase performance as new generations of systems emerge). The second step might be to open a competition in FY2005 and FY2006 for additional centers.

Local clusters of computers are meritorious alternatives to centralized large systems for many needs, and in some research areas special-purpose hardware is the best option.   The ACP will contribute to the creation of a grid environment (including middleware and tools) that will make all three options accessible to researchers at all institutions and facilitate the migration of applications from one to another. As with the number of centers, the balance of funding among these options should be based on additional analysis and community input.

A similar issue arises with professional support personnel. Budgets should seek (based on prior analysis) to achieve the best balance between local support (which can give more discipline-specific and intensive assistance) and centralized support (which benefits from economies of scale and scope and can usefully transfer expertise from one institution to another and from one discipline to another). A valuable middle ground is to locate discipline-specific groups at large centers.

Data repositories

Well curated data repositories are increasingly important to science and engineering research, allowing data gathered and created at great expense to be preserved over time and accessed by researchers around the world, including by disciples of other disciplines. The ACP should provide long-term and sustained support of such repositories. This involves much more than simply running large storage facilities. Supported by research into cyberinfrastructure, better ways to organize and manage such large repositories will be developed, and software infrastructure and tools will be developed, distributed, maintained, and supported. Appropriate standards will be developed that allow data to be self-documenting and discoverable through automated tools, and to insure the interoperability necessary to incorporate data acquired in one discipline into applications serving other disciplines.

To illustrate some detailed issues, data need to be organized in appropriate ways, metadata (machine readable and searchable descriptions of the data) must be systematically created, and basic manipulation and analysis tools provided. Data must be structured in ways that support both intra- and inter-discipline interoperability. Useful data repositories are also highly dynamic, requiring reclassification based on reanalysis of content. Migration of data to new media for preservation, and exploitation of higher capacity media is required. High-speed access to repositories by remote users raises capacity and scalability issues, with implications for their network, storage, and I/O subsystems.

As with computing, the cost of data repositories (done correctly) will be dominated by the recurring costs of personnel performing curation, maintenance and upgrade, and providing user advice, assistance, and support. The most sophisticated of these personnel need professional skills in the relevant aspects of information management and information technology (e.g., data bases, archival file systems, building portals), and will be developing and maintaining custom software. By using a combination of high-speed networks and local high-speed caches, there is no hard requirement to co-locate professional staff with physical storage particularly staff performing data acquisition and curation functions as opposed to disk partitioning, regeneration, and backup functions. As with computing, there is need for support personnel at local institutions, in discipline-specific groups (often located in centers), and centralized in centers. Although further analysis is needed, we expect that the most efficient approach will be to have relatively centralized storage hardware (with supporting staff) but distributed data acquisition and curation personnel. The balance of funding across these options should be determined by analysis and community input.

The challenge of data acquisition, curation, and access cannot be addressed solely by NSF, since other agencies in the United States and internationally also support repositories. For example, NIH supports certain biology and biomedical data collections and NASA funds many archives of astronomy and remote sensing data. NSF should support repositories for a number of disciplines, such as astronomy, atmospheric and oceanic sciences, biology, biomedicine, climate modeling and observations, engineering of many variations, environmental and earth sciences, geophysics, high-energy physics, neuroscience, nuclear physics, and space sciences, among others. One can easily envision 50 to 100 such repositories. Indeed, a Web search quickly yields scores of existing repositories, many of which will not scale to future demands, interoperate well among disciplines, nor guarantee long-term access. Based on current experience, each repository will require $1.5 million to $3 million annually, not even including the substantial additional effort required to produce clean, well-documented data that retains long-term access and value. Overall these repositories may require $150 million annually, assuming 75 such repositories with an average yearly budget of $2 million. The number of physical locations for storage farms and supporting personnel may be considerably smaller than the number of disciplinary repositories, based on analysis of tradeoffs between community responsiveness and the availability of discipline-specific expertise vs. economies of scale and scope.

In addition, it is important to maintain ongoing development centers that address issues spanning all disciplines and ensure that the latest outcomes from the research community (including research funded under this ACP) and the commercial sector are applied to the expanding data storage and management challenge. These centers would be primarily responsible for spreading the latest technologies and best practices and insuring interoperability across disciplines through appropriate standardization. They are the primary point of connection to the computer and information sciences research communities (including the digital library, knowledge management, and knowledge mining communities) for the derivation, description, and management of the knowledge derived from computations and observational data. We recommend that approximately five such centers be established at an estimated cost of $3 million per year each, for a total of $15 million per year. In some cases these centers may be co-located with significant data repositories.

The Panel also recommends the creation of teams that would work on discipline-specific metadata standards, data formats, tools, access portals, etc., as well as help to select and install software, e.g., for the grid and databases. If one such effort is supported at $2 million per year for each of the ten disciplines listed above, a combined funding level of $20 million per year will be required.

Digital libraries

An integral component of cyberinfrastructure includes the nation's digital libraries, an area where NSF is already providing intellectual and organizational leadership. These libraries contain (much more so in the future than today) our intellectual legacy, a fundamental resource for our scientific and engineering research and engineering practice.

NSF digital library initiatives have created new infrastructure and content of value to specific disciplines (including many in the humanities). It is important to continue such efforts through ongoing research, prototyping and experimentation with digital library technologies, development and deployment of proven solutions, and support for specific digital library repositories in disciplines represented at NSF. The potential has been barely tapped, and there is an opportunity to find and implement new mechanisms for sharing, annotating, reviewing, and disseminating knowledge. We suggest that the topic of digital libraries be broadened to consider even larger questions about the transformation of scholarly communication, including not only the accessing and sharing of knowledge, but also including this expanding knowledge as an integral element of the active collaboration among scholars.

The soon-to-conclude second phase of the NSF digital library initiative is investing about $10 million per year. Given the success of the initiatives, and the promise and critical importance of the area, we believe the budget should be at least $30 million per year for digital libraries activities, with a mix of project sizes from $1-3 million annually.

Networking

High-speed networks are a critical cyberinfrastructure facilitating access to the large, geographically distributed computing resources, data repositories, and digital libraries. As the commodity Internet is clearly not up to the task for high-end science and engineering applications, especially where there is a real-time element (e.g. remote instrumentation and collaboration), a high-speed research network backbone should be established and the current connections program extended to support access to this backbone as well as to provide international connections. Today we could aim for a 40 Gb/s (gigabit per second throughput) backbone with large center or user sites connecting at 10 – 40 Gb/s. Over time these numbers could increase rapidly with advances in technology and sustained funding. Assuming that 50 sites connect at 10 Gb/s and 40 sites at 40 Gb/s, a cost estimate of the backbone and connections is about $60 million per year.

As with computing, the primary issues in the backbone network are peak speed of data transfer and total capacity. The peak speed should be determined primarily by currently available production network equipment, and capacity upgrades will require ongoing monitoring and analysis to avoid significant congestion-induced communication latencies. However, from the perspective of applications and users, the performance of the backbone network is only one element of overall performance, which is also affected by local area networks, various processing and caching bottlenecks, processing delays in middleware and operating system layers, and computer I/O bandwidths, among others. For this reason, the research and development addressing performance issues within the ACP should focus its attention on overall system performance, seeking out bottlenecks and removing those bottlenecks through research into underlying technology advances, system architectures (e.g. the strategic location of caching), and development of more advanced hardware and software solutions. The adequate funding of facilities upgrades and the funding of these research and development activities are equally important in providing the research community with state-of-the-art facilities.

Within this operations portion of the ACP, system measurement instruments and software should be deployed, a knowledge database of issues and solutions should be developed and maintained, and professional support staff should advise, assist and support researchers and applications developers encountering difficult performance issues.

A number of unique scientific facilities utilized by U.S. science and engineering communities are located outside the United States -- some even funded by the NSF (e.g. the Gemini South Observatory[53] in Chile). As noted elsewhere in this report, international collaboration is essential in research, and the United States has a vital interest in ensuring that its science and engineering community has high-speed access to the international infrastructure. NSF needs to connect the national backbone to similar infrastructure in other countries, and cooperate in other ways through research, development, standardization, and operations.

While these budget estimates may seem low, as throughout this section, this estimate is in addition to current NSF network research and infrastructure networking expenditures (as we have specified throughout this section), which is currently about $40 million annually for networking infrastructure. We also expect that individual states (e.g., California, Illinois, Indiana, and North Carolina) and individual universities will make coordinated investments to ensure that institutional infrastructure provide appropriate connectivity from the national backbone all the way to researchers' desktops.

Application service centers

In the budget categories already addressed, there are clearly some unmet needs, such as support services for non-computational applications, visualization, collaboration, or distributed and cluster computing, among others. These services may be provided through a combination of a utility model (making them available on the network) and by providing software distributions and support personnel to aid in their installation and use. As the research and development portions of the ACP yield successful outcomes, the needs in this area will expand. Initially the Panel recommends funding a modest number of centers to initiate these activities (five with an annual funding of $2 million each would be reasonable). Over time this budget (and the size and number of centers) would grow, guided by the successful models and expanding needs and opportunities.

6.4 Summary

The scope and scale of the ACP will require an annual budget of about $1 billion over and above the current PACI and network infrastructure programs in NSF.

We estimate that about 65% of the total budget is for the recurring costs of professional staff and researchers, as opposed to the acquisition of hardware and software. A substantial portion of these recurring costs is devoted to developing, maintaining, distributing, upgrading, and supporting software. This emphasis is consistent with past President's Information Technology Advisory Committee[33] (PITAC) recommendations for substantially greater investments in software research and production.

The implementation recommendations and budgets sketched in this chapter are based on experience with related projects and activities and reflect numerous comments and suggestions received from community leaders. Nevertheless, our recommendations should be considered as only a beginning. The ACP will require ongoing planning, implementation and adequate resources if it is to achieve its goal of revolutionizing the conduct of science and engineering research. All NSF directorates must participate in the planning and in the implementation in order to ensure that the cyberinfrastructure that is built is effective in bringing about this revolution.