3.0 Trends and Issues

In this section we describe the converging and reinforcing trends and issues derived from our surveys, testimony sessions, readings, and deliberations; these formed the basis for the vision, findings, and recommendations presented in the preceding section. As mentioned earlier and illustrated in Figure 3.1, the ACP opportunity derives from a combination of the push of technology trends and the pull of vision and needs for its application in research communities. The impact of these trends is not necessarily linear. As certain thresholds of functionality or price-performance are crossed, disruptive changes occur. The trends may also reinforce one another, magnifying their impact.

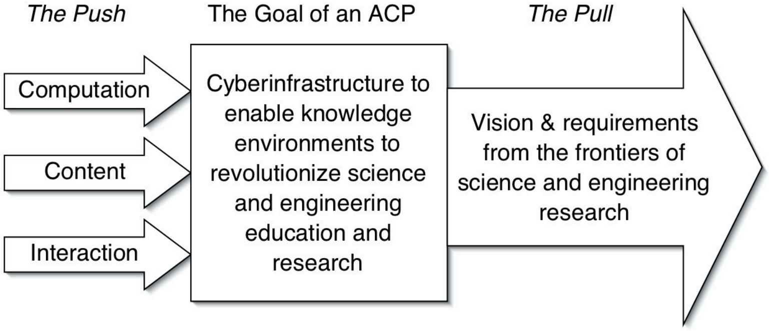
The Push
The Goal of an ACP
The Pull
Computation
Content
Interaction
Cyberinfrastructure to enable knowledge environments to revolutionize science and engineering education and research
Vision & requirements from the frontiers of science and engineering research

Figure 3.1. The push and pull for an ACP.

We have clustered these trends and issues into three areas: *computation*, *content*, and *interaction*. The substrate for all of these trends is the familiar exponential increase in the capacity of the base computation, storage, and communication technologies.

3.1 Computation

The measures of computing, networking and storage capacity continue to grow geometrically. We take for granted that computer speeds will rise radically with each new hardware generation, that machines will have more memory than before, that disks will hold ever more information, that the network will be faster, and that partially as a result software will provide ever more complexity and features. We should not hit physical limits for current basic chip and disk technologies before 2010 (and probably much later), so we assume continuation of this golden age of information technology through the period addressed by this report. (Consideration of other technologies such as quantum computing is beyond the scope of this report, but research underway suggests that technology may move onto even higher performance curves in the future.)

As we have ridden these smooth exponential curves for several decades, what has changed? We have passed several practical thresholds, resulting in qualitative breakthroughs. Scientific research that would have been prohibitively expensive or previously demanded national-scale resources can be done in local facilities. Workstations can now do computations that only the biggest and most expensive supercomputers could attack a few machine generations ago. Thus, serious computations demanding real-time visualization, simulation of interactions of thousands of particles, and 2D- and even 3D fluid dynamics are possible on the desktop. Combining commodity hardware (PC boards and networks) into a laboratory cluster permits computations that only national labs could attempt a decade ago. The entire scientific literature can fit on a few hundred disks, with material costs under $25K. (Disk storage became cheaper than paper years ago and is also competitive with microfilm.) There are individual civilian laboratories and state universities that are installing computers in the teraflops range and data server clusters in the 100 terabyte range.

But the demand for highest-performance computation is also increasing, and thus we continue to need a hierarchy (or “pyramid”) of connected computation resources of varying capacity and cost. In a few more years, we will cross the “peta” (10^{15}) line: there will be some supercomputers in the 0.1-1 petaflops range, some scientific databases will exceed 1 petabyte, and networks will exceed 1 petabits/s.

Hardware components – The hardware components underlying computation, storage, and communication have been improving exponentially (at a compound rate of growth) for many years; this is expected to continue over the scope of this report. Although the rate of growth of circuit speed may begin to flatten, major research directions have potential to break these barriers. The current speed growth directions are depending less on circuit speedup, and more on increasing circuit density and the number of parallel processing units on a chip or wafer. We will achieve petaflops not with a femtosecond clock, but rather by having a million processors on a nanosecond clock (give or take a factor of 10).

Processing – Computer speed is usually expressed as arithmetic calculations, or floating-point operations, per second (flops). In 1999, two machines in the world had a theoretical capacity of 1 teraflops. By now we estimate a dozen universities and laboratories have or have ordered computing clusters with theoretical capacities exceeding 1 teraflops, and by 2005 machines up to 10 teraflops will be relatively commonplace (a teraflops machine may even be affordable for some individual researchers). These changes are due to continued improvements in chip technology and the ability to utilize clusters of chips and mass-produced computers. We benefit from not only parallelism, but also speed; in late 2002, a clock rate of 350 GHz was announced for a silicon-based experimental device.

Storage – Many applications depend on manipulating masses of data, far more than can reside inside the processors. These data can be observational inputs, experimental values, or results of calculations, images, or videos. Such information is usually kept on disk (though the largest archives are stored on removable optical disks or magnetic tapes). The highest performance (measured variously as total number of characters of information stored, number of characters per volume of lab space, or number of characters retrieved per second) is generally found in the most recent commercial disks. Increasing overall storage capacity comes from utilizing many disks to store massive amounts of information and accessing them in parallel.

Disk capacities (measured as bits per square inch of magnetic material) have historically increased at 60% per year, but in the past few years bit storage density has increased by about 100% per year. Prices of individual units have fallen more slowly, so most of the economic improvement has come from larger capacities. The most capacious disks in late 2002 store about $3x10^{12}$ bits (320 gigabytes, or 0.33 terabytes) of information. Databases of a few terabytes are common; only ones over 100 terabytes are considered remarkable.

Networks – A major shift in computing has come from the practical availability of high-bandwidth data networks. Network connections up to 45 megabits/s are easily available, connections over 155 megabits/s are still aggressive, and some research institutions are beginning to connect at 2.5 gigabits/s and faster. Available technology can support far higher bandwidths. Deployments have already demonstrated 1.6 terabits/s on a single fiber (40 channels at 40 gigabits/s), while laboratory experiments have reached over 11 terabits/s. Switching data at these speeds remains relatively expensive, but technologies have been demonstrated.

Network researchers and providers are also introducing a new paradigm –*optical networks* based on the emergence of an optical layer, operating entirely in the optical domain (and avoiding electronic bottlenecks), to enable very high capacity end-to-end wavelength ("lambda") services that provide (through wave division multiplexing) many virtual fibers on a single physical fiber. Optical networks, for example, are being explored for linking widely distributed high-performance machines together in grids. Optical networking is an important emerging technology to explore and use in the ACP.

These improvements make it plausible to move huge files between sites, so that computing and storage facilities can be split or combined in a number of ways. However, the speed of light (~1 ns/ft, or about 20 ms to cross the U.S.) is not increasing, and networking switches add further delays. This puts a fundamental limitation on the use of widely dispersed processing and storage resources for tightly coupled computations and is one of the reasons that supercomputers remain indispensable for many scientific applications.

In-building networks are improving in two ways – bandwidth and mobility. Local area network (LAN) technology is now moving to high-speed Ethernets able to deliver 100 megabits/s or 1000 megabits/s to the individual server or desktop. Few current computers can handle such data rates effectively, nor can typical laboratory switches manage many full-speed streams, but this situation will improve rapidly.

The use of wireless (radio) access to the network is exploding, both within buildings and in general public uses. Very local access (using for example the IEEE 802.11 family of standards) can provide many megabits per second to a single device (laptop or PDA), and new generations of cellular telephone technology will permit 0.1-1 megabits/s to the roaming device in the next half dozen years. This has great promise for many mobile applications, such as gathering scientific data in the field and geographic-independent group collaboration.

Displays –Typical commercial displays offer about 1 square foot of useful visual information and present around 1 megapel (million picture elements). This is another technology that is rapidly advancing. Many labs (especially those on the Access Grid) combine between 3 and 15 typical displays to present a single large image. Recent special displays have higher density and brightness; desktop devices with over 9 megapel are now commercially available. Displays are also configured to provide 3D and immersive virtual reality experiences in CAVES or ImmersaDesks. Costs continue to decline and very useful 3D interaction is now available below $10K.

Provision and use of high performance computing – World leadership in the highest-capacity computing has been, and continues to be, a significant factor in research and national security. The federal government has been the primary investor in, and user of, the highest capacity machines. Mission agencies such as the Department of Energy and the Department of Defense have used supercomputers in mission-specific domains, including some use in basic research. The NSF, however, is specifically charged with fostering and supporting broad development and use of computers and other scientific methods and technologies for broad research and education in the sciences and engineering.

As computers evolved, various NSF directorates supported research in components, theory, software, systems, and applications of computers. An Advanced Scientific Computing (ASC) Program, situated in the Office of the Director, provided the NSF research community access to the highest-performance supercomputers of the day. In 1985, ASC activities and several other programs were merged into the Directorate for Computer and Information Science and Engineering (CISE). CISE supports investigator-initiated research in all areas of computer and information science and engineering and also supports high-performance national computing and information infrastructure for research and education generally. It has done this through co-investment in computational infrastructure in academia, and at the high end, through a series of centers and alliances. The development and operation of high-performance computational centers was also instrumental in the creation of the NSFNET, the precursor of the commercial Internet. In addition, the recommendations of the 1995 *Hayes Report (Report of the Task Force on the Future of the NSF Supercomputer Centers Program)*[36] along with the predecessor *Branscomb Report*[37] *(NSF Blue Ribbon Panel on High Performance Computing)* formed the basis for the development of the Partnerships for Advanced Computational Infrastructure (PACI) program.

Two PACI[2] partnerships established in 1997 are currently operating under the principles set forth in the Hayes Report by (1) providing access to high-end computing, (2) affording knowledge transfer of enabling technology and applications research results into the practice of high-performance computing, and (3) supporting education, outreach and training activities. Each partnership consists of a leading-edge site, the National Center for Supercomputing Applications in Urbana-Champaign and the San Diego Supercomputer Center in San Diego, and a significant number of partners. The highest-capacity machines are located at the two centers in Champaign-Urbana and San Diego, and they are networked with various other mid-level performance centers at other universities.

More recently the NSF made awards for terascale-capacity facilities to the Pittsburgh Supercomputing Center[5] and for the Distributed Terascale Facility[6] (providing tera*grid* capacity) to a consortium including National Center for Supercomputing Applications (NCSA) at the University of Illinois, Urbana-Champaign; the San Diego Supercomputer Center (SDSC) at the University of California, San Diego; Argonne National Laboratory in Argonne, IL; and the Center for Advanced Computing Research (CACR) at the California Institute of Technology in Pasadena. In October 2002, the Pittsburgh Supercomputer Center (PSC) was added to the Terascale Facility.

This evolution of high-performance computing programs at NSF is at the leading edge of evolving architectural diversity in high-capacity computing. In earlier years, the fastest computers used fundamentally faster components (newer technologies, higher cooling and powering, more complex processor designs). The current state is different – the fastest chips are now also among the most common, and they have very complicated internal structures. Only very specialized problems currently benefit from use of nonstandard parts. (Some of the most technologically impressive processors are found in game machines.) The commercial world continues to demand more computing power, and this huge demand for machines supports investment in new manufacturing processes and designs. High-end computing now depends more strongly on combining very large numbers of these commercially available devices, rather than trying to make unusually fast individual processors.

Parallelism is a recursive notion. Single-chip microprocessors using various forms of internal parallelism are the heart of a computational *node*. For much greater speedup, nodes are combined through switches into physically proximate (to minimize speed-of-light delays) *clusters* of nodes. Now, cluster supercomputers are being distributed over high-speed networks to form *grid computing* environments. The Terascale Initiative is building a large, fast, distributed infrastructure for open scientific research. When completed, the TeraGrid will include 20 teraflops of computing power connected at 40 Gb/s over five geographically distant sites. It will also include facilities for storing and managing nearly 1 petabyte of data, high-resolution visualization environments, and toolkits to support grid computing.

The demand for advanced computing is no longer restricted to a few research groups in a few fields, such as weather prediction and high-energy physics. Advanced computing now pervades scientific and engineering research, including the biological, chemical, social, and environmental sciences. However, the entry barrier continues to be very high. Numerous of our survey respondents observed that, in some areas, the state of the art in computer technology is outpacing tools and best practices from the user perspective. For example, the relatively straightforward and efficient autovectorizing and autoparallelizing compilers of the previous hardware era have given way to complicated messaging directives that must be inserted manually; to many users these are as intimidating and time consuming as programming in assembly language. Industry and academia should work together to remedy this problem and bring greater parity between the available facilities and the tools available for their use.

This issue becomes even more important with the move toward Grid-based capabilities. There is growing mismatch between *theoretical peak* and *actually realized* performance for production codes, as well as a growing investment of time required for users to achieve reasonably good performance. Researchers commented that although the theoretical peak performance of current machines is much higher, they obtain a smaller fraction of theoretical peak today than 10 years ago for many applications. Greater effort is needed to automate the conversion of code for efficient execution on various machine architectures, including clusters and grids, and to minimize massive code changes as the underlying machines evolve and change.

Many respondents to the Panel's Web survey (details of the survey are in Appendix B) indicated the importance to their work of research-group and departmental-scale computing facilities. We define such facilities as having a factor of 100 to 1000 less capability (e.g., computing, storage) than is provided by the national-scale centers. The proliferation and importance of such resources suggest the need for an effective mechanism – now lacking – to create, nurture, and support them as well as link them into the national cyberinfrastructure. Further, the results suggest that users view national centers as needing to provide capability of order 100 to 1000 times the power of systems generally available to individual academic departments and research groups. Such centers not only dramatically expand the capabilities available to individual projects, but also ensure that all university researchers have equal opportunity. At this point the promise of grids of computers cannot replace the need for both local mid-level facilities and highest-end national resources. Grids are extremely valuable for some types of computations but fail for others because of network latencies and other reasons.

Scientific and engineering applications are covering and will continue to cover even greater time and space scales (e.g., weather, which involves a coupling of scales ranging from planetary waves that last for more than a week, and individual thunderstorms, which are at subcity scale and last for one to a few hours). Such multi-scale problems, often involving the coupling of different models, are exceedingly complex and computationally intensive and thus need sustained high-end computing for the foreseeable future. For example, emerging community climate system models require a sustained 25 teraflops and involve computations closely coupled and thus susceptible to network latencies. But this is only the beginning, as the earth science community moves to comprehensive, high-resolution simulations of combined biological and geoscience models of the environment.

Although many important problems require the highest available processing power, cyberinfrastructure should not concentrate solely on team projects using only the largest and most powerful resources. Rather, it should support a hierarchy spanning a pyramid of machine capacities and the spectrum from small grants to large multidisciplinary centers and projects. As was pointed out in testimony concerning the National Virtual Observatory[10], large team efforts are required to build federations of data and tools to explore them; but smaller groups working independently and given access to these data and tools can (and likely will) make fundamental discoveries.

The current NSF-supported centers remain largely a batch-oriented environment, whereas many future problems will require on-demand supercomputing for steered calculations and a dynamic environment where the machine needs to respond to the calculation (e.g., dynamic adaptive nesting and the ingest of real-time data that impacts a real-time calculation; such as adaptive sensors in field biology). The current centers are not configured and administered to provide, in most cases, significant fractions of their resources in a dedicated fashion to support the most challenging research problems. Although machines may have the *capacity* to solve huge problems, the users may not have the *capability* to use them effectively because of lack of support for mapping their code efficiently onto specific parallel machines or because of restrictive machine allocation policies.

We received frequent strong input to the effect that the National Resource Allocation Committee (NRAC)[38] allocation process is no longer effective and must be overhauled. For example, users are subjected to double jeopardy by having to prepare both research grant (agency) proposals and proposals for computer resources. Funding of the former with a negative decision for the latter clearly creates a problem. NSF considered coupling the two processes in the early 1990s but chose to leave them separate. Mechanisms for requesting resources should be streamlined as well, and the reviewer base must be broadened to ensure an adequate understanding of the needs being expressed. Moreover, the new allocation process will likely need to include additional types of resources such as federated data repositories and remote instruments.

Even more fundamental issues of resource allocation are intrinsic in cyberinfrastructure concepts of large interoperating grids of computers, instruments, and data repositories. Human committees will not be capable of doing the complex dynamic allocation processes required to balance the supply and demand over thousands of users, hundreds of machines, and numerous variations of computational size and requirements for real-time response. Automated allocation mechanisms are themselves a research challenge and another example of the need for social scientists – in this case economists – to participate.

Both the capacity and demand for high performance computing continue to grow in depth and breadth of use. There continues to be constructive diversity in how this computing is provided and the need for continued experimentation and investment in new machine architecture and supporting software: operating systems, middleware, and application frameworks. On the other hand we need balance (and better yet, real synergy) between extending the frontiers of computing and extending the frontiers of science using computing. The challenge is to both break new ground and bring current and new users along.

A note about sustaining access to highest end computing – The continuing exponential improvement of the hardware underlying cyberinfrastructure provides accelerating opportunities for exercising creativity but can be daunting in terms of managing the attendant rapid obsolescence of facilities. Maintaining leading-edge cyberinfrastructure requires continuing investment, not one-time purchase. Cyberinfrastructure ("bit-based") investments differ from most other, more "atom-based" kinds. Delaying the start of construction of an accelerator or telescope or research vessel normally increases the cost of the acquisition. Frequently, the opposite is true for computing equipment, which becomes cheaper by waiting a year but becomes obsolete soon thereafter. One way to quantify this is through replacement schedules. Major research equipment may have a realistic lifetime of 10-25 years. The appropriate replacement interval for information technology at the frontiers of performance is closer to 3-5 years. Furthermore, there are changes in the ways machines are used and the types of computations that are needed. As the basic unit costs of information and calculation fall, new ways to get better answers or to displace scientists' time are discovered, and the appropriate levels of local and national computing and the appropriate balance between them will change.

The scientific research world pushes the limits of a number of technologies and acts as a driver for improvement. Collaboration between high end users and commercial providers has been effective and should continue. But the commercial mass markets will continue to determine the computing equipment and services that are most readily available, including the best programming language implementations, fastest chips, and largest disks. The research world has driven very high end networking and the largest computing clusters. There are commercial organizations that specialize in running large computers and disk farms or in taking over entire business functions. They have developed tools and methods for efficient operation to exacting contractual service level agreements, so they provide benchmarks or alternatives for deploying some of the cyberinfrastructure.

3.2 Content

As familiar as the exponential growth in computing, storage, and networking power is the exponential growth in digital information and data. Most all scientific and technical literature is now created in digital form, and large quantities have been converted to digital retrospectively. Scientific, engineering, and medical journal publishing is now done in a hybrid of digital and paper formats with digital taking dominance, although pricing and terms and conditions for use continue as major issues. Some presenters to our panel expressed deep concern about the increasing price of commercially published scientific literature that is forcing academic libraries to collect a smaller and smaller fraction of the overall literature.

The primary access to the latest findings in a growing number of fields is through the Web, then later through classic preprints and conferences, and only after that through refereed archival papers. The traditional linear, batch processing approach to scholarly communication is changing to a process of continuous refinement as scholars write, review, annotate, and revise in near-real time using the Internet. Major research libraries have switched from microfilm to digitization for both preservation and access.

Crucial data collections in the social, biological, and physical sciences are coming online and becoming remotely accessible; modern genome research would be impossible without such databases, and astronomical research is being similarly redefined through the National Virtual Observatory.[10] Enormous streams of data are arising from observational instruments and computational models. The high energy physics community, for examples, estimates that by about 2012 it will need an exabyte (10^{18} bytes) archive for data from four major large hadron collider (LHC) experiments. The National Center for Atmospheric Research (NCAR)[28] currently has 1000 terabytes of online data and is growing at 10 terabytes per month.

The NSF CISE Directorate through a series of Digital Libraries Initiatives[8, 12] has been instrumental in grounding and informing the emergence of digital libraries in basic computer science and engineering. It has produced important subsystems and institutions and has created synergy among researchers, practitioners (libraries and archivists), and production organizations (libraries, archives, museums). It has enabled research to help define the possibilities, pilot projects to help validate and make concepts real, and partnerships and startups to create new production services. It is a good example of interdisciplinary research, focused by test bed construction, which is needed in a broader cyberinfrastructure program.

A significant need exists in many disciplines for long-term, distributed, and stable data and metadata repositories that institutionalize community data holdings. These repositories should provide tutorials and documents on data format, quality control, interchange formatting, and translation, as well as tools for data preparation, fusion, data mining, knowledge discovery, and visualization. Increasingly powerful data mining techniques are creating greater demand for access to cross-disciplinary data archives. Through data mining new knowledge is being discovered in problem areas never intended at the time of the original data acquisition.

Other trends include the growing need to confederate data from multiple sources and disciplines. The emergence of supercomputing environments capable of executing comprehensive, multilevel simulations (for example, of the environment) requires interoperability between both computational models and the associated observational data from various fields. It was mentioned at a recent meeting of the environmental research and education community that some scientists are spending up to 75% of their time finding and converting data from other fields. Much of the data being sought is “preserved” in ad hoc and fragmented ways, and all too often ends up in “data mortuaries” rather than archives.

Repeatedly the Panel heard members of the research community citing the need for trusted and enduring organizations to assume the stewardship for scientific data. Stewardship includes ongoing creation and improvement of the metadata (machine-readable and interpretable descriptions of the data itself) by people cross-trained in scientific domains and knowledge management. A key element associated with filling this need is the development of middleware, standard or interoperable formats, and related data storage strategies. Although each discipline is likely best suited to creating and managing such repositories and tools, interoperability with other disciplines is essential, through the creation and adherence to standards, and other means. Additionally, greater emphasis needs to be given to the digitization and stewardship of legacy data (data archeology) and to digital libraries preserving and giving access to past scholarly work.

More and more disciplines are also expressing a compelling need for nearly instantaneous access to databases (both local and distributed) as well as to high-speed streams of near-real-time data from observation and monitoring instruments. Applications such as numerical weather prediction models need to be used in control loops to drive the remote sensors to optimize the data actually being collected; the linkage between data acquisition and processing is now two-way. It is important to note, however, that the technologies for such databases do not yet exist and that many needs of the research community are not accommodated by existing systems (e.g., commercial relational databases). This is a concrete example of how software for large-scale scientific use must extend well beyond the procurement of commercial technology, and often even beyond our current understanding. Thus, both coordinated research into the information technologies and the development of customized technologies for the research community are needed.

Online scientific instruments, or arrays of instruments, are a growing source of digital content for both huge quantities of primary data and the derivative processed datasets. Modern large instruments such as supercolliders and telescopes produce huge streams of data as well as growing numbers of ubiquitous arrays of small sensors. For example, in air and water pollution or seismological monitoring, satellites continue to beam back huge data sets and a growing interdisciplinary community intends to examine practically every aspect of the Earth system from space this decade using data from these satellites.

The emergence of ubiquitous wireless networks offers another big opportunity. Billions of Internet connected cell phones, embedded processors, hand-held devices, sensors, and actuators will lead to radical new applications in biomedicine, transportation, environmental monitoring, and interpersonal communication and collaboration. The combination of wireless LANs, the third generation of cellular phones, satellites, and the increasing use of unlicensed wireless bands will cover the world with connectivity enabling both scientific research and emergency preparedness to utilize a wide variety of "sensornets". Building on advances in micro-electronic mechanical systems (MEMS) and nanotechnology, smart sensors can be deployed widely, will be capable of multiple types of detection, and can survive for long periods of time[38]. The integration of real-time multisensor data with data mining across large distributed data archives opens further avenues for adaptive monitoring/observation, situational awareness, and emergency response.

3.3 Interaction

We use the term *interaction* in the broad sense of (1) communication between or joint activity involving two or more people and (2) the combined action of two or more entities that affect one another and work together. Higher-performance *computation* provides more powerful tools for discovery through analysis and more systemic and realistic simulations. Acquisition, curation, and ready access to vast and varied types of digital *content* provide the raw ingredients for discovery and dissemination of knowledge. Computation and content, integrated through networking, offer new modes of *interaction* among people, information, computational-based tools/services, and instruments.

Working together in the same time and place continues to be important, but through cyberinfrastructure this can be augmented to enable collaboration between people at different locations, at the same (synchronous) or different (asynchronous) times. The distance dimension can be generalized to include not only geographical but also organizational and/or disciplinary distance. Our surveys confirmed that collaboration among disciplines is increasingly necessary and now requires, in some cases, hundreds of scientists working on a single project around the globe. Cyberinfrastructure should support this type of collaboration in a reliable, flexible, easy-to-use, and cost-effective manner. Groups collaborate across institutions and time zones, sharing data, complementary expertise, ideas, and access to special facilities. This can greatly expand the possibilities for synergy and is especially important to those researchers who are more isolated due to geographic or institutional circumstances.

We also heard that because of converging advances in computation, digital content, and networking, the research community is poised to pursue its work in a much more connected and interactive way. We have the opportunity to extend networked systems to provide *comprehensive* and increasingly *seamless* functional services for research and learning – to create virtual laboratories, research organizations, indeed technology-enabled research environments that offer a full spectrum of activities in the process of scientific discovery and the education of the next generation. We are at a threshold where a *collaboratory* or *grid community* can become “the place” where a research community interacts with colleagues, data, literature, and observational systems together with very powerful computational models and services. Although many technical, social, and economic challenges remain, the potential exists for facilitating both deeper and broader scientific and engineering research and education.

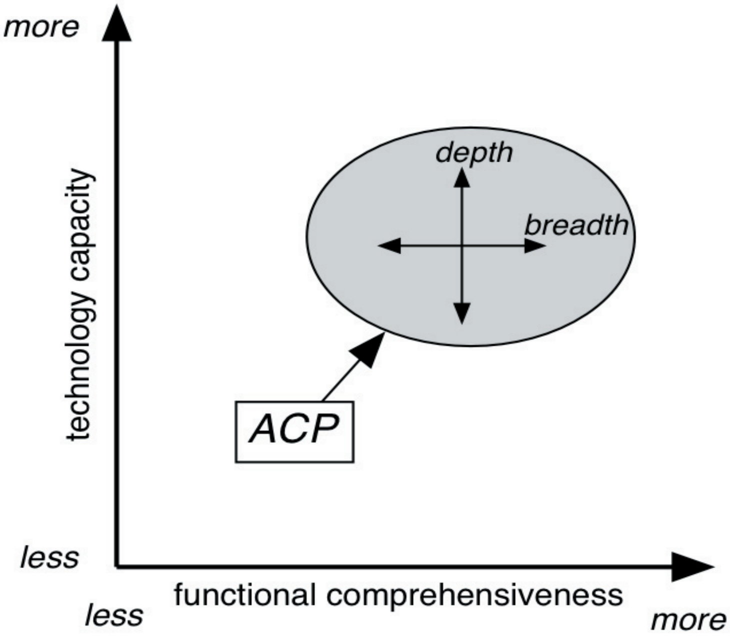

Figure 3.2 – Increasing capacity and functional comprehensiveness of cyberinfrastructure enable both depth and breadth approaches to discovery.

Figure 3.2 is an abstract and qualitative representation of two related dimensions emerging from advances in the nature and application of cyberinfrastructure. The vertical axis is a relative measure of the aggregate basic capability of the technology measured in terms of computation rates, storage capacity, and network bandwidth. The horizontal axis is a measure of breadth of use, or functional comprehensiveness – that is, how completely a cyberinfrastructure-based environment provides the resources and functions that researchers depend upon. To what extent can researchers readily find and effectively interact in a seamless way with all the colleagues, the data, the literature, the appropriate computational services, and the instruments necessary to meet their individual and community aspirations?

Technological capabilities expand rapidly. The Panel also heard, albeit more slowly and less predictably, that cyberinfrastructure is playing a more pervasive role in affecting how scientists do their work. Various fields begin the application of cyberinfrastructure in various ways. For example, some fields are building comprehensive collections of digital science literature; some communities have critical community data repositories and shared libraries of simulation codes; instruments and sensors arrays provide new types of observational data to widely dispersed research teams. The opportunity and challenge are to expand, integrate, and exploit the commonality among these applications of cyberinfrastructure. The shaded area of the graph represents a state of being or state of practice in this cyberinfrastructure capacity vs. comprehensiveness space. The goal of an Advanced Cyberinfrastructure Program (ACP) is to move the state of being region up and to the right (more comprehensive at higher capacity) – both *within* and *among* more and more fields of science and engineering.

As the combined state of capacity and functional comprehensiveness increases, and is adopted more broadly, the payoff will likely derive from enhancing both “depth” and “breadth” approaches to discovery.

In a depth approach, for example, atmospheric scientists could use higher-performance computation (together perhaps with denser and smarter distributed networks of sensors and with higher quality archival data) to improve the resolution and accuracy of a weather prediction model. Astronomers could use a more capable telescope to look more deeply into their favorite region of the universe.

In a breadth approach, a multidisciplinary team of earth scientists could use the availability of more computational power, more complete multi-dimensional data, enhanced observation capability, and more effective remote collaboration services to bring together an entire earth system simulation framework capable of supporting usefully predictive environmental simulations. Astronomers, given access to a federated “digital sky,” could explore the breadth of the known universe over the entire available electromagnetic spectrum to seek, for example, rare or new objects or phenomena. We can only begin to glimpse the impact of blended depth and breath approaches, especially as they weave together complementary expertise from multiple disciplines.

Another theme concerning knowledge environments for science based on cyberinfrastructure arose from the testimony we gathered. The theme is one of design of knowledge environments for *multiple uses*. In some cases this means to design such environments with the intent to (at least eventually) support both research and education and build further synergy between them. Others, in a similar context, encouraged intentional activity to use cyberinfrastructure to enhance broader participation (“democratization”) in science and engineering. The other variation of a multiple uses environment, sometimes called a *rapid-*

response collaboratory, is to support both basic science and, when necessary, the identification and rapid deployment of scientific and engineering resources to address natural or man made disasters (for example, earthquakes or bioterrorism attacks).