

Revolutionizing Science and Engineering through Cyberinfrastructure:

***Report of the National Science Foundation
Blue-Ribbon Advisory Panel on Cyberinfrastructure***

Draft 1.0

Released: April 19, 2002

Printed: 4/19/02 12:09 PM

Daniel E. Atkins, Chair
University of Michigan

Kelvin K. Droegemeier
University of Oklahoma.

Stuart I. Feldman
IBM

Hector Garcia-Molina
Stanford University

Michael L. Klein
University of Pennsylvania

Paul Messina
California Institute of Technology

David G. Messerschmitt
UC-Berkeley

Jeremiah P. Ostriker
Princeton University

Margaret H. Wright
New York University

*Comments on this Draft are welcome and should be sent to the
Panel, NSF-PCI-all@umich.edu **not later than May 1, 2002.***

1
2
3
4
5
6
7
8
9
10
11
12

Executive Summary

An Executive Summary will be prepared for the final report. For now please get a quick overview of the document by reviewing the Table of Contents.

Acknowledgments

In the final report we will definitely acknowledge the many people who have helped us in the endeavor.

Preface to Draft 1.0

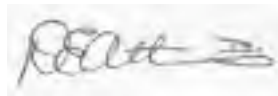
The activities of the Panel have been underway since May 2001. The process has included extensive interaction with the NSF community as well other US and international research funding agencies through hearings, surveys, readings, web browsing, and ad hoc input. The detailed work of the Panel has proceeded through a sub-committee structure responsible for analysis, framing, and writing of our findings and recommendations. The members of the Panel by design come from a variety of complementary perspectives on the topic at hand, and from a variety of backgrounds. The Panel has worked hard to take advantage of this diversity of complementary perspectives to develop views and recommendations that reflect strong consensus among the Panel members.

We are now anxious to share with the community-at-large a draft of what we have learned and concluded. This DRAFT represents the first integration of the writing of the various sub-committees. It has all of the properties of a first draft and is not yet polished as an act of writing or publishing. It does however, cover the breadth of our deliberations and what we expect to be our major new findings and recommendations in the final report. Please also note that the report is intended to be a four-tiered document and ultimately a web-document with rich hyper-linking. The four tiers are *executive summary, core report, appendices, and references/links* with more details and supporting resources. The DRAFT is largely the core report and is intended as the document that people with interest in the products of this Panel will actually read.

We are proposing that the NSF establish and lead a major strategic INITIATIVE with the goal of revolutionizing science and engineering through cyberinfrastructure enlisting the research interests and skills of computer scientists with those of other sciences, both in the NSF and other agencies. This initiative will need a carefully selected name and we have so far left that choice to others. In the DRAFT report will use the placeholder “the INITIATIVE.”

We are anxious to hear from the community on any topic relevant to the report. We have included line numbers in the DRAFT to facilitate pointing to specific content. (Please refer to section and line number.) What is clear; what is not? Have we left out something significant? How could we improve the report? Etc. etc.

Our intent is to produce the final report not later than June 1, 2002 and so we are asking that any input you wish to make on this Draft be sent via email to the Chair (atkins@umich.edu) not later than May 1, 2002. Thank you.



Daniel E. Atkins
Ann Arbor, April 2002

13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55

Table of Contents

1. THE VISION	1
1.1. A Nascent Revolution.....	1
1.2. Thresholds and Opportunities.....	2
1.3. Proposal for an Initiative	5
1.4. Funding and Organizational Considerations	5
1.4.1. Organization Proposal.....	6
1.4.2. Summary.....	7
2. BACKGROUND AND CHARGE	9
2.1. Background.....	9
2.2. The Charge.....	11
3. CHALLENGES AND OPPORTUNITIES FOR THE SCIENTIFIC RESEARCH COMMUNITY.....	13
3.1. Goals and Methodology.....	13
3.2. Assessment Findings	14
3.2.1. Philosophy and Process.....	14
3.2.2. Current Resources	15
3.2.3. Future Infrastructure.....	16
3.2.4. Emerging Paradigms and Activities	18
4. THE NEW CYBERINFRASTRUCTURE: WHAT CHANGED IN COMPUTING.....	20
4.1. An Embarrassment of Riches.....	20
4.2. Commercial Computing and the Needs of the Scientific Research Community	20
4.2.1. Commercial Products and Services	20
4.2.2. The Commercial World Has Become Far More Sophisticated	21
4.2.3. Distributed Computing is Suddenly Real.....	21
4.2.4. Needs Unique to IT for Scientific Computing	21
4.3. Cyberinfrastructure and Technology	22
4.4. Hardware Trends.....	22
4.4.1. Computational Processing	22
4.4.2. Memory	23
4.4.3. Storage.....	23
4.4.4. Networking – Wide Area Network.....	24
4.4.5. Networking – Local Area.....	24

4.4.6.	Displays	25
4.5.	Software and Content Trends - Software is Still the Bottleneck	25
4.5.1.	Shared Middleware	25
4.5.2.	Security	25
4.5.3.	Content Management	25
4.5.4.	Information Networking	26
4.5.5.	Collaborative Capabilities	27
4.6.	Investment Models	27
4.6.1.	Hardware Costs	27
4.6.2.	Human Capital	27
5.	THE LANDSCAPE OF RELATED ACTIVITIES	28
5.1.	Computing Industry	28
5.2.	Computing Research	28
5.3.	Other Sciences	28
5.4.	Other Federal Agencies	28
5.5.	Non-US Activities	29
5.6.	The Ecology of Scientific Computing	29
6.	PARTNERSHIPS FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE: PAST AND FUTURE ROLES	31
6.1.	The Past and Present	31
6.2.	Rationale for the Future	32
6.3.	The Future of the PACI program	34
7.	ACHIEVING THE VISION	35
7.1.	Slicing the Pie	35
7.2.	A Foundation of Technology Research and Technology Transfer	36
7.3.	Some Challenges	38
7.4.	Organization within NSF	39
7.4.1.	Organizational Principles	40
7.4.2.	Processes	42
7.4.3.	Incentives	43
7.4.4.	Continuity	44
8.	SCOPE AND BUDGET ESTIMATES	46

- 8.1. Scope of the INITIATIVE..... 46**
- 8.2. Fundamental, Longer-term Research in Information Technology and its Applications. 46**
- 8.3. Applications of Information Technology 47**
- 8.4. Cyberinfrastructure Supporting Applications..... 47**
 - 8.4.1. High-end general-purpose centers 47
 - 8.4.2. Data repositories..... 48
 - 8.4.3. Service centers..... 49
 - 8.4.4. Digital Libraries 49
 - 8.4.5. Networks..... 49
- 8.5. Core Technologies Incorporated into Cyberinfrastructure 49**
- 8.6. Table 7.1 Summary of Scope and Budget Estimates for the INITIATIVE..... 51**

1 1. The Vision

2 1.1. A Nascent Revolution

3 A new age has dawned in scientific and engineering research, driven by progress in
4 computing and communication technology. Scientists in many disciplines are already
5 revolutionizing their fields by using computers and digital data to replace and extend
6 their traditional efforts. The amounts of calculation and the extent of information that can
7 be stored and used are exploding at a rate that should shake up assumptions throughout
8 the research community. In a few years, the contents of the historic scientific literature
9 will fit on a rack of disks, and a computer that fits in a small office will provide more
10 computing than all the supercomputing centers together today. The results of today's
11 largest calculations and most sizable collections will take just seconds to transmit using
12 the fastest known network technologies. Thus, even today,

- 13 • The primary access to the latest findings in many fields such as physics is through
14 the Web, then later through classic preprints and conferences, and only after that
15 through refereed archival papers.
- 16 • Crucial data collections in the social, biological, and physical sciences are now
17 online and remotely accessible – modern genome research would be impossible
18 without such databases, and soon astronomical research will be similarly
19 redefined through the National Virtual Observatory.
- 20 • The classic two approaches to scientific research, theoretical/analytical and
21 experimental/observational have been extended to “*in silico*” simulation/modeling
22 to explore a larger number of possibilities and to achieve new levels of precisions.
- 23 • The enormous speedups of computers and networks have enabled simulations of
24 far more complex systems and phenomena, and visualizing the results in many
25 dimensions.
- 26 • Groups collaborate across institutions and time zones, sharing data and ideas and
27 access to special equipment without wasteful travel.
- 28 • Advanced computing use is no longer restricted to a few research groups in a few
29 fields such as weather prediction and high-energy physics, but pervades scientific
30 and engineering research, including the biological, chemical, social, and
31 environmental sciences as well as nano-technology.

32 In the future, we expect researchers

- 33 • To combine raw data and new models from many sources, and to utilize the most
34 up to date tools to analyze, visualize, and simulate complex interrelations.
- 35 • To collect and make generally available far more information (the outputs of all
36 major observatories and astronomical satellites, satellite and land-based weather
37 data, 3-d images of anthropologically important objects), and that this will lead to
38 a qualitative change in the way research is done and the type of science that will
39 result.
- 40 • To work across traditional boundaries: for environmental scientists to take
41 advantage of climate models, for physicists to make direct use of astronomical
42 observations, for social scientists to analyze interactive behavior of scientists as
43 well as others.

- 44 • To simulate more complex and exciting systems (cells, organisms rather than
45 proteins and DNA; the entire earth system, rather than air, water, land, and snow
46 independently).
- 47 • To access the entire published record of science online, and for future
48 publications to be much richer (hypertext, video, photographic images).
- 49 • To visualize the results of complex data sets in new and exciting ways, and to
50 create techniques for understanding and acting on the observations.
- 51 • To work routinely with colleagues at distant institutions, even ones that are not
52 traditionally considered Research universities, and with junior scientists and
53 students as genuine peers, despite differences in age, experience, race, or physical
54 ability.

55 **1.2. Thresholds and Opportunities**

56 These benefits are just the start of a revolution. Why act now? Computers have been
57 improving for decades, and a few researchers have tried to do many of the things in the
58 list. We believe that several key thresholds have recently been reached in the use of IT, in
59 part because NSF has made large and successful investments in a number of research
60 areas, including networking, supercomputing, human interfaces, collaboration
61 environments, and information management.

62 The Internet and the Web were invented to support the work of researchers, and their use
63 permeates all of science. Broadband networks connect all research centers and enable the
64 rapid communication of ideas, the sharing of resources, and remote access to data. The
65 next generations of the net promise even greater benefits to the research community.

66 Most modern researchers are fully conversant with and dependent on advanced
67 computing for their daily activity, and have a thirst for more. Older scientists are learning
68 to take advantage of the new technologies. First generations of discipline-specific
69 computing platforms and environments have been widely used and successful.

70 Closed form analytic solutions are available for a decreasing fraction of interesting
71 problems; only a numeric computation can produce useful results.

72 Moore's Law has led to simulations which begin to match the complexity of the real
73 world, having crossed the threshold that allows fully three dimensional, time-dependent
74 modeling with realistic physics and opening up a vast range of problems to qualitative
75 attacks. They range from cosmology to protein folding – problems formerly considered
76 far too complex to address head on.

77 In an increasing fraction of cases, it is faster, cheaper and more accurate to simulate a
78 model than to build and measure a physical object.

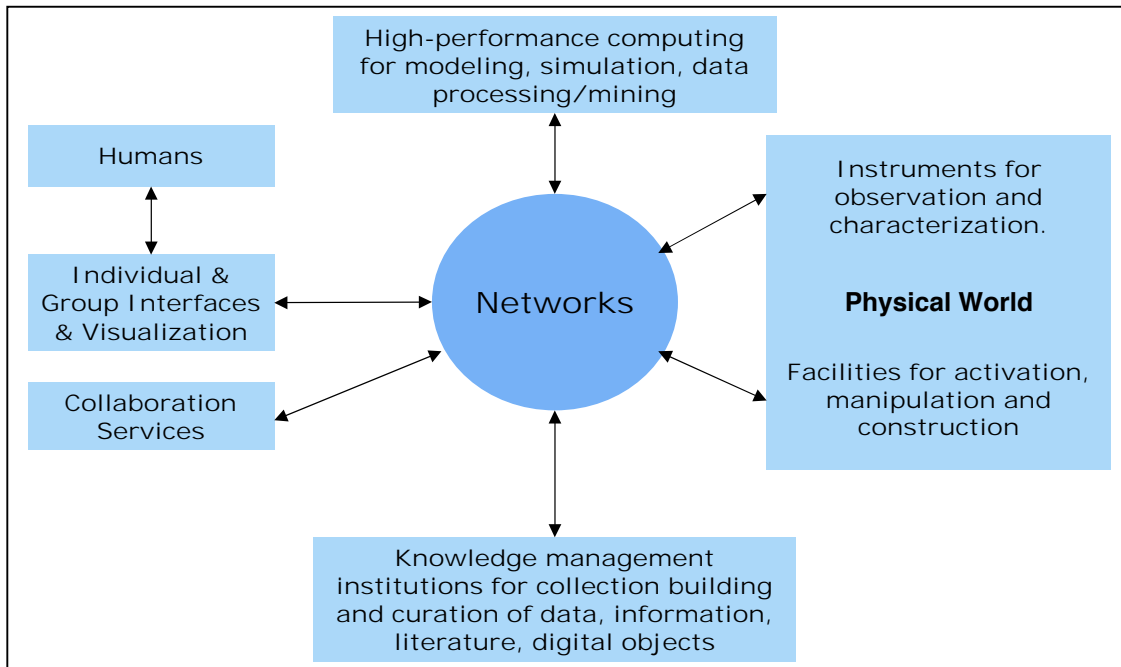
79 Widely accepted standards for information formats and access make collaboration
80 computationally feasible.

81 Storing terabytes of information is common and inexpensive; warehouses containing
82 hundreds or even thousands of terabytes of data will soon be affordable and necessary for
83 storing scientific and engineering information.

84 Computing power that was unavailable a few years ago – trillions of operations a seconds
85 – can now be found on a number of campuses.

86 Computational and visualization techniques have progressed enormously, and account for
87 at least as much speedup and scientific value as improved hardware.

88 Most researchers would not be able to function without e-mail or access to the web. They
 89 certainly would have fewer contacts with distant, especially international, scientists and
 90 be much less able to keep on the cutting edge of their field.
 91 These capabilities will in combination allow access to complex services as well as raw
 92 computing resources through the network, enabling both collaboration and sharing over
 93 distance and time. This picture is consistent with the vision of the Grid¹ [], the modern
 94 Internet [], distributed computing, and collaboratories[]. It is the basis for what some are
 95 calling *e-science*. A schematic of such services is shown in figure 1.1.



96 **Figure 1.1 – Schematic of services for cyberinfrastructure enabled research.**

- 97 There are also significant risks and costs if we do not make a major move at this time:
- 98 • Researchers in different fields and at different sites will adopt different formats
 99 and representations of key information, which will make it forever difficult or
 100 impossible to combine or reconcile.
 - 101 • If no decision is made to curate and store indefinitely raw and intermediate
 102 research results as well as the polished and reduced publications, irreplaceable
 103 data will be lost.
 - 104 • Effective use of cyberinfrastructure can break down artificial field boundaries,
 105 while differing tools and structures can isolate scientific communities for years.

¹ The meaning of the term “Grid” is evolving. It was first used as an analogy to the electrical power grid – dynamic pooling of computation power across distance and administrative domains. The “access grid” refers to a set of video, audio, and collaborative tools used in the research community over the Internet. More recently the term Grid has evolved to often mean a more comprehensive structure linking people, information, and tools/facilities as indicated in Figure 1.1. The term xGrid where x can be a discipline of a place is now coming into use, e.g. the BioGrid or the MGrid (Michigan Grid). In this sense the term “Grid” and collaboratory (or co-laboratory) are similar ideas.

- 106 • Groups are already building their own application and middleware software
107 without awareness of comparable needs elsewhere, both within the NSF and
108 across all of science. Much of this software will be of limited long-term value or
109 usefulness because of lack of a consistent computer science perspective. Time and
110 talent will be wasted that could have led to much better computing *and* much
111 better science.
- 112 • Very rapid changes are coming in computing and application architectures; lack
113 of consideration of work in other sciences and in the commercial world could
114 render projects obsolete before they deliver.

115 The time is ripe for NSF to accelerate the revolution for the benefit of all researchers. A
116 confluence of events and possibilities makes this the right time. Researchers are ramping
117 up their use of computing resources, starting to store enormous amounts of information
118 and sharing it. Distributed computing, large clusters, data farms, and broadband networks
119 (typified by Internet 2, Grid, and Web Services directions) have moved from research to
120 practical use. We anticipate a phase change, where a moderate effort can have a highly
121 desirable and nonlinear effect.

122

123 We envision an environment in which raw data and recent results are easily shared, not
124 just within a research group or institution but also between scientific disciplines and
125 locations. There is an exciting opportunity to share insights, software, and knowledge, to
126 reduce wasteful re-creation and repetition. Key applications and software that are used to
127 analyze and simulate phenomena in one field can be utilized broadly. This will only take
128 place if standards and underlying technical infrastructures are shared by all. Although
129 many of the mechanisms to support the best scientific computing are becoming available
130 through commercial channels, there continue to be needs that the commercial sector is
131 unlikely to meet directly because of the market size and technological risks.

132

133 Scientists must have easy access to the finest tools from the commercial and advanced
134 research sectors, without dampening their creativity and ardor to do even better.
135 Individual researchers expend too much effort, frequently with insufficient
136 knowledgeable computing assistance, to create and re-create computing resources, to
137 save and access information, to protect the assets. Much of this work could be done by
138 computing experts and used across the scientific research community. The initiative will
139 encourage groups of scientists to undertake large coordinated information-intensive
140 projects that can radically change the way they and their peers work, that will support the
141 sharing and long-term use of information that results from their work.

142

143 There are many possibilities that cannot be pursued today because of lack of fundamental
144 understanding in computing. We envision radical improvements in cyberinfrastructure
145 and the work in all the sciences over time, as work ripens in a number of areas of
146 computing research. Computer scientists will be encouraged to work on problems that
147 will further extend the range of the possible.

148 **1.3. Proposal for an Initiative**

149 We propose a large and concerted new effort, not just a linear extension of the current
150 investment level and resources. NSF must recognize that the scope of shared
151 cyberinfrastructure must be far broader than in the past: It includes computing cycles, but
152 also greater bandwidth networking, massive storage, and managed information. Even
153 these are not sufficient: there must be leadership on shared standards, middleware, and
154 basic applications for scientific computation. The individual disciplines must take the
155 lead on defining certain specialized software and hardware configurations, but in a
156 context that encourages them to give back results for the general good of the research
157 enterprise, and that facilitates innovative cross-disciplinary activities.

158
159 To succeed, NSF must institute a broad and deep program that supports the true needs of
160 all the sciences and NSF missions by committing

- 161 • to make the fruits of this research as well as related work from other agencies and
162 companies available in an integrated and easy fashion to support new approaches
163 to doing scientific and engineering research, and ensuring that the exponentially
164 growing amounts of data are collected, curated, managed, and stored for broad
165 long-term access by scientists everywhere;
- 166 • to create and continually renovate a new “high end”, so that selected research
167 projects can use centralized resources 100-1000 times faster and bigger than are
168 available locally. The continued (literally) exponential improvements in
169 computing speeds and disk capacity mean that research groups and universities
170 now have immediate access to far more computing than ever before, but the
171 limited national investment in massive resources means that the most aggressive
172 research projects frequently cannot move to the next level of complexity and
173 resolution. National needs for supercomputing capabilities will drive new
174 generations of work in architecture, and NSF needs to take advantage and
175 participate in such efforts to have a continually improving research
176 cyberinfrastructure;
- 177 • to extend the research base to benefit future generations of scientists by
178 supporting research in areas of computing science that are likely to have largest
179 impact;
- 180 • to use the new infrastructure to educate the next generations of scientists using the
181 best techniques, and to ensure broader participation without respect to field
182 boundaries, institutional wealth, personal origin or bodily ability;
- 183 • to maximize international collaboration and resource-sharing through
184 standardization and networking.

185 **1.4. Funding and Organizational Considerations**

186 We estimate that an additional \$650M per year will be needed to sustain the revolution.
187 (More detail in Section 8.) This is not a one-time charge like buying a building or a ship.
188 Computing technology is advancing rapidly, so we should plan to buy newer and better
189 equipment, and also to fund the considerable personnel involvement for maintenance,
190 extension, and assistance. Over time, people costs will constitute a growing fraction as
191 hardware and software unit costs decline.

192

193 There is already a significant base of effort and capability in the PACIs, which were
194 created in response to the Hayes Report. They run computing and data centers, create
195 important middleware and scientific software, and coordinate activities with other
196 scientists. We anticipate that they will play a continuing but evolving substantial role in
197 the greatly enlarged activity we propose.

198 Related activities are underway in other Federal agencies (e.g., Department of Energy,
199 National Institutes of Health), and in other countries (UK, EU, Japan). The planning
200 ought to be coordinated with them, and all efforts made to achieve consistent interfaces in
201 a timely fashion. (These other parties are making very large investments. The Japanese
202 Earth Simulator computer for climate research will be the world's fastest civilian
203 computer. NIH spends as much on IT-related activities as the total NSF research
204 budget².)

205
206 NSF has unique breadth of scientific scope and the mandate for the health of the
207 scientific research enterprise in the US. It therefore ought to take the lead to ensure that
208 our researchers have continuing access to the best resources as well as the ability to work
209 with their peers in other nations. An NSF initiative should improve the quality and
210 quantity and efficiency of scientific research and of the researchers. It can be catalytic
211 and provide over-the-horizon views for other agencies, research labs, and education-at-
212 large.

213
214 The effort must be managed carefully to reach these results, providing a rapidly growing
215 access to the most massive resources for the most advanced needs and improving the
216 capabilities of all. NSF must attack the problems of creating and supporting technical
217 applications, of managing the evolution of software and the changing economics of
218 computing with the same vigor that it supports research in more classic areas of
219 computing and other sciences.

220 **1.4.1. Organization Proposal**

221 The INITIATIVE must not be an incidental supplement to the current structure, but
222 demands a new coherent central organization, an INITIATIVE office, that will drive
223 major new discipline-specific work as well as provide massive new resources and
224 maintain consistency. A single leader must have authority to organize many concurrent
225 activities, create new organizational structures, allocate and manage significant resources,
226 and align incentives. The INITIATIVE is intended to serve all of the NSF community as
227 well as to coordinate with other similar initiatives in the U.S. and internationally. The
228 main activities that must be driven in a consistent manner include:

229
230 **Discipline-changing Cyberinfrastructure-based Projects:** The scientific and
231 engineering disciplines select and pursue large IT-intensive new efforts that can change
232 the range of the possible, and can affect the way that scientists in those fields view
233 research problems. These are managed by the senior people in the directorates, but the
234 Initiative Leader must ensure that the computational aspects are consistent with the
235 cyberinfrastructure vision. The projects should make maximum use of existing

² See more details in section 5.4.

236 cyberinfrastructure, and also contribute to it by ensuring that access to information and
237 applications access is open, that results in one field positively benefit others, and that
238 implementation and interfaces are consistent with standards established across NSF. As a
239 condition of access to national resources, these innovation projects must include groups
240 or individuals who are experts in the architectures and interfaces.

241 To guide the decisions, the Leader would chair a very high level committee (at the level
242 of the Foundation) that establishes policy and allocations across fields and projects.
243 The Initiative should lead that advance of research through computing across the Federal
244 government, as well as coordinating with related thrusts in other countries. This would
245 necessitate liaison across agencies.

246
247 **Shared middleware and applications:** Basic software must be planned and developed.
248 Standards for sharing of information metadata must be agreed across large swaths of
249 users, and perhaps among agencies and continents. The decisions about development
250 resources and quality must be made with a high level of application development,
251 computer science and computing industry knowledge as well as being informed by needs
252 of the other sciences. The Initiative would be responsible for establishing priorities,
253 avoiding undesired duplication, and monitoring quality of work. The development teams
254 themselves must have significant engineering competence, experience, and stability since
255 they will be responsible for long-term assets of the community.

256
257 **Shared physical resources:** Shared computing, storage, and networks must be managed
258 with a continuing investment stream and a plan for both aggregate capacity and
259 maximum tightly-bound capacity. (Some computations will need physically integrated
260 support because of latency constraints and shared memory needs; many others will be
261 capable of support by more loosely and distantly connected resources.) The systems must
262 be available, secure, well managed, even though they will utilize unusual configurations.
263 This function demands significant operational and management skill as well as very large
264 and ongoing capital investment.

265
266 **Computing research:** New work in certain areas should be especially encouraged,
267 including those likely to improve modern applications and the efficacy with which they
268 are produced, as well as research aimed at the new generations of distributed resources.
269 The Initiative must be able to support fundamental research in areas relating to
270 applications.

271
272 **Workforce and Education:** The initiative must encourage projects that try new
273 approaches in education, encourage broad institutional and personal inclusion, and
274 inculcate new technologies in the work of young researchers.

275 **1.4.2. Summary**

276 There is a once-in-a-generation opportunity to revamp and improve the process of
277 scientific research. A successful revolution requires long-term focus and commitment, as
278 well as innovative organizational structures, continuing high level of buy-in from the
279 upper levels of the National Science Foundation and Congress, as well as committed
280 champions and executives to execute well.

281 We recommend a strategic new investment to make this happen now.

2. Background and Charge

2.1. Background

The National Science Foundation through its 50 some years of “promoting the progress of science” has contributed significantly to the development of computers and computing both as the *object of research* as well as a *means of research* serving in all communities of science and engineering. The vision and initiative sketched in Section 1 presents NSF the opportunity, indeed we believe the responsibility, to take these complementary dual activities to the next frontier across all of science and engineering. It will strengthen scientific and engineering research and eventually education at all levels. It will accelerate the meaningful adoption of cyberinfrastructure in broad areas of higher education.

NSF is specifically charged with fostering and supporting the development and use of computers and other scientific methods and technologies for research and education in the sciences. As computers emerged, various Directorates of the NSF managed programs to support research in components, theory, software, systems, and applications of computers. An Advanced Scientific Computing (ASC) Program, situated in the Office of the Director, provided access to the highest performance super computers. In 1985, programs from other Directorates as well as the ASC activities were merged into the Directorate for Computer and Information Science and Engineering (CISE). [CISE has three goals:](#)

- To enable the U.S. to uphold a position of world leadership in computing, communications, and information science and engineering;
- To promote understanding of the principles and uses of advanced computing, communications and information systems in service to society; and
- To contribute to universal, transparent and affordable participation in an information-based society.

To achieve these, CISE supports investigator initiated research in all areas of computer and information science and engineering; helps develop and maintain cutting-edge national computing and information infrastructure for research and education generally; and contributes to the education and training of the next generation of computer scientists and engineers. CISE is currently organized in five divisions, three of which focus principally on research, and two of which combine both infrastructure and research functions.

Recent conversations with Eric Bloch, Director of NSF at the time CISE was formed, and Gordon Bell, the first Associate Director for CISE, confirmed that CISE was specifically given dual roles in the belief that there is significant synergy between research about computer and information science and engineering and the development, deployment and use of advanced computer and information systems environments to support their use in science and engineering broadly.

44
45 Specific CISE programs to develop and maintain cutting-edge national computing and
46 information infrastructure have been informed over the years by a series of advisory
47 panel inquiries and reports. The current high performance computing activity, for
48 example, is guided by the recommendations of the 1995 [Hayes Report \(Report of the](#)
49 [Task Force on the Future of the NSF Supercomputer Centers Program\)](#). These
50 recommendations, along with the predecessor [Branscomb Report \(NSF Blue Ribbon](#)
51 [Panel on High Performance Computing\)](#), formed the basis for the development of the
52 [Partnerships for Advanced Computational Infrastructure \(PACI\) program](#).
53

54 Two PACI partnerships established in 1997 are presently operating under the principles
55 set forth in the Hayes Report by (1) providing access to high-end computing, (2)
56 affording knowledge transfer of enabling technology and applications research results
57 into the practice of high-performance computing, and (3) supporting education, outreach
58 and training activities. Each partnership consists of a leading edge site, the [National](#)
59 [Center for Supercomputing Applications](#) in Urbana-Champaign and the [San Diego](#)
60 [Supercomputer Center](#) in San Diego, and a significant number of partners.
61

62 The term “cyberinfrastructure” was recently coined by NSF management to connote not
63 only advanced scientific computing but a more comprehensive infrastructure for research
64 and education based upon distributed but federated networks of computers, information
65 resources, on-line instruments, and human interfaces. It provides a convenient way to talk
66 about IT-based infrastructure in contrast to more traditional science infrastructure.
67 Specific projects built upon cyberinfrastructure are using names such as *GRID*, *E-*
68 *science communities*, and *collaboratory*.
69

70 During our inquires some people associated infrastructure narrowly with *equipment*
71 rather than with the broader concept of *equipment/facilities, people, and organizations*.
72 Highways are infrastructure, but so are the people and organizations to build, maintain
73 and police their use. In the more specific case of IT-based infrastructure, we define it to
74 be *a set of functions, capabilities, and/or services that make it easier, quicker, and less*
75 *expensive to develop, provision, and operate a relatively broad range of applications*.
76 *This can include facilities, software, tools, documentation, and associated human support*
77 *organizations*.
78

79 This Panel on Cyberinfrastructure was convened to explore a framework for an advanced
80 infrastructure initiative building upon, but going beyond, the Hayes report and the PACI
81 model. More specifically issues such as the following motivated the formation of this
82 Panel:
83

- 84 1. The current five-year cooperative agreements with the two PACI centers will end
85 soon, and thus questions about their effectiveness and future roles and funding are
86 apropos.
- 87 2. The emergence of an additional set of large infrastructure-type initiatives,
88 including the [Pittsburgh Super Computing Center](#), the [Distributed Terascale](#)
89 [Initiative](#), the [NSF Middleware Initiative](#) , and perhaps others may benefit from
90 being placed in a more strategic conceptual framework.

- 91 3. Increasing science/engineering community-based initiatives and budget demands
92 in all Directorates for greater investment in information technology to support
93 domain-specific research. The nature of these investments includes, but is not
94 limited to, high-end computation. Initial versions of such requirements and
95 supported research projects have bubbled up through the NSF [Information](#)
96 [Technology Research](#) initiative or have been launched from non-CISE
97 Directorates, for example, the [Network for Earthquake Engineering Simulation](#)
98 (NEES) in Engineering or the [National Science Digital Library](#) in Education and
99 Human Resources.
- 100 4. A recognition that funding models, funding levels, and organizational structures
101 for IT-based research infrastructure needed to be examined. Funding of IT-based
102 infrastructure --cyberinfrastructure--presents special challenges and opportunities
103 over the funding of more traditional scientific facilities and instruments. On the
104 one hand is the challenge of rapid depreciation. On the other hand, IT-based
105 infrastructure offers new opportunities for pooling and sharing resources, often in
106 place-independent ways.
- 107 5. A growing sense that science and engineering research and practice are reaching
108 thresholds in performance and adoption of IT that could radically transform the
109 “what”, “how” and “who” of scientific research on a truly global scale. Other
110 countries have begun major initiatives to create advanced cyper-infrastructure and
111 apply it at the frontiers of science.
- 112 6. Questions about the extent to which the rapid rate of change in computation,
113 storage, and communication technologies will change the nature of the
114 investments by NSF in this area. For example, will centralized highest-
115 performance computing centers continue to be needed, or will grids of distributed
116 machines meet the needs of the very highest-end computing communities.
117

118 **2.2. The Charge**

119 The formal [Charge to the Panel](#) has three parts:

- 120
- 121 1. Evaluate the performance of the PACI Program in meeting the needs of the
122 scientific research and engineering community.
 - 123 2. Recommend new areas of emphasis for the NSF Directorate for Computer and
124 Information Science and Engineering that will respond to the future needs of this
125 community.
 - 126 3. Recommend an implementation plan to enact any changes anticipated in the
127 recommendations for new areas of emphasis.
128

129 The [full charge](#) to the Panel includes 5 or 6 sub-items and is included in Appendix # and
130 is also available at http://www.cise.nsf.gov/b_ribbon/index.html. Although we have
131 addressed the full range of our Charge, this report is not organized topic by topic
132 according to the Charge. In the Appendix we also provide a mapping between specific
133 topics/questions in the Charge and the relevant findings and recommendations in this
134 report.
135

136 The Panel approached its critique of the PACI program in a way that made extensive use

137 of the normal review processes underway as prescribed by the PACI cooperative
138 agreements. We enriched our discovery processes through extensive survey and hearings
139 described in Section 3. We have focused primarily on the PACIs as a case study to
140 inform what should be done in the future. We base our recommendations for new areas of
141 emphasis largely on an analysis of what is happening already as documented in hearings
142 and relevant other reports and upon the personal expertise of the Panel members. The
143 implementation plan, emerging from integrating across all of our discovery processes,
144 focuses on defining the broad scope of an initiative, requisite organizational models, and
145 suggested funding levels.

146

147 The Panel's consideration of "new areas of emphasis that will respond to the future
148 needs of this community" is broader than infrastructure *per se*. Cyberinfrastructure is
149 built upon the fundamental technologies of computation, storage, and communication.
150 And cyberinfrastructure is, in turn, the substrate for building IT-based resources, projects,
151 and organizations that serve specific scientific communities. The ultimate goal of our
152 inquiry is to revolutionize science and engineering research and education – to broaden
153 participation, enhance discovery and understanding, and accelerate application to
154 important problems in our world. The creation, deployment, and application of advanced
155 cyberinfrastructure is the basis for an initiative but not an end in itself. It should not be
156 used only to do faster and better what we are now doing; it should be used to do new
157 things, in new ways.

158

159 The scope of an initiative includes not only the assets of the PACI program but also other
160 recent initiatives including the Pittsburgh Terascale Center, the Distributed Terascale
161 Project, some of the large ITR projects, the Digital Library Initiatives, the National
162 Science Digital Library, Networking and Middleware Initiatives, IT application projects
163 in various Directorates, needs and initiatives in other Federal R&D agencies, related
164 projects in other countries, and related projects in U.S. research universities. We also
165 expect our work to be complemented by the concurrent review by the National Science
166 Board of the state of U.S. research infrastructure, both "traditional" and "cyber".

167

168 In recent years the NSF management has used the triad *People, Ideas, and Tools* to
169 describe its strategic goals and corresponding investment areas. In this framework,
170 cyberinfrastructure may be thought of as the *tooling* for revolutionizing both the
171 education of *people* ("to develop a diverse, internationally competitive and globally
172 engaged workforce of scientists, engineers and well-prepared citizens) and their
173 production of new *ideas* ("discovery across the frontier of science and engineering,
174 connection to learning, innovation, and service to society.") The scope of the proposed
175 initiative not only relates to the NSF responsibilities *to foster and support the*
176 *development and use of computers*, but also to a wide range of NSF charter
177 responsibilities, including correlating its research and educational programs with *other*
178 *Federal programs*; supporting *international cooperation* in scientific and engineering
179 activities; strengthening research and education *innovation* in the sciences and
180 engineering; fostering the *interchange of scientific information* globally; and *increasing*
181 *the participation* of women and minorities and others under-represented in science and
182 technology

3. Challenges and Opportunities for the Scientific Research Community

3.1. Goals and Methodology

The first step toward developing a long-range strategic plan for cyberinfrastructure is to critically assess key challenges – scientific, technological, and sociological – that now exist or are anticipated to be faced by the research community during the next several years. Closely linked is the identification of opportunities that can be enabled through the development of a comprehensive national cyberinfrastructure. We have undertaken both these tasks using five assessment methodologies with a primary goal of obtaining direct input from the broadest elements of the domestic and, to the extent practicable and relevant, the international scientific and engineering communities – which includes academia, private industry, government agencies and laboratories, and state, regional, and national centers.

The two principal instruments for gathering information were a web-based survey and formal oral testimonies, the latter taken on three separate occasions. The survey was patterned after that developed for the Hayes report, though was expanded considerably to capture recent topics. Links to the survey, which was active on the web for approximately 3 months during the fall of 2001, were provided at the PACI sites, their affiliates, and numerous other regional and national centers and laboratories.

Additionally, email messages announcing the survey were sent to hundreds of individuals on community-wide mailing lists. Overall, more than 700 individuals responded to the survey, and their input is woven throughout this report. A detailed quantitative analysis of responses to specific questions, and a comparison with their counterparts in the Hayes survey, are presented in section 2.

The second principal assessment tool consisted of invited oral testimony (with the opportunity of providing written supplements) collected in public sessions held at the NSF on 28-29 November 2001 and 15 February 2002, and at Cal-Tech on 22 January 2002. Sixty-two participants were drawn from a broad spectrum of expertise and included domain research scientists and engineers; computer and computational scientists; research and operational center directors; NSF assistant directors and program managers; leaders from other agencies and programs; computer system administrators; students and post-doctoral scientists; and technicians and user consultants. A complete list of participants is included in Appendix #. Specific emphasis was given to persons from disciplines that are now or only recently have begun to use high performance systems; to the physically challenged; and to females and other traditionally underrepresented groups, the latter represented by tribal colleges and universities, historically black colleges and universities, and Hispanic colleges and universities.

43 The witnesses were asked to provide a 10 to 15 minute overview of their views regarding
44 the needs of the scientific and engineering communities during the coming decade, with
45 emphasis on broad issues such as data collection and management, the modalities of
46 providing computational power across a large spectrum of systems, collaborations, and
47 distributed resources. They also were asked to comment on the role of the federal
48 government in providing such infrastructure. Each witness then answered questions from
49 the Panel during the remainder of the 30-minute time period. Although most of the
50 witnesses attended the testimony sessions in person, some testimony was given via the
51 Access Grid. Verbatim transcripts were prepared by the National Science Foundation
52 and, along with associated visuals, are now available on the web at [https://lapp1.cise-
53 nsf.gov/rhilderb](https://lapp1.cise-nsf.gov/rhilderb) and will soon be moved to the Panel Report website.

54
55 In addition to the web-based survey and testimony, the Panel utilized three other general
56 sources of information in creating this report: ad hoc personal conversations; a wide
57 variety of written materials (see References); and their own knowledge and experiences.

58 **3.2. Assessment Findings**

59
60 In presenting the results of our community-wide assessment, which includes a blending
61 of information from all sources described above. *Although these findings are clustered,*
62 *we are not ranking the importance of the findings against each other.* Rather, the
63 remarkable degree of consistency among individual responses, and within and among
64 disciplinary communities, makes clear that *all of the issues described below* are vitally
65 important for ensuring an effective and sustainable cyberinfrastructure for the foreseeable
66 future. Furthermore, these findings represent a synthesis of all information collected, or
67 interpretation of it, rather than independent conclusions of the Committee.
68

69 **3.2.1. Philosophy and Process**

- 70
71 • Because of its breadth, flexibility, effective peer review system, and broad
72 mission orientation, the NSF is the singularly appropriate government agency to
73 chart a national course for cyberinfrastructure. This finding echoes the
74 recommendation of the Presidents Information Technology Advisory Council
75 (PITAC). However, an effective cyberinfrastructure will come about only
76 through significant multi-agency cooperation and coordination.
77
- 78 • Cyberinfrastructure now lies at the core of revolutionary science in most every
79 discipline, and thus all directorates, divisions and programs within the NSF must
80 take a direct programmatic role in its development and sustenance. This is
81 particularly important given that many of the major science advances in the future
82 are likely to occur at boundaries *among* disciplines.
83
- 84 • The NSF should consider human capital and software as co-equals with hardware
85 in the context of cyberinfrastructure. In general, the NSF places too little
86 emphasis on funding personnel, and on software development and especially

- 87 maintenance, in comparison with hardware and traditional physical infrastructure.
88
- 89 • Cyberinfrastructure requires continuity, consistency and sufficient funding. The
90 NSF should consider the effects of periodic full re-competition with uncertain
91 success and its associated impacts on human capital. Of course, continued high
92 quality is critical, but can be achieved less disruptively via mechanisms of
93 periodic peer review, as now is performed at the National Center for Atmospheric
94 Research.
 - 95
 - 96 • The NSF needs to provide a framework, motivation, and clear direction for
97 building and sustaining effective linkages between academia and industry, and
98 thus for bringing the benefits of basic research to society. Never before have
99 partnerships between the academic and corporate communities been more
100 important, or the complexities associated with their consummation more
101 formidable.
 - 102
 - 103 • The NSF should give attention to the sociological, economic and cultural issues
104 that come from having an all-encompassing information technology
105 infrastructure. Social and behavioral research can be involved as both users and
106 creators of this infrastructure. They need to play a role in understand the non-
107 technical barriers to adoption, human-centered design principles, and the long-
108 term outcomes for research communities.
 - 109
 - 110 • Open source software strategies continue to provide significant benefit to all
111 communities and represent an effective strategy for the future. The NSF should
112 continue to support this mode of development.
 - 113

114 **3.2.2. Current Resources**

- 115
- 116 • The entry barrier into high performance computing continues to be high,
117 representing a disincentive to use by new communities. Further, greater
118 investments need to be made in software development, as well as training/support,
119 to facilitate the use of parallel and distributed architectures by all users. This
120 issue will become even more important with the move toward Grid-based
121 capabilities. Indeed, there exists a growing mismatch between raw, theoretical
122 peak and realized performance for production codes, and the investment of time
123 required for users to achieve reasonably good performance. Numerous survey
124 respondents noted that, in some areas, the state of the art in computer technology
125 is outpacing tools and best practices from the user perspective. For example, the
126 relatively straightforward and efficient auto-vectorizing and -parallelizing
127 compilers of the previous hardware era have given way to complicated messaging
128 directives that must be inserted manually, and that to many users are as
129 intimidating and time consuming as assembly language. Industry and academia
130 must work together to remedy this problem and bring greater parity between
131 hardware and the tools available for its use.

- 132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
- The PACI centers were created to bring high-performance computing to the masses, and to broaden the spectrum of users, in essentially the single investigator mode of research. They have done so with notable success, and also have responded to dramatic changes in the user base and types of applications being run. However, the PACI centers remain largely a batch-oriented environment, whereas future problems will require 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; adaptive sensors in field biology). Further, they are not configured to provide, in most cases, significant fractions of their resources in a dedicated fashion to support the most challenging research problems. The NSF needs to broaden its vision for and increase the resources made available to cyberinfrastructure.
 - The National Resource Allocation Committee (NRAC) process to allocate computing resources to users no longer is effective and must be overhauled. For example, users are subjected to double jeopardy by having to prepare both research grant (agency) proposals as well as proposals for computer resources. Funding of the former with a negative decision for the latter clearly creates a problem! The NSF examined coupling the two processes in the early 1990s but chose to leave them separate. The allocation of cyber resources on a yearly basis to support multi-year grant awards also is a significant problem that must be corrected. Mechanisms for requesting resources should be streamlined as well, and the reviewer base broadened to ensure an adequate understanding of the needs being expressed. Finally, the new allocation process must be flexible to include new resources, such as distributed federated data repositories and remote visualization facilities.
 - The PACI centers have been highly successful at developing visionary, innovative technologies and prototype tools. However, they have been notably less successful in taking these visions and prototypes to the next stages, i.e., in deploying and supporting sustainable, practical tools that bring direct benefit to faculty, graduate students, and researchers. The main problem appears to be insufficient funding and the lack of mechanisms to selectively invest in promising activities -- not the lack of creative expertise or the desire to provide benefit to the community. Consequently, the NSF needs to initiate a well defined strategy for addressing this problem.

171 **3.2.3. Future Infrastructure**

- 172
173
174
175
- The “last mile” or “end to end” problem (i.e., of obtaining high-performance connectivity from backbones and regional networks to the desktop) exists in virtually every community and represents a serious problem for the future of

- 176 research and education, particularly with regard to traditionally underrepresented
177 groups.
178
- 179 • Numerous respondents to the web survey indicated the importance to their work
180 of research-group and departmental-scale computing facilities. We define such
181 facilities as having a factor of 100 to 1000 less capability (e.g., computing,
182 storage) than is provided by the national centers. The proliferation and
183 importance of such resources suggests the need for an effective mechanism – now
184 largely lacking -- to create, nurture and support them as well as link them into the
185 national cyberinfrastructure. Further, it suggests that users view national centers
186 as needing to provide capability of order 100 to 1000 times the power of systems
187 that generally are available to individual academic departments and research
188 groups.
189
 - 190 • Comprehensive environments are needed for linking models (broadly defined to
191 include, for example, models of physical processes, data, and data movement as
192 well as learning models) from multiple disciplines and synthesizing results in
193 interoperable frameworks.
194
 - 195 • The so-called Grid, built around the Internet and World Wide Web, is an
196 infrastructure designed to provide scalable, secure, high-performance mechanisms
197 for discovering and negotiating access to remote resources. Ultimately, it should
198 allow scientific collaborators to share resources on an unprecedented scale, and
199 allow geographically distributed groups to work together very effectively.
200 Although some are critical of the Grid, particularly in light of similar and
201 apparently faster advances being made by the private sector as well as current
202 limitations in using the parallel computers now available, the Grid in some sense
203 epitomizes the type of bold, risk-laden research that could pay huge dividends if
204 successful. Not taking risk is itself a risky proposition (D. Reed, NCSA), and the
205 NSF is encouraged to continue supporting efforts such as the Grid, which among
206 other things holds the promise for democratizing computing and removing the so-
207 called digital divide between the “haves” and the “have-nots”. However, such
208 support should not come at the expense of very high-end resources.
209
 - 210 • Significant need now exists, and will be increasing significantly in the future, for
211 on-demand (i.e., not pre-scheduled or via standard queueing systems) access to
212 networks, data bases, and high-performance computers, sometimes in a dedicated
213 fashion. The current national infrastructure -- in terms of physical resources,
214 middleware, and policy -- is not equipped to handle this need.
215
 - 216 • Inexpensive and effective/reliable tools are needed to support distance
217 collaboration among multiple sites (e.g., a desktop version of the Access Grid that
218 is affordable to everyone), particularly those that allow for manipulation of data,
219 instruments, documents, etc.
220

- 221 • Higher levels of security are needed to buffer vulnerability against cyber attacks,
222 protect intellectual property, and possibly avoid compromising national interests.
223

224 **3.2.4. Emerging Paradigms and Activities**

- 225
- 226 • Across all disciplines, the need for a comprehensive cyberinfrastructure is
227 growing at a rapid pace and rapidly becoming the essential lynchpin for research
228 at boundaries *among* disciplines. Consequently, cyberinfrastructure development
229 must be driven by the needs of disciplines and multi-disciplinary interactions.
230
- 231 • The need for a new workforce -- a new information technology professional – is
232 emerging. This workforce will have expertise in a particular “domain” science
233 area, or perhaps several, but also considerable expertise in computer science.
234 Consequently, they will provide the capabilities needed to develop, maintain, and
235 integrate complex software and hardware systems – understanding both the
236 cyberinfrastructure and scientific components. Such individuals often are referred
237 to as computational scientists and represent an important link to the end user. As
238 noted below, educational institutions must develop strategies for creating this new
239 workforce and ensuring its effective integration into traditional disciplinary
240 activities.
241
- 242 • Scientific and engineering applications are covering and will continue to cover
243 even greater time and space scales (e.g., weather, which involves a coupling of
244 scales ranging from planetary waves, that last for more than a week, to individual
245 thunderstorms, which are at sub-city scale and last for one to a few hours). Such
246 multi-scale problems, often involving the coupling of different models, are
247 exceedingly complex and computationally intensive and thus reflect the clear
248 need for sustained high-end computing for the foreseeable future. However,
249 cyberinfrastructure must not be construed as only the largest and most powerful
250 resources, but rather must span the spectrum from small grants to large
251 multidisciplinary centers and projects. The NSF cyberinfrastructure portfolio
252 should reflect this philosophy.
253
- 254 • Collaboration among disciplines is growing at an unprecedented pace and now
255 includes, in some cases, hundreds of scientists working on a single project across
256 the globe. Cyberinfrastructure must support this type of collaboration in a
257 reliable, flexible, and cost-effective manner.
258
- 259 • A significant need exists in many disciplines for long-term, distributed and stable
260 data and meta data repositories that institutionalize public-domain data holdings.
261 These repositories must provide tutorials and documents on data format, quality
262 control, interchange formatting, and translation, as well as tools for data
263 preparation, fusion, mining/knowledge discovery, and visualization. A key
264 element associated with filling this need is the development of middleware and
265 related data storage strategies. Although each discipline is likely best suited to

- 266 creating and managing such repositories and tools, interoperability with other
267 disciplines is essential, perhaps through the creation of standards. Additionally,
268 greater emphasis needs to be given to the digitization and stewardship of legacy
269 data (data archeology), and to digital libraries containing collections of scholarly
270 work.
271
- 272 • More and more disciplines are expressing a compelling need for nearly
273 instantaneous access to selected data bases (both local and distributed) and related
274 services, particularly because such access often drives the collection of data
275 themselves. It is important to note, however, that the technologies for such data
276 bases do not yet exist, and that user needs cannot be accommodated by existing
277 systems (e.g., Oracle) because they generally are not suited for scientific
278 applications. This need represents a concrete example of how INITIATIVE must
279 extend well beyond the procurement of commercial technology. There is also a
280 need for on going curation of data by professionals cross trained in information
281 management and science discipline..
282
 - 283 • Users are expressing the need for nearly instantaneous access to real time data
284 streams from observing platforms or computations, where such information feeds
285 prediction models and decision support tools used in time-critical decision
286 making. The also require an ability to remotely control complex instruments with
287 exceptional network quality of service.

4. The New Cyberinfrastructure: What Changed in Computing

4.1. *An Embarrassment of Riches*

The measures of computing capability continue to grow at a literally exponential rate. We take for granted that computer speeds will rise radically with each new hardware generation, that these machines will have more memory than before, that disks will hold even more amounts of information, and that software will change annoyingly but provide more features. We will not hit physical limits for current basic chip and disk technologies until about 2010, so we assume continuation of this golden age of information technology through the period addressed by this report.

Since we have been riding these smooth exponential curves for several decades, what has changed? We have passed several practical thresholds, which means there have been qualitative breakthroughs. Suddenly, scientific research that would have been prohibitively expensive or demanded nation-scale resources can be done in an ordinary lab. PCs and workstations in the \$1000 price range can now do computations that only the biggest and most expensive supercomputers could attack ten years ago. Thus, serious computations demanding real-time visualization, simulation of interactions of thousands of particles, 2D and even 3D fluid dynamics are all possible on an ordinary desktop. Combining commodity (cheap) hardware (PC boards and networks) into a laboratory cluster costing under \$100K permits computations that only national labs could attempt 5 years ago. The entire scientific literature would fit a few hundred disks, with materials costing 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 TF³ range and data farms in the 100 TB range.

In a few more years, we will cross the “peta” (10¹⁵) line: there will be some supercomputers in the 0.1-1 PF range, there will be scientific databases containing at least 1 PB, and backbone networks will have theoretical capacities exceeding 1 Pb/s.

4.2. *Commercial Computing and the Needs of the Scientific Research Community*

4.2.1. Commercial Products and Services

The scientific research world still pushes the limits of a number of technologies and acts as a driver for some changes, but the commercial mass markets determine the computing equipment and services that are easily available, including the best programming

³ The prefix “mega” (abbreviated M) means 10⁶, the prefix “giga” (abbreviated G) means 10⁹, “tera” (abbreviated T) means 10¹², and “peta”(abbreviated P) means 10¹⁵.

Data is measured in bytes (abbreviated B), containing 8 bits of information. By tradition, networking speeds are usually measured in bits per second (abbreviated b/s). Numerical computing power is measured in Floating Point Operations per Second (abbreviated F or FLOPS).

35 language implementations, fastest chips, and largest disks. The research world has driven
36 very high-end networking and the largest computing clusters.

37 There are commercial organizations that specialize in running large computers and disk
38 farms, or in taking over entire business functions. They have developed tools and
39 methods for efficient operation to exacting contractual Service Level Agreements, so they
40 provide benchmarks or alternatives for deploying some of the cyberinfrastructure.

41 **4.2.2. The Commercial World Has Become Far More Sophisticated**

42 Computing to support businesses is far more sophisticated than it was a few years ago.
43 Businesses use very large web servers (often hundreds or thousands of processors),
44 routinely manage data warehouses (containing over 100 TB) and depend on distributed
45 computing to provide both capacity and reliability. Many of the products and services
46 that advanced businesses demand and pay for are suited to improving scientific research
47 environments. Since enterprises are increasingly network-based, security-conscious, and
48 data-intensive, there is an increasing amount of sophisticated middleware and application
49 software available that should be appropriated by the research community.

50 **4.2.3. Distributed Computing is Suddenly Real**

51 The installation of broadband standardized IP networks and growing acceptance of
52 standards for describing data and services present realistic choices about how we
53 organize our information technology resources geographically and organizationally. The
54 original ARPANet was designed to permit sharing of scarce computing resources and
55 remote access to special applications. (Its main use was e-mail and later the Web, of
56 course.)

57 The World Wide Web made distributed data a reality for most users, who don't care
58 where the information resides. The recent focus on Web Service architectures is a way to
59 share applications, and to utilize them in a way that is location insensitive i.e. distributed
60 computing.

61 The rise of the Grid [], and considerable expansion of its protocols, is another major step
62 in the direction of separating physical resources from their computational use. There is a
63 promise of setting most users free of concerns of distance between systems, and choices
64 of operating system and programming language. Grid ideas are likely to dominate the
65 high end within 5 years (even being cautious about the rate at which new concepts
66 penetrate common use). These technologies are driven jointly by the needs of scientific
67 and of commercial computing.

68 **4.2.4. Needs Unique to IT for Scientific Computing**

69 Many of the needs of the scientific research community cannot be met by the commercial
70 computing sector. Some examples are:

- 71 • Numerical algorithms and systems
- 72 • Systems assistance for scientists
- 73 • Real-time ...
- 74 • Curating of collections
- 75 • Maximal computing capabilities

76 Most commercial applications are not limited by raw computing power, while
77 simulations have literally insatiable needs for cycles and memory.

78 **4.3. Cyberinfrastructure and Technology**

79 Earlier reports [Lax, Hayes, etc.] focused on providing single computers that would
80 permit groundbreaking calculations to be performed that would enable major progress in
81 science. Such capabilities will continue to be important for maintaining the momentum of
82 scientific research as well as the U. S. leadership. However, progress in computing, and
83 changing demands by new generations of scientists who are taking new approaches to a
84 wide variety of fields are changing our views of what the national cyberinfrastructure
85 should encompass to maximize value to the scientific research community. In the future,
86 this infrastructure must include massive computing, storage, content, networking,
87 collaborative capabilities, and the software to support the needs of researchers.

88 **4.4. Hardware Trends**

89 We present more information in Appendix ???, but briefly summarize expectations here.
90 Each of the physical components has been improving exponentially (at a compound rate
91 of growth) for many years, and is expected to continue doing so for at least 6 years.

92 **4.4.1. Computational Processing**

93 Computer speeds are usually expressed in terms of how many arithmetic calculations
94 (floating point operations) they can do per second (FLOPS). In 1999, two machines in the
95 world had a theoretical capacity of 1 TeraFLOPS. By 2002, a number of universities and
96 laboratories will own computing clusters over 1 TF, and by 2005 machines up to 10 TF
97 will be relatively commonplace (and TF machines may be affordable for some individual
98 researchers). These changes are due to continued improvements of chip technology and
99 the ability to utilize clusters of chips and inexpensive computers. The speed of chips is
100 increasing rapidly, as has been true for many decades. (In early 2002, a clock rate of 110
101 GHz was announced for a silicon-based device; the fastest commercial microprocessors
102 at the same time were clocked 50 times slower.)

103

104 In earlier years, the fastest computers used fundamentally faster components (newer
105 technologies, higher cooling and powering, more complex processor designs). The
106 current state is different – the fastest chips are also among the most common, and they
107 have very complicated internal structures. Only very specialized problems currently
108 benefit from use of non-standard parts. (Some of the most technologically impressive
109 processors are found in game machines.) The commercial world continues to demand
110 more computing power, and the huge volumes support investment in new manufacturing
111 processes and designs. High-end computing is now achieved by combining the efforts of
112 very large numbers of such devices rather than trying to make unusually fast single
113 pieces.

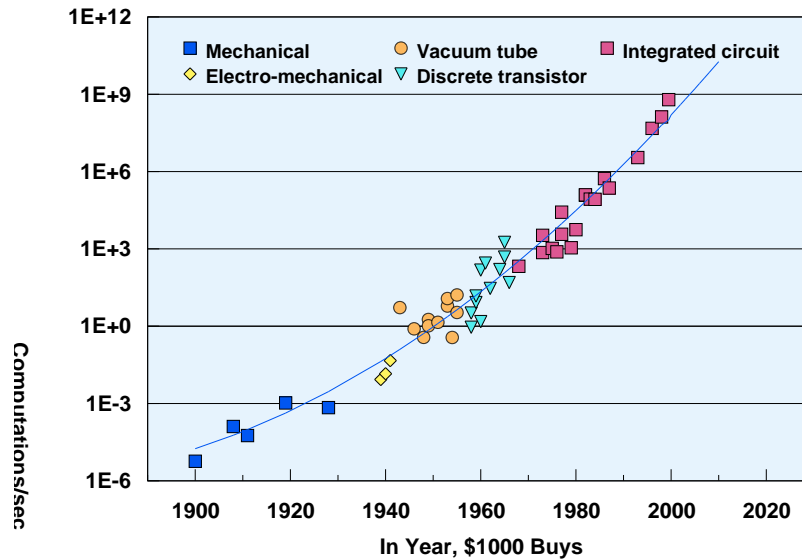
114

115 Simply counting floating point operations is not sufficient. Some “trivially parallel”
116 problems can easily be split into a huge number of independent pieces (many simulations
117 and optimizations have this property), others need a great deal of sharing and
118 synchronization among pieces (common for many engineering problems), and for some

119 there is no known way to break them up usefully. These properties of the solution
 120 technique determine the structure of the computers that can be used. The amount of
 121 memory available to a computation and the amount that can be usefully shared across
 122 simultaneous parts of the computation are another important consideration that affects
 123 what sort of computer and capacity are usable.

124

125 Computer power has been growing at a consistent, phenomenal rate for a century:



126

127

128 (The vertical grid lines on this semilog plot represent a factor of 1000, and the curve
 129 bends upward!)

130 4.4.2. Memory

131 The traditional rule of thumb calls for one byte of random-access memory (RAM) for
 132 each instruction per second of processor speed. A number of designs have much less
 133 memory than that, since there is a direct tradeoff in power and space between processing
 134 and memory.

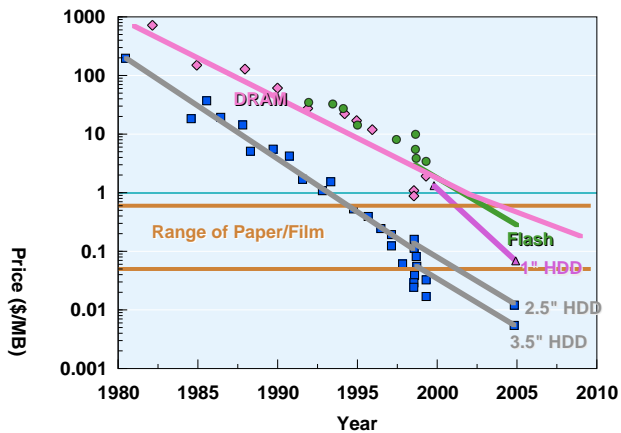
135 4.4.3. Storage

136 Many modern uses of computers depend on manipulating masses of data, far more than
 137 can be kept inside the processors. These can be observational inputs, experimental values,
 138 or results of other calculations. They can also be images or videos. Such information is
 139 usually kept on disk (though the largest archives are stored on removable optical disks or
 140 magnetic tapes). The highest performance (measured variously in total number of
 141 characters of information stored on a single device, or number of characters per volume
 142 of lab space, or in number of characters that can be retrieved per second) is generally
 143 found in the most recent standard disks. Increasing overall performance comes from
 144 utilizing many disks to store massive amounts of information, and accessing many of
 145 them at the same time.

146

147 Disk capacities (measured in bits per square inch of magnetic material) have historically
 148 increased at 60% per year, but in the past few years density has increased by about 100%
 149 per year. Prices of individual units have fallen more slowly, so most of the economic
 150 improvement has come from larger capacities. The most capacious disks in mid-2002
 151 hold around 10^{12} bits (100 GB, or 0.1 TB) of information. Databases of a few TB are
 152 common; only ones over 100 TB are remarkable.

Average Price of Storage



153
 154 It is not however sufficient to have a lot of storage: it is equally important to be able to
 155 organize, find, and combine the information on it.

156 4.4.4. Networking – Wide Area Network

157 A major shift in computing has come from the practical availability of high bandwidth
 158 data networks. Network connections up to 45 Mb/s are easily available, connections over
 159 155 Mb/s are still aggressive, and some research institutions are beginning to connect at
 160 2.5 Gb/s and more.

161
 162 Available technology can support far higher bandwidths. Deployments have already
 163 demonstrated carrying 1.6 Tb/s on a single fiber (40 channels at 40 Gb/s). Laboratory
 164 experiments have reached about 10 Tb/s on a single fiber. Switching at these speeds is
 165 not yet inexpensive, but technologies have been demonstrated. The dominant protocol is
 166 TCP/IP.

167
 168 These improvements make it plausible to move as much information as we need between
 169 sites, so that computing and storage facilities could be split or combined in a number of
 170 ways. However, the speed of light (~1 ns/ft, or about 20 ms to cross the US) is not
 171 improving, and networking technology adds further delays. For some purposes, such
 172 delays causes little problem, but some computations require much lower latency.

173 4.4.5. Networking – Local Area

174 Networks within a building are improving in two ways, bandwidth and mobility.

175 Local area network (LAN) technology is now moving to high speed Ethernets able to
176 deliver 100 Mb/s or 1000 Mb/s to the individual server or desktop. The dominant
177 protocol is TCP/IP, except for specialized protocols used to connect directly to processors
178 and storage.

179 Most current stations cannot handle such data rates effectively in 2002, nor can typical
180 laboratory switches manage many full-speed streams, but those situations will improve
181 rapidly.

182
183 The use of wireless (radio) connections is rising, both within buildings and in general
184 public. Very local access (using for example the IEEE 802.11 standards) can provide
185 many Mb/s to a single device (laptop or PDA), and new generations of cellular telephone
186 technology will permit 0.1-1 Mb/s to the roaming device in the next half dozen years.

187 **4.4.6. Displays**

188 Typical commercial displays continue to be about 1 square foot and present around 1
189 Mpel (million picture elements). These limits are being relieved. Many labs (especially
190 those on the Access Grid) combine between 3 and 15 typical displays to present a single
191 large image. Recent special displays have higher density and brightness; devices with
192 over 9 Mpel are commercially available.

193 **4.5. Software and Content Trends - Software is Still the** 194 **Bottleneck**

195 Our demands on software continue to grow, and the production of quality software to
196 meet the specialized needs of the research scientist is always later and more expensive
197 than expected. Technical computing is a relatively small market with difficult and
198 extremely challenging needs.

199 **4.5.1. Shared Middleware**

200 <MORE on types, sources, and engineering of Middleware>

201 **4.5.2. Security**

202 The massive investment we are recommending will be a critical aspect of the entire
203 research enterprise, and must be protected against accidental or malicious attack. The
204 distributed and networked nature means that problems can propagate widely and rapidly,
205 and that scientists will depend on capabilities at many sites. Furthermore, protection of
206 information is impossible without basic system security.

207
208 Modern authentication and authorization systems must be used uniformly throughout the
209 cyberinfrastructure. This requirement is unusual in traditional research environments, but
210 need not impede open collaboration and sharing.

211

212 **4.5.3. Content Management**

213 The growth of raw storage enables us to save ever larger amounts of data, but that does
214 not become useful information until it can be found, accessed, understood, and combined

215 with other information to produce scientific knowledge. It is useful to distinguish the
216 roles of storage management and content management:

217

218 Storage management focuses on data integrity (through access control, audit
219 trails, backups), availability (through operational efficiency and replication) and
220 performance (through caching, migration and reorganization, monitoring). These
221 are the basic roles of the modern (distributed) data center.

222

223 Content management focuses on collecting, standardizing, structuring, and
224 annotating data, as well as creating linkages and protecting the valuable content.

225 Curating a collection is an arduous task that requires deep knowledge of the
226 scientific discipline as well as information science and tools.

227

228 Digital libraries represent the largest scale application of these disciplines.

229

230 To be generally usable, information must be structured and described. Metadata (data
231 about the data) need to be defined so that multiple users and tools can find and utilize
232 what is in a distributed collection. The metadata themselves should be standardized,
233 ideally across disciplines, to encourage the broadest use of scientific information.

234

235 Rapid search also depends on indices, cross-references, and other types of links.

236

237 Certain types of information need to be protected specially. Personally identifiable
238 information (e.g., names of medical patients) must usually be shielded from most types of
239 access because of concerns (and laws) about confidentiality. Links between data sources
240 introduce further requirements for maintaining privacy.

241

242 Other data must be protected because of legal, proprietary, or customary constraints. It is
243 traditional to give first access to certain types of observational information to the person
244 or group who gathers it. Certain discoveries must be protected from general view until
245 patent or other intellectual property rights have been secured. Certain databases contain
246 trade secrets and are available only for a fee or by explicit permission. Such protections
247 are controversial in the research world, but the infrastructure must support them,
248 otherwise key types of information will be totally inaccessible to the community.

249 **4.5.4. Information Networking**

250 In the past few years, there have been important shifts in the ways we use networks.
251 Rather than just transmitting lots of raw bits, we are sending information that is formatted
252 in standard ways that express content (data and data that describes it) and actions to be
253 done (messages, service protocols). These lead toward the “Semantic Web” and the
254 “Grid”. A key issue will be maintaining (or introducing) the ability to share information
255 among loosely-coupled groups.

256

257 XML appears likely to be the default representation for information passed between
258 applications or systems. It has rapidly become the most common way to transmit such
259 information "over the wire". Although the basic syntax has been standardized, different

260 application domains customize their vocabularies and share dictionaries (DTDs). There is
261 thus an opportunity to improve communication between systems and among the scientific
262 disciplines, but also a chance that a Tower of Babel of DTDs will spring up.

263 **4.5.5. Collaborative Capabilities**

264 <MORE>

265 **4.6. *Investment Models***

266 Cyberinfrastructure calls for a continuing investment stream, not a one-time purchase.

267 **4.6.1. Hardware Costs**

268 Information Technology capital investments differ from most other kinds. Delaying the
269 start of construction of an accelerator or telescope or research vessel normally increases
270 the cost of the acquisition. Frequently, the opposite is true for computing equipment,
271 which becomes cheaper by waiting a year, but soon becomes obsolete. One way to
272 quantify this is through the depreciation times – major research equipment may have a
273 realistic lifetime of 10-25 years. Appropriate depreciation times for computing equipment
274 are closer to 3-5 years. The equipment does not actually wear out; it is just not worth
275 operating after that time. Furthermore, there are changes in the ways machines are used
276 and the types of computations that are wanted. As the basic unit costs of information and
277 calculation fall, new ways to get better answers or to displace scientists' time are
278 discovered, and the appropriate levels of local and national computing will change.

279 **4.6.2. Human Capital**

280 On the other hand, expertise continues to become more expensive. People with the right
281 blend of skills, drives, and training are rare, and in many cases they are the real
282 bottleneck for first-class IT use. Keeping current on computing technology is a full time
283 job; researchers in other areas are unlikely to make such a personal investment. Serious
284 software engineering to produce maintainable robust software takes focus and expert
285 knowledge, from people who have many competing job opportunities. It is a waste of
286 talent to use physics researchers as system administrators or software engineers, and
287 rarely results in high quality IT. On the other hand, computing experts are unlikely to be
288 aware of the hottest trends in other fields. The obvious solution is genuine partnerships,
289 based on pooling of needs and respect for different skills.

5. The Landscape of Related Activities

There are many different organizations defining, creating, or using forms of cyberinfrastructure. We propose that NSF should take a leading role in coordinating many of these efforts to avoid undesired duplication or conflict. Luckily, there are many opportunities for positive reinforcement. This section briefly describes several of these other activities.

5.1. Computing Industry

The previous chapter discussed some of the important computing trends. Much of the basic hardware and software will be defined by the larger worldwide computing industry. The advantage is relatively low unit cost (amortized over a huge economic base), the disadvantage is limited focus on the needs of scientific cyberinfrastructure.

5.2. Computing Research

Computer Science research activities at NSF will continue to contribute to the future needs of all users. The infrastructure centers (PACIs and Pittsburgh Terascale Computing System) continue to provide significant leadership to the community in a variety of important technologies, and should continue to do so in the INITIATIVE. In addition, federally funded national and industrial research laboratories have been highly influential in defining the directions of computer architectures, languages, databases, networks, and other core areas of computer science, as have many university and industrial projects funded by other agencies such as DARPA, DOE, and NASA.

5.3. Other Sciences

A number of grids are being established in the research world. Some of them are geographically defined, but others are aimed specifically at particular disciplines such as the North Carolina BioGrid and the University of Pennsylvania Breast Cancer Grid. There are also multi laboratory projects in fields such as environmental sciences, geophysics, and astronomy (National Virtual Observatory).

5.4. Other Federal Agencies

Several other agencies have very large scientific computing activities. Each of these requires supercomputing to satisfy their missions, and each has a larger computing budget than NSF's current plan. DOE's national laboratories (including Argonne, Livermore, Los Alamos, Pacific Northwest, and Sandia) each have used supercomputers for years (often serial number 1), and cooperate on languages, middleware, and scientific middleware. They also support significant civilian use. Many researchers use DOE facilities in addition to NSF and university facilities. The DOE National Laboratories do research with and have experience managing the largest computing systems anywhere. The DOE national laboratories have often driven the development of computer architectures for scientific computing, most recently through the Accelerated Strategic Computing Initiative (ASCI) which has commissioned the development of a succession of the world's fastest computers since 1996.

41 DoD and related agencies have purchased large computers ever since they existed, and
42 have pushed the limits of architecture and capacity through DARPA and the intelligence
43 agencies. Their needs are growing rapidly, and will influence the architectures both for
44 computing intensive and data intensive applications.

45
46 The National Institutes of Health support major initiatives in computing. Genomics and
47 modern molecular and cellular biology research depend on databases and simulations.
48 NIH researchers also spend huge sums on information technology related activities, by
49 some estimates⁴ at least 25% of the total budget. Potential organizational and technical
50 synergies are obvious. NASA has always depended on scientific computing for modeling
51 and project operations as well as for support of astrophysics research. They also took the
52 lead in production grid development in creating the Information Power Grid.

53
54 The Library of Congress and the National Archives are both pursuing large initiatives to
55 guarantee long-term access to our growing body of digital culture – especially that which
56 is born digital. They are developed advanced distributed architectures and curation
57 strategies to support this goal. There are very specific opportunities for linkages between
58 these projects and the INITIATIVE and in fact people from PACIs are already involved.

59 **5.5. Non-US Activities**

60 Major scientific laboratories elsewhere have contributed significantly to scientific
61 computing, and continue to do so. (The Web was born at CERN, just as the browser was
62 born at NCSA). For example, the UK National Grid is part of their overall eScience
63 effort, and the Netherlands National Grid has a similar purpose.

64
65 In 2001 the UK launched a major initiative called e-Science that involves the creation of
66 a national grid, development of grid middleware, and funding for applications
67 development within all of the Science Research Councils (the disciplines included are
68 roughly those in the purview of NSF and NIH combined). The European Union has
69 funded nine substantial (millions of Euros each) grid-related projects in the last two
70 years, some for infrastructure development and some for applications that take advantage
71 of the grids. In the upcoming Sixth Framework Programme, the EU is considering
72 funding a number of broader and larger scale (up to 100 million Euros) Grid projects.
73 Individual countries have significant grid efforts in place or in the early stages of
74 planning, including for example Canada, China, Italy, Japan, Korea, and Switzerland.

75 **5.6. The Ecology of Scientific Computing**

76 Many organizations will be building grids, linking them together, and supporting a wide
77 variety of computing for scientific and engineering research. In most cases, sharing and
78 cooperation are in everyone's interest – the larger the amount of information that

⁴ In recent testimony before the Committee on Science, Engineering, and Public Policy (CSEPP), NIH officials estimated that at least 25% of their annual budget is allocated to IT-related activities. In FY03 the total NIH budget is expected to be \$27B and therefore the IT-related portion is much larger than the entire NSF annual budget of about \$5B. This testimony was actually in the context of needing greater coordination and understanding of effectiveness given the huge magnitude of expenditures in this area.

79 scientists can examine, the greater the number of researchers who can work on a problem,
80 the better. It is important to induce consistent models of information, standard protocols
81 of resource sharing, common behaviors of necessary scientific services. Accomplishing
82 this will be a difficult and continual management task.

6. Partnerships for Advanced Computational Infrastructure: Past and Future Roles

6.1. *The Past and Present*

NSF has sponsored a series of initiatives to advance U.S. science and engineering by providing computational resources, the most recent of which is the Partnerships for Advanced Computational Infrastructure (PACI) program. The first initiatives began in the early 1980s, when the most powerful machines at that time---`supercomputers"--- were not generally accessible by the scientific community. Hence the predominant need was for access to computing cycles at the highest end, and as a result five NSF Supercomputer Centers were founded in 1986 and 1987.

The PACI program, established in 1997, was the next step. The PACI partnerships--- NCSA (National Computational Science Alliance) and NPACI (National Partnership for Advanced Computational Infrastructure)---were intended to fulfill significantly broader goals than only access to high-end compute power; their missions included provision of data storage and networking, education and outreach, and fostering of interdisciplinary research. At the center of each PACI partnership is a leading-edge site---the National Center for Supercomputing Applications for NCSA, and the San Diego Supercomputer Center for NPACI. The PACI program is explicitly not allowed to support basic research.

Following the guidelines of the original PACI solicitation, the activities of the PACI partnerships address multiple needs and serve multiple purposes.

- In the intervening five years, the two PACI partnerships have fulfilled their mission of providing high-end computing cycles. This conclusion is based on systematic, regularly conducted user surveys that are reported to NSF, and on the survey conducted as part of this panel's information-gathering process (Section 3).
- The PACIs have supported, engendered, and supplied software tools to help users take advantage of architecturally diverse, increasingly complex, and distributed hardware. In addition to joining and enhancing pre-existing software activities such as Globus and Condor, the PACIs have initiated diverse projects involving all aspects of high-end computing. As two examples, we mention the Access Grid, used at more than 100 sites worldwide, and the Cactus programming framework, an open-source environment that enables parallel computation on different architectures along with collaborative code development.
- Through a joint Education, Outreach, and Training activity, the PACIs have broadened access to computational science and engineering by encouraging women and under-represented groups at all educational levels.

- 44 • Some PACI-enabled collaborations between domain scientists and computer
45 scientists have been exemplars of interdisciplinary interactions in which information
46 technology is a creative, close partner with science. Selected examples of scientific
47 accomplishments associated with PACI partnerships are given in Appendix VII. To
48 name one among many, the recently funded National Virtual Observatory (see
49 <http://www.us-vo.org/nvo-proj.html>), which includes participants from NCSA and
50 NPACI, was described as a top priority in the 1999 U.S. National Academy of
51 Sciences decadal survey of astronomy and astrophysics. To a degree beyond anything
52 anticipated even five years earlier, the National Virtual Observatory links astronomy
53 with cyberinfrastructure in the forms of grid computing and federated access to
54 massive data collections. The National Virtual Observatory concept grew from
55 collaborations associated with the PACI program, and illustrates how advances in
56 computer science and information technology can inspire qualitatively new science,
57 not just traditional science that is bigger and faster.
58
- 59 • International collaboration is an inherent part of computational science and
60 engineering, and the PACIs are regularly involved with leading international
61 consortia such as the Global Grid Forum. Individual scientists supported in part by
62 PACI are major figures in visible international projects such as GridLab, which
63 involves Grid computing and numerical relativity.
64

65 The PACI Partnerships have been reviewed annually by a program review panel
66 convened by NSF. These reviews have been generally positive with respect to the
67 achievements of NCSA and NPACI as defined by the criteria of the PACI program.
68 However, despite numerous individual successes, there have been repeated concerns
69 expressed in the annual reviews about the overall effectiveness of PACI activities in
70 generic software and infrastructure for high-end computing (‘enabling technologies’)
71 and discipline-specific codes and infrastructure (‘application technologies’).
72

73 Part A of the charge to this Panel was to evaluate the performance of the PACI program
74 in meeting the needs of the scientific and engineering research communities. Given our
75 broad definition of cyberinfrastructure (see Section 2), we have interpreted this charge as
76 an opportunity to consider potential roles for the PACI partnerships in a greatly expanded
77 context. Since the Pittsburgh Supercomputing Center (PSC) was selected by NSF in
78 2000 as the site for the Terascale Computing System, we include PSC as well as the
79 PACIs in our discussion of the future. (A point-by-point response to part A of our charge
80 is contained in Appendix VIII.)

81 **6.2. Rationale for the Future**

82
83 The panel believes that today's science and engineering research requires computing
84 resources at ever-higher levels and in ever-wider dimensions (see Section 3). The need
85 remains, exactly as described in the 1995 Hayes report, for the U.S. science and
86 engineering research community to have access to machines that are 100-1000 times as
87 powerful as those available at typical research universities, and for support services to
88 enable those machines to be used most effectively. No end is in sight to the increasing

89 demand for advanced networking capabilities (including speed, bandwidth, quality of
90 service, and security). The importance of data in science and engineering continues on a
91 path of exponential growth; some even assert that the major science driver of high-end
92 computing will soon be data rather than cycles. It is crucial to provide major new
93 resources for handling and understanding data; the National Virtual Observatory (briefly
94 described in Section 6.1) emerged from a recognition that the data avalanche in
95 astronomy requires digital archives, metadata management tools, data discovery tools,
96 and adaptable programming interfaces. Finally, sustained work is needed on software
97 tools and infrastructure that enable effective general use of computing at the highest end
98 as well as on discipline-specific codes and infrastructure. It is universally agreed that
99 producing and maintaining widely usable, reliable software is at least one, possibly
100 several, orders of magnitude more difficult than generating an initial high-quality
101 prototype.

102

103 As described in Section 1, the panel is recommending a broad INITIATIVE whose goal
104 is to transform the conduct of science and engineering research. We believe strongly that
105 this initiative will succeed in addressing the needs of the scientific and engineering
106 communities only if funding of enabling and application infrastructure is organizationally
107 and functionally separate from the other activities; this is a fundamental change from the
108 all-in-one structure of the PACI partnerships. Our view is based on both philosophical
109 and practical reasons.

110

111 As a matter of principle, we are convinced that major, sustained new funding is needed
112 for discipline-specific infrastructure, and that disciplinary scientists, in close partnership
113 with computer scientists, are best able to judge the work that is most important.
114 Similarly, the quality of enabling technology projects should be assessed by experts with
115 a broad view of high-end computing who will pay attention both to opportunities for
116 complementary activities and to concerns about potentially excessive replication.

117

118 The practical motivation rests on the observation, frequently made during the panel's
119 information-gathering phase, that the PACI partnerships have made noticeably less
120 progress in producing enabling and application technologies than in providing high-end
121 resources. This disparity may have various causes:

122

- 123 • the difficulties of assembling academic teams willing to undertake the long-term
124 effort of producing software usable by others;
- 125 • the somewhat inflexible partnership structure of the PACIs, in which there are
126 large numbers of partners (leading to management complexity), as well as
127 difficulty phasing out activities of the original partners or adding new partners;
- 128 • limited review of enabling and application technology activities, particularly in
129 assessing their impact on the relevant users and communities.

129

130 Since the ultimate drivers of cyberinfrastructure are the needs of the scientific and
131 engineering research communities, the panel believes that the best enabling and
132 application infrastructure will be produced by projects that have been subjected to
133 appropriate peer review. We stress that such peer review must always include
134 consideration of the quality of each proposal's computer science and information

135 technology aspects. To be specific, infrastructure projects in application areas need to be
136 peer-reviewed by both domain and computer scientists, as are the current Information
137 Technology Research (ITR) proposals, to assess their quality based on criteria defined by
138 the needs of cyberinfrastructure for the particular scientific community. In this regard, it
139 is important that there should be no artificial distinction, as there was in the original
140 PACI program, between research and development; the best enabling and application
141 infrastructure projects, almost without exception, include both of these elements.

142

143 Enabling and application infrastructure projects can be proposed by teams from any
144 eligible institution or group of institutions, including, of course, the leading-edge PACI
145 sites. Given the expertise accumulated by the PACI partnerships, our expectation is that
146 they will be exceptionally well positioned to compete. Our hope is that NSF's
147 commitment of funds and energy to the INITIATIVE will lead to significant benefits
148 across science and engineering from the lessons learned in the PACI program.

149

150 **6.3. *The Future of the PACI program***

151

152 The panel recommends a two-year extension of the current PACI cooperative
153 agreements. After those two years, until the end of the original 10-year lifetime of the
154 PACI program, the panel believes that the two existing leading-edge sites (NCSA and
155 NPACI) and PSC should continue to be assured of stable, protected funding to provide
156 the highest-end computing resources. In addition, the two PACI partnerships should
157 continue their activities in education, training, and outreach. At the end of this period,
158 there should be another competition for the roles of "leading-edge sites", possibly
159 renamed, with (if appropriate) revised missions and structures. We also recommend
160 establishment (see Section 8) of additional high-end resource sites.

161

162 Based on the assumption that significant new funding is in place, the new, separately
163 peer-reviewed enabling and application infrastructure part of the INITIATIVE would
164 begin in 2004, after the two-year extension of the current cooperative agreements. New
165 funding is absolutely essential in order to retain highly skilled PACI staff and to maintain
166 already-established successful collaborations in enabling and application technologies.
167 As observed in Section 3, experienced and knowledgeable people are the single most
168 important component of cyberinfrastructure.

169

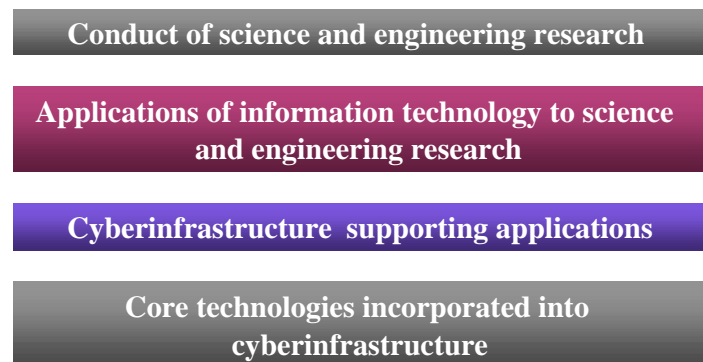
170 With this timeline---a two-year extension of the current agreements, a major infusion of
171 new funding in 2004 coupled with a partial disaggregation of functions through 2007---
172 the panel believes that stability will be ensured for the parts of the PACI program that
173 most need it. Our further hope is that this schedule will reduce the energy and anxiety
174 associated with submission of the annual program plan.

1 **7. Achieving the Vision**

2 The Panel has considered the essential elements of the INITIATIVE and appropriate
3 organizational structures to achieve it. We avoid being overly prescriptive, but rather
4 recommend basic principles, processes, and incentives that we believe will underpin its
5 success.

6 **7.1. Slicing the Pie**

7 Numerous elements must come together to achieve the goals of the INITIATIVE. To be
8 clear about what these are and how they fit together, a decomposition into major elements
9 and an associated terminology is needed (see Figure 1).
10



11 **Figure 1. A layered architectural view of the INITIATIVE.**

12
13
14 The most fundamental goal is to empower radical new ways of conducting science and
15 engineering through the *applications* of information technology. Science and engineering
16 is thus built (in part) on these applications, which are tailored to the specific needs of
17 people, groups, organizations, and communities conducting research in science and
18 engineering. Thus, the INITIATIVE directly funds activities resulting in the
19 conceptualization, implementation, and use of such applications, and is not focused on
20 cyberinfrastructure alone. Some of these applications are generic (such as those
21 supporting distributed collaboration) and many others are domain-specific (like one
22 supporting distributed community access to a large scientific instrument).

23
24 Applications are enabled and supported by the *cyberinfrastructure*, which incorporates a
25 set of equipment, facilities, tools, software, and services that support a range of
26 applications. Cyberinfrastructure makes applications dramatically easier to develop and

27 deploy, thus expanding the feasible scope of applications possible within budget and
28 organizational constraints, and shifting the scientist's and engineer's effort away from
29 information technology development and concentrating it on scientific and engineering
30 research. Cyberinfrastructure also increases efficiency and quality and reliability by
31 capturing commonalities among application needs, and facilitates the efficient sharing of
32 equipment and services.

33

34 Historically, infrastructure was viewed largely as raw resources like compute cycles or
35 communication bandwidth. As illustrated by many activities in the current PACI centers
36 and by the recent NSF middleware program, the scope of infrastructure is expanding
37 dramatically beyond this narrow definition. For purposes of the INITIATIVE,
38 infrastructure will comprise of a diverse set of technologies, facilities, and services and
39 intangibles like design processes and best practices and shared knowledge. A major
40 technological component is software that participates directly in applications and
41 software tools that aid in the development and management of applications. A critical
42 non-technological element is people and organizations that develop and maintain
43 software, operate equipment and software as it is used, and directly assist end-users in the
44 development and use of applications.

45

46 This INITIATIVE seeks to bring about dramatic and beneficial change in the conduct of
47 science and engineering research. Applications will greatly expand their role and become
48 increasingly integral to the conduct of science and engineering research.

49 Cyberinfrastructure, as it captures commonalities of need across applications,
50 incorporates more and more capabilities integral to the methodologies and processes of
51 science and engineering research. Cyberinfrastructure will become as fundamental and
52 important as an enabler for the enterprise as laboratories and instrumentation, as
53 fundamental as classroom instruction, and as fundamental as the system of conferences
54 and journals for dissemination of research outcomes. Through cyberinfrastructure we
55 strongly influence the conduct of science and engineering research (and ultimately
56 engineering development) in the coming decades.

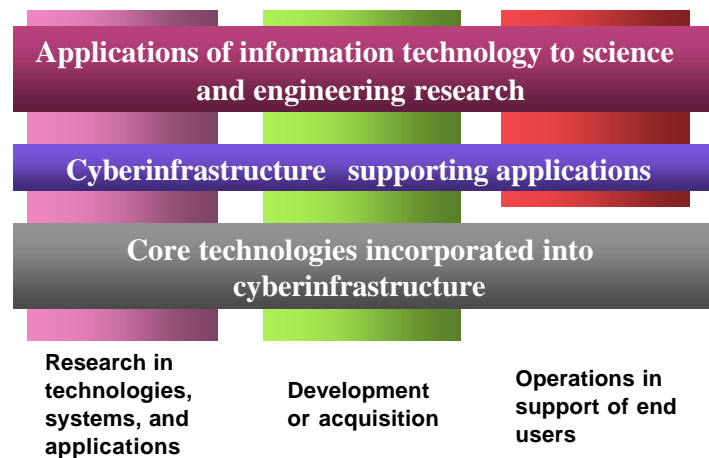
57

58 Technologists are naturally the first to embed leading-edge technologies integrally with
59 their research. The Internet—an inspirational example of this—was a new infrastructure
60 defined initially with the narrow purpose of enabling new research in distributed systems,
61 but which has now deeply impacted all research disciplines. The INITIATIVE seeks to
62 replicate this type of dramatic change across a wide spectrum of disciplines and a wide
63 spectrum of applications.

64 ***7.2. A Foundation of Technology Research and Technology*** 65 ***Transfer***

66 While this INITIATIVE is about revolutionizing the conduct of science and engineering,
67 an equally important goal is to transform information technology itself through research
68 strongly informed by the needs of science and engineering practice, and to transfer the
69 technologies so generated new uses in the science and engineering communities. To
70 illustrate this important aspect of the INITIATIVE, a second technology-transfer
71 dimension is added to Figure 1 to yield Figure 2. There are three major phases of

72 technology transfer (further elaborated and subdivided in Appendix XX): research
 73 (conceptualizing and bringing new ideas to practice), development (creating new
 74 software artifacts ready for deployment), and operations (installing these software
 75 artifacts and enabling facilities and equipment, keeping them running, and supporting
 76 end-users). These phases are all relevant to applications, to cyberinfrastructure, and to
 77 core technologies.
 78



79
 80 **Figure 2. Technology transfer adds another dimension, where operations are**
 81 **supported by development, which is based on research outcomes.**
 82

83 Our vision cannot be achieved by simply procuring existing commercial technologies. Of
 84 course, to the extent that commercial technologies and services are available off-the-
 85 shelf, they should be incorporated. But information technology is hardly mature; in fact,
 86 it is always evolving toward greater capabilities. Its applications are even less mature,
 87 and there are many opportunities to mold it more fruitfully to better meet the needs of end
 88 users. While there are many commonalities, there are also many distinctive needs of
 89 science and engineering, and these needs are not nearly as well served by commercial
 90 products as other application areas (like business processes or personal productivity or
 91 military operations). Further, science and engineering applications are often technology
 92 drivers, requiring extremes of processing and communication rates or storage capacities
 93 and longevity. Thus, research in new information technologies and applications utilizing
 94 those technologies often have important commercial spin-offs (as illustrated by
 95 supercomputing, first applied to scientific and later many commercial purposes).
 96

97 The NSF mission includes advancing information technologies and their effective
 98 application to societal needs through basic and applied research in information
 99 technology. The INITIATIVE offers a significant opportunity for research into the more

100 effective applications of information technology and opportunities for identifying and
101 refining its supporting cyberinfrastructure. Just as supercomputing and numerical
102 methods have been greatly advanced (and will continue in the future to be advanced) by
103 addressing the needs of the scientific and engineering communities, this INITIATIVE
104 will be a significant driver for a diverse suite of technologies including (but not limited
105 to) collaborative technologies, massive distributed databases, digital libraries, and the
106 preservation and exploitation of data. We expect many commercial spin-offs from this
107 research, impacting commercial science and engineering research and development and
108 other application areas.

109
110 The conduct of science and engineering is a social activity, pursued by individuals,
111 collaborations, and formal organizations. Any enlightened application of information
112 technology must take into account not only the mission of science and engineering, but
113 also the organizations and processes adopted in seeking these missions. A major
114 opportunity of the INITIATIVE is to rethink and redesign these organizations and
115 processes to make best use of information technology. In fact, this is more than an
116 opportunity, it is a requisite for success. Experience has shown that to automate existing
117 methodologies and processes is not the most effective use of technology; it is necessary
118 to fundamentally rethink how research is conducted in light of new technological
119 capabilities. Doing this effectively requires a holistic attention to mission, organization
120 and processes, and technology. This creates the need to involve social scientists as well as
121 natural scientists and technologists in a joint quest for better ways to conduct research.

122 **7.3. Some Challenges**

123 This INITIATIVE is ambitious, and as a starting point for considering the organization it
124 is helpful to recognize the most serious challenges.

125
126 **Only domain scientists and engineers can revolutionize their own fields.** At its core
127 the INITIATIVE involves rethinking the processes and methodologies underlying
128 individual scientific and engineering fields. Domain scientists and engineers must step up
129 and enthusiastically create and pursue a vision.

130
131 **Computer scientists (and allied technological fields) must be involved.** The
132 substantial and ongoing involvement of information technology specialists is required to
133 ensure that innovative new uses of technologies are identified, existing technologies are
134 molded in new ways, and research into new technologies and new applications of
135 technology is informed by opportunities and experiences in science and engineering
136 research.

137
138 **Commonalities across science and engineering disciplines must be captured.** Absent
139 appropriate levels of coordination and sharing of facilities and expertise, there would be
140 considerable duplication of effort, inefficiency, and excess costs.

141
142 **Collaboration across science and engineering disciplines must be enabled, not**
143 **impeded.** Too often information technology becomes a source of balkanization and an
144 obstacle to collaboration and to change. The goal is to make the cyberinfrastructure and

145 applications an enabler of (rather than an obstacle to) opportunistic and unanticipated
 146 forms of collaboration across disciplines, as well as encourage the natural formation of
 147 new disciplines. As in achieving commonalities, this requires a largely collective effort.

148

149 **Social scientists must work constructively with scientists and technologists.** The
 150 social scientists can assist in understanding social issues underlying the direction of the
 151 INITIATIVE, and like technologists can inform research into their own disciplines based
 152 on the experience gained.

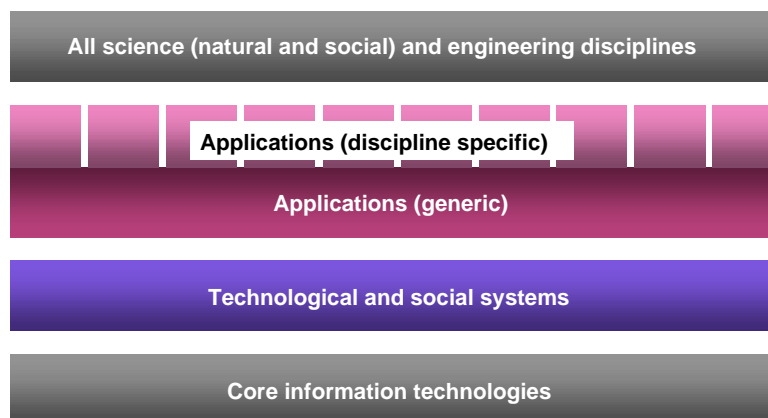
153 **7.4. Organization within NSF**

154 The INITIATIVE will be retrofitted to an NSF organization whose primary mission, the
 155 conduct of science and engineering research and education, remains unchanged. A
 156 challenge is to avoid disrupting this organization too much while successfully pursuing
 157 major changes in the organization and processes underlying its primary missions.

158

159 As a starting point, the structure of Figure 1 is modified to make it more coherent to the
 160 research disciplines represented at NSF in Figure 3.

161



162

163

Figure 3. Relationship of the layers of Figure 1 to underlying disciplines.

164

Applications are a hybrid case, as they share responsibility between technological and end-user disciplines.

165

166

167

168

169

170

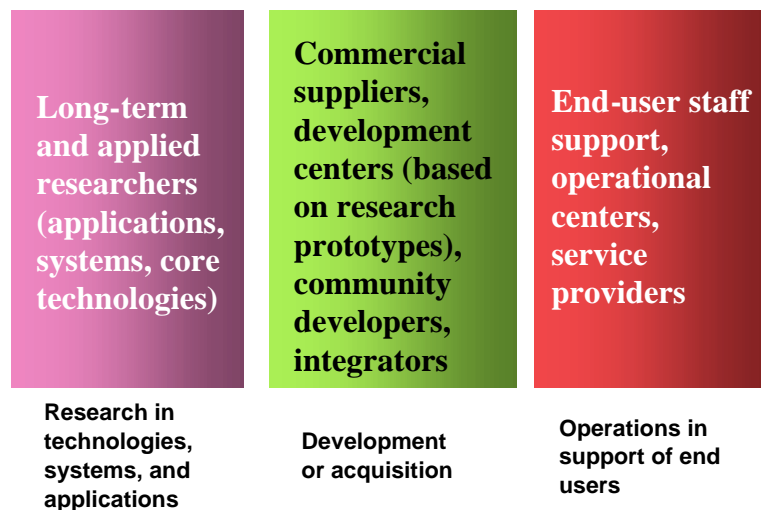
171

172 cyberinfrastructure and applications. Figure 3 also emphasizes that, in the context of the
 173 fundamentally social enterprise of science and engineering, technological systems (as
 174 defined here) and social systems (like groups, organizations, and communities) are
 175 fundamentally intertwined.

176
 177 Applications are divided into two groupings. Insofar as possible, applications should be
 178 generic, seeking to serve a variety of disciplines, but with sufficient flexibility and
 179 configurability to accommodate local variations. This contributes to both commonality
 180 (enabling future cross-discipline collaboration) and efficiency (through sharing of
 181 resources and expertise). On the other hand, there are clearly discipline-specific
 182 applications as well, and many of the organizational and process changes that accompany
 183 these applications are also specific. In this case, we rely heavily on a common
 184 cyberinfrastructure to encourage commonality and hold open the door to future cross-
 185 disciplinary collaboration.

186 **7.4.1. Organizational Principles**

187 The first division is between vision and governance of the INITIATIVE, which is the
 188 responsibility of NSF, and performance on the constituent parts (research, development,
 189 and operations), the latter illustrated in Figure 4.

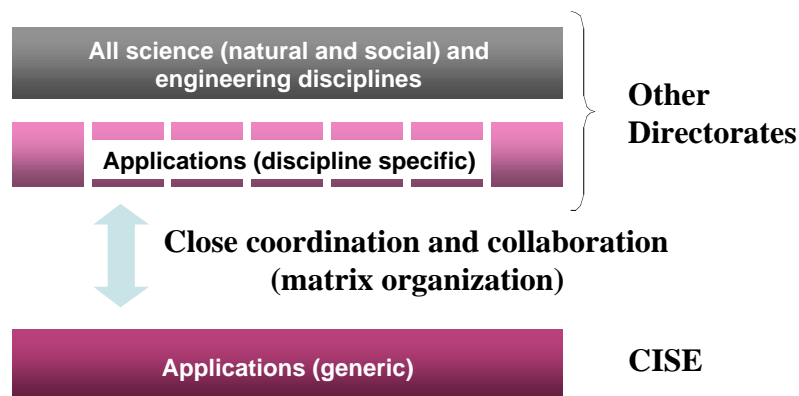


190
 191 **Figure 4. Summary of specific parties who deliver parts of the INITIATIVE.**

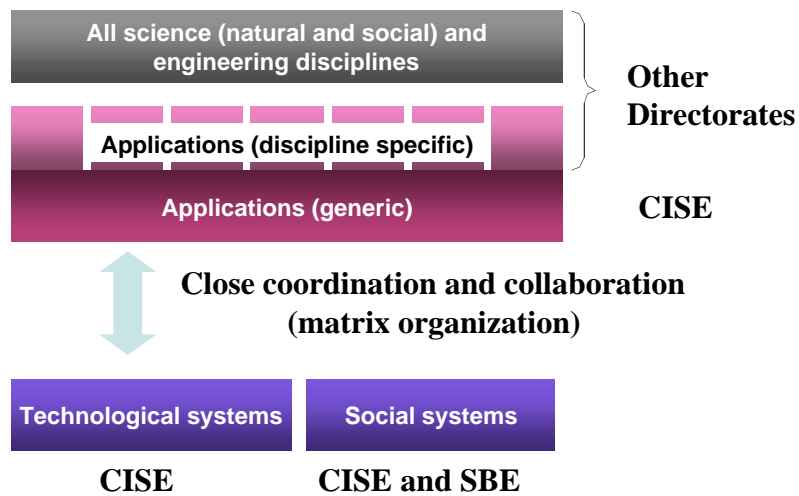
192
 193 In terms of the internal organization, our proposed division of responsibility is illustrated
 194 in Figure 5 (for applications) and Figure 6 (for cyberinfrastructure). While avoiding an
 195 overly prescriptive approach to the organization of the initiative, the Panel has identified
 196 some overriding principles, again referencing Figures 5 and 6.
 197

198 **Domain science and engineering Directorates must take the lead in revolutionizing**
 199 **their respective fields through new research organization and processes, supported**
 200 **by new applications of information technology.** We envision a program in each
 201 interested Directorate that takes primary responsibility for formulating and implementing
 202 a vision, fostering buy-in and participation of its respective scientific community, and
 203 creating a coherent program. Such efforts need to be open and oriented toward
 204 coordination with other Directorates, and emphasizing common standards and employing
 205 a common cyberinfrastructure.

206
 207 **CISE must be deeply involved as both a technology leader for the overall initiative**
 208 **and also in using scientific applications and user experience to inform its own**
 209 **technology research.** CISE should be primarily responsible for both cyberinfrastructure
 210 and generic applications, much as it has managed the PACI program. A major goal of
 211 cyberinfrastructure is to capture the major technology requirements and provide tools to
 212 aid in application development, thus minimizing technology-specific activities in other
 213 Directorates. CISE would be responsible for another major goal, insuring that the
 214 foundation of the INITIATIVE is a vibrant research agenda in cyberinfrastructure and
 215 applications rather than largely the procurement of commercial technologies. Finally, it
 216 should include and cooperate with SBE in conducting underlying research in the social
 217 aspects of both applications and cyberinfrastructure.



218 **Figure 5. Assignment of responsibility for the vision and governance of applications**
 219 **to the NSF Directorates.**
 220
 221



222
223 **Figure 6. Assignment of responsibility for the vision and governance of**
224 **cyberinfrastructure to NSF Directorates.**
225

226 These principles define an initiative distributed across most or all of the NSF
227 Directorates, including CISE, engineering, the natural and social sciences. To meet the
228 challenges of achieving commonalities and collaboration, it is critical that the constituent
229 programs within each Directorate each view themselves as a constituent within a
230 Foundation-wide initiative. The necessity for coherent overall coordination leads to the
231 third principle:
232

233 **A single leader must have fundamental responsibility for achieving these goals, with**
234 **sufficient credibility, power, and authority to succeed.** This highly qualified person
235 must be visible and highly placed, able to manage a large and complex operation with
236 very significant budget.

237 **7.4.2. Processes**

238 As emphasized in Figure 2, there are distinct activities, each making an essential
239 contribution to the INITIATIVE. One of these activities is research—a traditional
240 emphasis of the Foundation—but there are others, broadly defined as development,
241 operations, and use. These are decidedly not independent activities. Technology transfer
242 (left to right) seeks to benefit science and engineering research by employing the best
243 ideas arising from research. It needs to work the other way too—research agendas should
244 be influenced by the vision for the future conduct of science and engineering research.
245 Similarly, there is a vertical flow of ideas and influence. Applications should be
246 influenced by emerging or anticipated capabilities in cyberinfrastructure, which is
247 influenced in turn by advances in core technologies. But core technology research should

248 be informed by anticipated cyberinfrastructure requirements, which in turn is influenced
249 by capturing commonalities among application opportunities.

250
251 The research supporting applications in Figure 2 will desirably increase the collaboration
252 between computer scientists (and related disciplines) with domain scientists and
253 engineers, and also with social scientists in pursuing new applications of information
254 technology. Similarly the research supporting technological and social systems will have
255 the desirable impact of increasing the visibility of research into information technology
256 systems in the broad sense, in collaboration with social scientists—incorporating
257 processing, storage, and communication into holistic social-technical systems solutions.

258
259 It is entirely appropriate to revisit the internal organization of CISE in light of these
260 changing and magnified responsibilities. In particular, we believe that an organization
261 that mirrors the vertical structure of Figure 3 should be considered, as this would focus
262 the organization most squarely on the greatest challenges mentioned earlier. However,
263 care should also be exercised that research efforts devoted to advancing core technologies
264 receive continued high priority, as these efforts remains a critical underpinning of this
265 INITIATIVE as well as the nation’s industry and economy.

266
267 The key functions in moving technology from the research stage to uses were shown in
268 Figure 2. We expect that, following the successful Internet experience and the more
269 recent NSF middleware initiative, the development stage will focus on the productization
270 and integration of a combination of commercially available software (where available)
271 and research prototypes. The INITIATIVE must maintain a balance between deploying
272 and gaining experience with emerging technologies, while providing users with a stable
273 environment that is well documented and supported. The goal of development is to create
274 and evolve a unified software distribution that is well maintained and supported. Of
275 course, the development and operations are undertaken by experienced organizations
276 funded by NSF, normally under cooperative agreements. The longer-term goal should be
277 the commercialization of successful cyberinfrastructure and applications, always focusing
278 NSF efforts at the frontiers.

279
280 The operations stage will mix two models, as appropriate: a software distribution that can
281 be installed, operated, and supported within the end-user organizational context, and
282 software that is centrally operated to provide services over the network. NSF will support
283 organizations prepared to develop, maintain, and upgrade software distributions made
284 available to end-user organizations, and also organizations that operate
285 cyberinfrastructure and/or applications, providing services invoked over the network. A
286 proper and evolving balance should be maintained between centralized and end-user
287 operations, taking into account tradeoffs between the greater accountability and service to
288 the end-user with local staff vs. the efficiency and sharing of resources that comes with
289 centralization.

290 **7.4.3. Incentives**

291 There are three primary activities identified in Figure 2: research, development, and
292 operations (use is synonymous with research, except targeting domain science and

293 engineering rather than information technology). These have very different metrics for
294 evaluating proposals and outcomes.

295

296 **Research is a competition of ideas.** Allocation of resources starts with the program
297 announcement and evaluation of the resulting proposals. This is bottom-up, stating the
298 evaluation criteria up-front, but detailed initiatives arising from the research community.
299 Overlap or duplication is acceptable where different researchers pursue competing
300 visions for accomplishing similar ends. Post-evaluation is based on the intellectual
301 quality and impact of the research outcomes.

302

303 **Development is a competition of plans.** An overriding goal of development is to limit
304 duplication of effort, and concentrate resources on a set of integrated and maintained
305 software distributions collectively covering the scope of the INITIATIVE. Thus,
306 development is partitioned and assigned to organizations based on the responsiveness to
307 needs and credibility of their plan for pre-defined concrete outcomes. Post-evaluation
308 should be based on how effectively the plan has been implemented, and also on how
309 extensively the software is adopted and used.

310

311 **Operations is a competition for users.** Operations serve end-users, domain scientists
312 and engineering researchers, responsively providing services and support. There should
313 be two or more competitive operational options available to users, and one point of
314 evaluation should be which option attracts the most ‘customers’. Post-evaluation should
315 be based in large measure on input from the user community as to how well their needs
316 have been met.

317

318 These distinct evaluation criteria should not suggest that these activities must be strongly
319 separated organizationally; to the contrary, there may be advantages to grouping applied
320 research, development, and operations (or some subset of these activities) within a
321 common organization and geographic location.

322 **7.4.4. Continuity**

323 Human resources are critical to getting cyberinfrastructure and applications working,
324 keeping them working, and providing user support. In the past NSF has arguably under-
325 supported the recurring costs of permanent staff, preferring to focus resources on
326 acquiring ‘hard’ or ‘tangible’ assets or direct research costs. In this INITIATIVE, human
327 resources are the *primary* requirement in development and operations, and success is
328 clearly dependent on adequate funding.

329

330 Where possible off-the-shelf commercial technologies and services should be acquired,
331 but advanced and experimental capabilities will require NSF support of applied research,
332 development, and operations. Success depends on specialized skills not readily available
333 in the job market; rather, the most valuable staff will arrive with generalized
334 programming and system administration skills but learn valuable specialized skills after
335 years on the job. A starting assumption in the funding of development and operations
336 organizations should be continuity and long-term commitment. Absent significant
337 problems and negative evaluations, funding initiatives in these areas should work from a

338 base assumption of at least a ten-year lifetime for each participating organization. This is
339 not to minimize the importance of ongoing evaluation and feedback, nor is it intended to
340 preclude the redirection of funding from poorly performing organizations.
341

342

1 **8. Scope and Budget Estimates**

2 **8.1. Scope of the INITIATIVE**

3 Achieving the vision of this INITIATIVE will require *coordinated* NSF support of a
4 broader spectrum of activities and facilities than has been the case in the past and those
5 that have been supported will need substantially higher funding levels than at present. As
6 is described in Section 7, revolutionizing the conduct of science and engineering through
7 information technology and cyberinfrastructure requires the full involvement of
8 disciplinary applications teams with participation by computer scientists, and basic as
9 well as applied computer science research aimed at providing ever more effective
10 environments for supporting the applications and aimed at fundamental advances in
11 generic applications serving all science disciplines. Collecting, organizing, storing, and
12 providing access to vast quantities of data (such as observations from instruments,
13 simulation outputs, and visualization products) and other information (such as scholarly
14 publications) is becoming as important as simulation has been and will likely grow faster
15 over the next decade. To achieve the greatest benefits and broadest use of the
16 information technologies, teams should be formed whose mission is to identify and devise
17 solutions for common issues, approaches that facilitate interoperability across disciplines,
18 and capture common requirements across disciplines. The activities might be organized
19 in similar ways to the Grand Challenge projects of the last decade or the more recent
20 application-oriented ITR grants.

21
22 In what follows, each major component is described briefly and funding levels are
23 estimated for each. These components are keyed to the categories in Figure 1 of section
24 7. Appendices will include more detail on assumptions and derivation of these numbers.
25 and some “sanity checks” based upon experience in other countries or agencies.

26
27 **Our estimate is that the INITIATIVE could quickly ramp up to the effective**
28 **investment of \$650M per year of additional funding.** These funding recommendations
29 are for NSF programs only and assume that other Federal agencies and institutions will
30 continue to invest in related research and development.

31 **8.2. Fundamental, Longer-term Research in Information** 32 **Technology and its Applications.**

33 This type of research pursues revolutionary new ideas and fundamental understanding
34 specifically regarding new uses of information technology in science and engineering
35 disciplines and supporting infrastructure, without feeling constrained by the current
36 environment. [The preceding two sentences are verbatim from Dave’s draft, need to
37 reword to avoid repetition.] Ten projects at \$2 million per year each would be a good
38 starting point. More would be beneficial but it is not clear that there are enough people
39 with the requisite expertise and interests to support more efforts. These projects would
40 yield new ideas and, in some cases, research prototype implementations.

41 **8.3. Applications of Information Technology**

42 An important component of the INITIATIVE is support for scientists and engineers to
43 invest the effort required to take advantage of the new technologies and infrastructure for
44 the conduct of their research or, even more important, to find ways to use the new
45 methods and facilities to tackle research challenges that were previously out of reach.
46 The former projects will require applied research conducted over a few years. The latter
47 will be long-term research in the discipline science and possibly in computer science as
48 well. In both types of projects, discipline scientists will partner with computer scientists
49 in devising approaches to advance knowledge in new ways.
50 Models for this include the Grand Challenge awards of the mid 1990s and the
51 application-oriented large ITRs of recent years. The large number of fundable ITR
52 proposals in the last several years is a strong indicator of the latent demand for such
53 activities. Fifty additional grants at \$2 million per year will cost \$100 million. While
54 this is a substantial sum of money, these activities are absolutely crucial to achieving the
55 revolution in science and engineering that is envisioned. Time and effort and
56 collaboration with IT experts is required to learn to use new tools and to experiment with
57 new ways of applying them to specific research tasks. Unless there is support for *people*,
58 the benefits of the new technologies will not be gained.

59 **8.4. Cyberinfrastructure Supporting Applications**

60 There are many components in this category.

61 **8.4.1. High-end general-purpose centers**

62 Centers that operate very powerful computing resources for the US academic community
63 will continue to be needed. These centers will be similar to the leading-edge sites of the
64 current PACI program and would feature some or all of the facilities currently found in
65 such centers: high-end computers, large data archives, sophisticated visualization
66 systems, telecollaboration, licensed application packages, software libraries, digital
67 libraries, very high-speed connections to a national research network backbone, and a
68 cadre of highly skilled people who help users take advantage of the facilities. Since the
69 technologies deployed in these centers will be cutting edge, the support staff will usually
70 also have to develop some software to provide missing functionality in the environment
71 and to integrate the various resources and services.
72 Since progress on many applications is often paced by the size of the systems that are
73 available and by the allocation and scheduling policies, there need to be more and bigger
74 systems than the PACI program provides. The US academic research community should
75 have access to the most powerful computers that can be built at any point in time, instead
76 of an order of magnitude smaller as has often been the case in the last decade.
77 Furthermore, the number of such systems should be sufficiently large that individual
78 projects can be granted enough resource units that they could run many jobs per year that
79 use a large fraction (say at least 25% of the processors) of the computers for tens or
80 hundreds of hours. Such jobs usually access or produce vast amounts of data that needs
81 to be stored and visualized, hence the entire environment needs to be scaled accordingly.
82 The panel recommends that five to ten such centers be supported, of which three would
83 likely be the leading-edge sites of the PACI program plus the Pittsburgh Supercomputer
84 Center. While there are substantial economies of scale in operating large computers – a

85 modestly larger staff can support a much larger computer or several systems – there are
86 other advantages to having more than two or three centers. Each site tends to have
87 affinity with different disciplines or strengths in different aspects of information
88 technology. In addition, centers of this type are good training grounds for computational
89 scientists and engineers. There should also be some competition among centers for the
90 same user group as the primary mechanism for evaluating their effectiveness.
91 The estimated yearly budget for each such center is \$35 million, which is somewhat more
92 than existing NSF-funded centers receive in order to acquire much larger computers and
93 ancillary systems. The combined yearly budget for these centers would thus be about
94 \$280 million (\$210 million more than the current level).

95 **8.4.2. Data repositories**

96 Providing access to observational and other data entails far more than attaching a lot of
97 disks to a server that is on the Internet. The data need to be organized in appropriate
98 ways, metadata about many aspects of the data must be supplied, and basic manipulation
99 and analysis tools should be provided, to name a few tasks. Since access to data
100 repositories will enable important new investigations, sites should be funded to operate
101 such repositories. These data repository sites will be highly distributed because in
102 general they are best created and supported by teams from the discipline communities
103 that create and analyze the data. Those teams need to include people with professional
104 skills in the relevant aspects of IT (e.g., data bases, archival file systems, building
105 portals). Of course, as more and more research is multidisciplinary, the users of the data
106 will not necessarily be from the community that produced the data, thus increasing the
107 need to develop general data formats and interfaces to analysis tools.

108
109 Of course NSF is not the only funding source for such data repositories. For example,
110 NIH supports some biology and biomedical data collections and NASA funds many
111 archives of astronomy and remote sensing data. However, there are many data
112 repositories that should properly be supported by the NSF. One can readily imagine the
113 need for 50 to 100 data repositories. Based on current experience, each such repository
114 will require \$1.5 to \$3 million of funding per year. Note that this support does not
115 include the substantial effort required to produce clean, well-documented data that is
116 included in the holdings of the repository. This component of the INITIATIVE is
117 estimated to require \$140 million per year.

118
119 Through its strong computer science constituency, NSF is well positioned to support
120 basic and applied research on generally applicable tools and methods for
121 multidisciplinary access to diverse data collections. R&D centers could be established
122 for addressing common issues that arise in the creation and use of data collections,
123 interoperability across disciplines, and capturing common requirements across
124 disciplines. These activities will benefit data oriented research funded by all agencies.
125 Five such teams might be established at an estimated cost of \$2 million per year each, for
126 a total of \$10 million per year.

127
128 The Panel also recommends the creation of teams that would work on discipline-specific
129 metadata standards, data formats, tools, access portals, etc. They would also help select

130 and install software, e.g., for the Grid, databases. If one such effort is supported for each
131 of ten disciplines, a combined funding level of \$10 million per year will be required.
132

133 **8.4.3. Service centers**

134 In addition to sites that focus on providing access and analysis tools for data collections,
135 centers will be established that provide access to services such as applications software or
136 visualization services or perhaps just raw computing cycles on farms of workstations.
137 [NO COST ESTIMATE FOR THIS YET]

138 **8.4.4. Digital Libraries**

139 We note that the NSF played a leadership role in establishing a major R&D community
140 around the concept of digital libraries. These initiatives, especially the second round,
141 included numerous projects with the goal to both contribute to new knowledge and
142 create infrastructure and content of real use to specific disciplines (including some in the
143 humanities). We have not explicitly included funding for a digital library initiative in this
144 estimate but we assume some will continue and can be made relevant to the
145 INITIATIVE. We suggest that the topic of digital libraries should be broadened to
146 consider even larger questions about the transformation of scholarly communication in
147 the large.

148 **8.4.5. Networks**

149 Fast networks will be required to provide adequate access to the large, geographically
150 distributed data repositories and computing resources that this INITIATIVE will put in
151 place. A high-speed research network backbone should be established and the current
152 connections program extended to support suitable connections. Today one would aim for
153 a backbone capable of 40 Gb/s with large resource or user sites connecting at 10 Gb/s.
154 Over time these numbers would increase rapidly. In addition to the physical
155 infrastructure (that will be provided by commercial carriers) network R&D activities
156 must be established that concern themselves with ensuring *end-to-end* network
157 performance, data repository or computer to user, instrument to data repository, etc.
158 While some research of this type is underway, more efforts should be funded.
159

160 Assuming that on the order of 50 sites connect at 10 Gb/s and 30 at 30 Gb/s, a rough cost
161 estimate of the backbone and connections is \$42 million per year. An additional \$8
162 million for research on end-to-end performance would result in a \$50 million per year
163 investment in this component of the INITIATIVE.

164 **8.5. Core Technologies Incorporated into Cyberinfrastructure**

165 Cyberinfrastructure to support the myriad scientific and engineering applications will
166 comprise many software tools, system software components, and other software building
167 blocks. Examples of the software components include grid middleware, parallelizing
168 compilers, highly scalable parallel file systems, and sophisticated schedulers. Research
169 projects yield the concepts behind these components and usually produce research
170 prototypes. Specific examples are the Titanium compiler project at UC Berkeley, the
171 Storage Resource Broker project at SDSC, the data-cutter project at Ohio State, and the

172 Network Weather Service project at UC Santa Barbara. Such projects would be
173 supported by the first and second components of the INITIATIVE (“Conduct of science
174 and engineering research” and “Applications of information technology”). Also included
175 in this category might be software that provides higher level services such as solvers,
176 visualization, and data mining.

177

178 This part of the INITIATIVE would provide funding to turn the research prototypes into
179 widely usable products or elements of cyberinfrastructure. The NSF Middleware
180 Initiative is exemplary of the type of program required.

181

182 Much bigger investments need to be made in the task of turning research prototype
183 software into products that can be widely deployed and used. The effort required to do
184 this is much greater than that expended to create the prototype. While only a few
185 research prototypes would be selected for such development, the effort required to turn
186 them into products is likely to be at least an order of magnitude greater than was
187 expended to create the prototype.

188

189 Twenty projects at \$2 million per year each would be a good starting point. More would
190 be beneficial but it is not clear that there are enough people with the requisite expertise
191 and interests to support more efforts.

192

193 The table below provides a summary of the estimate scenario described above.

194

195 **8.6. Table 7.1 Summary of Scope and Budget Estimates for the**
 196 **INITIATIVE**

				Costs are in \$Ks				
				Yr.Cost per site or project	No. Cost	Total cost major categories		
Research in IT and its applications (in addition to ITR)				2,000	10	20,000	20,000	
Applications of IT (in addition to current ITR)				2,000	50	100,000	100,000	
Cyberinfrastructure supporting applications								
Large centers				35,000	8	280,000	280,000	
large computers				25,000				
system managers & programmers				4,000				
data archive				3,000				
visualization				750				
telecollaboration, e.g., Access Grid				150				
user support				1,500				
software development				600				
Networks								
national backbone				20,000		20,000	50,000	
end connections @ 10 Gb/s				200	50	10,000		
end connections @ 30 Gb/s				400	30	12,000		
network research (end-to-end bw)				8,000		8,000		
Data repositories				2,000	70	140,000	120,000	
Coord center for data repositories (discipline specific)				1,000	10	10,000	20,000	
STCs for data collections				2,000	5	10,000	20,000	
Digital Libraries				TBD				
Core technologies into cyberinfrastructure								
System software and tools development				2,000	20	40,000	40,000	
INITIATIVE Total								
								\$650,000

197
 198
 199

200
201
202
203
204
205
206
207
208
209
210

END OF CORE OF DRAFT 1.0