Chapter 4

Building a Scalable America

Larry Smarr*

4.1 Introduction

Digital library projects are beginning to appear everywhere across the world. They are an incredibly potent force. This chapter illustrates the importance of recognizing that with respect to the information infrastructure, many of the details currently being explored in library research cannot happen unless they are embedded in a much larger infrastructure that enables them to be used by people remotely and that enables collaboration among people who will use them.

It is important first to understand a ten year perspective of how far we have come at this point and where we are going in building not just a scalable America, but a scalable global civilization. It is not uncommon to question the existence of a connection between supercomputers and a discussion on digital libraries, but there is a connection nonetheless.

The national Science Foundation Supercomputer Center Program has been at work for about ten years. In fact, the early 1980s there was a graining of a national consensus of reports in which the supercomputer centers were created as a result of the Cold War and many have been very mission supportive like the NASA centers and the weather centers. The idea was to take this psychology and make it widely available to the university researchers or the research corporations. The problem was that there had been a period of building up the university computer centers in the 1970's, at the end of the Sputnik era. But until 1985, there had been 15 years of federal benign neglect of support for the means of employing high performance computer systems in American universities. That was rectified with the start of the Centers Program, but by that time the pool of people doing advanced applications using the latest in supercomputers had dwindled to the point where there came a need to completely replenish the national human resource pool. In fact, the NSF Centers within a few years brought 20 thousand academic researchers to state-of-the-art high performance computing systems over the Net, which was built as a result of the centers, to do supercomputing. There were no more than a few hundred that were doing it before the centers opened. At that time the perception was that going from

*National Center for Supercomputing Applications, University of Illinois

a few hundred to 20 thousand was a big step forward. Yet that was before we reached the age of Mosaic.

Currently there is a coming together of the HPCC idea with a notion of the NII, National Information Infrastructure of the digital libraries and networks, etc. There is a tremendous transition caused by the victory of the microprocessor across the board with a complete restructuring of the computer industry as a result of that victory. There is also a transition in the direction of a kind of lego approach in which the microprocessor is capable of building a scalable America, which is a network infrastructure for information in which the systems from the desktop to the teraflop are built out of the same basic units. There is software and a software transparency across this structure that enables knowledge to go beyond the boundaries that were originally anticipated.

Many think that the High Computing Communications Program is detached from the NII. Such a perception would be a reality expect for one fundamental difference between computing and any other kind of technology. The fact is that computers are exponentially increasing in power every year, while exponentially decreasing in cost per operation (addition, multiply, etc.). This trend has continued for decades and will live on for decades more. The result of such a trend is that the high performance machines that are sold for $20 million today will be in the range of $2000 to $20 thousand in the future. These follow a trend of approximately a factor of 1000 per decade.

We are witnessing the evolution of a single national agenda in which high performance computing is analogous to a time machine where high performance computing makes it possible to see today the capabilities of tomorrow's desktop systems. This eliminates the need of having to wait for the commercial vendors to bring these capabilities to the desktop before starting to develop appropriate applications. And since economic competitiveness comes form expediently moving from a new idea to an application and ultimately into the market, shortening that time scale between the two stages of idea to application can make a considerable difference in the economics of the situation.

Congress recently asked the National Academy to have the National Research Council do a study of the high performance computing and communications initiative. Dr. Larry Smarr was among the participants of that study, which heard a report in which 25 witnesses from all parts of the Federal Government said that a collision is taking place between the HPCC and the NII in which the HPCC has become increasingly supportive and complementary with respect to what is taking place in the NII. The new Blue Book, the Federal Government's blueprint for the HPCC contains many examples of this current materialization. A number of well documented examples have been published with hyperlinks to far more detailed examples throughout the world of the amazing diversity HPCC applications to information technology.

This is the famous image that Donna Cox, computer artist at NCSA, created years ago. It shows the complexity of the 1990 1.5 megabit per second backbone network by visualizing a database of dataflow over the internet. Every dot represents a university, each with thousands of computers hooked into it. The

complexity of the Net even as far back as a few years ago is staggering. But the question is the origin of the development and connection of this development to supercomputer centers.

Although ARPA developed protocols with the computer science community for many years, it was only after the formation of the NSF supercomputer centers that there became a need to hook together a fat pipe that would be the backbone of what emerged as the NSF Net from which stemmed the regional nets. These were the catalysts for the campuses digging up the quads and putting in fiber optics. It was a ripple effect that resulted in the Federal Government's investment which was necessary for continued advancement, even though the total sum of the investment was a few per cent of the actual amount that was spent overall to make the plan a reality.

Yet this pipe, the national superhighway, was only 56 kilobits per second which is equivalent to an ISDN connection that can be installed today in any home. This is a good example of the relevance of HPCC to NII, the backbone begun by NSF would not have become a reality, or would have taken many more years to develop, if it had not been for the supercomputer centers. Furthermore, software and protocols developed for that early advanced backbone at that time now make it possible for people to obtain that level of individual performance to their offices or homes. In fact, Ameritech is working with Champaign County as the first community in which anyone who has a telephone can call and get the telephone company to install an ISDN connection with a Motorola modem that makes it easy to use. By 1988 NSFnet progressed from 56 kilobits to 1.5 megabits per second. As the regional nets began to appear, it is clear that the universities were in a competitive race to figure out who could get their campus network up because they did not want to be left behind. By 1991 the NSFnet backbone capacity had gone from 1.5 megabits per second to 45 megabits per second.

We are still in the process of transitioning to the new internet in which there is a new backbone connecting the NSF centers called the VBNS. This new network is now 155 megabits per second and contains network access at four points awarded by commercial bid. In fact, in the US all of the nets have to hook into the commercial providers. Once again, the NSFnet and now all of the NSF supercomputers have their fat pipe in between each other doing experiments evaluating and developing applications, laying the groundwork for the day that this kind of bandwidth exists throughout the internet.

Although the supercomputer centers' systems are much faster than they were ten years ago, the increase in supercomputer speed is nowhere near the 3000 fold increase in bandwidth of the backbone of the internet. This kind of disproportional increase is a true example of exponential growth and what the supercomputer centers worked with to create that situation in every desktop in universities throughout the world. The public and corporations are more frequently hooking into the internet, therefore proving the presence of a relay synergy that historically created the first part of the infrastructure consisting of the data pipes and the connectivity. Yet there is still no middleware and no software that glues it all together.

Since the creation of Mosaic, which was a software tool developed at the NCSA and given to the international community on the internet, the status has changed from what existed at the creation of the first part of the infrastructure. Although the first cross platform Mac, Unix, and Windows version of Mosaic was released only 18 months ago, the transformation has been stunning. The internet itself is a network of networks. There is a new network that joins the internet every half hour somewhere in the world.

There are between 80 and 100 countries that have access to the internet. In fact, Mosaic was introduced as the first multimedia browser for the World Wide Web, which was developed by Tim Berners-Lee at CERN, a particle physics laboratory in Geneva, to help hyperlink information. That the Web and Mosaic were very easy to use in a cross platform environment lead to an unbelievable increase in their use. Now there is a new Web server every two hours thereby creating the need to browse continuously.

Another way of seeing the effect of Mosaic is to translate it into what a physicist would call a phase transition. In the short period of approximately 18 months, from the time the original Unix version of Mosaic was introduced (when Gopher was a dominant internet file organization tool), advancing to the three-platform Mosaic versions were distributed, and ultimately the release of commercial versions, there was a 100,000 fold increase of internet Web traffic.

In fact, World Wide Web traffic has now exceeded all other forms of transmission on the Net. While Mosaic was only one piece of beta or pre-release software that was put onto the Net, it made a significant change in the world and completely altered NCSA. With commercialization, Mosaic created an entire industry of the computer market segment with various companies competing. In March there were 250,000 copies of Mosaic downloaded just from the server at NCSA over the Net. There are currently over 4 million connections per week on the NCSA web server.

The NCSA Web Server Software which is also free from the is used in over 10,000 of the 20,000 to 30,000 Web servers around the world. And just the Mosaic licensees, there are about 45 companies with 120 products and we only started licensing in the summer of 1994. These figures do not take into account the Netscapes, the IBM Warps, the Net Manages, the American On-line book links and all the others that do not license. There is an entire competing industry built up around web software.

The speed of change has been unprecedented. Yet, again arises the question of what connection this phenomenon has with supercomputers. The fact is that NCSA web server growth went from zero to 4 million requests per week. Hewlett-Packard (HP) gave us a million dollar cluster of a dozen HP machines to use as a supercomputer to perform science and engineering calculations using a network of workstations. At that time the Web server was at the beginning of its development, and the fast growth in web traffic lead us to reassemble the HP supercomputer as a Web server. It is difficult to distinguish between high performance computing and the Net in this synergy.

Together, Mosaic and the Web have enabled a mass of information which currently exists online but which on one can find. While Mosaic may be compared to a digital library, it lacks a digital library's minimal requirement which would be to have the capability of finding something. Although the Web has created an interesting diversion, it does not have the capability that is expected of a real digital library. Thus, what is currently taking place and what will hopefully continue is the development of software that will organize the knowledge structure of the Web in a way that will enable one to find the desired information.

This highly technical slide shows that simply employing a Web server, a piece of software and a file system sitting on a Unix workstation somewhere with the HTTP protocol, is not enough. What is most piquant about, for instance, bringing the whole world of database technology and research to be able to help in the searching problem, is the variety of search engines that are being developed. URLs and hyperlinks represent the first stage. They do not have an object characteristic about them. The current move is towards URNs, URCs, metadata, etc. The idea is to get to a point where the user feeds information and the system organizes and indexes that information. The real enigma is in linking a distributed set of these repositories everywhere with enough smarts built around the Web server so that someone can go out onto the Web and request information on whatever topic they are researching. There is a variety of work in the digital library testbeds. There are many provocative ideas about how the world of distributed repositories is going to make it possible to find things. Again, the recurring theme is to understand the connection to supercomputers.

Another testbed being set up at NCSA, is meant for business applications employing databases. The databases that will be required to keep track of all this indexing are fairly large, particularly in the case of multimedia databases. If you have a small database and you do not need the answer back immediately, a workstation is appropriate. In fact, if you have a small amount of data and you need it back promptly, such servers as AS400s, etc. are effective. If there is a lot of data and promptness is not a priority, there is the world wide market for main frames. However, if there is a lot of data and promptness is a top priority, the only choice is to get access to a supercomputer or a large network of workstations or something to that effect. This is yet another area where high performance computing is involved.

The presence of high performance computing is clear in the area of classification. One approach to classification is traditional. It involves using key words and string searches. There is, however, another approach which is the development of the ability of indexing 400,000 abstracts of scientific papers by simply allowing the knowledge structure to emerge from it by examining within the same sentence which words seem to occur together. The result is something very distinct in the way of a classification.

Dr. Hsinchun Chen (University of Arizona), who is working with the digital library testbed in Illinois, was the largest user of the Silicon Graphics supercomputer in the two weeks leading up to the site visit. He used hundreds of hours of supercomputer time to do a concept space creation out of 400,000 abstracts.

This is not a lot of information. Yet it is highly unlikely to find a supercomputer anywhere that is big enough to do a concept creation of the whole space when one starts the process and finds that in addition to the Web servers one also has the repositories which have to be distributed to 100,000 Web servers each of which have many gigabytes of information.

Currently we have the publishing cycle which involves the notion of the user who wants information. A library has a lot of information which it references. There is also an indexer which classifies it and academic journals which ensure sure that only quality science gets published. This leads to the client who sits with the browser while the servers have the knowledge. It is clear that eventually we advance to a point where any computer on the Net can do all of these tasks.

Everyone is going to become a publisher, but quality may be sacrificed in the course of the process because of the lack of quality control. Unfortunately, there is no way to stop that. This is a peer to peer world in which these computers are peer to peer with each other, and once all the indexing and classification has been distributed out into them, a different environment will rapidly emerge.

Looking at the broad picture of the evolution of cyberspace, there is now the ability to retrieve files in a very user-friendly way. This ability stemmed from first getting the net hooked together, followed by Mosaic and the Net which ultimately lead to the current capabilities.

Now the trend is to move into this new world in which searching has become the thing and where Yahoo, Lycos and the like perform the tasks of preclassification for you. The predication, however, is that this trend, although it is extremely useful now, will be short-lived because it cannot possibly scale to the level which is necessary. The Net itself will have to scale to the appropriate level with the peer to peer model. Yet there is still a need for more research on that because thus far the main focus has been on the text, whereas text is not really what is going to be in the library. In fact, words actually occupy a minimal amount of space. Future searching will be required to handle multimedia objects, such as animations in time and space, and much other fine grained, high resolution image data.

NASA has done a lot to point us in the direction of the real issue. We are funded by one of the NASA CAN projects, the goal of which is to develop interfaces to earth observation data appropriate for, among others, school children and the general public. At the back end of are vast repositories, such as the earth observation system will produce. The earth is really big, if you can image it down to a resolution of square meters, and it is always changing. The digital library is going to have to deal with a high level of detail of granularity, in both space and time, which is many orders of magnitude beyond our capabilities today with current systems and applications.

Stepping back and looking at the broad picture of the digital library concept, it is apparent that there has developed a common misconception that it is scientists who do the astrophysics or earth sciences that are behind this. The reality, however, is that the people and society as a whole are the main players. For many decades the scientific and university community has been able to have the

luxury of isolation from that real world. That is no longer going to be sufficient for the digital library ideas as is evident by observing the Web. Science is becoming a participatory activity which calls for interaction between the scientific and the real worlds.

There were approximately one billion people that watched in real time as Neil Armstrong stepped onto the moon's surface when the Apollo moon landing took place in 1969. When the latest shuttle was up the ultraviolet telescope was one of the experiments on board. It took scientists 15 years to make that telescope a reality. But the goal of information technology is to achieve a system which makes it possible for the elementary school student, who happens to be on the internet, to get the images and an explanation of what is in the images at the same time and out of the same digital library that the scientist receives this information.

The goal is to build testbeds that will develop computer science to a point where it has practical applications for real people. If such a goal is attained, it will not matter which party is in power or whether there is restructuring of the Federal Government because people will support what they deem to be in their own best interests.

Among the beneficial capabilities of the supercomputer centers is that they can organize testbeds on multiple scales. Thus, we have gone from a machine room to the university campus to our community to large cities to the state to the country to the world and organized testbeds of which many other cities are doing their own versions. Years have been spent working in communities to create one of these testbeds.

Several months ago during a press conference led by the Illinois Governor and Lieutenant Governor, the two clicked their way through the Web world-wide in front of all the assembled representatives of the Chicago press. Last year the White House participated in bringing up in one uniform infrastructure on Mosaic and the Web all of the federal cabinet offices and independent agencies. Now Vice President Gore and his office are incorporating the new capabilities of the technology into the government reinvention and regulatory reform efforts of the National Performance Review. There are cities, towns, states that are using the latest developments in technology in a variety of ways and many countries are doing the same.

In bringing up a map of Germany and its Web sites which are represented by dots, the map is so dense with dots that it is almost impossible to map them all. This illustrates further that Web sites are a world-wide phenomenon and are a part of the notion of building a scalable America which takes place within the context of building a scalable America which takes place within the context of building a global system. Every participating country, of which there are about 80 to 100, is going through similar steps. They recognize that they can take the lead of a knowledge economy without going through the intermediate stages.

Often we have invested several hundred billion dollars into our infrastructure whereas other countries can avoid equal expense because they do not have to undo what we did, or repeat our development steps, rather they can go directly

to a wireless communication system, for example. The only advantage from the economic point of view is to be the first to get access to the newest technological advancement and to get the most number citizens involved in using this technology in a myriad of innovative ways to create value, whether it is value in the economy or value in the knowledge or value in arts, and to collaborate to form vast virtual teams which we are now seeing and of which 18 months ago we could not even conceive.

The fundamental building block for progress is the technologies that are being developed. Dreams and goals for the future with respect to digital libraries can be achieved only by using standards and the collaborative process which must interoperate within the other schemes. The current situation is atypical of the situation in regular research and universities. On the other hand, it is a situation where taking the right measures can potentially have far more pervasive support from the society as a whole than anything we have ever done in the past.