1 Grids in Context

CHAPTER

Larry Smarr

This second edition of *The Grid* marks a major turning point in the evolution of the Grid concept. The notion of linking people, computers, sensors, and data with networks is decades old, and we have seen increasingly sophisticated Grid applications in recent years. However, here we have for the first time a coherent description of the hardware, software, and applications required to create a functioning and persistent Grid.

For the Grid to be successful, we must consider not only technologies but also the *social aspects* of what will induce researchers, educators, businesses, and consumers to use the Grid as an everyday part of their work. One approach is to study the development of previous infrastructures, such as railroads and the electrical power distribution system, for clues as to how they were jump-started. Another approach is to envision how diverse subgroups of our society would benefit from the Grid and how people would change their working environment as a result of the new enabling technologies. I briefly pursue both approaches in this chapter.

1.1 LESSONS FROM HISTORY

To understand the future of the Grid, it helps to study the history of other infrastructures. The Corporation for National Research Initiatives has specifically considered the development of infrastructures for railroads, telephones and telegraphs, power and light, and banking (293–296). These are a few examples of infrastructures that are distributed, have different end-user devices and

intermediate reservoirs, and have created hundreds of billions of dollars in today's world market. Without such infrastructures, life as we know it would be impossible. As we study these infrastructures, which are all quite different on the surface, we begin to notice that they have in common a number of striking features.

The evolution of each infrastructure was incredibly complicated. There is no simple story of how each came into being; indeed, there is a good deal of disagreement on the historical details. However, one point is clear: these infrastructures did not come into being just because a bunch of researchers got together and thought it would be a cool thing to do. Private market forces and the government played fundamental roles, as did invention, research, and standardization. These infrastructures had an important local/global ordering to their time development that is not necessarily apparent, but was critical to the dynamics. When these infrastructures started, almost all traffic was local. For instance, there were robust city telephone systems, but long distance was rare. County roads were prevalent for many decades before the interstate highway system was started in the mid-1950s.

Also critical is a distribution of capacity throughout the infrastructure. When we think of the physical distribution of goods, there are local warehouses, regional warehouses, and national warehouses to enable efficient just-in-time resupply. This distributed system can work only because of the regional "caching" or storing of physical goods. The same is true, for example, in the national electrical distribution system, which ranges from huge-capacity cables leaving dams and nuclear reactors, to cross-country power lines, to local neighborhood lines, to the 110-V lines in private homes.

Such a capacity distribution is already becoming apparent in the developing Grid. In the United States, the Extensible Terascale Facility (ETF) is a good example, with five supernodes at the National Center for Supercomputing Applications (NCSA), the San Diego Supercomputer Center (SDSC), the Pittsburgh Supercomputing Center, Argonne National Laboratory, and the California Institute of Technology, with plans to extend it to a broad range of regional centers, hosting medium-sized nodes with capacity and functions integrated across the ETF by a common Grid middleware infrastructure. Networking capacity is also distributed, or hierarchical—with a 40 gigabit/sec network connecting major supercomputer centers, while the average person is using 50 kilobit/sec over a dial-up modem—a span of almost one million-fold in bandwidth. Similar structures are found within the UK, EU, Japanese, Chinese, and other national Grids. Such distributed differential capacity is characteristic of many of these other infrastructures.

Grids are almost fractal in nature: we need only to look at the billions of appliances in homes worldwide and work upward to the nuclear power generators. Grids are not flat.

1.1.1 Why Is There Chicago?

We tend to think of the world into which we were born as the way it always was, yet history reveals that this is not the case. Much of the development of societal structures is bound up with the development of infrastructure. A fascinating book on this topic is *Nature's Metropolis: Chicago and the Great West* (1992), written by William Cronon, a professor of history at Yale University. Cronon is a quantitative historian—a historian who takes data from shipping records, land sales, and so forth and reconstructs what actually happened. The issue he addresses is "Why is there Chicago?"

Chicago is an artifact of the emergence of infrastructure. Until the early 1800s, Chicago was a small field of onions on a very large lake with a lot of mosquitoes. The development of the modern metropolis of Chicago is a microstudy in the phenomenal story of how the emergence of railroads, enabling the movement of physical goods around the country, completely changed the United States in a few short decades. Cronon emphasizes that we cannot study the development of Chicago in isolation; rather, we must take a national viewpoint to capture the scale of the phenomenon. Specifically, Chicago can be understood only by considering the total transformation of the West.

Just as is happening in Brazil today, the native ecosystems of the center of the United States were rapidly annihilated and replaced by artificial ones in the last half of the nineteenth century. Bison were killed and replaced by cattle. The prairie was destroyed and replaced by monoculture agriculture (wheat or corn). The northern forests were destroyed to provide wood to build homes on the prairie where there were no trees. Cronon points out that none of this could have happened had it not been for the new railroad transportation infrastructure that allowed for goods to be shipped from deep inland to markets far away.

The Great Lakes and the rivers were the previous transportation infrastructure. St. Louis was the dominant midwestern city for a long time before Chicago came into existence. Yet as the railroads emerged and linked with the Great Lakes lines, Chicago grew much more rapidly than St. Louis did. Chicago emerged first of all as an intermediate "cache" for agricultural products. Before then, each farmer would grow his or her own corn, take bags on a wagon to a levee in St. Louis, and sell the bags individually. Chicago enabled farmers to pool their corn into a common elevator.

This fundamental change of pooling and "caching" mixed-origin grain allowed for new groups of middle people who bought and sold goods that they had not grown themselves. Financial institutions, such as the Chicago Board of Trade, came into being because a new class of financial instruments and interactions was required. Stockyards arose for the same "caching" purpose, but with livestock instead of grain. Inventions of "Grid technology" like the refrigerated railroad car,

without which meat would spoil before it got from the Chicago stockyards to the eastern consumers, allowed meat for the first time to be sold far from where the livestock was raised or butchered.

The new "middleware" infrastructure enabled the creation of new industries that were not anticipated and not derivative. For example, Chicago quickly became one of the great retailing centers of the world. Today, Chicago hosts the largest number of corporate headquarters outside New York City. Chicago has the world's second busiest airport, which emerged out of the rise of the air transport infrastructure. If there was ever an example of a city that came into existence as a result of the emergence of infrastructure, it is Chicago.

1.1.2 Why Is There Champaign–Urbana?

Champaign-Urbana, located in east central Illinois, had an analogous history. One-hundred and fifty years after the laying of the railroad tracks through Champaign County, we can find the living reminders of the social impact of the railroad infrastructure. Across Champaign County, a number of small towns are strung along the railroad track: Homer, Sydney, Philo, Tolono, Savoy, St. Joseph, Thomasboro, Rantoul, and Mahomet, each located where the grain elevators were built as grain "caches" along the railroad track. Their spacing was set by a one-day's wagon drive to bring corn to the grain elevator.

And how did Champaign-Urbana come into being? Urbana townspeople and the Illinois Central could not agree on a price for the railroad station, so the railroad simply built its station a few miles to the west of Urbana and went on south. That railroad station grew into Champaign. Today a number of shipping companies locate in Champaign–Urbana because it is on the intersection of two major east–west and north–south interstate highways. Again and again, when we look through the "lens" of infrastructure development, we understand just how powerful a social force it is.

1.2 THE GRID ON STEROIDS

We see that infrastructure has serious social consequences. Grids are going to have a similar revolutionary effect as railroads. However, instead of waiting 30 or 40 years to see the changes brought about by this infrastructure, we are seeing the changes much faster. It is not clear that people are sociologically ready to deal with this rate of change.

The world's computer and communications infrastructure is driven by exponentials: it's on steroids. Every aspect—the microprocessor, bandwidth and fiber, disk storage—is all on exponentials, a situation that has never happened before in the history of infrastructure. For example, NCSA's first Cray X-MP, purchased in 1986, cost $8 million, had its own electrical line to the Illinois Power substation, had special cooling, and had absolutely no graphics. It was connected to other supercomputers by the first National Science Foundation (NSF) network backbone with a capacity of only 56 kilobit/sec. A child's video game, such as a Microsoft Xbox, has an MIPS microprocessor with roughly 20 times the compute power of an X-MP processor and has a list price of $200. It uses 250 W, instead of 60,000. It has incredibly interactive three-dimensional graphics. Those who have 500 kilobit/sec DSL in their homes have more computer power and bandwidth than did all five NSF supercomputer centers put together only 17 years ago.

This situation is not like the railroads. It is not like anything in history. So, while we look back on history to get some guiding principles, we need to be a little humble in the face of what we are launching. Grids are going to change the world so quickly that we are not going to have much of a chance—on a human, political, or social time scale—to react and change our institutions. Therefore, there will be a lot of noise, a lot of angry people, and a lot of controversy.

1.3 FROM NETWORKS TO THE GRID

ARPANET, the precursor of the Internet, which forms backbone for the emerging information Grid, started in the early 1970s. ARPANET was an experimental network used by a few computer scientists and the Department of Defense (DoD) community. It developed critically important protocols such as TCP/IP and the notion of packet switching (analogous to the standardization of the electrical power distribution industry on AC versus DC). Various production and research networks evolved from this beginning. One of these was NSFNET, created in 1986 with a 56 kilobit/sec backbone that tied together the five (then new) NSF supercomputer centers. The backbone bandwidth of 56 kilobit/sec was upgraded to 1.5 megabit/sec, and then to 45 megabit/sec during the late 1980s and early 1990s. In 1995 the NSF transferred NSFNET to the commercial sector, which evolved it into today's Internet.

The Internet could never have arisen so quickly had it not been for the federal government funding 100% of the NSFNET backbone, partially funding the regionals, and occasionally funding the academic networking efforts. Within five years, however, the total funding by everybody else involved in the Internet was

probably a hundred times more than the federal government's initial investment. (How this worked in the case of railroads is still controversial, but the federal funding for the land ultimately turned out to be a fairly small percentage of the private funding (293).) Thus, I believe that the federal government has a critical role in bringing today's Grid into being.

Similarly, high-performance Grids will not spring into being just by waiting for private industry to do it. In 2002, NSF funded the creation of the TeraGrid, which links five major supercomputing sites with a 40-gigabit/sec dedicated optical network. With its Extensible Terascale Facility program, it is now looking to expand this new infrastructure to other research institutions. Major projects are also under way in Europe and Asia-Pacific to establish the necessary network and computational infrastructure. The cycle of initial government funding leading to a new set of commercial services has started again.

1.4 WHO WILL USE THE GRID?

We cannot assume that the Grid we are building is just going to work and instantly fill with users. We cannot make a *Field of Dreams* assumption: Build it and they will come. If nobody shows up to use it, the Grid is going to be a failure. People will use it only because they want to communicate with resources or people that they care about and need to do their work. This is an example of the local/global phenomenon mentioned in regard to previous infrastructures.

We must stimulate the use of the Grid by applications researchers, drawn from both computational science and experimental and observational science, as well as educators and visionary corporations. We must find groups of people who are globally or nationally dispersed but who have a preexisting need to interact with each other or remote resources—and are totally committed to doing so over the Grid. Their use of the Grid from the beginning will greatly shorten the time it takes for this new infrastructure to develop. If we rely on natural evolution, as witnessed by the other major infrastructures, we will wait decades for new capabilities to appear.

Computational scientists and engineers need the Grid. Computational scientists and engineers would like to visualize their applications in real time and, for many applications, steer or navigate the computation as well. It is still common practice for computational scientists to transfer the results of complex simulations by shipping tapes—and as a result, it can take days from when a simulation starts to when an error is detected. This is unacceptable. Many other researchers just cannot get

access to enough computational horsepower. They need the ability to reach out across the Grid to remote computers.

Experimental scientists need the Grid. It is important to remember that for every theoretical or computational scientist, there are 10 experimental and observational scientists. If we wish to generally influence science, we need to influence experimental and observational science. Experimental scientists want to hook up their remote instrumentation to supercomputers or to advanced visualization devices, and to use advanced user interfaces, such as Java or voice command for instrumental functions. Thus, for example, brain surgeons in Chicago could control MRI machines in Urbana and collaborate with colleagues in San Francisco while viewing and manipulating the data in real time in three dimensions.

The Berkeley–Illinois–Maryland Radio Telescope Array (BIMA), the fastest millimeter radio telescope array in the world, is located in the high desert of northern California. It is a beautiful region but quite remote from research universities. BIMA uses networks to send data to NCSA supercomputers, which effectively become the computational "lens" of this telescope. Because BIMA is an aperture synthesis telescope, the phase shifts of the radio waves coming in arrive slightly differently at each of the antennas, and a supercomputer is needed to reconstruct what the original object in the sky must have looked like. The telescope takes 1000 two-dimensional images of the sky at one time at different wavelengths, and the computer produces data cubes from them (two dimensions of space and one dimension of frequency). These visualizations can be used, in turn, to steer the instrument or to create a digital library on the fly, which people can look at remotely through Web browsers. This is a perfect example of the type of science that we would like to be able to do at much higher bandwidth, with much more robust software for a broad range of scientific instruments.

Corporations need the Grid. Most large corporations today are global in extent. While the Web has allowed for corporate intranets to arise, these are still fairly primitive in terms of the types of functionality this book discusses. For example, in the 1990s NCSA worked for several years with Caterpillar on the development of virtual environments to support virtual prototyping. Linking to computer-aided design files that define new heavy earth-moving equipment, Caterpillar has moved from standard computer graphics to heads-up displays to single stereo walls to full CAVE virtual worlds. Caterpillar prototyped teleimmersion with multiple sites, including some internationally. They have suppliers, manufacturers, and customers spread out all over the world and could imagine linking these people over the Grid.

Allstate Insurance, another NCSA industrial partner, has explored using high-speed networks to link its Menlo Park research lab with the data-mining activities going on at NCSA, where large-claims datasets reside on high-performance computers. Such a link would allow for new types of pattern recognition using a total corporate database approach. Given that Allstate has 15,000 claims offices, it imagines using Grid capabilities to unite the company into a single collaborative team. Similar stories will arise for almost any large corporation.

The environment needs the Grid. Although society is becoming politically ready to deal with large-scale environmental problems such as ozone depletion, global warming, and air and water pollution, researchers are lagging in creating trusted interactive knowledge bases that integrate all known science about these issues. Given the multidisciplinary nature of such problems, it is clear that to gather all the experts needed to study the problem, we will need the Grid.

For example, the dominant environmental threat to the Chesapeake Bay system has moved from industrial pollution to agricultural runoff. Farmers put fertilizer on their fields, which runs off in the spring into the rivers, which feeds large algae blooms, which later produce massive decaying organic masses near the bottom, which ultimately reduce the dissolved oxygen for the shellfish. This is a dynamic, nonlinear, highly coupled, multitimescale, and multispacescale multidisciplinary problem, and the only possible way to deal with it scientifically is computationally. In fact, researchers should be able to work in a collaborative computational framework that links researchers and remote sensors with the Grid, allowing for the integration of much more detailed models of chemical, biological, and physical effects and the testing of the model against field data.

Training and education need the Grid. One of the first applications of Grid technologies will be in remote training and education. Imagine the productivity gains if we had routine access to virtual lecture rooms! Currently, we often must spend several days flying across the country to deliver a lecture using an antiquated viewgraph projector in a meeting room with only 20 people in the audience. What if we were able to walk up to a local "power wall" and give a lecture fully electronically in a virtual environment with interactive Web materials to an audience gathered from around the country—and then simply walk back to the office instead of going back to a hotel or an airplane?

Given the rapid rate at which K–12 schools are joining community colleges and universities online, it will not be long before some of the more advanced schools begin to use these Grid technologies. The question is not how we use these technologies to redo classic education but what new types of collaborative education will arise using a national-scale Grid? Informal social experiments that

show how easily children can work together over the Internet are being run today with shared virtual world games, such as "Quake." The immersive game world has exploded (see Chapter 13). Children may teach us more about collaboration than we learn from university research scientists!

Nations and states need the Grid. Many countries are installing dedicated fiber or "lambda" networks internally, as if the rest of the world were not there. Within the United States, for example, states are installing "dark" fiber networks without regard for the rest of the country. As discussed previously, that is how infrastructure developments start—locally. As these regional "Tera POPs" and national and state dark fiber networks come online, the next question is "Can we interconnect them?" In the early days of the railroad, a number of different track gauges were adopted. When the train tracks extended from their local strongholds, trains could not pass from one system to another. People who are part of the statewide University of California system can interact at high speed with one another today over experimental fiber networks because they are linked institutionally by CENIC. However, what if they need to interact with researchers across the country using experimental networks? This is one of the first capabilities that a continental-scale LambdaGrid will bring to university researchers.

Of course, the Grid will rapidly evolve from research into the private sector. Already, massive networks linking supplier chains are being constructed in industries such as automobile manufacturing, and wireless extensions using RFID tags are exploding. Electronic commerce is just beginning to take off, and eliminating "speed bumps" from one state to the other will be essential for the free flow of information on which commerce is based.

The world needs the Grid. Naturally, Grid development will not stop at national boundaries. Recent proposals have advocated that goods traded over the Internet should be made tax and tariff free. Such a situation could radically change the patterns of world trade over the next few decades. Thus, there is great interest internationally in bringing advanced Grid technologies to all countries. NSF initiatives such as STAR-TAP and StarLight are facilitating the long-term interconnection and interoperability of advanced international networking in support of applications, performance measuring, and technology evaluations, while bodies such as the Global Grid Forum are establishing connections at the levels of protocols and software.

Consumers need the Grid. The functions described previously will all gradually move down the exponential to the consumer. Already we can enter virtual stores, click on merchandise, and find information about various items. Soon we will be

able to click on "Sales Help" to have a live streaming video of a store person come up on our screen, from anywhere in the country. Serious research is under way on personal avatars that, given personal measurements, can try on clothes (on a three-dimensional computer screen) to help a customer decide whether to buy an item.

1.5 SUMMARY

All the projects described in this chapter are under development and will come together as part of the Grid projects described in this book. The Grid, however, will require that we adopt new ways of working. The Grid is about collaboration, about working together. Fortunately, there is a whole new set of information technologies that come from the Web and from companies building tools to enable us to work together in collaborative online spaces. Old technologies such as proprietary video teleconferencing are giving way to open-system versions linked over the Internet, surrounded by white boards, interactive software, and recording capabilities.

More advanced efforts tailored to the needs of researchers also are under way, such as the Access Grid technology developed at Argonne National Laboratory. This technology enables shared collaborative spaces. A Web site allows all Access Grid software developers or users to interact with each other via hypernews, application galleries, and shared programs. This new approach to creating a virtual community of software developers working on a common software system (in this case, the Access Grid libraries) allows all participants to work together regardless of place. Presumably participants in these virtual communities can have a rate of progress far greater than by working in isolation.

As the new Grid technologies come into widespread use, we must shift our social patterns of interaction and our reward structure so that we can realize the potential gains from the nonlinear advancements that collaboration will create. Collaboration can be an almost magical amplifier for human work. The success of the Grid will both enable and depend on this amplification.