DISTRIBUTED AND MOBILE COMPUTING

Michael F. Goodchild[1], Douglas M. Johnston[2], David J. Maguire[3], and Valerian T. Noronha[4]

Abstract

Technology is moving rapidly to the point where computing will be available everywhere, will be fully mobile, and will provide access to widely distributed resources. This trend to itinerant, distributed, and ubiquitous computing is the latest in a long series of major architectural changes, with associated implications for where computing is performed. Production of geographic data is similarly moving to a new set of arrangements focused on local agencies and individuals. These changes are placed within an economic framework, as a basis for development of a new set of theoretical principles regarding the location of computing, and its implications for geographic information science. Digital libraries are a further instance of these trends, with associated implications for the spatial organization of libraries. The chapter ends by identifying a series of research issues, and the benefits that can be anticipated if these issues are solved.

1. INTRODUCTION

Over the past few years a large number of advances in computing and communications technology have made it possible for computing to occur virtually anywhere. Battery-powered laptops were one of the first of these, beginning in the mid 1980s, and further advances in miniaturization and battery technology have reduced the size of a full-powered but portable personal computer dramatically—in essence, it is the keyboard, the battery, and the screen that now limit further miniaturization and weight reduction in the laptop. More recently, the evolution of palmtop computers and other portable devices (Portable Digital Assistants or PDAs), as well as enhanced telecommunication devices have further stimulated the trend to mobile computing. The evolution of new operating systems (MS Windows CE and JavaSoft Java OS) and software components (MS COM and JavaSoft Java Beans are the main standards) have also had major impacts. The Internet has vastly improved inter-computer connectivity, making it possible for people to share data, visualizations, and methods while separated by great distances. Wireless communication technologies, both ground- and satellite-based, now make it possible to connect from places that have no conventional connectivity, in the form of copper or fiber. In short, we are rapidly approaching a time when computing will be:

- *itinerant*, maintaining full connectivity and computational power while moving with a

[1] National Center for Geographic Information and Analysis, and Department of Geography, University of California, Santa Barbara, CA 93106-4060, USA. Phone +1 805 893 8049, FAX +1 805 893 3146, Email good@ncgia.ucsb.edu

[2] Departments of Landscape Architecture and Geography, University of Illinois, Urbana, IL 61801, USA.

[3] Environmental Systems Research Institute, 380 New York St, Redlands, CA 92373, USA.

[4] Digital Geographics Research Corporation, 31-3600 Colonial Dr., Mississauga, Ontario, Canada L5L 5P5.

person, in a vehicle, or on an aircraft or ship;

- *distributed*, integrating functions that are performed at different places in a way that is transparent to the user; and
- *ubiquitous*, providing the same functionality independent of the user's location.

Of course we are still some way from achieving this ideal, and it is still easier and cheaper to compute in the conventional manner at a desktop using only the data, software, and hardware present in the workstation. Wireless communication is still more expensive, and more constrained by bandwidth, at least in those areas of the developed world that are well-served by copper and fiber. But the pace of change is rapid, and in some countries we are already approaching the point where the majority of telephone communication will be wireless.

New devices and methods of communication suggest the need for a fundamental rethinking of the principles of geographic information systems design. Computers are likely to be used in new kinds of places that are very different from the conventional office; to communicate information using new modes of interaction, such as the pen; and to require much better use of restricted displays and computing power. On the other hand, devices are already available that offer the power of a desktop machine in a wearable package. All of these developments challenge the research community to investigate entirely new applications, and to address technical problems.

This discussion is motivated largely by technological change, for three major reasons. First, technological change induces change in society, its practices and arrangements, and in the conduct of science, and it is important to anticipate such changes. By moving quickly in response to anticipated change, society and science will be better able to take advantage of the benefits of new technology, following the pattern that has typified the history of geographic information systems over the past three decades (for histories of GIS and the role of technology in their development see Coppock and Rhind, 1991; Foresman, 1998). Second, it is to be expected that generic technologies such as those of itinerant, distributed, and ubiquitous computing (IDU for short) will require specialized adaptation to exploit their potential fully in different domains. The geographic domain is sufficiently distinct and complex that substantial research and development activity will be needed, and substantial efforts will have to be made to introduce these new technologies into the domain's many application areas. Finally, research will be needed to explore the impacts these new technologies will have, and the issues they will raise, in areas such as personal privacy, community-based decision-making, and the accumulation of power.

IDU computing is by its nature generic, so it is not immediately clear that it merits special attention by the geographic information science community, or that substantial research problems exist that if solved will have value to the GIS application domain. But two arguments seem especially compelling in this regard. First, all three characteristics of IDU computing—the ability to be itinerant, distributed, and ubiquitous—are inherently geographic. Mobility specifically means with respect to spatial location; there are clear advantages to being able to integrate computing functions across many locations; and being able to compute anywhere in space is clearly an advantage. It is implicit in this chapter, therefore, that IDU refers to capabilities in space, rather than to capabilities in time or in any other framing dimensions. IDU computing is a geographic problem, and part of this chapter is devoted to the associated question: if computing can occur anywhere, *but must occur somewhere*, where should it occur? Answers to this question are likely to have profound influence on the geographic organization of society and its activities as we enter the information age.

Second, GIS analysis is focused on geographic location, and the ability to distribute computing across geographic locations is clearly valuable. It helps to make better decisions if the associated computing can occur where it is most helpful. Effective emergency management, for example, is likely to require that decisions be made at or close to the emergency, rather than in places more traditionally associated with computing, such as the desktop or the computer center. Collection of data on geographic phenomena is often best conducted in the presence of the phenomena, where more information is available to be sensed, than remotely, though satellite remote sensing is often a cost-effective option. Scientific research is often better done through controlled experiments in the field than through simulation in the laboratory. Thus applications of geographic information technologies stand to benefit disproportionately from IDU computing.

In summary, geographic information science and the UCGIS in particular recognize distributed and mobile computing as a significant area of research because the problems that geographic information technologies are designed to address are better solved in some places than others, and because in a distributed world it is possible to distribute the software, data, communications, and hardware of computing in ways that can convey substantial benefits.

This topic of distributed and mobile computing clearly has much in common with interoperability, another UCGIS research agenda topic, and there have been discussions over the past few years over the advantages and disadvantages of merging the two. Interoperability is clearly an important requirement for IDU computing. As will become clear from reading both chapters, the technical issues of standards and problems of semantics are discussed in the chapter on interoperability, while the emphasis in this chapter is on the geographic dimensions of computing, and on how new technologies are allowing computing to occur in different places, and in different environments, with significant implications for geographic information science and GIS.

The chapter is structured as follows. The next section addresses the nature of computing, proposing that it is best understood today as a form of communication. The following section briefly reviews the history of computing from a locational perspective, and the changes that have led to substantial IDU capabilities. The fourth section discusses the technologies of IDU computing, including hardware, software, and communications. The fifth section places IDU computing within an economic framework, to address questions of costs and benefits. This is followed by a section on distributed computing from the perspectives of libraries and central facilities location theory, in an effort to cast IDU computing within the framework of traditional arrangements for the production and dissemination of information. The chapter ends with a summary of research issues, and the benefits to be anticipated if they are solved.

2. THE NATURE OF COMPUTING

2.1 What is computing?

Computers began as calculating machines, designed to process very large numbers of arithmetic operations at speeds far beyond the capabilities of humans using tables of logarithms, slide rules, or mechanical calculators (Wilkes, 1957). Massive calculations had become necessary in many human activities, including surveying, nuclear weapons research, physical chemistry, and statistics, and the computers that began to appear on university campuses and in government laboratories in the late 1940s and 1950s permitted very rapid advances in many areas of science and engineering. Cryptography provided another well-funded and highly motivated application, in which computers were used not for arithmetic calculations but for very large numbers of simple operations on codes. In essence, the development of computing was the result of a convergence of interests between the military and intelligence communities that flourished in WWII and the early days of the Cold War, and the more general needs of science. Languages like FORTRAN reflected these priorities in their emphasis on calculation, and the representation of mathematical formulae in computer languages.

Until the 1980s the community of computer users was small, and the vast majority of the population had little need to interact directly with computing machinery. Even on campuses, the community of users formed a small elite, with its own largely inaccessible language, unusual hours of work, and seemingly endless demands for funds to purchase computing time. But the advent of the personal computer changed this dramatically. Academics with no interest in calculation suddenly found the need for a computer; and computers entered the classroom, in areas far removed from science and engineering. Computers entered the family household, and became a significant source of entertainment. Today, of course, numerical processing occupies very little of the capacity of society's computers, especially if weighted by number of users, and very few users are concerned with the code transformations of cryptography. Instead, computing is about accessing information, presenting it in visual form for easy comprehension, searching through databases, and sending information to others. Early computing was dominated by processing and transformation in the service of the user; today it is dominated by *communication,* either between users, or between the user and the machine as information storehouse.

Today, one *computes* if one uses a computing system to:

- perform a series of arithmetic operations (e.g., run a population forecasting model);
- display or print information (e.g., print a map or driving directions);
- transmit information to another computer (e.g., share a database with another user);
- search a network of computers for information (e.g., browse the World Wide Web);
- request information from a local or remote database (e.g., download an image from an archive);
- transform information using a set of rules (e.g., convert data from one map projection

to another);

and many other possibilities. All of these in some way involve communication of information, possibly combined with transformation, between computers and users. Users may communicate remotely with computers, and computers may communicate remotely with each other.

2.2 The location of computing

In the early days of computing there were no communication networks, and there were very strict limitations on the possible distance between the input and output peripherals, typically card readers and printers, and the central processing unit. While a user could travel away from the location of the input and output peripherals, there was a distinct cost to doing so: input and output had to be communicated physically, by shipping cards, tapes, or printed output, and thus incurred a substantial time penalty. Nevertheless this was often considered worthwhile, since computing capacity was so expensive relative to the costs associated with delay. Thus a user might willingly suffer a one week delay in order to obtain the processing services of a computing system at a remote campus.

Today, of course, such delays are no longer normal or acceptable. Computing has become so ubiquitous and cheap that delays are rarely experienced, and one expects to be able to connect to significant computing resources from almost anywhere. The locational pattern of computing has changed substantially in thirty years.

Nevertheless, every bit of information and every operation in a central processing unit has a well-defined location, at the scales of geographic experience (Heisenberg's uncertainty principle applies only at scales well below these). It is clear where the user is located, where bits are located on hard drives, where communications networks run, and where transformations of data occur as a result of the operation of software. From the perspective of communication, the important locations include:

1. the location of the user, where information finally resides as a result of computing;
2. the location of the user's workstation, which may travel with the user or may be fixed at some defined location such as an office desk;
3. the locations of the network used to transmit information to and from the user's workstation;
4. the locations where information is transformed or processed into the form requested by the user;
5. the locations where the necessary data are stored;
6. the locations where the data are input to the network or storage locations;
7. the locations where data are interpreted, processed, compiled, or otherwise prepared for storage;
8. the locations where the data are defined or measured;
9. the locations that are represented, in the special case of geographic data.

The last two bullets are of course particularly relevant to geographic data, since such locations always exist.

In the early days of computing all of these locations except the last must have been the same, or penalties would have been incurred through delays in shipping data on physical media. Today they can be widely separated, since data can be communicated effectively instantaneously and at little cost between locations connected by copper or fiber to the Internet, and at low cost between any other locations. A database could be in one place, fully distributed (spread over several locations), or federated (partitioned into several separate, but linked databases). In other words, computing has evolved from an activity associated with a single location to one of multiple locations, raising the interesting question: what factors determine the locations of computing?

The value of computing accrues only when information is provided to the user. As computing and communication costs fall (Moore's Law, propounded by Intel Corporation co-founder Gordon Moore, predicts that the power of a central processor chip will double every eighteen months at constant cost, and similar observations apply to storage devices), the locations of the human actors in the computing task become more and more important. At 1960s prices, it was cost-effective to move the user to the computer; but at 1990s prices, it is far more cost-effective to move the computer to the user. Today, a high-end workstation costs little more than a plane ticket, and the value of the user's time is far higher than the cost of computer rental. Of the eight tasks listed above, those involving human intelligence (1, 7, and possibly 8 and 9) now dominate the locational equation to a degree that would have been inconceivable three decades ago.

In a sense, then, the locations of the remaining tasks in the list do not matter. It makes little difference to the costs of computing whether data are stored in Chicago or Paris, given the existence of a high-speed network between them and the minimal costs of its use. However, there are still significant *latencies* or delays on the Internet, and they are strongly correlated with distance, at least at global scales (Murnion and Healey, 1998). Often the time-delay costs associated with long-distance communication on the Internet are sufficient to justify *mirroring* storage, or duplication of information at closer sites. No site on the Internet is 100% reliable, and there are therefore costs associated with expected site failure. In such cases, the costs of providing duplicate sites are perceived to be less than the costs associated with latency and down-time. Mirroring is rare within the U.S. (the Federal Geographic Data Committee's National Geographic Data Clearinghouse is a notable exception, *www.fgdc.gov*), but mirroring is more common internationally, reflecting the fact that much apparently distance-based latency is actually attributable to crossing international borders.

But while economic criteria may have little significance, there are nevertheless strong locational criteria associated with computing. Goodchild (1997) has argued that the following factors are important in determining where geographic data are stored and served to the Internet:

- *jurisdiction*, or the association between information about an area and the governmental responsibility for that area: information about a state, for example, is likely to be provided by a state-sponsored server;
- *funding*: since a server requires funding and creates a certain amount of employment, servers are likely to be located within the jurisdiction of the agency that funds them;

- *interest*: since geographic data are most likely to be of interest to users within the spatial extent or *footprint* of the data, they are most likely to be served from a location within that spatial extent;
- *legacy*: since data servers often replace traditional services, such as map libraries or stores, they often inherit the locations of those services, and the institutions that sponsored them.

Nevertheless, with the trend to out-sourcing and facility management contracts it is sometimes the case that computing occurs at a third-party location unaffected by any othese factors.

Each of the other tasks listed above also has its associated locational criteria. Compilation of geographic data often occurs in public sector agencies, such as the U.S. Geological Survey, at its national headquarters or at regional facilities. Increasingly, it is possible and cost-effective to compile geographic data locally, in county offices of the Department of Agriculture, or in City Halls, or even in the farm kitchen. Ultimately, such locations are constrained by the locations of the human intelligence that is an indispensible part of the compilation process. But with wireless communication and portable hardware, that intelligence may be best located in the field, where phenomena are best observed and where observations are most comprehensive and uninhibited.

In a world of IDU computing, therefore, the locational patterns of computing are likely to adapt to those of the human institutions and intelligence that define the need for computing and use its products. This is in sharp contrast to the historic pattern, when the scarcity and cost of computing were the defining elements. Although computing in an IDU world *can* occur anywhere, its various stages and tasks must occur somewhere, and there are both economic and less tangible criteria to provide the basis for locational choice. Exactly how, and with what long-term consequences, is a research issue that geographic information science should solve, so as to anticipate the long-term effects of IDU computing on the distribution of human activities. IDU computing may alter the locational patterns that evolved prior to the use of computers to communicate information; or further alter patterns that adapted to earlier and less flexible forms of computing, such as the mainframe.

3. THE LOCATIONAL HISTORY OF COMPUTING

IDU is the latest of a series of forms of computing technology that have followed each other in rapid succession since the early days of computing in the 1940s. In this section three phases of development are identified, each with a distinct set of locational imperatives.

3.1 Phase I: the single-user mainframe

From the 1940s through the mid 1960s computing technology was limited to mainframes, each costing upwards of $1 million, and financed by heavy charges on users based on the number of cycles consumed. Each user would define a number of instructions, to be executed in a batch during an interval while the user had been granted control of the machine.

High-speed communication was expensive and limited to very short distances. In essence, then, the locations of computers were determined by the distributions of their users in patterns that are readily understood within the theoretical framework of *central facilities location theory*. In classical central place theory (Berry, 1967; Christaller, 1966; Lösch, 1954)) and its more recent extensions, a central facility exists to serve a dispersed population if the demand for the service within its service area is sufficient to support the operation of the service. The minimum level of demand needed for operation is termed the *threshold*, measured in terms of sales for commercial services or size of population served for public services. The distance consumers are willing to travel to obtain the service or good is termed its *range*.

In principle, mainframes were spread over the geographic landscape in response to the distribution of demand. In practice, this demand was clustered in a few locations, such as university campuses. A few users not located in such clusters were willing to travel to obtain computing service, or to wait for instructions and results to be sent and returned by mail, because no alternative was available. The pattern of mainframes that emerged was thus very simple: one mainframe was located wherever a cluster was sufficiently large or influential to be able to afford one. In time as demand grew and the costs of mainframes fell it became possible to locate multiple mainframes where clusters were sufficiently large. In summary, the characteristics of Phase I were:

- very high fixed costs of computers;
- a highly clustered pattern of users;
- costs of travel for those users not located in large clusters that were low in relation to the costs of computing.

These conditions became increasingly invalid beginning in the mid 1960s, and today it is difficult to identify any legacy of this period, with the possible exception of certain large scientific data centers which still occupy locations that were determined initially by the presence of mainframes, and which still benefit from substantial economies of scale through co-location of staff and servers.

3.2 Phase II: the timesharing era

By the mid 1960s, developments in operating systems had made it possible for computers to handle many users simultaneously, through the process known as time-sharing. Although very large numbers of instructions could be executed per second, only a fraction of these would be the instructions issued by one user. As a result, it became possible for users to issue instructions and receive results *interactively* over an extended period of time, without being constrained to batch operation. This mode of operation required only relatively slow communication speeds between users and computers, speeds that could be supported by existing teletype technology over standard telephone lines. *Terminals*, consisting initially of simple teletype machines and evolving into combinations of keyboards and cathode ray tube displays, provided for local input and output. More sophisticated displays appeared in the 1970s that could display simple graphics. Only batch and advanced graphics applications required high-speed communication and thus were restricted to very short separation between user and computer.

Time-sharing changed the locational criteria of computing substantially. Computers were still massive mainframes, representing very large investments with high thresholds. But the range of their services increased dramatically. Users were no longer required to pay the costs of travel, but could obtain computing service through a terminal and a simple telephone line connection. Because a dedicated connection was required it was difficult to justify toll charges for interactive computing, but connections were free within local calling areas.

Nevertheless, batch interaction remained attractive well into the 1980s. Large data sets still had to be stored on site at the mainframe computer, since it was not possible to communicate large amounts of data in reasonable time using slow teletype technology. Remote use was feasible only for programming, the input of parameters, and relatively small amounts of output. As a result, there was little incentive to change the locational patterns that had evolved in Phase I, except in a few specialized cases.

3.3 Phase III: the workstation era

The early computers of the 1940s used vacuum-tube technology, required massive cooling, and were highly unreliable because of limited tube life. Very high costs were incurred because every component required manual assembly. Reliability improved substantially with the introduction of solid-state devices in the 1950s, but costs remained high until the invention of integrated components, and their widespread adoption beginning in the 1970s. Today, of course, millions of individual components are packaged on a single chip, and chips are manufactured at costs comparable to those of a single component of the 1950s.

Very-large-scale integration of components led to rapid falls in the costs and sizes of computers through the 1970s, until by 1980 it had become possible to package an entire computer in a device not much larger than a terminal, and to market it at a cost of a few thousand dollars. The *threshold* of computing fell by three or four orders of magnitude, until a single user could easily justify the acquisition of a computer, and the *personal computer* was born. The *range* also fell close to zero, because computing resources had become so common that it was almost never necessary for a user to travel to use a computer. Portable and laptop computers, which appeared quickly on the heels of the desktop workstation, removed the need to travel to a computer altogether. Vast numbers of new users were recruited to computing, the range of applications expanded to include such everyday tasks as word processing, and computers appeared as standard equipment in the office. Because of the economic advantages it took very little time for the necessary changes to be made in work habits: the stenography positions of the 1960s quickly disappeared, and keyboard skills became an essential part of most desk-job descriptions.

Nevertheless the mainframe computer survived well into this era. Early workstations had much less power than their mainframe contemporaries, could store and process relatively small amounts of data, and had limited software.

3.4 Phase IV: the networked era

Although the idea of connecting computers had its origins in the 1950s, and although many of the technical problems had been solved by the 1960s, high-speed communication networks capable of carrying bits at rates far above those of conventional telephone networks finally became widely available at reasonable costs only in the 1980s, with the widespread installation of fiber, microwave, and satellite links. Fiber networks were installed on university campuses, through industrial sites, and between major cities, although connections to the individual household are still largely restricted to telephone networks operating at a few thousand characters per second. Internet technology permitted bits to flow over these hybrid networks in ways that were essentially transparent to the user.

Computer applications evolved quickly to take advantage of the development of networking. *Client-server* architectures emerged to divide computing tasks between simple client systems owned by users, and more powerful servers owned by a range of providers. For many applications, software and data could reside with the server, while instructions were provided by the client and results returned. The World Wide Web represents the current state of evolution of client-server technology, with powerful servers providing services that range from sales of airline tickets and information about movies to geocoding and mapping.

Today's communication networks present a very complex geography (Hayes, 1997). A computer that is located at a node on the network *backbone* may be able to communicate at speeds of billions of characters per second, while another location may require telephone connection to a high-speed node, restricting communication to a few thousand characters per second. Other locations may lack fixed telephone connection, so communication will have to be carried over wireless telephone links. The most remote locations will be outside the coverage of wireless telephone systems, and will have to use relatively expensive and unreliable satellite communication. Economies of scale are highest for the backbone, resulting in very low costs per user or per bit; while the costs per user or bit of the so-called *last mile* or most peripheral connection may be far higher. The location theory of networks (e.g., Current, 1981) provides a comprehensive framework for optimization of communication network design.

In summary, four phases of evolution of computing have completed a transition from a location pattern based on provision of service from point-like central facilities of high fixed cost, to a pattern of almost ubiquitous, low-cost facilities located with respect to a fixed communications network. In Phase I, computers established themselves wherever a sufficient number of users existed; in Phase IV it is connectivity, rather than the existence of users, that provides the most important economic determinant of location, along with a large number of less tangible factors. Over the forty-year interval the costs of computing have fallen steadily; in the context of GIS, the cost of hardware and software to support a single user has fallen from around $100,000 to $100 in today's dollars.

The next section discusses the nature of communication technologies in more detail, and also introduces new technologies of computing that extend its significance in IDU applications.

4. IDU TECHNOLOGIES

4.1 Communications

As noted earlier, the first widely available form of communication with computers was provided by a simple adoption of existing teletype technology, which allowed characters to be sent over standard telephone lines using ASCII code, at rates of a few hundred bits per second (essentially determined by the rates of typing and printing achievable in mechanical teletypes). The coded signals of the teletype were made compatible with the analog technology of telephone networks by use of *modems* (modulator/demodulator), which converted streams of bits into acoustic signals. Today, rates as high as several thousand characters per second have become routine through incremental improvements in telephone systems.

Local area networks (LANs) are communication systems that connect computers over limited geographic domains, such as a single office complex or a campus, using combinations of copper wire and fiber, and achieving rates of millions of bits per second. They are largely transparent to users, to whom the network appears as an extension of the desktop workstation, with additional storage devices and processors.

Over the past ten years the Internet has become the most widely known instance of a *wide area network* (WAN), with services that approach those of the LAN in connectivity and transparency. Indeed it has become common to compare the Internet and the computers connected to it to a single, vast computer. Internet services such as the WWW provide additional layers of functionality.

At the periphery, connection is provided by a series of technologies that are much less reliable, and less and less transparent to the user. When computers move out of the office and into the vehicle or field, data communication can take place only over wireless media. This currently presents implementation hurdles, increased cost, and constraints on data communication rates. But the field of wireless communication is developing rapidly, both in the technologies available, and in the number of subscribers. In developing countries where the telephone cable infrastructure is not extensively developed, wireless voice telephony is attractive, and increasingly cost-competitive with traditional wire line. It is reasonable to expect that within a decade, wireless technologies will be far more advanced and more readily available, facilitating an explosion of IDU computing.

4.1.1 Technology

Wireless communication relies on radio waves, which are part of the spectrum of electromagnetic radiation (microwaves, visible light and x-rays are examples of waves in other ranges of the spectrum). The frequencies currently employed for data communication are about 300 kHz to 6 GHz, corresponding to wavelengths of 5cm–1000m. Weak electrical signals—music broadcasts, taxi dispatcher instructions, GPS transmissions from satellites, or cell phone conversations—are loaded onto stronger radio waves by a process called *modulation*. The wave is transmitted at a certain frequency; a receiver tunes in to this frequency, demodulates the wave and retrieves the electrical signal. The signal fades while traveling through the air, weakened by interference with other radio waves, and confused by bouncing off physical obstacles. Intermediate *repeater* stations may therefore amplify and relay the transmission as required. Low frequencies travel further; higher frequencies require more repeaters, which translates to more physical infrastructure. The history of wireless communication reflects a progression from lower to higher frequencies.

In early 20th century equipment, a *channel* width was about ±60 kHz; current equipment is more precise, with widths in the range of ±10 kHz, with ±60 kHz widths reserved for data-rich transmissions such as video. Transmissions on adjacent frequencies interfere with each other, hence even a two-way communications channel employs two well separated frequencies, one for each direction. Clearly there is a limit on the number of simultaneous transmissions that can be accommodated within the confines of the radio spectrum. A government body (the Federal Communications Commission or FCC in the U.S.) governs the use of frequencies. In the early days of radio, frequencies were assigned only to emergency services and public agencies. It is only over the last 30–40 years that the public have been allowed to transmit in wireless media, and this has led to unprecedented demand for finite spectrum space. Two broad approaches are used to conserve the radio spectrum: (a) a frequency is assigned to a defined geographic area, typically 10–50 km in radius, and the same frequency can then be re-used at several other distant locations—this is the basis of cellular telephony; and (b) messages are multiplexed, that is, divided into time or other slices, and combined into a single stream with other messages on the same frequency.

An important distinction is that between analog and digital media. In *analog* transmission, the electrical signal most closely resembles the original acoustic profile as spoken into the microphone—picture this as a smooth sine wave. Interference degrades the signal as it travels, and while repeaters can amplify the signal, they do not improve the signal-to-noise ratio. In *digital* transmission, the signal is quantized into discrete values by a process called Pulse Amplitude Modulation—picture the result as a wave stair-cased into a rectilinear path. About 8000 samples are taken each second, so that acoustic degradation, while perceptible, is not overwhelming (by comparison, music on a CD is sampled 44,000 times per second). The next step is Pulse Code Modulation, whereby the sampled wave amplitude values are converted into binary strings. The strings are organized into frames or packets, with origin and destination tags, and auxiliary data to enable error detection and correction. Due to the error correction ability, fidelity of the signal can be maintained. Repeaters receive a corrupted signal, correct it digitally and re-transmit the clarified signal. Digital signals are easily encrypted and stored, and fraudulent use can be limited.

Note that there are few boundaries between voice and data in terms of mode of carriage. Voice can be transmitted digitally, and conversely data can be transmitted by analog technology. There are several operational technologies for wireless communication: radio beacons, AM/FM, shortwave and TV broadcasts, two-way radios (as used in taxi dispatch), walkie-talkies and citizen’s band radios, pagers, cordless and cellular telephones. They are distinguished by the portion of the radio spectrum they occupy, the power of transmission, and the effective speed of data transfer. Below we discuss three of the most likely technologies for digital data exchange: cellular telephones, spread spectrum, and FM subcarrier.

Cellular. Cellular telephony is currently the most popular medium for private two-way voice and data communication. An area is organized into honeycomb-like cells, and a base station in each cell operates a transceiver (transmitter and receiver) with an operating radius of 10–50 km. Micro-cells and pico-cells can be set up in specific zones such as tunnels and buildings. Mobile phones communicate with the transceiver, which in turn is connected to the wired network. As a mobile unit nears the cell boundary, the base station senses a reduced signal strength, and hands off control to the neighboring cell. Within a fraction of a second, the frequency of communication changes, and the call resumes, the switch being transparent to the user.

The first wireless telephones were offered in the 1940s, but the concepts of cells and frequency re-use were developed later, and it was only in the late 1970s that automatic switching was sufficiently developed, and licensing authorities permitted widespread public use. North American service began in 1983 with the ~800 MHz Advanced Mobile Phone System (AMPS), which remains the most popular service. It is primarily an analog system, but recently there have been digital outgrowths.

Analog Cellular. To transmit digital data over an analog network, the data are modulated into audio signals (the "chirps" heard when a modem or fax machine establishes a connection); the audio is then transmitted in exactly the same way as a voice. Transmission rates are low, in the 2400–9600 bits per second (bps) range, and analog is subject to fading and other problems described above.

An alternative was introduced in 1993: Cellular Digital Packet Data (CDPD) or Wireless IP. CDPD is carried mostly over the analog cellular network. Data are organized into Internet Protocol (IP) *packets*, which are transmitted in short bursts over analog lines during lulls in voice traffic. Transmissions hop between channels in search of vacant slots. CDPD is currently one of the most effective forms of wireless data exchange, particularly for intermittent transmission, up to 1 kb at a time (circuit switching is more appropriate for continuous data communication). CDPD operates at 19.2 kbps; actual throughput rates are 8–11 kbps. Encryption, combined with channel hopping, make CDPD extremely secure. Service is billed by the number of packets transmitted, rather than air time, making it particularly appropriate for multiple mobile units (e.g. vehicle fleets). CDPD also operates on digital networks, on dedicated frequencies.

Digital Cellular. Digital service is new in North America, largely because the continent was relatively well served by analog AMPS in the 1980s; by contrast, in Europe, multiple analog protocols were implemented simultaneously, and a lack of standards inhibited interoperation between countries. Europe therefore took an early lead in the switch to digital technology. The Groupe Speciale Mobile (GSM) established a standard that has now been adopted in many countries. Unfortunately GSM is only marginally compatible with AMPS. Hence North America went a different route with digital cellular in the 1990s, employing voice digitization and multiplexing to create Digital-AMPS (D-AMPS), which was backward-compatible with AMPS. There is also some GSM in America, under the label PCS-1900. GSM is currently the only digital cellular technology in America that supports data transmission (D-AMPS does not), albeit at a relatively slow 9.6 kbps. In this context it is worth noting that the term PCS is used with liberty, and that some services sold under the title PCS are based on D-AMPS technology. Figure 1 summarizes the cellular options in North America today.

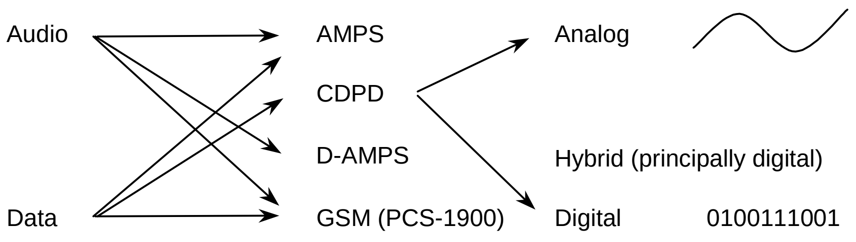

Figure 1. Cellular options in North America. AMPS is the most widely available.

Spread spectrum. The principle of spread spectrum (SS) is to spread a signal very rapidly over a wide range of frequencies (similar to frequency hopping in CDPD), according to a specific pattern. If a receiver knows this pattern, it can de-spread the signal and recover the data. Current technology allows frequency hopping at a rate of about 40 per second. The method is resistant to jamming and interference, and is inherently secure because the receiving party must know the frequency hopping schedule. For this reason, although SS was originally developed about 50 years ago, it was largely restricted to the military until about 1985. It is now growing rapidly because of the high data rates it enables, and because it allows multiple users to share the same frequency space. Wireless wide and local area networks are being developed based on spread spectrum, and it is being proposed as a generic means for users to access Information Service Providers (ISPs). The drawback is cost, which is currently high. Three radio bands are now reserved for spread spectrum, at approximately 900 MHz, 2400 MHz, and 5.8 GHz, and licensing requirements are minimal.

FM sub-carrier technology is the basis of the Radio Data System (RDS) in Europe, or Radio Data Broadcast System (RDBS) in the U.S. Just as CDPD makes use of unused voice cellular bandwidth, RDS exploits an unused portion of commercial FM broadcast waves. Unlike cellular and spread spectrum, this is a one-way broadcast technology, with data targeted to subscribers to an information service, rather than to the individual user. RDS is used for broadcasting digital traffic reports from Traffic Management Centers (TMCs), hence the acronym RDS-TMC. A decoder in the vehicle can be programmed to filter out traffic messages that do not pertain to the current route, and to accept text data transmissions only in a given language. RDS also has a feature to override an audio station, and to wake up a radio from sleep mode. It is therefore well suited to disaster warning. Another popular RDS application is real-time differential GPS correction, where 100 MHz RDS competes against 300 kHz radio beacons.

RDS is not yet widely established in the U.S. There are message standards issues to be resolved, and location referencing in particular is an area in which standards are still under development (e.g., Noronha *et al.*, 1999). Currently there are two RDS protocols vying for adoption in the U.S.: Sub-Carrier Traffic Information Channel (STIC), developed specifically for use in Intelligent Transportation Systems; and Data Audio Radio Channel (DARC). STIC receivers are expensive and require more power, but enable slightly higher data rates, 18 kbps versus DARC's 16 kbps.

4.1.2 Implementation issues

It is clear from the discussion above that there are numerous considerations in selecting a technology for wireless communication: (a) is communication one-way or two-way; (b) is it directed to an individual, or a community of subscribers, or is it broadcast to the general public; (c) is encryption required, and are there security concerns; (d) are voice and data both required; (e) how portable is the transceiver; and what are its power requirements; (f) what data speeds does the technology support; and (g) what are the costs and benefits compared with local data storage (as opposed to real-time data transmission), wireline, and other wireless technologies.

Consider a police fleet of say 100 vehicles, communicating with a base station in an Automated Vehicle Location (AVL) application. With some technologies, each mobile unit requires say 10 seconds to connect, transmit data, and sign off. If all 100 vehicles use the same communications channel, it would be 15 minutes between location reports for a given vehicle—this does not qualify as real-time tracking. On the other hand, technologies such as CDPD do not require sign-in for each transmission, and are clearly more appropriate to the application.

4.2 Field computing technologies

There are many reasons for wanting to access powerful computing facilities in the field. The field is where many geographic data are collected, by direct measurement, interview, observation, photography, or other suitable means. These data must be made available to their eventual users, either by processing in the field and transmitting interpreted information, or by transmitting raw data to some other site for processing. Data collection may require other supporting information, such as base maps or results of earlier observation to support change detection, and these data may be in digital form, allowing them to be downlinked and used in the field if suitable technology is available. Decisions of a geographic nature are often best made in the field, and in cases such as emergency management may require extensive and rapid processing of appropriate data.

Technologies suitable for field computing have advanced rapidly in recent years. The first truly portable computers appeared in the mid 1980s, following important developments in battery technology, lowered power consumption, and greater computational power and storage. Laptops are now commonly used in field settings, by scientists and decision makers who require access to information technology. More recently, improvements in wireless communications have made it possible to operate virtually anywhere, though at relatively low communications speeds, and reliability of communications remains an issue.

Field computing technologies now include many devices besides the laptop. To allow further size reductions, many systems have dispensed with the keyboard, replacing it with other interaction modalities such as the pen and sensitive screen, and making use of software for limited recognition of handwriting. Egenhofer (1997) has explored the potential for communication of geographic data by sketch, and voice recognition and speech synthesis are also promising technologies for human-computer interaction in the field. Screen size remains an issue for geographic data, though, given the need for visual communication and high visual resolution.

Clarke (1998) reviews recent advances in various types of wearable field devices. Entire computing systems have been constructed to be worn under clothing, and have become notorious for their use in increasing the odds of winning in various forms of gambling. It is possible, for example, to input data by pressing sensitive pads with the toes, and to receive output through miniature devices concealed in eyeglasses. Several gigabytes of storage can be concealed in clothing, along with powerful processing units. Of particular relevance to field GIS are devices that make it possible to see visual data displayed *heads up* in a unit that resembles a heavy pair of eyeglasses, and combines actual and virtual views. These devices are now used routinely in assembly plants, since they allow their users to see blueprints while working on assemblies. They could be used in field GIS to provide visual access to base maps, images, maps showing the landscape as it existed at some earlier time, or simulations of future landscapes. Systems such as these that *augment* reality are similar in some respects to *immersive* technologies, which replace reality with virtual renderings (e.g., Earnshaw *et al.*, 1995).

Goodchild (1998) argues that GIS should be seen as an interaction not only between human and computer, but between human, computer, and geographic reality (HCRI rather than HCI). A GIS database cannot be a perfect rendition of reality, since reality is infinitely complex but the database must always be finite, and thus some degree of approximation, generalization, or abstraction will always be necessary. Given this, effective GIS will always require interaction with the reality that the database imperfectly represents. Field GIS is an instance of HCRI, in which human and computer interact in the presence of the phenomenon, allowing the human to interact with reality while interacting with the computer. Field GIS allows direct access to ground truth, as well as to any digital representations of existing and prior conditions. As such, it promises to vastly improve the effectiveness of field research, and to open much more efficient channels of communication between field workers and the eventual users of their data.

4.3 Distributed computing

The networking functions provided by the Internet and its services support a wide range of modes of interaction between computers (for a review of GIS applications of the Internet see Plewe, 1997). But these are largely limited today to binary interactions between two computers. In a client-server environment, operations are divided between the client, which is commonly a machine of limited processing, storage, and software capabilities, and a server, which may have access to large resources of data, and much more powerful software and processing. Suppose, however, that a GIS user wishes to obtain two sets of data for the same area, and these are located on different servers. It is possible to download one set of data from one server, disconnect, and reconnect to the second server to download its data. It is not possible, however, to access both servers simultaneously, or to make use of services that avoid download of data. If a map is needed, it must be computed at the client from two downloaded data sets. This is in contrast to a map based on a single data set, which can be computed at the data set's server, avoiding the need to download data or to have mapping software at the client.

Thus while the Internet and WWW offer many powerful functions, they fail at this time to support many important but advanced capabilities that are part of the vision of truly distributed computing:

- simultaneous access to multiple servers from a single client;

- support for true distributed databases, such that the user sees a single database, but tables or even parts of tables are resident at different server sites;
- support for truly distributed software, such that the user sees a single software environment but modules remain resident at different server sites.

In all three of these areas there are active research projects and prototypes, but no comprehensive solution yet exists.

The WWW Mapping Special Interest Group of the Open GIS Consortium (*http://www.opengis.org/wwwmap/index.htm*) seeks to develop prototype demonstrations of solutions to the first bullet above, for GIS applications. Its aim is to make it possible for a client to access layers of data resident on multiple servers, while allowing the user to work with them as if they were all resident locally. For example, a user should be able to display and analyze a layer of soils and a layer of roads as if they were in the user's own GIS, without downloading them from their respective servers. Many problems will have to be overcome, including specification and adoption of standards, before this vision becomes a practical reality, despite the fact that it is already reality in the case of data on a single server.

In the case of the second bullet, there is much interest in the GIS community in making it possible for different agencies to own and maintain different parts of a single, unified database. For example, responsibility for the fields in a streets database might be divided between a mapping agency, with responsibility for defining and maintaining the basic geometric and topological framework of streets; and a transportation agency responsible for adding dynamic information on levels of congestion and construction projects. Since the latter will be represented as attributes in tables defined by the former, it is clearly preferable that parts of the database exist on different servers. But there are numerous problems in developing the technology of distributed databases (Burleson, 1994). Maintenance of integrity is particularly difficult, if different users are to be allowed to access and modify parts of the same tables simultaneously. An open question is the extent to which distributed GIS databases will be based on generic distributed database technology, or supported using conventional technology through administrative arrangements and protocols.

In the case of the third bullet, much effort is currently under way in the GIS software industry to exploit modern capabilities for distributed components. The underlying standards and protocols are being developed by the Open GIS Consortium (*http://www.opengis.org*), a group of companies, agencies, and universities dedicated to greater interoperability in GIS. Already it is possible to integrate GIS and other software environments, by allowing:

- a user in a GIS environment to access directly the services of some other environment, provided both are resident locally;
- the reverse, for a user in a non-GIS environment, such as a spreadsheet, to access GIS functions without leaving the current environment;
- a user to combine the services of a remote host by integrating them with the services of a local client.

Much more research needs to be done, however (see, for example, Goodchild *et al.*, 1999) before true distributed processing will be possible in GIS.

The benefits of distributed processing are clear, however. GIS applications are often complex, in areas such as environmental modeling or vehicle routing. A developer of capabilities in such areas may find it much more attractive to make software available in modules than as part of monolithic software, and may even insist on retaining complete control, requiring users to send data to the owner's host. The possibility of "sending the data, not the software" raises numerous issues of ownership, institutional arrangements, and protection of intellectual property.

4.4 Distributed production

Traditionally, production of high-quality, reliable geographic data has been the almost exclusive domain of central governments. Every country has invested in a national mapping agency, sometimes under military command and sometimes in the civilian sector (Rhind, 1997). Today, several trends suggest that we are moving away from that model into a more complex, distributed set of arrangements with an emphasis on local production and use. The simple model of a flow outwards from the center to a dispersed user community is being replaced by a much more complex model involving various forms of data sharing (Onsrud and Rushton, 1995). This is occurring for several reasons:

- The economics of geographic data production have changed dramatically. New sensors, new instruments, and new software make it possible for geographic data production to take on the characteristics of a cottage industry, with farmers, for example, able to afford sophisticated technology for mapping and monitoring their own fields and crops (Wilson, 1999). The fixed costs of mapping in particular have fallen as the set of applications has expanded, leaving little economic incentive or advantage in centralized production.

- Contemporary political trends are against large central government, and for privatization of what previously were largely government functions. In the UK, for example, the Ordnance Survey has been forced to operate on commercial lines, and similar patterns can be observed in most countries. In the US, commercial production of data is now considered profitable even though much government data is available free, because the private sector is seen as more responsive to the needs of users.

- Modern technology allows much more flexible approaches to geographic data production. The "wall-to-wall" coverage mandated for government mapping agencies is no longer appropriate when most users access data through technologies that can accommodate data of varying accuracy and resolution. Future coverage is likely to be on a *patchwork* basis, rather than the uniform coverage that is the expressed objective of national mapping agencies.

- If much geographic information is produced locally, and much of it is also consumed locally, there is little reason to integrate data at a broader scale. The interests of national governments are more likely to be satisfied by coarse-scale generalizations, leaving detailed data to be produced, distributed, and consumed locally.

5. AN ECONOMIC FRAMEWORK

The previous section on locational history has already hinted at how the location of computing might be placed within an economic framework. This section expands on that theme, and presents a basis for research on the costs, benefits, and economic value of distributed and mobile computing.

From the communications perspective established earlier, computing is seen as a process of information transfer from one person, group, or agency to another. Locations are associated with both sender and receiver, since the human intelligence associated with both must be located somewhere. Locations are also associated with storage, processing, and all of the other stages identified in Section 2.2.

5.1 Transport costs

Various costs are associated with the need to overcome the separation between the locations of sender and receiver, and with other separations such as that between the location of geographic ground truth and that of the sender. In classical location theory these costs are tangible, being determined by the costs of moving people or goods from place to place. In the case of information, however, there are both tangible costs, related to renting communication links and the fixed costs of establishing them, and intangible costs related to delay and unreliability.

Consider the costs associated with sending information between locations *i* and *j*. Several distinct conditions are relevant in today's environment:

- There exists a fixed communications link of high capacity, whose fixed costs have been absorbed by some agency and for which rental is free. Tangible costs are zero, as are intangible costs. This is the normal case when the Internet is accessible and both sender and receiver are in the U.S.
- There exists a fixed high-capacity communications link, but it has significant latency; while tangible costs are zero, there are substantial intangible costs to using the link.
- No fixed high-capacity communications link exists. Communications must rely on low-capacity links such as telephone lines, which may incur rental charges; or on wireless links with rental charges and substantial unreliability. There are both tangible and intangible costs associated with communication of information.

When costs exist, they begin to influence locational decisions, and either the sender or receiver may choose to relocate, other locations needed for the communication of information may be affected, or communication links may be chosen, to minimize transport costs.

5.2 Facility costs

Various facilities are needed for communication to occur between sender and receiver. Computer processing may be needed to modify or transform data, or to provide analysis or modeling. Processing will also be needed at the locations of servers, and at other nodes in the communications network. As in the previous section, many of these processing resources are available free because they are part of the Internet, or provided through other arrangements to the sender or receiver. Others must be covered by the users through rental or fixed cost charges. Locational decisions may be involved if there is the possibility of selection among alternative processing locations, or alternative computing resources. Costs may be tangible, when processing power must be rented or purchased, but they may also be intangible, as when choices exist between the computing resources of different agencies, and issues such as security, privacy, or intellectual property are important.

5.3 Human intelligence as a locational factor

Finally, the locational decision will be influenced by the need to consider the locations of human actors in the system. If GIS is a communication problem, as suggested here, then the locations of sender and receiver are both important. Other human actors may also be involved, as interpreters, custodians of ancillary data, or developers of software. In Section 3 it was argued that in earlier phases of the history of computing it was common for human intelligence to move in response to the availability of computing resources. The decision to compute in the office rather than the field may also be an instance of moving human intelligence to the location of computing resources, rather than the reverse. As computing becomes cheaper and the costs of communication lower, it will be increasingly common to move computing to human intelligence, rather than the opposite; and arguably that process is already almost complete with respect to locations provided with power supplies and Internet connections. Changing economics of computing and emerging field technologies will have substantial influence on the locational decisions made by GIS users.

Locational decisions such as those discussed in this chapter will clearly impact where computing is done, and where users choose to locate, both for scientific and for decision-making applications of GIS. An important area of research is emerging, in the development of models and frameworks that allow such decisions and options to be explored in a rigorous, well-informed framework that can make use of our increasing understanding of the costs of computing and communications.

6. LIBRARIES AND ARCHIVES

Like any other institution, libraries are feeling the influence of the shift to digital communication, and the concept of a *digital library* has received much attention in recent years. In principle a digital library is usable entirely through the Internet, and thus achieves universal access. Its catalog is digital, and all of its books and other information resources are also in digital form. Novel forms of searching are possible; information can be sorted and delivered independently of the media on which it was originally stored; and information can be processed and analyzed either by the library or by its users. In short, the digital library holds enormous promise for increasing humanity’s access to its information resources. This section discusses several aspects of digital libraries of relevance to geographic information, in the context of distributed and mobile computing.

6.1 The search problem

Much of the information used in GIS analysis is *framework* data—largely static data produced and disseminated for general purposes by government agencies and corporations (MSC, 1995). Framework data sets include digital imagery and digital topographic maps; more specifically, they include representations of terrain, hydrography, major cultural features, and other information used to provide a geographic frame of reference.

Such data sets are ideally suited to dissemination through libraries and similar institutions, where they can be stored, maintained, and lent in a reliable manner. Many libraries maintain extensive *collections* of maps and their digital equivalents, and increasingly the services of libraries are available online via the Internet. Many data archives, data centers, and clearinghouses have emerged in recent years, largely following the library model, for the purpose of serving specialized communities of users, and the National Geospatial Data Clearinghouse (NGDC; *www.fgdc.gov*) is an excellent example.

Today the WWW includes some 10^7 servers, any one of which might contain and be willing to provide a collection of geographic data sets. Any given data set may be available from several sites. In effect, the WWW serves as a distributed library of data resources, and is expected to grow massively in this role in the coming years. At the same time the WWW presents a growing problem for its users: how to find relevant information among the myriad sites of the network? Users must work through a two-stage process of search, first finding a site within the set of Internet sites, and then finding data within the site.

The second step in this process is often relatively straightforward, because owners of sites are able to set up effective cataloging and searching mechanisms within the sites under their control. But the first step is much more problematic, since few standards and protocols exist for the WWW as a whole. How, for example, is a GIS user in need of a digital base map for a small area of Utah to find one among the vast array of servers and services?

At this point in time, and in the absence of effective search mechanisms for geographic data, the user is forced to rely on one or another of the following heuristics:

- Assume that all servers possess the needed data. This is a reasonable strategy for major libraries, since every major library attempts to include all of the most important books in its collection, but is clearly absurd for the WWW.
- Use one of the WWW search services. The current generation of services provides powerful means for searching across the WWW, but relies almost entirely on key words. Since it is difficult to express location in words, and since these services are not effective at detecting the existence of geographic data, this strategy is generally ineffective.
- Go to a site that serves as a clearinghouse. The NGDC provides a mechanism for searching across a large number of collections, using common protocols for data description. If all geographic data could be registered with it, this would be an effective method of search. But it is very unlikely that all owners of geographic data

will be willing to invest the necessary time and effort; and it is likely that more than one clearinghouse will emerge, and that protocols will vary among them.

- Rely on personal knowledge. In practice most searches for data rely on some form of personal knowledge, personal network, or other form of informal communication.

Effective mechanisms for searching distributed archives of geographic data would be very useful, in helping users with little personal knowledge to exploit the massive resources of the WWW. The National Research Council (MSC, 1999) has defined the concept of a *distributed geolibrary*, a network of data resources that is searchable for information about a given geographic location, or *place-based search*. Search based on location is already possible within certain sites, including the Alexandria Digital Library (*alexandria.ucsb.edu*), but not across the WWW itself. Development of such methods should be a major research priority of the geographic information research community.

6.2 Libraries as central places

Although Section 6.1 has laid out the principles underlying search in the era of the WWW, the distribution of geographic data sets among WWW sites reflects far more the legacy of previous technologies and approaches. For example, the Alexandria Digital Library has as one of its objectives the provision of access to the rich holdings of the University of California, Santa Barbara's Map and Imagery Laboratory, a large collection of paper and digital maps and imagery. This collection has been built up over the years to serve a population of users largely confined to the UCSB campus and its immediate surroundings. It reflects the fact that it is the only accessible store of geographic data for that user community, and thus the collection has attempted to prioritize acquisitions on that basis. But this strategy is the exact opposite of what is needed in the WWW era, where distance is relatively unimportant, and where the user's problem is to find the collection most likely to contain a given data set. In the earlier era, it made sense for every major collection to try to include all important items; today, it makes more sense for collections to specialize, so that an item is present in a very small number of collections, and so that there are clear guidelines regarding where to search.

This transition is readily understood within the context of central place theory, as discussed earlier in Section 3. The WWW has increased the range of archival service, and has reduced its threshold. Rather than every site offering the same good, it is now possible for each site to specialize, and specialization also helps the search process by providing ready definitions of, and limits to, the contents of each collection. In the language of central place theory (Berry, 1967), a dense pattern of offerings of a low-order good is being replaced by a sparse pattern of offerings of a large number of high-order goods.

The literature of central place theory provides few clues as to how this transition will occur. The *adoptive hypothesis* (Bell, 1970) argues that it will occur by a form of Darwinian selection, in which those sites that are most progressive will ultimately force out the others. The speed with which this process operates depends very much on the far-sightedness of individuals, because the economic and politicla forces underlying it are comparatively weak, allowing inefficient and inappropriate sites to survive well into the new era.

7. CONCLUSION

7.1 Research issues

Many research issues have been identified in the previous discussion. This section summarizes them, as a series of priority areas for research.

- Examine the status and compatibility of standards across the full domain of distributed computing architectures and geographic information at national and international levels; identify important gaps and duplications; examine the adaptability of standards to rapid technological change; evaluate the degree to which geographic information standards and architectures are compliant with and embedded in such emerging frameworks as Java, CORBA, and COM/OLE; recommend appropriate actions.
- Build models of GIS activities as collections of special services in distributed object environments to support their integration into much broader modeling frameworks. This will help promote the longer-term objective of making GIS services readily accessible within the general computing environments of the future.
- Develop an economic model of the distributed processing of geographic information; include various assumptions about the distribution of costs, and use the economic model to develop a model of distributed GIS computing.
- Modify commonly used teaching materials in GIS to incorporate new material about distributed computing architectures.
- Develop methods for the efficient use of bandwidth in transmitting large volumes of geographic data, including progressive transmission and compression; investigate the current status of such methods for raster data; research the use of parallel methods for vector data.
- Develop improved models (i.e., structure and format) of geographic metadata to facilitate sharing of GIS data, to increase search and browse capabilities, and to allow users to evaluate appropriateness of use or allow compilers to judge fitness of data for inclusion in GIS.
- Develop theory that addresses the optimal location of computing activity, building on existing theories of the location of economic and other activities and on the economic model described earlier.
- Study the nature of human–computer interaction in the field, and the effects of different interaction modalities, including speech, sketch, and gesture.
- Develop new adaptive methods of field sampling that are directed by real-time analysis in the field.
- Study the role of contextual information gathered in the field by new technologies and used to inform subsequent analysis of primary data.
- Examine the implications of IDU computing with respect to intellectual property

rights to geographic information and within the context of broader developments in this area.

- Examine the social implications of IDU computing and its impacts on existing institutions and institutional arrangements.
- Conduct case studies examining the application of IDU computing in GISs, including horizontal applications (with data distributed across different locations), and vertical applications (with data distributed at different levels in the administrative hierarchy).
- Monitor the progress of research addressing the technical problems that IDU computing architectures pose with regard to geographic information, including maintenance of data integrity, fusion and integration of data, and automated generalization.
- Examine the various architectures for distributed computing and their implications for GIS. This will include consideration of distributed database design, client-server processing, database replication and versioning, and efficient data caching.

7.2 Anticipated benefits from research

- *Access*. By decentralizing control, distributed computing offers the potential for significant increases in the accessibility of information technology, and associated benefits. There have been many examples in recent years of the power of the Internet, wireless communication, and other information technologies to bypass the control of central governments, linking citizens in one country with those with common interests around the world. Wireless communication avoids the restrictions central governments impose through control over the installation of copper and fiber; digital communication avoids many of the restrictions imposed over the use of mail.
- *Cost reductions*. Modern software architectures, with their emphasis on modularity and interoperability, work to reduce the cost of GIS by increased competition and sharing, and by making modules more affordable than monolithic packages.
- *Improved decision-making*. Current technologies virtually require decisions that rely on computing support to be made at the desktop, where powerful hardware and connectivity are concentrated. IDU computing offers the prospect of computing anywhere, resulting in more timely and more accurate data and decisions.
- *Distributed custodianship*. The National Spatial Data Infrastructure (NSDI) calls for a system of partnerships to produce a future national framework for data as a patchwork quilt collected at different scales and produced and maintained by different governments and agencies. NSDI will require novel arrangements for framework management, area integration, and data distribution. Research on distributed and mobile computing will examine the basic feasibility and likely effects of such distributed custodianship in the context of distributed computing architectures, and will determine the institutional structures that must evolve to support such custodianship.
- *Data integration*. This research will help to integrate geographic information into the mainstream of future information technologies.

- *Missed opportunities*. By anticipating the impact that rapidly advancing technology will have on GIS, this research will allow the GIS community to take better advantage of the opportunities that the technology offers, in timely fashion.

Acknowledgment

The authors thank Dawn Wright of Oregon State University for her help in synthesizing the discussions that occurred on this topic in the UCGIS Virtual Seminar in Fall 1998.

References

Bell, T.L. (1970) A test of the adoptive hypothesis of spatial-economic pattern development: the case of the retail firm. *Proceedings of the Association of American Geographers* 2: 8–12.

Berry, B.J.L. (1967) *Geography of Market Centers and Retail Distribution*. Englewood Cliffs, NJ: Prentice-Hall.

Burleson, D.K. (1994) *Managing Distributed Databases: Building Bridges between Database Islands*. New York: Wiley.

Christaller, W. (1966) *Central Places in Southern Germany* (Translated by C.W. Baskin). Englewood Cliffs, NJ: Prentice-Hall.

Clarke, K.C. (1998) Visualising different geofutures. In P.A. Longley, S.M. Brooks, R. McDonnell, and W. Macmillan, editors, *Geocomputation: A Primer*. London: Wiley, pp. 119–138.

Coppock, J.T., and D.W. Rhind (1991) The history of GIS. In D.J. Maguire, M.F. Goodchild, and D.W. Rhind, editors, *Geographical Information Systems: Principles and Applications*. Harlow, UK: Longman Scientific and Technical, Vol. 1, pp. 21–43.

Current, J.R. (1981) Multiobjective design of transportation networks. Unpublished PhD Dissertation, Johns Hopkins University.

Earnshaw, R.A., J.A. Vince, and H. Jones, editors (1995) *Virtual Reality Applications*. San Diego: Academic Press.

Egenhofer, M.J. (1997) Query processing in spatial-query-by-sketch. *Journal of Visual Languages and Computing* 8(4): 403–424.

Foresman, T.W., editor (1998) *The History of Geographic Information Systems: Perspectives from the Pioneers*. Upper Saddle River, NJ: Prentice Hall PTR.

Goodchild, M.F. (1997) Towards a geography of geographic information in a digital world. *Computers, Environment and Urban Systems* 21(6): 377–391.

Goodchild, M.F. (1998) Rediscovering the world through GIS: Prospects for a second age of geographical discovery. *Proceedings, GISPlaNET 98, Lisbon*. CD

Goodchild, M.F., M.J. Egenhofer, R. Fegeas, and C.A. Kottman, editors (1999) *Interoperating Geographic Information Systems*. Norwell, MA: Kluwer Academic Publishers.

Hayes, B. (1997) The infrastructure of the information infrastructure. *American Scientist* 85 (May–June): 214–218.

Lösch, A. (1954) *The Economics of Location* (Translated by W.H. Woglom). New Haven, CT: Yale University Press.

Mapping Science Committee, National Research Council (1995) *A Data Foundation for the National Spatial Data Infrastructure*. Washington, DC: National Academies Press.

Mapping Science Committee, National Research Council (1999) *Distributed Geolibraries: Spatial Data Resources*. Washington, DC: National Academies Press.

Murnion, S., and R.G. Healey (1998) Modeling distance decay effects in Web server information flows. *Geographical Analysis* 30(4): 285–303.

Noronha, V.T., M.F. Goodchild, R.L. Church, and P. Fohl (1999) Location expression standards for ITS applications: testing the Cross Streets Profile. *Annals of Regional Science,* in press.

Onsrud, H.J., and G. Rushton, editors (1995) *Sharing Geographic Information*. New Brunswick, NJ: Center for Urban Policy Research, Rutgers University.

Plewe, B. (1997) *GIS Online: Informatio, Retrieval, Mapping, and the Internet*. Santa Fe: OnWord Press.

Rhind, D.W., editor (1997) *Framework for the World*. New York: Wiley.

Wilkes, M.V. (1957) *Automatic Digital Computers*. London: Methuen.

Wilson, J.P. (1999) Local, national, and global applications of GIS in agriculture. In P.A. Longley, M.F. Goodchild, D.J. Maguire, and D.W. Rhind, editors, *Geographical Information Systems: Principles, Techniques, Management and Applications,* Second Edition. New York: Wiley, Vol. 2, pp. 981–998.