# 4

# Towards a Location Theory of Distributed Computing and E-Commerce

**MICHAEL F. GOODCHILD**

*University of California, Santa Barbara, CA, USA*

## INTRODUCTION

Humanity has always struggled with the impediments created by distance, through its ability to reduce or truncate communication, add cost and delays to the movement of goods, and limit the spaces of direct experience; and such impediments play an essential role in our ability to understand the uneven distribution of phenomena on the Earth's surface, and the various strategies that humans adopt to decide where to locate, whether it be for commerce, residence, employment, or recreation.

Each new innovation in transportation technology has helped to reduce the impediments associated with distance, and many such innovations have been hailed as having profound effects, whether they be in helping to reduce human misunderstandings through more frequent contact, or in enhancing trade and the assumed benefits of an integrated global economy. Terms such as 'time–space convergence' (Brunn and Leinbach 1991; Forer 1974; Janelle 1968) and 'the global village' date from transportation and communication innovations that occurred several decades ago. But of all of these stages the most recent is surely one of the most significant – the phenomenon of networked digital communication that has made it possible to communicate at close to the speed of light and at very low cost across a network that increasingly reaches every part of the planet, based on the Transmission Control Protocol/Internet Protocol (TCP/IP). Many prior innovations in transportation technology pale in comparison to the impact of the Internet. Today more than ever 'there is no more *there*, everywhere is *here*', to cite a recent advertising campaign of a major telecommunications company.

An obvious reaction to such notions is to believe that location is now irrelevant, at least as far as communication of information is concerned. Virtual and immersive technologies already allow environments to be represented in astonishing detail at

remote sites, in effect transporting the individual across space without any of the impediments associated with moving matter. Yet we know that these environments are merely representations, and that there is no prospect of perfect representation – we can never build immersive environments that completely replace the human experience of being there, because representations can never fully replace the senses. Moreover, even digital information must exist somewhere, at some well-defined location on the network of hardware, software, and communication links. Pronouncements to the contrary, then, location in the digital world will be of great significance, and the impacts of digital technology on human society will be of profound importance.

Location theorists have long worked to understand and predict the locations of human activities. Many such efforts have been firmly grounded in economics, and have given primacy to transport cost as an explanatory factor. If the economic costs of transporting information have been reduced to near zero, it follows that predicting the locations of information-based activities will be increasingly difficult, since theories of location will have to give greater importance to factors that are less tangible, and more difficult to formalize and operationalize.

This chapter attempts to lay the groundwork for a theory of the location of computing, and more generally of activities that are associated with the information economy and e-commerce, and depend heavily on computing. The term *computing* is used throughout as shorthand for the acquisition, manipulation, storage, and analysis of information in digital form. The chapter is structured as follows. The next section briefly reviews classical location theory, which derives expectations about the locations adopted by the components of a system on the Earth's surface from statements about the behavior of actors controlling the components, and assumptions about relevant prior conditions. The subsequent section examines computing, and the degree to which it is influenced by location. This is followed by a discussion of the factors likely to influence the location of computing, given that the costs of overcoming distance are likely to be very small or nonexistent. The last major section of the chapter examines research libraries, as an instance of a system of physical facilities and associated services whose spatial structure will be profoundly influenced by the transition to a digital world, and as a case study for other types of facilities and services that are being similarly affected.

Any discussion of the location of computing must address the distinction between information that is and is not geographically referenced. We define geographically referenced information as having a defined geographic footprint, enclosing the area on the Earth's surface that is described by the information. Maps and Earth images clearly fit this definition, but so do certain books, reports, works of art, and other items that are associated in one way or another with some geographic area, and for which geographic location might form a basis of search. For example, footprints are essential to answer queries of the form 'what information is available about restaurants in Paris?' or 'who else wrote about the Bath of Jane Austen's time?'

## CLASSICAL LOCATION THEORY

In classical location theory the locations of activities on the Earth's surface are established by decision-makers, who are assumed to behave according to certain



rational principles. The central place theory of Christaller (1966) and Lösch (1954), for example, deals with the behaviors of consumers who occupy residential locations that are scattered over the surface, and entrepreneurs, who sell goods to consumers. Because entrepreneurs must invest in the high fixed costs of building stores, it is not possible to place stores at the locations of every consumer; instead, stores are established at a few central locations. In recent decades central place theory has been generalized to theories about the locations of all kinds of central services, in both the public and private sectors (Love *et al.* 1988; Mirchandani and Francis 1990). For example, location theory is now routinely used to locate such services as schools or fire stations, as well as retail stores. The classical location theory of Weber (1929) deals with the location of industrial activities, based on the costs of transporting inputs and outputs. A steel mill, for example, would locate at the least-cost point between its sources of supply (coal, iron ore, and limestone) and the markets for its products.

In classical central place theory the behavior of consumers and entrepreneurs is reduced to two simple premises: first, that consumers will travel to the central facility only if it is within some distance termed the range, and will otherwise opt not to purchase the service offered; and second, that entrepreneurs require some minimum number of consumers, known as the threshold, to continue offering the service. Much more complex models of both types of behavior have been proposed (see, for example, Berry and Parr 1988).

Clearly the predictions of the theory will be impacted if the range changes, which it can be expected to do following major innovations in transportation that reduce travel cost or increase travel speed. A consumer who had to walk to the nearest village might well be tempted to travel many times further to a larger town if able to do so at low cost by bus, and in time this change in behavior could drive stores in the village out of business. Similarly, local stores could be driven out of business if a sufficient number of consumers chose to make their requests via the Internet, and to receive goods via express delivery services. The costs of sending requests via the Internet, and shipping goods via express delivery services are both largely independent of distance, so there are no longer any incentives for entrepreneurs to locate close to consumers. Predictions will also change if the threshold changes, as a result of changed expenditure per consumer, or changes in the economies of scale of store operation or the costs of store establishment. A *category killer*, a large store operating at low margin in a market such as sporting goods, might be able to drive a number of smaller stores out of business because of its stronger scale economies, and even stronger scale economies may be associated with e-commerce operations like Amazon.com. Similarly, the predictions of Weber's industrial location theory will change if innovations in transportation result in different transport costs, or if technological innovations change the economies of scale of plant operation.

In addition to the effects of scale economies and transportation costs, locations will be influenced by prior conditions, for example concerning the availability and costs of land, or constraints on development imposed by planning agencies. Decisions will also be determined by variations in the spatial density of consumers, the presence of existing central facilities, competition between entrepreneurs, and a host of other factors. Historically, each new spatial organization has inherited the legacy of past spatial organizations, and many businesses have made their first forays into e-commerce from their existing locations, whether or not these are now optimal. The

classical theorists made simplifying assumptions about many of these additional factors, and were able to derive simple predictions about the patterns of activities that would be found under such ideal circumstances. After much research in the 1960s had failed to find even vestiges of such simple patterns (Berry and Parr 1988), and after demonstrations that simple patterns would fail to emerge if the assumptions were even slightly untrue (e.g. Goodchild 1972), the emphasis in location research moved sharply away from prediction to the more normative framework of site selection.

## LOCATION AND COMPUTING

### Computers as Central Facilities

Computing is defined here as the set of processes associated with digital technology: the input, storage, transformation, communication, and output of digital information – in short, the core processes of e-commerce. In the earliest days of computing, in the 1950s, there were virtually no capabilities for long-distance communication of digital information at high speed, and all computing activities were therefore necessarily co-located. Input was created on paper tape or punched cards, fed into the computing system, stored there on magnetic tape until processed, and the results were output using mechanical printers. Large volumes of digital information had to be sent by mail on magnetic tape. The fixed costs of establishing a computing facility were extremely high, and as a result computing services could be offered only from a few locations where high levels of demand existed within short distances. Thus early computing resembled the classical model of a central facility providing a service to a dispersed population. The earliest computers were installed in universities, government laboratories, and large corporations where sufficient demand could be found to meet the facility's threshold, within the service's range. A major university computing center of this period might have served several thousand users with a single system. Users traveled to the facility to obtain its services, though in some instances input and output were sent by mail or express delivery. The solution of making the facility mobile, which is sometimes used in order to bring a larger number of consumers within range (as in mobile libraries or periodic markets, Bromley 1980) was not available because of the extreme weight and power requirements.

By the late 1960s the computing industry had adopted the communication technology of the telegraph industry, which allowed coded information to be transmitted over standard telephone wires at speeds of several hundred characters per second, using input and output devices similar to typewriters. Users could now process data that had been previously stored on a remote computer, and obtain simple results immediately, and more detailed results when paper output arrived by mail or express delivery. But long distance telephone charges were still a significant impediment to computing from truly remote sites, and users remained closely clustered around the computer that served them.

By the early 1980s the fixed costs of computers had begun to fall rapidly, first with the introduction of the 'superminicomputer', typified by the Digital VAX 11/750, and later with the personal computer, typified by the IBM PC and the Macintosh. The



threshold demand for a computer's services fell by orders of magnitude, until it became possible to think of a *personal* computer, a system with a threshold of one user. Computing had now reached the equivalent of a retail system that provides a store for every consumer, instead of a few central stores, obviating any need for consumers to travel to obtain service. Today, approximately 40 percent of US households have at least one computer. The remaining 60 percent have decided either that the benefits of owning a computer are less than the costs, or that the costs of traveling to and using a central service, in a library or cybercafé or school, are less than the costs of owning.

### Computers as Communication Media

The previous analysis was appropriate for the paradigm of computing that prevailed until the early 1990s. In this view, computing is a service that meets certain human needs: for example, the need to write letters and reports or to perform tax calculations. To a scientist, this view implies the ability to analyze, using statistical methods, or to model and predict. In all such cases the computer is performing much as a servant, responding to the user's commands and inputs with various forms of output, and exploiting its ability to perform arithmetic and logical operations much faster than a human. The computer is controlled by its users, whether these be a large group of researchers at a university, or a single user of a home PC, and everything stored in the computer has been placed there by its users.

In the late 1960s a small group of researchers began to explore the notion of connecting computers electronically, so that a user of one machine could access the services of another machine. If this could be done, it would allow software, data, and the results of analysis to be shared across machines. The ARPANet project, funded by the military Advanced Research Projects Agency, devised the necessary protocols for high-speed communication, which by the 1980s had been implemented over an extensive and rapidly growing network. The designers made the decision to adopt protocols that were largely independent of the specific communication technology, making it easy to take advantage of advances such as fiber-optic links that vastly increased the amount of information that could be communicated in a given period of time. By the early 1990s the network had evolved into the Internet, and communication speeds were reaching above a million bits per second.

By linking computers, it became possible to develop *client–server* computing, in which functions are shared between two computers. The user interacts directly with the client computer, entering commands and observing displayed output, while the much more powerful server computer might contain the database and the software used to access and process its contents. The network allows a client to switch quickly from one server to another, and supports the sending of information between any pair of computers. Client–server computing exploits the strong scale economies of computing technology, which make it cheaper to serve many users from a single machine, but to reserve certain operations associated with user interaction for local processing at the client machine because of the network's bandwidth limitations.

In this new environment the resources available to the user are those of the entire network, not merely the user's own client. The network has become in effect one massive computing resource, co-located with the user because of its high degree of

connectivity and minimal transport costs. Moreover, the network has promoted entirely new applications that have little to do with data processing and much more to do with communication between users. Services such as electronic mail, chatrooms, push technologies, and electronic commerce had no equivalent in the prior paradigm, which provided only for communication between user and machine.

The measures of success of the old paradigm focused on quantity and quality of service: how much data could be processed, at what cost, and in what period of time? The new computing paradigm implies quite different measures, however. At the technical level it is important to know:

- the costs of communicating information;
- the delays or *latency* associated with communication;
- the reliability of the communication channel, in terms of the proportion of information lost, or the proportion of down time;
- the speed of the channel, or the amount of information communicated in a given time, in bits per second (bps).

But other more conceptual aspects may be equally or more important:

- whether the channel acts as a filter, bypassing only certain types of information;
- whether sender and receiver share the same systems of understanding (definitions of terms, language, coding systems);
- whether software systems are interoperable—capable of accepting and processing information from each other.

### Computers as Augmentations of the Senses

Concepts of virtual reality and immersive environments are based on the notion that communication is an alternative to direct sensory perception, and the basis on which we learn about parts of the world that are beyond our immediate surroundings, either in space or in time. Thus the Virtual Field Course project (Dykes *et al.* 1999; Raper *et al.* 1999) sees technology as a substitute for the field experience that is capable not only of avoiding the high costs of travel, but also of representing information about distant environments in ways that help students to learn.

The concept of *augmented* reality is similarly based in the relationship between computer-based communication and sensory perception, but differs in that its purpose is to add to sensory perception. For example, a traveler equipped with a laptop might query a database to find information about his or her environment that is not directly available through sight, hearing, touch, or smell, such as the telephone numbers of businesses on the street that the traveler is driving along, or the availability of rooms in an adjacent hotel. This concept of augmenting the senses through technology has great commercial potential, as it allows the proximity of enterprises to be brought to the customer's attention even though they are not visible. In a commercial world in which all of the corner locations are already taken, such sensory augmentation can make other locations more attractive and competitive. Spohrer (1999) describes the World-Board project as exploring the use of devices that know where they are, and can present information accordingly. Reality can be augmented by devices that appear in the corner of the visual field—technology that is now commonly used on factory floors to allow workers to see the blueprints of structures they are assembling. Geographers might one day use such devices to display maps in the field, or data collected in prior visits. Reality can also be augmented by sound (Loomis *et al.* 1998).



### The Four Locations of Computing

The previous sections allow us to identify four important types of locations in any computing task, contrasting with the two (consumer and entrepreneur) that are important in classical location theory:

- the location of the user, and of the device with which the user is interacting (the *user* location);
- the locations of the servers with which the device is interacting through a communication network (the *server* locations);
- the location of the recipient of communication from the user (the *receiver* location);
- the location that is described by the information which the user is processing (the *subject* location).

The third type of location will be relevant only for information communicated between people, and the fourth only for geographically referenced information, by definition.

In addition to these four, we should perhaps add various forms of intermediate stores, or caches. The academic's bookshelf contains books that are both more likely to be of interest than the average book in the research library, and also accessible in a shorter time. A researcher working on a given area might choose to retrieve a comprehensive collection of imagery for the area into a personal store, knowing that it could be retrieved from the personal store when needed much more rapidly than from the Internet as a whole, because of the time required to search the Internet, and the slow retrieval speed from remote stores compared with local ones, and could perhaps be retrieved using more specialized search mechanisms.

The user location will affect computing in many ways. First, it determines the nature of communication with the Internet. While the Internet is now accessible globally, the 'last mile' to the user's location can often reduce connectivity substantially. While two-way connections as fast as hundreds of megabytes per second may be available at certain campus and corporate locations, the average household one mile away and connected via the telephone network and acoustic modem may be restricted to 28.8 Kbps, and certain household communication technologies such as a digital subscriber line (DSL) provide much higher bandwidth to the user than from the user. A wireless modem operating via the cellular telephone network can provide access anywhere within cellular coverage, but at even lower communication rates. Finally, various forms of satellite-based communication extend coverage to virtually anywhere on the planet, but at much higher cost. Connectivity has become a complex function of location that is very different from the simple relationships between location and access to facilities that are assumed by classical location theories. If $C(x)$ denotes the connectivity available at location $x$, then the surface $C$ has a very rugged appearance,

with sharp gradients, contrasting sharply with the smooth isodapanes of Weberian theory. Moreover, if $r$ denotes the amount a user is willing to pay for connectivity, per unit of time, then $C$ is clearly also a function of $r$. Thus we write the bandwidth available at a location $x$ at cost $r$ as $C(x, r)$.

Suppose a user is located at distance $d$ from a point of large bandwidth. In classical location theory the costs of overcoming this distance would be assumed to increase with $d$. But this seems not to be true of Internet connectivity, at least over certain important ranges of distance. The cost of a cellular connection is independent of distance within certain ranges, as is the cost of a connection through the conventional telephone network. On the other hand, very remote locations are likely to offer connections only at very high cost, so a slight negative correlation between connection cost and distance may be found at very large distances.

Second, user location affects computing through its impact on modes of interaction. At the desk the user expects to interact mostly through the keyboard, the screen, and the mouse pointing device. Laptops extend these modes to airline seats and other even less comfortable locations. But outdoors the standard laptop screen is too dim, battery power too limiting, and keyboards too prone to failure. Instead, interaction with devices outdoors is likely to be through pens and small hand-held devices, both of which restrict the flow of information between human and device. Again, ease of interaction varies over space in very complex ways, differing sharply between locations inside and outside the same building.

We assume that servers are connected to the network with large bandwidth. Murnion and Healey (1998) have shown that the bandwidth available between two locations on the network is only weakly dependent on the distance between them, and similarly that latency is only weakly correlated with distance. Because some countries exert control over Internet access, and access is generally less available in some countries, the bandwidth available between two servers tends to be unique to the countries in which they are located, that is, $B(s_1, s_2) = E[i(s_1), i(s_2)]$, where $i(s)$ is the country containing server $s$, $E$ is the bandwidth available between servers in two countries, and $s_1$ and $s_2$ are the two servers. This structure induces a slight negative correlation between distance and bandwidth, because two servers further apart are more likely to be in different countries. There may also be similar effects if the two servers are located on different subnets that are relatively poorly connected to each other (see Wheeler and O'Kelly, 1999, for an analysis of the effects of network topology).

Now consider communication between a user and a receiver. The bandwidth available between them will be determined by their connectivity with the network, and also by whether they are located in different countries. The apparent connectivity between locations $x_1$ and $x_2$, using servers $s_1$ and $s_2$, and between individuals willing to pay $r_1$ and $r_2$ on connectivity will be determined by

$$\min[C(x_1, r_1), C(x_2, r_2), B(s_1, s_2)].$$

Distance between $x_1$ and $x_2$ will be a very poor predictor of this function.

Finally, any location theory of computing must consider the locational influences of the information that is being processed, if that information is geographically



referenced. Goodchild (1997) has used the term *information of geographically determined interest* (IGDI) to refer to the property of geographically referenced information that interest in it is likely to vary, being highest within the footprint of the information. For example, a street map of New York City is clearly of greatest interest to a user in that city. The same effect may be reflected in variation in the spatial resolution of information: a user in New York might require a detailed street map, whereas users in Paris might be satisfied only with generalized information about New York. The concept of computers as augmentations of reality is clearly relevant here, with its emphasis on devices that access information about their immediate geographic environments.

### Towards a Locationally Enabled Internet

Although the Internet is embedded in geographic space, its original conception had much to do with overcoming the impediments of geographic separation, at least within the USA. Domain names and IP addresses are assigned largely without reference to geographic location, although national domains predominate outside the USA, and state and local governments often adopt names that code geographic location (e.g. state.nc.us). Thus there is no obvious mapping between IP address and geographic location, no way to send e-mail to all users within a geographic region, and no way to readily identify the closest IP addresses. Search engines make it much easier to find World Wide Web (WWW) sites that have similar content than sites with similar locations. In short, the Internet's organizational mechanisms are predominantly functional rather than spatial.

A locationally enabled Internet would have significant benefits. A mandate from the US Congress will require cell phones to be locationally enabled by the end of 2001, so that the locations of emergencies reported using 911 from cell phones can be identified easily. As communications technologies become more and more integrated, the same capability in palmtops, laptops, and other devices connected via the Internet will be particularly useful. The concept of computers as augmentations of reality also implies an ability on the part of the device to determine its location, in order to present appropriate information. An Internet startup company, Go2 (http://www.go2online.com), is promoting a system of coding geographic locations as hierarchically ordered pairs of numerical digits, similar in appearance to IP addresses, to support the spatial organization of Internet information and services. Finally, there are obvious marketing benefits to enabling the sending of e-mail to addresses within a given geographic area.

## FACTORS AFFECTING THE SPATIAL ORGANIZATION OF COMPUTING

As noted at the outset, although many of the impediments to interaction at a distance have been reduced or removed by the high connectivity of the Internet, nevertheless every bit of digital information must be located somewhere, and rational decisions must be made about where to compute. We will argue in this section that while economic factors have a relatively weak impact on the spatial organization of computing,

other factors have risen in importance. Thus a location theory of computing and e-commerce is more likely to be driven by noneconomic factors than was classical location theory. This section examines those factors in detail.

### Accessibility

Accessibility is a familiar concept in geographic research (Pirie 1979). Generally, a service is considered more accessible the closer it is, and a large grouping of services is considered more accessible than a small grouping. Thus the accessibility to a collection of services each of size $z_i$ and distance $d_i$ is often characterized by a summation:

$$V = \sum_i z_i/d_i$$

a measure termed *potential* (Warntz 1967).

As we have seen, however, distance is a poor indicator of accessibility to Internet services, and the comparative irrelevance of distance in ordering opportunities suggests that other approaches are needed. Consider a set of servers, $S$, and suppose that the needed service is known to be available from one of them, but that the user must search among them for the correct one. One interpretation of accessibility in this context is that it is measured by the inverse of the set size, $1/|S|$. If servers vary in magnitude, such that large servers are more likely to contain the service, then an appropriate measure is

$$V = \sum_i z_i$$

where $z_i$ is the size of the $i$th server. In other words, the effect of making distance irrelevant to accessibility is captured by making all distances equal, and by redefining measures of accessibility based on set size rather than distance.

### Geographic Footprints

Consider information related to a specific footprint. It has already been argued that interest in such information is highest within the footprint. In addition:

- an agency is more likely to be mandated to collect and disseminate such information the more closely its jurisdiction approximates the footprint;
- access to ground truth, for checking and maintaining information, is highest within the footprint.

Thus Goodchild (1997) concluded that geographically referenced information is most likely to be found on a server located in or near the footprint. Although these arguments are not as strong in the case of other types of information, it seems reasonable to assume that information of specialized interest will be more likely found on a server located where that specialized interest is densest, whether or not the



information is geographically referenced. For example, special collections in research libraries tend to co-locate with research interests.

The collection-level metadata (CLM) of a server is defined as the description of its information contents. For servers containing geographically referenced information, CLM includes a description of geographic coverage. For example, Figure 4.1 shows the CLM of the geographically referenced information cataloged by the Alexandria Digital Library (http://alexandria.ucsb.edu). There is a clear concentration of information around the location of the library in Santa Barbara, but also much information about many other areas of the Earth's surface.

### Ubiquity

In classical location theory the solutions to problems of location are almost always fixed: an entrepreneur seeks a permanent central location to serve a fixed residential population. The mobile service discussed earlier is an exception in that the entrepreneur adopts a roving location, and similar solutions are often suggested in locating ambulances, rescue vehicles, and other services.

Wireless services offer the potential for high mobility in computing, because they allow a user to remain connected while his or her location changes. The term *ubiquitous computing* is used to describe an environment in which it is possible to compute anywhere, and in which geographic location therefore imposes no constraints. Battery-powered devices, such as palmtops, laptops, and in-vehicle computing systems, offer enormous potential for such ubiquity. In the area of intelligent transportation systems (ITS), the ability to compute in a moving vehicle is essential for services that allow drivers to modify their travel plans based on real-time information on congestion. Similarly, there are many e-commerce applications such as search for hotels and restaurants where the ability to compute while moving is of substantial benefit.

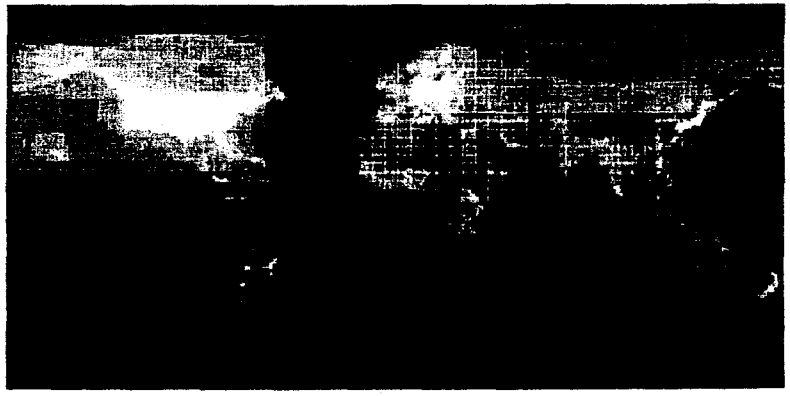

***Figure 4.1*** *An example of Collection-Level Metadata (CLM): the geographic distribution of data sets cataloged by the Alexandria Digital Library [http://alexandria.ucsb.edu]. Light shading indicates the highest density of data set footprints, and dark shading the lowest*

### Links to Other Activities

Many applications of IT involve direct interaction between computing and other mechanical or physical operations. For example, precision agriculture (Wilson 1999) requires the presence of IT on the farm, and even on moving farm equipment. Similarly, many resource industries, such as forestry, require co-location of IT and the resource itself. In such cases the user location is determined by external factors, rather than by issues of computing per se.

In e-commerce, IT is often intimately associated with the handling of physical goods, including express delivery operations and warehouses, and the economics of transportation may therefore have an important effect on the location of computing. Image of place is also an important factor, as reflected in the way the catalog company Lands' End markets its location in rural Wisconsin.

Computing is well known as an instance of an activity that is largely *footloose*, meaning that its functions can be carried out virtually anywhere. Telecommuting, the practice of working at home by making use of high connectivity, is in many ways a product of advances in information technology, though many 'cottage industries' have a long history of allowing their workers to produce at home. Telecommuting allows workers to select locations based on amenity or other factors that have nothing necessarily to do with information technology, by making use of near-universal connectivity.

### Face-to-face Communication

We have already noted that digital representations can never fully replicate the experience of the senses. Although user and receiver can often make use of very high connectivity, sufficient to support video for example, it seems that the Internet has done little to dampen the need for face-to-face meeting. Cities like New York, London, and Tokyo, which were built on the need for face-to-face business contact, continue to flourish, and tourism moves larger and larger numbers of people around the planet.

One of the major successes of the Internet has derived from its ability to transmit information independent of content, as universal packets that may contain anything from highly encrypted military signals, to e-mail text, to music. Thus it is possible to think of the Internet as a network whose ability to communicate information between people is limited only by bandwidth. But this raises an important question: are there fundamental limits to communication by electronic means that increased bandwidth will never solve? In a geographic context, it is common to note that certain sophisticated concepts such as 'sense of place' are much more difficult to express through digital channels than simple concepts such as distance, or location (Egenhofer *et al.* 1999). The issue appears intimately related to the more basic scientific issue of shared meaning: 'sense of place' is difficult to communicate, whether face-to-face or by digital means, because the sender and receiver do not enjoy shared understanding of the relevant terms.

### Distance and Relevance

The need to communicate is driven by many factors, including the mutual acquaintance of sender and receiver, the likelihood of shared interests, and the existence of a common language. The probability that any two people are acquainted clearly declines with the distance between them, despite the Internet's ability to create widely scattered communities. Similarly, the probability that any two people share interest is negatively correlated with distance because many interests and information media have geographic footprints. Finally, the strongly positive spatial autocorrelation of language induces a negative correlation between distance and the probability of shared language. Thus while the impediments to communication may now be only weakly correlated with distance, the propensity to communicate is likely to continue to exhibit strong negative correlation. Wheeler and O'Kelly (1998) demonstrate this effect clearly with their analysis of Internet traffic to and from Ohio State University.



### Attention

Media such as television acquire their value to commercial interests through their ability to capture attention remotely, augmenting the direct attention that is captured by outdoor advertising or direct mail. Similarly, the business case for Internet search engines, map services, and other popular applications is often based on the attention that such services can direct to advertising.

People are inevitably interested in their surroundings, and in issues of local importance. Moreover, people are more likely to avail themselves of services that require physical presence if these are close by. It follows that geographically enabled Internet services, such as directories to local businesses, are good ways to advertise other local services. The ability to target information based on the geographic location of the user and of the subject has substantial value in e-commerce, so much so that the term g-commerce is increasingly used to refer to Internet services that are geographically enabled.

### Summary

This section has identified various ways in which geographic location impinges on computing. Despite the weak correlations between distance and economic cost associated with Internet communication, then, there are solid grounds for believing that location will continue to be a factor in computing, driving complex decision-making processes.

This discussion has led to the following conclusions regarding the four locations introduced earlier:

- Subject locations are likely to be strongly correlated with server locations, since georeferenced information is most likely to be found on a server within its footprint.
- Interactions between user and receiver locations are likely to be strongly negatively correlated with distance, despite the lack of impediments to communication, because of factors driving the need for two people to communicate.
- The costs associated with communicating currently show strong geographic patterns, due to the current lack of ubiquitous access, and affecting both user and receiver locations. Cost surfaces show very strong local gradients, and also strong variation between countries.

## THE CASE OF THE RESEARCH LIBRARY

In this section we focus on a specific case – the location theory of the services provided by the research library. The arguments are based in large part on those presented by Goodchild (1997). The library is an immensely complex institution, and any discussion must inevitably simplify and even caricature it. This section focuses on the functions associated with the research library, as exemplified by the institutions that exist at all major research universities, as well as in some major cities and in national capitals. In addition it focuses on the functions such research libraries perform to provide information to their users, by acquiring and building collections, providing the mechanisms that allow users to search these collections for items of interest, allowing users to broaden their searches through browsing, and finally allowing users to retrieve information for use. Thus it ignores the functions of archiving, through which research libraries attempt to preserve knowledge that might otherwise be lost; as well as the provision of workspace, and direct involvement in the instructional programs of their host institutions.

Currently the number of such institutions is of order $10^4$. Although the collections of the largest research libraries might include order $10^7$ bound volumes, let us assume that the average collection is of order $10^6$. If the average volume contains order $10^5$ words of text, then a rough estimate of the number of words of text in the average research library might be $10^{11}$, or order $10^{12}$ bytes if we allow 10 bytes per word. Thus the text content of an average research library could be captured in 1 Tb of digital storage. Today, 1 Tb file stores are readily available, and occupy a space comparable to a small room. Of course we have ignored in this calculation all of the nontext content of bound volumes, including photographs and maps, and all of the annotation that can be so valuable to historical research.

Suppose order $10^3$ readers wish to access such a storage device using the Internet. If the average reader moves through text at 10 words per second, the rate at which the contents must be served is of order $10^5$ bytes per second, a rate well within the capacity of large contemporary servers. In summary, it is well within the capacity of current technology to provide the information dissemination services of a research library online. To do so it would be necessary to digitize the entire collection, but this again is well within the capacity of current technology, and almost all new text arriving in the research library is already in digital form. Of course major issues of intellectual property are involved in this scenario, and much debate about them is occurring at this time (see, for example, National Research Council 1999a), but such issues are beyond the scope of this chapter.

Since its widespread inception in approximately 1993, the WWW has rapidly become a major source of information, in many ways competing with the research library. Students searching for information for papers will often use the WWW exclusively, despite the relative lack of mechanisms for ensuring quality or completeness in information discovered in this way. Much of this section derives from a comparison between the WWW and the system of research libraries, so it would be helpful at the outset to establish a consistent terminology.

The WWW currently accesses some order $10^7$ servers or sites, distributed in most countries of the world. The geographic distribution of servers will be compared throughout the section to the geographic distribution of research libraries. The term 'collection' is used to describe both the set of volumes stored and cataloged by the research library, and the set of information objects accessible through a server. But while virtually all information in a research library is accessed in the form of bound volumes, there is no comparable control over the granularity of information accessible through a server.



Recent estimates put the total volume of information accessible through the WWW at order 1 Pb ($10^{15}$ bytes), though this figure may be unreliable by at least one order of magnitude. Each information object is accessed through a universal resource locator (URL), which is roughly equivalent in function to the call number assigned to a volume by the research library. But while call numbers carry useful information about subject, a URL carries almost no reliable information to guide the user of the WWW in searching for specific content. Instead, searching on the WWW is facilitated by a number of search engines, sites that use a combination of manual and automated means to build catalogs of WWW content, or relations linking words that are indicative of content to URLs in which those words figure prominently. Nevertheless, current estimates place only 10 percent of WWW content within the reach of search engines, leaving 90 percent essentially uncataloged.

Although the position of a volume on the shelves of a research library is established by subject, it is the catalog that provides the primary search mechanism. All materials acquired by the library are identified by catalog staff using author, title, and a controlled vocabulary of subjects. In the pre-digital era, these three keys were used to build card catalogs, physical systems that arrayed volumes in alphabetical order using card records. Thus the library user could easily find volumes knowing either author, title, or subject. Without the catalog it would be very difficult to make effective use of a collection of a million volumes or more.

These three sort keys are unidimensional (each key provides a basis for a simple sort, usually in alphabetical order), finite (the number of possible key values is limited), and discrete (the key values are countable), all requirements for a system based on the physical sorting of cards. In an early stage of library automation, now almost complete, the contents of the cards were transferred to digital databases, allowing much faster searching and easier update, as well as remote access. Users were now able to determine whether a given volume was in the library without having to visit. More importantly, however, the digital catalog enabled new methods of search. Full titles could be searched for key words, a much more powerful mechanism than relying on a title card catalog that sorted only by each title's first word. But most importantly, a digital database can be searched on many keys at once, and on dimensions that are continuous and infinite rather than discrete and finite.

Location and time are two obvious candidates for such expanded search capabilities. Location is multidimensional (two-dimensional for most geographic applications), and there are an infinite number of possible locations on the Earth's surface. Thus physical catalogs could support search by location only if that property could be captured by a limited number of place names in the controlled subject vocabulary. A digital catalog can easily support search by latitude and longitude (for example, find all information relevant to the area between 120 and 121 degrees West, 34 and 35 degrees North). The term *geolibrary* has been coined to describe a

library whose primary search key is geographic location (National Research Council 1999b).

Several instances of geolibraries have appeared in recent years to exploit the power of geographic location as a means of organizing information. The Alexandria Digital Library (ADL; http://alexandria.ucsb.edu) began as an effort to make the services of a major map and imagery collection available via the Internet, taking advantage of the new medium's ability to vastly increase their effective range. It rapidly became clear, however, that the ability to search by location could be applied to many other types of information that were not normally collected by map libraries, including books, photographs, and even pieces of music. ADL now contains over a million items, amounting to over 1 Tb of data, and includes a gazetteer with over 4 million entries to support users who must identify geographic locations by place name rather than by coordinates. Other instances of geolibraries include the US National Geospatial Data Clearinghouse (http://www.fgdc.gov) and Terraserver (http://www.terraserver.com).

### Libraries as Central Services

In central facilities location theory each facility provides a similar service. Competition between facilities leads to a hierarchical pattern in which facilities at the same level in the hierarchy offer the same goods. Similarly, major research libraries attempt to provide similar services to their users, modified to some extent by the special or unique collections that many libraries possess, and by the varying disciplinary strengths of their supporting institutions. We can represent this in a simple model in which every library strives to acquire the largest possible proportion of all significant published works – the largest and most successful research libraries, such as the Library of Congress, succeed in acquiring the largest proportion, while all other research libraries fall below this proportion to varying degrees.

The system of scholarly publication of which libraries are a very important part relies on the economies of scale that exist in the processes of printing and publication. The unit cost of printing a book falls dramatically with the number printed, and the marginal cost of printing one copy of a popular book is far less than its retail value. These economies of scale make it inevitable that the publishing industry will favor popular books over less popular ones – and because libraries are designed as physical repositories of books, it follows that the information found in libraries will be information that is of interest to the largest number of people, and the product of the largest possible print runs. By contrast, material that is of limited interest tends to be expensive to acquire, and is only found in libraries that have specialized in that respective subject area.

### Impacts of the Digital Transition

This system of research libraries has evolved within a particular technological environment, dominated by the printing press, metal shelves, card catalogs, etc. Arguably very little changed in this technological environment (a library built in the 17th century is still recognizable as such) until the advent of digital technology in the late 20th century. But over the past 20 years computerization has wrought very profound changes on this environment. Almost all information is now expressed in digital form at some point in its life, during the various stages of composition, editing, printing, dissemination, and distribution from the library. The transition to digital communication has brought enormous benefits, deriving from the economies of scale of a technology that is able to handle all kinds of information without respect to content, copy and transmit information almost instantaneously, and transform information through analysis at very low effort.



First, the digital transition has dramatically changed the parameters of the library as central facility. It has driven the effective range of the service much higher, since distance from the library is largely irrelevant if the institution is accessed and used via the Internet. Similarly the low costs of establishing and maintaining Internet servers have driven down the effective threshold, allowing services similar to those of a library to be provided at small scales by almost anyone.

Second, the digital transition has changed the economics of publishing, and made it feasible to publish information that is of comparatively low interest. An individual can create and publish a map or a paper using the digital technology of the WWW, even though it is clear that only a very small number of users are potentially interested. While the earlier environment favored publication of material of widespread interest, the much lower fixed costs of digital publication mean that publishing information of very limited interest can be economically viable.

Third, the digital transition has made new search mechanisms feasible. As noted earlier, it is now possible to think of searching library collections by location, by time, or by many new dimensions that were previously impossible. If a library is now able to support search by geographic location, it becomes a much more attractive repository for the kinds of information that are amenable to that particular search mechanism – specifically, information about places. Thus the development of new search mechanisms is likely to have a strong influence over the kinds of information we choose to store in libraries.

### Towards a New Library Geography

The combined influence of these factors is likely to be profound, affecting the spatial organization of libraries, their contents, their base of users, and the services they provide. Thus the new geography of libraries and of geographic information is likely to be fundamentally different from the old. The arguments presented in the previous sections lead to the following conclusions regarding the future spatial organization of research library services, as illustrated in Figure 4.2:

1. Increased range will remove the need for a large number of research libraries with similar content. In principle a single large facility, such as the Library of Congress, could satisfy the entire need, though issues of reliability and institutional legacy will likely maintain some redundancy.
2. Decreased thresholds and lower fixed publication costs will make it economically possible to organize library services for small and highly specialized collections.
3. New search mechanisms, and the ability of digital systems to handle information independently of content, will permit the serving of information not normally found in traditional research libraries, including geographically referenced information.

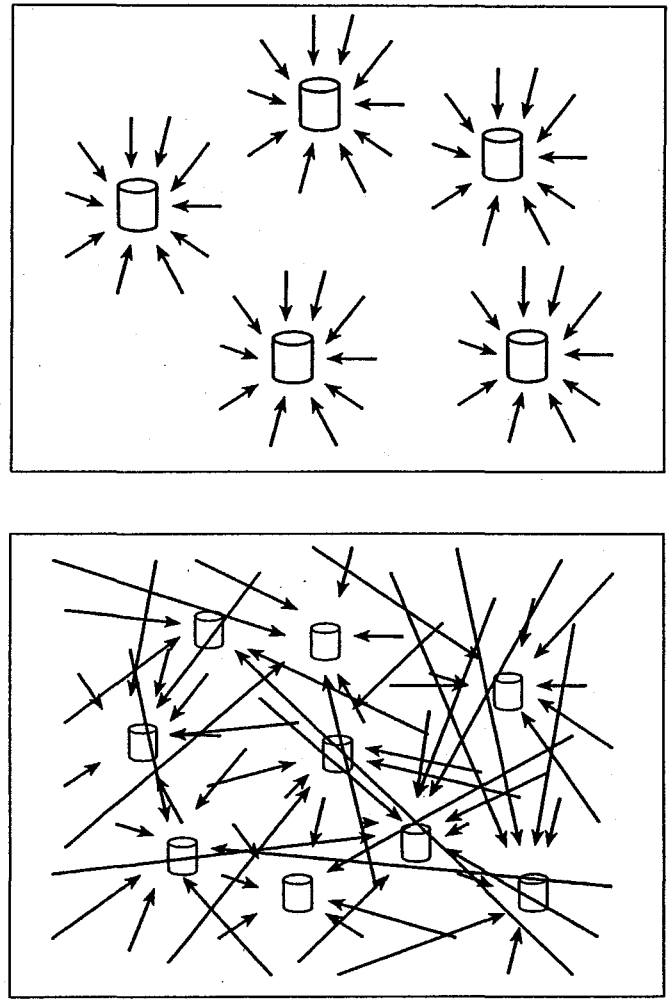


***Figure 4.2*** *Schematic rendering of the impacts of the digital transition on the distribution of research libraries. (a) The prior system, characterized by high threshold, short range, and uniform services, with substantial areas left unserved. (b) The emergent system of much lower thresholds, much larger ranges, universal access, and service diversification*



4. Specialized collections are most likely to be found on servers located close to the greatest number of interested users.

In summary, the current network of general-purpose research libraries that strive toward uniformity of services will be replaced by a much denser network of small collections, each representing some specialized field of interest, including geographically specialized interest, and thus striving to provide unique services. These changes are consistent with a system of central places that experience massive increases in range and massive decreases in threshold, and responds by diversifying its services. The potential for a geographic basis to that diversification suggests that the library of the future will place much more emphasis on geographically referenced information.

The transition to a new geography is likely to be slow. The existing system has many stakeholders, with economic as well as other interests in maintaining it. Changes will be incremental, as the existing system attempts to support new services by grafting them onto the old. In the long term, however, it will be impossible to sustain a system that reflects earlier realities, and is already badly out of agreement with underlying fundamentals.

## CONCLUSIONS

In the early days of computing, the high fixed costs of hardware ensured a spatial organization that resembled many other instances of central facilities location, where a few facilities are located to serve a large and dispersed population. Institutions responded by creating computing centers with responsibility to install and run computing services. Through time, declines in fixed costs led to decentralization, and eventually to the personal computer, though central computing services were able to sustain themselves by emphasizing residual services of high fixed costs, such as consulting and standards setting.

The advent of high-bandwidth connectivity through the Internet has allowed users to obtain computing services from virtually anywhere on the planet, driving the effective range of computing to infinity. The impediments to computing no longer increase with distance from the nearest computer, since computers are virtually everywhere and connectivity is largely independent of distance. But the map of connectivity remains highly irregular, with massive variations in the bandwidth that can be obtained at a given location, and at a given cost.

The notion of computers as central facilities is consistent with the paradigm of computing that prevailed until the early 1990s. Since then, however, the integration of computers into a massive network has changed the paradigm, and given increasing weight to the communication function. Today, we tend to see the Internet as a means of communicating information between people, with processing as a way of adding value to what is communicated. This perspective gives much greater importance to the efficiencies of the communication channel, and to the value of digital communication vis-à-vis direct face-to-face communication.

From a geographic perspective, it is also important that computing be seen as a means of augmenting the experience of the senses, by providing access to stored

representations of the user's surroundings. Such means are of great interest in geographic field work, since they allow direct connection from the field to previous work, representations of prior states, and computed predictions of future states. They require that devices be able to sense location, and respond accordingly. Other g-commerce services become possible when these capabilities are integrated with others, such as the ability to identify the geographic locations of IP addresses.

Because distance fails to impede Internet interaction as directly as it does travel, the spatial organization of computing and e-commerce will likely be influenced by a number of less tangible factors. These include the tendency of geographically referenced information to be served from locations in or near its footprint; linkages between information processing and other activities that are not so footloose; the perceived commercial value that is conferred through association with places; and the persistent tendency for human communication to be negatively correlated with distance.

The chapter used the case of the research library to examine the impacts that digital technology is likely to have on spatial organization. These impacts are likely to be profound, and to force a transition from a system of central facilities, providing essentially similar services to a much denser system of facilities providing highly specialized services, with locations determined largely by concentration of specialized user interest. However, they address only certain functions of research libraries, and examination of other functions may lead to somewhat different conclusions.

According to recent reports, fully 30 percent of recent expansion of the US economy is directly attributable to IT. This alone would be sufficient to justify detailed examination of the technology's impact on spatial organization. But IT also impacts other activities, through its effects on the costs of communication. Arguably, then, the impacts of IT on spatial organization will be comparable to those of major transportation innovations of the past, including canals and railroads, and may even exceed them in significance. As geographers, we should be in an excellent position to anticipate these impacts, and to explore their ramifications for communities, economies, and institutions.

## ACKNOWLEDGEMENT

The National Center for Geographic Information and Analysis, the Alexandria Digital Library, and the Center for Spatially Integrated Social Science are supported by the National Science Foundation under cooperative agreements with, and grants to, the University of California, Santa Barbara.