COMPUTING MILIEUX

The Challenges of Mobile Computing

George H. Forman and John Zahorjan
University of Washington

Recent advances in technology have provided portable computers with wireless interfaces that allow networked communication even while a user is mobile. Whereas today's first-generation notebook computers and personal digital assistants (PDAs) are self-contained, networked mobile computers are part of a greater computing infrastructure. Mobile computing — the use of a portable computer capable of wireless networking — will very likely revolutionize the way we use computers.

Wireless networking greatly enhances the utility of a portable computing device. It allows mobile users versatile communication with other people and expedient notification of important events, yet with much more flexibility than with cellular phones or pagers. It also permits continuous access to the services and resources of land-based networks. The combination of networking and mobility will engender new applications and services, such as collaborative software to support impromptu meetings, electronic bulletin boards whose contents adapt to the current viewers, lighting and heating that adjust to the needs of those present, and navigation software to guide users in unfamiliar places and on tours.[1]

The technical challenges that mobile computing must surmount to achieve this potential are hardly trivial, however. Some of the challenges in designing software for mobile computing systems are quite different from those involved in the design of software for today's stationary networked systems. In this article we focus on the issues pertinent to software designers without delving into the lower level details of the hardware realization of mobile computers. We look at some promising approaches under investigation and also consider their limitations.

The many issues to be dealt with stem from three essential properties of mobile computing: communication, mobility, and portability. Of course, special-purpose systems may avoid some design pressures by doing without certain desirable properties. For instance, portability would be less of a concern for mobile computers installed in the dashboards of cars than with hand-held mobile computers. However, we concentrate on the goal of large-scale, hand-held mobile computing as a way to reveal a wide assortment of issues.

Computers are typically configured for use in a single location. The shift toward mobility and wireless communication is testing the abilities of designers to adapt traditional system structures.

0018-9162/ $4.00 © 1994 IEEE

Wireless communication

Mobile computers require wireless network access, although sometimes — when in meeting rooms or at a user's desk — they may remain stationary long enough to be physically attached to the network for a better or cheaper connection.

Wireless networks communicate by modulating radio waves or pulsing infrared light. Wireless communication is linked to the wired network infrastructure by stationary transceivers. The area covered by an individual transceiver's signal is known as a cell. Cell sizes vary widely; for example, an infrared transceiver can cover a small meeting room, a cellular telephone transceiver has a range of a few miles, and a satellite beam can cover an area more than 400 miles in diameter.

Wireless communication faces more obstacles than wired communication because the surrounding environment interacts with the signal, blocking signal paths and introducing noise and echoes. As a result, wireless communication is characterized by lower bandwidths, higher error rates, and more frequent spurious disconnections. These factors can in turn increase communication latency resulting from retransmissions, retransmission time-out delays, error-control protocol processing, and short disconnections.

Mobility can also cause wireless connections to be lost or degraded. Users may travel beyond the coverage of network transceivers or enter areas of high interference. Unlike typical wired networks, the number of devices in a network cell varies dynamically, and large concentrations of mobile users, say, at conventions and public events, may overload network capacity.

The need for wireless communication leads to design challenges in several areas.

Disconnection. Today's computer systems often depend heavily on a network and may cease to function during network failures. For example, distributed file systems may lock up waiting for other servers, and application processes may fail altogether if the network stays down too long.

Network failure is a greater concern in mobile computing than in traditional computing because wireless communication is so susceptible to disconnection. Designers must decide whether to spend available resources on the network, trying to prevent disconnections, or to spend them trying to enable systems to cope with disconnections more gracefully and work around them where possible.

The more autonomous a mobile computer, the better it can tolerate network disconnection. For example, certain applications can reduce communication by running entirely locally on the mobile unit rather than by splitting the application and the user interface across the network. In environments with frequent disconnections, it is better for a mobile device to operate as a stand-alone computer than as a portable terminal.

In some cases, both round-trip latency and short disconnections can be hidden by asynchronous operation. The X11 Window System uses this technique to achieve good performance. With the synchronous remote procedure call paradigm, the client waits for a reply after each request; in asynchronous operation, a client sends multiple requests before asking for acknowledgment. Similarly, prefetching and delayed write-back also decouple the act of communication from the actual time a program consumes or produces data, allowing it to proceed during network disconnections. These techniques, therefore, have the potential to mask some network failures.

The Coda file system provides a good example of how to handle network disconnections, although it is designed for today's notebook computers in which disconnections may be less frequent, more predictable, and longer lasting than in mobile computing.[2] Information from the user's profile helps in keeping the best selection of files in an on-board cache. It is important to cache whole files rather than blocks of files so that entire files can be read during a disconnection. When the network reconnects, Coda attempts to reconcile the cache with the replicated master repository.

With Coda, files can be modified even during disconnections. More conservative file systems disallow this to prevent multiple users from making inconsistent changes. Coda's optimism is justified by studies showing that only rarely are files actually shared in a distributed system; fewer than 1 percent of all writes are followed by a write from a different user.[2] If strong consistency guarantees are needed, clients can ask for them explicitly. Hence, providing flexible consistency semantics can allow greater autonomy.

Of course, not all network disconnec-

Portable terminal versus stand-alone computer

The role of a portable device has considerable latitude within the notion of mobile computing. Is it a terminal or an independent, stand-alone computer? How many purposes should the device serve? Should it incorporate a telephone like the AT&T EO? Should its work environment be that of a general-purpose workstation or of something more restrictive, say, the Apple Newton MessagePad? These design choices greatly affect the way we approach the important issues of communication, mobility, and portability. For example, a portable terminal such as Xerox PARC's Tab[1] is more dependent on a network but less prone to loss of storage media than a stand-alone computer. Such questions require careful consideration when designing software for mobile computing.

Reference

1. M. Weiser, "Some Computer Science Issues in Ubiquitous Computing," *Comm. ACM*, Vol. 36, No. 7, July 1993, pp. 75-84.

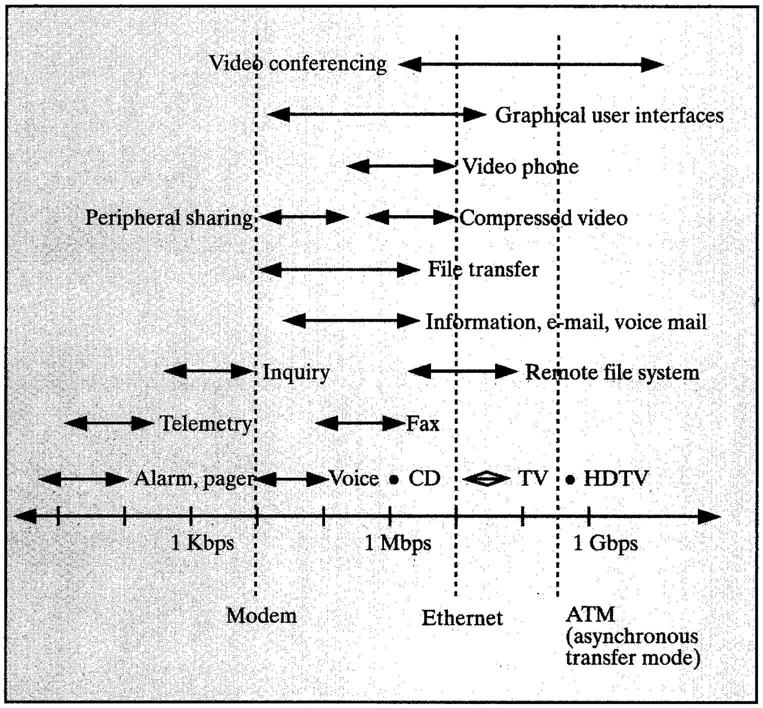


Figure 1. Application bandwidth requirements in bits per second. The vertical dashed lines show the bandwidth capability of certain network technologies. Cellular modems are becoming fast enough for mobile users' everyday information needs, such as e-mail, and someday may be able to support remote file systems.

tions can be masked. In these cases, good user interfaces can help by providing feedback about which operations are unavailable because of network disconnections.

Low bandwidth. Mobile computing designs need to reflect a greater concern for bandwidth consumption and constraints than do designs for stationary computing. Wireless networks deliver lower bandwidth than wired networks: Cutting-edge products for portable wireless communications achieve only 1 megabit per second for infrared communication, 2 Mbps for radio communication, and 9-14 kilobits per second for cellular telephony, while Ethernet provides 10 Mbps, fast Ethernet and FDDI 100 Mbps, and ATM (asynchronous transfer mode) 155 Mbps (see Figure 1). Even nonportable wireless networks, such as the Motorola Altair, barely achieve 5.7 Mbps.

Network bandwidth is divided among the users sharing a cell. The deliverable bandwidth *per user*, therefore, is an important measure of network capacity in addition to the raw transmission bandwidth. But because this measure depends on the size and distribution of a user population, Weiser and others recommend measuring a wireless network's capacity by its bandwidth per cubic meter.[1]

Improving network capacity means installing more wireless cells to service a user population. There are two ways to do this: Overlap cells on different wavelengths, or reduce transmission ranges so that more cells fit in a given area (see Figure 2).

The scalability of the first technique is limited because the electromagnetic spectrum available for public consumption is scarce. This technique is more flexible, however, because it allows (in fact, requires) software to allocate bandwidth among users.

The second technique is generally preferred. It is arguably simpler, reduces power requirements, and may decrease signal corruption because there are fewer objects in the environment to interact with. Also, it involves a hardware tradeoff between bandwidth and coverage area: Transceivers covering less area can achieve higher bandwidths.

Certain software techniques can also help cope with the low bandwidth of wireless links. Modems typically use compression to increase their effective bandwidth, sometimes almost doubling throughput. Because bulk operations are usually more efficient than many short transfers, buffering can improve bandwidth usage by making large requests out of many short ones. Buffering in conjunction with compression can further improve throughput because larger blocks compress better.

Certain software techniques for coping with disconnection can also help cope with low bandwidth. Network usage typically occurs in bursts, and disconnections are similar to bursts in that demand temporarily exceeds available bandwidth. For example, delayed write-back and prefetching use the periods of low network activity to reduce demand at the peaks. Delayed write-back can even reduce overall communication if the data to be transmitted is further mutated or deleted before it is transmitted. Prefetching involves knowing or guessing which files will be needed soon and downloading them over the network before they are demanded.[3] Bad guesses can waste network bandwidth, however.

System performance can be improved by scheduling communication intelligently. When available bandwidth does not satisfy the demand, processes the user is waiting for should be given priority. Backups should be performed only with "leftover" bandwidth. Mail can be trickle fed onto the mobile computer slowly before the user is notified. Although these techniques do not increase effective bandwidth, they improve user satisfaction just the same.

High bandwidth variability. Mobile computing designs also contend with much greater variation in network bandwidth than do traditional designs. Band-

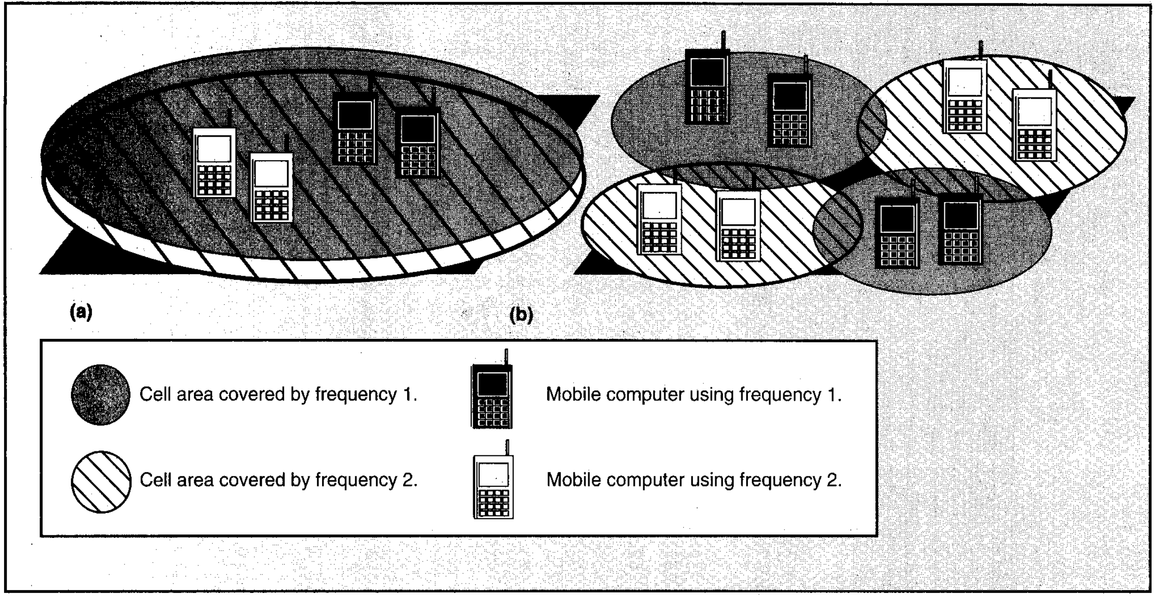

(a)
(b)
Cell area covered by frequency 1.
Mobile computer using frequency 1.
Cell area covered by frequency 2.
Mobile computer using frequency 2.

Figure 2. Suppose that a single frequency provides only enough wireless bandwidth for two users. Then two frequencies can support (a) four users with two large coincident cells or (b) eight users with four small noninterfering cells that use the same frequency in nonadjacent cells. The latter scheme requires more transceivers and installation effort but is more scalable and allows higher bandwidth technology and lower transmission power.

width can shift one to four orders of magnitude, depending on whether the system is plugged in or using wireless access. Fluctuant traffic load seldom causes this much variation in available bandwidth on today's networks.

An application can approach this variability in one of three ways: It can assume high-bandwidth connections and operate only while plugged in, it can assume low-bandwidth connections and not take advantage of higher bandwidth when it is available, or it can adapt to currently available resources, providing the user with a variable level of detail or quality. For example, a video-conferencing application could display only the current speaker or all the participants, depending on the available bandwidth. Different choices make sense for different applications.

Heterogeneous network. In contrast to most stationary computers, which stay connected to a single network, mobile computers encounter more heterogeneous network connections in several ways. First, as they leave the range of one network transceiver and switch to another, they may also need to change transmission speeds and protocols. Second, in some situations a mobile computer may have access to several network connections at once, for example, where adjacent cells overlap or where it can be plugged in for concurrent wired access.

Also, mobile computers may need to switch interfaces, for example, when going between indoors and outdoors. Infrared interfaces cannot be used outside because sunlight drowns out the signal. Even with radio frequency transmission, the interface may still need to change access protocols for different networks, for example, when switching from cellular coverage in a city to satellite coverage in the country. This heterogeneity makes mobile networking more complex than traditional networking.

Security risks. Precisely because connection to a wireless link is so easy, the security of wireless communication can be compromised much more easily than that of wired communication, especially if transmission extends over a large area. This increases pressure on mobile computing software designers to include security measures.

Security is further complicated if users are allowed to cross security domains. For example, a hospital may allow patients with mobile computers to use nearby printers but prohibit access to distant printers and resources designated for hospital personnel only.

Secure communication over insecure channels is accomplished by encryption, which can be done in software or, more quickly, by specialized hardware such as the recently proposed Clipper chip. Security depends on a secret encryption key known only to the authorized parties. Managing these keys securely is difficult, but it can be automated by software such as the Massachusetts Institute of Technology's Kerberos.[4]

Kerberos provides secure authentication services, as long as the Kerberos server itself is trusted. It authenticates users without exposing their passwords on the network and generates secret encryption keys that can be selectively shared between mutually suspicious parties. It also allows mobile units to authenticate themselves in domains where they are unknown, thus enhancing the scale of mobility. Methods have also been devised to use Kerberos for authorization control and accounting. Its security is limited, however. For example, the current version is susceptible to off-line password-guessing attacks and to replay attacks for a limited time window.

Mobility

The ability to change locations while connected to the network increases the volatility of some information. Certain data considered static for stationary computing becomes dynamic for mobile computing. For example, a stationary computer can be configured statically to prefer the nearest server, but a mobile computer needs a mechanism for determining which server to use.

As volatility increases, cost-benefit trade-off points shift, calling for appropriate modifications in the design. For example, a highly volatile data object has fewer uses per modification. For such objects it makes little sense to cache the data. As another example, consider static information, which is often managed by hand; to handle higher rates of change, automated methods are required. However, even where such methods exist, they may be ill-suited for the dynamism of mobile computing.

Mobility introduces several problems: A mobile computer's network address changes dynamically, its current location affects configuration parameters as well as answers to user queries, and the communication path grows as it wanders away from a nearby server.

Address migration. As people move, their mobile computers will use different network access points, or "addresses." Today's networking is not designed for dynamically changing addresses. Active network connections usually cannot be moved to a new address. Once an address for a host name is known to a system, it is typically cached with a long expiration time and with no way to invalidate out-of-date entries. In the Internet Protocol, for example, a host IP name is inextricably bound with its network address; moving to a new location means acquiring a new IP name. Human intervention is commonly required to coordinate the use of addresses.

To communicate with a mobile computer, messages must be sent to its most recent address. Four basic mechanisms determine a mobile computer's current address: broadcast,[5,6] central services,[7] home base,[8] and forwarding pointers.[5] These are the building blocks of the current proposals for "mobile-IP" schemes.

As people move, their mobile computers will use different network access points, or "addresses."

Selective broadcast. With the broadcast method, a message is sent to all network cells asking the mobile computer sought to reply with its current address. This becomes too expensive for frequent use in a large network, but if the mobile computer is known to be in some small set of cells, selectively broadcasting in just those cells is workable. Hence, the methods described below can use selective broadcast to obtain the current address when only approximate location information is known. For example, a slightly out-of-date cell address may suffice if adjacent cells are known.

Central services. With the central service method, the current address for each mobile computer is maintained in a logically centralized database. Each time a mobile computer changes its address, it sends a message to update the database. Even with the database's centralized location, the common techniques of distribution, replication, and caching can be used to improve availability and response time.

Home bases. The home base method is essentially the limiting case of distributing a central service; that is, the location of a given mobile computer is known by a single server. This aggressive distribution without replication can lead to low availability of information. For example, if a home base is down or inaccessible, the mobile computers it tracks cannot be contacted. If users sometimes change home bases, the address migration problem arises again, albeit with much lower volatility.

Forwarding pointers. With the forwarding pointer method, each time a mobile computer changes its address, a copy of the new address is deposited at the old location. Each message is forwarded along the chain of pointers until it reaches the mobile computer. To avoid the inefficient routing that can result from long chains, pointers at message forwarders can be updated gradually to reflect more recent addresses.

Although the forwarding pointer method is among the fastest, it is prone to failures anywhere along the trail of pointers, and in its simplest form it does not allow forwarding pointers to be deleted unless all possible message sources have been updated. Hence, forwarding pointers are often used only to speed the common case, and another method is used to fall back on for failures and to allow reclamation of old pointers.

The forwarding pointer method requires an active entity at the old address to receive and forward messages. This does not fit standard networking models, where either a network address is a passive entity, such as an Ethernet cable, or it's specific to the mobile computer, which cannot remain to forward its own messages. This mismatch introduces subtle difficulties in implementing forwarding efficiently (for example, with intracell traffic or when multiple gateways are attached to a network address).

Location-dependent information. Because traditional computers do not move, information that depends on location, such as the local name server, available printers, and the time zone, is typically configured statically. One challenge for mobile computing is to factor out this information intelligently and provide mechanisms for obtaining configuration data appropriate to each location. Additionally, a mobile computer carried with a user is likely to be used in a wide variety of administrative domains. Dealing with the multitude of conventions that current computing systems rely on is another challenge to building mobile systems.

Besides this dynamic configuration problem, mobile computers need access to more location-sensitive information than stationary computers do. If they are to serve as guides in places unfamiliar to their users, mobile computers may need to answer queries such as "where is the fiction section (in this particular library)?" or "where is the nearest open gas station heading north?"

Queries of this sort require static location information about the world. Other information needs can be even more complex: Badrinath and Imielinski are studying a related class of queries that depend on the dynamic locations of other mobile objects, for example, the location of the nearest taxi.[6]

Privacy. Answering dynamic location queries requires knowing the location of another mobile user. In some cases this may be sensitive information, more so if given at a fine resolution. Even where it is not particularly sensitive, such information should be protected against misuse; for example, we do not want a burglar to be able to determine when the inhabitants of a house are far away.

Privacy can be ensured by denying users the ability to know another's location. The challenge for mobile computing is to allow more flexible access to this information without violating privacy. Legitimate uses of location information include contacting colleagues, routing telephone calls, logging meetings in personal diaries, and tailoring the content of electronic announcement displays to the current viewers.[1]

Migrating locality. Mobile computing engenders a new kind of locality that migrates as users move. Even if a mobile computer is equipped to find the nearest server for a given service, over time migration may alter this condition. Because the physical distance between two points does not necessarily reflect the network distance, the communication path can grow disproportionately to actual movement. For example, a small movement can result in a much longer path when crossing network administrative boundaries, and a longer network path means communication traverses more intermediaries, resulting in longer latency and greater risk of disconnection. A longer communication path also consumes more network capacity, even though the bandwidth between the mobile unit and the server may not degrade.

Mobile computers need access to more location-sensitive information than do stationary computers.

To avoid these disadvantages, service connections may be dynamically transferred to servers that are closer.[9] When many mobile units converge, during meetings, for example, load-balancing concerns may outweigh the importance of communication locality.

Portability

Today's desktop computers are not meant to be carried, so designers take a liberal approach to space, power, cabling, and heat dissipation. In contrast, designers of hand-held mobile computers should strive for the properties of a wristwatch: small, light, durable, operational under wide environmental conditions, and requiring minimal power usage for long battery life. Concessions can be made in each of these areas to enhance functionality, but ultimately the user must receive value that exceeds the trouble of carrying the device. Similarly, any specialized hardware to offload such tasks as data compression or encryption from the CPU should justify its consumption of power and space.

Below, we describe the design pressures caused by portability constraints. These pressures are evident in the designs of the recent PDA products listed in Table 1.

Low power. Batteries are the largest single source of weight in a portable computer. While reducing battery weight is important, too small a battery can undermine the value of portability by causing users to recharge frequently, carry spare batteries, or use their mobile computers less. Minimizing power consumption can improve portability by reducing battery weight and lengthening the life of a charge.

Power consumption of dynamic components is proportional to CV^2F, where C is the capacitance of the circuit, V is the voltage swing, and F is the clock frequency. This function suggests three ways to save power:

(1) Capacitance can be reduced by greater levels of VLSI integration and multichip module technology.
(2) Voltage can be reduced by redesigning chips to operate at lower voltages. Historically, chips operate at 5 volts, but some, like those in the Apple MessagePad, save power by operating at 3 volts. Manufacturers are rapidly developing a line of low-power chip sets for 2.5V and 3.3V operation.
(3) Clock frequency can be reduced, thereby trading computational speed for power savings. PDA products have adopted this concession, as shown in Table 1. In some notebook computers, the clock frequency can be changed dynamically, providing a flexible trade-off; for example, the Sharp PC 6785 can save power by dynamically shifting its clock from 25 MHz to 10 MHz or even 5 MHz, as shown in Table 2. To

Table 1. Characteristics of personal digital assistant products and the AT&T EO tablet computer. Each has a pen interface and a black-and-white reflective LCD screen. The portable PC is included for comparison.

Product	RAM	MHz	CPU	Batteries		Weight	Display	
				(No. hours)	(type)	(lbs.)	(pixels)	(sq. in.)
Amstrad Pen Pad PDA600	128 Kbytes	20	Z-80	40	3 AA's	0.9	240 x 320	10.4
Apple Newton MessagePad	640 Kbytes	20	ARM*	6-8	4 AAA's	0.9	240 x 336	11.2
Apple Newton MessagePad 110	1 Mbyte	20	ARM*	50	4 AA's	1.25	240 x 320	11.8
Casio Z-7000 PDA	1 Mbyte	7.4	8086	100	3 AA's	1.0	320 x 256	12.4
Sharp Expert Pad	640 Kbytes	20	ARM*	20	4 AAA's	0.9	240 x 336	11.2
Tandy Z-550 Zoomer PDA	1 Mbyte	8	8086	100	3 AA's	1.0	320 x 256	12.4
AT&T EO 440 Personal Communicator	4-12 Mbytes	20	Hobbit	1-6	NiCad	2.2	640 x 480	25.7
Portable PC	4-16 Mbytes	33-66	486	1-6	NiCad	5-10	640 x 480 to 1,024 x 768 (color)	40

*Advanced RISC microprocessor

retain more computational power at lower frequencies, processors are being designed that perform more work on each clock cycle.[10]

Power can be conserved not only by the design but also by efficient operation. Power management software can power down individual components when they are idle, for example, spinning down the internal disk or turning off screen lighting. Recently, Li et al. determined that for today's notebook computing it is worthwhile to spin down the internal disk drive after it has been idle for just a few seconds.[11]

Applications can conserve power by reducing their appetite for computation, communication, and memory, and by performing their periodic operations infrequently to amortize the start-up overhead. Since radio modem transmission typically requires about 10 times as much power as reception, power can be saved by trading transmission for reception. For example, base stations might periodically broadcast information that otherwise would have to be requested frequently. In this way, mobile computers can obtain this information without expending power to transmit a request.

The potential savings of these techniques can be evaluated using Tables 2 and 3, which break down power consumption in notebook computers by component and subsystem, respectively. Although screen lighting consumes a large amount of power, it has been found to greatly improve readability; for example, on EO models it enhances contrast from 6:1 to 13:1. Nevertheless, PDA products have elected to omit screen lighting in favor of longer battery life.

Risks to data. Making computers portable increases the risk of physical damage, unauthorized access, loss, and theft. Breaches of privacy or total loss of data become more likely. These risks can be reduced by minimizing the essential data kept on board. Obviously, a mobile device that serves only as a portable terminal is less prone to data loss than a stand-alone computer. This is the approach taken for Xerox PARC's Tabs and the portable multimedia terminal project at the University of California, Berkeley.[10]

To help prevent unauthorized disclosure of information, data stored on disks and removable memory cards can be encrypted. For this to be effective, users must not leave authenticated sessions (logins) unattended.

Keeping a copy that does not reside on the portable unit can safeguard against data loss. However, users often neglect to make backup copies, and even when they do, data modified between backups is not protected. With the addition of wireless networks to portable computers, new or modified data can be copied immediately to secure, remote media. This can be accomplished with replicated file systems such as Echo and Coda.[2]

Table 2. Power consumption of portable-computer components and accessories.*

Device	Power (watts)
Base system (2 Mbytes, 25-MHz CPU)	3.650
Base system (2 Mbytes, 10-MHz CPU)	3.150
Base system (2 Mbytes, 5-MHz CPU)	2.800
Screen backlight	1.425
Hard drive motor	1.100
Math coprocessor	0.650
Floppy drive	0.500
External keyboard	0.490
LCD screen	0.315
Hard drive active (head seeks)	0.125
IC card slot	0.100
Additional memory (per Mbyte)	0.050
Parallel port	0.035
Serial port	0.030
Accessories	
1.8-inch PCMCIA hard drive	0.7-3.0
Cellular telephone (active)	5.400
Cellular telephone (standby)	0.300
Infrared network, 1 Mbit per second**	0.250
PCMCIA modem, 14,400 bits per second	1.365
PCMCIA modem, 9,600 bits per second	0.625
PCMCIA modem, 2,400 bits per second	0.565
Global positioning receiver**	0.670

*Data for computer components was derived from the Sharp PC 6785 manual; data for accessories was obtained from the manufacturers.
**Estimate for soon-to-be-released product.

Small user interface. Size constraints on a portable computer require a small user interface. Desktop windowing environments may be sufficient for today's notebook computers, but for smaller, more portable devices, current windowing technology is inadequate. On small displays it is impractical to have several windows open at a time regardless of screen resolution, and it can be difficult to locate windows or icons deeply stacked one on another. Also, window title bars and borders either consume significant portions of screen space or, if reduced, become difficult to operate with the pointing device.

Duchamp, Feiner, and Maguire investigated the use of head-mounted virtual reality displays for portable computers.[9] As the user's head turns, the image displayed to the eye shifts to give the sensation of a surrounding screen. This effectively increases the screen area available for windowing systems; however, wearing head gear is cumbersome, resolution is low (one-tenth that of conventional displays), eyes succumb to fatigue, and dim lighting is required.

Buttons versus analog input. The shortage of surface area on a small computer leads designers to sacrifice buttons in favor of analog input devices for communicating user commands. These communication mechanisms include handwriting recognition, gesture recognition, and voice recognition. Although on average handwriting is about three times slower than typing, it allows the keyboard to be eliminated, thus reducing size and improving durability. This approach has been adopted by all the PDA products listed in Table 1.

Handwriting recognition rates for high-end systems are typically 96-98 percent accurate when trained to a specific user.[12] With context information, recognition rates can be enhanced effectively to 100 percent, but context constraints do not help for all kinds of input, for example, when entering words that are not in the on-line dictionary. The Apple Newton's handwriting recognition, while among the best of the PDAs, is nevertheless reputedly a source of frustration. Recognizing a user's intention in a general setting is inherently hard because the interpretation of pen strokes is ambiguous. For example, a user drawing a circle may intend to select an object or an area, or write a zero, a degree sign, or the letter *o*.

Speech generation and rec-

Table 3. Power consumption of the major components in a portable computer.*

System	Power (percent)
Display edge-light	35
CPU/memory	31
Hard disk	10
Floppy disk	8
Display	5
Keyboard	1

*Data was obtained from the Compaq LTE 386/s20 manual.

ognition seem an ideal user interface for a mobile computer in that they require no surface area and allow hands-free and even eye-free operation. The voice-commanded VCR programmer by Voice Powered Technology demonstrates the feasibility of speech input to a hand-held device for a narrow domain. The Sphinx research project at Carnegie Mellon University has reported speaker-independent recognition rates of nearly 96 percent, and 98 percent for speaker-trained recognition. However, general-purpose speech input and output places substantial storage and processing demands on a mobile device. Also, speech may often be inappropriate: It disturbs others in quiet environments, it cannot be recognized clearly in noisy environments, and it can compromise privacy. Finally, because of its sequential nature, speech is ill-suited for skimming data.

Pointing devices. The mouse is the standard pointing device for desktop computers, but it doesn't suit mobile computers. Pens have become the standard input device for PDAs because of their ease of use while mobile, their versatility, and their ability to supplant the keyboard.

Switching to pens requires changing both the user interface and the software interface because a mouse and a pen are really quite different.[9] Users can jump to absolute screen positions and enter path information more easily with a pen than with a mouse, and it is nearly impossible to write with a mouse. Pen-positioning resolution on current tablet computers is several times more accurate than screen resolution; for example, pen resolution on the EO is 0.10 mm, while screen resolution is 0.23-0.30 mm. Also, parallax between the pen tip and the screen image can mislead when pointing; with a mouse there is no parallax because the mouse cursor provides feedback in the image plane. Finally, the mouse cursor obscures much less of the screen than the user's hand does when writing with a pen.

Small storage capacity. Storage space on a portable computer is limited by physical size and power requirements. Traditionally, disks provide large amounts of nonvolatile storage. In a mobile computer, however, disk drives are a liability. They consume more power than memory chips, except when off line, and they may not really be nonvolatile when subject to the indelicate treatment a portable device receives. Hence, none of the PDA products have disk drives.

Coping with limited storage is not a new problem. Solutions include compressing files automatically, accessing remote storage over the network, sharing code libraries, and compressing virtual memory pages. Although today's networked computers have had great success with distributed file systems and remote paging, mobile computers that regularly encounter network disconnections are less capable of relying on a network.

A novel approach to reducing the size of program code is to interpret script languages instead of executing compiled object codes, which are typically many times larger than the source code. This approach is embodied by General Magic's Telescript and Apple Technology Group's Dylan and NewtonScript. An equally important goal of such languages is to enhance portability by supporting a common programming model across different machines.

Mobile computing is a technology that enables access to digital resources at any time, from any location. From a narrow viewpoint, mobile computing represents a convenient addition to wire-based local area distributed systems. Taken more broadly, mobile computing represents the elimination of time-and-place restrictions imposed by desktop computers and wired networks.

In forecasting the impact of mobile technology, we would do well to observe recent trends in the use of the wired infrastructure, in particular, the Internet. In the past year, the advent of convenient mechanisms for browsing Internet resources has engendered an explosive growth in the use of those resources. The ability to access them at all times through mobile computing will allow their use to be integrated into all aspects of life and will accelerate the demand for network services. The challenge for computing designers is to adapt the system structures that have worked well for traditional computing so that mobile computing can be integrated as well. ■

Acknowledgments

Support for this work was provided in part by the National Science Foundation (grants CCR-9123308 and CCR-9200832), Tektronix Inc. (a graduate fellowship), the Washington Technology Center, and Digital Equipment Corp. (Systems Research Center and External Research Program). We thank Robert Bedichek, Brian Bershad, Blake Hannaford, Marc Fiuczynski, Brian Pinkerton, and Stefan Savage for helpful pointers and clarifying discussions that significantly improved this article.

References

1. M. Weiser, "Some Computer Science Issues in Ubiquitous Computing," *Comm. ACM*, Vol. 36, No. 7, July 1993, pp. 75-84.

2. J.J. Kistler and M. Satyanarayanan, "Disconnected Operation in the Coda File System," *ACM Trans. Computer Systems*, Vol. 10, No. 1, Feb. 1992, pp. 3-25.

3. C.D. Tait and D. Duchamp, "Detection and Exploitation of File Working Sets," *Proc. 11th Int'l Conf. Distributed Computing Systems*, IEEE CS Press, Los Alamitos, Calif., Order No. 2144, 1991, pp. 2-9.

4. B.C. Neuman, "Protection and Security Issues for Future Systems," in *Workshop on Operating Systems of the 90s and Beyond*, Lecture Notes in Computer Science #563, Springer-Verlag, New York, 1991, pp. 184-201.

5. J. Ioannisdis, D. Duchamp, and G.Q. Maguire Jr., "IP-Based Protocols for Mobile Internetworking," *Proc. SIGComm 91 Symp.*, ACM, New York, 1991, pp. 235-245.

6. T. Imielinski and B.R. Badrinath, "Data Management for Mobile Computing," *SIGMOD Record*, Vol. 22, No. 1, Mar. 1993, pp. 34-39.

7. C. Ma, "On Building Very Large Naming Systems," *Proc. Fifth ACM SIGOPS Workshop Models and Paradigms for Distributed Systems Structuring*, ACM, New York, 1992, 5 pp.

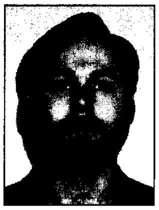
8. F. Teraoka and M. Tokoro, "Host Migration Transparency in IP Networks: The VIP Approach," *Computer Comm. Rev.*, Vol. 23, No. 1, Jan. 1993, pp. 45-65.

9. D. Duchamp, S.K. Feiner, and G.Q. Maguire Jr., "Software Technology for Wireless Mobile Computing," *IEEE Network Magazine*, Vol. 5, No. 6, Nov. 1991, pp. 12-18.

10. A. Chandrakasan, S. Sheng, and R.W. Brodersen, "Design Considerations for a Future Portable Multimedia Terminal," in *Third-Generation Wireless Information Networks*, S. Nanda and D.J. Goodman, eds., Kluwer Academic Publishers, Hingham, Mass., 1992, pp. 75-97.

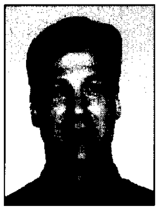
11. K. Li et al., "A Quantitative Analysis of Disk Drive Power Management in Portable Computers," tech. report, Computer Science Division, University of California, Berkeley, Calif., 1993.

12. C.C. Tappert, C.Y. Suen, and T. Wakahara, "On-Line Handwriting Recognition — A Survey," *Proc. Ninth Int'l Conf. Pattern Recognition*, Vol. 2, IEEE CS Press, Los Alamitos, Calif., Order No. 878, 1988, pp. 1,123-1,132.



George Forman is a PhD candidate in the Department of Computer Science and Engineering at the University of Washington. His research interests include mobile computing and compilers for parallel computers.

Forman received his BA in mathematics from Pomona College, California, in 1988. The following year he received a Fulbright fellowship for study at the Swiss Federal Institute of Technology, Zurich. He is a member of Sigma Xi and Phi Beta Kappa.



John Zahorjan is a professor of computer science and engineering at the University of Washington. His research interests include performance modeling and experimental evaluations, as well as issues in mobile computing, runtime support for parallel computing, and resource scheduling for continuous-media applications.

Zahorjan received an ScB in applied mathematics from Brown University in 1975 and MSc and PhD degrees in computer science from the University of Toronto in 1976 and 1980, respectively. In 1984 he received a Presidential Young Investigator Award from the National Science Foundation. He is a member of the IEEE Computer Society.

The authors can be contacted at the Department of Computer Science and Engineering, FR-35, University of Washington, Seattle, WA 98195, e-mail {forman, zahorjan}@cs.washington.edu.