SECTION 1

Introductions to the Main Topics

CHAPTER 1

Introduction to Mobile Computing

> Where is the life we have lost in living? Where is the knowledge we have lost in information? Where is the wisdom we have lost in knowledge?
>
> T. S. Elliot

1.1 INTRODUCTION

Mobile computing systems are computing systems that may be easily moved physically and whose computing capabilities may be used while they are being moved. Examples are laptops, personal digital assistants (PDAs), and mobile phones. By distinguishing mobile computing systems from other computing systems we can identify the distinctions in the tasks that they are designed to perform, the way that they are designed, and the way in which they are operated. There are many things that a mobile computing system can do that a stationary computing system cannot do; these added functionalities are the reason for separately characterizing mobile computing systems.

Among the distinguishing aspects of mobile computing systems are their prevalent wireless network connectivity, their small size, the mobile nature of their use, their power sources, and their functionalities that are particularly suited to the mobile user. Because of these features, mobile computing applications are inherently different than applications written for use on stationary computing systems. And, this brings me to the central motivation behind authoring this book.

The application development and software engineering disciplines are very young engineering disciplines compared to those such as structural, mechanical, and electrical engineering. Software design and implementation, for the most part,

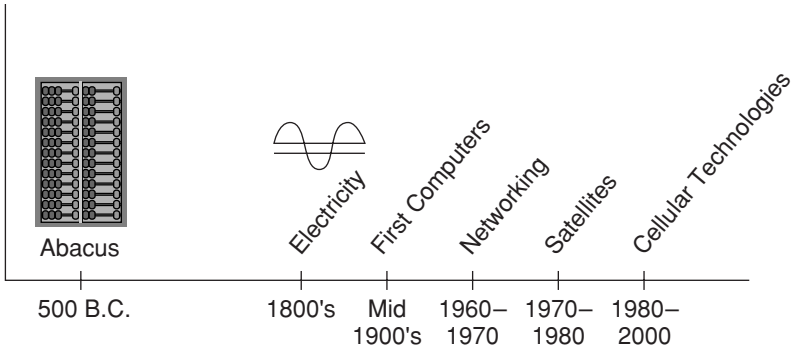

FIGURE 1.1. A Timeline of Mobile Computing.

remain part art and part science. However, there are definite signs of maturation with the development of architectures, metrics, proven tools, and other methodologies that give an engineering discipline its structure. Whereas there are a variety of methodologies, techniques, frameworks, and tools that are used in developing software for stationary systems, there are very few for mobile systems. Although mobile computing systems have existed as long as their stationary counterparts, most of the mature tools, methodologies, and architectures in software engineering today address the needs of stationary systems. One of our goals in this book will be to reflect on the research being done today to help evolve mobile application development and to outline some of the early proven techniques and technologies being tried in the commercial and academic environments.

In this text, we will look at those things that make the functional nature of mobile applications different than their stationary counterparts, take a survey of various development techniques that can be used to address these differences, and look at various basic technologies that allow us, as software developers, to create meaningful mobile applications in an extensible, flexible, and scalable manner.

1.1.1 A Brief History of Mobile Computing

Figure 1.1 shows a timeline of mobile computing development. One of the very first computing machines, the abacus, which was used as far back as 500 B.C., was, in effect, a mobile computing system because of its small size and portability. As technology progressed, the abacus evolved into the modern calculator. Most calculators today are made with an entire slew of mathematical functions while retaining their small size and portability. The abacus and calculators became important parts of technology not only because of their ability to compute but also because of their ease of use and portability. You can calculate the proceeds of a financial transaction anywhere as long as you had an abacus in 500 B.C. or have a calculator today. But, calculating numbers is only one part of computing.

Other aspects of computing, namely storage and interchange of information, do not date as far back as the abacus. Though writing has always been a way of storing information, we can hardly call a notebook a computing storage mechanism. The first mobile storage systems can be traced back only as far as the advent of the age of electronics.

FIGURE 1.2. Wireless Communication Systems.

A mobile computing system, as with any other type of computing system, can be connected to a network. Connectivity to the network, however, is not a prerequisite for being a mobile computing system. Dating from the late 1960s, networking allowed computers to talk to each other. Networking two or more computers together requires some medium that allows the signals to be exchanged among them. This was typically achieved through wired networks. Although wired networks remain the predominant method of connecting computers together, they are somewhat cumbersome for connecting mobile computing devices. Not only would network ports with always-available network connectivity have to be pervasive in a variety of physical locations, it would also not be possible to be connected to the network in real time if the device were moving. Therefore, providing connectivity through a wired system is virtually cost prohibitive. This is where wireless communication systems come to the rescue (Figure 1.2).

By the 1960s, the military had been using various forms of wireless communications for years. Not only were wireless technologies used in a variety of voice communication systems, but the aviation and the space program had created great advances in wireless communication as well. First, the military developed wireless communication through line of sight: If there were no obstacles between point A and point B, you could send and receive electromagnetic waves. Then

came techniques that allowed for wireless communication to encompass larger areas, such as using the atmosphere as a reflective mechanism. But, there were limitations on how far a signal could reach and there were many problems with reliability and quality of transmission and reception.

By the 1970s, communication satellites began to be commercialized. With the new communication satellites, the quality of service and reliability improved enormously. Still, satellites are expensive to build, launch, and maintain. So the available bandwidth provided by a series of satellites was limited. In the 1980s cellular telephony technologies became commercially viable and the 1990s were witness to advances in cellular technologies that made wireless data communication financially feasible in a pervasive way.

Today, there are a plethora of wireless technologies that allow reliable communication at relatively high bandwidths. Of course, bandwidth, reliability, and all other qualitative and quantitative aspects of measuring wireless technologies are relative to time and people's expectations (as seems to be with everything else in life!). Though most wireless networks today can transmit data at orders of magnitude faster speeds than just ten years ago, they are sure to seem archaically slow soon. It should, however, be noted that wired communication systems will almost certainly always offer us better reliability and higher data transmission bandwidths as long as electromagnetic communications is the primary means of data communications. The higher frequency sections of the electromagnetic spectrum are difficult to use for wireless communications because of natural noise, difficulty of directing the signal (and therefore high losses), and many other physical limitations. Since, by Nyquist's principle [Lathi 1989], the bandwidth made available by any communication system is bound by the frequencies used in carrying the signal, we can see that lack of availability of higher frequency ranges places a limitation on wireless communication systems that wired communication systems (such as fiber optic–based systems) do not have to contend with.

Because the greatest advances in mobile communications originated in the military, it is no surprise that one of the first applications of wireless communication for mobile computing systems was in displaying terrain maps of the battlefield. From this, the global positioning system (GPS) evolved so that soldiers could know their locations at any given time. Portable military computers were provided to provide calculations, graphics, and other data in the field of battle. In recent years, wireless telephony has become the major provider of a revenue stream that is being invested into improving the infrastructure to support higher bandwidth data communications.

1.1.2 Is Wireless Mobile or Is Mobile Wireless?

In wireless connectivity, mobile computing devices found a great way to connect with other devices on the network. In fact, this has been a great source of confusion between *wireless communications* and *mobile computing*. Mobile computing devices need not be wireless. Laptop computers, calculators, electronic watches, and many other devices are all mobile computing devices. None of them use any sort of wireless communication means to connect to a network. Even some hand-held personal assistants can only be synchronized with personal computers through

a docking port and do not have any means of wireless connectivity. So, before we embark on our journey in learning about mobile computing, it should be clear that *wireless communication systems are a type of communication system. What distinguishes a wireless communication system from others is that the communication channel is space itself.* There are a variety of physical waveguide channels such as fiber optics or metallic wires. Wireless communication systems do not use a waveguide to guide along the electromagnetic signal from the sender to the receiver. They rely on the mere fact that electromagnetic waves can travel through space if there are no obstacles that block them. Wireless communication systems are often used in mobile computing systems to facilitate network connectivity, but they are not mobile computing systems.

Recently, computer networks have evolved by leaps and bounds. These networks have begun to fundamentally change the way we live. Today, it is difficult to imagine computing without network connectivity. Networking and distributed computing are two of the largest segments that are the focus of current efforts in computing. Networks and computing devices are becoming increasingly blended together. Most mobile computing systems today, through wired or wireless connections, can connect to the network. Because of the nature of mobile computing systems, network connectivity of mobile systems is increasingly through wireless communication systems rather than wired ones. And this is quickly becoming somewhat of a nonmandatory distinguishing element between mobile and stationary systems. Though it is not a requirement for a mobile system to be wireless, most mobile systems are wireless. Nevertheless, let us emphasize that wireless connectivity and mobility are orthogonal in nature though they may be complementary. For example, we can have a PDA that has no wireless network connectivity; however, most PDAs are evolving into having some sort of wireless connectivity to the network.

Also, though it is important to understand that stationary and mobile computing systems are inherently different, this does not mean that they do not have any commonalities. We will build on existing software technologies and techniques used for stationary systems where these commonalities exist or where there is a logical extension of a stationary technique or technology that will mobilize it.

Because of the constant comparison between mobile systems and other types of systems, we will have to have a way to refer to the "other" types of systems. We will use the terms *nonmobile* and *stationary* interchangeably. Although mobile is an industry-wide accepted terminology to distinguish a group of systems with the characteristics that we have just mentioned, there is no consensus on a system that is not a mobile system. For this reason, we will simply use the terms stationary or nonmobile when speaking of such systems. It is also important to note the there is probably no system that is truly not mobile because just about any system may be moved. We will assume that cranes, trucks, or other large vehicles are not required for moving our mobile systems! A mobile system should be movable very easily by just one person.

There are four pieces to the mobile problem: the mobile user, the mobile device, the mobile application, and the mobile network. We will distinguish the mobile user from the stationary user by what we will call the *mobile condition:*

the set of properties that distinguishes the mobile user from the user of a typical, stationary computing system. We will wrap the differences between typical devices, applications, and networks with mobile devices, applications, and networks into a set of properties that we will call the *dimensions of mobility*: *the set of properties that distinguishes the mobile computing system from the stationary computing system*. Once we have some understanding of the mobile problem, we will look at some established nonproprietary methodologies and tools of the software industry trade such as Unified Modeling Language (UML) as well as some commercial proprietary tools such as Sun Microsystem's Java, Microsoft's Windows CE, Symbian, and Qualcomm's BREW. Once we have looked at these tools, we will set out to solve the problem of architecting, designing, and implementing solutions for mobile computing problems.

Let us start by looking at some of those variables that create a distinction between mobile and stationary computing systems.

1.2 Added Dimensions of Mobile Computing

It should be obvious that any mobile computing system can also be stationary! If we stop moving it, it is stationary. So, we can say that mobile computing systems are a superset of stationary computing systems. Therefore, we need to look at those elements that are outside of the stationary computing subset. These added dimensions will help us pick out variables that in turn allow us to divide and conquer the problems of mobile computing. The *dimensions of mobility*, as we will refer to them in this text, will be the tools that allow us to qualify our problem of building mobile software applications and mobile computing systems. Although these dimensions of mobility are not completely orthogonal with respect to each other, they are separate enough in nature that we can distinguish them and approximate them as orthogonal variables. Also, keep in mind that some of these dimensions are limitations; nevertheless, they are still added dimensions that need not be considered when dealing with the typical stationary application. These dimensions of mobility (Figure 1.3) are as follows:

1. location awareness,
2. network connectivity quality of service (QOS),
3. limited device capabilities (particularly storage and CPU),
4. limited power supply,
5. support for a wide variety of user interfaces,
6. platform proliferation, and
7. active transactions.

It is absolutely crucial that the reader understands these dimensions of mobility and keeps them in mind throughout the process of design and implementation of the mobile application. Too often, engineers begin with attention to design and get bogged down in details of the tools that they use and small focused problems within the bigger picture of the system, its design, and its architecture. The definition of the word "mobile" reveals the first dimension we will consider: location.

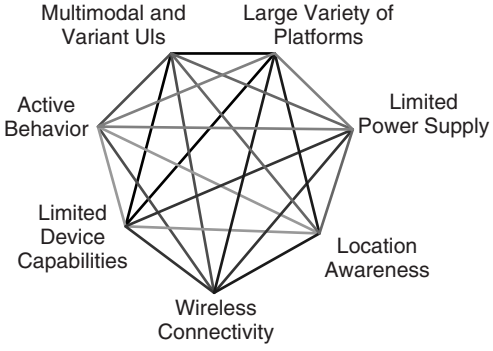

FIGURE 1.3. Dimensions of Mobility.

1.2.1 Location

A mobile device is not always at the same place: Its location is constantly changing. The changing location of the mobile device and the mobile application presents the designers of the device and software applications with great difficulties. However, it also presents us with an opportunity of using the location and the change in location to enhance the application. These challenges and opportunities can be divided into two general categories: *localization* and *location sensitivity*.

Localization is the mere ability of the architecture of the mobile application to accommodate logic that allows the selection of different business logic, level of work flow, and interfaces based on a given set of location information commonly referred to as locales. Localization is not exclusive to mobile applications but takes a much more prominent role in mobile applications. Localization is often required in stationary applications where users at different geographical locations access a centralized system. For example, some point-of-sale (POS) systems and e-commerce Web sites are able to take into account the different taxation rules depending on the locale of the sale and the location of the purchase. Whereas localization is something that stationary applications can have, location sensitivity is something fairly exclusive to mobile applications.

Location sensitivity is the ability of the device and the software application to first obtain location information while being used and then to take advantage of this location information in offering features and functionality. Location sensitivity may include more than just the absolute location of the device (if there is such a thing as absolute location—Einstein must be rolling in his grave now!). It may also include the location of the device relative to some starting point or a fixed point, some history of past locations, and a variety of calculated values that may be found from the location and the time such as speed and acceleration.

There are a variety of methods for collecting and using the location of the user and the device. The user may simply be prompted for his or her location, but this wouldn't make a very user-friendly application. Imagine a system that can only give you directions to where you want to go if you know where you are: It will be useful often, but occasionally, you won't know where you are or it would be too difficult to figure out your location. The device may be reset for a relative location if it has the ability to sense motion and can keep track of the change of location

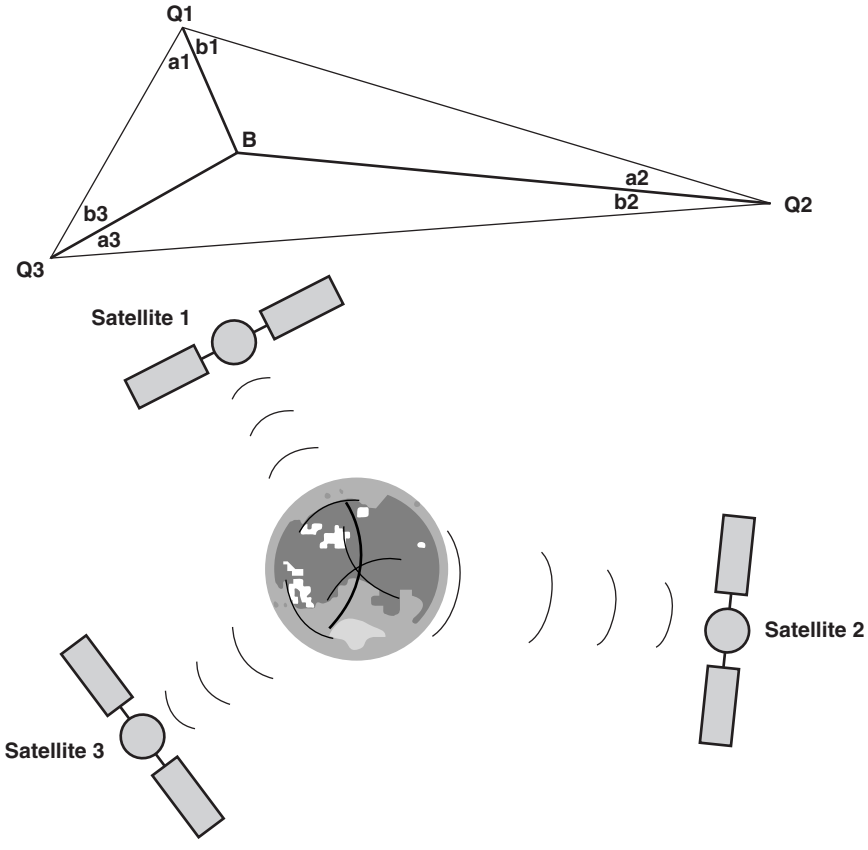

FIGURE 1.4. Determining Position Based on Triangulation.

for some period of time after this reset. Most location-sensing technologies (the particulars of which will be discussed in Chapter 12) use one or more of three categories of techniques: *triangulation*, *proximity*, and *scene analysis* [Hightower and Borriello 2001].

Triangulation (Figure 1.4) relies on age-old geometric methods that allow calculation of the location of a point that lies in the middle of three other points whose exact locations are known. If the distance to each one of the three points is known, we can use geometric techniques to calculate the exact location of the unknown point. *Proximity*-based methods measure the relative position of the unknown point to some known point. *Scene analysis* relies on image processing and topographical techniques to calculate the location of the unknown point based on a view of the unknown point from a known point.

The most well known location sensing system today is GPS. GPS-enabled devices can obtain latitude and longitude with accuracy of about 1–5 m. GPS has its roots in the military; until recently, the military placed restrictions on the accuracy of GPS available for public use. Most of these restrictions have now been lifted. GPS devices use triangulation techniques by triangulating data points from the satellite constellation that covers the entire surface of the earth. If a device does not have GPS capabilities but uses a cellular network for wireless connectivity,

signal strength and triangulation or other methods can be used to come up with some approximate location information, depending on the cellular network.

Regardless of how location information is obtained, it is one of the major differences between mobile and stationary systems. Location information can be to mobile applications what depth can be to two-dimensional pictures; it can give us an entirely new tool to automate tasks. An example of a stand-alone mobile software application that uses location information could be one that keeps track of the route that a user drives from home to work every day without the user entering the route manually; this could then be used to tell the user which route is the fastest way to get to work on a particular day or which route may result in the least amount of gas consumed. An example of a wirelessly networked mobile application taking advantage of location could be one that shows a field service worker where to go next, once he or she is finished with a task at one site, based on the requests for work in the queue and the location of the field service worker. It should be noted that acquiring position information requires connectivity to some network-based infrastructure. This infrastructure is typically isolated from the other network-based application infrastructures. Therefore, when we say stand-alone, we mean an application that may use some specific network-based infrastructure, such as GPS, for obtaining location information but is not connected to any other networks as a part of a distributed or network-based application.

Location information promises to be one of the biggest drivers of mobile applications as it allows for the introduction of new business models and fundamentally new methods of adding productivity to business systems.

1.2.2 Quality of Service

Whether wired or wireless connectivity is used, mobility means loss of network connectivity reliability. Moving from one physical location to another creates physical barriers that nearly guarantee some disconnected time from the network. If a mobile application is used on a wired mobile system, the mobile system must be disconnected between the times when it is connected to the wired docking ports to be moved. Of course, it is always a question whether a docking port is available when required let alone the quality and type of the available network connectivity at that docking port. In the case of wireless network connectivity, physical conditions can significantly affect the quality of service (QOS). For example, bad weather, solar flares, and a variety of other climate-related conditions can negatively affect the (QOS). This unreliability in network connectivity has given rise to the QOS field and has led to a slew of accompanying products. QOS tools and products are typically used to quantify and qualify the reliability, or unreliability, of the connectivity to the network and are mostly used by network operators. Network operators control the physical layer of the network and provide the facilities, such as Internet Protocol (IP), for software application connectivity.

Usually, the QOS tools, run by the network operators, provide information such as available bandwidth, risk of connectivity loss, and statistical measurements that allow software applications to make smart computing decisions. The key to designing and implementing mobile applications is that network connectivity and QOS need to be taken into account with an expanded scope. Most software