Challenges of Mobile Computing

Abhishek Kumar Gupta
Department of Computer Science & Engineering
College of Engineering Roorkee, Roorkee-247667, India

Abstract- The need for "information anywhere anytime" has been a driving force for the increasing growth in Web and Internet technology, wireless communication, and portable computing devices. The field of mobile computing is the merger of these advances in computing and communication with the aim of providing seamless computing environment for mobile users. Increasingly, we are dependent on information that is available only by accessing a network. This dependency implies that users would need access to this information even while on the move. Hence, mobility is additional parameter that needs to be considered in the design of networks, network protocols and information services. Mobility implies that networks need to cope with moving users. During a single session, users may connect from different network attachment points or use different networks or even more than one network simultaneously. Mobile computing environments are characterized by severe resources constraints and frequent changes in operating conditions. Mobile computing is gaining wide acceptance due to the rapid enhancement in wireless communication technologies. This has led to an increase in the demand for mobile information access. 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 conventions and public events, may overload network capacity. As the use and functionality of mobile device increase, IT organizations face several serious challenges in developing strategies for integrating and managing them, for maximizing their effectiveness.

Keywords- Bandwidth, Error-Control Protocol, Mobile Computing, Wireless Communication.

I. INTRODUCTION

The combination of wireless communication infrastructure and portable computing devices has laid the foundation for a new network computing paradigm, called mobile computing, which allows the users access information and collaborate with others while on the move [1]. Wireless mobile networks are typically characterized by severe constraints on resources, such as bandwidth and battery power, and by rapid fluctuations in availability of these resources; this makes it difficult for the system software to provide guaranteed quality-of-service at levels required by many distributed and collaborative applications. Also, due to mobility of the clients or the users, users may be disconnected from the network often and the users may also voluntarily switch off to save battery power; thus, management of this disconnection is a critical issue in designing mobile networks. Further, user mobility adds a new dimension to be distributed operating systems which has implications for specification, design, verification and implementation of both system and application software [2]. A challenging issue is to determine the interface and the guarantees that the system software must provide to the developers of both location-independent and location-dependent applications on mobile networks [3]. This has resulted in research on adaptive applications and system software which can gracefully respond to changes in operating conditions [4].

There are several articles which have identified the fundamental challenges in mobile computing. Mobile systems are (i) resource poor (ii) less secure (iii) have poor connectivity to the wired infrastructure and (iv) have less energy since they are powered by battery. In order to deal with these characteristics the mobile systems should employ dynamic adaptation schemes. One of the implications is that the solutions developed for mobile systems should be interoperable since as mobile clients move one domain to another they should be able to operate in the new domain. The mission of mobile computing is to allow users to access any information using any device over any network at anytime. However, wireless networks and mobile devices are introducing new requirements to software engineers due to their limited resources. A new architecture design model for software development for mobile devices needs to be adopted. The new model has to address the mobile computing constraints to success in supporting mobile information access.

In general, a mobile-computing network may be characterized as follows. It consists of multiple mobile agents that require access to (i) information generated at multiple geographically dispersed sites and (ii) computing engines to execute their decisions. It may be include one or more stationary agents that perform information acquisition and propagation to the mobile agents. While a static interconnection network may link the stationary agents, a dynamic interconnection network will connect the mobile agents to the stationary nodes. The mobile nodes may connect to specific nodes asynchronously, i.e., at irregular intervals of time, to acquire information, and following completion they will disconnect. The use of the term connection in this context refers to the transport layer in the ISO-OSI terminology. The underlying physical layer, however, is at liberty to utilize either wired or wireless transmission. The mobile and stationary agents are located at geographically dispersed sites. While both stationary and mobile nodes may have computing and communication needs, the relative weights and frequency are problem-specific. In addition, the system must be designed to accommodate evolutionary growth. That is, the system must continue to function and deliver relatively undiminished performance as the cumulative number of stationary and mobile entities increases with time.

II. MOBILE COMPUTING CONSTRAINTS

Mobile computing environment has three main components which affect their performance. These are:

A. Mobile devices constraints

Due to their limited size, mobile devices will continue to have limited resources such as processor speed, memory size, display size and resolution. Hence, designers of mobile applications are faced with a real challenge on how to meet these constraints while satisfying users demand of fast responsive mobile computing environment. Some mobile devices may have different input device such as microphone or a pen point screen to interact with users. These limited input devices must be addressed carefully when designing a mobile user interface. Power limitations of mobile devices are also to be addressed in terms of what an application should do when it recognizes a very limited power to use.

B. Networks Constraints

Wireless networks will continue to have limited bandwidth, high latency and frequent disconnection due to power limitations, available spectrum and mobility. The most critical aspect of these is the bandwidth. As bandwidth increases, power consumption also increases which shorten the battery life of mobile devices. Hence, energy restrictions of mobile devices will limit the effectiveness of data throughput to and from the device even if wireless networks connections deliver stable high bandwidth. Therefore, data access technique must be designed for users to overcome these limitations of bandwidth, latency, disconnected operation and still satisfy user's expectations [5].

C. Mobility Constraints

Mobility is about moving around. This movement can be in one of the following spaces with respect to users.

1) Physical space movement: in this space is when a user changes his/her physical location. This can affect network connection (disconnection, Poor connectivity) while the user is working. Network connection has to adapt to this behavior of user movement by re-connecting user with respect to new location. Data and file operations have also to adapt by supporting disconnection file operations as in the Coda file system [6].

2) Information space: consists of a large number of applications and data files scatter in the Internet. Movement here corresponds to selecting an application or a data file. Application and their data have to adapt to user's operating environment to perform their task efficiently. One way to accomplish this is by caching of data on the local machine as being done in the HTTP/1.1 protocol to reduce redundant www information traffic and hence improve network utilization.

3) Connection space: is the huge network of links that connect between various computer platforms. Movement in this space corresponds to selecting route between links and selecting a specific type platform. Adaptation in this space can be achieved by avoiding congested links and selecting shortest path between two points when delivering information.

III. CHALLENGES OF MOBILE COMPUTING

The need for mobile computing leads to design challenges in several areas.

A. Disconnection

Today's computer systems often depend heavily on a network and may cease to function during network failures. Foe example, distributed file systems may lock up waiting for other servers, and applications process 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 acknowledgement. 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. 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. Of course, not all network disconnections can be masked. In these cases, good user interfaces can help by providing feedback about which operations are unavailable because of network disconnections.

B. Low Bandwidth

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. 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 [7].

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 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.

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.

C. High bandwidth variability

Mobile computing designs also contend with much greater variation in network bandwidth than do traditional designs. Bandwidth can shift one to four orders of magnitude, depending on whether the system is plugged in or using wireless access. 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 senses for different applications.

D. 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 the country. This heterogeneity makes mobile networking more complex than traditional networking [8].

E. 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 [9]. Secure communication over insecure channels is accomplished by encryption, which can be done in software. 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.

IV. COMMUNICATION CONCEPTS ON MOBILITY

The ability to change location 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. 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.

A. Address migration

As people move, their mobile computers will use different network access points, or "addresses." Today's networking is not designated 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 address. 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, central services, home base and forwarding pointers. These are the building blocks of the current proposals for "mobile-IP" schemes [10].

1) 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 address may suffice is adjacent cells are known.

2) 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.

3) Home base: 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 with much lower volatility.

4) 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. Hence, forwarding pointers are often used only to speed the common case, and another method is used to allow reclamation of old pointers. The forwarding pointer method requires an active entity at the old address to receive and forward messages.

B. Location-dependent information

Because traditional computers do not move, information that depends on location such as 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.

1) 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. 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 [11].

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

V. 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 wrist watch: 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.

A. 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 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.

B. 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. 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.

CONCLUSION

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.

REFERENCES

[1] G.S. Blair, “The role of open implementation and reflection supporting mobile applications”, in Proceedings of the 9th International Workshop on Database and Expert Systems Applications, 1998.

[2] J. Flinn and M. Satyanarayanan, “Powerscope: A tool for profiling the energy usage of mobile applications”, in Proceedings of the Second IEEE Workshop on Mobile Computing Systems and Applications, New Orleans LA, February, 1999.

[3] J. R. Lorch and A. J. Smith. “Software strategies for portable computer energy management”, IEEE Personnel Communication, 5(3): 60-73, June 1998.

[4] Adnan Al-bar and Ian Wakeman, “A Survey of Adaptive Applications in Mobile Computing”, in Proceedings of the Second IEEE Workshop on Mobile Computing Systems and Applications, 2001.

[5] G.H. Forman and John Zahorjan, “The Challenges of Mobile Computing”, IEEE Computer, 27(4): 38-47, April 1994.

[6] S.K.S. Gupta and P.K. Srimani, “Experience in Teaching a Graduate Course in Mobile Computing”, in Proceedings of the IEEE Workshop on Mobile Computing Systems, 2000.

[7] M. Satyanarayanan, “Fundamentals Challenges in Mobile Computing”, in ACM Symposium on Principles of Distributed Computing, 1995.

[8] P.J. McCann and G.C. Roman, “Compositional Programming Abstractions for Mobile Computing”, IEEE Trans. on Software Engineering, 24(2): 97-110, February 1998.

[9] Anup Talukdar, et al., “On accommodating mobile hosts in an Integrated Services Packet Network”, IEEE Infocom, April 1997, pp.1048-1055.

[10] Jun Wei, et al., “Challenges of Communication in Mobile Computing”, IEEE Computer, 1999.

[11] G.P. Picco, “Lime: Linda Meets Mobility”, ICSE98 International Workshop on Computing and Communication in the Presence of Mobility, April 25, 1998.