DISTRIBUTED CACHE FOR A WIRELESS COMMUNICATION SYSTEM

Background of the Invention

Field of the Invention

The invention relates generally to wireless communication systems and, more particularly, to digital data transfer in a wireless communication system.

Description of the Related Art

Digital data communication systems are becoming more and more pervasive. For example, the Internet is commonly used to transfer digital data in the form of e-mail messages, web pages as well as audio, graphic and video information in digital format between many users and servers. Figure 1 is a block diagram showing a typical Internet digital data system. The Internet "cloud" 10 interconnects a large number of users and content servers. In order to gain access to the Internet cloud 10, an Internet service provider 12 acts as a gateway between the Internet cloud 10 and a group of users 14A-14N.

In the normal course of "surfing the web", a user accesses a series of web pages stored within a variety of content servers 8A-8N coupled to the Internet cloud 10. In general, a user requests a web page via browser software. The browser software retrieves the web page using a group of protocols defining the Internet. For example, the browser software uses the HyperText Transfer Protocol (HTTP) on top of Transmission Control Protocol/Internet Protocol (TCP/IP) to retrieve the web page. During the retrieval process, the user must wait as the browser software first makes a request for a Domain Name System (DNS) to find the appropriate content server, and then individually requests and receives the set of objects that make up the web page.

Each HTTP request to the content server opens a new TCP connection. After the connection is established, the user's browser software sends the HTTP request for the specific hypertext markup language (HTML) object. The wait experienced by the user is caused by the round-trip delay from the HTTP latency as all the HTML objects which compose the hypertext markup language (HTML) file are retrieved from the Internet.

The content server can be located many thousands of miles away from the user. Over such great distances, even if the messages were able to travel the speed of light, a significant latency would be accumulated by the numerous round-trip delays associated with retrieving each object.

The HTML file is digital data file that provides information to the browser such as display information. Generally, the HTML file comprises a set of embedded HTML objects. For example, the embedded objects may include Java applets, JPEG or GIF graphical objects, video files or sound clips. For example, a typical web page, such as the YAHOO!® home page, is an HTML file which designates many embedded objects such as advertisement banners, headline banners, the YAHOO!® logo and linking soft buttons. As an entire web page is retrieved, the browsers used by the users 14A-14N individually request each of the embedded objects. Thus, the response time associated with creating a fully displayed web page includes the time required to retrieve both the HTML file and all of the embedded objects referenced therein.

In order to decrease the waiting time of the system as well as decrease the amount of data that is transferred between the Internet 10 and the Internet service provider 12 of Figure 1, the Internet service provider 12 may incorporate a cache 16. Generally, the cache 16 may be embodied as a fast storage buffer or memory that can be accessed by a central processing unit within the Internet service provider 12. The cache 16 can be used to store HTML objects and other files such as graphic files or sound clips that are commonly requested by the users 14A-14N. For example, the YAHOO!® home page HTML file is likely to be accessed by a large number of the users 14A-14N on a daily basis. Therefore, in order to avoid downloading the YAHOO!® HTML file with its many HTML embedded objects multiple times during the day, the common objects which constitute the page can be stored in the cache 16 and retrieved from the cache 16 by the Internet service provider 12 when requested by one of the users 14A-14N. Because it is faster for the Internet service provider 12 to retrieve the files from the cache 16 than through the Internet 10 from the content servers 8, the waiting time perceived by the user is decreased.

Several factors must be determined when designing an efficient cache system. For example, because the cache 16 has a finite memory storage capacity, the Internet service provider 12 must determine which HTML objects are most likely to be requested by the users 14A-14N. Therefore, usage pattern algorithms have been developed which determine which HTML objects should be stored in the cache 16 based upon the usage pattern of the users 14A-14N. In addition, the information in the cache 16 must be updated at regular time intervals. For example, a web page which provides stock quotes at approximately real time must be updated every several seconds. An advertising banner, such as might be shown on the YAHOO!® homepage, may be updated every hour. Other information on the YAHOO!® homepage, such as the morning headlines, may be updated once or twice a day. Other objects, such as the YAHOO!® logo, may remain valid for much longer periods of time. The usage pattern algorithm must also determine which HTML objects are worthy of caching and which are updated with such frequency as to be unworthy. In addition, the usage pattern algorithm can be used to determine when a HTML object stored in the cache 16 should be deleted and retrieved once again from the Internet 10 in order to update the file.

Based upon these factors, using an efficient usage pattern algorithm, as well as a manageable sized cache, it is reasonable to expect that approximately 60% of the data requested by the users 14A-14N can be stored within the cache 16. Even if the storage capacity of the cache 16 is greatly increased, it is difficult to achieve a caching rate at greater than 60% based upon modern Internet usage patterns.

The Internet service provider 12 can be coupled to the users 14A-14N using a variety of well-known techniques. For example, a copper line such as a standard, directly coupled plain old telephone service (POTS) can be used to connect the Internet service provider 12 to the remote unit 14A. Also cable modems and other digital subscriber lines (DSL) have been developed in order to increase the data rate of the connection between the Internet service provider and the user.

Using modern wireless communication techniques, the Internet service provider 12 can be coupled to one or more of the users 14A-14N using a wireless link rather than a wired line link. Typically, the finite capacity of the wireless link as well as the limited transmission rate of the wireless link can cause additional delays in the

transmission of data between the Internet service provider and the user. In such a system, the benefits of reducing latencies in other areas of the system becomes more pronounced as cumulative delays increase the response time perceived by the users.

Figure 2 is a block diagram of a prior art digital data system providing Internet connection over a wireless link. In Figure 2, the system is configured as a terrestrial, system. A series of radio base stations 20A-20N are distributed throughout a geographic area where wireless Internet access is provided by the internet service provider 12. Each radio base station 20A-20N provides wireless communications to and from remote users within a corresponding physical coverage area. For example, in Figure 2, users 14A-14N are shown to be within the coverage area of the base station 20A. Thus, when a remote unit requests digital data, such as a web page, a request for the web page is passed from the user 14 to the associated base station 20A. The base station 20A passes the request to the Internet service provider 12 incurring yet additional delays. The Internet service provider 12 provides the requested information from the cache 16 or through the Internet cloud 10 from the content servers 8A – 8N, if the HTML objects are not available within the cache 16. The reverse process carries the objects back to the user, incurring yet additional delays. As noted above, the cumulative delay associated with transmission of the request over the wireless link as well as the other delays associated with response to the request can become intolerably high to the end-user.

Therefore, there has been a long felt need in the art to provide a means and method for providing efficient digital data access in a wireless communication system.

Summary of the Invention

In a wireless system comprising a set of limited physical coverage areas, a base station provides service over a wireless link to a plurality of users within one of the limited physical coverage areas. The base station is coupled to an Internet service provider via a backhaul. The Internet service provider is coupled to many content servers via a digital data network. Each base station has an associated cache. The cache is used to store files and objects which are frequently requested by the users within the limited physical coverage area associated with the corresponding base station. When a user sends a message comprising a request for a file or object to the base station, the base station parses the message. If the requested file or object is available from the cache, the base station forwards the file or object to the user from the cache. In this way, the request need not be sent over the backhaul or over the digital data network, thereby greatly improving the response time of the wireless system.

Brief Description of the Drawings

The features, objectives, and advantages of the invention will become more apparent from the detailed description set forth below when taken in conjunction with the drawings wherein like parts are identified with like reference numerals throughout, and wherein:

Figure 1 is a block diagram showing a typical Internet digital data system.

Figure 2 is a block diagram of a prior art digital data system providing Internet connection over a wireless link.

Figure 3 is a block diagram of a digital data system providing Internet connection over a wireless link according to the invention.

Figure 4 is a flow chart illustrating operation in accordance with one embodiment invention.

Detailed Description of the Invention

Figure 3 is a block diagram of one embodiment of the invention. In Figure 3, a central controller, such as an Internet service provider 112, is coupled to a digital data network, such as an Internet cloud 100 to provide connection to a group of content servers 98A – 98N. In addition, the Internet service provider 112 is coupled to a series of base stations 114A-114N. Generally, the Internet service provider 112 is coupled to the series of base stations 114A-114N via a base station controller as well as a series of high capacity digital data links.

Each base station 114A-114N provides service to a corresponding physical coverage area through the use of distributed antenna sites. In one embodiment, the base stations are terrestrial base stations such as those found in a typical mobile wireless or stationary wireless telecommunications system. In another embodiment, the base stations may be co-located with one another and may provide service to distinct physical coverage areas by means of satellites, broadband fixed wireless or directional antennas. The users 118A-118N are located within the coverage area of the base station 114A. The users 120A-120N are located in the coverage area of base station 114B. The users 124A-124N are located in the coverage area of the base station 114N.

Each base station 114A-114N has a corresponding cache 116A-116N. The caches 116A–116N may be embodied as fast storage buffers or memories which can be accessed by a router and central processing unit within the corresponding base stations 114A–114N. The cache 116 is used to store information frequently requested by the users associated with the corresponding base station. The base station 114A may comprise a usage pattern algorithm which monitors the usage pattern of the users 118A-118N within its corresponding coverage area in order to determine which files and objects to store within the cache 116A, according to well know techniques. In one embodiment, the cache 116 stores frequently requested HTML files and objects. When a user such as the user 118A requests a web page, the corresponding base station, such as the base station 114A, intercepts and analyzes the series of requests for files and objects which are generated by the user 120. Thus, in contrast to the prior art, the base station 114A does not simply indiscriminately pass the remote unit requests to the Internet service provider but, instead, examines the content of the requests.

If one or more of the requested files or objects are available within the cache 116A, the base station 114A responds by retrieving the requested file or object from the cache 116A and forwarding it to the user 118A. Thus, the request need not be passed over the Internet service provider 112 or to the Internet cloud 100. If any requested file or object which is not available within the cache 116A, the base station 114A passes the corresponding request to the Internet service provider 112, which in turn may pass the request over Internet cloud 100 to one of the content servers 98A –98N. In one embodiment, the Internet service provider 112 also comprises a cache 126 which can be accessed by the Internet service provider 112 and, if a requested object is available in the cache 126, the Internet service provider 112 need not retrieve the file from the Internet cloud 100.

The configuration shown in Figure 3 has many advantages. The connection between the Internet service provider 112 and the base stations 114A-114N is referred to as the backhaul. According to current costing structures, the cost of operating the backhaul is a substantial portion of the cost of providing a wireless network. By retrieving the information from the cache 116A, the base station 114A need not pass a message over the backhaul to the Internet service provider 112. Therefore, the configuration shown in Figure 3 reduces the loading on the backhaul, thus reducing the cost of providing wireless service. In addition, procuring backhaul capacity on a timely basis can limit the ability of a wireless carrier to deploy a wireless system. By reducing the loading on the backhaul, the carrier can operate the system using a lower capacity backhaul, thus, reducing the backhaul capacity required to deploy a system.

In addition, by associating the cache 116A with a limited coverage area, the usage patterns of the users are more likely to be correlated with one another. For example, a typical Internet service provider provides coverage for a rather large geographic region. More particularly, an Internet service provider may provide coverage for an entire county covering many hundreds of square miles comprising several cities, smaller towns, and rural areas. In contrast, a typical terrestrial base station could provide coverage for an approximately 3 – 100 square mile coverage area. When a larger region is broken down into the smaller regions, the probability of a correlation in the requested pages among the users increases. For example, a coverage area which includes a downtown region is more likely to have a high number of users requesting legal, business, financial and tax information than a coverage area servicing a college campus. A coverage area providing coverage to an affluent neighborhood is more likely to have a higher concentration of requests for information concerning an upcoming equestrian event than a coverage area providing service to an industrialized factory campus.

Thus, the usage pattern algorithm which selects the files which are stored within the cache 116 is likely to be able to increase the percentage of requests which can be serviced directly from the cache 116 rather than through the cache 126 or from the content servers 98. Thus, the configuration shown in Figure 3 further reduces the backhaul traffic as well as the latency perceived by the user.

In one embodiment, the base station executes anticipatory caching. Anticipatory caching occurs when a file or object is retrieved and stored in the cache before the users request it. For example, many users access the newspaper in the morning. Anticipatory caching can be used to retrieve the files and embedded objects associated with the newspapers as they are the released by the publishers in the early morning when the demand on the wireless system is minimum. In this way, when the users begin to access the papers in early morning, at least some of the files and objects are available from the cache. Caching according to the invention also improves the efficiencies achieved by anticipatory caching. Residents of a collage campus may be more likely to access a local or campus-wide newspaper rather than a national newspaper. Business users are more likely to access business and financial newspapers such as the Wall Street Journal. These regional differences are accommodated by the local caching system of the invention.

Another advantage associated with the configuration shown in Figure 3 is that the wireless link also operates more efficiently than the prior art. If the wireless link is configured as a circuit switch system, when a user 118 makes a request to the base station, a resource is dedicated to that remote unit until a response is received. By decreasing the latency associated with the response to the request, the wireless link resources consumed by a remote unit user to make a request and receive a response can be reduced, thus, increasing the number of users which can be serviced by the system at any one time.

Using modern wireless communication techniques, the rate at which data can be transferred over the wireless link between the user 118 and the base station 114 can be faster than the rate at which the Internet cloud 100 can provide information. Thus, by providing information directly from the cache 116, the rate at which data is passed over the wireless link need not be artificially lowered in order to accommodate the slower transmission rate of the Internet cloud 100. This in turn can increase overall system capacity so that the system can accommodate more users.

Figure 4 is a flow chart illustrating operation in accordance with one embodiment of the invention. In block 140, the user requests a file or object. The user forwards the request in a message to the base station over the wireless link. In block 142, the base station intercepts and parses the request. For example, the base station examines the contents of the request in the same manner as the Internet service provider, according to well-known techniques to identify the message. The base station also determines if the requested file or object is available from the base station cache. As noted above, in a preferred environment, the cache is co-located with the base station or is associated specifically with the base station. In block 144, if the file or object is available from the cache, the base station forwards the file or object to the user from the cache. The user receives the file or object in block 146. Thus, flow from block 140 to block 142 to block 144 and back to block 146 provides a rapid response to the user's request without utilizing backhaul resources, Internet service provider resources or digital data cloud resources.

If the requested file or object is not available at the cache at the base station, in block 144, the base station forwards the request to the Internet service provider over the backhaul as shown by the dashed line in Figure 4. In response, in block 148, the Internet service provider determines if the requested file or object is available from a cache associated with the Internet service provider. If so, in block 150, the Internet service provider forwards the file or object to the base station as shown by the dashed line in Figure 4. In block 152, the base station forwards the file or object to the remote unit. As shown by the dashed line, the user receives the file or object in block 154.

If the file or object is not available in the cache associated with the Internet service provider, in block 150, the Internet service provider requests the file or object from a content server via the internet cloud according to well known techniques as indicated by the dotted line in Figure 4. In block 156, the Internet service provider receives the file or object and forwards it to the base station as shown by the dotted line in Figure 4. The base station receives the file or object and in block 158 forwards the file or object to the user. In block 160, the user receives the file or object.

In parallel with the operation shown in Figure 4, a usage pattern algorithm is executed at the base station. For example, in block 142, in which the base station intercepts and parses the request, the base station may also

forward a designation of the requested file or object to the usage pattern algorithm for use by the usage pattern algorithm in determining the contents of the associated cache.

In one embodiment, a single cache is associated with more than one antenna site. For example, two base stations may share a cache. In another embodiment, some of the base stations comprise distinct sectors that provide service to a subsection of the limited physical coverage area associated with the base station. In such a case, a single cache may be provided for use by users in each sector, or one or more caches may be provided for use by users in a subset of the total number of sectors. In yet another embodiment, only a subset of the base stations within a wireless system have a local cache.

The invention may be embodied in other specific forms without departing from its spirit or central characteristics. The described embodiment is to be considered in all respects only as illustrative and not restrictive and the scope of the invention is, therefore, indicated by the appended claims rather than by the foregoing description. All changes which come within the meaning and range of equivalency of the claims are to be embraced within their scope.

WHAT IS CLAIMED IS:

1. A method of wireless data transfer comprising:

receiving a message over a wireless link at a base station;

identifying a requested file from said message at said base station;

determining whether said requested file is available from a memory associated with said base station;

forwarding said requested file from said base station to a requesting remote unit if said requested file is available within said memory; and

forwarding said message to a central controller if said requested file is not available from said memory.

2. The method of Claim 1, further comprising selecting files for storage in said memory based upon usage patterns of requested files received by said base station.

3. The method of Claim 1, further comprising:

receiving said message from said base station at said central controller;

identifying a requested file from said message at said central controller;

determining whether said requested file is available from a memory associated with said central controller;

forwarding said requested file to said base station if said requested file is available within said memory associated with said central controller; and

forwarding said message to a network if said requested file is not available from said memory associated with said central controller.

4. The method of Claim 1, wherein said base station provides service to a limited portion of a wireless system.

5. A wireless system comprising:

a plurality of base stations, each of said plurality of base stations servicing remote units located within a corresponding limited physical coverage area;

a central controller coupled to each of said plurality of base stations;

a network coupled to said central controller and coupled to each of said plurality of base stations via said central controller; and

a cache associated with a first base station of said plurality of base stations and storing digital data retrieved over said network;

wherein when said first base station receives a message from a wireless user requesting a file and said file is stored within said cache, said first base station identifies said file from said message, retrieves said file from said cache and forwards said file to said wireless user.

* C L M S *

6. The wireless system of Claim 5 wherein each of said plurality of base stations is located within said corresponding limited physical coverage area.

7. The wireless system of Claim 5 wherein each of said plurality of base stations provides service to said limited physical coverage area via the use of a satellite link.

8. The wireless system of Claim 5 further comprising a usage pattern algorthm for identifyin said digital data to be stored in said cache.

9. The wireless system of Claim 5 further comprising an anticipatory caching algorthm for identifying said digital data to be stored in said cache.

10. The wireless system of Claim 5 wherein said file is an embedded Internet object.

11. A system for wireless data transfer comprising:

means for receiving a message over a wireless link at a base station;

means for identifying a requested file from said message at said base station;

means for determining whether said requested file is available from a memory associated with said base station;

means for forwarding said requested file from said base station to a requesting remote unit if said requested file is available within said memory; and

means for forwarding said message to a central controller if said requested file is not available from said memory.

12. The system of Claim 11, further comprising selecting files for storage in said memory based upon usage patterns of requested files received by said base station.

13. The method of Claim 11, further comprising:

means for receiving said message from said base station at said central controller;

means for identifying a requested file from said message at said central controller;

means for determining whether said requested file is available from a memory associated with said central controller;

means for forwarding said requested file to said base station if said requested file is available within said memory associated with said central controller; and

means for forwarding said message to a network if said requested file is not available from said memory associated with said central controller.

