

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
9 February 2012 (09.02.2012)

PCT

(10) International Publication Number
WO 2012/018430 A1

- (51) International Patent Classification:
G06F 15/16 (2006.01) H04W 28/02 (2009.01)
H04W 88/02 (2009.01)
- (21) International Application Number:
PCT/US2011/037932
- (22) International Filing Date:
25 May 2011 (25.05.2011)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
61/367,870 26 July 2010 (26.07.2010) US
61/367,871 26 July 2010 (26.07.2010) US
61/408,846 1 November 2010 (01.11.2010) US
61/408,820 1 November 2010 (01.11.2010) US
61/408,826 1 November 2010 (01.11.2010) US
61/408,839 1 November 2010 (01.11.2010) US
61/408,829 1 November 2010 (01.11.2010) US
61/408,854 1 November 2010 (01.11.2010) US
61/408,858 1 November 2010 (01.11.2010) US
61/416,020 22 November 2010 (22.11.2010) US
61/416,033 22 November 2010 (22.11.2010) US
61/430,828 7 January 2011 (07.01.2011) US
- (72) Inventors; and
(75) Inventors/Applicants (for US only): BACKHOLM, Ari [FI/FI]; Koivistontie 14 E, FIN-02140 Espoo (FI). LUNA, Michael [US/US]; 2100 Seaport Boulevard, Redwood City, CA 94063 (US). YLINEN, Heikki [FI/FI]; Numersinkatu 11 A 8, FI-02600 Espoo (FI).
- (74) Agents: FU, Yenyun et al.; Perkins Coie LLP, P.O. Box 1208, Seattle, WA 98111-1208 (US).
- (81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.
- (84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK,
- (71) Applicant (for all designated States except US): SEVEN NETWORKS, INC.; 2100 Seaport Boulevard, Redwood City, California 94063, United States of America (US).

[Continued on next page]

(54) Title: MOBILE NETWORK TRAFFIC COORDINATION ACROSS MULTIPLE APPLICATIONS

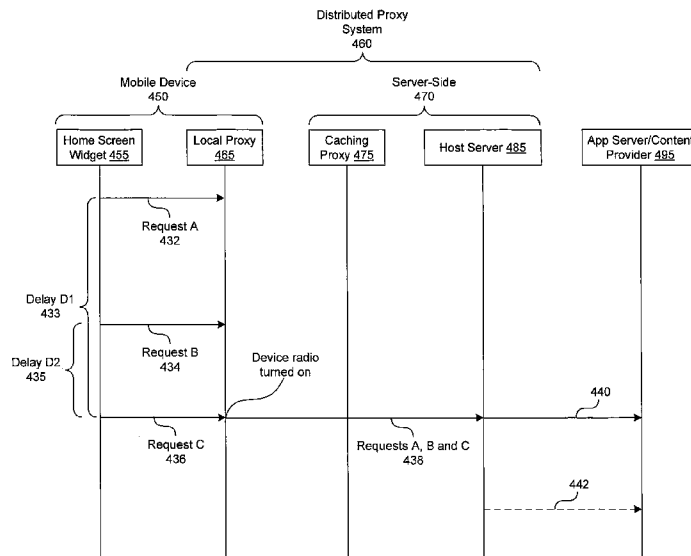


FIG. 4B

(57) Abstract: Systems and methods for mobile network traffic coordination across multiple applications are disclosed. In one aspect, embodiments of the present disclosure include a distributed proxy and cache system, including, a local proxy on a mobile device for intercepting a data request made via a mobile device, and a proxy server coupled to the mobile device and a content server to which the data request is directed. One embodiment includes, delaying transfer of a first data transfer request initiated by a first application until another data transfer request initiated by a second application is detected on the mobile device and transferring, the first data transfer request of the first application and the other data transfer request of the second application a single transfer operation over the network.

WO 2012/018430 A1



SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, —
GW, ML, MR, NE, SN, TD, TG).

*before the expiration of the time limit for amending the
claims and to be republished in the event of receipt of
amendments (Rule 48.2(h))*

Published:

— *with international search report (Art. 21(3))*

**MOBILE NETWORK TRAFFIC COORDINATION ACROSS MULTIPLE
APPLICATIONS**

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Patent Application No. 61/367,871 entitled “CONSERVING POWER CONSUMPTION IN APPLICATIONS WITH NETWORK INITIATED DATA TRANSFER FUNCTIONALITY”, which was filed on July 26, 2010, U.S. Provisional Patent Application No. 61/367,870 entitled “MANAGING AND IMPROVING NETWORK RESOURCE UTILIZATION, PERFORMANCE AND OPTIMIZING TRAFFIC IN WIRE LINE AND WIRELESS NETWORKS WITH MOBILE CLIENTS”, which was filed on July 26, 2010, U.S. Provisional Patent Application No. 61/408,858 entitled “CROSS APPLICATION TRAFFIC COORDINATION”, which was filed on November 1, 2010, U.S. Provisional Patent Application No. 61/408,839 entitled “ACTIVITY SESSION AS METHOD OF OPTIMIZING NETWORK RESOURCE USE”, which was filed on November 1, 2010, U.S. Provisional Patent Application No. 61/408,829 entitled “DISTRIBUTED POLICY MANAGEMENT”, which was filed on November 1, 2010, U.S. Provisional Patent Application No. 61/408,846 entitled “INTELLIGENT CACHE MANAGEMENT IN CONGESTED WIRELESS NETWORKS”, which was filed on November 1, 2010, U.S. Provisional Patent Application No. 61/408,854 entitled “INTELLIGENT MANAGEMENT OF NON-CACHABLE CONTENT IN WIRELESS NETWORKS”, which was filed on November 1, 2010, U.S. Provisional Patent Application No. 61/408,826 entitled “ONE WAY INTELLIGENT HEARTBEAT”, which was filed on November 1, 2010, U.S. Provisional Patent Application No. 61/408,820 entitled “TRAFFIC CATEGORIZATION AND POLICY DRIVING RADIO STATE”, which was filed on November 1, 2010, U.S. Provisional Patent Application No. 61/416,020 entitled “ALIGNING BURSTS FROM SERVER TO CLIENT”, which was filed on November 22, 2010, U.S. Provisional Patent Application No. 61/416,033 entitled “POLLING INTERVAL FUNCTIONS”, which was filed on November 22, 2010, U.S. Provisional Patent Application No. 61/430,828 entitled “DOMAIN NAME SYSTEM WITH NETWORK TRAFFIC HARMONIZATION”, which was filed on January 7, 2011, the contents of which are all incorporated by reference herein.

BACKGROUND

[0002] When WCDMA was specified, there was little attention to requirements posed by applications whose functions are based on actions initiated by the network, in contrast to functions initiated by the user or by the device. Such applications include, for example, push email, instant messaging, visual voicemail and voice and video telephony, and others. Such applications typically require an always-on IP connection and frequent transmit of small bits of data. WCDMA networks are designed and optimized for high-throughput of large amounts of data, not for applications that require frequent, but low-throughput and/or small amounts of data. Each transaction puts the mobile device radio in a high power mode for considerable length of time – typically between 15-30 seconds. As the high power mode can consume as much as 100x the power as an idle mode, these network-initiated applications quickly drain battery in WCDMA networks. The issue has been exacerbated by the rapid increase of popularity of applications with network-initiated functionalities, such as push email.

[0003] Lack of proper support has prompted a number of vendors to provide documents to guide their operator partners and independent software vendors to configure their networks and applications to perform better in WCDMA networks. This guidance focuses on: configuring networks to go to stay on high-power radio mode as short as possible and making periodic keep alive messages that are used to maintain an always-on TCP/IP connection as infrequent as possible. Such solutions typically assume lack of coordination between the user, the application and the network.

[0004] Furthermore, application protocols may provide long-lived connections that allow servers to push updated data to a mobile device without the need of the client to periodically re-establish the connection or to periodically query for changes. However, the mobile device needs to be sure that the connection remains usable by periodically sending some data, often called a keep-alive message, to the server and making sure the server is receiving this data. While the amount of data sent for a single keep-alive is not a lot and the keep-alive interval for an individual application is not too short, the cumulative effect of multiple applications performing this individually will amount to small pieces of data being sent very frequently. Frequently sending bursts of data in a wireless network

also result in high battery consumption due to the constant need of powering/re-powering the radio module.

BRIEF DESCRIPTION OF THE DRAWINGS

[0005] FIG. 1A illustrates an example diagram of a system where a host server facilitates management of traffic between client devices and an application server or content provider in a wireless network for resource conservation.

[0006] FIG. 1B illustrates an example diagram of a proxy and cache system distributed between the host server and device which facilitates network traffic management between a device and an application server/content provider for resource conservation.

[0007] FIG. 2A depicts a block diagram illustrating an example of client-side components in a distributed proxy and cache system residing on a mobile device that manages traffic in a wireless network for resource conservation.

[0008] FIG. 2B depicts a block diagram illustrating another example of components in the application behavior detector and the traffic shaping engine in the local proxy on the client-side of the distributed proxy system shown in the example of FIG. 2A.

[0009] FIG. 3 depicts a block diagram illustrating an example of server-side components in a distributed proxy and cache system that manages traffic in a wireless network for resource conservation.

[0010] FIG. 4A depicts a timing diagram showing how data requests from a mobile device to an application server/content provider in a wireless network can be coordinated by a distributed proxy system in a manner such that network and battery resources are conserved through using content caching and monitoring performed by the distributed proxy system.

[0011] FIG. 4B depicts a timing diagram showing how data requests from a mobile device to an application server/content provider in a wireless network can be aligned at the local proxy in the distributed proxy system to optimize network and radio use.

[0012] FIG. 5 depicts a diagram showing one example process for implementing a hybrid IP and SMS power saving mode on a mobile device using a distributed proxy and cache system (e.g., such as the distributed system shown in the example of FIG. 1B).

[0013] FIG. 6 depicts a flow chart illustrating example selection processes through which data transfer requests of multiple applications can be coordinated into a single transfer operation.

[0014] FIG. 7 depicts an example of triggering events that would cause a data request to be transferred without alignment with another data request.

[0015] FIG. 8 depicts a flow chart illustrating an example process for using timing characteristics of data requests made by individual applications to delay transfer of one or more of the data requests made by one of the individual applications.

[0016] FIG. 9 depicts a flow chart illustrating an example process for using application behavior of multiple applications to align their content requests made over the network.

[0017] FIG. 10 depicts an example of processes through which the time of delay for content requests can be determined to align content requests over the wireless network.

[0018] FIG. 11 shows a diagrammatic representation of a machine in the example form of a computer system within which a set of instructions, for causing the machine to perform any one or more of the methodologies discussed herein, may be executed.

DETAILED DESCRIPTION

[0019] The following description and drawings are illustrative and are not to be construed as limiting. Numerous specific details are described to provide a thorough understanding of the disclosure. However, in certain instances, well-known or conventional details are not described in order to avoid obscuring the description. References to one or an embodiment in the present disclosure can be, but not necessarily are, references to the same embodiment; and, such references mean at least one of the embodiments.

[0020] Reference in this specification to “one embodiment” or “an embodiment” means that a particular feature, structure, or characteristic described in connection with the embodiment is included in at least one embodiment of the disclosure. The appearances of the phrase “in one embodiment” in various places in the specification are not necessarily all referring to the same embodiment, nor are separate or alternative embodiments mutually exclusive of other embodiments. Moreover, various features are described which may be exhibited by some embodiments and not by others. Similarly, various requirements are described which may be requirements for some embodiments but not other embodiments.

[0021] The terms used in this specification generally have their ordinary meanings in the art, within the context of the disclosure, and in the specific context where each term is used. Certain terms that are used to describe the disclosure are discussed below, or elsewhere in the specification, to provide additional guidance to the practitioner regarding the description of the disclosure. For convenience, certain terms may be highlighted, for example using italics and/or quotation marks. The use of highlighting has no influence on the scope and meaning of a term; the scope and meaning of a term is the same, in the same context, whether or not it is highlighted. It will be appreciated that same thing can be said in more than one way.

[0022] Consequently, alternative language and synonyms may be used for any one or more of the terms discussed herein, nor is any special significance to be placed upon whether or not a term is elaborated or discussed herein. Synonyms for certain terms are provided. A recital of one or more synonyms does not exclude the use of other synonyms. The use of examples anywhere in this specification including examples of any terms discussed herein is illustrative only, and is not intended to further limit the scope and meaning of the disclosure or of any exemplified term. Likewise, the disclosure is not limited to various embodiments given in this specification.

[0023] Without intent to limit the scope of the disclosure, examples of instruments, apparatus, methods and their related results according to the embodiments of the present disclosure are given below. Note that titles or subtitles may be used in the examples for convenience of a reader, which in no way should limit the scope of the disclosure. Unless otherwise defined, all technical and scientific terms used herein have the same meaning as

commonly understood by one of ordinary skill in the art to which this disclosure pertains. In the case of conflict, the present document, including definitions will control.

[0024] Embodiments of the present disclosure include systems and methods for mobile network traffic coordination across multiple applications.

[0025] One embodiment of the disclosed technology includes, a system that optimizes multiple aspects of the connection with wired and wireless networks and devices through a comprehensive view of device and application activity including: loading, current application needs on a device, controlling the type of access (push vs. pull or hybrid), location, concentration of users in a single area, time of day, how often the user interacts with the application, content or device, and using this information to shape traffic to a cooperative client/server or simultaneously mobile devices without a cooperative client. Because the disclosed server is not tied to any specific network provider it has visibility into the network performance across all service providers. This enables optimizations to be applied to devices regardless of the operator or service provider, thereby enhancing the user experience and managing network utilization while roaming. Bandwidth has been considered a major issue in wireless networks today. More and more research has been done related to the need for additional bandwidth to solve access problems — many of the performance enhancing solutions and next generation standards, such as those commonly referred to as 4G, namely LTE, 4G, and WiMAX are focused on providing increased bandwidth. Although partially addressed by the standards a key problem that remains is lack of bandwidth on the signaling channel more so than the data channel.

[0026] Embodiments of the disclosed technology includes, for example, alignment of requests from multiple applications to minimize the need for several polling requests; leverage specific content types to determine how to proxy/manage a connection/content; and apply specific heuristics associated with device, user behavioral patterns (how often they interact with the device/application) and/or network parameters.

[0027] Embodiments of the present technology can further include, moving recurring HTTP polls performed by various widgets, RSS readers, etc., to remote network node (e.g., Network operation center (NOC)), thus considerably lowering device battery/power

consumption, radio channel signaling, and bandwidth usage. Additionally, the offloading can be performed transparently so that existing applications do not need to be changed.

[0028] In some embodiments, this can be implemented using a local proxy on the mobile device which automatically detects recurring requests for the same content (RSS feed, Widget data set) that matches a specific rule (e.g. happens every 15 minutes). The local proxy can automatically cache the content on the mobile device while delegating the polling to the server (e.g., a proxy server operated as an element of a communications network). The server can then notify the mobile/client proxy if the content changes, and if content has not changed (or not changed sufficiently, or in an identified manner or amount) the mobile proxy provides the latest version in its cache to the user (without need to utilize the radio at all). This way the mobile device (e.g., a mobile phone, smart phone, etc.) does not need to open up (e.g., thus powering on the radio) or use a data connection if the request is for content that is monitored and that has been not flagged as new/changed.

[0029] The logic for automatically adding content sources/application servers (e.g., including URLs/content) to be monitored can also check for various factors like how often the content is the same, how often the same request is made (is there a fixed interval/pattern?), which application is requesting the data, etc. Similar rules to decide between using the cache and request the data from the original source may also be implemented and executed by the local proxy and/or server.

[0030] For example, when the request comes at an unscheduled/unexpected time (user initiated check), or after every (n) consecutive times the response has been provided from the cache, etc., or if the application is running in the background vs. in a more interactive mode of the foreground. As more and more mobile applications base their features on resources available in the network, this becomes increasingly important. In addition, the disclosed technology allows elimination of unnecessary chatter from the network, benefiting the operators trying to optimize the wireless spectrum usage.

Cross Application Traffic Coordination

[0031] In one embodiment of the present disclosure, a group of applications [A, B, C, ...] may have a timeline of transfers of data from the mobile device (or client (e.g., mobile application, software agent, widget) on the mobile device) to the network, or from the

network to the mobile device (for receipt by the client). The time of the transfers can be represented as:

[0032] Application A: tA1, tA2, tA3, ...

[0033] Application B: tB1, tB2, tB3, ...

[0034] Application C: tC1, tC2, tC3, ...

[0035] Each of the times 't' can have a natural point of occurring based upon the independent activity of the corresponding application as operations are executed at the application server/provider and/or on the software client on the mobile device. For example, an application can transfer a message, an event, or other types of data to the network (or vice versa) at a regular or semi-regular series of times as part of polling, satisfying a device, application, or user request, application maintenance, or other operation.

[0036] Similarly, an application can transfer a message, event, or other types data via the network (or vice versa) at a regular, semi-regular, or irregular series of times to perform its inherent functions or operations, such as synchronizing two data stores, determining the contents of a data store, accessing new data from the application server/service provider, communicating with a peer device (e.g., another device with the same application or another application with which the requesting application interacts), exchanging control messages, etc.

[0037] In some instances, there is typically no correlation or weak correlation between the times at which data transfers or event transactions occur for one application as compared to a second application on a given mobile device, or for different data requests for the same application. In some cases, there may be a stronger correlation between the times at which a transfer occurs for one application as compared to a second application (e.g., where an operation of a first application is dependent upon or triggers an operation of a second application, or where a user typically executes an operation of one application in conjunction with an operation of a second application). Note that in some instances, the second application may be the same application as the first application and that

correlations can be tracked and determined for multiple requests sent by one application in a similar fashion.

[0038] In some embodiments, in order to optimize (e.g., typically to minimize) the number of times that a device (e.g., a mobile device or smart phone) radio is turned on to decrease the consumption of power (and hence conserve its battery or other power source), a distributed proxy system including a local proxy and/or proxy server can operate to intercept the events or transactions (or requests for transfer) of information. When intercepted, the local proxy and/or the proxy server can delay (or expedite) the time at which one or more of these transfers would normally occur in order to perform multiple transfers together as part of a single transfer operation (i.e., instead of performing multiple, individual transfers). Alternatively, the local and/or proxy server can pre-retrieve data for a non-priority application or less important/time sensitive application whose polls are typically expected to happen before another application having higher priority, for example. In other words, a delay could be negative resulting in content pre-retrieval for alignment with an anticipated data request which typically happens before the request of the lesser priority application.

[0039] The delay time (D) can represent a maximum time delay value (or in some instances, expedited time value) after receipt of a request to make such a transfer, with the value of D determined so as to enable the collection of as many of the transfers as feasible in a single, optimized data transfer. The delay times or expedited times of one or more transfers are determined so as to factor in any potential impact on performance and user experience. Ideally, the system determines D to prevent undesired penalties or inefficiencies, and to prevent undesired impact on the user experience. Note that as described above, the delay 'D' could be negative or positive for alignment purposes (e.g., to implement a delayed or an expedited transfer).

[0040] In some embodiments, delay time 'D' (use to represent both positive and negative delays (effectively and expedited transfer)) can be determined based on consideration of one or more of the following factors: the priority of the application (or the relative priority of one application in comparison to another), the nature or amount of data involved in the transfer (e.g., whether it represents fresh data, a housekeeping function, a control instruction, etc.), the status of the application (e.g., active, inactive, background,

foreground, etc.), a useable lifetime of the data to be transferred (a period before it becomes stale), the interval between the transfer times of multiple data requests for a single application, the interval between the transfer times across more than one application (e.g., the largest transfer time interval based on consideration of all active applications), network characteristics (available bandwidth, network latency, etc.), or another relevant factor.

[0041] In some embodiments, the delay time 'D' of specific events/transactions can be controlled by the mobile device (e.g., platform, device settings, or OS specifications), network service provider, and/or the user as part of optimizing the battery life to align data transfer requests across multiple applications or the same application, as opposed to performing each data transfer individually. In some instances, the user can manually configure a setting specifying that requests across multiple applications or the same application are to be batched. The user can enable the setting, and allow the system to configure the details. In addition, the user can specify preferences, priorities, or any other constraints related to alignment of data request transfer of the mobile network.

[0042] FIG. 1A illustrates an example diagram of a system where a host server 100 facilitates management of traffic between client devices 102 and an application server or content provider 110 in a wireless network for resource conservation.

[0043] The client devices 102A-D can be any system and/or device, and/or any combination of devices/systems that is able to establish a connection, including wired, wireless, cellular connections with another device, a server and/or other systems such as host server 100 and/or application server/content provider 110. Client devices 102 will typically include a display and/or other output functionalities to present information and data exchanged between among the devices 102 and/or the host server 100 and/or application server/content provider 110.

[0044] For example, the client devices 102 can include mobile, hand held or portable devices or non-portable devices and can be any of, but not limited to, a server desktop, a desktop computer, a computer cluster, or portable devices including, a notebook, a laptop computer, a handheld computer, a palmtop computer, a mobile phone, a cell phone, a smart phone, a PDA, a Blackberry device, a Palm device, a handheld tablet (e.g. an iPad or any other tablet), a hand held console, a hand held gaming device or console, any

SuperPhone such as the iPhone, and/or any other portable, mobile, hand held devices, etc. In one embodiment, the client devices 102, host server 100, and app server 110 are coupled via a network 106 and/or a network 108. In some embodiments, the devices 102 and host server 100 may be directly connected to one another. .

[0045] The input mechanism on client devices 102 can include touch screen keypad (including single touch, multi-touch, gesture sensing in 2D or 3D, etc.), a physical keypad, a mouse, a pointer, a track pad, motion detector (e.g., including 1-axis, 2-axis, 3-axis accelerometer, etc.), a light sensor, capacitance sensor, resistance sensor, temperature sensor, proximity sensor, a piezoelectric device, device orientation detector (e.g., electronic compass, tilt sensor, rotation sensor, gyroscope, accelerometer), or a combination of the above.

[0046] Signals received or detected indicating user activity at client devices 102 through one or more of the above input mechanism, or others, can be used in the disclosed technology in acquiring context awareness at the client device 102. Context awareness at client devices 102 generally includes, by way of example but not limitation, client device 102 operation or state acknowledgement, management, user activity/behavior/interaction awareness, detection, sensing, tracking, trending, and/or application (e.g., mobile applications) type, behavior, activity, operating state, etc.

[0047] Context awareness in the present disclosure also includes knowledge and detection of network side contextual data and can include network information such as network capacity, bandwidth, traffic, type of network/connectivity, and/or any other operational state data. Network side contextual data can be received from and/or queried from network service providers (e.g., cell provider 112 and/or Internet service providers) of the network 106 and/or network 108 (e.g., by the host server and/or devices 102). In addition to application context awareness as determined from the client 102 side, the application context awareness may also be received from or obtained/queried from the respective application/service providers 110 (by the host 100 and/or client devices 102).

[0048] The host server 100 can use, for example, contextual information obtained for client devices 102, networks 106/108, applications (e.g., mobile applications), application server/provider 110, or any combination of the above, to manage the traffic in the system to satisfy data needs of the client devices 102 (e.g., to satisfy application or any other

request including HTTP request). In one embodiment, the traffic is managed by the host server 100 to satisfy data requests made in response to explicit or non-explicit user 103 requests and/or device/application maintenance tasks. The traffic can be managed such that network consumption, for example, use of the cellular network is conserved for effective and efficient bandwidth utilization. In addition, the host server 100 can manage and coordinate such traffic in the system such that use of device 102 side resources (e.g., including but not limited to battery power consumption, radio use, processor/memory use) are optimized with a general philosophy for resource conservation while still optimizing performance and user experience.

[0049] For example, in context of battery conservation, the device 150 can observe user activity (for example, by observing user keystrokes, backlight status, or other signals via one or more input mechanisms, etc.) and alters device 102 behaviors. The device 150 can also request the host server 100 to alter the behavior for network resource consumption based on user activity or behavior.

[0050] In one embodiment, the traffic management for resource conservation is performed using a distributed system between the host server 100 and client device 102. The distributed system can include proxy server and cache components on the server 100 side and on the client 102 side, for example, as shown by the server cache 135 on the server 100 side and the local cache 150 on the client 102 side.

[0051] Functions and techniques disclosed for context aware traffic management for resource conservation in networks (e.g., network 106 and/or 108) and devices 102, reside in a distributed proxy and cache system. The proxy and cache system can be distributed between, and reside on, a given client device 102 in part or in whole and/or host server 100 in part or in whole. The distributed proxy and cache system are illustrated with further reference to the example diagram shown in **FIG. 1B**. Functions and techniques performed by the proxy and cache components in the client device 102, the host server 100, and the related components therein are described, respectively, in detail with further reference to the examples of **FIG. 2-3**.

[0052] In one embodiment, client devices 102 communicate with the host server 100 and/or the application server 110 over network 106, which can be a cellular network. To facilitate overall traffic management between devices 102 and various application

servers/content providers 110 to implement network (bandwidth utilization) and device resource (e.g., battery consumption), the host server 100 can communicate with the application server/providers 110 over the network 108, which can include the Internet.

[0053] In general, the networks 106 and/or 108, over which the client devices 102, the host server 100, and/or application server 110 communicate, may be a cellular network, a telephonic network, an open network, such as the Internet, or a private network, such as an intranet and/or the extranet, or any combination thereof. For example, the Internet can provide file transfer, remote log in, email, news, RSS, cloud-based services, instant messaging, visual voicemail, push mail, VoIP, and other services through any known or convenient protocol, such as, but is not limited to the TCP/IP protocol, UDP, HTTP, DNS, Open System Interconnections (OSI), FTP, UPnP, iSCSI, NSF, ISDN, PDH, RS-232, SDH, SONET, etc.

[0054] The networks 106 and/or 108 can be any collection of distinct networks operating wholly or partially in conjunction to provide connectivity to the client devices 102 and the host server 100 and may appear as one or more networks to the serviced systems and devices. In one embodiment, communications to and from the client devices 102 can be achieved by, an open network, such as the Internet, or a private network, such as an intranet and/or the extranet. In one embodiment, communications can be achieved by a secure communications protocol, such as secure sockets layer (SSL), or transport layer security (TLS).

[0055] In addition, communications can be achieved via one or more networks, such as, but are not limited to, one or more of WiMax, a Local Area Network (LAN), Wireless Local Area Network (WLAN), a Personal area network (PAN), a Campus area network (CAN), a Metropolitan area network (MAN), a Wide area network (WAN), a Wireless wide area network (WWAN), enabled with technologies such as, by way of example, Global System for Mobile Communications (GSM), Personal Communications Service (PCS), Digital Advanced Mobile Phone Service (D-Amps), Bluetooth, Wi-Fi, Fixed Wireless Data, 2G, 2.5G, 3G, 4G, IMT-Advanced, pre-4G, 3G LTE, 3GPP LTE, LTE Advanced, mobile WiMax, WiMax 2, WirelessMAN-Advanced networks, enhanced data rates for GSM evolution (EDGE), General packet radio service (GPRS), enhanced GPRS, iBurst, UMTS, HSPDA, HSUPA, HSPA, UMTS-TDD, 1xRTT, EV-DO, messaging

protocols such as, TCP/IP, SMS, MMS, extensible messaging and presence protocol (XMPP), real time messaging protocol (RTMP), instant messaging and presence protocol (IMPP), instant messaging, USSD, IRC, or any other wireless data networks or messaging protocols.

[0056] FIG. 1B illustrates an example diagram of a proxy and cache system distributed between the host server 100 and device 150 which facilitates network traffic management between the device 150 and an application server/content provider 100 (e.g., a source server) for resource conservation.

[0057] The distributed proxy and cache system can include, for example, the proxy server 125 (e.g., remote proxy) and the server cache, 135 components on the server side. The server-side proxy 125 and cache 135 can, as illustrated, reside internal to the host server 100. In addition, the proxy server 125 and cache 135 on the server-side can be partially or wholly external to the host server 100 and in communication via one or more of the networks 106 and 108. For example, the proxy server 125 may be external to the host server and the server cache 135 may be maintained at the host server 100. Alternatively, the proxy server 125 may be within the host server 100 while the server cache is external to the host server 100. In addition, each of the proxy server 125 and the cache 135 may be partially internal to the host server 100 and partially external to the host server 100.

[0058] The distributed system can also, include, in one embodiment, client-side components, including by way of example but not limitation, a local proxy 175 (e.g., a mobile client on a mobile device) and/or a local cache 185, which can, as illustrated, reside internal to the device 150 (e.g., a mobile device).

[0059] In addition, the client-side proxy 175 and local cache 185 can be partially or wholly external to the device 150 and in communication via one or more of the networks 106 and 108. For example, the local proxy 175 may be external to the device 150 and the local cache 185 may be maintained at the device 150. Alternatively, the local proxy 175 may be within the device 150 while the local cache 185 is external to the device 150. In addition, each of the proxy 175 and the cache 185 may be partially internal to the host server 100 and partially external to the host server 100.

[0060] In one embodiment, the distributed system can include an optional caching proxy server 199. The caching proxy server 199 can be a component which is operated by the application server/content provider 110, the host server 100, or a network service provider 112, and or any combination of the above to facilitate network traffic management for network and device resource conservation. Proxy server 199 can be used, for example, for caching content to be provided to the device 150, for example, from one or more of, the application server/provider 110, host server 100, and/or a network service provider 112. Content caching can also be entirely or partially performed by the remote proxy 125 to satisfy application requests or other data requests at the device 150.

[0061] In context aware traffic management and optimization for resource conservation in a network (e.g., cellular or other wireless networks), characteristics of user activity/behavior and/or application behavior at a mobile device 150 can be tracked by the local proxy 175 and communicated, over the network 106 to the proxy server 125 component in the host server 100, for example, as connection metadata. The proxy server 125 which in turn is coupled to the application server/provider 110 provides content and data to satisfy requests made at the device 150.

[0062] In addition, the local proxy 175 can identify and retrieve mobile device properties including, one or more of, battery level, network that the device is registered on, radio state, whether the mobile device is being used (e.g., interacted with by a user). In some instances, the local proxy 175 can delay, expedite (prefetch) and/or modify data prior to transmission to the proxy server 125, when appropriate, as will be further detailed with references to the description associated with the examples of **FIG. 2-3**.

[0063] The local database 185 can be included in the local proxy 175 or coupled to the proxy 175 and can be queried for a locally stored response to the data request prior to the data request being forwarded on to the proxy server 125. Locally cached responses can be used by the local proxy 175 to satisfy certain application requests of the mobile device 150, by retrieving cached content stored in the cache storage 185, when the cached content is still valid.

[0064] Similarly, the proxy server 125 of the host server 100 can also delay, expedite, or modify data from the local proxy prior to transmission to the content sources (e.g., the app server/content provider 110). In addition, the proxy server 125 uses device properties

and connection metadata to generate rules for satisfying request of applications on the mobile device 150. The proxy server 125 can gather real time traffic information about requests of applications for later use in optimizing similar connections with the mobile device 150 or other mobile devices.

[0065] In general, the local proxy 175 and the proxy server 125 are transparent to the multiple applications executing on the mobile device. The local proxy 175 is generally transparent to the operating system or platform of the mobile device and may or may not be specific to device manufacturers. In some instances, the local proxy 175 is optionally customizable in part or in whole to be device specific. In some embodiments, the local proxy 175 may be bundled into a wireless model, into a firewall, and/or a router.

[0066] In one embodiment, the host server 100 can in some instances, utilize the store and forward functions of a short message service center (SMSC) 112, such as that provided by the network service provider 112, in communicating with the device 150 in achieving network traffic management. As will be further described with reference to the example of **FIG. 3**, the host server 100 can forward content or HTTP responses to the SMSC 112 such that it is automatically forwarded to the device 150 if available, and for subsequent forwarding if the device 150 is not currently available.

[0067] In general, the disclosed distributed proxy and cache system allows optimization of network usage, for example, by serving requests from the local cache 185, the local proxy 175 reduces the number of requests that need to be satisfied over the network 106. Further, the local proxy 175 and the proxy server 125 may filter irrelevant data from the communicated data. In addition, the local proxy 175 and the proxy server 125 can also accumulate low priority data and send it in batches to avoid the protocol overhead of sending individual data fragments. The local proxy 175 and the proxy server 125 can also compress or transcode the traffic, reducing the amount of data sent over the network 106 and/or 108. The signaling traffic in the network 106 and/or 108 can be reduced, as the networks are now used less often and the network traffic can be synchronized among individual applications.

[0068] With respect to the battery life of the mobile device 150, by serving application or content requests from the local cache 185, the local proxy 175 can reduce the number of times the radio module is powered up. The local proxy 175 and the proxy

server 125 can work in conjunction to accumulate low priority data and send it in batches to reduce the number of times and/or amount of time when the radio is powered up. The local proxy 175 can synchronize the network use by performing the batched data transfer for all connections simultaneously.

[0069] FIG. 2A depicts a block diagram illustrating an example of client-side components in a distributed proxy and cache system residing on a device 250 that manages traffic in a wireless network for resource conservation.

[0070] The device 250, which can be a portable or mobile device, such as a portable phone, generally includes, for example, a network interface 208, an operating system 204, a context API 206, and mobile applications which may be proxy unaware 210 or proxy aware 220. Note that the device 250 is specifically illustrated in the example of FIG. 2 as a mobile device, such is not a limitation and that device 250 may be any portable/mobile or non-portable device able to receive, transmit signals to satisfy data requests over a network including wired or wireless networks (e.g., WiFi, cellular, Bluetooth, etc.).

[0071] The network interface 202 can be a networking module that enables the device 250 to mediate data in a network with an entity that is external to the host server 250, through any known and/or convenient communications protocol supported by the host and the external entity. The network interface 208 can include one or more of a network adaptor card, a wireless network interface card (e.g., SMS interface, WiFi interface, interfaces for various generations of mobile communication standards including but not limited to 1G, 2G, 3G, 3.5G, 4G, LTE, etc.), Bluetooth, or whether or not the connection is via a router, an access point, a wireless router, a switch, a multilayer switch, a protocol converter, a gateway, a bridge, bridge router, a hub, a digital media receiver, and/or a repeater.

[0072] Device 250 can further include, client-side components of the distributed proxy and cache system which can include, a local proxy 275 (e.g., a mobile client of a mobile device) and a cache 285. In one embodiment, the local proxy 275 includes a user activity module 215, a proxy API 225, a request/transaction manager 235, a caching policy manager 245, a traffic shaping engine 255, and/or a connection manager 265. The traffic shaping engine 255 may further include an alignment module 256 and/or a batching module 257, the connection manager 265 may further include a radio controller 266. The

request/transaction manager 235 can further include an application behavior detector 236 and/or a prioritization engine 241, the application behavior detector 236 may further include a pattern detector 237 and/or an application profile generator 239. Additional or less components/modules/engines can be included in the local proxy 275 and each illustrated component.

[0073] As used herein, a “module,” “a manager,” a “handler,” a “detector,” an “interface,” or an “engine” includes a general purpose, dedicated or shared processor and, typically, firmware or software modules that are executed by the processor. Depending upon implementation-specific or other considerations, the module, manager, handler, or engine can be centralized or its functionality distributed. The module, manager, handler, or engine can include general or special purpose hardware, firmware, or software embodied in a computer-readable (storage) medium for execution by the processor. As used herein, a computer-readable medium or computer-readable storage medium is intended to include all mediums that are statutory (e.g., in the United States, under 35 U.S.C. 101), and to specifically exclude all mediums that are non-statutory in nature to the extent that the exclusion is necessary for a claim that includes the computer-readable (storage) medium to be valid. Known statutory computer-readable mediums include hardware (e.g., registers, random access memory (RAM), non-volatile (NV) storage, to name a few), but may or may not be limited to hardware.

[0074] In one embodiment, a portion of the distributed proxy and cache system for network traffic management resides in or is in communication with device 250, including local proxy 275 (mobile client) and/or cache 285. The local proxy 275 can provide an interface on the device 250 for users to access device applications and services including email, IM, voice mail, visual voicemail, feeds, Internet, other applications, etc.

[0075] The proxy 275 is generally application independent and can be used by applications (e.g., both proxy aware and proxy-unaware mobile applications 210 and 220) to open TCP connections to a remote server (e.g., the server 100 in the examples of **FIG. 1A-1B** and/or server proxy 125/325 shown in the examples of **FIG. 1B** and **FIG. 3**). In some instances, the local proxy 275 includes a proxy API 225 which can be optionally used to interface with proxy-aware applications 220 (or mobile applications on a mobile device).

[0076] The applications 210 and 220 can generally include any user application, widgets, software, HTTP-based application, web browsers, video or other multimedia streaming or downloading application, video games, social network applications, email clients, RSS management applications, application stores, document management applications, productivity enhancement applications, etc. The applications can be provided with the device OS, by the device manufacturer, by the network service provider, downloaded by the user, or provided by others.

[0077] One embodiment of the local proxy 275 includes or is coupled to a context API 206, as shown. The context API 206 may be a part of the operating system 204 or device platform or independent of the operating system 204, as illustrated. The operating system 204 can include any operating system including but not limited to, any previous, current, and/or future versions/releases of, Windows Mobile, iOS, Android, Symbian, Palm OS, Brew MP, Java 2 Micro Edition (J2ME), Blackberry, etc.

[0078] The context API 206 may be a plug-in to the operating system 204 or a particular client application on the device 250. The context API 206 can detect signals indicative of user or device activity, for example, sensing motion, gesture, device location, changes in device location, device backlight, keystrokes, clicks,, activated touch screen, mouse click or detection of other pointer devices. The context API 206 can be coupled to input devices or sensors on the device 250 to identify these signals. Such signals can generally include input received in response to explicit user input at an input device/mechanism at the device 250 and/or collected from ambient signals/contextual cues detected at or in the vicinity of the device 250 (e.g., light, motion, piezoelectric, etc.).

[0079] In one embodiment, the user activity module 215 interacts with the context API 206 to identify, determine, infer, detect, compute, predict, and/or anticipate, characteristics of user activity on the device 250. Various inputs collected by the context API 206 can be aggregated by the user activity module 215 to generate a profile for characteristics of user activity. Such a profile can be generated by the module 215 with various temporal characteristics. For instance, user activity profile can be generated in real-time for a given instant to provide a view of what the user is doing or not doing at a given time (e.g., defined by a time window, in the last minute, in the last 30 seconds, etc.), a user activity profile can also be generated for a 'session' defined by an application or

web page that describes the characteristics of user behavior with respect to a specific task they are engaged in on the device 250, or for a specific time period (e.g., for the last 2 hours, for the last 5 hours).

[0080] Additionally, characteristic profiles can be generated by the user activity module 215 to depict a historical trend for user activity and behavior (e.g. 1 week, 1 mo, 2 mo, etc.). Such historical profiles can also be used to deduce trends of user behavior, for example, access frequency at different times of day, trends for certain days of the week (weekends or week days), user activity trends based on location data (e.g., IP address, GPS, or cell tower coordinate data) or changes in location data (e.g., user activity based on user location, or user activity based on whether the user is on the go, or traveling outside a home region, etc.) to obtain user activity characteristics.

[0081] In one embodiment, user activity module 215 can detect and track user activity with respect to applications, documents, files, windows, icons, and folders on the device 250. For example, the user activity module 215 can detect when an application or window (e.g., a web browser) has been exited, closed, minimized, maximized, opened, moved into the foreground, or into the background, multimedia content playback, etc.

[0082] In one embodiment, characteristics of the user activity on the device 250 can be used to locally adjust behavior of the device (e.g., mobile device) to optimize its resource consumption such as battery/power consumption and more generally, consumption of other device resources including memory, storage, and processing power. In one embodiment, the use of a radio on a device can be adjusted based on characteristics of user behavior (e.g., by the radio controller 266 of the connection manager 265) coupled to the user activity module 215. For example, the radio controller 266 can turn the radio on or off, based on characteristics of the user activity on the device 250. In addition, the radio controller 266 can adjust the power mode of the radio (e.g., to be in a higher power mode or lower power mode) depending on characteristics of user activity.

[0083] In one embodiment, characteristics of the user activity on device 250 can also be used to cause another device (e.g., other computers, a mobile device, or a non-portable device) or server (e.g., host server 100 and 300 in the examples of **FIG. 1A-B** and **FIG. 3**) which can communicate (e.g., via a cellular or other network) with the device 250 to modify its communication frequency with the device 250. The local proxy 275 can use

the characteristics information of user behavior determined by the user activity module 215 to instruct the remote device as to how to modulate its communication frequency (e.g., decreasing communication frequency, such as data push frequency if the user is idle, requesting that the remote device notify the device 250 if new data, changed, data, or data of a certain level of importance becomes available, etc.).

[0084] In one embodiment, the user activity module 215 can, in response to determining that user activity characteristics indicate that a user is active after a period of inactivity, request that a remote device (e.g., server host server 100 and 300 in the examples of **FIG. 1A-B** and **FIG. 3**) send the data that was buffered as a result of the previously decreased communication frequency.

[0085] In addition, or in alternative, the local proxy 275 can communicate the characteristics of user activity at the device 250 to the remove device (e.g., e.g., host server 100 and 300 in the examples of **FIG. 1A-B** and **FIG. 3**) and the remote device determines how to alter its own communication frequency with the device 250 for network resource conservation and conservation of device 250 resources.

[0086] One embodiment of the local proxy 275 further includes a request/transaction manager 235, which can detect, identify, intercept, process, manage, data requests initiated on the device 250, for example, by applications 210 and/or 220, and/or directly/indirectly by a user request. The request/transaction manager 235 can determine how and when to process a given request or transaction, or a set of requests/transactions, based on transaction characteristics.

[0087] The request/transaction manager 235 can prioritize requests or transactions made by applications and/or users at the device 250, for example by the prioritization engine 241. Importance or priority of requests/transactions can be determined by the manager 235 by applying a rule set, for example, according to time sensitivity of the transaction, time sensitivity of the content in the transaction, time criticality of the transaction, time criticality of the data transmitted in the transaction, and/or time criticality or importance of an application making the request.

[0088] In addition, transaction characteristics can also depend on whether the transaction was a result of user-interaction or other user initiated action on the device (e.g.,

user interaction with a mobile application). In general, a time critical transaction can include a transaction resulting from a user-initiated data transfer, and can be prioritized as such. Transaction characteristics can also depend on the amount of data that will be transferred or is anticipated to be transferred as a result of the request/requested transaction. For example, the connection manager 265, can adjust the radio mode (e.g., high power or low power mode via the radio controller 266) based on the amount of data that will need to be transferred.

[0089] In addition, the radio controller 266/connection manager 265 can adjust the radio power mode (high or low) based on time criticality/sensitivity of the transaction. The radio controller 266 can trigger the use of high power radio mode when a time-critical transaction (e.g., a transaction resulting from a user-initiated data transfer, an application running in the foreground, any other event meeting a certain criteria) is initiated or detected.

[0090] In general, the priorities can be determined set by in default, for example, based on device platform, device manufacturer, operating system, etc. Priorities can alternatively or in additionally be set by the particular application; for example, the Facebook mobile application can set its own priorities for various transactions (e.g., a status update can be of higher priority than an add friend request or a poke request, a message send request can be of higher priority than a message delete request, for example), an email client or IM chat client may have its own configurations for priority. The prioritization engine 241 may include set of rules for assigning priority.

[0091] The prioritization engine 241 can also track network provider limitations or specifications on application or transaction priority in determining an overall priority status for a request/transaction. Furthermore, priority can in part or in whole be determined by user preferences, either explicit or implicit. A user, can in general, set priorities at different tiers, such as, specific priorities for sessions, or types, or applications (e.g., a browsing session, a gaming session, versus an IM chat session, the user may set a gaming session to always have higher priority than an IM chat session, which may have higher priority than web-browsing session). A user can set application-specific priorities, (e.g., a user may set Facebook related transactions to have a higher priority than LinkedIn related transactions), for specific transaction types (e.g., for all send message requests

across all applications to have higher priority than message delete requests, for all calendar-related events to have a high priority, etc.), and/or for specific folders.

[0092] The prioritization engine 241 can track and resolve conflicts in priorities set by different entities. For example, manual settings specified by the user may take precedence over device OS settings, network provider parameters/limitations (e.g., set in default for a network service area, geographic locale, set for a specific time of day, or set based on service/fee type) may limit any user-specified settings and/or application-set priorities. In some instances, a manual sync request received from a user can override some, most, or all priority settings in that the requested synchronization is performed when requested, regardless of the individually assigned priority or an overall priority ranking for the requested action.

[0093] Priority can be specified and tracked internally in any known and/or convenient manner, including but not limited to, a binary representation, a multi-valued representation, a graded representation and all are considered to be within the scope of the disclosed technology.

Change (initiated on device)	Priority	Change (initiated on server)	Priority
Send email	High	Receive email	High
Delete email	Low	Edit email	Often not possible to sync (Low if possible)
(Un)read email	Low		
Move message	Low	New email in deleted items	Low
Read more	High		
Down load attachment	High	Delete an email	Low
		(Un)Read an email	Low
New Calendar event	High	Move messages	Low
Edit/change Calendar event	High	Any calendar change	High
Add a contact	High	Any contact change	High
Edit a contact	High	Wipe/lock device	High
Search contacts	High	Settings change	High
Change a setting	High	Any folder change	High
		Connector restart	High (if no

Change (initiated on device)	Priority	Change (initiated on server)	Priority
Manual send/receive	High		changes nothing is sent)
IM status change	Medium	Social Network Status Updates	Medium
Auction outbid or change notification	High	Sever Weather Alerts	High
Weather Updates	Low	News Updates	Low

Table I

[0094] Table I below shows, for illustration purposes, some examples of transactions with examples of assigned priorities in a binary representation scheme. Additional assignments are possible for additional types of events, requests, transactions, and as previously described, priority assignments can be made at more or less granular levels, e.g., at the session level or at the application level, etc.

[0095] As shown by way of example in the above table, in general, lower priority requests/transactions can include, updating message status as being read, unread, deleting of messages, deletion of contacts; higher priority requests/transactions, can in some instances include, status updates, new IM chat message, new email, calendar event update/cancellation/deletion, an event in a mobile gaming session, or other entertainment related events, a purchase confirmation through a web purchase or online, request to load additional or download content, contact book related events, a transaction to change a device setting, location-aware or location-based events/transactions, or any other events/request/transactions initiated by a user or where the user is known to be, expected to be, or suspected to be waiting for a response, etc.

[0096] Inbox pruning events (e.g., email, or any other types of messages), are generally considered low priority and absent other impending events, generally will not trigger use of the radio on the device 250. Specifically, pruning events to remove old email or other content can be 'piggy backed' with other communications if the radio is not otherwise on, at the time of a scheduled pruning event. For example, if the user has preferences set to 'keep messages for 7 days old,' then instead of powering on the device radio to initiate a message delete from the device 250 the moment that the message has

exceeded 7 days old, the message is deleted when the radio is powered on next. If the radio is already on, then pruning may occur as regularly scheduled.

[0097] The request/transaction manager 235, can use the priorities for requests (e.g., by the prioritization engine 241) to manage outgoing traffic from the device 250 for resource optimization (e.g., to utilize the device radio more efficiently for battery conservation). For example, transactions/requests below a certain priority ranking may not trigger use of the radio on the device 250 if the radio is not already switched on, as controlled by the connection manager 265. In contrast, the radio controller 266 can turn on the radio such a request can be sent when a request for a transaction is detected to be over a certain priority level.

[0098] In one embodiment, priority assignments (such as that determined by the local proxy 275 or another device/entity) can be used cause a remote device to modify its communication with the frequency with the mobile device. For example, the remote device can be configured to send notifications to the device 250 when data of higher importance is available to be sent to the mobile device.

[0099] In one embodiment, transaction priority can be used in conjunction with characteristics of user activity in shaping or managing traffic, for example, by the traffic shaping engine 255. For example, the traffic shaping engine 255 can, in response to detecting that a user is dormant or inactive, wait to send low priority transactions from the device 250, for a period of time. In addition, the traffic shaping engine 255 can allow multiple low priority transactions to accumulate for batch transferring from the device 250 (e.g., via the batching module 257). In one embodiment, the priorities can be set, configured, or readjusted by a user. For example, content depicted in Table I in the same or similar form can be accessible in a user interface on the device 250 and for example , used by the user to adjust or view the priorities.

[00100] The batching module 257 can initiate batch transfer based on certain criteria. For example, batch transfer (e.g., of multiple occurrences of events, some of which occurred at different instances in time) may occur after a certain number of low priority events have been detected, or after an amount of time elapsed after the first of the low priority event was initiated. In addition, the batching module 257 can initiate batch transfer of the cumulated low priority events when a higher priority event is initiated or

detected at the device 250. Batch transfer can otherwise be initiated when radio use is triggered for another reason (e.g., to receive data from a remote device such as host server 100 or 300). In one embodiment, an impending pruning event (pruning of an inbox), or any other low priority events, can be executed when a batch transfer occurs.

[00101] In general, the batching capability can be disabled or enabled at the event/transaction level, application level, or session level, based on any one or combination of the following: user configuration, device limitations/settings, manufacturer specification, network provider parameters/limitations, platform specific limitations/settings, device OS settings, etc. In one embodiment, batch transfer can be initiated when an application/window/file is closed out, exited, or moved into the background; users can optionally be prompted before initiating a batch transfer; users can also manually trigger batch transfers.

[00102] In one embodiment, the local proxy 275 locally adjusts radio use on the device 250 by caching data in the cache 285. When requests or transactions from the device 250 can be satisfied by content stored in the cache 285, the radio controller 266 need not activate the radio to send the request to a remote entity (e.g., the host server 100, 300, as shown in **FIG. 1** and **FIG. 3** or a content provider/application server such as the server/provider 110 shown in the examples of **FIG. 1A** and **FIG. 1B**). As such, the local proxy 275 can use the local cache 285 and the cache policy manager 245 to locally store data for satisfying data requests to eliminate or reduce the use of the device radio for conservation of network resources and device battery consumption.

[00103] In leveraging the local cache, once the request/transaction manager 225 intercepts a data request by an application on the device 250, the local repository 285 can be queried to determine if there is any locally stored response, and also determine whether the response is valid. When a valid response is available in the local cache 285, the response can be provided to the application on the device 250 without the device 250 needing to access the cellular network.

[00104] If a valid response is not available, the local proxy 275 can query a remote proxy (e.g., the server proxy 325 of **FIG. 3**) to determine whether a remotely stored response is valid. If so, the remotely stored response (e.g., which may be stored on the server cache 135 or optional caching server 199 shown in the example of **FIG. 1B**) can be

provided to the mobile device, possibly without the mobile device 250 needing to access the cellular network, thus relieving consumption of network resources.

[00105] If a valid cache response is not available, or if cache responses are unavailable for the intercepted data request, the local proxy 275, for example, the caching policy manager 245, can send the data request to a remote proxy (e.g., server proxy 325 of **FIG. 3**) which forwards the data request to a content source (e.g., application server/content provider 110 of **FIG. 1**) and a response from the content source can be provided through the remote proxy, as will be further described in the description associated with the example host server 300 of **FIG. 3**. The cache policy manager 245 can manage or process requests that use a variety of protocols, including but not limited to HTTP, HTTPS, IMAP, POP, SMTP and/or ActiveSync. The caching policy manager 245 can locally store responses for data requests in the local database 285 as cache entries, for subsequent use in satisfying same or similar data requests.

[00106] The manager 245 can request that the remote proxy monitor responses for the data request, and the remote proxy can notify the device 250 when an unexpected response to the data request is detected. In such an event, the cache policy manager 245 can erase or replace the locally stored response(s) on the device 250 when notified of the unexpected response (e.g., new data, changed data, additional data, etc.) to the data request. In one embodiment, the caching policy manager 245 is able to detect or identify the protocol used for a specific request, including but not limited to HTTP, HTTPS, IMAP, POP, SMTP and/or ActiveSync. In one embodiment, application specific handlers (e.g., via the application protocol module 246 of the manager 245) on the local proxy 275 allows for optimization of any protocol that can be port mapped to a handler in the distributed proxy (e.g., port mapped on the proxy server 325 in the example of **FIG. 3**).

[00107] In one embodiment, the local proxy 275 notifies the remote proxy such that the remote proxy can monitor responses received for the data request from the content source for changed results prior to returning the result to the device 250, for example, when the data request to the content source has yielded same results to be returned to the mobile device. In general, the local proxy 275 can simulate application server responses for applications on the device 250, using locally cached content. This can prevent

utilization of the cellular network for transactions where new/changed data is not available, thus freeing up network resources and preventing network congestion.

[00108] In one embodiment, the local proxy 275 includes an application behavior detector 236 to track, detect, observe, monitor, applications (e.g., proxy aware and/or unaware applications 210 and 220) accessed or installed on the device 250. Application behaviors, or patterns in detected behaviors (e.g., via the pattern detector 237) of one or more applications accessed on the device 250 can be used by the local proxy 275 to optimize traffic in a wireless network needed to satisfy the data needs of these applications.

[00109] For example, based on detected behavior of multiple applications, the traffic shaping engine 255 can align content requests made by at least some of the applications over the network (wireless network) (e.g., via the alignment module 256). The alignment module can delay or expedite some earlier received requests to achieve alignment. When requests are aligned, the traffic shaping engine 255 can utilize the connection manager to poll over the network to satisfy application data requests. Content requests for multiple applications can be aligned based on behavior patterns or rules/settings including, for example, content types requested by the multiple applications (audio, video, text, etc.), mobile device parameters, and/or network parameters/traffic conditions, network service provider constraints/specifications, etc.

[00110] In one embodiment, the pattern detector 237 can detect recurrences in application requests made by the multiple applications, for example, by tracking patterns in application behavior. A tracked pattern can include, detecting that certain applications, as a background process, poll an application server regularly, at certain times of day, on certain days of the week, periodically in a predictable fashion, with a certain frequency, with a certain frequency in response to a certain type of event, in response to a certain type user query, frequency that requested content is the same, frequency with which a same request is made, interval between requests, applications making a request, or any combination of the above, for example.

[00111] Such recurrences can be used by traffic shaping engine 255 to offload polling of content from a content source (e.g., from an application server/content provider 110 of **FIG. 1**) that would result from the application requests that would be performed at the

mobile device 250 to be performed instead, by a proxy server (e.g., proxy server 125 of **FIG. 1B** or proxy server 325 of **FIG. 3**) remote from the device 250. Traffic engine 255 can decide to offload the polling when the recurrences match a rule. For example, there are multiple occurrences or requests for the same resource that have exactly the same content, or returned value, or based on detection of repeatable time periods between requests and responses such as a resource that is requested at specific times during the day. The offloading of the polling can decrease the amount of bandwidth consumption needed by the mobile device 250 to establish a wireless (cellular) connection with the content source for repetitive content polls.

[00112] As a result of the offloading of the polling, locally cached content stored in the local cache 285 can be provided to satisfy data requests at the device 250, when content change is not detected in the polling of the content sources. As such, when data has not changed, application data needs can be satisfied without needing to enable radio use or occupying cellular bandwidth in a wireless network. When data has changed and/or new data has been received, the remote entity to which polling is offloaded, can notify the device 250. The remote entity may be the host server 300 as shown in the example of **FIG. 3**.

[00113] In one embodiment, the local proxy 275 can mitigate the need/use of periodic keep alive messages (heartbeat messages) to maintain TCP/IP connections, which can consume significant amounts of power thus having detrimental impacts on mobile device battery life. The connection manager 265 in the local proxy (e.g., the heartbeat manager 267) can detect, identify, and intercept any or all heartbeat (keep-alive) messages being sent from applications.

[00114] The heartbeat manager 267 can prevent any or all of these heartbeat messages from being sent over the cellular, or other network, and instead rely on the server component of the distributed proxy system (e.g., shown in **FIG. 1B**) to generate the and send the heartbeat messages to maintain a connection with the backend (e.g., app server/provider 110 in the example of **FIG. 1**).

[00115] The local proxy 275 generally represents any one or a portion of the functions described for the individual managers, modules, and/or engines. The local proxy 275 and

device 250 can include additional or less components; more or less functions can be included, in whole or in part, without deviating from the novel art of the disclosure.

[00116] FIG. 2B depicts a block diagram illustrating another example of components in the application behavior detector 236 and the traffic shaping engine 255 in the local proxy 275 on the client-side of the distributed proxy system shown in the example of FIG. 2A.

[00117] In one embodiment, the pattern detector 237 of the application behavior detector 236 further includes a correlation detector 238 and the application profile generator 239 further includes an application status detector 240.

[00118] The correlation detector 238, in one embodiment, can detect, determine, identify, compute, track, any correlations in the timing of data transfer requests made by applications, agents, and/or widgets accessed via (e.g., application streaming or accessed through a cloud) or running on the device 250.

[00119] In general, correlation types include event-level correlations and application-level correlations, and can include system/application. The correlation detector 238 can track and monitor system/application triggered events and/or user-triggered events/transactions. In addition, the correlation detector 238 can identify or track correlations between events for a given application or across different applications.

[00120] A correlation can be detected, for example, when a first event/transaction type of a first application triggers the initiation of a second event/transaction type of the same application. The triggering can be detected by the correlation detector 238 through identifying patterns in the timing characteristics of such events occurring within the first application. For example, a correlation can include, an identification of the ordering of the first and second event/transaction types (e.g., the second event type always occurs after the first event type). A correlation can also include, determining that the first and second event/transaction types occur within a timeframe of one another (e.g., the first event type occurs within a 10 ms. time window of the second event type), etc.

[00121] A correlation can also be detected, for example, when a specific event type of one application triggers a specific event type of another application (or has a timing or

ordering relationship thereof). Similarly, the correlation detector 238 can detect that operation of one application (e.g., such as the launching of one application or any other activity) is related to the operation of another application. For example, the detector can determine that if one application is launched, the other application is also launched.

[00122] The detector can also detect that one application is always launched within a certain time window of another application being launched/accessed, or the activity status of one application is linked to the application status of the activity status of another application (e.g., when one application moves into the foreground/background, the other application moves into the foreground/background, or when one application becomes active/inactive, the other application also changes state and becomes either active or inactive, etc.). Generally, such events/transactions detected and tracked for correlation can include system or application-initiated events, or user-triggered events (e.g., including explicit user requests or implicit user requests).

[00123] In some instances, the correlation detector 238 uses user activity module 215 and can also detect application or application event correlations in relation to and/or in conjunction with user activity. For example, the correlation detector 238 can determine that an occurrence of a first event type frequently causes the user to perform an action which triggers a second event (of the same application or different application). While examples are described herein for two events and examples given for two applications, note that the correlation detector 238 can track, detect, and identify correlations in occurrences of events/transactions, correlations for multiple events (e.g., 2, 3, 4, 5, etc.) that can be detected and identified.

[00124] In general, the correlation detector 238 tracks the timing characteristics of requests made by applications to detect correlations. The correlations can be incorporated into an application's profile by the profile generator 239.

[00125] In one embodiment, the profile generator includes an application status detector 240. The application status detector 240 can detect an activity state of an application on the device 250. An activity state can indicate, by way of example but not limitation, whether a specific application is operating in the background or foreground on the mobile device 250, whether the application is active or inactive, whether the application is being interacted with (e.g., by a user, or another application or device). The

activity state or status of an application on the device 250 can also be included in an application's profile (e.g., by the profile generator 239) along with any correlation with other events or applications and used for data request alignment.

[00126] For example, one embodiment of the alignment module 256 of the traffic shaping engine 255 which is able to use the application behavior (e.g., as determined by the application behavior detector 236) of one or multiple applications on a device 250 to align some of the content requests (e.g., aligned by the alignment module 256) made by the same application or at least a portion the multiple applications from the mobile device over the network. The application behaviors can be indicated in application profiles generated by the application profile generator 239, for example.

[00127] In some instances the content requests (made by the same application or different applications) are aligned by delaying or expediting a time at which some of the content requests would occur without alignment and the traffic shaping engine 256 can transfer the content requests that are delayed or expedited in a single transfer operation over the network (cellular or other wireless network). The amount of time that a request can be delayed (time 'D') is generally determined (e.g. by the delay module 258) to optimize a number of content requests able to be aligned in the single transfer operation. Hence delay module 258 can utilize the correlations in event/transaction occurrences within an application or across multiple applications as identified by the correlation detector 238 in determining delay time for aligning multiple requests.

[00128] The delay time 'D' (e.g., refers to the time by which a request is expedited or delayed) is generally determined based on application behavior (e.g., as determined by the application behavior detector 236 and/or indicated in an application profile). More specifically, the time that is delayed in transfer of a given content request can be determined based on priority of a specific application (e.g., as determined by the prioritization engine 241 of the application behavior detector 236) making the given content request, or based on the priority of the specific application relative to other applications on the mobile device 250. In addition, the delay module 258 can determine delays in transfer further based on, one or more of amount of data involved in the given content request, a nature of data involved in the given content request, usable lifetime of

data to be transferred in the given content request, and/or network characteristics including available bandwidth or network latency.

[00129] Thus in one example of a system level operation in aligning requests, the local proxy 275 of the device 250 detects a first data request made via a first application and a second data request made via a second application, or where first and second data requests are made by the same application. The alignment module 256 of the traffic shaping engine 255 in the local proxy 275 on the device 250 can delay the transfer of the first data transfer request made via the first application to the proxy server (e.g., proxy server 125 or 325 in the examples of **FIG. 1B** and **FIG. 3**) until another data transfer request made via the second application is detected by the local proxy 275 such that the transfer the first data transfer request and the second data transfer request occur in a single transfer operation over the network (thus needing to enable radio use only once, assuming that the radio was off on the device 250 when the first data request occurred).

[00130] The delay of the first data transfer request can be determined by the delay module 258 and the decision can be made when, for example, the second application is of a higher priority relative to the first application, or that the second application is running in the foreground, or that a user is interacting with the second application and the second data request is initiated in response to the user interaction, or that the second application is more data intensive than the first application.

[00131] **FIG. 3** depicts a block diagram illustrating an example of server-side components in a distributed proxy and cache system residing on a host server 300 that manages traffic in a wireless network for resource conservation.

[00132] The host server 300 generally includes, for example, a network interface 308 and/or one or more repositories 312, 314, 316. Note that server 300 may be any portable/mobile or non-portable device, server, cluster of computers and/or other types of processing units (e.g., any number of a machine shown in the example of **FIG. 11**) able to receive, transmit signals to satisfy data requests over a network including any wired or wireless networks (e.g., WiFi, cellular, Bluetooth, etc.).

[00133] The network interface 308 can include networking module(s) or device(s) that enable the server 300 to mediate data in a network with an entity that is external to the

host server 300, through any known and/or convenient communications protocol supported by the host and the external entity. Specifically, the network interface 308 allows the server 308 to communicate with multiple devices including mobile phone devices 350, and/or one or more application servers/content providers 310.

[00134] The host server 300 can store information about connections (e.g., network characteristics, conditions, types of connections, etc.) with devices in the connection metadata repository 312. Additionally, any information about third party application or content providers can also be stored in 312. The host server 300 can store information about devices (e.g., hardware capability, properties, device settings, device language, network capability, manufacturer, device model, OS, OS version, etc.) in the device information repository 314. Additionally, the host server 300 can store information about network providers and the various network service areas in the network service provider repository 316.

[00135] The communication enabled by 308 allows for simultaneous connections (e.g., including cellular connections) with devices 350 and/or connections (e.g., including wired/wireless, HTTP, Internet connections, LAN, Wifi, etc.) with content servers/providers 310, to manage the traffic between devices 350 and content providers 310, for optimizing network resource utilization and/or to conserve power (battery) consumption on the serviced devices 350. The host server 300 can communicate with mobile devices 350 serviced by different network service providers and/or in the same/different network service areas. The host server 300 can operate and is compatible with devices 350 with varying types or levels of mobile capabilities, including by way of example but not limitation, 1G, 2G, 2G transitional (2.5G, 2.75G), 3G (IMT-2000), 3G transitional (3.5G, 3.75G, 3.9G), 4G (IMT-advanced), etc.

[00136] In general, the network interface 308 can include one or more of a network adaptor card, a wireless network interface card (e.g., SMS interface, WiFi interface, interfaces for various generations of mobile communication standards including but not limited to 1G, 2G, 3G, 3.5G, 4G type networks such as , LTE, WiMAX, etc.), Bluetooth, WiFi, or any other network whether or not connected via a a router, an access point, a wireless router, a switch, a multilayer switch, a protocol converter, a gateway, a bridge, bridge router, a hub, a digital media receiver, and/or a repeater.

[00137] The host server 300 can further include, server-side components of the distributed proxy and cache system which can include, a proxy server 325 and a server cache 335. In one embodiment, the server proxy 325 can include an HTTP access engine 345, a caching policy manager 355, a proxy controller 365, a traffic shaping engine 375, a new data detector 386, and/or a connection manager 395.

[00138] The HTTP access engine 345 may further include a heartbeat manager 346, the proxy controller 365 may further include a data invalidator module 366, the traffic shaping engine 375 may further include a control protocol 276 and a batching module 377. Additional or less components/modules/engines can be included in the proxy server 325 and each illustrated component.

[00139] As used herein, a “module,” “a manager,” a “handler,” a “detector,” an “interface,” a “controller,” or an “engine” includes a general purpose, dedicated or shared processor and, typically, firmware or software modules that are executed by the processor. Depending upon implementation-specific or other considerations, the module, manager, handler, or engine can be centralized or its functionality distributed. The module, manager, handler, or engine can include general or special purpose hardware, firmware, or software embodied in a computer-readable (storage) medium for execution by the processor. As used herein, a computer-readable medium or computer-readable storage medium is intended to include all mediums that are statutory (e.g., in the United States, under 35 U.S.C. 101), and to specifically exclude all mediums that are non-statutory in nature to the extent that the exclusion is necessary for a claim that includes the computer-readable (storage) medium to be valid. Known statutory computer-readable mediums include hardware (e.g., registers, random access memory (RAM), non-volatile (NV) storage, to name a few), but may or may not be limited to hardware.

[00140] In the example of a device (e.g., mobile device 350) making an application or content request to an app server or content provider 310, the request may be intercepted and routed to the proxy server 325, which is coupled to the device 350 and the provider 310. Specifically, the proxy server is able to communicate with the local proxy (e.g., proxy 175 and 275 of the examples of **FIG. 1** and **FIG. 2** respectively) of the device 350, the local proxy forwards the data request to the proxy server 325 for, in some instances,

further processing, and if needed, for transmission to the content server 310 for a response to the data request.

[00141] In such a configuration, the host 300, or the proxy server 325 in the host server 300 can utilize intelligent information provided by the local proxy in adjusting its communication with the device in such a manner that optimizes use of network and device resources. For example, the proxy server 325 can identify characteristics of user activity on the device 350 to modify its communication frequency. The characteristics of user activity can be determined by, for example, the activity/behavior awareness module 366 in the proxy controller 365, via information collected by the local proxy on the device 350.

[00142] In one embodiment, communication frequency can be controlled by the connection manager 396 of the proxy server 325, for example, to adjust push frequency of content or updates to the device 350. For instance, push frequency can be decreased by the connection manager 396 when characteristics of the user activity indicate that the user is inactive. In one embodiment, when the characteristics of the user activity indicate that the user is subsequently active after a period of inactivity, the connection manager 396 can adjust the communication frequency with the device 350 to send data that was buffered as a result of decreased communication frequency, to the device 350.

[00143] In addition, the proxy server 325 includes priority awareness of various requests, transactions, sessions, applications, and/or specific events. Such awareness can be determined by the local proxy on the device 350 and provided to the proxy server 325. The priority awareness module 367 of the proxy server 325 can generally assess the priority (e.g., including time-criticality, time-sensitivity, etc.) of various events or applications; additionally, the priority awareness module 367 can track priorities determined by local proxies of devices 350.

[00144] In one embodiment, through priority awareness, the connection manager 395 can further modify communication frequency (e.g., use of radio as controlled by the radio controller 396) of the server 300 with the devices 350. For example, the server 300 can notify the device 350, thus requesting use of the radio if it is not already in use, when data or updates of an importance/priority level which meets a criteria becomes available to be sent.

[00145] In one embodiment, the proxy server 325 can detect multiple occurrences of events (e.g., transactions, content, data received from server/provider 310) and allow the events to accumulate for batch transfer to device 350. Batch transfer can be cumulated and transfer of events can be delayed based on priority awareness and/or user activity/application behavior awareness, as tracked by modules 366 and/or 367. For example, batch transfer of multiple events (of a lower priority) to the device 350 can be initiated by the batching module 377 when an event of a higher priority (meeting a threshold or criteria) is detected at the server 300. In addition, batch transfer from the server 300 can be triggered when the server receives data from the device 350, indicating that the device radio is already in use and is thus on. In one embodiment, the proxy server 324 can order the each messages/packets in a batch for transmission based on event/transaction priority, such that higher priority content can be sent first, in case connection is lost or the battery dies, etc.

[00146] In one embodiment, the server 300 caches data (e.g., as managed by the caching policy manager 355) such that communication frequency over a network (e.g., cellular network) with the device 350 can be modified (e.g., decreased). The data can be cached, for example in the server cache 335, for subsequent retrieval or batch sending to the device 350 to potentially decrease the need to turn on the device 350 radio. The server cache 335 can be partially or wholly internal to the host server 300, although in the example of **FIG. 3**, it is shown as being external to the host 300. In some instances, the server cache 335 may be the same as and/or integrated in part or in whole with another cache managed by another entity (e.g., the optional caching proxy server 199 shown in the example of **FIG. 1B**), such as being managed by an application server/content provider 110, a network service provider, or another third party.

[00147] In one embodiment, content caching is performed locally on the device 350 with the assistance of host server 300. For example, proxy server 325 in the host server 300 can query the application server/provider 310 with requests and monitor changes in responses. When changed or new responses are detected (e.g., by the new data detector 347), the proxy server 325 can notify the mobile device 350, such that the local proxy on the device 350 can make the decision to invalidate (e.g., indicated as out-dated) the relevant cache entries stored as any responses in its local cache. Alternatively, the data invalidator module 368 can automatically instruct the local proxy of the device 350 to

invalidate certain cached data, based on received responses from the application server/provider 310. The cached data is marked as invalid, and can get replaced or deleted when new content is received from the content server 310.

[00148] Note that data change can be detected by the detector 347 in one or more ways. For example, the server/provider 310 can notify the host server 300 upon a change. The change can also be detected at the host server 300 in response to a direct poll of the source server/provider 310. In some instances, the proxy server 325 can in addition, pre-load the local cache on the device 350 with the new/updated data. This can be performed when the host server 300 detects that the radio on the mobile device is already in use, or when the server 300 has additional content/data to be sent to the device 350.

[00149] One or more the above mechanisms can be implemented simultaneously or adjusted/configured based on application (e.g., different policies for different servers/providers 310). In some instances, the source provider/server 310 may notify the host 300 for certain types of events (e.g., events meeting a priority threshold level). In addition, the provider/server 310 may be configured to notify the host 300 at specific time intervals, regardless of event priority.

[00150] In one embodiment, the proxy server 325 of the host 300 can monitor/track responses received for the data request from the content source for changed results prior to returning the result to the mobile device, such monitoring may be suitable when data request to the content source has yielded same results to be returned to the mobile device, thus preventing network/power consumption from being used when no new/changes are made to a particular requested. The local proxy of the device 350 can instruct the proxy server 325 to perform such monitoring or the proxy server 325 can automatically initiate such a process upon receiving a certain number of the same responses (e.g., or a number of the same responses in a period of time) for a particular request.

[00151] In one embodiment, the server 300, for example, through the activity/behavior awareness module 366, is able to identify or detect user activity, at a device that is separate from the mobile device 350. For example, the module 366 may detect that a user's message inbox (e.g., email or types of inbox) is being accessed. This can indicate that the user is interacting with his/her application using a device other than the mobile device 350 and may not need frequent updates, if at all.

[00152] The server 300, in this instance, can thus decrease the frequency with which new or updated content is sent to the mobile device 350, or eliminate all communication for as long as the user is detected to be using another device for access. Such frequency decrease may be application specific (e.g., for the application with which the user is interacting with on another device), or it may be a general frequency decrease (e.g., since the user is detected to be interacting with one server or one application via another device, he/she could also use it to access other services) to the mobile device 350.

[00153] In one embodiment, the host server 300 is able to poll content sources 310 on behalf of devices 350 to conserve power or battery consumption on devices 350. For example, certain applications on the mobile device 350 can poll its respective server 310 in a predictable recurring fashion. Such recurrence or other types of application behaviors can be tracked by the activity/behavior module 366 in the proxy controller 365. The host server 300 can thus poll content sources 310 for applications on the mobile device 350, that would otherwise be performed by the device 350 through a wireless (e.g., including cellular connectivity). The host server can poll the sources 310 for new or changed data by way of the HTTP access engine 345 to establish HTTP connection or by way of radio controller 396 to connect to the source 310 over the cellular network. When new or changed data is detected, the new data detector can notify the device 350 that such data is available and/or provide the new/changed data to the device 350.

[00154] In one embodiment, the connection manager 395 determines that the mobile device 350 is unavailable (e.g., the radio is turned off) and utilizes SMS to transmit content to the device 350, for instance via the SMSC shown in the example of **FIG. 1B**. SMS is used to transmit invalidation messages, batches of invalidation messages, or even content in the case the content is small enough to fit into just a few (usually one or two) SMS messages. This avoids the need to access the radio channel to send overhead information. The host server 300 can use SMS for certain transactions or responses having a priority level above a threshold or otherwise meeting a criteria. The server 300 can also utilize SMS as an out-of-band trigger to maintain or wake-up an IP connection as an alternative to maintaining an always-on IP connection.

[00155] In one embodiment, the connection manager 395 in the proxy server 325 (e.g., the heartbeat manager 398) can generate and/or transmit heartbeat messages on behalf of

connected devices 350, to maintain a backend connection with a provider 310 for applications running on devices 350.

[00156] For example, in the distributed proxy system, local cache on the device 350 can prevent any or all heartbeat messages needed to maintain TCP/IP connections required for applications, from being sent over the cellular, or other network, and instead rely on the proxy server 325 on the host server 300 to generate and/or send the heartbeat messages to maintain a connection with the backend (e.g., app server/provider 110 in the example of FIG. 1). The proxy server can generate the keep-alive (heartbeat) messages independent of the operations of the local proxy on the mobile device.

[00157] The repositories 312, 314, and/or 316 can additionally store software, descriptive data, images, system information, drivers, and/or any other data item utilized by other components of the host server 300 and/or any other servers for operation. The repositories may be managed by a database management system (DBMS), for example but not limited to, Oracle, DB2, Microsoft Access, Microsoft SQL Server, PostgreSQL, MySQL, FileMaker, etc.

[00158] The repositories can be implemented via object-oriented technology and/or via text files, and can be managed by a distributed database management system, an object-oriented database management system (OODBMS) (e.g., ConceptBase, FastDB Main Memory Database Management System, JDOInstruments, ObjectDB, etc.), an object-relational database management system (ORDBMS) (e.g., Informix, OpenLink Virtuoso, VMDS, etc.), a file system, and/or any other convenient or known database management package.

[00159] FIG. 4A depicts a diagram showing how data requests from a mobile device 450 to an application server/content provider 496 in a wireless network can be coordinated by a distributed proxy system 460 in a manner such that network and battery resources are conserved through using content caching and monitoring performed by the distributed proxy system 460.

[00160] In satisfying application or client requests on a mobile device 450 without the distributed proxy system 460, the mobile device 450, or the software widget executing on the device 450 performs a data request 402 (e.g., an HTTP GET, POST, or other request)

directly to the application server 495 and receives a response 404 directly from the server/provider 495. If the data has been updated, the widget on the mobile device 450 can refreshes itself to reflect the update and waits for small period of time and initiates another data request to the server/provider 495.

[00161] In one embodiment, the requesting client or software widget 455 on the device 450 can utilize the distributed proxy system 460 in handling the data request made to server/provider 495. In general, the distributed proxy system 460 can include a local proxy 465 (which is typically considered a client-side component of the system 460 and can reside on the mobile device 450), a caching proxy (475, considered a server-side component 470 of the system 460 and can reside on the host server 485 or be wholly or partially external to the host server 485), a host server 485. The local proxy 465 can be connected to the proxy 475 and host server 485 via any network or combination of networks.

[00162] When the distributed proxy system 460 is used for data/application requests, the widget 455 can perform the data request 406 via the local proxy 465. The local proxy 465, can intercept the requests made by device applications, and can identify the connection type of the request (e.g., an HTTP get request or other types of requests). The local proxy 465 can then query the local cache for any previous information about the request (e.g., to determine whether a locally stored response is available and/or still valid). If a locally stored response is not available or if there is an invalid response stored, the local proxy 465 can update or store information about the request, the time it was made, and any additional data, in the local cache. The information can be updated for use in potentially satisfying subsequent requests.

[00163] The local proxy 465 can then send the request to the host server 485 and the server 485 can perform the request 406 and returns the results in response 408. The local proxy 465 can store the result and in addition, information about the result and returns the result to the requesting widget 455.

[00164] In one embodiment, if the same request has occurred multiple times (within a certain time period) and it has often yielded same results, the local proxy 465 can notify 410 the server 485 that the request should be monitored (e.g., steps 412 and 414) for result changes prior to returning a result to the local proxy 465 or requesting widget 455.

[00165] In one embodiment, if a request is marked for monitoring, the local proxy 465 can now store the results into the local cache. Now, when the data request 416, for which a locally response is available, is made by the widget 455 and intercepted at the local proxy 465, the proxy 465 can return the response 418 from the local cache without needing to establish a connection communication over the wireless network.

[00166] In addition, the server proxy performs the requests marked for monitoring 420 to determine whether the response 422 for the given request has changed. In general, the host server 485 can perform this monitoring independently of the widget 455 or local proxy 465 operations. Whenever an unexpected response 422 is received for a request, the server 485 can notify the local proxy 465 that the response has changed (e.g., the invalidate notification in step 424) and that the locally stored response on the client should be erased or replaced with a new response.

[00167] In this case, a subsequent data request 426 by the widget 455 from the device 450 results in the data being returned from host server 485 (e.g., via the caching proxy 475). Thus, through utilizing the distributed proxy system 460 the wireless (cellular) network is intelligently used when the content/data for the widget or software application 455 on the mobile device 450 has actually changed. As such, the traffic needed to check for the changes to application data is not performed over the wireless (cellular) network. This reduces the amount of generated network traffic and shortens the total time and the number of times the radio module is powered up on the mobile device 450, thus reducing battery consumption, and in addition, frees up network bandwidth.

[00168] FIG. 4B depicts a timing diagram showing how data requests from a mobile device 450 to an application server/content provider 495 in a wireless network can be aligned at the local proxy 465 in the distributed proxy system 460 to optimize network and radio use.

[00169] When a data request A 432 is detected at the local proxy 465 on the mobile device 450, the local proxy 465 can determine that the radio on the mobile device 450 is currently off and decide to wait to transfer the request A 432 over the network. When data request B 434 is detected, the proxy 465 can determine (e.g., based on conditions and/or processes illustrated in the flow charts in the examples of FIG. 6-10) whether to transfer the data request B 434 and/or whether also to transfer the impending data request A 432.

[00170] In the example shown in **FIG. 4B**, the local proxy 465 does not transfer data requests A and B until data request C436 is detected at the device 450 by the proxy 465. The time 'D1' 433 with which request A 432 and time 'D2' 435 with which request B 434 are delayed, can be determined by the local proxy 465 (e.g., as described in conjunction with the traffic shaping module and delay module in the example of **FIG. 2B**). The radio can be turned on when request C 436 is received and data transfers for requests A, B, and C 438 can be aligned at this point to be transferred to the host server 480 on the server-side 470 of the distributed proxy 460 which can forward the request to the server/provider 495, immediately in transfer 440, or possibly with some delay, in transfer 442. Note that the data requests A, B, and C can all originate from the same application or different applications on the mobile device 450.

[00171] **FIG. 5** depicts a diagram showing one example process for implementing a hybrid IP and SMS power saving mode on a mobile device 550 using a distributed proxy and cache system (e.g., such as the distributed system shown in the example of **FIG. 1B**).

[00172] In step 502, the local proxy (e.g., proxy 175 in the example of **FIG. 1B**) monitors the device for user activity. When the user is determined to be active, server push is active. For example, always-on-push IP connection can be maintained and if available, SMS triggers can be immediately sent to the mobile device 550 as it becomes available.

[00173] In process 504, after the user has been detected to be inactive or idle over a period of time (e.g., the example is shown for a period of inactivity of 20 min.), the local proxy can adjust the device to go into the power saving mode. In the power saving mode, when the local proxy receives a message or a correspondence from a remote proxy (e.g., the server proxy 135 in the example of **FIG. 1B**) on the server-side of the distributed proxy and cache system, the local proxy can respond with a call indicating that the device 550 is currently in power save mode (e.g., via a power save remote procedure call). In some instances, the local proxy take the opportunity to notify multiple accounts or providers (e.g., 510A, and 510B) of the current power save status (e.g., timed to use the same radio power-on event).

[00174] In one embodiment, the response from the local proxy can include a time (e.g., the power save period) indicating to the remote proxy (e.g., server proxy 135) and/or the

app server/providers 510A/B when the device 550 is next able to receive changes or additional data. A default power savings period can be set by the local proxy. Consecutive power saving periods can increase in duration. For example, if a first power saving period has elapsed without an activity occurring, the device 550 can continue into a second power saving mode with a longer time period (e.g., see periods one 503 and period two 505). In general, any activity on the device takes the client out of power saving mode and ends that particular power save event.

[00175] In addition, if new data or event is received before the end of any one power saving period, then the wait period communicated to the servers 510A/B can be the existing period, rather than an incremented time period. For example, in step 506, since new content was received during the power saving mode, the next wait period communicated in step 508 to servers 510A/B is again the same time saving period. In response, the remote proxy server, upon receipt of power save notification from the device 550, can stop sending changes (data or SMS's) for the period of time requested (the wait period). At the end of the wait period, any notifications received can be acted upon and changes sent to the device 550 as a single batched event. If no notifications come in, then true push can be resumed with the data or an SMS being sent immediately to the device 550. To optimize batch sending content to the mobile device 550, the proxy server can start the poll or data collect event earlier (before the end of a power save period) in order to increase the chance that the client will receive data at the next radio power on event.

[00176] In one embodiment, whenever new data or content comes into the device 550 while it is in a power saving mode, it can respond with the power saving remote procedure call to all end points currently registered (e.g., server/providers 510A/B). Note that the wait period can be updated in operation in real time to accommodate operating conditions. For example, as the mobile device 550 sends additional power saving calls (e.g., with updated wait times) if multiple servers 510A/B or others, respond to the end of a wait period with different delays, the local proxy can adjust the wait period on the fly to accommodate the different delays.

[00177] Detection of user activity 512 at the device 550 causes the power save mode to be exited. When the device 550 exits power save mode, it can send power save cancel call to the proxy server and immediately receives any changes associated with any pending

notifications. This may require a poll to be run by the proxy server after receiving the power saving cancel call. If the latest power saving period has expired, then no power save cancel call may be needed as the proxy server will already be in traditional push operation mode.

[00178] In one embodiment, power save mode is not applied when the device 550 is plugged into a charger. This setting can be reconfigured or adjusted by the user or another party. In general, the power save mode can be turned on and off, for example, by the user via a user interface on device 550. In general, timing of power events to receive data can be synced with any power save calls to optimize radio use.

[00179] FIG. 6 depicts a flow chart illustrating example selection processes through which data transfer requests of multiple applications can be coordinated into a single transfer operation.

[00180] In process 602, a first data transfer request initiated by a first application on a device or mobile device is received. One or more selection processes can be performed to determine whether to delay transfer of the first data transfer request, as shown in steps 604-608, including, determining whether the user is interacting with the first application, whether the first application is in the foreground, or whether the radio of the mobile device on which the request is initiated is already on. Note that although the selection processes are illustrated and identified in a specific order, the order in which the system checks for applicability is not limited to such, any of the above conditions can be checked in any ordering or any combination with one of the other conditions.

[00181] If one of the above applies, in process 610, in general, a decision can be made to transfer the data request over the wireless (cellular or others) network, or to power on the radio (e.g., via the radio controller 256 shown in the example of FIG. 2A) to perform the transfer without or with minimal delay. Additional conditions that are not shown here which may cause the transfer of the data request to occur immediately or upon receipt may be included.

[00182] If none of the conditions in 604-608 applies (or other suitable conditions depending on the specific implementation), in process 612, the transfer of the first data

transfer request can be delayed. In process 614, it is determined whether another data transfer request has been initiated. If not, the process continues at flow 'A' in **FIG. 7**.

[00183] If so, the system performs one or more of several condition checks shown in decision flows 616-624 including determining whether the second application is of higher priority than the first application, whether the second application running in the foreground, whether the user interacting with the second application, whether the other data request initiated in response to user interaction, and/or whether the second application is more data intensive than the first application. The flow chart continues at 'A' in **FIG. 7** if none of the above conditions apply.

[00184] If any of the above conditions apply, then in process 626, the first data transfer request of the first application and the other data transfer request of the second application are transferred in a single transfer operation over the network. Note that the conditions shown in 616-624 can be applied in any order or any combination with one another, although illustrated in the example flow chart as having a particular order.

[00185] **FIG. 7** depicts an example of triggering events that would cause a data request to be transferred without alignment with another data request.

[00186] The events shown in processes 702-710 can occur independently or in conjunction with one another and cause the data request to be transferred over the wireless network without or with minimal delay upon occurrences of these events. For example, in process 702, the mobile device radio is turned on due to another event. In process 704, a certain time period has elapsed, in process 706, a user trigger is detected, and/or the first application exits in process 708 and/or moves into the background in process 710. When any of the above conditions are detected, in process 721, the data request is transferred over the wireless (cellular or others) network.

[00187] **FIG. 8** depicts a flow chart illustrating an example process for using timing characteristics of data requests made by individual applications to delay transfer of one or more of the data requests made by one of the individual applications.

[00188] In processes 802 and 804, data transfer requests made by first and second application on the mobile device are tracked. For example, the transaction/request

manager 235 of the local proxy 275 shown in the example of **FIG. 2A** can detect the occurrence of the data transfer requests and request that the application behavior detector 236 begin to track correlations or other types of behaviors of the requesting applications.

[00189] In process 806, a first timing characteristic of data transfer requests made by the first application is determined and in process 808, a second timing characteristic of data transfer requests made by the second application is determined, and can be used, for example, by the correlation detector 238 to identify any correlations in the requests.

[00190] Based on the identified timing characteristics and any determined correlations or applicable priorities (e.g., as determined by the prioritization engine 238 in the example of **FIG. 2A**). In process 810, the transfer of the first data request can be delayed using the first and second timing characteristics or in process 812, the transfer of the second data request can be delayed using the first and second timing characteristics. Such delay (e.g., determined by the delay module 258 of the local proxy shown in the example of **FIG. 2B**) can be until another event occurs, until a manual (user) trigger is detected, or after a certain amount of time, or when another triggering event (e.g., user interaction) occurs. Note that the delays are user-configurable and can be tracked and factored into consideration by the delay module 258.

[00191] While the example is illustrated and described for different applications (e.g., first and second applications are different applications), the process can similarly be applied to different requests within the same application (e.g., first and second applications may be the same application).

[00192] **FIG. 9** depicts a flow chart illustrating an example process for using application behavior of multiple applications to align their content requests made over the network.

[00193] In process 902, application behavior of multiple applications are detected on a mobile device. In general, the behaviors of any number of applications on the mobile device that are detected can be tracked. Any or all mobile applications on a device can be monitored for the potential for its requests to be aligned for traffic coordination. In addition, the user can select the applications to be aligned, or specify applications not to be tracked for traffic coordination. Furthermore, the device platform, manufacturer, OS

settings, and/or network provider may have additional specifications or conditions for aligning traffic requests and selection of applications for traffic coordination.

[00194] In process 904, some of the content requests made by at least a portion the multiple applications from the mobile device over the network, are aligned. Process flow continues to step 'B' as shown in the example of **FIG. 10** which depicts an example of processes through which the time of delay 'D' for content requests can be determined to align content requests over the wireless network. In process 906, some of the content requests that are delayed in a single transfer operation are transferred over the network.

[00195] Example processes applied to determine delay time 'D' include, by way of example but not limitation: Determine priority of a specific application or priority of the application relative to other applications in process 1002, Determine an amount of data involved in a given content request in step 1004, Determine the useable lifetime of data to be transferred in the content request in step 1006, Determine a nature of data involved in a given content request in step 1008, Determine a status of the application making the content request in step 1010, and/or Determine the network characteristics when the request is made in step 1012. Any number of the above conditions can be applied in any order. Additional conditions which can be used may not be illustrated in the example above. In process 1014, any user configuration or overriding settings can be factored into consideration in determining delay to align content request, in process 1016.

[00196] **FIG. 11** shows a diagrammatic representation of a machine in the example form of a computer system within which a set of instructions, for causing the machine to perform any one or more of the methodologies discussed herein, may be executed.

[00197] In alternative embodiments, the machine operates as a standalone device or may be connected (e.g., networked) to other machines. In a networked deployment, the machine may operate in the capacity of a server or a client machine in a client-server network environment, or as a peer machine in a peer-to-peer (or distributed) network environment.

[00198] The machine may be a server computer, a client computer, a personal computer (PC), a user device, a tablet PC, a laptop computer, a set-top box (STB), a personal digital assistant (PDA), a cellular telephone, an iPhone, an iPad, a Blackberry, a

processor, a telephone, a web appliance, a network router, switch or bridge, a console, a hand-held console, a (hand-held) gaming device, a music player, any portable, mobile, hand-held device, or any machine capable of executing a set of instructions (sequential or otherwise) that specify actions to be taken by that machine.

[00199] While the machine-readable medium or machine-readable storage medium is shown in an exemplary embodiment to be a single medium, the term "machine-readable medium" and "machine-readable storage medium" should be taken to include a single medium or multiple media (e.g., a centralized or distributed database, and/or associated caches and servers) that store the one or more sets of instructions. The term "machine-readable medium" and "machine-readable storage medium" shall also be taken to include any medium that is capable of storing, encoding or carrying a set of instructions for execution by the machine and that cause the machine to perform any one or more of the methodologies of the presently disclosed technique and innovation.

[00200] In general, the routines executed to implement the embodiments of the disclosure, may be implemented as part of an operating system or a specific application, component, program, object, module or sequence of instructions referred to as "computer programs." The computer programs typically comprise one or more instructions set at various times in various memory and storage devices in a computer, and that, when read and executed by one or more processing units or processors in a computer, cause the computer to perform operations to execute elements involving the various aspects of the disclosure.

[00201] Moreover, while embodiments have been described in the context of fully functioning computers and computer systems, those skilled in the art will appreciate that the various embodiments are capable of being distributed as a program product in a variety of forms, and that the disclosure applies equally regardless of the particular type of machine or computer-readable media used to actually effect the distribution.

[00202] Further examples of machine-readable storage media, machine-readable media, or computer-readable (storage) media include but are not limited to recordable type media such as volatile and non-volatile memory devices, floppy and other removable disks, hard disk drives, optical disks (e.g., Compact Disk Read-Only Memory (CD

ROMS), Digital Versatile Disks, (DVDs), etc.), among others, and transmission type media such as digital and analog communication links.

[00203] Unless the context clearly requires otherwise, throughout the description and the claims, the words “comprise,” “comprising,” and the like are to be construed in an inclusive sense, as opposed to an exclusive or exhaustive sense; that is to say, in the sense of “including, but not limited to.” As used herein, the terms “connected,” “coupled,” or any variant thereof, means any connection or coupling, either direct or indirect, between two or more elements; the coupling of connection between the elements can be physical, logical, or a combination thereof. Additionally, the words “herein,” “above,” “below,” and words of similar import, when used in this application, shall refer to this application as a whole and not to any particular portions of this application. Where the context permits, words in the above Detailed Description using the singular or plural number may also include the plural or singular number respectively. The word “or,” in reference to a list of two or more items, covers all of the following interpretations of the word: any of the items in the list, all of the items in the list, and any combination of the items in the list.

[00204] The above detailed description of embodiments of the disclosure is not intended to be exhaustive or to limit the teachings to the precise form disclosed above. While specific embodiments of, and examples for, the disclosure are described above for illustrative purposes, various equivalent modifications are possible within the scope of the disclosure, as those skilled in the relevant art will recognize. For example, while processes or blocks are presented in a given order, alternative embodiments may perform routines having steps, or employ systems having blocks, in a different order, and some processes or blocks may be deleted, moved, added, subdivided, combined, and/or modified to provide alternative or subcombinations. Each of these processes or blocks may be implemented in a variety of different ways. Also, while processes or blocks are at times shown as being performed in series, these processes or blocks may instead be performed in parallel, or may be performed at different times. Further any specific numbers noted herein are only examples: alternative implementations may employ differing values or ranges.

[00205] The teachings of the disclosure provided herein can be applied to other systems, not necessarily the system described above. The elements and acts of the various embodiments described above can be combined to provide further embodiments.

[00206] Any patents and applications and other references noted above, including any that may be listed in accompanying filing papers, are incorporated herein by reference. Aspects of the disclosure can be modified, if necessary, to employ the systems, functions, and concepts of the various references described above to provide yet further embodiments of the disclosure.

[00207] These and other changes can be made to the disclosure in light of the above Detailed Description. While the above description describes certain embodiments of the disclosure, and describes the best mode contemplated, no matter how detailed the above appears in text, the teachings can be practiced in many ways. Details of the system may vary considerably in its implementation details, while still being encompassed by the subject matter disclosed herein. As noted above, particular terminology used when describing certain features or aspects of the disclosure should not be taken to imply that the terminology is being redefined herein to be restricted to any specific characteristics, features, or aspects of the disclosure with which that terminology is associated. In general, the terms used in the following claims should not be construed to limit the disclosure to the specific embodiments disclosed in the specification, unless the above Detailed Description section explicitly defines such terms. Accordingly, the actual scope of the disclosure encompasses not only the disclosed embodiments, but also all equivalent ways of practicing or implementing the disclosure under the claims.

[00208] While certain aspects of the disclosure are presented below in certain claim forms, the inventors contemplate the various aspects of the disclosure in any number of claim forms. For example, while only one aspect of the disclosure is recited as a means-plus-function claim under 35 U.S.C. §112, ¶6, other aspects may likewise be embodied as a means-plus-function claim, or in other forms, such as being embodied in a computer-readable medium. (Any claims intended to be treated under 35 U.S.C. §112, ¶6 will begin with the words “means for”.) Accordingly, the applicant reserves the right to add additional claims after filing the application to pursue such additional claim forms for other aspects of the disclosure.

CLAIMS

What is claimed is:

1. A method of coordinating traffic among applications on a mobile device, the method, comprising:
delaying transfer of a first data transfer request initiated by a first application until another data transfer request initiated by a second application is detected on the mobile device;
transferring, the first data transfer request of the first application and the other data transfer request of the second application in a single transfer operation over a network.
2. The method of claim 1, further comprising,
tracking data transfer requests made by the first and second applications on the mobile device;
determining a first timing characteristic of data transfer requests made by the first application and a second timing characteristic of data transfer requests made by the second application;
using the first and second timing characteristics in delaying the transfer of the first data request.
3. The method of claim 1, wherein, the second application is of higher priority than the first application; or the second application is more data intensive than the first application.
4. The method of claim 1, wherein, the second application is running in the foreground.
5. The method of claim 1, wherein, a user is interacting with the second application and the other data request is initiated in response to the user interaction.

6. The method of claim 1, wherein, a user is not interacting with the first application.

7. The method of claim 1, wherein, the first and second applications are the same application.

8. The method of claim 1, determining an amount of time with which the first data transfer request is delayed based on whether a response to the first data transfer request has changed.

9. A method of management of application traffic on a mobile device, the method, comprising:

using application behavior of multiple applications on a mobile device to align some of the content requests made by at least a portion of the multiple applications from the mobile device over the network;

wherein, the some of the content requests are aligned by delaying a time at which the some of the content requests would occur without alignment, and transferring the some of the content requests that are delayed in a single transfer operation over the network.

10. The method of claim 9, wherein, the using the application behavior includes detecting correlation in data transfer time that occurs for a first application as compared to a second application.

11. The method of claim 10, wherein, operation of the first application triggers an operation of the second application.

12. The method of claim 9, wherein, the some of the content request identified for alignment exhibit a correlation in the times at which data transfer occurs.

13. The method of claim 9, wherein the time that is delayed to align each of the some of the content requests is determined to optimize a number of content requests able to be aligned in the single transfer operation.

14. The method of claim 9, wherein the time that is delayed in transfer of a given content request is determined based on priority of a specific application making the given content request, or based on the priority of the specific application relative to other applications on the mobile device.

15. The method of claim 9, wherein the time that is delayed in transfer of a given content request is determined based on amount of data involved in the given content request.

16. The method of claim 9, wherein the time that is delayed in transfer of a given content request is determined based on a nature of data involved in the given content request.

17. The method of claim 9, wherein the time that is delayed in transfer of a given content request is determined based on a status of a specific application making the given content request, the status indicating whether the specific application is operating in the background or foreground on the mobile device.

18. The method of claim 9, wherein the time that is delayed in transfer of a given content request is determined based on a status of a specific application making the given content request, the status indicating whether the specific application is active or inactive.

19. The method of claim 9, wherein the time that is delayed in transfer of a given content request is determined based on usable lifetime of data to be transferred in the given content request.

20. The method of claim 9, wherein the time that is delayed in transfer of a given content request is determined based on network characteristics including available bandwidth or network latency.

21. The method of claim 9, wherein, the time that is delayed is user-configurable.

22. The method of claim 9, wherein, the single transfer occurs when user interaction with the mobile device results in content request needing to be made over the network.

23. A system with a distributed proxy for cross application traffic management in a wireless network, the system, comprising:

a mobile device having a local proxy which detects a first data request made via a first application and a second data request made via a second application;

a proxy server coupled to the mobile device and one or more servers to which the first and second data requests are directed, the proxy server being able to communicate with the local proxy,

wherein, transfer of the first data transfer request made via the first application to the proxy server is delayed by the local proxy until another data transfer request made via the second application is detected;

wherein, the local proxy transfers the first data transfer request and the second data transfer request in a single transfer operation over the network.

24. The system of claim 23, wherein, the first application is configured by a user of the mobile device to be eligible for delayed data transfer requests.

25. The system of claim 23, wherein, the local proxy determines that the second application is of a higher priority relative to the first application.

26. The system of claim 23, wherein, the proxy server is aware of the priority of the first application relative to the second application.

27. A system with a distributed proxy for cross application traffic management in a wireless network, the system, comprising:

a mobile device having a local proxy which detects a first data request made via a first application and a second data request made via a second application;

a proxy server coupled to the mobile device and one or more servers to which the first and second data requests are directed, the proxy server being able to communicate with the local proxy,

wherein, transfer of the first data transfer request made via the first application to the proxy server is delayed by the local proxy until another data transfer request made via the second application is detected by the proxy server;
wherein, the proxy server transfers the first data transfer request and the second data transfer request in a single transfer operation over the network.

28. The system of claim 27, wherein, the proxy server transfers the first data transfer request and the second data transfer request in a single transfer operation over the network when content for the second data request for the second application has changed.

29. A system for management of application traffic on a mobile device, the system, comprising:

means for, using application behavior of multiple applications on a mobile device to align some of the content requests made by at least a portion the multiple applications from the mobile device over the network;

means for, aligning the some of the content requests by delaying a time at which the some of the content requests would occur without alignment;

wherein the time that is delayed in aligning each of the some of the content requests is determined to optimize a number of content requests able to be aligned in the single transfer operation;

means for, transferring the some of the content requests that are delayed in a single transfer operation over the network.

30. The system of claim 29, further comprising, means for, determining the time that is delayed based on a status of a specific application making the given content request, the status indicating whether the specific application is in the background or foreground.

31. The system of claim 29, further comprising, means for, determining the time that is delayed based on network characteristics including available bandwidth or network latency.

32. A machine-readable storage medium having stored thereon instructions which when executed by a processor performs a method of coordinating traffic among applications on a mobile device, the method, comprising:

delaying transfer of a first data transfer request initiated by an application until a second data transfer request initiated by the application is detected on the mobile device;

determining that the second data request is initiated in response to the user interaction;

transferring, the first data transfer request and the second data transfer request of the application a single transfer operation over the network.

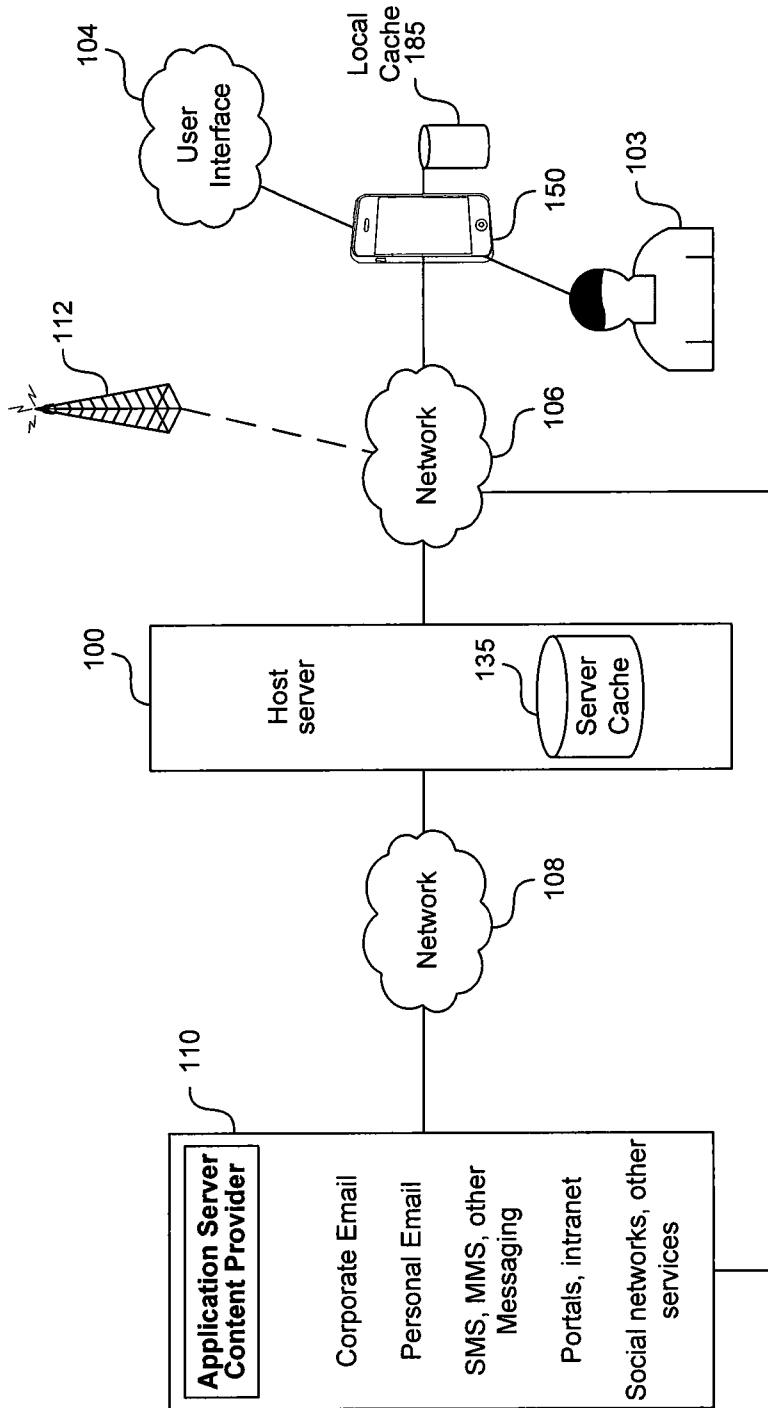


FIG. 1A

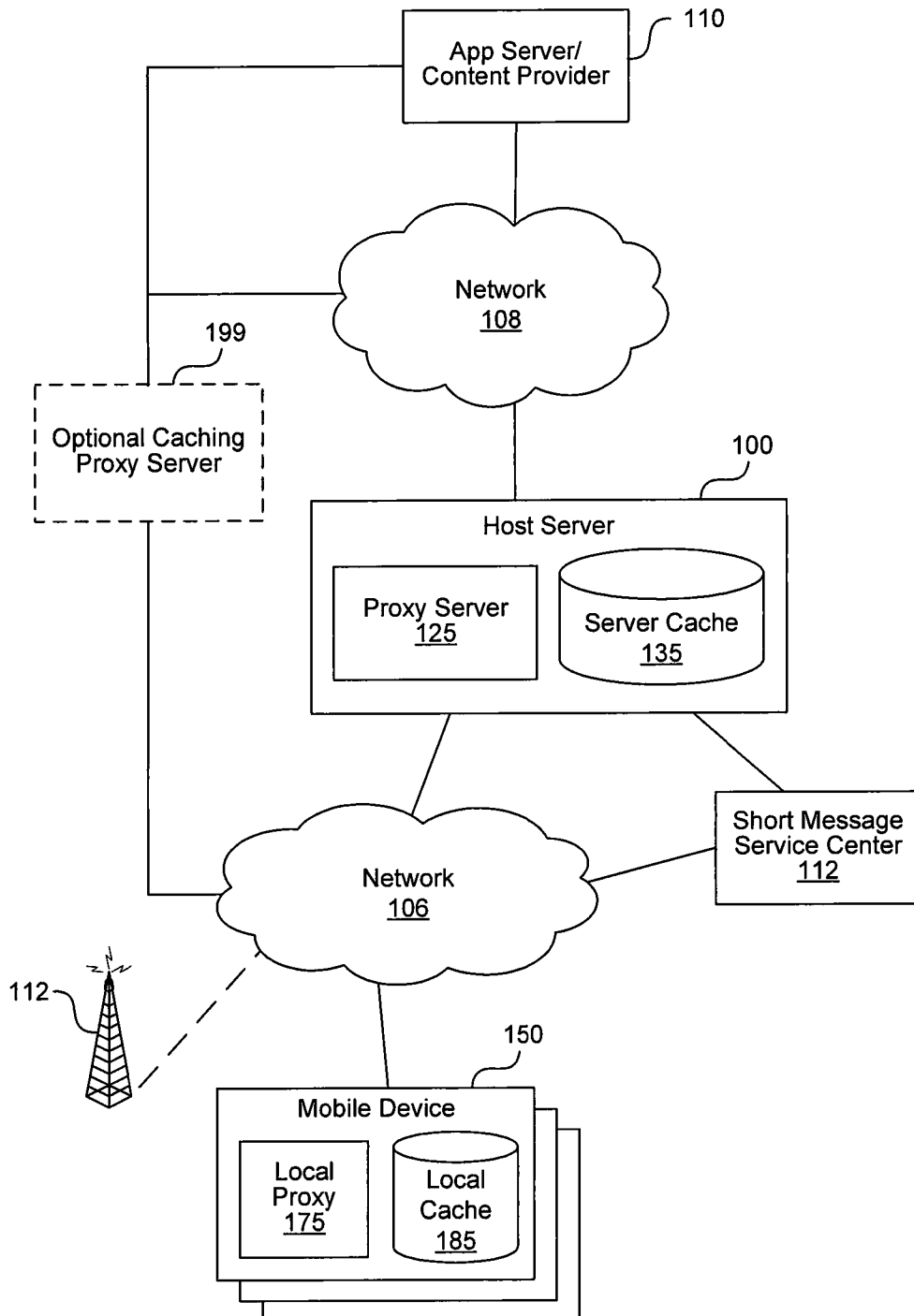


FIG. 1B

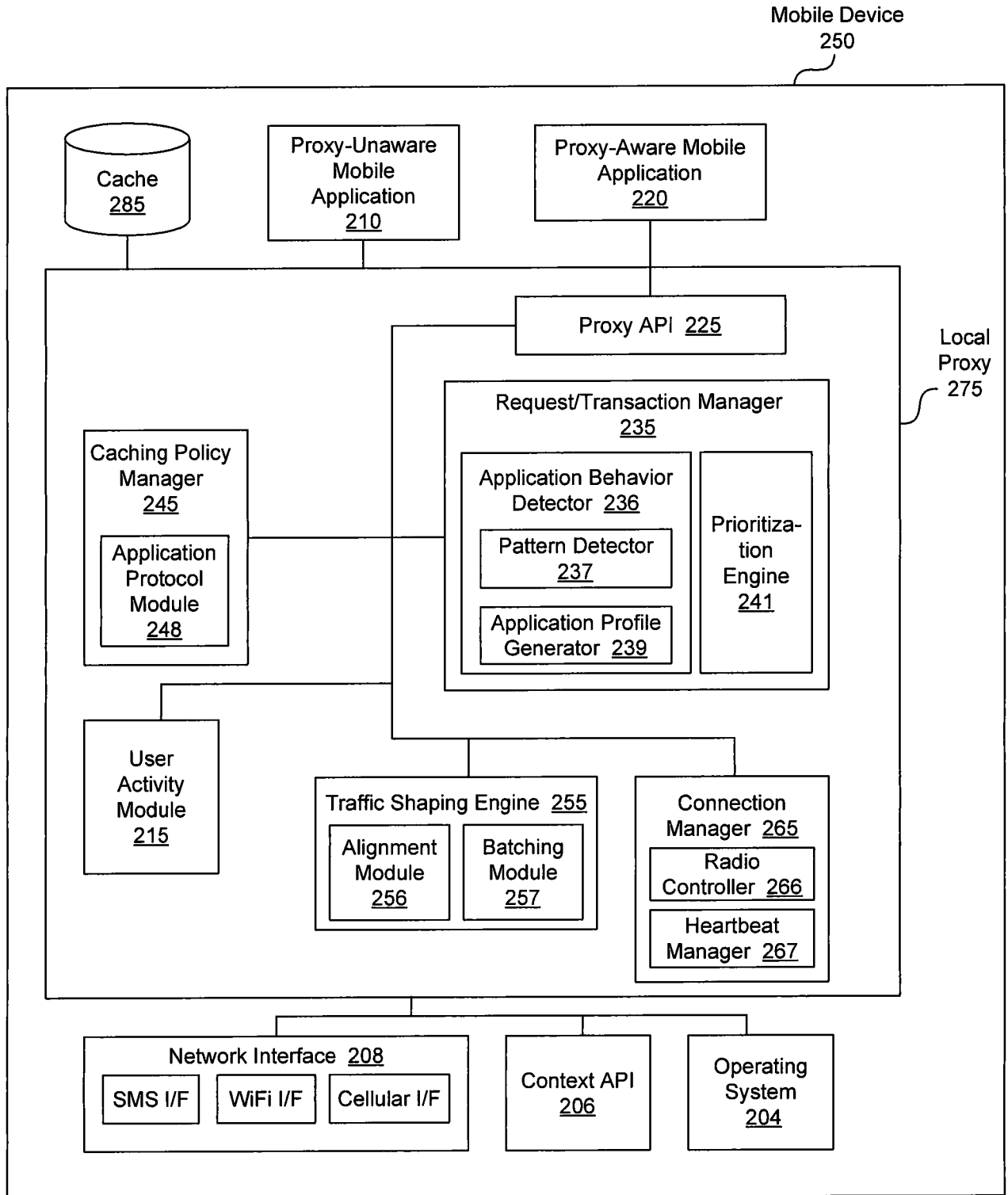


FIG. 2A

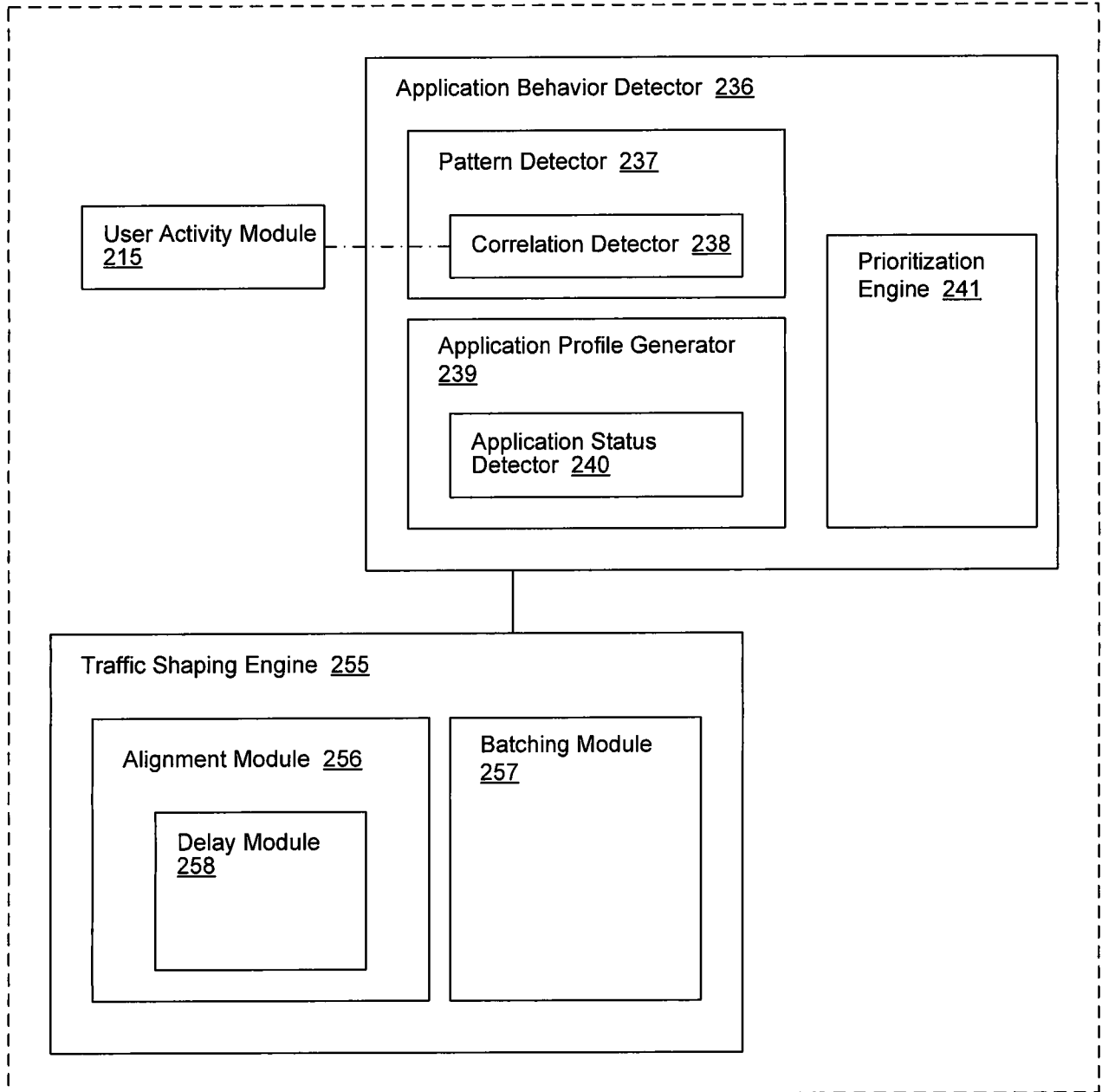


FIG. 2B

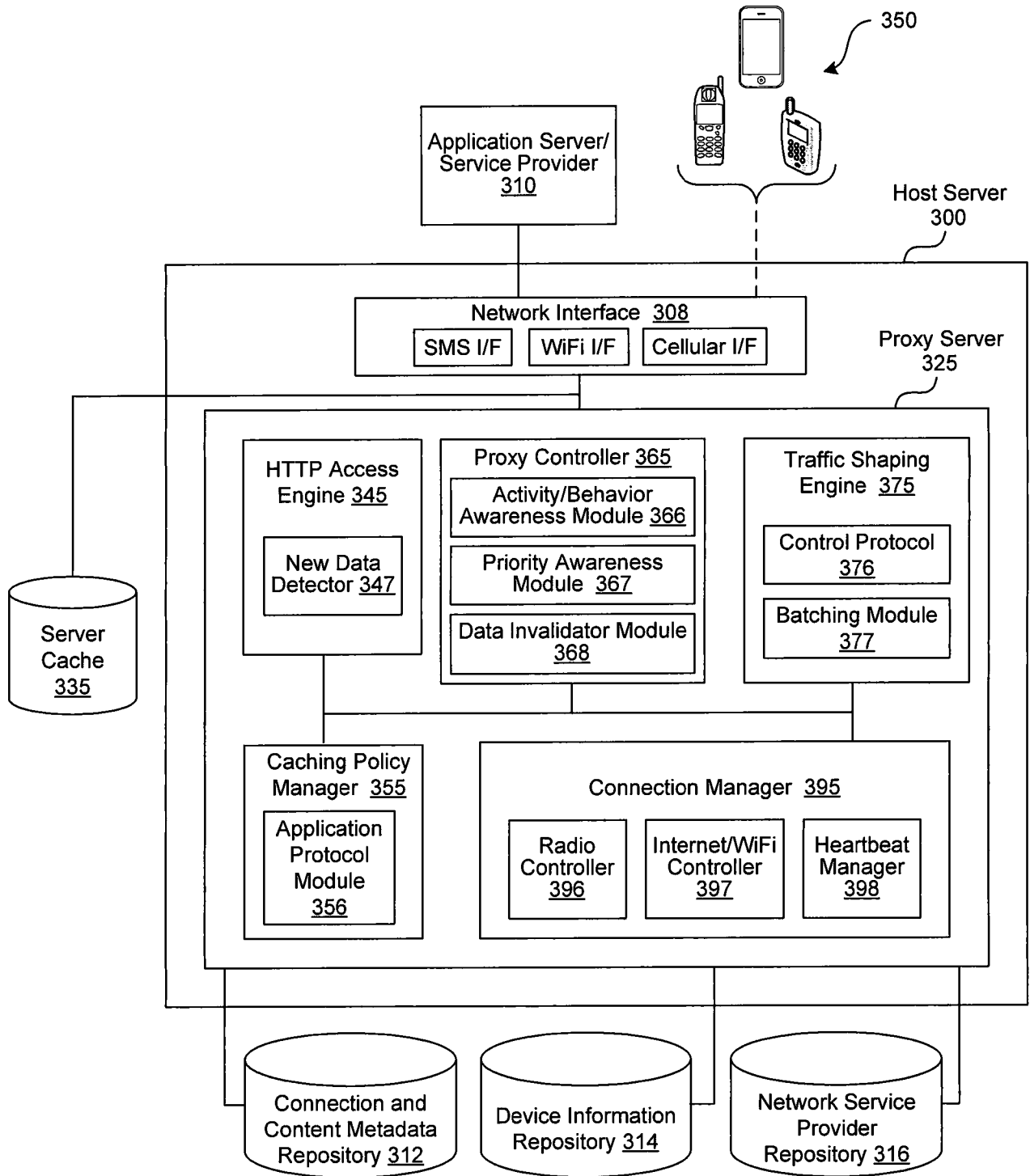


FIG. 3

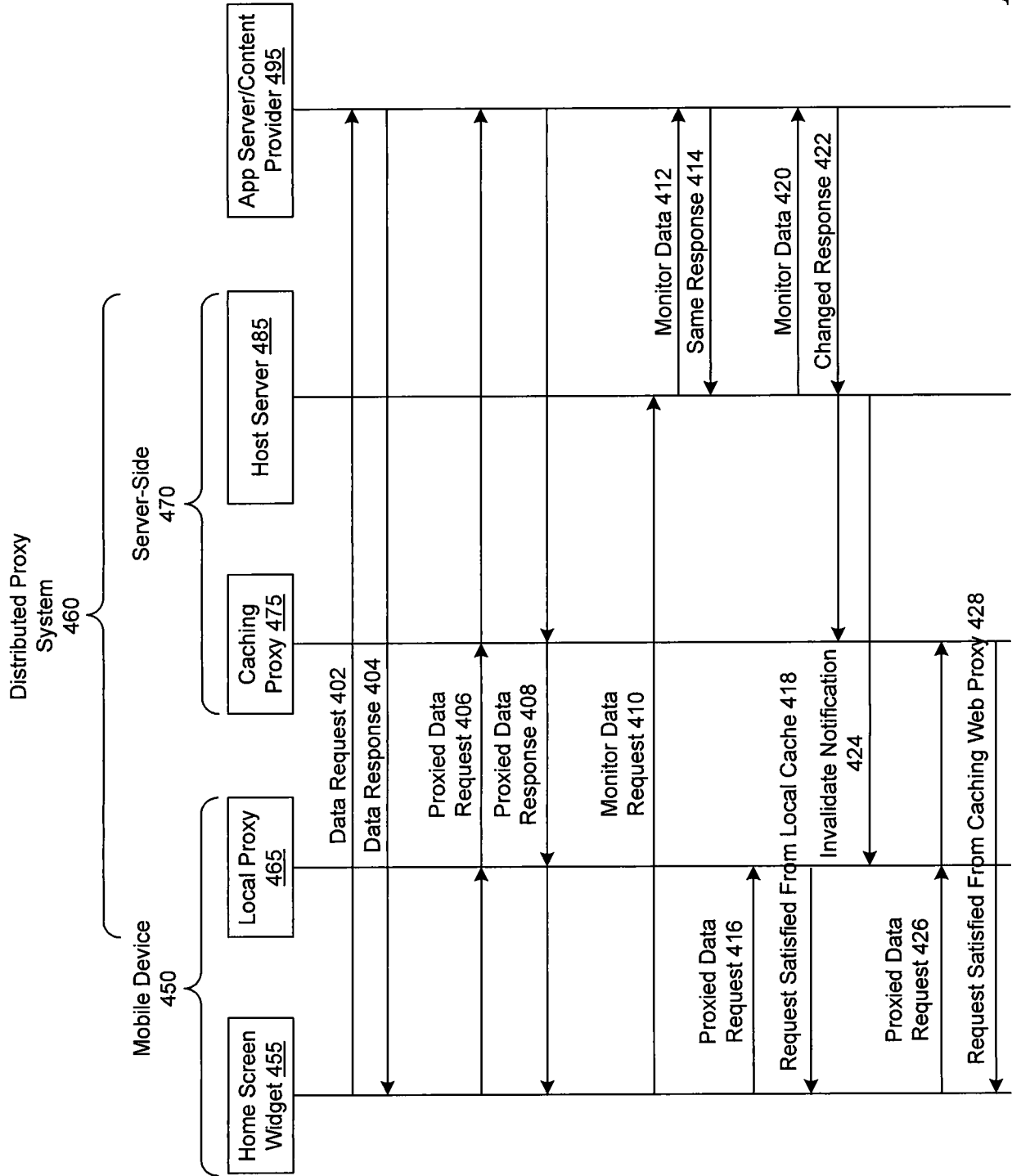


FIG. 4A

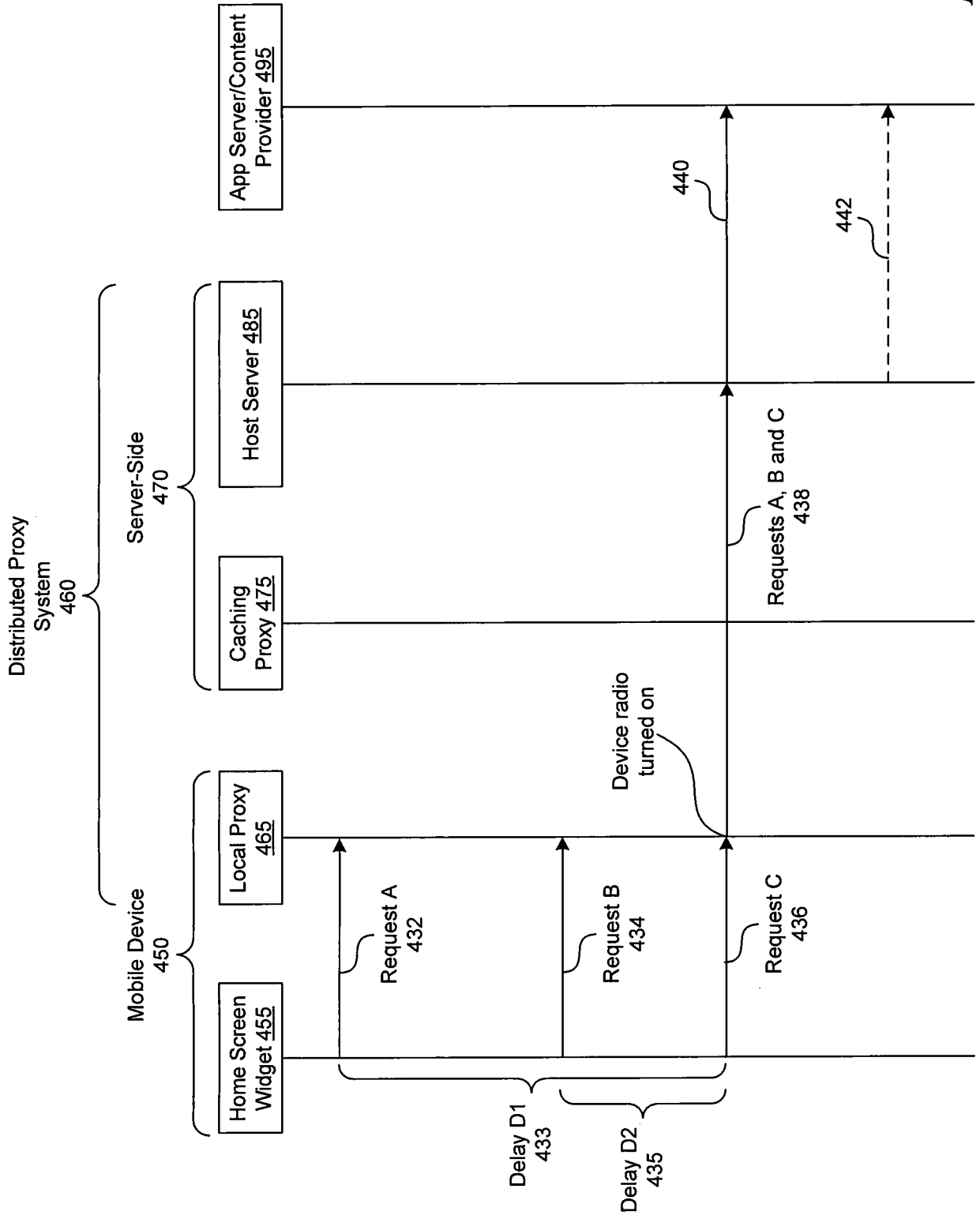


FIG. 4B

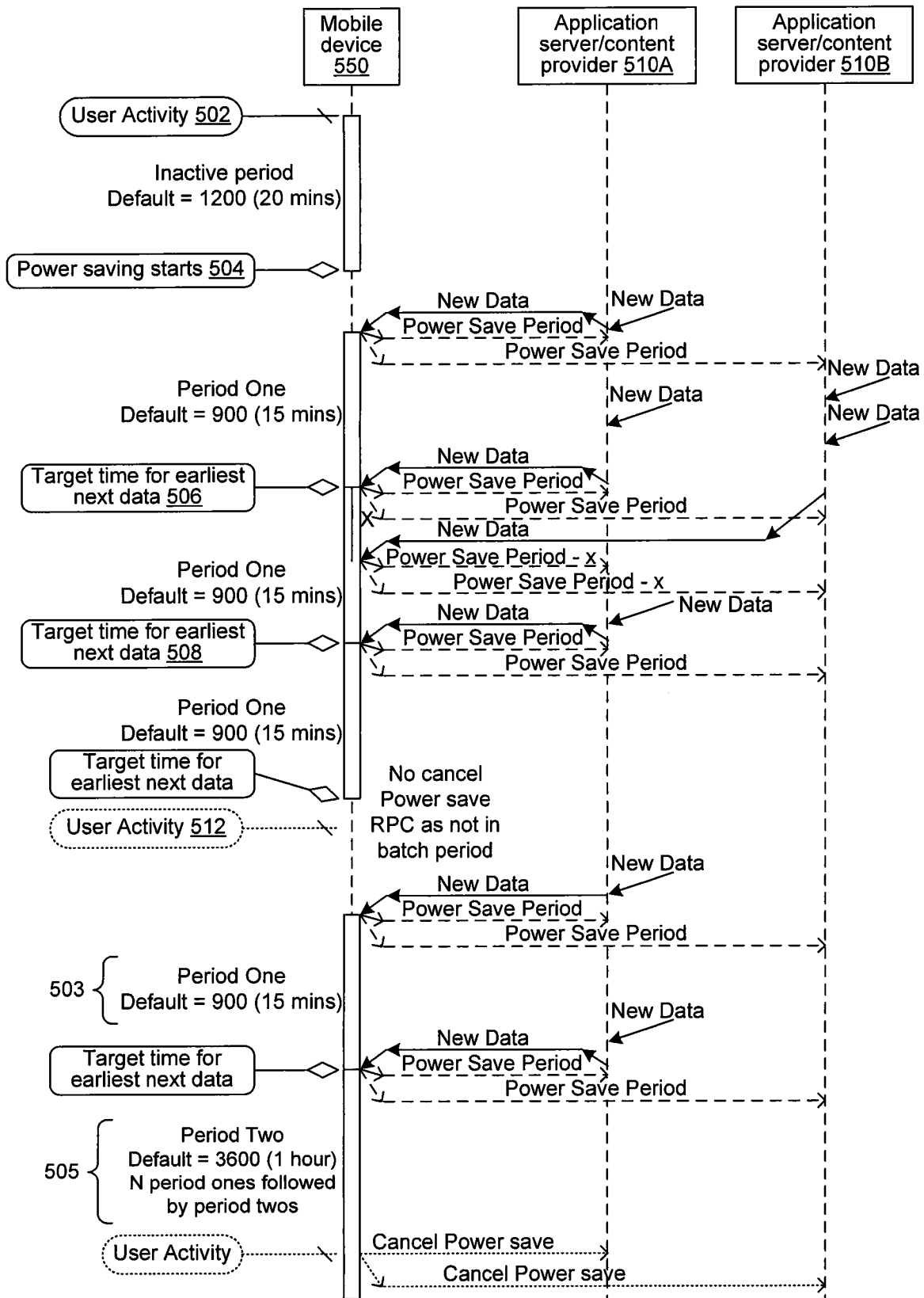


FIG. 5

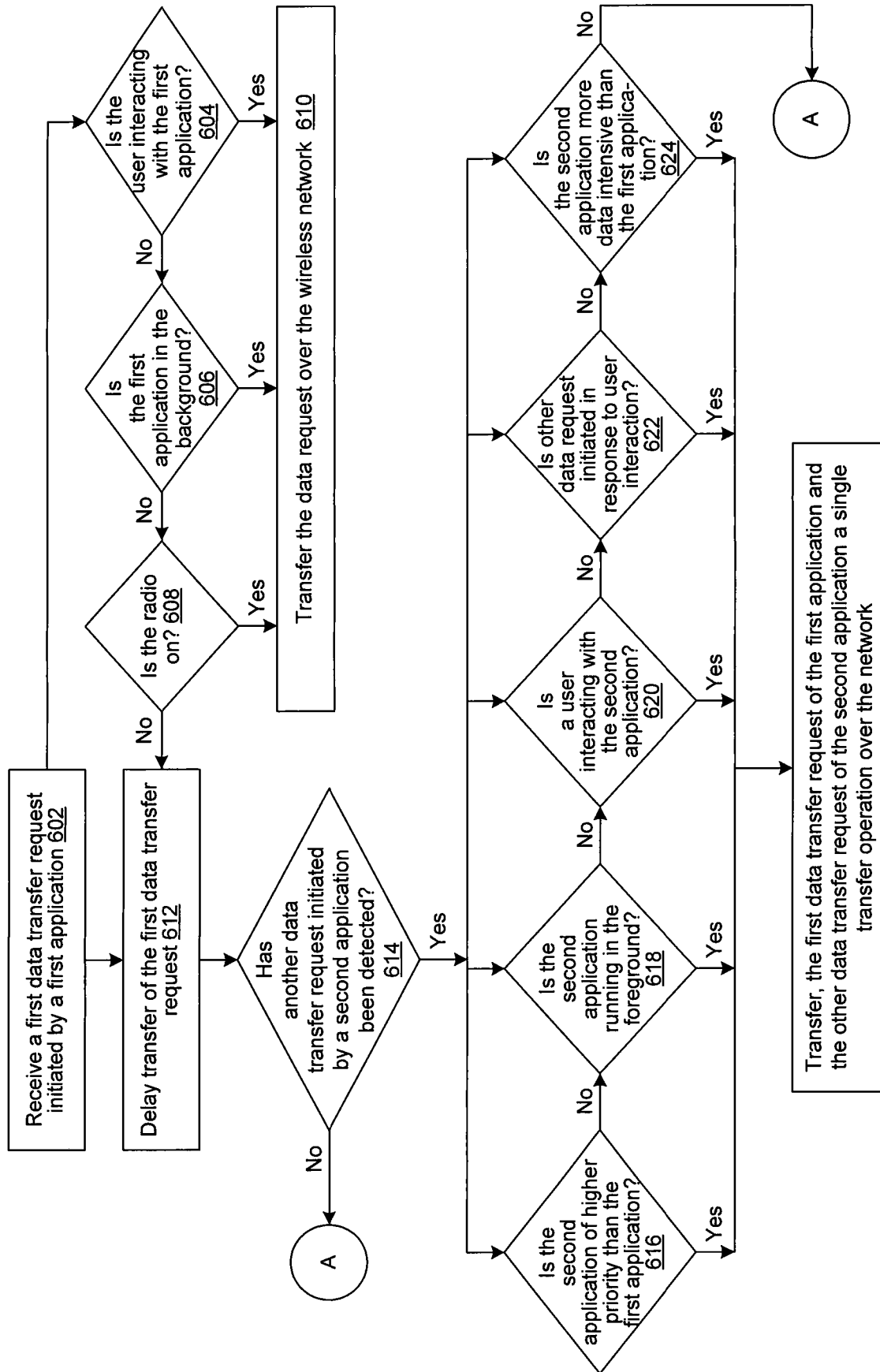


FIG. 6

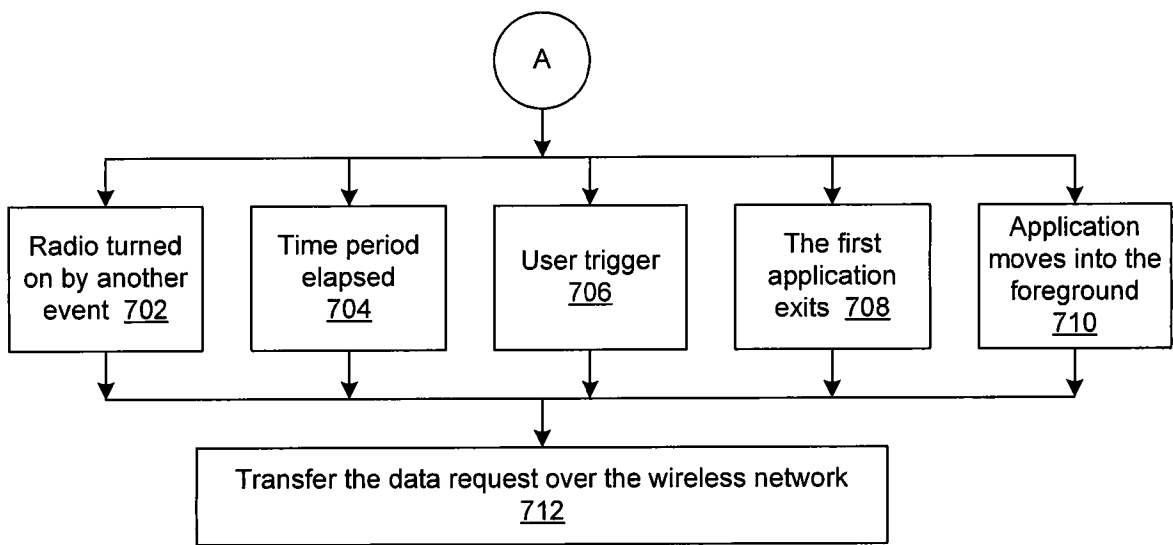
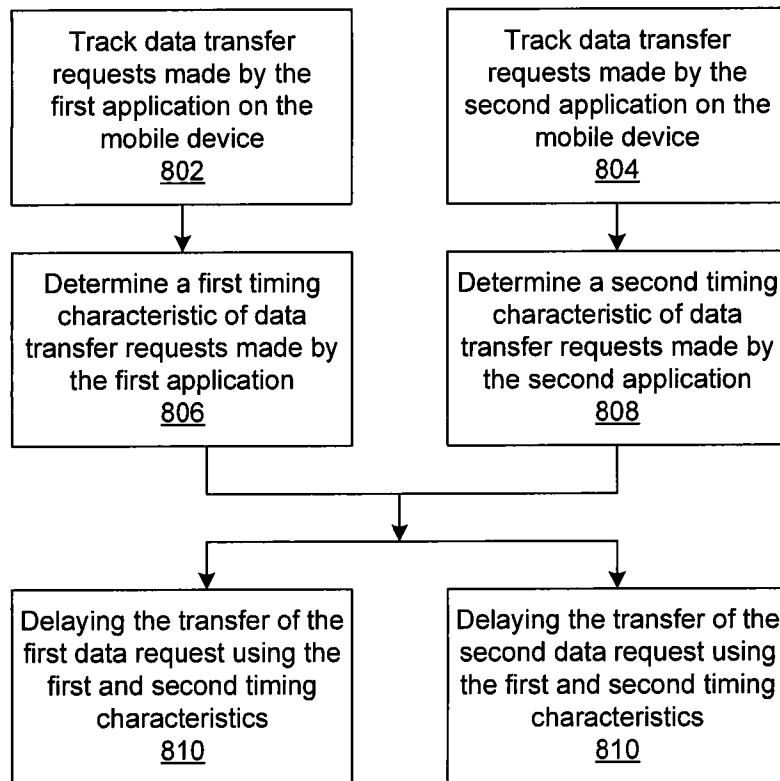


FIG. 7

**FIG. 8**

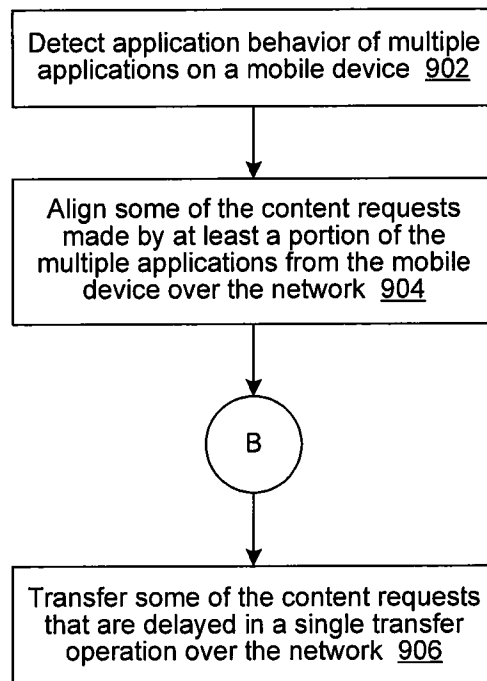


FIG. 9

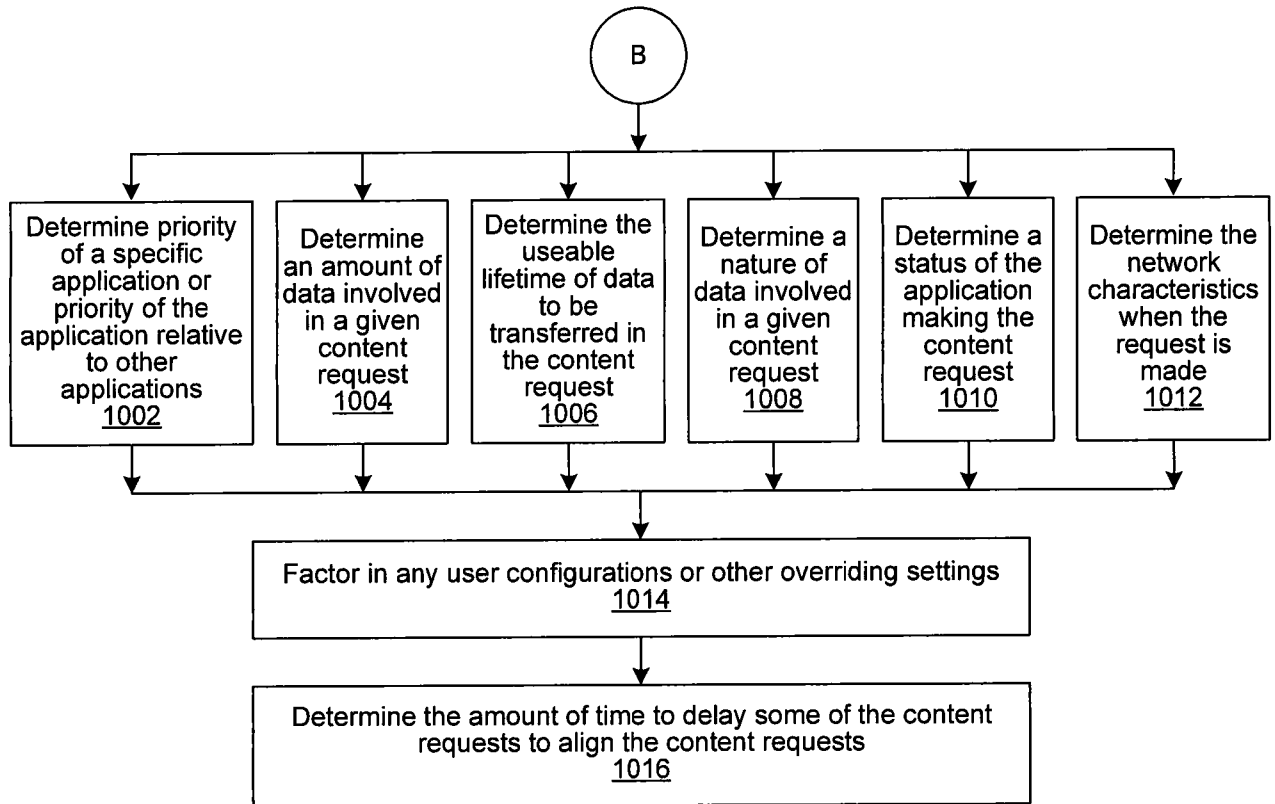


FIG. 10

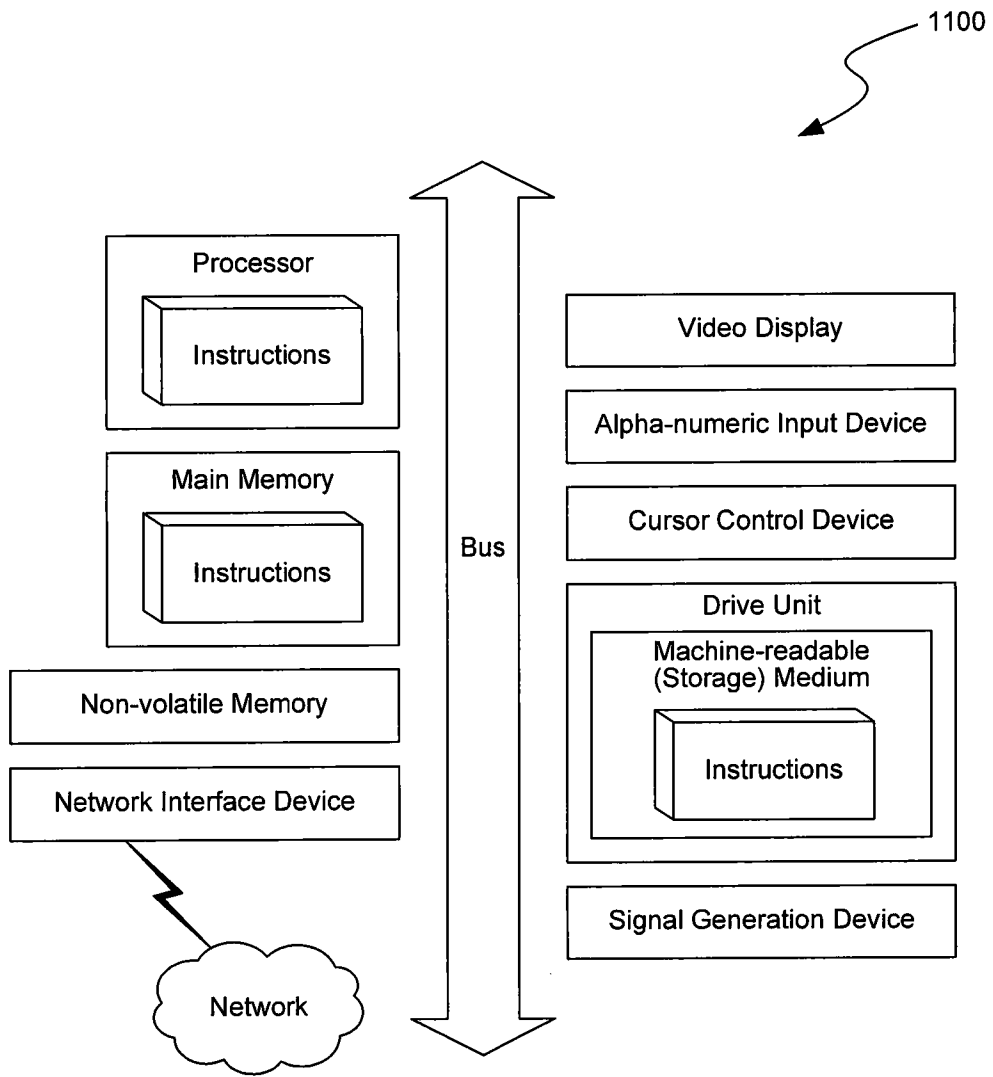


FIG. 11

A. CLASSIFICATION OF SUBJECT MATTER**G06F 15/16(2006.01)i, H04W 88/02(2009.01)i, H04W 28/02(2009.01)i**

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F 15/16; G06F 17/30; H04L 12/28; H04L 12/56

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models

Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS(KIPO internal) & Keywords: mobile, application, request, delay

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2007-0288469 A1 (SHENFIELD MICHAEL) 13 December 2007 See abstract; paragraphs [0023]-[0094] and figures 1-6.	1-32
A	US 2006-0209842 A1 (CREAMER THOMAS E. et al.) 21 September 2006 See abstract; paragraphs [0018]-[0047] and figures 1-2.	1-32
A	US 2010-0077083 A1 (TRAN AHN P. et al.) 25 March 2010 See abstract; paragraphs [0012]-[0044] and figures 1-6.	1-32
A	US 2010-0131617 A1 (OSBORNE JOHN et al.) 27 May 2010 See abstract; paragraphs [0024]-[0033] and figures 1-3.	1-32

 Further documents are listed in the continuation of Box C. See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

30 DECEMBER 2011 (30.12.2011)

Date of mailing of the international search report

02 JANUARY 2012 (02.01.2012)

Name and mailing address of the ISA/KR

Korean Intellectual Property Office
Government Complex-Daejeon, 189 Cheongsa-ro,
Seo-gu, Daejeon 302-701, Republic of Korea

Facsimile No. 82-42-472-7140

Authorized officer

CHOI, Jung Kwon

Telephone No. 82-42-481-8507



INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2011/037932

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2007-0288469 A1	13.12.2007	None	
US 2006-0209842 A1	21.09.2006	CN 100547993 C CN 1832470 A0 US 7957271 B2	07.10.2009 13.09.2006 07.06.2011
US 2010-0077083 A1	25.03.2010	CN 102165818 A EP 2340673 A2 KR 10-2011-0061578 A US 7966410 B2 WO 2010-036768 A2 WO 2010-036768 A3	24.08.2011 06.07.2011 09.06.2011 21.06.2011 01.04.2010 01.04.2010
US 2010-0131617 A1	27.05.2010	None	