



(51) International Patent Classification:

H04W 48/18 (2009.01) H04W 72/02 (2009.01)
H04W 88/06 (2009.01) H04W 36/36 (2009.01)

(21) International Application Number:

PCT/US2019/035082

(22) International Filing Date:

31 May 2019 (31.05.2019)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/678,810 31 May 2018 (31.05.2018) US

(71) Applicant: **MOBOPHILES, INC.** [US/US]; DBA

Mobilize, 2716 Ocean Park Blvd., Suite 3055, Santa Monica, CA 90405 (US).

(72) Inventors: **CHOW, William, Weiyeh**; 2716 Ocean Park Blvd., Suite 3055, Santa Monica, CA 90405 (US). **TSUIE, Mark, Lea**; 2716 Ocean Park Blvd., Suite 3055, Santa Monica, CA 90405 (US). **TRUONG, Brian, Alex**; 2716 Ocean Park Blvd., Suite 3055, Santa Monica, CA 90405 (US).

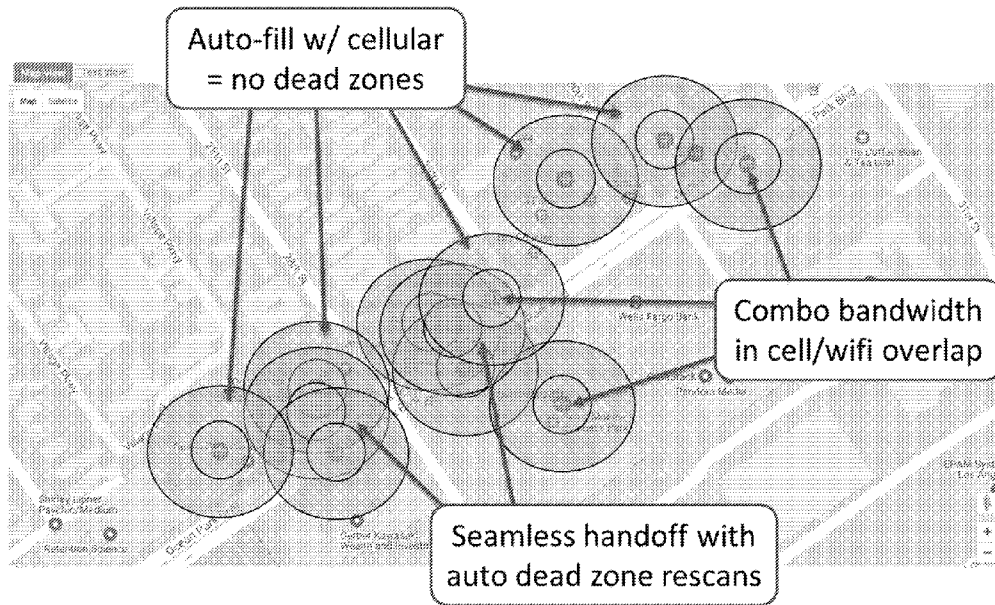
(74) Agent: **LEE, Shaun, P.**; Lewis Roca Rothgerber Christie LLP, P.O. Box 29001, Glendale, CA 91209 (US).

(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, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA,

(54) Title: SYSTEMS AND METHODS FOR DYNAMIC CHANNEL BONDING

FIG. 7



(57) Abstract: A method for managing network traffic on a portable communication device including a processor, a memory, and a plurality of network interfaces configured to connect to a plurality of networks includes: processing network traffic of an application running on the processor via a first network of the plurality of networks, the first network being specified as a primary network by an operating system running on the processor; monitoring a plurality of network status information associated with the first network; detecting a problem of the first network when one or more parameters of the received network status information is outside of one or more threshold values; in response to detecting the problem in the first network, selecting a second network of the plurality of networks as the primary network; and processing the network traffic via the second network as the updated primary network.



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, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, 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, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

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

1 **SYSTEMS AND METHODS FOR DYNAMIC CHANNEL BONDING**

CROSS-REFERENCE TO RELATED APPLICATIONS

5 **[0001]** This application claims the benefit of U.S. Provisional Patent Application No. 62/678,810, filed in the U.S. Patent and Trademark Office on May 31, 2018, the entire disclosure of which is incorporated by reference herein.

FIELD

10 **[0002]** Aspects of embodiments of the present invention relate to the field of computer networking.

BACKGROUND

15 **[0003]** Connection bonding or link aggregation relates to methods for combining or aggregating multiple network connections.

20 **[0004]** Under some network conditions, network connectivity on computing devices can be poor, such as when the computing devices are near or beyond the usable ranges of the networks' reach, or near or beyond the usable limits of the networks' capacity. For example, when a smartphone that is connected to a wireless local area network or WLAN (e.g., a Wi-Fi network) is moved to the edge of the effective range of the wireless network (such as when sitting in a car in the driveway at home), the smartphone will generally continue to try to transmit and request data via the wireless network rather than switch to its cellular connection, even when the connection to the wireless network is poor. This may result in a poor user
25 experience, such as a temporary loss of network connectivity when a user is at the edges of local wireless networks, even when cellular network signal is strong. Another example is when a smartphone is connected to a wireless network that has more traffic than it can handle, such as a congested Wi-Fi network in a busy coffee shop or in a stadium during a sporting event. This may result in a poor user
30 experience, such as hangs or timeouts when trying to access internet content.

35 **[0005]** Examples of WLAN connections include networks such as those based on IEEE 802.11 standards, also known as Wi-Fi. Examples of cellular network connections include 3G wireless cellular networks (sometimes referred to as, for example, Universal Mobile Telecommunications Service or UMTS, Global Systems for Mobile or GSM, and Code Division Multiple Access 2000 or CDMA2000) or 4G wireless cellular networks (sometimes referred to as, for example, Long Term Evolution Advanced or LTE Advanced) and the like.

1 [0006] There are additional issues with connecting to other types of networks like
Wi-Fi, such as: performance (e.g., may be slower than the cellular connection),
accessibility (e.g., may require login before granting internet access), or quality (e.g.,
there may be another access point with a better signal or performance). Typically,
5 these issues are not handled well or at all by the device (including the software
running on it), so users often have to manually manage these connections, including
disabling or disconnecting Wi-Fi on their device to manually force the device to use
the cellular connection. This results in a very poor user experience (e.g., requiring
the user to remember when to turn Wi-Fi back on) and has additional consequences,
10 such as higher cellular costs for either the user (e.g., when the user has a metered
data plan) or the mobile network operator (e.g., when user has an unlimited data
plan).

[0007] Devices can also connect to other network types, such as Bluetooth, or
possibly future network types, any of which can cause similar issues at their edges.
15 In general, when there are multiple, possibly different, heterogeneous networks that
a device can connect to, there are challenges in how and when the device should
use each of those available networks.

Related Art

20 [0008] The issues with managing the usage of multiple, possibly heterogeneous
networks that are available to a device have seen other related approaches in
addressing them. These range from providing user interface (UI) settings to enable
manual control by the user, to software-based approaches to automate the
connection to the network (e.g., Passpoint® for Wi-Fi), to controlling how and when
25 multiple available networks are used (e.g., Multipath TCP).

[0009] For example, some comparative network protocols have been designed to
leverage multiple networks concurrently, such as Multipath TCP (MPTCP), by
splitting network traffic across the multiple available networks and then using an
associated intermediary to recombine the split traffic downstream, enabling possibly
30 higher performance (e.g., by load balancing data across cellular and Wi-Fi networks
concurrently) and higher resiliency (e.g., by retrying/resending lost Wi-Fi packets to
the cellular network). There are many such network protocols developed over the
years, including proprietary ones, that follow this integrated client-server model of
splitting traffic at the client and then having a server (or equivalent intermediary)
35 downstream to recombine the traffic. However, these approaches require all device
traffic to be directed through an intermediary, which can have a significant impact
(e.g., high operating cost). This cost would generally be borne by the operator of one
of the networks being used by the device, such as the mobile network operator

1 (MNO), where the additional cost includes new servers and additional bandwidth (e.g., of the Wi-Fi traffic) that the MNO did not previously need to handle while that user's device was connected to the Wi-Fi network.

5 **[0010]** Another comparative approach taken is for an application to be multi-network aware, such that the application will direct its traffic over both Wi-Fi and/or cellular, either simultaneously (for performance via load balancing) or selectively (for resiliency against one network not working). For example, the Samsung® Download Booster can speed up large file downloads (>30 MB) via load balancing across cellular and Wi-Fi. However, this approach does not scale to provide general
10 improvement to network performance across different applications (e.g., this approach does not improve web browsing or watching videos) because it is only used specifically for file downloads performed by certain apps such as a web browser). These approaches also lack the awareness to dynamically adjust the load between the available networks based on the dynamically changing performance
15 characteristics of each network. Another comparative example is the Wi-Fi Assist feature in Apple® iOS, which can switch an app to use cellular if it detects that the Wi-Fi is unresponsive. However, this implementation needs network requests to hang or timeout before switching from Wi-Fi to cellular because it lacks the ability to quickly and seamlessly detect (e.g., in near real-time) which network is actually
20 responsive. Also, the Apple® Wi-Fi Assist feature is restricted to the "foreground" app (i.e., the application running in the foreground that the user is currently interacting with) because it lacks policy control for either the user or MNO to control how much cellular data is used.

25 **SUMMARY**

[0011] Aspects of embodiments of the present invention relate to systems and methods for automatically managing multiple network connections in a computing device to improve network connectivity.

30 **[0012]** According to one embodiment of the present invention, a method for managing network traffic on a portable communication device including a processor, a memory, and a plurality of network interfaces configured to connect to a plurality of networks includes: intercepting, by a traffic manager running on the processor, network data to and from an application running on the processor; transmitting, by the traffic manager, an idempotent request of the network data to a server via the
35 plurality of networks; receiving, by the traffic manager, a response to the idempotent request from the server via a first network on the plurality of networks; and selecting, by the traffic manager, the first network of the plurality of networks to use for receiving and transmitting the response to the application.

- 1 **[0013]** The transmitting the idempotent request to the server via the plurality of networks may include transmitting the idempotent request on a second network of the plurality of networks, and the method may further include terminating the idempotent request on the second network.
- 5 **[0014]** The transmitting the idempotent request to the server via the plurality of networks may include: transmitting the idempotent request to the server via one network of the plurality of networks; and transmitting the idempotent request to the server via one or more other networks of the plurality of networks in accordance with a delay.
- 10 **[0015]** The delay may be shorter than an application-level timeout of the application running on the portable communication device.
- [0016]** The delay may be set based on a typical response time to the idempotent request.
- [0017]** The idempotent request may be associated with a network protocol, and the delay may be set based on the network protocol associated with the idempotent request.
- 15 **[0018]** The delay may be set based on a size of a response to the idempotent request.
- [0019]** The delay may be set based on a network quality metric.
- 20 **[0020]** The plurality of networks may be arranged in a preference ranking, and the one or more other networks may be selected in accordance with the preference ranking.
- [0021]** Each of the networks may be associated with a different delay in accordance with the preference ranking.
- 25 **[0022]** The method may further include rearranging the plurality of networks in the preference ranking in accordance with one or more dynamic factors.
- [0023]** The one or more dynamic factors may include network performance.
- [0024]** The one or more dynamic factors may include network traffic cost.
- [0025]** The plurality of networks may include a plurality of different types of networks.
- 30 **[0026]** The types of networks may include one or more of: cellular, Bluetooth, and Wi-Fi networks.
- [0027]** According to one embodiment of the present invention, a method for managing network traffic on a portable communication device including a processor, a memory, and a plurality of network interfaces configured to connect to a plurality of networks includes: processing network traffic of an application running on the processor via a first network of the plurality of networks, the first network being specified as a primary network by an operating system running on the processor;
- 35

1 monitoring a plurality of network status information associated with the first network;
detecting a problem of the first network when one or more parameters of the
received network status information is outside of one or more threshold values; in
response to detecting the problem in the first network, selecting a second network of
5 the plurality of networks as the primary network; and processing the network traffic
via the second network as the updated primary network.

[0028] The network traffic may include a request and a response to the request,
and wherein the problem in the first network is detected based on a response time
between a timestamp of the request and a timestamp of the response exceeding a
10 maximum threshold.

[0029] The detecting the problem in the first network may include: monitoring at
least one of network statistics on the first network; and detecting the problem when a
change in the network statistics exceeds a threshold.

[0030] The network statistics may include a packet loss rate or a bad packet rate.

15 **[0031]** The first network may be a wireless network, and the detecting the
problem in the first network may include: monitoring a signal strength of the wireless
network; and detecting the problem when the signal strength drops below a
threshold signal strength.

[0032] The second network of the plurality of networks may be selected in
20 accordance with a preference ranking of the plurality of networks.

[0033] The problem in the first network may be detected based on responses
received on other networks before a response is received on the first network.

[0034] According to one embodiment of the present invention, a portable
communication device includes a processor, a memory, and a plurality of network
25 interfaces configured to connect to a plurality of networks, the memory storing
instructions that, when executed by the processor, cause the processor to manage
network traffic of the portable communication device by: intercepting, by a traffic
manager running on the processor, network data to and from an application running
on the processor; transmitting, by the traffic manager, an idempotent request of the
30 network data to a server via the plurality of networks; receiving, by the traffic
manager, a response to the idempotent request from the server via a first network on
the plurality of networks; and selecting, by the traffic manager, the first network of the
plurality of networks to use for receiving and transmitting the response to the
application.

35 **[0035]** The transmitting the idempotent request to the server via the plurality of
networks may include transmitting the idempotent request on a second network of
the plurality of networks, and the instructions may further include instructions that
cause the processor to terminate the idempotent request on the second network.

1 **[0036]** The transmitting the idempotent request to the server via the plurality of
networks may include: transmitting the idempotent request to the server via one
network of the plurality of networks; and transmitting the idempotent request to the
server via one or more other networks of the plurality of networks in accordance with
5 a delay.

[0037] The delay may be shorter than an application-level timeout of the
application running on the portable communication device.

[0038] The delay may be set based on a typical response time to the idempotent
request.

10 **[0039]** The idempotent request may be associated with a network protocol, and
the delay may be set based on the network protocol associated with the idempotent
request.

[0040] The delay may be set based on a size of a response to the idempotent
request.

15 **[0041]** The delay may be set based on a network quality metric.

[0042] The plurality of networks may be arranged in a preference ranking, and the
one or more other networks may be selected in accordance with the preference
ranking.

[0043] Each of the networks may be associated with a different delay in
20 accordance with the preference ranking.

[0044] The memory may further store instructions that, when executed by the
processor, cause the processor to rearrange the plurality of networks in the
preference ranking in accordance with one or more dynamic factors.

[0045] The one or more dynamic factors may include network performance.

25 **[0046]** The one or more dynamic factors may include network traffic cost.

[0047] The plurality of networks may include a plurality of different types of
networks.

[0048] The types of networks may include one or more of: cellular, Bluetooth, and
Wi-Fi networks.

30 **[0049]** According to one embodiment of the present invention, a portable
communication device includes a processor, a memory, and a plurality of network
interfaces configured to connect to a plurality of networks, the memory storing
instructions that, when executed by the processor, cause the processor to manage
network traffic of the portable communication device by: processing network traffic of
35 an application running on the processor via a first network of the plurality of
networks, the first network being specified as a primary network by an operating
system running on the processor; monitoring a plurality of network status information
associated with the first network; detecting a problem of the first network when one

1 or more parameters of the received network status information is outside of one or
more threshold values; in response to detecting the problem in the first network,
selecting a second network of the plurality of networks as the primary network; and
5 processing the network traffic via the second network as the updated primary
network.

[0050] The network traffic may include a request and a response to the request,
and the problem in the first network may be detected based on a response time
between a timestamp of the request and a timestamp of the response exceeding a
maximum threshold.

10 **[0051]** The instructions for detecting the problem in the first network may include
instructions that, when executed by the processor, cause the processor to: monitor
at least one of network statistics on the first network; and detect the problem when a
change in the network statistics exceeds a threshold.

[0052] The network statistics may include a packet loss rate or a bad packet rate.

15 **[0053]** The first network may be a wireless network, and the detecting the
problem in the first network may include: monitoring a signal strength of the wireless
network; and detecting the problem when the signal strength drops below a
threshold signal strength.

[0054] The second network of the plurality of networks may be selected in
20 accordance with a preference ranking of the plurality of networks.

[0055] The problem in the first network may be detected based on responses
received on other networks before a response is received on the first network.

[0056] Aspects embodiments of the present invention relate to providing
improved performance by sending requests on different networks based on an active
25 request count, including: distributing load based on ratio of active requests between
different networks; dynamically adjusting ratio of active requests based on factors
including: latency and bandwidth; detecting and tracking active requests of encrypted
traffic, including detecting new active requests via presence of outbound/request
data and request completion if sufficiently long delay since last inbound/response
30 data.

[0057] According to one embodiment of the present invention, a method of
network traffic management on a portable communication device having a computer
processor and a plurality of network interfaces for transmitting or receiving data to a
computer server includes: intercepting, by a traffic manager application, electronic
35 traffic of the first data to or from a first application; tracking, by the traffic manager
application, of the number of active data requests or responses for each network
interface; and selecting, by the traffic manager application, the network interface to

1 use for the delivery of first data to or from the first application to or from the server based on the number of active data exchanges.

5 **[0058]** According to one embodiment of the present invention, a method for managing traffic on a portable communication device including a processor, a memory, and a plurality of network interfaces configured to connect to a plurality of networks includes: intercepting, by a traffic manager running on the processor, network data to and from an application running on the processor; transmitting, by the traffic manager, a plurality of requests to a server, the requests being transmitted via the plurality of networks; tracking, by the traffic manager, a plurality of numbers of active data requests associated with each network interface; and selecting, by the traffic manager, a network interface of the plurality of network interfaces for transmitting the network data to the server and for receiving the network data from the server in accordance with the plurality of numbers of active data requests associated with the network interfaces.

15 **[0059]** One of the plurality of networks may be designated as a primary network, and the selecting the network interface may include: determining a number of active data requests associated with the primary network exceeds a threshold value; and selecting the network interface corresponding to another network of the plurality of networks for which the number of active data requests does not exceed the threshold value.

20 **[0060]** The selecting the network interface of the plurality of network interfaces may include: selecting a second network interface of the plurality of network interfaces; calculating a ratio of active requests in accordance with a number of active data requests associated with the network interface; and a number of active data requests associated with the second network interface; and distributing network traffic of the network data between the network interface and the second network interface in accordance with the ratio of active requests.

25 **[0061]** The method may further include dynamically adjusting the ratio of active requests in accordance with a plurality of network conditions of the plurality of networks.

30 **[0062]** The network conditions may include latency measurements corresponding to the plurality of networks.

[0063] The network conditions may include maximum bandwidth measurements corresponding to the plurality of networks.

35 **[0064]** The method may further include: detecting when a first network associated with the network interface or a second network associated with the second network interface is an unresponsive network; and dropping the unresponsive network.

1 [0065] Aspects of embodiments of the present invention relate to connection
management, including: switching to another network when bonding (e.g., doubletap
or load-balancing) detects degradation of the primary/default network; switching to
another network based on doubletap (e.g. when it switches off of the primary
5 network); and switching to another network based on load-balancing, e.g. when
active requests are not completing quickly enough on the current primary network

BRIEF DESCRIPTION OF THE DRAWINGS

10 [0066] The accompanying drawings, together with the specification, illustrate
exemplary embodiments of the present invention, and, together with the description,
serve to explain the principles of the present invention.

[0067] FIG. 1 is a schematic diagram of an example portable communication
device architecture, such as a smartphone, suitable for use with an on-device
channel bonding implementation according to an embodiment of the present
15 invention.

[0068] FIG. 2 is a block diagram of the software components within a proxy to
enable monitoring and control of network traffic between an app and a server
according to embodiments of the present invention.

20 [0069] FIG. 3 is a high level flowchart diagram illustrating a method for setting a
primary network according to one embodiment of the present invention.

[0070] FIG. 4 is a high level flowchart diagram illustrating a method for leveraging
advance signals to control network access according to one embodiment of the
present invention.

25 [0071] FIG. 5 is a high level flowchart diagram illustrating an example method of
the load-balancing approach according to one embodiment of the present invention.

[0072] FIG. 6 shows an example of a neighborhood with multiple Wi-Fi networks
where each one has inner green circular region indicating the area of good signal
quality and an outer red circular region indicating the area of bad signal quality (i.e.
the Wi-Fi dead zone).

30 [0073] FIG. 7 shows the same example as FIG. 6 but leveraging the techniques
according to embodiments of the present invention.

DETAILED DESCRIPTION

35 [0074] In the following detailed description, only certain exemplary embodiments
of the present invention are shown and described, by way of illustration. As those
skilled in the art would recognize, the invention may be embodied in many different

1 forms and should not be construed as being limited to the embodiments set forth herein.

5 **[0075]** One challenge in mobile networking is that handoff between different networks is often not seamless, such as in the case of heterogeneous networks. Use of heterogeneous networks commonly occurs when switching between different types of networks supported by the mobile device, such as switching between Wi-Fi and cellular. For example, when a mobile device is joined/attached to a Wi-Fi network, the device will stay on that Wi-Fi network and use it for network communications, even when the Wi-Fi signal is too weak (or poor) to be used for reliable communications, because it is difficult to detect that the Wi-Fi is unusable until all attempts to use it fail (e.g., after a long timeout). This state is often known as the “dead zone,” where the device has no internet connectivity since it is still connected to and using (or attempting to use) a network that is not usable. This problem also occurs when switching between networks of the same type because there is often no coordination between those networks. For example, when the device is in the dead zone of one Wi-Fi network but there is another, better and usable Wi-Fi network within range, the device will not switch to the other Wi-Fi network without some kind of explicit coordination between those Wi-Fi networks. An example of explicit coordination is an “enterprise” mesh network with multiple Wi-Fi access points, typically operated by the company, and leverages handoff techniques, such as IEEE 802.11r or 802.11v. However, the user is often using unrelated Wi-Fi networks, such as in public spaces where there may be one or more overlapping or adjoining Wi-Fi networks, which are not aware of one another (e.g., not cooperating), and thus, are unable to proactively handoff a device between them. In other words, this is another case of heterogeneous networks preventing seamless handoffs.

25 **[0076]** One challenge in distributing load across multiple networks is in determining the optimal distribution of load when each network has different performance characteristics. For example, a simple load-balancing policy can often yield no improvement in performance since overall performance is dictated by the slowest path. In Table 1, below, it is assumed that a Wi-Fi Network has a speed of 40 mbps and a cellular network has a speed of 20 mbps. The table shows various calculations for elapsed time to download a 40 MB file and a 100 MB using various distributions of the load between the faster, Wi-Fi connection and the slower, cellular connection:

35

1
5
10
15

Wi-Fi speed	Cell speed	File size	LB %	Elapsed Time
40 mbps (5 MB/s)	20 mbps (2.5 MB/s)	40 MB	100% Wi-Fi	8 s
			50/50: 20 MB on Wi-Fi = 4 s 20 MB on cell = 8 s	8 s
			Proportional: 27 MB on Wi-Fi = 5.4 s 13 MB on cell = 5.2 s	5.2 s
		100 MB	100% Wi-Fi	20 s
			50/50: 50 MB on Wi-Fi = 10 sec 50 MB on cell = 20 sec	20 s
			Proportional: 67 MB on Wi-Fi = 13.4 s 33 MB on cell = 13.2 s	13.4 s

Table 1

[0077] Table 1 illustrates that it's desirable to distribute load proportionally to the relative performance of the available networks to be able to achieve actual performance gains, e.g., a Wi-Fi network that is twice as fast as the cellular network should get twice as much traffic to significantly reduce the total wait time.

[0078] However, one problem in properly distributing load across multiple networks is determining when that is actually needed, since it is difficult for a client device to determine when it has reached the capacity of a network without attempting to generate more network load and/or traffic than the network can handle, which rarely occurs during normal usage, such as browsing web pages. Some approaches to determine that the network's capacity ceiling has been reached is to generate test traffic, such as via a "speed test" that continuously adds multiple concurrent downloads of a large test file until the total bandwidth rate no longer increases. However, this approach cannot generally be applied during normal user activity since it creates unnecessary additional load (which may impact battery life or network usage), can interfere/compete with the user's activity, and can add additional network usage cost (e.g., when using a metered network). Also, this technique does not account for dynamic changes in the available network capacity (e.g., congestion varying based on other users on the same network), and it is generally undesirable to constantly generate test traffic to continuously recalculate the current network capacity.

1 **[0079]** Accordingly, aspects of embodiments of the present invention are directed
to approaches that have the ability to detect, in real-time (e.g., as needed, at the time
of network usage), the quality and performance of the available networks at the end
user (client) device, without requiring traffic to be split and recombined by an
5 intermediary/server and without needing to generate network traffic to test if the
current capacity of the network has been reached. Some aspects of embodiments of
the present invention relate to fine-grained awareness of the available networks such
that load-balancing can be adjusted dynamically to speed up different types of
network traffic (e.g., not just large file downloads but also web browsing and the like),
10 as well as to be able to failover from Wi-Fi to cellular quickly to avoid hangs and
timeouts at the client.

[0080] Aspects of embodiments of the present invention relate to systems and
methods for automatically managing multiple network connections in a computing
device to improve network connectivity. For the sake of convenience, embodiments
15 of the present invention will be described herein with regard to improving network
connectivity in mobile devices having both wireless local area network (e.g., WLAN
or Wi-Fi) connectivity and having cellular network (e.g., 3G GSM, 4G LTE)
connectivity. However, embodiments of the present invention are not limited thereto
and may also be applied to devices having two different network connections (e.g., a
20 wireless local area network connection and a wired local area network connection)
and devices having more than two different network connections (e.g., a third
network connection such as wired local area network connection or a satellite data
connection).

[0081] Example embodiments of the present invention will now be described with
25 reference to the accompanying drawings. In the drawings, the same or similar
reference numerals refer to the same or similar elements throughout. Herein, the use
of the term “may,” when describing embodiments of the present invention, refers to
“one or more embodiments of the present invention.” In addition, the use of
alternative language, such as “or,” when describing embodiments of the present
30 invention, refers to “one or more embodiments of the present invention” for each
corresponding item listed.

[0082] FIG. 1 is a schematic diagram of an example portable communication
device (such as a smartphone) architecture 100 suitable for use with an on-device
channel bonding implementation according to an embodiment of the present
35 invention. For purposes of illustration, the portable communication device or mobile
device 100 will be assumed to be an Android® smartphone. However, embodiments
of the present invention are not limited thereto and may be applied to other
computing devices such as smartphones and tablet computers running operating

1 systems other than Android®, laptop computers, and the like. Further, while such
mobile devices may be capable of supporting many users, for ease of description, it
will be assumed that a mobile device is dedicated to a particular user, so the term
“user” and “mobile device” (or any computing device used in a personal or portable
5 manner) may be used synonymously throughout.

[0083] According to one or more embodiments of the present invention, the
general architecture on portable communication devices or mobile devices (such as
architecture 100) provides for a centralized proxy 130 (or traffic manager) that is
configured to monitor or control the data traffic originating from applications (e.g.,
10 mobile apps, or “apps”) to, for example, an application server (or app server) 250
that the mobile device accesses, e.g., via a Wi-Fi or cellular network. This approach
enables channel bonding to be performed across multiple networks (e.g., Wi-Fi and
cellular) and across multiple apps, and allows the channel bonding to be centrally
managed, although embodiments of the present invention are not limited thereto. In
15 other embodiments, the channel bonding may be performed privately within each
app or for a particular subset of apps.

[0084] The apps and other programmable components of portable
communication device 100 may be implemented, for example, as sets of computer
instructions stored on a non-transitory storage device (e.g., flash memory 170) of
20 portable communication device 100, and configured to be executed on one or more
processors of the portable communication device 100. The proxy 130 may also
manage traffic for particular web sites, such as from a web browser. Accordingly, for
ease of description, terms such as “application,” “app,” “web site,” or “site” may be
used somewhat interchangeably throughout the present application when referring to
25 categories of content being managed by the proxy 130.

[0085] The proxy 130 may be engaged from a number of different mechanisms,
such as a proxy server (e.g., via operating system (OS) network settings) using a
sockets layer 120, a virtual private network (VPN) service 140 (e.g., via OS network
settings) using network tunnel (TUN) device 230, or embedded within an app using
30 interception layer 150. The proxy 130 may be run on a Java virtual machine (JVM),
Android® Runtime (ART), or other managed runtime environment 160 or may run
directly on the operating system without a managed runtime environment. The proxy
130 may include a cache engine 110 for managing cached content on a physical
storage device, such as flash memory 170 or other nonvolatile storage device.

35 Without loss of generality, such a device may sometimes be referred to as a “disk,”
though it could be any type of non-transitory storage device, such as a solid-state
drive (e.g., NAND flash memory). In addition, the cached or any other stored content
may be stored, in whole or in part, on volatile storage, such as dynamic random

1 access memory (DRAM), and this volatile storage may be used in combination with
nonvolatile storage, such as in a tiered manner where the most recently accessed
content is stored in faster volatile storage before it is transitioned to slower
nonvolatile storage.

5 **[0086]** The proxy 130 may run in a variety of form factors, such as an application,
kernel driver, or within the OS on the mobile device, and be configured (for example,
via OS network settings) to receive network connections from internal apps 180 and
external application servers 250. In one or more embodiments, the proxy server may
run in a managed runtime environment such as a JVM 160. The proxy 130 may act
10 as an intermediary on behalf of client applications 180. For example, the proxy 130
may service the request of an app 180 running in another managed runtime
environment such as JVM 165.

[0087] As one example of operation, the app 180 may request access the Internet
using, for example, an Android service such as HttpURLConnection 190. (Here,
15 HTTP stands for hypertext transfer protocol and URL stands for uniform resource
locator, e.g., a web address.) HttpURLConnection 190 may then invoke network
services 200 provided by the OS to access the Internet. Network services 200 may
access the Internet, for example, using access point name (APN) 210 (e.g., a mobile
network such as 3G or 4G cellular network) or Wi-Fi connection 220. Network
20 services 200 may be configured to route requests from app 180 to proxy server 130
using a proxy configuration applied globally to the system, or applied to the APN 210
or Wi-Fi connection 220, as shown by dotted lines in FIG. 1. Network services 200
may also route requests from the app 180 to the proxy 130 using a variety of other
ways, for example, via network tunnel (TUN) device 230 or IP routing tables (also
25 known as “iptables”), as shown by dashed lines in FIG. 1.

[0088] Network services 200 may be configured to specify a proxy directly or
indirectly (e.g., as a global system proxy directly detected and used by apps running
on the device, or indirectly through a setting on the APN 210 or Wi-Fi connection
220) to access the Internet, such that a request may be sent through a standard
30 communications layer, such as sockets 120 (e.g., a network socket for connecting to
the Internet), which is received by the proxy 130. The proxy 130, in turn, may make a
request to the app server 250 through the network services 200 (while bypassing the
APN or Wi-Fi proxy configuration to avoid looping back to itself) and over an external
network 240, where the app server 250 services the request and returns any
35 responding communications to the proxy 130 via the external network 240 and the
network services 200. Accordingly, the proxy 130 may then monitor or control the
communications between apps 180 and servers 250. The proxy 130 may also cache
some, none, or all of the response received from the servers 250 via the caching

1 engine 110 before returning the response through the network socket 120 to the app 180 through the same described stages in reverse (e.g., the APN 210 connection / Wi-Fi 220 connection, the network services 200, and the HttpURLConnection library 190).

5 **[0089]** Instead of using a proxy configuration on the APN or Wi-Fi connection, the network services 200 may also be configured to route requests to proxy server 130 through a variety of other means. For example, another approach is using a network tunnel (TUN) 230 to establish a VPN connection, which may route network activity to VPN service 140 to handle the network transmission. The VPN service 140 may then
10 route the request to the proxy 130 to manage the traffic between the app 180 and app server 250 using the sockets 120 (as appropriate) to service the request and return the response via the network tunnel 230.

[0090] Another mechanism for engaging the proxy 130 is to use an interception layer (such as interception layers 150 and 155) within an app to redirect traffic to the proxy process. For example, in the above example, before or in place of having
15 HttpURLConnection 190 invoke network services 200 to access the Internet, HttpURLConnection 190 may have an interception layer 150 intercept the request from app 180 and directly forward its traffic to the proxy 130. Forwarding to the proxy 130 from intercept 150 may be performed through network services 200 or using
20 standard inter-process communications (IPC) mechanism as would be apparent to one of ordinary skill in the art, such as message queues, named pipes, or shared memory.

[0091] In addition to the proxy 130 operating in a separate process, such as within JVM 160, in other embodiments, the proxy 130 may be embedded within the
25 requesting process, such as JVM 165 or Browser 185 (such as a web browser). The proxy 130 may then manage the app's network traffic without the need for any inter-process communications.

[0092] In another example, the web browser 185 seeks to access the Internet (e.g., external network 240). Similar to the app 180 above, the web browser 185 may
30 take advantage of the proxy 130 by a number of different approaches. For example, the web browser 185 may be configured to access the Internet by using network sockets 125, which could then use network services 200 to access the app server 250 and/or the proxy 130 using, for example, sockets 120 or VPN service 140 as described above. In a similar fashion, interception layer 155 may be added to the
35 web browser 185, which may then intercept the request from the web browser 185 and forward its traffic to the proxy 130.

[0093] In further detail, the above techniques may be integrated into existing interfaces, where, in some embodiments, the techniques are differentiated between

1 Secure Sockets Layer (SSL, e.g., encrypted) communications and non-SSL (e.g.,
unencrypted) communications. Integration with applications may be enabled for non-
SSL communications, for instance, in a centralized location in the network stack. For
example, proxy 130 may be configured as the proxy for all or a subset of network
5 protocols, such as only for HTTP, HTTPS, or both. Similarly, proxy 130 may be
configured as the proxy for all or a subset of network interfaces, such as for cellular,
Wi-Fi, or both. For example, for APN 210 access, the cellular access point may be
set to the proxy 130. For iptables access, the corresponding Internet Protocol (IP)
routing table entries may be set. For VPN service, the VPN client (such as VPN
10 service 140) may route traffic to the proxy 130. For Wi-Fi, the proxy 130 may be set
for each Wi-Fi access point (AP). For a global system proxy, the system may direct
traffic for all application traffic to the proxy 130.

[0094] In addition, integration with applications that use encrypted
communications, such as SSL or TLS, may require access to unencrypted network
15 data (e.g., the network data prior to its being encrypted). There are a number of
approaches that may be used here. For a man-in-the-middle approach, access to
encrypted data may be obtained by impersonating the server via a trusted certificate
authority (CA). For a software development kit (SDK) approach (such as with the
interception layer 155 in FIG. 1), build-time linking may be used with hooks to the
20 networking API above (e.g., before) the encryption layer. For a relink approach,
existing apps may be decompiled and relinked to use custom replacement
application programming interfaces (API's), such as with HttpURLConnection 190.
For a substitute approach, such as with a browser like web browser 185, an
alternative version of the app may be provided where the interception is already
25 wired in. This may be particularly appropriate for widely used open source apps.

[0095] While FIG. 1 is directed mostly to the architecture of a portable
communication device or a mobile device 100, on-device channel bonding may also
entail other components, such as software components configured to run on one or
more processors of mobile device 100. FIG. 2 is a block diagram of the software
30 components within a proxy 130 to enable monitoring and control of network traffic
between an app 280 and a server 250 according to some example embodiments of
the present invention.

[0096] In FIG. 2, the app 280 running within the mobile device 100 communicates
with app server 250, and proxy 130 would intercept the app's network traffic using
35 any of the methods previously discussed, such as through a system proxy setting or
VPN. Within proxy 130, in one or more embodiments of the present invention, there
are logical software components that perform the network traffic monitoring and
control, and these software components may include a ClientHandler 132 that may

1 handle the internal data path 133 (internal to the mobile device 100) with app 280
and a RequestHandler 134 that may handle the external network data path 135
(external to the mobile device 100, e.g., over an external network such as the
Internet) with app server 250. FIG. 2 depicts several arrows between the proxy 130
5 and the app server 250, representing different networks (e.g., different Wi-Fi
networks, different cellular networks, and the like) accessible to the portable
communication device 100 via different network interfaces of the portable
communication device 100.

[0097] The data path 133 between app 280 and ClientHandler 132 may take
10 place over different mechanisms, depending on the method used to intercept the
app's network traffic, such as a network socket or any other standard inter-process
communications mechanism as would be apparent to one of ordinary skill. The data
path 135 between RequestHandler 133 and app server 250 may also take place
over different mechanisms, depending on how the app 280 would normally
15 communicate with app server 250, such as with a network socket using TCP/IP. By
decoupling the internal data path 133 with app 280 and the external data path 135
with app server 250, this allows proxy 130 to separate the network path used by app
280 from the network path(s) used in communicating with app server 250, such as to
20 deliver data to and/or from app 280 (over internal data path 133) while using one or
more available network connections to and/or from app server 250 (over external
network data paths 135).

[0098] According to another embodiment of the present invention, a PathManager
136 may be used to manage the external network data paths 135 used by
RequestHandler 133. When there are multiple available networks that can be used,
25 PathManager 136 can implement one or more policies to direct RequestHandler on
which network to use for communicating with app server 250. For example, there
may be a policy for determining how to handle a slow network (e.g., a congested Wi-
Fi network at a busy coffee shop) by directing RequestHandler 133 to distribute (aka
load-balance) multiple requests across multiple networks (e.g., across the multiple
30 external network data paths 135). As another example, there may be a policy for
determining when to issue a data request redundantly on multiple networks to
provide a seamless and responsive data connection even though one of the
networks may be unresponsive.

[0099] Any of the aforementioned policies may be, for example, provided by the
35 user, pre-configured by the app, or received from an external system, such as a
management server. Any of the aforementioned policies may also be extended to
have multiple or variable parameters, such as which performance metrics, cost
factors, relative weightings of each parameter, etc. For example, a load-balancing

1 policy may consider the network speed/bandwidth, latency and cost per byte, to
calculate a policy on how/when to use another network (e.g., cellular) to augment the
default, primary network (e.g., Wi-Fi).

5 ***DoubleTap***

[00100] Some aspects of embodiments of the present invention relate to selecting
from among the available networks or external network data paths 135 by sending
the same request for data on more than one of the network connections or network
10 interfaces available to the computing device. Continuing with the example described
above, a smartphone may send a redundant or “doubletap” request (e.g., DNS
query, or a request for a web page) on both its wireless local area network (or Wi-Fi)
connection and its cellular network connection. The term “request” is used
generically here to refer to any network request for which there is a corresponding
15 response. To doubletap a request, the request simply needs to be idempotent (e.g.,
can be safely re-sent on another connection, or can be safely received by an app
server 250 multiple times). For example, a request for account status is generally
idempotent because no user data is changed at the app server 250 as a result of the
request, but a request to charge a credit card would generally not be idempotent
20 because the app server 250 may charge the credit card each time it received that
request. This approach can be taken for virtually any type of network traffic, since the
initial request of many network protocols is generally idempotent. Examples of such
idempotent requests include a Domain Name System (DNS) query, SYN packet for a
Transmission Control Protocol (TCP) connection, ClientHello for a Transport Layer
Security (TLS) handshake or an HTTP GET request. Examples of requests that are
25 generally not idempotent (or non-idempotent) include HTTP POST.

[00101] Embodiments of the present invention may immediately doubletap (e.g.,
send the same data request redundantly) on one or more other available networks,
such as when the cost of the redundant request/response is negligible/small (e.g., a
DNS request, where the size of the response is defined, by the protocol, to be small),
30 or it may first send the request on the preferred network (e.g., Wi-Fi) before
doubletapping the same request on any of the other available networks (e.g.,
cellular) after an initial delay, such as if the cost is large (e.g., a request that is
expected to generate a response with a large size, such as a file download or a
request that consumes significant server resources, such as establishing a TCP
35 connection). Accordingly, the length of the delay may be set based on the known or
expected size of the response to the idempotent request (e.g., short delays for small
sized responses and longer delays for large sized responses) or based on other cost
considerations (e.g., shorter delays for small costs and longer delays for higher

1 costs). The computing device 100 can then detect which connection has better
performance (e.g., which of the connections delivers a response first) and continue
using the better connection while dropping (e.g., refraining from receiving the
response and/or transmitting further requests) the other connection. In the above
5 example, when in an area with a weak Wi-Fi connection and using a doubletap
technique according to embodiments of the present invention, the computing device
100 may receive a response via the cellular connection more quickly than through
the Wi-Fi connection and may serve the results immediately. In contrast, in
10 comparative, the computing device 100 may attempt the Wi-Fi connection first and
only transmit a request along the cellular connection after the Wi-Fi connection has
not responded within a brief delay (e.g., a period much shorter than an application-
level timeout), herein referred to as the doubletap delay.

[00102] The sooner the redundant request is sent, the sooner it can be detected
whether one network is more responsive than the other. Typically, when there are
15 multiple networks, there is a preference or priority in determining which one is used.
For example, Wi-Fi is typically an unlimited or unmetered network, while cellular can
often be a metered or throttled network, so Wi-Fi may be preferred as the “primary”
network to use by default, while cellular and possibly other networks may be used
secondarily when Wi-Fi is slow or unresponsive. Different networks may have
20 different costs and performance characteristics so there may be different priorities
associated with each one, such as preferring cellular over Bluetooth, or only allowing
cellular as a secondary network when it has an unlimited data plan. In other words,
the order of preference for the available networks can be determined based on one
or more characteristics of the network, such as performance or cost. In some
25 embodiments of the present invention, the order of preference for the available
networks is set through configuration data provided by a user. Embodiments of the
present invention can leverage a combination of secondary networks in different
orders of preference to use for redundant requests.

[00103] For example, when an application requests data from a server, it may
30 send a DNS request for the IP address associated with the app server’s domain
name, followed by a TCP SYN packet to establish a connection with the app server,
possibly followed by a TLS ClientHello (e.g., if for an HTTPS request) and finally, it
sends an HTTP GET to request the data. Each of these requests is idempotent
and can be issued redundantly on the secondary network to provide resiliency
35 against any problems with the primary network. However, each of these requests
has different costs and impacts to the network and/or server. For example, DNS
requests/responses are small and stateless, so they can be sent redundantly on
multiple networks without incurring significant cost or overhead on the network or the

1 DNS servers. However, a TCP connection is a more limited resource (e.g., servers
have a limit on concurrent connections), so, in some embodiments, the Proxy 130
delays the redundant request for a TCP connection to the server (e.g., sending a
SYN packet) on the secondary network until it detects that it has not received a
5 response from the app server 250 (e.g., receiving SYN-ACK) within the typical time
via the primary network (e.g., the typical time may often be within 1-2 seconds). As
noted above, this waiting period may be referred to herein as a doubletap delay
(e.g., a period slightly longer than the typical time for a response, e.g., 2 seconds).

[00104] In general, when delaying the redundant doubletap request on the
10 secondary network(s), the request on the primary network is effectively given a small
head start and it will complete successfully most of the time (e.g., a response will
usually be received via the primary network within the doubletap delay, without the
need to send a SYN request to the server on the secondary network). But after the
doubletap delay, the redundant request is sent on the secondary network(s), so it
15 would race against the request that was sent on the primary network. Any network
for which a response is seen can be used to receive the response, and it is possible
that the response is received via more than one network. To reduce or minimize
network traffic that would be duplicative, in some embodiments the Proxy 130
selects just one (e.g., exactly one) network from which to receive the response. In
20 one embodiment of the present invention, the network from which the Proxy 130
received the response first is selected as the doubletap “winner,” and the Proxy 130
terminates the data request on the other network or networks (e.g., by closing the
other connection or connections). Terminating the doubletap “loser” or “losers”
prevents further processing and network usage (e.g., bandwidth consumption) for
25 handling the redundant response, and how a redundant request is terminated varies
based on the type of request. For example, DNS is stateless so it is sufficient to
simply ignore any redundant “loser” responses, while an HTTP request can be
terminated by closing the redundant connection on the network that lost the
doubletap race.

30 **[00105]** There are many examples of protocols that can be “doubletapped” as
described herein, such as previously described for the DNS and TCP protocols, and
it should be apparent to someone of ordinary skill in the art to see that other
protocols can similarly leverage this doubletap approach to provide redundancy. For
example, a TLS session is commonly established over a TCP connection and the
35 TLS ClientHello request can be doubletapped on the secondary network, preferably
also after a delay during which the Proxy 130 does not receive a TLS ServerHello
response from the app server 250 on the primary network within the typical response
time (e.g., also typically within 1-2 seconds). Another example is QUIC (Quick UDP

1 Internet Connection), which establishes a logical secure connection using UDP
instead of TCP and leverages the same TLS handshake, so its idempotent requests
can be sent redundantly using the doubletap technique (e.g., send the same
ClientHello redundantly on the secondary network, possibly after a delay). As
5 another example, an HTTP request is commonly performed after establishing a TCP
or QUIC connection, and any idempotent HTTP request (e.g., HTTP requests where
the method is GET, HEAD, PUT, or DELETE) can be sent redundantly on the
secondary network(s), also preferably with a delay after initially sending on the
primary network. As described in these examples, doubletap can be applied to
10 different protocol layers, such as those defined by the OSI network protocol model,
and doubletap can even be performed in more than one protocol layer for the same
application-level data request. Performing the doubletap at multiple protocol layers
provides multiple points of redundancy, where a failure (e.g., due to a bad network
signal) can occur at any of these multiple points in the client-server interaction and
15 recovered on a secondary network.

[00106] Because the redundant request on the secondary network is delayed only
by the typical response time seen for most responses (e.g., the doubletap delay
time), in some embodiments of the present invention, these doubletap delays are
much shorter thresholds than would typically be set for application timeouts, such as
20 those that result in errors when exceeded (e.g., timeouts that are set based on the
maximum or worst-case response time). Using short delays (e.g., a few seconds) for
issuing redundant requests allows these redundant requests to provide responses
from the secondary network when the primary network is bad/unresponsive without
exceeding the timeouts in the application that are used to detect network problems.
25 In some instances, this can prevent an application timeout message from being
displayed to a user (because a response is received from the secondary network). In
contrast, in comparative systems, a user might need to wait a long time until a
timeout error occurred at the application level (e.g., tens of seconds) and it would
often be unclear to the user what is the cause of the problem (e.g. an unresponsive
30 Wi-Fi network is often indistinguishable from an unresponsive server, unless
communications with that server was also attempted on another network). An
embodiment of the present invention is to configure an appropriate doubletap delay
for each of the possible requests that may be sent redundantly on the secondary
network, such that each delay allows for most responses of a given type of request
35 to be received within that timeframe whenever the primary network is good or
responsive or healthy.

[00107] FIG. 3 is a high level flowchart diagram illustrating an example method of
the doubletap approach according to one embodiment of the present invention. This

1 flowchart can be performed by an intermediary in the network data path, such as
described in FIG. 1 by proxy 130 between app 280 and app server 250. The proxy
130 may be waiting at step 310 for data requests to be sent from an app to a server.
When a data request is received from an app at step 315, the proxy determines at
5 step 320 whether the request is idempotent, to determine whether it can apply the
doubletap approach. If no, then the normal network processing by a proxy is applied
at step 325 without performing the doubletap. If yes, then the proxy 130 will apply the
doubletap, first by saving the current timestamp at step 330 as the beginning of the
doubletap delay, if applicable (e.g., based on the type of request, and possibly other
10 factors discussed previously). The data request is then sent to the server at step 335
on the current primary network. The proxy 130 waits at step 340 for the response to
be received from the server (e.g., if the network is good/responsive), up to the
doubletap delay. If the proxy determines at step 345 that the doubletap delay was
exceeded, then the proxy 130 proceeds to step 350 to send the request redundantly
15 on one or more of the available secondary networks. If, in step 345, the proxy 130
determines that the response was received successfully within the doubletap delay,
then the proxy 130 proceeds to step 375 to complete the response by sending the
response to the app 280.

[00108] If a redundant request was sent at step 350, then the proxy 130 waits for
20 the response to be received on any of the networks that the request was sent
redundantly on. At step 360, if the proxy 130 determines that a response was never
received via any network (e.g., within a timeout period), then the proxy 130 fails the
request back to the app at step 365 (e.g., sends a failure, such as an error response
or connection reset, to the app 280). However, if the first response is received (e.g.,
25 within the timeout period), then this first response is the “winner” of the doubletap
race and the proxy 130 can abandon the request on the other “losing” networks at
step 370 (e.g., either explicitly by closing their sockets, or implicitly by simply
ignoring later responses). Now that the proxy 130 has a successful response, it can
send that response from the doubletap winner to the app 280 at step 375.

30 Depending on the protocol and the state of the connection with the app, there may
be additional processing at step 380 to be performed to complete the logical request.
For example, a TLS handshake begins with a ClientHello but may take additional
request/response round-trips with the server to complete the process. Conversely, a
DNS query or TCP connection would be completed by a single response from the
35 server and would not need to have additional processing at step 380.

[00109] When the primary network has been determined to be problematic (e.g.,
slow or not fully unresponsive), it may be desirable to preemptively switch to a
different network as the primary network before the current primary network has

1 explicitly failed (e.g., timed out, disconnected), since the delays on the current
primary network can often continue beyond the current request/connection, such as
when the user is at the edge of the optimal range of a Wi-Fi network (aka a “dead
zone”). Continuing to process network data using the problematic primary network by
5 default may still work but can result in the user experiencing slower overall network
performance than normal, since the redundant data requests sent on the secondary
network(s) may be delayed to minimize duplicative network traffic. Switching to
another, perhaps more responsive, network as our primary network for processing
the network traffic means that requests are sent on that network first by default
10 without any delay, thus allowing the user to experience the full performance of that
more responsive network. Different policies are possible when deciding how and
when to change the primary network, such as by exceeding a threshold of
consecutive doubletap winners, or exceeding a maximum threshold for the response
time of one or more responses. The threshold used for these policies can be
15 configured manually or dynamically set by heuristics/machine learning based on
prior results seen for a particular network (e.g., it may be normal for some Wi-Fi
networks to respond to DNS more slowly than others) or for a particular server (e.g.,
it may be normal for some servers to respond to TCP SYN or TLS ClientHello
requests slower than others).

20 **[00110]** Another way to improve performance when the primary network is
problematic (or diminished or degraded) is to leverage other signals to either
determine how much to delay the redundant requests to the secondary network(s) or
pre-emptively switch to another network as the primary network. For example, if the
device 100 (e.g., the operating system) provides access to information about
25 network quality (e.g., signal strength, rate of packet failure, (e.g., lost/bad packets),
and the like), then these can be used to decide how and when to use the secondary
networks. For example, if the primary network is Wi-Fi and its signal quality (e.g.,
Received Signal Strength Indicator aka RSSI) falls below a particular threshold, then
the doubletap delay for the redundant requests can be reduced, possibly to none, to
30 allow the secondary network to handle requests sooner and reduce or minimize any
user-visible delays due to waiting for a timeout on a degraded network connection.
This approach can be gradual or tiered, such that different delays can be configured
for different signal quality levels (e.g., doubletap delay goes down as the signal
quality gets worse). As the signal quality continues to fall and/or degrade, it can also
35 be used to switch to a new primary network when the previous primary network’s
signal quality level falls below a worst-case level. Any quality level or indicator
available for the network can be used, in any combination, to apply the
aforementioned approach to reduce the doubletap delay or switch to a new primary

1 network. In some embodiments, the signal strength of the current primary network
(i.e. prior to switching to another network as the new primary network) may still be
within an operable range, although the performance may be degraded. For example,
the signal strength may be degraded to a point where the maximum bandwidth of the
5 current primary network is reduced, but the connection is still usable (e.g., partial
packet loss).

[00111] FIG. 4 is a high level flowchart diagram illustrating a method of leveraging
advance signals to control network access according to one embodiment of the
present invention. The term “advance signals” may be used herein to refer to early
10 warning signs detected from monitoring network behavior (e.g., before the current
primary network has explicitly failed or timed out). In some embodiments, detection
of the advance signals is used to adjust the doubletap delay and/or preemptively
switch the primary/default network. Referring to FIG. 4, according to one
embodiment, in step 410, the proxy 130 registers for signal strength updates (e.g.
15 RSSI) or similar network quality data from the operating system (this step may be
omitted in embodiments where network quality data is not available, such as when
the operating system doesn’t support it or it’s not available for the current primary
network). In step 415, the proxy 130 may also begin monitoring network metrics,
such as packet statistics (e.g., counts and/or rates of dropped packets or
20 retransmitted packets detected on the network). In step 420, the proxy 130 waits for
changes or updates to the subscribed network signal quality data and/or network
statistics data. If a change is detected (e.g., operating system sends a notification via
a callback or equivalent, if supported, and/or a thread or equivalent periodically
samples and monitors for changes in network data), then in step 425, the proxy 130
25 receives the updated network data (e.g. RSSI data and/or network statistics data).

[00112] In step 430, the proxy 130 determines whether the received new value
(e.g., of the RSSI data or the network statistics data) is “below” (or otherwise better
than) a given threshold indicating a “good” network. If so, then the proxy returns to
step 420 to wait for more data. For example, if the packet loss rate is below a “good”
30 threshold, then no change is needed and the proxy 130 returns to step 420. The
term “below” (or “<” in FIG. 4) does not necessarily refer to numerical values, but
rather refers to whether the value is within a range of “good” performance. For
example, an RSSI value is better when the signal strength is higher, so a high signal
strength would satisfy being within the condition of a “good” threshold’ as shown in
35 FIG. 4. Conversely, packet loss is better when it is lower, so a low value (or rate of
change in this value) would satisfy being within the condition of a “good threshold”.

[00113] If the new value is not within the “good” threshold or range, then the proxy
130 proceeds to determine in step 435 if the new value is within a “usable” range

1 (because the proxy 130 reached step 435, the new value is outside the “good”
range). If so, then in step 440, the proxy 130 reduces the doubletap delay discussed
above. This is because the current network appears to be unreliable, and therefore
requests should be sent on alternative networks sooner than later. After reducing the
5 doubletap delay in step 440, the proxy returns to wait for the next network status
update in step 420.

[00114] If the proxy 130 determines in step 435 that the new value is not within the
“usable” range, then in step 445 the proxy 130 sets another network of the networks
available to the portable communications device 100 as the primary network. The
10 selection of another network may be done based on a preference ranking among the
other networks (e.g., a highest ranked network), as described above.

[00115] Some aspects of embodiments of the present invention relate to selecting
another network to be the primary network based on the response time of the
doubletap request as discussed above with respect to FIG. 3. Still referring to FIG. 4,
15 in some embodiments of the present invention, after the proxy 130 performs, in step
450, a doubletap via a current primary network, the proxy 130 then calculates the
response time of the doubletap winner at step 455 (e.g., the difference between a
timestamp of a request and a timestamp of a corresponding response from the
network). In step 460, the proxy 130 determines whether the response time
20 exceeded a “usable” threshold, such as a time threshold that is low enough to avoid
failures/timeouts detected by app 280 but high enough to indicate the network is not
sufficiently responsive to be preferred as the default network for future requests (e.g.
2-4 seconds). If so, then the proxy 130 selects a new primary network in step 445, as
discussed above.

25 **[00116]** If the response time of the doubletap winner is within the maximum
“usable” time threshold allowed for the primary network, then proxy 130 checks at
step 465 whether the primary network was the doubletap winner. If yes, then the
primary network is confirmed to be sufficiently responsive and proxy 130 continues
to use it by default for requests at step 450. However, if the doubletap winner was a
30 secondary network, then we increment at step 470 the number of consecutive
doubletap losses observed for the primary network. Next, at step 475, proxy 130
checks whether the number of consecutive doubletap losses by the primary network
exceeds a “usable” threshold indicating that the primary network is sufficiently
degraded or unresponsive (e.g. 2-4 consecutive losses). If no, then the primary
35 network may still be sufficiently response and proxy 130 continues to use it by
default for requests at step 450. If not, then proxy 130 selects a new primary network
in step 445, as discussed above.

1 **[00117]** When switching to another network to use by default as the primary
network, this generally means the preferred network (e.g., Wi-Fi) is problematic (e.g.,
unresponsive or too slow) and it is better to use another, less preferred network
(e.g., cellular). It may then be possible later for the more preferred network to
5 become responsive enough again, such as if the user moves back closer to where
the Wi-Fi signal quality is good again. The doubletap approach can be used to
determine that a better network is available to use as the primary network than the
current one, such as by continuing to allow doubletap requests to race across the
available networks, leverage network quality indicators, or a combination thereof.
10 Performing the doubletap even after switching to a new primary network allows us to
see the responsiveness of the previous primary network, and to monitor for a change
in the conditions that initially triggered the switch. For example, if the primary
network is switched to cellular due to exceeding consecutive doubletap loses, the
proxy 130 can simply continue sending redundant requests to Wi-Fi, and if the proxy
15 130 detects consecutive doubletap winners on Wi-Fi, the proxy 130 can then switch
back to Wi-Fi as the primary network.

[00118] In some embodiments, the preference ranking among available networks
is determined based on a combination of factors, such as cost, performance,
business relationship, etc. These factors can be variable, such as network
20 performance (e.g., temporary congestion) or network traffic cost (e.g., higher during
peak periods), so the preference ranking may change dynamically. Embodiments of
the present invention can leverage any combination of static or dynamic factors to
consider which network to prefer at any time. For example, to consider performance
as a factor, the throughput of the available networks can be monitored. Since the
25 system (e.g. proxy 130) is handling the network traffic, an embodiment of the present
invention can then actively measure the current network bandwidth or latency of the
traffic flowing through the system (e.g., current rate, peak rate within recent time
period), and then adjust the preference ranking accordingly (e.g., a Wi-Fi network
that is significantly slower than the cellular network can reduce its relative ranking,
30 possibly such that the cellular may be considered more preferred). As another
example, to consider cost as a factor, the system may query for the current cost for
using a particular network. An embodiment of the present invention may make an
application programming interface (API) call to the network or some backend
management and/or pricing system to query for the current cost of using an available
35 network (e.g., current Wi-Fi or cellular network), and then adjust the preference
ranking accordingly. For example, these costs can sometimes be dynamic, such as
when a user is roaming onto 3rd-party cellular network, attempting to access a Wi-Fi
network that charges for access, or possibly varying based on current on-

1 demand/spot pricing that increases based on congestion. In some aspects of
embodiments of the present invention, an entity such as a user, mobile network
operator, or device manufacturer can supply a particular ranking (e.g., pin particular
5 networks higher in the preference ranking) or may configure the weights of particular
factors used in combining the preference ranking of the networks.

[00119] The aforementioned embodiments of the present invention may rely on
network traffic to trigger their respective action, such as switching to a different
primary network or re-ranking (or rearranging) the preference of the available
networks. In some cases, there may not be sufficient network traffic to trigger a
10 desired action, such as when there is only a single long-lived encrypted connection
receiving data from the server without any new data requests that can be
doubletapped. In one embodiment of the present invention, the proxy 130 generates
its own doubletapped requests to generate a minimum level of activity to help trigger
detection of an unresponsive primary network. For example, one approach is to
15 monitor for a minimum time threshold (e.g., a few seconds) since the last DNS
request received from the inside that was doubletapped, and if that was exceeded,
then the proxy 130 performs a doubletap of a test DNS request (e.g., for a well-
known hostname) that helps detect an unresponsive primary network within that
minimum time threshold. It should be apparent to someone of ordinary skill that any
20 data request that can be doubletapped can then be used, such as a TCP connection
request (e.g., SYN). Similarly, these simple and lightweight doubletap tests can then
be used to trigger other necessary actions, such as re-ranking of the preferred order
of the available networks.

25 ***Load Balancing***

[00120] Some aspects of embodiments of the present invention relate to
enhancing the performance of slow connections on a network (e.g., a slow Wi-Fi
connection) with another connection on another network (e.g., a cellular connection).
In various embodiments of the present invention, the load balancing technique
30 distributes load across one or more available networks to increase overall network
performance (e.g., increase bandwidth, reduce latency). Some embodiments of the
present invention may measure the performances of the different network
connections (similar to the doubletap technique described above) and select the
network connection or a combination of network connections having higher
35 throughput or less congestion. This approach allows multiple requests or
connections to be distributed across the available networks based upon their
performance, such as distributing based upon the maximum achievable throughput
of each network, the number of outstanding requests/connections of each network,

1 or a combination thereof. In other embodiments of the present invention, all of the network connections (e.g., both Wi-Fi and cellular connections) may be used concurrently to download different portions of the data, where the data received along the different connections is recombined at the device.

5 **[00121]** Embodiments of the present invention are directed towards detecting when to distribute network traffic to a secondary, less preferred network when the primary, more preferred network is busy or slow. Generally, it is desirable to use the preferred network when it is fast and has sufficient capacity, and only use other, less preferred networks when the preferred network's capacity has been reached.

10 However, this can be difficult to determine accurately unless there is enough data being transferred to completely saturate the network connection. Also, when the network is shared with other users, which is common for both Wi-Fi and cellular networks, the capacity used by other users will reduce the peak capacity that can be measured. As such, some embodiments relate to determining that the preferred

15 network's capacity has been reached without the generation of test traffic and works independently of the network load generated by other users on that network. An embodiment of the present invention is to track the currently active requests on the preferred network and distribute requests to other, less preferred networks when that threshold has been reached. The notion of an active request here is any logical

20 request/response pair, and this can apply to both cleartext traffic (e.g., an HTTP GET) as well as to encrypted traffic (e.g., receiving encrypted data in response to sending outbound encrypted data). In the case of cleartext traffic, determining a request is no longer active can be based upon receiving the complete response data (e.g., an HTTP response can specify a content-length header or an end of data

25 marker for chunked transfers). For encrypted traffic, it may not be possible to identify an explicit end of a response, so in some embodiments, a request is treated as being complete upon receipt of the initial data response (e.g., most responses from the server are small and complete within a few seconds). Alternatively, there are other signals that can be leveraged to identify the end of an encrypted data request, such

30 as when the time since receiving the last encrypted data chunk exceeds the typical minimum time between data chunks from the server (e.g., 2-3 seconds). This approach of tracking active requests has the advantage of only considering connections that have actual requests being actively processed, and avoids considering long-lived connections that remain but are inactive (e.g., pooling unused

35 connections). This approach can also handle pipelined requests going over a single connection, since each new pipelined request is sent out as a consecutive sequence of outbound data to the server, and each of these can increment the active request

1 count until a corresponding sequence of inbound data from the server is received to decrement the active request count.

5 **[00122]** With a count of “active requests”, the proxy 130 can then shift new requests to another, less preferred network (e.g., cellular) whenever a threshold for the primary, preferred network (e.g., Wi-Fi) is exceeded. This threshold (e.g., minimum threshold) controls the selection of networks such that the primary network is tried first for new requests, and if these active requests are completed (e.g., responded to) quickly without exceeding this threshold, then the proxy 130 uses the primary network to service the bulk of the network traffic. Whenever this threshold is
10 reached, then a proxy 130, according to embodiments of the present invention, directs new requests to another network and correspondingly tracks the number of active requests for that other network to establish a balanced distribution of load among the available networks.

15 **[00123]** Another embodiment of the present invention is to use the number of active requests to establish a ratio of requests to distribute to each network. With a relative ratio of active requests between the available networks, the system can maintain a distribution of active requests among multiple networks to balance (e.g., for optimal performance) the network load across those networks. For example, a slower network may have requests that remain active for a longer timeframe than
20 requests sent to the faster network, so this approach provides a way to distribute traffic among multiple networks without using test traffic to determine the capacity of each of the networks. In various embodiments of the present invention, the ratio of active requests to maintain between each of the available networks is configured statically, computed dynamically, or a combination thereof. In some embodiments,
25 the active requests ratio is combined with the minimum threshold for the primary network, such that the ratio is used to distribute new requests among the networks when the minimum threshold is currently exceeded.

30 **[00124]** FIG. 5 is a high level flowchart diagram illustrating an example method of the load-balancing approach. This flowchart can be performed by an intermediary in the network data path, such as by proxy 130 between app 280 and app server 250 described in FIG. 1. According to one embodiment of the present invention, the proxy 130 waits at step 410 for data requests to be sent from an app 280 to an app server 250. When a data request is received from an app at step 55, the proxy 130 determines at step 520 whether the primary network has exceeded its minimum
35 threshold of active requests. If no, then the primary network is selected at step 525 for the new request and then proceeds to step 555. If yes, then the proxy 130 iterates through each of the secondary networks, such as in order of preference, to identify or select a network to use by checking its ratio of active requests relative to

1 the primary network. If the active requests ratio of the primary network relative to the
current secondary network at step 535 is greater the target ratio, then the proxy 130
selects the current secondary network at step 545 to use for this request as if it were
the primary and then the proxy 130 proceeds to step 555. Otherwise, if the active
5 requests ratio between the primary and current secondary network is less than the
target ratio, then the proxy 130 continues to the next secondary network at step 530.
If there are no other secondary networks at step 540, then the primary network is
selected at step 550 and proceeds to step 555. At step 555, a network has been
selected for the request so the active request count for the network is incremented.
10 In one embodiment of the present invention, the load-balancing approach is
combined with the doubletap approach at step 560, where the request is then
processed according to the flowchart of FIG. 3 to provide redundancy against the
selected network being unresponsive. When the request is completed (e.g., when a
response is received from the selected network), the number of active requests for
15 the selected network is decremented.

[00125] Some aspects of embodiments of the present invention relate to
measuring the speed/bandwidth of each network without using test traffic by
monitoring the bandwidth used by the network activity currently flowing through the
system (e.g., average and peak speeds). This approach allows the system (e.g., the
20 proxy 130) to also use these bandwidth measurements of each network to calculate
an appropriate ratio of active requests among these available networks, and then to
distribute new requests to each network to maintain this ratio. For example, Table 1,
above, shows how overall performance can be increased by distributing the load
across multiple networks based on the relative speed of each one. Because
25 networks are generally shared by multiple mobile devices among multiple users
(e.g., cellular and Wi-Fi), the performance of these shared networks can vary over
time (e.g., due to contention between the users). As such, these bandwidth
measurements can be used to dynamically monitor and update the active requests
ratio between the available networks.

30 **[00126]** Some aspects of embodiments of the present invention relate to tracking
the maximum bandwidth seen for each network (e.g., specific Wi-Fi SSID and
BSSID, or a specific cell tower) as a possible bandwidth ceiling for each network,
which can be used for calculating the active request ratio. For example, if the
number of active requests is increasing for a network but the measured bandwidth is
35 not increasing, then the currently measured bandwidth may be a possible current
ceiling for the network. Also, some aspects of embodiments of the present invention
relate to using a minimum or threshold amount of data transferred during the
timeframe of the bandwidth measurement to increase the confidence that the

1 measured bandwidth is a possible ceiling and to avoid false positives. In some
embodiments of the present invention, the proxy 130 calculates a confidence level
(or confidence score) in the current bandwidth, such as based on the proximity of the
current measurement to previous measurements and/or based on how much data
5 was transferred during the timeframe of the current measurement. Having a
confidence level allows the system to determine whether the currently measured
bandwidth should be used to update the active requests ratio, such as by
establishing a minimum confidence level required to update the ratio. Then, when
selecting a network for a data request, the proxy 130 also determines whether a
10 currently selected network is at its bandwidth ceiling and choose another network
that is below its bandwidth ceiling, possibly doing so in spite of (e.g., overruling) the
active requests ratio.

[00127] Tracking the active requests for each network can also be used to
dynamically detect changes in the capacity of a network. Some aspects of
15 embodiments of the present invention relate to monitoring how long requests remain
active for each network and adjusting the active requests threshold of each network
to allow each network to be able to complete the active requests (e.g., decrement
the number of active requests) in a similar amount of time. For example, if one
network is found to be completing active requests twice as fast (e.g., 2 times faster)
20 than another network, then the ratio of active requests can be adjusted to send twice
as many (e.g., 2 times more) new requests to the faster network than the slower one.
The request completion time for each network may change over time (e.g., due to
activity by other users on the shared network increasing or decreasing over time), so
this can be used to dynamically adjust the active requests ratio between the
25 available networks. In some embodiments, this approach is applied in circumstances
where there is insufficient network activity to determine the peak capacity or speed of
the available networks (e.g., without needing to use test traffic to determine that peak
capacity). This approach can also be combined with the previously discussed
approach of using speed measurements of each network to calculate an appropriate
30 ratio, such as by assigning a weighting to each approach and calculating a ratio of
active requests based on the combined weighting. In some embodiments, the
relative weighting of each approach is adjusted based on their respective
confidence. For example, the currently measured speed of a network is more
accurate when the amount of data transferred is high, while measuring the active
35 time of requests is more accurate when the amount of data transferred is low, so the
weighting of each can be adjust dynamically based on these factors (e.g., bytes
transferred per request).

1 **[00128]** Some embodiments of the above approach use response time or latency
to calculate a ratio of requests to send to each network, such as by measuring the
time from sending the request to receiving the first byte of response (e.g., time to first
byte without waiting for the active request to fully complete). This approach can also
5 be combined in a similar manner (e.g. calculate a combined weighted) with the
aforementioned approaches of factoring in speed measurements and active time per
request. It should be apparent to someone of ordinary skill in the art that a variety of
different performance metrics can be leveraged to determine an optimal ratio of
active requests to distribute to each available network.

10 **[00129]** Each of the aforementioned thresholds or ratios can also be adjusted on a
per network basis (e.g., track separate values based on historical results seen for
each distinct Wi-Fi SSID), such that the system learns thresholds (e.g., the best
thresholds) to use for each network. This allows the system to establish operating
settings (e.g., the optimal operating settings) for a network without having to start
15 from scratch every time that network is joined. The various measurements and
calculated ratios discussed previously can be stored for each network (e.g., in
accordance with the Wi-Fi SSID of the network) and restored whenever rejoining that
specific network, then continuously updated over time.

20 ***Wi-Fi connection management***

[00130] Some aspects of embodiments of the present invention relate to
automating the Wi-Fi connection process, such as checking for other Wi-Fi networks
with better signal quality and either generating notifications when the device is within
range of a “good” Wi-Fi connection or pre-emptively detaching from the currently
25 connected Wi-Fi network to force the device to join the better Wi-Fi network. Some
aspects of embodiments of the present invention relate to automating the process of
logging in to captive portals (e.g., Wi-Fi connections provided by businesses but
subject to log-in, password, or payment requirements).

[00131] One aspect of embodiments of the present invention relates to leveraging
30 the portable communication device’s 100 scanning of available networks (e.g., Wi-Fi
rescans) to connect to different networks to improve performance. These rescans
occur regularly in most operating systems, such as when a user moves, when the
RSSI of the current network changes, or when the user views the list of available
networks. For example, when doubletap, described above with respect to, for
35 example, FIG. 3, triggers a primary network switch (e.g., when in the dead zone of
the Wi-Fi network), it can be beneficial to check for a different network of the
preferred type (e.g., another Wi-Fi network with a better RSSI), since the current
network has been deemed to be slow or unresponsive. Another example would be

1 when the load-balancing determines that the active requests are finishing too slowly (e.g., average or worse case times are below some minimum threshold).

5 **[00132]** Regardless of the trigger, any detected degradation of a currently connected network can be used to trigger switching to another network of the same type. In other words, the earlier discussion about the doubletap approach was to switch from one network type (e.g., Wi-Fi) to another network type (e.g., for a single request or to use as the default/primary for all requests), but this can be further generalized to switch between different networks of the same type. For example, this would be helpful for switching to another Wi-Fi network (e.g., having a different SSID, or same SSID but different BSSID), since devices generally do not switch to another Wi-Fi network until the current one is unusable because they typically only detect problems using timeouts that are set high enough that there is certainty of a problem. However, as discussed above, these timeouts are generally too high and thus too slow to seamlessly move users to a better network (e.g., the app may hang for a long time and/or the user may see an error message before the device switches to a better network). With the doubletap approach, for example, the system can make a faster decision to change to another network of the preferred type (e.g., disconnect from the current Wi-Fi network and join another Wi-Fi network) well before a user sees a timeout failure, enabling a faster, more seamless transition while maximizing the quality/availability of the preferred network type. Also, the handoff between different unrelated Wi-Fi networks can be slow, since the portable communication device 100 may need to perform a long multi-step process to join and authenticate to the Wi-Fi network, so the doubletap approach can ensure a more seamless Wi-Fi network to Wi-Fi network transition by using cellular during this transition as previously discussed.

25 **[00133]** FIG. 6 shows an example of a neighborhood with multiple Wi-Fi networks where each one has inner circular region indicating the area of good signal quality and an outer circular region indicating the area of bad signal quality (e.g., the Wi-Fi dead zone). For example, when the outer zone of one Wi-Fi network (e.g., Wi-Fi network "A") overlaps with the inner zone of another Wi-Fi network (e.g., Wi-Fi network "B"), then that can result in Wi-Fi network A preventing the use of Wi-Fi network B, such as when a mobile device first attaches to Wi-Fi network A but then stays attached to it even though Wi-Fi network B may offer a better network connection.

35 **[00134]** FIG. 7 shows the same example as FIG. 6 but leveraging the techniques according to embodiments of the present invention. With the doubletap technique, when the mobile device 100 is in the outer unusable zones of each Wi-Fi network, it can leverage a cellular network to automatically fill in network connectivity while in

1 the dead zone of an associated Wi-Fi network. With the load-balancing technique,
the inner usable zones of each Wi-Fi can leverage the cellular network to combine
the bandwidth of both to increase overall performance. With the connection
management technique, the mobile device 100 can be automatically switched to
5 another Wi-Fi network by performing scanning for another Wi-Fi network, such as
either triggered by the doubletap logic in the outer dead zone or by the load-
balancing logic in the inner good zone.

[00135] While the present invention has been described in connection with certain
exemplary embodiments, it is to be understood that the invention is not limited to the
10 disclosed embodiments, but, on the contrary, is intended to cover various
modifications and equivalent arrangements included within the spirit and scope of
the appended claims, and equivalents thereof.

15

20

25

30

35

1 **WHAT IS CLAIMED IS:**

1. A method for managing network traffic on a portable communication device comprising a processor, a memory, and a plurality of network interfaces configured to connect to a plurality of networks, the method comprising:

5 intercepting, by a traffic manager running on the processor, network data to and from an application running on the processor;

transmitting, by the traffic manager, an idempotent request of the network data to a server via the plurality of networks;

10 receiving, by the traffic manager, a response to the idempotent request from the server via a first network on the plurality of networks; and

selecting, by the traffic manager, the first network of the plurality of networks to use for receiving and transmitting the response to the application.

2. The method of claim 1, wherein the transmitting the idempotent request to the server via the plurality of networks comprises transmitting the idempotent request on a second network of the plurality of networks, and

wherein the method further comprises terminating the idempotent request on the second network.

3. The method of claim 1, wherein the transmitting the idempotent request to the server via the plurality of networks comprises:

transmitting the idempotent request to the server via one network of the plurality of networks; and

25 transmitting the idempotent request to the server via one or more other networks of the plurality of networks in accordance with a delay.

4. The method of claim 3, wherein the delay is shorter than an application-level timeout of the application running on the portable communication device.

30 5. The method of claim 3, wherein the delay is set based on a typical response time to the idempotent request.

6. The method of claim 3, wherein the idempotent request is associated with a network protocol, and

35 wherein the delay is set based on the network protocol associated with the idempotent request.

- 1 7. The method of claim 3, wherein the delay is set based on a size of a response
to the idempotent request.
- 5 8. The method of claim 3, wherein the delay is set based on a network quality
metric.
9. The method of claim 3, wherein the plurality of networks are arranged in a
preference ranking, and
10 wherein the one or more other networks are selected in accordance with the
preference ranking.
10. The method of claim 9, wherein each of the networks is associated with a
different delay in accordance with the preference ranking.
- 15 11. The method of claim 9, further comprising rearranging the plurality of
networks in the preference ranking in accordance with one or more dynamic factors.
12. The method of claim 11, wherein the one or more dynamic factors comprise
network performance.
20
13. The method of claim 11, wherein the one or more dynamic factors comprise
network traffic cost.
14. The method of claim 1, wherein the plurality of networks comprise a plurality
25 of different types of networks.
15. The method of claim 14, wherein the types of networks comprise one or more
of: cellular, Bluetooth, and Wi-Fi networks.
- 30 16. A method for managing network traffic on a portable communication device
comprising a processor, a memory, and a plurality of network interfaces configured
to connect to a plurality of networks, the method comprising:
processing network traffic of an application running on the processor via a first
network of the plurality of networks, the first network being specified as a primary
35 network by an operating system running on the processor;
monitoring a plurality of network status information associated with the first
network;

1 detecting a problem of the first network when one or more parameters of the
received network status information is outside of one or more threshold values;
in response to detecting the problem in the first network, selecting a second
network of the plurality of networks as the primary network; and
5 processing the network traffic via the second network as the updated primary
network.

17. The method of claim 16, wherein the network traffic comprises a request and
a response to the request, and
10 wherein the problem in the first network is detected based on a response time
between a timestamp of the request and a timestamp of the response exceeding a
maximum threshold.

18. The method of claim 16, wherein the detecting the problem in the first network
15 comprises:
monitoring at least one of network statistics on the first network; and
detecting the problem when a change in the network statistics exceeds a
threshold.

19. The method of claim 18, wherein the network statistics comprise a packet loss
20 rate or a bad packet rate.

20. The method of claim 16, wherein the first network is a wireless network, and
wherein the detecting the problem in the first network comprises:
25 monitoring a signal strength of the wireless network; and
detecting the problem when the signal strength drops below a
threshold signal strength.

21. The method of claim 16, wherein the second network of the plurality of
30 networks is selected in accordance with a preference ranking of the plurality of
networks.

22. The method of claim 16, wherein the problem in the first network is detected
based on responses received on other networks before a response is received on
35 the first network.

23. A portable communication device comprising a processor, a memory, and a
plurality of network interfaces configured to connect to a plurality of networks, the

1 memory storing instructions that, when executed by the processor, cause the
processor to manage network traffic of the portable communication device by:
intercepting, by a traffic manager running on the processor, network data to
and from an application running on the processor;
5 transmitting, by the traffic manager, an idempotent request of the network
data to a server via the plurality of networks;
receiving, by the traffic manager, a response to the idempotent request from
the server via a first network on the plurality of networks; and
selecting, by the traffic manager, the first network of the plurality of networks
10 to use for receiving and transmitting the response to the application.

24. The portable communication device of claim 23, wherein the transmitting the
idempotent request to the server via the plurality of networks comprises transmitting
the idempotent request on a second network of the plurality of networks, and
15 wherein the instructions further comprise instructions that cause the processor
to terminate the idempotent request on the second network.

25. The portable communication device of claim 23, wherein the transmitting the
idempotent request to the server via the plurality of networks comprises:
20 transmitting the idempotent request to the server via one network of the
plurality of networks; and
transmitting the idempotent request to the server via one or more other
networks of the plurality of networks in accordance with a delay.

25 26. The portable communication device of claim 25, wherein the delay is shorter
than an application-level timeout of the application running on the portable
communication device.

27. The portable communication device of claim 25, wherein the delay is set
30 based on a typical response time to the idempotent request.

28. The portable communication device of claim 25, wherein the idempotent
request is associated with a network protocol, and
wherein the delay is set based on the network protocol associated with the
35 idempotent request.

29. The portable communication device of claim 25, wherein the delay is set
based on a size of a response to the idempotent request.

1

30. The portable communication device of claim 25, wherein the delay is set based on a network quality metric.

5

31. The portable communication device of claim 25, wherein the plurality of networks are arranged in a preference ranking, and wherein the one or more other networks are selected in accordance with the preference ranking.

10

32. The portable communication device of claim 31, wherein each of the networks is associated with a different delay in accordance with the preference ranking.

15

33. The portable communication device of claim 31, wherein the memory further stores instructions that, when executed by the processor, cause the processor to rearrange the plurality of networks in the preference ranking in accordance with one or more dynamic factors.

20

34. The portable communication device of claim 33, wherein the one or more dynamic factors comprise network performance.

35. The portable communication device of claim 34, wherein the one or more dynamic factors comprise network traffic cost.

25

36. The portable communication device of claim 23, wherein the plurality of networks comprise a plurality of different types of networks.

37. The portable communication device of claim 36, wherein the types of networks comprise one or more of: cellular, Bluetooth, and Wi-Fi networks.

30

38. A portable communication device comprising a processor, a memory, and a plurality of network interfaces configured to connect to a plurality of networks, the memory storing instructions that, when executed by the processor, cause the processor to manage network traffic of the portable communication device by:

35

processing network traffic of an application running on the processor via a first network of the plurality of networks, the first network being specified as a primary network by an operating system running on the processor;

monitoring a plurality of network status information associated with the first network;

1 detecting a problem of the first network when one or more parameters of the
received network status information is outside of one or more threshold values;
in response to detecting the problem in the first network, selecting a second
network of the plurality of networks as the primary network; and
5 processing the network traffic via the second network as the updated primary
network.

39. The portable communication device of claim 38, wherein the network traffic
comprises a request and a response to the request, and
10 wherein the problem in the first network is detected based on a response time
between a timestamp of the request and a timestamp of the response exceeding a
maximum threshold.

40. The portable communication device of claim 38, wherein the instructions for
15 detecting the problem in the first network comprise instructions that, when executed
by the processor, cause the processor to:
monitor at least one of network statistics on the first network; and
detect the problem when a change in the network statistics exceeds a
threshold.

20 41. The portable communication device of claim 40, wherein the network statistics
comprise a packet loss rate or a bad packet rate.

42. The portable communication device of claim 38, wherein the first network is a
25 wireless network, and
wherein the detecting the problem in the first network comprises:
monitoring a signal strength of the wireless network; and
detecting the problem when the signal strength drops below a
threshold signal strength.

30 43. The portable communication device of claim 38, wherein the second network
of the plurality of networks is selected in accordance with a preference ranking of the
plurality of networks.

35 44. The portable communication device of claim 38, wherein the problem in the
first network is detected based on responses received on other networks before a
response is received on the first network.

FIG. 1

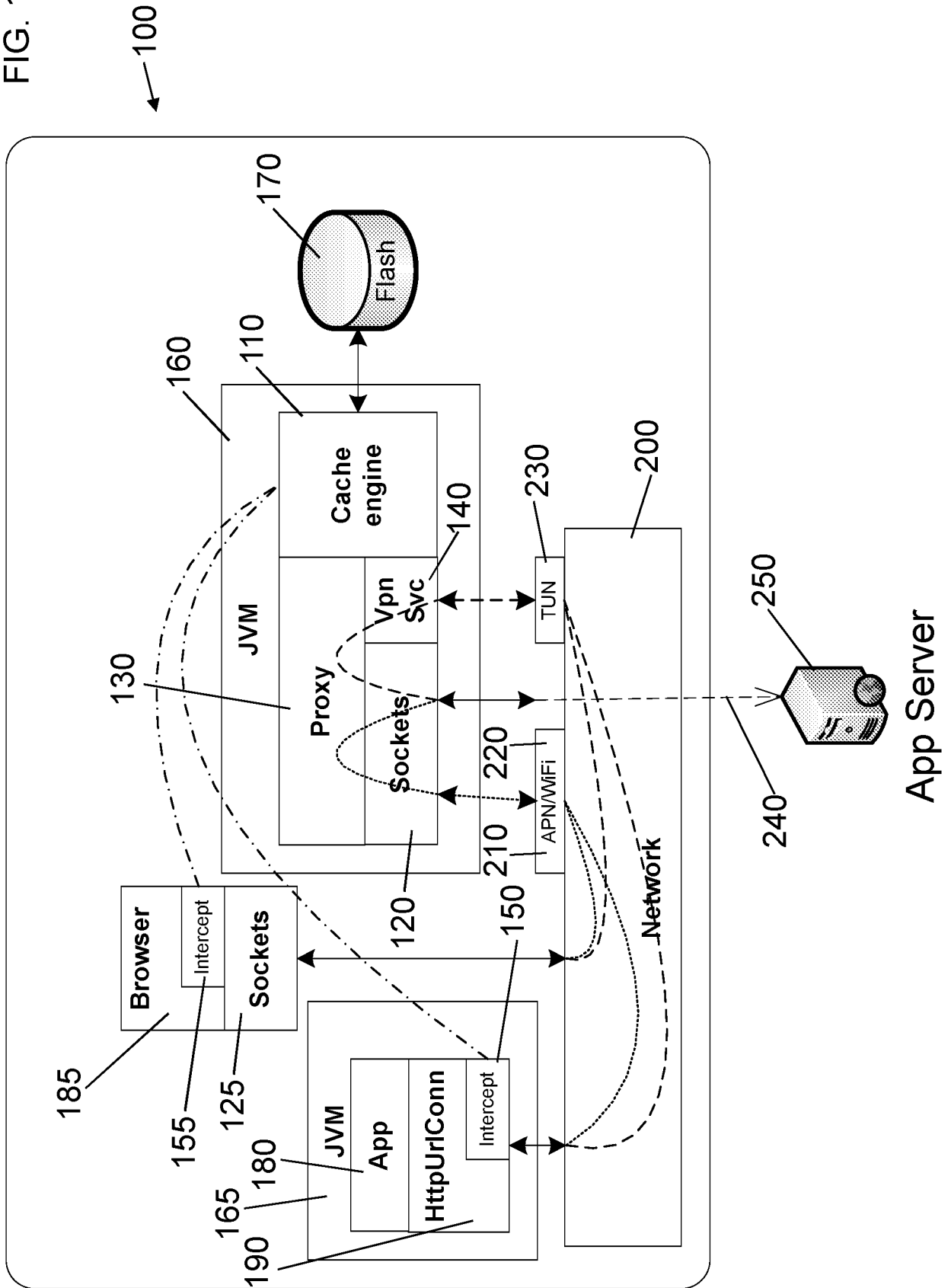


FIG. 2

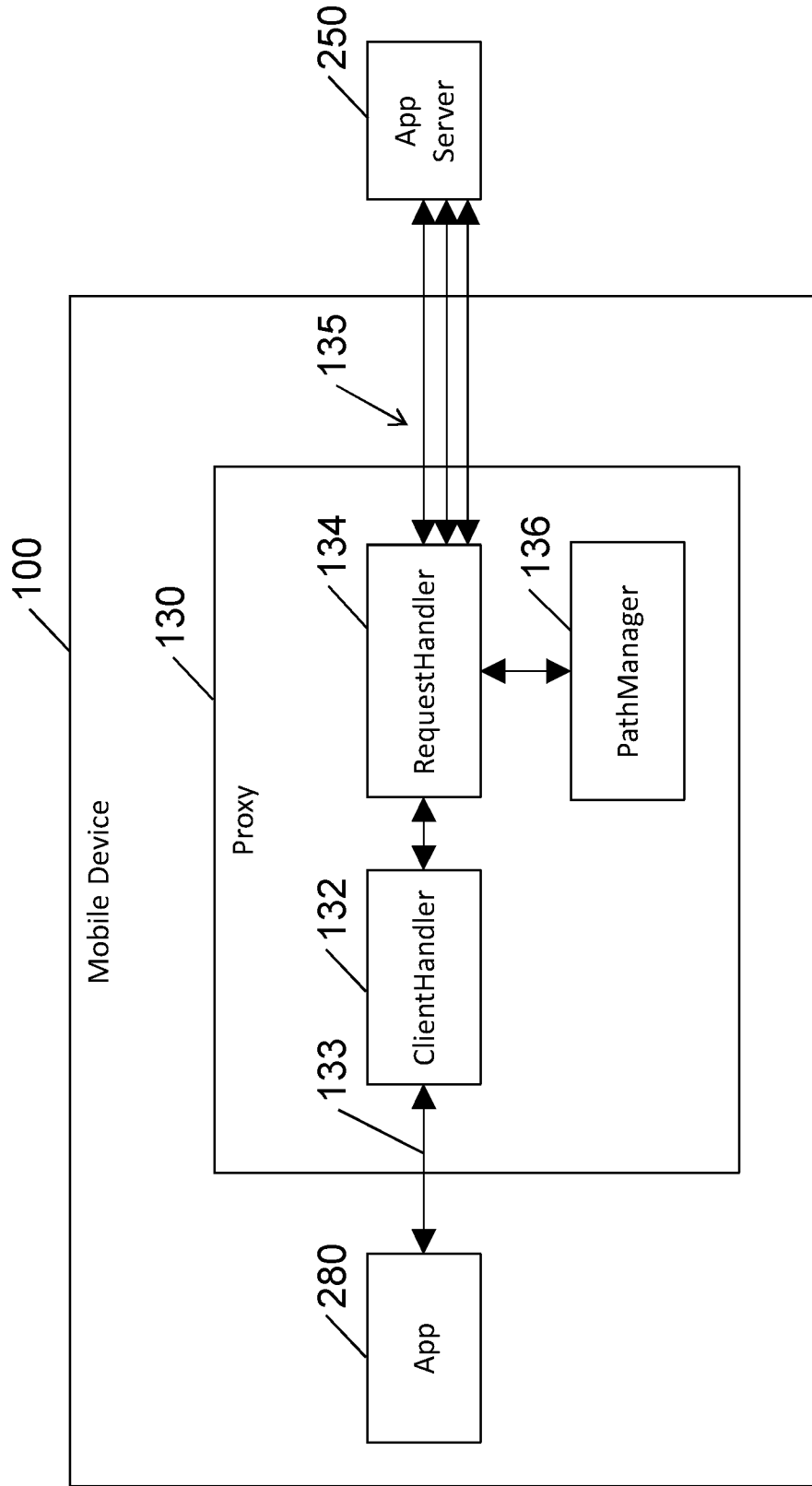


FIG. 3

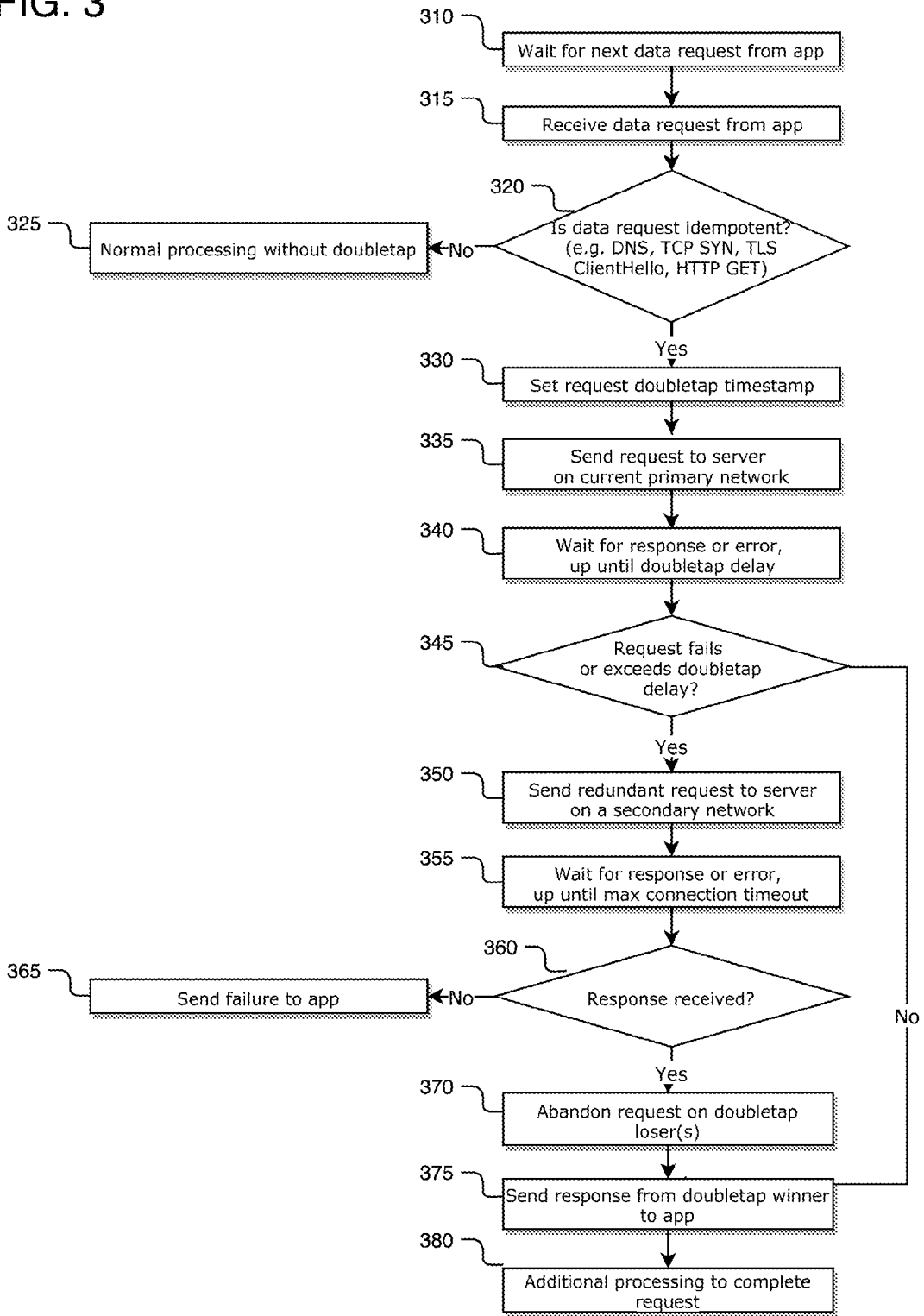


FIG. 4



FIG. 5

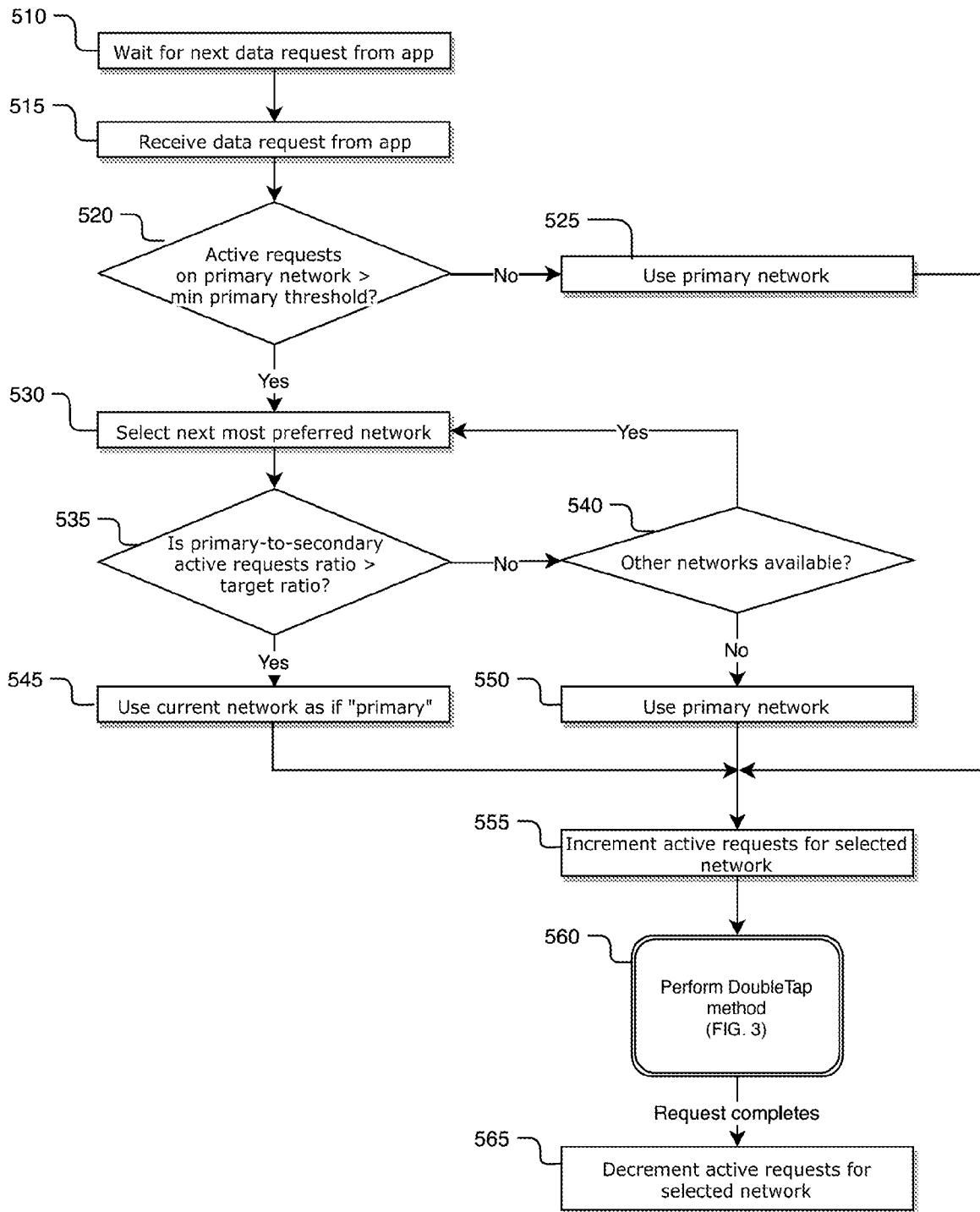
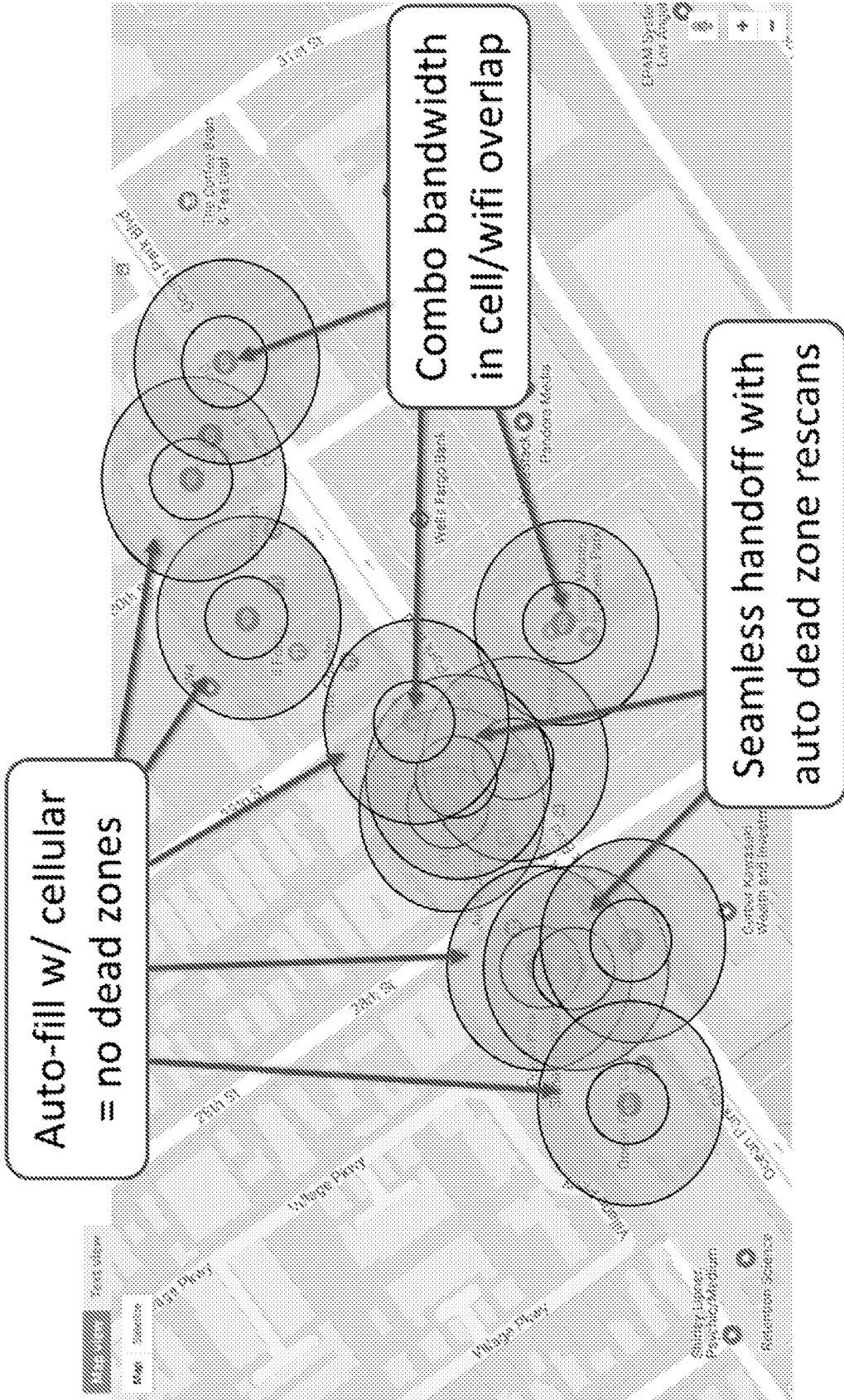


FIG. 7



INTERNATIONAL SEARCH REPORT

International application No.

PCT/US19/35082

A. CLASSIFICATION OF SUBJECT MATTER

IPC - H04W 48/18, 88/06, 72/02, 36/36 (2019.01)

CPC - H04W 48/18, 88/06, 72/02, 36/36

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)

See Search History document

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

See Search History document

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

See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	EP 1962471 A1 (ALCATEL LUCENT) 27 August 2008, abstract, paragraphs [0002], [0034], [0037]-[0038]	1-15, 23-37
Y	US 6,141,686 A (JACKOWSKI, S et al.) 31 October 2000, abstract	1-15, 23-37
Y	WO 2010/007556 A1 (TELEFONAKTIEBOLAGET LM ERICSSON) 21 January 2010, page 4 lines 5-20, page 10 lines 1-10, claims 8-9	3-13, 25-35
Y	US 2017/0373804 A1 (CONVIDA WIRELESS, LLC) 28 December 2017, abstract, paragraph [0037], [0066], claim 1	4-7, 26-29

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

"D" document cited by the applicant in the international application

"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 another 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

31 July 2019 (31.07.2019)

Date of mailing of the international search report

30 SEP 2019

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-8300

Authorized officer

Shane Thomas

Telephone No. PCT Helpdesk: 571-272-4300

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US19/35082

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

*** Continued in Extra Sheet ***

1. As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.
3. As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1-15, 23-37

Remark on Protest

- The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- No protest accompanied the payment of additional search fees.

-Continued from Box III -

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be examined, the appropriate additional examination fee must be paid.

Group I: Claims 1-15 and 23-37 are directed towards a method and device for handling idempotent requests.

Group II: Claims 16-22 and 38-44 are directed towards a method and device for responding to problem in the network based on received status information.

The inventions listed as Groups I-II do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons:

The special technical features of Group I include at least intercepting, by a traffic manager running on the processor, network data to and from an application running on the processor; transmitting, by the traffic manager, an idempotent request of the network data to a server via the plurality of networks; receiving, by the traffic manager, a response to the idempotent request from the server via a first network on the plurality of networks; and selecting, by the traffic manager, the first network of the plurality of networks to use for receiving and transmitting the response to the application, which are not present in Group II.

The special technical features of Group II at least processing network traffic of an application running on the processor via a first network of the plurality of networks, the first network being specified as a primary network by an operating system running on the processor; monitoring a plurality of network status information associated with the first network; detecting a problem of the first network when one or more parameters of the received network status information is outside of one or more threshold values; in response to detecting the problem in the first network, selecting a second network of the plurality of networks as the primary network; and processing the network traffic via the second network as the updated primary network, which are not present in Group I.

The common technical features shared by Groups I-II are a method for managing network traffic on a portable communication device comprising a processor, a memory, and a plurality of network interfaces configured to connect to a plurality of networks, the method comprising: A portable communication device comprising a processor, a memory, and a plurality of network interfaces configured to connect to a plurality of networks, the memory storing instructions that, when executed by the processor, cause the processor to manage network traffic of the portable communication device by: network traffic; and selecting a network.

However, these common features are previously disclosed by US 2015/0105074 A1 to HU, Q et al. (hereinafter "Hu"). Hu discloses a method for managing network traffic on a portable communication device comprising a processor, a memory, and a plurality of network interfaces configured to connect to a plurality of networks, the method comprising (improving the network traffic distribution at a multimode terminal having GSM, LTE and WI-FI capability (a plurality of network interfaces configured to connect to a plurality of networks), a processor and storage for programmable codes, fig. 4, paragraph [0004], [0081]): a portable communication device comprising a processor, a memory, and a plurality of network interfaces configured to connect to a plurality of networks, the memory storing instructions that, when executed by the processor, cause the processor to (improving the network traffic distribution at a multimode terminal having GSM, LTE and WI-FI capability (a plurality of network interfaces configured to connect to a plurality of networks), a processor and storage for programmable codes, fig. 4, paragraph [0004], [0081]): manage network traffic of the portable communication device by (improve network traffic distribution, paragraph [0004]): selecting a network (selecting a network, abstract).

Since the common technical features are previously disclosed by the Hu reference, these common features are not special and so Groups I-II lack unity.