



(51) International Patent Classification:
H04L 12/24 (2006.01)

(21) International Application Number:
PCT/EP2019/061992

(22) International Filing Date:
09 May 2019 (09.05.2019)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
18171643.2 09 May 2018 (09.05.2018) EP

(71) Applicant: NEC LABORATORIES EUROPE GMBH
[DE/DE]; Kurfürsten-Anlage 36, 69115 Heidelberg (DE).

(72) Inventors: LIEBSCH, Marco; c/o NEC Laboratories
Europe GmbH, Kurfürsten-Anlage 36, 69115 Heidelberg
(DE). YOSHIZAWA, Takahito; c/o NEC Laboratories

Europe GmbH, Kurfürsten-Anlage 36, 69115 Heidelberg
(DE). SINGH, Shubhranshu; c/o NEC Laboratories Eu-
rope GmbH, Kurfürsten-Anlage 36, 69115 Heidelberg
(DE).

(74) Agent: SMITH, Jeremy; Mathys & Squire LLP, The
Shard, 32 London Bridge Street, London Greater London
SE1 9SG (GB).

(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: LEVERAGING DATA ANALYTICS FOR RESOURCES OPTIMISATION IN A CLOUD-NATIVE 5G SYSTEM ARCHITECTURE WHICH USES SERVICE-BASED INTERFACES

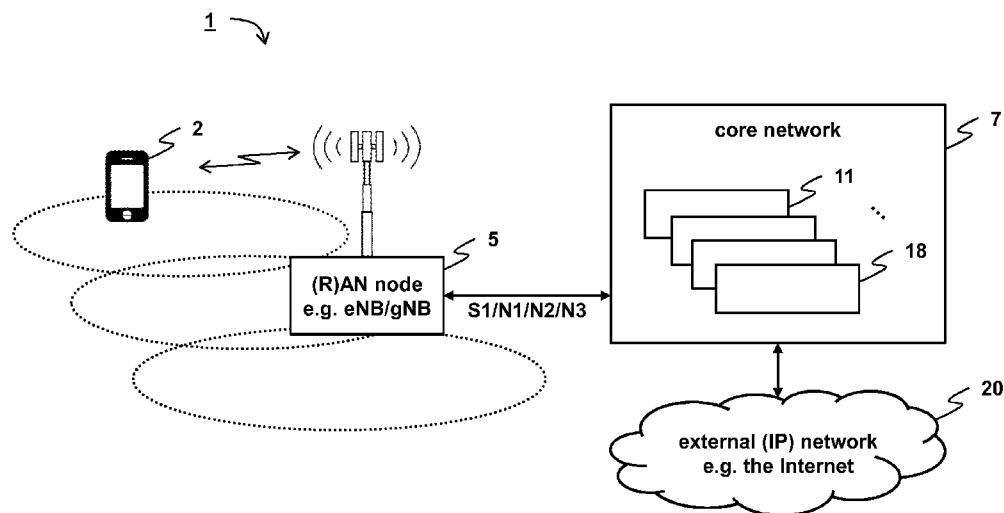


Figure 1

(57) Abstract: A method is described that leverages data analytics and machine learning to enable continued adjustment and optimisation of allocated and used resources for the communication in a 5G System Architecture per the service-based architecture principles, while meeting expected performance figures. This is of particular value in a cloud deployment of control plane functions per the 5G System Architecture, which utilise service-based interfaces and follow the associated principles of a Service Consumer and Service Producer. Optimisation is based on selected features from mobile subscriber/device context, such as mobility and communication patterns, connection status and delay tolerance on the control-and/or data plane.



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))*

**Leveraging data analytics for resources optimisation in a cloud-native 5G system
architecture which uses service-based interfaces**

The present invention relates to a communication system. The invention has particular but not exclusive relevance to wireless communication systems and devices thereof operating according to the 3rd Generation Partnership Project (3GPP) standards or equivalents or derivatives thereof. The invention has particular although not exclusive relevance to data analytics for resources optimisation in the so-called '5G' (or 'Next Generation') systems.

The next generation mobile communication system adopted the service-based architecture principles for Representational State Transfer (REST) communication in between functions of the core network control plane. Selected protocols utilise Hyper Text Transfer Protocol (HTTP) and a JavaScript Object Notation ('json') message body for message serialisation, whereas the communication is protected by Transport Layer Security (TLS) and transported over Transport Control Protocol (TCP). Future adoption of alternative transport protocols, such as Quick UDP Internet Connections (QUIC), is also considered.

Adoption of service-based interfaces implies that each network function represents a self-contained service which can be accessed through open interfaces per Open API description. Instances of network functions announce their availability by registration with a repository function, which is considered part of the 5G system architecture (as shown in Figure 5) [1]: the Network Repository Function (NRF).

Figure 5 illustrates schematically a mobile (cellular or wireless) telecommunication system to which embodiments of the invention may be applied (5G System Architecture – Service-based interface representation). Functions of the 5G Core control plane include the Policy Control Function (PCF), the Session Management Function (SMF), the Access and Mobility Management Function (AMF), as well as some others, such as for network exposure to third parties (NEF) and an Application Function (AF), to name a few. These functions are further defined and explained in 3GPP Technical Specification (TS) 23.501 and TS 23.502. Notation-wise, service-based interfaces are denoted by function name with a preceding capital letter N, e.g. Nnrf, Naf, Nsmf, etc. Numbered reference points, such as N4 in between the SMF and the a User Plane Function (UPF) still consider other communication protocols, which are not service-based but can be protocols such as the General Packet Radio Services (GPRS) Tunneling Protocol (GTP-C).

By utilising open interfaces and a service registry, REST-based communication enables a lot of flexibility, such as ease of integration of new services, enriching the features of a

service, enriching the semantics by means of extensions to the API, or scaling by creating more instances of a service and balance the load in between them.

Traditional mobile communication networks, such as the Evolved Packet System (EPS), utilise proven protocols in the core network, such as GTP-C [3] or Diameter [4], in between two dedicated functions, which represent the protocol endpoints. Moving the communication in between core network functions to REST per the service-based architecture principles implies various challenges in terms of performance, mainly transactions per second and transaction latency. Use of connected-oriented transport layer protocol, most notably TCP, requires establishment and maintenance of the connection between two endpoints. In addition, inclusion of security mechanism, such as TLS, introduces extra processing and resource usage. Therefore, communication services that are based on TLS over TCP, such as Hyper Text Transfer Protocol Secure (HTTPS) require processing and bring about various resource usage implications. In this sense, scaling and associated provisioning of additional instances of a network/service function enables load balancing but potentially increases the number of functions that maintain transport and connection states (http session states, TLS session states, TCP session states).

As discussed above, the set up and maintenance of https sessions is costly (latency, resources), as depicted in Figure 6, which illustrates schematically an exemplary and generalised session establishment (a TLS setup sequence including a TCP connection setup part, a TLS initial crypto negotiation part, a certificate exchange and encryption start part, and a http(s) message exchange part).

Web-based communication may tear down an http/TLS/TCP session between a client and a server after the requested service has been completed, e.g. all requested information has been transferred to the client, or all information has been posted from the client to the server. However, client and server may keep an http and the associated TLS session open for a certain period according to a configurable lifetime, even if there is no further transaction ongoing.

Communication between control plane instances of the 5G system architecture assumes peers of Service Consumer and Service Producer. The following two types of interactions are defined in [1]: Request – Response, and Asynchronous Subscribe – Notify. These service consumer – producer interactions are illustrated in Figure 7.

In terms of performance, REST-based communication may be constrained in the transaction rate which can be partially compensated by distributing/balancing clients'

service request load to multiple instances of the same serving function type. Balancing of requests to a Service Producer (server role) can be either performed by the Service Consumer (client role) or by a load balancer. This is of particular relevance for a cloud-native design and deployment, where Service Consumer/Producer instances are realised by virtual machines or containers and the number or instances can be increased or reduced per the deployed scaling procedure.

In case the Service Consumer performs load balancing, it may receive multiple candidate service producers' identifiers or addresses (e.g. from the repository/registry) along with complementary data for each service producer instance, such as available capacity, current load, etc. for proper selection of one instance of a Service Producer where to send the request to.

Alternatively, the Service Consumer's request is received by a balancing function, which distributes requests between multiple instances of the addressed service producer. Different algorithms can apply how to distribute requests. Figure 8 illustrates schematically an exemplary way of balancing transaction traffic between multiple stateful service producer instances. Specifically, Figure 8 depicts one example how a Service Producer's request may be treated by a balancer function and forwarded to one selected instance of the addressed service producer type. A Service Consumer handles states (session, location, status) of multiple subscribers and/or devices, and a Service Producer instance may need to hold such state for the subscriber/device, too. Storing such state at the Service Producer instance (as depicted in Figure 8) creates a binding between the subscriber/device and the selected Service Producer instance, as subsequent requests in the context of the same subscriber/device need to be processed at the same, previously selected Service Producer instance. Such locally stored states are advantageous in enabling fast access to the information, but increase demand for resources on each of the service producer instances. In addition, if there is a need for relocating a particular Service Consumer's service from one Service provider instance to another, proper maintenance and management of such resources becomes necessary.

Externalisation of states is considered to mitigate the resources issues, where, for example, Service Producer instances read/write states from/to and external database for each transaction (SDL in [2], UDR in [1]). This potentially enables re-selection of a service producer instance for each transaction for a single subscriber/device as the state is externalised. However, multiple Service Producer instances may create a lot of load for maintaining (e.g. writing and acquiring) states during each transaction for millions of subscribers/devices. This situation is depicted in Figure 9. Specifically, Figure 9 illustrates

schematically an exemplary way of balancing transaction traffic between multiple transport connections to stateless service producer instances.

Due to the constraints of a transport connection between a Service Consumer and a Service Producer instance, multiple parallel connections can be setup and maintained.

5 Such constraints are in particular a result of head-of-line (HOL) blocking issues with TCP but also exhaustion of stream identifiers in http/2, where a new transport connection needs to be setup in order to support the re-use of already used stream identifiers. Current documentation per 3GPP considers 'a few transport connections' in between functions.

10 The following list summarises the problems associated with scalable deployment of service-based interfaces in between Service Consumer- and Service Producer instances for a 5G mobile communication system:

- 15 • Long-term binding between Service Consumer and stateful Service Producer instances (instance-local session states) for a subscriber/device. Hinders intended state depletion at service instances, e.g. to scale-in. High resources requirements at service instances, in particular for storage.
- Increased transaction latency and traffic load when states are externalised (stateless Service Producer instances)
- Negative performance impact of too few parallel transport connections between Service Consumer and Service Producer instances
- 20 • Superfluous resources (transport connection states) at service instances in case too many transport connections are maintained in parallel
- High signaling load and additional latency for frequent setup/teardown of secure transport connections in between Service Consumer and service producer instances.

25 Accordingly, the present invention seeks to provide methods and associated apparatus that address or at least alleviate (at least some of) the above described problems.

Although for efficiency of understanding for those of skill in the art, the invention will be described in detail in the context of a 3GPP system (5G networks), the principles of the invention can be applied to other systems as well.

30 In one aspect, the invention provides a method performed by a network function, the method comprising: providing at least one transport connection, between a service

consumer function and a respective service producer function, for communicating data over said transport connection; and determining whether to store, at an external storage, subscriber/session data relating to said at least one transport connection at said service producer function based on information associated with a subscriber / user equipment to which said data relates.

In one aspect, the invention provides a method performed by a network function, the method comprising: obtaining information identifying whether a current number of available transport connections between a service producer function and at least one service consumer function meets a delay budget associated with subscribers/devices served by the service producer function; and determining, based on said obtained information, whether to add or remove at least one transport connection between the service producer function and at least one service consumer function.

In one aspect, the invention provides a method performed by a network function, the method comprising: collecting and analysing data relating to transport connections in a network comprising a plurality of service consumer functions and a plurality of service producer functions; maintaining rules relating to said transport connections based on said data; and providing said rules to at least one of a service consumer function, a service producer function, and a balancer function for managing at least one transport connection.

In one aspect, the invention provides a network function comprising: means for providing at least one transport connection, between a service consumer function and a respective service producer function, for communicating data over said transport connection; and means for determining whether to store, at an external storage, subscriber/session data relating to said at least one transport connection at said service producer function based on information associated with a subscriber / user equipment to which said data relates.

In one aspect, the invention provides a network function comprising: means for obtaining information identifying whether a current number of available transport connections between a service producer function and at least one service consumer function meets a delay budget associated with subscribers/devices served by the service producer function; and means for determining, based on said obtained information, whether to add or remove at least one transport connection between the service producer function and at least one service consumer function.

In one aspect, the invention provides a network function comprising: means for collecting and analysing data relating to transport connections in a network comprising a plurality of service consumer functions and a plurality of service producer functions; means for

maintaining rules relating to said transport connections based on said data; and means for providing said rules to at least one of a service consumer function, a service producer function, and a balancer function for managing at least one transport connection.

5 In one aspect, the invention provides a network function comprising a controller, and a transceiver, wherein the controller is configured to: provide at least one transport connection, between a service consumer function and a respective service producer function, for communicating data over said transport connection; and determine whether to store, at an external storage, subscriber/session data relating to said at least one transport connection at said service producer function based on information associated with a
10 subscriber / user equipment to which said data relates.

In one aspect, the invention provides a network function comprising a controller, and a transceiver, wherein the controller is configured to: obtain information identifying whether a current number of available transport connections between a service producer function and at least one service consumer function meets a delay budget associated with
15 subscribers/devices served by the service producer function; and determine, based on said obtained information, whether to add or remove at least one transport connection between the service producer function and at least one service consumer function.

In one aspect, the invention provides a network function comprising a controller, and a transceiver, wherein the controller is configured to: collect and analyse data relating to
20 transport connections in a network comprising a plurality of service consumer functions and a plurality of service producer functions; maintain rules relating to said transport connections based on said data; and provide said rules to at least one of a service consumer function, a service producer function, and a balancer function for managing at least one transport connection.

25 Aspects of the invention extend to corresponding systems, and computer program products such as computer readable storage media having instructions stored thereon which are operable to program a programmable processor to carry out a method as described in the aspects and possibilities set out above or recited in the claims and/or to program a suitably adapted computer to provide the apparatus recited in any of the claims.

30 Each feature disclosed in this specification (which term includes the claims) and/or shown in the drawings may be incorporated in the invention independently of (or in combination with) any other disclosed and/or illustrated features. In particular but without limitation the features of any of the claims dependent from a particular independent claim may be introduced into that independent claim in any combination or individually.

The invention will now be described, by way of example only, with reference to the accompanying drawings in which:

Figure 1 schematically illustrates a mobile (cellular or wireless) telecommunication system to which the described embodiments are applicable;

- 5 Figure 2 is a block diagram illustrating the main components of the UE shown in Figure 1;
- Figure 3 is a block diagram illustrating the main components of an exemplary (R)AN node shown in Figure 1; and
- Figure 4 is a block diagram illustrating the main components of a generic core network node shown in Figure 1;
- 10 Figure 5 illustrates further details of the system shown in Figure 1;
- Figure 6 illustrates schematically an exemplary and generalised session establishment procedure;
- Figure 7 illustrates schematically the difference between two types of interactions between a service customer and a service producer;
- 15 Figure 8 illustrates schematically an exemplary procedure for balancing transaction traffic between multiple stateful service producer instances;
- Figure 9 illustrates schematically an exemplary procedure for balancing transaction traffic between multiple transport connections to stateless service producer instances;
- Figure 10 illustrates schematically an exemplary way of performing data analytics of subscriber/device personalized data and statistics;
- 20 Figure 11 illustrates schematically an exemplary procedure for context-based balancing of subscriber/device states, and scaling of transport connections;
- Figure 12 illustrates schematically exemplary network and service functions and the orchestrator in accordance with an embodiment of the present invention; and
- 25 Figure 13 is a schematic signalling (timing) diagram illustrating an exemplary procedure for analytics-based decisions to store session states and to scale transport connections in between service functions.

Overview

- 30 Figure 1 schematically illustrates a mobile (cellular or wireless) telecommunication system 1 to which the above embodiments are applicable.

In this network, users of mobile devices 2 (UEs) can communicate with each other and other users via respective base stations 5 and a core network 7 using an appropriate 3GPP radio access technology (RAT), for example, an E-UTRA and/or 5G RAT. It will be appreciated that a number of base stations 5 form a (radio) access network or (R)AN. As
35 those skilled in the art will appreciate, whilst one mobile device 2 and one base station 5

are shown in Figure 1 for illustration purposes, the system, when implemented, will typically include other base stations and mobile devices (UEs).

Each base station 5 controls one or more associated cells (either directly or via other nodes such as home base stations, relays, remote radio heads, and/or the like). A base station 5 that supports E-UTRA/4G protocols may be referred to as an 'eNB' and a base station 5 that supports NextGeneration/5G protocols may be referred to as a 'gNBs'. It will be appreciated that some base stations 5 may be configured to support both 4G and 5G, and/or any other 3GPP or non-3GPP communication protocols.

The mobile device 2 and its serving base station 5 are connected via an appropriate air interface (for example the so-called 'Uu' interface and/or the like). Neighbouring base stations 5 are connected to each other via an appropriate base station to base station interface (such as the so-called 'X2' interface, 'Xn' interface and/or the like). The base station 5 is also connected to the core network nodes via an appropriate interface (such as the so-called 'S1', 'N1', 'N2', 'N3' interface, and/or the like).

The core network 7 typically includes logical nodes (or 'functions') for supporting communication in the telecommunication system 1. Typically, for example, the core network 7 of a 'Next Generation' / 5G system will include, amongst other functions, control plane functions and user plane functions. In this example, the core network 7 includes, amongst others, the network nodes described above with reference to Figure 12: a Service Consumer function 11, a balancer 13, one or more nodes providing transport connections 14, a plurality of Service Producers 15 (per type), an external storage/database 17, a network 16 connecting the external storage/database 17 with the Service Producer instances 15, and a data analytics and resources orchestration function 18. From the core network 7, connection to an external IP network 20 (such as the Internet) is also provided.

It will be appreciated that the balancer 13 and the orchestration function 18 may not necessarily be implemented as functions within the core network 7. Either one of these may be provided as a component of a cloud network or a platform (coupled to the core network 7), which may be implementation specific. However, the balancer 13 and the orchestration function 18 are network functions to enable/optimize the service based communication in between the functions (including core network functions) of the system 1.

The network functions (the above described network nodes) are configured to carry out REST-based communication with each other (at least for the control plane). For example, a network function may be configured to operate as a service producer or as a service

consumer (although it will be appreciated that a given network function may have different roles with respect to different services).

The network functions (nodes) are configured to perform optimisation, for example:

- (A) at service function instances for the storage and maintenance of subscriber/device states, such as session states, and
- (B) at service function instances and/or load balancers for the setup and maintenance of secure transport connections, e.g. TCP/TLS or QUIC, under consideration of expected service quality and performance from the mobile subscriber and a device's point of view.

User equipment (UE)

Figure 2 is a block diagram illustrating the main components of the UE (mobile device 2) shown in Figures 1 and 12. As shown, the UE includes a transceiver circuit 31 which is operable to transmit signals to and to receive signals from the connected node(s) via one or more antenna 33. Although not necessarily shown in Figure 2, the UE will of course have all the usual functionality of a conventional mobile device (such as a user interface 35) and this may be provided by any one or any combination of hardware, software and firmware, as appropriate. A controller 37 controls the operation of the UE 2 in accordance with software stored in a memory 39. The software may be pre-installed in the memory 39 and/or may be downloaded via the telecommunication network 1 or from a removable data storage device (RMD), for example. The software includes, among other things, an operating system 41 and a communications control module 43. The communications control module 43 is responsible for handling (generating/sending/receiving) signalling messages and uplink/downlink data packets between the UE 2 and other nodes, including (R)AN nodes 5 and core network nodes.

(R)AN node

Figure 3 is a block diagram illustrating the main components of an exemplary (R)AN node 5 (base station) shown in Figure 1. As shown, the (R)AN node 5 includes a transceiver circuit 51 which is operable to transmit signals to and to receive signals from connected UE(s) 2 via one or more antenna 53 and to transmit signals to and to receive signals from other network nodes (either directly or indirectly) via a network interface 55. The network interface 55 typically includes an appropriate base station – base station interface (such as X2/Xn) and an appropriate base station – core network interface (such as S1/N1/N2/N3). A controller 57 controls the operation of the (R)AN node 5 in accordance with software stored in a memory 59. The software may be pre-installed in the memory 59 and/or may be downloaded via the telecommunication network 1 or from a removable data storage device (RMD), for example. The software includes, among other things, an operating system 61

and a communications control module 63. The communications control module 63 is responsible for handling (generating/sending/receiving) signalling between the (R)AN node 5 and other nodes, such as the UE 2 and the core network nodes. The communications control module 63 is also responsible for communicating data packets for each UE 2 using a network slice that is appropriate for that UE 2.

Core network node

Figure 4 is a block diagram illustrating the main components of a generic core network node (or function), for example, the Service Consumer function 11, the Balancer 13, a node of the transport connection 14, the Service Producer 15, a node of the network 16, the external storage/database 17, and the data analytics and resources orchestration function 18 shown in Figure 1. As shown, the core network node includes a transceiver circuit 71 which is operable to transmit signals to and to receive signals from other nodes (including the UE 2 and the (R)AN node 5) via a network interface 75. A controller 77 controls the operation of the core network node in accordance with software stored in a memory 79. The software may be pre-installed in the memory 79 and/or may be downloaded via the telecommunication network 1 or from a removable data storage device (RMD), for example. The software includes, among other things, an operating system 81 and a communications control module 83. The communications control module 83 is responsible for handling (generating/sending/ receiving) signaling between the core network node and other nodes, such as the UE 2, (R)AN node 5, and other core network nodes.

Detailed description

The following exemplary embodiments enable optimisation of resources, which are used:

- (A) at service function instances for the storage and maintenance of subscriber/device states, such as session states, and
- (B) at service function instances and/or load balancers for the setup and maintenance of secure transport connections, e.g. TCP/TLS or QUIC, under consideration of expected service quality and performance from the mobile subscriber and a device's point of view.

The following is a detailed description of:

- a system and a method to enable and optimise the operation per the communication principles of a service-based architecture by using context information
- a system and method to efficiently determine the time when to externalise some subscribers'/devices' states, e.g. session states, from a network/service function

instance, in particular states associated with delay tolerant subscribers/devices, to save storage resources at the service/network function instance and to ease state depletion at the service/network function instance, while other subscribers'/devices' states are kept and maintained at the network/service function instance to minimise latency to the states for delay-intolerant subscribers/devices

- This enables better long-term balancing of load in between service/network function instances by re-selecting a new instance for a subsequent service request, which pulls the session state from the external database
- This eases intentional depletion of states at a service/network function instance for scale-in (e.g. shut down a service/network function instance)
- Decisions to not externalise some subscriber/device states at all or for some time enables faster access to the states. Such decisions are in particular taken for delay-intolerant subscribers/devices

- procedures for context-based balancing of service requests between multiple available instances of a receiving service function by forwarding new requests, which are associated with subscribers/devices whose states can be externalised (delay-tolerance) or can be deprecated (short connection lifetime, frequent detachment) soon, to dedicated service/network function instances.

- This enables immediate or early externalisation of associated session states and eases state depletion at these dedicated service/network function instances in support of scale-in

- a system and a method to scale the number of parallel transport connections in between a Service Consumer and a Service Producer or an associated load balancer based (A) on the current load of the Service Consumer instance and/or the Service Producer instance as well as (B) on the delay-tolerance of subscribers/devices which result in the communication between a Service Consumer and a Service Producer.

- This enables guaranteed transport capacity and expected performance (e.g. low latency) in particular for communication in the context of subscribers/ service transactions in the subscriber
- This enables the service function instance to have a set of transport connections available and perform transmission of a service request (at the Service Consumer) or a Notification (at the Service Producer) per a scheduling algorithm, e.g. select one transport connection which suits the expected service

level for the request per the analysed context of the associated subscriber/device.

Figure 10 illustrates schematically exemplary data analytics of subscriber/device personalised data and statistics as well as service function load (input) to optimise the balance between performance and used resources (output).

System and method for optimisation is based on the following facts:

- Devices, with frequent handover and communication patterns (frequent and lengthy data sessions), create frequent signaling and require frequent access to the subscriber's/device's session states.
- Frequent access (e.g. read/write) to session states, e.g. for each transaction in the context of a subscriber/device, results in increased network load in case states are externalised, e.g. using an external storage or database.
- Access to externalised states introduces additional latency to a signaling transaction
- Frequent access to externalised states can create high load on the external storage/database and introduce additional queueing delay.
- Externalisation of states reduced the need for local storage at service function instances
- Localised session states minimise latency for access (e.g. read/write) to the states
- Subscribers/Devices with low delay tolerance expect low latency communication, fast signaling procedure for paging and data plane restoration (e.g. bearer setup), as well as packet forwarding
- Multiple parallel transport connections in between two instances of service functions, e.g. Service Consumer and Service Producer, or in between a Service Consumer and a Service Producer load balancer increase the load at the service function instance or the load balancer respectively due to the maintenance of transport connection states and signaling to lifecycle manage each transport connection.
- Multiple parallel transport connections in between two instances of service functions, e.g. Service Consumer and service producer, or in between a Service Consumer and a Service Producer load balancer increase the throughput/capacity and reduce the latency for the service-based communication in between a Service Consumer and a Service Producer.

System and method operation apply the following principles:

- States associated with delay-tolerant subscribers/devices in connected state are externalised early
- States associated with subscribers/devices which are delay-tolerant on the control plane only if it has no impact on the data plane (tolerant to larger signaling latency, e.g. paging and activation from DRX mode) but delay-intolerant on the data plane (fast handover signaling) are externalised when the device is in idle mode (no data communication) and/or in DRX mode but kept on the service instance as residential state otherwise if the expected delay budget for accessing externalised states exceeds the tolerable delay, e.g. in dependency of the network performance and load in between service/network function instances and the external storage. This principle is in particular followed for subscribers/device with large idle periods (long idle phases in between two consecutive communication phases)
- States associated with subscribers/devices which are delay-intolerant are kept on the service instance as residential state if the expected (anticipated) or predicted delay budget for accessing externalised states exceeds the tolerable delay, e.g. in dependency of the network performance and load in between service/network function instances and the external storage, as well as load on the external storage/database.
- States associated with subscribers/devices which have frequent data sessions and/or cause frequent handover/location updates can be kept on the service instance as residential state if the expected traffic to access the externalised states results in exceeding a given load budget on the network, on the external storage/database and/or on the service function instance.
- Service requests for delay-intolerant subscribers/devices are directed/balanced (e.g. by context-based balancing) to service/network function instances which are permanent or long-lived and not meant for (frequent) shutdown, e.g. during scale-in.
- Service requests for delay-tolerant subscribers/devices are directed (e.g. by context-based balancing) to service/network function instances which can hold residential states for a limited time but can either externalise or delete states frequently. This enables complete state deletion and teardown of such instance, e.g. for scale-in.
- Peers of service/network function instances (e.g. Service Consumer and Service Producer), which experience high load on the service-based interfaces increase the number of transport connections in between them.
- Peers of service/network function instances, which treat service communication for delay-intolerant subscribers/devices, increase the number of transport connections

in between them to have sufficient transport capacity with low transport delay available.

- Peers of service/network function instances, which experience low load on the service-based interfaces and probability of unexpected peaks in such load is low, can reduce the number of transport connections.
- Service/network function instances export residential states in bulk to the external state storage/database, either even-based, on request or periodically. This keeps network load to minimum but enables holding a snapshot of residential states as a copy and supports recovery procedures in case of service function instance full or partial failure.
- The described principles can go hand in hand with additional decision algorithms and associated operations for the externalisation of states, the scale of transport connections or the scale of service/network function instances.

Processing of data that reveals subscribers'/devices' mobility- and communication patterns, load information and delay-tolerance is performed in an advantageous way on one or multiple dedicated functions for analytics and resources orchestration, which controls multiple service/network function instances and, if applicable, (context-based) load balancer. The resource orchestrator provides rules to the service/network functions for resources optimisation per the above description. This includes decisions or rules when to scale the transport connections, e.g. by adding/removing transport connections or changing the characteristics (e.g. capacity, performance) of a transport connections, as well as rules that apply to a scheduler, which places requests, or notifications respectively, on a suitable transport connection. Furthermore, the resources orchestrator provides decisions or rules for taking decisions on a service/network function instance whether and when to externalise subscriber/device states. Such decision or rules can address a group of subscriber/device states or states associated with an individual subscriber/device. Some information, e.g. transport connection RTT, window sizes, etc, which is valuable to determine the need to scale the number of transport connections, may be available on a service/network function instance or a transport controller. In such case, the information can either be provided to the resource orchestrator, or the service/network (or load balancer) function takes the local information into account together with the resources optimisation rules as provided by the resource orchestrator. The association between resource orchestrator and service/network functions is exemplarily depicted in Figure 11. A preferred deployment considers also provisioning of decisions or rules for balancing of requests or subscriptions to one of multiple available receiving service/network function

instances to a load balancer. This enables the balancer to make smart decisions based on context. Such balancer is not covered in Figure 11 but can be seen in Figure 8.

Figure 11 illustrates schematically an exemplary way of performing context-based balancing of subscriber/device states, which are hold at the service/network function instance as residential states or externalised, and scaling of transport connections.

Determination of load figures (service function instance load, transport connection load, transport connection latency, network load/latency that has impact to the access to an external states storage) and delay tolerance of subscribers/devices can be based on information which is collected from different functional elements. For example, service/network functions, load/context balancers or resources orchestrator can retrieve information about service/network function load from the VNF Management and Orchestration System via and API gateway, or from a dedicated data analytics function. Information about subscriber/device delay tolerance may be, for example, retrieved from a profile database or statistics database, e.g. analysing events of violated service level agreement (SLA) for a subscriber/device. Information about subscriber/device communication patters (data session frequency/length, idle periods, DRX intervals and mode duration, rate of incoming/originating data sessions), as well as mobility patterns can come from an associated monitoring function and statistics database. Information about network load can be determined, for example, via netstat tools, transport network/SDN controllers or from functions associated with Virtual Infrastructure Management (VIM).

Resources optimisation per this present application is based on data analytics and can be enhanced by machine learning algorithms to enable on the one hand side an optimised balance between performance expectation and allocated/used resources, on the other side means to detect and predict anomalies, such as unexpected load peaks, function failures or network congestion, and enable in-advance scaling of resources (e.g. add transport connections, add service/network function instances, add storage to service/network function instance).

The exemplary embodiments aim to:

- Optimise resources usage in alignment with performance expectations
- Optimise treatment of transaction-latency sensitive vs. delay-tolerant functions/subscribers/applications
- Dynamically adapt to the system load

It will be appreciated that resources (service/network function instance storage, and transport connections) may be tailored per at least one of the following attributes:

- Subscriber/device mobility pattern (e.g. statistics). Handover – data analytics, prediction, explicit signaling
- Communication pattern (e.g. statistics), session duration, idle periods, Active/Dormant transitions, paging activity – data analytics, prediction, explicit signaling
- Network load, service function instance load (compute and storage), external storage/database load and associated latency – data analytics, prediction, explicit signaling
- Delay tolerance - per subscription/device or per group
- Profile information (SLA, subscription contract, ...) - per subscriptions
- Service requirement (e.g. QoS)
- Traffic information (e.g. application being used, indicating the traffic type and intensity),
 - Shorter/longer Transmission Time Interval (TTI)
 - Shorter/longer average packet size
 - Constant/bursty

Figure 12 illustrates schematically exemplary network- and service functions and the orchestrator in accordance with an embodiment of the present invention.

In more detail, Figure 12 shows a Service Consumer function 11, which handles multiple subscribers/devices 2, a balancer 13, transport connections 14 in between the Service Consumer 11 and the balancer 13 or the Service Producer instances, multiple instances of a type of Service Producer 15 (Service Producer 15-1 and 15-2) whose service is required by the Service Consumer 11, an external storage/database 17, a network 16, which connects the external storage/database 17 with the Service Producer instances 15, and a data analytics and resources orchestration function 18 (short: Orchestrator).

In the following exemplary sequence, it is assumed that the balancer 13 terminates transport connections, hence enforces rules from the orchestrator 18 about scaling transport connections. Alternatively, transport connections can be terminated in Service Producer instances 15. In both cases, the orchestrator 18 may alternatively or in addition enforce rules for scaling up/down transport connections' characteristics, such as performance, latency, priority, at a transport controller (not shown) to have directly impact to the transport connection performance in between Service Consumers and Service Producers 15. Details how the Orchestrator 18 collects and analyses data points are not shown, but multiple options exist. The Orchestrator 18 may already receive analysed data or collect data from one or multiple sources about one or multiple values associated with

subscriber/device profiles, mobility- and/or communication patterns/statistics, status patterns/statistics, network load/statistics, current performance and performance statistics of the transport connections which are used for the communication between Service Producer 15 and Service Consumer. Two types of subscribers/devices service levels in terms of delay intolerance are assumed, which are 'no' and 'medium' delay tolerant. A sequence with an exemplary embodiment is depicted in Figure 13 and described below.

Figure 13 illustrates schematically an exemplary embodiment of analytics-based decisions to store session states at a local or external (e.g. database) storage and to scale transport connections in between service functions.

10 [1] Based on collected and analysed data points, the orchestrator 18 creates rules or takes decisions for traffic balancing, session state treatment (local storage vs. external storage) and transport connections 14 for enforcement at the service functions and the balancer 13. The orchestrator 18 may enforce rules or concrete decisions to, for example, a network resources controller to scale/configure the characteristics of existing transport connections.

15 [2] The orchestrator 18 conveys rules/decisions to the enforcement points.

[3] The Service Consumer 11 sends a request in the context of a low delay tolerant subscriber/device towards the Service Producer 15, which is balanced according to the applied balancing rules to Service Producer instance 15-1.

20 [4] Per the applied storage treatment rules, the Service Producer 15-1 stores the session state on its local storage.

[5] The Service Consumer 11 sends a request in the context of a medium delay tolerant subscriber/device towards the Service Producer 15, which is balanced according to the applied balancing rules to Service Producer instance 15-2.

25 [6] Per the applied storage treatment rules, the Service Producer 15-2 stores the session state on the external database/storage 17, as per the rule the overall transaction latency is anticipated to meet the associated delay bounds.

30 [7] Based on further collected and analysed data points, the orchestrator 18 creates new or updated rules, which take, for example, into account a more loaded network and database/external storage 17, that result in an increased delay portion to the service transaction latency.

[8] The orchestrator 18 conveys rules/decisions to the enforcement points.

[9] The Service Consumer 11 sends a request in the context of yet another medium delay tolerant subscriber/device to the Service Producer 15, which is balanced according to the applied balancing rules to Service Producer instance 15-2.

[10] Per the applied updated storage treatment rules, the Service Producer 15-2 stores the session state on its local storage, as the anticipated delay bounds may likely exceed an acceptable value according to the changed situation in the network, service functions and database.

[11] The orchestrator 18 anticipates an improved performance by scaling the transport connections 14, which can lead to an optimisation in the use of resources at Service function instances 15. This may be due to heavy load on too few or too performance constrained transport connections 14.

[12] The orchestrator 18 conveys rules/decisions to the enforcement points.

[12'] The orchestrator 18 may update transport connection endpoints (Service Consumer 11 and load balancer 13 or Service Producer instances 15) to scale transport connections 14 by setting up additional transport connections or changing the configuration associated with one or multiple transport connections 14.

[12''] The orchestrator 18 may update the transport resources controller to scale up the existing transport connections 14 to achieve better performance.

[13] The Service Consumer 11 sends a request in the context of yet another medium delay tolerant subscriber/device to the Service Producer 15, which is balanced according to the applied balancing rules to Service Producer instance 15-2.

[14] Per the applied updated storage treatment rules, the Service Producer 15-2 stores the session state on the external database/storage 17, as per the rule the overall transaction latency is anticipated to meet the associated delay bounds after scaling the transport connections 14.

Summary

Beneficially, the above described exemplary embodiments include, although they are not limited to, one or more of the following functionalities:

- 1) A functionality to determine whether to store subscriber/device states on an external storage/database or on a service/network function instance based on the subscriber's/device's mobility- and communication pattern, as well as on its delay-tolerance on the control- and data plane.

- 2) A functionality to determine when to externalise subscriber/device states based on subscriber's/device's connection- and communication status (Idle, DRX mode) and delay-tolerance.
- 3) A functionality to determine whether to increase or decrease the number of transport connections in between a Service Consumer instance and a Service Producer Instance or in between a Service Consumer instance and a load balancer, which balances requests between multiple instances of a Service Producer.

It can be seen that in the above embodiments describe exemplary methods for the optimisation of used resources in the deployment and operation of a service-based architecture while meeting expected performance figures.

These exemplary methods may comprise at least one of the following steps:

- 1) Determining when to externalise a subscriber/device state based on the subscriber's/device's mobility- and communication characteristics, and its delay tolerance.
- 2) Determining when to add or remove transport connections and associated capacity in between a Service Consumer instance and a Service Producer instance, or in between a Service Consumer instance and a further function, such as a load balancer or proxy, which terminates transport connections on behalf of one or multiple Service Producer instances, based on the anticipation/prediction of whether or not the load on the current number of available transport connections meets the delay budget of all treated subscribers/devices with low delay-tolerance.

Benefits

It can be seen that the above embodiments provide at least one of the following benefits:

- Provides solution to recognised problems with REST communication for a 5G System Architecture control plane per service-based architecture principles.
- Enables carrier-grade performance for mobile control plane with stringent requirements on transaction rate and transaction latency while using REST communication principles with Service Consumer and Service Producer instances, which are deployed and managed using cloud computing.
- Leverages data analytics and machine learning to dynamically adjust allocated and used resources for the deployment and operation of Service Consumers and Service Producers under consideration of (mobile) subscriber characteristics.
- Enables optimisation without interfering with 3GPP standardisation → vendor differentiation.

Modifications and Alternatives

Detailed embodiments have been described above. As those skilled in the art will appreciate, a number of modifications and alternatives can be made to the above embodiments whilst still benefiting from the inventions embodied therein. By way of illustration only a number of these alternatives and modifications will now be described.

In the above description, the UE, the (R)AN node, and the core network node are described for ease of understanding as having a number of discrete modules (such as the communication control modules). Whilst these modules may be provided in this way for certain applications, for example where an existing system has been modified to implement the invention, in other applications, for example in systems designed with the inventive features in mind from the outset, these modules may be built into the overall operating system or code and so these modules may not be discernible as discrete entities. These modules may also be implemented in software, hardware, firmware or a mix of these.

Each controller may comprise any suitable form of processing circuitry including (but not limited to), for example: one or more hardware implemented computer processors; microprocessors; central processing units (CPUs); arithmetic logic units (ALUs); input/output (IO) circuits; internal memories / caches (program and/or data); processing registers; communication buses (e.g. control, data and/or address buses); direct memory access (DMA) functions; hardware or software implemented counters, pointers and/or timers; and/or the like.

In the above embodiments, a number of software modules were described. As those skilled in the art will appreciate, the software modules may be provided in compiled or un-compiled form and may be supplied to the UE, the (R)AN node, and the core network node as a signal over a computer network, or on a recording medium. Further, the functionality performed by part or all of this software may be performed using one or more dedicated hardware circuits. However, the use of software modules is preferred as it facilitates the updating of the UE, the (R)AN node, and the core network node in order to update their functionalities.

The above embodiments are also applicable to 'non-mobile' or generally stationary user equipment. The UE may be a device or a part of a system that provides applications, services, and solutions for "Internet of Things" (IoT) or Machine-Type Communications (MTC), using a variety of wired and/or wireless communication technologies. It will be appreciated that the UE may operate without requiring human supervision or interaction.

The method may further comprise the network function requesting the service producer function to store said subscriber/session data at the service producer function when it is determined that said subscriber/session data is not to be stored at an external storage. The determining may be based on at least one rule relating to said at least one transport connection. In this case, the method may further comprise the network function obtaining at least one rule from an orchestrator function configured to collect and analyse data relating to transport connections in a network comprising said service consumer function and said service producer function.

The network function may comprise a balancer function and/or a service producer function.

The at least one transport connection may be terminated at the balancer function which may be located between the service consumer function and the service producer function.

The subscriber/session data may comprise at least one of: data identifying a state of the subscriber / user equipment, data identifying mobility characteristics of the subscriber / user equipment, data identifying communication characteristics of the subscriber / user equipment, and data identifying delay tolerance associated with data communicated for said subscriber / user equipment.

The determining may comprise determining whether to add or remove at least one transport connection and/or capacity between the service consumer function and a load balancer function or proxy which terminates transport connections on behalf of at least one service producer function.

The information identifying whether a current number of available transport connections meets a delay budget may comprise information identifying a current load or a remaining capacity associated with said transport connections.

The transport connection may utilise at least one of: Hyper Text Transfer Protocol (HTTP); JavaScript Object Notation (json); Transport Layer Security (TLS); Transport Control Protocol (TCP); and Quick UDP Internet Connections (QUIC).

The service producer function may comprise: a User Plane Function (UPF); a Network Slice Selection Function (NSSF); a Network Exposure Function (NEF); a Network Exposure Function (NRF); a Policy Control Function (PCF); a Unified Data Management (UDM) function; an Application Function (AF); an Authentication Server Function (AUSF); an Access and Mobility Management Function (AMF); a Session Management Function (SMF); a Service Control Point (SCP); or a (Radio) Access Node. Similarly, the service

consumer function may comprise: a UPF; an NSSF; a NEF; a NRF; a PCF; a UDM; an AF; an AUSF; an AMF; a SMF; a SCP; or a (R)AN node.

The data relating to transport connections may comprise at least one of: data relating to a state of a subscriber / user equipment, data relating to mobility characteristics of a subscriber / user equipment, data relating to communication characteristics of a subscriber / user equipment, data relating to a delay tolerance associated with a subscriber / user equipment; and data relating to a current load or a remaining capacity associated with at least one transport connection and/or a network function.

The rules may specify at least one of: whether to store control data relating to a particular transport connection at a service producer function associated with that transport connection or to store said control data at an external storage different to said service producer function associated with that transport connection; and whether to change a number of transport connections employed between a service producer function and a service consumer function in order to meet a delay budget associated with subscribers/devices served by that service producer function.

Various other modifications will be apparent to those skilled in the art and will not be described in further detail here.

References

- [1] 3GPP System Architecture for 5G System, TS 23.501 (f10), 03/2018
- [2] NGMN Alliance, Working Document, Service-based Architecture in 5G, v0.3.12, 12/2017
- [3] 3GPP TS 29.274 Tunneling Protocol for Control Plane (GTPv2-C) v15.3.0, 03/2018
- [4] IETF RFC 3588 Diameter based protocol, 09/2003

Abbreviations and terminology

3GPP	3rd Generation Partnership Project
AF	Application Function
AMF	Access and Mobility Management Function
API	Application Programming Interfaces
AUSF	Authentication Server Function
EPS	Evolved Packet System
GPRS	General Packet Radio Services
GTP-C	GPRS Tunneling Protocol
HOL	Head of Line
HTTPS	Hyper Text Transfer Protocol Secure

	NEF	Network Exposure Function
	NF	Network Function
	NF instance	An identifiable instance of the NF.
5	NF service	A functionality exposed by a NF through a service based interface and consumed by other authorized NFs.
	NF service instance	An identifiable instance of the NF service.
	NF service operation	An elementary unit a NF service is composed of.
10	NF Service Set	A group of interchangeable NF instances of the same type, supporting the same services and the same Network Slice(s). The NF instances in the same group of interchangeable NF service instances of the same service type within an NF instance. The NF service instances in the same NF Service Set have access to the same context data.
15	NF Set	NF Set may be geographically distributed but have access to the same context data.
	NRF	Network Repository Function
	NSSF	Network Slice Selection Function
	PCF	Policy Control Function
	QoS	Quality-of-Service
20	QUIC	Quick UDP Internet Connections
	REST	Representational State Transfer
	SCP	Service Control Point
	SDL	Shared Data Layer
	SLA	Service level agreement
25	SMF	Session Management Function
	TCP	Transport Control Protocol
	TLS	Transport Layer Security
	TTI	Transmission Time Interval
	UDM	Unified Data Management
30	UDR	Unified Data Repository
	UPF	User Plane Function

CLAIMS

1. A method performed by a network function, the method comprising:

providing at least one transport connection, between a service consumer function and a respective service producer function, for communicating data over said transport
5 connection; and

determining whether to store, at an external storage, subscriber/session data relating to said at least one transport connection at said service producer function based on information associated with a subscriber / user equipment to which said data relates.

2. The method according to claim 1, further comprising requesting said service
10 producer function to store said subscriber/session data at said service producer function when it is determined that said subscriber/session data is not to be stored at an external storage.

3. The method according to claim 1 or 2, wherein said determining is based on at least one rule relating to said at least one transport connection.

4. The method according to claim 3, further comprising obtaining said at least one rule
15 from an orchestrator function configured to collect and analyse data relating to transport connections in a network comprising said service consumer function and said service producer function.

5. The method according to any of claims 1 to 4, wherein said network function
20 comprises at least one of a balancer function and a service producer function.

6. The method according to any of claims 1 to 5, wherein the at least one transport connection is terminated at a balancer function between the service consumer function and the corresponding service producer function.

7. The method according to any of claims 1 to 6, wherein said subscriber/session data
25 comprises at least one of: data identifying a state of the subscriber / user equipment, data identifying mobility characteristics of the subscriber / user equipment, data identifying communication characteristics of the subscriber / user equipment, and data identifying delay tolerance associated with data communicated for said subscriber / user equipment.

8. A method performed by a network function, the method comprising:

obtaining information identifying whether a current number of available transport connections between a service producer function and at least one service consumer function meets a delay budget associated with subscribers/devices served by the service producer function; and

determining, based on said obtained information, whether to add or remove at least one transport connection between the service producer function and at least one service consumer function.

9. The method according to claim 8, wherein said determining comprises determining whether to add or remove at least one transport connection and/or capacity between the service consumer function and a load balancer function or proxy which terminates transport connections on behalf of at least one service producer function.

10. The method according to claim 8 or 9, wherein said information identifying whether a current number of available transport connections meets a delay budget comprises information identifying a current load or a remaining capacity associated with said transport connections.

11. The method according to any of claims 8 to 10, wherein the network function comprises at least one of a balancer function and the service producer function.

12. The method according to any of claims 1 to 11, wherein said transport connection utilises at least one of: Hyper Text Transfer Protocol (HTTP); JavaScript Object Notation (json); Transport Layer Security (TLS); Transport Control Protocol (TCP); and Quick UDP Internet Connections (QUIC).

13. The method according to any of claims 1 to 12, wherein said service producer function comprises: a User Plane Function (UPF); a Network Slice Selection Function (NSSF); a Network Exposure Function (NEF); a Network Exposure Function (NRF); a Policy Control Function (PCF); a Unified Data Management (UDM) function; an Application Function (AF); an Authentication Server Function (AUSF); an Access and Mobility Management Function (AMF); a Session Management Function (SMF); a Service Control Point (SCP); or a (Radio) Access Node.

14. The method according to any of claims 1 to 13, wherein said service consumer function comprises: a UPF; an NSSF; a NEF; a NRF; a PCF; a UDM; an AF; an AUSF; an AMF; a SMF; a SCP; or a (R)AN node.

15. A method performed by a network function, the method comprising:

collecting and analysing data relating to transport connections in a network comprising a plurality of service consumer functions and a plurality of service producer functions;

maintaining rules relating to said transport connections based on said data; and

5 providing said rules to at least one of a service consumer function, a service producer function, and a balancer function for managing at least one transport connection.

16. The method according to claim 15, wherein said data relating to transport connections comprises at least one of: data relating to a state of a subscriber / user equipment, data relating to mobility characteristics of a subscriber / user equipment, data
10 relating to communication characteristics of a subscriber / user equipment, data relating to a delay tolerance associated with a subscriber / user equipment; and data relating to a current load or a remaining capacity associated with at least one transport connection and/or a network function.

17. The method according to claim 15 or 16, wherein said rules specify at least one of:

15 whether to store subscriber/session data relating to a particular transport connection at a service producer function associated with that transport connection or to store said subscriber/session data at an external storage different to said service producer function associated with that transport connection; and

whether to change a number of transport connections employed between a service
20 producer function and a service consumer function in order to meet a delay budget associated with subscribers/devices served by that service producer function.

18. A network function comprising:

means for providing at least one transport connection, between a service consumer function and a respective service producer function, for communicating data over said
25 transport connection; and

means for determining whether to store, at an external storage, subscriber/session data relating to said at least one transport connection at said service producer function based on information associated with a subscriber / user equipment to which said data relates.

19. A network function comprising:

means for obtaining information identifying whether a current number of available transport connections between a service producer function and at least one service consumer function meets a delay budget associated with subscribers/devices served by the service producer function; and

means for determining, based on said obtained information, whether to add or remove at least one transport connection between the service producer function and at least one service consumer function.

20. A network function comprising:

10 means for collecting and analysing data relating to transport connections in a network comprising a plurality of service consumer functions and a plurality of service producer functions;

means for maintaining rules relating to said transport connections based on said data; and

15 means for providing said rules to at least one of a service consumer function, a service producer function, and a balancer function for managing at least one transport connection.

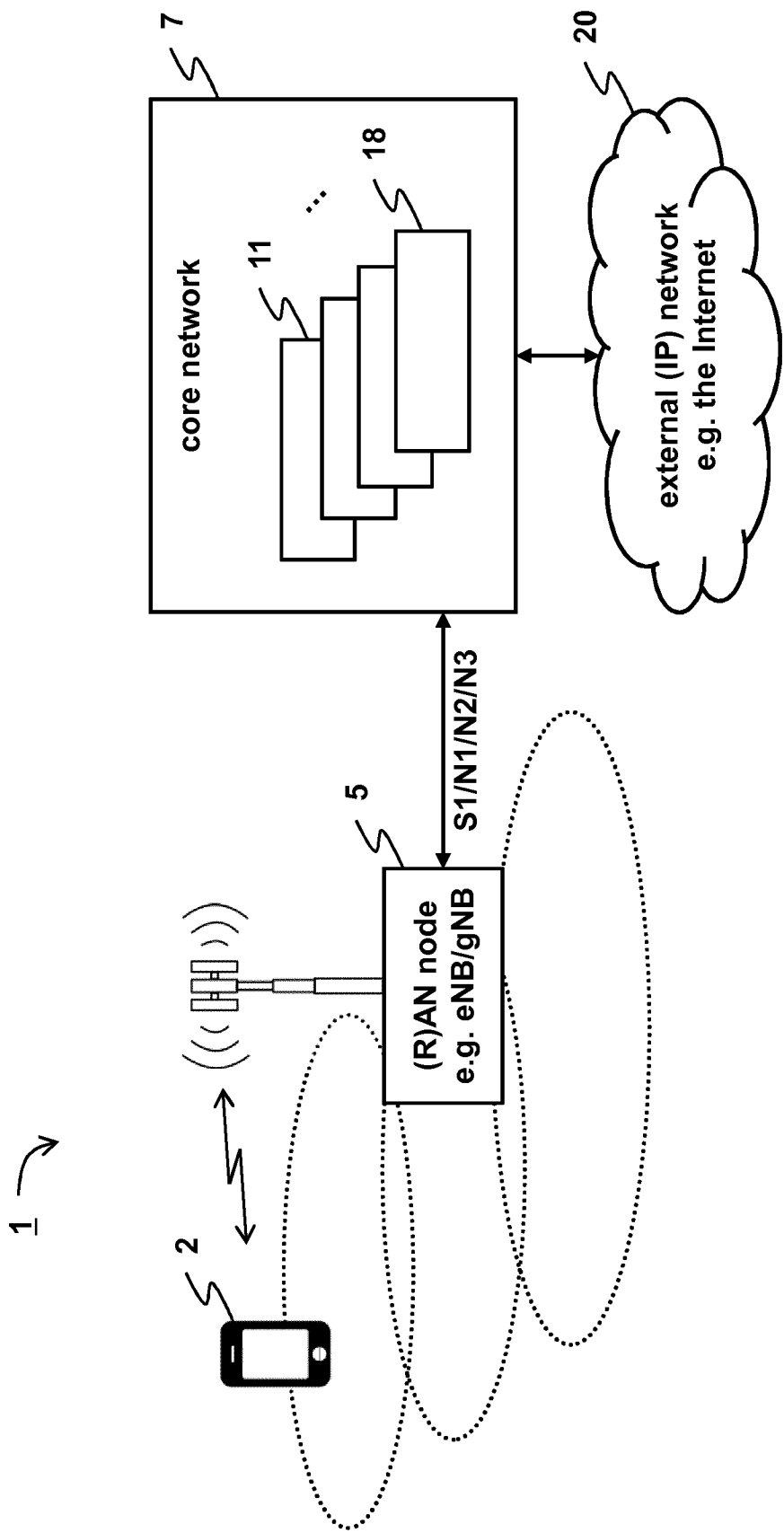
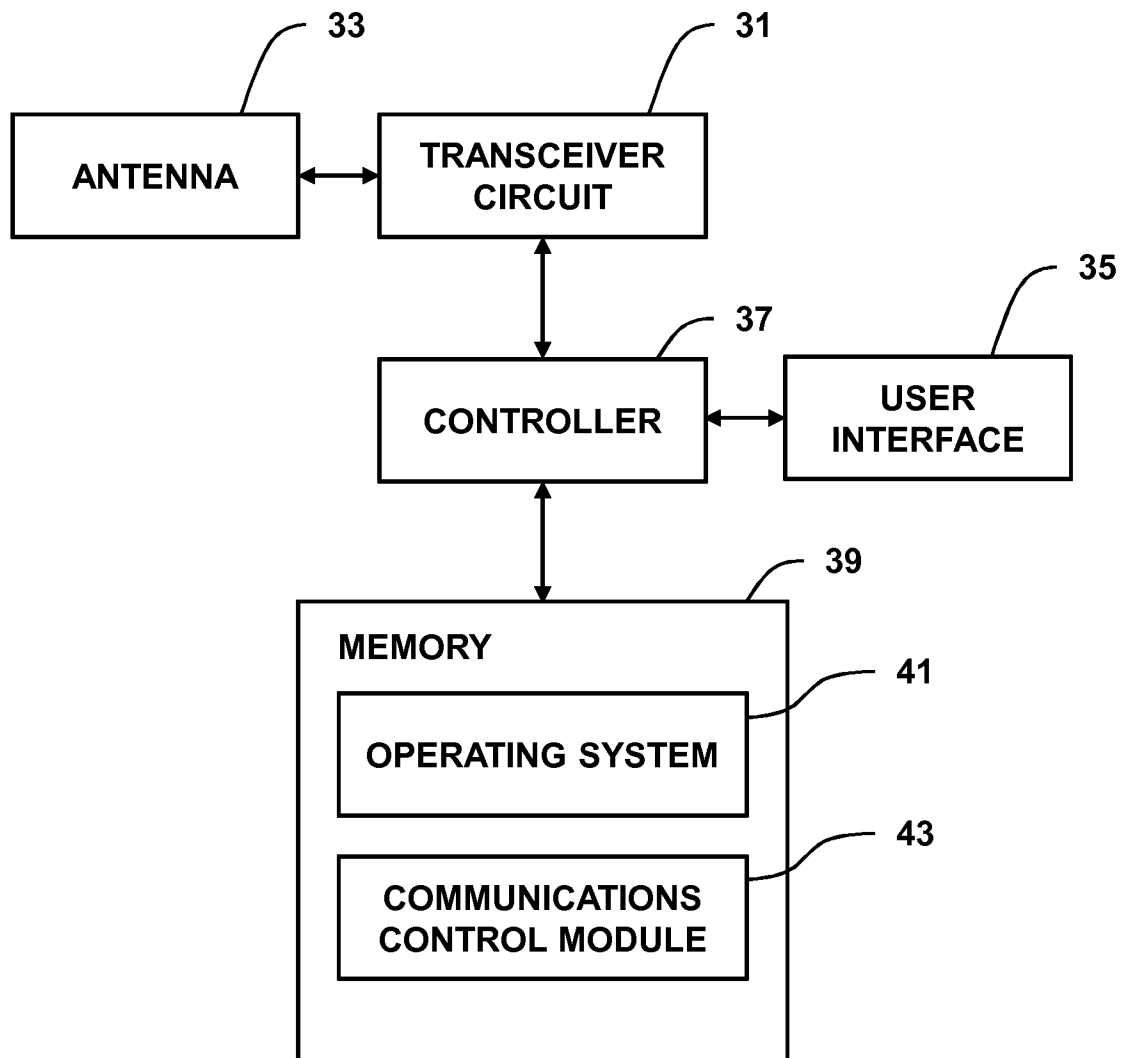
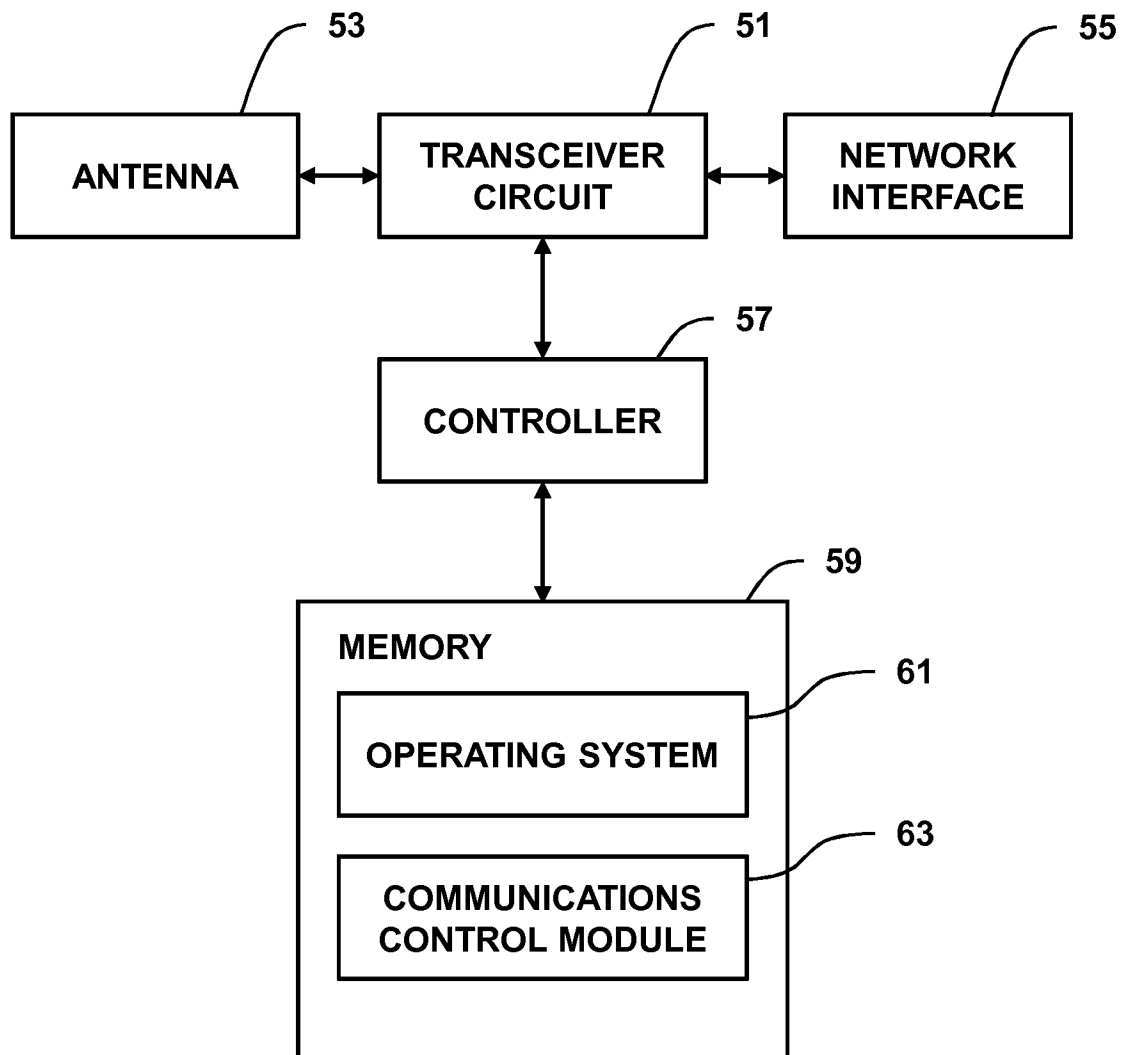


Figure 1

2/13

**Figure 2**

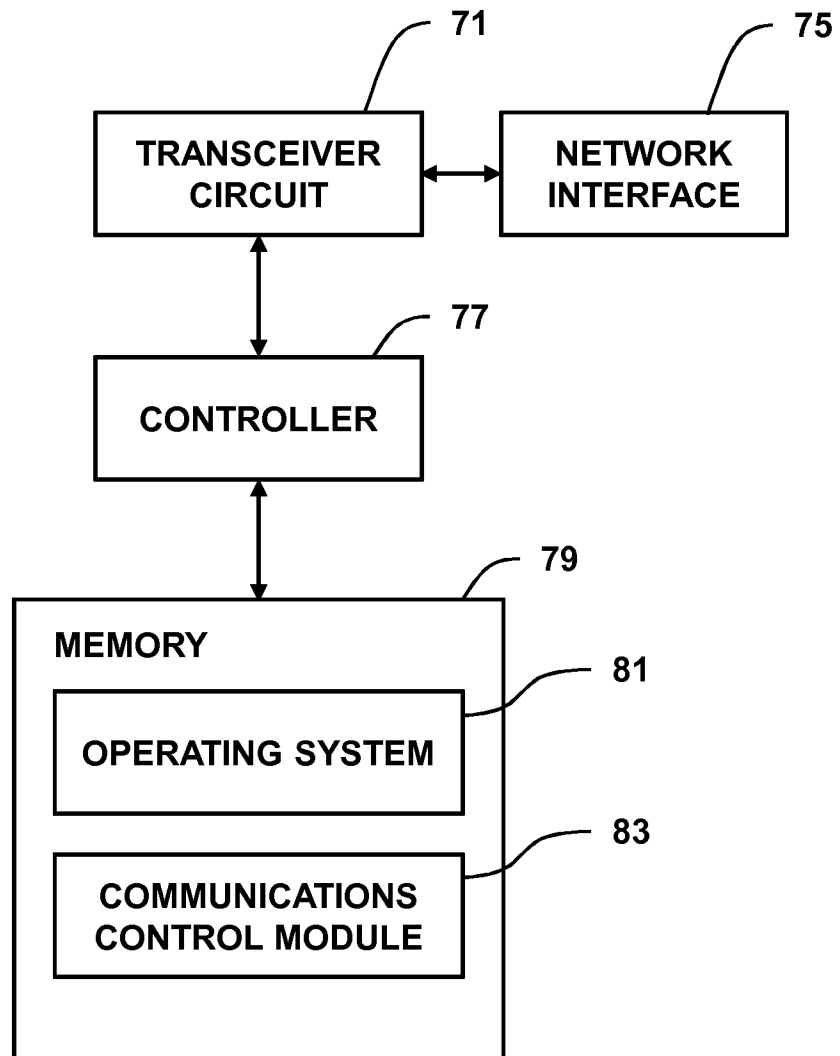
3/13



5

Figure 3

4/13



11-18

Figure 4

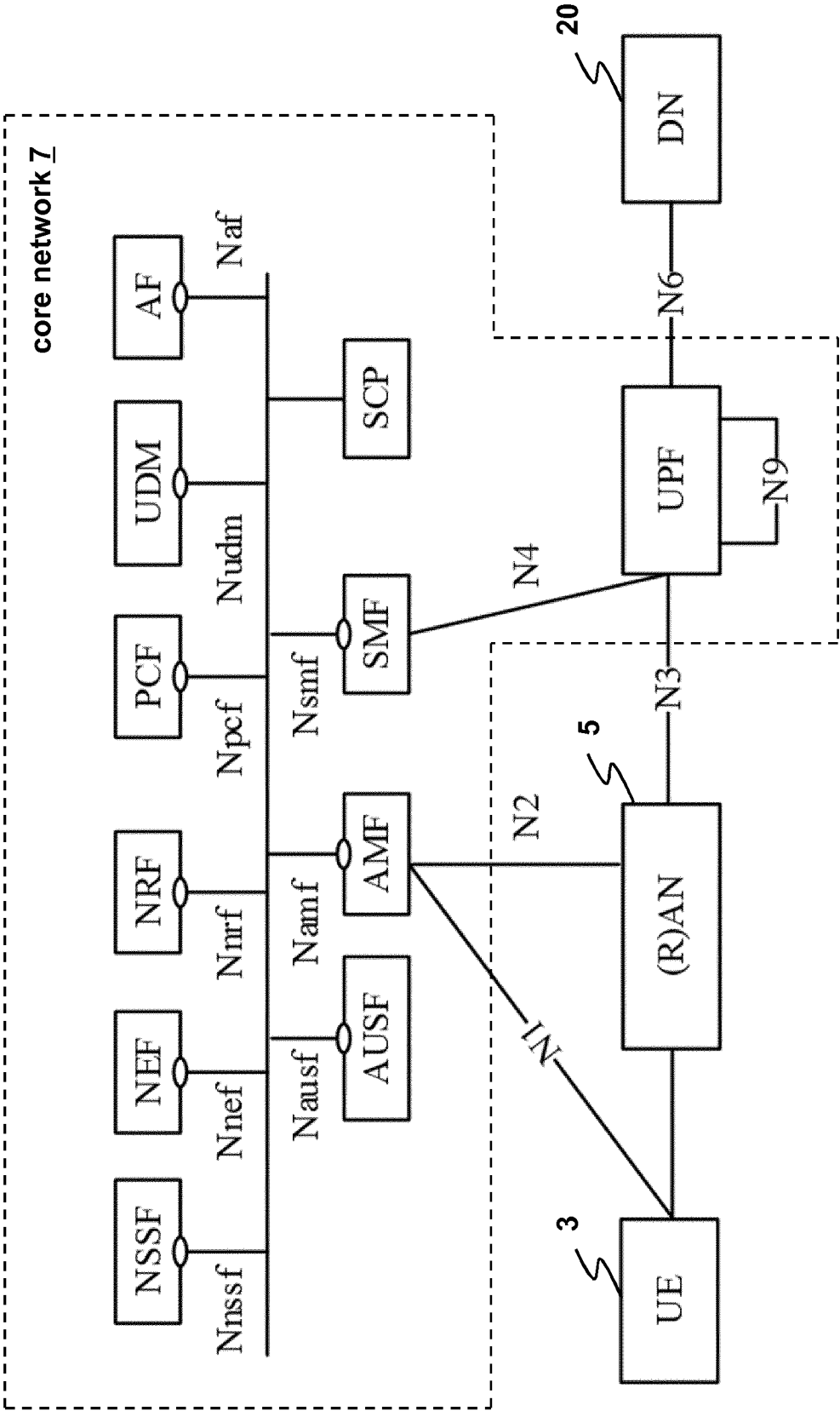


Figure 5

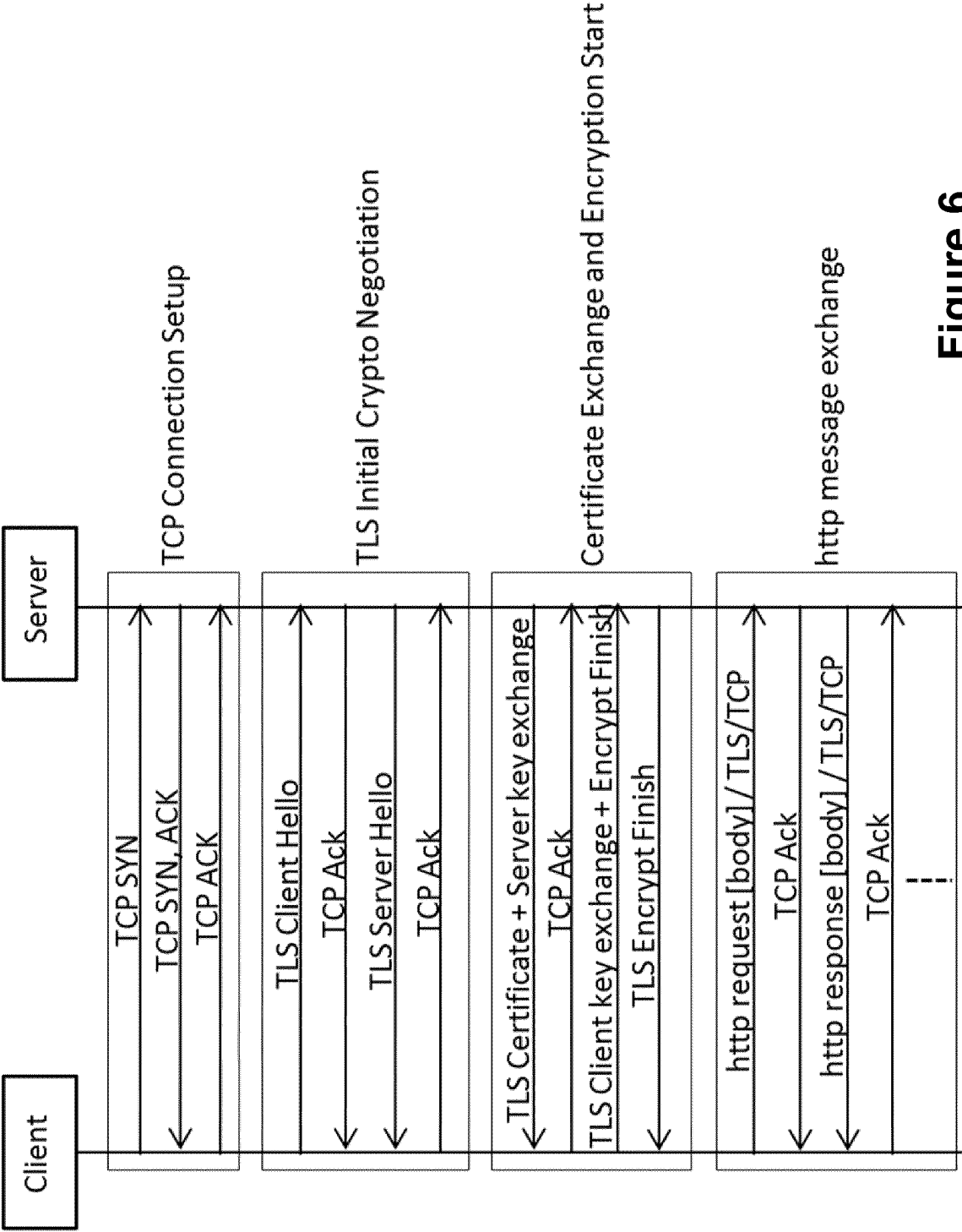


Figure 6

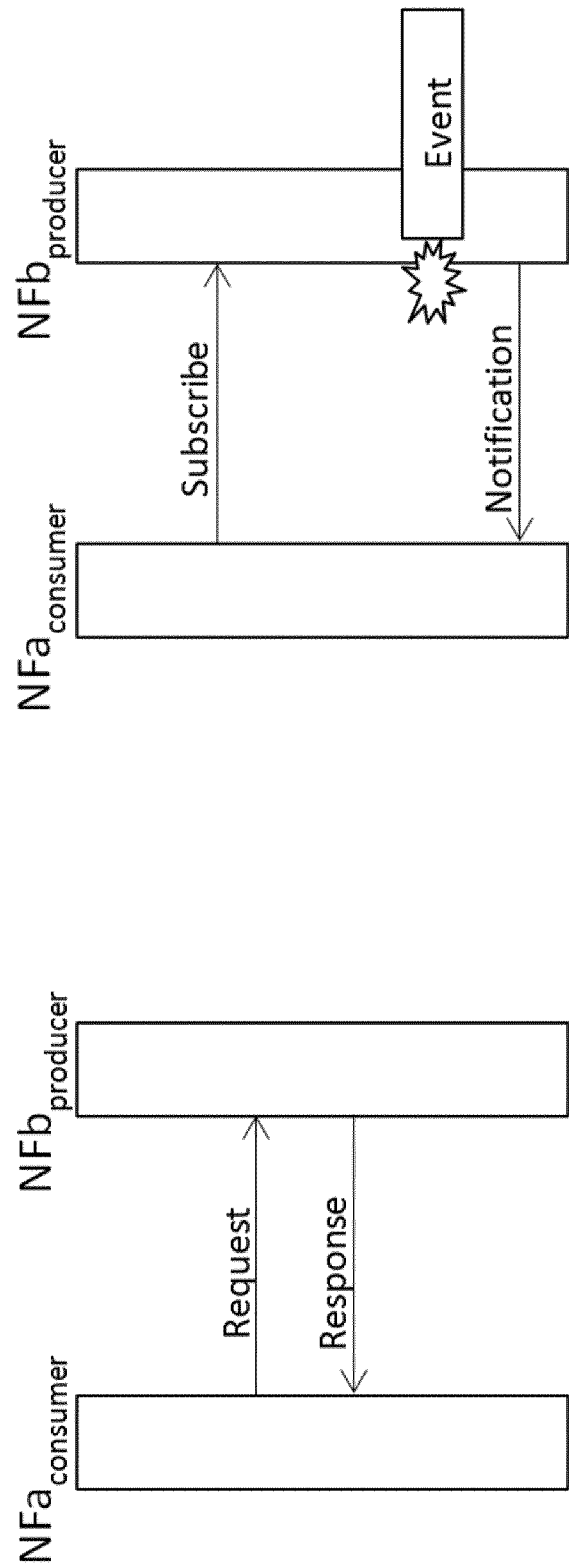
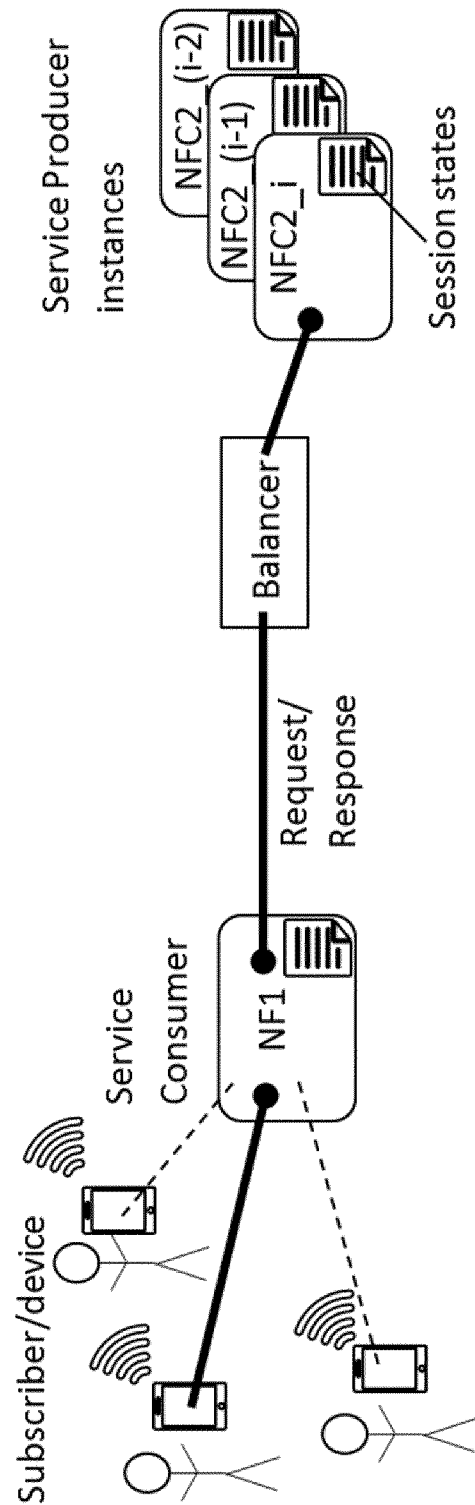
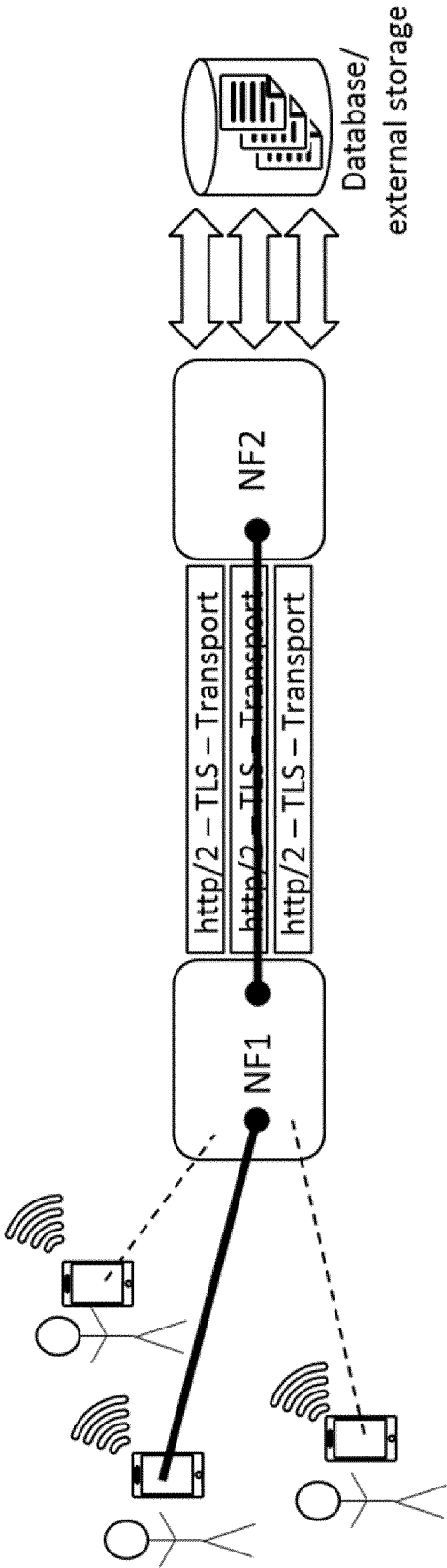


Figure 7



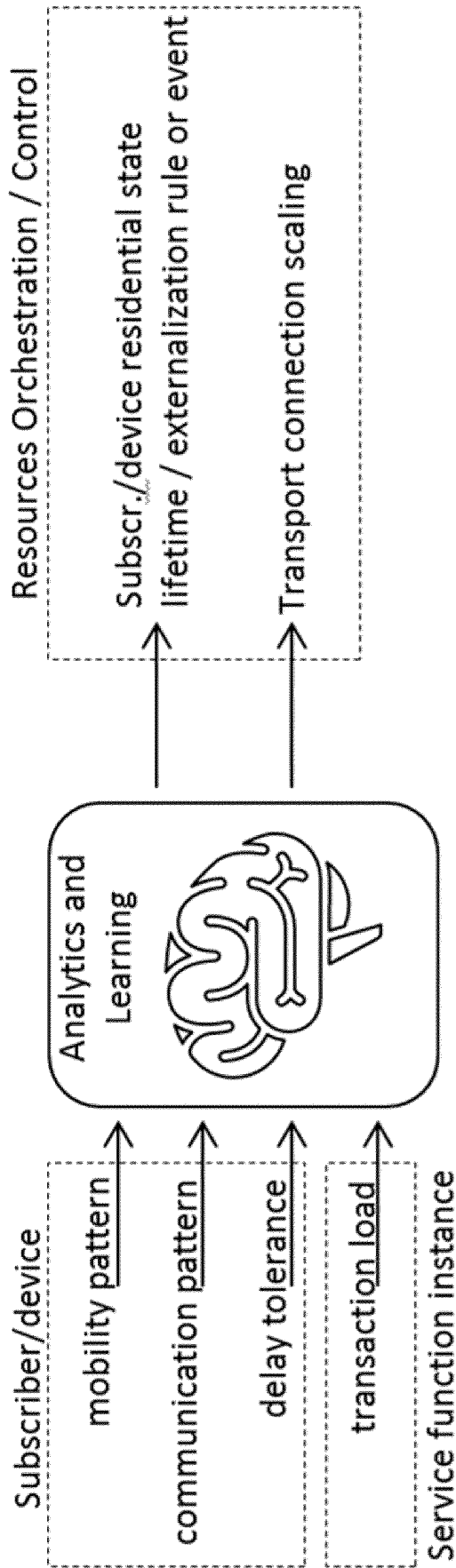
balancing transaction traffic between multiple stateful service producer instances

Figure 8



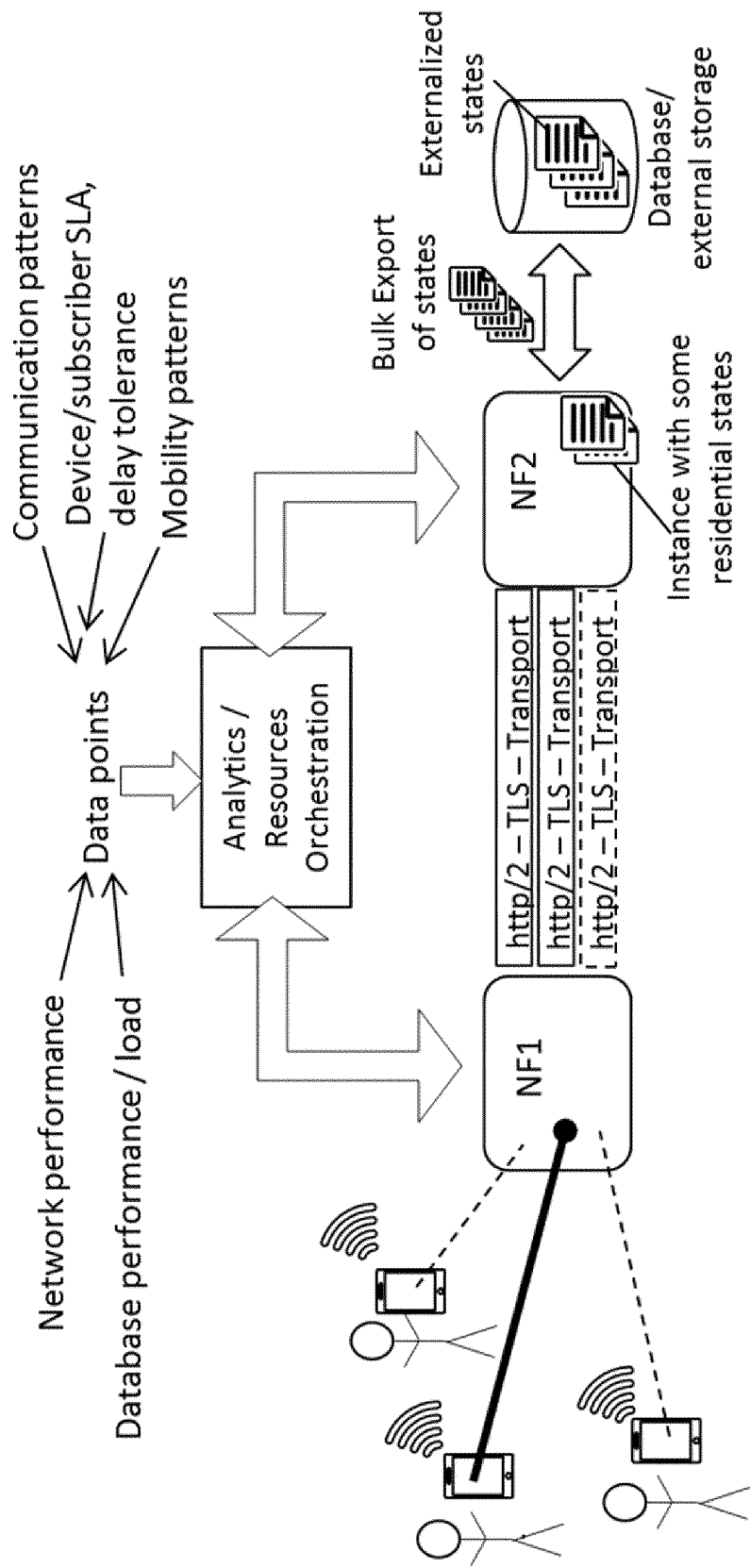
balancing transaction traffic between multiple transport connections to stateless service producer instances

Figure 9



Data analytics of subscriber/device personalized data and statistics as well as service function load (input) to optimize the balance between performance and used resources (output)

Figure 10



context-based balancing of subscriber/device states, which are hold at the service/network function instance as residential states or externalized, and scaling of transport connections

Figure 11

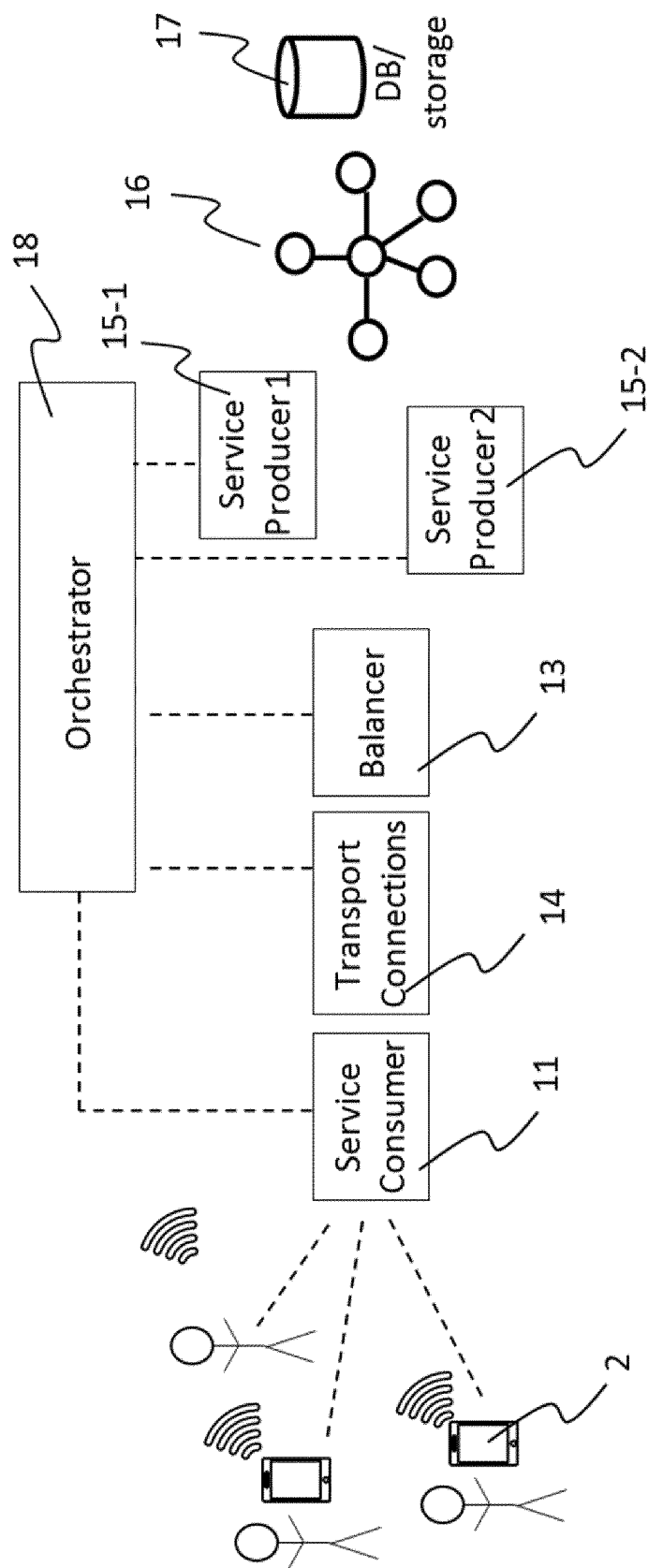


Figure 12

13/13

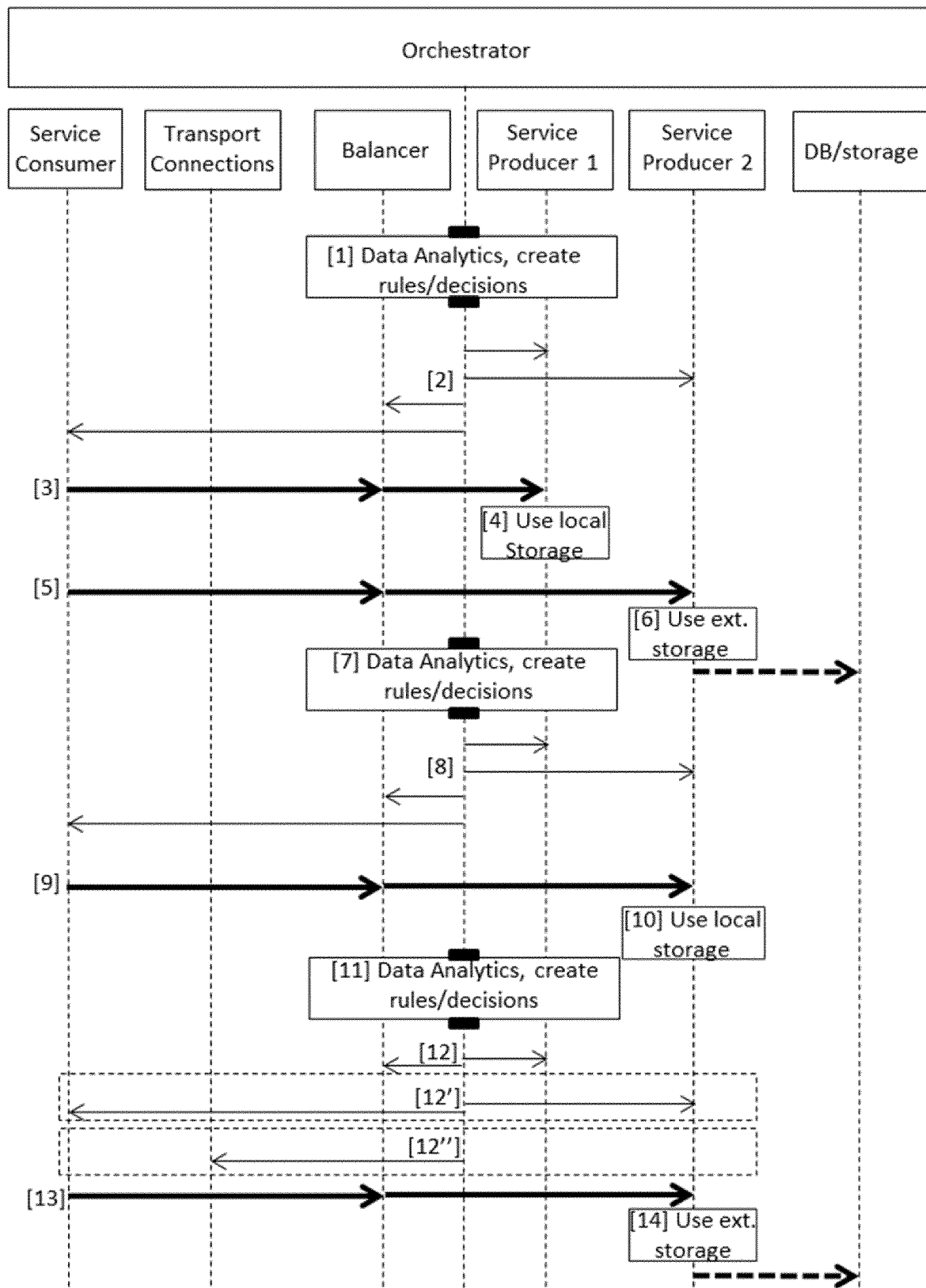


Figure 13

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2019/061992

A. CLASSIFICATION OF SUBJECT MATTER
INV. H04L12/24
ADD.

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)
H04L G06F

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

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

EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2011/202658 A1 (OKUNO MICHITAKA [JP] ET AL) 18 August 2011 (2011-08-18) abstract paragraph [0008] paragraph [0019] paragraph [0091] - paragraph [0107]; figures 4-9, 13,14 -----	1-7,18
Y	US 2015/213075 A1 (COLLINGE JAMES [US] ET AL) 30 July 2015 (2015-07-30) abstract paragraph [0001] paragraph [0011] - paragraph [0012]; figure 1A -----	1-7,18



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents :

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

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

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of 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

26 June 2019

Date of mailing of the international search report

02/09/2019

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Von Der Straten, G

INTERNATIONAL SEARCH REPORT

International application No.
PCT/EP2019/061992

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:

see additional 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 an 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.:

1-7, 18

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.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. claims: 1-7, 18

Method and network function for externally storing
subscriber/session data of a transport connection

2. claims: 8-14, 19

Method and network function for determining based on a delay
budget of at least one transport connection

3. claims: 15-17, 20

Method and network function for providing rules to network
functions for managing at least one transport connection

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2019/061992

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2011202658	A1	18-08-2011	EP 2362606 A1 31-08-2011
			JP 5506444 B2 28-05-2014
			JP 2011170591 A 01-09-2011
			US 2011202658 A1 18-08-2011

US 2015213075	A1	30-07-2015	BR 112015002319 A2 04-07-2017
			CN 104509059 A 08-04-2015
			EP 2893670 A1 15-07-2015
			JP 2015530021 A 08-10-2015
			KR 20150054758 A 20-05-2015
			TW 201424315 A 16-06-2014
			US 2015213075 A1 30-07-2015
			WO 2014039057 A1 13-03-2014
