(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
22 August 2024 (22.08.2024)

WIPO | PCT

(10) International Publication Number
WO 2024/171202 A1

(51) International Patent Classification:
H04L 67/01 (2022.01) *H04L 67/14* (2022.01)

(21) International Application Number:
PCT/IN2023/050154

(22) International Filing Date:
17 February 2023 (17.02.2023)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **TELEFONAKTIEBOLAGET LM ERICSSON (PUBL)** [SE/SE]; SE-164 83, Stockholm (SE).

(72) Inventor; and
(71) Applicant *(for SC only)*: **ROY, Arup Kumar** [IN/IN]; Flat 2A, Oasis Nature, 266 Hossainpur Road, KOLKATA 700107 (IN).

(72) Inventors: **PEREPU, Satheesh Kumar**; Q 201, Oxygen by Urban Tree, Pushkin Street, Nookampalyam Road, Chennai 600131 (IN). **M, Saravanan**; 44/77 3rd Main Road, Venkateswara Nagar, Velachery, CHENNAI 600042 (IN). **DUTTA, Sudipta**; AB/28 Prafulla Kanan, PO -Prafulla Kanan, Kolkata 700101 (IN).

(74) Agent: **SINGH, Manisha** et al.; LEXORBIS, 709/710, Tolstoy House 15-17, Tolstoy Marg, New Delhi 110 001 (IN).

(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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, 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, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, 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, ME, 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).

Declarations under Rule 4.17:
— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

Published:
— *with international search report (Art. 21(3))*
— *in black and white; the international application as filed contained color or greyscale and is available for download from PATENTSCOPE*

(54) **Title:** METHODS AND NODES IN A COMMUNICATIONS NETWORK

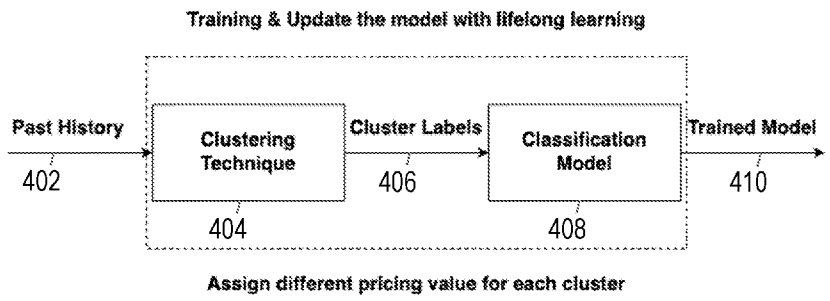

Fig. 4a

(57) **Abstract:** A computer implemented method performed by a first node in a communications network. The method comprises clustering (202) session data according to network usage pattern, to obtain a plurality of clusters; associating (204) each cluster in the plurality of clusters with an identifier; and training (206) a model using a machine learning process, to take session data of a session as input and output an identifier associated with the cluster to which the session data belongs.

WO 2024/171202 A1

METHODS AND NODES IN A COMMUNICATIONS NETWORK

Technical Field

This disclosure relates to methods, nodes and systems in a communications network. More particularly but non-exclusively, the disclosure relates to identifying traffic types in a communications network.

Background

In fifth generation (5G) communications networks, products and services have become more complex, as Communications Service Providers (CSPs) are gearing up to address varied enterprise models, such as ultra-reliable low-latency communication (URLLC), enhanced mobile broadband (eMBB), and massive machine type communication (mMTC).

To monetize next generation services for consumers and across industry verticals to serve enterprises, Business Support Systems (BSS) need to be more agile, flexible, robust and resilient in order to effectively be able to meter and charge appropriately for the innovative business models which are emerging with the advent of 5G.

5G has heralded new ways of accessing Radio Access Networks (RANs). One such model is the "Radio Access Network (RAN) Wholesaler" model, where a CSP offers 5G RAN as a service towards its partner Mobile Network Operators (MNOs) / Mobile Virtual Network Operators (MVNOs). In such configurations, the MNOs and MVNOs have their own 5G Core Networks. By taking RAN as a service from the "RAN Wholesaler" CSP, the MNOs and MVNOs can launch different services for their end customers.

In such configurations, the "RAN Wholesaler" CSP may act more as service enabler for the Business to Business to End-customer services, known as "B2B2X" services, launched by partner MNO/MVNOs. This setup is also known as a Multiple Operator Core Network (MOCN) setup.

Summary

RAN Wholesaler CSPs are looking to segregate their shared RAN resources into different RAN partitions and offer each MNO/ MVNO partners different dedicated RAN partitions. Partitioning the RAN ensures that each partner MNO/MVNOs gets the RAN resources dedicated and committed to them to carry their traffic.

Over each of the RAN partitions, RAN wholesalers are also looking to offer differentiated Quality of service (QoS) based RAN traffic flows towards the MNO/ MVNO partners. These QoS differentiated traffic flows are segregated by configuring scheduling and prioritization configurations applied at the baseband side, to treat the radio traffic packets at

the expected priority treatment as per the overlay service class and priority (e.g., Gold, Silver and Bronze etc.)

These QoS differentiated RAN traffic flows from the RAN Wholesaler side need to be stitched back at the MNO/MVNO Core side appropriately over specific Public Land Mobile Network (PLMN) IDs and Single – Network Slice Selection Assistance Information (SNSSAI) based end-to-end (e2e) network slices that can carry the overlay service user plane traffic to deliver the QoS promises (e.g., bandwidth, latency, reliability, and the number of supported end-users) of a specific slice type, i.e., gold-, silver-, and bronze-based slices.

There is a desire amongst CSPs to be able to identify traffic slices having different quality of traffic flow thereon. There are a range of applications for such information, one of which is the ability to charge different amounts for different QoS differentiated traffic flows. However current technology limits distinguishing such traffic flows and generating metered records which can be used by the BSS Charging/Rating pipeline. This is because in the MOCN set-up, the RAN wholesaler generally does not own the 5G core (which is provided by the MNO/MNVO partners) which is usually responsible for charging and billing. Consequently, the CSPs don't have the standard Call Detail Record (CDR) source for processing and aggregating subscriber usage on specific slices, as the CDRs are generated by the Charging Function (CHF) of the 5G-Core. As such, the CSPs have access to very limited information on session data.

Thus, according to a first aspect herein there is a computer implemented method performed by a first node in a communications network. The method comprises clustering session data according to network usage pattern, to obtain a plurality of clusters; associating each cluster in the plurality of clusters with an identifier; and training a model using a machine learning process, to take session data of a session as input and output an identifier associated with the cluster to which the session data belongs.

In some embodiments, the method may further comprise associating each identifier with a traffic type, each traffic type having a different quality of service, QoS profile.

According to a second aspect there is a first node in a communications network, the first node comprising: a memory comprising instruction data representing a set of instructions; and a processor configured to communicate with the memory and to execute the set of instructions. The set of instructions, when executed by the processor, cause the processor to: cluster session data according to network usage pattern, to obtain a plurality of clusters; associate each cluster in the plurality of clusters with an identifier; and train a model

using a machine learning process, to take session data of a session as input and output an identifier associated with the cluster to which the session data belongs.

In a third aspect there is a first node in a communications network. The first node is configured to: cluster session data according to network usage pattern, to obtain a plurality of clusters; associate each cluster in the plurality of clusters with an identifier; and train a model using a machine learning process, to take session data of a session as input and output an identifier associated with the cluster to which the session data belongs.

In a fourth aspect there is a computer program comprising instructions which, when executed on at least one processor, cause the at least one processor to carry out the method of the first aspect.

In a fifth aspect there is a carrier containing a computer program according to the fourth aspect, wherein the carrier comprises one of an electronic signal, optical signal, radio signal or computer readable storage medium.

In a sixth aspect there is a computer program product comprising non transitory computer readable media having stored thereon a computer program according to the fourth aspect.

Thus, in this manner, when implemented by a CSP, a CSP can determine, in an automated manner, a QoS profile for a given traffic flow from its network usage pattern, enabling a quick classification to be performed. The use of clustering and unsupervised learning allows the method to be adaptable to the ever-changing conditions and requirements of communications networks. When implemented in a MOCN setup, this provides a technical manner in which a CSP of a RAN wholesaler can determine QoS being provided to MOCN partners on each of its differentiated network slices, despite the CSP not having access to the core network data of the MNO/MNVO partners.

Brief Description of the Drawings

For a better understanding and to show more clearly how embodiments herein may be carried into effect, reference will now be made, by way of example only, to the accompanying drawings, in which:

Fig. 1 shows an example first node according to some embodiments herein;

Fig. 2 shows an example method in a first node according to some embodiments herein;

Fig. 3 shows example clusters according to some embodiments herein;

Fig. 4a shows an example block diagram of a method of training a machine learning model according to some embodiments herein;

Fig. 4b shows an example block diagram of a method of using a machine learning model according to some embodiments herein;

Fig. 5 shows clusters obtained during testing of an example method herein; and

Fig. 6 shows an example signal diagram according to an embodiment herein.

Detailed Description

The disclosure herein relates to a communications network (or telecommunications network). A communications network may comprise any one, or any combination of: a wired link (e.g. ASDL) or a wireless link such as Global System for Mobile Communications (GSM), Wideband Code Division Multiple Access (WCDMA), Long Term Evolution (LTE), New Radio (NR), WiFi, Bluetooth or future wireless technologies. The skilled person will appreciate that these are merely examples and that the communications network may comprise other types of links. A wireless network may be configured to operate according to specific standards or other types of predefined rules or procedures. Thus, particular embodiments of the wireless network may implement communication standards, such as Global System for Mobile Communications (GSM), Universal Mobile Telecommunications System (UMTS), Long Term Evolution (LTE), and/or other suitable 2G, 3G, 4G, or 5G standards; wireless local area network (WLAN) standards, such as the IEEE 802.11 standards; and/or any other appropriate wireless communication standard, such as the Worldwide Interoperability for Microwave Access (WiMax), Bluetooth, Z-Wave and/or ZigBee standards.

Fig 1 illustrates a network node 100 in a communications network according to some embodiments herein. Generally, the node 100 may comprise any component or network function (e.g. any hardware or software module) in the communications network suitable for performing the functions described herein. For example, a node may comprise equipment capable, configured, arranged and/or operable to communicate directly or indirectly with a UE (such as a wireless device) and/or with other network nodes or equipment in the communications network to enable and/or provide wireless or wired access to the UE and/or to perform other functions (e.g., administration) in the communications network. Examples of nodes include, but are not limited to, access points (APs) (e.g., radio access points), base stations (BSs) (e.g., radio base stations, Node Bs, evolved Node Bs (eNBs) and NR NodeBs (gNBs)).

The first node may be operated (e.g. owned and maintained) by a Radio Access Network (RAN) Wholesaler as part of a Multiple Operator Core Network (MOCN) setup. In a MOCN set up, sessions may be operated by different Mobile Network Operators, MNOs, accessing services through the RAN wholesaler. In such embodiments, the first node may be separated from the 5G core network associated with said sessions. As such, the first node may not have access to data usually collected by the core for processes such as billing.

The first node may be the node that will perform a QoS based classification of aggregate session data obtained at the edge of the RAN (for example, a cell site router). RAN nodes may not be considered as a source of CDRs as specified by 3GPP spec TS32.240. The charging data function (CDF) (that receives charging events from the Charging Trigger Function and uses the information to construct CDRs events) is hosted at the core side.

The node 100 is configured (e.g. adapted, operative, or programmed) to perform any of the embodiments of the method 200 as described below. It will be appreciated that the node 100 may comprise one or more virtual machines running different software and/or processes. The node 100 may therefore comprise one or more servers, switches and/or storage devices and/or may comprise cloud computing infrastructure or infrastructure configured to perform in a distributed manner, that runs the software and/or processes.

The node 100 may comprise a processor (e.g. processing circuitry or logic) 102. The processor 102 may control the operation of the node 100 in the manner described herein. The processor 102 can comprise one or more processors, processing units, multi-core processors or modules that are configured or programmed to control the node 100 in the manner described herein. In particular implementations, the processor 102 can comprise a plurality of software and/or hardware modules that are each configured to perform, or are for performing, individual or multiple steps of the functionality of the node 100 as described herein.

The node 100 may comprise a memory 104. In some embodiments, the memory 104 of the node 100 can be configured to store program code or instructions 106 that can be executed by the processor 102 of the node 100 to perform the functionality described herein. Alternatively or in addition, the memory 104 of the node 100, can be configured to store any requests, resources, information, data, signals, or similar that are described herein. The processor 102 of the node 100 may be configured to control the memory 104 of the node 100 to store any requests, resources, information, data, signals, or similar that are described herein.

It will be appreciated that the node 100 may comprise other components in addition or alternatively to those indicated in Fig. 1. For example, in some embodiments, the

node 100 may comprise a communications interface. The communications interface may be for use in communicating with other nodes in the communications network, (e.g. such as other physical or virtual nodes). For example, the communications interface may be configured to transmit to and/or receive from other nodes or network functions requests, resources, information, data, signals, or similar. The processor 102 of node 100 may be configured to control such a communications interface to transmit to and/or receive from other nodes or network functions requests, resources, information, data, signals, or similar.

Briefly, in one embodiment, the node 100 may be configured to cluster session data according to network usage pattern, to obtain a plurality of clusters, associate each cluster in the plurality of clusters with an identifier, and train a model using a machine learning process, to take session data of a session as input and output an identifier associated with the cluster to which the session data belongs.

Clustering and training a model on the clustered data in this manner is a type of self-supervised machine learning. The skilled person will be familiar with machine learning and self-supervised machine learning, which can be used to detect patterns or clusters in data (that may not be readily apparent to a human-observer). In embodiments herein, as will be described in more detail below, self-supervised machine learning is used to group sessions according to network usage pattern and train a model to classify session data according to the determined groups. The groups as a whole can then be labelled e.g. according to a communication type or service quality (e.g. URLLC, MMTC etc). Thus, the node 100 can be used to identify session type in a semi-automated manner, particularly in MOCN scenarios where core data to this effect may be unavailable to the first node.

Fig. 2 which shows a computer implemented method 200 performed by a first node in a communications network, according to some embodiments herein. The method 200 may be performed by a node such as the node 100 described above.

In a first step 202 the method 200 comprises clustering session data according to network usage pattern, to obtain a plurality of clusters.

In some embodiments, step 202 comprises requesting and/or obtaining the session data. The session data may be obtained, for example, from one or more Cell Site Routers (CSRs). RAN CSRs are configured to expose the RAN user plane traffic (uplink and downlink) for each partner MNOs/ MVNOs as separate unique VLAN traffic. In RAN CSR session logs, each line item represents RAN traffic during a specific day and duration for a specific MNO/MVNO partner identified by a unique VLAN identifier.

In some embodiments, the CSR routers supplying the session data are configured to reflect key QoS parameters (in this case the observations), including but not limited to bandwidth and latency experienced, as part of each session item entry. Thus, in some embodiments, the session data is obtained from CSR session logs. As noted above, the session data may comprise QoS information such as bandwidth and latency for each session.

The session data is clustered in step 202 using any known clustering method (k-means, hierarchical clustering etc). As an example, the clustering may be performed as part of a self-supervised machine learning method to learn the clusters dynamically. The idea in such embodiments is to use unlabelled data and use it to arrive at clusters in an automated manner.

In step 204 the method 200 comprises associating 204 each cluster in the plurality of clusters with an identifier, which may otherwise be referred to as a label. The identifier can be, for example, a numerical identifier, such as an enumeration. For example, each cluster can be labelled 1, 2, 3 etc.

In some embodiments, each identifier may be associated with a traffic type, each traffic type having a different QoS profile. For example each identifier may be associated with a traffic type by comparing network usage patterns of sessions in the clusters to reference QoS profiles for said traffic types. The profiles may be retrieved e.g. from a database and/or set by a human engineer. In this manner, the process of associating clusters with particular traffic types may be performed with minimal input from a human engineer.

Generally, the traffic types may correspond to different network slices with different Quality of Service traffic flows. Examples of traffic types include, but are not limited to: enhanced mobile broadband, eMBB, massive machine type communication, mMTC, critical machine type communication, cMTC, and ultra-reliable low-latency communication, URLLC.

In other examples, the clusters can be labelled (e.g. by a human expert) with an identifier corresponding (e.g. directly) to a traffic type, QoS type, response pattern for the customer, or according to any other cluster characteristic. In one example, the clusters may be labelled according to a charging/billing profile to be used for sessions falling in said cluster.

A sample cluster diagram for various different traffic types is illustrated in Fig. 3, which shows, for latency (on the x-axis) and throughput (on the y-axis) different clusters 302;304 associated with different traffic types.

As an example:

• eMBB traffic session observations will exhibit a higher bandwidth and mid level latency (might have variance within eMBB traffic based on configured QoS profile to segregate between eMBB-Normal vs. eMBB-Premium traffic)

• mMTC traffic on the contrary will exhibit low bandwidth and higher latency observations (primarily to support IoT sensor data)

• cMTC traffic sessions on the contrary will be very low on latency observations and exhibit mid level throughput observations

Based on the QoS profile configurations applied to support each of these different QoS flows, theoretical upper and lower limits of probabilistic observations is available from the RAN engineering side for the traffic types. These theoretical probabilistic QoS parameter limits can be used as references for comparing against the QoS clusters observed based on the actual CSR logs collected for a specific day range to arrive at a confidence limit to indicate the accuracy of the observed clusters. In this way, semi-automatic mapping of clusters to QoS profiles may be performed.

In step 206 the method 200 comprises training a model using a machine learning process, to take session data of a session as input and output an identifier associated with the cluster to which the session data belongs.

The model may be, for example, a neural network model. The skilled person will be familiar with neural networks models and machine learning processes (e.g. such as gradient descent and backpropagation) used to train neural networks. There are many Open Source Libraries that may be used to create and train neural networks, such as, for example, Scikit-learn described in the paper: "Scikit-learn: Machine Learning in Python", by Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

The skilled person will appreciate that neural networks are merely an example, however and that other types of classification machine learning model might equally be used, such as, for example, Decision Trees, Random Forests and Support Vector Machines (SVM).

The session data obtained in step 202 may be used as training data for the model. Each session data example in the session data in step 202 may act as an example input for the model. The identifier of the cluster to which the respective session data example was associated with in step 204 may be used as the corresponding ground truth (e.g. "correct") output for said example input. Thus, the training data may be in the form of [session data example; ground truth identifier] input-output pairs.

In one example, the session data input to the model may comprise the features Latency and Throughput and the output of the model may be traffic-type as described above.

The skilled person will appreciate however that these are examples and that other input features may equally be used.

In one example, the model may be trained as follows: let N be the number of clusters initially chosen for clustering. The session data is then divided into these clusters based on their usage pattern and geographical location. Now, the cluster level data is annotated with the label information and the model is trained using this label information. An example loss function that can be used in the training of the model is:

$$J = -\sum_{i=1}^{N} y_i \log(h_\theta(x_i)) + (1 - y_i) \log(1 - h_\theta(x_i))$$

where x_i is the feature space data 'i', y_i is the label for the sample 'i' and $h_\theta(.)$ is the model output for the sample. In this way, the model learns how to optimize the parameter θ such the loss function is minimized.

In summary, the training of the model comprises the following steps: 1) use a clustering method like K-means, hierarchial clustering etc. to cluster the data. 2) Use the learned clusters and label the data using the cluster names like 1, 2, etc. 3) A model is trained using a loss function, such as that described above. In this way, self-supervised learning machine learning is used to classify different sessions according to network usage pattern in order to classify sessions according to a traffic type associated with a particular QoS profile.

In use, the trained model may be used to classify a first new session by inputting the session data (e.g. for example, latency and throughput) into the trained model and receiving a label as output.

The label may be used, for example, in a billing process. E.g. a billing process may use different charging rates for sessions of different traffic types. As another example, the traffic type data may be used in a statistical manner to understand the sessions operating on the communications network, for the purpose of maintenance or planning. As another example, the observed QoS parameters in session data may be forwarded to a Data Analytics application to find out patterns of quality degradations/ deviations of specific QoS flow over time and according take corrective actions in RAN configurations.

Turning now to other embodiments, the validity of the trained model may be assessed and the model may be updated over time. Trained models may become less accurate

or fail over time for various reason, such as, because of new patterns emerging in the data. What is proposed herein to mitigate such effects is the following:

For every new sample, the following steps are performed:

1. For every sample, predict a confidence score for each predicted label (e.g. instead of softmax output).
2. If the confidence score is greater than a threshold α, use the predicted label and assign the sample to the cluster.
3. If the confidence score is less than a threshold α, assign this new point to a new cluster.
4. The model is updated/retrained when number of new samples reaches some number N.

Put another way, the method 200 may further comprise i) receiving first session data for the first new session. In this sense, the first session data is new data (e.g. that may not have been previously used to train the model). The method may then comprise ii) providing the first session data as input to the model and obtaining as output a first identifier (e.g. a first label) predicted for the first new session. E.g. the first session data is input to the model and a corresponding identifier is predicted for the first session data. The method 200 may then comprise iii) determining a first confidence score reflecting a confidence with which the first identifier is predicted for the first session data. The first confidence score may be a confidence output by the model in its prediction of the first identifier. In a fourth step, iv) the method 200 may further comprise labelling the first session data with the first identifier if the first confidence score satisfies a first criteria. The first criteria may be, for example, a confidence threshold for which a predicted identifier is used if predicted with a confidence above the confidence threshold.

If the first confidence score does not satisfy the first criteria, then the first session data may be labelled with a first new identifier (e.g. indicating that the first session data should belong to a different, new cluster).

This process (e.g. steps i)-iv) above) may be repeated for a plurality of new sessions and if the number of the new sessions associated with the first new identifier satisfies a second criteria (which may be e.g. a threshold number) then this may trigger re-training of the model. In this way, if e.g. more than n of the new sessions cannot be labelled with an identifier with sufficient confidence, then the method 200 may be repeated (e.g. on, or taking

into account of, newer session data), since this may be an indication that the clusters obtained in step 204 are out of date.

In some embodiments, where the model is a neural network, the model may be updated using a plastic neural network approach to update the self-supervised model at every interval. Plastic neural networks are described in the paper by Miconi, Thomas, Kenneth Stanley, and Jeff Clune. "Differentiable plasticity: training plastic neural networks with backpropagation." In International Conference on Machine Learning, pp. 3559-3568. PMLR, 2018. Plastic neural networks aim to autonomously design and create learning systems and also introduce lifelong learning to the existing systems by bootstrapping learning from scratch, recovering performance in unseen conditions, testing the computational advantages of neural components, and deriving hypotheses on the emergence of biological learning. Plastic Neural networks cannot be directly used for clustering as they can be used for classification problems because of the nature of the loss function. Hence, herein in this disclosure, the new type of loss function described above is introduced to enable the use of plastic neural networks for life long learning of clusters.

Herein, plasticity may be applied by randomly updating some weights of the neural network by adding noise to them. Adding of noise to existing weights can randomize the model scores so that the model can learn new things instead of being overly-specific.

$$\theta_{new} = \theta + U$$

where θ is the initial trained model, U is the noise weights need to add to the model and θ_{new} is the updated trained model. New samples (e.g. new session data) can also be used in the training with the new weights so that the model can perform robustly for new samples.

Thus, the method 200 above may comprise perturbing values of one or more weights or biases in the model (e.g. with noise) to obtain a perturbed model and performing (further) training on the perturbed model. As noted above, the advantage of this is to disturb the model so that new data can be more readily be taken into account.

Turning now to an example embodiment, as described above, RAN wholesalers typically enable multiple RAN slices with QoS differentiated traffic flows to support a wide variety of services demanding, for example, eMBB, mMTC and cMTC type user plane traffic. As such, the session logs from cell site routers at the edge of the RAN is expected to reflect session records that have a wide variance of observations along the key QoS dimensions like bandwidth, latency etc. This was illustrated in Fig. 3 described above.

The method 200 described above, can be used to determine the predicted volume of data against each of the observed QoS clusters and generate metered records against individual traffic flows. The generated metered records can then be used by a Business Support System (BSS) charging/rating pipeline to apply differentiated rates during B2B charging/rating.

In one example, the following steps may be performed:

1) First collect RAN CSR's and extract features for every unique UE id in a certain time like, for example, one hour. Example features are throughput, and latency, but other features could alternatively or additionally be used.

2) Then, use the method 200 described above to train a model to output a cluster identifier for a given input set of feature values.

3) For each cluster, decide upon a pricing based on (e.g. average) throughput and latency values associated with the respective cluster.

4) For every unique UE id in future for a given time, the trained model may be used to obtain an identifier (e.g. label) for the session and a confidence score output by the model for said predicted identifier. If the confidence score is greater than a threshold confidence, α, the new sample is assigned to the cluster (e.g. corresponding to the predicted identifier) and the UE id is charged according to the rate that that cluster has.

4a) If the confidence score is less than threshold, the new sample is assigned to a new cluster.

4b) If the number of new cluster samples reaches N, a pricing rate is assigned to the new cluster, or else, the new point is charged at an average pricing (or default pricing) until the pricing is decided.

5) For every time interval i.e. 1 day or such, the model is randomized by adding random values to the weights (e.g. according to the plasticity principles described above with respect to the method 200).

6) Steps 4, 4a and 4b are repeated for every execution of step 5.

7) If the model performance is poor e.g. if more number of new samples goes to unassigned clusters, the model is updated with the new data by performing steps 1 to 3.

In some cases, existing clusters may be removed (deleted) based on new patterns in the data. For this, when predicting labels for the new samples, in the event that the predicted label is not present in true set of labels for successive samples (like 100 or some threshold number), then the unused label may be removed from the data and the model may be retrained

using self-supervised learning with lifelong learning. In this way, the clusters can be added/deleted in real-time to handle new patterns in the data.

A block diagram of an embodiment of the method 200 is illustrated in Fig. 4. Fig. 4a illustrates the training of a model using the method 200 described above. In step 402 (previous) session data is obtained and this is clustered in 404 (e.g. step 202 is performed). The cluster labels are assigned to each of the lines of session data 406; 204 and the session data and respective labels are used as training data to train a classification model (according to step 206).

Fig. 4b illustrates an embodiment of the use of the model output from the process shown in Fig. 4a. In inference, new session data 412 is input to the model 414 and a predicted identifier (or label) 416 is output. If a confidence score associated with the label is greater than a threshold confidence, then the predicted label is used 420 as the label for the new session data. If the confidence score is less than the threshold confidence, then a counter is incremented 422. When the counter reaches a threshold, N, then the points may be added to a new cluster in 426 and the model may be updated with further training in 428. The updated model 430 is then used in subsequent classifications.

Test Data

To test the efficacy of the method 200, sample RAN CSRs from an operator over a one hour time period were obtained. First, these were combined to arrive at aggregated CRSs per unique id. For the purpose of this test, the throughtput_out and Latency_out features were collected and the self-supervised learning discussed above was used according to the method 200) to obtain clusters. The sample clusters 502 obtained are illustrated in Fig. 5. As can be seen, the data can be divided into clusters based on these features and thus, different pricing values can be allocated to each different cluster. These can then be charged differently.

Fig. 6 shows the end-to-end system flow for metering record generation through the proposed model hosted within the BSS mediation system.

Step 61) RAN Operations push necessary configurations to the RAN domain 604 through RAN Domain Manager 602, to apply necessary scheduling and priority configurations against specific 5G quality indicator (5QI) values in the 5QI mapping table. 5QI table can be further linked to a combination of Public Land Mobile Network Identifier (PLMN ID) and Single Network Slice Selection Assistance Information (S-NSSAI) values configured in a MOCN setup. RAN Domain Manager applies the QoS profile configurations into targeted managed baseband objects and activates the desired RAN QoS profiles identified by unique 5QI values.

Step 62) Applied RAN QoS configurations can be exported and pushed into a "Traffic Clustering & Predictive Metering Model" 606 that performs the method 200 described above. Based on the RAN QoS configurations applied, along with traffic pattern data available from CSP's network operations, the theoretical probabilistic QoS dimension thresholds can be calculated for different traffic types.

Step 63) Once the configured QoS flows carry user plane traffic in the RAN, RAN CSR 610 generates the session logs that identifies the RAN traffic volume under various configured VLAN identifiers. It is also assumed here the QoS dimensions to be observed in traffic clustering (e.g. bandwidth, latency etc,) are also available as part of the CSR logs

Step 64) RAN CSR session logs are also exported and used by the "Traffic Clustering & Predictive Metering Model" to determine the traffic QoS clusters by applying the ML algorithm(s) on the observed QoS dimensions in the CSR logs.

Step 65) Based on the QoS clusters derived, the "Traffic Clustering & Predictive Metering Model" further generates the metering records for each observed cluster along with predictive uplink and downlink traffic volume.

Step 66) Generated metering records are transferred to BSS rating/ charging system 612 where differential rates can be applied on the data volumes based on the QoS cluster identifier.

Thus, in this way, the method 200 may be used to determine QoS differentiated billing based on CSR log data. The signal diagram in Fig. 6 may be used to determine QoS clusters, which can be used to determine the volume of usage against each cluster and derive the accuracy and dependability based on reference theoretical limits computed based on RAN QoS profile (slice) configurations applied.

As discussed above, one of the fundamental reasons for 5G adoption among operators is to increase revenue growth by offering connectivity towards enterprises to fulfil industry specific use cases. Depending on the nature of the targeted industry vertical and use cases, the variety of QoS demands from the 5G connectivity providers are huge and CSPs want to apply QoS differentiation to be able to price the connectivity differentially to increase their share of revenue and RAN Wholesale business in the 5G era is no exception.

With traditional business models CSPs have looked upon the differential monetization aspects from a Core perspective, and as such, earlier generations of mobile

technology and hence the 5G Core has ample levers (e.g., Charging Function (CHF) and Charging Access Function (CAF) embedded in core) to track volumetric usage at a per subscriber and service level that can be aggregated upwards towards B2B monetization like inter operator settlements.

However, with recent business models like "RAN as a Service" (such as MOCN as described above), where operators don't have the subscriber visibility (without ownership of the core), the ability to track and differentiate RAN traffic by QoS at the edge of the RAN and meter them using differentiated pricing rule is a relatively new concept that CSPs are exploring.

While there can be a "bottoms-up" approach of handling the problem through investment in portfolio upgrade and by building the capability at a commercial deployment level to track and segregate QoS based traffic volume, what has been discussed herein is a "top-down" view that applies a predictive approach to determining the QoS differentiated traffic clusters with sufficient level of confidence that can give RAN Wholesalers a manner in which to perform QoS differentiated monetization.

Turning now to other embodiments, there is also provided a computer program product comprising a computer readable medium, the computer readable medium having computer readable code embodied therein, the computer readable code being configured such that, on execution by a suitable computer or processor, the computer or processor is caused to perform the method or methods described herein.

Thus, it will be appreciated that the disclosure also applies to computer programs, particularly computer programs on or in a carrier, adapted to put embodiments into practice. The program may be in the form of a source code, an object code, a code intermediate source and an object code such as in a partially compiled form, or in any other form suitable for use in the implementation of the method according to the embodiments described herein.

It will also be appreciated that such a program may have many different architectural designs. For example, a program code implementing the functionality of the method or system may be sub-divided into one or more sub-routines. Many different ways of distributing the functionality among these sub-routines will be apparent to the skilled person. The sub-routines may be stored together in one executable file to form a self-contained program. Such an executable file may comprise computer-executable instructions, for example, processor instructions and/or interpreter instructions (e.g. Java interpreter instructions). Alternatively, one or more or all of the sub-routines may be stored in at least

one external library file and linked with a main program either statically or dynamically, e.g. at run-time. The main program contains at least one call to at least one of the sub-routines. The sub-routines may also comprise function calls to each other.

The carrier of a computer program may be any entity or device capable of carrying the program. For example, the carrier may include a data storage, such as a ROM, for example, a CD ROM or a semiconductor ROM, or a magnetic recording medium, for example, a hard disk. Furthermore, the carrier may be a transmissible carrier such as an electric or optical signal, which may be conveyed via electric or optical cable or by radio or other means. When the program is embodied in such a signal, the carrier may be constituted by such a cable or other device or means. Alternatively, the carrier may be an integrated circuit in which the program is embedded, the integrated circuit being adapted to perform, or used in the performance of, the relevant method.

Variations to the disclosed embodiments can be understood and effected by those skilled in the art in practicing the claimed invention, from a study of the drawings, the disclosure and the appended claims. In the claims, the word "comprising" does not exclude other elements or steps, and the indefinite article "a" or "an" does not exclude a plurality. A single processor or other unit may fulfil the functions of several items recited in the claims. The mere fact that certain measures are recited in mutually different dependent claims does not indicate that a combination of these measures cannot be used to advantage. A computer program may be stored/distributed on a suitable medium, such as an optical storage medium or a solid-state medium supplied together with or as part of other hardware, but may also be distributed in other forms, such as via the Internet or other wired or wireless telecommunication systems. Any reference signs in the claims should not be construed as limiting the scope.

CLAIMS:

1. A computer implemented method performed by a first node in a communications network, the method comprising:

clustering (202) session data according to network usage pattern, to obtain a plurality of clusters;

associating (204) each cluster in the plurality of clusters with an identifier; and

training (206) a model using a machine learning process, to take session data of a session as input and output an identifier associated with the cluster to which the session data belongs.

2. A method as in claim 1 further comprising:

associating each identifier with a traffic type, each traffic type having a different quality of service, QoS profile.

3. A method as in claim 2 wherein the step of associating is performed by comparing network usage patterns of sessions in the clusters to reference QoS profiles for said traffic types.

4. A method as in any one of the preceding claims wherein the method further comprises:

i) receiving first session data for a first new session;

ii) providing the first session data as input to the model and obtaining as output a first identifier predicted for the first new session;

iii) determining a first confidence score reflecting a confidence with which the first identifier is predicted for the first session data; and

iv) labelling the first session data with the first identifier if the first confidence score satisfies a first criteria.

5. A method as in claim 4 further comprising:

labelling the first session data with a first new identifier if the first confidence score does not satisfy the first criteria.

6. A method as in claim 5 further comprising:

repeating steps i) to iv) for a plurality of new sessions; and

re-training the model if a number of the new sessions associated with the first new identifier satisfies a second criteria.

7. A method as in claim 6 wherein the step of re-training comprises:

perturbing values of one or more weights or biases in the model to obtain a perturbed model; and

performing training on the perturbed model.

8. A method as in any one of the preceding claims further comprising: associating each identifier with a charging profile.

9. A method as in claim 8 further comprising repeating the steps of:

i) receiving second session data for a second new session;

ii) providing the second session data as input to the model and obtaining as output a second identifier predicted for the second new session;

iii) associating a second charging profile associated with the second identifier with the second new session.

10. A method as in any one of claims 2 to 9 wherein the traffic types comprise one or more of:

enhanced mobile broadband, eMBB;

massive machine type communication, mMTC;

critical machine type communication, cMTC; and

ultra-reliable low-latency communication, URLLC.

11. A method as in any one of claims 2 to 9 wherein the traffic types correspond to different network slices with different Quality of Service traffic flows.

12. A method as in any one of the preceding claims wherein the first node is operated by a Radio Access Network, RAN, Wholesaler as part of a Multiple Operator Core Network, MOCN, setup; and

wherein the sessions are operated by different Mobile Network Operators, MNOs, accessing services through the RAN wholesaler.

13. A method as in any one of the preceding claims wherein the method further comprises obtaining the session data from cell-site router, CSR, session logs.

14. A first node in a communications network, the first node comprising:

a memory comprising instruction data representing a set of instructions; and

a processor configured to communicate with the memory and to execute the set of instructions, wherein the set of instructions, when executed by the processor, cause the processor to:

cluster session data according to network usage pattern, to obtain a plurality of clusters;

associate each cluster in the plurality of clusters with an identifier; and

train a model using a machine learning process, to take session data of a session as input and output an identifier associated with the cluster to which the session data belongs.

15. A first node as in claim 14 further configured to perform the method of any one of claims 2 to 13.

16. A first node in a communications network, wherein the first node is configured to:

cluster session data according to network usage pattern, to obtain a plurality of clusters;

associate each cluster in the plurality of clusters with an identifier; and

train a model using a machine learning process, to take session data of a session as input and output an identifier associated with the cluster to which the session data belongs.

17. A first node as in claim 16 further configured to perform the method of any one of claims 2 to 13.

18. A computer program comprising instructions which, when executed on at least one processor, cause the at least one processor to carry out a method according to any of claims 1 to 13.

19. A carrier containing a computer program according to claim 18, wherein the carrier comprises one of an electronic signal, optical signal, radio signal or computer readable storage medium.

20. A computer program product comprising non transitory computer readable media having stored thereon a computer program according to claim 18.

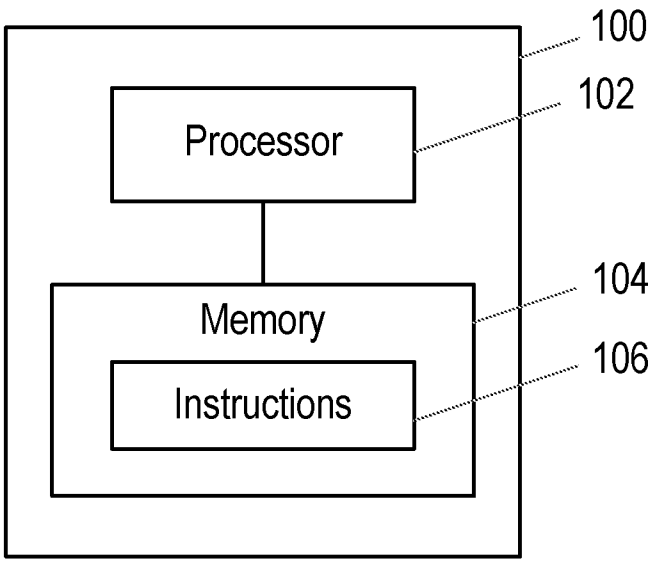

Fig. 1

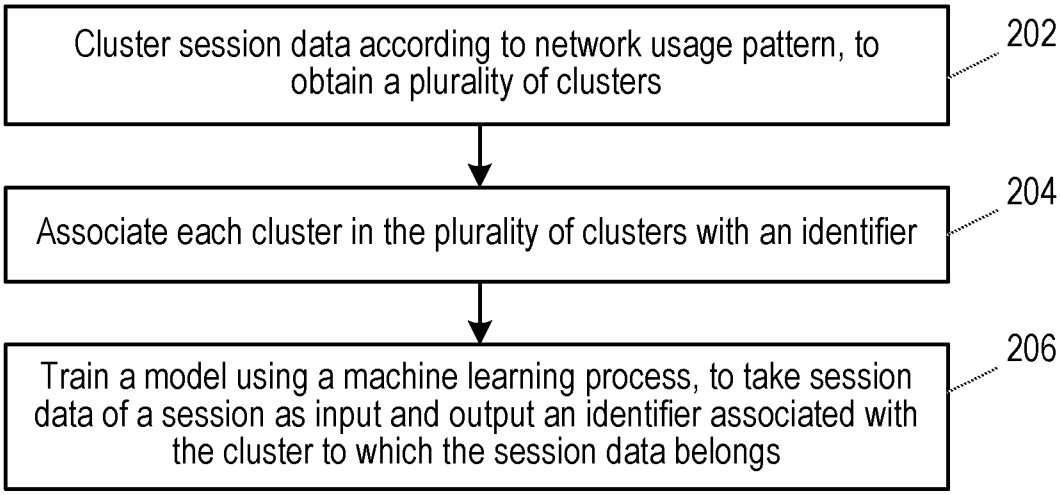
Cluster session data according to network usage pattern, to obtain a plurality of clusters
202
Associate each cluster in the plurality of clusters with an identifier
204
Train a model using a machine learning process, to take session data of a session as input and output an identifier associated with the cluster to which the session data belongs
206

Fig. 2

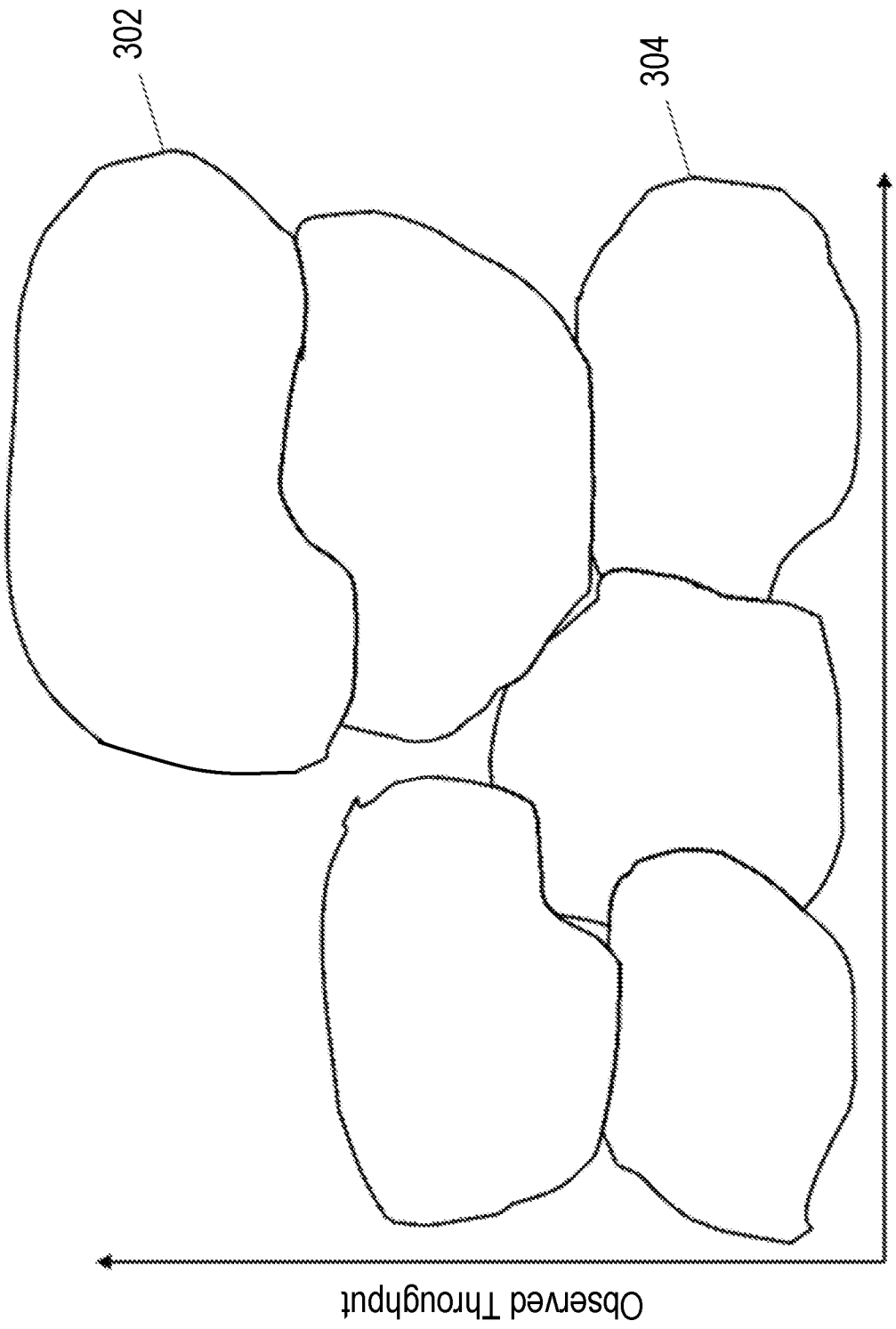

Fig. 3

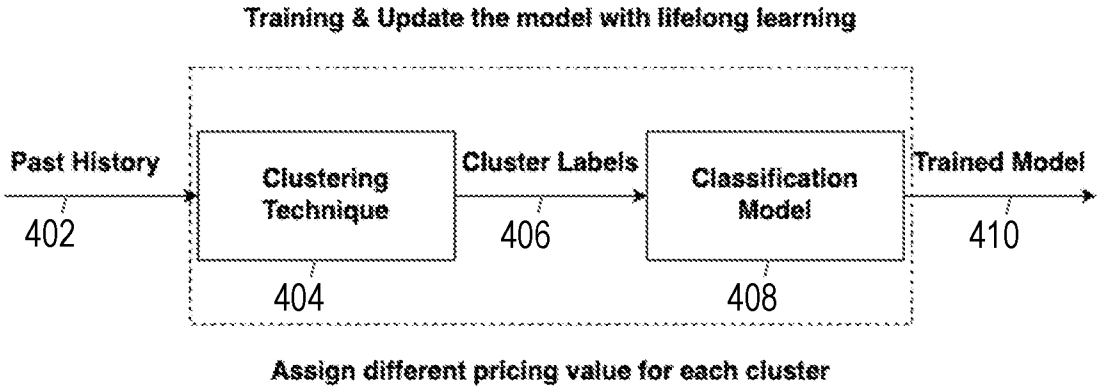
Training & Update the model with lifelong learning
Past History
402
Clustering Technique
404
Cluster Labels
406
Classification Model
408
Trained Model
410
Assign different pricing value for each cluster

Fig. 4a

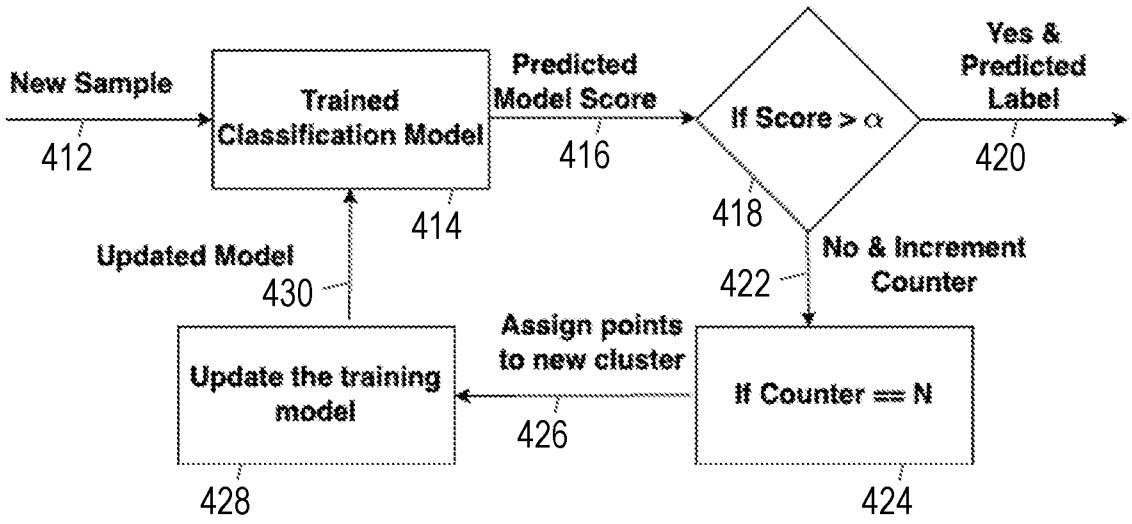
New Sample
412
Trained Classification Model
414
Predicted Model Score
416
If Score > α
418
Yes & Predicted Label
420
No & Increment Counter
422
If Counter == N
424
Assign points to new cluster
426
Update the training model
428
Updated Model
430

Fig. 4b

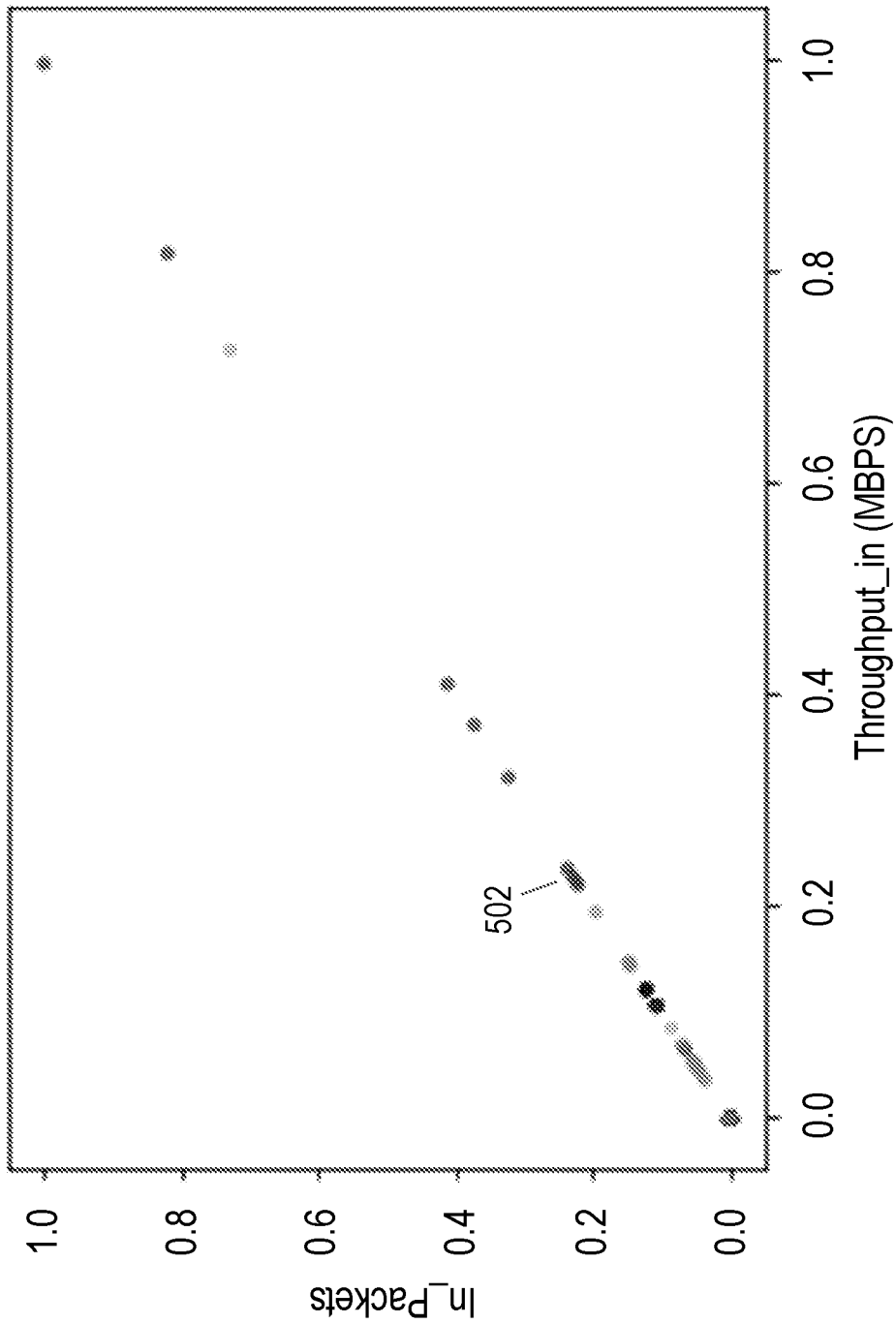
In_Packets
1.0
0.8
0.6
0.4
0.2
0.0
502
0.0
0.2
0.4
0.6
0.8
1.0
Throughput_in (MBPS)

Fig. 5

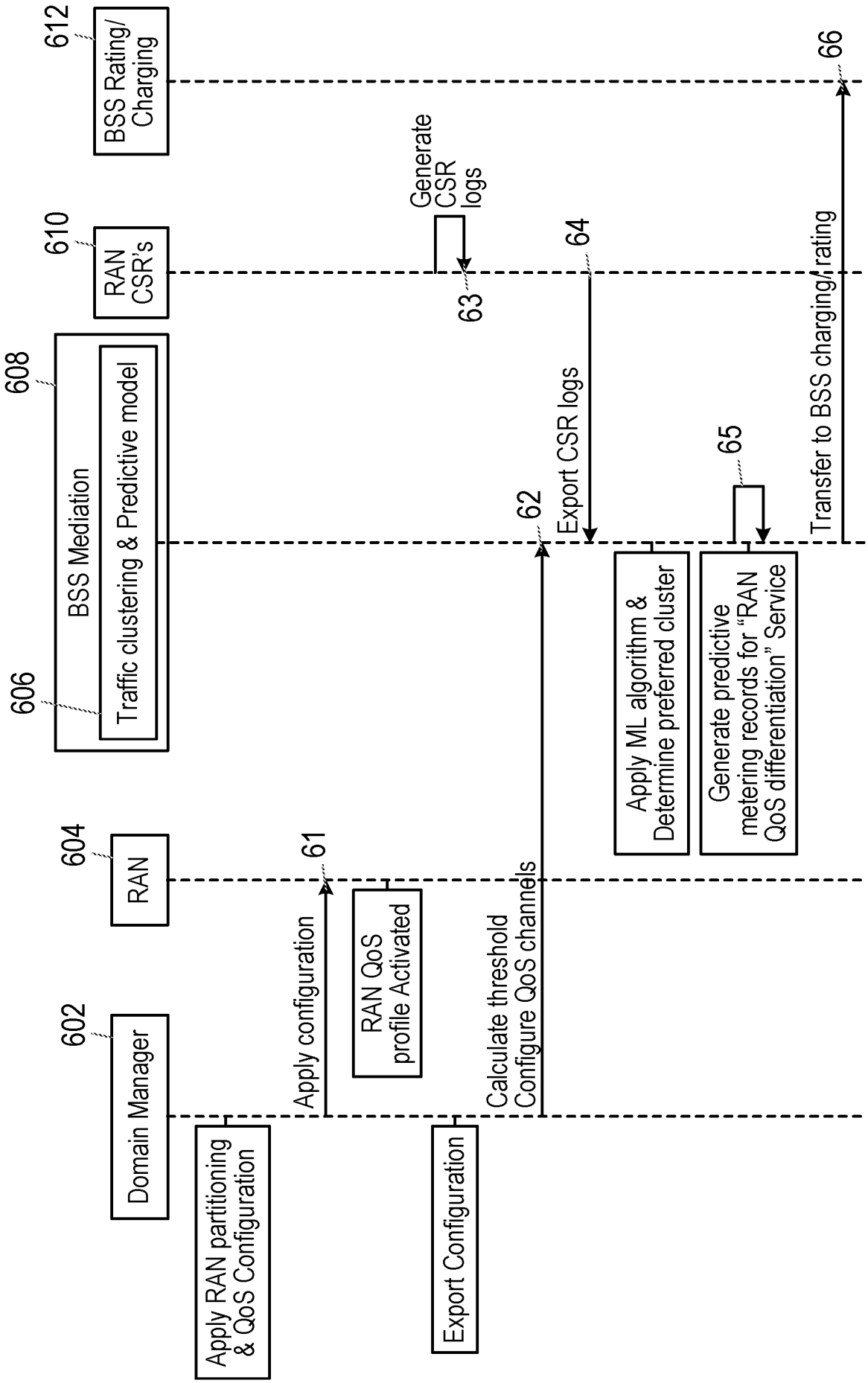
602
Domain Manager
604
RAN
606
Traffic clustering & Predictive model
608
BSS Mediation
610
RAN CSR's
612
BSS Rating/ Charging
Apply RAN partitioning & QoS Configuration
Apply configuration
61
RAN QoS profile Activated
Export Configuration
Calculate threshold Configure QoS channels
62
Generate CSR logs
63
Export CSR logs
64
Apply ML algorithm & Determine preferred cluster
Generate predictive metering records for "RAN QoS differentiation" Service
65
Transfer to BSS charging/ rating
66

Fig. 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/IN2023/050154

A. CLASSIFICATION OF SUBJECT MATTER
H04L67/01,H04L67/14 Version=2023.01

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

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

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

Databases: Google Patents, PatSeer, IPO Internal Database
Keywords: cluster, traffic type, RAN, network usage pattern

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Walelgne, Ermias Andargie, et al. "Clustering and predicting the data usage patterns of geographically diverse mobile users." Computer Networks 187 (2021): 107737. 14 March 2021 (14/03/2021) Abstract, 3. Clustering mobile users, 4. Mobile user classification and prediction	1-3, 14-17
Y	2.1. Measurement platform	4-13
Y	US 2023037228 A1 (Tata Consultancy Services Limited) 02 February 2023 (02/02/2023) Whole Document	4-13

☐ 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	Date of mailing of the international search report
12-06-2023	12-06-2023

Name and mailing address of the ISA/	Authorized officer
Indian Patent Office Plot No.32, Sector 14,Dwarka,New Delhi-110075 Facsimile No.	Nikhil Katiyar Telephone No. +91-1125300200

INTERNATIONAL SEARCH REPORT

International application No.
PCT/IN2023/050154

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.: 18-20
because they relate to subject matter not required to be searched by this Authority, namely:
The subject matter of claims 18-20 relates to computer program product, which does not require an international search by the International Searching Authority in accordance with PCT Article 17(2)(a)(i) and [Rule 39.1(vi)].

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:

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.:

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.

Citation	Pub.Date	Family	Pub.Date
US 20230037228 A1	02-02-2023	EP 4120777 A1	18-01-2023