



- (51) International Patent Classification:
H04L 12/715 (2013.01) *H04L 29/12* (2006.01)
- (21) International Application Number:
PCT/EP2016/065242
- (22) International Filing Date:
30 June 2016 (30.06.2016)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
16167292.8 27 April 2016 (27.04.2016) EP
- (71) Applicant: NEC EUROPE LTD. [DE/DE]; Kurfürsten-Anlage 36, 69115 Heidelberg (DE).
- (72) Inventors: SCHNEIDER, Fabian; Langer Anger 71, 69115 Heidelberg (DE). BIFULCO, Roberto; Bergheimerstraße 38, 69115 Heidelberg (DE).
- (74) Agent: ULLRICH & NAUMANN; Schneidmühlstraße 21, 69115 Heidelberg (DE).
- (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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, 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, 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.

(54) Title: METHOD FOR CONTROLLING A NETWORK

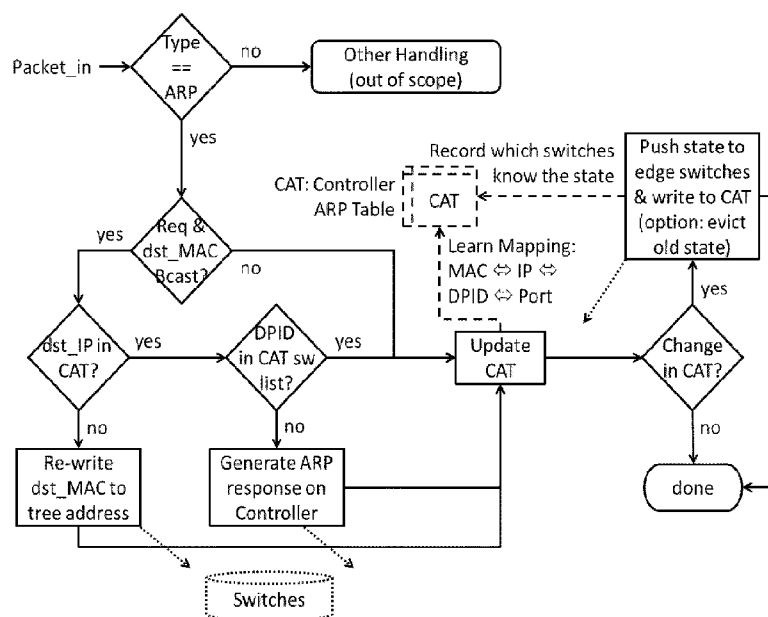


Fig. 1

(57) Abstract: The present invention relates to a method for controlling a network, said network comprising a plurality of forwarding elements, 'FE', being connected with each other, one or more end hosts, 'EH', connected to one or more of said FE, and a controller for controlling said FE, wherein packet processing rules for end-host control protocols, 'ECP', like the address resolution protocol, 'ARP', are installed on said FE, wherein a) When an ECP Request, 'ECPRQ', is received by an FE and a1) said ECPRQ was not processed by said controller, said ECPRQ is provided to said controller and an ECP response is computed by said receiving FE based on extracted information from said ECPRQ mapped onto forwarding information based on mapping information if provided, otherwise a2) if said ECPRQ was processed by said controller, said ECPRQ is forwarded according to forwarding information of said ECPRQ, b) When an ECP response, 'ECPR' is received by an FE Said ECPR is forwarded according to forwarding information and if said receiving FE is



(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))*

an edge FE, said ECPR is provided to said controller, and wherein c) an ECP database is maintained and updated by said controller, said ECP database comprising for each EH its network location, protocol network location and network attachment information and at least in part a certain state of said ECP database is installed on some of said FE.

METHOD FOR CONTROLLING A NETWORK

The EP patent application EP 16 167 292.8 is herein incorporated by reference.

5

The present invention relates to a method for controlling a network like a software-defined network, said network comprising a plurality of forwarding elements, 'FE', being connected with each other, one or more end hosts, 'EH', connected to one or more of said FE, and a controller for controlling said FE, wherein packet processing rules for end-host control protocols, 'ECP', like the address resolution protocol, 'ARP', are installed on said FE.

10

The present invention further relates to a controller for controlling a network like a software-defined network, said network comprising a plurality of forwarding elements, 'FE', being connected with each other, one or more end hosts, 'EH', connected to one or more of said FE, wherein packet processing rules for end-host control protocols, 'ECP', like the address resolution protocol, 'ARP', are installed on said FE.

15

The present invention further relates to a forwarding element, 'FE', controlled by a controller of a network like a software-defined network, said network comprising said FE and one or more further FE being connected with each other, and one or more end hosts, 'EH', connected to said FE, wherein packet processing rules for end-host control protocols, 'ECP', like the address resolution protocol, 'ARP', are installed on said FE.

20

25

The present invention further relates to a network like a software-defined network, said network comprising a plurality of forwarding elements, 'FE', one or more end hosts connected to one or more of said FE on the controller.

30

The present invention even further relates to a non-transitory computer readable medium storing a program causing a computer to execute a method for controlling a network like a software-defined network, said network comprising a plurality of forwarding elements, 'FE', being connected with each other, one or more end

- 2 -

hosts, 'EH', connected to one or more of said FE, and a controller for controlling said FE, wherein packet processing rules for end-host control protocols, 'ECP', like the address resolution protocol, 'ARP', are installed on said FE.

5 The present invention even further relates to a non-transitory computer readable medium storing a program causing a computer to execute a method for operating a controller for controlling a network like a software-defined network, said network comprising a plurality of forwarding elements, 'FE', being connected with each other, one or more end hosts, 'EH', connected to one or more of said FE, wherein
10 packet processing rules for end-host control protocols, 'ECP', like the address resolution protocol, 'ARP', are installed on said FE.

Although applicable in general to any kind of network, the present invention will be described with regard to Software-defined networks.

15 Although applicable in general to any kind of end host and/or user-device generated control protocols that perform the mapping e.g. between network addresses at different levels of the network's protocol stack, the present invention will be described with regard to the address resolution protocol ARP according to
20 RFC 826.

One of the main principles of Software-defined networks - SDN - is to decouple the packet processing and forwarding functions (execution, often called data plane) from the functions that determine what to do to packets and where to forward them
25 (decision, often called control plane). A second main principle is then to centralize the control functions and package them in a so called SDN controller that in turn programs the packet processing and forwarding functions in SDN switches (network elements).

30 This is a difference to the well-known and still used distributed network control protocols, where control and execution is co-located in network elements and each network element determines what to do to packets and where to forward them.

However one of the problems with SDN today is the lack of support for legacy network components and the integration of the different control methods e.g. centralized vs. distributed. A conventional method to deal with legacy control protocols is to relay them to the SDN controller and let it

- (1) inspect the legacy control packets,
- (2) update its databases,
- (3) calculate required changes, and if required
- (4) re-program the SDN switches and finally
- (5) respond to the incoming control traffic.

However this is only appropriate for low-volume and inter-network (across network boundaries) legacy control protocols such as Routing (BGP, OSPF, etc.), DNS, Multicast. In contrast thereto end-host and user-device generated control protocols, such as ARP (Address Resolution Protocol, RFC 826), cause a lot of traffic and besides the need to be supported, they cannot be efficiently handled.

In WO 2015/000517 A1 packet templates are shown for programming in-switch response generation to request packets e.g. for legacy control packets. Other conventional handling of ARP packets in software-defined networks is for example disclosed in the non-patent literature of Roberto Bifulco, Julien Boite, Mathieu Bouet, Fabian Schneider. Improving SDN with InSPired Switches, In proceedings ACM Symposium on SDN Research (SOSR), 2016 and in the non-patent literature of Hyunjeong Cho, Saehoon Kang and Younghee Lee, "Centralized ARP proxy server over SDN controller to cut down ARP broadcast in large-scale data center networks," 2015 International Conference on Information Networking (ICOIN), Cambodia, 2015, pp. 301-306. doi: 10.1109/ICOIN.2015.7057900, the latter showing an interworking between a virtual machine instantiation and an SDN controller, wherein ARP is implemented via a proxy in the SDN controller.

One of the problems addressed by embodiments of the present invention is to reduce the ARP traffic and to reduce interactions between the controller and forwarding elements while retaining flexibility.

In an embodiment the present invention provides a method for controlling a virtual network like a software-defined network, said virtual network being based on an underlying physical network, said physical network comprising a plurality of forwarding elements, 'FE', being connected with each other, one or more end hosts, 'EH', connected to one or more of said FE, and a controller for controlling said FE, wherein packet processing rules for end-host control protocols, 'ECP', like the address resolution protocol, 'ARP', are installed on said FE, wherein

a) When an ECP Request, 'ECPRQ', is received by an FE and

a1) said ECPRQ was not processed by said controller, said ECPRQ is provided to said controller and an ECP response is computed by said receiving FE based on extracted information from said ECPRQ mapped onto forwarding information based on mapping information if provided, otherwise

a2) if said ECPRQ was processed by said controller, said ECPRQ is forwarded according to forwarding information of said ECPRQ,

b) When an ECP response, 'ECPR', is received by an FE

Said ECPR is forwarded according to forwarding information and if said receiving FE is an edge FE, said ECPR is provided to said controller,
and wherein

c) an ECP database is maintained and updated by said controller, said ECP database comprising for each EH its network location, protocol network location and network attachment information and at least in part a certain state of said ECP database is installed on some of said FE.

In a further embodiment the present invention provides a controller for controlling a virtual network like a software-defined network, said virtual network being based on an underlying physical network, said physical network comprising a plurality of forwarding elements, 'FE', being connected with each other, one or more end hosts, 'EH', connected to one or more of said FE, wherein packet processing rules for end-host control protocols, 'ECP', like the address resolution protocol, 'ARP', are installed on said FE, said controller being adapted to

- 5 -

A) Receive an ECP Request, 'ECPRQ', from an FE when said ECPRQ was not processed by said controller, to

B) Receive an ECP response, 'ECPR' from an FE if said receiving FE is an edge FE, to

5 C) Maintain and update an ECP database said ECP database comprising for each EH its network location, protocol network location and network attachment information and to

D) Install at least in part a certain state of said ECP database on some of said FE.

10

In a further embodiment the present invention provides a forwarding element, 'FE', controlled by a controller of a virtual network like a software-defined network, said virtual network based on an underlying physical network, said physical network comprising a plurality of forwarding elements, 'FE', being connected with each other, one or more end hosts, 'EH', connected to said FE, wherein packet processing rules for end-host control protocols, 'ECP', like the address resolution protocol, 'ARP', are installed on said FE, wherein said FE being adapted to

15

A1) Receive an ECP Request, 'ECPRQ', and when said ECPRQ was not processed by said controller, to provide said ECPRQ to said controller, and to

20

B1) Compute an ECP response based on extracted information from said ECPRQ mapped onto forwarding information based on mapping information if provided, otherwise

C1) if said ECPRQ was processed by said controller, to forward said ECPR according to forwarding information of said ECPRQ, and to

25

D1) Receive an ECP response, 'ECPR' and to forward said ECPR according to forwarding information and if said FE is an edge FE, to provide said ECPR to said controller.

30

In a further embodiment the present invention provides a virtual network like a software-defined network, said virtual network being based on an underlying physical network, said physical network comprising a plurality of forwarding elements, 'FE' according to claim 12, one or more end hosts, 'EH', connected to one or more of said FE, and a controller according to claim 11.

In an even further embodiment the present invention provides a non-transitory computer readable medium storing a program causing a computer to execute a method for controlling a virtual network like a software-defined network, said virtual network being based on an underlying physical network, said physical network comprising a plurality of forwarding elements, 'FE', being connected with each other, one or more end hosts, 'EH', connected to one or more of said FE, and a controller for controlling said FE, wherein packet processing rules for end-host control protocols, 'ECP', like the address resolution protocol, 'ARP', are installed on said FE, wherein

- a) When an ECP Request, 'ECPRQ', is received by an FE and
 - a1) said ECPRQ was not processed by said controller, said ECPRQ is provided to said controller and an ECP response is computed by said receiving FE based on extracted information from said ECPRQ mapped onto forwarding information based on mapping information if provided, otherwise
 - a2) if said ECPRQ was processed by said controller, said ECPRQ is forwarded according to forwarding information of said ECPRQ,

- b) When an ECP response, 'ECPR' is received by an FE
 - Said ECPR is forwarded according to forwarding information and if said receiving FE is an edge FE, said ECPR is provided to said controller,
- and wherein

- c) an ECP database is maintained and updated by said controller, said ECP database comprising for each EH its network location, protocol network location and network attachment information and at least in part a certain state of said ECP database is installed on some of said FE.

In an even further embodiment the present invention provides a non-transitory computer readable medium storing a program causing a computer to execute a method for operating a controller for controlling a virtual network like a software-defined network, said virtual network being based on an underlying physical

network, said physical network comprising a plurality of forwarding elements, 'FE', being connected with each other, one or more end hosts, 'EH', connected to one or more of said FE, wherein packet processing rules for end-host control protocols, 'ECP', like the address resolution protocol, 'ARP', are installed on said FE, said
5 method comprising

A1) Receiving an ECP Request, 'ECPRQ', from an FE when said ECPRQ was not processed by said controller,

B1) Receiving an ECP response, 'ECPR' from an FE if said receiving FE is an edge FE,

10 C1) Maintaining and updating an ECP table said ECP table comprising for each EH its network location, protocol network location and network attachment information and

D1) Installing at least in part a certain state of said ECP table on some of said FE.

15
The terms "forwarding element", "end host", "controller", etc. refer in particular in the claims, preferably in the specification each to a device adapted to perform
20 computing like a personal computer, a tablet, a mobile phone, a server, or the like and comprises one or more processors having one or more cores and may be connectable to a memory for storing an application which is adapted to perform corresponding steps of one or more of the embodiments of the present invention. Any application may be software-based and/or hardware-based installed in the
25 memory on which the processor(s) can work on. The computing devices may be adapted in such a way that the corresponding steps to be computed are performed in an optimized way. For instance different steps may be performed in parallel with a single processor on different of its cores. Further the devices may be identical forming a single computing device. The devices or devices may also be
30 instantiated as a virtual device running on a physical computing resource. Different devices may therefore be executed on said physical computing resource.

The term "forwarding element" further refers in particular in the claims, preferably in the specification to a device adapted to perform forwarding of information, data

etc. in the network, like a switch or a router or the like: Said device may have dedicated hardware support for forwarding and/or processing of packets in networks like computer networks.

5 The term “computer readable medium” may refer to any kind of medium, which can be used together with a computation device or computer and on which information can be stored. Said information may be any kind of data which can be read into a memory of a computer. For example said information may include program code for executing with said computer. Examples of a computer readable
10 medium are tapes, CD-ROMs, DVD-ROMs, DVD-RAMs, DVD-RWs, BluRay, DAT, MiniDisk, solid state disks SSD, floppy disks, SD-cards, CF-cards, memory-sticks, USB-sticks, EPROM. EEPROM or the like.

15 The term “rule” is to be understood in its broadest sense and refers in particular in the claims, preferably in the specification to one or more conditions to which one or more actions are assigned or linked, wherein said actions being executed or performed when said conditions are matched. For instance when pre-defined parameters like a type of the packet and size of the packet forming a condition are matched, then an action like forwarding said packet to a certain end host are
20 performed,

The terms “forwarding information” and “network location” are each to be understood in its broadest sense and refers in particular in the claims, preferably in the specification to a physical address, e.g. a hardware address like a MAC
25 address or the like.

The term “edge forwarding element” or “edge FE” is to be understood in its broadest sense and refers in particular in the claims, preferably in the specification to a forwarding element which is directly connected to at least one end host EH
30 said EH may run a legacy IP/TCP stacks and which may also generate ECP requests.

The term “topology” refers in particular in the claims, preferably in the specification to information about at least part of all network connections, the network comprising forwarding elements and end hosts being connected with each other.

5 The term “protocol network location” is to be understood in its broadest sense and refers in particular in the claims, preferably in the specification to protocol network addresses like IP addresses.

10 The term “network attachment point” is to be understood in its broadest sense and refers in particular in the claims, preferably in the specification to information about a data path, like data identification specifying one or more interfaces of the forwarding element.

15 The term “state” with regard to the ECP database is to be understood in its broadest sense and refers in particular in the claims, preferably in the specification to a relation, relationship, mapping or the like between hardware location information and protocol location information like hardware addresses or protocol addresses, e. g. a MAC-IP mapping.

20 The term “tree” refers in particular in the claims, preferably in the specification to at least two network devices being connected to each other and organized in a tree-like structure and being brought into an ordering. In particular the term “tree” with regard to “forwarding” refers to information which are forwarded according to the forwarding trees, said forwarding trees comprising of different forwarding elements
25 for forwarding the data along the forwarding elements of the tree to end hosts.

30 The term “exception” with regard to the term “situation” is to be understood in its broadest sense and refers in particular in the claims, preferably in the specification to any kind of situation, scenario, state, etc. within the virtual or physical network which is e.g. defined by one or more thresholds for certain parameters, said thresholds being exceeded or fallen below.

At least one embodiment may have at least one of the following advantages:

- 10 -

- fewer interactions between forwarding elements and the controller thereby reducing load on the corresponding control channel and the controller; this applies in particular when compared with a conventional ECP request like an ARP request being broadcast through the network
- 5
- fewer ECP traffic like ARP traffic in the controlled network; this applies in particular when compared to a ARP proxy in or used with the controller
- 10
- retained flexibility of the virtual network control to change the processing and forwarding at any time; this means for example that the controller decides which network element replies to the ECP requests like ARP requests which is being selected or chosen at closest to the source of the request as possible.
- 15
- Further features, advantages and further embodiments are described or may become apparent in the following:
- Said ECP may be the ARP.
- 20
- A topology of the network may be detected and forwarding information may be computed and provided in form of forwarding trees. This enables in a very effective way to prepare the virtual network and the underlying physical network for operation.
- 25
- To decide if an ECPRQ was processed by said controller it may be checked whether the extracted information corresponds to broadcast information like a broadcast address or the like. This enables an easy way to differentiate whether a request has been processed by the controller before or not. In case the extracted information corresponds to broadcast information then this request is defined to be
- 30
- not processed by the controller.

One or more states in the ECP database may be removed after certain periods of time. This allows during operation to keep the ECP database “clean” by removing

entries which are not used anymore. Thus resources for said database can be saved.

5 A state of said ECP database may be only installed on some of said edge FE. This enables in an optimized way to use computing resources within edge switches for example since a memory in a FE is usually very limited and for example in large networks it might be not desirable to replicate all IP-to-MAC mappings at each individual edge switch.

10 Said state of said ECP database for an EH may be only installed on edge FE at which an ECPRQ for said EH has been received. This allows to further enhance the efficiency since only rules and states are installed on edge switches receiving specific ECPRQ that match to the corresponding installed state.

15 Installing information of states installed on FE may be stored in the ECP database. This enhances further the operation and the installation of rules by the controller since the controller knows the states on corresponding FEs.

20 In case of an exception situation of said network, a higher priority rule in one or more of said FE may be installed omitting transmitting an ECPRQ and/or ECPR to said controller. This enables for example in case of an overload situation as an exception situation to avoid sending of ECPRQ and ECPR to the controller. Thus overload within the network is reduced.

25 Said higher priority rule may be provided with a hard timeout. This ensures that such a higher priority rule is without exception – “hard” - limited to a certain amount of time since for example such a higher priority rule prevents the controller from learning mappings and thus the controller is temporary not able to recognize the current situation with regard to installed states, etc. in the network.

30 States to be installed on FE may be pushed to said FE. This ensures that the controller can install at every time and continuously states on the FE without having to wait until switches request.

There are several ways how to design and further develop the teaching of the present invention in an advantageous way. To this end it is to be referred to the patent claims subordinate to the independent claims on the one hand and to the following explanation of further embodiments of the invention by way of example, illustrated by the figure on the other hand. In connection with the explanation of the further embodiments of the invention by the aid of the figure, generally further embodiments and further developments of the teaching will be explained.

In the drawings

Fig. 1 shows part of a method according to an embodiment of the present invention;

Fig. 2 shows part of a method according to a further embodiment of the present invention;

Fig. 3 shows part of a method according to a further embodiment of the present invention;

Fig. 4 shows part of a method according to a further embodiment of the present invention and

Fig. 5 shows part of a method according to a further embodiment of the present invention.

Fig. 1 shows part of a method according to an embodiment of the present invention.

In Fig. 1 a method for operating a controller of the virtual network is shown. Prior to perform the method for the controller according to Fig. 1 a preparation method or phase is performed:

In the preparation phase default ARP packet processing rules will be installed in the SDN switches, like shown in table 1 below with rules #1 - #4 the topology is detected, forwarding trees are calculated and corresponding forwarding entries will be installed in the switches flow tables, wherein a forwarding tree can be calculated e.g. using mechanisms such as spanning tree algorithms or multicast tree calculation and MAC addresses may be reserved to identify such forwarding trees. In a simple star topology network forwarding to such tree would e.g. correspond to broadcasting. In this case an additional MAC address for the broadcast tree beyond the default may be required. In more complex topologies one already uses multiple spanning trees, multicast trees or esp. in data centers virtual networks each with its own broadcast like forwarding tree. For some rules edge switches and other switches are differentiated. Edge switches are characterized in that they are directly connected to end-hosts that run legacy IP/TCP network stacks and therefore also generate ARP requests.

Table 1:

#	Match	Action
1	ARP_op==Request + ARP_MAC_dst==BCast	To_Controller + ARP_PacketTemplate* *If no template is available, skip this Action
2	ARP_op==Request + ARP_MAC_dst==Tree	Output (according to forwarding tree)
3	ARP_op==Reply @ Edge Switch	To_Controller + Output (according to forwarding tree)
4	ARP_op==Reply @ other Switch	Output (according to forwarding tree)

ARP_op refers to the ARP operation (either request or reply); ARP_MAC_dst refers to the ARP target hardware address; To_Controller means to send the received packet to the controller (in OpenFlow this is for example possible using the packet_in instruction); Output (according to forwarding tree) means to forward the packet along the typical L2 forwarding rules for the selected forwarding tree.

The in-switch packet generation is invoked with the ARP_PacketTemplate action e. g. shown in EP 16 167 292.8. Each edge switch holds a state table in which the SDN controller can install MAC address to IP address mappings. If the ARP_PacketTemplate action is triggered the following processing occurs:

- 5 1. The destination IP address is extracted from the received ARP request packet
2. This IP is looked up in the MAC $\leftrightarrow$ IP state table.
3. If an entry exists an ARP response packet is generated by
 - 10 a. Loading the template ARP response packet stored as part of the ARP_PacketTemplate
 - b. Copying the source/sender MAC and IP address to the destination/target MAC and IP address fields of the template
 - c. Copying the previously extracted destination IP address and the corresponding MAC address from the MAC $\leftrightarrow$ IP state table to the
 - 15 source/sender MAC and IP address fields of the template
 - d. Calculate the Ethernet checksum for the generated packet
 - e. Sending this packet out according to regular L2 forwarding.
4. If no entry exists, no further action is performed. Optionally in this case the
- 20 switch could notify the controller of a lack of entry.

20

Here ARP requests that have been processed by the SDN controller and those that have not are differentiated. This is reflected by the first two rules #1, #2. If the ARP_MAC_dst is the broadcast address, then it is defined this request has not been processed by the controller. Accordingly the request is sent to the controller.

25

If the ARP_MAC_dst is another address e.g. an identifier for a multicast tree it is assumed the packet was processed by the controller already, and will not sent it over again.

30 In a next step the operation of the controller is performed together with a optimization if needed:

In the operation phase the controller method will run in a loop together with the optional optimization. The controller algorithm or method maintains a controller ARP Table (CAT) which serves as a network-wide MAC learning table and records

mappings between Hardware Addresses (e.g. Ethernet MAC address), Protocol Addresses (e.g. IP Addresses), and the corresponding network attachment point. The controller method also installs part of this collected state into the switches so that the locally executed in-switch packet generation function can create ARP response packets close to the ARP requestor.

In the optimization phase, special forwarding rules can be installed in order to further reduce the load of packets sent to the controller and CAT state can be selectively shared with only subsets of all the switches.

Embodiments of invention are primarily targeted towards data center networks however they are applicable to all virtual networks like SDN networks that allow in-switch packet generation.

In the operation phase the SDN controller maintains the Controller-ARP table CAT and runs the method depicted in Fig. 1. An example of the CAT is given below:

Table 2: Example Controller-ARP Table (CAT)

MAC	IP	DPID	PortNr	Switches w/ State	Comment
00:00:00:00:00:01	192.168.0.1	0001	#1		Entry No. 1
01:02:03:04:05:06	192.168.0.2	0002	#1	0001	Entry No. 2
01:23:45:67:89:AB	10.11.12.13	0003	#52	0001,0002,0003	Entry No. 2

In this network-wide ARP table the controller maintains its knowledge about all end hosts EH attached to the network. For each end host EH it stores the MAC address of the end host, the IP address of the end-host and the network attachment point in form of the DPID (data path ID, specifies the switch) and port number in the switch with that DPID. In addition the switches are recorded to which this state (MAC $\leftrightarrow$ IP mapping) was transferred.

The content of the CAT table and instructions to the switches are governed by the steps as shown in Fig. 1.

When an ARP packet is sent to the Controller (according to the To_Controller action from the default rules), the method goes through a two step process. Both steps could be performed independently or subsequently. The first step (shown on the left Fig. 1 before entering the "Update CAT" process) is responsible for ensuring that an ARP response is produced. The second step (on the right in Fig. 1 starting with the "Update CAT" process) is responsible for learning the mappings and distributing this knowledge to the switches.

The first step is looking at the destination/target addresses of the packet. It begins with checking if the ARP Request is observed for the first time in the network, this is true when the dst_MAC address is the broadcast address. If this is not the case, as well as if the packet is an ARP reply, the second step is performed. If the dst_MAC address is the broadcast address there are two options, either the corresponding mapping (i.e. the dst_IP is in the CAT) was learned or not. If it was not learned yet, the packet is broadcasted along the forwarding tree and therefore replaces the dst_MAC address with the forwarding tree address. If it was learned already it is checked if this learned state has been installed at the receiving switch (i.e. is the DPID of the receiving switch in the switch list of the CAT). If so the receiving switch has already generated the ARP response according to the default rules and then CAT update is performed. If not, the response is generated on the controller (this case should be the exception, which can be handled in the optimization phase).

In the second step the controller updates the CAT with the source/sender information from the packet. If the learned this update may be limited to a timer reset of the corresponding entry in the CAT. If however in case of a change in the CAT, e.g. new entry or change of an entry resulted from the update, the controller needs to push this state to the edge switches, and to maintain the list of switches to which the corresponding MAC $\leftrightarrow$ IP mapping has been pushed. Optionally the old state may be evicted from the CAT in this step.

In a deployment scenario of data centers where virtual machines, 'VM', are instantiated by a cloud management system which likely knows details such as

the IP & MAC address and location of the instantiated VM, such information can be exchanged with the SDN controller to pre-fill the CAT. This is optional.

5 Gracious ARP replies, i.e. ARP replies that are sent deliberately without a corresponding ARP request in order to update ARP mapping tables, are properly handled by embodiments of the present invention. According to the rules all ARP replies on edge switches are forwarded to the controller. There the source host mapping is learned for ARP replies.

10 In the following Figures 2-5 an example of the operation of the controller according to the steps of Fig. 1 is shown.

Fig. 2 shows part of a method according to a further embodiment of the present invention.

15 In Fig. 2 an initial ARP request and no entries at the controller are shown.

In the beginning the CAT is assumed to be empty, because no ARP packet has been processed. As shown in Fig. 2 an ARP request is received by and at switch S0 and the packet is sent to the controller (rule #1). As no state is installed, no ARP reply is generated at the switch.

Fig. 3 shows part of a method according to a further embodiment of the present invention.

25 In Fig. 3 controller processing ARP request forwarding for an initial ARP request is shown.

30 Fig. 3 shows that according to the controller the dst_IP (in this case H99) has not been learned yet, the dst_MAC is re-written to the address of Tree1 (D0) and the packet is distributed along the specified forwarding tree. Concurrently, the mapping for H1 is learned and stored in the CAT (see top right). H99 will eventually receive the ARP request for itself. According to the default rules only S0 will send the packet to the controller, while all other switches forward the packet

along the forwarding tree because the dst_MAC has been re-written to the tree address D0 (rule #2).

Fig. 4 shows part of a method according to a further embodiment of the present invention.

In Fig. 4 an ARP reply to an initial ARP request is shown.

Fig. 4 shows the distribution of the resulting ARP reply. Upon reception of the ARP reply switch S9 sends a copy to the controller (rule #3), so that the mapping for H99 can be learned. As the packet is an ARP reply, only the second step of the controller operating method is processed. Simultaneously, the ARP reply is forwarded along the unicast path to H1, where it enables standard TCP/IP connections between H1 and H99.

Here the mapping state is distributed to all edge switches, as is shown in the rightmost column in the CATs in the top right corner. This means that after these two ARP packets (the request and the reply) all edge switches (S0-S9) will directly generate the ARP replies for H1 and H99, which is shown in Fig. 5.

To optimize the aforementioned method an optional optimization phase can be performed. Possible optimizations are as follows:

- Switch limitations: Often memory at switches is limited, therefore in large networks it might not be desirable to replicate the all IP to MAC mappings at each individual edge switch, esp. when for example all hosts in a specific virtual network are attached at only two switches (i.e. assume a virtual network of hosts H3,H5,H6,H7,H80, H81,H82 in the topology shown in Fig. 5) it would be useless to store the mappings for these hosts at other switches (i.e. only store them at S0 and S8).

Another example could be a more generic method or procedure that only installs mapping state for a host on edge switches at which an ARP request for that host arrived, or one that evicts state at switches after a certain

period of not having received ARP requests for that state. In order to enable such optimization the controller may keep track of where it installed the mapping state and needs the fallback to generate ARP replies it self. If such a fallback was triggered it may be advisable to install the corresponding mapping again.

- Controller limitations: SDN controllers need to process large numbers of packet_in's. Basically a copy of each ARP message received at an edge switch is sent to the controller, in order to update the CAT. In a steady state network this task degrades to a maintenance task, which however still generated a lot of processing load at the controller and the controller to switch communication channel. In order to address such potential overload situations the controller could temporarily choose to specific higher priority rules like the one in Table 3, which omit the packet_in to the controller.

Table 3: Special Rules (higher priority)

#	Match	Action
5	ARP_op==Request + ARP_MAC_dst==BCast + dst_IP==X	ARP_PacketTemplate*

In case the controller chooses to use this option stale state should be avoided. This rule basically prevents the controller from learning CAT mappings, i.e. both new mappings and updates to existing mappings. This could be addressed by deploying hard timeouts for such entries. Anyway outdated entries may cause other failure notifications, such as ICMP destination unreachable.

A further embodiment of the present invention provides a method for SDNized APR handling with in-switch packet generation

Comprising the steps of

- 1) Configuring default SDN rules for ARP processing, using programmable in-switch packet generation mechanisms (as defined in Table 1)
- 2) Running a ARP processing control algorithm in the SDN controller during network operations (as shown in Fig. 1)

In summary embodiments of the present invention enable or provide in particular

- 5 1) Design of an ARP handling method that leverages the concepts of SDN in general and in-switch packet generation (see earlier patent application) in particular. More specifically by
 - a. Using an additional L2 address for "broadcasting" ARP requests for unknown destination L3 addresses in order to avoid always notifying the SDN controller while forwarding on the Bcast tree
 - 10 b. Synchronizing the state learned at the controller with selected the edge switches, that can then use PacketTemplates to generate reply on switch rather than on controller.
 - c. Push state to selected switches, and store which switch has PacketTemplates installed
- 15 2) Combination of packet processing rules and controller algorithm to replace legacy end-host based ARP processing with network assisted ARP processing, to reduce network load.
- 3) Optional optimization approach to address switch memory limitations and controller and control channel load.

20 Embodiments of the present invention may have in particular inter alia one or more of the following advantages:

- 25 - Fewer ARP traffic in the controlled network. Esp. when compare to a basic all ARP requests are broadcast through the network case
- Fewer interactions between the SDN switch and the SDN controller (thereby reducing load on the SDN control channel and the SDN controller). Esp. when compared to a ARP proxy in the Controller deployment.
- 30 - While retaining the flexibility of SDN control to change the processing and forwarding at any time. I.e. the controller decides which network element replies to the ARP requests, ideally choosing the closest to the source.

Many modifications and other embodiments of the invention set forth herein will come to mind to the one skilled in the art to which the invention pertains having the

benefit of the teachings presented in the foregoing description and the associated drawings. Therefore, it is to be understood that the invention is not to be limited to the specific embodiments disclosed and that modifications and other embodiments are intended to be included within the scope of the appended claims. Although
5 specific terms are employed herein, they are used in a generic and descriptive sense only and not for purposes of limitation.

Claims

- 5 1. A method for controlling a network comprising a plurality of forwarding elements, 'FE', being connected with each other, one or more end hosts, 'EH', connected to one or more of said FE, and a controller for controlling said FE, wherein packet processing rules for end-host control protocols, 'ECP', like the address resolution protocol, 'ARP', are installed on said FE, wherein
- 10 a) When an ECP Request, 'ECPRQ', is received by an FE and
- a1) said ECPRQ was not processed by said controller, said ECPRQ is provided to said controller and an ECP response is computed by said receiving FE based on extracted information from said ECPRQ mapped onto forwarding information based on mapping information if provided, otherwise
- 15 a2) if said ECPRQ was processed by said controller, said ECPRQ is forwarded according to forwarding information of said ECPRQ,
- b) When an ECP response, 'ECPR' is received by an FE
- 20 Said ECPR is forwarded according to forwarding information and if said receiving FE is an edge FE, said ECPR is provided to said controller,
- and wherein
- c) an ECP database is maintained and updated by said controller, said ECP database comprising for each EH its network location, protocol network location and network attachment information and at least in part a certain state of said ECP database is installed on some of said FE.
- 25
- 30 2. The method according to claim 1, wherein a topology of the network is detected and forwarding information is computed and provided in form of forwarding trees.

3. The method according to one of the claims 1-2, wherein to decide if an ECPRQ was processed by said controller it is checked whether the extracted information corresponds to broadcast information like a broadcast address.
- 5 4. The method according to one of the claims 1-3, wherein one or more states in the ECP database are removed after certain periods of time.
- 5 5. The method according to one of the claims 1-4, wherein a state of said ECP database is only installed on some of said edge FE.
- 10 6. The method according to claim 5, wherein said state of said ECP database for an EH is only installed on edge FE at which an ECPRQ for said EH has been received.
- 15 7. The method according to one of the claims 1-6, wherein installing information of states installed on FE in the ECP database is stored.
- 20 8. The method according to one of the claims 1-7, wherein in case of an exception situation of said network, a higher priority rule in one or more of said FE is installed omitting transmitting an ECPRQ and/or ECPR to said controller.
9. The method according to claim 8, wherein said higher priority rule is provided with a hard timeout.
- 25 10. The method according to one of the claims 1-9, wherein states to be installed on FE are pushed to said FE.
- 30 11. A controller for controlling a network, said network comprising a plurality of forwarding elements, 'FE', being connected with each other, one or more end hosts, 'EH', connected to one or more of said FE, wherein packet processing rules for end-host control protocols, 'ECP', like the address resolution protocol, 'ARP', are installed on said FE, said controller being adapted to
 - A) Receive an ECP Request, 'ECPRQ', from an FE when said ECPRQ was not processed by said controller, to

- 5 B) Receive an ECP response, 'ECPR' from an FE if said receiving FE is an edge FE, to
- C) Maintain and update an ECP database said ECP database comprising for each EH its network location, protocol network location and network attachment information and to
- D) Install at least in part a certain state of said ECP database on some of said FE.
- 10 12. A forwarding element, 'FE', controlled by a controller of a network, said network comprising a plurality of forwarding elements, 'FE', being connected with each other, one or more end hosts, 'EH', connected to said FE, wherein packet processing rules for end-host control protocols, 'ECP', like the address resolution protocol, 'ARP', are installed on said FE, wherein said FE being adapted to
- 15 A1) Receive an ECP Request, 'ECPRQ', and when said ECPRQ was not processed by said controller, to provide said ECPRQ to said controller, and to
- B1) Compute an ECP response based on extracted information from said ECPRQ mapped onto forwarding information based on mapping information if provided, otherwise
- 20 C1) if said ECPRQ was processed by said controller, to forward said ECPR according to forwarding information of said ECPRQ, and to
- D1) Receive an ECP response, 'ECPR' and to forward said ECPR according to forwarding information and if said FE is an edge FE, to provide said ECPR to said controller.
- 25 13. A network comprising a plurality of forwarding elements, 'FE' according to claim 12, one or more end hosts, 'EH', connected to one or more of said FE, and a controller according to claim 11.
- 30 14. A non-transitory computer readable medium storing a program causing a computer to execute a method for controlling a network, said network comprising a plurality of forwarding elements, 'FE', being connected with each other, one or more end hosts, 'EH', connected to one or more of said

FE, and a controller for controlling said FE, wherein packet processing rules for end-host control protocols, 'ECP', like the address resolution protocol, 'ARP', are installed on said FE, wherein

a) When an ECP Request, 'ECPRQ', is received by an FE and

a1) said ECPRQ was not processed by said controller, said ECPRQ is provided to said controller and an ECP response is computed by said receiving FE based on extracted information from said ECPRQ mapped onto forwarding information based on mapping information if provided, otherwise

a2) if said ECPRQ was processed by said controller, said ECPRQ is forwarded according to forwarding information of said ECPRQ,

b) When an ECP response, 'ECPR' is received by an FE

Said ECPR is forwarded according to forwarding information and if said receiving FE is an edge FE, said ECPR is provided to said controller,

and wherein

c) an ECP database is maintained and updated by said controller, said ECP database comprising for each EH its network location, protocol network location and network attachment information and at least in part a certain state of said ECP database is installed on some of said FE.

15. A non-transitory computer readable medium storing a program causing a computer to execute a method for operating a controller for controlling a network, said network comprising a plurality of forwarding elements, 'FE', being connected with each other, one or more end hosts, 'EH', connected to one or more of said FE, wherein packet processing rules for end-host control protocols, 'ECP', like the address resolution protocol, 'ARP', are installed on said FE, said method comprising

A) Receiving an ECP Request, 'ECPRQ', from an FE when said ECPRQ was not processed by said controller,

B) Receiving an ECP response, 'ECPR' from an FE if said receiving FE is an edge FE,

- 26 -

- C) Maintaining and updating an ECP database said ECP database comprising for each EH its network location, protocol network location and network attachment information and
- D) Installing at least in part a certain state of said ECP database on some of said FE.

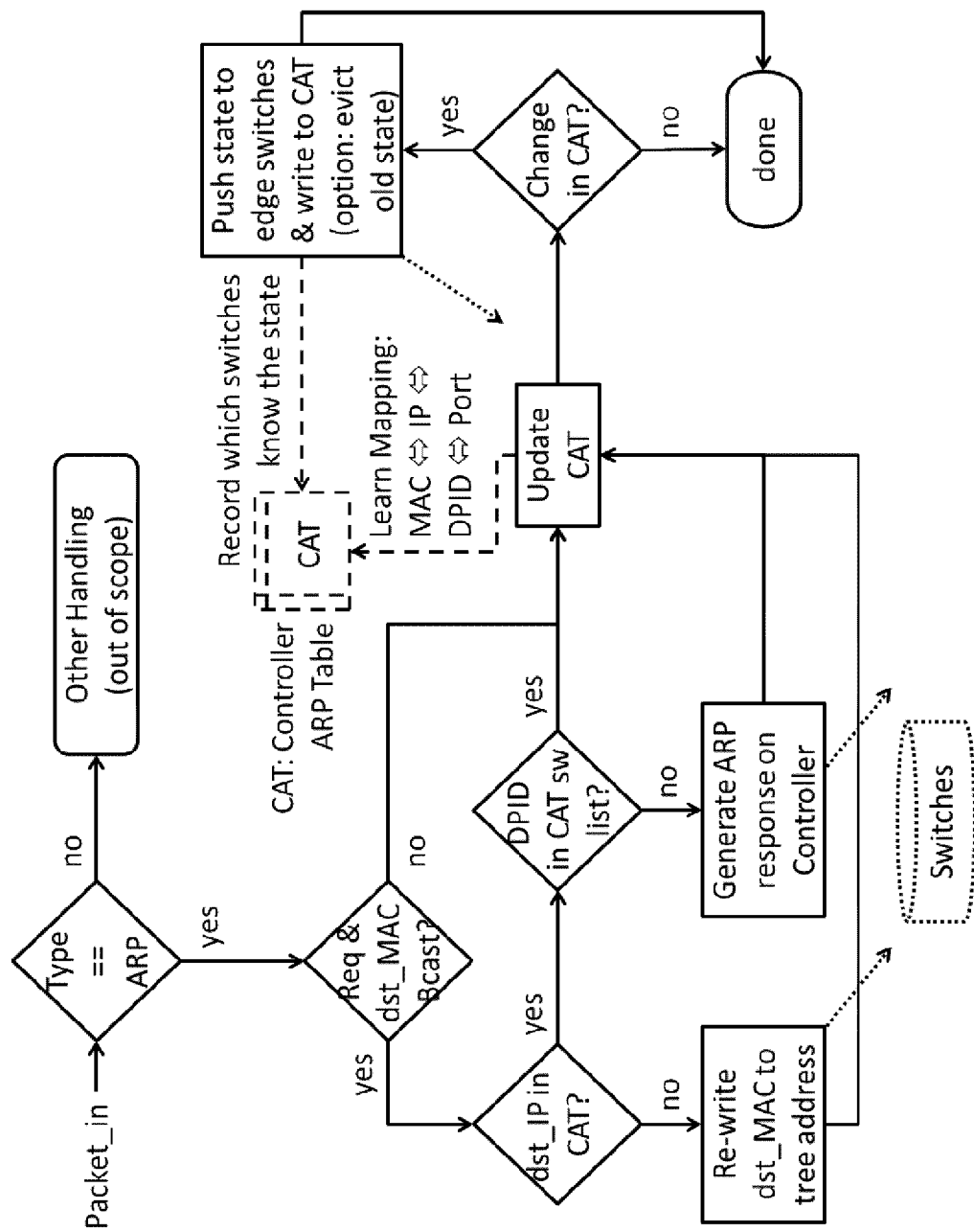


Fig. 1

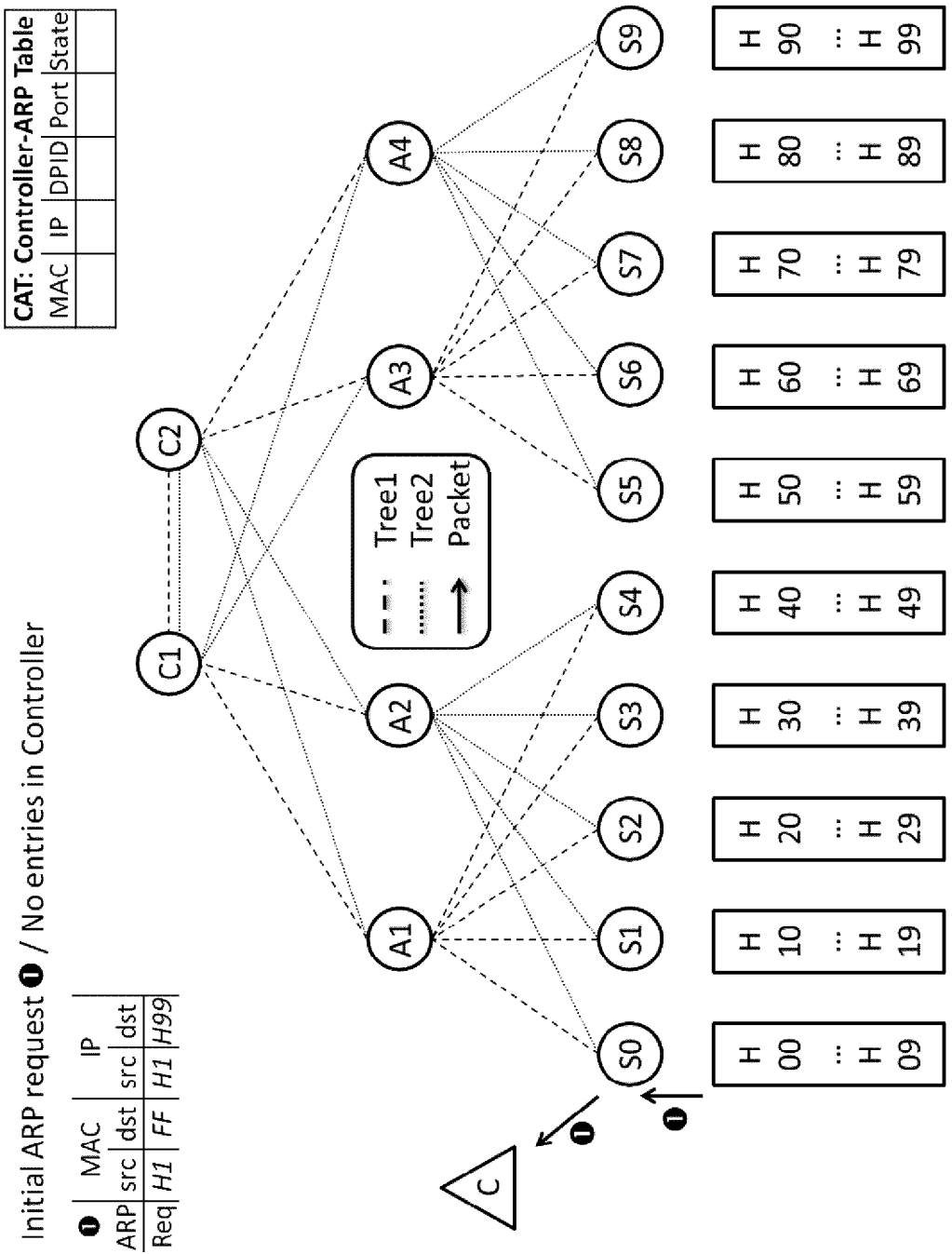


Fig. 2

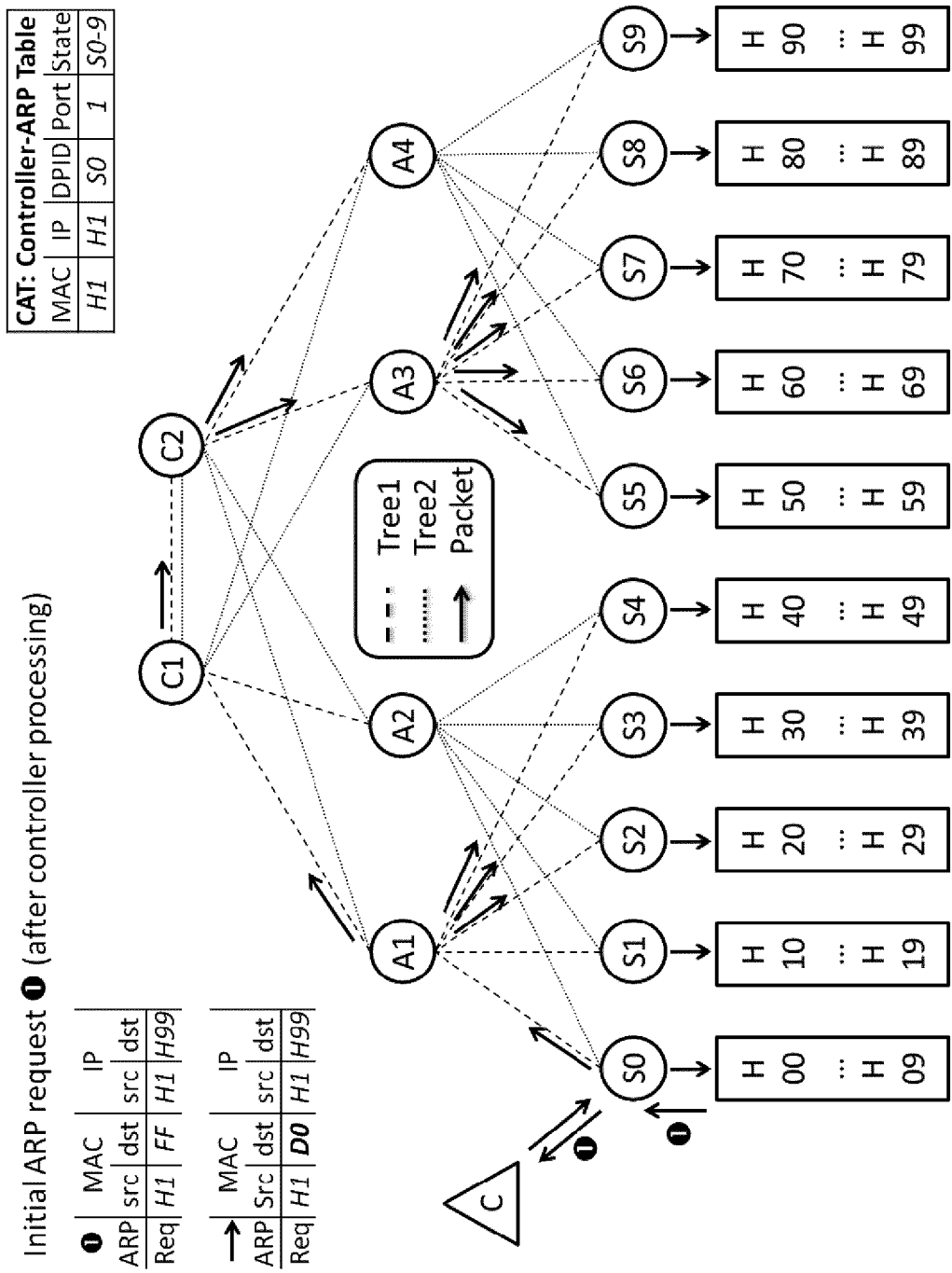


Fig. 3

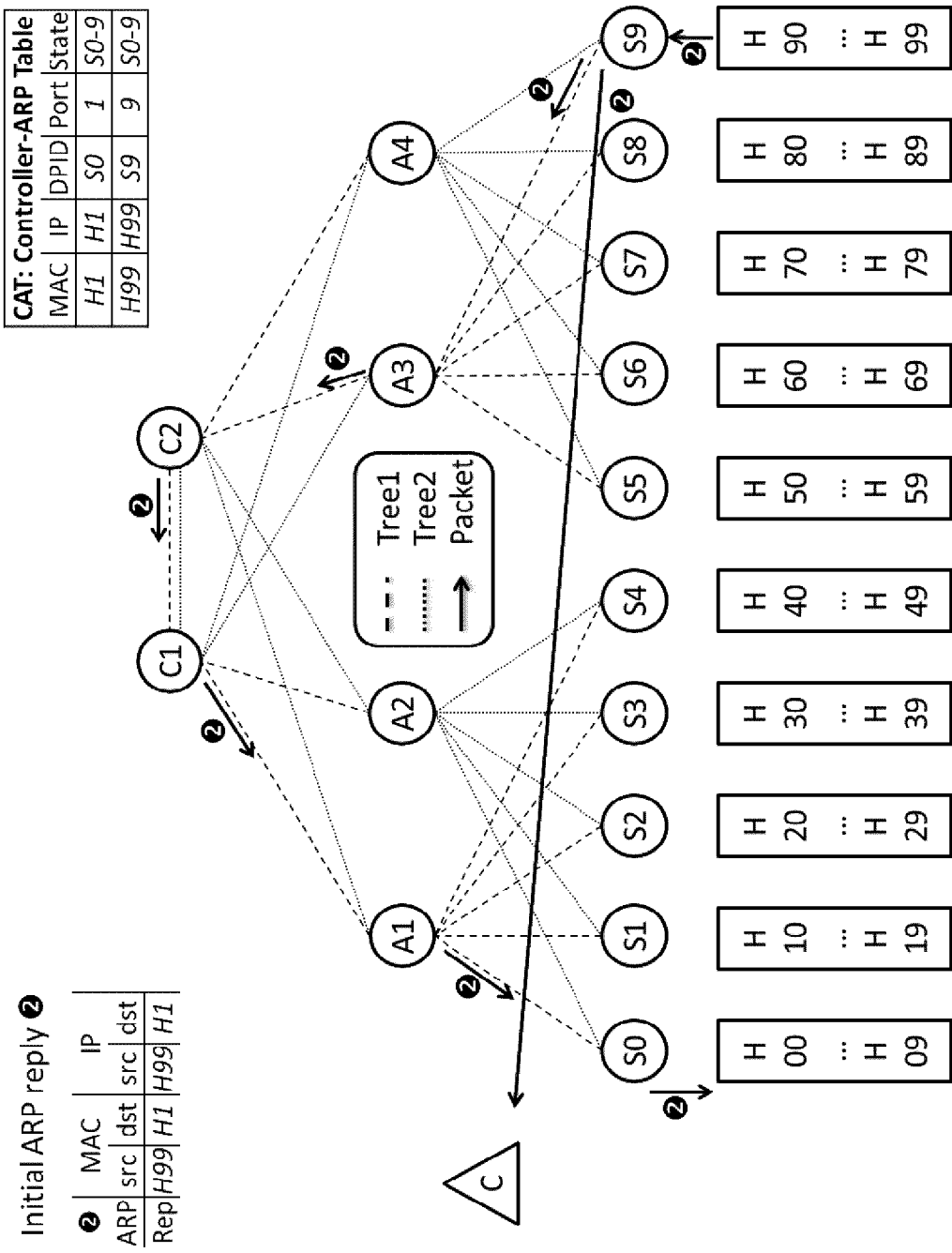


Fig. 4

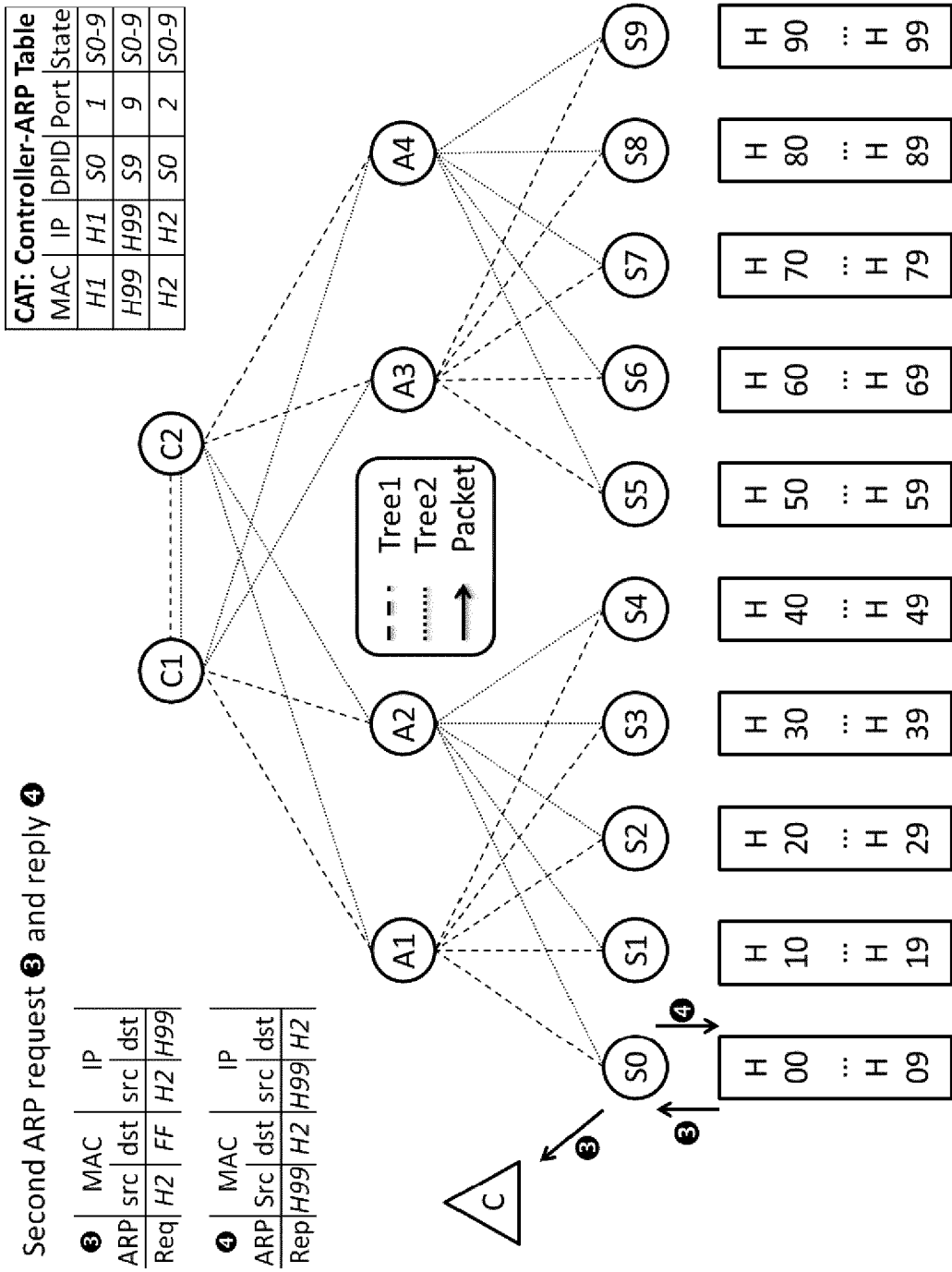


Fig. 5

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2016/065242

A. CLASSIFICATION OF SUBJECT MATTER

INV. H04L12/715 H04L29/12
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

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)

EP0-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2015/109478 A1 (HUAWEI TECH CO LTD [CN]) 30 July 2015 (2015-07-30)	11, 15
A	the whole document	1-10, 12-14
E	& EP 3 091 713 A1 (HUAWEI TECH CO LTD [CN]) 9 November 2016 (2016-11-09) paragraphs [0042] - [0053], [0083] - [0088], [0110] - [0120]; figures 1A, 2, 3	1-15
A	US 2013/250958 A1 (WATANABE YOSHIKAZU [JP]) 26 September 2013 (2013-09-26) paragraphs [0173] - [0204]; figures 13-15	1-15
A	US 2015/071289 A1 (SHIN JI SOO [KR] ET AL) 12 March 2015 (2015-03-12) paragraphs [0016], [0017], [0049] - [0080]	1-15



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

10 January 2017

Date of mailing of the international search report

20/01/2017

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

Engmann, Steffen

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2016/065242

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2015109478 A1	30-07-2015	CN 105379228 A EP 3091713 A1 US 2016330167 A1 WO 2015109478 A1	02-03-2016 09-11-2016 10-11-2016 30-07-2015
US 2013250958 A1	26-09-2013	JP 5874726 B2 JP 2014505379 A US 2013250958 A1 WO 2012093429 A1	02-03-2016 27-02-2014 26-09-2013 12-07-2012
US 2015071289 A1	12-03-2015	KR 20150030099 A US 2015071289 A1	19-03-2015 12-03-2015