



(51) International Patent Classification:
H04L 29/06 (2006.01)

(21) International Application Number:
PCT/EP2016/074496

(22) International Filing Date:
12 October 2016 (12.10.2016)

(25) Filing Language: English

(26) Publication Language: English

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

(72) Inventors: BIFULCO, Roberto; Bergheimerstraße 38,
69115 Heidelberg (DE). SCHNEIDER, Fabian; Langer
Anger 71, 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, DJ, 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, KW,
KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK,
MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA,
PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD,
SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT,
TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every
kind of regional protection available): ARIPO (BW, GH,
GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ,
UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ,

(54) Title: METHOD AND SYSTEM FOR ACCELERATION OF TCP CONNECTION ESTABLISHMENT

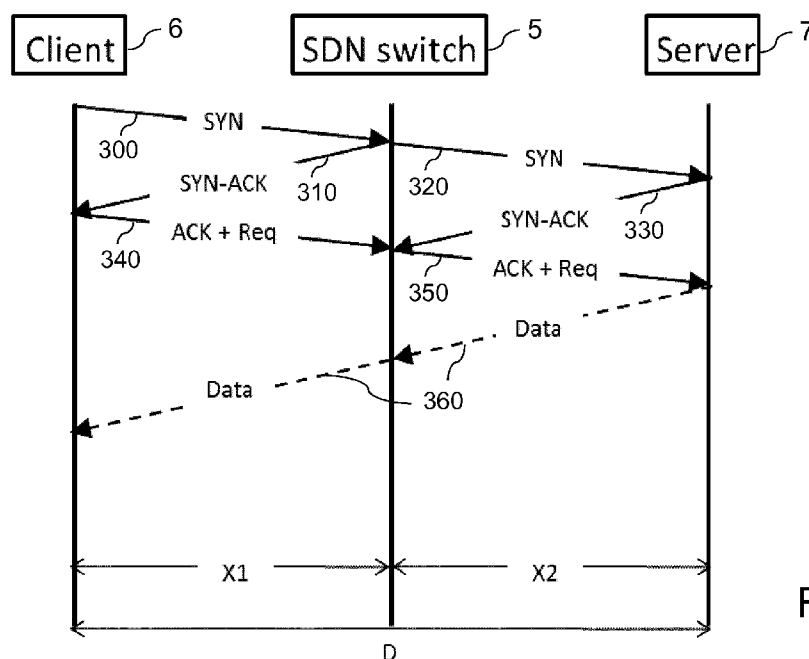


Fig. 5

(57) Abstract: The invention relates to a method and system for acceleration of TCP connection establishment between a client and a server in a network, comprising at least one stateful SDN switch (5) with packet generation capabilities, including input/output ports (8a, 8b), a state table (1), a finite state machine, FSM, table (2), and a feedback loop (9) implemented between said state table (1) and said FSM table (2), and an SDN controller (4) for programming rules into said state table (1) and said FSM table (2) of said at least one stateful SDN switch (5). The at least one stateful SDN switch (5) is configured to answer a TCP SYN segment received from the client (6) with a corresponding SYN ACK segment containing a sequence number generated by said at least one stateful SDN switch (5) on behalf of the server (7) and to forward said TCP SYN segment received from the client (6) to the server (7).

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))*

METHOD AND SYSTEM FOR ACCELERATION OF TCP CONNECTION ESTABLISHMENT

The present invention relates to a method and a system for acceleration of TCP connection establishment between a client and a server in a network.

Short TCP flows, i.e. flows that terminate a few round trip times (RTT) after establishment of the TCP session, are the prevalent flows in today's web services. For these TCP flows, the time required for TCP session establishment, also called three-way handshake, can take a considerable part of the overall connection time, as described in Sivasankar Radhakrishnan, Yuchung Cheng, Jerry Chu, Arvind Jain, and Barath Raghavan: "TCP Fast Open", in Proceedings of the 7th International Conference on emerging Networking EXperiments and Technologies (CoNEXT '11), ACM, New York, NY, USA, Article 21. Furthermore, the time required for TCP session establishment is dependent on the RTT between the two parties involved in the communication (generally referred to as client and server), since it requires three TCP segments to travel between the two parties, in any case.

In the state of the art, the problem has been addressed by deploying a TCP proxy close to the TCP client of the connection (i.e., the initiator of the TCP connection), as described in Abhinav Pathak, Y. Angela Wang, Cheng Huang, Albert Greenberg, Y. Charlie Hu, Randy Kern, Jin Li, and Keith W. Ross: "Measuring and evaluating TCP splitting for cloud services", in Proceedings of the 11th International Conference on Passive and Aactive Measurement (PAM '10), Arvind Krishnamurthy and Bernhard Plattner (Eds.). Springer-Verlag, Berlin, Heidelberg, 41-50. The client establishes a new connection with the proxy. In principle, such establishment is a fast operation that guarantees short network delays due to the proximity of the client to the proxy. However, the solution comes along with the disadvantage that the proxy has a permanent, proactively established connection towards the TCP connection's destination server.

A second way of achieving a reduction of the time required for TCP connection establishment is by deploying a TCP proxy that provides early replies to the TCP

segments sent for the connection establishment. If the proxy is on the path between the client and the server, ideally in the middle of the path delay-wise, then the connection is effectively accelerated. This technique requires the proxy to reply to TCP SYN and TCP SYN ACK segments as soon as they are received. Also, the proxy should forward the TCP SYN as soon as possible to the TCP destination server. The details of the approach are described in Giuseppe Siracusano, Roberto Bifulco, Simon Kuenzer, Stefano Salsano, Nicola Blefari Melazzi, Felipe Huici: "On-the-Fly TCP Acceleration with Miniproxy", in ACM SIGCOMM HotMiddlebox, 2016, and in Sameer Ladiwala, Ramaswamy Ramaswamy, and Tilman Wolf: "Transparent TCP acceleration", in Comput. Commun. 32, Issue 4, March 2009, 691-702.

Another type of optimization can be instead applied to the case of TCP connection "re-establishment". In particular, when a connection is established a second time towards the same server, the client can include a TCP session cookie generated during the first connection establishment. Such cookie allows the client to perform a direct data sending to the server without performing a full three-way handshake. This approach is usually called TCP fast open (described in the above cited document "TCP Fast Open").

In view of the above it is an objective of the present invention to improve and further develop a method and a system for acceleration of TCP connection establishment between a client and a server in a network in such a way that, by employing means that are readily to implement and operable with low effort, a significant shortening of the time required for TCP connection establishment is achieved.

In accordance with the invention, the aforementioned objective is accomplished by a method for acceleration of TCP connection establishment between a client and a server in a network, wherein the method comprises deploying at least one stateful switch with packet generation capabilities in said network, and configuring said at least one stateful switch

- to receive a TCP SYN segment from the client,

- 3 -

- to generate a sequence number in a manner coordinated with the server,
- to answer said TCP SYN segment, on behalf of the server, with a corresponding SYN ACK segment containing said sequence number,
- to forward said TCP SYN segment received from the client to the server, and
- to act as a mere forwarding element for segments exchanged between the client and the server once a TCP connection is established such that no state relating to the TCP connection is held by said at least one stateful switch.

Furthermore, the above objective is accomplished by a system for acceleration of TCP connection establishment between a client and a server in a network, the system comprising:

at least one stateful switch with packet generation capabilities, including input/output ports, a state table, a finite state machine, FSM, table, and a feedback loop implemented between said state table and said FSM table, and

a controller for programming rules into said state table and said FSM table of said at least one stateful switch,

wherein said at least one stateful switch is configured

- to receive a TCP SYN segment from the client,
- to generate a sequence number in a manner coordinated with said server,
- to answer said TCP SYN segment, on behalf of the server, with a corresponding SYN ACK segment containing said sequence number,
- to forward said TCP SYN segment received from the client to the server, and
- to act as a mere forwarding element for segments exchanged between the client and the server once a TCP connection is established such that no state relating to the TCP connection is held by said at least one stateful switch.

According to the invention it has been recognized that a significant shortening of the time required for TCP connection establishment can be achieved by using intermediate stateful switches (e.g. stateful SDN switches) with in-switch packet generation capabilities that are configured to work as TCP handshake

accelerators. According to embodiments a state machine is configured for the handling of TCP's three-way handshake acceleration in the stateful switch. Specifically, these stateful switches are coordinated with TCP servers to perform early generation of SYN-ACK TCP segments. According to embodiments of the invention the method/system lets a switch answer a TCP SYN segment with the corresponding SYN ACK, while the SYN is also forwarded to the final TCP destination server. The switch performs then two parallel TCP connection establishments, one with the client and the other with the server. In this context it is important to note that in accordance with the present invention the stateful switch is not handling internally the state for the full TCP connection. Instead, it just stores the minimal state required for the execution of the three-way handshake. Therefore it is possible that, once the TCP connection establishment is performed, the switch is configured to act as a 'normal' switch, i.e. it becomes just an intermediate forwarding element of the segments exchanged between the TCP client and the TCP server.

Accordingly, the present invention accelerates TCP connection establishments without deploying full-fledged TCP proxies, but by using intermediate nodes which only keep state during the three-way handshake. Specifically, no state about the TCP connection is/has to be stored in the switch, and there is no involvement of the SDN switch in the TCP connection after the connection establishment. Furthermore, there is no need for buffers or for closing/opening TCP connections. According to embodiments the sequence number generation procedure is offloaded from a TCP server to a stateful (e.g. SDN) switch.

Embodiments of the invention relate to a method for the generation of a TCP SYN-ACK segment on behalf of a TCP server without proxying the TCP connection that comprising the steps of

- 1) Deploying one or more SDN switches in the network. These switches need to support stateful forwarding and in-switch packet generation.
- 2) Configuring the switches to generate SYN-ACK packet.
- 3) Configuring the TCP servers to use the TCP sequence number generated by the switch.

Embodiments of the invention can be part of sophisticated SDN solutions, providing enhancements of both the SDN switches and the controller. In general, the invention can be used also in Telecom operators' deployments as an advance TCP acceleration service. In this sense, it may be part of a next generation SDN-enhanced TMS.

According to an embodiment the stateful switch, in order to answer a TCP SYN segment received from the client with a corresponding SYN ACK segment, may perform the steps of copying the TCP SYN segment received from the client in a new segment, including the generated sequence (acknowledgment) number into the new segment, and forwarding said new segment to the client. According to embodiments of the invention the switch may generate the SYN ACK segment according to the principles of in-switch packet generations.

According to an embodiment the stateful switch may be configured to communicate the generated sequence number to the server. For instance, in order to communicate the sequence number to the TCP server, the stateful switch may include the sequence number into a custom additional TCP option that is added to the TCP SYN segment received from the client, before forwarding the TCP SYN segment to the server. Alternatively, the stateful switch can encode the sequence number in a packet's header field.

An alternative way of how the server can gain knowledge of the sequence number generated by the stateful switch can be that the coordination between the server and the stateful switch includes a mechanism in which the server agrees with the stateful switch in a common scheme to generate the sequence number. In other words, the stateful switch generates the sequence number in cooperation with the server. For instance, in a straightforward implementation, the server and the stateful switch may agree, in advance, on a list of sequence numbers that are employed according to a specified order.

As already mentioned above, the stateful switch is in charge of generating the sequence number on behalf of the TCP server (and the TCP server has knowledge of this generated sequence number, either by virtue of the agreed generation scheme or by virtue of explicit information from the switch). Therefore, according to an embodiment, it may be provided that the server should use the sequence number generated by the stateful switch for the newly initiated TCP connection, i.e. for generating the server's SYN ACK segment. Typically, this embodiment requires agreement between the TCP servers and the stateful network switches that perform the acceleration.

Alternatively, the server may be allowed to use an arbitrary sequence number for generating the server's SYN ACK segment. In this case, it may be provided that the SDN switch is configured to perform a translation between sequence numbers (i.e. between this arbitrary sequence number and the sequence number generated by the SDN switch on behalf of the server) for all remaining segments exchanged by the client and the server in the respective TCP connection.

With respect to the sequence number generation by the SDN switch several different implementations may be realized. For instance, according to one embodiment the SDN switch may generate the sequence number by copying the sequence number from the TCP SYN segment received from the client. This implementation comes along with the advantage that the generated sequence number does not have to be communicated to the server separately since, in fact, the server can read it directly from the SYN packet.

According to alternative embodiments the sequence number may be derived from a counter stored in the SDN switch, or it may be generated by an arbitrary algorithm the SDN switch executes. According to still another embodiment the sequence number may be determined by the packet template used for the generation. As will be easily appreciated by those skilled in the art, the above listing is not exhaustive, and further alternative ways of how the SDN switch generates the sequence number may be envisioned. In any case, however, the switch either communicates the generated sequence number to the server, or the server and

the switch have agreed, in advance, on the respectively applied sequence number generation mechanism (such that the server is enabled to derive the sequence number from this generation mechanism).

The TCP server whose connection establishment procedure has been accelerated should be aware that the connection establishment underwent such process. In fact, for any received SYN from a TCP client, if the SYN triggered a SYN-ACK generation by any of the SDN stateful switches in the path, then it is advantageous that the server uses the sequence number generated by such switch. As mentioned earlier, the sequence number generation and communication-to-server procedures can happen in any of the previously described ways.

In case only a subset of the new TCP connection establishments received by the server has been actually accelerated, the server needs to be able to distinguish between TCP SYN segments for which a new sequence number has to be generated and for which this is not the case (since the new sequence number has already been generated by any of SDN stateful switches on the path between the TCP client and the TCP server. According to an embodiment this can straightforwardly happen by tagging the TCP SYN segment in case acceleration was provided, i.e. the SDN switch introduces a tag into the TCP SYN segment received from the client before forwarding them to the server. To this end, any known tagging technique may be employed.

Similar to the generation of the sequence number on behalf of the TCP server, the switch may use additional appropriate information for the SYN-ACK packet generation. Accordingly, according to an embodiment, the SDN controller may be configured to program the stateful SDN switch with respect to the admission or permission of TCP options for the SYN ACK packet generation. For instance, the programming may relate to adding (or not adding) the Kind 4 TCP option for selective acknowledgements permitted (as specified in IETF RFC 2018: "TCP Selective Acknowledgment Options", October 1996). This TCP option is only allowed in SYN packets and indicates that the sender is able to process Selective Acknowledgements. Whether to add the option can be programmed by the SDN

controller depending on the requested server, or probed/learned by the switch from previous SYN-ACKs.

Additionally or alternatively, the programming may relate to adding (or not adding) the Kind 3 TCP option for window scaling (as specified in IETF RFC 7323: "TCP Extensions for High Performance", September 2014). This TCP option is required to have correct window scaling information in the SYN-ACK, i.e. matching the TCP Servers window scaling factor, as otherwise reliable TCP transfers are not ensured. Again whether to add the option can be programmed by the SDN controller depending on the requested server, or probed/learned by the switch from previous SYN-ACKs.

According to still another embodiment the programming may relate to adding (or not adding) the Kind 2 TCP option for maximum segment size announcements (as specified in IETF RFC 793: "TRANSMISSION CONTROL PROTOCOL – DARPA INTERNET PROGRAM PROTOCOL SPECIFICATION", September 1981). This option is used to inform the other side of the of the maximum TCP segment size that is expected to work well for the connection. This option is not crucial to have. Yet, again whether to add the option can be programmed by the SDN controller depending on the requested server, or probed/learned by the switch from previous SYN-ACKs.

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 dependent patent claims on the one hand and to the following explanation of preferred embodiments of the invention by way of example, illustrated by the drawing on the other hand. In connection with the explanation of the preferred embodiments of the invention by the aid of the drawing, generally preferred embodiments and further developments of the teaching will be explained. In the drawing

Fig. 1 is a schematic view illustrating a two-phase port knocking application FSM,

- Fig. 2 is a schematic view illustrating a stateful switching architecture implementing the port knocking application of Fig. 1,
- Fig. 3 is a schematic view illustrating TCP connection establishment acceleration using a TCP proxy according to prior art,
- Fig. 4 is a schematic view illustrating TCP connection establishment acceleration in accordance with an embodiment of the present invention,
- Fig. 5 is a diagram illustrating TCP connection establishment acceleration with an SDN switch that replies to TCP SYN segments in accordance with embodiments of the present invention, and
- Fig. 6 is a schematic view illustrating an implementation of a TCP connection establishment acceleration proxy using an SDN stateful switch in accordance with embodiments of the present invention.

Since embodiments of the present invention described hereinafter in detail rely on the concepts of Software-Defined Networking (SDN), OpenFlow and OpenState, for ease of understanding first some essential aspects of these concepts will be briefly summarized, while it is generally assumed that those skilled in the art are sufficiently familiar with the respective technologies.

The Software-Defined Networking (SDN) paradigm (as specified in https://www.opennetworking.org/images/stories/downloads/sdn-resources/technical-reports/TR_SDN-ARCH-Overview-1.1-11112014.02.pdf) brings a separation of packet forwarding (data plane) and the control plane functions.

Being widely adopted, the OpenFlow protocol provides flow-level abstractions for remote programming of a switch's data plane from a centralized controller. A controller instructs an underlying switch with per-flow rules by means of specific *Flow_Mod* messages. Such message contains match and action parts with the first

specifying the packet headers to match and with the second applying a particular processing decision to all the packets belonging to the specified flow. These forwarding rules are translated into forwarding table statements and become installed into one or several forwarding tables of a table pipeline (for reference, cf. <https://www.opennetworking.org/technical-communities/areas/specification>)

Such forwarding abstraction implies stateless packet processing in the table pipeline, meaning that any possible state transitions of any particular flow may be implemented only on the controller's side. This means that any forwarding logic that needs stateful packet processing requires several interactions between the switch and the controller. Examples of stateful processing applications which may involve such interactions are: traffic anomalies detection, firewall-like primitives (e.g. reflexive ACL), adaptive QoS enforcement, etc.

OpenState (as described, e.g., in Giuseppe Bianchi, Marco Bonola, Antonio Capone, and Carmelo Cascone: "OpenState: programming platform-independent stateful openflow applications inside the switch", in SIGCOMM Comput. Commun. Rev. 44, 2 (April 2014), 44-51) enables an extension to flow-oriented stateless processing which enhances forwarding engines with an eXtended Finite State Machine (XFSM). This extension allows to track the state of each flow in the switching pipeline and to perform transitions of this state by adding to the stateless forwarding architecture two tables: state table and XFSM table. Figs. 1 and 2 illustrate this approach.

Fig. 1 schematically illustrates the concept of a stateful SDN (Software Defined Network) switch that provides an extension to flow-oriented stateless processing by means of implementing an FSM (Finite State Machine) table in addition to the state table. This extension allows to track the state of each flow in the switching pipeline and to perform transitions of this state.

Specifically, Fig. 1 represents an example of a simple stateful packet processing application known as "Port knocking" (for reference, see Giuseppe Bianchi, Marco Bonola, Antonio Capone, and Carmelo Cascone: "OpenState: Programming

Platform-independent Stateful OpenFlow Applications Inside the Switch”, in SIGCOMM Comput. Commun. Rev. 44, 2 (April 2014), 44-51). This application opens a transport port in the system for a particular remote host (e.g. TCP port for an SSH connection) after incoming flow’s packets were destined to a predefined port sequence. If the port sequence matches a predefined one in the FSM (which in the illustrated example is the sequence of TCP destination ports = [P1, P2]), the connection moves from an established state (i.e. A/Drop) to another established state (i.e. B/Fwd), i.e. all the subsequent packets of the same connection are forwarded.

In detail, if the match M on port P1 fails, the respective packet is dropped and the system remains in state A. If this match is successful, the respective packet is still dropped, but the system performs a switch to state B. Once in state B, if a match on port P2 is successful, the system remains in state B and all subsequent packets of the respective connection are forwarded. As soon as the match on port P2 fails the respective packet is dropped, and the system returns to state A.

To compile the described logic, the switching pipeline needs to be modified as depicted in Fig. 2. Here, the state table 1 is the first of the two tables of the architecture. In this state table 1, which is a key-value first-match-lookup table, a first-match lookup is performed to match some packet header(s) of an incoming flow against the state associated with this flow. The records in the table 1 are ordered e.g. by means of OpenFlow priority attributes. For instance, in the port knocking example of Fig. 1, such a field may match against the source IP address of every remote host which initiates the connection. The lookup returns a value, which is a state associated with the incoming flow (host) or a default state value associated with the wildcard match (*MATCH_ALL* in OpenFlow terms). This state is then passed together with the corresponding packet headers to the second stage – the FSM table 2. The state transfer between the tables 1, 2 of a switching pipeline may be performed by using metadata fields.

In the FSM table 2, a combination of packet headers and the transferred state is a key for the next lookup. In this context it should be noted that match fields in the

FSM table 2 (denoted 'Match*' in Fig. 2, as well as in all subsequent figures) are generally independent from the match field(s) in the state table 1 (denoted 'Match'). In the port knocking example, the match field in the FSM table 2 is the TCP destination port. Depending on the combination of the state and the Match* fields, actions for flow processing and its state update (if needed) are taken, which are in accordance with the FSM described in connection with Fig. 1. If a flow's state update is needed, a feedback loop becomes involved between the two tables 1, 2 passing a *set_state: pkt_headers+next_state* message with an update for a particular flow. A key extractor module 3 derives a particular value(s) for the match field(s) of the state table 1 from the packet headers and passes it together with *next_state* to the state table 1, thus either updating an existing record or installing a new one. The initial programming of both tables 1, 2 happens from the centralized controller 4 by means of *State_mod* and *Flow_mod* messages, as indicated by the dotted lines in Fig. 2. In this regard it should be noted that the state table 1 requires in this example only a single *State_mod* message with a default match/state decision (shown at the bottom of the state table), while all the subsequent updates happen through the feedback loop.

Fig. 3 is a schematic view illustrating TCP connection establishment acceleration using a TCP proxy 10 according to the prior art. The TCP proxy 10 performs the three-way handshake initiated by the client 5, then, it initiates a three-way handshake with the server 7. In both cases, the TCP proxy 10 creates state associated with the connection, such as TCP's finite state machine tracking, timers, buffers, etc. Such state is represented with the box "TCP Session State" 11 in Fig. 3. Notice that the TCP Session State 11 is kept for the entire duration of the established TCP session.

Fig. 4 is a schematic view illustrating TCP connection establishment acceleration using an embodiment of the present invention. Here, the intermediary box 12, e.g., a stateful SDN switch 5, performs the three-way handshake initiated by the client 6 and in parallel performs also a three-way handshake with the server 7. Given the described method, the intermediary 12 keeps the state only required for the conclusion of the three-way handshake with the client 6 and the server 7. This

TCP three-way handshake state is denoted by reference number 13 in Fig. 4. Such state is minimal and mainly includes the TCP's finite state machine tracking for the connection establishment part. That is, it monitors the handling of the SYN, SYNACK, ACK TCP frames used by the three-way handshake, but e.g., no buffers are created and/or maintained. Notice that once the TCP session establishment is concluded, there is no state maintained in the intermediary 12 for the established TCP session, independently from the session duration.

Turning now to Fig. 5, this figure illustrates a message flow diagram of a TCP connection establishment accelerated with a SDN switch 5 in accordance with an embodiment of the present invention. The TCP connection is established between a client 6 and a server 7 with the SDN switch 5 in between, communicating both with the client 6 and the server 7.

In the illustrated embodiment the SDN switch 5 is assumed to be located such that the delay between the client 6 and the SDN switch 5 (denoted delay 'X1' in Fig. 5) is approximately of the same size as the delay between the SDN switch 5 and the server 7 (denoted delay 'X2'), i.e. $X1 \approx X2 \approx D/2$, where 'D' denotes the overall delay between the client 6 and the server 7. While this is the most favorable situation in terms of the amount of acceleration that can be achieved, the method can be applied likewise in situations with $X1 \neq X2$ (for instance, due to the fact that the SDN switch 5 is topologically closer to either the client 6 or the server 7), as will be easily appreciated by those skilled in the art.

Essentially, in the embodiment of Fig. 5 the SDN switch 5, which is a stateful SDN switch with packet generation capabilities, is configured to work as TCP connection establishment accelerator. The basic principle of the acceleration is to let the SDN switch 5 answer a TCP SYN segment with the corresponding SYN ACK, while the SYN is also forwarded to the final TCP destination server 7, as shown in Fig. 5. The SDN switch 5 performs then two parallel TCP connection establishments, one with the client 5 and the other with the server 7. Once the establishment is performed, the SDN switch 5 acts as a normal switch becoming just an intermediate forwarding element between the client 6 and the server 7.

In detail the procedure illustrated in Fig. 5 unfolds as follows:

To initiate the TCP connection, as shown at step 300, the client 6 sends a TCP SYN segment.

Upon reception of this TCP SYN segment, the SDN switch 5 performs the following actions:

It generates a sequence (acknowledgement) number for the server side, as will be described in more detail below in connection with Fig. 6. Further, as shown at step 310, it responds to the client 6 with a TCP SYN ACK which includes the generated sequence number as acknowledgement number. Further, as shown at step 320, the SDN switch 5 forwards a TCP SYN to the server using the client's 6 sequence number (i.e. the sequence number contained in the client's 6 TCP SYN segment). Finally, the SDN switch 5 informs the server 7 about the generated sequence number the server 7 should use for the newly initiated TCP connection.

Next, as shown at step 330, the server 7 sends a SYN-ACK. According to an embodiment of the present invention, this step may be implemented as an optional step, i.e. it can be omitted. In cases, in which the server 7 is configured to send the SYN-ACK, the SDN switch 5 may simply drop this SYN-ACK.

Upon reception of the TCP SYN ACK from the SDN switch 5, the client 6 sends the ACK (and the first data segment), as shown at step 340. It is noted that in the illustrated embodiment the ACK is assumed to contain as first data segment a request for content, the message being correspondingly denoted 'ACK + Req'.

The SDN switch 5 forwards the ACK from the client 6 and any other TCP segment sent by the client 6 to the server 7, as illustrated at step 350.

With the exchange of the TCP ACK segment, the TCP handshake is completed, i.e. the TCP connection is successfully established. From now on, the SDN switch

5 becomes just a forwarder of the segments between the client 6 and the server 7, as exemplarily shown at step 360 for one data segment.

Fig. 6, in which like numerals denote like components/elements as in the previous Fig. 2, is a schematic view illustrating an OpenState based SDN stateful switch 5 configured for deployment as a TCP connection establishment accelerator in accordance with embodiments of the present invention. The general assembly of this switch 5 (i.e. being composed of input/output ports 8a/b, a state table 1 and a FSM table 2, with the two tables 1, 2 being programmable by a controller 4, and a feedback loop 9 with key extractor module 3 implemented between the two tables 1, 2) is the same as that of the switch described above in connection with Fig. 2. However, the configuration/programming of the switch 5 is different, as will be described in detail hereinafter.

Fig. 6 is a snapshot showing the SDN switch 5 at a point in time where two flows, denoted Flow1 and Flow2, are assumed to be currently processed by the switch 5. Hence, these flows have corresponding flow table entries, FTEs, in the state table 1, with Flow1 having 'SYNACK' as its currently associated state and with Flow2 having 'ESTB' as its currently associated state. By default, the state table 1 contains another record, which has a wildcard match and its associated state is 'DEF'. As will become apparent from the following description, the state 'SYNACK' is generally used to handle flows during the TCP handshake procedure, the state 'ESTB' (referring to 'established') is generally used to handle flows after the TCP handshake procedure has been successfully terminated and the TCP connection is thus established, and the state 'DEF' (referring to 'default') is generally used to handle any new flows.

When a new flow, which has no associated fine-grained record in the state table 1, is received at the SDN switch 5, it will be handled by the default record and gets associated the state 'DEF'. Consequently, when transferred to the FSM table 2, the flow will be handled by the first entry of the FSM table 2, which contains a wildcard match. This is appropriate when it is assumed that only TCP SYN segments are directed to the SDN switch 5. In a configuration, however, where the

SDN switch 5 may also receive other packets, the respective FSM table 2 entry may be designed to specifically match on SYN packets.

The first entry of the FSM table 2 indicates in its action field '*Set SYNACK, gen SYN-ACK, fwd*'. Accordingly, in case of the reception of a new TCP SYN segment from a client 6 at the SDN switch 5, the following set of actions will be executed by the SDN switch 5:

First, by executing the '*Set SYNACK*' action, the state for the corresponding flow will be switched from 'DEF' to 'SYNACK'. This happens by means of the feedback loop between the two tables 1, 2 passing a *set_state: pkt_headers+next_state* message and via key extractor module 3. As a result, a new record will be installed in state table 1 that associates the corresponding flow the state 'SYNACK'. Furthermore, the '*gen SYN-ACK*' will be executed, which corresponds to an in-switch packet generation action and which will be described in detail below. Finally, execution of the action '*fwd*' means that the TCP SYN segment is forwarded to the server 7, corresponding to step 320 of Fig. 5.

Turning now to the '*gen SYN-ACK*' action, this action comprises the following operations:

First, the client's 6 TCP SYN segment is copied into a new segment S. Then, a sequence number for the server-side TCP connection is generated, and this generated sequence number is included into the new segment S. Finally, the new segment S is forwarded to the client 6.

Essentially, the '*gen SYN-ACK*' action corresponds to an in-switch packet generation (as described e.g. in Roberto Bifulco, Julien Boite, Mathieu Bouet, and Fabian Schneider: "Improving SDN with InSPired Switches", in Proceedings of the Symposium on SDN Research (SOSR '16). ACM, New York, NY, USA, Article 11). In this context, an In-Switch Packet generation (InSP) Application Programming Interface (API) allows the SDN controller 4 to program the autonomous generation of packets in the switch 5. The packet generation operation can be specified by

providing three pieces of information: the trigger, the content and the actions. The trigger tells the switch 5 when a packet should be generated; the content specifies what are the packet's header and payload; the actions specify how the switch 5 should use the packet. An OpenFlow-based implementation of the API leverages the OpenFlow abstractions such as Flow Tables and Flow Table Entries (FTEs), and defines new ones for supporting the specification of the trigger, content and actions information. In particular, it defines a Packet Template Table to store the content of the packets that will be generated by the switch 5. Each Packet Template Table Entry (PTE) specifies the content of one packet and has a unique identifier that is used as a reference in other parts of the API. Second, the implementation adds a new OpenFlow instruction, the InSP instruction, that specifies the actions using standard OpenFlow actions. Finally, the trigger is provided by defining a FTE that contains the InSP instruction. In fact, the InSP instruction contains also a PTE's identifier that points to the corresponding PTE. Whenever a packet is matched by the FTE, the InSP instruction is triggered and the pointed PTE is used to generate the packet to which the instruction's actions are applied.

When the next packet of the respective flow arrives at the SDN switch 5, this packet will be associated with the state 'SYNACK' and will thus be handled by either the second or the third entry of FSM table 2. In case this packet is the SYNACK segment from the server 7 (cf. step 330 of Fig. 5), the second entry of FSM table 2 applies and the corresponding action is to simply drop this packet, as described above in connection with Fig. 5. On the other hand, if the next packet is the ACK segment sent by the client 6 (cf. step 340 of Fig. 5), the wildcard match of the third entry of FSM table 2 matches on this packet. Consequently, the corresponding action would be '*Set ESTB*', i.e. to switch the state of the flow from 'SYNACK' to 'ESTB' (since after the reception of the ACK segment the TCP connection is established) and ' *fwd* ', i.e. to forward the ACK segment to the server 7, corresponding to step 350 of Fig. 5.

All subsequent packets, upon arriving at the SDN switch 5, will be associated with the state 'ESTB' at the state table 1 and will therefore be handled by the fourth

and/or fifth entry of FSM table 2. The fifth entry reflects step 360 of Fig. 5, i.e. the SDN switch 5 simply functions as an intermediate forwarding element between the client 6 and the switch 7. The fourth entry considers the situation in which the ACK from the client 6 is received at the SDN switch 5 before the SYNACK from the server 7 is received at the SDN switch 5. The entry ensures that the SYNACK from the server 7 is excluded from the general forwarding of packets to the client 6 and that this packet is dropped instead.

Many modifications and other embodiments of the invention set forth herein will come to mind 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 specific terms are employed herein, they are used in a generic and descriptive sense only and not for purposes of limitation.

Claims

1. Method for acceleration of TCP connection establishment between a client (6) and a server (7) in a network, comprising:
deploying at least one stateful switch (5) with packet generation capabilities in said network, and
configuring said at least one stateful switch (5)
 - to receive a TCP SYN segment from the client (6),
 - to generate a sequence number in a manner coordinated with the server (7),
 - to answer said TCP SYN segment, on behalf of the server (7), with a corresponding SYN ACK segment containing said sequence number,
 - to forward said TCP SYN segment received from the client (6) to the server (7), and
 - to act as a mere forwarding element for segments exchanged between the client (6) and the server (7) once a TCP connection is established such that no state relating to the TCP connection is held by said at least one stateful switch (5).
2. Method according to claim 1, wherein said at least one stateful switch (5), in order to answer a TCP SYN segment received from the client (6) with a corresponding SYN ACK segment, performs the steps of
copying the TCP SYN segment received from the client (6) in a new segment, including said generated sequence number into said new segment, and forwarding said new segment to the client (6).
3. Method according to claim 1 or 2, wherein the server (7) gains knowledge of the sequence number generated by said at least one stateful switch (5) either by being informed by said at least one stateful switch (5) about said generated sequence number or by agreeing with said at least one switch (5) in a common scheme to generate the sequence number.

4. Method according to any of claims 1 to 3, wherein the server (7) uses the sequence number generated by said at least one stateful switch (5) for generating the server's (7) SYN ACK segment.
5. Method according to any of claims 1 to 3, wherein the server (7) uses an arbitrary sequence number for generating the server's (7) SYN ACK segment, and wherein said at least one stateful switch (5) performs a translation between said arbitrary sequence number and said sequence number generated by said at least one stateful switch (5) on behalf of the server (7).
6. Method according to any of claims 1 to 5, wherein said at least one stateful switch (5) generates said sequence number by copying the sequence number from the TCP SYN segment received from the client (6).
7. Method according to any of claims 1 to 5, wherein said at least one stateful switch (5) generates said sequence number by deriving it from a counter stored in said at least one stateful switch (5).
8. Method according to any of claims 1 to 5, wherein said at least one stateful switch (5) generates said sequence number by executing an arbitrary algorithm, or wherein said sequence number is determined by the packet template used for the generation.
9. Method according to any of claims 1 to 8, wherein said at least one stateful switch (5) introduces a tag into TCP SYN segments received from the client (6) before forwarding them to the server (7).
10. Method according to any of claims 1 to 9, wherein said at least one stateful switch (5) is a stateful SDN switch.
11. System for acceleration of TCP connection establishment between a client (6) and a server (7) in a network, preferably for carrying out a method according to any of claims 1 to 10, said system comprising:

- 21 -

at least one stateful switch (5) with packet generation capabilities, including input/output ports (8a, 8b), a state table (1), a finite state machine, FSM, table (2), and a feedback loop (9) implemented between said state table (1) and said FSM table (2), and

a controller (4) for programming rules into said state table (1) and said FSM table (2) of said at least one stateful switch (5),

wherein said at least one stateful switch (5) is configured

- to receive a TCP SYN segment from the client (6),
- to generate a sequence number in a manner coordinated with said server (7),
- to answer said TCP SYN segment, on behalf of the server (7), with a corresponding SYN ACK segment containing said sequence number,
- to forward said TCP SYN segment received from the client (6) to the server (7), and
- to act as a mere forwarding element for segments exchanged between the client (6) and the server (7) once a TCP connection is established such that no state relating to the TCP connection is held by said at least one stateful switch (5).

12. System according to claim 11, wherein said at least one stateful switch (5), in order to answer a TCP SYN segment received from the client (6) with a corresponding SYN ACK segment, is configured to perform the steps of

copying the TCP SYN segment received from the client (6) in a new segment, including said generated sequence number into said new segment, and forwarding said new segment to the client (6).

13. System according to claim 11 or 12, wherein said controller (4) is configured to program said at least one stateful switch (5) with respect to the admission or permission of TCP options for the SYN ACK packet generation, including at least one of the TCP options for selective acknowledgment, window scaling, or maximum segment size announcements.

14. Flow-based programmable network device, in particular stateful switch (5), configured for being employed in a method and/or a system according to any of claims 1 to 13.

15. Controller entity, in particular SDN controller (4), configured for being employed in a method and/or in a system according to any of claims 1 to 13.

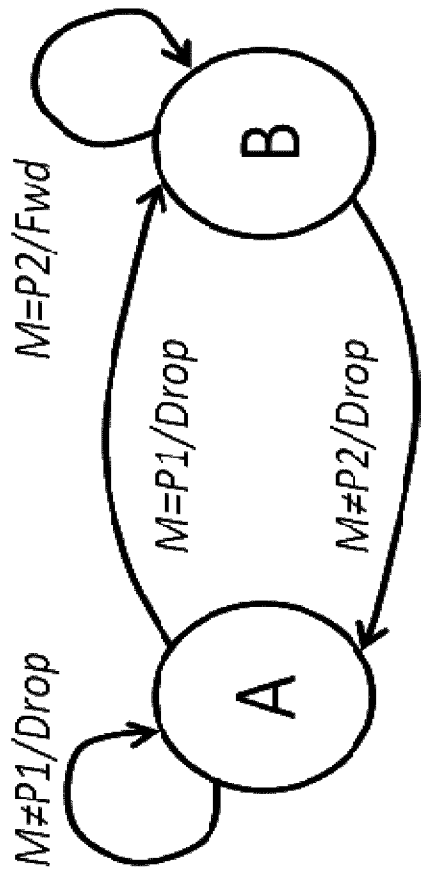


Fig. 1

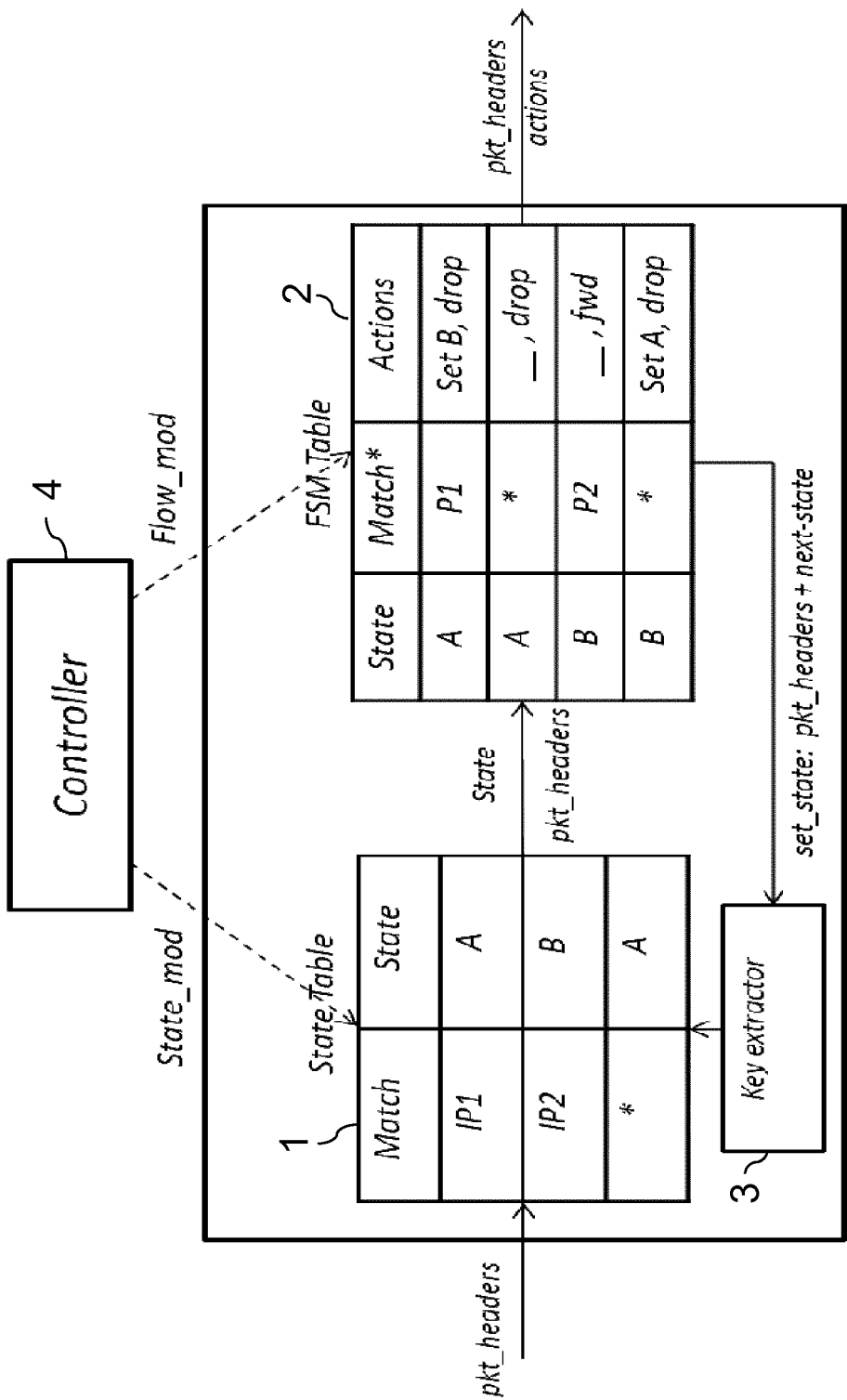


Fig. 2

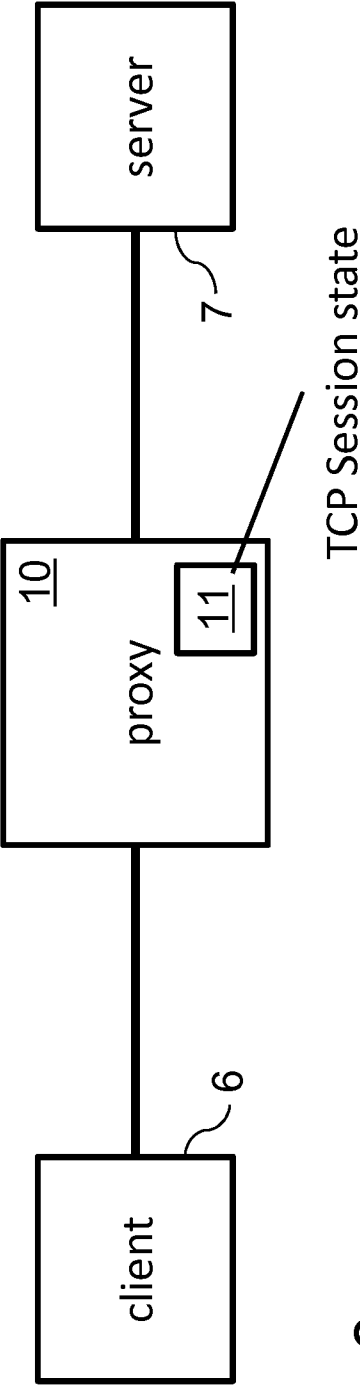


Fig. 3

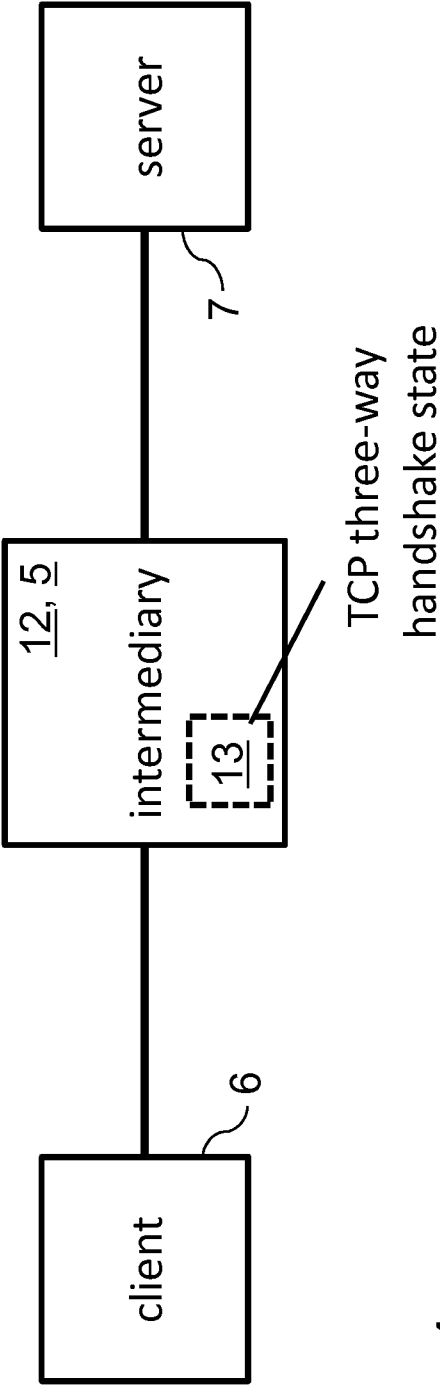


Fig. 4

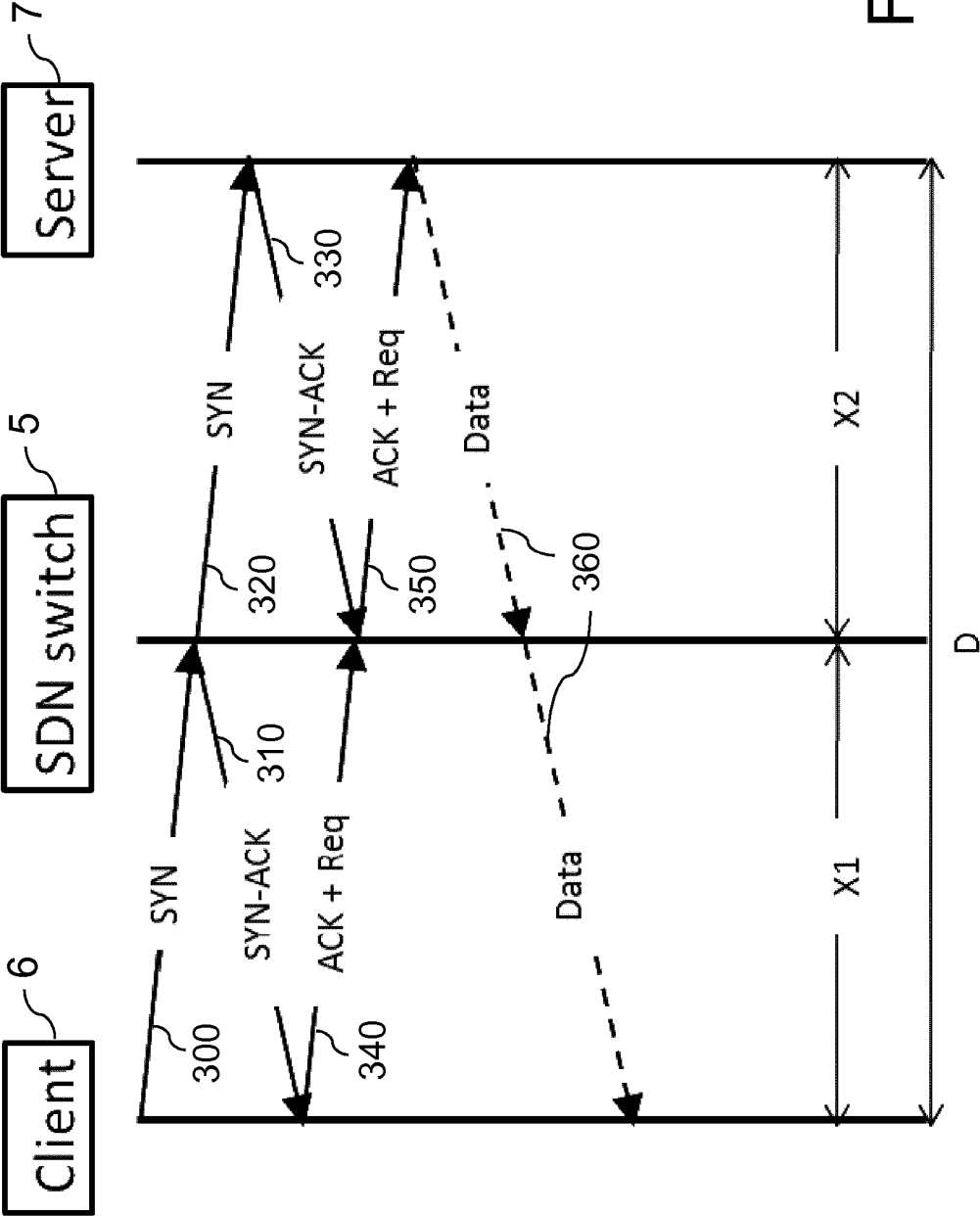


Fig. 5

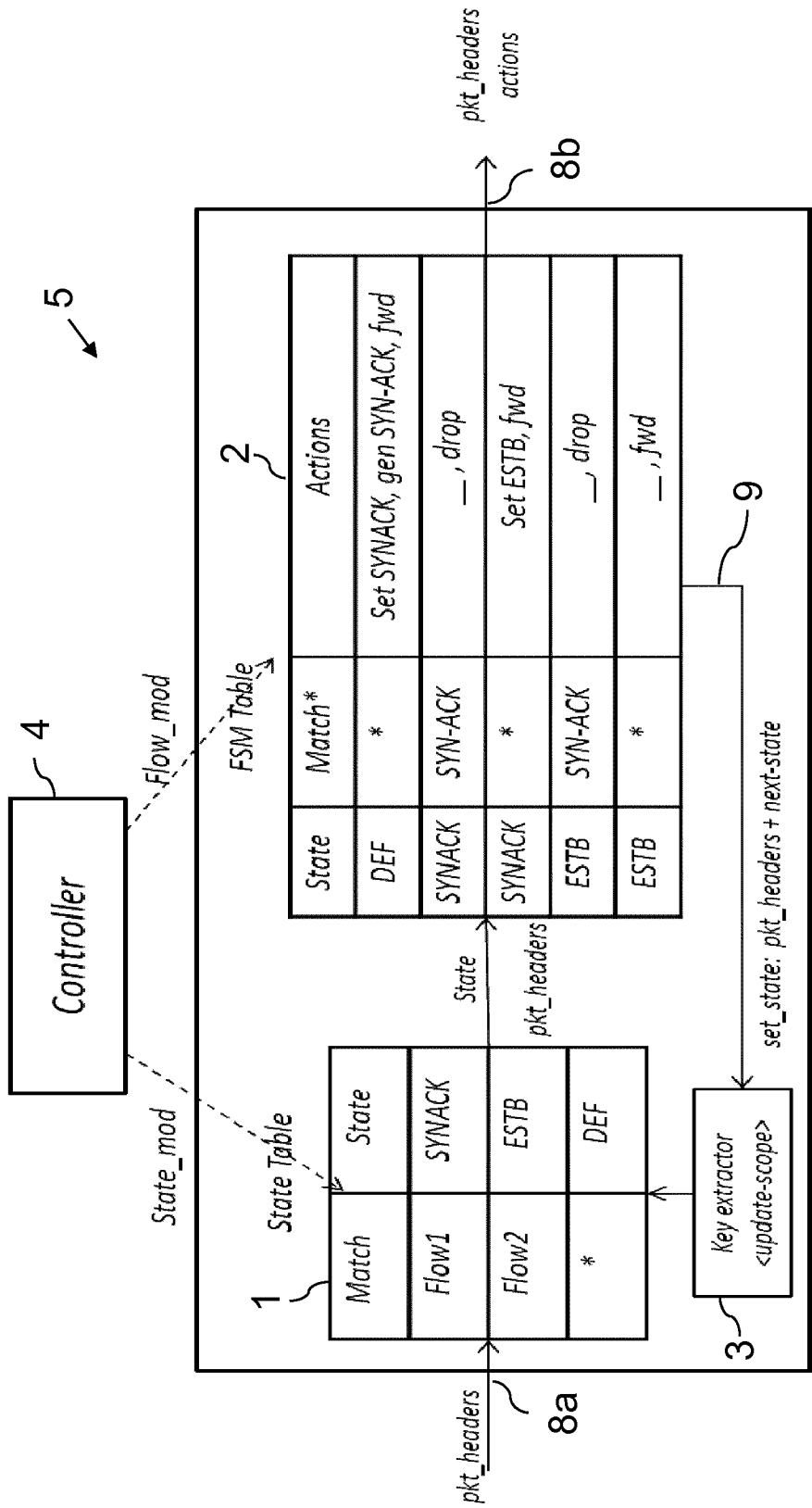


Fig. 6

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2016/074496

A. CLASSIFICATION OF SUBJECT MATTER
INV. H04L29/06
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)

EPO-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	US 8 489 670 B1 (FLETCHER GREGORY [US] ET AL) 16 July 2013 (2013-07-16) column 4, line 42 - column 7, line 45 column 10, line 11 - column 12, line 65 -----	1-15
X	GIUSEPPE SIRACUSANO ET AL: "On-the-Fly TCP Acceleration with Miniproxy", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 20 May 2016 (2016-05-20), XP080702448, page 4 -----	1-15
X	US 8 180 902 B1 (DAY MARK STUART [US] ET AL) 15 May 2012 (2012-05-15) column 22, line 43 - column 25, line 18 ----- -/--	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

31 May 2017

Date of mailing of the international search report

13/06/2017

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

Dupuis, Hervé

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2016/074496

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 103 491 065 A (ZTE CORP) 1 January 2014 (2014-01-01) pages 1-60 -----	1-15
A	US 2016/197823 A1 (SREERAMOJU AKSHAY KUMAR [US]) 7 July 2016 (2016-07-07) paragraph [0034]; figure 4 -----	1-15
A	JP 2014 003459 A (HITACHI LTD) 9 January 2014 (2014-01-09) paragraph [0043]; figure 6 -----	1-15
A	US 2007/280273 A1 (YUN WON-DONG [KR] ET AL) 6 December 2007 (2007-12-06) paragraph [0011]; figure 2 -----	1-15
A	WO 2015/131929 A1 (HUAWEI TECH CO LTD [CN]; PORAT HAYIM [DE]) 11 September 2015 (2015-09-11) page 21, line 23 - page 22, line 7 -----	1-15
X	US 2016/218960 A1 (SUNDARABABU PREMNATH [IN] ET AL) 28 July 2016 (2016-07-28) paragraph [0038]; figure 2 -----	1-15
X	ABDELMONIEM AHMED M ET AL: "Incast-Aware Switch-Assisted TCP Congestion Control for Data Centers", 2015 IEEE GLOBAL COMMUNICATIONS CONFERENCE (GLOBECOM), IEEE, 6 December 2015 (2015-12-06), pages 1-6, XP032872733, DOI: 10.1109/GLOCOM.2014.7417567 [retrieved on 2016-02-23] paragraph [03.B] -----	1-15
X	SEUNGWON SHIN ET AL: "AVANT-GUARD", COMPUTER & COMMUNICATIONS SECURITY, ACM, 2 PENN PLAZA, SUITE 701 NEW YORK NY 10121-0701 USA, 4 November 2013 (2013-11-04), pages 413-424, XP058034060, DOI: 10.1145/2508859.2516684 ISBN: 978-1-4503-2477-9 paragraph [5.2.1] -----	1-15

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2016/074496

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 8489670	B1	16-07-2013	US 8489670 B1 16-07-2013
		US 2013304809 A1	14-11-2013
		US 2013304892 A1	14-11-2013
		US 2015381771 A1	31-12-2015
US 8180902	B1	15-05-2012	US 8180902 B1 15-05-2012
		US 8688844 B1	01-04-2014
CN 103491065	A	01-01-2014	NONE
US 2016197823	A1	07-07-2016	NONE
JP 2014003459	A	09-01-2014	JP 2014003459 A 09-01-2014
		US 2013336213 A1	19-12-2013
US 2007280273	A1	06-12-2007	BR PI0711617 A2 06-12-2011
		CN 101449489 A	03-06-2009
		EP 2022192 A1	11-02-2009
		JP 2009539297 A	12-11-2009
		US 2007280273 A1	06-12-2007
		WO 2007139286 A1	06-12-2007
WO 2015131929	A1	11-09-2015	CN 106063202 A 26-10-2016
		EP 3047613 A1	27-07-2016
		US 2016373349 A1	22-12-2016
		WO 2015131929 A1	11-09-2015
US 2016218960	A1	28-07-2016	NONE