POLYTECHNIC OF TURIN - STOCKHOLM'S KTH

Faculty of Engineering

Master of Science in Computer Engineering

Master Thesis

MPTCP Security Evaluation

Analysing and fixing critical MPTCP vulnerabilities, contributing to the Linux kernel implementation of the protocol

Advisors:
prof. Antonio Lioy
prof. Peter SjÃűdin

Candidate:
Fabrizio Demaria

Company tutors
Intel Corporation
Henrik Svensson
Joakim Nordell
Shujuan Chen

March 2016

Acknowledgements

Thanks to...

Summary

Abstract goes here...

Contents

Chapter 1

Introduction

1.1 Motivation

The last few decades have seen the most pronounced technology evolution in history, in many different research areas and consumer markets: from robotics to smartphones, from medicine to cars, etc. One of the pillars upon which all these advancements have been made possible is the Internet, or more generally the entire set of networking technologies that allow software to communicate.

The process towards interconnected devices saw a big leap forward in the early 1960s with the first research into packet switching as an alternative to the old circuit switching. But it is 1982 the year of standardization for the TCP/IP protocol suite, which permitted the expansion of interconnected networks [wiki]. The Internet grew rapidly, passing from a few tens of million users in the 1990s to almost 3 billions users in 2014 [ref]. Even more astonishing is the number of networked devices and connections globally, around 14 billion in 2014 [ref].

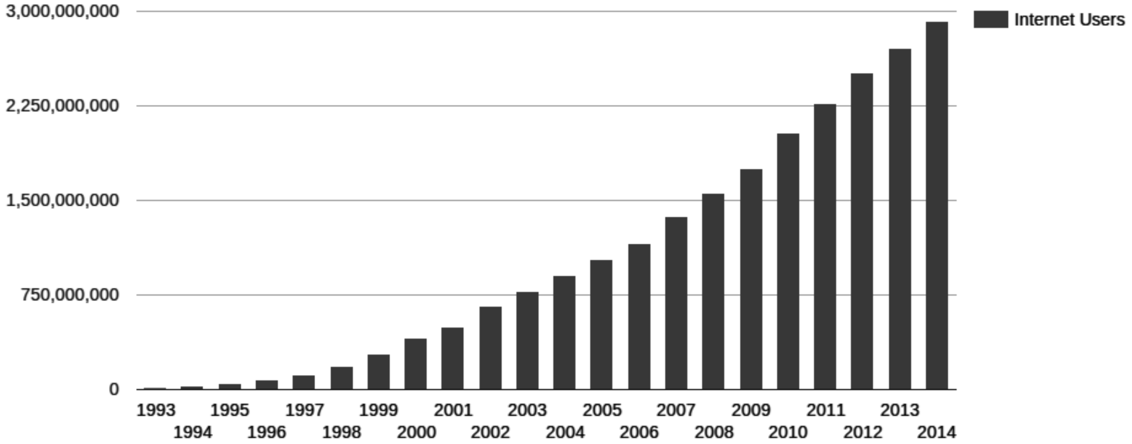

Figure 1.1: The expansion of the Internet

"Networks" is a very generic term. In the IT context, a computer network is set of connected nodes adopting common protocols to exchange data. The most widespread protocol for networking communication is the above-mentioned TCP/IP protocol, that is used in the vast majority of services like the World Wide Web, email, file transfer, remote system access, etc. It is also often used as a communication protocol in private networks and datacenters. The reason for its wide adoption is not the fact that there aren't good alternatives: TCP/IP is not to most performing protocol in every network environment, but it is fairly simple and it introduces a relatively low complexity in the overall architecture, still meeting all the basic security and reliability requirements. Back in the 1980s, TCP/IP was the simpler way for applications to use most networks, and eventually it was chosen as the protocol for the Internet, thus quickly becoming a de-facto standard [ref].

During its life, the TCP/IP protocol suite have seen updates and additional components to reach the desired levels of network congestion, traffic load balancing, handling of unpredictable behaviors, security, user-experience and so on. Such aspects became more and more challenging with the uncontrollable expansion of the Internet. Albeit, after all these years the core components of the TCP/IP protocol design haven't changed at all, mainly for retro-compatibility reasons. This inevitably causes some aspects of the old protocol to look very limited in the current networking reality. A striking example is the scarcity of available IPv4 addresses: when TCP/IP was designed in the early stages, a 32-bit number seemed to be a very high number to encompass all the users of the network. Nevertheless, due to the unexpected increase rate in the number of Internet users (and also due to inefficient IP allocation policies), the available IPv4 addresses run out quickly, forcing the introduction of the lengthy 126-bit address format, known as IPv6, formalized in 1998. IPv6 is intended to replace IPv4, but the transition to the new format turned out to be a remarkably complicated procedure overall: IPv6 is not designed to be directly interoperable with IPv4, and even if nowadays the majority of the systems are IPv6-compatible, it took about 20 years to reach the current percentage of overall adoption: 10% [percentage of IPv6 users accessing Google ref]. This should give an idea of the big challenge that is modifying the core design aspects of the TCP/IP architecture; such issue is a recurrent topic in this paper.

When the TCP protocol was first developed in the 1970s, it was certainly difficult to predict the rate of growth of the networks all around the globe, not only in terms of the number of nodes involved, but also in terms of the quantity and type of the transmitted data, the increasing need of low latency for new streaming applications, the advancement in the hardware adopted to carry the data and the computing power of the interconnected devices. Today we can count billions of interconnected devices, and we have just started the era of the IoT (Internet of Things) which aims at giving communication capabilities to virtually every object commonly used in our daily life. As a result of this process, the networks are becoming more complex and devices often use multiple interfaces to stay connected. Common appliances like smartphones provide both cellular connectivity and Wi-Fi modules (figure 1.2); same technologies can be often found in tablets; laptops have at least Wi-Fi capabilities plus an Ethernet port, and they support third-party receivers for connectivity through cellular networks. The argumentation is much more complex in the backend infrastructures' scenario, which is rapidly evolving due to a new interest in BigData storage and analysis, as well as the flourishing of wide-scale low-latency streaming services (video streaming, VOIP, multiplayer videogames, etc.). Datacenters often count tens of thousands of interconnected nodes, including content-delivery servers that are capable of handling a huge number of network interfaces simultaneously.

The implications of this new reality include the possibility of establishing multiple paths to transmit data between two applications running on the communicating hosts, since they are now often equipped with multiple network interfaces, each configured with an active IP address. Back in 1970s TCP was designed to create a virtual connection

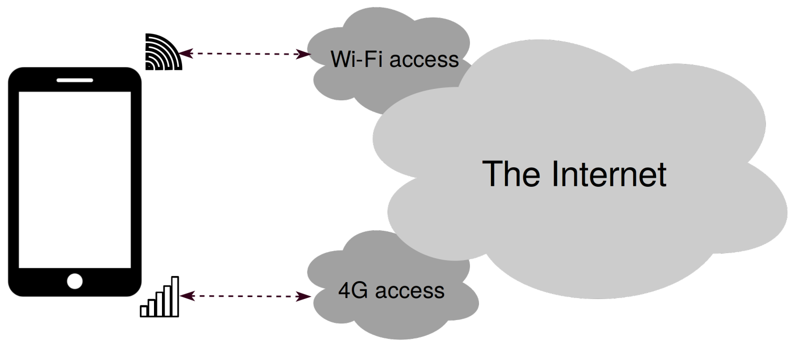

Figure 1.2: The smartphone connectivity

between exactly two IP addresses and two port values, with almost no flexibility or dynamism in address/port addition and/or removal within the duration of the connection. In the multipath reality of the infrastructures of today, to old point-to-point singlepath connection provided by TCP looks quite limiting. This led to various projects aiming at exploiting the multipath concept, and Multipath TCP is one of them.

Multipath TCP (MPTCP) is an ongoing project managed by the Internet Engineering Task Force (IETF), whose specifications have been published as Experimental standard in January 2013 [ref RFC-6824]; such protocol extends the current TCP to introduce multipathing capabilities, maintaining retro-compatibility at the end-points and undertaking a major endeavor to avoid disrupting of middleboxes' behavior. MPTPC can communicate with the application layer via standard TCP interface and it automatically splits data at the sender, it sends the data through different subflows (each being basically a regular TCP connection) according to the IP-addresses/interfaces available at the hosts and finally reassembles the data at the receiver, in fact enabling multipathing.

1.1.1 Benefits of MPTCP

Multipathing provides hosts with the resource pooling concept applied to networking access. Resource pooling allows dynamism and flexibility in requesting and handling resources and it is a positive trend in many services and architectures, like Content Delivery Networks (CDNs), Peer-to-Peer (P2P) networks, Cloud Computing, etc. The very concept of packet-switching, the core aspect of the modern Internet, is based on a resource pooling technique: circuit utilization is no more performed by allocating isolated channels in the link (static multiplexing) as it was the case with circuit switching, but the traffic is fragmented into small addressed packets that can share the overall link capacity (statistical multiplexing) [ref]. MPTCP aims at taking this concept to the next level, by grouping a set of separate links into a pool of links (figure 1.3).

The benefits include better resource utilization, better throughput and smoother reaction to failures, leading to an overall improved user experience, as shown in the following four major use-cases:

- Combining MPTCP multipath and multihoming (the connection to the Internet via multiple providers), it is possible to achieve higher throughput by exploiting multiple simultaneous connections to transfer different portions of the same piece of data. For example, a typical smartphone could use its cellular module and its Wi-Fi module

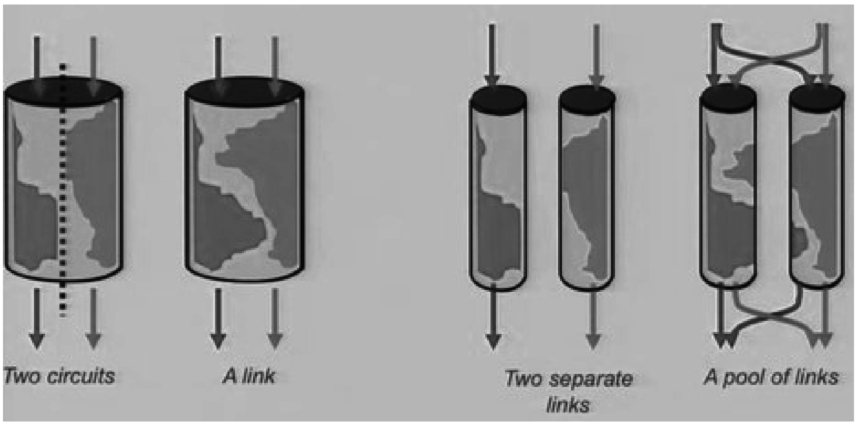

Figure 1.3: MPTCP pooling principle

simultaneously in downloading a file from a remote server, despite them having two different IP addresses;

- It is possible to introduce failure handover for the connection with no special mechanism at network or link layer. If one of the interfaces goes down or the flow of data gets interrupted for any reasons, data transfer can seamlessly continue through other interfaces;
- By assigning different priorities to the various flows, it is possible to better handle data transfer through the different interfaces; this could be useful if some connectivity modules drain more battery than others, or if any interfaces are associated to a limited-capacity data-plan. For example, consider the case of a file download on a smartphone via 4G connectivity: it would be advantageous to seamlessly switch the whole data transfer to the Wi-Fi interface if that becomes available in the middle of the download, starting from the point left by the cellular connection and without the need to restart the session;
- Providing multipath awareness to current network stacks can improve load balancing and exploitation of the network resources in datacenters; this is a valuable aspect, considering that the network performance in datacenters is usually critical for maintaining low latency of the overall system. A similar concept applies to load balancing in ISPs' network backbones.

1.1.2 Multipathing solutions

MPTCP aims at achieving all the benefits mentioned in the previous paragraph by operating at the transport layer of the traditional Internet architecture (figure 1.4). Before MPTCP, other proposals have been elaborated to achieve multipath benefits by introducing new technologies at the link layer, network layer and transport layer (the latter being the layer on which TCP operates). Even at the application layer developers can create custom frameworks on top of TCP to achieve benefits similar to those that would come by exploiting multipath natively at the lower layers. For example, most modern browsers open many TCP connection simultaneously to download the various elements of a Web Page to improve user experience. Another example could be Skype and similar VOIP services, which try to automatically reconnect hosts in case of problems with minimum impact on the user experience. Albeit all the solutions at the application layer are just clever workarounds on top of regular TCP and they fall only marginally in our discussion regarding multipath.

```
+-------------+                                     +-------------+
| Application |<------------ end-to-end ------------>| Application |
+-------------+                                     +-------------+
|  Transport  |<------------ end-to-end ------------>|  Transport  |
+-------------+   +-------------+   +-------------+   +-------------+
|   Network   |<->|   Network   |<->|   Network   |<->|   Network   |
+-------------+   +-------------+   +-------------+   +-------------+
   End Host           Router            Router          End Host
```

Figure 1.4: The traditional Internet architecture

The following list gives a general overview of the most important multipathing solutions other than MPTCP, grouped according to the architectural layer they operate in:

- *Link layer*: there are link aggregation techniques to combine the capacities of different interfaces to the same switch [add a ref]. There are different ways to achieve resource pooling through link aggregation, but the basic concept is to setup multiple interfaces with the same IP address (and usually the same MAC address) so that the link aggregation is transparent to the higher-level applications and then various algorithms can be used to redistribute the data packets to the various links. In order for this to work, proper configuration is needed at the host and at the next-hop switch. Despite being a common solution in ISPs' inner networks and datacenters to improve throughput and achieve higher network-access, end-users cannot directly take advantage of this technology.

- *Network layer*: there exist multiple solutions to better exploit multipathing at this layer, most notably *Mobile IP* and *Shim6*. Without going into the details, they both provide hot-handover capabilities with no interruption of the higher-level services, with some limitations: Mobile IP requires extensive support by the underlying infrastructure and Shim6 is an IPv6 only solution. More importantly, there is a fundamental problem in confining resource pooling at the network layer: TCP operates at the transport layer but it is closely related to the network layer because it statefully inspects various properties of the underlying network paths to provide performance optimizations (this is why referring to TCP often implies taking into consideration the whole TCP/IP protocol stack): in most cases, transparent modifications at the network layer would cause TCP malfunctioning.

- *Transport layer*: the most notable experiment in multipath exploitation prior to MPTCP is the Stream Control Transmission Protocol (SCTP). Such protocol is, in many ways, similar to MPTCP: the first version of SCTP provided fail-over capabilities by exploiting different interfaces, and successive versions introduced multistreaming capabilities to increase throughput. The major problem with SCTP is that it was thought to be an alternative, enhanced version of TCP, and the two protocol are indeed incompatible with each other. This means that a wide adoption of SCTP would require to upgrade the network to be SCTP aware. Moreover, all the applications would need to be upgraded to explicitly switch to the new protocol for communication. The vast global networking scenario of today, mainly based on TCP, makes these requirements virtually impossible to meet, and SCTP remains a technology of very limited adoption.

All these previous solutions didn't get widespread adoption. Link layers and network layers solutions require extensive modifications in the underlying network configurations in order to achieve the desired results; introducing a new multipath-aware protocol at the transport layer requires to change all the applications in order for them to communicate over the new protocol, thus allowing this solution in very limited scenarios; workarounds at the application layer, despite being quite effective, are far from the purpose of MPTCP.

MPTCP primary goal is to automatically introduce the multipath benefits to infrastructures and devices currently adopting TCP, with the minimum possible effort from users, developers, network maintainers. Engineers decided that the best way to achieve all these requirements was to still use TCP as fundamental block for communication, extending it to support multipath: the entire protocol design works by adding MPTCP custom options into regular TCP packets and each subflow in MPTCP is indeed seen by the lower infrastructure as a regular TCP connection.

MPTCP got a lot of attention in the Internet community in the last few years, and many consider MPTCP as a valuable step forward for the whole global network currently relying on TCP. The final goal of MPTCP is to replace the majority of the current TCP implementations, which is a very delicate process in which all the current TCP standards in terms of robustness, performance and security have to be maintained, if not improved. This paper is an evaluation of the security aspects of MPTCP, with an analysis of current threats and vulnerabilities affecting the protocol.

1.2 Problem statement

MPTCP is a big effort from the IETF working group to unlock multipath networking capabilities worldwide, with many subtle implications for current infrastructures. Hence the importance of evaluating the current security status of MPTCP, by inspecting its implications on external middleboxes and security equipment and also by analyzing internal design flaws that might allow attacks to the MPTCP sessions. The reference implementation for the new protocol is available for the Linux kernel and currently maintained in an off-tree open-source repository. The main focus of this paper is related to the main vulnerability currently known for the protocol, concerning the ADD_ADDR component. Such vulnerability is tested and studied; the solution for it is implemented and evaluated. In the process, patches for the Linux kernel implementation of the protocol have been developed to fix the vulnerability and mark the first step to towards the new version of MPTCP.

A comprehensive list of all the objectives for the thesis work is the following:

- Studying the security implications of adopting MPTCP on current infrastructures;
- Listing the known vulnerabilities affecting the current version of the protocol;
- Studying and exploiting the ADD_ADDR vulnerability of the protocol;
- Evaluating the possible solutions for the ADD_ADDR vulnerability;
- Assessing the best solution for the ADD_ADDR vulnerability and developing it for the Linux kernel implementation of MPTCP;
- Developing effective and powerful simulation scenarios in order to test MPTCP (and possibly other networking protocols);
- Contributing to the upstreaming of MPTCP into the Linux kernel by developing patches and contributing to official RFC documentation.

1.3 Methodology

The thesis work has been carried out at the Intel Corporation offices in Lund (Sweden). The process took six months in total, with a main focus on testing and developing. The entire work has been closely followed by major stakeholders in the MPTCP community, located in Sweden, Romania and the United States. Such cooperation involved patch reviewing and weekly meetings.

The workflow started with an overall study of MPTCP and how it interacts with the most common middleboxes. The next step was a more focused evaluation of the current threats for the protocol, mainly referencing to the document RFC-7430; within the document, only one vulnerability is considered a blocking issue in the advancement of MPTCP standardization, known as the ADD_ADDR vulnerability. The document also proposes a change in the protocol design that fixes the problem. With such background, the actual development stage of the work started. At first, it was necessary to sync with the development status by interacting with the official MPTCP mailing list for developers [ref]; this allowed to make sure that the ADD_ADDR solution proposed in RFC-7430 was indeed the preferred one and that nobody started developing a patch for it already. Before starting to work on the fix, a first stage of the work involved a deeper analysis of the ADD_ADDR vulnerability. A connection hijacking has been executed by exploiting such vulnerability in a testing environment. This allowed to better validate the criticality of the problem and it was a useful experiment to get acquainted with MPTCP. Moreover, it was a good way to setup a proper testing environment that was indeed used during the whole patch-development process that followed. After having reproduced the attack, it followed an analysis of the MPTCP source code within the Linux kernel in order to understand how the protocol implementation works inside the TCP stack. This step was required to get the required knowledge to start developing patches.

The entire code developed during the stage, around 400 additions, was eventually merged into the official MPTCP repository for the Linux kernel. Some additional contributions have been performed in order to improve RFC documentation about the protocol and to upgrade related networking tools to be compatible with the new version of MPTCP. The final evaluations and the writing of the report took place in the last two months of the working period.

1.3.1 Document structure

The structure of this paper mainly follows the workflow explained in the previous section. After the introductory first chapter, the discussion is mainly subdivided into two parts: first, an analysis about MPTCP background and working principles (chapters 2 and 3); second, a discussion about the original work on simulating and fixing the ADD_ADDR vulnerability (chapters 4 and 5):

- Chapter 2 starts with a broad explanation of the basic concepts of TCP to introduce how MPTCP has been developed on top of it. All the technical details of the new protocol can be found in this chapter. In this chapter it is also included an analysis on the MPTCP deployment status in the real world and the problematics associated in upstreaming the protocol (mainly incompatibilities with current middleboxes).
- Chapter 3 is again a background analysis on MPTCP, with a narrowed focus on its security aspects. The chapter includes a comprehensive threats analysis, with an

overview of the current security issues affecting the new protocol. An entire section is dedicated to the ADD_ADDR vulnerability. In such section all the details regarding the vulnerability are presented: how to exploit it to hijack an MPTCP connection and what are the requirements an attacker needs to execute the attack.

- Chapter 4 is the first part that introduces the original work carried out during the thesis period. Taking as reference the theory behind the ADD_ADDR attack explained in the previous chapter, this section explains the development of the script capable of exploiting the vulnerability in a simulated environment. The script code is explained step-by-step, as well as the entire procedure to setup the virtual machines to execute the attack. This entire chapter aims at validating the criticality of the ADD_ADDR vulnerability and in doing so it also provide setup guidelines for a powerful simulating environment that can be useful for future MPTCP testing and development.

- Chapter 5 contains the core part of the thesis work. It starts with a theoretical evaluation of the accepted fix for the ADD_ADDR vulnerability and goes on with its development for the Linux kernel implementation of MPTCP. All the issues encountered during the project, as well as the required side-feature that needed to be implemented for proper functioning, are reported in this chapter. The two last sections cover the remaining part of the work: the set of contributions not mentioned in the previous sections and a final evaluation of the performance of the produced patches.

- Chapter 6 is the conclusive part of the paper, where related work and proposals for future work are present, together with some final thoughts.

Chapter 2

Multipath TCP

2.1 Transmission Control Protocol (TCP)

MPTCP is an extension of regular TCP, the ubiquitous protocol for highly reliable host-to-host communication in a packet-switched computer network. A proper introduction of the fundamentals of TCP is due. TCP is a host-to-host communication protocol operating at a layer in between the application and the Internet Protocol. TCP abstracts all the details of the network connection to the application and it is used at the sender to split the application data stream into segments that can be efficiently routed through the network after being encapsulated into an IP packet. At the receiver, the segments are reassembled before being sent to the application layer.

The reasons why TCP became a de-facto standard in modern computer communication has been briefly mentioned in the introductory part of the paper. A more technical analysis shows that TCP maintains good levels of reliability for the connection independently from the lower layers it depends on for the raw transmission of bits. TCP is indeed able to handle possible data loss, data damaging, data duplication, out-of-order delivery of data. In order to do this, the data to be transmitted is split into a sequence of TCP segments, each containing an additional *TCP header* with the information needed to operate the protocol functionalities at the nodes. Such functionalities are [ref]:

- *Basic data transfer*: sending continuous stream of octets in each direction between its users, using as endpoints the 4-tuple: source IP address, source port, destination IP address, destination port. The IP address allows to route packets to the destination machine, while the port values direct the content of the packet to the right application within a host;

- *Reliability*: in-order, reliable data transfer is achieved by adding a sequence number to each transmitted octet and using ACK signals and timeouts to possibly trigger retransmission of lost packets. TCP assures that no transmission errors will affect the delivery of the data if the network is not completely partitioned;

- *Flow control*: the receiver can control the amount of data sent by the sender in a certain moment of the connection by returning a "window" value in the TCP header, so that it is possible to avoid buffer congestion;

- *Multiplexing*: a single host is allowed to use multiple independent TCP connections simultaneously thanks to the port value available in the protocol. This value, together with the host address assigned at the internet communication layer, forms a socket, that is the actual endpoint of a TCP connection;

- *Connections*: TCP initializes and maintains status information regarding each connection and the data stream between a pair of sockets in order to provide all its functionalities. Such status information is initialized during a first handshake procedure, and released only upon connection termination. TCP is indeed known as a virtual-connection protocol;

- *Precedence and Security*: these aspects refer to the possibility of prioritize connections and assign security properties to them. Both precedence and security can be configured by users, but default values are provided.

As noted above, all these features are made possible by processing the bits at the TCP header. Such structured set of information is mostly static and predefined (with the exception of the custom TCP Option field), so that at each position in the header corresponds a well known portion of the protocol data. The TCP header looks like the one in Figure 2.1.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          Source Port          |       Destination Port        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Sequence Number                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Acknowledgment Number                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Data |           |U|A|P|R|S|F|                               |
| Offset| Reserved  |R|C|S|S|Y|I|            Window             |
|       |           |G|K|H|T|N|N|                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           Checksum            |         Urgent Pointer        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Options                    |    Padding    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             data                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2.1: The TCP header format

Regarding the basic operation of regular TCP, a connection is divided into three steps: *connection establishment*, *data transfer* and *connection release*. Different fields in the TCP header are used for the different phases of the connection.

Connection enstablishment

During the connection establishment, a three-way handshake is performed between the client and the server: the client sends a SYN packet to the port the on which the server is listening; after that, the server answers with a SYN/ACK packet to acknowledge the connection request; as a third and final step, the client acknowledge the SYN/ACK packet by sending to the server an ACK packet (hence the name "three-way handshake"). In order to define the kind of TCP segment received, single-bit fields in the TCP header are used (SYN and ACK flags are shown in figure 2.1). The three-way handshake is important for various reasons: first of of all, both hosts declare their willingness to open the TCP connection using the addresses and ports indicated in the packets: the *Source Port* and *Destination Port*, together with the source and destination IP addresses provided in the upper IP header (not shown in figure 2.1), are the means for identifying the two endpoints of the TCP connection. **These fixed fields clearly shows the single-path fundamental design of the protocol**. Moreover, during the initial handshake both client and server declare the supported *Options* and the initial *Sequence Number* values to be used for both directions of the connection.

```
    TCP A                                                TCP B

1.  CLOSED                                               LISTEN

2.  SYN-SENT    --> <SEQ=100><CTL=SYN>               --> SYN-RECEIVED

3.  ESTABLISHED <-- <SEQ=300><ACK=101><CTL=SYN,ACK>  <-- SYN-RECEIVED

4.  ESTABLISHED --> <SEQ=101><ACK=301><CTL=ACK>       --> ESTABLISHED
```

Figure 2.2: Basic TCP three-way handshake procedure

Data transfer

TCP splits the data to be transferred into multiple packets that are independently routed in the network and it is possible that they arrive at destination unordered, or that some of them are lost on the way. TCP is a protocol that offers bidirectional communication, so that the active downloader (the receiver) can communicate to the sender information for data transfer control. *Sequence Number* and *Acknowledgment Number* are used to number each transferred octet, so that it is possible at the receiver to put them in sequence and acknowledge them in a cumulative way: by acknowledging sequence number X to the sender, the receiver is signalling that all packets up to but not including X have been received. This system, together with timeouts and sliding window mechanisms, also allows for retransmission of lost packets. There is a flag bits used to determine if a TCP segment is an ACK segment (also used during the initial handshake), meaning that the *Acknowledgment Number* field in the current packet indeed represents the next *Sequence Number* that the receiver is expecting. The *Window* field is used to indicate to the sender the range of sequence numbers that the receiver is prepared to accept for a particular moment of the connection. In this way, the receiver can tune the the data flow and slow it down if the application is slow at consuming data and buffers tend to fill up quickly. The *Checksum* field guarantees that data has not been modified on its way to the destination, intentionally or unintentionally.

Connection release

In normal cases, each participant terminates its end of the TCP connection by using a specific bit available in the TCP header: the FIN bit. The FIN message is indeed a way for a host to signal the request for connection termination, but such request has to be acknowledged with a ACK for both endpoints before reaching final termination. Connection termination differs from the three-way handshake mechanism used for for connection establishment, and it can be better described as a pair of two-way handshakes between client and server. There are also cases in which something goes wrong during the connection, compromising the correct functioning of the TCP protocol for data transfer. In these cases, the RST flag is used in a message to force abrupt closure of the connection.

```
TCP A                                                  TCP B

ESTABLISHED                                            ESTABLISHED

(Close)
FIN-WAIT-1  --> <SEQ=100><ACK=300><CTL=FIN,ACK>  --> CLOSE-WAIT

FIN-WAIT-2  <-- <SEQ=300><ACK=101><CTL=ACK>      <-- CLOSE-WAIT

                                                     (Close)
TIME-WAIT   <-- <SEQ=300><ACK=101><CTL=FIN,ACK>  <-- LAST-ACK

TIME-WAIT   --> <SEQ=101><ACK=301><CTL=ACK>      --> CLOSED
```

Figure 2.3: Normal TCP connection termination procedure

A component of the TCP header that is fundamental for MPTCP is the *Options* field, which was firstly introduced as a free space for future additions to the protocol. In this specific case, the TLV solution is adopted to process the data inside the field. "TLV" stands for *type-length-value*, where the type is the ID value uniquely identifying the option, the length is the number of bytes of the option, whereas the value represents the actual option content. This particular design allows to skip unknown options at the receiver by simply checking the length value and moving the pointer accordingly. An important limitation for this field is that its total length cannot be more than 40 bytes [ref].

2.2 MPTCP design

MPTCP *functional goals* are to increase resilience of the connectivity and efficiency of the resource usage by exploiting multiple paths (subflows) for the connection. Similar goals can be found in other multipathing solutions as the ones described in section 1.1.2, but what is really unique about MPTCP design is the set of its *compatibilities goals* [ref]:

- *Application compatibility* aims at instantiating a protocol that can be fully operational with no modifications for the applications using it. This means that the networking APIs and the overall service model of regular TCP has to be maintained with MPTCP; the entire MPTCP functioning is handled transparently by the underlying system. Such transparency must be maintained also in terms of throughput, resilience and security for the connection, that cannot be deteriorated with respect to the current TCP standards;

- *Network compatibility* is a goal similar to the previous one, since MPTCP is supposed to work out of the box with the current underlying network layer and the ones below it. The main reason still resides in the possibility of achieving a smooth wide-spread deployment of the protocol on the current infrastructure;

- *Users compatibility* is a corollary to both network and application compatibility, which states that MPTCP flows must be fair to regular TCP connection in case of shard bottlenecks. If MPTCP would adopt a congestion control that is the same of the one for regular TCP, each subflow would get the same amount of resources as a single, regular TCP connection. Specific MPTCP congestion-control schemes have been studied to avoid such problems [refs].

All these compatibilities requirements should explain the very fundamental decision of developing the new multipath protocol at the transport layer. Let's take into consideration the traditional TCP protocol stack and compare it to the new MPTCP stack (figure 2.4). To achieve the required compatibility goals, changes had to be applied to the layers lower than the application layer, so that current applications do not have to be upgraded to make use of multipath; on the other side, the new protocol had to be placed at layers above the network layer: the network layer operates within the network infrastructure, a segment of the overall networking architecture that shouldn't be modified for MPTCP deployment. The transport layer, right above the network layer, is indeed the first component operating at the end systems: in order to get the smoothest possible widespread transition towards MPTCP, the protocol is intended to be deployed as a simple upgrade of the end systems' OS, with no modifications applied to the network infrastructure.

```
                             +-------------------------------+
                             |           Application         |
+---------------+            +-------------------------------+
|  Application  |            |             MPTCP             |
+---------------+            + - - - - - - - + - - - - - - - +
|      TCP      |            | Subflow (TCP) | Subflow (TCP) |
+---------------+            +-------------------------------+
|      IP       |            |       IP      |       IP      |
+---------------+            +-------------------------------+
```

Figure 2.4: The TCP and MPTCP protocol stacks

The choice of working at the transport layer was indeed the only available option. Within that option, the choice of maintaining TCP as the fundamental operating protocol for MPTCP was still straightforward for similar compatibility reasons; for this very purpose engineers decided to add all the required options for MPTCP inside the TCP *Option* field in the TCP header. In this way, MPTCP-aware systems can process the MPTCP options for multipathing, but if a system that is not MPTCP-aware receives a MPTCP connection-request, it would simply discard the MPTCP options and treat such message as a plain TCP connection-request. MPTCP design maintains the behavior of the subflows to be compliant with regular TCP, while splitting data and send it through different paths as well as reassembling the same data at the receiver are processes executed by the end systems only. MPTCP subflows are indeed seen by middleboxes as regular and independent TCP connections carrying some additional options. If security policies at the middleboxes is not too restrictive against unknown options, MPTCP-unaware intermediate nodes would still be compatible with the new protocol. MPTCP was indeed designed to be as compatible as possible with all the most common middleboxes of the Internet of today. For what regards applications, they don't need to be changed either since MPTCP would be added into the network stack at the operating system level: MPTPC transparently splits the data buffered from the application layer and send it through different subflows, according to the number of available endpoints at the connected hosts. Communication with the application layer can be performed through the old TCP APIs, even if MPTCP specific options can be used by upgraded user applications to take advantage of more advanced options in MPTCP.

A functional decomposition of MPTCP brings up four core functions the protocol needs to function:

- *Path management*: MPTCP has to provide a mechanism to detect and use multiple paths between two hosts;

- *Packet scheduling*: MPTCP fragments the byte stream received from the application in order to transmit it through different subflows, adding the required sequenced mapping used to reconstruct the same byte stream at receiver. This function of MPTCP adopts the information from the path management component to exploit the different paths;
- *Subflow interface*: as mentioned multiple times, MPTCP uses TCP to send data in a single subflow;
- *Congestion control*: a congestion control mechanism at the MPTCP connection layer is needed to make sure that MPTCP wouldn't starve a regular TCP flow in a shared bottleneck. The congestion control component of MPTCP implements the algorithms used to decide how to schedule the various data segments (which paths and rate adopt).

All the MPTCP functions are implemented internally inside the specific operating system in use on the device, and they use a relatively compact set of TCP Options to operate between two hosts. Technically, there is only a single generic MPTCP option, to which has been assigned the value 30 as the TCP "Option-Kind" identifier; at a lower level it is possible to identify a set of eight MPTCP option subtypes, each identified by a 4-bit value (this classification, reported in figure 2.5, references to RFC-6824).

Value	Symbol	Name
0x0	MP_CAPABLE	Multipath Capable
0x1	MP_JOIN	Join Connection
0x2	DSS	Data Sequence Signal (Data ACK and data sequence mapping)
0x3	ADD_ADDR	Add Address
0x4	REMOVE_ADDR	Remove Address
0x5	MP_PRIO	Change Subflow Priority
0x6	MP_FAIL	Fallback
0x7	MP_FASTCLOSE	Fast Close

Figure 2.5: The set of MPTCP options [RFC-6824]

2.2.1 Control plane

The control plane for MPTCP takes into consideration all the options used in MPTCP to handle connection initiation, addition and removal of subflows, priority assignment to specific subflows, error handling via 'fallback' mechanism. These options are reported in the following subsections, adopting as reference documentation the RFC-6824.

MP_CAPABLE

The connection initiation of an MPTCP connection is very similar to the standard TCP initial three-way handshake, involving a SYN, SYN/ACK and ACK exchange on a single path between host A and host B. In a regular TCP connection establishment these three packets are used to guarantee that both hosts have received an acknowledgment of the connection and also to exchange the two random initial sequence numbers that will be used to acknowledge data delivery for the two directions of the connection. Despite working as regular TCP packets, if MPTCP is enabled the SYN packet from host A will have a MP_CAPABLE option in the *Options* field of the TCP header; if the receiver host B is not MPTCP-compatible it will simply discard the MP_CAPABLE option and proceeds instantiating a regular TCP connection. In case both hosts are MPTCP-compatible, the MP_CAPABLE option is inserted in the three packets of the initial handshake for two purposes: advertise that both hosts are indeed MPTCP-compatible and allow them to exchange two 64-bit keys (Key-A and Key-B), according to the scheme in figure 2.6.

```
Host A                                  Host B
------                                  ------
MP_CAPABLE             ->
[A's key, flags]
                       <-               MP_CAPABLE
                                        [B's key, flags]
ACK + MP_CAPABLE       ->
[A's key, B's key, flags]
```

Figure 2.6: MPTCP connection initiation

These keys are sent in clear inside the MP_CAPABLE option only during the initial handshake and their purpose is to identify a specific MPTCP connection within a host (useful when associating a new subflow to an existing MPTCP connection, for example) and to provide shared security material that is used in MPTCP for authorization mechanisms (more on this later in this section). The *Option* field in the TCP header can only be 40 bytes long, and it is not reserved to MPTCP only. For this reason it is of primary importance to keep the amount of MPTCP related metadata as low as possible. In fact, the original 64 bit keys are exchanged only during initial handshake; subsequently, shorter 32 bit tokens (Token-A and Token-B) derived from such keys will be used to address a specific MPTCP connection, even if this procedure requires additional checks in case of collisions with other tokens already assigned to other MPTCP connections in the same machine. There is another fundamental motivation for using this mechanism of shorter tokens: the full keys, that represent security material used in the protocol for authentication purposes (look at MP_JOIN and ADD_ADDR2 messages), are exposed only during connection setup in the MP_CAPABLE messages; sending the full keys each time a new subflow has to be started would diminish the overall security level of the protocol. Therefore, an implementation will require a mapping from each token to the corresponding connection, and in turn to the keys for the connection.

Regarding the hashing algorithm used to produce the tokens, this can be negotiated by using a portion of the flag bits inside the MP_CAPABLE option. In this paper, the SHA1 (and HMAC-SHA1 in case a key element is needed) is considered [ref to SHA1]. Note that the SHA1 algorithm produces a 160-bit / 20-octet resulting value, that is then truncated to its leftmost 32 or 64 bits according to the different cases in the MPTCP operations.

MP_JOIN

Suppose that after the first subflow is operational host A initiates a new subflow between one of its addresses and one of host B's addresses (figure 2.11). Host A sends a TCP SYN packet to host B containing the MP_JOIN option, which includes Token-B (the token

```
                     1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---------------+---------------+-------+-------+---------------+
|     Kind      |    Length     |Subtype|Version|A|B|C|D|E|F|G|H|
+---------------+---------------+-------+-------+---------------+
|                   Option Sender's Key (64 bits)               |
|                                                               |
|                                                               |
+---------------------------------------------------------------+
|                  Option Receiver's Key (64 bits)              |
|                     (if option Length == 20)                  |
|                                                               |
+---------------------------------------------------------------+
```

Figure 2.7: MP_CAPABLE option

derived B's key) and a nonce value used to prevent replay attacks. An additional field in the MP_JOIN option is called Address ID, an identifier for the original addresses in use within a single MPTCP connection; this additional value allows to refer to a specific subflow without the need to use the addresses as identifiers, which is very useful when middleboxes like NATs alter the IP header during the transit of the packets.

At the lower layers of the network, the SYN packet sent in this way looks like a legitimate request from host A to initiate a new TCP connection with host B, being the SYN packet the first of the regular TCP initial handshake. Host B treats such packet as a new MPTCP subflow request, and it uses the Token-B in the packet to associate the request to the specific ongoing MPTCP connection with host A.

The handshake flow for MP_JOIN includes HMAC values for authentication purposes, and it is structured as follow:

- Token-B is added in the SYN packet from host A to host B in order to address a specific MPTCP connection; a random nonce (R-A) is also sent along;

```
                     1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---------------+---------------+-------+-----+-+---------------+
|     Kind      |  Length = 12  |Subtype|     |B|   Address ID  |
+---------------+---------------+-------+-----+-+---------------+
|                   Receiver's Token (32 bits)                  |
+---------------------------------------------------------------+
|                Sender's Random Number (32 bits)               |
+---------------------------------------------------------------+
```

Figure 2.8: MP_JOIN option - SYN

- Host B processes the request and sends back a truncated HMAC value calculated by using as *key* the concatenation of Key-B followed by Key-A, and as the *message* the concatenation of a new nonce computed at host B (R-B) and the one received from host A (R-A). R-B is also added to the packet, since it is needed by host A in the next step;
- The last ACK form host A to host B only has to contain the HMAC calculated using as key the concatenation of Key-A and Key-B, and as message the concatenation R-A and R-B. This time, the HMAC value is sent in its full length of 160-bit.
- Note that the HMAC in the ACK packet from host A to host B has to be acknowledged for the subflow to be finally established. This because the third ACK is the

```
                     1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---------------+---------------+-------+-----+-+---------------+
|     Kind      |  Length = 16  |Subtype|     |B|   Address ID  |
+---------------+---------------+-------+-----+-+---------------+
|                                                               |
|                Sender's Truncated HMAC (64 bits)              |
|                                                               |
+---------------------------------------------------------------+
|                Sender's Random Number (32 bits)               |
+---------------------------------------------------------------+
```

Figure 2.9: MP_JOIN option - SYN/ACK

```
                     1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---------------+---------------+-------+-----------------------+
|     Kind      |  Length = 24  |Subtype|      (reserved)       |
+---------------+---------------+-------+-----------------------+
|                                                               |
|                                                               |
|                   Sender's HMAC (160 bits)                    |
|                                                               |
|                                                               |
+---------------------------------------------------------------+
```

Figure 2.10: MP_JOIN option - ACK

only packet where the HMAC from host A is sent, and it has to be acknowledged or retransmitted is the fourth ACK from host B is not received.

These HMAC values are used to authenticate the participants in the subflow establishment, since both have to know the keys for the MPTCP connection. If the creation of the new subflow is not possible because A sends an unknown Token-B to host B or the HMAC material exchanged is not recognized by either hosts or the SYN/ACK received at host A misses the MP_JOIN option, then the operation is stopped sending a TCP RST.

If everything works properly, the entire procedure of instantiating a MPTCP connection and add a subflow is represented with the example in figure **??**.

ADD_ADDR

Even if a host can directly instantiate a new subflow using the MP_JOIN option as previously described, another possibility is for the host to advertise an available address to the other machine, thus allowing the latter to instantiate the subflow. This functionalities can be useful, for example, in a typical client-server configuration in which only the client is allowed to open new connections with the server: if a new interface becomes available at the server, it can dynamically advertise it to the client which in turns can send the SYN+MP_JOIN packet for subflow initiation.

This functionality is provided in MPTCP by the ADD_ADDR option, that indeed contains the additional address (and, optionally, port) to be advertised. To cope with NATs, the option also includes the previously mentioned Address ID 8-bit integer, that has to be bounded to the new address used to create the subflow. The ADD_ADDR option is treated as a soft component of the overall architecture, with no need to be sent reliably and/or be acknowledged by the receiver. The option can be added to any packet in the MPTCP connection if there is enough space in the *Option* field of the

```
          Host A                                    Host B
   ------------------------                       ----------
   Address A1    Address A2                       Address B1
   ----------    ----------                       ----------
       |              |                               |
       |            SYN + MP_CAPABLE(Key-A)           |
       |--------------------------------------------->|
       |<---------------------------------------------|
       |          SYN/ACK + MP_CAPABLE(Key-B)         |
       |              |                               |
       |        ACK + MP_CAPABLE(Key-A, Key-B)        |
       |--------------------------------------------->|
       |              |                               |
       |              |   SYN + MP_JOIN(Token-B, R-A) |
       |              |------------------------------>|
       |              |<------------------------------|
       |              | SYN/ACK + MP_JOIN(HMAC-B, R-B)|
       |              |                               |
       |              |     ACK + MP_JOIN(HMAC-A)     |
       |              |------------------------------>|
       |              |<------------------------------|
       |              |              ACK              |

 HMAC-A = HMAC(Key=(Key-A+Key-B), Msg=(R-A+R-B))
 HMAC-B = HMAC(Key=(Key-B+Key-A), Msg=(R-B+R-A))
```

Figure 2.11: MPTCP authentication example

TCP header, with no guarantee that such option will be received or that the receiver will indeed use the advertised information to start a new subflow. This low priority assigned to ADD_ADDR is reasonable since the malfunctioning of this option would not break the overall data transmission, but it might only cause a missed opportunity for better multipath exploitation. For similar reasons, there is no need to ensure a proper ordering for ADD_ADDR and REMOVE_ADDR at the receiver (REMOVE_ADDR, explained in the following subsection, is similar to ADD_ADDR but it indicates which subflow to shut down instead). Albeit the typical TCP validity test will be performed before inspecting the ADD_ADDR option.

The content of the ADD_ADDR option is shown in figure 2.12. The IPVer indicates if the advertised address is of kind IPv4 or IPv6, while the other fields contain the Address ID, Address and optional port as previously stated.

```
                     1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---------------+---------------+-------+-------+---------------+
|     Kind      |    Length     |Subtype| IPVer |  Address ID   |
+---------------+---------------+-------+-------+---------------+
|          Address (IPv4 - 4 octets / IPv6 - 16 octets)         |
+-------------------------------+-------------------------------+
|   Port (2 octets, optional)   |
+-------------------------------+
```

Figure 2.12: ADD_ADDR option

REMOVE_ADDR

If an address becomes unavailable during a MPTCP connection, the affected host should announce this so that any subflow currently using that address can be terminated. For security purposes, when a REMOVE_ADDR is received, a test is performed to make sure that the address is not available anymore, by sending a TCP keepalive on the path. The Address ID is used to identify the path to be shut down, so that no explicit address is needed (and no address is indeed present in the REMOVE_ADDR option), and the option would work through NATs as well. A subflow that is working properly must not use this option to close the connection, but a FIN exchange similar to regular TCP is performed instead.

```
                    1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---------------+---------------+-------+-------+---------------+
|     Kind      |  Length = 3+n |Subtype|(resvd)|   Address ID  | ...
+---------------+---------------+-------+-------+---------------+
                           (followed by n-1 Address IDs, if required)
```

Figure 2.13: REMOVE_ADDR option

MP_FASTCLOSE

This option can be thought as the MPTCP-level counterpart of the RST signal for the TCP connections: it permits the abrupt closure of the whole MPTCP connection. The RST signals couldn't trigger such behavior, since they are confined to work against a single TCP flow (i.e. MPTCP subflow).

This option can be sent by host A to trigger MPTCP closure at host B. In this case, MP_FASTCLOSE must contain the value of Key-B. When host B receives the option through one of the subflows, it will send a TCP RST answer via the same subflow and then tears down all the subflow. Host A is waiting for the TCP RST answer from host B before tearing down all the subflows. This generic behavior might change slightly if both hosts send an MP_FASTCLOSE at the same time, or if the awaited TCP RST signal is not received within a certain timeout (these would trigger a limited number of retransmissions for this option).

```
                    1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---------------+---------------+-------+-----------------------+
|     Kind      |    Length     |Subtype|      (reserved)       |
+---------------+---------------+-------+-----------------------+
|                      Option Receiver's Key                    |
|                            (64 bits)                          |
|                                                               |
+---------------------------------------------------------------+
```

Figure 2.14: MP_FASTCLOSE option

MP_FAIL

There are various cases in which things might go wrong for a MPTCP connection, and the right procedure to handle such cases is to 'fall back', meaning either falling back to regular TCP or removing the subflow generating the issue. The first solution we saw already for the MP_CAPABLE case, where TCP fall back is guaranteed in case a host is not MPTCP compatible. Similarly, subflow addition will be blocked if anything goes wrong in the MP_JOIN packets' exchange procedure. However, there are other cases in which problems occur after this initiation phases, on regular packets.

As explained in section [ref to section "Data plane"], data acknowledgment with MPTCP requires a DSS option present in the ACK packets. If that option is missing, the path is not considered MPTCP capable. The consequences are different according to the subflow: if the affected path is the first instantiated with the MP_CAPABLE option then it must fall back to regular TCP; any other subflow showing such problem would be closed with a RST. Fallback can be required at any point during the connection if a middlebox modifies the data stream. This case would be detected thanks to the checksum properties of MPTCP data transfer. If checksum fails, all data from the failing segment onwards cannot be trusted anymore. When this happens to a subflow, that has to be immediately closed with a RST and a MP_FAIL option that indicates the data sequence number that failed the checksum. Such option indeed contains a single main field storing the full 64-bit sequence number. The receiver can then avoid to acknowledge untrusted data, that will be sent again through a different subflow following the retransmission features of the data plane part of the MPTCP protocol. After the fallback to regular TCP, it is mandatory not to revert back to MPTCP later on in the connection.

```
                     1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---------------+---------------+-------+----------------------+
|     Kind      |   Length=12   |Subtype|      (reserved)      |
+---------------+---------------+-------+----------------------+
|                                                              |
|                 Data Sequence Number (8 octets)              |
|                                                              |
+--------------------------------------------------------------+
```

Figure 2.15: MP_FAIL option

MP_PRIO

It is possible to indicate if a path has to be used regularly or just as backup in case there no other available regular paths. This preference can be advertised at subflow creation via a flag in the MP_JOIN option, but it is also possible to signal a priority change during the whole time of the MPTCP connection. In fact, it is enough to send the MP_PRIO option to the targeted subflow to switch priority, to signal the other host about the change; it is also possible to add an Address ID to the option to target all the subflows using the addresses associated to such Address ID. This option is only sent from the receiver to the sender, even if the sender can discard such priority preference for any reasons.

```
                     1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---------------+---------------+-------+-----+-+--------------+
|     Kind      |     Length    |Subtype|     |B| AddrID (opt) |
+---------------+---------------+-------+-----+-+--------------+
```

Figure 2.16: MP_PRIO option

2.2.2 Data plane

This part concerns all the MPTCP options used to manage the data flow in a MPTCP connection, including how the byte stream is subdivided into different subflow and how the original order of the packets is provided at the receiver.

DSS option

In order for the receiver to reassemble the correct data stream in MPTCP, a specific option called DSS is used. This option contains the data mapping and Data ACK: data mapping is used to map the sequence space at each subflow (which is independently handled by the TCP protocol) to the overall sequence space at connection level; Data ACK is the procedure to acknowledge data receipt at connection level in MPTCP. The DSS option can also signal the equivalent of a TCP FIN for the overall MPTCP connection, meaning that the current mapping covers the final data from the sender (figure 2.17).

```
Host A                                 Host B
------                                 ------
DATA_SEQUENCE_SIGNAL      ->
[Data FIN]

                          <-           (MPTCP DATA_ACK)
```

Figure 2.17: Closing a MPTCP connection

This option might also include the checksum field to perform integrity checks on the payload (if this was enabled when instantiating the connection via the MP_CAPABLE option). DSS is indeed a versatile packet, whose per-packet behavior is mainly defined by the appropriate flags (figure 2.18).

Regarding the data sequence mapping in MPTCP, the general idea is to maintain TCP-compliant and independent sequence numbers for the single subflows, while using a mapping functionality at the MPTCP-level, provided by the DSS option, to properly rearrange the data at the receiver at guarantee in-order and reliable overall transmission as in the case of legacy TCP. The alternative approach would have been to have a single MPTCP-level sequence number used for the entire set of subflows, meaning that a single subflow inspected by middleboxes at the network infrastructure would look like a TCP connection with holes in the payload delivery; this could trigger unwanted behaviors that would be against the compatibility goals towards the current networks.

The DSS option achieves data sequence mapping with the combination of three fields: for a certain number of bytes (indicated in the *Data-Level Length field*, and starting from the reported subflow sequence number (*Subflow Sequence Number* field), the TCP-level sequence maps to the MPTCP-level sequence with starting value indicated in the *Data sequence number* field. The *Data ACK* field works as regular TCP ACK value, but it refers to the MPTCP-level acknowledgment of the received data. Note that subflow-level acknowledgement is still provided by regular TCP, but a second acknowledgement mechanism at connection-level is desired, since there might cases in which data that has been acknowledges at the subflow-level can be discarded in the buffers before reaching the application. By following the core principles of MPTCP, retransmission of packets can occur at different paths.

2.3 MPTCP deployment

A seamless transition to MPTCP on current infrastructures is a major requirement for MTPCP deployment. Despite the big effort in designing a protocol compliant with strict

```
                     1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---------------+---------------+-------+----------------------+
|     Kind      |    Length     |Subtype| (reserved) |F|m|M|a|A|
+---------------+---------------+-------+----------------------+
|           Data ACK (4 or 8 octets, depending on flags)       |
+--------------------------------------------------------------+
|   Data sequence number (4 or 8 octets, depending on flags)   |
+--------------------------------------------------------------+
|              Subflow Sequence Number (4 octets)              |
+-------------------------------+------------------------------+
|  Data-Level Length (2 octets) |      Checksum (2 octets)     |
+-------------------------------+------------------------------+
```

Figure 2.18: DSS option

compatibility requirements, assuring correct functioning in all the current network scenarios is not a viable possibility for MPTCP. The main problematics are related to unwanted behavior of middleboxes processing unknown MPTCP packets, but that is not the only aspect currently limiting the deployment status of the new protocol. MPTCP has to guarantee the same levels of reliability, performance and security of regular TCP (including the cases in which the fallback mechanism is adopted to switch to plain TCP). For these reasons, MPTCP includes various mechanisms to cope with the most common middleboxes of today's Internet, including the possibility to detect when external boxes operate on the traffic in a way that cannot be handled by MPTCP thus triggering fallback to regular TCP.

2.3.1 Middleboxes compatibility

The Internet at its core was designed to provide end-to-end connectivity across an infrastructure of interconnected routers. Nevertheless, the growing rate of adoption and increasing complexity of the Internet brought up a wide set of new requirements inherent the intermediate stages of the communication rather then the end-hosts. Such requirements include the need to instantiate protection techniques against potential attacks, more flexibility in content delivery, caching for more efficient communication; they can even be more specific, like the need to rapidly overcome the previously mentioned IPv4 addresses depletion. Middleboxes are pieces of equipment that operate on the network traffic to meet these requirements. The most common middleboxes are NATs, proxies and firewalls, but nowadays there is a huge variety of deployed middleboxes that inevitably break the end-to-end principle of the Internet. Despite their usage is often required, they are more or less intrusive at different layers of operation (according to where they operate within the OSI architectural model), thus causing malfunctioning of many protocols, and MPTCP is no exception. Middleboxes can indeed inspect packets, re-route them, drop them, split them into multiple fragments, and even modify single fields in packets' headers (like rewriting sequence number or removing TCP options) as well as change their payload.

The various operations and purposes of middleboxes are many and often mixed together to achieve more complex policies, and it is also very common that different kind of operations are performed inside the same physical machine. Despite this, it is possible to define a set of most common distinct middleboxes' operations [href].

Firewalls

A simple example can be the case of a standard firewall that is not MPTCP-aware and its default policy is set to "Deny". In this case, all the traffic is blocked unless the set of custom rules configured in the firewall explicitly allows a specific typology of packets. In this case, specific for MPTCP must be added to support the new protocol, and this might cause a considerable effort for network maintainers. For example, it is often not straightforward to operate on legacy firewall configurations for big companies with many access points. A more subtle problem with firewalls might be derived by the fact that they can sometimes manipulate the sequence numbers of a TCP connection, thus moving the sequence number space with respect to the initial value in use by the end-hosts. This feature has been introduced in the past to improve security with older TCP/IP stacks, but the concept can disrupt MPTCP mapping between subflow sequence number and MPTCP-level sequence number. To avoid such problems, MPTCP is designed so that the mapping in the DSS option is using relative values compared to the initial sequence number and functioning is not jeopardized but changes of the absolute values performed by the firewalls. Yet another case concerns firewalls that remove unknown TCP options for security purposes. If such operation is symmetric, TCP segments would lose the MP_CAPABLE option and fallback seamlessly to regular TCP. However, there are middleboxes that operates asymmetrically thus removing unknown TCP options only fro non-SYN segments. To cope with this, MPTCP requires that for the first window of data, each segment must include an MPTCP option, otherwise fallback is assured [href].

NATs

Another ubiquitous piece of equipment is the "Network Address Translation" (NAT). As the name suggests, NATs modify the IP addresses within the packets on their way towards the destination. The main purpose is to group addresses of an internal private network and map them to a single public address before forward the traffic to the Internet. NATs are also able to redirect the response from the Internet to the right host in the internal network. This procedure became very common with the depletion of IPv4 addresses, since in many cases the address space assigned to a outer portion of the Internet is not large enough to cover the number of hosts willing to acquire connectivity. NATs turned out to be a very effective way to temporarily solve the problem of IPv4 addresses, but their mode of operation is intrusive at the network and transport layer, since the IP addresses are not fixed anymore. For example, even if NATs use internal tables to keep track of the mappings and are able to redirect replies from the external network to the right internal host, it is no more possible for the external hosts to instantiate a new connection with a specific host residing behind a NAT. This is true also in MPTCP, such that a server often cannot open a new subflow with a client if the latter is behind a NAT, even if a valid MPTCP session between client and server is already active. This is one of the main use cases in which an ADD_ADDR message can be sent on live subflows in order to trigger a new subflow connection request from the other side. Moreover, to cope with NATs that might be operational on the paths and might change the source address of the packets, MPTCP refers to addresses by using an Address ID, as previously mentioned in the section [add ref to control plane].

Segment splitting and coalescing

There are middleboxes that split segments on the Internet as required by the MTU (maximum transmission unit). This means that the payload of a single TCP packet can be scattered across multiple smaller TCP packets and regrouped back together by using the appropriate TCP fields in the header. This operation usually copies TCP options unchanged into each of the smaller packets that are generated. By simply adopting data-sequence numbers in a MPTCP option, the receiver might receive different packets with the same data-sequence numbers and it would be unable to reconstruct the original data. MPTCP takes care of segment splitting and coalescing by mapping the subflow-level TCP sequence number with the MPTCP-level sequence, by providing both the beginning (with respect to the subflow sequence number) and the length of the of the data-sequence mapping [section about DSS]. MPTCP would work also in the more uncommon cases in which segment splitters copy the original TCP option in only one of the generated smaller segments. If the first data-segment does not contain an MPTCP option, fallback to regular TCP is performed (as explained in 'Firewalls' section), otherwise MPTCP would work seamlessly even under these circumstances [href].

Application-level gateways

There are middleboxes that operate at higher layer in OSI model, modifying the payload of the packets: adding and removing bytes can change the boundaries of the data-sequence mapping and MPTCP information about it would become inconsistent. The only way to cope with this case is to fallback to regular TCP. In order to that, MPTCP has to detect when the payload has been changed by middleboxes and that is the main reason for which the checksum field has been added inside the DSS option [ref to DSS section]. The checksum calculation is optional in MPTCP and can be negotiated during connection establishment with a flag in the MP_CAPABLE option. Nevertheless, it is recommended for operations on the open Internet.

2.3.2 Deployment status

MPTCP proves to be a major TCP extension, and in this regards its design required a lot of efforts and several interconnected research projects. The European Commission funded the work at the UniversitÃĩ catholique de Louvain with the FP7 Trilogy project in 2007 [href], followed by CHANGE [href] and Trilogy 2 [href]. Fundings have been instantiated by Google and Nokia, too [href].

Also by analyzing the main steps in MPTCP evolution we can detect the big interested in the protocol: six month after the Experimental standard for MPTCP has been published in January 2013 by the IETF, it was possible to count three major independent MPTCP implementations other than the Linux kernel implementation [href], including a FreeBSD implementation from Swinburne University of Technology [href] and a NetScalar Firmware implementation from Citrix Systems [href]. Moreover, recent versions of MPTCP kernel (from 0.89.5) are now compatible with Android (with some limitations), and many porting projects have been developed to test older versions of MPTCP on various Android devices [href]. As of June 2015, a Solaris implementation is reportedly under development by Oracle [href]. All these implementations follows the standard RFC documentation for MPTCP, and they have shown good interoperability capabilities while operating with the standard MPTCP-compatible Linux kernel, especially for what regards the core MPTCP signaling messages (secondary MPTCP features like ADD_ADDR address advertisement are not always implemented [href]).

The very first large scale commercial deployment of MPTCP dates back to 2013, when Apple introduced the new protocol in iOS7 to work with the intelligent personal assistant Siri. Apple's mobile operating system implements MPTCP as in RFC-6824 (excluding some features) in order to use cellular data subflow in case the Wi-Fi connectivity becomes unavailable during a Siri request processing [href]. This is indeed the first example of wide adoption of MPTCP over the Internet even if limited to a specific Apple service connecting to proprietary servers. Nevertheless, the news was helpful in spreading the awareness about the protocol also to a more consumer-oriented audience. Apple also added MPTCP capabilities to Mac OS X 10.10 in October 16, 2014 [href], proving to be very active in developing and testing MPTCP.

In studying the protocol's deployment process, it is very important to analyze the relation between costs and benefits that MPTCP would bring to each and every group of MPTCP stakeholders. The success of MPTCP depends on its deployment, and its deployment strongly depends on endpoints. We have already discussed about the considerable interest shown by OS authors, which naturally fits the pre-deployment stage. But eventually it will be the end-users to decide the future for MPTCP: they are the ones directly accessing the biggest part of MPTCP benefits as described in section [add MPTCP benefits section]. Leaving aside middleboxes interference, there is conceptually no need for technical modifications at the intermediate infrastructure to make MPTCP available at the end-users. Nevertheless, connectivity providers (ISPs) still represent an important part of the entire set of stakeholders that might benefit from MPTCP wide adoption: multipathing can directly improve resource utilization and congestion bottlenecks within the overall infrastructure, but it can also be seen by ISPs as an enabler of new business models, since end-users might show an increase interest in multihoming solutions [href]. End-users' feedback and ISPs' feedback for MPTCP do and will drive the interest of infrastructure vendors to better support the protocol or not with their middleboxes. Yet another case study involves data infrastructure maintainers, that can be considered a smaller but important subset of end-users. In this case it is fundamental the value that MPTCP can bring to data centers of today as well as the possibilities enabled by MPTCP for the design of the data centers of the future [href].

All these considerations are difficult to analyze in the real world, thus making it hard to predict future trends for MPTCP adoption. Current applications of MPTCP rarely detach from experimental branches and little is known on how the new protocol would behave in the Internet if globally enabled. Excluding the MPTCP usage for Siri and Apple's server, the closest example of real world usage of MPTCP has been setup and analyzed by the UniversitÃľ catholique de Louvain: the experiment consisted in collecting a dataset about traffic usage for an MPTCP-enabled Web server exposed to the open Internet in November 2014 [href]. The Web server was running th stable version 0.89 of the MPTCP implementation in the Linux kernel and using a single physical network interface supporting both IPv4 and IPv6. As for the content, the Web server was hosting the Multipath TCP implementation in the Linux kernel, a common destination for early adopters of the new protocol. After on week of monitoring, the dataset included around 122 millions of TCP packets destined to the Web server and roughly a quarter of those were MPTCP packets for a total of 5098 observed MPTCP connections. An interesting fact about the analyzed ADD_ADDR packets showed that clients advertised mostly private addresses (79% of the IPv4 advertised addresses), thus confirming the importance of

MPTCP being able to pass through NATs. As a side consideration, this exposure of private addresses enabled by ADD_ADDR could raise some security concerns, since it might allow to discover and enumerate private networks. The final evaluation for this experiment demonstrated that MPTCP works properly in the open Internet if the Application Level Gateway are handled by protecting the payload using the checksum in the DSS option (a feature enabled on server side for the entire set of 5098 MPTCP connections).

For what regards the current numbers MPTCP-enabled clients and servers around the world, such information is not easy to retrieve. For this purpose, a service has been built by NICTA (Sidney) and Simula Research Laboratory (Oslo), to scan the most common Web servers for the websites retrieved from the Alexa Top 1M list and check for MPTCP compatibility. This test is run between once a day and once a week, so that a live dashboard showing the retrieved data over time is maintained [href]. According to their latest results, the current rate of adoption of MPTCP from the scanned IP addresses and domains is around 0.1% [href], showing that the current status is far from large scale adoption, despite the surprising number of implementations.

Chapter 3

MPTCP security

3.1 Threats analysis

A complete security evaluation of MPTCP can be subdivided into two main categories.

A first perspective is to study of the vulnerabilities in the current MPTCP design that can be exploited to carry out flooding or hijacking attacks on an MPTCP session. This is an assessment on how consistently the MPTCP extension would impact the security standards of a plain TCP connection.

A second perspective is to understand how the new protocol affects the functioning and behavior of external security gears. This evaluation might include compatibility issues for middleboxes not yet aware of MPTCP [ref section of middleboxes] as well as more fundamental problematics related to security monitoring solutions that wouldn't work anymore with MPTCP: by splitting the logic flow of data into different paths, potentially belonging to different ISPs, it would be much harder to keep track of the content of the transmitted data over the networks. Moreover, the MPTCP ability to reroute traffic on the fly, adding and removing addresses and interfaces, would per se cause major problems with current intrusion detection and intrusion prevention equipment.

This paper focuses on the first point: MPTCP enables data transmission using multiple source-destination address pairs per endpoint and this generates *new* scenarios in which an attacker can exploit the way subflows are generated, maintained and destroyed to perform flooding or hijacking attacks. Flooding attacks are Denial-of-Service procedures that aim at overloading an MPTCP host with connection requests in order to quickly consume its resources. Hijacking attacks aim at taking total control of the MPTCP session, thus being considered the ultimate example of those threats falling in the *active attacks* category (more and this later on, in this section).

MPTCP security mechanism was designed with the primary goal of being at least as good as the one currently available for standard TCP [RFC6181]. The official MPTCP documentation and analysis reports don't cover common threats affecting both TCP and MPTCP, but only the vulnerabilities introduced by the new protocol alone. Nevertheless, it is of paramount importance that the various security mechanisms deployed as part of standard TCP, for example mitigation techniques for reset attacks, are still compatible with Multipath TCP.

Apart from the fundamental objective of keeping MPTCP at least as reliable and secure as TCP, official documents offer another set of requirements mainly related to securing subflow management in MPTCP [RFC6824bis]. These requirements are:

- Provide a mechanism to confirm that the parties in a subflow handshake are the same as in the original connection setup.
- Provide verification that the peer can receive traffic at a new address before using it as part of a connection.
- Provide replay protection, i.e., ensure that a request to add/remove a subflow is fresh.

MPTCP involves an extensive usage of hash-based handshake algorithms to achieve the required security specifications, as described in chapter 2.

Once the security requirements are clear, it follows a set of related problematics due to the way MPTCP is added to the normal TCP stack. Most notably, the entire behavior of the protocol relies on the TCP Options field, which is of limited length of 40 bytes. This factor plays an important role in the definition of the security material to be exchanged during an MPTCP session (truncating the HMAC values and tokens is a common technique). Moreover, TCP Options field has been designed to accept any custom protocol extending TCP and for security reasons many middleboxes would discard or modify packets containing unknown options. As a last point, MPTCP approach to subflows' creation implies that a host cannot rely on other established subflows to support the addition of a new one [RFC6182-5.8]; this last requirement follows the *break-before-make* property of MPTCP, that must be able to react to a subflow failure a posteriori by establishing new subflows and automatically sending again the undelivered data. All these considerations define the fundamental boundaries and the context in which the security design of MPTCP has to be developed to meet the requirements.

3.1.1 Threats classifications

Introducing the support of multiple addresses per endpoint in a single TCP connection does result in additional vulnerabilities compared to single-path TCP. These new vulnerabilities need proper investigation in order to determine which of them can be considered critical and might require modifications in the protocol design in order to meet the required specifications. In order to classify how critical each security threat is, it is a good starting point to define the various typologies of attack according to their requirements, rate of success and what power they can provide to the attacker.

The general requirements for an attack to be executed might be grouped into the following categories:

- *Off-path attacker*: the attacker does not need to be located in any of the paths of the MPTCP connection at any time in order to execute the attack;
- *Partial-time (time-shifted) on-path attacker*: the attacker has to be able to eavesdrop a specific set of information during the lifetime of the MPTCP connection in order to execute the attack. It doesn't need to eavesdrop the entire communication in between the hosts, and the specific direction and/or subflow for the sniffing procedure are attack specific;
- *On-path attacker*: this attacker has to be on at least one of the paths during the entire lifetime of the MPTCP session in order to execute the attack.

We can clearly state that the critical case concerns off-path attacks, which do not require any eavesdrop procedure in order to be executed. In fact, on-path attacks are not considered part of the MPTCP work, since they allows for a significant number of attacks on regular TCP already. A primary goal in the design of MPTCP is not to introduce new ways to perform off-path attacks or time-shifted attacks.

The effects of an attack over an MPTCP connection and the power that the attack can provide to the attacker can be divided into two main categories:

- *Passive attacker*: the attacker is able to capture some or all of the packets of the MPTCP session but it can't manipulate, drop or delay them, and it can't inject new packets in the current session either.
- *Active attacker*: the attacker can pretend to be someone else, introduce new messages, delete existing messages, substitute one message for another, replay old messages, interrupt a communication's channel, or alter stored information in a computer.

The rate of success of a certain attack over a MPTCP connection strongly depends on the specific requirements: two attacks falling in the same categories in terms of attacker eavesdrop capabilities and passive/active typologies might have rather different rates of success. For example, a certain kind of attack might require IP spoofing, thus being unfeasible in a network with ingress filtering [add reference]. There are no general thresholds to define when an attack can be considered a real threat according to the success rate, but this is an important factor to be studied in an attack analysis.

3.2 Minor threats

In this section are presented the other residual threats under analysis by the IETF community at the time of writing. They all fall into two main kinds of attacks: flooding attacks and hijacking attacks.

3.2.1 DoS attack on MP_JOIN

This kind of DoS attack would prevent hosts from creating new subflows. In order to be executed, the attacker has to know a valid token value of an existing MPTCP session. This 32-bit value can be eavesdropped or the attacker has to guess it.

This attack exploits the fact that a host B receiving a SYN+MP_JOIN message will create a state before answering with the SYN/ACK+MP_JOIN packet. This means that some resources will be consumed at the host to keep in memory information regarding this connection request from the other party; in this way, when the host B receives the third ACK+MP_JOIN packet, it can correctly associate it to the initial request and complete the handshake procedure. The creation of such state is required because there is no information in the ACK+MP_JOIN packet that links it to the first SYN+MP_JOIN request, so it is up to the host to remember all the ongoing requests. An attacker can exploit this by sending SYN+MP_JOIN packets to a host without providing the final acknowledge packets. This can be done until the attacked host runs out of available spots for initiating additional subflows. The initial number of such available spots depends on the implementation and configuration at the host machine.

This attack can be exploited to perform a typical TCP flooding attack. This is the perfect example of how MPTCP might introduce new vulnerabilities that might affect the underlying TCP protocol. SYN flooding attacks for TCP have been studied for many years and current implementations use mitigation techniques like SYN cookies [reference] in order to allow stateless connection initiations. But each SYN+MP_JOIN packet received at the host would trigger the creation of an associated state, while this is not the case for the attacker machine that can simply forge these packet in stateless manner. Exploiting this unbalance in resource utilization is referred to as *amplification attack*.

A possible solution to this problem is to extend the MP_JOIN option format to include the information required to identify a specific request throughout the 3-way handshake, without requiring hosts to create associated states.

3.2.2 Keys eavesdrop

An attacker can obtain the keys exchanged at the beginning of the MPTCP session, exploiting the fact that those are sent in clear. This is in fact a partial-time on-path eavesdropper attack, whose success would enable a vast set of attacking scenarios, even if the attacker itself has moved away from the session after sniffing the aforementioned keys. The keys associated to an MPTCP session are sensitive pieces of information, used to identify a specific connection at the hosts and used as keying material for all the HMAC computations in the protocol. With such pieces of information an attacker can potentially execute a connection hijacking. This problem is encountered again when analyzing the ADD_ADDR attack, Section 3.3 and a more in-depth analysis of the possible mitigation techniques for this attack is presented.

The maximum length of the TCP Option field brings strong limitations for security implementations: for example, using certificates in TCP Options would be impossible. Moreover, strong cryptographic computation is also discouraged inside TCP for performance reasons. Nevertheless, many techniques can be used to prevent the keys' eavesdrop attack other than the more obvious possibility of adopting pre-shared keys. Such techniques are mentioned in RFC-7430 and briefly explained in the following sections.

Hash chains

Hash chain is a way to obtain many one-time keys applying a cryptographic hash function recursively, starting from a random seed S:

$$H[0] = H(S); H[1] = H(H[0]); H[2] = H(H[1]); ...; H[n] = H(H[n-1])$$

This technique allows to authenticate end hosts without the need to exchange keying material upfront. In the following explanation we will discuss a simplified scenario in which only host A needs to authenticate itself to host B. If host A initially identifies itself giving H[n], it can later on send H[n-1] for authentication towards host B. In fact, hash chains cannot be reversed (i.e. it is impossible to find H[n-1] by just knowing H[n]), meaning only host A could have generated a valid H[n-1]; host B can verify its authenticity by simply hashing it and checking that it is equal to the previously seen H[n]. The main issues with this operation is that, once H[n-1] is sent by host A, it cannot be reused, since this might have eavesdropped, leading to a situation not dissimilar to the original problem. With key chains, hosts would continue scaling down the chain, meaning that a second authentication would require host A to send H[n-2] (of course, host A can calculate any value in the chain since it knows the original seed), and host B must have saved previously acknowledged H[n-1] to verify that H[H[n-1]] equals the received H[n-2] value. The main issue with this solution is that the value 'n' will reach 0 sooner or later, meaning that host A needs to compute a new hash chain from a new seed and also signal host B about this operation, thus requiring the definition of a new MPTCP option containing the final entry of the old chain (for authentication purposes) and the first entry of the new chain. Another direct consequence of such solution is that hash chains would add computational complexity to MPTCP operations, despite it being still reasonably acceptable. An unverified proposal for the new message exchange using hash chain can be found on the IETF mailing list, and it is shown in figure 3.1 [href].

```
         Host A                                       Host B
------------------------                     ------------------------
Address A1     Address A2                    Address B1     Address B2
----------     ----------                    ----------     ----------
    |              |                             |              |
    |     SYN + MP_CAPABLE(Key-A, H-A[n])        |              |
    |------------------------------------------->|              |
    |<-------------------------------------------|              |
    |    SYN/ACK + MP_CAPABLE(Key-B, H-B[n])     |              |
    |              |                             |              |
    |              |                             |              |
    |              |     SYN + MP_JOIN(Token-B, H-A[n-1])       |
    |              |------------------------------------------->|
    |              |<-------------------------------------------|
    |              |   SYN/ACK + MP_JOIN(Token-A, H-B[n-1])     |
    |              |                             |              |
    |              | HMAC(K=(K-A+K-B), M=(H-A[n-1]+H-B[n-1]))   |
    |              |------------------------------------------->|
    |              |<-------------------------------------------|
    |              | HMAC(K=(K-B+K-A), M=(H-B[n-1]+H-A[n-1]))   |
    |              |                             |              |


Token-A = Hash(IDSN-A); IDSN-A = Hash(Key-A)
Token-B = Hash(IDSN-B); IDSN-B = Hash(Key-B)
H-A[n] = nth entry in A's hash chain
```

Figure 3.1: Hash chain message exchange proposal

The first two messages for the MP_JOIN operation are used for authentication purposes (the hash chain values are transmitted there), but the last two messages operate in a similar fashion with respect to the current MPTCP solution, using an HMAC whose key depends on the original keys exchanged via the MP_CAPABLE option. In this way, eavesdropping the original keys is not enough to operate on the connection, but knowing the original keys is still required to validate subflows' creation.

SSL/TLS and SSH

Another well rated proposal is to use application-layer protocol like SSL/TLS or SSH to negotiate a shared key between the end-points. For example, SSL/TLS already provides a mechanism to negotiate shared secret by using a Diffie-Hellman algorithm [href]. An RFC draft can be found to describe a possible prototype for this solution. A bit field in the MP_CAPABLE option would signal the intent of using application-provided keys for the connection (maintaining retro-compatibility with older versions of MPTCP that do not support this feature). The main back draw of asking the application layer to provide the security mechanism, is that the application itself has to be upgraded to provide the necessary MPTCP socket options [href]:

- MPTCP_ENABLE_APP_KEY: When this option is enabled, MP_CAPABLE is sent with the proper bit in order to signal the usage of an application supplied key for authentication;
- MPTCP_KEY: This socket option is used to pass the actual key to the MPTCP layer.

Despite being the message exchange almost unchanged under these circumstances, some synchronization concerns might arise due the fact that it's possible the client's application has already called the socket with the proper options while the server is still waiting for the key. In this case, silently dropping the SYN packets from the client, together with the usual TCP retransmission mechanism, should solve the problem.

Secure MPTCP

Secure MPTCT[href] (SMTCP) refers to the integration of MPTCP with tcpcrypt, the latter being a protocol that attempts to encrypt almost the entire content of the traffic [href]. SMTCP has been proposed as more secure version of MPTCP that would protect the data stream itself rather than addressing each and every security flaw in the signalling component of the protocol. Indeed, all the MPTCP signalling data would be encrypted and integrity protected as well, thus meaning that the overall protection for MPTCP would be achieved by the tcpcrypt extensions alone. An interesting factor of this solution, is that tcpcrypt also require sharing keying material to provide encryption, thus being tcpcrypt itself vulnerable to Man-in-the-Middle attacks during the initial key negotiation.

Other proposals

Use of Cryptographically Generated Addresses (CGAs) [RFC-3972]; Use of DNSSEC [href].

3.2.3 SYN/ACK attack

This is a partial-time on-path active attack. An attacker that can intercept and alter the MP_JOIN packets is able to add any address it wants to the session. This is possible because there is no relation between the source addresses and the security material in the MP_JOIN packets. But securing the source address in MP_JOIN is not feasible if MPTCP is supposed to work through NATs: these middle-boxes operate exactly as described in this attack procedure.

Possible solutions have to reside on a different layer, perhaps securing the payload as a technique to limit the impact of such attack in a MPTCP session.

3.3 ADD_ADDR attack

This paper is mainly focused on studying and testing the ADD_ADDR vulnerability of MPTCP, as well as providing an analysis of the commonly accepted fix and its implementation in Linux kernel. This section describes the attack procedure in details, while the considerations about the possible solutions for the ADD_ADDR vulnerabilities can be found in chapter 5.

3.3.1 Concept

The ADD_ADDR attack is an *off-path active attack* that exploits a major vulnerability in the MPTCP version 0 design [ref to RFC version 0]. As previously mentioned, the attacks falling into this category are usually the most critical and can easily jeopardize the protocol security requirements. With the current MPTCP model, an attacker can forge and inject an ADD_ADDR message into an MPTCP session to achieve a complete hijacking of the connection, placing itself as a man-in-the-middle. Being this an off-path attack, the attacker can *conceptually* send the forged ADD_ADDR message from anywhere in the network (as allowed by routing), with no need to be physically close to the victim machines. At the end of the attacking procedure, the attacker will be able to operate in any way on the ongoing data transmission, with no clear warning given to the original parties involved in the MPTCP session. If no protection system is used at the application layer (like data encryption), the attacker can eavesdrop all the information and even modify or generate the exchanged content. The attack vector enabled by such exploit is huge.

The culprit is indeed the format of the ADD_ADDR option, whose behaviour will be changed with the new ADD_ADDR2 for the very purpose of fixing this vulnerability [last chapter ref]. This vulnerability is entirely related to the MPTCP design, and due to its characteristics it is considered a blocking issue in the MPTCP progress towards Standard Track [7430].

3.3.2 Procedure

Let's consider a scenario in which two machines, host A and host B, are communicating over an MPTCP session involving one or more subflows. The attacker is called host C and it is operating remotely with no eavesdrop capabilities. The attacker is using address IPC and targeting a single MPTCP subflow between host A (address IPA and port PA) and host B (address IPB and port B). The scenario is reported in figure 3.2.

Here is reported the step-by-step procedure to carry out the ADD_ADDR attack:

1. The first step performed by the attacker is to forge an ADD_ADDR message as follows: it is an ACK TCP packet with source address IPA, destination address IPB and the advertised address in the ADD_ADDR option is IPC. The ADD_ADDR option also contains the *Address ID* field, that the attacker can set to a number high enough in order not to collide with existing identifiers for the ongoing subflows between hosts A and B. The format of all the various MPTCP option can be found in chapter 2. The forged packet is then sent to host B.

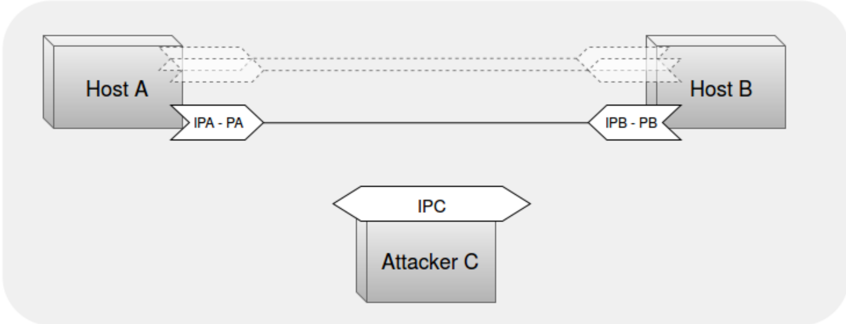

Figure 3.2: Attack scenario

2. Host B will process the forged packet as a legit request by host A of advertising a new available interface with address IPC. This most likely triggers the creation of a new subflow towards the new IP address, meaning that host B sends a SYN+MP_JOIN packet to the attacker. This packet contains all the security material needed in the first phase of the MP_JOIN three-way handshake, and the attacker does NOT need to operate over that portion of data: the attacker C simply manipulate the SYN+MP_JOIN packet by changing the source IP to IPC and the destination IP to IPA; then, it forwards such packet to host A.

3. Host A will process the incoming packet as a legit request by host B of starting a new subflow from host B's new available interface having address IPC. All the required information is present in the MP_JOIN option, like the token of host A that identifies the specific MPTCP session to which attach the new subflow to. Host A computes all the needed parameters (like a valid HMAC value), generate the SYN/ACK+MP_JOIN packet and send it to IPC. The attacker, similarly to the previous steps, manipulate the IP addresses of the packet from A by changing the source endpoint from IPA to IPC and the destination endpoint from IPC to IPB. At this point, attacker C sends the packet to host B.

4. All the parameters in the received packet looks correct to host B, which replies with an ACK+MP_JOIN packet to attacker C. The attacker changes the source address to IPC and the destination address to IPA and sends the modified packet to host A. Upon acknowledge reception, host A will verify all the parameters in the packet (which will be correct since properly calculated by host B), and create a new subflow towards IPC. At this point the attacker has managed to place itself as man-in-the-middle.

5. As a further, optional step, the attacker can send RST packets to the other subflow in order to close them thus being able to perform a full hijack of the MPTCP session between host A and host B. The attacker can now operate upon the connection in any possible way, modifying, delaying, dropping, forging packets between the two parties.

By exploit the ADD_ADDR option, the attack procedure is relatively straightforward. Albeit there are some important requirements and limitations that consistently limit the rate of success of such attack, which are discussed in the following section.

3.3.3 Requirements

A first, basic prerequisite needed by the attacker to inject the ADD_ADDR message into an ongoing MPTCP session is to know the IP addresses and port values adopted by host A and host B for the targeted subflow. It is reasonable to assume that the IP addresses are known. In a typical client-server configuration, the server's port for a certain application protocol is fixed and can be assumed to be known, too. For the client counterpart, the port value can cause problems in the presence of protection techniques like port randomization [ref]: in these cases the attacker has to start a guessing procedure whose rate of success also depends on the ephemeral port range employed [ref].

The knowledge about the above-mentioned four-tuple is a basic requirement for obvious reasons, but knowing the endpoint details is not enough to inject valid packets into an ongoing TCP session (that, in this case, can be also seen as an MPTCP subflow session): these packets have to contain SEQ and ACK sequence numbers that are compatible with the current ones within the stream. SEQ and ACK values are used in TCP to provide reliable, in-order transmission of data as well as services related to flow and congestion control [ref]. A very common protection technique is to randomize those 32-bit values at TCP connection setup, forcing the attacker (who acts off-path) to blindly guess them. TCP provides a window mechanism to deal with possible transmission's misalignments: at any given time, the accepted ACK values are those between the last ACK received and the same value plus the receiving window parameter. As a result, the number of packets to be sent in the attempt of guessing the right SEQ and ACK values and consequently the rate of success of the attack are strongly influenced by the TCP receive windows size at the targeted TCP host.

The requirements listed so far all pertain to the underlying TCP protocol. The only MPTCP specific parameter that can cause the failure of the ADD_ADDR attack procedure is the Address ID field in the option. The purpose of this value has been previously explained, and it doesn't actually offer an overall protection improvement. It is enough for the attacker to chose an ID value that is not in use by other subflow in the MPTCP session. In usual scenarios with a relatively limited number of subflow with the MPTCP session, applying a random value to this field should work just fine.

Moving away from the inner parameters evaluation and taking into consideration external protection mechanisms, it is worth mentioning that the attacker has to be able to manipulate and forge packets, including changing their source address field. This process, known as IP spoofing [ref], is a well known technique for which protection technologies have been developed, most notably the ingress filtering [2827] or source address validation [6056]. However, these methods are not vastly deployed and cannot be considered a sufficient mitigation for the ADD_ADDR vulnerability [ref on ingress filtering usage].

Lastly, the attacker has to be able to direct the malicious ADD_ADDR packet to a host that is actually capable of starting a new subflow, namely the client in a client-server model. The current Linux kernel implementation prohibits the server to instantiate a new subflow and only the client does so.

Chapter 4

ADD_ADDR attack simulation

4.1 Environment setup

In order to achieve a reliable reproduction of a real world scenario, the simulation involves the setup of two User Mode Linux (UML) virtual machines running a Linux kernel with enabled support for MPTCP. These two machines act as client and server, carrying on an MPTCP connection that is the target for the ADD_ADDR attack. Using UML to proceed with the experiments allows for very fast setup and boot-up time, with good emulation of real machines and giving the possibility to work on a single hosting machine with no risk of damaging or crashing its underlying kernel.

A good resource in terms of tools, configuration files and kernel images is the official mptcp website: *http://www.multipath-tcp.org*. In particular, the website offers a python script that downloads all the necessary files to run the two virtual machines. Considering our purpose of verifying the ADD_ADDR attack feasibility, there is no need to modify or debug the Linux kernel source code, and the above mentioned components can be used out of the box. At this stage of the analysis it is actually advised to perform the attack on the official distribution as is, and develop external tools for injecting packets and monitoring the status of the connections. More specifically, the MPTCP version adopted for the tests is: *Stable release v0.89.0-rc*.

When executing the script *setup.py* retrieved from the official Website, a few files are downloaded. A *vmlinux* executable file with the MPTCP compatible Linux kernel, two file-systems for the client and the server (*fs_client* and *fs_server*) and two shell scripts to configure and run the virtual machines (*client.sh* and *server.sh*). No manual configuration is needed, and client and server should be able to connect via MPTCP right away. Here it follows the content of the *client.sh* (a similar shell script that is not reported here can be found for the server counterpart, including a single *tap2* interface setup in that case):

```
#!/bin/bash

USER=whoami

sudo tunctl -u $USER -t tap0
sudo tunctl -u $USER -t tap1

sudo ifconfig tap0 10.1.1.1 netmask 255.255.255.0 up
sudo ifconfig tap1 10.1.2.1 netmask 255.255.255.0 up
```

```

sudo sysctl net.ipv4.ip_forward=1
sudo iptables -t nat -A POSTROUTING -s 10.0.0.0/8 ! -d 10.0.0.0/8 -j
    MASQUERADE

sudo chmod 666 /dev/net/tun

./vmlinux ubda=fs_client mem=256M umid=umlA eth0=tuntap,tap0
    eth1=tuntap,tap1

sudo tunctl -d tap0
sudo tunctl -d tap1

sudo iptables -t nat -D POSTROUTING -s 10.0.0.0/8 ! -d 10.0.0.0/8 -j
    MASQUERADE
```

Listing 4.1: *client.sh*

These scripts call the *tunctl* command to create the tap interfaces and later assign an IP address to them by using *ifconfig*. A *tap* (namely network tap or tap interface) simulates a link layer device and it can be used to create a network bridge [wiki]. How taps are used in our simulation will become clear when observing the final network scenario. In order for the new tap interfaces to recognize each other and being able to send packets to each other it is necessary to enable the *ip forwarding* option using the corresponding *sysctl* command. It is also necessary to configure the *iptables* upon startup... The virtual machine is launched by executing the *vmlinux* file with some options to define various properties as well as attaching the newly created tap interfaces that will be used locally to sniff and inject packets (acting, in this specific case, as a physical man in the middle).

The resulting network scenario is graphically depicted in Figure 4.1.

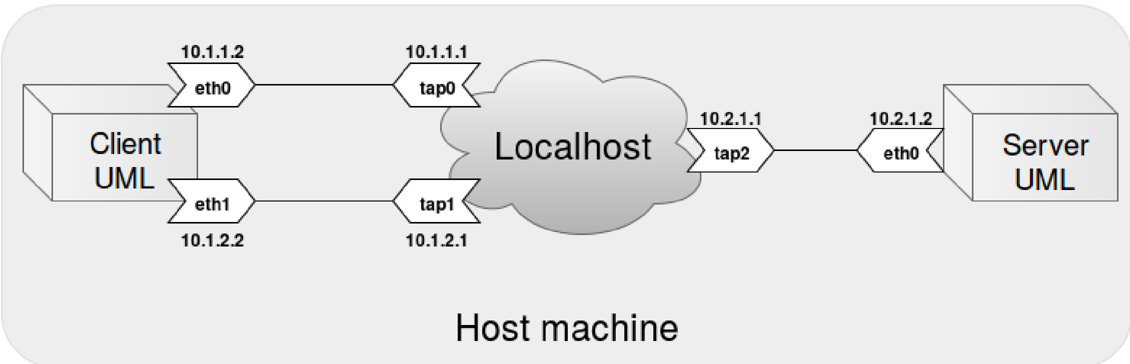

Figure 4.1: Network scenario

In order to carry out the ADD_ADDR attack it is necessary to inject forged packets into the existing MPTCP flow. In order to do this it is possible to use Scapy, a powerful interactive packet manipulation program that is able to forge or decode packets of a wide number of protocols, send them on the wire, capture them, match requests and replies, and much more [*http://www.secdev.org/projects/scapy*]. Moreover, there exists an unofficial version of Scapy that supports MPTCP and it can be found at the following repository: *https://github.com/nimai/mptcp-scapy*. The Python script that can be used to carry out the ADD_ADDR attack can be found here: *https://github.com/fabriziodemaria/MPTCP-Exploit*.

It is appropriate to mention here some of the limitations of the tool (that are examined more in details in Section 4.4: *Limitations and future work*): the tool has been designed to hijack a specific kind of communication involving client and server sending each others text messages using the tool *netcat*. It is very unlikely that the procedure completes with other kind of MPTCP connection setups between client and server. Nevertheless, this specific exploit serves well our purpose of assessing the danger and feasibility of the ADD_ADDR attack in general terms. Moreover, this tool simplify the attack procedure by sniffing the SEQ and ACK numbers of the ongoing connection instead of starting a procedure to try and guess the values. Also, the ports in use by the client and the server are retrieved automatically by inspecting the sniffed packets, while the IP addresses have to be provided by the user when launching the attack script. Further considerations about these simplifications can be found in Section 4.4.

The python module *test_add_address.py* in the root of the GitHub repository follows the analysis in RFC7430[] to perform the various steps necessary to hijack the MPTCP connection. All the requirements and theoretical details about this procedure have been reported in Section 3.3, and this section is limited to show and investigate the actual code implementation.

By establishing a testing connection via the *iperf* tool, two subflows are automatically generated by MPTCP, from the two interfaces of the client (ip addresses: *10.1.1.2* and *10.1.2.2*) and the single server's interface (with ip address: *10.2.1.2*).

In order to carry out the ADD_ADDR attack it is necessary to inject forged packets into the existing MPTCP flow. In order to do this it is possible to use Scapy, a powerful interactive packet manipulation program that is able to forge or decode packets of a wide number of protocols, send them on the wire, capture them, match requests and replies, and much more [*http://www.secdev.org/projects/scapy*]. Moreover, there exists an unofficial version of Scapy that supports MPTCP and it can be found at the following repository: *https://github.com/nimai/mptcp-scapy*. The Python script that can be used to carry out the ADD_ADDR attack can be found here: *https://github.com/fabriziodemaria/MPTCP-Exploit*.

It is appropriate to mention here some of the limitations of the tool (that are examined more in details in Section 4.4: *Limitations and future work*): the tool has been designed to hijack a specific kind of communication involving client and server sending each others text messages using the tool *netcat*. It is very unlikely that the procedure completes with other kind of MPTCP connection setups between client and server. Nevertheless, this specific exploit serves well our purpose of assessing the danger and feasibility of the ADD_ADDR attack in general terms. Moreover, this tool simplify the attack procedure by sniffing the SEQ and ACK numbers of the ongoing connection instead of starting a procedure to try and guess the values. Also, the ports in use by the client and the server are retrieved automatically by inspecting the sniffed packets, while the IP addresses have to be provided by the user when launching the attack script. Further considerations about these simplifications can be found in Section 4.4.

The python module *test_add_address.py* in the root of the GitHub repository follows the analysis in RFC7430[] to perform the various steps necessary to hijack the MPTCP connection. All the requirements and theoretical details about this procedure have been

reported in Section 3.3, and this section is limited to show and investigate the actual code implementation.

4.2 Attack script

The very first step to be executed is the following: all the RST outgoing packets that can be generated by the hosting machine (the attacker) must be blocked during the process, when the first phases are completing and no finalized TCP connection can be actually detected by the system. This is done with the commands in Listing 4.2.

```
execCommand("sudo iptables -I OUTPUT -p tcp --tcp-flags ALL RST,ACK -j
    DROP", shell = True)
execCommand("sudo iptables -I OUTPUT -p tcp --tcp-flags ALL RST -j
    DROP", shell = True)
```

Listing 4.2: *Disable RST outgoing packets*

The Scapy built-in *sniff* function allows to retrieve packets from a specific interface, according to a custom filter function *filter_source* that inspects the source address. In this way it is possible to retrieve the IP addresses, ports, SEQ and ACK numbers of the ongoing connection between client and server. The first constructive step of the whole procedure consists in forging of the ADD_ADDR packet using the method *forge_addaddr* (Listing 4.3).

```
def forge_addaddr(myIP, srcIP, srcPort, dstIP, dstPort, sniffedSeq,
    sniffedAck):
    pkt = (IP(version=4L, src=srcIP, dst=dstIP)/ TCP(sport=srcPort,
        dport=dstPort, flags="A", seq=sniffedSeq, ack=sniffedAck,
        options=[TCPOption_MP(mptcp=MPTCP_AddAddr(address_id=ADDRESS_ID,
        adv_addr=myIP))]))
    return pkt
```

Listing 4.3: *forge_addaddr method*

Here comes the first consideration about the script design: once the ADD_ADDR is sent to the victim client, the tool has to be already listening for the MP_JOIN sent back as a response; in order to make sure this happens, multithreading is used to start looking for the MP_JOIN packet even before ADD_ADDR is sent, with the thread named *SYNThread*. *SYNThread* just calls the method *get_MPTCP_syn* in the module *sniff_script.py*, that uses *tcpdump* with a specific filter option. In fact the Scapy *sniff* functionality proves to be unreliable in case of a high flow of packets to be processed and often skips some when the buffers reach their limits. Even if this is fine in other parts of the script where any packet capture is fine to retrieve ACK and SEQ numbers, it is mandatory not to miss the single MP_JOIN+SYN packet sent by the client upon ADD_ADDR reception. This problem concerning the sniffing function of Scapy is also reported in the official website under the section "Known bugs": *May miss packets under heavy load.* Note that this wouldn't be a problem with the slow message exchange of *netcat*, but the script can be also tested with high throughput applications like *iperf*, hence the usage of the more reliable *tcpdump*.

In order to filter out exactly the MP_JOIN packet we are looking for, the following command in Listing 4.4 is used, where *tf* is just a temporary file to store the information and *i* is the interface name passed as a parameter.

```
execCommand("sudo tcpdump -c 1 -w " + tf.name + ".cap -i " + i + "
    \"tcp[tcpflags] & tcp-syn != 0\" 2>/dev/null", shell = True)
```

Listing 4.4: *tcpdump for MP_JOIN*

A similar sniffing procedure is used for the next steps regarding SYN/ACK and ACK MP_JOIN packets, as it can be seen for the threads named *SYNACKThread* and *ACK-Thread*. Each time these sniffing threads are started, a sleep function is called for a time expressed in *THREAD_SYNC_TIME*, as a poor but effective mechanism that ensures that *tcpdump* is called and running in the new threads before proceeding.

The MP_JOIN packets generated and received in this way are manipulated to change the IP addresses and ports (and possibly other fields) as described in the attack procedure and then forwarded to the right host. Note that manipulating packet's fields in Scapy is different with respect to the case of ADD_ADDR where the packet is forged from scratch. All the functions *forge_ack*, *forge_synack* and *forge_syn* actually don't forge a new packet but slightly modify a copy of the received packet. While doing this it is necessary to eliminate the *checksum* value so that Scapy automatically recalculate the correct value for it, taking into consideration the updated values. Similar considerations hold for the Ethernet layer of the manipulated packets.

Once the ACK has been sent to the server, the new subflow is set up. In order to make it more visible, the next steps in the script enable again the outgoing RST packets and forge some of them to close all the subflows apart from the malicious one. By following the *print* messages in the script, this corresponds to *Phase 5*. Now, all the messages from the server to the client are sent to the attacker instead, without an explicit way for the victim to notice.

The very last portion of the script runs the method *handle_payload* that both prints the text messages (payload) received from the server and generate *data_ack* packets for the server in order to keep the connection alive.

The final tool and a step-by-step guide on how to use it can be found in the repository: *https://github.com/fabriziodemaria/MPTCP-Exploit*.

4.3 Reproducing the attack

This procedure has been tested on a Ubuntu 14.04 LTS machine.

1. Open two terminal windows and run the client.sh and server.sh scripts to launch the UML virtual machines (user/password: root)
2. On the server machine, run the following (you can use a TCP Port of your choice here):

   ```
   netcat -l -p 33443
   ```

3. On the client machine, we first need to disable one of the two network interfaces, namely eth1. This is necessary due to some limitations currently affecting the Scapy tool and the attacking script (the connection will still be MPTCP, with a single subflow):

```
ifdown eth1
```

4. Now you can run netcat on the client, too:

```
netcat 10.2.1.2 33443
```

5. Try to exchange messages between client and server to verify that communication is active.

6. Now we can start the attack opening a new terminal on our local machine (it is necessary to start the Scapy script AFTER having established the netcat connection).

7. Go to the folder were you downloaded the Scapy tool and type the following:

```
sudo python test\_add\_address.py 10.1.1.1 \
10.2.1.2 10.1.1.2 tap2 tap0
```

   NOTE: If an import error appears, try to install the missing dependencies with:

```
sudo apt-get install python-netaddr
```

8. Go back to the client UML terminal and start sending messages to the server. You should notice that while the messages exchange goes on, the attacking script progresses.

   IMPORTANT: it might be that the script gets stuck (it shouldn't take more than a few seconds to complete). If that is the case, close netcat and start again from step 2.

9. If you reach 100% in the attack process, just try to send a message from the server to the client and you will notice that the messages are now sent to the attacking machine instead. Further improvements would allow to also answer back to the server, thus impersonating the client.

4.4 Conclusions

The Scapy tool developed for this research targets a specific scenario to exploit the ADD_ADDR vulnerability. It is not intended to be general enough to break all the existing MPTCP implementations. Nevertheless, by succeeding in this specific case involving a *netcat* communication between two hosts, it is indeed proved the feasibility and gravity of the problem, and it should be relatively easy to extend the portability of the attacking tool to act in new scenarios.

This section mainly investigates the workarounds used to simplify the attacking process, to prove that they are not critical enough to devalue the results of the tool itself.

All the requirements for the succeeding of the attack have been already listed in Section 3.3. Here is reported a short summary:

- the four-tuple: IP and port for both source and destination;
- valid ACK/SEQ numbers for the targeted subflow;
- valid address identifier for the malicious IP address used to hijack the connection;

Regarding the last point, the Address ID chosen for the new subflow initiated by the attacker must be different from all the other IDs already used by the other subflows. It is fairly easy to choose a value quite high that has very low probability of being in use already. This value is set to 6 by default in the attack Scapy script.

It is a fair assumption that the four-tuples identifying the connection endpoints are known by the attacker[RFC7430], apart from the client side port value: in that case the difficulty in guessing the right port in use very much depends on the port randomization technique deployed at the client host [RFC6056]. Since it is anyway possible to guess the port, it is a fair simplification to simply provide it to the application in our tests: for this reason the tool has been designed to accept the IP addresses as arguments and automatically gets the ports in use to increase the rate of success in different testing scenarios, without the need for the user to provide that kind of information.

Guessing the SEQ and ACK numbers is by far more complex. Again, all the considerations about this have been reported in previous sections: it is possible to generate a big number of packets trying to guess the acceptable values for packet injection. This is out of scope in this research, so it is acceptable to simplify the attack by providing the SEQ and ACK values (by sniffing them from the ongoing connection).

It is important to emphasize that despite these workarounds, that require to act as a physical man-in-the-middle, no other information apart from the IP addresses, ports, SEQ and ACK values have been retrieved using Scapy's sniff or tcpdump, and no packet originally sent to the trusted hosts have been discarded or modified. All the sniffed values can be guessed and, despite the reduced chance of success, the exploit could be executed via a 100% off-path attack. That is why this is considered a major vulnerability for MPTCP deployment as of RFC7430 indications. In the next sections the solution to this problem and its Linux kernel implementation are discussed in the details.

Chapter 5

Fixing ADD_ADDR

5.1 The ADD_ADDR2 format

There is an ongoing effort to move the current MPTCP specification **??** from Experimental to Standard Track. Solving the ADD_ADDR vulnerability is believed to be a fundamental step to reach the required security standards for the transition to happen. By analyzing the nature of the vulnerability, various proposals have been elaborated to modify the design of the ADD_ADDR option [RFC-7430]. The conceptual flaw behind the option is that no secret material related to the ongoing MPTCP is included. The only security mechanism connected to such message is indeed the TCP-level sequence and acknowledge numbers, that an attacker has to know in order to inject such message into an ongoing session. A possible solution could be to add the receiver token of the connection as a field in the ADD_ADDR option. Such token, exchanged only during connection establishment via the MP_CAPABLE option, is supposed to be unknown to the attacker that in turns would not be able to forge a valid ADD_ADDR message. This solution wouldn't be effective if the attacker is able to eavesdrop the keys during the initial handshake; keys' eavesdrop is indeed a security concern related to MPTCP [ref to section on Keys' Eavesdrop], and for this reason it is not advisable to add such information in clear inside the ADD_ADDR option, since that would give more opportunities for eavesdropping. Another possibility would be to maintain the ADD_ADDR format unchanged but to block the attack at a later stage. For example, if the destination address of the SYN packet is added as part of the message used to calculate the HMAC value, the attacker wouldn't be able to recompute the HMAC value after modifying the destination address. However, since addresses are not a stable piece of information in a network with NATs, using the destination address to calculate the HMAC might not work. In order to achieve higher security levels maintaining NAT compatibility, a third option has been proposed with positive feedback. The idea is to add to the ADD_ADDR option a new field containing the truncated HMAC value (rightmost 64 bits) calculated as follow: the *key* is the MPTCP key of the sender as originally agreed in the MP_CAPABLE handshake; the *message* is the concatenation of the previous three fields in packet: Address ID, advertised IP address, and Port. The new format (figure 5.1 has been formally specified for the first time in RFC-6824bis-04.

Such format would require the attacker to know the key in order to forge a valid ADD_ADDR2 message, but such key is not exposed as in the case of the previous solution. Albeit, if the attacker is able to eavesdrop the keys during connection initiation it would be possible to exploit the same vulnerability even with the new address format. More experiments about this case are reported in section [ref to the experimental evaluation

```
                    1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---------------+---------------+-------+-------+---------------+
|     Kind      |     Length    |Subtype| IPVer |  Address ID   |
+---------------+---------------+-------+-------+---------------+
|          Address (IPv4 - 4 octets / IPv6 - 16 octets)         |
+-------------------------------+-------------------------------+
|   Port (2 octets, optional)   |                               |
+-------------------------------+                               |
|                    Truncated HMAC (8 octets)                  |
|                               +-------------------------------+
|                               |
+-------------------------------+
```

Figure 5.1: ADD_ADDR2 option

section]. Possible mitigations for such threat are explained in section [ref to keys' eavesdrop section 3.3.2].

The keys' eavesdrop threat is a partial-time on-path eavesdrop, a category that is considered less critical in terms of security concerns. Such keys' eavesdrop procedure in MPTCP has an almost identical counterpart in SCTP, when the SCTP-AUTH extension is used without pre-shared keys [RFC-5061]. In these regards the same security levels of SCTP would be reached in MPTCP by upgrading ADD_ADDR to ADD_ADDR2. Since SCTP is Standard Track, ADD_ADDR2 is indeed considered a sufficient modification of the MPTCP first design to reach the security levels required for the transition to Standard Track.

5.2 Implementing ADD_ADDR2

The current MPTCP patch added to the TCP stack in the Linux kernel currently counts around 12000 lines of code [href]. It is considered the reference implementation for MPTCP and it closely follows RFC standards and set of features. Moreover, a lot of effort has been put into the implementation design in order to make the new protocol acceptable for upstream to the official Linux kernel. For such purpose, it is of paramount importance to keep the added complexity into the TCP stack as low as possible, in order not to jeopardize performance and stability of regular TCP. Nevertheless, high performance is expected for MPTCP. The main architectural concepts related to the control plane of the protocol are now explained, before introducing the modifications related to the new ADD_ADDR2 format.

5.2.1 MPTCP in Linux

With MPTCP in the Linux kernel, three main layers are introduced to guarantee multipath management and retro-compatibility with regular TCP [href]. The first element is the *master subsocket*, which provides the interface used by the applications to communicate with the TCP stack. The structure of the master subsocket follows the regular TCP standards, in order to maintain retro-compatibility towards the application layer: in fact this is the only element used by the kernel in case of regular TCP connectivity. The second element is called *multi-path control block (mpcb)* and it is the main brain of MPTCP, handling MPTCP-specific functionalities: the mpcb runs the algorithms that determine when to start or stop subflows, which subflow to chose in order to send a particular piece of data over the network and how to reconstruct the original data from the scattered segments coming from different subflows at the receiver. all the reordering algorithms in the mpcb work at the data-level, while the reordering of the data at the single subflows is handled by the underlying regular TCP. The final element of the MPTCP architecture is the set of *slave subsockets*, the actual endpoints for the multiple MPTCP subflows. Such elements are not visible by the application, but they are handled by the mpcb. The master subsocket and the slave subsockets form the pool of subflows used in MPTCP.

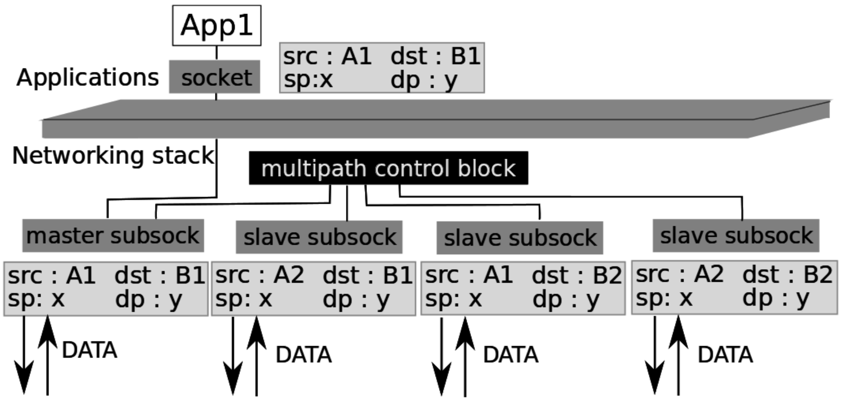

Figure 5.2: MPTCP Linux architecture

Analyzing to the actual code implementation related to such architecture, it is mainly composed of several structures linked by pointers. In order to maintain the design-goal if minimizing the impact over regular TCP, when a TCP structure would need additional elements to handle MPTCP-related functionalities, the common choice is to define a new MPTCP-specific structure to store those elements. In this way, upon regular TCP operations, there would be no increase in memory-footprint and all the standard TCP structures would be in place. On top of that, having specific structures for MPTCP code makes it easier to read and understand the MPTCP parts inserted into the TCP stack. For example, a fundamental structure in TCP is the *tcp_sock* structure, that is used to store the state of a single TCP connection. In MPTCP, additional information for each TCP subflow is needed (for example the Address ID associated to each subflow). A new *mptcp_tcp_sock* struct has been defined and each subflow contains a pointer to such new structure. Also the previously mentioned main architectural element that is the multi-path control block is implemented in code using a new structure called *mptcp_cb*.

The allocation policy for all the new MPTCP structures is lazy-allocation, meaning that MPTCP structures are allocated only if it is detected that both hosts support the new protocol. This choice is again related to the main purpose of not affecting regular TCP when MPTCP fails during negotiation (or later on during the connection lifetime). A downside of this approach is related to the fact that the TCP stack operations are often executed in a soft-interrupt context, that does not allow functions to sleep in order to wait for available memory: this means that memory allocations might fail, forcing a fallback to regular TCP. Nevertheless, connection setup in an MPTCP-compatible environment requires the client to send a first MP_CAPABLE segment: this means that, even if no data structure is allocated during this first stage, the client has to generate a random key, and the related token is also calculated to check that it is not already used to identify another MPTCP connection. A reference to the originated *tcp_sock* structure is saved inside the hashtable used to keep track of the ongoing MPTCP connections. At this point, such *tcp_sock* is called "meta-socket". After that, if the SYN/ACK from the server does not

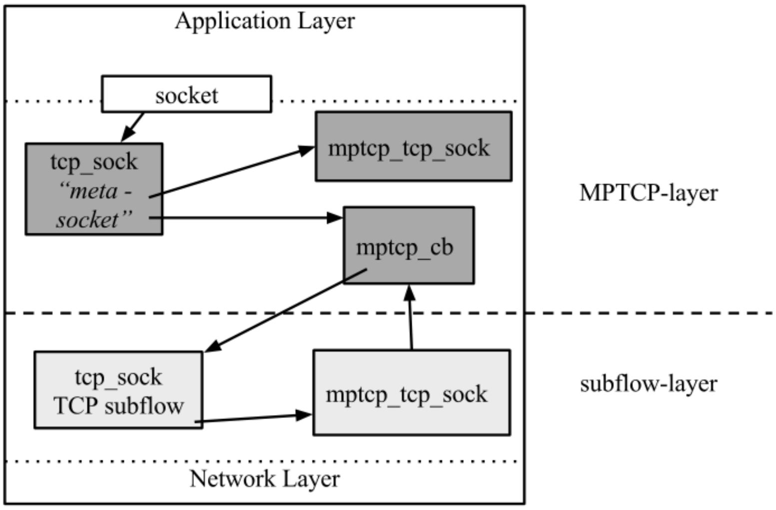

Figure 5.3: MPTCP high-level data structures with relative references

contain a valid MP_CAPABLE option, then the client simply removes the reference to the meta-socket from the hashtable before proceeding with a regular TCP handshake procedure. If the server side is MPTCP-compatible and the MP_CAPABLE is indeed present in the incoming TCP segment, then the MPTCP structures are allocated: *mptcp_cb* and *mptcp_tcp_sock*. At the server, the status of the connection is not fully operational until the final ACK from the client as required by the TCP three-way handshake. When the SYN packet with MP_CAPABLE option is processed, a *request_sock* data structure is allocated that has some additional space for MPTCP-related information with respect to the same structure used for regular TCP option. In case no MP_CAPABLE option is present in the received SYN packet, the regular TCP version of the *request_sock* data structure is used, following the same design principles previously explained. The random key and uniqueness of the token are procedures executed at the server in a similar way with respect to the client. If all the MPTCP initialization procedure proceeds as expected, when the server receives the ACK from the client the master-socket is ready and linked to the *mptcp_cb*. With the Linux kernel implementation of MPTCP, only the client is allowed to start the establishment of new subflows. There are two main reasons for this: if both client and server start a subflow at the same time, it can be that multiple path are established between the same pair of IP addresses, which can cause problems in some cases. Moreover, clients are often operating behind a NAT, which don't allow server to start new TCP sessions and consequently, they wouldn't allow the server to start a new subflow with the client.

After the connection has been established, multiple subflows can be used with MPTCP and there is a modular path-manager interface to allow flexibility in the heuristic adopted to decide which interfaces can be used and in which manner. In creating a new subflow, the client has to add the MP_JOIN option inside the SYN packet and, differently from the MP_CAPABLE scenario, the MPTCP-related structures like *mptcp_tcp_sock* are created early on. At this point, failure wouldn't cause fallback to regular TCP and there is no need to risk memory allocation failures upon reception of the SYN/ACK from the server. Even if the subflows in MPTCP resemble regular TCP connections, the initial handshake differs in a way that it now requires four steps to reach fully operational status. The reason for this is that the third ACK now contains the HMAC value calculated by the client that has to be verified and acknowledged by the server before any data can be transmitted on such subflow. Regarding the operational flow in the stack upon reception of a SYN packet, there is no early inspection aimed at determining if an MP_JOIN option is present: that would causes performance degradation in case of regular TCP SYN packets. Instead, the packet is processed with regular TCP stack until, in case of matching with a listening socket, the function *tcp_v4_conn_request()* is called: here the TCP options are scanned, and if MP_JOIN is present then redirection to MPTCP happens, and the lookup in the hashtable is performed to determine which MPTCP connection the new SYN packet is addressing to. As a new addition required by MPTCP, if there is no matching socket found for the incoming packet, MPTCP still checks if the SYN message contains the MP_JOIN option via the *mptcp_lookup_join*. At this point, the server creates a request socket that is saved into the hashtable so that it can be retrieved when the client answers with the ACK message during the last stage of the subflow handshake.

This section presented an overview of the most important data structures and functions used in MPTCP to handle connection establishment and subflow management. The following sections will deal with the ADD_ADDR functionality in the Linux kernel and how this was modified to implement the new format ADD_ADDR2.

5.2.2 Retro-compatibility

Version control mechanism was not present but it is needed to negotiate which format to use in a MPTCP session: ADD_ADDR or ADD_ADDR2.

5.2.3 Port advertisement

Port advertisement in ADD_ADDR is possible according to RFC specifications but it was not part of the implementation at the beginning of the thesis work, so it has been added.

5.2.4 IPv6 considerations

Longer addresses brought some issues related to TCP option fields limitations.

5.2.5 Crypto-API in MPTCP

A major problem was how to deal with the new hashing requirements introduced by ADD_ADDR2. Extending the current MPTCP hashing function to deal with input messages of arbitrary size is a first point to explain. The second part has to deal with the whole analysis related to adopting the kernel CRYPTO APIs to calculate the HMAC values in MPTCP and why this is not advisable.

5.3 Other contributions

Another minor part of the thesis work on MPTCP is related to some small contributions to the official RFC documentation and other open-source projects.

5.4 Experimental evaluation

This part should include performance analysis regarding the new format introduced with ADD_ADDR2. A discussion on how the new format (and all the other modifications introduced with the patches) could impact any aspect of the protocol should be present in this section. It is possible to add here the other possible solutions for ADD_ADDR fix, and why they are not good enough.

Chapter 6

Conclusions

6.1 Related work

References to all the related work, including all the efforts to make MPTCP secure and stable, can be reported here.

6.2 Future work

Here goes the list of the next steps to be taken care of in terms of MPTCP security, in order to facilitate the protocol's upstream and its widespread deployment. This section can also include a more specific discussion on the aspects to be still analysed regarding the new format ADD_ADDR2.

6.3 Final thoughts

Some final conclusion.

Appendix A

An appendix

Appendix content goes here...