

Multipath TCP Working Group  
Internet-Draft  
Intended status: Experimental  
Expires: May 7, 2020

M. Amend  
Deutsche Telekom  
J. Kang  
Huawei  
November 4, 2019

Multipath TCP Extension for Robust Session Establishment  
draft-amend-mptcp-robe-01

Abstract

Multipath TCP extends the plain, single-path limited, TCP towards the capability of multipath transmission. This greatly improves the reliability and performance of TCP communication. For backwards compatibility reasons the Multipath TCP was designed to setup successfully an initial path first, after which subsequent paths can be added for multipath transmission. For that reason the Multipath TCP has the same limitations as the plain TCP during connection setup, in case the selected path is not functional.

This document proposes a set of implementations and possible combinations thereof, that provide a more Robust Establishment (RobE) of MPTCP sessions. It includes RobE\_TIMER, RobE\_SIM, RobE\_eSIM and RobE\_IPS.

RobE\_TIMER is designed to stay close to MPTCP in that standard functionality is used wherever possible. Resiliency against network outages is achieved by modifying the SYN retransmission timer: If one path is defective, another path is used.

RobE\_SIM and RobE\_eSIM provides the ability to simultaneously use multiple paths for connection setup. They ensure connectivity if at least one functional path out of a bunch of paths is given and offers beside that the opportunity to significantly improve loading times of Internet services.

RobE\_IPS provides a heuristic to select properly an initial path for connection establishment with a remote host based on empirical data derived from previous connection information.

In practice, these independent solutions can be complementary used. This document also presents the design and protocol procedure for those combinations in addition to the respective stand-alone solutions.

## Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on May 7, 2020.

## Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

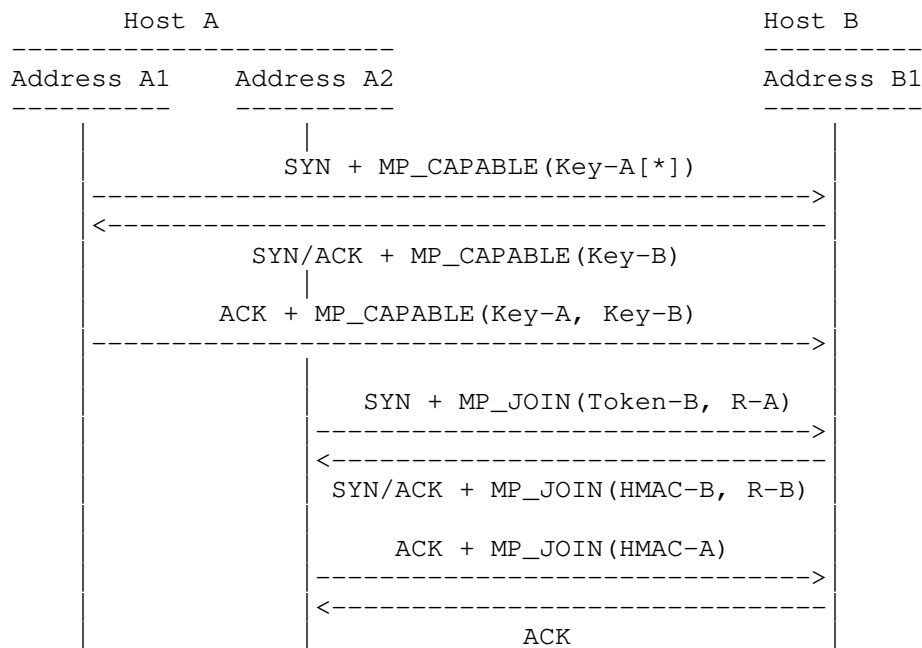
1. Introduction . . . . .	3
1.1. Terminology . . . . .	6
2. Implementation without MPTCP protocol adaptation . . . . .	7
2.1. Re-transmission Timer(RobE_TIMER) . . . . .	7
2.2. Simultaneous Initial Paths Simple Version (RobE_SIM) . . . . .	8
2.3. Heuristic Initial Path Selection (RobE_IPS) . . . . .	9
2.3.1. Architecture . . . . .	9
2.3.2. Typical Scenarios . . . . .	10
2.3.3. Path decision information . . . . .	13
2.3.4. Initial Path Selection use local RTT information . . . . .	14
2.4. Combination of RobE_SIM and RobE_IPS . . . . .	14
2.5. Combination of RobE_TIMER and RobE_IPS . . . . .	15
3. Implementation with Bi-directional MPTCP Support . . . . .	16
3.1. Simultaneous Initial Paths Extended Version (RobE_eSIM) . . . . .	16

3.1.1.	RobE_eSIM implicit Negotiation and Procedure . . . .	17
3.1.2.	RobE_eSIM explicit Negotiation and Procedure . . . .	18
3.1.3.	Protocol Adaptation . . . . .	19
3.1.4.	Fallback Mechanisms . . . . .	20
3.1.5.	Comparison RobE_SIM and RobE_eSIM . . . . .	22
3.1.6.	Security Consideration . . . . .	23
3.2.	Heuristic Initial Path Selection with remote RTT Measurement . . . . .	23
3.2.1.	Description . . . . .	23
3.2.2.	Protocol Adaptation . . . . .	24
3.2.3.	Fallback Mechanism . . . . .	25
3.2.4.	Security Consideration . . . . .	25
4.	IANA Considerations . . . . .	25
5.	Informative References . . . . .	26
	Authors' Addresses . . . . .	26

## 1. Introduction

Multipath TCP Robust Session Establishment (MPTCP RobE) is a set of extensions to regular MPTCP [RFC6824] and its next upcoming version [I-D.ietf-mptcp-rfc6824bis], which releases single path limitations during the initial connection setup. Several scenarios require and benefit from a reliable and in time connection setup which is not covered by [RFC6824] and [I-D.ietf-mptcp-rfc6824bis] so far. MPTCP was designed to be compliant with the TCP standard [RFC0793] and introduced therefore the concept of an initial TCP flow while adding subsequent flows after successful multipath negotiation on the initial path. While fulfilling its purpose, MPTCP is however fully dependent on the transmission characteristics of the communication link selected for initiating MPTCP.

Figure 1 shows the traditional way of MPTCP handshaking with a MP\_CAPABLE exchanged first, followed when successful negotiated by additional flows engaging MP\_JOIN. [RFC6824] and the upcoming next MPTCP [I-D.ietf-mptcp-rfc6824bis] differ in that a Key-A is sent with the first MP\_CAPABLE or not.



[\*] Key-A in the first MP-capable is related to RFC6824 only and does not exist in RFC6824bis.

Figure 1: MPTCP connection setup

Multipath TCP itself enables hosts to exchange packets belonging to a single connection over several paths. Implemented in mobile phones (UEs), these paths are usually assigned to different network interfaces within the UE and corresponds to different networks such as cellular and WiFi. The path or network interface for initiating the initial subflow setup is most often provided by the operation system of the UE. For example, if a cellular connection and WiFi are present in a mobile phone, WiFi is usually the interface offered to initiate the MPTCP session.

This design falls short in situations where the default path does not provide the best performance compared to other available paths. In a worst case the default path is not even capable of setting up the initial flow letting any other functional path unused. For example, if the WiFi signal is weak, broken or cannot forward traffic to the destination, the establishment of the subflow will be delayed or impossible. This in turn, leads to a longer startup delay or no communication at all for services using MPTCP even if other functional paths are available. Even in scenarios where all paths

are functional but services would benefit from a setup over the path with the lowest latency, MPTCP has no mean to support this demand.

It can be concluded, that sequential path establishment relying with an initial path establishment over an external given default route will result in experience reduction when using MPTCP. So this document proposes solutions to overcome the aforementioned limitations and provides a more robust connection setup compared to traditional MPTCP.

RobE\_SIM and RobE\_eSIM aims to overcome the limitations of [RFC6824] and [I-D.ietf-mptcp-rfc6824bis], using one initial flow and introduces the concept of potential initial flows triggered simultaneously. Potential initial flows gives the freedom to use more than one path to request multipath capability and select the initial flow at a later point. RobE\_SIM is a break-before-make mechanism, guaranteeing at least the robust connection establishment, however the RobE\_eSIM reuses every potential initial flow request to combine it with less overhead and accelerated multipath availability, leveraging a new MPTCP option MP\_JOIN\_CAP. From a standardization perspective, the RobE\_SIM is fully compliant with [RFC6824] and [I-D.ietf-mptcp-rfc6824bis] and is herein more of a descriptive and procedural nature. The RobE\_eSIM requires a new MPTCP option but with the potential to significantly improve the MPTCP experience.

For the limitation of the default initial path, RobE\_IPS makes no changes to standard MPTCP procedure and improves the performance of connection establishment by introducing an initial path selection strategy and required algorithms. The input for strategy and algorithms is the transmission status information which represents the transmission performance of each available path or network interface. The transmission status information is characterized by at least one of the parameters: signal strength, throughput, round-trip time (RTT) and link success rate. In this way, a path with better transmission performance can be learned and determined and the respective network interface can be used for connection establishment.

The most simple approach for a robust MPTCP session establishment is RobE\_TIMER, iterating the process of initial path establishment over all available paths, if the previous try has failed. Triggering a new try on a next path is depending on an expiration timer, preferably re-use TCP's in-built expiration timer.

Table 1 summarizes the impact of RobE\_TIMER, RobE\_SIM, RobE\_eSIM and RobE\_IPS compared to [RFC6824] and [I-D.ietf-mptcp-rfc6824bis].

Scenario	MPTCP	RobE_TIMER	RobE_SIM	RobE_eSIM	RobE_IPS
IP packet loss	Delayed connection	In the scope of timer	No impact	No impact	Delayed connection
IP broken	No connection	In the scope of timer	No impact	No impact	No connection
IP setup duration dependency	Default route	Default route (+ path 1..n)	Fastest path	Fastest path	Selected path
MP availability duration	MP_CAPABLE HS + MP_JOIN HS	sum <sub>1..n</sub> (MP_CAPABLE <sub>n</sub> HS) + MP_JOIN HS	MP_CAPABLE HS + MP_JOIN HS	max(MP_CAPABLE <sub>1</sub> .. MP_CAPABLE <sub>n</sub> HS)	MP_CAPABLE HS + MP_JOIN HS
Guaranteeing session setup	Depends on the default route	Yes	Yes	Yes	Depends on selection

IP: Initial Path; MP: Multi-Path; HS: Handshake

Table 1: Overview RobE features during initial connection setup

### 1.1. Terminology

This document makes use of a number of terms that are either MPTCP-specific or have defined meaning in the context of MPTCP, as follows:

**Path:** A sequence of links between a sender and a receiver, defined in this context by a 4-tuple of source and destination address/port pairs.

**Subflow:** A flow of TCP segments operating over an individual path, which forms part of a larger MPTCP connection. A subflow is started and terminated similar to a regular TCP connection.

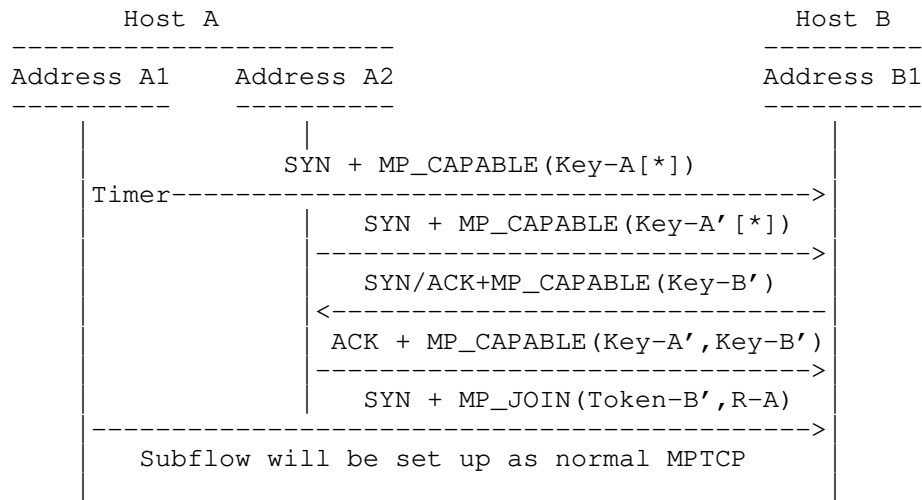
## 2. Implementation without MPTCP protocol adaptation

RobE\_TIMER, RobE\_SIM and RobE\_IPS are compatible with the current MPTCP protocol definitions in [RFC6824] and [I-D.ietf-mptcp-rfc6824bis] but may lack of the full optimization potential which require protocol adaptation in Section 3. Following sections will describe them in detail.

### 2.1. Re-transmission Timer (RobE\_TIMER)

In RobE\_TIMER, a new connection is initiated by sending a SYN+MP\_CAPABLE along the initial path. If this path is functional, the solution will perform identical to classic MPTCP: the initial flow will be established, and subsequent flows can be created afterwards. If however the initial path is faulty, the retransmission will be triggered on another path. This path might circumvent the dysfunctional network, and allow the client to create an initial subflow. The first path is now seen as a subsequent path and the client sends SYN+MP\_JOIN messages to create a subsequent flow.

In high latency networks, the initial SYN+MP\_CAPABLE might be delayed until the client retries on another path. Once the second SYN arrives at the server, it will try to complete the three-way handshake. If the first SYN was delayed by more than the retransmission time plus half a Round Trip Time (RTT) of the second path, it will arrive at the server after the second SYN. The server could now treat the segment as obsolete and drop it.



[\*] Key-A in the first MP-capable is related to RFC6824 only and does not exist in RFC6824bis.

Figure 2: The RobE\_TIMER Solution

Immediately after sending the final ACK of the initial handshake, subflows are established on the remaining paths as defined in [RFC6824] and [I-D.ietf-mptcp-rfc6824bis]

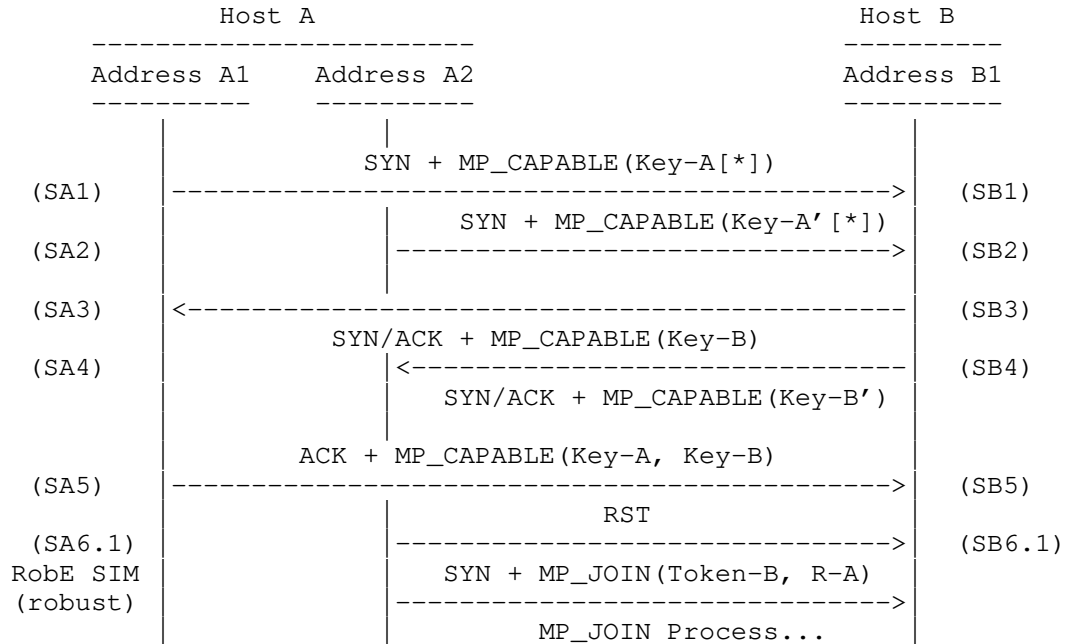
[Notes: How to set the Timer is TBD. If there is the case that the first SYN on default path arrives earlier than that from the second path, the MPTCP connection will be initialized on the path of the first SYN. The server could treat the second SYN as obsolete and drop it.]

## 2.2. Simultaneous Initial Paths Simple Version (RobE\_SIM)

RobE\_SIM is a sender only implementation and no negotiation is required. In RobE\_SIM, the MPTCP connection setup benefits from the fastest path. As shown in Figure 3, host A initiates the connection handshake on more than one path independently (SA1 and SA2). The paths selected for RobE\_SIM and referred to as potential initial flows, can belong to the number of interfaces on the device or a subset selected on experience. When Host A receives the first SYN/ACK back from Host B (SA3), the path carrying this message is identified as the normal initial path. Host A sends then immediately a TCP RST message (SA6.1) on any other path used for simultaneous connection setup causing an immediate termination of assigned flows (break-before-make). The terminated ones are merged as subsequent subflows following the JOIN procedure described in [RFC6824] and



[I-D.ietf-mptcp-rfc6824bis]. The process is equivalent to any other scenario where the SYN/ACK arrives on an other path than depicted in Figure 3.



[\*] Key-A in the first MP-capable is related to RFC6824 only and does not exist in RFC6824bis.

Figure 3: MPTCP RobE\_SIM Connection Setup

### 2.3. Heuristic Initial Path Selection (RobE\_IPS)

#### 2.3.1. Architecture

Figure 4 provides the architecture for RobE\_IPS and employs an "Initial Path Selection" logic which can be integrated into the MPTCP stack or exists as an isolated module in the terminal. The IPS logic has access to a set of transmission status information for each available path or its belonging network interfaces. When an application starts a first communication, IPS selects based on the available path transmission characteristics the path with the highest probability to succeed.

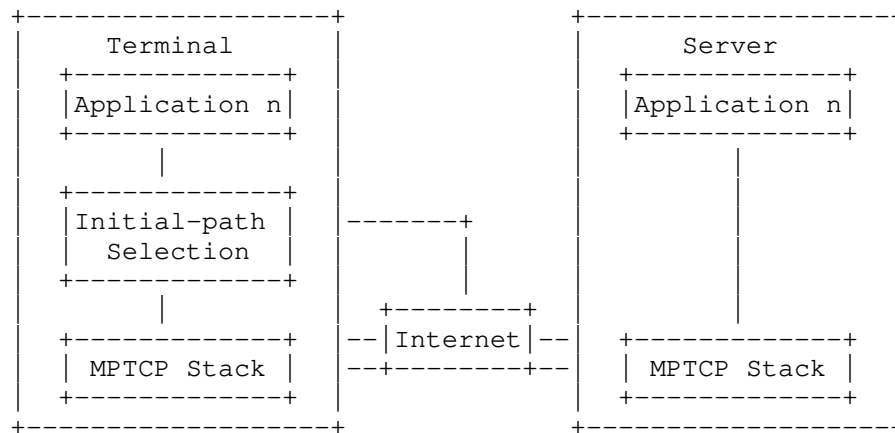


Figure 4: Architecture for Initial-path Selection

### 2.3.2. Typical Scenarios

Two typical RobE\_IPS scenarios are presented in this section. Figure 5 shows that the "Initial Path Selection" logic executed for each MPTCP connection establishment. On the other hand Figure 6 describes that "Initial Path Selection" in case no path information are available. Considering the fact that no heuristics are given before a recent MPTCP connection was established, the default initial path can be adopted. Further combinations and implementations with more or less sophisticated heuristics are possible.

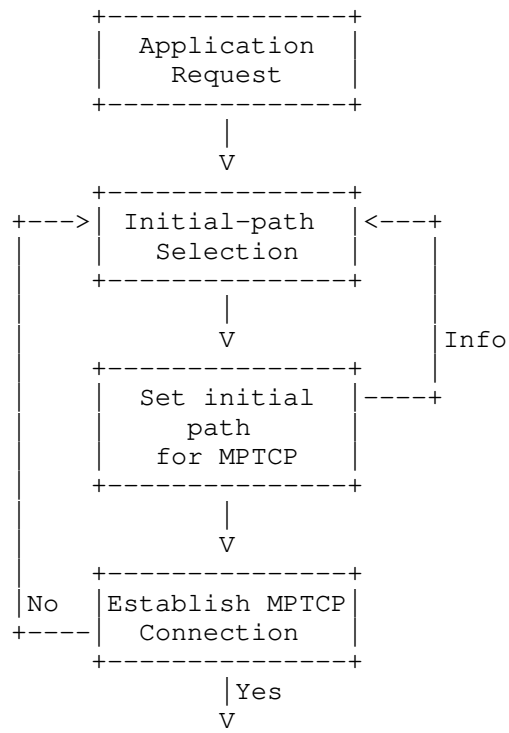


Figure 5: RobE\_IPS for each connection establishment

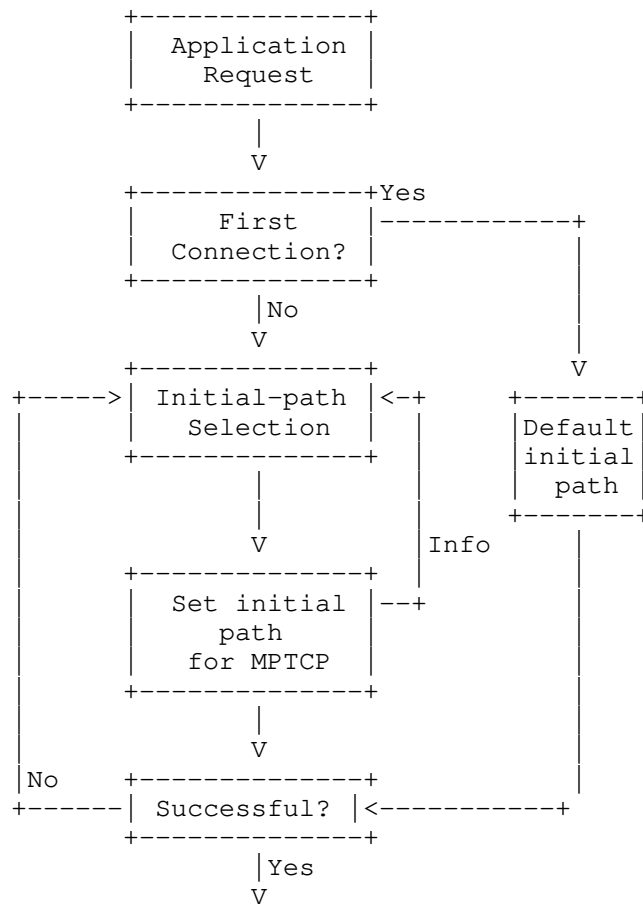


Figure 6: RobE\_IPS using default route when no meaningful heuristic available

Figure 7 shows the process flow of "Initial Path Selection". Upon a request from an application, the IPS logic will acquire transmission status information which represents the transmission performance of each available path or network interface and evaluate it. The transmission status information is characterized by at least one of the parameters: signal strength, throughput, round-trip time (RTT) and link success rate. In this way, the path with the best transmission performance can be determined and used for connection establishment.

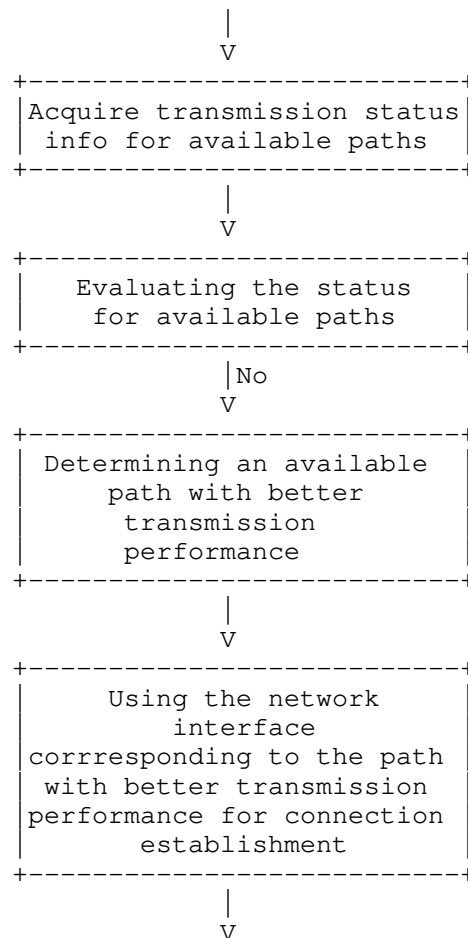


Figure 7: Implementation process for Initial Path Selection

### 2.3.3. Path decision information

The level of heuristic can be mainly divided into three layer: application level, transport-layer level and link-layer level based on the information acquisition method. For example, RTT can be calculated for each path within a MPTCP connection and belongs thereof to the transport-layer level. The transmission status information for each available path SHOULD be characterized by at least one of the parameters: signal strength, throughput, RTT and link success rate. Application level information are more seen for statistical purposes.

- o Application level: application name, domain name, port number and location.
- o Transport-layer level: RTT, CWND, Error rate.

#### 2.3.4. Initial Path Selection use local RTT information

Figure 8 presents a "Initial Path Selection" logic based on RTT, e.g. assuming two paths over LTE and WiFi access. RTT calculation on the transport layer usually reflect the time when an information is sent and an related acknowledgment received. For an asymmetric usage (e.g. download only) of a communication it might happen that recent RTT calculation is only available on sender side which is possibly not the side which employs the IPS logic. A solution for this can be found in Section 3.2. Instead of using the most recent RTT value of a path a filtered value consisting of several measured RTTs can be used. A RTT can also be derived from link layer information but may has a limited meaning when it does not picture the end-to-end latency.

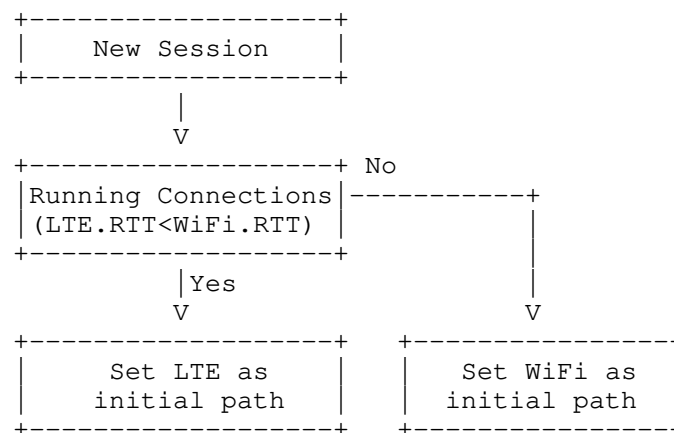


Figure 8: Initial-path Selection based on RTT

#### 2.4. Combination of RobE\_SIM and RobE\_IPS

In an implementation, a single solution may not be sufficient to get an expected behavior. Combination of approaches to improve robustness is recommended therefore. Figure 9 shows the combination of RobE\_SIM and RobE\_IPS. RobE\_SIM can be used at the very beginning when the sender is absent of path information followed by RobE\_IPS for consecutive connections.

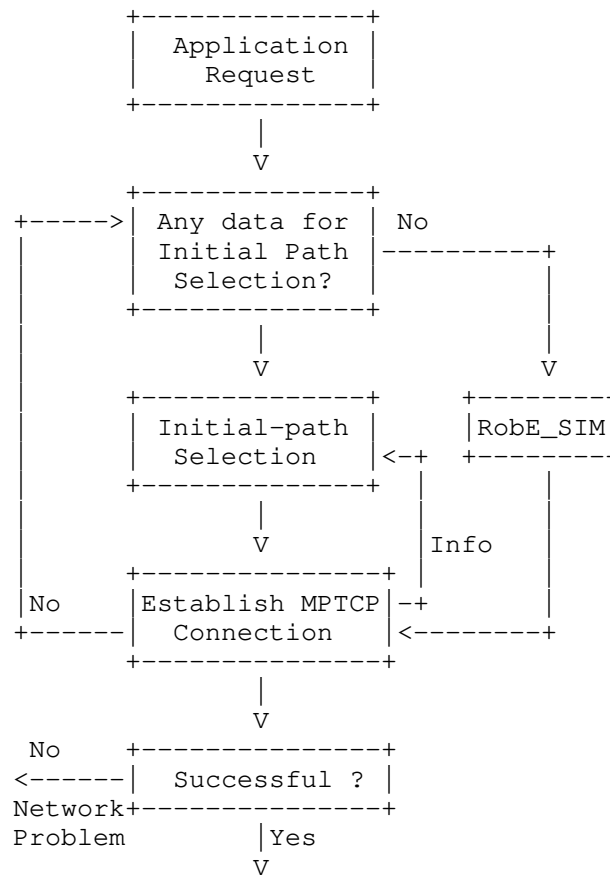


Figure 9: Combination of RobE\_SIM and RobE\_IPS

## 2.5. Combination of RobE\_TIMER and RobE\_IPS

Since RobE\_IPS solely does not guarantee that session can be set up on the selection of initial path, it can also be combined with RobE\_TIMER which generates less overhead compared to the combination with RobE\_SIM in Section 2.4 and guarantees session setup. RobE\_TIMER can be introduced to optimize the control of path switching when the initial path selected by RobE\_IPS is dysfunctional. When the system enables RobE\_IPS and uses the selected initial path for session establishment, it sets the timer for path switching. When timer is expired, the system will change to another path to re-establish connection according to Section 2.1.

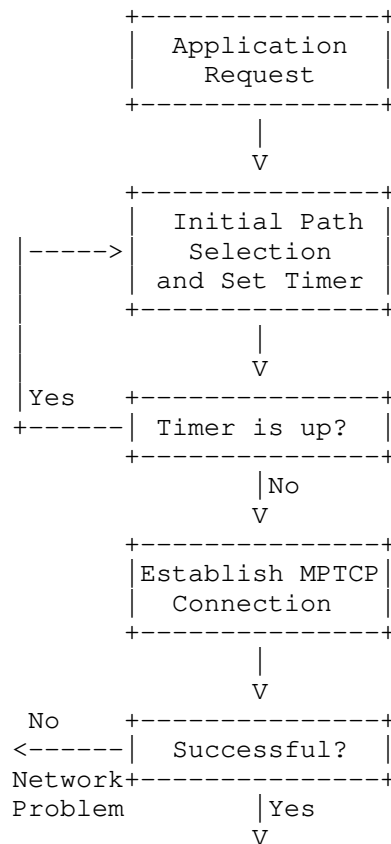


Figure 10: Combination of RobE\_Timer and RobE\_IPS

### 3. Implementation with Bi-directional MPTCP Support

Solutions which requires bi-directional support between two MPTCP hosts promise to have better and possibly more features. However, they cannot be defined without extending current standards in [RFC6824] and [I-D.ietf-mptcp-rfc6824bis]. The RobE\_SIM and RobE\_IPS approach are both capable of profit from an explicit support of the remote end host and defined within this section.

#### 3.1. Simultaneous Initial Paths Extended Version (RobE\_eSIM)

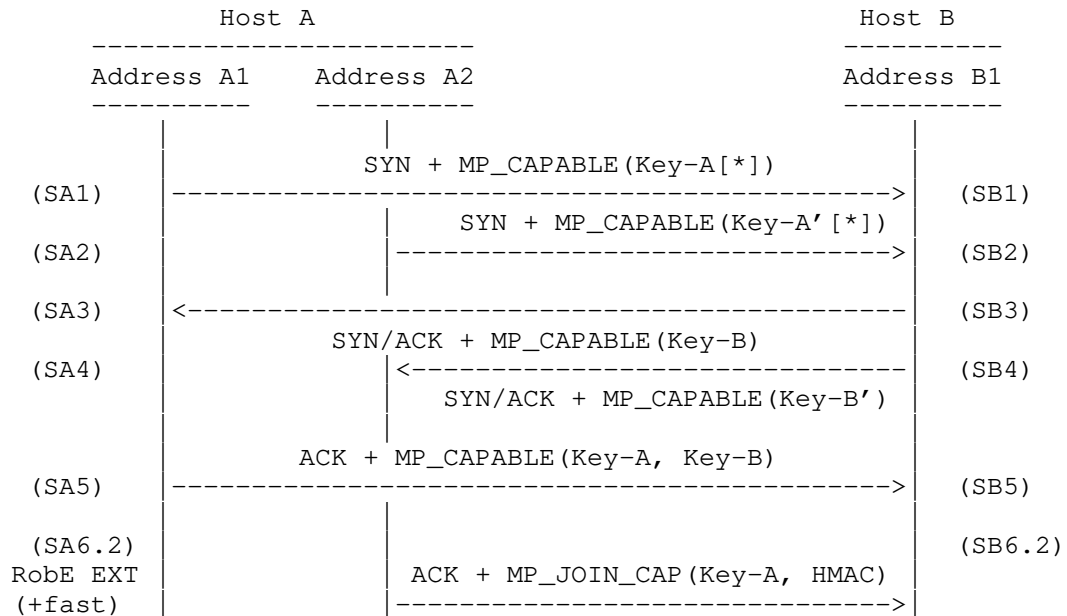
RobE\_eSIM extends RobE\_SIM by reusing the potential initial flows. This eliminates the overhead from RobE\_SIM by introducing a new option MP\_JOIN\_CAP and accelerate the transmission speed by early availability of multiple paths. Further it relaxes the dependency on a reliable third ACK of the 3-way handshake in



[I-D.ietf-mptcp-rfc6824bis]. Remote endpoint support can be negotiated in two ways, an implicit in Section 3.1.1 or explicit in Section 3.1.2

### 3.1.1. RobE\_eSIM implicit Negotiation and Procedure

Similar to RobE\_SIM in Section 2.2 the establishment process of [RFC6824] or [I-D.ietf-mptcp-rfc6824bis] is applied independently on multiple path simultaneously. In Figure 11 this is shown in SA1 and SA2. The first path which returns a SYN/ACK (e.g. SA3) is selected as the initial path and proceed with the traditional establishment process (SA5). Any other path which has to send the final ACK of the 3-way handshake includes a new option MP\_JOIN\_CAP (see definition in Section 3.1.3.2) instead of a MP\_CAPABLE (SA6.2).



[\*] Key-A in the first MP-capable is related to RFC6824 only and does not exist in RFC6824bis.

Figure 11: MPTCP RobE\_eSIM implicit Connection Setup

Following the possible process in Figure 11, two further constellations are imaginable and elaborated below.

1. In the flow diagram Figure 11, A1<->B1 is assumed to be the initial flow. A2<->B1 shall be recycled and the ACK is sent with MP\_JOIN\_CAP. Furthermore, the MP\_CAPABLE arrives first at Host B

(SB5) and the MP\_JOIN\_CAP afterwards (SB6.2). When the MP\_JOIN\_CAP is received, Host B has to iterate over the connection list once (like MP\_JOIN) and check for Key-A availability. If a Key-A connection is found, this one is validated against the HMAC value. The validation has two reasons: first, several Key-A can exist, because different hosts may choose the same Key-A by accident. Furthermore, no one can join a connection by just recording/brute-forcing Key-A and duplicating the request.

2. Like above, but MP\_JOIN\_CAP arrives before last MP\_CAPABLE at Host B

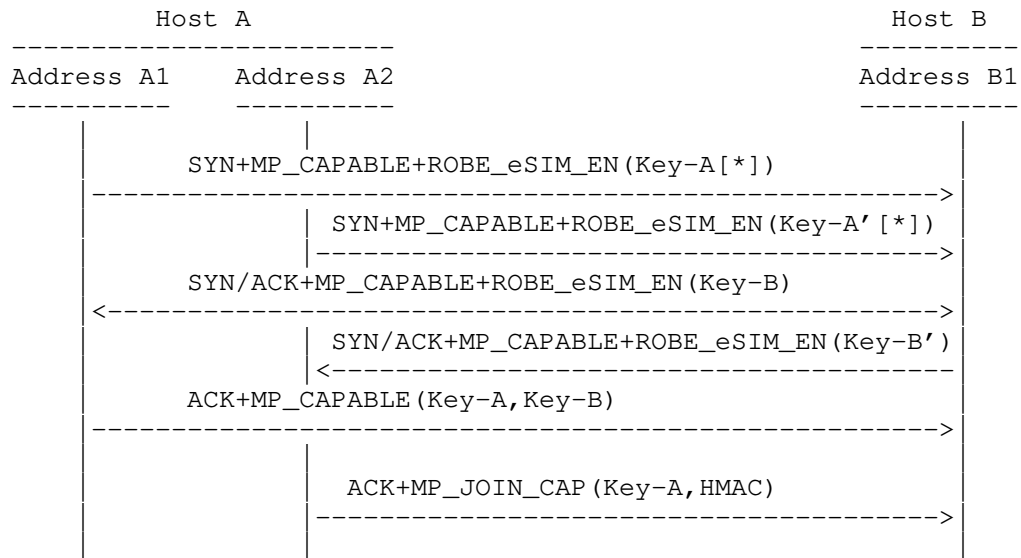
- \* [I-D.ietf-mptcp-rfc6824bis]; Based on Key-A, Host B will iterate over the connection list, but it will not find a match, because Key-A of the previous selected initial flow (SA3, SA5) has not arrived yet. So it will continue with a fast iteration only over the connections which are still in establishment phase using the 10 bit Key-B fast hash ( $\text{crc16}(\text{Key-B}) \& 0x3FF$ ). If it matches against a (precomputed) existing Key-B\_fast\_hash in the connection list, it will validate the request using the  $\text{HMAC}(\text{Key-A+B+B'})$  to ensure legitimation. If successful, both, the initial flow and the MP\_JOIN\_CAP flow, can be immediately established. This is true, because without the knowledge of Key-B, Host A could not calculate the HMAC. So it is clear, that Host A had received the SYN/ACK (SB3). This also mitigates the exchange of a reliable ACK during the handshake process. MPTCP sends the Key-A only with the last ACK and therefore prevents subsequent flow establishment until successful reception at Host B. Using RobE\_EXT, the reception of a MP\_JOIN\_CAP ([I-D.ietf-mptcp-rfc6824bis]) is sufficient to establish both, the path carrying Key-B and Key-B'.

- \* [RFC6824]; Can match based on Key-A, same effort as for a MP JOIN.

3. A2<->B1 is selected as initial flow, because the respective SYN/ACK returns earlier at Host A. It is the same as above, just the other way round.

### 3.1.2. RobE\_eSIM explicit Negotiation and Procedure

The process of an explicit negotiation of RobE\_eSIM follows Figure 11 but uses the ROBE\_eSIM\_EN option Figure 13 additionally during the handshake procedure.



[\*] Key-A in the first MP-capable is related to RFC6824 only and does not exist in RFC6824bis.

Figure 12: MPTCP RobE\_eSIM explicit Connection Setup

### 3.1.3. Protocol Adaptation

#### 3.1.3.1. ROBE\_eSIM\_EN Option

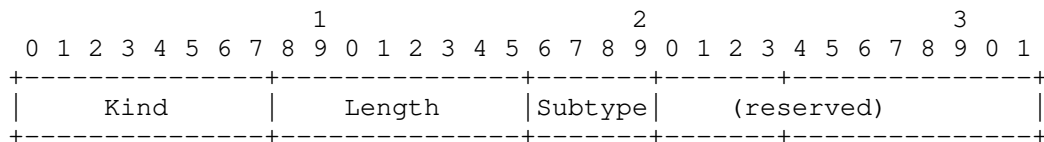


Figure 13: ROBE\_eSIM\_EN\_OPTION

#### 3.1.3.2. MP\_JOIN\_CAP Option

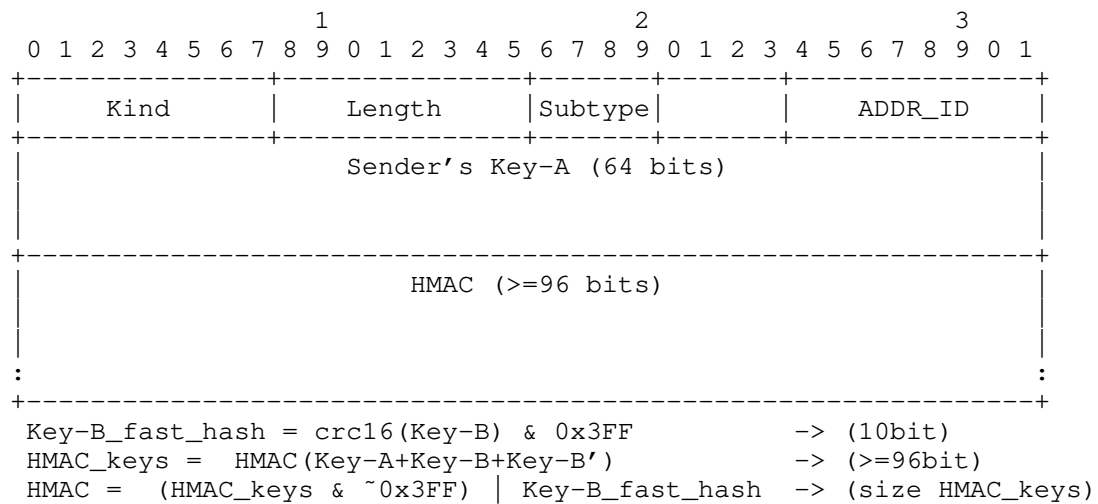


Figure 14: MP\_JOIN\_CAP

Computational effort on receiver side is most often expected to be the same as with MP\_JOIN. Key-A ensures identification of related flows Key-B\_fast\_hash enables MP session even when selected initial flow is not fully established yet (slight computational overhead). HMAC authenticates relationship of initial and potential initial flows.

#### 3.1.4. Fallback Mechanisms

##### 3.1.4.1. Fallback mechanism for implicit RobE\_eSIM

[TBD]

##### 3.1.4.2. Fallback mechanism for explicit RobE\_eSIM

This mechanism considers that both sides support MPTCP capability but the receiver does not equipped with RobE\_eSIM. MPTCP session with RobE\_eSIM negotiation will seamlessly fallback to normal MPTCP process.

[Requires further check how an unaware Host B reacts on possible ROBE\_eSIM\_EN; Ignore or RST? See also RFC6824 Sec. 3.6 "Should fallback [...] the path does not support the MPTCP options"]

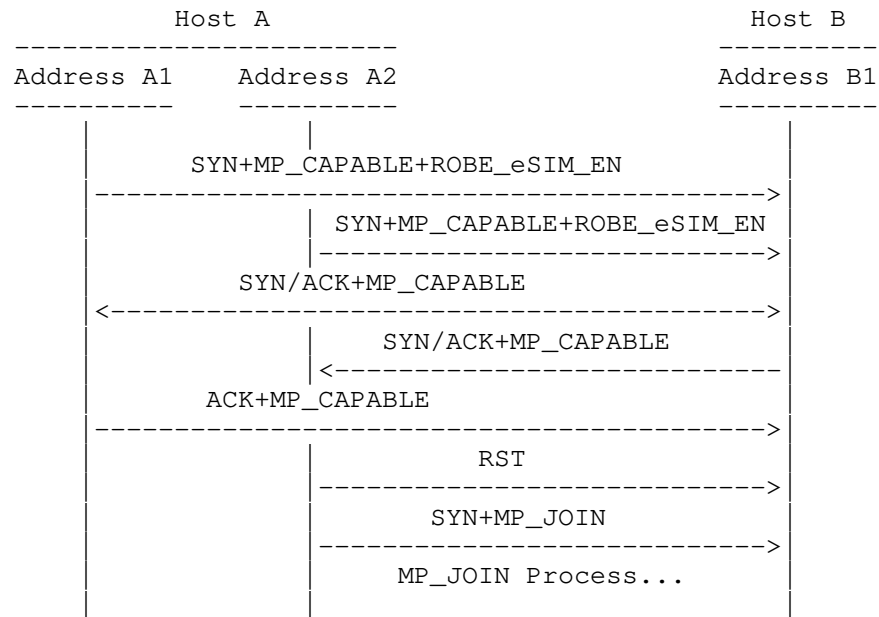


Figure 15: Fallback to MPTCP when missing RobE\_eSIM support

#### 3.1.4.3. Fallback to regular TCP when missing MPTCP support

When the receiver is not MPTCP enabled, MPTCP session with RobE\_eSIM negotiation will seamlessly fallback to regular process which is illustrated in this section.

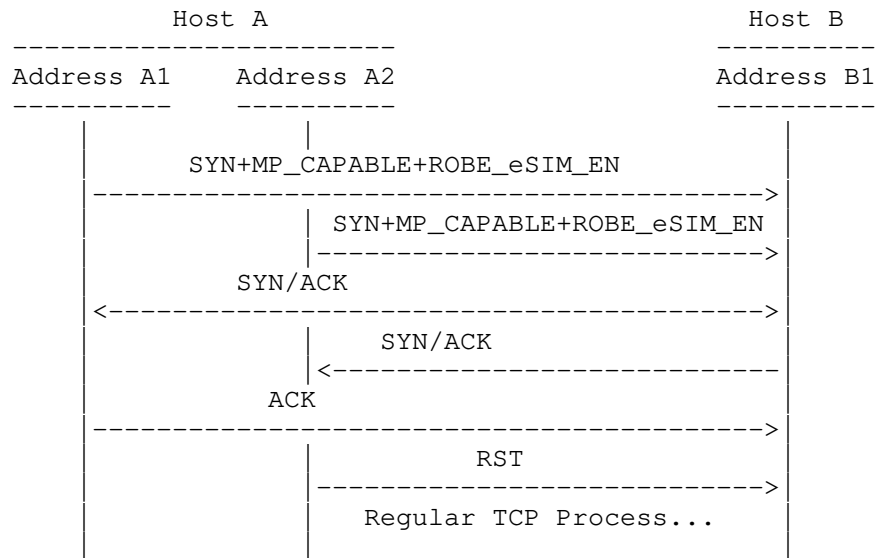
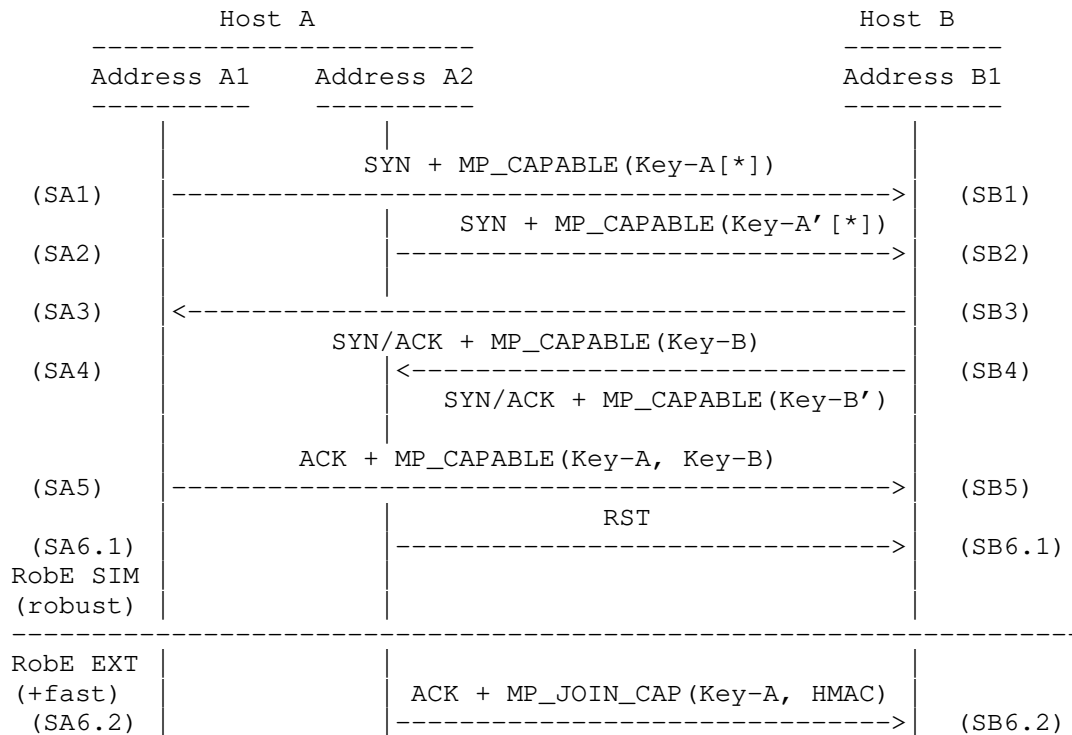


Figure 16: Fallback to TCP without MPTCP support

### 3.1.5. Comparison RobE\_SIM and RobE\_eSIM

Potential initial flows in RobE\_SIM Section 2.2 and RobE\_eSIM Section 3.1 guarantee MPTCP session establishment if at least one selected path for session establishment is functional. Figure 17 makes the differences between both approaches visible and points to the latest decision possibility during session setup when RobE\_SIM or RobE\_eSIM can be selected. Until SA5 in Figure 17 traditional MPTCP connection setup is independently applied on multiple paths simultaneously and offers to select the initial flow later (potential initial flows). The final decision which path is selected as the main one and the handling of the remaining flow(s) differs in SA6.1 when RobE\_SIM is applied or instead SA6.2 RobE\_eSIM.



[\*] Key-A in the first MP-capable is related to RFC6824 only and does not exist in RFC6824bis.

Figure 17: MPTCP RobE\_SIM and RobE\_eSIM connection setup

### 3.1.6. Security Consideration

[Tbd, however no differences to [RFC6824] and [I-D.ietf-mptcp-rfc6824bis] are expected]

## 3.2. Heuristic Initial Path Selection with remote RTT Measurement

### 3.2.1. Description

Usually the path RTT can be determined by a time difference between sending a package and an ACK and is integrated into the TCP protocol. For asymmetric transmission, the latest RTT for TCP flows is calculated by the side which sends data at latest and possible does not correspond to the site which employs RobE\_IPS. This problem is already elaborated in Section 2.3.4 and can be solved by transmitting the RTT information per subflow. The negotiation procedure is

depicted in Figure 18 and uses the MPTCP option `L_RTT_EN` defined in Section 3.2.2.

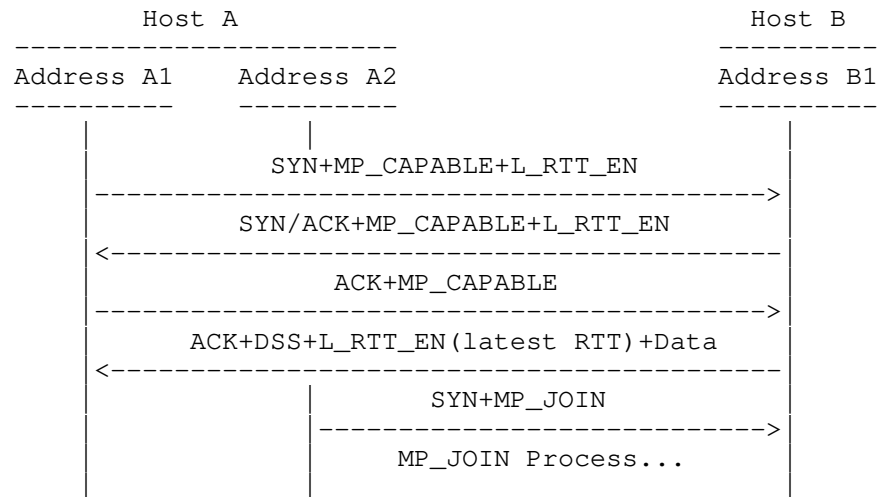


Figure 18: Negotiation procedure for RTT exchange

A successful negotiation allows the exchange of the measured RTT value from one subflow of a MPTCP host to another using the "Latest RTT" field within the `L_RTT_EN` option.

### 3.2.2. Protocol Adaptation

Calculating the "Latest RTT" by a remote host in an asymmetry transmission scenario should be transferred from remote host to the client running RobE\_IPS. So a new MPTCP subtype option named `L_RTT_EN` is allocated for this function. During the three-way handshake `L_RTT_EN` is used for negotiation of remote RTT measurement capability between client and server (in Section 3.2.1). When both parts support the usage of remote RTT measurement, the "Latest RTT" field in `L_RTT_EN` is applied for carrying the value of latest RTT computed by the remote host.



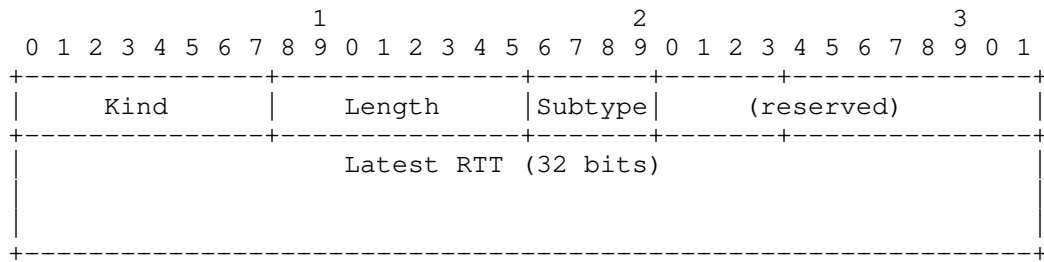


Figure 19: ROBE\_L\_RTT\_EN OPTION

### 3.2.3. Fallback Mechanism

When the receiver is not L\_RTT\_EN capable, MPTCP session with L\_RTT\_EN negotiation will seamlessly fallback to normal MPTCP process.

[TBD, Need same checks as Section 3.1.4.2]

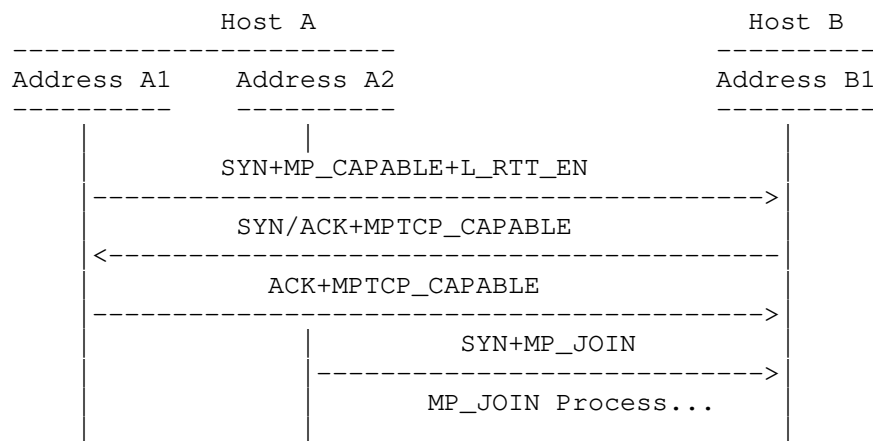


Figure 20: Fallback to MPTCP without RobE\_IPS

### 3.2.4. Security Consideration

[Tbd]

## 4. IANA Considerations

This document defines three new values to MPTCP Option Subtype as following.

Value	Symbol	Name	Reference
TBD	ROBE_eSIM_EN	RobE_eSIM enabled	Section 3.1
TBD	MP_JOIN_CAP	Join connection directly in RobE_eSIM	Section 3.1
TBD	L_RTT_EN	Server RTT enabled	Section 3.2

RobE Option Subtypes

Table 2: RobE Option Subtypes

## 5. Informative References

- [I-D.ietf-mptcp-rfc6824bis]  
 Ford, A., Raiciu, C., Handley, M., Bonaventure, O., and C. Paasch, "TCP Extensions for Multipath Operation with Multiple Addresses", draft-ietf-mptcp-rfc6824bis-18 (work in progress), June 2019.
- [RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, DOI 10.17487/RFC0793, September 1981, <<https://www.rfc-editor.org/info/rfc793>>.
- [RFC6824] Ford, A., Raiciu, C., Handley, M., and O. Bonaventure, "TCP Extensions for Multipath Operation with Multiple Addresses", RFC 6824, DOI 10.17487/RFC6824, January 2013, <<https://www.rfc-editor.org/info/rfc6824>>.

## Authors' Addresses

Markus Amend  
 Deutsche Telekom  
 T-Online-Allee 7  
 64295 Darmstadt  
 Germany

Email: [Markus.Amend@telekom.de](mailto:Markus.Amend@telekom.de)

Jiao Kang  
Huawei  
D2-03, Huawei Industrial Base  
Longgang District  
Shenzhen  
China

Email: kangjiao@huawei.com

Network Working Group  
Internet-Draft  
Intended status: Experimental  
Expires: January 9, 2020

M. Bagnulo  
UC3M  
A. Andersdotter  
Article 19  
C. Paasch  
Apple  
July 8, 2019

Privacy threats and possible countermeasures for Multipath-TCP (MPTCP)  
draft-bagnulo-mptcp-privacy-00.txt

#### Abstract

This note performs a differential analysis of the threats regarding privacy of the Multipath TCP protocol compared to regular TCP and proposes a set of countermeasures for the threats identified.

#### Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on January 9, 2020.

#### Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of

the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	2
2. Threat Analysis . . . . .	3
2.1. Types of attackers . . . . .	3
2.2. Detailed attack mechanics. . . . .	4
2.2.1. Attacks using MP_CAPABLE and MP_JOIN. . . . .	4
2.2.2. Attacks using ADD_ADDR. . . . .	4
3. Countermeasures. . . . .	4
4. MPTCP privacy features. . . . .	5
5. Security Considerations . . . . .	5
6. IANA Considerations . . . . .	6
7. Acknowledgements . . . . .	6
8. Informative References . . . . .	6
Authors' Addresses . . . . .	6

## 1. Introduction

Multipath-TCP (MPTCP) [RFC6824] [I-D.ietf-mptcp-rfc6824bis] is a set of extensions to TCP that enable the use of multiple IP addresses throughout the lifetime of a (MP)TCP connection. The use of multiple addresses in a connection allows two main uses cases, namely mobility and multihoming. In the case of multihoming, if an endpoint is connected to the Internet through multiple interfaces simultaneously (each ones having a different IP address), the use of MPTCP allow additional fault tolerance as the connection can be preserved by using an alternative IP address even if the IP address originally used to establish the connection is rendered unavailable. In the case of mobility, as an endpoint changes its attachment to the Internet, it acquires a new IP address associated to its new attachment point. By using MPTCP, connections can be preserved throughout the changes of attachment points and their respective IP addresses by adding the new IP addresses to the ongoing MPTCP connections.

Because of its very nature, the operation of MPTCP presents privacy implications, as other protocols that bind multiple IP addresses to a given endpoint [I-D.nordmark-id-loc-privacy]. Because MPTCP explicitly associated multiple IP addresses to a given connection and hence to a given endpoint, it discloses information about the node whereabouts to third parties. In this note, we perform an analysis of the privacy implications of the operation of the MPTCP compared to regular TCP and we provide a set of countermeasures to address the identified threats. It is out of the scope of this document to identify privacy threats that equally affect both MPTCP and TCP, such

as the ones resulting from exchanging unencrypted data that can be observed by eavesdroppers along the path. As mentioned earlier, we only identify threats against privacy introduced by MPTCP which are not present in TCP.

## 2. Threat Analysis

The threats concerning privacy of the use of MPTCP are essentially two:

**Movement tracking:** In the case that MPTCP is used for mobility, the use of multiple addresses in the same MPTCP connection can be used by an attacker to track the movement of the victim. Since IP addresses can be related to location (in a more or less accurate manner), by observing the different addresses used in a MPTCP connection, the attacker can track the itinerary of the victim.

**More accurate positioning.** If multiple address are used simultaneously in a MPTCP connection, this implies that the endpoint is connected at the multiple attachment points at the same time, potentially providing the means for a more accurate positioning of the endpoint (e.g. if an endpoint is exposing the IP address of the cellular access it is providing less information than when it also exposes the IP address of an Internet coffee wifi access).

### 2.1. Types of attackers

We classify the types of attackers based on their topological location, which determines the amount of information they have access to. the different types of attackers are:

**Partially on-path attacker.** This attacker is located along one or more, but not all the paths used to exchange MPTCP signaling information. As such, it is able to see some but not all the MPTCP packets.

**Full On-path attacker.** This attacker is able to eavesdrop all the MPTCP signaling message exchange, but it is not the other endpoint of the information.

**Endpoint:** in this case, the other endpoint of the MPTCP connection is the attacker (in the sense that it will use the information revealed through MPTCP for other purposes beyond the MPTCP operation e.g. the endpoint may decide to sell the location and tracking information of the MPTCP endpoints to third parties).

## 2.2. Detailed attack mechanics.

### 2.2.1. Attacks using MP\_CAPABLE and MP\_JOIN.

A MPTCP endpoint initiates a MPTCP connection by including the MP\_CAPABLE option in the SYN message. After that, it uses the MP\_JOIN option to add subsequent subflows using other IP address to the existent connection. The MP\_JOIN message include a token that is used by the MPTCP receiver to identify which of the ongoing MPTCP connections this particular subflow is being added to. In order for an attacker to bind the different address together, it must be able to observe at least two messages carrying two different addresses. In particular, by observing the initial MP\_CAPABLE SYN and a following MP\_JOIN message, the attacker can relate these two IP addresses. Also, by observing two MP\_JOIN messages carrying different IP addresses but the same token, the attacker can also relate the two IP addresses together.

This attack can be executed by any attacker that is capable of observing the different MP\_CAPABLE and MP\_JOIN messages. So, for a partially on-path attacker, the attack will be as effective as the number of used path the attacker has access to. If it only has access to one path, the attacker would not gather any information. Both full on-path attackers and the endpoint would have access to all the information, so the attack effectiveness is complete.

Both versions of MPTCP, i.e. [RFC6824] [I-D.ietf-mptcp-rfc6824bis] are equally affected by this attack.

### 2.2.2. Attacks using ADD\_ADDR.

The ADD\_ADDR option allows the sender of the message to add an IP address to the existing connection. From a privacy perspective, the packet containing the ADD\_ADDR information already discloses a binding between two addresses, the address used a source address of the packet and the address included in the ADD\_ADDR option. This attack can be performed by any attacker who is able to observe the message, including partial and full on-path attackers and the endpoint itself. This attack can be combined with the attack done using MP\_CAPABLE AND MP\_JOIN messages described in the previous section, to retrieve a larger set of addresses. This attack affects both version [RFC6824] [I-D.ietf-mptcp-rfc6824bis] of MPTCP.

## 3. Countermeasures.

It is possible to design countermeasures to prevent the described attacks.

ADD\_ADDR attack.

In order to prevent the ADD\_ADDR based attack, it would be possible to encrypt the address carried in the ADD\_ADDR message, for example with the key exchanged in the MP\_CAPABLE exchange. By doing this, only the attackers who have observed the initial MP\_CAPABLE message would be able to decrypt the content of the ADD\_ADDR message, significantly limiting the attack surface.

MP\_CAPABLE and MP\_JOIN.

In order to prevent the MP\_CAPABLE/MP\_JOIN attack, it would be necessary to change the token in every MP\_JOIN message. The difficulty with this of course is that the token is used as a key to identify which MPTCP connection this new subflow belongs to. Using different tokens would be possible as long as the receiver would be able to decrypt it and find the ongoing connection that this new subflow belongs to.

For instance, the token could be the hash of the concatenation of a trail of  $n$  zeros, the key and the new IP address of the flow. This token would change with every new subflow, since the IP address would change (we could also add the source port, to support the case of multiple subflows with the same source IP address). Upon the reception of an MP\_JOIN message, the receiver would need to try with all the keys of ongoing connections. It will know it has succeeded, because the correct one will result in a trail of  $n$  zeros. The problem with this mechanism is that it imposes an additional cost in terms of computation upon the establishment of a new subflow.

Additional countermeasures could be in the form of a recommendation about when to establish a new subflow or when to announce new addresses using ADD\_ADDR. Generating awareness that doing so discloses private information of the endpoint would make implementations more conservative when advertising IP addresses.

#### 4. MPTCP privacy features.

MPTCP also provides some positive side effects with regard to privacy. In particular, because the information is spread across multiple paths, in order to be able to eavesdrop all the content of a MPTCP connection, the attacker needs to be present in all used paths, making more challenging for the attacker to fulfill its goal.

#### 5. Security Considerations



## 6. IANA Considerations

## 7. Acknowledgements

This work was supported by the Spanish Ministry of Economy and Competitiveness through the 5G-City project (TEC2016-76795-C6-3-R).

## 8. Informative References

[I-D.ietf-mptcp-rfc6824bis]

Ford, A., Raiciu, C., Handley, M., Bonaventure, O., and C. Paasch, "TCP Extensions for Multipath Operation with Multiple Addresses", draft-ietf-mptcp-rfc6824bis-18 (work in progress), June 2019.

[I-D.nordmark-id-loc-privacy]

Nordmark, E., "Privacy issues in ID/locator separation systems", draft-nordmark-id-loc-privacy-00 (work in progress), July 2018.

[RFC6824] Ford, A., Raiciu, C., Handley, M., and O. Bonaventure, "TCP Extensions for Multipath Operation with Multiple Addresses", RFC 6824, DOI 10.17487/RFC6824, January 2013, <<https://www.rfc-editor.org/info/rfc6824>>.

## Authors' Addresses

Marcelo Bagnulo  
UC3M

Email: [marcelo@it.uc3m.es](mailto:marcelo@it.uc3m.es)

Amelia Andersdotter  
Article 19

Email: [amelia@article19.org](mailto:amelia@article19.org)

Christoph Paasch  
Apple

Email: [cpaasch@apple.com](mailto:cpaasch@apple.com)

MPTCP Working Group  
Internet-Draft  
Intended status: Informational  
Expires: January 9, 2020

V. Tran  
O. Bonaventure  
Universite catholique de Louvain  
July 08, 2019

Multipath TCP Inactivity Time Option  
draft-hoang-mptcp-inactivity-time-00

Abstract

This document discusses the lifetime of idle MPTCP sessions, and defines an MPTCP Option subkind for hosts to announce and request this value.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on January 9, 2020.

Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	2
2. Terminology . . . . .	3
3. Lifetime of Multipath TCP connections . . . . .	3
4. The Multipath TCP Inactivity Timer Option . . . . .	4
4.1. Option Format . . . . .	4
4.2. ITO Option and Local Policies . . . . .	5
5. Implementation and Interoperability . . . . .	5
5.1. Interaction with TCP Keepalives . . . . .	5
5.2. Interaction with User Timeout . . . . .	6
6. Security Considerations . . . . .	6
7. IANA Considerations . . . . .	6
8. Acknowledgements . . . . .	6
9. References . . . . .	6
9.1. Normative References . . . . .	6
9.2. Informative References . . . . .	7
Authors' Addresses . . . . .	7

## 1. Introduction

Multipath TCP [RFC6824] [I-D.ietf-mptcp-rfc6824bis] is used in various use cases [RFC8040]. In several of these deployments, a Multipath TCP connection is composed of a number of subflows that varies during the lifetime of the connection. TCP implementations create state when a connection starts and maintain this state until the abrupt or graceful termination of the connection according to the TCP state machine [RFC0793]. Multipath TCP implementations need to maintain two types of states:

- o the per-Multipath TCP connection state that includes the connection keys, data sequence numbers, ...
- o the per-TCP subflows connection states

The per-TCP state of each subflow is managed according to [RFC0793]. The Multipath TCP connection state also needs to be managed. We consider a subflow to be active if its TCP state machine exists and is in the Established state. We call inactive a Multipath TCP connection that currently has no active subflow. It is important to note that it is possible for one host to consider a given Multipath TCP connection to be inactive while the other endpoint considers that the connection is still active. This can happen for example when some packets are lost or when one of the hosts has received a spurious RST on the only active subflow. [RFC6824] and [I-D.ietf-mptcp-rfc6824bis] specify how this state can be removed upon transmission or reception of a FASTCLOSE and after having exchanged DATA\_FINs. However, [RFC6824] and

[I-D.ietf-mptcp-rfc6824bis] do not specify precisely how the Multipath TCP connection state must be managed when all the TCP subflows associated with this Multipath TCP connection have been released. A similar problem existed in TCP for idle connections. It was clarified in [RFC1122].

Given the cost of maintaining state for inactive Multipath TCP connections, hosts may want to limit the time during which they maintain state for inactive Multipath TCP connections. Neither [RFC6824] nor [I-D.ietf-mptcp-rfc6824bis] propose a default duration to maintain this state. Although some applications such as http/1.1 include a mechanism to terminate idle connections, many applications do not do this. We fill this gap in this document and propose a Multipath TCP option that enables hosts to agree on a common minimum duration to maintain inactive Multipath TCP connections.

## 2. Terminology

In this document, the key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described in BCP 14, RFC 2119 [RFC2119].

## 3. Lifetime of Multipath TCP connections

A Multipath TCP connection starts with a three-way-handshake that uses the MP\_CAPABLE option and ends with either the transmission of DATA\_FINs or a FASTCLOSE. Between these two events, the number of subflows that compose the connection may vary and during some periods of time, it is possible that no active subflow is associated with an existing Multipath TCP connection.

There are several types of events which can occur during the lifetime of a Multipath TCP connection:

- o establishment of a new subflow
- o graceful termination of subflow with the exchange of FINs
- o abrupt termination of a subflow due to multiple expirations of its retransmission timer, the reception of a RST or an abnormal packet, ...

To support the establishment of a new subflow with the MP\_JOIN option, a host must maintain state for all established Multipath TCP connection. Neither [RFC6824] nor [I-D.ietf-mptcp-rfc6824bis] specify how a host should behave upon (graceful or abrupt) termination of the last subflow associated with an existing Multipath

TCP connection. One possibility would be to consider that since no subflow is associated with this connection, its state could be freed. A consequence of this policy is that it would then be impossible to establish a new subflow to this connection to recover from the failure of the previous one. Another possibility is to preserve the state of this Multipath TCP connection for some period of time. This would better match the expectations of mobile use cases where subflows can fail abruptly when devices move.

This document proposes the utilisation of an Inactivity timer in Multipath TCP implementations. This timer bounds the time during which there is no active subflow associated with a given Multipath TCP connection. It starts when the last active subflow associated with a connection is terminated. It is reset when one active subflow is associated with this Multipath TCP connection. Upon expiration of this timer, the host SHOULD remove state for the associated Multipath TCP connection. If a host wants to maintain a Multipath TCP connection as active, it SHOULD try to reestablish a subflow associated with this Multipath TCP connection before the expiration of its inactivity timer. This document suggests implementations to start the reestablishment of such a subflow after half the Inactivity timer.

#### 4. The Multipath TCP Inactivity Timer Option

The Multipath TCP Inactivity Timer Option (MPTCP ITO) is used to exchange the Inactivity Timer associated with this connection. This documents recommends a default value of 2 hours and 4 minutes – the same as default NAT timeout as discussed in [RFC5382]. The reason for this value is to allow the peer with default keepalive timer of 2 hours ([RFC1122]) to probe successfully.

##### 4.1. Option Format

The format of the ITO option is depicted in Fig. Figure 1:

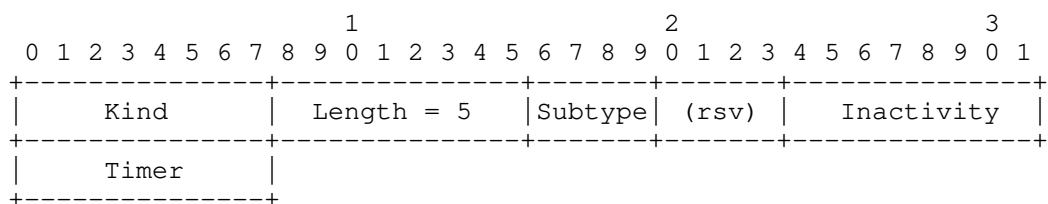


Figure 1: MPTCP Inactivity Timer Option Format

In the ITO option, the Inactivity Timer (16 bits) specifies the requested Timer for the MPTCP session in seconds. A value of zero indicates that the state for the Multipath TCP connection will be removed as soon as there is no active subflow.

#### 4.2. ITO Option and Local Policies

The ITO option is used to specify the value of the Inactivity timer associated with the connection by the host that sends the option. Upon reception of this option, the receiver SHOULD update its ITO option based on the ITO option received from its peer. However, it MAY choose to apply and transmit its own ITO option, depending on its local policies. Note that the ITO option is an indicative value. A host may terminate long inactive connections before the expiration of the ITO timer due to the increasing memory resource pressure.

Like all Multipath TCP options, the ITO Option is exchanged without any protection from TCP's reliability mechanisms. Therefore, implementations MUST NOT assume that it is transferred reliably. Implementations that use the ITO option can transmit the ITO option at any time. Since the utilisation of this option is not negotiated during the connection handshake, a host MUST NOT send more than three ITO options on a connection where it has not received any ITO option.

After the expiration of the Inactivity timer, the host MAY choose to close the MPTCP session with MPTCP\_RST and all of its subflows with TCP\_RST, and report the inactivity timeout error to the user. This is a common case on the servers that want to free the resource occupied by unused sessions as soon as possible so that they could serve other users.

### 5. Implementation and Interoperability

#### 5.1. Interaction with TCP Keepalives

The mechanism specified in this document operates above the TCP Keepalives defined in [RFC1122]. If TCP Keepalives are enabled on at least one of the subflows of a Multipath TCP connection, then this subflow will remain active and the Inactivity timer of the associated Multipath TCP connection will not expire. In this case, the Inactivity timer should be longer than the TCP Keepalive timer. If TCP Keepalives are disabled, then the mechanism described in this document will remove the state of the inactive connections.

On some MPTCP implementations like Linux, the TCP Keepalive is supported at the MPTCP layer. To implement the idle timeout option, these implementations may use the MPTCP KeepAliveTime value to keep track of the current idle time. However, the idle timeout mechanism,

when activated, should override the keepalive mechanism. This means that after the idle timeout fired, the host should abort the connection instead of sending the TCP keepalives.

The MPTCP Inactivity Timer Option MAY be controlled on a system-wide setting or on a per-connection basis. Specific APIs and mechanisms for controlling the ITO option are out of the scope of this document.

## 5.2. Interaction with User Timeout

The interactions between the User Timeout option [RFC5482] and the ITO option will be discussed in a subsequent revision of this document.

## 6. Security Considerations

The ITO option enables hosts to exchange the value of a timer that protects against some types of resource exhaustion attacks. Multipath TCP implementations should define a range of safe values for the ITO option and prevent applications from configuring an inactivity timer outside this range. They should also ignore received ITO options that are outside this range.

Since the ITO option is neither encrypted nor authenticated, on-path attackers and middleboxes could remove, add or modify the ITO option on observed Multipath TCP connections.

## 7. IANA Considerations

IANA is requested to assign an MPTCP option subtype for the ITO option from the "MPTCP Option Subtypes" available at <https://www.iana.org/assignments/tcp-parameters/tcp-parameters.xhtml>

## 8. Acknowledgements

This work was supported by the ARC-SDN project and the Walinnov MQUIC project, No. 1810018.

## 9. References

### 9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.

## 9.2. Informative References

- [I-D.ietf-mptcp-rfc6824bis]  
Ford, A., Raiciu, C., Handley, M., Bonaventure, O., and C. Paasch, "TCP Extensions for Multipath Operation with Multiple Addresses", draft-ietf-mptcp-rfc6824bis-18 (work in progress), June 2019.
- [RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, DOI 10.17487/RFC0793, September 1981, <<https://www.rfc-editor.org/info/rfc793>>.
- [RFC1122] Braden, R., Ed., "Requirements for Internet Hosts - Communication Layers", STD 3, RFC 1122, DOI 10.17487/RFC1122, October 1989, <<https://www.rfc-editor.org/info/rfc1122>>.
- [RFC5382] Guha, S., Ed., Biswas, K., Ford, B., Sivakumar, S., and P. Srisuresh, "NAT Behavioral Requirements for TCP", BCP 142, RFC 5382, DOI 10.17487/RFC5382, October 2008, <<https://www.rfc-editor.org/info/rfc5382>>.
- [RFC5482] Eggert, L. and F. Gont, "TCP User Timeout Option", RFC 5482, DOI 10.17487/RFC5482, March 2009, <<https://www.rfc-editor.org/info/rfc5482>>.
- [RFC6824] Ford, A., Raiciu, C., Handley, M., and O. Bonaventure, "TCP Extensions for Multipath Operation with Multiple Addresses", RFC 6824, DOI 10.17487/RFC6824, January 2013, <<https://www.rfc-editor.org/info/rfc6824>>.
- [RFC8040] Bierman, A., Bjorklund, M., and K. Watsen, "RESTCONF Protocol", RFC 8040, DOI 10.17487/RFC8040, January 2017, <<https://www.rfc-editor.org/info/rfc8040>>.

## Authors' Addresses

Viet-Hoang Tran  
Universite catholique de Louvain

Email: [hoang.tran@uclouvain.be](mailto:hoang.tran@uclouvain.be)



Olivier Bonaventure  
Universite catholique de Louvain  
Pl. Ste Barbe, 2  
Louvain-la-Neuve 1348  
Belgium

Email: [olivier.bonaventure@uclouvain.be](mailto:olivier.bonaventure@uclouvain.be)

MPTCP Working Group  
Internet-Draft  
Intended status: Informational  
Expires: January 9, 2020

V. Tran  
O. Bonaventure  
Universite catholique de Louvain  
July 08, 2019

Multipath TCP Subflow Rate Limit Option  
draft-hoang-mptcp-sub-rate-limit-00

Abstract

This document defines a new MPTCP Option that enables hosts to request their peers to limit the maximum transfer rate on a per-subflow basis.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <https://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on January 9, 2020.

Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<https://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	2
2. Terminology . . . . .	3
3. The Multipath TCP Subflow Rate Limit (SRL) Option . . . . .	3
3.1. Option Format . . . . .	3
3.2. SRL Option and Local Policies . . . . .	4
4. Implementation and Interoperability . . . . .	4
5. Security Considerations . . . . .	4
6. IANA Considerations . . . . .	5
7. Acknowledgements . . . . .	5
8. References . . . . .	5
8.1. Normative References . . . . .	5
8.2. Informative References . . . . .	5
Authors' Addresses . . . . .	6

## 1. Introduction

Multipath TCP [RFC6824] [I-D.ietf-mptcp-rfc6824bis] is used in various use cases [RFC8040]. In several situations, a Multipath TCP host would like its peer to limit the sending rate over a specific subflow.

It is common for mobile clients to have limited cellular data subscription. Even if this is not the case, mobile network operators may still silently throttle the networking capacity of the customers who have used up a large amount of cellular data. A good LTE or 5G connection running at full speed during less than a few hours could consume the entire monthly budget cellular quota of many users. This is even more important when the mobile clients are roaming abroad where the monetary cost for cellular data can be very high. A common scenario is that mobile users want to limit the monetary cost of using cellular networks or to avoid running out of their mobile data quota. Smartphones can easily rate limit their upstream bandwidth, but unfortunately, most smartphone applications mainly receive data. For these applications, a rate limit must happen on the server side. This rate limit could be enforced by the application, e.g. by selecting a specific video coding scheme, but applying it at the transport layer would be more generic and could be done from the system level automatically.

As discussed on the multipathtcp IETF mailing list [paasch\_mptcpwg\_2019], this rate-control mechanism can also be used when a client wants to inform a server to close a subflow gracefully by requesting a zero transfer rate. Though the client may send a TCP-RST on this subflow instead, in-flight data would be lost and must be reinjected over other subflows. Another solution is to send an MP\_PRIO option to the sender to put the cellular subflow into

backup mode, but this request could be overridden by the sender's local policy.

## 2. Terminology

In this document, the key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described in BCP 14, RFC 2119 [RFC2119].

## 3. The Multipath TCP Subflow Rate Limit (SRL) Option

This document proposes a Subflow Rate Limit option that indicates a maximum receive rate for the subflow it is sent. Like other MPTCP options, this option is not sent reliably. Hosts SHOULD resend it several times, but not more frequently than once per second.

### 3.1. Option Format

The format of the SRL option is depicted in Fig. Figure 1:

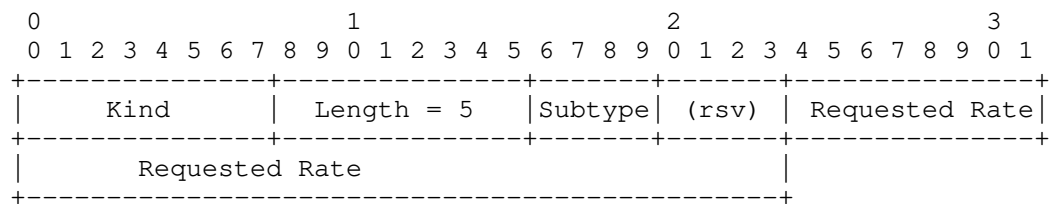


Figure 1: MPTCP Subflow Rate Limit Option Format

All reserved bits MUST be set to zero.

In the SRL option, the Requested Rate (32 bits) is specified in IEEE 754 binary32 format (single-precision binary floating-point format). The same format is used in [RFC3630] to specify the maximum bandwidth of a link. This format is as follows:

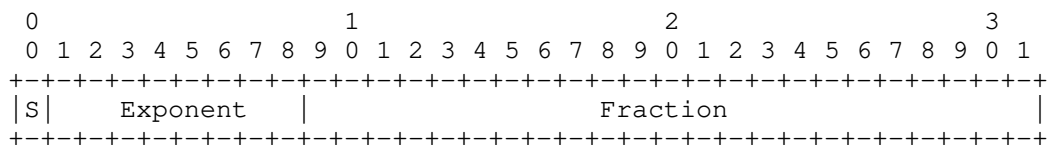


Figure 2: IEEE binary32 format

- o Sign bit (1 bit) MUST be set to zero.
- o Exponent (8 bits) is the exponent base 2 in "excess 127" notation.
- o Fraction (23 bits) is the mantissa - 1, with an implied binary point in front of it.

Thus, the above represents the value:

$$(-1)^{(S)} * 2^{(Exponent-127)} * (1 + Fraction)$$

The unit of this value is Kilobits per second (Kbps).

### 3.2. SRL Option and Local Policies

Note that the SRL option is an indicative value. Upon reception of this option, the receiver SHOULD set the maximum rate on the subflow over which the option was received.

Like all Multipath TCP options, the SRL Option is exchanged without any protection from TCP's reliability mechanisms. Therefore, implementations MUST NOT assume that it is transferred reliably. Implementations that use the SRL option can transmit the SRL option at any time. Since the utilisation of this option is not negotiated during the connection handshake, a host MUST NOT send more than three SRL options on a connection where it has not received any SRL option.

### 4. Implementation and Interoperability

Implementations MAY use various mechanisms to implement the rate control policy, for example using TCP Pacing or clamping the subflow congestion window.

### 5. Security Considerations

Since the SRL option is neither encrypted nor authenticated, on-path attackers and middleboxes could remove, add or modify the SRL option on observed Multipath TCP connections. However, manipulating this option does not open new attacks compared to the ones documented in [RFC6181] [RFC7430].

For example, an on-path middle man could insert an option to throttle the rate on a subflow to nearly zero, effectively stalling the subflow. However, if an attacker has that capability, it could instead drop all packets or inject the TCP-RST/MP-FASTCLOSE.

On the other hand, on-path middleboxes may increase the rate-limit value in the exchanged option to a very high value. This effectively has the same effect as filtering out the option.

## 6. IANA Considerations

IANA is requested to assign an MPTCP option subtype for the SRL option from the "MPTCP Option Subtypes" available at <https://www.iana.org/assignments/tcp-parameters/tcp-parameters.xhtml>

## 7. Acknowledgements

This work was supported by the ARC-SDN project and the Walinnov MQUIC project, No. 1810018.

## 8. References

### 8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC6181] Bagnulo, M., "Threat Analysis for TCP Extensions for Multipath Operation with Multiple Addresses", RFC 6181, DOI 10.17487/RFC6181, March 2011, <<https://www.rfc-editor.org/info/rfc6181>>.
- [RFC6824] Ford, A., Raiciu, C., Handley, M., and O. Bonaventure, "TCP Extensions for Multipath Operation with Multiple Addresses", RFC 6824, DOI 10.17487/RFC6824, January 2013, <<https://www.rfc-editor.org/info/rfc6824>>.
- [RFC7430] Bagnulo, M., Paasch, C., Gont, F., Bonaventure, O., and C. Raiciu, "Analysis of Residual Threats and Possible Fixes for Multipath TCP (MPTCP)", RFC 7430, DOI 10.17487/RFC7430, July 2015, <<https://www.rfc-editor.org/info/rfc7430>>.

### 8.2. Informative References

- [I-D.ietf-mptcp-rfc6824bis] Ford, A., Raiciu, C., Handley, M., Bonaventure, O., and C. Paasch, "TCP Extensions for Multipath Operation with Multiple Addresses", draft-ietf-mptcp-rfc6824bis-18 (work in progress), June 2019.

[paasch\_mptcpwg\_2019]

Paasch, C., "Regarding rate control at a subflow level",  
May 2019,  
<<https://mailarchive.ietf.org/arch/msg/multipathtcp/fyhIpWnXCr0sImRECGh8Lx1zWdc>>.

[RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering  
(TE) Extensions to OSPF Version 2", RFC 3630,  
DOI 10.17487/RFC3630, September 2003,  
<<https://www.rfc-editor.org/info/rfc3630>>.

[RFC8040] Bierman, A., Bjorklund, M., and K. Watsen, "RESTCONF  
Protocol", RFC 8040, DOI 10.17487/RFC8040, January 2017,  
<<https://www.rfc-editor.org/info/rfc8040>>.

#### Authors' Addresses

Viet-Hoang Tran  
Universite catholique de Louvain

Email: [hoang.tran@uclouvain.be](mailto:hoang.tran@uclouvain.be)

Olivier Bonaventure  
Universite catholique de Louvain  
Pl. Ste Barbe, 2  
Louvain-la-Neuve 1348  
Belgium

Email: [olivier.bonaventure@uclouvain.be](mailto:olivier.bonaventure@uclouvain.be)

Internet Engineering Task Force  
Internet-Draft  
Obsoletes: 6824 (if approved)  
Intended status: Standards Track  
Expires: December 10, 2019

A. Ford  
Pexip  
C. Raiciu  
U. Politechnica of Bucharest  
M. Handley  
U. College London  
O. Bonaventure  
U. catholique de Louvain  
C. Paasch  
Apple, Inc.  
June 8, 2019

TCP Extensions for Multipath Operation with Multiple Addresses  
draft-ietf-mptcp-rfc6824bis-18

Abstract

TCP/IP communication is currently restricted to a single path per connection, yet multiple paths often exist between peers. The simultaneous use of these multiple paths for a TCP/IP session would improve resource usage within the network and, thus, improve user experience through higher throughput and improved resilience to network failure.

Multipath TCP provides the ability to simultaneously use multiple paths between peers. This document presents a set of extensions to traditional TCP to support multipath operation. The protocol offers the same type of service to applications as TCP (i.e., reliable bytestream), and it provides the components necessary to establish and use multiple TCP flows across potentially disjoint paths.

This document specifies v1 of Multipath TCP, obsoleting v0 as specified in RFC6824, through clarifications and modifications primarily driven by deployment experience.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any



time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on December 10, 2019.

#### Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

#### Table of Contents

1. Introduction . . . . .	3
1.1. Design Assumptions . . . . .	4
1.2. Multipath TCP in the Networking Stack . . . . .	5
1.3. Terminology . . . . .	6
1.4. MPTCP Concept . . . . .	7
1.5. Requirements Language . . . . .	8
2. Operation Overview . . . . .	8
2.1. Initiating an MPTCP Connection . . . . .	9
2.2. Associating a New Subflow with an Existing MPTCP Connection . . . . .	10
2.3. Informing the Other Host about Another Potential Address	11
2.4. Data Transfer Using MPTCP . . . . .	12
2.5. Requesting a Change in a Path's Priority . . . . .	13
2.6. Closing an MPTCP Connection . . . . .	13
2.7. Notable Features . . . . .	14
3. MPTCP Protocol . . . . .	15
3.1. Connection Initiation . . . . .	16
3.2. Starting a New Subflow . . . . .	23
3.3. General MPTCP Operation . . . . .	28
3.3.1. Data Sequence Mapping . . . . .	30
3.3.2. Data Acknowledgments . . . . .	33
3.3.3. Closing a Connection . . . . .	34
3.3.4. Receiver Considerations . . . . .	36
3.3.5. Sender Considerations . . . . .	37
3.3.6. Reliability and Retransmissions . . . . .	38
3.3.7. Congestion Control Considerations . . . . .	39

3.3.8. Subflow Policy . . . . .	39
3.4. Address Knowledge Exchange (Path Management) . . . . .	40
3.4.1. Address Advertisement . . . . .	42
3.4.2. Remove Address . . . . .	45
3.5. Fast Close . . . . .	46
3.6. Subflow Reset . . . . .	48
3.7. Fallback . . . . .	49
3.8. Error Handling . . . . .	53
3.9. Heuristics . . . . .	53
3.9.1. Port Usage . . . . .	54
3.9.2. Delayed Subflow Start and Subflow Symmetry . . . . .	54
3.9.3. Failure Handling . . . . .	55
4. Semantic Issues . . . . .	56
5. Security Considerations . . . . .	57
6. Interactions with Middleboxes . . . . .	60
7. Acknowledgments . . . . .	63
8. IANA Considerations . . . . .	64
8.1. MPTCP Option Subtypes . . . . .	64
8.2. MPTCP Handshake Algorithms . . . . .	65
8.3. MP_TCPRST Reason Codes . . . . .	66
9. References . . . . .	67
9.1. Normative References . . . . .	67
9.2. Informative References . . . . .	68
Appendix A. Notes on Use of TCP Options . . . . .	71
Appendix B. TCP Fast Open and MPTCP . . . . .	72
B.1. TFO cookie request with MPTCP . . . . .	72
B.2. Data sequence mapping under TFO . . . . .	73
B.3. Connection establishment examples . . . . .	74
Appendix C. Control Blocks . . . . .	76
C.1. MPTCP Control Block . . . . .	76
C.1.1. Authentication and Metadata . . . . .	76
C.1.2. Sending Side . . . . .	77
C.1.3. Receiving Side . . . . .	77
C.2. TCP Control Blocks . . . . .	77
C.2.1. Sending Side . . . . .	78
C.2.2. Receiving Side . . . . .	78
Appendix D. Finite State Machine . . . . .	78
Appendix E. Changes from RFC6824 . . . . .	79
Authors' Addresses . . . . .	81

## 1. Introduction

Multipath TCP (MPTCP) is a set of extensions to regular TCP [RFC0793] to provide a Multipath TCP [RFC6182] service, which enables a transport connection to operate across multiple paths simultaneously. This document presents the protocol changes required to add multipath capability to TCP; specifically, those for signaling and setting up multiple paths ("subflows"), managing these subflows, reassembly of

data, and termination of sessions. This is not the only information required to create a Multipath TCP implementation, however. This document is complemented by three others:

- o Architecture [RFC6182], which explains the motivations behind Multipath TCP, contains a discussion of high-level design decisions on which this design is based, and an explanation of a functional separation through which an extensible MPTCP implementation can be developed.
- o Congestion control [RFC6356] presents a safe congestion control algorithm for coupling the behavior of the multiple paths in order to "do no harm" to other network users.
- o Application considerations [RFC6897] discusses what impact MPTCP will have on applications, what applications will want to do with MPTCP, and as a consequence of these factors, what API extensions an MPTCP implementation should present.

This document is an update to, and obsoletes, the v0 specification of Multipath TCP (RFC6824). This document specifies MPTCP v1, which is not backward compatible with MPTCP v0. This document additionally defines version negotiation procedures for implementations that support both versions.

### 1.1. Design Assumptions

In order to limit the potentially huge design space, the mptcp working group imposed two key constraints on the Multipath TCP design presented in this document:

- o It must be backwards-compatible with current, regular TCP, to increase its chances of deployment.
- o It can be assumed that one or both hosts are multihomed and multiaddressed.

To simplify the design, we assume that the presence of multiple addresses at a host is sufficient to indicate the existence of multiple paths. These paths need not be entirely disjoint: they may share one or many routers between them. Even in such a situation, making use of multiple paths is beneficial, improving resource utilization and resilience to a subset of node failures. The congestion control algorithms defined in [RFC6356] ensure this does not act detrimentally. Furthermore, there may be some scenarios where different TCP ports on a single host can provide disjoint paths (such as through certain Equal-Cost Multipath (ECMP) implementations

[RFC2992]), and so the MPTCP design also supports the use of ports in path identifiers.

There are three aspects to the backwards-compatibility listed above (discussed in more detail in [RFC6182]):

**External Constraints:** The protocol must function through the vast majority of existing middleboxes such as NATs, firewalls, and proxies, and as such must resemble existing TCP as far as possible on the wire. Furthermore, the protocol must not assume the segments it sends on the wire arrive unmodified at the destination: they may be split or coalesced; TCP options may be removed or duplicated.

**Application Constraints:** The protocol must be usable with no change to existing applications that use the common TCP API (although it is reasonable that not all features would be available to such legacy applications). Furthermore, the protocol must provide the same service model as regular TCP to the application.

**Fallback:** The protocol should be able to fall back to standard TCP with no interference from the user, to be able to communicate with legacy hosts.

The complementary application considerations document [RFC6897] discusses the necessary features of an API to provide backwards-compatibility, as well as API extensions to convey the behavior of MPTCP at a level of control and information equivalent to that available with regular, single-path TCP.

Further discussion of the design constraints and associated design decisions are given in the MPTCP Architecture document [RFC6182] and in [howhard].

## 1.2. Multipath TCP in the Networking Stack

MPTCP operates at the transport layer and aims to be transparent to both higher and lower layers. It is a set of additional features on top of standard TCP; Figure 1 illustrates this layering. MPTCP is designed to be usable by legacy applications with no changes; detailed discussion of its interactions with applications is given in [RFC6897].

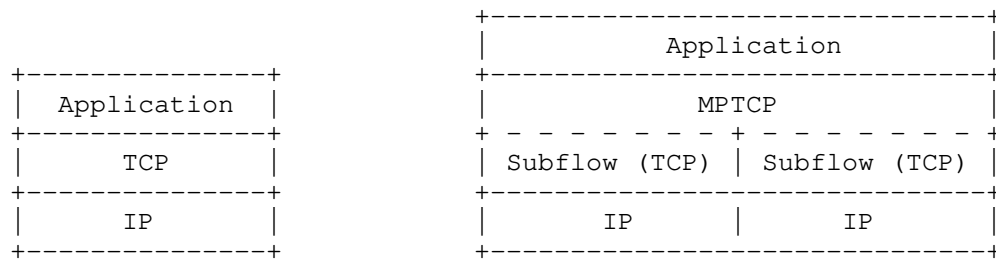


Figure 1: Comparison of Standard TCP and MPTCP Protocol Stacks

### 1.3. Terminology

This document makes use of a number of terms that are either MPTCP-specific or have defined meaning in the context of MPTCP, as follows:

**Path:** A sequence of links between a sender and a receiver, defined in this context by a 4-tuple of source and destination address/port pairs.

**Subflow:** A flow of TCP segments operating over an individual path, which forms part of a larger MPTCP connection. A subflow is started and terminated similar to a regular TCP connection.

**(MPTCP) Connection:** A set of one or more subflows, over which an application can communicate between two hosts. There is a one-to-one mapping between a connection and an application socket.

**Data-level:** The payload data is nominally transferred over a connection, which in turn is transported over subflows. Thus, the term "data-level" is synonymous with "connection level", in contrast to "subflow-level", which refers to properties of an individual subflow.

**Token:** A locally unique identifier given to a multipath connection by a host. May also be referred to as a "Connection ID".

**Host:** An end host operating an MPTCP implementation, and either initiating or accepting an MPTCP connection.

In addition to these terms, note that MPTCP's interpretation of, and effect on, regular single-path TCP semantics are discussed in Section 4.

#### 1.4. MPTCP Concept

This section provides a high-level summary of normal operation of MPTCP, and is illustrated by the scenario shown in Figure 2. A detailed description of operation is given in Section 3.

- o To a non-MPTCP-aware application, MPTCP will behave the same as normal TCP. Extended APIs could provide additional control to MPTCP-aware applications [RFC6897]. An application begins by opening a TCP socket in the normal way. MPTCP signaling and operation are handled by the MPTCP implementation.
- o An MPTCP connection begins similarly to a regular TCP connection. This is illustrated in Figure 2 where an MPTCP connection is established between addresses A1 and B1 on Hosts A and B, respectively.
- o If extra paths are available, additional TCP sessions (termed MPTCP "subflows") are created on these paths, and are combined with the existing session, which continues to appear as a single connection to the applications at both ends. The creation of the additional TCP session is illustrated between Address A2 on Host A and Address B1 on Host B.
- o MPTCP identifies multiple paths by the presence of multiple addresses at hosts. Combinations of these multiple addresses equate to the additional paths. In the example, other potential paths that could be set up are A1<->B2 and A2<->B2. Although this additional session is shown as being initiated from A2, it could equally have been initiated from B1 or B2.
- o The discovery and setup of additional subflows will be achieved through a path management method; this document describes a mechanism by which a host can initiate new subflows by using its own additional addresses, or by signaling its available addresses to the other host.
- o MPTCP adds connection-level sequence numbers to allow the reassembly of segments arriving on multiple subflows with differing network delays.
- o Subflows are terminated as regular TCP connections, with a four-way FIN handshake. The MPTCP connection is terminated by a connection-level FIN.

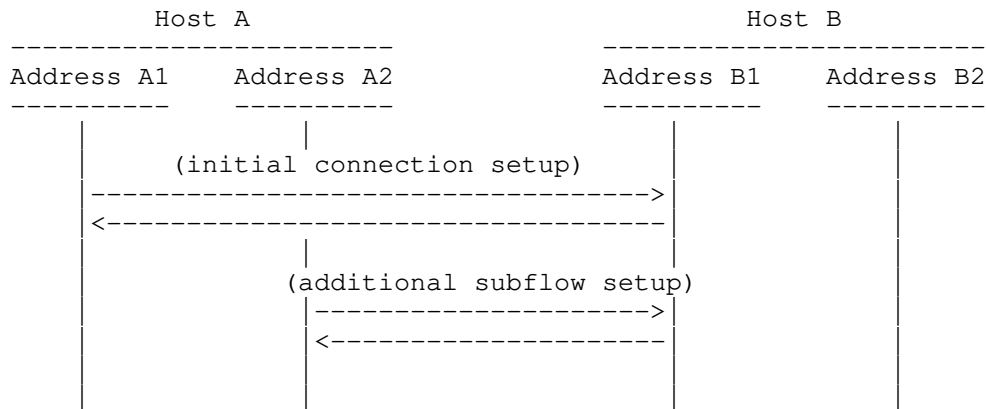


Figure 2: Example MPTCP Usage Scenario

### 1.5. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

## 2. Operation Overview

This section presents a single description of common MPTCP operation, with reference to the protocol operation. This is a high-level overview of the key functions; the full specification follows in Section 3. Extensibility and negotiated features are not discussed here. Considerable reference is made to symbolic names of MPTCP options throughout this section -- these are subtypes of the IANA-assigned MPTCP option (see Section 8), and their formats are defined in the detailed protocol specification that follows in Section 3.

A Multipath TCP connection provides a bidirectional bytestream between two hosts communicating like normal TCP and, thus, does not require any change to the applications. However, Multipath TCP enables the hosts to use different paths with different IP addresses to exchange packets belonging to the MPTCP connection. A Multipath TCP connection appears like a normal TCP connection to an application. However, to the network layer, each MPTCP subflow looks like a regular TCP flow whose segments carry a new TCP option type. Multipath TCP manages the creation, removal, and utilization of these subflows to send data. The number of subflows that are managed within a Multipath TCP connection is not fixed and it can fluctuate during the lifetime of the Multipath TCP connection.

All MPTCP operations are signaled with a TCP option -- a single numerical type for MPTCP, with "sub-types" for each MPTCP message. What follows is a summary of the purpose and rationale of these messages.

## 2.1. Initiating an MPTCP Connection

This is the same signaling as for initiating a normal TCP connection, but the SYN, SYN/ACK, and initial ACK (and data) packets also carry the MP\_CAPABLE option. This option has a variable length and serves multiple purposes. Firstly, it verifies whether the remote host supports Multipath TCP; secondly, this option allows the hosts to exchange some information to authenticate the establishment of additional subflows. Further details are given in Section 3.1.

Host A		Host B
-----		-----
MP_CAPABLE	->	
[flags]		
	<-	MP_CAPABLE
		[B's key, flags]
ACK + MP_CAPABLE (+ data)	->	
[A's key, B's key, flags, (data-level details)]		

Retransmission of the ACK + MP\_CAPABLE can occur if it is not known if it has been received. The following diagrams show all possible exchanges for the initial subflow setup to ensure this reliability.



```

Host A (with data to send immediately)  Host B
-----
MP_CAPABLE                               ->
[flags]                                  <-
                                         MP_CAPABLE
                                         [B's key, flags]

ACK + MP_CAPABLE + data                  ->
[A's key, B's key, flags, data-level details]

Host A (with data to send later)         Host B
-----
MP_CAPABLE                               ->
[flags]                                  <-
                                         MP_CAPABLE
                                         [B's key, flags]

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

ACK + MP_CAPABLE + data                  ->
[A's key, B's key, flags, data-level details]

Host A                                   Host B (sending first)
-----
MP_CAPABLE                               ->
[flags]                                  <-
                                         MP_CAPABLE
                                         [B's key, flags]

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

                                         <-
                                         ACK + DSS + data
                                         [data-level details]

```

## 2.2. Associating a New Subflow with an Existing MPTCP Connection

The exchange of keys in the MP\_CAPABLE handshake provides material that can be used to authenticate the endpoints when new subflows will be set up. Additional subflows begin in the same way as initiating a normal TCP connection, but the SYN, SYN/ACK, and ACK packets also carry the MP\_JOIN option.

Host A initiates a new subflow between one of its addresses and one of Host B's addresses. The token -- generated from the key -- is used to identify which MPTCP connection it is joining, and the HMAC is used for authentication. The Hash-based Message Authentication Code (HMAC) uses the keys exchanged in the MP\_CAPABLE handshake, and

the random numbers (nonces) exchanged in these MP\_JOIN options. MP\_JOIN also contains flags and an Address ID that can be used to refer to the source address without the sender needing to know if it has been changed by a NAT. Further details are in Section 3.2.

```

Host A                               Host B
-----                               -----
MP_JOIN                               ->
[B's token, A's nonce,
 A's Address ID, flags]
                                     <-
                                     MP_JOIN
                                     [B's HMAC, B's nonce,
                                     B's Address ID, flags]

ACK + MP_JOIN                         ->
[A's HMAC]

                                     <-
                                     ACK

```

### 2.3. Informing the Other Host about Another Potential Address

The set of IP addresses associated to a multihomed host may change during the lifetime of an MPTCP connection. MPTCP supports the addition and removal of addresses on a host both implicitly and explicitly. If Host A has established a subflow starting at address/port pair IP#-A1 and wants to open a second subflow starting at address/port pair IP#-A2, it simply initiates the establishment of the subflow as explained above. The remote host will then be implicitly informed about the new address.

In some circumstances, a host may want to advertise to the remote host the availability of an address without establishing a new subflow, for example, when a NAT prevents setup in one direction. In the example below, Host A informs Host B about its alternative IP address/port pair (IP#-A2). Host B may later send an MP\_JOIN to this new address. The ADD\_ADDR option contains a HMAC to authenticate the address as having been sent from the originator of the connection. The receiver of this option echoes it back to the client to indicate successful receipt. Further details are in Section 3.4.1.

```

Host A                                     Host B
-----                                     -----
ADD_ADDR                                ->
[Echo-flag=0,
 IP#-A2,
 IP#-A2's Address ID,
 HMAC of IP#-A2]

                                     <-
                                     ADD_ADDR
                                     [Echo-flag=1,
                                     IP#-A2,
                                     IP#-A2's Address ID,
                                     HMAC of IP#-A2]

```

There is a corresponding signal for address removal, making use of the Address ID that is signaled in the add address handshake. Further details in Section 3.4.2.

```

Host A                                     Host B
-----                                     -----
REMOVE_ADDR                             ->
[IP#-A2's Address ID]

```

#### 2.4. Data Transfer Using MPTCP

To ensure reliable, in-order delivery of data over subflows that may appear and disappear at any time, MPTCP uses a 64-bit data sequence number (DSN) to number all data sent over the MPTCP connection. Each subflow has its own 32-bit sequence number space, utilising the regular TCP sequence number header, and an MPTCP option maps the subflow sequence space to the data sequence space. In this way, data can be retransmitted on different subflows (mapped to the same DSN) in the event of failure.

The Data Sequence Signal (DSS) carries the Data Sequence Mapping. The Data Sequence Mapping consists of the subflow sequence number, data sequence number, and length for which this mapping is valid. This option can also carry a connection-level acknowledgment (the "Data ACK") for the received DSN.

With MPTCP, all subflows share the same receive buffer and advertise the same receive window. There are two levels of acknowledgment in MPTCP. Regular TCP acknowledgments are used on each subflow to acknowledge the reception of the segments sent over the subflow independently of their DSN. In addition, there are connection-level acknowledgments for the data sequence space. These acknowledgments track the advancement of the bytestream and slide the receiving window.

Further details are in Section 3.3.

```

Host A                               Host B
-----                               -----
DSS                                  ->
[Data Sequence Mapping]
[Data ACK]
[Checksum]

```

## 2.5. Requesting a Change in a Path's Priority

Hosts can indicate at initial subflow setup whether they wish the subflow to be used as a regular or backup path -- a backup path only being used if there are no regular paths available. During a connection, Host A can request a change in the priority of a subflow through the MP\_PRIO signal to Host B. Further details are in Section 3.3.8.

```

Host A                               Host B
-----                               -----
MP_PRIO                             ->

```

## 2.6. Closing an MPTCP Connection

When a host wants to close an existing subflow, but not the whole connection, it can initiate a regular TCP FIN/ACK exchange.

When Host A wants to inform Host B that it has no more data to send, it signals this "Data FIN" as part of the Data Sequence Signal (see above). It has the same semantics and behavior as a regular TCP FIN, but at the connection level. Once all the data on the MPTCP connection has been successfully received, then this message is acknowledged at the connection level with a Data ACK. Further details are in Section 3.3.3.

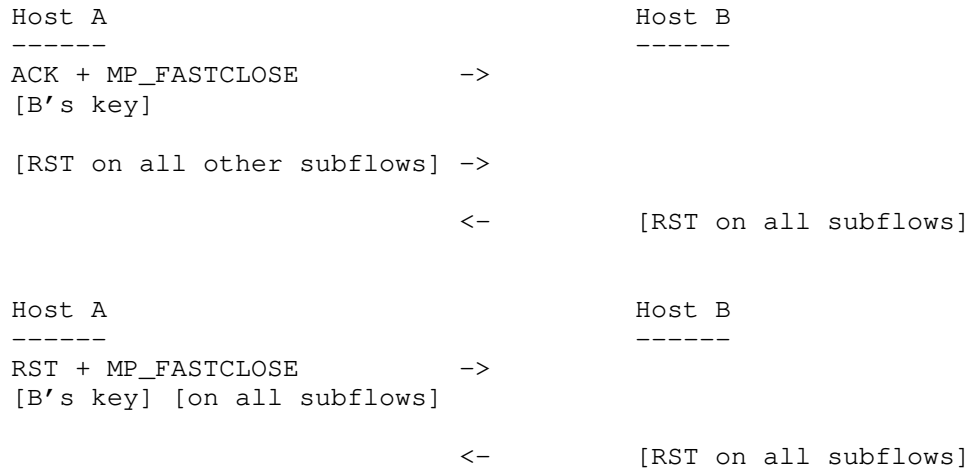
```

Host A                               Host B
-----                               -----
DSS                                  ->
[Data FIN]
<-
DSS
[Data ACK]

```

There is an additional method of connection closure, referred to as "Fast Close", which is analogous to closing a single-path TCP connection with a RST signal. The MP\_FASTCLOSE signal is used to indicate to the peer that the connection will be abruptly closed and no data will be accepted anymore. This can be used on an ACK (ensuring reliability of the signal), or a RST (which is not). Both

examples are shown in the following diagrams. Further details are in Section 3.5.



## 2.7. Notable Features

It is worth highlighting that MPTCP's signaling has been designed with several key requirements in mind:

- o To cope with NATs on the path, addresses are referred to by Address IDs, in case the IP packet's source address gets changed by a NAT. Setting up a new TCP flow is not possible if the receiver of the SYN is behind a NAT; to allow subflows to be created when either end is behind a NAT, MPTCP uses the ADD\_ADDR message.
- o MPTCP falls back to ordinary TCP if MPTCP operation is not possible, for example, if one host is not MPTCP capable or if a middlebox alters the payload. This is discussed in Section 3.7.
- o To address the threats identified in [RFC6181], the following steps are taken: keys are sent in the clear in the MP\_CAPABLE messages; MP\_JOIN messages are secured with HMAC-SHA256 ([RFC2104], [RFC6234]) using those keys; and standard TCP validity checks are made on the other messages (ensuring sequence numbers are in-window [RFC5961]). Residual threats to MPTCP v0 were identified in [RFC7430], and those affecting the protocol (i.e. modification to ADD\_ADDR) have been incorporated in this document. Further discussion of security can be found in Section 5.

### 3. MPTCP Protocol

This section describes the operation of the MPTCP protocol, and is subdivided into sections for each key part of the protocol operation.

All MPTCP operations are signaled using optional TCP header fields. A single TCP option number ("Kind") has been assigned by IANA for MPTCP (see Section 8), and then individual messages will be determined by a "subtype", the values of which are also stored in an IANA registry (and are also listed in Section 8). As with all TCP options, the Length field is specified in bytes, and includes the 2 bytes of Kind and Length.

Throughout this document, when reference is made to an MPTCP option by symbolic name, such as "MP\_CAPABLE", this refers to a TCP option with the single MPTCP option type, and with the subtype value of the symbolic name as defined in Section 8. This subtype is a 4-bit field -- the first 4 bits of the option payload, as shown in Figure 3. The MPTCP messages are defined in the following sections.

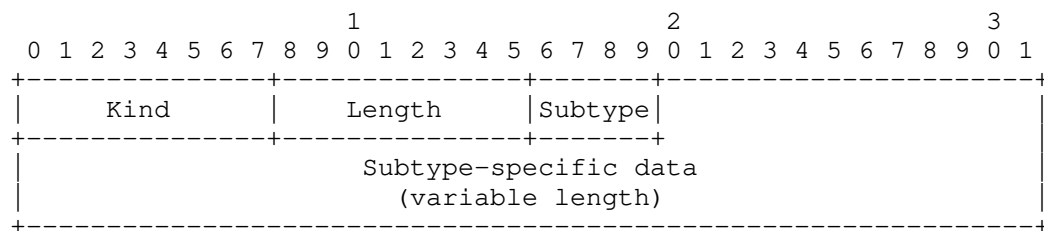


Figure 3: MPTCP Option Format

Those MPTCP options associated with subflow initiation are used on packets with the SYN flag set. Additionally, there is one MPTCP option for signaling metadata to ensure segmented data can be recombined for delivery to the application.

The remaining options, however, are signals that do not need to be on a specific packet, such as those for signaling additional addresses. Whilst an implementation may desire to send MPTCP options as soon as possible, it may not be possible to combine all desired options (both those for MPTCP and for regular TCP, such as SACK (selective acknowledgment) [RFC2018]) on a single packet. Therefore, an implementation may choose to send duplicate ACKs containing the additional signaling information. This changes the semantics of a duplicate ACK; these are usually only sent as a signal of a lost segment [RFC5681] in regular TCP. Therefore, an MPTCP implementation receiving a duplicate ACK that contains an MPTCP option MUST NOT treat it as a signal of congestion. Additionally, an MPTCP

implementation SHOULD NOT send more than two duplicate ACKs in a row for the purposes of sending MPTCP options alone, in order to ensure no middleboxes misinterpret this as a sign of congestion.

Furthermore, standard TCP validity checks (such as ensuring the sequence number and acknowledgment number are within window) MUST be undertaken before processing any MPTCP signals, as described in [RFC5961], and initial subflow sequence numbers SHOULD be generated according to the recommendations in [RFC6528].

### 3.1. Connection Initiation

Connection initiation begins with a SYN, SYN/ACK, ACK exchange on a single path. Each packet contains the Multipath Capable (MP\_CAPABLE) MPTCP option (Figure 4). This option declares its sender is capable of performing Multipath TCP and wishes to do so on this particular connection.

The MP\_CAPABLE exchange in this specification (v1) is different to that specified in v0. If a host supports multiple versions of MPTCP, the sender of the MP\_CAPABLE option SHOULD signal the highest version number it supports. In return, in its MP\_CAPABLE option, the receiver will signal the version number it wishes to use, which MUST be equal to or lower than the version number indicated in the initial MP\_CAPABLE. There is a caveat though with respect to this version negotiation with old listeners that only support v0. A listener that supports v0 expects that the MP\_CAPABLE option in the SYN-segment includes the initiator's key. If the initiator however already upgraded to v1, it won't include the key in the SYN-segment. Thus, the listener will ignore the MP\_CAPABLE of this SYN-segment and reply with a SYN/ACK that does not include an MP\_CAPABLE. The initiator MAY choose to immediately fall back to TCP or MAY choose to attempt a connection using MPTCP v0 (if the initiator supports v0), in order to discover whether the listener supports the earlier version of MPTCP. In general a MPTCP v0 connection is likely to be preferred to a TCP one, however in a particular deployment scenario it may be known that the listener is unlikely to support MPTCPv0 and so the initiator may prefer not to attempt a v0 connection. An initiator MAY cache information for a peer about what version of MPTCP it supports if any, and use this information for future connection attempts.

The MP\_CAPABLE option is variable-length, with different fields included depending on which packet the option is used on. The full MP\_CAPABLE option is shown in Figure 4.

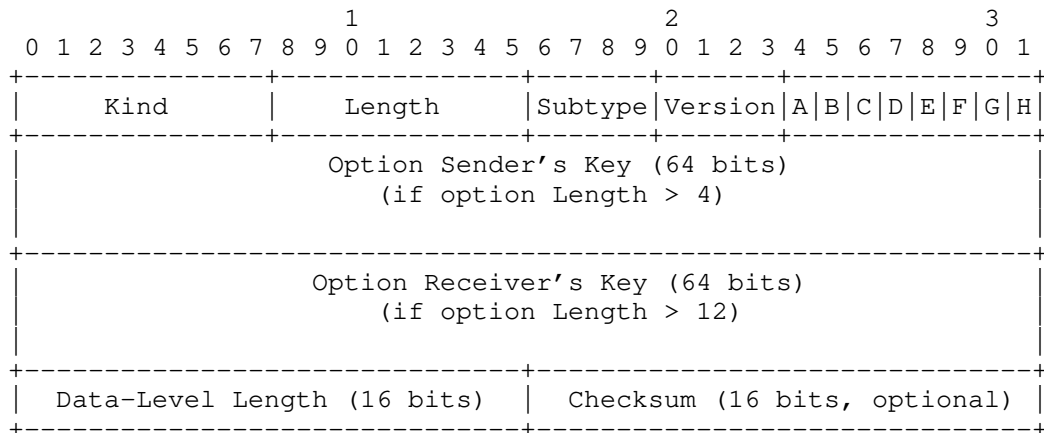


Figure 4: Multipath Capable (MP\_CAPABLE) Option

The MP\_CAPABLE option is carried on the SYN, SYN/ACK, and ACK packets that start the first subflow of an MPTCP connection, as well as the first packet that carries data, if the initiator wishes to send first. The data carried by each option is as follows, where A = initiator and B = listener.

- o SYN (A->B): only the first four octets (Length = 4).
- o SYN/ACK (B->A): B's Key for this connection (Length = 12).
- o ACK (no data) (A->B): A's Key followed by B's Key (Length = 20).
- o ACK (with first data) (A->B): A's Key followed by B's Key followed by Data-Level Length, and optional Checksum (Length = 22 or 24).

The contents of the option is determined by the SYN and ACK flags of the packet, along with the option's length field. For the diagram shown in Figure 4, "sender" and "receiver" refer to the sender or receiver of the TCP packet (which can be either host).

The initial SYN, containing just the MP\_CAPABLE header, is used to define the version of MPTCP being requested, as well as exchanging flags to negotiate connection features, described later.

This option is used to declare the 64-bit keys that the end hosts have generated for this MPTCP connection. These keys are used to authenticate the addition of future subflows to this connection. This is the only time the key will be sent in clear on the wire (unless "fast close", Section 3.5, is used); all future subflows will identify the connection using a 32-bit "token". This token is a



cryptographic hash of this key. The algorithm for this process is dependent on the authentication algorithm selected; the method of selection is defined later in this section.

Upon reception of the initial SYN-segment, a stateful server generates a random key and replies with a SYN/ACK. The key's method of generation is implementation specific. The key **MUST** be hard to guess, and it **MUST** be unique for the sending host across all its current MPTCP connections. Recommendations for generating random numbers for use in keys are given in [RFC4086]. Connections will be indexed at each host by the token (a one-way hash of the key). Therefore, an implementation will require a mapping from each token to the corresponding connection, and in turn to the keys for the connection.

There is a risk that two different keys will hash to the same token. The risk of hash collisions is usually small, unless the host is handling many tens of thousands of connections. Therefore, an implementation **SHOULD** check its list of connection tokens to ensure there is no collision before sending its key, and if there is, then it should generate a new key. This would, however, be costly for a server with thousands of connections. The subflow handshake mechanism (Section 3.2) will ensure that new subflows only join the correct connection, however, through the cryptographic handshake, as well as checking the connection tokens in both directions, and ensuring sequence numbers are in-window. So in the worst case if there was a token collision, the new subflow would not succeed, but the MPTCP connection would continue to provide a regular TCP service.

Since key generation is implementation-specific, there is no requirement that they be simply random numbers. An implementation is free to exchange cryptographic material out-of-band and generate these keys from this, in order to provide additional mechanisms by which to verify the identity of the communicating entities. For example, an implementation could choose to link its MPTCP keys to those used in higher-layer TLS or SSH connections.

If the server behaves in a stateless manner, it has to generate its own key in a verifiable fashion. This verifiable way of generating the key can be done by using a hash of the 4-tuple, sequence number and a local secret (similar to what is done for the TCP-sequence number [RFC4987]). It will thus be able to verify whether it is indeed the originator of the key echoed back in the later MP\_CAPABLE option. As for a stateful server, the tokens **SHOULD** be checked for uniqueness, however if uniqueness is not met, and there is no way to generate an alternative verifiable key, then the connection **MUST** fall back to using regular TCP by not sending a MP\_CAPABLE in the SYN/ACK.

The ACK carries both A's key and B's key. This is the first time that A's key is seen on the wire, although it is expected that A will have generated a key locally before the initial SYN. The echoing of B's key allows B to operate statelessly, as described above. Therefore, A's key must be delivered reliably to B, and in order to do this, the transmission of this packet must be made reliable.

If B has data to send first, then the reliable delivery of the ACK+MP\_CAPABLE can be inferred by the receipt of this data with a MPTCP Data Sequence Signal (DSS) option (Section 3.3). If, however, A wishes to send data first, it has two options to ensure the reliable delivery of the ACK+MP\_CAPABLE. If it immediately has data to send, then the third ACK (with data) would also contain an MP\_CAPABLE option with additional data parameters (the Data-Level Length and optional Checksum as shown in Figure 4). If A does not immediately have data to send, it MUST include the MP\_CAPABLE on the third ACK, but without the additional data parameters. When A does have data to send, it must repeat the sending of the MP\_CAPABLE option from the third ACK, with additional data parameters. This MP\_CAPABLE option is in place of the DSS, and simply specifies the data-level length of the payload, and the checksum (if the use of checksums is negotiated). This is the minimal data required to establish a MPTCP connection - it allows validation of the payload, and given it is the first data, the Initial Data Sequence Number (IDSN) is also known (as it is generated from the key, as described below). Conveying the keys on the first data packet allows the TCP reliability mechanisms to ensure the packet is successfully delivered. The receiver will acknowledge this data at the connection level with a Data ACK, as if a DSS option has been received.

There could be situations where both A and B attempt to transmit initial data at the same time. For example, if A did not initially have data to send, but then needed to transmit data before it had received anything from B, it would use a MP\_CAPABLE option with data parameters (since it would not know if the MP\_CAPABLE on the ACK was received). In such a situation, B may also have transmitted data with a DSS option, but it had not yet been received at A. Therefore, B has received data with a MP\_CAPABLE mapping after it has sent data with a DSS option. To ensure these situations can be handled, it follows that the data parameters in a MP\_CAPABLE are semantically equivalent to those in a DSS option and can be used interchangeably. Similar situations could occur when the MP\_CAPABLE with data is lost and retransmitted. Furthermore, in the case of TCP Segmentation Offloading, the MP\_CAPABLE with data parameters may be duplicated across multiple packets, and implementations must also be able to cope with duplicate MP\_CAPABLE mappings as well as duplicate DSS mappings.

Additionally, the MP\_CAPABLE exchange allows the safe passage of MPTCP options on SYN packets to be determined. If any of these options are dropped, MPTCP will gracefully fall back to regular single-path TCP, as documented in Section 3.7. If at any point in the handshake either party thinks the MPTCP negotiation is compromised, for example by a middlebox corrupting the TCP options, or unexpected ACK numbers being present, the host MUST stop using MPTCP and no longer include MPTCP options in future TCP packets. The other host will then also fall back to regular TCP using the fall back mechanism. Note that new subflows MUST NOT be established (using the process documented in Section 3.2) until a Data Sequence Signal (DSS) option has been successfully received across the path (as documented in Section 3.3).

Like all MPTCP options, the MP\_CAPABLE option starts with the Kind and Length to specify the TCP-option kind and its length. Followed by that is the MP\_CAPABLE option. The first 4 bits of the first octet in the MP\_CAPABLE option (Figure 4) define the MPTCP option subtype (see Section 8; for MP\_CAPABLE, this is 0x0), and the remaining 4 bits of this octet specify the MPTCP version in use (for this specification, this is 1).

The second octet is reserved for flags, allocated as follows:

- A: The leftmost bit, labeled "A", SHOULD be set to 1 to indicate "Checksum Required", unless the system administrator has decided that checksums are not required (for example, if the environment is controlled and no middleboxes exist that might adjust the payload).
- B: The second bit, labeled "B", is an extensibility flag, and MUST be set to 0 for current implementations. This will be used for an extensibility mechanism in a future specification, and the impact of this flag will be defined at a later date. It is expected, but not mandated, that this flag would be used as part of an alternative security mechanism that does not require a full version upgrade of the protocol, but does require redefining some elements of the handshake. If receiving a message with the 'B' flag set to 1, and this is not understood, then the MP\_CAPABLE in this SYN MUST be silently ignored, which triggers a fallback to regular TCP; the sender is expected to retry with a format compatible with this legacy specification. Note that the length of the MP\_CAPABLE option, and the meanings of bits "D" through "H", may be altered by setting B=1.
- C: The third bit, labeled "C", is set to "1" to indicate that the sender of this option will not accept additional MPTCP subflows to the source address and port, and therefore the receiver MUST NOT

try to open any additional subflows towards this address and port. This is an efficiency improvement for situations where the sender knows a restriction is in place, for example if the sender is behind a strict NAT, or operating behind a legacy Layer 4 load balancer.

D through H: The remaining bits, labeled "D" through "H", are used for crypto algorithm negotiation. In this specification only the rightmost bit, labeled "H", is assigned. Bit "H" indicates the use of HMAC-SHA256 (as defined in Section 3.2). An implementation that only supports this method MUST set bit "H" to 1, and bits "D" through "G" to 0.

A crypto algorithm MUST be specified. If flag bits D through H are all 0, the MP\_CAPABLE option MUST be treated as invalid and ignored (that is, it must be treated as a regular TCP handshake).

The selection of the authentication algorithm also impacts the algorithm used to generate the token and the Initial Data Sequence Number (IDSN). In this specification, with only the SHA-256 algorithm (bit "H") specified and selected, the token MUST be a truncated (most significant 32 bits) SHA-256 hash ([RFC6234]) of the key. A different, 64-bit truncation (the least significant 64 bits) of the SHA-256 hash of the key MUST be used as the IDSN. Note that the key MUST be hashed in network byte order. Also note that the "least significant" bits MUST be the rightmost bits of the SHA-256 digest, as per [RFC6234]. Future specifications of the use of the crypto bits may choose to specify different algorithms for token and IDSN generation.

Both the crypto and checksum bits negotiate capabilities in similar ways. For the Checksum Required bit (labeled "A"), if either host requires the use of checksums, checksums MUST be used. In other words, the only way for checksums not to be used is if both hosts in their SYNs set A=0. This decision is confirmed by the setting of the "A" bit in the third packet (the ACK) of the handshake. For example, if the initiator sets A=0 in the SYN, but the responder sets A=1 in the SYN/ACK, checksums MUST be used in both directions, and the initiator will set A=1 in the ACK. The decision whether to use checksums will be stored by an implementation in a per-connection binary state variable. If A=1 is received by a host that does not want to use checksums, it MUST fall back to regular TCP by ignoring the MP\_CAPABLE option as if it was invalid.

For crypto negotiation, the responder has the choice. The initiator creates a proposal setting a bit for each algorithm it supports to 1 (in this version of the specification, there is only one proposal, so bit "H" will be always set to 1). The responder responds with only 1

bit set -- this is the chosen algorithm. The rationale for this behavior is that the responder will typically be a server with potentially many thousands of connections, so it may wish to choose an algorithm with minimal computational complexity, depending on the load. If a responder does not support (or does not want to support) any of the initiator's proposals, it **MUST** respond without an MP\_CAPABLE option, thus forcing a fallback to regular TCP.

The MP\_CAPABLE option is only used in the first subflow of a connection, in order to identify the connection; all following subflows will use the "Join" option (see Section 3.2) to join the existing connection.

If a SYN contains an MP\_CAPABLE option but the SYN/ACK does not, it is assumed that sender of the SYN/ACK is not multipath capable; thus, the MPTCP session **MUST** operate as a regular, single-path TCP. If a SYN does not contain a MP\_CAPABLE option, the SYN/ACK **MUST NOT** contain one in response. If the third packet (the ACK) does not contain the MP\_CAPABLE option, then the session **MUST** fall back to operating as a regular, single-path TCP. This is to maintain compatibility with middleboxes on the path that drop some or all TCP options. Note that an implementation **MAY** choose to attempt sending MPTCP options more than one time before making this decision to operate as regular TCP (see Section 3.9).

If the SYN packets are unacknowledged, it is up to local policy to decide how to respond. It is expected that a sender will eventually fall back to single-path TCP (i.e., without the MP\_CAPABLE option) in order to work around middleboxes that may drop packets with unknown options; however, the number of multipath-capable attempts that are made first will be up to local policy. It is possible that MPTCP and non-MPTCP SYNs could get reordered in the network. Therefore, the final state is inferred from the presence or absence of the MP\_CAPABLE option in the third packet of the TCP handshake. If this option is not present, the connection **SHOULD** fall back to regular TCP, as documented in Section 3.7.

The initial data sequence number on an MPTCP connection is generated from the key. The algorithm for IDSN generation is also determined from the negotiated authentication algorithm. In this specification, with only the SHA-256 algorithm specified and selected, the IDSN of a host **MUST** be the least significant 64 bits of the SHA-256 hash of its key, i.e., IDSN-A = Hash(Key-A) and IDSN-B = Hash(Key-B). This deterministic generation of the IDSN allows a receiver to ensure that there are no gaps in sequence space at the start of the connection. The SYN with MP\_CAPABLE occupies the first octet of data sequence space, although this does not need to be acknowledged at the connection level until the first data is sent (see Section 3.3).

### 3.2. Starting a New Subflow

Once an MPTCP connection has begun with the MP\_CAPABLE exchange, further subflows can be added to the connection. Hosts have knowledge of their own address(es), and can become aware of the other host's addresses through signaling exchanges as described in Section 3.4. Using this knowledge, a host can initiate a new subflow over a currently unused pair of addresses. It is permitted for either host in a connection to initiate the creation of a new subflow, but it is expected that this will normally be the original connection initiator (see Section 3.9 for heuristics).

A new subflow is started as a normal TCP SYN/ACK exchange. The Join Connection (MP\_JOIN) MPTCP option is used to identify the connection to be joined by the new subflow. It uses keying material that was exchanged in the initial MP\_CAPABLE handshake (Section 3.1), and that handshake also negotiates the crypto algorithm in use for the MP\_JOIN handshake.

This section specifies the behavior of MP\_JOIN using the HMAC-SHA256 algorithm. An MP\_JOIN option is present in the SYN, SYN/ACK, and ACK of the three-way handshake, although in each case with a different format.

In the first MP\_JOIN on the SYN packet, illustrated in Figure 5, the initiator sends a token, random number, and address ID.

The token is used to identify the MPTCP connection and is a cryptographic hash of the receiver's key, as exchanged in the initial MP\_CAPABLE handshake (Section 3.1). In this specification, the tokens presented in this option are generated by the SHA-256 [RFC6234] algorithm, truncated to the most significant 32 bits. The token included in the MP\_JOIN option is the token that the receiver of the packet uses to identify this connection; i.e., Host A will send Token-B (which is generated from Key-B). Note that the hash generation algorithm can be overridden by the choice of cryptographic handshake algorithm, as defined in Section 3.1.

The MP\_JOIN SYN sends not only the token (which is static for a connection) but also random numbers (nonces) that are used to prevent replay attacks on the authentication method. Recommendations for the generation of random numbers for this purpose are given in [RFC4086].

The MP\_JOIN option includes an "Address ID". This is an identifier generated by the sender of the option, used to identify the source address of this packet, even if the IP header has been changed in transit by a middlebox. The numeric value of this field is generated by the sender and must map uniquely to a source IP address for the

sending host. The Address ID allows address removal (Section 3.4.2) without needing to know what the source address at the receiver is, thus allowing address removal through NATs. The Address ID also allows correlation between new subflow setup attempts and address signaling (Section 3.4.1), to prevent setting up duplicate subflows on the same path, if an MP\_JOIN and ADD\_ADDR are sent at the same time.

The Address IDs of the subflow used in the initial SYN exchange of the first subflow in the connection are implicit, and have the value zero. A host MUST store the mappings between Address IDs and addresses both for itself and the remote host. An implementation will also need to know which local and remote Address IDs are associated with which established subflows, for when addresses are removed from a local or remote host.

The MP\_JOIN option on packets with the SYN flag set also includes 4 bits of flags, 3 of which are currently reserved and MUST be set to zero by the sender. The final bit, labeled "B", indicates whether the sender of this option wishes this subflow to be used as a backup path (B=1) in the event of failure of other paths, or whether it wants it to be used as part of the connection immediately. By setting B=1, the sender of the option is requesting the other host to only send data on this subflow if there are no available subflows where B=0. Subflow policy is discussed in more detail in Section 3.3.8.

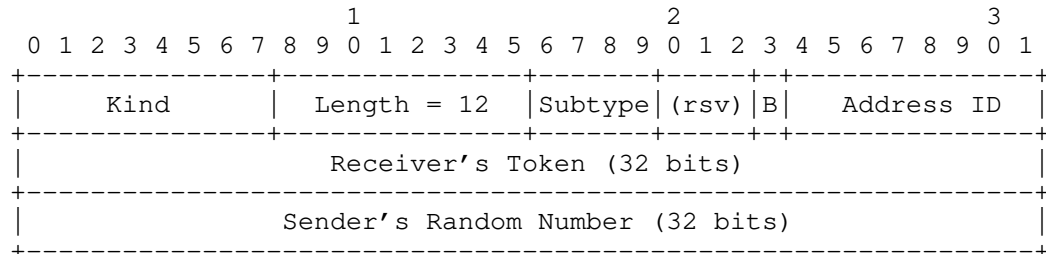


Figure 5: Join Connection (MP\_JOIN) Option (for Initial SYN)

When receiving a SYN with an MP\_JOIN option that contains a valid token for an existing MPTCP connection, the recipient SHOULD respond with a SYN/ACK also containing an MP\_JOIN option containing a random number and a truncated (leftmost 64 bits) Hash-based Message Authentication Code (HMAC). This version of the option is shown in Figure 6. If the token is unknown, or the host wants to refuse subflow establishment (for example, due to a limit on the number of subflows it will permit), the receiver will send back a reset (RST) signal, analogous to an unknown port in TCP, containing a MP\_TCP\_RST

option (Section 3.6) with a "MPTCP specific error" reason code. Although calculating an HMAC requires cryptographic operations, it is believed that the 32-bit token in the MP\_JOIN SYN gives sufficient protection against blind state exhaustion attacks; therefore, there is no need to provide mechanisms to allow a responder to operate statelessly at the MP\_JOIN stage.

An HMAC is sent by both hosts -- by the initiator (Host A) in the third packet (the ACK) and by the responder (Host B) in the second packet (the SYN/ACK). Doing the HMAC exchange at this stage allows both hosts to have first exchanged random data (in the first two SYN packets) that is used as the "message". This specification defines that HMAC as defined in [RFC2104] is used, along with the SHA-256 hash algorithm [RFC6234], and that the output is truncated to the leftmost 160 bits (20 octets). Due to option space limitations, the HMAC included in the SYN/ACK is truncated to the leftmost 64 bits, but this is acceptable since random numbers are used; thus, an attacker only has one chance to correctly guess the HMAC that matches the random number previously sent by the peer (if the HMAC is incorrect, the TCP connection is closed, so a new MP\_JOIN negotiation with a new random number is required).

The initiator's authentication information is sent in its first ACK (the third packet of the handshake), as shown in Figure 7. This data needs to be sent reliably, since it is the only time this HMAC is sent; therefore, receipt of this packet MUST trigger a regular TCP ACK in response, and the packet MUST be retransmitted if this ACK is not received. In other words, sending the ACK/MP\_JOIN packet places the subflow in the PRE\_ESTABLISHED state, and it moves to the ESTABLISHED state only on receipt of an ACK from the receiver. It is not permitted to send data while in the PRE\_ESTABLISHED state. The reserved bits in this option MUST be set to zero by the sender.

The key for the HMAC algorithm, in the case of the message transmitted by Host A, will be Key-A followed by Key-B, and in the case of Host B, Key-B followed by Key-A. These are the keys that were exchanged in the original MP\_CAPABLE handshake. The "message" for the HMAC algorithm in each case is the concatenations of random number for each host (denoted by R): for Host A, R-A followed by R-B; and for Host B, R-B followed by R-A.



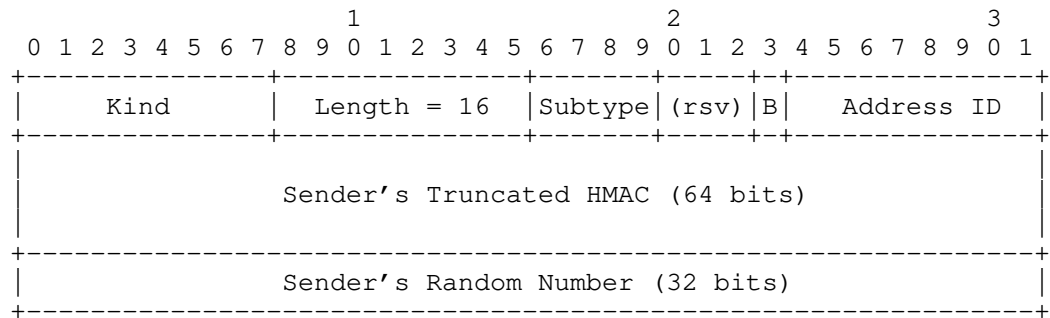


Figure 6: Join Connection (MP\_JOIN) Option (for Responding SYN/ACK)

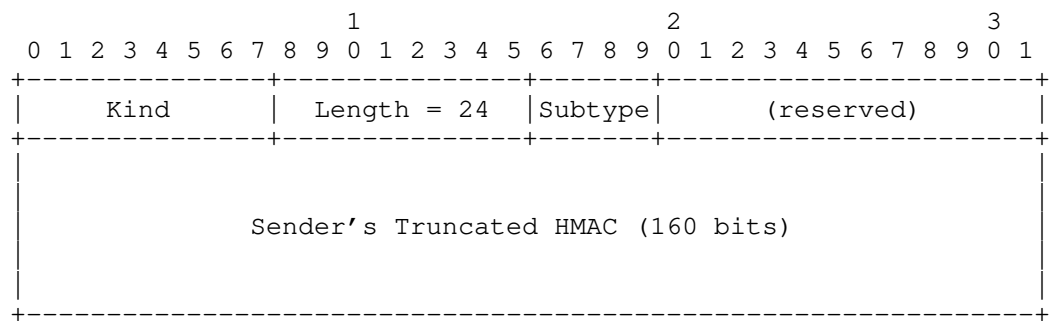
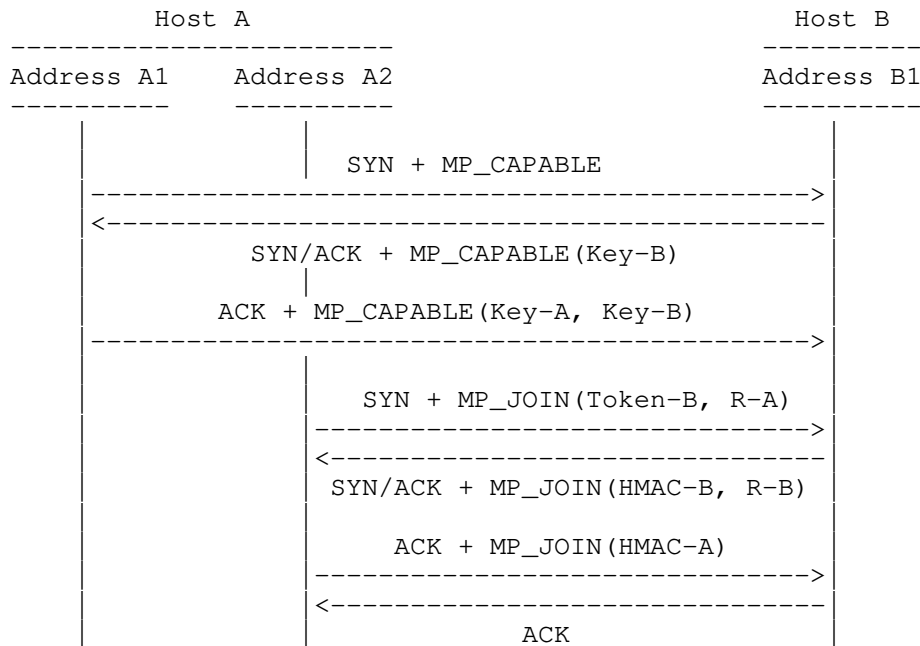


Figure 7: Join Connection (MP\_JOIN) Option (for Third ACK)

These various MPTCP options fit together to enable authenticated subflow setup as illustrated in Figure 8.



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 8: Example Use of MPTCP Authentication

If the token received at Host B is unknown or local policy prohibits the acceptance of the new subflow, the recipient MUST respond with a TCP RST for the subflow. If appropriate, a MP\_TCPRST option with a "Administratively prohibited" reason code (Section 3.6) should be included.

If the token is accepted at Host B, but the HMAC returned to Host A does not match the one expected, Host A MUST close the subflow with a TCP RST. In this, and all following cases of sending a RST in this section, the sender SHOULD send a MP\_TCPRST option (Section 3.6) on this RST packet with the reason code for a "MPTCP specific error".

If Host B does not receive the expected HMAC, or the MP\_JOIN option is missing from the ACK, it MUST close the subflow with a TCP RST.

If the HMACs are verified as correct, then both hosts have verified each other as being the same peers as existed at the start of the connection, and they have agreed of which connection this subflow will become a part.

If the SYN/ACK as received at Host A does not have an MP\_JOIN option, Host A MUST close the subflow with a TCP RST.

This covers all cases of the loss of an MP\_JOIN. In more detail, if MP\_JOIN is stripped from the SYN on the path from A to B, and Host B does not have a listener on the relevant port, it will respond with a RST in the normal way. If in response to a SYN with an MP\_JOIN option, a SYN/ACK is received without the MP\_JOIN option (either since it was stripped on the return path, or it was stripped on the outgoing path but Host B responded as if it were a new regular TCP session), then the subflow is unusable and Host A MUST close it with a RST.

Note that additional subflows can be created between any pair of ports (but see Section 3.9 for heuristics); no explicit application-level accept calls or bind calls are required to open additional subflows. To associate a new subflow with an existing connection, the token supplied in the subflow's SYN exchange is used for demultiplexing. This then binds the 5-tuple of the TCP subflow to the local token of the connection. A consequence is that it is possible to allow any port pairs to be used for a connection.

Demultiplexing subflow SYNs MUST be done using the token; this is unlike traditional TCP, where the destination port is used for demultiplexing SYN packets. Once a subflow is set up, demultiplexing packets is done using the 5-tuple, as in traditional TCP. The 5-tuples will be mapped to the local connection identifier (token). Note that Host A will know its local token for the subflow even though it is not sent on the wire -- only the responder's token is sent.

### 3.3. General MPTCP Operation

This section discusses operation of MPTCP for data transfer. At a high level, an MPTCP implementation will take one input data stream from an application, and split it into one or more subflows, with sufficient control information to allow it to be reassembled and delivered reliably and in order to the recipient application. The following subsections define this behavior in detail.

The data sequence mapping and the Data ACK are signaled in the Data Sequence Signal (DSS) option (Figure 9). Either or both can be signaled in one DSS, depending on the flags set. The data sequence mapping defines how the sequence space on the subflow maps to the connection level, and the Data ACK acknowledges receipt of data at the connection level. These functions are described in more detail in the following two subsections.

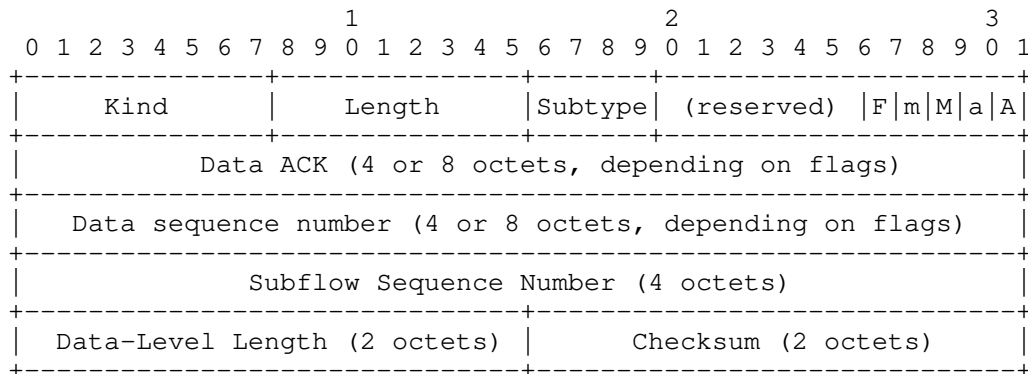


Figure 9: Data Sequence Signal (DSS) Option

The flags, when set, define the contents of this option, as follows:

- o A = Data ACK present
- o a = Data ACK is 8 octets (if not set, Data ACK is 4 octets)
- o M = Data Sequence Number (DSN), Subflow Sequence Number (SSN), Data-Level Length, and Checksum (if negotiated) present
- o m = Data sequence number is 8 octets (if not set, DSN is 4 octets)

The flags 'a' and 'm' only have meaning if the corresponding 'A' or 'M' flags are set; otherwise, they will be ignored. The maximum length of this option, with all flags set, is 28 octets.

The 'F' flag indicates "Data FIN". If present, this means that this mapping covers the final data from the sender. This is the connection-level equivalent to the FIN flag in single-path TCP. A connection is not closed unless there has been a Data FIN exchange, a MP\_FASTCLOSE (Section 3.5) message, or an implementation-specific, connection-level send timeout. The purpose of the Data FIN and the interactions between this flag, the subflow-level FIN flag, and the data sequence mapping are described in Section 3.3.3. The remaining reserved bits MUST be set to zero by an implementation of this specification.

Note that the checksum is only present in this option if the use of MPTCP checksumming has been negotiated at the MP\_CAPABLE handshake (see Section 3.1). The presence of the checksum can be inferred from the length of the option. If a checksum is present, but its use had not been negotiated in the MP\_CAPABLE handshake, the receiver MUST close the subflow with a RST as it not behaving as negotiated. If a

checksum is not present when its use has been negotiated, the receiver MUST close the subflow with a RST as it is considered broken. In both cases, this RST SHOULD be accompanied with a MP\_TCPRST option (Section 3.6) with the reason code for a "MPTCP specific error".

### 3.3.1. Data Sequence Mapping

The data stream as a whole can be reassembled through the use of the data sequence mapping components of the DSS option (Figure 9), which define the mapping from the subflow sequence number to the data sequence number. This is used by the receiver to ensure in-order delivery to the application layer. Meanwhile, the subflow-level sequence numbers (i.e., the regular sequence numbers in the TCP header) have subflow-only relevance. It is expected (but not mandated) that SACK [RFC2018] is used at the subflow level to improve efficiency.

The data sequence mapping specifies a mapping from subflow sequence space to data sequence space. This is expressed in terms of starting sequence numbers for the subflow and the data level, and a length of bytes for which this mapping is valid. This explicit mapping for a range of data was chosen rather than per-packet signaling to assist with compatibility with situations where TCP/IP segmentation or coalescing is undertaken separately from the stack that is generating the data flow (e.g., through the use of TCP segmentation offloading on network interface cards, or by middleboxes such as performance enhancing proxies). It also allows a single mapping to cover many packets, which may be useful in bulk transfer situations.

A mapping is fixed, in that the subflow sequence number is bound to the data sequence number after the mapping has been processed. A sender MUST NOT change this mapping after it has been declared; however, the same data sequence number can be mapped to by different subflows for retransmission purposes (see Section 3.3.6). This would also permit the same data to be sent simultaneously on multiple subflows for resilience or efficiency purposes, especially in the case of lossy links. Although the detailed specification of such operation is outside the scope of this document, an implementation SHOULD treat the first data that is received at a subflow for the data sequence space as that which should be delivered to the application, and any later data for that sequence space SHOULD be ignored.

The data sequence number is specified as an absolute value, whereas the subflow sequence numbering is relative (the SYN at the start of the subflow has relative subflow sequence number 0). This is to allow middleboxes to change the initial sequence number of a subflow,

such as firewalls that undertake Initial Sequence Number (ISN) randomization.

The data sequence mapping also contains a checksum of the data that this mapping covers, if use of checksums has been negotiated at the MP\_CAPABLE exchange. Checksums are used to detect if the payload has been adjusted in any way by a non-MPTCP-aware middlebox. If this checksum fails, it will trigger a failure of the subflow, or a fallback to regular TCP, as documented in Section 3.7, since MPTCP can no longer reliably know the subflow sequence space at the receiver to build data sequence mappings. Without checksumming enabled, corrupt data may be delivered to the application if a middlebox alters segment boundaries, alters content, or does not deliver all segments covered by a data sequence mapping. It is therefore RECOMMENDED to use checksumming unless it is known the network path contains no such devices.

The checksum algorithm used is the standard TCP checksum [RFC0793], operating over the data covered by this mapping, along with a pseudo-header as shown in Figure 10.

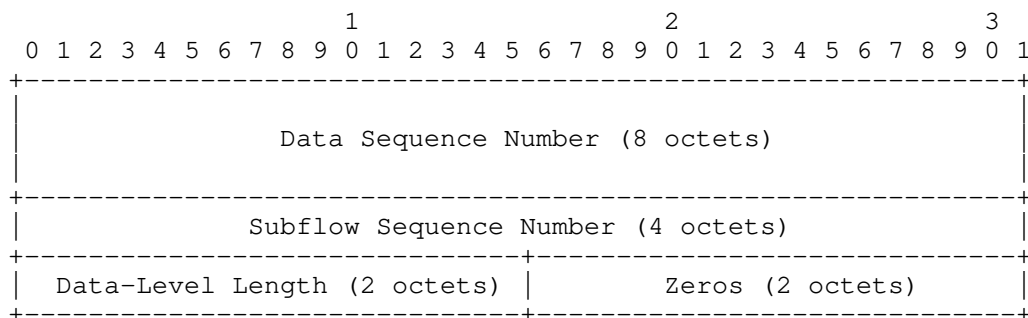


Figure 10: Pseudo-Header for DSS Checksum

Note that the data sequence number used in the pseudo-header is always the 64-bit value, irrespective of what length is used in the DSS option itself. The standard TCP checksum algorithm has been chosen since it will be calculated anyway for the TCP subflow, and if calculated first over the data before adding the pseudo-headers, it only needs to be calculated once. Furthermore, since the TCP checksum is additive, the checksum for a DSN\_MAP can be constructed by simply adding together the checksums for the data of each constituent TCP segment, and adding the checksum for the DSS pseudo-header.

Note that checksumming relies on the TCP subflow containing contiguous data; therefore, a TCP subflow MUST NOT use the Urgent

Pointer to interrupt an existing mapping. Further note, however, that if Urgent data is received on a subflow, it SHOULD be mapped to the data sequence space and delivered to the application analogous to Urgent data in regular TCP.

To avoid possible deadlock scenarios, subflow-level processing should be undertaken separately from that at connection level. Therefore, even if a mapping does not exist from the subflow space to the data-level space, the data SHOULD still be ACKed at the subflow (if it is in-window). This data cannot, however, be acknowledged at the data level (Section 3.3.2) because its data sequence numbers are unknown. Implementations MAY hold onto such unmapped data for a short while in the expectation that a mapping will arrive shortly. Such unmapped data cannot be counted as being within the connection level receive window because this is relative to the data sequence numbers, so if the receiver runs out of memory to hold this data, it will have to be discarded. If a mapping for that subflow-level sequence space does not arrive within a receive window of data, that subflow SHOULD be treated as broken, closed with a RST, and any unmapped data silently discarded.

Data sequence numbers are always 64-bit quantities, and MUST be maintained as such in implementations. If a connection is progressing at a slow rate, so protection against wrapped sequence numbers is not required, then an implementation MAY include just the lower 32 bits of the data sequence number in the data sequence mapping and/or Data ACK as an optimization, and an implementation can make this choice independently for each packet. An implementation MUST be able to receive and process both 64-bit or 32-bit sequence number values, but it is not required that an implementation is able to send both.

An implementation MUST send the full 64-bit data sequence number if it is transmitting at a sufficiently high rate that the 32-bit value could wrap within the Maximum Segment Lifetime (MSL) [RFC7323]. The lengths of the DSNs used in these values (which may be different) are declared with flags in the DSS option. Implementations MUST accept a 32-bit DSN and implicitly promote it to a 64-bit quantity by incrementing the upper 32 bits of sequence number each time the lower 32 bits wrap. A sanity check MUST be implemented to ensure that a wrap occurs at an expected time (e.g., the sequence number jumps from a very high number to a very low number) and is not triggered by out-of-order packets.

As with the standard TCP sequence number, the data sequence number should not start at zero, but at a random value to make blind session hijacking harder. This specification requires setting the initial data sequence number (IDSN) of each host to the least significant 64

bits of the SHA-256 hash of the host's key, as described in Section 3.1. This is required also in order for the receiver to know what the expected IDSN is, and thus determine if any initial connection-level packets are missing; this is particularly relevant if two subflows start transmitting simultaneously.

A data sequence mapping does not need to be included in every MPTCP packet, as long as the subflow sequence space in that packet is covered by a mapping known at the receiver. This can be used to reduce overhead in cases where the mapping is known in advance; one such case is when there is a single subflow between the hosts, another is when segments of data are scheduled in larger than packet-sized chunks.

An "infinite" mapping can be used to fall back to regular TCP by mapping the subflow-level data to the connection-level data for the remainder of the connection (see Section 3.7). This is achieved by setting the Data-Level Length field of the DSS option to the reserved value of 0. The checksum, in such a case, will also be set to zero.

### 3.3.2. Data Acknowledgments

To provide full end-to-end resilience, MPTCP provides a connection-level acknowledgment, to act as a cumulative ACK for the connection as a whole. This is the "Data ACK" field of the DSS option (Figure 9). The Data ACK is analogous to the behavior of the standard TCP cumulative ACK -- indicating how much data has been successfully received (with no holes). This is in comparison to the subflow-level ACK, which acts analogous to TCP SACK, given that there may still be holes in the data stream at the connection level. The Data ACK specifies the next data sequence number it expects to receive.

The Data ACK, as for the DSN, can be sent as the full 64-bit value, or as the lower 32 bits. If data is received with a 64-bit DSN, it MUST be acknowledged with a 64-bit Data ACK. If the DSN received is 32 bits, an implementation can choose whether to send a 32-bit or 64-bit Data ACK, and an implementation MUST accept either in this situation.

The Data ACK proves that the data, and all required MPTCP signaling, has been received and accepted by the remote end. One key use of the Data ACK signal is that it is used to indicate the left edge of the advertised receive window. As explained in Section 3.3.4, the receive window is shared by all subflows and is relative to the Data ACK. Because of this, an implementation MUST NOT use the RCV.WND field of a TCP segment at the connection level if it does not also carry a DSS option with a Data ACK field. Furthermore, separating



the connection-level acknowledgments from the subflow level allows processing to be done separately, and a receiver has the freedom to drop segments after acknowledgment at the subflow level, for example, due to memory constraints when many segments arrive out of order.

An MPTCP sender **MUST NOT** free data from the send buffer until it has been acknowledged by both a Data ACK received on any subflow and at the subflow level by all subflows on which the data was sent. The former condition ensures liveness of the connection and the latter condition ensures liveness and self-consistence of a subflow when data needs to be retransmitted. Note, however, that if some data needs to be retransmitted multiple times over a subflow, there is a risk of blocking the sending window. In this case, the MPTCP sender can decide to terminate the subflow that is behaving badly by sending a RST, using an appropriate MP\_TCPRST (Section 3.6) error code.

The Data ACK **MAY** be included in all segments; however, optimizations **SHOULD** be considered in more advanced implementations, where the Data ACK is present in segments only when the Data ACK value advances, and this behavior **MUST** be treated as valid. This behavior ensures the sender buffer is freed, while reducing overhead when the data transfer is unidirectional.

### 3.3.3. Closing a Connection

In regular TCP, a FIN announces the receiver that the sender has no more data to send. In order to allow subflows to operate independently and to keep the appearance of TCP over the wire, a FIN in MPTCP only affects the subflow on which it is sent. This allows nodes to exercise considerable freedom over which paths are in use at any one time. The semantics of a FIN remain as for regular TCP; i.e., it is not until both sides have ACKed each other's FINs that the subflow is fully closed.

When an application calls close() on a socket, this indicates that it has no more data to send; for regular TCP, this would result in a FIN on the connection. For MPTCP, an equivalent mechanism is needed, and this is referred to as the DATA\_FIN.

A DATA\_FIN is an indication that the sender has no more data to send, and as such can be used to verify that all data has been successfully received. A DATA\_FIN, as with the FIN on a regular TCP connection, is a unidirectional signal.

The DATA\_FIN is signaled by setting the 'F' flag in the Data Sequence Signal option (Figure 9) to 1. A DATA\_FIN occupies 1 octet (the final octet) of the connection-level sequence space. Note that the DATA\_FIN is included in the Data-Level Length, but not at the subflow

level: for example, a segment with DSN 80, and Data-Level Length 11, with DATA\_FIN set, would map 10 octets from the subflow into data sequence space 80-89, the DATA\_FIN is DSN 90; therefore, this segment including DATA\_FIN would be acknowledged with a DATA\_ACK of 91.

Note that when the DATA\_FIN is not attached to a TCP segment containing data, the Data Sequence Signal MUST have a subflow sequence number of 0, a Data-Level Length of 1, and the data sequence number that corresponds with the DATA\_FIN itself. The checksum in this case will only cover the pseudo-header.

A DATA\_FIN has the semantics and behavior as a regular TCP FIN, but at the connection level. Notably, it is only DATA\_ACKed once all data has been successfully received at the connection level. Note, therefore, that a DATA\_FIN is decoupled from a subflow FIN. It is only permissible to combine these signals on one subflow if there is no data outstanding on other subflows. Otherwise, it may be necessary to retransmit data on different subflows. Essentially, a host MUST NOT close all functioning subflows unless it is safe to do so, i.e., until all outstanding data has been DATA\_ACKed, or until the segment with the DATA\_FIN flag set is the only outstanding segment.

Once a DATA\_FIN has been acknowledged, all remaining subflows MUST be closed with standard FIN exchanges. Both hosts SHOULD send FINs on all subflows, as a courtesy to allow middleboxes to clean up state even if an individual subflow has failed. It is also encouraged to reduce the timeouts (Maximum Segment Lifetime) on subflows at end hosts after receiving a DATA\_FIN. In particular, any subflows where there is still outstanding data queued (which has been retransmitted on other subflows in order to get the DATA\_FIN acknowledged) MAY be closed with a RST with MP\_TCP\_RST (Section 3.6) error code for "too much outstanding data".

A connection is considered closed once both hosts' DATA\_FINS have been acknowledged by DATA\_ACKs.

As specified above, a standard TCP FIN on an individual subflow only shuts down the subflow on which it was sent. If all subflows have been closed with a FIN exchange, but no DATA\_FIN has been received and acknowledged, the MPTCP connection is treated as closed only after a timeout. This implies that an implementation will have TIME\_WAIT states at both the subflow and connection levels (see Appendix D). This permits "break-before-make" scenarios where connectivity is lost on all subflows before a new one can be re-established.

### 3.3.4. Receiver Considerations

Regular TCP advertises a receive window in each packet, telling the sender how much data the receiver is willing to accept past the cumulative ack. The receive window is used to implement flow control, throttling down fast senders when receivers cannot keep up.

MPTCP also uses a unique receive window, shared between the subflows. The idea is to allow any subflow to send data as long as the receiver is willing to accept it. The alternative, maintaining per subflow receive windows, could end up stalling some subflows while others would not use up their window.

The receive window is relative to the `DATA_ACK`. As in TCP, a receiver **MUST NOT** shrink the right edge of the receive window (i.e., `DATA_ACK + receive window`). The receiver will use the data sequence number to tell if a packet should be accepted at the connection level.

When deciding to accept packets at subflow level, regular TCP checks the sequence number in the packet against the allowed receive window. With multipath, such a check is done using only the connection-level window. A sanity check **SHOULD** be performed at subflow level to ensure that the subflow and mapped sequence numbers meet the following test:  $SSN - SUBFLOW\_ACK \leq DSN - DATA\_ACK$ , where `SSN` is the subflow sequence number of the received packet and `SUBFLOW_ACK` is the `RCV.NXT` (next expected sequence number) of the subflow (with the equivalent connection-level definitions for `DSN` and `DATA_ACK`).

In regular TCP, once a segment is deemed in-window, it is put either in the in-order receive queue or in the out-of-order queue. In Multipath TCP, the same happens but at the connection level: a segment is placed in the connection level in-order or out-of-order queue if it is in-window at both connection and subflow levels. The stack still has to remember, for each subflow, which segments were received successfully so that it can ACK them at subflow level appropriately. Typically, this will be implemented by keeping per subflow out-of-order queues (containing only message headers, not the payloads) and remembering the value of the cumulative ACK.

It is important for implementers to understand how large a receiver buffer is appropriate. The lower bound for full network utilization is the maximum bandwidth-delay product of any one of the paths. However, this might be insufficient when a packet is lost on a slower subflow and needs to be retransmitted (see Section 3.3.6). A tight upper bound would be the maximum round-trip time (RTT) of any path multiplied by the total bandwidth available across all paths. This permits all subflows to continue at full speed while a packet is

fast-retransmitted on the maximum RTT path. Even this might be insufficient to maintain full performance in the event of a retransmit timeout on the maximum RTT path. It is for future study to determine the relationship between retransmission strategies and receive buffer sizing.

### 3.3.5. Sender Considerations

The sender remembers receiver window advertisements from the receiver. It should only update its local receive window values when the largest sequence number allowed (i.e., `DATA_ACK` + receive window) increases, on the receipt of a `DATA_ACK`. This is important to allow using paths with different RTTs, and thus different feedback loops.

MPTCP uses a single receive window across all subflows, and if the receive window was guaranteed to be unchanged end-to-end, a host could always read the most recent receive window value. However, some classes of middleboxes may alter the TCP-level receive window. Typically, these will shrink the offered window, although for short periods of time it may be possible for the window to be larger (however, note that this would not continue for long periods since ultimately the middlebox must keep up with delivering data to the receiver). Therefore, if receive window sizes differ on multiple subflows, when sending data MPTCP SHOULD take the largest of the most recent window sizes as the one to use in calculations. This rule is implicit in the requirement not to reduce the right edge of the window.

The sender MUST also remember the receive windows advertised by each subflow. The allowed window for subflow *i* is (`ack_i`, `ack_i` + `rcv_wnd_i`), where `ack_i` is the subflow-level cumulative ACK of subflow *i*. This ensures data will not be sent to a middlebox unless there is enough buffering for the data.

Putting the two rules together, we get the following: a sender is allowed to send data segments with data-level sequence numbers between (`DATA_ACK`, `DATA_ACK` + `receive_window`). Each of these segments will be mapped onto subflows, as long as subflow sequence numbers are in the allowed windows for those subflows. Note that subflow sequence numbers do not generally affect flow control if the same receive window is advertised across all subflows. They will perform flow control for those subflows with a smaller advertised receive window.

The send buffer MUST, at a minimum, be as big as the receive buffer, to enable the sender to reach maximum throughput.

### 3.3.6. Reliability and Retransmissions

The data sequence mapping allows senders to resend data with the same data sequence number on a different subflow. When doing this, a host **MUST** still retransmit the original data on the original subflow, in order to preserve the subflow integrity (middleboxes could replay old data, and/or could reject holes in subflows), and a receiver will ignore these retransmissions. While this is clearly suboptimal, for compatibility reasons this is sensible behavior. Optimizations could be negotiated in future versions of this protocol. Note also that this property would also permit a sender to always send the same data, with the same data sequence number, on multiple subflows, if desired for reliability reasons.

This protocol specification does not mandate any mechanisms for handling retransmissions, and much will be dependent upon local policy (as discussed in Section 3.3.8). One can imagine aggressive connection-level retransmissions policies where every packet lost at subflow level is retransmitted on a different subflow (hence, wasting bandwidth but possibly reducing application-to-application delays), or conservative retransmission policies where connection-level retransmits are only used after a few subflow-level retransmission timeouts occur.

It is envisaged that a standard connection-level retransmission mechanism would be implemented around a connection-level data queue: all segments that haven't been `DATA_ACKed` are stored. A timer is set when the head of the connection-level is `ACKed` at subflow level but its corresponding data is not `ACKed` at data level. This timer will guard against failures in retransmission by middleboxes that proactively `ACK` data.

The sender **MUST** keep data in its send buffer as long as the data has not been acknowledged at both connection level and on all subflows on which it has been sent. In this way, the sender can always retransmit the data if needed, on the same subflow or on a different one. A special case is when a subflow fails: the sender will typically resend the data on other working subflows after a timeout, and will keep trying to retransmit the data on the failed subflow too. The sender will declare the subflow failed after a predefined upper bound on retransmissions is reached (which **MAY** be lower than the usual TCP limits of the Maximum Segment Life), or on the receipt of an ICMP error, and only then delete the outstanding data segments.

If multiple retransmissions are triggered that indicate that a subflow performs badly, this **MAY** lead to a host resetting the subflow with a RST. However, additional research is required to understand the heuristics of how and when to reset underperforming subflows.

For example, a highly asymmetric path may be misdiagnosed as underperforming. A RST for this purpose SHOULD be accompanied with an "Unacceptable performance" MP\_TCPRST option (Section 3.6).

### 3.3.7. Congestion Control Considerations

Different subflows in an MPTCP connection have different congestion windows. To achieve fairness at bottlenecks and resource pooling, it is necessary to couple the congestion windows in use on each subflow, in order to push most traffic to uncongested links. One algorithm for achieving this is presented in [RFC6356]; the algorithm does not achieve perfect resource pooling but is "safe" in that it is readily deployable in the current Internet. By this, we mean that it does not take up more capacity on any one path than if it was a single path flow using only that route, so this ensures fair coexistence with single-path TCP at shared bottlenecks.

It is foreseeable that different congestion controllers will be implemented for MPTCP, each aiming to achieve different properties in the resource pooling/fairness/stability design space, as well as those for achieving different properties in quality of service, reliability, and resilience.

Regardless of the algorithm used, the design of the MPTCP protocol aims to provide the congestion control implementations sufficient information to take the right decisions; this information includes, for each subflow, which packets were lost and when.

### 3.3.8. Subflow Policy

Within a local MPTCP implementation, a host may use any local policy it wishes to decide how to share the traffic to be sent over the available paths.

In the typical use case, where the goal is to maximize throughput, all available paths will be used simultaneously for data transfer, using coupled congestion control as described in [RFC6356]. It is expected, however, that other use cases will appear.

For instance, a possibility is an 'all-or-nothing' approach, i.e., have a second path ready for use in the event of failure of the first path, but alternatives could include entirely saturating one path before using an additional path (the 'overflow' case). Such choices would be most likely based on the monetary cost of links, but may also be based on properties such as the delay or jitter of links, where stability (of delay or bandwidth) is more important than throughput. Application requirements such as these are discussed in detail in [RFC6897].

The ability to make effective choices at the sender requires full knowledge of the path "cost", which is unlikely to be the case. It would be desirable for a receiver to be able to signal their own preferences for paths, since they will often be the multihomed party, and may have to pay for metered incoming bandwidth.

To enable this, the MP\_JOIN option (see Section 3.2) contains the 'B' bit, which allows a host to indicate to its peer that this path should be treated as a backup path to use only in the event of failure of other working subflows (i.e., a subflow where the receiver has indicated B=1 SHOULD NOT be used to send data unless there are no usable subflows where B=0).

In the event that the available set of paths changes, a host may wish to signal a change in priority of subflows to the peer (e.g., a subflow that was previously set as backup should now take priority over all remaining subflows). Therefore, the MP\_PRIO option, shown in Figure 11, can be used to change the 'B' flag of the subflow on which it is sent.

Another use of the MP\_PRIO option is to set the 'B' flag on a subflow to cleanly retire its use before closing it and removing it with REMOVE\_ADDR Section 3.4.2, for example to support make-before-break session continuity, where new subflows are added before the previously used ones are closed.

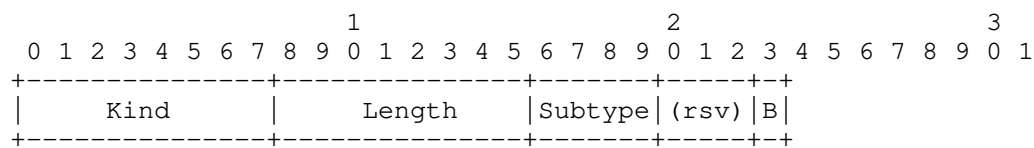


Figure 11: Change Subflow Priority (MP\_PRIO) Option

It should be noted that the backup flag is a request from a data receiver to a data sender only, and the data sender SHOULD adhere to these requests. A host cannot assume that the data sender will do so, however, since local policies -- or technical difficulties -- may override MP\_PRIO requests. Note also that this signal applies to a single direction, and so the sender of this option could choose to continue using the subflow to send data even if it has signaled B=1 to the other host.

### 3.4. Address Knowledge Exchange (Path Management)

We use the term "path management" to refer to the exchange of information about additional paths between hosts, which in this design is managed by multiple addresses at hosts. For more detail of

the architectural thinking behind this design, see the MPTCP Architecture document [RFC6182].

This design makes use of two methods of sharing such information, and both can be used on a connection. The first is the direct setup of new subflows, already described in Section 3.2, where the initiator has an additional address. The second method, described in the following subsections, signals addresses explicitly to the other host to allow it to initiate new subflows. The two mechanisms are complementary: the first is implicit and simple, while the explicit is more complex but is more robust. Together, the mechanisms allow addresses to change in flight (and thus support operation through NATs, since the source address need not be known), and also allow the signaling of previously unknown addresses, and of addresses belonging to other address families (e.g., both IPv4 and IPv6).

Here is an example of typical operation of the protocol:

- o An MPTCP connection is initially set up between address/port A1 of Host A and address/port B1 of Host B. If Host A is multihomed and multiaddressed, it can start an additional subflow from its address A2 to B1, by sending a SYN with a Join option from A2 to B1, using B's previously declared token for this connection. Alternatively, if B is multihomed, it can try to set up a new subflow from B2 to A1, using A's previously declared token. In either case, the SYN will be sent to the port already in use for the original subflow on the receiving host.
- o Simultaneously (or after a timeout), an ADD\_ADDR option (Section 3.4.1) is sent on an existing subflow, informing the receiver of the sender's alternative address(es). The recipient can use this information to open a new subflow to the sender's additional address. In our example, A will send ADD\_ADDR option informing B of address/port A2. The mix of using the SYN-based option and the ADD\_ADDR option, including timeouts, is implementation specific and can be tailored to agree with local policy.
- o If subflow A2-B1 is successfully set up, Host B can use the Address ID in the Join option to correlate this with the ADD\_ADDR option that will also arrive on an existing subflow; now B knows not to open A2-B1, ignoring the ADD\_ADDR. Otherwise, if B has not received the A2-B1 MP\_JOIN SYN but received the ADD\_ADDR, it can try to initiate a new subflow from one or more of its addresses to address A2. This permits new sessions to be opened if one host is behind a NAT.



Other ways of using the two signaling mechanisms are possible; for instance, signaling addresses in other address families can only be done explicitly using the Add Address option.

#### 3.4.1. Address Advertisement

The Add Address (ADD\_ADDR) MPTCP option announces additional addresses (and optionally, ports) on which a host can be reached (Figure 12). This option can be used at any time during a connection, depending on when the sender wishes to enable multiple paths and/or when paths become available. As with all MPTCP signals, the receiver MUST undertake standard TCP validity checks, e.g. [RFC5961], before acting upon it.

Every address has an Address ID that can be used for uniquely identifying the address within a connection for address removal. The Address ID is also used to identify MP\_JOIN options (see Section 3.2) relating to the same address, even when address translators are in use. The Address ID MUST uniquely identify the address for the sender of the option (within the scope of the connection), but the mechanism for allocating such IDs is implementation specific.

All address IDs learned via either MP\_JOIN or ADD\_ADDR SHOULD be stored by the receiver in a data structure that gathers all the Address ID to address mappings for a connection (identified by a token pair). In this way, there is a stored mapping between Address ID, observed source address, and token pair for future processing of control information for a connection. Note that an implementation MAY discard incoming address advertisements at will, for example, for avoiding updating mapping state, or because advertised addresses are of no use to it (for example, IPv6 addresses when it has IPv4 only). Therefore, a host MUST treat address advertisements as soft state, and it MAY choose to refresh advertisements periodically. Note also that an implementation MAY choose to cache these address advertisements even if they are not currently relevant but may be relevant in the future, such as IPv4 addresses when IPv6 connectivity is available but IPv4 is awaiting DHCP.

This option is shown in Figure 12. The illustration is sized for IPv4 addresses. For IPv6, the length of the address will be 16 octets (instead of 4).

The 2 octets that specify the TCP port number to use are optional and their presence can be inferred from the length of the option. Although it is expected that the majority of use cases will use the same port pairs as used for the initial subflow (e.g., port 80 remains port 80 on all subflows, as does the ephemeral port at the client), there may be cases (such as port-based load balancing) where

the explicit specification of a different port is required. If no port is specified, MPTCP SHOULD attempt to connect to the specified address on the same port as is already in use by the subflow on which the ADD\_ADDR signal was sent; this is discussed in more detail in Section 3.9.

The Truncated HMAC present in this Option is the rightmost 64 bits of an HMAC, negotiated and calculated in the same way as for MP\_JOIN as described in Section 3.2. For this specification of MPTCP, as there is only one hash algorithm option specified, this will be HMAC as defined in [RFC2104], using the SHA-256 hash algorithm [RFC6234]. In the same way as for MP\_JOIN, the key for the HMAC algorithm, in the case of the message transmitted by Host A, will be Key-A followed by Key-B, and in the case of Host B, Key-B followed by Key-A. These are the keys that were exchanged in the original MP\_CAPABLE handshake. The message for the HMAC is the Address ID, IP Address, and Port which precede the HMAC in the ADD\_ADDR option. If the port is not present in the ADD\_ADDR option, the HMAC message will nevertheless include two octets of value zero. The rationale for the HMAC is to prevent unauthorized entities from injecting ADD\_ADDR signals in an attempt to hijack a connection. Note that additionally the presence of this HMAC prevents the address being changed in flight unless the key is known by an intermediary. If a host receives an ADD\_ADDR option for which it cannot validate the HMAC, it SHOULD silently ignore the option.

A set of four flags are present after the subtype and before the Address ID. Only the rightmost bit - labelled 'E' - is assigned in this specification. The other bits are currently unassigned and MUST be set to zero by a sender and MUST be ignored by the receiver.

The 'E' flag exists to provide reliability for this option. Because this option will often be sent on pure ACKs, there is no guarantee of reliability. Therefore, a receiver receiving a fresh ADD\_ADDR option (where E=0), will send the same option back to the sender, but not including the HMAC, and with E=1, to indicate receipt. The lack of this echo can be used by the initial ADD\_ADDR sender to retransmit the ADD\_ADDR according to local policy.

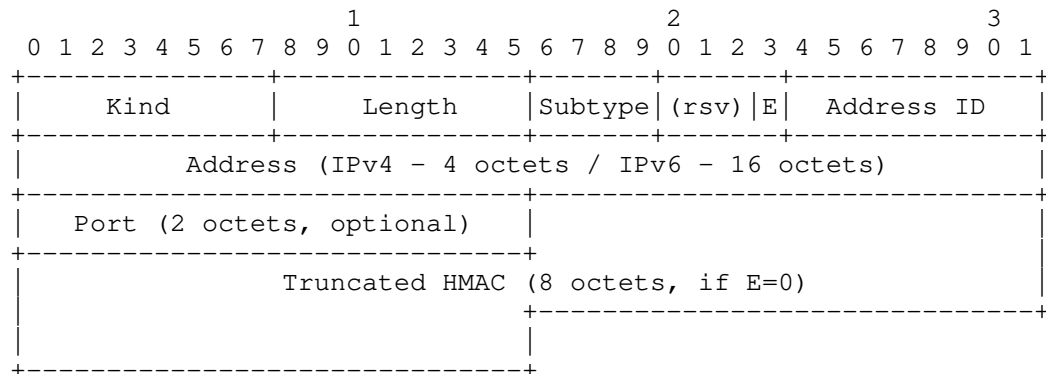


Figure 12: Add Address (ADD\_ADDR) Option

Due to the proliferation of NATs, it is reasonably likely that one host may attempt to advertise private addresses [RFC1918]. It is not desirable to prohibit this, since there may be cases where both hosts have additional interfaces on the same private network, and a host MAY advertise such addresses. The MP\_JOIN handshake to create a new subflow (Section 3.2) provides mechanisms to minimize security risks. The MP\_JOIN message contains a 32-bit token that uniquely identifies the connection to the receiving host. If the token is unknown, the host will return with a RST. In the unlikely event that the token is valid at the receiving host, subflow setup will continue, but the HMAC exchange must occur for authentication. This will fail, and will provide sufficient protection against two unconnected hosts accidentally setting up a new subflow upon the signal of a private address. Further security considerations around the issue of ADD\_ADDR messages that accidentally misdirect, or maliciously direct, new MP\_JOIN attempts are discussed in Section 5.

A host that receives an ADD\_ADDR but finds a connection set up to that IP address and port number is unsuccessful SHOULD NOT perform further connection attempts to this address/port combination for this connection. A sender that wants to trigger a new incoming connection attempt on a previously advertised address/port combination can therefore refresh ADD\_ADDR information by sending the option again.

A host can therefore send an ADD\_ADDR message with an already assigned Address ID, but the Address MUST be the same as previously assigned to this Address ID. A new ADD\_ADDR may have the same, or different, port number. If the port number is different, the receiving host SHOULD try to set up a new subflow to this new address/port combination.

A host wishing to replace an existing Address ID MUST first remove the existing one (Section 3.4.2).

During normal MPTCP operation, it is unlikely that there will be sufficient TCP option space for ADD\_ADDR to be included along with those for data sequence numbering (Section 3.3.1). Therefore, it is expected that an MPTCP implementation will send the ADD\_ADDR option on separate ACKs. As discussed earlier, however, an MPTCP implementation MUST NOT treat duplicate ACKs with any MPTCP option, with the exception of the DSS option, as indications of congestion [RFC5681], and an MPTCP implementation SHOULD NOT send more than two duplicate ACKs in a row for signaling purposes.

#### 3.4.2. Remove Address

If, during the lifetime of an MPTCP connection, a previously announced address becomes invalid (e.g., if the interface disappears, or an IPv6 address is no longer preferred), the affected host SHOULD announce this so that the peer can remove subflows related to this address. Even if an address is not in use by a MPTCP connection, if it has been previously announced, an implementation SHOULD announce its removal. A host MAY also choose to announce that a valid IP address should not be used any longer, for example for make-before-break session continuity.

This is achieved through the Remove Address (REMOVE\_ADDR) option (Figure 13), which will remove a previously added address (or list of addresses) from a connection and terminate any subflows currently using that address.

For security purposes, if a host receives a REMOVE\_ADDR option, it must ensure the affected path(s) are no longer in use before it instigates closure. The receipt of REMOVE\_ADDR SHOULD first trigger the sending of a TCP keepalive [RFC1122] on the path, and if a response is received the path SHOULD NOT be removed. If the path is found to still be alive, the receiving host SHOULD no longer use the specified address for future connections, but it is the responsibility of the host which sent the REMOVE\_ADDR to shut down the subflow. The requesting host MAY also use MP\_PRIO (Section 3.3.8) to request a path is no longer used, before removal. Typical TCP validity tests on the subflow (e.g., ensuring sequence and ACK numbers are correct) MUST also be undertaken. An implementation can use indications of these test failures as part of intrusion detection or error logging.

The sending and receipt (if no keepalive response was received) of this message SHOULD trigger the sending of RSTs by both hosts on the

affected subflow(s) (if possible), as a courtesy to cleaning up middlebox state, before cleaning up any local state.

Address removal is undertaken by ID, so as to permit the use of NATs and other middleboxes that rewrite source addresses. If there is no address at the requested ID, the receiver will silently ignore the request.

A subflow that is still functioning **MUST** be closed with a FIN exchange as in regular TCP, rather than using this option. For more information, see Section 3.3.3.

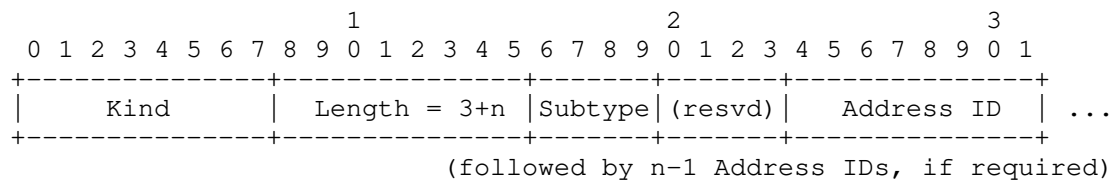


Figure 13: Remove Address (REMOVE\_ADDR) Option

### 3.5. Fast Close

Regular TCP has the means of sending a reset (RST) signal to abruptly close a connection. With MPTCP, a regular RST only has the scope of the subflow and will only close the concerned subflow but not affect the remaining subflows. MPTCP's connection will stay alive at the data level, in order to permit break-before-make handover between subflows. It is therefore necessary to provide an MPTCP-level "reset" to allow the abrupt closure of the whole MPTCP connection, and this is the MP\_FASTCLOSE option.

MP\_FASTCLOSE is used to indicate to the peer that the connection will be abruptly closed and no data will be accepted anymore. The reasons for triggering an MP\_FASTCLOSE are implementation specific. Regular TCP does not allow sending a RST while the connection is in a synchronized state [RFC0793]. Nevertheless, implementations allow the sending of a RST in this state, if, for example, the operating system is running out of resources. In these cases, MPTCP should send the MP\_FASTCLOSE. This option is illustrated in Figure 14.

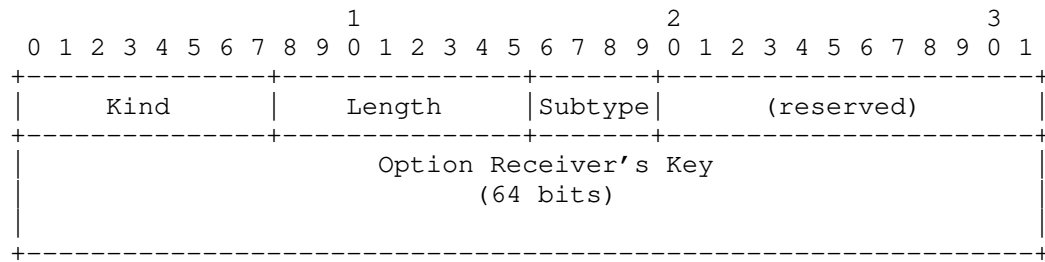


Figure 14: Fast Close (MP\_FASTCLOSE) Option

If Host A wants to force the closure of an MPTCP connection, it has two different options:

- o Option A (ACK) : Host A sends an ACK containing the MP\_FASTCLOSE option on one subflow, containing the key of Host B as declared in the initial connection handshake. On all the other subflows, Host A sends a regular TCP RST to close these subflows, and tears them down. Host A now enters FASTCLOSE\_WAIT state.
- o Option R (RST) : Host A sends a RST containing the MP\_FASTCLOSE option on all subflows, containing the key of Host B as declared in the initial connection handshake. Host A can tear the subflows and the connection down immediately.

If host A decides to force the closure by using Option A and sending an ACK with the MP\_FASTCLOSE option, the connection shall proceed as follows:

- o Upon receipt of an ACK with MP\_FASTCLOSE by Host B, containing the valid key, Host B answers on the same subflow with a TCP RST and tears down all subflows also through sending TCP RST signals. Host B can now close the whole MPTCP connection (it transitions directly to CLOSED state).
- o As soon as Host A has received the TCP RST on the remaining subflow, it can close this subflow and tear down the whole connection (transition from FASTCLOSE\_WAIT to CLOSED states). If Host A receives an MP\_FASTCLOSE instead of a TCP RST, both hosts attempted fast closure simultaneously. Host A should reply with a TCP RST and tear down the connection.
- o If Host A does not receive a TCP RST in reply to its MP\_FASTCLOSE after one retransmission timeout (RTO) (the RTO of the subflow where the MP\_FASTCLOSE has been sent), it SHOULD retransmit the MP\_FASTCLOSE. The number of retransmissions SHOULD be limited to avoid this connection from being retained for a long time, but

this limit is implementation specific. A RECOMMENDED number is 3. If no TCP RST is received in response, Host A SHOULD send a TCP RST with the MP\_FASTCLOSE option itself when it releases state in order to clear any remaining state at middleboxes.

If however host A decides to force the closure by using Option R and sending a RST with the MP\_FASTCLOSE option, Host B will act as follows: Upon receipt of a RST with MP\_FASTCLOSE, containing the valid key, Host B tears down all subflows by sending a TCP RST. Host B can now close the whole MPTCP connection (it transitions directly to CLOSED state).

### 3.6. Subflow Reset

An implementation of MPTCP may also need to send a regular TCP RST to force the closure of a subflow. A host sends a TCP RST in order to close a subflow or reject an attempt to open a subflow (MP\_JOIN). In order to inform the receiving host why a subflow is being closed or rejected, the TCP RST packet MAY include the MP\_TCPRST Option. The host MAY use this information to decide, for example, whether it tries to re-establish the subflow immediately, later, or never.

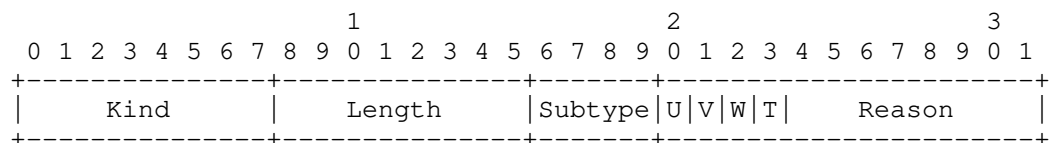


Figure 15: TCP RST Reason (MP\_TCPRST) Option

The MP\_TCPRST option contains a reason code that allows the sender of the option to provide more information about the reason for the termination of the subflow. Using 12 bits of option space, the first four bits are reserved for flags (only one of which is currently defined), and the remaining octet is used to express a reason code for this subflow termination, from which a receiver MAY infer information about the usability of this path.

The "T" flag is used by the sender to indicate whether the error condition that is reported is Transient (T bit set to 1) or Permanent (T bit set to 0). If the error condition is considered to be Transient by the sender of the RST segment, the recipient of this segment MAY try to reestablish a subflow for this connection over the failed path. The time at which a receiver may try to re-establish this is implementation-specific, but SHOULD take into account the properties of the failure defined by the following reason code. If the error condition is considered to be permanent, the receiver of the RST segment SHOULD NOT try to reestablish a subflow for this

connection over this path. The "U", "V" and "W" flags are not defined by this specification and are reserved for future use. An implementation of this specification MUST set these flags to 0, and a receiver MUST ignore them.

The "Reason" code is an 8-bit field that indicates the reason for the termination of the subflow. The following codes are defined in this document:

- o Unspecified error (code 0x0). This is the default error implying the subflow is no longer available. The presence of this option shows that the RST was generated by a MPTCP-aware device.
- o MPTCP specific error (code 0x01). An error has been detected in the processing of MPTCP options. This is the usual reason code to return in the cases where a RST is being sent to close a subflow for reasons of an invalid response.
- o Lack of resources (code 0x02). This code indicates that the sending host does not have enough resources to support the terminated subflow.
- o Administratively prohibited (code 0x03). This code indicates that the requested subflow is prohibited by the policies of the sending host.
- o Too much outstanding data (code 0x04). This code indicates that there is an excessive amount of data that need to be transmitted over the terminated subflow while having already been acknowledged over one or more other subflows. This may occur if a path has been unavailable for a short period and it is more efficient to reset and start again than it is to retransmit the queued data.
- o Unacceptable performance (code 0x05). This code indicates that the performance of this subflow was too low compared to the other subflows of this Multipath TCP connection.
- o Middlebox interference (code 0x06). Middlebox interference has been detected over this subflow making MPTCP signaling invalid. For example, this may be sent if the checksum does not validate.

### 3.7. Fallback

Sometimes, middleboxes will exist on a path that could prevent the operation of MPTCP. MPTCP has been designed in order to cope with many middlebox modifications (see Section 6), but there are still some cases where a subflow could fail to operate within the MPTCP requirements. These cases are notably the following: the loss of



MPTCP options on a path, and the modification of payload data. If such an event occurs, it is necessary to "fall back" to the previous, safe operation. This may be either falling back to regular TCP or removing a problematic subflow.

At the start of an MPTCP connection (i.e., the first subflow), it is important to ensure that the path is fully MPTCP capable and the necessary MPTCP options can reach each host. The handshake as described in Section 3.1 SHOULD fall back to regular TCP if either of the SYN messages do not have the MPTCP options: this is the same, and desired, behavior in the case where a host is not MPTCP capable, or the path does not support the MPTCP options. When attempting to join an existing MPTCP connection (Section 3.2), if a path is not MPTCP capable and the MPTCP options do not get through on the SYNs, the subflow will be closed according to the MP\_JOIN logic.

There is, however, another corner case that should be addressed. That is one of MPTCP options getting through on the SYN, but not on regular packets. This can be resolved if the subflow is the first subflow, and thus all data in flight is contiguous, using the following rules.

A sender MUST include a DSS option with data sequence mapping in every segment until one of the sent segments has been acknowledged with a DSS option containing a Data ACK. Upon reception of the acknowledgment, the sender has the confirmation that the DSS option passes in both directions and may choose to send fewer DSS options than once per segment.

If, however, an ACK is received for data (not just for the SYN) without a DSS option containing a Data ACK, the sender determines the path is not MPTCP capable. In the case of this occurring on an additional subflow (i.e., one started with MP\_JOIN), the host MUST close the subflow with a RST, which SHOULD contain a MP\_TCP\_RST option (Section 3.6) with a "Middlebox interference" reason code.

In the case of such an ACK being received on the first subflow (i.e., that started with MP\_CAPABLE), before any additional subflows are added, the implementation MUST drop out of an MPTCP mode, back to regular TCP. The sender will send one final data sequence mapping, with the Data-Level Length value of 0 indicating an infinite mapping (to inform the other end in case the path drops options in one direction only), and then revert to sending data on the single subflow without any MPTCP options.

If a subflow breaks during operation, e.g. if it is re-routed and MPTCP options are no longer permitted, then once this is detected (by the subflow-level receive buffer filling up, since there is no

mapping available in order to DATA\_ACK this data), the subflow SHOULD be treated as broken and closed with a RST, since no data can be delivered to the application layer, and no fallback signal can be reliably sent. This RST SHOULD include the MP\_TCPRST option (Section 3.6) with a "Middlebox interference" reason code.

These rules should cover all cases where such a failure could happen: whether it's on the forward or reverse path and whether the server or the client first sends data.

So far this section has discussed the loss of MPTCP options, either initially, or during the course of the connection. As described in Section 3.3, each portion of data for which there is a mapping is protected by a checksum, if checksums have been negotiated. This mechanism is used to detect if middleboxes have made any adjustments to the payload (added, removed, or changed data). A checksum will fail if the data has been changed in any way. This will also detect if the length of data on the subflow is increased or decreased, and this means the data sequence mapping is no longer valid. The sender no longer knows what subflow-level sequence number the receiver is genuinely operating at (the middlebox will be faking ACKs in return), and it cannot signal any further mappings. Furthermore, in addition to the possibility of payload modifications that are valid at the application layer, there is the possibility that such modifications could be triggered across MPTCP segment boundaries, corrupting the data. Therefore, all data from the start of the segment that failed the checksum onwards is not trustworthy.

Note that if checksum usage has not been negotiated, this fallback mechanism cannot be used unless there is some higher or lower layer signal to inform the MPTCP implementation that the payload has been tampered with.

When multiple subflows are in use, the data in flight on a subflow will likely involve data that is not contiguously part of the connection-level stream, since segments will be spread across the multiple subflows. Due to the problems identified above, it is not possible to determine what adjustment has done to the data (notably, any changes to the subflow sequence numbering). Therefore, it is not possible to recover the subflow, and the affected subflow must be immediately closed with a RST, featuring an MP\_FAIL option (Figure 16), which defines the data sequence number at the start of the segment (defined by the data sequence mapping) that had the checksum failure. Note that the MP\_FAIL option requires the use of the full 64-bit sequence number, even if 32-bit sequence numbers are normally in use in the DSS signals on the path.

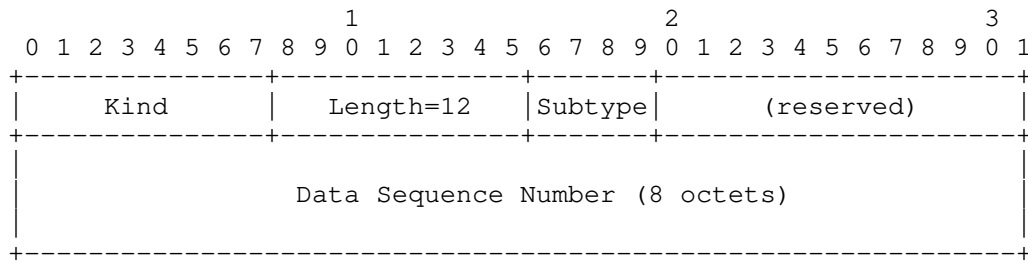


Figure 16: Fallback (MP\_FAIL) Option

The receiver of this option MUST discard all data following the data sequence number specified. Failed data MUST NOT be DATA\_ACKed and so will be retransmitted on other subflows (Section 3.3.6).

A special case is when there is a single subflow and it fails with a checksum error. If it is known that all unacknowledged data in flight is contiguous (which will usually be the case with a single subflow), an infinite mapping can be applied to the subflow without the need to close it first, and essentially turn off all further MPTCP signaling. In this case, if a receiver identifies a checksum failure when there is only one path, it will send back an MP\_FAIL option on the subflow-level ACK, referring to the data-level sequence number of the start of the segment on which the checksum error was detected. The sender will receive this, and if all unacknowledged data in flight is contiguous, will signal an infinite mapping. This infinite mapping will be a DSS option (Section 3.3) on the first new packet, containing a data sequence mapping that acts retroactively, referring to the start of the subflow sequence number of the most recent segment that was known to be delivered intact (i.e. was successfully DATA\_ACKed). From that point onwards, data can be altered by a middlebox without affecting MPTCP, as the data stream is equivalent to a regular, legacy TCP session. Whilst in theory paths may only be damaged in one direction, and the MP\_FAIL signal affects only one direction of traffic, for implementation simplicity, the receiver of an MP\_FAIL MUST also respond with an MP\_FAIL in the reverse direction and entirely revert to a regular TCP session.

In the rare case that the data is not contiguous (which could happen when there is only one subflow but it is retransmitting data from a subflow that has recently been uncleanly closed), the receiver MUST close the subflow with a RST with MP\_FAIL. The receiver MUST discard all data that follows the data sequence number specified. The sender MAY attempt to create a new subflow belonging to the same connection, and, if it chooses to do so, SHOULD place the single subflow immediately in single-path mode by setting an infinite data sequence

mapping. This mapping will begin from the data-level sequence number that was declared in the MP\_FAIL.

After a sender signals an infinite mapping, it MUST only use subflow ACKs to clear its send buffer. This is because Data ACKs may become misaligned with the subflow ACKs when middleboxes insert or delete data. The receiver SHOULD stop generating Data ACKs after it receives an infinite mapping.

When a connection has fallen back with an infinite mapping, only one subflow can send data; otherwise, the receiver would not know how to reorder the data. In practice, this means that all MPTCP subflows will have to be terminated except one. Once MPTCP falls back to regular TCP, it MUST NOT revert to MPTCP later in the connection.

It should be emphasized that MPTCP is not attempting to prevent the use of middleboxes that want to adjust the payload. An MPTCP-aware middlebox could provide such functionality by also rewriting checksums.

### 3.8. Error Handling

In addition to the fallback mechanism as described above, the standard classes of TCP errors may need to be handled in an MPTCP-specific way. Note that changing semantics -- such as the relevance of a RST -- are covered in Section 4. Where possible, we do not want to deviate from regular TCP behavior.

The following list covers possible errors and the appropriate MPTCP behavior:

- o Unknown token in MP\_JOIN (or HMAC failure in MP\_JOIN ACK, or missing MP\_JOIN in SYN/ACK response): send RST (analogous to TCP's behavior on an unknown port)
- o DSN out of window (during normal operation): drop the data, do not send Data ACKs
- o Remove request for unknown address ID: silently ignore

### 3.9. Heuristics

There are a number of heuristics that are needed for performance or deployment but that are not required for protocol correctness. In this section, we detail such heuristics. Note that discussion of buffering and certain sender and receiver window behaviors are presented in Sections 3.3.4 and 3.3.5, as well as retransmission in Section 3.3.6.

### 3.9.1. Port Usage

Under typical operation, an MPTCP implementation SHOULD use the same ports as already in use. In other words, the destination port of a SYN containing an MP\_JOIN option SHOULD be the same as the remote port of the first subflow in the connection. The local port for such SYNs SHOULD also be the same as for the first subflow (and as such, an implementation SHOULD reserve ephemeral ports across all local IP addresses), although there may be cases where this is infeasible. This strategy is intended to maximize the probability of the SYN being permitted by a firewall or NAT at the recipient and to avoid confusing any network monitoring software.

There may also be cases, however, where a host wishes to signal that a specific port should be used, and this facility is provided in the ADD\_ADDR option as documented in Section 3.4.1. It is therefore feasible to allow multiple subflows between the same two addresses but using different port pairs, and such a facility could be used to allow load balancing within the network based on 5-tuples (e.g., some ECMP implementations [RFC2992]).

### 3.9.2. Delayed Subflow Start and Subflow Symmetry

Many TCP connections are short-lived and consist only of a few segments, and so the overheads of using MPTCP outweigh any benefits. A heuristic is required, therefore, to decide when to start using additional subflows in an MPTCP connection. Experimental deployments have shown that MPTCP can be applied in a range of scenarios so an implementation is likely to need to take into account factors including the type of traffic being sent and duration of session, and this information MAY be signalled by the application layer.

However, for standard TCP traffic, a suggested general-purpose heuristic that an implementation MAY choose to employ is as follows.

If a host has data buffered for its peer (which implies that the application has received a request for data), the host opens one subflow for each initial window's worth of data that is buffered.

Consideration should also be given to limiting the rate of adding new subflows, as well as limiting the total number of subflows open for a particular connection. A host may choose to vary these values based on its load or knowledge of traffic and path characteristics.

Note that this heuristic alone is probably insufficient. Traffic for many common applications, such as downloads, is highly asymmetric and the host that is multihomed may well be the client that will never fill its buffers, and thus never use MPTCP according to this

heuristic. Advanced APIs that allow an application to signal its traffic requirements would aid in these decisions.

An additional time-based heuristic could be applied, opening additional subflows after a given period of time has passed. This would alleviate the above issue, and also provide resilience for low-bandwidth but long-lived applications.

Another issue is that both communicating hosts may simultaneously try to set up a subflow between the same pair of addresses. This leads to an inefficient use of resources.

If the same ports are used on all subflows, as recommended above, then standard TCP simultaneous open logic should take care of this situation and only one subflow will be established between the address pairs. However, this relies on the same ports being used at both end hosts. If a host does not support TCP simultaneous open, it is RECOMMENDED that some element of randomization is applied to the time to wait before opening new subflows, so that only one subflow is created between a given address pair. If, however, hosts signal additional ports to use (for example, for leveraging ECMP on-path), this heuristic is not appropriate.

This section has shown some of the considerations that an implementer should give when developing MPTCP heuristics, but is not intended to be prescriptive.

### 3.9.3. Failure Handling

Requirements for MPTCP's handling of unexpected signals have been given in Section 3.8. There are other failure cases, however, where a hosts can choose appropriate behavior.

For example, Section 3.1 suggests that a host SHOULD fall back to trying regular TCP SYNs after one or more failures of MPTCP SYNs for a connection. A host may keep a system-wide cache of such information, so that it can back off from using MPTCP, firstly for that particular destination host, and eventually on a whole interface, if MPTCP connections continue failing. The duration of such a cache would be implementation-specific.

Another failure could occur when the MP\_JOIN handshake fails. Section 3.8 specifies that an incorrect handshake MUST lead to the subflow being closed with a RST. A host operating an active intrusion detection system may choose to start blocking MP\_JOIN packets from the source host if multiple failed MP\_JOIN attempts are seen. From the connection initiator's point of view, if an MP\_JOIN fails, it SHOULD NOT attempt to connect to the same IP address and

port during the lifetime of the connection, unless the other host refreshes the information with another ADD\_ADDR option. Note that the ADD\_ADDR option is informational only, and does not guarantee the other host will attempt a connection.

In addition, an implementation may learn, over a number of connections, that certain interfaces or destination addresses consistently fail and may default to not trying to use MPTCP for these. Behavior could also be learned for particularly badly performing subflows or subflows that regularly fail during use, in order to temporarily choose not to use these paths.

#### 4. Semantic Issues

In order to support multipath operation, the semantics of some TCP components have changed. To aid clarity, this section collects these semantic changes as a reference.

**Sequence number:** The (in-header) TCP sequence number is specific to the subflow. To allow the receiver to reorder application data, an additional data-level sequence space is used. In this data-level sequence space, the initial SYN and the final DATA\_FIN occupy 1 octet of sequence space. This is to ensure these signals are acknowledged at the connection level. There is an explicit mapping of data sequence space to subflow sequence space, which is signaled through TCP options in data packets.

**ACK:** The ACK field in the TCP header acknowledges only the subflow sequence number, not the data-level sequence space. Implementations SHOULD NOT attempt to infer a data-level acknowledgment from the subflow ACKs. This separates subflow- and connection-level processing at an end host.

**Duplicate ACK:** A duplicate ACK that includes any MPTCP signaling (with the exception of the DSS option) MUST NOT be treated as a signal of congestion. To limit the chances of non-MPTCP-aware entities mistakenly interpreting duplicate ACKs as a signal of congestion, MPTCP SHOULD NOT send more than two duplicate ACKs containing (non-DSS) MPTCP signals in a row.

**Receive Window:** The receive window in the TCP header indicates the amount of free buffer space for the whole data-level connection (as opposed to for this subflow) that is available at the receiver. This is the same semantics as regular TCP, but to maintain these semantics the receive window must be interpreted at the sender as relative to the sequence number given in the DATA\_ACK rather than the subflow ACK in the TCP header. In this way, the original flow control role is preserved. Note that some

middleboxes may change the receive window, and so a host SHOULD use the maximum value of those recently seen on the constituent subflows for the connection-level receive window, and also needs to maintain a subflow-level window for subflow-level processing.

FIN: The FIN flag in the TCP header applies only to the subflow it is sent on, not to the whole connection. For connection-level FIN semantics, the DATA\_FIN option is used.

RST: The RST flag in the TCP header applies only to the subflow it is sent on, not to the whole connection. The MP\_FASTCLOSE option provides the fast close functionality of a RST at the MPTCP connection level.

Address List: Address list management (i.e., knowledge of the local and remote hosts' lists of available IP addresses) is handled on a per-connection basis (as opposed to per subflow, per host, or per pair of communicating hosts). This permits the application of per-connection local policy. Adding an address to one connection (either explicitly through an Add Address message, or implicitly through a Join) has no implication for other connections between the same pair of hosts.

5-tuple: The 5-tuple (protocol, local address, local port, remote address, remote port) presented by kernel APIs to the application layer in a non-multipath-aware application is that of the first subflow, even if the subflow has since been closed and removed from the connection. This decision, and other related API issues, are discussed in more detail in [RFC6897].

## 5. Security Considerations

As identified in [RFC6181], the addition of multipath capability to TCP will bring with it a number of new classes of threat. In order to prevent these, [RFC6182] presents a set of requirements for a security solution for MPTCP. The fundamental goal is for the security of MPTCP to be "no worse" than regular TCP today, and the key security requirements are:

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



In order to achieve these goals, MPTCP includes a hash-based handshake algorithm documented in Sections 3.1 and 3.2.

The security of the MPTCP connection hangs on the use of keys that are shared once at the start of the first subflow, and are never sent again over the network (unless used in the fast close mechanism, Section 3.5). To ease demultiplexing while not giving away any cryptographic material, future subflows use a truncated cryptographic hash of this key as the connection identification "token". The keys are concatenated and used as keys for creating Hash-based Message Authentication Codes (HMACs) used on subflow setup, in order to verify that the parties in the handshake are the same as in the original connection setup. It also provides verification that the peer can receive traffic at this new address. Replay attacks would still be possible when only keys are used; therefore, the handshakes use single-use random numbers (nonces) at both ends -- this ensures the HMAC will never be the same on two handshakes. Guidance on generating random numbers suitable for use as keys is given in [RFC4086] and discussed in Section 3.1. The nonces are valid for the lifetime of the TCP connection attempt. HMAC is also used to secure the ADD\_ADDR option, due to the threats identified in [RFC7430].

The use of crypto capability bits in the initial connection handshake to negotiate use of a particular algorithm allows the deployment of additional crypto mechanisms in the future. This negotiation would nevertheless be susceptible to a bid-down attack by an on-path active attacker who could modify the crypto capability bits in the response from the receiver to use a less secure crypto mechanism. The security mechanism presented in this document should therefore protect against all forms of flooding and hijacking attacks discussed in [RFC6181].

The version negotiation specified in Section 3.1, if differing MPTCP versions shared a common negotiation format, would allow an on-path attacker to apply a theoretical bid-down attack. Since the v1 and v0 protocols have a different handshake, such an attack would require the client to re-establish the connection using v0, and this being supported by the server. Note that an on-path attacker would have access to the raw data, negating any other TCP-level security mechanisms. Also a change from RFC6824 has removed the subflow identifier from the MP\_PRIO option (Section 3.3.8), to remove the theoretical attack where a subflow could be placed in "backup" mode by an attacker.

During normal operation, regular TCP protection mechanisms (such as ensuring sequence numbers are in-window) will provide the same level of protection against attacks on individual TCP subflows as exists for regular TCP today. Implementations will introduce additional

buffers compared to regular TCP, to reassemble data at the connection level. The application of window sizing will minimize the risk of denial-of-service attacks consuming resources.

As discussed in Section 3.4.1, a host may advertise its private addresses, but these might point to different hosts in the receiver's network. The MP\_JOIN handshake (Section 3.2) will ensure that this does not succeed in setting up a subflow to the incorrect host. However, it could still create unwanted TCP handshake traffic. This feature of MPTCP could be a target for denial-of-service exploits, with malicious participants in MPTCP connections encouraging the recipient to target other hosts in the network. Therefore, implementations should consider heuristics (Section 3.9) at both the sender and receiver to reduce the impact of this.

To further protect against malicious ADD\_ADDR messages sent by an off-path attacker, the ADD\_ADDR includes an HMAC using the keys negotiated during the handshake. This effectively prevents an attacker from diverting an MPTCP connection through an off-path ADD\_ADDR injection into the stream.

A small security risk could theoretically exist with key reuse, but in order to accomplish a replay attack, both the sender and receiver keys, and the sender and receiver random numbers, in the MP\_JOIN handshake (Section 3.2) would have to match.

Whilst this specification defines a "medium" security solution, meeting the criteria specified at the start of this section and the threat analysis ([RFC6181]), since attacks only ever get worse, it is likely that a future version of MPTCP would need to be able to support stronger security. There are several ways the security of MPTCP could potentially be improved; some of these would be compatible with MPTCP as defined in this document, whilst others may not be. For now, the best approach is to get experience with the current approach, establish what might work, and check that the threat analysis is still accurate.

Possible ways of improving MPTCP security could include:

- o defining a new MPCTP cryptographic algorithm, as negotiated in MP\_CAPABLE. A sub-case could be to include an additional deployment assumption, such as stateful servers, in order to allow a more powerful algorithm to be used.
- o defining how to secure data transfer with MPTCP, whilst not changing the signaling part of the protocol.

- o defining security that requires more option space, perhaps in conjunction with a "long options" proposal for extending the TCP options space (such as those surveyed in [TCPL0]), or perhaps building on the current approach with a second stage of MPTCP-option-based security.
- o revisiting the working group's decision to exclusively use TCP options for MPTCP signaling, and instead look at also making use of the TCP payloads.

MPTCP has been designed with several methods available to indicate a new security mechanism, including:

- o available flags in MP\_CAPABLE (Figure 4);
- o available subtypes in the MPTCP option (Figure 3);
- o the version field in MP\_CAPABLE (Figure 4);

## 6. Interactions with Middleboxes

Multipath TCP was designed to be deployable in the present world. Its design takes into account "reasonable" existing middlebox behavior. In this section, we outline a few representative middlebox-related failure scenarios and show how Multipath TCP handles them. Next, we list the design decisions multipath has made to accommodate the different middleboxes.

A primary concern is our use of a new TCP option. Middleboxes should forward packets with unknown options unchanged, yet there are some that don't. These we expect will either strip options and pass the data, drop packets with new options, copy the same option into multiple segments (e.g., when doing segmentation), or drop options during segment coalescing.

MPTCP uses a single new TCP option "Kind", and all message types are defined by "subtype" values (see Section 8). This should reduce the chances of only some types of MPTCP options being passed, and instead the key differing characteristics are different paths, and the presence of the SYN flag.

MPTCP SYN packets on the first subflow of a connection contain the MP\_CAPABLE option (Section 3.1). If this is dropped, MPTCP SHOULD fall back to regular TCP. If packets with the MP\_JOIN option (Section 3.2) are dropped, the paths will simply not be used.

If a middlebox strips options but otherwise passes the packets unchanged, MPTCP will behave safely. If an MP\_CAPABLE option is

dropped on either the outgoing or the return path, the initiating host can fall back to regular TCP, as illustrated in Figure 17 and discussed in Section 3.1.

Subflow SYNs contain the MP\_JOIN option. If this option is stripped on the outgoing path, the SYN will appear to be a regular SYN to Host B. Depending on whether there is a listening socket on the target port, Host B will reply either with SYN/ACK or RST (subflow connection fails). When Host A receives the SYN/ACK it sends a RST because the SYN/ACK does not contain the MP\_JOIN option and its token. Either way, the subflow setup fails, but otherwise does not affect the MPTCP connection as a whole.

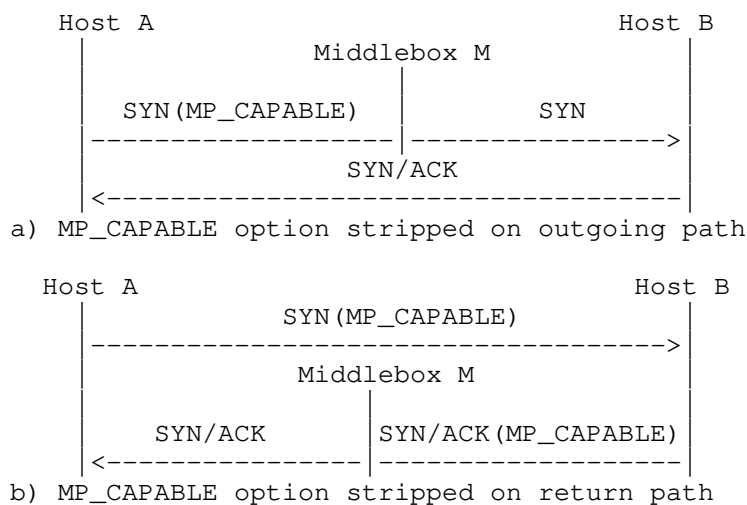


Figure 17: Connection Setup with Middleboxes that Strip Options from Packets

We now examine data flow with MPTCP, assuming the flow is correctly set up, which implies the options in the SYN packets were allowed through by the relevant middleboxes. If options are allowed through and there is no resegmentation or coalescing to TCP segments, Multipath TCP flows can proceed without problems.

The case when options get stripped on data packets has been discussed in the Fallback section. If only some MPTCP options are stripped, behavior is not deterministic. If some data sequence mappings are lost, the connection can continue so long as mappings exist for the subflow-level data (e.g., if multiple maps have been sent that reinforce each other). If some subflow-level space is left unmapped, however, the subflow is treated as broken and is closed, through the process described in Section 3.7. MPTCP should survive with a loss

of some Data ACKs, but performance will degrade as the fraction of stripped options increases. We do not expect such cases to appear in practice, though: most middleboxes will either strip all options or let them all through.

We end this section with a list of middlebox classes, their behavior, and the elements in the MPTCP design that allow operation through such middleboxes. Issues surrounding dropping packets with options or stripping options were discussed above, and are not included here:

- o NATs [RFC3022] (Network Address (and Port) Translators) change the source address (and often source port) of packets. This means that a host will not know its public-facing address for signaling in MPTCP. Therefore, MPTCP permits implicit address addition via the MP\_JOIN option, and the handshake mechanism ensures that connection attempts to private addresses [RFC1918], since they are authenticated, will only set up subflows to the correct hosts. Explicit address removal is undertaken by an Address ID to allow no knowledge of the source address.
- o Performance Enhancing Proxies (PEPs) [RFC3135] might proactively ACK data to increase performance. MPTCP, however, relies on accurate congestion control signals from the end host, and non-MPTCP-aware PEPs will not be able to provide such signals. MPTCP will, therefore, fall back to single-path TCP, or close the problematic subflow (see Section 3.7).
- o Traffic Normalizers [norm] may not allow holes in sequence numbers, and may cache packets and retransmit the same data. MPTCP looks like standard TCP on the wire, and will not retransmit different data on the same subflow sequence number. In the event of a retransmission, the same data will be retransmitted on the original TCP subflow even if it is additionally retransmitted at the connection level on a different subflow.
- o Firewalls [RFC2979] might perform initial sequence number randomization on TCP connections. MPTCP uses relative sequence numbers in data sequence mapping to cope with this. Like NATs, firewalls will not permit many incoming connections, so MPTCP supports address signaling (ADD\_ADDR) so that a multiaddressed host can invite its peer behind the firewall/NAT to connect out to its additional interface.
- o Intrusion Detection/Prevention Systems (IDS/IPS) observe packet streams for patterns and content that could threaten a network. MPTCP may require the instrumentation of additional paths, and an MPTCP-aware IDS/IPS would need to read MPTCP tokens to correlate data from multiple subflows to maintain comparable visibility into

all of the traffic between devices. Without such changes, an IDS would get an incomplete view of the traffic, increasing the risk of missing traffic of interest (false negatives), and increasing the chances of erroneously identifying a subflow as a risk due to only seeing partial data (false positives).

- o Application-level middleboxes such as content-aware firewalls may alter the payload within a subflow, such as rewriting URIs in HTTP traffic. MPTCP will detect these using the checksum and close the affected subflow(s), if there are other subflows that can be used. If all subflows are affected, multipath will fall back to TCP, allowing such middleboxes to change the payload. MPTCP-aware middleboxes should be able to adjust the payload and MPTCP metadata in order not to break the connection.

In addition, all classes of middleboxes may affect TCP traffic in the following ways:

- o TCP options may be removed, or packets with unknown options dropped, by many classes of middleboxes. It is intended that the initial SYN exchange, with a TCP option, will be sufficient to identify the path capabilities. If such a packet does not get through, MPTCP will end up falling back to regular TCP.
- o Segmentation/Coalescing (e.g., TCP segmentation offloading) might copy options between packets and might strip some options. MPTCP's data sequence mapping includes the relative subflow sequence number instead of using the sequence number in the segment. In this way, the mapping is independent of the packets that carry it.
- o The receive window may be shrunk by some middleboxes at the subflow level. MPTCP will use the maximum window at data level, but will also obey subflow-specific windows.

## 7. Acknowledgments

The authors gratefully acknowledge significant input into this document from Sebastien Barre and Andrew McDonald.

The authors also wish to acknowledge reviews and contributions from Iljitsch van Beijnum, Lars Eggert, Marcelo Bagnulo, Robert Hancock, Pasi Sarolahti, Toby Moncaster, Philip Eardley, Sergio Lembo, Lawrence Conroy, Yoshifumi Nishida, Bob Briscoe, Stein Gjessing, Andrew McGregor, Georg Hampel, Anumita Biswas, Wes Eddy, Alexey Melnikov, Francis Dupont, Adrian Farrel, Barry Leiba, Robert Sparks, Sean Turner, Stephen Farrell, Martin Stiernerling, Gregory Detal, Fabien Duchene, Xavier de Foy, Rahul Jadhav, Klemens Schragel, Mirja

Kuehlewind, Sheng Jiang, Alissa Cooper, Ines Robles, Roman Danyliw, Adam Roach, Barry Leiba, Alexey Melnikov, Eric Vyncke, and Ben Kaduk.

## 8. IANA Considerations

This document obsoletes RFC6824 and as such IANA is requested to update the TCP option space registry to point to this document for Multipath TCP, as follows:

Kind	Length	Meaning	Reference
30	N	Multipath TCP (MPTCP)	This document

Table 1: TCP Option Kind Numbers

### 8.1. MPTCP Option Subtypes

The 4-bit MPTCP subtype sub-registry ("MPTCP Option Subtypes" under the "Transmission Control Protocol (TCP) Parameters" registry) was defined in RFC6824. Since RFC6824 was an Experimental not Standards Track RFC, and since no further entries have occurred beyond those pointing to RFC6824, IANA is requested to replace the existing registry with Table 2 and with the following explanatory note.

Note: This registry specifies the MPTCP Option Subtypes for MPTCP v1, which obsoletes the Experimental MPTCP v0. For the MPTCP v0 subtypes, please refer to RFC6824.

Value	Symbol	Name	Reference
0x0	MP_CAPABLE	Multipath Capable	This document, Section 3.1
0x1	MP_JOIN	Join Connection	This document, Section 3.2
0x2	DSS	Data Sequence Signal (Data ACK and data sequence mapping)	This document, Section 3.3
0x3	ADD_ADDR	Add Address	This document, Section 3.4.1
0x4	REMOVE_ADDR	Remove Address	This document, Section 3.4.2
0x5	MP_PRIO	Change Subflow Priority	This document, Section 3.3.8
0x6	MP_FAIL	Fallback	This document, Section 3.7
0x7	MP_FASTCLOSE	Fast Close	This document, Section 3.5
0x8	MP_TCPRST	Subflow Reset	This document, Section 3.6
0xf	MP_EXPERIMENTAL	Reserved for private experiments	

Table 2: MPTCP Option Subtypes

Values 0x9 through 0xe are currently unassigned. Option 0xf is reserved for use by private experiments. Its use may be formalized in a future specification. Future assignments in this registry are to be defined by Standards Action as defined by [RFC8126]. Assignments consist of the MPTCP subtype's symbolic name and its associated value, and a reference to its specification.

## 8.2. MPTCP Handshake Algorithms

The "MPTCP Handshake Algorithms" sub-registry under the "Transmission Control Protocol (TCP) Parameters" registry was defined in RFC6824. Since RFC6824 was an Experimental not Standards Track RFC, and since



no further entries have occurred beyond those pointing to RFC6824, IANA is requested to replace the existing registry with Table 3 and with the following explanatory note.

Note: This registry specifies the MPTCP Handshake Algorithms for MPTCP v1, which obsoletes the Experimental MPTCP v0. For the MPTCP v0 subtypes, please refer to RFC6824.

Flag Bit	Meaning	Reference
A	Checksum required	This document, Section 3.1
B	Extensibility	This document, Section 3.1
C	Do not attempt to establish new subflows to the source address.	This document, Section 3.1
D-G	Unassigned	
H	HMAC-SHA256	This document, Section 3.2

Table 3: MPTCP Handshake Algorithms

Note that the meanings of bits D through H can be dependent upon bit B, depending on how Extensibility is defined in future specifications; see Section 3.1 for more information.

Future assignments in this registry are also to be defined by Standards Action as defined by [RFC8126]. Assignments consist of the value of the flags, a symbolic name for the algorithm, and a reference to its specification.

### 8.3. MP\_TCPRST Reason Codes

IANA is requested to create a further sub-registry, "MPTCP MP\_TCPRST Reason Codes" under the "Transmission Control Protocol (TCP) Parameters" registry, based on the reason code in MP\_TCPRST (Section 3.6) message. Initial values for this registry are given in Table 4; future assignments are to be defined by Specification Required as defined by [RFC8126]. Assignments consist of the value of the code, a short description of its meaning, and a reference to its specification. The maximum value is 0xff.

As guidance to the Designated Expert [RFC8126], assignments should not normally be refused unless codepoint space is becoming scarce, providing that there is a clear distinction from other, already-

existing codes, and also providing there is sufficient guidance for implementors both sending and receiving these codes.

Code	Meaning	Reference
0x00	Unspecified TCP error	This document, Section 3.6
0x01	MPTCP specific error	This document, Section 3.6
0x02	Lack of resources	This document, Section 3.6
0x03	Administratively prohibited	This document, Section 3.6
0x04	Too much outstanding data	This document, Section 3.6
0x05	Unacceptable performance	This document, Section 3.6
0x06	Middlebox interference	This document, Section 3.6

Table 4: MPTCP MP\_TCPRST Reason Codes

## 9. References

### 9.1. Normative References

- [RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, DOI 10.17487/RFC0793, September 1981, <<https://www.rfc-editor.org/info/rfc793>>.
- [RFC2104] Krawczyk, H., Bellare, M., and R. Canetti, "HMAC: Keyed-Hashing for Message Authentication", RFC 2104, DOI 10.17487/RFC2104, February 1997, <<https://www.rfc-editor.org/info/rfc2104>>.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC5961] Ramaiah, A., Stewart, R., and M. Dalal, "Improving TCP's Robustness to Blind In-Window Attacks", RFC 5961, DOI 10.17487/RFC5961, August 2010, <<https://www.rfc-editor.org/info/rfc5961>>.
- [RFC6234] Eastlake 3rd, D. and T. Hansen, "US Secure Hash Algorithms (SHA and SHA-based HMAC and HKDF)", RFC 6234, DOI 10.17487/RFC6234, May 2011, <<https://www.rfc-editor.org/info/rfc6234>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.

## 9.2. Informative References

- [deployments] Bonaventure, O. and S. Seo, "Multipath TCP Deployments", IETF Journal 2016, November 2016, <<https://www.ietfjournal.org/multipath-tcp-deployments/>>.
- [howhard] Raiciu, C., Paasch, C., Barre, S., Ford, A., Honda, M., Duchene, F., Bonaventure, O., and M. Handley, "How Hard Can It Be? Designing and Implementing a Deployable Multipath TCP", Usenix Symposium on Networked Systems Design and Implementation 2012, 2012, <<https://www.usenix.org/conference/nsdi12/how-hard-can-it-be-designing-and-implementing-deployable-multipath-tcp>>.
- [norm] Handley, M., Paxson, V., and C. Kreibich, "Network Intrusion Detection: Evasion, Traffic Normalization, and End-to-End Protocol Semantics", Usenix Security 2001, 2001, <[http://www.usenix.org/events/sec01/full\\_papers/handley/handley.pdf](http://www.usenix.org/events/sec01/full_papers/handley/handley.pdf)>.
- [RFC1122] Braden, R., Ed., "Requirements for Internet Hosts - Communication Layers", STD 3, RFC 1122, DOI 10.17487/RFC1122, October 1989, <<https://www.rfc-editor.org/info/rfc1122>>.
- [RFC1918] Rekhter, Y., Moskowitz, B., Karrenberg, D., de Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, DOI 10.17487/RFC1918, February 1996, <<https://www.rfc-editor.org/info/rfc1918>>.
- [RFC2018] Mathis, M., Mahdavi, J., Floyd, S., and A. Romanow, "TCP Selective Acknowledgment Options", RFC 2018, DOI 10.17487/RFC2018, October 1996, <<https://www.rfc-editor.org/info/rfc2018>>.
- [RFC2979] Freed, N., "Behavior of and Requirements for Internet Firewalls", RFC 2979, DOI 10.17487/RFC2979, October 2000, <<https://www.rfc-editor.org/info/rfc2979>>.
- [RFC2992] Hopps, C., "Analysis of an Equal-Cost Multi-Path Algorithm", RFC 2992, DOI 10.17487/RFC2992, November 2000, <<https://www.rfc-editor.org/info/rfc2992>>.

- [RFC3022] Srisuresh, P. and K. Egevang, "Traditional IP Network Address Translator (Traditional NAT)", RFC 3022, DOI 10.17487/RFC3022, January 2001, <<https://www.rfc-editor.org/info/rfc3022>>.
- [RFC3135] Border, J., Kojo, M., Griner, J., Montenegro, G., and Z. Shelby, "Performance Enhancing Proxies Intended to Mitigate Link-Related Degradations", RFC 3135, DOI 10.17487/RFC3135, June 2001, <<https://www.rfc-editor.org/info/rfc3135>>.
- [RFC4086] Eastlake 3rd, D., Schiller, J., and S. Crocker, "Randomness Requirements for Security", BCP 106, RFC 4086, DOI 10.17487/RFC4086, June 2005, <<https://www.rfc-editor.org/info/rfc4086>>.
- [RFC4987] Eddy, W., "TCP SYN Flooding Attacks and Common Mitigations", RFC 4987, DOI 10.17487/RFC4987, August 2007, <<https://www.rfc-editor.org/info/rfc4987>>.
- [RFC5681] Allman, M., Paxson, V., and E. Blanton, "TCP Congestion Control", RFC 5681, DOI 10.17487/RFC5681, September 2009, <<https://www.rfc-editor.org/info/rfc5681>>.
- [RFC6181] Bagnulo, M., "Threat Analysis for TCP Extensions for Multipath Operation with Multiple Addresses", RFC 6181, DOI 10.17487/RFC6181, March 2011, <<https://www.rfc-editor.org/info/rfc6181>>.
- [RFC6182] Ford, A., Raiciu, C., Handley, M., Barre, S., and J. Iyengar, "Architectural Guidelines for Multipath TCP Development", RFC 6182, DOI 10.17487/RFC6182, March 2011, <<https://www.rfc-editor.org/info/rfc6182>>.
- [RFC6356] Raiciu, C., Handley, M., and D. Wischik, "Coupled Congestion Control for Multipath Transport Protocols", RFC 6356, DOI 10.17487/RFC6356, October 2011, <<https://www.rfc-editor.org/info/rfc6356>>.
- [RFC6528] Gont, F. and S. Bellovin, "Defending against Sequence Number Attacks", RFC 6528, DOI 10.17487/RFC6528, February 2012, <<https://www.rfc-editor.org/info/rfc6528>>.
- [RFC6897] Scharf, M. and A. Ford, "Multipath TCP (MPTCP) Application Interface Considerations", RFC 6897, DOI 10.17487/RFC6897, March 2013, <<https://www.rfc-editor.org/info/rfc6897>>.

- [RFC7323] Borman, D., Braden, B., Jacobson, V., and R. Scheffenegger, Ed., "TCP Extensions for High Performance", RFC 7323, DOI 10.17487/RFC7323, September 2014, <<https://www.rfc-editor.org/info/rfc7323>>.
- [RFC7413] Cheng, Y., Chu, J., Radhakrishnan, S., and A. Jain, "TCP Fast Open", RFC 7413, DOI 10.17487/RFC7413, December 2014, <<https://www.rfc-editor.org/info/rfc7413>>.
- [RFC7430] Bagnulo, M., Paasch, C., Gont, F., Bonaventure, O., and C. Raiciu, "Analysis of Residual Threats and Possible Fixes for Multipath TCP (MPTCP)", RFC 7430, DOI 10.17487/RFC7430, July 2015, <<https://www.rfc-editor.org/info/rfc7430>>.
- [RFC8041] Bonaventure, O., Paasch, C., and G. Detal, "Use Cases and Operational Experience with Multipath TCP", RFC 8041, DOI 10.17487/RFC8041, January 2017, <<https://www.rfc-editor.org/info/rfc8041>>.
- [RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <<https://www.rfc-editor.org/info/rfc8126>>.
- [TCPLO] Ramaiah, A., "TCP option space extension", Work in Progress, March 2012.

## Appendix A. Notes on Use of TCP Options

The TCP option space is limited due to the length of the Data Offset field in the TCP header (4 bits), which defines the TCP header length in 32-bit words. With the standard TCP header being 20 bytes, this leaves a maximum of 40 bytes for options, and many of these may already be used by options such as timestamp and SACK.

We have performed a brief study on the commonly used TCP options in SYN, data, and pure ACK packets, and found that there is enough room to fit all the options we propose using in this document.

SYN packets typically include Maximum Segment Size (MSS) (4 bytes), window scale (3 bytes), SACK permitted (2 bytes), and timestamp (10 bytes) options. Together these sum to 19 bytes. Some operating systems appear to pad each option up to a word boundary, thus using 24 bytes (a brief survey suggests Windows XP and Mac OS X do this, whereas Linux does not). Optimistically, therefore, we have 21 bytes spare, or 16 if it has to be word-aligned. In either case, however, the SYN versions of Multipath Capable (12 bytes) and Join (12 or 16 bytes) options will fit in this remaining space.

Note that due to the use of a 64-bit data-level sequence space, it is feasible that MPTCP will not require the timestamp option for protection against wrapped sequence numbers (PAWS [RFC7323]), since the data-level sequence space has far less chance of wrapping. Confirmation of the validity of this optimisation is for further study.

TCP data packets typically carry timestamp options in every packet, taking 10 bytes (or 12 with padding). That leaves 30 bytes (or 28, if word-aligned). The Data Sequence Signal (DSS) option varies in length depending on whether the data sequence mapping and DATA\_ACK are included, and whether the sequence numbers in use are 4 or 8 octets. The maximum size of the DSS option is 28 bytes, so even that will fit in the available space. But unless a connection is both bidirectional and high-bandwidth, it is unlikely that all that option space will be required on each DSS option.

Within the DSS option, it is not necessary to include the data sequence mapping and DATA\_ACK in each packet, and in many cases it may be possible to alternate their presence (so long as the mapping covers the data being sent in the following packet). It would also be possible to alternate between 4- and 8-byte sequence numbers in each option.

On subflow and connection setup, an MPTCP option is also set on the third packet (an ACK). These are 20 bytes (for Multipath Capable)

and 24 bytes (for Join), both of which will fit in the available option space.

Pure ACKs in TCP typically contain only timestamps (10 bytes). Here, Multipath TCP typically needs to encode only the DATA\_ACK (maximum of 12 bytes). Occasionally, ACKs will contain SACK information. Depending on the number of lost packets, SACK may utilize the entire option space. If a DATA\_ACK had to be included, then it is probably necessary to reduce the number of SACK blocks to accommodate the DATA\_ACK. However, the presence of the DATA\_ACK is unlikely to be necessary in a case where SACK is in use, since until at least some of the SACK blocks have been retransmitted, the cumulative data-level ACK will not be moving forward (or if it does, due to retransmissions on another path, then that path can also be used to transmit the new DATA\_ACK).

The ADD\_ADDR option can be between 16 and 30 bytes, depending on whether IPv4 or IPv6 is used, and whether or not the port number is present. It is unlikely that such signaling would fit in a data packet (although if there is space, it is fine to include it). It is recommended to use duplicate ACKs with no other payload or options in order to transmit these rare signals. Note this is the reason for mandating that duplicate ACKs with MPTCP options are not taken as a signal of congestion.

## Appendix B. TCP Fast Open and MPTCP

TCP Fast Open (TFO) is an experimental TCP extension, described in [RFC7413], which has been introduced to allow sending data one RTT earlier than with regular TCP. This is considered a valuable gain as very short connections are very common, especially for HTTP request/response schemes. It achieves this by sending the SYN-segment together with the application's data and allowing the listener to reply immediately with data after the SYN/ACK. [RFC7413] secures this mechanism, by using a new TCP option that includes a cookie which is negotiated in a preceding connection.

When using TCP Fast Open in conjunction with MPTCP, there are two key points to take into account, detailed hereafter.

### B.1. TFO cookie request with MPTCP

When a TFO initiator first connects to a listener, it cannot immediately include data in the SYN for security reasons [RFC7413]. Instead, it requests a cookie that will be used in subsequent connections. This is done with the TCP cookie request/response options, of respectively 2 bytes and 6-18 bytes (depending on the chosen cookie length).

TFO and MPTCP can be combined provided that the total length of all the options does not exceed the maximum 40 bytes possible in TCP:

- o In the SYN: MPTCP uses a 4-bytes long MP\_CAPABLE option. The MPTCP and TFO options sum up to 6 bytes. With typical TCP-options using up to 19 bytes in the SYN (24 bytes if options are padded at a word boundary), there is enough space to combine the MP\_CAPABLE with the TFO Cookie Request.
- o In the SYN+ACK: MPTCP uses a 12-bytes long MP\_CAPABLE option, but now TFO can be as long as 18 bytes. Since the maximum option length may be exceeded, it is up to the listener to solve this by using a shorter cookie. As an example, if we consider that 19 bytes are used for classical TCP options, the maximum possible cookie length would be of 7 bytes. Note that the same limitation applies to subsequent connections, for the SYN packet (because the initiator then echoes back the cookie to the listener). Finally, if the security impact of reducing the cookie size is not deemed acceptable, the listener can reduce the amount of other TCP-options by omitting the TCP timestamps (as outlined in Appendix A).

#### B.2. Data sequence mapping under TFO

MPTCP uses, in the TCP establishment phase, a key exchange that is used to generate the Initial Data Sequence Numbers (IDSNs). In particular, the SYN with MP\_CAPABLE occupies the first octet of the data sequence space. With TFO, one way to handle the data sent together with the SYN would be to consider an implicit DSS mapping that covers that SYN segment (since there is not enough space in the SYN to include a DSS option). The problem with that approach is that if a middlebox modifies the TFO data, this will not be noticed by MPTCP because of the absence of a DSS-checksum. For example, a TCP (but not MPTCP)-aware middlebox could insert bytes at the beginning of the stream and adapt the TCP checksum and sequence numbers accordingly. With an implicit mapping, this would give to initiator and listener a different view on the DSS-mapping, with no way to detect this inconsistency as the DSS checksum is not present.

To solve this, the TFO data must not be considered part of the Data Sequence Number space: the SYN with MP\_CAPABLE still occupies the first octet of data sequence space, but then the first non-TFO data byte occupies the second octet. This guarantees that, if the use of DSS-checksum is negotiated, all data in the data sequence number space is checksummed. We also note that this does not entail a loss of functionality, because TFO-data is always only sent on the initial subflow before any attempt to create additional subflows.



## B.3. Connection establishment examples

The following shows a few examples of possible TFO+MPTCP establishment scenarios.

Before an initiator can send data together with the SYN, it must request a cookie to the listener, as shown in Figure 18. This is done by simply combining the TFO and MPTCP options.

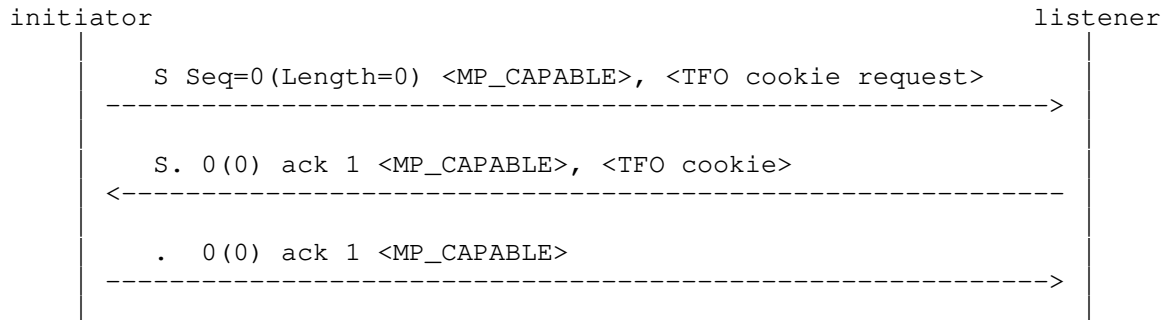


Figure 18: Cookie request - sequence number and length are annotated as Seq(Length) and used hereafter in the figures.

Once this is done, the received cookie can be used for TFO, as shown in Figure 19. In this example, the initiator first sends 20 bytes in the SYN. The listener immediately replies with 100 bytes following the SYN-ACK upon which the initiator replies with 20 more bytes. Note that the last segment in the figure has a TCP sequence number of 21, while the DSS subflow sequence number is 1 (because the TFO data is not part of the data sequence number space, as explained in Section Appendix B.2).

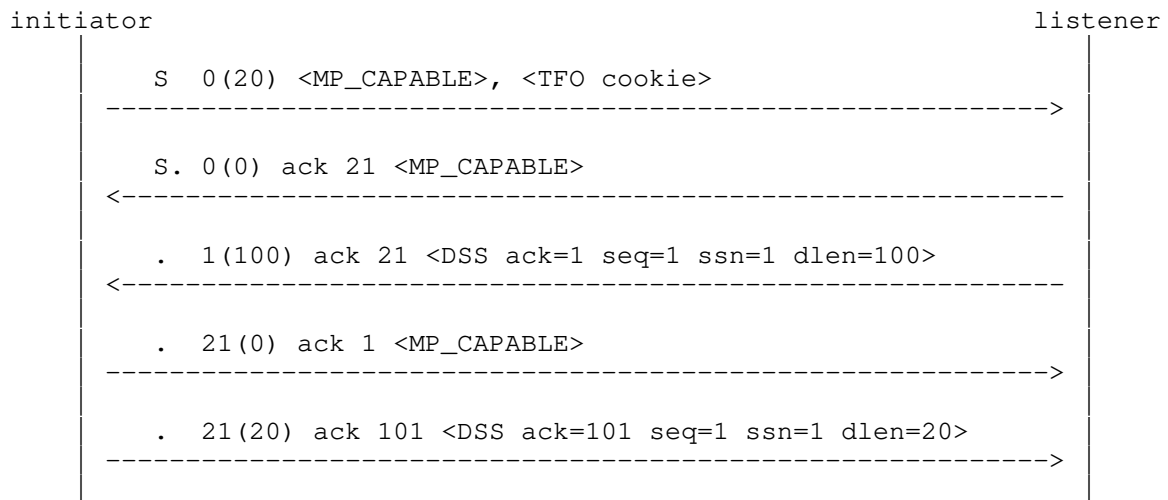


Figure 19: The listener supports TFO

In Figure 20, the listener does not support TFO. The initiator detects that no state is created in the listener (as no data is acked), and now sends the MP\_CAPABLE in the third ack, in order for the listener to build its MPTCP context at the end of the establishment. Now, the tfo data, retransmitted, becomes part of the data sequence mapping because it is effectively sent (in fact re-sent) after the establishment.

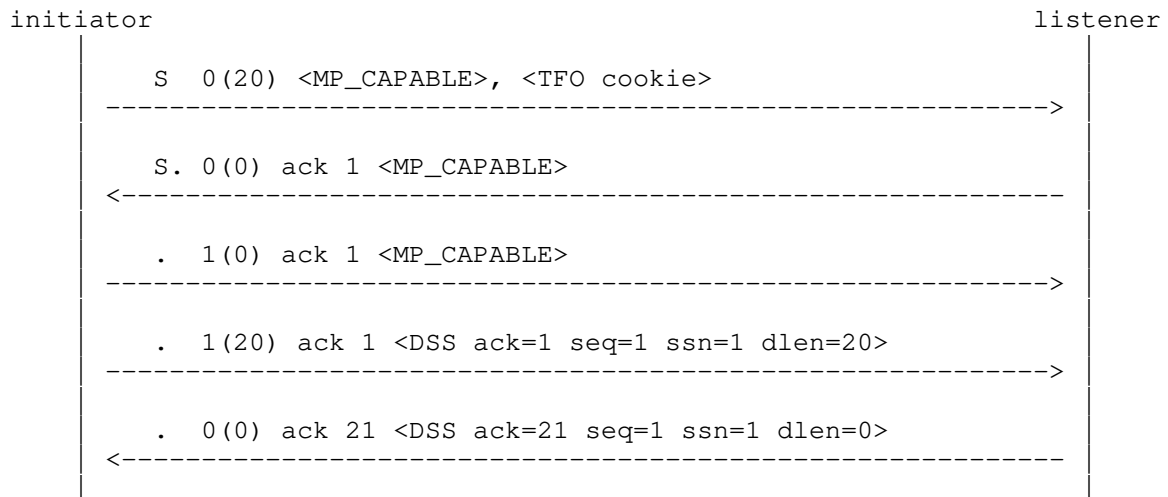


Figure 20: The listener does not support TFO

It is also possible that the listener acknowledges only part of the TFO data, as illustrated in Figure 21. The initiator will simply retransmit the missing data together with a DSS-mapping.

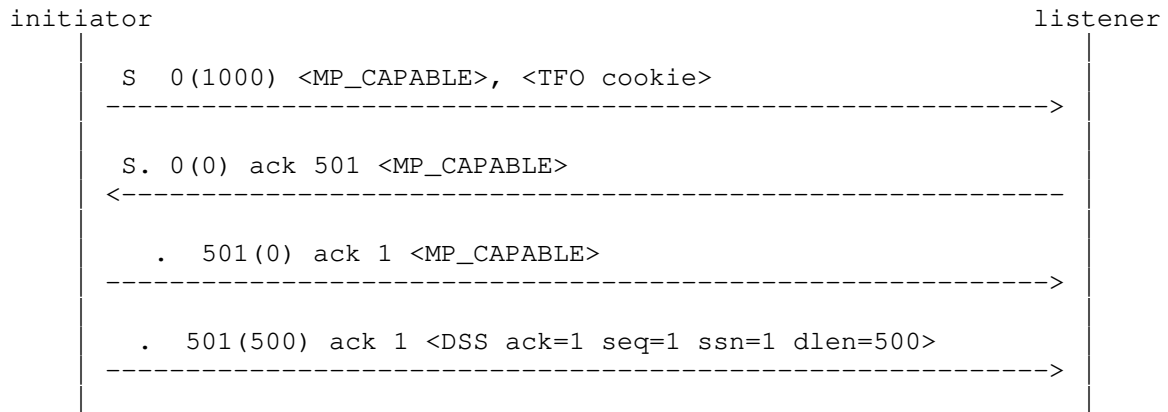


Figure 21: Partial data acknowledgement

## Appendix C. Control Blocks

Conceptually, an MPTCP connection can be represented as an MPTCP protocol control block (PCB) that contains several variables that track the progress and the state of the MPTCP connection and a set of linked TCP control blocks that correspond to the subflows that have been established.

RFC 793 [RFC0793] specifies several state variables. Whenever possible, we reuse the same terminology as RFC 793 to describe the state variables that are maintained by MPTCP.

### C.1. MPTCP Control Block

The MPTCP control block contains the following variable per connection.

#### C.1.1. Authentication and Metadata

**Local.Token (32 bits):** This is the token chosen by the local host on this MPTCP connection. The token must be unique among all established MPTCP connections, and is generated from the local key.

**Local.Key (64 bits):** This is the key sent by the local host on this MPTCP connection.

Remote.Token (32 bits): This is the token chosen by the remote host on this MPTCP connection, generated from the remote key.

Remote.Key (64 bits): This is the key chosen by the remote host on this MPTCP connection

MPTCP.Checksum (flag): This flag is set to true if at least one of the hosts has set the A bit in the MP\_CAPABLE options exchanged during connection establishment, and is set to false otherwise. If this flag is set, the checksum must be computed in all DSS options.

#### C.1.2. Sending Side

SND.UNA (64 bits): This is the data sequence number of the next byte to be acknowledged, at the MPTCP connection level. This variable is updated upon reception of a DSS option containing a DATA\_ACK.

SND.NXT (64 bits): This is the data sequence number of the next byte to be sent. SND.NXT is used to determine the value of the DSN in the DSS option.

SND.WND (32 bits with RFC 7323, 16 bits otherwise): This is the sending window. MPTCP maintains the sending window at the MPTCP connection level and the same window is shared by all subflows. All subflows use the MPTCP connection level SND.WND to compute the SEQ.WND value that is sent in each transmitted segment.

#### C.1.3. Receiving Side

RCV.NXT (64 bits): This is the data sequence number of the next byte that is expected on the MPTCP connection. This state variable is modified upon reception of in-order data. The value of RCV.NXT is used to specify the DATA\_ACK that is sent in the DSS option on all subflows.

RCV.WND (32 bits with RFC 7323, 16 bits otherwise): This is the connection-level receive window, which is the maximum of the RCV.WND on all the subflows.

#### C.2. TCP Control Blocks

The MPTCP control block also contains a list of the TCP control blocks that are associated with the MPTCP connection.

Note that the TCP control block on the TCP subflows does not contain the RCV.WND and SND.WND state variables as these are maintained at the MPTCP connection level and not at the subflow level.

Inside each TCP control block, the following state variables are defined.

#### C.2.1. Sending Side

SND.UNA (32 bits): This is the sequence number of the next byte to be acknowledged on the subflow. This variable is updated upon reception of each TCP acknowledgment on the subflow.

SND.NXT (32 bits): This is the sequence number of the next byte to be sent on the subflow. SND.NXT is used to set the value of SEG.SEQ upon transmission of the next segment.

#### C.2.2. Receiving Side

RCV.NXT (32 bits): This is the sequence number of the next byte that is expected on the subflow. This state variable is modified upon reception of in-order segments. The value of RCV.NXT is copied to the SEG.ACK field of the next segments transmitted on the subflow.

RCV.WND (32 bits with RFC 7323, 16 bits otherwise): This is the subflow-level receive window that is updated with the window field from the segments received on this subflow.

### Appendix D. Finite State Machine

The diagram in Figure 22 shows the Finite State Machine for connection-level closure. This illustrates how the DATA\_FIN connection-level signal (indicated in the diagram as the DFIN flag on a DATA\_ACK) interacts with subflow-level FINs, and permits "break-before-make" handover between subflows.

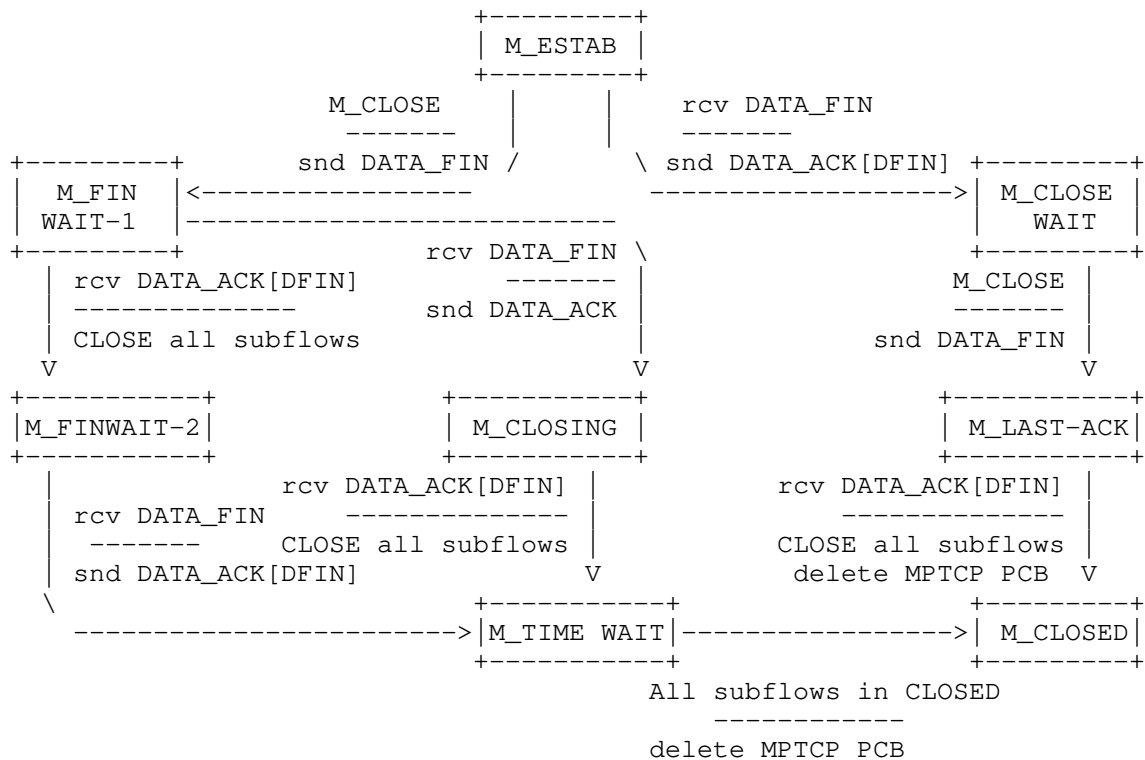


Figure 22: Finite State Machine for Connection Closure

## Appendix E. Changes from RFC6824

This section lists the key technical changes between RFC6824, specifying MPTCP v0, and this document, which obsoletes RFC6824 and specifies MPTCP v1. Note that this specification is not backwards compatible with RFC6824.

- o The document incorporates lessons learnt from the various implementations, deployments and experiments gathered in the documents "Use Cases and Operational Experience with Multipath TCP" [RFC8041] and the IETF Journal article "Multipath TCP Deployments" [deployments].
- o Connection initiation, through the exchange of the MP\_CAPABLE MPTCP option, is different from RFC6824. The SYN no longer includes the initiator's key, allowing the MP\_CAPABLE option on the SYN to be shorter in length, and to avoid duplicating the sending of keying material.

- o This also ensures reliable delivery of the key on the MP\_CAPABLE option by allowing its transmission to be combined with data and thus using TCP's in-built reliability mechanism. If the initiator does not immediately have data to send, the MP\_CAPABLE option with the keys will be repeated on the first data packet. If the other end is first to send, then the presence of the DSS option implicitly confirms the receipt of the MP\_CAPABLE.
- o In the Flags field of MP\_CAPABLE, C is now assigned to mean that the sender of this option will not accept additional MPTCP subflows to the source address and port. This is an efficiency improvement, for example where the sender is behind a strict NAT.
- o In the Flags field of MP\_CAPABLE, H now indicates the use of HMAC-SHA256 (rather than HMAC-SHA1).
- o Connection initiation also defines the procedure for version negotiation, for implementations that support both v0 (RFC6824) and v1 (this document).
- o The HMAC-SHA256 (rather than HMAC-SHA1) algorithm is used, as the algorithm provides better security. It is used to generate the token in the MP\_JOIN and ADD\_ADDR messages, and to set the initial data sequence number.
- o A new subflow-level option exists to signal reasons for sending a RST on a subflow (MP\_TCP\_RST Section 3.6), which can help an implementation decide whether to attempt later re-connection.
- o The MP\_PRIO option (Section 3.3.8), which is used to signal a change of priority for a subflow, no longer includes the AddrID field. Its purpose was to allow the changed priority to be applied on a subflow other than the one it was sent on. However, it has been realised that this could be used by a man-in-the-middle to divert all traffic on to its own path, and MP\_PRIO does not include a token or other security mechanism.
- o The ADD\_ADDR option (Section 3.4.1), which is used to inform the other host about another potential address, is different in several ways. It now includes an HMAC of the added address, for enhanced security. In addition, reliability for the ADD\_ADDR option has been added: the IPVer field is replaced with a flag field, and one flag is assigned (E) which is used as an 'Echo' so a host can indicate that it has received the option.
- o An additional way of performing a Fast Close is described, by sending a MP\_FASTCLOSE option on a RST on all subflows. This

allows the host to tear down the subflows and the connection immediately.

- o In the IANA registry a new MPTCP subtype option, MP\_EXPERIMENTAL, is reserved for private experiments. However, the document doesn't define how to use the subtype option.
- o A new Appendix discusses the usage of both the MPTCP and TCP Fast Open on the same packet (Appendix B).

#### Authors' Addresses

Alan Ford  
Pexip

EMail: alan.ford@gmail.com

Costin Raiciu  
University Politehnica of Bucharest  
Splaiul Independentei 313  
Bucharest  
Romania

EMail: costin.raiciu@cs.pub.ro

Mark Handley  
University College London  
Gower Street  
London WC1E 6BT  
UK

EMail: m.handley@cs.ucl.ac.uk

Olivier Bonaventure  
Universite catholique de Louvain  
Pl. Ste Barbe, 2  
Louvain-la-Neuve 1348  
Belgium

EMail: olivier.bonaventure@uclouvain.be



Christoph Paasch  
Apple, Inc.  
Cupertino  
US

EMail: [cpaasch@apple.com](mailto:cpaasch@apple.com)