Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 3, 2014

M. Bagnulo
UC3M

February 12, 2014

Secure MPTCP
draft-bagnulo-mptcp-secure-00

Abstract

This memo contains some initial thoughts about how to secure MPTCP. As currently defined, MPTCP provides basic security features to protect the MPTCP signaling and the data flows unprotected. In this note, we explore the possible use to tcpcrypt to provide enhanced security to MPTCP.

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 August 3, 2014.

Copyright Notice

Copyright (c) 2014 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 2
2. Initial SMTCP connection 3
3. Adding new subflows 4
4. Exchanging data 5
5. Backward compatibility 6
6. Concluding remarks 7
7. Security Considerations 7
8. IANA Considerations 7
9. Acknowledgements 7
10. References 7
10.1. Normative References 8
10.2. Informative References 8
Author's Address 8

1. Introduction

Multi-path TCP (MPTCP) [RFC6824] defines the extensions to TCP that allow transmitting data over multiple paths in a single TCP connection. This is achieved by opening multiple subflows within the same TCP connection. Each subflow is associated to a different address/port pair. As currently defined, MPTCP provides basic security for the signaling used to establish the subflows. A threat analysis for MPTCP is presented in [RFC6181] and a residual threat analysis is presented in [I-D.bagnulo-mptcp-attacks]. From these analysis we can extract that MPTCP as currently defined is vulnerable to attackers that can eavesdrop the initial connection establishment exchange and also to attackers that can intercept any subflow establishment exchange. In addition, MPTCP does not provide any protection to the data stream (other than splitting the data stream over multiple paths), as this was a non goal of the MPTCP design. In [I-D.bagnulo-mptcp-attacks] it is concluded that if a more secure version of MPTCP should be pursued, the path to follow would be to protect the data stream rather than trying to provide additional security to the signaling. The reader is referred to the aforementioned reference for additional insight why this is the case. The goal of this document is provide initial considerations about how to provide enhanced security to MPTCP by securing the data stream.

In this note, we analyze the use of tcpcryp [I-D.bittau-tcp-crypt] to secure MPTCP. tcpcrypt defines extensions to TCP so opportunistically encrypt the data stream of a TCP connection. By using tcpcrypt in MPTCP, we would be able to provide enhanced security to MPTCP. We note however, that the resulting solution would still be vulnerable to Man-in-the-Middle attacks during the initial key negotiation. However, the attacker in this case must be active and must remain located along the path during the whole lifetime of the connection.

We call SMTCP to the integration of MPTCP with tcpcryp. This would provide stronger security of MPTCP, both for the signaling and for the data. Since all the MPTCP signaling will be protected by tcpcrypt (i.e. encrypted and its integrity protected) there is no need for the existent MPTCP security mechanisms when used with tcpcrypt. This means that there is no need to negotiate a the current MPTCP key and that the HMAC protection provided by the MPTCP protocol is not needed (except for backward compatibility issues). All the protection will be achieved with the tcpcrypt extensions.

2. Initial SMTCP connection

Suppose both A and B are SMTCP capable. Suppose A has both IPA1 and IPA2. Suppose A initiates a SMTCP connection with B. The exchange would look like as follows:

A -> B: SYN + MP_CAPABLE (including A's key (ka) and C bit set) + CRYPT/Hello This contains 15 bytes of options (the motivation for including both the MP_CAPABLE and the CRYPT option is for backward compatibility, see Section 5). SYN packets usually carry as well MSS (4 bytes), SACK (2 bytes) and Window-Scale (3 bytes). Negotiating timestamps would be 10 more bytes. As options has a maximum length of 40 bytes, this would be compatible with all the mentioned options.

B -> A: SYN/ACK + MP_CAPABLE (including B's key (kb) and the C bit set) + CRYPT/PKCONF (with pub-cipher-list) Assuming we have 2 algorithms in the list, this is 10 bytes, making a total of 22 bytes. This packet also usually carries the same options than the SYN packet, so in this case, not all the mentioned additional options would fit.

A -> B: INIT (3 bytes of options) plus crypto data in the payload (The MP_CAPABLE option is needed since the keys are generated by tcpcrypt)

B -> A: INIT (3 bytes of options) plus crypto data in the payload

During the 3-way handshake MPTCP generates the following values:

The key: session to key to protect the signaling (one per each side)

The Token: used as connection identifier (one per each side)

The IDSN: Initial Data Seq Number (one per each side)

The idea would be to derive them from the tcpcrypt values.

- The keys: tcpcrypt generates 4 keys, kec, kac, kes and kas. These will be used to secure MPTCP, as discussed later on. for the purposes of MPTCP signaling, the keys that will be used are the authentication keys, so the keys for MPTCP are kac and kas.

- The tokens: tcpcrypt generates a Session ID, which is the full length of a hash output. The thing is that tcpcrypt generates a single SID for both endpoints, while MPTCP generates one per endpoint. In addition, MPTCP needed to generate a pair of ISDNs.

We could then generate the MPTCP values out of the tcpcrypt values as follows

Key A= kac

Key B= Kas

Token A= 32 most significant bits of (hash(ka))

Token B= 32 most significant bits of (hash(kb))

IDSN A= 64 least significant bits of (hash(kac + SID))

IDSN B= 64 least significant bits of (hash(kas + SID))

DISCUSSION: We need to exchange ka in the MP_CAPABLE message because of backwards compatibility issues (see Section 5), so there is no easy way around it. However, the reason to exchange kb in the MP_CAPABLE is because we need to be able to generate Token B in a way that is unique in host B. This seems an overkill (exchanging a 64 bit key to achieve a 32 bit unique token). It could be possible to define a new MP_CAPABLE option that would exchange a 32 bit long token directly. We also need to understand the security implications of exchanging the token in the clear. Another alternative would be to generate the token as a hash of kas and the SID and if it happens to clash, restart the connection. Tis may make sense depending how likely is this to happen. TODO: work out the other flavors of tcpcrypt connection initiation, including the use of cached keys

3. Adding new subflows

The options for this is whether to treat a new subflow as a new tcpcrypt connection or not, the implication being that a new tcpcrypt connection uses a different shared secret and hence different keys (even though not new public key operations are needed). Probably this is not the way to go. All the flows of a MPTCP connection should be part of the same tcpcrypt session. (the other option would

imply that there are different keys for the same MPTCP session, which may be cumbersome?)

So, one simple way of doing this would be to simply use the existent MPTCP exchange to add a new subflow with the MPTCP security measures. This implies sending an MP_JOIN containing the receiver's token and a random number which will be responded with another MP_JOIN and the final JOIN message. The tcpcrypt keys are used instead of the regular MPTCP keys.

(similar approach can be used for the ADD_ADDR option, which is secure using an HMAC using the tcpcryp derived key)

A alternative approach would be to drop completely the MPTCP security mechanisms and use the tcpcryp MAC option to secure the MPTCP signaling. This implies that the tcpcryp MAC option would need also to protect the MPTCP MP_JOIN option

4. Exchanging data

Once the keys and the other values have been negotiated, data can flow. All data in the MPTCP connection will be encrypted with tcpcrypt keys and its integrity protected using the tcpcrypt MAC option. This adds 22 bytes of options (assuming 160 bits long hash).

Question: do we need to have a 160 bit long hash or can we live with less?

Now, MPTCP includes the DSS option in order to synchronize the data sequence number with the sequence numbers of the subflows. The DSS option max length is 28 bytes. The results it that the MAC option plus the DSS option are 50 bytes, which is a problem.

The good news is that the DSS does not need to be sent in every segment and that 28 bytes is the maximum length.

Currently the DSS option includes information both about DSN mapping to subflow seq number and data ack. In order to limit the size of the option, one option is to prevent that both Data Ack and DSN to subflow seq numbers mappings are sent in the same option. This would result that when Data acks are sent, the DSS option has a maximum of 12 bytes and when DSN to subflow seq number mapping are sent, the max length is 20 bytes. This is still 2 bytes too long. There are two ways we can shrink this. One option is to prevent the use of Checksum when tcpcrypt is used. checksum is optional, so this could be done. Moreover, it makes sense to do this, because all the information protected in this checksum is protected by the tcpcrypt MAC option. this results that the DSS option is now 18 bytes, which

with the tcpcrypt MAC option will make up to 40 bytes of tcp options. The other possible way to shrink this is to use the 4 bytes seq numbers rather than the 8 ones. This would reduce the DSS option to 14 bytes for the DSN to subflow seq number mapping and to 8 bytes in the case of Data acks.

The DSS option will be sent in the clear i.e. not encrypted by tcpcrypt. The MAC option must cover the DSS option (correct?). This implies that we need to add to the MAC data structure the DSS option.

Question: how often the DSS needs to be sent? I mean, if we send the DSS and the MAC options, we will be using all the TCP option room, so, not SACK can be sent, which is bad.

5. Backward compatibility

There will be the following 5 types of node:

MPTCP nodes: supports MPTCP as defined in RFC6824 or RFC6824bis but does not support tcpcrypt,

tcpcryp nodes: supports tcpcrypt but does not support MPTCP,

SMTCP capable nodes: support MPTCP and tcpcrypt and the use of tcpcrypt to secure MPTCP,

legacy nodes: dont support neither tcpcrypt nor MPTCP,

MPTCP/tcpcrypt nodes: supports both tcpcrypt and MPTCP but does not support the use of tcpcrypt to protect MPTCP.

The expected behavior is as following:

a. SMTCP contacts a SMTCP node, SMTCP should be used

b. SMTCP contacts a MPTCP node, MPTCP should be used

c. SMTCP contacts a tcpcrypt node, tcpcrypt should be used

d. SMTCP contacts a MPTCP/tcpcrypt node, not sure what should happen... should MPTCP and tcpcrypt be used in a non integrated fashion? (not sure if there is enough space in the TCP options for this...)

e. SMTCP contacts a legacy node, TCP should be used.

In order to achieve, we use the following approach. In the initial SYN of the initial 3-way handshake, both the CRYPT/Hello option (3

bytes) plus the MP_CAPABLE option including the initiator's key (12 bytes) should be sent. This allows supporting b), c) and e) i.e. the receiver can discard either of the two options or both of them resulting in each of the mentioned cases.

In order to support case a) (and to distinguish it from case d), we need to signal it in an explicit way. I guess the easiest way is to use one of the flags C to H in the MP_CAPABLE message. Let's assume it is the C(rypt) flag. If the C flag is set and the CRYPT/Hello option is present, this means SMTCP (i.e. use tcpcrypt to protect MPTCP signaling and data).

6. Concluding remarks

One main challenge in order to use tcpcrypt to secure MPTCP is the option space. There is little room for TCP options and this approach would consume most of it, which would prevent the use of other options like SACK. One way to address this would be that tcpcrypt is changed to send the MAC as part of the data stream rather than an option. As tcpcrypt is being discussed, this can be an option.

A second issue to consider is how this would work with TSO. Currently MPTCP is compatible with TSO and it would be important that SMPTCP is also compatible.

Another comment is that it would be possible to secure MPTCP using something like TLS opportunistically and transparently to the application. This is TBD as an alternative approach.

7. Security Considerations

This whole document is about securing MPTCP. In future versions of the document, this section could include a residual threat analysis.

8. IANA Considerations

TBD

9. Acknowledgements

The authors thank ...

10. References

10.1. Normative References

[I-D.bittau-tcp-crypt]
Bittau, A., Boneh, D., Hamburg, M., Handley, M., Mazieres, D., and Q. Slack, "Cryptographic protection of TCP Streams (tcpcrypt)", draft-bittau-tcp-crypt-03 (work in progress), September 2012.

[RFC6824] Ford, A., Raiciu, C., Handley, M., and O. Bonaventure, "TCP Extensions for Multipath Operation with Multiple Addresses", RFC 6824, January 2013.

10.2. Informative References

[I-D.bagnulo-mptcp-attacks]
Bagnulo, M., Paasch, C., Gont, F., Bonaventure, O., and C. Raiciu, "Analysis of MPTCP residual threats and possible fixes", draft-bagnulo-mptcp-attacks-01 (work in progress), October 2013.

[RFC6181] Bagnulo, M., "Threat Analysis for TCP Extensions for Multipath Operation with Multiple Addresses", RFC 6181, March 2011.

Author’s Address

Marcelo Bagnulo
Universidad Carlos III de Madrid
Av. Universidad 30
Leganes, Madrid 28911
SPAIN

Phone: 34 91 6249500
Email: marcelo@it.uc3m.es
URI: http://www.it.uc3m.es

MPTCP Working Group
INTERNET-DRAFT
Intended Status: Informational
Expires: August 14, 2014

L. Deng
D. Liu
T. Sun
China Mobile

February 14, 2014

MPTCP Proxy for Mobile Networks
draft-deng-mptcp-mobile-network-proxy-00

Abstract

This document discusses the motivation and usecases for ISP deployed MPTCP proxies in mobile networks.

Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

Copyright and License Notice

Copyright (c) 2013 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
2 Terminology 3
3 Considerations for MPTCP Proxy 3
4 Use-cases for Network deployed MPTCP Proxy 4
4.1 Dynamic traffic offloading based on network information . . 4
4.2 Resource pooling for reduced expense 4
5 Requirements for MPTCP Proxy 5
5.1 Protocol transition 5
5.2 Traffic mediation 5
5 Security Considerations 5
6 IANA Considerations 5
7 References 7
7.1 Normative References 7
Authors' Addresses 8

1 Introduction

Due the scarcity of wireless frequency resources and the instability of wireless signals, combined with the operators' strong motive to preserve service upgrade with smooth network evolution, make full use of mobile terminal's multi-homing capability has long been a quest for mobile networks.

In particular, the motivations include resource pooling for better performance (where the network could provide a better performance for resource-intensive services by allowing them to transparently using combined capacities from different RATs) as well as intelligent selection for better accommodation and seamless handover for better mobility.

Since R6, 3GPP network defined GAN, interfaces for non-3GPP RATs through GERAN simulation. In R7, I-WLAN was introduced to 3GPP network, for inter-working of PLMN with WLAN RAT. In R8, it is specified that a shared anchor could be used for both I-WLAN and PS RATs, yielding seamless handover. Since R8, there have been work on EPS's mobility support for simultaneous multiple RATs through different PDN connections (MAPCON). Most recently, in R10, it is possible to use EPS's mobility support for simultaneous multiple RATs through a single PDN connection (IFOM).

However, there is still not possible for a single IP flow to make full use of multiple interfaces simultaneously.

2 Terminology

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

3 Considerations for MPTCP Proxy

MPTCP[RFC6824] offers transparent wireless resource pooling for a single "IP flow" for multi-homing UEs with least network complications, as it effectively implements automatic RAT selection/handover/pooling through TCP's adaptive end-to-end rating mechanism[RFC6356].

However, end-to-end MPTCP solution deprives network's control over service/RAT preference, which is considered to be essential for better operation and service provision in 3GPP networks. As the same time, it has to suffer from compatibility issues with legacy application SPs who are reluctant to support MPTCP natively.

Therefore, network deployed MPTCP proxy comes as a compromise, which would certainly benefit MPTCP-enabled UEs without SP's MPTCP deployment by providing protocol adaptation, and at the same time maintain as the wireless network operator's policy enforcement point for their preferred network selection/usage strategies.

4 Use-cases for Network deployed MPTCP Proxy

For 3rd party service provider who does not supporting MPTCP in their servers, the network deployed proxy could be used to enable MPTCP capability in resource pooling from various radio access networks for enhanced QoE/mobility.

As for 3rd party service providers supporting MPTCP, the network deployed proxy could also bring benefits to both the operator and the users by enabling the following benefits.

4.1 Dynamic traffic offloading based on network information

For real-time interactive services with higher QoS requirements it is expected that 3GPP network can provide better guarantees on the average case. For bulk data transfer who is satisfied with best-effort delivery, Wi-Fi would be a great choice. But the vertical partition does not fit everywhere for the wireless condition itself is quite dynamic and hard to predict. It is important to implement adaptive offloading mechanisms in order to achieve higher resource utility with ever changing radio environment for a possibly moving terminal based on network status, e.g. cell load, AP's signal intensity, user's subscription type, etc.

4.2 Resource pooling for reduced expense

Due to its low construction and operation expenses, Wi-Fi has been adopted by mobile operators as a complementary RAT for their traditional 3GPP networks. However, different construction and operation expenses of various radio networks result in differences in charging rates/policies for different RATs.

For instance, Wi-Fi access may be charged by the access duration, while the 3GPP access may be charged by the consumed data volume. Even if using the same policy, Wi-Fi service is expected to be much cheaper than 3GPP data service.

Moreover, different subscription packages may offer various data plans for various RATs. For instance, a basic 4G package may contain free data volume as well free Wi-Fi access too.

By enabling MPTCP session between UE and network proxy, via mediating sub-flow data traffic based on their Radio access types and the user's subscription package, it is possible to further reduce the usage expenses from both sides of the network and user.

5 Requirements for MPTCP Proxy

In order to realize the above use-cases, it is expected that a network deployed MPTCP proxy provide the following functionality:

5.1 Protocol transition

To allow a MPCTP-enabled UE to make full use of the multiple radio interfaces even if it is communicating with a non-MPTCP server, the proxy should support

(a) Detection of UE's MPTCP capability;

(b) Negotiation with MPTCP UE on behalf of non-MPTCP SP;

(c) Translation/Mapping between TCP and MPTCP sessions.

5.2 Traffic mediation

(a) Anchoring of sub-flow traffic: On one hand, it is not always possible for a single GW be sitting on the path of every sub-flow from a MPTCP session, hence explicit traffic anchoring to enable a single point of general control over MPTCP sub-flows should be considered.

(b) Mediation of sub-flow traffic: On the other hand, for fine-grained mediation of sub-flow traffic, both static and dynamic selection/offloading/pooling policies should be allowed. For instance, "always prefer Wi-Fi over 3GPP" could be a static policy for bulk data transfer services, while "use 3GPP only for backup unless Wi-Fi is congested" could be a dynamic offloading policy for a un-prioritized VoIP service.

5 Security Considerations

TBA.

6 IANA Considerations

There is no IANA action in this document.

7 References

7.1 Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC3550] Schulzrinne, H., Casner, S., Frederick, R., and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", STD 64, RFC 3550, July 2003.

[RFC3611] Friedman, T., Ed., Caceres, R., Ed., and A. Clark, Ed., "RTP Control Protocol Extended Reports (RTCP XR)", RFC 3611, November 2003.

[IFOM] IP Flow Mobility and seamless WLAN offload. 3GPP work item 450041.

[RFC6824] Ford, A., Raiciu, C., Handley, M., and O. Bonaventure, "TCP Extensions for Multipath Operation with Multiple Addresses", RFC 6824, January 2013.

[RFC6356] Raiciu, C., Handley, M., and D. Wischik , "Coupled Congestion Control for Multipath Transport Protocols", RFC 6356, October 2011.

Authors' Addresses

Lingli Deng
China Mobile

Email: Email: denglingli@chinamobile.com

Dapeng Liu
China Mobile

Email: Email: liudapeng@chinamobile.com

Tao Sun
China Mobile

Email: Email: suntao@chinamobile.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: September 28, 2015

M. Bagnulo
UC3M
C. Paasch
UCLouvain
F. Gont
SI6 Networks / UTN-FRH
O. Bonaventure
UCLouvain
C. Raiciu
UPB

March 27, 2015

Analysis of MPTCP residual threats and possible fixes
draft-ietf-mptcp-attacks-04

Abstract

This documents performs an analysis of the residual threats for MPTCP and explores possible solutions to them.

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 September 28, 2015.

Copyright Notice

Copyright (c) 2015 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 2
2. ADD_ADDR attack 4
2.1. Possible security enhancements to prevent this attack . . 10
3. DoS attack on MP_JOIN 10
3.1. Possible security enhancements to prevent this attack . . 11
4. SYN flooding amplification 11
4.1. Possible security enhancements to prevent this attack . . 12
5. Eavesdropper in the initial handshake 12
5.1. Possible security enhancements to prevent this attack . . 13
6. SYN/JOIN attack 13
6.1. Possible security enhancements to prevent this attack . . 14
7. Reccomendations 14
7.1. Security enhancements for MPTCP 15
8. Security considerations 15
9. IANA Considerations 15
10. Acknowledgments 15
11. References 15
11.1. Normative References 16
11.2. Informative References 16
Authors' Addresses 17

1. Introduction

This document provides a complement to the threat analysis for Multipath TCP (MPTCP) [RFC6824] documented in RFC 6181 [RFC6181]. RFC 6181 provided a threat analysis for the general solution space of extending TCP to operate with multiple IP addresses per connection. Its main goal was to leverage previous experience acquired during the design of other multi-address protocols, notably SHIM6 [RFC5533], SCTP [RFC4960] and MIPv6 [RFC6275] for designing MPTCP. Thus, RFC 6181 was produced before the actual MPTCP specification (RFC6824) was completed, and documented a set of reccommendations that were considered during the production of such specification.

This document complements RFC 6181 with a vulnerability analysis of the specific mechanisms specified in RFC 6824. The motivation for this analysis is to identify possible security issues with MPTCP as currently specified and propose security enhancements to address the identified security issues.

The goal of the security mechanisms defined in RFC 6824 were to make MPTCP no worse than currently available single-path TCP. We believe that this goal is still valid, so we will perform our analysis on the same grounds. This document describes all the threats identified that are specific to MPTCP (as defined in RFC6824) that are not possible with (single-path) TCP. This means that threats that are common to TCP and MPTCP are not covered in this document.

Types of attackers: for all attacks considered in this document, we identify the type of attacker. We can classify the attackers based on their location as follows:

- Off-path attacker. This is an attacker that does not need to be located in any of the paths of the MPTCP session at any point in time during the lifetime of the MPTCP session. This means that the Off-path attacker cannot eavesdrop any of the packets of the MPTCP session.

- Partial time On-path attacker. This is an attacker that needs to be in at least one of the paths during part but not during the entire lifetime of the MPTCP session. The attacker can be in the forward and/or backward directions, for the initial subflow and/or other subflows. The specific needs of the attacker will be made explicit in the attack description.

- On-path attacker. This attacker needs to be on at least one of the paths during the whole duration of the MPTCP session. The attacker can be in the forward and/or backward directions, for the initial subflow and/or other subflows. The specific needs of the attacker will be made explicit in the attack description.

We can also classify the attackers based on their actions as follows:

- Eavesdropper. The attacker is able to capture some of the packets of the MPTCP session to perform the attack, but it is not capable of changing, discarding or delaying any packet of the MPTCP session. The attacker can be in the forward and/or backward directions, for the initial subflow and/or other subflows. The specific needs of the attacker will be made explicit in the attack description.

- Active attacker. The attacker is able to change, discard or delay some of the packets of the MPTCP session. The attacker can be in the forward and/or backward directions, for the initial subflow and/or other subflows. The specific needs of the attacker will be made explicit in the attack description.

In this document, we consider the following possible combinations of attackers:

- an On-path eavesdropper
- an On-path active attacker
- an Off-path active attacker
- a Partial-time On-path eavesdropper
- a Partial-time On-path active attacker

In the rest of the document we describe different attacks that are possible against the MPTCP protocol specified in RFC6824 and we propose possible security enhancements to address them.

2. ADD_ADDR attack

Summary of the attack:

Type of attack: MPTCP session hijack enabling Man-in-the-Middle.

Type of attacker: Off-path, active attacker.

Description:

In this attack, the attacker uses the ADD_ADDR option defined in RFC6824 to hijack an ongoing MPTCP session and enables himself to perform a Man-in-the-Middle attack on the MPTCP session.

Consider the following scenario. Host A with address IPA has one MPTCP session with Host B with address IPB. The MPTCP subflow between IPA and IPB is using port PA on host A and port PB on host B. The tokens for the MPTCP session are TA and TB for Host A and Host B respectively. Host C is the attacker. It owns address IPC. The attack is executed as follows:

1. Host C sends a forged packet with source address IPA, destination address IPB, source port PA and destination port PB. The packet has the ACK flag set. The TCP sequence number for the segment is i and the ACK sequence number is j. We will assume all these are valid, we discuss what the attacker needs to figure these ones later on. The packet contains the ADD_ADDR option. The ADD_ADDR option announces IPC as an alternative address for the connection. It also contains an eight bit address identifier which does not bring any strong security benefit.

2. Host B receives the ADD_ADDR message and it replies by sending a TCP SYN packet. (Note: the MPTCP specification states that the host receiving the ADD_ADDR option may initiate a new subflow. If the host is configured so that it does not initiate a new subflow the attack will not succeed. For example, on the current Linux implementation, the server does not create subflows. Only the client does so.) The source address for the packet is IPB, the destination address for the packet is IPC, the source port is PB' and the destination port is PA' (It is not required that PA=PA' nor that PB=PB'). The sequence number for this packet is the new initial sequence number for this subflow. The ACK sequence number is not relevant as the ACK flag is not set. The packet carries an MP_JOIN option and it carries the token TA. It also carries a random nonce generated by Host B called RB.

3. Host C receives the SYN+MP_JOIN packet from Host B, and it alters it in the following way. It changes the source address to IPC and the destination address to IPA. It sends the modified packet to Host A, impersonating Host B.

4. Host A receives the SYN+MP_JOIN message and it replies with a SYN/ACK+MP_JOIN message. The packet has source address IPA and destination address IPC, as well as all the other needed parameters. In particular, Host A computes a valid HMAC and places it in the MP_JOIN option.

5. Host C receives the SYN/ACK+MP_JOIN message and it changes the source address to IPC and the destination address to IPB. It sends the modified packet to IPB impersonating Host A.

6. Host B receives the SYN/ACK+MP_JOIN message. Host B verifies the HMAC of the MP_JOIN option and confirms its validity. It replies with an ACK+MP_JOIN packet. The packet has source address IPB and destination address IPC, as well as all the other needed parameters. The returned MP_JOIN option contains a valid HMAC computed by Host B.

7. Host C receives the ACK+MP_JOIN message from B and it alters it in the following way. It changes the source address to IPC and the destination address to IPA. It sends the modified packet to Host A impersonating Host B.

8. Host A receives the ACK+MP_JOIN message and creates the new subflow.

   At this point the attacker has managed to place itself as a MitM for one subflow for the existing MPTCP session. It should be noted that there still exists the subflow between

address IPA and IPB that does not flow through the attacker, so the attacker has not completely intercepted all the packets in the communication (yet). If the attacker wishes to completely intercept the MPTCP session it can do the following additional step.

9. Host C sends two TCP RST messages. One TCP RST packet is sent to Host B, with source address IPA and destination address IPB and source and destination ports PA and PB, respectively. The other TCP RST message is sent to Host A, with source address IPB and destination address IPA and source and destination ports PB and PA, respectively. Both RST messages must contain a valid sequence number. Note that figuring the sequence numbers to be used here for subflow A is the same difficulty as being able to send the initial ADD_ADDR option with valid Sequence number and ACK value. If there are more subflows, then the attacker needs to find the Sequence Number and ACK for each subflow.

   At this point the attacker has managed to fully hijack the MPTCP session.

Information required by the attacker to perform the described attack:

In order to perform this attack the attacker needs to guess or know the following pieces of information: (The attacker need this information for one of the subflows belonging to the MPTCP session.)

- the four-tuple {Client-side IP Address, Client-side Port, Server-side Address, Servcer-side Port} that identifies the target TCP connection
- a valid sequence number for the subflow
- a valid ACK sequence number for the subflow
- a valid address identifier for IPC

TCP connections are uniquely identified by the four-tuple {Source Address, Source Port, Destination Address, Destination Port}. Thus, in order to attack a TCP connection, an attacker needs to know or be able to guess each of the values in that four-tuple. Assuming the two peers of the target TCP connection are known, the Source Address and the Destination Address can be assumed to be known.

We note that in order to be able to successfully perform this attack, the attacker needs to be able to send packets with a forged source address. This means that the attacker cannot be located in a network where techniques like ingress filtering

[RFC2827] or source address validation [RFC7039] are deployed. However, ingress filtering is not as widely implemented as one would expect, and hence cannot be relied upon as a mitigation for this kind of attack.

Assuming the attacker knows the application protocol for which the TCP connection is being employed, the server-side port can also be assumed to be known. Finally, the client-side port will generally not be known, and will need to be guessed by the attacker. The chances of an attacker guessing the client-side port will depend on the ephemeral port range employed by the client, and whether the client implements port randomization [RFC6056].

Assuming TCP sequence number randomization is in place (see e.g. [RFC6528]), an attacker would have to blindly guess a valid TCP sequence number. That is,

RCV.NXT =< SEG.SEQ < RCV.NXT+RCV.WND or RCV.NXT =< SEG.SEQ+SEG.LEN-1 < RCV.NXT+RCV.WND

As a result, the chances of an attacker to succeed will depend on the TCP receive window size at the target TCP peer.

We note that automatic TCP buffer tuning mechanisms have been become common for popular TCP implementations, and hence very large TCP window sizes of values up to 2 MB could end up being employed by such TCP implementations.

According to [RFC0793], the Acknowledgement Number is considered valid as long as it does not acknowledge the receipt of data that has not yet been sent. That is, the following expression must be true:

SEG.ACK <= SND.NXT

However, for implementations that support [RFC5961], the following (stricter) validation check is enforced:

SND.UNA - SND.MAX.WND <= SEG.ACK <= SND.NXT

Finally, in order for the address identifier to be valid, the only requirement is that it needs to be different than the ones already being used by Host A in that MPTCP session, so a random identifier is likely to work.

Given that a large number of factors affect the chances of an attacker of successfully performing the aforementioned off-path attacks, we provide two general expressions for the expected number of packets the attacker needs to send to succeed in the attack: one

for MPTCP implementations that support [RFC5961], and another for MPTCP implementations that do not.

Implementations that do not support RFC 5961

Packets = (2^32/(RCV_WND)) * 2 * EPH_PORT_SIZE/2 * 1/MSS

Where the new :

Packets:
Maximum number of packets required to successfully perform an off-path (blind) attack.

RCV_WND:
TCP receive window size (RCV.WND) at the target node.

EPH_PORT_SIZE:
Number of ports comprising the ephemeral port range at the "client" system.

MSS:
Maximum Segment Size, assuming the attacker will send full segments to maximize the chances to get a hit.

Notes:
The value "2^32" represents the size of the TCP sequence number space.
The value "2" accounts for 2 different ACK numbers (separated by 2^31) that should be employed to make sure the ACK number is valid.

The following table contains some sample results for the number of required packets, based on different values of RCV_WND and EPH_PORT_SIZE for a MSS of 1500 bytes.

Ports \ Win	16 KB	128 KB	256 KB	2048 KB
4000	699050	87381	43690	5461
10000	1747626	218453	109226	13653
50000	8738133	1092266	546133	68266

Table 1: Max. Number of Packets for Successful Attack

Implementations that do support RFC 5961

Packets = (2^32/(RCV_WND)) * (2^32/(2 * SND_MAX_WND)) * EPH_PORT_SIZE/2 * 1/MSS

Where:

Packets:
Maximum number of packets required to successfully perform an off-path (blind) attack.

RCV_WND:
TCP receive window size (RCV.WND) at the target MPTCP endpoint.

SND_MAX_WND:
Maximum TCP send window size ever employed by the target MPTCP end-point (SND.MAX.WND).

EPH_PORT_SIZE:
Number of ports comprising the ephemeral port range at the "client" system.

Notes:
The value "2^32" represents the size of the TCP sequence number space.
The parameter "SND_MAX_WND" is specified in [RFC5961].
The value "2*SND_MAX_WND" results from the expresion "SND.NXT - SND.UNA - MAX.SND.WND", assuming that, for connections that perform bulk data transfers, "SND.NXT - SND.UNA == MAX.SND.WND". If an attacker targets a TCP endpoint that is not actively transferring data, "2 * SND_MAX_WND" would become "SND_MAX_WND" (and hence a successful attack would typically require more packets).

The following table contains some sample results for the number of required packets, based on different values of RCV_WND, SND_MAX_WND, and EPH_PORT_SIZE. For these implementations, only a limited number of sample results are provided, just as an indication of how [RFC5961] increases the difficulty of performing these attacks.

Ports \ Win	16 KB	128 KB	256 KB	2048 KB
4000	45812984490	715827882	178956970	2796202

Table 2: Max. Number of Packets for Successful Attack

Note:

In the aforementioned table, all values are computed with RCV_WND equal to SND_MAX_WND.

2.1. Possible security enhancements to prevent this attack

1. To include the token of the connection in the ADD_ADDR option. This would make it harder for the attacker to launch the attack, since he needs to either eavesdrop the token (so this can no longer be a blind attack) or to guess it, but a random 32 bit number is not so easy to guess. However, this would imply that any eavesdropper that is able to see the token, would be able to launch this attack. This solution then increases the vulnerability window against eavesdroppers from the initial 3-way handshake for the MPTCP session to any exchange of the ADD_ADDR messages.

2. To include the HMAC of the address contained in the ADD_ADDR option. The key used for the HMAC is the concatenation of the key of the receiver and the key of the sender (in the same way they are used for generating the HMAC of the MP_JOIN message). This makes it much more secure, since it requires the attacker to have both keys (either by eavesdropping it in the first exchange or by guessing it). Because this solution relies on the key used in the MPTCP session, the protection of this solution would increase if new key generation methods are defined for MPTCP (e.g. using SSL keys as has been proposed).

3. To include the destination address of the SYN packet in the HMAC of the MP_JOIN message. As the attacker requires to change the destination address to perform the described attack, protecting it would prevent the attack. It wouldn't allow hosts behind NATs to be reached by an address in the ADD_ADDR option, even with static NAT bindings (like a web server at home).

Out of the options described above, option 2 is reccommended as it achieves a higher security level while preserving the required functionality (i.e. NAT compatibility).

3. DoS attack on MP_JOIN

Summary of the attack:

Type of attack: MPTCP Denial-of-Service attack, preventing the hosts from creating new subflows.

Type of attacker: Off-path, active attacker

Description:

As currently specified, the initial SYN+MP_JOIN message of the 3-way handshake for additional subflows creates state in the host receiving the message. This is because the SYN+MP_JOIN contains the 32-bit token that allows the receiver to identify the MPTCP-session and the 32-bit random nonce, used in the HMAC calculation. As this information is not resent in the third ACK of the 3-way handshake, a host must create state upon reception of a SYN+MP_JOIN.

Assume that there exists an MPTCP-session between host A and host B, with token Ta and Tb. An attacker, sending a SYN+MP_JOIN to host B, with the valid token Tb, will trigger the creation of state on host B. The number of these half-open connections a host can store per MPTCP-session is limited by a certain number, and it is implementation-dependent. The attacker can simply exhaust this limit by sending multiple SYN+MP_JOINs with different 5-tuples. The (possibly forged) source address of the attack packets will typically correspond to an address that is not in use, or else the SYN/ACK sent by Host B would elicit a RST from the impersonated node, thus removing the corresponding state at Host B. Further discussion of traditional SYN-flood attacks and common mitigations can be found in [RFC4987]

This effectively prevents the host A from sending any more SYN+MP_JOINs to host B, as the number of acceptable half-open connections per MPTCP-session on host B has been exhausted.

The attacker needs to know the token Tb in order to perform the described attack. This can be achieved if it is a Partial-time On-path eavesdropper , observing the 3-way handshake of the establishment of an additional subflow between host A and host B. If the attacker is never on-path, it has to guess the 32-bit token.

3.1. Possible security enhancements to prevent this attack

The third packet of the 3-way handshake could be extended to contain also the 32-bit token and the random nonce that has been sent in the SYN+MP_JOIN. Further, host B will have to generate its own random nonce in a reproducible fashion (e.g., a Hash of the 5-tuple + initial sequence-number + local secret). This will allow host B to reply to a SYN+MP_JOIN without having to create state. Upon the reception of the third ACK, host B can then verify the correctness of the HMAC and create the state.

4. SYN flooding amplification

Summary of the attack:

Type of attack: The attacker can use the SYN+MP_JOIN messages to amplify the SYN flooding attack.

Type of attacker: Off-path, active attacker

Description:

SYN flooding attacks [RFC4987] use SYN messages to exhaust the server's resources and prevent new TCP connections. A common mitigation is the use of SYN cookies [RFC4987] that allow the stateless processing of the initial SYN message.

With MPTCP, the initial SYN can be processed in a stateless fashion using the aforementioned SYN cookies. However, as we described in the previous section, as currently specified, the SYN+MP_JOIN messages are not processed in a stateless manner. This opens a new attack vector. The attacker can now open a MPTCP session by sending a regular SYN and creating the associated state but then send as many SYN+MP_JOIN messages as supported by the server with different source address source port combinations, consuming server's resources without having to create state in the attacker. This is an amplification attack, where the cost on the attacker side is only the cost of the state associated with the initial SYN while the cost on the server side is the state for the initial SYN plus all the state associated to all the following SYN+MP_JOIN.

4.1. Possible security enhancements to prevent this attack

1. The solution described for the previous DoS attack on MP_JOIN would also prevent this attack.

2. Limiting the number of half open subflows to a low number (e.g. 3 subflows) would also limit the impact of this attack.

5. Eavesdropper in the initial handshake

Summary of the attack

Type of attack: An eavesdropper present in the initial handshake where the keys are exchanged can hijack the MPTCP session at any time in the future.

Type of attacker: a Partial-time On-path eavesdropper

Description:

In this case, the attacker is present along the path when the initial 3-way handshake takes place, and therefore is able to learn the keys

used in the MPTCP session. This allows the attacker to move away from the MPTCP session path and still be able to hijack the MPTCP session in the future. This vulnerability was readily identified at the moment of the design of the MPTCP security solution and the threat was considered acceptable.

5.1. Possible security enhancements to prevent this attack

There are many techniques that can be used to prevent this attack and each of them represents different tradeoffs. At this point, we limit ourselves to enumerate them and provide useful pointers.

1. Use of hash-chains. The use of hash chains for MPTCP has been explored in [hash-chains]

2. Use of SSL keys for MPTCP security as described in [I-D.paasch-mptcp-ssl]

3. Use of Cryptographically-Generated Addresses (CGAs) for MPTCP security. CGAs [RFC3972] have been used in the past to secure multi addressed protocols like SHIM6 [RFC5533].

4. Use of TCPCrypt [I-D.bittau-tcp-crypt]

5. Use DNSSEC. DNSSEC has been proposed to secure the Mobile IP protocol [dnssec]

6. SYN/JOIN attack

Summary of the attack

Type of attack: An attacker that can intercept the SYN/JOIN message can alter the source address being added.

Type of attacker: a Partial-time On-path eavesdropper

Description:

The attacker is present along the path when the SYN/JOIN exchange takes place, and this allows the attacker to add any new address it wants to by simply substituting the source address of the SYN/JOIN packet for one it chooses. This vulnerability was readily identified at the moment of the design of the MPTCP security solution and the threat was considered acceptable.

6.1. Possible security enhancements to prevent this attack

It should be noted that this vulnerability is fundamental due to the NAT support requirement. In other words, MPTCP must work through NATs in order to be deployable in the current Internet. NAT behavior is unfortunately indistinguishable from this attack. It is impossible to secure the source address, since doing so would prevent MPTCP to work though NATs. This basically implies that the solution cannot rely on securing the address. A more promising approach would be then to look into securing the payload exchanged, limiting the impact that the attack would have in the communication (e.g. TCPCrypt [I-D.bittau-tcp-crypt] or similar).

7. Reccomendations

Current MPTCP specification [RFC6824] is experimental. There is an ongoing effort to move it to Standards track. We believe that the work on MPTCP security should follow two threads:

- The work on improving MPTCP security so that is enough to become a Standard Track document.
- The work on analyzing possible additional security enhancements to provide a more secure version of MPTCP.

We will expand on these two next.

MPTCP security for a Standard Track specification.

We believe that in order for MPTCP to progress to Standard Track, the ADD_ADDR attack must be addressed. We believe that the solution that should be adopted in order to deal with this attack is to include an HMAC to the ADD_ADDR message (with the address being added used as input to the HMAC, as well as the key). This would make the ADD_ADDR message as secure as the JOIN message. In addition, this implies that if we implement a more secure way to create the key used in the MPTCP connection, then the security of both the MP_JOIN and the ADD_ADDR messages is automatically improved (since both use the same key in the HMAC).

We believe that this is enough for MPTCP to progress as a Standard track document, because the security level is similar to single path TCP, as results from our previous analysis. Moreover, the security level achieved with these changes is exactly the same as other Standard Track documents. In particular, this would be the same security level as SCTP with dynamic addresses as defined in [RFC5061]. The Security Consideration section of RFC5061 (which is a Standard Track document) reads:

> The addition and or deletion of an IP address to an existing association does provide an additional mechanism by which existing associations can be hijacked. Therefore, this document requires the use of the authentication mechanism defined in [RFC4895] to limit the ability of an attacker to hijack an association. Hijacking an association by using the addition and deletion of an IP address is only possible for an attacker who is able to intercept the initial two packets of the association setup when the SCTP-AUTH extension is used without pre-shared keys. If such a threat is considered a possibility, then the [RFC4895] extension must be used with a preconfigured shared endpoint pair key to mitigate this threat.

This is the same security level that would be achieved by MPTCP plus the ADD_ADDR security measure reccommended.

7.1. Security enhancements for MPTCP

We also believe that is worthwhile exploring alternatives to secure MPTCP. As we identified earlier, the problem is securing JOIN messages is fundamentally incompatible with NAT support, so it is likely that a solution to this problem involves the protection of the data itself. Exploring the integration of MPTCP and approaches like TCPCrypt [I-D.bittau-tcp-crypt] or integration with SSL seem promising venues.

8. Security considerations

This whole document is about security considerations for MPTCP.

9. IANA Considerations

There are no IANA considerations in this memo.

10. Acknowledgments

We would like to thank Mark Handley for his comments on the attacks and countermeasures discussed in this document. thanks to Alissa Copper, Phil Eardley, Yoshifumi Nishida, Barry Leiba, Stephen Farrell, Stefan Winter for the comments and review. Marcelo Bagnulo, Christophe Paasch, Oliver Bonaventure and Costin Raiciu are partially funded by the EU Trilogy 2 project.

11. References

11.1. Normative References

[RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, September 1981.

[RFC3972] Aura, T., "Cryptographically Generated Addresses (CGA)", RFC 3972, March 2005.

[RFC5961] Ramaiah, A., Stewart, R., and M. Dalal, "Improving TCP's Robustness to Blind In-Window Attacks", RFC 5961, August 2010.

[RFC6056] Larsen, M. and F. Gont, "Recommendations for Transport-Protocol Port Randomization", BCP 156, RFC 6056, January 2011.

[RFC6528] Gont, F. and S. Bellovin, "Defending against Sequence Number Attacks", RFC 6528, February 2012.

[RFC5061] Stewart, R., Xie, Q., Tuexen, M., Maruyama, S., and M. Kozuka, "Stream Control Transmission Protocol (SCTP) Dynamic Address Reconfiguration", RFC 5061, September 2007.

[RFC4895] Tuexen, M., Stewart, R., Lei, P., and E. Rescorla, "Authenticated Chunks for the Stream Control Transmission Protocol (SCTP)", RFC 4895, August 2007.

[RFC6824] Ford, A., Raiciu, C., Handley, M., and O. Bonaventure, "TCP Extensions for Multipath Operation with Multiple Addresses", RFC 6824, January 2013.

11.2. Informative References

[RFC6181] Bagnulo, M., "Threat Analysis for TCP Extensions for Multipath Operation with Multiple Addresses", RFC 6181, March 2011.

[RFC5533] Nordmark, E. and M. Bagnulo, "Shim6: Level 3 Multihoming Shim Protocol for IPv6", RFC 5533, June 2009.

[RFC4960] Stewart, R., "Stream Control Transmission Protocol", RFC 4960, September 2007.

[RFC6275] Perkins, C., Johnson, D., and J. Arkko, "Mobility Support in IPv6", RFC 6275, July 2011.

[RFC2827] Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", BCP 38, RFC 2827, May 2000.

[RFC7039] Wu, J., Bi, J., Bagnulo, M., Baker, F., and C. Vogt, "Source Address Validation Improvement (SAVI) Framework", RFC 7039, October 2013.

[RFC4987] Eddy, W., "TCP SYN Flooding Attacks and Common Mitigations", RFC 4987, August 2007.

[I-D.paasch-mptcp-ssl] Paasch, C. and O. Bonaventure, "Securing the MultiPath TCP handshake with external keys", draft-paasch-mptcp-ssl-00 (work in progress), October 2012.

[I-D.bittau-tcp-crypt] Bittau, A., Boneh, D., Hamburg, M., Handley, M., Mazieres, D., and Q. Slack, "Cryptographic protection of TCP Streams (tcpcrypt)", draft-bittau-tcp-crypt-04 (work in progress), February 2014.

[hash-chains] Diez, J., Bagnulo, M., Valera, F., and I. Vidal, "Security for multipath TCP: a constructive approach", International Journal of Internet Protocol Technology 6, 2011.

[dnssec] Kukec, A., Bagnulo, M., Ayaz, S., Bauer, C., and W. Eddy, "OAM-DNSSEC: Route Optimization for Aeronautical Mobility using DNSSEC", 4th ACM International Workshop on Mobility in the Evolving Internet Architecture MobiArch 2009, 2009.

Authors' Addresses

Marcelo Bagnulo
Universidad Carlos III de Madrid
Av. Universidad 30
Leganes, Madrid 28911
SPAIN

Phone: 34 91 6249500
Email: marcelo@it.uc3m.es
URI: http://www.it.uc3m.es

Christoph Paasch
UCLouvain
Place Sainte Barbe, 2
Louvain-la-Neuve, 1348
Belgium

Email: christoph.paasch@uclouvain.be

Fernando Gont
SI6 Networks / UTN-FRH
Evaristo Carriego 2644
Haedo, Provincia de Buenos Aires 1706
Argentina

Phone: +54 11 4650 8472
Email: fgont@si6networks.com
URI: http://www.si6networks.com

Olivier Bonaventure
UCLouvain
Place Sainte Barbe, 2
Louvain-la-Neuve, 1348
Belgium

Email: olivier.bonaventure@uclouvain.be

Costin Raiciu
Universitatea Politehnica Bucuresti
Splaiul Independentei 313a
Bucuresti
Romania

Email: costin.raiciu@cs.pub.ro

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:

- 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.

- 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.

- 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:

- It must be backwards-compatible with current, regular TCP, to increase its chances of deployment.

- 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].

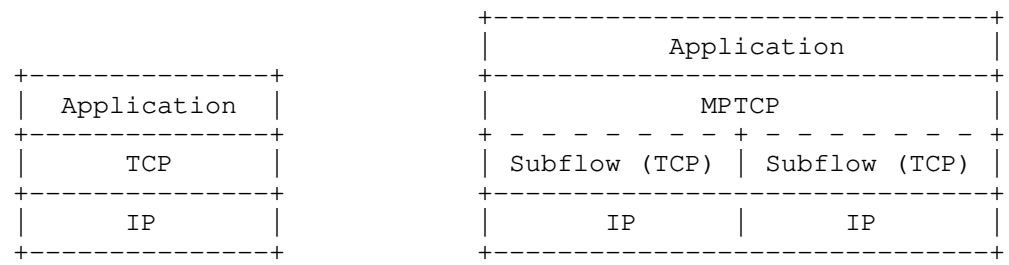


Figure 1: Comparison of Standard TCP and MPTCP Protocol Stacks

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.

- 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.

- 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.

- 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.

- 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.

- 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.

- MPTCP adds connection-level sequence numbers to allow the reassembly of segments arriving on multiple subflows with differing network delays.

- Subflows are terminated as regular TCP connections, with a four-way FIN handshake. The MPTCP connection is terminated by a connection-level FIN.

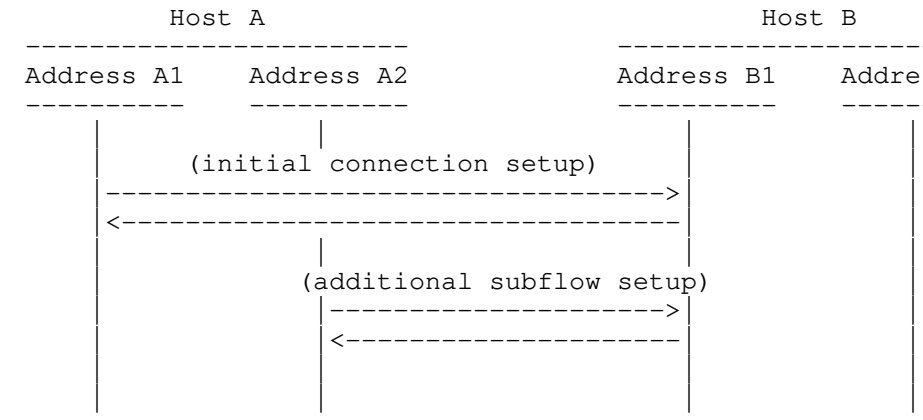


Figure 2: Example MPTCP Usage Scenario

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]
```

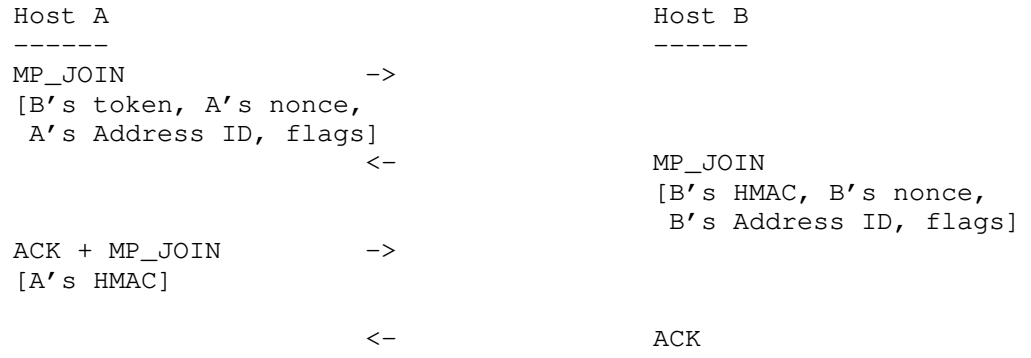
```
Host A                                  Host B (sending first)
------                                  ------
MP_CAPABLE                 ->
[flags]
                           <-           MP_CAPABLE
                                        [B's key, flags]
ACK + MP_CAPABLE           ->
[A's key, B's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



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.

```
   Host A                                  Host B
   ------                                  ------
   ACK + MP_FASTCLOSE          ->
   [B's key]

   [RST on all other subflows] ->

                               <-          [RST on all subflows]
```

```
   Host A                                  Host B
   ------                                  ------
   RST + MP_FASTCLOSE          ->
   [B's key] [on all subflows]

                               <-          [RST on all subflows]
```

2.7. Notable Features

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

- 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.

- 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.

- 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.

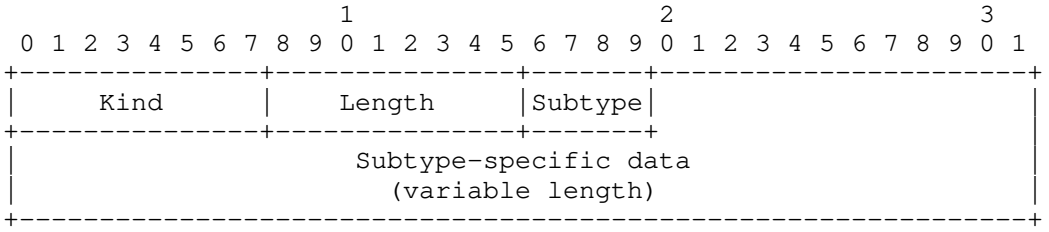


Figure 3: MPTCP Option Format

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.

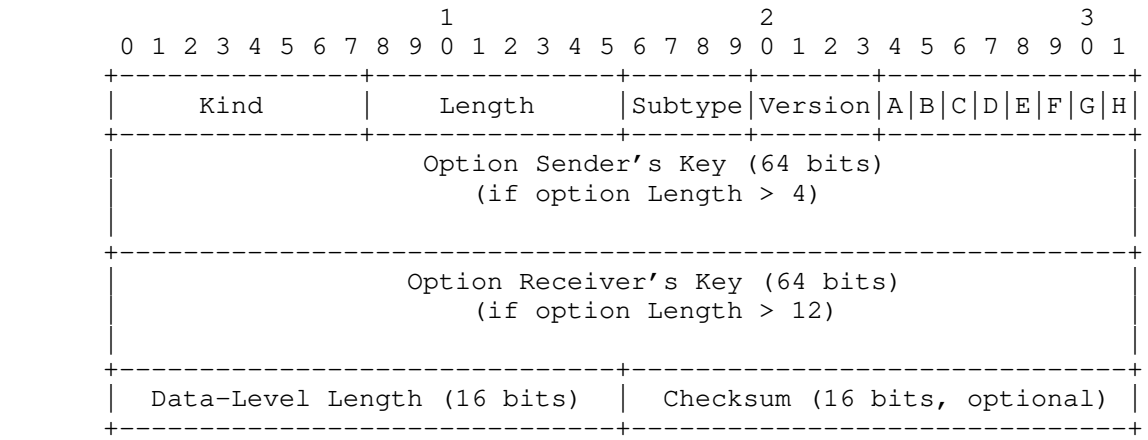


Figure 4: Multipath Capable (MP_CAPABLE) Option

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.

- SYN (A->B): only the first four octets (Length = 4).
- SYN/ACK (B->A): B's Key for this connection (Length = 12).
- ACK (no data) (A->B): A's Key followed by B's Key (Length = 20).
- 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.

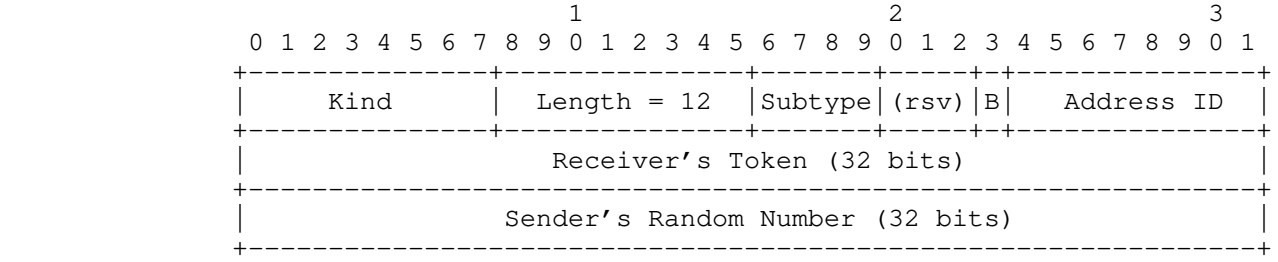


Figure 5: Join Connection (MP_JOIN) Option (for Initial SYN)

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_TCPRST

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.

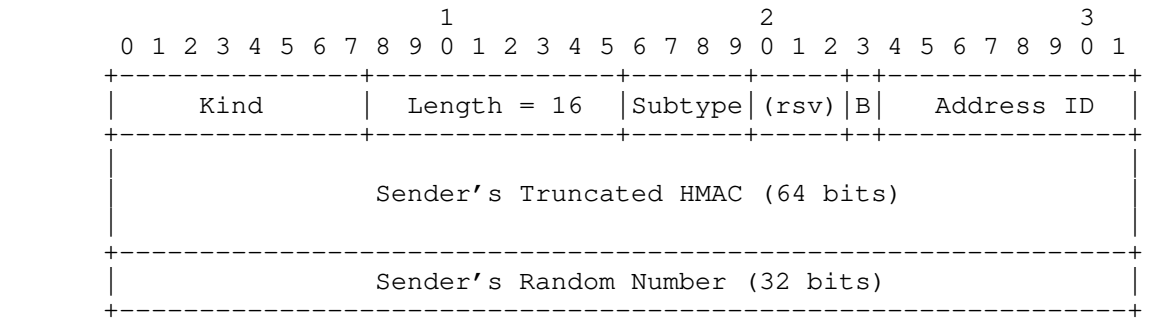


Figure 6: Join Connection (MP_JOIN) Option (for Responding SYN/ACK)

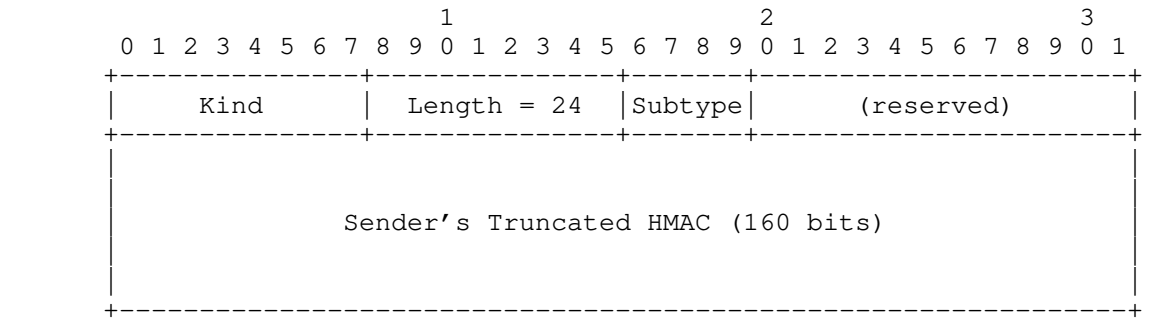
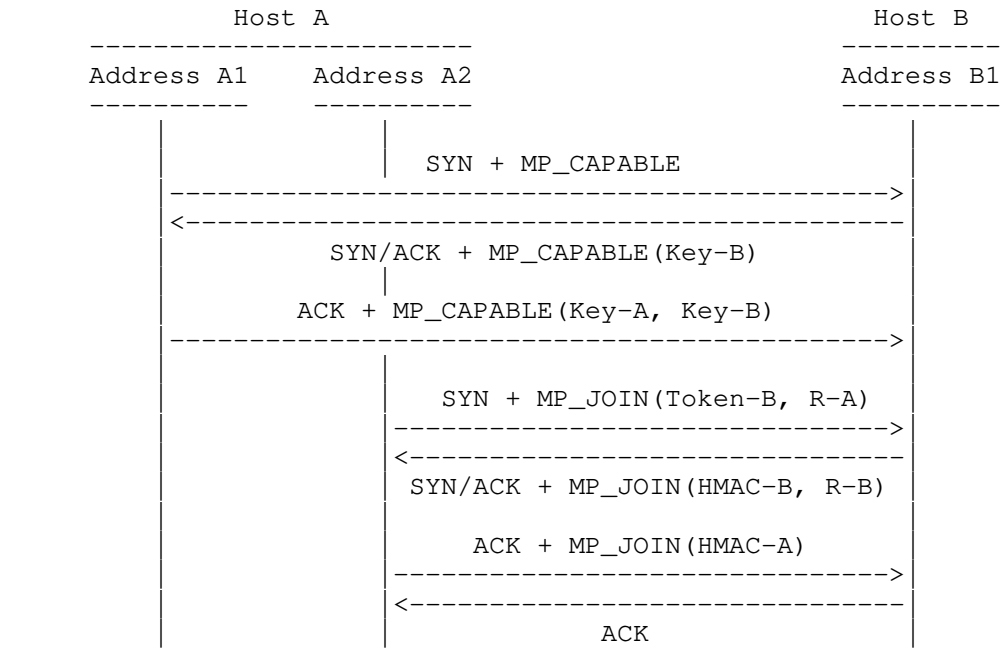


Figure 7: Join Connection (MP_JOIN) Option (for Third ACK)

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.

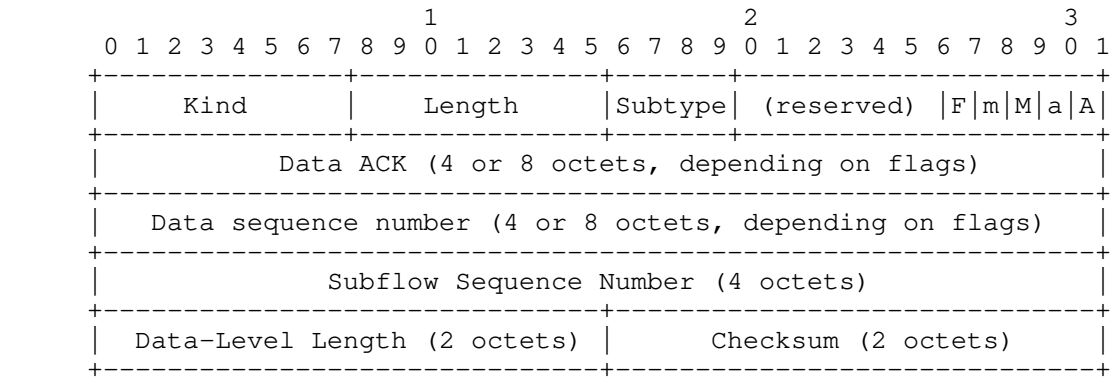


Figure 9: Data Sequence Signal (DSS) Option

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.

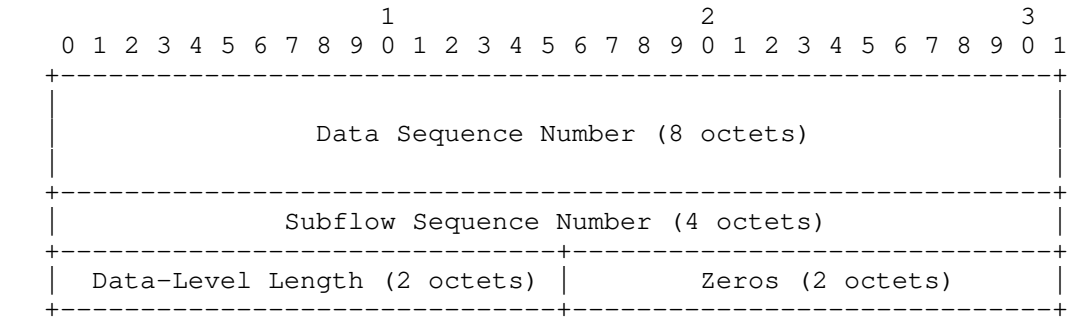


Figure 10: Pseudo-Header for DSS Checksum

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_TCPRST (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 <= 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.

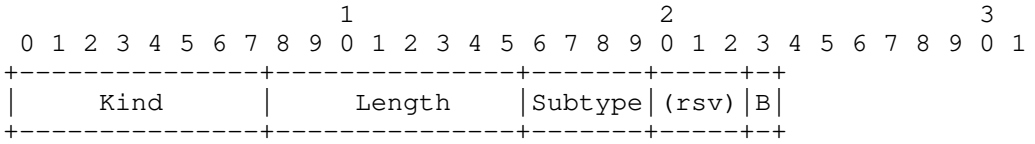


Figure 11: Change Subflow Priority (MP_PRIO) Option

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:

- 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.

- 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.

- 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.

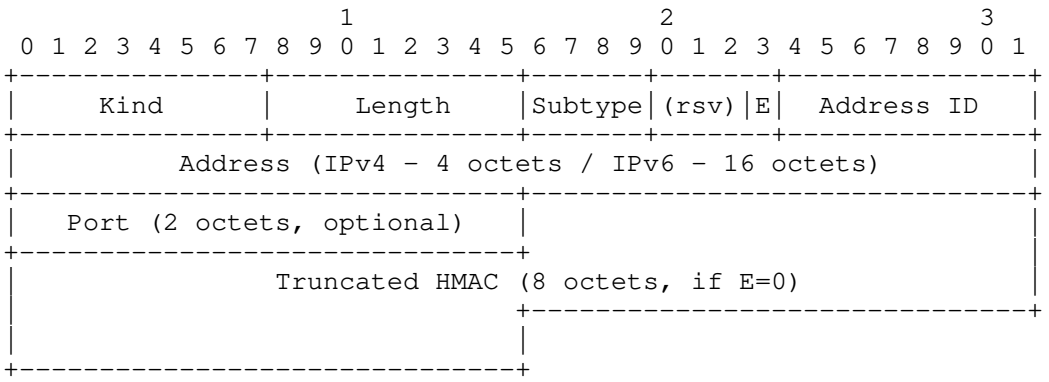


Figure 12: Add Address (ADD_ADDR) Option

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.

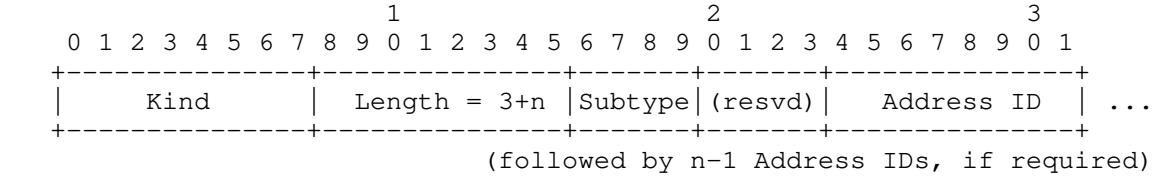


Figure 13: Remove Address (REMOVE_ADDR) Option

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.

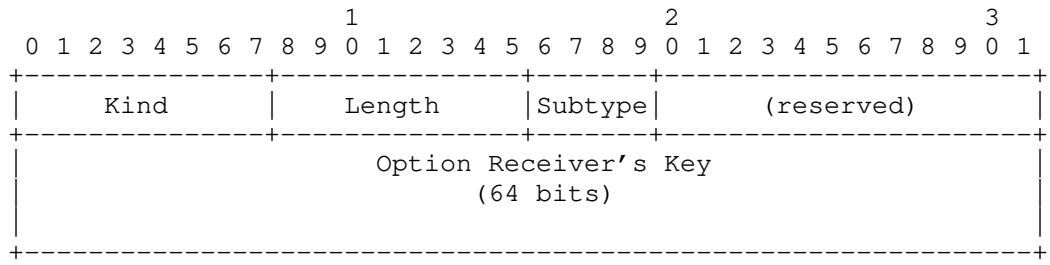


Figure 14: Fast Close (MP_FASTCLOSE) Option

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

- 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.

- 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:

- 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).

- 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.

- 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.

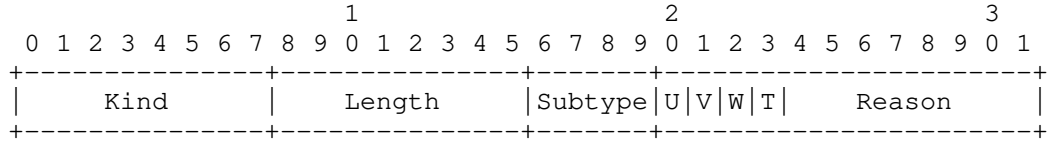


Figure 15: TCP RST Reason (MP_TCPRST) Option

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:

- 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.

- 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.

- Lack of resources (code 0x02). This code indicates that the sending host does not have enough resources to support the terminated subflow.

- Administratively prohibited (code 0x03). This code indicates that the requested subflow is prohibited by the policies of the sending host.

- 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.

- 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.

- 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_TCPRST 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.

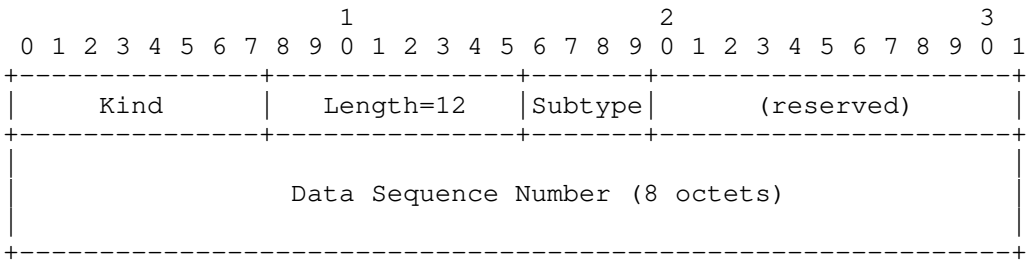


Figure 16: Fallback (MP_FAIL) Option

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 receive 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:

- 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)

- DSN out of window (during normal operation): drop the data, do not send Data ACKs

- 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:

- Provide a mechanism to confirm that the parties in a subflow handshake are the same as in the original connection setup.

- Provide verification that the peer can receive traffic at a new address before using it as part of a connection.

- Provide replay protection, i.e., ensure that a request to add/remove a subflow is 'fresh'.

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:

- 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.

- 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 [TCPLO]), 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.

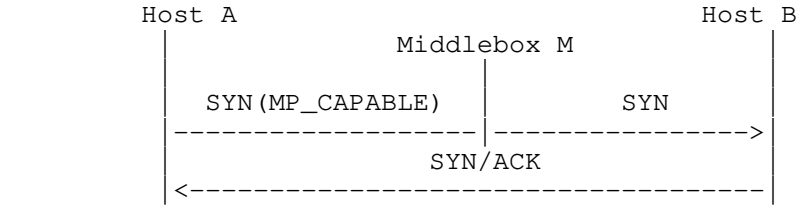
MPTCP uses a single new TCP option "Kind", and all message types are defined by "subtype" values (see Section 8). This should reduce the chances of only some types of MPTCP options being passed, and instead the key differing characteristics are different paths, and the presence of the SYN flag.

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.

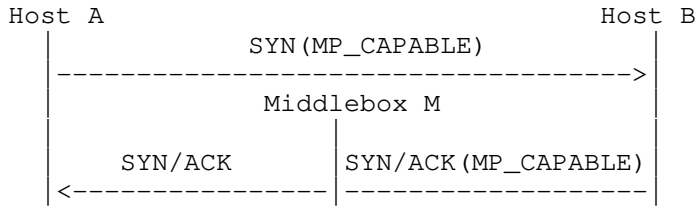
If a middlebox strips options but otherwise passes the packets unchanged, MPTCP will behave safely. If an MP_CAPABLE option is

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.



a) MP_CAPABLE option stripped on outgoing path



b) MP_CAPABLE option stripped on return path

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:

- 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.

- 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).

- 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.

- 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.

- 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 mutliple 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).

- 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:

- 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.

- 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.

- 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 Stiemerling, 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>.

[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:

- 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.

- 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.

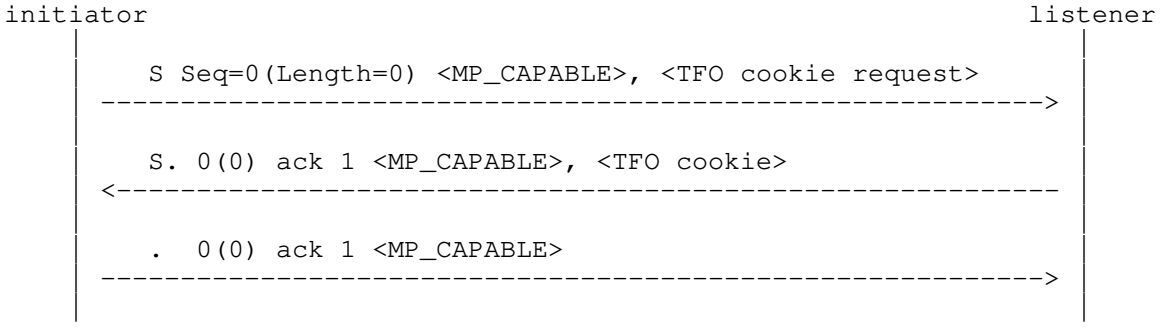


Figure 18: Cookie request - sequence number and length are annotated as Seq(Length) and used hereafter in the figures.

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.

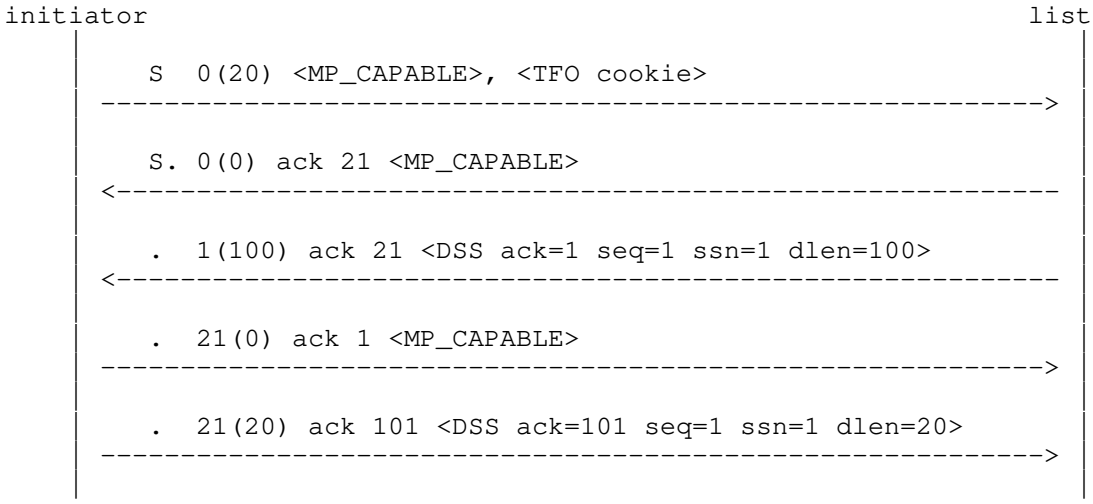


Figure 19: The listener supports TFO

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 then 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.

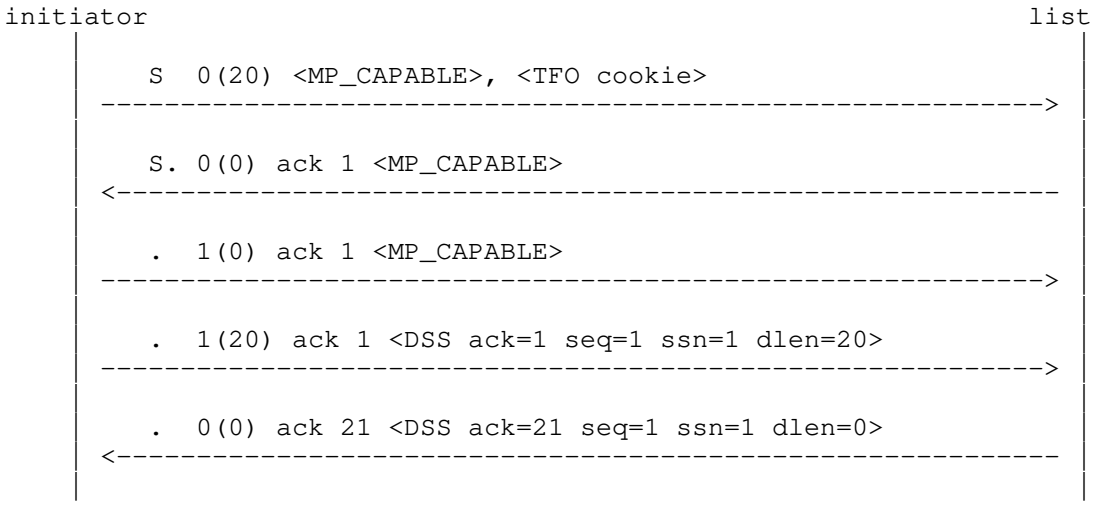


Figure 20: The listener does not support TFO

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.

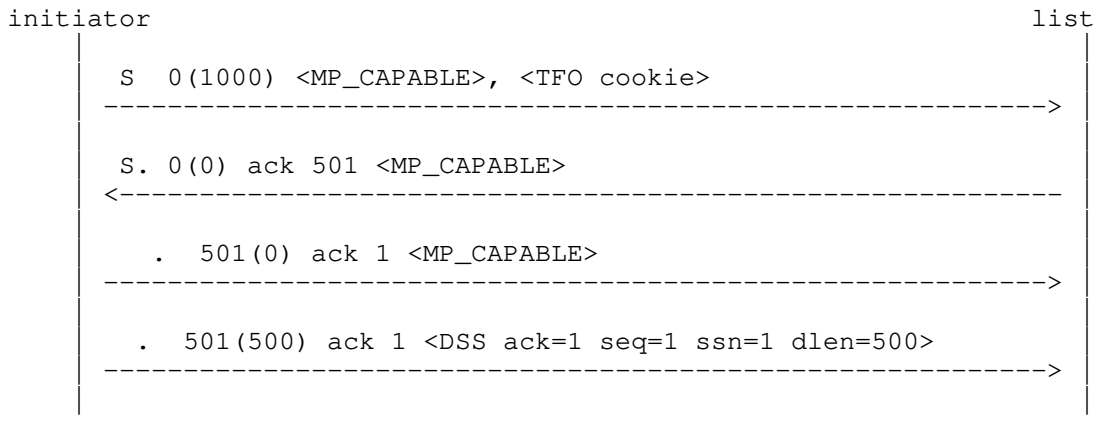


Figure 21: Partial data acknowledgement

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.

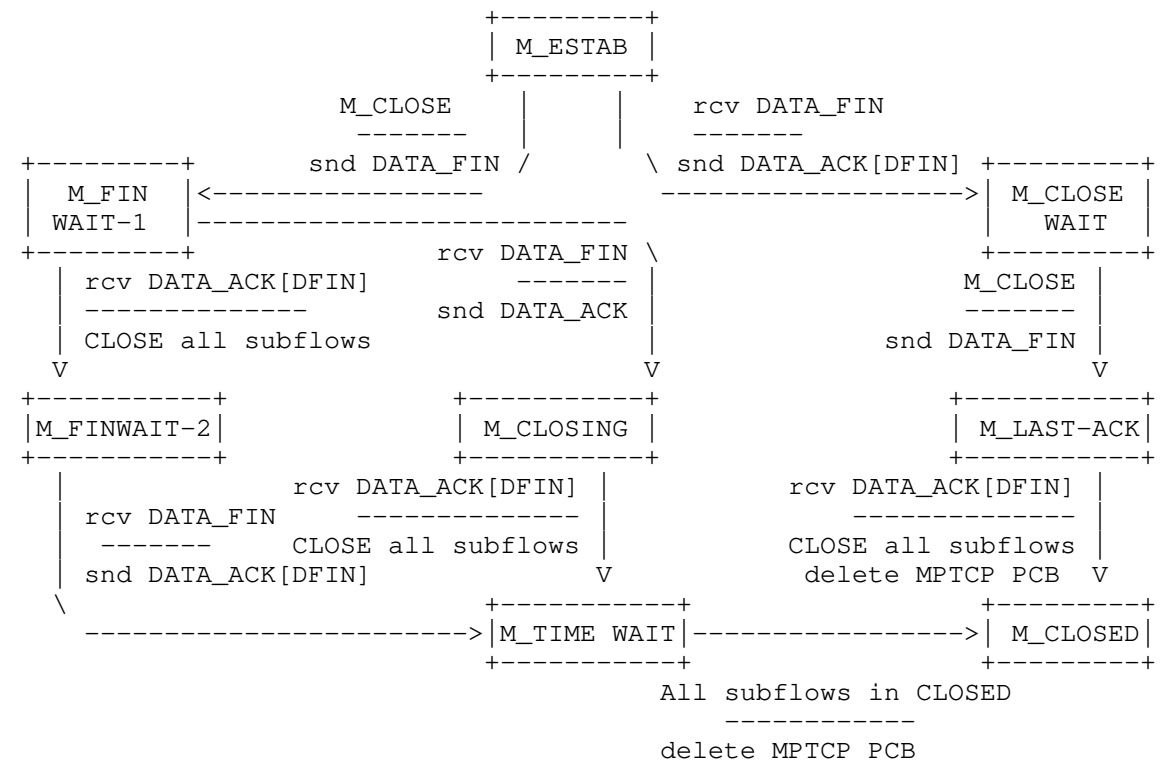


Figure 22: Finite State Machine for Connection Closure

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.

- 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].

- 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.

- 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.

- 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.

- In the Flags field of MP_CAPABLE, H now indicates the use of HMAC-SHA256 (rather than HMAC-SHA1).

- Connection initiation also defines the procedure for version negotiation, for implementations that support both v0 (RFC6824) and v1 (this document).

- 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.

- A new subflow-level option exists to signal reasons for sending a RST on a subflow (MP_TCPRST Section 3.6), which can help an implementation decide whether to attempt later re-connection.

- 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.

- 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.

- 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.

- 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.

- 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

Multipath TCP
INTERNET-DRAFT
Intended Status: Standard Track
Expires: January 5, 2015

Ramin Khalili
T-Labs/TU-Berlin
Nicolas Gast
Miroslav Popovic
Jean-Yves Le Boudec
EPFL-LCA2
July 4, 2014

Opportunistic Linked-Increases Congestion Control Algorithm for MPTCP
draft-khalili-mptcp-congestion-control-05

Abstract

This document describes the mechanism of OLIA, the "Opportunistic Linked Increases Algorithm". OLIA is a congestion control algorithm for MPTCP. The current congestion control algorithm of MPTCP, LIA [4], forces a tradeoff between optimal congestion balancing and responsiveness. OLIA's design departs from this tradeoff and provide these properties simultaneously. Hence, it solves the identified performance problems with LIA while retaining non-flappiness and responsiveness behavior of LIA, as shown by different studies [5, 6, 7, 8]. OLIA is now part of the UCLouvain's MPTCP implementation [9, 11].

Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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 5, 2015.

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html

The list of Internet-Draft Shadow Directories can be accessed at

http://www.ietf.org/shadow.html

Copyright and License Notice

Copyright (c) 2014 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 Requirements Language 4
1.2 Terminology 4
2 The set of best paths, paths with maximum windows, and collected paths 5
3 Opportunistic Linked-Increases Algorithm 6
4 Practical considerations 8
5 Discussion 9
6 References 9
6.1 Normative References 9
6.2 Informative References 9
Authors' Addresses 11

1 Introduction

The current MPTCP implementation uses a congestion control algorithm called LIA, the "Linked-Increases" algorithm [4]. The design of LIA forces a tradeoff between optimal congestion balancing and responsiveness. Hence, to provide good responsiveness, LIA's current implementation must depart from optimal congestion balancing. This leads to important performance issues (refer to [5] and [6]): (i) in some scenarios upgrading TCP users to MPTCP results in a significant drop in the aggregate throughput in the network without any benefit for anybody; and (ii) MPTCP users can be excessively aggressive toward TCP users.

In this draft, we introduce OLIA, the "opportunistic linked increases algorithm", as an alternative to LIA. Contrary to LIA, OLIA's design is not based on a trade-off between responsiveness and optimal congestion balancing; it can provide both simultaneously [5].

Similarly to LIA, OLIA couples the additive increases and uses unmodified TCP behavior in the case of a loss. The difference between LIA and OLIA is in the increase part. OLIA is an adaptation of TCP New Reno to support multiple paths. OLIA's increase part, Equation (1), has two terms:

- The first term is an adaptation of the increase term of Kelly and Voice's algorithm [10]. This term is essential to provide optimal resource pooling.

- The second term guarantees responsiveness and non-flappiness of OLIA. By measuring the number of transmitted bytes since the last loss, it reacts to events within the current window and adapts to changes faster than the first term.

By adapting the window increases as a function of RTTs, OLIA also compensates for different RTTs. As OLIA is rooted on the optimal algorithm of [10], it provides fairness and optimal congestion balancing. Because of the second term, it is responsive and non-flappy.

OLIA is implemented in the Linux kernel and is now a part of UCLouvain's MPTCP implementation [9, 11]. In [5], we study the performance of MPTCP with OLIA over a testbed, by simulations and by theoretical analysis. We prove theoretically that OLIA is Pareto-optimal and that it satisfies the design goals of MPTCP described in [4]. Hence, it can provide optimal congestion balancing and fairness in the network. Our measurements and simulations indicate that MPTCP with OLIA is as responsive and non-flappy as MPTCP with LIA and that it solves the identified problems with LIA. Recent studies show that

MPTCP with OLIA always outperforms MPTCP with LIA and is very responsive to the changes in the environment [7, 8].

The rest of the document provides a description of OLIA. For an analysis of its performance, we refer to [5, 7, 8].

1.1 Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [1].

1.2 Terminology

Regular TCP: The standard version of TCP that operates between a single pair of IP addresses and ports [2].

Multipath TCP: A modified version of the regular TCP that allows a user to spread its traffic across multiple paths.

MPTCP: The proposal for multipath TCP specified in [3].

LIA: The Linked-Increases Algorithm of MPTCP (the congestion control of MPTCP) [4].

OLIA: The Opportunistic Linked-Increases Algorithm for MPTCP proposed in [5].

all_paths: The set of all the paths established by the MPTCP connection.

best_paths: The set of paths in all_paths that are presumably the best paths for the MPTCP connection.

max_w_paths: The set of paths in all_paths with largest congestion windows.

collected_paths: The set of paths in all_paths that are presumably the best paths but do not have largest congestion window (i.e. the paths of best_paths that are not in max_w_paths).

w_r: The congestion windows on a path r.

rtt_r: The Round-Trip Time on a path r.

MSS_r: The Maximum Segment Size that specifies the largest amount of data can be transmitted by a TCP packet on the path r.

2 The set of best paths, paths with maximum windows, and collected paths

A MPTCP connection has access to one or more paths. Let all_paths be the set of all the paths established by the MPTCP connection and r be one of these paths.

We denote by l_{1r} the number of bytes that were successfully transmitted over path r between the last two losses seen on r, and by l_{2r} the number of bytes that are successfully transmitted over r after the last loss. We denote by l_r=max{l_{1r},l_{2r}} the smoothed estimation of number of bytes transmitted on path r between last two losses.

l_{1r} and l_{2r} can be measured by using information that is already available to a regular TCP user:

- For each ACK on r: l_{2r} <- l_{2r} + (number of bytes that are acknowledged by ACK),

- For each loss on r: l_{1r} <- l_{2r} and l_{2r} <- 0.

l_{1r} and l_{2r} are initially set to zero when the connection is established. If no losses have been observed on r until now, then l_{1r}=0 and l_{2r} is the total number of bytes transmitted on r.

Let rtt_r be the round-trip time observed on path r (e.g. the smoothed round-trip time used by regular TCP) and w_r be the congestion windows on path r. We denote by best_paths the set of paths r in all_paths that have the maximum value of l_r*l_r/rtt_r, by max_w_paths the set of paths r in all_paths with largest w_r, and by collected_paths the set of best paths that do not have maximum window size, i.e.:

- best_paths = { r | r = arg max_{p in all_paths} (l_p*l_p/rtt_p) }

- max_w_paths = { r | r = arg max_{p in all_paths} (w_p) }

- collected_paths = { r | r in best_paths and not in max_w_paths }.

where arg max is the argument of maximum, the set of points of the given argument for which the given function is maximum. arg max is applied over all paths p in all_paths.

best_paths represents the set of paths that are presumably the best paths (in term of transmission rate) for the user: 1/l_r can be considered as an estimate of byte loss probability on path r, and hence the rate that path r can provide to a TCP user can be estimated by (2*l_r)^{1/2}/rtt_r. A collected path is a path that is presumably

good but is not fully used. The set collected_paths can be empty.

Note that l_{1r}, l_{2r}, l_r, rtt_r, w_r, best_paths, max_w_paths and collected_paths are all functions of time.

3 Opportunistic Linked-Increases Algorithm

In this section, we introduce OLIA. OLIA is a window-based congestion-control algorithm. It couples the increase of congestion windows and uses unmodified TCP behavior in the case of a loss. OLIA is an alternative for LIA, the current congestion control of MPTCP.

The algorithm only applies to the increase part of the congestion avoidance phase. The fast retransmit and fast recovery algorithms, as well as the multiplicative decrease of the congestion avoidance phase, are the same as in TCP [2]. We also use a similar slow start algorithm as in TCP, with the modification that we set the ssthresh (slow start threshold) to be 1 MSS if multiple paths are established. In the case of a single path flow, we use the same minimum ssthresh as in TCP (i.e. 2 MSS). The purpose of this modification is to avoid transmitting unnecessary traffic over congested paths when multiple paths are available to a user.

For a path r, we denote by w_r the congestion windows on this path (also called subflow). We denote by MSS_r be the maximum segment size on the path r. We assume that w_r is maintained in bytes.

Our proposed "Opportunistic Linked-Increases Algorithm" (OLIA) must:

- For each ACK on path r, increase w_r by

$$\left(\frac{w_r/rtt_r^2}{(SUM_{p \text{ in all_paths}} (w_p/rtt_p))^2} + \frac{alpha_r}{w_r} \right) \quad (1)$$

multiplied by MSS_r * bytes_acked.

The summation in the denominator of the first term is over all the paths p in all_paths. Recall that w_p and rtt_p denote the window size and the round trip time of a path p.

alpha_r is calculated as follows:

- If r is in collected_paths, then

$$alpha_r = \frac{1/number_of_paths}{|collected_paths|}$$

- If r is in max_w_paths and if collected_paths is not empty, then

$$alpha_r = - \frac{1/number_of_paths}{|max_w_paths|}$$

- Otherwise, alpha_r=0.

|collected_paths| and |max_w_paths| are the number of paths in collected_paths and in max_w_paths. Note that the sum of all alpha_r is equal to 0.

The first term in (1) is an adaptation of Kelly and Voice's increase term [10] and provides the optimal resource pooling (Kelly and Voice's algorithm is based on scalable TCP; the first term in (1) is a TCP compatible version of their algorithm that compensates also for different RTTs). The second term, with alpha_r, guarantees responsiveness and non-flappiness of our algorithm.

By definition of alpha_r, if all the best paths have the largest window size, then alpha_r=0 for any r. This is because we already use the capacity available to the user by using all the best path.

If there is any best path with a small window size, i.e. if collected_paths is not empty, then alpha_r is positive for all r in collected_paths and negative for all r in max_w_paths. Hence, our algorithm increases windows faster on the paths that are presumably best but that have small windows. The increase will be slower on the paths with maximum windows. In this case, OLIA re-forwards traffic from fully used paths (i.e. paths in max_w_paths) to paths that have free capacity available to the users (i.e. paths in collected_paths).

In [4], three goals have been proposed for the design of a practical multipath congestion control algorithm : (1) Improve throughput: a multipath TCP user should perform at least as well as a TCP user that uses the best path available to it. (2) Do no harm: a multipath TCP user should never take up more capacity from any of its paths than a TCP user. And (3) balance congestion: a multipath TCP algorithm should balance congestion in the network, subject to meeting the first two goals.

Our theoretical results in [5] show that OLIA fully satisfies these three goals. LIA, however, fails to fully satisfy the goal (3) as discussed in [5] and [6]. Moreover, in [5], we show through measurements and by simulation that our algorithm is as responsive and non-flappy as LIA and that it can solve the identified problems with LIA. In [7], Chen et al. study how MPTCP with LIA and OLIA performs in the wild with a common wireless environment, namely using

both WiFi and Cellular simultaneously. Their results show that MPTCP with OLIA is very responsive to the changes in the environment and always outperforms MPTCP with LIA. Furthermore, using Experimental Design, Paasch et al. [8] show that MPTCP with OLIA satisfy the design goal of MPTCP in a very wide range of scenarios and always outperform MPTCP with LIA.

4 Practical considerations

Calculation of alpha requires performing costly floating point operation whenever an ACK received over path r. In practice, however, we can integrate calculation of alpha and Equation (1) together. Our algorithm can be therefore simplified as the following.

For each ACK on the path r:

- If r is in collected_paths, increase w_r by

$$\frac{w_r/rtt_r^2}{(SUM_p\ (w_p/rtt_p))^2} + \frac{1}{w_r * number_of_paths * |collected_paths|} \quad (2)$$

multiplied by MSS_r * bytes_acked.

- If r is in max_w_paths and if collected_paths is not empty, increase w_r by

$$\frac{w_r/rtt_r^2}{(SUM_p\ (w_p/rtt_p))^2} - \frac{1}{w_r * number_of_paths * |max_w_paths|} \quad (3)$$

multiplied by MSS_r * bytes_acked.

- Otherwise, increase w_r by

$$\frac{(w_r/rtt_r^2)}{(SUM_p\ (w_p/rtt_p))^2} \quad (4)$$

multiplied by MSS_r * bytes_acked.

The summation in the dominator of the first term of equations (2), (3), and (4) is over the path p in all_paths. To compute the increase, we only need to determine the sets collected_paths and max_w_paths when an ACK is received on the path r. We can further simplify the algorithm by updating the sets collected_paths and max_w_paths only once per round-trip time or whenever there is a drop

on the path.

We can see from above that in some cases (i.e. when r is max_w_paths and collected_paths is not empty) the increase could be negative. This is a property of our algorithm as in this case OLIA re-forwards traffic from paths in max_w_paths to paths in collected_paths. It is easy to show that using our algorithm, w_r >= 1 for any path r.

5 Discussion

Our results in [5] show that the identified problems with current MPTCP implementation are not due to the nature of a window-based multipath protocol, but rather to the design of LIA. OLIA shows that it is possible to build an alternative to LIA that mitigates these problems and that is as responsive and non-flappy as LIA.

Our proposed algorithm can provide similar resource pooling as Kelly and Voice’s algorithm [10] and fully satisfies the design goals of MPTCP described in [4]. Hence, it can provide optimal congestion balancing and fairness in the network [5]. Moreover, it is as responsive and non-flappy as LIA and outperforms LIA in realistic scenarios such as wireless networks (refer to [5, 7, 8]).

We therefore believe that mptcp working group should revisit the congestion control part of MPTCP and that an alternative algorithm, such as OLIA, should be considered.

6 References

6.1 Normative References

[2] Allman, M., Paxson, V., and E. Blanton, "TCP Congestion Control", RFC 5681, September 2009.

[3] Ford, A., Raiciu, C., Greenhalgh, A., and M. Handley, "Architectural Guidelines for Multipath TCP Development", RFC 6182, March 2011.

6.2 Informative References

[4] Raiciu, C., Handley, M., and D. Wischik, "Coupled Congestion Control for Multipath Transport Protocols", RFC 6356, October 2011.

[5] R. Khalili, N. Gast, M. Popovic, U. Upadhyay, and J.-Y. Le Boudec. "MPTCP is not Pareto-optimality: Performance issues and a possible solution", ACM CoNEXT 2012 (The

extended version is published at IEEE/ACM Transaction of Networking, volume 5, issue 25, October 2013).

[6] R. Khalili, N. Gast, M. Popovic, and J.-Y. Le Boudec. "Performance Issues with MPTCP", draft-khalili-mptcp-performance-issues-06.

[7] Chen Y.-C, Y.-S. Lim, R. J. Gibbens, E. M. Nahum, R. Khalili, and D. Towsley, "A Measurement-based Study of Multipath TCP Performance over Wireless Networks." ACM IMC 2013.

[8] C. Paasch, R. Khalili, and O. Bonaventure, "On the Benefits of Applying Experimental Design to Improve Multipath TCP." ACM CoNEXT 2013.

[9] UCL, Louvain-la-Neuve, Belgium, "MultiPath TCP-Linux kernel implementation," 2013 [Online]. Available: http://mptcp.info.ucl.ac.be/.

[10] F. Kelly, and T. Voice, "Stability of end-to-end algorithms for joint routing and rate control", ACM SIGCOMM CCR vol. 35 num. 2, pp. 5-12, 2005.

[11] O. Bonaventure, and C. Paasch, "Experience with Multipath TCP", draft-bonaventure-mptcp-experience-00

Authors' Addresses

Ramin Khalili
T-Labs/TU-Berlin
TEL 18, Ernst-Reuter-Platz 7
10587 Berlin
Germany

Phone: +49 16 0972 28074
EMail: ramin@net.t-labs.tu-berlin.de

Nicolas Gast
EPFL IC ISC LCA2
Station 14
CH-1015 Lausanne
Switzerland

Phone: +41 21 693 1254
EMail: nicolas.gast@epfl.ch

Miroslav Popovic
EPFL IC ISC LCA2
Station 14
CH-1015 Lausanne
Switzerland

Phone: +41 21 693 6466
EMail: miroslav.popovic@epfl.ch

Jean-Yves Le Boudec
EPFL IC ISC LCA2
Station 14
CH-1015 Lausanne
Switzerland

Phone: +41 21 693 6631
EMail: jean-yves.leboudec@epfl.ch

Multipath TCP
INTERNET-DRAFT
Intended Status: Standard Track
Expires: January 5, 2015

Ramin Khalili
T-Labs/TU-Berlin
Nicolas Gast
Miroslav Popovic
Jean-Yves Le Boudec
EPFL-LCA2
July 4, 2014

Performance Issues with MPTCP
draft-khalili-mptcp-performance-issues-06

Abstract

We show, by measurements over a testbed and by mathematical analysis, that the current MPTCP suffers from two problems: (P1) Upgrading some TCP users to MPTCP can reduce the throughput of others without any benefit to the upgraded users; and (P2) MPTCP users can be excessively aggressive towards TCP users. We attribute these problems to the "Linked Increases" Algorithm (LIA) of MPTCP [4], and more specifically, to an excessive amount of traffic transmitted over congested paths. Our results show that these problems are important and can be mitigated. We believe that the design of the congestion control of MPTCP should be improved.

Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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 5, 2015.

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

Copyright and License Notice

Copyright (c) 2014 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 Requirements Language 4
1.2 Terminology 4
2 MPTCP's LIA 5
3 Testbed Setup 5
4 Scenario A: MPTCP can penalize regular TCP users 6
5 Scenario B: MPTCP can penalize other MPTCP users 8
6 Scenario C: MPTCP is excessively aggressive towards TCP users . 10
7 Can the suboptimality of MPTCP with LIA be fixed? 12
8 Conclusion 15
9 References 15
9.1 Normative References 15
9.2 Informative References 15
Authors' Addresses 17

1 Introduction

Regular TCP uses a window-based congestion-control mechanism to adjust the transmission rate of users [2]. It always provides a Pareto-optimal allocation of resources: it is impossible to increase the throughput of one user without decreasing the throughput of another or without increasing the congestion cost [5]. It also guarantees a fair allocation of bandwidth among the users but favors the connections with lower RTTs [6].

Various mechanisms have been used to build a multipath transport protocol compatible with the regular TCP. Inspired by utility maximization frameworks, [7, 8] propose a family of algorithms. These algorithms tend to use only the best paths available to users and are optimal in static settings where paths have similar RTTs. In practice, however, they suffer from several problems [9]. First, they might fail to quickly detect free capacity as they do not probe paths with high loss probabilities sufficiently. Second, they exhibit flappiness: When there are multiple good paths available to a user, the user will randomly flip its traffic between these paths. This is not desirable, specifically, when the achieved rate depends on RTTs, as with regular TCP.

Because of the issues just mentioned, the congestion control part of MPTCP does not follow the algorithms in [7, 8]. Instead, it follows an ad-hoc design based on the following goals [4]. (1) Improve throughput: a multipath TCP user should perform at least as well as a TCP user that uses the best path available to it. (2) Do no harm: a multipath TCP user should never take up more capacity from any of its paths than a regular TCP user. And (3) balance congestion: a multipath TCP algorithm should balance congestion in the network, subject to meeting the first two goals.

MPTCP compensates for different RTTs and solves many problems of multipath transport [10, 11]: It can effectively use the available bandwidth, it improves throughput and fairness compared to independent regular TCP flows in many scenarios, and it solves the flappiness problem.

We show, however, by measurements over our testbed and mathematical analysis, that MPTCP still suffers from the following problems:
(P1) Upgrading some regular TCP users to MPTCP can reduce the throughput of other users without any benefit to the upgraded users. This is a symptom of non-Pareto optimality.
(P2) MPTCP users can be excessively aggressive towards TCP users.
We attribute these problems to the "Linked Increases" Algorithm (LIA) of MPTCP [4] and specially to an excessive amount of traffic transmitted over congested paths.

These problems indicate that LIA fails to fully satisfy its design goals, especially goal number 3. The design of LIA forces a trade off between optimal resource pooling and responsiveness, it cannot provide both at the same time. Hence, to provide good responsiveness, LIA's current implementation must depart from Pareto-optimality, which leads to problems (P1) and (P2).

This document provides a number of examples and scenarios (Sections 4 to 6) in which MPTCP with LIA exhibits problems (P1) and (P2). Our results show that the identified problems with LIA are important. Moreover, we show in Section 7 that these problems are not due to the nature of a window-based multipath protocol, but rather to the design of LIA; it is possible to build an alternative to LIA that mitigates these problems and is as responsive and non-flappy as LIA. Hence, we believe that the design of the congestion control of MPTCP should be improved.

1.1 Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [1].

1.2 Terminology

Regular TCP: The standard version of TCP that operates between a single pair of IP addresses and ports [2].

Multipath TCP: A modified version of the regular TCP that allows a user to spread its traffic across multiple paths.

MPTCP: The proposal for multipath TCP specified in [3].

LIA: The "Linked Increases" Algorithm of MPTCP (the congestion control of MPTCP) [4].

OLIA: The Opportunistic "Linked Increases" Algorithm for MPTCP proposed in [12].

RTT: The Round-Trip Time seen by a user.

MSS: The Maximum Segment Size that specifies the largest amount of data can be transmitted by a TCP packet.

AP: Access Point.

2 MPTCP's LIA

The design of the congestion control algorithm of MPTCP has been described in details by Mark Handly et al. at the 77th IETF meeting in Anaheim [13].

The actual implementation of the congestion control algorithm, called "Linked Increases" Algorithm (LIA), increases the window size w_r by a quantity proportional to w_r /w_tot for each ACK received on path r (see [13, slide 3]). w_tot is the sum of the congestion windows over all paths available to the user.

LIA is one of a family of multipath congestion control algorithms that can be indexed by a parameter 0<phi<2 (see [13, slide 9]). These algorithms results in transmitting a traffic proportional to (1/p_r)^{1/phi} a path that has a probability loss p_r:

* If phi is very close to zero, then only path with smallest loss rate will be used (e.g. if p2>p1then w2=0 and w1>0). It correspond to the fully coupled algorithm of [7] and is flappy [9].
* phi=2 corresponds to having uncoupled TCP flows on each of the path. It is very responsive and non-flappy, but does not balance congestion in the network.
* phi=1 correspond to LIA and is described in RFC6356 [4]. It provides a compromise between congestion balancing and responsiveness.

3 Testbed Setup

To investigate the behavior of MPTCP in practice, three testbed topologies are created representing scenarios in Sections 4, 5, and 6. We use server-client PCs that run MPTCP enabled Linux kernels. We use MPTCP for the Linux kernel 3.0.0 released in February 2012. In all our scenarios, laptop PCs are used as routers. We install "Click Modular Router" software [14] on the routers to implement topologies with different characteristics. Iperf is used to create multiple connections.

In our scenarios, we are able to implement links with configurable bandwidth and delay. We are also able to set the parameters of the RED queues following the structure in [15]. For a 10 Mbps link, we set the packet loss probability equal to 0 up to a queue size of 25 Maximum Segment Size (MSS). Then it grows linearly to the value 0.1 at 50 MSS. It again increases linearly up to 1 for 100 MSS. The parameters are proportionally adapted when the link capacity changes.

To verify that the problems observed are caused by the congestion-control algorithm of MPTCP and not by some unknown problems in our testbed, we perform a mathematical analysis of MPTCP. This analysis

is based on the fix point analysis of MPTCP. As we will see, our mathematical results confirm our measurement results. The details of these mathematical analyses are available in [12].

4 Scenario A: MPTCP can penalize regular TCP users

Consider a network with two types of users. There are N1 users of type1, each with a high-speed private connection. These users access different files on a media streaming server. The server has a network connection with capacity limit of N1C1 Mbps. Type1 users can activate a second connection through a shared AP by using MPTCP. There are also N2 users of type2; they are connected to the shared AP. They download their contents from the Internet. The shared AP has a capacity of N2C2 Mbps.

We implement this scenario in a testbed similarly to Figure 1. Within router-PCs R1 and R2, we implement links with capacities N1C1 and N2C2 and RTTs of 150 ms (including queuing delay), modeling the bottleneck at the server side and the shared AP, respectively. High speed connections are used to implement private connections of type1 users.

When type1 users use only their private connection, each type1 user receives a rate of C1 and each type2 user receives a rate of C2. By upgrading type1 users to MPTCP, we observe, through measurement and by mathematical analysis, that the throughput of type2 users significantly decreases. However, type1 users receive the same throughput as before, because the bottleneck for their connections is at the server side. We report the throughput received by users after upgrading type1 users to MPTCP on Table 1 for C1=C2=1 Mbps. For each case, we take 5 measurements. In each case, the confidence intervals are very small (less than 0.01Mbps).

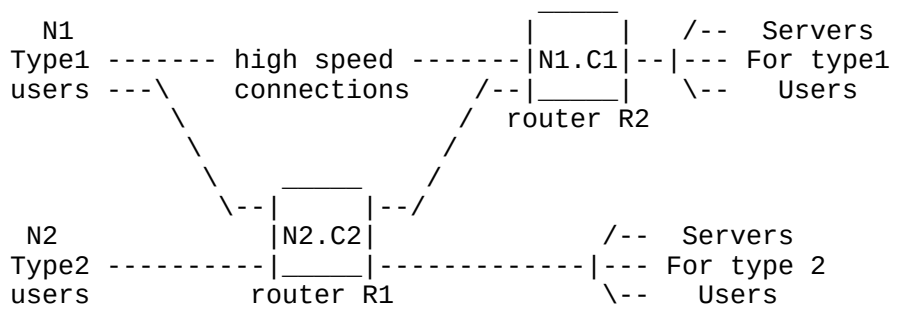


Figure 1: Testbed implementation of Scenario A: router R1 implements the bottleneck at the server side and router R2 implements the shared AP bottleneck.

In [12 Section3.2], we provide a mathematical analysis of MPTCP that confirm our measurements: The predicted rate of type1 users is always 1 and the predicted rate for type2 users is, respectively, 0.74 when N1=N2=10 and 0.50 when N1=30 and N2=10.

		Type1 users are single path (measurement)	Type1 users are multipath: MPTCP (meas.)	Type1 users are multipath: optimal algorithm with p. cost (theory)	Type1 users are multipath: optimal algorithm w/out p. cost (theory)
N1=10	type1	0.98	0.96	1	1
N2=10	type2	0.98	0.70	0.94	1
N1=30	type1	0.98	0.98	1	1
N2=10	type2	0.98	0.44	0.8	1

meas.=measurements, p.=probing, w/out=without, Values are in Mbps.

Table 1: Throughput obtained by type1 and type2 users in Scenario A: upgrading type1 users to MPTCP decrease the throughput of type2 with no benefit for type1 users. The problem is much less critical using optimal algorithm with probing cost.

We observe that MPTCP exhibits problem (P1): upgrading type1 users to MPTCP penalizes type2 users without any gain for type1 users. As the number of type1 users increases, the throughput of type2 users decreases, but the throughput of type1 users does not change as it is limited by the capacity of the streaming server. For N1=N2, type2 users see a decrease of 30%. When N1=3N2, this decrease is 55%.

We compare the performance of MPTCP with two theoretical baselines. They serve as references to see how far from the optimum MPTCP with LIA is. We show in Section 7 that it is possible to replace LIA by an alternative that keeps the same non-flappiness and responsiveness and performs closer to the optimum.

The first baseline is an algorithm that provides theoretical optimal resource pooling in the network (as discussed in [7] and several other theoretical papers). We refer to it as "optimal algorithm without probing cost".

In practice, however, the value of the congestion windows are bounded below by 1 MSS. Hence, with a window-based congestion-control algorithm, a minimum probing traffic of 1 MSS per RTT will be sent

over a path. We introduce a second theoretical baseline, called "optimal algorithm with probing cost"; it provides optimal resource pooling in the network given that a minimum probing traffic of 1 MSS per RTT is sent over each path.

We show the performance of these optimal algorithms in Table 1. Using an optimal algorithm with probing cost, the entire capacity of the shared AP is allocated to type2 users. Hence, all the users in the network receives a throughput of 1 Mbps. By using an optimal algorithm with probing cost, type1 users will send only 1MSS per RTT over the shared AP. Hence, we observe a decrease on the throughput of type2 users. However, the decrease is much less than what we observe using MPTCP. The performance of our proposed alternative to LIA is shown in Section 7, Table 4.

This performance problem of MPTCP can be explained by the fact that LIA does not fully balance congestion. For N1=N2, we observe through measurements that p1=0.009 and p2=0.02 (the probability of losses at routers R1 and R2). For N1=3N2, the value of p1 remains almost the same and p2 increases to p2=0.05. LIA excessively increases congestion on the shared AP, which is not in compliance with goal 3. In [12], we propose an alternative to LIA. Using this algorithm, we have p1=0.009 and p2=0.012 for N1=10 and 0.018 for N1=30. Hence, it is possible to provide a better congestion balancing in the network.

Because of some practical issues, we did not study larger number of connections in our testbed. However, we have mathematical results (using LIA’s loss-throughput formula [12]) as well as simulation results (using a flow-level simulator) that confirm our observation for larger number of connections.

5 Scenario B: MPTCP can penalize other MPTCP users

Consider a multi-homing scenario as follows. We have four Internet Service Providers (ISPs), X, Y, Z, and T. Y is a local ISP in a small city, which connects to the Internet through Z. X, Z, and T are nation-wide service providers and are connected to each other through high speed links. X provides Internet services to users in the city and is a competitor of Y. They have access capacity limits of CX, CY, CZ, and CT. Z and T are hosts of different video streaming servers.

There are two types of users: Blue users download contents from servers in Z and Red users download from servers in ISP T. To increase their reliability, Blue users use multi-homing and are connected to both ISPs X and Y. Red users can connect either only to Y or to both X and Y .

We implement the scenario in a testbed similar to Figure 2. b1 and b2

are the paths available to Blue users. Red users use the path r1, but can upgrade to MPTCP by establishing a second connection through path r2. The measurement results are reported in Table 2 for a setting with CX=27, CT=36, and CY=CZ=100, all in Mbps, where we have 15 Red and 15 Blue users. RTTs are around 150 ms (including queuing delay) over all paths. We also show the performance of theoretically optimal algorithms with and without probing cost.

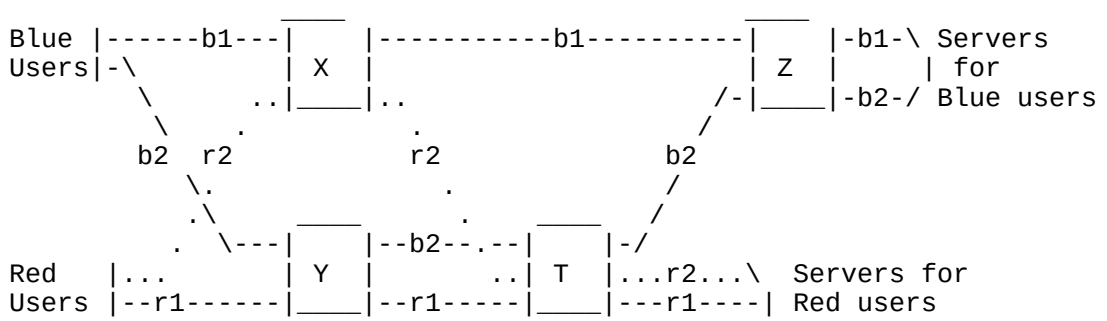


Figure 2. Testbed implementation of Scenario B. Blue users transmit over path b1 and b2. Red users use path r1, but can upgrade to MPTCP by establishing a second connection through path r2.

	Red users are single-path			Red users are multipath		
	MPTCP (meas.)	Blue users use optimal algorithm with p. cost (theory)	Blue users use optimal algorithm w/out p. cost (theory)	MPTCP (meas.)	Blue and Red users use optimal algorithm with p. cost (theory)	Blue and Red users use optimal algorithm w/out p. cost (theory)
Red users	1.5	2.1	2.1	1.4	2.04	2.1
Blue users	2.5	2.1	2.1	2.0	2.04	2.1

meas.=measurements, p.=probing, w/out=without, Values are in Mbps.

Table 2: Throughput received by users before and after upgrading Red users to MPTCP. We have 15 Red and 15 Blue users. By upgrading Red users to MPTCP, the aggregate throughput of users decreases by 13% with no benefit for Red users.

We observe that when Red users only connect to ISP Y, the aggregate throughput of users is close to the cut-set bound, 63 Mbps. However, Blue users get a higher share of the network bandwidth. Now, let’s consider that Red users upgrade to MPTCP by establishing a second

connection through X (showed by pointed-line in Figure 2). Our results in Table 1 show that Red users do not receive any higher throughput. However, the average rate of Blue users drops by 20%, which results in a drop of 13% in aggregate throughput.

In [12 Section 3.3], we also provide a mathematical analysis of MPTCP. Our mathematical results predict that by upgrading the Red user to MPTCP the rate of Blue users will be reduced by 21%. This results in 14% decrease in the aggregate throughput. Hence, our mathematical results confirm our observations from the measurement. Similar behavior is predicted for other values of CX and CT [12 Figure 4a].

Using an optimal algorithm without probing cost, Red users transmit only over path r1 and Blue users split their traffic over paths b1 and b2 to equalize the rate of blue and red users. Upgrading Red users to multipath does not change the allocation. Hence, we observe no decrease in the aggregate throughput and the rate of each user. By using an optimal algorithm with probing cost, the rate of Blue and Red users decreases by 3% when we upgrade the Red users to multipath users since red users are forced to send 1 MSS per RTT over path r2. This decrease is much less than what we observe using MPTCP with LIA. The performance of our proposed alternative to LIA is shown in Section 7, Table 5.

Similarly to Scenario A, the problem can be attributed to the excessive amount of traffic sent over the congested paths. This illustrates that MPTCP fails to balance the congestion in the network.

6 Scenario C: MPTCP is excessively aggressive towards TCP users

Consider a scenario with N1 multipath users, N2 single-path users, and two APs with capacities N1C1 and N2C2 Mbps. Multipath users connect to both APs and they share AP2 with single-path users. The users download their contents from the Internet. This scenario is implemented in a testbed similar to Figure 3.

If the allocation of rates is proportionally fair, multipath users will use AP2 only if C1<C2 and all users will receive the same throughput. When C1>C2, a fair multipath user will not transmit over AP2. However, using MPTCP with LIA, multpath users receive a disproportionately larger share of bandwidth.

We implement the scenario in our testbed and measure the performance of MPTCP with LIA. We report the throughput received by single-path and multipath users in Table 3. We present the results for N2=10 and two values of N1=10 and 30, where C1=C2=1 Mbps. RTTs are around 150

ms (including queuing delay). We also present the performance of optimal (proportionally fair) algorithms.

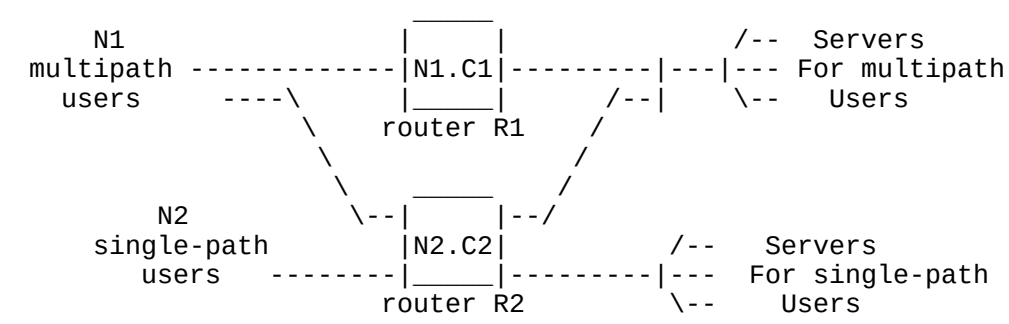


Figure 3: Testbed implementation of Scenario C: routers R1 and R2 implement AP1 and AP2 with capacities N1C1 and N2C2 Mbps.

		multipath users use MPTCP (measurement)	multipath users use optimal algorithm with p. cost (theory)	multipath users use optimal algorithm w/out p. cost (theory)
N1=10	multipath users	1.3	1.04	1
N2=10	single-path users	0.68	0.94	1
N1=30	multipath users	1.19	1.04	1
N2=10	single-path users	0.38	0.8	1

p.=probing, w/out=without, Values are in Mbps.

Table3: Throughput obtained by single-path and multipath users in Scenario C: MPTCP is excessively aggressive toward TCP users and performs far from how an optimal algorithm would perform.

As C1=C2, for any fairness criterion, multipath users should not use AP2. Our results show that, MPTCP users are disproportionately aggressive and exhibit problem (P2). Hence, single-path users receive a much smaller share than they should. For N1=N2, single-path users see a decrease of about 30% in their received throughput compared to a fair allocation. When N1=3N2, this decrease is around 55%.

These measurements are confirmed by our mathematical analysis as shown in [12 Section 3.4]. The predicted rate of type1 users is 1.3

for N1=N2=10 and is 1.17 when N1=30 and N2=10. For type2 users, the predicted rate is 0.7 when N1=N2=10 and 0.48 when N1=30 and N2=10.

An optimal algorithm with probing cost provide a proportional fairness among the users. By using an optimal algorithm with probing cost, single-path users receive a rate less than what a proportionally fair algorithm will provide them. However, as we observe, the problem is much less critical compared to the case we use MPTCP. The performance of our proposed alternative to LIA is shown in Section 7, Table 6.

These results clearly show that MPTCP suffers from fairness issues. The problem occurs because LIA fails to fully satisfy goal 3. As in Scenarios A and B, MPTCP sends an excessive amount of traffic over the congested paths.

7 Can the suboptimality of MPTCP with LIA be fixed?

We have shown in Section 4, 5, and 6 that MPTCP with LIA performs far behind an optimal algorithm. The question is, "is it possible to modify the congestion control algorithm of MPTCP to perform closer to the optimum". To answer this question, we implement a new congestion control algorithm for MPTCP, called Opportunistic "Linked Increases" Algorithm (OLIA). OLIA is described in [13] and its performance is analyzed in [12]. In this section, we show that in Scenarios A, B and C OLIA performs close to an optimal algorithm with probing cost. Moreover, as shown in [12, Sections 4.3 and 6.2], OLIA keeps the same non-flappiness and responsiveness as LIA.

Contrary to LIA, OLIA's design is not based on a trade-off between responsiveness and optimal resource pooling. OLIA couples the additive increases and uses unmodified TCP behavior in the case of a loss. The difference between LIA and OLIA is in the increase part. OLIA's increase part two has terms:

* The first term is an adaptation of the increase term of the optimal algorithm in [7]. This term is essential to provide congestion balancing and fairness.
* The second term guarantees responsiveness and non-flappiness of OLIA. By measuring the number of transmitted bytes since the last loss, it reacts to events within the current window and adapts to changes faster than the first term.

Because OLIA is rooted in the optimal algorithm of [7], it can provide fairness and congestion balancing. Because of the second term, it is responsive and non-flappy.

We implemented OLIA by modifying the congestion control part of the MPTCP implementation based on the Linux Kernel 3.0.0. For conciseness, we do not describe OLIA in this paper and refer to [12]

for details about the algorithm and its implementation.

	Type1 users are single path (measurement)	Type1 users are multipath: MPTCP w. OLIA [with LIA] (measurement)	Type1 users are multipath: optimal algorithm with p. cost (theory)	Type1 users are multipath: optimal algorithm w/out p. cost (theory)
N1=10	type1	0.98	0.98 [0.96]	1	1
N2=10	type2	0.98	0.86 [0.70]	0.94	1
N1=30	type1	0.98	0.98 [0.98]	1	1
N2=10	type2	0.98	0.75 [0.44]	0.8	1

p.=probing, w.=with, w/out=without, Values are in Mbps.

Table 4. Performance of MPTCP with OLIA compared to MPTCP with LIA in scenario A. We show the throughput obtained by users before and after upgrading type1 users to MPTCP. The values in brackets are the values for MPTCP with LIA (taken from table 1).

	Red users are single-path			Red users are multipath		
	Blue users use			Blue and Red users use		
	MPTCP with OLIA [with LIA] (meas.)	optimal algorithm with p. cost (theory)	optimal algorithm w/out p. cost (theory)	MPTCP with OLIA [with LIA] (meas.)	optimal algorithm with p. cost (theory)	optimal algorithm w/out p. cost (theory)
Red users	1.8 [1.5]	2.1	2.1	1.7 [1.4]	2.04	2.1
Blue users	2.2 [2.5]	2.1	2.1	2.2 [2.0]	2.04	2.1

meas.=measurements, p.=probing, w/out=without, Values are in Mbps.

Table 5. Performance of MPTCP with OLIA compared to MPTCP with LIA in scenario B. We show the throughput received by users before and after upgrading Red users to MPTCP. The values in brackets are the values for MPTCP with LIA (taken from Table 2).

		multipath users use MPTCP with OLIA [with LIA] (measurement)	multipath users use optimal algorithm with probing cost (theory)	multipath users use optimal algorithm w/out probing cost (theory)
N1=10	multipath users	1.11 [1.30]	1.04	1
N2=10	single-path users	0.88 [0.68]	0.94	1
N1=10	multipath users	1.1 [1.19]	1.04	1
N2=10	single-path users	0.72 [0.38]	0.8	1

meas.=measurements, w/out=without, Values are in Mbps.

Table 6. Performance of MPTCP with OLIA compared to MPTCP with LIA in scenario C. We show the throughput obtained by single-path and multipath users. The values in brackets are the values for MPTCP with LIA (taken from Table 3).

We study the performance of MPTCP with OLIA through measurements in Scenarios A, B, and C. The results are reported in Tables 4, 5 and 6. We compare the performance of our algorithm with MPTCP with LIA and with optimal algorithms. We observe that in all cases, MPTCP with OLIA provide a significant improvement over MPTCP with LIA. Moreover, it performs close to an optimal algorithm with probing cost.

The results show that it is possible to perform close to an optimal algorithm with probing cost by using a TCP-like algorithm. Moreover, we show in [12, Section 4.3 and Section 6.2] that MPTCP with OLIA is as responsive and non-flappy as MPTCP with LIA. This shows that it is possible to build a practical multi-path congestion control that works close to an optimal algorithm with probing cost.

In [17], Chen et al. study how MPTCP with LIA and OLIA performs in the wild with a common wireless environment, namely using both WiFi and Cellular simultaneously. Their results show that MPTCP with OLIA is very responsive to the changes in the environment and always outperforms MPTCP with LIA. Furthermore, using Experimental Design, Paasch et al. [8] show that MPTCP with OLIA satisfy the design goal of MPTCP in a very wide range of scenarios and always outperform MPTCP with LIA.

8 Conclusion

We have shown that MPTCP with LIA suffers from important performance issues. Moreover, it is possible to build an alternative to LIA that performs close to an optimal algorithm with probing cost while being as responsive and non-flappy as LIA. Hence, we believe that mptcp working group should revisit the congestion control part of MPTCP and that an alternative algorithm, such as OLIA [12], should be considered.

9 References

9.1 Normative References

[2] Allman, M., Paxson, V., and E. Blanton, "TCP Congestion Control", RFC 5681, September 2009.

[3] Ford, A., Raiciu, C., Greenhalgh, A., and M. Handley, "Architectural Guidelines for Multipath TCP Development", RFC 6182, March 2011.

9.2 Informative References

[4] Raiciu, C., Handley, M., and D. Wischik, "Coupled Congestion Control for Multipath Transport Protocols", RFC 6356, October 2011.

[5] F.P. Kelly. Mathematical modelling of the internet. Mathematics unlimited-2001 and beyond.

[6] F.P. Kelly, A.K. Maulloo, and D.K.H. Tan. Rate control for communication networks: shadow prices, proportional fairness and stability. Journal of the Operational Research society, 49, 1998.

[7] Kelly, F. and T. Voice, "Stability of end-to-end algorithms for joint routing and rate control", ACM SIGCOMM CCR vol. 35 num. 2, pp. 5-12, 2005.

[8] H. Han, S. Shakkottai, CV Hollot, R. Srikant, and D. Towsley. "Multi-path tcp: a joint congestion control and routing scheme to exploit path diversity in the internet", Trans. on Net., 14, 2006.

[9] D. Wischik, M. Handly, and C. Raiciu. "Control of multipath tcp and optimization of multipath routing in the internet", NetCOOP, 2009.

[10] D. Wischik, C. Raiciu, A. Greenhalgh, and M. Handly. "Design, implementation and evaluation of congestion control for multipath tcp", Usenix NSDI, 2011.

[11] C. Raiciu, S. Barre, C. Pluntke, A. Greenhalgh, D. Wischik, and M. Handly. "Improving datacenter performance and robustness with multipath tcp", ACM Sigcomm, 2011.

[12] R. Khalili, N. Gast, M. Popovic, U. Upadhyay, and J.-Y. Le Boudec. "MPTCP is not Pareto-optimality: Performance issues and a possible solution", ACM CoNEXT 2012 (The extended version is published at IEEE/ACM Transaction of Networking, volume 5, issue 25, October 2013).

[13] M. Handly, D. Wischik, C. Raiciu, "Coupled Congestion Control for MPTCP", presented at 77th IETF meeting, Anaheim, California. <http://www.ietf.org/proceedings/77/slides/mptcp-9.pdf>.

[14] B. Chen J. Jannotti E. Kohler, R. Morris and M. F. Kaashoek. "The click modular router", Trans. Comput. Syst., 18, August 2000.

[15] S. Floyd and V. Jacobson. "Random early detection gateways for congestion avoidance", Trans. on Net., 1, August, 1993.

[16] R. Khalili, N. Gast, M. Popovic, J.-Y. Le Boudec, "Opportunistic Linked-Increases Congestion Control Algorithm for MPTCP", draft-khalili-mptcp-congestion-control-05.

[17] Chen Y.-C, Y.-S. Lim, R. J. Gibbens, E. M. Nahum, R. Khalili, and D. Towsley, "A Measurement-based Study of Multipath TCP Performance over Wireless Networks." ACM IMC 2013.

[18] C. Paasch, R. Khalili, and O. Bonaventure, "On the Benefits of Applying Experimental Design to Improve Multipath TCP." ACM CoNEXT 2013.

Authors' Addresses

Ramin Khalili
T-Labs/TU-Berlin
TEL 18, Ernst-Reuter-Platz 7
10587 Berlin
Germany

Phone: +49 16 0972 28074
EMail: ramin@net.t-labs.tu-berlin.de

Nicolas Gast
EPFL IC ISC LCA2
Station 14
CH-1015 Lausanne
Switzerland

Phone: +41 21 693 1254
EMail: nicolas.gast@epfl.ch

Miroslav Popovic
EPFL IC ISC LCA2
Station 14
CH-1015 Lausanne
Switzerland

Phone: +41 21 693 6466
EMail: miroslav.popovic@epfl.ch

Jean-Yves Le Boudec
EPFL IC ISC LCA2
Station 14
CH-1015 Lausanne
Switzerland

Phone: +41 21 693 6631
EMail: jean-yves.leboudec@epfl.ch

MPTCP
Internet-Draft
Intended status: Experimental
Expires: August 15, 2014

C. Paasch
O. Bonaventure
UCLouvain

February 11, 2014

A generic control stream for Multipath TCP
draft-paasch-mptcp-control-stream-00

Abstract

Multipath TCP's extensive use of TCP options to exchange control information consumes a significant part of the TCP option space. Extending MPTCP to add more control information into the session becomes cumbersome as the TCP option space is limited to 40 bytes.

This draft introduces a control stream that allows to send control information as part of the subflow's payload. The control stream is mapped into a separate sequence number space and uses a TLV-format for maximum extensibility. It is left to future documents to specify how the TLV-format might be used to exchange control information. As the control stream is sent as part of the subflow's payload, it is not subject to the 40 bytes limitation of the TCP option space.

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 August 15, 2014.

Copyright Notice

Copyright (c) 2014 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 2
2. The control stream 3
2.1. Window considerations 5
3. Connection initiation 5
4. Starting a new subflow 8
5. Examples of key negotiation through the control stream . . . 9
5.1. Reusing the application's TLS key 10
5.2. TLS-like key exchange 10
5.3. Tcpcrypt-like key exchange 10
6. Other example use cases of the control stream 10
6.1. Address signaling 10
7. Security Considerations 10
8. Acknowledgments 11
9. Informative References 11
Authors' Addresses 11

1. Introduction

Multipath TCP [RFC6824] uses the TCP options to exchange control information between the communication hosts. [RFC6824] defines several new TCP options that are used during the three-way handshake and the data transfer. Using options is the standard method to extend the TCP protocol. Unfortunately, the maximum length of the TCP options field is 40 bytes. This severely limit the utilisation of options to exchange control information between communicating hosts. During the three-way handshake, the TCP options space is further limited by the other TCP options that are also included in the SYN and SYN+ACK segments. [RFC6824] did its best to minimize the size of the MP_CAPABLE option inside the SYN and SYN+ACK segments given the presence of other options (typically MSS, timestamp, selective acknowledgements and window scale). However, this has been at the cost of a reduced security due to the utilization of security keys that are too short.

The security requirements for MPTCP ask for a strong authentication of additional subflows [RFC6181]. Given the restriction in the size of the MPTCP options, it seems very difficult to provide strong security by relying only on TCP options that cannot be longer than 40

bytes and are not exchanged reliably. Although a design to overcome these problems would probably be possible, it would add a lot of complexity to the protocol.

Furthermore, today's MPTCP control information is sent in an unreliable manner. This means that control information like MP_PRIO, ADD_ADDR or REMOVE_ADDRESS might get lost, resulting in potential suboptimal performance of Multipath TCP.

In this document, we show that another design is possible. Instead of using only TCP options to exchange control information, we show how it is possible to define a control stream in parallel with the data stream that is used to exchange data over the established subflows. By using this control stream, two MPTCP hosts can reliably exchange control information without being restricted by TCP option space. The control stream can be used to exchange cryptographic material to authenticate the handshake of additional subflows or for any other purpose.

Together with the control stream, we propose to modify the MPTCP-handshake so that no crypto information is exchanged within the TCP options. We suggest to use the control stream instead. Within the control stream, different key-negotiation schemes can be specified (e.g., reuse SSL-key, tcpcrypt-style, Diffie-Hellman,...)

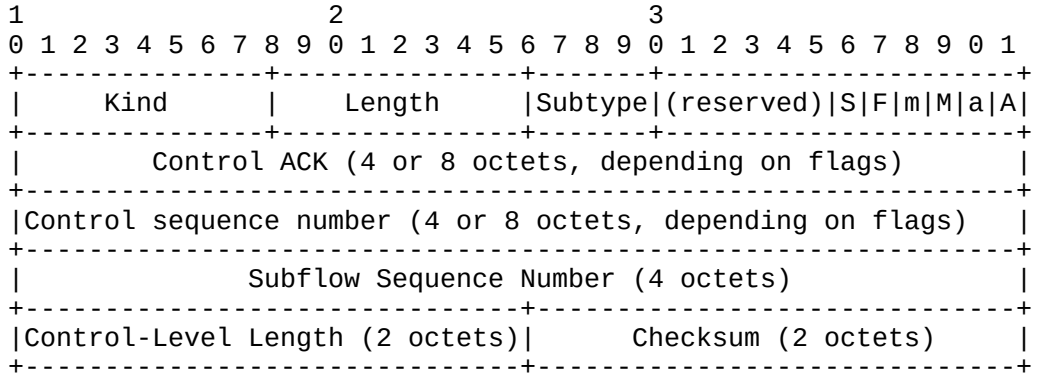
This document is structured as follows. First, we define how the control stream can be used within an MPTCP session. Section 3 presents the modified MPTCP handshake of the initial subflow, while Section 4 specifies the handshake of additional subflows. Section 5 gives some example use-cases for the key negotiation through the control stream. Finally, Section 6 gives another example on how to use the control stream to conduct the MPTCP session.

2. The control stream

In contrast with SCTP [RFC4960], TCP and Multipath TCP [RFC6824] only support one data stream. SCTP uses chunks to allow the communicating hosts to exchange control information of almost unlimited size. As explained earlier, having a control stream in Multipath TCP would enable a reliable delivery of the control information without strict length limitations.

This section defines a control stream that allows to exchange MPTCP control information of arbitrary length besides the regular data stream. The control stream holds data in a TLV-format and thus any type of data can be added to it. Further, the control stream provides a reliable and in-order delivery of the control data.

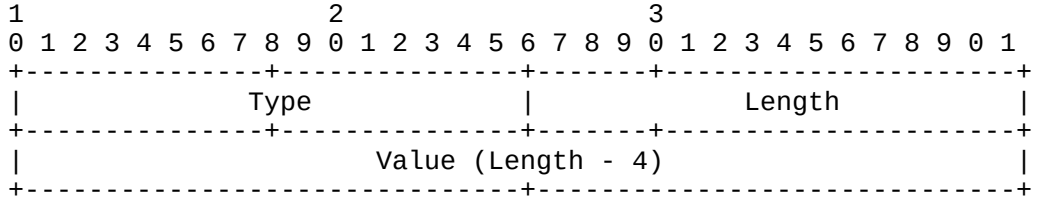
The control stream is sent within the payload of the TCP segments. This ensures a reliable delivery of the TLVs exchanged in the control stream. Further, a separate control-sequence number space is defined for the control stream to ensure in-order delivery of the control stream. The Initial Control stream Sequence Number (ICSN) is the same as the IDSN in the respective directions. A DSS-mapping is used within the TCP option space to signal the control stream sequence numbers as well as a control stream acknowledgement. This DSS-mapping option is the same as the one defined in [RFC6824]. To differentiate the control stream from the data stream, we use the last bit of the 'reserved' field of the MPTCP DSS option. We call this bit the Stream (S) bit. When the DSS option is used to map regular data, this bit is set to 0. When the DSS option is used to map one TLV on the control stream, it is set to 1 (see Figure Figure 1)



The S bit of the 'reserved' field is set to 1 when sending on the control stream.

Figure 1

The control information exchanged in the control stream is encoded by using a TLV format, where the type and length are 16-bit values. This allows for maximum extensibility and to use very long data within the control stream. The format of the TLV option is shown in Figure 2



The TLV option format

Figure 2

2.1. Window considerations

MPTCP uses the receive-window to do flow-control at the receiver. The receive-window within MPTCP is being used at the data sequence level, however any segment sent on a subflow must obey to the last window-announcement received on this particular subflow with respect to the subflow-level sequence number.

The control stream is no different with respect to this last point. The subflow-sequence numbers used for control stream data must fit within the window announced over this specific subflow. However, to avoid issues of receive-window handling at the control stream sequence number level, a host may never have more than one unacknowledged TLV-field in-flight. This effectively limits the amount of memory required to support the control-stream down to 64KB (the maximum size of a TLV-field).

TCP uses the congestion-window to limit the amount of unacknowledged in-flight data within a TCP connection. The control stream must also obey to this limitation. As the control stream uses regular TCP sequence numbers, the congestion-window limitations apply too.

3. Connection initiation

The control stream allows to negotiate the crypto material to authenticate new subflows. Thus, the handshake of the initial subflow does not need anymore to send the 64-bit key in plaintext. The suggested modification to the initial handshake is detailled in this section.

MultiPath TCP uses the MP_CAPABLE option in the handshake for the initial subflow. This handshake was designed to meet several requirements. When designing another variant of the Multipath TCP handshake, it is important to have these requirements in mind. These requirements are :

1. Detect whether the peer supports MultiPath TCP.

2. Exchange locally unique tokens that unambiguously identify the Multipath TCP connection

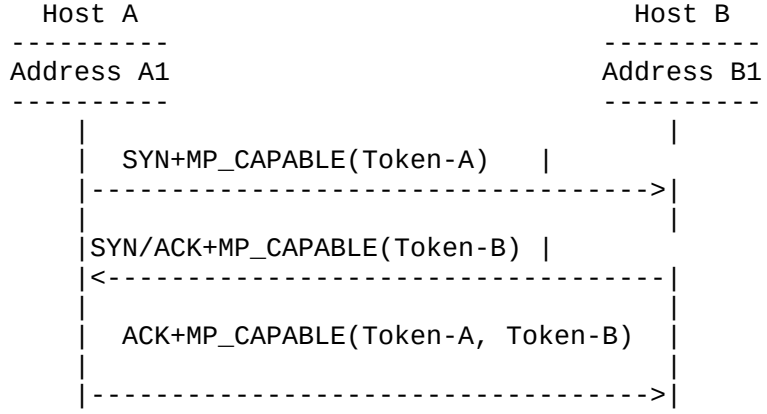
3. Agree on an Initial Data Sequence Number to initialize the MPTCP state on each direction of the Multipath TCP connection

Before discussing the proposed handshake, it is important to have in mind how [RFC6824] meets the three requirements above.

The first requirement is simply met by using a Multipath TCP specific option, like all TCP extensions.

To meet the second requirement, a simple solution would have been to encode the token inside the MP_CAPABLE option. However, this would have increased its size. This would have limited the possibility of extending Multipath TCP later by adding new TCP options that require space inside the SYN segments. To minimize the number of option bytes consummed in the SYN segment, [RFC6824] uses a hash function to compute the token based on the keys exchanged in clear. However, using hash functions implies that implementations must handle the possible collisions which increases the complexity of implementing the Multipath TCP handshake.

In this document we suggest a simplified handshake that meets the above three goals. This simplified handshake avoids negotiating the crypto-material during the three-way handshake. Instead, security information is exchanged reliably by relying on the control stream. The figure below provides an overview of the proposed handshake.



Handshake of the initial subflow.

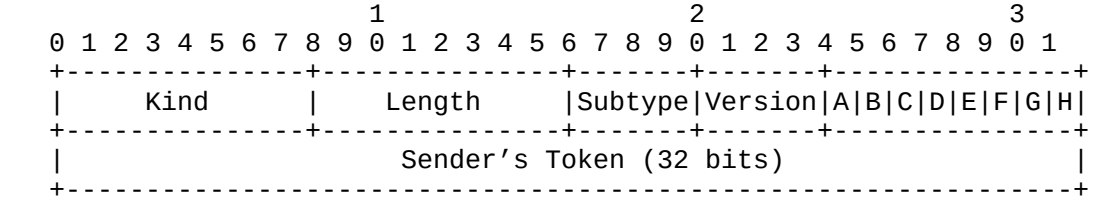
Figure 3

MPTCP's establishment of the initial subflow follows TCP's regular 3-way handshake, but the SYN, SYN/ACK and ACK packets contain the MP_CAPABLE-option. The proposed MP_CAPABLE option contains one 32 bits token in the SYN and SYN/ACK segments. The third ACK includes an MP_CAPABLE option that contains the two tokens. Echoing all the information back in the third ACK allows stateless operation of the

server. The tokens are used to explicitly exchange the identifiers of the Multipath TCP connection.

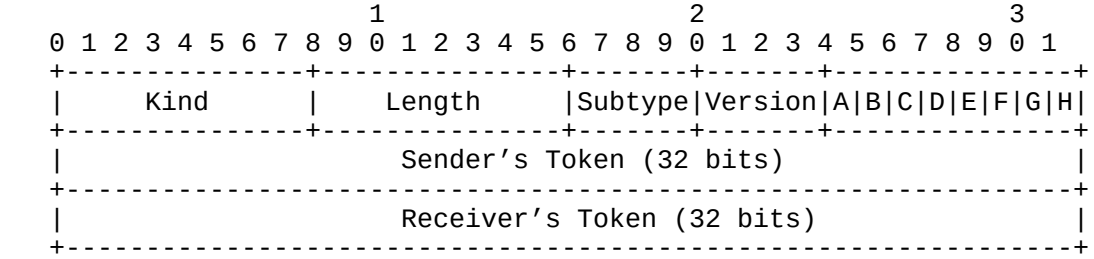
It is required that the server, upon reception of the SYN generates a token different from the client's token. This is necessary to protect against reflection attacks when establishing additional subflows.

The format of the proposed MP_CAPABLE option is proposed in the figures below.



Format of the MP_CAPABLE-option in the SYN and SYN/ACK packets

Figure 4



Format of the MP_CAPABLE-option in the third ACK of the handshake

Figure 5

The format of the MP_CAPABLE option is shown in Figure 4. To indicate that this MP_CAPABLE contains tokens numbers and not keys (as in [RFC6824]), the Version-field is set to 1. The message format of the third ACK's MP_CAPABLE option is show in Figure 5.

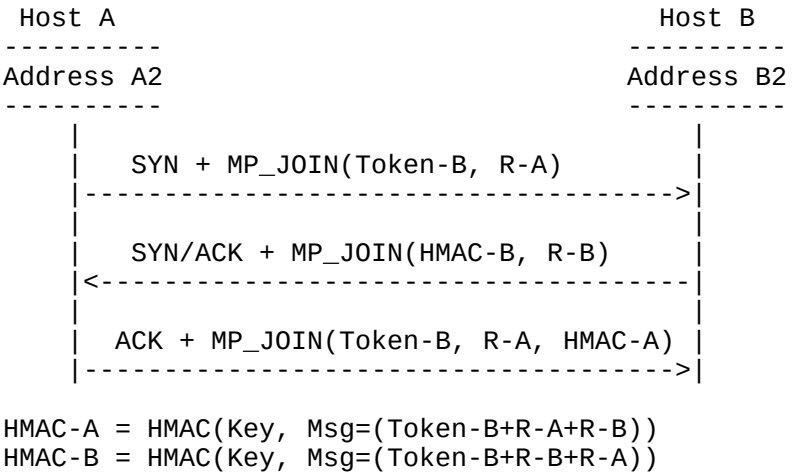
The Initial Data Sequence Number (IDSN) serves to initialize the MPTCP state on the end-hosts in the same way as TCP's sequence numbers do during the 3-way handshake. There is one IDSN for each direction of the data-stream. The IDSN for the data from the client to the server is the 64 low-order bits of the hash (SHA1) of the concatenation of the tokens (Token-A || Token-B). For the data from

server to client, the IDSN is 64 low-order bits of the hash (SHA1) of the reverse concatenation (Token-B || Token-A). The tokens should be generated with sufficient randomness so that they are hard to guess. Recommendations for generating random numbers are given in [RFC4086].

The meaning of the other fields and behavior of the end-hosts during the MP_CAPABLE exchange is the same as specified in [RFC6824].

4. Starting a new subflow

The handshake for the establishment of a new subflow is similar to the one specified in [RFC6824]. There are three important differences. First, the HMAC is computed by using the keys negotiated over the control stream. Second, the token and the client's random numbers are included inside the third ack to allow stateless operation of the passive opener of an additional subflow. Finally, the token is used within the message of the HMAC. This protects against reflection attacks, as the HMAC cannot be sent in the reverse direction anymore, because the tokens are ensured to be different on both end-hosts.

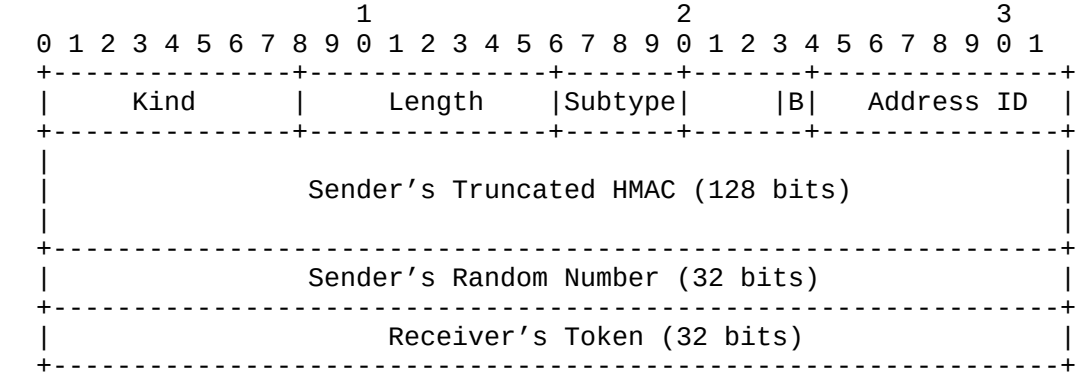


Handshake of a new subflow.

Figure 6

In order to allow the Token-B and R-A inside the third ack, the HMAC-A must also be a truncated version of the 160-bit HMAC-SHA1. Thus, HMAC-A is the truncated (leftmost 128 bits) of the HMAC as shown in Figure 6.

The message-format of the MP_JOIN-option in the SYN and the SYN/ACK is the same as in [RFC6824]. As the third ACK includes the Token and the random nonce, the MP_JOIN message format of the third ack is as shown in Figure 7. The length of the MP_JOIN-option in the third ACK is 28 bytes. Thus, there remains enough space to insert the timestamp option in the third ACK.



Format of the MP_JOIN-option

Figure 7

The semantics of the backup-bit "B" and the Address ID are the same as in [RFC6824].

5. Examples of key negotiation through the control stream

The control stream’s primary goal is to negotiate the crypto-material to authenticate additional subflows. Both hosts must agree on which key-negotiation scheme to use over the control stream. The option "key select" of the control stream is of type 1 and it negotiates the available key-negotiation schemes. The value-field of the "key select"-option contains a bitmask of available key-negotiation schemes. The bitmask remains to be defined as the schemes are being defined. The bits within the bitmask are numbered, starting from the leftmost as being ’1’.

The key-select must be initiated by one host and answered by the other one. During the initiation, the host offers the available schemes, and the answering host selects one of the offered ones. The hosts need thus to ensure an order among themself of who initiates the "key select" option. A possibility would be that the host with the smaller token initiates the "key select" option.

The following are examples of how the control stream could be used to negotiate the cryptographic material. A proper specification is probably needed for each of them.

5.1. Reusing the application's TLS key

Within the "key select"-option, this negotiation scheme takes the bit number 1. It signals to the peer that the connection should use a derivate of TLS's master key to authenticate new subflows with this "MPTCP key". It is required that indeed TLS is being used within the data stream.

As TLS allows to modify the key being used during a TLS session, the control stream might be used to ensure that both end hosts agree on the "MPTCP key" being used at a specific moment in time through the exchange of the hash of the "MPTCP key".

5.2. TLS-like key exchange

It enables a key-negotiation in an TLS-like manner, thus authenticating the client/server through a certificate.

5.3. Tcpcrypt-like key exchange

It uses the control stream, to exchange a secret key in a tcpcrypt-like manner. Optionally, it may include a data-sequence number to define from which moment on the data stream should be encrypted.

6. Other example use cases of the control stream

This shows one example of how the control stream can be used within MPTCP.

6.1. Address signaling

In RFC6824, the address-signaling is achieved through the ADD_ADDRESS and REMOVE_ADDRESS options. These options are sent within the TCP options-space and thus do not benefit from reliable delivery. Further, security-concerns have rosen concerning the ADD_ADDRESS-option. Using the control stream to signal the addition or removal of addresses allows to make these options reliable and provides the space to add any kind of cryptographic material to enhance their security.

7. Security Considerations

TBD

8. Acknowledgments

This work is supported by the European FP7 Project "Trilogy2" under grant agreement 317756.

9. Informative References

[RFC4086] Eastlake, D., Schiller, J., and S. Crocker, "Randomness Requirements for Security", BCP 106, RFC 4086, June 2005.

[RFC4960] Stewart, R., "Stream Control Transmission Protocol", RFC 4960, September 2007.

[RFC6181] Bagnulo, M., "Threat Analysis for TCP Extensions for Multipath Operation with Multiple Addresses", RFC 6181, March 2011.

[RFC6824] Ford, A., Raiciu, C., Handley, M., and O. Bonaventure, "TCP Extensions for Multipath Operation with Multiple Addresses", RFC 6824, January 2013.

Authors' Addresses

Christoph Paasch
UCLouvain
Place Sainte Barbe, 2
Louvain-la-Neuve 1348
BE

Email: christoph.paasch@uclouvain.be

Olivier Bonaventure
UCLouvain
Place Sainte Barbe, 2
Louvain-la-Neuve 1348
BE

Email: olivier.bonaventure@uclouvain.be

MPTCP
Internet-Draft
Intended status: Informational
Expires: April 18, 2013

C. Paasch, Ed.
O. Bonaventure
UCLouvain
October 15, 2012

Securing the MultiPath TCP handshake with external keys
draft-paasch-mptcp-ssl-00

Abstract

Multipath TCP currently relies on the exchange of keys in clear during the initial handshake to authenticate the establishment of additional subflows. This document proposes a variant of the Multipath TCP handshake that allows Multipath TCP to reuse keys negotiated by the Application layer protocol above it such as SSL/TLS to authenticate the establishment of additional subflows.

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 April 18, 2013.

Copyright Notice

Copyright (c) 2012 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

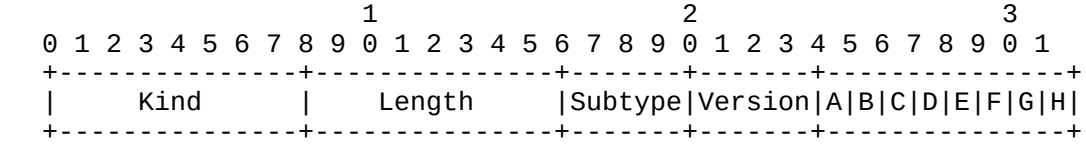
1. Introduction 3
2. Connection initiation 3
3. Multipath TCP API 4
4. Starting a new subflow 4
5. Deployment 6
6. Security Considerations 7
7. Informative References 7
Authors' Addresses 7

1. Introduction

Multipath TCP is an extension to TCP that enables hosts to use multiple paths to exchange data for a single connection. [I-D.ietf-mptcp-multiaddressed] describes the current design of the Multipath TCP protocol. The design of Multipath TCP has been influenced by various factors including the backward compatibility with regular TCP, the fallback to TCP when middleboxes interfere with the Multipath TCP options, ... The design of Multipath TCP has also been affected by security requirements. The security threats against Multipath TCP are documented in [RFC6181]. Multipath TCP aims at being no worse than TCP from a security viewpoint. Other approaches such as [I-D.bittau-tcp-crypt] or [RFC5925] have been proposed to reduce the vulnerability of TCP to attacks. Multipath TCP currently addresses the security threats identified in [RFC6181] by exchanging keys during the handshake for the initial subflow. These keys are then used to generate HMACs to authenticate the establishment of subsequent TCP subflows. Exchanging keys in clear during the initial handshake has obvious shortcomings from a security viewpoint. However, some application-layer protocols like SSL/TLS or ssh already negotiate a shared key between the end-points. In this document we propose a modification to the handshake used by Multipath TCP for the initial and subsequent subflows that enables Multipath TCP to rely on an application-supplied key to authenticate the establishment of the subflows.

2. Connection initiation

The handshake of the initial subflow is a small variation to the handshake of [I-D.ietf-mptcp-multiaddressed] or draft-paasch-mptcp-lowoverhead-00. The header of the MP_CAPABLE option of these two MPTCP-versions has the format as shown in the below figure.



Header of the MP_CAPABLE option

Figure 1

We propose to use the B bit in this option to indicate whether the host that sent the MP_CAPABLE option will use an application supplied key to authenticate the additional subflows or not. When the B bit

is set, it indicates that the authentication key is supplied by the application. If the B bit has not been set in both directions, the authentication mechanism is used as defined by the MPTCP version ([I-D.ietf-mptcp-multiaddressed] or draft-paasch-mptcp-lowoverhead-00).

In MPTCP version 0, even if the B bit is set the end-hosts still have to generate a key that fulfills the requirements as defined in MPTCP version 0. This is necessary to handle the case where the client supports the B bit, but the server not yet. For a more in-depth analysis of this kind of deployment scenario, have a look at Section 5.

By using the same handshake as draft-paasch-mptcp-lowoverhead-00, the proposed handshake can also benefit from the lower overhead for generating the token and thus the faster establishment of the initial subflow.

3. Multipath TCP API

The proposed mechanism requires an interaction between the application and the MPTCP layer. This can be achieved by the means of socket options. Two socket options are necessary:

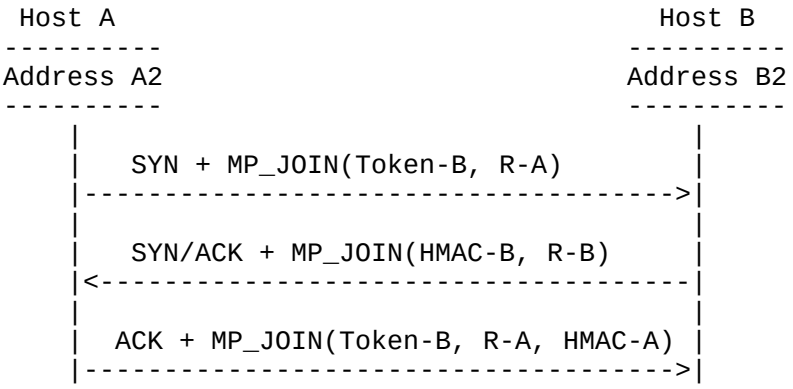
- MPTCP_ENABLE_APP_KEY : This socket option tells the socket layer that an application supplied key will be used to secure the establishement of additional subflows. This socket option MUST be used before establishing the initial subflow, or before starting to listen on a socket to accept new connections. When this socket option is used, the MP_CAPABLE option is sent with the "B"-bit set to 1.

- MPTCP_KEY : This socket option allows the application to provide a key to the MPTCP layer. Both end-points MUST use this socket option in order to allow the MPTCP-layer to create new subflows. It is up to the application to negotiate the key between the end-points. E.g., in the case of SSL/TLS, the key can be a hash of the shared secret that has been negotiated with the SSL exchange. Separate documents will describe in details how applications such as TLS or SSH can pass a shared secret to Multipath TCP by using this option.

4. Starting a new subflow

The handshake for the establishment of a new subflow is similar to the one specified in [I-D.ietf-mptcp-multiaddressed]. There are two

important differences. First, the HMAC is computed by using the keys provided by the application. Second, the token and the client's random number are included inside the third ack to allow stateless operation of the passive opener of an additional subflow.

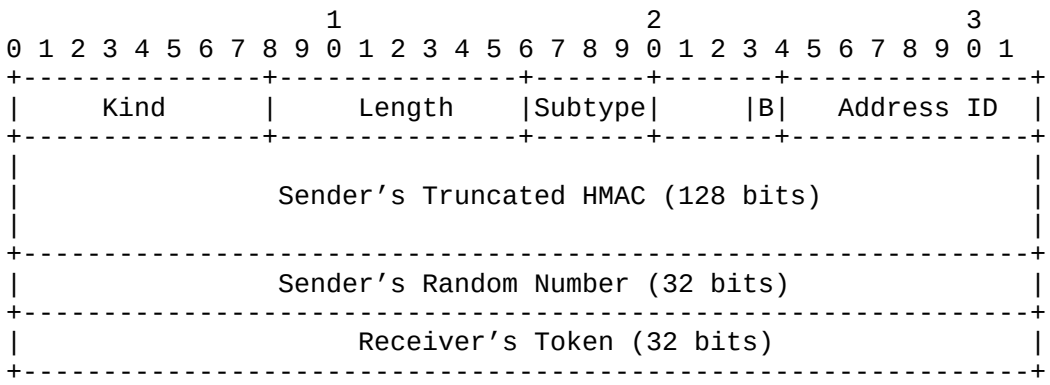


Handshake of a new subflow.

Figure 2

In order to allow the Token-B and R-A inside the third ack, the HMAC-A must also be a truncated version of the 160-bit HMAC-SHA1. Thus, HMAC-A is the truncated (leftmost 128 bits) of the HMAC as shown in Figure 2.

The message-format of the MP_JOIN-option in the SYN and the SYN/ACK is the same as in [I-D.ietf-mptcp-multiaddressed]. As the third ACK includes the Token and the random nonce, the MP_JOIN message format of the third ack is as show in Figure 3. The length of the MP_JOIN-option in the third ACK is 28 bytes. There remains thus enough space to insert the timestamp option in the third ACK.



Format of the MP_JOIN-option

Figure 3

The semantics of the backup-bit "B" and the Address ID are the same as in [I-D.ietf-mptcp-multiaddressed].

5. Deployment

This proposed mechanism assumes that the application uses new socket-options to provide the key to the MPTCP-layer. Thus, the first requirement for deploying this MPTCP handshake is that the TLS/SSL-layer has been modified. There may of course be scenarios, where the client is supporting the proposed solution, but the server not. Thus, the client sends out the MP_CAPABLE with the B bit set, but the server replies without enabling the B bit. Upon reception of the SYN/ACK, it is up to the client's policy how to react. It can either continue with the negotiated version of MPTCP but without using the key from the application or fallback to regular TCP.

The applications will have to pass the shared key to the MPTCP-layer by the means of a socket-option. It may be that the client's application has already done the call to the socket-option but the server's application not yet. The server will receive a SYN with the MP_JOIN-option, without knowing the key. In that case the server should silently drop the SYN. The TCP retransmission mechanism on the client-side will retransmit the SYN after the initial RTO expired (after 1 second). And the server's application potentially will have finally set the key via the socket-option.

6. Security Considerations

It is recommended that the applications do not pass the plain shared key to the MPTCP layer. They should rather pass a hash of their shared secret to the MPTCP layer. These security considerations will be discussed in documents that describe how applications such as TLS/SSL or SSH can interact efficiently with Multipath TCP.

7. Informative References

[I-D.bittau-tcp-crypt]
Bittau, A., Boneh, D., Hamburg, M., Handley, M., Mazieres, D., and Q. Slack, "Cryptographic protection of TCP Streams (tcpcrypt)", draft-bittau-tcp-crypt-03 (work in progress), September 2012.

[I-D.ietf-mptcp-api]
Scharf, M. and A. Ford, "MPTCP Application Interface Considerations", draft-ietf-mptcp-api-05 (work in progress), April 2012.

[I-D.ietf-mptcp-multiaddressed]
Ford, A., Raiciu, C., Handley, M., and O. Bonaventure, "TCP Extensions for Multipath Operation with Multiple Addresses", draft-ietf-mptcp-multiaddressed-10 (work in progress), October 2012.

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, June 2010.

[RFC6181] Bagnulo, M., "Threat Analysis for TCP Extensions for Multipath Operation with Multiple Addresses", RFC 6181, March 2011.

Authors' Addresses

Christoph Paasch (editor)
UCLouvain
Place Sainte Barbe, 2
Louvain-la-Neuve, 1348
BE

Email: christoph.paasch@uclouvain.be

Olivier Bonaventure
UCLouvain
Place Sainte Barbe, 2
Louvain-la-Neuve, 1348
BE

Email: olivier.bonaventure@uclouvain.be