An MPTCP Option for Network-Assisted MPTCP Deployments: Plain Transport Mode

Mohamed Boucadair, Christian Jacquenet, Denis Behaghel, Stefano Secci, Wim Henderickx, Robert Skog, Olivier Bonaventure, Suresh Vinapamula, Sunghoon Seo, Wouter Cloetens, et al.

Network Working Group
Internet-Draft
Intended status: Experimental
Expires: January 5, 2017

M. Boucadair
C. Jacquenet
Orange
D. Behaghel
OneAccess
S. Secci
UPMC
W. Henderickx
Nokia/Alcatel-Lucent
R. Skog
Ericsson
O. Bonaventure
Tessares
S. Vinapamula
Juniper
S. Seo
Korea Telecom
W. Cloetens
SoftAtHome
U. Meyer
Vodafone
LM. Contreras
Telefonica

July 4, 2016

An MPTCP Option for Network-Assisted MPTCP Deployments: Plain Transport Mode

draft-boucadair-mptcp-plain-mode-08

Abstract

Because of the lack of Multipath TCP (MPTCP) support at the server side, some service providers now consider a network-assisted model that relies upon the activation of a dedicated function called MPTCP concentrator. This document focuses on a deployment scheme where the identity of the MPTCP concentrator(s) is explicitly configured on connected hosts.

This document specifies an MPTCP option that is used to avoid the encapsulation of packets and out-of-band signaling between the CPE and the MPTCP concentrator. Also, this document specifies how UDP traffic, in particular, can be distributed among available paths by leveraging MPTCP capabilities.

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 [RFC2119].

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 January 5, 2017.

Copyright Notice

Table of Contents

1. Introduction

One of the promising deployment scenarios for Multipath TCP (MPTCP, [RFC6824]) is to enable a Customer Premises Equipment (CPE) that is connected to multiple networks (e.g., DSL, LTE, WLAN) to optimize the usage of such resources. This deployment scenario is called a network-assisted MPTCP model, and relies upon MPTCP proxies located on both the CPE and network sides (Figure 1). The latter plays the role of an MPTCP concentrator. Such concentrator terminates the MPTCP sessions established from CPEs, before redirecting traffic into legacy TCP sessions.

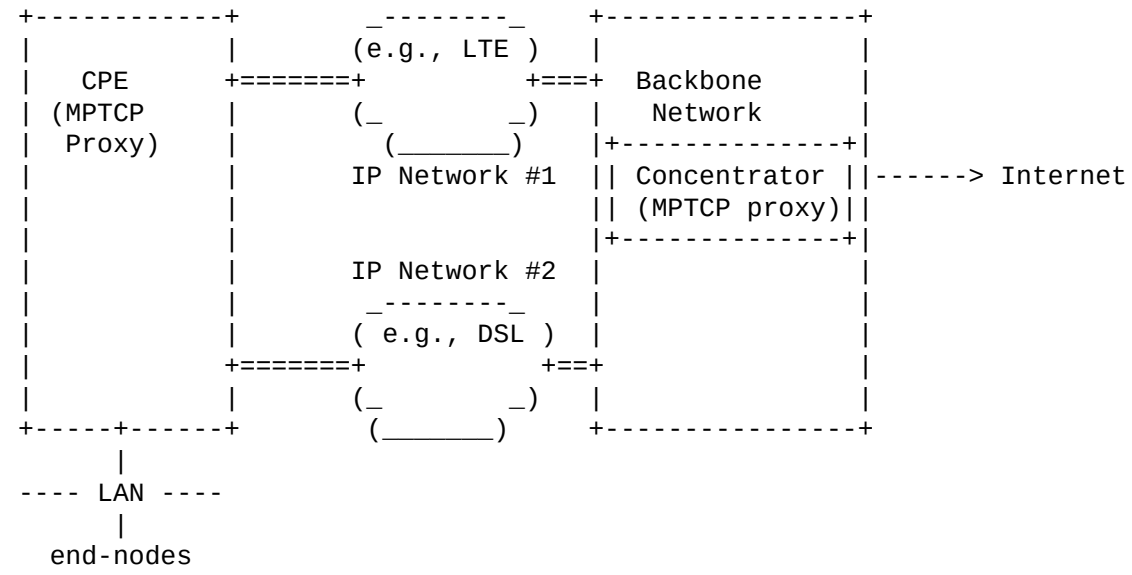
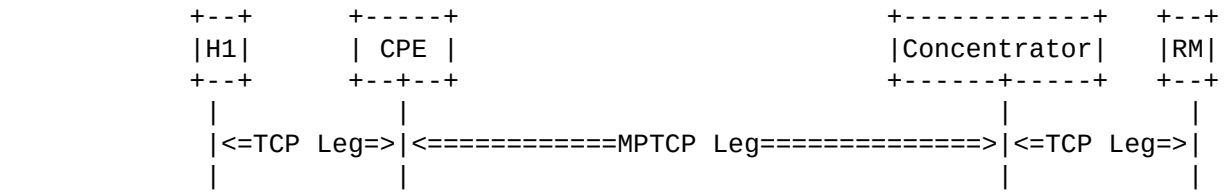


Figure 1: "Network-Assisted" MPTCP Design

Network-assisted MPTCP deployment models are designed to facilitate the adoption of MPTCP for the establishment of multi-path communications without making any assumption about the support of MPTCP by the communicating peers. Thus, MPTCP proxies deployed in CPEs and in concentrators located in the network are responsible for establishing multi-path communications on behalf of endpoints, thereby taking advantage of MPTCP capabilities to optimize resource usage to achieve different goals that include (but are not limited to) bandwidth aggregation, primary/backup communication paths, and traffic offload management. Figure 2 depicts the various TCP connection legs in network-assisted MPTCP deployment models.



Legend:
H1: Host 1
RM: Remote Machine

Figure 2: Connection Legs (CPE-based Model)

There are also MPTCP deployments to assist hosts that are directly connected to multiple networks to establish multi-path

communications. The communication legs that are involved in such deployments are shown in Figure 2.

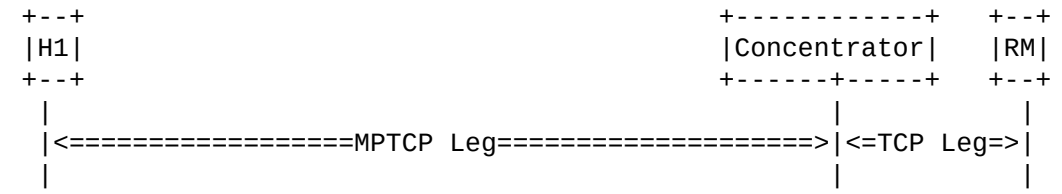


Figure 3: Connection Legs (Host-based Model)

Most of the current operational deployments that take advantage of multi-interfaced devices rely upon the use of an encapsulation scheme (such as GRE [I-D.zhang-gre-tunnel-bonding]). The use of encapsulation is motivated by the need to steer traffic towards the concentrator and also to allow the distribution of any kind of traffic besides TCP (e.g., UDP) among the available paths without requiring any advanced traffic engineering tweaking technique in the network side to intercept traffic and redirect it towards the appropriate concentrator.

This specification assumes an MPTCP concentrator is reachable by means of one or multiple IP addresses. Also, it assumes the various network attachments provided to an MPTCP-enabled device (CPE or host) are managed by the same administrative entity. The IP reachability information of an MPTCP concentrator can be explicitly configured on a device, e.g., by means of a specific DHCP option [I-D.boucadair-mptcp-dhc]. This document assumes such explicit configuration. Additional assumptions are listed in Section 3.

Current operational MPTCP deployments by network operators are focused on the forwarding of TCP traffic. In addition, the design of such deployments sometimes assumes the use of extra signalling provided by SOCKS [RFC1928], at the cost of additional management complexity and possible service degradation (e.g., up to 8 SOCKS messages may need to be exchanged between two MPTCP proxies before an MPTCP connection is established, thereby yielding several tens of milliseconds of extra delay before the connection is established) .

To avoid the burden of encapsulation and additional signalling between MPTCP proxies, this document explains how a plain transport mode is enabled, so that packets are exchanged between the CPE and the concentrator without requiring the activation of any encapsulation scheme (e.g., IP-in-IP [RFC2473], GRE [RFC1701]). This plain transport mode also avoids the need for out-of-band signalling.

The solution described in this document also works properly when NATs are present in the communication path between the CPE and the concentrator, unlike solutions that rely upon GRE tunneling. In particular, the solution proposed in this document accommodates deployments that involve CGN (Carrier Grade NAT) upstream the concentrator.

The plain transport mode is characterized as follows:

- No encapsulation required (no tunnels, whatsoever).
- No out-of-band signaling for each MPTCP subflow required.
- Carries any protocol (incl. UDP) for the benefit of massive MPTCP adoption (Section 5).
- Accommodates various deployment contexts (Section 6).

2. Terminology

The reader should be familiar with the terminology defined in [RFC6824].

This document makes use of the following terms:

Customer-facing interface: is an interface of the MPTCP concentrator that is visible to a CPE or a host directly connected to the operator's network, and which is used for communication purposes between a CPE/host and the MPTCP concentrator.

Internet-facing interface: is an interface of the MPTCP concentrator that is visible to a remote host on the Internet.

IP transport address: refers to an IP address and transport port number.

MPTCP proxy: is a software module that is responsible for transforming a TCP connection into an MPTCP connection, and vice versa. Typically, an MPTCP proxy is embedded in a CPE and a concentrator.

MPTCP leg: refers to a network segment where MPTCP is used to establish TCP connections (see Figure 2).

MPTCP concentrator (or concentrator): refers to a functional element that is responsible for aggregating traffic pertaining to a group of CPEs. This element is typically located upstream in the network, e.g., beyond a Broadband Network Gateway (BNG) or a PDN Gateway (PGW) in wired and wireless access network environments, respectively. One or multiple concentrators can be deployed in

the network to help MPTCP-enabled CPEs establish MPTCP connections via available network attachments.

On the uplink path, the concentrator terminates the MPTCP connections received from its customer-facing interfaces and transforms these connections into legacy TCP connections towards upstream servers.

On the downlink path, the concentrator converts the legacy server's TCP connections into MPTCP connections towards its customer-facing interfaces.

3. Assumptions & Scope

The following assumptions are made:

- The logic for mounting network attachments by a CPE (or a host directly connected to the operator's network) is deployment- and implementation-specific and is out of scope of this document.

- Policies can be enforced by a concentrator instance operated by the Network Provider to manage both upstream and downstream traffic. These policies may be subscriber-specific, connection-specific, system-wide, or else.

- The concentrator may be notified about monitoring results (e.g., provided by passive or active probes) that detail the status of the various network legs available to service a customer, a group of customers, a whole region, etc. No assumption is made in this document about how these monitoring operations are executed.

- An MPTCP-enabled, multi-interfaced CPE or host that is directly connected to one or multiple access networks is allocated addresses/prefixes via legacy mechanisms (e.g., DHCP) supported by the various available network attachments. The CPE/host may be assigned the same or distinct IP address/prefix via the various available network attachments.

- The location of the concentrator(s) is deployment-specific. Network Providers may choose to adopt centralized or distributed designs. Nevertheless, in order to take advantage of MPTCP, the location of the concentrator should not jeopardize packet forwarding performance overall.

- The logic of traffic distribution over multiple paths is deployment-specific. This document does not require nor preclude any particular traffic distribution schemes.

- No assumption is made whether one single or multiple IP addresses/ prefixes are assigned to host connected to a CPE.

It is out of the scope of this document to discuss criteria for selecting traffic to be eligible to MPTCP service. It is out of scope of the document to specify how a CPE selects its concentrator(s), too.

Likewise, methods to avoid TCP fragmentation, such as rewriting the TCP Maximum Segment Size (MSS) option, are out of scope for this document.

This document focuses on the CPE-based model (i.e., the CPE embeds a MPTCP proxy that behaves on behalf of terminal devices), but plain transport mode can also apply to host-based models.

TCP/MPTCP session tracking by the MPTCP proxy is implementation-specific. Readers may refer to Section 2 of [RFC7857].

This specifications focuses on TCP and UDP. Future documents may specify the exact behavior for transporting other protocols over MPTCP connections.

Also, this specification focuses on a stateful design; stateless approaches that rely on including the Plain Mode option in all packets are out of scope.

4. Plain Transport Mode Behavior

As shown in Figure 2, TCP connections initiated by a host are converted by the CPE into MPTCP connections towards the concentrator. Then, the concentrator converts these connections into legacy TCP connections towards the final destinations. Since the concentrator can be located anywhere in the operator’s network, Section 4.1 introduces a new TCP option to supply the concentrator with required information to forward the traffic to its final destination. When a CPE receives a SYN segment from a host of the LAN, it rewrites the destination address of that segment to an address of the concentrator, and places the original destination (and possibly source) addresses in this TCP option. Further details are specified in the following sub-sections.

Specific UDP processing is discussed in Section 5.

4.1. Plain Mode MPTCP Option

The Plain Mode (PM) option carries the source/destination IP addresses and/or port numbers of the origin source and destination nodes. It is also used to indicate whether the data carried in the packet is relayed from a native TCP connection or refers to the use of another transport protocol. The format of the option is shown in Figure 4.

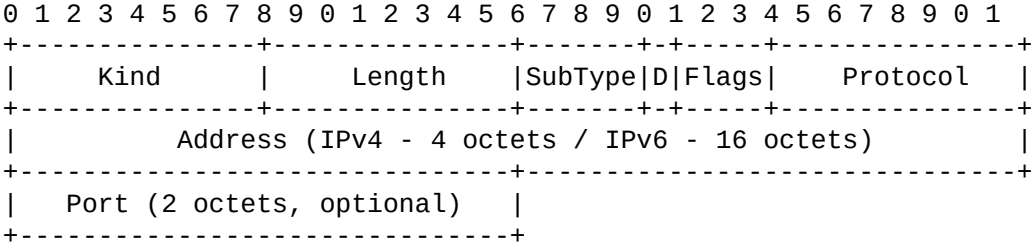


Figure 4: Plain Mode MPTCP Option

The description of the fields is as follows:

- Kind and Length: are the same as in Section 3 of [RFC6824].
- Subtype: to be defined by IANA (Section 8). Implementations may use "0xe" subtype encoding for early deployment purposes in managed networks.
- D-bit (direction bit): this flag indicates whether the enclosed IP address (and port number) reflects the source or destination IP address (and port). When the D-bit is set, the enclosed IP address must be interpreted as the source IP address. When the D-bit is unset, the enclosed IP address must be interpreted as the destination IP address.
- "Flag" bits: are reserved bits for future assignment as additional flag bits. These additional flag bits MUST each be set to zero and MUST be ignored upon receipt.
- Protocol: conveys the protocol number as assigned by IANA [proto_numbers]. For example, this field is set to 17 for UDP traffic, or 6 for TCP. The processing of UDP flows is further discussed in Section 5.
- Address: includes a source or destination IP address. The address family is determined by the "Length" field. Concretely, a PM option containing an IPv4 address has a Length field of 8 bytes (or 10 if a port number is included). A PM option containing an

IPv6 address has a Length of 20 bytes (or 22 if a port number is included).

- Port: If the D-bit is set (resp. unset), a source (resp. destination) port number may be associated with the IP address. This field is valid for protocols that use a 16 bit port number (e.g., UDP, TCP, SCTP).

4.2. Carrying the Plain Mode Option

When using an MPTCP connection to forward traffic (whether it's TCP traffic or any other traffic), the CPE (resp. the concentrator) MUST insert a Plain Mode option in a SYN packet sent to the concentrator (resp. the CPE). The Plain Mode option MUST be included in the SYN payload.

Note: Given the length of the PM option, especially when IPv6 addresses are used, and the set of TCP options that are likely to be included in a SYN message, it will not always be possible to place the PM option inside the dedicated TCP option space. Given that this option is designed to be used in a controlled environment, this specification recommends to always place the PM options inside the payload of a SYN segment. Including data in a SYN payload is allowed as per Section 3.4 of [RFC0793].

If the original SYN message contains data in its payload (e.g., [RFC7413]), that data MUST be placed right after PM and "End of Options List" (EOL) options when generating the SYN in the MPTCP leg. The EOL option serves as a marker to delineate the end of the TCP options and the beginning of the data included in the original SYN .

The Plain Mode option MUST only appear in SYN segments that contain the MP_CAPABLE option. SYN messages to create subsequent subflows of a given MPTCP connection MUST NOT include any PM option (Figure 5).

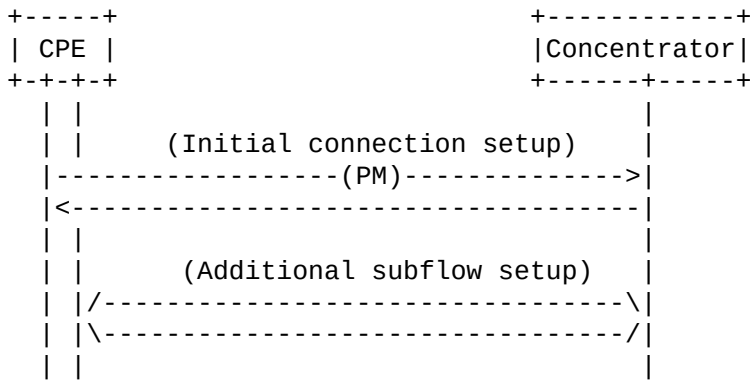


Figure 5: Carrying the Plain Mode Option (Focus on the MPTCP Leg)

By default, source IP address preservation is assumed for IPv6 while global address sharing is assumed for IPv4. This means that, by default, two plain mode option instances MUST be included in a SYN segment for IPv6 (both source and destination) and one instance MUST be present for IPv4 (either the source or the destination). The CPE and the concentrator MUST support a configurable parameter to modify this default behavior to accommodate alternate deployment models (see Section 6).

An implementation MUST ignore PM options that include multicast, broadcast, and host loopback addresses [RFC6890].

The 'Protocol' field of the PM option MUST be copied from the 'Protocol' field of the IPv4 header or set to the type of the transport header of the IPv6 packet that will be forwarded along MPTCP subflows. The CPE and the concentrator MAY be configured to disable traffic aggregation for some transport protocols because of the nature of the service they relate to (e.g., IP multicast traffic typically specific of live TV broadcasting services). By default, TCP and UDP traffic bonding MUST be enabled.

4.3. Binding Tables

4.3.1. On the Need to Maintain a State

Because the source IP and/or destination IP addresses are communicated only during the SYN exchange of the initial subflow, the CPE and the concentrator MUST maintain a state that binds the MPTCP transport coordinates to the destination/source IP address, ports, and protocol. This specification discusses the external behavior of this stateful design; the internal behavior for maintaining that state is implementation-specific.

This document uses 'Internal transport session identifier' to identify a particular transport session on the LAN side of the CPE and 'External transport session identifier' to identify a particular transport session on the Internet-facing Interface of the concentrator. An implementation could use the classical 4-tuple (source and destination addresses and ports) as such an identifier.

An MPTCP proxy also needs to identify a particular MPTCP connection. We refer to it as the 'MPTCP transport coordinates'. An implementation could, for example, use the token assigned to a specific connection to identify an MPTCP connection. The 4-tuple of each subflow that belong to an MPTCP connection can also be part of the MPTCP transport coordinates.

Binding entries can be created as a result of a packet or be configured directly on the CPE or the concentrator.

4.3.2. Binding & Transport Session Entries

An implementation may maintain distinct binding tables, each for a given transport protocol, or maintain one single binding table to handle all supported transport protocols.

Subflows can be added or deleted during the lifetime of an MPTCP connection based on triggers that are local to the CPE/concentrator, based on signals received from the concentrator/CPE, or as a result of processing a packet. These triggers are outside the scope of this specification.

The CPE must maintain a binding entry that allows to associate the internal transport address (IP address, port number) with one or a set of external IP transport addresses, that are assigned in the WAN interfaces of the CPE in the context of a given MPTCP connection. Each of the external transport addresses points to a subflow that is created between the CPE and the concentrator. For each binding entry, one or multiple transport session entries are maintained by the CPE and the concentrator. These session entries are meant to store the information that is required for rewriting packets. A session entry is created as a result of handling a packet.

A session entry maintained by the CPE may be structured as follows:

Internal transport session identifier: This information typically include the source IP transport address (IPl, Pl) and the destination IP transport address (IPd, Pd) of the connection.

MPTCP transport coordinates: These coordinates include information about the subflows that compose this MPTCP connection. When a

packet matches an existing binding entry, the CPE may decide whether existing subflows can be used to forward the packet, or whether a new subflow is to be created.

The following information is maintained for each MPTCP subflow:

* (IPwi, Pwi): The source IP address and port for this subflow.

* (IPci, Pci): IPci is one of the concentrator's IP addresses, while Pci is a port number selected to establish this subflow. This information is used as the destination IP address and port of a packet matching this entry.

Transport protocol: This information is typically retrieved from the outgoing packet that will be placed in MPTCP connections. The transport protocol specifies the protocol that is used in the LAN side.

Lifetime: This information indicates the remaining validity lifetime for the session entry. When the lifetime expires, this session entry is deleted from the table. If all sessions bound to a given binding entry expired, that binding entry must be deleted. Recommendations for setting this parameter are defined in [RFC7857][RFC5382][RFC4787].

For example:

o An outgoing packet {src=(IPl, Pl); dst=(IPd,Pd)} will be transformed by the CPE to {src=(IPwi,Pwi); dst=(IPci,Pci)}.

o An incoming packet {src=(IPci,Pci); dst=(IPwi,Pwi) will be transformed by the CPE to {src=(IPd, Pd); dst=(IPl,Pl)};

The structure of a session entry maintained by the concentrator may be as follows:

External transport session identifier: This information typically include the external IP transport address (IPe, Pe) and the destination IP transport address (IPd, Pd) of the connection. The external IP transport address is set to the (IPl, Pl) if and only if the concentrator is configured to preserve the source IP address and port. In such case, this information is retrieved from the PM option included in a SYN packet. Otherwise, the external IP address and port are selected by the concentrator from a local pool.

MPTCP transport coordinates: These coordinates include information about the subflows that compose this MPTCP connection. When a

packet matches an existing binding entry, the concentrator may decide whether existing subflows can be used to forward the packet, or whether a new subflow is to be created.

The following information is maintained for each MPTCP subflow:

* (IPwi, Pwi): The source IP address and port for this subflow.

* (IPci, Pci): The destination IP address and port for this subflow. It can be set by the CPE or selected by the concentrator.

Transport protocol: This information is retrieved from the PM option included in a SYN packet. The transport protocol specifies the protocol that must be used when sending the packet through the Internet-facing interface.

Lifetime: This information indicates the remaining validity lifetime for the session entry. When the lifetime expires, this session entry is deleted from the table. If all sessions bound to a given binding entry expired, that binding entry must be deleted. Recommendations for setting this parameter are defined in [RFC7857][RFC5382][RFC4787].

For example:

o An outgoing packet {src=(IPwi,Pwi); dst=(IPci,Pci)} will be transformed by the concentrator to {src=(IPe,Pe); dst=(IPd,Pd)}.

o An incoming packet {src=(IPd,Pd); dst=(IPe,Pe) will be transformed by the concentrator to {src=(IPd, Pd); dst=(IPci,Pci)}.

4.3.3. Expiration of a Binding Entry

A configurable parameter MAY be supported by the CPE and the concentrator to terminate MPTCP connections with the FASTCLOSE procedure (Section 3.5 of [RFC6824]) when a binding entry expires.

If there is no binding state that matches a received non-SYN segment, the CPE/concentrator SHOULD reply with a RST segment. This behavior aims to synchronize the binding tables between the CPE and the concentrator by clearing bindings that are present either in the CPE or in the concentrator.

The configurable parameter is set by default to 'Disable'.

4.4. Theory of Operation: Focus on TCP

4.4.1. Processing an Outgoing SYN

PM option usage for an outgoing TCP SYN (i.e., from the CPE to the concentrator) is as follows:

(1) Outgoing TCP SYNs that can be forwarded by a CPE along MPTCP subflows are transformed by the CPE into TCP packets carried over an MPTCP connection.

The decision-making process to decide whether a given flow should be MPTCP-serviced or not is local to the CPE, and reflects the service-inferred policies as defined by the bonding service provider. As such, the decision-making process is policy-driven, implementation- and deployment-specific.

(2) As a result, SYNs packets are sent over an MPTCP connection according to the plain transport mode (i.e., without any encapsulation header), and the related instructions carried in the PM option.

The source IP address and port number are those assigned to one of the CPE WAN interfaces. Because multiple IP addresses may be available to the CPE, the address used to rewrite the source IP address for an outgoing packet forwarded through a given network attachment (typically, a WAN interface) MUST be associated with that network attachment. It is assumed that ingress traffic filtering policies ([RFC2827]) are enforced at the network boundaries to prevent any spoofing attack.

The destination IP address is replaced by the CPE with one of the IP addresses of the concentrator.

The destination port number may be maintained as initially set by the host or altered by the CPE.

The original destination and/or source IP address are copied into Plain Mode options. The option is inserted as per the guidelines documented in Section 4.2.

A session entry (including the protocol) MUST be maintained by the CPE for that outgoing packet (Section 4.3). A timeout is associated with this entry as per the recommendations in [RFC5382].

(3) Upon receipt of a SYN packet on its MPTCP leg, the concentrator extracts the IP address(es) included in the PM option and uses

it as the destination (and possibly the source) IP address of the corresponding SYN packet that it will forward towards its final destination. The 'Protocol' field of the PM option indicates the transport protocol that must be used when sending the packet through the Internet-facing interface.

The source IP address and port belong to a pool that is configured to the concentrators if address or prefix rewriting is enabled (see Section 6). A session entry MUST be instantiated by the concentrator to record the state (see Section 4.3).

The concentrator may be configured to behave as either a 1:1 IPv4 address translator or a N:1 IPv4 address translator where a given global IPv4 address is therefore shared by multiple CPEs. Network Providers should be aware of the complications that may arise if a given IP address/prefix is shared by multiple customers (see [RFC6269][RFC6967]). Whether these complications apply or not to a network-assisted MPTCP environment is deployment-specific.

The concentrator should preserve the same external IP address that was assigned to a given CPE for all its outgoing connections when forwarding traffic from an MPTCP connection to the Internet (i.e., use an "IP address pooling" behavior of "Paired") [RFC4787]. The port allocation policy configured on the concentrator (e.g., port set assignment, deterministic NAPT, etc.) is implementation and deployment-specific.

4.4.2. Processing an Incoming SYN

In order to appropriately handle incoming SYN packets, the concentrator (resp. CPE) are supposed to be configured with instructions that allows to redirect the traffic to the appropriate CPE (resp. Internal host).

Plain transport mode operation for an incoming TCP SYN (i.e., when traffic is forwarded from the concentrator towards the CPE) is as follows:

(1) If the incoming TCP SYN matches a binding entry (Section 4.3), the concentrator rewrites some of the packet's fields according to the information maintained in this entry. In addition, the concentrator records the source IP address and port in the PM option. Also, the 'Protocol' field of the PM option is set according to the guidelines in Section 4.2.

The source IP address is replaced with one of the IP addresses listed in the binding information base maintained by the concentrator.

The destination IP address is replaced with one of the CPE's IP addresses.

A session entry is instantiated to record the transport-related information to rewrite the packet.

(2) Upon receipt of the TCP SYN by the CPE, it extracts the IP address included in the Plain Mode option and uses it as the source IP address of the packet that the CPE will forward through its LAN interface until the packet reaches its final destination.

The destination IP address, port, and protocol are retrieved from a binding entry maintained by the CPE.

4.4.3. Processing Subsequent Outgoing/Incoming Non-SYNs

The required information to rewrite non-SYN packets that match an existing binding entry, is retrieved from the Binding Information Bases (BIB) maintained by the CPE and the concentrator (see Section 4.3). The MPTCP proxy may decide at any time to create or terminate subflows associated to an MPTCP connection. When a packet arrives, its content is transported over one of the subflows of a bound MPTCP connection.

Non-SYN messages exchanged in the context of an existing subflow and all messages for non-initial subflows do not include the PM option.

4.4.4. Handling TCP RST Messages

RST messages may be received from the LAN side of the CPE or by the concentrator in its Internet-facing interface. When the CPE or the concentrator receive a TCP RST matching an existing entry, it MUST apply the FASTCLOSE procedure defined in Section 3.5 of [RFC6824]) to terminate the MPTCP connection and the associated subflows. The transport coordinates of the FASTCLOSE messages are set according to the information maintained in the binding table.

The CPE and the concentrator SHOULD wait for 4 minutes before deleting the session and removing any state associated with it if no packets are received during that 4-minute timeout [RFC7857].

5. Processing UDP Traffic

This document leverages the ability to create MPTCP connections on the CPE/concentrator to also carry data conveyed in UDP datagrams. A UDP flow can be defined as a series of UDP packets that have the same source and destination address and ports. Upon receipt of the first packet of such a flow, a binding entry (Section 4.3) is created to map this flow onto an MPTCP connection between the CPE and the concentrator. All the subsequent UDP segments of this UDP flow are transported over that MPTCP connection. The MPTCP connection is released when no traffic is exchanged for this flow (Section 5.1.3).

5.1. Behavior

From an application standpoint, there may be a value to distribute UDP datagrams among available network attachments for the sake of network resource optimization, for example. This document uses MPTCP features to control how UDP datagrams are distributed among existing network attachments. The data carried in UDP datagrams belonging to a given UDP flow are therefore transported in an MPTCP connection. An MPTCP connection is bound to one UDP flow. New MPTCP connections are created in order to handle additional UDP flows.

The management of MPTCP connections that are triggered by UDP datagrams follows the guidelines documented in [RFC6824].

The following sub-sections exclusively focus on the external behavior to achieve UDP to TCP conversion (Section 5.1.1), and vice versa (Section 5.1.2).

5.1.1. UDP to TCP Conversion

This function is applied to UDP traffic received by the CPE from the LAN, and to UDP traffic received by the concentrator from one of its Internet-facing interfaces.

When the CPE (or the concentrator) receives a UDP datagram to be distributed over MPTCP subflows, it MUST check whether the packet matches an existing binding entry (Section 4.3).

If an entry is found, and the packet is to be placed on an existing subflow, the packet is processed according to the corresponding session entry. If an entry is found, but the packet should be placed on a new subflow, a session entry MUST be instantiated by the CPE for that outgoing packet. The information about the transport protocol (UDP, in this case) MUST also be included in this binding entry. In both cases, the CPE (or the concentrator) MUST proceed as follows:

1. Extract the payload and its length from the UDP datagram.

2. Send the length (as a 16 bits field in network byte order) followed by the payload of the UDP datagram over the bound MPTCP connection.

UDP packets that are received by the concentrator, but do not match an existing binding, MUST be silently dropped.

UDP packets that are received by the CPE, but do not match an existing binding, MUST be proceed as follows:

1. Instantiate a new binding entry for this outgoing packet. The information about the transport protocol (UDP, in this case) MUST also be included in this binding entry.

2. Initiate the MPTCP connection that will be used to carry the UDP datagrams of this flow towards the chosen concentrator. For this, the CPE MUST create a SYN segment containing the following information :

   * The MP_CAPABLE option and possibly other TCP options.

   * The payload contains the following information (in this order):

     + A PM option indicating the original source address and port if source address preservation is enabled.

     + A PM option indicating the original destination address and port.

     + The EOL TCP option.

     + The Length of the UDP payload in network byte order.

     + The payload of the UDP datagram.

When setting the source IP address, the destination IP address, and the IP address enclosed in the Plain Mode MPTCP option of the corresponding TCP packet, the same considerations as specified in Section 4.3 MUST be applied.

Whether one or multiple UDP payloads are included in the same TCP segment is implementation- and deployment-specific.

5.1.2. TCP to UDP Conversion

Upon receipt of a SYN segment containing the PM option specifying the UDP protocol, the concentrator MUST proceed as follows:

- Create a binding entry to map this MPTCP connection to a UDP flow (Section 4.3).
- Extract the destination, and possibly source, transport addresses from the PM option and complete the session entry with this information.
- Extract the UDP payload.
- Generate a UDP datagram with the corresponding IP addresses and ports and the UDP payload.

Upon receipt of a SYN segment containing the PM option specifying the UDP protocol, the CPE MUST proceed as follows:

- If no binding is found, the packet MUST be silently dropped.
- If a binding is found:
  - Extract the source (optionally, destination) transport addresses from the PM option.
  - Create a session entry to map this MPTCP connection to a UDP flow (Section 4.3).
  - Extract the UDP payload.
  - Generate a UDP datagram with the corresponding IP addresses and ports and the UDP payload.

Upon receipt of data over an MPTCP connection that is bound to a UDP flow, the 'Length' field is used to extract the UDP payloads from the bytestream and generates the corresponding UDP datagrams.

The concentrator (or the CPE) MUST follow the same procedure as mentioned in Section 4.3 for address and port rewriting purposes.

5.1.3. Terminating UDP-Triggered Subflows

UDP-triggered subflows SHOULD be terminated by an MPTCP endpoint (CPE or concentrator) if no UDP packet matching the corresponding binding entry is received for at least 5 minutes (see Section 4.3 of [RFC4787]). Consequently, the procedure in Section 4.4.4 MUST be

followed to terminate the MPTCP connection and the associated subflows. The transport coordinates of the FASTCLOSE messages are set according to the information maintained in the binding table.

5.2. Examples

A flow example is shown in Figure 6 to illustrate how TCP packets are generated to relay UDP datagrams using several subflows. Non-SYN messages that belong to a given subflow do not include any PM option. Also, this example shows how subsequent UDP datagrams of this flow are transported over the existing subflow or how a new subflow is created. In this example, the SYN segment issued to add a new subflow also includes data received in the original UDP datagram.

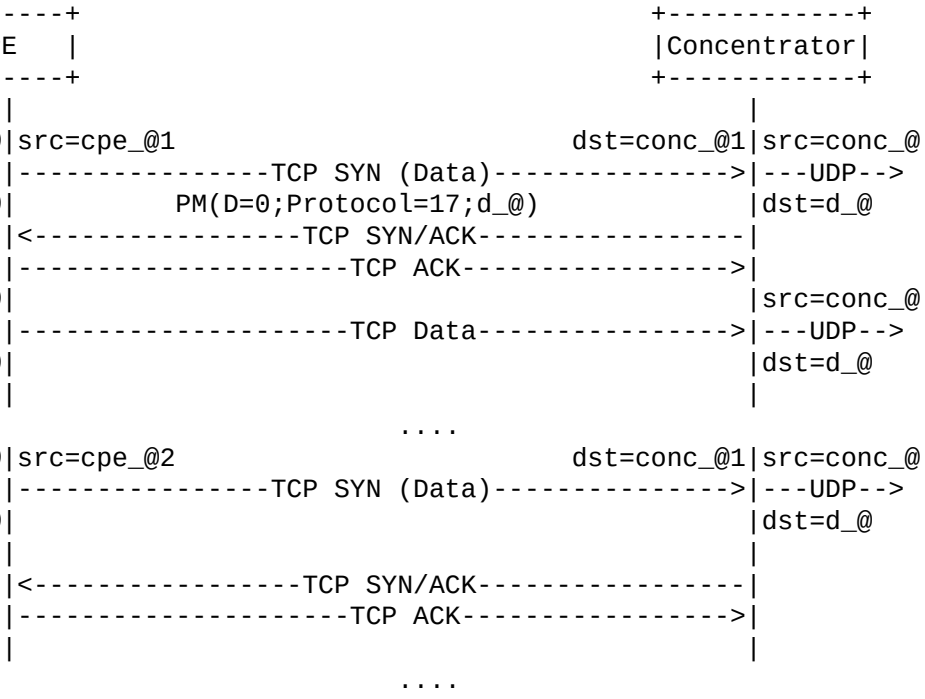


Figure 6: Distributing UDP packets over multiple paths (1)

Figure 7 shows an example of UDP datagrams that are transported over MPTCP subflows. Unlike the previous example, additional subflows to transport UDP datagrams of the same flow are established in advance between the CPE and the concentrator.

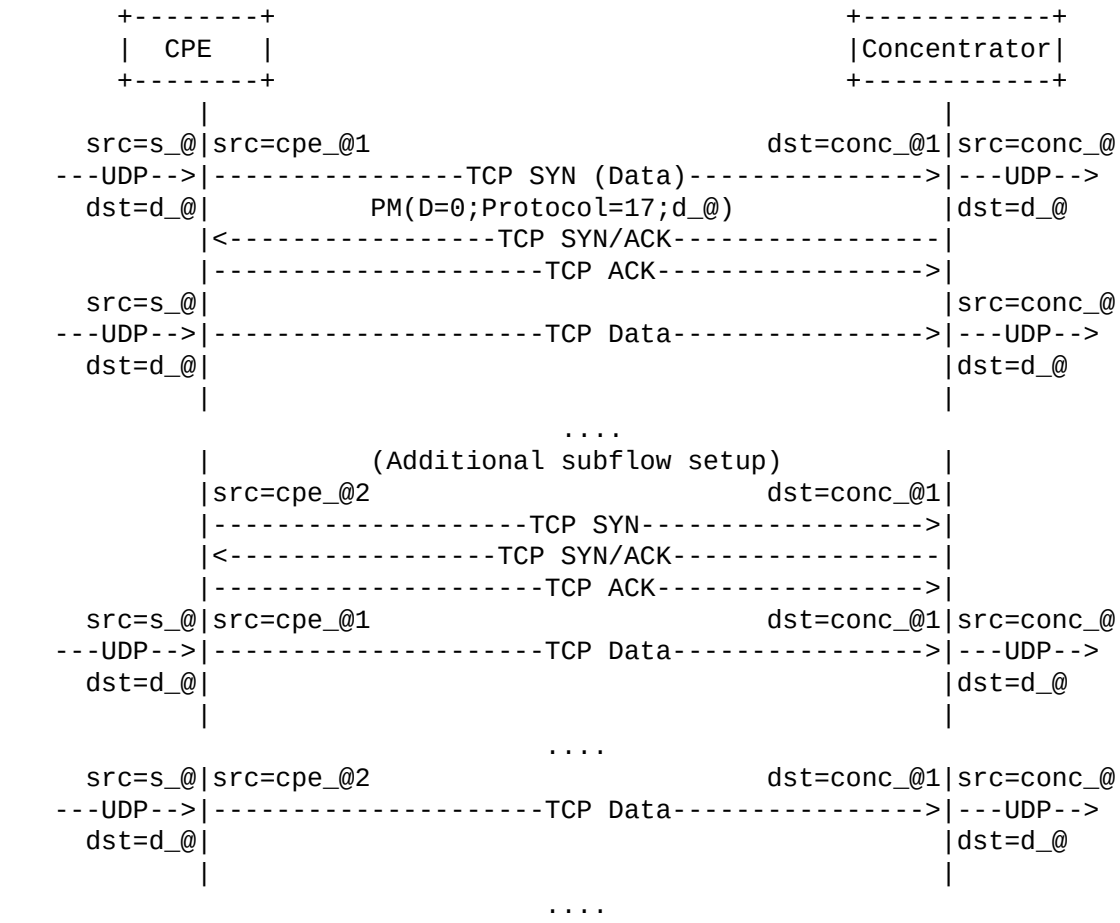


Figure 7: Distributing UDP packets over multiple paths (2)

5.3. Fragmentation & Reassembly Considerations

The subsequent UDP/TCP header swapping introduced in Section 5.1 represents an overhead that is equal to the difference between TCP and UDP header sizes. To avoid fragmentation when processing large UDP datagrams, it is RECOMMENDED to increase the MTU of all links between the CPE and the concentrator to accommodate this overhead.

Nevertheless, in deployments where increasing the MTU of all links is not possible for some reason, the CPE and the concentrator SHOULD be configurable to enable/disable fragmentation and reassembly of UDP datagrams. The decision to enable or disable this parameter is deployment-specific. This parameter is set to 'Disabled' by default.

If this configurable parameter is set to 'Disabled', large UDP datagrams that may thus be fragmented MUST NOT be forwarded along the

MPTCP connection, i.e., the bonding service MUST NOT be applied to such large packets.

If this configurable parameter is set to 'Enabled', the CPE and the concentrator MUST perform IPv4 fragmentation and reassembly for packets that exceed the link MTU. Concretely, IPv4 fragmentation MUST be performed once UDP/TCP header swapping is completed. Packet reassembly MUST occur before TCP/UDP header swapping. The behavior to adopt whenever the swapping of UDP/TCP headers leads to IPv4 fragmentation is as follows:

- Present the packet to the MPTCP proxy as per Section 5.1.1.
- Fragment the transformed packet (TCP), and then forward the fragments.

The remote MPTCP endpoint (CPE or concentrator) then adopts the following behavior:

- Reassemble the TCP packet,
- Present the packet to the MPTCP proxy as per Section 5.1.2.

In order to protect the CPE and the concentrator and minimize the risk of degrading the overall bonding service performance, dedicated resources SHOULD be reserved for handling fragments (e.g., by limiting the amount of resources to process out-of-order packets).

5.3.1. Receiving IPv4 Fragments on the Internet-Facing Interface of the Concentrator

The forwarding of an IPv4 packet received on the Internet-facing interface of the concentrator requires the IPv4 destination address and the transport-protocol destination port for binding lookup purposes. If the first packet received contains the transport-protocol information, the concentrator uses a cache and forwards the fragments unchanged (i.e., without reassembly). A description of such a caching algorithm is outside the scope of this document. If subsequent fragments arrive before the first fragment, the concentrator SHOULD queue these fragments till the first fragment is received.

The processing of the first fragment MUST follow the same procedure as in Section 5.1.1. The rewriting of the IP addresses of subsequent fragments MUST follow the instructions maintained in the binding table and the fragmentation cache. The MF (More Fragments) bit and 'Fragment offset' field MUST NOT be modified by the concentrator.

5.3.2. Receiving IPv4 Fragments from the LAN

If the first packet received contains the transport-protocol information, the CPE uses a cache and forwards the fragments unchanged (i.e., without reassembly). If subsequent fragments arrive before the first fragment, the concentrator SHOULD queue these fragments till the first fragment is received.

The processing of the first fragment MUST follow the same procedure as in Section 5.1.2. The rewriting of the IP addresses of subsequent fragments MUST follow the instructions maintained in the binding table and the fragmentation cache. The MF bit and 'Fragment offset' field MUST NOT be modified by the CPE.

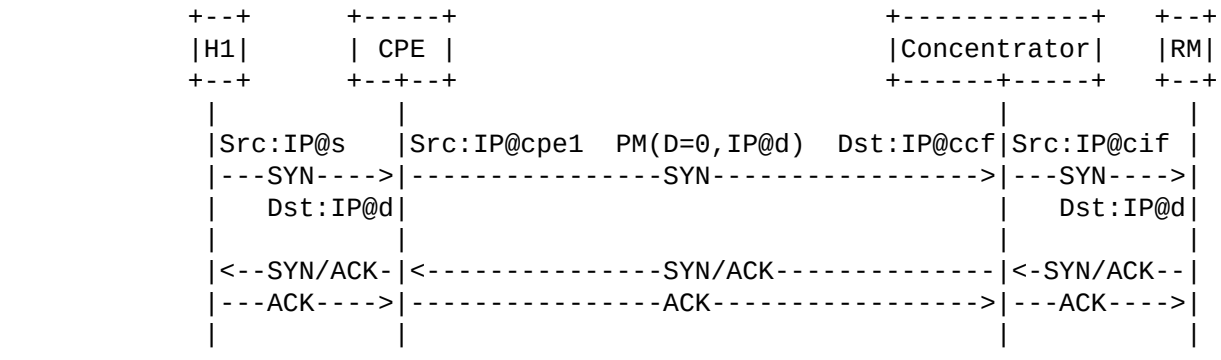
5.3.3. Distinct Address Families

If distinct address families are used in the UDP and MPTCP legs, fragmentation SHOULD be handled as described in Sections 4 and 5 of [RFC7915].

6. Deployment Scenarios

The Plain Transport Mode accommodates various deployment contexts such as:

IPv4 address sharing: Because of global IPv4 address depletion, optimization of the IPv4 address usage is mandatory, and this includes IPv4 addresses that are assigned by the concentrator at its Internet-facing interfaces (Figure 8). A pool of global IPv4 addresses is provisioned to the concentrator along with possible instructions about the address sharing ratio to apply (see Appendix B of [RFC6269]). Adequate forwarding policies are enforced so that traffic destined to an address of such pool is intercepted by the appropriate concentrator.



Legend:
ccf: Concentrator Customer-facing Interface
cif: Concentrator Internet-facing Interface

Figure 8: Example of Outgoing SYN without Source Address Preservation

IPv4 address 1:1 translation: For networks that do not face global IPv4 address depletion yet, the concentrator can be configured so that source IPv4 addresses of the CPE are replaced with other (public) IPv4 address. A pool of global IPv4 addresses is then provisioned to the concentrator for this purpose. Rewriting source IPv4 addresses may be used as a means to redirect incoming traffic towards the appropriate concentrator.

Source IPv6 address preservation: Some IPv6 deployments may require the preservation of the source IPv6 address (Figure 9). This model avoids the need for the concentrator to support ALGs to accommodate applications with IPv6 address referrals. In order to intercept incoming traffic, specific IPv6 routes are injected so that traffic is redirected towards the concentrator.

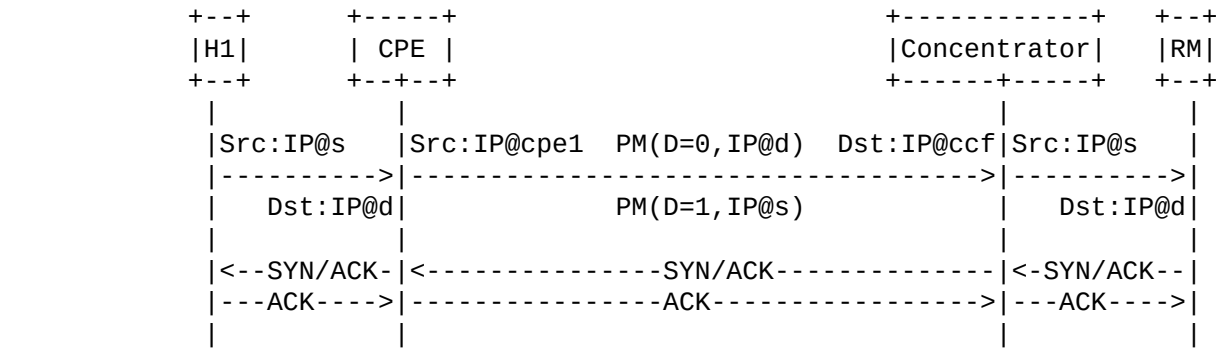


Figure 9: Example of Outgoing SYN with Source Address Preservation

IPv6 prefix sharing (NPTv6): Rewriting the source IPv6 prefix ([RFC6296]) may be needed to redirect incoming traffic towards the appropriate concentrator. A pool of IPv6 prefixes is then provisioned to the concentrator for this purpose.

Subflows of a given MPTCP connection can be associated to the same address family or may be established with different address families. Also, the plain transport mode applies regardless of the addressing scheme enforced by each CPE network attachment. In particular, the plain transport mode indifferently accommodates the following combinations.

LAN Leg	CPE-Concentrator Legs	Concentrator-RM Leg
IPv4	IPv4	IPv4
IPv4	IPv6	IPv4
IPv4	IPv6 & IPv4	IPv4
IPv6	IPv6	IPv6
IPv6	IPv4	IPv6
IPv6	IPv6 & IPv4	IPv6

Also, the CPE and the concentrator may be configured to preserve the same DSCP marking or enforce DSCP re-marking policies, and the plain transport mode described in this document fully respects these DSCP marking policies. Those considerations are deployment-specific.

7. Additional Considerations

7.1. Authorization

The Network Provider that manages the various network attachments (including the concentrators) may enforce authentication and authorization policies using appropriate mechanisms. For example, a non-exhaustive list of methods to achieve authorization is provided hereafter:

- The network provider may enforce a policy based on the International Mobile Subscriber Identity (IMSI) to verify that a user is allowed to benefit from the aggregation service. If that authorization fails, the PDP context /bearer won't be mounted. This method does not require any involvement from the concentrator.

- The network provider may enforce a policy based on Access Control Lists (ACLs), e.g., at the Broadband Network Gateway (BNG) to control the CPEs that are authorized to communicate with a concentrator. These ACLs may be installed as a result of RADIUS

exchanges, for instance. This method does not require any involvement from the concentrator.

- The concentrator may implement an Ident interface [RFC1413] to retrieve an identifier that will be used to assess whether that client is authorized to make use of the aggregation service. Ident exchanges will take place only when receiving the first subflow from a given source IP address.

- The concentrator may embed a RADIUS client that will solicit an AAA server to check whether connections received from a given source IP address are authorized or not.

A first safeguard against the misuse of the concentrator resources by illegitimate users (e.g., users with access networks that are not managed by the same operator owning the concentrator) is to reject MPTCP connections received on the Internet-facing interfaces. Only MPTCP connections received on the customer-facing interfaces of a concentrator will be accepted.

Because only the CPE is entitled to establish MPTCP connections with a concentrator, ACLs may be installed on the CPE to avoid that internal terminals issue MPTCP connections towards one of the concentrators.

7.2. Checksum Adjustment

Given that the TCP and UDP checksum covers the pseudo-header that contains the source and destination IP addresses, the checksum should be updated to reflect the change of these addresses. For the particular case of UDP/TCP conversion (Section 5), the UDP checksum can be computed from the TCP one and vice versa.

7.3. Logging

If the concentrator is used in global IPv4 address sharing environments, the logging recommendations discussed in Section 4 of [RFC6888] need to be considered. Security-related issues encountered in address sharing environments are documented in Section 13 of [RFC6269].

7.4. Middlebox Interference

The use of the Plain Transport Mode option is primarily meant for MPTCP designs that involve access networks managed by the same operator. Appropriate setup is required before MPTCP with the Plain Transport Mode option is activated, so that possible middleboxes

located in these access networks do not strip MPTCP signals, nor remove data contained in the SYN payload.

The plain transport mode may be deployed at large but some complications may arise, e.g., if an in-path middlebox removes the MPTCP option or data from the SYN payload. These complications not specific to the Plain Mode, and are encountered whenever MPTCP is deployed.

7.5. EPC Billing & Accounting

In case that one of MPTCP subflow between CPE and concentrator includes mobile (e.g., LTE, 3G, etc), billing and accounting of the traffic may be considered per subflow, per subscriber, or else. Since packets generated from/to the subscriber (CPE) are destined/ sourced to/from the concentrator, the EPC nodes may need to inspect, in some deployments, the destination/source address and/or port included in the plain mode option to check and make billing and accounting actions. Alternate deployment approaches may be adopted to avoid inspecting L3/4 information (e.g., rely on application-based filters, correlate flow characteristics retrieved using Policy and Charging Control (PCC) interfaces, etc.).

It is out of the scope of this document to make any recommendation in that area.

8. IANA Considerations

This document requests an MPTCP subtype code for this option:

- Plain Mode MPTCP Option

  NOTE: Implementations may use "0xe" subtype encoding for early deployment purposes in managed networks.

9. Security Considerations

MPTCP-related security threats are discussed in [RFC6181] and [RFC6824]. Additional considerations are discussed in the following sub-sections.

9.1. Privacy

The concentrator may have access to privacy-related information (e.g., IMSI, link identifier, subscriber credentials, etc.). The concentrator MUST NOT leak such sensitive information outside a local domain.

9.2. Denial-of-Service (DoS)

Means to protect the MPTCP concentrator against Denial-of-Service (DoS) attacks MUST be enabled. Such means include the enforcement of ingress filtering policies at the network boundaries [RFC2827].

In order to prevent the exhaustion of concentrator's resources, by establishing a large number of simultaneous subflows for each MPTCP connection, the administrator SHOULD limit the number of allowed subflows per CPE for a given connection. Means to protect against SYN flooding attacks MUST also be enabled ([RFC4987]).

Attacks that originate outside of the domain can be prevented if ingress filtering policies are enforced. Nevertheless, attacks from within the network between a host and a concentrator instance are yet another actual threat. Means to ensure that illegitimate nodes cannot connect to a network should be implemented.

9.3. Illegitimate Concentrator

Traffic theft is a risk if an illegitimate concentrator is inserted in the path. Indeed, inserting an illegitimate concentrator in the forwarding path allows traffic intercept and can therefore provide access to sensitive data issued by or destined to a host. To mitigate this threat, secure means to discover a concentrator should be enabled.

9.4. High Rate Reassembly

The CPE and the concentrator may perform packet reassembly. Some security-related issues are discussed in [RFC4963][RFC1858][RFC3128].

10. Acknowledgements

Many thanks to Chi Dung Phung, Mingui Zhang, Rao Shoaib, Yoshifumi Nishida, and Christoph Paasch for the comments.

Thanks to Ian Farrer, Mikael Abrahamsson, Alan Ford, Dan Wing, and Sri Gundavelli for the fruitful discussions in IETF#95 (Buenos Aires).

Special thanks to Pierrick Seite, Yannick Le Goff, Fred Klamm, and Xavier Grall for their valuable comments.

Thanks also to Olaf Schleusing, Martin Gysi, Thomas Zasowski, Andreas Burkhard, Silka Simmen, Sandro Berger, Michael Melloul, Jean-Yves Flahaut, Adrien Desportes, Gregory Detal, Benjamin David, Arun Srinivasan, and Raghavendra Mallya for the discussion.

11. References

11.1. Normative References

[proto_numbers] http://www.iana.org/assignments/protocol-numbers, "Protocol Numbers".

[RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, DOI 10.17487/RFC0793, September 1981, <http://www.rfc-editor.org/info/rfc793>.

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

[RFC4787] Audet, F., Ed. and C. Jennings, "Network Address Translation (NAT) Behavioral Requirements for Unicast UDP", BCP 127, RFC 4787, DOI 10.17487/RFC4787, January 2007, <http://www.rfc-editor.org/info/rfc4787>.

[RFC5382] Guha, S., Ed., Biswas, K., Ford, B., Sivakumar, S., and P. Srisuresh, "NAT Behavioral Requirements for TCP", BCP 142, RFC 5382, DOI 10.17487/RFC5382, October 2008, <http://www.rfc-editor.org/info/rfc5382>.

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

[RFC6890] Cotton, M., Vegoda, L., Bonica, R., Ed., and B. Haberman, "Special-Purpose IP Address Registries", BCP 153, RFC 6890, DOI 10.17487/RFC6890, April 2013, <http://www.rfc-editor.org/info/rfc6890>.

[RFC7857] Penno, R., Perreault, S., Boucadair, M., Ed., Sivakumar, S., and K. Naito, "Updates to Network Address Translation (NAT) Behavioral Requirements", BCP 127, RFC 7857, DOI 10.17487/RFC7857, April 2016, <http://www.rfc-editor.org/info/rfc7857>.

[RFC7915] Bao, C., Li, X., Baker, F., Anderson, T., and F. Gont, "IP/ICMP Translation Algorithm", RFC 7915, DOI 10.17487/RFC7915, June 2016, <http://www.rfc-editor.org/info/rfc7915>.

11.2. Informative References

[I-D.boucadair-mptcp-dhc]
Boucadair, M., Jacquenet, C., and T. Reddy, "DHCP Options for Network-Assisted Multipath TCP (MPTCP)", draft-boucadair-mptcp-dhc-05 (work in progress), May 2016.

[I-D.zhang-gre-tunnel-bonding]
Leymann, N., Heidemann, C., Zhang, M., Sarikaya, B., and M. Cullen, "Huawei's GRE Tunnel Bonding Protocol", draft-zhang-gre-tunnel-bonding-03 (work in progress), May 2016.

[RFC1413] St. Johns, M., "Identification Protocol", RFC 1413, DOI 10.17487/RFC1413, February 1993, <http://www.rfc-editor.org/info/rfc1413>.

[RFC1701] Hanks, S., Li, T., Farinacci, D., and P. Traina, "Generic Routing Encapsulation (GRE)", RFC 1701, DOI 10.17487/RFC1701, October 1994, <http://www.rfc-editor.org/info/rfc1701>.

[RFC1858] Ziemba, G., Reed, D., and P. Traina, "Security Considerations for IP Fragment Filtering", RFC 1858, DOI 10.17487/RFC1858, October 1995, <http://www.rfc-editor.org/info/rfc1858>.

[RFC1928] Leech, M., Ganis, M., Lee, Y., Kuris, R., Koblas, D., and L. Jones, "SOCKS Protocol Version 5", RFC 1928, DOI 10.17487/RFC1928, March 1996, <http://www.rfc-editor.org/info/rfc1928>.

[RFC2473] Conta, A. and S. Deering, "Generic Packet Tunneling in IPv6 Specification", RFC 2473, DOI 10.17487/RFC2473, December 1998, <http://www.rfc-editor.org/info/rfc2473>.

[RFC2827] Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", BCP 38, RFC 2827, DOI 10.17487/RFC2827, May 2000, <http://www.rfc-editor.org/info/rfc2827>.

[RFC3128] Miller, I., "Protection Against a Variant of the Tiny Fragment Attack (RFC 1858)", RFC 3128, DOI 10.17487/RFC3128, June 2001, <http://www.rfc-editor.org/info/rfc3128>.

[RFC4963] Heffner, J., Mathis, M., and B. Chandler, "IPv4 Reassembly Errors at High Data Rates", RFC 4963, DOI 10.17487/RFC4963, July 2007, <http://www.rfc-editor.org/info/rfc4963>.

[RFC4987] Eddy, W., "TCP SYN Flooding Attacks and Common Mitigations", RFC 4987, DOI 10.17487/RFC4987, August 2007, <http://www.rfc-editor.org/info/rfc4987>.

[RFC6181] Bagnulo, M., "Threat Analysis for TCP Extensions for Multipath Operation with Multiple Addresses", RFC 6181, DOI 10.17487/RFC6181, March 2011, <http://www.rfc-editor.org/info/rfc6181>.

[RFC6269] Ford, M., Ed., Boucadair, M., Durand, A., Levis, P., and P. Roberts, "Issues with IP Address Sharing", RFC 6269, DOI 10.17487/RFC6269, June 2011, <http://www.rfc-editor.org/info/rfc6269>.

[RFC6296] Wasserman, M. and F. Baker, "IPv6-to-IPv6 Network Prefix Translation", RFC 6296, DOI 10.17487/RFC6296, June 2011, <http://www.rfc-editor.org/info/rfc6296>.

[RFC6888] Perreault, S., Ed., Yamagata, I., Miyakawa, S., Nakagawa, A., and H. Ashida, "Common Requirements for Carrier-Grade NATs (CGNs)", BCP 127, RFC 6888, DOI 10.17487/RFC6888, April 2013, <http://www.rfc-editor.org/info/rfc6888>.

[RFC6967] Boucadair, M., Touch, J., Levis, P., and R. Penno, "Analysis of Potential Solutions for Revealing a Host Identifier (HOST_ID) in Shared Address Deployments", RFC 6967, DOI 10.17487/RFC6967, June 2013, <http://www.rfc-editor.org/info/rfc6967>.

[RFC7413] Cheng, Y., Chu, J., Radhakrishnan, S., and A. Jain, "TCP Fast Open", RFC 7413, DOI 10.17487/RFC7413, December 2014, <http://www.rfc-editor.org/info/rfc7413>.

Authors' Addresses

Mohamed Boucadair
Orange
Rennes 35000
France

Email: mohamed.boucadair@orange.com

Christian Jacquenet
Orange
Rennes
France

Email: christian.jacquenet@orange.com

Denis Behaghel
OneAccess

Email: Denis.Behaghel@oneaccess-net.com

Stefano Secci
Universite Pierre et Marie Curie
Paris
France

Email: stefano.secci@lip6.fr

Wim Henderickx
Nokia/Alcatel-Lucent
Belgium

Email: wim.henderickx@alcatel-lucent.com

Robert Skog
Ericsson

Email: robert.skog@ericsson.com

Olivier Bonaventure
Tessares
Belgium

Email: olivier.bonaventure@tessares.net

Suresh Vinapamula
Juniper
1137 Innovation Way
Sunnyvale, CA 94089
USA

Email: Sureshk@juniper.net

SungHoon Seo
Korea Telecom
Seoul
Korea

Email: sh.seo@kt.com

Wouter Cloetens
SoftAtHome
Vaartdijk 3 701
3018 Wijgmaal
Belgium

Email: wouter.cloetens@softathome.com

Ullrich Meyer
Vodafone
Germany

Email: ullrich.meyer@vodafone.com

Luis M. Contreras
Telefonica
Spain

Email: luismiguel.contrerasmurillo@telefonica.com