Internet Engineering Task Force	M. Handley
Internet-Draft	C. Raiciu
Intended status: Experimental	University College London
Expires: April 23, 2010	M. Bagnulo
	Universidad Carlos III de Madrid
	October 20, 2009

TOC

Outgoing Packet Routing with MP-TCP
draft-handley-mptcp-routing-00.txt

Status of this Memo

Copyright Notice

Abstract

Multipath TCP extends the TCP protocol to allow multiple paths to be used simultaneously for the same TCP connection. The different paths

are typically provided using multiple IP addresses for the same end system, each address taken from a subnet that is routed differently. In this document we describe a set of conventions for how to ensure that outgoing packets are routed in a manner consistent with the network topology and constraints on use of that topology such as those imposed by ingress filtering on IP address prefixes.

Table of Contents

---

1. Requirements Language

TOC

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 (Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels," March 1997.) [1].

---

2. Introduction

TOC

Multipath TCP (Ford, A., Raiciu, C., and M. Handley, "TCP Extensions for Multipath Operation with Multiple Addresses," March 2010.) [2] is an extension to the regular TCP protocol to allow multiple subflows to be established between the same pair of end systems, and for a single

TCP connection to stripe its data across these subflows. The intended benefits are improved performance, robustness, and pooling of network capacity. In principle there are many ways to identify and distinguish the packets of these subflows, and to guide them towards different paths through the network. One simple way to do this is to use multiple IP addresses at each endpoint.

If a host is on a multi-homed network, or if it has multiple interfaces (e.g. 3G and WiFi on a smart phone), then each of these addresses can be routed via a different network provider giving path diversity. For incoming traffic to the multi-addressed host, conventional IP routing will guide packets to the correct network link. For outgoing traffic however, destination-based routing by itself is insufficient to ensure that packets are sent over the appropriate paths. Not only could this reduce the diversity of paths available, but ingress filtering by ISPs may cause inappropriately routed packets to be dropped. This document describes a set of conventions that multipath-capable end-systems can follow to maximize the probability that packets reach their destination and to ensure that multiple paths can in fact be utilized.

In the sections that follow, we will assume a particular model for how an end-system routing table should function. This is both a strawman and an idealized model, and it is not necessarily expected that hosts will directly implement this model. The intent though is to describe what we believe to be reasonable behavior rather than how to implement this behavior.

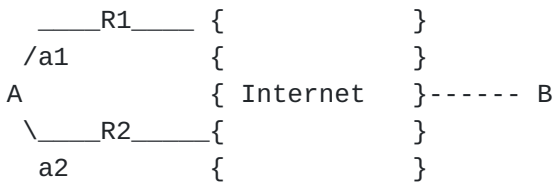
---

3. Multi-addressed Hosts

TOC

Consider a host that has more than one network interface and that wishes to send and receive regular TCP flows over these interfaces. To be able to receive packets on all of these interfaces, they are given IP addresses from different IP subnets. These subnets will be advertised via IP routing so that they are reachable by the host's intended correspondents.

For outgoing packets a host typically has a host routing table that defines which prefixes should be routed via each possible next hop router, and the choice of next hop router then determines the network interface used to reach that router. For a multi-addressed host this can be problematic. For the sake of example, consider the following network topology:



Host A is directly multihomed to two ISPs, and is given address a1 by one ISP and a2 by the second ISP. Such a scenario might occur when A is a smart phone connected simultaneously via a 3G network and via WiFi. It is communicating with server B which has a single network link and a single IP address.

If A initiates a TCP connection to B, A's IP stack will choose the next hop router based on the best path to B as determined by the host routing table (often this will be via a default route). If the application does not bind the local IP address, then if R1 is chosen as the next hop router then address a1 will be used as the source address for this connection, otherwise a2 will be used. Under such circumstances, the TCP connection functions correctly.

If B initiates a TCP connection to A and sends the SYN to address a1, then routing will route the incoming packet via R1 and hence to a1. If A's best route to B is via R1, then there is no problem. However, if A's best route to B is via R2, then what happens next depends on the host stack implementation. Two common models are in use:

* If A implements a strong host model, the connection will be rejected because the incoming packet arrived on the incorrect network interface.

* If A implements a weak host model, the connection will be accepted and the SYN/ACK from A to be will be sent via router R2, but with source address a1. As this address does not come from the address prefix allocated by ISP 2, then there is a good chance that ISP 2 will drop the packet in its ingress filter, believing the source address to be spoofed.

Clearly neither of these behaviors is desirable. As a result, configurations such as that shown above are generally not configured unless it is expected that host A will only act in the role of a client.

Unfortunately the configuration shown above is also the simplest case where Multipath TCP, using multiple addresses to distinguish its subflows, will gain any significant benefit. Not only must the configuration above work for a TCP flow that successfully negotiates multipath capability, but also it must work for regular TCP flows to and from that multi-addressed host.

In fact, for Multipath TCP to be effective, even as a client, modifications to the local host routing mechanism will be needed. Even if A initiates two subflows with B, addressed using a1 and a2 respectively, if the operating system determines the next-hop router (R1 or R2) purely using the host routing table, then only one outgoing path to B will be used. Suppose R1 is used. Not only does this fail to load-balance across the two outgoing paths, packets from the a2 subflow risk being dropped as spoofed by ISP 1's ingress filters.

To summarise: it does not make sense to configure current hosts with such an addressing scheme unless they are expected to only act as

clients. However, for Multipath TCP to be effective, such configurations will be necessary. Thus hosts implementing Multipath TCP will need to also implement modifications to the local host routing mechanism, so as to avoid the undesirable scenario above.

---

4. Idealized Host Routing Model

TOC

The idealized host routing table assumed in this document changes the model described above for hosts implementing MP-TCP. This model is also safe for hosts that do not implememt MP-TCP, so it may make sense to make it the default behavior on some operating systems, even if MP-TCP is not implemented or not configured.
The main change is simple, and corresponds to common routing behavior found in routers: an MP-TCP host MAY have more than one host routing table entry for the same IP prefix (default is just a special case of a very short prefix), so long as they specify different next hop routers. Each routing table entry MAY have an associated metric, where a lower metric indicates that routing table entry is preferred.
Packets from multipath and non-multipath flows are forwarded identically. The following procedure SHOULD be followed:

1. Identify the set of routing table entries that match the destination address. These main include default routes. Of these, eliminate all that do not have the longest prefix length.

2. If no route matches, drop the packet and inform TCP of the loss. MP-TCP may be able to re-send the packet's data to a different destination address.

3. If none of the routing table entries has a next hop on the same IP subnet as the source address TCP put in the packet, send the packet using the route with the lowest metric.

4. Otherwise at least one routing table entry has a next hop on the same IP subnet as the packet's source address. Of these routes, send the packet using the route with the lowest metric.

The motivation is that a packet should only ever be sent via a next hop that has a route to the destination, but where possible a packet should be sent via a subnet that is compatible with the source address in the packet. Sometimes though it may not be possible to do this, and we discuss these cases below.
An alternate behaviour for rule 3 is also acceptable, and corresponds to the strong host philsophy:

*If none of the routing table entries has a next hop on the same IP subnet as the source address TCP put in the packet, drop the packet and inform TCP of the loss.

This alternative rule only affects behavior in a corner case that can be regarded as either misconfiguration or routing failure (depending on whether or not the host runs a dynamic routing protocol), and so does not substantially affect the overall behavior.
This section has presented a strawman for how host routing should behave in an MP-TCP system. This behavior is not intended to be definitive; other host behaviors can be devised that will have the same or similar effects when paired with a multipath transport protocol. Rather, the intent of this section is to define baseline behavior within which we can then define how MP-TCP should behave.

4.1. Interaction with NATs

TOC

The existence of Network Address Translators (NATs) in the network does not change the forwarding behavior described above. However, if a NAT is present on one of the paths out of a site, it is important that a subflow continues to traverse that NAT for its entire lifetime, or else never traverses that NAT at all. Thus NATs provide an additional constraint on the host routing rules:
The routing of an existing MP-TCP subflow should not be affected by the subsequent establishment of additional subflows to the same destination.

---

5. MP-TCP Interaction with Host Routing

TOC

Having defined a strawman for how host routing should behave in a MP-TCP system, we can now define how MP-TCP should interact with that host routing mechanism.

5.1. TCP Active End-System Behaviour

TOC

When a regular TCP connection sends the first SYN packet to a destination, the application can either bind the socket to a local IP address or leave it unbound. If it is bound, the source address of the SYN is chosen by the application, and the TCP session is subsequently bound to this IP address. If the application leaves the source address

unbound, the TCP implementation typically looks at the routing table to determine the next-hop router, and chooses its local IP address to be the one from the subnet of the next-hop router. The TCP session is then bound to this dynamically chosen address, even if the host routing changes and packets are subsequently sent from a different interface.
When a multipath TCP connection sends the first SYN packet on the first subflow to a destination address, it SHOULD follow precisely the same procedure as for a regular TCP connection. This applies to both bound and unbound sockets.
When a multipath TCP wishes to establish an additional subflow to the same destination address, it MUST use a either a different local IP address or a different port from those of its existing subflows on that connection, otherwise the new subflow cannot be distinguished from the existing subflows. MP-TCP SHOULD choose a different source address, if one is available, as this maximises the path diversity for incoming traffic.
It might also be possible to establish an additional subflow using an existing source address, so a different route exists via a different nexthop router on that subnet. Such behavior is OPTIONAL, and requires additional state to be held that binds a subflow to a particular next hop router. The rules below assume a new source address is always used.
To establish a new subflow, MP-TCP will first examine the host routing table to determine the set of routes to that destination. The same basic procedure is followed, similar to that used by the host routing:

1. Identify the set of routing table entries that match the destination address. Of these eliminate all that do not have the longest prefix length.

2. If no route matches, the destination is currently unreachable, and the attempt to establish a new subflow fails. The MP-TCP implementation SHOULD retry with a different destination address if the other end has indicated more than one.

3. Take the set of local IP addresses already used by the subflows of this connection to this destination address. Eliminate from the remaining routing table entries those where the next-hop router is on the same IP subnet as any of these addresses.

4. If no route remains, there are no more local addresses to try to this destination address, and the attempt to establish a new subflow fails. The MP-TCP MAY retry with a different destination address if the other end has indicated more than one.

5. Of the remaining routes, choose the one with the lowest metric. Bind the subflow to the host's local IP address on the subnet of the next hop router from this route.

After a subflow has been established, the IP addresses it uses are fixed. The source address of all packets sent by an established subflow is set by TCP, and the packets are routed using the basic procedure in Section 4 (Idealized Host Routing Model).

5.2. Passive Open of MP-TCP Subflows

TOC

When a regular TCP passive listener receives a TCP SYN packet, if it chooses to accept the connection, the destination address in the SYN packet is bound to the connection. All subsequent packets the host sends on this connection will use this IP address as the source address. Routing for these outgoing packets is determined by the usual unipath forwarding mechanisms.

An MP-TCP passive listener behaves in basically the same way. If the subflow is accepted, the destination address of the incoming SYN packet binds the subflow to that address. All subsequent packets on that subflow will be sent with that source address.

With an active opener, the procedure in Section 5.1 (TCP Active End-System Behaviour) ensures subflows are only established with a source addresses for which there is an active (i.e., longest prefix match) route that leaves via a subnet with that source address. In other words, additional subflows will only be established when the host believes it can use the source address in a way that (from its point of view) is congruent with routing.

A passive listener does not have this luxury. The destination address of the incoming SYN packet determines the local IP address bound to the subflow. There are two distinct cases to consider, depending on the addresses in the SYN packet and the active routes on the listening host.

* Congruent Routing: The incoming SYN binds the connection to a local IP address, and there is an active route back to the destination via a next-hop router on that subnet. In this case the host routing is congruent with the local address chosen, so the forwarding rules present no problem.

* Incongruent Routing: The incoming SYN binds the connection to a local IP address, but there is no active route back to the destination via any next-hop router on that subnet. If there is no route to the destination at all, then the connection cannot be established. However, if there is a route via some other subnet then the OS has the option of using it, even though it knows the routing is not congruent with the addressing. This is less than ideal: although the OS knows that incoming packets can still reach it at the address in question, it does not have the control it would wish over outgoing packets, nor can it be sure that

outgoing packets will not be filtered by an ISP's ingress filtering. If the incoming SYN packet is attempting to establish a new subflow on an existing MP-TCP connection that already has a congruently routed and active subflow, then MP-TCP SHOULD reject the new subflow, as the connection is already functioning acceptably. If there is no congruent active subflow, the OS has the option of either dropping the connection or accepting it. If the OS chooses to accept the connection, it SHOULD also immediately attempt to establish a second subflow using the correct source address for the route to the destination.

Discussion: the non-congruent routing case might be considered to be a case of misconfigured routing on the host. It would also be reasonable behavior to fail to establish such TCP connections, multipath or otherwise. If the OS implementor chooses to allow such connections, then it might also be reasonable to pin the connection to the outgoing interface upon which the connection was successfully established. There is a strict tradeoff here between fragility in the presence of NATs and the ability for a host to re-route connections based on dynamic routing information. This problem is not specific to MP-TCP but occurs with regular TCP too. The behavior above chooses neither to drop not to pin, and seems a reasonable compromise in this tradeoff space.
Aside: depending on the final MP-TCP protocol spec, it may be possible for an MP-TCP passive lister to send a SYN/ACK from an IP address that is different from that in the initial SYN, and for the client to correctly bind the subflow to the TCP state. If this is possible, it solves the second scenario above. However it raises security questions, as it may make it simpler to hijack TCP sessions, and so we do not currently recommend such behavior.

---

6. Example Scenarios

TOC

The forwarding rules and MP-TCP behavior described above can be applied, no matter what the configuration of the MP-TCP host. However, it is worth examining several specific scenarios that are likely to be common to examine how the routing can be applied.

6.1. Multi-interface Host

TOC

A common scenario is one where a host has more than one interface over which it can route to the destination. This is typified by a smart phone (or other wireless device) that has both 3G and WiFi connectivity.

In such a case, it is expected that each interface is given a unique address from the subnet on which that interface resides. If each interface also has a route (of the same longest prefix length) that allows the host to reach the destination, then MP-TCP can be applied precisely as described in Section 4 (Idealized Host Routing Model) and Section 5 (MP-TCP Interaction with Host Routing).

6.2. Single-interface Host at Multihomed Site

TOC

Another common usage scenario is where a host has only a single interface, but it is located at a site that is multihomed to more than one ISP. For MP-TCP to balance in-bound traffic across the access links, the multiple links must be associated with different IP prefixes, and the hosts within the site must have more than one IP address.
There are two distinct scenarios to consider:

- Different next-hop IP routers on the host's LAN are associated with each prefix.

- The same physical router on the host's LAN is associated with all the prefixes.

For simplicity, it is worth considering these two cases separately.

6.2.1. Different Next-hop Routers

TOC

In this scenario the host has more than one IP address and logically resides on more than one subnet. It sees different outgoing routers on each of these subnets. These subnets behave as if they were different virtual interfaces from the point of view of routing, then MP-TCP can be applied precisely as described in Section 4 (Idealized Host Routing Model) and Section 5 (MP-TCP Interaction with Host Routing).
Although this scenario is quite limited, we believe it is also very common. For flexibility reasons, it appears than many data centers consist of a hierarchical L2 switch fabric on which the servers and routers reside.

6.2.2. Single Next-hop Router

TOC

In this scenario the host also has more than one IP address and logically resides on more than one subnet. However the topology is such

that only a single physical router is used to forward outgoing traffic. The actual routers used to connect to the organization's ISPs can be multiple IP hops away from the MP-TCP-capable server.
In such a scenario the host cannot itself directly control the path taken by the outgoing traffic. If such a host naively uses the forwarding rules from Section 4 (Idealized Host Routing Model) and Section 5 (MP-TCP Interaction with Host Routing), then outgoing traffic will not be balanced across the outgoing links, as it will all be forwarded within the site purely on the destination address in the packets. Perhaps worse, it is possible that packets with an IP address from one ISP are sent via the link from the other ISP, and that ISP implements ingress filtering and discards the packets.
We note that this scenario is actually worse with regular TCP, as such a host cannot retry with a different address. Thus such scenarios tend not to be configured in practice. However, it is clearly desirable for such sites to be able to take advantage of the benefits of MP-TCP; under such a scenario regular TCP must also work well. A number of possibilities seem to be available:

*Deploy source-address-based routing within the site for outgoing traffic. The normal MP-TCP host routing behavior can then be used.

*Configure more than one virtual-router instance on the physical router. From the host's point of view, the network then appears to be one with multiple routers, one for each subnet, so normal MP-TCP host routing behavior can be used. It then becomes the router's responsibility to ensure that the packets reach the correct outgoing routers. This is simple if the router is directly connected to the exit routes, or if MPLS is used within the site. Tunneling might also be used to direct the traffic to the correct exit router.

*Configure the hosts to tunnel their outgoing traffic to the exit routers. These tunnels would appear as virtual interfaces, so the normal MP-TCP host routing behavior can be used over these virtual interfaces.

*Use IP loose source routing to direct the traffic via the correct exit router. This would require a configuration change on the hosts. In addition, the LSRR option frequently causes traffic to be dropped in firewalls. Thus if this option were used, it would be advisable for the site exit routers to strip the option before forwarding off-site.

It is not yet clear whether some of these options are preferable to others. It is likely that different solutions may make most sense at different sites. Some sites might even find it simplest to change the

topology so that the existing routers are on the same L2 infrastructure as the MP-TCP hosts.

---

7. IPv6 Considerations

TOC

The descriptions above are intended to be agnostic as to whether IPv4 or IPv6 is used. However, IPv6 adds some additional complexity.
In IPv6, router advertisement messages are sent using link-local IPv6 addresses. Thus even though a host may have a globally routable address on an interface, and may know that this interface corresponds to a particular IPv6 subnet, the router's address in the host routing table is not useful to identify the subnet address and hence to determine the choice of the host's routable address.
The solution to this problem is for the host to maintain a binding table that maps the router's host address to the subnet's routable prefix. This binding table MAY be filled in when the host receives a router advertisement message from the router indicating the subnet prefix.
We note that this slightly overloads the purpose of a router advertisement message, to indicate that this router is a valid next hop for packets sourced from this prefix. This does not seem to be a significant departure from current practice, but it is possible that it may change the outgoing routing on existing deployments.

---

8. Security Considerations

TOC

This document discusses the binding of TCP and MP-TCP connections to IP addresses, which has the potential to change the way traffic is routed in networks. This does not introduce any new security risks per-se, but any change to how traffic is routed might cause network administrators assumptions about where traffic flows to be incorrect. However, the traffic only flows via routers for which the hosts have route table entries, so the emphasis for administrators is to ensure that host routing is configured in a way that matches security assumptions.
The use of network-based mechansims to route outgoing traffic might open up new avenues for attack. This document does not discuss these mechanisms in sufficient detail to merit a discussion of their security or other properties here.

---

9. Normative References

TOC

[1]

	Bradner, S., “Key words for use in RFCs to Indicate Requirement Levels,” BCP 14, RFC 2119, March 1997 (TXT, HTML, XML).
[2]	Ford, A., Raiciu, C., and M. Handley, “TCP Extensions for Multipath Operation with Multiple Addresses,” draft-ford-mptcp-multiaddressed-03 (work in progress), March 2010 (TXT).

Authors' Addresses

TOC

	Mark Handley
	University College London
	Department of Computer Science
	Gower St.
	London WC1E 6BT
	UK
Phone:	+44 20 7679 7296
Email:	m.handley@cs.ucl.ac.uk
	Costin Raiciu
	University College London
	Department of Computer Science
	Gower St.
	London WC1E 6BT
	UK
Phone:	+44 20 7679 3666
Email:	C.Raiciu@cs.ucl.ac.uk
	Marcelo Bagnulo
	Universidad Carlos III de Madrid
	Av. Universidad 30
	Leganes, Madrid 28911
	Spain
Phone:	+34 91 6248814
Email:	marcelo@it.uc3m.es