TAPS Working Group
Internet-Draft
Intended status: Informational
Expires: 13 January 2022

A. Brunstrom, Ed.
Karlstad University
T. Pauly, Ed.
Apple Inc.
T. Enghardt
Netflix
K-J. Grinnemo
Karlstad University
T. Jones
University of Aberdeen
P. Tiesel
SAP SE
C. Perkins
University of Glasgow
M. Welzl
University of Oslo
12 July 2021

Implementing Interfaces to Transport Services
draft-ietf-taps-impl-10

Abstract

The Transport Services system enables applications to use transport protocols flexibly for network communication and defines a protocol-independent Transport Services Application Programming Interface (API) that is based on an asynchronous, event-driven interaction pattern. This document serves as a guide to implementation on how to build such a system.

Status of This Memo

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

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

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

This Internet-Draft will expire on 13 January 2022.

Copyright Notice

Copyright (c) 2021 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction

The Transport Services architecture [I-D.ietf-taps-arch] defines a system that allows applications to use transport networking protocols flexibly. The interface such a system exposes to applications is defined as the Transport Services API [I-D.ietf-taps-interface]. This API is designed to be generic across multiple transport protocols and sets of protocols features.

This document serves as a guide to implementation on how to build a system that provides a Transport Services API. It is the job of an implementation of a Transport Services system to turn the requests of an application into decisions on how to establish connections, and how to transfer data over those connections once established. The terminology used in this document is based on the Architecture [I-D.ietf-taps-arch].

2. Implementing Connection Objects

The connection objects that are exposed to applications for Transport Services are:

* the Preconnection, the bundle of Properties that describes the application constraints on, and preferences for, the transport;

* the Connection, the basic object that represents a flow of data as Messages in either direction between the Local and Remote Endpoints;

* and the Listener, a passive waiting object that delivers new Connections.

Preconnection objects should be implemented as bundles of properties that an application can both read and write. A Preconnection object influences a Connection only at one point in time: when the Connection is created. Connection objects represent the interface between the application and the implementation to manage transport state, and conduct data transfer. During the process of establishment (Section 4), the Connection will not be bound to a specific transport protocol instance, since multiple candidate Protocol Stacks might be raced.

Once a Preconnection has been used to create an outbound Connection or a Listener, the implementation should ensure that the copy of the properties held by the Connection or Listener is not affected when the application makes changes to a Preconnection object. This may involve the implementation performing a deep-copy, copying the object with all the objects that it references.

Once the Connection is established, its interface maps actions and events to the details of the chosen Protocol Stack. For example, the same Connection object may ultimately represent a single instance of one transport protocol (e.g., a TCP connection, a TLS session over TCP, a UDP flow with fully-specified Local and Remote Endpoints, a DTLS session, a SCTP stream, a QUIC stream, or an HTTP/2 stream). The properties held by a Connection or Listener is independent of other connections that are not part of the same Connection Group.

Once Initate has been called, the Selection Properties and Endpoint information are immutable (i.e, an application is not able to later modify Selection Properties on the original Preconnection object). Listener objects are created with a Preconnection, at which point their configuration should be considered immutable by the implementation. The process of listening is described in Section 4.7.

3. Implementing Pre-Establishment

During pre-establishment the application specifies one or more Endpoints to be used for communication as well as protocol preferences and constraints via Selection Properties and, if desired, also Connection Properties. Generally, Connection Properties should be configured as early as possible, because they can serve as input to decisions that are made by the implementation (e.g., the Capacity Profile can guide usage of a protocol offering scavenger-type congestion control).

The implementation stores these properties as a part of the Preconnection object for use during connection establishment. For Selection Properties that are not provided by the application, the implementation must use the default values specified in the Transport Services API ([I-D.ietf-taps-interface]).

3.1. Configuration-time errors

The Transport Services system should have a list of supported protocols available, which each have transport features reflecting the capabilities of the protocol. Once an application specifies its Transport Properties, the transport system matches the required and prohibited properties against the transport features of the available protocols.

In the following cases, failure should be detected during pre-establishment:

* A request by an application for Protocol Properties that include requirements or prohibitions that cannot be satisfied by any of the available protocols. For example, if an application requires "Configure Reliability per Message", but no such feature is available in any protocol the host running the transport system on the host running the transport system this should result in an error, e.g., when SCTP is not supported by the operating system.

* A request by an application for Protocol Properties that are in conflict with each other, i.e., the required and prohibited properties cannot be satisfied by the same protocol. For example,

if an application prohibits "Reliable Data Transfer" but then requires "Configure Reliability per Message", this mismatch should result in an error.

To avoid allocating resources that are not finally needed, it is important that configuration-time errors fail as early as possible.

3.2. Role of system policy

The properties specified during pre-establishment have a close relationship to system policy. The implementation is responsible for combining and reconciling several different sources of preferences when establishing Connections. These include, but are not limited to:

1. Application preferences, i.e., preferences specified during the pre-establishment via Selection Properties.

2. Dynamic system policy, i.e., policy compiled from internally and externally acquired information about available network interfaces, supported transport protocols, and current/previous Connections. Examples of ways to externally retrieve policy-support information are through OS-specific statistics/measurement tools and tools that reside on middleboxes and routers.

3. Default implementation policy, i.e., predefined policy by OS or application.

In general, any protocol or path used for a connection must conform to all three sources of constraints. A violation of any of the layers should cause a protocol or path to be considered ineligible for use. For an example of application preferences leading to constraints, an application may prohibit the use of metered network interfaces for a given Connection to avoid user cost. Similarly, the system policy at a given time may prohibit the use of such a metered network interface from the application's process. Lastly, the implementation itself may default to disallowing certain network interfaces unless explicitly requested by the application and allowed by the system.

It is expected that the database of system policies and the method of looking up these policies will vary across various platforms. An implementation should attempt to look up the relevant policies for the system in a dynamic way to make sure it is reflecting an accurate version of the system policy, since the system's policy regarding the application's traffic may change over time due to user or administrative changes.

4. Implementing Connection Establishment

The process of establishing a network connection begins when an application expresses intent to communicate with a Remote Endpoint by calling Initiate. (At this point, any constraints or requirements the application may have on the connection are available from pre-establishment.) The process can be considered complete once there is at least one Protocol Stack that has completed any required setup to the point that it can transmit and receive the application's data.

Connection establishment is divided into two top-level steps: Candidate Gathering, to identify the paths, protocols, and endpoints to use, and Candidate Racing (see Section 4.2.2 of [I-D.ietf-taps-arch]), in which the necessary protocol handshakes are conducted so that the transport system can select which set to use.

This document structures the candidates for racing as a tree as terminological convention. While a a tree structure is not the only way in which racing can be implemented, it does ease the illustration of how racing works.

The most simple example of this process might involve identifying the single IP address to which the implementation wishes to connect, using the system's current default interface or path, and starting a TCP handshake to establish a stream to the specified IP address. However, each step may also differ depending on the requirements of the connection: if the endpoint is defined as a hostname and port, then there may be multiple resolved addresses that are available; there may also be multiple interfaces or paths available, other than the default system interface; and some protocols may not need any transport handshake to be considered "established" (such as UDP), while other connections may utilize layered protocol handshakes, such as TLS over TCP.

Whenever an implementation has multiple options for connection establishment, it can view the set of all individual connection establishment options as a single, aggregate connection establishment. The aggregate set conceptually includes every valid combination of endpoints, paths, and protocols. As an example, consider an implementation that initiates a TCP connection to a hostname + port endpoint, and has two valid interfaces available (Wi-Fi and LTE). The hostname resolves to a single IPv4 address on the Wi-Fi network, and resolves to the same IPv4 address on the LTE network, as well as a single IPv6 address. The aggregate set of connection establishment options can be viewed as follows:

```
Aggregate [Endpoint: www.example.com:80] [Interface: Any]   [Protocol: TCP]
|-> [Endpoint: 192.0.2.1:80]       [Interface: Wi-Fi] [Protocol: TCP]
|-> [Endpoint: 192.0.2.1:80]       [Interface: LTE]   [Protocol: TCP]
|-> [Endpoint: 2001:DB8::1.80]     [Interface: LTE]   [Protocol: TCP]
```

Any one of these sub-entries on the aggregate connection attempt would satisfy the original application intent. The concern of this section is the algorithm defining which of these options to try, when, and in what order.

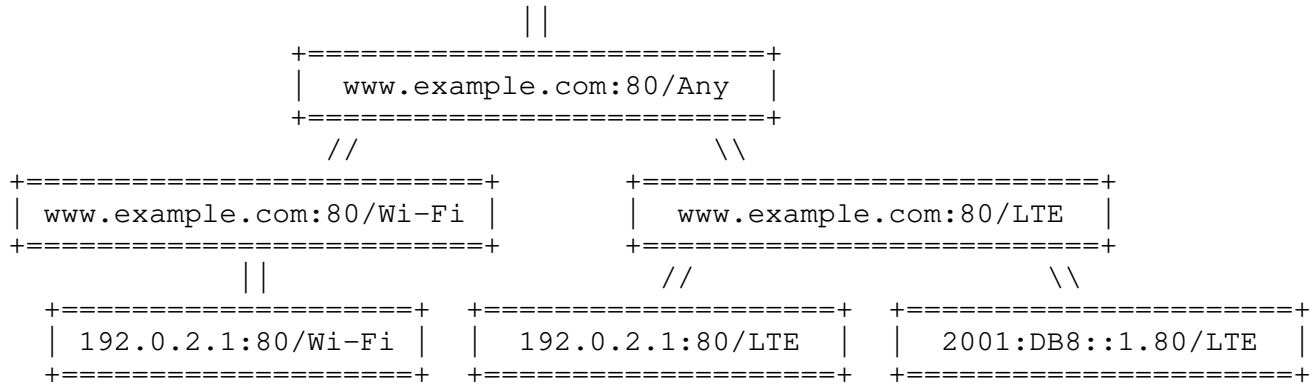
During Candidate Gathering, an implementation first excludes all protocols and paths that match a Prohibit or do not match all Require properties. Then, the implementation will sort branches according to Preferred properties, Avoided properties, and possibly other criteria.

4.1. Structuring Candidates as a Tree

As noted above, the considereration of multiple candidates in a gathering and racing process can be conceptually structured as a tree; this terminological convention is used throughout this document.

Each leaf node of the tree represents a single, coherent connection attempt, with an Endpoint, a Path, and a set of protocols that can directly negotiate and send data on the network. Each node in the tree that is not a leaf represents a connection attempt that is either underspecified, or else includes multiple distinct options. For example, when connecting on an IP network, a connection attempt to a hostname and port is underspecified, because the connection attempt requires a resolved IP address as its Remote Endpoint. In this case, the node represented by the connection attempt to the hostname is a parent node, with child nodes for each IP address. Similarly, an implementation that is allowed to connect using multiple interfaces will have a parent node of the tree for the decision between the paths, with a branch for each interface.

The example aggregate connection attempt above can be drawn as a tree by grouping the addresses resolved on the same interface into branches:



The rest of this section will use a notation scheme to represent this tree. The parent (or trunk) node of the tree will be represented by a single integer, such as "1". Each child of that node will have an integer that identifies it, from 1 to the number of children. That child node will be uniquely identified by concatenating its integer to it's parents identifier with a dot in between, such as "1.1" and "1.2". Each node will be summarized by a tuple of three elements: Endpoint, Path, and Protocol. The above example can now be written more succinctly as:

```
1 [www.example.com:80, Any, TCP]
  1.1 [www.example.com:80, Wi-Fi, TCP]
    1.1.1 [192.0.2.1:80, Wi-Fi, TCP]
  1.2 [www.example.com:80, LTE, TCP]
    1.2.1 [192.0.2.1:80, LTE, TCP]
    1.2.2 [2001:DB8::1.80, LTE, TCP]
```

When an implementation views this aggregate set of connection attempts as a single connection establishment, it only will use one of the leaf nodes to transfer data. Thus, when a single leaf node becomes ready to use, then the entire connection attempt is ready to use by the application. Another way to represent this is that every leaf node updates the state of its parent node when it becomes ready, until the trunk node of the tree is ready, which then notifies the application that the connection as a whole is ready to use.

A connection establishment tree may be degenerate, and only have a single leaf node, such as a connection attempt to an IP address over a single interface with a single protocol.

```
1 [192.0.2.1:80, Wi-Fi, TCP]
```

A parent node may also only have one child (or leaf) node, such as a when a hostname resolves to only a single IP address.

```
1 [www.example.com:80, Wi-Fi, TCP]
  1.1 [192.0.2.1:80, Wi-Fi, TCP]
```

4.1.1. Branch Types

There are three types of branching from a parent node into one or more child nodes. Any parent node of the tree must only use one type of branching.

4.1.1.1. Derived Endpoints

If a connection originally targets a single endpoint, there may be multiple endpoints of different types that can be derived from the original. The connection library creates an ordered list of the derived endpoints according to application preference, system policy and expected performance.

DNS hostname-to-address resolution is the most common method of endpoint derivation. When trying to connect to a hostname endpoint on a traditional IP network, the implementation should send DNS queries for both A (IPv4) and AAAA (IPv6) records if both are supported on the local interface. The algorithm for ordering and racing these addresses should follow the recommendations in Happy Eyeballs [RFC8305].

```
1 [www.example.com:80, Wi-Fi, TCP]
  1.1 [2001:DB8::1.80, Wi-Fi, TCP]
  1.2 [192.0.2.1:80, Wi-Fi, TCP]
  1.3 [2001:DB8::2.80, Wi-Fi, TCP]
  1.4 [2001:DB8::3.80, Wi-Fi, TCP]
```

DNS-Based Service Discovery [RFC6763] can also provide an endpoint derivation step. When trying to connect to a named service, the client may discover one or more hostname and port pairs on the local network using multicast DNS [RFC6762]. These hostnames should each be treated as a branch that can be attempted independently from other hostnames. Each of these hostnames might resolve to one or more addresses, which would create multiple layers of branching.

```
1 [term-printer._ipp._tcp.meeting.ietf.org, Wi-Fi, TCP]
  1.1 [term-printer.meeting.ietf.org:631, Wi-Fi, TCP]
    1.1.1 [31.133.160.18.631, Wi-Fi, TCP]
```

4.1.1.2. Alternate Paths

If a client has multiple network interfaces available to it, e.g., a mobile client with both Wi-Fi and Cellular connectivity, it can attempt a connection over any of the interfaces. This represents a branch point in the connection establishment. Similar to a derived endpoint, the interfaces should be ranked based on preference, system policy, and performance. Attempts should be started on one interface, and then on other interfaces successively after delays based on expected round-trip-time or other available metrics.

```
1 [192.0.2.1:80, Any, TCP]
  1.1 [192.0.2.1:80, Wi-Fi, TCP]
  1.2 [192.0.2.1:80, LTE, TCP]
```

This same approach applies to any situation in which the client is aware of multiple links or views of the network. Multiple Paths, each with a coherent set of addresses, routes, DNS server, and more, may share a single interface. A path may also represent a virtual interface service such as a Virtual Private Network (VPN).

The list of available paths should be constrained by any requirements or prohibitions the application sets, as well as system policy.

4.1.1.3. Protocol Options

Differences in possible protocol compositions and options can also provide a branching point in connection establishment. This allows clients to be resilient to situations in which a certain protocol is not functioning on a server or network.

This approach is commonly used for connections with optional proxy server configurations. A single connection might have several options available: an HTTP-based proxy, a SOCKS-based proxy, or no proxy. These options should be ranked and attempted in succession.

```
1 [www.example.com:80, Any, HTTP/TCP]
  1.1 [192.0.2.8:80, Any, HTTP/HTTP Proxy/TCP]
  1.2 [192.0.2.7:10234, Any, HTTP/SOCKS/TCP]
  1.3 [www.example.com:80, Any, HTTP/TCP]
    1.3.1 [192.0.2.1:80, Any, HTTP/TCP]
```

This approach also allows a client to attempt different sets of application and transport protocols that, when available, could provide preferable features. For example, the protocol options could involve QUIC [I-D.ietf-quic-transport] over UDP on one branch, and HTTP/2 [RFC7540] over TLS over TCP on the other:

```
1 [www.example.com:443, Any, Any HTTP]
  1.1 [www.example.com:443, Any, QUIC/UDP]
    1.1.1 [192.0.2.1:443, Any, QUIC/UDP]
  1.2 [www.example.com:443, Any, HTTP2/TLS/TCP]
    1.2.1 [192.0.2.1:443, Any, HTTP2/TLS/TCP]
```

Another example is racing SCTP with TCP:

```
1 [www.example.com:80, Any, Any Stream]
  1.1 [www.example.com:80, Any, SCTP]
    1.1.1 [192.0.2.1:80, Any, SCTP]
  1.2 [www.example.com:80, Any, TCP]
    1.2.1 [192.0.2.1:80, Any, TCP]
```

Implementations that support racing protocols and protocol options should maintain a history of which protocols and protocol options successfully established, on a per-network and per-endpoint basis (see Section 9.2). This information can influence future racing decisions to prioritize or prune branches.

4.1.2. Branching Order-of-Operations

Branch types must occur in a specific order relative to one another to avoid creating leaf nodes with invalid or incompatible settings. In the example above, it would be invalid to branch for derived endpoints (the DNS results for www.example.com) before branching between interface paths, since there are situations when the results will be different across networks due to private names or different supported IP versions. Implementations must be careful to branch in an order that results in usable leaf nodes whenever there are multiple branch types that could be used from a single node.

The order of operations for branching should be:

1. Alternate Paths

2. Protocol Options

3. Derived Endpoints

where a lower number indicates higher precedence and therefore higher placement in the tree. Branching between paths is the first in the list because results across multiple interfaces are likely not related to one another: endpoint resolution may return different results, especially when using locally resolved host and service names, and which protocols are supported and preferred may differ across interfaces. Thus, if multiple paths are attempted, the overall connection can be seen as a race between the available paths or interfaces.

Protocol options are next checked in order. Whether or not a set of protocol, or protocol-specific options, can successfully connect is generally not dependent on which specific IP address is used. Furthermore, the protocol stacks being attempted may influence or altogether change the endpoints being used. Adding a proxy to a connection's branch will change the endpoint to the proxy's IP address or hostname. Choosing an alternate protocol may also modify the ports that should be selected.

Branching for derived endpoints is the final step, and may have multiple layers of derivation or resolution, such as DNS service resolution and DNS hostname resolution.

For example, if the application has indicated both a preference for WiFi over LTE and for a feature only available in SCTP, branches will be first sorted accord to path selection, with WiFi at the top. Then, branches with SCTP will be sorted to the top within their subtree according to the properties influencing protocol selection. However, if the implementation has current cache information that SCTP is not available on the path over WiFi, there is no SCTP node in the WiFi subtree. Here, the path over WiFi will be tried first, and, if connection establishment succeeds, TCP will be used. So the Selection Property of preferring WiFi takes precedence over the Property that led to a preference for SCTP.

1. [www.example.com:80, Any, Any Stream]
1.1 [192.0.2.1:80, Wi-Fi, Any Stream]
1.1.1 [192.0.2.1:80, Wi-Fi, TCP]
1.2 [192.0.3.1:80, LTE, Any Stream]
1.2.1 [192.0.3.1:80, LTE, SCTP]
1.2.2 [192.0.3.1:80, LTE, TCP]

4.1.3. Sorting Branches

Implementations should sort the branches of the tree of connection options in order of their preference rank, from most preferred to least preferred. Leaf nodes on branches with higher rankings represent connection attempts that will be raced first. Implementations should order the branches to reflect the preferences expressed by the application for its new connection, including Selection Properties, which are specified in [I-D.ietf-taps-interface].

In addition to the properties provided by the application, an implementation may include additional criteria such as cached performance estimates, see Section 9.2, or system policy, see Section 3.2, in the ranking. Two examples of how Selection and Connection Properties may be used to sort branches are provided below:

* "Interface Instance or Type": If the application specifies an interface type to be preferred or avoided, implementations should accordingly rank the paths. If the application specifies an interface type to be required or prohibited, an implementation is expeceted to not include the non-conforming paths.

* "Capacity Profile": An implementation can use the Capacity Profile to prefer paths that match an application's expected traffic pattern. This match will use cached performance estimates, see Section 9.2:

  - Scavenger: Prefer paths with the highest expected available capacity, but minimising impact on other traffic, based on the observed maximum throughput;

  - Low Latency/Interactive: Prefer paths with the lowest expected Round Trip Time, based on observed round trip time estimates;

  - Low Latency/Non-Interactive: Prefer paths with a low expected Round Trip Time, but can tolerate delay variation;

  - Constant-Rate Streaming: Prefer paths that are expected to satisy the requested Stream Send or Stream Receive Bitrate, based on the observed maximum throughput;

  - Capacity-Seeking: Prefer adapting to paths to determine the highest available capacity, based on the observed maximum throughput.

Implementations process the Properties in the following order: Prohibit, Require, Prefer, Avoid. If Selection Properties contain any prohibited properties, the implementation should first purge branches containing nodes with these properties. For required properties, it should only keep branches that satisfy these requirements. Finally, it should order the branches according to the preferred properties, and finally use any avoided properties as a tiebreaker. When ordering branches, an implementation can give more weight to properties that the application has explicitly set, than to the properties that are default.

The available protocols and paths on a specific system and in a specific context can change; therefore, the result of sorting and the outcome of racing may vary, even when using the same Selection and Connection Properties. However, an implementation ought to provide a consistent outcome to applications, e.g., by preferring protocols and paths that are already used by existing Connections that specified similar Properties.

4.2. Candidate Gathering

The step of gathering candidates involves identifying which paths, protocols, and endpoints may be used for a given Connection. This list is determined by the requirements, prohibitions, and preferences of the application as specified in the Selection Properties.

4.2.1. Gathering Endpoint Candidates

Both Local and Remote Endpoint Candidates must be discovered during connection establishment. To support Interactive Connectivity Establishment (ICE) [RFC8445], or similar protocols that involve out-of-band indirect signalling to exchange candidates with the Remote Endpoint, it is important to query the set of candidate Local Endpoints, and provide the protocol stack with a set of candidate Remote Endpoints, before the Local Endpoint attempts to establish connections.

4.2.1.1. Local Endpoint candidates

The set of possible Local Endpoints is gathered. In the simple case, this merely enumerates the local interfaces and protocols, and allocates ephemeral source ports. For example, a system that has WiFi and Ethernet and supports IPv4 and IPv6 might gather four candidate Local Endpoints (IPv4 on Ethernet, IPv6 on Ethernet, IPv4 on WiFi, and IPv6 on WiFi) that can form the source for a transient.

If NAT traversal is required, the process of gathering Local Endpoints becomes broadly equivalent to the ICE candidate gathering phase (see Section 5.1.1. of [RFC8445]). The endpoint determines its server reflexive Local Endpoints (i.e., the translated address of a Local Endpoint, on the other side of a NAT, e.g via a STUN sever [RFC5389]) and relayed Local Endpoints (e.g., via a TURN server [RFC5766] or other relay), for each interface and network protocol. These are added to the set of candidate Local Endpoints for this connection.

Gathering Local Endpoints is primarily a local operation, although it might involve exchanges with a STUN server to derive server reflexive Local Endpoints, or with a TURN server or other relay to derive relayed Local Endpoints. However, it does not involve communication with the Remote Endpoint.

4.2.1.2. Remote Endpoint Candidates

The Remote Endpoint is typically a name that needs to be resolved into a set of possible addresses that can be used for communication. Resolving the Remote Endpoint is the process of recursively performing such name lookups, until fully resolved, to return the set of candidates for the Remote Endpoint of this connection.

How this resolution is done will depend on the type of the Remote Endpoint, and can also be specific to each Local Endpoint. A common case is when the Remote Endpoint is a DNS name, in which case it is resolved to give a set of IPv4 and IPv6 addresses representing that name. Some types of Remote Endpoint might require more complex resolution. Resolving the Remote Endpoint for a peer-to-peer connection might involve communication with a rendezvous server, which in turn contacts the peer to gain consent to communicate and retrieve its set of candidate Local Endpoints, which are returned and form the candidate remote addresses for contacting that peer.

Resolving the Remote Endpoint is not a local operation. It will involve a directory service, and can require communication with the Remote Endpoint to rendezvous and exchange peer addresses. This can expose some or all of the candidate Local Endpoints to the Remote Endpoint.

4.3. Candidate Racing

The primary goal of the Candidate Racing process is to successfully negotiate a protocol stack to an endpoint over an interface--to connect a single leaf node of the tree--with as little delay and as few unnecessary connections attempts as possible. Optimizing these two factors improves the user experience, while minimizing network load.

This section covers the dynamic aspect of connection establishment. The tree described above is a useful conceptual and architectural model. However, an implementation is unable to know the full tree before it is formed and many of the possible branches ultimately might not be used.

There are three different approaches to racing the attempts for different nodes of the connection establishment tree:

1. Simultaneous

2. Staggered

3. Failover

Each approach is appropriate in different use-cases and branch types. However, to avoid consuming unnecessary network resources, implementations should not use simultaneous racing as a default approach.

The timing algorithms for racing should remain independent across branches of the tree. Any timers or racing logic is isolated to a given parent node, and is not ordered precisely with regards to other children of other nodes.

4.3.1. Simultaneous

Simultaneous racing is when multiple alternate branches are started without waiting for any one branch to make progress before starting the next alternative. This means the attempts are effectively simultaneous. Simultaneous racing should be avoided by implementations, since it consumes extra network resources and establishes state that might not be used.

4.3.2. Staggered

Staggered racing can be used whenever a single node of the tree has multiple child nodes. Based on the order determined when building the tree, the first child node will be initiated immediately, followed by the next child node after some delay. Once that second child node is initiated, the third child node (if present) will begin after another delay, and so on until all child nodes have been initiated, or one of the child nodes successfully completes its negotiation.

Staggered racing attempts can proceed in parallel. Implementations should not terminate an earlier child connection attempt upon starting a secondary child.

If a child node fails to establish connectivity (as in Section 4.4.1) before the delay time has expired for the next child, the next child should be started immediately.

Staggered racing between IP addresses for a generic Connection should follow the Happy Eyeballs algorithm described in [RFC8305]. [RFC8421] provides guidance for racing when performing Interactive Connectivity Establishment (ICE).

Generally, the delay before starting a given child node ought to be based on the length of time the previously started child node is expected to take before it succeeds or makes progress in connection establishment. Algorithms like Happy Eyeballs choose a delay based on how long the transport connection handshake is expected to take. When performing staggered races in multiple branch types (such as racing between network interfaces, and then racing between IP addresses), a longer delay may be chosen for some branch types. For example, when racing between network interfaces, the delay should also take into account the amount of time it takes to prepare the network interface (such as radio association) and name resolution over that interface, in addition to the delay that would be added for a single transport connection handshake.

Since the staggered delay can be chosen based on dynamic information, such as predicted round-trip time, implementations should define upper and lower bounds for delay times. These bounds are implementation-specific, and may differ based on which branch type is being used.

4.3.3. Failover

If an implementation or application has a strong preference for one branch over another, the branching node may choose to wait until one child has failed before starting the next. Failure of a leaf node is determined by its protocol negotiation failing or timing out; failure of a parent branching node is determined by all of its children failing.

An example in which failover is recommended is a race between a protocol stack that uses a proxy and a protocol stack that bypasses the proxy. Failover is useful in case the proxy is down or misconfigured, but any more aggressive type of racing may end up unnecessarily avoiding a proxy that was preferred by policy.

4.4. Completing Establishment

The process of connection establishment completes when one leaf node of the tree has successfully completed negotiation with the Remote Endpoint, or else all nodes of the tree have failed to connect. The first leaf node to complete its connection is then used by the application to send and receive data.

Successes and failures of a given attempt should be reported up to parent nodes (towards the trunk of the tree). For example, in the following case, if 1.1.1 fails to connect, it reports the failure to 1.1. Since 1.1 has no other child nodes, it also has failed and reports that failure to 1. Because 1.2 has not yet failed, 1 is not considered to have failed. Since 1.2 has not yet started, it is started and the process continues. Similarly, if 1.1.1 successfully connects, then it marks 1.1 as connected, which propagates to the trunk node 1. At this point, the connection as a whole is considered to be successfully connected and ready to process application data.

```
1 [www.example.com:80, Any, TCP]
  1.1 [www.example.com:80, Wi-Fi, TCP]
    1.1.1 [192.0.2.1:80, Wi-Fi, TCP]
  1.2 [www.example.com:80, LTE, TCP]
...
```

If a leaf node has successfully completed its connection, all other attempts should be made ineligible for use by the application for the original request. New connection attempts that involve transmitting data on the network ought not to be started after another leaf node has already successfully completed, because the connection as a whole has now been established. An implementation may choose to let certain handshakes and negotiations complete in order to gather metrics to influence future connections. Keeping additional connections is generally not recommended since those attempts were slower to connect and may exhibit less desirable properties.

4.4.1. Determining Successful Establishment

Implementations may select the criteria by which a leaf node is considered to be successfully connected differently on a per-protocol basis. If the only protocol being used is a transport protocol with a clear handshake, like TCP, then the obvious choice is to declare that node "connected" when the last packet of the three-way handshake has been received. If the only protocol being used is an connectionless protocol, like UDP, the implementation may consider the node fully "connected" the moment it determines a route is present, before sending any packets on the network, see further Section 4.6.

For protocol stacks with multiple handshakes, the decision becomes more nuanced. If the protocol stack involves both TLS and TCP, an implementation could determine that a leaf node is connected after the TCP handshake is complete, or it can wait for the TLS handshake to complete as well. The benefit of declaring completion when the TCP handshake finishes, and thus stopping the race for other branches of the tree, is reduced burden on the network and Remote Endpoints from further connection attempts that are likely to be abandoned. On the other hand, by waiting until the TLS handshake is complete, an implementation avoids the scenario in which a TCP handshake completes quickly, but TLS negotiation is either very slow or fails altogether in particular network conditions or to a particular endpoint. To avoid the issue of TLS possibly failing, the implementation should not generate a Ready event for the Connection until TLS is established.

If all of the leaf nodes fail to connect during racing, i.e. none of the configurations that satisfy all requirements given in the Transport Properties actually work over the available paths, then the transport system should notify the application with an InitiateError event. An InitiateError event should also be generated in case the transport system finds no usable candidates to race.

4.5. Establishing multiplexed connections

Multiplexing several Connections over a single underlying transport connection requires that the Connections to be multiplexed belong to the same Connection Group (as is indicated by the application using the Clone call). When the underlying transport connection supports multi-streaming, the Transport System can map each Connection in the Connection Group to a different stream. Thus, when the Connections that are offered to an application by the Transport System are multiplexed, the Transport System may implement the establishment of a new Connection by simply beginning to use a new stream of an already established transport connection and there is no need for a connection establishment procedure. This, then, also means that there may not be any "establishment" message (like a TCP SYN), but the application can simply start sending or receiving. Therefore, when the Initiate action of a Transport System is called without Messages being handed over, it cannot be guaranteed that the other endpoint will have any way to know about this, and hence a passive endpoint's ConnectionReceived event may not be called upon an active endpoint's Inititate. Instead, calling the ConnectionReceived event may be delayed until the first Message arrives.

4.6. Handling connectionless protocols

While protocols that use an explicit handshake to validate a Connection to a peer can be used for racing multiple establishment attempts in parallel, connectionless protocols such as raw UDP do not offer a way to validate the presence of a peer or the usability of a Connection without application feedback. An implementation should consider such a protocol stack to be established as soon as the Transport Services system has selected a path on which to send data.

However, if a peer is not reachable over the network using the connectionless protocol, or data cannot be exchanged for any other reason, the application may want to attempt using another candidate Protocol Stack. The implementation should maintain the list of other candidate Protocol Stacks that were eligible to use.

4.7. Implementing listeners

When an implementation is asked to Listen, it registers with the system to wait for incoming traffic to the Local Endpoint. If no Local Endpoint is specified, the implementation should use an ephemeral port.

If the Selection Properties do not require a single network interface or path, but allow the use of multiple paths, the Listener object should register for incoming traffic on all of the network interfaces

or paths that conform to the Properties. The set of available paths can change over time, so the implementation should monitor network path changes, and change the registration of the Listener across all usable paths as appropriate. When using multiple paths, the Listener is generally expected to use the same port for listening on each.

If the Selection Properties allow multiple protocols to be used for listening, and the implementation supports it, the Listener object should support receiving inbound connections for each eligible protocol on each eligible path.

4.7.1. Implementing listeners for Connected Protocols

Connected protocols such as TCP and TLS-over-TCP have a strong mapping between the Local and Remote Endpoints (four-tuple) and their protocol connection state. These map into Connection objects. Whenever a new inbound handshake is being started, the Listener should generate a new Connection object and pass it to the application.

4.7.2. Implementing listeners for Connectionless Protocols

Connectionless protocols such as UDP and UDP-lite generally do not provide the same mechanisms that connected protocols do to offer Connection objects. Implementations should wait for incoming packets for connectionless protocols on a listening port and should perform four-tuple matching of packets to either existing Connection objects or the creation of new Connection objects. On platforms with facilities to create a "virtual connection" for connectionless protocols implementations should use these mechanisms to minimise the handling of datagrams intended for already created Connection objects.

4.7.3. Implementing listeners for Multiplexed Protocols

Protocols that provide multiplexing of streams into a single four-tuple can listen both for entirely new connections (a new HTTP/2 stream on a new TCP connection, for example) and for new sub-connections (a new HTTP/2 stream on an existing connection). If the abstraction of Connection presented to the application is mapped to the multiplexed stream, then the Listener should deliver new Connection objects in the same way for either case. The implementation should allow the application to introspect the Connection Group marked on the Connections to determine the grouping of the multiplexing.

5. Implementing Sending and Receiving Data

The most basic mapping for sending a Message is an abstraction of datagrams, in which the transport protocol naturally deals in discrete packets. Each Message here corresponds to a single datagram. Generally, these will be short enough that sending and receiving will always use a complete Message.

For protocols that expose byte-streams, the only delineation provided by the protocol is the end of the stream in a given direction. Each Message in this case corresponds to the entire stream of bytes in a direction. These Messages may be quite long, in which case they can be sent in multiple parts.

Protocols that provide the framing (such as length-value protocols, or protocols that use delimiters) may support Message sizes that do not fit within a single datagram. Each Message for framing protocols corresponds to a single frame, which may be sent either as a complete Message in the underlying protocol, or in multiple parts.

5.1. Sending Messages

The effect of the application sending a Message is determined by the top-level protocol in the established Protocol Stack. That is, if the top-level protocol provides an abstraction of framed messages over a connection, the receiving application will be able to obtain multiple Messages on that connection, even if the framing protocol is built on a byte-stream protocol like TCP.

5.1.1. Message Properties

* Lifetime: this should be implemented by removing the Message from the queue of pending Messages after the Lifetime has expired. A queue of pending Messages within the transport system implementation that have yet to be handed to the Protocol Stack can always support this property, but once a Message has been sent into the send buffer of a protocol, only certain protocols may support removing a message. For example, an implementation cannot remove bytes from a TCP send buffer, while it can remove data from a SCTP send buffer using the partial reliability extension [RFC8303]. When there is no standing queue of Messages within the system, and the Protocol Stack does not support the removal of a Message from the stack's send buffer, this property may be ignored.

* Priority: this represents the ability to prioritize a Message over other Messages. This can be implemented by the system re-ordering Messages that have yet to be handed to the Protocol Stack, or by

giving relative priority hints to protocols that support priorities per Message. For example, an implementation of HTTP/2 could choose to send Messages of different Priority on streams of different priority.

* Ordered: when this is false, this disables the requirement of in-order-delivery for protocols that support configurable ordering. When the protocol stack does not support configurable ordering, this property may be ignored.

* Safely Replayable: when this is true, this means that the Message can be used by a transport mechanism that might transfer it multiple times - e.g., as a result of racing multiple transports or as part of TCP Fast Open. Also, protocols that do not protect against duplicated messages, such as UDP (when used directly, without a protocol layered atop), can only be used with Messages that are Safely Replayable. When a transport system is permitted to replay messages, replay protection could be provided by the application.

* Final: when this is true, this means that the sender will not send any further messages. The Connection need not be closed (in case the Protocol Stack supports half-close operation, like TCP). Any messages sent after a Final message will result in a SendError.

* Corruption Protection Length: when this is set to any value other than "Full Coverage", it sets the minimum protection in protocols that allow limiting the checksum length (e.g. UDP-Lite). If the protocol stack does not support checksum length limitation, this property may be ignored.

* Reliable Data Transfer (Message): When true, the property specifies that the Message must be reliably transmitted. When false, and if unreliable transmission is supported by the underlying protocol, then the Message should be unreliably transmitted. If the underlying protocol does not support unreliable transmission, the Message should be reliably transmitted.

* Message Capacity Profile Override: When true, this expresses a wish to override the Generic Connection Property "Capacity Profile" for this Message. Depending on the value, this can, for example, be implemented by changing the DSCP value of the associated packet (note that the guidelines in Section 6 of [RFC7657] apply; e.g., the DSCP value should not be changed for different packets within a reliable transport protocol session or DCCP connection).

* No Fragmentation: When set, this property limits the message size to the Maximum Message Size Before Fragmentation or Segmentation (see Section 10.1.7 of [I-D.ietf-taps-interface]). Messages larger than this size generate an error. Setting this avoids transport-layer segmentation or network-layer fragmentation. When used with transports running over IP version 4 the Don't Fragment bit will be set to avoid on-path IP fragmentation ([RFC8304]).

5.1.2. Send Completion

The application should be notified whenever a Message or partial Message has been consumed by the Protocol Stack, or has failed to send. The time at which a Message is considered to have been consumed by the Protocol Stack may vary depending on the protocol. For example, for a basic datagram protocol like UDP, this may correspond to the time when the packet is sent into the interface driver. For a protocol that buffers data in queues, like TCP, this may correspond to when the data has entered the send buffer. The time at which a message has failed to send is after the Protocol Stack or the Transport Services implementation itself has not successfully sent the entire Message content or partial Message content on any open candidate connection; this may depend on protocol-specific timeouts.

5.1.3. Batching Sends

Since sending a Message may involve a context switch between the application and the transport system, sending patterns that involve multiple small Messages can incur high overhead if each needs to be enqueued separately. To avoid this, the application can indicate a batch of Send actions through the API. When this is used, the implementation can defer the processing of Messages until the batch is complete.

5.2. Receiving Messages

Similar to sending, Receiving a Message is determined by the top-level protocol in the established Protocol Stack. The main difference with Receiving is that the size and boundaries of the Message are not known beforehand. The application can communicate in its Receive action the parameters for the Message, which can help the implementation know how much data to deliver and when. For example, if the application only wants to receive a complete Message, the implementation should wait until an entire Message (datagram, stream, or frame) is read before delivering any Message content to the application. This requires the implementation to understand where messages end, either via a supplied deframer or because the top-level protocol in the established Protocol Stack preserves message

boundaries. If the top-level protocol only supports a byte-stream and no framers were supported, the application can control the flow of received data by specifying the minimum number of bytes of Message content it wants to receive at one time.

If a Connection finishes before a requested Receive action can be satisfied, the implementation should deliver any partial Message content outstanding, or if none is available, an indication that there will be no more received Messages.

5.3. Handling of data for fast-open protocols

Several protocols allow sending higher-level protocol or application data during their protocol establishment, such as TCP Fast Open [RFC7413] and TLS 1.3 [RFC8446]. This approach is referred to as sending Zero-RTT (0-RTT) data. This is a desirable feature, but poses challenges to an implementation that uses racing during connection establishment.

The amount of data that can be sent as 0-RTT data varies by protocol and can be queried by the application using the "Maximum Message Size Concurrent with Connection Establishment" Connection Property. An implementation can set this property according to the protocols that it will race based on the given Selection Properties when the application requests to establish a connection.

If the application has 0-RTT data to send in any protocol handshakes, it needs to provide this data before the handshakes have begun. When racing, this means that the data should be provided before the process of connection establishment has begun. If the application wants to send 0-RTT data, it must indicate this to the implementation by setting the "Safely Replayable" send parameter to true when sending the data. In general, 0-RTT data may be replayed (for example, if a TCP SYN contains data, and the SYN is retransmitted, the data will be retransmitted as well but may be considered as a new connection instead of a retransmission). Also, when racing connections, different leaf nodes have the opportunity to send the same data independently. If data is truly safely replayable, this should be permissible.

Once the application has provided its 0-RTT data, an implementation should keep a copy of this data and provide it to each new leaf node that is started and for which a 0-RTT protocol is being used.

It is also possible that protocol stacks within a particular leaf node use 0-RTT handshakes without any safely replayable application data. For example, TCP Fast Open could use a Client Hello from TLS as its 0-RTT data, shortening the cumulative handshake time.

0-RTT handshakes often rely on previous state, such as TCP Fast Open cookies, previously established TLS tickets, or out-of-band distributed pre-shared keys (PSKs). Implementations should be aware of security concerns around using these tokens across multiple addresses or paths when racing. In the case of TLS, any given ticket or PSK should only be used on one leaf node, since servers will likely reject duplicate tickets in order to prevent replays (see section-8.1 [RFC8446]). If implementations have multiple tickets available from a previous connection, each leaf node attempt can use a different ticket. In effect, each leaf node will send the same early application data, yet encoded (encrypted) differently on the wire.

6. Implementing Message Framers

Message Framers are functions that define simple transformations between application Message data and raw transport protocol data. A Framer can encapsulate or encode outbound Messages, and decapsulate or decode inbound data into Messages.

While many protocols can be represented as Message Framers, for the purposes of the Transport Services interface these are ways for applications or application frameworks to define their own Message parsing to be included within a Connection's Protocol Stack. As an example, TLS is exposed as a protocol natively supported by the Transport Services interface, even though it could also serve the purpose of framing data over TCP.

Most Message Framers fall into one of two categories:

* Header-prefixed record formats, such as a basic Type-Length-Value (TLV) structure

* Delimiter-separated formats, such as HTTP/1.1.

Common Message Framers can be provided by the Transport Services implementation, but an implementation ought to allow custom Message Framers to be defined by the application or some other piece of software. This section describes one possible interface for defining Message Framers as an example.

6.1. Defining Message Framers

A Message Framer is primarily defined by the code that handles events for a framer implementation, specifically how it handles inbound and outbound data parsing. The function that implements custom framing logic will be referred to as the "framer implementation", which may be provided by the Transport Services implementation or the application itself. The Message Framer refers to the object or function within the main Connection implementation that delivers events to the custom framer implementation whenever data is ready to be parsed or framed.

When a Connection establishment attempt begins, an event can be delivered to notify the framer implementation that a new Connection is being created. Similarly, a stop event can be delivered when a Connection is being torn down. The framer implementation can use the Connection object to look up specific properties of the Connection or the network being used that may influence how to frame Messages.

```
MessageFramer -> Start(Connection)
MessageFramer -> Stop(Connection)
```

When a Message Framer generates a "Start" event, the framer implementation has the opportunity to start writing some data prior to the Connection delivering its "Ready" event. This allows the implementation to communicate control data to the Remote Endpoint that can be used to parse Messages.

```
MessageFramer.MakeConnectionReady(Connection)
```

Similarly, when a Message Framer generates a "Stop" event, the framer implementation has the opportunity to write some final data or clear up its local state before the "Closed" event is delivered to the Application. The framer implementation can indicate that it has finished with this.

```
MessageFramer.MakeConnectionClosed(Connection)
```

At any time if the implementation encounters a fatal error, it can also cause the Connection to fail and provide an error.

```
MessageFramer.FailConnection(Connection, Error)
```

Should the framer implementation deem the candidate selected during racing unsuitable it can signal this by failing the Connection prior to marking it as ready. If there are no other candidates available, the Connection will fail. Otherwise, the Connection will select a different candidate and the Message Framer will generate a new "Start" event.

Before an implementation marks a Message Framer as ready, it can also dynamically add a protocol or framer above it in the stack. This allows protocols that need to add TLS conditionally, like STARTTLS [RFC3207], to modify the Protocol Stack based on a handshake result.

```
otherFramer := NewMessageFramer()
MessageFramer.PrependFramer(Connection, otherFramer)
```

A Message Framer might also choose to go into a passthrough mode once an initial exchange or handshake has been completed, such as the STARTTLS case mentioned above. This can also be useful for proxy protocols like SOCKS [RFC1928] or HTTP CONNECT [RFC7230]. In such cases, a Message Framer implementation can intercept sending and receiving of messages at first, but then indicate that no more processing is needed.

```
MessageFramer.StartPassthrough()
```

6.2. Sender-side Message Framing

Message Framers generate an event whenever a Connection sends a new Message.

```
MessageFramer -> NewSentMessage<Connection, MessageData, MessageContext, IsEndOfM
```

Upon receiving this event, a framer implementation is responsible for performing any necessary transformations and sending the resulting data back to the Message Framer, which will in turn send it to the next protocol. Implementations SHOULD ensure that there is a way to pass the original data through without copying to improve performance.

```
MessageFramer.Send(Connection, Data)
```

To provide an example, a simple protocol that adds a length as a header would receive the "NewSentMessage" event, create a data representation of the length of the Message data, and then send a block of data that is the concatenation of the length header and the original Message data.

6.3. Receiver-side Message Framing

In order to parse a received flow of data into Messages, the Message Framer notifies the framer implementation whenever new data is available to parse.

MessageFramer -> HandleReceivedData<Connection>

Upon receiving this event, the framer implementation can inspect the inbound data. The data is parsed from a particular cursor representing the unprocessed data. The application requests a specific amount of data it needs to have available in order to parse. If the data is not available, the parse fails.

MessageFramer.Parse(Connection, MinimumIncompleteLength, MaximumLength) -> (Data,

The framer implementation can directly advance the receive cursor once it has parsed data to effectively discard data (for example, discard a header once the content has been parsed).

To deliver a Message to the application, the framer implementation can either directly deliver data that it has allocated, or deliver a range of data directly from the underlying transport and simultaneously advance the receive cursor.

MessageFramer.AdvanceReceiveCursor(Connection, Length)
MessageFramer.DeliverAndAdvanceReceiveCursor(Connection, MessageContext, Length,
MessageFramer.Deliver(Connection, MessageContext, Data, IsEndOfMessage)

Note that "MessageFramer.DeliverAndAdvanceReceiveCursor" allows the framer implementation to earmark bytes as part of a Message even before they are received by the transport. This allows the delivery of very large Messages without requiring the implementation to directly inspect all of the bytes.

To provide an example, a simple protocol that parses a length as a header value would receive the "HandleReceivedData" event, and call "Parse" with a minimum and maximum set to the length of the header field. Once the parse succeeded, it would call "AdvanceReceiveCursor" with the length of the header field, and then call "DeliverAndAdvanceReceiveCursor" with the length of the body that was parsed from the header, marking the new Message as complete.

7. Implementing Connection Management

Once a Connection is established, the Transport Services system allows applications to interact with the Connection by modifying or inspecting Connection Properties. A Connection can also generate events in the form of Soft Errors.

The set of Connection Properties that are supported for setting and getting on a Connection are described in [I-D.ietf-taps-interface]. For any properties that are generic, and thus could apply to all protocols being used by a Connection, the Transport System should store the properties in storage common to all protocols, and notify all protocol instances in the Protocol Stack whenever the properties have been modified by the application. For protocol-specfic properties, such as the User Timeout that applies to TCP, the Transport System only needs to update the relevant protocol instance.

If an error is encountered in setting a property (for example, if the application tries to set a TCP-specific property on a Connection that is not using TCP), the action should fail gracefully. The application may be informed of the error, but the Connection itself should not be terminated.

The Transport Services implementation should allow protocol instances in the Protocol Stack to pass up arbitrary generic or protocol-specific errors that can be delivered to the application as Soft Errors. These allow the application to be informed of ICMP errors, and other similar events.

7.1. Pooled Connection

For protocols that employ request/response pairs and do not require in-order delivery of the responses, like HTTP, the transport implementation may distribute interactions across several underlying transport connections. For these kinds of protocols, implementations may hide the connection management and only expose a single Connection object and the individual requests/responses as messages. These Pooled Connections can use multiple connections or multiple streams of multi-streaming connections between endpoints, as long as all of these satisfy the requirements, and prohibitions specified in the Selection Properties of the Pooled Connection. This enables implementations to realize transparent connection coalescing, connection migration, and to perform per-message endpoint and path selection by choosing among these underlying connections.

7.2. Handling Path Changes

When a path change occurs, e.g., when the IP address of an interface changes or a new interface becomes available, the Transport Services implementation is responsible for notifying the Protocol Instance of the change. The path change may interrupt connectivity on a path for an active connection or provide an opportunity for a transport that supports multipath or migration to adapt to the new paths. Note that, from the Transport Services API point of view, migration is considered a part of multipath connectivity; it is just a limiting policy on multipath usage. If the "multipath" Selection Property is set to "Disabled", migration is disallowed.

For protocols that do not support multipath or migration, the Protocol Instances should be informed of the path change, but should not be forcibly disconnected if the previously used path becomes unavailable. There are many common user scenarios that can lead to a path becoming temporarily unavailable, and then recovering before the transport protocol reaches a timeout error. These are particularly common using mobile devices. Examples include: an Ethernet cable becoming unplugged and then plugged back in; a device losing a Wi-Fi signal while a user is in an elevator, and reattaching when the user leaves the elevator; and a user losing the radio signal while riding a train through a tunnel. If the device is able to rejoin a network with the same IP address, a stateful transport connection can generally resume. Thus, while it is useful for a Protocol Instance to be aware of a temporary loss of connectivity, the Transport Services implementation should not aggressively close connections in these scenarios.

If the Protocol Stack includes a transport protocol that supports multipath connectivity, the Transport Services implementation should also inform the Protocol Instance of potentially new paths that become permissible based on the "multipath" Selection Property and the "multipath-policy" Connection Property choices made by the application. A protocol can then establish new subflows over new paths while an active path is still available or, if migration is supported, also after a break has been detected, and should attempt to tear down subflows over paths that are no longer used. The Transport Services API's Connection Property "multipath-policy" allows an application to indicate when and how different paths should be used. However, detailed handling of these policies is still implementation-specific. For example, if the "multipath" Selection Property is set to "active", the decision about when to create a new path or to announce a new path or set of paths to the Remote Endpoint, e.g., in the form of additional IP addresses, is implementation-specific. If the Protocol Stack includes a transport protocol that does not support multipath, but does support migrating between paths, the update to the set of available paths can trigger the connection to be migrated.

In case of Pooled Connections Section 7.1, the Transport Services implementation may add connections over new paths to the pool if permissible based on the multipath policy and Selection Properties. In case a previously used path becomes unavailable, the transport system may disconnect all connections that require this path, but should not disconnect the pooled connection object exposed to the application. The strategy to do so is implementation-specific, but should be consistent with the behavior of multipath transports.

8. Implementing Connection Termination

With TCP, when an application closes a connection, this means that it has no more data to send (but expects all data that has been handed over to be reliably delivered). However, with TCP only, "close" does not mean that the application will stop receiving data. This is related to TCP's ability to support half-closed connections.

SCTP is an example of a protocol that does not support such half-closed connections. Hence, with SCTP, the meaning of "close" is stricter: an application has no more data to send (but expects all data that has been handed over to be reliably delivered), and will also not receive any more data.

Implementing a protocol independent transport system means that the exposed semantics must be the strictest subset of the semantics of all supported protocols. Hence, as is common with all reliable transport protocols, after a Close action, the application can expect

to have its reliability requirements honored regarding the data it has given to the Transport System, but it cannot expect to be able to read any more data after calling Close.

Abort differs from Close only in that no guarantees are given regarding data that the application has handed over to the Transport System before calling Abort.

As explained in Section 4.5, when a new stream is multiplexed on an already existing connection of a Transport Protocol Instance, there is no need for a connection establishment procedure. Because the Connections that are offered by the Transport System can be implemented as streams that are multiplexed on a transport protocol's connection, it can therefore not be guaranteed that one Endpoint's Initiate action provokes a ConnectionReceived event at its peer.

For Close (provoking a Finished event) and Abort (provoking a ConnectionError event), the same logic applies: while it is desirable to be informed when a peer closes or aborts a Connection, whether this is possible depends on the underlying protocol, and no guarantees can be given. With SCTP, the transport system can use the stream reset procedure to cause a Finish event upon a Close action from the peer [NEAT-flow-mapping].

9. Cached State

Beyond a single Connection's lifetime, it is useful for an implementation to keep state and history. This cached state can help improve future Connection establishment due to re-using results and credentials, and favoring paths and protocols that performed well in the past.

Cached state may be associated with different Endpoints for the same Connection, depending on the protocol generating the cached content. For example, session tickets for TLS are associated with specific endpoints, and thus should be cached based on a Connection's hostname Endpoint (if applicable). On the other hand, performance characteristics of a path are more likely tied to the IP address and subnet being used.

9.1. Protocol state caches

Some protocols will have long-term state to be cached in association with Endpoints. This state often has some time after which it is expired, so the implementation should allow each protocol to specify an expiration for cached content.

Examples of cached protocol state include:

* The DNS protocol can cache resolution answers (A and AAAA queries, for example), associated with a Time To Live (TTL) to be used for future hostname resolutions without requiring asking the DNS resolver again.

* TLS caches session state and tickets based on a hostname, which can be used for resuming sessions with a server.

* TCP can cache cookies for use in TCP Fast Open.

Cached protocol state is primarily used during Connection establishment for a single Protocol Stack, but may be used to influence an implementation's preference between several candidate Protocol Stacks. For example, if two IP address Endpoints are otherwise equally preferred, an implementation may choose to attempt a connection to an address for which it has a TCP Fast Open cookie.

Applications can request that a Connection Group maintain a separate cache for protocol state. Connections in the group will not use cached state from connections outside the group, and connections outside the group will not use state cached from connections inside the group. This may be necessary, for example, if application-layer identifiers rotate and clients wish to avoid linkability via trackable TLS tickets or TFO cookies.

9.2. Performance caches

In addition to protocol state, Protocol Instances should provide data into a performance-oriented cache to help guide future protocol and path selection. Some performance information can be gathered generically across several protocols to allow predictive comparisons between protocols on given paths:

* Observed Round Trip Time

* Connection Establishment latency

* Connection Establishment success rate

These items can be cached on a per-address and per-subnet granularity, and averaged between different values. The information should be cached on a per-network basis, since it is expected that different network attachments will have different performance characteristics. Besides Protocol Instances, other system entities may also provide data into performance-oriented caches. This could for instance be signal strength information reported by radio modems like Wi-Fi and mobile broadband or information about the battery-level of the device. Furthermore, the system may cache the observed maximum throughput on a path as an estimate of the available bandwidth.

An implementation should use this information, when possible, to influence preference between candidate paths, endpoints, and protocol options. Eligible options that historically had significantly better performance than others should be selected first when gathering candidates (see Section 4.2) to ensure better performance for the application.

The reasonable lifetime for cached performance values will vary depending on the nature of the value. Certain information, like the connection establishment success rate to a Remote Endpoint using a given protocol stack, can be stored for a long period of time (hours or longer), since it is expected that the capabilities of the Remote Endpoint are not changing very quickly. On the other hand, the Round Trip Time observed by TCP over a particular network path may vary over a relatively short time interval. For such values, the implementation should remove them from the cache more quickly, or treat older values with less confidence/weight.

[I-D.ietf-tcpm-2140bis] provides guidance about sharing of TCP Control Block information between connections on initialization.

10. Specific Transport Protocol Considerations

Each protocol that can run as part of a Transport Services implementation should have a well-defined API mapping. API mappings for a protocol apply most to Connections in which the given protocol is the "top" of the Protocol Stack. For example, the mapping of the "Send" function for TCP applies to Connections in which the application directly sends over TCP.

Each protocol has a notion of Connectedness. Possible values for Connectedness are:

* Connectionless. Connectionless protocols do not establish explicit state between endpoints, and do not perform a handshake during Connection establishment.

* Connected. Connected protocols establish state between endpoints, and perform a handshake during Connection establishment. The handshake may be 0-RTT to send data or resume a session, but bidirectional traffic is required to confirm connectedness.

* Multiplexing Connected. Multiplexing Connected protocols share properties with Connected protocols, but also explictly support opening multiple application-level flows. This means that they can support cloning new Connection objects without a new explicit handshake.

Protocols also define a notion of Data Unit. Possible values for Data Unit are:

* Byte-stream. Byte-stream protocols do not define any Message boundaries of their own apart from the end of a stream in each direction.

* Datagram. Datagram protocols define Message boundaries at the same level of transmission, such that only complete (not partial) Messages are supported.

* Message. Message protocols support Message boundaries that can be sent and received either as complete or partial Messages. Maximum Message lengths can be defined, and Messages can be partially reliable.

Below, terms in capitals with a dot (e.g., "CONNECT.SCTP") refer to the primitives with the same name in section 4 of [RFC8303]. For further implementation details, the description of these primitives in [RFC8303] points to section 3 of [RFC8303] and section 3 of [RFC8304], which refers back to the relevant specifications for each protocol. This back-tracking method applies to all elements of [RFC8923] (see appendix D of [I-D.ietf-taps-interface]): they are listed in appendix A of [RFC8923] with an implementation hint in the same style, pointing back to section 4 of [RFC8303].

This document defines the API mappings for protocols defined in [RFC8923]. Other protocol mappings can be provided as separate documents, following the mapping template Appendix A.

10.1. TCP

Connectedness: Connected

Data Unit: Byte-stream

API mappings for TCP are as follows:

Connection Object: TCP connections between two hosts map directly to Connection objects.

Initiate: CONNECT.TCP. Calling "Initiate" on a TCP Connection causes it to reserve a local port, and send a SYN to the Remote Endpoint.

InitiateWithSend: CONNECT.TCP with parameter "user message". Early safely replayable data is sent on a TCP Connection in the SYN, as TCP Fast Open data.

Ready: A TCP Connection is ready once the three-way handshake is complete.

InitiateError: Failure of CONNECT.TCP. TCP can throw various errors during connection setup. Specifically, it is important to handle a RST being sent by the peer during the handshake.

ConnectionError: Once established, TCP throws errors whenever the connection is disconnected, such as due to receiving a RST from the peer.

Listen: LISTEN.TCP. Calling "Listen" for TCP binds a local port and prepares it to receive inbound SYN packets from peers.

ConnectionReceived: TCP Listeners will deliver new connections once they have replied to an inbound SYN with a SYN-ACK.

Clone: Calling "Clone" on a TCP Connection creates a new Connection with equivalent parameters. These Connections, and Connections generated via later calls to "Clone" on one of them, form a Connection Group. To realize entanglement for these Connections, with the exception of "Connection Priority", changing a Connection Property on one of them must affect the Connection Properties of the others too. No guarantees of honoring the Connection Property "Connection Priority" are given, and thus it is safe for an implementation of a transport system to ignore this property. When it is reasonable to assume that Connections traverse the same path (e.g., when they share the same encapsulation), support for it can also experimentally be implemented using a congestion control coupling mechanism (see for example [TCP-COUPLING] or [RFC3124]).

Send: SEND.TCP. TCP does not on its own preserve Message boundaries. Calling "Send" on a TCP connection lays out the bytes on the TCP send stream without any other delineation. Any Message marked as Final will cause TCP to send a FIN once the Message has been completely written, by calling CLOSE.TCP immediately upon

successful termination of SEND.TCP. Note that transmitting a Message marked as Final should not cause the "Closed" event to be delivered to the application, as it will still be possible to receive data until the peer closes or aborts the TCP connection.

Receive: With RECEIVE.TCP, TCP delivers a stream of bytes without any Message delineation. All data delivered in the "Received" or "ReceivedPartial" event will be part of a single stream-wide Message that is marked Final (unless a Message Framer is used). EndOfMessage will be delivered when the TCP Connection has received a FIN (CLOSE-EVENT.TCP) from the peer. Note that reception of a FIN should not cause the "Closed" event to be delivered to the application, as it will still be possible for the application to send data.

Close: Calling "Close" on a TCP Connection indicates that the Connection should be gracefully closed (CLOSE.TCP) by sending a FIN to the peer. It will then still be possible to receive data until the peer closes or aborts the TCP connection. The "Closed" event will be issued upon reception of a FIN.

Abort: Calling "Abort" on a TCP Connection indicates that the Connection should be immediately closed by sending a RST to the peer (ABORT.TCP).

10.2. MPTCP

Connectedness: Connected

Data Unit: Byte-stream

API mappings for MPTCP are identical to TCP. MPTCP adds support for multipath properties, such as "Multipath Transport" and "Policy for using Multipath Transports".

10.3. UDP

Connectedness: Connectionless

Data Unit: Datagram

API mappings for UDP are as follows:

Connection Object: UDP connections represent a pair of specific IP addresses and ports on two hosts.

Initiate: CONNECT.UDP. Calling "Initiate" on a UDP Connection

causes it to reserve a local port, but does not generate any traffic.

InitiateWithSend: Early data on a UDP Connection does not have any special meaning. The data is sent whenever the Connection is Ready.

Ready: A UDP Connection is ready once the system has reserved a local port and has a path to send to the Remote Endpoint.

InitiateError: UDP Connections can only generate errors on initiation due to port conflicts on the local system.

ConnectionError: Once in use, UDP throws "soft errors" (ERROR.UDP(-Lite)) upon receiving ICMP notifications indicating failures in the network.

Listen: LISTEN.UDP. Calling "Listen" for UDP binds a local port and prepares it to receive inbound UDP datagrams from peers.

ConnectionReceived: UDP Listeners will deliver new connections once they have received traffic from a new Remote Endpoint.

Clone: Calling "Clone" on a UDP Connection creates a new Connection with equivalent parameters. The two Connections are otherwise independent.

Send: SEND.UDP(-Lite). Calling "Send" on a UDP connection sends the data as the payload of a complete UDP datagram. Marking Messages as Final does not change anything in the datagram's contents. Upon sending a UDP datagram, some relevant fields and flags in the IP header can be controlled: DSCP (SET_DSCP.UDP(-Lite)), DF in IPv4 (SET_DF.UDP(-Lite)) and ECN flag (SET_ECN.UDP(-Lite)).

Receive: RECEIVE.UDP(-Lite). UDP only delivers complete Messages to "Received", each of which represents a single datagram received in a UDP packet. Upon receiving a UDP datagram, the ECN flag from the IP header can be obtained (GET_ECN.UDP(-Lite)).

Close: Calling "Close" on a UDP Connection (ABORT.UDP(-Lite)) releases the local port reservation.

Abort: Calling "Abort" on a UDP Connection (ABORT.UDP(-Lite)) is identical to calling "Close".

10.4. UDP-Lite

Connectedness: Connectionless

Data Unit: Datagram

API mappings for UDP-Lite are identical to UDP. Properties that require checksum coverage are not supported by UDP-Lite, such as "Corruption Protection Length", "Full Checksum Coverage on Sending", "Required Minimum Corruption Protection Coverage for Receiving", and "Full Checksum Coverage on Receiving".

10.5. UDP Multicast Receive

Connectedness: Connectionless

Data Unit: Datagram

API mappings for Receiving Multicast UDP are as follows:

Connection Object: Established UDP Multicast Receive connections represent a pair of specific IP addresses and ports. The "unidirectional receive" transport property is required, and the Local Endpoint must be configured with a group IP address and a port.

Initiate: Calling "Initiate" on a UDP Multicast Receive Connection causes an immediate InitiateError. This is an unsupported operation.

InitiateWithSend: Calling "InitiateWithSend" on a UDP Multicast Receive Connection causes an immediate InitiateError. This is an unsupported operation.

Ready: A UDP Multicast Receive Connection is ready once the system has received traffic for the appropriate group and port.

InitiateError: UDP Multicast Receive Connections generate an InitiateError if Initiate is called.

ConnectionError: Once in use, UDP throws "soft errors" (ERROR.UDP(-Lite)) upon receiving ICMP notifications indicating failures in the network.

Listen: LISTEN.UDP. Calling "Listen" for UDP Multicast Receive binds a local port, prepares it to receive inbound UDP datagrams from peers, and issues a multicast host join. If a Remote Endpoint with an address is supplied, the join is Source-specific

Multicast, and the path selection is based on the route to the Remote Endpoint. If a Remote Endpoint is not supplied, the join is Any-source Multicast, and the path selection is based on the outbound route to the group supplied in the Local Endpoint.

ConnectionReceived: UDP Multicast Receive Listeners will deliver new connections once they have received traffic from a new Remote Endpoint.

Clone: Calling "Clone" on a UDP Multicast Receive Connection creates a new Connection with equivalent parameters. The two Connections are otherwise independent.

Send: SEND.UDP(-Lite). Calling "Send" on a UDP Multicast Receive connection causes an immediate SendError. This is an unsupported operation.

Receive: RECEIVE.UDP(-Lite). The Receive operation in a UDP Multicast Receive connection only delivers complete Messages to "Received", each of which represents a single datagram received in a UDP packet. Upon receiving a UDP datagram, the ECN flag from the IP header can be obtained (GET_ECN.UDP(-Lite)).

Close: Calling "Close" on a UDP Multicast Receive Connection (ABORT.UDP(-Lite)) releases the local port reservation and leaves the group.

Abort: Calling "Abort" on a UDP Multicast Receive Connection (ABORT.UDP(-Lite)) is identical to calling "Close".

10.6. SCTP

Connectedness: Connected

Data Unit: Message

API mappings for SCTP are as follows:

Connection Object: Connection objects can be mapped to an SCTP association or a stream in an SCTP association. Mapping Connection objects to SCTP streams is called "stream mapping" and has additional requirements as follows. The following explanation assumes a client-server communication model.

Stream mapping requires an association to already be in place between the client and the server, and it requires the server to understand that a new incoming stream should be represented as a new Connection Object by the Transport Services system. A new SCTP stream is

created by sending an SCTP message with a new stream id. Thus, to implement stream mapping, the Transport Services system MUST provide a newly created Connection Object to the application upon the reception of such a message. The necessary semantics to implement a Transport Services system's Close and Abort primitives are provided by the stream reconfiguration (reset) procedure described in [RFC6525]. This also allows to re-use a stream id after resetting ("closing") the stream. To implement this functionality, SCTP stream reconfiguration [RFC6525] MUST be supported by both the client and the server side.

To avoid head-of-line blocking, stream mapping SHOULD only be implemented when both sides support message interleaving [RFC8260]. This allows a sender to schedule transmissions between multiple streams without risking that transmission of a large message on one stream might block transmissions on other streams for a long time.

To avoid conflicts between stream ids, the following procedure is recommended: the first Connection, for which the SCTP association has been created, MUST always use stream id zero. All additional Connections are assigned to unused stream ids in growing order. To avoid a conflict when both endpoints map new Connections simultaneously, the peer which initiated association MUST use even stream ids whereas the remote side MUST map its Connections to odd stream ids. Both sides maintain a status map of the assigned stream ids. Generally, new streams SHOULD consume the lowest available (even or odd, depending on the side) stream id; this rule is relevant when lower ids become available because Connection objects associated with the streams are closed.

SCTP stream mapping as described here has been implemented in a research prototype; a desription of this implementation is given in [NEAT-flow-mapping].

Initiate: If this is the only Connection object that is assigned to the SCTP association or stream mapping is not used, CONNECT.SCTP is called. Else, unless the Selection Property "activeReadBeforeSend" is Preferred or Required, a new stream is used: if there are enough streams available, "Initiate" is a local operation that assigns a new stream id to the Connection object. The number of streams is negotiated as a parameter of the prior CONNECT.SCTP call, and it represents a trade-off between local resource usage and the number of Connection objects that can be mapped without requiring a reconfiguration signal. When running out of streams, ADD_STREAM.SCTP must be called.

InitiateWithSend: If this is the only Connection object that is

assigned to the SCTP association or stream mapping is not used, CONNECT.SCTP is called with the "user message" parameter. Else, a new stream is used (see "Initiate" for how to handle running out of streams), and this just sends the first message on a new stream.

Ready: "Initiate" or "InitiateWithSend" returns without an error, i.e. SCTP's four-way handshake has completed. If an association with the peer already exists, stream mapping is used and enough streams are available, a Connection Object instantly becomes Ready after calling "Initiate" or "InitiateWithSend".

InitiateError: Failure of CONNECT.SCTP.

ConnectionError: TIMEOUT.SCTP or ABORT-EVENT.SCTP.

Listen: LISTEN.SCTP. If an association with the peer already exists and stream mapping is used, "Listen" just expects to receive a new message with a new stream id (chosen in accordance with the stream id assignment procedure described above).

ConnectionReceived: LISTEN.SCTP returns without an error (a result of successful CONNECT.SCTP from the peer), or, in case of stream mapping, the first message has arrived on a new stream (in this case, "Receive" is also invoked).

Clone: Calling "Clone" on an SCTP association creates a new Connection object and assigns it a new stream id in accordance with the stream id assignment procedure described above. If there are not enough streams available, ADD_STREAM.SCTP must be called.

Priority (Connection): When this value is changed, or a Message with Message Property "Priority" is sent, and there are multiple Connection objects assigned to the same SCTP association, CONFIGURE_STREAM_SCHEDULER.SCTP is called to adjust the priorities of streams in the SCTP association.

Send: SEND.SCTP. Message Properties such as "Lifetime" and "Ordered" map to parameters of this primitive.

Receive: RECEIVE.SCTP. The "partial flag" of RECEIVE.SCTP invokes a "ReceivedPartial" event.

Close: If this is the only Connection object that is assigned to the SCTP association, CLOSE.SCTP is called, and the "Closed" event will be delivered to the application upon the ensuing CLOSE-EVENT.SCTP. Else, the Connection object is one out of several Connection objects that are assigned to the same SCTP assocation, and RESET_STREAM.SCTP

must be called, which informs the peer that the stream will no longer be used for mapping and can be used by future "Initiate", "InitiateWithSend" or "Listen" calls. At the peer, the event RESET_STREAM-EVENT.SCTP will fire, which the peer must answer by issuing RESET_STREAM.SCTP too. The resulting local RESET_STREAM-EVENT.SCTP informs the Transport Services system that the stream id can now be re-used by the next "Initiate", "InitiateWithSend" or "Listen" calls, and invokes a "Closed" event towards the application.

Abort: If this is the only Connection object that is assigned to the SCTP association, ABORT.SCTP is called. Else, the Connection object is one out of several Connection objects that are assigned to the same SCTP assocation, and shutdown proceeds as described under "Close".

11. IANA Considerations

RFC-EDITOR: Please remove this section before publication.

This document has no actions for IANA.

12. Security Considerations

[I-D.ietf-taps-arch] outlines general security consideration and requirements for any system that implements the Transport Services archtecture. [I-D.ietf-taps-interface] provides further discussion on security and privacy implications of the Transport Services API. This document provides additional guidance on implementation specifics for the Transport Services API and as such the security considerations in both of these documents apply. The next two subsections discuss further considerations that are specific to mechanisms specified in this document.

12.1. Considerations for Candidate Gathering

Implementations should avoid downgrade attacks that allow network interference to cause the implementation to select less secure, or entirely insecure, combinations of paths and protocols.

12.2. Considerations for Candidate Racing

See Section 5.3 for security considerations around racing with 0-RTT data.

An attacker that knows a particular device is racing several options during connection establishment may be able to block packets for the first connection attempt, thus inducing the device to fall back to a secondary attempt. This is a problem if the secondary attempts have

worse security properties that enable further attacks. Implementations should ensure that all options have equivalent security properties to avoid incentivizing attacks.

Since results from the network can determine how a connection attempt tree is built, such as when DNS returns a list of resolved endpoints, it is possible for the network to cause an implementation to consume significant on-device resources. Implementations should limit the maximum amount of state allowed for any given node, including the number of child nodes, especially when the state is based on results from the network.

13. Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644334 (NEAT).

This work has been supported by Leibniz Prize project funds of DFG - German Research Foundation: Gottfried Wilhelm Leibniz-Preis 2011 (FKZ FE 570/4-1).

This work has been supported by the UK Engineering and Physical Sciences Research Council under grant EP/R04144X/1.

This work has been supported by the Research Council of Norway under its "Toppforsk" programme through the "OCARINA" project.

Thanks to Stuart Cheshire, Josh Graessley, David Schinazi, and Eric Kinnear for their implementation and design efforts, including Happy Eyeballs, that heavily influenced this work.

14. References

14.1. Normative References

[I-D.ietf-taps-arch]
Pauly, T., Trammell, B., Brunstrom, A., Fairhurst, G., Perkins, C., Tiesel, P. S., and C. A. Wood, "An Architecture for Transport Services", Work in Progress, Internet-Draft, draft-ietf-taps-arch-10, 30 April 2021, <https://www.ietf.org/archive/id/draft-ietf-taps-arch-10.txt>.

[I-D.ietf-taps-interface]
Trammell, B., Welzl, M., Enghardt, T., Fairhurst, G., Kuehlewind, M., Perkins, C., Tiesel, P. S., Wood, C. A., Pauly, T., and K. Rose, "An Abstract Application Layer

Interface to Transport Services", Work in Progress, Internet-Draft, draft-ietf-taps-interface-12, 9 April 2021, <https://www.ietf.org/archive/id/draft-ietf-taps-interface-12.txt>.

[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>.

[RFC7540] Belshe, M., Peon, R., and M. Thomson, Ed., "Hypertext Transfer Protocol Version 2 (HTTP/2)", RFC 7540, DOI 10.17487/RFC7540, May 2015, <https://www.rfc-editor.org/info/rfc7540>.

[RFC8303] Welzl, M., Tuexen, M., and N. Khademi, "On the Usage of Transport Features Provided by IETF Transport Protocols", RFC 8303, DOI 10.17487/RFC8303, February 2018, <https://www.rfc-editor.org/info/rfc8303>.

[RFC8304] Fairhurst, G. and T. Jones, "Transport Features of the User Datagram Protocol (UDP) and Lightweight UDP (UDP-Lite)", RFC 8304, DOI 10.17487/RFC8304, February 2018, <https://www.rfc-editor.org/info/rfc8304>.

[RFC8305] Schinazi, D. and T. Pauly, "Happy Eyeballs Version 2: Better Connectivity Using Concurrency", RFC 8305, DOI 10.17487/RFC8305, December 2017, <https://www.rfc-editor.org/info/rfc8305>.

[RFC8421] Martinsen, P., Reddy, T., and P. Patil, "Guidelines for Multihomed and IPv4/IPv6 Dual-Stack Interactive Connectivity Establishment (ICE)", BCP 217, RFC 8421, DOI 10.17487/RFC8421, July 2018, <https://www.rfc-editor.org/info/rfc8421>.

[RFC8446] Rescorla, E., "The Transport Layer Security (TLS) Protocol Version 1.3", RFC 8446, DOI 10.17487/RFC8446, August 2018, <https://www.rfc-editor.org/info/rfc8446>.

[RFC8923] Welzl, M. and S. Gjessing, "A Minimal Set of Transport Services for End Systems", RFC 8923, DOI 10.17487/RFC8923, October 2020, <https://www.rfc-editor.org/info/rfc8923>.

14.2. Informative References

[I-D.ietf-quic-transport]
Iyengar, J. and M. Thomson, "QUIC: A UDP-Based Multiplexed and Secure Transport", Work in Progress, Internet-Draft,

draft-ietf-quic-transport-34, 14 January 2021, <https://www.ietf.org/archive/id/draft-ietf-quic-transport-34.txt>.

[I-D.ietf-tcpm-2140bis] Touch, J., Welzl, M., and S. Islam, "TCP Control Block Interdependence", Work in Progress, Internet-Draft, draft-ietf-tcpm-2140bis-11, 12 April 2021, <https://www.ietf.org/archive/id/draft-ietf-tcpm-2140bis-11.txt>.

[NEAT-flow-mapping] "Transparent Flow Mapping for NEAT", IFIP NETWORKING 2017 Workshop on Future of Internet Transport (FIT 2017) , 2017.

[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, <https://www.rfc-editor.org/info/rfc1928>.

[RFC3124] Balakrishnan, H. and S. Seshan, "The Congestion Manager", RFC 3124, DOI 10.17487/RFC3124, June 2001, <https://www.rfc-editor.org/info/rfc3124>.

[RFC3207] Hoffman, P., "SMTP Service Extension for Secure SMTP over Transport Layer Security", RFC 3207, DOI 10.17487/RFC3207, February 2002, <https://www.rfc-editor.org/info/rfc3207>.

[RFC5389] Rosenberg, J., Mahy, R., Matthews, P., and D. Wing, "Session Traversal Utilities for NAT (STUN)", RFC 5389, DOI 10.17487/RFC5389, October 2008, <https://www.rfc-editor.org/info/rfc5389>.

[RFC5766] Mahy, R., Matthews, P., and J. Rosenberg, "Traversal Using Relays around NAT (TURN): Relay Extensions to Session Traversal Utilities for NAT (STUN)", RFC 5766, DOI 10.17487/RFC5766, April 2010, <https://www.rfc-editor.org/info/rfc5766>.

[RFC6525] Stewart, R., Tuexen, M., and P. Lei, "Stream Control Transmission Protocol (SCTP) Stream Reconfiguration", RFC 6525, DOI 10.17487/RFC6525, February 2012, <https://www.rfc-editor.org/info/rfc6525>.

[RFC6762] Cheshire, S. and M. Krochmal, "Multicast DNS", RFC 6762, DOI 10.17487/RFC6762, February 2013, <https://www.rfc-editor.org/info/rfc6762>.

[RFC6763] Cheshire, S. and M. Krochmal, "DNS-Based Service Discovery", RFC 6763, DOI 10.17487/RFC6763, February 2013, <https://www.rfc-editor.org/info/rfc6763>.

[RFC7230] Fielding, R., Ed. and J. Reschke, Ed., "Hypertext Transfer Protocol (HTTP/1.1): Message Syntax and Routing", RFC 7230, DOI 10.17487/RFC7230, June 2014, <https://www.rfc-editor.org/info/rfc7230>.

[RFC7657] Black, D., Ed. and P. Jones, "Differentiated Services (Diffserv) and Real-Time Communication", RFC 7657, DOI 10.17487/RFC7657, November 2015, <https://www.rfc-editor.org/info/rfc7657>.

[RFC8260] Stewart, R., Tuexen, M., Loreto, S., and R. Seggelmann, "Stream Schedulers and User Message Interleaving for the Stream Control Transmission Protocol", RFC 8260, DOI 10.17487/RFC8260, November 2017, <https://www.rfc-editor.org/info/rfc8260>.

[RFC8445] Keranen, A., Holmberg, C., and J. Rosenberg, "Interactive Connectivity Establishment (ICE): A Protocol for Network Address Translator (NAT) Traversal", RFC 8445, DOI 10.17487/RFC8445, July 2018, <https://www.rfc-editor.org/info/rfc8445>.

[TCP-COUPLING] "ctrlTCP: Reducing Latency through Coupled, Heterogeneous Multi-Flow TCP Congestion Control", IEEE INFOCOM Global Internet Symposium (GI) workshop (GI 2018) , n.d..

Appendix A. API Mapping Template

Any protocol mapping for the Transport Services API should follow a common template.

Connectedness: (Connectionless/Connected/Multiplexing Connected)

Data Unit: (Byte-stream/Datagram/Message)

Connection Object:

Initiate:

InitiateWithSend:

Ready:

InitiateError:

ConnectionError:

Listen:

ConnectionReceived:

Clone:

Send:

Receive:

Close:

Abort:

Appendix B. Additional Properties

This appendix discusses implementation considerations for additional parameters and properties that could be used to enhance transport protocol and/or path selection, or the transmission of messages given a Protocol Stack that implements them. These are not part of the interface, and may be removed from the final document, but are presented here to support discussion within the TAPS working group as to whether they should be added to a future revision of the base specification.

B.1. Properties Affecting Sorting of Branches

In addition to the Protocol and Path Selection Properties discussed in Section 4.1.3, the following properties under discussion can influence branch sorting:

* Bounds on Send or Receive Rate: If the application indicates a bound on the expected Send or Receive bitrate, an implementation may prefer a path that can likely provide the desired bandwidth, based on cached maximum throughput, see Section 9.2. The application may know the Send or Receive Bitrate from metadata in adaptive HTTP streaming, such as MPEG-DASH.

* Cost Preferences: If the application indicates a preference to avoid expensive paths, and some paths are associated with a monetary cost, an implementation should decrease the ranking of such paths. If the application indicates that it prohibits using expensive paths, paths that are associated with a cost should be purged from the decision tree.

Appendix C. Reasons for errors

The Transport Services API [I-D.ietf-taps-interface] allows for the several generic error types to specify a more detailed reason as to why an error occurred. This appendix lists some of the possible reasons.

* InvalidConfiguration: The transport properties and endpoints provided by the application are either contradictory or incomplete. Examples include the lack of a Remote Endpoint on an active open or using a multicast group address while not requesting a unidirectional receive.

* NoCandidates: The configuration is valid, but none of the available transport protocols can satisfy the transport properties provided by the application.

* ResolutionFailed: The remote or local specifier provided by the application can not be resolved.

* EstablishmentFailed: The Transport Services system was unable to establish a transport-layer connection to the remote endpoint specified by the application.

* PolicyProhibited: The system policy prevents the transport system from performing the action requested by the application.

* NotCloneable: The protocol stack is not capable of being cloned.

* MessageTooLarge: The message size is too big for the transport system to handle.

* ProtocolFailed: The underlying protocol stack failed.

* InvalidMessageProperties: The message properties are either contradictory to the transport properties or they can not be satisfied by the transport system.

* DeframingFailed: The data that was received by the underlying protocol stack could not be deframed.

* ConnectionAborted: The connection was aborted by the peer.

* Timeout: Delivery of a message was not possible after a timeout.

Appendix D. Existing Implementations

This appendix gives an overview of existing implementations, at the time of writing, of transport systems that are (to some degree) in line with this document.

* Apple's Network.framework:
  - Network.framework is a transport-level API built for C, Objective-C, and Swift. It a connect-by-name API that supports transport security protocols. It provides userspace implementations of TCP, UDP, TLS, DTLS, proxy protocols, and allows extension via custom framers.
  - Documentation: https://developer.apple.com/documentation/network (https://developer.apple.com/documentation/network)
* NEAT and NEATPy:
  - NEAT is the output of the European H2020 research project "NEAT"; it is a user-space library for protocol-independent communication on top of TCP, UDP and SCTP, with many more features such as a policy manager.
  - Code: https://github.com/NEAT-project/neat (https://github.com/NEAT-project/neat)
  - NEAT project: https://www.neat-project.org (https://www.neat-project.org)
  - NEATPy is a Python shim over NEAT which updates the NEAT API to be in line with version 6 of the TAPS interface draft.
  - Code: https://github.com/theagilepadawan/NEATPy (https://github.com/theagilepadawan/NEATPy)
* PyTAPS:
  - A TAPS implementation based on Python asyncio, offering protocol-independent communication to applications on top of TCP, UDP and TLS, with support for multicast.
  - Code: https://github.com/fg-inet/python-asyncio-taps (https://github.com/fg-inet/python-asyncio-taps)

Authors' Addresses

Anna Brunstrom (editor)
Karlstad University
Universitetsgatan 2
651 88 Karlstad
Sweden

Email: anna.brunstrom@kau.se

Tommy Pauly (editor)
Apple Inc.
One Apple Park Way
Cupertino, California 95014,
United States of America

Email: tpauly@apple.com

Theresa Enghardt
Netflix
121 Albright Way
Los Gatos, CA 95032,
United States of America

Email: ietf@tenghardt.net

Karl-Johan Grinnemo
Karlstad University
Universitetsgatan 2
651 88 Karlstad
Sweden

Email: karl-johan.grinnemo@kau.se

Tom Jones
University of Aberdeen
Fraser Noble Building
Aberdeen, AB24 3UE
United Kingdom

Email: tom@erg.abdn.ac.uk

Philipp S. Tiesel
SAP SE
Konrad-Zuse-Ring 10

14469 Potsdam
Germany

Email: philipp@tiesel.net

Colin Perkins
University of Glasgow
School of Computing Science
Glasgow G12 8QQ
United Kingdom

Email: csp@csperkins.org

Michael Welzl
University of Oslo
PO Box 1080 Blindern
0316 Oslo
Norway

Email: michawe@ifi.uio.no