ST Working Group
Internet-Draft
File: draft-ietf-st2-spec-00.txt

L. Delgrossi

October 1994
Expires: January 1995

Internet Stream Protocol Version 2 (ST2)

Protocol Specification - Version ST2Plus

Status of this Memo

Abstract:

This memo contains a revised specification of the Internet STream Protocol Version 2 (ST2). ST2 is a resource reservation protocol intended to provide end-to-end real-time guarantees over an internet. It allows its applications to build multi-destination simplex data streams with a desired quality of service. The revised version of ST2 specified in this memo is called ST2Plus.

Editor's Note:

This memo is available both in ASCII format (file: draft-ietf-st2-spec-00.txt) and in PostScript (file: draft-ietf-st2-spec-00.ps). The PostScript version contains some additional pictures that help to clarify the text, and it is therefore recommended.

Introduction

1 What is ST2?

The Internet Stream Protocol, Version 2 (ST2) is a connection-oriented internetworking protocol that operates at the same layer as connectionless IP. It has been developed to support the efficient delivery of data streams to single or multiple destinations in applications that require guaranteed data throughput and controlled delay characteristics. The main application area of the protocol is the real-time transport of digital audio and video packet streams across internets.

ST2 can be used to reserve bandwidth for multimedia streams across network routes. This reservation, together with appropriate network access and packet scheduling mechanisms in all nodes running the protocol, guarantees a well-defined quality of service to ST2 applications. It ensures that each multimedia packet is delivered within its deadline, that is, at the time where it needs to be presented. This facilitates a smooth playout of digital audio and video that is essential for this time-critical data, but can typically not be provided by best-effort IP communication.

Just like IP, ST2 actually consists of two protocols: ST for the data transport and SCMP, the Stream Control Message Protocol, for all control functions, mainly those for resource reservation. ST is simple and contains only one PDU that is designed for fast and efficient data forwarding in order to achieve low communication delays. SCMP, however, is quite complex. As with ICMP and IP, SCMP packets are transferred within ST packets as shown in Figure 1.

1.1 Protocol History

The first version of ST was published in the late 1970's and was used throughout the 1980's for experimental voice and video transmission. The experience gained in these applications led to the development of the revised protocol version ST2. The revision extends the original protocol to make it more complete and more applicable to emerging multimedia environments. The specification of this protocol version is contained in Internet RFC 1190 which was published in October 1990 [RFC1190].

With more and more developments of commercial distributed multimedia applications underway and with a growing dissatisfaction at the transmission quality for audio and video over IP in the MBONE, interest in ST2 has grown over the last years. Companies such as BBN have products available incorporating the protocol. The BERKOM project of the German PTT uses ST2 as its core protocol for the provision of multimedia teleservices such as conferencing and mailing. Among others, Digital, HP, IBM, and Siemens-Nixdorf

participate in this project. In addition, implementations of ST2 for Sun, Silicon Graphics, Macintosh, NeXT, and PC platforms are available.

In 1993, the IETF has started a new working group on ST2. Its mission is to clean up the current protocol specification to ensure better interoperability between the existing and emerging implementations. It shall also reflect the experiences gained with the current ST2 implementations and applications. This has led to the specification of the ST2Plus version contained in this document.

1.2 Streams

Streams form the core concepts of ST2. They are established between a sending origin and one or more receiving targets in the form of a routing tree. Nodes in the tree represent so-called ST agents, entities executing the ST2 protocol; links in the tree are called hops.

Figure 2 illustrates a stream from an origin to four targets, where the ST agent on Target 2 also functions as a router. Let us use this Target 2/Router node to explain some basic ST2 terminology: the direction of the stream from this node to Target 3 and 4 is called downstream, the direction towards the Origin node upstream. ST agents that are one hop away from a given node are called previous-hops in the upstream, and next-hops in the downstream direction.

Streams are maintained using SCMP messages. Typical SCMP messages are CONNECT and ACCEPT to build a stream, DISCONNECT and REFUSE to close a stream, or CHANGE to modify the stream characteristics, for example the set of targets or the quality of service.

Each ST agent maintains state information describing the streams flowing through it. It can actively gather and distribute such information. If, for example, an intermediate ST agent fails, the neighboring agents can recognize this via HELLO messages that are periodically exchanged between ST agents that share streams. STATUS packets can be used to ask other ST agents about a particular stream. These agents then send back a STATUS-RESPONSE message. NOTIFY messages serve to inform ST agents of changes such as a route change.

ST2 offers a wealth of functionalities for stream management. Streams can be grouped together to minimize allocated resources or to process them in the same way in case of failures. During audio conferences, for example, only one person should speak at a time. Using the group mechanism, resources for only one audio stream of the group need to be reserved. Using the same concept, an entire group of related audio and video streams can be dropped if one of them fails.

1.3 Data Transmission

Data transfer in ST2 is simplex in the downstream direction. Data transport through streams is very efficient. ST2 puts only a small header in front of the user data. The header contains a protocol identification that distinguishes ST2 from IP packets, an ST2 version number, a priority field (specifying a relative importance of streams in cases of conflict), a length counter, a stream identification, and a checksum. These elements form an 8-byte header which can be extended by an optional 8-byte timestamp.

Efficiency is also achieved by avoiding fragmentation and reassembly on router nodes. Negotiations at stream establishment time yield a maximum transmission unit (MTU) for data packets on a stream. This MTU is communicated to the upper layers, so that they provide data packets of suitable size to ST2.

Communication with multiple next-hops can be made even more efficient using MAC Layer multicast. If a subnet supports multicast, a single multicast packet is sufficient to reach all next- hops connected to this subnet. This leads to a significant reduction of the bandwidth requirements of a stream. If multicast is not provided, separate packets need to be sent to each next-hop.

As ST2 relies on reservation, it does not contain error correction mechanisms features for data exchange such as retransmission known from TCP. It is assumed that digital audio and video require partially correct delivery only. In many cases, retransmitted packets would arrive too late to meet their real-time delivery requirements. On the other hand, depending on the data encoding and the particular application, a small number of errors in audio and video streams are acceptable. In any case, reliability can be provided by layers on top of ST2 if needed.

1.4 Flow Specifications

As part of establishing a connection, SCMP negotiates quality-of-service parameters for a stream. In ST2 terminology, these parameters form a flow specification (FlowSpec, for short) which is associated with the stream. Different versions of FlowSpecs exist and can be distinguished by a version number. Typically, they contain parameters such as average and maximum throughput, end-to-end delay, and delay variance of a stream.

Three kinds of entities participate in the quality-of-service negotiation: application entities on the origin and target sites as the service users, ST agents, and local resource managers (LRM). The origin application supplies the initial FlowSpec requesting a

particular service quality. Each ST agent which obtains the specification as part of a connection establishment message initiates the reservation of local resources by the corresponding resource manager. These resource managers control the usage of CPU capacity for protocol processing, buffer space for storing messages, and bandwidth in the outgoing network. ST2 does not determine how resource managers make reservations and how resources are scheduled according to these reservations; ST2, however, assumes these mechanisms as its basis.

The FlowSpec negotiation procedure is illustrated in Figure 3. Depending on the success of its local reservations, an ST agent updates the FlowSpec while the connection establishment message passes downstream (for example, keeping track of accumulated delay). The final FlowSpec is communicated to the target application which may base its accept/reject decision for establishing the connection on it and may finally also modify the FlowSpec. If a target accepts the connection, the (possibly modified) FlowSpec is propagated back to the origin which can then calculate an overall service quality for all targets. If all targets in a particular ST2 connection need to adhere to the same FlowSpec, the origin may - during a second phase of connection establishment - issue a CHANGE request to adjust reservations.

1.5 ST2 and IP

ST2 is designed to coexist with IP on each node. A typical distributed multimedia application would use both protocols: IP for the transfer of traditional data and control information, and ST2 for the transfer of digital audio and video. Whereas IP typically will be accessed from TCP or UDP, ST2 will have new multimedia end-to-end protocols on top of it.

Both ST2 and IP apply the same addressing schemes to identify different hosts and use ARP for address resolution. ST2 can easily be modified to include the longer host addresses of the next generation IP. ST2 uses the same Layer 2 SAPs as IP. ST2 and IP packets differ in the first four bits, containing the internetwork protocol version number: number 5 is reserved for ST2 (IP itself has version number 4). An ST agent receives a packet over the IP SAP using the first 4 bits of the frame to select ST2 packets.

As a special function, ST2 messages can be encapsulated in IP packets. This allows them to pass through routers which do not run ST2. Resource management is typically not available for these IP route segments. IP encapsulation is, therefore, suggested only for portions of the network which do not constitute a system bottleneck.

1.6 Outline of This Document

This document contains the specification for the ST2Plus version of the ST2 protocol. In the rest of the document, whenever the terms "ST" or "ST2" are used, they refer to ST2Plus.

The document is organized as follows: Section 2 describes the ST user service; Section 3 through Section 7 describe stream setup, modification, and tear down; exceptional cases are handled in Section 8; failure detection and groups of streams respectively in Section 9 and Section 10; the FlowSpec is presented in Section 12; finally, the formats of the different protocol elements and PDUs are defined by Section 14 through Section 20. Figure 1: ST2 Data and Control Path Figure 2: The Stream Concept Figure 3: Quality-of-Service Negotiation with FlowSpecs

Figure 6: ST Service at the Target Figure 4: Primitives for the OPEN Stream Operation Figure 5: ST Service at the Origin

2 ST User Service Description

This section describes the ST user service from the high-level point of view of an application. It defines the ST stream operations and primitive functions. It specifies which operations on streams can be invoked by the applications built on top of ST and when the ST primitive functions can be legally executed. Note that the ST primitives do not form an API. They are used here with the only purpose of illustrating the service model for ST.

2.1 Stream Operations and Primitive Functions

An ST application at the origin may create, expand, reduce, change, send data to, and delete a stream. When a stream is expanded, new targets are added to the stream; when a stream is reduced, some of the current targets are dropped from it. When a stream is changed, the associated quality of service is modified.

An ST application at the target may join, receive data from, and leave a stream.

This translates into the following stream operations:

o OPEN: create new stream [origin], CLOSE: delete stream [origin],

o ADD: expand stream, i.e. add new targets to it [origin],

o DROP: reduce stream, i.e. drop targets from it [origin],

- JOIN: join a stream [target], LEAVE: leave a stream [target],
- DATA: send data through stream [origin],
- CHG: change a stream's qos [origin],

Each stream operation may require the execution of several primitive functions to be completed. For instance, to open a new stream, a request is first issued by the sender and an indication is generated at the receiver; then, the receiver may accept or refuse the request and the correspondent indication is generated at the sender. This is shown in Figure 4 below.

Table 1 defines the ST service primitive functions associated to each stream operation. The column labelled "O/T" indicates whether the primitive is executed at the origin or at the target.

2.2 State Diagrams

It is not sufficient to define the set of ST stream operations. It is also necessary to specify when the operations can be legally executed. For this reason, a set of states are now introduced and the transitions from one state to the others are specified. States are defined with respect to a single stream. The previously defined stream operations can be legally executed only from an appropriate state.

An ST agent may, with respect to an ST stream, be in one of the following states:

- IDLE: the stream has not been created yet.
- PENDING: the stream is in the process of being established.
- ACTIVE: the stream is established and active.
- ADDING: the stream is established. A stream expansion is underway.
- CHGING: the stream is established. A stream change is underway.

Previous experience with ST suggested to impose limits on the stream operations that can be executed at the same time. These restrictions are:

1. A single ADD or CHG operation can be processed at one time. If another ADD or CHG is already underway, further requests are

queued by the ST agent and handled only after the previous operation has been completed. It also applies to two subsequent requests of the same kind, e.g. two ADD or two CHG operations. The second operation is not executed until the first one has been completed.

2. Deleting a stream, leaving a stream, or dropping targets from a stream is possible only after stream establishment has been completed. A stream is considerate to be established when all the next-hops of the origin have either accepted or refused the stream. Note that stream refuse is automatically forced after timeout if no reply comes from a next-hop.

3. An ST agent forwards data only along already established paths to the targets. A path is con- sidered to be established when the next-hop on the path has explicitly accepted the stream. This implies that the target and all other intermediate ST agents are ready to handle the incoming data packets. In no cases an ST agent will forward data to a next-hop agent that has not explicitly accepted the stream. To be sure that all targets receive the data, an application should send the data only after all paths have been established, i.e. the stream is established.

4. It is allowed to send data from the CHGING and ADDING states. When sending data from the CHGING state the quality of service to the targets affected by the change is undefined. When sending data from the ADDING state the targets that receive the data include at least all the targets that were already part of the stream at the time the ADD operation was invoked.

The rules introduced above require ST agents to queue incoming requests when the current state does not allow to process them immediately. In order to preserve the semantics, ST agents have to maintain the order of the requests, i.e. implement FIFO queuing. Exceptionally, the CLOSE request at the origin and the LEAVE request at the target may be immediately processed: in this cases, the queue is deleted and it is possible that requests in the queue are not processed.

The following state diagrams define the ST service. Separate diagrams are presented for the origin and the targets. To keep the figure simple, only the primitives that cause state transitions are represented.

The symbol (a/r)* indicates that all targets in the target list have explicitly accepted or refused the stream, or refuse has been forced after timeout. If the target list is empty, i.e. it contains no targets, the (a/r)* condition is immediately satisfied, so the empty stream is created and state ESTBL is entered.

2.3 State Transition Tables

Table 2 and Table 3 define which primitives can be processed from which states and the possible state transitions.

Figure 7: Sample Topology for an ST Stream

SCMP Functional Description

ST agents create and manage streams using the ST Control Message Protocol (SCMP). Conceptually, SCMP resides immediately above ST (as does ICMP above IP). SCMP follows a request-response model. SCMP messages are made reliable through the use of retransmission after timeout, cf. Section 11.2.

This section contains a functional description of SCMP. To help clarify the SCMP exchanges used to setup and maintain ST streams, we include an example of a simple network topology, represented in Figure 7. The topology is used to illustrate the protocol interactions during the execution of stream operations. For instance, an ST application may:

- Create a stream from A to the peers at B, C and D,
- Add a peer at E,
- Drop peers B and C, and
- Let F join the stream
- Delete the stream.

We begin with a description of stream setup, see Section 3; stream option are presented in Section 4; data transfer in Section 5; Section 6 illustrates stream modification including stream expansion, reduction, changes of the quality of service associated to a stream. Finally, stream deletion is handled in Section 7.

3 Stream Setup

This section presents a description of stream setup. For simplicity, we assume that everything succeeds, e.g. any required resources are available, and the routing is correct. Possible failures in the setup phase are handled in Section 8.1.

3.1 Initial Setup at the Origin

Before stream setup can be started, the application has to collect

the necessary information to determine the structure of the communication. This includes identifying the participants and selecting the characteristics of the data flow. Such information is passed to the ST agent at the stream's origin. The ST agent performs the following operations:

- allocates a stream ID (SID) for the stream, cf. Section 11.1,
- invokes the routing function to determine the set of next-hops for the stream, cf. Section 3.1.1,
- invokes the Local Resource Manager (LRM), cf. Section 3.1.2, to reserve local and network resources
- creates local database entries to store information on the new stream,
- propagates the stream creation request to the next-hops determined by the routing function, see Section 3.2.

3.1.1 Invoking the Routing Function

An ST agent that is setting up a stream invokes the routing function to find a path to reach each of the targets specified by the target list provided by the application. This is similar to the routing decision in IP. However, in this case the route is to a multitude of targets rather than to a single destination. The routing function is not part of the ST protocol and therefore it is not specified by this document.

The result of the routing function is a set of next-hop ST agents. The set of next-hops selected by the routing function is not necessarily the same as the set of next-hops that IP would select given a number of independent IP datagrams to the same destinations. The routing algorithm may attempt to optimize parameters other than the number of hops that the packets will take, such as delay, local network bandwidth consumption, or total internet bandwidth consumption.

3.1.2 Reserving Resources

An ST agent helps reserving both local and network resources. Local resources may include CPU processing time and buffer space at the local host. Network resources may comprise bandwidth over the outgoing links to the next-hops determined by the routing function. Resource reservation is not part of the ST protocol and therefore it is not specified by this document. ST invokes at every host the Local Resource Manager (LRM) to perform the appropriate reservations.

Functions as resource scheduling and reservation enforcement are part of the LRM's tasks and not of an ST agent's.

The ST FlowSpec contains all the information needed to allocate the necessary resources. The information contained in the FlowSpec is passed to the LRM as parameter of the reservation functions. The LRM updates the FlowSpec information before it passes it back to the ST agent. Further information on the ST FlowSpec can be found in Section 12.

Note that if the data has to be sent across a network to a single next-hop, then only the point-to- point bandwidth needs to be reserved. If the data has to be sent to multiple next-hop agents across a single network and network layer multicasting is not available, the ST agent replicates the data to each next-hop agent and therefore bandwidth has to be reserved by the LRM for all the next-hops. If network layer multicast is supported, its use reduces the bandwidth required since one single copy of the data is received by all next-hop agents. The membership of a stream in a Group may also affect the amount of resources that have to be allocated by the LRM, cf. Section 10.

Effects similar to reservation of the necessary resources may be obtained even when the network cannot provide direct support for the reservation. Certainly if total reservations are a small fraction of the overall resources, such as packet switch processing bandwidth, buffer space, or network bandwidth, then the desired performance can be honoured if the degree of confidence is consistent with the requirements as stated in the FlowSpec. Other solutions can be designed for specific networks.

3.2 Sending CONNECT Messages

The ST agent sends a CONNECT message to each of the next-hop ST agents identified by the routing function. Each CONNECT message contains the SID, an updated FlowSpec, and a TargetList. In general, the FlowSpec and TargetList depend on both the next-hop and the intervening network. Each TargetList is a subset of the original TargetList, identifying the targets that are to be reached through the next-hop to which the CONNECT message is being sent.

The TargetList may be empty, see Section 3.2.1.; if the TargetList causes a too long CONNECT message to be generated, the CONNECT message is partitioned as explained in Section 3.2.2. If multiple next-hops are to be reached through a network that supports network level multicast, a different CONNECT message must nevertheless be sent to each next-hop since each will have a different TargetList.

Let us consider the network topology in Figure 7 on page 16. Suppose that the original TargetList contains targets B, C, and D. The routing function invoked at A returns that B is reachable via Router 1 and C and D are reachable via Router 2. Thus, A generates two CONNECT messages, one for Router 1 and one for Router 2. The CONNECT message for Router 1 contains a TargetList including target B only; the CONNECT message for Router 2 contains a TargetList including targets C and D.

3.2.1 Empty Target List

An application at the origin may request the local ST agent to create empty streams. It does so by passing an empty TargetList to the local ST agent during the initial stream setup. When the local ST agent receives request to create an empty stream, it allocates the stream ID (SID), updates its local database entries to store information on the new stream and notifies the application that stream setup is complete. The local ST agent does not generate any CONNECT message for streams with an empty TargetList.

3.2.2 Long Target Lists

Each ST agent knows the MTU of the networks to which it is connected, and those MTUs restrict the size of the SCMP message it can send. SCMP messages with long TargetList can cause the size of the SCMP message to exceed the network MTU. The ST agent which receives an SCMP message bigger than its MTU must break the original message into multiple fragments, each carrying part of the TargetList. The effect of this partition is to compromise the performance but still carry out the function of the SCMP message. If the original SCMP message contains any Userdata parameters, these parameters are replicated in each fragment for delivery to all targets. Applications that support a large number of receivers may avoid using long target lists by exploiting the stream joining functions, cf. Section 6.2.

3.3 Processing CONNECT Messages

3.3.1 CONNECT Processing by an Intermediate Agent

An ST agent receiving a CONNECT message, assuming no errors, responds to the previous-hop with an ACK. The ACK message must identify the CONNECT to which it corresponds by including the reference number indicated by the Reference field of the CONNECT message. The intermediate ST agent invokes the routing function, reserves resources via the LRM, and then propagates the CONNECT messages to its next-hops, as described in the previous section.

3.3.2 Setup at the Targets

An ST agent that is the target of a CONNECT message, assuming no errors, responds to the previous-hop with an ACK. The ST agent reserves local resources and inquires from the specified application process whether or not it is willing to accept the connection.

In particular, the application must be presented with parameters from the CONNECT, such as the SID, FlowSpec, Options, and Group, to be used as a basis for its decision. The application is identified by a combination of the NextPcol field and the SAP field included in the correspondent (usually single remaining) Target of the TargetList. The contents of the SAP field may specify the port or other local identifier for use by the protocol layer above the host ST layer. Subsequently received data packets will carry the SID, that can be mapped into this information and be used for their delivery.

Finally, based on the application's decision, the ST agent sends to the previous-hop from which the CONNECT was received an ACCEPT or REFUSE message. Since the ACCEPT (or REFUSE) message has to be acknowledged by the previous-hop, it is assigned a new Reference number that will be returned in the ACK. The CONNECT to which the ACCEPT (or REFUSE) is a reply is identified by placing the CONNECT's Reference number in the LnkReference field of the ACCEPT (or REFUSE). The ACCEPT message contains the FlowSpec as accepted by the application at the target.

3.4 Processing ACCEPT Messages

3.4.1 ACCEPT Processing by an Intermediate Agent

When an intermediate ST agent receives an ACCEPT, it first verifies that the message is a response to an earlier CONNECT. If not, it responds to the next-hop ST agent with an ERROR message, with ReasonCode (LnkRefUnknown). Otherwise, it responds to the next-hop ST agent with an ACK, and propagates the ACCEPT message to the previous-hop along the same path traced by the CONNECT but in the reverse direction toward the origin.

The FlowSpec is included in the ACCEPT message so that the origin and intermediate ST agents can gain access to the information that was accumulated as the CONNECT traversed the internet. Note that the resources, as specified in the FlowSpec in the ACCEPT message, may differ from the resources that were reserved by the agent when the CONNECT was originally processed. However, the agent does not adjust the reservation in response to the ACCEPT. It is expected that any excess resource allocation will be released for use by other stream or datagram traffic through an explicit CHANGE message initiated by the application at the origin if it does not wish to be charged for any excess resource allocations.

3.4.2 ACCEPT Processing by the Origin

The origin will eventually receive an ACCEPT (or REFUSE) message from each of the targets. As each ACCEPT is received, the application is notified of the target and the resources that were successfully allocated along the path to it, as specified in the FlowSpec contained in the ACCEPT message. The application may then use the information to either adopt or terminate the portion of the stream to each target. When ACCEPT (or REFUSE) from all targets have been received at the origin, the application is notified that stream setup is complete. For problems due to CONNECT timeout, please refer to Section 8.1.1.

When an ACCEPT is received by the origin, the path to the target is considered to be established and the ST agent is allowed to forward the data along this path as explained in Section 5 and in the ST user service description in Section 2.

3.5 Processing REFUSE Messages

3.5.1 REFUSE Processing by the Intermediate Agent

If an application at a target does not wish to participate in the stream, it sends a REFUSE message back to the origin with ReasonCode (ApplDisconnect). An intermediate ST agent that receives a REFUSE message with ReasonCode (ApplDisconnect) acknowledges it by sending an ACK to the next-hop, considers which resources are to be released, deletes the target entry from the internal database, and propagates the REFUSE message back to the previous-hop ST agent.

If, after deleting the specified target, the next-hop has no remaining targets, then those resources associated with that next-hop agent may be released. Note that network resources may not actually be released if network multicasting is being used since they may still be required for traffic to other next-hops in the multicast group.

3.5.2 REFUSE Processing by the Origin

When the REFUSE reaches the origin, the origin sends an ACK and notifies the application that the target is no longer part of the stream and also if the stream has no remaining targets. If there are no remaining targets, the application may wish to terminate the stream or keep the stream active to allow stream joining as described in Section 6.2.

4 Stream Options

An application may select among some stream options. The desired options are indicated to the ST agent at the origin when a new stream is created. Options apply to single streams and are valid during the whole stream's lifetime. The options chosen by the application at the origin are included into the initial CONNECT message(s). When a CONNECT message reaches a target, the application at the target is notified of the stream options that have been selected.

4.1 No Recovery

The NoRecovery option is used to indicate that ST agents should not attempt recovery in case of network or component failure. If a failure occurs, the origin will be notified via a REFUSE message and the targets via a DISCONNECT, with an appropriate ReasonCode indicating the reason of the failure. The application at the origin may decide whether to rebuild the deleted portion of the stream by sending a CONNECT message. The NoRecovery option is specified by setting the S-bit in the CONNECT message, see Section 16.4.

4.2 Join Authorization

To Be Written

5 Data Transfer

An application is not guaranteed that the data reaches its destinations: ST is unreliable and it does not make any attempt to recover from packet loss, e.g. due to the underlying network. In case the data reaches its destination, it does it accordingly to the negotiated quality of service.

An ST agent forwards the data only along already established paths to targets. A path is considered to be established when the ST next-hop agent on the path sends an ACCEPT message. This implies that the target and all other intermediate ST agents on the path to the target are ready to handle the incoming data packets. In no cases an ST agent will forward data to a next-hop agent that has not explicitly accepted the stream.

To be fairly sure that all targets receive the data with the desired quality of service, an application should send the data only after the whole stream has been established. Depending on the local API, an application may not be prevented to send data before the completion of stream setup, but it should be aware that the data could be lost or not reach all the intended targets.

At the end of the connection setup phase, the origin, each target, and each intermediate ST agent has a database entry that allows it to

forward the data packets from the origin to the targets and to recover from failures of the intermediate agents or networks. The database should be optimized to make the packet forwarding task most efficient. The time critical operation is an intermediate agent receiving a packet from the previous-hop agent and forwarding it to the next- hop agents. The database entry must also contain the FlowSpec, utilization information, the address of the origin and previous-hop, and the addresses of the targets and next-hops, so it can perform enforcement and recover from failures.

An ST agent receives data packets encapsulated by an ST header. A data packet received by an ST agent contains the SID. This SID was selected at the origin so that it is globally unique and thus can be used as an index into the database, to obtain quickly the necessary replication and forwarding information.

The forwarding information will be network and implementation specific, but must identify the next-hop agents. It is suggested that the cached information for a next-hop agent include the local network address of the next- hop. If the data packet must be forwarded to multiple next- hops across a single network that supports multicast, the database may specify the next-hops by a (local network) multicast address. If the network does not support multicast, or the next-hops are on different networks, multiple copies of the data packet must be sent.

No data fragmentation is supported during the data transfer phase. The application is expected to segment its PDUs according to the minimum MTU over all paths in the stream. The application receives information on the MTUs relative to the paths to the targets as part of the FlowSpec contained in the ACCEPT message, see also Section 12. The minimum MTU over all paths has to be calculated from the MTUs relative to the single paths. If the application at the origin sends a too large data packet, the ST agent at the origin generates an error and it does not forward the data.

6 Modifying an Existing Stream

Some applications may wish to modify a stream after it has been created. Possible changes include expanding a stream, reducing it, and changing its FlowSpec. In ST, changes to a stream may be initiated both by the origin and the targets. Targets may be added by the origin as described in Section 6.1 or they may request to join the stream as described in Section 6.2. The origin can reduce a stream by dropping some or all of its targets. This is described in Section 6.3. Targets may spontaneously decide to leave a stream as described in Section 6.4. Section 6.5 explains how to change a stream's FlowSpec.

As defined by the ST service model, see Section 2, an ST agent can handle only one stream modification at a time. If a stream modification operation is already underway, further requests are queued and handled when the previous operation has been completed. This also applies to two subsequent requests of the same kind, e.g. two subsequent changes to the FlowSpec.

6.1 The Origin Adding New Targets

It is possible for an application at the origin to add new targets to an existing stream any time after the stream has been established. Before new targets are added, the application has to collect the necessary information on the new targets. Such information is passed to the ST agent at the origin.

The ST agent at the origin issues a CONNECT message that contains the SID, the FlowSpec, and the TargetList specifying the new targets. This is similar to sending a CONNECT message during stream establishment, with the following exceptions: the origin checks that a) the SID is valid, b) the targets are not already members of the stream, c) the FlowSpec of the new target, if present, matches the FlowSpec of the existing stream, i.e it requires an equal or smaller amount of resources to be allocated. If the FlowSpec of the new target does not match the FlowSpec of the existing stream, it is simply ignored.

An intermediate ST agent that is already a node in the stream looks at the SID and verifies that the stream is the same. It then checks if the intersection of the TargetList and the targets of the established stream is empty. If this is not the case, it responds with an ERROR message with the appropriate ReasonCode (RouteLoop) that contains a TargetList of those targets that were duplicates.

For each new target in the TargetList, processing is much the same as for the original CONNECT. The CONNECT is acknowledged, propagated, and network resources are reserved. However, it may be possible to route to the new targets using previously allocated paths or an existing multicast group. In that case, additional resources do not need to be reserved but more next-hops might have to be added to an existing multicast group.

Intermediate or target ST agents that are not already nodes in the stream behave as in case of stream setup (see Section 3.3.1 and Section 3.3.2).

6.2 A Target Joining a Stream

An application may request to join an existing stream. It has to

collect information on the stream including the stream ID (SID) and the IP address of the stream's origin. This can be done out-of- band, e.g. via regular IP. The information is then passed to the local ST agent together with the FlowSpec. The ST agent generates a JOIN message containing the application's request to join the stream and sends it toward the stream origin.

An ST agent receiving a JOIN message, assuming no errors, responds with an ACK. The ACK message must identify the JOIN message to which it corresponds by including the Reference number indicated by the Reference field of the Join message. If the ST agent is not traversed by the stream that has to be joined, it propagates the JOIN message toward the stream's origin. Eventually, an ST agent traversed by the stream or the stream's origin itself is reached. This ST agent responds to the join request based on the join authorization level associated with the stream, cf. Section 4.2.:

o level 0 (refuse join)

It is not allowed to join the stream. No further actions are taken.

o level 1 (ask origin)

The JOIN message is propagated back until the origin is reached. At the origin, the appli- cation is requested to either grant or deny the permission to join the stream. If the permis- sion is denied, no further actions are taken. Otherwise, the origin issues a CONNECT message with a TargetList including the target that requested to join the stream. The target is then added as in normal stream setup.

o level 2 (ok, notify origin)

The ST agent sends a CONNECT message with a TargetList including the target that requested to join the stream. This results in adding the target to the stream. When the ST agent which is already part in the stream receives the ACCEPT message indicating that the new target has been added, it does not propagate the ACCEPT message backwards. Instead, it issues a NOTIFY message with ReasonCode(TargetJoined) to inform the origin of the new target.

o level 3 (ok)

The ST agent sends a CONNECT message with a TargetList including the target that requested to join the stream. This results in adding the target to the stream. When the ST agent which is already part in the stream receives the ACCEPT message indicating that the new target has been added, it does not propagate the ACCEPT message backwards, nor it notifies the origin.

6.2.1 FlowSpec

To Be Written

6.2.2 Router as Origin

To Be Written

6.3 The Origin Removing Targets

The application at the origin specifies a set of targets that are to be removed from the stream and an appropriate ReasonCode (ApplDisconnect). The targets are partitioned into multiple DISCONNECT messages based on the next-hop to the individual targets. If the TargetList is too long to fit into one DISCONNECT message, it is partitioned as described in Section 3.2.2.

An ST agent that receives a DISCONNECT message acknowledges it by sending an ACK back to the previous-hop. The DISCONNECT is also propagated to the relevant next-hop ST agents. Before propagating the message, the TargetList is partitioned based on next-hop ST agents.

If, after deleting the specified targets, any next-hop has no remaining targets, then those resources associated with that next-hop agent may be released. Note that network resources may not actually be released if network multicasting is being used since they may still be required for traffic to other next-hops in the multicast group.

When the DISCONNECT reaches a target, the target sends an ACK and notifies the application that it is no longer part of the stream and for which reason. The ST agent at the target deletes the stream from its database after performing any necessary management and accounting functions. Note that the stream is not deleted if the ST agent is also a router for the stream and there are remaining downstream targets.

6.4 A Target Deleting Itself

The application at the target may inform ST that it wants to be removed from the stream and the appropriate ReasonCode (ApplDisconnect). The agent then forms a REFUSE message with itself as the only entry in the TargetList. The REFUSE is sent back to the origin via the previous-hop. If a stream has multiple targets and one target leaves the stream using this REFUSE mechanism, the stream to the other targets is not affected; the stream continues to exist.

An ST agent that receives such a REFUSE message acknowledges it by

sending an ACK to the next-hop. The target is deleted and, if the next-hop has no remaining targets, then the resources associated with that next-hop agent may be released. Note that network resources may not actually be released if network multicasting is being used since they may still be required for traffic to other next-hops in the multicast group. The REFUSE is also propagated back to the previous-hop ST agent.

When the REFUSE reaches the origin, the origin sends an ACK and notifies the application that the target is no longer part of the stream.

6.5 Changing a Stream's FlowSpec

The application at the sender may wish to change the FlowSpec of an established stream. To do so, it informs the ST agent at the origin of the new FlowSpec and of the list of targets relative to the change. The origin then issues one CHANGE message with the new FlowSpec per next-hop and sends it to the relevant next-hop agents. CHANGE messages are structured and processed similarly to CONNECT messages.

A next-hop agent that is an intermediate agent and receives a CHANGE message similarly determines if it can implement the new FlowSpec along the hop to each of its next-hop agents, and if so, it propagates the CHANGE messages along the established paths. If this process succeeds, the CHANGE messages will eventually reach the targets, which will each respond with an ACCEPT (or REFUSE) message that is propagated back to the origin.

If the change to the FlowSpec is in a direction that makes fewer demands of the involved networks, then the change has a high probability of success along the path of the established stream. Each ST agent receiving the CHANGE message makes the necessary requested changes to the network resource allocations, and if successful, propagates the CHANGE message along the established paths. If the change cannot be made then the ST agent must recover using DISCONNECT and REFUSE messages as in the case of a network failure, see Section 9.2. Note that a failure to change the resources requested for specific targets should not cause other targets in the stream to be deleted.

7 Stream Tear Down

A stream is usually terminated by the origin when it has no further data to send, but may also be partially torn down by the individual targets. These cases will not be further discussed since they have already been described above.

A stream is also torn down if the application should terminate abnormally. Processing in this case is identical to the previous descriptions except that the ReasonCode (ApplAbort) is different.

When all targets have left a stream, the origin notifies the application of that fact, and the application then is responsible for terminating the stream. Note, however, that the application may decide to add targets to the stream instead of terminating it.

8 Exceptional Cases

The previous descriptions covered the simple cases where everything worked. We now discuss what happens when things do not succeed. Included are situations where messages are lost, the requested resources are not available, the routing fails or is inconsistent.

8.1 Setup Failures

8.1.1 Setup Failure due to CONNECT Timeout

When sending a CONNECT message, an ST agent expects an ACK from the next hop ST agent. If the CONNECT fails due to timeout (see Section 11.2), the ST agent sends a REFUSE message back in the direction of the origin with the appropriate ReasonCode (ConnectTimeout).

8.1.2 Setup Failure due to ACCEPT Timeout

An ST agent that propagates an ACCEPT message backward toward the origin expects an ACK from the previous hop ST agent. If the ACCEPT fails due to timeout (see Section 11.2), the ST agent replaces the ACCEPT with a REFUSE and sends a DISCONNECT in the direction toward the target. Both REFUSE and DISCONNECT must identify the affected targets and specify the appropriate ReasonCode (AcceptTimeout).

8.1.3 Setup Failure due to Routing Failures

It is possible for an agent to receive a CONNECT message that contains a known SID, but from an agent other than the previous-hop agent of the stream with that SID. This may be:

1. that two branches of the tree forming the stream have joined back together,

2. the result of an attempted recovery of a partially failed stream, or

3. an erroneous routing loop.

The TargetList contained in the CONNECT is used to distinguish the different cases by comparing each newly received target with those of the previously existing stream:

o if the IP address of the targets differ, it is case 1;

o if the target matches a target in the existing stream, it may be case #2 or #3.

Case #1 is handled in Section 8.2.2 on path convergence. The remaining cases requiring recovery, a partially failed stream and an erroneous routing loop, are not easily distinguishable. In attempting recovery of a failed stream, an agent may issue new CONNECT messages to the affected targets. Such a CONNECT may reach an agent downstream of the failure before that agent has received a DISCONNECT from the neighbourhood of the failure. Until that agent receives the DISCONNECT, it cannot distinguish between a failure recovery and an erroneous routing loop. That agent must therefore respond to the CONNECT with a REFUSE message with the affected targets specified in the TargetList and an appropriate ReasonCode (StreamExists).

The agent immediately preceding that point, i.e., the latest agent to send the CONNECT message, will receive the REFUSE message. It must release any resources reserved exclusively for traffic to the listed targets. If this agent was not the one attempting the stream recovery, then it cannot distinguish between a failure recovery and an erroneous routing loop. It should repeat the CONNECT after a ToConnect timeout, cf. Section 11.2 and Section 8.1.1. If after NConnect retransmissions it continues to receive REFUSE messages, it should propagate the REFUSE message toward the origin, with the TargetList that specifies the affected targets, but with a different error code (RouteLoop).

The REFUSE message with this error code (RouteLoop) is propagated by each ST agent without retransmitting any CONNECT messages. At each agent, it causes any resources reserved exclusively for the listed targets to be released. The REFUSE will be propagated to the origin in the case of an erroneous routing loop. In the case of stream recovery, it will be propagated to the ST agent that is attempting the recovery, which may be an intermediate agent or the origin itself. In the case of a stream recovery, the agent attempting the recovery may issue new CONNECT messages to the same or to different next-hops.

If an agent receives both a REFUSE message and a DISCONNECT message with a target in common then it can release the relevant resources and propagate neither the REFUSE nor the DISCONNECT.

If the origin receives such a REFUSE message, it should attempt to send a new CONNECT to all the affected targets. Since routing errors in an internet are assumed to be temporary, the new CONNECTs will eventually find acceptable routes to the targets, if one exists. If no further routes exist after NRetryRoute tries, the application should be informed so that it may take whatever action it seems necessary.

8.2 Further Issues

8.2.1 Problems due to Routing Inconsistency

When an intermediate agent receives a CONNECT, it invokes the routing algorithm to select the next-hop agents based on the TargetList and the networks to which it is connected. If the resulting next-hop to any of the targets is across the same network from which it received the CONNECT (but not the previous-hop itself), there may be a routing problem. However, the routing algorithm at the previous-hop may be optimizing differently than the local algorithm would in the same situation. Since the local ST agent cannot distinguish the two cases, it should permit the setup but send back to the previous-hop agent an informative NOTIFY message with the appropriate ReasonCode (RouteBack), pertinent TargetList, and in the NextHopIPAddress element the address of the next-hop ST agent returned by its routing algorithm.

The agent that receives such a NOTIFY should ACK it. If the agent is using an algorithm that would produce such behaviour, no further action is taken; if not, the agent should send a DISCONNECT to the next-hop agent to correct the problem.

Alternatively, if the next-hop returned by the routing function is in fact the previous-hop, a routing inconsistency has been detected. In this case, a REFUSE is sent back to the previous- hop agent containing an appropriate ReasonCode (RouteInconsist), pertinent TargetList, and in the NextHopIPAddress element the address of the previous-hop. When the previous-hop receives the REFUSE, it will recompute the next-hop for the affected targets. If there is a difference in the routing databases in the two agents, they may exchange CONNECT and REFUSE messages again. Since such routing errors in the internet are assumed to be temporary, the situation should eventually stabilize.

8.2.2 Path Convergence

It is possible for an agent to receive a CONNECT message that contains a known SID, but from an agent other than the previous hop agent of the stream with that SID. This might be the result of two

branches of the tree forming the stream have joined back together. Other cases are discussed in Section 8.1.3.

This version of ST does not allow streams which have converged path, i.e streams are always tree-shaped and not graph-like. The ST agent which detects this condition informs the previous hop ST agent (the latest ST agent to send the CONNECT message) by sending a NOTIFY message with ReasonCode(PathConverge). Upon receipt of the NOTIFY message, the previous hop ST agent will find alternate route to the listed targets with a different next hop ST agent. If there is no next hop ST-agent other than the one it receives the NOTIFY message from, the ST agent must release any resources reserved for the listed targets and send a REFUSE message with ReasonCode(PathConverge) to its previous hop ST agent. In the same way, the REFUSE message is possibly propagated back by each ST agent. At each agent, it causes any resources reserved exclusively for the listed targets to be released. When the REFUSE reaches the origin, the ST agent at the origin should attempt to send a CONNECT with the listed targets to a different route. If no route exists, or after NRetryRoute tries, the application should be informed so that it may take whatever actions it seems necessary.

8.2.3 Problems in Reserving Resources

If the local or network resources are not available, an ST agent may:

o try alternative paths to the targets: the ST agent calls the routing function to find a different path to the targets. If an alternative path is found, stream connection setup continues in the usual way, as described in Section 3.

o preempt one or more of the already established streams: this way, the ST agent attempts to free enough resources to allow for the new stream to be established. Stream preemption is discussed in Section 9.3.

o refuse to establish the stream along this path: the origin ST agent informs the application of the stream setup failure; an ST agent at a router or target issues a REFUSE message (as described in Section 3.5) with ReasonCode (CantGetResrc).

It depends on the local implementations whether an ST agent tries alternative paths or preempts other streams. Also, the order of the actions taken is not defined here. In any case, if enough resources cannot be found over different paths or as a consequence of stream preemption, the agent has to explicitly refuse to establish the stream.

8.2.4 Problems Caused by CHANGE Messages

A CHANGE might fail for several reasons, including:

o the request may be for a larger amount of network resources when those resources are not available;

o it might be required that all the former resources are released before the new ones are requested and, due to unlucky timing, an unrelated request for network resources might be processed between the time the resources are released and the time the new resources are requested, so that the former resources are no longer available.

If the attempt to change the FlowSpec fails then the ST agent where the failure occurs must intentionally break the affected portion of the stream. This is done by sending REFUSE and DISCONNECT messages with ReasonCode (ChgFailed).

9 Failure Detection and Recovery

9.1 Failure Detection

The ST failure detection mechanism is based on two assumptions:

1. If a neighbor of an ST agent is up, and has been up without a disruption, and has not notified the ST agent of a problem with streams that pass through both, then the ST agent can assume that there has not been any problem with those streams.

2. A network through which an ST agent has routed a stream will notify the ST agent if there is a problem that affects the stream data packets but does not affect the control packets.

The purpose of the robustness protocol defined here is for ST agents to determine that the streams through a neighbor have been broken by the failure of the neighbor or the intervening network. This protocol should detect the overwhelming majority of failures that can occur. Once a failure is detected, the recovery procedures described in Section 9.2 are initiated by the ST agents.

9.1.1 Network Failures

An ST agent can detect network failures by two mechanisms:

o the network can report a failure, or

o the ST agent can discover a failure by itself.

They differ in the amount of information that an ST agent has available to it in order to make a recovery decision. For example, a network may be able to report that reserved bandwidth has been lost and the reason for the loss and may also report that connectivity to the neighboring ST agent remains intact. In this case, the ST agent may request the network to allocate bandwidth anew. On the other hand, an ST agent may discover that communication with a neighboring ST agent has ceased because it has not received any traffic from that neighbor in some time period. If an ST agent detects a failure, it may not be able to determine if the failure was in the network while the neighbor remains available, or the neighbor has failed while the network remains intact.

9.1.2 Detecting ST Agents Failures

Each ST agent periodically sends each neighbour with which it shares one or more streams a HELLO message. This message exchange is between ST agents, not entities representing streams or applications. That is, an ST agent need only send a single HELLO message to a neighbour regardless of the number of streams that flow between them. All ST agents (host as well as intermediate) must participate in this exchange. However, only agents that share active streams need to participate in this exchange and it is an error to send a HELLO message to a neighbour ST agent with no streams in common, e.g. to check whether it is active. Note that STATUS messages can be used to poll neighbour ST agents.

A HELLO message is ACKed if the Reference field is non-zero. As well as identifying the sender, the HELLO message has two fields:

o a HelloTimer field that is in units of milliseconds modulo the maximum for the field size, and

o a Restarted-bit specifying that the ST agent has been restarted recently.

The HelloTimer must appear to be incremented every millisecond whether a HELLO message is sent or not, but it is allowable for an ST agent to create a new HelloTimer only when it sends a HELLO message. The HelloTimer wraps around to zero after reaching the maximum value. Whenever an ST agent suffers a catastrophic event that may result in it losing ST state information, it must reset its HelloTimer to zero and must set the Restarted-bit for the following HelloTimerHoldDown seconds.

Each ST stream has a RecoveryTimeout value associated with it. This value is assigned by the origin and carried into the CONNECT message, see Section 16.4.

An ST agent must send HELLO messages to its neighbour with a period shorter than the smallest RecoveryTimeout of all the active streams that pass between the two agents, regardless of direction. This period must be smaller by a factor, called HelloLossFactor, which is at least as large as the greatest number of consecutive HELLO messages that could credibly be lost while the communication between the two ST agents is still viable.

An ST agent may send simultaneous HELLO messages to all its neighbors at the rate necessary to support the smallest RecoveryTimeout of any active stream. Alternately, it may send HELLO messages to different neighbors independently at different rates corresponding to RecoveryTimeouts of individual streams.

The agent that receives a HELLO message expects to receive at least one new HELLO message from a neighbor during the RecoveryTimeout of every active stream through that neighbor. It can detect duplicate or delayed HELLO messages by saving the HelloTimer field of the most recent valid HELLO message from that neighbor and comparing it with the HelloTimer field of incoming HELLO messages. It will only accept an incoming HELLO message from that neighbor if it has a HelloTimer field that is greater than the most recent valid HELLO message by the time elapsed since that message was received plus twice the maximum likely delay variance from that neighbor. If the ST agent does not receive a valid HELLO message within the RecoveryTimeout of a stream, it must assume that the neighboring ST agent or the communication link between the two has failed and it must initiate stream recovery activity.

Furthermore, if an ST agent receives a HELLO message that contains the Restarted-bit set, it must assume that the sending ST agent has lost its ST state. If it shares streams with that neighbor, it must initiate stream recovery activity. If it does not share streams with that neighbor, it should not attempt to create one until that bit is no longer set. If an ST agent receives a CONNECT message from a neighbor whose Restarted-bit is still set, it must respond with ERROR with the appropriate ReasonCode (RemoteRestart). If it receives a CONNECT message while its own Restarted-bit is set, it must respond with ERROR with the appropriate ReasonCode (RestartLocal).

9.2 Failure Recovery

If an intermediate agent fails or a network or part of a network fails, the previous-hop agent and the various next-hop agents will discover the fact by the failure detection mechanism described in Section 9.1.

The recovery of an ST stream is a relatively complex and time

consuming effort because it is designed in a general manner to operate across a large number of networks with diverse characteristics. Therefore, it may require information to be distributed widely, and may require relatively long timers. On the other hand, since a network is a homogeneous system, failure recovery in the network may be a relatively faster and simpler operation. Therefore an ST agent that detects a failure should attempt to fix the network failure before attempting recovery of the ST stream. If the stream that existed between two ST agents before the failure cannot be reconstructed by network recovery mechanisms alone, then the ST stream recovery mechanism must be invoked.

If stream recovery is necessary, the different ST agents may need to perform different functions, depending on their relation to the failure:

o An agent that is a next-hop of a failure should first verify that there was a failure. It can do this using STATUS messages to query its upstream neighbor. If it cannot communicate with that neighbor, then it should first send a REFUSE message with the appropriate ReasonCode ("failure") to the neighbor to speed up the failure recovery in case the hop is unidirectional, i.e., the neighbor can hear the agent but the agent cannot hear the neighbor. The ST agent detecting the failure must then send DISCONNECT messages with the same ReasonCode toward the targets.

The intermediate agents process this DISCONNECT message just like the DISCON- NECT that tears down the stream. However, a target ST agent that receives a DISCONNECT message with the appropriate ReasonCode ("failure") will maintain the stream state and notify the next higher protocol of the failure. In effect, these DISCONNECT messages tear down the stream from the point of the failure to the targets, but inform the targets that the stream may be fixed shortly.

o An agent that is the previous-hop before the failed component first verifies that there was a failure by querying the downstream neighbor using STATUS messages. If the neighbor has lost its state but is available, then the ST agent may reconstruct the stream if the NoRecovery option is not selected. If it cannot communicate with the next-hop, then the agent detecting the failure releases any resources that are dedicated exclusively to sending data on the broken branch and sends a DISCONNECT message with the appropriate ReasonCode ("failure") toward the affected targets. It does so to speed up failure recovery in case the communication may be unidirectional and this message might be delivered successfully.

The agent that is the previous-hop before the failed component can attempt to recover the streams for which the NoRecovery option is not

selected:

o If the NoRecovery option is selected, then the ST agent sends a REFUSE message with the appropriate ReasonCode ("failure") to the previous-hop. The TargetList in these messages contains all the targets that were reached through the broken branch. Multiple REFUSE mes- sages may be required if the PDU is too long for the MTU of the intervening network. The REFUSE message is propagated all the way to the origin, which can attempt recovery of the stream by sending a new CONNECT to the affected targets. The new CONNECT will be treated by intermediate ST agents as an addition of new targets into the established stream.

o If the NoRecovery option is not selected, the ST agent can attempt recovery of the stream. It does so by issuing a new CONNECT message to the affected targets. If the ST agent can- not find new routes to some targets, or if the only route to some targets is through the previ- ous-hop, then it sends one or more REFUSE messages to the previous-hop with the appropriate ReasonCode ("failure") specifying the affected targets in the TargetList. The pre- vious-hop can then attempt recovery of the stream by issuing a CONNECT to those targets. If it cannot find an appropriate route, it will propagate the REFUSE message toward the ori- gin.

Regardless of which agent attempts recovery of a damaged stream, it will issue one or more CONNECT messages to the affected targets. These CONNECT messages are treated by intermediate ST agents as additions of new targets into the established stream. The FlowSpecs of the new CONNECT messages are the same as the ones contained in the most recent CONNECT or CHANGE messages that the ST agent had sent toward the affected targets when the stream was operational.

9.2.1 Problems in Stream Recovery

The reconstruction of a broken stream may not proceed smoothly. Since there may be some delay while the information concerning the failure is propagated throughout an internet, routing errors may occur for some time after a failure. As a result, the ST agent attempting the recovery may receive ERROR messages for the new CONNECTs that are caused by internet routing errors. The ST agent attempting the recovery should be prepared to resend CONNECTs before it succeeds in reconstructing the stream. If the failure partitions the internet and a new set of routes cannot be found to the targets, the REFUSE messages will eventually be propagated to the origin, which can then inform the application so it can decide whether to terminate or to continue to attempt recovery of the stream.

The new CONNECT may at some point reach an ST agent downstream of the

failure before the DISCONNECT does. In this case, the agent that receives the CONNECT is not yet aware that the stream has suffered a failure, and will interpret the new CONNECT as resulting from a routing failure. It will respond with an ERROR message with the appropriate ReasonCode (StreamExists). Since the timeout that the ST agents immediately preceding the failure and immediately following the failure are approximately the same, it is very likely that the remnants of the broken stream will soon be torn down by a DISCONNECT message with the appropriate ReasonCode ("failure"). Therefore, the ST agent that receives the ERROR message with ReasonCode (StreamExists) should retransmit the CONNECT message after the ToConnect timeout expires. If this fails again, the request will be retried for NConnect times. Only if it still fails will the ST agent send a REFUSE message with the appropriate ReasonCode (RouteLoop) to its previous-hop. This message will be propagated back to the ST agent that is attempting recovery of the damaged stream. That ST agent can issue a new CONNECT message if it so chooses. The REFUSE is matched to a CONNECT message created by a recovery operation through the LnkReference field in the CONNECT.

ST agents that have propagated a CONNECT message and have received a REFUSE message should maintain this information for some period of time. If an agent receives a second CONNECT message for a target that recently resulted in a REFUSE, that agent may respond with a REFUSE immediately rather than attempting to propagate the CONNECT. This has the effect of pruning the tree that is formed by the propagation of CONNECT messages to a target that is not reachable by the routes that are selected first. The tree will pass through any given ST agent only once, and the stream setup phase will be completed faster.

If a CONNECT message reaches a target, the target should as efficiently as possible use the state that it has saved from before the stream failed during recovery of the stream. It will then issue an ACCEPT message toward the origin. The ACCEPT message will be intercepted by the ST agent that is attempting recovery of the damaged stream, if not the origin. If the FlowSpec contained in the ACCEPT specifies the same selection of parameters as were in effect before the failure, then the ST agent that is attempting recovery will not propagate the ACCEPT. If the selections of the parameters are different, then the agent that is attempting recovery will send the origin a NOTIFY message with the appropriate ReasonCode (FailureRecovery) that contains a FlowSpec that specifies the new parameter values. The origin may then have to change its data generation characteristics and the stream's parameters with a CHANGE message to use the newly recovered subtree.

9.3 Stream Preemption

An intermediate ST agent may decide to break a stream intentionally. This is called stream preemption. Usually streams are preempted in order to free resources for a new stream which has a higher priority. ST does not define when stream preemption should be used but it provides the means to implement it.

If an ST agent decides that it is necessary to preempt one or more of the stream traversing it, the decision on which streams have to be preempted has to be made. ST provides two ways of optimizing such decision:

1. Streams can be assigned an StreamImportance value from 0 (most important) to 7 (least important). This value is carried in the CONNECT message when the stream is setup, see Section 16.4.

2. An application may specify that a set of streams are related to each other and that they are all candidate for preemption if one of them gets preempted. It can be done by using the fate- sharing relationship defined in Section 10. This helps making a good choice when more than one stream have to be preempted, because it leads to breaking a single application as oppo- site to as many applications as the number of preempted streams.

Stream preemption requires the following actions from the ST agents:

o An intermediate agent that breaks the stream intentionally sends DISCONNECT mes- sages with the appropriate ReasonCode (StreamPreempted) toward the affected targets. It sends a REFUSE message with the appropriate ReasonCode (StreamPreempted) to the previ- ous-hop.

o A previous-hop agent of the preempted stream acts as in case of failure recovery, cf. Sec- tion 9.2. If the NoRecovery option is set, is propagates the REFUSE message back to the ori- gin. If the NoRecovery option is not set, it attempts to rebuild the deleted paths and, in case this does not work, it propagates the REFUSE message to the previous-hop.

o A target or next-hop agent of the preempted stream acts as in case of failure recovery, cf. Section 9.2. It releases resources that are allocated to the stream, but it maintains the internal state information describing the stream for some time in case the stream is quickly fixed.

Note that, as opposite to failure recovery, there is no need to verify that the failure actually occurred, because this is explicitly indicated by the ReasonCode (StreamPreempted).

10 A Group of Streams

There may be need to associate related streams. The group mechanism is simply an association technique that allows ST agents to identify the different streams that are to be associated.

A group consists of a set of streams and a relationship. The set of streams may be empty. The relationship applies to all group members. Each group is identified by a group name. The group name is unique across the Internet.

Streams belong to the same group if they have the same GroupName in the GroupName field of the Group parameter. The relationship is defined by the Relationship field. Group membership must be specified at stream creation time and persists for the whole stream lifetime. A single stream may belong to multiple groups.

The ST agent that creates a new group is called group initiator. Any ST agent can be a group initiator. The initiator allocates the GroupName and the Relationship among group members. The initiator may or may not be the origin of a stream belonging to the group. The group name has to be generated as described in Section 10.1. Relationships defined by this version of the protocol are listed in Section 10.2.

10.1 Group Name Generator

The GroupName includes a 16-bit unique identifier, a 32-bit IP address, and a 32-bit creation timestamp. It is defined in Section 10.3. An ST implementation has to provide a group name generator facility, so that an application or higher layer protocol can obtain a unique GroupName from the ST layer. This is a mechanism for the application to request the allocation of a GroupName that is independent of the request to create a stream. The GroupName is used by the application or higher layer protocol when creating the streams that are to be part of the group.

For instance, the following two functions could be made available:

- AllocateGroupName() -> result, GroupName

- ReleaseGroupName() -> result

10.2 Basic ST Relationships

This version of ST defines four basic relationships. An ST2PLus implementation must support all four basic relationships. The basic relationships are described in detail below in Section 10.2.1 -

Section 10.2.4.

ST provides the means to define new relationships as the need for them becomes clear in the future. This can be done by assigning one of the unused bits of the Relationship field of the Group parameter.

The next sections describe the four basic relationships.

10.2.1 Bandwidth Sharing

Streams belonging to this group share the same network bandwidth. This is intended to support applications as audio conferences where, of all participants, only some at a time are allowed to speak. In such a scenario, global bandwidth utilization can be optimized, e.g. it is sufficient to reserve bandwidth for a small set of audio streams.

The N parameter indicates for how many streams at the same time bandwidth should be allocated. An ST agent allocates N times the bandwidth required by the most demanding stream in the group, say Bmax. If the application intends for instance to allow three participants to speak at the same time, N has a value of three and the ST agent will allocate for the group an amount of bandwidth equal to 3*Bmax.

This mechanism does not always allocate an optimal amount of bandwidth (as when a stream requires 4 Mbits/s and all the other streams in the same group require 1 Mbits/s only: N=3 causes the allocation of 12 Mbits/s). However, it is simple to implement and it works well with streams that have homogeneous requirements. An alternative would be to keep track of the single streams requirements and allocate the exact amount of bandwidth.

An ST agent always attempts to reserve N*Bmax bandwidth. If less bandwidth than N*Bmax is available, the new stream is not built. If bandwidth for the group has already been allocated and a new stream with a bandwidth demand inferior to Bmax is being established, the ST agent, depending on the local implementation, may not need to contact the local resource manager and it can proceed directly with the stream setup.

Note that ST agents become aware of a group's requirements only when the streams belonging to the group are created. In case of the bandwidth sharing relationship, an application should attempt to establish the most demanding streams first to minimize stream setup efforts. If on the contrary the less demanding streams are built first, it will be always necessary to allocate additional bandwidth in consecutive steps as the most demanding streams are built.

10.2.2 Fate Sharing

Streams belonging to this group share the same fate. If a stream is deleted, the other members of the group are also deleted. This is intended to support stream preemption by indicating which streams are mutually related. If preemption of multiple streams is necessary, this information can be used to delete a set of related streams, e.g. with impact on a single application, instead of making a random choice with the possible effect of interrupting several different applications.

This relationship provides a hint on which streams should be preempted. Still, the entity responsible for the preemption is not forced to behave accordingly, and other streams could be preempted first based on different criteria.

10.2.3 Route Sharing

Streams belonging to this group share the same paths. This can be desirable for several reasons, e.g. to exploit the same allocated resources or in the attempt to maintain the transmission order. An ST agent attempts to select the same path although the way this is implemented depends heavily on the routing algorithm which is used.

If the routing algorithm is sophisticated enough, an ST agent can suggest that a stream is routed over an already established path. Otherwise, it can ask the routing algorithm for a set of legal routes to the destination and check whether the desired path is included in those feasible.

Route sharing is a hint to the routing algorithm used by ST. Failing to route a stream through the shared path does not normally cause the deletion of the stream: the stream is built over an alternative path whenever possible.

10.2.4 Subnet Resources Sharing

Streams belonging to this group share the same MAC layer subnetwork addresses. As an example, the same MAC layer multicast address can be used for all the streams in a given group. This mechanism allows for a better utilization of MAC layer multicast addresses and it is especially useful when used with network adapters that offer a very small number of MAC layer multicast addresses.

This relationship provides a hint to the data link layer functions.

10.3 Relationships Orthogonality

The four basic relationships, as they have been defined, are orthogonal. This means, any combinations of the basic relationships are allowed. For instance, let's consider an application that requires full-duplex service for a stream with multiple targets. Also, let's suppose that only N targets are allowed to send data back to the origin at the same time. In this scenario, all the reverse streams could belong to the same group. They could be sharing both the paths and the bandwidth. The Path&Bandwidth sharing relationship is obtained from the basic set of relationships. This example is important because it shows how full-duplex service can be obtained in ST.

As new relationships are defined, it should be indicated whether they are or not orthogonal with respect to the previously defined ones. This will be reflected by illegal values for the Relationship field of the Group parameter (see Section 15.2.3).

11 Ancillary Functions

11.1 Stream IDs Generation

To Be Written

11.2 SCMP Reliability

The ST Control Message Protocol is made reliable through the use of retransmission when response is not received in a timely manner. In general, when sending a SCMP messages which requires an ACK back, the sending ST agent needs to set the Toxxxx timer (where xxxx is the SCMP message type, e.g. ToConnect). If it does not receive an ACK back before the Toxxxx timer expires, the ST agent should retransmit the SCMP message. If no ACK has been received within Nxxxx retransmissions, then a SCMP timeout condition occurs and the ST agent enters its SCMP timeout recovery state. The actions performed by the ST agent as the result of the SCMP timeout condition differ for different SCMP message. In some cases (CONNECT,ACCEPT) the ST agent handles the timeout by sending additional SCMP message (REFUSE/DISCONNECT) to its neighbour ST agents (see Section 8.1.1 & Section 8.1.2), while in other cases (REFUSE, DISCONNECT) it simply gives up sine there is nothing else it can do.

For some SCMP messages (CONNECT,CHANGE) the sending ST agent also expects a response back (ACCEPT/REFUSE) after ACK has been received. For these cases, the ST agent needs to set the Rtoxxxx timer after it receives the ACK. If it does not receive the appropriate response back when Rtoxxxx expires, the ST agent updates its state data and perform appropriate recovery action as described in other sections.

Timeout and retransmission algorithm is implementation dependent and it is outside the scope of this document. However, it must be reasonable enough not to cause excessive retransmission of SCMP message while maintain the robustness of the protocol. Algorithms on this subject are described in [RFC1122], [Jaco88], [KaPa87].

11.3 IP Encapsulation of ST

ST packets may be encapsulated in IP to allow them to pass through routers that don't support the ST Protocol. Of course, ST resource management is precluded over such a path, and packet overhead is increased by encapsulation, but if the performance is reasonably predictable this may be better than not communicating at all.

IP-encapsulated ST packets begin with a normal IP header. Most fields of the IP header should be filled in according to the same rules that apply to any other IP packet. Three fields of special interest are:

o Protocol is 5 to indicate an ST packet is enclosed, as opposed to TCP or UDP, for example. The assignment of protocol 5 to ST is an arranged coincidence with the assignment of IP Version 5 to ST [RFC1190].

o Destination Address is that of the next-hop ST agent. This may or may not be the target of the ST stream. There may be an intermediate ST agent to which the packet should be routed to take advantage of service guarantees on the path past that agent. Such an intermediate agent would not be on a directly-connected network (or else IP encapsulation wouldn't be needed), so it would probably not be listed in the normal routing table. Additional routing mechanisms, not defined here, will be required to learn about such agents.

o Type-of-Service may be set to an appropriate value for the service being requested (usually low delay, high throughput, normal reliability). This feature is not implemented uniformly in the Internet, so its use can't be precisely defined here.

IP encapsulation adds little difficulty for the ST agent that receives the packet. However, when IP encapsulation is performed it must be done in both directions. To process the encapsulated IP message, the ST agents simply remove the IP header and proceed with ST header as usual.

The more difficult part is during setup, when the ST agent must decide whether or not to encapsulate. If the next-hop ST agent is on a remote network and the route to that network is through a router that supports IP but not ST, then encapsulation is required. The ST agents make encapsulation decision based on information provided by

routing function to indicate whether the routers in the path support ST. It is outside the scope of this document to address routing function and therefore neither its algorithm nor implementation is specified here. ST assumes that appropriate routing algorithm exists to which ST has access.

On forwarding, the (mostly constant) IP Header must be inserted and the IP checksum appropriately updated.

11.4 IP Multicasting

To Be Written

11.5 Routing

To Be Written

11.6 Security

To Be Written

12 FlowSpec

The FlowSpec is used to convey stream service requirements end-to-end. The contents of the FlowSpec are transparent to the ST agents. An ST agent extracts the FlowSpec from the correspondent incoming SCMP message and passes it to the LRM as required. The LRM updates the FlowSpec values based on the amount of resources that it has allocated to the stream.

12.1 FlowSpec Versions

ST is not dependent on a particular FlowSpec format and it is expected that other versions of the FlowSpec than those introduced below in this section will be needed in the future. Different FlowSpec formats are distinguished by the value of the Version field. The following values are reserved:

0 - Null FlowSpec /* mandatory */

1 - ST Version 1

2 - ST Vesrion 1.5

3 - RFC 1190 FlowSpec

4 - HeiTS FlowSpec

5 - BerKom FlowSpec

6 - RFC 1363 FlowSpec

7 - ST2Plus FlowSpec /* mandatory */

A single stream is always associated to a single FlowSpec format. Changes to the FlowSpec are also relative to the same FlowSpec format, i.e. the value of the Version field cannot be changed during the lifetime of the stream.

12.2 The Null FlowSpec (#0)

The FlowSpec identified by a value of 0 for its Version field is called the "Null FlowSpec". An ST agent that receives the Null FlowSpec always assumes that sufficient resources for the stream are available. The Null FlowSpec fields values are never updated. Stream setup takes place in the usual way, but no resources are actually reserved.

The main purpose of the Null FlowSpec is that of facilitating interoperability tests by allowing streams to be built without actually allocating the correspondent amount of resources. The Null FlowSpec may also be used for testing and debugging purposes.

The complete format is specified in Section 15.2.2.

12.3 The ST Current FlowSpec (#7)

FlowSpec #7 is the FlowSpec to be used by the current version of ST. It contains values for 3 basic QoS parameters: message size, throughput, and delay. Also, it is possible to specify a QoS class, e.g. guaranteed. Each parameter has to be expressed via a set of values:

o the "desired" values are assigned by the application and never changed by the LRM

o the "limit" values are assigned by the application and never changed by the LRM

o the "actual" values indicate the guarantees that the system is able to provide. They are updated by the LRM at each node. The "actual" values are always bounded by the "limit" values.

12.3.1 Qos Classes

We also define two QoS classes:

1. QOS_GUARANTEED

2. QOS_PREDICTIVE

o The QOS_GUARANTEED service class implies that the negotiated QoS for the stream is never violated during the data transfer. For instance, the desired rate is the peak rate for the transmission. This may sometimes lead to overbooking of resources, but it provides strict real-time guarantees.

o The QOS_PREDICTIVE service class implies that the negotiated QoS may be violated for short time intervals. Reservations are done for the average case as opposite to the peak case required by the QOS_GUARANTEED service class.

If a LRM that doesn't support class QOS_PREDICTIVE (QOS_GUARANTEED) receives a FlowSpec containing a QOS_PREDICTIVE (QOS_GUARANTEED) class, it informs the local ST agent. The ST agent may try different paths or delete the correspondent portion of the stream with ReasonCode (QoSClassUnknown).

12.3.2 Maximum Message Size

This parameter is expressed in bytes. It represents the maximum size allowed for messages sent as part of the stream. The LRM first checks whether it is possible to get the value desired by the application (DesMaxSize). If not, it updates the actual value (ActMaxSize) with the available size unless this value is inferior to the minimum allowed by the application (LimitMaxSize), in which case it informs the local ST agent that it is not possible to build the stream along this path.

12.3.3 Rate or Throughput

This parameter is expressed in bytes/seconds. It represents the transmission rate for the stream. The LRM first checks whether it is possible to get the value desired by the application (DesRate). If not, it updates the actual value (ActRate) with the available rate unless this value is inferior to the minimum allowed by the application (LimitRate), in which case it informs the local ST agent that it is not possible to build the stream along this path.

12.3.4 Maximum Delay and Delay Jitter

This parameter is expressed in milliseconds. It represents the maximum end-to-end for the stream. The LRM first checks whether it is possible to get the value desired by the application (DesMaxDelay). If not, it updates the actual value (ActMaxDelay) with the available

rate unless this value is greater than the maximum delay allowed by the application (LimitMaxDelay), in which case it informs the local ST agent that it is not possible to build the stream along this path.

The LRM also updates the MinDelay field by adding the minimum possible delay to the next- hop. Information on the minimum possible delay allows to calculate another important QoS parameter, the delay jitter.

The complete format is specified in Section 15.2.2.

13 ST State Machines

To Be Written

ST Protocol Data Units

The ST PDUs sent between ST agents consist of an ST Header encapsulating either a higher layer PDU or an ST Control Message. Since ST operates as an extension of IP, the packet arrives at the same network service access point that IP uses to receive IP datagrams, e.g., ST would use the same ethertype (0x800) as does IP. The two types of packets are distinguished by the IP Version Number field (the first four bits of the packet); IP currently uses a value of 4, while ST has been assigned the value 5 (see [RFC791]). There is no requirement for compatibility between IP and ST packet headers beyond the first four bits.

The ST Header also includes an ST Version Number, a total length field, a header checksum, a unique id, and the origin IP address as shown in Figure 8. See Section 18 for an explanation of the notation.

Figure 8: ST Header

o ST is the IP Version Number assigned to identify ST packets. The value for ST is 5.

o Ver is the ST Version Number. This document defines ST Version 3.

o ??? (TBD)

o D (bit 8) is set to 1 in all ST data packets and to 0 in all SCMP control messages.

o TotalBytes is the length, in bytes, of the entire ST packet, it includes the ST Header but does not include any local network headers or trailers. In general, all length fields in the ST Proto-

col are in units of bytes.

o HeaderChecksum covers only the ST Header (12 bytes). The ST Protocol uses 16-bit checksums here in the ST Header and in each Control Message. The standard Internet check- sum algorithm is used: "The checksum field is the 16-bit one's complement of the one's com- plement sum of all 16-bit words in the header. For purposes of computing the checksum, the value of the checksum field is zero (0)." See [RFC1071], [RFC1141], and [RFC791] for sug- gestions for efficient checksum algorithms.

o UniqueID is the first element of the stream id (SID). It is locally unique at the origin.

o OriginIPAddress is the second element of the SID. It is the IP address of the origin.

14 ST Data Packets

ST packets whose D-bit is non-zero are data packets. Their interpretation is a matter for the higher layer protocols and consequently is not specified here. The data packets are not protected by an ST checksum and will be delivered to the higher layer protocol even with errors. ST agents will not pass data packets over a new hop whose setup is not complete.

14.1 Stream ID

The UniqueID and OriginIPAddress fields form the Stream ID (SID), which is used by the ST agents to identify which stream the data packet belongs to. The same SID is used in data packets and control messages.

In certain situations, e.g. usually due to a crash and subsequent reboot, it is possible that an ST agent receives a data packet belonging to a stream of which the agent has lost state information. In this case, the agent is not able to forward the packet and has to discard it. SIDs include the IP address of the origin. This allows to request the origin that the unidentified data flow is stopped.

15 SCMP Protocol Data Units

ST Control Messages are between a previous-hop agent and its next-hop agents using a D-bit of zero (0). The control protocol follows a request-response model with all requests expecting responses. Retransmission after timeout (see Section 11.2) is used to allow for lost or ignored messages. Control messages do not extend across packet boundaries; if a control message is too large for the MTU of a

hop, its information is partitioned and a control message per partition is sent (see Section 3.2.2). All control messages have the following format:

Figure 9: ST Control Message Format

o OpCode identifies the type of control message. Each is described in detail in following sec- tions.

o Options is used to convey OpCode-specific variations for a control message.

o TotalBytes is the length of the control message, in bytes, including all OpCode specific fields and optional parameters. The value is always divisible by four (4).

o Reference is a transaction number. Each sender of a request control message assigns a Refer- ence number to the message that is unique with respect to the stream. The Reference number is used by the receiver to detect and discard duplicates. Each acknowledgment carries the Reference number of the request being acknowledged. Reference zero (0) is never used, and Reference numbers are assumed to be monotonically increasing with wraparound so that the older-than and more-recent-than relations are well defined.

o LnkReference contains the Reference field of the request control message that caused this request control message to be created. It is used in situations where a single request leads to multiple responses from the same ST agent. Examples are CONNECT and CHANGE mes- sages that are first acknowledged hop-by-hop and then lead to an ACCEPT or REFUSE response from each target.

o SenderIPAddress is the 32-bit IP address of the network interface that the ST agent used to send the control message. This value changes each time the packet is forwarded by an ST agent (hop-by-hop).

o Checksum is the checksum of the control message. Because the control messages are sent in packets that may be delivered with bits in error, each control message must be checked before it is acted upon.

o ReasonCode is set to zero (0 = NoError) in most SCMP messages. Otherwise, it can be set to an appropriate value to indicate an error situation as defined in Section 17.3.

o OpCode Specific Data contains any additional information that is associated with the con- trol message. It depends on the specific

control message and is explained further below. In some response control messages, fields of zero (0) are included to allow the format to match that of the corresponding request message. The OpCode Specific Data may also contain any of the optional Parameters defined in Section 15.2.

15.1 ST Control Messages

The CONNECT message is used to establish a stream. It is an end-to-end message created by the origin. It propagates all the way to the targets, and require an ACK in response. It causes the targets to issue ACCEPT or REFUSE messages. The CONNECT message is also used to add one or more targets to an existing stream and during recovery of a broken stream.

The CHANGE message is used to change the characteristics of an established stream. It is processed similarly to the CONNECT message, but it propagates along an already established stream.

The ACCEPT message is an end-to-end message generated by a target and is used to signify the successful completion of the setup of a stream or part of a stream, or the change of the FlowSpec. There are no other messages that are similar to it.

The REFUSE message is sent by a target to refuse the setup of a stream or the change of the FlowSpec. In these cases, it is an end-to-end message. An intermediate ST agent issues a REFUSE if it cannot find a route to a target, can only find a route to a target through the previous-hop, preempts a stream, or detects a failure in a next-hop ST agent or network. In all cases a REFUSE propagates in the direction toward the origin.

The JOIN-REQUEST message is used to request to join an already established stream. It propagates in the upstream direction until either the origin is hit or a router that is traversed by the stream. As a consequence of this message, a CONNECT will be delivered to the target that requested to join. An authorization scheme prevents undesired destinations to join the stream.

The DISCONNECT message is used to tear down streams or parts of streams. It propagates in the direction from the origin toward the targets. It is either used as an end-to-end message generated by the origin that is used to completely tear down a stream, or is generated by an intermediate ST agent that preempts a stream or detects the failure of its previous-hop agent or network in the stream. In the latter case, it is used to tear down the part of the stream from the failure to the targets, thus the message propagates all the way to the targets.

Usually, SCMP messages are acknowledged by the receiver. This is done via the ACK message. If an SCMP message contains errors and it cannot be identified or interpreted, an ERROR message is issued. Other SCMP messages include HELLO, NOTIFY, STATUS, and STATUS- RESPONSE.

The following sections contain descriptions of common fields and parameters, followed by descriptions of the individual control messages, both listed in alphabetical order. A brief description of the use of the control message is given. The packet format is shown graphically.

15.2 Common SCMP Elements

Several fields and parameters (referred to generically as elements) are common to two or more PDUs. They are described in detail here instead of repeating their description several times. In many cases, the presence of a parameter is optional. To permit the parameters to be easily defined and parsed, each is identified with a PCode byte that is followed by a PBytes byte indicating the length of the parameter in bytes (including the PCode, PByte, and any padding bytes). If the length of the information is not a multiple of 4 bytes, the parameter is padded with one to three zero (0) bytes. PBytes is thus always a multiple of four (4). Parameters can be present in any order.

15.2.1 ErroredPDU

The ErroredPDU parameter (PCode = 1) is used for diagnostic purposes to encapsulate a received ST PDU that contained an error. It may be optionally included in the ERROR message. Its use is primarily diagnostic.

Figure 10: ErroredPDU

o PDUBytes indicates how many bytes of the PDUInError are actually present.

o PDUInError is the PDU in error, beginning with the ST Header.

15.2.2 FlowSpec

The FlowSpec is used to convey stream service requirements end-to-end. We expect that other versions of FlowSpec will be needed in the future, which may or may not be subsets or supersets of the version described here. PBytes will allow new constraints to be added to the end without having to simultaneously update all implementations in the field.

The FlowSpec parameter (PCode = 2) is used in several messages to convey the FlowSpec. The format of the FlowSpec field depends on the FlowSpec version. For details on the current Flowspec version, see Section 12.

Figure 11: FlowSpec

15.2.3 Group

The Group parameter (PCode = 3) is an optional argument used to indicate that the stream is a member of the specified group.

Figure 12: Group Parameter

o GroupUniqueID, GroupInitiatorIPAddress, and GroupCreationTime are allocated by the group name generator functions (see Section 10.1). These three fields together form the GroupName field.

o Relationship has the following format:

Figure 13: Relationship Field

The B, F, P, S bits correspond to Bandwidth, Fate, Path, and Subnet resources sharing. A value of 1 indicates that the relationship exists. All combinations of these four bits are allowed because the four basic relationships are orthogonal. Bits 0-11 of the Relationship field are reserved for future use and must be set to 0.

N contains a legal value only if the B-bit is set. It is the value of the N parameter to be used as explained in Section 10.2.1.

15.2.4 MulticastAddress

The MulticastAddress parameter (PCode = 4) is an optional parameter that is used when setting up a network level multicast group, to communicate an IP and/or local network multicast address to the next-hop agents that should become members of the group.

Figure 14: MulticastAddress

o LocalNetBytes is the length of the Local Net Multicast Address.

o IPMulticastAddress is described in [RFC1112]. This field is zero (0) if no IP multicast address is known or is applicable. The block of addresses 224.1.0.0 - 224.1.255.255 has been allocated for use by ST, see Section 17.4.

o Local Net Multicast Address is the multicast address to be used on the local network. It cor- responds to the IPMulticastAddress when the latter is non-zero.

15.2.5 NextHopIPAddress

The NextHopIPAddress parameter (PCode = 5) is an optional parameter of NOTIFY or REFUSE and contains the IP address of a suggested next-hop ST agent.

Figure 15: NextHopIPAddress

15.2.6 Origin

The Origin parameter (PCode = 6) is used to identify the next higher protocol, and the SAP being used in conjunction with that protocol.

Figure 16: Origin

o NextPcol is an 8-bit field used in demultiplexing operations to identify the protocol to be used above ST. The values of NextPcol are in the same number space as the IP Header's Pro- tocol field and are consequently defined in the Assigned Numbers RFC [RFC791].

o OriginSAPBytes specifies the length of the OriginSAP, exclusive of any padding required to maintain 32-bit alignment.

o OriginSAP identifies the origin's SAP associated with the NextPcol protocol.

Note that the IP address of the origin is not included in this parameter because it is always available as part of the ST header.

15.2.7 RecordRoute

The RecordRoute parameter (PCode = 7) may be used to request that the route between the origin and a target be recorded and returned to the origin. It is included in the CONNECT and ACCEPT messages.

Figure 17: RecordRoute

o FreeOffset is the offset to the position where the next next-hop IP address should be inserted. It is initialized to four (4) and incremented by four each time an agent inserts its IP address.

15.2.8 Target and TargetList

Several control messages use a parameter called TargetList (PCode =

13), which contains information about the targets to which the message pertains. For each Target in the TargetList, the information includes the IP address of the target, the SAP applicable to the next higher layer protocol, and the length of the SAP (SAPBytes). Consequently, a Target structure can be of variable length. Each entry has the format shown in Figure 18.

Figure 18: Target

o TargetIPAddress is the IP Address of the Target.

o TargetBytes is the length of the Target structure, beginning with the TargetIPAddress and including any SrcRoute parameters.

o SAPBytes is the length of the SAP, excluding any padding required to maintain 32-bit align- ment.

o SAP may be longer than 2 bytes and it includes a padding when required. There would be no padding required for SAPs with lengths of 2, 6, 10, etc., bytes.

Figure 19: TargetList

15.2.9 UserData

The UserData parameter (PCode = 14) is an optional parameter that may be used by the next higher protocol or an application to convey arbitrary information to its peers. Note that since the size of control messages is limited by the smallest MTU in the path to the targets, the maximum size of this parameter cannot be specified a priori. If the parameter is too large for some network's MTU, a UserDataSize error will occur. The parameter must be padded to a multiple of 32 bits.

Figure 20: UserData

o UserBytes specifies the number of valid UserInformation bytes.

o UserInformation is arbitrary data meaningful to the next higher protocol layer or applica- tion.

16 ST Control Message PDUs

Each control message is described in a following section. See Section 18 for an explanation of the notation.

16.1 ACCEPT

ACCEPT (OpCode = 1) is issued by a target as a positive response to a CONNECT message. It implies that the target is prepared to accept data from the origin along the stream that was established by the CONNECT. ACCEPT is also issued as a positive response to a CHANGE message. It implies that the target accepts the proposed stream modification.

The ACCEPT includes the FlowSpec that contains the cumulative information that was calculated by the intervening ST agents as the CONNECT (or CHANGE) made its way from the origin to the target, as well as any modifications made by the application at the target. The FlowSpec is not modified on this trip from the target back to the origin.

The ACCEPT is relayed by the ST agents from the target to the origin along the path established by the CONNECT (or CHANGE) but in the reverse direction. The ACCEPT must be acknowledged with an ACK at each hop.

Since the cumulative FlowSpec information can be different for different targets, no attempt is made to combine the ACCEPTs from the various targets. The TargetList included in each ACCEPT contains the IP address of a single target, i.e. the one that issued the ACCEPT.

Figure 21: ACCEPT Control Message

Reference contains a number assigned by the agent sending the ACCEPT for use in the acknowledging ACK.

LnkReference is the Reference number from the corresponding CONNECT (or CHANGE).

16.2 ACK

ACK (OpCode = 2) is used to acknowledge a request. The ACK message is not propagated beyond the previous-hop or next-hop agent.

Reference is the Reference number of the control message being acknowledged.

ReasonCode is usually NoError, but other possibilities exist, e.g., DuplicateIgn.

Figure 22: ACK Control Message

16.3 CHANGE

CHANGE (OpCode = 3) is used to change the FlowSpec of an established stream. The CHANGE message is processed similarly to the CONNECT message, except that it travels along the path of an established stream. The CHANGE must be propagated until it reaches all the stream's targets. It must be ACKed at every hop.

G (bit 8) is used to request a global, stream-wide change; the TargetList parameter may be omitted when the G bit is specified.

Figure 23: CHANGE Control Message

16.4 CONNECT

CONNECT (OpCode = 5) requests the setup of a new stream or an addition to or recovery of an existing stream. Only the origin can issue the initial set of CONNECTs to setup a stream, and the first CONNECT to each next-hop is used to convey the SID.

The CONNECT message must fit within the maximum allowable packet size (MTU) for the intervening network. If a CONNECT message is too large, it must be fragmented into multiple CONNECT messages by partitioning the TargetList (see Section 3.2.2). Any UserData parameter will be replicated in each fragment for delivery to all targets.

The next-hop initially responds with an ACK, which implies that the CONNECT was valid and is being processed. The next-hop will later relay back either an ACCEPT or REFUSE from each target.

An intermediate ST agent that receives a CONNECT selects the next-hop ST agents, partitions the TargetList accordingly, reserves network resources in the direction toward the next-hop, updates the FlowSpec accordingly, and sends the resulting CONNECTs.

If the intermediate ST agent that is processing a CONNECT fails to find a route to a target, it responds with a REFUSE with the appropriate reason code, e.g., NoRouteToDest. If the next- hop to a target is by way of the network from which it received the CONNECT, then it sends a NOTIFY with the appropriate reason code, e.g., RouteBack. In either case, the TargetList specifies the affected targets. The intermediate ST agent will only route to and propagate a CONNECT to the targets for which it does not issue a REFUSE.

If a received CONNECT contains a new SID, a new stream should be created. If the SID is known, there are four cases

TargetList is the list of IP addresses of the target processes. It is of arbitrary size up to the maximum allowed for packets travelling across the specific network.

Figure 24: CONNECT Control Message

16.5 DISCONNECT

DISCONNECT (OpCode = 6) is used by an origin to tear down an established stream or part of a stream, or by an intermediate agent that detects a failure between itself and its previous-hop, as distinguished by the ReasonCode. The DISCONNECT message specifies the list of targets that are to be disconnected. An ACK is required in response to a DISCONNECT message. The DISCONNECT message is propagated all the way to the specified targets. The targets are expected to terminate their participation in the stream.

Note that in the case of a failure it may be advantageous to retain state information as the stream should be repaired shortly, see Section 9.2.

G (bit 8) is used to request a DISCONNECT of all the stream's targets; the TargetList parameter may be omitted when the G bit is set (1).

Figure 25: DISCONNECT Control Message

16.6 ERROR

ERROR (OpCode = 7) is sent in acknowledgment to a request in which an error is detected. No action is taken on the erroneous request. No ACK is expected. The ERROR message is not propagated beyond the previous-hop or next-hop agent.

An ERROR is never sent in response to another ERROR. The receiver of an ERROR is encouraged to try again without waiting for a retransmission timeout.

Reference is the Reference number of the erroneous request.

Figure 26: ERROR Control Message

16.7 HELLO

HELLO (OpCode = 8) is used as part of the ST failure detection mechanism, see Section 9.1.

R (bit 8) is used for the Restarted-bit.

Reference is non-zero to inform the receiver that an ACK should be promptly sent so that the sender can update its round-trip time estimates. If the Reference is zero, no ACK should be sent.

TBD: HelloTimer

Figure 27: HELLO Control Message

16.8 JOIN-REQUEST

TBD

Figure 28: JOIN-REQUEST Control Message

16.9 NOTIFY

NOTIFY (OpCode = 10) is issued by an agent to inform other agents, the origin, or targets of events that may be significant. The action taken by the receiver of a NOTIFY depends on the ReasonCode. Possible events are suspected routing problems or resource allocation changes that occur after a stream has been established. These changes occur when network components fail and when competing streams preempt resources previously reserved by a lower precedence stream. We also anticipate that NOTIFY can be used in the future when additional resources become available, as is the case when network components recover or when higher precedence streams are deleted.

NOTIFY may contain a FlowSpec that reflects that revised guarantee that can be promised to the stream. NOTIFY may also identify those targets that are affected by the change. In this way, NOTIFY is similar to ACCEPT.

When NOTIFY is received at the origin, the application should be notified of the target and the change in resources allocated along the path to it, as specified in the FlowSpec contained in the NOTIFY message. The application may then use the information to either adjust or terminate the portion of the stream to each affected target.

The NOTIFY may be propagated beyond the previous-hop or next-hop agent; it must be acknowledged with an ACK.

Reference contains a number assigned by the agent sending the NOTIFY for use in the acknowledging ACK.

ReasonCode identifies the reason for the notification.

LnkReference, when non-zero, is the Reference number from a command that is the subject of the notification.

NextHopIPAddress is an optional parameter and contains the IP address of a suggested next- hop ST agent.

TargetList is present when the notification is related to one or more targets.

Figure 29: NOTIFY Control Message

16.10 REFUSE

REFUSE (OpCode = 11) is issued by a target that either does not wish to accept a CONNECT message or wishes to remove itself from an established stream. It might also be issued by an intermediate agent in response to a CONNECT or CHANGE either to terminate a routing loop, or when a satisfactory next-hop to a target cannot be found. It may also be a separate command when an existing stream has been preempted by a higher precedence stream or an agent detects the failure of a previous-hop, next-hop, or the network between them. In all cases, the TargetList specifies the targets that are affected by the condition. Each REFUSE must be acknowledged by an ACK.

The REFUSE is relayed back by the agents to the origin (or intermediate agent that created the CONNECT or CHANGE) along the path traced by the CONNECT. The agent receiving the REFUSE will process it differently depending on the condition that caused it, as specified in the ReasonCode field. In some cases, such as if a next-hop cannot obtain resources, the agent can release any resources reserved exclusively for transmissions in the stream in question to the target specified in the TargetList, and the previous-hop can attempt to find an alternate route. In some cases, such as a routing failure, the previous-hop cannot determine where the failure occurred, and it propagates the REFUSE back to the origin, which can attempt recovery of the stream by issuing a new CONNECT.

No special effort is made to combine multiple REFUSE messages since it is considered most unlikely that separate REFUSEs will happen to both pass through an agent at the same time and be easily combined, e.g., have identical ReasonCodes and parameters.

Reference contains a number assigned by the agent sending the REFUSE for use in the acknowledging ACK.

LnkReference is either the Reference number from the corresponding CONNECT or CHANGE, if it is the result of such a message, or zero when the REFUSE was originated as a separate command.

Figure 30: REFUSE Control Message

16.11 STATUS

STATUS (OpCode = 12) is used to inquire about the existence of a

particular stream identified by the SID.

Use of STATUS is intended for diagnostic purposes and to assist in stream cleanup operations. It is possible in cases of multiple failures or network partitioning for an ST agent to have information about a stream after the stream has either ceased to exist or has been rerouted around the agent. When an agent concludes that a stream has not been used for a period of time and might no longer be valid, it can probe the stream's previous-hop or next-hops to see if they believe that the stream still exists through the interrogating agent.

When a stream has been identified, a STATUS-RESPONSE is returned that will contain no optional parameters if the specified stream is unknown, or will otherwise contain the current FlowSpec, TargetList, and possibly Groups of the stream.

Q (bit 9) is set to one (1) for remote diagnostic purposes when the receiving agent should return a stream's parameters, whether or not the source of the message is believed to be a previous-hop or next-hop in the specified stream. Note that this use has potential for disclosure of sensitive information.

Figure 31: STATUS Control Message

16.12 STATUS-RESPONSE

STATUS-RESPONSE (OpCode = 13) is the reply to a STATUS message. If the stream specified in the STATUS message is not known, the STATUS-RESPONSE will contain the specified SID but no other parameters. It will otherwise contain the current SID, FlowSpec, TargetList, and possibly Groups of the stream.

Figure 32: STATUS-RESPONSE Control Message

17 Suggested Protocol Constants

The ST Protocol uses several fields that must have specific values for the protocol to work, and also several values that an implementation must select. This section specifies the required values and suggests initial values for others. It is recommended that the latter be implemented as variables so that they may be easily changed when experience indicates better values. Eventually, they should be managed via the normal network management facilities.

ST uses IP Version Number 5.

When encapsulated in IP, ST uses IP Protocol Number 5.

17.1 SCMP Messages

1. ACCEPT
2. ACK
3. CHANGE
4. CONNECT
5. DISCONNECT
6. ERROR
7. HELLO
8. JOIN
9. NOTIFY
10. REFUSE
11. STATUS
12. STATUS-RESPONSE

17.2 SCMP Parameters

1. ErroredPDU
2. FlowSpec
3. Group
4. MulticastAddress
5. NextHopIPAddress
6. Origin
7. RecordRoute
8. TargetList
9. UserData

17.3 ReasonCode

Several errors may occur during protocol processing. All ST error codes are taken from a single number space. The currently defined values and their meaning is presented in the list below. Note that new error codes may be defined from time to time. All implementations are expected to handle new codes in a graceful manner. If an unknown ReasonCode is encountered, it should be assumed to be fatal. The ReasonCode is an 8-bit field. Following values are defined:

To Be Written

17.4 IP Multicast Addresses

The following permanent IP multicast addresses have been assigned to ST:

224.0.0.7 All ST routers

224.0.0.8 All ST hosts

In addition, a block of transient IP multicast addresses, 224.1.0.0 - 224.1.255.255, has been allocated for ST multicast groups. Note that in the case of Ethernet, an ST Multicast address of 224.1.cc.dd maps to an Ethernet Multicast address of 01:00:5E:01:cc:dd, see [RFC1112].

18 Notation

To Be Written

19 Further Study

To Be Written

20 References

[RFC1071]

Braden, Borman, Partridge: Computing the Internet Checksum, RFC 1071, USC/Information Sciences Institute, Cray Research, BBN Laboratories, Sep- tember 1988.

[RFC1112]

Deering, S.: Host Extensions for IP multicasting, RFC 1112, Stanford Univer- sity, August 1989.

[RFC1122]

Braden, R.: Requirements for Internet Hosts -- Communication Layers,

RFC 1122, USC/Information Sciences Institute, October 1989.

[Jaco88]

Jacobson, V.: Congestion Avoidance and Control, ACM SIGCOMM-88, August 1988.

[KaPa87]

Karn, P. and C. Partridge: Round Trip Time Estimation, ACM SIGCOMM-87, August 1987.

[RFC1141]

Mallory, T. and A. Kullberg: Incremental Updating of the Internet Checksum, RFC 1141, BBN, January 1990.

[RFC 1363]

C. Partridge: A Proposal Flow Specification, RFC 1363.

[RFC791]

Postel: Internet Protocol, RFC 791, DARPA, September 1981.

[RFC1060]

Reynolds, Postel: Assigned Numbers, RFC 1060, USC/ISI, March 1990.

[RFC1190]

Topolcic C.: Internet Stream Protocol Version 2 (ST2), October 1990.