WHITE PAPER

H.323 Firewall/NAT Traversal

H.460.17, H.460.18, H.460.19

RADVISION
Delivering the Visual Experience

NOTICE

For further information contact RADVISION or your local distributor or reseller.

USA/Americas
17-17 State Highway 208,
Suite #300
Fair Lawn, NJ,
07410-2819 USA
Tel (201) 689-6300
Fax (201) 689-6301
info@radvision.com

EMEA
24 Raul Wallenberg St.
Tel Aviv, Israel 69719
Tel +972.3.767.9300
Fax +972.3.767.9313
infointernational@radvision.com

Asia/Pacific
Suite 2901 29/F China
Resources Building,
26 Harbour Road,
Wanchai, Hong Kong
Tel +852.2801.4070
Fax +852.2801.4071
apacinfo@radvision.com

About This Document

This document first presents the problems of firewall/NAT traversal and then describes the issues that are resolved by using the three new firewall/NAT traversal standards: H.460.17, H.460.18 and H.460.19.

TABLE of CONTENTS

Introduction

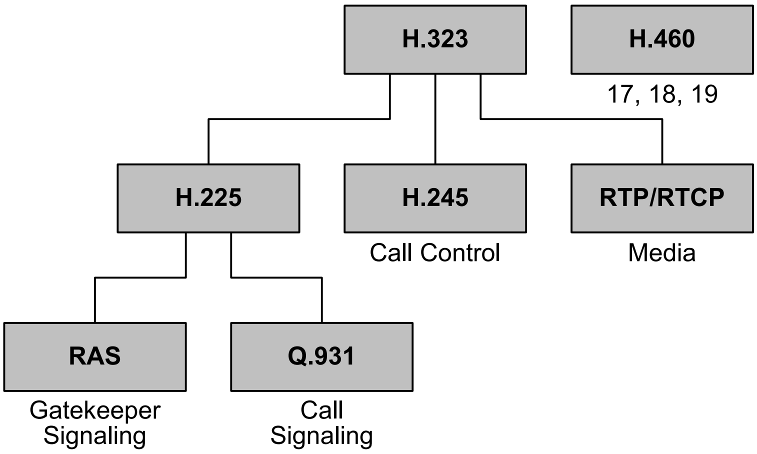
H.323's generic extensibility framework provides a simple method for adding features to the basic standard. Many of these new features are added through the H.460 set of standards, as is the case with three new firewall and NAT (Network Address Translator) traversal documents that were recently added to the standard: H.460.17, H.460.18 and H.460.19.

This paper describes how each of the standards operate, the pros and cons of using each of these solutions and how to best deploy them in order to resolve specific aspects of firewall/NAT traversal. This overview provides you with the flexibility to choose the solution most suited for the scenarios that you deploy.

A typical call scenario is described in the appendix at the end of this paper.

Connection Addresses

The following shows an overview of the protocols applied in the call setup process. Each protocol is sent via a different network connection.



In the most basic scenario:

- RAS is sent over UDP.
- Q.931 and H.245 are each sent on different TCP connections.
- RTP and RTCP are each sent on different UDP connections.

The addresses of each of these connections are not known beforehand and are usually sent inside the various message layers. For example:

- Q.931 addresses are stated inside RAS (Registration, Admission and Status) messages.
- H.245 addresses are stated inside Q.931 messages.
- RTP (Realtime Transport Protocol) and RTCP (Realtime Transport Control Protocol) addresses are stated inside H.245 messages.

Note: Sometimes fewer connections can be used, for example, when H.245 tunneling is employed. In this case, the Q.931 connection can tunnel H.245 messages without the need of a dedicated H.245 connection over TCP (Transmission Control Protocol). Another example is when Q.931 multiplexing is employed, where several calls being made to the same immediate destination share the same Q.931 connection.

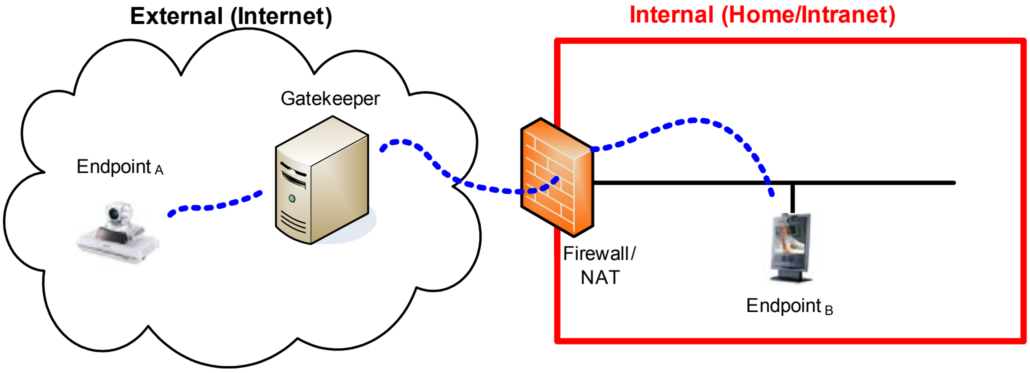
The Problems of Firewall/NAT Traversal

The purpose of a firewall is to ensure that only authorized users are allowed access. Today, most homes possess a firewall-protected connection through cable or ADSL.

Enterprise or home users who are protected behind their own firewall are communicating from what we call an *internal* network for the purpose of this white paper. The user's endpoint communicates with a gatekeeper via the Internet through a firewall or a NAT (usually both) to another endpoint. The gatekeeper and the other endpoint are on what we call an *external* network. When the other endpoint is on a different internal network, it should also use a similar solution to traverse through its own firewall/NAT.

The following presents the main problems of an endpoint on an external network trying to setup a call with an endpoint that is on an internal network, or behind a firewall. In diagram 1, the enterprise or home endpoint on the internal network is named **Endpoint B** and the endpoint on the external network is named **Endpoint A**.

Diagram1:



Problem 1: Lack of symmetry. Endpoints that communicate from behind a firewall can open TCP connections and send TCP or UDP messages to an endpoint that is located on an external network. However, endpoints on an external network cannot send TCP or UDP messages to an internal endpoint without having an open connection through the firewall already in place, and so, cannot call an endpoint on an internal network. For example, in the previous drawing, Endpoint A cannot dial into Endpoint B.

This blockage of the external endpoint's attempts is because of the fact that the inherent purpose of a firewall is to block incoming connections that are not recognized.

For example: A Q.931 SETUP message is sent by an external endpoint to the gatekeeper to open a new connection and from there to the endpoint on the internal network. Firewalls will usually block this type of message because the protocol of the message (for example: H.323, H.245, RTP or RTCP) is not recognized. Since the ports used for H.323 communications are dynamically allocated and used, an H.323-aware firewall is hard to implement.

The preferred method for an endpoint on an external network to call an endpoint on an internal network is when the endpoint on the internal network opens a special connection, called a *pinhole,* though the firewall. We will refer only to dynamic pinholes in this paper, where the pinhole is created by an internal endpoint to an external endpoint through the firewall and maintained as long as the internal endpoint keeps the connection open. This pinhole will allow traffic to flow in both directions. However, there are various issues that arise when doing this which are discussed in the following sections.

Problem 2: NAT – Internal Network Address Exposure

A NAT replaces internal endpoint addresses with public endpoint addresses on the Internet for two reasons:

- Security: to prevent outside entities from acquiring the addresses of the internal endpoint network and thus detecting the internal network structure.
- Reduction of the required number of Public IPv4 addresses.

Each one of the protocols described in the previous sections send internal addresses inside their messages. Thus, either the endpoint in the internal network must know the public addresses for its internal addresses (which is not always possible), or the NAT must be able to replace the address inside the H.323 messages payload.

The problem with this solution is the complexity of the H.323 protocol and the difficulty of a firewall to recognize the internals of the H.323 protocol. A firewall that modifies H.323 messages in order to route them properly also causes problems for authentication and security protocols that are used by H.323. The implementation of any suitable solution would require detailed knowledge of the H.323 protocol.

Comparing H.460.17 / H.460.18 / H.460.19

During the past few years, the International Telecommunications Union (ITU), which is responsible for the standardization of the H.323 protocol, determined that the problem of firewall/NAT traversal is of the utmost urgency. However, the ITU decided to wait for the Internet Engineering Task Force (IETF) to create a generic standard which would serve as a long-term solution.

Now, two years later, it has become obvious that a long-term solution is not imminent, and that a specific firewall/NAT traversal solution for H.323 should be defined without further delay.

Several different proposals were submitted to the ITU, three of which have recently been standardized (during the ITU meeting in the summer of 2005).

These solutions were created by several leading companies in the VoIP industry, one of them being RADVISION, and will shortly be adopted by many other companies.

The three standards (H.460.17, H.460.18 and H.460.19) provide two categories of solutions:

- H.460.17 and H.460.18 standards dealing with signaling.
- H.460.19 standard dealing with media.

These standards start with some initial assumptions regarding the firewall. Any TCP connection or UDP packet sent from the internal network through the firewall opens a pinhole dynamically in the firewall. This pinhole allows incoming messages to be sent from the destination of the TCP connection or the UDP packet. The pinhole stays open as long as the internal network sends information through the pinhole to the same destination.

These standards only deal with a simplified scenario, where an internal network wishes to dial an endpoint on the public network or vice versa. Solutions that expand the scope of this simple scenario for real life situations where both endpoints are located behind different firewalls in two separate internal networks can be achieved using border elements, but are out of the scope of this whitepaper.

Note: The following simplified descriptions suffice to demonstrate the behavior of these solutions, even though more complex scenarios are also possible.

H.460.17

The H.460.17 standard was initially called RAS over H.225. This name indicated the fact that all the RAS messages are tunneled on top of Q.931 FACILITY messages. Using this standard, the calls are also multiplexed over Q.931 and H.245 messages are tunneled over Q.931 messages.

In the typical case scenario of H.460.17, a single TCP connection is required between the endpoint and a gatekeeper in order to handle all H.323 signaling. The endpoint must register with the gatekeeper in order to join the network after initialization. This means that the endpoint on the internal network (Endpoint B) opens a TCP connection to the gatekeeper through the firewall/NAT and that this connection is used to send Q.931 messages with H.245 and RAS messages tunneled on top of it in both directions. The initial message transmitted will be the RAS RRQ message, tunneled over a Q.931 FACILITY message.

The addresses of the RAS, Q.931 and H.245 protocols do not need to pass through a NAT and, therefore, do not need to be translated.

This solution resolves all call setup signaling issues by simply opening a TCP connection from the internal network to the external one, tunneling

RAS and using the other building blocks of the standard (Q.931 multiplexing and tunneling of H.245 messages) to create a single connection that is always open for both directions.

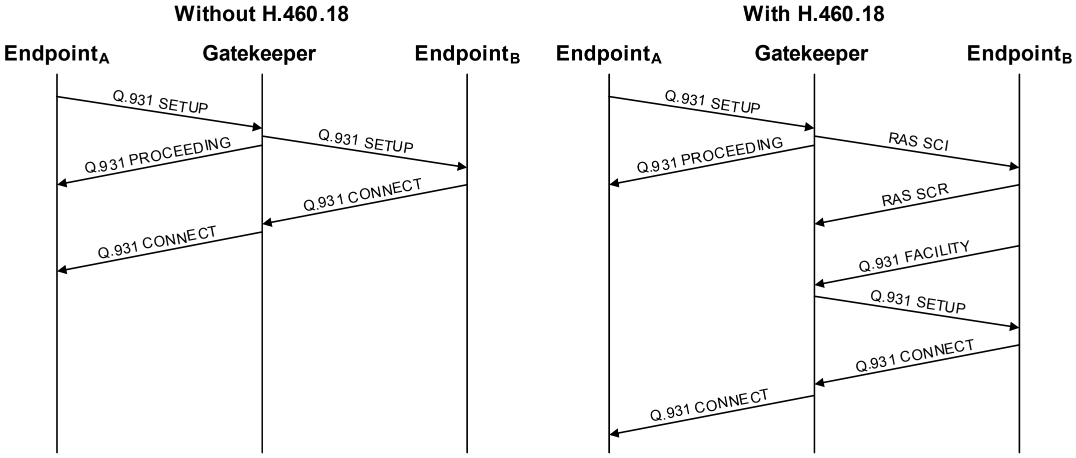
In this case, if the endpoint on the external network (Endpoint A) attempts to dial a registered endpoint on the internal network (Endpoint B); the Q.931 SETUP message will use the same Q.931 connection that was initially used for registration.

H.460.18

The H.460.18 solution takes a different approach. It does not employ tunneling, but instead tries to emulate basic H.323 messages by always attempting to open pinholes from the internal network to the external one.

In other words, to use an analogy: if the external endpoint cannot let itself into the internal network, it will ask the internal endpoint to open the door for it (a pinhole). An SCI message is the knock on the door and an SCR response indicates that the recipient is coming to open the door. A FACILITY message opens the door and invites the external endpoint to enter with its SETUP message so that the call can begin.

Initially, one pinhole is opened for the RAS messages. RAS messages are transmitted in the usual manner, but they contain extra signaling that indicates the usage of the H.460.18 standard, and the pinhole is maintained by "keep-alive" messages. Afterwards, when an incoming call arrives, meaning a Q.931 SETUP message intended for the internal network, a different procedure is used between the gatekeeper and the internally registered endpoint, utilizing the existing UDP pinhole as shown below:



The diagram above shows the messages sent between the gatekeeper and the internal Endpoint B both **with** and **without** the usage of the H.460.18 protocol.

- An incoming call arrives at the gatekeeper. Typically, without the usage of H.460.18 and without a firewall, the gatekeeper sends the Q.931 SETUP message to the internal endpoint and that endpoint sends back a Q.931 CONNECT message if a connection was made successfully
- This cannot be done from an external endpoint to an internal endpoint (behind a firewall), because a new TCP connection must be opened and the firewall will block such a connection. Therefore, a RAS Service Control Indication (SCI) message will be sent instead on the opened RAS pinhole between the gatekeeper and the internal Endpoint B. This message indicates to the internal endpoint that an incoming call is being attempted, which means a Q.931 connection is required.
- The gatekeeper will then receive a Service Control Response (SCR) message back from the internal Endpoint B.
- The internal Endpoint B then sends a special Q.931 FACILITY message, which opens a pinhole through the firewall.
- By recognizing the content of the message, the gatekeeper can link this newly opened connection to the incoming call and can send back the Q.931 SETUP message.
- The internal Endpoint B can then return the Q.931 CONNECT message.

H.460.18 Summary: Without using the H.460.18 solution, which enables the gatekeeper to open a connection, the external Endpoint A could not communicate with internal Endpoint B, because the firewall would block its attempt to setup a call. Therefore, the gatekeeper, on the behalf of the external endpoint, must signal the internal Endpoint B to open the connection instead. For this reason, the extra signaling was added (the call is always routed through the gatekeeper).

Another mechanism is used to open multiple H.245 connections to the same server address: the server passes the same H.245 address on several calls, and incoming connections are distinguished by an H.245 generic message sent from the endpoint on the new H.245 connection as the first message.

H.460.19

H.460.19 adds two required features to RTP and RTCP by using a slightly modified version of these protocols.

Using keep-alive streams empty UDP packets (keep-alive messages) are sent back from the address which receives media inside the firewall/NAT to the address from which media is sent. In this way the internal endpoint effectively opens a pinhole through the firewall. This pinhole is used for incoming media.

Multiplexing enables several calls to be handled by the same connection, so that the same pair of UDP ports can be reused for several different RTP or RTCP sessions. This ensures the reduction of the resources required when developing a server that must handle several calls. This also reduces the number of pinholes that must be open on the firewall because the same number of pinholes can be used for a larger number of calls.

Comparative Summary

The H.460.17, H.460.18 and H.460.19 provide two solutions for signaling and one for media. Both of the signaling solutions can be used with the media solution, meaning that either H.460.17 or H.460.18 can be used together with H.460.19.

Using H.460.17 reduces the number of connections from the endpoint to a gatekeeper. Using the H.460.18 solution requires that the external Endpoint A or the gatekeeper will signal to open a linking connection so that the internal Endpoint B will open a connection on its behalf in the reverse direction.

H.460.17 and H.460.18 deal with the signaling aspect of call setup, but do not present any long-range solutions since they do not solve all the issues. H.460.19 provides a solution for opening RTP and RTCP pinholes and a method for maintaining them using a keep-alive mechanism.

The major deterrent to the deployment of these solutions is the prevalence of legacy H.323 equipment, which has a long deployment history. Upgrading these video conferencing terminals will ensure massive expense in addition to the fact that some legacy equipment cannot be upgraded. For these enterprises, the only viable solution is to deploy proxies (which function like gateways) on the internal network side and traversal servers on the external network side to work on behalf of the internal endpoints and the external gatekeepers.

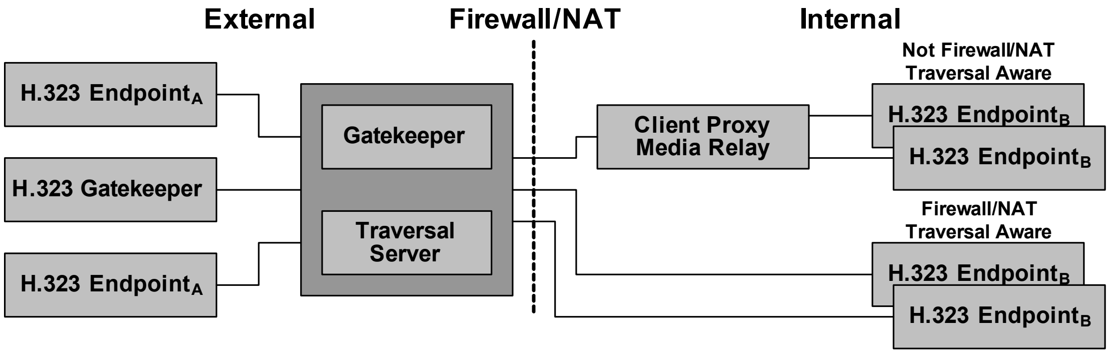
The best possible solution is for the internal endpoint and the gatekeeper to be firewall/NAT aware, meaning to have the ability to signal these new additions to the standard. Whenever possible, this should be the preferred solution over the use of proxies and traversal servers.

Deploying

There are two issues to consider when deploying these solutions.

- Support for legacy endpoints that are not firewall/NAT traversal aware.
- Required capacity and security needs.

The first issue is easily solved if the endpoints can be upgraded to be firewall/NAT traversal aware. Otherwise, a proxy and/or traversal server must be created on the side of the endpoint and/or the gatekeeper that can communicate with the legacy endpoints that are not firewall/NAT traversal aware, as shown below:



Another deployment option is to upgrade some of the endpoints to be firewall/NAT traversal aware and to support other legacy endpoints for which an upgrade is not financially feasible or impossible through a proxy that is made to be firewall/NAT traversal aware.

The client proxy described above has two roles:

- For the legacy endpoint in the external network, the proxy relays messages from the gatekeeper.
- For the gatekeeper (or traversal server), the client gateway serves as a firewall/NAT traversal aware endpoint on behalf of the legacy endpoint.
- The gatekeeper on the external network must also be firewall/NAT aware. This can be achieved in either of two ways:
    - By modifying the gatekeeper to be firewall/NAT aware.
    - By adding a traversal server that will act as a firewall/NAT aware gatekeeper (actually relays messages from the real gatekeeper) for the client proxy or the firewall/NAT aware H.323 endpoint, and acts as a client to the gatekeeper.

The client proxy and traversal server, discussed above, handle both the signaling and the media. These two components can be separated into two other building blocks, one for signaling and one for media relay.

It is clear that it is extremely beneficial to use H.460.19 for handling the media relay. The following discusses which call signaling standard, H.460.17 or H.460.18, is best for handling signaling.

H.460.17

H.460.17 uses RAS over TCP to reduce the number of network connections while H.460.18 opens pinholes for the next signaling protocol.

When comparing the advantages of using H.460.17 versus the advantages of using H.460.18, a prominent benefit of using H.460.17 is that it only requires a single pinhole through the firewall for all the calls made from a specific client to the gatekeeper. The same pinhole is used for RAS, Q.931 and H.245 messages. This reduces the networking resource requirements and the number of pinholes in the firewall. Reducing the number of pinholes is considered to be a major security benefit.

The only issue is the fact that an endpoint must know, before registering, whether the gatekeeper supports this standard. This can be done by a Domain Name Server (DNS) lookup of the gatekeeper.

If the DNS indicates that the gatekeeper is a TCP service, it can then be assumed that the gatekeeper supports RAS over TCP.

Due to the fact that currently not all endpoints are able to lookup a gatekeeper using DNS, this solution may not be applicable for all scenarios.

H.460.18

H.460.18 is a more complex solution that requires more resources and more signaling, but performs in a similar manner to the initial H.323 standard. This means that none of the advanced features of H.323 are needed (meaning that no H.245 tunneling, no Q.931 multiplexing or such are needed).

This can be applied to situations when no DNS lookup is possible. You should note, however, that this solution requires many more pinholes through the firewall/NAT, which may be considered a security risk. It also requires more round trips for the initial call setup, increasing the overhead on the network, the H.323 entities and the time required for the setup negotiation.

The standard itself also enables the reduction of the number of pinholes by using an H.245 server.

Conclusion

This document has described the features and benefits of using each of the three solutions. We conclude that it is highly beneficial to support all three standards.

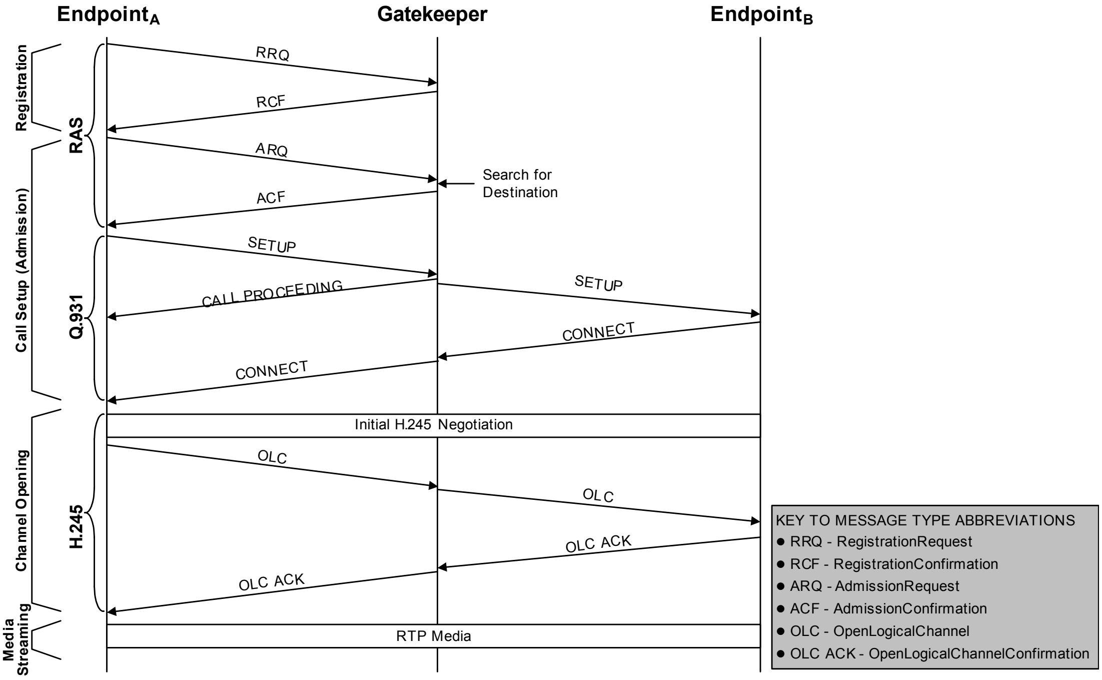
This comprehensive support provides the flexibility to interact with a variety of gatekeepers. During the process of registration, endpoints should first try to detect an H.460.17 supporting gatekeeper using DNS, and failing that, try to register while indicating support for H.460.18 and then determine the gatekeeper's support options by its response.

If the gatekeeper supports both H.460.17 and H.460.18, then H.460.17 should be preferred because, generally, all solutions strive to use as few pinholes as possible through the firewall. H.460.17 reduces the resources required by the client, the gatekeeper and the firewall itself while at the same time increasing security.

The same solutions can be used even when the communicating endpoints each reside inside a different internal network and behind a different firewall. In this case, the routing should be handled by the gatekeepers, preferably using URL addresses for the endpoints and DNS lookups for the gatekeepers.

Appendix: A Typical Call Scenario

The following shows a typical call scenario and points out the firewall/NAT traversal problems that it presents.



The umbrella standard H.323, which is used for VoIP and any other rich multimedia application, uses RAS (Registration, Admission and Status) for gatekeeper signaling and Q.931 for call signaling. Call control is handled by H.245. Realtime Transport Protocol (RTP) and Realtime Transport Control Protocol (RTCP) are used to send media. Each of the above standards uses either Transmission Control Protocol (TCP) or User Datagram Protocol (UDP) connections.

Typically, call setup goes through the following stages:

- **Endpoint Registration:** RAS messages handle endpoint registration. It includes the Registration Request (RRQ) and Registration Confirm (RCF) messages. This procedure is not an actual part of a call setup, but it is stated here due to its relevance to firewall/NAT issues.

- **Call Setup Admission:** The RAS protocol also handles the Call Setup Admission, which includes the Admission Request (ARQ) and the receipt of the Admission Confirm (ACF) by an endpoint from the gatekeeper.

- **Call Setup Signaling:** The call setup signaling is performed using SETUP, CALL PROCEEDING and the CONNECT messages, which are Q.931 messages. Other Q.931 messages may also be used during call setup.
- **Control Negotiation:** Before media channels can be opened, the control channel itself needs to be connected and capabilities have to be negotiated.
- **Channel Opening:** Media channel opening and acknowledgement is performed using Open Logical Channel (OLC) and OLC ACK messages, which are handled by the H.245 protocol. A call might have more than a single media channel opened in each direction.
- **Media Transmission:** Media transmission is handled over two UDP connections: one RTP and one RTC.

The following shows a typical call scenario between Endpoint A and Endpoint B, which are two terminals in a network that communicate via a gatekeeper.

Registration

- Initially, each endpoint registers with the gatekeeper indicating that it is online and ready to send or receive calls. This is performed by each endpoint sending a RRQ message. This message includes the listening Q.931 address of each endpoint that will later be used when a Q.931 SETUP message needs to be received.
- The gatekeeper generally answers with a RCF message. The gatekeeper stores all these registrations, which it can then use to connect registered endpoints to each other. The gatekeeper can also access an endpoint that is not registered by searching for it in the network's Domain Name Server (DNS) directory services or by locating the registration of this endpoint with a different gatekeeper.

Call Setup (Admission)

- Call setup is started by Endpoint A (for instance), which sends an ARQ to the gatekeeper. This message indicates that Endpoint A would like to communicate with Endpoint B by specifying Endpoint B as the destination of the call.
- The gatekeeper searches for this destination (Endpoint B) and once it is found, will return an ACF message.

- In this scenario we assume that the gatekeeper is using a routed mode, meaning that this gatekeeper will receive and route all signaling of this call between both endpoints. Endpoint A now sends a Q.931 SETUP message to the gatekeeper.
- The gatekeeper then returns a Q.931 CALL PROCEEDING message indicating that it is now going to attempt to connect the call to the requested destination (Endpoint B).
- The gatekeeper then sends a Q.931 SETUP message to Endpoint B.
- If setup is successful, a Q.931 CONNECT message is sent from Endpoint B to the gatekeeper and then the gatekeeper routes the Q.931 CONNECT message to Endpoint A.

  One of the Q.931 messages may contain an H.245 address, which will later be used to open an H.245 connection.

Opening Call Channels

- The media channels can now be opened by OLC messages using an H.245 connection opened by using the H.245 address. In this example scenario one of the endpoints will open a single channel.
- Each endpoint acknowledges the channel opening with an OLC ACK message.
  Note: Many other H.245 messages that are not shown here are transmitted.

Media Transmission

- Media can now stream through RTP between these endpoints.