Internet Engineering Task Force
Internet Draft
draft-mankin-im-session-guide-00.txt
November, 2001
Expires: May, 2002

Allison Mankin
USC/ISI
Jon Peterson
NeuStar

Guidelines for Instant Message Sessions

Status of this Memo

This document is an individual submission to the IETF. Comments should be directed to the authors.

Abstract

This document recommends a set of guidelines for session-based instant messaging, focusing particularly on security properties, the selection of transport protocols and the effects of network intermediaries.

1. Introduction

As the standardization of instant messaging systems in the IETF has progressed, a model has developed that decouples the signaling used to establish an instant messaging session from session data itself. A set of guidelines with regard to protocol and architecture selection for decomposed sessions of instant messages are proposed in this document.

After introducing the distinction between the 'session model' of instant message transmission and the traditional 'paging model', the authors propose a set of rules for selecting underlying transport protocols for instant messaging sessions and describe some session-layer characteristics that are required for proper management and security of instant messages. These principles are complicated by the introduction of network intermediaries that operate on instant messaging sessions; the repercussions of intermediaries for the transport and session layer mechanisms are explored, and measures to diminish the impact of intermediaries on instant message sessions are recommended. Finally, a set of normative guidelines derived from the preceding text are enumerated.

The guidelines presented in this document have resulted from discussions in the SIMPLE WG of the IETF pursuant to the use of the Session Initiation Protocol (SIP, [2543]) for instant messaging applications. The guidelines below have been somewhat generalized to apply more broadly to any instant messaging system that admits of a signaling/session distinction. SIP is however used in examples throughout the document to illustrate typical protocol behavior.

2. Session Model and Paging Model

2.1 Session Model

Some solutions for exchanging instant messages propose a two-layer approach in which preliminary signaling is used to characterize an instant messaging session that will be established separately. Usually the traffic of the instant messaging session, when it has been initiated, will follow a different path through the network than the signaling that preceded it. Architectures using this signaling/session distinction will hereafter be referred to as examples of the 'session model.' An example of the session model is given in [SESSION].

```
                   1)Request   +-------+
                    +--------->|  IM   |-----------+
                    | Session  |Session|           |
                    |          |Router | 2) Accept |
         SIGNALING  | +--------|       |<--------+ |
                    | |        +-------+  Session| |
                    | |                          | |
                    | V                          | V
            +---------+         SESSION          +---------+
            |IM Client|<------------------------>|IM Client|
            +---------+     3) Exchange IMs      +---------+
```

<Fig 1 Session model>

The purpose of the initial signaling in the session model is twofold:

First, to locate the party with whom the originator would like to have a session. This may include transmitting messages through proxy servers or other signaling intermediaries before the message arrives at the host with whom an instant messaging session will be shared.

Second, to negotiate capabilities associated with the exchange of instant messages. These would include acceptable MIME types that might appear in messages, as well as necessary information about how and where instant messages should be sent (IP addresses and ports, transports, and so forth).

Once preliminary signaling has completed, the instant messaging session begins in accordance with the characteristics described in the preliminary signaling. Usually, this will entail establishing a new network connection specifically for instant messages separate from the network connection that was used for signaling. Ideally, this instant messaging connection will go directly end-to-end between the participants in a session.

This signaling/session distinction is common in Internet telephony systems (such as SIP), Internet gaming and many other real-time communications applications, although commonly in these applications the session is described as 'media' and is transmitted over RTP ([1889]).

2.2 Paging Model

In contrast to the session model is the 'paging model' for instant messaging, in which the preliminary signaling and the transmission of actual instant messages are conflated. Rather than sending any preliminary signaling, endpoints send instant messages without preamble; a set of headers containing routing and capability information is prepended to each individual instant message. An example of the paging model can be found in [MESSAGE].

```
                   1) Transmit +-------+
                    +--------->|  IM   |-----------+
                    |   IM     |       |           |
                    |          |Router | 2) ACK    |
         SIGNALING  | +--------|       |<--------+ |
            IMs     | |        +-------+     IM  | |
                    | |                          | |
                    | V                          | V
            +---------+                          +---------+
            |IM Client|                          |IM Client|
            +---------+                          +---------+
```

<Fig 2 Paging model>

2.3 Comparison of paging and session

The session model arises from concerns that the paging model is not sufficiently scalable. When large numbers of users are sending many messages simultaneously through the same signaling infrastructure, the signaling infrastructure becomes strained. In the paging model, each instant message contains its own routing and capability data; the signaling infrastructure must therefore essentially make a new forwarding decision each time an instant message is sent.

Additionally, from a sheer network capacity perspective instant messages sent using the paging model are larger than instant messages sent using the session model. The size of any headers containing routing and capability information may significantly exceed the size of an instant message. In the session model, only session management information adds to the length of the instant messaging content.

Finally, the separation of signaling and session greatly facilitates the implementation of complex services like advanced call-control (transfer and redirection) and conferencing. Note that the examples below assume simple two-participant IM sessions.

3. Transport for Instant Messaging Sessions

3.1 Transport

In the session model, when an instant messaging session is requested, the preliminary signaling proposes the characteristics of the connection over which instant messages will be sent. These characteristics include the underlying transport protocol that will be used to carry instant messages.

Any IM system that supports the session model needs a means for

specifying in the preliminary signaling what transport protocol should be used in the instant messaging session.

Only protocols supporting congestion control are suitable for carrying sessions of instant messages. Therefore, protocols such as UDP cannot be specified for this purpose. Transport protocols such as TCP and SCTP, which have well-understood congestion control properties, should be used instead.

For example, when SIP is used to set up an instant messaging session, an INVITE is sent containing a Session Description Protocol (SDP, [2327]) body that characterizes the desired session; the SDP extensions for instant messaging ([IM-SDP]) allow for the specification of a transport protocol.

3.2 Session

Merely sending the body of an instant message (perhaps wrapped in MIME headers) over the selected transport layer is, however, not sufficient to create an instant messaging session. Some amount of information is needed at the session layer in order to properly manage a session once it has been established.

For example, there needs to be enough information available in each instant message that both participants can determine which session the instant message belongs to (and with that, the identity of the other participant(s) in the session), where this particular instant message fits into the sequence of messages transmitted in the course of this session, and so forth.

In simple end-to-end cases much of this information could be inferred from transport or network-layer qualities (from what IP address did I receive this message, what port did it arrive on, etc). But the explicit presence of this information in session messaging becomes important when multiple instant messaging sessions are established between a pair of endpoints. This can occur when the endpoints in question are aggregating instant messaging sessions on behalf of a number of participants (possibly for scalability reasons). Each aggregating endpoint must know, when it receives an instant message from its peer aggregator, which IM client is associated with that particular session.

```
      sender
    +---------+ individual                                +---------+
    |IM Client|----------+                      +---------|IM Client|
    +---------+ session  |                      |         +---------+
                       +-------+          +-------+
                       |  IM   |          |  IM   |
    +---------+        |       |          |       |       +---------+
    |IM Client|--------| Muxer |----------| Muxer |-------|IM Client|
    +---------+        |       | aggregate|       |       +---------+
                       +-------+  circuit +-------+
                         |                      |          receiver
    +---------+          |                      |         +---------+
    |IM Client|----------+                      +---------|IM Client|
    +---------+                                 individual+---------+
```

<Fig 3 Aggregate Circuits>

It may also be desirable for the session layer to manage flow control between competing instant messaging session within an aggregated 'application circuit' in order to ensure that each session receives an equitable share of network and processing resources.

Finally, the session layer protocol should have some means of ensuring end-to-end instant message integrity and confidentiality, as well as mutual authentication for session participants. While few would dispute that these are important qualities for an instant messaging system, it is important to note that they apply both to the signaling and the session components of an IM system when the two are decomposed. Even if mutual authentication was performed in the application layer signaling, it is still important that authenticate the remote side of the instant messaging session as well. Obviously confidentiality and integrity as important, if not more so, for the instant messages themselves as they are for session establishment signaling.

3.3 Case study of SIP for instant messaging sessions

Some have argued within the SIMPLE working group that SIP should be used both to signal the request for a session and then within the session to carry IMs; i.e. that SIP itself should be used as a session-layer protocol to carry instant messages during an instant messaging session.

This also raises concerns about the applicability of SIP to the problem of session layer management. If at all possible, the session layer should not carry any superfluous information. While clearly SIP headers provide some of the information described above, they also contain a great deal of routing data (in Via headers, Record-Route and Route, for example) that don't immediately seem necessary in an instant messaging

system.

Some known problems arise from using SIP as a session layer protocol when session intermediaries are introduced; these problems are detailed further below.

4. Session Intermediaries

Ideally, when the session model is used, after the preliminary signaling has been completed session traffic can travel end-to-end between the participants in the session without any further interaction with intermediary network elements. However, in some instances service providers may wish to introduce session intermediaries through which instant messaging session traffic is transmitted. The presence of intermediaries can, however, greatly impact transport and session layer activity in an instant messaging system.

4.1 Why Intermediaries?

```
                               +-------+
                    +--------->|  IM   |<----------+
                    |          |Session|           |
                    |          |Router |           |
         SIGNALING  |          |       |           |
                    |          +-------+           |
                    |                              |
                    V           SESSION            V
            +---------+      +------------+      +---------+
            |IM Client|<---->|Intermediary|<---->|IM Client|
            +---------+      +------------+      +---------+
```

<Fig 4 A Simple Intermediary>

The most common reason for introducing a session intermediary is network address translation (NAT, [NAT]). As is detailed in [NAT-G], protocols that have separate signaling and session layers have some significant problems traversing NATs. For the most part these problems result from the citation within signaling of IP addresses and ports that are intended for subsequent use in establishing the session - if a signaling message containing these citations crosses a NAT boundary, the addresses to which the message refers may no longer be meaningful (or routable) to a recipient.

Application Layer Gateways (ALGs) that analyze and modify signaling in order to facilitate the traversal of specific applications are in widespread use today. Some work has been done towards a more

sophisticated solution to this problem within the MIDCOM working group. In the MIDCOM model (see [x]), an element positioned as a session router can re-write certain aspects of the signaling and control, through an external protocol, an intermediary (or 'middlebox') like a NAT in order to allow a session to traverse that intermediary seamlessly. In many MIDCOM architectures, it is desirable for the addition of a middlebox to a network to be transparent to applications that traverse it - in other words, an application has no way of knowing, based on its conventional inspection of signaling and session traffic, that a middlebox is in its session path. ALGs, MIDCOM and pre-MIDCOM architectures are becoming increasingly common elements in service provider networks.

```
                 NAT                 NAT
     +---------+  ||  +------------+  ||  +---------+
     |IM Client|<-||->|Intermediary|<-||->|IM Client|
     +---------+  ||  +------------+  ||  +---------+
```

<Fig 5 Drat! A NAT!>

But even aside from the necessity of NAT traversal there are a number of reasons why a service provider might introduce session intermediaries. The service provider might wish to enforce certain policies at a session layer (regarding the size of messages, their payload type, perhaps even their content). In some regions lawful intercept of instant messages sent by certain participants might be required. Service providers might want to monitor instant messages statistically for network management or capacity planning. Aggregating many individual sessions into 'application circuits' containing instant messages from multiple sessions (as shown above) also requires intermediaries.

4.2 Effects of intermediaries on security

The first and most obvious concern with session intermediaries is their potential interference with the secure end-to-end transmission of instant messages. Regardless of whether security is assured in the network, transport or application layer, session establishment is jeopardized if intermediaries need to access the encrypted portions messages in order to fulfill their purpose. Authentication mechanisms may similarly fail if an IM client unknowingly challenges an intermediary in place of a participant.

Intermediaries must explicitly be made a part of any desired security associations if session establishment is to be successful.

An intermediary also introduces a new point in the network that attackers might attempt to compromise. The security of the end-to-end

session is therefore predicated on the security of these intermediaries.

Note that in some architectures it might be desirable to introduce intermediaries specifically to terminate security associations (like TLS proxies/aggregators) for scalability reasons. Not all intermediaries have negative effects on security - but if they are deployed in ignorance of security requirements then they may lead to widespread system failures.

4.3 Effects of intermediaries on transport

The introduction of intermediaries also potentially allows the characteristics of sessions to be altered in mid-network without the knowledge or consent of the endpoints.

```
  +---------+  TCP +------------+  UDP +------------+  TCP +---------+
  |IM Client|<---->|Intermediary|<---->|Intermediary|<---->|IM Client|
  +---------+      +------------+      +------------+      +---------+
```

<Fig 6 UDP in the middle>

SIP, for example, only permits transport protocols to be set hop-by-hop rather than end-to-end. Were SIP to be used as a session-layer protocol for an instant messaging session in a network with session intermediaries, this could lead to certain hops in a session reverting to undesirable protocols (e.g. UDP). However, if transport is set globally for a session, there is no risk of this.

Even if the transport selected by the endpoints supports congestion control and remains unchanged by intermediaries, network flows that are controlling congestion only over short sequential hops can inhibit competing longer path flows and can use more than a fair share of path resources. A large service provider fielding many intermediaries might thereby inadvertantly (or intentionally) shut out traffic traversing its network that it doesn't intermediate (for further information see [CONG]).

Finally, note that setting up multiple 'application circuits' between two hosts is undesirable regardless of their congestion control properties. This is especially important in architectures in which intermediaries aggregate requests for a number of clients. A pair of intermediaries each responsible for a number of users initiating sessions with one another must not establish one circuit per session, obviously. But moreover proxies also should not establish, say, five circuits with one another and load-balance session traffic across

them.

4.4 Proliferation of Intermediaries

All of the above effects are compounded by the proliferation of intermediaries. In the worst case each administrative domain and/or each NAT boundary which session traffic traverses could conceivably introduce its own intermediary or intermediaries to a session.

The proliferation of intermediaries is undesirable as it leads to fate sharing among many unrelated elements in the network. This becomes especially problematic as sessions traverse different administrative domains each of which controls intermediaries.

Proliferation makes it much more likely that an individual session will fail, and much more difficult to diagnose failures when they occur.

4.5 Discovery of Unknown Intermediaries

It may not always be obvious to clients initiating a session that intermediaries have inserted themselves into the session path. However, because of the concerns raised above, endpoints may wish to know of the presence of intermediaries.

One mechanism that can be used to determine whether or not any intermediaries are in the session path is to send encrypted instant messages. If any intermediaries require access to the content of the messages in order to perform their function, then session establishment will fail. However, it may not be clear to either endpoint where or why session establishment has failed if this occurs.

It is therefore desirable that an intermediary have a mechanism for informing both participants in a session of the intermediary's presence.

Discovery will not by itself solve any of the concerns with intermediaries, of course. If an intermediary is broken, or its disposition prevents the creation of necessary security associations, then hopefully there is some way that clients can get around it in order to establish a session. This would only be possible if one of the clients had control over whether or not the intermediary is in the session path.

Following the recommendation of a 'one party consent' model given in [OPES], one of the principal participants in the session is required to explicitly authorize an intermediary to enter the session stream.

Service providers should not interpose an intermediary into a instant messaging session unless a client requests that the presence of an intermediary.

5. Normative Guidelines

o Preliminary signaling used to establish a session of instant messages MUST be capable of negotiating a common underlying transport protocol for instant messaging.

o Session messaging MUST NOT use UDP as a transport layer because of UDP's lack of congestion control. Transport protocols used for session messaging MUST support congestion control.

o A transport/session layer protocol for instant messaging sessions MUST explicitly specify the identities of the participants in the session, a unique session identifier, and sequencing information for messages in a session.

o A transport/session layer solution for instant messaging sessions MUST support end-to-end confidentiality and integrity mechanisms for instant messages.

o A transport/session layer solution for instant messaging sessions MUST support end-to-end mutual authentication.

o A transport/session layer solution for instant messaging sessions MUST support a mechanism through which a participant in a session can discover the presence of an intermediary.

o A transport/session layer solution for instant messaging sessions SHOULD support a mechanism for specifying a single transport protocol end-to-end.

o End-to-end security mechanisms for instant messaging sessions SHOULD accommodate network intermediaries that are have been authorized to act on the session. For example, headers on which intermediaries need to operate might be kept in cleartext while the remainder of an instant message was encrypted from end-to-end.

o Intermediaries SHOULD NOT interpose themselves between endpoints in a session unless they are specifically authorized to do so by one of the principal participants in a session.

o Any intermediaries interposing themselves in instant messaging sessions themselves MUST notify all participants of their presence through either the preliminary signaling or subsequent session messaging.

o Service providers SHOULD NOT deploy intermediaries where they are not absolutely necessary.

6. Security Considerations

Security considerations for instant messaging sessions are discussed in some detail in Sections 3.2 and 4.2.

7. IANA Considerations

This document does not have any implications for IANA.

8. References

[2543] M. Handley, H. Schulzrinne, E. Schooler, and J. Rosenberg, "SIP: session initiation protocol," RFC2543, Internet Engineering Task Force, Mar. 1999.

[1889] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: a transport protocol for real-time applications," RFC1889, Internet Engineering Task Force, Jan. 1996.

[MESSAGE] Rosenberg, J. , Willis, D. , Rosenberg, J. , Sparks, R. , Campbell, B. , Schulzrinne, H. , Lennox, J. , Huitema, C. , Aboba, B. , Gurle, D. and D. Oran, "SIP Extensions for Instant Messaging", draft-ietf-simple-im-01.txt (work in progress), July 2001.

[SESSION] Campbell, B. and J. Rosenberg, "SIP Instant Message Sessions", draft-ietf-simple-im-session-00.txt (work in progress), July 2001.

[IM-SDP] Campbell, B. and J. Rosenberg, "SDP Extensions for SIP Instant Message Sessions", internet-draft draft-ietf-simple-im-sdp-00.txt, July 2001

[2327] Handley, M. and V Jacobson, "SDP: Session Description Protocol", RFC 2327, April 1998.

[NAT] M. Holdrege and P. Srisuresh, "Protocol complications with the IP network address translator," Request for Comments 3027, Internet Engineering Task Force, Jan. 2001.

[NAT-G] D. Senie, "NAT friendly application design guidelines," Internet Draft, Internet Engineering Task Force, Mar. 2001. Work in progress.

[MIDCOM] P. Srisuresh, J. Kuthan, and J. Rosenberg, "Middlebox communication architecture and framework," Internet Draft, Internet

Engineering Task Force, Feb. 2001. Work in progress.

[STUN] J. Rosenberg, J. Weinberger, C. Huitema, R. Mahy, "STUN - Simple Traversal of UDP Through NATs", Internet Draft, Internet Engineering Task Force, Oct. 2001. Work in progress.

[OPES] Internet Architecture Board, "IAB Architectural and Policy Considerations for OPES", Internet-Draft, Internet Engineering Task Force, Oct. 2001. Work in progress.

[CONG] S. Floyd and K. Fall, "Promoting the Use of End-to-End Congestion Control in the Internet", IEEE/ACM Transactions on Networking, May 3 1999 (http://www.aciri.org/floyd/papers/collapse.may99.pdf)

9. Authors' Addresses

Allison Mankin
USC/ISI
4350 N. Fairfax Drive, Suite 620
Arlington VA 22203
Email: mankin@isi.edu
Phone: +1-703-812-3706

Jon Peterson
NeuStar, Inc.
1800 Sutter Street, Suite 570
Concord, CA 94520
Jon.Peterson@NeuStar.com

Full Copyright Statement