Network Working Group
Request for Comments: 3322
Category: Informational

H. Hannu
Ericsson

January 2003

Signaling Compression (SigComp) Requirements & Assumptions

Status of this Memo

Copyright Notice

Abstract

The purpose of this document is to outline requirements and motivations for the development of a scheme for compression and decompression of messages from signaling protocols. In wireless environments and especially in cellular systems, e.g., GSM (Global System for Mobile communications) and UMTS (Universal Mobile Telecommunications System), there is a need to maximize the transport efficiency for data over the radio interface. With the introduction of SIP/SDP (Session Initiation Protocol/Session Description Protocol) to cellular devices, compression of the signaling messages should be considered in order to improve both service availability and quality, mainly by reducing the user idle time, e.g., at call setup.

Table of Contents

1. Introduction

In wireless environments, and especially in cellular systems, such as GSM/GPRS, there is a need to maximize the transport efficiency of data over the radio interface. The radio spectrum is rather expensive and must be carefully used. Therefore, the cellular systems must support a sufficient number of users to make them economically feasible. Thus, there is a limitation in the per user bandwidth.

Compressing the headers of the network and transport protocols used for carrying user data is one way to make more efficient use of the scarce radio resources [ROHC]. However, compression of the messages from signaling protocols, such as SIP/SDP, should also be considered to increase the radio resource usage even further. Compression will also improve the service quality by reducing the user idle time at e.g., call setup. When IP is used end-to-end, new applications, such as streaming, will be brought to tiny end-hosts, such as cellular devices. This will introduce additional traffic in cellular systems. Compression of signaling messages, such as RTSP [RTSP], should also be considered to improve both the service availability and quality.

New services with their corresponding signaling protocols make it reasonable to consider a scheme that is generic. The scheme should be generic in the meaning that the scheme can efficiently be applied to arbitrary protocols with certain characteristics, such as the ASCII based protocols SIP and RTSP.

1.1. Protocol Characteristics

The following application signaling protocols are examples of protocols that are expected to be commonly used in the future. Some of their characteristics are described below.

1.1.1 SIP

The Session Initiation Protocol [SIP] is an application layer protocol for establishing, modifying and terminating multimedia sessions or calls. These sessions include Internet multimedia conferences, Internet telephony and similar applications. SIP can be used over either TCP [TCP] or UDP [UDP]. SIP is a text based protocol, using ISO 10646 in UTF-8 encoding.

1.1.2 SDP

The Session Description Protocol [SDP] is used to advertise multimedia conferences and communicate conference addresses and conference tool specific information. It is also used for general real-time multimedia session description purposes. SDP is carried in the message body of SIP and RTSP messages. SDP is text based using the ISO 10646 character set in UTF-8 encoding.

1.1.3 RTSP

The Real Time Streaming Protocol [RTSP] is an application level protocol for controlling the delivery of data with real-time properties, such as audio and video. RTSP may use UDP or TCP (or other) as a transport protocol. RTSP is text based using the ISO 10646 character set in UTF-8 encoding.

1.1.4 Protocol Similarities

The above protocols have many similarities. These similarities will have implications on solutions to the problems they create in conjunction with e.g., cellular radio access. The similarities include:

- Requests and reply characteristics. When a sender sends a request, it stays idle until it has received a response. Hence, it typically takes a number of round trip times to conclude e.g., a SIP session.

- They are ASCII based.

- They are generous in size in order to provide the necessary information to the session participants.

- SIP and RTSP share many common header field names, methods and status codes. The traffic patterns are also similar. The signaling is carried out primarily under the set up phase. For SIP, this means that the majority of the signaling is carried out to set up a phone call or multimedia session. For RTSP, the majority of the signaling is done before the transmission of application data.

1.2. Cellular System Radio Characteristics

Partly to enable high utilization of cellular systems, and partly due to the unreliable nature of the radio media, cellular links have characteristics that differ somewhat from a typical fixed link, e.g.,

copper or fiber. The most important characteristics are the lossy behavior of cellular links and the large round trip times.

The quality in a radio system typically changes from one radio frame to another due to fading in the radio channel. Due to the nature of the radio media and interference from other radio users, the average bit error rate (BER) can be 10e-3 with a variation roughly between 10e-2 to 10e-4. To be able to use the radio media with its error characteristics, methods such as forward error correction (FEC) and interleaving are used. If these methods were not used, the BER of a cellular radio channel would be around 10 %. Thus, radio links are, by nature, error prone. The final packet loss rate may be further reduced by applying low level retransmissions (ARQ) over the radio channel; however, this trades decreased packet loss rate for a larger delay. By applying methods to decrease BER, the system delay is increased. In some cellular systems, the algorithmic channel round trip delay is in the order of 80 ms. Other sources of delays are DSP-processing, node-internal delay and transmission. A general value for the RTT is difficult to state, but it might be as high as 200 ms.

For cellular systems it is of vital importance to have a sufficient number of users per cell; otherwise the system cost would prohibit deployment. It is crucial to use the existing bandwidth carefully; hence the average user bit rate is typically relatively low compared to the average user bit rate in wired line systems. This is especially important for mass market services like voice.

2. Motivation for Signaling Reduction

The need for solving the problems caused by the signaling protocol messages is exemplified in this chapter by looking at a typical SIP/SDP Call Setup sequence over a narrow band channel.

2.1 Estimation of Call Setup Delay Using SIP/SDP

Figure 2.1 shows an example of SIP signaling between two termination points with a wireless link between, and the resulting delay under certain system assumptions.

It should be noted that the used figures represent a very narrow band link. E.g., a WCDMA system can provide maximum bit rates up to 2 Mbits/s in ideal conditions, but that means one single user would consume all radio resources in the cell. For a mass market service such as voice, it is always crucial to reduce the bandwidth requirements for each user.

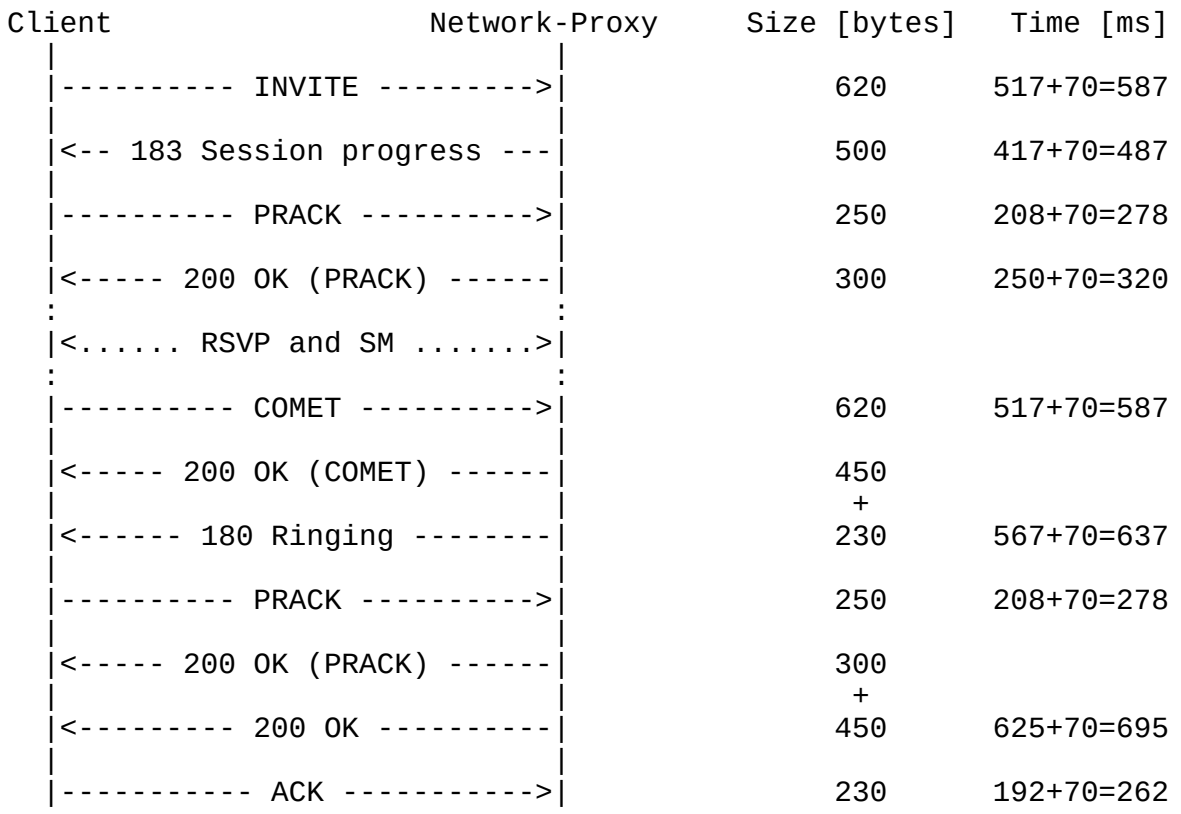

Figure 2.1. SIP signaling delays assuming a link speed of 9600 bits/sec and a RTT of 140 ms.

The one way delay is calculated according to the following equation:

OneWayDelay =
MessageSize[bits]/LinkSpeed[bits/sec] + RTT[sec]/2 (eq. 1)

The following values have been used:

RTT/2: 70 ms
LinkSpeed 9.6 kbps

The delay formula is based on an approximation of a WCDMA radio access method for speech services. The approximation is rather crude. For instance, delays caused by possible retransmissions due to errors are ignored. Further, these calculations also assume that there is only one cellular link in the path and take delays in an eventual intermediate IP-network into account. Even if this approximation is crude, it is still sufficient to provide representative numbers and enable comparisons. The message size given in Figure 2.1, is typical for a SIP/SDP call setup sequence.

2.1.1 Delay Results

Applying equation 1 to each SIP/SDP message shown in the example of Figure 2.1 gives a total delay of 4131 ms from the first SIP/SDP message to the last. The RSVP and Session Management (Radio Bearer setup), displayed in Figure 2.1, will add approximately 1.5 seconds to the total delay, using equation 1. However, there will also be RSVP and SM signaling prior to the SIP INVITE message to establish the radio bearer, which would add approximately another 1.5 seconds.

In [TSG] there is a comparison between GERAN call setup using SIP and ordinary GSM call setup. For a typical GSM call setup, the time is about 3.6 seconds, and for the case when using SIP, the call setup is approximately 7.9 seconds.

Another situation that would benefit from reduced signaling is carrying signaling messages over narrow bandwidth links in mid-call. For GERAN, this will result in frame stealing with degraded speech quality as a result.

Thus, solutions are needed to reduce the signaling delay and the required bandwidth when considering both system bandwidth requirements and service setup delays.

3. Alternatives for Signaling Reduction

More or less attractive solutions to the previously mentioned problems can be outlined:

- Increase the user bit rate

  An increase of the bit rate per user will decrease the number of users per cell. There exist systems (for example WCDMA) which can provide high bit rates and even variable rates, e.g., at the setup of new sessions. However, there are also systems, e.g., GSM/EDGE, where it is not possible to reach these high bit rates in all situations. At the cell borders, for example, the signal strength to noise ratio will be lower and result in a lower bit rate. In general, an unnecessary increase of the bit rate should be avoided due to the higher system cost introduced and the possibility of denial of service. The latter could, for example, be caused by lack of enough bandwidth to support the sending of the large setup message within a required time period, which is set for QoS reasons.

- Decrease the RTT of the cellular link

  Decreasing the RTT would require substantial system changes and is thus not feasible in the short term. Further, the RTT-delay caused by interleaving and FEC will always have to be present regardless of which system is used. Otherwise the BER will be too high for the received data to be useful, or alternatively trigger retransmissions giving an average total delay of the same or higher magnitude.

- Optimize message sequence for the protocols

  If the request/response pattern could be eased up, then "keeping the pipe full" could be a way forward. Thus, instead of following the message sequence described in Figure 4.2, more than one message would be sent in a row, even though no response has been received. However, this would entail protocol changes and may be difficult at the current date.

- Protocol stripping

  Removing fields from a message would decrease the size of the messages to some extent. However, this would cause the loss of transparency and thus violate the End-to-End principle and is thus not desirable.

- Compression

  By compressing messages, the impact of the mentioned problems could be decreased. Compared to the other possible solutions compression can be made, and must be, transparent to the end-user application. Thus, compression seems to be the most attractive way forward.

4. Assumptions

- Negotiation

  How the usage of compression is negotiated is out of the scope for this compression solution and must be handled by e.g., the protocol the messages of which are to be compressed.

- Reliable transport

  With reliable transport, it is assumed that a transport recovered from data that is damaged, lost, duplicated, or delivered out of order, e.g., [TCP].

- Unreliable transport

  With unreliable transport, it is assumed that a transport does not have the capabilities of a reliable transport, e.g., [UDP].

5. Requirements

This chapter states requirements for a signaling compression scheme to be developed in the IETF ROHC WG.

The requirements are divided into two parts. Section 5.1 sets general requirements concerning the Internet infrastructure, while Section 5.2 sets requirements on the scheme itself.

5.1. General Requirements

1. Transparency: When a message is compressed and then decompressed, the result must be bitwise identical to the original message.

   Justification: This is to ensure that the compression scheme will not cause problems for any current or future part of the Internet infrastructure.

   Note: See also requirement 9.

2. Header compression coexistence: The compression scheme must be able to coexist with header compression, especially the ROHC protocol.

   Justification: Signaling compression is used because there is a need to conserve bandwidth usage. In that case, header compression will likely be needed too.

3a. Compatibility: The compression scheme must be constructed in such a way that it allows the above protocols' mechanisms to negotiate whether the compression scheme is to be applied or not.

   Justification: Two entities must be able to communicate regardless if the signaling compression scheme is implemented at both entities or not.

3b. Ubiquity: Modifications to the protocols generating the messages that are to be compressed, must not be required for the compression scheme to work.

   Justification: This will simplify deployment of the compression scheme.

Note: This does not preclude making extensions, which are related to the signaling compression scheme, to existing protocols, as long as the extensions are backward compatible.

4. Generality: Compression of arbitrary message streams must be supported. The signaling compression scheme must not be limited to certain protocols, traffic patterns or sessions. It must not assume any message pattern to be able to perform compression.

   Justification: There might be a future need for compression of different ASCII based signaling protocols. This requirement will minimize future work.

   Note: This does not preclude optimization for certain streams.

5. Unidirectional routes: The compression scheme must be able to operate on unidirectional routes, i.e., without explicit feedback messages from the decompressor.

   Note: Implementations on unidirectional routes might possibly show a degraded performance compared to implementations on bi-directional routes.

6. Transport: The solution must work for both unreliable and reliable underlying transport protocols, e.g., UDP and TCP.

   Justification: The protocols, which generate the messages that are to be compressed, may use either an unreliable or a reliable underlying transport.

   Note: This should not be taken to mean that the same set of solution mechanisms must be used over both unreliable and reliable transport.

5.2. Performance Requirements

The performance requirements in this section and the following subsections are valid for both unreliable and reliable underlying transport.

7. Scalability: The scheme must be flexible to accommodate a range of compressors/decompressors with varying memory and processor capabilities.

   Justification: A primary target for the signaling compression scheme is cellular systems, where the mobile terminals have varying capabilities.

8. Delay: The signaling compression must not noticeably add to the delay experienced by the end user.

   Justification: Reduction of the user experienced delay is the main purpose of signaling compression.

   Note: This requirement is intended to prevent schemes that achieve compression efficiency at the expense of delay, i.e., queuing of messages to improve the compression efficiency should be avoided.

The following requirements are grouped into two subsections, a robustness section and a compression efficiency section.

5.2.1. Robustness

The requirements in this section concern the issue of when compressed messages should be correctly decompressed. The transparency requirement (first requirement) covers the issue with faulty decompressed messages.

9. Residual errors: The compression scheme must be resilient against errors undetected by lower layers, i.e., the probability of incorrect decompression caused by such undetected errors must be low.

   Justification: A primary target for the signaling compression scheme is cellular systems, where undetected errors might be introduced on the cellular link.

10. Error propagation: Propagation of errors due to signaling compression should be kept at an absolute minimum. Loss or damage to a single or several messages, between compressor and decompressor should not prevent compression and decompression of later messages.

    Justification: Error propagation reduces resource utilization and quality.

11. Delay: The compression scheme must be able to perform compression and decompression of messages under all expected delay conditions.

5.2.2. Compression Efficiency

This section states requirements related to compression efficiency.

12. Message loss: Loss or damage to a single or several messages, on the link between compressor and decompressor, should not prevent the usage of later messages in the compression and decompression process.

13. Moderate message misordering: The scheme should allow for the correct decompression of messages, that have been moderately misordered (1-2 messages) between compressor and decompressor. The scheme should not prevent the usage of later messages in the compression and decompression process.

    Justification: Misordering is frequent on the Internet, and this kind of misordering is common.

6. Security Considerations

A protocol specified to meet these requirements must be able to cope with packets that have undergone security measures, such as encryption, without adding any security risks. This document, by itself however, does not add any security risks.

7. IANA Considerations

A protocol which meets these requirements may require the IANA to assign various numbers. This document by itself however, does not require any IANA involvement.

8. References

[ROHC] Bormann, C., Burmeister, C., Degermark, M., Fukushima, H., Hannu, H., Jonsson, L-E., Hakenberg, R., Koren, T., Le, K., Liu, Z., Martensson, A., Miyazaki, A., Svanbro, K., Wiebke, T., Yoshimura, T. and H. Zheng, "RObust Header Compression (ROHC): Framework and four profiles: RTP, UDP, ESP, and uncompressed", RFC 3095, July 2001.

[RTSP] Schulzrinne, H., Rao, A. and R. Lanphier, "Real Time Streaming Protocol (RTSP)", RFC 2326, April 1998.

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

[SIP] Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M. and E. Schooler, "SIP: Session Initiation Protocol", RFC 3261, June 2002.

[UDP] Postel, J., "User Datagram Protocol", STD 6, RFC 768, August 1980.

[TCP] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, September 1981.

[TSG] Nortel Networks, "A Comparison Between GERAN Packet-Switched Call Setup Using SIP and GSM Circuit-Switched Call Setup Using RIL3-CC, RIL3-MM, RIL3-RR, and DTAP", 3GPP TSG GERAN #2, GP-000508, 6-10 November 2000.

9. Author's Address

Hans Hannu
Box 920
Ericsson AB
SE-971 28 Lulea, Sweden

Phone: +46 920 20 21 84
EMail: hans.hannu@epl.ericsson.se

10. Full Copyright Statement

Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.