Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 16, 2009

T. Asveren
Sonus Networks
V. Fajardo
Toshiba America Research Inc.
July 15, 2008

Diameter Congestion Signaling
draft-asveren-dime-cong-03.txt

Status of this Memo

Abstract

Diameter base protocol defines the network layer functionality to be used by applications. This document adds hop-to-hop congestion notification mechanism to that functionality.

Table of Contents

1. Introduction

Diameter base protocol defines the network layer functionality to be used by applications. Requests are routed based on Destination-Host AVP, Destination-Realm AVP, Application-Id values and the status of Diameter connections to neighboring peers.

This document defines a new AVP to be used by peers to notify their neighbors about their congestion status, so that this information can be used while routing requests. It is left up to the implementation to decide for local congestion levels but some guidelines are provided to prevent undesirable situations like oscilliation.

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

The following terms defines the functionality used in describing entities in this document.

Congestion

The situation on a Diameter node, where the current load is above the normal operational limits and effects processing of messages. A node which suffers from congestion should not be sent certain messages depending on the severity of congestion to prevent it to be overloaded further.

Congestion level

A numeric value used to quantify the severity of congestion on a Diameter node. This value is generalized in this document and not meant to be an exhaustive indicator of all possible variables that can define congestion. The possible numeric values for congestion level is defined in Section 5.2.

Congestion state

Congestion state pertains to the congestion level and other relevant information that a Diameter node keeps about each of its neighboring peers.

3. Motivation

When routing Diameter messages, it is preferable to consider the congestion status of peers to increase the response time of answers and to prevent overloading of nodes. A method of signaling the congestion state among adjacent peers independent of any applications will allow for a more real-time, self correcting method of reducing and or distributing load among Diameter neighbors.

There are several scenarios where relying on an application level result code is not efficient to notify peers about congestion status changes. The succeeding sections provides applicability scenarios for requiring per hop congestion status signaling.

3.1. Congestion of Intermediaries

In a Diameter network, intermediate nodes such as relay agents, proxies, may be present. Such nodes may get congested and it is desirable to consider their congestion status when selecting the next hop node.

Intermediate nodes do not host the application logic to process a request completely and do not generate answers except routing failures for the requests they receive. Generating a result code of "TOO_BUSY_HERE" to notify about intermediate congestion is not appropriate because it indicates congestion of the server specified in the Destination-Host AVP or in general the logical application service identified by Destination-Realm AVP and Application-Id. It is therefore limited to Diameter application end-points and does not consider the congestion state of intermediaries and other application traffic routed through them.

3.2. Multiple Applications On the Same Node

A Diameter node may host multiple applications simultaneously. Although it is possible to aggregate congestion status of the node on application logic level, it may be preferable to do this on a centralized logical entity like the Diameter base protocol in a layered architecture. A hop-to-hop message generated and consumed by base protocol layer would be more suitable for such a task split between different layers.

3.3. Congestion Detection Time

Relying on congestion notification via application level result code is inherently a reactive mechanism. This requires that an application level request needs to be received for congestion notification to be sent on the answer.

This problem can be aggravated in configurations such as Diameter servers which are communicating with multiple peers. A highly congested server can signal its congestion state to other peers only when those peers send a request to the server.

3.4. Notification of Congestion Abatement

Currently, there is no existing method of to signal end of congestion. Peers may probe the congested node periodically with new requests and can decide based on the result code of the corresponding answer whether congestion has abated. However, such method is not effective if peers have no application level request to send and therefore suffers the same drawbacks as Section 3.3. A hop-to-hop congestion indication message could provide notification of congestion abatement immediately.

3.5. Multiple Congestion Levels

A congestion result code provides only a single congestion level of "congested." For certain configurations it may be desirable to provide multiple congestion levels. Especially for the cases where load information is to be used for loadsharing purposes, multiple levels are desirable.

3.6. Shortcomings of Transport Layer Congestion Indications

Diameter uses TCP and SCTP as transport protocols between adjacent entities. Although the concept of receiver congestion is present in those protocols there are some reasons, which make their use unsuitable for Diameter congestion detection purposes:

- It is not straightforward to learn the current status of receiver windows with sockets API, which is the defacto standard for applications to access services of TCP and SCTP in common operating systems. For TCP, applications can be aware of congestion only when the receiver window is full. For SCTP, application need to poll the status of receiver window, there is no trigger mechanism present.

- Propogation of congestion may take longer than desired. Congestion will be visible on sender side only after it propogated to transport protocol layer, which may require multiple queues to be filled first on receiver side.

- When congested node has multiple connections, the receiver window for each connection needs to be full, i.e. if a node does not send messages to the congested node, it won't be able to learn about its congestion status or not before sending enough messages to

fill the receiver window.

4. Scope

This document defines mechanisms to communicate congestion information in a hop-by-hop fashion. This information can be used to protect overloaded nodes against further traffic being sent to them and to loadshare requests among multiple endpoints. Although the strategy mainly relies on hop-by-hop communication, it also defines new result codes to be used in an end-to-end fashion to prevent unnecessary request retransmission on base protocol level in case of congested nodes/services.

5. Hop-To-Hop Congestion Notification Mechanism

5.1. Congestion Level Signaling Procedures

5.1.1. Sending Congestion Information

A Diameter node sends a Congestion-Level AVP in a Device Watchdog Request message to its adjacent neighbors to indicate its current congestion level.

A node's congestion level SHOULD fall into one of the congestion levels defined in Section 5.2. A Diameter node SHOULD send a DWR to its neighboring peers as soon as it determines that its congestion level changes. Sending the DWR message with Congestion-Level AVP as soon as congestion level changes is important so that adjacent nodes can stop sending new requests to the congested node to prevent it to get further overloaded.

In the case where a new peer attempts to connect to an existing node supporting congestion control signaling, the Congestion-Level AVP Section 5.2 may also be sent by the node in the CEA message to immediately indicate to the new peer of the nodes congestion state. This is in the case where the existing peer is already experiencing high levels of congestion and would want to notify any new peer immediately rather than sending a DWR which has an inherent latency.

When a node receives a request and the node already notified its neighbors that it is unable to handle new requests, the node MAY silently drop the request or MAY send back an error answer with result code DIAMETER_CONGESTED.

5.1.2. Receiving Congestion Information

A peer receiving a Congestion-Level AVP in a DWR SHOULD create and maintain congestion state for the sender of the DWR if it has not already done so. The congestion state should at least contain the currently advertised congestion state of the peer.

The receiver of the DWR should react according to the congestion level information provided by the Congestion-Level AVP, i.e. it SHOULD NOT send messages which are not allowed by the corresponding congestion level.

It SHOULD be expected that Nodes MAY advertise congestion levels non-sequentially, e.g. a node may first advertise CONGESTION_LEVEL_1 and then CONGESTION_LEVEL_3.

In the case where a peer does not support congestion level based request routing, it SHOULD ignore the presence of Congestion-Level AVP in DWR, CER and CEA messages. Considering that M-bit is not set for Congestion-Level AVP, this behavior is guaranteed by nodes compliant to Diameter Base Protocol.

5.1.3. Local Congestion Level Determination Guidelines

Considering the vast amount of criteria which may be used as metrics when determining congestion levels and different architectures, this document does not mandate a mechanism to decide for different congestion levels.

For sender of the Congestion-Level AVP, it is left up to the implementation on determining the current congestion level of a Diameter node. The implementation may rely on the traffic rate, processing load, backend call latency, storage/resource availability etc. or any such combinations to determine the appropriate congestion level. Deciding when congestion level changes on a node is also implementation dependent but nodes SHOULD provide hysteresis between onset and abatement values of the congestion levels. Note that schemes to determine congestion level changes should not be very sensitive so as not to trigger sending many DWR message causing congestion control flapping among neighboring peers.

It is recommended that triggering of onset and abatement levels should be deterministic. It should be noted that nodes MAY also choose to use only a subset of the defined congestion level values, e.g. a node MAY use only CONGESTION_LEVEL_0 and CONGESTION_LEVEL_3 values to indicate a binary state of congested or not congested.

Diameter nodes SHOULD NOT send DWR messages with Congestion-Level AVP

very frequently, for example more than once a second. Frequent DWR transmissions has the adverse side effect of triggering false disconnection indication if the receiver is highly congested and cannot send a DWA within the appropriate time. In the case that a disconnection indication does occur due to failed DWR/DWA exchanges even if the DWR transmissions are set to an acceptable frequency, then the peers should follow the normal disconnection process specified in RFC3588.

It is RECOMMENDED that nodes change their congestion state and notify their neighbors before congestion gets severe enough to cause significant problems for the processing of pending and on the flight requests.

5.1.4. Preventing Unnecessary Retransmission

If an adjacent node to an endpoint, e.g. a relay agent or a proxy, is notified that the endpoint is unable to handle new requests, there is no need that the same request is retransmited via an alternate route as shown in Figure 1. In such a situation, the adjacent node SHOULD reply back with an error answer with result code DIAMETER_ENDPOINT_CONGESTED. The Origin-Host AVP MUST be populated with the identity of the congested endpoint and Error-Reporting-Host AVP MUST contain the identity of the error reporting host.

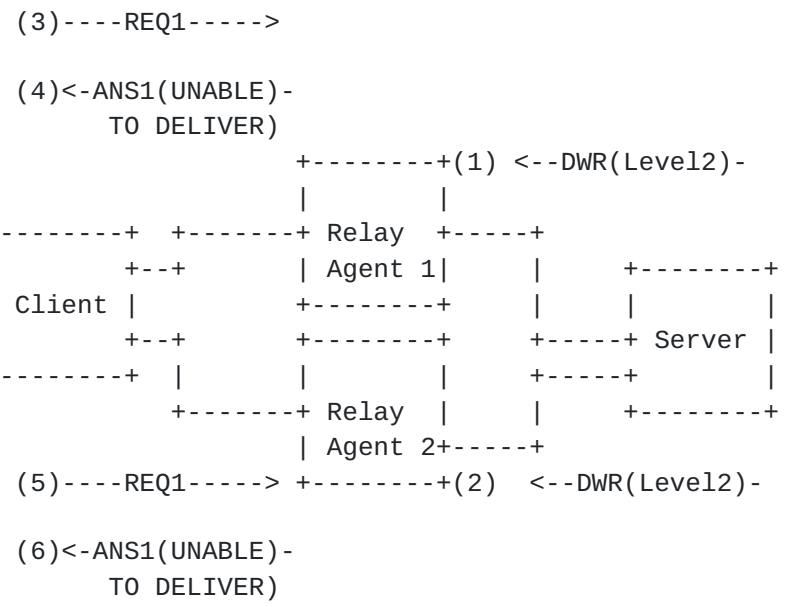

Figure 1: Unnecessary message retransmission during congestion

Similarly if a node adjacent to all endpoints providing service for a specific application in a realm has received congestion level updates from all of them indicating that they are unable to handle new

requests, the node SHOULD reply back with an DIAMETER_SERVICE_CONGESTED error answer, if it receives a request without Destination-Host AVP. The Origin-Host AVP MUST be populated with the identity of the error reporting host. It should be noted that nodes should generate this error answer if and only if they are sure that they have a connection to all of the endpoints providing the service for the corresponding realm. Otherwise an error answer with result code "UNABLE_TO_DELIVER" SHOULD be returned.

5.2. Congestion-Level AVP

Congestion-Level AVP is of type Enumerated and indicates the congestion level of a node. The following values are defined for Congestion-Level AVP:

CONGESTION_LEVEL_0 0

This value indicates that the load on the sender node is below the manageable limit and the node is ready to handle new messages.

CONGESTION_LEVEL_1 1

This value indicates that the load on the sender node is below the manageable limit but requests for new sessions SHOULD be sent preferrably to other nodes.

CONGESTION_LEVEL_2 2

This value indicates that no requests for new sessions SHOULD be sent to the node. A node in this state MAY drop request messages for new sessions. However, requests for existing sessions and answer messages still SHOULD be sent to the node.

CONGESTION_LEVEL_3 3

This value indicates that no new requests SHOULD be sent to the node even if they are requests for existing sessions. A node in this state MAY drop received request messages.

CONGESTION_LEVEL_4 4

This value indicates that no new messages (neither requests nor answers) SHOULD be sent to the node. A node in this state MAY drop any received message.

5.3. Error Answers

This document defines a new result code of protocol error class:

DIAMETER_CONGESTED 3011 A request has been received and the congestion state of the receiver node is not suitable to process the request.

DIAMETER_ENDPOINT_CONGESTED 3012 A request with a Destination-Host AVP is received, the destination of the request is an adjacent node and already declared that it is unable to handle new requests.

DIAMETER_APPLICATION_CONGESTED 3013 A request without a Destination-Host AVP is received, it is known that all potential endpoints for the request declared that they are unable to handle a new request.

6. Examples

6.1. Providing Hysteresis for Local Congestion Level Decision

This example assumes a local congestion level decision policy based on the number of messages in an incoming queue. The node decides for a maximum number of 1024 pending requests. This number could be based on the processing power of the node, the nature of the application and the expected message rate. The following figure displays possible onset/abatement values for different congestion levels.

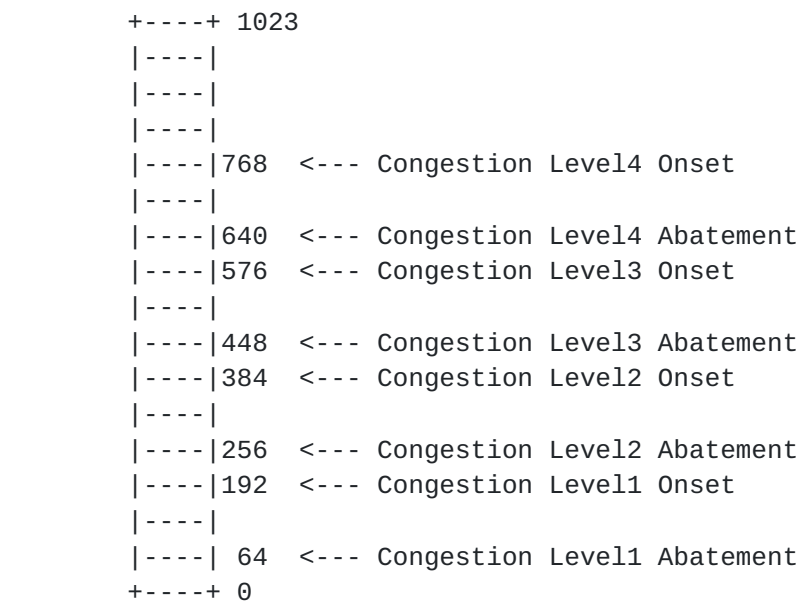

Figure 2: Congestion Onset/Abatement Levels

The difference between onset and abatement levels for the same congestion level is necessary to provide hysteresis so that congestion level does not change frequently between two levels.

6.2. Congestion Level Used For Loadsharing

The following scenario assumes a configration, where a relay agent is distributing traffic to two servers. It is assumed that the service consists of a single transaction, i.e. all requests belong to different sessions.

After Server2 declares that if there is some other server present, it should be preferred for new requests, relay agent stops loadsharing new requests among Server1 and Server2 and sends all new requests to Server1. When congestion state on Server2 is back to normal, Relay Agent continues to loadshare new requests among both servers.

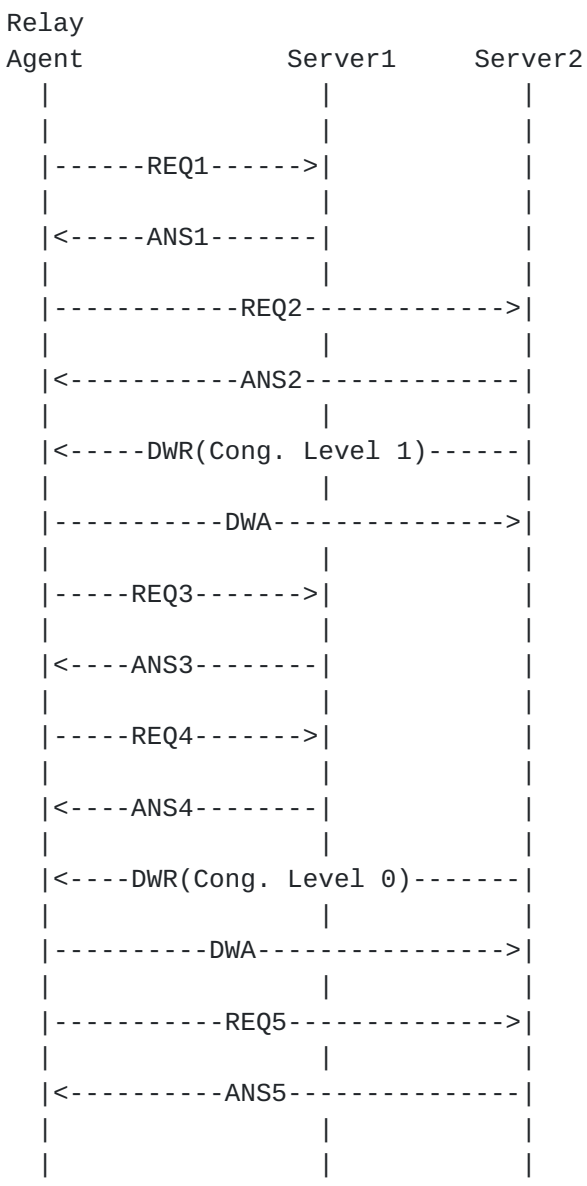

Figure 3: Congestion Level 1 Being Used in Loadsharing

7. IANA Considerations

IANA is to assign new AVP codes for Congestion-Level AVP defined in Section 5.2.

8. Security Considerations

This document does not contain a security protocol; it describes extensions to the existing Diameter protocol. All security issues of

DIAMETER protocol must be considered in implementing this specification. This extension does not add any unique concerns.

9. Acknowledgments

The authors wish to thank Bernard Aboba for his invaluable comments.

10. Normative References

[RFC3588] Calhoun, P., Loughney, J., Guttman, E., Zorn, G., and J. Arkko, "Diameter Base Protocol", RFC 3588, September 2003.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

Authors' Addresses

Tolga Asveren
Sonus Networks
4400 Route 9 South
Freehold, NJ, 07728
USA

Email: tasveren@sonusnet.com

Victor Fajardo
Toshiba America Research Inc.
One Telcordia Drive
Piscataway, NJ 08854
USA

Email: vfajardo@tari.toshiba.com

Full Copyright Statement

Intellectual Property