Diameter Maintenance and Extensions (DIME)
Internet-Draft
Intended status: Standards Track
Expires: November 27, 2015

S. Donovan
Oracle

May 26, 2015

Diameter Routing Message Priority
draft-donovan-dime-drmp-01.txt

Abstract

When making routing and resource allocation decisions, Diameter nodes currently have no generic mechanism to determine the relative priority of Diameter messages. This document defines a mechanism to allow Diameter endpoints to indicate the relative priority of Diameter transactions. With this information Diameter nodes can factor that priority into routing, resource allocation and overload abatement decisions.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on November 27, 2015.

Table of Contents

1. Introduction

The DOIC solution [I-D.ietf-dime-ovli] for Diameter overload control introduces scenarios where Diameter routing decisions made by Diameter nodes can be influenced by the overload state of other Diameter nodes. This includes the scenarios where Diameter endpoints and Diameter agents can throttle requests as a result of the target for the request being overloaded.

With currently available mechanisms these Diameter nodes do not have a clean mechanism to differentiate request message priorities when making these throttling decisions. As such, all requests are treated the same meaning that all requests have the same probability of being throttled.

There are scenarios where treating all requests the same can cause issues. For instance it might be considered important to reduce the probability of transactions involving first responders during a period of heavy signaling resulting from a natural disaster being throttled during overload scenarios.

This document defines a mechanism that allows Diameter nodes to indicate the relative priority of Diameter transactions. With this information other Diameter nodes can factor the relative priority of requests into routing and throttling decisions.

2. Terminology and Abbreviations

Diversion

As defined in [I-D.ietf-dime-ovli]. An overload abatement treatment where the reacting node selects alternate destinations or paths for requests.

DOIC

Diameter Overload Indication Conveyance.

DRMP

Diameter Routing Message Priority.

Overload Abatement

As defined in [I-D.ietf-dime-ovli]. Reaction to receipt of an overload report resulting in a reduction in traffic sent to the reporting node. Abatement actions include diversion and throttling.

Priority

The relative importance of a Diameter message. A higher priority value implies a higher relative importance of the message.

Throttling

As defined in [I-D.ietf-dime-ovli]. An abatement treatment that limits the number of requests sent by the DIOC reacting node. Throttling can include a Diameter Client choosing to not send requests, or a Diameter Agent or Server rejecting requests with appropriate error responses. In both cases the result of the throttling is a permanent rejection of the transaction.

3. Conventions Used in This Document

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 RFC 2119 [RFC2119].

RFC 2119 [RFC2119] interpretation does not apply for the above listed words when they are not used in all-caps format.

4. Problem Statement

With the introduction of overload control mechanisms, Diameter nodes will be required to make decisions regarding which Diameter request messages should be throttled as a result of overloaded Diameter nodes.

There is currently no generic mechanism to indicate which request messages should be given preferential treatment when these throttling decisions are made.

As a result, all messages are treated equally and, as such, have an equal probability of being throttled.

There are a number of scenarios where it is appropriate for an application to mark a request as being of a higher priority than other application requests. These are discussed in the next section.

This document defines a mechanism for applications to indicate priority for individual transactions, reducing the probability of those transactions being throttled if there are other lower priority transactions that are eligible for throttling treatment.

While the primary usage of DRMP defined priorities is for input to Diameter overload control related throttling decisions, it is also expected that the priority information could also be used for other routing related functionality. This might include giving higher priority transactions preferential treatment when selecting routes.

It is also envisioned that DRMP priority information could be used by Diameter endpoints to make resource allocation decisions. For instance, a Diameter Server might choose to use the priority information to treat higher priority requests ahead of lower priority requests.

> Note: There are a number of application specific definitions indicating various views of application level priority for different requests. Using these application specific priority AVPs as input to throttling and other Diameter routing decisions

would require Diameter agents to understand all applications and do application specific parsing of all messages in order to determine the priority of individual messages. This is considered an unacceptable level of complexity to put on elements whose primary responsibility is to route Diameter messages.

5. Use Cases

This section discussed various scenarios where Diameter transactions can benefit from the use of priority information.

5.1. First Responder Related Signaling

Natural disasters can result in a considerable increase in usage of network resources. This can be made worse if the disaster results in a loss of network capacity.

The combination of added load and reduced capacity can lead to Diameter nodes becoming overloaded and, as a result, the use of DOIC mechanisms to request a reduction in traffic. This in turn results in requests being throttled in an attempt to control the overload scenario and prevent the overloaded node from failing.

There is the need for first responders and other individuals responsible for handling the after effects of the disaster to be assured that they can gain access to the network resources in order to communicate both between themselves and with other network resources.

Signaling associated with first responders needs to be given a higher priority to help ensure they can most effectively do their job.

The United States Wireless Priority Services (WPS) and Government Emergency Telecommunications Service (GETS) are examples of systems designed to address these first responder needs.

5.2. Emergency Call Related Signaling

Similar to the first responder scenario, there is also signaling associated with emergency calls. Given the critical nature of these emergency calls, this signaling should also be given preferential treatment when possible.

5.3. Differentiated Services

Operators may desire to differentiate network-based services by providing a service level agreement that includes preferential

Diameter routing behavior. This might, for example, be modeled as Platinum, Gold and Silver levels of service.

In this scenario an operator might offer a Platinum SLA the includes ensuring that all signaling for a customer who purchases the Platinum service being marked as having a higher priority than signaling associated with Gold and Silver customers.

5.4. Application Specific Priorities

There are scenarios within Diameter applications where it might be appropriate to give a subset of the transactions for the application a higher priority than other transactions for that application.

For instance, when there is a series of transactions required for a user to gain access to network services, it might be appropriate to mark transactions that occur later in the series at a higher priority than those that occur early in the series. This would recognize that there was potentially significant work done by the network already that would be lost if those later transactions were throttled.

There are also scenarios where an agent cannot easily differentiate a request that starts a session from requests that update or end sessions. In these scenarios it might be appropriate to mark the requests that establish new sessions with a lower priority than updates and session ending requests. This also recognizes that more work has already taken place for established sessions and, as a result, it might be more harmful if the session update and session ending requests were to be throttled.

There are also scenarios where the priority of requests for individual command codes within an application depends on the context that exists when the request is sent. There isn't always information in the message from which this context can be determined by Diameter nodes other than the node that originates the request.

This is similar to the scenario where a series of requests are needed to access a network service. It is different in that the series of requests involve different application command-codes. In this scenario it is requests with the same command-code that have different implied priorities.

One example of this is in the 3GPP application [S6a] where a ULR request resulting from an MME restoration procedure might be given a higher priority than a ULR resulting from an initial attach.

6. Theory of Operation

This section outlines the envisioned usage of DRMP.

The expected behavior depends on the role (request sender, agent or request handler) of the Diameter node handling the request.

The following behavior is expected during the flow of a Diameter transaction.

1. Request sender - The sender of a request, be it a Diameter Client or a Diameter Server, determines the relative priority of the request and includes that priority information in the request. The method for determining the relative priority is application specific and is outside the scope of this specification. The request sender also saves the priority information with the transaction state. This will be used when handling the answer messages.

2. Agents handing the request - Agents use the priority information when making routing decisions. This can include determining which requests to route first, which requests to throttle and where the request is routed. For instance, requests with higher priority might have a lower probability of being throttled. The mechanism for how the agent determines which requests are throttled is implementation dependent and is outside the scope of this document. The agent also records the transaction priority in the transaction state. This will be used when handling the associated answer message for the transaction.

3. Request handler - The handler of the request, be it a Diameter Server or a Diameter Client, can use the priority information to determine how to handle the request. This could include determining the order in which requests are handled and resources that are applied to handling of the request.

4. Answer sender - The handler of the request is also the sender of the answer. The answer sender uses the priority information received in the request message when sending the answer. This implies that answers for higher priority transactions are given preferential treatment to lower priority transactions.

5. Agent handling the answer - Agents handling answer messages use the priority information stored with the transaction state to determine the priority of relaying the answer message. This implies that answers for higher priority transactions are given preferential treatment to lower priority transactions.

6. Answer handler - The handler of the answer message uses the priority of the transaction when allocating resources for processing that occurs after the receipt of the answer message.

7. Normative Behavior

This section contains the normative behavior associated with Diameter Resource Message Priority (DRMP).

When routing priority information is availble, Diameter nodes SHOULD include Diameter routing message priority in all Diameter request messages.

> Note: The method of determining the priority value included in the request is application specific and is not in the scope of this specification.

The priority marking scheme SHOULD NOT require the Diameter Agents to understand application specific AVPs.

When routing priority information is availble, Diameter nodes SHOULD use DRMP information when making Diameter overload related throttling decisions.

Diameter agents MAY use DRMP information when relaying messages. This includes the selection of routes and the ordering of messages relayed.

> The priority information included applies to both the request message and answer message associated with the transaction. As such it is used in the processing of both types of messages.

Diameter endpoints MAY use DRMP information when making resource allocation decisions for the transaction associated with the request message that contains the DRMP information.

Diameter endpoints MAY use DRMP information when making resource allocation decisions for the transaction associated with the answer messages using the DRMP information associated with the transaction.

When there is a mix of transactions specifying priority in request messages and transactions that do not have the priority specified, transactions that do not have a specified priority SHOULD be treated as having the PRIORITY_0 priority.

When setting and using priorities, PRIORITY_0 MUST be treated as the lowest priority.

When setting and using priorities, PRIORITY_1 MUST be treated as a higher priority than PRIORITY_0 and a lower priority than PRIORITY_2.

When setting and using priorities, PRIORITY_2 MUST be treated as a higher priority than PRIORITY_1 and a lower priority than PRIORITY_3.

When setting and using priorities, PRIORITY_3 MUST be treated as a higher priority than PRIORITY_2 and a lower priority than PRIORITY_4.

When setting and using priorities, PRIORITY_4 MUST be the highest priority.

> Editor's note: It is likely that there are other considerations for setting and using priorities. For instance, it might be good to use priority 1 to indicate elevated priority for strictly protocol reasons (e.g.; the S6a use case). Priorities 3, 4 and 5 would then be used for non protocol reasons.

8. Attribute Value Pairs

This section describes the encoding and semantics of the Diameter Overload Indication Attribute Value Pairs (AVPs) defined in this document.

8.1. DRMP AVP

The DRMP (AVP code TBD1) is of type Enumerated. The value of the AVP indicates the routing message priority for the transaction. The following values are initially defined:

PRIORITY_0 0 Priority 0 is the lowest priority.

PRIORITY_1 1 Priority 1 is the second lowest priority.

PRIORITY_2 2 Priority 2 is the middle priority.

PRIORITY_3 3 Priority 3 is the second higest priority.

PRIORITY_4 4 Priority 4 is the higest priority.

8.2. Attribute Value Pair flag rules

Attribute Name	AVP Code	Section Defined	Value Type	AVP flag rules: MUST	AVP flag rules: MUST NOT
DRMP	TBD1	8.1	Grouped		V

9. IANA Considerations

9.1. AVP codes

New AVPs defined by this specification are listed in Section 8. All AVP codes are allocated from the 'Authentication, Authorization, and Accounting (AAA) Parameters' AVP Codes registry.

9.2. New registries

There are no new IANA registries introduced by this document.

> Editor's Note: The current assumption is that there is no need to extend the number of priorities beyond the five defined in this specification. This assumption needs to be verified. If there is the need for extensability then a new IANA registry would be required. This new registry would be established as part of the standardization effort associated with the definition of new priority values.

10. Security Considerations

The DRMP could be used to get better access to services. This could result in one segment of a Diameter network limiting service to another segment of a Diameter network.

11. Contributors

The following people contributed substantial ideas, feedback, and discussion to this document:

- Janet P. Gunn

12. References

12.1. Normative References

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

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

[RFC6733] Fajardo, V., Arkko, J., Loughney, J., and G. Zorn, "Diameter Base Protocol", RFC 6733, October 2012.

12.2. Informative References

[I-D.ietf-dime-ovli] Korhonen, J., Donovan, S., Campbell, B., and L. Morand, "Diameter Overload Indication Conveyance", draft-ietf-dime-ovli-08 (work in progress), February 2015.

[RFC4412] Schulzrinne, H. and J. Polk, "Communications Resource Priority for the Session Initiation Protocol (SIP)", RFC 4412, February 2006.

[S6a] 3GPP, "Evolved Packet System (EPS); Mobility Management Entity (MME) and Serving GPRS Support Node (SGSN) related interfaces based on Diameter protocol", 3GPP TS 29.272 10.8.0, June 2013.

Appendix A. Design Considerations and Questions

This section contains a list of questions that will influence the design of the DRMP mechanism. It is expected that this section will be removed once the DRMP mechanism is defined.

A.1. Relationship with SIP Resource Priority

Question 1: Is there value with aligning the Diameter Routing Message Priority design with the SIP Resource Priority [RFC4412]work?

Current thoughts: SIP Resource Priority is considered to be addressing a superset of the requirements that DRMP addresses. The concensus seems to be that there is no need for multiple name spaces with DRMP.

Question 2: If so, is there value in reusing the existing SIP Resource Priority name spaces and request handling strategies?

Current thoughts: Given that the direction for DRMP is to have a single set of priority values, DRMP will not reuse name spaces.

A.2. Priority Encoding Method

Question 3: Is there a preference for handling DRMP by introducing AVPs or by using existing bits in the Diameter Command Flags field?

Current thoughts: The advantage of using bits in the Command Flags field is that it would reduce parsing overhead for elements that need access to the routing priority information. The question is whether this optimization in parsing overhead is worth the expense of using the reserved bits.

There are four bits remaining in the Command Flags header. If this approach is taken then the expectation would be that three of the bits would be used, allowing for eight priority levels.

This approach has questionable utility if multiple namespaces are to be used as the namespace identity would still require an AVP. Once the requirement for parsing the namespace AVP is introduced the incremental savings from utilizing the Command Flags would be minimal.

The current direction is to use AVPs to communicate priority. This gives the ability to extend the DRMP mechanism if additional functionality, such as name spaces, is determined to be required.

A.3. Base Protocol versus Application Extension

Question 4: Should DRMP be base protocol behavior or should Diameter applications be required to explicitly incorporate DRMP behavior?

The direction is to make the behavior generic across all applications.

A.4. Scope of Priority Setting

Question 5: Which of the following does the DRMP priority apply to:

Messages - meaning that a separate priority can be set for request messages and answer messages?

Transactions - meaning that the priority set in the request message also applies to the answer messages?

Request messages - meaning that answer message priority always has an implied higher priority than all request messages?

Current thoughts: The consensus is to have the DRMP priority apply to transactions.

Author's Address

Steve Donovan
Oracle
7460 Warren Parkway
Frisco, Texas 75034
United States

Email: srdonovan@usdonovans.com