Internet Engineering Task Force (IETF)
Request for Comments: 7944
Category: Standards Track
ISSN: 2070-1721

S. Donovan
Oracle

August 2016

Diameter Routing Message Priority

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 addresses this by defining 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 is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc7944.

Table of Contents

1. Introduction

The Diameter Overload Indication Conveyance (DOIC) solution [RFC7683] 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 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 being throttled during overload scenarios caused, for example, by a period of heavy signaling resulting from a natural disaster.

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.

1.1. Applicability

There are two primary considerations that must be addressed for the mechanism described in this document to work effectively. The first takes into consideration the fact that the Diameter base protocol defined in [RFC6733] is designed to transport multiple Diameter applications and that Diameter nodes can be implemented that support multiple applications. In order for the Diameter Routing Message Priority (DRMP) mechanism to work, the priorities defined for all messages across all applications used in a Diameter administrative domain must be defined in a consistent and coordinated fashion, taking the default priority into account. See Section 10 for a discussion of some of the considerations that need to be factored into the setting of DRMPs used by Diameter applications.

> Note that this consideration does not apply to Diameter networks where all Diameter nodes only support a single application.

Without this cross application priority design taken into consideration, it is possible for messages for one application to gain unwarranted preferential treatment over messages for other applications.

This mechanism also depends on all of the messages that carry the DRMP Attribute Value Pair (AVP) that are inserted into Diameter messages by trusted nodes within the Diameter administrative domain. As discussed in Section 12, misbehaving nodes have the ability to use the DRMP mechanism to gain unwarranted preferential treatment.

When messages cross Diameter administrative boundaries, care should be taken to either strip or modify the DRMP values in these messages. If the priority definitions vary between the two Diameter administrative domains, then it is possible for messages from a foreign domain to gain unwarranted preferential treatment.

2. Terminology and Abbreviations

Diversion

As defined in [RFC7683]. 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 [RFC7683]. 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 lower-priority value implies a higher relative importance of the message.

Throttling

As defined in [RFC7683]. An abatement treatment that limits the number of requests sent by the DOIC 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 [RFC2119].

The interpretation from RFC 2119 does not apply for the above listed words when they are not used in all caps.

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 throttling decisions related to Diameter overload control, 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 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. It might also use the priority information as a reason to fail a request as a result of insufficient resources.

> 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 discusses various scenarios where Diameter transactions can benefit from the use of priority information.

It is important to note that for priority information to be reliably usable, the Diameter nodes sending and consuming DRMP AVPs must have pre-established trust relationships of the sort described in Section 12.

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. In turn, this 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 jobs.

The United States Wireless Priority Services (WPS) and Government Emergency Telecommunications Service (GETS) are examples of systems designed to address the command and control aspects of 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 (SLA) 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 that includes ensuring that all signaling for a customer who purchases the Platinum service is 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 from a signaling point of view 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 involves different application command codes. In this scenario, requests with the same command code have different implied priorities.

One example of this is in the 3GPP application [S6a] where an Update Location Request (ULR) resulting from a Mobility Management Entity (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 handling 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 candidates to be throttled is implementation dependent and is outside the scope of this document. Before forwarding request messages, agents generally do not modify the priority information present in the received request message nor include the priority information when absent in the received request message. However, in some scenarios, agents can modify the priority information, for example, edge agents modifying the priority values set by an adjacent operator. There might be other scenarios where a Diameter endpoint does not support the DRMP mechanism, and agents insert the priority information in the request messages for that non-supporting endpoint. When forwarding the request messages, the agent also saves the transaction priority in the transaction state either as locally managed state or using the Proxy-Info mechanism defined in [RFC6733]. 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 the 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 over lower-priority transactions. The answer sender also has the option of including priority information in the answer message. This is done when the answer message needs to have a different priority than the priority carried in the request message. The priority included by the answer sender is application specific.

5. Agent handling the answer -- By default, agents handling answer messages use the priority information stored with the transaction state to determine the priority of relaying the answer message. However, priority information included in the answer message, when present, is used in place of the stored priority information. The use of priority information implies that answers for higher-priority transactions are given preferential treatment over lower-priority transactions. When forwarding the answer messages, agents generally do not modify the priority information present in the received answer messages nor include the priority information when absent in the received answer messages. However, in some scenarios, agents can modify the priority information, for example, edge agents modifying the priority values set by an adjacent operator. There might be other scenarios where a Diameter endpoint does not support the DRMP mechanism, and agents insert the priority information for that non-supporting endpoint.

6. Answer handler -- The answer handler uses the same method as the agent to determine the priority of the answer message. By default, the handler of the answer message uses the priority saved in the transaction’s state. Priority information in the answer message is used when present. The priority is used when allocating resources for processing that occurs after the receipt of the answer message.

7. Extensibility

This document does not define extensibility mechanisms that are specific to the DRMP mechanism. As a result, any extension that requires new AVPs will be required to use existing Diameter extensibility mechanisms defined in [RFC6733].

8. Normative Behavior

This section contains the normative behavior associated with DRMP.

When routing priority information is available, Diameter nodes SHOULD include Diameter routing message priority in the DRMP AVP 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 does not require the Diameter Agents to understand application-specific AVPs.

When available, Diameter nodes SHOULD use routing priority information included in the DRMP AVP when making Diameter overload throttling decisions.

Diameter Agents MAY use routing priority information included in the DRMP AVP when relaying request and answer messages. This includes the selection of routes and the ordering of messages relayed.

> Note: The priority information included in the DRMP AVP in request messages applies to both the request message and, by default, the answer message associated with the transaction.

While done only in exceptional circumstances, Diameter Agents MAY modify priority information when relaying request and answer messages.

> Note: There might be scenarios where a Diameter Agent does modify priority information. For instance, an edge agent might need to modify the priority values set by an adjacent operator.

While done only in exceptional circumstances, Diameter Agents MAY add priority information when relaying request and answer messages.

> Note: There might be scenarios where a Diameter endpoint does not support the DRMP mechanism, and agents insert priority information for that non-supporting endpoint.

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

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

Diameter endpoints MAY include the DRMP AVP in answer messages. This is done when the priority for the answer message needs to have a different priority than the priority carried in the request message.

When determining the priority to apply to answer messages, Diameter nodes SHOULD use the priority indicated in the DRMP AVP carried in the answer message, if it exists. If there is not DRMP AVP in the answer message, then the Diameter node SHOULD use the priority indicated in the DRMP AVP of the associated request message.

> Note: One method to determine what priority to apply to an answer when there is no DRMP AVP in the answer message is to save the priority included in the request message in the state associated with the Diameter transaction. Another is to use the Proxy-Info mechanism defined in [RFC6733].

Diameter nodes MUST have a default priority to apply to transactions that do not have an explicit priority set in the DRMP AVP.

In order to guarantee consistent handling of messages from non-upgraded Diameter Clients, Diameter nodes SHOULD use the PRIORITY_10 priority as this default priority value.

> PRIORITY_10 is a midrange priority that corresponds to "normal" traffic and thus would be a suitable default for most deployments, while still allowing different Diameter applications to designate other priorities for lower- and higher-priority traffic.
>
> Note: This does not imply that a DRMP AVP is added to the message. Rather, the message is treated the same as a message that has a DRMP AVP with a priority value of PRIORITY_10.

Diameter nodes MUST support the ability for the default priority to be modified through local configuration interfaces.

> Note: There are scenarios where operators might want to specify a different default value for transactions that do not have an explicit priority. In this case, the operator-defined local

policy would override the use of PRIORITY_10 as the default priority.

When using DRMP information, Diameter nodes MUST use the default priority for transactions that do not have priority specified in a DRMP AVP.

Note: This guidance on the handling of messages without a priority does not result in a Diameter Agent inserting a DRMP AVP into the message. Rather, it gives guidance on how that specific transaction should be treated when its priority is compared with other requests. When a Diameter Agent relays the request, it will not insert a DRMP AVP with a priority value of 10.

When setting and using priorities, for all integers x,y in [0,15], treat PRIORITY_<x> as lower priority than PRIORITY_<y> when y<x.

Note: As a result, PRIORITY_0 is the highest priority.

9. Attribute Value Pairs

This section describes the encoding and semantics of the Diameter Routing Message Priority AVP defined in this document.

9.1. DRMP AVP

The DRMP (AVP code 301) is of type Enumerated. The value of the AVP indicates the routing message priority for the transaction. The following values are defined:

PRIORITY_15 15 PRIORITY_15 is the lowest priority.

PRIORITY_14 14 PRIORITY_14 is a higher priority than PRIORITY_15 and a lower priority than PRIORITY_13.

PRIORITY_13 13 PRIORITY_13 is a higher priority than PRIORITY_14 and a lower priority than PRIORITY_12.

PRIORITY_12 12 PRIORITY_12 is a higher priority than PRIORITY_13 and a lower priority than PRIORITY_11.

PRIORITY_11 11 PRIORITY_11 is a higher priority than PRIORITY_12 and a lower priority than PRIORITY_10.

PRIORITY_10 10 PRIORITY_10 is a higher priority than PRIORITY_11 and a lower priority than PRIORITY_9.

PRIORITY_9 9 PRIORITY_9 is a higher priority than PRIORITY_10 and a lower priority than PRIORITY_8.

PRIORITY_8 8 PRIORITY_8 is a higher priority than PRIORITY_9 and a lower priority than PRIORITY_7.

PRIORITY_7 7 PRIORITY_7 is a higher priority than PRIORITY_8 and a lower priority than PRIORITY_6.

PRIORITY_6 6 PRIORITY_6 is a higher priority than PRIORITY_7 and a lower priority than PRIORITY_5.

PRIORITY_5 5 PRIORITY_5 is a higher priority than PRIORITY_6 and a lower priority than PRIORITY_4.

PRIORITY_4 4 PRIORITY_4 is a higher priority than PRIORITY_5 and a lower priority than PRIORITY_3.

PRIORITY_3 3 PRIORITY_3 is a higher priority than PRIORITY_4 and a lower priority than PRIORITY_2.

PRIORITY_2 2 PRIORITY_2 is a higher priority than PRIORITY_3 and a lower priority than PRIORITY_1.

PRIORITY_1 1 PRIORITY_1 is a higher priority than PRIORITY_2 and a lower priority than PRIORITY_0.

PRIORITY_0 0 Priority 0 is the highest priority.

9.2. Attribute Value Pair Flag Rules

Attribute Name	AVP Code	Section Defined	Value Type	AVP Flag Rules		
				MUST		MUST NOT
DRMP	301	9.1	Enumerated			V

10. Considerations When Defining Application Priorities

As discussed in Section 1.1, it is important that the definition of priority values used by all applications within a single Diameter administrative domain be done in a consistent and coordinated manner.

The following are some things to be considered when defining the DRMPs to be used in Diameter networks that support Diameter nodes handling multiple applications.

1. As with any prioritization scheme, it is possible for higher-priority messages to block lower-priority messages from ever being handled. In a Diameter network, this will often result in those Diameter transactions being retried. This can result in more traffic than the network would have handled without use of the DRMP mechanism.

   One potential guideline to prevent unwanted starving of lower-priority messages is to have higher-priority messages represent a relatively small portion of messages handled by the Diameter network under normal scenarios.

   > Note that there are scenarios, such as first responder messages, where the blocking of lower-priority messages is a requirement.

2. When setting priorities for any of the use cases outlined in Section 5, it is important to use the same priority values across applications. For instance, when defining priority for the first responder use case discussed in Section 5.1 and the emergency call use case discussed in Section 5.2, one high-priority value might be used for all first responder messages, say PRIORITY_2, and a slightly lower-priority value, say PRIORITY_3, might be used for emergency-call-related messages. These values should be specified for these use cases across all applications used within the Diameter administrative domain.

   > Note that the values mentioned here are strictly for illustrative purposes. The actual values used for these use cases are likely to be different.

3. Messages without the DRMP AVP will be given default priority value treatment. This will include messages from Diameter Clients that have not been updated to support the DRMP mechanism. It might also include messages from foreign administrative domains if the DRMP AVPs are stripped from messages crossing the Diameter administrative domains.

4. The process used to introduce the DRMP mechanism into a Diameter network should also be taken into consideration. Messages of the same type within the same application might get different treatment depending on whether those messages are sent from nodes that are upgraded to support the DRMP mechanism versus nodes that have not yet been upgraded to support the DRMP mechanism.

11. IANA Considerations

11.1. AVP Codes

The new AVP defined by this specification is listed in Section 9. All AVP codes are allocated from the "AVP Codes" subregistry of the "Authentication, Authorization, and Accounting (AAA) Parameters" registry.

12. Security Considerations

DRMP gives Diameter nodes the ability to influence which requests are throttled during overload scenarios. In addition, DRMP can be used in determining the routing decisions for request messages. Improper use of the DRMP mechanism could result in the malicious Diameter node gaining preferential treatment, by reducing the probability of its requests being throttled, over other Diameter nodes. This would be achieved by the malicious node inserting priority values that are artificially high.

Diameter does not include features to provide end-to-end authentication, integrity protection, or confidentiality. This opens the possibility that malicious or compromised agents in the path of a request could modify the DRMP AVP to reflect a priority different than that asserted by the sender of the request.

12.1. Potential Threat Modes

The Diameter protocol involves transactions in the form of requests and answers exchanged between clients and servers. These clients and servers may be peers; that is, they may share a direct transport (e.g., the Transmission Control Protocol (TCP) or Stream Control Transmission Protocol (SCTP)) connection, or the messages may traverse one or more intermediaries, known as Diameter Agents. Diameter nodes use Transport Layer Security (TLS), Datagram Transport Layer Security (DTLS), or IPsec to authenticate peers and to provide confidentiality and integrity protection of traffic between peers. Nodes can make authorization decisions based on the peer identities authenticated at the transport layer.

When agents are involved, this presents an effectively transitive trust model. That is, a Diameter Client or Server can authorize an agent for certain actions, but it must trust that agent to make appropriate authorization decisions about its peers, and so on. Since confidentiality and integrity protection occurs at the transport layer, agents can read, and perhaps modify, any part of a Diameter message, including the DRMP AVP.

There are several ways an attacker might attempt to exploit the DRMP mechanism. A malicious or compromised Diameter node might insert invalid priority values resulting in either preferential treatment, resulting from higher values, or degraded treatment resulting from lower values, for that node.

A similar attack involves a malicious or compromised Diameter Agent changing the priority value resulting in the sending Diameter node getting either preferential or degraded service.

The DRMP mechanism can be used to aid in overload throttling decisions. When this is the case, then the above attacks are limited in scope to when one or more Diameter nodes are in an overloaded state.

The DRMP mechanism can also be used to influence the order in which Diameter messages are handled by Diameter nodes. The above attacks have a potentially greater impact in this scenario as the priority indication impacts the handling of all requests at all times, independent of the overload status of Diameter nodes in the Diameter network.

12.2. Denial-of-Service Attacks

The DRMP mechanism does not open direct denial-of-service attack vectors. Rather, it introduces a mechanism where a node can gain unwarranted preferential treatment. It also introduces a mechanism where a node can get degraded service in the scenario where a rogue agent changes the priority value included in messages.

12.3. End-to-End Security Issues

The lack of end-to-end integrity features in Diameter [RFC6733] makes it difficult to establish trust in DRMP AVPs received from non-adjacent nodes. Any agents in the message path may insert or modify DRMP AVPs. Nodes must trust that their adjacent peers perform proper checks on overload reports from their peers, and so on, creating a transitive-trust requirement extending for potentially long chains of nodes. Network operators must determine if this transitive trust requirement is acceptable for their deployments. Nodes supporting

DRMP MUST give operators the ability to select which peers are trusted to deliver DRMP AVPs, and whether they are trusted to forward the DRMP AVPs from non-adjacent nodes. Diameter nodes MUST strip DRMP AVPs from messages received from peers that are not trusted for DRMP purposes.

It is expected that work on end-to-end Diameter security might make it easier to establish trust in non-adjacent nodes for DRMP purposes. Readers should be reminded, however, that the DRMP mechanism allows Diameter Agents to modify AVPs in existing messages that are originated by other nodes. If end-to-end security is enabled, there is a risk that such modification could violate integrity protection. The details of using any future Diameter end-to-end security mechanism with DRMP will require careful consideration and are beyond the scope of this document.

13. References

13.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC6733] Fajardo, V., Ed., Arkko, J., Loughney, J., and G. Zorn, Ed., "Diameter Base Protocol", RFC 6733, DOI 10.17487/RFC6733, October 2012, <http://www.rfc-editor.org/info/rfc6733>.

13.2. Informative References

[RFC7683] Korhonen, J., Ed., Donovan, S., Ed., Campbell, B., and L. Morand, "Diameter Overload Indication Conveyance", RFC 7683, DOI 10.17487/RFC7683, October 2015, <http://www.rfc-editor.org/info/rfc7683>.

[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, 14.0.0, June 2016, <http://www.3gpp.org/ftp/Specs/html-info/29272.htm>.

Contributors

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

- Janet P. Gunn

Author's Address

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

Email: srdonovan@usdonovans.com