Diameter Overload Control Design Team Report

DIME WG – IETF88

draft-docdt-dime-ovli-01

Design Team Report

Background

- A design team formed after IETF87 to work on the Diameter Overload Control solution proposal
  - Jouni Korhonen, (Hannes Tscofenig), Steve Donovan, Ben Campbell, Nirav Salot, Lionel Morand, Susan Shishufeng, Maria Cruz Bartolome, Martin Dolly, Jean-Jacques Trottin, Ulrich Wiehe

- Mail list doc-dt@ietf.org , archives available

- Weekly calls

- One f2f meeting after the 3GPP CT4#62bis

- Solution wanted/needed for 3GPP Release-12

Main Solution Principles

- Piggybacking
  - Can be used on top of existing applications..
  - The context of the overload control is determined by the "underlying" application the overload control is piggybacked on.

- Capability announcement
  - "Client" announces what it is capable of and "server" does the same.. At least one of the capabilities have to match.

- Extensibility
  - New functionality, algorithms, etc can added and registered with IANA.. and then announced as new capabilities.

Main Solution Principles cont'd

- Default (loss-like) algorithm and traffic abatement
  - Left for the "client" to figure out based on the Overload Report sent by the "server".
  - The report is only a "server" indicated "reduction percentage".

- The "endpoint" principle
  - Overload control is considered as an overlay on top of an arbitrary Diameter deployment.
  - The overload control information is exchanged two between "endpoints" capable of overload control solution.
  - Specific "reacting node" and "reporting node" roles, not to tie the solution specifically to "client-server" solution.

Decisions..

- The Diameter overload control "baseline solution" is not going to fulfill all requirement document requirements:
  - Separate documents will be needed for features that did not fit into the base line.. Take the agent overload as an example.

- Intentional separation between the overload reporting and overload control:
  - The baseline only solves the reactive reporting part i.e. the "Diameter Overload Indication Conveyance".
  - Pro-active overload controlling left for future work.

- No explicit algorithm identifiers
  - The algorithms can be deducted from the capability announcements and per capability/feature specific AVPs.

Open Issues and parts under discussion in -01

- Several "bigger" open issues
  - Extensibility and capability announcement details to be nailed down.
  - Destination-Realm and Destination-Host routed requests details missing.

- Missing
  - Basic overload report processing description missing/stale for the reacting and reporting endpoints (e.g. for client/server).

- Features under discussion
  - Inserting throttling information into requests.

- Loads of cleanup for -02 ahead.

Issue: Extensibility and capability announcement

- Plain feature vector is not really enough
  - Change the "flag vector" to a grouped AVP.
  - Need to add timestamp/sequence number to indicate validity of the announced features.

- Remove the existing "negotiation" part
  - It is a bidirectional announcement of capabilities.
  - Obviously at least one of the announced capabilities need to overload for endpoints to be able to perform overload control information conveyance..

Missing: overload report processing

- Just write it down..

- Would "detail" the use of the default algorithm..

Issue: Destination-Realm and -Host routed requests details

- Proposal sent to the list by Ben..

- Review it and tell whether it is acceptable

Proposals under discussion: throttling information into requests

- A request would contain information that a specific request survived throttling done by the reacting node.

- Indicates to on path nodes / reporting node that someone is _doing_ traffic abatement..
  - Additional knowledge to announced features..

- Not decided whether this is needed for the baseline.

Next step..

- Ship -02 asap incorporating the resolution for known issues and filling the missing text pieces.
  - Above changes could also be incorporated to WG adopted -00 revision..

- Adopt as a WG I-D ?
  - We admit -01 is still incomplete but from the design team point of view mature enough to serve as a base for the baseline solution.
  - We need to get a WG solution document out of the working group fast that our "waiting customer" (3GPP) can proceed with their work.

Comments / Questions?