Session Initiation Protocol (SIP) overload control
draft-gurbani-soc-overload-control-01
(V. K. Gurbani (Ed.), V. Hilt and H. Schulzrinne)

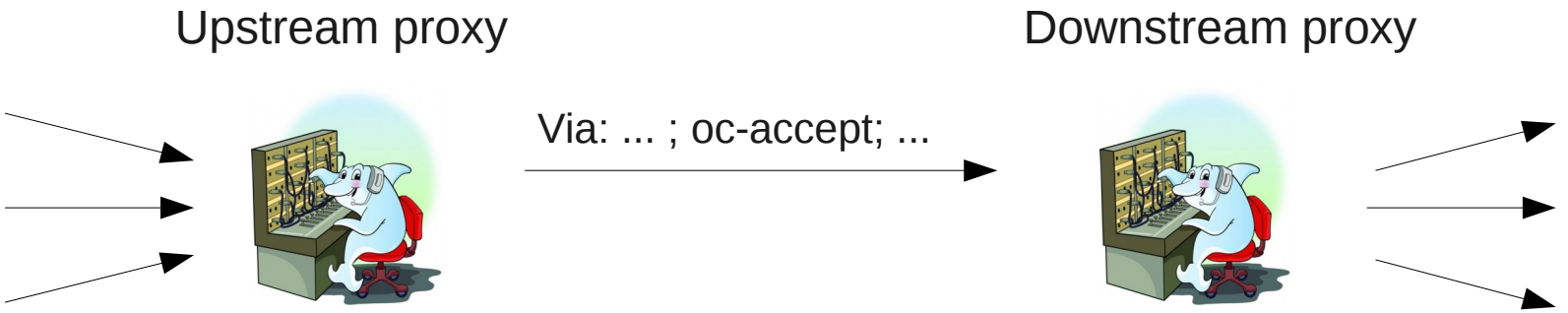
Presenter: Vijay K. Gurbani
IETF 78, Maastricht, Netherlands
July 26, 2010

Alcatel·Lucent

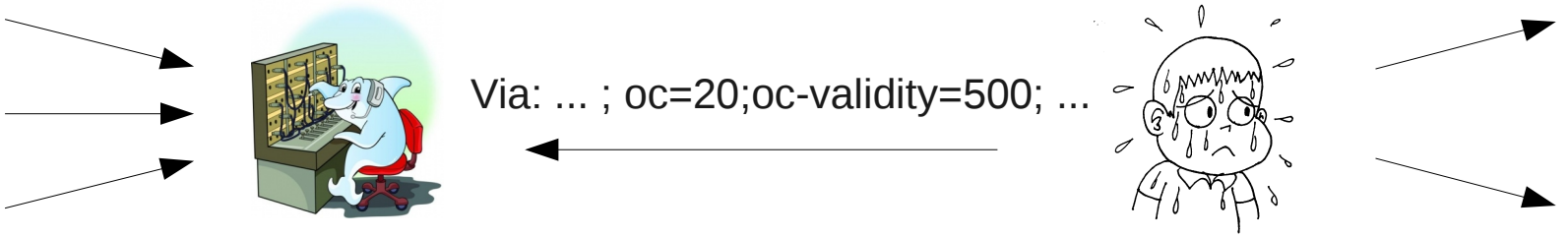
Draft background

- Work started in October 2006 (draft-hilt-sipping-overload).
- SOC working group formed in June-July 2010.
- 8 revisions later, draft-hilt-sipping-overload transitions to draft-gurbani-soc-overload-control (-00, June 2010) under the newly formed SOC working group.

Brief overview of operations



Downstream proxy gets overloaded

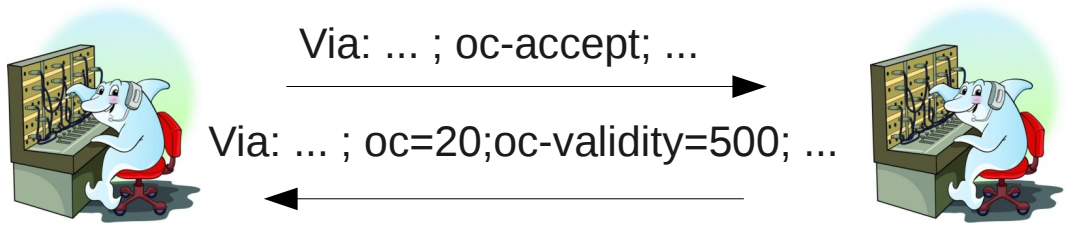


Traffic to downstream server reduced by 20% for 500ms.

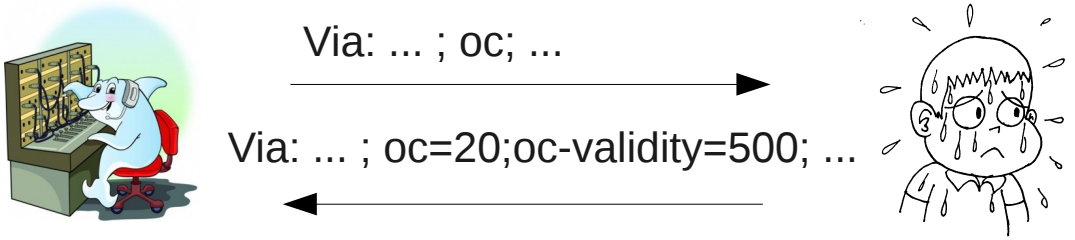
vkg@bell-labs.com

List discussion

- Reduction of Via parameters
  - oc-accept – Why not follow rport semantics?
  - Instead of:



  - How about:



List discussion

- Reduction of Via parameters
  - oc-port: current semantics are that overload applies to all ports on IP address. If want to choose a single port, use the “oc-port” parameter.
  - Why not make ip+port the default, and get rid of the “oc-port” parameter altogether?

List discussion

- Simplification of “oc-seq” parameter.
  - Current ABNF:

    oc-seq = (1*DIGIT) / (1*DIGIT “.” 1*DIGIT)

    Forces upstream server to support two formats.

  - Consider instead:

    oc-seq = (1*DIGIT “.” 1*DIGIT)

List discussion

- R-P header summary.
  - Upstream and downstream servers need to agree on which R-P namespace they will support.
  - Some out-of-band (read: “business arrangement”) method will be required to converge on the chosen namespace.

Open issue 1

- Message prioritization (discussed extensively on list).
  - Problem: How does the upstream node determine which messages to drop?
  - Provide guidelines to determine which messages can be dropped.
  - Strawman guideline from list discussion between Paul K., Volker H., Janet G., Keith D., and Vijay G.*:

* See list discussions between Jul 02, 2010 to Jul 22, 2010.

Open issue 1

- Messages that must not be dropped:
  - R-P header with an agreed-upon (by the upstream and downstream server) namespace take precedence and must not be dropped during overload;
  - In-dialogue messages must not be dropped to the extent possible.
- Messages that can be dropped:
  - Out-of-dialogue messages --- which for this purpose we can define as messages that does not have a To tag --- get dropped under overload;
  - some in-dialogue messages may have to be dropped depending on the arrival rate of in-dialogue messages and the loss rate preferred by the downstream server.

Open issue 2

If a downstream neighbor does not respond to a request at all, the upstream SIP server will stop sending requests to the downstream neighbor. The upstream SIP server will periodically forward a single request to probe the health of its downstream neighbor. It has been suggested --- see http://www.ietf.org/mail-archive/web/sip-overload/current/msg00229.html --- that we have a notification mechanism in place for the downstream neighbor to signal to the upstream SIP server that it is ready to receive requests. This notification scheme has advantages, but comes with obvious disadvantages as well.

Open issue 3

The 'oc' parameter value specified in this document is defined to contain a loss rate. However, other types of overload control feedback exist, for example, a target rate for rate-based overload control or message confirmations and window-size for window-based overload control.

While it would in theory be possible to allow multiple types of overload control feedback to co-exist (e.g., by using different parameters for the different feedback types) it is very problematic for interoperability purposes and would require SIP servers to implement multiple overload control mechanisms.

Next steps

- Draft has been discussed extensively on the list.
- I believe the draft comprehensively documents majority of pertinent issues that will arise in a solution for overload control.
- Adopt as a working group item and refine going forward?