

draft-ietf-soc-overload-design-02

Volker Hilt, Eric Noel,
Charles Shen, Ahmed Abdelal

*(volker.hilt@alcatel-lucent.com, eric.noel@att.com,
charles@cs.columbia.edu, aabdelal@sonusnet.com)*

SOC Interim, Dec 2010

Status & Changes

WGLC completed Aug 22 to Sept 5th

- Thanks to everyone who has provided comments!

Version -02 released

- Addresses the comments raised during WGLC.
- Added numerous clarifications and fixes.
- Changes as discussed in the following slides.

Load not Caused by Processing SIP Messages

A SIP server can be overloaded for reasons that do not involve the processing of SIP messages

- E.g., processing of RTP packets, database queries, software updates and event handling

If the server detects overload, it applies SIP overload control mechanisms to avoid a congestion collapse on the SIP signaling plane.

- May not significantly reduce the load on the server if the resource shortage was created by another service.
- Expected that the server uses appropriate methods of controlling the resource usage of other services.
- Specifics of controlling the resource usage of other services and their coordination is out of scope for this document.

Separation of Directions

In a realistic deployment, SIP messages will flow in both directions, from server B to server A as well as server A to server B.

The overload control mechanisms in each direction can be considered independently.

Server Farms



In some cases, the servers D, E, and F are in a server farm and configured to appear as a single server.

- Server A reports overload on behalf of the server farm.

In cases where A is not a SIP entity:

- Servers D, E, and F can report the overall load of the server farm.
- One of the servers (e.g., server E) can report overload on behalf of the server farm.
 - Not all messages will contain overload control information and it needs to be ensured that all upstream neighbors are periodically served by server E.

Performance Metrics

Added computational complexity as a metric:

- What is the (cpu) load created by the overload "monitor" and "actuator"

Conclusion

All issues raised in WGLC addressed.