SIP Load balancing Charter

IETF81 Dispatch MEETING

Monday, July 25, 2011

Quebec City, Canada

Vijay K. Gurbani

R Parthasarathi

Agenda

- Problem statement
- Current solutions
- SIP LB considerations
- Next steps

Problem statement

- Definition of problem: Distribute SIP requests to a collection of servers to effectively utilize the resources at those servers.
  - Prevent excessive oscillation at the servers (i.e., toggle between on-off state).

Problem statement

- SIP load balancing (LB) is performed without any agreed upon common principle
- Varying SIP server capability and capacity in single load balancing farm call for generic mechanism
- Resource usage varies from (B2BUA) server to (PSTN GW) server.

Problem Statement

- A SIP load balancer may be:
  - SIP-aware (proxy)
  - SIP-unaware (operates on rules derived from source/destination IP address tuples, or use DNS updates)
  - Minimally SIP-aware (may be able to parse enough to get the Call-ID)

Current Solution - 1

- Load balance based on an invariant (Call-ID or H(Call-ID))
  - Assumes all servers of equal capacity
  - Invariant service time
  - No feedback from downstream entity

Current Solution - 2

- Round-robin based solution.
  - Assumes all servers of equal capacity
  - Invariant service time
  - No feedback from downstream entity
- Will work for low traffic arrival rates, but may not at higher traffic arrival rates.

Current Solution - 3

- Round-robin with 503 feedback based solution.
  - Works for a small set of downstream entities; will not scale.
  - May conflate overload control with load balancing.

Current Solution - 4

- DNS SRV based with weights updated dynamically through rfc2136.
  - Will not work if IP addresses are used in SIP URIs (enterprises)
  - Need for a logical entity to collect load information from all servers and updates DNS.

SIP LB consideration

- A closed loop model appears to be beneficial
- Diversity of SIP downstream servers
- Information to be provisioned in Load balancer and in downstream
- In-path or out-path or both?
- How does LB play with overload control?
- Do we need separate solution for signaling servers and media servers?

Split signaling and media LB

- As SIP request resource consumption in SIP signaling only server varies drastically from SIP media servers, should the solution be split such that load balancing of a pure signaling server is different than that of a SIP server that handles signaling as well as media?

Split signaling and media LB

- IMPORTANT: Should we have different deliverables for media and signaling-only servers?
  - Yes. Current charter deliverables reflect this:

    Feb 2013 Submit signaling based SIP load-balancing solution to IESG as Proposed Standard RFC

    Feb 2013 Submit signaling and media based SIP overload solution to IESG as Proposed Standards RFC
  - No. Modify charter to reflect this.

Charter milestones

- Mar 2012 Survey document for SIP load balancing strategies to IESG as an Informational document.
- Jun 2012 Use cases and requirement document to IESG as an Informational document.
- Aug 2012 Design & Architecture to IESG as Informational RFC.
- Feb 2013 Submit signaling based SIP load balancing solution to IESG as Proposed Standard RFC.
- Feb 2013 Submit signaling and media based SIP load balancing solution to IESG as Proposed Standard RFC.

Next steps

- Ready to answer the question on "Where to do this work?"
  - New WG?
  - Existing WG?