SoC
SIP Overload Control based on Resource Availability

IETF-78 MEETING

R.Parthasarathi

on behalf of the team

Draft:draft-partha-soc-overload-resource-availability-00

Authors: R.Parthasarathi, Sheshadri Shalya

Agenda

- Resource Availability (RA) based Overload control requirements
- Resource Availability based overload control mechanism
- Explicit Feedback mechanism using SUB/NOT in Resource Availability
- Resource Availability Package
- Next Steps

Resource Availability Requirements

- Same Device plays the role of different SIP entities simultaneously - SBC, 3PCC B2BUA, PSTN GW, Media Server, transcoder, conferencing server
- Other than CPU & Memory, each SIP entity require resources like DSP, DS0, Disk

SIP Entity	Protocol	Resource
SIP only B2BUA	SIP	CPU, Memory
B2BUA with topology hiding (SBC)	SIP, RTP	CPU, Memory
B2BUA & transcoding	SIP, RTP	CPU, Memory, DSP
B2BUA & PSTN GW	SIP, RTP, ISDN,CAS	CPU, Memory, DSP, DS0
B2BUA with Recording	SIP, RTP	CPU, Memory, Disk

- Overloading any one resource may potentially impact some / all of the SIP Entities – Ex: DS0 impacts only PSTN GW while CPU impacts all.

Resource Availability Requirements

- Each SIP entity provides different services. Each service consumes different quantity of the same resources.
  - For eg.,: B2BUA with address hiding service (SBC), memory requirements vary for different types of call as indicated below:

Call type (1 call)	**Resource (Memory)**
SIP Audio call	0.2%
SIP Video call	1%
SIP Telepresence	10%

  - With the above nature of the call types, it is not possible to have overload control based on the single capacity number for all types of services
  - Enterprise Business and Call Centre deployments need this proposal
- In call centre deployment, Load balancer to SIP Entity shall be in the order of 1:75. Say 75 SIP Entities are overloaded, the single call has to retried in all 75 servers without overload information

Resource Availability (RA) Mechanism

- Informing Resource Availability information to the neighboring SIP entities based on the change in the status of resource availability or in a periodic manner.
- Resources of the system include:
  - CPU
  - Memory
  - DS0
  - DSP
  - Disk
  - Available Bandwidth of the system
- Based on the RA information, the neighboring Entity shall do
  - Intelligent Routing
  - Monitor the Resource Usage
- Helps Administrator to do better planning and design of Network resources and its optimal utilization

Explicit Feedback mechanism using SUB/NOT in RA

- In case Subscriber knows in advance that the required resource is not available in the Notifier, there is no need to send any message till Notifier indicates its available status.
- In SUB/NOT mechanism, Subscriber has to process only NOTIFY with overload information rather than looking into all the message.
- SUB/NOT works well when there is no traffic between two servers
- As it is a separate overload channel, the overload message shall be prioritized using special mechanism like DSCP.
- SUBSCRIBE dialog is linked with network failures seen in other SIP messages by which subscriber will know the health of the Notifier immediately.

Resource Availability Package

- New package urn:ietf:params:xml:ns:rai is introduced
- Schema contains “System” tuple which indicates the overall devices status
- Individual resource information is under the tuple “Resource” which contains the specified resource information.
- Each resource shall have separate unit.

Next Steps

- Is there agreement that a problem exists with overload of call stateful devices (such as GWs, SBCs, ...) that is not currently being addressed by this WG?
- Is there interest in working on that problem in this WG, now or in the future?
- Would more information characterizing this class of problem be of interest?

Discussion