
**Open Community Specification**
***POCS-2***

# Transmission of a Session Capacity Estimate (SCE) to Prevent Session Initiation Protocol (SIP) Server Overload

*October 2011*

## Summary

This memo describes a method of preventing overload of Session Initiation Protocol (SIP) servers through the calculation, conveyance, and usage of a value referred to as the Session Capacity Estimate (SCE), a value that summarily represents the capacity of a SIP server.

## Author(s)

Paul E. Jones <paulej@packetizer.com>
Gonzalo Salgueiro <gsalguei@cisco.com>
Victor Pascual <vpascual@acmepacket.com>

## Acknowledgements

None.

# Table of Contents

# Transmission of a Session Capacity Estimate (SCE) to Prevent Session Initiation Protocol (SIP) Server Overload

## 1 Introduction

Traditionally, reporting resource capacity has entailed reporting information related to discrete resources, such as available DS0s, free memory, or CPU utilization. The problem with this approach is that, more often than not, the only useful information is the number of available DS0s, as that equates to physical components in a system that are clearly a limited resource. Exhaustion of such tangible resources has definite meaning, unlike changes in the reported values of memory or CPU utilization. While there is clearly a point at which memory or CPU utilization will have an effect on the ability of a network element to process communication requests, it is impossible for other entities in the network to effectively use that information directly, since 90% CPU utilization, for example, does not necessarily mean that the device has more or less capacity than a device reporting 50% CPU utilization.

Herewith, we introduce a different method for reporting capacity. This method is one that, at a high level, entails reporting an estimation of the number of additional communication sessions a device can accept. This value may be used by other SIP [2] devices to facilitate the selection of a SIP server to which to direct traffic in an attempt to avoid overloading a server.

## 2 Conventions used in this document

The key words “must”, “must not”, “required”, “shall”, “shall not”, “should”, “should not”, “recommended”, “may”, and “optional” in this document are to be interpreted as described in RFC 2119 [3].

## 3 Session Capacity Estimate

Any element within a communication network has the ability to determine when it is nearing an overloaded state. That determination might be based on, as examples, CPU utilization, memory utilization, bandwidth, or a combination of all of these.

Rather than reporting raw values for discrete elements of the system, a better method for reporting capacity and preventing system overload is to report an estimated value of the number of additional communication sessions a network element can accept. This calculation would be locally computed based on these discrete values and is most useful for network elements that do not have a fixed number of resources, such as SIP proxies or Session Border Controllers (SBCs). We refer to this calculation as the Session Capacity Estimate (SCE).

# 4 Computing the Session Capacity Estimate

This Session Capacity Estimate (SCE) value might be computed using a variety of input variables, usually based on measurements of discrete resources. A SIP server may locally monitor memory utilization, CPU usage, and bandwidth, as examples, and use that data to arrive at an estimate for the number of additional sessions it can accept.

A means of computing the Session Capacity Estimate (SCE) might be to observe how much of any given resource is consumed based on the resources already consumed for the currently active sessions. For example, assume each communication session consumes 5% of available resources, whatever those resources might be. In this case, the maximum number of sessions that could be supported would be 20. With each additional session, the SCE value would be reduced by one.

Most of the time, however, a given session consumes more or less resources than another session and, as such, the SCE value is dynamic in nature. For example, an SBC that processes media may be able to processes fewer video sessions than voice sessions, perhaps due to bandwidth restrictions or perhaps DSP resources required to perform transcoding of media. As each new session is introduced in the system, a new SCE value can be computed that reflects the element's best guess as to how much capacity remains, taking into consideration measurable resources like memory, CPU, bandwidth, etc.

# 5 Conveyance of the Session Capacity Estimate

The SCE value is conveyed through normal SIP signaling exchanges between devices.

The SCE value reported by the SIP entity sending a response upstream is included in the Via header as a new parameter called “sce”. For example, a SIP server might include a Via header like the following to indicate that it is capable of accepting an additional 275 sessions:

```
Via: SIP/2.0/UDP 192.168.1.10:5060;
     branch=z9hG4bK776asdhds;sce=275
```

During periods when no other SIP exchanges take place between two servers, a SIP User Agent might send an OPTIONS “ping” message to a peer entity to determine the status of that peer as per POCS-1 [1]. The response to that OPTIONS message would include the SCE value in addition to other information that might be returned.

# 6 Utilizing the Session Capacity Estimate

A SIP server or SIP User Agent attempting to identify a next hop to which it may direct a session can use the SCE information collected from its peers in order to select a next hop that has the highest likelihood of successfully accepting the sessions. To do this, it merely compares the SCE values reported by its peers.

The actual selection might be based on the highest SCE value reported by peer elements. Alternatively, the server might use a round-robin approach to cycle through peer elements

that report an SCE value at or above some minimum value or above the average value reported by peer entities.

Whatever method is used is a matter for implementation. In any case, the objective is to identify a device that is least likely to return a 503 "Service Unavailable" status code, indicating the peer device has exceeded capacity limits.

# 7 References


[1] Jones, et al., "Using OPTIONS to Query for Operational Status in the Session Initiation Protocol (SIP)", POCS-1, October 2011.

[2] Rosenberg, J., et al, "SIP: Session Initiation Protocol", RFC 3261, June 2002.

[3] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.