Chapter 19
An Architectural and Evaluative Review of Implicit and Explicit SIP Overload Handling

Marco Happenhofer  
Vienna University of Technology, Austria

Joachim Fabini  
Vienna University of Technology, Austria

Christoph Egger  
Vienna University of Technology, Austria

Michael Hirschbichler  
Vienna University of Technology, Austria

ABSTRACT

Last year’s trend to migrate circuit-switched voice networks to packet switched Internet Protocol (IP) based networks has favored wide deployment of Session Initiation Protocol (SIP) based systems and networks. As a reaction to large-scale SIP deployment experiences in the field and the need to implement high availability and reliability within these new networks, the focus of SIP extension standardization has shifted from adding new SIP signaling functionality to operational and maintenance aspects, a particular importance being attributed to overload control. Overload denotes a situation in which the traffic injected into a system exceeds the system’s designed capacity. The authors present a detailed categorization of overload architectures and outline main reasons why SIP-based networks are at high risk to collapse when operating at overload. Using measurements in a real SIP infrastructure this paper compares the performance of two overload protection schemes, namely implicit and explicit overload protection, against the performance of non-protected systems. The measurement results recommend overload protection as a mandatory component of commercial SIP deployments to safeguard operation and prevent system collapse in case of overload.

DOI: 10.4018/978-1-4666-4165-5.ch019
1. INTRODUCTION

The evolution of IP networks, the continuous increase of bandwidth and decrease of latency enable new applications for IP networks, which go far beyond the scope of what IP was initially designed for. One of these new applications is the classical voice service, which was mainly operated by telecom operators in the past. Nowadays we have the technology to operate voice services on top of dedicated IP networks or even, with some limitations, in the public Internet. Particularly the user perceived audio and video quality may suffer under the Internet’s packet switched delay characteristics. Several mechanisms have been introduced to minimize the impact of best-effort transport on media streams, some of them being, e.g., smart codecs or dedicated bandwidth reservation. However the overall perceived quality of the service depends also on other aspects, as for example availability and the response times of session setup. Not being able to establish a media connection in a timely manner over an IP network can be even more frustrating than media quality impairments. Operators must safeguard that their signaling network, the system which is responsible for setting up the voice and video calls, is available at any time and provides almost constant response times. This is a challenging task as under certain circumstances a huge amount of incoming calls might overload the signaling network, for example due to catastrophic incidents like earthquakes or common events like power failures for specific districts. The signaling network typically does not have sufficient spare capacities to process all incoming service requests in such rare disaster scenarios, main reason being that a network dimensioning to handle such overload cases is not economically feasible. It is also highly difficult to predict realistic load requirements for such scenarios, as communication infrastructure or hardware might be affected by the catastrophic incident, resulting in a potentially dramatic load increase on still functional systems and components.

A few decades ago, system designers in the IP domain have faced similar challenges, namely the overload of performance-limited IP routers. This issue has been resolved with the design and deployment of the Transmission Control Protocol (TCP). This protocol senses continuously the round trip time in the IP network. It concludes from lost messages or late message arrivals that an overload condition has occurred and reduces the injected load accordingly. Mainly because of this network-friendly behavior, the TCP/IP protocol has become the most successful transport protocol in the Internet.

For signaling networks that use the Session Initiation Protocol (SIP) (Rosenberg et al., 2002) no TCP-equivalent solution to ensure proper operation in overload conditions has been defined so far. The basic SIP protocol has been standardized in 1999 by the IETF and was initially supposed as developed protocol for voice calls over the Internet. Growing popularity in the Internet community and the adoption of SIP as the main signaling protocol of the 3rd Generation Partnership Project IP Multimedia Subsystem has boosted the wide-spread deployment of SIP. However, the main focus of SIP development is nowadays shifting from functionality extensions, which were required for large-scale deployments, to operation and management extensions like overload control.

The IETF proposed already in 2008 an exhaustive RFC (Rosenberg, 2008) that mentions several conditions under which the current approach with a simple reject message might fail. This was also a trigger for designing an architecture to communicate within the SIP network overload conditions and to protect the system for overload. However, IETF standardization work on SIP overload control is approaching its final stages. Therefore we have implemented a prototype of these overload protection mechanisms and evaluated the effectiveness of the suggested architecture by comparing it against unprotected SIP networks.

The rest of the paper is organized as follows. Section two presents and categorizes overload
Related Content

Measurment of End-User Computing Satisfaction
www.igi-global.com/chapter/measurment-end-user-computing-satisfaction/20245?camid=4v1

Measurements through Rectenna System
www.igi-global.com/article/measurements-through-rectenna-system/109648?camid=4v1a

About Processing of Exponentially Damped Signals: A Hardware for Biomedical Applications
www.igi-global.com/article/processing-exponentially-damped-signals/74698?camid=4v1a

Measuring Internet Dependency
www.igi-global.com/chapter/measuring-internet-dependency/20225?camid=4v1a