Session Initiation Protocol (SIP) Server
Overload Control: Design and Evaluation

Charles Shen¹, Henning Schulzrinne¹, and Erich Nahum²

¹ Columbia University, New York, NY 10027, USA
{charles,hgs}@cs.columbia.edu
² IBM T.J. Watson Research Center, Hawthorne, NY 10532, USA
nahum@watson.ibm.com

Abstract. A Session Initiation Protocol (SIP) server may be overloaded by emergency-induced call volume, “American Idol” style flash crowd effects or denial of service attacks. The SIP server overload problem is interesting especially because the costs of serving or rejecting a SIP session can be similar. For this reason, the built-in SIP overload control mechanism based on generating rejection messages cannot prevent the server from entering congestion collapse under heavy load. The SIP overload problem calls for a pushback control solution in which the potentially overloaded receiving server may notify its upstream sending servers to have them send only the amount of load within the receiving server’s processing capacity. The pushback framework can be achieved by either a rate-based feedback or a window-based feedback. The centerpiece of the feedback mechanism is the algorithm used to generate load regulation information. We propose three new window-based feedback algorithms and evaluate them together with two existing rate-based feedback algorithms. We compare the different algorithms in terms of the number of tuning parameters and performance under both steady and variable load. Furthermore, we identify two categories of fairness requirements for SIP overload control, namely, user-centric and provider-centric fairness. With the introduction of a new double-feed SIP overload control architecture, we show how the algorithms can meet those fairness criteria.

1 Introduction

The Session Initiation Protocol [1] (SIP) is a signaling protocol standardized by IETF for creating, modifying, and terminating sessions in the Internet. It has been used for many session-oriented applications, such as calls, multimedia distributions, video conferencing, presence service and instant messaging. Major standards bodies including 3GPP, ITU-I, and ETSI have all adopted SIP as the core signaling protocol for Next Generation Networks predominately based on the Internet Multimedia Subsystem (IMS) architecture.

The widespread popularity of SIP has raised attention to its readiness of handling overload [2]. A SIP server can be overloaded for many reasons, such as emergency-induced call volume, flash crowds generated by TV programs (e.g.,
American Idol), special events such as “free tickets to third caller”, or even denial of service attacks. Although server overload is by no means a new problem for the Internet, the key observation that distinguishes the SIP overload problem from others is that the cost of rejecting a SIP session usually cannot be ignored compared to the cost of serving a session. Consequently, when a SIP server has to reject a large amount of arriving sessions, its performance collapses. This explains why using the built-in SIP overload control mechanism based on generating a rejection response messages does not solve the problem. If, as is often recommended, the rejected sessions are sent to a load-sharing SIP server, the alternative server will soon also be generating nothing but rejection responses, leading to a cascading failure. Another important aspect of overload in SIP is related to SIP’s multi-hop server architecture with name-based application level routing. This aspect creates the so-called “server to server” overload problem that is generally not comparable to overload in other servers such as web server.

To avoid the overloaded server ending up at a state spending all its resources rejecting sessions, Hilt et al. [3] outlined a SIP overload control framework based on feedback from the receiving server to its upstream sending servers. The feedback can be in terms of a rate or a load limiting window size. However, the exact algorithms that may be applied in this framework and the potential performance implications are not obvious. In particular, to our best knowledge there has been no published work on specific window-based algorithms for SIP overload control, or comprehensive performance evaluation of rate-based feedback algorithms that also discusses dynamic load conditions and overload control fairness issues.

In this paper, we introduce a new dynamic session estimation scheme which plays an essential role in applying selected control algorithms to the SIP overload environment. We then propose three new window-based algorithms for SIP overload. We also apply two existing load adaption algorithms for rate-based overload control. We thus cover all three types of feedback control mechanisms in [3]: the absolute rate feedback, relative rate feedback and window feedback. Our simulation evaluation results show that although the algorithms differ in their tuning parameters, most of them are able to achieve theoretical maximum performance under steady state load conditions. The results under dynamic load conditions with source arrival and departure are also encouraging. Furthermore, we look at the fairness issue in the context of SIP overload and propose the notion of user-centric fairness vs. service provider-centric fairness. We show how different algorithms may achieve the desired type of fairness. In particular, we found that the user-centric fairness is difficult to achieve in the absolute rate or window-based feedback mechanisms. We solve this problem by introducing a new double-feed SIP overload control architecture.

The rest of this paper is organized as follows: Section 2 presents background on the SIP overload problem, and discusses related work. In Section 3, we propose three window-based SIP overload control algorithms and describe two existing load adaptation algorithm to be applied for rate-based SIP overload control. Then we present the simulation model and basic SIP overload results without feedback control in Section 4. The steady load performance evaluation of the