Scaling Internet Telephony Signaling

Henning Schulzrinne, Wenyu Jiang, Xiatao Wu
Columbia University
{hgs,}@cs.columbia.edu

December 18, 1998

Abstract

Scalability is a primary concern when considering replacing traditional, SS7-based systems with IP-based signaling. This note compares some aspects of signaling performance, focusing on performance of servers implementing the Session Initiation Protocol (SIP).

1 Introduction

Internet-based system may well largely replace traditional circuit-switched telephony in the next decade. However, achieving this goal requires that Internet-based solutions can offer existing services with at least the same performance and reliability as the more traditional technology. In short, Internet telephony has to be able to "scale" to the size of the telephone network.

The problem of scaling can be usefully divided into issues of packet transmission, resource control, network management and signaling. Scaling of packet transmission depends on the availability of routers with a sufficient number of high-speed interfaces and the ability of these routers to forward packets, including small voice packets, at line rate, with QOS-based scheduling if needed. Scaling of resource control depends on whether the resource reservation mechanisms, operating either on a per-flow or aggregate basis, and the associated policy control protocols can handle the number of concurrently active flows. Scaling of the network management system requires that the number of carriers can manage the multitude of devices as well as monitor and aggregate per-flow performance measurements.

This paper focuses on the problems associated with scaling signaling to the scale of a current local or long-distance network.

2 Scope of Problem

In order to assess the scope of the scaling problems, it might be helpful to gain some understanding of the call volume that is handled by the current telephone network.

2.1 Call Volume

According to the FCC [**?**], terminating minutes are the best measure of long-distance conversation minutes. [1] In the second quarter of 1998, there were 81.1 billion terminating minutes in the United State. Assuming an average call duration of 2.1 minutes, this equals 10,300 long distance calls/second for all long-distance carriers.

2.1.1 Call Duration

An FCC report [**?**, Table 2.6] indicates that 55.5 billion interstate toll calls generated 467.7 billion access minutes in 1997, for an estimated average call duration of 4.2 minutes. (This estimate is rough, as it assumes that each conversation minute generates two access minutes. However, in 1997, terminating minutes were only about 63% of access minutes.) Similarly, 23.3 billion intrastate calls generated 178.2 billion access minutes, for an estimated average duration of 3.82 minutes per call.

[1] "1 800 and 888 calls only generate originating access minutes, ... WATS calls generate only terminating access minutes. International calls only generate originating access minutes when exiting the United States and only terminating access minutes when entering the United States. Access minutes are not measured for calls made on private telecommunications systems and on leased lines. Terminating minutes do not equal conversation minutes because terminating minutes for international calls exiting the United States are not counted and terminating minutes for some personal 800 numbers are counted twice." [**?**] In other words, access minutes = 2 * long-distance + 800; terminating minutes = WATS + incoming international + long-distance; originating minutes = long distance + outgoing international + 800 + 2 * personal 800 minutes.

The estimate of average call duration can also be derived from the following calculation. The *AT&T Factbook* [?] notes that AT&T completed 68 billion calls in 1996. According to [?], it terminated 142.1 billion minutes during that year, for an average call duration of 2.1 minutes. However, this number is apparently off by a significant factor, as the AT&T market share is at 46.3% of terminating minutes, while the call count indicates a value of 21%. It is also likely that the call volume is inflated by in-state (inter-LATA) calls carried by AT&T. This appears likely given that per-line interstate and intrastate calls are roughly equal. Taking the market share into account, we also arrive at roughly four minutes per call.

2.2 Signaling Traffic

We estimate the total number of interstate calls by dividing the terminating minutes in 1997 (316.5 billion) [?] by the estimated call duration, yielding 2509 calls/second.

For local calls, the call duration estimate is probably too low, but given the total number of 1997 dial equipment minutes (two for each conversation minute) of 2,407 billion, we obtain a local call rate of 19,081 calls/second.

Clearly, calls are not distributed evenly across time; for example, AT&T notes that it completed 230 million calls on an average *business* day in 1996 (2662 calls/second), while the average daily volume is only 186 million calls (2152 calls/second).

As an order of magnitude, let us assume that half of all interstate calls are 800-number calls. Thus, the signaling system needs to accomodate roughly 1200 calls/second.

The signaling load depends on how much mapping effort is required for each call. At one extreme, if we implement full number portability, every phone call requires a database lookup. At the other, only 800-number calls require servers in the network. (We will discuss the PBX case later.)

We assume that each SIP packet requires 100 bytes, including 32 bytes of IP and UDP overhead. We also assume that each call requires one number mapping via a proxy server and thus generates a total of six messages. Thus, for all 800-number traffic, all servers combined would see a total incoming and outgoing traffic volume of 1200 messages or 5.76 Mb/s.

2.3 Switch Performance

Busy hours generate roughly 5 to 6 times the call load compared to the daily average [2]. As a rule of thumb, the busy hour rates need to be multiplied by a factor of 2 or 3 to arrive at peak calling loads [3].

A 4ESS switch has a capacity of 1.4 million BHCA (busy hour call attempts), equivalent to 390 call attempts/s. The older 1A processor had a capacity of 650,000 BHCA, equivalent to 180 call attempts/s.

A typical call requires 6 ISUP messages per call, with about 80 octets of signaling load.

2.4 Server Performance

Server performance

Tcl INVITE being written to socket and the 3XX RESPONSE being totally read from the socket. 10 times totally cost 1.270255s. The cost for the the program used for getting the time is about 0.03s each time.

The running time of INVITE response is still 0.1s. For 100 response, the cost is about 0.003 0.02s The time between 100 response and 3XX response is 0.1s 0.15s. The possible cause is fork and exec on aliases (0.04s) and namemap.

3 Packet Efficiency

References

[1] J. Zolnierek, K. Rangos, and J. Eisner, "Long distance market shares second quarter 1998," Report MKSH2Q98, Industry Analysis Division of the Common Carrier Bureau, Federal Communications Commission, Washington, D.C., Sept. 1998.

[2] AT&T, "At&t fact book," 1997. http://www.att.com/factbook.

[2] Mike Hluchyj, personal communications

[3] Alan Leon Varney, personal communications