InterNetNews: Usenet transport for Internet sites

Rich Salz – Open Software Foundation

ABSTRACT

NNTP, the Network News Transfer Protocol, has been labelled the most widely implemented elective protocol in the Internet. The growth of the Internet has meant more sites exchanging NNTP data. While the explosive growth in Usenet traffic places demands on all sites, the goal of fast network access puts particular demands on NNTP hosts.

InterNetNews is an implementation of the Usenet transport layer designed to address this situation. It replaces the standard UNIX server architecture with a single long-running server that handles all incoming connections. It has proven to be quite successful, providing quick and efficient news transfer.

Introduction

Usenet is a distributed bulletin board system, built as a logical network on top of other networks and connections. By design, messages resemble standard Internet electronic mail messages as defined in RFC822 [Crocker82]. The Usenet message format is described in RFC1036 [Adams87]. This defines some additional headers. It also limits the values of some of the standard headers as well as giving some of them special semantics.

Newsgroups are the classification system of Usenet. The required Newsgroups header specifies where a message, or article, should be filed upon reception. Sites are free to carry whatever groups they want. Most sites carry the core set of so-called ‘‘mainstream’’ groups. There are currently about 730 of these groups, and one or two new ones is created every week.

Messages generated at a site are sent to the site’s ‘‘neighbors’’ who process them and relay them to their neighbors, and so on. Sites can be interconnected – indeed, on the Internet, this is quite common. See Figure 1.

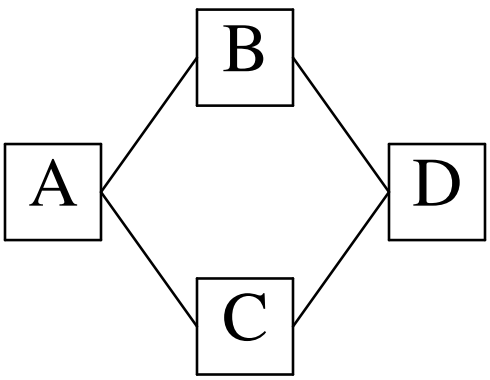

Figure 1: Small Usenet topology (all links are two-way).

The Path header is used to prevent message loops. For example, an article written at *A* could get sent to *B*, *D*, *C*, and then back to *A*. Before propagating an article, a site prepends its own name to the Path header. Before propagating an article to a site, the receiving host checks to make sure that the site that would receive the article does not appear in the Path line. For example, when the article arrived at site *C*, the Path would contain *A!B!D*, so site *C* would know not to send the article to *A*.

Sites also keep a record of the Message-ID’s of all articles they currently have. If *D* receives an article from *B*, it will reject the article if *C* offers it later. For self-protection, most sites keep a record of recent articles that they no longer have. This is very useful when another site dumps a (usually quite large) batch of old news back out to Usenet.

For the past few years, the amount of data generated by Usenet sites has been doubling every year. A site that receives all the mainstream groups is receiving over 17 megabytes a day spread out over 11,000 articles [Adams92]. About 20% of the data is article headers, and while all of them must be scanned only half of it is must be processed by the Usenet software.[1]

The number of sites participating in Usenet has been growing almost as quickly. Based on articles his site receives and survey data sent in by participating sites, Brian Reid