ISSN 2229-5518

Methods Used to Handle Overloading of Information in Usenet

Deepika Saxena, Monika Saxena

Abstract— Usenet is the name of a worldwide network of servers for group communication between people. From 1979 and onwards, it has seen a near exponential growth in the amount of data transported, which has been a strain on bandwidth and storage. There has been a wide range of academic research with focus on the WWW, but Usenet has been neglected. Instead, Usenet's evolution has been dominated by practical solutions. This paper describes the history of Usenet in a growth perspective, and introduces methods for collection and analysis of statistical data for testing the usefulness of various caching strategies. A set of different caching strategies are proposed and examined in light of bandwidth and storage demands as well as user perceived performance.the advanced caching methods for news offers relief for reading servers' storage and bandwidth capacity by exploiting usage patterns for fetching or prefetching articles .i have shown the problems occurs in this type of methods with little bit solutions.users may want to read, but it will not solve the problem of near exponential growth nor the problems of Usenet'sbackbone peers.

INTRODUCTION

Usenet was created in 1979 . Since, Usenet is the name of a worldwide network of servers for group communication between people. It has seen an impressive rowth from a small academic community to a network used by millions of people from a wide variety of backgrounds all over the world. The total size of the data flowing through Usenet has been more than tripling every year between 1993 and 2001.This growth has not been without problems, and has raised significant challenges in how to handle the ever-increasing volume of Usenet data flow. Very few are able to handle all of Usenet, and as the amount of users and data they produce increase, as do the challenges with having enough network bandwidth and storage capacity. Spending great sums of money on hardware components relieves the situation, but it does not solve it. My motivation for this thesis was to find a way to reduce the problems we see today. I have introduced the idea of advanced caching methods as a general improvement for parts of the Usenet distribution network, as well as discussed other work that has been done to relieve network bandwidth and storage capacity. I also introduce methods for analyzing and evaluating caching strategies based on statistical data from news servers.

I first provide an introduction to Usenet architecture and technology, followed by Usenet's history from the perspective of growth and the challenges of this growth, as well as a brief mention of some other trends and suggestions for dealing with the volume of Usenet data traffic. I present advanced caching strategies that may help handling these challenges. I then pose questions about what 1 way these methods may improve on Usenet.To my knowledge; there are no peer-reviewed sources for the growth of Usenet in a historical perspective or for caching of news in particular. The development of Usenet technology has been a community effort rather than an academic one, and many of the conventions and standards have been informal at first to be standardized later. I have attempted to structure and word the paper for an audience and people that is not familiar with Usenet, its historical background, how it used to work, what the protocols are, or how it works today. It is an advantage to have some familiarity with the Internet, the WWW, e-mail and networks. Readers familiar with how Usenet works, its history of growth, and the problems arising from. Definitions that I introduce are marked clearly, while Usenet specific terminology is explained as it is used with the terms *emphasized*.

What Is Usenet?

News is a distributed platform for group communication mainly between human based on a network of servers all around the world."Usenet" is an Abbreviation for "Unix User Network", but is also known under other names, specifically "NetNews", simply "News" or "Usenet News". News is a slightly misleading name for what Usenet is meant for: asynchronous communication between people, as opposed to news items distributed by mass media.

Usenet defines the following way:

Usenet is the set of people who exchange articles tagged with one or more universally recognized labels, called "newsgroups" (or "groups" for short). There is often confusion about the precise set of newsgroups that constitute Usenet; one commonly accepted definition is that it consists of newsgroups listed in the periodic "List of Active Newsgroups" postings, which appear regularly in news.lists.misc and other newsgroups. An even broader definition includes even newsgroups that are restricted to specific geographic regions or organizations.Each Usenet site makes its own decisions about the set of groups available to its users; this set differs from site to site.

The communication between users is largely controlled by local administrators of the news service the *news administrators* at a news service provider (NSP). An NSP can also be a full Internet Service Provider (ISP).
While Usenet is today mostly a part of the Internet, using the same basic network protocols for communicaion between servers, it has been common to say that “Usenet is not the Internet”. The reason for this is that the transport of news itself is not fundamentally dependent on the Internet; it just is the most used platform today.
There is much more to Usenet than I mention in this chapter, which is intended as an introduction and overview of what I consider relevant for understanding
this paper. Some parts have been simplified in order to avoid too much excruciating detail.

The Usenet Model

The Usenet News model has the following major aspects to consider:
_ Message format
_ Message distribution
_ Message storage
The main flow of Usenet is commonly through the Internet, using the Network News Transfer Protocol (NNTP) , a TCP[1] based protocol for transmission. Most Internet standards are described in RFCs[2], and the IETF[3] is working on several new standards. Usenet’s standards are described in RFCs, but there are de facto Usenet standards not included in the RFCs, although the IETF is working on standardising these enhancements.

1.2.1 Message Format

Message format are logically divided into two separate parts, *head* (also called *headers*) and *body*. The headers contain meta-information about the article,such as who allegedly posted the article, from where, at what time, to which

```
From:Jan           Ingvoldstad           <ja-
ni+news@tsathoggua.rlyeh.net>
Subject: How do newsadmins deal with news
traffic today?
Newsgroups: news.software.nntp
Date: 30 Apr 2001 13:03:01 +0000
Message-ID:                          <ygtlmoiw-
nei.fsf@tsathoggua.rlyeh.net>
Sender: jani@tsathoggua.rlyeh.net
Path:
nntp.uio.no!uio.no!news.tele.dk!148.122.208.6
81!
news2.oke.nextra.no!nextra.com!news.klingenbe
rg.no!
tsathoggua.rlyeh.net!not-for-mail
```

Some example news headers

newsgroups, with what subject of discussion, a unique message ID, and the path through which servers the article has been passed to avoid re-relaying to those servers. Other headers may be used, but these are not relevant here; I will discuss some of these when necessary.The article’s body contains the actual message, which must be plain text, including quotations of former articles in the same discussion. casually the author adds a *signature*, which contains information about the author, a quip,a quote from a book or movie, or all of these at the same time. This signature is considered part of the body.**A Note on the Path Header** The Path header has a syntax from before the DNS[4] was created, and each news server identifier is separated by a “bang”– ’!’. This ID is either a name registered in the UUCP[5] maps or, since the introduction of DNS, the full DNS name of the server. The identifier must only be in place for relaying servers where the article passed as a news article, so if it passes via e.g. an e-mail server, there should be no entry for that. The last entry is not considered part of the path entry, and is in the case of a user agent normally the *local part* of an e-mail address, and the “not for-mail” entry is there in case it is difficult or impractical to supply that local part. With this last exception, it is supposed to be possible to send an e-mail to each entry in the path list, plus the local part after the path list.

Message Distribution and Storage

While the news article format is compliant with the Internet mail message format, news distribution is significantly different from mail distribution.

Many mailreaders are also newsreaders[6], which causes some initial confusion for users on this issue.News articles are commonly spread by a flooding algorithm between news servers, also known as *news feedrs/feeding servers* or *peers*. Where each *downstream peer* gets a *newsfeed* of articles from their *upstream peer*. This is called a "pushed" stream, similar to the "push" technology used for WWW The receiving servers reject articles they already have instead of requesting the ones they do not have. This is called a *pull stream*, like clients pulling documents off the WWW. Note that it is possible for the downstream peer to request articles from their upstream peer, but it is not commonly used.

So far, this is deceptively similar to Internet mail, which is also can be sent from server to server until finally received by the user's mailbox, although the current practice is to send e-mail directly to the server local to the receiving user. However, users do not get this feed directly in their own mailbox, as would be the case with Internet mail and mailing lists. Instead, their newsreader fetches a list of newsgroups and articles from the news server, using NNRP[7]. This kind of news server is called a *reader/reading server*. I will refer to this function as *reading server* from now on, in order to avoid confusion between human reader, newsreader program and reader server.The user then chooses which newsgroups to read articles from from his newsreader's *subscription list*. This list of *subscribed* newsgroups is updated by the user. When the user has chosen a newsgroup, he then can choose which articles to read within that newsgroup.**Note on Built-in Filtering in Newsreaders** Many newsreaders offer filtering methods based on patterns in article headers and body in the form of a so-called *kill file*. If an article matches this pattern, the newsreader will not download the article, and if it is already downloaded, it will not display it.

Some newsreaders offer additional functionality in form of a *score file*. This is also a kind of filter, but unlike a kill file, the choice is not black or white. *Scoring* is more flexible, and allows the user to set positive or negative values for various patterns. These values are cumulative. In addition to setting values for patterns, the user specifies a score threshold for which articles should be displayed. This way, it is possible to e.g. ignore certain authors, unless they post with a subject the user finds more interesting (high score for the subject pattern) than he finds the author uninteresting or annoying (low score for the author pattern). Articles are stored on these central news servers, making them shared as opposed to mailing lists, where each user effectively stores his own message copy in his mailbox. This does not prevent the user from downloading and storing his own copy.

Another way to explain the distribution of news from peer to peer, is to compare it to the message transfer system (MTS) of relaying mail transfer agents (MTA) in OSI's message handling system (MHS, from the X.400 recommendations . This is close enough to Internet e-mail in how it works that the comparison makes sense; Internet e-mail only issues receipts upon failed delivery, and then to the sender of the message. Messages are stored and forwarded for each node on the path from the sending UA to the receiving UA in both models. The important difference is that for Usenet news, messages are not distributed directly to the end users; they have to request them from their local reading server.

As opposed to e-mail, is that news is not a reliable medium of transport for messages. News was not designed for reliability, and there are control mechanisms that allow people to remove their own articles after they were posted.It is possible for one reading server to offer articles within one same newsgroup that another news server does not, yet responses to these articles may show up on both servers. This will typically happen if an article is attempted sent from one of the servers to the other, and the other does not respond or accept it before a predefined timeout at the first server. Where e-mail via SMTP usually will generate a response to the message sender if the message could not be delivered, news offers no such service. This is good for the users, whose mailboxes would be overflowing with such responses if there should be one generate for each of the news servers that could not receive it, considering that there are tens of thousands of news servers the article may have been attempted distributed to.

Even, it does not show that peers do not transport exactly the same amount of articles everywhere to everyone. What really happens is that each of the news administrators has made agreements with one or several news administrators about which newsgroups or hierarchies they will distribute between themselves. Some of these transport as much data as they can get a full newsfeed and can be considered part of a Usenet "backbone". Others transport other amounts of data. In addition to these differences in newsfeed size, these peers do not necessarily connect with the closest other peer. These issues are attempted visualized in the fairly complex . The real Usenet distribution network is far more complex.

Information Structure

Articles are organized in *newsgroups* (discussion groups), similar to mailing lists in that they each have a name and a particular topic of discussion. In difference from mailing lists, newsgroups are organized in named hierarchies. It is possible for an article to be posted to several newsgroups simultaneously; this is called *crossposting*. The newsgroup names are on the form:

Hierarchy What

comp	-Computers
humanities	-Arts and humanities
misc	-Miscellaneous
news	-Usenet
rec	-Recreational
sci	-Science
soc	-Social/Sociology
talk	-General discussions

The names bear some significance to what the topic of discussion on that particular group is, both in that it has influence on what is to be discussed there, and in that it shows what actually is discussed. In addition to its name, most groups have a brief description stored at the reading server.

hierarchy *alt*, which is more "free" in how groups are created and organized, is also considered part of the core by many users and news administrators.

There are also national and local hierarchies that not necessarily followthis organization scheme for choice of top level names, but use similar schemes for their own subhierarchies.

Handling the Challenges

Compared to Usenet, documents on the web live a long time. The web site Deja[19], later bought by Google[20] and renamed Google Groups[21], have attempted to store all news articles, with the exception of most binaries, for eternity. They have failed in that they do not have all news articles for the time period they are covering. Some of these are missing because authors have reserved themselves against being stored by use of the optional **X-No-Archive** header, which Google honors by not storing these articles. It is not uncommon for regular news servers to not get all articles that are posted to Usenet,but it is regrettable that those who set out to store and provide "everything" are unable to do so. Note that Google Groups does not try to store binary articles, which makes their task more manageable. The National Library of Norway preserves articles posted to the *no* hierarchy as a part of Norway's cultural heritage.

A news article cannot be changed once it is posted, but it can be cancelled and replaced by other articles, or simply be expired (deleted) because the news server attempts to conserve storage space. Such removal of articles happen all the time, since news administrators want to limit the use of storage space, and partially because there are automated utility news programs which cancel spam .To handle the ever-increasing traffic on Usenet, WWW for a few years, this has not been the case for Usenet .

DEFINITION 1 (PROXY)

In the context of Usenet, a proxy is an intermediate server that transparently to user agents or downstream peers provides articles that it itself doesnot have, but are available from one of the proxy's upstream peers.

DEFINITION 2 (CACHING)

For Usenet, caching means copying and storing incoming data, and keepingthat data for a period of time.

In terms of usability, flow and group control.

DEFINITION 3 (PREFETCHING)

Fetching data from an upstream peer before it is requested by user agentsor downstream peers.

It is useful to note that Usenet's flooding algorithm can be viewed as a time based prefetching caching mechanism, in that everybody gets a recently posted article as soon as possible after it is posted, and that it is then only up to the leaf nodes the reading servers to decide how many of these are available to their users.

The History and Development of Usenet

In 1999, Usenet News turned 20 years. In those 20 years, many things have changed, but some underlying principles have remained. When BBSes (Bulletin Board Systems) were very popular, many people expressed that Usenet was just another BBS.Where BBSes (with few exceptions) were limited to single computers and people connected with their modems (or whatever means they had) to post their messages and discuss with others of like or different mind, Usenet was from the beginning a distributed system, where messages were transmitted between different computers to be available from more servers. Usenet was probably best compared with a network of BBSes, each carrying the same discussions.In 1999, Usenet News turned 20 years. In those 20 years, many things havechanged, but some underlying principles have remained. When BBSes (Bulletin Board Systems) were very popular, many people expressed that Usenet was just another BBS.Where BBSes (with few exceptions) were limited to single computers and people connected with their

modems (or whatever means they had) to post their messages and discuss with others of like or different mind, Usenet was from the beginning a distributed system,where messages were transmitted between different computers to be available from more servers. Usenet was probably best compared with a network of BBSes, each carrying the same discussions.

The Beginning of Usenet

The birth of Usenet is linked to a single event: An operating system upgrade rendered existing bulletin board software non-functional, which caused two graduate students at Duke University in North Carolina, Tom Truscott and Jim Ellis, to develop the idea of a distributed news system. This was in the fall of 1979 [Hauben and Hauben, 1995].At first, Usenet was a substitute for a broken bulletin board system, an experiment with UUCP, based on a 3-page Unix shell script. The script allowed people to subscribe to different groups, post and read notes in sequence, and also post to different groups at the same time (crossposting) [Hauben and Hauben, 1995].Steve Bellovin, one of the people who Truscott and Ellis presented their design to, wrote the shell script using Unix V7 to test the design concept. The first Usenet was a two-server setup, but it evolved quickly.

Conclusion

I have presented the history of Usenet from a growth perspective, and shown that there are technical problems with the its continued growth. Smaller sites cannot afford to offer their users all the newsgroups they might want to read, and the problem seems to be growing. While other solutions than caching such as filtering greatly reduce the size of a full newsfeed, they are rigid and do not adapt the incoming flow depending on usage, as caching will. The world wide web has used various caching methods for years, and a lot of work and research has been done to optimize caching for the web.However, nobody has worked with solutions for news.My proposed advanced caching methods for Usenet will help the smaller sites to appear to offer a greater amount of newsgroups and articles, but does not address the problem of the seemingly exponential growth. However, even a linear reduction in newsfeed size will buy the news administrators time to postpone the next hardware upgrade, which means they will save money.

References:

Cand Scient Thesis 4th August 2001
Handling Information Overload on Usenet Advanced Caching Methods for News - Jan Ingvoldstad
[Assange et al., 2001] Assange, J., Bowker, L., and nntpcache crew, T. (2001).
What is NNTPCache? http://www.nntpcache.org/about.html.
[Barber, 2000] Barber, S. (2000). RFC 2980: Common NNTP extensions.RFC.
[Barber, 2001] Barber, S. (2001). Network news transport protocol. Internet Draft.
[Bumgarner, 1995] Bumgarner, L. S. (1995). USENET — The Great Renaming — 1985–1988. http://www.vrx.net/usenet/history/rename.html.
[Cidera Inc., 2001] Cidera Inc. (2001). Cidera usenet news service.
http://www.cidera.com/services/usenet_news/index.shtml.
[Collyer, 1992] Collyer, G. (1992). newsoverview - netnews overview files. newsoverview(5) man page.
[Crocker, 1982] Crocker, D. H. (1982). RFC 822: Standard for the Format of ARPA Internet Text Messages. RFC.
[Danzig, 1998] Danzig, P. (1998). Netcache architecture and deployment.
Computer Networks and ISDN Systems, 30:2081–2091.
[Delany and Herbert, 1993] Delany, M. and Herbert, A. (1993). gup - a group
update program. ftp.mira.net.au:/unix/news/gup-0.4.tar.gz.
[Freed and Borenstein, 1996a] Freed, N. and Borenstein, N. (1996a). RFC
2045: Multipurpose Internet Mail Extensions (MIME) Part One: Format of
Internet Message Bodies. RFC.
[Freed and Borenstein, 1996b] Freed, N. and Borenstein, N. (1996b). RFC 2046: Multipurpose Internet Mail Extensions (MIME) Part Two: Media
Types. RFC.[Freed and Borenstein, 1996c] Freed, N. and Borenstein, N. (1996c). RFC 2049: Multipurpose Internet Mail Extensions (MIME) Part Five:
Conformance Criteria and Examples. RFC.
87 [Freed et al., 1996] Freed, N., Klensin, J., and Postel, J. (1996). RFC 2048:Multipurpose Internet Mail Extensions (MIME) Part Four: Registration
Procedures. RFC. [Freenix, 2001] Freenix (2001). Top 1000 Usenet sites.http://www.top1000.org.
[Hardy, 1993] Hardy, H. E. (1993). The Usenet System. *ITCA Yearbook*.
[Hauben and Hauben, 1995] Hauben, R. and Hauben, M. (1995). On the Early Days of Usenet: The Roots of the Cooperative Online Culture.

http://www.columbia.edu/~rh120/ch106.x10.
[Horton and Adams, 1987] Horton, M. and Adams, R. (1987). RFC 1036: Standard for USENET Messages. RFC.
[Ingvoldstad, 1998] Ingvoldstad, J. (1998). Usenet news som plattform for effektiv gruppekommunikasjon. Essay written in Norwegian as a part of
studies at Ifi.
[Inktomi Corporation, 2000] Inktomi Corporation (2000). NNTP Caching for Usenet Services.
http://www.inktomi.com/products/network/traffic/tech/nntp/index.html.
[Kantor and Lapsley, 1986] Kantor, B. and Lapsley, P. (1986). RFC 977:
Network News Transfer Protocol — A Proposed standard for the Stream-Based Transmission of News. RFC.
[Kondou, 2001] Kondou, K. (2001). Daily Usenet Article statistics on newsfeed.mesh.ad.jp. http://newsfeed.mesh.ad.jp/flow/.
[Konstan et al., 1997] Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L.,
Gordon, L. R., and Reidl, J. (1997). GroupLens: Applying Collaborative Filtering to Usenet News. *Communications of the ACM*, 40(3):77–87.
[Krasel, 2001] Krasel, C. (2001). Leafnode, an NNTP server for small sites.
http://www.leafnode.org.