USENET

David Conrad

drc@isc.org

Internet Software Consortium

Overview

- USENET Introduction and Theory
- History of USENET
- USENET Structure and Operation
- USENET Issues
- Summary

Why Should You Care?

- USENET News is typically provided as a matter of course by Internet service providers
  - A check-box item
- USENET "push" model of content transmission is still useful
  - as the proliferation of "groupware" would demonstrate
- USENET can be very resource intensive
  - Bandwidth, hardware, management personnel
- USENET articles can get you into trouble

USENET Introduction and Theory

- USENET is
  - A content transport system
    - Like electronic mail, only different
  - A logical network layered on top of other networks
  - A broadcast, one-to-many medium
- Derived from very early Unix networking technology
  - "Unix to Unix Copy (UUCP)"
- Internet USENET hosts normally runs its own protocol (NNTP) over TCP/IP, but can also use UUCP over TCP
  - UUCP over TCP useful in **very** bad network conditions

Digression #1: UUCP

- UUCP -- Unix to Unix Copy
  - Actually a suite of programs to facilitate transfer of files from one machine to another machine over a network
    - Either a dialup network (my machine calls yours) or an Internet(-like) network
  - Important commands:
    - uux -- execute a command on another system
    - uucp -- queue a file for copying
    - uucico -- copy in/copy out queued files
    - uusched -- the scheduler for UUCP commands
- See Unix manual pages for more information

Digression #2: UUCP Addressing

- UUCP Addressing is position-relative
  - The address varies depending on where you are in the network
  - Uses a path concept to trace route from originating machine to destination
    - `inn.isc.org!usenet.dec.com!usenet.sony.com!user`
      - originator is `user@usenet.sony.com`
      - message got to `inn.isc.org` via `usenet.dec.org`
    - implies very little flexibility if any of the machines in a path are broken
- USENET still uses UUCP addressing in places

Short History of USENET

- First started at Duke University in USA in late 1970’s
  - Conceptually, similar to posting a note on a subject specific bulletin board
- First software was called “A News Software”
  - “B News” and then “C News” soon followed
    - Both B and C News still found on the Internet today
- Originally, USENET consisted of two sets of bulletin boards, mod.* and net.*
  - mod.* was moderated, net.* wasn’t

History (cont’d)

- In mid-1980’s Network News Transport Protocol (NNTP) was developed
  - An application layer protocol using TCP
  - Internet Network News (INN) and other TCP/IP based news servers followed
- In 1986, the Great USENET Renaming occurred
  - Splitting mod.* and net.* into “the big 8”
- With the explosion of the Internet since the early 1990s, traffic has grown from a few megabytes per day to many gigabytes per day
  - Unfortunately, the signal to noise ratio is pretty poor
    - Very few sites carry full newsfeeds anymore

USENET Structure and Operation

- A distributed Bulletin Board System
  - Take your message and “post” it on the BBS
- Users post messages (“articles”) to areas called “newsgroups”
  - Newsgroups have themes or topics
  - Each article is given
    - a site relative article number
    - a globally unique message identifier
- Articles can be posted to multiple newsgroups at one time
  - Frowned upon, but common
- Articles are copied to other USENET news servers
  - All servers willing to accept the article on the entire Internet

USENET Structure & Operation (cont’d)

- No central control or authority
  - Anyone can create and post a news article
- New newsgroups can be created by anyone
  - simply post a specially formatted article called a “control message”
    - Control messages are easily forged
      - Can be cryptographically signed using PGP
- Local policy determines how long articles are kept in storage

USENET Structure & Operation (cont’d)

- Newsgroups are hierarchical
  - `comp.` -- articles related to computers
  - `comp.protocols` -- articles related to computer (networking) protocols
  - `comp.protocols.tcp-ip` -- articles related to TCP/IP networking
  - `comp.protocols.tcp-ip.dns` -- articles related to the DNS (which uses TCP/IP and allows computers to talk to each other)

Newsgroups

- Newsgroup hierarchies vary wildly
  - addition/deletion of newsgroups in "the big eight" hierarchies controlled by the "USENET Cabal"
    - The "big eight" are `comp, humanities, misc, news, rec, sci, soc, & talk`
      - carried by most news servers
  - The "alt" hierarchy established because some people didn't like the USENET Cabal
  - Other hierarchies are "private" but propagated
    - e.g., news hierarchies for a corporation's products
      - e.g., `microsoft.public.*`
- Acceptance of a particular hierarchy is a local policy decision

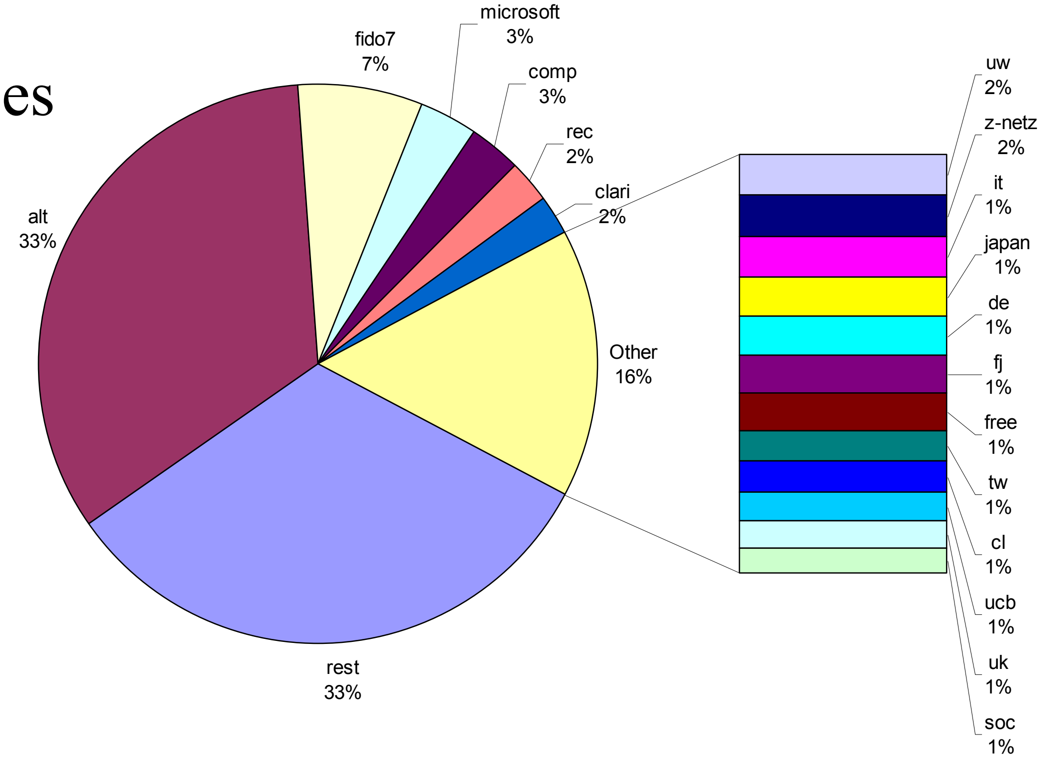
Newsgroups (cont'd)

- Newsgroups can be moderated or unmoderated
  - Articles posted to a moderated newsgroup must have an `approved` header field
    - The moderator is supposed to be the one to do this
      - Easy to forge the appropriate magic to get past this check
    - Moderators are volunteers interested in the subject of the newsgroup
  - Most newsgroups are not moderated
    - However, moderated ones usually have better signal to noise ratios

Newsgroups (cont'd)

- Currently there are
  - 460 newsgroup hierarchies
  - 28,948 newsgroups
- Top hierarchies are:

– alt	33.97%
– fido7	7.21%
– microsoft	3.34%
– comp	3.31%
– rec	2.49%
– clari	2.09%



Flood Fill Article Propagation

- USENET articles are propagated using“flood fill”
  - Each USENET news server has one or more peers
  - Each article received from a peer or from a user of that server (a posting) is sent to all other peers that haven’t yet seen the article
  - A “push” model of data transmission
- The full article is copied all over the Internet
  - when using real-time feeders, the article can reach all major news hosts on the Internet in a matter of minutes

Article Format

- Plain 7-bit ASCII text
  - non-ASCII encoded into ASCII
    - typically using *uuencode*
    - MIME encoding becoming more and more popular
- Resembles an email message
  - News article format a subset of e-mail (RFC 822) format
- Described in RFC 1036
  - "Son of RFC 1036" is in progress
- Content of articles moving more and more to HTML

- Has 6 required headers
  - `From` -- who wrote the article
  - `Date` -- the date the article was posted
  - `Newsgroups` -- the newsgroup(s) the article was posted to
  - `Subject` -- the subject of the article
  - `Message-ID` -- a globally unique identifier for the article
  - `Path` -- the UUCP path the article has take to reach the current system
- Other headers optional
  - Unknown headers passed unchanged

A USENET Article

```
Path: papaya.bbn.com!rsalz
From: rsalz@bbn.com (Rich Salz)
Newsgroups: news.software.nntp,news.admin,comp.org.usenix
Subject: Seeking beta-testers for a new NNTP transfer system
Message-ID: <3632@litchi.bbn.com>
Date: 18 Jun 91 15:47:21 GMT
Followup-To: poster
Organization: Bolt, Beranek and Newman, Inc.
Lines: 72
Xref: papaya.bbn.com news.software.nntp:1550 news.admin:15565
comp.org.usenix:418

InterNetNews, or INN, is a news transport system.  The core part of the
package is a single long-running daemon that handles all incoming NNTP
connections.  It files the articles and arranges for them to be forwarded
to downstream sites.  Because it is long-running, it can be directed to
spawn other long-running processes, telling them exactly when an article
should be sent to a feed.
<…>
/r$
```

The `Path` Field

- When a server receives an article, it adds its own name to the front of the `Path`, e.g.:

  An article with a `path` of:

  `Path: usenet.dec.com!usenet.sony.com!user`

  would be modified to

  `Path:` **`inn.isc.org!`**`usenet.dec.com!usenet.sony.com!user`

  when it is sent from usenet.dec.com to inn.isc.org

- Before sending an article to a peer, the news server checks the `Path` to see if the peer is already listed
  - Stops loops

Control Message Propagation

- Control messages come in several flavors
  - `Cancel` removes a previously posted article
  - `Newgroup` creates a newsgroup
  - `Rmgroup` removes a newsgroup
  - `Checkgroups` asks the server to check its list of newsgroups against an official list
  - `Sendsys` request a copy of the configuration describing the server's peers
  - `Version` request information about the type and version of the software being run

Control Messages (cont’d)

- Only `Cancel`, `Newgroup`, and `Rmgroup` are in common usage now
  - `Checkgroups`, `Sendsys`, and `Version` considered security risks
- `Cancel` control messages are by far the most common
  - And the most frequently forged
- `Newgroup` and `Rmgroup` are important to track
  - Should not be blindly executed
    - Use PGP header verification if possible

NNTP and Its Use

- NNTP is a simple application layer protocol
  - “Standard” verb/numeric response code format
  - Described in RFC 977
- Mostly a command/response protocol
  - One server sends “I have article <number>” to peer
  - Peer sends “no thanks, seen it already” or “OK, send it”

Internet Network News

- ISC’s INN is an open source USENET news system
  - Available from ftp://ftp.isc.org/isc/inn/inn.tar.gz
- INN is a transport system
  - Will use an appropriate application layer transport mechanism
    - NNTP (by preference)
    - UUCP
    - even SMTP
  - Can also handle compressed batches of news
  - Can be extended easily to handle other transport mechanisms as needed

History of INN

- Created in the early 1990’s
  - Originally written by Rich Salz
    - First beta release June 18, 1991
  - Current version 2.2
    - Released January 21, 1999
- Was the first real-time News transporter
  - C news used the NNTP reference implementation, but incoming articles were put into batch files for later processing

What INN Does

- Transport news articles
- Implements NNTP (RFC 977)
- Primarily uses TCP/IP
  - Can use UUCP or other transport mechanisms
- Provides network client (reader) interface
- Feeds in real-time or in batch mode
  - Compressed or uncompressed articles

What INN Doesn’t Do

- No client software (news readers)
  - Gobs of news readers exist
    - Old style: rn, trn, vnews,
    - New style: Netscape Communicator, Microsoft Explorer
- No extra support for large-scale “reverse” (sucking) feeds
  - “Pull” model instead of “push”
- No web interfaces for users or administrators (yet)
  - Management of INN is a painful
- INN is middleware and not a vertical solution
  - Vertical solutions such as Netscape’s Collabra exist

Types of News Servers

- Transit servers
  - Usually at enterprise gateways
  - Have no regular reader clients
  - Don’t keep articles around for long
  - Less resource requirements than readers
  - Easier to secure

- Reader Servers
  - Require significantly more resources than transit servers
  - Require more management resources
  - Usually stores articles for long periods
  - Targets for spammers

Caching NNTP Servers

- Provides some level of scalability
  - Reduced resource requirements, higher performance
- When a reader requests articles, the caching server first checks local storage and (if article isn’t found) requests the article from an upstream server using NNTP reader commands
  - Upstream server treats the request like any other reader request
- Articles typically fetched on demand, but large numbers of articles can be pre-fetched

Caching NNTP Servers (cont’d)

- Lets the site running the caching server avoid accepting a full feed
  - Full feeds demand large amounts of disk space
- Useful for sites with inconsistent or sparse reading patterns
- Not a good idea for sites with poor network connections
  - Reader performance affected by upstream server

Futures

- “Groupware” such as Lotus Notes and Netscape Collabra are the next evolutionary of USENET News
  - Very pretty user interfaces on the news reading clients
  - Much more easily managed servers
  - Tighter integration of transport / user interface / article store
    - Includes database retrieval mechanisms for article content
- USENET messages will likely become more HTML rich
  - Newsreaders unable to handle HTML will likely fade away
- USENET will continue to evolve

USENET Evolution

- Current USENET technology results in tremendous resource utilization
  - Disk, network, CPU, management, etc.
- Gigabytes / day of messages
  - Typically, only a tiny percentage of these messages are ever read
  - Large percentage of messages are spam

USENET Evolution (Cont’d)

- USENET articles will likely move to a header/pointer format
  - Content only fetched if article is read
  - Gateways to “old” USENET that fetch the content a create/post a “legacy” article
- Likely permits a reduction in the amount of resources consumed
  - Can be aided by integration with WWW caches
- Can help in the reduction of spam
  - Integration with tools like MAPS/RBL

USENET Issues

- As with any service which provides content, “inappropriate” content can be found
  - Hateful literature, pornography, libel/slander
  - There are constant calls to censor this content
    - ISPs often get caught in the middle
      - Easy targets
      - Little control
  - Technological advances may “help” content control issues
- USENET growth will continue to be an issue
  - New technology many help this as well

Summary

- USENET has been around since the beginning of the Internet
- News is still useful for pushing information to a wide audience
    - A flood fill model of information propagation assures global distribution
- USENET News hierarchy is largely chaotic
- USENET articles are similar to mail is format
- USENET will likely evolve to a header/pointer format
    - Will reduce the resource requirements and (hopefully) help the signal to noise ratio

Where to Get More Information

- RFC 1036 -- Standard for Interchange of USENET Messages
  - http://www.isi.edu/in-notes/rfc1036.txt
- RFC 977 -- Network News Transfer Protocol
  - http://www.isi.edu/in-notes/rfc1036.txt
- Henry Spencer & David Lawrence, *Managing Usenet*, 1st Edition January 1998, O'Reilly & Associates
- Internet Network News (INN)
  - http://www.isc.org/inn.html