

INTERNET-DRAFT  
Expires: September 1998

Carsten Bormann  
Universitaet Bremen TZI  
March 1998

Network News Distribution Protocol: Architecture and Design Guidelines  
draft-bormann-mnnp-nndp-00.txt

## Status of this memo

This document is an Internet-Draft. Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as ``work in progress.''

To learn the current status of any Internet-Draft, please check the ``1id-abstracts.txt' listing contained in the Internet-Drafts Shadow Directories on ftp.is.co.za (Africa), nic.nordu.net (Europe), munnari.oz.au (Pacific Rim), ds.internic.net (US East Coast), or ftp.isi.edu (US West Coast).

Distribution of this document is unlimited.

## Abstract

This document describes an architecture and a set of protocols for distributing Netnews [RFC0977, [RFC1036](#)] via IP multicast enabled networks. The architecture is designed to be useful in the global Internet. In particular, it allows multiple news servers to cooperate on multicasting each new article only once. To facilitate scalability to tens of thousands of news servers, it also provides for receive-only multicast participants (that continue to send articles via conventional NNTP).

This document is a submission to the IETF MNMP working group. Comments are solicited and should be addressed to the working groups' mailing list at [ietf-mnnp@va.pubnix.com](mailto:ietf-mnnp@va.pubnix.com) and/or the author.

## [1.](#) Introduction

Netnews (or Usenet news) is one of the more important systems for electronic communication that make up what is now loosely called ``the Internet' in the media. Usenet operates by flood-distributing

messages called articles between participating systems, called news servers. The Usenet is experiencing growth problems as with any other element of the thriving Internet environment.

It is widely recognized that NNTP, the article distribution system in use in the Usenet, is running into scaling problems. Some ISPs are reporting numbers of between 7 and 12 % for the NNTP contribution to their backbone traffic -- this for a data stream that is less than 64 kbit/s in total (see below).

As Usenet is fundamentally a multicasting system, an obvious approach is to apply the emerging Internet network layer multicasting technology to Usenet distribution. One experiment described in the literature, MUSE [firehose paper], transmitted Usenet articles as UDP multicast packets between participating sites. While this experiment was moderately successful, it suffered from packet loss problems (that increase exponentially with the number of fragments generated from one article). Also, a scalable security architecture was not defined for this experiment.

This document defines an architecture and sketches two protocols to make network layer multicasting more useful for news distribution. The architecture will, in reference to an earlier experiment [newscaster] be called Newscaster-2 or simply Newscaster; the two protocols will be called NNDP (Network News Distribution Protocol) and NNDCP (Network News Distribution Coordination Protocol), respectively.

### 1.1. Benefits of multicasting Netnews

Distributing Netnews via network layer multicast provides a number of benefits. For ISPs, Newscaster can help to significantly reduce the backbone NNTP load: Each article traverses each link (in the best case) only once instead of traversing the backbone links multiple times, once to each target news server.

One other benefit of Newscaster will be reduced article propagation times -- while current NNTP servers can be very fast, Newscaster replaces multiple unicast hops between news servers by a single multicast hop. As propagation times currently measure on the order of hours, a reduction to the order of minutes would be a nice achievement; a reduction below that (to seconds) is, however, not intended. (As a side benefit, Newscaster will reduce the link bandwidth consumed by a leaf news receiver by using batching and compression and by reducing the NNTP/TCP/IP overhead incurred per

article.)

## 1.2. Basic Assumptions

This document makes a number of assumptions about the basic technical parameters of the Netnews system. We assume a total number of new news articles to be distributed per day in the few hundred thousands, i.e., one to a few articles per second. We also assume that the total volume of those articles is on the order of hundreds of megabytes per day, i.e., tens to a few hundreds of kbit/s. Newscaster-2 is scalable beyond those numbers, but not infinitely so. [In particular, ``similar'' problems with different technical

Bormann

[Page 2]

---

INTERNET-DRAFT NNDP: Architecture and Design Guidelines

March 1998

parameters (such as live stock price feeds) are not necessarily supported as efficiently as the actual worldwide Netnews system; solving such similar problems is explicitly a non-goal of the architecture.]

In addition, we assume that the concept of News servers that receive a full feed of news articles continues to be useful. On-demand retrieval of news articles from neighboring servers is an interesting concept but outside the scope. We believe that most News servers will want to receive most of the articles in the Netnews system; Newscaster does not support elaborate mechanisms to receive a specific subset of articles that cover exactly the newsgroups that are ``subscribed'' by a News server. (Newscaster does support partitioning the global news-feed into a few general subsets, such as alt.\* and comp.\*/sci.\*.)

One very important point in the design of a multicast Netnews distribution system is that, even if it takes off quickly, News server administrators will not simply turn off their existing, well-understood and robust system of NNTP feeds. To make a feature out of what could be considered a bug, the Newscaster system is intended to work with and be supplementary to the NNTP system. Newscaster-based news servers continue to speak NNTP to neighboring systems, using NNTP as a background scheme to fill in articles that it might have missed in the multicast distribution. Therefore, Newscaster can be a much more light-weight protocol as it needs not be 100 % reliable.

## 1.3. The multiple-entry problem

Given that Newscaster is not replacing, but supplementing NNTP, and that the Newscaster system will for a long time be only a subset of the global Netnews system, the two distribution mechanisms need to

cooperate. The most significant problem here is that a single news article may be flood-distributed from its source via NNTP and reach multiple Newscaster systems at about the same time (observations in the live network show that this now often happens for multiple well-connected news servers within a second). As, in a multicast scenario, there is no way to ask all the receivers whether they already have received an article, this, without further mechanisms, would mean that Newscasters regularly send multiple redundant copies of a single article.

This document proposes a coordination protocol between Newscaster systems to decide which Newscaster system distributes a particular article. The coordination protocol is separate from the distribution protocol; receive-only sites need not be involved in the coordination protocol. Note that correctness of the coordination protocol is not a prerequisite to correctness of the overall system, only to its efficiency, i.e., an occasional slip (multiple transmission of one article) is tolerable.

## [2.](#) The Newscaster Architecture

### [2.1.](#) Protocols

Newscaster assumes an underlying IP multicast network such as the experimental Mbone and/or the operational IP multicast networks being deployed by many ISPs. The multicast network is assumed to be able to sustain a rate-controlled low-bandwidth stream of packets for extended periods; the only form of congestion control envisaged is that receivers can drop out if they experience consistent congestion.

To achieve a degree of performance in the presence of losses in the experimental Mbone, some form of error control is required. To achieve good scalability without router support, the distribution protocol only uses forward error correction; as news servers gain multicast connectivity, they simply can start listening to the feed without having to send any (unicast or multicast) data.

The coordination protocol does not need to be as scalable as the distribution protocol: It will be hard to impossible to coordinate between a few tens of thousand news servers, and various features of the distribution protocol (batching, compression, digital signatures)

argue for limiting the number of active Newscaster servers. We assume that new articles travel via NNTP to the nearest active Newscaster system and are multicast from there to the rest of the world.

[Appendix A](#) defines a preliminary coordination protocol based on a multicast transport protocol called MTP-2. (This protocol is a version of MTP ([RFC1301](#)) that was developed further to be more useful in WANs. It allows multicasting a sequence of arbitrary size messages, each of which can consist of one or more multicast packets. The MTP-2 protocol provides a global sequencing of the messages, as well as global rate control.)

Other coordination protocols may be defined. Passive, receive-only Newscaster systems need not be aware of the coordination protocol being used -- they only need to understand the distribution protocol. In particular, the distribution protocol can be used from a single source to a local (e.g., per-ISP) set of receivers; the coordination protocol then becomes trivial.

## [2.2.](#) Operation of active Newscasters

A news server actively participating in the Newscaster system is simply called a Newscaster. The set of cooperating Newscasters is called the Newscaster Web. The entire Web is a single news system from the point of view of [RFC1036](#) Path headers. For the global Newscaster Web, the name of the news system as it occurs in the Path header is "newscaster-2.mcast.net". Additional local Newscaster Webs can be created, if needed, under different names.

Each Newscaster examines each article it receives via NNTP or other means whether it already contains a Newscaster Path header entry and immediately removes it from further consideration in the Newscaster Web if this is the case (in the INN implementation of the Netnews protocols, this is done automatically if the outgoing link is identified by the Web name, e.g. "newscaster-2.mcast.net").

Those articles that do not contain a Newscaster Path header entry are then prepared for being multicast into the Web. Several such articles will generally be sent together as a batch. The coordination protocol is used to decide, for each article, whether it is actually this Newscaster which will distribute the article. At the service interface, an implementation of a coordination protocol receives a set of message-ids (a tentative batch) as input and

returns a (possibly empty) subset of the message-ids to be sent in an actual batch. In general, each Newscaster should have only one set of articles in progress with the coordination protocol at any point in time. Further articles arriving during processing by the coordination protocol should be collected for a future tentative batch. Also, Newscasters should wait a few seconds for further articles to arrive before submitting a new batch to the coordination protocol.

Actual batches are then formed out of the articles selected according to [RFC 1036, section 4.3](#). They are then compressed using the gzip format ([RFC1952](#)) and digitally signed (see below). Finally, they are distributed using the distribution protocol.

### [2.3](#). Security

Any system that transports Netnews must provide some basic security against spoofing attacks. Since the multicasting system itself provides only very limited assurances that a source address is correct, we resort to cryptographic measures.

Simple shared-secret authentication is not scalable -- in a production version, thousands of News server administrators would have to be in possession of the key. Instead, a public key system is used, based on a web-of-trust security policy.

In the current NNTP system, each news server administrator trusts its neighbor news server administrators to institute a good local usage policy and to respond to incidents in a manner that helps to preserve the integrity of the news system. The transitive closure of this web of trust equals the actual connectivity of the news system. If a news administrator misbehaves, he runs the risk of being disconnected.

The Newscaster security policy attempts to mimic this existing policy by cryptographic means. Instead of creating NNTP links to ``neighboring'' systems, a news administrator creates certificates for all the Newscasters that she trusts. These certificates are regularly distributed in a newsgroup that is reserved for this

purpose (such as, news.config.newscaster), ensuring they can be received even by sites that are not yet in possession of all the certificates. Every receive-only system has to trust one or more sites (e.g., the Newscaster equivalent of a ``well-connected site'') to root its certificate chain. If a receiver of a Newscaster batch

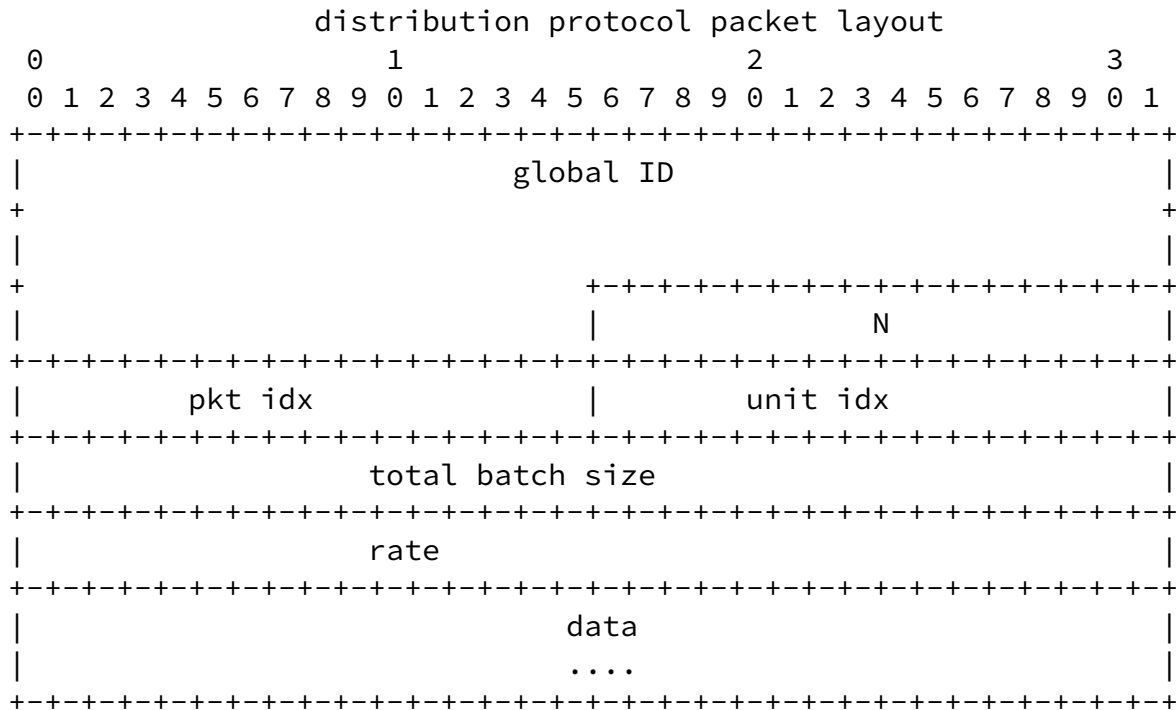
does not find a certificate chain that verifies the signature of the batch, it discards the batch.

\* Issue \*: What type of key system and digital signature is used? Newscaster should provide relatively fast signature checking with modest, but (due to batching) not necessarily stellar signing performance. The author would tend to use [RFC1991](#) type (PGP) formats, using RSA and MD5.

### 3. NNDP: The distribution protocol

The NNDP distribution protocol is used to distribute payloads to all receivers. Payloads will generally be small to a few dozen kilobytes, but may be much larger in case a large article needs to be transferred. The job of the distribution protocol is to:

- partition the payload into packets that can be multicast without being fragmented on the way. We assume an Internet-wide MTU of 1280 (based on the IPv6 MTU) and save 80 bytes for header overhead (IP, UDP, other), leaving 1200 bytes for the distribution protocol data.
- add forward error correction. We use Vandermonde matrices as implemented by Luigi Rizzo [<http://www.iet.unipi.it/~luigi/vdm.tgz>]. The amount of error correction to be added is a system parameter: For small batches, we always add at least one FEC packet. For larger batches, the FEC overhead is defined by a constant expansion factor. (This factor could be chosen to match the TCP equation at the rate intended.) For very large batches, the batch is split into units which are independently subjected to FEC (packets from all units of a batch are interleaved to spread out the transmission).
- multicast the data at a defined rate (leaky bucket model). It is the job of the coordination protocol to assign a rate to each batch to be sent. (The rate should be relatively low to space out the packets, allowing FEC to work around burst losses.)
- enable reassembly/erasure processing at the receiver. The batches are tagged by a unique, 80-bit global ID, which is assigned by the coordination protocol (e.g., global source ID/sequence number). (Note that reassembly errors are not catastrophic, as an incorrectly reassembled batch will be rejected at signature check.) Each packet carries a total batch size, a unit number within the batch, a packet number within the unit, and the number of packets to be sent per unit (N).



(For a discussion of the rate parameter, see NNDP below.)

\* Issue \*: What is a good unit size? E.g., 128 KB? Should we actually use the TCP equivalence equation to compute an expansion factor from the rate?

#### [4.](#) Acknowledgments

This document has been prompted by the discussions in the MNNP BOF at the Washington IETF. In particular, the author would like to thank Joe Malcolm for the thought-provoking discussions at this IETF.

#### [5.](#) References

TBD

#### [6.](#) Addresses

##### [6.1.](#) Working Group

[The MNNP working group is in creation.]

##### [6.2.](#) Author's address



Carsten Bormann  
Universitaet Bremen FB3 TZI  
Postfach 330440  
D-28334 Bremen, GERMANY  
cabo@tzi.org  
phone +49.421.218-7024  
fax +49.421.218-7000

## 7. Annex A: MTP-2 based coordination protocol

When a batch is being prepared, a short MTP-2 message (an announcement) is sent that just contains the message IDs of the articles in the batch. When this message has been transmitted in the MTP-2 Web and all lower-numbered messages have arrived, the Newscaster removes those articles from the batch that have been announced in lower-numbered announcements. This, in the steady state case, makes it unlikely that two Newscasters will be transmitting the same article concurrently. However, Newscasters that return after a multicast outage would start to transmit old articles (that they have received via NNTP while other systems got them via Newscaster). To minimize the impact of such late-comers on the Newscast efficiency, Newscasters only newscast articles they have newly received while being active in the Web (i.e., no spooling).

For IPv4, the global ID of a batch is composed of the concatenation of the IP address of the MTP-2 master at the time of receiving the announcement and the 24-bit MTP-2 sequence number, filled with zeroes at the end.

Rate control is performed in the following way: Each Newscaster is aware of the total system rate defined for the Web (e.g., 128 kbit/s). Newscasters that are transmitting batches share this bandwidth by setting up short-term reservations. Each Newscaster also maintains a running idea of all the reservations currently in effect. Upon reception of an announcement, the receiving newscaster

considers half the unreserved system rate to be reserved for the announcer. This reservation is corrected by the actual rate used by the sender, once an NNDP packet is received for this batch (rate field). The sender of a batch is allowed to use up to half of what it considers to be the unreserved rate at the time it receives its own announcement for this batch. Each Newscaster deletes a reservation for a batch once the sender should have stopped sending data, according to its actual chosen rate and the size of the batch as indicated in the NNDP packets, or (if no NNDP packets were received at all), after a timeout of T\_SEND (T\_SEND is initially set to 15 seconds). Newscasters avoid using silly rates (i.e., less than a very small fraction of the system rate for a large batch).

## [8.](#) Annex B: Newscasters: Active vs. Passive

Given that there are tens of thousands of news servers in operation, and that NNDCP is intended to work between maybe a thousand active Newscasters, the question immediately comes to mind which news servers should be active Newscasters and which should only listen to the global Netnews distribution. In essence, this is of course a judgment call, which may be guided by:

- Multicast connectivity. An active Newscaster obviously needs to be able to source multicast traffic, not just receive it. Given the current tendency of ISPs to charge extra for multicast sourcing, many news servers may not want to become active Newscasters.
- Path lengths. While the Newscaster architecture takes out many hops from the Netnews distribution paths, an article needs to traverse NNTP hops up to the first active Newscaster before it can be efficiently multicast to the rest of the world. Often, a (topological) region will want to maintain at least one active Newscaster to minimize those path lengths.
- Maintaining the web of trust. Maintainers of active Newscasters need to actively work on maintaining their position in the web of trust that is used as the security foundation of Newscaster.

