9.5

USING NEWS SERVER TECHNOLOGY FOR DATA DELIVERY

Anne Wilson*
Unidata Program Center, Boulder, Colorado

1. INTRODUCTION

For many years, Unidata has assisted the atmospheric community with data access and related tools. We are known for, among other things, providing data in near real time using the LDM (Local Data Manger). The LDM is the driver for our IDD (Internet Data Distribution) network, which has provided data at no cost in near real time to universities and research labs for almost ten years. LDM is a push based package, meaning that data is relayed as soon as it arrives. It constantly streams data to local sites using a manually established routing topology. At the time the LDM was created, the Internet was young and data volumes were significantly smaller than they are now. Sites could receive data and archive them themselves. This is the approach still employed by many sites.

2. E-SCIENCE

However, with advances in sensor technology, data storage, and CPU performance, we are seeing data unprecedented in volume and quality. Satellite data alone is expected to increase by two orders of magnitude in the next ten years. With such increases in data volume the job of managing and using the data becomes harder. Bandwidth and storage requirements make it impractical to stream all data everywhere all the time. Yet, essential requirements persist. For example, in the atmospheric community, it remains critical that high quality data be available on demand in near real time, as in the case of severe weather forecasting.

Fortunately, bandwidth has grown exponentially over the past ten years, so that geographically remote sites can now be "networkologically" close. Indeed, computer scientists are researching "distributed virtual computers" in which the central architectural element is optical networking, not computers, creating networks faster than the computers attached to them [Smarr, 2003]. Similarly, storage technology has improved and prices have dropped dramatically, making it possible for multi-terabyte and even petabyte (1,000TB) data sets to exist.

At the same time, advanced computing technologies are emerging such as improved transport protocols [Falk, 2003], distributed system middleware such as Grid services [Foster, 2003], and distributed file storage methods [Plank, 2003]. Increasing use of metadata and sophisticated subsetting methods such as data mining allow improved use of the data.

These developments provide the foundation for "e-science", network-centric, large-scale science "carried out through global collaborations, requiring access to very large data collections, very large-scale computing resources, and high-performance visualization" [Newman, 2003].

3. FUTURE DATA DELIVERY REQUIREMENTS

Unidata is a participant in Linked Environments for Atmospheric Discovery (LEAD), a cyberinfrastructure for mesoscale meteorology research and education [Droegemeier, 2003]. In order to meet LEAD goals, in addition to streaming data to local sites and facilitating local management, data delivery must meet the following requirements:

- Routing must be robust and highly flexible. Data delivery must support both long term, static routes and dynamic temporary routes, and everything in between.
- Access to both real time and retrospective data is required.
- Through data integration and modeling, LEAD users will be able to generate new data types to be shared with others. Thus, data delivery mechanisms must be able to handle the dynamic introduction of new data types.

**Corresponding author address:* Anne Wilson, Unidata/UCAR, PO Box 3000, Boulder, CO 80307;e-mail <anne@unidata.ucar.edu>

- Sites must be able to dynamically change the subscriptions that define which data products they wish to receive.

- The cost must be minimal.

4. USENET AND NNTP

Usenet (for "users network"), also known as Netnews, is a network existing for the interchange of just about any kind of information. Although originally text based, now the majority of news volume is binary, such as images or music.

The first Usenet messages were exchanged in 1979. NNTP (Network News Transport Protocol), the protocol upon which the interchange is based, was developed in 1986. Although web technology has provided new options for communication, Usenet remains extremely popular. Currently there are over 100,000 newsgroups, and, although the exact number is impossible to determine, it is thought that there are hundreds of thousands of sites that relay news. A recent search shows one site reporting at times over 800 gigabytes per day [Berkeley Daily Usenet Report]. NNTP data transfers are the largest application by volume on Internet2, significantly larger than HTTP. Usenet has proved reliable and robust in the face of vast, freewheeling, heterogeneous anarchy.

NNTP provides the following features that are relevant to data delivery:

- NNTP routing relies on the "flooding algorithm" in which sites are highly interconnected. Articles "flow" to sites using massive redundancy such that an article will reach a site by the fastest route possible at the moment.

- Under NNTP articles are classified using a virtually unlimited number of hierarchically structured newsgroups. Articles can be cross posted to more than one newsgroup, providing multiple views of an article.

- NNTP supports pull based article retrieval, so that clients can connect to a server and retrieve articles of interest on demand as long as they are available at the server.

- NNTP also supports "control" messages, messages that may initiate processing at a remote site, depending on how that remote site is configured. This provides a limited degree of network-level configuration. For example, control messages are used to inform sites about additions and deletions to newsgroup hierarchies.

4. INN (Internet News)

Unidata has been testing data relay using INN, a popular, open source implementation of NNTP. INN first appeared in 1992 [Salz, 1992] and has been under development ever since. INN features pertinent to data delivery include

- local management of articles by filing them or sending them to other programs,

- both batch and streaming transmission,

- handling of peer outages by maintaining backlogs,

- dynamic creation and destruction of connections to peers based on relay volume,

- multiple spooling methods that can be configured in any combination to address a variety of goals, such as a combination of short and long term storage,

- an overview mechanism: a database describing current holdings based on header fields, plus a method to query the format of the metadata,

- authentication and PGP verification.

6. NLDM

The project using INN to relay data is called NLDM (NNTP based LDM). The first step in NLDM evaluation was to ensure that data can be streamed and managed locally as easily and as well as our current LDM. Toward that end, since January we have been using INN to relay the CONDUIT data stream from Boulder to Washington, D.C. The CONDUIT data stream is by far the largest data stream in the IDD, consisting of at times just under three gigabytes of data per hour. Being model output from NCEP, this stream shows large bursts of volume as models are disseminated. The experimental results have been positive with good product

latencies. Recently level II radar data was added to the flow, and it is anticipated that the rest of the IDD volume will be incorporated shortly.

No changes have been made to the INN software to relay data. However, a statistical reporting package has been built on top of INN. NLDM statistics are gathered every second, reported every five minutes, and show

- number of products received per second
- number of bytes received per second
- average and maximum product latencies
- percentage of products sent that were received
- cumulative latencies, i.e., latencies sorted into bins
- number of connections over time
- distributions of paths taken by products

These statistics are reported for different window and bin size combinations ranging from a five minute window with a one second bin size to a two day window with a ten minute bin size.

7. JNLDM

Another part of the effort to relay data using the NNTP protocol is JNLDM, a Java version of software to receive and manage products sent via INN or any other NNTP based implementation. JNLDM is intended for sites with minimal computing and system administration resources. Although intended as a "receive-only" client, JNLDM does have relay out capability in order to report statistics

8. NLDM EXPERIMENTAL RESULTS

We are using NLDM and JNLDM to build a prototype network in order to test data relay more fully. Currently the NLDM network consists of

- A Linux box in Boulder running NLDM that ingests data
- A Linux box in Boulder running JNLDM that receives data
- Two Windows boxes in Boulder running JNLDM that receive data
- A Solaris box in Washington, D.C., that receives data

The above are all computers maintained by Unidata. However, we also peer with a Usenet site at the University of Oregon[1], one of the top 25 sites in the world in terms of volume. This site peers with other Usenet participants, thus our data is being relayed around Usenet and around the world.

Peering with a Usenet site and observing the behavior of the automated routing provided by the NNTP flooding algorithm is very interesting. Although the vast majority of products reach the Washington, D.C. destination via a direct path from Boulder, there are infrequent but not uncommon times when products travel beyond the University of Oregon, out to Usenet, then back to Washington. This means that at those points in time that was the fastest path available.

An interesting, although unplanned demonstration of the flooding algorithm occurred on the night of the 18th of September when hurricane Isabel struck Washington, D.C. At that time the NLDM host there lost Internet2 connectivity for about eighteen hours, although it maintained connectivity to the commodity internet. During this difficulty, the statistics clearly showed that the percentage of products taking the direct path dropped by 20% to 50%, while the other available paths rose in frequency. Most significantly, there was no observable change in product latencies.

9. NEXT STEPS

We plan to continue to build the prototype network testing both NLDM and JNLDM. The following sites have volunteered to be our first beta sites: Texas A&M University, the University of Illinois, the University of Iowa, and the University of Wisconsin-Madison. We expect to be relaying the entire IDD volume between these sites in the near future and will likely solicit additional beta sites after that. The results from these experiments will be reported at the conference in January.

Although the flooding algorithm minimizes sensitivity to network congestion and site failure, routing is not completely "hands free" in that when a site joins the network peers must be chosen. We will choose two to five peers for each site, the

[1] Special thanks to Dr. Joe St Sauver at the University of Oregon for assistance, support and good cheer.

number depending on the resources and bandwidth available to that site. Peers will initially be chosen based on geographic proximity. Later, if fast routes show up in the site's path statistics, we can modify its set of peers accordingly.

We also plan to investigate improvement of the routing efficiency. The flooding algorithm achieves robust routing through massive redundancy. It may be that this routing can be made more efficient by applying techniques learned in the area of application level multicasting [Rew, 2004], which would automate the process of peer selection and updating in such a way that bandwidth usage is improved.

10. CONCLUSION

News server technology is showing promise for meeting future data delivery needs. In particular, NDLM provides the following:

- The flooding algorithm is very robust and flexible in that products seek the fastest route available and are automatically routed around failed sites.
- It is possible that control messages can be used to reconfigure peering so that routes can be dynamically changed.
- In addition to streaming data, pull based article retrieval allows access to retrospective data.
- New data types can be dynamically introduced by creating new newsgroups on the fly and using control messages to announce them.
- Control messages (or something similar) can be relayed to change subscription lists at remote sites.
- Being freely available, INN can be acquired at no cost.

8. REFERENCES

Berkeley Daily Usenet Report, Berkeley Daily Usenet Report.

Droegemeier, K., Brewster, K., Weber, D., Xue, M., Chandrasekar, V., Graves, S., Ramachandran, R., Rushing, J., Wilhelmson, R., Reed, D., Joseph, E., Morris, V., Clark, R., Yalda, S., Gannon, D., Plale, B., Ramamurthy, M., Domenico, B., Murray, D., and Wilson, A., 2003: Linked Environments for Atmospheric Discover (LEAD): A Cyberinfrastructure for Mesoscale Meteorology Research and Education, *20th International Conference on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, Phoenix, AZ, January 11 - 15, 2004.

Falk, A., Faber, T., Bannister, J., Chien, A., Grossman, R,, Leigh, J., Transport Protocols for High Performance, *Communications of the ACM*, Vol. 46, No. 11, November, 2003, p 43.

Foster, I., and Grossman, R., Data Integration in A Bandwidth-Rich World, *Communications of the ACM*, Vol. 46, No. 11, November, 2003, p 51.

Newman, H., Ellisman, M., Orcutt, J., Data-Intensive E-Science Frontier Research, *Communications of the ACM*, Vol. 46, No. 11, November, 2003, p 69.

Plank, J., Atchley, S., Ding, Y., Beck, M. Algorithms for High Performance Wide-Area Distributed File Downloads, *19th ACM Symposium on Operating Systems Principles*, Bolton Landing, New York, October 19 - 22, 2003.

Rew, R. K., Fishing for Data, Pier to Pier, *20th International Conference on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, Phoenix, AZ, January 11 - 15, 2004.

Salz, R., InterNetNews: Usenet Transport for Internet Sites, *Proceedings of the USENIX Summer 1992 Technical Conference*, San Antonio, TX, June, 1992.

Smarr, L., Chien, A., DeFanti, T., Leigh, J., Papadopoulos, P., The Optiputer, *Communications of the ACM*, Vol. 46, No. 11, p. 59, November, 2003.