9.6 TOWARD REAL-TIME COLLECTION OF INTERNET DATA DISTRIBUTION STATISTICS FOR AUTOMATED TOPOLOGY AND PERFORMANCE MONITORING

Steven R. Chiswell *

Unidata/UCAR, Boulder, Colorado

1. INTRODUCTION

The Unidata Internet Data Distribution (IDD) has been a successful innovation in the delivery of meteorological and related data to universities through scaleable data routing topologies (Baltuch, 1997). The increasing volume of data that can be provided to universities in the realm of competitive bandwidth considerations provides interesting challenges for the distribution of data in real-time (Rew and Wilson, 2001). As the utility of the IDD becomes evident for distributing increasingly specialized data sets, new paradigms in automated routing and discovery must be explored (Chiswell, 2001). In order to move the current IDD system toward an automated topology, an updated statistics collection mechanism must be developed which is able to provide both real-time data feed and latency statistics for any node in the topology. The statistics collection must also be able to provide the routing information necessary to determine what the current status of the topology tree is. The ability to reliably deliver the soaring volume of data is paramount to the successful integration of these products into teaching and research activities.

2. COLLECTING REAL-TIME STATISTICS

In order to ascertain the quality and timeliness of the Unidata IDD, participants are asked to provide automated statistical feedback on the performance of their data reception. Currently, statistics are collected by the Local Data Manager (LDM) software on each machine participating in the IDD, with an hourly summary delivered to a central location via standard internet mail protocols. The growing volume of data has lead to the continued advancement of the LDM to become increasingly efficient. However, the collection and reporting of statistics has remained relatively unchanged. In order to move toward automated topology and failover, a more timely and reliable method for reporting statistics must be created.

Many universities employ firewalls and other security methods for protecting their networked hosts. The current use of sendmail for delivery of statistics is well suited to this end since universities typically provide campus mail servers. However, due to the handing off of mail messages to other servers, the timeliness of this method can be a major factor when considering delivering statistics of greater temporal resolution. Generally, firewalls can be configured to allow data packets to known hosts and port addresses through as is typical for the LDM protocol. Other highly reliable TCP based services can fulfill the need for proxy or gateway delivery as well. Here, the LDM protocol has been employed for a general proof of concept.

The current hourly statistics summary has been augmented to provide an instantaneous latency value of the most recent product received for each feedtype and upstream host feeding data. Periodic delivery of this information to a central statistics collection server provides a real-time statistical catalog for the IDD network. The utility of this database is being explored in order to make automated topology decisions.

3. DETERMINING TOPOLOGY

The IDD routing topologies are currently manually maintained based on site data requirements and the ability for upstream hosts to deliver the requested data to the recipient. Local network conditions can necessitate a feed site to change its upstream host for a particular data set. Other conditions effecting host computer reliability and performance can also be taken into account for failover decisions. As the topology becomes increasingly dynamic, topology decisions must become automated in order to respond to local network conditions and individual users of data sets.

At present, topology information is inferred from the upstream connection logs. In order to maintain the topology history, the data logs employ a viscosity of several days. As a result, the instantaneous topology of the IDD has traditionally been unknown. By tagging each product with the upstream host identifier before it is inserted into the local data queue, the routing history of each product can be maintained. The current statistics collection has further been augmented to provide the upstream host from which each product is received. In this manner, not only is the current status of the network known, but the instantaneous quality of each feed route can also be determined.

4. NEW DIRECTIONS

The new development of statistical information and collection mechanisms provides the basis for the development and testing of algorithms to detect trouble spots in the distribution topology and provide alternative feed hosts for failover decisions. The central statistics collection provides the database necessary for individual IDD nodes to discover possible upstream hosts and access their current quality. Methods for accessing the network statistics and making automated decisions based on the current state will be discussed.

5. REFERENCES

Baltuch, M, 1997: Unidata's Internet Data Distribution. *13TH International Conference on IIPS for Meteorology, Oceanography, and Hydrology*. **6.6**.

Chiswell, S., 2001: Evolving Data Requirements for Research and Education. *17TH International Conference on IIPS for Meteorology, Oceanography, and Hydrology*. **12.15**

Rew, R., and A. Wilson, 2001: The Unidata LDM System: Recent Improvements for Scalability. *17TH International Conference on IIPS for Meteorology, Oceanography, and Hydrology*. **4.19**.