Dynamics of Internet Routing Information *

Bilal Chinoy
San Diego Supercomputer Center
San Diego, CA 92186-9784

bac@sdsc.edu

Abstract

The Internet is a complex mesh of networks that use a common suite (TCP/IP) of networking protocols. A key feature of the Internet is that all of these constituent networks are interconnected, thereby providing system wide communication. The magnitude and pattern of the flow of routing information directly represents the connectivity stability of the Internet. The NSFNET backbone network provides transit services to a large portion of the global Internet and maintains routing tables reflecting this current connectivity. These routing tables are constantly updated based on information received by the attached networks. This paper investigates the dynamics of routing information flow as presented to the NSFNET backbone network.

1 Introduction

The global Internet is a rapidly growing community of interconnected networks. In the U.S, these networks are loosely organized as a tri-level hierarchy, with the continental backbone networks at the top of the hierarchy. Networks at this level provide transit services to large portions of the Internet and changes in client network connectivity are reflected in these backbone routing tables. The volatility of these changes is a direct indicator of the state of flux in Internet connectivity. By analyzing this routing information and quantifying its dynamics, one can get a clearer picture of the degree of stability in Internet connectivity.

This paper examines changing routing information in an effort to better characterize Internet system-wide connectivity. The NSFNET system of regional and campus networks, which represent a large fraction of the global Internet, are serviced by the NSFNET wide-area backbone network [1]. We analyze routing fluctuations seen at the NSFNET backbone to measure the connectivity stability of the attached networks.

2 Motivation

Our motivation for this study comes from our recognition of the lack of quantitative information about the dynamics of the Internet routing system, in spite of its key role in the stability of Internet connectivity. A wide variety of routing protocols and techniques are in use today and most networks are custom-engineered and configured. While some researchers have studied the behavior of routing protocols within the confines of a homogeneous network, no one has tackled the system. This study will, we hope, represent a first step in quantifying Internet-wide routing stability.

3 Internet routing

The TCP/IP Internet is organized as an interconnection of *Autonomous Systems* (hereafter known as AS). An AS is a collection of internetwork routers managed and administered by a single authority or organization. A variety of routing schemes and protocols are used to maintain state information and compute paths, both within the confines of an AS and between adjacent ASs. Based on the above model, Internet routing protocols fall into 2 classes. *Intra-AS* protocols are used within the boundaries of an AS, and *inter-AS* protocols are used between ASs. Typically an AS uses a single intra-AS protocol within its boundaries to generate and propagate routing information, though is not unusual to have ASs using multiple protocols.

The DARPA sponsored Arpanet [2] served as the principal wide-area transit network for the Internet for many years. The Arpanet maintained a complete and exhaustive list of IP networks reachable at any given time in its routing tables. Any network not contained in the routing tables was deemed unreachable at that time, and all datagrams destined for that network were not

*Supported by National Science Foundation Grant NCR-9119473

forwardable. This routing information omniscience and the requirement for a single transit network to know about all connected networks at all times was referred to as the *core* routing model, after the observation that the Arpanet was effectively the core of the Internet.

Although the Arpanet has since been replaced by a combination of several national backbone networks serving as major transit paths and no single core network exists, elements of the model persist. In general, packet forwarding within the major backbone networks still requires explicit knowledge of the destination network address. *Default* routing, whereby packets destined to networks not explicitly contained in routing tables are forwarded on a predetermined default route, is not the accepted practice in these national transit networks. Many smaller networks that wish to constrain the size of their routing tables prefer this scheme of default routing, and direct all their default traffic to a chosen transit network. This does not preclude them, of course, from choosing different (and arguably better) routes for networks for which they do maintain explicit knowledge.

Connectivity changes for a network are first processed and then propagated by the parent AS using the appropriate intra-AS protocol. If this information needs to be known outside the boundaries of the parent AS (as is the general connectivity case), inter-AS protocols pick up the change and propagate it to neighboring ASs. This process continues until all Internet ASs that have been configured to explicitly receive state information about this network have been notified. Since complete knowledge in national backbone routing tables means that this information must be received by the backbone AS, a dynamic picture of the state of connectivity of the attached networks can be formed by observing changes to the backbone routing tables.

4 NSFNET system routing

The NSFNET routing architecture and model is described in [3] and its essence is described here. In the NSFNET system of networks, the NSFNET backbone network occupies the top-most level in the tri-level hierarchy, while regional networks form the second tier. These nets are themselves composed of smaller campus and institution networks, which then form the third tier.

For routing purposes, the NSFNET backbone is modeled as a single AS into which regional ASs connect. At each connection point, one or more regional ASs interact with the backbone in order to share routing and reachability information. The backbone and its attached regional networks use either EGP (Exterior Gateway Protocol) [4] or BGP (Border Gateway Protocol) [5] as the inter-AS routing protocol. The NSFNET backbone itself uses a subset of the ANSI standard IS-IS protocol, adapted for IP networks, for its intra-AS routing protocol.

4.1 Routing information flow

The NSFNET backbone and regional network border routers periodically exchange information at the attachment points. This information flow is bidirectional, with each AS updating its peer about reachable networks. Consider the typical case shown in Figure 4, where net `X.Y.Z` belongs to AS A and net $X_1.Y_1.Z_1$ belongs to AS B. If `X.Y.Z` and $X_1.Y_1.Z_1$ wish to exchange packets, then all routers along the path must have appropriate entries pointing to the next router in the AS path, which in this example is A→L→M→N→B. [1]. ASs L and N could be regional systems and M could represent the wide-area backbone interconnecting them.

Consider a change in state for network `X.Y.Z`. This is detected and disseminated within the parent AS A by the intra-AS routing protocol that A uses. It is then propagated across AS boundaries until all ASs have been made aware of this change. Note that every routing protocol has a finite hold-down time for state information changes, and updates are not propagated onward unless these hold-down timers expire. This is done to prevent short-lived changes from unnecessary dissemination.

Within the NSFNET backbone, the IS-IS intra-AS routing protocol floods the change to all backbone routers, which then proceed to update their routing tables. As this new information is incorporated in the backbone, it is communicated to all the other attached networks via inter-AS protocols at the backbone boundaries. The change is propagated throughout the system of connected nets.

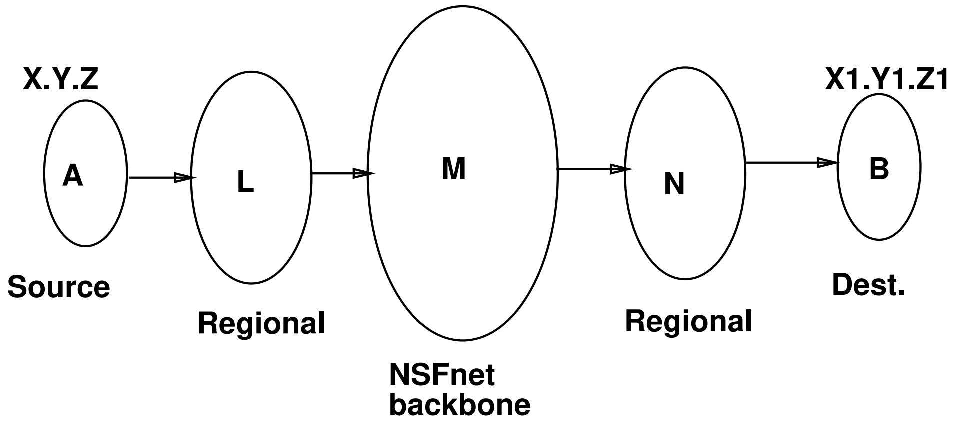
4.2 Routing information content

The essential information carried by any routing protocol is the state of the network which it serves and computation of routes to be used by packet forwarders is made on the basis of this information. Since we are mostly concerned with trans-AS information, we will focus our attention on inter-AS routing protocols.

Every AS has a set of routers that participate in inter-AS routing on behalf of that AS, sharing knowledge about the state of that AS with the neighboring systems. These routers are called *border* routers, because they are at the borders of the parent AS.

[1] ASs A and L could use a default route pointing to backbone AS M, and assume that M has the appropriate entry for the destination network

Fig 1: Routing Information Flow



Within the NSFNET system of networks, every IP network is required to belong to atleast one AS. This allows route aggregation to be done on the basis of AS groups and indeed backbone routers route datagrams to appropriate exit points based on the primary parent AS of the destination IP network address. For the EGP protocol, information carried is in the form

$$< net >< parentAS >< reachable?(Y/N) >$$

That is, EGP only provides binary network reachability information.

The BGP protocol added a number of improvements to EGP, notably attaching a meaningful *metric* of reachability and an *AS path* attribute for each network carried. The metric allows border gateways to select between multiple paths to the same network, and the AS path allows rapid detection of routing loops. The BGP information, therefore, is as

$$< net >< parentAS >< metric >< ASpath >$$

Both protocols are used within the NSFNET system, but a transition toward using only BGP is underway.

5 Measurements

5.1 Trace collection

The entire trace collection for this study was done on the NSFNET backbone routing control processors. These machines are UNIX-based and the routing process is a daemon that handles all communications between the router and its intra and inter-AS peers. This software timestamps and logs all routing related traffic sent and received, allowing reconstruction of routing events and quick detection of routing anomalies and problems.

We chose a 12 hour period on Tuesday, Aug 18 1992 between 10:00 am and 10:00 pm EDT for our study. All NSFNET nodes on the T1 backbone were used to collect data for this study. Each routing processor logged the relevant information to local disk and filters applied to the logged information allowed us to constrain the volume of trace information. At the end of the data collection period we collated and aggregated all traces on a post-processing machine for analysis. It is important to note that data collection was performed during a period of time when there were no known or planned network outages. This means that the volatility seen in the collected data represents normal routing fluctuations within the NSFNET system of networks.

5.2 Trace contents

The trace data collected were of the form

`<Collector node>`: Node collecting data

`<timestamp>`: Timestamp associated with routing message

`<seq.num>`: Sequence number tag

`<sent:Destination>`: Receiving node, if sending message

`<rcvd:Source>`: Source node, if receiving message

`<message type and content>`

where the message type and content was one of

- inter-AS routing update with a list of updated networks
- intra-AS routing update with a list of updated

networks

Each routing message was identified by a sequence number allowing individual messages to be tracked as they traversed the NSFNET backbone network.

6 Analysis

Since the state of connectivity of Internet networks as seen by the NSFNET backbone is the focus of our study, we will define some terms in this context.

A *connectivity transition* event is defined as a network event that causes a client network to either be added to the backbone routing tables while previously absent, or be deleted from the routing tables while previously present. That is, any event signaling a change in reachability of a network as seen by the NSFNET backbone. Note that a network may suffer more outages than are recorded at the NSFNET backbone. This is because most Internet routing protocols employ a hold-down period during which they do not propagate changed information about the state of a network. This serves as a means of damping routing information changes in an attempt to prevent oscillations through the set of connected networks. It is conceivable that networks regain connectivity within the time span of routing protocol hold down periods, effectively hiding the change from the system at large.

An *unreachability cycle* is defined as the interval of time between two successive connectivity transition events, the first being the deletion of a network and the next being the subsequent addition of the same network in the backbone routing tables. The unreachability cycle interval represents the amount of time a network is unreachable as viewed by the backbone and packets destined to that network cannot be forwarded.

A *cluster* of networks is a group of networks that undergo an unreachability cycle together. The cluster represents the aggregation of networks that suffer connectivity outages together, again as seen by the NSFNET backbone. As an example, consider the case where a group of networks share a common path to the NSFNET backbone. If that path becomes unusable then all these networks will appear to become unreachable simultaneously, and will apppear in the backbone routing tables together when the shared path is usable again. This group of networks, then, constitute a cluster.

6.1 EGP information content

As we noted earlier, the NSFNET system of networks use both EGP and BGP as inter-AS routing protocols. One key observation made about the use EGP in large network environments is the large overhead associated with the update process of the protocol. EGP requires that all networks associated with the participating ASs be exchanged in *every* update, as long as those networks are reachable. Every update thus contains the entire list of currently reachable networks, and unreachability is signaled by the absence of a network from this list. In this study we attempted to quantify the magnitude of new information contained in a given EGP update.

The collected data allowed us to build a time history of network fluctuations as reported by EGP. We could then determine how many EGP- exchanged networks actually changed state by checking against previous updates. The information content of a particular EGP update was then the fraction of networks that actually changed state. Figure 2 charts this fraction for all updates seen in our trace. The graph shows that almost 90% of the updates recorded contained close to 0% new information. This signifies that almost none of the EGP updates conveyed any new information about the participating networks, implying a tremendous waste of network bandwidth, router CPU and memory resources.

BGP built upon this observation by requiring only *incremental* updates, whereby only those networks which undergo state transitions are exchanged in routing updates. This limited the size and frequency of update messages, conserving network resources.

6.2 Cluster size distribution

The cluster size observed in a routing fluctuation is a good indicator of the nature of the event that causes the outage. Recall that the Internet is loosely organized as a tri-level hierarchy and that regional networks that are themselves composed of smaller networks are usually directly connected to the backbone networks. Larger cluster sizes seen during a routing fluctuation typically indicate a problem closer to the backbone level. These events could be caused by a distribution network breakdown or an outage involving a border router serving multiple smaller ASs. Smaller cluster sizes typically indicate network outages farther away from the wide-area backbone.

Figure 3 shows the cluster size frequency distribution for routing fluctuations seen in the trace. The cluster sizes varied from a single network to a maximum of approximately 450 networks. Since the total number of networks reachable using the NSFNET backbone at the time of the trace collection was approximately 6500, the upper bound represents about 7% of the total set of networks. The graph clearly shows that fluctuations involving less than 10 networks in a cluster was the dominant behavior. Indeed, the frequency of fluctu-

Fig. 2: EGP information content

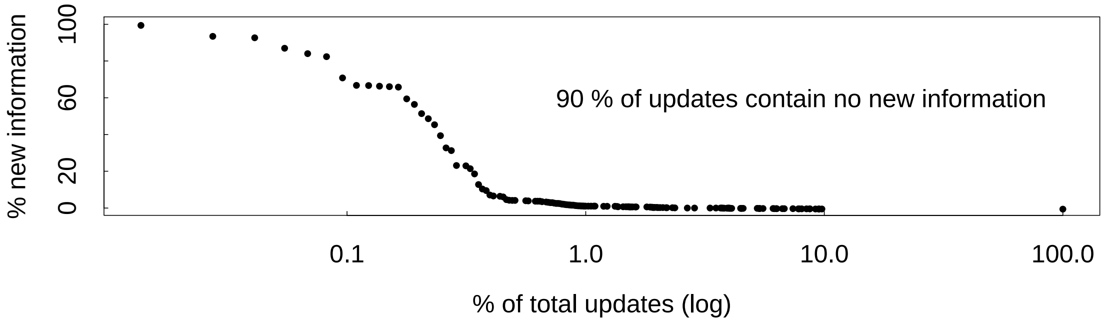
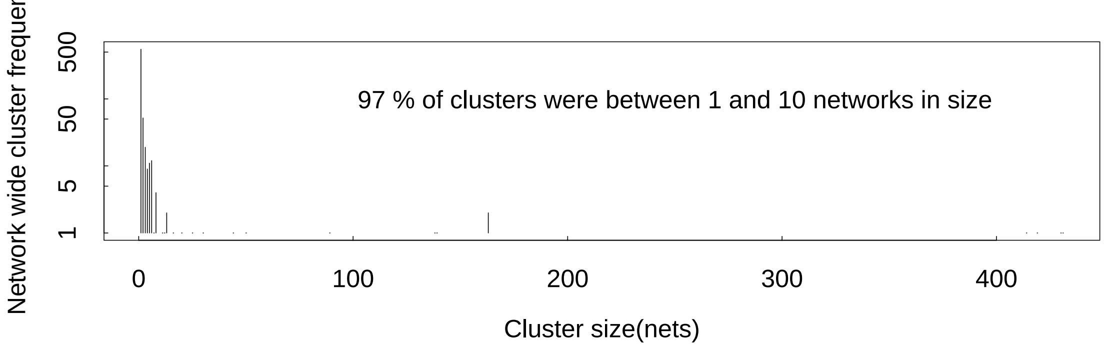


Fig. 3: Cluster Size frequency distribution



ations involving only 1 network was the greatest. This seems to indicate that the Internet has good overall *system* stability and large scale oscillations are relatively uncommon.

6.3 Unreachability cycle distribution

A key question we investigated in this study was the length of time a network or a cluster of networks remained unreachable as viewed from the backbone. This is a first order measure of the availability of a given network to the rest of the system. To that end, a scatter plot of the unreachability cycle interval versus the observed cluster size is plotted in Figure 4. Intervals range from the low value of 30 seconds to a high of 466 minutes (or 7.7 hours) [2]. The larger cluster sizes show two distinct patterns. In one case, the cycle time is about 3 to 10 minutes. This is typically the case when there is a glitch in either physical connectivity or in routing protocols caused by excessive loads. This behavior is usually self-correcting within a few minutes. In the second case we see much larger intervals, signifying connectivity fluctuations of a more longer term nature. These typically require human intervention to correct and the cycle times are thus correspondingly longer and events of this type are rarer.

The smaller cluster behavior can be explained in similar terms, the significant difference being the greater volatility in individual network connectivity. That is, there are many more events that affect the connectivity of individual nets than there are for larger groups of networks, or even entire ASs. Consequently, we see a much larger range of cycle times for the smaller cluster sizes. Some networks show cycle times on the order of many hours while other come in and go out of backbone routing tables relatively quickly.

6.4 Connectivity Transitions

In addition to the cluster size and cycle time distributions, it is instructive to examine the distribution of the number of state transitions a network experiences. It is intuitive to expect that all networks will not exhibit the same degree of volatility, and Figure 5 shows this differential. The graph clearly shows the high degree of inequality in this distribution. We see that a larger fraction (about 97%) of networks experienced less than 10 transition events in the trace period of 12 hours. About 40% of networks showed only one transition, while less

[2] Whereas EGP updates occur every 3 minutes , BGP updates are event-driven and thus we do see cycles of less than 3 minutes duration

Fig. 4: Scatter Plot of Cycle interval v/s Cluster size

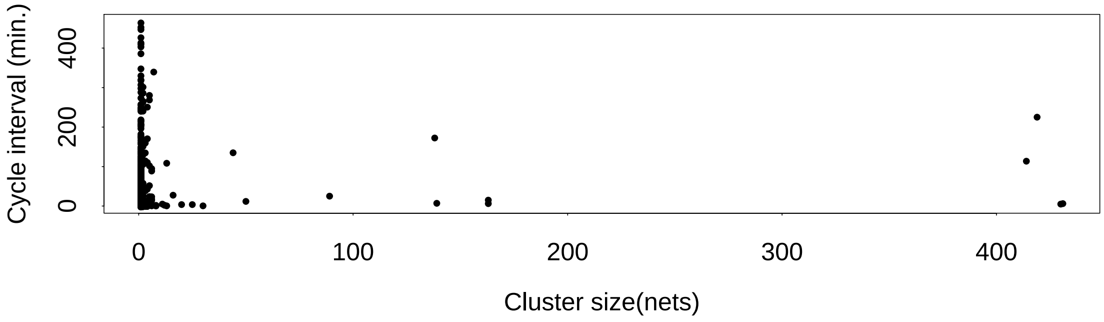
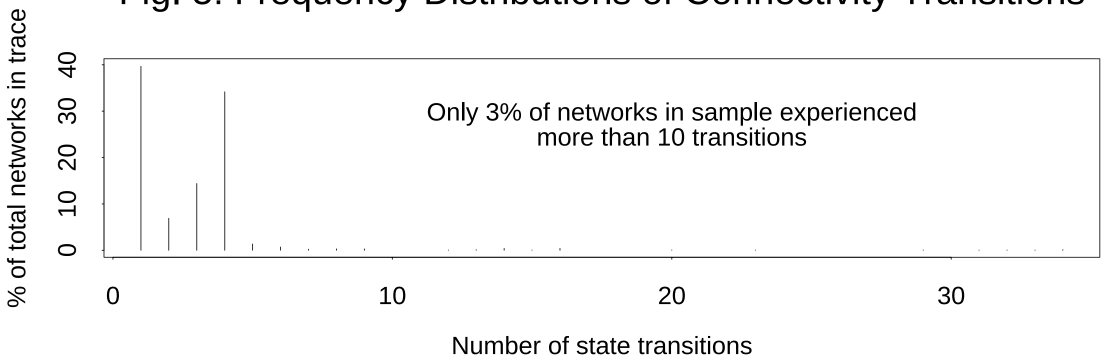


Fig. 5: Frequency Distributions of Connectivity Transitions



than 1% of networks experienced more than 30 transitions

This uneven distribution clearly demonstrates that a small number of networks have much more volatile connectivity to the NSFNET backbone than the large majority. In fact, this evidence suggests that there are regular disturbances in Internet connectivity affecting small numbers of networks consistently.

6.5 Update propagation times

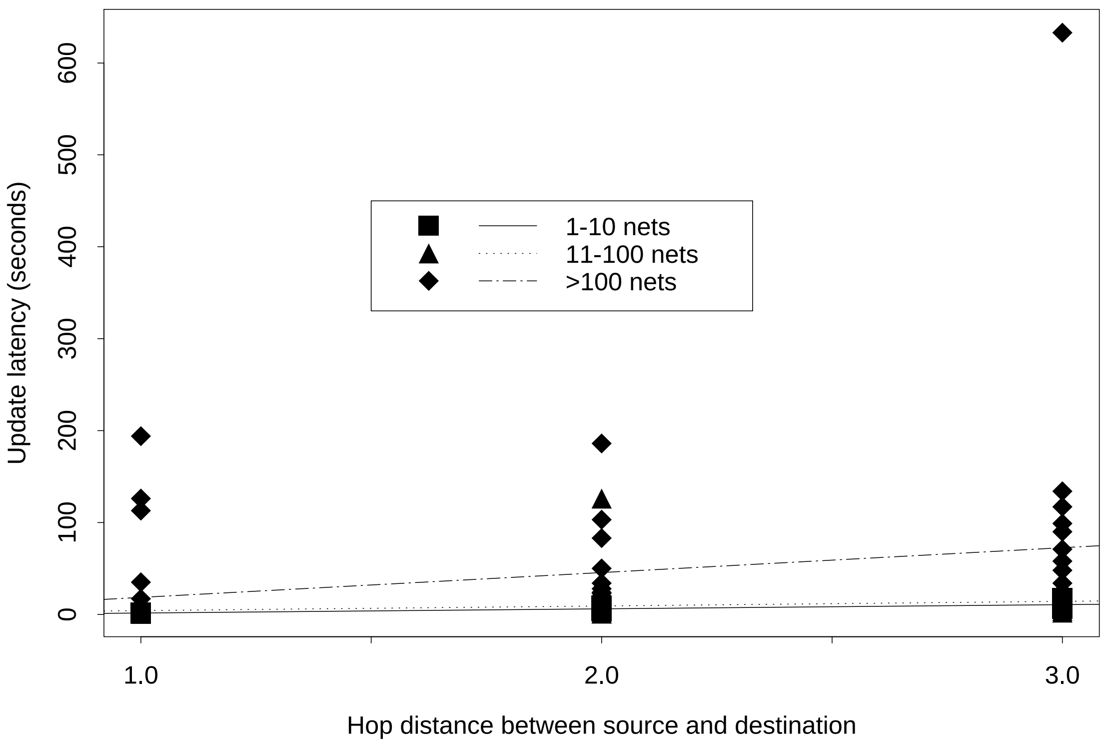
The NSFNET T1 backbone routers are interconnected using DS1 leased circuits, operating at 1.544 Mbps. External IP networks that change state are tagged by the local backbone border router and are then flooded to the rest of the network. This process requires all routers to process incoming update datagrams and to then forward them on to the neighboring routers, thus effectively flooding the information. Since multiple copies of the same information can be received, each update is sequenced and its receipt is time-stamped to ensure information integrity and a single, consistent image of the network at each router.

Consider the case of an attached IP network undergoing a connectivity transition. The time taken for this information to reach an intermediate network is the sum of the times each intermediate AS takes to process and propagate the update. This time is function of many parameters such as the size of the AS, the routing protocol used and the traffic load. While the NSFNET backbone transit time for routing information is but one component of the total delay, examining it provides insight into the magnitudes involved in this process.

Since the link-state based IS-IS update algorithm requires each router to process all information and then forward the updates on, we parameterize the measurements based on the node hop distance between the destination and source routers. Additionally, past experiences indicate that the propagation times depend on the size of the routing updates.

For this study, we measure the time elapsed between a backbone router sourcing a routing update and other backbone routers receiving the same update. Since each update is indexed by a sequence number, we can track the progress of an individual update through the network. This time is termed the *update latency* and is parameterized by the number of external networks the update carries. We must stress at this point that software implementation of the IS-IS mechanism as well as router hardware (memory and CPU) play a role in determining the update latencies observed in this study. However, observations have indicated that the routers do not suffer from lack of either memory or CPU resources during

Fig. 6: Scatter plot of hop distance v/s update latency



update processing and route computation. Variations in traffic loads *do not* affect this study as the NSFNET backbone packet switches maintain dual queues and prefer routing and management packets over user traffic.

We group the observations into 3 classes.

- Updates carrying between 1 and 10 external nets
- Updates carrying between 11 and 100 external nets
- Updates carrying more than 100 external nets

Figure 6 shows the results of these measurements, with the scatter points plotting individual observations. Robust straight line fits are added to show trends in the observations.

The update latencies seem to increase quite rapidly with the update size. Some of the larger updates take on the order of 100 seconds to permeate through the network. The smaller updates, on the contrary, take a modest amount of time to flow through the backbone, typically between 10 and 20 seconds to reach a router 3 hops away.

The graph also shows that update latencies increase with the hop distance between source and destination. This, while intuitively obvious, is nevertheless interesting to study further. We can see that for larger updates, the distance away from the source is a critical factor in determining latencies. A backbone router that is 1 hop away receives a small update approximately 1 to 3 seconds after it was sent, while a 3 hop distant router may receive the update in about 5 to 15 seconds. For larger updates, 1 hop routers receive them in 20 to 50 seconds while 3 hop routers receive them in 50 to 150 seconds.

The average one-way delay for a user packet is approximately 60 ms across the NSFNET T1 backbone. This means that the above update latencies, on the order of seconds, are 3 orders of magnitude greater than the average delay a user packet experiences in transiting the backbone. This ratio of routing latencies to packet delays is an important design issue in high speed Internets. As high as this ratio is today, it could get even higher as transmission speeds increase and inter-AS connectivity grows denser. Inter and Intra-AS routing protocol designers should recognize this disparity and attempt to optimize update algorithms and information aggregation in order to minimize this ratio.

7 Conclusions

This study represents an initial attempt to characterize and quantify the dynamics associated with routing information flow in the Internet. The most illustrative results show the unequal distribution of fluctuation cluster sizes and cycle intervals. It is seen that most of the fluctuations involve small numbers of networks and cycle intervals are on the order of 10 minutes. Most networks experience only a few transitions while a very small number of nets show a high degree of volatility. Finally, we see that the NSFNET backbone update process is sensitive to the size of the update as well as the hop distance between source and destination. Larger updates may take many minutes to permeate through the backbone. This causes a high update latency to user packet latency ratio, which has consequences on overall system stability and network reachability.

8 Future work

Our paper raises some important questions regarding Internet system stability. The various causes of network connectivity fluctuations need to be further investigated and characterized. While this study presented a backbone centered view of the global routing system, studies quantifying the end-to-end dynamics of routing information are also needed. These would include the parameters associated with propagating information down to the leaf nodes of the Internet. Finally, a comprehensive model of Internet wide routing flow dynamics would help in predicting end-to-end performance and reliability.

References

[1] Chinoy, B., Braun, H.W., "The New NSFNET Backbone Network", TR GA-A21029, SDSC 1992

[2] McQuillan, J., Walden, D., "The ARPA Network Design Decisions", Computer Networks, Vol. 1, No. 5, Aug 1977, pp 243-289.

[3] Rekhter, Y., "EGP and Policy Based Routing in the New NSFNET Backbone", RFC 1092, Feb 1989.

[4] Mills, D.L., "Autonomous Confederations", RFC 975, Feb 1986.

[5] Lougheed, K., Rekhter, Y., "A Border Gateway Protocol", RFC 1163, June 1990

[6] Rekhter, Y., "The NSFNET Backbone SPF based Interior Routing Protocol", RFC 1074, Oct 1988.

[7] Postel, J., "Internet Protocol", RFC 791, Sep 1981.