A Measurement Study of Changes in Service-Level Reachability in the Global TCP/IP Internet: Goals, Experimental Design, Implementation, and Policy Considerations

Michael F. Schwartz

Internet Request For Comments 1273   November 1991

Department of Computer Science
Campus Box 430
University of Colorado
Boulder, Colorado 80309-0430
(303) 492-3902
Internet: schwartz@cs.colorado.edu

Abstract

In this report we discuss plans to carry out a longitudinal measurement study of changes in service-level reachability in the global TCP/IP Internet. We overview our experimental design, considerations of network and remote site load, mechanisms used to control the measurement collection process, and network appropriate use and privacy issues, including our efforts to inform sites measured by this study. A list of references and information on how to contact the Principal Investigator are included.

An Earlier version of this report appeared as University of Colorado Technical Report CU-CS-551-91, October 1991.

Introduction

The global TCP/IP Internet interconnects millions of individuals at thousands of institutions worldwide, offering the potential for significant collaboration through network services and electronic information exchange. At the same time, such powerful connectivity offers many avenues for security violations, as evidenced by a number of well publicized events over the past few years. In response, many sites have imposed mechanisms to limit their exposure to security intrusions, ranging from disabling certain inter-site services, to using external gateways that only allow electronic mail delivery, to gateways that limit remote interactions via access control lists, to disconnection from the Internet. While these measures are preferable to the damage that could occur from security violations, taken to an extreme they could eventually reduce the Internet to little more than a means of supporting certain pre-approved point-to-point data transfers. Such diminished functionality could hinder or prevent the deployment of important new types of network services, impeding both research and commercial advancement.

To understand the evolution of this situation, we have designed a study to measure changes in Internet service-level reachability over a period of one year. The study considers upper layer service reachability instead of basic IP connectivity because the former indicates the willingness of organizations to participate in inter-organizational computing, which will be an important component of future wide area distributed applications.

The data we gather will contribute to Internet research and engineering planning activities in a number of ways. The data will indicate the mechanisms sites use to distance themselves from Internet connectivity, the types of services that sites are willing to run (and hence the type of distributed collaboration they are willing to support), and variations in these characteristics as a function of geographic location and type of institution (commercial, educational, etc.). Understanding these trends will allow application designers and network builders to more realistically plan for how to support future wide area distributed applications such as digital library systems, information services, wide area distributed file systems, and conferencing and other collaboration-support systems. The measurements will also be of general interest, as they represent direct measurements of the evolution of a global electronic society.

Clearly, a study of this nature and magnitude raises a number of potential concerns. In this note we overview our experimental design, considerations of network and remote site load, mechanisms used to control the measurement collection process, and our efforts to inform sites measured by this study, along with concomitant network appropriate use and privacy issues.

A point we wish to stress from the outset is that this is not a study of network security. The experiments do not attempt to probe the security mechanisms of any machine on the network. The study is concerned solely with the evolution of network connectivity and service reachability.

Experimental Design

The study consists of a set of runs of a program over the span of one to two days each month, repeated bimonthly for a period of one year (in January 1992, March 1992, May 1992, July 1992, September 1992, and November 1992). Each program run attempts to connect to 13 different TCP services at each of approximately 12,700 Internet domains worldwide, recording the failure/success status of each attempt. The program will attempt no data transfers in either direction. If a connection is successful, it is simply closed and counted. (Note in particular that this means that the security mechanism behind individual network services will not be tested.) The machines on which connections are attempted will be selected at random from a large list of machines in the Internet, constrained such that at most 1 to 3 machines is contacted in any particular domain.

The services to which connections will be attempted are:

Port Number	Service	Port Number	Service
13	daytime	111	Sun portmap
15	netstat	513	rlogin
21	FTP	514	rsh
23	telnet	540	UUCP
25	SMTP	543	klogin
53	Domain Naming System	544	krcmd, kshell
79	finger		

This list was chosen to span a representative range of service types, each of which can be expected to be found on any machine in a site (so that probing random machines is meaningful). The one exception is the Domain Naming System, for which the machines to probe are selected from information obtained from the Domain system itself. Only TCP services are tested, since the TCP connection mechanism allows one to determine if a server is running in an application-independent fashion.

As an aside, it would be possible to retrieve "Well Known Service" records from the Domain Naming System, as a somewhat less "invasive" measurement approach. However, these records are not required for proper network operation, and hence are far from complete or consistent in the Domain Naming System. The only way to collect the data we want is to measure them in the fashion described above.

Network and Remote Site Load

The measurement software is quite careful to avoid generating unnecessary internet packets, and to avoid congesting the internet with too much concurrent activity. Once it has successfully connected to a particular service in a domain, the software never attempts to connect to that service on any machine in that domain again, for the duration of the current measurement run (i.e., the current 60 days). Once it has recorded 3 connection refusals at any machines in that domain for a service, it does not try that service at that domain again during the current measurement run. If it experiences 3 timeouts on any machine in a domain, it gives up on the domain, possibly to be retried again a day later (to overcome transient network problems). In the worst case there will be 3 connection failures for each service at 3 different machines, which amounts to 37 connection requests per domain (3 for each of the 12 services other than the Domain Naming System, and one for the Domain Naming System). However, the average will be much less than this.

To quantify the actual Internet load, we now present some measurements from test runs of the measurement software that were performed in August 1991. In total, 50,549 Domain Naming System lookups were performed, and 73,760 connections were attempted. This measurement run completed in approximately 10 hours, never initiating more than 20 network operations (name lookups or connection attempts) concurrently. The total NSFNET backbone load from all traffic sources that month was approximately 5 billion packets. Therefore, the traffic from our measurement study amounted to less than .5% of this volume on the day that the measurements were collected. Since the Internet contains several other backbones besides NSFNET, the proportionate increase in total Internet traffic was significantly less than .5%.

The cost to a remote site being measured is effectively zero. From the above measurements, on average we attempted 5.7 connections per remote domain. The cost of a connection open/close sequence is quite small, particularly when compared to the cost of the many electronic mail and news transmissions that most sites experience on a given day.

Control Over Measurement Collection Process

The measurement software evolved from an earlier set of experiments used to measure the reach of an experimental Internet white pages tool called netfind [Schwartz & Tsirigotis 1991b], and has been evolved and tested extensively over a period of two years. During this time it has been used in a number of experiments of increasing scale. The software uses several redundant checks and other mechanisms to ensure that careful control is maintained over the network operations that are performed [Schwartz & Tsirigotis 1991a]. In addition, we monitor the progress and network loading of the measurements during the measurement runs, observing the log of connection requests in progress as well as physical and transport level network status (which indicate the amount of concurrent network activity in progress). Finally, because the measurements are controlled from a single centralized location, it is quite easy to stop the measurements at any time.

Network Appropriate Use and Privacy Issues

When we performed our initial test runs of this study, we attempted to inform site administrators at each study site about this study, by posting a message on the USENET newsgroup "alt.security" and by sending individual electronic mail messages to site administrators. We also informed the Computer Emergency Response Team (CERT) at CMU of the study. As a practical matter, informing all sites turned out to be quite difficult. Part of the problem was that no channels exist to allow such information to be easily disseminated. Approximately half of the messages we sent to site administrators were returned by remote mail systems as undeliverable. Moreover, the network traffic and remote site administrative load caused by the study announcement messages far outstripped the network and administrative load required by the study itself. Some sites felt that the announcement was an unnecessary imposition of their time.

In addition to these practical problems, a broad announcement of this study could affect the measurements it attempts to gather. Some sites would likely react to the announcement by changing the reachability of their services. Asking for explicit permission from sites would yield even worse methodological problems, as this would have provided a self-selected study group consisting of sites that are less likely to disconnect from the Internet.

In contrast with our attempts to announce the study, running the study without announcing it caused only a small number of site administrators to notice the traffic and inquire about it to either the CERT or to one of the responsible network contacts at the University of Colorado. The remote site administrator and network overhead of announcing the the study, coupled with the practical and methodological problems of announcing the study, lead us to prefer to run the study without further broad announcements. Yet, to avoid causing alarm at a site detecting our network measurement activity, it makes sense to announce the study.

To resolve this problem, we discussed the study with the Internet Activities Board, Internet Engineering Steering Group, National Science Foundation, representatives of several U.S. regional networks, and a number of individuals involved with network security, including the Computer Emergency Response Team, members of the Internet Engineering Task Force Security and Advisory Group, and a member of the Lawrence Livermore National Laboratory Computer Incident Advisory Capability. The first part of our efforts resulted in the production of Internet Request For Comments (RFC) number 1262 [Cerf 1991]. Beyond this, we have agreed that the appropriate action at this point is to announce the study well ahead of running it via the current RFC, augmented with an electronic posting that briefly describes the study goals and methodology and points to this RFC. That announcement will be posted to the Internet Engineering Task Force mailing list, the comp.protocols.tcp-ip USENET bulletin board, and the Computer Emergency Response Team's cert-tools mailing list. Moreover, in case a site misses these announcements, we will run the measurement software in a fashion intended to minimize the effort a site administrator might expend to determine the nature of the activity after detecting it. In particular, we will run the program from an account called "testnet" on a machine with few other users logged in. "Fingering" [Zimmerman 1990] this machine will indicate the testnet login. "Fingering" the testnet login will return information about this study.

The data collected by this study is somewhat sensitive to privacy and security concerns, in the sense that it might be used as a "road map" of accessible network services. We will treat the raw data as private information, publishing measurements only in global statistical terms, divorced from the actual sites that make up the underlying data points. We previously carried out a study with much larger privacy implications than the current study [Schwartz & Wood 1991], and successfully masked the data to protect individual privacy.

For Further Information

Information about the general research program within which this study fit is available by anonymous FTP from latour.cs.colorado.edu, in pub/RD.Papers. This directory contains a "README" file that describes the overall research project (which focuses on resource discovery), and includes a bibliography. Particularly relevant are:

- [Schwartz 1991b], a project overview;
- [Schwartz 1991a], about an earlier, simpler version of the current study;
- [Schwartz & Tsirigotis 1991b], about the netfind white pages tool;
- [Schwartz & Tsirigotis 1991a], which considers a number of the techniques used in this experiment, including those for controlling the progress of the measurements;

and

- [Schwartz & Wood 1991], about an earlier study we carried out that raises significant potential privacy questions, for which we carefully masked the underlying data, presenting the results without sacrificing individual privacy.

Also:

- [Cerf 1991], IAB guidelines for Internet measurement activity.

Once the results of this study are complete, we will publish them in a conference or journal, as well as by anonymous FTP.

Communication With Principal Investigator

If you would like to have your site removed from this study, or you would like to be added to the list of people who receive results from this study, or you would like to communicate with the Principal Investigator for some other reason, please send electronic mail to schwartz@cs.colorado.edu.

References

[Cerf 1991] V. G. Cerf, editor. Guidelines for Internet Measurement Activities. Request For Comments 1262, Internet Activities Board, October 1991.

[Schwartz & Tsirigotis 1991a]
M. F. Schwartz and P. G. Tsirigotis. Techniques for Supporting Wide Area Distributed Applications. Technical Report CU-CS-519-91, Department of Computer Science, University of Colorado, Boulder, Colorado, February 1991; Revised August 1991. Submitted for publication.

[Schwartz & Tsirigotis 1991b]
M. F. Schwartz and P. G. Tsirigotis. Experience with a Semantically Cognizant Internet White Pages Directory Tool. *Journal of Internetworking: Research and Experience*, 2(1), pp. 23-50, March 1991.

[Schwartz 1991a]
M. F. Schwartz. The Great Disconnection? Technical Report CU-CS-521-91, Department of Computer Science, University of Colorado, Boulder, Colorado, February 1991.

[Schwartz & Wood 1991]
M. F. Schwartz and D. C. M. Wood. A Measurement Study of Organizational Properties in the Global Electronic Mail Community. Technical Report CU-CS-482-90, Department of Computer Science, University of Colorado, Boulder, Colorado, August 1990; Revised July 1991. Submitted for publication.

[Schwartz 1991b]
M. F. Schwartz. Resource Discovery in the Global Internet. Technical Report CU-CS-555-91, Department of Computer Science, University of Colorado, Boulder, Colorado, November 1991. Submitted for publication.

[Zimmerman 1990]
D. Zimmerman. The Finger User Information Protocol. Request For Comments 1194, Center for Discrete Mathematics and Theoretical Computer Science, November 1990.