Electrical Engineering Department University of Delaware
Technical Report 90-7-1 July 1990

Support of the National Science Foundation Supercomputer Network Program

Final Report on NSF/DNCRI Grant NCR-86-12015

David L. Mills

Abstract

This is the final report on the NSF Division of Network and Communications Research and Infrastructure grant NCR-86-12015: "Support of the National Science Foundation Supercomputer Network Program," for the period September 1986 through February 1989.

1. Introduction

This is the final report on the NSF Division of Network and Communications Research and Infrastructure grant NCR-86-12015: "Support of the National Science Foundation Supercomputer Network Program," for the period September 1986 through February 1989. The grant provided services in support of the National Science Foundation program for Supercomputer Networks and included advanced architectural planning, identification and study of related technical problems, performance evaluation and assistance on internet configuration and protocol-convergence issues.

2. Background

The DARPA Internet program was conceived about twelve years ago with the goal of interoperability over a wide spectrum of network technologies, including proliferated mesh networks such as ARPANET and MILNET, satellite networks such as SATNET and WIDEBAND, as well as packet-radio networks and many other local-area technologies. In many ways the goals and technologies of the Internet program parallel those of the NSF Supercomputer Network program. In addition, a large body of supporting hardware and software for the Internet protocols is available, along with an extensive technical corps familiar with its operation and maintenance. It is a stated goal of the NSF Supercomputer Network program to utilize Internet technology to achieve interoperability between the existing and planned Supercomputer networks.

In 1986 the near-term goals of the NSF Supercomputer Network program was the establishment of the Phase-I Backbone network, consortia networks and pilot projects and their integration as an interoperable network or internet. Since many of the resources to accomplish these goals were shared among several services (e.g., workstations) this required integration with campus networks and the Internet system as well.

Experience has shown repeatedly that the success of a large-scale integration such as this requires intricate planning and the participation of technical experts with a thorough understanding of the requirements, alternatives and solution methodologies. This effort encompasses the spectrum from network architecture and protocol selection through traffic forecasting, resource acquisition, system configuration and operational support.

The specific problem addressed in the proposal was the identification, study and resolution of the technical issues that arise as the result of the NSFNET Phase-I Backbone integration effort. For example, it was necessary to establish specific guidelines to assist vendors of host and gateway systems in the resolution of implementation issues and strategies peculiar to the NSF community. It was particularly important to provide technical advice to the users and operating organizations on the resolution of specific issues, such as system configuration, problem tracking and monitoring/control. In addition, it was necessary to establish service specifications and protocol designs for new services such as access control and resource accounting specifically in the NSF context. Finally, it was necessary to incorporate in the study and design cycles considerations for new architectures and protocols, including the eventual transition to ISO protocols.

The specific effort proposed in the grant included the following:

1. Preparation of technical planning documents and architecture projections relevant to existing or proposed NSF internets.

2. Study and recommendation on specific problems in protocol interoperability, such as strategies for the interconnection of new and experimental networks or protocols to an NSF internet.

3. Attendance as chair or member of various planning meetings held at NSF and participating organizations, including agenda contributions and study papers.

4. Monitoring of the planning processes of NSF and the participating organizations in order to identify and resolve potential problems that may affect interoperabiilty or the efficient utilization of resources.

A particularly important aspect of this effort was the collaboration with other projects, included a separately funded DARPA contract. This contract provided several items of equipment, including workstations, peripheral equipment and connections to ARPANET.

3. Accomplishments

Much of the effort involved in this grant was in the area of engineering studies and support of the NSFNET Phase-I Backbone network, which used packet switches based on the LSI-11 processor and Fuzzball operating system. The network design is documented in [MIL87e] and the Fuzzball design and evolution in [MIL88b]. The protocol used for the retrieval of statistics information from the Fuzzball is presented in [MIL87a], while the protocol for time synchronization is presented in [MIL88a]. During its lifetime from September 1986 to its dismantling in July 1988, the traffic carried on the net escalated far beyond the original design capabilities, which required extensive re-engineering of the routing algorithms and congestion-avoidance mechanisms. However, the experience gained, documented in [MIL87b], [MIL87d], [MIL87e] and [MIL88b], has led to new and deeper understanding of the issues involved in building very large, very fast and very complex nets.

The Principal Investigator continued to attend meetings of the Internet Activities Board (IAB) and Internet Architecture (INARC) Task Force, as well as several ad hoc meetings as called. Graduate students Mike Minnich and Paul Schragger attended some of these meetings in the capacity of secretary/scribe and took voluminous notes on the proceedings. In June 1989 a workshop was held sponsored by the University and the Internet Architecture Task Force titled “Future of the Internet System Architecture and TCP/IP Protocols.” The workshop was exceptionally well attended and reported in [MIL90b]. Graduate students Paul Schragger and Mike Davis collaborated on the minutes and notes cited in this document.

While the NSFNET Phase-I Backbone net has been dismantled, all six of the Fuzzball packet switches remain in operation as Internet time servers supporting the Network Time Protocol [MIL89b], [MIL89c] and [MIL89d]. Since these and about a dozen other Fuzzballs also employed as time servers and measurement hosts are directly connected to regional and backbone nets such as the ARPANET and NSFNET Phase-II Backbone, they serve as ideal observation and performance evaluation platforms. Work proceeding now is directed at the refinement and extension of time-transfer techniques [MIL90d], the design of strategic experiments that collect detailed one-way delay information over long periods of time on a non-intrusive basis [MIL90a] and on innovative routing strategies [MIL89a].

In addition to the above effort, a number of students have been supported in varying degrees under the grant. In Summer 1988 Erik Perkins completed a study program on high-speed multiprocessor architecture. In Summer 1989 Mark Shaw and Stefan Levie studied issues in time-transfer technology and completed related hardware and software projects. Also in Summer 1989 Mike Davis completed a traffic study on the NSFNET Phase-I Backbone network using data collected over the period 1986-1988. While not directly supported under the grant, several students completed term

papers on subjects related to NSFNET issues, including the Domain Name System, congestion control, routing and policy management.

4. Future Plans

A follow-on proposal has been submitted to the NSF Division of Network and Communications Research and Infrastructure. It continues the effort begun with the previous NSF grant and is designed to assist the DNCRI program in four ways:

1. Attend designated meetings on behalf of NSF/DNCRI interests, including IAB and INARC meetings and related workshops.

2. Analyze and evaluate data collected on the NSFNET Phase-I Backbone, study certain periods when the net became violently overloaded and fell into a state of near congestive collapse and evaluate the effectiveness of the Fuzzball congestion-avoidance mechanisms during these periods.

3. Maintain and update specifications and reference implementations for the Network Time Protocol (NTP), including certain public NSFNET time services, radio clocks and measurement platforms.

4. Explore selected performance issues with the NSFNET Phase-II Backbone, including the collection of relevant performance data, analysis of experiments and investigation of anomalous behavior.

In addition to the above, a renewal proposal has been submitted to DARPA to continue the research in network architecture, congestion control, routing and policy management.

With Professors C.S. Ih, C.G. Boncelet and P. Amer and our students, we have proposed a white paper [MIL88c] for a research program in high-speed networking. This paper has been approved for inclusion in the Corporation for National Research Initiatives Program on Gigabit Networks. The paper describes an architecture for a high-speed net and packet switch capable of operation in the gigabit range over distances spanning the country. It is intended for applications involving very high traffic rates typical of supercomputer and motion-video applications. The net is assumed composed of fiber trunks, while the switch architecture is based on an inexpensive, bit-slice crossbar switch which can be configured for speeds from 25 Mbps to 1 Gbps by simply adding slices. The switches are externally controlled by a microprogrammed sequencer using a distributed reservation protocol and scheduling algorithm.

This project involves the development of a high-speed reservation-TDMA (R-TDMA) crossbar-switch and scheduling algorithm, together with a prototype network operating at speeds of about 100 Mbps and scalable to well over 1 Gbps. The emphasis in this work is on the design of distributed algorithms for reservation and scheduling to achieve very high burst rates over long distances together with efficient use of network resources.

In addition to the above, a renewal proposal has been submitted to DARPA to continue the research in network architecture, congestion control, routing and policy management.

5. Publications

The following publications have been supported by (1) NSF Division of Network and Communications Research and Infrastructure grant NCR-86-12015: “Support of the National Science

Foundation Supercomputer Network Program“ (September 1986 - February 1989, $29K) and (2) DARPA Information Processing Technology Office order 6036, contract N00140-87-C-8901: “Research in Internet Architectures and Protocols“ (June 1987 - December 1989, $282K). Each citation is accompanied by the abstract of the paper or an extract from the introduction.

[MIL87a] Mills, D.L. Statistics server. DARPA Network Working Group Report RFC-996. University of Delaware, February 1987.

This RFC specifies a standard for the ARPA Internet community. Hosts and gateways on the DARPA Internet that choose to implement a remote statistics monitoring facility may use this protocol to send statistics data upon request to a monitoring center or debugging host.

[MIL87b] Mills, D.L. Dissimilar Gateway Protocol. Memorandum distributed to the Internet Engineering Task Force and NSF Network Working Group, February 1987.

Information useful for routing between autonomous systems is presently exchanged between them using the Exterior Protocol (EGP), described in RFC-904. A new protocol called the Dissimilar Gateway Protocol (DGP) has been proposed and a feasibility prototype demonstrated. While EGP provides the basic architectural model and protocols to interconnect sets of systems in a topologically flat space, DGP provides a hierarchical structure in which EGP could be used within a single level and DGP would be used between levels.

[MIL87c] Mills, D.L. A distributed-protocol authentication scheme. DARPA Network Working Group Report RFC-1004. University of Delaware, April 1987.

The purpose of this RFC is to focus discussion on authentication problems in the Internet and possible methods of solution. The proposed solutions this document are not intended as standards for the Internet at this time. Rather, it is hoped that a general consensus will emerge as to the appropriate solution to authentication problems, leading eventually to the adoption of standards.

[MIL87d] Mills, D.L. Metric Transformations. Memorandum distributed to the Internet Engineering Task Force and NSF Network Working Group, June 1987.

This paper describes a system of pragmatic methodologies used in the DARPA/NSF Internet system to interconnect networks and network systems using different routing architectures.

[MIL87e] Mills, D.L., and H.-W. Braun. The NSFNET Backbone Network. *Proc. ACM SIGCOMM 87 Symposium* (Stoweflake Vermont, August 1987), 191-196.

The NSFNET Backbone Network interconnects six supercomputer sites, several regional networks and ARPANET. It supports the DARPA Internet protocol suite and DCN subnet protocols, which provide delay-based routing and very accurate time-synchronization services. This paper describes the design and implementation of this network, with special emphasis on robustness issues and congestion-control mechanisms.

[MIL88a] Mills, D.L. Network Time Protocol (version 1) specification and implementation. DARPA Network Working Group Report RFC-1059, University of Delaware, May 1988. Also published as: Electrical Engineering Department Report 88-4-1, University of Delaware, July 1988.

Note: This document has been superseded by [MIL90c].

[MIL88b] Mills, D.L. The Fuzzball. *Proc. ACM SIGCOMM 88 Symposium* (Palo Alto, California, August 1988), 115-122.

The Fuzzball is an operating system and applications library designed for the PDP11 family of computers. It was intended as a development platform and research pipewrench for the DARPA/NSF Internet, but has occasionally escaped to earn revenue in commercial service. It was designed, implemented and evolved over a seventeen-year era spanning the development of the ARPANET and TCP/IP protocol suites and can today be found at Internet outposts from Hawaii to Italy standing watch for adventurous applications and enduring experiments. This paper describes the Fuzzball and its applications, including a description of its novel congestion avoidance/control and timekeeping mechanisms.

[MIL88c] Mills, D.L., C.S. Ih, C.G. Boncelet and P. Amer. A high-speed reservation-TDMA wide-area network. White paper submitted to Corporation for National Research Initiatives, 16 December 1988.

This is a white paper describing an architecture for a high-speed network and packet switch capable of operation in the gigabit range over distances spanning the country. It is intended for applications involving very high traffic rates typical of supercomputer and motion video applications. The network is assumed composed of fiber trunks, while the switch architecture is based on an inexpensive, bit-slice crossbar switch which can be configured for speeds from 25 Mbps to 1 Gbps by simply adding slices. The switches are externally controlled by a microprogrammed sequencer using a distributed reservation protocol and scheduling algorithm.

[MIL89a] Mills, D.L. Wiretap: an experimental multiple-path routing algorithm. *ACM Computer Communication Review 19, 1* (January 1989), 85-98.

This paper introduces Wiretap, an experimental routing algorithm which computes maximum-likelihood diversity routes for packet-radio stations sharing a common broadcast channel, but with some stations hidden from others. The wiretapper observes the paths (source routes) used by other stations sending traffic on the channel and, using a heuristic set of factors and weights, constructs speculative paths for its own traffic. The algorithm is presented as an example of maximum-likelihood routing and database management techniques useful for richly connected networks of mobile stations. Of particular interest are the mechanisms to compute, select, rank and cache a potentially large number of speculative routes when only limited computational resources are available.

A prototype implementation has been constructed and tested for the AX.25 packet-radio channel now in widespread use in the amateur-radio community. Its design is similar in many respects to the SPF algorithm used in the ARPANET and NSFNET backbone networks, and is in fact a variation of the Viterbi algorithm, which constructs maximum-likelihood paths on a graph according to a weighted sum of factors assigned to the nodes and edges.

[MIL89b] Mills, D.L. Network Time Protocol (Version 2) specification and implementation. DARPA Network Working Group Report RFC-1119, University of Delaware, September 1989. Also published as: Electrical Engineering Department Report 89-9-2, University of Delaware, September 1989.

Note: This document has been superseded by [MIL90c].

[MIL89c] Mills, D.L. Measured performance of the Network Time Protocol in the Internet system. DARPA Network Working Group Report RFC-1128, University of Delaware, October 1989. Also published as: Electrical Engineering Department Report 89-9-3, University of Delaware, September 1989.

See [MIL90a].

[MIL89d] Mills, D.L. Internet time synchronization: the Network Time Protocol. DARPA Network Working Group Report RFC-1129, University of Delaware, October 1989. Also published as: Electrical Engineering Department Report 89-9-1, University of Delaware, September 1989.

See [MIL90d].

[MIL90a] Mills, D.L. On the accuracy and stability of clocks synchronized by the Network Time Protocol in the Internet system. *ACM Computer Communication Review 20, 1* (January 1990), 65-75.

This paper describes a series of experiments involving over 100,000 hosts of the Internet system and located in the U.S., Europe and the Pacific. The experiments are designed to evaluate the availability, accuracy and reliability of international standard time distribution using the Internet and the Network Time Protocol (NTP), which has been designated an Internet Standard protocol. NTP is designed specifically for use in a large, diverse internet system operating at speeds from mundane to lightwave. In NTP a distributed subnet of time servers operating in a self-organizing, hierarchical, master-slave configuration exchange precision timestamps in order to synchronize host clocks to each other and national time standards via wire or radio.

The experiments are designed to locate Internet hosts and gateways that provide time by one of three time distribution protocols and evaluate the accuracy of their indications. For those hosts that support NTP, the experiments determine the distribution of errors and other statistics over paths spanning major portions of the globe. Finally, the experiments evaluate the accuracy and reliability of precision timekeeping using NTP and typical Internet paths involving ARPANET, NSFNET and regional networks. The experiments demonstrate that timekeeping throughout most portions of the Internet can be maintained to an accuracy of a few tens of milliseconds and a stability of a few milliseconds per day, even in cases of failure or disruption of clocks, time servers or networks.

[MIL90b] Mills, D.L., Schragger, P., and Davis, M. Internet Architecture Workshop: Future of the Internet System Architecture and TCP/IP Protocols. *ACM Computer Communication Review 20, 1* (January 1990).

The Internet Architecture Task Force (INARC) is a unit of the Internet Activities Board (IAB), which oversees the research, development and standards of the Internet community and TCP/IP protocol suite. This is a report on the INARC workshop held at the University of Delaware on 1-2 June 1989 at which the future of the Internet and its protocols, as well as their relevance to other protocol communities, was explored and debated.

[MIL90c] Mills, D.L. A comparison of the Network Time Protocol and Digital Time Service. Memorandum distributed to the Network Working Group, February 1990.

This memorandum examines the models, architectures and protocols used in NTP and DTS and compares them in timekeeping accuracy, stability and robustness.

[MIL90d] Mills, D.L. Network Time Protocol (Version 3) specification, implementation and analysis. Electrical Engineering Department Report 90-6-1, University of Delaware, June 1990.

This document describes the Network Time Protocol (NTP), specifies its formal structure and summarizes information useful for its implementation. NTP provides the mechanisms to synchronize time and coordinate time distribution in a large, diverse internet operating at rates from mundane to lightwave. It uses a returnable-time design in which a distributed subnet of time servers operating in a self-organizing, hierarchical-master-slave configuration synchronizes local clocks within the subnet and to national time standards via wire or radio. The servers can also redistribute reference time via local routing algorithms and time daemons.

[MIL90e] Mills, D.L. Internet time synchronization: the Network Time Protocol. *IEEE Trans. Communications* (to appear).

This paper describes the Network Time Protocol (NTP), which is designed to distribute time information in a large, diverse internet system operating at speeds from mundane to lightwave. It uses a symmetric architecture in which a distributed subnet of time servers operating in a self-organizing, hierarchical configuration synchronizes local clocks within the subnet and to national time standards via wire, radio or calibrated atomic clock. The servers can also redistribute time information within a network via local routing algorithms and time daemons.

This paper also discusses the architecture, protocol and algorithms, which were developed over several years of implementation refinement and resulted in the designation of NTP as an Internet Standard protocol. The NTP synchronization system, which has been in regular operation in the Internet for the last several years, is described along with performance data which shows that timekeeping accuracy throughout most portions of the Internet can be ordinarily maintained to within a few milliseconds, even in cases of failure or disruption of clocks, time servers or networks.