# Gigabit Network Communications Research

## Quarterly R&D Status Report No. 9

**Period: 1 OCT 92- 31 DEC 92**



*Sponsored by*
*Defense Advanced Research Projects Agency (DoD)*
*Computer Systems Technology Office*
*Gigabit Network Communications Research*
*Under Contract #DABT63-91-C-0001*
*PR&C: HR0011-0218-0001*
*AAP No. DAR1010*
*Issued by Directorate of Contracting*
*Fort Huachuca, AZ*

# GIGABIT NETWORK COMMUNICATIONS RESEARCH

## QUARTERLY R&D STATUS REPORT for DABT63-91-C-0001

## (covering the period 10/1/92-12/31/92)

# 1. ATOMIC

The ATOMIC project utilizes inexpensive fine-grain multicomputer components to create a multi-gigabit per second local area network. This unique approach leverages recent DARPA supported research at Caltech by Prof. Charles Seitz and his students. That research produced dense and extremely fast Mosaic single-chip multicomputer nodes. ATOMIC is a unique example of network development that grows upwards from multicomputer machine networking into the domain of a LAN.

The principal concerns for this quarter were performance improvement and testing of the SPARCstation ATOMIC SBus card and modification to allow multicast.

## 1.A. SOFTWARE

To create a new LAN requires a large amount of software to be developed. We continue to make extensions to the BSD UNIX kernel to enhance the operation of the LAN.

We have brought up SBus host interfaces and Address Consultant (AC) code. Much time was spent dealing with the more stringent alignment requirements of the SPARC processors vs. the earlier 68000 series processors of the Sun 3/xx model machines. We believe that all code now is converted to be 64-bit aligned. Theoretically, we won't have problems if we later need to move to a 64-bit machine.

A large amount of debugging and error-checking code was added to the AC and Mosaics to facilitate problem fixes and implementation of new features. We added a stack checking routine to the ATOMIC code. If the stack overflows, hosts outputs a pstring and switch nodes flash the LED at a faster than normal rate. Checksumming was added also.

Steve Deering's IP multicast was integrated into our ATOMIC Sun OS 4.1.1 kernel. Thus ATOMIC has been extended to provide broadcast and multicast services. This is a simple implementation and probably will be changed to look more like the Ethernet implementation for storage of multicast addresses etc.

Working with Bob Braden and Liming Wei, the Transaction TCP is now up and running within the ATOMIC LAN.

### *1.A.1 Performance Testing of SBus Interface*

We have tested the speed of SBus interface. As we expected, it is exactly 80% of the VME-version ATOMIC host interface performance. That's due to 20Mhz SBus clock frequency versus the 25Mhz clock rate used on the VME board.

TCP and UDP on ATOMIC now provide performance in the 17Mb/s range. That is about double the Ethernet performance. This is reaching the limit of the 'standard' TCP/IP implementation within our current Sun–class machines. Hand–coding key loops and use of jumbo TCP/IP packets would improve performance.

## 1.B. HARDWARE

Meetings between our group and the Mosaic staff at Caltech took place regarding Slack/Dialog schedules. Additional discussions covered longer channels and fiber optics.

Three key areas were discussed and are being developed:

### *1.B.1 Cable Technology*

SLACK has been adequately demonstrated as far as Caltech/Mosaic staff is concerned. They will develop a follow–on to SLACK, called Dialog, that will also incorporate fault detection and recovery logic. The Dialog chips will be used by the ATOMIC project in its initial implementation of 30 meter and longer copper–cable Mosaic channels.

When ATOMIC fiber–optic cable development begins, the HP HDMP 1002/1004 series 800 Mb/s encoder/decoder chips are the best current choice. In keeping with the low data–link layer overhead of ATOMIC/Mosaic, these parts emulate virtual ribbon cables and are ideally suited to carry Mosaic channel data. The obvious alternative is to use Fibre Channel parts. These would require alteration of the ATOMIC/Mosaic packets to meet Fibre Channel specifications and that is unattractive, as the necessary pocket translation would entail a great deal of overhead.

Finisar laser drivers have been determined to be the best current choice for discrete laser transmitter/receiver chips. Finisar avoids Class–I OFC requirements by transmitting at lower power levels over multimode fiber. As a result this transmitter/receiver combination is limited to 500 meters.

Lasertron Corp. makes a $5,000 module that contains the HP HDMP 1002/1004 parts and their own transmitter/receiver. They can transmit up to 10 kilometers using 1300nm single–mode fiber and include a Class–I OFC safety circuit, which is required by FCC regulations for office installation.

### *1.B.2 General Host Interface Design*

Caltech and ISI now concur that a DMA engine that copies ATOMIC packets to/from a host system bus is the most realistic model for a series of host interface designs that would share a common interface design. The only section of hardware to change from machine type to machine type would be the DMA engine interface to the particular system bus.

Although a DMA copy engine is not the highest performance alternative, it seems to be a reasonable performance/cost compromise.

*1.B.3 Network Reliability*

Cooperation with the Caltech design staff is leading to the incorporation of network reliability enhancements for the Mosaic–to–channel interface and ATOMIC mesh router–to–network interface. Definition of these features continues and they will be incorporated and tested in upcoming Dialog chips created by Caltech.

Desired features are the ability to reset a Mosaic channel to recover from a blocked path and the ability to isolate an ATOMIC mesh router from the network when it is being loaded or diagnosed.

## 2. Personal Conferencing

It is crucial that teleconferencing and telecollaboration be supported across a variety of session modes and across a wide scale in several dimensions including session size, population size, and geographic distribution. Our past work on connection/session management has concentrated on sessions with a relatively small number of participants wherein tight control is feasible to enable features such as authentication and confidentiality. We developed a connection management architecture and the Connection Control Protocol (CCP) for this mode of operation.

However, the increasing popularity of audio/video multicasts of IETF meetings across the Internet clearly demonstrates the need to also support larger scale sessions with loose control. We have explored some of these issues in the paper, "The Impact of Scaling on a Multimedia Connection Architecture", presented by Eve Schooler at the Third International Workshop on Network and Operating System Support for Digital Audio and Video to be held in San Diego, CA.

Our platform for testing connection management protocols, including CCP, is the multimedia conference control program (MMCC). Our original implementation was under the SunView window system, but conversion to X windows is a practical requirement for continued development and use. This quarter we have made a preliminary conversion to XView except for a few pieces of functionality that don't convert directly.

We are participating in IETF working groups to develop broader solutions for telecollaboration across the Internet. At the November IETF meeting, Eve Schooler led two sessions on Conference Control. The aim of these discussions was to understand how a new working group on the topic might contribute to the remote conferencing architecture effort. It was agreed that there is a need for a session layer control protocol to perform higher layer functions than the transport protocol proposed in the Audio/Video Transport (AVT) WG chaired by Steve Casner. The beginnings of design criteria for this protocol were identified.

In the AVT WG, a draft specification for the Realtime Transport Protocol (RTP) was presented, based on previous WG discussion and substantial email discussion between

us and the primary author, Henning Schulzrinne. The working group reached consensus on most of the open issues, and produced a list of changes that have been incorporated into the specification. It has been released as a collection of Internet–Drafts.

One of the functions of RTP is to provide synchronization between transmitters and receivers of realtime media streams, including synchronization among several streams. At ISI, we are working with BBN to determine how the Synchronization Protocol they designed should be incorporated into our multimedia teleconferencing programs, and how it impacts both RTP and connection management.

We were involved in several trial demonstrations of a distributed performance over DARTnet of multi–part music synchronized using the protocol. ISI served as an endpoint for a distributed music demonstration; we performed one of the instrument parts of a Haydn trio in realtime and provided feedback about the sound quality and synchronization accuracy.

The November IETF meeting was the third meeting to be "audiocast" and the second to include video. This time we transmitted two simultaneous channels of audio and video during some of the working group breakout sessions, but had to cut back to a single channel when it was determined that the high load of multicast traffic was causing some backbone network nodes to crash. This time the IP multicast network supporting the transmission was built up in a much more coordinated manner compared to the ad–hoc collection of multicast tunnels used in July. We have played a major role in the coordination and construction of this virtual multicast backbone, dubbed MBONE, to support future IETF audio/videocasts and other experiments with an even larger number of participants and countries. This volunteer effort has been very effective in distributing the workload of supporting the IETF audiocast to a larger group of people at regional networks and participating sites.

In early November, we set up a demonstration for the ARPA Technology Council of the audio and video technology used in the IETF transmissions.

## 3. Integrated Services Protocols

During this quarter, we have continued to monitor, and where possible to facilitate, the research effort to develop an integrated service architecture for the Internet. The components of this problem are those laid out in our October 1991 paper in the High–Performance Network Research Report: resource model and flow specification, traffic control mechanism, admission control algorithm, and classifier. The research community represented by DARTnet and the End–to–End Research Group has made significant progress on these tasks.

These issues were extensively discussed by the major players during a meeting of the End–to–End Research Group held at MIT in October 1992. It became apparent that the two major divergent views on traffic control (represented by the Clark, Shenker, Zhang model of Guaranteed and Predicted service, and the Jacobson, Floyd model of

hierarchical link sharing) were converging towards a common point. We expect rapid progress towards a single traffic control model during the next quarter, with a demonstration for ARPA.

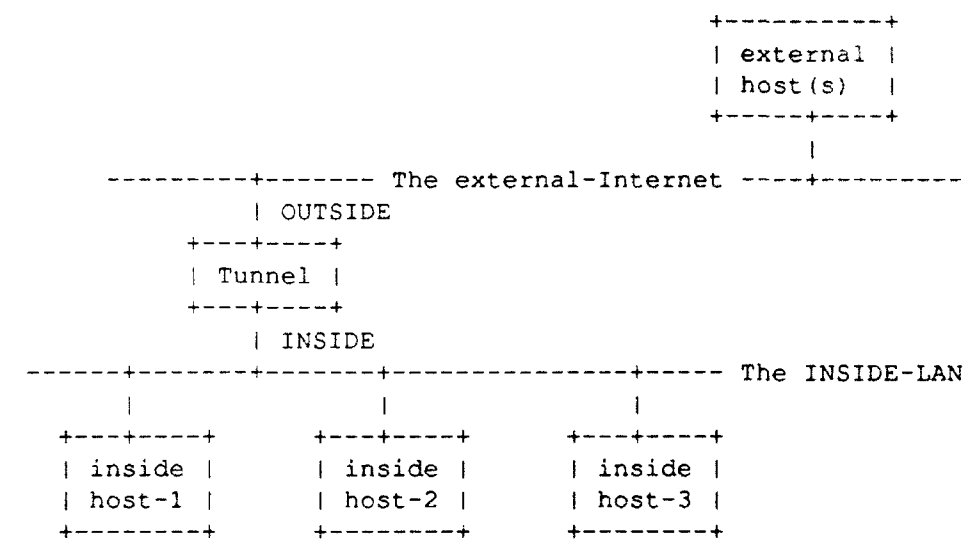
Also at this meeting, significant progress was made towards designing a "soft state" setup protocol. This work, now titled "RSVP", has been the subject of intensive effort at Xerox PARC by Zhang, Shenker, and Jamin. ISI has submitted a proposal for Estrin (USC, on sabbatical at ISI) and Braden to collaborate with Xerox PARC on this work.

## 4. IP/SQ Congestion Control

Conversion of IP/SQ code to reside in the new DARTnet kernel environment occurred during the first part of December. This was more difficult than had been anticipated. It has been completed and clears the way for debugging and testing with the improved round–trip estimator. Tests should take place in January or February of 1993.

## 5. Tunnel

The Tunnel is a "diode–gateway" that allows smooth seamless initiation of communication from the INSIDE to the OUTSIDE, and while keeping explicit access control on communication initiated by the OUTSIDE, to the INSIDE.




The operation of the tunnel is as follows:

OUTWARDS: The tunnels allows inside users (i.e., on the INSIDE–LAN) to implicitly initiate communication with external hosts (i.e., on the external Internet), without any additional effort. This communication may support any IP–based communication (such as with TCP or UDP), and is not limited to Telnet connections only.

INWARDS: However, external users (such as local personnel on travel, using external hosts) must use explicit access control to initiate communication with internal hosts.

This is implemented by the authorized user first using Telnet to the Tunnel, then logging into it to create a "visa" to allow direct communication (of any IP traffic) between the external host and a specific internal host. Hence, in order to have direct communication with the inside, the external user must have access privileges (e.g., a password), on the Tunnel. This procedure supports any IP–based communication including applications that cannot run over Telnet such as remote X–windows, FTP, packet–audio, and NeXT/NeXT communication.

This quarter we received the GFE equipment, and completed most of the basic functionality of the Tunnel. We have a working prototype installed in a lab at ISI.

Next quarter we will be working on the utility programs that are required for the operation and the management of the Tunnel, and we will be adding the logging functionality to the Tunnel software.

## 5. Automated Cluster Teleconferencing (ACT)

The Touring Machine (TM) has been installed at ARPA. The substitution of the Confertech audio bridge with a less expensive TEAC unit proved successful. There have been some difficulties with faulty switching equipment that are currently being addressed in coordination with the switch manufacturer.

The system is currently verified in point–to–point mode only, and is limited to 8 users and one teleconferencing bridge until the hardware faults in the larger, 20x20 switch have been corrected. Further tests in early January 1993 will verify the availability of the teleconferencing (multipoint) mode.

A multi–domain extension for the Touring Machine has been designed, including a strategy for minimal (no) alteration of the Bellcore system, with a maximum of functionality. It will support both interoperation of separate Touring Machines and connection with heterogeneous remote services.

The next quarter's plans will be determined after the hardware faults are corrected and a project evaluation has taken place. Possible plans include implementation of the multi–domain extension, adding external fault–tolerance and compensation mechanisms, evaluating digital transport methods, and determining the expected lifetime of the current system software.

During the next quarter, a document will be prepared describing ISI's contribution to the extension of TM, in the implementation of testbed–independent teleconferencing. Bellcore's TM installation provides teleconferencing using a proprietary interface, complicated back–end control software, and expensive bridging equipment; ISI has developed a replacement teleconferencing module for the TM that uses inexpensive hardware and simple software.

The next quarter will also include the documentation of a new multipoint multimedia model (currently called M3) that integrates components of the Bellcore TM and ISI

MMCC models, and is augmented to address failure mode considerations emanating from ACT project feedback.

Equipment for DARTnet teleconferencing (for ACT project management use) has been installed at ARPA. The ARPA router is up, and current with the DARTnet configuration (see DARTnet summary).

## 7. DARTnet Network Operations Center (DARTNOC)

ISI continues to schedule test times for experimenters.

ISI purchased sufficient disks to put one 200 MB disk or two 100 MB disks on each DARTnet router. Disk space on the NASA–Ames, Anaheim–POP (A.K.A. LA–POP), DC–POP, and ARPA has been expanded or installed to 200 MB. These disks are configured with two root partitions. The routers with restricted physical access (NASA–Ames and the POPs) have been equipped with two disk drives and are configured so that they can be booted from either drive. Thus, if the primary drive should fail, the machine can still be booted remotely from the other drive, and service can be restored without local access.

The other reason for a second root partition is to support the use of another operating system in parallel with SunOS. We had planned to use 4.4 BSD Unix for this purpose, although at the present time prospects for the availability of 4.4 BSD Unix do not seem good.

A new base system was prepared from SunOS Release 4.1.2, by deleting the many Sun modules that are not required on a router — window support, compilers, etc. This system was installed on NASA–Ames, Anaheim–POP, DC–POP, and ARPA routers. Work will soon begin on propagating this new base to all the routers, with a uniform configuration of disk partition sizes. At present, many of the on–site routers have small root partitions, which causes operational difficulties.

New DARTnet SunOS kernels have been built with fixes for some problems plus added features for clock synchronization, IP encapsulation, S–bus expansion, additional BPF channels, raw bytesync support for video codecs, and others. All source file modifications were logged with RCS. Source and object trees were updated for distribution to DARTnet experimenters.

ISI brought up DARTnet II using the new carrier, Sprint. DARTnet II has three new sites, ARPA, BellCore, and Sun. Part of the process of bringing up DARTnet II included installing an S–bus expansion chassis at Ames and DC and removing the expansion chassis from the LA–POP. Two of the CSU/DSU's with older firmware used in DARTnet were determined to be incompatible with Sprint's network. One of the CSU/DSUs was swapped with the CSU/DSU on the MIT microwave link. The other one was replaced with a spare CSU/DSU that was being used for testing at LBL.

Bringing up DARTnet II on the new carrier took much longer than originally anticipated. Sprint did not have the circuits operational when they said that they would. They also did not run the tests they said they would nor provide female RJ45 jacks until after ISI arrived to install the equipment. Of the 10 circuits Sprint provided, only three worked initially and remained operational during the installation phase.

There have been three outages since the installation of the network. Sprint was able to fix all of the problems with a minimum of hassle. Repairs were completed within a couple of hours. This is a good sign for future operational stability and maintainability.

# 8. Infrastructure

## 8.A. USER SERVICES

As the USAC Chair, Joyce Reynolds participated in IESG Teleconferences from October – December 1992 and attended the IETF meeting in Washington, D.C., November 16–20, 1992.

Eleven working groups in the User Services Area of the IETF met in Washington, D.C. One BOF (Birds of a Feather) was held in the User Services area regarding a working group formation on Training Materials.

During this period, one new Working Group was formed in the User Services Area of the IETF: Network Training Materials (trainmat).

There are currently 14 active Working Groups in the User Services Area of the IETF.

## 8.B. INTERNET MONTHLY REPORT

The Internet Monthly Report (IMR) is the status report on the operation of the Internet and the research and development activities of the Internet community. It features reports from the IAB, the Internet Research Task Force and its research groups, and the Internet Engineering Task Force and its working groups in addition to the reports from approximately 30 regional networks and individual sites. A typical monthly report is approximately 40 pages.

During this reporting period, three Internet Monthly Reports for September 1992, October 1992, and November 1992 were assembled, edited, and distributed directly (via electronic mail) to over 375 mailboxes, some of which are exploder mailboxes where the report is sent to a sublist of people. In particular, the mailbox "IETF@isi.edu", which is one of the mailboxes on the IMR list, goes out to an additional 935 mailboxes, many of which are further exploders.

## 8.C. HIGH PERFORMANCE NETWORK RESEARCH REPORT

The High Performance Network Research Report (HPNRR) discusses research and development activities in the Gigabits program and the advanced networking research

community. A typical report is about 25 pages. During this reporting period, three reports for September 1992, October 1992, and November 1992 were assembled, edited, and distributed directly (via electronic mail) to over 130 people.

## 8.D. REQUEST FOR COMMENTS

ISI serves as the technical editor and "publisher" of the Internet document series called "Requests for Comments" (RFCs). 20 RFCs were published this quarter:

RFC 1366: Gerich, E., "Guidelines for Management of IP Address Space", Merit, October 1992.

RFC 1367: Topolcic, C., "Schedule for IP Address Space Management Guidelines", CNRI, October 1992.

RFC 1368: McMaster, D. (Synoptics Communications, Inc.), K. McCloghrie (Hughes LAN Systems, Inc.), "Definitions of Managed Objects for IEEE 802.3 Repeater Devices", October 1992.

RFC 1369: Kastenholz, F., "Implementation Notes and Experience for The Internet Ethernet MIB", FTP Software, October 1992.

RFC 1370: Chapin, L. (IAB, Chair), "Applicability Statement for OSPF", October 1992.

RFC 1371: Gross, P.,(IETF/IESG Chair), "Choosing a "Common IGP" for the IP Internet (The IESG's Recommendation to the IAB)", October 1992.

RFC 1372: Hedrick, C. (Rutgers), and D. Borman (Cray Research, Inc.), "Telnet Remote Flow Control Option", October 1992.

RFC 1373: Tignor, T., "Portable DUAs", USC/ISI, October 1992.

RFC 1374: Renwick, J., and A. Nicholson, "IP and ARP on HIPPI", Cray Research Inc., October 1992.

RFC 1375: Robinson, P., "Suggestion for New Classes of IP Addresses", Tansin A. Darcos & Co., October 1992.

RFC 1376: Senum, S., "The PPP DECnet Phase IV Control Protocol (DNCP)", Network Systems Corporation, November 1992.

RFC 1377: Katz, D., "The PPP OSI Network Layer Control Protocol (OSINLCP)", Cisco, November 1992.

RFC 1378: Parker, B., "The PPP AppleTalk Control Protocol (ATCP)", Cayman Systems, November 1992.

RFC 1379: Braden, B., "Extending TCP for Transactions — Concepts", USC/ISI, November 1992.

RFC 1380: Gross, P. (IESG Chair), and P. Almquist (IESG Internet AD), "IESG Deliberations on Routing and Addressing", November 1992.

RFC 1381: Throop, D. (Data General Corporation), and F. Baker (Advanced Computer Communications), "SNMP MIB Extension for X.25 LAPB", November 1992.

RFC 1382: Throop, D., Editor, "SNMP MIB Extension for the X.25 Packet Layer", Data General Corporation, November 1992.

RFC 1383: Huitema, C., "An Experiment in DNS Based IP Routing" INRIA, December 1992.

RFC 1385: Wang, Z., "EIP: The Extended Internet Protocol A Framework for Maintaining Backward Compatibility", University College London, November 1992.

RFC 1386: Cooper, A., and Jon Postel, "The US Domain", USC/ISI, December 1992.

### 8.E. VISITORS

Michael StJohns visited ISI to discuss the future of Gigabit Network Communications Research.

### 8.F. TRAVEL

Walt Prue went to Washington, DC to install DARTnet equipment, October 7–9, 1992.

Steve Casner visited Sun Microsystems in San Jose, October 14, 1992.

Bob Braden chaired the End-to-End Research Group meeting at MIT, October 14–18, 1992.

Danny Cohen attended a Telemedia Review in Cambridge, MA, October 26–27, 1992.

Bob Braden , Joyce Reynolds, Peter Will, and Jon Postel attended Interop '92, October 28–30, 1992, in San Francisco, California. Joyce Reynolds was a session leader/speaker, Bob Braden gave a talk, and both Bob and Jon Postel attended IAB meetings.

Greg Finn and Danny Cohen attended meetings at Digital Equipment Corporation in Palo Alto, November 12, 1992.

Bob Braden, Steve Casner, Eve Schooler, Jon Postel, Peter Will, and Joyce Reynolds, attended IETF meetings November 17–22, 1992 in Washington D.C. Bob Braden and Jon Postel attended IAB meetings held at the IETF.

Joe Touch installed ACT equipment at DARPA in Washington, DC, and then gave a presentation in Philadelphia, November 16–30, 1992.

Danny Cohen travelled to Los Alamos National Labs to present a seminar on ATOMIC, December 3–4, 1992.

Danny Cohen met with DARPA officials and NeXT personnel to discuss future collaboration, December 6–8, 1992.

## 8.G. SEMINARS

EVE SCHOOLER gave a seminar on "The Impact of Scaling on a Multimedia Conferencing Architecture". As the last two meetings of the Internet Engineering Task Force (IETF) have shown, Internet teleconferencing has arrived — whether or not we are ready for it. Packet audio and video have now been "mediacast" to approximately 170 different hosts in 10 countries, and for the November meeting the number of remote participants is likely to increase by a factor of ten. Yet, the underlying technology to support wide scale packet teleconferencing is barely in place.

Eve lead led a working discussion on the impact of scaling on our efforts to define a teleconferencing architecture. Three scaling issues of particular interest include: scaling up in size to support (i) very large numbers of participants, (ii) many simultaneous teleconferences, and (iii) a widely dispersed user population (i.e., inter–domain). Eve described ongoing work in this area, the pieces that are missing, and the beginnings of options for solutions. Also discussed were multicast addressing concerns, techniques for bandwidth reduction, session management, conference directory services, heterogeneity, and system performance and robustness.

GREG FINN gave a working talk on "Issues in Internetwork Addressing and Routing for Many–Address Workstations" (i.e. if a "host" has many IP addresses ... possibly a dynamic set of them ... then how do we talk to it). What affect will such hosts have on the Internet addressing and routing ... how might one wish to incorporate them into the Internet without breaking it ... and so on.

JOE TOUCH gave a seminar on the topic, "Parallelizing Protocols: Do's and Don'ts". As communication rates increase into the gigabit range, there is increasing concern about the capability of existing protocols to keep pace. Many similar bottlenecks are alleviated by the use of parallelism, so one hypothesis is to "parallelize" protocols. In this talk, Joe discussed the pros and cons of this hypothesis. He examined the dimensions to which parallelism might be applied, and distinguish the unique communication issues that result. In conclusion, Joe indicated that conventional parallelism techniques may not be applicable to protocols. New techniques, sometimes considered unconventional, become more significant in this light. These include information parallelism (Parallel Communication) and subsuming the workstation into the network (NetStation). Parallel Communication was also discussed briefly.

## 9. Publications, papers, and presentations

PAPERS

Tignor, T., "Portable DUAs", USC/ISI, RFC 1373, October 1992.

Braden, B., "Extending TCP for Transactions — Concepts", USC/ISI, RFC 1379, November 1992.

Cooper, A., and Jon Postel, "The US Domain", USC/ISI, RFC 1386, December 1992.

We submitted a paper on the novel approach taken in creating the ATOMIC LAN to the *Journal of High-Speed Networks.*

A paper entitled "Communication Parallelism" (J. Touch, ISI) was accepted to IEEE InfoCom '93. A letter to IEEE Communications Magazine, regarding an article on gigabit network protocol issues, will be printed in the February 1993 issue.

PRESENTATIONS

Joe Touch presented his paper on "Physics Analogs in Communication Models" at the Physics of Computation Workshop in Addison, Texas on October 10, 1992.

Danny Cohen gave a seminar on ATOMIC at Los Alamos National Lab.

Danny Cohen and Gregory Finn gave an oral presentation to Digital Equipment's Systems Research Center.

Bob Braden gave a talk on Internet Integrated Service at Interop '92.