UNIVERSITY OF SOUTHERN CALIFORNIA *INFORMATION SCIENCES INSTITUTE*
4676 Admiralty Way Marina del Rey, CA 90292

Netstation Architecture
and Advanced ATOMIC Network
Final Report

Period of performance: 1 OCT 93 - 31 DEC 97

Sponsored by
Defense Advanced Research Projects Agency (DoD)
Information Technology Office
Netstation Architecture and Advanced ATOMIC Network
Under Contract #DABT63-93-C-0062
PR&C: HJ1500-3215-0588
AAP No. DAR3A698
Issued by Directorate of Contracting
Fort Huachuca, AZ

NETSTATION ARCHITECTURE AND ADVANCED ATOMIC NETWORK

FINAL REPORT for DABT63-93-C-0062

(covering the period 10/1/93-12/31/97)

Project Leaders: Gregory Finn, Joseph Touch
Theodore V. Faber, Steve Hotz, Rod Van Meter, Anne Hutton,
Annette DeSchon, Bob Felderman, Paul Mockapetris, Hong Xu
Graduate Students: S. K. Munnangi, Vivek Goyal, Avneesh Sachdev, Tom Fisher,
Reza Rejaie, Wei Yue, Stephen Suryaputra, Simon Walton, Darshan Jani, Nehal Bhau
ISI Computer Networks Division

Bruce Parham, Mike Gorman, Jerry Wills, Jeff LaCoss
ISI Integrated Systems Laboratory

1. Introduction

The NAAAN project is composed of two major tasks - Netstation and Advanced Atomic Networking (the latter informally called ATOMIC-2). The Netstation task concerns itself with issues that arise when a gigabit network is substituted in workstations in place of a system bus. The ATOMIC-2 task focuses on generalizing the ATOMIC gigabit LAN technology, so that it interoperates with workstations and other networks in much the same way that Ethernet-based LANs do.

The intention of the Netstation task was to demonstrate that a workstation can be architected around a set of networked devices that inhabit a gigabit cyberspace, freeing the workstation from bus-induced configuration limitations. Toward that end stand-alone network devices and interfaces were prototyped, device control protocols were studied and developed, and performance limits were explored in the context of the ATOMIC/Myricom gigabit LAN technology.

The Netstation task was revised to focus more directly on technology transition. This included rearchitecting the prototypes to use commercially standard interfaces and internetwork protocols. The intention was to make the concept commercially practical, to demonstrate that network peripheral devices could be directly and safely controlled via the internetwork and to make possible widespread third-party device control and data transfer. Widened access however required development of and demonstration of a methodology that could ensure the integrity of internet-accessible peripheral devices. The resulting integrity protection methodology was successfully developed and tested in the prototypes.

The ATOMIC-2 task's original goal was to develop and replicate production-grade instances of the prototype gigabit ATOMIC LAN. Myricom, a commercial company marketing ATOMIC LAN components, subsumed the task of developing a production-grade system from research prototype components, both in hardware and software. The ATOMIC-2 task was revised to focus on bottlenecks that remain after a production ATOMIC LAN is installed. These include an fast ethernet and ATM gateway, a high-performance networked file server, and high-speed authentication. It also includes issues of service provision and application performance enhancement.

2. Background

The ATOMIC LAN is a gigabit network developed from the Mosaic mesh supercomputer chips developed at the California Institute of Technology (CalTech) [2]. The Mosaic chips included a processor and a small routing and data transfer module, the latter to support their mesh interconnection. In 1992-3, Mosiac was used by a joint project of CalTech and USC/ISI, to implement a prototype LAN called ATOMIC (originally meaning “ATM Over MosaIC”). A simple 3x3 switch and two variants of Sun host interfaces were developed, using byte-wide copper ribbon cable for 3-6’ links, resulting in a lab-sized proof of concept. This was later refined to an 8x8 switch, and cable drivers were added to support links as long as 100’. In late 1993, principles in the ATOMIC project left both CalTech and USC/ISI to form Myricom, a private effort to commercialize the ATOMIC technology, which they called “Myrinet” [29]. Myricom is aimed at the commercial network of workstation / cluster of workstation (NOW/COW) market. The NAAAN project is a simultaneous effort to examine the use of ATOMIC technology to develop production LAN services and to replace the backplane of a host computer (see the timeline in Figure 1).

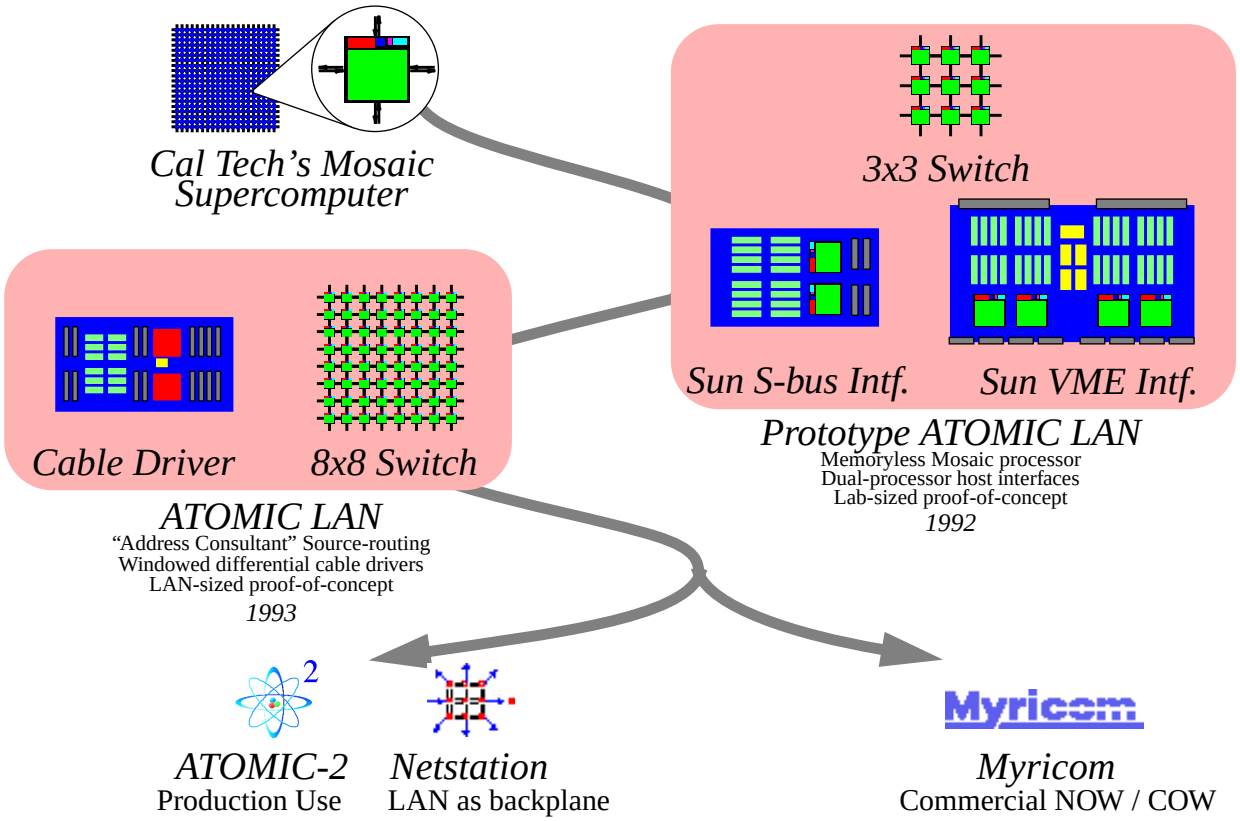

Figure 1 Evolution of the ATOMIC into Netstation, ATOMIC-2, and Myricom

3. Netstation

The objective of the Netstation Architecture effort was to investigate what happens when the system bus in a workstation is replaced by a high-speed local area network. To answer that question required investigation and software development on a broad front that covered five major areas:

- Protocols
- Performance
- Prototyping and Proof-of-Concept
- Integration and Technology Transfer
- Peripheral Device Integrity

The Netstation effort progressed through two major phases. In the first phase work concentrated on demonstration of an intra-LAN proof-of-concept; the focus of which was performance. Research, prototyping, software and protocol development focussed on determining the practicality of the Netstation concept using the ATOMIC/Myricom gigabit network to connect a workstation to peripherals.

During the second phase, work concentrated on extending the proof-of-concept architecture to the unrestricted commercial networking environment of the Internet. This implied first, that standard protocols must be used, and second, that the network-attached peripherals (NAPs) of a workstation could be accessed by sources outside its network. Developing a methodology for ensuring NAP integrity thus became a new and paramount concern.

Protocols

A Netstation architecture splits a workstation apart from its major peripheral devices. Whereas communication and control over devices in a bus-oriented traditional architecture is centralized, a Netstation architecture is decentralized. Its devices must be controlled via packets sent across a network.

The reliance upon networking to control devices implies that devices become clients of some external owner or owners. That in turn implies that a protocol must exist between the client device and its owners. The design and requirements of such protocols was an early and major topic of investigation.

Performance

Breaking the workstation architecture apart and connecting its various pieces via a high-speed network necessarily affects a workstation’s performance. Separating a device from its controlling owner increases latency between the issuing of a command to a device and its response. This raises two issues: (1) can devices be effectively used when device command/response latency is increased, and (2) what techniques can be employed to reduce end-to-end latency.

While the latency due to distance is a fixed property of separation, latencies introduced by protocol or operating system architectures, either in the device or in its owner, can be studied and reduced. The design of low-latency transport protocols and low-message-latency operating system architectures was a major early topic of investigation.

Prototyping and Proof-of-Concept

To demonstrate that the underlying Netstation architectural concept was sound required development of a prototype for experimentation and performance characterization. To make possible the evaluation of different networks and peripheral devices, a generic motherboard was designed and fabricated, as were several peripheral device interface cards for the motherboard. Creating the operating system, protocols and device control software for the attached peripherals was a major topic of effort.

Integration and Technology Transfer

Once the underlying architectural concept of a Netstation is shown to be practical, the focus of effort shifted somewhat toward how best to integrate these results and transfer them to industry. Toward that end project staff became active in the IEEE community studying and developing standards for mass-storage network-attached peripherals (NAPs).

A key obstacle to the greater use of NAPs is a reliance upon link-layer specific transport protocols. These provide best performance, but also condemn NAPs to remain isolated on their particular networks. Removing that isolation through adoption of standardized Internet protocols became a major topic for the project.

The mass-storage community has believed that Internet protocols are too overhead intensive. Therefore, the Netstation prototype was reworked to employ only Internet protocols and the resulting performance behavior of its prototype peripherals was characterized to highlight how well NAPs that used Internet protocols performed.

Peripheral Device Integrity

A consequence of using Internet protocols for NAPs is that they become widely accessible. This can threaten device access control and integrity, but it also makes widespread third-party data transfer capability. A major topic of investigation of the project was to develop a model and methodology for protecting NAP access control and integrity while simultaneously allowing safe third-party transfers.

3.a. Phase One

The first phase work of Netstation concentrated on demonstration of an intra-LAN proof-of-concept; the focus of which was performance. Research, prototyping, software and protocol development focussed on determining the practicality of the Netstation concept using the ATOMIC/Myricom gigabit network to connect a workstation to peripherals.

The primary accomplishments in pursuit of this proof of concept of performance included:

- Protocol Performance Evaluation: Included implementation and analysis of both IP-level and our custom ATOMIC/Myrinet transport-level Device Transport Protocol (DTP) protocol.
- X-Server Evaluation and Development: Included analysis of X-server implementation structure for decomposition, evaluation of data transfer requirements to/from non-resident frame buffer, and resulting design and implementation of Netstation X-server.
- Display Device API: Included specification of a remote frame buffer access API, analogous to and compatible with SCSI-III family of commands.

- Display Hardware Development: Included design and development of display motherboard, network interface card, Myricom cable driver card, frame buffer card and flash memory card.
- Demonstration of Display System: Demonstrated Netstation X-server running on Sun Workstation, controlling the remote Netstation Display frame buffer over our custom DTP transport protocol running on ATOMIC/Myrinet.
- Zero-Pass Embedded IP Checksumming: Included identification of remaining performance bottlenecks for high-performance device communication, and design of a mechanism that completely eliminated the problem, while addressing issues of backward compatibility and cost of implementation.

Discussion

The first phase emphasis on performance focused primarily on communication at the network, transport, and application layer; specifically, Netstation was concerned with reducing per packet delay. At the network layer, our design and high-performance implementation of IP over Myrinet achieved a rate of over 60,000 packets/second, which corresponded to a very low delay of 16 microseconds per packet. The custom DTP transport layer achieved a similar per packet delay to provide reliable transaction streams of over 30,000 transactions per second. The round trip delay of 33 microseconds between two Sun workstations on a Myrinet was faster than any similar efforts at the time. The concern for application layer performance was whether the data transfer requirements for a network attached frame buffer could be supported by the per-packet communication rates; we considered the frame buffer as the target device to be supported due to the performance demands for updating large window frames. Our profile evaluation of an X-Server under several typical usage conditions showed that the communication stack must support packet rates of up to 2000 packets/sec (or 1000 acknowledged transactions), and throughput rates of up to 2 Mbits/second. Our conclusion was that the Netstation architecture was, in fact, feasible.

The hardware development for Phase-I concentrated first on creating a general purpose motherboard that would be connected to a gigabit network and to which a variety of prototype peripheral devices could be connected. The motherboard included a TI 320C80 DSP/CPU as its core processor, a 64-bit memory bus, support for DRAM, SRAM and Flash memory, little/big endian conversion, 16- and 64-bit network interface paths, twin 16-bit D-to-A converters, and ISA bus support. Because a display frame buffer was deemed the most demanding peripheral to demonstrate as a Netstation device, a 32-bit per pixel color frame buffer and display controller card was developed as the first peripheral to be attached to the motherboard. The network initially chosen to demonstrate the Netstation Display was the ATOMIC/Myrinet gigabit network. A network interface card to attach that network to the motherboard was created along with a necessary cable driver card.

The demonstration of this phase of Netstation brought all of the efforts, thus far, together. In this demonstration, the Netstation X-server running on Sun Workstation controlled the remote Netstation Display frame buffer running on our custom hardware. Control between the two systems used our frame buffer command API transferred across the Myrinet LAN, using our custom DTP transport protocol. The Netstation X-server interface ran at a pace that was comparable to a local Xserver display. The one area where performance was noticeably slower, but acceptable, was for scrolling of large regions. This demonstration system did not include explicit support for making a single “scroll request” across the network, so performance was slowed by the need to make multiple individual frame buffer changes for each pixel of vertical scrolling. The fact that the display

performed acceptably with this high-level handicap further demonstrated the performance capabilities of the system.

The final development in this phase of the Netstation effort was both (1) a study to identify performance bottlenecks that would impede a more standardized (Phase II) Netstation system, and (2) a mechanism to alleviate what was perceived as the most critical bottleneck: IP checksumming. As the project changed its focus from performance to issues of shared access control running in a wide area (i.e. TCP/IP) environment, the first issue addressed was whether performance would suffer. The performance degradation for "touching" all of a packet payload that is needed to calculate the IP checksum was well known, and the area identified as the most critical bottleneck. The Netstation group developed a completely backward-compatible mechanism for calculating and inserting an embedded IP checksum, and documented and introduced this mechanism to the networking community.

3.b. Phase Two

The second phase of Netstation concentrated on extending the proof-of-concept architecture to the unrestricted commercial networking environment of the Internet. This implied first, that standard protocols must be used, and second, that the network-attached peripherals (NAPs) of a workstation could be accessed by sources outside its network. Developing a methodology for ensuring NAP integrity thus became a new and paramount concern.

The primary accomplishments in pursuit of providing standard protocols and shared resource protection in the internetworking environment included:

- Derived Virtual Device (DVD) Abstraction and Implementation: Included design and development of methodology for providing safe shared access to internetwork attached devices, and a prototype implementation in the Phase-II Netstation Display system.
- Netstation Kernel and Communication Stack: Included a multitasking communication-oriented kernel, a tuned TCP/IP network stack, Ethernet device driver, PCI bridge device driver for the Netstation motherboard.
- Phase-II Display Hardware: Include PCI bridge, JPEG daughter card, burst-mode frame buffer (note: only the PCI bridge hardware was used in final demonstration).
- Implementation of VISA adaptation layer: Included SunOS implementation to convert UNIX SCSI disk device access to Netstation DVD network commands to a remote disk.
- Demonstration of Phase-II Display System: Demonstrated Phase-II display system including DVD support for third-party I/O running on Netstation motherboard, kernel and communication stack.
- Demonstration of Netstation Disk System: Demonstrated standard Sun workstation accessing prototype Netstation IPdisk, using VISA adaptation layer.

Discussion

The second phase primary concern was making network-attached devices available within a shared wide-area environment, such as an internetwork. This shifted the emphasis from performance to issues of safe shared accesses and use of standard protocols. The primary contribution was a new abstraction/methodology for allowing devices to be shared in an "unsafe" environment.

The Derived Virtual Device abstraction provided mechanisms to address the primary areas of shared access: (1) resource bounds checking, (2) user authentication, and (3) restricted/controlled operations, e.g. read-only derived devices. As proof of concept, a DVD implementation was created for the Netstation Display. This was achieved by porting a Scheme interpreter that ran in the motherboard. The display device ‘owner’ created the Scheme code that implemented a safe execution environment for third-party connections to the display. The demonstration consisted of the owner creating a DVD that provided third-party clients with a constrained frame buffer. Subsequently, a third-party client that opened a connection to the display freely used the frame buffer subject to the owner’s pre-defined constraints, thus preserving the integrity of the owner’s private portion of the frame buffer.

The hardware and software development for Phase-II concentrated primarily on providing a demonstration environment that would be commercially practical. This required conversion of the Netstation motherboard to make use of commercially standard network interface. The 100 Base-T Ethernet was chosen as the Phase-II network interface. Since most commercially-available 100 Base-T Ethernet cards interfaced to a PCI Bus, a PCI Bus capability was needed for the motherboard. To that end both PCI Bridge and PCI Bus daughter cards were developed for the motherboard. On the software side, considerable effort was made to provide a high-performance communication oriented kernel and TCP/IP stack.

The objective of the second Phase-II Netstation demonstration was to determine what were performance levels could be realistically obtained when driving a commercially standard SCSI disk remotely across the internetwork from a workstation. To that end an emulated disk drive, IPdisk, was created that implemented a SCSI block device command set. IPdisk could be accessed via either TCP or UDP. IPdisk supported the DVD model, allowing an ‘owner’ to download small programs that act as filters on the SCSI RPCs prior to their actual execution.

To use IPdisk required the creation of a SCSI device adaptation driver that resided within a commercial kernel. SunOS was chosen for this purpose. A new adaptation layer driver, VISA, was created for the SunOS kernel. VISA acted as a standard part of the scsi_transport structure within the kernel, accepting SCSI device commands and sending them across the internetwork to IPdisk for actual execution. Detailed file system and network CPU utilization and performance data gathered during use of IPdisk/VISA indicated that it was possible when using UDP as the transport protocol to reach more than 80% of SCSI’s maximum throughput. We were thus able to demonstrate that IP is a viable alternative to special-purpose storage network protocols.

4. ATOMIC-2

ATOMIC-2 addressed the production use of a gigabit LAN. In a low-bandwidth LAN (e.g., 10 Mbps ethernet), the LAN is the bottleneck, affecting host-host, host-file server, and host-gateway interactions (Figure 2, left). The advent of gigabit LAN technologies, such as ATOMIC, alleviates the LAN bandwidth, but other limitations persist (Figure 2, right). The goal of the ATOMIC-2 project is to locate and address the limitations that remain, to allow gigabit LANs to provide enhanced production-level services [18] [19].

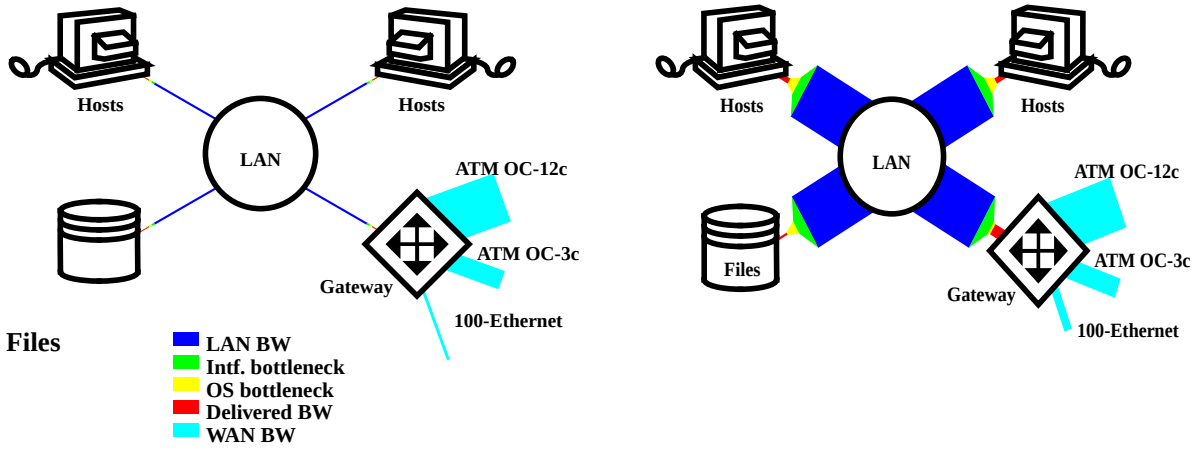

Figure 2 Before ATOMIC, the network is the bottleneck (left); after ATOMIC, bottlenecks remain at higher layers (right)

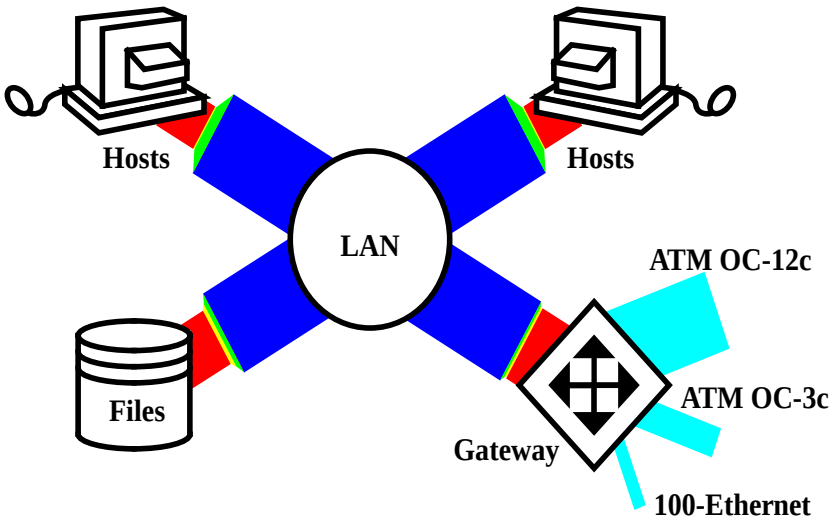

Figure 3 ATOMIC-2 alleviates the bottlenecks that remain after gigabit LANs are installed

When the ATOMIC-2 project began, there was no available production gigabit LAN technology. Earlier work at USC/ISI and CalTech helped develop lab prototypes of a candidate technology, the ATOMIC LAN. Prototype switches, including 3x3 and 8x8 meshes, and prototype host interface boards had been implemented. Early work had begun on the development of link driver hardware,

to extend the link length from several meters to tens, or even hundreds of meters. As a result, the ATOMIC-2 project's original goals included developing and producing production-grade cables, switches, and host interfaces. The final phases of this original goal included advanced development, addressing fast data transport and fast gateway design. Optionally, issues of second-generation host interfaces, link encryption, bulk storage, and router interfaces were considered.

During the first year of the NAAAN effort, the ATOMIC LAN technology was successfully spun-off from USC/ISI and CalTech, to a start-up company called Myricom, based in Arcadia, CA. Myricom's mission was to develop and produce production-grade ATOMIC LAN equipment, which they call Myrinet [29]. The advent of Myricom overtook many of the original ATOMIC-2 project goals, obviating the need to develop production cables, switches, and host interfaces. At that time (last quarter 1994), the ATOMIC-2 project was revised to focus on the latter components of its original goals, those of advanced development. Rather than developing and replicating production equipment, the project then assumed the availability of such equipment, and focused on the remaining performance issues.

The revised ATOMIC-2 project's goal was to identify and address performance limitations that remain after a gigabit LAN is installed, that impede its use for everyday users. This differentiates it from corresponding graphics or supercomputing research, for which optimizations exist, but also which operate under conditions that are unrealistic for everyday users. For example, supercomputer networks optimize bulk transfer assuming files are 10-100 Mbytes long; ATOMIC-2 needed to address more typical file transfer sizes of 100 Kbytes. As a result, ATOMIC-2 provides a perspective on gigabit LAN performance issues that will impact the bulk of future users, rather than only a niche domain.

The ATOMIC-2 project is a coordinated effort to increase the benefit of using a gigabit LAN for everyday networking. It consist of a set of target issues, each addressing a dimension of this space:

- *Installation* - install a gigabit LAN for production use by researchers in the networking division of USC/ISI, and (pending review) at DARPA.
- *Gateway* - develop a production-quality gateway from a host. This is required because Myricom is not developing router line cards; a host-based gateway utilizes the available Myricom host interfaces to support WAN interoperation.
- *Service provision* - examine whether and how quality-of-service can be included in ATOMIC LAN technology.
- *Application demo* - develop an application exhibiting functionality not feasible in conventional LAN technology. Multimedia teleconferencing, parallel applications, and web servers are considered.
- *File server* - develop a high-performance shared file system, for typical file sizes.
- *Security* - develop high-performance authentication and/or encryption.

This set of target issues represents a wide variety of domains, but remains a concerted effort to provide gigabit LAN service to everyday users. For example, the gateway, file server, and security results can be composed to provide secure remote access.

The result of the ATOMIC-2 project was to complete a task in each of its target areas:

- *Installation* -
  USC/ISI maintained the largest known production gigabit LAN, including over 40 office workstations, a file server, and a gateway to both fast ethernet local services, and a regional ATM OC-3c testbed. This LAN provided production-grade connectivity for the entire Computer Networks Division, and supported high-capacity experiments as well as conventional network service. It has been in continuous operation since Feb. 1995. The installation of a corresponding system at DARPA was declined, due to shift in workstation and operating system use there, and reliability concerns. ISI also performed extensive tests comparing Myrinet, ATM, and fast ethernet technology.

- *Gateway* -
  A SunOS/SPARC host gateway was implemented, optimized, and debugged, and provided WAN and local resource connectivity for production access to both fast ethernet and ATM. A PC host gateway was also prototyped, and its performance increased by over 50%.

- *Service provision* -
  Experiments conducted by the ATOMIC-2 project were instrumental in proving that the Myrinet technology could not support differentiated services. The switches lack internal queues, and QoS can be established only by varying packet lengths; this requires global coordination, and is difficult to guarantee.

- *Application demo* -
  The ATOMIC-2 LAN supported extensive video teleconferencing, far in excess of that which even switched fast ethernet supports. Variants of teleconferencing were developed that avoided computationally intensive compression, by transmitting raw video data using large packets. Full frame rate video (over 20 frames/second) was feasible using conventional hosts. PVM was supported by a custom transport protocol, ATM (Atomic Transport Protocol). At one parallel systems conference, the ATOMIC-2 PVM had the highest bandwidth and lowest latency of all PVM implementations present. Finally, ATOMIC-2 identified a design flaw in TCP that caused state accumulation at servers on gigabit LANs, and developed a variety of techniques to alleviate that problem.

- *File server* -
  A driver was developed for the Texas Memory System's SAM 300 RAM disk. The availability of that disk allowed performance analysis of NFS in the absence of disk bandwidth bottlenecks. Subsequent analysis of NFS's remaining bottlenecks yielded a 2
  0% improvement in NFS bandwidth, where much of the improvement is exhibited for file sizes as low as 100 KB.

- *Security* -
  The default IPSEC authentication algorithm, MD5, was analyzed, and shown to have performance limits that may impact its utility for gigabit networks. An alternate algorithm was developed that is up to twice as fast, when used on a per-packet basis.

4.a. Discussion

The following is a more detailed discussion of the project results, contributions, and implications, based on each target area of effort.

Installation

The ATOMIC-2 project completed an installation of Myricom's Myrinet LAN throughout its Computer Networks Division, serving 57 offices, using 11 in-ceiling switches, and a Sun SPARC gateway (Figure 4, left). The network provided primary network service for 38 office workstations (Sun SPARCs), and another 10 lab workstations (PCs and SPARCs) (Figure 4, right) [18] [19]. This represented the largest Myrinet in the world, and one of the largest production gigabit LANs ever developed, and the one in longest continuous operation (since Feb. 1995, to the present) The network was designed with dual independent paths, so that a link or port failure inside the network did not disable the entire net. The Myricom products did not exhibit production-level reliability, so additional redundancy, in the form of a conventional ethernet, was required. A novel routing configuration was developed which automatically switched to the backup ethernet in the event of a port or link failure (Figure 5) [17].

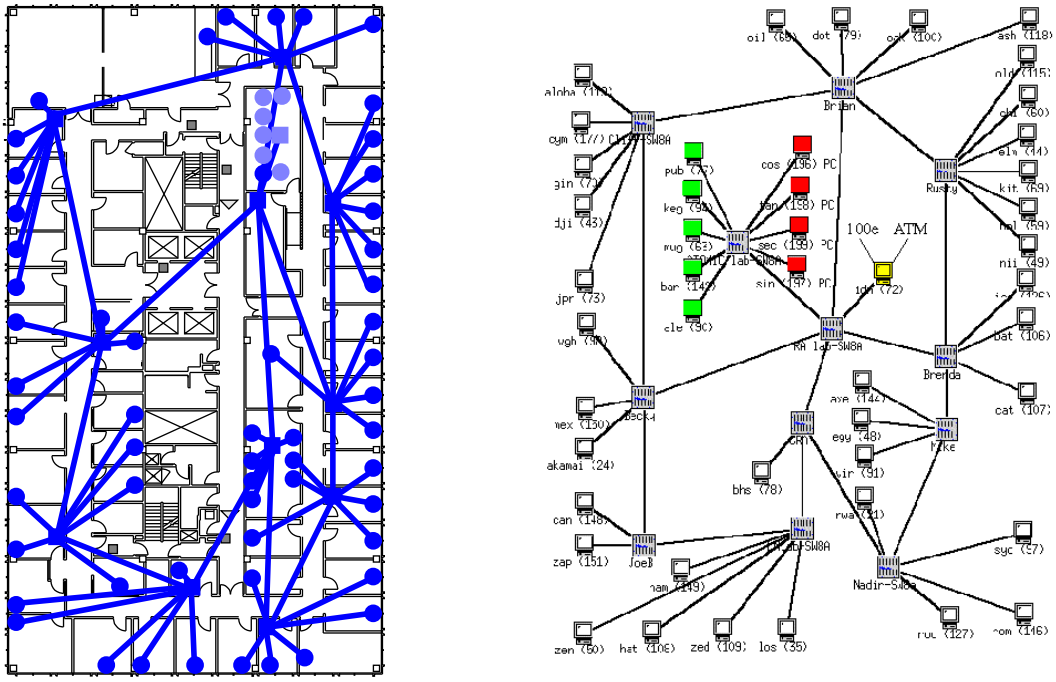

Figure 4

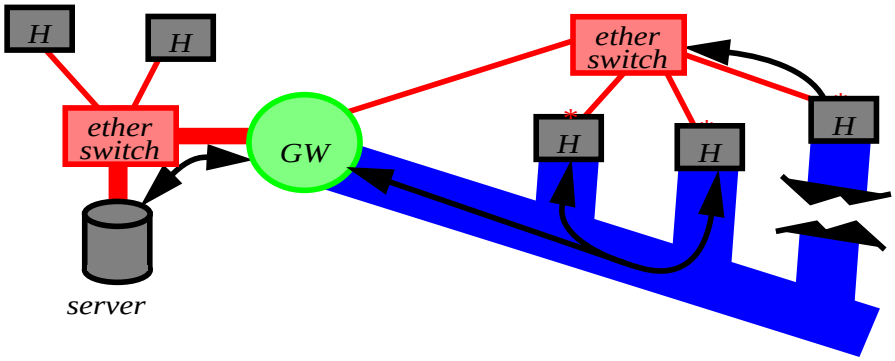

Figure 5 ATOMIC-2 hosts (right) are dual-connected, via the ATOMIC LAN (lower) and switched ethernet. Routing follows whatever path is currently available.

Techniques were also developed to accommodate office-area installation of the Myricom products, which are designed primarily for lab use. This included wrapping the cables to allow them to be installed in ceiling plenums, developing suspension shelves for the switches which suspended them upside-down to allow monitoring link lights, and planning a distribution topology that served all 57 offices on an entire floor within the 80-foot cable length limits (Figure 6). Installation was complicated by the delayed (and eventually abandoned) development of fiber-optic cables, which would allow conventional wiring-closet topologies, or the inclusion of services on other floors within ISI. In the course of this development, network monitoring tools were developed. Myricom provided on-line graphical network analysis tools; other, more SNMP-like interfaces were required to collect and compile statistics off-line. These statistics were made available on the ATOMIC-2 web page.

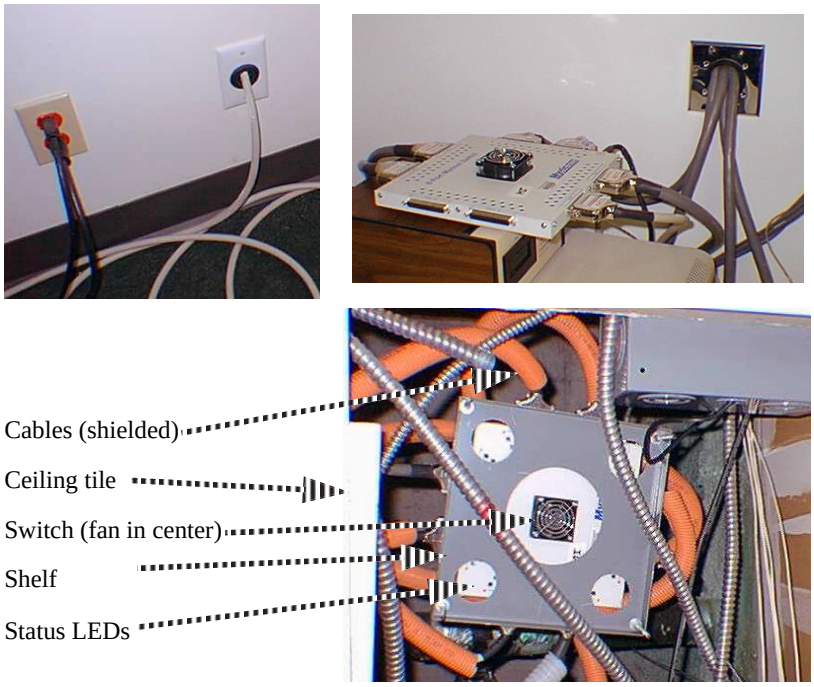

Figure 6

A similar network was scheduled to be installed at DARPA, but was not, due to the high failure rate of the components, the safety ratings of the equipment, and the fact that, for the vast majority of uses, fast ethernet represented a more efficient, cost-effective alternative. ISI compared the Myricom (640 Mbps, $2000/host) ATM OC-3c (155 Mbps, $3000/host), and fast ethernet (100 Mbps, $500/host) technologies, and found that, for packets smaller than 1 KB, fast ethernet was more effective (Figure 7). ATOMIC outperforms ATM and fast ethernet significantly only for packets over 2KB. The implication is that, for wide-area traffic, dominated by 512-byte packets, fast ethernet is more effective. ATOMIC is better for LAN services, such as in-house teleconferencing or file access.

ISI continues to await Myricom equipment upgrades, provided gratis. These upgrades allow switches to remain connected to link cables when the host is disconnected; in our existing equipment, this situation causes switch lockup and eventual port failure.

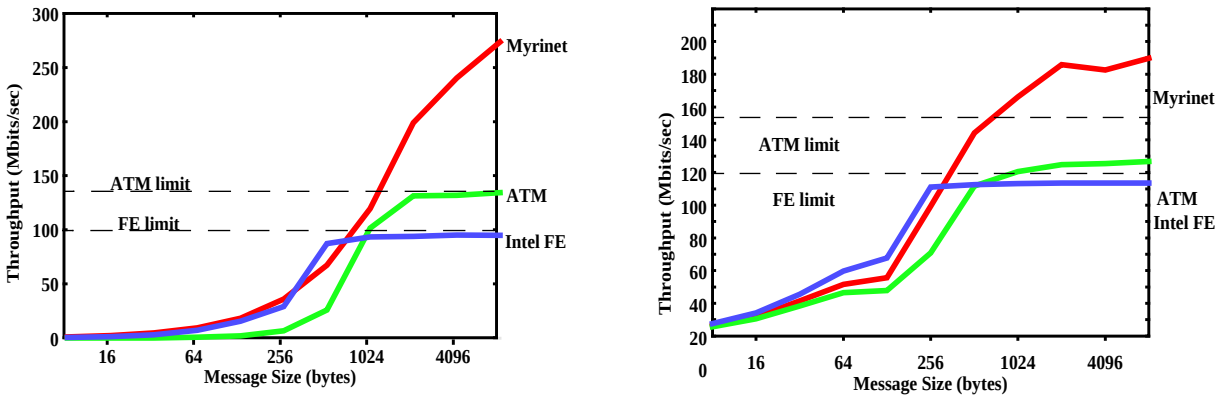

Figure 7 Host bandwidth is limited by fast ethernet, in UDP (left) and TCP (right); ATM is link limited for UDP, but processor limited for TCP; ATOMIC exceeds both, but only for packets over 1KB

This production network uses a Sun SPARC gateway, connected to fast ethernet to access local file servers on other floors of ISI, and a regional ATM testbed, called SCAN/ARC. The ATM testbed is provided by GTE (SCAN, the Southern California ATM Network) and PacBell (ARC, the ATM Research Consortium). The ATOMIC-2 project participated on these networks as one of their first ATM sites, and the one in longest continuous connectivity. We also developed the network architecture, including ATM VCI tables for PVC allocation, and provided a DNS (scan.net), and managed the IP address allocation within the testbed.

As part of this work, ISI also contributed significant feedback to Myricom and the its user community. We were instrumental in detecting and correcting a variety of software and hardware bugs, many of which would not have been noticed in a smaller or non-production system. We also compiled a a table of the interoperation of various versions Myricom hardware, and procedures for the safe handling of the equipment.

Gateway

A gateway was developed to allow the ATOMIC LAN to interoperate with other LAN and WAN technologies. Myricom's instance of the ATOMIC LAN is representative, in that high performance host interfaces were available early in the development of the technology, and router interfaces came later, or, as in the case of Myricom, not at all. As a result, host-based routing is required, using only host interfaces to bridge the technologies.

An initial production router was implemented using a Sun SPARC running SunOS 4.1.3. This host interconnected the ATOMIC LAN, ATM OC-3c (to a regional testbed), and fast ethernet (for shared file servers and Internet access). Performance measurements of this gateway indicated that its conventional forwarding technique limited throughput to OC-3c bandwidths (155 Mbps). This was due to a combination of limited peripheral bus bandwidth, processor speed, and memory bandwidth.

A second-generation router was developed, using a 200 Mhz Pentium Pro running FreeBSD 2.2.5. This platform offered higher CPU performance, memory bandwidth, and a much higher performance peripheral bus (PCI). A driver was developed for the Adaptec PCI ATM OC-3c card, which was also distributed to the participants of the DARPA CAIRN testbed. A novel technique

for peer-DMA was implemented in a modified Myrinet driver, so packets could be routed among the host interfaces without incurring an additional copy into host memory; this technique improved router throughput by 50% [9] [20] [21].

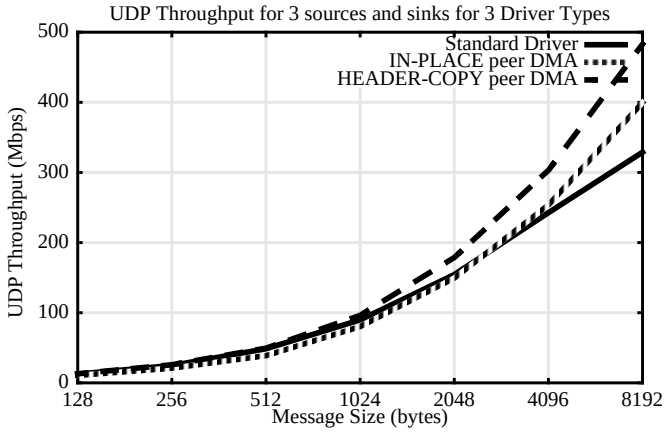

Figure 8 Peer-DMA increases host router bandwidth by up to 45% (shown here for UDP traffic)

Service Provision

ATOMIC-2 investigated the opportunity for service provision (QoS) in the ATOMIC LAN. The ATOMIC LAN achieves high performance switches inexpensively by trading complex, costly buffering and queuing mechanisms with overprovisioning (excess capacity) and a deterministic prioritization mechanism. In the typical case, resource reservation is not an issue, because there is enough excess capacity to effectively ignore bandwidth as a critical resources.

However, in the case where hard real-time guarantees are required, it is important to understand how and whether QoS can be guaranteed in wormhole-routed networks like ATOMIC. ISI performed experiments validating analysis performed later at UCLA, which proved that QoS would require that reserved paths did not intersect at switches [15]. This can be accomplished by severe overprovisioning (a network per call, effectively), or by using the switches only for single hops, and requiring host-based routers (with buffers and queuing mechanisms) in-between. The latter replaces the wormhole routed network with a host-router network of simple switched hubs, defeating the gain of using simple, cheap switches. We found that there was a way to perform QoS reservation, but that it was sensitive to the global coordination of packet length and path, deemed infeasible.

Applications

The ATOMIC-2 project identified several applications to be used as proof-of-concept to demonstrate the utility of a production gigabit LAN. We tested PVM (Parallel Virtual Machine) applications, multimedia teleconferencing, and web service as candidate applications [33].

A version of PVM was installed and measured over the ATOMIC LAN. We found that PVM achieved lower performance than conventional TCP over the same link. A modified version of PVM was developed that increased throughput by 130%, over 70% higher throughput than TCP (Figure 9). This version used a custom transport protocol, ATM (ATOMIC Transport Protocol), which that pipelined the data copies in the OS and driver [25].

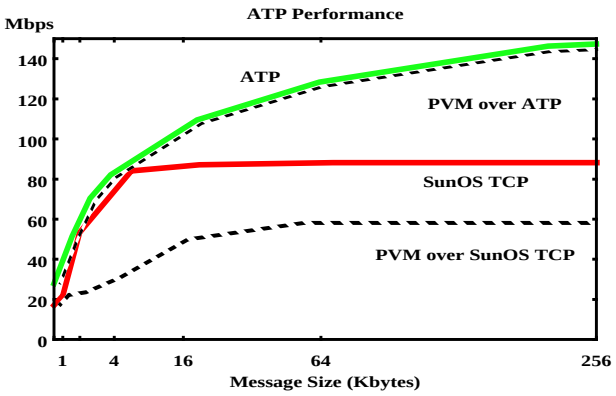

Figure 9 ATP allows PVM to achieve bandwidths higher than even raw TCP

We also examined the development of a user space TCP, to extend the results of ATP to generic application use. The goal was to design a library that would intercept calls to the socket, and replace TCP requests with ATP within the LAN. This work was overtaken by a similar implementation done at several other sites.

The ATOMIC-2 project also tested high-performance multimedia teleconferencing. We modified the source code of nv, Xerox PARC's video tool, to omit data compression and emit raw video in very large (8KB) packets. Conventional nv was CPU limited on Sun SPARCs, generating only 8 frames/second, at 1-2 Mbps. The modified nv was capable of near full-motion video (20-22 frames/second), and consumed 6-8 Mbps. This data rate relined on the use of large packets; when the packet size was reduced to 512 bytes (for Internet WAN use), the gains were not significant. The result is that high performance teleconferencing can be achieved, trading bandwidth for CPU capacity, but only where large packets can be used end-to-end.

Finally, the ATOMIC-2 project considered the effect of high bandwidth on web service. The web is the dominant application on the Internet, and ATOMIC provided a way to examine the effect of gigabit networking on this application far in advance. We found that a design property of TCP, which causes state to accumulate at the server rather than at the client, inhibited web throughput (connections/second), validating an earlier hypothesis (Figure 10) [32]. We developed patches for the Apache web server, and alternative modifications to TCP to avoid the state accumulation [10] [11]. We were able to enhance the performance of a Sun SPARC 20/71 (75 Mhz) to achieve data throughputs previously attainable only on a Dec Alpha (500 Mhz) workstation.

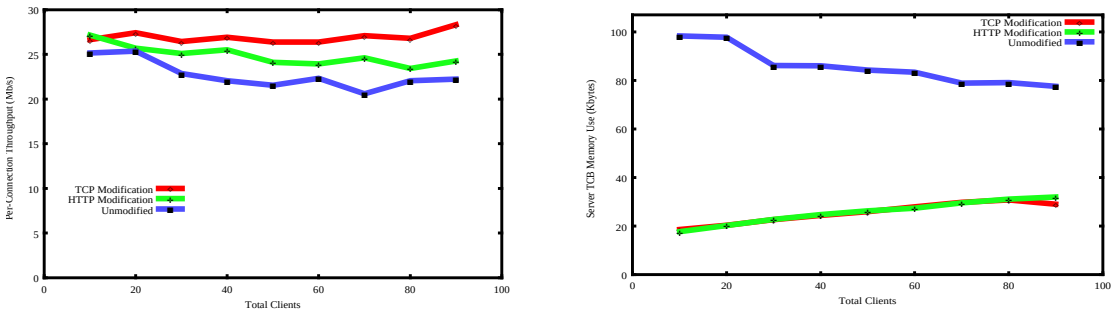

Figure 10 HTTP state alleviated by TCP modifications support the highest BW (left); HTTP modifications are not as effective, though both reduce server state (right)

File Server

The ATOMIC-2 project developed a high-performance file server. There have been a number of similar fast file server projects, but ATOMIC-2 focused on files of typical size for average workstation users, around 100 KB. The conventional bandwidth bottleneck of disk systems was omitted by using a high-performance RAM disk system, the Texas Memory Systems SAM-300. We developed SunOS drivers for the SAM-300, to allow it to be mounted as a traditional unix file system.

NFS was examined, in light of this RAM disk, and found to have performance limits [1]. We developed a pipelined RPC, to avoid the stop-and-go thrashing that occurs because each NFS biod (block I/O daemon) handles only a single outstanding request. The result was a 30% improvement in the performance of the NFS protocol, which translated to a 20% improvement in the write-to-disk performance of NFS (Figure 11). We also found that increasing the number of biods did not help.

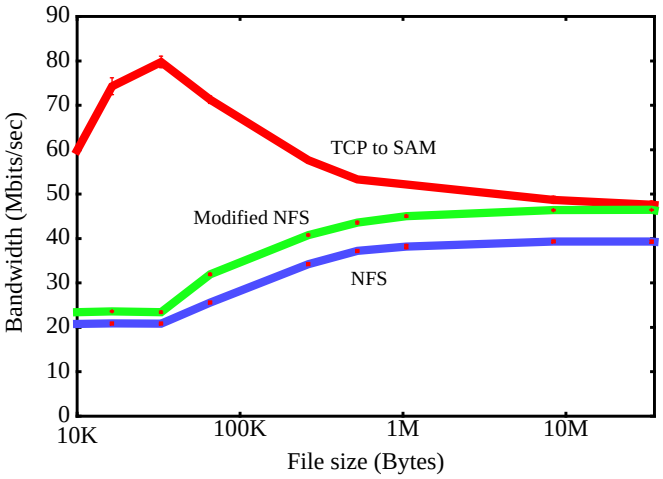

Figure 11 Modified NFS achieves 20% higher bandwidth than NFS for large files; for small files, TCP remains more efficient than NFS

Security

The ATOMIC -2 project measured the performance of authentication algorithms, both stand-alone and in the IPv4 processing path. We found that MD5, the default IPSEC authentication algorithm, has performance limits that would severely inhibit the bandwidth available on the ATOMIC LAN [5] [16]. TCP/IPv4 with MD5 was capable only 30 Mbps throughput, on systems that typically achieved 90 Mbps for TCP/IPv4 without authentication (Figure 12).

We also characterized the effect of various components of the IPSEC authentication algorithm, including header modification, data touching overheads, byte-reordering, and a variety of candidate hash algorithms. Together with IBM, we developed an alternate hash that achieves 55 Mbps, an 80% improvement. We also distributed an optimized MD5 assembler that is within several percent of a provably optimal implementation. Our test suite was distributed to the IPSEC community.

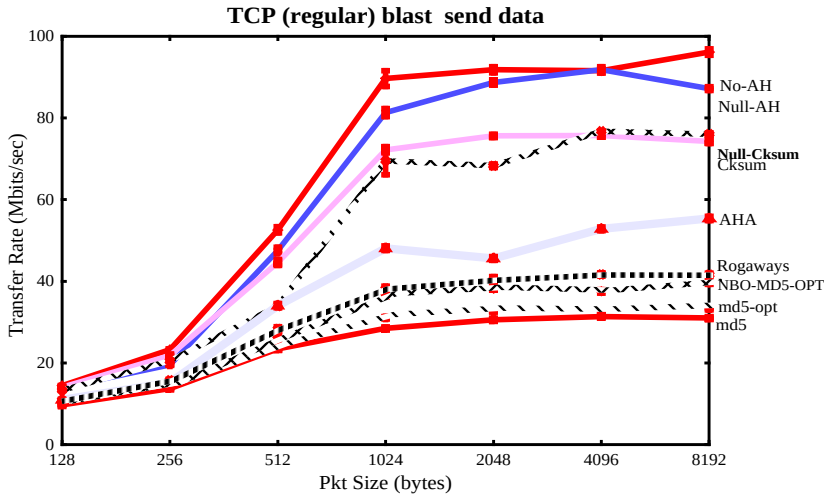

Figure 12 Optimized MD5 severely limits bandwidth on ATOMIC-2; ISI's Alternate Hash Algorithm is faster than other known algorithms, though even data touching is a significant issue at these speeds

The algorithms are shown in Figure 12 are:

• No- AH	No AH headers or hash
• Null-AH	An AH header, but no hash algorithm (all below include AH)
• Cksum	Internet checksum as a hash algorithm
• Null-Cksum	Sender hash only (emulates ILP for hash processing)
• AHA	An Alternate Hash Algorithm [5] [16]
• Rogaways	Phil Rogaway's, uses bin-scatter and an intra-bin AHA hash
• NBO-MD5-OPT	Optimized MD5, in network-standard byte order [5] [16]
• md5-opt	Optimized MD5 [5] [16]
• md5	From RFC 1321 [30]

5. Technology Transfer

5.a. Software releases:

- SunOS IPSEC AH performance suite, including optimized MD5 C and Pentium assembler, Alternate hash algorithm (used by IETF IPSEC working group)
- SunOS ATOMIC-PVM and ATP (ATOMIC Transport Protocol)
- SunOS TMS SAM-300 driver
- FreeBSD Adaptec ATM driver (used by DARPA CAIRN participants)
- SunOS patches to use Myrinet as primary interface (to install driver in the kernel)
- Blast performance tool
- Apache TIME_WAIT reduction patch
- SunOS TCP TIME_WAIT shift patch
- Optimized FreeBSD Myrinet driver (rolled into product)
- FreeBSD Myrinet driver patches to support Peer-DMA forwarding
- IPDisk code to implement a network-attached SCSI disk

5.b. Hardware prototypes:

- Netstation prototype node motherboard
- ATOMIC/Myricom network interface board
- ATOMIC/Myricom cable driver board
- Display board
- JPEG board
- PCI Bus bridge board
- Flash memory board
- Display board, burst capable

5.c. Other

- bug fixes and design enhancements, both hardware and software, back to Myricom
- procedures for the safe handling and multi-version interoperation of Myricom equipment

6. Travel

1/94 Greg Finn attended the High Definition Systems IEC in Washington, D.C., where he presented the current research plans for the Netstation portion of the NAAAN contract.

2/94 Bob Felderman spoke at the Naval Research Lab in Washington DC. The topic was “ATOMIC: The Evolution of a LAN”.

3/94 Bob Felderman gave an invited presentation at the University of Toronto Computer Systems Research Institute titled, “ATOMIC: Evolution and Performance”.

5/94 Greg Finn gave a presentation on presented a talk on early results of RPC experiments at INTEROP’94 in Las Vegas, NV/

5/94 Paul Mockapetris attended an ATM Switch briefing at Net, Inc. in San Jose, CA.

7/94 Gregory Finn and Joe Touch attended an ARPA PI conference in Santa Fe.

7/95 The Los Angeles area NPR station KPCC did a one hour program on the Internet during their Friday Airtalk program. The local guest speaker was Gregory Finn from USC/ISI.

9/94 Gregory Finn gave a short presentation on performance results to the High Performance Communication Principal Investigator meeting at Santa Fe, NM.

11/94 Ted Faber attended the CNRI Gigabit Testbed Workshop in Washington D.C. His expenses were covered by another contract.

11/94 Paul Mockapetris met in Washington DC with Gary Minden and Michael St. Johns to discuss the ATOMIC2 and NETSTATION plan revisions.

11/94 Hong Xu presented the paper “Optimal Software Multicast in Wormhole-Routed Multistage Networks” at Supercomputing '94 in Washington D.C.

4/95 Hong Xu and Ted Fisher attended the International Parallel Processing Symposium in Santa Barbara, CA, to present their work on improving PVM performance.

4/95 Joe Touch attended IEEE Infocom in Boston. During that trip, he also visited GTE Labs in Waltham, MA, in conjunction with the SCAN/ARC ATM network research consortia.

4/95 Joe Touch gave a presentation entitled “Progress on the GBN ‘94 ‘Five Challenges That Define High-Speed Protocols’” at the IEEE Gigabit Networking Workshop in Boston/

4/95 Joe Touch gave a presentation written by the entire ATOMIC-2 group entitled “ATOMIC-2: Production Use of a Gigabit LAN” at the IEEE Gigabit Networking Workshop in Boston.

4/95 Hong Xu gave a presentation at the first High-Speed Network Computing workshop in the 9th International Parallel Processing Symposium (Santa Barbara).

5/95 Joe Touch gave a presentation entitled “Performance of MD5” at IBM T.J. Watson Research Lab, Hawthorne NY.

6/95 Hong Xu attended the SSN meeting at UCLA, to make a presentation on link-layer quality of service provision in the Myrinet.

7/95 Joe Touch presented a poster on the status of the ATOMIC-2 task at the ARPA CSTO Principal Investigator's Meeting in Ft. Lauderdale, FL.

8/95 Hong Xu attended the 3rd Hot Interconnects conference at Stanford University to meet with researchers at UCLA and Stanford on PVM performance optimization.

8/95 Joe Touch attended the invited Sigcomm Middleware Workshop in Boston. He presented a discussion of "Communications Middleware at a panel, "High-Level Communications Middleware," with Dave Clark, Christian Huitema, and Van Jacobson.

8/95 Joe Touch attended the Sigcomm 1995 conference in Boston, and presented a paper entitled, "Performance Analysis of MD5."

9/95 Rod Van Meter presented a talk on the Netstation Project to the Network Attached Storage Device working group of the National Storage Industry Consortium.

10/95 Ted Faber presented a talk on ATOMIC-2 status at the Very High-Speed Protocols Workshop at the Univ. of Maryland, Baltimore County.

11/95 Ted Faber attended Mobicom '95, at UC Berkeley. There he attended the mobile IP tutorial, and attended the conference sessions, to determine the effect of emerging mobility issues on the ATOMIC-2 tasks.

11/95 Ted Faber presented a talk at the Washington University St. Louis on NFS performance.

2/96 Rod Van Meter attended Supercomputing '95 in San Diego.

3/96 Greg Finn, Joe Touch, and Ted Faber attended the ARPA Networking PI Meeting in Charlottesville, SC. Greg gave a presentation on the status of the NAAAN project.

3/96 Joe Touch presented a talk on the status of the ATOMIC-2 project at the IEEE Gigabit Networking Workshop, held in conjunction with Infocom '96 in San Francisco.

3/96 Joe Touch also attended Infocom in San Francisco, and participated in the Infocom Program Committee meetings, as well as the IEEE TCCC meeting.

1/96 Ted Faber hosted a visit by Jeff Hollingsworth, Univ. of Maryland at College Park, who gave a talk at ISI on "Parallel I/O Optimization via Application Tuning".

5/96 Rod van Meter attended the Federated Computing Research Conferences in Philadelphia. He attended workshops on Input/Output in Parallel and Distributed Systems and SIGMETRICS. He also attended portions of the International Conference on Supercomputing and the International Symposium on Computer Architecture.

7/96 Joe Touch attended the Gigabit Network Technology Workshop at the University of Washington at St. Louis, and presented a white paper on "ISI's High-Performance Networking Research."

9/96 Rodney Van Meter attended the Fifth NASA Goddard Conference on Mass Storage Systems and Technologies, where he presented a paper “Derived Virtual Devices: A Secure Distributed File System Mechanism.”

11/96 Joe Touch presented the ATOMIC-2 research during travel funded on another project, giving an invited lecture to the high-speed net research group at the Univ. College at London.

1/97 Rodney Van Meter travelled to the Usenix annual technical conference to present a paper.

1/97 Rodney Van Meter attended a Network Storage Architecture Task Force meeting at the Hawaii International Conference on System Sciences to present a position paper on Internet-attached storage devices.

3/97 Rodney Van Meter attended the Globus workshop at the San Diego Supercomputing center. The workshop discussed approaches to improve the network I/O performance of very high-end systems.

3/97 Joe Touch attended the DARPA Nets PI meeting in Baltimore, MD and presented a summary of the NAAAN project.

3/97 Joe Touch met with Scott Michel of UCLA in Los Angeles to coordinate the reconfiguration of the SCAN ATM system.

4/97 Joe Touch co-chaired the IEEE Gigabit Networking Workshop at Kobe, Japan, and presented paper on Atomic-2, chaired a session on “High-Performance Web”, and attended Infocom.

5/97 Joe Touch participated in the Next-Generation Internet Workshop, in Washington D.C. This travel was supported by an indirect grant from the NSF.

5/97 Greg Finn travelled to Las Vegas, NV to present a paper at the IEEE + Interop Engineers Conference.

5/97 Rodney Van Meter travelled to Washington, D.C. to attend the Digital Government workshop.

7/97 Joe Touch and Joe Bannister visited DARPA and the DOE to discuss the status of the ATOMIC-2 task with our project manager, and the impact of this work on the DOE cluster computer user community.

7/97 Joe Touch visited Teledesic to present a talk and discuss applying ATOMIC-2 to their satellite routing system with Hans-Werner Braun. This travel was funded by Teledesic.

7/97 The ATOMIC-2 project hosted visitors from Genuity, a regional network service provider, on July 20, 1997. They discussed the implications of the ATOMIC-2 dynamic configuration and high-performance host-based routing on network backbone architecture with Rodney Joffee and others.

7/97 Rodney Van Meter travelled to HP Laboratories in July 1997 to present a paper.

9/97 Joe Touch, Ted Faber, and Joe Bannister attended the IEEE Computer Communication Workshop, in Phoenix. Joe Bannister chaired the workshop, and Joe and Ted gave presentations, and discussed the ATOMIC LAN results with attendees.

10/97 Joe Touch attended a meeting of the Cluster Interconnects working group at Sandia National Labs, in Livermore, CA. He discussed the application of ATOMIC LAN and ATOMIC-2 technology to the cluster interconnects domain there.

10/97 Joe Touch and Ted Faber attended the IEEE ICNP conference to present a paper on dynamic host routing.

10/97 Joe Touch co-chaired the DARPA Optical Internets Workshop in Arlington, VA.

10/97 Joe Touch presented a summary of the ATOMIC-2 results at Columbia University, hosted by Y. Yemini, and Univ. of Maryland at College Park, hosted by J. Hollingsworth.

12/97 Joe Touch, Joe Bannister, and Ted Faber attended the 40th IETF, in Washington DC. Joe and Ted presented a summary of the TCP TIME_WAIT avoidance work, to enhance the performance of web servers, to the TCP-IMPL working group.

12/97 Joe Touch attended a second meeting of the Cluster Interconnects working group at Sandia National Labs, to continue earlier discussions on the transition of ATOMIC-2 results.

7. Publications and presentations

7.a. Journals

[1] Faber, T., “Optimizing Throughput in a Workstation-Based Network File System over a High Bandwidth Local Area Network,” SIGOPS Operating Systems Review, January 1998.

[2] Felderman R., DeSchon A., Cohen, D., and Finn, G., “ATOMIC: A High Speed Local Communication Architecture” *Journal of High Speed Networks*, 1994, No. 1, pp. 1-28.

[3] Sterbenz, J., Schulzrinne, H., and Touch, J., “Report and Discussion on the IEEE ComSoc TCGN Gigabit Networking Workshop 1995,” *IEEE Network,* July 1995, pp. 9-21.

[4] Touch, J., “Defining ‘High-Speed’ Protocols: Five Challenges & an Example That Survives the Challenges,” *IEEE JSAC*, special issue on Applications Enabling Gigabit Networks, V. 13, N. 5, June 1995, pp. 828-835. Also available as ISI/RS-95-408.

[5] Touch, J., “Report on MD5 Performance,” RFC-1810, USC/ISI, June 1995.

[6] Touch, J., and Parham, B., “Implementing the Internet Checksum in Hardware,” USC/ISI, Network Working Group RFC-1936, April 1996.

[7] Finn, G., Hotz, S., Van Meter, R., “The Impact of a Zero-Scan Internet Checksumming Mechanism”, ACM Computer Communication Review, October 1996.

[8] Van Meter, R., “A Brief Survey of Current Work on Network Attached Peripherals” ACM Operating Systems Review, Jan. 1996.

[9] Walton, S., Hutton, A., and Touch, J., “High-Speed Data Paths in Host-Based Routers,” IEEE Computer, Nov. 1998, pp. 2-8.

7.b. Workshops

[10] Faber, T., Touch, J., and Yue, W., “The TIME-WAIT state in TCP and Its Effect on Busy Servers,” (submitted to Infocom ’99).

[11] Faber, T., Touch, J., and Yue, W., “Avoiding the TCP TIME_WAIT state at Busy Servers,” (working draft), ISI, Sept. 1997.

[12] Finn, G., and Mockapetris, P., “Netstation Architecture Multi-Gigabit Workstation Network Fabric,” *1994 Spring Interop Engineering Conference*.

[13] Finn, G.G., Van Meter, R., Rogers, C.M., “Datagram Forwarding via Stateless Internetwork Switching,” Proceedings of the IEEE + Interop Engineers Conference, 7-8/5/97, Las Vegas, NV.

[14] Hotz, S., Van Meter, R., and Finn, G.G., “Internet Protocols for Network-Attached Peripherals”, Sixth NASA Goddard Conference on Mass Storage Systems and Technologies in conjunction with 15th IEEE Symposium on Mass Storage Systems, March 1998.

[15] Kwan, B., Hu, P., Bambos, N., Kleinrock, L., Touch, J., and Xu, H., “Best-Effort Bandwidth Reservation in High Speed LANs using Wormhole Routing,” *Proc. International. Conf. on Computer Communications and Networks 1996*, Oct. 16-19, 1996, Washington D.C.

[16] Touch, J., “Performance Analysis of MD5,” *Proceedings of Sigcomm ‘95*, pp. 77-86.

[17] Touch, J., and Faber, T., “Dynamic Host Routing for Production Use of Developmental Networks,” Proc. of IEEE International Conference on Network Protocols, 1997.

[18] Touch, J., Faber, T., DeSchon, A., Sachdev, A., “ATOMIC-2: Going the Last Meter for Gigabit LANs,” IEEE Gigabit Networking Workshop, San Francisco, CA, 1996.

[19] Touch, J., Faber, T., Hutton, A., Jani, D., Yue, W., “Experiences with a Production Gigabit LAN”, IEEE Gigabit Networking Workshop, Kobe, Japan, 1997.

[20] Touch, J., Hutton, A., Walton, S., “Host-based Routing Using Peer DMA,” Presented at the IEEE Gigabit Networking Workshop, San Francisco, April 1998.

[21] Walton, S., Hutton, A., and Touch, J., “Efficient High-Speed Data Paths for IP Forwarding Using Host-Based Routers,” 9th IEEE Workshop on Local and Metropolitan Area Networks (LANMAN), Banff, Alberta, Canada, May 17-20, 1998.

[22] Van Meter, R., “Observing the Effects of Multi-Zone Disk Drives,” at the Usenix 1997 Annual Technical Conference, Anaheim, Jan. 1997

[23] Van Meter, R., Finn, G.G., and Hotz, S., “VISA: Netstation’s Virtual Internet SCSI Adapter”, ASPLOS VIII, October 1998.

[24] Van Meter, R., Hotz, S., and Finn, G.G., “Derived Virtual Devices: A Secure Distributed File System Mechanism” *Proceedings of the Fifth NASA Goddard Conference on Mass Storage Systems and Technologies*, 1995.

[25] Xu, H. and Fisher, T., “Improving PVM Performance Using the ATOMIC User-Level Protocol”, accepted to appear in *Proceedings of High-Speed Network Computing Workshop at the 9th International Parallel Processing Symposium*, April 1995.

[26] Xu, H., Gui, Y.D., and Ni, L.M., “Optimal Software Multicast in Wormhole-Routed Multistage Networks”, Proceedings of Supercomputing'94, pp. 703-712, November 1994.

7.c. Other

[27] Finn, G., Van Meter, R., Hotz, S., “Interfacing High-Definition Displays via the Internet”, August 1995. (available at http://www.isi.edu/netstation/)

[28] Finn, G., Hotz, S., Van Meter, R., “NVD Research Issues and Preliminary Models”, March 1995. (available at http://www.isi.edu/netstation/)

8. Additional References

[29] Boden, N., Cohen, D., *et al*, “Myrinet – A Gigabit-per-Second Local-Area Network,” IEEE-Micro,Vol.15, No.1, February 1995, pp.29-36.

[30] Rivest, R., “The MD5 Message-Digest Algorithm,” Network Working Group RFC 1321, MIT Laboratory for Computer Science and RSA Data Security, Inc., April 1992.

[31] Clark, D., D. Tennenhouse, D., “Architectural Considerations for a New Generation of Protocols,” Proc. Sigcomm 1990, IEEE, Sept. 1990, pp. 200-208.

[32] Moskowitz, R., “Why in the World is the Web So Slow?”, in *Network Computing*, March 15, 1996, pp. 22-24.

[33] Geist, A., *et al.*, P*VM 3 User’s Guide and Reference manual*. Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, Sept. 1994.