LA-UR-97-7-3501

Approved for public release;
distribution is unlimited

Title:

Super-speed Computer Interfaces and Networks

Author(s):

D. Tolmie, CIC-5
W. St. John, CIC-5
D. DuBois, CIC-5
G. Dornhoff, CIC-5
J. Morrison, CIC-5
A. DuBois, CIC-5

Submitted to:

DOE Office of Scientific and Technical Information (OSTI)

RECEIVED

OCT 01 1997

OSTI

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED *ph*

MASTER

Los Alamos
NATIONAL LABORATORY

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Form 836 (10/96)
ST 2629

DISCLAIMER

**Portions of this document may be illegible
in electronic image products. Images are
produced from the best available original
document.**

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, make any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Super-Speed Computer Interfaces and Networks

Donald E. Tolmie*, Wallace St. John, David H. DuBois,
Arlo G. Dornhoff, John F. Morrison, and Andrew J. DuBois
Computing, Information, and Communications Division

Abstract

This is the final report of a one-year, Laboratory Directed Research and Development (LDRD) project at Los Alamos National Laboratory (LANL). Research into super-speed computer interfaces has been directed towards identifying networking requirements from compute-intensive applications that are crucial to DOE programs. In particular, both the DOE Energy Research High Performance Computing Research Centers (HPCRC) and the DOE Defense Programs Accelerated Strategic Computing Initiative (ASCI) have planned applications that will require large increases in network bandwidth. This project was set up to help network researchers identify those networking requirements and to plan the development of such networks. Based on studies, research, and LANL-sponsored workshops, this project helped forge the beginnings for multi-gigabit/sec network research and development that today is being lead by Los Alamos in the American National Standards Institute (ANSI) 6.4 gigabit/sec specification called HIPPI-6400.

Background and Research Objectives

In 1995, the team of Wally St. John and David DuBois were awarded an R&D 100 award for the development of the HIPPI/SONET Gateway. This system provided a way for high performance, geographically separate local supercomputer networks using the 800 megabit/sec High Performance Parallel Interface (HIPPI) to interconnect at these data rates using the Synchronous Optical Network (SONET) infrastructure operated by regional and long distance telephone companies.

The HIPPI/SONET Gateway epitomized the quality and research, design, and development that Los Alamos National Laboratory (LANL) engineers have produced since the late 70s when high-performance network hardware was required to support the first supercomputers coming into the Laboratory. Engineers in the Computing, Information, and Communications (CIC) Division at LANL were responsible for the High Speed Parallel Interface development from 1978 to 1993 that produced a 50 megabit/sec network

*Principal Investigator (E-mail: det@lanl.gov)

physical layer. It was after the commercial networking manufacturers started to provide 100 megabit links (FDDI, e.g.) in the late 1980s that the researchers moved to the development of gigabit/sec links. HIPPI is the result of this transition, and the current high speed HIPPI network in the Laboratory's production computing network is another example of the capabilities within CIC Division to develop and deploy technology crucial to the Laboratory's mission.

Importance to LANL's Science and Technology Base and National R&D Needs

Several programs at Los Alamos have compute-intensive applications that are driving the need for more bandwidth in the production networks. These include the stockpile stewardship, grand challenge (weather modeling, etc.), and crisis management applications, and each of these requires tremendous computing resources that also generate or process terabytes of data. In addition, the growth of technology in commodity computing (i.e., high end PCs) also creates demands for more bandwidth. Some of these demands and other criteria are summarized as follows:

- Rapidly increasing microprocessor speeds create larger network demands
- Grand challenge problems that require interconnected supercomputers
- Increasing networking needs with upcoming MPPs (multiple parallel processors) and SMPs (synchronous multi-processors)
- Decreasing granularity places a greater demand on network resources

Current demands for networking speeds with bandwidths ranging from 1 Mbit/s to 1 Tbit/s and latency ranging from nanoseconds to seconds are increasing. The teraflop machines, made from interconnected smaller machines, need bandwidths from 10-1000 Gbit/s and latencies from nanoseconds to seconds. Data fusion, e.g., accessing and compiling city information, used a digital library with terabytes of data and millions of users, with 30-40 Gbit/s LANs and 10 Gbit/s WANs. Data fusion needed bandwidths across the whole range, but latencies were in the second range. In between were video editing, in which 20 video streams with 1200 Mbytes per stream were concatenated, and medical imaging. Both video editing and medical imaging needed bandwidths in the 1-100 Gbit/s range, but latencies were in the microsecond to second range. Sensor data, e.g., physics data, also had very large bandwidth requirements, but latency requirements were in the seconds range. The finer the granularity across the interconnected computers, the higher the bandwidth required in the network.

Scientific Approach and Accomplishments

Based on the above, it became clear to CIC network researchers that a new standard for high-performance computer networks was needed. The work covered by this LDRD effort focused on looking at the available high-end network technologies and using the knowledge gained, coupled with the requirements detailed above, to help Los Alamos chart its future path for network research. The work was mostly performed by the authors as well as other contacts within and outside the Laboratory.

Various network technologies were analyzed including the work done previously at Los Alamos on HIPPI. Other technologies that have been studied include the following:

- ATM - Asynchronous Transfer Mode: A technology that uses small, 53-byte, fixed-sized packets (cells) to transport information across an ATM network.
- SCI - Scalable Coherent Interface: A technology to expand the backplane of computers to higher speeds and longer distances.
- FCS - Fiber Channel Standard: A technology that combines channel connectivity for disks and other peripherals with fiber transport capabilities covering the ranges of SCSI, IPI3, and HIPPI.
- WDM- Wave Division Multiplexing: A technique to utilize more of the inherent bandwidth of fiber optics by sending various "channels" over different wavelengths of light.

The result of the studies of these technologies yielded the following conclusions based on the requirements for high-end networks to meet DOE needs:

- ATM - Will become the defacto wide area networking scheme but has congestion and cell loss problems that will affect its performance in local networks.
- SCI - Will be used to replace the internal computer systems interconnect bus and possibly to interconnect clusters but does not have the error-correcting capability required to minimize latency in general network use if errors occur.
- FCS - Too complex to be easily deployed for networking. A limited version may become the defacto standard for channel interconnect.
- WDM - Not mature enough to be useful for high-speed packet switched networks but holds tremendous promise for solving network demands in the future when links above 10 Gbit/sec are required.

In addition, a significant amount of effort was spent on design reviews of the E-Systems advanced multi-gigabit programs. This was felt to be useful since E-Systems had to develop a system that came close to matching the speed and latency requirements of DOE programs. The E-Systems switch may provide the switching requirement for DOE efforts as well, so LANL's participation is considered beneficial especially under the goals of the LDRD project. A Memorandum of Intent was executed with E-Systems for future exploration of gigabit networking. From this effort, new insight into the complexity of multi-gigabit switching was derived.

The remaining effort on this project was used to help define a new network research strategy that would provide the technology needed to meet the requirements and specifications in bandwidth and latency that were found above. This effort complimented the initial effort to provide an ANSI-standard proposal called HIPPI-64(X), which describes a new network based on 6.4 Gbit/sec links. This proposed ANSI standard is incorporating many of the useful technologies derived from the list of existing high-performance networks listed previously while also enhancing the key performance parameters, bandwidth and latency. A preliminary version of this emerging standard is included in this report.

Summary

The goal of this project was to identify the requirements for high-performance networks in the near future (5-10 years). Once the requirements were identified, the remaining work was to build on past expertise in high-end networking at Los Alamos through collaborations with other institutions and testbeds to evaluate network components, to develop relationships with computer and network vendors in support of these requirements, and to conduct early experiments on network switch and diagnostic techniques if any were available prior to the end of the project.

The new super-speed network requirements that have been determined fit in the 5 to 10 gigabit/sec range. This was based on input from application developers both at LANL and at other research institutes. In addition, another parameter, network latency, has also been identified as critical to these applications. A network latency of less than 1 microsecond is considered the upper bound for the applications to perform well.

An evaluation of current network technologies has been done under this project. Based on the criteria above, it was found that no one technology meets all of the requirements. Therefore, LANL researchers have been collaborating with computer and

networking vendors to develop a 6.4 gigabit network specification called HIPPI-6400.

Some simulation of error recovery techniques for such high-speed links has been done, but since no HIPPI-6400 hardware exists yet, no evaluation of components could be done.

Follow-on work to complete the HIPPI-6400 specification and efforts to build HIPPI-6400 analysis tools and switches have been started under separate funding. The latest version of the HIPPI-6400 specification, which came out of some work on this project, is included with this report.

preprint removed