The Supercomputer Supernet Testbed: A WDM-Based Supercomputer Interconnect

Leonard Kleinrock, *Fellow, IEEE,* Mario Gerla, *Member, IEEE,* Nicholas Bambos, Jason Cong, Eli Gafni, Larry Bergman, *Member, IEEE,* Joseph Bannister, *Senior Member, IEEE,* Steve P. Monacos, Theodore Bujewski, Po-Chi Hu, B. Kannan, Bruce Kwan, Emilio Leonardi, John Peck, Prasasth Palnati, and Simon Walton

Abstract—**Current fiber optic networks effectively provide local connectivity among end user computing devices, and can serve as backbone fabric between LAN subnets across campus and metropolitan areas. However, combining both stream service (in which ATM excels) and low latency datagram service (in which cluster networks like Myrinet and POLO excel) has been difficult to realize. This paper describes a new wavelength division multiplexed (WDM) fiber optic network that supports both stream and datagram service and extends reach and functionality of low-latency, high bandwidth workstation clusters to a campus and MAN setting. The novel concept is based on combining the rich interconnect structure of WDM fiber optics with the fast, low-latency mesh of crossbar switches recently developed for workstation groups. This system, called the Supercomputer Supernet (SSN) achieves a high level of performance by replacing the point-to-point copper wire links with a parallel channel (WDM) fiber optic interconnect system. The novel scheme interconnects asynchronous wormhole routing switches used in parallel supercomputers via multi-channel WDM fiber optic links embedded in to an optical star (or tree) "physical" topology. WDM will be used to subdivide the very large fiber bandwidth into several channels, each of Gb/s bandwidth. WDM channels (supporting also time division multiplexing) will be established between modules, thus defining a dense "virtual" interconnection topology, which is dynamically reconfigurable and responds to changing traffic patterns. A pool of channels will be set aside for direct, end-to-end connections between crossbars, providing circuit-switched service for real-time traffic applications.**

I. Introduction

THE Supercomputer Supernet (SSN) currently being developed at UCLA, JPL and Aerospace under ARPA support is a novel, high-performance, scalable optical interconnection network for supercomputers and workstation clusters based on asynchronous wormhole routing crossbar switches.

Manuscript received April 18, 1995. This work was supported by the Advanced Research Projects Agency (ARPA) of the U.S. Department of Defense under Contract DABT63-93-C-0055 and the University of California through an agreement with the National Aeronautics and Space Administration. . This paper was presented in part under the title "The Distributed Supercomputer Supernet—A Multi-Service Optical Intelligent Network." This work was performed by the Center for Space Microelecronics Technology at the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, and the University of California, Los Angeles.

L. Kleinrock, M. Gerla, N. Bambos, J. Cong, E. Gafni, P.-C. Hu, B. Kannan, B. Kwan, E. Leonardi, J. Peck, P. Palnati, and S. Walton are with the University of California, Los Angeles, CA 90024 USA.

L. Bergman and S. P. Monacos are with the Jet Propulsion Laboratory, Pasadena, CA 91109 USA.

J. Bannister and T. Bujewski are with the Aerospace Corporation, El Segundo, CA 90245 USA.

Publisher Item Identifier S 0733-8724(96)04553-7.

The WDM fiber optics extends the geographic coverage range from interdepartmental to campus and even to metropolitan areas. The network provides very high-speed multiple services, supporting hybrid circuit-switched and datagram traffic, and direct or multi-hop connections that are dynamically reconfigurable. At a first networking level, the crossbars locally interconnect workstations, supercomputers, peripheral devices, mass memory, etc. through host interfaces. At a higher networking level, the crossbars are fully interconnected with optical fibers supporting multiple wavelength division multiplexed channels, allowing communication between devices connected to distinct crossbars.

The resulting distributed SSN will be very fast—up to one gigabit per second (Gb/s) per channel—and possess a low latency approaching the physical media propagation time. It will scale up in the number of hosts connected and in geographical coverage for LAN and MAN situations. Using today's technologies, and being guided by emerging ones, the network design integrates the high throughput and parallelism of optics with the high intelligence of electronic processing, being clearly in line with, as well as at the front of modern networking trends.

A. New Enabling Applications

Just as conventional LAN technology enabled such present day applications as network file system (NFS), low-latency, high bandwidth <1 μs, >60 Mbytes/s [MB/s]) workstation cluster networks will likely enable new applications as well. These include the following:

- low cost arrayable video servers,
- distributed memory parallel supercomputers,
- network based memory management,
- real-time data acquisition and processing systems (e.g., radar), and
- video display walls.

One common characteristic of these applications is that they depend on extending the RAM memory of any one workstation in a cluster to any of the others in the group. Such network based distributed memory permit large applications to run beyond the confines of any one machine's memory space on a demand basis.

These functions have long been accepted as a minimum requirement to implement efficient message passing commu-

nications on MPP supercomputers, but only recently have been made possible over network interconnected workstations using a new breed of low latency (<1 μs) high bandwidth (>60 MB/s) networks that match the memory bandwidth and responsiveness of workstations.

Adding WDM fiber optics enhances this system in two ways. First, these cluster networks are extended from a machine room (100's m) to a campus setting (LAN), and secondly, the multiple channels of WDM fiber optics make it possible to support circuit switched type information services (e.g., voice, video) concurrently with conventional datagram services with maximum isolation.

In high end supercomputer networks, SSN will also provide the underlying low latency network fabric for interconnecting meta-supercomputer machines with scalable I/O; that is, I/O which is dynamically scalable in degree of parallelism from the host interface, network fabric, tertiary storage, and the application itself. This is particularly important for large data flow applications, such as radar or image processing, where the memory contents of MPP machines must be exchanged or updated on short time intervals commensurate with the real-time data source frame rate. The CASA gigabit testbed demonstrated the power of this meta-computing method using a single channel HIPPI network, but was not capable of setting up multiple channels between MPP machines on a dynamic basis over long distances. Hence, finer grain parallel applications with massive I/O requirements could not be attempted. Likewise, conventional telecom services, such as ATM, cannot efficiently provide this high bandwidth on demand without long setup delay.

B. Optical WDM Network Alternatives

Supercomputer networking and high-speed optical communications are active areas of research. Several optical networks have been proposed [2], and a few have been or are being implemented. Optical WDM network testbeds include LAMBDANET [9], Rainbow [6], the All Optical Network (AON) [8], and Lightning [7]. Two main alternatives have emerged in the design of WDM networks: single-hop and multihop networks.

Instead of using a direct path from source to destination, multihop networks [16] may require some packets to travel across several hops. For this reason, multihop networks are not suitable for high-throughput, real-time, delay-sensitive traffic. In fact, high data rates require loose flow control, which on the other hand gives limited protection against congestion. Alleviating congestion by dropping or deflecting messages is not an acceptable solution. Dropping messages is problematic in supercomputing since, at the high data rates involved, losing the contents of even a single buffer (which can exceed 64 kilobytes) is potentially disastrous. Deflection routing, on the other hand, introduces unpredictable delay and out-of-order delivery, which is intolerable given the high data rates used. Finally, multihop networks do not naturally support broadcast and multicast.

Single-hop networks [15] provide a dedicated, switchless path between each communicating pair of nodes. Each two-party communication requires one party to be aware of the other's request to communicate and to find a free virtual channel over which to communicate. This requires frequency-agile lasers and detectors over a broad range of the optical spectrum. The devices must also be capable of nanosecond reaction times. Furthermore, single-hop networks require substantial control and coordination overhead e.g. rendezvous control and dedicated out-of-band control channels). With current technology single-hop networks cannot readily accommodate bursty, short-lived communications.

Single-hop and multihop networks both suffer from limitations. The major single limitation of single-hop networks is the "complexity" of scaling up to large user populations and therefore high throughputs. If a single wavelength is used, then the throughput is limited by the maximum data rate achievable with affordable digital circuit technology. Capacity can be enhanced by using multiple wavelengths and implementing time and frequency division access schemes as in LAMBDANET and Rainbow. However, to achieve good efficiency in bursty traffic environments, these schemes require frequency-agile, rapidly tuned lasers and detectors over broad ranges of the optical spectrum. Such devices are not yet commercially available, although rapid progress of the technology in this direction has been reported [3]. Still, a major challenge is the production of components with both high tuning speed and broad wavelength range. Furthermore, the coordination of transceivers for short burst exchanges introduces considerable control overhead.

Like earlier testbeds, SSN employs WDM technology and high-speed transmission. However, the key feature that distinguishes the proposed SSN testbed from other approaches is the combination of multiple single-hop on-demand circuits with a multihop virtual embedded network to realize a hybrid architecture. This way, both stream traffic and low latency datagram traffic can be efficiently supported using the appropriate transport mechasnism. The SSN also allows for the dynamic reconfiguration of the virtual topology of its multihop component by slow retuning of its transceivers.

C. Outline of the Paper

In Section II, the SSN architecture will be described in more detail, beginning with the Myrinet high speed low latency LAN that has been developed by Myricom Inc. for workstation clustering application. SSN builds on this basic switching fabric by adding the WDM fiber links and an optical channel interface (OCI) controller for extending the connectivity from a building to a campus setting. In Section III, the overall SSN network protocol suite will be described for datagram and stream services, including datagram transfer, flow control, routing, multicasting, deadlock prevention, and dynamic reallocation of wavelengths for different services on the Myrinet system and the WDM optical backbone. In Section IV, performance studies will be described that analyze the topology tradeoffs, the scaling problem associated with wormhole routing (and a deadlock free routing technique) and the performance of multihop virtual topologies. Finally, in Section V, the SSN testbed for the LAN and MAN setting will be described.

II. Architecture

A. Myrinet

The Myrinet [24] high-speed LAN is an integral part of the SSN. Myrinet, manufactured by Myricom, Inc., is a high-speed, switch-based LAN intended to provide access over a limited geographical area. Myrinet has its roots in the multicomputer world [25], [4], where it was used as the interconnection network for a prototype parallel computer. It uses eight-bit–wide data paths between LAN elements, operating at a data rate of 640 megabits/second. The data channel is a full-duplex point-to-point link from a host interface to a switching node or between switching nodes. The Myrinet LAN transmits nine-bit symbols, eight of which carry data and one of which carries control information. Thus, in addition to data octets, several other nondata symbols are possible. The topology is arbitrary, being any configuration of interconnected host interfaces and switching nodes. Each switching node can have up to 16 ports. The Myrinet LAN has a limited spatial coverage, since the maximum link length is 25 m.

The Myrinet switches are simple, nonblocking switches that make switching decisions for a message by examining the source-supplied routing information in the message header. The complete route from the source node to a destination node is supplied to the requesting source node by a special route-manager software entity. Myrinet uses a form of cut-through routing called wormhole routing, in which the head of the message may arrive at its destination node before the tail has even left source node. This keeps latency very low. If an in-transit message is blocked at a switch, then the progress of the entire message is halted by backpressure. To reduce message latency these switches can switch a message in less than 600 ns.

The full-duplex channels use symbol-by-symbol stop-and-go flow control. Special STOP, GO, and IDLE symbols are available for controlling the flow of messages. Every Myrinet host and switch interface has a so-called slack buffer, which holds a small number of in-transit symbols. The size of the slack buffer is enough to hold twice as many symbols as can propagate simultaneously on a maximum-length link (27 symbols). When the receiving slack buffer has filled beyond a threshold, the receiver sends a STOP symbol to halt the incoming flow. Since it could take up to a full link-propagation delay for the STOP to arrive, the buffer must be able to absorb at least two link's worth of symbols beyond the threshold. When the sender receives the STOP, it immediately throttles its flow. When the switch's port unblocks and the slack buffer has drained below the threshold, the receiver sends a GO symbol to restart the flow.

A Myrinet message consists of a maximum of 5 600 000 symbols. Multicast is not supported in Myrinet. A source node would have to transmit to each destination node a copy of the multicast message.

B. OPTIMIC WDM Fabric

In the two level network architecture of the Supercomputer SuperNet (SSN) project, the second level is an optical backbone network that interconnects several high speed Myrinet LAN's. The optical channel interface (OCI) acts as an interface between the optical backbone and the high speed Myrinet LAN's. The physical optical backbone network can be any architecture–a single passive star coupler, a tree or a star of stars [11]. Space division multiplexing (using multiple fibers) is employed too.

The optical backbone network is viewed as a collection of a pool of wavelengths. By employing wavelength division multiplexing (WDM), the optical backbone network is designed to provide support for circuit switching, packet switching, multicasting and broadcasting.

Circuit switched service can be provided by dedicating a wavelength between two OCI's after an arbitration protocol. Packet switched service is provided by configuring the network as a multihop network [16]. Any scalable multihop virtual topology (like the traditional shufflenet [12] or the bidirectional shufflenet [18]) can be configured on the physical topology.

Since one of the goals of the SSN project is to extend the low-latency, high bandwidth supporting protocols of the high speed Myrinet LAN to the optical backbone network, the optical backbone should be able to support wormhole routing and the backpressure hop-by-hop flow control mechanism. The optical backbone can be configured into any topology that satisfies these requirements. Wormhole routing has a potential for deadlocks. The issue of deadlock prevention in SSN is addressed later.

C. OCI Design

An important contribution of the SSN project is the *optical channel interface* card. The goal of the OCI card is a modular design which focuses on the optical backbone protocol by using off-the-shelf Myrinet and optical components where possible. The OCI card extends the slack buffer concept of the Myrinet scheme to allow for long distance optical links between Myrinet clusters. Additionally, the OCI is responsible for arbitration and routing functions over the optical backbone.

The OCI card is a three port device. Two ports are used to connect one Myrinet port to one optical backbone port. The third port is a standard 9U VME connection for OCI configuration and system monitoring functions. A SPARC 32 b CPU card and one or more OCI boards will be packaged in a single VME enclosure to realize an integrated package for interfacing multiple Myrinet ports to the SSN optical backbone. Use of a VME based system allows for ease of integration by using an industry standard backplane for OCI integration. The OCI chassis is a 9U VME card cage used to house multiple OCI boards and a SPARC based CPU card for monitor and control. Fig. 1 shows the basic configuration of the OCI chassis.

The detailed OCI design provides the basis for the SSN optical backbone. The OCI consists of the dual LANai circuitry used as a Myrinet destination, the low-level data path monitoring logic, the flow control and routing logic, the VME interface, the worm buffers and the WDM fiber optic transceiver. By using *field programmable gate arrays* (FPGA)

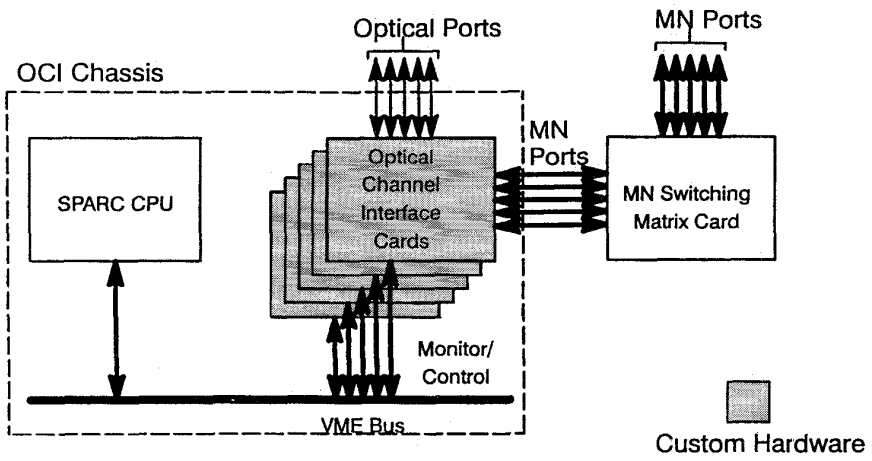

Fig. 1. Optical channel interface (OCI) chassis.

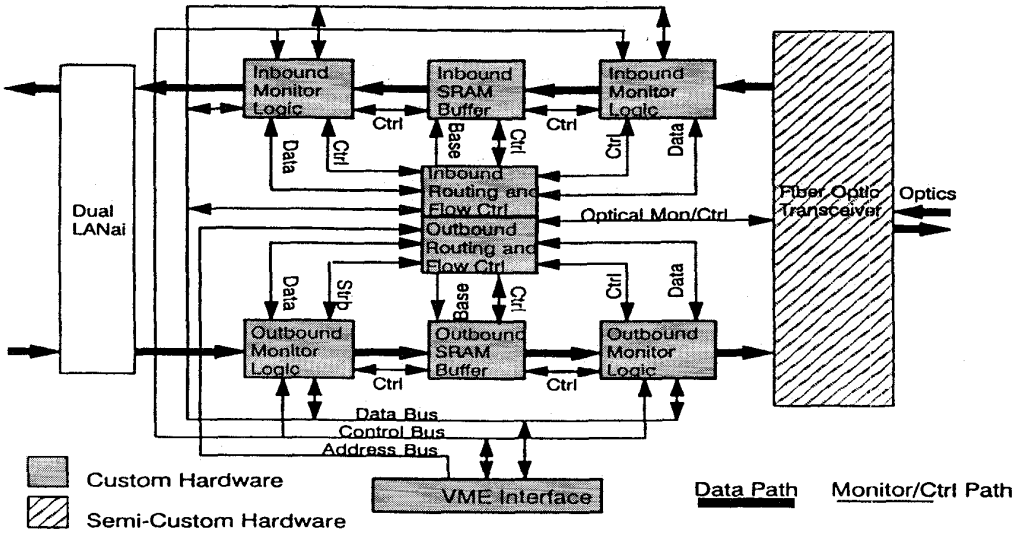

Fig. 2. OCI board block diagram.

for the monitor and control logic of the OCI, we allow for evolution of the design with the existing platform. The detailed block diagram of the OCI is shown in Fig. 2.

The dual LANai circuitry is a two port board which uses a pair of LANai routing processors to emulate a Myrinet source/destination while passing worms to/from the OCI board in a Myrinet format. One of the LANai processors of this circuitry provides the standard repertoire of Myrinet data and control bytes used to transmit and receive worms as part of a Myrinet [23]. The backend of this LANai looks like a DMA engine [23]. To simplify the interface to the backend of this LANai, a second LANai is used to generate a Myrinet like port [23] for transmission of worms to/from an OCI card.

The VME interface circuitry implements the VME protocol to communicate with the SPARC card. This circuitry also provides the data, address and interrupt information path ways from the various OCI logic blocks to a SPARC VME card for configuration of an OCI board and statistic collection operations.

The OCI board also contains worm buffers which are extensions of the slack buffers in the LANai. These buffers are needed due to the long propagation delay of the fiber optic links of the optical backbone. The OCI card also has the capability to handle worm priorities. This functionality is achieved by using random access memory (RAM) to allow for selection of worms based on worm priority. For a simple first-in-first-out buffering scheme, the delay through these buffers is in the neighborhood of one micro second.

The final component of the OCI is the WDM fiber optic transceiver. This hardware is similar to a standard transceiver but also provides the capability to transmit or receive at different optical wavelengths. Tuning at the transmitter will use tunable lasers or stepped wavelength laser arrays, while receiver tuning is accomplished with tunable filters. By pro-

viding the flow control and routing logic with the capability to use different wavelengths, we allow for multiple simultaneous communications over the same fiber optic media for improved optical backbone throughput and functionality.

III. Protocols

The key services provided by SSN are as follows.

1) low latency datagram service, to support fine grain distributed supercomputing. Variable size datagrams are allowed (as opposed to fixed size) in order to avoid segmentation/reassembly delay and overhead in origin and destination hosts.
2) high bandwidth, connection oriented service to support scientific visualization, large file transfers and more generally, time critical stream transmissions.

In the SSN project, protocols have been developed both in the Myrinet and in the optical backbone, in order to support the above basic services as well as additional services. The main function of the optical backbone is to extend geographically the reach of a Myrinet, and to permit the "transparent interconnection" of a large number of Myrinet islands.

In the remainder of the section, we briefly review the Myrinet protocols, and then focus on the optical WDM network protocols and network control and management procedures.

A. Myrinet Protocols

The commercial Myrinet already comes equipped with protocols for the support of datagram service. Source routing and backpressure flow control allow efficient transfer of datagrams in the Myrinet. In the SSN program, additional protocols are being implemented to provide:

1) integrated packet and circuit switched service support
2) bandwidth allocation to circuit switched traffic
3) "intelligent", alternate routing to minimize blocking and reduce latency (more generally "congestion management")
4) multicasting
5) priority support and QoS enforcement for different traffic classes.

Some of these protocols have been extensively evaluated via analysis and simulation. Performance results (along with more detailed description of the protocols) are reported in Section IV.

B. WDM Optical Backbone Protocols

Optical backbone protocols must extend transparently the Myrinet services end-to-end while achieving efficient utilization (and reallocation) of expensive backbone resources, efficient scaling, congestion protection and fault tolerance. The WDM optical star/tree architecture is ideally suited to this set of requirements in that it allows high bandwidth interconnection, efficient integration of multiple services and flexible reallocation of channel/bandwidth resources. The optical fabric will be initially a combination of space and wavelength division multiplexing; later in the project, it will be enriched to support also T/WDMA (time and wavelength division multiple access).

The following are the key protocols supported in the optical backbone.

1) **C/S Protocol:** Initially, a separate wavelength/fiber will be allocated to each C/S connection. When T/WDMA will be available, multiple connections with possibly different data rates will be carried in each WDM channel. An efficient technique for supporting multirate connections (based on receiver pipelining and slot retuning) was reported in [13].
2) **Datagram Transfer Protocol:** Two basic options are available here: namely, the single hop scheme (with wavelength retuning at transmitter and receiver on a datagram by datagram basis) and the multihop scheme. In our testbed, we will pursue the multihop scheme, which is less demanding in terms of transmitter/receiver tunability. Later, we will also consider single hop schemes. Section IV presents a comparison of these alternatives.
3) **Flow Control:** The Myrinet backpressure type flow control is extended to the optical backbone by using large slack buffers in the OCI's. Furthermore, virtual channels have been defined on each individual link of the multihop network, so that a single backpressured worm does not clog the link.
4) **Routing Protocol:** Two options are available for routing: the "flat" source routing option, which is an end-to-end extension of the Myrinet routing scheme, and the two-level routing option, where separate routing schemes are used for Myrinet and optical backbone. The latter scheme is more scalable, and offers better flexibility in backbone routing, at the expense of additional implementation complexity (in the OCI). As part of the latter scheme, deflection routing will also be explored.
5) **Multicasting:** is supported both for C/S connections and P/S transfers. Recall that C/S connections are single hop ("broadcast and select"). Thus, the signal can be received by all the OCI's in the multi-cast group, by simply tuning to the transmission wavelength at the proper slot. In the multihop network, multicasting can be achieved by define a proper multicast tree (embedded in the virtual multihop topology). Alternatively, an hamiltonian loop, visiting all the OCI's in the multicast group, can be used.
6) **Priorities:** are implemented in order to handle datagram traffic with different QoS. Furthermore, priority is given to transit traffic (over entry traffic) in the multihop network, in order to maintain the backbone clear of congestion, and stop the overload at the entry points.
7) **Deadlock Prevention:** Deadlocks may occur in the transfer of datagrams in the backbone network. These would bring the entire network to a halt, and therefore must be prevented. Several schemes are now being considered, and are discussed and evaluated in more detail in Section IV.
8) **Dynamic Reallocation of Wavelengths:** to different services. The WDM architecture offers the unique op-

portunity to reallocate bandwidth resources to different services (in our case, C/S and P/S service) based on user requirements. Furthermore, the multihop virtual topology can be dynamically "tuned" to obtain the best match with the current traffic pattern. Previous studies have shown that topology readjustments can lead to significant throughput and delay improvements. Initially, the dynamic reallocation and topology tuning will be in terms of actual wavelength. With the introduction of T/WDMA, finer grain, more efficient reallocation can be achieved. Protocols for dynamic reallocation are currently under development.

IV. Protocol and Performance Studies

There are several research issues that are brought about by the unique two level architecture of SSN. To start, performance models and simulation studies have been developed for improving the performance of the low-latency, high-bandwidth electronic network (Myrinet). We have developed several algorithms that extend the operating region of the Myrinet network to better match the large aggregate throughput of the attached WDM optical network. The algorithms involve refinement of the the timeout scheme to help increase throughput.

Another important component of research is the development of protocols for implementation in the optical WDM fabric. The issue of providing deadlock-free routes in the optical backbone as well as across the whole network has been studied [17]. Further, the properties of the bidirectional shufflenet multihop topology (a derivative of the traditional shufflenet with bidirectional links) have been studied [19], [18]. Bidirectional shufflenet allows for easy extension of the hop-by-hop backpressure flow control mechanism of the electronic LAN to the optical WDM fabric besides increased throughput and shorter average hops.

In this section, we present a few results from the studies on the performance of Myrinet and enhancements to improve its performance. Then, the issue of deadlock-free routing in SSN is addressed. Finally, some results for virtual multihop topologies embedded in the WDM optical backbone are presented. All performance results were obtained using a simulator written in Maisie [20], [1].

A. Enhancing Performance of the Myrinet Network

We have been developing algorithms to enhance the performance of the low-latency, high speed network that supports cluster computing. This must be done in order to adequately harness the large aggregate throughput of the attached WDM optical network. Myrinet implements wormhole switching. Due to low latency, Myrinet provides an effective interface for the hosts and supercomputers to the optical portion of the network. The ideal operating region for wormhole switching networks is in the low load traffic region. Blocking is minimal and thus latency remains low. However, as the traffic load increases, more worms block. A feedback blocking effect occurs and message delay rises rapidly.

To increase the network's effective operating region, we have developed different algorithms to handle the timeout parameter. In Myrinet, timeout is used to break deadlocks that may occur in the system due to spurious network errors resulting in misrouted worms. Whenever a worm arrives at a switch, a counter is activated if its desired output port is busy. After some timeout value, a reset signal is sent out that resets the entire network. Currently, the timeout value is set to a very large value (50ms) and the reset mechanism clears all worms in the network. Studies have been done on a refined timeout mechanism where smaller timeout values are used to selectively clear only the worm that has been blocked for a time longer than the timeout value of the system ([10]). The results show that small timeout values better optimize the system (see Fig. 3).

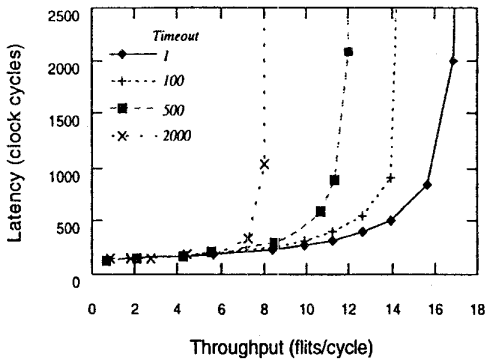

Fig. 3. Performance of a 3×3 torus network with different timeout values (worm length = 100 flits).

In a different timeout parameter refinement, we implement a scheme using switch state information to determine the timeout value for a blocked worm [14]. When a worm arrives at a switch and the output port it desires is busy with a worm that is itself blocked, the newly arriving worm times out and is retransmitted. Otherwise, if the arriving worm sees that the worm currently using the output port is flowing, it blocks and waits for the flowing worm to be completely transmitted. The motivation behind switch state dependent timeouts (SSD TO) comes from the observation of the feedback blocking effect of a wormhole switching network under medium to high traffic loads. As the traffic load rises, more worms encounter blocked output ports. After a certain threshold, several worms clog the network and prevent other worms from flowing. Using the SSD TO algorithm, only worms that are waiting behind other worms that are making forward progress are allowed to block and wait. All other worms are forced to timeout. The SSD TO improves delay performance over the original large timeout scheme where all worms must block and wait unless a deadlock occurs (see Fig. 4).

B. Deadlock Free Routing in SSN

Wormhole routing can cause deadlocks to happen in the network if a cycle of worms blocking each other exists. Deadlock resolution using timeouts is a costly procedure requiring a network reset. Deadlock free routing is thus desirable for SSN.

Deadlock free routing can be done on any bidirectional topology by forming a spanning tree of the topology and then

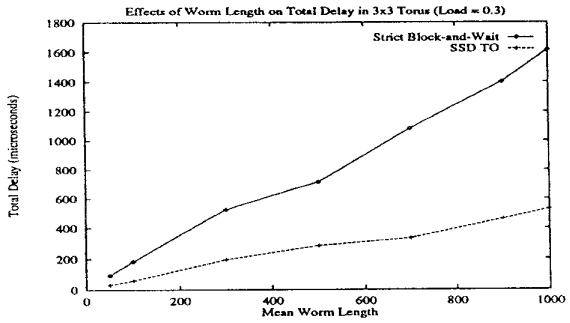

Fig. 4. Delay performance of the 3 × 3 torus network with Switch State Dependent Timeout with varying worm lengths.

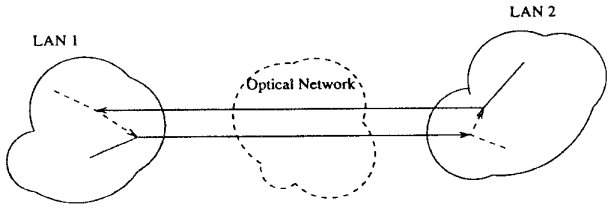

Fig. 5. An example showing that deadlock free routing in each part of the network does not imply a deadlock free network. The solid lines represent a worm destined to a remote LAN that are in a remote LAN; the dashed lines represent local/remote traffic still in the origin LAN. The arrows indicate where blocking occurs.

routing worms on this spanning tree such that the worm first travels zero or more links in the Up direction (toward the root) and then zero or more links in the Down direction (away from the root) toward the destination. This technique is called Up/Down routing.

Another technique for deadlock free routing, on unidirectional as well as bidirectional topologies, is the use of virtual channels [5]. In this approach, each link is decomposed into several virtual channels (all sharing the same bandwidth but each having its own slack buffer). These virtual channels are then connected together to form several levels of virtual networks. After eliminating cycles from each level, a fixed ordering is made among these virtual networks. This fixed ordering determines the routing. An advantage of using virtual channels is that shortest path routing is possible unlike in the Up/Down routing approach.

For the multihop virtual topologies of the traditional shufflenet and the bidirectional shufflenet (in which the links of the shufflenet are bidirectional), deadlock free routing techniques employing Up/Down routing (as in Autonet [22], [21]) and virtual channels have been described elsewhere [17]. Here we present the approach used to perform deadlock free routing in the two level architecture of the SSN project, assuming that the virtual multihop topology implemented in the optical backbone network is the bidirectional shufflenet.

For the high speed electronic mesh network, deadlock free routing can be achieved by employing the technique of the Up/Down routing on a spanning tree of the network. For the optical network, we can use Up/Down routing or virtual channels to achieve deadlock free routing in the bidirectional shufflenet. However, having deadlock free routing schemes in the different parts of the network, does not guarantee deadlock freedom for the network as a whole. An example is shown in Fig. 5.

Several alternatives for having deadlock free routing across the network exist. One way is to use Up/Down routing on a single spanning tree for the whole network. In this approach, either one of the OCI's or one of the stations on the net can behave as the root of the spanning tree. This alternative works correctly since we have bidirectional links in SSN. However, this approach has the drawback that shortest paths are not guaranteed.

Second, we could extend the virtual channel mapping method across the whole network. Namely, we achieve deadlock free routing in the electronic LAN's by dedicating separate sets of virtual channels to origin and remote LAN traffic.[1] Thus, origin LAN traffic does not interfere with remote LAN traffic providing deadlock free routing across the whole SSN network. Unfortunately, Myrinet LAN's currently do not support virtual channels with separate slack buffers for each virtual channel.

A third solution consists of requiring that the OCI in each subnet be the root of the spanning tree used to define the up/down links for that subnet. Then, deadlock prevention can be achieved for the subnet with up/down routing as discussed earlier. For the bidirectional shufflenet deadlock prevention can be achieved using either the Up/Down approach or the virtual channel approach. For deadlock free routing across the whole network, it suffices to require that remote traffic (i.e. a connection between two distinct subnets) be routed UP in the origin LAN and DOWN in the destination LAN. Of course, choosing the OCI as the root of the spanning tree for the subnet is key to this solution. Deadlock prevention is evident from inspection of Fig. 5 where conflict between local and remote worms could not exist with this new rule. Note that this deadlock free routing scheme can be easily extended to the case in which an electronic LAN is connected to the optical backbone via multiple OCI's. In this case, one of the OCI's is elected as the root of the spanning tree (using, for example, a distributed procedure similar to the IEEE 802.1 spanning tree bridge algorithm). Each OCI can then handle remote traffic to and from sources and destinations in its subtree. Of course, the 'root' OCI could handle all remote traffic by itself. However, hosts will be assigned to different OCI's using appropriate criteria e.g. shortest distance) in order to balance the traffic across the electronic LAN, OCI and the bidirectional shufflenet.

One important advantage of the last solution (over the previous two) is that it allows the independent choice of the most cost effective deadlock free routing scheme for each level of the SSN architecture.

C. Performance of Multihop Topologies

Since fast tunable receivers are not yet widely available, studies of the performance of multihop topologies of shuf-

[1] Here, traffic originating within a LAN destined to a host in the same LAN is called local traffic. If the destination is a host on a remote LAN then the traffic is called remote traffic. If the worm is traveling through the LAN to which the origin host belongs then it is said to be in the origin LAN otherwise it is said to be in a remote LAN.

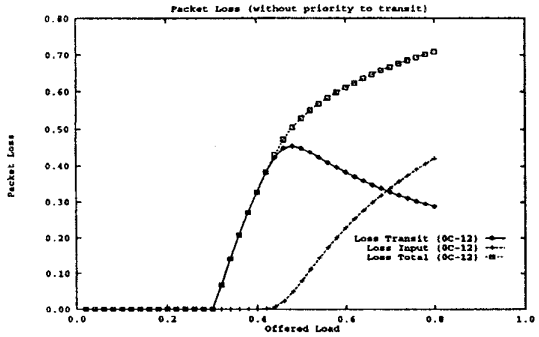

Fig. 6. Packet loss vs Offered load in the case when no priority is given to transit traffic. An 8 node shufflenet topology was used.

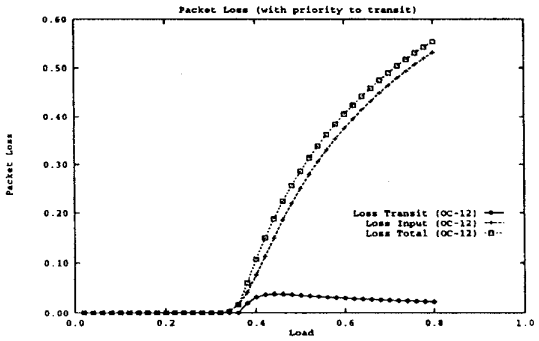

Fig. 7. Packet loss vs Offered load in the case when priority is given to transit traffic. An 8 node shufflenet topology was used.

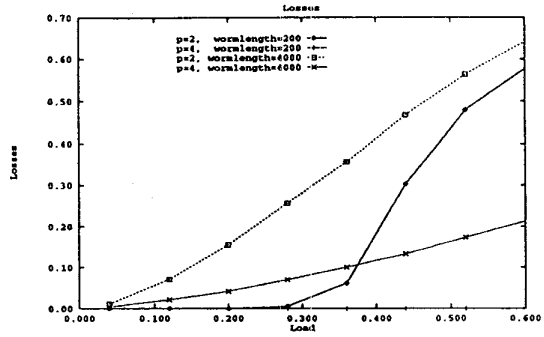

Fig. 8. Worm Loss Percentage versus Offered Load for a 24 node shufflenet and a 32 node shufflenet. The loss is shown for worms of average length 200 and 4000 bytes.

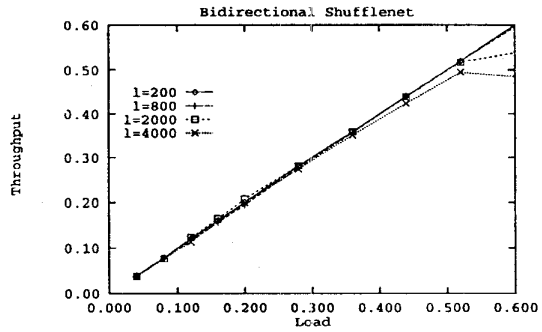

Fig. 9. Throughput versus Offered Load for a 24 node bidirectional shufflenet. Average worm lengths range from 200 to 4000 bytes.

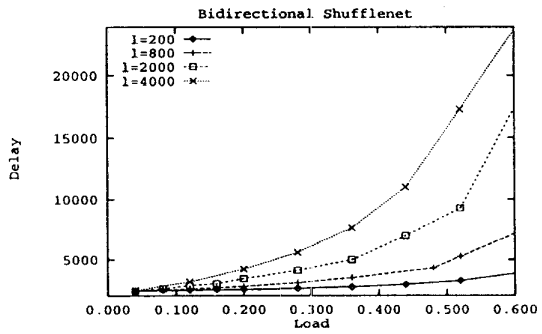

Fig. 10. Delay versus Offered Load for a 24 node bidirectional shufflenet. Average worm lengths range from 200 to 4000 bytes.

flenet and bidirectional shufflenet have been carried out via simulation and analysis as well. Though the WDM optical backbone network can be configured as any virtual multihop topology, we have only studied the traditional shufflenet and the bidirectional shufflenet topologies.

1) Performance of Shufflenet: We studied the effect of giving priority to transit traffic (i.e., traffic which is passing through an intermediate station) over input traffic (i.e, traffic entering the station from the Myrinet LAN. In this study, no flow control was assumed, thus, there was packet loss due to buffer overflow. The speed for the optical link was assumed to be the OC-12 rate (622 Mb/s). In Figs. 6 and 7, we plot the results for the two cases. Fig. 6 shows that in the absence of priority to transit traffic, packet loss starts at about 0.3 offered load and the transit packets are dropped first. However, when priority is given to transit traffic, the loss starts at about 0.35 offered load and almost all the loss is at the input. Since the hop-by-hop backpressure flow control mechanism employed by Myrinet is being extended to the WDM optical backbone fabric, there would be no loss at the input. Thus, the significant result of this study is that giving priority to transit traffic in the traditional shufflenet gives higher throughputs and lower losses.

2) Performance of Bidirectional Shufflenet: A bidirectional shufflenet is a traditional shufflenet with bidirectional links. Each station in the bidirectional shufflenet has twice the number of fixed tuned transmitters and receivers than a station in the traditional shufflenet. The bidirectional nature of this multihop virtual topology makes it easy to extend the hop-by-hop backpressure flow control mechanism to the WDM optical backbone. Also, this topology has a shorter average hop length. In the absence of flow control, the traditional shufflenet shows worm loss (see Fig. 8). In the bidirectional shufflenet, worm loss does not occur because of the flow control mechanism. In Fig. 6, we show the throughput obtained for a 24 node bidirectional shufflenet topology for different average worm lengths. In Fig. 9, we show the delay performance for different worm lengths. The significant results of this study are that the use of bidirectional shufflenet's natural support for flow control eliminates loss and the throughput and delay performances are quite good.

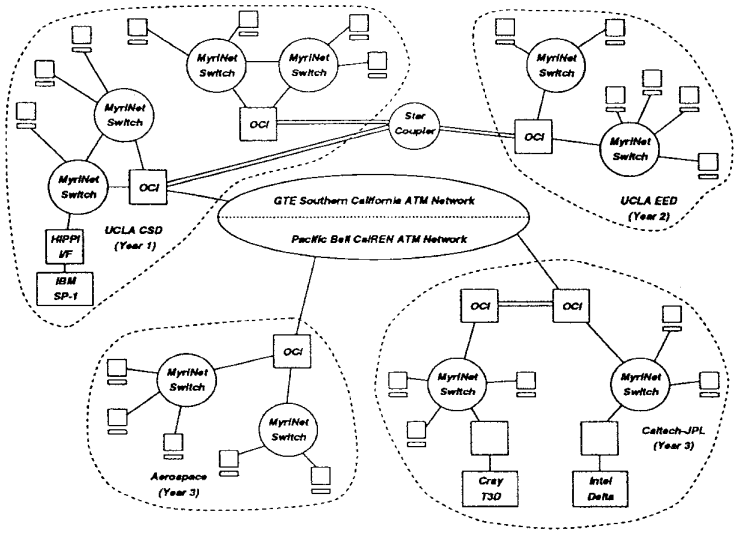

Fig. 11. The SSN testbed consists of four clusters of Myrinet switches: four at the UCLA computer science department, two at the electrical engineering department, two in Pasadena (one at JPL and one at Caltech), and two at Aerospace Corporation.

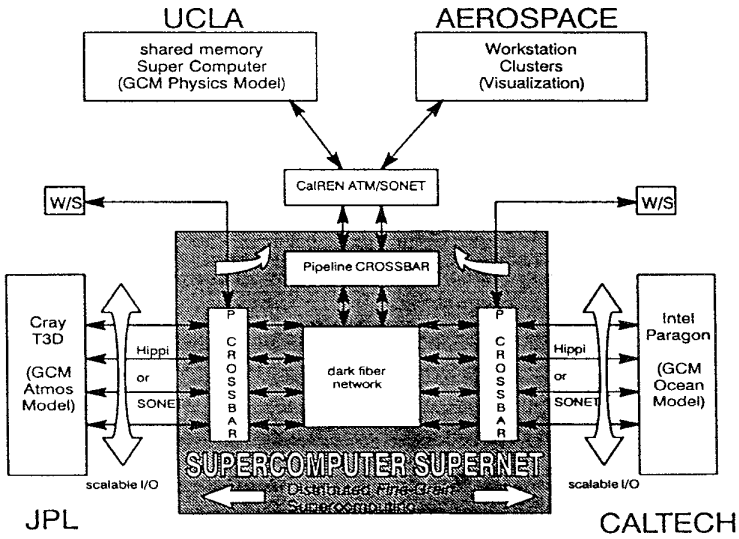

Fig. 12. SSN supports a fine-grain distributed supercomputing and visualization GCM application using Myrinet on one segment of the CASA testbed between JPL and Caltech.

V. Applications and Testbeds

A. Applications

The low-latency, dynamic reconfigurability, and scalability of SSN are expected to enable several new applications in the area of distributed supercomputing and visualization:

1) *Fine Grain Meta-Supercomputer:* The SSN attributes would accelerate the evolution of a network-based operating system with precise synchronization of dispersed processes, fine grain process management on 100's–1000's of processor elements, distributed checkpointing of jobs, and dynamic entry of new hosts.
2) *Real Time Distributed Network Operating System:* Low and predictable (bounded) latency makes SSN well suited for wide area network control and data acquisition applications. Examples include missile tracking, radar telemetry, remote robot control, and in the commercial arena, oil refinery and power plant control, avionics and spacecraft control systems, control of electrical power distribution systems, and factory automation.
3) *Distributed Image Data Base Perusal:* Scientific image-based data-base archival and perusal systems are now being developed in several efforts, such as the UC Sequoia effort and the MAGIC testbed. NASA applications, such as EOS, will require the capability of perusing through terabytes of data very quickly and interactively. A low latency high throughput network will be essential for responding quickly to interactive

control from the user (datagram) and sending image bursts back to the user (streams/circuit switched).

B. Testbeds

1) Campus: The basic SSN testbed topology is shown in Fig. 11. The Myrinet switches are placed in three clusters: a group of four in the UCLA Computer Science department building, a group of two in the UCLA Electrical Engineering department building, two at JPL/Caltech (between two supercomputers), and finally, two at the Aerospace Corporation. OCI's interconnect the clusters as well as selected ports within the largest cluster at the UCLA Computer Science Department.

One fiber optic link segment (14 km) of the CASA gigabit network between JPL and Caltech in the Pasadena area is proposed as the target SSN testbed demonstration site using scalable I/O supercomputers (see Fig. 12). The proposed SSN application that combines elements of (1) and (2) above is the UCLA Global Climate Model (GCM) being developed by R. Mechoso for the CASA project. On the present CASA network, a single HIPPI channel only permits a coarse-grain coupling of the ocean/atmosphere model between the Caltech Intel DELTA (running the ocean model) and JPL Cray YMP (running the atmospheric model). Running over the existing dark fiber, SSN would provide four times the capacity (3.2 Gbit/s) and lower latency routing between the two supercomputers than the present single HIPPI channel with Crossbar Interfaces. This would provide a foundation for a finer grain decomposition of the GCM application. Simultaneously, high performance workstations can interactively capture image results of the running GCM model and peruse through new data sets that would be staged for later GCM runs. The SSN network dynamically allocates/deallocates optical channel bandwidth as workstations or massively parallel processor (MPP) nodes enter/leave the network. The Myrinet APCS network node also accommodates instantaneous reconfiguration of the MPP I/O channels from asynchronous I/O for separate partitioned jobs (e.g., one per quadrant of the MPP) to coherently striped I/O for one large single job.

2) MAN: Between UCLA and Pasadena (a distance of about 30 miles), and between UCLA and Aerospace (a distance of about 15 miles) the network fabric will consist of a single ATM/SONET OC-3 channel provided by the Pacific Bell CalREN (California Research and Education Network) and GTE consortiums (see Fig. 11). At each location, an OCI will be configured to provide a gateway function by incorporating a SONET/ATM network interface with the OCI SPARC CPU controller. A higher performance solution is being explored with a leading routing vendor. Initially, permanent virtual circuits (PVC) will be used between the sites. Striped channel performance, which SSN provides via WDM in a LAN campus setting, can be provided by setting up multiple PVC's and/or SONET OC-3 channels.

VI. Conclusion

As fine grain, closely coupled real-time distributed system applications begin to mature for cluster workstation computing and networking of metamassively parallel processor supercomputers, low-latency rapidly reconfigurable networks with high Gb/s per channel capacity will be required. SSN provides one such network fabric for binding these systems together that is easily scalable in both physical size and number of ports per host. It is also adaptable to a variety of optical transmission techniques, providing multiple growth paths as WDM and spatial optical multiplexing optoelectronics becomes commercially available. Such networks also raise a host of new issues in network management, flow and congestion control, and error recovery that will be the subject of future work.

References

[1] R. Bagrodia, Y. Chen, M. Gerla, B. Kwan, J. Martin, P. Palnati, and S. Walton, "Scalable simulation of a high-speed wormhole network in a parallel language," to be published.

[2] J. Bannister, M. Gerla, and M. Kovačvić, "Routing in optical networks," in *Routing in Communications Networks*, M. Steenstrup, Ed. Englewood Cliffs, NJ: Prentice-Hall, 1995, pp. 187–225.

[3] C. A. Bracket, "Dense wavelength division multiplexing networks: Principles and applications," *IEEE J. Select. Areas Commun.*, vol. 8, pp. 948–964, Aug. 1990.

[4] D. Cohen, G. Finn, R. Felderman, and A. DeSchon, "The ATOMIC LAN," in *IEEE Workshop High Perform. Commun. Subsyst.*, Tucson, AZ, Feb. 1992.

[5] W. J. Dally and C. L. Seitz, "Deadlock-free message routing in multiprocessor interconnection networks," *IEEE Trans. Comput.*, vol. C-36, pp. 547–553, May 1987.

[6] N. R. Dono, P. E. Green, K. Liu, R. Ramaswami, and F. F. Tong, "A wavelength division multiple access network for computer communication," *IEEE J. Select. Areas Commun.*, vol. 8, pp. 983–994, Aug. 1990.

[7] P. W. Dowd, K. Bogineni, K. A. Aly, and J. Perreault, "Hierarchical scalable photonic architectures for high-performance processor interconnection," *IEEE Trans. Comput.*, vol. 42, pp. 1105–1120, Sept. 1993.

[8] S. B. Alexander *et al.* "A precompetitive consortium on wide-band all-optical networks,"*J. Lightwave Technol.*, vol. 11, pp. 714–735, May/June 1993.

[9] M. S. Goodman, H. Kobrinski, M. Vecchi, R. M. Bulley, and J. L. Gimlett, "The lambdanet multiwavelength network: Architecture, applications, and demonstrations," *IEEE J. Select. Areas Commun.*, vol. 8, pp. 995–1004, Aug. 1990.

[10] P.-C. Hu and L. Kleinrock, "A queueing model for wormhole routing with timeout," 1995, presented at the Fourth Int. Conf. Comput. Commun. Networks.

[11] B. Kannan, S. Fotedar, and M. Gerla, "A two level optical star WDM metropolitan area network," in *Proc. GLOBECOM 94 Conf.*, June 1994.

[12] M. Karol and S. Shaikh, "A simple adaptive routing scheme for schufflenet multihop lightwave networks," in *Proc. GLOBECOM 88 Conf.*, 1988, pp. 1640–1647.

[13] M. Kovacevic, M. Gerla, and J. Bannister, "Time and wavelength division multiple access with acoustooptic tunable filters," *Fiber and Integr. Opt.*, vol. 12, no. 2, pp. 113–132, 1993.

[14] B. Kwan and N. Bambos, "Performance of a switch state dependent timeout scheme in a wormhole switching lan," Univ. California, Los Angeles, Tech. Rep. UCLA ENG-95-121, Apr. 1995.

[15] B. Mukherjee, "Wdm-based Local lightwave networks—Part I: Single-hop systems," *IEEE Network*, vol. 6, pp. 12–27, May 1992.

[16] ——, "Wdm-based Local lightwave networks—Part II: Multi-hop systems," *IEEE Network*, vol. 6, pp. 20–32, July 1992.

[17] P. Palnati, M. Gerla, and E. Leonardi, "Deadlock-free Routing in a Hierarchical Supercomputer Interconnection Network," 1995. Submitted for publication.

[18] P. Palnati, E. Leonardi, B. Kannan, and M. Gerla, "Bidirectional Shufflenet: A Multihop Topology for Backpressure Flow Control," to be presented at the Fourth International Conf. on Computer Communications and Networks, 1995.

[19] ——, "Performance analysis of bidirectional shufflenet: A multihop topology for backpressure flow control." to be published.

[20] R. Bagrodia, K. M. Chandy, and J. A Misra, "Message-based approach to discrete-event simulation," *IEEE Trans. Software Eng.*, vol. 13, no. 6, June 1987.

[21] T. L. Rodeheffer, "Experience with Autonet," *Comput. Networks ISDN Syst.*, pp. 623–629, 1993.

[22] M. D. Schroeder, A. D. Birrell, M. Burrows, H. Murray, R. M. Needham, T. L. Rodeheffer, E. H. Satterthwaite, and C. P. Thacker, "Autonet: A high-speed self-configuring local area network using point-to-point links," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 1318–1335, Oct. 1991.
[23] C. Seitz, Private Communication.
[24] C. Seitz, D. Cohen, and R. Felderman, "Myrinet—A gigabit-per-second local-area network," *IEEE Micro*, vol. 15, no. 1, pp. 29–36, Feb. 1995.
[25] C. L. Seitz, J. Seizovic, and W.-K. Su, "The design of the Caltech Mosaic C multicomputer," in *Proc. Symp. Integr. Syst.*, Mar. 1993.

Leonard Kleinrock (S'55–M'64–SM'71–F'73) received the B.S. degree in electrical engineering from the City College of New York in 1957 and the M.S.E.E. and Ph.D.E.E. degrees from the Massachusetts Institute of Technology in 1959 and 1963, respectively.

He has been a Professor of Computer Science, at the University of California, Los Angeles, since 1963. His research interests focus on performance evaluation of high speed networks and parallel and distributed systems. He has had over 180 papers published and is the author of five books. He is the principal investigator for the ARPA Advanced Networking and Distributed Systems grant at UCLA. He is also founder and CEO of Technology Transfer Institute, a computer-communications seminar and consulting organization located in Santa Monica, CA.

Dr. Kleinrock is a member of the National Academy of Engineering, is a Guggenheim Fellow, and was a founding member of the Computer Science and Telecommunications Board of the National Research Council. He has received numerous best paper and teaching awards, including the ICC 1978 Prize Winning Paper Award, the 1976 Lanchester Prize for outstanding work in Operations Research, and the Communications Society 1975 Leonard G. Abraham Prize Paper Award. In 1982, he received the Townsend Harris Medal. Also in 1982, he was co-winner of the L. M. Ericsson Prize, presented by His Majesty King Carl Gustaf of Sweden, for his outstanding contribution in packet switching technology. In July of 1986, Dr. Kleinrock received the 12th Marconi International Fellowship Award, presented by His Royal Highness Prince Albert, brother of King Baudoin of Belgium, for his pioneering work in the field of computer networks. In the same year, he received the UCLA Outstanding Teacher Award. In 1990, he received the ACM SIGCOMM award recognizing his seminal role in developing methods for analyzing packet network technology.

Mario Gerla (M'75) received the graduate degree in electrical engineering from Politecnico di Milano, Italy, in 1966 and the M.S. and Ph.D. degrees in computer science from University of California, Los Angeles in 1970 and 1973, respectively.

He is a Professor of Computer Science at UCLA. From 1973 to 1976, he was a manager in Network Analysis Corporation, Glen Cove, NY, where he was involved in several computer network design projects for both government and industry, including performance analysis and topological updating of the ARPANET under a contract from DoD. From 1976 to 1977, he was with Tran Telecommunication, Los Angeles, CA, where he participated in the development of an integrated packet and circuit network. Since 1977, he has been on the Faculty of the Computer Science Department of UCLA. His research interests include the design, performance evaluation, and control of distributed computer communication systems and networks. His current research projects cover the following areas: topology design and bandwidth allocation in ATM networks, design and implementation of optical interconnects for supercomputer applications, design and performance evaluation of air/ground wireless communications for the Aeronautical Telecommunications Network, and network protocol design and implementation for a mobile, integrated services wireless radio network.

Nicholas Bambos received the Ph.D. degree in electrical engineering and computer science from the University of California, Berkeley in 1989.

He joined the Electrical Engineering Department of the University of California, Los Angeles in 1989, and he is currently an Associate Professor. His research interests include performance analysis and evaluation of communication networks and distributed systems, queueing systems, stochastic process, and resource allocation in random environemnts.

Jason Cong received the Ph.D. degree in computer science from the University of Illinois at Urbana-Champaign in 1990.

Currently, he is an Associate Professor and co-Director of the VLSI CAD Laboratory in the Computer Science Department of University of California, Los Angeles. His research interests include computer-aided design of VLSI circuits and systems, VLSI interconnect design and optimization, rapid system prototyping, and configurable computing using FPGA's.

Eli Gafni received the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology, Cambridge in 1982. Since then, he has been on the faculty of the University of California at Los Angeles. His research interests are in distributed systems and protocols and computer science theory.

Larry Bergman (S'72–M'77) received the B.S.E.E degree from California State university, San Luis Obispo, in 1973, the M.S. degree from California Institute of Technology, Pasadena, in 1974, and the Ph.D. degree from Chalmers University of Technology, Gothenburg, Sweden, in 1983, all in electrical engineering.

In 1975, he joined the Jet Propulsion Laboratory, Pasadena, CA, where he contributed to the early development of low power spacecraft data buses and computer networks. Since 1981, he has focused on the development of multigigabit fiber optic local area networks, terabit all-optical computer networks, fiber optic sensors, and holographic optical interconnects for VLSI chips. He has authored over 65 papers in the fields of fiber optics and high-speed communications, received four patents, and also has lectured on optics and telecommunications are UCLA. In 1993, he earned the Technology and Applications Program (TAP) Directorate Exceptional Service Award for sustained research contributions to the fields of fiber optic network and supercomputer communications. He presently supervises a research group in high-speed optical systems, and is also the project engineer for the JPL Supercomputer.

Dr. Bergman is a member of AES, SMPTE, Tau Beta Pi, Eta Kappa Nu, Phi Kapa Phi, and Sigma Xi.

Joseph Bannister (M'80–SM'95) received the Ph.D. degree in computer science, the M.S. degree in computer science and the M.S. degree in electrical engineering, all from University of California, Los Angeles. He received the B.A. degree in mathematics from the University of Virginia.

Since 1988 he has been with The Aerospace Corporation, a Federally Funded Research and Development Center in Los Angeles, CA, where he is the Director of the Computer Systems Research Department, managing a group that performs applied research in high-speed computing, programming environments, computer architecture, and advanced networking. His current research concentrates on the design, analysis, and implementation of communication systems to support high-speed computation.

Dr. Bannister is a member of Sigma Xi, AAAS, and ACM SIGCOMM.

Steve P. Monacos for a biography, see this issue, p. 1369.

Theodore Bujewski received the B.S. and M.S. degrees in operations research from Case Western Reserve, and the M.S. degree in computer science from the University of California at Los Angeles.

Prior to starting graduate school at the University of Chicago, he was employed by the Aerospace Corporation.

Po-Chi Hu received the M.S. degree with a study of the protocol design for optical networks in 1993. He is currently pursuing the Ph.D. degree as a doctoral student in the Computer Science Department at the University of California at Los Angeles. His Ph.D. dissertation will focus on the performance evaluation of wormhole routing.

He has attended the graduate program of UCLA Computer Science Department since 1991. His research interests include the design of optical networks, ATM switching, wireless communication, and queueing theory.

B. Kannan is pursuing the Ph.D. degree in the Computer Science Department at the University of California at Los Angeles.

His research interests include computer communications networks modeling and performance evaluation, distributed systems, and queueing theory.

Bruce Kwan received the M.S. degree from the University of California at Los Angeles (UCLA) in 1991 and is currently pursuing the Ph.D. degree as a doctoral student in the Department of Electrical Engineering at UCLA. His Ph.D. work is on the performance evaluation of wormhole routing network protocols.

His research interests include high-speed network protocols, queueing theory, and parallel and distributed systems.

Emilio Leonardi is currently pursuing the Ph.D. degree in the Electronics Department at Politecnico di Torino, Turin, Italy.

Since September 1994, he spent one year at Computer Science Department, the University of California at Los Angeles (UCLA), being involved in the SSN (Supercomputer-Supernet Project). His current research interests are in the field of performance evaluation of optical networks, queueing theory, and formal description techniques.

John Peck is currently pursuing the Ph.D. degree in the Computer Science Department at the University of California at Los Angeles (UCLA).

His research topics include logic synthesis systems for SRAM-based FPGA's and custom computing using FPGA's.

Prasasth Palnati received the M.Sc(Tech.) degree in computer science degree from the Birla Institute of Technology and Science, Pilani, India, in 1990, and the M.S. degree in computer science from the University of Maryland, Baltimore County, in 1992. He is currently pursuing the Ph.D. degree as a doctoral student in the computer science Department at the University of California at Los Angeles (UCLA) working on his dissertation titled "Protocols for an optical multihop virtual topology with support for wormhole routing."

His research interests include the design and analysis of optical networks and protocols for high-speed networks.

Simon Walton received the M.S. degree in computer science from the University of California at Los Angeles in 1994. He is currently pursuing the Ph.D. degree as a doctoral student with the computer science department.

His research interests include efficient intrahost data-copying and multicasting.