

- [3] J. A. Bannister, M. Gerla, and M. Kovačević. Routing in Optical Networks. In Martha Steenstrup, editor, *Routing in Communications Networks*, pages 187–225. Prentice Hall, Englewood Cliffs, New Jersey, 1995.
- [4] D. Cohen, G. Finn, R. Felderman, and A. DeSchon. ATOMIC: A Very-High-Speed Local Area Network. In *IEEE Workshop on High-Speed Communication systems*, Tuscon AZ, February 1992.
- [5] D. Cohen, G. Finn, R. Felderman, and A. DeSchon. The ATOMIC LAN. In *IEEE Workshop on High Performance Communications Subsystems*, Tucson, Arizona, February 1992.
- [6] W. J. Dally and C. L. Seitz. Deadlock-Free Message Routing in Multiprocessor Interconnection Networks. *IEEE Transactions on Computers*, C-36(5):547–553, May 1987.
- [7] N.R. Dono, P.E. Green, K. Liu, R. Ramaswami, and F.F.Tong. A Wavelength Division Multiple Access Network for Computer Communication. *IEEE Journal on Selected Areas in Communications*, 8:983–994, August 1990.
- [8] P. W. Dowd, K. Bogineni, K. A. Aly, and J. Perreault. Hierarchical Scalable Photonic Architectures for High-performance Processor Interconnection. *IEEE Transactions on Computers*, 42(9):1105–1120, September 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 Journal on Selected Areas in Communications*, 8:995–1004, August 1990.
- [10] B. Kannan, S. Fotedar, and M. Gerla. A Two Level Optical Star WDM Metropolitan Area Network. In *Proceedings of GLOBECOM 94*, pages 563–566, November 1994.
- [11] M. Karol and S. Shaikh. A Simple Adaptive Routing Scheme for Shufflenet Multihop Lightwave Networks. In *Proceedings of GLOBECOM 88*, pages 1640–1647, 1988.
- [12] M Kovacevic, M. Gerla, and J. Bannister. Time and Wavelength Division Multiple Access with Acoustooptic Tunable Filters. *Fiber and Integrated Optics* **12(2)**, pages 113–132, 1993.
- [13] E. Leonardi, F. Neri, P. Palnati, and M. Gerla. Congestion Control Techniques in Asynchronous Wormhole Routing Networks. Technical Report CSD-950065, UCLA, 1995.
- [14] B. Mukherjee. WDM-based Local Lightwave Networks Part I: Single-Hop Systems. *IEEE Network*, 6(3):12–27, May 1992.
- [15] B. Mukherjee. WDM-based Local Lightwave Networks Part II: Multi-Hop Systems. *IEEE Network*, 6(4):20–32, July 1992.
- [16] L. M. Ni and P. K. McKinley. A Survey of Wormhole Routing Techniques in Direct Networks. *IEEE Computer*, 26(2):62–76, February 1993.
- [17] P. Palnati, M. Gerla, and E. Leonardi. Deadlock-free Routing in an Optical Interconnect for High-Speed Wormhole Routing Networks. Technical Report CSD-950064, UCLA, 1995.
- [18] P. Palnati, E. Leonardi, and M. Gerla. Bidirectional Shufflenet: A Multihop Topology for Backpressure Flow Control. In *Proceedings of 4th International Conference on Computer Communications and Networks*, pages 74–81, September 1995.
- [19] 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 Journal on Selected Areas in Communications*, 9(8):1318–1335, October 1991.
- [20] C. Seitz. Private Communication.
- [21] C. Seitz, D. Cohen, and R. Felderman. Myrinet—A Gigabit-per-second Local-Area Network. *IEEE Micro*, 15(1):29–36, February 1995.
- [22] C. Seitz, J. Seizovic, and W. Su. The Design of the Caltech Mosaic C Multicomputer. In *Proceedings of the Symposium on Integrated Systems*, March 1993.

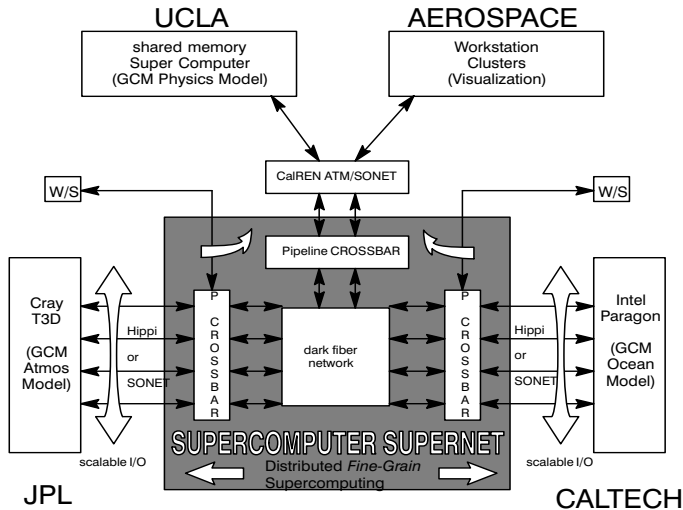


Figure 7: SSN supports a fine-grain distributed supercomputing and visualization GCM application using Myrinet on one segment of the CASA testbed between JPL and Caltech.

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 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.

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 point-to-point ATM/SONET OC-3 channels provided by the Pacific Bell CalREN (California Research and Education Network) and GTE consortiums (see Fig. 6). At each location, an OCI can be configured to provide a gateway function by incorporating a SONE/ATM network interface with the OCI SPARC CPU controller. A higher performance solution is also being explored with a separate Myrinet/ATM gateway. Initially, permanent virtual circuits (PVC) will be used between the sites. Striped channel perfor-

mance, which SSN provides via WDM in a LAN campus setting, can be provided by setting up multiple PVCs and/or SONE OC-3 channels.

6 Conclusion

As fine grain, closely coupled real-time distributed system applications begin to mature for cluster workstation computing and networking of meta-massively 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.

Acknowledgment

The research described in this paper was carried out at the Jet Propulsion Laboratory, California Institute of Technology, the University of California, Los Angeles, and The Aerospace Corporation, and was sponsored by the Advanced Research Projects Agency (ARPA) of the U.S. Department of Defense under Contract DABT63-93-C-0055 and University of California through an agreement with the National Aeronautics and Space Administration.

Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government, University of California, Jet Propulsion Laboratory, California Institute of Technology, or The Aerospace Corporation.

References

- [1] S. Alexander et al. A Precompetitive Consortium on Wide-Band All-Optical Networks. *Journal of Lightwave Technology*, 11(5/6):714-735, May/June 1993.
- [2] R. Bagrodia, K.M. Chandy, and J. Misra. A Message-Based Approach to Discrete-Event Simulation. *IEEE Transactions on Software Engineering*, 13(6), June 1987.

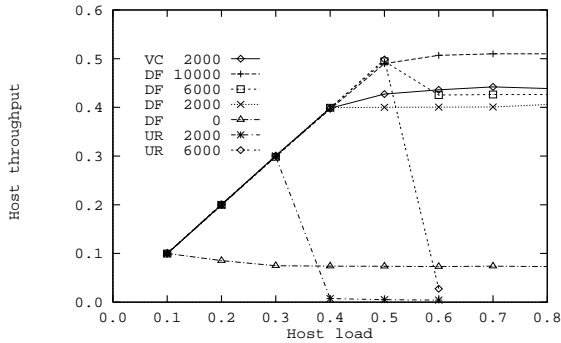


Figure 5: Host throughput for an 8×8 bidirectional grid. Average worm size is 500; maximum worm size is 10000.

the links (which is 400 bytes for a 1 Km link with speed of light assumed to be 2×10^8 m/s). Fig. 5 shows host throughputs for this network configuration. Seven curves are shown, for different flow control techniques, and different slack buffer sizes. One curve (labeled VC) refers to the virtual channel approach, with slack buffers of 2000 bytes (slightly more than four times the link propagation delay) on each virtual channel. Three virtual channels per link are needed in this topology to avoid deadlocks while allowing shortest path routing. Two other curves (labeled UR) refer again to backpressure flow control, but with unrestricted shortest path routing, i.e., without virtual channels or Up/Down routing. In this situation deadlocks occur regardless of buffer size. The remaining 4 curves (labeled DF) refer to deflection routing, with deflection buffer size equal to 0 (pure asynchronous deflection), 2000, 6000, and 10000 bytes (delayed deflection), respectively.

The conclusion is that unrestricted shortest path routing results in deadlocks and pure asynchronous deflection gives bad throughput. However, delayed deflection routing compares quite favorably with the virtual-channel based deadlock-free shortest path routing with backpressure flow control. Similar results were obtained for the bidirectional shufflenet topology. This implies that the SSN could use either approach to control congestion in the optical backbone network – the deciding factors would be the implementation complexity and tolerance for the drawbacks of deflection routing – unbounded delays and out-of-order delivery.

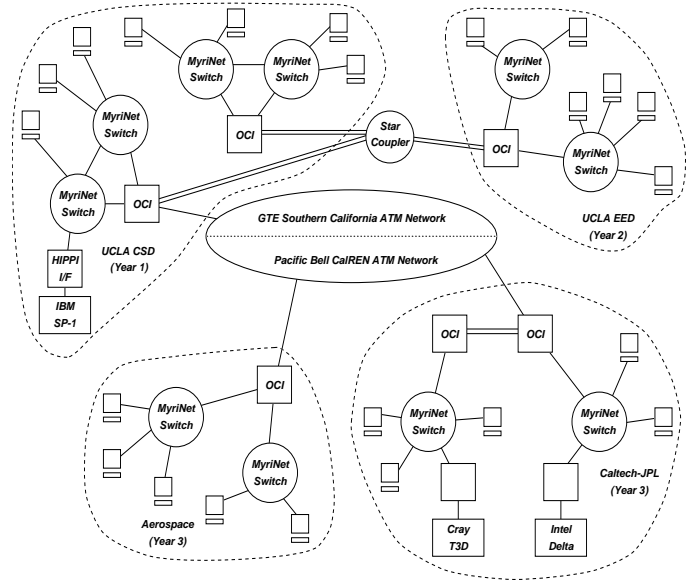


Figure 6: The SSN testbed consists of four clusters of MyriNet switches: four at the UCLA computer science department, two at the electrical engineering department, two in Pasadena (one at JPL and one at Caltech), and two at Aerospace Corporation.

5 Testbed

The basic SSN testbed topology is shown in Fig. 6. 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. OCIs interconnect the clusters as well as selected ports within the largest cluster at the UCLA Computer Science Department.

One fiber optic link segment (14km) 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. 7). The proposed SSN application 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

the normal way, and in addition, will be retransmitted to the next host in the sequence. If the worm is transmitted back to the originator of the multicast message, this serves as an explicit acknowledgement of the successful reception of the multicast message by all hosts of the multicast group. This last transmission could be omitted too.

- **Unrooted Tree:** In this scheme multicasting is performed over an unrooted tree. The tree is formed over the host connectivity graph. Each node of the tree represents a host, and an incoming multicast worm is retransmitted by the host to its neighbors in the tree. There is one spanning tree for each multicast group.

A Hamiltonian circuit multicast scheme such as the one just outlined has been implemented. In order to reduce the load on the host’s kernel and peripheral bus the retransmission of multicast worms is done entirely within the network interface card by using the software that runs on the card’s processor. A control process running on the host as a normal application informs the network interface (via its device driver) of the pertinent multicast information like the multicast group, next host in the multicast group etc. The remainder of the operation is performed by the device driver (for originating hosts only) and the network interface software. This control process, the multicast group manager, is currently a stub process but it is expected to develop into a more complex process that will interact with multicast group managers on other hosts and with the IP group management protocol. Under this implementation, multicast worms are not returned to their originating host but stop at the previous node in the circuit.

The two alternative schemes have been simulated using the SSN simulator. The average multicast latency was used as a measure of network loading and performance. An 8×8 torus was simulated. The average worm length was 400 bytes. 10% of the worms generated were multicast worms. Ten multicast groups, each with ten members, were simulated, with the members chosen at random. The resultant latencies are shown in Figure 4. These results indicate that the unrooted tree scheme produces less loading of the network than the Hamiltonian-circuit-based scheme, as expected.

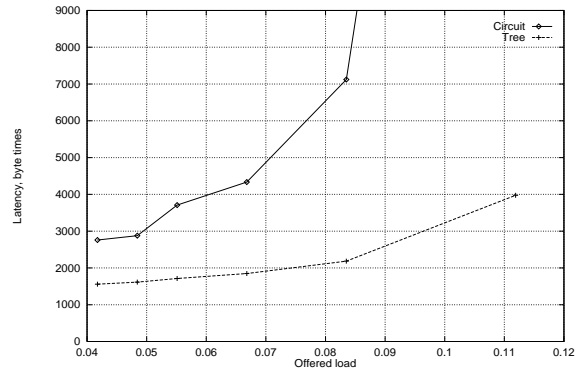


Figure 4: Average multicast latency against network load on 8×8 torus

4.2.2 Congestion Control in the Optical Backbone

Backpressure flow control spreads information about congestion by ‘freezing’ the worm in place (possibly stretching across several links and slack buffers (in switches)). In deflection routing, a worm for which the required output port is busy, is deflected on a free output port. This can cause out-of-sequence delivery and unbounded delays. However, deflection routing implicitly informs hosts about congestion in the network by monitoring the traffic at the local switch and observing the number of hops that the passing worms have undergone as a result of deflection. A higher number of hops traveled indicates a higher level of congestion. Thus, backpressure flow control and deflection routing are possible alternative solutions to the problem of congestion in the network.

Wormhole routing with backpressure flow control can lead to deadlocks [6, 17]. Hence, we use the virtual channel approach or the Up/Down routing approach [17]. Deflection routing could result in a livelock, thus we use a deflection buffer (DF) to store some of the worm and probabilistically break livelocks (a procedure that we call delayed deflection) [13]. With delayed deflection, the routing decision on an incoming worm is delayed until an output port becomes available.

We have compared deflection and backpressure, using simulation. The experimental optical network configuration is the 8×8 bidirectional grid with an average worm size equal to 500. This means that worms are short with respect to the propagation delay on

optic transceiver. The stepped wavelength laser diode arrays (see Fig. 3) have four side by side single mode DFB lasers made on a single substrate, with each laser emitting light at a slightly different wavelength in the 1.55 μm region. The laser design is an InP-based ridge waveguide laser. Due to the simplicity of fabrication and less stringent fabrication tolerances compared to buried heterostructure lasers, ridge waveguide lasers are seen to have a strong potential for commercial use.

The laser wafers were prepared by atmospheric pressure metal-organic chemical vapor deposition (MOCVD) on (100)-oriented n+ InP substrates. The active region consists of four compressively strained (e=1%) InGaAsP quantum wells, each 94 Angstrom wide, with 150 Angstrom barriers of InGaAsP. Fabrication of this material into 4-element DFB laser arrays requires e-beam writing of the diffraction gratings, an MOCVD regrowth, and the fabrication of the ridge waveguide structure. The top four layers of the laser structure (contact, 2 InP layers, and etch stop) are removed in order to define the distributed feedback grating in the SCH region. The pitch of the grating for the individual lasers is determined by the modal index and the design criteria of four wavelengths in the range from 1.54-1.56 μm (to be compatible with erbium doped fiber amplifiers). This leads to four grating pitches in the range from 2375 - 2400 Angstroms.

This hardware provides the capability to transmit or receive at different optical wavelengths. By providing the flow control and routing logic with the capability to use different wavelengths, multiple simultaneous communications over the same fiber optic media for improved optical backbone throughput and functionality are possible.

4 Performance Results

In this section, we describe the simulation platform that has been developed for evaluating protocols for SSN. We also present a few selected performance results that are illustrative of the evolutionary methodology prior to design.

4.1 Simulation Platform

A modular simulator of the SSN has been developed to experiment with various routing, flow-control, and interconnection strategies in a hierarchical, reconfigurable network. The detailed model comprises more than 5000 lines of Maisie (a C-based message-passing discrete event simulation language [2]) code. Innova-

tive features of this simulator, based on the capabilities of Maisie, include:

- byte-level rather than packet-level simulation for accurate modeling of wormhole routing
- dynamic network reconfigurability
- potential for parallel execution to improve performance
- scalable and modular design

The SSN simulator operates mainly through emulation; that is, the operation of the code mirrors the actual operation of the network components being simulated. Although not necessarily the most efficient approach, this method is simple to implement and makes modification of the simulator to reflect protocol enhancements straightforward. Another advantage of this approach is that it makes it easy to port algorithms from simulation to actual implementation.

4.2 Results

The SSN simulator has been used extensively to compare and evaluate protocols for the high-speed electronic network, the optical backbone network and the entire two-level network. Results from two such studies are included here. The first study compares two alternative approaches to providing multicast service support in Myrinet. The second one evaluates deflection routing in the optical backbone network as a congestion control technique and contrasts it with the backpressure flow control approach.

4.2.1 Multicasting in Myrinet

As Myrinet does not implement multicasting, we wish to provide multicasting support that efficiently utilizes link capacity and host resources. The basic approach is that the host retransmits received multicast worms to other hosts within the same multicast group. This approach requires no modification of the Myrinet hardware. Two schemes for such retransmissions were investigated:

- **Hamiltonian circuit:** In this scheme a directed circuit amongst all hosts in a multicast group is defined using a graph representing the network (called host connectivity graph). To perform multicasting a source host sends each multicast worm to the next host in the circuit; at this host the worm will pass up through the protocol stack in

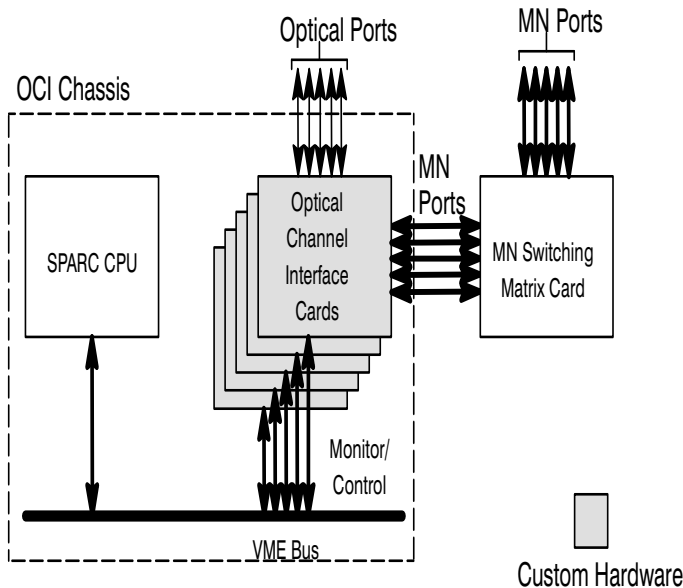


Figure 1: Optical Channel Interface (OCI) Chassis.

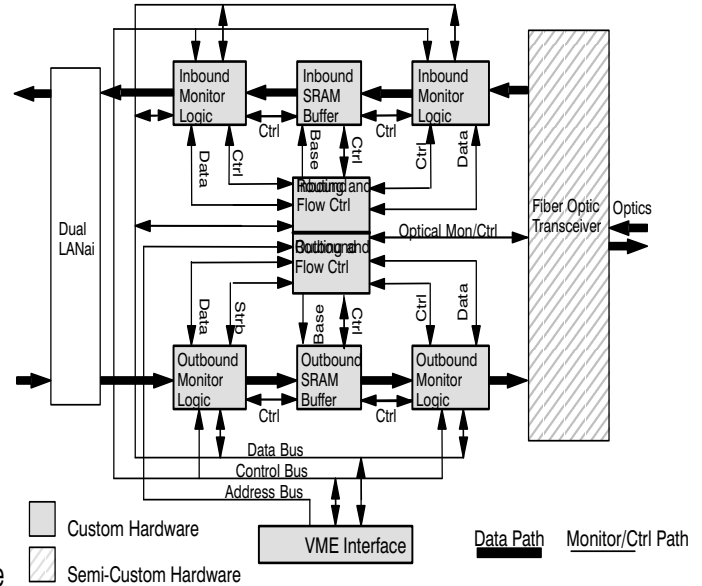


Figure 2: OCI board block diagram.

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.

The OCI consists of the dual-LANai circuitry¹ 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) for the monitor and control logic of the OCI, evolution of the design is possible. 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 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 [20]. The backend of this LANai looks like a DMA engine [20]. To simplify the interface to the backend of this LANai, a second LANai is used to generate a Myrinet like port [20] 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 inter-

¹LANai is a Myricom, Inc. processor that is used in the Myrinet host interface cards.

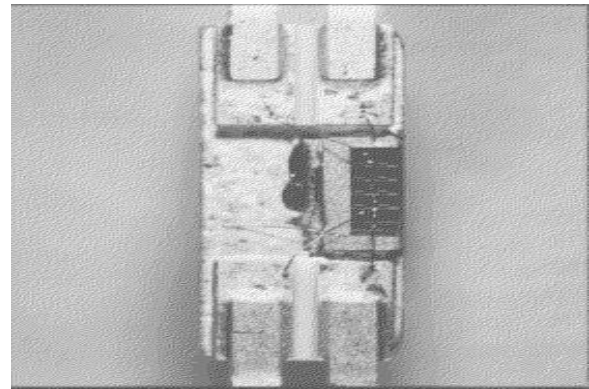


Figure 3: Stepped wavelength laser diode array.

rupt information pathways 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 provides 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 microsecond.

The final component of the OCI is the WDM fiber

to provide a term of comparison.

- (c) **flow control.** The Myrinet backpressure type flow control is extended to the optical backbone by using large slack buffers in the OCIs. 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.
- (d) **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 2-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 has also been explored [13].
- (e) **multicasting** is supported both for guaranteed bandwidth services and datagram transfers. Recall that guaranteed bandwidth connections are single hop (“broadcast and select”). Thus, the signal can be received by all the OCIs in the multicast group, by simply tuning to the transmission wavelength at the proper slot. In the multihop network, multicasting can be achieved by defining a proper multicast tree (embedded in the virtual multihop topology). Dynamic tree maintenance with receiver initiated join is being considered. Another alternative is an hamiltonian loop, visiting all the OCIs in the multicast group.
- (f) **priorities** are implemented in order to handle datagram traffic with different QoS. Furthermore, priority is given to transit traffic (over entry traffic), in order to maintain the multihop backbone clear of congestion, and stop the overload at the entry points.
- (g) **deadlock prevention.** Wormhole routing is prone to deadlocks. A deadlock would bring the entire network to a halt, and therefore must be prevented. The main options are to perform up/down routing on a spanning tree and to use virtual channels to allow deadlock-free shortest paths. These schemes are discussed and evaluated in more detail in [17].

- (h) **dynamic reallocation of wavelengths** to different services. The WDM architecture offers the unique opportunity to reallocate bandwidth resources to different services (in our case, guaranteed bandwidth and datagram 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.

3.4 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.

3.4.1 Functionality

The OCI ‘bridges’ the two levels of the SSN network architecture. The OCI’s basic function is to interface the high-speed Myrinet LAN with the optical backbone network. Thus, the OCI has a Myrinet interface and an optical network interface. Since wormhole routing and backpressure flow control are extended to the optical backbone network, the OCI must provide sufficient buffering to prevent data loss due to buffer overflow. The OCI is responsible for electro-optic and opto-electronic conversions; it provides controls for optical transmitters/receivers; it implements the WDM optical backbone protocols (including arbitration and routing) and, it collects performance measurements.

3.4.2 Hardware Description

The OCI card is a three port device (see Fig. 1). One port connects to Myrinet. Another port connects to the optical backbone. The third port is a standard 9U VME connection for OCI configuration and system monitoring functions. A SPARC 32-bit 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

throttles its flow. When the switch's port unblocks and the slack buffer has drained below another threshold, the receiver sends a GO symbol to restart the flow.

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. Myrinet does not, however, provide support for guaranteed bandwidth services. Further, multicast is not implemented in Myrinet. A source node would have to transmit to each destination node a separate copy of the multicast message. To overcome these deficiencies, additional protocols are being implemented in the SSN program to provide:

- (a) integrated datagram and guaranteed bandwidth service support
- (b) bandwidth allocation to guaranteed bandwidth traffic
- (c) "intelligent", alternate routing to minimize blocking and reduce latency (more generally "congestion management")
- (d) multicasting
- (e) priority support and QoS enforcement for different traffic classes.

Some of these protocols have been extensively evaluated via analysis and simulation. Selected performance results are reported in section 4.

3.2 Optical WDM Fabric

The second level of the Supercomputer SuperNet (SSN) is an optical backbone network that interconnects several high-speed Myrinet LANs. The Optical Channel Interface (OCI) acts as an interface between the optical backbone and the high-speed Myrinet LANs. The physical optical backbone network can be any architecture – a single passive star coupler, a tree or a star of stars [10]. Space Division Multiplexing (using multiple fibers) is employed too.

The optical backbone network can be viewed as a pool of wavelengths. By employing WDM, the optical backbone "virtual" topology is configured to provide support for guaranteed bandwidth traffic, packet switching, multicasting and broadcasting.

Guaranteed bandwidth service can be provided by dedicating a wavelength between two OCIs after an

arbitration protocol. Datagram service is provided by creating a virtual multihop network [15]. Any scalable multihop topology (like the traditional shufflenet [11] or the bidirectional shufflenet [18]) can be configured on the physical topology.

Since the primary goal of the SSN project is to extend the Myrinet low-latency, high-bandwidth protocols to the optical backbone, the latter must be equipped to support wormhole routing and backpressure flow control. The optical backbone can be arbitrary as long as it satisfies these requirements. One consequence of enforcing backpressure is that bidirectional links are required. Thus, the bidirectional shufflenet is preferable to the traditional shufflenet.

3.3 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:

- (a) **Guaranteed bandwidth protocol.** Initially, a separate wavelength/fiber will be allocated to each 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 multi-rate connections (based on receiver pipelining and slot retuning) was reported in [12].
- (b) **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. Single hop will be studied via simulation

2.5 The Integration Challenge

From the review of the electronic and optical supercomputer interconnection options, we can conclude that each type of network is best suited to certain functions. In SSN, for example, the high-speed electronic network provides a low-latency, high-bandwidth datagram service employing wormhole routing and backpressure flow control. Support for guaranteed bandwidth service, multicasting and scalable I/O is limited. The geographic coverage is limited to a few hundred meters. Also, the congestion caused by traffic imbalance cannot be effectively prevented. The SSN optical backbone network supports both guaranteed bandwidth service (via single-hop) and datagram service (via multi-hop) but with different degrees of implementation complexity. Reconfigurability of the optical network helps to overcome unbalanced traffic patterns. Again, a robust, distributed reconfiguration algorithm must be developed.

Thus, the integration of the electronic LAN and the optical backbone for achieving the design goals outlined earlier is a challenging task. The challenge is two-fold: to design the appropriate hardware interface (optical channel interface (OCI)) and to develop protocols that enable the key requirements of distributed supercomputer or workstation cluster computing. For example, the low-latency service of the electronic LAN is extended to the optical backbone by implementing wormhole routing in the backbone as well. Likewise, the backpressure flow control mechanism is extended to the backbone to prevent loss of data due to buffer overflows. We describe the design of the OCI and the protocols in the following section.

3 SSN Architecture

The Supercomputer Supernet (SSN) has a two-level architecture. The lower level consists of a high-speed wormhole-routing electronic LAN, called Myrinet. The higher level of the SSN architecture consists of an optical WDM backbone network that interconnects several lower level LANs. In this section, we present, in some detail, the architecture of SSN.

3.1 Myrinet

Myrinet [21], 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 [22, 5], where it was

used as the interconnection network for the Cosmic Cube, a prototype parallel computer. It uses eight-bit-wide data paths between LAN elements, operating at a data rate of 640 Mb/s. 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.

Messages are transmitted in Myrinet as variable sized worms. Worms are composed of flow-control digits, flits (in Myrinet, a flit is a byte). Worms can be up to 5.6 MB long. Each worm has a head portion, a data portion and a tail portion. The head of the worm contains a source route (i.e., list of switch port numbers). This information is used by the simple, non-blocking Myrinet switches to make switching decisions. 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. Since Myrinet uses a form of cut-through routing called wormhole routing [16], 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 achieve low latency, these switches switch a message (i.e., process and route the header) in less than 600 nanoseconds.

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

(Gb/s) per channel) that integrates guaranteed bandwidth service and datagram service. In high-end supercomputer networks, SSN will also provide the underlying low latency network fabric for interconnecting meta-supercomputer machines with scalable I/O. The CASA gigabit testbed demonstrated the power of this meta-computing method using a single channel HIPPI network, but lacked the capability 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.

2.3 Optical Interconnection Options

Supercomputer high-speed optical interconnection is an active area of research. Several optical networks have been proposed [3], and a few have been or are being implemented. Optical WDM network testbeds include LAMBDANET [9], Rainbow [7], the All Optical Network (AON) [1], and Lightning [8]. Two main alternatives have emerged in the design of WDM networks: single-hop and multihop networks. Here we briefly outline them since they are important to SSN's approach to provide services.

Single-hop networks [14] 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 are adequate for circuit-switched type traffic, but cannot readily accommodate bursty, short-lived communications.

Instead of using a direct path from source to destination, **multihop** networks [15] may require some packets to travel across two or more hops with buffering at intermediate nodes. Because of random queue fluctuations, multihop networks are not suitable for high-throughput, real-time, delay-sensitive traffic. In fact, high data rates require a very streamlined flow

control, which in turn gives limited protection against congestion and buffer overflow. Alleviating congestion by dropping or deflecting messages is not a desirable solution. Dropping messages is problematic in supercomputing since, at high data rates, 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. In summary, multihop networks are adequate for datagram traffic, but are not well suited for high speed real time streams.

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 mechanism. The SSN also allows for the dynamic reconfiguration of the multihop topology by slow retuning of its transceivers.

2.4 Electronic Interconnection Options

Several electronic supercomputer interconnection networks have been researched in the recent past. The Los Alamos MCN (Multiple Crossbar Network) consists of a collection of Cross Bar Switches connected by HIPPI channels. High throughput and low latency goals are achieved by using a fast packet switching physical layer, connectionless protocol to set up a path for each packet (or group of packets). Scalability and high HIPPI connection set up latency are drawbacks of this scheme. Nectar is a network of crossbar switches that supports both guaranteed bandwidth and datagram service (for small packets ≤ 1 Kbyte). The short packet size and uncontrolled sharing of bandwidth between the guaranteed bandwidth service and the datagram service are potential problems. ATOMIC [4] is a network of crossbar switches interconnected by point-to-point links supporting wormhole routing. Autonet [19] is also a crossbar based network that supports datagram traffic with low latency and high throughput by employing cut-through routing. Thus neither network provides support for guaranteed bandwidth service.

In Section 2, the supercomputer interconnection environment is reviewed with emphasis on supercomputer applications as well as interconnection options both in the optical and electronic domain. In Section 3, the SSN architecture is described, beginning with the Myrinet high-speed, wormhole-routing LAN. SSN builds on this basic LAN switching fabric by adding the WDM fiber links and an Optical Channel Interface (OCI) controller which extends the range from building to campus. The overall SSN network protocol suite for datagram and stream services (i.e., datagram transfer, flow control, routing, multicasting etc.) is described. In Section 4, representative performance results obtained via simulation are reported. The SSN testbed for the LAN and MAN setting is described in Section 5. Section 6 concludes the paper.

2 The Supercomputer Interconnection Environment

In this section, we present some representative applications enabled by a supercomputer or workstation network. The requirements of such applications are then identified and candidate networking approaches in the electronic and optical domain are reviewed.

2.1 New Enabling Applications

Just as conventional LAN technology enabled such present day applications as Network File System (NFS), the emerging low-latency, high-bandwidth ($<1\mu s$, $>60\text{Mbytes/s}$ [MB/s]) workstation cluster networks will likely enable new applications. These include:

- low cost arrayable video servers
- distributed memory parallel supercomputers (employing fine grain process management on 100's-1000's of processor elements, distributed checkpointing of jobs, and dynamic entry of new hosts)
- network based memory management
- real-time data acquisition and processing systems (e.g., radar, missile tracking, remote robot control)
- video display walls (with capability of perusing through huge database archives very quickly and interactively)

One common characteristic of these applications is that they "extend" the RAM memory of any one

workstation to any other in the cluster. Such network based distributed memory permits large applications to run beyond the confines of any one machine's memory space on a demand basis. This function has long been viewed as a minimum requirement to implement efficient message passing communications on MPP supercomputers, but only recently has been made possible over network interconnected workstations using a new breed of low-latency ($<1\mu s$) high-bandwidth ($>60\text{MB/s}$) 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 WDM channels make it possible to support guaranteed bandwidth services (e.g., voice, video) concurrently with conventional datagram services, with maximum isolation.

2.2 Network Requirements

The above mentioned applications require that the supercomputer network provide support for some key services. These services include

1. low-latency datagram service, to support fine grain distributed supercomputing. Variable size (as opposed to fixed size) datagram handling is required 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.
3. 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.

In the SSN project, protocols have been developed both in the high-speed LAN and in the optical backbone, in order to support the above basic services. SSN is designed to provide a very high-speed interconnection (up to one gigabit per second

The Supercomputer Supernet (SSN): A High-Speed Electro-Optic Campus and Metropolitan Network*

Nicholas Bambos, Joseph Bannister, Larry Bergman, Jason Cong,
Eli Gafni, Mario Gerla, Leonard Kleinrock, Steve Monacos,
Po-Chi Hu, B. Kannan, Bruce Kwan, Prasasth Palnati, John Peck and Simon Walton
Computer Science Department
University of California, Los Angeles CA 90095-1596.

Abstract

The Supercomputer Supernet (SSN) is a high-performance, scalable optical interconnection network for supercomputers and workstation clusters based on asynchronous, wormhole-routing switches. The WDM optical backbone extends the geographic coverage range from interdepartmental to campus and even to metropolitan areas with dynamically reconfigurable direct or multi-hop connections. The network provides very high-speed integrated services, supporting connection oriented, guaranteed bandwidth traffic as well as datagram traffic. The first networking level of the two-level SSN architecture is electronic and consists of crossbar meshes locally interconnecting workstations, supercomputers, peripheral devices and mass memory. At a higher networking level, an optical backbone network supporting multiple wavelength division multiplexed (WDM) channels allows communication between devices connected to distinct crossbar meshes. In this paper, we focus on the protocols of the WDM optical backbone network and address the issue of integration of the electronic wormhole-routing LAN with the optical backbone.

Keywords: Supercomputer interconnection network, Optical WDM network, Wormhole routing.

1 Introduction

The Supercomputer Supernet (SSN) currently being developed at UCLA, JPL and Aerospace under ARPA support is a novel, high-performance, scal-

able optical interconnection network for supercomputers and workstation clusters based on asynchronous wormhole routing crossbar switches. SSN has a two-level architecture in which high-speed electronic LANs are interconnected via a wavelength division multiplexed (WDM) optical network. The high-speed electronic LAN uses crossbars to interconnect workstations, supercomputers, peripheral devices and mass memory. The high-speed LAN employs wormhole routing and backpressure flow control to provide a low-latency, high-bandwidth connection over limited distance (maximum link length is 25m). The WDM fiber optic backbone network extends the geographic range from interdepartmental to campus and even to metropolitan areas via dynamically reconfigurable single-hop or multi-hop connections.

Distributed supercomputing and cluster computing impose critical demands on the network. These include low-latency, high-bandwidth connections, support for guaranteed bandwidth service and mechanisms to deal with unbalanced traffic patterns. The high-speed electronic LAN employs wormhole routing to provide low-latency datagram traffic support. The WDM fabric, on the other hand, can efficiently support guaranteed bandwidth services because of the enormous bandwidth available. Thus, the challenge is to effectively integrate the high intelligence and flexibility of electronic switching with the high throughput and parallelism of optics. The main thrust of this paper is to show how the two-level SSN architecture actually achieves this integration and provides support for distributed supercomputing and cluster computing.

*This work was supported by the USDOD ARPA/CSTO under Contract DABT63-93-C-0055. The Distributed Supercomputer Supernet – A Multi Service Optical Intelligent Network.