The Supercomputer Supernet: A Scalable Distributed Terabit Network

Leonard Kleinrock,
Mario Gerla, Nicholas Bambos,
Jason Cong, and Eli Gafni
University of California, Los Angeles, CA

Larry Bergman
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA

Joseph Bannister
Aerospace Corporation
El Segundo, CA

ABSTRACT — Conventional supercomputer interconnection networks consist of crossbar modules, which are connected by point–to–point copper or fiber links to create distributed mesh topologies (e.g., CP, Nectar). This type of ”physical networking” topology creates cable layout problems, dealing with bundles of cables/fibers between various pairs of modules. It also introduces several routing hops, increasing the probability of interference between connections and making it difficult to guarantee quality of service to real time applications. We describe a new network called the Supercomputer Supernet (SSN) that attempts to overcome these problems by replacing the point–to–point links with an fiber optic interconnect system. The novel scheme employs asynchronous pipeline crossbar switches (APCS) used in parallel supercomputers to interconnect multi-channel wavelength division multiplexed (WDM) fiber optic links 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, responding 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.*

1. INTRODUCTION

The Supercompuer Supernet (SSN) is a novel, high–performance, scalable *optical* interconnection network for supercomputers, which is based on asynchronous wormhole routing crossbar switches. The geographic coverage ranges 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 (WDM) channels, allowing communication between devices connected to distinct crossbars. These asynchronous crossbars, consisting of two–dimensional arrays of processors, primarily perform the communication switching; however, SSN may also be able to capture the large, latent, distributed computational power of the routers, to be used for network control and management, leading to an *intelligent* network.

The resulting distributed SSN, will be very fast—up to one gigabit per second (Gb/s) per channel speeds. It will scale up in the number of hosts connected and in geographical coverage. 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.

2. BACKGROUND

Supercomputer Networking

Supercomputer networking and high–speed optical communications are two very active areas of research. A more detailed discussion of research relevant to the distributed SSN and the corresponding prototype SSN can be found in Section 4. Here, we limit the discussion to a few key observations that position SSN with respect to other approaches.

Current research efforts (CFFD92, Sch91, Am89, Hoel90, Da91, Brac90, GKVBG90, DGLRT90, AKH87) can be classified basically into three categories, based on topology and provided service: point–to–point fiber, virtual point–to–point embedded in a fully broadcast physical topology, or multiple single–hop on–demand circuits. Among other things that distinguish the proposed research from other approaches, is that it combines multiple single–hop on–demand circuits with a multihop virtual embedded network.

Both point–to–point and virtual point–to–point nets are not suitable for high–volume, real–time, delay–sensitive traffic. High speeds require loose flow control, which on the other hand gives limited protection against congestion. Alleviating congestion by dropping or deflecting messages is not a suitable solution. Dropping messages is unwise since, at the high rates involved, losing even the content of a single buffer (64KB) is disastrous. Deflection, on the other hand, introduces unpredictable delay and out-of-order reception, again, something that, given the rates involved, is intolerable. Finally multihop networks do not naturally support broadcast and multicast.

Single hop networks cannot readily with current technology accommodate bursty short lived communication. Each two party communication requires the one party to be aware of the others request to communicate, together with the need to find a free virtual channel on which to communicate. This requires frequency agile lasers and detectors over a broad range of the optical spectrum and with nanosecond reaction times. Furthermore, it involves a substantial control and coordination overhead (e.g., rendezvous control and dedicated control channel). The SSN two-level architecture, which combines a single–hop subnet (for stream, circuit switched traffic) and a multihop subnet (for datagram traffic), effectively combines the benefits of the two types of networks, yet avoiding their shortcomings.

Prior Art in All-Optical Networks

Many varieties of optical networks have been investigated, proposed and prototyped (see [Brac90]). The class of optical networks which appears most suitable for high speed campus and metro interconnects is that of Passive Optic Networks (PONs), based on a broadcast medium (star, tree or bus) and exploiting WDM. In this class, the research has proceeded in two different directions, namely: single hop, and multiple hop optical networks. In the following, we briefly review these two approaches and discuss their limitations.

In single hop networks, all inputs are combined in a star coupler and broadcast to all outputs. To permit multiple, simultaneous transmissions, WDM is used, and is often combined with Time Division Multiplexing (TDM). The user must thus *select* the wavelength and time slot at each transmission. Several different possibilities exist, depending on whether transmitters, receivers, or both are tunable. It is also possible to have multiple fixed transmitters or receivers at each node instead of tunable ones. Numerous schemes falling into this category have recently been proposed. Some have been prototyped. A few representative examples are reported below.

LAMBDANET: The Bellcore's LAMBDANET system [GKVBG90] uses a combination of TDM and WDM. Each node has a fixed transmitter and an array of receivers. A grating demultiplexer is used to separate different optical channels. Each transmitter time–division multiplexes the traffic destined to all other nodes in a high–speed single wavelength data stream. Each receiving node simultaneously receives all the traffic, buffers it, and selects—using electronic circuits—the traffic destined for it. Two sets of experiments were performed, with 18 and 16 wavelengths, running at 1.5 Gb/s and 2 Gb/s, respectively.

Rainbow: Rainbow [DGLRT90] is a research prototype network designed at IBM. It is a circuit–switched metropolitan area network (MAN) backbone consisting of 32 IBM PS/2's as gateway stations, communicating with each other at 200–Mb/s data rates and submillisecond switching times. Rainbow has a passive broadcast star topology with fixed transmitters and tunable receivers. It uses both wavelength- and time-division multiplexing. A decentralized in-band signaling protocol was chosen for coordinating the tuning of the receiver filters in the network. Each transmitter, when it has a packet to transmit, repeatedly sends requests to transmit to a particular destination, until it receives an acknowledgment (ACK). Each receiving station, when idle, polls all transmitters (by tuning on their wavelengths) to see if there is one that requests transmission, and returns an ACK to it. This protocol is not suitable for packet–switched traffic, because of long request—response delays required prior to each packet transmission. To solve this problem a packet–switching protocol has been also proposed [CDR90]. It requires out–of–band signaling and introduces an additional fixed transmitter and receiver at each node. Other broadcast–and–select type systems and protocols are reported in [AGKV88], [CG87], [GK 91], [LGA90], [OS91], [CF91] and [GK91].

Instead of using a direct path from source to destination, multihop networks may require some packets to travel across several hops. In general, each hop incurs the penalty of an electro-optical conversion. Obviously, the virtual topology should provide routes with as few hops as possible.

Manhattan Street Network: One of the early proposals in the area of multihop lightwave networks was the Manhattan Street Network (MSN) [Max85], a multihop, mesh–connected network that uses unidirectional links between adjacent stations. Routing in the MSN is simplified by its regular structure, and, using the technique of deflection (or hot–potato) routing, it can operate with as few as one buffer per output port. The toroidal topology of the MSN ensures that a deflected packet will nominally take four extra hops to travel ‘‘around the block’’ if it needs to return to its point of deflection.

ShuffleNet: ShuffleNet, proposed in [AKH87], embeds a perfect–shuffle interconnection within a fully broadcast physical topology. This can be accomplished with stations having two sets of independently tuned (fixed) transceivers and at least twice as many WDM channels as the number of stations. The resulting multichannel, multihop network can achieve very high throughput with low delay. Bannister et al. [BG89, Bann90, BG90, BFG90a, BFG90b] and later Labourdette and Acampora [LA90a,LA91] studied the problems of designing virtual topologies for these networks, especially in the case of nonuniform traffic. By using various optimization techniques, the authors were able to achieve substantial improvements in performance. In [LA90b] the problem was further refined by considering the more realistic case in which transceivers have a limited tuning range.

Single and multiple hop networks both suffer some limitations. Starting with single hop networks, we note that these networks perform very well under most criteria. The major single limitation 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, that is, in the order of a few Gb/s [AKH87]. Capacity can be upgraded by using multiple wavelengths and implementing time and frequency division access schemes as shown in LAMBDANET, SWIFT and Rainbow. However, to achieve good efficiency in bursty traffic environments, these schemes require frequency agile lasers and detectors over broad ranges of the optical spectrum and with nanosecond reaction times. Such devices are not yet commercially available, although rapid progress of the technology in this direction has been reported [Brac90]. Still, a major challenge is the production of components with both high tuning speed, and broad wavelength range [Brac90]. Furthermore, the coordination of transceivers for short burst exchanges introduces considerable control overhead.

If we now consider multihop networks, we discover that such networks scale up rather well with network size by exploiting the parallelism of the mesh virtual topology. Furthermore, high aggregate throughputs are achieved with very simple station configurations (typically two fixed wavelength transmitters and receivers per node) which are more readily available and much less costly than their single–hop network counterparts [AKH87]. Furthermore, channel access control is straightforward. On the negative side, multihop networks perform rather poorly with respect to other criteria. They are prone to congestion, due to the lack of network flow control. Packet loss can be avoided by using deflection routing [Max85]. This, however, tends to cause large delay fluctuations and out-of-sequence packet deliveries, which cannot be tolerated by real-time traffic. Schemes have been proposed to support synchronous type connections [BFT91], and to enforce fairness [Max90]. These schemes, however, tend to increase network control overhead. Multihop networks cannot readily and efficiently implement broadcast and multicast, unless the simple routing structure implemented in the nodes is radically modified, at the cost of additional complexity. Finally, single node additions may require major topology reconfigurations, if the regular topology structure must be preserved. A particularly relevant multihop network example follows.

ATOMIC: One of the first networks to apply multiprocessor computer communications technology to local area network (LAN) switching is ATOMIC. Initially, the ATOMIC switching element was the mesh router module, which is made up of many 8-by-8 Caltech MOSAIC *mesh router* ICs [CIT90]. The mesh router is a board with 64 MOSAIC chips organized in an 8-by-8 matrix. Each MOSAIC chip contains a general purpose 11 MIPS processor, RAM, ROM and a DMA channel interface. The chip is equipped with eight half-duplex, bit-parallel, electronic channels, which can operate at the nominal rate of 480 Mb/s each. The 8-by-8 mesh thus has 32 full-duplex MOSAIC channels available at its edges. These channels will be used for host connections or for connections to other mesh routers. Functionally, the 8-by-8 mesh can be viewed as a crossbar switch. Packets are *source routed* from input to output port based on their X-Y coordinates. Connections to hosts are provided by the Host Interface (HI) board. This board has been developed by Caltech and the USC Information Sciences Institute [CFFD92]. It is based on four MOSAIC chips, plus memory and bus interface logic chips. It provides direct network access to workstations through its I/O bus. It has been demonstrated that the HI board can support data rates approaching 400Mb/s (for 1500-byte packets).

Enhanced ATOMIC: Recently, in mid-1994, the MOSAIC technology in ATOMIC was replaced by the Myrinet technology, developed by Myricom. Namely, MOSAIC mesh routers are replaced by Myrinet 4x4 or 8x8 crossbar switches. Functionally, the Mryinet transport mechanism is very similar to the MOSAIC mesh router mechanism, i.e., 640 Mb/s source routing, pipelining, backpressure flow control, etc. In addition, Myrinet supports a proprietary routing scheme which is much more flexible than the X-Y routing of MOSAIC and yet is deadlock free. An ATOMIC subnet could be a single Myrinet crossbar, or a mesh of crossbars interconnected with each other.

Scalable All-Optical Network Technologies

The main motivation for SSN stems from the limitations observed in current supercomputer networks, and from the opportunities offered by emerging communication technologies such as WDM optoelectronics for novel, feasible system architectures. Exploited will be current advances in wavelength division fiber optic networks to provide a large number of high bandwidth channels between switching nodes and fast asynchronous electronic-crossbar switches with inherent processing power and very low latency. Also, the distributed processing power of the network, can be exploited to support various network management operations and possibly other computational tasks. The goal is to develop a high–performance, intelligent supercomputer network, which *integrates* the bandwidth and connectivity advantages of *optical* interconnection with the *intelligence and low latency* of *electronic* processing for network control and management. We call the resulting system the *Distributed Supercomputer Supernet*.

One of the challenges in the design of a supercomputer interconnect is the layout of the topology which connects the various switching modules (e.g. hubs in Nectar-Net, or CP*s in the Los Alamos Multiple Crossbar Network). Ideally, the topology should be a fully interconnected mesh, with all modules directly connected to each other. This simplifies routing, and permits one to establish connections with low store–and–forward delay, minimal en route interference and guaranteed quality of service. Indeed, guaranteed quality of service (i.e. bandwidth) is particularly important for real time connections (e.g. visualization streams). On the other hand, the full mesh topology is very impractical from the standpoint of cable / fiber installation especially over large distances (campus and metropolitan areas). A linear or loop topology is much easier to install (and expand). In this case, however, the problem is the large number of hops that the packet must traverse. A satisfactory compromise is difficult to reach, especially in large nets. In the face of this difficult tradeoff, we see that current supercomputer networks are likely to be limited in scaling and in geographical growth.

Another challenge to supercomputer interconnect scaling is the control and management complexity, which rapidly increases with size. In SSN, the solution is to develop an intelligent interconnect network, which relies on the processing capacity of the switching nodes and network interfaces.

Recent advances in optical device technology (transmitters, receivers, amplifiers) also make it possible to multiplex several Gb/s channels on a single fiber using combined WDM (Wavelength Division Multiplexing) and TDM (Time Division Multiplexing) techniques. This capability can be exploited to overcome some of the aforementioned problems. Namely, we propose to use, as a starting point, the ATOMIC network developed by the USC Information Sciences Institute [CFFD92]. In ATOMIC, the switches are connected by point–to–point physical links. In SSN, each link of the point–to–point physical interconnection network is replaced with a *passive optical star* (or *tree*) network which we call OPTIMIC—for OPTical Interconnect of Myrinet ICs (the switching elements)

Thus, the switching modules (i.e., APCS) will be interconnected by an optical star. In particular, each module will interface to only one fiber. WDM channels will be maintained between selected pairs of modules, thus creating a *virtual* multihop topology. In brief, OPTIMIC interconnects several ATOMIC subnets via an optical WDM backplane. Packets will be transmitted on the multihop topology in the same way they were transmitted in the original ATOMIC network (except that the number of hops is typically smaller). In addition, a set of WDM channels will be set aside, and made available (on demand) for direct end–to–end connections between modules. Real time traffic will use such direct connections to avoid store–and–forward delays and to reduce en route blocking. One of the design goals of OPTIMIC is to extend the Myrinet transport protocols (e.g., pipelining, backpressure flow control, deadlock free routing, etc) transparently though the optical backplane.

Numerous interconnection configurations are possible between OPTIMIC switching modules and the passive optical star. One typical scenario is as follows. Of the 32 available ports, eight ports will be used for network interconnection. Of these, three will be connected (via WDM channels) to three ports on remote modules. Recall that these WDM channels are part of the ”virtual” topology. The remaining five ports will be used for circuit switched (C/S) connections. They will be equipped with tunable lasers and receivers. At connection set up time, lasers will be tuned to the desired wavelength. Initially, two laser/receiver pairs will be assigned to each module. The remaining 24 ports of the 8 by 8 module are available for host interconnections. Up to four hosts can be connected to the same port (using a daisy chain arrangement) at the cost of a reduction in throughput.

Alternate Optics Technologies

It should also be noted that several alternate fiber optic technologies may be employed in the network to either enlarge the channel pool between switching nodes or to extend the SSN to metropolitan networks. These include using tunable Fabry–Perot narrow–band filters at each optical receiver or monolithic stepped wavelength laser arrays at each transmitter combined with fast electronic channel switches. Using laser arrays enhances channel concurrency (i.e., transmitting on multiple wavelengths simultaneously), and also relaxes the demands on the electronic multiplexer rates by transmitting data in parallel byte or word frames across the fiber. It also eliminates the overhead normally required for framing and synchronizing a serial signal, more closely matches typical computer system bus interfaces, and enhances scalability of system capacity beyond the switching speed limit of single laser diode and transistor devices. Present laser array dimensions are about 4–8 for 1550nm devices, but are expected to increase to 64–128 in 3–5 years.

The channel selection speed of tunable laser diodes varies from 15ns for three-section DBR lasers operating at 1531nm (2.2nm continuous tuning range and 7.3nm quasicontinuous tuning range) to millisecond rates for thermally/mechanically tuned devices. The Fabry–Perot narrow band filter can also switch in millisecond speeds over a broad wavelength. Stepped wavelength laser arrays can be switched as fast as the electronic switches that front–end them—typically a few nanoseconds to 100 picoseconds. Current off–the–shelf technology supports data rates to 2.4Gb/s.

3. SSN

Architecture

Architecturally, the optical fabric of SSN—which we call OPTIMIC—has been directly conceived to support both circuit–switched and multi–hop traffic, achieve virtual topology reconfigurable interconnection through an optical star (or tree), and base its networking operations on the distributed processing capacity of the network itself. Although it can be based on already existing technologies, it is also well positioned to absorb the new exciting technologies that are currently emerging in optoelectronics and high–speed intelligent networking, showing the future directions in high performance computing and communications.

The conceptual SSN architecture is shown in Fig. 1. Each network node consists of an APCS constructed from the Myrinet pipeline crossbar ICs (that are asynchronous) and multiple optical channel interfaces. More generally, the APCS is a mesh interconnection of Myrinet crossbar switches which support several hosts. The APCS establishes fast connections from one of several local hosts to one of the available optical channels. Typically, datagram connections remain permanent (solid lines) while stream based services (such as video) are made on demand. Since the number of available optical channels (>24) greatly exceeds the number of ports on a given crossbar node (≤8), many different virtual topology configurations are possible. Also, the probability of encountering a blocked state among the circuit switched channels is greatly reduced as well.

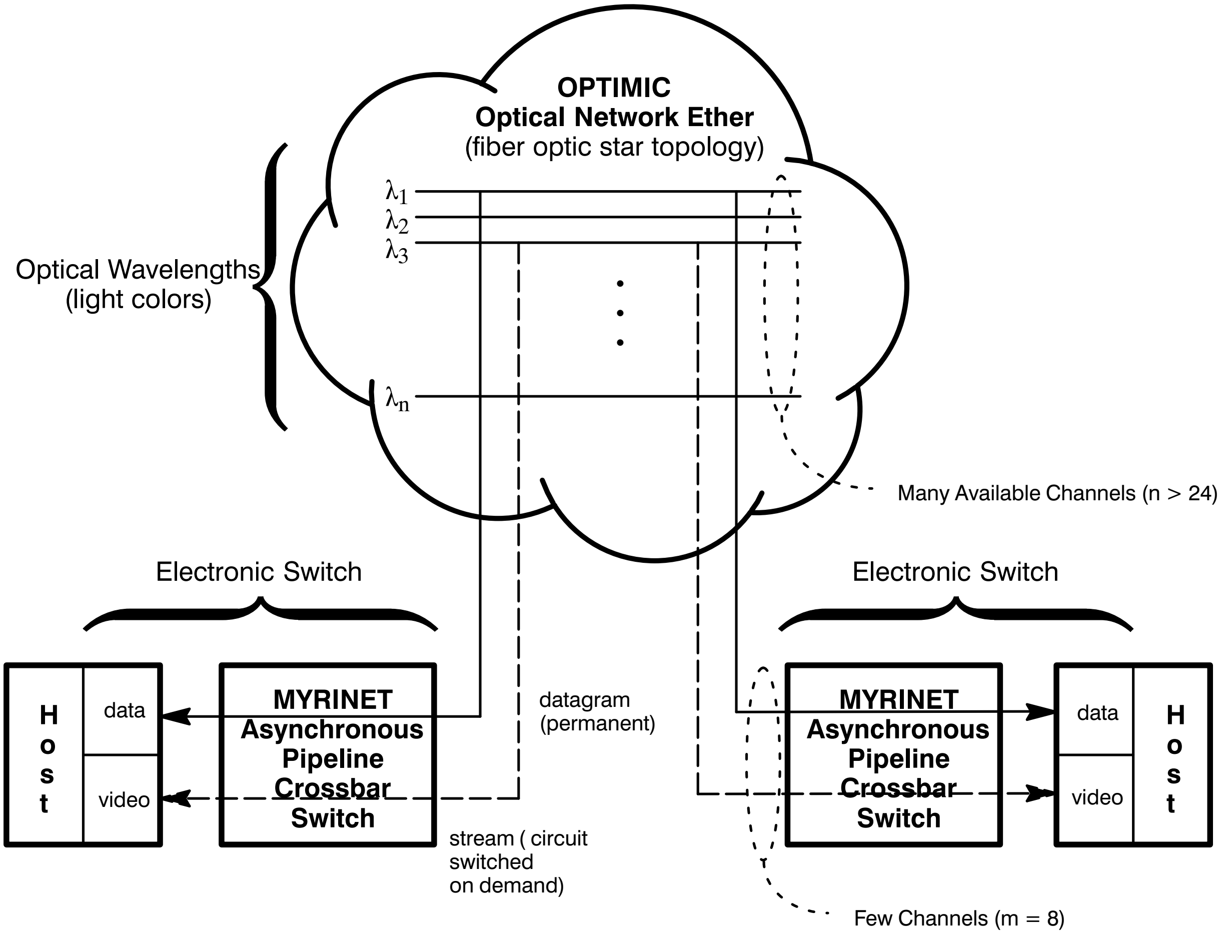
OPTIMIC
Optical Network Ether
(fiber optic star topology)
Optical Wavelengths
(light colors)
λ1
λ2
λ3
λn
Many Available Channels (n > 24)
Electronic Switch
Electronic Switch
H o s t
data
video
MYRINET Asynchronous Pipeline Crossbar Switch
datagram
(permanent)
stream (circuit switched on demand)
MYRINET Asynchronous Pipeline Crossbar Switch
data
video
H o s t
Few Channels (m = 8)

Fig 1. The basic elements of SSN include Myrinet based asynchronous pipelined crossbar switches that implement very low latency circuit switched and datagram connections from a multi-media host to a multi-channel optical network (OPTIMIC), typically based on a star topology. Generally, the number of optical channels is made much larger than the number of crossbar ports to enhance non-blocking of the circuit switched channels and enrich the virtual topology of the datagram channels. While many fiber optic technologies can be used (including space, time, and frequency division multiplexing), tunable laser diodes and Fabry Perot receivers maximize performance and provide the most cost-effective utilization of the cable plant.

An initial testbed implementation of OPTIMIC (in Fig. 2) shows four Myrinet crossbar (APCS) nodes and five isolated STAR–based fiber optic networks. Since tunable laser technology is still quite experimental, fiber optic ribbon cable (with one fiber representing one wavelength) could also be used initially to implement an optical space division multiplexing network with equivalent functionality.

Eventually, it is planned that the OPTIMIC testbed configuration will consist of as many as 8 APCS elements. The *virtual* topology will initially be a perfect shuffle, with maximum path length of 3 hops. Assuming three dedicated wavelengths (i.e., three ports) per module, the number of wavelengths required is $3 \times 8 = 24$. This number, however, can be reduced to 8, by time division multiplexing three 800 Mb/s subchannels on a single WDM channel operating at 2.4 Gb/s. The 800 Mb/s channel rate is adequate for our purposes since it exceeds the 640 Mb/s Myrinet data rate. In addition, a pool of 24 channels, at 800 Mb/s each (i.e., 8 wavelengths) will be set aside for circuit-switched connections. The total number of required wavelengths is 16, each wavelength supporting 3 TDM

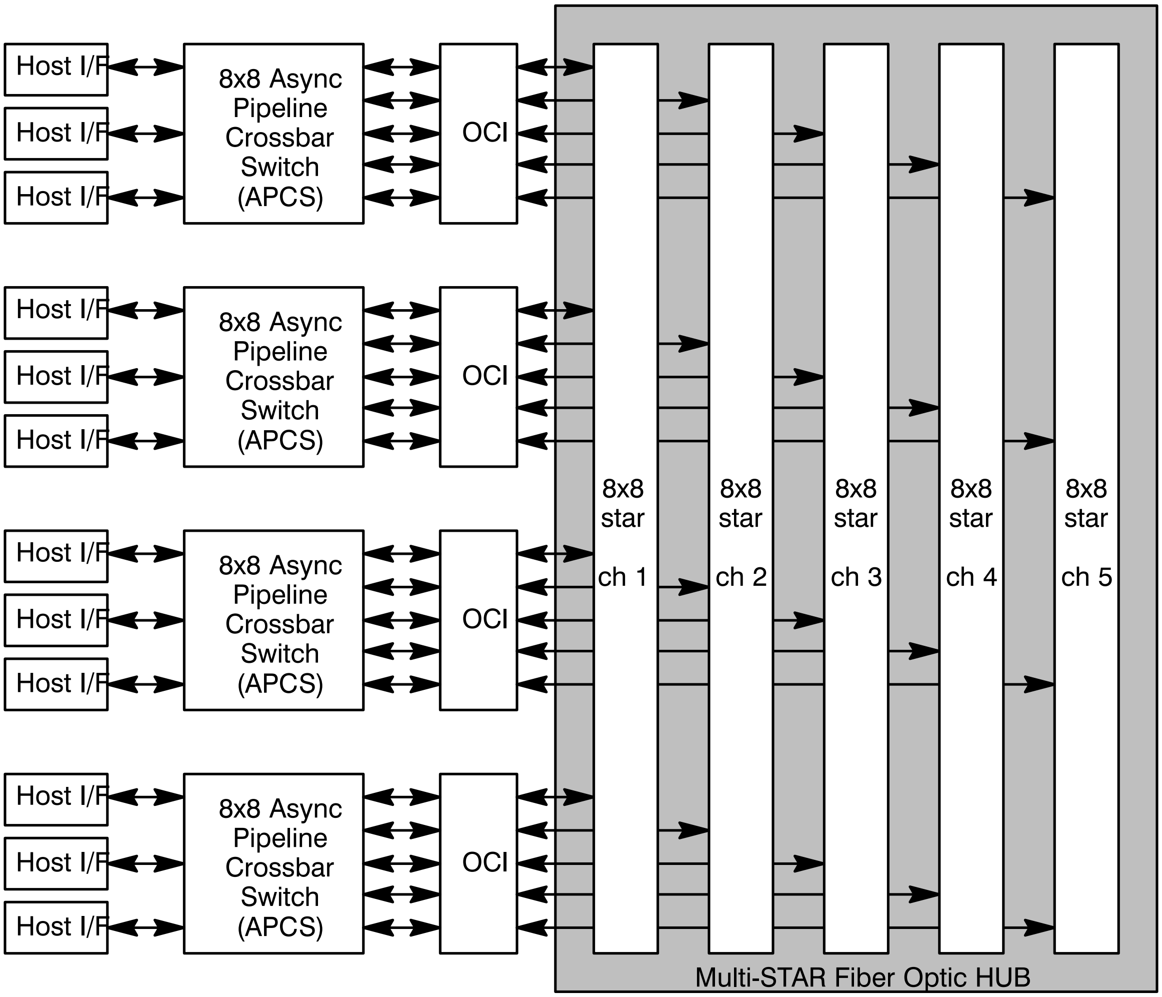

Fig 2. Phase 1 OPTIMIC hardware architecture consists of 4 network pipeline crossbar nodes and five optical channels. Each optical channel of an 8x8 optical star (3 ports are unused) plus associated fiber optic full-duplex fiber optic transceivers. The media is fiber optic ribbon cable (multimode for campus and single-mode for MAN).

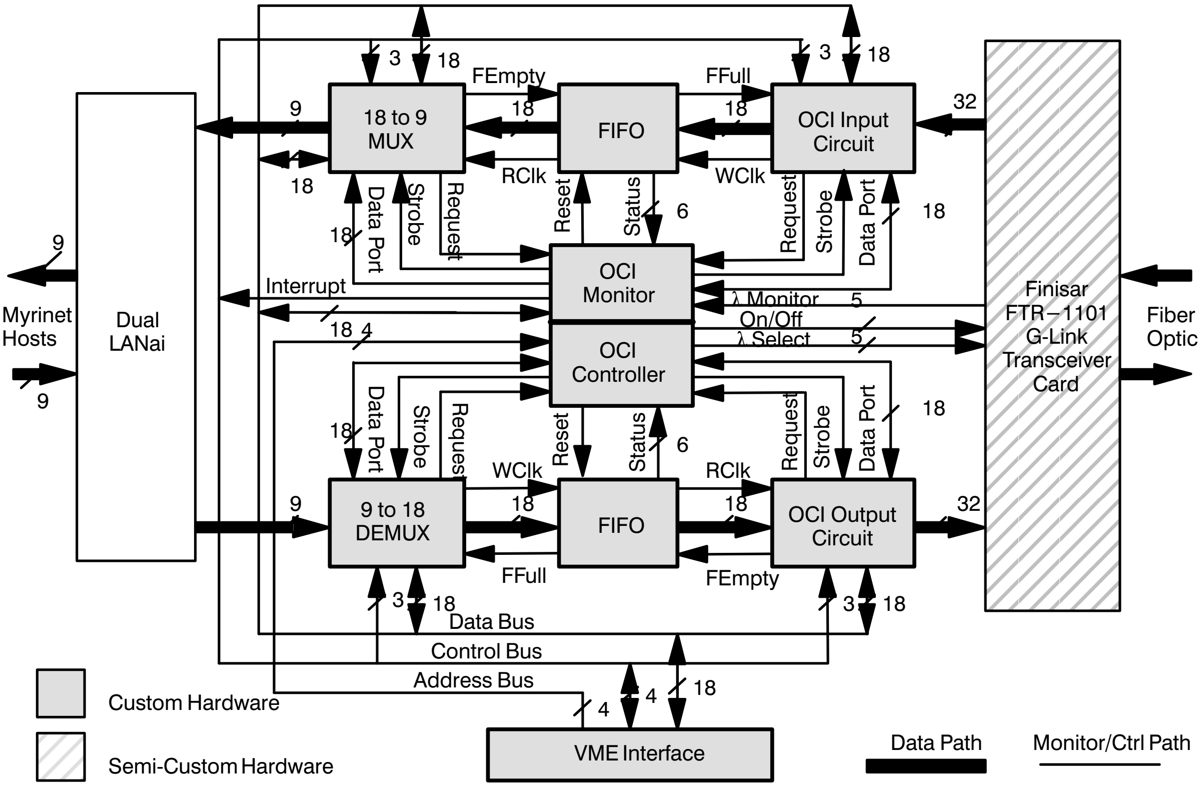

Fig 3. OPTMIC Optical Channel Interface (OCI) board. This board buffers, multiplexes, and switches asynchronous Myrinet crossbar ports to a given optical fiber channel. Control, network management, and setup of the OCI is via a VME interface to a local SPARC CPU card.

channels at 800 Mb/s each. Hosts are connected to Myrinet crossbar switches with host interfaces (HI). In principle, there is no limit on the number of hosts, since the APCS can be replaced by an arbitrary Myrinet mesh. In our target configuration, up to 50 hosts must be supported, thus requiring 50 HIs.

Non–real–time traffic (file transfers, interactive communications, etc.) will travel on the virtual multihop network (at most 3 hops). Real–time traffic will use circuit switched connections. Signaling and control traffic (e.g., call set up messages) will travel on the virtual, multihop network.

Optical Channel Interface (OCI).

The Optical Channel Interface board (Fig. 3), or OCI, is responsible for buffering and switching between APCS ports and the fiber optic links. Although only five (5) fiber optic links are planned to be built in the early testbed, considerably more fibers could be added later.

Asynchronous Clock Recovery

There are several types of clock recovery, framing, and multiplexer/demultiplexer ICs commercially available today that operate at Gb/s rates. Some are also integrated with fiber optic transceivers to minimize layout problems. This will be the method used to synchronize streams between adjacent OPTIMIC nodes.

Fiber Optic Links

The circuit switched (C/S) mode of OPTIMIC requires the availability of many dedicated optical channels (where many is defined as a number larger than the number of APCS ports). This enhances the scalability and reconfigurability of the network and reduces the possibility of blocked paths.

In all, there are four potential technologies that may be employed either singly or in combination: (1) spatial multiplexing (via fiber ribbon cable), (2) spectral multiplexing via dense wavelength division multiplexing (WDM) optical components (either tunable lasers, Fabry Perot receivers, or stepped wavelength laser arrays), (3) optical frequency division multiplexing (FDM) via sub-carrier multiplexing, and finally, (4) electronic time division multiplexing (TDM). The lowest risk technology is the fiber optic ribbon cable driven by mono-wavelength laser diodes (spatial multiplexing) as shown in Fig. 4 below. It is also the least expensive for a small number of channels (<16). Its disadvantages are that multiple fiber media plants are required, limiting scalability. An advantage is that it can always be augmented with WDM at a later date.

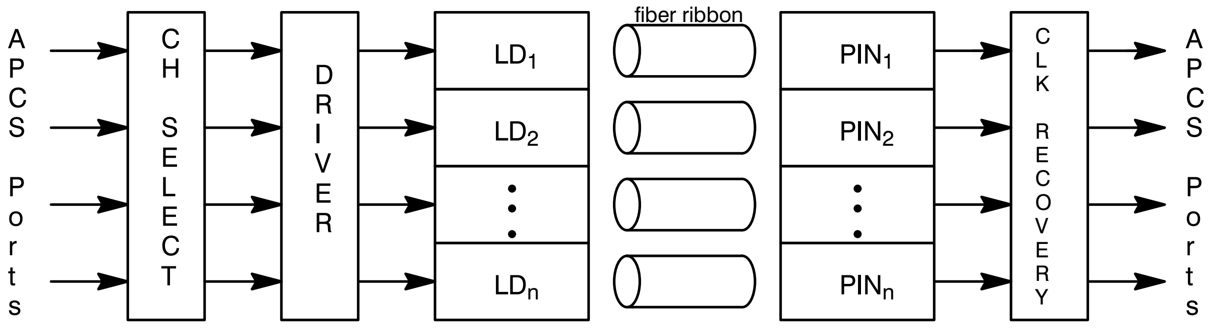

Fig. 4. Fiber optic ribbon cable provides multiplex channels via spatial multiplexing.

The next most viable technique is to assign fixed wavelengths to transmitter/receiver pairs using stepped-wavelength laser diode arrays, one single-mode fiber, and grating front-end loaded fiber optic receivers. This technique lends itself to fast switching (<10ns) and can be made very stable. An integrated optics implementation would probably be required for large quantities (Fig. 5). Disad-

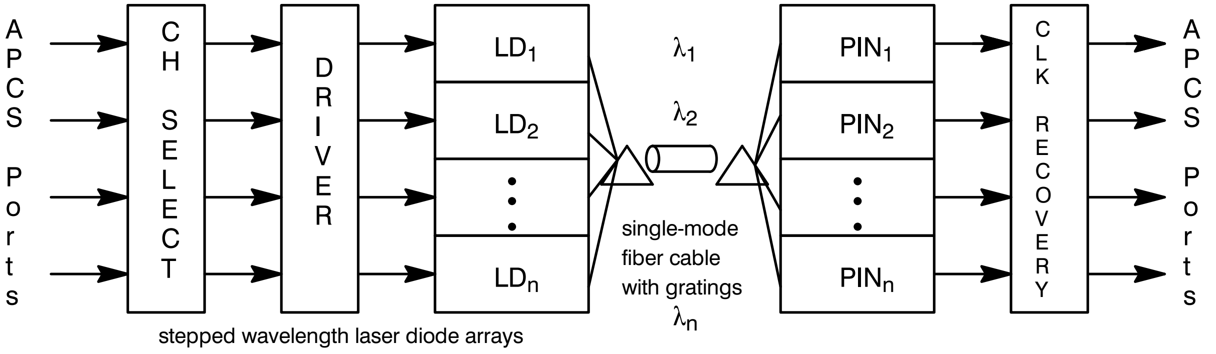

Fig. 5. Stepped wavelength laser diode (LD) arrays and PIN-diode receivers provide tens of Gb/s channels per single optical fiber, but scales poorly and offers poor LD array utilization when many 2-user communications must take place across the star ether.

vantages of the approach are that the laser diode array utilization decreases rapidly as the channel spectrum capacity grows larger compared to the APCS port dimension.

The most effective utilization of optical spectrum space with the least optoelectronic array complexity occurs when either the optical transmitters or receivers can be tuned to a given channel slot. Hence, the number of optoelectronic devices (n) exactly equal the APCS port dimension. In Fig. 6, the sources are individually tuned while in Fig. 7 the receivers are individually tuned. The tuning stability of current laser arrays (Fig. 6) are slow enough (10's of μs to ms) and coarse enough that probably only a few devices could be attempted in the next 3 years (<8).

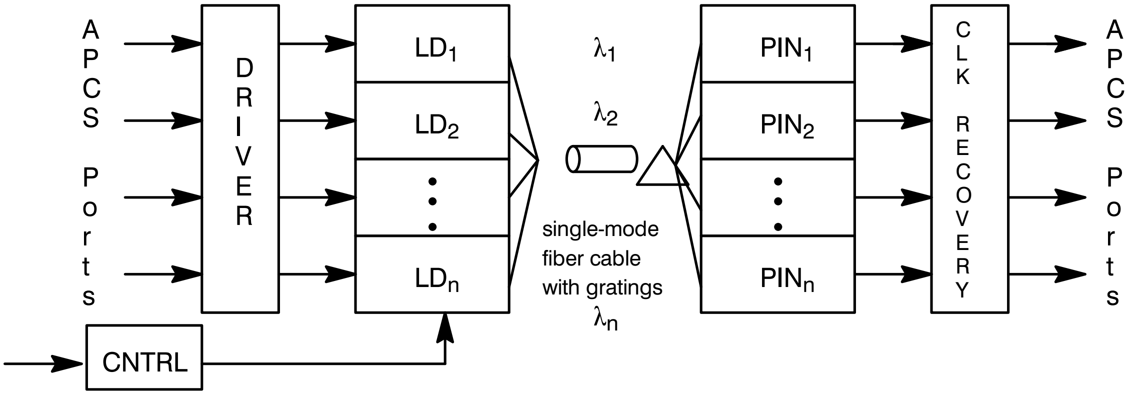

Fig. 6. Tunable laser diode arrays permit a few number of physical LD devices to cover the entire optical channel spectrum.

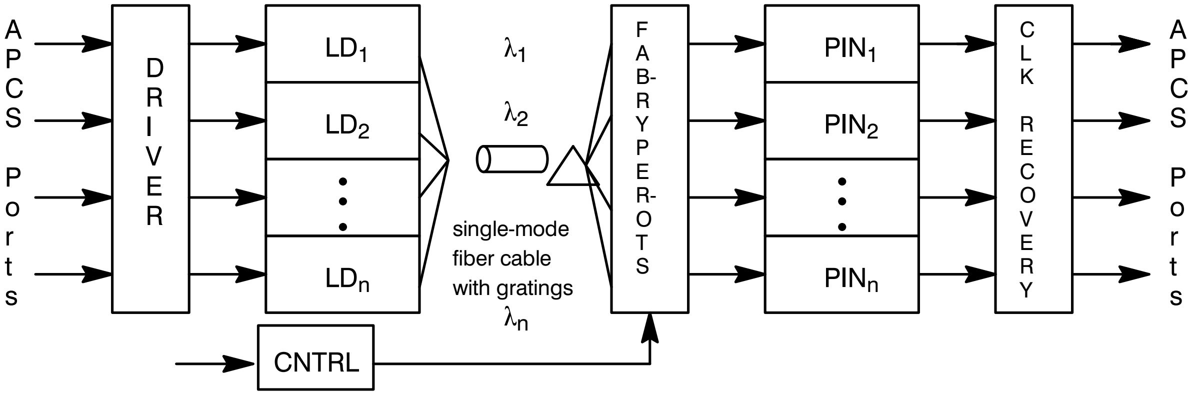

Fig. 7. Tunable acousto-optic Fabry Perot receivers provide similar capability to tunable LD arrays except that the control and management of the spectrum space is somewhat simpler, especially for multicast transmissions.

The most effective technology in terms of maximizing system performance would be to utilize tunable laser diodes combined with tunable Fabry Perot receivers. This would produce the richest network virtual topology, maximize aggregate capacity, and minimize probability of blocked states.

OPTIMIC Software

As in any communication network software is needed to support the transfer of messages. OPTIMIC's exclusive use of source routing and its flexible virtual topology suggest a prominent role for software that implements communication protocols and network–management functions. The principal software modules in the OPTIMIC network are (1) Address Consultant, (2) Topology Manager, (3) OCI Topology Manager Proxy Agents, (4) Communication Protocols, and (5) Test Tools, which are all described below.

The Address Consultant function was introduced in [ATOMIC] as a means of binding destination address to routes when switches support only source routing, as in OPTIMIC. We have therefore adopted the Address Consultant function for OPTIMIC. The Address Consultant thus plays the role of the Address Resolution Protocol, which maps Internet addresses to media access control addresses in broadcast LANs.

Copies of the Address Consultant reside in hosts attached to the network. To promote scalability while keeping overhead low, there typically is one Address Consultant in each network cluster (i.e., the electronic Myinet subnetworks that are connected to the optical subnetwork). The Address Consultant is responsible for providing each host of its cluster with a route specification from the originating host to the destination host. By means of probe messages each Address Consultant discovers the best route between its cluster's hosts and other clusters. Conferring with the Address Consultants in foreign clusters, the Address Consultant can then find the remainder of the route to the destination host. Clearly, the Address Consultant is responsible for knowing the topology of its native cluster as well as the optical subnet. A subtle point is that all Address Consultants should be aware of the virtual topology of the optical subnet——reliance on a single Address Consultant's knowledge by using source–specified route as the return route will not suffice, since the virtual topology of the optical subnet need not be based on bidirectional links. It is also the case that each host must know the path to its Address Consultant in order to make requests of it. The Address Consultant insures this by explicitly informing each host how to reach it.

The tunable transceivers of the OCI provide the capability of defining different virtual topologies for the network. The virtual topology of the network is defined and controlled by the Topology Manager, of which there is only one active copy in the network at a given time. A backup Topology Manager may be provided to increase network dependability. The Topology Manager resides on a network–management host, which may also perform additional duties, such as address consultation and other network–management functions. Cooperating with Address Consultants, the Topology Manager knows the current topology of the entire network. Conversely, the Topology Manager can redefine the network's virtual topology at any time by sending commands to OCIs. Moreover, the Topology Manager keeps track of special attributes of the topology, such as which resources are dedicated to packet– and circuit–switched traffic. These attributes can also be communicated to the Address Consultant, which uses this information to inform a requesting node of the best path to use for a given class of traffic.

An important function of the Topology Manager is to determine the best virtual topology for the prevailing network conditions. In [BannisterFrattaGerla90] it was shown that simple optimization algorithms can result in significant performance gains when applied to the problem of virtual–topology design. Such algorithms will be incorporated into the Topology Manager. The integration of packet– and circuit–switched traffic introduces new issues into the virtual–topology design problem, and these will be addressed as part of the OPTIMIC project.

The Topology Manager controls and monitors the state of the OCIs by means of a special protocol that allows it to communicate commands to proxy agents that reside in the OCIs. The OCI incorporates a simple host based on the Myricom LANai chip, a 11–MIPS microprocessor and Myrinet adaptor logic. An OCI Topology Manager Agent has read and write access to registers used to tune the OCI's optical transceivers. The OCI Topology Manager Agent executes on the LANai processors and programmed I/O operations are effected through memory–mapped registers. Acting on behalf of the Topology Manager, the OCI Agents configure the OCI to realize a specific topology. Although not strictly a topology–management function, an auxiliary role of the Topology Manager Agent is to assist in maintaining configuration parameters that control the field–programmable gate arrays on the OCI.

The Myrinet product supports TCP/IP–based protocols. However, we envision the need for a ‘‘raw’’ packet interface that provides direct access to the OPTMIC source–routing layer. It is also expected that a simple network–management protocol (essentially based on SNMP) will be employed by the Topology Manager and its OCI Proxy Agents.

To support the measurement of network traffic, software is required to capture and characterize packets unobtrusively. Measurements can be consolidated into a global traffic matrix and used by the Topology Manager to find an optimal virtual topology, and they can be used in our performance studies of OPTIMIC. Other test software, such as artificial–traffic generators, will also be provided in the network.

4. RESEARCH DIRECTIONS

The efficient operation of the optical interconnect will require the development of several algorithms, tools and specialized protocols including:

(a) routing and congestion control procedures for the multihop network.

(b) dynamic reconfiguration tools for the virtual topology (using tunable lasers/receivers).

(c) efficient optical channel/fiber scaling techniques to handle large user populations.

(d) low latency communication over OPTIMIC.

Wormhole Routing

Various research issues arise in the area of routing and congestion control. For example, given the nature of message routing (wormhole routing) in a Myrinet network, a basic issue is the following: Suppose that there is a short path in the network on which a given message can be routed from its source to the destination. However, when the source node requests a network manager for a path to send its message to the destination the above mentioned short path has a section of it busy, because some other message (worm) is partially occupying it. The routing algorithm, after searching for an alternative path, identifies the shortest free one. Unfortunately, the latter path is much longer than the very short one which is busy. The network manager has then the following dilemma: should it let the message go on the very long path or block it until the short one becomes free again. This decision making problem can be resolved based on the statistics of the message lengths. Basically, the routing algorithm, if it anticipates (based on past statistical measurements of the traffic) that the short path will become available soon (depending on how long the longer path is) it will let the message use it (and be wormhole blocked on it until it clears); otherwise, it will force it to use the longer path. The mathematical analysis of the problem is quite involved, but we have managed to resolve the issue in an adequately general case.

We have also been studying other problems related to congestion control as well as dynamic reconfiguration of the virtual topology of the network using tunable lasers/receivers etc. We plan to build software tools incorporating the solutions of the issues we have been studying in these areas. Moreover, we will do extensive experimentation on the testbed itself.

Wavelength Channel Reduction

The testbed will allow experimentation with recently proposed WDM/TDM techniques, which could further reduce the number of required wavelengths. In fact, in a separate, ongoing research project, the use of WDM in a multiaccess environment has been investigated [LK92b]. These algorithms will allow very efficient use of the fiber bandwidth using new access schemes. Also, the performance has been evaluated under various assumptions regarding the number of fixed and tunable receivers and transmitters, and have done so for arbitrary traffic profiles. With the use of a star-structured glass switch connection, the OPTIMIC testbed will permit verification of the theory by demonstrating these algorithms and access schemes in real operational environments.

Architecture Scaling Properties

An important research direction in OPTIMIC is the scaling to large user populations. It is conceivable that in a metropolitan OPTIMIC configuration, several thousands of users may be connected to the network. The basic OPTIMIC design must therefore be extended to handle large numbers of user ports (and correspondingly, large numbers of optical ports on the optical backbone). The major limitations to scaling in OPTIMIC are the small number of wavelengths available in a fiber (up to 50, say, using direct detection techniques), the slow tuning time of optical transceivers (relative to packet transmission time) and the optical power loss through the various couplers stages in a multilevel tree, or through the stages of a modular star coupler. To overcome these scaling problems in OPTIMIC, we have developed a three–pronged strategy, exploiting (a) T/WDMA techniques to make the single hop, C/S access more efficient; (b) channel sharing to reduce the number of wavelengths required by the multihop scheme, and (b) multifiber cables to reduce the number of wavelengths required, and to improve power budget. These techniques are briefly described below.

We have developed a T/WDMA access scheme which allows the efficient sharing of the optical channels (i.e., wavelengths) by users with different data rates. With T/WDMA, users are time division multiplexed on each channel. The requirement for a fast retunable receiver (i.e., nanosecond tuning time) is relaxed by the use of a strategy called *subframe tuning and pipelining*. The details are reported in the [Kov93].

We have studied the effect of channel sharing in multihop networks. Channel sharing helps reduce the number of wavelengths required by the multihop virtual topology. This reduction, in turn, permits an increase in the number of stations connected by the multihop network. Since OPTIMIC uses a multihop subnet for datagram traffic, it will clearly benefit from channel sharing. The results of this study are reported in [Ger93].

We have evaluated the use of multifiber passive optical networks with the purpose of ” trading wavelengths for fibers.” Namely, we have shown that by using multifiber ribbon cables we can drastically reduce the number of wavelength per fiber required to support a give user population. In fact, we have shown that with a fiber cable of 250 fibers, and 3 wavelengths per fiber, we can interconnect up to 2000 stations without the need for optical amplification or fast receiver tuning [Bann93].

Using a combination of the above scaling techniques, it is possible to achieve aggregate throughputs well in excess of one terabit per second. For example, using a fiber plant with 128 fibers per cable and 20 wavelengths per fiber, we have at our disposal a pool of over 2500 optical channels. With TDM, we can transmit up to 3 streams (at 800 Mb/s each) on each wavelength. Thus, the total aggregate bandwidth is about 6 terabits per second (Tb/s).

Low latency is one of the most important requirements in distributed supercomputing. In many applications, the efficiency depends very critically on the instantaneous transfer of datagrams across the network. In contrast to other interconnection schemes (e.g., ATM, HIPPI, etc) which require a connection setup before delivering the data, SSN allows immediate transfer without prior setup (datagram mode). The use of cut-through (i.e., wormhole) switching eliminates the need of buffering the datagram at intermediate nodes, thus minimizing the delay impact of multiple hops. Delays, of course, may build up if either the Myrinet or the OPTIMIC backbone become congested. To overcome this problem, we are now investigating congestion prevention schemes which combine dynamic routing, transit packet priorities, and virtual topology tuning.

5. ENABLED APPLICATIONS

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

Examples

Fine Grain Meta-Supercomputer: The OPTIMIC attributes would accelerate the evolution of a network-based operating system (OS) with precise synchronization of dispersed processes, fine grain process management on 100's–1000's of processor elements (PEs), distributed checkpointing of jobs, and dynamic entry of new hosts.

Real Time Distributed Network Operating System: Low and predictable (bounded) latency makes OPTIMIC ideal for wide area network control and data acquisition applications. Examples in the government include Air Force satellite communication (SATCOM) network, Ballistic Missile Defense Organization (BMDO) missile tracking and wargaming, remote robot control for NASA applications, 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.

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).

Target Demonstration Application

The basic OPTIMIC testbed topology is shown in Fig. 8. APCS switching nodes 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 OPTIMIC testbed demonstration site using scalable I/O supercomputers (see Fig. 9). The proposed OPTIMIC 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 channel high performance parallel interface (HIPPI) 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). In late FY'94, the Caltech Intel DELTA will be upgraded into a Paragon and the JPL Cray YMP to a T3D,

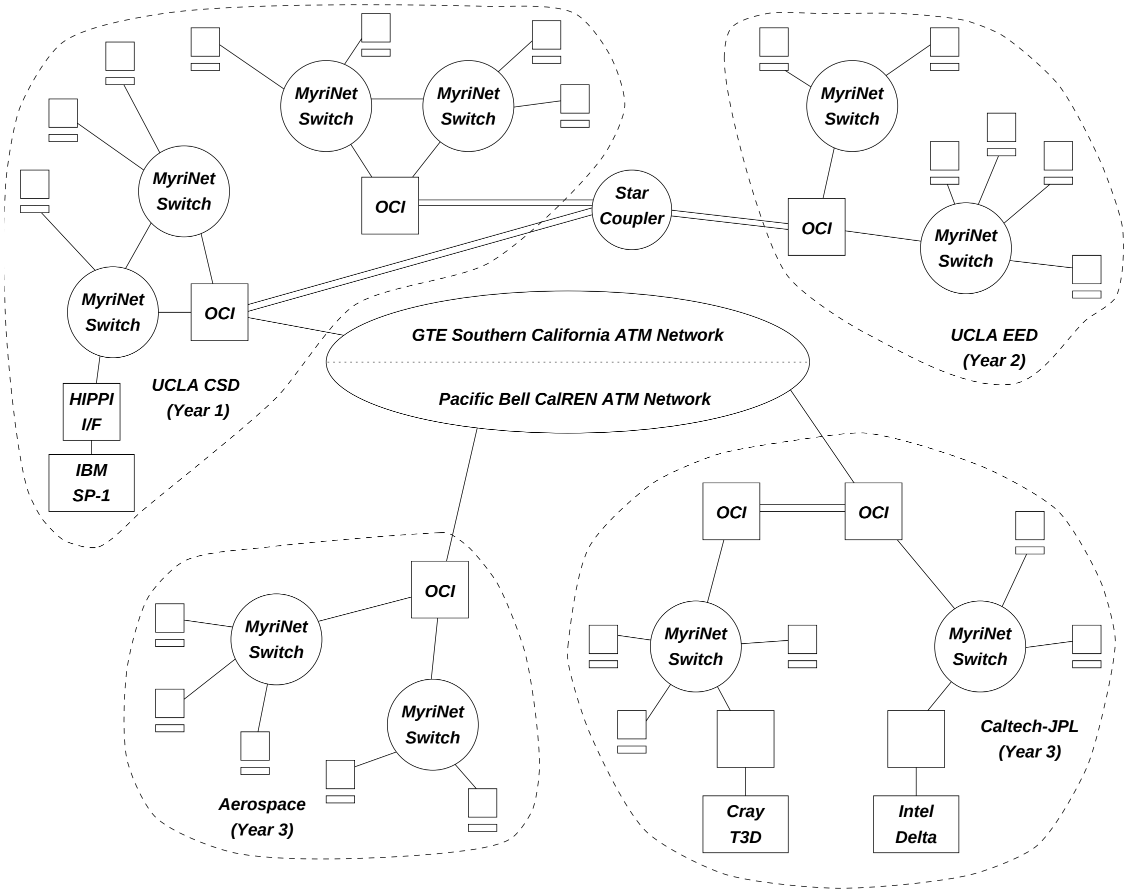

Fig. 8 – The SSN testbed consists of four clusters of APCS switches: four at the UCLA computer science department, two at the electrical engineering department, two in Pasadena (one at JPL and one at Caltech), and two at Aerospace Corporation.

both with multiple HIPPI ports. Running over the existing dark fiber, OPTIMIC 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 (CBI). 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 OPTIMIC 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.

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 (MPP) 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 scal-

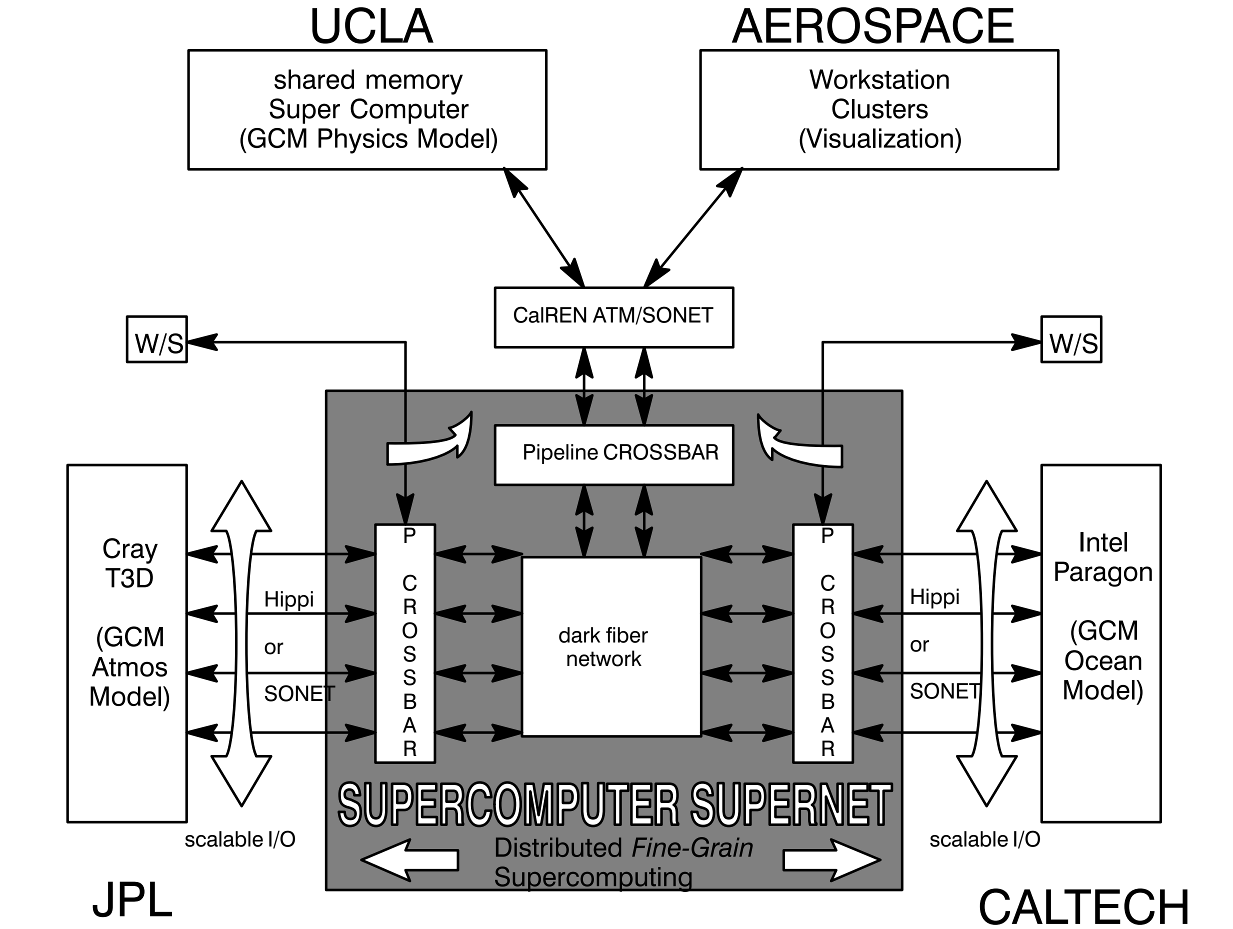
UCLA
AEROSPACE
shared memory Super Computer (GCM Physics Model)
Workstation Clusters (Visualization)
CalREN ATM/SONET
W/S
W/S
Pipeline CROSSBAR
Cray T3D (GCM Atmos Model)
Hippi or SONET
P CROSSBAR
dark fiber network
P CROSSBAR
Hippi or SONET
Intel Paragon (GCM Ocean Model)
SUPERCOMPUTER SUPERNET
scalable I/O
Distributed *Fine-Grain* Supercomputing
scalable I/O
JPL
CALTECH

Fig. 9 – OPTIMIC supports a fine-grain distributed supercomputing and visualization GCM application using MOSAIC on one segment of the CASA testbed between JPL and Caltech.

able 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.

7. 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.

8. REFERENCES

[AKH87] Acampora, A.S., Karol, M. J. and M.G.Hluchyj, ”Terabit Lightwave Networks: The Multihop Approach,” *AT&T Technical Journal*, Vol. 66, No.6, pp. 21–34, November/December 1987.

[AGKV88] Arthurs, E., Goodman, M.S. Kobrinski and M. P. Vecchi, ”HYPASS: An Optoelectronic Hybrid Packet–Switching System,” *IEEE Journal on Selected Areas in Communications*, Vol. 6, pp.1500–1510, 1988.

[Am89] Amould, E., et al, ”The Design of Nectar: A Network Backplane for Heterogeneous Multicomputers,” Carnegie Mellon Univ. Tech. Report CMU–CS–89–101, 1989.

[BG89] Bannister, J.A., and M. Gerla, ”Design Of The Wavelength–Division Optical Network,” Technical Report, CSD–890022, UCLA Computer Science Department, Los Angeles, CA, May 1989.

[Bann90] Bannister, J.A., ”The Wavelength–Division Optical Network: Architectures, Topologies, and Protocols,” Technical Report CSD–900007, UCLA Computer Science Department, Los Angeles, CA, March 1990.

[Bann93] J.Bannister et al, ” All optical multifiber tree network,” *IEEE/OSA Journal of Lightwave Technology*, Vol. 11, no. 5/6, pp. 985–1005, May/June 1993.

[BG90] Bannister, J.A. and M. Gerla, ”Design Of The Wavelength–Division Optical Network,” *Proceedings of IEEE ICC ’90*, pp.962–967, Atlanta, GA, April 1990.

[BFG90a] Bannister, J.A., Fratta, L., and M. Gerla, ”Topological Design Of The Wavelength Division Optical Network,” *Proceedings of IEEE INFOCOM ’90*, Vol. 3, pp.1005–1013, San Francisco, CA, June 1990.

[BFG90b] Bannister, J.A., Fratta, L. and M. Gerla, ”Optimal Topologies for the Wavelength Division Optical Network,” *Proceedings of EFOC/LAN ’90*, pp.53–57, Munich, Federal Republic of Germany, June 1990.

[BFT91] Borgonovo, F., Fratta, L. and F. Tonelli, ”Circuit Service In Deflection Networks,” *Proceedings of IEEE INFOCOM ’91*, Bal Harbour, FL, pp.69–76, April 1991.

[Brac90] Bracket, C.A., ”Dense Wavelength Division Multiplexing Networks: Principles And Applications,” *IEEE Journal on Selected Areas in Communications*, Vol. 8, pp.948–964, August 1990.

[CDR90] Chen, M., Dono, N.R., and R. Ramaswami, ”A Media-Access Protocol For Packet–switched Wavelength Division Multiaccess Metropolitan Area Networks,” *IEEE Journal on Selected Areas in Communications*, Vol. 8, pp.1048–1057, August 1990.

[CF91] Chlamtac, I. and A. Fumagalli, ”QUADRO–Stars: High Performance Optical WDM Star Networks,” *Proceedings, IEEE GLOBECOM ’91*, Phoenix, AZ, pp.1224–1229, December 1991.

[CFFD92] Cohen, D., G. Finn, R. Felderman, and A. DeSchon, ”ATOMIC: A Very–High–Speed Local Area Network,” *IEEE Workshop on High–Speed Communication Subsystems*, Tuscon AZ., Feb. 1992.

[CG87] Chlamtac, I. and A. Ganz, ”Toward Alternative High Speed Networks: The SWIFT Architecture,” *Proc. of IEEE INFOCOM’87*, San Francisco, CA., pp. 1102–1108, March 1987.

[CIT90] California Institute of Technology, ”Submicron Systems Architecture Project Semiannual Technical Report”, Computer Science Tech. Report Caltech–CS–TR–90–05, 1990.

[Da91] Davie, B., ”A Host–Network Interface Architecture for ATM,” Proc. of SIGCOMM’91, pp. 307–315.

[DGLRT90] Dono, N.R., Green, P.E., Liu, K., Ramaswami, R. and F.F.Tong, ”A Wavelength Division Multiple Access Network For Computer Communication,” *IEEE Journal on Selected Areas in Communications*, Vol. 8, pp.983–994, August 1990.

[FSK89] Felderman, R., E. Schooler, and L. Kleinrock, ”The Benevolent Bandit Laboratory: A Testbed For Distributed Algorithms,” *IEEE Journal on Selected Areas in Communications,* Vol. 7, No. 2 (ISSN 0733–8716), pp. 303–311, February 1989.

[Ger93] M. Gerla, et al., ”Channel sharing in multihop lightwave networks,” presented at the LAN/MAN Workshop, San Diego, Oct 1993.

[GK91] Ganz, A. and Z. Koren, ”WDM Passive Start-Protocols And Performance Analysis,” *Proceedings of IEEE INFOCOM ’91,* Bal Harbour, FL, pp.991–1000, April 1991

[GKVBG90] Goodman, M.S., Kobrinski, H., Vecchi, M., Bulley, R.M., and J.L. Gimlett, ”The LAMBDANET Multiwavelength Network: Architecture, Applications, and Demonstrations,” *IEEE Journal on Selected Areas in Communications*, Vol. 8, pp.995–1004, August 1990.

[Kov93] M.Kovacevic and M.Gerla, ”T/WDMA strategies in passive optic networks,” *ICC Proceedings,* Geneva, Switzerland, May 1993.

[LA90a] Labourdette, J.F.P. and A.S. Acampora, ”Wavelength Agility In Multihop Lightwave Networks,” *Proceedings of INFOCOM ’90*, pp.1022–1029, San Francisco, CA, June 1990

[LA90b] Labourdette, J.F.P. and A.S. Acampora, ”Partially Reconfigurable Multihop Lightwave Networks,” *Proceedings of GLOBECOM ’90,* San Diego, CA, December 1990

[LA91] Labourdette, J.F.P. and A.S. Acampora, ”Logically Rearrangeable Multihop Lightwave Networks,” *IEEE Transactions On Communications*, Vol.39, No.8, pp.1223–1230, August 1991.

[LGA90] Lee, T.T., Goodman, M.S., and E. Arthurs, ”A Broadband Optical Multicast Switch,” *ISS ’90*, 1990.

[LK92a] Lu, J. and L. Kleinrock, ”On the Performance of Wavelength Division Multiple Access Networks”, International Conference on Communications, Chicago, IL., June 1992.

[LK92b] Lu, J. and L. Kleinrock, ”A Wavelength Division Multiple Access Protocol for High–Speed Local Area Networks with a Passive Start Topology” to appear in Performance Evaluation, 1992.

[Max85] Maxemchuk, N.F., ”Regular Mesh Topologies In Local And Metropolitan Area Networks,” *AT&T Technical Journal*, 64(7), pp.1659–1685, September 1985.

[Max90] Maxemchuk, N.F., ”Deflection Routing In The Manhattan Street Network,” *Proceedings of NATO Workshop on High Speed Networks*, Sophia Antipolis, France, June 1990.

[OS91] Ofek, Y. and M. Sidi, ”Design And Analysis Of A Hybrid Access Control To An Optical Star Using WDM,” *Proceedings of INFOCOM ’91*, Bal Harbour, FL, pp.20–31, April 1991.

[Sch91] Schroeder, M., et al, ”Autonet: A High–Speed, Self–Configuring Local Area Network Using Point–to–Point Links,” *IEEE Journal on Selected Areas in Communications*, Vol. 9, No. 8, Oct. 1991.