High Speed Computing, LANs, and WANS

Larry A. Bergman and Steve Monacos
Jet Propılsion Laboratory
California Institute of Technology
Pasadena, California 91109

1. ABSTRACT

Optical fiber networks may one day offer potential capacities exceeding 10 terabits/sec. This paper describes present gigabit network techniques for distributed computing as illustrated by the CASA gigabit testbed, and then explores future all-optic network architectures that offers increased capacity, more optimized level of service for a given application, high fault tolerance, and dynamic reconfigurability.

Keywords: all-optic networks, terabit fiberoptic networks, scalable supercomputer I/O systems

2. INTRODUCTION

High speed fiber optic networks are envisioned as one of the key enabling technologies that will herald in the information age through the creation of the national information highway. As network capacity increases from present DS3 (45 Mbit/s) rates up to a few gigabits per second, a number of exciting applications will appear, such as closely coupled distributed supercomputing. This capability is being explored with the CASA gigabit testbed interconnects seven large computers at JPL, Caltech, San Diego SuperComputer Center (SDSC), and Los Alamos National Laboratory (LANL) through a SONET OC-48 (2.4 Gbit/s) fiber optic network, This paper will present some of he early results of the CASA gigabit testbed, and conclude on future directions in all-optical networking technology that might eventually achieve data rates beyond 50 Gbit/s. Two examples described will be the spectrally encoded ShuffleNet and the Supercomputer Super Network (SSN).

3. CASA

The CASA Gigabit Network Testbed is one of five testbeds involved in the CNRI led Gigabit Testbed Initiative, a collaboration of numerous research institutions and industrial organizations working together toward the goal of a gigabit network capability for the research and education communities. The objective of the CASA gigabit testbed (Fig. 1.) is to demonstrate that distributed supercomputing over gigabit networks can provide new levels of computational resources for leading-edge scientific problems-in spite of high communications latency. Distributing large computations among several supercomputers provides the opportunity both to bring to bear greater computing power than is available in any single machine and to use the most suitable machine for each step of the task. Two research issues are: (1) to devise algorithms that hide latency and (2) how to write or modify applications software to run in a distributed fashion, These in turn may lead to new programming models or identify the need for new network services and functionality. CASA will use algorithms and software environments that have been developed for parallel computing (e.g., Express). These systems provide mechanisms for sending data between processes, for distributing data among several processors, and for overlapping communication with computation. The applications are organized so that communications latency over the wide-area network can be overlapped with computations. The algorithms are also designed to minimize the number of messages to be sent by predicting in each processor the values that are being computed by other processors.

Three applications were chosen from the areas of chemistry, geophysics, and climate modeling to illustrate different styles of parallel decomposition and traffic patterns. The chemical reaction dynamics model is important in the study of the reaction of fluorine and hydrogen, which is relevant to powerful chemical lasers. These computations involve operations on very large matrices and require frequent communication of large blocks of data between the computers that participate in the calculation, The second application will develop an interactive visualization program for geological applications that takes input from Landsat, seismic, and topographic databases, Among the benefits of such analysis will be much clearer identification of fault zones, plate thrusts, and surface erosion effects (Fig. 2.). The. climate modeling application will combine ocean and atmospheric models simultaneously running in separate computers and continually exchanging boundary data across the CASA network.

The CASA network consists of long–haul Synchronous Optical NETwork (SONET) OC-48 fibers between the four sites (provided by public carriers). These fibers will interconnect HIPPI-based (High Performance Parallel

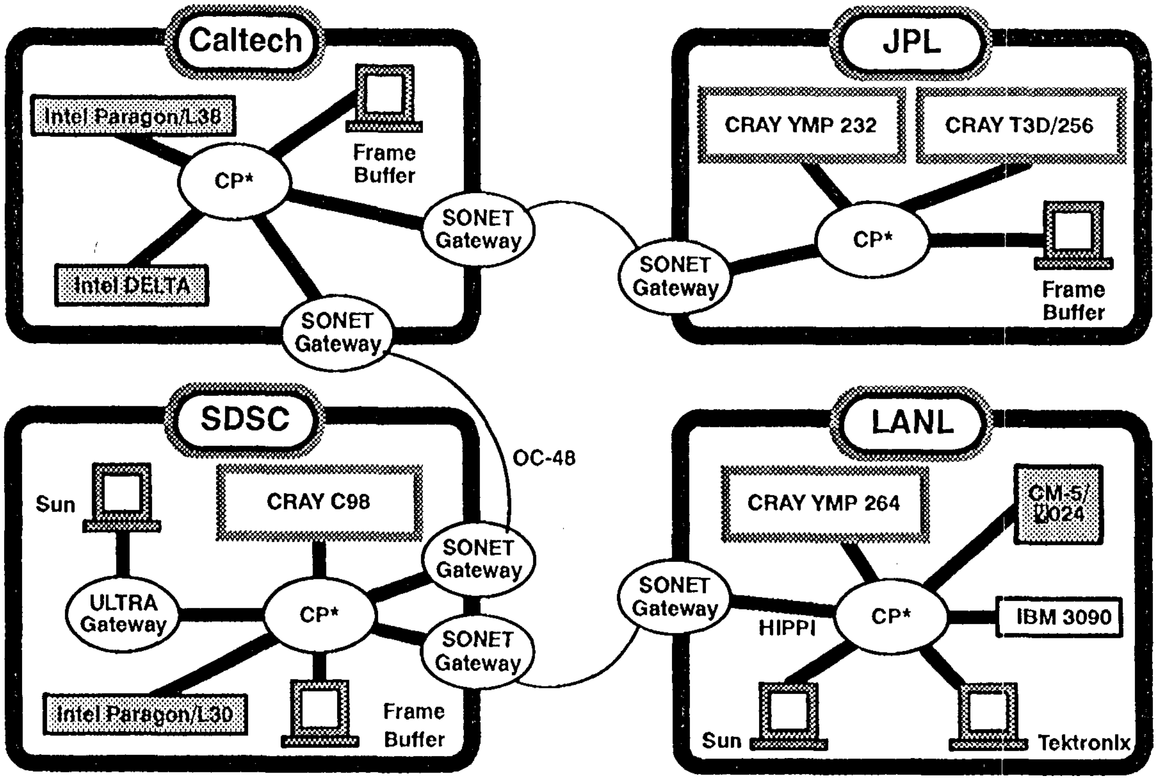


Fig. 1. The CASA gigabit testbed will demonstrate that distributed super computing using wide - area high –speed networks can provide new levels of computational resources for leading– edge scientific problems. The total aggregate meta-supercomputer capacity exceeds 27□GigaFLOPS.

interface) local area networks at each site through HIPPI –SONET gateway developed by LANL. Some crossbar switch ports may also have an outboard TCP/I P accelerator attached for hosts without fast external I/O engines.

4. ALL OPTICAL NETWORKS

All-optic networks typically are defined as networks where the signal is maintained in the photonic domain from source to destination without ever incurring a electronic conversion along the way (e.g., for regeneration). Accrued advantages typically include increased bandwidth with fewer components, massive parallelism suitable for scalable 1/0 in supercomputers, low latency (approaching media propagation speed), and dynamic

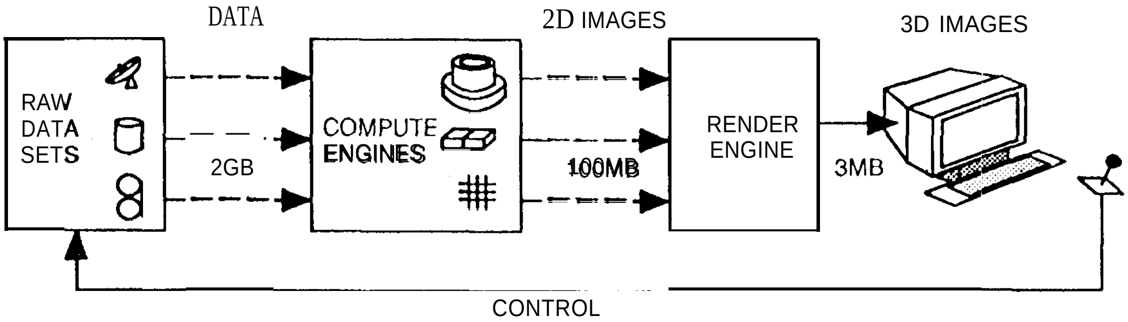


Fig. 2. Data flow for CASA CALCRUST

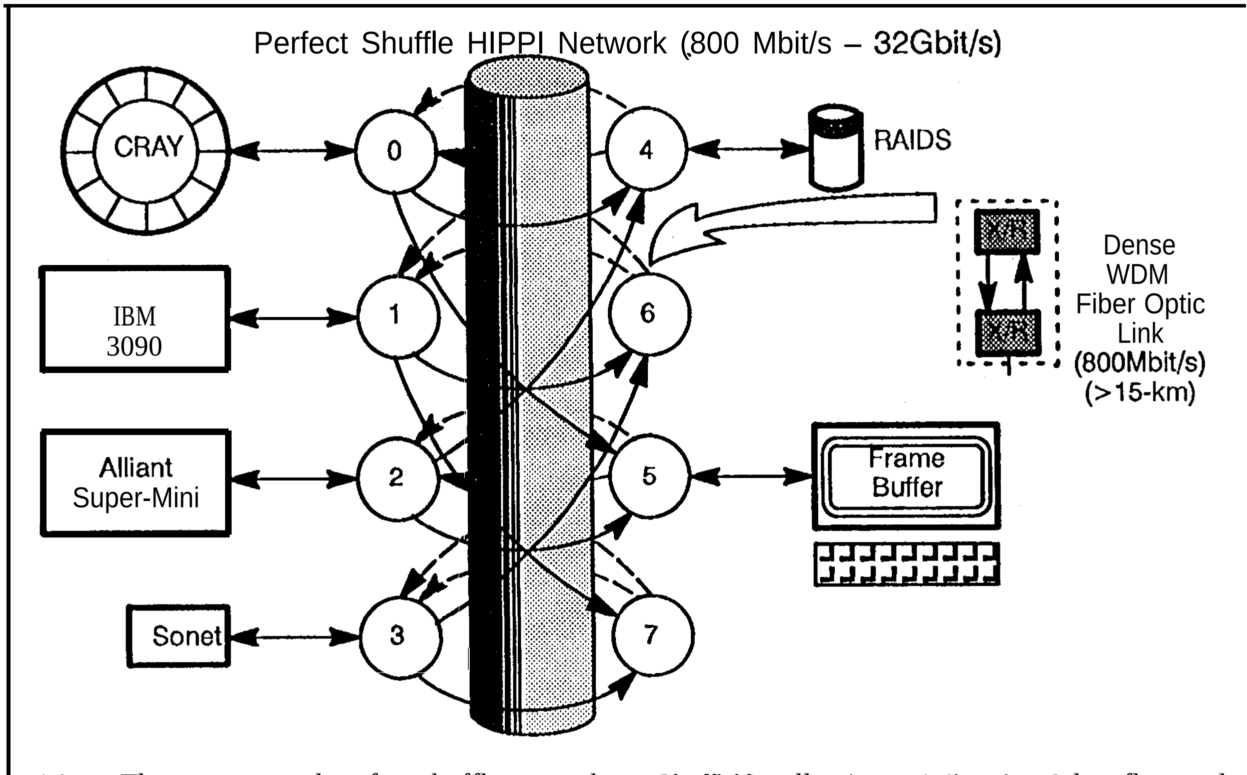

Perfect Shuffle HIPPI Network (800 Mbit/s – 32Gbit/s)
CRAY
IBM 3090
Alliant Super-Mini
Sonet
0
1
2
3
4
6
5
7
RAIDS
X/R
X/R
Dense WDM Fiber Optic Link (800Mbit/s) (>15-km)
Frame Buffer

Fig. 3. The wrap-around perfect shuffle network, or *ShuffleNet,* allowing uni-directional data flow and data storage within the network, Data rates scales to 32 Gbit/s or host interface speed.

reconfigurability. Most all-optical networks currently being studied are based on wavelength division multiplexing (WDM) and designed to support stream services such as voice and video. Typically, their channel switching properties are too slow (~ms) for handling packet switched services, and especially short packets typical of fine-grain distributed computing applications. Two such all optical (or mostly optical) wide area . networks networks that are optimized for low latency and rapid reconfigurability are drawn from the massively parallel processor (MPP) supercomputer internal communications networks family: ShuffleNet and SSN. Both systems offer very low latency in a local area network (LAN) setting, but are also extensible to wide area network (WAN) settings.

S. SHUFFLENET

ShuffleNet (Fig. 3.) is a contraction for shuffle-exchange network. It refers to a cylindrical, wrap–around network of nodes with unidirectional optical external links and a bidirectional link to the node host, Each node host, implemented in conventional electronics, is both a source and sink for network traffic, which makes the network highly symmetric. This results in several desirable network properties: simple packet self-routing procedure at each node, suitable for optics implementation; efficient use of point-to-point bit –serial bandwidth, ideal for fiber data transmission; concurrent packets and thus high network capacity; network capacity that grows monotonically with network size; and multiple paths between nodes raising burst capacity, easing contention and failure modes.

The *Hot Potato* contention resolution algorithm is used at an intermediate node in a multi-hop network. Packets entering a node are only held long enough to determine where they wish to go. Then the packets to be forwarded leave the node, as many as possible on one of the shortest (of possibly several equivalent) paths to its destination. Some packets may be miss-routed, but the penalty for this can be made relatively small. The delay until a packet gets another chance, for networks with moderate internode separation, is small for packets going at the speed of light. Furthermore, miss-routed packets can have increased priority in later contention and thus decreased chances of further miss-routing. In some reasonable ranges of network parameters-traffic asymmetry and loading, internode distance, packet size, and network size—this algorithm can compare in efficiency with the conventional store-and-forward technique. Since the hot potato technique does not require asynchronous storage of the through – going packets, this node protocol is well adapted to lightwave flow-through architectures. It also simplifies host flow buffering requirements since the network possesses *memory.*

Typical applications include fine-grain distributed supercomputing for data fusion on the Global Grid, and implementing a distributed fault tolerant ATM switch in the network. Key ShuffleNet components include. dense WDM fiber optics, stepped wavelength laser diode arrays, photodiode detector arrays, and integrated optic (grating) coupling and wavelength control optics. ShuffleNet then to combines these three interlocking ideas, all of which take advantage of the strengths and mitigate the weaknesses of lightwave transmission.

6. SUPERCOMPUTER SUPERNET (SSN)

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, SSN is a new network (Fig. 4.) that attempts to overcome these problems by replacing the point-to-point links with an all –optical interconnect system. The novel scheme employs asynchronous pipeline crossbar switches used in parallel supercomputers to interconnect multi-channel WDM fiber optic links to an optical star (or tree) "physical" topology. Wavelength Division Multiplexing (WDM) will be used to subdivide the very large fiber bandwidth into several channels, each of Gigabit/sec 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. The low switching latency also makes SSN ideal for many fine grain applications.

6.1. Architecture

Architecturally, 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 intelligent fabric 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 general OPTIMIC architecture is shown in Fig. 4. Each network node consists of an APC constructed from the MyriNet pipeline crossbar ICs (that are asynchronous) and multiple optical channels. The APC 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.

A initial testbed implementation of OPTIMIC (in Fig. 5.) shows 4 MyriNet crowbar (APC) nodes and five isolated STAR -based fiber optic networks. Since tunable laser technology is still quite experimental, fiberoptic 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. 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 x *8* = 24. This number, however, can be reduced to 8, by time division multiplexing three 800 Mbps subchannels cm a single WDM channel operating at 2.4 Gbps. The 800 Mbps channel rate is adequate for our purposes since it exceeds the 640 Mbps MyriNet chip rate. In addition, a pool of 24 channels, at 800 Mbps 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 channels at 800 Mbps each. The maximum number of hosts that can be connected to the system is 8 *x* 24 x 4 = 768. In our target configuration, up to 50 hosts could 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.

6.2. Optical Channel Interface (OCI).

The Optical Channel Interface board (Fig. 6.) or OCI is responsible for buffering and switching between APC 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,

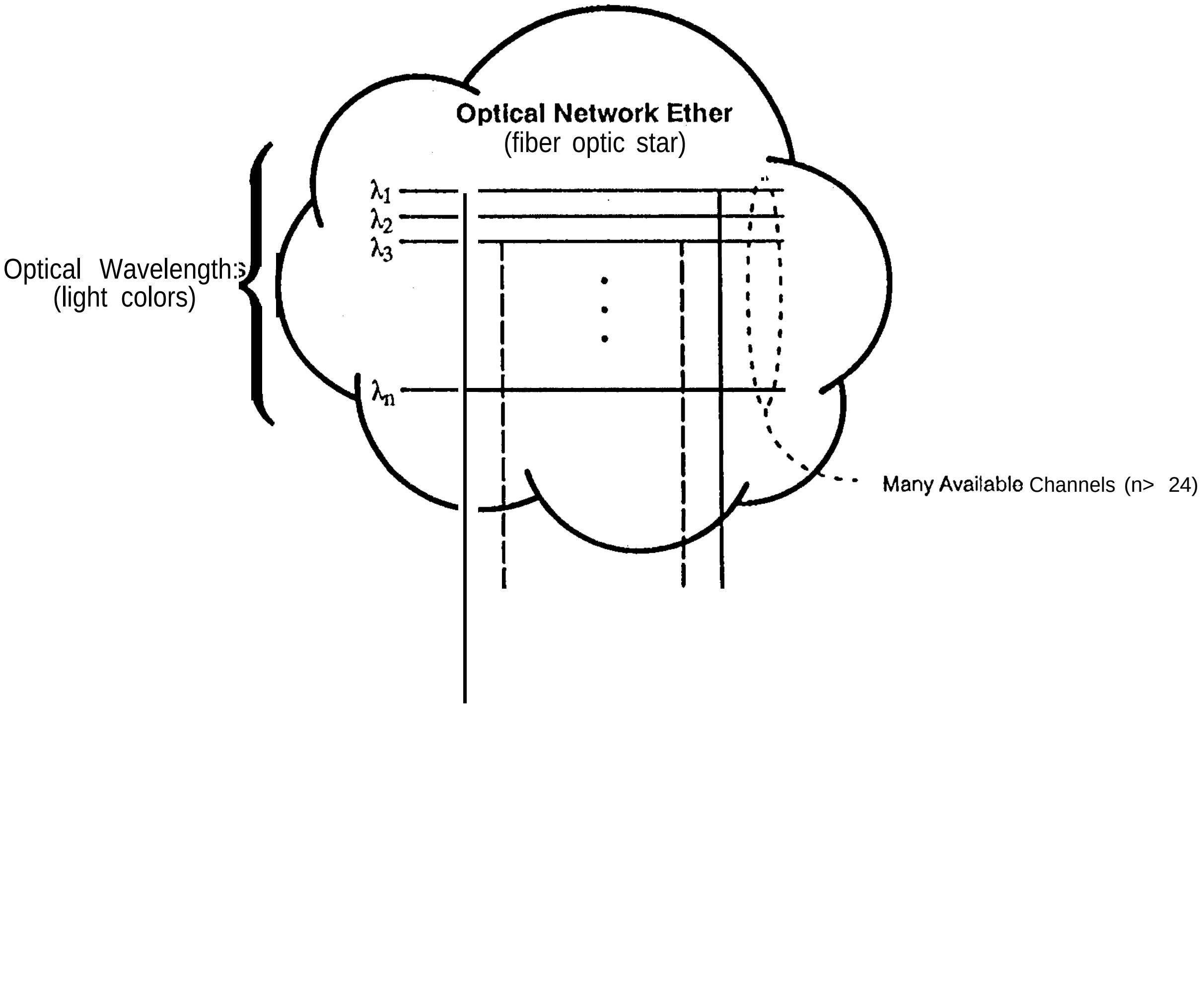
6.3. Asynchronous Clock Recovery

There are several types of clock recovery, framing, and multiplexer/demultiplexer ICs commercially available today that operate at Gbit/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.

6.4. 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 APC 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 maybe 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 (spatial) described as (1) 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.

The most effective technology in terms of maximizing system performance would be to utilize tunable laser diodes combined with tunable Fabry Perot receivers (Fig. 7.). This would produce the richest network virtual topology, maximize aggregate capacity, and minimize probability of blocked states.

7. ENABLING APPLICATIONS

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

7.1. Examples

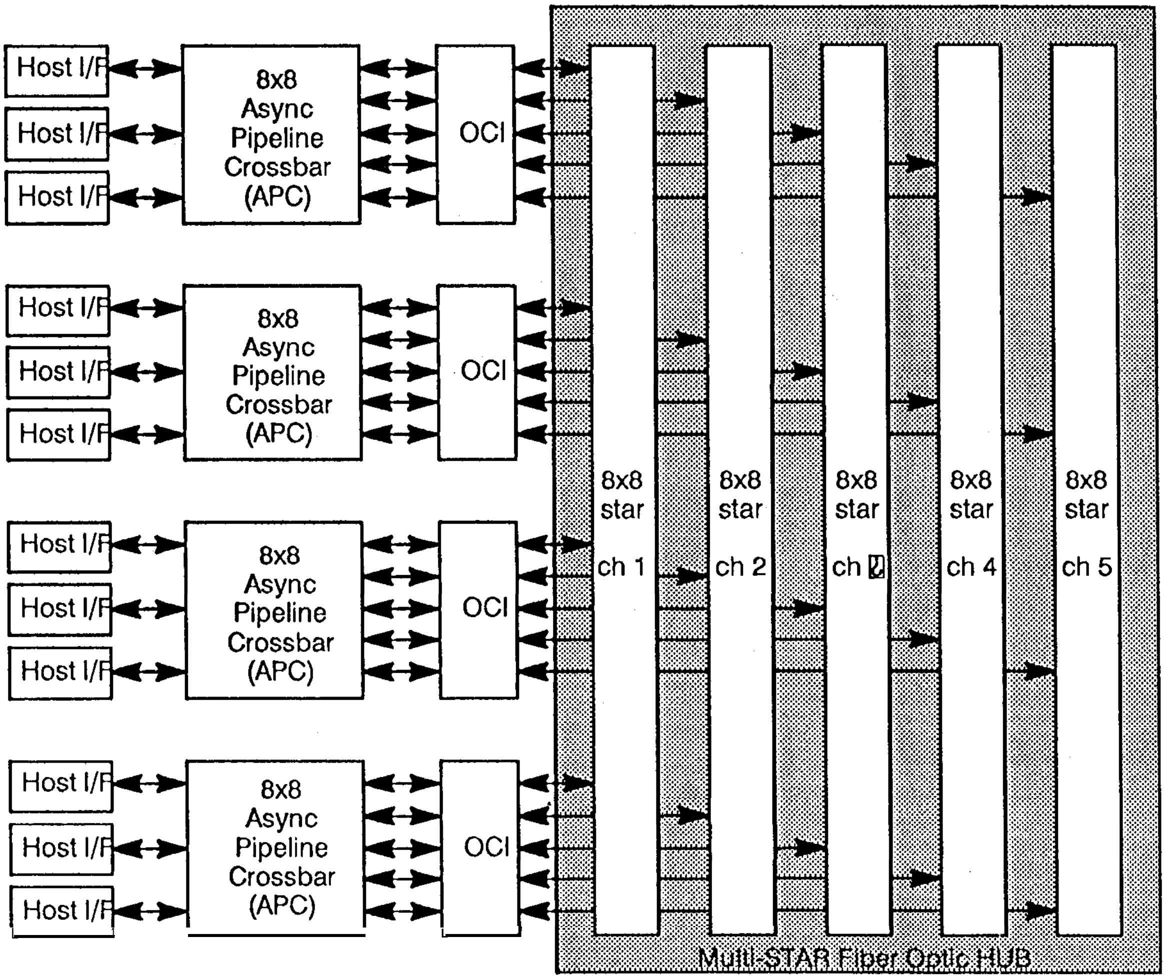


Fig. 5. Phase 1 OPTIMIC hardware architecture consists of 4 network pipeline crossbar nodes and five optical channels. Each optical channel of an 8x8 optical star (3 ports are unused) plus associated fiber optic full-duplex fiber optic transceivers. The media is fiberoptic ribbon cable (multimode for campus and single-mode for MAN).

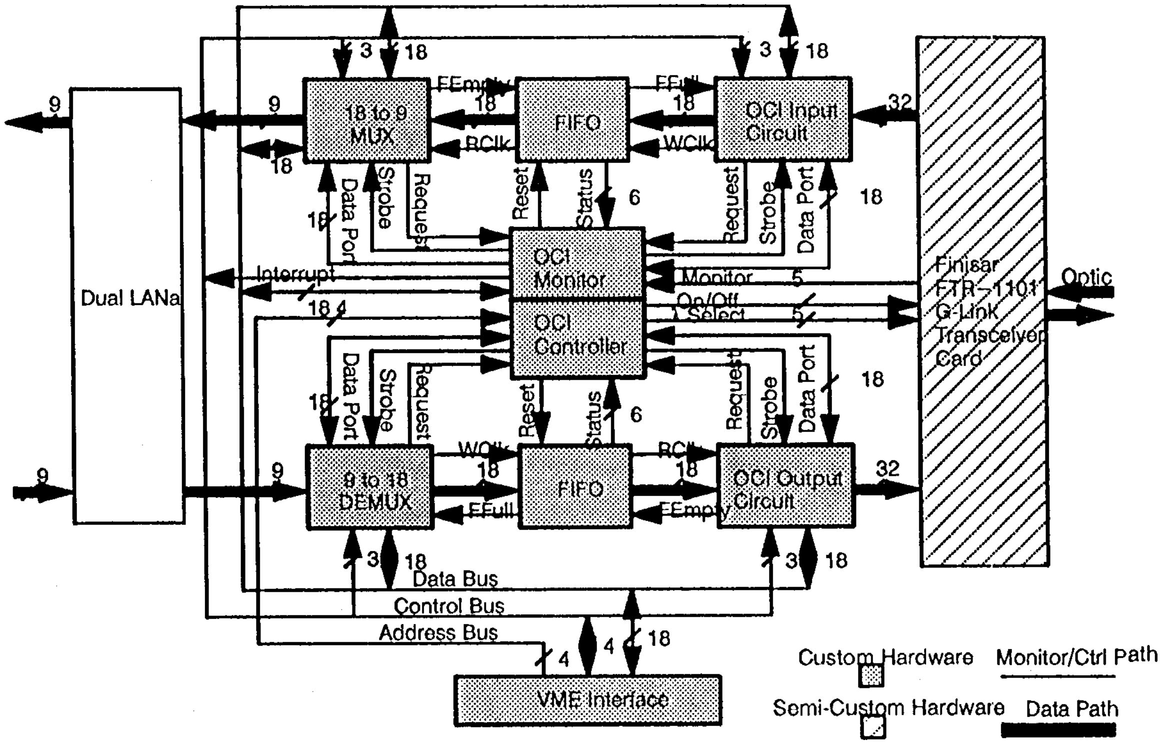


Fig. 6. The OPTIMIC Optical Channel Interface (OCI) acts as a interface between the Myri-Net asynchronous switched network and the synchronous optical links. It also transparently sets up and tears down WDM channels as needed.

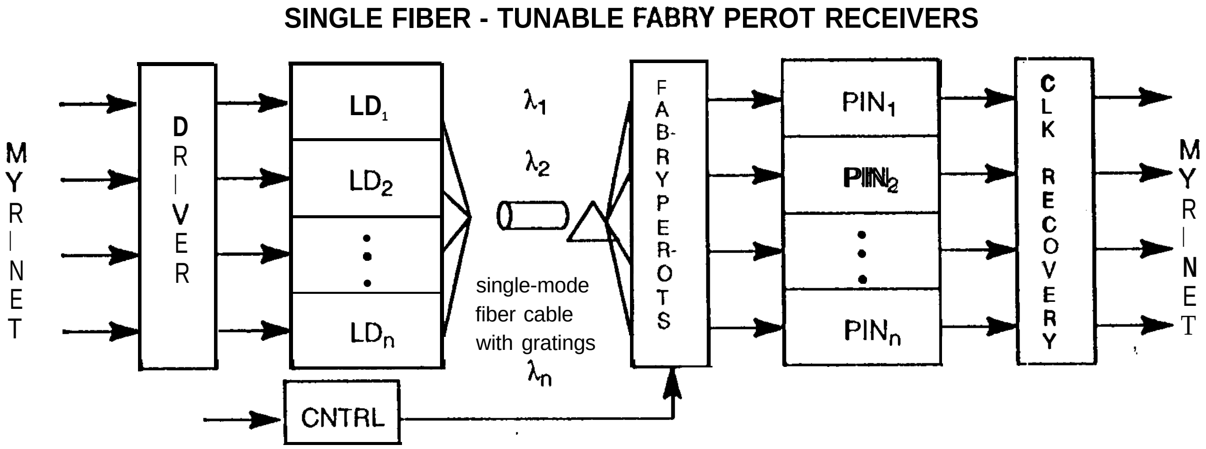

SINGLE FIBER - TUNABLE FABRY PEROT RECEIVERS
MYRINET
DRIVER
LD1
LD2
LDn
λ1
λ2
λn
single-mode fiber cable with gratings
FABRY-PEROTS
PIN1
PIN2
PINn
CLK RECOVERY
MYRINET
CNTRL

Fig. 7, WDM link technology using tunable Fabry Perot receivers.

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 SATCOM network, SDI BE, 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.

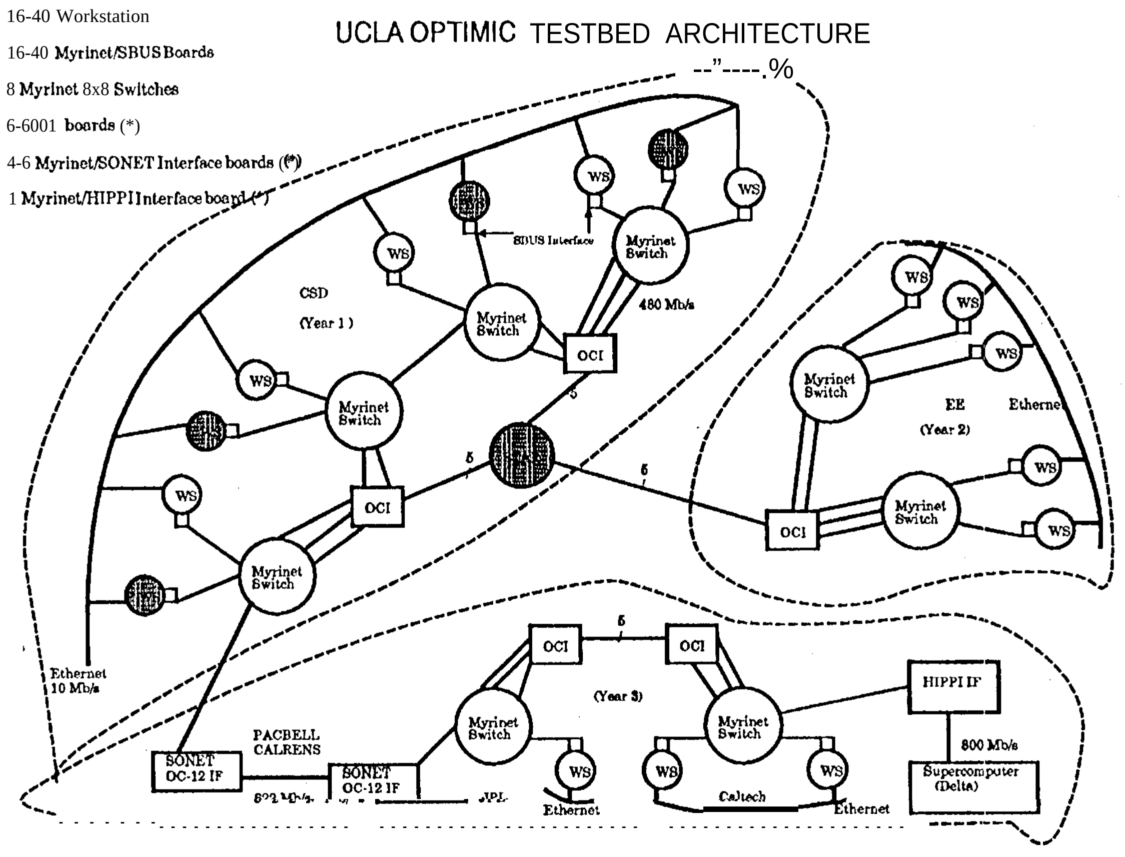


Fig. 8. The OPTIMIC testbed consists of three clusters of APC switches: four at the UCLA engineering and computer science building, two at the atmospheric physics building, and two in Pasadena (one at JPL and one at Caltech).

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).

7.2. Target Demonstration Application

The basic OPTIMIC testbed topology is shown in Fig. 8.. APC switching nodes are placed in three clusters: a group of four in the UCLA engineering and computer science building, a group of two in the UCLA Atmospheric Physics building, and a group of two at JPL/Caltech. OCIs interconnect the three clusters as well as selected ports within the largest cluster at the UCLA Computer Science Department.

One fiber optic link segment (14km) of 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 (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 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, 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

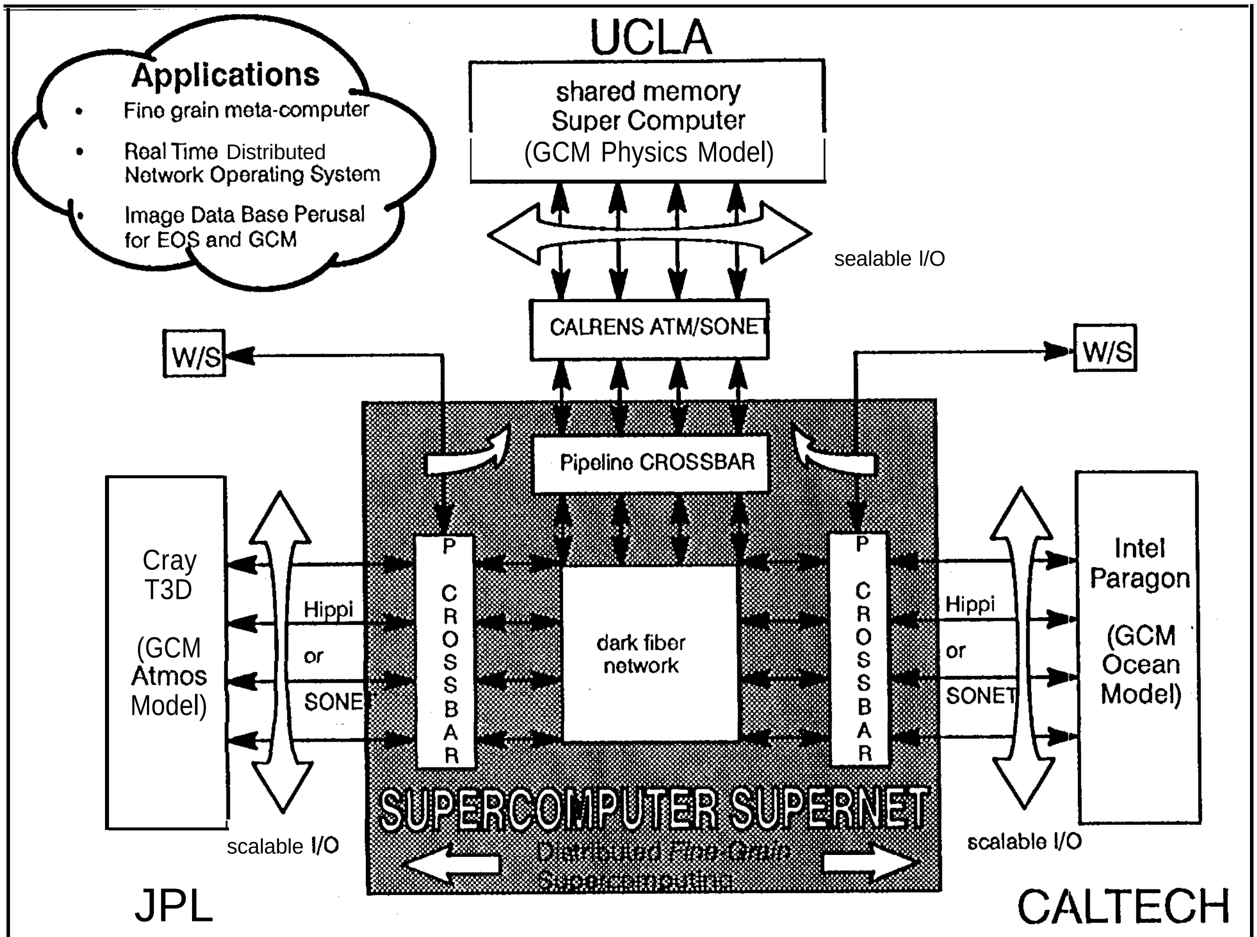

UCLA
Applications
Fine grain meta-computer
Real Time Distributed Network Operating System
Image Data Base Perusal for EOS and GCM
shared memory Super Computer (GCM Physics Model)
sealable I/O
CALRENS ATM/SONET
W/S
W/S
Pipeline CROSSBAR
Cray T3D (GCM Atmos Model)
Hippi or SONET
P CROSSBAR
dark fiber network
P CROSSBAR
Hippi or SONET
Intel Paragon (GCM Ocean Model)
SUPERCOMPUTER SUPERNET
Distributed Fine-Grain Supercomputing
scalable I/O
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.

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 MPP nodes enter/leave the network, The MyriNet APC network node also accommodates instantaneous reconfiguration of the MPP 1/0 channels from asynchronous 1/0 for separate partitioned jobs (e.g., one per quadrant of the MPP) to coherently striped I/O for one large single job.

8. ACKNOWLEDGMENT

The author would like to acknowledge the contributions R. Hartmayer and J. Morookian at JPL, J. Sauer at University of Colorado, and L. Kleinrock and M. Gerla at UCLA. The research described in this paper was performed by the Center for Space Microelectronics Technology, Jet Propulsion Laboratory, California Institute of Technology, and was jointly sponsored by ARPA (in part, under contract DABT63–93–C–0055), NSF, CNRI, and BMDO/IST through an agreement with the National Aeronautic and Space Administration and the University of California, Los Angeles. 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, the University of California, Jet Propulsion Laboratory, or the California Institute. of Technology.

9. REFERENCES

[1] Acampora, AS., 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/Decem ber 1987.

[2] J,Bannister et al, " All optical multifiber tree network", IEEE Journal of Lightwave Technology, May/June. 1993.

[3] Bracket, C.A., "Dense Wavelength Division Multiplexing Networks: Principles And Applieations", IEEE, Journal on Selected Areas in Communications, Vol. 8, pp.948-964, August 1990.

[4] M.Kovacevic, M.Gerla and J. Bannister " T/WDMA strategies in passive optic networks", ICC Proceedings, Geneva, Switzerland, May 1993,