M. Jonsson, "Optical interconnections in parallel radar signal processing systems," Research Report CCA - 9909, Centre for Computer Systems Architecture (CCA), Halmstad University, Sweden, Apr. 1999.

Optical Interconnections in Parallel Radar Signal Processing Systems

Magnus Jonsson

Halmstad University, Halmstad, Sweden
email: magnus.jonsson@cca.hh.se

Abstract

Optical interconnection networks is a promising design alternative for future parallel computer systems. Numerous configurations with different degrees of optics, optoelectronics, and electronics have been proposed. In this paper, some of these interconnection networks and technologies are briefly surveyed. Also, a discussion of their suitability in radar signal processing systems is provided, where several different ways of coarse algorithm mapping are considered.

Contents

1. Introduction

Future radar signal processing systems will have high computational demands, which implies parallel computer systems. In turn, the performance of parallel computers is highly dependent on the performance of their interconnection networks. In this paper, optical interconnection networks are reviewed from a signal processing perspective. The networks are evaluated according to how suitable they are to map signal processing chains in, e.g., radar systems. Different ways of mappings are considered. These kinds of mappings are simplified in their nature to ease evaluation of the diverse range of networks. More detailed discussions on algorithm mapping in similar signal processing systems can be found in [Liu and Prasanna 1998].

Two categories, according to Flynn's classification [Flynn 1966] [Flynn 1972], for which parallel computers may be classified into are MIMD (Multiple Instruction streams Multiple Data streams) [Hord 1993] and SIMD (Single Instruction stream Multiple Data streams) [Hord 1990]. MIMD computers are typically more coarse grained than SIMD computers which, instead, have more tightly coupled processing elements. MIMD computers may be further divided into distributed-memory multicomputers and shared-memory multiprocessors. Shared-memory multiprocessors of today normally employ mechanisms to retain cache-coherency among the processing elements [Stenström 1990] [Tomaševic and Milutinovic 1994] [Tomaševic and Milutinovic 1994B]. Parallelism can exist on a lower level too (i.e., on a more fine grained level), e.g., instruction level parallelism (ILP).

Most of the interconnection networks mentioned in this paper are more suitable for coarse-grained systems, but not all. Referring to a node in this paper means a computer system with one interface to the network. A node can actually be a parallel computer itself. For example, it can be a SIMD computer, i.e., the system is a MIMD computer on the high level, connecting multiple autonomous SIMD computers [Taveniku et al. 1998]. This configuration is related to the MSIMD (multiple-SIMD) architecture of PASM [Siegel et al. 1981]. Optical interconnection technology can actually be employed on several levels in a system, e.g., fiber-optics connecting nodes on a coarse level, where each node contains free-space optics to connect processing elements internally in the node.

After an introduction to interconnection networks in parallel and distributed computers in Section 2, a general discussion on optical interconnections in parallel computers is given in Section 3. Then, in Section 4, different system configurations and requirements for, especially, radar signal processing systems are explained. In Section 5, optical interconnection systems and technologies are briefly surveyed and evaluated according to the developed "criterias" explained in Section 4. The paper is then concluded in Section 6.

2. Interconnections in Parallel Computers

Interconnection networks are often divided into static (also called direct) and dynamic (also called indirect) networks. Static networks have a defined static topology where the nodes are directly connected to nearest neighbours via point-to-point links. This forms a static topology of the network, e.g., a 2-dimensional mesh. In dynamic networks the traffic is routed through a switched-based network. Therefore, we can say that the nodes are indirectly connected to each other.

Not all networks fall into one of the two categories given above. Therefore, two more categories can be added: shared-medium networks and hybrid networks [Duato et al. 1997]. After a presentation of different parameters of interconnection networks in Subsection 2.1, static, dynamic, shared-medium, and hybrid networks are described in Subsections 2.2, 2.3, 2.4, and 2.5, respectively. Then, in Subsection 2.6, different kinds of group communication are presented. After a discussion of routing in Subsection 2.7, the section ends with an overview of different high-performance networks in Subsection 2.8.

2.1 Design and performance parameters

From text books covering the area of parallel computing [Almasi and Gottlieb 1994] [Casavant et al. 1996] [Decegama 1989] [Hockney and Jesshope 1988] [Hwang 1993] [Hwang and Briggs 1985] [Lawson et. al. 1992] and tutorial texts on interconnection networks for parallel computers [Bhuyan et al. 1989] [Duato et al. 1997] [Reed and Grunwald 1987] [Siegel 1990] [Varma and Raghavendra 1994], we can find various design and performance parameters and desired features of interconnection networks for parallel computers. Moreover, more general network discussions and concept definitions are found in computer communication text books [Halsall 1995] [Peterson and Davie 1996] [Stallings 1997] [Tanenbaum 1996]. We will explain some of these below for which several have influence on the analysis.

Fault tolerance
Redundant paths in the network can bring fault tolerance.

Latency and delay
Latency and delay are terms that can be defined in a number of ways. One common definition, normally called message delay, is the time from message generation in the source node until the message is

fully received at the destination node and available for use by the application running on it.

Transmission capacity

Transmission capacity is measured in bit/s or Byte/s and denotes the maximum amount of data that can be transferred per time unit over a link, or aggregately over a whole interconnection network. Transmission capacity is often, but somewhat misleading [Freeman 1998], called bandwidth.

Throughput

The term throughput is used when describing efficiently used transmission capacity at different assumptions like certain traffic patterns and message generation rates. It is usually measured in bit/s or number of delivered messages per time unit. In the ideal parallel computer system a PE can have a sustained throughput as high as the peak throughput, independent of the traffic pattern from other PEs.

Bisection bandwidth

The bisection of a system is the section that divides the system into two halves with equal number of nodes. The bisection bandwidth is the aggregated bandwidth over the links that cross the bisection. In asymmetric systems the number of links across the bisection depends on where the bisection is drawn. However, since the bisection bandwidth is a worst-case metric, the bisection leading to the smallest bisection bandwidth should be chosen [Hennessy and Patterson 1996].

Cost

An interconnection network designer always want to bring the cost down as long as possible. Instead of measuring cost in a monetary value, it can be measured in, e.g., number of links or switches.

Conflict free

The ideal network should be conflict-free. If there exist a physical connection between each pair of nodes (or a full crossbar is used), conflicts in the network are avoided. Conflicts caused by, e.g., limited amount buffers must, however, still be considered. Since N^2 connections (or cross points in a crossbar) are needed, a conflict-free network will be too expensive in larger systems.

Uniformity

Sometimes it is desirable to have uniform latency and bandwidth, independent of which pair of nodes that communicates with each other. However, parallel computers with a non-uniform network can scale to a large number of PEs for applications with a high degree of local communication [Agarwal 1991].

Circuit switching vs. packet switching

When using circuit switching, a "physical" channel is allocated before communication between a pair of PEs can start. The "physical" channel do not need to be pure physical but can, e.g., be a cyclically available time-slot on a time multiplexed channel. The advantage with circuit-switching is the guaranteed bandwidth, while disadvantages are long setup times and low bandwidth utilization when the channel is idle for a long time, since the bandwidth normally can't be reused. When using packet switching, the data to be transferred is split into packets that compete with packets from other nodes for the bandwidth. Traffic situations with temporary bursts of large data volumes from one or a few PEs can therefore be handled better in a packet-switched network than if much of the bandwidth were allocated by other nodes using circuit-switching. Also, sporadic traffic often experience a shorter latency than if a circuit should be setup each time. Handling real-time traffic is, however, harder in a packet-switched network.

Connectionless vs. connection-oriented service

When using connection oriented service the two communicating parties first agree upon establishing a logical "error free" connection. After the phase of data transferring, the logical connection is disconnected. During data transfer, each correctly received packet (or similar) is acknowledged to the transmitting party. If not acknowledged, or retransmission is explicitly requested, a packet is retransmitted. Also, to guarantee that packets arrives in correct (transmitted) order, each packet is stamped with a sequence number. Connectionless service do not involve connection establishment and is not an error free service. Because the protocol entity in a destination PE is not aware of forthcoming packets it cannot ask for retransmission of packets that are corrupted and discarded before arriving to it.

Scalability — When more PEs are added to a scalable system, performance in terms of, e.g., network capacity should increase proportionally to avoid bottlenecks. Sometimes, scalability only refers to scalability in one dimension. In this way, *size scalability* can denote the ability to build arbitrarily big systems with a chosen (network) architecture and increase the number of nodes with no or small modifications of the architecture.

Incremental expandability

The possibility of expanding a system with another small subsystem instead of, e.g., being forced to double the number of nodes to maintain a certain topology, is often desirable. It can also be the case that a parallel computer system is optimized for a certain size which can lead to, e.g., unused communication links at smaller system sizes. The term *modularity* is sometimes used instead of, or at least coupled to, the term incremental expandability.

Control strategy — The control strategy can be either centralized (global) or decentralized (local) and is normally referring to the way of steering switches in the network. In a network with centralized control a single controller steers the whole network. This might imply that all switches in a network are set at the same time to optimize for a certain global communication pattern. In a network with decentralized control each switch decides on its own how to route incoming messages.

2.2 Static networks

In static networks a certain static topology is chosen. Some common topologies are linear array, ring, 2-dimensional mesh, 2-dimensional torus, and binary hypercube (Figure 1). The parameters below are used to describe static networks.

Diameter — For each possible pair of nodes in the network there exist a shortest path. The diameter is defined as the number of hops over the longest of these shortest paths.

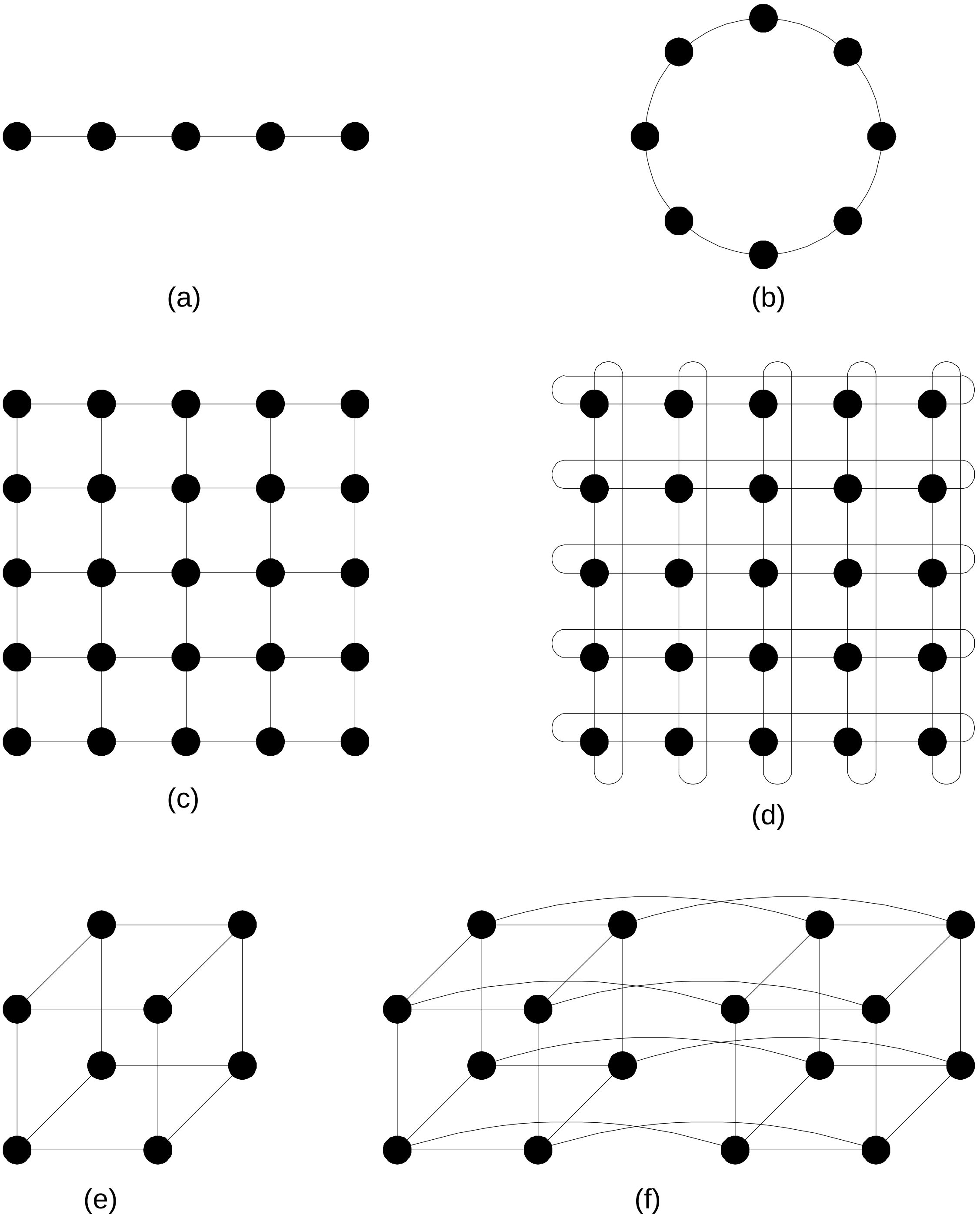

Figure 1: Some static topologies: (a) linear array, (b) ring, (c) 2-dimensional mesh, (d) 2-dimensional torus, (e) 3-dimensional binary hypercube, and (e) 4-dimensional binary hypercube.

Node degree   Number of links that connects a node to its nearest neighbors. The node degree can either be constant for the whole network or differ between the nodes. As an

> example, the boundary nodes in a 2-dimensional mesh has a node degree of 3 and corner nodes a node degree of 2, while the rest of nodes has a node degree of 4. A constant node degree is an example of a feature that might make expansions of the system easier.

If we let N denote the number of nodes or PEs in a system and assume bidirectional links, we can, with these parameters, characterize the different topologies in the figure. The linear array has a node degree of two, except for the end nodes, and a diameter of $N - 1$. The linear array topology is employed in, e.g., REMAP-β [Bengtsson et al. 1993].

The ring has a constant node degree of 2, while the diameter is $N / 2$. It is cheap to expand a ring network but since the diameter increases with N, only small systems are practical. An example of a computer using a ring network is KSR-1 [Almasi and Gottlieb 1994]. For larger systems, however, a hierarchy of rings must be used.

The node degree of the 2-dimensional mesh is discussed above, while the diameter is $2(\sqrt{N} - 1)$. If a mesh network is extended with wrap-around connections it is called a torus. A torus has lower diameter than the mesh, $\sqrt{N}$ for the 2-dimensional case, and a uniform node degree. Examples of parallel computers with a 2-dimensional mesh interconnection network are the distributed shared memory multiprocessor from MIT, MIT Alewife machine [Agarwal et al. 1995], and the MPP SIMD computer (with the extension of reconfigurable function of the boundary nodes) [Batcher 1980] [Batcher 1980B]. A 3-dimensional mesh network is used in the J-machine at MIT [Dally et al. 1993].

The MasPar MP-1 [Blank 1990] [Nickolls 1992], MP-2 [MasPar 1992], and the embedded version by Litton/MasPar [Smeyne and Nickolls 1995] has a 2-dimensional torus network where each node is connected to its eight nearest neighbors by the use of shared X connections. Moreover, the Fujitsu AP3000 distributed-memory multicomputer has a 2-dimensional torus network [Ishihata et al. 1997], while the CRAY T3D [Kessler and Schwarzmeier 1993] [Koeninger et al. 1994] and T3E are both computers that use the 3-dimensional torus topology.

A binary hypercube has two nodes along the side in each dimension, i.e., a total of 2^k nodes, where k is the dimension. The diameter is k since the maximum distance to travel is one hop in each dimension. The constant node degree of a binary hypercube is also k. Examples of computers with hypercube interconnection networks are the Cosmic Cube with a 6-dimensional hypercube [Seitz 1985], and the Connection Machine CM-2 with a 12-dimensional hypercube where each node in the hypercube consists

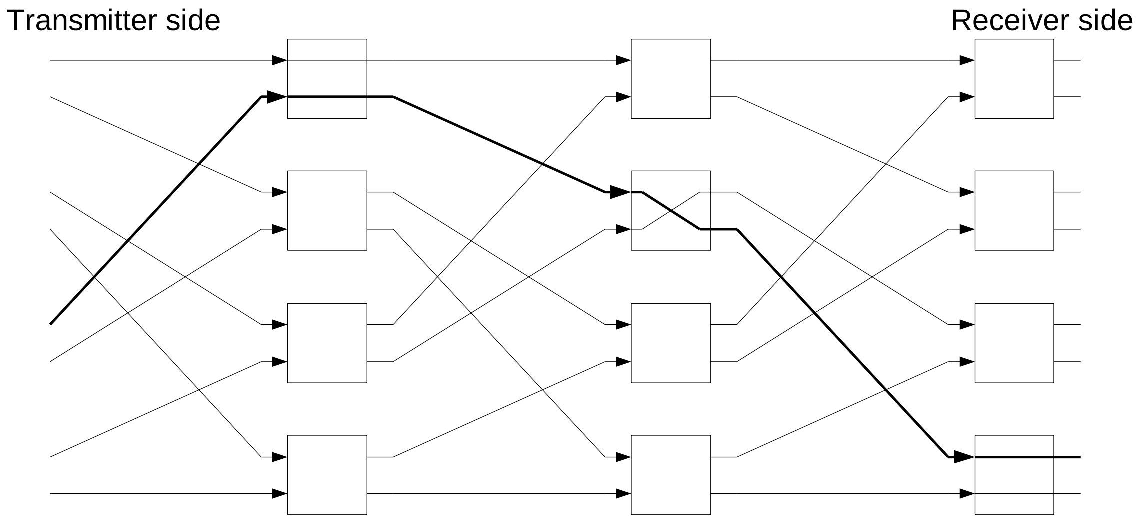

Figure 2: Omega network for an eight-node system. One path through the network is highlighted. As an example, the upper-most connections to the network on the transmitter and receiver sides are connected to the same node.

of 16 PEs [Thinking Machines 1991]. Replacement of 1024 wires in the CM-2 by two optical fibers has been demonstrated [Lane et al. 1989].

More references and a discussion on static networks are found in [Stojmenovic 1996].

2.3 Dynamic networks

The crossbar is the most flexible dynamic network and can be compared with a fully connected topology, i.e., point-to-point connections between all possible combinations of two nodes. The drawback, however, is the increases by N^2 in cost/complexity of the switch, where N is the number of nodes. Systems with a single true crossbar are therefore limited to small systems. Starfire from Sun Microsystems is a symmetric multiprocessor system with a 16 x 16 crossbar network for transferring data transactions [Charlesworth 1998]. For the snoopy cache-coherence protocol, a bus is also used in the system. Other computers with a crossbar network include the VPP500 [Miura et al. 1993] and VPP700 [Uchida 1997] from Fujitsu.

In multistage shuffle-exchange networks, the cost is reduced to increase by $N \log_2 N$, but where $\log_2 N$ stages must be traversed to reach the destination. An example of such a network is the Omega network (Figure 2) which provides exactly one path from every input to every output. The four different switch functions of the 2 × 2 switches that are used as building blocks are shown in Figure 3, where the two rightmost configurations are used for broadcast. Larger switches than 2 × 2 can also be used when

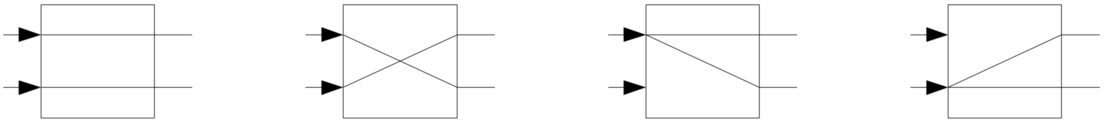

Figure 3: Possible states of a 2 ´ 2 switch.

building multistage networks but such systems are not treated here. Each stage of switches in an Omega network is preceded by a perfect-shuffle pattern. In contrast with a crossbar network, which is a *nonblocking* network, an Omega network is a *blocking* network. This means that there might not always exist a path through the network due to already existing paths that block. An Omega network is used in the NYU Ultracomputer which is a globally shared memory multiprocessor [Gottlieb et al. 1982] [Gottlieb et al. 1983]. The network connects *N* processing elements to *N* memory modules. The Cedar system is another globally shared memory multiprocessor where a multistage shuffle-exchange network connects processors to memory modules [Kuck et al. 1993].

Rearrangeable networks is a third category of multistage networks where it always is possible to find a path through the network. However, if not all paths are routed at the same time it might be needed to reroute already existing paths. An example of a rearrangeable network is the Benes network shown in Figure 4.

A bi-directional multistage network [Stunkel et al. 1994] [Stunkel et al. 1995] [Sethu et al. 1998] is used in the IBM SP2 [Agerwala et al. 1995]. Switch boards each having multiple 8 × 8 switches, coupled as 4 × 4 bi-directional switches, are used when configuring a system. Other multistage

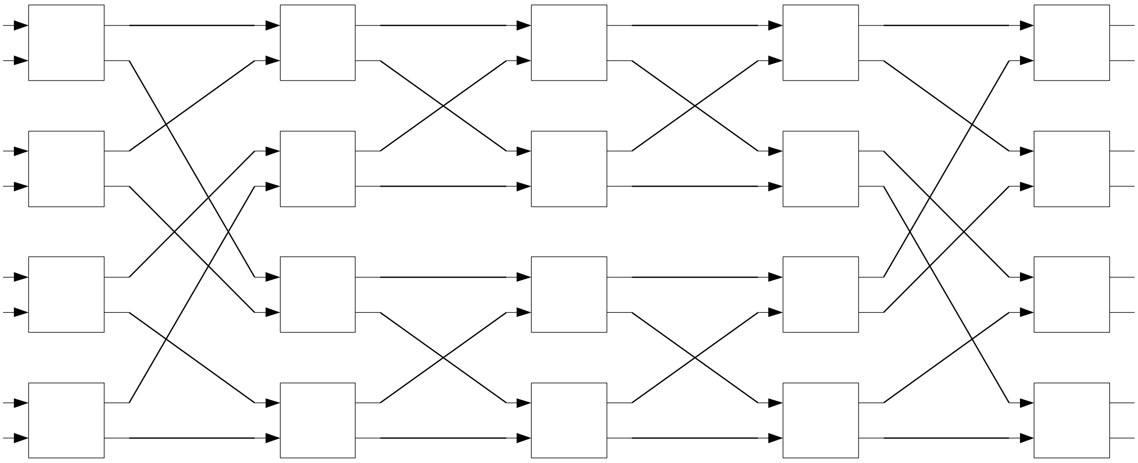

Figure 4: An rearrangeble Benes network.

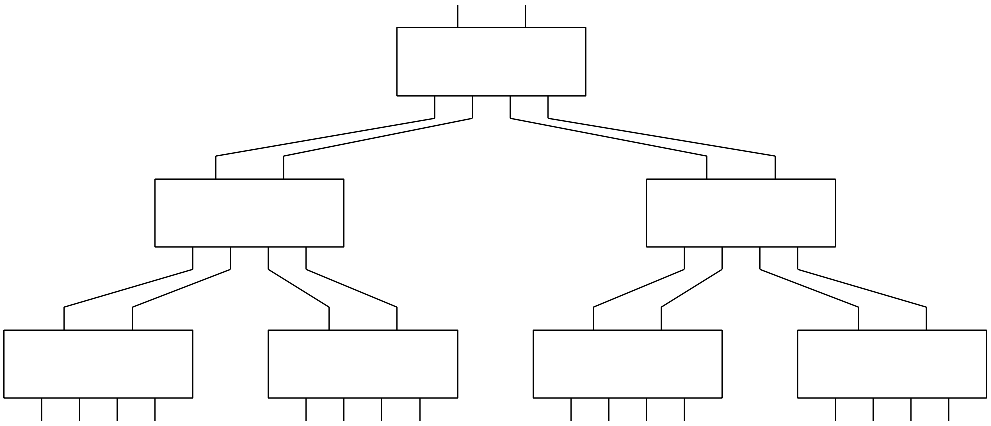

Figure 5: Fat-tree of switches where nodes are leaves in the tree.

networks include Banyan networks [Goke and Lipovski 1973].

A tree network can be seen as a static network topology. However, as long as there are nodes only at the leaves, the tree of switches can be seen as a self-standing system connecting the nodes indirectly with each other. The root nodes in a tree network easily becomes a bottleneck. This problem is solved in the scalable fat-tree network where higher-bandwidth links are used as closer to the root they are [Leiserson 1985]. An example of a (not fully scalable) switched fat-tree network built up of 6 × 6 switches is shown in Figure 5. Each link in the figure is bi-directional. The CM-5 is an example of a parallel computer with a fat-tree network [Hillis and Tucker 1993] [Leiserson et al. 1992]. The fat-tree topology is also used in the Meiko CS-2 [Beecroft et al. 1994].

2.4 Shared-medium networks

A common way of implementing a shared-medium network is to use the bus topology but it can also be, e.g., a ring where only one node is allowed to send at a time. The big advantage of a shared-medium network is the easy implementation of broadcast which is useful in many situations. The disadvantage is that the bandwidth does not scale at all with the number of nodes. The bus is commonly used in small systems using the snooping-on-the-bus cache-coherence protocol, e.g., in Silicon Graphics Power Challenge [Silicon 1994]. Multiple buses can be used to enhance performance relative single-bus systems [Mudge et al. 1987].

2.5 Hybrid networks

An example of a hybrid network is the hierarchical network in Stanford DASH [Lenoski et al. 1992] [Lenoski et al. 1992B] [Lenoski et al. 1993]. A 2-

dimensional mesh network connects bus-based clusters. The aim of the configuration is to get a scalable cache-coherent shared memory multiprocessor. The cache-coherence protocols used are snoopy-on-the-bus inside each cluster and a distributed directory-based protocol between the clusters. Paradigm, instead, uses a hierarchical bus network in its cache-hierarchy implementation [Cheriton et al. 1991]. The CRAY APP, in turn, groups the processing elements in groups of up to 12 processing elements connected to a common bus. Up to seven such buses of processing elements can be connected, via a crossbar, to a globally shared memory [Carlile 1993].

2.6 Group communication

Many parallel programs can take advantage of special support for group communication, or collective communication, i.e., communication where many nodes (or processes) are collectively involved. The nodes involved in a group communication operation are said to be members of a group. Some kinds of group communication are:

- *Multicast:* One-to-many communication where one node sends the same message to all members of the group. The special case where all nodes in the system are members of the group is called broadcast.
- *Scatter:* One-to-many communication where one node sends different messages to different members of the group.
- *Reduction (global combining):* Many-to-one communication where different messages from different members of the group are combined into one message for delivery to one destination node. Some common operators used when combining are SUM, OR, and AND.
- *Gather:* Many-to-one communication where different messages from different members of the group are concatenated in a defined order for delivery to one destination node.
- *Reduce and spread:* Variant of the reduction operator, where the result is spread to all group members.
- *Barrier synchronization:* A synchronization point is defined in the program code, for which all members must arrive to before any of the members may continue beyond the synchronization point. This is a special case of "reduce and spread" where no data is involved.
- *Scan:* For each member of the group, m_i, where $1 \leq i \leq M$, a reduction is made where the node is chosen to be the destination node. Each such reduction is made from a sub-group of $N \leq M$ nodes, e.g., nodes m_j, where $i - N \leq j \leq i$.

The possibility to define groups and call group communication routines is supported in, e.g., PVM (Parallel Virtual Machine) [Geist et al. 1994]. The underlying group communication mechanisms can be implemented in several different ways: (*i*) in software using the same network as for ordinary traffic, (*ii*) by the use of a more or less general network dedicated for group communication as in the CM-5 [Leiserson et al. 1992], and (*iii*) by dedicated hardware specialized for, e.g., barrier synchronization [O'Keefe and Dietz 1990] [O'Keefe and Dietz 1990B].

2.7 Routing

The routing decision, or path selection, in a packet-switched interconnection network can be made either inside the network by routers or by the end nodes, so called source routing. In parallel computers, source-routing can be more easily used than in, e.g., internet communication [Comer 1995], because the topology does not change so often and is normally not so complex and/or irregular. Source-routing is used in, e.g., Myrinet [Boden et al. 1995] and the IBM SP2 communication system [Stunkel et al. 1995], where the header of a packet includes the desired switch-setting of each router on the way from source to destination. Each router drops its corresponding switch-setting field in the header when the packet is forwarded.

If each router should make routing-decisions based on a single destination-address in the header of a packet, this can be done in two ways. First, *static* routing tables can be loaded into the routers and only rarely changed. The second method is more sophisticated and involves *adaptive* routing. Whenever a packet arrives at a router, the best next-hop is adaptively chosen based on, e.g., congestion statistics. Adaptive routing can better utilize the resources in the network compared to static routing and source-routing even though some support for different paths to chose between can be incorporated into source-routing systems.

When a packet arrives at a router it can be stored and error-checked before trying to forwarding it. This method is called store-and-forward and might be simple to implement but requires some buffer memory and adds significant latency for each router that is passed on the path from source to destination. The alternative is to use cut-through switching, i.e., only the header of the packet, with the destination address, need to arrive before the packet can be forwarded to an output port [Kermani and Kleinrock 1979]. This means the whole packet do not have to be stored and it will only experience a low latency because the router begins to forward the packet before it is fully received. If there is no suitable output port free, the rest of the packet can be received and stored as in store-and-forward.

Wormhole routing is a kind of cut-through where it is not needed to store the whole packet in a router if an output port is busy [Dally and Seitz 1987]. Instead, a flow-control signal is propagated back to the transmitter to order it, and intermediate routers, to stop sending. The transmission of the packet is resumed when the busy port becomes free. Methods to avoid deadlocks in wormhole-routing networks (two or more "worms" block each other) have been proposed [Ni and McKinley 1993], e.g., dividing physical channels into multiple virtual-channels, each with dedicated buffers in the routers [Dally and Seitz 1987]. Virtual-channels can also be used to, among other things, decrease blocking in a network and to guarantee bandwidth to virtual-circuits [Dally 1992].

2.8 High-performance networks

There are several alternatives of general high-performance networks that can be used to connect rather powerful and possibly heterogeneous computing nodes, that might be physically separated by several tens of meters or more. Both standards and ongoing research projects exist. Often, this kind of networks are used in NOWs (Network of Workstations) or COWs (Cluster of Workstations) [Anderson et al. 1995]. Another possibility is to have a heterogeneous system of both workstations and supercomputers.

Myrinet is a switch-based solution with support for arbitrary topologies [Boden et al. 1995]. In the current version (November 1998), full-duplex 1.28 + 1.28 Gbit/s links connect switches and nodes in the selected topology. Electrical signals on the cables carry data, control-information, and flow-control for the reversed direction. Host interfaces for PCI and SBus are available, while switches with 4, 8, 12, and 16 ports exist. Myrinet is based on earlier work performed at Caltech [Seitz and Su 1993] and University of Southern California [Felderman et al. 1994]. A parallel computer architecture with an hierarchy of Myrinet switches is reported in [Boggess and Shirley 1997]. The lowest level in the hierarchy is having a switch connecting several processors on a single board. The same switch is an interface to the next level which connects several boards in a backplane. Other works related to Myrinet have been reported [Prylli and Tourancheau 1998], and includes a multicast protocol where the Myrinet network interfaces forwards multicast packets along a spanning tree [Bhoedjang et al. 1998]. More references to reports on communication systems where Myrinet is used, can be found [Bhoedjang et al. 1998B].

An interconnection system similar to Myrinet, but especially developed for embedded systems, is RACEway from Mercury Computer Systems [Kuszmaul 1995] [Einstein 1996] [Isenstein 1994] [Mercury 1998]. A RACEway system is built up with 6-port crossbar switches to get an active backplane. Several different topologies can be chosen but the typical topology is fat-tree of switches, where each switch has four children and two parents. Circuit-switching and source routing are used. Support of real-time traffic is obtained by using priorities, where a higher-priority transmission preempts a lower-priority transmission. The link bandwidth is 160 MByte/s.

Other networks for NOWs or similar systems include:

- *Nectar:* switched-based source-routing network [Arnould et al. 1989] [Steenkiste 1996]
- *Fibre Channel:* standardized network supporting different link speeds and topologies, e.g., switch-based [Anderson and Cornelius 1992] [Boisseau et al. 1994] [Emerson 1995] [Sachs and Varma 1996] [Saunders 1996]
- *HIPPI:* standardized network where switches can be used to switch point-to-point links, each with 800 Mbit/s or 1.6 Gbit/s, simplex or duplex [Saunders 1996] [Tolmie and Renwick 1993]
- *TNet:* switch-based wormhole-routing network [Horst 1995]
- *SCI:* standardized network supporting cache-coherence in different topologies of the network, e.g., ring or switch-based [Gustavson and Li 1996] [IEEE 1993]. Used in, e.g., a system from Sequent [Lovett and Clapp 1996]
- *Spider:* short-distance (few meters) switch-based network with 2 × 1 GByte/s full duplex links [Galles 1997], used in SGI's Origin computer systems [Laudon and Lenoski 1997]
- *HAL's Mercury Interconnect Architecture:* network based on crossbars with six 1.6 + 1.6 Gbyte/s full duplex ports [Weber et al. 1997]

Experiments with ATM networks in parallel and distributed computing systems have also been reported [Eicken et al. 1995].

3. Optical Interconnections in Parallel Computers

The broad field of optical interconnections not only include traditional single-channel fiber-optics. Novel technologies allow for features like multiple high-speed channels in a single fiber and 2-dimensional arrays of optical free-space channels. Some work on comparative technology studies and classifications are reviewed below. The classification made in the survey reported in Section 5 is, to some extent, influenced by work referred to here.

Some drawbacks or limitations of electrical interconnects stated when arguing for optical interconnections in parallel computing systems and similar systems are [Tooley 1996] [Yatagai et al. 1996]:

- pin bottlenecks, both on chip and board level
- clock/signal skew
- bandwidth limitation
- high power consumption
- limited fanouts

Some advantages of optical interconnects are [Caulfield 1998] [Guilfoyle et al. 1998] [Irakliotis and Mitkas 1998] [Jahns 1994] [Jahns 1998] [Lund 1997] [Ozaktas 1997B] [Stunkel 1997] [Yatagai et al. 1996]:

- large spatial and temporal bandwidth
- optical beams do not affect each other
- light is immune to electromagnetic interference
- light beams do not suffer from signal frequency-dependent attentuation
- parallel global high speed interconnects
- possibility of nonplanar interconnections
- multiplexing in multiple domains, e.g., the time and wavelength domains
- high-density interconnects where light beams can cross each other
- avoiding large electrical backplane contacts and thereby reducing chassis size
- low energy per bit and high speed-power product
- Low skew

Of course, there are arguments for not using optics too [Bohr 1998]. For example, it is argued that performance of electrical interconnections will continue to scale [Horowitz et al. 1998] and it has been demonstrated that a bit rate of 4 Gbit/s over 100 m of twisted-pair copper cable is possible [Dally et al. 1998]. Comparative studies of optical and electrical interconnection have been reported, e.g., comparing energy consumption and system speed [Feldman et al. 1987] [Tooley 1996] [Yayla et al. 1998]. The best thing, however, is not to choose either electronics or optics, but to use both technologies in their respective areas they are best suited for [Caulfield 1998].

Classification of optical interconnection systems can be done in a number of ways. Some possible groups of systems are listed below [Kurokawa et al. 1998]:

- parallel optical fiber links (e.g., fiber-ribbons and image fibers)
- free-space optical interconnects
- optical waveguide circuits

Ozaktas view alternative optical interconnection architectures for parallel computing as a tree where the top-level alternatives are two- and three-dimensional systems [Ozaktas 1997B] [Ozaktas 1997C]. Two-dimensional systems are further divided into planar free space and waveguides, while three-dimensional systems are further divided into free space and fibers or waveguides. Ozaktas then further divides three-dimensional systems and argues for such systems where devices arrayed on planes are globally connected with a regular connection pattern. The focus is obviously on building dense parallel computing systems and not distributed systems like clusters of workstations. Many references to reports on work related to free-space systems are given by Ozaktas [Ozaktas 1997] [Ozaktas 1997B]. One possible classification of free-space optical interconnection systems is [Yatagai et al. 1996]:

- stacked optics
- planar optics
- stacked and planar optics

for which examples of systems are described and referred in Section 5. More or less general discussions on the use of optical interconnections in parallel computers have also been publicized [Goodman et al. 1984] [Rudolph 1998] [Schenfeld 1995] [Schenfeld 1996].

4. System Configurations and Requirements

The optical interconnection systems surveyed are only to a little extent evaluated in terms of quantitative performance parameters such as bandwidth and latency. Instead, the evaluation has been done according to how the systems perform in different system configurations and processing modes. Of course, quantitative parameters have influence on the performance but will not be exactly quantified, just roughly commented for the different systems.

In Figure 6, a radar signal processing chain similar to those described in [Jonsson et al. 1996] [Taveniku et al. 1996] is shown together with its bandwidth demands. The chain will only figure as a sample system with different communication requirements. No details of the chain are therefore

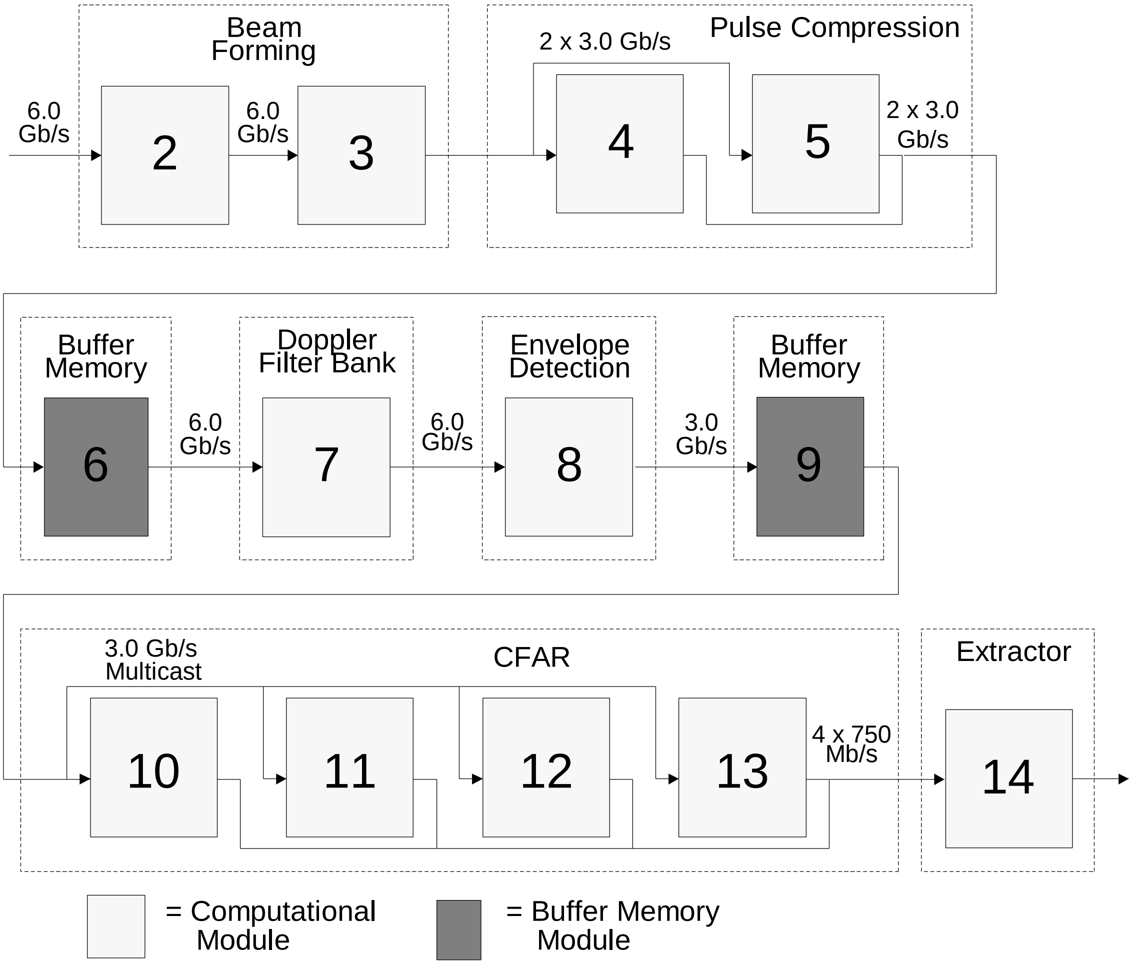

Figure 6: Data flow between the modules in the sample radar signal processing chain.

covered here. The figure shows how the work is split in a coarse grained MIMD fashion where each computational module is powerful itself, e.g., containing multiple processors. A data cube, that initially comes from the antenna, contains data in three dimensions (channel, pulse, and distance). After the processing of one stage, the new data cube is forwarded to the next node in the chain. As a pipelined system a module can start processing new data as soon as it has sent the results of the former calculation to the next node. The aggregated throughput demand is about 45 Gb/s, including the data from the antenna (Node 1) feeding the chain. As seen in the figure, the chain contains both multicast, one-to-many, and many-to-one communication patterns.

The data flow must not be disturbed by, for example, status information that the network also has to transport. A network that can guarantee delivery of semi-static high-bandwidth traffic at the same time as carrying rapidly changing control and status traffic is therefore valuable if not required.

The memory modules are used when corner-turning the data cube. A memory module stores incoming messages from the communication system in a way to finally have the whole corner-turned data cube in memory. The memory modules may not be needed if the computational modules have enough memory to store a whole data cube. Also in this case, the communication system does a main part of the corner-turning.

A number of different working modes are possible in a radar system. The task of one mode can, for example, be to scan the whole working range, while the task of another mode can be to track a certain object. Normally, the algorithm mapping and communication patterns are different for two different modes. The signal processing chain discussed above represents one mode.

In the following subsections different kinds of parallel algorithm mapping and performance criteria for use in the analysis are described.

4.1 Pure pipeline chain

The simplest case considered is a single, straight forward pipeline chain. Such a system is shown in Figure 7, where each shaded box represents a computational module and the cubes represent the data flow. Each computational module runs one or several pipeline stages and the arcs only represent the dataflow between modules. The physical topology can, e.g., be a ring or a crossbar switch. The chain is a purified form of the chain in Figure 6, without, e.g., multicasting to simplify the discussion on performance.

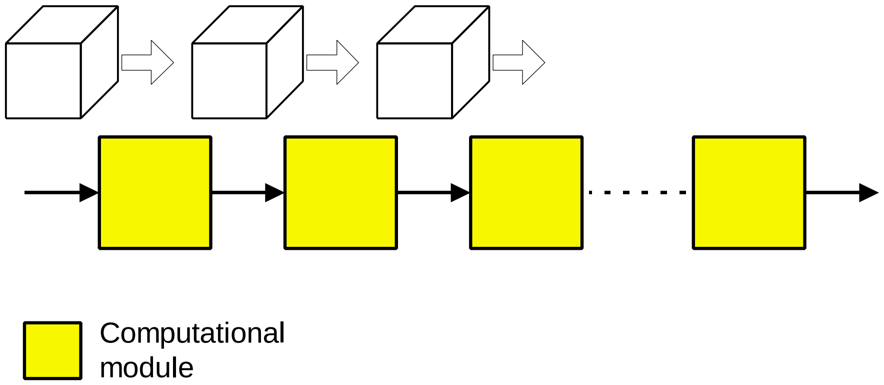

Figure 7: A pure pipeline chain where one or several stages in the chain are mapped on each module.

4.2 Same program multiple data

A common way of mapping radar signal processing algorithms onto parallel computers is to let each processor do the same operation but on a different set of data, i.e., SPMD (Same Program Multiple Data). All the PEs in Figure 8 then work together on one of the pipeline stages in Figure 6 at the time. After the processing of one stage, the data cube is redistributed if needed.

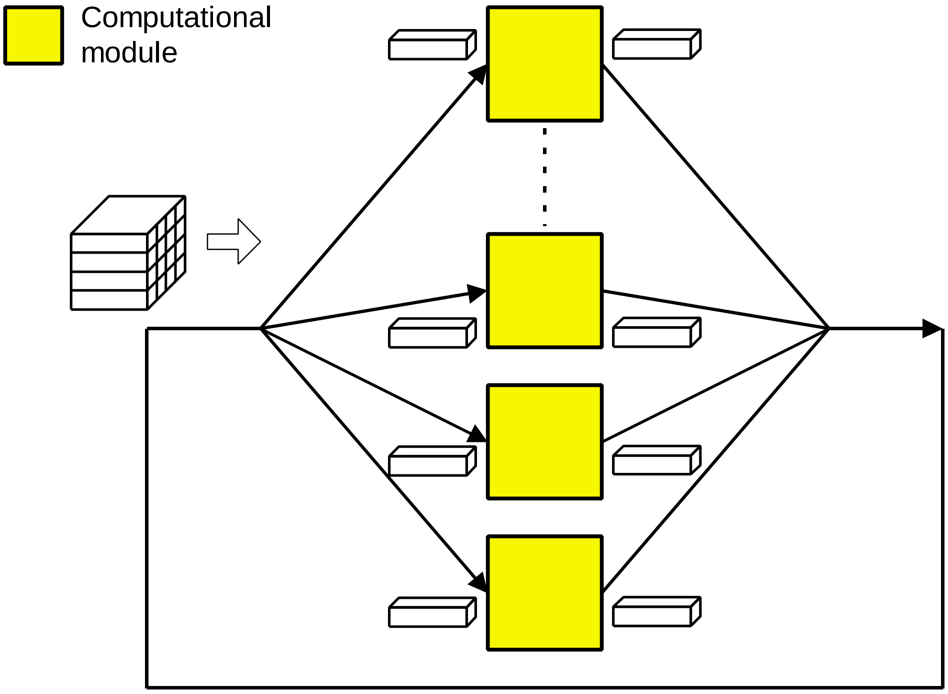

Figure 8: If the SPMD model is used, all PEs work together on one stage in the signal processing chain at the time.

Since this is not a pure pipelined system it might be harder to overlap communication and computation.

Although not considered here, there might occur communication during the processing of one stage, e.g., nearest neighbor communication if a static topology like the 2-dimensional mesh is used. The number of times it is needed to redistribute the data cube between the PEs might however be reduced compared to the number of times the data cube must be transferred to next module for processing in a pipeline chain. Instead, it might be harder to overlap communication and computation, which implies extra bandwidth requirements [Teitelbaum 1998]. When corner turning the data cube, one half of the data cube has to be transferred across the network bisection [Teitelbaum 1998]. The bisection bandwidth is therefore an essential performance parameter when using this strategy.

Incoming data from the antenna is not shown in Figure 8, but is assumed to be fed into the PEs not effecting the communication pattern shown in the figure. For example, special I/O channels can be used instead of the communication system carrying the traffic indicated by the arcs in the figure.

4.3 Parallel hardware for multiple modes

When having several concurrent operation modes of the radar instead of one, these can be mapped in a number of different ways. One way is to have several groups of processing elements, each group living its own life and being dedicated for one mode. Assuming one of the two kinds of process mapping described in Subsections 4.1 and 4.2, we have two ways of mapping concurrent modes totally in parallel.

Figure 9 shows a system of parallel pipeline chains, each chain with it's own computational modules. The incoming data from the antenna is multicasted to the first node in each chain, i.e., all modes operates on all data cubes (instead of operating in an interleaved way which is also possible).

In Figure 10, a system with parallel groups of PEs is shown. Each group executes an SPMD program which corresponding to the dedicated working mode. In this case also, some form of multicasting to spread the incoming data is assumed.

4.4 Concurrent modes on same hardware

The system configurations where several modes run concurrently on the same set of nodes puts demand on the network to be reconfigurable. Two different ways of mapping are considered here and described below.

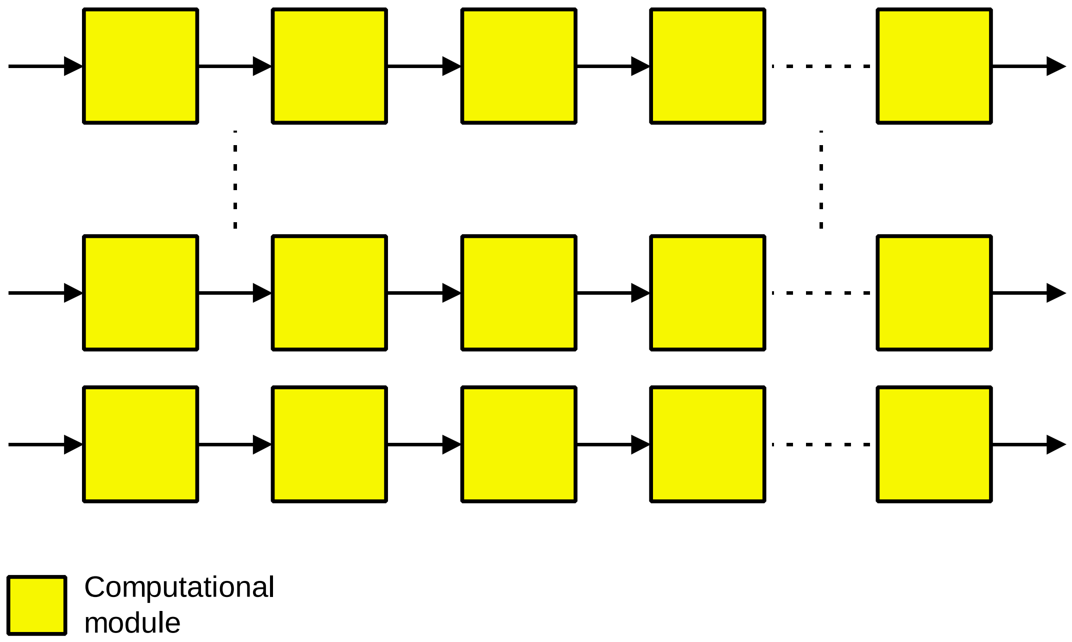

Figure 9 : Several parallel groups of PEs, each running a pipeline-chain program.

The first way is to have concurrent modes where all nodes switch mode at each new data batch. This means that the signal processing chain for each mode is only fed with one of N data batches, where N is the number of modes. Each node switch mode of operation in the order of once every 10 ms. It should hence be possible to reconfigure the network in a few hundred microseconds since reconfiguration might be needed for each mode change. The reason is that the communication pattern can differ from mode to mode. As an example, a node running several pipeline chains like the one shown in Figure 7 (but not pure pipeline) concurrently, might perform a multicast in one mode and a one-to-many or a one-to-one in the next mode. A similar reasoning can be done about the system shown in Figure 8 that then runs several concurrent SPMD programs.

The second way to have concurrent modes on the same hardware is to let each node switch mode several times per data batch. In this way the signal processing chain for each mode is fed with all data batches. Of course, this requires, in the general case, more processing power. Also, which is of more interest here, reconfiguration must be done faster. Depending on the number of working modes and how the algorithms are mapped onto the computational modules, reconfiguring the network in a few tens of microseconds or less might be desirable.

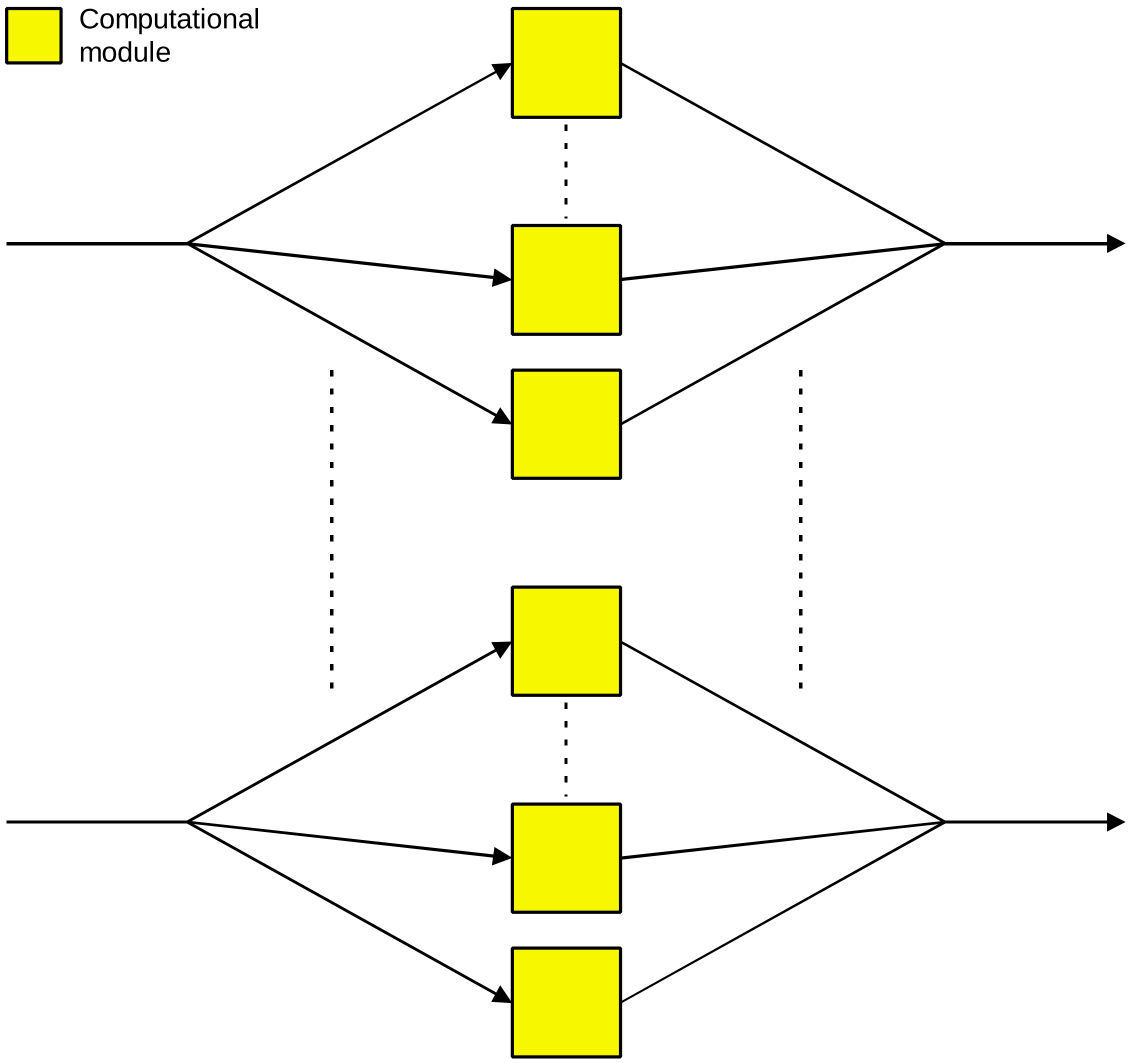

Figure 10: Several parallel groups of PEs, each running an SPMD program.

4.5 One data cube per processing element

One can let the incoming data be distributed so each data cube is given to a different processor, as shown in Figure 11. All calculations in the whole signal processing chain is then performed by the same processor for each data cube. Therefore, no communication is needed before the results are gathered. Although this way of mapping minimizes communication it will typically not be a good solution. The reason is that the computational latency will be too large. For example, consider a system where the data is distributed to 100 processors, one data cube to each processor at a time. If 100 percent utilization is assumed and the CPI (Coherent Processing Interval, i.e., the time between the start of two subsequent data cubes) is 10 ms, then the computational latency will be 1 s which is not acceptable.

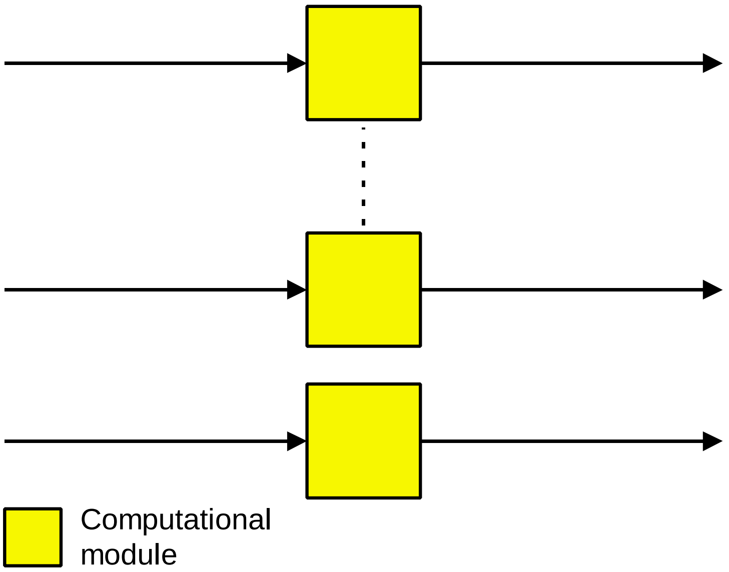

Figure 11: Several PEs in parallel where each data cube is processed by a single PE.

Hence, this kind of mapping will not be further investigated. More reasons why this is not a good choice of mapping are found in [Liu and Prasanna 1998].

4.6 Traffic types

For many of the cases described above, circuit switching might work well for the data flow in the signal processing chain. The reason is that the mode is not changed so often, e.g., not for every single data cube. However, packet switching is needed to carry short messages like control and status messages. There is no time to setup a circuit for this kind of message and always having circuits connected is not viable for sporadic traffic. Nevertheless, packet-switching can be supported by a totally different subsystem, e.g., having an optical subsystem for circuit switching and an electrical subsystem for packet switching.

4.7 System sizes and communication distances

Different communication systems may fit for different ranges of system sizes or communication distances. An interconnection network might, for example, not offer sufficient throughput over a long distance. On the other hand it might be too large physically or be too expensive compared to other solutions for small sized computer systems. Evaluation is made to put qualitative judgements about the communication systems' suitability for the different system sizes and communication distances given in Table 1.

Kind of communication	**Communication distances**
Intra chip	0.1 - 2 cm
Intra MCM	1 - 10 cm
Intra board	5 - 50 cm
Inter board	0.1 - 1 m
Inter cabinet	0.5 - 10 m
Inter and intra room	10 - 100 m
Intra and inter building	100 m - 10 km

Table 1 : Classification of system sizes and communication distances.

5. Evaluation of Optical Interconnection Systems

A number of proposals of communication systems that can be suitable in a radar signal processing system is given in Subsections 5.1 through 5.9. Some of the proposals are hybrids of several other systems. Although many other optical interconnection architectures that might be candidates exist, only some selected groups, or concepts, are selected to give a reasonably broad view of possible solutions. A little bit more focus is given to pure fiber-optic solutions than, e.g., free-space systems. The survey ends with a summarizing evaluation in Subsection 5.10. Other surveys and tutorial texts in the field of optical interconnects exist [Goldberg 1997], e.g., focusing on parallel computers and ATM switches [Kurokawa and Ikegami 1996].

5.1 Fiber-ribbon ring network

If fiber-ribbon cables/links [Hahn 1995] [Karstensen et al. 1995] [Schwartz et al. 1996] [Siala et al. 1994] [Wong et al. 1995] are used to connect the nodes in a point-to-point linked ring network, bit-parallel transfer can be utilized. In such a network one of the fibers in each ribbon is dedicated to carry the clock signal. Therefore, no clock-recovery circuits are needed in the receivers. Other fibers can be utilized for, e.g., frame synchronization.

As seen in Figure 12, aggregated throughputs higher than 1 can be obtained in ring networks with support for spatial bandwidth reuse (sometimes

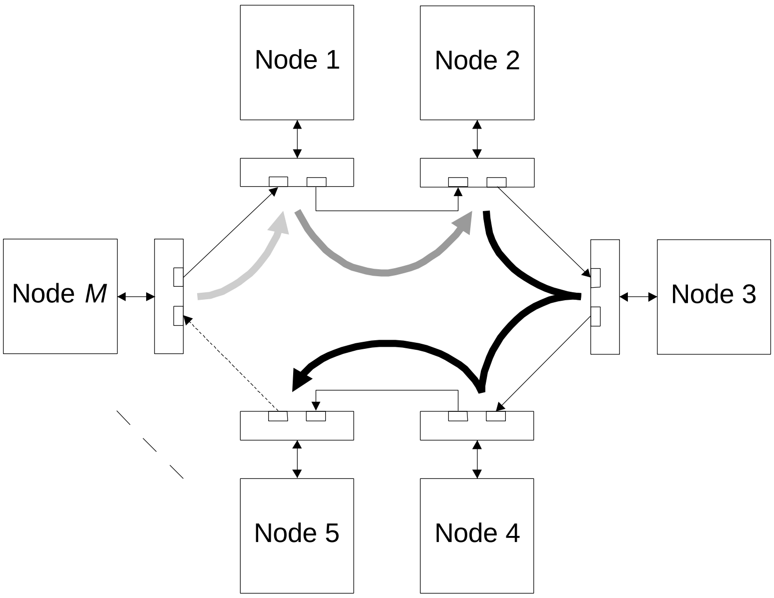

Figure 12: Example of spatial bandwidth reuse. Node M sends to Node 1 at the same time as Node 1 sends to Node 2, and Node 2 sends a multicast packet to Node 3, 4, and 5.

called pipeline rings) [Wong and Yum 1994]. This feature can be highly used in signal-processing applications with a pipelined dataflow, i.e., most of the communication is to the nearest down-stream neighbor. Two fiber-ribbon pipeline ring networks have been reported recently [Jonsson 1998B]. The first network has support for circuit-switching on 8+1 fibers (data and clock) and packet-switching on an additional fiber [Jonsson et al. 1997B]. The second network is more flexible, but with a little more complexity, and has support for packet-switching on 8+1 fibers and uses a tenth fiber for control packets (see Figure 13) [Jonsson 1998]. The control packets carries MAC (Medium Access Control) information for the collision-less MAC protocol with support for slot-reserving. Slot-reserving can be used to get RTVCs (Real-Time Virtual Channels) [Arvind et al. 1991] for which guaranteed bandwidth and a worst-case latency are specified (compare to circuit-switching).

Another fiber-ribbon ring network is the PONI network (formerly USC POLO), which is proposed to be used in COWs (Cluster of Workstations) and similar systems [Raghavan et al. 1999] [Sano and Levi 1998]. Integrated circuits have been developed for the network and tests have been performed [Sano et al. 1996] [USC 1997].

5.2 WDM star network

A passive fiber-optic star distributes all incoming light, on the input ports, to all output ports. A network with the logical function of a bus is obtained

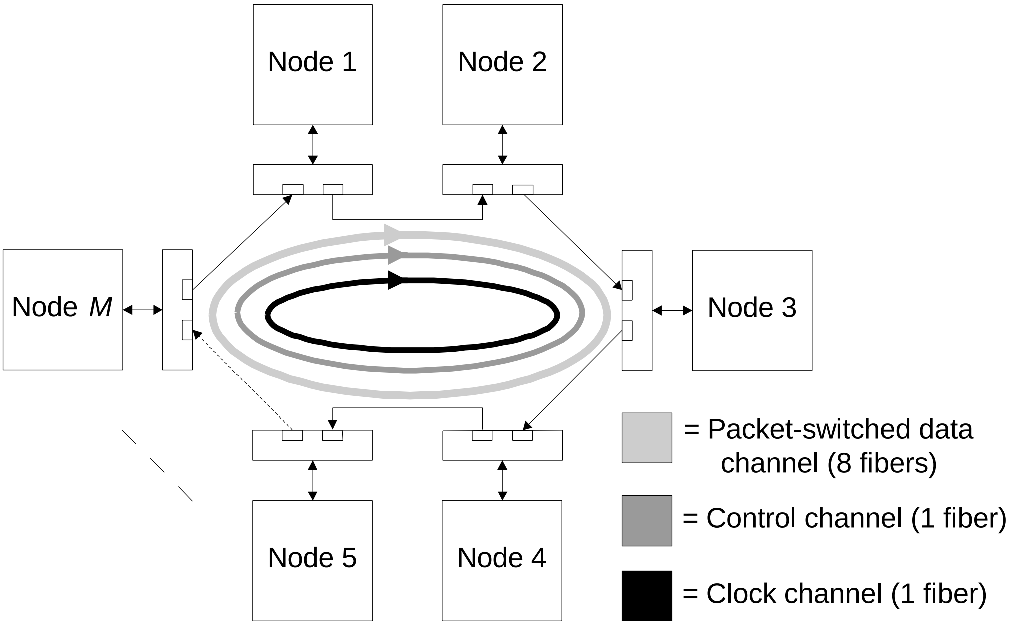
Node 1
Node 2
Node *M*
Node 3
Node 5
Node 4
= Packet-switched data channel (8 fibers)
= Control channel (1 fiber)
= Clock channel (1 fiber)

Figure 13: Control-channel based network built up with fiber-ribbon point-to-point links.

when connecting the transmitting and receiving side of each node to one input and output fiber of the star, respectively. By using WDM (Wavelength Division Multiplexing), multiple wavelength channels can carry data simultaneously in the network [Brackett 1990]. In other words, each channel has a specific color of light. To get a flexible WDMA (Wavelength Division Multiple Access) network, we need tunable receivers and/or transmitters, i.e., it should be possible to send/listen on an arbitrary channel [Mukherjee 1992].

In Figure 14, an example of a WDM star network configuration is shown. Each node transmits on a wavelength unique to the node, while the receiver can listen to an arbitrary wavelength. The configuration is used in the TD-TWDMA network (Time-Deterministic Time and Wavelength Division Multiple Access) [Jonsson et al. 1996], which has support for guaranteeing real-time services, both in single-star networks [Jonsson et al. 1997] and in star-of-stars networks [Jonsson and Svensson 1997]. One can say that this kind of network architectures implements a distributed crossbar. The flexibility is hence high, and multicast and single-destination traffic can co-exist. The number of wavelengths is practically limited to 16-32 [Brackett 1996], but, as stated above, hierarchical networks with wavelength reuse can be built.

Tunable components (e.g., filters) with tuning latencies in the order of a nanosecond have been reported, but they often have a limited tuning range [Kobrinski et al. 1988]. At the expense of longer tuning latencies, however, components with a broader tuning range exist [Cheung 1990]. Such

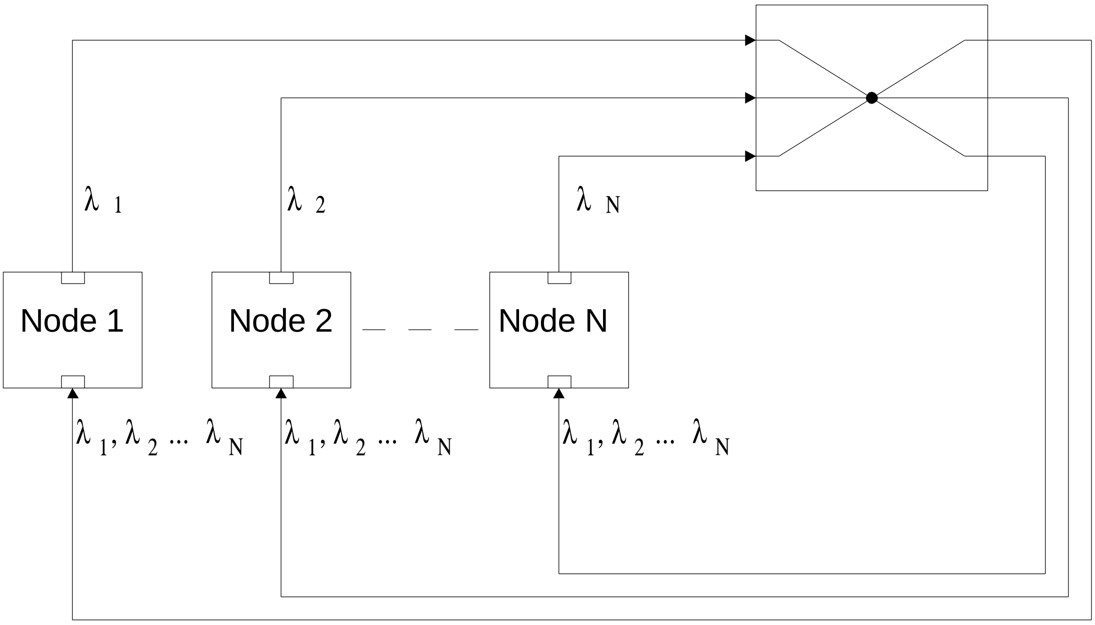

Figure 14: WDM star network.

components can be used to get a cheaper network for systems where much of the communication patterns remain constant for a longer period, e.g., during the processing of a data cube in a radar system with a pipelined mapping of the processing stages. To support for some more rapidly changing traffic patterns, the nodes can be extended with transmitters and receivers fixed-tuned to a special broadcast channel. This configuration can be compared to having support for both circuit-switching and packet-switching.

An example, based on the signal processing chain shown in Figure 6, is given below to show that only one input and one output channel (in addition to the broadcast channel) are needed during the processing of a data cube, i.e., the normal minimum time running one pipeline chain (working mode of the radar). In Figure 6, there are 13 nodes. In addition, the antenna is seen as one node (feeds the first node in the chain with data), Node 1, and there is one master node responsible for supervising the whole system and interacting with the user, Node 15. For simplicity we assume an efficient channel bandwidth of 6.0 Gb/s. A feasible allocation scheme of eight channels (in addition to the broadcast channel) is shown in Table 2.

Totally removing the ability of tuning in a WDM star network leads to a multi-hop network [Mukherjee 1992B]. Each node in a multi-hop network transmits and receives on one or a few dedicated wavelengths. If a node does not have the capability of sending on one of the receiver wavelengths of the destination node, the traffic must pass one or several intermediate

Node	Input channel	Output channel
1	–	λ_1
2	λ_1	λ_2
3	λ_2	λ_3
4	λ_3	λ_4
5	λ_3	λ_4
6	λ_4	λ_5
7	λ_5	λ_6
8	λ_6	λ_7
9	λ_7	λ_7
10	λ_7	λ_8
11	λ_7	λ_8
12	λ_7	λ_8
13	λ_7	λ_8
14	λ_8	λ_8
15	λ_8	–

Table 2 : A feasible allocation scheme of the wavelength channels in a WDM star network for the sample radar system.

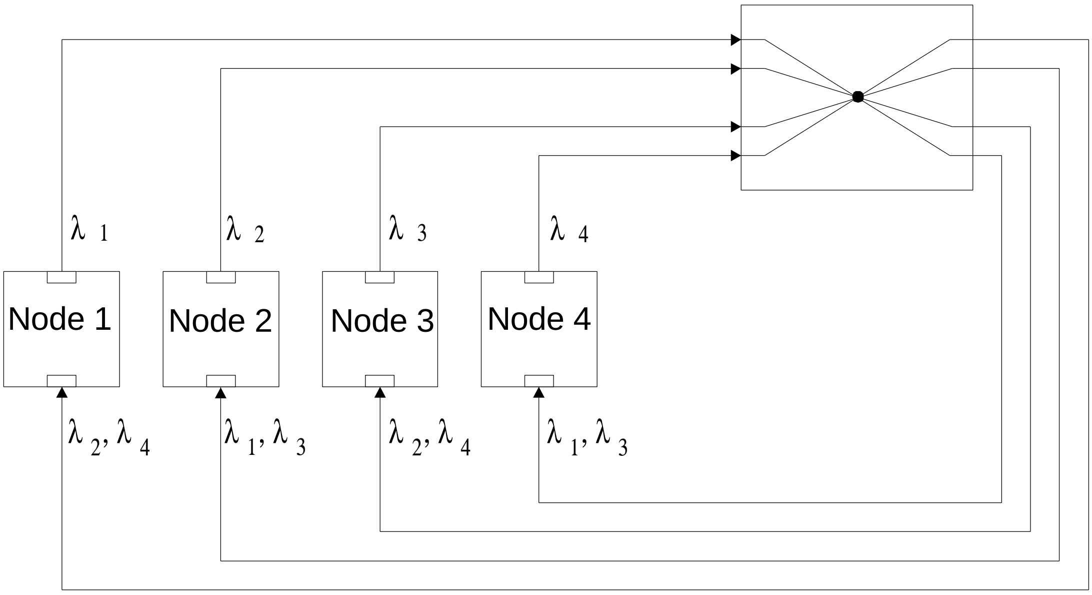

Figure 15: WDM star multi-hop network.

nodes. The wavelengths can be chosen to get, e.g., a perfect-shuffle network [Acampora and Karol 1989], or a network with a pattern similar to the dataflow in pipelined system. One can also choose to have a network with several topologies embedded, e.g., a ring and a hypercube. An example of a multi-hop network is shown in Figure 15. This configuration of wavelength assignments corresponds to the topology shown in Figure 16. Dynamic real-time scheduling can be done in a multi-hop network [Yu and Bhattacharya 1997]. The method works like static scheduling where a central node runs the scheduling algorithm, but where high-priority messages might preempt low-priority messages. The highest priority level is used for messages with hard deadlines while the other levels are used for messages with soft deadlines. Lower priority levels are used for less important messages.

5.3 WDM ring network

A WDM ring network utilizes ADMs (Add-Drop Multiplexers) in all nodes to insert, listen, and remove wavelength channels from the ring. In the WDMA ring network described in [Irshid and Kavehrad 1992], each node is assigned a node-unique wavelength to transmit on. The other nodes can then tune in an arbitrary channel to listen on. This configuration is logically the same as that for the WDM star network with fixed transmitters and tunable receivers. The distributed crossbar again gives good performance for general communication patterns, such as in a radar system with the SPMD model.

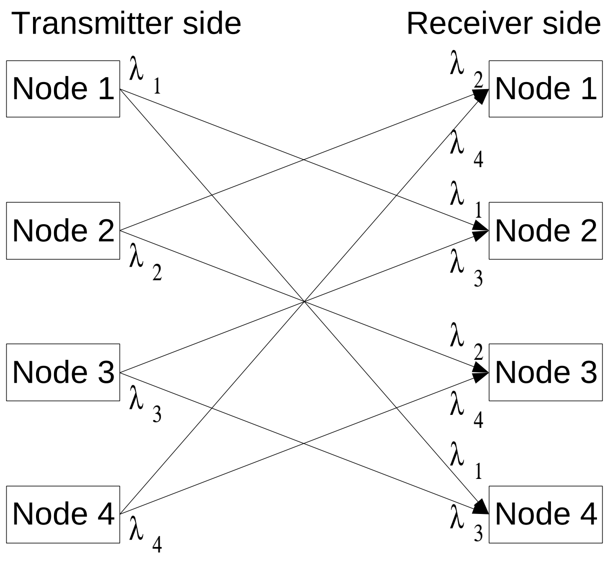

Figure 16: Multi-hop topology.

Spatial wavelength reuse can be achieved by removing the transmitted light at the destination node (last destination node for multicast). At high degrees of nearest-downstream-neighbor communication, as in systems with a pipelined mapping, throughputs significantly higher than 1 can be achieved for a single wavelength. This leads to a less number of wavelength channels needed.

As discussed for the WDM star network, components with long tuning latencies (ADMs in the case of a ring) can be used for traffic patterns that do not change rapidly. An additional broadcast wavelength dedicated for packet-switching keeps the network flexible. A single 6 Gbit/s channel (in addition to the broadcast channel) is enough for the signal processing chain shown in Figure 6, if the ADMs in Node 1, 2, 3, 6, 7, 8, and 9 terminates the channel for wavelength reuse.

5.4 Integrated fiber and waveguide solutions

Fibers or other kinds of waveguides (hereafter commonly denoted as channels) can be integrated to form a more or less compact system of channels. Fibers can be laminated to form a foil of channels, for use as intra-PCB (Printed Circuit Board) or back-plane interconnection systems [Eriksen et al. 1995] [Robertsson et al. 1995] [Shahid and Holland 1996]. Fiber-ribbon connectors are applied to fiber end-points of the foil. An example is shown in Figure 17, where four computational nodes are connected in a ring topology. In addition, there is a clock node that distributes clock signals to the four computational nodes via equal-length fibers to keep the clock signals in phase. In other words, a fiber-optic clock

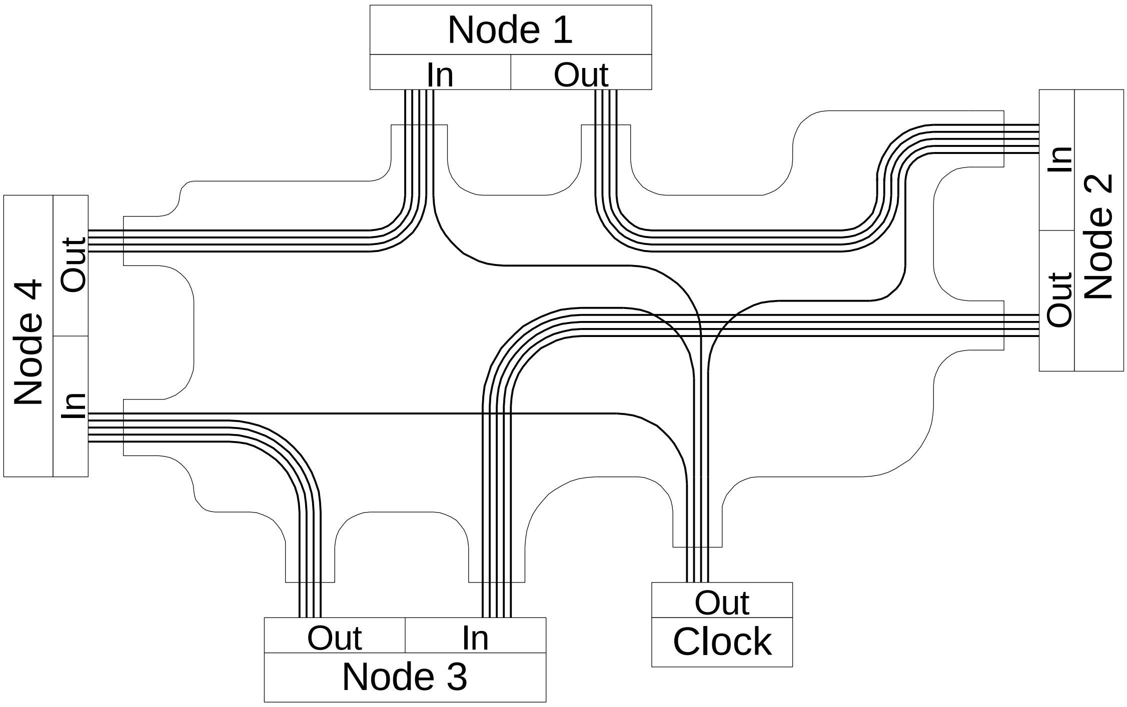

Figure 17: A foil of fibers connects four computational nodes. In addition, a clock node distributes clock signals to the computational nodes.

distribution network [Kiefer and Swanson 1995] and a data network are integrated into one system. If one foil is placed on each PCB in a rack, they can be passively connected to each other via fiber-ribbon cables. Using polymer waveguides instead of fibers brings advantages such as the possibility of integrating splitters and combiners into the foil, and the potential for more cost-effective mass-production [Eriksen et al. 1995].

Integrated systems of channels can be setup and used in a number of configurations, for which some are discussed below. One way is to embed a ring with bit-parallel transmission and the possibility of spatial bandwidth reuse as described in Subsection 5.1. Of course, this leads to the same good performance for pipelined data flows as the fiber-ribbon ring network, only changing the medium to a more compact form. Besides pure communication purposes, channels for, e.g., clock distribution (as seen in the example) and flow control can be integrated into the same system.

Another way is to follow the proposed use of an array of passive optical stars to connect processor boards in a multiprocessor system, via fiber-ribbon links, for which experiments with 6 x 700 Mb/s fiber-ribbon links were done (see Figure 18) [Parker et al. 1992]. As indicated above, such a configuration can be integrated by the use of polymer waveguides. The power budget can, however, be a limiting factor to the number of nodes

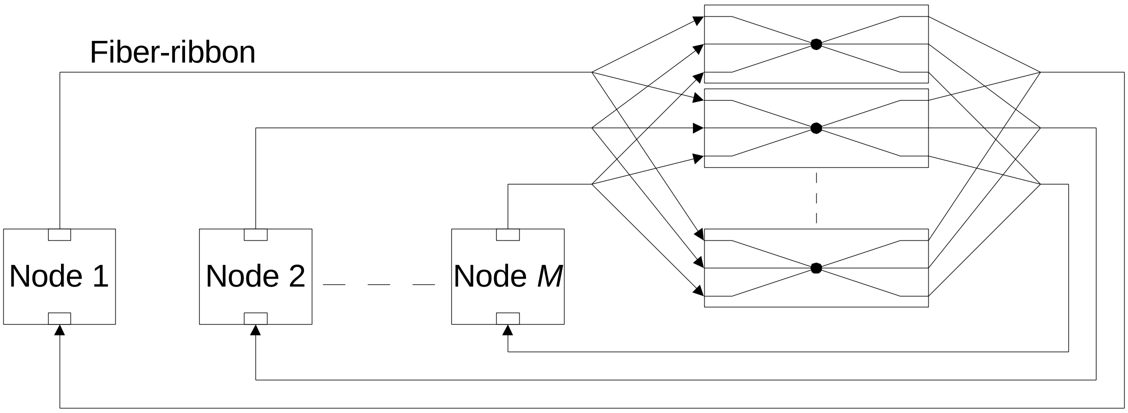

Figure 18: Array of passive optical stars connects a number of nodes via fiber-ribbon cables.

and/or the distance. Advantages are simple hardware due to bit-parallel transmission (like other fiber-ribbon solutions), and the broadcast nature. Many-to-many communication patterns, as used when corner-turning in SPMD mode, map easily on the broadcast architecture as long as the star-array not becomes a bottle-neck. In a similar system the star array is exchanged by a chip (with optoelectronics) that has one incoming ribbon from each node and one output ribbon [Lukowicz et al. 1998]. The output ribbon is coupled to an array of $1 \times N$ couplers so each node has a ribbon connected to its receiver. The chip couples the incoming traffic together in a way that simulates a bus. At contention, the chip can temporarily store packets.

Electronic crossbars can be distributed on the PCBs and/or placed on a special switch-PCB in a back-plane system, and be connected by integrated parallel channels. A distributed crossbar, instead, can be realized by bit-serial transmissions over a fully connected topology, i.e., there is one channel between each pair of nodes (see Figure 19) [Li et al. 1998C]. Broadcast is done by driving all the laser-diodes of a node with the same bit-stream. A simple solution is to always drive all the laser-diodes in the transmitting node and couple the photo-diodes together as one incoming channel. This brings an architecture with the support of a single broadcast at the time. By only sending on one fiber when performing single-destination communication will, however, give the opportunity of having multiple transmissions in the system at the same time. The number of nodes, N, in this configuration is limited because the number of fibers grows by N^2. Also, clock-recovery circuits must be involved. After all, the distressed power budget brings an advantage over a system with passive splitters or stars (this holds for, e.g., the point-to-point connected ring also). The flexibility of a crossbar makes the network good for radar systems with the SPMD mode.

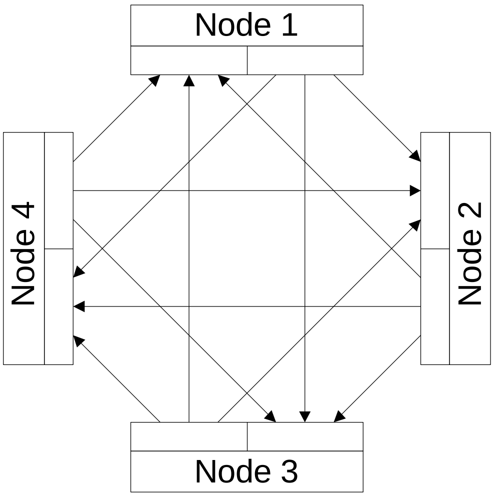

Figure 19: Fully connected topology where each node has an array of N - 1 laser diodes and an array of N - 1 photo diodes, where N is the number of nodes.

Other similar systems include the integration of fibers into a PCB for the purpose of clock distribution [Li et al. 1998]. Distribution to up to 128 nodes was demonstrated. The fibers are laminated on one side of the PCB while integrated circuits are placed on the reversed side. The end section of each fiber is bent 90 degrees to lead the light through a via hole to the reversed side of the PCB.

5.5 Optical interconnections and electronic crossbars

Communication systems like, e.g., Myrinet [Boden et al. 1995], where arbitrary switched topologies can be built, can support a number of different traffic patterns possible in radar systems. Fiber-ribbons can be used to increase bandwidth while still sending in bit-parallel mode. Bit rates in the order of 1 Gbit/s over each fiber in the ribbon is possible over tens of meters using standard fiber-ribbons. As noted in Section 5.4, foils of fibers or waveguides (e.g., arranged as ribbons) can be used to interconnect nodes and crossbars on the PCB and/or back-plane level.

An alternative to ribbons is bit-parallel transfer over a single fiber by using WDM. In such configurations each bit in the word, plus the clock signal, is given a dedicated wavelength. Wavelengths (or fibers in a fiber-ribbon cable) can also be dedicated to other purposes like frame synchronization and flow control. Significantly higher bandwidth-distance products can be achieved when using bit-parallel WDM over dispersion-shifted fiber, instead of fiber-ribbons [Bergman et al. 1998] [Bergman et al. 1998B]. If, however, only communication over shorter distances exist (e.g., a few meters), the bandwidth-distance product is not necessarily a limiting factor.

Transmission experiments with an array of eight pie-shaped VCSELs arranged in a circular area with a diameter of 60 um, to match the core of a multimode fiber, has been reported [Coldren et al. 1998]. Other works on the integration of components for short-distance (non telecom) WDM links have been reported, e.g., a 4 × 2.5 Gbit/s transceiver with integrated splitter, combiner, filters, etc. [Aronson et al. 1998].

The switch itself can also be modified to increase performance or packing density. A single-chip switch core where fiber-ribbons is coupled directly to optoelectronic devices on the chip is possible [Szymanski et al. 1998]. Attaching 32 incoming and 32 outgoing fiber-ribbons with 800 Mb/s per fiber translates to an aggregated bandwidth of 204 Gbit/s through the switch when eight fibers on each link are used for data.

A 16×16 crossbar switch-chip, with integrated optoelectronic I/O, have been implemented for switching of packets transferred using bit-parallel WDM [Krisnamoorthy et al. 1996]. Each node has two single-mode fibers coupled to the switch, one for input and one for output.

Another switch chip with integrated optoelectronic I/O is intended to be attached to each node in a static topology like a multidimensional torus [Pinkston et al. 1998]. A special feature of the chips is that potential deadlocks is handled by a global mechanism. The mechanism brings mutual exclusion to let one packet at a time use dedicated hardware to recover from a potential deadlock.

Finally, it can be noted that both Myricom (Myrinet) and Mercury (RACEway) are looking at optical technology for their future products [Lund 1997].

5.6 Systems with smart-pixel arrays

In smart-pixel based systems the interconnection network can normally not be seen as a stand-alone subsystem. Instead, processors and optoelectronic devices for communication are integrated on the same substrate [Neff et al. 1996]. Typically, smart-pixels are organized in a 2-dimensional array (e.g., on a chip) where each smart-pixel consists of a processor, a laser diode, and a photo diode. Similar configurations exist, for example, where incoming light is modulated by a modulator in the smart-pixel.

Several smart-pixel arrays can be arranged in a row where data is transferred stage by stage (see Figure 20). In between the arrays, holographic interconnects or other optics might be used to steer or switch the optical channels. The row arrangement is especially suitable in applications where computations can be mapped in a pipeline fashion with one pipeline stage per array, e.g., image processing [McArdle et al. 1996]

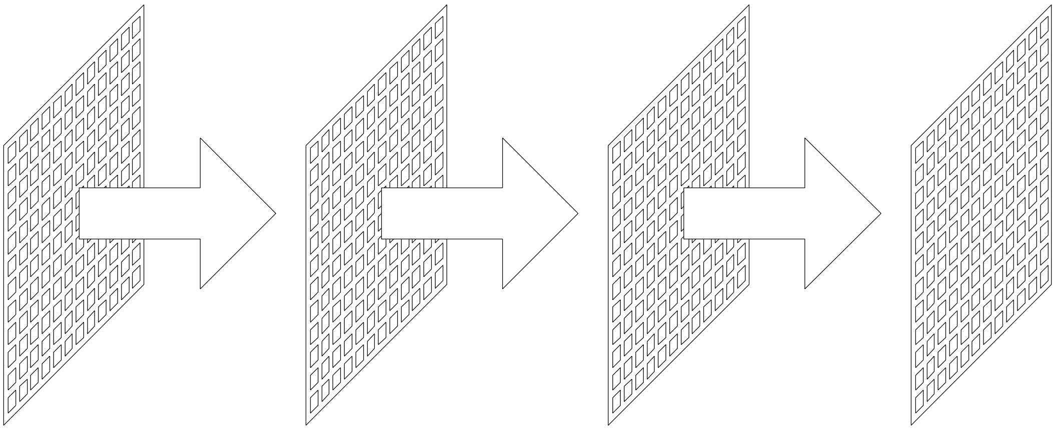

Figure 20: Row of smart-pixel arrays.

[McArdle et al. 1997] [Ishikawa and McArdle 1998], sorting [Gourlay et al. 1998], and applications using FFT [Betzos and Mitkas 1998]. The row of smart-pixel arrays can be placed on a PCB with the bottom side of each array electrically connected to the PCB [Neff 1994].

A system where smart-pixel arrays are connected in a ring has been reported [Chen et al. 1998B] [Chen et al. 1998]. Each array operates in SIMD (Single Instruction stream Multiple Data streams) mode on 2-dimensional data. A modified CSMA/CD (Carrier Sense Multiple Access, Collision Detect) protocol is used for arbitration in the ring where some of the pixels in each array are used for data and some for addressing and clocking.

In [Kurokawa et al. 1998], an estimation of maximum number of pixels and I/O throughput allowed on a chip is reported for an VCSEL-based smart-pixel array. Power consumption and pixel homogeneity (variation of the VCSEL threshold current) are considered and compact 150 MHz processor elements each with 200 gates are assumed. With a threshold current variation of 10 % a single chip with more than 1000 processor elements (smart pixels) is possible, each processor with an 300 Mb/s I/O port. The estimated maximum total I/O throughput for such a system is 600 Gb/s. VCSELs are better than edge-emitting laser diodes for free-space communication mainly because of the possibility of 2-dimensional VCSEL arrays, and because integration with electronic circuits is simpler [Kurokawa et al. 1998].

5.7 Optical and optoelectronic switch-fabrics

The architecture with optical interconnections and electronic crossbars is flexible and powerful. Optics and optoelectronics can, however, be used internally in a switch fabric also, i.e., more than just in the I/O interface. A broad spectra of solutions has been proposed for which some examples are given below.

SDM (Space Division Multiplexing) switches [Kato et al. 1998] [Guilfoyle et al. 1998] [Sawchuk et al. 1987] and WDM switches (consisting of, e.g., wavelength converters and wavelength selective components) [Pedersen et al. 1998] [Flipse 1998] can be used both as stand-alone switches and as building components in bigger switch fabrics [Reif and Yoshida 1994]. As an example, a Banyan multistage network built of 2 x 2 switch elements has been described [Chamberlain et al. 1998]. Another multistage network uses both WDM and SDM switches, but in different stages [Kawai et al. 1995]. A multistage network can also be implemented by having chips with processing elements placed on a 2-dimensional plane [Christensen and Haney 1997]. The processors then communicate with each other by using a mirror bouncing back the optical signal to the plane but to another processor. Switching is made on the chips while each pass between two switch stages corresponds to a bounce on the mirror.

A multistage switch incorporating both electrical and optical switching, but in different stages, has also been reported [Duan and Wilmsen 1998]. Some work has been focused on the communication between stages, e.g., perfect shuffle with lenses and prisms [Lohmann et al. 1986]. Switch times for SDM switches in the order of 1 ns have been reported [Kato et al. 1998], while some SDM switches have switch times in the order of 1 ms [Tajima et al. 1998]. A switch can, e.g., be placed on a dedicated board in a cabinet and be connected to processor boards via fibers or an optical backplane [Maeno et al. 1997].

A system that implements a distributed crossbar, or a fully connected system, connecting N nodes with only passive optics in between the transmitters and receivers has been demonstrated [Li et al. 1998B]. All optical channels turned-on from a transmitter's 2-dimensional $\sqrt{N} \times \sqrt{N}$ VCSEL (Vertical Cavity Surface Emitting Laser) array are inserted into a fiber image guide. The fiber image guides from all transmitters end at a central free-space system with lenses. The lenses are arranged in a way that each VCSEL pixel in a VCSEL array is focused on a single spot together with the corresponding pixels in all other arrays. In this way, there are N spots for which each is focused into a single fiber leading to a receiver. Hence, selecting a pixel in a VCSEL array to turn on corresponds to addressing a destination node.

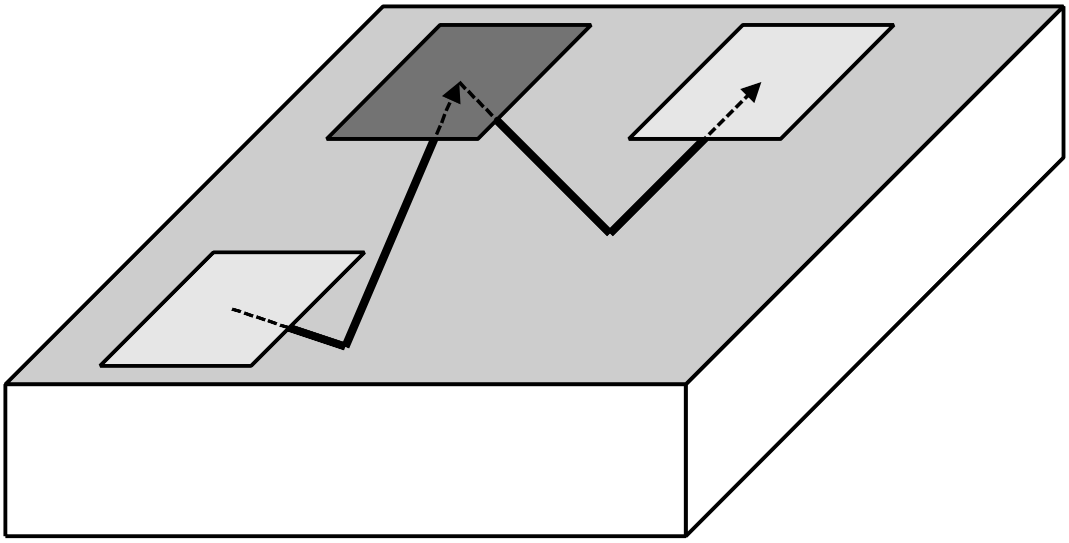

Figure 21: Example of a planar free-space system. The direction of the beam is steered, by the optical element, on the way between two chips.

5.8 Planar free-space optics

By placing electronic chips (including optoelectronic devices) and optical elements on a substrate where light-beams can travel, we get a planar free-space system [Jahns 1994] [Jahns 1998] [Sinzinger 1998]. Electronic chips are placed in a 2-dimensional plane, while light-beams travel in a 3-dimensional space. In this way, optical systems can be integrated monolithically, which brings compact, stable, and potentially inexpensive systems [Jahns 1998].

The interconnection pattern in a planar free-space system can, for example, be chosen with respect to a pipelined dataflow between chips. Another possibility is to have a more general topology, like the 2-dimensional mesh, or to have special optical or optoelectronic devices dedicated for switch functions. The latter configurations can be the right choice if the SPMD program model is used.

5.9 Free-space optical backplanes

Several different optical backplanes have been proposed, for which three different types are discussed below. As shown in Figure 22a, using planar free-space optics is one way to transport optical signals between PCBs (Printed Circuit Boards). Holographic gratings can be used to insert and extract the optical signals to/from the waveguide which might be a glass substrate [Zhao et al. 1995]. Several beams, or bus lines, can be used, i.e., each arrow in the figure represents several parallel beams [Zhao et al.1996].

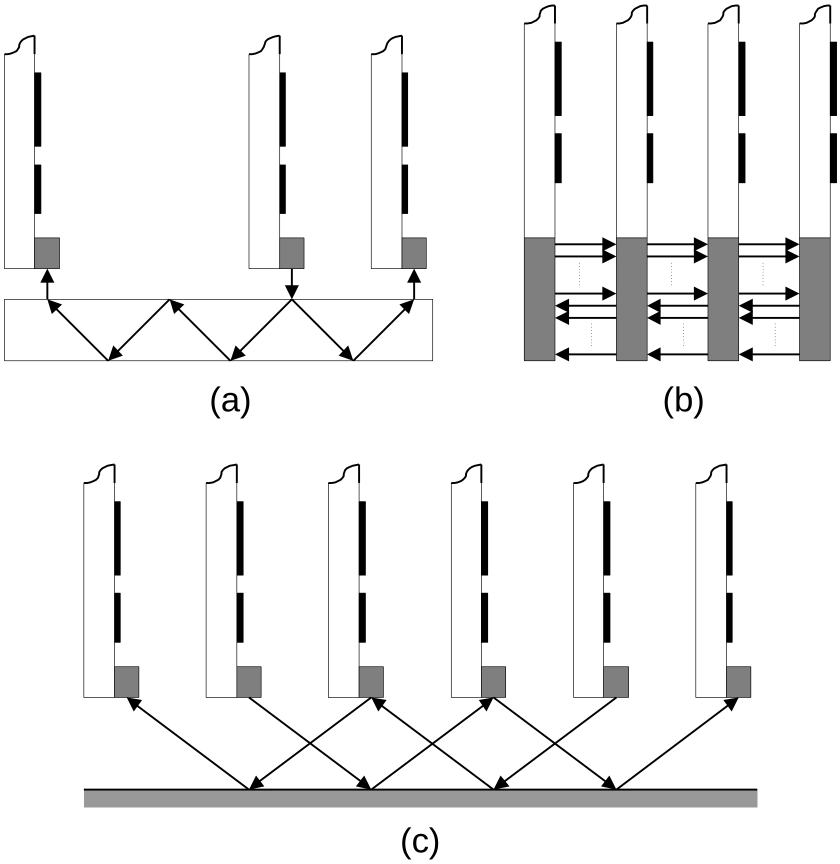

Figure 22: Optical backplane configurations: (a) with planar free-space optics, (b) with smart-pixel arrays, and (c) with a mirror.

In the system shown in Figure 22b, 2-dimensional arrays of optical beams (typically 10 000) link neighboring PCBs together in a point-to-point fashion [Szymanski 1995] [Hinton and Szymanski 1995]. Smart-pixel arrays then act as intelligent routers that can, e.g., bypass data or perform data extraction operations where some data is passes to the local PCB and some data is retransmitted to the next PCB [Supmonchai and Szymanski 1998]. Each smart-pixel array can typically contain 1 000 smart-pixels arranged in a 2-dimensional array, where each pixel has a receiver, a transmitter, and a simple processing unit. One way of configuring the system is to connect the smart-pixel arrays in a ring, but where the ring can be reconfigured to embed other topologies [Szymanski and Hinton 1995] [Szymanski and Supmonchai 1996].

The configuration shown in Figure 22c is similar to the optical backplane based on planar free-space interconnects. The difference is the replacement of the waveguide by a mirror [Hirabayashi et al. 1998]. An optical beam leaving a transmitter is simply bounced once on the mirror before arriving at the receiver. Regeneration of the optical signal (multi-hop) might be needed on the way from source to final destination.

Network	**Pipeline**	**SPMD**	**Notes**
Fiber-ribbon pipeline ring	Good	Moderate	Good for SPMD too if enough bandwidth
WDM star, rapid tuning	Good	Good	Flexible passive network
WDM star, slow tuning and broadcast channel	Good	Poor	WDM star alternative that might be cheaper
Multi-hop WDM star	Moderate	Moderate	Can be optimized for pipelined mapping
WDM ring with rapid tuning	Good	Good	More channels might be needed for SPMD
WDM ring, slow tuning and broadcast channel	Good	Poor	WDM ring alternative that might be cheaper
Fiber ribbons and array of stars	Moderate	Moderate	Power and bandwidth limited
Fully connected topology with broadcast driving	Moderate	Moderate	Bandwidth limited. Grows with N^2
Fully connected topology with flexible driving	Good	Good	Grows with N^2
Optical fibers and electronic crossbars	Good	Good	Optolectronics needed in switch too

Table 3: Performance summary of some of the discussed networks with respect to pipeline and SPMD mapping.

Of the three discussed types of optical backplanes, the one with smart-pixel arrays seems to be the most powerful. On the other hand, a simple passive optical backplane might have other advantages. More optical backplanes have been proposed, e.g., a bus where optical signals can pass through transparent photo-detectors or be modulated by spatial light modulators [Hamanaka 1991].

5.10 Summary

Some of the mentioned networks that incorporates fiber-optics are summarized in Table 3, where remarks on suitability/performance for the two basic cases of mapping (pipeline and SPMD) are made. The pipeline mapping fits on a larger variety of networks because of the absence of all-to-all traffic pattern. Limiting factors on the use of SPMD mapping varies from network to network and can, e.g., be tuning speed when switching from many different sources and destinations, shared resources that becomes bottlenecks, and a topology that favors nearest-neighbor communication.

Performance estimations for the case when mapping several parallel working modes, each with either pipeline or SPMD mapping, are summarized in Table 4. The networks should essentially handle the same kind of traffic patterns as in the single-mode case. However, the incoming data from the antenna must, in the pipeline case, be multicasted to the first

Network	**Pipeline**	**SPMD**	**Notes**
Fiber-ribbon pipeline ring	Moderate	Poor	Extra bandwidth needed to distribute input data
WDM star, rapid tuning	Good	Good	Broadcast support is used
WDM star, slow tuning and broadcast channel	Good	Poor	Broadcast support is used
Multi-hop WDM star	Moderate	Moderate	Depends on which virtual topology that is chosen
WDM ring with rapid tuning	Good	Good	Broadcast support is used
WDM ring, slow tuning and broadcast channel	Good	Poor	Broadcast support is used
Fiber ribbons and array of stars	Moderate	Moderate	Broadcast support is used
Fully connected topology with broadcast driving	Moderate	Moderate	Broadcast support is used
Fully connected topology with flexible driving	Good	Good	Broadcast support is used
Optical fibers and electronic crossbars	Good	Good	Broadcast support is used

Table 4 : Performance summary of some of the discussed networks with respect to pipeline and SPMD mapping, when several concurrent modes run in parallel, each on dedicated hardware.

node in each group of nodes dedicated to a working mode. In the case of SPMD, the incoming data is distributed by multiple multicast transmissions, each carrying a subset of the data cube.

If several concurrent working-modes are time-interleaved, all nodes work together on the data cube, one mode at a time. In this way, the communication patterns change several times per CPI. The tuning latencies, in appropriate networks, must therefore be reduced proportionally to the number of working modes. For most networks this is not a problem.

Having optics inside a switch gives the same flexibility as electronic crossbars but it might be able to build larger switch fabrics with high transmission capacities using optics. The suitability of the different free-space systems for the mapping-cases discussed depends a lot on the more detailed configurations of the systems. For example, planar free-space systems can be arranged in arbitrary topologies. However, the free-space technology chosen has some influence on possibilities of topologies etc depending on, e.g., the linear order (or similar arrangements) of smart-pixel arrays and the 2-dimensional plane for which components are restricted to when using planar optics.

Suitability of different technologies/networks in different system sizes are summarized in Table 5. The lower bounds on system sizes arises from such things as miniaturization problems (e.g., fiber-ribbon connectors) and

Kind of communication	**WDM star/ring**	**Fiber-ribbon**	**Foil of fibers**	**free-space**
Intra chip				Good
Intra MCM				Good
Intra board			Good	Good
Inter board	Expensive	Good	Good	(Good)
Inter cabinet	Expensive	Good		
Inter and intra room	Good	Good		
Intra and inter building	Good	Moderate		

Table 5: Suitability in different system sizes. Empty cells in a column means the technology/network is not suitable for the corresponding system sizes.

expensive components that today are primarily developed for long-distance communication (e.g., WDM components). Examples of reasons to upper bounds are channel-to-channel skew in fiber-ribbons (especially when having a dedicated clock channel), high signal-losses (e.g., foils of polymer waveguides), and alignment problems in free-space systems.

The increasingly good price/performance ratio for fiber-ribbon links indicates a great success potential for several of the networks discussed. On the other hand, the WDM technique offer flexible multi-channel networks that can be passively implemented. Integrated fiber and waveguide solutions makes the building of compact systems possible, especially for networks like those using fiber-ribbons. The same reasoning about compactness can be argued for free-space systems. Optical backplanes might have its success in the similarities with current rack-based systems.

6. Conclusions

A broad range of optical interconnection networks have been surveyed. It has been shown that several of these networks can benefit from a pipelined dataflow, both as a high ratio of spatial reuse and as decreased reconfiguration latencies. On the other side, several of the surveyed systems offer the high flexibility which is needed in systems with a high degree of many-to-many communication. Compactness is another parameter that have influence on the choice of network. Last, further developments and commercialization of components for the mass-market is of big importance for success of the networks.

7. References

[Acampora and Karol 1989] A. S. Acampora and M. J. Karol, "An overview of lightwave packet networks," *IEEE Network*, pp. 29-41, Jan. 1989.

[Agarwal 1991] A. Agarwal, "Limits on interconnection network performance," *IEEE Transactions on Parallel and Distributed Systems*, vol. 2, no. 4, pp. 398-412, Oct. 1991.

[Agarwal et al. 1995] A. Agarwal, R. Bianchini, D. Chaiken, K. L. Johnson, D. Kranz, J. Kubiatowicz, B.-H. Lim, K. Mackenzie, and D. Yeung, "The MIT-Alewife machine: architecture and performance," *Proc. 22nd International Symposia on Computer Architecture (ISCA'94)*, Santa Margherita Ligure, Italy, June 22-24, 1995.

[Agerwala et al. 1995] T. Agerwala, J. L. Martin, J. H. Mirza, D. C. Sadler, D. M. Dias, and M. Snir, "SP2 system architecture," *IBM Systems Journal*, vol. 34, no. 2, pp. 152-184, 1995.

[Almasi and Gottlieb 1994] G. S. Almasi and A. Gottlieb, Eds., *Highly Parallel Computing.* The Benjamin/Cummings Publishing Company, Inc., Redwood City, CA, USA, 1994, ISBN 0-8053-0443-6.

[Anderson and Cornelius 1992] T. M. Anderson and R. S. Cornelius, "High-performance switching with Fibre Channel," *Proc. 38th Annual IEEE International Computer Conference (COMPCON Spring 92) Digest*, San Francisco, CA, USA, Feb. 24-28, 1992, pp. 261-264.

[Anderson et al. 1995] T. E. Anderson, D. E. Culler, and D. A. Patterson, "A case for NOW (Networks of Workstations)," *IEEE Micro*, vol. 15, no. 1, pp. 54-64, Feb. 1995.

[Arnould et al. 1989] E. A. Arnould, F. J. Bitz, E. C. Cooper, H. T. Kung, R. D. Sansom, and P. A. Steenkiste, "The design of Nectar: a network backplane for heterogeneous multicomputers," *Proc. ASPLOS-III*, pp. 205-216, 1989.

[Aronson et al. 1998] L. B. Aronson, B. E. Lemoff, L. A. Buckman, and D. W. Dolfi, “ Low-cost multimode WDM for local area networks up to 10 Gb/s,” *IEEE Photonics Technology Letters*," vol. 10, no. 10, pp. 1489-1491, Oct. 1998.

[Arvind et al. 1991] K. Arvind, K. Ramamritham, and J. A. Stankovic, “A local area network architecture for communication in distributed real-time systems,” *Journal of Real-Time Systems*, vol. 3, no. 2, pp. 115-147, May 1991.

[Batcher 1980] K. E. Batcher, “Design of a massively parallel processor,” *IEEE Transactions on Computers*, vol. 29, no. 9, pp. 836-840, Sept. 1980.

[Batcher 1980B] K. E. Batcher, "Architecture of a massively parallel processor," *Proc. 7th International Symposia on Computer Architecture (ISCA'80)*, May 1980.

[Beecroft et al. 1994] J. Beecroft, "Meiko CS-2 interconnect ELAN-ELITE design," *Parallel Computing*, vol. 20, pp. 1627-1638, no. 10, 1994.

[Bengtsson et al. 1993] L. Bengtsson, A. Linde, B. Svensson, and A. Åhlander, “The REMAP massively parallel computer platform for neural computations," *Proc. Third International Conference on Microelectronics for Neural Networks (MICRONEURO 1993)*, Edinburgh, Scotland, UK, Apr. 6-8, 1993.

[Bergman et al. 1998] L. Bergman, J. Morookian, and C. Yeh, “An all-optical long-distance multi-Gbytes/s bit-parallel WDM single-fiber link,” *Journal of Lightwave Technology*, vol. 16, no. 9, pp. 1577-1582, Sept. 1998.

[Bergman et al. 1998B] L. A. Bergman, C. Yeh, and J. Morookian, ”Towards the realization of multi-km × Gbyte/sec bit-parallel WDM single fiber computer links,” *Proc. 5th International Conference on Massively Parallel Processing using Optical Interconnections (MPPOI 98)*, Las Vegas, NV, USA, June 15-17, 1998, pp. 218-223.

[Betzos and Mitkas 1998] G. A. Betzos and P. A. Mitkas, ”Performance evaluation of massively parallel processing architectures with three-dimensional optical interconnections," *Applied Optics*, vol. 37, no. 2, pp. 315-325, Jan. 10, 1998.

[Bhoedjang et al. 1998] R. Bhoedjang, T. Rühl, and H. Bal, "Efficient multicast on Myrinet using link-level flow control," *Proc. of the International Conference on Parallel Processing (ICPP 98)*, Minneapolis, MN, USA, Aug. 10-14, 1998.

[Bhoedjang et al. 1998B] R. A. F. Bhoedjang, T. Rühl, and H. E. Bal, “User-level network interface protocols,” *Computer*, vol. 31, no. 11, pp. 53-60, Nov. 1998.

[Bhuyan et al. 1989] L. N. Bhuyan, Q. Yang, and D. P. Agrawal, “Performance of multiprocessor interconnection networks,” *Computer*, vol. 22, no. 2, pp. 25-37, Feb. 1989.

[Blank 1990] T. Blank, "The MasPar MP-1 architecture," *Proc. 35th Annual IEEE International Computer Conference (COMPCON Spring '90) Digest*, San Francisco, CA, USA, Feb. 26 - Mar. 2, 1990, pp. 20-24.

[Boden et al. 1995] N. J. Boden, D. Cohen, R. E. Felderman, A. E. Kulawik, C. L. Seitz, J. N. Seizovic, and W.-K. Su, "Myrinet: a gigabit-per-second local area network," *IEEE Micro*, vol. 15, no. 1, pp. 29-36, Feb. 1995.

[Boggess and Shirley 1997] T. Boggess and F. Shirley, "High-performance scalable computing for real-time applications," *Proc. of the Sixth International Conference on Computer Communications and Networks (IC^{3}N '97)*, Las Vegas, NV, USA, Sept. 22-25, 1997, pp. 332-335.

[Bohr 1998] M. Bohr, “Silicon trends and limits for advanced microprocessors,” *Communications of the ACM*, vol. 41, no. 3, pp. 80-87, Mar. 1998.

[Boisseau et al. 1994] M. Boisseau, M. Demange, and J.-M. Munier, *High Speed Networks*. John Wiley & Sons, Ltd., West Sussex, England, UK, 1994, ISBN 0-471-95109-9.

[Brackett 1990] C. A. Brackett, "Dense wavelength division multiplexing networks: principles and applications," *IEEE Journal on Selected Areas in Communications*, vol. 8, no. 6, pp. 948-964, Aug. 1990.

[Brackett 1996] C. A. Brackett, “Foreword: Is there an emerging consensus on WDM networking?,” *Journal of Lightwave Technology*, vol. 14, no. 6, pp. 936-941, June 1996.

[Carlile 1993] B. R. Carlile, "Algorithms and design: the CRAY APP shared-memory system," *Proc. 38th Annual IEEE International Computer Conference (COMPCON Spring '93) Digest*, San Francisco, CA, USA, pp. 312-320, 1993.

[Casavant et al. 1996] T. L. Casavant, P. Tvrdík, and F. Plášil, Eds., *Parallel Computers: Theory and Practice*. IEEE Computer Society Press, Los Alamitos, CA, USA, 1996, ISBN 0-8186-5162-8.

[Caulfield 1998] H. J. Caulfield, “Perspectives in optical computing,” *Computer*, vol. 31, no. 2, pp. 22-25, Feb. 1998.

[Chamberlain et al. 1998] R. D. Chamberlain, M. A. Franklin, R. B. Krchnavek, and B. H. Baysal, ”Design of an optically-interconnected multiprocessor,” *Proc. 5th International Conference on Massively Parallel Processing using Optical Interconnections (MPPOI'98),* Las Vegas, NV, USA, June 15-17, 1998, pp. 114-122.

[Charlesworth 1998] A. Charlesworth, "Starfire: extending the SMP envelope," *IEEE Micro*, vol. 18, no. 1, pp. 39-49, Jan./Feb. 1998.

[Chen et al. 1998] C.-H. Chen, B. Hoanca, C. B. Kuznia, A. A. Sawchuk, and J.-M. Wu, ”Architecture and optical system design for TRANslucent Smart Pixel ARray (TRANSPAR) chips," *Proc. Optics in Computing (OC'98),* Brugge, Belgium, June 17-20, 1998, pp. 316-319.

[Chen et al. 1998B] C.-H. Chen, B. Hoanca, C. B. Kuznia, A. A. Sawchuk, and J.-M. Wu, ”TRANslucent Smart Pixel ARray (TRANSPAR) chips for high throughput networks and SIMD signal processing,” *Proc. 5th International Conference on Massively Parallel Processing using Optical Interconnections (MPPOI'98),* Las Vegas, NV, USA, June 15-17, 1998, pp. 42-49.

[Cheriton et al. 1991] D. R. Cheriton, H. A. Goosen, and P. D. Boyle, “Paradigm: a highly scalable shared-memory multicomputer architecture,” *Computer,* vol. 24, no. 2, pp. 33-46, Feb. 1991.

[Cheung 1990] K.-W. Cheung, "Acoustooptic tunable filters in narrowband WDM networks: system issues and network applications," *IEEE Journal on Selected Areas in Communications*, vol. 8, no. 6, pp. 1015-1025, Aug. 1990.

[Christensen and Haney 1997] M. P. Christensen and M. W. Haney, ”Two-bounce free-space arbitrary interconnection architecture,” *Proc. Massively Parallel Processing using Optical Interconnections (MPPOI'97),* Montreal, Canada, June 22-24, 1997, pp. 61-67.

[Coldren et al. 1998] L. A. Coldren, E. R. Hegblom, Y. A. Akulova, J. Ko, E. M. Strzelecka, and S. Y. Hu, ”Vertical-cavity lasers for parallel optical interconnects,” *Proc. 5th International Conference on Massively Parallel Processing using Optical Interconnections (MPPOI'98),* Las Vegas, NV, USA, June 15-17, 1998, pp. 2-10.

[Comer 1995] D. E. Comer, *Internetworking with TCP/IP, Vol I: Principles, Protocols, and Architecture.* third edition, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1995, ISBN 0-13-227836-7.

[Dally 1992] W. J. Dally, "Virtual-channel flow control," *IEEE Transactions on Parallel and Distributed Systems*, vol. 3, no. 2, pp. 194-205, Mar. 1992.

[Dally and Seitz 1987] W. J. Dally and C. L. Seitz, “Deadlock-free message routing in multiprocessor interconnection networks,” *IEEE Transactions on Computers*, vol. 36, no. 5, pp. 547-553, May 1987.

[Dally et al. 1993] W. J. Dally, J. S.Keen, and M. D. Noakes, "The J-Machine architecture and evaluation," *Proc. 38th Annual IEEE International Computer Conference (COMPCON Spring ’93) Digest*, San Francisco, CA, USA, pp. 183-188, 1993.

[Dally et al. 1998] W. J. Dally, M.-J. E. Lee, F.-T. An, J. Poulton, and S. Tell, ”High-performance electrical signaling,” *Proc. 5th International Conference on Massively Parallel Processing using Optical Interconnections (MPPOI’98)*, Las Vegas, NV, USA, June 15-17, 1998, pp. 11-16.

[Decegama 1989] A. L. Decegama, *The Technology of Parallel Processing: Parallel Processing Architectures and VLSI Hardware Volume I*. Prentice-Hall, Inc., Englewood Cliffs, NJ, USA, 1989, ISBN 0-13-898438-7.

[Duan and Wilmsen 1998] C. Duan and C. W. Wilmsen, "Optoelectronic ATM switch using VCSEL and smart detector arrays," *Proc. Optics in Computing (OC'98)*, Brugge, Belgium, June 17-20, 1998, pp. 103-106.

[Duato et al. 1997] J. Duato, S. Yalamanchili, and L. Ni, *Interconnection Networks: an Engineering Approach*. IEEE Computer Society Press, Los Alamitos, CA, USA, 1997, ISBN 0-8186-7800-3.

[Eicken et al. 1995] T. von Eicken, A. Basu, and V. Buch, "Low-latency communication over ATM networks using active messages," *IEEE Micro*, vol. 15, no. 1, pp. 46-53, Feb. 1995.

[Einstein 1996] T. Einstein, "RACEway Interlink - a real-time multicomputing interconnect fabric for high-performance VMEbus systems," *VMEbus Systems*, Spring 1996.

[Emerson 1995] S. Emerson, "Evaluation of a data communication model for switched fibre channel," *IEEE Network*, pp. 38-44, Nov./Dec. 1995.

[Eriksen et al. 1995] P. Eriksen, K. Gustafsson, M. Niburg, G. Palmskog, M. Robertsson, and K. Åkermark, "The Apollo demonstrator – new low-cost technologies for optical interconnects," *Ericsson Review*, vol. 72, no. 2, 1995.

[Felderman et al. 1994] R. Felderman, A. DeSchon, D. Cohen, and G. Finn, “ATOMIC: a high speed local communciation architecture,” *Journal of High Speed Networks*, vol. 3, no. 1, pp. 1-30, 1994.

[Feldman et al. 1987] M. R. Feldman, S. C. Esener, C. C. Guest, and S. H. Lee, "Comparison between optical and electrical interconnects based on power and speed considerations," *Applied Optics*, vol. 27, no. 9, pp. 1742-1751, May 1, 1988.

[Flipse 1998] R. Flipse, "Optical switches ease bandwidth crunch," *EuroPhotonics*, vol. 3, no. 5, pp. 44-45, Aug./Sept. 1998.

[Flynn 1966] M. J. Flynn, "Very high-speed computing systems," *Proceedings of the IEEE*, vol. 54, no. 12, pp. 1901-1909, Dec. 1966.

[Flynn 1972] M. J. Flynn, "Some computer architectures and their effectiveness," *IEEE Transactions on Computers*, vol. 21, no. 9, pp. 948-960, Sept. 1972.

[Freeman 1998] R. L. Freeman, "Bits, symbols, bauds, and bandwidth," *IEEE Communications Magazine*, vol. 36, no. 4, pp. 96-99, Apr. 1998.

[Galles 1997] M. Galles, "Spider: a high-speed network interconnect," *IEEE Micro*, vol. 17, no. 1, pp. 34-39, Jan./Feb. 1997.

[Geist et al. 1994] A. Geist, A. Beguelin, J. Dongarra, W. Jiang, R. Manchek, and V. Sunderam, *PVM: Parallel Virtual Machine - A Users' Guide and Tutorial for Networked Parallel Computing*. The MIT Press, Cambridge, MA, USA, 1994.

[Goke and Lipovski 1973] L. R. Goke and G. J. Lipovski, "Banyan networks for partitioning multiprocessor systems," *Proc. 1st International Symposia on Computer Architecture (ISCA'73)*, 1973.

[Goldberg 1997] L. Goldberg, "Moving toward the light: new optics developments reshaping electronics," *Electronic Design*, vol. 45, no. 28, pp. 65-74, Dec. 15, 1997.

[Goodman et al. 1984] J. W. Goodman, F. I. Leonberger, S.-Y. Kung, and R. A. Athale, "Optical interconnections for VLSI systems," *Proceedings of the IEEE*, vol. 72, no. 7, pp. 850-866, July 1984.

[Gottlieb et al. 1982] A. Gottlieb, R. Grishman, C. P. Kruskal, K. P. McAuliffe, L. Rudolph, and M. Snir, "The NYU Ultracomputer – designing a MIMD, shared-memory parallel machine," *Proc. 9th International Symposia on Computer Architecture (ISCA'82)*, 1982.

[Gottlieb et al. 1983] A. Gottlieb, R. Grishman, C. P. Kruskal, K. P. McAuliffe, L. Rudolph, and M. Snir, "The NYU Ultracomputer – designing an MIMD shared memory parallel computer," *IEEE Transactions on Computers*, vol. 32, no. 2, pp. 175-189, Feb. 1983.

[Gourlay et al. 1998] J. Gourlay, T.-Y. Yang, J. A. B. Dines, J. F. Snowdon, and A. C. Walker, "Development of free-space digital optics in computing," *Computer*, vol. 31, no. 2, pp. 38-44, Feb. 1998.

[Guilfoyle et al. 1998] P. S. Guilfoyle, J. M. Hessenbruch, and R. V. Stone, "Free-space interconnects for high-performance optoelectronic switching," *Computer*, vol. 31, no. 2, pp. 69-75, Feb. 1998.

[Gustavson and Li 1996] D. B. Gustavson and Q. Li, "The scalable coherent interface (SCI)," *IEEE Communications Magazine*, no. 8, pp. 52-63, Aug. 1996.

[Hahn 1995] K. H. Hahn, "POLO – Parallel optical links for gigabyte/s data communications," *Proc. LEOS 95*, San Francisco, CA, USA, Oct. 30 – Nov. 2, 1995, vol. 1, pp. 228-229.

[Halsall 1995] F. Halsall, *Data Communications, Computer Networks and Open Systems*. fourth edition, Addison-Wesley Longman Ltd., Essex, UK, 1995, ISBN 0-201-42293-X.

[Hamanaka 1991] K. Hamanaka, "Otical bus interconnection system using selfloc lenses," *Optics Letters*, vol. 16, no. 6, pp. 1222-1224, Aug. 15, 1991.

[Hennessy and Patterson 1996] J. L. Hennessy and D. A. Patterson, *Computer Architecture: A Quantitative Approach*. Morgan Kaufmann Publishers, Inc., San Francisco, CA, USA, 1996, ISBN 0-55860-329-8.

[Hillis and Tucker 1993] W. D. Hillis and L. W. Tucker, "The CM-5 Connection Machine: a scalable supercomputer," *Communications of the ACM*, vol. 36, no. 11, pp. 31-40, Nov. 1993.

[Hinton and Szymanski 1995] H. S. Hinton and T. H. Szymanski, "Intelligent optical backplanes," *Proc. 2nd International Conference on Massively Parallel Processing using Optical Interconnections (MPPOI 95)*, San Antonio, TX, USA, Oct 23-24, 1995, pp. 133-143.

[Hirabayashi et al. 1998] K. Hirabayashi, T. Yamamoto, and S. Hino, "Optical backplane with free-space optical interconnections using tunable beam deflectors and a mirror for bookshelf-assembled terabit per second class asynchronous transfer mode switch," *Optical Engineering*, vol. 37, no. 4, pp. 1332-1342, Apr. 1998.

[Hockney and Jesshope 1988] R. W. Hockney and C. R. Jesshope, *Parallel Computers 2: architecture, programming and algorithms*. Adam Hilger, Bristol, UK, 1988, ISBN 0-85274-812-4.

[Hord 1990] R. M. Hord, *Parallel Supercomputing in SIMD Architectures*. CRC Press, Inc., Boca Raton, FL, USA, 1990, ISBN 0-8493-4271-6.

[Hord 1993] R. M. Hord, *Parallel Supercomputing in MIMD Architectures*. CRC Press, Inc., Boca Raton, FL, USA, 1993, ISBN 0-8493-4417-4.

[Horowitz et al. 1998] M. Horowitz, C.-K. K. Yang, and S. Sidiropoulos, "High-speed electrical signaling: overview and limitations," *IEEE Micro*, vol. 18, no. 1, pp. 12-24, Jan./Feb. 1998.

[Horst 1995] R. W. Horst, "TNet: a reliable system area network," *IEEE Micro*, vol. 15, no. 1, pp. 37-45, Feb. 1995.

[Hwang 1993] K. Hwang, *Advanced Computer Architecture*. McGraw-Hill, Inc., 1993, ISBN 0-07-031622-8.

[Hwang and Briggs 1985] K. Hwang and F. A. Briggs, *Computer Architecture and Parallel Processing*. McGraw-Hill, Inc., 1985, ISBN 0-07-031556-6.

[IEEE 1993] *IEEE Standard for Scalable Coherent Interface (SCI)*. IEEE, New York, NY, USA, 1993, ISBN 1-55937-222-2.

[Irakliotis and Mitkas 1998] L. J. Irakliotis and P. A. Mitkas, "Optics: a maturing technology for better computing," *Computer*, vol. 31, no. 2, pp. 36-37, Feb. 1998.

[Irshid and Kavehrad 1992] M. I. Irshid and M. Kavehrad, "A fully transparent fiber-optic ring architecture for WDM networks," *Journal of Lightwave Technology*, vol. 10, no. 1, pp. 101-108, Jan. 1992.

[Isenstein 1994] B. Isenstein, "Scaling I/O bandwidth with multiprocessors," *Electronic Design*, June 13, 1994.

[Ishihata et al. 1997] H. Ishihata, M. Takahashi, and H. Sato, " Hardware of AP3000 scalar parallel server," *Fujitsu Scientific and Technical Journal*, vol. 33, no. 1, pp. 24-30, June 1997.

[Ishikawa and McArdle 1998] M. Ishikawa and N. McArdle, "Optically interconnected parallel computing systems," *Computer*, vol. 31, no. 2, pp. 61-68, Feb. 1998.

[Jahns 1994] J. Jahns, "Planar packaging of free-space optical interconnects," *Proceedings of the IEEE*, vol. 82, no. 11, pp. 1623-1631, Nov. 1994.

[Jahns 1998] J. Jahns, ”Integrated free-space optical interconnects for chip-to-chip communications,” *Proc. 5th International Conference on Massively Parallel Processing using Optical Interconnections (MPPOI'98)*, Las Vegas, NV, USA, June 15-17, 1998, pp. 20-23.

[Jonsson 1998] M. Jonsson, “Control-channel based fiber-ribbon pipeline ring network," *Proc. Massively Parallel Processing using Optical Interconnections (MPPOI'98)*, Las Vegas, NV, USA, June 15-17, 1998, pp. 158-165.

[Jonsson 1998B] M. Jonsson, “Two fiber-ribbon ring networks for parallel and distributed computing systems,” *Optical Engineering*, vol. 37, no. 12, pp. 3196-3204, Dec. 1998.

[Jonsson and Svensson 1997] M. Jonsson and B. Svensson, “On inter-cluster communication in a time-deterministic WDM star network,” *Proc. 2nd Workshop on Optics and Computer Science (WOCS)*, Geneva, Switzerland, Apr. 1, 1997.

[Jonsson et al. 1996] M. Jonsson, A. Åhlander, M. Taveniku, and B. Svensson, “Time-deterministic WDM star network for massively parallel computing in radar systems,” *Proc. Massively Parallel Processing using Optical Interconnections, MPPOI'96*, Lahaina, HI, USA, Oct. 27-29, 1996, pp. 85-93.

[Jonsson et al. 1997] M. Jonsson, K. Börjesson, and M. Legardt, “Dynamic time-deterministic traffic in a fiber-optic WDM star network,” *Proc. 9th Euromicro Workshop on Real Time Systems*, Toledo, Spain, June 11-13, 1997.

[Jonsson et al. 1997B] M. Jonsson, B. Svensson, M. Taveniku, and A. Åhlander, “Fiber-ribbon pipeline ring network for high-performance distributed computing systems,” *Proc. International Symposium on Parallel Architectures, Algorithms and Networks (I-SPAN'97)*, Taipei, Taiwan, Dec. 18-20, 1997, pp. 138-143.

[Karstensen et al. 1995] H. Karstensen, C. Hanke, M. Honsberg, J.-R. Kropp, J. Wieland, M. Blaser, P. Weger, and J. Popp, “Parallel optical interconnection for uncoded data transmission with 1 Gb/s-per-channel capacity, high dynamic range, and low power consumption,” *Journal of Lightwave Technology*, vol. 13, no. 6, pp. 1017-1030, June 1995.

[Kato et al. 1998] T. Kato, J. Sasaki, T. Shimoda, H. Hatakeyama, T. Tamanuki, M. Yamaguchi, M. Kitamura, and M. Itoh, "10 Gb/s photonic cell switching with hybrid 4×4 optical matrix switch module on silica based planar waveguide platform," *Optical Fiber Communication Conference, OFC'98 Technical Digest*, San Jose, CA, USA, Feb. 22-27, 1998, pp. 437-440.

[Kawai et al. 1995] S. Kawai, H. Kurita, and K. Kubota, "Design of electro-photonic computer-networks with non-blocking and self-routing functions," *Optical Computing, vol. 10, 1995 OSA Technical Digest Series*, Salt Lake City, Utah, Mar. 13-16, 1995, pp. 263-265.

[Kermani and Kleinrock 1979] P. Kermani and L. Kleinrock, "Virtual cut-through: a new computer communication switching technique," *Computer Networks*, vol. 3, pp. 267-286, 1979.

[Kessler and Schwarzmeier 1993] R. E. Kessler and J. L. Schwarzmeier, "CRAY T3D: a new dimension for Cray Research," *Proc. 38th Annual IEEE International Computer Conference (COMPCON Spring '93) Digest,* San Francisco, CA, USA, pp. 176-182, 1993.

[Kiefer and Swanson 1995] D. R. Kiefer and V. W. Swanson, "Implementation of optical clock distribution in a supercomputer," *Optical Computing, vol. 10, 1995 OSA Technical Digest Series*, Salt Lake City, Utah, Mar. 13-16, 1995, pp. 260-262.

[Kobrinski et al. 1988] H. Kobrinski, M. P. Vecchi, E. L. Goldstein, and R. M. Bulley, "Wavelength selection with nanosecond switching times using distributed-feedback laser amplifiers," *Electronics Letters*, vol. 24, no. 15, pp. 969-971, July 21, 1988.

[Koeninger et al. 1994] R. K. Koeninger, M. Furtney, and M. Walker, "A shared memory MPP from Cray Research," *Digital Technical Journal,* vol. 6, no. 2, pp. 8-21, 1994.

[Krisnamoorthy et al. 1996] A. V. Krisnamoorthy, J. E. Ford, K. W. Goosen, J. A. Walker, B. Tseng, S. P. Hui, J. E. Cunningham, W. Y. Jan, T. K. Woodward, M. C. Nuss, R. G. Rozier, F. E. Kiamilev, and D. A. B. Miller, "The AMOEBA chip: an optoelectronic switch for multiprocessor networking using dense-WDM," *Proc. 3rd International Conference on Massively Parallel Processing using Optical Interconnections (MPPOI'96),* Maui, HI, USA, Oct. 27-29, 1996, pp. 94-100.

[Kuck et al. 1993] D. Kuck, E. Davidson, D. Lawrie, A. Sameh, C.-Q. Zhu, A. Vaidenbaum, J. Konicek, P. Yew, K. Gallivan, W. Jallby, H. WijshoffR. Bramley, U. M. Yang, P. Emrath, D. Padua, R. Eigenmann, J. Hoeflinger, G. Jaxon, Z. Li, T. Murphy, J. Andrews, and S. Turner, "The Cedar system and an initial performance study," *Proc. 20th International Symposia on Computer Architecture (ISCA'93),* San Diego, CA, USA, May 16-19, 1993.

[Kurokawa and Ikegami 1996] T. Kurokawa and T. Ikegami, ”Optical interconnection technologies based on vertical-cavity surface-emitting lasers and smart pixels,” *Proc. 3rd International Conference on Massively Parallel Processing using Optical Interconnections (MPPOI'96),* Maui, HI, USA, Oct. 27-29, 1996, pp. 300-305.

[Kurokawa et al. 1998] T. Kurokawa, S. Matso, T. Nakahara, K. Tateno, Y. Ohiso, A. Wakatsuki, and H. Tsuda, ”Design approaches for VCSEL's and VCSEL-based smart pixels toward parallel optoelectronic processing systems," *Applied Optics,* vol. 37, no. 2, pp. 194-204, Jan. 10, 1998.

[Kuszmaul 1995] B. C. Kuszmaul, "The RACE network architecture," *Proc. 9th International Parallel Processing Symposium (IPPS'95),* Santa Barbara, CA, USA, Apr. 25-28, 1995, pp. 508-513.

[Lane et al. 1989] T. A. Lane, J. A. Quam, B. O. Khale, and E. C. Parish, "Gigabit optical interconnections for the Connection Machine," in *Optical Interconnects in the Computer Environment, Proc. SPIE vol. 1178,* J. Pazaris and G. R. Willenbring, Eds., pp. 24-35, 1989.

[Laudon and Lenoski 1997] J. Laudon and D. Lenoski, "The SGI Origin: a ccNUMA highly scalable server," *Proc. 24th International Symposia on Computer Architecture (ISCA'97),* Denver, CO, USA, June 2-4, 1997.

[Lawson et. al. 1992] H. W. Lawson; with contributions by B. Svensson and L. Wanhammar, *Parallel Processing in Industrial Real-Time Applications.* Prentice-Hall, Inc., 1992.

[Leiserson 1985] C. E. Leiserson, “Fat-trees: universal networks for hardware-efficient supercomputing,” *IEEE Transactions on Computers,* vol. 34, no. 10, pp. 892-901, Oct. 1985.

[Leiserson et al. 1992] C. E. Leiserson, Z. S. Abuhamdeh, D. C. Douglas, C. R. Feynman, M. N. Ganmukhi, J. V. Hill, W. D. Hillis, B. C. Kuszmaul, M. A. St. Pierre, D. S. Wells, M. C. Wong, S.-W. Yang, and R. Zak, "The network architecture of the Connection Machine CM-5," *Proc. of the Fourth Annual ACM Symposium on Parallel Algorithms and Architectures,* June 1992.

[Lenoski et al. 1992] D. Lenoski, J. Laudon, T. Joe, D. Nakahira, L. Stevens, A. Gupta, and J. Hennessy, "The DASH prototype: implementation and performance," *Proc. 19th International Symposia on Computer Architecture (ISCA'92),* Gold Coast, Australia, May 19-21, 1992.

[Lenoski et al. 1992B] D. Lenoski, J. Laudon, K. Gharachorloo, W.-D. Weber, A. Gupta, J. Hennessy, M. Horowitz, and M. S. Lam, "The Stanford DASH multiprocessor," *Computer,* vol. 25, no. 3, pp. 63-79, Mar. 1992.

[Lenoski et al. 1993] D. Lenoski, J. Laudon, T. Joe, D. Nakahira, L. Stevens, A. Gupta, and J. Hennessy, "The DASH prototype: logic overhead and performance," *IEEE Transactions on Parallel and Distributed Systems,* vol. 4, no. 1, pp. 41-61, Jan. 1993.

[Li et al. 1998] Y. Li, J. Popelek, J.-K. Rhee, L. J. Wang, T. Wang, and K. Shum, "Demonstration of fiber-based board-level optical clock distributions," *Proc. 5th International Conference on Massively Parallel Processing using Optical Interconnections (MPPOI 98),* Las Vegas, NV, USA, June 15-17, 1998, pp. 224-228.

[Li et al. 1998B] Y. Li, T. Wang, and S. Kawai, "Distributed crossbar interconnects with vertical-cavity surface-emitting laser-angle multiplexing and fiber image guides," *Applied Optics,* vol. 37, no. 2, pp. 254-263, Jan. 10, 1998.

[Li et al. 1998C] Y. Li, J. Ai, and T. Wang, "100×100 opto-electronic cross-connector using OPTOBUS™," *Proc. Optics in Computing (OC'98),* Brugge, Belgium, June 17-20, 1998, pp. 282-284.

[Liu and Prasanna 1998] W. Liu and V. K. Prasanna, "Utilizing the power of high-performance computing," *IEEE Signal Processing Magazine,* vol. 15, no. 5, pp. 85-100, Sept. 1998.

[Lohmann et al. 1986] A. W. Lohmann, W. Stork, and G. Stucke, "Optical perfect shuffle," *Applied Optics,* vol. 25, no. 10, pp. 1530-1531, May 15, 1986.

[Lovett and Clapp 1996] T. Lovett and R. Clapp, "STiNG: a CC-NUMA computer system for the commercial marketplace," *Proc. 23rd International Symposia on Computer Architecture (ISCA'96),* May 1996.

[Lukowicz et al. 1998] P. Lukowicz, S. Sinzinger, K. Dunkel, and H.-D. Bauer, "Design of an opto-electronic VLSI/parallel fiber bus," *Proc. Optics in Computing (OC'98),* Brugge, Belgium, June 17-20, 1998, pp. 289-292.

[Lund 1997] C. Lund, ”Optics inside future computers,” *Proc. Massively Parallel Processing using Optical Interconnections (MPPOI'97),* Montreal, Canada, June 22-24, 1997, pp. 156-159.

[Maeno et al. 1997] Y. Maeno, A. Tajima, Y. Suemura, and N. Henmi, ”8.5 Gbit/s/port synchronous optical packet-switch,” *Proc. Massively Parallel Processing using Optical Interconnections (MPPOI'97),* Montreal, Canada, June 22-24, 1997, pp. 114-119.

[MasPar 1992] “The design of the MasPar MP-2: a cost effective massively parallel computer,” *White paper, MasPar Computer Corporation, Sunnyvale, CA, USA*, 1992.

[McArdle et al. 1996] N. McArdle, M. Naruse, T. Komuro, H. Sakaida, M. Ishikawa, Y.Kobayashi, and H. Toyoda, ”A smart-pixel parallel optoelectronic computing system with free-space dynamic interconnections,” *Proc. 3rd International Conference on Massively Parallel Processing using Optical Interconnections (MPPOI'96),* Maui, HI, USA, Oct. 27-29, 1996, pp. 146-157.

[McArdle et al. 1997] N. McArdle, M. Naruse, T. Komuro, and M. Ishikawa, ”Realisation of a smart-pixel parallel optoelectronic computing system,” *Proc. Massively Parallel Processing using Optical Interconnections (MPPOI'97),* Montreal, Canada, June 22-24, 1997, pp. 190-195.

[Mercury 1998] *RACE® Series RACEway Interlink Modules Data Sheet.* Mercury Computer Systems, Inc., 1998.

[Miura et al. 1993] K. Miura, M. Takamura, Y. Sakamoto, and S. Okada, "Overview of the Fujitsu VPP500 supercomputer," *Proc. 38th Annual IEEE International Computer Conference (COMPCON Spring '93) Digest*, San Francisco, CA, USA, pp. 128-130, 1993.

[Mudge et al. 1987] T. N. Mudge, J. P. Haynes, and D. C. Winsor, “Multiple bus architectures,” *Computer*, vol. 20, no. 6, pp. 42-48, June 1987.

[Mukherjee 1992] B. Mukherjee, "WDM-based local lightwave networks part I: single-hop systems," *IEEE Network*, pp. 12-27, May 1992.

[Mukherjee 1992B] B. Mukherjee, "WDM-based local lightwave networks part II: multihop systems," *IEEE Network*, pp. 20-32, July 1992.

[Neff 1994] J. A. Neff, ”Optical interconects based on two-dimensional VCSEL arrays,” *Proc. Massively Parallel Processing using Optical Interconnections (MPPOI'94),* Cancun, Mexico, Apr. 26-27, 1994, pp. 202-212.

[Neff et al. 1996] J. A. Neff, C. Chen, T. McLaren, C.-C. Mao, A. Fedor, W. Berseth, Y. C. Lee, and V. Morozov, ”VCSEL/CMOS smart pixel arrays for free-space optical interconnects,” *Proc. 3rd International Conference on Massively Parallel Processing using Optical Interconnections (MPPOI 96),* Maui, HI, USA, Oct. 27-29, 1996, pp. 282-289.

[Ni and McKinley 1993] L. M. Ni and P. K. McKinley, “A survey of wormhole routing techniques in direct networks,” *Computer*, vol. 26, no. 2, pp. 62-76, Feb. 1993.

[Nickolls 1992] J. R. Nickolls, "Interconnection architecture and packaging in massively parallel computers," *Proc. Packaging, Interconnects, and Optoelectronics for the Design of Parallel Computers Workshop*, Mar. 18-19, 1992.

[O'Keefe and Dietz 1990] M. T. O'Keefe and H. G. Dietz, "Hardware barrier synchronization: static barrier MIMD (SBM)," *Proc. of the International Conference on Parallel Processing (ICPP 90)*, pp. 35-42, Aug. 1990.

[O'Keefe and Dietz 1990B] M. T. O'Keefe and H. G. Dietz, "Hardware barrier synchronization: dynamic barrier MIMD (DBM)," *Proc. of the International Conference on Parallel Processing (ICPP 90)*, pp. 43-46, Aug. 1990.

[Ozaktas 1997] H. M. Ozaktas, ”Fundamentals of optical interconnections – a review,” *Proc. Massively Parallel Processing using Optical Interconnections (MPPOI 97)*, Montreal, Canada, June 22-24, 1997, pp. 184-189.

[Ozaktas 1997B] H. M. Ozaktas, ”Toward an optimal foundation architecture for optoelectronic computing. Part I. Regularly interconnected device planes," *Applied Optics*, vol. 36, no. 23, pp. 5682-5696, Aug. 10, 1997.

[Ozaktas 1997C] H. M. Ozaktas, ”Toward an optimal foundation architecture for optoelectronic computing. Part II. Physical construction and application platforms," *Applied Optics*, vol. 36, no. 23, pp. 5697-5705, Aug. 10, 1997.

[Parker et al. 1992] J. W. Parker, P. J. Ayliffe, T. V. Clapp, M. C. Geear, P. M. Harrison, and R. G. Peall, "Multifibre bus for rack-to-rack interconnects based on opto-hybrid transmitter/receiver array pair," *Electronics Letters*, vol. 28, no. 8, pp. 801-803, April 9, 1992.

[Pedersen et al. 1998] R. J. S. Pedersen, B. Mikkelsen, B. F. Jørgensen, M. Nissov, K. E. Stubkjaer, K. Wünstel, K. Daub, E. Lach, G. Laube, W. Idler, M. Schilling, P. Doussiere, and F. Pommerau, "WDM cross-connect cascade based on all-optical wavelength converters for routing and wavelength slot interchanging using a reduced number of internal wavelengths," *Optical Fiber Communication Conference, OFC'98 Technical Digest*, San Jose, CA, USA, Feb. 22-27, 1998, pp. 58-59.

[Peterson and Davie 1996] L. L. Peterson and B. S. Davie, *Computer Networks: a Systems Approach*. Morgan Kaufmann Publishers, Inc., San Francisco, CA, USA, 1996, ISBN 0-55860-368-9.

[Pinkston et al. 1998] T. M. Pinkston, M. Raksapatcharawong, and Y. Choi, "WARP II: an optoelectronic fully adaptive network router chip," *Proc. Optics in Computing (OC'98)*, Brugge, Belgium, June 17-20, 1998, pp. 311-315.

[Prylli and Tourancheau 1998] L. Prylli and B. Tourancheau, " BIP: a new protocol designed for high performance networking on Myrinet," *Proc. of the 1st Workshop on Personal Computer based Networks of Workstations (PC-NOW'98)*, Orlando, FL, USA, Apr. 3, 1998.

[Raghavan et al. 1999] B. Raghavan, Y.-G. Kim, T.-Y. Chuang, B. Madhavan, and A. F. J. Levi, "A Gbyte/s parallel fiber-optic network interface for multimedia applications," *IEEE Network*, Jan./Feb. 1999.

[Reed and Grunwald 1987] D. A. Reed and D. C. Grunwald, "The performance of multicomputer interconnection networks," *Computer*, vol. 20, no. 6, pp. 63-73, June 1987.

[Reif and Yoshida 1994] J. H. Reif and A. Yoshida, "Free space optical message routing for high performance parallel computers," *Proc. Massively Parallel Processing using Optical Interconnections (MPPOI'94)*, Cancun, Mexico, Apr. 26-27, 1994, pp. 37-44.

[Robertsson et al. 1995] M. Robertsson, K. Engberg, P. Eriksen, H. Hesselbom, M. Niburg, and G. Palmskog, "Optical interconnects in packaging for telecom applications," *Proc. of the 10th European Microelectronics Conference*, pp. 580-591, 1995.

[Rudolph 1998] L. Rudolph, "Do parallel computers really need optical interconnection networks?," *Proc. 5th International Conference on Massively Parallel Processing using Optical Interconnections (MPPOI'98)*, Las Vegas, NV, USA, June 15-17, 1998, pp. 176-183.

[Sachs and Varma 1996] M. W. Sachs and A. Varma, "Fibre channel and related standards," *IEEE Communications Magazine,* no. 8, pp. 40-50, Aug. 1996.

[Sano and Levi 1998] B. J. Sano and A. F. J. Levi, "Networks for the professional campus environment," in *Multimedia Technology for Applications.* B. Sheu and M. Ismail, Eds., McGraw-Hill, Inc., pp. 413-427, 1998, ISBN 0-7803-1174-4.

[Sano et al. 1996] B. Sano, B. Madhavan, and A. F. J. Levi, "8 Gbps CMOS interface for parallel fiber-optic interconnects," *Electronics Letters,* vol. 32, pp. 2262-2263, 1996.

[Saunders 1996] S. Saunders, *The McGraw-Hill High-Speed LANs Handbook*. McGraw-Hill, New York, NY, USA, 1996, ISBN 0-07-057199-6.

[Sawchuk et al. 1987] A. A. Sawchuk, B. K. Jenkins, C. S. Raghavendra, and A. Varma, "Optical crossbar networks," *Computer,* vol. 20, no. 6, pp. 50-60, June 1987.

[Schenfeld 1995] E. Schenfeld, "Massively parallel processing with optical interconnections: what can be done, should be and must not be done by optics," *Optical Computing, vol. 10, 1995 OSA Technical Digest Series,* Salt Lake City, Utah, Mar. 13-16, 1995, pp. 16-18.

[Schenfeld 1996] E. Schenfeld, "Massively parallel processing with optical interconnections," in *Parallel and Distributed Computing Handbook.* A. Y. Zomaya Ed., McGraw-Hill, Inc., pp. 780-810, 1996.

[Schwartz et al. 1996] D. B. Schwartz, K. Y. Chun, N. Choi, D. Diaz, S. Planer, G. Raskin, and S. G. Shook, "OPTOBUS™ I: performance of a 4 Gb/s optical interconnect," *Proc. Massively Parallel Processing using Optical Interconnections (MPPOI'96),* Maui, HI, USA, Oct. 27-29, 1996, pp. 256-263.

[Seitz 1985] C. L. Seitz, "The Cosmic Cube," *Communications of the ACM,* vol. 28, no. 1, pp. 22-33, Jan. 1985.

[Seitz and Su 1993] C. L. Seitz and W.-K. Su, "A family of routing and communication chips based on the Mosaic," *Research on Integrated Systems: Proceedings of the 1993 Symposium*, pp. 320-337, 1993.

[Sethu et al. 1998] H. Sethu, C. Stunkel, and R. Stucke, "IBM RS/6000 SP interconnection network topologies for large systems," *Proc. of the International Conference on Parallel Processing (ICPP'98),* Minneapolis, MN, USA, Aug. 10-14, 1998.

[Shahid and Holland 1996] M. A. Shahid and W. R. Holland, ”Flexible optical backplane interconnections,” *Proc. 3rd International Conference on Massively Parallel Processing using Optical Interconnections (MPPOI'96),* Maui, HI, USA, Oct. 27-29, 1996, pp. 178-185.

[Siala et al. 1994] S. Siala, A. P. Kanjamala, R. N. Nottenburg, and A. F. J. Levi, "Low skew multimode ribbon fibres for parallel optical communication," *Electronics Letters*, vol. 30, no. 21, pp. 1784-1786, Oct. 13, 1994.

[Siegel 1990] H. J. Siegel, *Interconnection Networks for Large-Scale Parallel Processing.* McGraw-Hill, Inc., 1990, ISBN 0-669-03594-7.

[Siegel et al. 1981] H. J. Siegel, L. J. Siegel, F. C. Kemmerer, P. T. Mueller, Jr., H. E. Smalley, Jr., and S. D. Smith, “PASM: a partionable SIMD/MIMD system for image processing and pattern recognition,” *IEEE Transactions on Computers,* vol. 30, no. 12, pp. 934-947, Dec. 1981.

[Silicon 1994] *Power Challenge Technical Report.* Silicon Graphics, Inc., 1994.

[Sinzinger 1998] S. Sinzinger, "Planar optics as the technological platform for optical interconnects," *Proc. Optics in Computing (OC'98),* Brugge, Belgium, June 17-20, 1998, pp. 40-43.

[Smeyne and Nickolls 1995] A. L. Smeyne and J. R. Nickolls, "A rugged scalable parallel system," *Proc. 9th International Parallel Processing Symposium (IPPS'95),* Santa Barbara, CA, USA, Apr. 25-28, 1995, pp. 502-507.

[Stallings 1997] W. Stallings, *Data and Computer Communications.* fifth edition, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1997, ISBN 0-13-571274-2.

[Steenkiste 1996] P. Steenkiste, "Network-based multicomputers: a practical supercomputer architecture," *IEEE Transactions on Parallel and Distributed Systems*, vol. 7, no. 8, pp. 861-875, Aug. 1996.

[Stenström 1990] P. Stenström, “A survey of cache coherence schemes for multiprocessors,” *Computer*, vol. 23, no. 6, pp. 12-24, June 1990.

[Stojmenovic 1996] I. Stojmenovic, "Direct interconnection networks," in *Parallel and Distributed Computing Handbook.* A. Y. Zomaya Ed., McGraw-Hill, Inc., pp. 537-567, 1996.

[Stunkel 1997] C. B. Stunkel, ”Commercial MPP networks: time for optics?,” *Proc. Massively Parallel Processing using Optical Interconnections (MPPOI 97)*, Montreal, Canada, June 22-24, 1997, pp. 90-95.

[Stunkel et al. 1994] C. B. Stunkel, D. G. Shea, B. Abali, M. Atkins, C. A. Bender, D. G. Grice, P. H. Hochschild, D. J. Joseph, B. J. Nathanson, R. A. Swetz, R. F. Stucke, M. Tsao, and P. R. Varker, “The SP2 communication subsystem,” *White paper, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA*, Aug. 1994.

[Stunkel et al. 1995] C. B. Stunkel, D. G. Shea, B. Abali, M. G. Atkins, C. A. Bender, D. G. Grice, P. Hochschild, D. J. Joseph, B. J. Nathanson, R. A. Swetz, R. F. Stucke, M. Tsao, P. R. Varker, "The SP2 high-performance switch," *IBM Systems Journal*, vol. 34, no. 2, pp. 185-204, 1995.

[Supmonchai and Szymanski 1998] B. Supmonchai and T. Szymanski, "High speed VLSI concentrators for terabit intelligent optical backplanes," *Proc. Optics in Computing (OC'98)*, Brugge, Belgium, June 17-20, 1998, pp. 306-310.

[Szymanski 1995] T. H. Szymanski, ”Intelligent optical backplanes," *Optical Computing, vol. 10, 1995 OSA Technical Digest Series*, Salt Lake City, Utah, Mar. 13-16, 1995, pp. 11-13.

[Szymanski and Hinton 1995] T. Szymanski and H. S. Hinton, ”Design of a terabit free-space photonic backplane for parallel computing,” *Proc. 2nd International Conference on Massively Parallel Processing using Optical Interconnections (MPPOI 95)*, San Antonio, TX, USA, Oct 23-24, 1995, pp. 16-27.

[Szymanski and Supmonchai 1996] T. H. Szymanski and B. Supmonchai, ”Reconfigurable computing with optical backplanes – an economic argument for optical interconnects,” *Proc. 3rd International Conference on Massively Parallel Processing using Optical Interconnections (MPPOI 96)*, Maui, HI, USA, Oct. 27-29, 1996, pp. 321-328.

[Szymanski et al. 1998] T. H. Szymanski, A. Au, M. Lafrenière-Roula, V. Tyan, B. Supmonchai, J. Wong, B. Zerrouk, and S. T. Obenaus, ”Terabit optical local area networks for multiprocessing systems," *Applied Optics*, vol. 37, no. 2, pp. 264-275, Jan. 10, 1998.

[Tajima et al. 1998] A. Tajima, N. Kitamura, S. Takahashi, S. Kitamura, Y. Maeno, Y. Suemura, and N. Henmi, “10-Gb/s/port gated divider passive combiner optical switch with single-mode-to-multimode combiner,” *IEEE Photonics Technology Letters*," vol. 10, no. 1, pp. 162-164, Jan. 1998.

[Taveniku et al. 1996] M. Taveniku, A. Åhlander, M. Jonsson, and B. Svensson, “A multiple SIMD mesh architecture for multi-channel radar processing,” *Proc. International Conference on Signal Processing Applications & Technology, ICSPAT´96*, Boston, MA, USA, Oct. 7-10, 1996, pp. 1421-1427.

[Taveniku et al. 1998] M. Taveniku, A. Ahlander, M. Jonsson, and B. Svensson, "The VEGA moderately parallel MIMD, moderately parallel SIMD, architecture for high performance array signal processing," *Proc. 12th International Parallel Processing Symposium & 9th Symposium on Parallel and Distributed Processing (IPPS/SPDP 98)*, Orlando, FL, USA, Mar. 30 - Apr. 3, 1998, pp. 226-232.

[Tanenbaum 1996] A. S. Tanenbaum, *Computer Networks*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1996, ISBN 0-13-349945-6.

[Teitelbaum 1998] K. Teitelbaum, ”Crossbar tree networks for embedded signal processing applications,” *Proc. Massively Parallel Processing using Optical Interconnections (MPPOI 98)*, Las Vegas, NV, USA, June 15-17, 1998, pp. 200-207.

[Thinking Machines 1991] *The Connection Machine CM-200 Series Technical Summary*. Thinking Machines, Corp., June 1991.

[Tolmie and Renwick 1993] D. Tolmie and J. Renwick, "HIPPI: simplicity yields success," *IEEE Network*, pp. 28-32, Jan. 1993.

[Tomaševic and Milutinovic 1994] M. Tomaševic and V. Milutinovic, "Hardware approaches to cache coherence in shared-memory multiprocessors, Part 1," *IEEE Micro*, vol. 14, no. 5, pp. 52-59, Oct. 1994.

[Tomaševic and Milutinovic 1994B] M. Tomaševic and V. Milutinovic, "Hardware approaches to cache coherence in shared-memory multiprocessors, Part 2," *IEEE Micro*, vol. 14, no. 5, pp. 61-66, Oct. 1994.

[Tooley 1996] F. A. P. Tooley, ”Optically interconnected electronics – challanges and choices,” *Proc. Massively Parallel Processing using Optical Interconnections (MPPOI 96)*, Maui, HI, USA, Oct. 27-29, 1996, pp. 138-145.

[Uchida 1997] N. Uchida, "Hardware of VX/VPP300/VPP700 series of vector-parallel supercomputer systems," *Fujitsu Scientific and Technical Journal*, vol. 33, no. 1, pp. 6-14, June 1997.

[USC 1997] *POLO Technical Summary*. University of Southern California, Oct. 1997.

[Varma and Raghavendra 1994] A. Varma and C. S. Raghavendra, Eds., *Interconnection Networks for Multiprocessors and Multicomputers: Theory and Practice.* IEEE Computer Society Press, Los Alamitos, CA, USA, 1994, ISBN 0-8186-4972-2.

[Weber et al. 1997] W.-D. Weber, S. Gold, P. Helland, T. Shimizu, T. Wicki, and W. Wilcke, "The Mercury interconnect architecture: a cost-effective infrastructure for high-performance servers," *Proc. 24th International Symposia on Computer Architecture (ISCA'97),* Denver, CO, USA, June 2-4, 1997.

[Wong and Yum 1994] P. C. Wong and T.-S. P. Yum, "Design and analysis of a pipeline ring protocol," *IEEE Transactions on communications*, vol. 42, no. 2/3/4, pp. 1153-1161, Feb./Mar./Apr. 1994.

[Wong et al. 1995] Y.-M. Wong, D. J. Muehlner, C. C. Faudskar, D. B. Buchholz, M. Fishteyn, J. L. Brandner, W. J. Parzygnat, R. A. Morgan, T. Mullally, R. E. Leibenguth, G. D. Guth, M. W. Focht, K. G. Glogovsky, J. L. Zilko, J. V. Gates, P. J. Anthony, B. H. Tyrone, Jr., T. J. Ireland, D. H. Lewis, Jr., D. F. Smith, S. F. Nati, D. K. Lewis, D. L. Rogers, H. A. Aispain, S. M. Gowda, S. G. Walker, Y. H. Kwark, R. J. S. Bates, D. M. Kuchta, J. D. Crow, "Technology development of a high-density 32-channel 16-Gb/s optical data link for optical interconnection applications for the optoelectronic technology consortium (OETC)," *Journal of Lightwave Technology*, vol. 13, no. 6, pp. 995-1016, June 1995.

[Yatagai et al. 1996] T. Yatagai, S. Kawai, and H. Huang, "Optical computing and interconnects," *Proceedings of the IEEE*, vol. 84, no. 6, pp. 828-852, June 1996.

[Yayla et al. 1998] G. I. Yayla, P. J. Marchand, and S. C. Esener, "Speed and energy analysis of digital interconnections: comparison of on-chip, off-chip, and free-space technologies," *Applied Optics*, vol. 37, no. 2, pp. 205-227, Jan. 10, 1998.

[Yu and Bhattacharya 1997] C.-C. Yu and S. Bhattacharya, "Dynamic scheduling of real-time messages over an optical network," *Proc. of the Sixth International Conference on Computer Communications and Networks (IC^{3}N 97),* Las Vegas, NV, USA, Sept. 22-25, 1997, pp. 336-339.

[Zhao et al. 1995] C. Zhao, T.-H. Oh, and R. T. Chen, "General purpose bidirectional optical backplane: high-performance bus for multiprocessor systems," *Proc. 2nd International Conference on Massively Parallel Processing using Optical Interconnections (MPPOI'95),* San Antonio, TX, USA, Oct 23-24, 1995, pp. 188-195.

[Zhao et al.1996] C. Zhao, J. Liu, and R. T. Chen, ”Hybrid optoelectronic backplane bus for multiprocessor-based computing systems,” *Proc. 3rd International Conference on Massively Parallel Processing using Optical Interconnections (MPPOI '96)*, Maui, HI, USA, Oct. 27-29, 1996, pp. 313-320.