M. Jonsson, "Fiber-optic interconnections in high-performance real-time computer systems," Licentiate Thesis Summary, Technical Report CCA, Halmstad University, Sweden, May 1997.

Fiber-Optic Interconnections in High-Performance Real-Time Computer Systems

By

Magnus Jonsson

Submitted to the School of Electrical and Computer Engineering,
Chalmers University of Technology in partial fulfillment of the requirements for the degree of Licentiate of Engineering.

Department of Computer Engineering
Chalmers University of Technology
S-412 96 Göteborg
Sweden

ISBN 91-71997-494-6, Göteborg 1997

Preface

My work concerning the use of fiber-optic interconnections in computer systems was initiated already in my master thesis work at the Centre for Computer Systems Architecture (CCA), Halmstad University, finished in September 1994 and supervised by Kenneth Nilsson. I was then, in October 1994, registered as a doctoral candidate at the Department of Computer Engineering (CE), Chalmers University of Technology. In January 1995, my employment at Halmstad University was converted to the form of a doctoral candidate and the work towards this thesis could really begin.

The work reported in this thesis has been part of two projects: (i) the REMAP project, financed by NUTEK, the Swedish National Board for Industrial and Technical Development, and (ii) the PARAD project, financed by the Swedish Ministry of Education in cooperation with Ericsson Microwave Systems AB (EMW). Both CCA and CE have taken parts in the two projects.

I want to express my gratitude to my supervisor, Professor Bertil Svensson, especially for encouraging my work. A special thank goes to my co-authors: Klas Börjesson, Magnus Legardt, Kenneth Nilsson, Bertil Svensson, Mikael Taveniku, and Anders Åhlander. I also want to thank all other people at CCA, CE, and EMW that have fed this work with valuable support and discussions.

Last but not least, I direct a very big thank to my wife Eva, for her support and understanding all these late nights of work.

Magnus Jonsson

Abstract

Future parallel computer systems for embedded real-time applications, where each node in itself can be a parallel computer, are predicted to have very high bandwidth demands on the interconnection network. Other important properties are time-deterministic latency and guarantees to meet deadlines. In this thesis, a fiber-optic passive optical star network with a medium access protocol for packet switched communication in distributed real-time systems is proposed. By using WDM (Wavelength Division Multiplexing), multiple channels, each with a capacity of several Gb/s, are obtained.

A number of protocols for WDM star networks have recently been proposed. However, the area of real-time protocols for these networks is quite unexplored. The protocol proposed in this thesis is based on TDMA (Time Division Multiple Access) and uses a new distributed slot-allocation algorithm with real-time properties. Services for both guarantee-seeking messages and best-effort messages are supported for single destination, multicast, and broadcast transmission. Slot reserving can be used to increase the time-deterministic bandwidth, while still having an efficient bandwidth utilization due to a simple slot release method.

By connecting several clusters of the proposed WDM star network by a backbone star, thus forming a star-of-stars network, we get a modular and scalable high-bandwidth network. The deterministic properties of the network are theoretically analyzed for both intra-cluster and inter-cluster communication, and computer simulations of intra-cluster communication are reported. Also, an overview of high-performance fiber-optic communication systems is presented.

List of Appended Papers

This thesis contains a thesis summary, followed by five technical papers. The first paper is a tutorial overview of high-performance fiber-optic communication systems [Paper A], while the other four are refereed research papers published in conference proceedings [Paper B - Paper E]. Related, but not appended, papers are: a paper [Jonsson et al. 1995], which is an earlier version of [Paper B]; a paper [Taveniku et al. 1996], corresponding to [Paper C], but more focused on algorithm mapping instead of fiber-optic communication; and two papers, targeted for journal publication, summarizing the work reported in this thesis, one focused on protocol design and analysis [Jonsson 1997], and the other focused on network and system aspects [Jonsson et al. 1997].

Appended papers

[Paper A] M. Jonsson, "High-performance fiber-optic communication networks for distributed computing systems," *Research Report CCA – 9709, Centre for Computer Systems Architecture (CCA), Halmstad University, Sweden*, Apr. 1997.

[Paper B] M. Jonsson, K. Nilsson, and B. Svensson, "A fiber-optic interconnection concept for scaleable massively parallel computing," *Proc. Massively Parallel Processing using Optical Interconnections (MPPOI'95)*, San Antonio, TX, USA, Oct. 23-24, 1995, pp. 313-320.

[Paper C] 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)*, Maui, HI, USA, Oct. 27-29, 1996, pp. 85-93.

[Paper D] 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.

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

Other related papers

[Jonsson et al. 1995] M. Jonsson, K. Nilsson, and B. Svensson, "Fiber-optic interconnections in massively parallel systems," *Proc. Sixth Swedish Workshop on Computer System Architecture (DSA'95)*, Stockholm, Sweden, June 1-2, 1995, pp. 51-52.

[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.

[Jonsson 1997] M. Jonsson, "WDM star network for high performance distributed real-time systems," *submitted for reviewing*.

[Jonsson et al. 1997] M. Jonsson, B. Svensson, M. Taveniku, and A. Åhlander, "Embedded supercomputing using fiber-optic interconnections," *to be submitted for reviewing*.

Contents

Thesis Summary

1. Introduction

Fiber-optic communication systems with multiple channels obtained through the use of WDM (Wavelength Division Multiplexing) has reached the commercial stage in the area of telecommunication. In the future, the WDM technique is foreseen to be applicable also in local multiple-access networks. At the same time, the bandwidth demands in parallel and distributed computing systems are increasing due to new evolving applications. Fiber-optic multiple-channel communication systems can give new perspectives to these systems.

In this thesis, the development of a high-performance fiber-optic multiple-channel network for distributed real-time systems is reported. The network architecture used is either a star-of-stars configuration or a single-star configuration. Each cluster (or single-star network) employs the WDM star configuration where transmitted data on all optical wavelength channels are broadcasted to all other nodes in the cluster (single-star network) by a passive optical star. Each node in the cluster (single-star network) is connected to the star by two fibers, one for each direction. All clusters in the star-of-stars network is connected to a backbone cluster via electronic gateway-nodes (Figure 1). The network is described on the conceptual level in [Paper B] [Jonsson et al. 1995], where also a point-to-point linked network with an electronic star, in the form of a true crossbar, was proposed as future alternative to the passive optical star.

Massively parallel embedded computer systems, where parallelism may be exploited in each node itself, are specially targeted. A typical system is the radar signal processing system described in [Paper C] [Taveniku et al. 1996], where each module consists of a SIMD (Single Instruction stream Multiple Data streams) computer and a network interface. In this way, a MIMSIMD (Multiple Instruction Streams for Multiple SIMD arrays) computer system is formed. Other applications where the MIMSIMD architecture with a high-performance interconnection network might be required are described in [Davis et al. 1992] [Svensson and Wiberg 1993]. Although the network is targeted for massively parallel processing systems, it can be used in other high-performance systems too.

In a real-time system, the correct functioning of the system depends on the time when a result is produced as well as the correctness of the result [Stankovic 1988]. In many real-time systems, timing must be guaranteed in order to avoid life-threatening situations. An example is the automatic

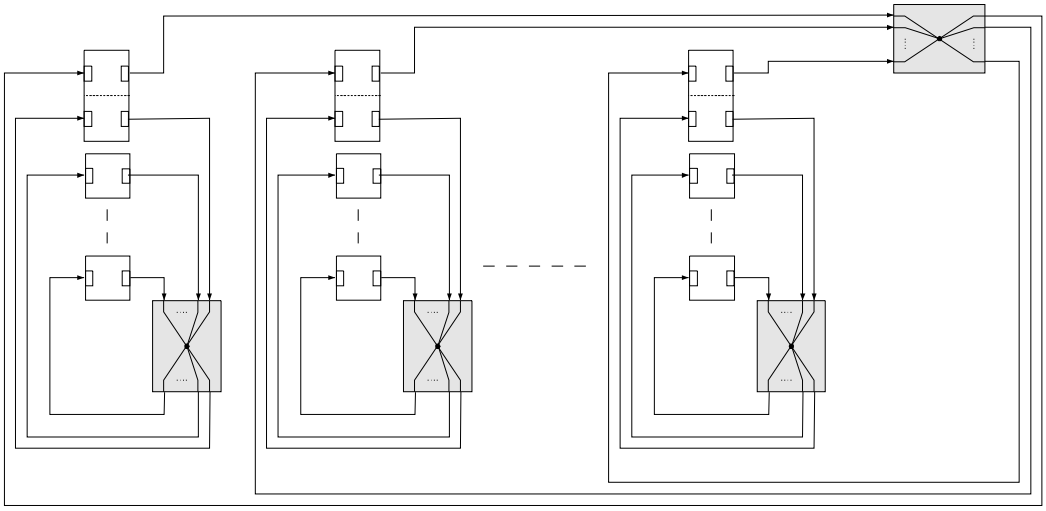

Figure 1: Multiple passive optical stars topology.

brake system in a car. In distributed real-time systems, the interconnection network plays a very important role to fulfill the required system functioning.

A number of protocols for WDM star networks have been proposed. However, the area of real-time protocols for these networks is quite unexplored, with a few exceptions. The Rainbow network, described in [Dono et al. 1990] [Janiello et al. 1993], and the *N*-DT-WDMA protocol, described in [Humblet et al. 1993], support guaranteed bandwidth for circuit switched and virtual circuit switched traffic, respectively. However, the bandwidth utilization is reduced when there is no traffic on an established connection, because the bandwidth cannot be reused by other nodes. The I-TDMA [Sivalingam et al. 1992] and I-TDMA* [Bogineni et al. 1993] protocols are other examples where guaranteed bandwidth cannot be dynamically reused. These two protocols are multiple-channel extensions to static uniform TDMA.

A real-time protocol for packet switched communication in WDM star networks is described in [Yan et al. 1996]. The QOS (quality of service) associated with a real-time packet in the network is related to the probability of missing its deadline. The protocol tries to globally minimize the number of packets not managing their QOS, by adaptively changing the priority of queued packets. Although dynamic real-time properties are supported, the matter of success or not of a packet transmission depends on the global state of the network, and transmission success can not be guaranteed in advance.

In this thesis, a medium access protocol for packet switched real-time communication in WDM star networks is proposed. The protocol is called TD-TWDMA (Time-Deterministic Time and Wavelength Division Multiple Access) and supports guaranteed real-time services for both single destination, multicast (messages destined to more than one node but not to all), and broadcast (messages destined to all nodes) transmission. Slot reservation is also supported, while an efficient bandwidth utilization is obtained due to a simple slot release method. This dynamic protocol is developed from the conceptual thoughts, described in [Paper B], of combining dynamic and static (off-line) scheduling. The deterministic properties of the protocol are theoretically analyzed for intra-cluster communication [Paper C] [Paper E], while computer simulations show the performance for general traffic patterns [Paper E]. In [Paper C], it is also shown how a network with the TD-TWDMA protocol can be used in a massively parallel radar signal processing system.

By the use of electronic gateway nodes we retain the popular WDM star network architecture in each cluster, for which cheap components can be foreseen to appear in the future. With electronic gateway nodes we also achieve wavelength reuse in each cluster and in the backbone, and the TD-TWDMA protocol can be used separately in each cluster and in the backbone. The deterministic properties for inter-cluster communication are analyzed in [Paper D].

Other hierarchical WDM star networks include the wavelength-flat (all nodes share the same wavelength space) tree-of-stars network [Dowd et al. 1993], the tree-of-stars network (called LIGHTNING) that has wavelength routing elements between each level [Dowd et al. 1996], and the multiple star network where each node is directly connected to both a local star and a remote star [Ganz and Gao 1992B]. The star-of-stars network proposed in this thesis can be seen as a two-level tree-of-stars network.

In each cluster of the proposed network, fixed-wavelength transmitters and tunable receivers are used (Figure 2). A fixed unit is always tuned to one and the same wavelength channel, while a tunable unit can be tuned to an arbitrary wavelength channel. Each transmitter has a specific wavelength (home channel) and the network architecture can be described as FT^1-TR^1 using the classification scheme given in [Mukherjee 1992]. FT^1-TR^1 stands for one Fixed Transmitter and one Tunable Receiver per node. Other FT-TR networks are described in [Dono et al. 1990] [Bracket 1991], and general information on WDM star networks is found in [Bracket 1990] [Mukherjee 1992] [Mestdagh 1995]. Components for WDMA networks are reviewed in [Green 1993]. The receivers are tunable over the whole range of channels used in the cluster. This makes the cluster a single-hop network where any receiver can be reached by any transmitter in a single hop [Mukherjee 1992] [Ramaswami 1993]. The main reason why FT^1-TR^1 is chosen is the naturally

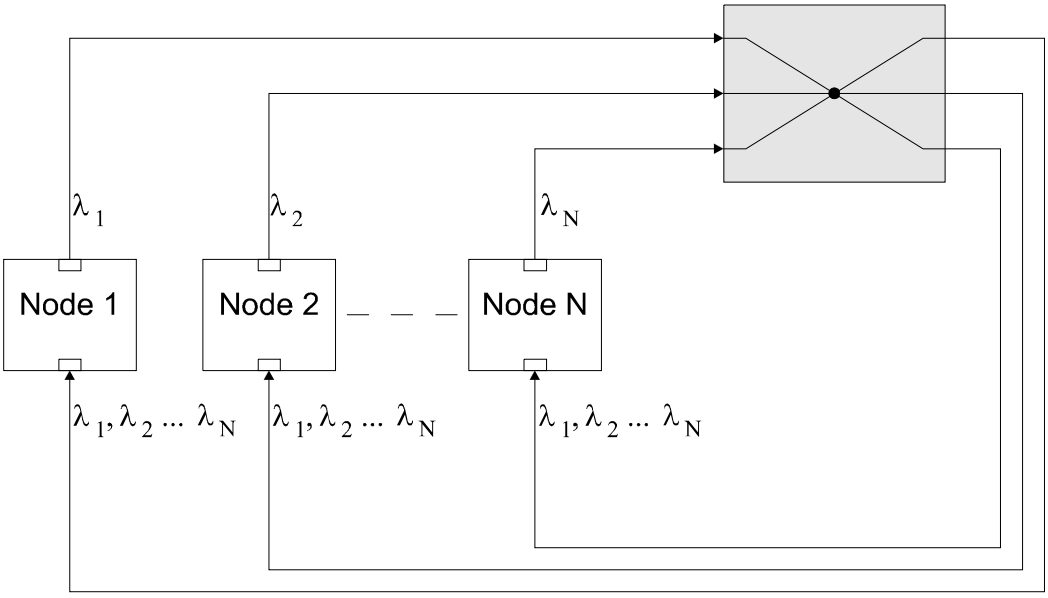

Figure 2: Passive optical star network with fixed transmitters and tunable receivers.

embedded broadcast function obtained when all receivers are tuned to the same transmitter channel.

A 100-channel WDM system has been demonstrated [Toba et al. 1993] and systems with 1000 channels are possible [Wailes and Meyer 1991]. However, the practical limit in number of wavelengths in the kind of networks reported in this paper is expected to be somewhere between 16 and 32 [Bracket 1996]. This translates into star-of-stars networks of maximum sizes between 256 and 1024 nodes, gateway-nodes included.

In dynamic distributed real-time systems, messages may be classified into two categories [Arvind et al. 1991]: best effort messages and guarantee-seeking messages. While best effort messages normally have soft deadlines, such that the system need only try its best to meet the deadlines, guarantee-seeking messages have harder timing constraints. If the communication system cannot guarantee the timing constraints of a guarantee-seeking message, the owner of the message should be made aware of it immediately.

In a network using the TD-TWDMA protocol, real-time services, in the form of guarantee-seeking and best effort messages, are supported for both single destination, multicast, and broadcast transmission. By the use of TDMA (Time Division Multiple Access) the access to each channel is divided into cycles of time-slots. Each node has a number of guaranteed slots to support

guarantee-seeking messages. However, if there are no such messages queued in a node the slots will be released for best effort messages from other nodes (or the same node) according to a predetermined scheme. This is a simple method to obtain an efficient bandwidth utilization and at the same time have time-deterministic bandwidth. A node can increase its time-deterministic bandwidth, used for guarantee-seeking messages, by slot reserving. The slot release method is also used for the reserved slots. The implementation of real-time services is described in [Paper E].

The main function of the TD-TWDMA protocol is to allocate time-slots for either guarantee-seeking messages or best effort messages. The allocation is done using a deterministic distributed slot-allocation algorithm. The algorithm temporarily changes the predetermined scheme according to the current slot demands from each node. These slot demands are transmitted in advance on the same channels as the data is transmitted on, and contain information about which guaranteed slots to keep and which to release. This type of WDM networks without a separate control channel are denoted as *non-control channel* based networks. Networks where a separate control channel is used to reserve access to the data channels are denoted as *control channel* based networks. Other non-control channel based networks are found in [Dono et al. 1990] [Ganz and Koren 1991] [Ganz and Gao 1992], while control channel based networks are found in [Habbab et al. 1987] [Chen et al. 1990] [Chipalkatti et al. 1992] [Bogineni and Dowd 1992]. FatMAC, presented in [Sivalingam and Dowd 1995], is another protocol for non-control channel based WDM star networks where the access to the channels is divided into a control phase and a data phase. However, FatMAC has no support for real-time services. Tutorial overviews covering high-capacity fiber-optic networks for computer communication are found in [Paper A] [Acampora and Karol 1989] [Green 1991] [Ramaswami 1993].

The main contributions reported in this thesis can be summarized as: (i) development of a scalable fiber-optic communication concept supporting high-bandwidth dynamic real-time traffic, (ii) a WDM star network protocol with deterministic features for packet switched communication in real-time computer systems, (iii) analysis and improvement of the deterministic properties in a star-of-stars network with electrically separated WDM star clusters, and (iv) development of mechanisms that implement guaranteeing real-time services for the proposed network.

The rest of the thesis summary is organized as follows. The protocol is presented in the second section, while inter-cluster communication is described in Section 3. Some of the simulation results are presented in Section 4. Section 5 is a conclusion and summary.

2. The TD-TWDMA protocol

Because each transmitter in a cluster (or single-star network) has a specific wavelength, the number of wavelengths in the cluster, C, hereafter denoted as channels, equals the number of nodes in the cluster, M, i.e., $C = M$. The transmitter and receiver parts of the transceiver are independent and can work concurrently. The description of the protocol that we will give is for the single-star network but holds also for a cluster or the backbone in the star-of-stars network.

There are $2M$ queues in each transmitter, M for best effort messages and M for guarantee-seeking messages. For each of the two types of messages, one queue is for broadcast and $M - 1$ queues are for single destination messages (one for each of the other nodes). The broadcast queues are used for both true broadcast messages and for multicast messages. We define the size of an entry in the queues as that of a packet, i.e., a part of a message corresponding to one slot.

In the next subsection the function of the cycles is explained. In 2.2, the distributed slot-allocation algorithm is presented, followed by 2.3 where the implementation of real-time services is described. In 2.4, slot reserving is discussed.

2.1 Transmitter and receiver cycles

Because every transmitter has its own home-channel the only conflict that can appear is when two or more nodes want to transmit to the same node at the same instant. To prevent conflicts, slots in the network are therefore allocated in the receiver cycles where each slot will have a specific owner. This allocation is done by the distributed slot-allocation algorithm.

There is also one cycle running in each transmitter, but only to tell when and to whom the node is allowed to transmit. The transmitter cycle reflects the slots that the node owns in the receiver cycle of every other node (exemplified in Figure 3). In the receiver cycles, shown in the upper table in the figure, each slot in each receiver cycle can only be assigned to one transmitter at the same time. The lower table shows how the slots of a transmitter cycle are built up by copying its entries in the corresponding slots in each receiver cycle.

In Figure 4 it is shown how a *receiver* cycle with S slots is partitioned into data slots and control slots. Each node m_i, $1 \le i \le M$, is assigned one of the M control slots where it broadcasts control information to all other nodes m_j, $1 \le j \le M$ and $j \ne i$. The control slots are therefore identically assigned in every receiver cycle. When control slots have been gathered from all other

Receiver cycles	*Data slots*											
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Node 1	2	2	3	4	3	2	3	4	4	2	3	4
Node 2	1	3	3	3	4	4	3	4	1	1	3	4
Node 3	1	4	4	4	1	2	1	4	1	1	2	4
Node 4	1	2	1	1	2	2	3	2	1	2	3	3

⇓

	Data slots											
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Transmitter cycle in node 1	2 3 4	-	4	4	3	-	3	-	2 3 4	2 3	-	-

Figure 3: The transmitter cycle, lower table, is filled by taking all owned slots in the receiver cycles, upper table, in all other nodes. Note that a multicast is possible in Slots 1, 9, and 10.

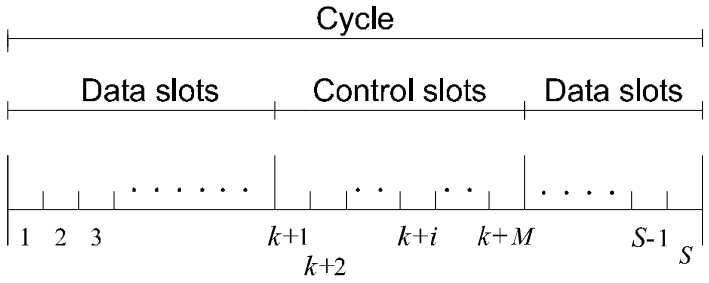

Figure 4: A receiver cycle is partitioned into data slots and control slots.

nodes, allocation of the data slots in the *next* cycle can be calculated using the distributed slot-allocation algorithm described in the next subsection.

To reduce the latency, the control slots are placed as late as possible in the cycle. However, there must be time for the distributed slot-allocation algorithm to be calculated before the beginning of the next cycle.

2.2 Distributed slot allocation algorithm

The TD-TWDMA protocol consists of three steps:

1. Each node transmits a control slot

Receivers	***Data slots***											
Priority	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***
1: High	-	2	3	4	-	2	3	4	-	2	3	4
1: Low	*2*	*2*	*2*	*2*	*3*	*3*	*3*	*3*	*4*	*4*	*4*	*4*
2: High	1	-	3	4	1	-	3	4	1	-	3	4
2: Low	*3*	*3*	*3*	*3*	*4*	*4*	*4*	*4*	*1*	*1*	*1*	*1*
3: High	1	2	-	4	1	2	-	4	1	2	-	4
3: Low	*4*	*4*	*4*	*4*	*1*	*1*	*1*	*1*	*2*	*2*	*2*	*2*
4: High	1	2	3	-	1	2	3	-	1	2	3	-
4: Low	*1*	*1*	*1*	*1*	*2*	*2*	*2*	*2*	*3*	*3*	*3*	*3*

Figure 5: Allocation scheme for the receiver cycles in a four-node system.

2. Each node separately runs the distributed slot-allocation algorithm

3. The nodes transmit and receive data slots

How the contents of the control slots is calculated and how incoming messages and buffers are handled, partly determines the real-time services and is described in the next subsection. When describing the distributed slot-allocation algorithm it is assumed, for simplicity, that all broadcasted control slots are received before the algorithm starts. However, this is not a requirement in a real implementation because, for each slot, the outcome of the distributed slot-allocation algorithm only depends on the control slot information from one node.

The slot-allocation algorithm is based on a predetermined allocation scheme that can be partly overloaded. As an example, the allocation scheme for a four-node system is shown in Figure 5. We do not call this scheme "reservation", because that term is used when describing the overloading of the scheme. Only data slots are shown and even if the control slots might be in the middle of the cycle we here assume they have index 13 to 16.

The total number of slots in a cycle is set to $S = M^2$ and the number of data slots is $M(M - 1)$. Each pair of rows represents one receiver cycle, where each number is the index of the transmitter that owns the corresponding slot. The high-priority row is the default scheme, but if the high-priority slot owner does not need the slot it is temporarily released for the current cycle, i.e., the cycle where the data should have been sent. This is done by a release message contained in the owner's control slot. The low-priority owner will then get the slot. If neither the high-priority nor the low-priority

owner needs the slot, it will be unused. This is the cost of having a *simple* algorithm. However, compared to a static TDM system an *efficient* bandwidth utilization is achieved due to the slot-release method.

Since each node independently can perform the computations of the slot-allocation scheme we call it a distributed algorithm. No extra latency is required to return the result of the algorithm, which had been the case if, e.g., a master had calculated it.

2.3 Real-time services

The description of how guarantee-seeking and best effort messages are passed through the transmitter includes the following parts:

1. Treatment of arriving guarantee-seeking messages
2. Treatment of arriving best effort messages
3. The moment of transmitting a control slot
4. The moment when all control slots have been received
5. Action at the beginning of a data slot

The ability to give deadline guarantees to a message relies on knowing when we have guaranteed slots in the forthcoming cycles. This is achieved by having a matrix that holds the number of high-priority slots, belonging to the node, in each of a number of forthcoming cycles next to the current one. For each cycle, one element in the matrix is dedicated to each type of guarantee-seeking messages (corresponding to the buffer for that type), i.e., the matrix is two-dimensional. Broadcast and multicast do not have to be separately treated here since the protocol does not allow the default scheme to be changed (by reservation) to have high-priority multicast to less than all other nodes.

By scanning the matrix, we can see if there are enough high-priority slots for the actual kind of an arriving message, before deadline, to guarantee the message to be sent in time. If not, the message will be rejected immediately and the owner will have time to handle the situation. If, instead, a guarantee can be given, each element in the matrix corresponding to required slots is decremented by the number of required slots for that element, i.e., the sum of all decrements equals the number of packets in the message. In the case of a single destination message with too few available slots, the broadcast elements can also be scanned. Packets are always put in

the guarantee-seeking queue corresponding to the element in the matrix that was decremented, and in order of transmission to ease reassembling of the message.

Best effort messages might be transmitted in order according to, e.g., the *earliest deadline first* algorithm, but in the scope of this thesis they are assumed to be transmitted in order of arrival. An arriving best effort message is hence just put in the right best effort destination queue. Multicast messages are put in the broadcast queue.

When it is time for a node to send the control slot, all packets in the queues for guarantee-seeking messages are counted. All high-priority slots in the next cycle that are not needed for buffered guarantee-seeking packets are released, except for slots with broadcast capability which are given to best effort broadcast and multicast messages if there are any buffered.

When all control slots have been received, the allocation scheme for the next cycle is temporarily modified to contain either the high-priority or the low-priority owner for each slot to each receiver, depending on if the slot was released or not.

At the beginning of each data slot in the next cycle, the modified allocation scheme is used to determine, in the transmitter, which queue to take a packet from for transmission, and, in the receiver, which channel to tune in for reception. By the definition of the protocol, there will always be a packet for transmission in the addressed queue when it is a guarantee-seeking queue or the best effort broadcast queue. However, the best effort single destination queues might be empty. In that case an empty packet is sent, only telling the receiver that there was nothing to send.

2.4 Slot reserving

A node can reserve slots to increase its guaranteed bandwidth. A maximum of $M(M-2)$ slots per cycle, slot 5 to 12 in the example in Figure 5, are allowed to be reserved. Slots 1 through M are not allowed to be reserved and M slots are control slots. When reserving slots, the corresponding high-priority entry (in the predetermined allocation scheme) in the receiver cycle or cycles if broadcast is used are exchanged with the index of the reserving node. To keep complexity down in the transmitters the corresponding slots in all receivers must be reserved (i.e., for broadcast) if the slots are to be used for multicast messages.

The assignment of reserved slots can be changed during run-time either by higher layers, by a development system, or by having slot-assignment schemes for several working modes stored in the nodes. If the reservation is

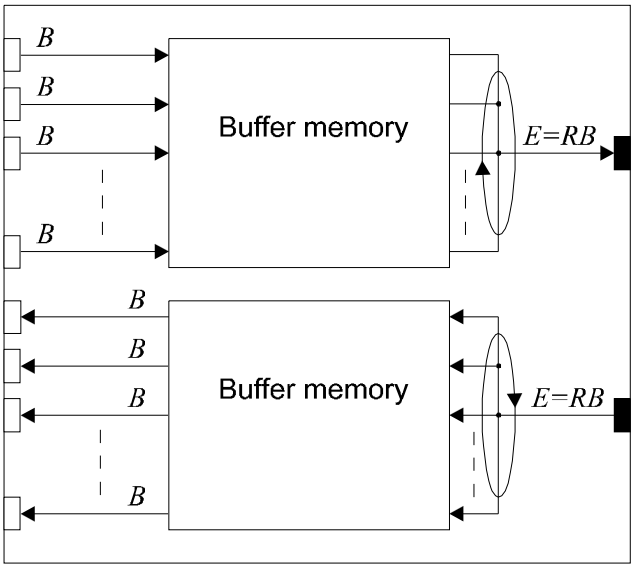

Figure 6: Gateway node with higher channel-bandwidth on the backbone side (right side).

done by higher layers, a reservation request to the nodes in question is sent in a control slot together with the ordinary contents of the control slot.

3. Inter-cluster communication

Real-time services are implemented in a similar way for inter-cluster communication as for intra-cluster communication. However, the larger number of nodes sharing the backbone slots must be considered. An analysis of this slot sharing in the backbone is reported in [Paper D], where also synchronization issues in star-of-stars systems are presented.

A gateway node in the star-of-stars network shown in Figure 1 contains network interfaces to both the backbone star and its dedicated cluster. There are buffer memories for both upward and downward traffic. The size of the buffer memories is significantly larger than the possible traffic in one cycle. Status information on the buffers is always sent together with the other information in the control slots.

In some systems, it is desirable that the bandwidth per channel in the backbone be higher than in the clusters. The increase in bandwidth may be implemented either by a higher bit rate or by having several wavelengths per channel. If R is the ratio of backbone channel-bandwidth, E, to cluster channel-bandwidth, B (i.e., $R = E/B$), then the gateway nodes (Figure 6) are designed to have R transceiver modules each on the cluster side, in order to

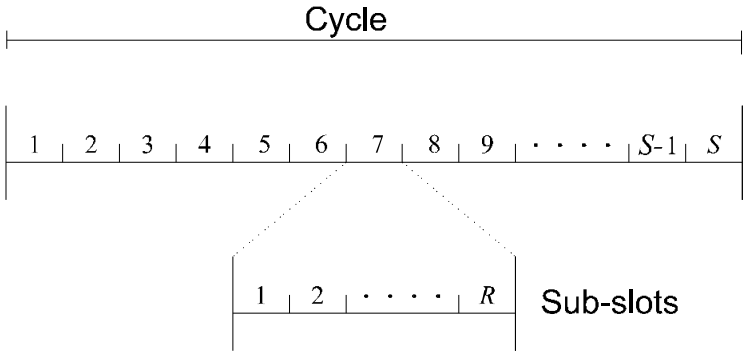

Figure 7: Each backbone slot is divided into R sub-slots with the same pair of gateway nodes as source and destination.

achieve the same aggregated bandwidth as on the backbone side. Also, a gateway node has R dedicated home-channels on the cluster side, one for each transmitter.

The cycle length in the backbone is always the same as that in the clusters, both when measured in time and when measured in number of slots. Therefore, the number of bits in a backbone slot is R times higher than that in a cluster slot. Each backbone slot is divided into R sub-slots, all with the same pair of gateway nodes as source and destination (Figure 7). However, the sub-slots can have different pairs of end-nodes. A sub-slot has the same number of bits as a cluster slot, which makes the design of the gateway nodes easier. Also, the worst-case latency may decrease when using sub-slots if several slots in the same source cluster and with the same destination cluster can be packed together.

4. Performance

The average performance of the network has been analyzed through computer simulations for general traffic. Single star networks with 8, 16, and 32 nodes were simulated. Other assumptions were:

- A gap of one data slot between the last control slot and the next cycle was assumed. To give an example of a real system, the slot duration was set to 1 μs.
- All guarantee-seeking messages have a deadline 5 ms from the moment of generation.
- Uniform traffic was assumed, i.e., all nodes had equal probability of message generation and uniformly distributed destination addresses.

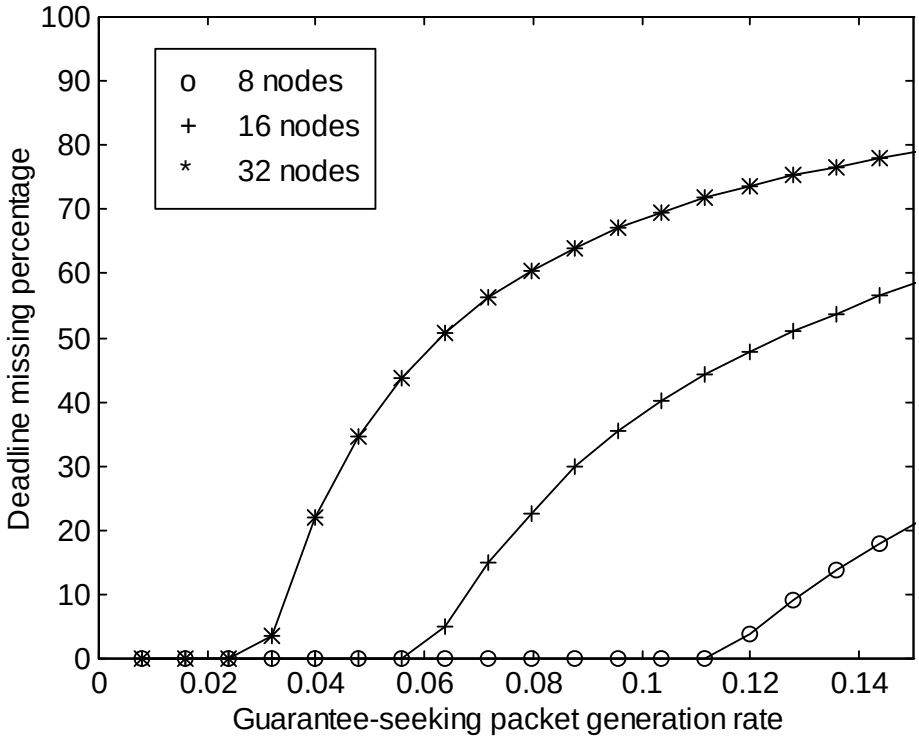

Figure 8: Real-time performance plotted as fraction of messages that miss their deadlines, versus traffic intensity.

Messages were generated according to a Poisson process and all messages were of single-destination type.

- Message lengths were exponentially distributed between 1 and 10 slots (discrete number of slots), with a length of 1 slot as the highest probability.
- For simplicity, the propagation delay was neglected and no slot reservation was done.
- Infinite queue lengths were assumed.
- Packet generation rate is set to the message generation rate through the mean message length.
- Latency is defined as the time elapsed from the arriving moment of a message until the last packet of the message leaves the transmitter.

First, the deadline missing percentage versus packet generation rate of guarantee-seeking messages is plotted in Figure 8. At moderate traffic

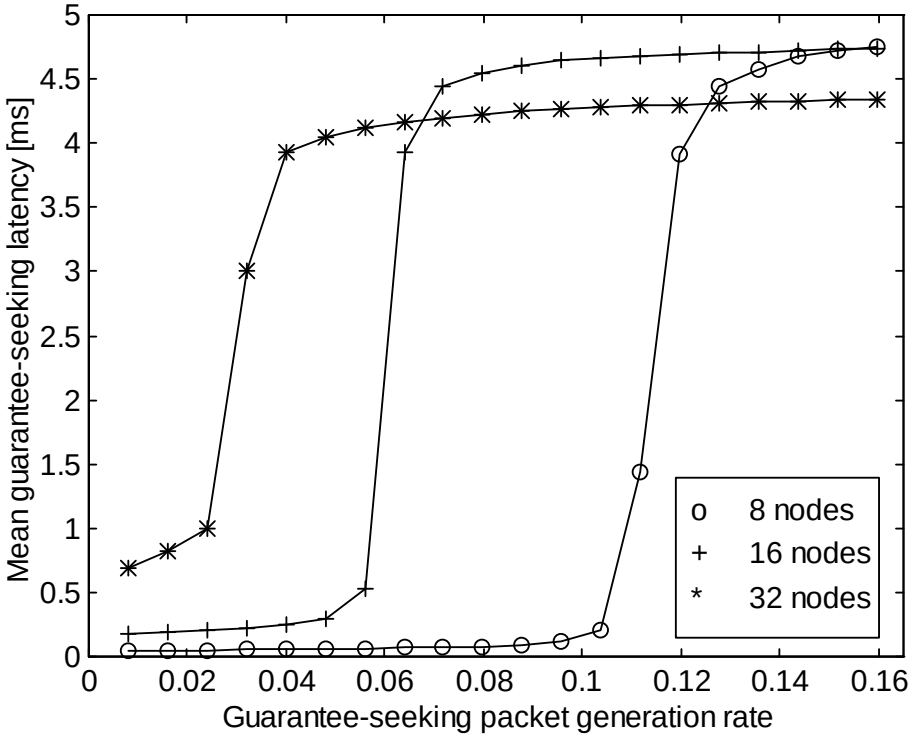

Figure 9: Latency for guarantee-seeking messages plotted versus traffic intensity.

intensities, no messages miss their deadlines for the given assumptions. The deterministic bandwidth fraction is $(M - 1)/M^2$. This gives 10.9, 5.9, and 3.0 percent for 8, 16, and 32 nodes, respectively. As shown in the figure, guarantee-seeking messages begin to be partly rejected around these values. The plot is independent of the packet generation rate of best effort messages, since guarantee-seeking messages are always prioritized.

In Figure 9, the mean latency for guarantee-seeking messages is plotted against packet generation rate. Again, the performance is independent of the amount of best effort traffic. Because messages that can not be guaranteed to meet their deadlines are discarded, the guarantee-seeking latency is upper bounded. The latency is rather uniform at low traffic intensities, but starts to grow earlier for larger networks because of the lower fraction of deterministic bandwidth.

The plot in Figure 10 shows the latency for best effort messages versus total packet generation rate (guarantee-seeking plus best effort). The traffic consisted of 10 % guarantee-seeking messages and 90 % best effort

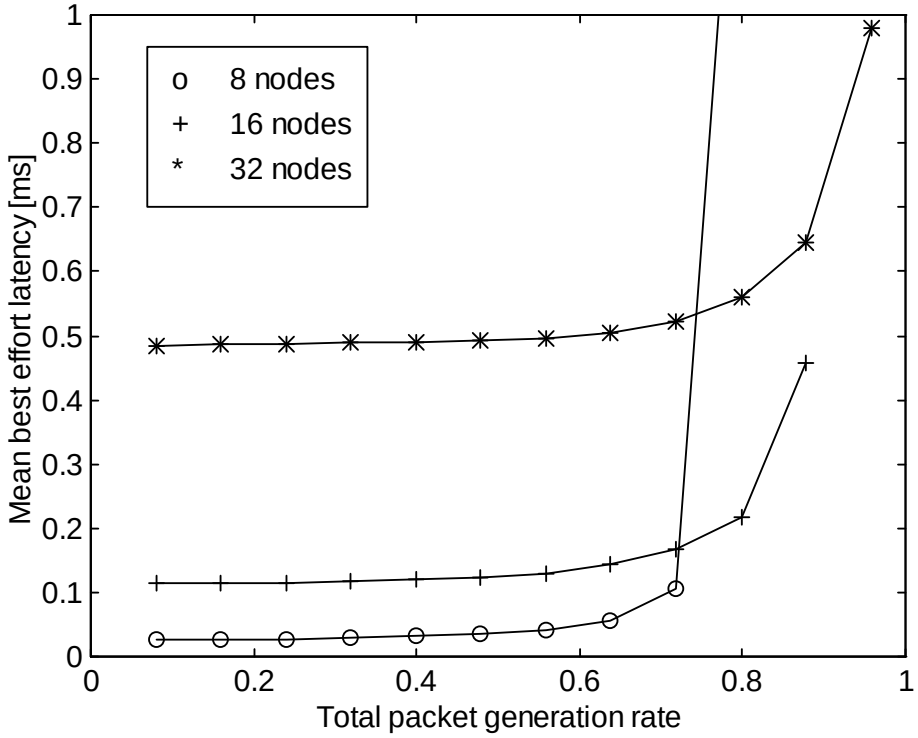

Figure 10: Mean latency for best effort messages in a network with 10 % guarantee-seeking and 90 % best effort messages.

messages. The fraction of the bandwidth available for data packets is $(M-1)/M$, which gives 87.5, 93.8, and 96.9 percent for 8, 16, and 32 nodes respectively. However, when the guarantee-seeking traffic gets saturated, some of the total number of generated packets are discarded guarantee-seeking messages. Below saturation, the best effort latency is almost uniform.

The mean best effort latency is plotted again in Figure 11, but versus the bandwidth utilization. The same ratio between guarantee-seeking messages and best effort messages was used. The plot looks very similar to the previous best effort latency plot in Figure 10. This is a consequence of the almost linear relationship between packet generation rate and bandwidth utilization below saturation.

5. Conclusions

The proposed network, with the TD-TWDMA real-time protocol, contributes to the rather unexplored area of WDM networks and protocols for

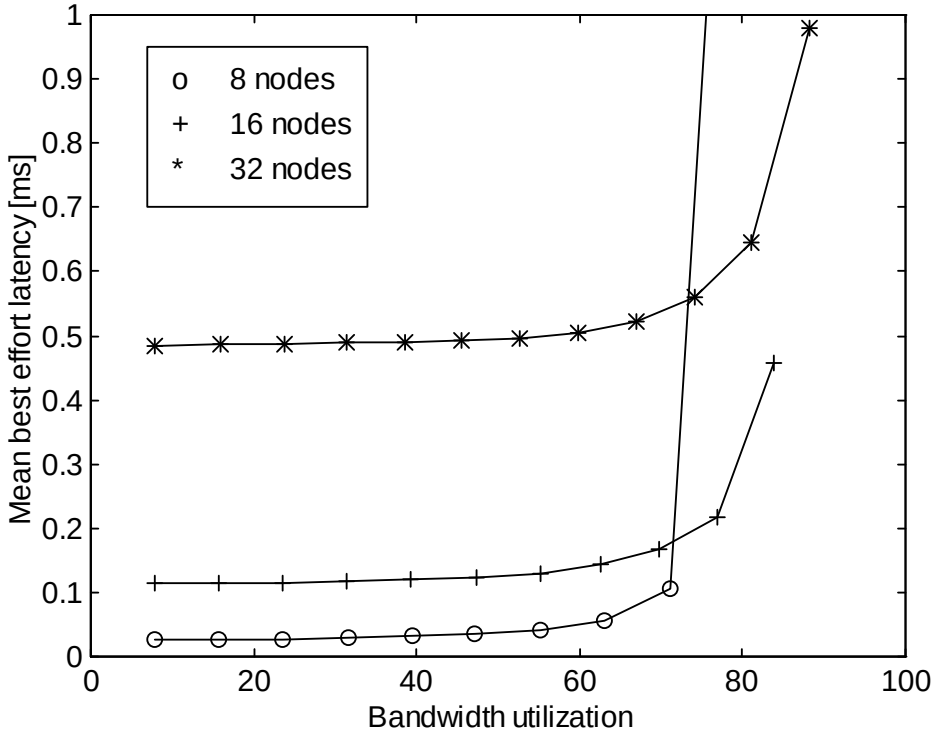

Figure 11: Mean latency for best effort messages plotted versus bandwidth utilization. The traffic consisted of 10 % guarantee-seeking and 90 % best effort messages.

distributed real-time systems. The network supports deadline guarantees for packet switched traffic, which is an important feature in dynamic hard real-time systems.

6. 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.

[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.

[Bogineni and Dowd 1992] K. Bogineni and P.W. Dowd, "A collisionless multiple access protocol for a wavelength division multiplexed star-coupled configuration: architecture and performance analysis," *Journal of Lightwave Technology*, vol 10, no. 11, pp. 1688-1699, Nov 1992.

[Bogineni et al. 1993] K. Bogineni, K. M. Sivalingam, and P. W. Dowd, "Low-complexity multiple access protocols for wavelength-division multiplexed photonic networks," *IEEE Journal on Selected Areas in Communications*, vol. 11, no. 4, pp. 590-604, May 1993.

[Bracket 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.

[Bracket 1991] C. A. Brackett, "On the capacity of multiwavelength optical-star packet switches," *IEEE LTS*, pp. 33-37, May 1991.

[Bracket 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.

[Chen et al. 1990] M. Chen, N. R. Dono, and R. Ramaswami, "A media-access protocol for packet-switched wavelength division multiaccess metropolitan area networks," *IEEE Journal on Selected Areas in Communications*, vol. 8, no. 6, pp. 1048-1057, Aug. 1990.

[Chipalkatti et al. 1992] R. Chipalkatti, Z. Zhang and A. S. Acampora, "High speed communication protocols for optical star coupler using WDM," *Proc. INFOCOM'92*, Florence, Italy, May 1992, pp. 2124-2133.

[Davis et al. 1992] E. W. Davis, T. Nordström, and B. Svensson, "Issues and applications driving research in non-conforming massively parallel processors," *Proc. of the New Frontiers, a Workshop of Future Direction of Massively Parallel Processing*. Scherson Ed., McLean, Virginia, pp. 68-78, 1992.

[Dono et al. 1990] N. R. Dono, P. E. Green, K. Liu, R. Ramaswami, and F. F. Tong, "A wavelength division multiple access network for computer communication," *IEEE Journal on Selected Areas in Communications*, vol. 8, no. 6, pp. 983-994, Aug. 1990.

[Dowd et al. 1993] P. W. Dowd, K. K. Bogineni, A. Aly, and J. A. Perreault, "Hierarchical scalable photonic architectures for high-performance processor interconnection," *IEEE Transactions on Computers*, vol. 42, no. 9, pp. 1105-1120, Sept. 1993.

[Dowd et al. 1996] P. Dowd et al., "LIGHTNING network and systems architecture," *Journal of Lightwave Technology*, vol. 14, no. 6, pp. 1371-1387, June 1996.

[Ganz and Gao 1992] A. Ganz and Y. Gao, “A time-wavelength assignment algorithm for a WDM star network,” *Proc. INFOCOM’92*, Florence, Italy, May, 1992, pp. 2144-2150.

[Ganz and Gao 1992B] A. Ganz and Y. Gao, "Traffic scheduling in multiple WDM star systems," *Proc. IEEE International Conference on Communications (ICC’92)*, Chicago, IL, USA, June 1992, pp. 1468-1472.

[Ganz and Koren 1991] A. Ganz and Z. Koren, “WDM passive star - protocols and performance analysis,” *Proc. INFOCOM’91*, Bal Harbour, FL, USA, Apr. 1991, pp. 991-1000.

[Green 1991] P. E. Green, “The future of fiber-optic computer networks,” *Computer*, pp.78-87, Sept. 1991.

[Green 1993] P. E. Green, *Fiber Optic Networks*. Prentice-Hall, Inc., 1993, ISBN 0-13-319492-2.

[Habbab et al. 1987] I. M. I. Habbab, M. Kavehrad, and C. W. Sundberg, "Protocols for very high-speed optical fiber local area networks using a passive star topology," *Journal of Lightwave Technology*, vol. 5, no. 12, pp. 1782-1794, Dec. 1987.

[Humblet et al. 1993] P. A. Humblet, R. Ramaswami, and K. N. Sivarajan, "An efficient communication protocol for high-speed packet-switched multichannel networks," *IEEE Journal on Selected Areas in Communications*, vol. 11, no. 4, pp. 568-578, May 1993.

[Janiello et al. 1993] F. J. Janiello, R. Ramaswami, and D. G. Steinberg, "A prototype circuit-switched multi-wavelength optical metropolitan-area network," *Proc. IEEE International Conference on Communications (ICC’92)*, Geneva, Switzerland, May 1993, pp. 818-823.

[Jonsson et al. 1995] M. Jonsson, K. Nilsson, and B. Svensson, “Fiber-optic interconnections in massively parallel systems,” *Proc. Sixth Swedish Workshop on Computer System Architecture (DSA’95)*, Stockholm, Sweden, June 1-2, 1995, pp. 51-52.

[Mestdagh 1995] D. J. G. Mestdagh, *Fundamentals of Multiaccess Optical Fiber Networks*. Artech House, Inc., 1995, ISBN 0-89006-666-3.

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

[Ramaswami 1993] R. Ramaswami, “Multiwavelength lightwave networks for computer communication,” *IEEE Communications Magazine*, vol. 31, no. 2, pp. 78-88, Feb. 1993.

[Sivalingam et al. 1992] K. M. Sivalingam, K. Bogineni, and P. W. Dowd, "Pre-allocation media access control protocols for multiple access WDM photonic networks," *Proc. ACM SIGCOMM'92*, Baltimore, MD, USA, Aug. 17-20, 1992, pp. 14-26.

[Sivalingam and Dowd 1995] K. M. Sivalingam and P. W. Dowd, "A multilevel WDM access protocol for an optically interconnected multiprocessor system," *Journal of Lightwave Technology*, vol. 13, no. 11, pp. 2152-2167, Nov. 1995.

[Stankovic 1988] J. A. Stankovic, "Misconceptions about real-time computing," *Computer*, vol. 21, no. 10, pp. 10-19, Oct. 1988.

[Svensson and Wiberg 1993] B. Svensson and P.-A. Wiberg, "Autonomous systems demand new computer system architectures and new development strategies," *Proc. of the 19th Annual Conference of the IEEE Industrial Electronics Society (IECON '93)*, Maui, Hawaii, USA, Nov. 15-19, 1993, pp. 27-31.

[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.

[Toba et al. 1993] H. Toba, K. Nakanishi, K. Oda, K. Inoue, and T. Kominato, "A 100-channel optical FDM six-stage in-line amplifier system employing tunable gain equalizers," *IEEE Photonics Technology Letters*," vol. 5, no. 2, pp. 248-251, Feb. 1993.

[Wailes and Meyer 1991] T. S. Wailes and D. G. Meyer, "Multiple channel architecture: a new optical interconnection strategy for massively parallel computers," *Journal of lightwave technology*, vol. 9, no. 12, pp. 1702-1716, Dec. 1991.

[Yan et al. 1996] A. Yan, A. Ganz, and C. M. Krishna, "A distributed adaptive protocol providing real-time services on WDM-based LAN's," *Journal of Lightwave Technology*, vol. 14, no. 6, pp. 1245-1254, June 1996.

Abstracts of Appended Papers

Paper A: High-Performance Fiber-Optic Communication Networks for Distributed Computing Systems

Interconnection networks have a key role in distributed processing systems of today but to follow the evolution in processing power, new technologies are needed. Multiple-channel fiber-optic communication can solve the emerging demand of bandwidth. New protocols must, however, be used to coordinate the use of multiple high-speed channels. This paper describes protocols and fiber-optic network architectures that are considered to be useful as flexible interconnection networks in future parallel and distributed computing systems. Also, a short overview of system components and system technologies is presented.

Paper B: A Fiber-Optic Interconnection Concept for Scaleable Massively Parallel Computing

One of the most important features of interconnection networks for massively parallel computer systems is scaleability. The fiber-optic network described in this paper uses both wavelength division multiplexing and a configurable ratio between optics and electronics to gain an architecture with good scaleability. The network connects distributed modules together to a huge parallel system where each node itself typically consists of parallel processing elements. The paper describes two different implementations of the star topology, one uses an electronic star and fiber optic connections, the other is purely optical with a passive optical star in the center. The medium access control of the communication concept is presented and some scaleability properties are discussed involving also a multiple-star topology.

Paper C: Time-Deterministic WDM Star Network for Massively Parallel Computing in Radar Systems

In massively parallel computer systems for embedded real-time applications there are normally very high bandwidth demands on the interconnection network. Other important properties are time-deterministic latency and services to guarantee that deadlines are met. In this paper we analyze how these properties vary with the design parameters for a passive optical star network, specifically when used in a massively parallel radar signal processing system. The aggregated bandwidth and computational power of the radar system are approximately 45 Gb/s and 100 GOPS, respectively.

The analysis is focused on the medium access control protocol, called TD-TWDMA, for the time and wavelength multiplexed network. It is concluded that the proposed network is very well suited to this kind of signal-processing applications. We also present a new distributed slot-allocation algorithm with real-time properties.

Paper D: **On Inter-Cluster Communication in A Time-Deterministic WDM Star Network**

Future real-time applications requiring massively parallel computer systems also put high demands on the interconnection network. By connecting several WDM star clusters by a backbone star, forming a star-of-stars network, we get a modular high-bandwidth network. In this paper we show how to achieve time-deterministic packet switched communication in such networks, even for inter-cluster communication. An analysis of how the deterministic latency and node bandwidth vary with design parameters is presented. We also propose a general clock-synchronization scheme, improving the worst-case latency with up to 33 percentages.

Paper E: **Dynamic Time-Deterministic Traffic in a Fiber-Optic WDM Star Network**

A number of protocols for WDM (Wavelength Division Multiplexing) star networks have been proposed. However, the area of real-time protocols for these networks is quite unexplored. In this paper, a real-time protocol, based on TDM (Time Division Multiplexing), for fiber-optic star networks is presented. By the use of WDM, multiple Gb/s channels are achieved. Services for both guarantee-seeking messages and best-effort messages are supported for single destination, multicast, and broadcast transmission. Slot reserving can be used to increase the time-deterministic bandwidth, while still having an efficient bandwidth utilization due to a simple slot release method. The deterministic properties of the protocol are analyzed and simulation results presented.