THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

High Performance Fiber-Optic Interconnection Networks for Real-Time Computing Systems

By

Magnus Jonsson

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

Centre for Computer Systems Architecture
Halmstad University
Box 823
SE-301 18 Halmstad
Sweden

Göteborg, November 1999
Preface and Acknowledgement

My work concerning fiber-optic network architectures and protocols for such networks 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. Recently after that, I was registered as a doctoral candidate at the Department of Computer Engineering (CE), Chalmers University of Technology, and the work towards a Ph.D. thesis could start. The whole work was performed at Halmstad University but with support in various ways from Chalmers.

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, first financed by the Swedish Ministry of Education in co-operation with Ericsson Microwave Systems AB (EMW), later by the KK Foundation in co-operation with EMW. Both CCA and CE have taken part in the two projects.

First, I want to express my gratitude to my wonderful wife Eva for endless patience and support. Of course, I also want to mention my two children, Marcus and Erica, who were born during this Ph.D. work. Many thanks go to my supervisor, Professor Bertil Svensson, especially for encouraging my work. A special thank you goes to my co-authors: Carl Bergenhem, Klas Börjesson, Magnus Legardt, Kenneth Nilsson, Jörgen Olsson, Bertil Svensson, Mikael Taveniku, and Anders Åhlander. I also want to thank all other people at CCA (Lars Bengtsson, Stefan Lund, Per-Arne Wiberg, and many more), CE (Håkan Forsberg, Jan Jonsson, Henrik Lönn, and others), and EMW (Richard Larsson, Joachim Strömbergson, Per Söderstam, and others) that have fed this work with valuable support, comments, and discussions.

I also express my gratitude to those (not mentioned above) for whom I have supervised master or bachelor thesis work in the scope of the projects: Henrik Arleving, Ali Jamshid Far, Bassam Mahmoud, and Andreas Sundgren. OPTOBUS and ECL components for the building of a prototype have been sponsored by Motorola and National Semiconductor, respectively.

Last, I want to encourage curious people who have in mind taking a Ph.D. It’s a lot of work but it gives experiences well worth it.
Parallel and distributed computing systems become more and more powerful and hence place increasingly higher demands on the networks that interconnect their processors or processing nodes. Many of the applications running on such systems, especially embedded systems applications, have real-time requirements and, with increasing application demands, high-performance networks are the hearts of these systems. Fiber-optic networks are good candidates for use in such systems in the future.

This thesis contributes to the relatively unexplored area of fiber-optic networks for parallel and distributed real-time computer systems and suggests and evaluates several fiber-optic networks and protocols. Two different technologies are used in the networks, WDM (Wavelength Division Multiplexing) and fiber-ribbon point-to-point links. WDM offers multiple channels, each with a capacity of several Gbit/s. A WDM star network in which protocols and services are efficiently integrated to support different kinds of real-time demands, especially hard ones, has been developed. The star-of-stars topology can be chosen to offer better network scalability.

The WDM star architecture is attractive but its future success depends on components becoming more commercially mature. Fiber-ribbon links, offering instead an aggregated bandwidth of several Gbit/s, have already reached the market with a promising price/performance ratio. This has motivated the development and investigation of two new ring networks based on fiber-ribbon links. The networks take advantage of spatial bandwidth reuse, which can greatly enhance performance in applications with a significant amount of nearest downstream neighbor communication. One of the ring networks is control channel based and not only has support for real-time services like the WDM star network but also low level support for, e.g., group communication.

The approach has been to develop network protocols with support for dynamic real-time services, out of time-deterministic static TDMA systems. The focus has been on functionality more than pure performance figures, mostly on real-time features but also on other types of functionality for parallel and distributed systems. Worst-case analyses, some simulations, and case studies are reported for the networks. The focus has been on embedded supercomputer applications, where each node itself can be a parallel computer, and it is shown that the networks are well suited for use in the radar signal processing systems studied. Other application examples in which these kinds of networks are valuable are distributed multimedia systems, satellite imaging and other image processing applications.
Contents

Preface and Acknowledgement ................................................................. 3
Abstract ............................................................................................................. 5
Contents............................................................................................................. 7
List of Figures ................................................................................................. 11
List of Tables................................................................................................... 15

Part I: Introduction and Tutorial Surveys ................................................. 17

1. Introduction ............................................................................................ 19
   1.1 Contributions.................................................................................... 19
   1.2 Published Papers and Research Reports......................................... 21
   1.3 Parallel Computers .......................................................................... 25
   1.4 Real-Time Communication .............................................................. 25
   1.5 Optical Interconnections in Parallel Computers ............................ 27
   1.6 Disposition of the Thesis.................................................................. 29

2. High Performance Fiber-Optic Networks.............................................. 31
   2.1 Introduction ...................................................................................... 31
   2.2 Multiple Access Methods.................................................................. 31
   2.2.1 Wavelength Division Multiple Access....................................... 31
   2.2.2 Time Division Multiple Access .................................................. 33
   2.2.3 Subcarrier Division Multiple Access ......................................... 33
   2.2.4 Code Division Multiple Access................................................... 34
   2.2.5 Discussion................................................................................... 34
   2.3 Components for WDMA Networks .................................................. 35
   2.3.1 Tunable Receivers...................................................................... 36
   2.3.2 Tunable Transmitters................................................................ 38
   2.3.3 Other WDM Components........................................................... 40
   2.4 Interconnection Architectures ......................................................... 41
   2.4.1 Point-to-Point Linked Networks ............................................... 41
   2.4.2 All-Optical Multi-Access Networks .......................................... 43
   2.5 Classifications of WDMA Networks ................................................ 46
   2.5.1 Broadcast-and-Select and Wavelength Routing ....................... 46
   2.5.2 Single-Hop and Multihop Networks ......................................... 47
   2.5.3 Networks with and without Control Channel .......................... 47
   2.5.4 Receiver and Transmitter Tunability ....................................... 48
   2.5.5 Random-Access Protocols........................................................... 49
   2.5.6 Time-Wavelength Assignment Schemes................................... 49
   2.6 Single-Hop WDMA Star Networks and Protocols........................... 50
   2.6.1 FOX............................................................................................. 53
   2.6.2 Lambdanet ................................................................................. 53
   2.6.3 DT-WDMA .................................................................................. 53
   2.6.4 A Distributed Adaptive Protocol ............................................... 54
Part I: Interconnection Systems

2.6.5 Interleaved TDMA ................................................................. 54
2.6.6 FatMAC .................................................................................. 55
2.7 Conclusions.................................................................................. 55

3. Interconnections in Parallel Computers........................................ 57
3.1 Design and Performance Parameters......................................... 57
3.2 Static Networks.......................................................................... 61
3.3 Dynamic Networks................................................................. 63
3.4 Shared Medium Networks ....................................................... 65
3.5 Hybrid Networks...................................................................... 66
3.6 Group Communication.............................................................. 66
3.7 Routing...................................................................................... 67
3.8 High Performance Networks for Coarse Grained Parallel
Computers.................................................................................... 68

Part II: Review of Radar Signal Processing Systems and Suitable
Interconnection Networks.......................................................... 71

4. A Sample Radar Signal Processing System................................. 73
4.1 Computation Module Design.................................................... 74
4.2 System Design Issues............................................................... 74
4.3 Inter-Module Communication Demands................................. 75

5. Configurations and Requirements of Signal Processing Systems... 79
5.1 Pure Pipeline Chain................................................................. 79
5.2 Same Program Multiple Data................................................... 80
5.3 Parallel Hardware For Multiple Modes................................. 81
5.4 Concurrent Modes on the Same Hardware............................ 82
5.5 One Data Cube per Processing Element................................. 83
5.6 Traffic Types......................................................................... 84
5.7 System Sizes and Communication Distances.......................... 84

6.1 Fiber-Ribbon Ring Network.................................................... 85
6.2 WDM Star Network................................................................. 86
6.3 WDM Ring Network............................................................... 89
6.4 Integrated Fiber and Waveguide Solutions............................ 90
6.5 Optical Interconnections and Electronic Crossbars............... 93
6.6 Systems with Smart Pixel Arrays......................................... 94
6.7 Optical and Optoelectronic Switch-Fabrics.......................... 96
6.8 Planar Free Space Optics......................................................... 97
6.9 Free Space Optical Backplanes............................................. 97
6.10 Summary .............................................................................. 99

Part III: Proposed Network Architectures and Protocols ........103

7. FT-TR WDM Star Network....................................................... 105
7.1 Protocol Overview and Related Networks............................. 106
7.1.1 Circuit Switched Networks ...................................................... 106
7.1.2 Packet Switched Real-Time WDM Networks ......................... 106
7.1.3 A New Protocol with Real-Time Support .............................. 107
7.1.4 Protocol and Real-Time Features ............................................ 109
7.2 Protocol Description ....................................................................... 110
7.2.1 Transmitter and Receiver Cycles ............................................ 112
7.2.2 Distributed Slot Allocation Algorithm ................................. 113
7.3 Real-Time Services ......................................................................... 116
7.3.1 Arriving Messages .................................................................... 116
7.3.2 Control Slots ............................................................................. 117
7.3.3 Data Slots ................................................................................. 118
7.3.4 Slot Reserving .......................................................................... 118
7.4 Implementation Aspects ................................................................ 119
7.4.1 Clock Synchronization ............................................................. 119
7.4.2 Clock Recovery ......................................................................... 119
7.4.3 Computational Complexity ...................................................... 120
7.4.4 Electronic Stars ........................................................................ 120
7.5 Deterministic Performance ............................................................ 121
7.6 Case Study ...................................................................................... 123
7.7 Simulation Results ......................................................................... 124
7.8 Summary ........................................................................................ 127
8. WDM Star-of-Stars Network ............................................................ 129
8.1 Network Architecture and Protocol ............................................... 129
8.2 Deterministic Performance ............................................................ 131
8.3 Clock Synchronization Aspects...................................................... 135
8.4 Summary ........................................................................................ 139
9. Control Channel Based Fiber-Ribbon Pipeline Ring Network ........... 141
9.1 Network Overview .......................................................................... 142
9.2 Related Networks ........................................................................... 143
9.3 The CC-FPR Protocol ..................................................................... 144
9.4 User Services .................................................................................. 148
9.4.1 Real-Time Virtual Channels.................................................... 149
9.4.2 Guarantee Seeking Messages .................................................. 150
9.4.3 Best Effort Messages ............................................................... 151
9.4.4 Barrier Synchronization .......................................................... 151
9.4.5 Global Reduction ...................................................................... 152
9.4.6 Low Level Support for Reliable Transmission........................ 153
9.5 Implementation Aspects ............................................................... 153
9.6 Case Study ...................................................................................... 155
9.7 Performance Analysis ................................................................. 157
9.8 Summary ........................................................................................ 163
10. Packet and Circuit Switched Fiber-Ribbon Pipeline Ring Network .. 165
10.1 Circuit Switched Traffic ................................................................. 165
10.2 Packet Switched Traffic ................................................................. 167
10.3 Circuit Establishment ................................................................... 167
List of Figures

Figure 1: Three passive optical network architectures: (a) ring, (b) (dual) bus, and (c) star..........................................................44

Figure 2: Folded bus. ............................................................................................................................................................45

Figure 3: 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..........................................................61

Figure 4: Omega network for an eight-node system. One path through the network is highlighted ..........................................................64

Figure 5: Possible states of a $2 \times 2$ switch. ..........................................................................................................................64

Figure 6: A rearrangeble Benes network .................................................................................................................................65

Figure 7: Fat-tree of switches where nodes are leaves in the tree. ..........65

Figure 8: Computation module with eight $8 \times 8$ meshes of PEs. ..........74

Figure 9: The MIMD system enables a functional decomposition and allocation of the program to the modules. Each function or block of functions is then mapped onto the processors in the computation module (CM). ..........................................................................................................................75

Figure 10: Description of one mode of operation in the 64-channel ground based radar system. ..........................................................................................................................................................................................76

Figure 11: Data flow between the modules in the signal processing chain. .................................................................................................77

Figure 12: A pure pipeline chain where one or more stages in the chain are mapped on each module..................................................................................................................................................................79

Figure 13: If the SPMD model is used, all PEs work together on one stage in the signal processing chain at a time........................................79

Figure 14: Several parallel groups of PEs, each running a pipeline chain program. ..........................................................................................................................81

Figure 15: Several parallel groups of PEs, each running an SPMD program. .................................................................................................81

Figure 16: Several PEs in parallel, where each data cube is processed by a single PE .................................................................82

Figure 17: 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 Nodes 3, 4, and 5..............................................................................................................................................85

Figure 18: Control channel based network built up with fiber-ribbon point-to-point links. ........................................................................86
Figure 39: Gateway node with higher channel bandwidth on the backbone side (right side)........................................................................................................... 130

Figure 40: Each backbone slot is divided into $R$ sub-slots with the same pair of gateway nodes as source and destination........................................ 131

Figure 41: Worst-case latency when full slot reservation by other nodes is assumed. The x-axis represents the number of ordinary nodes and the slot length is assumed to be $\gamma = 1.0 \mu s$.......................................................... 133

Figure 42: Worst-case latency when no slot reservation is assumed. The x-axis represents the number of ordinary nodes and the slot length is assumed to be $\gamma = 1.0 \mu s$......................................................................... 134

Figure 43: Node bandwidth, when full slot reservation by other nodes is assumed, in number of high priority slots per total number of slots in a cycle. The x-axis represents the number of ordinary nodes.............. 136

Figure 44: Node bandwidth, when no slot reservation is assumed, in number of high priority slots per total number of slots in a cycle. The x-axis represents the number of ordinary nodes........................ 137

Figure 45: With the synchronization scheme used, incoming traffic to a gateway node can be forwarded immediately except for the case of internal delay in the gateway node....................................................... 138

Figure 46: Worst-case latency comparison between usage and no usage of the proposed synchronization scheme. No slot reservation is assumed. The x-axis represents the number of ordinary nodes and the slot length is assumed to be $\gamma = 1.0 \mu s$..................................................................... 139

Figure 47: (a) Bi-directional fiber-ribbon link. (b) Unidirectional ring network built up of $M/2$ bi-directional links........................................... 141

Figure 48: The role of being slot initiator is cyclically repeated. Each of the $M$ nodes is the slot initiator in one slot per cycle.............................. 144

Figure 49: The node succeeding the slot initiator initiates the control packet transmission........................................................................................ 145

Figure 50: In each slot, a node passes/transmits one control packet and one data packet, where the control packet is used for the arbitration of the next slot................................................................. 146

Figure 51: A control packet contains a start bit, a link reservation field, and a destination field................................................................. 146

Figure 52: Conceptual view of the control channel part of the transceiver................................................................. 147

Figure 53: Conceptual view of the data channel part of the transceiver.... 147

Figure 54: A control packet travels around a network with five nodes. Node 1 is the slot initiator................................................................. 148
Figure 55: An example: Node 1 sends a single destination packet to Node 3, while Node 4 sends a multicast packet to Node 5 and Node 1.

Figure 56: The bandwidth utilization depends on the ratio of the total propagation delay around the ring to the cycle length. The boxes with bold text show the link through which each slot first propagates.

Figure 57: Detailed description of the control packet contents.

Figure 58: Data flow between the modules in the radar signal processing chain.

Figure 59: One cycle of 45 slots where Slots 16 through 45 are reserved. Each number in the table indicates, for Slots 16 through 45, the owner of the corresponding link. Neighboring segments in each slot have different background shading. The slot initiators are indicated in Slots 1 through 15.

Figure 60: Slot distribution according to Cases A and B. In the example, $N_r = 2$, $N_o = 1$, and $M = 3$.

Figure 61: Worst-case latency when the only slots a node gets are those ordinary slots for which it is the slot initiator. Case A is assumed.

Figure 62: Worst-case latency when the only slots a node gets are those ordinary slots for which it is the slot initiator. Case B is assumed.

Figure 63: Comparison of worst-case latency for Cases A and B.

Figure 64: Maximum aggregated throughput. Case A is assumed.

Figure 65: Maximum aggregated throughput. Case B is assumed.

Figure 66: Example of an allocation scheme for the links in a five-node system. The slot initiators are in bold type, and different segments have different background shading.

Figure 67: Network demonstrator with two nodes. The cable bundle in the upper part of the figure contains two fiber-ribbons ending in each node’s OPTOBUS module.
List of Tables

Table 1: Experimental single-hop WDMA star networks. M is the number of nodes...............................................................................................................................50

Table 2: Non-control channel based protocols for single-hop broadcast-and-select WDMA networks. M is the number of nodes, and N is a positive integer greater than zero..................................................................................................................50

Table 3: Control channel based protocols for single-hop broadcast-and-select WDMA networks. M is the number of nodes, and N is a positive integer greater than zero..................................................................................................................51

Table 4: Other proposed single-hop WDM star networks. M is the number of nodes, and N is a positive integer greater than zero..................................................52

Table 5: Classification of system sizes and communication distances. ....84

Table 6: A feasible allocation scheme of the wavelength channels in a WDM star network for the sample radar system.................................................................88

Table 7: Performance summary of some of the networks discussed with respect to pipeline and SPMD mapping........................................................................................................99

Table 8: Performance summary of some of the networks discussed with respect to pipeline and SPMD mapping when several concurrent modes run in parallel, each on dedicated hardware.................................................................100

Table 9: Suitability in different system sizes. Empty cells in a column mean that the technology/network is not suitable for the corresponding system sizes............................................................................................................101

Table 10: Notations for descriptions of the network architecture and protocol. .................................................................................................................................111

Table 11: Additional notation when describing the star-of-stars network architecture. ......................................................................................................................129

Table 12: Node bandwidth in number of high priority slots per total number of slots in a cycle. ...........................................................................................................135
Part I: Introduction and Tutorial Surveys
1. Introduction

Fiber-optic communication systems with multiple channels obtained through the use of WDM (Wavelength Division Multiplexing) have reached the commercial stage in the area of telecommunications (see Chapter 2 for a survey of fiber-optic communication technology). In the future, the WDM technique is expected also to be applicable in local multiple-access networks. At the same time, bandwidth demands in parallel and distributed computing systems are increasing because of new, evolving applications. Fiber-optic multiple-channel communication systems can give new perspectives in these systems.

Another technology that can be very useful for the development of new parallel and distributed computing systems is the use of fiber-ribbon links. Commercially available products have appeared very quickly, and a good price/performance ratio is expected.

An area of high-performance fiber-optic networks that is relatively unexplored is networks and protocols for hard real-time systems. This field has been the main target in the work reported in this thesis. The focus has also been on case studies and special support for signal processing applications, especially radar signal processing.

The contributions of the present work are summarized in Section 1.1, while Section 1.2 states which papers the thesis is based on. Introductions to parallel computers, real-time communication, and optical interconnections in parallel computers are given in Sections 1.3, 1.4, and 1.5, respectively. The disposition of the thesis is given in Section 1.6.

1.1 Contributions

One of the motivating factors for this work was the small amount of research done on high-performance fiber-optic networks for hard real-time computer systems (we do not focus on telecommunications and do not treat, e.g., FDDI, Fibre Distributed Data Interface, networks and extensions of FDDI networks for high-performance networks in the framework of this thesis). There is a need for this kind of network in, e.g., embedded supercomputers. The radar signal processing system described in Chapter 4 is a typical example of an application and computer system which we place in the category of embedded supercomputing.

The approach has been to develop network protocols with support for dynamic real-time services, out of traditional time-deterministic static TDMA systems (see, e.g., [Kopetz and Grünsteidl 1994] [Kopetz et al. 1989]
[Nilsson et al. 1993] [Nilsson et al. 1993B] [Nilsson 1994] [Wiberg 1993] for time-deterministic communication systems for hard real-time systems, while considering emerging promising, fiber-optic technologies. The focus has been on functionality more than pure performance figures, mostly on real-time features but, later in the work, on other types of functionality for parallel and distributed systems as well. The focus of the analyses made has been on predictable performance, including case studies with feasibility tests.

The WDM star network proposed in Chapter 7, and Chapter 8 for the star-of-stars variant, was developed to fill the gap in high performance real-time networking and supports real-time traffic for dynamic hard real-time systems where each node can have a sustained output dataflow of several Gbit/s. This is offered in combination with other desired, or even required, properties such as broadcast capabilities, scalability in terms of number of nodes, and scalability in terms of aggregated system bandwidth.

Although the WDM star network represented a breakthrough, there were several reasons to continue the work:

1. High demands on the optoelectronic components. While components with low tuning-latency have been demonstrated, it might not be reasonable to expect components combining low tuning latency, large tuning range (many channels possible), and fair price to appear in the near future.

2. The appearance of fiber-ribbon links on the market, which made possible new network architectures.

3. The pipelined dataflow that is common in many signal processing applications may not always need the flexibility of the distributed crossbar that the WDM star network implements.

The network architecture of a ring, built up of fiber-ribbon point-to-point links, was chosen as an alternative to the WDM star network. With the possibility of spatial reuse of the bandwidth in the ring (referred to as a pipeline ring in this thesis), the ring offers good suitability for applications with a pipelined dataflow. At the same, it is a flexible network if a good medium access method is offered. Two fiber-ribbon pipeline ring networks have been developed in the scope of this thesis. In the first network, the ring is divided into two sub-networks, one for data traffic and one for control information (see Chapter 9). The sub-network for control traffic is used primarily for medium access information (for the arbitration of the data sub-network) but also to implement low-level support for, e.g., group communication. Compared with other networks, this network offers a great of functionality. In the second network, the ring is divided into two sub-
networks, one for circuit-switched traffic and one for packet-switched traffic (see Chapter 10).

It is contradictionary to have the highest possible performance and flexibility at the same time that availability and the cost of network components are considered. Different network solutions must be offered for different application demands. The choice of network may depend on, e.g., availability of components, scalability requirements, and traffic patterns. A broad range of high-performance networks, with special consideration to the application of radar signal processing is briefly described in Chapter 6, including the WDM star (and star-of-stars) network and the two fiber-ribbon ring networks developed in the scope of this work (described in more detail in Chapters 7 through 10).

The main contributions of this thesis work can be summarized as:

1. Development, analysis, and case study (with chosen industrial application) of a WDM star network where protocols and services are integrated in an efficient way to support different kinds of real-time demands, especially hard real-time demands, for parallel and distributed computing systems.

2. Development of two new ring network architectures for which features of fiber-ribbon links are used advantageously.

3. Development, analysis, and case study (with chosen industrial application) of protocols and services for a fiber-ribbon ring network based on a control channel. Services for different kinds of real-time demands, especially hard real-time demands, are supported. Mechanisms for spatial reuse of bandwidth and low-level support for, e.g., barrier synchronization and other kinds of group communication are also efficiently integrated into the network.

1.2 Published Papers and Research Reports

The papers on which the thesis is based can be divided into four groups: (i) papers concerning the proposed WDM star and star-of-stars networks, (ii) papers concerning the two proposed fiber-ribbon ring networks, (iii) surveys of optical networks (mainly fiber-optic) and network components, including evaluations of network architectures for signal processing systems, and (iv) papers reporting research on radar signal processing systems. The papers in the first three groups have been written mainly by me, while my work represents a minor part of the papers in the fourth group. The papers are listed by group below.
Group 1:


Initial ideas for how to achieve a WDM star network with support for both hard real-time traffic and non real-time traffic by dividing a cycle of time slots into different parts with different roles.


Shorter version of Paper A presented at a national conference.


Presentation of the TD-TWDMA protocol together with a case study of the suitability of the network in a radar signal processing system.


Design alternatives and analysis of star-of-stars networks where each star cluster follows the TD-TWDMA protocol. An interpretation, made by Borella et al., of this article is found in [Borella et al. 1998B].


Implementation details on real-time services for the TD-TWDMA network, together with simulation results.

Group 2:


Presentation of a novel network architecture where fiber-ribbon point-to-point links form a ring network. The network is divided into two sub-networks, one for circuit-switched traffic and one for packet-switched traffic. A case study of the suitability of the network in a radar signal processing system is also provided.


Presentation of a new network architecture similar to that described in Paper F. The network is called CC-FPR (Control Channel based Fiber-ribbon Pipeline Ring) and is divided into one data channel and one control channel. The CC-FPR protocol uses the control channel to coordinate the arbitration of the data channel. A case study of the suitability of the network in a radar signal processing system is also provided.


Journal version of Papers F and G.


Services in the CC-FPR network for, e.g., real-time traffic and group communication are briefly described (work to be continued beyond the scope of this thesis). A worst-case analysis is also provided for the network.


Slightly longer version of Paper I presented at a national conference.
Group 3:


Survey of fiber-optic networks and components for such networks.


Survey of fiber-optic and free-space optical network architectures. The networks are briefly evaluated according to their suitability in different configurations of signal processing systems. A tutorial part on interconnection networks for parallel computers is also included.


Conference paper in which parts of Paper L were extracted (before the report was finished) so that the paper emphasized on fiber-optic networks for signal-processing applications.

Group 4:


A MSIMD (Multiple SIMD) system for radar signal processing is proposed. The TD-TWDMA network is one part of the system proposal.

In a continuation of the work presented in Paper N, several design alternatives for an MSIMD system for radar signal processing are presented and discussed. The TD-TWDMA network and the fiber-ribbon ring network described in Paper F are design alternatives for the interconnection network in the system proposal.

1.3 Parallel Computers

Two categories, according to Flynn's classification [Flynn 1966] [Flynn 1972], into which parallel computers may be classified are MIMD (Multiple Instruction streams Multiple Data streams) [Hord 1993] and SIMD (Single Instruction stream Multiple Data streams) [Hord 1990]. MIMD computers are typically coarser grained than SIMD computers, which instead have more tightly coupled processing elements that are all controlled by the same instruction flow. 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] [Tomašević and Milutinović 1994] [Tomašević and Milutinović 1994B]. Parallelism can exist on a lower level as well (i.e., on a more fine grained level), e.g., instruction level parallelism (ILP).

Most, but not all of the interconnection networks mentioned in this thesis are more suitable for coarse-grained systems. Referring to a node in this thesis 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 an 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.

1.4 Real-Time Communication

In a real-time computer system, the correct functioning of the system depends on the time at which 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 control systems for nuclear power plants. Other real-time systems include flight control systems, radar systems, robotics, and industrial control
systems. In distributed real-time systems, the interconnection network plays a very important role in fulfilling the system functioning requirements. Often, guaranteeing real-time services is much more important than performance such as throughput or average latency in these systems. Some classifications and concepts important for the area of real-time communications are briefly described below.

Whether the traffic is periodic or aperiodic (or mixed) plays an important role in real-time communication systems. An example of periodic traffic is messages triggered by the sampling of a sensor at a regular interval. Aperiodic traffic may occur as a result of non-predictable events such as user input or a signal indicating that a dangerous situation has been observed by a sensor.

There is a design choice as to whether communication traffic (or processes) should be scheduled statically, i.e., off-line, or dynamically. Communication schemes can also be setup during run-time semi-statically, e.g., having a schedule which can be exchanged with a new one when the system changes its mode of operation. The static approach allows for easier analysis but requires more knowledge of the system behavior at the design stage. The dynamic approach, on the other hand, allows for the interconnection network to adaptively change schedule or on the basis of the current system behavior. Interconnection systems with truly dynamic features often allow for treatment of real-time requirements for separate packets or messages. This thesis focuses on networks with both semi-static and dynamic features.

Real-time systems are often classified into soft or hard real-time systems depending on how critical the (timing) correctness of their behavior is. In terms of deadlines for processes or messages, we can introduce a third term, firm [Kopetz 1997]. If, e.g., a result of a calculation or an arriving message also has utility after its deadline, the deadline is classified as soft. It is otherwise classified as firm. If the missing of a deadline may end in a catastrophe, the deadline is classified as hard. Services for real-time traffic with hard and soft deadlines are often classified into guarantee-seeking and best-effort, respectively. This work has focused on networks with support for both guarantee-seeking and best-effort traffic.

Communication networks can be divided into multiple-access networks (the focus of this thesis) and switched networks. Multiple-access networks have some form of shared medium. The access to this medium is regulated by a medium access method. In the case of real-time multiple-access networks [Malcolm and Zhao 1995], it is often the medium access method that is specially designed to support real-time traffic. Examples of multiple-access networks with some form of support for real-time traffic are: Token Bus, Token Ring, CAN (Control Area Network), FDDI, and TTP (Time-Triggered Protocol) [Kopetz 1997].
With regard to (packet-) switched real-time networks [Zhang 1995], they are here assumed to be constructed of switches interlinked with point-to-point interconnections, i.e., so called mesh networks. Most reported switched real-time networks are wide area networks and are therefore not in focus in this thesis (which is focused on short distance systems). Real-time message streams, e.g., multimedia streams, are normally supported in the form of virtual circuits or the like. The handling of these message streams in the switches during the lifetime of a connection and at setup can be done in a deterministic or a statistical way. A deterministic service guarantees a certain performance even in the worst-case, while a statistical service does not have a hundred percent guarantee for this. Instead, a statistical service might have a specified probability of packet loss and deadline or jitter violation.

1.5 Optical Interconnections in Parallel Computers

Optics have been used for a while in communication systems where optical fibers each acts as a single point-to-point channel. However, the broad field of optical interconnections does not only include traditional single-channel fiber-optics. Novel technologies allow for features such as multiple high-speed channels in a single fiber and two-dimensional arrays of optical free-space channels. Some work on comparative technology studies and classifications are reviewed below.

In arguments for optical interconnections in parallel computing systems and similar systems, some stated drawbacks or limitations of electrical interconnects are [Tooley 1996] [Yatagai et al. 1996):

- pin bottlenecks, both on the chip and the 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] [J ahns 1994] [J ahns 1998] [Lund 1997] [Ozaktas 1997B] [Stunkel 1997] [Yatagai et al. 1996]:

- large spatial and temporal bandwidth are offered
- optical beams do not affect one another
- light is immune to electromagnetic interference
• light beams do not suffer from signal frequency-dependent attenuation
• parallel global high speed interconnects are offered
• the possibility of non-planar interconnections is offered
• multiplexing can be done in multiple domains, e.g., the time and wavelength domains
• high-density interconnects where light beams can cross each other are possible
• large electrical backplane contacts can be avoided, thereby reducing the chassis size
• optical interconnects offer low energy per bit and high speed-power product
• low skew is possible

Of course, there are also arguments for not using optics [Bohr 1998]. For example, it is argued that the 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 interconnections 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 the respective areas to which they are best suited [Caulfield 1998].

A classification of more advanced optical interconnection systems can be made 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 views alternative optical interconnection architectures for parallel computing as a tree in which 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 such
as clusters of workstations. Many references to work related to free-space systems are found in [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 to in Chapter 6. More or less general discussions on the use of optical interconnections in parallel computers have also been published [Goodman et al. 1984] [Rudolph 1998] [Schenfeld 1995] [Schenfeld 1996].

1.6 Disposition of the Thesis

The dissertation is divided into three parts as described below:

- Part I: background material and tutorial surveys. The introduction is followed by two chapters on fiber-optic communication systems and interconnection networks for parallel computers, respectively.

- Part II: optical interconnection networks for signal processing systems, especially for radar signal processing systems. A brief introduction is given to the radar signal processing systems developed in the projects of which this work has been a part of. Presented in this part are both selected network architectures of interest which were described in Part I, suggestions for how to configure these and other network architectures especially for signal processing applications, and the networks that are described in more detail in Part III.

- Part III: the main contributions of the thesis work presented. Four chapters present the WDM star network, the star-of-stars network, the control channel based fiber-ribbon ring network, and the simplified fiber-ribbon ring network, respectively. The part then ends with a concluding chapter of the thesis.

Because of the disposition chosen, and the fact that Parts II and III contain enough overlapping background information, one may begin reading in Parts II or III, especially if the contents of earlier chapters are well known to the reader.
2. High Performance Fiber-Optic Networks

2.1 Introduction

Optical fibers for communication systems offer a bandwidth of more than 30 THz per fiber [Brackett 1990], giving them great potential for future computer communication networks for data intensive applications. This survey describes representative examples of high-performance fiber-optic networks as well as MAC (Medium Access Control) protocols [Rom and Sidi 1990] for those networks. WDMA (Wavelength Division Multiple Access, described below) networks and passive optical networks (PONs) are given special attention. Previous reviews of high-capacity fiber-optic networks for data communication are found in [Acampora and Karol 1989] [Chatterjee and Pawlowski 1999] [DeCusatis 1998] [Green 1991] [Modiano 1999] [Ramaswami 1993] [Senior et al. 1998]. The review given here has a focus on optical LANs [Kazovsky et al. 1994] and similar networks that can be used for flexible communication in parallel and distributed computing systems. Many proposed fiber-optic networks, although developed with a specific application in mind, can be used in a wide range of applications; thus we do not necessarily sort those networks out. Nevertheless, e.g., pure telecommunication networks are not treated here.

A review of fiber-optic network technologies, components, and classifications is presented in Sections 2.2 through 2.5. Examples of WDMA networks are described in Section 2.6. Section 2.7 gives a conclusion and summary.

2.2 Multiple Access Methods

Accessing and sharing the huge amount of optical bandwidth among the multiple nodes in a fiber-optic network is a challenging problem. Four multiple access methods are described below: WDMA, TDMA (Time Division Multiple Access), SCMA (SubCarrier Division Multiple Access), and CDMA (Code Division Multiple Access). The most popular one for achieving high aggregated bandwidth seems to be WDMA, especially in end-user systems owing to their simple transceiver designs. Additional information on multiple-access methods is found in [Gagliardi and Karp 1995].

2.2.1 Wavelength Division Multiple Access

When using WDMA, multiplexing is done in the spectral domain of the light [Hill 1989] [Kaminow 1989] [Green and Ramaswami 1990] (WDMA systems use the WDM technology, and the two terms are therefore often used interchangeably here). In this way, several optical carriers, or channels, are
implemented in the network. For systems with denser wavelength spacing than 1 nm, the WDMA technique is also referred to as optical FDM (Frequency Division Multiplexing) [Agrawal 1992]. One way to use the wavelength channels is to allow several nodes to transmit simultaneously on different channels. Data are then sent bit-serially between nodes tuned to the same channel. The channel to which each transmitter and receiver are tuned at any time is controlled by the MAC protocol. Another way is to synchronously transmit on several channels in parallel, i.e., bit-parallel byte-serial transmission [Loeb and Stilwell 1988] [Loeb and Stilwell 1990]. However, compensation for bit-skew caused by group delay dispersion (different wavelength channels travel at different speeds in the fiber) may be needed in these systems [Jeong and Goodman 1996].

The practical limit (e.g., because of “the impracticality of administering a network with a very large number of differing lasers”) in the number of wavelengths in LAN networks and similar networks is expected to be somewhere between 16 and 32 [Brackett 1996]. However, a WDM system with 100 channels has been demonstrated [Toba et al. 1990], and systems with thousand 1 Gbit/s channels are likely in the future [Wailes and Meyer 1991]. Experimental WDM demonstrations include:

- Experiments with 16 channels at 2 Gbit/s [Lin et al. 1988]. Tunable Fabry-Perot filters (see Section 2.3.1) were used for channel selection.

- A WDM experiment with 16 channels, each at 622 Mbit/s [Toba et al. 1989]. One of the channels were selected by a tunable filter consisting of Mach-Zehnder interferometers (see Section 2.3.1).

- A system with 100 channels, each carrying data at a bit rate of 622 Mbit/s [Toba et al. 1990].

- A commercially available WDM link called MuxMaster [Janniello et al. 1995]. A MuxMaster link has 20 wavelengths on a single fiber which implement 10 full-duplex protocol-independent connections. Experiments with a bit rate of 200 Mbit/s on each wavelength were reported but bit rates in the order of Gbit/s are possible. Maximum distance is 50 km.

- A 160 Gbit/s system with $8 \times 20$ Gbit/s channels and a channel spacing of 4 nm [Sorel et al. 1996].

- 1 Tbit/s aggregated capacity with 50 channels, each modulated at a rate of 20 Gbit/s [Chraplyvy et al. 1996].

32
2.2.2 Time Division Multiple Access

TDMA networks can be divided into two groups, those using *packet-interleaving* and those using *bit-interleaving*. Networks using packet-interleaving, for example [Barry et al. 1996], divide the access to the medium into time slots. The length of each slot normally equals the transmission duration of a packet. Simple MAC protocols can be used, but each node must work at the speed of the aggregated bit rate in the network.

With bit-interleaving, each node sends only one bit at a time at regular intervals. If $N$ nodes transmit simultaneously, the bit stream from one node will have a bit-to-bit interval of $N$ bit slots. The width of the optical pulse is, however, normally several times shorter than the duration of a bit slot. In a bit-interleaved TDMA network, clock synchronization must be maintained with a precision in the order of less than a bit period, which is difficult when a bit rate of multiple Gbit/s is used. Experiments with a 250 Gbit/s network were reported in [Prucnal et al. 1994]. The optical pulse width was 1 ps, the bit slot was 4 ps and the bit-to-bit interval from one node was 10 ns. This resulted in the capacity to support a 100 Mbit/s bit-stream to each of 2 500 nodes. Bit-interleaved TDMA networks and technologies to implement such networks are reviewed in [Spirit et al. 1994] [Seo et al. 1996].

2.2.3 Subcarrier Division Multiple Access

The data from each node in a SCMA network are used to modulate node-specific microwave subcarriers, i.e., multiplexing is done in the microwave frequency domain [Darcie 1987]. The subcarrier then modulates an optical carrier. All subcarriers are detected at the receiver photo-diode, but only the desired one is demodulated using conventional microwave techniques. SCMA can be contrasted with WDMA, where multiplexing is also done in the spectral domain but on the lightwave carrier, about $10^{14}$ Hz, while the subcarrier microwave frequency is at $10^8$ to $10^{10}$ Hz [Mestdagh 1995].

A typical application suitable for SCMA is distribution of analog video channels [Chang et al. 1995]. The video channels can then remain analog through the whole network, and similar FM techniques as used today can be employed. However, SCMA has also been proposed for use in data communication networks [Darcie 1987].

In [Georges and Lau 1993], an SCMA network experiment using self-pulsating CD laser-diodes was reported. By changing the pulse frequency (in the range of 2 to 4 GHz), a node can tune to another subcarrier channel. A bit rate of 150 Mbit/s was used in the experiment.
2.2.4 Code Division Multiple Access

Several CDMA methods for fiber-optic communication have been proposed. The majority of work is being done on the method having a set of orthogonal code sequences of \( F_{KLSV} \), where each chip has a value of either “1” or “0” [Salehi 1989] [Salehi and Brackett 1989] [Marhic 1993]. Each code sequence then corresponds to the destination address of a specific node. During transmission, each “1” bit is encoded with the destination node's code sequence of chips, while a “0” bit is encoded with a sequence of only “0” chips. The drawback of this scheme is that each transceiver must work at a speed \( N \) times higher than the bit rate, where \( N \) is the length of the code, i.e., the number of chips. In wireless communication systems, this method is referred to as Direct Sequence Spread Spectrum (DS-SS) [Bantz and Bauchot 1994].

Decoders to be used in the receivers can be implemented optically [Prucnal et al. 1986]. CDMA experiments using an optical decoder were reported in [Prucnal et al. 1986B]. The optical decoder consisted of \( P \) delay lines where \( P \) is the number of chips that are set to 1 in the receiving node's code sequence. By matching the delay lines to the corresponding positions of the “1” chips in the code sequence, an auto correlation function sensitive to the code sequence is obtained. A similar method was proposed for use in the transmitter. A bit rate of 3.125 Mbit/s was achieved on a 100 Mbaud fiber-optic link, using 32 chips per bit.

Instead of having the code sequence spread in time, spectral encoding can be used. In [Zaccarin and Kavehrad 1993], the light from a LED (Light Emitting Diode) is spatially split into a discrete number of spectral components by a diffraction grating. The light is then passed through a spatial amplitude mask with the code sequence and combined in the fiber. In this way, each of the spectral components is encoded with one of the chips, and the LED need only to be modulated with the bit rate. A similar technique is used in the receiver. In [Weiner et al. 1988], the code sequence is also applied in the spectral domain but with a phase mask instead of an amplitude mask.

Another CDMA method called PFDM (Pulse Frequency Division Multiplexing) is described in [Frenkel 1992]. Short laser pulses with a node specific repetition rate are used to encode the data. The received signal can be decoded with optical delay lines, with delays proportional to the pulse frequency, and optical AND gates.

2.2.5 Discussion

The different properties of the multiple-access methods make them suitable for different applications/systems. In the kind of systems in focus in this thesis, a limited number of users share the network cost (the opposite of
telecommunication backbone networks). CDMA and bit-interleaved TDMA are therefore considered too complex. When a bit rate of several Gbit/s is needed from each node in the network simultaneously, packet-interleaved TDMA will also be too complex, because every node must work at the aggregated bit rate of the network. SCMA is also excluded because of the limited network capacity [Mestdagh 1995]. What is left is WDMA, which gives easy access to the channels when tunable components reach reasonable costs, and where each node need only work at the speed of its own bit rate.

Several proposed network architectures use WDMA in combination with one of the other multiple access methods, for example, TDMA [Jonsson et al. 1995] and SCMA [Shankaranarayanan et al. 1991]. As an example, data packets can be carried on wavelength channels while network control information is carried on subcarrier channels [Chiang et al. 1996]. These kinds of hybrids can give new, valuable properties to networks, for example, flexible real-time properties when combining TDMA and WDMA [Jonsson et al. 1996].

2.3 Components for WDMA Networks

A fiber-optic communication system typically consists of a transmitter, a receiver, and some form of transmission medium based on optical fibers. Common optoelectronic components used in the transmitter are LEDs and laser diodes [Ettenberg 1998], where laser diodes can be divided into multimode and single-mode laser diodes. Single-mode laser diodes have narrower spectral widths than multimode laser diodes by the incorporation of a grating filter inside or outside the cavity. The conventional multimode laser diode is called the Fabry-Perot laser, while two common single-mode laser diodes are the DBR (Distributed Bragg Region) and the DFB (Distributed FeedBack) laser diodes [Carroll et al. 1993]. In the receiver, the commonly used components are PIN (P-insulator-N) diodes and APDs (Avalanche Photo Diode). The three classic types of optical fibers are multimode step-index fiber, multimode graded-index fiber, and single-mode step-index fiber, each with its own core/cladding design and dispersion characteristics. Further reading on fiber-optic communication systems and their components is found in [Keiser 1991], while an overview of components specially needed in WDMA networks is presented here.

Components for WDMA networks is a large research field, and examples of proposed components are given below. The component descriptions are given from a systems view, explaining what system function they are to perform and what level of performance can be expected. The aim is to provide the necessary amount of knowledge before proceeding to the presentation of architectures and protocols and leaving implementation details. In most
parallel and distributed computing systems, the longest distance between two end-nodes is rather short. Topics related to long distance communication are therefore not described here, e.g.:

*optical amplifiers*: may be needed when the optical signal is split to many nodes but are otherwise used when the signal is passed and attenuated through long distances of fiber

*coherent detection*: although it offers narrow channel spacing, it is too complicated for end-user systems

*different kinds of fibers*: performance trimming of fibers (lower attenuation and dispersion) is important for telecommunication but less relevant for short-distance communication

More information on WDMA components is found in [Borella et al. 1997] [Brackett 1990] [Green 1993] [Sadot and Boimovich 1998] [Tong 1998] [Zirngibl 1998].

### 2.3.1 Tunable Receivers

A receiver in a WDMA network must tune in one of all the wavelengths from the incoming fiber. Since a photo diode itself detects a broad band of wavelengths, a wavelength filter must be used. Tunable optical filters include interferometer filters, filters based on mode coupling, filters based on resonant amplification, and grating based filters. Fast tuning and a broad tuning range is desired while the pass-bandwidth should be adapted to the incoming channels, i.e., it should pass the whole energy of one channel while preventing energy from other channels from being detected.

The Mach-Zehnder interferometer filter splits the incoming light into two beams, delays one of the beams slightly more than the other and then lets the two beams interact with each other [Hecht 1987]. Tuning is obtained when the delay is changed. By cascading several Mach-Zehnder interferometer filters, a greater wavelength selectivity is achieved. In [Wooten et al. 1996], a three-stage electro-optically tunable Mach-Zehnder interferometer with a 50-ns tuning latency over a tuning range of eight channels is demonstrated. Experiments with a 16-channel (four stages) interferometer was reported in [Oda et al. 1989], while a 128-channel (seven stages) interferometer were used in the 100-channel transmission experiment reported in [Toba et al. 1990].

The Fabry-Perot filter is another interferometer in which two mirrors form a resonant cavity [Hecht 1987]. With a mirror reflectivity of less than 100 percent, some of the resonated light will escape out. The Fabry-Perot filter is tuned by changing the distance between the mirrors, i.e., moving one of
the mirrors. The distance must be short to limit the number of resonating wavelengths to one in the working range of the communication system. Because of the need of mechanically moving one of the lenses, the tuning latency is large. Piezoelectric transducers are commonly used to change the distance [Miller and Janniello 1990], while another method is to rotate (relative to the incoming and outgoing beams) a Fabry-Perot filter with a fixed distance [Frenkel and Lin 1988]. Experiments with cascaded Fabry-Perot filters to enhance performance have also been reported [Kaminow et al. 1989]. The two-stage filter described in [Miller and Miller 1992], supports 1000 channels in a 40 nm range. A tunable Fabry-Perot filter without moving parts is described in [Patel et al. 1990]. The cavity was filled with liquid crystals, and wavelength tuning was obtained by applying an electric field to change the refractive index of the liquid crystals.

Tunable filters based on mode coupling refers to filters with wavelength dependent coupling between optical fields (modes) where either acousto-optic, electro-optic, or magneto-optic effects are used [Kobrinski and Cheung 1989]. The optical signal is first split into two orthogonal polarization states and then passed through a resonant structure [Green 1993]. For a wavelength corresponding to the resonant frequency, the polarization state will be changed 90 degrees. A second polarization splitter at the output will filter this wavelength out from the rest of the wavelength channels. The different kinds of mode coupling filters generate the resonant structure in different ways. Electro-optic filters have low tuning latencies, while acousto-optic filters have a broad tuning range and support multi-wavelength filtering by applying multiple acoustic frequencies. Acousto-optic filters are reviewed in [Cheung 1990] [Gupta 1997].

If a single-mode laser diode is biased below threshold, it will start emitting when incoming light, at the wavelength selected by the grating in the laser, is present. If a tunable single-mode laser (see below) is used, a wavelength tunable filter with amplification is achieved. A tuning latency as low as 1 ns has been demonstrated [Kobrinski et al. 1988], but the tuning range is very limited. As an example, the tuning range demonstrated in [Numai et al. 1989] was 0.95 nm, although an 18-channel wavelength selection was expected with this filter.

In a grating based filter, the incoming light is spatially split into its wavelength components. If a photo diode array is used to detect all of the wavelength components (one wavelength per photo diode), as proposed in [Kirkby 1990], a multi-channel receiver is achieved in which the electronic switching time sets the tuning latency, i.e., very fast. A grating integrated monolithically with a photo diode array detecting 42 wavelength channels spaced by 4 nm was reported in [Cremer et al. 1992]. In [Parker and Mears 1996], one digitally tunable wavelength is steered to the output fiber by using an SLM (Spatial Light Modulator) together with a fixed grating.
Tuning to discrete wavelengths spaced by 1.3 nm over a tuning range of 38.5 nm was demonstrated. The demultiplexing of 120 channels by the use of a concave grating was reported in [Sun et al. 1998].

### 2.3.2 Tunable Transmitters

With wavelength-tunable transmitters, a certain transmission wavelength can be chosen, either from a continuous range or from a number of discrete wavelengths. From a system engineer's point of view, an ideal tunable laser diode has a tuning range of about 100 nm and a tuning latency in the order of nanoseconds [Mestdagh 1995]. Many components aiming to meet one of the two wishes have been proposed, and some meet both rather well. The linewidth (spectral width; normally measured as the FWHM, Full Width Half Maximum) of the emitted light gives a hint of the possible number of channels, but other parameters such as receiver selectivity must be considered when designing a WDM communication system. Wavelength-tunable laser diodes are reviewed in [Lee and Zah 1989].

The simplest way to achieve a tunable transmitter is by temperature tuning, utilizing the fact that the refractive index of the active layer is temperature dependent. Although a tuning range of 10.8 nm has been demonstrated [Kameda et al. 1993], only limited tuning range and long tuning latency are achieved with this method.

In [Kobrinski et al. 1990], a demonstration of a tuning latency of less than 15 ns over a tuning range of 2.2 nm and experiments with 16 wavelength channels are reported. A three-section DBR laser was used where two of the sections were used for wavelength tuning by changing the values of the injected currents and, hence, the index of refraction. As a consequence of changing the index of refraction of the grating in a DBR, the lasing wavelength (optical frequency) is changed. In [Goodman et al. 1988], eight channels were achieved from a double-section DFB with a tuning speed of less than 5 ns. A four-wavelength DBR laser array is reported in [Delorme et al. 1996]. Each laser is a three-section DBR tunable over a range of 12 nm. The total tuning range of the component is 28 nm.

By activating one laser in an array of laser diodes emitting at different wavelengths, a discretely tunable laser is achieved. A $7 \times 11$ [Chang-Hasnain et al. 1991B] and a $7 \times 20$ [Chang-Hasnain et al. 1991] VCSEL (Vertical Cavity Surface Emitting Laser) array with each laser emitting a unique, non-redundant wavelength has been reported (see [Giboney et al. 1998] for general information on VCSELS). The latter array had a wavelength span of 43 nm, and WDM experiments were carried out in which four of the lasers were each coupled to a fiber. The fibers were combined by a star coupler, and each laser was simultaneously modulated at 155 Mbit/s. A $2 \times 8$ VCSEL array with a small overlap in wavelength...
range between the two rows but in which each laser was capable of 5 Gbit/s operation was reported in [Maeda et al. 1991]. Other WDM VCSEL arrays include a $2 \times 2$ densely-packed array [Huffaker and Deppe 1996], and other WDM array components include a DFB laser array with 20 lasers emitting at ten 2-nm spaced wavelengths (two diodes per wavelength) [Lee et al. 1996].

One of the disadvantages of array components is the difficulty in coupling the light from all laser diodes into one fiber. However, the integration of a 21-wavelength DFB laser array, a star coupler, and two optical amplifiers on one chip was reported in [Zah et al. 1988].

Another way to get multiple parallel wavelength channels is by spectral slicing of the output light from a number of broad-spectrum LEDs [Zirngibl et al. 1996]. By filtering out a different wavelength from each diode, a wavelength tunable transmitter is obtained when selecting one of the diodes for modulation. Experiments with a 16-channel 150 Mbit/s system employing spectral slicing were reported in [Chapuran 1991].

Several external cavity laser diodes with external filters for wavelength tuning have been reported. By antireflection coating one of the output facets on a laser diode and having an external mirror, the cavity length is extended. If a diffraction grating is used as the external mirror, tuning is obtained by moving the grating (fine-tuning by changing the cavity length) or rotating the grating (a different longitudinal mode is selected from the grating) [Mellis et al. 1988]. Tuning ranges exceeding 240 nm have been demonstrated with external-cavity grating-based laser diodes [Bagley et al. 1990] [Tabuchi and Ishikawa 1990]. The drawback of these grating-based lasers is the large tuning latency caused by the mechanical movement of the grating. However, devices with shorter tuning latencies have been demonstrated. One way to avoid the mechanical moving mechanism is to exchange the diffraction grating with an acousto-optic filter that selects the lasing wavelength and allows for tuning latencies in the range of microseconds [Coquin and Cheung 1988]. Another fully electronically tunable laser diode is the MAGIC (Multistripe Array Grating Integrated Cavity) laser diode [Soole et al. 1992]. The MAGIC laser chooses one wavelength for lasing by activating one of several waveguide stripes. A fixed diffraction grating couples a specific wavelength into each stripe.

In [Larson and Harris 1996], a 15-nm tuning range and a 0.14 nm linewidth are demonstrated for a single VCSEL laser. A movable mirror was placed on top of the laser with an air gap between. By moving the mirror, the cavity length and, hence, the wavelength of the emitted light are changed. A similar tunable VCSEL but with a tuning range of 31.6 nm was recently reported [Sun et al. 1998]. An LED with the same tuning mechanism, a
tuning range of 39 nm, and a linewidth of 1.9 nm is presented in [Larson and Harris 1995].

A diode-laser pumped tunable fiber laser is reported in [Chieng and Minasian 1994]. Light in the 1550 nm range goes through an Erbium-doped fiber ring, where it is amplified. A tunable Mach-Zehnder interferometer selects one wavelength which is re-amplified. A tuning range of 39 nm is achieved. A fiber laser with an electro-optic filter is reported in [Chollet et al. 1996]. The laser supports 50 wavelength channels over a spectral range of 10 nm with a 50-ns tuning latency.

2.3.3 Other WDM Components

Essential system components when building fiber-optic networks like those described in the next section include splitters, combiners, stars, wavelength converters [Nesset et al. 1998], WDM demultiplexers, and WDM multiplexers. Several of the components mentioned are commercially available with limited features, e.g., a small number of input/output ports.

A $1 \times N$ fiber-optic splitter splits the light from one input fiber to $N$ output fibers. Both symmetric and asymmetric splitters are available, where an asymmetric splitter splits an unequal amount of light to the different output ports. A combiner works in the opposite way combining input signals from several fibers to one output fiber. An $N \times 1$ fiber-optic star (often referred to as a passive optical star) can be viewed as an $N \times 1$ combiner followed by an $1 \times 1$ split. The conventional way to build a star is to use a number of $2 \times 1$ combiners and $1 \times 2$ splitters. However, it is difficult to build large stars using this technique, and other techniques have been proposed to overcome the problem [Okamoto et al. 1992] [Yun and Kavehrad 1992]. A 144 $\times$ 144 passive optical star is described in [Okamoto et al. 1992B] and [Kato et al. 1993].

A WDM demultiplexer is a splitter that splits the incoming signal on one fiber to several output fibers, each with a different spectral component of the input signal [Juma 1996]. The demultiplexer either splits the input signal into groups of wavelength channels or into one separate wavelength channel per output fiber. The most common methods for demultiplexing the input signal are said to be the use of an interference filter or a diffraction grating [Straus and Kawasaki 1987]. A WDM multiplexer couples together different wavelengths on different fibers into one outgoing fiber. Although there exist WDM multiplexers with wavelength selective inputs, the simplest way to get a WDM multiplexer is to use an ordinary $N \times 1$ combiner.
The ADM (Add-Drop Multiplexer) is another component in which one or several wavelengths may be added/dropped to/from a bypassing fiber. The component can be used in, e.g., WDM ring networks (described in the next section). An ADM based on two fixed wavelength, multi-layer Fabry-Perot filters is described in [Hamel et al. 1995]. A component in which the wavelength to be added/dropped can be tuned over a discrete number of wavelengths is described in [Glance 1996].

An arrayed waveguide grating (AWG) [Okamoto and Inoue 1995] can be used, e.g., as a demultiplexer in a wavelength router [McGreer 1998] (see Section 2.5.1) or as an ADM. In the case of a single-input AWG, the light from the input waveguide is split into a number of waveguides with different lengths. Where all these waveguides end, the light beams constructively interfere at the entrance of one of several output waveguides. The output waveguide into which the light is transferred depends on the wavelength of the light and the input waveguide, if several exist, into which the light entered.

### 2.4 Interconnection Architectures

The simplest fiber-optic transmission system is to have a point-to-point link between two nodes. A network can be built by using several point-to-point links. These so-called point-to-point linked networks have been commercially available for a while and are described in Section 2.4.1. However, all-optical multi-access networks are expected to be popular in the future (described in Section 2.4.2).

#### 2.4.1 Point-to-Point Linked Networks

In a fiber-optic point-to-point linked network, a number of optically-isolated links connect the nodes in the network. If, for example, wavelength division multiplexing should be used to increase the bandwidth, the technology must thus be implemented separately in each point-to-point link. The main types of point-to-point linked networks that fall into the scope of this report are ring networks [Davies and Ghani 1983] and switched networks.

In the FDDI network [Ross 1989] [Jain 1993], the nodes are connected in a unidirectional primary ring where each node receives from one link and transmits on another. A node acts like a repeater for all incoming messages except those addressed to the node and those sent by the node and should be removed from the ring after one round. In addition to the primary ring, some or all nodes are connected to a secondary ring which, for example, can be used as a backup ring. A transmission rate of 125 Mbaud is used, which gives a data rate of 100 Mbit/s because of the 4b/5b encoding that is used. A token protocol is used for medium access control. Guarantees can be given
for both latency and bandwidth in the FDDI network and its successor, FDDI-II [Ross 1989] [Jain 1993].

In the Wavelength Distributed Data Interface (WDDI) network described in [Ramamurthy et al. 1995], multiple FDDI networks are obtained on logical rings in a WDM ring network. Only a subset of the total number of nodes in the network is connected to each logical ring. Some of the nodes act as bridges to interconnect two or more logical rings.

In switched networks, one or several switches connect the nodes together. All traffic in the network passes one switch at least. One well known switched network is ATM (Asynchronous Transfer Mode). ATM was developed to carry a wide range of traffic such as video, voice, and data, both at the local and the global levels. To ease the transition from current LAN technologies, ATM LAN emulation has been developed to support emulation of a broadcast medium [Truong et al. 1995] [Finn and Mason 1996]. Examples of different services provided by an ATM network are [Reilly 1994] [Suzuki 1994]:

- Constant Bit Rate (CBR) with minimal and constant delay between the end-nodes. Can be used for, e.g., voice over ISDN.
- Variable Bit Rate (VBR) with minimal and constant delay between the end-nodes. Designed to handle video and audio where, e.g., the compression ratio can vary.
- Unspecified Bit Rate (UBR), e.g., for data transmissions without real-time requirements.

As indicated above, real-time services are supported, but guarantees are not made for separate cells (data units in ATM each consisting of 48-byte data and a 5-byte address) without a setup phase, due to the fact that ATM is a connection-oriented network.

Electronic ATM switches and integrated switching circuits are commercially available [Goldberg 1994], while optical switches have been proposed [Nishio et al. 1993] [Jajszczyk and Mouftah 1993] [Blumenthal et al. 1994]. The way in which ATM networks can be used in parallel and distributed computing systems is discussed in [Hariri and Lu 1996]. Another network for parallel processing systems, using fiber-optic point-to-point interconnections, is Nectar, where one or several electronic crossbars switch the traffic [Arnould et al. 1989].

A multi-stage switching system with WDMA interconnections between the stages and with wavelength conversion inside the switching stages was reported in [Fujiwara et al. 1988]. Wavelength conversion was obtained by
converting the optical signal first into electronic form and then back to optical again, but at another wavelength.

A system component that has reached the market recently [Bursky 1994] [Fibre Systems 1998] is the fiber-ribbon link [Buckman et al. 1998] [Engelbertsen et al. 1996] [Hahn 1995] [Hahn 1995B] [Hahn et al. 1996] [Hartman et al. 1990] [Jiang et al. 1998] [Karstensen et al. 1995] [Karstensen et al. 1998] [Kuchta et al. 1998] [Nagarajan et al. 1998] [Nishimura et al. 1997] [Nishimura et al. 1998] [Schwartz et al. 1996] [Siala et al. 1994] [Wickman et al. 1999] [Wong et al. 1995]. Several links can be used to build high bandwidth point-to-point linked networks [Hahn et al. 1995]. With ten parallel fibers, each carrying data at a bit rate of 400 Mbit/s, an aggregated bandwidth of 4 Gbit/s is achieved [Schwartz et al. 1996]. Bi-directional links, with some fibers in the fiber-ribbon cables dedicated for each direction, are also possible [Jiang et al. 1995]. Further discussions on fiber-ribbon aspects are presented in Section 9.5, page 153, while more references to reports on fiber-ribbon links are found in [Tooley 1996].

Modules that support multiple high-speed channels but are not specifically optimized for fiber-ribbons have been reported, e.g., receiver and transmitter modules with five channels, each channel with a bit rate of 2.8 Gbit/s [Nishikido et al. 1995].

2.4.2 All-Optical Multi-Access Networks

In an all-optical network, the data stream remains in the optical form all the way from the transmitter to the receiver. Three basic architectures for all-optical multi-access networks are the ring, the bus, and the star (see Figure 1). These network architectures will be discussed below.

An all-optical multi-access ring network differs from a traditional ring network in the sense that all other nodes can be reached in a single hop without any intermediate optoelectronic conversion. In contrast to the repeating function of a node in, for example, a FDDI network, messages simply pass a node through passive optics in an all-optical multi-access ring network. This is true for all messages except a node’s own messages that should be removed from the ring after one round. Just a fraction of the optical power contained in the bypassing fiber is tapped to the receiver, which gives all nodes the opportunity to read the message, i.e., a multicast (one to many) or a broadcast (one to all). Outgoing messages are inserted into the ring and, in a multi-channel system, mixed together with bypassing messages on other channels.

In the WDMA ring network described in [Irshid and Kavehrad 1992], each node is assigned a node-unique wavelength on which to transmit.
Transmitted messages are then accessed by tunable receivers. Experiments with a WDMA unidirectional ring network are reported in [Chawki et al. 1995], while a WDMA bi-directional ring network is described in [Elrefaie 1993]. WDMA ring networks with central electronic switching are described in [Wagner and Chapuran 1992]. WDMA ring networks for long distance communication have already reached the market [Grenfeldt 1998].

*Figure 1: Three passive optical network architectures: (a) ring, (b) (dual) bus, and (c) star.*
A hierarchical ring network is described in [Louri and Gupta 1996] [Louri and Gupta 1997]. One ring is dedicated to each cluster for intra-cluster communication, while a hierarchical all-optical (no optical-to-electrical conversion or vice versa in the gateway nodes) ring network is used for inter-cluster communication. A TDMA protocol is investigated for use in the network.

In an optical bus, the light travels only in one direction, making it necessary to have two buses (upper and lower), one for each direction (higher or lower node index of destination nodes). This kind of bus architecture is called dual bus. The disadvantage of the dual bus is that two transceivers are needed in each node. This is avoided in the folded bus, where the two buses are connected with a wrap-around connection at one end of the buses (see Figure 2) [Tseng and Chen 1982]. In the folded bus, transmitters are connected to the upper bus while receivers are connected to the lower bus.

Several bus architectures and hybrids in which the bus is part of the architectures are discussed in [Nassehi et al. 1985]. A WDMA dual bus network is described in [Cheung 1992], and a WDMA folded bus network is described in [Chlamtac and Ganz 1988]. WDMA dual bus networks in which messages do not always remain unchanged until the end of the bus, but in which the protocols support several transmissions on the same wavelength channel simultaneously (wavelength reuse), are presented in [Huang and Sheu 1996] [Huang and Sheu 1997].

In a star network, the incoming light waves from all nodes are combined and uniformly distributed back to the nodes. In other words, the optical power contained in the middle of the star is equally divided between all nodes. WDMA star networks are described in numerous papers and are specially treated in Section 2.6. In addition to LANs and similar networks, WDMA star networks have also been proposed for internal use in packet switches [Eng 1988] [Brackett 1991].

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], the star-of-stars network that has an electronic gateway node between each cluster and the backbone star [Jonsson and Svensson 1997] and the multiple star network where each node is directly connected to both a local star and a remote star [Ganz and Gao 1992B].

All of the three basic network architectures have different advantages. The ring has the least amount of fibers, a bus network’s medium access protocol can utilize the linear ordering of the nodes [Nassehi et al. 1985] and the attenuation for an (ideal) star only grows logarithmically with the number of nodes. However, star networks are the most popular, judging from the number of published papers.

2.5 Classifications of WDMA Networks

The classifications below give the terminology and the background knowledge assumed in the overview of proposed networks and protocols (medium access control protocols) in the next section. If not otherwise stated, multi-access, single-hop broadcast-and-select networks (explained below) are assumed in the following sections. More information on WDMA networks is found in [Borella et al. 1998] [Brackett 1990] [Green 1993] [Mestdagh 1995]. Predictions for future directions in WDMA networking are found in [Brackett 1996] [Green 1996].

2.5.1 Broadcast-and-Select and Wavelength Routing

The difference between broadcast-and-select networks and wavelength routing networks is whether or not there exists any wavelength dependent routing in the network [Gerstel 1996]. The path followed from a transmitter to a receiver in a wavelength routing network is determined by the selected wavelength. As long as two paths do not have any common fiber links, they can use the same wavelength simultaneously (wavelength reuse). This is not possible in a broadcast-and-select network, where all receivers always have all transmitted channels available on the incoming fiber. However, a broadcast-and-select network does not need any wavelength selective devices out in the network. Instead, the receivers decide (on the basis of the protocol used) when to receive and/or which channel to tune in.

Several wavelength routing networks have been proposed, for example, the three-level hierarchical network described in [Alexander et al. 1993]. Another hierarchical wavelength routing network is LIGHTNING [Dowd et al. 1995] [Dowd et al. 1996]. The topology in LIGHTNING is a tree-structure of passive optical stars where each level is separated by a lambda partitioner. The lambda partitioner selects the wavelengths that are to stay
in the levels beneath the partitioner and the wavelengths that will pass to
the levels above. In this way, the number of wavelengths in each fiber has
its minimum at the top level (only top level wavelengths exist) and
maximum at the bottom level (wavelengths for all levels exist). All
wavelengths that are not passed to a certain level can be reused in each
cluster on the level below.

The focus in this section is on broadcast-and-select networks and examples
of such networks are found in Section 2.6.

2.5.2 Single-Hop and Multihop Networks
In a single-hop network, all nodes can reach any other node in a single hop.
This means that the transmitted data are not passed through any
intermediate routing stages and remain in optical form all the way from the
source node to the destination node. A disadvantage of single-hop networks
is that each transmitter and/or receiver must be equipped with tunable
components. Advantages are flexibility and the absence of extra latency at
intermediate nodes.

In a multihop network, fixed-wavelength (or slowly tunable) transmitters
and receivers are configured to shape a virtual topology. As an example, a
node may be able to transmit on $\lambda_1$, $\lambda_2$, and $\lambda_3$ in a broadcast-and-select
network, where the three neighbors in a virtual binary three-dimensional
cube have receivers tuned to $\lambda_1$, $\lambda_2$, and $\lambda_3$, respectively. A number of virtual
topologies have been investigated, e.g., perfect shuffle in Shufflenet
[Acampora and Karol 1989]. An experimental multihop network was
reported in [Gidron 1992]. An advantage of multihop networks is that no
components with low tuning latency are needed. By having slowly tunable
components, the virtual topology can be reconfigured when the overall
traffic demands in the network are changed.

Although multihop networks can be reconfigurable, single-hop networks are
more flexible. This chapter focuses on single-hop networks. Single-hop
networks are reviewed in [Mukherjee 1992], and multihop networks are
reviewed in [Mukherjee 1992B].

2.5.3 Networks with and without Control Channel
A number of protocols for WDMA networks assume a separate control
channel. This control channel is normally used to reserve access on the data
channels. Hence, these protocols are sometimes called reservation protocols.
Protocols for networks with more than one control channel have also been
proposed, for example, $N$-DT-WDMA [Humblet et al. 1992] [Humblet et al.
1993]. Non-control channel based networks often assign a fixed home
channel to each node, either on the transmitter side or on the receiver side.
Protocols for these networks are then called *pre-allocation* protocols. Protocols for networks without control channel but where access to the data channels is divided into a control phase and a data phase are found in [Sivalingam and Dowd 1995] [Jonsson et al. 1996]. These protocols are called *hybrid* protocols.

Instead of having an additional wavelength channel for the control channel, SCMA can be used to achieve a number of control channels [Su and Olshansky 1993]. In the proposed network, the subcarrier carrying a node’s control data modulates the same laser diode as that on which the data are transmitted. Two photo diodes are used in the receiver, one with a filter that filters out the wavelength with the data and one without a filter to receive all subcarriers in the network. One of the subcarriers is then selected using microwave techniques.

### 2.5.4 Receiver and Transmitter Tunability

The most flexible networks are those with both tunable transmitters and tunable receivers. However, protocol complexity and node cost decrease when fixed wavelength units are used in either the transmitter or the receiver. Using the classification scheme given in [Mukherjee 1992], WDM networks can be divided into:

- **FT-FR** (Fixed Transmitters and Fixed Receivers)
- **TT-FR** (Tunable Transmitters and Fixed Receivers)
- **FT-TR** (Fixed Transmitters and Tunable Receivers)
- **TT-TR** (Tunable Transmitters and Tunable Receivers)

Some networks have more than one transmitter and/or receiver in each node. A network with \(i\) fixed transmitters, \(j\) tunable transmitters, \(m\) fixed receivers, and \(n\) tunable receivers can be described as:

\[
\text{FT}^i\text{TT}^j\text{FR}^m\text{TR}^n
\]

if no control channel is used or as:

\[
\text{CC-FT}^i\text{TT}^j\text{FR}^m\text{TR}^n
\]

if the network is control channel based. The number of nodes in a network is denoted as \(M\). As an example, a TT-\(\text{FR}^m\) network has one tunable transmitter and \(M\) fixed receivers in each node.

In a packet switched network, the tuning latency is critical for the network performance. This issue was addressed in the discussion on tunable
components in Section 2.3. Also, fast locking clock recovery circuits must be used instead of the slow PLL based circuits used in point-to-point links. Several methods to recover the clock signal on one or a few bits have been proposed [Banu and Dunlop 1992] [Jonsson and Moen 1994] [Cerisola et al. 1995].

2.5.5 Random-Access Protocols
In a network using a random-access protocol, all nodes compete for a transmission channel in a random (uncontrolled) way [Halsall 1995]. Random-access protocols can be contrasted with protocols with dynamic or static pre-assigned transmission patterns. The original single-channel random-access protocol is the Aloha protocol, where transmission is done with no regard to other nodes. If two messages from different nodes overlap in time, both are corrupted. Several protocols are more or less pure improvements of the Aloha protocol, e.g., Slotted Aloha, CSMA (Carrier-Sense Multiple-Access), and CSMA/CD (CSMA with Collision Detection). In slotted Aloha, all nodes are synchronized and transmissions are allowed to be started only at the beginning of a time slot. In this way, the probability of collision is reduced. In CSMA, the transmission medium is sensed before transmission starts. If a carrier is sensed, the transmission is postponed until the current message has reached its end. If two nodes sense the medium at the same time and find it free, they both begin to send and a collision will take place. In this case, the medium is busy for the whole duration of the corrupted transmissions. This is avoided in the CSMA/CD protocol, where a collision can be detected during transmission. If a collision is detected by two nodes, both nodes stop and wait for a random time before trying again, beginning with the carrier-sense mechanism. Common to the random-access protocols is that they do not perform well at high traffic-loads owing to the increased probability of collision but experience low latency at light loads.

In multiple-channel networks, a random-access protocol can be used both on the data channels and on the control channel if present. Random access protocols for control channel based networks are known as X/Y protocols, where X denotes the protocol used on the control channel while Y denotes the protocol on the data channels. An example of an X/Y protocol is the Slotted Aloha/Aloha protocol described in [Habbab et al. 1987].

2.5.6 Time-Wavelength Assignment Schemes
Several algorithms have been proposed to solve different kinds of time wavelength assignment problems [Rouskas and Ammar 1995]. One specific problem is to schedule a traffic matrix (containing the communication demands between every possible pair of nodes) on a limited number of wavelength channels, minimizing the number of time slots needed [Pieris
Other aspects may also be considered, for example, minimizing the number of tuning periods, that is, the periods in which the tuning of the transmitters and the receivers are not changed [Ganz and Gao 1992] [Ganz and Gao 1992B].

### 2.6 Single-Hop WDMA Star Networks and Protocols

A number of single-hop broadcast-and-select WDMA star networks and protocols for these kinds of networks have been proposed. They are suggested for a wide range of applications but in general can all be used in LANs and distributed computing systems. Several of these networks are summarized in Table 1, Table 2, Table 3, and Table 4. Table 1 lists network experiments, while Table 2 and Table 3 list MAC protocols for non-control channel based and control channel based networks, respectively. Although the protocols listed are for packet-switched traffic, protocols for circuit-switched traffic have been proposed [Dono et al. 1990]. Some papers

<table>
<thead>
<tr>
<th>Name/Reference</th>
<th>Tunability classification</th>
<th>Number of wavelengths</th>
<th>Channel bandwidth</th>
<th>Organization/other info.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOX [Arthurs et al. 1988]</td>
<td>TT-FR</td>
<td>2</td>
<td>1 Gbit/s</td>
<td>Bellcore; two stars, one for each direction</td>
</tr>
<tr>
<td>HYPASS [Kobrinski et al. 1988B]</td>
<td>TT-FR &amp; FT-TR</td>
<td>8</td>
<td>1.2 Gbit/s</td>
<td>Bellcore; two stars, one for each direction</td>
</tr>
<tr>
<td>Lambdanet [Kobrinski et al. 1987]</td>
<td>FT-FR Multiple</td>
<td>18</td>
<td>1.5 Gbit/s</td>
<td>Bellcore</td>
</tr>
<tr>
<td>Rainbow-I [Janniello et al. 1992]</td>
<td>FT-TR</td>
<td>6</td>
<td>300 Mbit/s</td>
<td>IBM T. J. Watson; circuit switched traffic</td>
</tr>
</tbody>
</table>

**Table 1: Experimental single-hop WDMA star networks. M is the number of nodes.**

and Sasaki 1994]. Other aspects may also be considered, for example, minimizing the number of tuning periods, that is, the periods in which the tuning of the transmitters and the receivers are not changed [Ganz and Gao 1992] [Ganz and Gao 1992B].

<table>
<thead>
<tr>
<th>Name/Reference</th>
<th>Tunability classification</th>
<th>Number of channels</th>
<th>Protocol type</th>
<th>Packet length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slotted Aloha [Dowd 1991]</td>
<td>TT-FR</td>
<td>N</td>
<td>Slotted Aloha with ACK-control slots</td>
<td>Fixed</td>
</tr>
<tr>
<td>I-TDMA [Sivalingam et al. 1992]</td>
<td>TT-FR</td>
<td>N</td>
<td>Static TDMA</td>
<td>Fixed</td>
</tr>
<tr>
<td>I-TDMA* [Bogineni et al. 1993]</td>
<td>TT-FR</td>
<td>N</td>
<td>Static TDMA</td>
<td>Fixed</td>
</tr>
<tr>
<td>FatMAC [Sivalingam and Dowd 1995]</td>
<td>TT-FR</td>
<td>N</td>
<td>TDMA with control phase for reserving</td>
<td>Variable (fixed units)</td>
</tr>
<tr>
<td>TD-TWDMA [Jonsson et al. 1996]</td>
<td>FT-TR</td>
<td>M</td>
<td>TDMA with control phase for reserving</td>
<td>Fixed</td>
</tr>
</tbody>
</table>

**Table 2: Non-control channel based protocols for single-hop broadcast-and-select WDMA networks. M is the number of nodes, and N is a positive integer greater than zero.**
describing single-hop WDM star networks that do not fit into the scope of Table 1, Table 2, or Table 3, are listed in Table 4. Several of these papers are application suggestions and early contributions of conceptual thoughts.

Of the network experiments in Table 1, Lambdanet [Kobrinski et al. 1987] [Goodman et al. 1990] was the first to reach “industrial strength” [Green 1996]. Other noteworthy experimental demonstrations include the two-star FOX network for shared memory multiprocessors [Arthurs et al. 1988] and the circuit switched Rainbow network [Dono et al. 1990]. The experimental demonstrations of Rainbow-I were reported in [Janniello et al. 1992].

<table>
<thead>
<tr>
<th>Name/Reference</th>
<th>Tunability classification</th>
<th>Num. of chan.</th>
<th>Control ch. protocol</th>
<th>Data ch. protocol</th>
<th>Packet length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aloha/Aloha [Habbab et al. 1987]</td>
<td>CC-TT-TR</td>
<td>$N + 1$</td>
<td>Aloha</td>
<td>Aloha</td>
<td>Fixed</td>
</tr>
<tr>
<td>Slotted Aloha/Aloha [Habbab et al. 1987]</td>
<td>CC-TT-TR</td>
<td>$N + 1$</td>
<td>Slotted Aloha</td>
<td>Aloha</td>
<td>Fixed</td>
</tr>
<tr>
<td>Aloha/CSMA [Habbab et al. 1987]</td>
<td>CC-TT-TR</td>
<td>$N + 1$</td>
<td>Aloha</td>
<td>CSMA</td>
<td>Fixed</td>
</tr>
<tr>
<td>CSMA/Aloha [Habbab et al. 1987]</td>
<td>CC-TT-TR</td>
<td>$N + 1$</td>
<td>CSMA</td>
<td>Aloha</td>
<td>Fixed</td>
</tr>
<tr>
<td>CSMA/N-Server Switch [Habbab et al. 1987]</td>
<td>CC-TT-TR</td>
<td>$N + 1$</td>
<td>CSMA</td>
<td>N-Server Switch</td>
<td>Fixed</td>
</tr>
<tr>
<td>DT-WDMA [Chen et al. 1990]</td>
<td>CC-FT$^2$-FRTR</td>
<td>$M + 1$</td>
<td>Uniform static TDMA</td>
<td>Distributed algorithm</td>
<td>Fixed</td>
</tr>
<tr>
<td>Improved Slotted Aloha/Aloha [Mehravari 1991]</td>
<td>CC-TT-TR</td>
<td>$N + 1$</td>
<td>Slotted Aloha</td>
<td>Aloha</td>
<td>Fixed</td>
</tr>
<tr>
<td>Slotted Aloha/N-Server Switch [Mehravari 1991]</td>
<td>CC-TT-TR</td>
<td>$N + 1$</td>
<td>Slotted Aloha</td>
<td>N-Server Switch</td>
<td>Fixed</td>
</tr>
<tr>
<td>Slotted Aloha [Sudhakar et al. 1991]</td>
<td>CC-TT-TR</td>
<td>$N + 1$</td>
<td>Slotted Aloha</td>
<td>Aloha</td>
<td>Fixed</td>
</tr>
<tr>
<td>Reservation Aloha [Sudhakar et al. 1991]</td>
<td>CC-TT-TR</td>
<td>$N + 1$</td>
<td>Slotted Aloha</td>
<td>Aloha</td>
<td>Fixed</td>
</tr>
<tr>
<td>DAS [Chipalkatti et al. 1992]</td>
<td>CC-FT$^2$-FRTR</td>
<td>$M + 1$</td>
<td>Uniform static TDMA</td>
<td>Distributed algorithm</td>
<td>Fixed</td>
</tr>
<tr>
<td>Hybrid TDM [Chipalkatti et al. 1992]</td>
<td>CC-FTTT$^N$-FR$^{N+1}$</td>
<td>$M + 1$</td>
<td>Uniform static TDMA</td>
<td>Two different protocols</td>
<td>Fixed</td>
</tr>
<tr>
<td>$N$-DT-WDMA [Humblet et al. 1993]</td>
<td>CC-FTTT-FRTR</td>
<td>$2M$</td>
<td>slotted</td>
<td>slotted</td>
<td>Fixed</td>
</tr>
<tr>
<td>MultiS-Net [Jia and Mukherjee 1996]</td>
<td>CC-TT-TR</td>
<td>$N + 1$</td>
<td>Slotted random access</td>
<td>Distributed algorithm</td>
<td>Fixed</td>
</tr>
</tbody>
</table>

Table 3: Control channel based protocols for single-hop broadcast-and-select WDMA networks. $M$ is the number of nodes, and $N$ is a positive integer greater than zero.
Of the protocols for non-control channel based networks (Table 2), several are TDMA variants. For example, I-TDMA [Sivalingam et al. 1992] and I-TDMA* [Bogineni et al. 1993] are multiple-channel variants of the traditional static TDMA. Random access protocols have also been proposed for non-control channel based networks. One example described in [Dowd 1991] is a variant of the Slotted Aloha protocol.

Numerous protocols for control channel based networks have been proposed (Table 3). One of the first papers, [Habbab et al. 1987], describing protocols for WDMA networks presents five random access protocols utilizing a control channel. Several variants and improvements of these protocols have later been proposed, for example, those described in [Mehravari 1991]. Several of the protocols for control channel based networks use uniform static TDMA on the control channel and some form of distributed algorithm to schedule access to the data channels. Some of the protocols support variable sized packets, for example, the TDMA-C protocol [Bogineni and Dowd 1992]. Protocols with other noteworthy features include: protocols utilizing multiple control channels, for example, N-DT-WDMA [Humblet et al. 1993], where \( N \) is the number of nodes and control channels; tell-and-go protocols (data are sent immediately after the control information, without waiting for the control information from the own and/or other nodes to return), for example, the Aloha/Aloha protocol [Habbab et al. 1987]; and protocols not requiring dedicated hardware for the control channel, for example, MultiS-Net [Jia and Mukherjee 1996]. Protocols for single-hop networks are reviewed in [Mukherjee 1992].

Some of the networks and protocols listed in Table 1 to Table 4 are described in Sections 2.6.1 - 2.6.6 for purposes of example. Sections 2.6.1 - 2.6.2 are network oriented, while Sections 2.6.3 - 2.6.6 are protocol oriented. The protocol descriptions are given at a level trying to explain the basic function and important features. If nothing else is stated, a passive optical star is assumed as the physical topology when describing the protocols. However, the passive optical star can often be exchanged with another multi-access network without any large modifications of the protocol. The

<table>
<thead>
<tr>
<th>Name/Reference</th>
<th>Tunability classification</th>
<th>Other features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photonic knockout switch</td>
<td>FT-FR(^N)</td>
<td>Application: centralized packet-switches.</td>
</tr>
<tr>
<td>[Eng 1988]</td>
<td></td>
<td>A separate electrical control network is used.</td>
</tr>
<tr>
<td>Broadcast SYMFONET</td>
<td>FT-FR(^M)</td>
<td>Application: shared memory multiprocessors.</td>
</tr>
<tr>
<td>[Westmore 1991]</td>
<td></td>
<td>Synchronization issues and timing are discussed.</td>
</tr>
<tr>
<td>MCA</td>
<td>TT(^N)-TR(^N)</td>
<td>Application: massively parallel computers.</td>
</tr>
<tr>
<td>[Brackett 1991]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 4: Other proposed single-hop WDM star networks. \( M \) is the number of nodes, and \( N \) is a positive integer greater than zero.*
focus on single-hop broadcast-and-select networks is motivated by the simple architecture needed to support flexible traffic between end-nodes. More WDM networks and protocols have been reported, e.g., those described in [Jia et al. 1995] [Levine and Akyildiz 1995].

2.6.1 FOX
The FOX network reported in [Arthurs et al. 1988] is an interconnection architecture for parallel computers with shared memory. Two stars are used, one for communication from the processing elements and to the memory modules and one for the opposite direction. Both the star networks are TT-FR networks with a unique wavelength for each receiver. If two processing elements transmit to the same memory module at the same time, a collision will occur. With the motivation of a low cache miss rate, collision detection and retransmission are proposed to be a feasible solution to this problem. Experiments with two wavelengths and less than 20 ns tuning latency were reported.

A related network is the HYPASS switching system [Kobrinski et al. 1988B] [Goodman et al. 1988]. HYPASS also employs one star for each direction, but internally in a switch. One star is used for the data transport, while the other passes back control information from the output stages.

2.6.2 Lambdanet
The Lambdanet WDMA star network presented in [Kobrinski et al. 1987] is classified as FT-FR^M, where M is the number of nodes. Although Lambdanet was designed to have one specific wavelength per node as a transmitter home channel, two extra wavelengths were used in the 16-node experiment, i.e., a total number of 18 wavelengths. Commercial DFB laser diodes were selected to obtain a channel spacing of 2 nm, while a grating wavelength demultiplexer was used in the receiver to select one of the incoming wavelengths. Further experiments was reported in [Goodman et al. 1990], where the channel bit-rate was increased from 1.5 Gbit/s to 2 Gbit/s.

2.6.3 DT-WDMA
The DT-WDMA (Dynamic Time-Wavelength Division Multi Access) protocol for CC-FT^2-FRTR networks, described in [Chen et al. 1990], divides the access to both control channel and data channels into slots of equal size. However, slots on the control channel are further divided into mini-slots, one to each transmitter. When node i wants to transmit to node j, it waits for its next mini-slot on the control channel. An address field in the control packet is set with the address of node j. Another field is set with a delay value related to the generation time of the message. The data message is then sent in the data slot succeeding the control slot. After each slot when
all mini-slots are received, a deterministic distributed algorithm (separately computed in each node with the same outcome) is run in each node. On the basis of the distributed algorithm, node \( j \) will choose to tune in the node with the largest delay value of all the nodes that have the address of node \( j \) in the address field. All other messages destined to node \( j \) will get lost. In a pipeline fashion, several messages can be sent before success or not of the first message is known.

The DT-WDMA protocol tries to minimize the maximum packet delay. It also tries to cope with large propagation delays by using the pipeline mechanism. In addition to packet-switched traffic, circuit-switched traffic is also supported by DT-WDMA.

### 2.6.4 A Distributed Adaptive Protocol

In [Yan et al. 1996] and [Yan et al. 1996B], a protocol, not named in the papers, supporting soft real-time traffic is described. The QoS (Quality of Service) associated with a real-time packet in the network is the probability of missing the deadline. The distributed algorithm used tries to globally minimize the number of packets not managing the QoS by adaptively changing the priority of the queued packets. The network architecture is \( \text{CC-FTTT}^N \cdot \text{FR}^N + 1 \), where the fixed transmitter and one of the fixed receivers are dedicated to the control channel. The \( N \) tunable transmitters and \( N \) of the fixed receivers are dedicated to the \( N \) data channels.

Access to the control channel is based on a circulating token. The token is always broadcasted to every other node so that they can take part of the status information about the sending node that is contained in the token. In each node, the latest version of this information from every other node is stored in a table. In this way, the current system state is known by all nodes, and an identical copy of the distributed algorithm can be run on each node. When a node has the token, it determines which data channels to acquire or release, in a way trying to globally meet the QoS demands. The determination is based on the current system state and on information about the own queued messages, e.g., deadline and acceptable deadline missing fraction. The node to which the token is passed is also determined according to the QoS demands of all messages in the network.

Even though the protocol has sophisticated methods for real-time messages, it is targeted only for soft real-time traffic. Guarantees can not be given that a message will meet its deadline.

### 2.6.5 Interleaved TDMA

The I-TDMA (Interleaved TDMA) protocol described in [Sivalingam et al. 1992] is an extension of the traditional static uniform TDMA protocol. The
protocol assumes a non-control channel based network with tunable transmitters and fixed receivers. The access to each channel is divided into $M$ slots, where $M$ is the number of nodes. The assignment is fully static and gives each node one slot per cycle and channel in which to transmit. Each node has access to a maximum of one channel at a time. If there are $M$ channels in the network, each node always has access to exactly one channel. An extension of I-TDMA, called I-TDMA*, is described in [Bogineni et al. 1993]. The only difference is that I-TDMA* has $C$ queues, where $C$ is the number of channels, for outgoing messages in each node instead of one queue. Head-of-line problems are hence avoided. The head-of-line problem means that a node has a packet to send but can not reach it in the queue because there are other packets (whose destination it is not possible to transmit to for the moment) in front of it. As with traditional TDMA, these protocols reach high bandwidth utilization but have relatively large latencies at low traffic intensities.

### 2.6.6 FatMAC

The FatMAC protocol is proposed for use in distributed shared memory multiprocessors [Sivalingam 1994] [Sivalingam and Dowd 1995]. No control channel is used, and the network is classified as TT-FT. By choosing a laser diode array as the tunable transmitter, broadcast is made possible through simultaneous activation of all laser diodes in the array. The access to the channels is divided into cycles of variable length. The cycles have two parts, a control phase followed by a data phase. Each node has a slot with broadcast capability in the control phase, where it transmits its transmission demands for the cycle. The packet is scheduled in the data phase among other demanded transmissions, following the same order as the control slots. The length of the data phase depends on the number of demanded transmissions. Positive features of the network include: support for variable length packets, no need for control channel, and collisionless transmission.

A related protocol is TD-TWDMA (see Chapter 7), which also has a control phase and a data phase instead of a separate control channel [Jonsson et al. 1996] [Jonsson et al. 1997]. The TD-TWDMA protocol is developed for distributed real-time systems, however, and has features for those systems.

### 2.7 Conclusions

This chapter has presented an introduction to high-performance fiber-optic networks. The emphasis has been on multiple-channel passive optical networks, especially WDMA networks, because these networks have the highest potential for meeting future bandwidth demands at a reasonable cost. Networks for numerous applications have been proposed in the
literature, but most of the networks and protocols referred to in this chapter can be used in distributed computing systems too. WDMA star networks are foreseen to have a key role in future high-performance computer communication networks, and examples of such networks and protocols for them have been described. Many of the components demonstrated for WDMA networks, as reviewed in this paper, further indicate that high-performance WDMA networks for end-user systems will be available in the near future.
3. 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 neighbors via point-to-point links. This forms a static topology of the network, e.g., a two-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. Thus 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 Section 3.1, static, dynamic, shared-medium, and hybrid networks are described in Sections 3.2, 3.3, 3.4, and 3.5, respectively. Section 3.6 then presents different kinds of group communication. A discussion of routing is given in Section 3.7, and the chapter ends with an overview of different high-performance networks for coarse grained systems in Section 3.8.

3.1 Design and Performance Parameters

The terminology used in the literature on interconnection networks can vary a bit. This chapter states definitions of terms for which some are used in the thesis, while others are included just to give a hint of which parameters that can characterize interconnection networks. Various design and performance parameters and desirable features of interconnection networks for parallel computers are found in textbooks 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], compilations of papers [Scherson and Youssef 1994] [Varma and Raghavendra 1994], and tutorial texts on interconnection networks for parallel computers [Bhuyan et al. 1989] [Duato et al. 1997] [Reed and Grunwald 1987] [Siegel 1990]. Furthermore, more general network discussions and concept definitions are found in computer communication textbooks [Halsall 1995] [Peterson and Davie 1996] [Stallings 1997] [Tanenbaum 1996]. We will explain some of these below, of which several have an 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 misleadingly [Freeman 1998], called bandwidth.

Throughput
The term throughput is used when describing efficiently used transmission capacity at different assumptions such as 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 (Processing Element) 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 an equal numbers 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 wishes to bring the cost down as far as possible. Instead of measuring cost in monetary terms, it can be measured in, e.g., number of links or switches.

Conflict free
The ideal network should be conflict-free. If there exists 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., a limited amount of buffers must still be considered, however. 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 two nodes that communicate 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 the communication between a pair of PEs can start. The "physical" channel does not need to be a purely physical but can, e.g., be a cyclically available time slot on a time multiplexed channel. The advantage of 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 cannot be reused. When using packet switching, the data to be transferred are split into packets that compete for bandwidth with packets from other nodes. Traffic situations with temporary bursts of large volumes of data from one or a few PEs can therefore be handled better in a packet-switched network than in the case that much of the bandwidth is allocated by other nodes using circuit switching. Also, sporadic traffic often experiences a shorter latency than in the case that a circuit must 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 transfer, the logical connection is disconnected. During data transfer, each correctly received packet (or similar) is acknowledged to the transmitting party. If it is not acknowledged, or retransmission is explicitly requested, a packet is retransmitted. Also, to guarantee that packets arrive in the correct
(transmitted) order, each packet is stamped with a sequence number. Connectionless service does 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**

Sometimes, scalability refers only to scalability in one dimension, e.g., size. In this way, *size scalability* can denote the ability to build arbitrarily large systems with a chosen (network) architecture and increase the number of nodes with no or small modifications of the architecture. Just saying scalability often denotes that performance (e.g., bandwidth) scales with the number of nodes. As an example, when more PEs are added to a scalable system, performance in terms of, e.g., network capacity should, in order to avoid bottlenecks, increase proportionally with the number of PEs.

**Incremental expandability**

It is often desirable to be able to expand a system with another small subsystem instead of, e.g., being forced to double the number of nodes to maintain a certain topology. 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 related to, the term incremental expandability.

**Control strategy**

The control strategy can be either centralized (global) or decentralized (local) and normally refers to the way in which switches are steered 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.
A certain static topology is chosen in static networks. Some common topologies are linear array, ring, 2-dimensional mesh, 2-dimensional torus, and binary hypercube (Figure 3). The parameters below are used to describe static networks.

3.2 Static Networks

A certain static topology is chosen in static networks. Some common topologies are linear array, ring, 2-dimensional mesh, 2-dimensional torus, and binary hypercube (Figure 3). The parameters below are used to describe static networks.

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

**Node degree**
The node degree, the number of links that connect a node to its nearest neighbors, can either be constant for the whole network or differ between the nodes. As an example, the boundary nodes in a two-dimensional mesh have a node degree of 3 and corner nodes a node degree of 2, while the rest of the nodes have 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 bi-directional 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-\( \beta \) [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. Multiple access protocols for LAN type ring and multiple ring networks are found in [Bhuyan et al. 1989B].

The node degree of the two-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 two-dimensional case, and a uniform node degree. Examples of parallel computers with a two-dimensional mesh interconnection network are the distributed shared memory multiprocessor from MIT, the MIT Alewife [Agarwal et al. 1995], and the MPP SIMD computer (with the extension of reconfigurable function of the boundary nodes) [Batcher 1980] [Batcher 1980B]. A three-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 two-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 two-dimensional torus network [Ishihata et al. 1997], while the CRAY T3D [Kessler and Schwarzmeier 1993] [Koeninger et al. 1994] and T3E both use the three-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 be travelled 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 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].

3.3 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 increase 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 $16 \times 16$ crossbar network for 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 function is reduced $N \log_2 N$, but where $\log_2 N$ stages must be traversed to reach the destination. An example of such a network is the Omega network (Figure 4) which provides exactly one path from every input to every output. The four different switch functions of the $2 \times 2$ switch that is used as building block are shown in Figure 5, where the two rightmost configurations are used for broadcast. Switches larger than $2 \times 2$ can also be used. Each stage of the switches in an Omega network is preceded by a perfect-shuffle pattern. In contrast to 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 as a result of already existing paths that block the way. 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].
networks is another category of multistage networks where it is always possible to find a path through the network. However, if not all paths are routed at the same time, it may be necessary to reroute already existing paths. An example of a rearrangeable network is the Benes network shown in Figure 6. Other multistage networks include Banyan networks [Goke and Lipovski 1973].

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]. Switchboards each having multiple $8 \times 8$ switches, coupled as $4 \times 4$ bi-directional switches, are used as building blocks.

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 become a bottleneck. This problem is solved in the scalable fat-tree network where links closer to the root have higher bandwidth [Leiserson 1985]. An example of a (not fully scalable) switched fat-tree network built of $6 \times 6$ switches is shown in Figure 7. Each link in the figure is bi-directional. The CM-5 is an example of a parallel

Figure 4: Omega network for an eight-node system. One path through the network is highlighted.

Rearrangeable networks is another category of multistage networks where it is always possible to find a path through the network. However, if not all paths are routed at the same time, it may be necessary to reroute already existing paths. An example of a rearrangeable network is the Benes network shown in Figure 6. Other multistage networks include Banyan networks [Goke and Lipovski 1973].

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]. Switchboards each having multiple $8 \times 8$ switches, coupled as $4 \times 4$ bi-directional switches, are used as building blocks.

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 become a bottleneck. This problem is solved in the scalable fat-tree network where links closer to the root have higher bandwidth [Leiserson 1985]. An example of a (not fully scalable) switched fat-tree network built of $6 \times 6$ switches is shown in Figure 7. Each link in the figure is bi-directional. The CM-5 is an example of a parallel

Figure 5: Possible states of a $2 \times 2$ switch.
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].

3.4 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 great 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, e.g., in Silicon Graphics Power Challenge [Silicon 1994]. Multiple buses can be used to enhance performance relative to single-bus systems [Mudge et al. 1987].

Figure 6: A rearrangeable Benes network.

Figure 7: Fat-tree of switches where nodes are leaves in the tree.
3.5 Hybrid Networks

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

3.6 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 in which 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, at which all members must arrive 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 \( 1 \leq i \leq M \) and \( M \) denotes the number of nodes, 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 $i - N + 1 \leq j \leq i$.

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].

3.7 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. Source routing can be more easily used in
parallel computers than in, e.g., internet communication [Comer 1995],
because the topology does not change as often and is not normally 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 must 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 as compared with static routing and
source routing, even though some support for different paths to choose
between can be incorporated into source routing systems.

When a packet arrives at a router, it can be stored and error checked before
an attempt is made to forward it. This method is called store-and-forward
and may 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 must arrive before the
packet can be forwarded to an output port [Kermani and Kleinrock 1979].
This means that it is not necessary to store the whole packet 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 free output port, 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
[Pinkston and Warnakulasuriya 1997] in wormhole routing networks (two
or more "worms" block each other) and similar networks 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, among other things, to decrease
blocking in a network and to guarantee bandwidth to virtual circuits [Dally
1992].

3.8 High Performance Networks for Coarse Grained
Parallel Computers

There are several different general high-performance networks that can be
used to connect rather powerful and possibly heterogeneous computing
nodes, which might be physically separated by several tens of meters or
more. Both standards and ongoing research projects exist. Often, these
kinds of networks are used in NOWs (Networks of Workstations) or COWs
(Clusters 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 four, eight, 12, and 16 ports exist. Myrinet
is based on earlier work performed at Caltech [Seitz and Su 1993] and the
University of Southern California [Felderman et al. 1994]. A parallel
computer architecture with a hierarchy of Myrinet switches is reported in
[Boggess and Shirley 1997]. The lowest level in the hierarchy has a switch
that connects several processors on a single board. The same switch is an
interface to the next level, which connects several boards in a backplane.
Other work related to Myrinet has been reported [Prylli and Tourancheau
1998] and includes a multicast protocol where the Myrinet network-
interfaces forward multicast packets along a spanning tree [Bhoedjang et al.
1998]. More references to reports on communication systems in which
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] [Iseinstein 1994] [Mercury 1998]. A RACEway system is built up of six-port (bi-directional) crossbar switches to get an active backplane. Several different topologies can be chosen, but the typical one is a fat-tree of switches, where each switch has four children and two parents. Circuit switching with source routing are used. Support for 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 high performance networks for rather coarse grained 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 $2 \times 1$ 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]
- **GigaRing**: a 600 + 600 MByte/s dual ring network developed by Cray Research for use as a supercomputer interconnect [Scott 1996].

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

The category of interconnection networks for rather coarse-grained systems (typically multicomputers) are especially in focus in this thesis.
Part II: Review of Radar Signal Processing
Systems and Suitable Interconnection
Networks
4. A Sample Radar Signal Processing System

The signal processing system under consideration is primarily developed for use in applications with a phased array antenna, i.e., an antenna with multiple fixed antenna elements (and digital beam forming) instead of a moveable antenna. The system has a number of different requirements depending on the application, but the algorithms are usually well known. They comprise mainly linear operations such as matrix-by-vector multiplication, matrix inversion, FIR-filtering, DFT etc. In general, the functions will work on relatively short vectors and small matrices, but at a fast pace and with large sets of vectors and matrices.

One of the goals of our research is to find a good scalable architecture which can give sufficiently high computing speed for this application without too great a loss in generality, i.e., the use of efficient programmable computers is preferred. A solution to this is to use the two-dimensional array SIMD machine as a building block. The two-dimensional array is well known in the literature, and a number of machines have been built, e.g., MasPar [Blank 1990] [Nickolls 1992] [MasPar 1992], Connection Machine [Thinking Machines 1991], and DAP [Hord 1990]. Numerous successful mappings of algorithms on these machines have been done. However, these machines are not efficient when the calculations are too small for the machine size, i.e., it is difficult to fit a 32 by 32 matrix problem on a 65k processor machine. Thus, it can be noted that the mesh is a promising architecture, but that the size of the mesh should be in the order of the size of the data structures in the calculations, i.e., the size of the matrices in the data set. To cope with the large number of matrices in the data, many computation modules, each with the two-dimensional array topology, can be interconnected to share the load. A very powerful interconnection network is needed for this, because each computation module (hereafter also called node) can produce a sustained data flow of several Gbit/s. Guaranteed bandwidth must also be supported in order not to disturb the dataflow.

This chapter briefly describes our proposal for a computer system, which is a MSIMD mesh system, intended to meet the imposed requirements in terms of computing power, generality, size, and power consumption. The computation module architecture is presented in Section 4.1, and system design issues are discussed in Section 4.2. The communication demands are treated in Section 4.3. More detailed descriptions of the radar signal processing systems developed are found in other publications by our research group [Taveniku et al. 1996] [Taveniku et al. 1998] [Ahlander 1996] [Taveniku 1997].
4.1 Computation Module Design

Several similar computation module designs have been developed and evaluated in the scope of the projects of which this thesis has been part. One is the computation module shown in a simplified form in Figure 8 [Taveniku et al. 1996]. Each module consists of eight 8-by-8 meshes working in an SIMD fashion. Internally in the meshes, the processors (PEs) are connected with nearest neighbor connections (x-grid), together with row and column broadcast lines. Computations and inter-PE communication can be performed simultaneously. In addition to the PE meshes, the module holds I/O buffers, memory, and control units.

4.2 System Design Issues

The overall system design is a scalable, moderately parallel MIMD system with moderately parallel SIMD modules interconnected with a scalable, optical real-time interconnection network. The applications are implemented on the system as follows. A multi-mode radar application can in general be described as a set of independent modes of operation. Furthermore, each mode is described as a series of transformations on the sampled data stream. In addition to this, there are control functions and registration functions controlling and monitoring the system.
The system implementation process is illustrated in Figure 9. First, the modes of operation can be designed separately. The modes are then functionally decomposed and laid out on the system (MIMD level). In this stage, only the computational demands and the network bandwidth demands need to be considered. When the functions are assigned to processing nodes, these can be further mapped onto the processors, and buffer space can be allocated. This task is simplified by the fact that only a limited number of algorithms are used, therefore making it possible to use and develop a function library. Furthermore, owing to the properties of the application, data are inherently parallel and therefore easy to map onto these small processor arrays. The chosen approach gives: (i) intuitive use of the MIMD system, (ii) understandable, high level functional decomposition, (iii) scalability on the MIMD level, and (iv) efficient mapping of algorithms because of small modules.

4.3 Inter-Module Communication Demands

As an example, we show how signal processing in a ground based surveillance radar system is implemented using the approach described in Section 4.2 [Taveniku et al. 1996]. In the system, 64 lobes are created that use 64 receiver channels. The data flow in one mode of operation can be described as shown in Figure 10, while the total computational demand is 40 GFLOPS. The nodes in this system are SIMD computers of the array-of-meshes type (see Section 4.1) with a sustained performance of 4 GFLOPS. The functions are mapped onto the nodes as shown, together with the bandwidth demands, in Figure 11. The algorithms are then individually mapped onto the specific processor array, in this case an 8-by-8-by-8 mesh PE array as shown in Figure 8. The chain will only figure as a sample.
system with different communication requirements. Thus no details about the chain are covered here.

The figure shows how the work is split in a coarse grained MIMD fashion, where each computational module is itself powerful (containing multiple processors in this case). 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 Gbit/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.

Corner-turning of the data cube is done when the PEs must process data along another dimension of the data cube. The memory modules are used for this task. A memory module stores incoming messages from the communication system in such a way that the whole corner-turned data cube is finally 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.
Although control I/O and high speed data I/O are logically separated in the module, a single network interface handles all intermodule communication. Since each module itself can have a sustained output data rate of several Gbit/s, a powerful interconnection network is needed. Another important feature of the network is the ability to guarantee that the time constraints are met, i.e., the data flow must not be disturbed by, for example, status information that the network must also transport. A network that can guarantee delivery of semi-static high bandwidth traffic at the same time that it carries rapidly changing control and status traffic is therefore valuable, or even required. We suggest two different ways of implementing the communication network, both employing fiber-optic technology and supporting guaranteed timely delivery. The first network is a WDM star network which scales to large systems, and the second is a fiber-ribbon ring network suitable for systems of a moderate size and/or systems with high degrees of nearest neighbor communication. These two main network architectures, together with different design alternatives, are described in Chapters 7 through 10.

Figure 11: Data flow between the modules in the signal processing chain.
5. Configurations and Requirements of Signal Processing Systems

Future radar signal processing systems will have high computational demands, which implies that parallel computer systems are needed. In turn, the performance of parallel computers is highly dependent on the performance of their interconnection networks. Next chapter, reviews optical interconnection networks from a signal processing perspective. The networks are evaluated according to their suitability for mapping signal processing chains in, e.g., radar systems. Different ways to map are explained. These kinds of mappings are simplified in their nature to facilitate an evaluation of the diverse range of networks. More detailed discussions of algorithm mapping in similar signal processing systems are found in [Liu and Prasanna 1998].

Sections 5.1 through 5.7, describe different kinds of parallel algorithm mapping and performance criteria for use in the analysis.

5.1 Pure Pipeline Chain

The simplest case considered is a single, straightforward pipeline chain. Such a system is shown in Figure 12, 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 represent only 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 sample chain in Figure 11 (Page 77), without, e.g., multicasting to simplify the discussion of performance.

![Figure 12: A pure pipeline chain where one or more stages in the chain are mapped on each module.](image)
A common way of mapping radar signal processing algorithms onto parallel computers is to let each processor carry out the same operation but on different sets of data, i.e., SPMD (Same Program Multiple Data). All the PEs in Figure 13 then work together on one of the pipeline stages in Figure 11 (Page 77) at a time. After the processing of one stage, the data cube is redistributed if necessary. 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 two-dimensional mesh is used. The number of times it is necessary to redistribute the data cube between the PEs may however be reduced as compared with the number of times the data cube must be transferred to the 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 must be transferred across the network bisection [Teitelbaum 1998]. The bisection bandwidth is therefore an essential performance parameter in this strategy.

Figure 13: If the SPMD model is used, all PEs work together on one stage in the signal processing chain at a time.
Incoming data from the antenna are not shown in Figure 13 but are assumed to be fed into the PEs in such a way that it does not affect 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.

5.3 Parallel Hardware For Multiple Modes

When there are several concurrent operation modes of the radar instead of only one, these can be mapped in a number of different ways. One way is to have several groups of processing elements, each living its own life and dedicated for one mode. Assuming one of the two kinds of process mapping described in Sections 5.1 and 5.2, we have two ways of mapping concurrent modes totally in parallel.

Figure 14 shows a system of parallel pipeline chains, chain with its own computational modules. The incoming data from the antenna are multicast to the first node in each chain, i.e., all modes operate on all data cubes (instead of operating in an interleaved way, which is also possible).

Figure 15 shows a system with parallel groups of PEs. Each group executes an SPMD program that corresponds to the dedicated working mode. In this case also, some form of multicasting to spread the incoming data is assumed.
The system configurations in which several modes run concurrently on the same set of nodes place a demand on the network to be reconfigurable. Two different ways of mapping are considered here and described below.

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 switches mode of operation in the order of once every ten 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 concurrently running several pipeline chains like the one shown in Figure 12 (but not pure pipeline) might perform a multicast in one mode and one-to-many or one-to-one in the next mode. Similar

![Figure 15: Several parallel groups of PEs, each running an SPMD program.](image)

5.4 Concurrent Modes on the Same Hardware

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

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 switches mode of operation in the order of once every ten 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 concurrently running several pipeline chains like the one shown in Figure 12 (but not pure pipeline) might perform a multicast in one mode and one-to-many or one-to-one in the next mode. Similar
communication reasoning can be used in the system shown in Figure 13, which 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, in the general case, this requires more processing power. Also, which is of greater 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.

5.5 One Data Cube per Processing Element

One can allow the incoming data to be distributed such that each data cube is given to a different processor, as shown in Figure 16. All calculations in the whole signal processing chain are then performed by the same processor for each data cube, therefore requiring no communication before the results are gathered. Although this way of mapping minimizes communication it will typically not be a good solution because the computational latency will be too great. For example, consider a system in which the data are distributed to 100 processors, one data cube at a time to each processor. If 100 percent processor utilization is assumed and the CPI (Coherent Processing Interval, i.e., the time between the start of two subsequent data cubes) is ten ms, then the computational latency will be one s, which is not
acceptable. Hence, this kind of mapping will not be further investigated. Further reasons for why this is not a good choice of mapping are found in [Liu and Prasanna 1998].

5.6 Traffic Types

For many of the cases described above, circuit switching might work well for the data flow in the signal processing chain because the mode is not changed very 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 set up a circuit for this kind of messages, and to always have circuits connected is not viable for sporadic traffic. Nevertheless, packet switching can be supported by a totally different subsystem, e.g., an optical subsystem for circuit switching and an electrical subsystem for packet switching.

5.7 System Sizes and Communication Distances

Different communication systems may fit different ranges of system sizes or communication distances. One interconnection network might, for example, not offer sufficient throughput over a long distance. Another might be physically too large or too expensive compared with other solutions for small computer systems. An evaluation is made to come to qualitative judgements about the communication systems’ suitability for the different system sizes and communication distances given in Table 5.

<table>
<thead>
<tr>
<th>Kind of communication</th>
<th>Communication distances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intra chip</td>
<td>0.1 - 2 cm</td>
</tr>
<tr>
<td>Intra MCM</td>
<td>1 - 10 cm</td>
</tr>
<tr>
<td>Intra board</td>
<td>5 - 50 cm</td>
</tr>
<tr>
<td>Inter board</td>
<td>0.1 - 1 m</td>
</tr>
<tr>
<td>Inter cabinet</td>
<td>0.5 - 10 m</td>
</tr>
<tr>
<td>Inter and intra room</td>
<td>10 - 100 m</td>
</tr>
<tr>
<td>Intra and inter building</td>
<td>100 m - 10 km</td>
</tr>
</tbody>
</table>

Table 5: Classification of system sizes and communication distances.
A number of proposals for optical or optoelectronic communication systems suitable in a radar signal processing system is given in Sections 6.1 through 6.9, some of which are hybrids of several other systems. Although there exist many other optical interconnection architectures that might be candidates, only some selected groups or concepts are selected here to give a reasonably broad view of possible solutions. Somewhat more emphasis is placed on pure fiber-optic solutions than, e.g., free space systems. The survey ends with a summarizing evaluation in Section 6.10. The classification made in the chapter is to some extent influenced by work referred to in Section 1.5. There are some previous surveys and tutorial texts in the field of optical interconnects [Goldberg 1997] [Kurokawa and Ikegami 1996], the latter focusing on parallel computers and ATM switches.

6.1 Fiber-Ribbon Ring Network

Bit-parallel transfer can be utilized when fiber-ribbon cables/links are used to connect the nodes in a point-to-point linked ring network. 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 17, aggregated throughputs higher than 1 can be obtained

![Diagram of a fiber-ribbon ring network](image)

*Figure 17: 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 Nodes 3, 4, and 5.*
in ring networks with support for spatial bandwidth reuse (sometimes called pipeline rings) [Wong and Yum 1994]. This feature can be effectively used in signal processing applications with a pipelined dataflow, i.e., most of the communication is to the nearest downstream neighbor. Two fiber-ribbon pipeline ring networks have recently been reported [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 is a little more complex, and has support for packet switching on 8+1 fibers and uses a tenth fiber for control packets (see Figure 18) [Jonsson 1998]. The control packets carries MAC 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 with circuit switching).

Another fiber-ribbon ring network is the PONI network (formerly USC POLO), which is proposed to be used in COWs 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].

### 6.2 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

---

**Figure 18:** 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, multiple wavelength channels can carry data simultaneously in the network [Brackett 1990]. In other words, each channel has a specific color of light. A flexible WDMA network requires tunable receivers and/or transmitters, i.e., it should be possible to send/listen on an arbitrary channel [Mukherjee 1992].

Figure 19 shows an example of a WDM star network configuration. 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 [Jonsson et al. 1996], which has support for guaranteeing real-time services, both in single-star networks [Jonsson et al. 1997] and star-of-stars networks [Jonsson and Svensson 1997]. One can say that this kind of network architecture 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 can be used [Cheung 1990]. Such components can be used to achieve a cheaper network in 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 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 11 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). Figure 11 shows 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 Gbit/s. A feasible allocation scheme of eight channels (in addition to the broadcast channel) is shown in Table 6.

Complete removal of the ability to tune in a WDM star network gives 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 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

<table>
<thead>
<tr>
<th>Node</th>
<th>Input channel</th>
<th>Output channel</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>–</td>
<td>$\lambda_1$</td>
</tr>
<tr>
<td>2</td>
<td>$\lambda_1$</td>
<td>$\lambda_2$</td>
</tr>
<tr>
<td>3</td>
<td>$\lambda_2$</td>
<td>$\lambda_3$</td>
</tr>
<tr>
<td>4</td>
<td>$\lambda_3$</td>
<td>$\lambda_4$</td>
</tr>
<tr>
<td>5</td>
<td>$\lambda_3$</td>
<td>$\lambda_4$</td>
</tr>
<tr>
<td>6</td>
<td>$\lambda_4$</td>
<td>$\lambda_5$</td>
</tr>
<tr>
<td>7</td>
<td>$\lambda_5$</td>
<td>$\lambda_6$</td>
</tr>
<tr>
<td>8</td>
<td>$\lambda_6$</td>
<td>$\lambda_7$</td>
</tr>
<tr>
<td>9</td>
<td>$\lambda_7$</td>
<td>$\lambda_7$</td>
</tr>
<tr>
<td>10</td>
<td>$\lambda_7$</td>
<td>$\lambda_8$</td>
</tr>
<tr>
<td>11</td>
<td>$\lambda_7$</td>
<td>$\lambda_8$</td>
</tr>
<tr>
<td>12</td>
<td>$\lambda_7$</td>
<td>$\lambda_8$</td>
</tr>
<tr>
<td>13</td>
<td>$\lambda_7$</td>
<td>$\lambda_8$</td>
</tr>
<tr>
<td>14</td>
<td>$\lambda_8$</td>
<td>$\lambda_8$</td>
</tr>
<tr>
<td>15</td>
<td>$\lambda_8$</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 6: A feasible allocation scheme of the wavelength channels in a WDM star network for the sample radar system.
A WDM ring network utilizes ADMs in all nodes to insert, listen, and remove wavelength channels to/from the ring. In the WDMA ring network described in [Irshid and Kavehrad 1992], each node is assigned a node-unique wavelength on which to transmit. The other nodes can then tune in an arbitrary channel on which to listen. 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.

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. In this way, a smaller number of wavelength channels is 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 sufficient for the signal processing chain shown in Figure 11, if the ADMs in Nodes 1, 2, 3, 6, 7, 8, and 9 terminate the channel for wavelength reuse.

6.4 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 22, 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 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, some of which are discussed below. One way is to embed a ring with bit-parallel transmission and the possibility for spatial bandwidth reuse, as described in Section 6.1. Of course, this leads to the same good performance for pipelined data flows as does the fiber-ribbon ring network; the medium is simply changed into 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 Mbit/s fiber-ribbon links were done (see Figure 23) [Parker 1991] [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 and/or the distance. Advantages are simple hardware owing 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 does not become 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 that 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 can instead be realized by bitserial transmissions over a fully connected topology, i.e., there is one channel between each pair of nodes (see Figure 24) [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. In this way we get an architecture with the support of a single broadcast at a time. Sending on only 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 $N^2$. Clock recovery circuits must also be involved. The distressed power

Figure 23: Array of passive optical stars connects a number of nodes via fiber-ribbon cables.
budget actually offers an advantage over a system with passive splitters or stars (this also holds for, e.g., the point-to-point connected ring). The flexibility of a crossbar makes the network good for radar systems with the SPMD mode.

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 so called via hole to the reversed side of the PCB.

### 6.5 Optical Interconnections and Electronic Crossbars

Communication systems such as in Myrinet [Boden et al. 1995], where arbitrary switched topologies can be built by using electrical switches, can support a number of different traffic patterns possible in radar systems. Fiber-ribbons can be used to increase bandwidth, compared to electrical systems, while still sending in bit-parallel mode. It is possible to have bit rates in the order of 1 Gbit/s over each fiber in the ribbon over tens of meters using standard fiber-ribbons. As noted in Section 6.4, 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 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 such as 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, there is only communication over shorter distances (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 µm, to match the core of a multimode fiber, have been reported [Coldren et al. 1998]. Other work on the integration of components for short distance (non telecom) WDM links has 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 are coupled directly to optoelectronic devices on the chip is possible [Szymanski et al. 1998]. Attaching 32 incoming and 32 outgoing fiber-ribbons with 800 Mbit/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 was implemented for switching 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 are handled by a global mechanism. The mechanism operates in order to get mutual exclusion among packets 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].

### 6.6 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 two-dimensional array (e.g., on a chip) where each smart pixel consists of a processor, a laser diode and a photo diode. Other, but 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 25). Between arrays, holographic interconnects or other optics might be used to steer or switch the optical channels. The row arrangement is especially suitable in applications in which computations can be mapped in a pipeline fashion with one pipeline stage per array, e.g., image processing [McArdle et al. 1996] [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 edge 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. 1998] [Chen et al. 1998B]. Each array operates in SIMD mode on two-dimensional data. A modified CSMA/CD 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 a VCSEL based smart pixel array. Power consumption and pixel homogeneity (lack of 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 having a 300 Mbit/s I/O port. The estimated maximum total I/O throughput for such a system is 600 Gbit/s. VCSELs are better than edge-emitting laser diodes for free space communication mainly because of the possibility of two-dimensional VCSEL arrays and because integration with electronic circuits is simpler [Kurokawa et al. 1998].
6.7 Optical and Optoelectronic Switch-Fabrics

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

SDM (Space Division Multiplexing) switches [Goh et al. 1998] [Guilfoyle et al. 1998] [Kato et al. 1998] [Lai et al. 1998] [Moosburger and Petermann 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 larger 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 using chips with processing elements placed on a two-dimensional plane [Christensen and Haney 1997]. The processors then communicate with each other by a mirror that bounces back the beam 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 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 be placed on a dedicated board in a cabinet and be connected to processor boards via fibers or via 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 between the transmitters and receivers has been demonstrated [Li et al. 1998B]. All optical channels turned on from a transmitter's two-dimensional \( \sqrt{N} \times \sqrt{N} \) VCSEL 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 such a way that the light from each VCSEL pixel in a VCSEL array is focused on a single spot together with the corresponding pixels in all other arrays. This gives \( N \) spots where each is focused into a single fiber leading to a receiver. Hence, selecting a pixel in a VCSEL array to be turned on corresponds to addressing a destination node.
6.8 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 (Figure 26) [Jahns 1994] [Jahns 1998] [Sinzinger 1998]. Electronic chips are placed in a two-dimensional plane, while light beams travel in a three-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, such as the two-dimensional mesh, or to have special optical or optoelectronic devices dedicated to switch functions. The latter configurations may be the best choice if the SPMD program model is used.

6.9 Free Space Optical Backplanes

Several different optical backplanes have been proposed, three of which are discussed below. As shown in Figure 27a, using planar free space optics is one means of transporting optical signals between PCBs. Holographic gratings can be used to insert and extract the optical signals to/from the waveguide, which may 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].
In the system shown in Figure 27b, two-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 pass to the local PCB and some data are retransmitted to the next PCB [Supmonchai and Szymanski 1998]. Each smart pixel array can typically contain 1 000 smart pixels arranged in a two-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, where the ring can be reconfigured to embed other topologies [Szymanski and Hinton 1995] [Szymanski and Supmonchai 1996].

The configuration shown in Figure 27c 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 it arrives at the receiver. A regeneration of the optical signal (multi-hop) might be needed on the way from the source to the final destination.

Figure 27: Optical backplane configurations: (a) with planar free space optics, (b) with smart pixel arrays, and (c) with a mirror.
Of the three types of optical backplanes discussed, the one with smart pixel arrays seems to be the most powerful. On the other hand, a simple passive optical backplane may have other advantages. Other 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].

<table>
<thead>
<tr>
<th>Network</th>
<th>Pipeline</th>
<th>SPMD</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fiber-ribbon pipeline ring</td>
<td>Good</td>
<td>Moderate</td>
<td>Good for SPMD too if enough bandwidth</td>
</tr>
<tr>
<td>WDM star, rapid tuning</td>
<td>Good</td>
<td>Good</td>
<td>Flexible passive network</td>
</tr>
<tr>
<td>WDM star, slow tuning and broadcast channel</td>
<td>Good</td>
<td>Poor</td>
<td>WDM star alternative that may be cheaper</td>
</tr>
<tr>
<td>Multi-hop WDM star</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Can be optimized for pipelined mapping</td>
</tr>
<tr>
<td>WDM ring with rapid tuning</td>
<td>Good</td>
<td>Good</td>
<td>More channels might be needed for SPMD</td>
</tr>
<tr>
<td>WDM ring, slow tuning and broadcast channel</td>
<td>Good</td>
<td>Poor</td>
<td>WDM ring alternative that may be cheaper</td>
</tr>
<tr>
<td>Fiber ribbons and array of stars</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Power and bandwidth limited</td>
</tr>
<tr>
<td>Fully connected topology with broadcast driving</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Bandwidth limited. Cost grows with $N^2$</td>
</tr>
<tr>
<td>Fully connected topology with flexible driving</td>
<td>Good</td>
<td>Good</td>
<td>Cost grows with $N^2$</td>
</tr>
<tr>
<td>Optical fibers and electronic crossbars</td>
<td>Good</td>
<td>Good</td>
<td>Optoelectronics also needed in switch</td>
</tr>
</tbody>
</table>

Table 7: Performance summary of some of the networks discussed with respect to pipeline and SPMD mapping.

Of the three types of optical backplanes discussed, the one with smart pixel arrays seems to be the most powerful. On the other hand, a simple passive optical backplane may have other advantages. Other 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].

6.10 Summary

Some of the networks that incorporate fiber-optics are summarized in Table 7 with remarks on their suitability/performance for the two basic cases of mapping (pipeline and SPMD). The pipeline mapping fits on a larger variety of networks because of the absence of all-to-all traffic patterns. Limiting factors to the use of SPMD mapping vary from network to network; these may be tuning speed when switching from many different sources and destinations, shared resources that become bottlenecks and a topology that favors nearest-neighbor communication.

Performance estimations for the case of mapping several parallel working modes, each with either pipeline or SPMD mapping, are summarized in Table 8. 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 node in each group of nodes dedicated to a working mode. In the case of SPMD, the incoming data are 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. This is not a problem for most networks.

Having optics inside a switch gives the same flexibility as electronic crossbars, but it might be possible 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 great deal 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 possible topologies etc depending on, e.g., the linear order (or similar arrangements) of smart-pixel arrays and the two-dimensional plane to which components are restricted to when using planar optics.

The suitability of different technologies/networks in different system sizes is summarized in Table 9. The lower bounds on system sizes arise from such

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

Table 8: Performance summary of some of the networks discussed with respect to pipeline and SPMD mapping when several concurrent modes run in parallel, each on dedicated hardware.
factors as miniaturization problems (e.g., fiber-ribbon connectors) and expensive components that are primarily developed today for long distance communication (e.g., WDM components). Some reasons for upper bounds are channel-to-channel skew in fiber-ribbons (especially when there is 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 fiber-ribbon networks discussed. On the other hand, the WDM technique offers flexible multi-channel networks that can be implemented using passive technology. Integrated fiber and waveguide solutions make the building of compact systems possible, especially for networks such as those using fiber-ribbons. The same reasoning about compactness can be argued for free space systems. Optical backplanes may earn their success from the similarities with current rack-based systems.

Not all optical interconnection networks are quite as easy to categorize into the groups above, e.g. those described in [Chiarulli et al. 1994] [Teza et al. 1995] [Louri et al. 1996], the latter describing a network using both free-space and fiber-ribbon interconnects. In another system, plastic modules performing different functions are snapped together to form a free space optical interconnection system [Neilson and Schenfeld 1998]. Examples of module functions are relaying and beam splitting of a two-dimensional array of incoming beams.

After describing a radar signal processing system in Chapter 4 and different requirements and configurations in Chapter 5, we have now, in Chapter 6, evaluated a number of optical network configurations.

<table>
<thead>
<tr>
<th>Kind of communication</th>
<th>WDM star/ribbon</th>
<th>Fiber-ribbon</th>
<th>Foil of fibers</th>
<th>free-space</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intra chip</td>
<td></td>
<td></td>
<td></td>
<td>Good</td>
</tr>
<tr>
<td>Intra MCM</td>
<td></td>
<td></td>
<td></td>
<td>Good</td>
</tr>
<tr>
<td>Intra board</td>
<td></td>
<td>Good</td>
<td></td>
<td>Good</td>
</tr>
<tr>
<td>Inter board</td>
<td>Expensive</td>
<td>Good</td>
<td></td>
<td>Good (Good)</td>
</tr>
<tr>
<td>Inter cabinet</td>
<td>Expensive</td>
<td>Good</td>
<td></td>
<td>Good</td>
</tr>
<tr>
<td>Intra and inter room</td>
<td>Good</td>
<td>Good</td>
<td>Moderate</td>
<td></td>
</tr>
<tr>
<td>Intra and inter building</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 9: Suitability in different system sizes. Empty cells in a column mean that the technology/network is not suitable for the corresponding system sizes.*
Part III: Proposed Network Architectures and Protocols
7. FT-TR WDM Star Network

High-performance interconnection networks can be foreseen to have a central role in future distributed real-time systems. If a number of computation modules, each very powerful or even parallel, are used to obtain a massively parallel or distributed system, a modular interconnection network able to carry a huge amount of data is needed. Other key features of the network are time deterministic latency and guarantees to meet deadlines. Application examples are future radar signal processing systems, distributed multimedia systems, and image processing systems. A typical system is the radar signal processing system described in [Jonsson et al. 1996] [Taveniku et al. 1996], where each module consists of a SIMD computer and a network interface. In this way, an MSIMD computer system is formed. Other applications in which the MSIMD architecture with a high performance interconnection network may be required are described in [Davis et al. 1992] [Svensson and Wiberg 1993]. It should noted that other MIMD systems can have similar network requirements as MSIMD systems.

The WDM technique [Hill 1989] [Green and Ramaswami 1990] offers extremely high aggregated bandwidth by the use of multiple wavelength channels in fiber-optic communication systems, and is expected to have a key role in future computer networks. Each channel has a specific wavelength, i.e., color, of the light. A promising network architecture for high performance systems, for which commercial components have already appeared, is the WDM star network with multiple Gbit/s channels. This is based on a passive optical star which implements a fiber-optic multi-access network [Mestdagh 1995] where all incoming packets, in the form of trains of light pulses, to the star are distributed to all nodes in the network by splitting the light.

The remainder of the chapter is organized as follows. First, an introduction to the network and protocol is presented in Section 7.1 together with related networks. The protocol is then presented in Section 7.2, while the implementation of real-time services is described in Section 7.3. Section 7.4 discusses some aspects of implementation. The analysis of the network is made in Sections 7.5, 7.6, and 7.7, which respectively present deterministic performance, the case study, and simulation results. Section 7.8 gives a conclusion and summary.
7.1 Protocol Overview and Related Networks

7.1.1 Circuit Switched Networks
A number of protocols for WDM star networks have been proposed. However, the area of real-time protocols for these networks is relatively unexplored, with a few exceptions. The Rainbow network, described in [Dono et al. 1990] [Janniello et al. 1992], 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 multichannel extensions to static uniform TDMA.

7.1.2 Packet Switched Real-Time WDM Networks
A real-time protocol for packet switched communication in WDM star networks is described in [Yan et al. 1996B]. The QoS 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 to keep their QoS by adaptively changing the priority of queued packets. Although dynamic real-time properties are supported, the matter of the success or failure of a packet transmission depends on the global state of the network, and transmission success can not be guaranteed in advance.

A class of real-time protocols for CC-FTTT-FRTR (see Section 2.5.4, page 48) WDM star networks is proposed in [Maode et al. 1998]. The protocols support global "optimization" of the real-time traffic, i.e., the order of transmission is calculated such that packets from all nodes are sorted at the same time according to their real-time requirements, instead of just being locally sorted in each node. However, each node in the network requires both a tunable transmitter and an additional transmitter/receiver pair for the control channel. Moreover, the protocols do not support the ability to guarantee timely delivery at the time of message generation.

Other work reported on real-time support in similar WDM networks include:

- A star network with support for allocation of isochronous channels (guaranteed bit-rate and bounded network delay with low jitter), but without support for aperiodic traffic, which is essential for, e.g., control traffic in parallel and distributed systems [Wang et al. 1997].
• A star network with support for virtual circuits, both with and without (best effort) guaranteed bit-rate, but with the need of a dedicated control channel and a central scheduling node and without support for aperiodic traffic [Kam et al. 1998] [Kam et al. 1998B].

• A WDM bus network with support for real-time streams but not for guarantees to separate messages [Cho et al. 1995].

• A network in which all nodes track already guaranteed but not sent messages to be able to decide whether to guarantee new traffic. However, there is no support for RTVCs or the like, and a control channel with \( M \) mini slots per data slot, where \( M \) is the number of nodes, is needed (can be difficult to implement owing to, e.g., clock synchronization problems) [Chen and Georganas 1993].

• A network with support for real-time traffic by the use of priorities, but without support to guarantee delivery of messages before their deadlines [Selvakennedy and Ramani 1996].

Research on pre-allocation of real-time message streams has also been carried out [Dong et al. 1998] [Tyan et al. 1996].

7.1.3 A New Protocol with Real-Time Support
This chapter gives a proposal for a medium access protocol for packet switched real-time communication in WDM star networks. The protocol, which is time deterministic, is called TD-TWDMA and supports guaranteed real-time services for both single destination, multicast, and broadcast transmission. Slot reservation is also supported, while bandwidth is used efficiently owing to a simple slot release method. The protocol uses both time and wavelength division multiplexing and is targeted for distributed real-time systems, especially very high performance systems, e.g., by parallelism within the nodes. The deterministic properties of the protocol are theoretically analyzed, while computer simulations show the performance in a network with general traffic.

We propose to expand the single-star network to a star-of-stars topology (Figure 28) for larger systems, where a backbone star connects several WDM star networks (clusters). Each cluster then has an electronic gateway-node as an interface to the backbone star. By the use of electronic gateway nodes, we retain the popular WDM star network architecture in each cluster, for which cheap components can be expected to appear in the future. With electronic gateway nodes, we also achieve wavelength reuse in each cluster and in the backbone.
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.

Fixed wavelength transmitters and tunable receivers are used in each cluster or single star network (Figure 29). 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 [Brackett 1991] [Dono et al. 1990], and general information on WDM star networks is given in [...].
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 for choosing FT1-TR1 is the naturally 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]. However, the practical limit in number of wavelengths in the kind of networks reported in this chapter is expected to be somewhere between 16 and 32 [Brackett 1996]. This translates into star-of-stars networks of maximum sizes between 256 and 1024 nodes, gateway nodes included.

The TD-TWDMA protocol is used separately in each cluster and in the backbone. Therefore, in this chapter, the protocol is described as running in a single-star stand-alone network. Inter-cluster communication [Jonsson and Svensson 1997] is covered in the next chapter.

7.1.4 Protocol and Real-Time Features

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 aware of it immediately.

In a network using the TD-TWDMA protocol, real-time services to support both guarantee seeking and best effort messages, exist for both single destination, multicast, and broadcast transmission. With TDMA, 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 guaranteed messages 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 at the same time deterministic bandwidth is achieved. A node can increase its time deterministic bandwidth, used for guarantee seeking messages, by slot reserving. Another way of using slot reserving is for the allocation of RTVCs [Arvind et al. 1991]. The slot release method is also used for the reserved slots.

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 are transmitted on and they contain information about which guaranteed slots should be kept and which should be released. These types of WDM networks without a separate control channel are denoted as non control channel based networks. Networks in which 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 than the one presented in this thesis are found in [Dono et al. 1990] [Ganz and Koren 1991] [Ganz and Gao 1992], while control channel based networks are found in [Bogineni and Dowd 1992] [Habbab et al. 1987] [Chen et al. 1990] [Chipalkatti et al. 1992]. FatMAC, presented in [Sivalingam and Dowd 1995], is another non control channel protocol for 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.

### 7.2 Protocol Description

The network and protocol proposed will now be described in detail. The notation used when describing the network is found in Table 10. This notation is used in the next chapter as well. Because each transmitter in the network has a specific wavelength, the number of wavelengths, $C$, hereafter denoted as channels, equals the number of nodes, $C = M$. The transmitter
and receiver parts of the transceiver are independent and can work concurrently.

There are $2M$ queues in each transmitter, $M$ queues 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 (messages destined to more than one node but not to all).
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.

Section 7.2.1 explains the function of the cycles, while the distributed slot allocation algorithm is presented in Section 7.2.2.

7.2.1 Transmitter and Receiver Cycles
Because every transmitter has its own home channel, the only possible conflict is when two or more nodes wish 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.

One cycle is also running in each transmitter, but this is 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 30). In the receiver cycles, shown in the upper table in the figure, each slot in each receiver cycle can be assigned to only 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.

Figure 31 shows how a receiver cycle is partitioned into data slots and control slots. Each node $m_i$, $1 \leq i \leq M$, is assigned one of the $M$ control slots, where it broadcasts control information to all other nodes $m_j$, $1 \leq j \leq M$ and
The control slots are therefore identically assigned in every receiver cycle. When control slots have been gathered from all other nodes, allocation of the data slots in the next cycle can be calculated using the distributed slot allocation algorithm described in Section 7.2.2.

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 beginning the next cycle.

### 7.2.2 Distributed Slot Allocation Algorithm

The TD-TWDMA protocol consists of three steps:

1. Each node transmits a control slot
2. Each node separately runs the distributed slot allocation algorithm
3. The nodes transmit and receive data slots

The ways in which the contents of the control slots is calculated and incoming messages and buffers are handled partly determine the real-time services and are described in Section 7.3. When describing the distributed slot allocation algorithm, it is assumed for simplicity that all broadcasted control slots are received before the algorithm is started. As described in Section 7.3, however, this is not a requirement in a real implementation.

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 32. 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 though the control slots may be in the middle of the cycle, we assume here they have index 13 to 16. The elements denoted with ‘-’ can be exchanged with node indexes but, for simplicity, this is not investigated here.
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 (described below) 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 TDMA system, efficient bandwidth utilization is achieved with the slot release method.

In the predetermined allocation scheme, $v_{ij}$ in matrix $V$ identifies the index of the high priority owner of slot $i$ in the receiver cycle in node $j$. In the same way, $w_{ij}$ in matrix $W$ identifies the index of the low priority owner. The high priority owners are determined by

$$v_{ij} = ((i - 1) \mod M) + 1$$  \hspace{1cm} (1)

for $1 \leq i \leq M(M - 1)$, $1 \leq j \leq M$ and $i \neq j$, while the low priority owners are determined by

$$w_{ij} = (((i - 1) \div M + j) \mod M) + 1$$  \hspace{1cm} (2)

The high priority slots are coordinated to allow for broadcasts but, as described in Section 7.3.4, this is not always the case.

<table>
<thead>
<tr>
<th>Receivers</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Priority</td>
<td>1: High</td>
<td>-</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>-</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>-</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1: Low</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>2: High</td>
<td>1</td>
<td>-</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>-</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>2: Low</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3: High</td>
<td>1</td>
<td>2</td>
<td>-</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>-</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>3: Low</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>4: High</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>4: Low</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Figure 32: Allocation scheme for the receiver cycles in a four-node system.
The release message mentioned above is actually a specified value in the control slot matrix, $X$, that each node transmits in the control slot. The matrix has the same size as matrix $V$, and each element in the matrix gives information about the corresponding high priority slot. All elements $x_{ij}$, $1 \leq i \leq S$, $1 \leq j \leq M$, in the matrix are set to zero except for those corresponding to high priority slots belonging to the node, which should not be released. Those elements are instead set to one. This means that a zero in the position of a high priority slot belonging to the node will release the slot.

When all the $X$ matrices (one from each node) have been gathered by a node, two new matrices are composed, $Y$ and $Z$. Each element in $Y$ is used to determine what queue to take a packet from, for transmission in a slot, while an element in $Z$ determines what channel to tune in at the beginning of a slot. The two matrices have elements corresponding to both data slots and control slots during one cycle but, for clearness, the control slot elements are not treated here. The way the elements in $Y$ are set and used to determine the current queue is described in Section 7.3. When the distributed slot allocation algorithm is run, an element states the kind of slot and is defined as:

$$y_i = \begin{cases} 
-1 - j & \text{Low priority slot to node } j \\
0 & \text{No transmission in slot } i \\
1 & \text{High priority multicast/broadcast} \\
1 + j & \text{High priority slot to node } j
\end{cases}$$

for slot $i$ where $1 \leq i \leq M(M - 1)$. Each element is set using the following algorithm for node $m_k$:

1. Let $y_i = -1 - j$, where $j$ refers to the element $w_{ij} = k$ in matrix $W$, i.e., the receiver at which the node has a default low priority slot

2. If $x_{ij} = 1$ in matrix $X$ from node $m_l$, where $j$ refers to the element $w_{ij} = k$ in matrix $W$ and $l = v_{ij}$, then set $y_i = 0$, i.e., the slot was not released by the high priority owner

3. If $\sum_{j=1}^{M} x_{ij} > 1$ for matrix $X$ from node $m_k$, then set $y_i = 1$, i.e., a high priority multicast/broadcast slot belonging to the node

4. If $\sum_{j=1}^{M} x_{ij} = 1$ for matrix $X$ from node $m_k$, then set $y_i = 1 + l$, where $l$ refers to the element $x_{il} = 1$ in matrix $X$ from node $m_k$, i.e., a high priority single destination slot belonging to the node
Each element $z_i, 1 \leq i \leq M(M - 1)$, in matrix $Z$ is set using the following algorithm for node $m_k$:

1. Let $z_i = w_{ik}$, i.e., the default low priority owner
2. If $x_{ik} = 1$ in matrix $X$ from node $m_j$, where $j = v_{ik}$, then set $z_i = v_{ik}$, i.e., the slot was not released by the high priority owner

Since each node can independently perform the computations of the slot allocation scheme, it is called 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.

### 7.3 Real-Time Services

In the description of how guarantee seeking and best effort messages are passed through the transmitter to obtain real-time services, the following parts will be explained:

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 at which all control slots are received
5. Action at the beginning of a data slot

Points 1 and 2 will first be explained. In 7.3.2, the third and fourth points are discussed, and the last point is described in 7.3.3. Finally, slot reserving and RTVCs are introduced in 7.3.4.

#### 7.3.1 Arriving Messages

The ability to give deadline guarantees for a message relies on knowing when there are guaranteed slots in the forthcoming cycles. Element $g_{ij}, 1 \leq i \leq P, 1 \leq j \leq M + 1$, in matrix $G$ holds the number of high priority slots belonging to the node in the $i$:th cycle next to the currently running one, where $j = 1$ for broadcast slots and $2 \leq j \leq M + 1$ for single destination slots to node $m_{j-1}$. $P$ is chosen at system design and tells us how far in the future, relative to the current time, deadlines can be guaranteed. Broadcast and multicast do not have to be treated separately 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 $g_{ij}$, starting with $i = 1$ and with $j$ set to the actual type of arriving message, we can see whether there are enough high priority slots,
before deadline, to guarantee that a message will 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 G 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 are also scanned. Packets are always put in the guarantee seeking queue that corresponds to the element in G that was decremented in the order of transmission to facilitate reassembling the message.

Best effort messages may be transmitted in order, according to, e.g., the EDF (Earliest Deadline First) algorithm, but in the scope of this thesis they are assumed not to have deadlines specified and to be transmitted in order of arrival. An arriving best effort message is hence simply put in the correct best-effort destination-queue. Multicast messages are put in the broadcast queue.

7.3.2 Control Slots
When it is time to send a control slot, the following algorithm is executed in node mk, 1 ≤ k ≤ M:

if g11 > 0
    K = MIN(i, g11), where i is the number of packets in the best effort broadcast queue
endif

count the packets in each guarantee seeking queue to set hi, 1 ≤ i ≤ M + 1, i ≠ k + 1
xij = 0, 1 ≤ i ≤ M, 1 ≤ j ≤ M(M - 1)
for j = 1 to M, j ≠ k
    scan xij starting with i = 1 and set xij = 1 in hj+1 positions (if possible),
    where the corresponding element in V is vij = k and vil ≠ k, 1 ≤ l ≤ M, l ≠ j
endfor

do the same scan of xij for broadcast high priority slots, first for the h1 guarantee seeking packets and then for the K best effort packets, setting xij = 1 for each corresponding destination that is set in the address field of a multicast (all destinations for broadcast), i.e., releasing slots to destinations not specified

send control slot
\( g_{ij} = g_{i+1,j}, 1 \leq i \leq P - 1, 1 \leq j \leq M + 1 \)

\( g_{ij}, i = P, 1 \leq j \leq M + 1, \) are set to the corresponding number of default (after reservation) high priority slots for each corresponding kind of destination

where \( K \) is the number of best effort multicast/broadcast packets that are to be sent in the next cycle. The elements in matrix \( H \) state the current number of packets in each guarantee seeking queue, where \( h_1 \) is the broadcast queue and \( h_i, 2 \leq i \leq M + 1, i \neq k + 1, \) is the single destination queue to node \( m_{i-1}. \)

When all control slots have been received, the elements in \( Y \) are calculated as described in the Section 7.2.2. However, each element is modified to state the queue from which a packet should be taken when transmitting in the corresponding slot, and is, for Slot \( i, \) defined as:

\[
y_i = \begin{cases} 
-1 - j & \text{Best effort to node } j \\
-1 & \text{Best effort broadcast} \\
0 & \text{No transmission in slot } i \\
1 & \text{Guarantee seeking broadcast} \\
1 + j & \text{Guarantee seeking to node } j 
\end{cases}
\]  

(4)

where \( 1 \leq i \leq M(M - 1). \) \( K \) of the elements in matrix \( Y, \) where \( y_i = 1, \) should be set to \( y_i = -1, \) beginning to scan the elements with \( i = M(M - 1), \) i.e., in descending order.

### 7.3.3 Data Slots

At the beginning of each data slot, \( y_i \) is used to determine the queue from which to take a packet. By the definition of the protocol, there will always be a packet in the addressed queue when it is a guarantee seeking queue or the best effort broadcast queue. However, the best effort single destination queues may be empty. In that case, an empty message is sent, telling the receiver only that there was nothing to send. A packet generated immediately before a best effort single destination queue is checked will be sent in that slot. The minimum latency for best effort single destination messages is therefore zero.

### 7.3.4 Slot Reserving

A node can reserve slots to increase its guaranteed bandwidth. A maximum of \( M(M - 2) \) slots per cycle, Slots 5 to 12 in the example in Figure 32, are allowed to be reserved. Slots \( s_i, 1 \leq i \leq M, \) are not allowed to be reserved and slots \( s_i, MP - M + 1 \leq i \leq MP, \) are control slots. When reserving slots, the
corresponding high priority entry (element in matrix \( V \)) in the receiver cycle, or cycles if broadcast is used, are exchanged with the index of the reserving node. To reduce complexity 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 storing slot assignment schemes for several working modes in the nodes. If the reservation is done by higher layers, a reservation request to the nodes in question is sent in a control slot (together with matrix \( X \)).

Slot reservation can also be used for the allocation of RTVCs where an RTVC, with its reserved slots, is typically dedicated to a specific application task. An RTVC then has a specified worst-case latency and a guaranteed bandwidth (see Section 7.5). The allocation of RTVCs can also be used as a service-primitive when reserving slots for increased bandwidth for guarantee seeking traffic. Each such RTVC is then dedicated to be used for guarantee seeking messages generated by the task owning the RTVC. The mechanism for guarantee seeking traffic described above must then be replicated for each RTVC but is somewhat simpler because only slots reserved for the RTVC in question must be considered.

7.4 Implementation Aspects

7.4.1 Clock Synchronization
The TDMA method used requires synchronized nodes. Harder synchronization will reduce the necessary gap between the time slots. A method to achieve harder synchronization in TDMA star networks is to account for the node-to-star propagation delay so that all packets transmitted in the same slot pass the star simultaneously [Jonsson et al. 1995] [Bengtsson et al. 1994]. A discussion of timing and dispersion in WDM star networks can be found in [Semaan and Humblet 1993]. In this thesis, the interconnection distances are assumed to be short, and the propagation delay is hence neglected.

7.4.2 Clock Recovery
A method to reduce the effect of clock-recovery and tuning latencies in the receivers will increase performance. By duplicating the optoelectronic and clock-recovery parts of the receiver, the time for clock-recovery and wavelength tuning can even be eliminated. This is done by locking one clock-recovery circuit to the currently used channel while the other one recovers bit synchronization for the channel that will be used in the next
slot. In this way, the tuning time of the receiver only needs to be shorter than the duration of one slot, minus the clock-recovery time. For this reason, we assume that the tuning latency in the receivers can be neglected.

7.4.3 Computational Complexity

For each slot, the outcome of the distributed slot allocation algorithm depends only on the control slot information from the corresponding high priority owner. Therefore, the algorithm can start as soon as the first control slot is received. A gap of only one data slot, between the control slots and the beginning of the next cycle, to finish the computation is hence assumed. The small gap positively affects the latency. Only table indexing is then used in the data slots, in transmitters to choose between buffered messages and in receivers to choose channel tuning.

The protocol and its real-time services may seem rather computationally intensive because, for the computation of each slot allocation, operations are done on every element in a matrix column or in a whole matrix. In a real implementation, however, the searches and scans of matrix $V$ and $W$ can be exchanged for simpler operations. The reason is that these matrices do not change for each cycle and, e.g., a matrix search must only be done when the default allocation scheme is changed. Similar reasoning can be used concerning matrix $G$, because only a few elements are changed in each cycle. When all $X$ matrices are received, they can be composed into one $X$ matrix to save memory space by doing an element wise OR. Since all elements in $X$ are binary, the composition can be accelerated using bit-parallel operations. Each element in $X$ corresponds to the element with the same indexes in matrix $V$ (and $W$) and, even if the elements in $X$ change values in each cycle, their relation to $V$ (and $W$) does not change. Hence, similar reasoning as is used with $V$ and $W$ to reduce the computational complexity can be used here as well. The counting of packets in the queues can be eliminated by continuously updating counters when packets are inserted or removed.

7.4.4 Electronic Stars

Electronic stars are also consistent with the network architecture; these give a network which can be implemented with cheap components today or in the near future depending on size and bandwidth (see Figure 33) [Jonsson et al. 1995]. The switching circuit should always be able to simultaneously assign a path to an arbitrary output channel from every single incoming channel (true crossbar). The switch should also be able to map one input channel to many output channels, multicasting, or one input to all outputs, broadcasting (these features are automatically implemented with a true crossbar switch). As shown in the figure, the control unit of the star is connected to the switch in the same way as the other channels are. By this connection, the control unit also receives control packets and can,
with the information in these packets, configure the star according to the TD-TWDMA protocol. The electronic star simulates the passive optical star in a way almost invisible to the nodes. Therefore, the nodes still run the distributed slot allocation algorithm.

Although the electronic star brings additional cost in, e.g., optoelectronic transceivers in the star, the point-to-point links have a positive effect. The difference as compared to the passive optical star is the simplified clock-recovery because bit synchronization can be maintained by continuously sending information, either data or synchronization patterns, over the serial link (see, e.g., [TriQuint 1992]). This can be done due to the fact that each fiber is actually a point-to-point connection and not a shared medium.

7.5 Deterministic Performance

The real-time services rely on time-deterministic communication. The deterministic bandwidth and the worst-case packet latency for this guaranteed bandwidth is analyzed below. As stated above, the analysis treats intra-cluster communication.

When a specific node is the only transmitter in the network, the bandwidth it receives is $B(M - 1)/M$, where $B$ is the channel bandwidth. This gives 87, 94, and 97 percent of the channel bandwidth for 8, 16, and 32 nodes, respectively. This bandwidth is then divided into $B(M - 1)/M^2$, which can
be used for broadcasting, and $B(M-1)^2 / M^2$ for point-to-point communication, i.e., $B(M-1) / M^2$ to each of all the other nodes. When competing with other nodes, the broadcast bandwidth is always guaranteed (deterministic) and can be used for either single destination, multicast, or broadcast messages. This deterministic bandwidth is 10.9, 5.9, and 3.0 percent of the channel bandwidth for 8, 16, and 32 nodes, respectively. If maximum allowed reservation in another node's receiver cycle is possible, the deterministic bandwidth is increased to $B(M^2 - 2M + 1) / M^2$ including the broadcast bandwidth. This gives 77, 88, and 94 percent of the channel bandwidth for 8, 16, and 32 nodes, respectively. When some of the reserved slots are reserved in all receiver cycles, the deterministic broadcast bandwidth is also increased. A node that leaves all slots allowed for reservation to other reserving nodes will have the minimum guaranteed bandwidth possible: $B / M^2$. This gives 1.6, 0.4, and 0.1 percent of the channel bandwidth for 8, 16, and 32 nodes, respectively. However, all this bandwidth has broadcast capability.

The worst-case latency for a single packet gives an essential performance measurement of networks for real-time systems and will therefore be analyzed for the proposed network. It is here assumed that only high priority slots can be used and the control packet must hence be transmitted first to claim the use of the high priority slot. The time it takes to finish the slot allocation algorithm, $\mu$, sets the limit for how late in the cycle the control slots can be placed. As shown in Figure 34, the worst-case latency for node $m_i$, $1 \leq i \leq M$, is

$$\tau_{max} = S\gamma + M\gamma + \mu = (M + 1)\gamma + \mu$$

which consists of: (i) the worst-case delay before the own control slot appears, i.e., one cycle as shown in the figure, (ii) the delay from the control slot to the first owned data slot. By fine grain interleaving of the slots, where each node is assured to have one of the first $M$ slots, we can minimize the second part of the latency so that node $m_i$, $1 \leq i \leq M$, will never have a slot later than slot $s_i$ as its first slot in the cycle. As explained above, $\mu$ can be assumed to be the duration of one slot. This gives, for a network with a
slot duration of 1 $\mu$s, a worst-case latency of 73 $\mu$s, 0.27 ms, and 1.1 ms for 8, 16, and 32 nodes, respectively. Note that the actual latency may be significantly lower, especially when many low priority slots are available since those slots can be used without first waiting for control slot transmission.

The worst-case latency of an RTVC belonging to node $m_i$ is

$$\tau_{\text{max}} = s_i \gamma + (M + j - i) \gamma + \mu = (M^2 + M + j - i) \gamma + \mu$$  \hspace{1cm} (6)

if $s_i$ is the first slot in the cycle reserved for the RTVC. The bandwidth of an RTVC is $NB/M^2$, where $N$ is the number of slots reserved for the RTVC.

### 7.6 Case Study

The signal processing system shown in Figure 11 on page 77 was chosen as a case study, where the bandwidth demands on the inter-module communication are also given. Although the data flow has a very simple structure, there can exist control messages in other directions than the data flow. Also, the algorithms, and hence the data flow, can be changed when switching to a different working mode. A general network is therefore required. For simplicity, the maximum data flow from one module is assumed to be 6.0 Gbit/s, excluding, e.g., error checking codes. The latency budget for the whole chain is 100 ms, in which 10 ms is included for inter-module communication. With the ten pipeline stages (some of the stages include several nodes) shown in the figure, including the antenna, the maximum allowed latency per link is $\tau = 1.0$ ms.

To guarantee a node bandwidth of 6.0 Gbit/s, maximum slot reservation is made for these pipeline stages (RTVCs can be used). This gives

$$6.0 \text{ Gbit/s} = BM(M - 2) / M^2$$ \hspace{1cm} (7)

which, e.g., for $M = 16$ leads to $B = 6.9$ Gbit/s. The network size of $M = 16$ implies that the calculation holds for the 14 nodes in the radar system, including the antenna, plus two other nodes (e.g., for control purposes). With the calculated value of $B$ we will, in addition to the reserved bandwidth, have a high priority bandwidth of $B/M^2 = 27$ Mbit/s. This bandwidth is contained in the part of the cycle that is not allowed to be reserved and can be used for control and status messages. According to the above calculations, we have guaranteed bandwidths for both the dataflow and for control messages in the radar system.

With assumed values for parameters, it is possible to get the maximum number of nodes in a single-cluster network from Equation 5. Assuming a slot length of $\gamma = 1 \mu$s, a gap of one data slot between the last control slot
and the next cycle, i.e., \( \mu = \gamma \) and a worst-case latency of \( \tau_{\text{max}} = 1.0 \text{ ms} \), \( M_{\text{max}} = 30 \) is found to be the maximum number of nodes. The value of \( M_{\text{max}} \) can be increased by decreasing \( \gamma \), since the number of bits in one slot is rather high. At an effective bandwidth of 6.0 Gbit/s, there will be 6000 bits in each slot, which can be compared to ATM with its 424-bit cells. Clustering the system into a star-of-stars topology can also increase \( M_{\text{max}} \) [Jonsson et al. 1996].

7.7 Simulation Results

In addition to the worst-case analysis (deterministic performance), the average performance of the network was 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, i.e., \( \mu = \gamma \). To give an example of a real system, the slot duration was set to \( \gamma = 1 \mu s \). This corresponds to a 1 kbit packet if the data rate is 1 Gbit/s.
- All guarantee seeking messages have a deadline at 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. Messages were generated according to a Poisson process, and all messages were of single destination type.
- Message lengths were exponentially distributed between one and ten slots (discrete number of slots), with a length of one 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 the message generation rate through the mean message length.
- Latency is defined as the time elapsed from the moment of arrival of a message until the last packet of the message leaves the transmitter.

First, the deadline missing percentage (percentage of the messages that are not accepted for transmission) versus packet generation rate of guarantee seeking messages is plotted in Figure 35. At moderate traffic 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 packet generation rate of best effort messages, since guarantee seeking messages are always given priority.

In Figure 36, 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 37 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 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 becomes saturated, some of the total number of generated packets are discarded guarantee seeking messages. Below saturation, the best effort latency is almost uniform.

Figure 35: Real-time performance plotted as fraction of messages that miss their deadlines versus traffic intensity.
Figure 36: Latency for guarantee seeking messages plotted versus traffic intensity.

Figure 37: Mean latency for best-effort messages in a network with 10% guarantee seeking and 90% best-effort messages.
The mean best effort latency is plotted again in Figure 38, this time 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 37. This is a consequence of the almost linear relationship between packet generation rate and bandwidth utilization below saturation.

**Figure 38:** Mean latency for best-effort messages plotted versus bandwidth utilization. The traffic consisted of 10% guarantee seeking and 90% best-effort messages.

The mean best effort latency is plotted again in Figure 38, this time 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 37. This is a consequence of the almost linear relationship between packet generation rate and bandwidth utilization below saturation.

### 7.8 Summary

A medium access protocol for WDM star networks in distributed real-time computer systems has been proposed. To obtain scalability, the network architecture can be extended to a star-of-stars configuration. At the same time as dynamic real-time properties are supported, the protocol gives almost uniform latency nearly up to the theoretical saturation point. Deadline guarantees are supported, where the underlying deterministic bandwidth can be changed dynamically through slot reserving. An efficient bandwidth utilization is achieved by means of a simple slot release method.
8. WDM Star-of-Stars Network

By using electronic gateway nodes we retain the popular WDM star network architecture in each cluster, for which cheap components can be expected to appear in the future [Jonsson and Svensson 1997]. With electronic gateway nodes, we also achieve wavelength reuse in each cluster.

The same MAC protocol, TD-TWDMA (see the previous chapter) [Jonsson et al. 1996] is used in every cluster and in the backbone. Real-time services are implemented for inter-cluster communication in a similar way as for intra-cluster communication. However, the larger number of nodes sharing the backbone slots must be considered when calculating the deterministic performance. We also present a new method for clock synchronization to reduce the worst-case latency in this kind of multi-cluster network using TDM. The same notation as was used in the previous chapter (see Table 10 on page 111) is used when describing the multi-cluster extension of the network, but with the additional notation found in Table 11.

The remainder of the chapter is organized as follows. The network architecture and associated protocols are described in Section 8.1. Section 8.2 then presents the deterministic performance of the network, which is followed by a discussion of clock synchronization in Section 8.3. The chapter is then summarized in Section 8.4.

8.1 Network Architecture and Protocol

Although the same MAC protocol is used separately in each cluster, the clusters can be coordinated to improve performance and to get time-deterministic communication for inter-cluster communication as well.

\[ M_j: \] Number of nodes in cluster \( j \), including all transceiver modules on the cluster side in the gateway node

\[ Q_j: \] Number of ordinary nodes in cluster \( j \), i.e., number of physical end-nodes

\[ C_j: \] Number of channels (wavelengths) in cluster \( j \)

\[ L: \] Number of clusters, which is the same as the number of gateway nodes

\[ B: \] Effective bandwidth of each channel in the clusters

\[ E: \] Effective bandwidth of each channel in the backbone

\[ R: \] Ratio between the channel bandwidth in the backbone and the channel bandwidth in the clusters

*Table 11: Additional notation when describing the star-of-stars network architecture.*
A network consists of $L$ clusters. Each cluster has $M_i$, $1 \leq i \leq L$, nodes where one of the nodes is a gateway node. A gateway node contains network interfaces to both the backbone star and its dedicated cluster and 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 (flow control information) 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 are designed to each have $R$ transceiver modules on the cluster side in order to achieve the same aggregated bandwidth as on the backbone side (Figure 39). Also, a gateway node has $R$ dedicated home channels on the cluster side, one for each transmitter. Since the transceiver modules in a gateway node are seen as connections to separate nodes by the MAC protocol, the number of nodes is still defined as $M_i = C_i$, $1 \leq i \leq L$. In this way, we have $Q_i = M_i - R$ ordinary end-nodes in each cluster.

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 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 40). 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 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.

### 8.2 Deterministic Performance

Deterministic performance is important, e.g., for the ability to give guarantees for guarantee seeking messages. The worst-case latency for inter-cluster communication between two end-nodes is analyzed for two cases: (i) full slot reservation by other nodes and (ii) no slot reservation. When reservation is considered, full reservation is assumed in all clusters and in the backbone by other nodes than the analyzed transmitting node. In both cases, the network size, $M_{\text{total}}$, is assumed to be the total number of nodes in the network as seen by the MAC protocol. Also, it is assumed that each cluster has the same number of nodes as the number of clusters in the network:

$$M_{\text{total}} = \sum_{i=1}^{L} M_i = L^2$$

(8)

The gap between the last control slot and the next cycle is assumed to be one data slot, i.e., $\mu = \gamma$. We can then get the worst-case latency of a single cluster from Equation 5:

$$\tau_{\text{max}} = (M^2 + M + 1)\gamma$$

(9)
The guaranteed minimal bandwidth per source node, proportional to the number of high priority slots excluding reserved slots, is also analyzed. It is assumed in the analysis that no manipulation is made of the allocation scheme in order to utilize slots not having any function. These slots are those which are allocated for traffic between two transceiver modules in the same gateway node.

When full slot reservation by other nodes is assumed, the number of slots to which the source node has access in the receiver cycles in the destination node and in the gateway nodes of the source and destination clusters reaches its minimum. The first transit is from the source end-node and to the gateway node of the source cluster. Since the source node has direct access to its own cluster, the worst-case latency, \( \tau_1 \), for the first transit can be obtained by using Equation 9 for intra-cluster communication, with \( M = L \), i.e.:

$$\tau_1 = \gamma(L^2 + L + 1).$$ \hfill (10)

The second transit is between the gateway nodes of the source cluster and the destination cluster, through the backbone. Here, slots from all \( Q = L - R \) ordinary nodes in the source cluster must, in the worst-case, be multiplexed over several high priority backbone slots. At full slot reservation, a node has only one high priority slot per cycle in which to transmit. Therefore, \( L - R - 1 \) extra cycles, in addition to the normal intra-cluster latency (Equation 9), are needed. The source gateway node is responsible for carrying out this multiplexing using the round robin scheduling strategy. Hence, the source end-node is guaranteed its part of the bandwidth. The worst-case latency for the second transit is

$$\tau_2 = \gamma(L^2(L - R) + L + 1).$$ \hfill (11)

In the destination cluster, for transfer from the gateway-node to the end-node, slots from all \( (L - R)(L - 1) \) ordinary nodes outside the cluster must, in the worst-case, be multiplexed. The latency decreases when \( R > 1 \), because the multiple transceiver modules, in the gateway node, work in parallel:

$$\tau_3 = \gamma\left(\text{Trunc}\left(\frac{(L - R)(L - 1)}{R}\right) + 1\right)L^2 + L + 1)$$ \hfill (12)

The total worst-case latency, with full slot reservation, is

$$\tau_{max} = \tau_1 + \tau_2 + \tau_3 = \gamma\left(\left(\frac{(L - R)(L - 1)}{R} + 2 + L - R\right)L^2 + 3L + 3\right)$$ \hfill (13)
and is plotted in Figure 41, in which the horizontal axis represents the total number, $L(L - R)$, of ordinary nodes in the network and each curve represents a specific value of $R$. To give an example of a real system, the slot length is assumed to be $\gamma = 1.0 \, \mu s$ in all latency plots. Figure 41 indicates how latency decreases with $R$.

When no slot reservation in any receiver cycle is assumed, a node has $L - 1$ high priority slots in which to transmit. The worst-case latency, $\tau_1$, for the first transit is the same as when full reservation is considered (Equation 10). The latency of the rest of the transit will however decrease as compared to the full reservation latency. The second transit latency is $\tau_2 = \tau_1$, because the high priority slots in the transmitter of the gateway node, on the backbone side, will always be enough to multiplex one slot from each cluster node in one cycle:

$$L - 1 \geq L - R$$ (14)
In the third transit, to the destination end-node, the gateway node can multiplex $\frac{R(L - 1)}{1}$ slots per cycle time. The latency is

$$\tau_3 = \gamma\left(\frac{L}{R}\right)\left[\frac{L^2 + L + 1}{c}\right]$$

(15)

The total worst-case latency, with no slot reservation, is

$$\tau_{max} = \tau_1 + \tau_2 + \tau_3 = \gamma\left(\frac{L}{R} + 2\right)\left[\frac{L^2 + 3L + 3}{c}\right]$$

(16)

Figure 42 shows how the worst-case latency when no slot reservation is used varies with the total number of ordinary nodes. The latency is significantly lower as compared to the case of full-reservation. This effect is related to the higher number of slots in the gateway nodes that can be used for multiplexing incoming messages.

Figure 42: Worst-case latency when no slot reservation is assumed. The x-axis represents the number of ordinary nodes and the slot length is assumed to be $\gamma = 1.0$ $\mu$s.
We define the deterministic bandwidth per node as the minimum high priority bandwidth when there are no reserved slots. In this analysis we express the bandwidth as a ratio to the full channel bandwidth, $B$. Measured as the number of high priority slots per total number of slots in a cycle, this so called bandwidth is shown in Table 12 for the different cases. The last transit is the bottleneck; the node bandwidth for it, as a function of the number of ordinary nodes, is plotted in Figure 43 for full slot reservation. As seen in the figure, the deterministic node bandwidth is a rather low part of the full bandwidth but all of the bandwidth can be used for broadcasting. Nodes requiring more deterministic bandwidth can use slot reservation. The deterministic node-bandwidth, when no slot reservation is assumed, is plotted in Figure 44 and is higher than the former.

The real-time services offered by the MAC layer rely on the deterministic latency and bandwidth. A guarantee can be stated if the known minimum number of high priority slots along the whole path through the network is sufficient to transfer the message in time. In the calculation of this justification, the deterministic latency, the deterministic bandwidth, and the deadline of the message are used. A slot of a guaranteed message is tagged to indicate its priority over other slots. The gateway nodes then always transmit the tagged messages before buffered best effort messages.

### Table 12: Node bandwidth in number of high priority slots per total number of slots in a cycle.

<table>
<thead>
<tr>
<th></th>
<th>Full reservation</th>
<th>No reservation</th>
</tr>
</thead>
<tbody>
<tr>
<td>From source end-node to gateway node:</td>
<td>$\frac{1}{L^2}$</td>
<td>$\frac{L - 1}{L^2}$</td>
</tr>
<tr>
<td>Through backbone:</td>
<td>$\frac{R}{L^2(L - R)}$</td>
<td>$\frac{R(L - 1)}{L^2(L - R)}$</td>
</tr>
<tr>
<td>From gateway node to destination end-node:</td>
<td>$\frac{R}{L^2(L - R)(L - 1)}$</td>
<td>$\frac{R}{L^2(L - R)}$</td>
</tr>
</tbody>
</table>

8.3 Clock Synchronization Aspects

The nodes are synchronized to account for the propagation delay between transmitting and receiving nodes, i.e., receivers are synchronized to
transmitters, proportional to the propagation delays, and thus slots are first expected when they have traveled through the network [Jonsson et al. 1995] [Bengtsson et al. 1994]. This type of synchronization is always used inside a cluster.

The inter-cluster worst-case latency can be reduced by using the synchronization scheme shown in Figure 45. For clarity, in the example in the figure, all fibers are assumed to have the same length and propagation delay and to have a propagation delay significantly lower than the cycle time of $S$ slots. These restrictions do not directly apply to a real implementation.

The midpoint of the backbone star is used as the reference point. Each cluster is then synchronized to the backbone by its gateway node. In a gateway node, the receiver on the cluster side is synchronized so tightly to the transmitter on the backbone side that incoming messages are directly forwarded when possible. In this way, the information extracted from an incoming control slot (about next hop, i.e., the gateway node of the

Figure 43: Node bandwidth, when full slot reservation by other nodes is assumed, in number of high priority slots per total number of slots in a cycle. The x-axis represents the number of ordinary nodes.
destination cluster) can be transmitted to the other gateway nodes, like a pipeline mechanism, before the data slots have arrived. This will reduce the worst-case latency by $\gamma \frac{L^2 + L + 1}{c^2 + \frac{1}{c}}$, i.e., eliminate the latency for one hop as described by Equation 9. On the other hand, the receiver on the backbone side is not synchronized to the transmitter on the cluster side. This is because transmitters in a cluster are synchronized to the receivers in the same cluster. Therefore, the same latency is experienced as for the case in which the clusters and the backbone are not synchronized (described above).

The improved worst-case latency (Equation 13) when full slot reservation is assumed is

$$\tau_{max} = \gamma \left( \frac{(L-R)(L-1)}{R} + 1 + L - R \right) \frac{L^2 + 2L + 2}{c^2}$$  \hspace{1cm} (17)$$

The corresponding improved latency when no slot reservation is assumed (Equation 16) is
The relative latency improvement is best in the case of no slot reservation for networks with larger values of $R$ (Figure 46). The figure compares the latencies with and without synchronization between the clusters and the backbone. Note that the latter latency is the same as that described by Equation 16 above. Even better relative improvement is achieved when a slot can pass through the whole network without multiplexing over several cycles in the gateway nodes, e.g., by the use of reserved slots or at low traffic. In this case, the worst-case latency decreases from $3\gamma (L^2 + L + 1)$ to

$$\tau_{max} = \gamma \left( \frac{L}{R} + 1 \right) \left[ L^2 + 2L + 2 \right]$$ (18)
We have shown how to calculate the worst-case latency for inter-cluster communication in a WDM star network using the TD-TWDMA protocol. The analysis shows how the latency decreases for larger networks when the ratio between the backbone bandwidth and the cluster bandwidth increases. A calculation of the minimum deterministic bandwidth obtained in the case in which no reservation is used by the analyzed node was presented, and a synchronization scheme was proposed. At reserved traffic or low traffic, the improvement to the worst-case latency when using this synchronization scheme is 33 percent, as compared to the case in which the clusters are not synchronized with one another.

\[ 2\gamma (L^2 + L + 1), \] i.e., an improvement of 33 percent. The synchronization scheme is general and can be used in other similar networks to improve worst-case performance.

### 8.4 Summary

We have shown how to calculate the worst-case latency for inter-cluster communication in a WDM star network using the TD-TWDMA protocol. The analysis shows how the latency decreases for larger networks when the ratio between the backbone bandwidth and the cluster bandwidth increases. A calculation of the minimum deterministic bandwidth obtained in the case in which no reservation is used by the analyzed node was presented, and a synchronization scheme was proposed. At reserved traffic or low traffic, the improvement to the worst-case latency when using this synchronization scheme is 33 percent, as compared to the case in which the clusters are not synchronized with one another.
9. Control Channel Based Fiber-Ribbon Pipeline Ring Network

This chapter and the next chapter present two ring networks suitable for different situations, both based on fiber-ribbon links. The proposed networks are pipeline ring networks based on optical fiber-ribbon point-to-point links. In a pipeline ring network, several packets can be traveling through the network simultaneously, thus achieving an aggregated throughput higher than the capacity of a single link. Motorola OPTOBUS™ bi-directional links [Schwartz et al. 1996] with ten fibers per direction are used (similar links can be used, too) but the links are arranged in a unidirectional ring architecture (see Figure 47) where only \( \lceil M/2 \rceil \) bi-directional links are needed to close a ring of \( M \) nodes. Fiber-ribbon links offering an aggregated bandwidth of several Gbit/s have already reached the market [Bursky 1994]. The increasingly good price/performance ratio for fiber-ribbon links indicates a great potential for success of the proposed kind of networks.

The remainder of the chapter is organized as follows. An overview of the CC-FPR network is presented in Section 9.1, while a review of related work

\[ \text{Figure 47: (a) Bi-directional fiber-ribbon link. (b) Unidirectional ring network built up of } M/2 \text{ bi-directional links.} \]
is given in Section 9.2. The CC-FPR protocol is presented in Section 9.3. Section 9.4 describes different user services. Section 9.5 discusses implementation aspects, and a case study is presented in Section 9.6 to show the efficiency of the networks. This is followed by a performance analysis and a summary in Sections 9.7 and 9.8, respectively.

9.1 Network Overview

The first network (described in this chapter) is called CC-FPR [Jonsson 1998] [Jonsson 1998B] and has special features for both parallel processing in general and for distributed real-time systems [Jonsson et al. 1999] [Jonsson et al. 1999B]. The physical ring network is divided into two rings: a data ring and a control ring. In each fiber-ribbon link, eight fibers carry data and one fiber is used to clock the data, byte by byte. Together, these fibers form a data channel that carries data packets. The access is divided into slots as in an ordinary TDMA network. The tenth fiber is dedicated to bit-serial transmission of control packets that are used for the arbitration of data transmission in each slot. The clock signal on the dedicated clock fiber, which is used to clock data, also clocks each bit in the control packets. Separating clock and control fibers simplifies the transceiver hardware implementation, which is verified by the current prototype development. The control channel is also used for the implementation of low level support for barrier synchronization, global reduction, and reliable transmission.

Real-time services in the form of best effort messages, guarantee seeking messages and RTVCs [Arvind et al. 1991] are supported for single destination, multicast and broadcast transmission by the network. Application examples requiring a high performance network of this kind are radar signal processing systems [Jonsson et al. 1996] [Taveniku et al. 1998] and multimedia systems. An aggregated throughput of tens of Gbit/s is needed in soon-to-come radar systems, and this is achieved with the proposed network as shown in a case study (see Section 9.6) [Jonsson 1998] [Jonsson 1998B].

The node synchronization requirement is more relaxed than for a traditional TDMA network, and the network is somewhat similar to a slotted ring network (but without the need of a central controller). This is because the access to the network circulates among the nodes according to the physical order of the nodes in the ring. In addition, the ring can dynamically (for each slot) be partitioned into segments to obtain a pipeline ring network where several transmissions can take place simultaneously. Even simultaneous multicast transmissions are possible when the multicast segments do not overlap.
9.2 Related Networks

Two kinds of spatial reuse in ring networks can be identified. The intention of the first type is to overcome performance degradation owing to long propagation delays around the ring as compared to the frame length. Performance is increased if another node (or nodes) can start transmission before the current transmission is finished. One example is early token release in the 16 Mbit/s token ring network, where the token is released when the last bit of the frame has left the sending node. When early token release is not used, the sending node does not release the token until the frame has traveled around the ring and is completely removed from the ring. Another example is slotted ring (e.g., the Cambridge network [Greaves and Zielinski 1993] [Hopper and Needham 1988]) where several time slots, in which each can host a frame, travel around the ring at the same time.

The first type of spatial reuse increases the utilization, but it cannot exceed 1 as long as data frames are not removed for reuse of the bandwidth for new frames. This states the definition of the second type of spatial reuse. For example, suppose the frame contained in a slot in a slotted ring is removed by the destination node [Adams 1994]. The destination node, or another node downstream, can then reuse the slot. If the distribution of source and destination nodes is totally uniform, the average utilization can theoretically reach 2. This means that the average source to destination distance is half the ring, which gives an average of two transmitted frames per time unit. The term "pipeline ring network" is used in this thesis to denote a network with spatial reuse of the second type. It should be noted that spatial reuse of the bandwidth can be implemented in dual unidirectional bus networks as well [Garrett and Li 1991] [Ray and Jiang 1995] [Ray and Mukherjee 1995].

Other pipeline ring networks are described in [Chen et al. 1991] [Ofek 1994] (MetaRing), [Imai et al. 1994] (ATMR), [Wong and Yum 1994], [Jafari et al. 1980], and [Xu and Herzog 1988] and further references are given in [Wong and Yum 1994]. Advantages of the CC-FPR network over these other networks include the use of high bandwidth fiber-ribbon links and the close relation between a dedicated control channel and a data channel without disturbing the flow of data packets. In other words, control and data are overlapped in time. With less header overhead in the data packets the slot length can be shortened to reduce latency without too great a sacrifice of bandwidth utilization. The separate clock and control fibers also simplify the transceiver hardware implementation. Another fiber-ribbon ring network is the USC PONI (formerly called POLO) network, proposed to be used in multimedia applications [Sano and Levi 1998] [Raghavan et al. 1999]. However, spatial reuse of bandwidth is not a function of the USC PONI, i.e., it is not a pipeline ring network.
The network described by Jafari et al. also relies on a separate control channel but needs a central control node that brings additional cost in hardware and in latency when waiting for response from the central control node [Jafari et al. 1980]. The CC-FPR network is insensitive to propagation delay in the sense that no feedback is needed, from other nodes or from a central controller, between control packet and data packet transmissions.

A network with a similar slot reuse mechanism for reserved slot as for the CC-FPR network (see Section 9.4.1) has been reported [Marsan et al. 1997]. However, the network does not support concurrent transmissions in different segments of a single ring channel.

Other high performance ring networks include the WDM passive ring [Irshid and Kavehrad 1992] and the hierarchical WDM ring [Louri and Gupta 1997], which are more closely related to the WDM star network and star-of-stars network described in Chapters 7 and 8, respectively.

9.3 The CC-FPR Protocol

Throughout Section 9.3, slot reserving, as described in Section 9.4.1, is assumed not to be used. Before we explain the arbitration mechanism, we will describe how data packets travel on the ring.

The access to the network is cyclic; each cycle consists of \(M\) time slots, where \(M\) is the number of nodes. Each node is denoted \(m_i\), \(1 \leq i \leq M\). Each slot always has one node that is responsible for initiating the traffic around the ring. This node is called the slot initiator (SI). Each node is slot initiator in one slot per cycle, as shown in Figure 48. At the end of the slot, the role of slot initiator is asynchronously handed over to the next node downstream. This can be done implicitly simply by sensing the end of the slot, i.e., the last bit.
The CC-FPR medium access protocol is based on the use of a control packet that, for each slot, travels almost one round (over $M - 1$ links) on the control channel ring, as shown in Figure 49. The node that will be the slot initiator in the next slot initiates the transmission of the control packet, as shown in the figure. We denote this node $S_{I+1}$. In the time domain, the control packet always travels around the ring in the time slot preceding the one for which it controls the arbitration (see Figure 50). Accordingly, the control packet always passes each node one time slot before the data packet to which it is related.

The contents of the control packet are shown in Figure 51. The control packet consists of a start bit followed by an $M$-bit long link reservation field and an $M$ bit long destination field, where $M$ is the number of nodes. Each bit in the link-reservation field states whether the corresponding link is reserved for transmission in the next slot. In the same way, each bit in the destination field states whether the corresponding node has a data packet destined to it in the next slot. Additional information, such as flow control, is also included in the control packet; for clarity, this is not shown in the figure.

Each node succeeding $S_{I+1}$ checks the control packet as it passes to determine: (i) whether it will receive a data packet in the next slot, which is indicated by the node's bit in the destination field, and (ii) if a data packet will pass the node in the next slot, which is indicated by the bit in the link-
reservation field corresponding to the outgoing link of the node. If no data packet is to pass the node, i.e., the rest of the ring back to the slot initiator is free, then the node can transmit a data packet in the next time slot in this part of the ring.

When a node has a packet ready for transmission, it prepares in advance new link reservation and destination fields to reserve needed links and notify destination node(s). In this way, the node can immediately change the control packet when it passes, provided the bit in the link-reservation field corresponding to the outgoing link of the node is set to zero (see Figure 52 for a conceptual view of the control channel part of the transceiver). Since there is no data packet that will pass the node, succeeding nodes have no use for the overwritten information in the control packet.

Because all of the nodes succeeding the slot initiator repeat the procedure of checking the control packet, multiple transmissions in different segments of the ring can occur in the same slot. The transceivers are designed to allow for both reception and transmission at the same time (see Figure 53), which increases the possibility of spatial bandwidth reuse. An example of how the control packets travel around a five-node network is shown in Figure 54.

Figure 50: In each slot, a node passes/transmits one control packet and one data packet, where the control packet is used for the arbitration of the next slot.

<table>
<thead>
<tr>
<th>Link-reservation field</th>
<th>Destination field</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bit 0</td>
<td>Bit 1</td>
</tr>
<tr>
<td>Start bit</td>
<td>Link 1</td>
</tr>
</tbody>
</table>

Figure 51: A control packet contains a start bit, a link reservation field, and a destination field.
The arbitration results in two concurrent data packet transmissions in the next slot, one single-destination and one multicast packet, as shown in Figure 55. Node \( m_1 \) is SI+1 in the example; it therefore initiates the control packet transmission described in Figure 54. It reserves Link 1 and Link 2 for transmission to node \( m_2 \) and informs this node, by setting the corresponding bits in the link reservation field and the destination field, respectively, that it will have a data packet destined to it in the next slot. While node \( m_2 \) and node \( m_3 \) do not change the control packet, they check it to see whether there will be any data packets destined to them in the next slot. Node \( m_4 \) reserves Link 4 and Link 5 for a multicast transmission to node \( m_5 \) and node \( m_1 \). Node \( m_5 \) then receives the control packet and removes it from the ring.

The reason why the control packet travels only among the first \( M - 1 \) links after SI+1 is that the clock signal is interrupted by the SI (see Figure 49). The node that initiated the transmission of the control packet, SI+1, does not return the packet. Consequently, it will not be informed of whether or not there is a data packet destined to it in the next slot. However, the node

![Figure 52: Conceptual view of the control channel part of the transceiver.](image)

![Figure 53: Conceptual view of the data channel part of the transceiver.](image)
will receive either a packet destined to the node or an empty packet.

It is essential for desirable performance that the delay of the control packet in each node it bypasses be minimal, especially in large networks. One method is to organize the bits in the link-reservation field in the control packet for each slot, so that they appear in the same order as that in which the control packet travels. In other words, the first bit corresponds to the outgoing link from the slot initiator. Thus, when a node wishes to change the contents of a control packet, it does not have to store the whole packet before checking and possibly overwriting it. Instead, it can retransmit the packet bit by bit and exchange the remaining part of the packet (if transmission is possible) after reading the bit in the link reservation field corresponding to its outgoing link. The node's bit in the destination field in the incoming control packet must, however, be checked before it is thrown away. Using this method, the delay in each node can be reduced to only one or a few bits.

As indicated in Figure 56, the bandwidth utilization depends on the ratio of the total propagation delay around the ring to the cycle length. This is an effect related to the asynchronous passing mechanism of the slot initiator assignment. Further evaluation of this phenomena is provided in Section 9.7.

9.4 User Services

The user services described below are: real-time virtual channels (Section 9.4.1), guarantee seeking messages (Section 9.4.2), best effort messages (Section 9.4.3), barrier synchronization (Section 9.4.4), global reduction (Section 9.4.5), and reliable transmission (Section 9.4.6).
Many computer systems have real-time demands where the network must offer logical connections with guaranteed bandwidth and bounded latency. This can be done in the network by using slot reserving. We refer to such connections as RTVCs. Either the whole ring is reserved for a specific node in a slot, or several segments of the ring are dedicated to some specific nodes.

When slot reservation is allowed, the cycle is prolonged to contain \( M(N_{ord} + N_{res}) \) slots, where \( N_{ord} \) and \( N_{res} \) are the number of slots for ordinary use and for reservation (per node and cycle), respectively. The values of \( N_{ord} \) and \( N_{res} \) are chosen at system design and remain unchanged during operation of the network if the system function does not change radically. The \( N_{ord} \) ordinary slots cannot be reserved. However, all nodes in the network can try to reserve a segment of the ring in a reservation slot, not only the SI of that slot. In each node, a separate packet queue is provided for each of its RTVCs.

When a node wishes to reserve a slot for an RTVC, it searches for slots where the required links are free to allow allocation of a new segment. First, the node's own slots (i.e., where the node itself is the slot initiator) are searched. If enough slots (actually only a segment in each slot) cannot be allocated for the reservation, the search is continued in other slots. In this

Figure 55: An example: Node 1 sends a single destination packet to Node 3, while Node 4 sends a multicast packet to Node 5 and Node 1.

9.4.1 Real-Time Virtual Channels

Many computer systems have real-time demands where the network must offer logical connections with guaranteed bandwidth and bounded latency. This can be done in the network by using slot reserving. We refer to such connections as RTVCs. Either the whole ring is reserved for a specific node in a slot, or several segments of the ring are dedicated to some specific nodes.

When slot reservation is allowed, the cycle is prolonged to contain \( M(N_{ord} + N_{res}) \) slots, where \( N_{ord} \) and \( N_{res} \) are the number of slots for ordinary use and for reservation (per node and cycle), respectively. The values of \( N_{ord} \) and \( N_{res} \) are chosen at system design and remain unchanged during operation of the network if the system function does not change radically. The \( N_{ord} \) ordinary slots cannot be reserved. However, all nodes in the network can try to reserve a segment of the ring in a reservation slot, not only the SI of that slot. In each node, a separate packet queue is provided for each of its RTVCs.

When a node wishes to reserve a slot for an RTVC, it searches for slots where the required links are free to allow allocation of a new segment. First, the node's own slots (i.e., where the node itself is the slot initiator) are searched. If enough slots (actually only a segment in each slot) cannot be allocated for the reservation, the search is continued in other slots. In this
case, the node broadcasts a packet containing a request to all other nodes to allocate the desired segment in their slots. The packet contains information about the links required and the number of slots needed. Each node then checks whether any of its own slots have the required free links. All nodes send a packet back to the requesting node to notify which slots, if any, that have been allocated. When the requesting node has received the answers, it decides whether it is satisfied with the number of allocated slots. If not, it sends a release packet. Otherwise, it can start using the reserved slots immediately. However, it should still send a release packet if more slots than needed were allocated. The same medium access method as for ordinary slots is used for reservation slots, but with the restriction not to pass links in reserved segments of the ring. A reserved segment of the ring temporarily not in use, however, can be reused by nodes down stream of the owner.

9.4.2 Guarantee Seeking Messages

Guarantee seeking messages normally have hard 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 the CC-FPR network, a guarantee is given only if enough deterministic bandwidth (slots) owned by the node is free before the deadline of the message. The available deterministic bandwidth corresponds to ordinary slots where the node is the slot initiator and which

---

Figure 5.6: The bandwidth utilization depends on the ratio of the total propagation delay around the ring to the cycle length. The boxes with bold text show the link through which each slot first propagates.
have not already been encountered for other guarantee seeking messages queued in the node.

Each transmitter has \( M - 1 \) queues for guarantee seeking messages, one for each possible destination (the node itself excluded). When a multicast packet arrives for queuing, it is put in the queue corresponding to the multicast destination furthest away from the source node downstream. In this way, multicast packets are treated in the same way as single-destination packets, and multiple multicast packets can travel in the network at the same time whenever possible.

9.4.3 Best Effort Messages
Best effort messages can be sent in all ordinary slots where the node is SI but in which the node does not have any guarantee seeking messages that can be sent. In competition with the other nodes according to the CC-FPR protocol, other ordinary slots can also be used but, again, only as long as no guarantee seeking messages can be sent. The same method is used for reservation slots that are not (or are only partly) reserved for the moment or that are reserved but not used in the current slot.

As for guarantee seeking messages, each transmitter has \( M - 1 \) queues. The messages in each of these queues are sorted according to the EDF-algorithm. If the deadline expires, the sending process is notified. Other similar algorithms can also be used.

9.4.4 Barrier Synchronization
Barrier Synchronization (BS) is an operation to control the flow of processes in a distributed processing system. A logical point in the control flow of an algorithm is defined, at which all processes in a processing group must arrive before any of the processes in the group are allowed to proceed further. When, during execution, a BS point is encountered in the application program, the node broadcasts the encountered BS_id in the control packet when the node is SI +1. In this way, all nodes are notified that the node has reached the BS point. Nodes not belonging to the BS group can ignore the broadcast, but nodes belonging to the same group, i.e., have the same id, will make a note in an internal table. The control packet contains a field in which BS_id can be sent (see Figure 57). The id field contains eight bits, which permits ids ranging from one to 255. When the field is zero, no BS command is sent.

When a node participating in the BS group has received the correct BS_id from all the participants, it knows that all the other nodes are at the same executing point and may proceed. The worst case latency for a node that reaches the BS point until it can broadcast this to the other nodes is one
This assumes one slot per node and that each node is SI+1 only once per cycle. Clearly, the implications of sending BS information in the control channel are both bounded latency and better bandwidth utilization in the data channel. The whole BS mechanism is handled by the communication interface, transparent to the calling user processes.

In the description above, static allocation of barrier synchronization BS_ids is assumed. The programmer (or the compiler) allocates the required parameters for BS and GR (Global Reduction) off-line before run-time. With minor adjustments, dynamic allocation is also possible, although this is not investigated further in this paper.

9.4.5 Global Reduction

Global reduction is similar to barrier synchronization, where data are collected from distributed processes when they signal their arrival at the synchronization point. A global operation, e.g., sum or product, is performed on the collected data so that only a single value is returned. At the end of the GR, all participating nodes have access to the same data. As in the case of BS, we assume that the programmer (or the compiler) statically allocates the necessary parameters, off-line before run-time.

The GR command requires the following parameters: operation, length, data type, GR_id, and the ids of the participating nodes. The last two parameters are similar to the parameters in the BS command. The operation parameter tells the nodes what operation (sum, product, max, min, etc.) should be performed on the data received. The data type parameter indicates how the data field in the control packet (see Figure 57) is to be interpreted, and the length parameter gives the length of the data field. In the studied case, the data field is 320 bits long and may facilitate the transfer of, e.g., up to five double precision floating point numbers (IEEE-754). Other data types may also be distributed. Except for the additional fields and the global function, the nodes treat GR commands in the same way as BS commands. The same reasoning of performance advantages also holds for GR.
The type bit states whether the control packet contains a BS or a GR command and hence whether data are contained in the data field (see Figure 57). The data field is currently only used for data reduction.

9.4.6 Low Level Support for Reliable Transmission

The proposed network has low level support for reliable transmission [Bergenhem and Olsson 1999]. Network control information, acknowledge/negative acknowledge and flow control, is sent in the control channel instead of in ordinary data packets. This results in less or no overhead in the data channel, i.e., better bandwidth utilization. The field in the control packet named ACK/NACK contains bits that correspond to the $M$ packets that may have been received by the current SI+1 during the previous $M$ slots. The ACK/NACK information is therefore always sent when a node is SI+1. If a packet was correctly received (correct checksum), a “1” is written in the position of the ACK/NACK field that corresponds to the slot that the packet was received. If a faulty packet or no packet was received, a “0” is written. All nodes must keep track of their transmissions and can therefore resolve the meaning of the bits in the ACK/NACK field. In this way, the nodes can be notified as to whether their packet was correctly received or must be retransmitted. The latency for a node to send and receive ACK/NACK is bounded, which is desirable.

The $4(M - 1)$ bits in the flow control field relate to independent logical connections. Put simply, each node can have up to four logical connections, with low-level support for flow control, from each other node. The SI+1 sets the bit corresponding to a logical connection to “1” if it is to be halted temporarily; otherwise, the bit is set to “0”.

9.5 Implementation Aspects

In addition to the high bandwidth offered by a fiber-ribbon cable, it also offers a ten-fold increase in packing density as compared to electrical cables, resulting in less rigid cables [Karstensen et al. 1995]. Furthermore, it is not necessary for the designer to be concerned about electromagnetic emissions. These properties make possible new components such as the single-chip optoelectronic switch core reported in [Szymanski et al. 1998]. In addition to the switch function, the chip eliminates 32 OPTOBUS 800 Mbit/s per fiber transceivers. This translates to an aggregated bandwidth of 204 Gbit/s through the switch when eight of the ten fibers on each link are used for data. Such a switch can connect multiple ring clusters when building large networks. The high bit-rate of a fiber-ribbon link makes it possible to reduce the slot duration in the proposed networks, still keeping the same number of bits in a packet as on a slower link based on an electrical cable. This reduces the latency.
In scaling up the bandwidth of a fiber-ribbon link where a dedicated fiber carries the clock signal, the main problem is channel-to-channel skew. The skew is mainly the result of differences in propagation delay between different fibers and variations of lasing delay time among different laser diodes [Kurokawa et al. 1998]. The 400 Mbit/s OPTOBUS has a specified maximum skew of 200 ps, excluding the fiber-ribbon cable for which 6 ps/m is assumed for standard ribbons [Schwartz et al. 1996]. Since the data stream passing a node in the ring network is, at least, passing a pipeline register, the channel-to-channel skew is not accumulated over several nodes. The limited distance between two adjacent nodes, owing to skew, is in the same order of magnitude as in today’s LAN networks (a few hundred meters). In parallel and distributed computer systems, so called System Area Networks (SAN), the maximum required distance is normally lower than this limit. It may however be difficult to increase the bit rate to several Gbit/s per fiber without significantly reducing the distance. It can also be argued that it should be possible to construct networks with more physically distributed nodes, since this may be valuable in some applications. The latency of the network is not dependant on the distance, except that the propagation delay is added and the throughput is reduced as mentioned above. This motivation calls for the discussion of techniques to reduce the effect of the skew below.

One technique is to actually reduce the skew, either by using low skew ribbons or employing skew compensation. Fiber-ribbons with about 1 ps/m skew [Siala et al. 1994] and below [Kanjamala and Levi 1995] have been developed, which essentially increases the possible bandwidth distance product. All the fibers in the same ribbon were sequentially cut to reduce the variation of refractive index among the fibers. In the fiber-ribbon link described in [Wong et al. 1995], a dedicated fiber carries a clock signal used to clock data on 31 fibers. The transmitter circuitry for each channel has a programmable clock skew adjustment to adjust the clock in 80-ps increments.

Another technique is to extract the clock signal from the bit flow on each fiber instead of using a separate fiber to carry the clock signal. The disadvantage is increased hardware complexity when adding a clock recovery circuit and a buffer circuit for each channel in the receiver. A hybrid solution is to skip the separate clock channel and encode clock information on the data channels while still sending in bit-parallel mode, as reported in [Yoshikawa et al. 1997] [Yoshikawa et al. 1997B]. In this case, a deskew unit relying on FIFO registers (First In First Out) ensures that parallel data words that are output from the receiver are identical to those which were sent. A possible ± 15-ns deskew was reported. A similar system is reported in [Fujimoto et al. 1998].
The techniques mentioned above introduce either increased hardware complexity or a more sophisticated fiber-ribbon manufacturing process. If the manufacturing process allows for adding more fibers in each ribbon, this may be a cheaper alternative. For example, 12 channel links with 1 Gbit/s per channel [Karstensen et al. 1995] and 2 Gbit/s per channel [Karstensen 1995] have been reported, and array modules supporting 12 × 2.4 Gbit/s for, e.g., fiber-ribbon links were described in [Peall 1995]. A fiber-ribbon link with 32 fibers, each with a bit rate of 500 Mbit/s, was described in [Wong et al. 1995], and researchers at NEC have developed a module in which 8 × 2 lasers are coupled to two fiber-ribbons [Kasahara 1998]. Instead of fiber-ribbons, fiber imaging guides (FIGs) with thousands of pixels can be used [Li et al. 1995]. In the system described in [Li et al. 1998B], both a 14000-pixel FIG and a 3500-pixel FIG were coupled to an 8 × 8 VCSEL array in different setups.

### 9.6 Case Study

A typical application with high throughput requirements and a pipelined data flow between the computational modules is future radar signal processing systems. In Figure 58, a signal processing chain, similar to those described in [Jonsson et al. 1996] [Taveniku et al. 1996] is shown together with its bandwidth demands. Each computational module in the figure contains multiple processors. The chain is a good example containing both multicast, one-to-many, and many-to-one communication patterns. The aggregated throughput demand is 30 Gbit/s. Only the throughput requirements are treated here; all details on the chain are covered in the two papers referred to above. The data flow must not be disturbed by, for example, status information that the network must also transport. Slot reserving is therefore a good choice for the data flow in the signal processing chain when using the CC-FPR network.

We assume links with ten fibers and 800 Mbit/s per fiber in the case study. In the CC-FPR network, this translates to a bandwidth of 6.4 Gbit/s for data traffic on eight of the fibers. For simplicity, we assume an efficient bandwidth of 6.0 Gbit/s after, for example, check-sums have been excluded. Figure 58 shows 13 nodes. In addition, the antenna is seen as one node (feeds the first node in the chain with data) and one master node is responsible for supervising the whole system and interacting with the user. We denote the antenna as node $P_{1}$, the modules shown in the figure as node $m_{i}$, 2 ≤ i ≤ 14, and the master node as $m_{15}$. The numbers of the modules are also indicated in the figure. Hence, the number of ordinary slots is 15, but the cycle is extended so that it also contains 30 slots for reservation. Accordingly, there are 45 slots in a cycle, where one slot per cycle corresponds to a bandwidth of 133 Mbit/s at a total efficient bandwidth of 6.0 Gbit/s.
A feasible allocation scheme of the slots is shown in Figure 59. For clarity, all of the reservation slots are placed after the ordinary slots. In a real implementation, however, there are two ways of spreading the reservation slots, as explained below in Section 9.7. Care must be taken, however, in distributing the reservation slots, the reason being that, when there are intermediate nodes between the source and destination nodes, allocation is not possible in those slots in which one of the intermediate nodes is the slot initiator.

The maximum data flow from one module is 4.0 Gbit/s, which corresponds to reserving having a segment of the ring in all of the 30 slots (for reservation). Slots for both of the two 2 Gbit/s data flows to the pulse compression nodes can be allocated, since one of the two data flows is tapped before adding the data flow produced from the same node. The incoming data flow to the CFAR (Constant False Alarm Ratio) nodes is multicasted to all of these.
nodes. Although this multicast data flow must remain unchanged until the last CFAR node, it can coexist with the data flow produced from the CFAR nodes. The reason for this is that the multicast bandwidth is only 2 Gbit/s. The rest of the data flows are pure pipeline flows and fit easily on the network as long as the calculations are mapped on the nodes according to the pipeline order.

### 9.7 Performance Analysis

To be able to give guaranteed real-time services, worst-case performance must be determined. An analysis of how worst-case latency and deterministic throughput vary with the network design parameters is given below. Each node is assumed to have $N_{ord}$ ordinary slots and $N_{res}$ slots for which reservation is allowed, per cycle. Two ways of circulating the role of being slot initiator are treated and compared to each other (see Figure 60): Case A, where each node is slot initiator in $N_{ord} + N_{res}$ slots in sequence, and Case B, where the slots are interleaved in a way that each cycle is in practice divided into $N_{ord} + N_{res}$ sub-cycles, where each node is slot initiator once per sub-cycle. Throughout the section, the total propagation time
around the ring is denoted as $T_{prop}$, the skew between a control packet and the data packet it arbitrates for as $T_{skew}$, and $M$ is the number of nodes in the network. The value of $T_{skew}$ is assumed to be equal to the duration of one slot, $T_{slot}$, which is set to 1 µs in this analysis.

The worst-case latency a node experiences for guarantee seeking traffic is in the case in which all reservation-slots are reserved for other traffic and the only ordinary slots it gains access to are those in which it is the slot initiator. If a packet is generated just after the node had the chance to send, the node must wait until the next ordinary slot in which it is slot initiator. The latency for Case A will then be:

$$T_{lat} = T_{slot} (M(N_{ord} + N_{res}) - N_{ord} + 1) + T_{prop} + T_{skew}. \quad (19)$$

The equation holds for best effort traffic as well if no other (best effort or guarantee seeking) traffic is queued in the node to be sent before. The latency is plotted in Figure 61 against the number of nodes for different combinations of $N_{ord}$ and $N_{res}$, and for different values of $T_{prop}$. The same worst-case latency, but here for Case B, is:

$$T_{lat} = (T_{prop} + T_{slot} M)(1 + N_{res}) + T_{skew}. \quad (20)$$

and is plotted in Figure 62. As seen in the figures, Case B can be more sensitive to large propagation delays. With the same assumptions as above, further comparison between Cases A and B is shown in Figure 63. Case B
can offer lower latency in some situations in which $N_{ord} > 1$, especially for large values of $M$.

Even if worst-case parameters for guaranteed services must be determined, the worst-case latency in the case of no other traffic in the network is an essential parameter to form an idea about the more general performance. However, it is still assumed that all reservation slots are reserved for other traffic. For communication to the nearest neighbor down stream, it is possible to send in any ordinary slots. If the packet must instead travel over several links, the slots for which any of the intermediate nodes is slot initiator cannot be used because the slot initiator always terminates both clock and data. The worst-case latency for Case A is:

$$T_{hot} = T_{slot} (Q(N_{ord} + N_{res}) - N_{ord} + 1) + T_{prop} + T_{slay}$$

(21)

where $Q$ is the number of links a packet must pass to come to the destination node. The gap between each cycle (see Figure 56) is always
experienced by a node when it is handing over the role of being slot initiator to the next node. Because the worst-case latency appears when a node just misses a slot where it is the slot initiator (the last one if there are several in a sequence), the whole $T_{\text{prop}}$ is always a part of the latency, even if $Q < M$. Because of the similarity between Equations 19 and 21, a plot of Equation 21 would look the same as the plot of Equation 19 in Figure 61, except that $Q$ is represented on the x-axis. The same worst-case latency, but here for Case B, is:

$$T_{\text{lat}} = T_{\text{slot}} (N_{\text{res}} M + Q) + T_{\text{show}} + \frac{T_{\text{prop}} (1 + N_{\text{ord}})}{N_{\text{res}}}$$  \hspace{1cm} (22)$$

and varies with respect to both $Q$ and $M$ but not with respect to $N_{\text{ord}}$.

For the ability to give a bounded latency coupled to an RTVC, consideration must be given to the placement of the slots reserved for the RTVC in the cycle. The $N$ slots reserved for an RTVC are denoted as $s_i$, $1 \leq i \leq N$. The worst-case latency is:
where \( N_{\text{gap}_i} \) is the number of slots between the start of slot \( s_i \) and the start of slot \( s_{i+1} \) (i.e., in next cycle if \( i = N \)). The number of slots for which the source node is handing over the role of being slot initiator to the next node during this time is denoted as \( N_{\text{SI Pass}_i} \).

The maximum aggregated throughput in the network is:

\[
S_{\text{max}} = \frac{(N_{\text{ord}} + N_{\text{res}})M_P T_{\text{slot}}}{(N_{\text{ord}} + N_{\text{res}})MT_{\text{slot}} + T_{\text{prop}}} \tag{24}
\]

where \( P \) is the average number of transmissions possible per slot as a consequence of spatial slot reuse and Case A is assumed. The maximum aggregated throughput is plotted in Figure 64 for \( P = 1 \). When \( P > 1 \), the throughput plotted in the figure is simply multiplied by \( P \). At low values of \( T_{\text{prop}} \), throughputs near \( P \) can be achieved even for low values of \( M \). The corresponding throughput for Case B is:
and is plotted in Figure 65. As can be seen, the maximum throughput for Case B does not vary with respect to $N_{ord}$ and $N_{res}$. When $N_{ord} + N_{res} = 1$, Case A and Case B give the same aggregated throughput. If $N_{ord} + N_{res} > 1$, however, Case A gives better aggregated throughput.

The same performance difference as in the case of aggregated throughput is found when looking at the throughput corresponding to one slot per cycle. The single-slot throughput is relevant information for all kinds of traffic, but especially when calculating guaranteed bandwidth for an RTVC. It also determines the minimum throughput a node gets if it only can send in a single ordinary slot per cycle, i.e., the only ordinary slot for which the node is slot initiator if $N_{ord} = 1$. The single slot throughputs are:

$$S_{\text{max}} = \frac{T_{\text{slot}}}{(N_{ord} + N_{res})MT_{\text{slot}} + T_{\text{prop}}}$$ \hspace{1cm} (26)
We have presented a fiber-ribbon based ring network with services for parallel and distributed real-time systems. A key component of the network architecture is the flexible control channel that can be configured to be used for different types of network control. Examples of this are the low level support for barrier synchronization, global reduction, and reliable transmission. A great advantage is that the low level support can be implemented with little or no modifications to existing hardware. High throughputs can be achieved in the network, especially in systems with some kind of pipelined dataflow between the nodes. The network offers real-

\[
S_{max} = \frac{T_{slot}}{(N_{ord} + N_{res})(MT_{slot} + T_{prop})}
\]  

(27)

for Cases A and B, respectively.

9.8 Summary

We have presented a fiber-ribbon based ring network with services for parallel and distributed real-time systems. A key component of the network architecture is the flexible control channel that can be configured to be used for different types of network control. Examples of this are the low level support for barrier synchronization, global reduction, and reliable transmission. A great advantage is that the low level support can be implemented with little or no modifications to existing hardware. High throughputs can be achieved in the network, especially in systems with some kind of pipelined dataflow between the nodes. The network offers real-

Figure 65: Maximum aggregated throughput. Case B is assumed.
time services for logical connections with guaranteed performance, RTVCs, and for separate messages, best effort and guarantee seeking messages. An analysis of worst-case latency and deterministic throughput has been provided for two variants of time slot organization. One offers higher throughput while the other offers lower latency in some situations. Also worth mentioning is that the network can be built today using fiber-optic off-the-shelf components and that this is ongoing work.
The physical ring of the second ring network proposed in this thesis is subdivided into two networks carrying different kinds of traffic [Jonsson et al. 1997B]. Nine of the fibers are used for time multiplexed circuit switched traffic; eight fibers are for data and one is for clocking. The tenth fiber is dedicated to packet switched traffic using, for example, a token ring protocol. This fiber also carries control messages to reconfigure the TDMA schedule (i.e., circuit establishment) for the other nine fibers. This network is a good choice when the main data flow in the network does not change rapidly. Compared to the CC-FPR network, the network for both packet and circuit switched traffic is slightly simpler at the expense of somewhat reduced support of dynamic traffic patterns. However, in many systems, only a fraction of the traffic is, e.g., "bursty".

Circuit switched and packet switched traffic are discussed in Sections 10.1 and 10.2, respectively. Section 10.3 describes circuit establishment, while a case study is provided in Section 10.4. Mode changes are treated in Section 10.5, and a summary is given in Section 10.6.

10.1 Circuit Switched Traffic

For circuit switched traffic, the first nine fibers in each link form a high speed channel. All of the high speed channels together form a high speed ring network for circuit switched traffic. The access is divided into slots as in an ordinary TDMA network. However, in each slot, the network can be divided into segments as in the CC-FPR network. Also, for each slot, there is always a slot initiator node. The same kind of asynchronous slot synchronization method is also used.

The access is cyclic, and each cycle consists of $K$ slots. In a typical case, $K$ is a multiple of $M$, where $M$ is the number of nodes, and each node is the slot initiator in $K/M$ slots. An example of an agreed schedule for a network with $K=M=5$ slots per cycle is shown in Figure 66. Each column represents one time slot and contains information on how the ring is segmented in that slot. Each number in a column is the node index of the owner of the corresponding link. The bold type numbers indicate the current slot initiator. In each segment and slot, one, and only one, node can be the owner of the links and hence has the right to use the segment links for transmission. In the first slot in the example, node $m_1$ (slot initiator) owns the link between itself and node $m_2$. Hence, it can transmit to node $m_2$ but not to any other node. In the same slot, node $m_2$ can transmit to any of
nodes $m_3$, $m_4$, $m_5$, or $m_1$. The choice is made by the process that owns the circuit (logical connection) to which the slot segment is associated. A multicast to two or more of these nodes is also possible.

In the third slot, the link between node $m_1$ and node $m_2$ is free. Although the link is free, node $m_1$ must not disturb the asynchronous slot synchronization technique and therefore transmits an empty packet to node $m_2$. In the fifth slot, node $m_5$ has the capability of transmitting a broadcast packet (a packet to all other nodes in the ring).

The same reasoning about bandwidth utilization according to the asynchronous slot synchronization as in the CC-FPR network holds for this network as well. The maximum aggregated throughput of the network made possible by the asynchronous slot synchronization method, $S_{\text{max}}$, is:

$$S_{\text{max}} = \frac{KPT_{\text{slot}}}{KT_{\text{slot}} + T_{\text{prop}}}$$

(28)

where $K$ is the number of slots per cycle, $P$ is the average number of packet transmissions in each slot, $T_{\text{slot}}$ is the duration of one slot, and $T_{\text{prop}}$ is the total propagation delay around the ring.

The latency grows linearly with distance, measured in number of hops (repeating latency in each node). Adding to the latency is also the propagation delay between source and destination node, as well as the delay until the first available slot for transmission. By distributing tasks in such a way as to minimize the number of hops, both latency and remaining bandwidth will be improved.
10.2 Packet Switched Traffic

The tenth fibers from all of the links are combined to form a ring network totally dedicated to packet switched traffic. An ordinary ring protocol can be used. However, there are two requirements: (i) it must be possible to halt the protocol when special packets for circuit establishment are to be transmitted (see Section 10.3), and (ii) the latency must be upper bounded to assure transmission of the packets for circuit establishment. When using, e.g., a token ring protocol on the packet network, this network will support low latency communication for sporadic packets at moderate traffic rates. At the same time, it is assured that the circuit switched traffic (often real-time traffic) is not disturbed by packet switched traffic.

10.3 Circuit Establishment

When a node is to establish a new circuit, it searches for slots where the required links are free so that the allocation of a new segment can be made. First, the node's own slots (i.e., where the node itself is the slot initiator) are searched. When too few slots (actually only a segment in each slot) for the circuit can be allocated, the search is continued in other slots. In that case, a special request packet is transmitted on the packet network to ask other nodes to allocate the desired segment in their slots. This packet is immediately followed by a collect packet to collect information on the success of the slot segment allocations.

The request packet, which is broadcast to all other nodes, contains information about the links required and the number of slots needed. Each node then checks whether any of its own slots have the required free links. If so, it prepares to modify the collect packet when it arrives (before forwarding it), to notify the requesting node of which slots have been allocated. However, if any of the previous nodes have already allocated slot segments and modified the collect packet, the number of slots needed would be decreased by the corresponding number of allocated slots. The number of slots still needed is indicated by a dedicated field in the collect packet. In this way, allocation of more than the necessary number of slots is avoided. However, several nodes can each allocate some of the slots needed, and information about all of these allocations is added to the same collect packet.

When the requesting node receives the collect packet after one round, it decides whether the number of allocated slots is sufficient. If not, it sends a release packet. Otherwise, it can start using the established circuit immediately.
10.4 Case Study

With the same assumptions as are made in the case study described in Section 9.6 (Page 155), we will now show that the second pipeline ring network is also feasible for the chosen radar signal processing application. Since we still assume links with ten fibers and 800 Mbit/s per fiber, a bandwidth of 6.4 Gbit/s is dedicated to circuit switched traffic on eight of the fibers, while 800 Mbit/s is dedicated to packet switched traffic on one fiber.

We also assume there are still 45 slots per cycle, which gives a bandwidth of 133 Mbit/s per cycle and slot. In the case of the second network, this means that each cycle is divided into \( K = 45 \) slots per cycle for circuit switched traffic. The allocation scheme in Figure 59 on Page 157 then also holds for this network, leaving Slots 1 through 15 free. Another possibility is to have \( K = 12 \) slots per cycle. In that case, one slot corresponds to 500 Mbit/s, and all bandwidths in Figure 58 on Page 156 are divisible by the slot bandwidth.

10.5 Mode Changes

It is possible to use a number of different working modes 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 may be to track a certain object. Normally, the algorithm mapping and communication patterns are different for two different modes. The change of the circuits at mode changes can be performed in two different ways: \( (L) \) switching between the various slot allocation schemes for a known set of modes, schemes that are statically stored in each node, and \( (ll) \) dynamically changing the slot allocation scheme in each node at a mode change by establishing new circuits, as described in Section 10.3.

In the first case, a mode change request is broadcast by the subsystem responsible for mode changes (which is the master node). Immediately after transmission of the request packet, an acknowledge packet is transmitted. The acknowledge packet is halted in each node until the node is prepared for the requested mode change. In this way, the master node knows that all involved nodes are prepared for the mode change when it receives the acknowledge packet after one turn around the ring. Whether other packets are allowed during a mode change depends on the tolerable latency of the mode change.

In the second case, the mode change is initiated by the master node in the same manner as a broadcast packet. However, each of the nodes involved is then responsible for requesting its required bandwidth. Each node also sends its own acknowledge (or negative acknowledge if it failed to establish
the required circuits) packet to the master node, indicating that it is prepared for the mode change.

When all the nodes have been prepared for the mode change, the system will change to the new mode in the next batch. The packets coming from the antenna in the new batch will be tagged to indicate the new mode. In that way, the nodes will be triggered to change to the new mode. Nodes that are placed later in the signal processing chain are triggered by the packets generated by succeeding nodes. Even if a node has a totally different job to do (and a different communication pattern) in the new mode, it can be triggered in this way. This is possible as long as two different batches are data independent.

10.6 Summary

We have presented a ring network in which very high throughputs can be achieved, especially in systems having some kind of pipelined dataflow between the nodes. The network supports packet switched traffic at the same time as guaranteed bandwidth is supported through circuit switching. In a typical system, circuits can be set up for time critical dataflows, guaranteeing that they are not disturbed by, e.g., control information. These network features are highly appreciated in, e.g., radar signal processing systems. It is also worth mentioning that the network can be built today using fiber-optic off-the-shelf components.
11. Conclusions and Future Work

11.1 Conclusions

Two different kinds of networks have been in focus, WDM star networks and fiber-ribbon ring networks. Protocols and real-time services were proposed and evaluated for both networks (see Chapters 7 and 9, respectively). Because the WDM star network implements a distributed crossbar it has better support for general communication patterns, compared to the ring. From the perspective of fault-tolerance (which, however, not has been addressed in this thesis) the passive optical star seems to be more reliable than a uni-directional ring, due to its passive nature. On the other hand, the ring network allows for simpler transceiver design as a consequence of the relaxed clock synchronization (because of the asynchronous slot-synchronization method), lack of tuning latencies, and the fact that no propagation delay measurements has to be done. In practice, this can lead to shorter time-slots which, in turn, reduces the latencies. Two other important factors are what kind of traffic (communication pattern etc.) that should be carried by the network and how the different technologies will mature.

A star-of-stars WDM network has also been proposed, offering a scalable network architecture (see Chapter 8). Moreover, a simplified variant of the fiber-ribbon ring network with less support for dynamic traffic was proposed (Chapter 10). Again, the choice of network depends much upon things such as the communication patterns in the application. The WDM star network as well as the two ring networks have been shown to be feasible in a high-performance radar system, and the star-of-stars network is offered as a more scalable alternative. More general analyses of the networks, with a focus on support for real-time communication have also been presented.

11.2 Ongoing and Future work

A network demonstrator of the control channel based ring network is currently being built, and some initial tests with two nodes have been performed (see Figure 67). Some building blocks included in a node are:

- Optoelectronic conversion and fiber-ribbon attachment (Motorola OPTOBUS™).
- Protocol processor (a commercial DSP board together with an additional memory and I/O board).
Dual-ported memories and various control logic (the big cards on which other cards are mounted – see Figure 67). There are three memory banks for transmission, reception, and protocol interactions between the two processors, respectively.

Control channel logic implemented in an FPGA (not shown in the figure.

Application processor (a commercial DSP board together with an additional memory and I/O board).

A layered protocol hierarchy is also being developed for use in the demonstrator.

Other possible directions for future work are:

- Further development of methods for heterogeneous real-time services in high performance networks, i.e., services with different QoS. Funding for a project with this focus has recently been granted.

- Investigation of how similar protocols and network architectures can be used in free-space systems. Such activities have already been
initiated and include collaboration with research groups on photonics and micro-electronics at Chalmers University of Technology.

- More work with a focus on group communication and similar services.
- Further work on real-time services in hierarchical networks.
- Development of methods for, e.g., system startup, node insertion/deletion, and fault tolerance. Much research efforts have been paid to these problems and standards exist for specific LANs.

Finally, it should be noted that the work has been valuable for the industrial partner, Ericsson Microwave Systems, and may have impact on future computer system designs for, e.g., radar signal processing.
References


<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADM</td>
<td>Add-Drop Multiplexer</td>
</tr>
<tr>
<td>APD</td>
<td>Avalanche Photo Diode</td>
</tr>
<tr>
<td>ATM</td>
<td>Asynchronous Transfer Mode</td>
</tr>
<tr>
<td>AWG</td>
<td>Arrayed Waveguide Grating</td>
</tr>
<tr>
<td>BS</td>
<td>Barrier Synchronization</td>
</tr>
<tr>
<td>CAN</td>
<td>Control Area Network</td>
</tr>
<tr>
<td>CBR</td>
<td>Constant Bit Rate</td>
</tr>
<tr>
<td>CC-FPR</td>
<td>Control Channel based Fiber-ribbon Pipeline Ring</td>
</tr>
<tr>
<td>CDMA</td>
<td>Code Division Multiple Access</td>
</tr>
<tr>
<td>CFAR</td>
<td>Constant False Alarm Ratio</td>
</tr>
<tr>
<td>CM</td>
<td>Computation Module</td>
</tr>
<tr>
<td>COW</td>
<td>Cluster of Workstations</td>
</tr>
<tr>
<td>CPI</td>
<td>Coherent Processing Interval</td>
</tr>
<tr>
<td>CSMA</td>
<td>Carrier-Sense Multiple-Access</td>
</tr>
<tr>
<td>CSMA/CD</td>
<td>CSMA with Collision Detection</td>
</tr>
<tr>
<td>DBR</td>
<td>Distributed Bragg Region</td>
</tr>
<tr>
<td>DFB</td>
<td>Distributed Feedback</td>
</tr>
<tr>
<td>DS-SS</td>
<td>Direct Sequence Spread Spectrum</td>
</tr>
<tr>
<td>DT-WDMA</td>
<td>Dynamic Time-Wavelength Division Multi Access</td>
</tr>
<tr>
<td>EDF</td>
<td>Earliest Deadline First</td>
</tr>
<tr>
<td>FDDI</td>
<td>Fibre Distributed Data Interface</td>
</tr>
<tr>
<td>FDM</td>
<td>Frequency Division Multiplexing</td>
</tr>
<tr>
<td>FIFO</td>
<td>First In First Out</td>
</tr>
<tr>
<td>FIG</td>
<td>Fiber Imaging Guide</td>
</tr>
<tr>
<td>FWHM</td>
<td>Full Width Half Maximum</td>
</tr>
<tr>
<td>GR</td>
<td>Global Reduction</td>
</tr>
<tr>
<td>ILP</td>
<td>Instruction Level Parallelism</td>
</tr>
<tr>
<td>I-TDMA</td>
<td>Interleaved TDMA</td>
</tr>
<tr>
<td>LED</td>
<td>Light Emitting Diode</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>MAC</td>
<td>Medium Access Control</td>
</tr>
<tr>
<td>MAGIC</td>
<td>Multistripe Array Grating Integrated Cavity</td>
</tr>
<tr>
<td>MIMD</td>
<td>Multiple Instruction streams Multiple Data streams</td>
</tr>
<tr>
<td>MSIMD</td>
<td>Multiple SIMD arrays</td>
</tr>
<tr>
<td>NOW</td>
<td>Network of Workstations</td>
</tr>
<tr>
<td>PCB</td>
<td>Printed Circuit Board</td>
</tr>
<tr>
<td>PE</td>
<td>Processing Element</td>
</tr>
<tr>
<td>PFDM</td>
<td>Pulse Frequency Division Multiplexing</td>
</tr>
<tr>
<td>PIN</td>
<td>P-insulator-N</td>
</tr>
<tr>
<td>PON</td>
<td>Passive Optical Networks</td>
</tr>
<tr>
<td>PVM</td>
<td>Parallel Virtual Machine</td>
</tr>
<tr>
<td>QoS</td>
<td>Quality of Service</td>
</tr>
<tr>
<td>RTVC</td>
<td>Real-Time Virtual Channels</td>
</tr>
<tr>
<td>SAN</td>
<td>System Area Network</td>
</tr>
<tr>
<td>SCMA</td>
<td>SubCarrier Division Multiple Access</td>
</tr>
<tr>
<td>SDM</td>
<td>Space Division Multiplexing</td>
</tr>
<tr>
<td>SI</td>
<td>Slot Initiator</td>
</tr>
<tr>
<td>SIMD</td>
<td>Single Instruction stream Multiple Data streams</td>
</tr>
<tr>
<td>SLM</td>
<td>Spatial Light Modulator</td>
</tr>
<tr>
<td>SPMD</td>
<td>Same Program Multiple Data</td>
</tr>
<tr>
<td>TDMA</td>
<td>Time Division Multiple Access</td>
</tr>
<tr>
<td>TTP</td>
<td>Time-Triggered Protocol</td>
</tr>
<tr>
<td>UBR</td>
<td>Unspecified Bit Rate</td>
</tr>
<tr>
<td>VBR</td>
<td>Variable Bit Rate</td>
</tr>
<tr>
<td>VCSEL</td>
<td>Vertical Cavity Surface Emitting Laser</td>
</tr>
<tr>
<td>WDDI</td>
<td>Wavelength Distributed Data Interface</td>
</tr>
<tr>
<td>WDM</td>
<td>Wavelength Division Multiplexing</td>
</tr>
<tr>
<td>WDMA</td>
<td>Wavelength Division Multiple Access</td>
</tr>
</tbody>
</table>