

Arndt Bode Mario Dal Cin (Eds.)

*ccc 7-732*

# Parallel Computer Architectures

Theory, Hardware, Software, Applications

BIBLIOTHEQUE DU CERIST

**Springer-Verlag**

Berlin Heidelberg New York

London Paris Tokyo

Hong Kong Barcelona

Budapest

## Series Editors

Gerhard Goos  
 Universität Karlsruhe  
 Postfach 69 80  
 Vincenz-Priessnitz-Straße 1  
 D-76131 Karlsruhe, Germany

Juris Harmanis  
 Cornell University  
 Department of Computer Science  
 4130 Upson Hall  
 Ithaca, NY 14853, USA

## Volume Editors

Arndt Bode  
 Institut für Informatik, TU München  
 Arcisstr. 21, D-80333 München, Germany

Mario Dal Cin  
 Institut für Mathematische Maschinen und Datenverarbeitung  
 Lehrstuhl für Informatik III  
 Martensstr. 3, D-91058 Erlangen, Germany

CR Subject Classification (1991): C.1-4, D.1, D.3-4, F.1.3

6340

ISBN 3-540-57307-0 Springer-Verlag Berlin Heidelberg New York  
 ISBN 0-387-57307-0 Springer-Verlag New York Berlin Heidelberg

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

© Springer-Verlag Berlin Heidelberg 1993  
 Printed in Germany

Typesetting: Camera-ready by author  
 Printing and binding: Druckhaus Beltz, Hemsbach/Bergstr.  
 45/3140-543210 - Printed on acid-free paper

# Preface

Parallel computer architectures are now going to real applications! This fact is demonstrated by the large number of application areas covered in this book (see section on applications of parallel computer architectures). The applications range from image analysis to quantum mechanics and data bases. Still, the use of parallel architectures poses serious problems and requires the development of new techniques and tools.

This book is a collection of best papers presented at the first workshop on two major research activities at the Universität Erlangen-Nürnberg and Technische Universität München. At both universities, more than 100 researchers are working in the field of multiprocessor systems and network configurations and methods and tools for parallel systems. Indeed, the German Science Foundation (Deutsche Forschungsgemeinschaft) has been sponsoring the projects under grant numbers SFB 182 and SFB 342. Research grants in the form of a Sonderforschungsbereich are given to selected German Universities in portions of three years following a thoroughful reviewing process. The overall duration of such a research grant is restricted to 12 years. The initiative at Erlangen-Nürnberg was started in 1987 and has been headed since this time by Prof. Dr. H. Wedekind. Work at TU-München began in 1990, head of this initiative is Prof. Dr. A. Bode. The authors of this book are grateful to the Deutsche Forschungsgemeinschaft for its continuing support in the field of research on parallel processing.

The first section of the book is devoted to hardware aspects of parallel systems. Here, a number of basic problems has to be solved. Latency and bandwidths of interconnection networks are a bottleneck for parallel process communication. Optoelectronic media, discussed in this section, could change this fact. The scalability of parallel hardware is demonstrated with the multiprocessor system MEMSY based on the concept of distributed shared memory. Scalable parallel systems need fault tolerance mechanisms to guarantee reliable system behaviour even in the presence of defects in parts of the system. An approach to fault tolerance for scalable parallel systems is discussed in this section.

The next section is devoted to performance aspects of parallel systems. Analytical models for performance prediction are presented as well as a new hardware monitor system together with the evaluation software.

Tools for the automatic parallelization of existing applications are a dream, but not yet reality for the user of parallel systems. Different aspects for automatic treatment of parallel applications are covered in the next section on architectures and tools for parallelization. Dynamic load balancing is an application transparent mechanism of the operating system to guarantee equal load on the elements of a multiprocessor system. Randomized shared memory is one possible implementation of a virtual shared memory based on distributed memory hardware.

Finally, optimizing tools for superscalar, superpipelined and VLIW(very long instruction word)-architectures already cover automatic parallelization on the basis of individual machine instructions.

The section on modelling techniques groups a number of articles on different aspects of object oriented distributed systems. A distribution language, memory management and the support for types, classes and inheritance are covered. Formal description techniques are based on Petri nets and algebraic specification.

Finally, the section on applications covers knowledge based image analysis, different parallel algorithms for CAD tools for VLSI design, a parallel sorting algorithm for parallel data bases, quantum mechanics algorithms, the solution of partial differential equations, and the solution of a Navier Stokes solver based on multigrid techniques for fluid dynamic applications.

Erlangen and München, March 1993

Arndt Bode  
Chairman SFB 342

Hartmut Wedekind  
Chairman SFB 182

Mario Dal Cin  
SFB 182

# Contents

## Hardware Aspects of Multiprocessor Systems

### Optoelectronic Interconnections

J. Schwider, N. Streibl, K. Zühl (Universität Erlangen-Nürnberg) ..... 1

### MEMSY – A Modular Expandable Multiprocessor System

F. Hofmann, M. Dal Cin, A. Grygier, H. Hessenauer, U. Hildebrand,  
C.-U. Linster, T. Thiel, S. Turowski (Universität Erlangen-Nürnberg) ..... 15

### Fault Tolerance in Distributed Shared Memory Multiprocessors

M. Dal Cin, A. Grygier, H. Hessenauer, U. Hildebrand, J. Hönig,  
W. Hohl, E. Michel, A. Pataricza (Universität Erlangen-Nürnberg) ..... 31

## Performance of Parallel Systems

### Optimal Multiprogramming Control for Parallel Computations

E. Jessen, W. Ertel, Ch. Suttner (Technische Universität München) ..... 49

### The Distributed Hardware Monitor ZM4 and its Interface to MEMSY

R. Hofmann (Universität Erlangen-Nürnberg) ..... 66

### Graph Models for Performance Evaluation of Parallel Programs

F. Hartleb (Universität Erlangen-Nürnberg) ..... 80

## Architectures and Tools for Parallelization

### Load Management on Multiprocessor Systems

Th. Ludwig (Technische Universität München) ..... 87

### Randomized Shared Memory - Concept and Efficiency of a Scalable Shared Memory Scheme

H. Hellwagner (Siemens AG, München) ..... 102

### Methods for Exploitation of Fine-Grained Parallelism

G. Böckle, Ch. Störmann, I. Wildgruber (Siemens AG, München) ..... 118

## Modelling Techniques

### Causality Based Proof of a Distributed Shared Memory System

D. Gomm, E. Kindler (Technische Universität München) ..... 133

### Object- and Memory-Management Architecture

- A Concept for Open, Object-Oriented Operating Systems -

J. Kleinöder (Universität Erlangen-Nürnberg) ..... 150

An Orthogonal Distribution Language for Uniform Object-Oriented Languages M. Faustle (Universität Erlangen-Nürnberg) .....	166
Towards the Implementation of a Uniform Object Model F. Hauck (Universität Erlangen-Nürnberg) .....	180
FOCUS: A Formal Design Method for Distributed Systems F. Dederichs, C. Dendorfer, R. Weber (Technische Universität München) .....	190
<b>Applications of Parallel Systems</b>	
Parallelism in a Semantic Network for Image Understanding V. Fischer, H. Niemann (Universität Erlangen-Nürnberg) .....	203
Architectures for Parallel Slicing Enumeration in VLSI Layout H. Spruth, F. Johannes (Technische Universität München) .....	219
Application of Fault Parallelism to the Automatic Test Pattern Generation for Sequential Circuits P. Krauss, K. Antreich (Technische Universität München) .....	234
Parallel Sorting of Large Data Volumes on Distributed Memory Multiprocessors M. Pawlowski, R. Bayer (Technische Universität München) .....	246
Quantum Mechanical Programs for Distributed Systems: Strategies and Results H. Früchtl, P. Otto (Universität Erlangen-Nürnberg) .....	265
On the Parallel Solution of 3D PDEs on a Network of Workstations and on Vector Computers M. Griebel, W. Huber, T. Störtkuhl, C. Zenger (Technische Universität München) .....	276
Numerical Simulation of Complex Fluid Flows on MIMD Computers M. Perić, M. Schäfer, E. Schreck (Universität Erlangen-Nürnberg) .....	292
<b>Index</b> .....	307

# Optoelectronic Interconnections

Johannes Schwider, Norbert Streibl, Konrad Zühl  
Physikalisches Institut der Universität Erlangen-Nürnberg  
Staudtstr. 7, D-8520 Erlangen, Germany.

## 1 Bandwidth

The overall performance of data processing machines can be increased in two ways: (i) by using faster system clocks and (ii) by using parallel systems consisting of a multitude of interconnected processing elements. In the near future central aims of information technology are the development of teraflop ( $10^{12}$  floating point operations per second) supercomputers and switching networks for telecommunications with terabit bandwidth.

With an acceleration of the system clock alone both of these aims cannot be achieved. A data processing system contains three basic functions: (i) active combining and switching of data, (ii) passive transport of data and (iii) storage of data (which often is implemented by flip-flops, that is by active devices). In quantum-electronics the fastest components are resonant tunneling diodes with a response, measurable for example by nonlinear frequency mixing, beyond 1 THz [Sol 83]. These frequencies belong already to the far infrared region of the electromagnetic spectrum. The simplest circuit, a ring oscillator consisting of two connected active devices in a loop, runs at about 300 GHz [Bro 88]. Today somewhat more complex integrated circuits in GaAs-technology are in research with on the order of 30 GBit/s bandwidth. Still more complex high-speed components, for example multiplexers and demultiplexers for fiber optical links with some 10 GBit/s are commercial. Modern digital telephone exchanges handle data with several hundred MBit/s. The characteristic data rate of a personal computer, determined by the time required for memory access, is on the order of 10 MBit/s. Consequently, one observes the trend summarized in table 1: Although ultrafast devices do exist, complex systems are necessarily slow.

The reason are the fundamental electromagnetic properties of electrical interconnections at high frequencies. If the wavelength of the electromagnetic radiation (associated with the frequency content of the signals to be transported) and the length of the line have similar order of magnitude, a vast variety of problems arises. Efficient screening is required to prevent crosstalk through radiation. Standing waves, reflections and echoes occur unless all lines are correctly terminated (which by the way is expensive in terms of energy). Impedance matching and 'microwave design rules' are required for splitting and joining signal lines. As a consequence, the fastest standard bus system (FutureBus II) supports today only 100 MBit/s per line.

On the other hand, optics is good at communicating informations. Recently, optoelectronic interconnections are an area of active research [Hase 84, Good

Table 1: Complexity and speed of electronic systems

System	bandwidth	number of parts
resonant tunneling devices	3	THz 1
ring oscillator	0.3	THz 2
microwave IC (GaAs)	0.03	THz several
telecommunications, supercomputer	0.000 3	THz many
personal computer	0.000 03	THz cheap

84, Kos 85, Berg 87, Sto 87, Cha 91, Die 92, Bac 92, Par 92]. Optical beams can freely cross through each other without interaction. Optics supports parallel interconnections, either through fiber bundles or through imaging systems. Optocouplers are widely used for isolation and to prevent ground loops. Finally, with state of the art optoelectronic devices the heat dissipation at the beginning and the end of an interconnection can be very small, in contrast to the electronic line drivers, that must be large in order to move necessarily large currents. Thus, the basic impedance matching problem is alleviated by the use of optical interconnects [Mil 89]. Because performance or packing density (or both) in modern electronics are limited by heat dissipation, the use of optoelectronics should yield definite advantages.







Over long distances and at high data rates optical fibers have already supplanted electrical connections. The basic question of the research in optoelectronic interconnections is: how short can optical interconnections be and still offer advantages over electronics? It seems fairly clear, that high performance systems will use optics for connections between modules and boards, later on maybe even between hybrids and chips. On the other hand, an all-optical computer looks today as a worthwhile but maybe elusive aim of basic research.

## 2 Parallelism

Basically, parallel interconnections between a number of participants may have different dimensionality: Fibres (or wires) are one-dimensional connections. A printed circuit board, the wiring of gates on a chip or integrated optics offer essentially two-dimensional interconnections. The number of bonding pads at the edge of a chip scales with the linear dimension, the number of gates with the area. Therefore, quasiplanar interconnections cause a bottleneck in a complex system because much fewer connections are available than active devices. A good way out are three-dimensional interconnections through the 3D space above the plane, where the active devices are located. Then the number of interconnections and the number of devices scale essentially in the same way with the area. A numerical example is worthwhile: If each channel requires  $300 \mu\text{m}$  space and a chip or a connector has an overall size of  $1 \text{ cm}$ , then with two-dimensional connections we obtain a parallelism of 32 channels/cm, but with three-dimensional connections  $32 \times 32 = 1024 \text{ channels/cm}^2$ . Hence, three-dimensional interconnections support highly parallel systems.



**Table 2: Parallel interconnection topologies**

connection type	#sources	#receivers	example	schematic
ordered point to point	1	1	wire bundle	
random point to point	1	1	switch fabric	
broadcasting, fan-out	1	N	clock distrib.	
fan-in	N	1	interrupt	
bus	N	N	bus system	
reconfigurable	N	N	telephone	

Optical communications is used across long distances, but in these applications usually only one line is necessary. The shorter the distance the higher is the required parallelism. Between subsystems and modules in a computer the interconnections are provided by a bus, which has on the order of 100 parallel lines. Modern chip packages have several hundred pins, hence optical chip to chip interconnections are worthwhile only if a parallelism of on the order of 1000 lines is provided. Optical gate to gate interconnections become interesting only if some  $10^4$ – $10^6$  gates can be 'wired'.

As the degree of parallelism is increased, the data rate of the individual lines decreases: whereas a single fiber optical communications line might run at 20 GBit/s, a bus system should run at the systems clock rate, that is on the order of several 100 MBit/s in a high performance system. Otherwise different clocks must be used within a single subsystem for computing and outside communications, which is not practical in many cases. Also the cost for time multiplexing in terms of gate delays, space, heat dissipation and — last but not least — money is not negligible.

Another important feature that may serve to classify interconnection systems is topology: As shown in table 2 there are a number of different parallel interconnection topologies which are increasingly difficult to implement. The simplest approach are ordered parallel point to point connections. The electronic implementation is a number of parallel wires, optically they may be implemented by a fiber bundle, a fiber ribbon cable or by an imaging system that images an array of light sources onto a receiver array.

Somewhat more complicated are permutation elements, that is random point to point interconnections. They allow to change the geometrical order in a bundle of parallel connections. Such permutations are required in many algorithms and therefore in many special purpose machines, for example in sorting and searching and therefore in switching networks and telephone exchanges (packet switch), in fast data transformations such as the fast Fourier transform and therefore in signal processors.

Multipoint interconnection may allow (i) fan-out, i.e. broadcasting of a signal from one source to several receivers, or (ii) fan-in, i.e. the listening of one receiver into the signals transmitted by several receivers, or (iii) the combination of fan-out and fan-in, i.e. the sharing of a common communications line by several participants such as a bus line. Finally, and most complicated, there are reconfigurable interconnections, where the 'wiring' of the participants can be changed. A crossbar or a telephone exchange are examples of such a connection.

For all of these topologies optical implementations have been proposed in the past, see for example [Hase 84, Good 84, Kos 85, Berg 87, Sto 87, Cha 91, Die 92, Bac 92, Par 92], some of which will be presented in the following.

### 3 Optical backplane

Optoelectronic interconnections between boards (distance  $x$  up to 1 m) and integrated circuits (distance  $x$  on the order of 1 cm) based on a light guiding plate have been widely studied (for example: Hase 84, Brenn 88, Herr 89, Jah 90, Haum 90, Cha 91, Par 92, Stre 93). They have been proposed to serve as 'optical backplane' or 'optical printed circuit board'. Fig. 1 shows the basic optical setup for one single communication channel, which may be replicated for parallel (multichannel) interconnections.

A thick plate of glass or polymer material is used to guide the optical signals. It may be considered as an extremely multimodal waveguide or simply as free

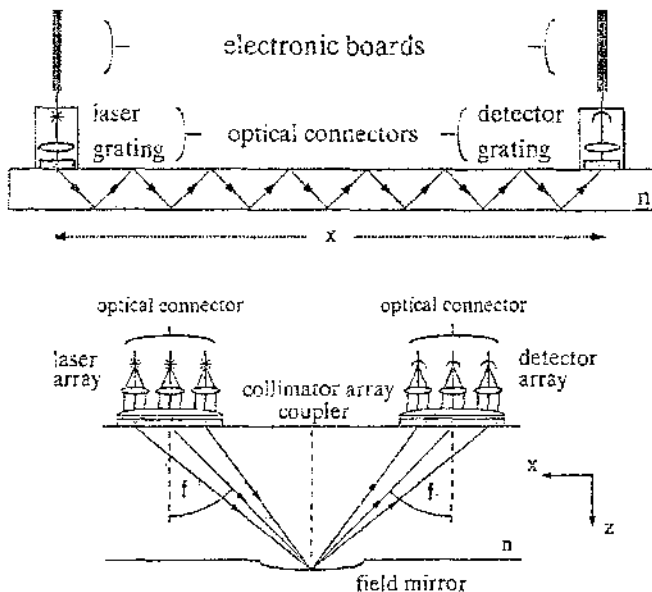


Figure 1: (a) Principle of a light guiding plate. (b) a curved mirror between the optical connectors on the light guiding plate images the transmitters onto the receivers.

space 'filled' with transparent material. Its advantage in comparison to free space communication is that it acts as a mechanically well defined base plate for the 'optical connectors' and at the same time protects the optical signals from external disturbances such as dust, air turbulence etc. Its advantage compared to single mode waveguides is that the required tolerances and the alignment problems are much less severe – at least as long as the detectors for the optical signals are not too small.

A collimated beam from a semiconductor laser located on the optical connector is deflected by a grating by a large angle  $\varphi$ , coupled into the light guiding plate, propagates towards the receiver via multiple reflections, is coupled out of the plate by another grating at the second optical connector and detected by the receiver. Holographically recorded volume phase gratings in dichromated gelatine were reported to have excellent coupling efficiency (loss on the order of less than 0.5 dB per coupler for slanted gratings with  $45^\circ$  deflection angle and 740 nm period at 786 nm wavelength) [Hlaum 90]. Light guiding is achieved with a loss on the order of 0.1 dB/cm. Point to point interconnections thus suffer from losses on the order of 3 – 10 dB, depending on the communication distance.

The simple setup of fig. 1a has two basic drawbacks: firstly, the beam is divergent due to diffraction at the aperture of the optical connector, which severely limits the packing density for parallel channels; secondly, the deflection angle is a function of the wavelength, which has as a consequence tight tolerance requirements for the laser wavelengths. More specifically: For an interconnection length  $L = x/\sin\varphi$  within a light guiding plate with refractive index  $n$  and for light with the wavelength  $\lambda$  in vacuo a beam diameter of at least  $d_{min} \approx (L\lambda/n)^{1/2}$  is required in order to avoid excessive spreading of the beam by diffraction. Adjacent parallel channels have to be separated from each other by a multiple of this minimum beam diameter  $d_{min}$  in order to avoid crosstalk. Thus, for board distances  $x$  of up to 1 m a packing density of not too much more than 10 channels/cm<sup>2</sup> can be implemented with reasonable signal to noise ratio. Such a low packing density is competitive with electronics only at extremely high data rates. Specifically it also prohibits the use of monolithically integrated and therefore densely packed laser and detector arrays.

The chromatic aberration (grating dispersion) makes the deflection angles wavelength dependent and causes problems with 'aiming' the beams at the output couplers. Fabrication tolerances, mode hopping and thermal drift may lead to significant differences in the wavelength of the individual semiconductor lasers. For a deflection angle  $\varphi \approx 45^\circ$ , which is preferred in order to eliminate the effects of thermal expansion of the light guiding plate, the distance of the optimum position of the coupler and the wavelength have the same relative deviation  $\delta x/x \approx 2\delta\lambda/\lambda$ . Hence, the cost for selecting and controlling the laser wavelength for interconnection distances of up to 1 m is prohibitive.

Both problems, diffractive broadening as well as chromatic aberration, can be overcome by imaging [Stre 93]. A field lens (or a mirror or a diffractive zone plate) between the optical connectors can be used to image the apertures of the transmitters onto those of the receivers. Fig. 1b shows the principle of such a parallel interconnection. The lasers of a laser array are collimated by a

geometrically similar array of microlenses. An additional lens, whose function may be incorporated into the coupling hologram, images all lasers onto one point of the light guiding plate. There the vertex of the mirror is located that performs the one to one imaging of the apertures of the microlenses and therefore acts as the field lens. At the receiver site a completely symmetrical setup is used to focus down onto the detectors. In practice the light guiding plate may be thinner than shown in fig. 1b, if the light path is folded by multiple reflections as in fig. 1a. For long interconnections a chain of several lenses may be used to relay the image.

Imaging guarantees, that light from each transmitter aperture is focused onto the receiver aperture independently from small errors  $\delta\varphi$  in its propagation direction provided that the field lens is sufficiently large to catch the beam. Therefore the chromatic aberration of the grating couplers is completely eliminated by this design. Also, the optical setup is completely symmetrical, which eliminates all odd order monochromatic aberrations: specifically, the image is free from geometrical distortion, which is important for array operation. Moreover, coma is corrected, which leads us to expect good image quality off axis. Hence, the size of the arrays and thus the number of possible parallel channels may be significant. At the optical connectors the system is 'telecentric', which means that the principal rays for all channels are parallel. Thus only the angular alignment of the connectors is critical, the tolerances for displacement are somewhat less severe. Also, the microlenses are used on axis which reduces aberrations of the focal spots on the detectors. A full design of the optical setup is given in [Stre 90] and includes aberration analysis. It is shown theoretically for all interconnection distances up to 1 m and practically in a feasibility experiment for an interconnection of about 20 cm length, that within a field (= cross section of the optical connector) of  $1 \text{ cm}^2$  some 1000 optical channels can be transmitted in parallel.

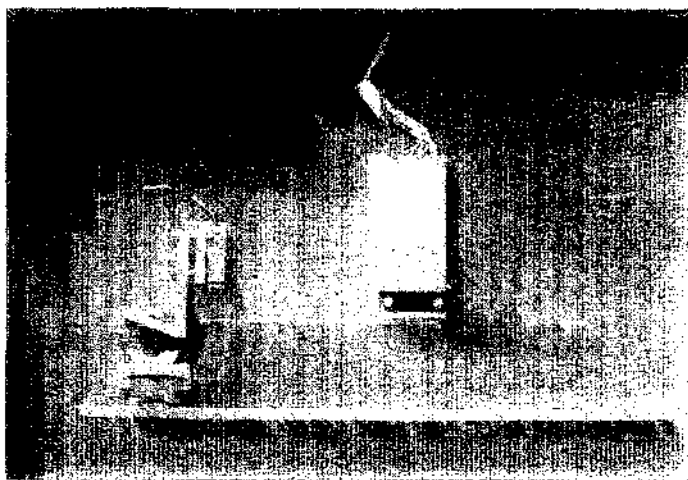


Figure 2: Experimental setup: Light guiding plate connecting two participants

## 4 Optical bus

In a preceding section the difficulties of impedance matching at high data rates were mentioned that are incurred with a bus system having many taps coupling signals in and out a common communications line. Consequently, at high data rates, beyond some 100 MBit/s star couplers, i. e. broadcasting topologies are employed instead of busses which is expensive. Also in much slower systems, such as a multiprocessor, optical implementations of a bus is worthwhile because of synchronicity: optical interconnections easily allow to control propagation delays down to picoseconds and are consequently very well suited for the implementation of global synchronisation modules or clock distribution within a multiprocessor system.

Bus lines are used in electronics to save complexity: instead of wiring  $N$  participants with  $(N^2 - N)/2$  individual communication lines, they all share one common bus line. The power of each emitter has to be split into  $N$  parts for the  $N$  listeners, which requires a multiple  $1 : N$  beam splitter. Also each listener must receive power from  $N$  different emitters, which might or might not involve a second  $1 : N$  beam splitting component. This beam combination on the receivers involves basic physical questions [Krack 92] regarding the possibility of lossless fan-in in singlemode systems. In a free space optical system an overall theoretical efficiency of  $1/(2N-1)$  compared to the absolute theoretical minimum of  $1/N$  is achieved, if only one single beamsplitter is used for transmitters and receivers simultaneously [Krack 92]. Fig. 3 shows an optical implementation involving a Dammann grating [Damm 71] or a similar phase-only diffraction grating as multiple beam splitter.

Each participant in a bus line has optical transmitters and receivers, and a fiber link transporting the optical signals from the electronic board to an all-

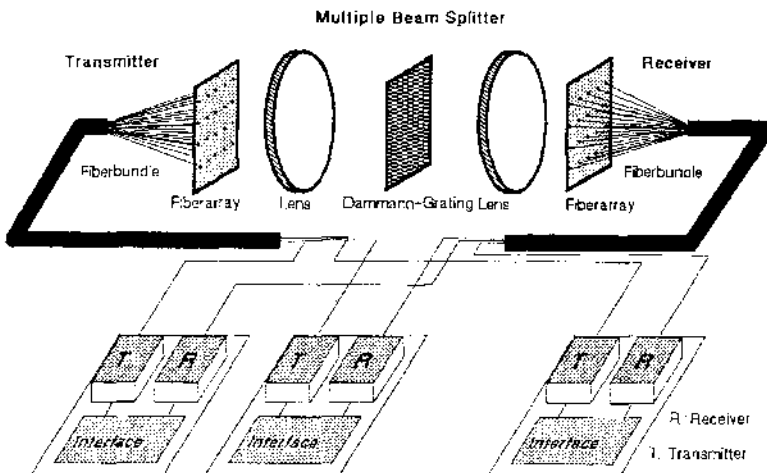


Figure 3: Principle of an optical bus.

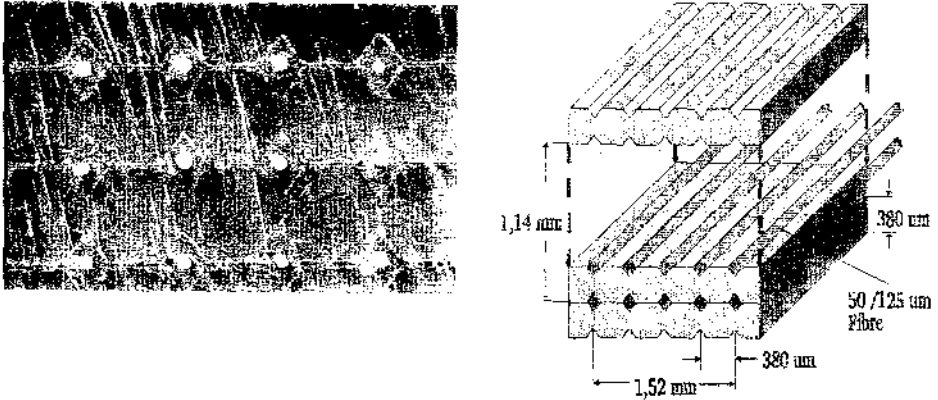


Figure 4: *Fiber end plate for coupling fibers to an optical interconnect stage.*

optical interconnection stage. The fibers end in a regular array in a fiber end plate. Fig. 4 shows a photograph of such a device. It was fabricated by wet chemical etching of V-grooves on both sides of silicon wafers in 100 orientation by using KOH. These grooves can be applied very accurately with microlithographic precision. In a selfaligned assembling procedure alternately a layer of fibers and a grooved silicon plate are stacked onto each other. All components are glued together and the fiber end plate is polished. In this way an accurate array of fibers can be fabricated.

The fibers allow a mechanically flexible coupling of many participants into a highly accurate optical system without putting tight geometrical constraints on the construction of the electronic system. Additionally it concentrates all signals coming from different, possibly widely spaced participants within a tight volume which can be handled with a compact optical system: for example the end plate could have a size of  $1\text{ cm}^2$ . Finally the fiber concentrator puts all the signals on a well defined place in the plate with small tolerances. By using a telecentric one to one imaging system the end plates are imaged onto each other. Without the multiple beam splitter this imaging system would implement an ordered point to point connection. However, in the filter plane a multiple beam splitter such as a Damman grating or a similar device is inserted, whose optical function is shown in fig 5.

Such a grating has rectangular surface corrugations that are fabricated by microlithographic methods. The component shown in fig. 5 was designed by using nonlinear optimization, its structure was plotted by using a laser beam writing system and it was etched into a fused silica substrate by using reactive ion etching [Hasel 92]. By performing a diffraction experiment it can be seen that one incoming beam of light is split into a multitude of beams, that is the grating performs a fan-out. Similarly, since the optical system in fig. 4 is completely symmetric the multiple beam splitter performs at the same time a fan-in function. Consequently, at each outgoing fiber of the second endplate signals of several input channels are superimposed, as well as the signal emerging from

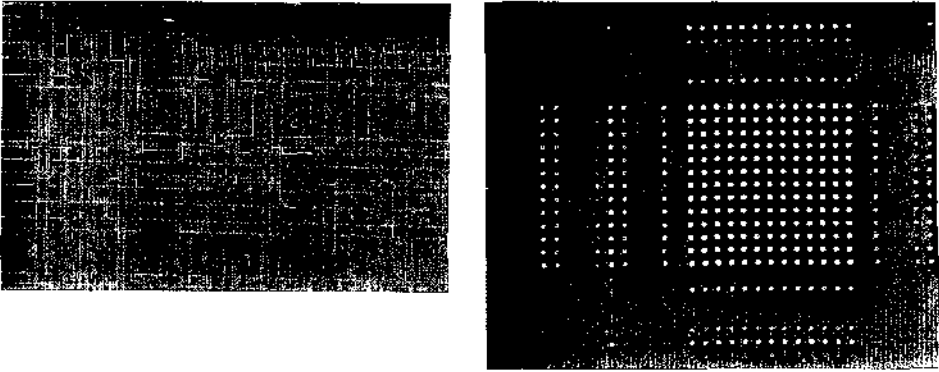


Figure 5: *Surface structure and diffraction pattern from a Dammann grating.*

each incoming fiber is distributed to several listeners. The grating implements a bus topology and serves as fan-out and fan-in element at the same time.

## 5 Optical switching

In telecommunications as well as in multiprocessors a problem of central importance is setting up reconfigurable interconnections. A switching matrix for  $N$  participants implemented as a crossbar requires  $N^2$  switches, as well as the ability of each line to drive  $N$  switches ( $\approx$  fan-out). In spite of this scaling, which makes crossbars attractive only for a moderate number of participants optical implementations have been proposed [Sto 87]. Massively parallel systems, such as a telephone exchange, require to build up complexity by combining a multitude of (simple) components. Such a switching fabric is schematically shown in fig 6.

Such multistage networks require much less switches than a crossbar, in some cases  $O(N \log N)$ , and have been widely studied [Feng 81]. Typically they consist of several stages with small crossbars (in the minimalistic case of the so called shuffle exchange network  $2 \times 2$  crossbars are used) and permutation elements implementing a random wiring in between them. These global interconnections between stages are fairly long and densely packed lines which are subject to crosstalk and therefore well suited for optical replacements. Holographic optical elements have been widely discussed for this purpose [Schw 92], fig. 7 shows the principle.

In a demonstration setup for an optoelectronic switching network at the University of Erlangen holographic permutation elements were used in conjunction with 'smart detectors'. A minimal network consisting of three subsequent stages with four parallel channels was implemented. From a PC, which served to generate the transmitted data and check the received data for bit errors, an array of lasers was driven to inject optical signals into the system. In the demonstration discrete lasers were used and by a geometrically similar array of microlenses collimated. In the near future it can be expected that monolithically integrated

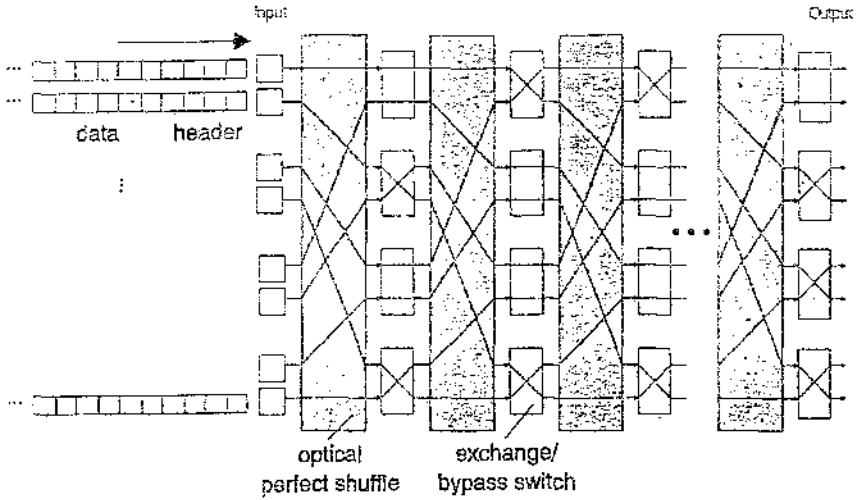


Figure 3: Multistage switching network (alternating permutation/switch stages).

arrays of microlasers become available [Cold 92]. The parallelly collimated laser beams pass a holographic permutation element realized as volume holograms in dichromated gelatine. This component implements firstly the perfect shuffle, i. e. the global interconnection pattern between stages, and secondly it changes the cross section by reducing the distances between adjacent channels. Thus the (large) pitch of the discrete laser array is matched to the ( $420\text{ }\mu\text{m}$  small) pitch of the receiver.

As a receiver an opto-ASIC, i. e. a custom designed optoelectronic CMOS integrated circuit [Zürli 92], was used as a 'smart detector'. It contains photodiodes, analog electronic amplification, digital circuits for implementing  $2 \times 2$  crossbars and line drivers for electronic output. The packet switching network was configured as a self-routing Batcher sorting network. This requires each  $2 \times 2$  crossbar to be able to extract its switch setting from the incoming data. For this purpose each node has a small finite state machine. Thus, the network does not require central control. The photodiodes on the opto-ASIC are an interesting part, because they do not belong to a standard CMOS process. Nevertheless a good responsivity of  $0.28\text{ A/W}$  can be achieved. After receiving and switching the signals are transmitted into the next optical permutation element by means of the next laser array. The complete setup is shown in fig. 8

The demonstration was running at  $1\text{ MBit/s}$  (limited by the PC-interface for data generation and measurement of bit error rate). It was possible to set arbitrary non-blocking signal paths up and to reconfigure them.

In the long term light emitters (laser diode arrays) and receivers ('smart detectors') should be mounted as near to each other as possible. Obviously, today this can be achieved by building hybrid setups. Suitable projects are in research [Bac 92].



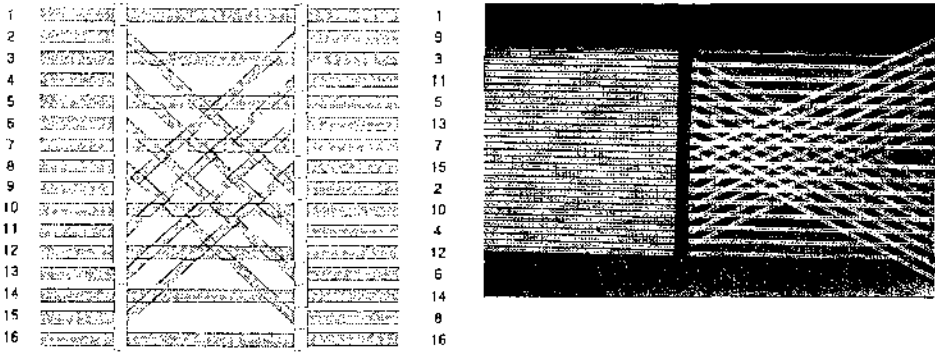


Figure 7: *Holographic permutation elements (a) principle (b) experiment with beams passing a faceted volume hologram visualized by fluorescence.*

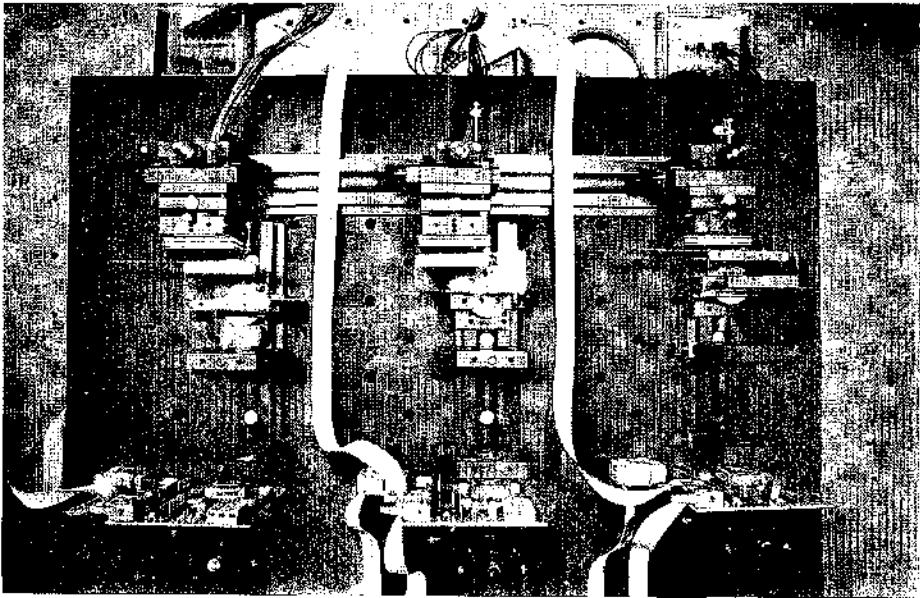


Figure 8: *Experimental setup for a selfrouting multistage shuffle-exchange network with three subsequent stages and four parallel channels.*

In the future it is hoped that light emitters, receivers and digital electronic can be integrated on the same common substrate in a monolithic fashion. Such so-called 'smart pixels' would be a major breakthrough. The philosophy of smart pixels is to have many small electronic processing 'islands' having optical I/Os with optical connections between them. Optoelectronic switches, such as the one shown above, could be implemented in a compact fashion as chip to chip interconnections. By adding functionality to the switching nodes, very

soon powerful processors can be envisioned. For example with the architecture similar to a multistage network, simply by giving each node the ability to add incoming numbers and to multiply them by a factor, a special purpose signal processor can be built. It could be used for Fourier transformation and other fast algorithms. In a similar approach (but with a different wiring diagram) other massively parallel computer architectures such as systolic arrays or cellular automata could be implemented optoelectronically by using smart pixels [Fey 92].

## 6 Conclusion

Optoelectronic interconnections are useful in short range interconnections, for example within multiprocessor systems, for several reasons:

- high bandwidth of each individual channel (in excess of 1 GBit/s)
- high parallelism and packing density (in excess of 1000 channels/cm<sup>2</sup>)
- three-dimensional topology and global interconnection patterns
- broadcasting, signal distribution and bus topology at high speed
- synchronisation because all delays are exactly known
- no crosstalk along the line (only at the optoelectronic terminals)
- isolation prevents ground loops
- high impedance devices reduce heat dissipation for communications
- hopefully: integration with electronics towards 'smart pixels'

Several examples for optoelectronic interconnections, namely an optical backplane, an optical bus and reconfigurable optoelectronic switches were presented. This survey was by no means complete but had a more exemplary character.

In the literature there is not much doubt, that optical interconnections will soon be useful within distributed systems, in multiprocessors and in telecommunications. On the other hand, all-optical computers still require some major breakthrough in optical switching devices or logic gates. As Chavel [Chav 91] puts it: *'The only reason we can see at present to talk about an optical computer is not that anyone might need it or that there is a visible advantage to it, but rather that nobody knows how much may still be expected from progress in nonlinear optics and technology; the only reason we can see not to provide optical interconnects to computers, at least at the board to board level is, that electronic has a well established interconnect technology and optics does not, so that meaningful practical issues like ruggedness, alignment tolerances have not yet been adequately worked out.'*