Halmstad University Post-Print

A fiber-optic interconnection concept for scaleable massively parallel computing

Magnus Jonsson, Kenneth Nilsson and Bertil Svensson

N.B.: When citing this work, cite the original article.

©1995 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Jonsson M, Nilsson K, Svensson B. A fiber-optic interconnection concept for scaleable massively parallel computing. In: Proceedings of the Second International Conference on Massively Parallel Processing Using Optical Interconnections : October 23 - 24, 1995, San Antonio, Texas. IEEE; 1995. p. 313-320.

DOI: <http://dx.doi.org/10.1109/MPPOI.1995.528658>

Copyright: IEEE

Post-Print available at: Halmstad University DiVA
<http://urn.kb.se/resolve?urn=urn:nbn:se:hh:diva-2766>

A Fiber-Optic Interconnection Concept for Scaleable Massively Parallel Computing

Magnus Jonsson¹, Kenneth Nilsson¹, Bertil Svensson^{1,2}

1. Centre for Computer Architecture, Halmstad University, Halmstad, Sweden

2. Department of Computer Engineering, Chalmers University of Technology, Göteborg, Sweden

email: Magnus.Jonsson@cca.hh.se, Kenneth.Nilsson@cca.hh.se, svensson@ce.chalmers.se

www: <http://www.hh.se/cca>

Abstract

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

This paper presents a fiber-optic communication concept where the bandwidth is scaleable to having multiple Gbps channels in each node incorporating WDM (wavelength division multiplexing) [8] [7] [4] and passive optical stars. The optical fiber not only supports the network with high bandwidth, it also gives a very robust communication with low bit error rates even in tough environments. Electronic stars are also consistent with the concept; these give a network which can be implemented with cheap components today or in the near future. The concept is primarily designed for embedded applications but inter-node distances of several hundred meters are no problem. By connecting several clusters by a backbone star and dual-ported nodes, Figure 1, huge modular networks can be built. In each dual-ported node there are buffers to synchronize the time-slots in the two stars involved.

1 Introduction

In many embedded applications incorporating distributed processing nodes there has to be a high bandwidth in the communication system. One example is a robot with several intelligent image sensors where extracted features or raw data are transferred in real time to SIMD-arrays (Single Instruction stream, Multiple Data streams) executing artificial neural network algorithms [17] [2] [13]. The data exchanged between the nodes can be multi-dimensional patterns and therefore the communication network must be able to carry large amounts of information (Gigabits per second in a typical system) [16]. For a system with 100 SIMD-nodes, architectures for which are developed in the ongoing REMAP project, the aggregated communication speed is about 100-1000 Gbps [3].

Figure 1: Network topology with a backbone star connecting several clusters.

The concept originates from a distributed SIMD architecture, MIMSIMD (Multiple Instruction streams for Multiple SIMD arrays), where each node can consist of thousands of processing elements. This makes the nodes require a very high communication and I/O bandwidth, but also low-bandwidth nodes can be connected to the network such as traditional digital signal processors, sensors and actuators. The bit-length of the frames, the smallest fraction of the communication messages, can with advantage be set to the same number as the number of processing elements in the SIMD-nodes.

By using Time Division Multiplexing (TDM) the bandwidth is split into time-slots that are allocated statically or dynamically. The time-slots can then be used by logical channels. In applications with partly predetermined behavior some of the scheduling scheme can be statically downloaded at system design with the possibility to change it later by connecting the development system to the target system via a gateway node, see Figure 2. A certain amount of bandwidth is thereby guaranteed to each node in the system according to the estimated requirements. Dynamic scheduling is achieved by sending the bandwidth demands of the nodes at regular intervals, and then giving the nodes their share of the bandwidth.

Figure 2: Development system connected to the network.

The scheduling scheme of each node is loaded into a time-slot memory [19] which is read in a cyclic way. Each position in the memory tells the communication unit what to do at a certain instance, if it should transmit/receive and if so, which channel that should be used. There can exist several time-slots belonging to the same logical channel. Because of the medium access method used all communication units in the system are well synchronized (described in Section 4) and the clock circuits can with advantage be accessed by the processing part of the nodes when global time is needed.

The rest of the paper is organized as follows. In the second and third section the electronic star and the passive optical star will be described, while the fourth section will focus on the medium access control and the fifth on the scalability properties. In Section 6 some typical applications are mentioned. Section 7 will then conclude the paper.

2 The Electronic Star

The electronic star in the concept, Figure 3, reminds of an ATM-switch (Asynchronous Transfer Mode) but they have nothing more in common than that they both switch messages between fiber optic point-to-point connections. However, several standard integrated circuits intended for ATM-networks may be used, both communication chips and switching chips, as long as they do not have any parts of the ATM-protocol included.

Each node in the system has two fibers, one for each direction, connecting it to the star. The transmitter and receiver modules can for example be TriQuint's Hot Rod chips [18] where bit synchronization is maintained by continuously sending information, either data or synchronization patterns, over the serial link. This can be done because of the fact that each fiber actually is a point to point connection and not a shared medium.

Figure 3: Electronic star.

To keep the scheduling complexity low it is an advantage if the switching circuit is always 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). Commercially available switch chips are listed in [6]. 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 it receives control messages and can, by the information in these messages, configure the star. This will be further described in Section 4.

3 The Passive Optical Star

Exchanging the electronic star in a cluster by a passive optical one, Figure 4, will enforce a broadcasting mechanism where all incoming messages to the star will be distributed to all nodes in the cluster by splitting the light. The passive optical star can be seen as a bus topology where, as long as only one wavelength is used, only one message can pass through the star at a time. This is in contrast with the electronic star which can handle messages from all nodes in the system at the same time. In the future, however, when WDM-components become available at reasonable prices, the passive star cluster can be very powerful with several parallel wavelength channels per node where, for full flexibility, each node in the cluster can have both tunable transmitters and receivers. Different WDM-components are described in [14].

If the number of wavelength channels is set to exactly the same as the number of nodes in the cluster, then the complexity can be kept down at the cost of the limitation in bandwidth with only one channel per node. This is

Figure 4: Passive optical star.

Figure 5: A passive optical star cluster with fixed transmitters and tunable receivers.

done by the architecture in Figure 5 where every transmitter has a unique wavelength and every receiver is tunable. Still there is full connectivity and a broadcast can be sent from any of the nodes by having all receivers of the cluster listen to the same wavelength.

The number of nodes is limited by the power that is launched by a transmitting node, the attenuation in the star and in the fibers, and by the sensitivity of the receiver. Dispersion is another limiting factor that appears in high speed fiber optic systems but is not so critical in embedded systems as in telecommunication because of the short distances.

Stars with a small number of inputs and outputs are commercially available and larger ones have been reported [20]. A 144×144 passive optical star is described in [10] and [15].

4 Medium Access Control

Here we describe the medium access method used in the network. We first describe the basic ideas, common to both implementations. Then, in Section 4.2 and 4.3, we explain details in the medium access control for the electronic star cluster and the passive optical star cluster, respectively. We then describe static and dynamic scheduling strategies.

4.1 Time Division Multiplexing

The access to the physical medium in a cluster, handled by the medium access control layer (MAC), is cyclically scheduled and based on TDM and well synchronized nodes [12]. To avoid collision there is a copy of the schedule in every node. The schedule is stored in time-slot memories where the access to the star, and hence to the network, is partitioned into time slots of equal length. At

system start-up all nodes measure their propagation delays to the star by sending messages to themselves. By this the nodes are able to know precisely when they should begin transmitting or receiving. The advantages with time-slot memories are that high bandwidth utilization can be achieved since collisions are avoided and propagation delays are accounted for, and that the network can be used by hard real-time systems that need time-deterministic communications. In [11] TDM is motivated by constant medium-access delay independent of external stimuli, and by the fact that it is time deterministic, load independent and collision-free.

The mechanism used to synchronize the nodes is based on adjustable hardware clocks. At appropriate intervals the nodes exchange their clock values [19]. Even if the accuracy of the synchronization method is very good there is always an uncertainty. To get around this problem a gap is inserted after each frame. The minimum size of the gap can be determined by:

$$t_{\text{gap}} = t_{\text{fiber}} + t_{\text{reset}} + t_{\text{synch}} + t_{\text{disp}}$$

t_{fiber} = Uncertainty in the measurement of the fiber lengths.

t_{reset} = Delays in the transmitters and receivers to reset after each frame.

t_{synch} = Uncertainty in the node synchronization.

t_{disp} = Maximum variation in propagation delay between two wavelengths due to dispersion.

4.2 The Electronic Star Cluster

In the electronic star cluster the medium access control is implemented in the following way: To avoid collision, every node, as well as the star coupler, has a copy of the schedule stored in a cyclically accessed time-slot memory where the access to the star is partitioned into time slots of equal length. The time slots can be visualized as parts

of the star coupler communication cycle as shown in Figure 6. The communication cycle of the star coupler is chosen as the nominal cycle because the star coupler is the bottleneck of the network and collisions should be avoided in it. Both in the star and in the transmitters and receivers, which have their own cycles, the cycles are implemented using cyclic time-slot memories where each position in the memory corresponds to a time-slot in which one frame can be transmitted or received. To account for the propagation delays in the fibers the transmitter and receiver communication cycles are time skewed relative to the communication cycle of the star, anti-clockwise in the transmitters and clockwise in the receivers, to transmit earlier and receive later.

Every receiver in the star has its own communication cycle and the contents of each slot in the time-slot memory tells the switch where to direct the frame. This is controlled by the control unit in the star, Figure 3. In the nodes there are two communication cycles, one for the transmitter and one for the receiver. When scheduling a transmitter to send at a specific time the time-slot for that particular instance is allocated both in the star cycle of the actual receiver in the star, in the transmitter cycle, and in the cycle of the receiver or receivers for which the message is destined.

4.3 The Passive Optical Star Cluster

To achieve more bandwidth in the future the electronic star can be exchanged by a passive optical star and WDM. Every wavelength then has its own virtual communication cycle in the star running in parallel with all other wavelengths. Transmitters and receivers, only capable of using one wavelength at the time, only need to have one cycle each, but still they have to store the information about which wavelength to use in each time slot. Therefore, in a node with no parallel wavelength channels, there are still only two communication cycles,

Figure 6: Communication cycles.

one for the transmitter and one for the receiver. The cycles have access to all wavelengths, if that particular device is tunable, but only one at a time.

When a receiver is to change to another wavelength it is important to have a fast clock recovery circuit to retain the high bandwidth utilization. One implementation of a fast clock recovery is described in [9] where the clock recovery circuit always has eight different phases of the clock to choose between. Another way to solve this problem is to have one clock recovery circuit for each transmitting node and its wavelength channel, or to have two circuits, one for the currently used channel and one to recover the clock for the next channel to be used, like a pipeline stage.

4.4 Scheduling

In some applications all communications can be scheduled off-line statically and then be downloaded to the target system, i.e. into the time-slot memories. The communication will then automatically be time-deterministic, which is a great advantage in real-time systems. However, dynamic properties of the interconnection network are often desired. Here we will discuss both static and dynamic properties of the network. For clearness the assumption is made that the receivers are capable only of receiving from one wavelength channel at the time.

In dynamic distributed real-time systems, messages may be classified into two categories [1], best effort messages and guarantee seeking messages. While best effort messages normally have soft deadlines where the system will 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 this concept some of the bandwidth is statically allocated at system design, the static part of the cycle in Figure 7. This gives all nodes a minimum number of time-slots, according to a rough estimation of the bandwidth needs, where data frames can be transmitted. The medium access control layer can then offer the higher layers, not described here, services for guarantee seeking messages. If the statically allocated bandwidth is sufficient for the message to meet its deadline, then a guarantee is stated. If not, the message will never be accepted for transmission.

In the control slots of the previous cycle a broadcast message is sent from each node to tell which node or nodes the data messages in the static part is intended for, seen as the arrow in Figure 8 pointing at the static part of the next cycle. Only one control slot is needed for each

Figure 7: The communication cycle is split into three parts: control, static, and dynamic part.

source node. Information about buffer space, dynamic allocation of time-slots (described below) and clock values for the synchronization mechanism are also contained in the control slot. Even the static time-slots are partly dynamic, and hence more flexible, because it is only the source that is statically allocated, the destination is chosen dynamically just one cycle in advance. However, the above layers are expected to implement services where the receiver also is involved in the setup of connections. The size of the static part depends on how good estimations of the communication behavior that can be done at system design or when the system is changed.

The dynamic part of the communication cycle in Figure 7 is also shared by all nodes but the shares are recalculated at each cycle. Every turn of the cycle the nodes ask for bandwidth by sending broadcast messages in the control slots. In this way each node can independently calculate the scheme for the next cycle since it has exactly the same information as every other node. The requests will take effect on the dynamic part of the next cycle, see Figure 8. Assuming that B_{dyn} is the total bandwidth of the dynamic part and N is the number of nodes in the cluster, if the total requested bandwidth exceeds the amount of bandwidth in the dynamic part of the cycle, only nodes that required less bandwidth than B_{dyn} / N get what they asked for. The rest of the bandwidth will be equally shared by the other nodes. The dynamic allocation scheme is aimed for best effort messages.

Figure 8: The control slots in cycle N refer to the data slots in cycle N+1.

This far only communication through one channel in the star at a time has been discussed. However, the time-slot allocation scheme can be extended to search for free resources after the main scheduling has been done. The free resources can in each time-slot position be in the form of: free wavelength channels in the passive optical star, free independent paths through the electronic star, transmitters that are not sending, and receivers that are not receiving in the actual time-slot position. In addition to the already mentioned free resources, there can exist time-slots in the static part that are not used in the actual cycle. These can be used to temporarily enlarge the dynamic part of the cycle. When searching for the free resources, time-slots for the whole path from the transmitter to the receiver must then be free in the same time-slot position in the cycles if it shall be possible to use the path. All cycles in the star must hence be scheduled at the same time and in the case of the electronic star the time-slots of the different channels through the switch must be started synchronously to avoid collision.

5 Scalability Properties

The fact that the medium access control is the same for electronic star clusters and passive optical star clusters makes it easy to upgrade from an active star to a passive one. By using WDM with fixed transmitters and the same number of wavelength channels as the number of nodes in the cluster the passive star will reach the same bandwidth as the electronic crossbar star but with $2N$ optoelectronic components instead of $4N$ where N is the number of nodes. The number of optoelectronic components related to the number of nodes in the cluster is shown in Figure 9. For the passive star, the number mentioned is only true when WDM-components are available as single components, e.g. tunable photo diodes or arrays of photo diodes with different wavelengths. Without this kind of components the number will be $N + N^2$, assuming fixed transmitters and tunable receivers. Thus the electronic star is the best alternative today and in the near future, while the passive star can offer a cheaper system with less components in a longer perspective. The concept is hence scalable to future technology.

Figure 9: The number of optoelectronic components, y-axis, related to the number of nodes in the cluster, x-axis.

The maximum size of a passive star cluster primarily depends on the attenuation in the star, which is proportional to the number of nodes, but also on dispersion. However, the dispersion is not so critical in short distance communication as in telecommunication and therefore the components in embedded systems have less requirements. In the case of the electronic star these parameters are even less critical and do not depend on the number of nodes, but when the size of the cluster grows too big the complexity of the star limits the size of the cluster. To take care of this limitation we introduce clustering, Figure 1, where a backbone star connects several clusters via dual-ported nodes. With passive optical stars sized 144×144 , networks with over 20 000 nodes can be built. Each of these nodes is typically a SIMD array with thousands of processing elements.

Figure 10: A network with both a passive optical star and an electronic star.

Both electronic and optical star clusters can exist in the same system, see Figure 10. The scheduling of inter-cluster communication is not described here but is ongoing work. When dividing a system into clusters the communication behavior should be estimated as far as possible to place nodes that communicate a lot with each other in the same cluster. In this way the inter-cluster communication can be kept at a low level. By clustering we achieve scalability to systems with a large number of nodes.

If all nodes in a passive star cluster are expanded to handle several wavelength channels at the same time we get higher peak bandwidth from the nodes. If we then insert more wavelength channels than nodes in the cluster we get higher total bandwidth too. In other words we can scale to higher bandwidth, both per node and overall, by using WDM.

6 Applications

The concept is primarily developed for embedded distributed real-time systems but can be used in non real-time applications too. In applications like autonomous robots where the bandwidth requirements of the

distributed nodes differ from node to node, e.g. depending on the kind of sensors that are attached to the nodes, the concept offers the possibility to allocate different amount of static bandwidth between the nodes at system design.

In some applications there are several working modes of the system, where each mode represents a static communication scheme. This could for example be a radar system with distributed SIMD-nodes where the modes are different sets of algorithms. The modes can be realized as different segments in the time-slot memories where a change of a segment pointer alters between the segments instantaneously.

Other applications that demand a high bandwidth are real-time image processing and applications using artificial neural networks. With massively parallel SIMD-nodes, multiple-Gbps communication is needed to satisfy the requirements of this kind of systems [3] [5].

7 Conclusions

In this paper a communication concept for embedded massively parallel real-time systems has been described. It meets the demands of multi-million processor-array systems organized in a modular way. A star network topology is used and the main feature of the concept is scalability which is offered in several ways. By having two kinds of stars, electronic and passive optical, it is scalable from currently available technology to future technology where WDM can be implemented with single optoelectronic components. WDM is used to make the system scale to higher bandwidth and clustering to make it scale to larger networks.

8 Acknowledgment

This work is part of the REMAP project, financed by NUTEK, the Swedish National Board for Industrial and Technical Development.

9 References

- [1] K. Arvind, K. Ramamritham, and J. A. Stankovic, "A local area network architecture for communication in distributed real-time systems," *Journal of Real-Time Systems*, vol. 3, no. 2, pp. 115-147, May 1991.
- [2] L. Bengtsson, K. Nilsson, and B. Svensson, "A processor array module for distributed, massively parallel, embedded computing," *The Euromicro Journal*, vol. 38, no. 1-5, pp. 529-537, Sept. 1993.
- [3] L. Bengtsson, K. Nilsson, and B. Svensson, "A high-performance embedded massively parallel processing system,"

IEEE and Euromicro Conference, MPC'S'94, Ischia, Italy, May 1994, pp. 201-206.

[4] D. E. A. Clarke and T. Kanada, "Broadband: the last mile," *IEEE Communications Magazine*, vol. 31, no. 3, pp. 94-100, Mar. 1993.

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

[6] L. Goldberg, "ATM switching: a brief introduction," *Electronic Design*, pp. 87-103, Dec. 16, 1994.

[7] P. E. Green and R. Ramaswami, "Direct detection lightwave systems: why pay more?," *IEEE LCS*, pp. 36-49, Nov. 1990.

[8] G. R. Hill, "Wavelength domain optical network techniques," *Proceedings of the IEEE*, vol. 77, no. 1, pp. 121-132, Jan. 1989.

[9] M. Jonsson, U. Moen, "A high performance fibre-optic time-deterministic communication network," *Master thesis, Technical Report CCA-9406, Halmstad University, Sweden*, 1994.

[10] K. Kato, K. Okamoto, H. Okazaki, Y. Ohmori, and I. Nishi, "Packaging of large-scale integrated-optic NxN star couplers," *IEEE Photonics Technology Letters*, vol. 4, no. 3, pp. 348-351, Mar. 1993.

[11] H. Kopetz, A. Damm, C. Koza, M. Mulazzani, W. Schwabl, C. Senft, and R. Zainlinger, "Distributed fault-tolerant real-time systems: the Mars approach," *IEEE Micro*, vol. 9, no. 1, pp. 25-40, Feb. 1989.

[12] K. Nilsson, B. Svensson and P. Wiberg, "A modular, massively parallel computer architecture for trainable real-time control systems," *Control Engineering Practice*, Vol. 1, No. 4, pp. 655-661, 1993.

[13] K. Nilsson, B. Svensson, and L. Bengtsson, "Massively parallel architectures for future action-oriented systems: possibilities, solutions, and implementation aspects," *MCPA'93, International Workshop on Mechatronical Computer Systems for Perception and Action*, Halmstad, Sweden, June 1-3, 1993, pp. 229-240.

[14] K. Nilsson, "Design of Time-deterministic communication in a computer system for embedded real-time applications," *Research Report CCA-9404, Halmstad University, Sweden*, June 1994.

[15] K. Okamoto, H. Okazaki, Y. Ohmori, and K. Kato, "Fabrication of large scale integrated-optic NxN star couplers," *IEEE Photonics Technology Letters*, vol. 4, no. 9, pp. 1032-1035, Sept. 1992.

[16] B. Svensson, T. Nordström, K. Nilsson, and P. Wiberg, "Towards modular, massively parallel neural computers," *SNCC-92, Swedish National Conference on Connectionism*, Skövde, Sweden, Sept. 9-10, 1992, pp. 213-226.

[17] B. Svensson and P. Wiberg, "Autonomous systems demand new computer system architectures and new development

strategies," *IECON '93: 19th Annual Conference of the IEEE Industrial Electronics Society*, Maui, Hawaii, USA, Nov. 15-19, 1993, pp. 27-31.

[18] *Data communications products data book*, TriQuint, 1992.

[19] P. Wiberg, "Change-oriented time-deterministic real-time systems," *Research Report CCA-9403, Halmstad University, Sweden*, 1994.

[20] G. Yun and M. Kavehrad, "NxN passive fiber star coupler using grating degeneration and sandwich structures," *IEEE Photonics Technology Letters*, vol. 4, no. 9, pp. 1035-1037, Sept. 1992.