Fibre-ribbon ring network with inherent support for earliest deadline first message scheduling

Carl Bergenhem and Magnus Jonsson

Bergenhem C, Jonsson M. Fibre-ribbon ring network with inherent support for earliest deadline first message scheduling. In: Proceedings: International Parallel and Distributed Processing Symposium : April 15-19, 2002, Ft. Lauderdale, Florida, USA. Los Alamitos, Calif.: IEEE; 2002. p. 157-163.

DOI: http://dx.doi.org/10.1109/IPDPS.2002.1016235

Bergenhem, C. and M. Jonsson, "Fibre-ribbon ring network with inherent support for earliest deadline first message scheduling," Proc. Workshop on Parallel and Distributed Real-Time Systems (WPDRTS'02) in conjunction with International Parallel and Distributed Processing Symposium (IPDPS'02), Fort Lauderdale, FL, USA, April 15-16, 2002.

Fibre-Ribbon Ring Network with Inherent Support for Earliest Deadline First Message Scheduling

Carl Bergenhem and Magnus Jonsson

{carl.bergenhem, magnus.jonsson}@ide.hh.se, Computers and Communications lab, Halmstad University, Halmstad, Sweden. Phone: +46-35-167100 Fax: +46-35-120348

Abstract

This paper presents a network with earliest deadline first (EDF) scheduling on a per slot basis. The network is called CCR-EDF (Control Channel based Ring network with EDF scheduling). The topology is a pipelined unidirectional fibre-ribbon ring that supports several simultaneous transmissions in non-overlapping segments and with dedicated fibres for clock and arbitration. In each slot the node that has highest priority generates the clock. The clock hand over strategy together with the scheduling feature gives the network the functionality for earliest deadline scheduling of periodic messages belonging to logical real-time connections. Logical real-time connections may be added and removed during runtime, through admission control. Guaranteed real-time communication service is supported to the user. Other services include best effort traffic and special services for parallel and distributed processing such as barrier synchronisation and global reduction.

1 Introduction

In this paper, a novel fibre-ribbon ring network and its medium access protocol is presented. The network is called CCR-EDF (Control Channel based Ring network with EDF scheduling). It is divided into three rings: a data ring, a clock ring, and a control ring. In each fibre-ribbon link, eight fibres carry data and one fibre is used to clock the data, byte for byte, see Figure 1. Together, these fibres form a data channel that carries data-packets. The access to the network is divided into fixed-sized time-slots. The tenth fibre 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 fibre, that is used to clock data also clocks each bit in the control-packets. The CCR-EDF network consists of a network architecture with the topology and a medium access protocol. Both aspects are discussed in this paper. Together they form a network that offers the interesting services to the user, such as scheduling of hard real time traffic.

In addition to local area networks, the CCR-EDF network is also suitable for parallel and distributed real-time systems. Application examples are future radar signal processing systems, distributed multimedia systems, satellite imaging and other image processing methods. A typical example is the radar signal processing system described in [1], [2]. Often, these systems are classified as real-time computer systems. In a real-time computer system, correct function depends both on the time at which a result is produced and on its accuracy [3]. In distributed real-time systems, the interconnection network is a very important part of the computer system. Often, guaranteeing real-time services is much more important in these systems than performance, e.g., average latency.

A network similar to the one presented here is described in [4], [5]. The results show that the network in [4] has a rather pessimistic worst-case schedulability bound. This makes it unsuitable for hard real time traffic, because of very low guaranteed utilisation. However, the network is suitable for best effort traffic because of the priority mechanism in the medium access protocol.

Other networks with a kind of EDF scheduling include [6] and [7]. Advantages of the class of fibre ribbon ring networks [8] (including the network presented here) 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 sacrificing too much in bandwidth utilization. Also, the separate clock and control fibers simplify the transceiver hardware implementation.

The medium access protocol provides the user with a service for sending periodic messages in logical real-time connections that have been checked for feasibility with an admission control mechanism. The messages in the logical real-time connections are then scheduled with earliest

Proceedings of the International Parallel and Distributed Processing Symposium (IPDPS'02)
1530-2075/02 $17.00 © 2002 **IEEE**

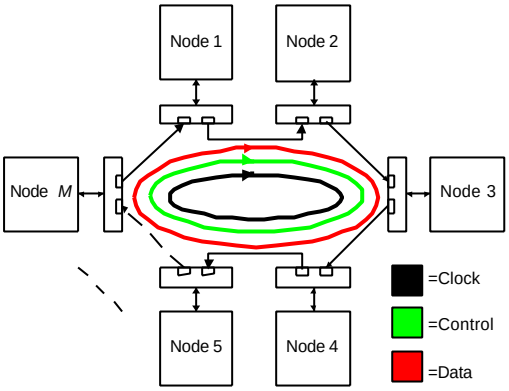

Figure 1: A Control Channel based Fiber Ribbon Pipeline Ring network.

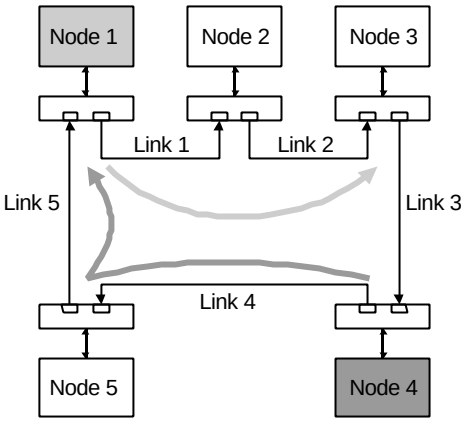

Figure 2: Example where Node 1 sends a single-destination packet to Node 3, and Node 4 sends a multicast packet to Node 5 and Node 1.

deadline first. Logical real-time connections may be added and removed from the system during runtime. The global deadline scheduling is a mechanism that is built into the medium access protocol. It allows for distributed scheduling. No further software in upper layers is required for this service. Other networks may have upper layer protocols added to them to give them better characteristics for real-time traffic, but it is difficult to achieve fine deadline granularity by using upper layer protocols.

Real-time services in the form of best effort messages, and logical real-time connections are supported for single destination, multicast and broadcast transmission by the network. There is also a service for non real time messages. The network also provides services for parallel and distributed computer systems such as short messages, barrier synchronisation, and global reduction. Support for reliable transmission service (flow control and packet acknowledgement) is also provided as an intrinsic part of the network [4].

The network with the proposed protocol is best suited for LANs and SANs (system area networks) where the number of nodes and network length is relatively small. This is important since the propagation delay adversely affects the medium access protocol. Examples of suitable applications are embedded systems (e.g., for use as an interconnection network in a radar signal processing system) and cluster computing.

The network presented in this paper has a distributed clocking strategy. During each slot, one node has the task of generating the clock signal for the entire network. This node is called the master node. The clock signal from the master node is propagated in the clock fibre around the network over N 1 hops, almost back to the source node that generated the clock signal. The end of a slot is detected implicitly by detecting that the clock signal has ended. This signals that the role of master is to be handed over to another node in the network. In the implementation of distributed clock strategy found in [9] and in [4], hand over is always to the next downstream node. The advantage of this is simplicity; the clock hand over time, between slots, is constant. The hand over time is constant since it is always the downstream node the will assume responsibility in the next slot. Therefore there will be the same gap between all slots, as long as the link length between each pair of neighbours is roughly the same. However, the simple clock hand over strategy has drawbacks concerning scheduling anomalies. It is shown in [5] that analysis of the network is complicated and that worst-case performance is pessimistic to such a degree that the worst-case analysis is of little use. The lack of good results is attributed to the unsuitable clocking strategy. The simple clocking strategy causes priority inversion because highest priority messages may be pre-empted in some situations due to clock interruption (see [5]). If the clocking node is in the path of the message, the message is unfeasible and cannot be send during that slot. This situation occurs because clock hand over is done in a round robin fashion that doesn't take into account the arrival of a highest priority message in a node. The clocking strategy proposed in this paper, does not suffer from the same pessimism.

The rest of the paper is organised as follows. Section 2 presents the network architecture. Section 3 presents the medium access protocol. Section 4 discusses some timing aspects of the network. Section 5 discusses some assumptions that are required for the scheduling framework. Section 6 presents the scheduling framework itself. Finally conclusions are presented in Section 7.

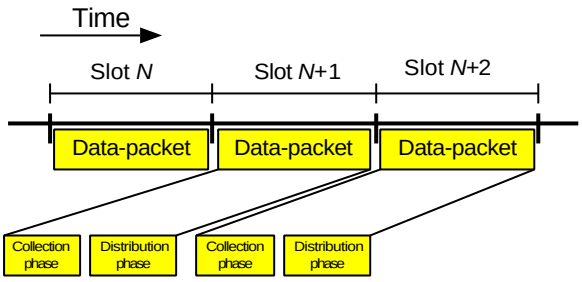

Figure 3: The two phases, collection and distribution, of the TCMA protocol. Notice that the network arbitration information, for data in slot *N+1,* is sent in the previous slot, slot *N.* Observe that the lengths of the phases, and placement in time, in the diagram are not to scale.

2 The CCR-EDF network architecture

Motorola OPTOBUS bi-directional links with ten fibres per direction are (in this paper) assumed to be used but the links are arranged in a unidirectional ring architecture where only $N / 2$ bi-directional links are needed to close a ring of N nodes. All links are assumed to be of the same length. Fibre-ribbon links offering an aggregated bit rate of several Gbits/s have reached the market [10]. The increasingly good price/performance ratio for fibre-ribbon links indicates a great success potential for the proposed type of networks.

The ring can dynamically (for each slot) be partitioned into segments to obtain a pipeline optical ring network [4]. Several transmissions can be performed simultaneously through spatial bandwidth reuse, thus achieving an aggregated throughput higher than the single-link bit rate (see Figure 2 for an example). Even simultaneous multicast transmissions are possible as long as multicast segments do not overlap. Although simultaneous transmissions are possible in the network because of spatial reuse, each node can only transmit one packet at a time and receive one packet at a time.

For the following discussion, the term master node is exchangeable with “the node with clocking responsibility etc.”. That is, the master node also clocks the network. The clocking strategy functions as follows. During a slot, the node that has the highest priority message, according to the arbitration process has the responsibility to clock the network. In the following slot, the clocking responsibility is handed over to the node that has the highest priority message in that slot. This may be another node or the same as in the previous slot. Thus clock hand over is always done in accordance with the result of the medium access arbitration process, described further in next section.

0	Nothing to send
1	Non-Real Time
2-16	Best Effort
17-31	Logical real-time connection

Table 1: The allocation of priority levels to user services. A higher priority within the traffic class implies shorter laxity and a more urgent message

The result of the arbitration process is knowledge of messages at the head of the local queues in all nodes, and therefore also knowledge about which node has the highest priority message in the entire system. The current master distributes this information to all nodes. A distribution packet is sent so that the end of the packet corresponds with the end of the slot. This implies that when the master stops the clock at the end of the slot, all nodes have the information that they need to perform clock hand over that takes place in the gap between slots. The node that has highest priority in the coming slot detects when the clock signal is stopped and assumes the master role. The highest priority node knows that it will be master because of the information received in the distributions phase packet.

Since the node that is master, also the node that has the highest priority message, has responsibility for generating the clock, then there cannot occur a situation where the node cannot send its message. This is because the node will at most send $N - 1$ hops (where N is the number of nodes) and will never have to transmit past a master, i.e. cross the clock break at the master node.

A drawback with this method is that the clock hand over time, the gap between slots, is not constant. The size of the gap between slots depends on the distance to the next master, which will vary between 1 and $N - 1$. See also Section 4, on timing analysis.

In the CCR-EDF network, access to the network is divided into slots. The minimum slot length is discussed in a later chapter. There is no concept of cycle since a cycle cannot be defined. In other cases a cycle would be e.g. when all node have been master once. However, in this network the master role is not shared equally among nodes but is given to the node with the highest priority message.

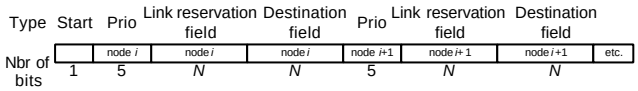

Figure 4: Contents of the TCM Collection phase packet. The figure shows that one request per node make up the complete packet. Each request is three fields, the priority field, the link reservation field and the destination field.

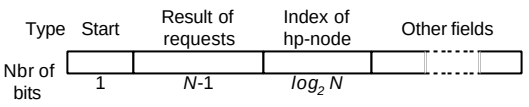

Figure 5: The modified TCMA distribution phase packet. The field labelled "index of hp-node" contains the index of the node that has the highest priority message. NB, several fields at the end of the packet have been omitted because they are not part of the discussion.

3 The CCR-EDF medium access protocol

The medium access protocol has two main tasks. The first is to decide and signal which packet(s) is to be sent during a slot. The second task is that the network must know exactly which node has the highest priority message in each slot. This is to be able to perform clock hand over to the correct node. Therefore, this information is included as an index in the distribution phase packet.

The two phases of medium access are collection phase and distribution phase (see Figure 3). As can be seen, medium access arbitration occurs in the time slot prior to the actual transmission. The protocol is time division multiplexed into slots to share access between nodes.

A disadvantage with the CC-FPR protocol presented in [9] is that a node only considers the time constraints of packets that are queued in it, and not in downstream nodes. As an example (see Figure 2), Node 1 decides that it will send and books Links 1 and 2, regardless of what Node 2 may have to send. This means that packets with very tight deadlines may miss their deadlines. The novel network presented here does not suffer from this problem.

In the collection phase, the current master initiates by generating an empty packet, with a start bit only, and transmits it on the control channel. Each node appends its own request to this packet as it passes. The master receives the complete request packet (see Figure 4) and sorts the requests according to priority (urgency). In the event priority ties the index (known by the master) of the node resolves the tie.

The network can handle three classes of traffic: logical real-time connection, best effort, and non-real time. Which class of traffic that a certain message belongs to, is signalled to the master with the priority field in the request (see Figure 4). Table 1 shows the allocation of the priority field to each user services in the network. The time until deadline (referred to as laxity) of a message is mapped, with a certain function, to be expressed within the limitation of the priority field, see Table 1. This applies to both logical real-time connection and best effort traffic. A shorter laxity of the packet implies a higher priority of the request. The result of the mapping is written to the priority field. One priority level is reserved (0 in the proposed implementation of the protocol) and used by a node to indicate that it does not have a request. If so, the node signals this to the master by using the reserved priority level and also writes zeros in the other fields of the request packet. Observe that messages that are part of logical real-time connections always have higher priority than any other service. However, a possible situation, considering spatial reuse, is that a best effort message uses the spatially reused capacity and may be transmitted simultaneously as a logical real-time connection message. The best effort message does not affect the logical real-time connection message. Observed locally in a node, best effort messages will only be requested to be sent if there is no logical real-time connection message queued. The same applies to non real-time message. They are only sent if there are no best effort and no logical real-time connection messages.

Request priority is a central mechanism of the CCR-EDF protocol. Deadlines are mapped with a function to priority. For the following discussion, a logarithmic mapping function is assumed. This mapping gives higher resolution of laxity, the closer to its deadline a packet gets. Further discussion of deadline to priority mapping function is out of the scope of this paper.

When the completed collection phase packet arrives back at the master, the requests are processed. There can only be N requests in the master, as each node gets to send one request per slot. The list of requests is sorted in the same way as the local queues. The master traverses the list, starting with the request with highest priority (closest to deadline) and then tries to fulfil as many of the N requests as possible.

The second phase, the distribution phase, is described as follows. The master sends a packet, see Figure 5, on the control channel that contains the result of the arbitration. This is either acceptance or denial of a node´s request and also which node contains the highest priority message in that slot. All nodes read the message. A request was

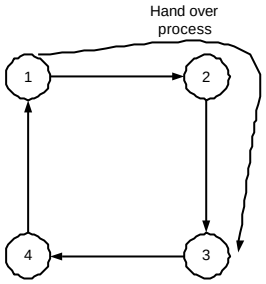

Figure 6: In the current slot, node 1 is master. The arbitration process is run during slot *i* - *1* and discovers that node 3 has highest priority and will be master during slot *i*.

granted if the corresponding bit in the "request result field" of the distribution phase packet contains a "1". The protocol also has a feature to permit several non-overlapping transmissions, that is grant several requests, during one slot. This is function is called spatial reuse and is used during run-time, but is not used in the analysis (see Section 5). Observe the distribution phase packet also may contain other information such as acknowledgement for transmission etc. These are further described in [11] and will not be part of the discussion here.

The addition of an index that points to the node that has highest priority in the current coming slot, see Figure 4, enables all nodes to know who will have the highest priority message in the coming slot and that that node will assume the role as master and clock the network. The index field needs to be

$\log_2 N$ bits wide to represent numbers up to N.

When all nodes have received the distribution phase packet, and hand over has taken place, the new slot commences and data may begin to flow in the data channel.

4 Timing properties

The worst-case hand over time is when hand over is to the node that is $N - 1$ hops or segments away. This corresponds to hand over to the upstream neighbour node. The general equation for hand over time is:

$$t_{handover} = P \times L \times D \qquad (1)$$

where P is the propagation delay of the light per meter in the fibre, L is the average length (in meters) of the segments and D is the number of segments that are traversed. In the worst-case, D is equal to $N - 1$.

An example of the clock hand over process is given in Figure 6 and 7. In Figure 6, Node 1 had the highest priority during slot $i - 1$. The network arbitration process discovers that node 3 will take over the role as master.

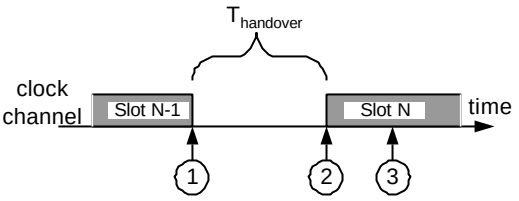

Figure 7: Shows some interesting points along the timeline of the hand over procedure. The numbers are for reference. Observe that the diagram is not to scale!

Figure 7 points out some interesting points in the timing of the hand over process. At point 1, the entire distribution packet has been sent. Node 1 stops clocking after one bit time more. At point 2, Node 3 receives the distribution packet and discovers that it will be master in the next slot. It senses that the clock has stopped after one more bit time. This is the trigger for the end of the slot and the node assumes the role as master. The node then starts clocking again with only one bit time gap with no clock. At point 3 Node 4 receives distribution the packet and discovers that it will not be master. Node four receives the clock signal again after one bit time after the distribution phase packet. Node four may then transmit if it was granted permission to send.

The general equation for minimum slot length, because the collections phase must be finished before the end of the data transmission in one slot, is:

$$t_{minslot} = N \times t_{nodes} + t_{prop} \qquad (2)$$

where N is the number of nodes in the network, t_{nodes} is the delay experienced by the control packet (during the collection phase) through each node and t_{prop} is the propagation delay through the whole ring.

5 Assumptions for the scheduling framework

The network provides a feature for spatial reuse that is used during run-time. However, the benefits of the feature is not taken into account in the analysis which assumes that only one message may be transmitted per slot. The reason for this is that the degree of spatial reuse in a specific slot can only be know statistically. One message per slot can always be guaranteed. During run-time spatial reuse can be used and always results in positive effects.

For scheduling it is assumed that the smallest time unit is a slot. An assumption concerning messages is that the relative deadline is equal to the period of the logical real-time connection. The maximum delay that a message may encounter is

$$t_{maxdelay} = t_{deadline} + t_{latency} \qquad (3)$$

where $t_{deadline}$ is the deadline of the message, and $t_{latency}$ is the worst-case latency that the message may experience. The worst-case latency is:

$$t_{latency} = 2 \times t_{slot} + t_{handover\_max} \quad (4)$$

The time slot delay assumes that one slot is just missed and one slot is needed for arbitration, while the hand over time assumes worst-case delay for the master to hand over to the correct node. On the user level, the deadline is $t_{maxdelay}$, i.e., the deadline of the message is used for scheduling, but the user perceives $t_{maxdelay}$ Thus the scheduling is not affected by $t_{latency}$

6 The scheduling framework

With the following discussion a test to find out if a new logical real-time connection can be accepted and guaranteed, will be presented. After the logical real-time connections are accepted their messages will be scheduled by earliest deadline first.

The basic EDF feasibility test is as follows:

$$\sum_i \frac{e_i}{P_i} \quad U_{\max} \quad (5)$$

where e_i is the size (number of slots) required for a message in logical real-time connection *i* and P_i is the period of messages in logical real-time connection i. Their quotient is the utilisation of logical real-time connection *i*. $U_{\max}$ is the worst-case maximum utilisation, i.e., the lowest utilisation that can be gained at full load, due to always experiencing a worst-case hand over delay between the slots (see Section 4). Since message parameters do not affect this, it is also considered to be the worst-case throughput at full load. However, actual throughput depends on the number of messages in the network. The worst-case is:

$$U_{\max} = \frac{t_{slot}}{t_{slot} + t_{\max\ handover}} \quad (6)$$

t_{slot} is the slot size and $t_{slot} + t_{\max\ handover}$ is the worst-case time for clock hand over, see also Section 4. Since there is a gap in between the slots, which cannot be used effectively, and considering the restriction of no spatial reuse, the quotient $U_{\max}$ is lower than one. The total bandwidth can be seen as the length of slots together with the gap. The effective bandwidth can be seen as the duration of the slots.

Below, an online schedulability analysis of periodic messages, referred to as logical real-time connections, will be discussed. The basic function is dynamic schedulability testing or online centralised admission control. The assumption for the analysis is that logical real-time connections arrive one at a time at any time even during run time. Another assumption is that logical real-time connections are well behaved, i.e., agreed parameters are always honoured by the transmitting node. A specific node in the system is designated to solely handle new logical real-time connections added to the system and to remove them when required. Communication with this node is handled with the best effort traffic user service.

The set ***Ma*** contains the logical real-time connections that have been tested for feasibility and are accepted. The admission test is as follows. If the utilisation of the logical real-time connections in ***Ma*** together with the new connection is below $U_{\max}$ (see Equation 6) then the new logical real-time connection is admitted into ***Ma***. The node that the new logical real-time connection belongs to is notified. The logical real-time connection can now be activated and may be used for traffic. If the utilisation of the new connection and Ma is higher than $U_{\max}$ then the new logical real-time connection is rejected. The requesting node is notified of the failure. The above procedure is repeated for every request for a logical real-time connection that arrives at the node responsible for admissions control.

7 Conclusions

This paper presents a novel network suitable for hard real-time traffic. A pipelined optical ring network forms the network architecture. It has a distributed clocking strategy that makes it suitable for global deadline scheduling. The node that has the highest priority message in each slot is also handed the role as master, which includes responsibility for clocking. Because of this, the highest priority message from any node, in the system, can always be sent to any destination. This forms the basis for the scheduling framework. Clock hand over is done in accordance with the result from the medium access protocol. The medium access protocol has two basic functions, arbitration of access to the network and deciding which node to take over the clock role. Each node in the network sends a request for transmission of its locally highest priority message to the current master. The result is a list of requests for transmission of messages, one from each node in the network.

The user services include best effort messages, non real time messages, logical real-time connections, group communication such as barrier synchronisation and global reduction, and services for reliable transmission. Logical

real-time connections are realised by admission control and earliest deadline first scheduling of messages.

The network with the presented protocol is suitable for applications with demands for real-time performance, such as in embedded systems, e.g., for use as interconnection network in a radar signal processing system, or as a high performance network for use in a LAN environment. Also worth mentioning is that the network can be built today using fibre-optic off-the-shelf components

8 Future work

Important questions that are not answered in this paper are among others, fault tolerance in case of node failure. The current study also assumes that the "token" is never lost. In a real implementation, using a time out and a designated node that always will start could solve this. This paper also lacks performance evaluation and simulation results of the network and protocol. This will be presented in a future paper. This study will also give "hard numbers" on e.g. hand over time and actual figures of utilisation.

Acknowledgement

This work is part of M-NET, a project financed by SSF (Swedish Foundation for Strategic Research) through ARTES (A Real-Time network of graduate Education in Sweden).

References

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

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

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

[4] C. Bergenhem, M. Jonsson, and J. Olsson, "Fibre-ribbon Pipeline Ring Network with Distributed Global Deadline Scheduling and Deterministic User Services," *Proc. of the Workshop on Optical Networks held in conjunction with International Conference on Parallel Processing 2001*, Valencia, Spain, Sept. 3-7, 2001.

[5] C. Bergenhem and M. Jonsson, " Analysis problems in a spatial reuse ring network with a simple clocking strategy," *Technical report IDE 0138, Halmstad University*, Halmstad Sweden, Nov. 15, 2001.

[6] K.K. Zuberi and K.G. Shin, "Design and implementation of efficient message scheduling for controller area network," *IEEE Transactions on compiters*, vol. 49, no. 2, pp. 182-188, Feb. 2000

[7] K. G. Shin, "Real-Time Communications in a Computer-Controlled Workcell, "*IEEE Transactions on robotics and automation*, vol. 7, no. 1, pp. 105-113, Feb. 1991

[8] M. Jonsson, and C. Bergenhem, "A Class of Fiber-Ribbon Pipeline Ring Networks for Parallel and Distributed Computing Systems", Proc. of CSREA International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA-2001), Las Vegas, NV, USA, Jun. 25-28, 2001, pp. 1869-1875.

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

[10] D. Bursky, "Parallel optical links move data at 3 Gbits/s," *Electronic Design*, vol. 42, no. 24, pp. 79-82, Nov. 21, 1994.

[11] Jonsson, M., C. Bergenhem, and J. Olsson, "Fibre-ribbon ring network with services for parallel processing and distributed real-time systems," *Proc. ISCA 12th International Conference on Parallel and Distributed Computing Systems (PDCS-99)*, Fort Lauderdale, FL, USA, Aug. 18-20, 1999, pp. 94-101.