Proc. 2nd Workshop on Optics and Computer Science (WOCS), Geneva, Switzerland, Apr. 1, 1997.

On Inter-Cluster Communication in A Time-Deterministic WDM Star Network

Magnus Jonsson[1] and Bertil Svensson[1, 2]

1. Centre for Computer Systems Architecture, Halmstad University, Halmstad, Sweden
2. Department of Computer Engineering, Chalmers University of Technology, Göteborg, Sweden

email: Magnus.Jonsson@cca.hh.se, svensson@ce.chalmers.se www: http://www.hh.se/cca

Abstract

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

1 Introduction

High-performance interconnection networks can be foreseen to have a central role in future distributed real-time systems. If a number of computation modules, each of them parallel, are used to obtain a massively parallel system, a modular interconnection network, able to carry a huge amount of data, is needed. Other key features of the network are time-deterministic latency and guarantees to meet deadlines. A typical system is the radar signal processing system described in [1] [2], where each module consists of a SIMD (Single Instruction stream Multiple Data streams) computer and a network interface. In this way, a MIMSIMD (Multiple Instruction Streams for Multiple SIMD arrays) computer system is formed. Other applications where the MIMSIMD architecture with a high-performance interconnection network might be required are described in [3] [4].

In this paper we present a modular network architecture with time-deterministic packet switched communication, both for intra-cluster communication and inter-cluster communication. The network supports both best-effort messages and guarantee-seeking messages [5]. The topology used is backbone-connected all-optical star clusters, i.e., star-of-stars, where electronic gateway-nodes separate the clusters (Figure 1). A cluster consists of a passive optical star where the WDM (Wavelength Division Multiplexing) technique is used to achieve multiple Gb/s channels [6] [7] [8]. By the use of electronic gateway nodes we retain the popular WDM star network architecture in each cluster, for which cheap components can be spawn to appear in the future. With electronic gateway nodes we also achieve wavelength reuse in each cluster. Other hierarchical WDM star networks include the wavelength-flat (all nodes share the same wavelength

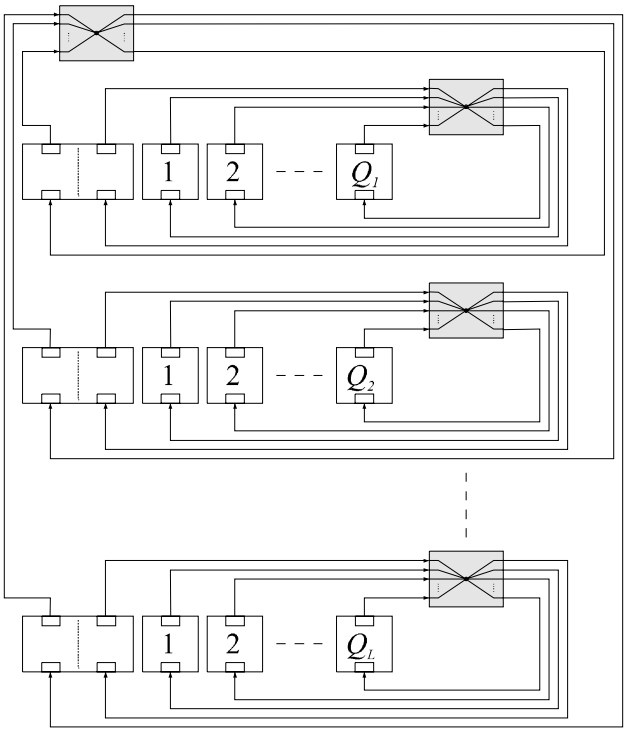

Figure 1: Multiple passive optical stars topology.

space) tree-of-stars network [9], the tree-of-stars network (called LIGHTNING) that has wavelength routing elements between each level [10], and the multiple star network where each node is directly connected to both a local star and a remote star [11].

In each cluster, fixed-wavelength transmitters and tunable receivers are used (Figure 2). A fixed unit is always tuned to the same wavelength channel, while a tunable unit can be tuned to an arbitrary wavelength channel. Components for WDMA networks are reviewed in [12]. Each transmitter has a specific wavelength and the network architecture can be described as FT^1-TR^1 using the classification scheme given in [13]. FT^1-TR^1 stands for one Fixed Transmitter and one Tunable Receiver per node. Other FT-TR networks are described in [14] [15], and general information on WDM star networks in [13] [16] [17]. The receivers are tunable over the whole range of channels used in the cluster. This makes the cluster a single-hop network where any receiver can be reached by any transmitter in a single hop [13] [18]. The main reason why FT^1-TR^1 is chosen is the naturally embedded broadcast function obtained when all receivers are tuned to the same transmitter channel.

A 100-channel WDM system has been demonstrated [19] and systems with 1000 channels are possible [20]. Although, the practical limit in number of wavelengths in the kind of networks reported in this paper is expected to be somewhere between 16 and 32 [21]. This translates into star-of-stars networks of maximum sizes between 256 and 1024 nodes, gateway-nodes included.

The same MAC (Medium Access Control) protocol, TD-TWDMA (Time-Deterministic Time and Wavelength Division Multiple Access) [1], is used in every cluster and in the backbone. We will show how the MAC protocol and the network architecture are configured to increase system performance and to get time-deterministic inter-cluster communication. In [1], intra-cluster communication was analyzed in detail, while only rough assumptions were made for inter-cluster communication. Here, we analyze how latency and node bandwidth for inter-cluster communication vary with design parameters of the network. We also present a new way of clock synchronization to reduce the worst-case latency in this kind of multi-cluster networks using TDM (Time Division Multiplexing).

The rest of the paper is organized as follows. In the second section the network concept is described. In Section 3 intra-cluster communication and TD-TWDMA are presented. Inter-cluster communication is described in Section 4, in which clock-synchronization is also discussed. Also, scalability issues are analyzed, and it is shown how low latency can be achieved for systems with high bandwidth in the backbone. Section 5 is a conclusion and summary.

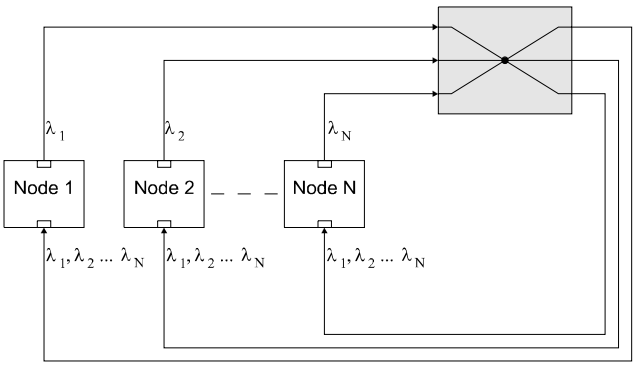

Figure 2: Passive optical star cluster with fixed transmitters and tunable receivers.

2 Network concept

In this section we will give an overview of the network. Each cluster uses WDM to get a home-channel for every node. In addition to WDM, TDM is used, so the access to each receiver is divided into cycles of equal length. Each cycle is then further divided into a fixed number of slots. Because every transmitter has its own home-channel the only conflict that can appear is when two or more nodes want to transmit to the same node at the same instant. To prevent conflicts, slots in the TD-TWDMA network are allocated in the receiver cycles, by a distributed slot-allocation algorithm, so that each slot has a specific owner. There is also one cycle running in each transmitter, but only to tell when and to whom the node is allowed to transmit. The transmitter cycle reflects the slots that the node owns in the receiver cycle of every other node, as seen in Figure 3. In the receiver cycles, shown in the upper table in the figure, each slot in each receiver cycle can only be assigned to one transmitter at the same time. The lower table shows how the slots of a transmitter cycle are built up by copying its entries in the corresponding slots in each receiver cycle. It is then up to the transmitter to decide, in each slot, if it should transmit to the node(s) it has access to or not.

The main function of the TD-TWDMA protocol is to allocate the time-slots to appropriate nodes based on what slot demands the nodes have. These slot demands are transmitted in advance on the same channels as the data. WDM networks with this type of protocol not using a separate control channel are denoted as *non-control channel based* networks. Networks where a separate control channel is used to reserve access to the data channels are denoted as *control channel based* networks. Other non-control channel based networks are found in [15] [22] [23], while control channel based networks are found in [24] [25-27].

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

⇓

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

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

In dynamic distributed real-time systems, messages may be classified into two categories [28]: best-effort messages and guarantee-seeking messages. While best-effort messages normally have soft deadlines, such that the system need only try its best to meet the deadlines, guarantee-seeking messages have harder timing constraints. If the communication system cannot guarantee the timing constraints of a guarantee-seeking message, the owner of the message should be aware of it immediately.

When using the TD-TWDMA protocol, each node has a number of guaranteed slots so that the medium access control layer can offer the higher layers services for guarantee-seeking messages. If the guaranteed bandwidth is sufficient for the message to meet its deadline, then a guarantee is given. If not, the message will be rejected immediately so that the owner will have time to handle the situation.

If there is no guaranteed messages in a node the slots will be released for best effort messages from other nodes (or the same node) according to a predetermined scheme. In this way a more *efficient bandwidth utilization* is achieved.

Each node can also reserve a number of slots in the receiver cycle of any other node, in order to increase the deterministic bandwidth for guaranteeing real-time services. The assignment of reserved slots can be changed during run-time either by higher layers, by a development system, or by having slot-assignment schemes for several working modes stored in the nodes. The change from one mode to another can then take place by just changing an offset pointer. Also reserved slots are released if not needed for guaranteed messages.

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

3 Intra-cluster communication

In this section, we describe intra-cluster communication. The notation used when describing the network is found in Table 1.

The passive optical star, in one cluster, implements a fiber-optic multi-access network [17] with M nodes. All incoming messages to the star are distributed to all nodes in the cluster by splitting the light. The FT^1-TR^1 transceiver-configuration is used and each transmitter in the cluster is assigned a unique wavelength. Hence, the number, C, of wavelength channels is equal to the number, M, of nodes. The transmitter and receiver parts of the transceiver are totally independent and can work concurrently.

Each cycle consists of S slots where s_i, $1 \le i \le S$, denotes a slot. If γ is the slot length, we use $S\gamma$ to denote the cycle time. Figure 4 shows how a receiver cycle is

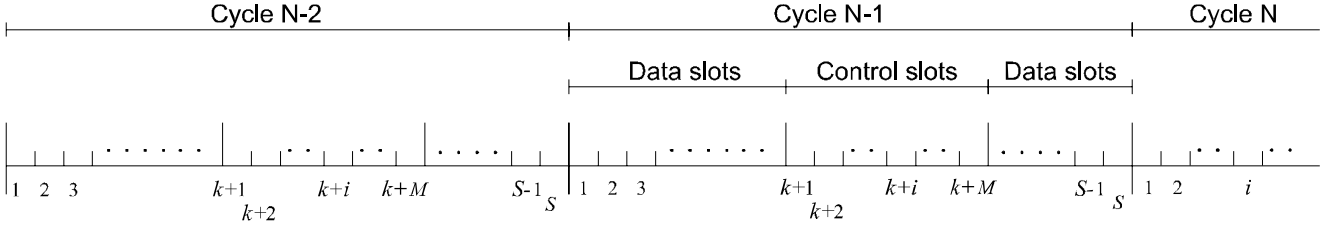

Figure 4: A cycle is partitioned into data slots and control slots, where the control slots are transmitted one cycle in advance related to the data slots that they are carrying information about.

γ:	Slot length, including gap
μ:	Computing time of the distributed slot-allocation algorithm
S:	Total number of slots in a cycle
s_i:	The i:th slot in a cycle
v_{ij}:	Index of the high-priority owner (transmitter) of slot i in the receiver cycle in node j
w_{ij}:	Index of the low-priority owner (transmitter) of slot i in the receiver cycle in node j
τ:	Latency: delay from the moment a message arrives at the transmitter buffer, in the source node, until the moment the transmission begins from the last gateway-node, excluding propagation delay
Intra-cluster communication:	
M:	Number of nodes in the cluster
C:	Number of channels (wavelengths) in the cluster
m_i:	Node i
Inter-cluster communication:	
M_j:	Number of nodes in cluster j, including all transceiver modules, on the cluster side, in the gateway node
Q_j:	Number of ordinary nodes in cluster j, i.e., number of physical end-nodes
C_j:	Number of channels (wavelengths) in cluster j
L:	Number of clusters, which is the same as the number of gateway nodes
B:	Effective bandwidth of each channel in the clusters
E:	Effective bandwidth of each channel in the backbone
R:	Ratio between the channel bandwidth in the backbone and the channel bandwidth in the clusters

Table 1: Notation when describing the network architecture.

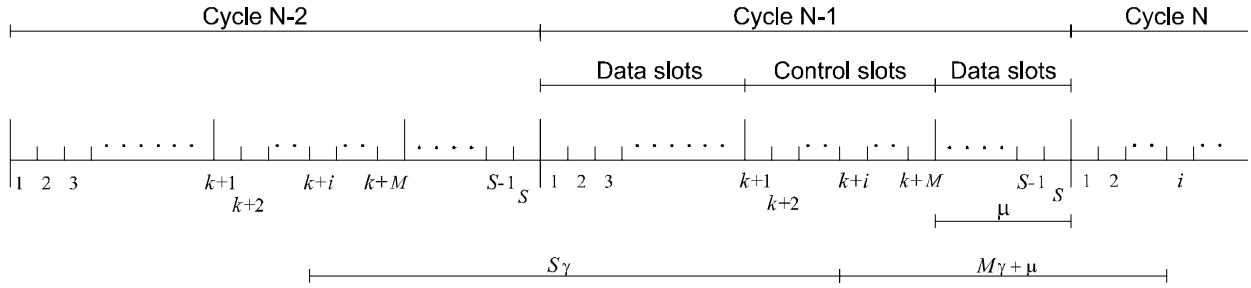

Figure 5: Latency calculation.

partitioned into data slots and control slots. Each node m_i, $1 \le i \le M$, is assigned one of the M control slots in which it broadcasts control information to all other nodes m_j, $1 \le j \le M$ and $j \ne i$. The control slots are therefore identically assigned in every node's receiver cycle. Control information that is sent in a control slot informs the other nodes about the slot demands for the next cycle. These slot demands contain information on which guaranteed slots (including reserved slots) to keep and which to temporarily release in the next cycle. When control slots from all the nodes has been gathered, allocation of the data slots in the next cycle can be calculated using the distributed slot-allocation algorithm described below.

The time, μ, it takes to run the slot-allocation algorithm sets the limit on how late in the cycle the control slots can be placed. By fine-grain interleaving of the slots, where node m_i, $1 \le i \le M$ is assured to have slot s_i as its first slot in the cycle, we can minimize the delay from a node's control slot until its first data slot in the next cycle. As shown in Figure 5, the worst-case latency for node m_i, $1 \le i \le M$, will be

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

where the first term in the middle expression is the worst-case delay before the node's own control slot appears, i.e. one cycle as shown in the figure.

The slot-allocation algorithm is based on a predetermined allocation scheme that can be partially overloaded, using the reservation method mentioned above. As an example, the allocation scheme for a four-node system is shown in Figure 6. We do not call this scheme "reservation" because that term is used when describing the overloading of the scheme. The total number of slots in a cycle is set to $S = M^2$ (in this case 16) and the number of data slots is $M(M - 1)$, i.e., 12. In the

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

Figure 6: Allocation scheme for the receiver cycles in a four-node system.

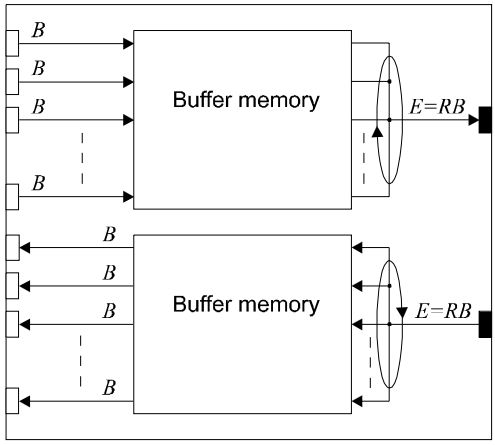

Figure 7: Gateway node with higher channel-bandwidth on the backbone side (right side).

figure, only the data slots are shown, and even if the control slots can be in the middle of the cycle we here assume they have index 13 to 16. Each pair of rows represents one receiver cycle, where each number is the index of the transmitter that owns the corresponding slot. The high-priority row is the default scheme, but if the high-priority slot owner does not need the slot it is temporarily released as described above. The low-priority owner will then get the slot. If neither the high-priority nor the low-priority owner needs the slot, it will be unused. This is the cost of having a simple algorithm.

In the predetermined allocation scheme, v_{ij} identifies the index of the high-priority owner of slot i in the receiver cycle in node j. In the same way, w_{ij} identifies the index of the low-priority owner. High-priority slots are coordinated to allow broadcast and the owners are determined by

$$v_{ij} = (i - 1) \bmod M + 1 \qquad (2)$$

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

$$w_{ij} = ((i - 1) \operatorname{div} M + j) \bmod M + 1 \qquad (3)$$

A maximum of $M(M - 2)$ slots, slot 5 to 12 in the example of Figure 6, are allowed to be reserved. The first M slots are not allowed to be reserved, and the last M slots are control slots. When reserving slots, the corresponding high-priority entry in the receiver cycle, or cycles if multicasting is used, is exchanged with the index of the reserving node.

Since each node can independently perform the computations of the slot-allocation scheme, we call it a distributed algorithm. With this simple algorithm, the updating of the table entries can start as soon as the first control slot is received. Only table indexing is then used in the data slots, in transmitters to choose between buffered messages, and in receivers to choose channel tuning. Hence, a calculation time equal to one slot-length, $\mu = \gamma$, can be assumed. Furthermore, since the number of slots in one cycle is $S = M^2$, Equation 1, describing the worst-case latency, can be rewritten as

$$\tau_{max} = \gamma (M^2 + M + 1) \qquad (4)$$

4 Inter-cluster communication

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

A network consists of L clusters. Each cluster has M_i, $1 \le i \le L$, nodes where one of the nodes is a gateway node. A gateway node contains network interfaces, to both the backbone star and its dedicated cluster, and buffer memories for both upward and downward traffic. The size of the buffer memories is significantly larger than the possible traffic in one cycle. Status information on the buffers is always sent together with the other information in the control slots.

In some systems, it is desirable that the bandwidth per channel in the backbone be higher than in the clusters. The increase in bandwidth may be implemented either by a higher bit rate or by having several wavelengths per channel. If R is the ratio of backbone channel-bandwidth, E, to cluster channel-bandwidth, B (i.e. $R = E / B$), then the gateway nodes are designed to each have R transceiver modules on the cluster side in order to achieve the same aggregated bandwidth as on the backbone side (Figure 7). Also, a gateway node has R dedicated home-channels on the cluster side, one for each transmitter. Since the transceiver modules in a gateway node are seen as

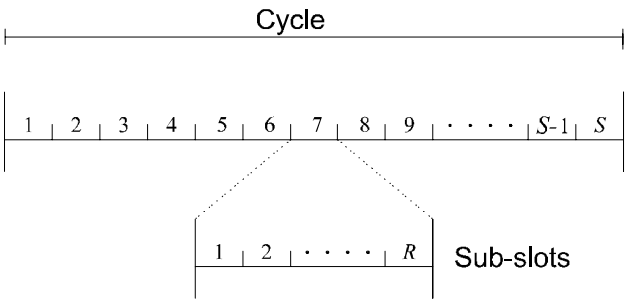

Figure 8: Each backbone slot is divided into R sub-slots with the same pair of gateway nodes as source and destination.

connections to separate nodes by the MAC protocol, the number of nodes is still defined as $M_i = C_i$, $1 \le i \le L$. In this way, we have $Q_i = M_i - R$ ordinary end-nodes in each cluster.

The cycle length in the backbone is always the same as that in the clusters, both when measured in time and when measured in number of slots. Therefore, the number of bits in a backbone slot is R times higher than in a cluster slot. Each backbone slot is divided into R sub-slots all with the same pair of gateway nodes as source and destination (Figure 8). However, the sub-slots can have different pairs of end-nodes. A sub-slot has the same number of bits as a cluster slot, which makes the design of the gateway nodes easier. Also, the latency may decrease when using sub-slots if several slots in the same source cluster and with the same destination cluster can be packed together.

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

$$M_{total} = \sum_{i=1}^{L} M_i = L^2 \qquad (5)$$

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

When *full slot-reservation* by other nodes is assumed, the number of slots that the source node has access to in the receiver cycles in the destination node and in the gateway nodes of the source and destination clusters reaches its minimum. The first transit is from the source end-node, and to the gateway node of the source cluster. Since the source node has direct access to its own cluster, the worst-case latency, τ_1, for the first transit can be obtained by using Equation 4 for intra-cluster communication, with $M = L$. I.e.:

$$\tau_1 = \gamma (L^2 + L + 1) \qquad (6)$$

The second transit is between the gateway nodes of the source cluster and the destination cluster, through the backbone. Here, slots from all $Q = L - R$ ordinary nodes in the source cluster must, in the worst-case, be multiplexed over several high-priority backbone-slots. At full slot-reservation, a node only has one high-priority slot per cycle to transmit in. Therefore, $L - R - 1$ extra cycles, in addition to the normal intra-cluster latency (Equation 4) are needed. The source gateway-node is responsible for carrying out this multiplexing, using the round-robin scheduling strategy. Hence, the source end-node is guaranteed its part of the bandwidth. The worst-case latency for the second transit is

$$\tau_2 = \gamma (L^2(L - R) + L + 1). \qquad (7)$$

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

$$\tau_3 = \gamma \left(Trunc\left(\frac{(L-R)(L-1)}{R} + 1 \right) L^2 + L + 1 \right) \qquad (8)$$

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

$$\tau_{max} = \tau_1 + \tau_2 + \tau_3 = \gamma \left(Trunc\left(\frac{(L-R)(L-1)}{R} + 2 + L - R \right) L^2 + 3L + 3 \right) \qquad (9)$$

and is plotted in Figure 9, in which the horizontal axis represents the total number, $L(L - R)$, of ordinary nodes in the network and each curve represents a specific value of R. To give an example of a real system the slot length is assumed to be $\gamma = 1.0$ μs in all latency plots. Figure 9 indicates how the latency decreases with R.

When *no slot-reservation* in any receiver cycle is assumed, a node has $L - 1$ high-priority slots to transmit in. The worst-case latency, τ_1, for the first transit is the same as when full reservation was considered (Equation 6). The

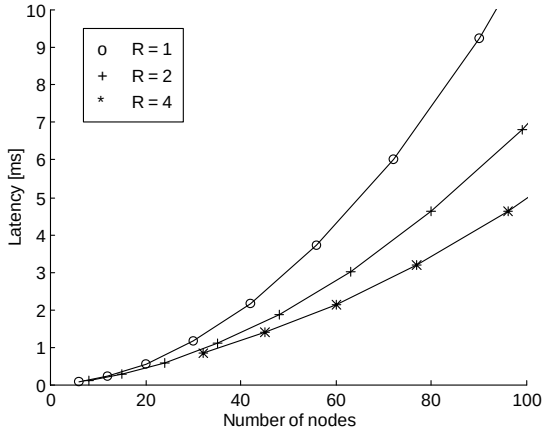

Figure 9: Worst case latency when full slot-reservation, by other nodes, is assumed. The x-axis represents the number of ordinary nodes and the slot length is assumed to be γ= 1.0 μs.

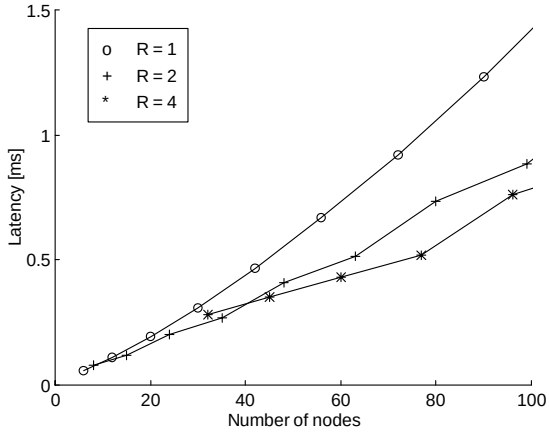

Figure 10: Worst-case latency when no slot-reservation is assumed. The x-axis represents the number of ordinary nodes and the slot length is assumed to be γ= 1.0 μs.

latency of the rest of the transit will, however, decrease compared to the full-reservation latency. The second-transit latency is $\tau_2 = \tau_1$, because the high-priority slots in the transmitter of the gateway-node, on the backbone side, will always be enough to multiplex one slot from each cluster-node in one cycle:

$$L - 1 \geq L - R \tag{10}$$

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

$$\tau_3 = \gamma\left(Trunc\left(\frac{L}{R}\right)L^2 + L + 1\right) \tag{11}$$

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

$$\tau_{max} = \tau_1 + \tau_2 + \tau_3 = \gamma\left(Trunc\left(\frac{L}{R} + 2\right)L^2 + 3L + 3\right) \tag{12}$$

Figure 10 shows how the worst-case latency, when no slot-reservation is used, varies with the total number of ordinary nodes. The latency is significantly lower compared to the case of full-reservation. This effect is related to the higher number of slots in the gateway nodes that can be used for multiplexing of incoming messages.

We define the deterministic bandwidth per node as the minimum high-priority bandwidth when not having any reserved slots. In this analysis we express the bandwidth as a ratio to the full channel-bandwidth, *B*. Measured as the number of high-priority slots per total number of slots in a cycle, this so-called bandwidth is shown in Table 2 for the different cases. The last transit is the bottleneck and the node bandwidth for it, as a function of the number of ordinary nodes, is plotted in Figure 11, for full slot-reservation. As seen in the figure, the deterministic node-bandwidth is a rather low part of the full bandwidth but all of the bandwidth can be used for broadcasting. Nodes requiring more deterministic bandwidth can use slot-reservation. The deterministic node-bandwidth, when no slot-reservation is assumed, is plotted in Figure 12 and is higher than the former.

The real-time services offered by the MAC-layer rely on the deterministic latency and bandwidth. A guarantee can be stated if the known minimum number of high-priority slots, along the whole path through the network, is enough to transfer the message in time. In the calculation of this justification, the deterministic latency, the deterministic bandwidth, and the deadline of the message are used. A slot of a guaranteed message is tagged to

Full reservation	**No reservation**
From source end-node to gateway node:	
$\frac{1}{L^2}$	$\frac{L-1}{L^2}$
Through backbone:	
$\frac{R}{L^2(L-R)}$	$\frac{R(L-1)}{L^2(L-R)}$
From gateway node to destination end-node:	
$\frac{R}{L^2(L-R)(L-1)}$	$\frac{R}{L^2(L-R)}$

Table 2: Node bandwidth in number of high-priority slots per total number of slots in a cycle.

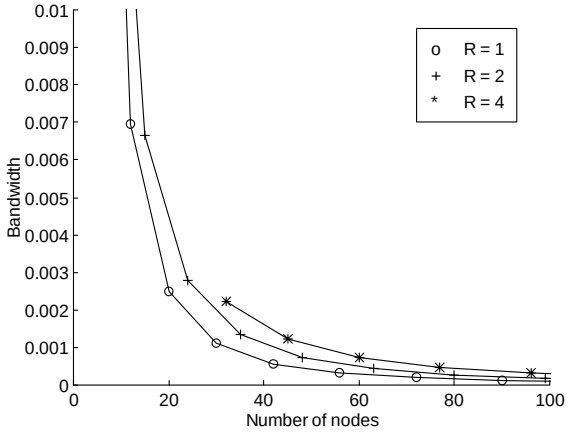

Figure 11: Node bandwidth, when full slot-reservation by other nodes is assumed, in number of high-priority slots per total number of slots in a cycle. The x-axis represents the number of ordinary nodes.

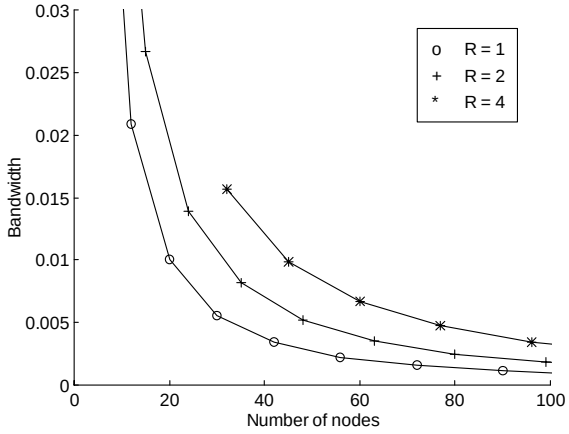

Figure 12: Node bandwidth, when no slot-reservation is assumed, in number of high-priority slots per total number of slots in a cycle. The x-axis represents the number of ordinary nodes.

indicate its priority over other slots. The gateway nodes then always transmit the tagged messages before buffered best-effort messages.

4.1 Clock synchronization aspects

The nodes are synchronized to account for the propagation delay between transmitting and receiving nodes, i.e., receivers are synchronized to transmitters, proportional to the propagation delays, so slots are expected first when they have traveled through the network [5] [29]. This type of synchronization is always used inside a cluster. A discussion of timing and dispersion in WDM star networks can be found in [30].

The inter-cluster worst-case latency can be reduced by using the synchronization scheme shown in Figure 13. For clarity, all fibers are assumed to have the same length and propagation delay and to have a propagation delay significantly lower than the cycle time of S slots, in the example in the figure. These restrictions do not apply to a real implementation.

The midpoint of the backbone star is used as the reference point. Each cluster is then synchronized to the backbone by its gateway node. In a gateway node, the receiver on the cluster side is synchronized so tightly to the transmitter on the backbone side so incoming messages is directly forwarded. In this way, the information in an incoming control slot can be transmitted to the other gateway nodes, like a pipeline mechanism, before the data slots have arrived. This will reduce the worst-case latency by $\gamma\,(L^2 + L + 1)$, i.e., eliminate the latency for one hop as described by Equation 4. On the other hand, the receiver

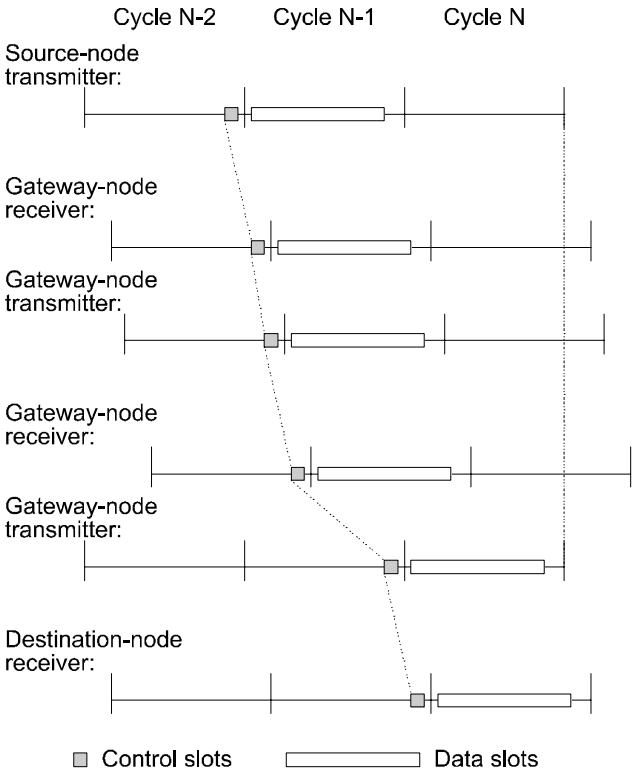

Figure 13: With the synchronization scheme used, incoming traffic to a gateway-node can be forwarded immediately except for internal delay in the gateway node.

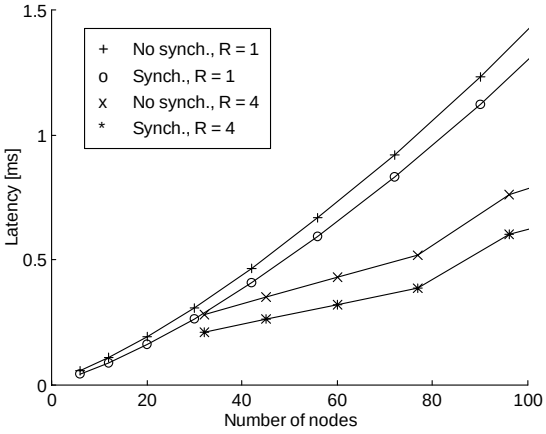

Figure 14: Worst-case latency comparison between usage and no usage of the proposed synchronization scheme. No slot-reservation is assumed. The x-axis represents the number of ordinary nodes and the slot length is assumed to be γ= 1.0 μs.

on the backbone side is not synchronized to the transmitter on the cluster side. This is because transmitters in a cluster are synchronized to the receivers in the same cluster. Therefore, the same latency is experienced as for the case when the clusters and the backbone are not synchronized (described above).

The improved worst-case latency (Equation 9) when full slot-reservation is assumed is

$$\tau_{max} = \gamma\left(Trunc\left(\frac{(L-R)(L-1)}{R} + 1 + L - R \right) L^2 + 2L + 2 \right) \quad (13)$$

The corresponding improved latency when no slot-reservation is assumed (Equation 12) is

$$\tau_{max} = \gamma\left(Trunc\left(\frac{L}{R} + 1 \right) L^2 + 2L + 2 \right) \quad (14)$$

The relative latency improvement is best in the case of no slot-reservation for networks with larger values of R (Figure 14). The figure compares the latencies with and without synchronization between the clusters and the backbone. Note that the latter latency is the same as that described above by Equation 12. Even better relative improvement is achieved when a slot can pass through the whole network without multiplexing over several cycles in the gateway nodes, e.g., by the use of reserved slots or at low traffic. In this case, the worst-case latency decreases from 3γ (L^2 + L + 1) to 2γ (L^2 + L + 1), i.e., an improvement of 33 percentages. The synchronization scheme is general and can be used in other similar networks also to improve performance.

5 Conclusions

We have shown how to calculate the worst-case latency for inter-cluster communication in a WDM star network using the TD-TWDMA protocol. The analysis shows how the latency, for larger networks, decreases when the ratio between the backbone bandwidth and the cluster bandwidth increases. A calculation of the minimum deterministic bandwidth obtained in the case when no reservation is used by the analyzed node, has been presented. Also, a synchronization scheme is proposed. At reserved traffic or at low traffic, the improvements to the worst-case latency when using this synchronization scheme is 33 percentages, compared to the case when the clusters are not synchronized with each other.

6 Acknowledgement

This work is part of the REMAP project, financed by NUTEK, the Swedish National Board for Industrial and Technical Development, and the PARAD project, financed by the Swedish Ministry of Education in cooperation with Ericsson Microwave Systems AB.

7 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] 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.

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

[5] M. Jonsson, K. Nilsson, and B. Svensson, "A fiber-optic interconnection concept for scaleable massively parallel computing," *Proc. Massively Parallel Processing using Optical*

Interconnections (MPPOI'95), San Antonio, TX, USA, Oct. 23-24, 1995, pp. 313-320.

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

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

[8] Chraplyvy et al., "1-Tb/s transmission experiment," *IEEE Photonics Technology Letters*," vol. 8, no. 9, pp. 1264-1266, Sept. 1996.

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

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

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

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

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

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

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

[16] C. A. Brackett, "Dense wavelength division multiplexing networks: principles and applications," *IEEE Journal on Selected Areas in Communications*, vol. 8, no. 6, pp. 948-964, Aug. 1990.

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

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

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

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

[21] C. A. Brackett, "Foreword: Is there an emerging consensus on WDM networking?," *Journal of Lightwave Technology*, vol. 14, no. 6, pp. 936-941, June 1996.

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

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

[24] K. Bogineni and P.W. Dowd, "A Collisionless Multiple Access Protocol for a Wavelength Division Multiplexed Star-Coupled Configuration: Architecture and Performance Analysis," *Journal of Lightwave Technology*, vol 10, no. 11, pp. 1688-1699, Nov 1992.

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

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

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

[28] 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.

[29] L. Bengtsson, K. Nilsson, and B. Svensson, "A high-performance embedded massively parallel processing system," *Proc. IEEE and Euromicro Conference (MPCS'94)*, Ischia, Italy, May 1994, pp. 201-206.

[30] G. Semaan and P. Humblet, "Timing and dispersion in WDM optical star networks," *Proc. INFOCOM'93*, San Francisco, CA, USA, 1993, pp. 573-577.