ATOLL: PERFORMANCE AND COST OPTIMIZATION OF A SAN INTERCONNECT

U. Brüning H. Fröning P. R. Schulz L. Rzymianowicz

Chair of Computer Architecture, University of Mannheim, Germany
D-68131 Mannheim, D7, 2-4
email: ulrich@ti.uni-mannheim.de

Abstract

The goal of the ATOLL System Area Network (SAN) is the optimization of a communication system for cluster computing with respect to the following qualities: performance, cost, simplicity of use and general applicability. The design provides many new and unique features, which makes ATOLL an enabling technology in the area of cluster computing. After finishing the ATOLL design space analysis our major effort concentrated on the high performance chip implementation. The following paper concentrates on the aspects of the hardware optimization process at the different levels of the design.

Keywords: System Area Networks, Cluster Computing, Network Interface

1 Introduction

Cluster computing has become a promising trend in high performance computing due to the steady cost-reduction and performance increase of the used computing resource, the PC. Mass production makes the PC a cheap, efficient and attractive building block for a parallel computing system. Computing can be performed at very high performance levels by these 'Gigahertz'-Processors but the communication between the PCs, used as node computers, is mostly underdeveloped. Using the cheap Fast-Ethernet or the Gigabit-Ethernet for communication may satisfy some applications but many demand much higher communication bandwidth and more important much lower latency as the Fast/Gigabit-Ethernet can provide [5]. This is the driving force behind the development of SANs, which tries to trade bandwidth versus interconnect length and costs. One very successful representative of such a System Area Network (SAN) interconnect is Myrinet 2000 [7]. Other players in the area of high-speed interconnects are SCI [8] and Quadrics Network QsNet [10]. The QsNet is the most sophisticated interconnect as it includes virtual memory support by a translation look aside buffer (TLB) and programability by two onchip processors, a thread CPU and a µCoded CPU. It would be an ideal interconnect besides the extraordinary high costs and the requirement to add crossbar chips to construct the network interconnect. A detailed comparison to SCI and Myrinet can be found in [16].

The basic architectural ideas for the ATOLL interconnect have been derived many years ago from the implementation of a parallel computer system named MANNA [1]. and PowerMANNA. A complete synthesizable simulation model was developped for the network interface [2] connecting directly to the bus of a 4 processor node architecture based on the PowerPC 620 CPU. The long delay in the availability of the CPU lead to a major redesign of the host interface now aimed to connect to a standard I/O bus (PCI-X).

The ATOLL development [3] tries to optimize the price/performance ratio of the network interface controller by implementing all required components (host interface, host ports, network interfaces, switch, link interfaces) for a SAN on a single chip (except the link cables) combined with very innovative architectural features at a price/performance range not yet achieved.

In the following chapter we will describe the basic architecture of ATOLL. Chapter 3 discusses the optimization for performance and cost. Chapter 4 provides examples of some special hardware features and how they are implemented. In chapter 5 we will present the performance data and the cost relations. Chapter 6 concludes our description and provides some remarks.

2 Basic architecture of ATOLL

The ATOLL Chip (figure 1) completely integrates a switch, network ports and network interfaces on a single chip. It mainly consists of a self-routing 8x8 crossbar switch where four ports are used as link ports connecting to the interconnection network and four network ports con-

necting to the host ports. The four host ports are completely replicated devices to directly support up to four processes [4]. This feature fits well to the popular SMP dual and four processor nodes.

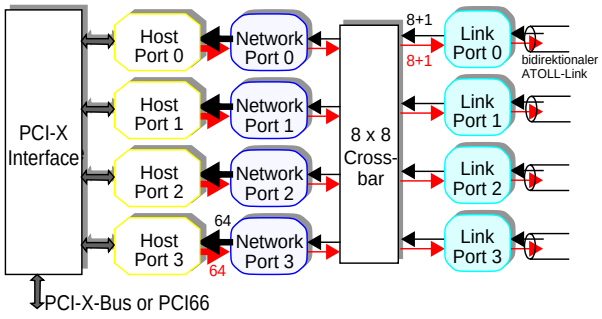

Figure 1: Top level ATOLL-Chip structure

The host ports can be mapped into the user's address space to give the user direct access to the communication device (user level communication). User level communication does not involve the operating system for each send/receive operation and thus significantly reduces communication latency, which also has an impact on the implementation of common parallel programming interfaces like MPI or VIA [15].

Two modes of operation are supported in order to reach a very low latency and also very high bandwidth.

The programmed I/O mode (PIO) is the ideal methode to inject short messages into the network, as the processor usually holds the data in its local cache or registers and only has to copy the data from its registers or the cache to the host port's fifo of the NIC. A special addressing of the fifo allows to utilize write combining (and also read-prefetching) and use burst transfers through the PCI-Bus.

The direct memory access mode (DMA) guarantees full concurrency between NIC and processor for message transactions and is descriptor based. A descriptor for a message is constructed and entered into the host port's descriptor queue located in main memory. After incrementing the write pointer in the NIC the ATOLL chip performs the communcation function fully autonomously by HW. Further details of the data transfer mechanisms can be found in [4].

As a connection to the host system the standard PCI-X interface [9] is used. The PCI-X core [13] connects the four host ports to the PCI-X bus providing a maximum bandwidth to the host of 1GBytes/s (133MHz x 8 Bytes). Slower bus specifications are also supported, e.g. PCI-X 100, PCI-X 66 and PCI 66/33 with a data width of 64 or 32 bits. The chip is restricted to 3.3V slots because 5V signals would exceed the maximum break through voltage of the input buffer transistors at the 0.18µm UMC logic CMOS technology.

The network port converts the 64bit data stream to and from the host port to a byte wide stream for the interconnect system. The link port adds the link level protocol to the message and controls the transmission from link port to link port through the link cable. This stream is automatically partitioned into Link Packets (LIPs) of 64 bytes each and extended by a CRC check to verify the transmission. In case of an error, an automatic retransmission is performed between two link cable endpoints.

One link cable houses a sending and a receiving data channel each 9bit wide, providing a bidirectional interconnect between two nodes. The bandwidth is directly related to the ATOLL onchip clock and thus results in 250MBytes/s for each direction on one bidirectional link, summing up to an aggregate bandwidth of 2GBytes/s for all four links. The typical cluster to be constructed using four links will be a grid or torus as depicted in figure 2.

The bisection bandwidth of this 16 node configuration is 8 x 500MBytes/s = 40GBytes/s and the longest path is 4 hops. Each hop adds merely 27 clock ticks to the pipelined message transport time which is about 100ns. This interconnect structure is scalable to a high number of nodes (8 x 8 = 256 or more) because adding nodes also adds crossbar switches for interconnectivity.

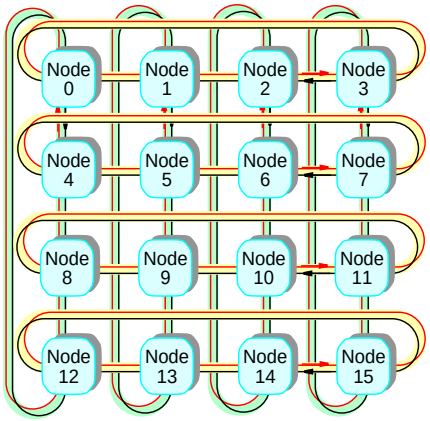

Figure 2: Cluster interconnect for 16 nodes

The limits of the torus will be reached, if long messages block a lot of intermediate crossbar stages. This problem can be deminished by reducing the maximum size of the packet length. FIFO input buffer in each channel in front of the crossbar minimize this effect and provide a smooth transition from wormhole routing to store-and-forward in case of a heavy loaded interconnect.

There is one remarkable architectural feature of the ATOLL chip which has been carefully evaluated by simulation. All required components for the interconnect are included and most functions are realized directly in hardware but **no processor and no large message buffer RAM** are added to the chip. The reason not to include a processor for message setup or other communication related functions is the speed advances of using one of the SMP-'Gigahertz' host processors for such functions. Including a 250MHz processor in the chip would have increased silicon area significantly and never would such a processor scale with the

technology as fast as the host CPU. Furthermore, a development environment had to be supported for the instruction set which would have consumed a lot of resources better used at other items in the design.

No large buffer RAM can be found on the chip (only the functional required FIFOs) because this would have reduced the yield of the chip production. Adding the buffer SRAM to the board is expensive, restricted in size and would have increased the pin count significantly. All these arguments made it obvious that all buffer space should be located in main memory. This solution is better scalable due to the large size of the main memory compared to onboard or onchip SRAM.

3 Optimization for performance and costs

The complete integration of the network interface (NI), consisting of a host interface, the host ports, the network port, the link interfaces and most important also the **crossbar switch** makes it possible to build SANs with only one single generic component, the ATOLL chip. Producing and supporting only one single integrated chip keeps costs low and let the design team concentrate on one chip to optimize performance.

But not only the chip, the whole system must be optimized for performance and every item must be checked for the most cost-effective implementation alternative. The following components have been deeply analysed:

- link cables
- link connectors at the printed circuit board
- printed circuit board for mounting the chip
- chip package
- chip

To begin with the interconnect cables, the design decision was to use a byte wide link with a cheap but high performance electrical transmission.

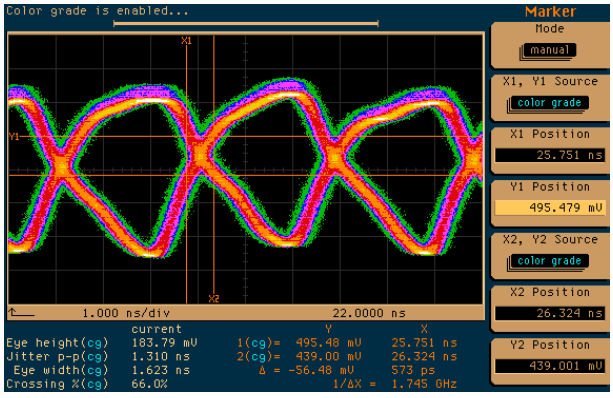

Figure 3: Eye diagram of LVDS transmission

The choise for serial electrical or serial optical transmission was rejected because it would have required additional transceivers with 2.5 GHz serializer and high costs in the transceivers and the cabling. Twisted pair cables with 68 signal lines and with controlled impedance of 100 Ohm and a maximum length of 10 m (30 ft) can transmit 2 x 250MBytes/s (equals a signal frequency of 125 MHz) with small technological effort and leaves space for future enhancements. The standard SCSI-3 connectors in a stacked version are used for the board connectors. Transmission simulations and tests have been carried out to check for reflections and signal integrity over the signal path (chip, package, printed circuit board (PCB), connector, cable, connector, PCB, package, chip) and have shown very low error rates and sufficient opened eye diagrams.

For long distance interconnects (100m) an optical transmission interface is under development which can be plugged to the board connector, converting the LVDS byte wide channel to a serial optical signal thus adding costs only when required.

The printed circuit board must interconnect the link cable connector with the chip utilizing differential traces with controlled impedance of 100 Ohm. The PCI-X edge connector must be connected to the chip by using traces with 57 Ohm ±10%. Controlled impedance requires a ground plane for the implementation of microstrip lines. These requirements defines a minimum of 4 layers, 2 outside layers (top and bottom) for running the traces and 2 inner layers as planes for VDD (3.3V) and GND for the microstrip reference planes. Staying with only two interconnect layers was very difficult to layout, but adding another two layers would have increased costs by 25%.

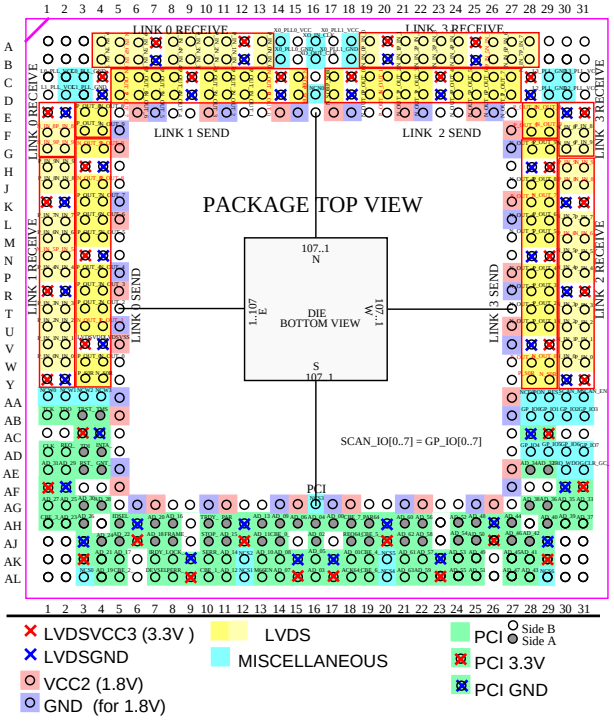

Figure 4: BGA signal to ball assignment

Now, the design challenge was to define the pin out of the chip and the link connectors in such a way, that no crossing of signals occurs. The same solution must be found for the PCI edge connector traces running to the chip, in which the PCI ordering is fixed. Exchanging crossed sig-

nals would have needed vias and additional layers on the board. A solution was found, where all transmitting traces run on the bottom layer with only one via directly located at the pad of the chip (the BGA of the chip has only pads on the top layer). All receiving traces run on the top layer without any via which avoids discontinuities in the impedance. A similar ordering was found for the PCI. The package of the chip was also included in this optimization process. We discovered that only a chip with 5 rows of a 31 x 31 matrix BGA (520 balls) fits our requirements as presented in figure 4. More than 5 rows showed massive routing congestions which could not be resolved in 2 routing layers. The inner ring of balls was used only for the core voltage supply (GND and 1.8V VDD). The package provides a GND plane and thus allows to control the impedance also in the package until the bond finger pad. The wire bond creates a disturbance which cannot be avoided in a cost-effective manner (expensive flip-chip mounting could avoid the wire bond discontinuity). 6 balls in the corners of the package carry no signals. This simplifies the routing of traces from ball to bond finger pad, which otherwise would have led to routing congestions in the corner area. So we could stay with the standard feature size of 30µm for the trace width and keep costs low.

LVDS signal traces require a termination resistor for every differential pair. These usually discrete components sum up to 40 terminaton resistors, which must be placed as close as possible to the end of the transmission line. For optimal performance, the termination resistors should be integrated directly into the receiver I/O-cell, avoiding route and place of 40 discrete components thus reducing costs. A LVDS receiver I/O cell with directly integrated termination resistors was developed by the University of Kaiserslautern for this project.

4 Special hardware features of ATOLL

The data transmission on the link cable uses the differential LVDS-Standard (‘low voltage differential signaling’) [11]. LVDS has many advantages over common used technologies, as there are:

- very low power consumption also at high frequencies
- use of alternating switched current sources (4mA)keeps power consumption of driver nearly constant
- simple parallel resistor termination (100 Ohm)
- high signal to noise supression ratio with 1V common mode voltage rejection
- low EMI by use of low voltage swing and differential transmission over twisted pair cables
- hot pluggable in power-on state
- use of transmission lines with controlled impedance featuring low skew and thus allows wave pipelining

Link data transmission is performed on 9 differential signal lines and one differential clock line, requiring 2 x 20 wires for one bidirectional link. The link clock signal is in phase with the data signals and has the same frequency as the fastest possible data changes. This requests a cable bandwidth of 125MHz when the ATOLL is operated at 250MHz. Both edges of the clock determine data changes (double data rate) and the midpoint in between the posedge and the negedge of the link clock are the optimal strobe position for data. The link clock is multiplied by 2 by an analog PLL and the phase difference is adjusted to zero. The PLL clock is then inverted to generate the sample clock, whereas the posedge of the sample clock is now the optimal strobe position for the data.

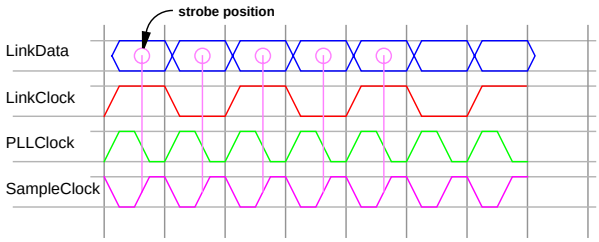

Figure 5: Link Data and Clock recovery

The ATOLL clock generation is optimized for performance and cost. Every ATOLL chip has an integrated clock oszillator controlled by an external crystal of 14.318 MHz, which is the cheapest precise clock base to be installed. The onchip clock frequency is generated from a PLL which can be programmed as shown in table 1.

The software can program the clock in the range from 114 to 501 MHz, depending on the speed grade of the chips in a cluster.

n	clk f.	n	clk f.	n	clk f.	n	clk f.
8	114,544	15	214,770	22	314,996	29	425,222
9	128,862	16	229,088	23	329,314	30	429,540
10	143,180	17	243,406	24	343,632	31	443,858
11	157,498	18	257,724	25	357,950	32	458,176
12	171,816	19	272,042	26	372,268	33	472,494
13	186,134	20	286,360	27	386,586	34	486,812
14	200,452	21	300,678	28	400,904	35	501,130

Table 1: Programmable clock frequencies

The typical operating frequency of ATOLL will be 250MHz and all ATOLL chips of a cluster must be operated at approximately the same frequency (same value of n). This is accomplished by using the same clock table entry for all ATOLL chips in a cluster and then only the small frequency difference (50ppm) between the crystals exists. This small difference is resynchronized at every input channel by a synchronization FIFO (every channel carries its own clock!), which can drop IDLE characters from the stream to compensate for a slightly faster sender. If we assume a max. deviation between two ATOLL chips, it is sufficient to insert one IDLE character into the data stream every ~700 clocks. The resulting bandwidth reduction is less than 0.1% and is therefore negligible for this compensation.

Another feature of the ATOLL is a special clock synchronization module. This module can switch the master clock under software control from the oszillator to one of the incoming link clocks without a clock phase shift. After switching to a link clock the two ATOLL chips connected by this link run in synchronous mode, having no further frequency deviation. If all chips are switched to this mode the whole cluster runs from the single master clock generated from the first chip in this "link clock tree". This mode avoids any overhead for resynchronization and keeps all counter register in the cluster synchronous. Then the 64bit global timer register provides a cluster wide timing reference with a resolution of one clock tick (4ns at 250MHz clock) with no frequency deviation between any of these registers. This can be used for global time stamping of messages.

For optimization of communication performance, it is necessary to have information about the link utilization. Hot spots on specific links can be avoided by changing the routing of messages, if the actual link utilization is known. Therefore, link performance counters (lpc) are added to each outgoing link, which counts the number of transfered bytes. The lpc is a 32bit counter incrementing monotonously for every transmitted byte. Overflow of this register must be handled by software. Reading a link performance counter (lpc) and the global timer register (gtr) at the start of a measurement interval (0) and at the end (1) gives the link utilization *U* by the following formula:

$$U = \frac{lpc1 - lpc0}{gtr1 - gtr0}$$

The time interval is defined by software and the calculation must also be performed by software. This provides a high flexibility in defining measurement intervalls. Furthermore, these operations can be executed much more cost effectively than by an onchip division unit.

An additional feature of the ATOLL chip is the 8 bit wide general purpose I/O-port (GPIO). It is fully software controllable through a 32 bit onchip register, which contains 4 x 8 bit fields for the concurrent write to the output enable bits and the output bits driven to the I/O-pads. The input can be read in the third field while the fourth field is kept internal. The byte arrangement is presented in the following figure.

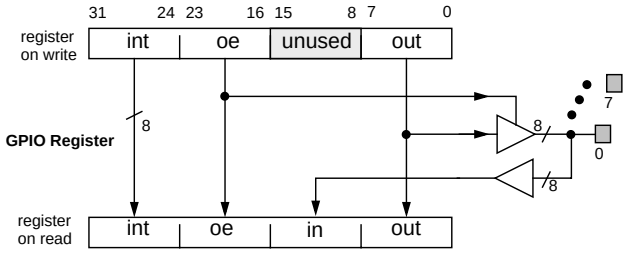

Figure 6: GPIO register and I/O-model

Only 2 pins of this port are used on the ATOLL board for the I^2C bus to pheriperals like temperature sensors and configuration memories. The 6 others are fed to a connector for user specific extensions or experimental purposes. For example, one pin can serve as special function pin for the implementation of wired or barriers [6] as depicted in figure 7.

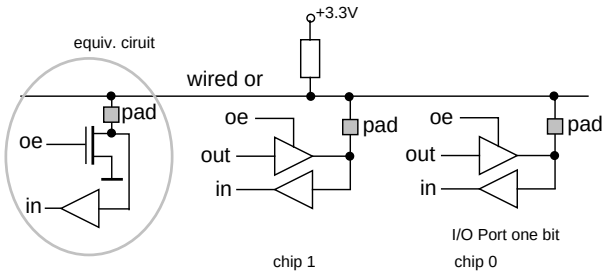

Figure 7: barrier line using GPIO

5 Performance results

A HW start up latency of 2.4 µs was found on small message sizes of 32 Bytes, slightly increasing to 3.3 µs for 256 Bytes. The link utilization in relation to the message size is presented in the following figure. Message sizes of 8k Bytes and larger nearly fully utilize one of the links.

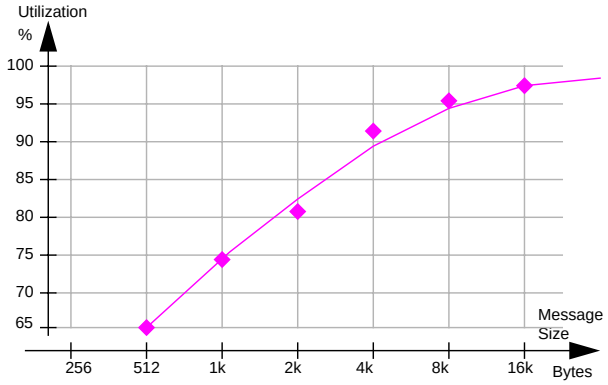

Figure 8: Link utilization [14]

The bandwidth asymptotically approaches a maximum sustained bandwidth of 225-230 Mbyte/s, or about 90 % of the theoretical maximum bandwidth. The PCI-X bandwidth is fully utilized when using 4 host posts, reaching a sustained bandwidth of 920 MBytes/s.

6 Conclusion

What we have presented in this paper are the ways we have found to optimise the cost and performance of a NIC. Also personnel costs (US$ 0,10 per transistor)are low to comparable industrial projects. Some of the main hardware aspects of the ATOLL SAN are described. Additionally we provide all required software layers to run the ATOLL chip with MPI and PVM as 'open source'. Furthermore the AT-

OLL architecture is documented giving researchers the possibility to work with us on the future of high performance cluster computing.

7 Acknowledgements

At the end, I (Prof. Brüning, chair) want to make some historical remarks to provide the reader with some insight to the ATOLL project. The whole project from the conceptual design to the fully functional chip has been carried out through a time of now more than 5 years and was not funded by third parties. Three years ago, right during the place & route phase of the design our former project partner canceled collaboration, which forced us to re-synthesize and port the chip to 0.18µm CMOS in order to stay cutting-edge. Getting funded for HW-developments like ATOLL is nearly impossible in Germany. Only our restricted budged and the manpower (3 research assistants) of the chair have been used to keep the project going.

At this point I want to thank the hardware design team, the software team and the students working with motivation on our dream.

Further I have to thank Greg Papadopolous from SUN Microsystems for his donation of 4 SUN Ultra 2 computers, which provides us with the computing power and the memory space to run the synthesis and the simulation of the ATOLL-Chip. Synopsys gave us many helpful advises for their tools and licensed the PCI-X core to our project. IMEC in Belgium has done a good job on the physical layout. Prof. Tielert with his expertise in analog design and his research assistant Marc Wegener from the University of KL developed the PCI-X and the LVDS I/O-Cells for the chip.

Many more people should have been named here, which provided us with intellectual support and helped us to let our dream come true, a 5x5 mm ‘small’ network on a chip with 1,6 Mio. gates (4,5 Mio. transistors).

Sometime in the mid of the project, we thought it might be a good idea to have an industrial partner. Astonishingly we found one, but after some time of intensive colaboration we lost our partner due to a merger in 1999. Thereafter we tried it ourselves.

8 References

[1] Wolfgang K. Giloi, Ulrich Brüning, Wolfgang Schröder-Preikschat, MANNA: Prototype of a Distributed Memory Architecture with Maximized Sustained Performance, *Proceedings Euromicro PDP Workshop*, 1996.

[2] Ulrich Brüning, Lambert Schaelicke, ATOLL: A high-performance communication device for parallel systems, *Proc. of the 1997 Conference on Advances in Parallel and Distributed Computing*, Shanghai, China, March 19-21, 1997.

[3] L. Rzymianowicz, U. Brüning, J. Kluge, P. Schulz and M. Waack, ATOLL: A Network on a Chip, *Cluster Computing Technical Session (CC-TEA) of the PDPTA'99 conference*, Las Vegas, NV, June 28 - July 1, 1999.

[4] Jörg Kluge, Ulrich Brüning, Markus Fischer, Lars Rzymianowicz, Patrick Schulz and Mathias Waack, The ATOLL approach for a fast and reliable System Area Network, *Third Intl. Workshop on Advanced Parallel Processing Technologies (APPT'99) conference*, Changsha, P.R. China, October 19-21 1999.

[5] Oak Ridge National Laboratory http://www.epm.ornl.gov/~sscott/presentations/HPCU99/img015.gif

[6] M. T. O’Keefe and H. G. Dietz, "Hardware barrier synchronization: static barrier MIMD (SBM)," *Proc. of 1990 Int’l Conf. on Parallel Processing*, St. Charles, IL, pp. I 35-42, August 1990.
see also:
http://aggregate.org/TechPub/ICPP95/icpp95.html

[7] Myricom: A Gigabit-per-Second Local Area Network, IEEE Micro, 1995.

[8] Dolphin: PCI-SCI-Adapter Card, http://www.dolphinics.com/products/pci64_adapter_card.html

[9] PCI-X Specification 1.0 from PCI-SIG http://www.pcisig.com/specifications/pci_x

[10] QsNet: The Quadrics Network, Quadrics Supercomputer Ltd., Jan. 1999.

[11] LVDS TIA Standards, Global Engineering Documents, http://global.ihs.com

[12] Internal Technical Report: ATOLL Hardware reference Manual, Chair of Computer Architecture, University of Mannheim, 2001.

[13] DesignWare DW_pcix, MacroCell Databook, Synopsys Inc., Release 2001.10, October 2001.

[14] L. Rzymianowicz, *Designing Efficient Network Interfaces For System Area Networks*, Ph.D. Thesis, University of Mannheim, 2002.

[15] M. Fischer, U. Bruening, J. Kluge, L. Rzymianowicz, P. Schulz, M. Waack, Impact of Configurable Network Features in ATOLL, *Proc. of HPCAsia2000*, APSCC 2000, Beijing, China, May 14-17, 2000.

[16] M. Fischer, U. Brüning, J. Kluge, L. Rzymianowicz, P. Schulz, M. Waack, ATOLL, a new switched, high speed Interconnect in Comparison to Myrinet and SCI, *Proc. of IPDPS 2000*, PC NOW Workshop, Cancun, Mexico, May 1-5 2000.