RHiNET-3/SW: an 80-Gbit/s high-speed network switch for distributed parallel computing

S. Nishimura[1], T. Kudoh[2], H. Nishi[2], J. Yamamoto[2], R. Ueno[3], K. Harasawa[4], S. Fukuda[4], Y. Shikichi[4], S. Akutsu[4], K. Tasho[5], and H. Amano[3]

[1] RWCP Optical Interconnection Hitachi Laboratory.
(c/o Central Research Laboratory, Hitachi, Ltd.)
E-mail: nisimura@crl.hitachi.co.jp
[2] Real World Computing Partnership, Tsukuba Research Center
[3] Dept. of Information and Computer Science, Keio University
[4] Hitachi Communication Systems, Inc.
[5] Synergetech, Inc.

Abstract

We have developed a prototype network switch, RHiNET-3/SW, for a RHiNET high-performance distributed parallel computing environment. It has eight I/O ports and each port provides high-speed, bi-directional 10-Gbit/s-per-port parallel optical data transmission in a distance of over 300 m. The aggregate throughput is 80 Gbit/s per board.

A switch consists of a one-chip CMOS ASIC 8x8 switch LSI (SW-LSI; a 784-pin BGA package), four deskew LSIs, and eight pairs of 1.25-Gbit/s x 12-channel optical interconnection modules on a single board. The switch uses a hop-by-hop retransmission mechanism and credit-based flow control to provide reliable and long-transmission-distance data communication. The deskew LSI has a skew compensation function for 10-bit parallel data channels and an 8B10B encoder/decoder. Its optical transmitter modules use an 850-nm VCSEL and a 12-channel MMF (multi-mode fiber) ribbon. RHiNET-3/SW enables high-throughput, long-distance and flexible-flow-control network communication by means of distributed parallel computing using commercial PCs.

1. Introduction

RHiNET (real world computing partnership , RWCP, high-performance network) represents a new class of networks that offer the advantages of both system-area networks (SANs) and local-area networks (LANs) [1]-[3]. RHiNET is basically designed to realize high performance parallel processing environment by connecting computers distributed on one or more floors of building. Specially designed network interfaces and network switches, in combination with high speed optical interconnections, will support fast communication that realizes high performance parallel processing in a local area distributed computing environment.

Most high-performance cluster systems consisting of PCs use a SAN, for example Myrinet [4], for interconnection. SANs provide low-latency, large-bandwidth communication by using wormhole- or virtual-cut-through routing. Therefore, SANs are suitable for interconnecting the nodes in a parallel computing system. However, SANs are designed to connect dedicated computers within a small area, so the length of the links and the topology of the network must be restricted in order to achieve high performance. On the other hand, high-speed LANs, e.g., Gbit-Ethernet and 10Gbit-Ethernet, with link bandwidths of over 1 Gbit/s are becoming available [5], [6]. LANs provide relatively free topology choices and longer links. However, the communication latency of most of today's commercial LANs tends to be greater than that of SANs because of their store-and-forward routing strategy. Moreover, today's LANs support the Internet protocol (IP), which introduces additional overhead for use in a node-to-node communication in parallel computing. RHiNET is therefore designed not only for cluster (i.e., SAN) environments but also for local-area distributed environments with diameters of several hundred meters (i.e., a LAN). RHiNET provides high-throughput, long-transmission-distance and low-latency node-to-node communication by using optical interconnections.

Aiming towards practical RHiNET implementation, we have developed three prototype systems. The first, a network switch called RHiNET-1, uses 1.33-Gbit/s optical interconnections [7]. The second, RHiNET-2/SW, uses 8-Gbit/s optical interconnec-

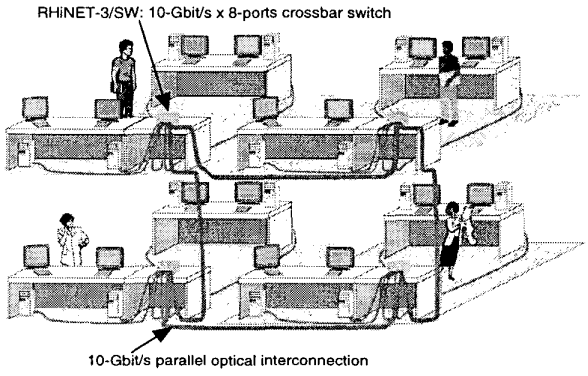

Fig. 1: Schematic structure of RHiNET-3

0-7695-1357-3/01 $10.00 © 2001 IEEE

tions [8], [9]. We demonstrated 16-node parallel processing by using sixteen PC nodes and five RHiNET-2/SWs. Based on RHiNET-2/SW, we have designed the third switch, RHiNET-3/SW (Fig. 1). RHiNET-3/SW can provide high-throughput (bi-directional 10 Gbit/s per port), long-distance (>300 m), and free-topology (32 virtual channels (VCs) with credit-based flow control) network communication.

2. RHINET-2/SW

RHiNET-2/SW is 8-Gbit/s x 8-port switch. Using RHiNET-2/SW, we have succeeded to demonstrate 16-node parallel computing. However, we found some issues to construct more nodes and long-transmission-distance parallel computing system, using RHiNET-2/SW. RHiNET-3/SW has been designed to improve the issues of RHiNET-2/SW.

RHiNET-2/SW has eight ports, and each with 8-Gbit/s transmission capacity, and its aggregate throughput is 64 Gbit/s [8]. The core of RHiNET-2/SW is one chip switch that consists of a 0.18-micron CMOS ASIC. It provides 800-Mbit/s-per-pin high-speed low-voltage differential signaling (LVDS) I/O. Its aggregate throughput is 64 Gbit/s per chip. It provides sixteen VCs at each port and a 4-Kbyte VC buffer in each VC on chip in order to avoid delay. The total amount of VC buffer memory is 512 Kbytes. RHiNET-2/SW provides flow control by using a slack buffer system. Special GO or STOP character is used in handshaking. Maximum link length is 100 m, which is limited by the size of the buffer memory. Routing is done according to the routing information statically stored in the routing table of each switch. Each switch has a routing table with 65,536 entries. To support slower network interfaces, each port can be set to bit rates of 8 Gbit/s, 2 Gbit/s or 1 Gbit/s.

Each I/O port uses 8-Gbit/s (800-Mbit/s x 10-bit) synchronized parallel optical interconnection to achieve high-speed, small-skew node-to-node interconnection [9]. The optical interconnection module used in RHiNET-2/SW enables small-skew (< 50 ps), synchronized parallel data transmission via highly-uniform LD and PD arrays and SMF (single-mode fiber) ribbon. To eliminate the skew, gate-latch is used in both the transmitter and receiver modules. However, the transmission length is limited to 100 m by the skew of the SMF ribbon and the size of an input buffer in the switch is determined based on this limitation. The bit error rate (BER) of each channel is in the order of 10^{-20} in order to maintain the simple data transmission without the need for skew compensation or complex error correction.

3. RHINET-3/SW

3.1. Concept

RHiNET-2/SW was designed based on the use of very-low-error-rate (BER: 10^{-20}) optical interconnections. Such low-error-rate links could be considered as error-free for actual system use. However, the optical interconnect in RHiNET-2/SW consists of expensive material and a precise structure in order to reduce skew, and provide low-error-rate DC-coupled communication up to 800 Mbit/s per channel over 100-m connection. Currently, low cost and limited-band optical interconnects that do not support skew-less communication are widely utilized. We have, therefore, designed a new switch with hard-wired error detection and hop-by-hop retransmission mechanism, called RHiNET-3/SW. This switch has eight input ports and eight output ports and each link has bi-directional 10-Gbit/s data throughput. It uses 1.25-Gbit/s x 12-channel AC-coupled optical interconnections in each data I/O port. In addition, to support links of more than 100 m, it uses credit-based flow control and a deskew function.

RHiNET-3/SW consists of one 0.14-micron CMOS ASIC 8x8 switch-LSI (SW-LSI), four 0.14-micron CMOS ASIC deskew LSIs (DS-LSIs), and eight pairs of 12-channel optical interconnection transmitter and receiver modules (Fig. 2). It has eight I/O ports and each port has a 1.25-Gbit/s x 10-bit 8B10B encoded-data channels with a 1-bit transmission clock. A 10-bit parallel optical signal is converted to a 10-bit electrical signal in the optical receiver and is sent to the RX-block of the DS-LSI. In the RX block, the channel-to-channel skew is corrected, a 10-bit signal is 8B10B decoded to an 8-bit synchronized signal, and it is sent to the SW-LSI. The 8-bit signal is switched by the SW-LSI, encoded to a 10-bit parallel signal in the TX block of the DS-LSI, re-converted to an optical signal in the optical transmitter, and transmitted through the 12-channel MMF ribbon. Thus, each port has a bi-directional 10-Gbit/s throughput, and the aggregate throughput is 80 Gbit/s. All electrical I/O interfaces have CMOS LVDS or CML (current mode logic) interfaces, which make all components easy to use in practical computer circuits.

3.2. Architecture

RHiNET-3/SW is provided with retransmission mechanism to realize reliable network. In addition, RHiNET-3/SW enables more nodes and a longer connection distance than RHiNET-2/SW. It thus provides the following five improved functions.

3.2.1. Hop-by-hop retransmission. RHiNET-3/SW supports low-cost optical interconnection modules, which may cause non-negligible transaction errors. Therefore, a hop-by-hop retransmission mechanism is provided to realize error-free communication between switches.
The unit of retransmission is a Micro Frame (MF) composed of two lines. Figure 3 shows the format of a MF. A MF contains a credit, a request sequence number, an acknowledge sequence number, and a CRC. The sequence numbers are provided in order to distinguish a bit error or a discarded line.

A buffer of 1024-lines for longer links (ports 0 and 1) or a buffer of 256-lines for shorter links (ports 2-7) is equipped in each output port as a retransmission buffer.

3.2.2. Credit-based flow control. RHiNET-3/SW uses the credit-based flow control method to eliminate logical limitation of fiber length and to utilize packet buffers effectively [10], [11].

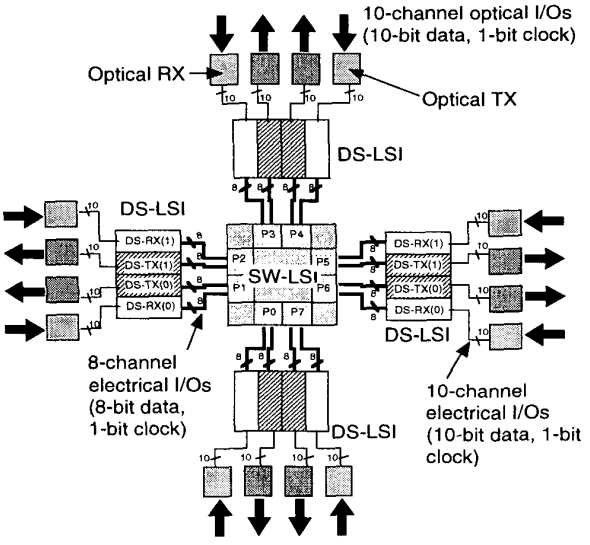

Fig. 2 Schematic structure of RHiNET-3/SW

It has an 80-bit-data handshaking unit consisting of an 8-byte payload and a 16-bit administration area. One-VC-buffer can store 128 lines; therefore, each port has a 32x7-bit credit pool.

3.2.3. 32 Virtual Channel (VC). RHiNET-3/SW has 32 VCs at each input port. RHiNET-3/SW supports a structured buffer pool and virtual networks using these VCs in a similar way to RHiNET-2/SW [8]. The structured buffer pool method[3] provides deadlock-free routing by providing a number of virtual channels at each input port. The number of virtual channels must be larger than the diameter of the network; i.e., larger than the maximum number of intermediate nodes between any two nodes in the network.

Virtual networks are fully independent networks. Virtual networks are realized by dividing the 32 VCs to some groups. A packet sent with a VC of a group will never use VCs belong to different groups.

The RHiNET-3 system needs four virtual networks in order to prevent deadlock among transactions. Therefore the VCs are divided into four groups. Eight VCs in each group are enough to connect a 1000-node system with 8x8 switches.

3.2.4. Cut-through function. When a packet cannot proceed in the network, another packet can 'cut-through' the suspended packet. In the network, a packet can be uncoupled or coupled at any line. Therefore, packets may be 'mixed' in the network. However, it is guaranteed that any packet arrives to the destination without mixture.

To prevent the mixture at the destination, packets that have the same destination and VC-ID cannot cut-through each other.

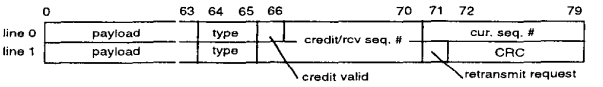

Fig. 3: Structure of Micro Frame (MF) packet

To manage this, RHiNET-3/SW uses two tables, S-Table and A-Table. The S-Table in each input port keeps an ID of a VC of the suspended packet. The A-Table in each output port keeps an ID of a VC in service. The table-look-up and packet-delivery are executed in parallel in order to reduce table-look-up overhead.

3.2.5. Routing. RHiNET-3/SW supports both table routing and source routing packet by packet. It assures no deadlocks in any routing algorithm by the structured buffer pool. The maximum number of hops in source routing is eight. Other deadlock-free algorithms such as the spanning-tree protocol [12], [13] can be used to enhance the size of the network.

3.3. SW-LSI

Figure 4 is a block diagram of the internal logic in a SW-LSI that has eight input ports and eight output ports. The input port receives a 1.25-Gbit/s x 8-bit packet and converts it to a 125-Mbit/s x 80-bit packet by using a 1:10 demultiplexer. After synchronizing the signal to an internal clock in the elastic buffer, the retransmit-rx module checks the MF by CRC or sequence number. If there is an error, it sends a retransmission request to the retransmit-rx module. A retransmission request from the link is also sent to the retransmit-tx module. A VC controller manages the VC memory and makes a request for arbitration. The retransmit-tx module adds an administration bit field consisting of CRC, sequence number, and credit number. Finally, a 10:1 multiplexer divides the 125-Mbit/s x 80-bit line into a 1.25-Gbit/s x 8-bit signal and sends it to the output port.

Table 1 lists the specifications of RHiNET-3/SW; internal logic frequency is defined according to memory-cell latency, and all its high-speed I/O interfaces use LVDS logic. Table 2 shows the gate sizes of SW-LSI.

The total latency of hop-by-hop transmission including two DS-LSIs is 192 + x ns (x: latency of link). The total latency of hop-by-hop retransmission is 390 + 2x ns.

3.4. DS-LSI

To provide long-transmission-distance of over 300 m, the DS-LSI can perform channel-to-channel skew compensation on the 10-bit receive side of each DS-RX. Also, the DS-LSI has 8B10B encoder and decoder for high-speed (1.25 Gbit/s) AC-coupled data transmission in the optical interconnection. It has two ports and each port has the functions of skew compensation and 8B10B encoding/decoding. Totally, four DS-LSIs are mounted around one SW-LSI, and in the RHiNET-3/SW, eight I/O ports can perform the skew compensation independently.

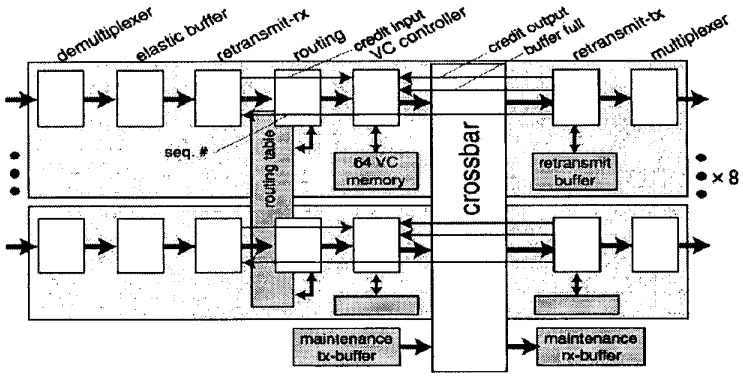

Fig. 4: Block diagram of internal logic of SW-LSI

Table 1: Specifications of SW-LSI

design rule	0.14 micron
silicon die size	272.91 mm^2
number of port	8
buffer size	80 Kbytes per port
number of VCs	32 per port
memory size of each VC	2 Kbytes
bandwidth	10 Gbit/s per port (payload: 8 Gbit/s per port)
routing table size	65536
latency	160 ns
I/O frequency	1.25 Gbit/s
internal logic frequency	125 Mbit/s
package	784 BGA

Table 2: Gate sizes of SW-LSI

name of part	number of gates
MUX	8k x 8
DEMUX, elastic buffer	8k x 8
ECC decoder	1k x 8
ECC encoder	2k x 8
VC controller	72k x 8
crossbar	50k
credit controller	150k
routing table controller	10k x 8
status register	25k
maintenance controller	39k
retransmission controller (TX)	27k x 8
retransmission controller (RX)	13k x 8
ports controller	8k x 8
etc	120k
total	1502k

A DS-LSI has two DS-TX blocks and two DS-RX blocks. The DS-TX block converts a 1.25-Gbit/s x 8-bit DC-coupled input signal (from SW-LSI) to a 1.25-Gbit/s x 10-bit 8B10B encoded output signal. The DS-RX block decodes a 1.25-Gbit/s x 10-bit 8B10B input signal from the optical receiver and outputs a 1.25-Gbit/s x 8-bit DC-coupled signal to the SW-LSI. The output signal of the DS-RX is guaranteed as synchronized parallel data.

The DS-LSI performs channel-to-channel skew compensation between the 10-bit optical parallel data channel and the 1-bit transmission clock channel, which optically connects the DS-TX to the DS-RX. Channel-to-channel skew is mainly caused by the skew from the 12-channel fiber ribbon (< 50 ps/m). In the initial state of data communication, the DS-TX block generates a special data pattern for skew compensation (the special pattern consists of consequent 8B10B special Comma bits). And the DS-RX block analyzes the received special data pattern and compensates the delay of each 10-bit data channel to the clock channel. The DS-LSI can compensate a skew of +/- 256 ns, which is larger than the worst case of a 1-km MMF ribbon (+/- 64 ns).

3.5. Optical link

RHiNET-3/SW uses 12-channel optical transmitter and receiver modules (MFT/MFR62340) produced by ZARLINK semiconductor [14]. The optical transmitter modules use 850-nm VCSEL (vertical-cavity surface-emitting laser diode) arrays. The channel configuration is made up of 12 CML signals. The CML output signals from the receiver module are converted to LVDS signals via a level converter. The data rate is 155 Mbit/s to 2.5 Gbit/s per channel and aggregate throughput is over 30

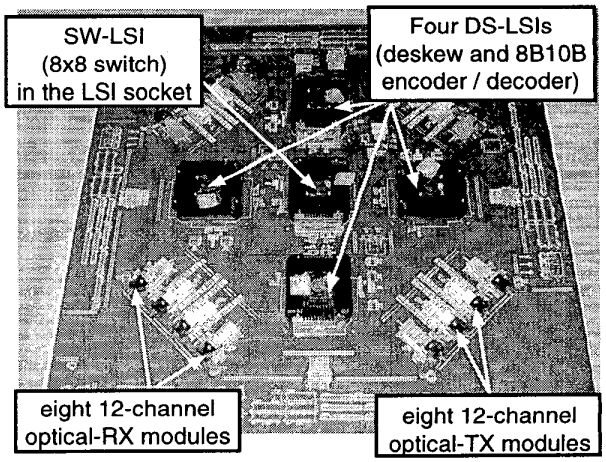

Fig. 5: Board layout of RHiNET-3/SW

Gbit/s per module. The switch uses a grated index (GI) 50/125 12-channel MMF ribbon. The transmission length is up to 300 m with a 500-MHz•km MMF fiber at a data rate of 2.5 Gbit/s (longer transmission length is available at lower data rate). The modules support MPO/MTP and MPX fiber connectors, and channel BER is about 10^{-12}.

3.6. Motherboard of RHiNET-3/SW

In the center of the motherboard of the RHiNET-3/SW eight-by-eight network switch, a 784-pin ball-grid-array (BGA) single-chip SW-LSI is mounted (Fig. 5). Four 784-pin BGA DS-LSIs are mounted around the SW-LSI. And eight pairs of optical transmitter and receiver modules are mounted near the LSIs. The board size is 550 by 550 mm. To achieve high-density implementation with high-speed (1.25 Gbit/s) signaling devices, we have to overcome many complex problems, as such as crosstalk, skew, and propagation loss [8]. We thus optimized the layout of the circuit board according to the experimental results to realize low-crosstalk, high-speed (1.25 Gbit/s), and high-density electrical I/O.

4. Summary

The developed RHiNET system enables high-performance parallel computing in a distributed environment. We have developed the RHiNET-3/SW network for high-performance computing using personal computers distributed in an office -floor environment. RHiNET-3/SW provides a deadlock-, topology-, and error-free network. To enable a reliable and long-transmission-distance (over 300 m) data communication, RHiNET-3/SW supports, hop-by-hop retransmission, credit-based flow control, and cut-through functions. Optical interconnection allows high-speed, long-transmission-distance data transmission. To achieve high-speed node-to-node interconnection, we implemented eight pairs of 1.25-Gbit/s x 12-channel optical interconnection modules, an 80-Gbit/s SW-LSI, and four DS-LSIs in a compact circuit board. RHiNET-3/SW has eight input and eight output optical data ports. The bandwidth of each port is bi-directional 10 Gbit/s (aggregate throughput of the switch is 80 Gbit/s). All electrical interfaces are composed of high-speed CMOS or LVDS. The DS-LSI has a skew compensation function for 10-bit parallel data channels and 8B10B encoding and decoding. Optical interconnection transmitter modules use 850-nm VCSEL and MMF ribbon. The structure and layout of the circuit board is optimized for high-speed (data rate: 1.25 Gbit/s per channel), high-density (using 784-pin BGA LSIs) implementation.

Acknowledgements

We are grateful for the assistance and advice of Kazuyoshi Satoh of the Device Development Center, Hitachi, Ltd., Yoshiteru Keikohin and Kozoh Oosugi of Hitachi Information Technology Co., Ltd., and Atsushi Takai and Atsushi Miura of the Telecommunication and Information Infrastructure Systems Group, Hitachi, Ltd.,

References

[1] T. Kudoh, J. Yamamoto, F. Sudoh, H. Amano, Y. Ishikawa, and M. Sato: "Memory based light weight communication architecture for local area distributed computing", Innovative architecture for future generation high-performance processors and systems, IEEE Computer Society Press, pp. 133-139, 1997.

[2] L.M. Ni, "Should Scalable Parallel Computers Support Efficient Hardware Multicast", Proceeding of 1995 Int'l Conference on Parallel Processing Workshop on Challenges for Parallel Processing, pp. 2-7, August 1995.

[3] T. Horie, H. Ishihara, T. Shimizu, and M. Ikesaka, "AP1000 Architecture and Performance of LU Decomposition", Proceedings of 1991 Int'l Conference on Parallel Processing, pp. 634-635, August 1991.

[4] http://www.myri.com/

[5] HIPPI-6400 working drafts, T11.1 maintenance drafts of ANSI NCITS.

[6] IEEE802.3 Higher Speed Study Group http://grouper.ieee.org/groups/802/3/10G_study/public/index.html

[7] H. Nishi, K. Tasho, T. Kudoh, and H. Amano, "RHiNET-1/SW: One-chip switch ASIC for a local area system network", Proc. COOL Chips III, poster 7, Apr. 2000.

[8] S. Nishimura, T. Kudoh, H. Nishi, J. Yamamoto, K. Harasawa, N. Matsudaira, S. Akutsu, K. Tasho, and H. Amano, "High-speed network switch RHiNET-2/SW and its implementation with optical interconnections", International Conference of Hot Interconnects 8, 31-38, Stanford U.S.A., August 2000.

[9] A. Takai, T. Kato, S. Yamashita, S. Hanatani, Y. Motegi, K. Ito, H. Abe, and H. Kodera, "200-Mb/s/ch 100-m Optical Subsystem Interconnections Using 8-Channel 1.3-um Laser Diode Arrays and Single-Mode Fiber Arrays", J. of Lightwave Technology 12, pp. 260-270, 1994.

[10] R. Ueno and S. Inasawa, H. Nishi, T. Kudoh, and H. Amano, "Flow control method in high speed transfer using optical interconnect", Proceedings of the IASTED International Symposia Applied Informatics, pp. 371-376, Feb. 2001.

[11] M. Katevenist, D. Serpanos, and P. Vatsolaki, "ATLAS I: A general-purpose, single-chip ATM switch with credit based flow control", International Conference of Hot Interconnects 6, 63-73, Stanford U.S.A., August 1998.

[12] M.D.Schroeder and others, "Autonet: A high-speed, self-configuring local area network using point-to-point links", DEC Technical Report, SRC 59, 1990.

[13] Jos Carlos Sancho, Antonio Robles, "Improving the Up*/Down* Routing Scheme for Networks of Workstations", Euro-Par 2000, pp. 882-889", 2000.

[14] http://www.zarlink.com/