IEEE 1355 HS-Links: Present Status and Future Prospects

C.R.Anderson[2,8], M.Boosten[1,7], R.W.Dobinson[1,2,6], S.Haas[1,2], R.Heeley[1], N.A.H.Madsen[1,6], B.Martin[1], J.Pech[1,4,5], D.A.Thornley[1,3], C.L.Ullod[4]

[1] *CERN, 1211 Geneva 23, Switzerland*
[2] *University of Liverpool, Liverpool, UK*
[3] *University of Kent, Canterbury, UK*
[4] *University of Zaragoza, Zaragoza, Spain*
[5] *Institute of Physics, Prague, Czech Republic*
[6] *RHBNC, University of London, UK*
[7] *Eindhoven University of Technology, Eindhoven, The Netherlands*
[8] *CCLRC, UK*

Abstract. IEEE 1355 HS-Links and their support devices have been investigated as part of the ESPRIT projects Macramé and Arches. A description of the HS-Link technology and initial experience with the RCube 8-way packet router and the Bullit HS-Link interface device are presented. A 64 node HS-Link switching network based using these devices is being constructed at CERN. We report on the the design and construction of the network testbed.

1 Introduction

Effective exploitation of multiple processors in a distributed computing environment relies on a low latency, high bandwidth, inter-processor communication network. The IEEE 1355 HS-Link technology allows such networks to be constructed. It can potentially also be applied in other applications which require high performance switching, such as LAN or WAN routers, ATM switches, or high performance data acquisition systems. We present initial experience with HS-Links and the associated silicon implementations. A 64 node HS-Link switching network based on these devices is currently being constructed at CERN. We report on the overall architecture of this network testbed and the design of the individual components.

The work presented here has been carried out within the framework of the European Union's ESPRIT[1] program as part of the OMI[2] Macramé[3] and ARCHES[4] projects.

2 IEEE 1355 HS-Link Technology

Two complementary high-speed serial link technologies have been developed within the framework of the OMI/HIC[5] Esprit project. They have been subsequently standardised and form the basis of the IEEE 1355 [1] standard:

- 100 MBaud Data-Strobe (DS) link

[1]European Strategic Program for Research and development in Information Technology
[2]Open Microprocessor Initiative
[3]Multiprocessor Architectures: Connectivity, Routers And Modelling Environment (ESPRIT project 8603)
[4]Application, Refinement and Consolidation of HIC, Exploiting Standards (ESPRIT project 20693)
[5]Heterogeneous InterConnect (ESPRIT project 7252)

- 1 GBaud High Speed (HS) link

The standard allows modular scalable interconnects to be constructed based on high-speed point-to-point links and switches. Using the lightweight protocols of IEEE 1355 these networks can provide a transparent transport layer for a range of higher level protocols.

The IEEE 1355 protocol stack defines four protocol layers: bit, character, exchange and packet layers. Characters are groups of consecutive bits which represent data or control information. The exchange layer controls the exchange of characters in order to ensure the proper functioning of a link. It includes functions such as link flow control and the link startup mechanism. A credit based flow control scheme is used which operates on a per link basis. This scheme ensures that no characters will be lost due to buffer overflow.

The devices associated with the HS-Link are fabricated with a standard 0.5 μm CMOS process technology. They achieve GBaud performance from a serial link macrocell that occupies only 1 mm^2 of silicon and consumes less than 300 mW [2]. This is achieved through the use of an innovative delay-locked loop (DLL) technique for the serialiser and deserialiser circuits. The use of the DLL simplifies clock recovery and enables a high level of integration by avoiding the need for high frequency on-chip phase-locked loop (PLL) circuits.

2.1 Physical Media

The HS-Link is a high-speed serial interconnect technology. The current implementations are specified to operate at serial line speeds between 700 MBaud and 1 GBaud. HS-Links are intended to be used as a high-speed interconnect between chips on a printed-circuit board, between printed circuit boards over a backplane or between racks using coaxial cable connections. The connection length for the specified coaxial cable is limited to 5 meters due to the cable attenuation. Transmission over single-mode and low-cost multimode fibre optic cable is also possible, the length of the link over which the full data rate can be sustained is then restricted by the amount of buffering for the flow-control protocol (see 2.4).

2.2 HS-Link Signals

HS-Links are designed for bidirectional point-to-point connections, each link consists of one signal in each direction. HS-Links use single-ended transmission over 50 Ohm matched impedance transmission lines. These can be implemented as controlled impedance tracks on a printed circuit board or as coaxial cables for connections between boards. The serial line signal is DC balanced, i.e. it has a constant DC component. This enables AC coupling to be used for improved common mode rejection and also allows a simple interface to standard fibre optic transceivers with pseudo ECL inputs and outputs, since the link uses a serial line swing of 800mV, which is centered between the positive power supply and ground.

2.3 Character Encoding

HS-Links use an 8B/12B DC balanced encoding scheme, where 8 bits of data are encoded into 12 code bits, i.e. the encoding overhead is 50%. Table 1 shows the encoding of the HS-Link characters. The clock recovery using the delay-locked loop scheme requires a positive going synchronisation pulse at the beginning of every character. Therefore a *start* and a *stop* bit need to be added to every byte in addition to the *parity* bit and the *inversion* bit, which is required to maintain DC balance. The data can be sent inverted to ensure that the disparity, i.e. the difference between the number of ones and zeroes, tends towards zero. The inversion bit indicates the polarity of the data bits in the character. The odd parity check will detect all single bit errors.

Start	Parity	Data/Inverted Data								Invert	Stop
1	P	D0	D1	D2	D3	D4	D4	D6	D7	I	0

Table 1: HS-Link Character Encoding

The encoding scheme allows the transmission of the 256 data values plus 126 control characters. 8 of these control characters are reserved for the low level link protocol. In order to ensure a continuous stream of characters, which is required to keep the receiver calibrated, *IDLE* characters are sent when no data is available. The *flow control character (FCC)* is used for the flow control mechanism. The *end-of-packet (EP)* character is used to terminate packets and can be replaced by the *exceptional end-of-packet (EEP)* character to indicate that an error has occurred. The other four control characters are used during the link startup and shutdown procedure.

2.4 Flow Control

IEEE 1355 links use a credit based flow control scheme that works on a per link basis. HS-Link use the same flow-control mechanism as DS-Links, the only difference being that the flow control unit size, or flit, has been increased from 8 to 32 characters.

2.5 Link Startup

When a device is started, the transmitter section will start to emit a continuous stream of IDLE characters. This generates a square wave signal with 50% duty-cycle at 1/12 of the Baud rate on the outgoing link. The receiver on the other side of the link will use this character frequency to calibrate the receiver delay-locked-loop. Following the link calibration there is an exchange of control characters to insure that both ends of the link have been started successfully.

After correct operation of the link has been established, each end sends a number of flow control characters corresponding to its receive buffer size. The current implementations have buffering for at least two flits, i.e. 64 characters.

2.6 Packet Format

Information in IEEE 1355 networks is transmitted in packets, a packet consists of a destination header, which is used to route the packet through a switching fabric, followed by the actual payload data, and an end-of-packet character. There are no restrictions on the packet size. Figure 1 illustrates the packet format.

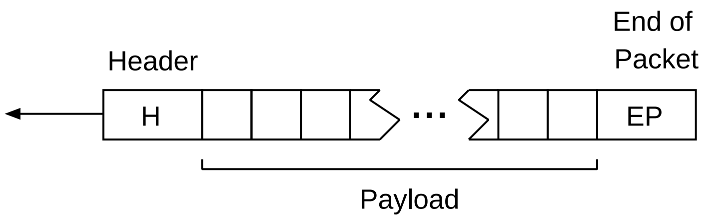

Figure 1: Packet Format

3 HS-Link Devices

The HS-Link technology can already be used to build scalable high-performance switching networks using the two supporting devices which are currently available:

- the Bullit chip provides a parallel interface to an HS-Link,
- the RCube is an 8-way crossbar switch.

3.1 *The Bullit HS-Link Interface Chip*

The Bullit chip [3] provides a parallel interface to an HS-Link. Figure 2 shows the block diagram of the device. It consists of a transmitter/receiver pair, FIFO buffers and a low level protocol engine.

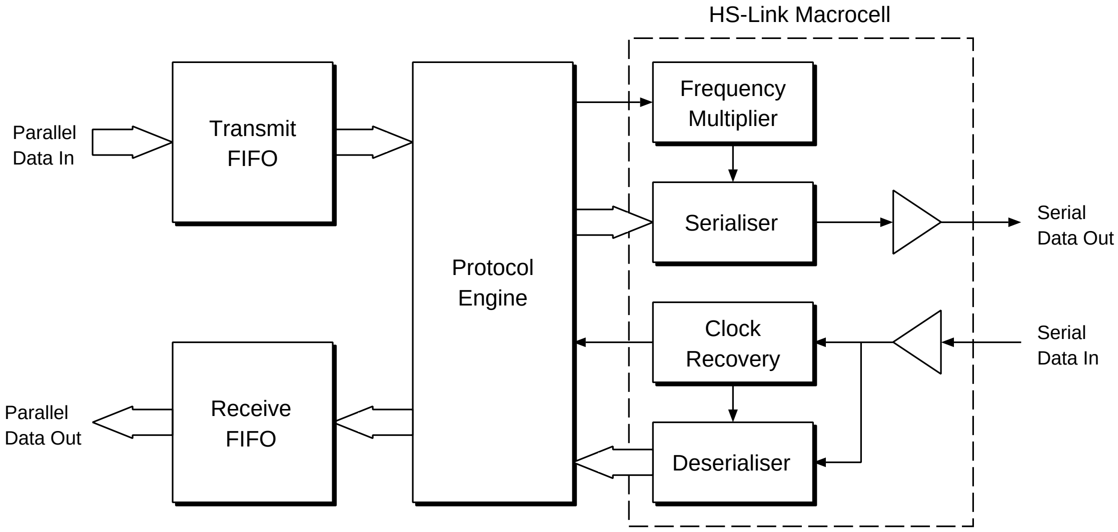

Figure 2: Block Diagram of the Bullit Chip

The device has internal 80 character deep input and output FIFOs. The access to these FIFOs is by separate two byte wide parallel interfaces. The protocol engine implements the low-level link protocol and performs functions such as 8B/12B encoding/decoding, flow control to avoid receiver FIFO overrun, automatic insertion and deletion of IDLE characters and acknowledged link startup and shutdown. The HS-Link macrocell performs transmit clock frequency multiplication, character clock recovery and character serialisation and deserialisation. The speed of the transmitter section is controlled by a single byte rate clock.

3.2 *The RCube 8-way HS-Link Router*

The RCube [4] is an 8x8 router for IEEE 1355 HS-Link networks. Its is based on an 8x8 non-blocking crossbar switch and 8 bidirectional 1 GBaud serial links. This results in a total cross-sectional data bandwidth of 640 Mbyte/s. The RCube uses “Wormhole Routing” [5], which allows packets of unlimited length to be routed. It also provides very low latency in lightly loaded networks, each RCube has a latency of only 150 ns. The device also provides adaptive routing which enables efficient load balancing in multistage networks, as demonstrated for the DS-Link technology [6].

Figure 3 shows a block diagram of the RCube. The chip is built from three different functional units. The High Speed Link macrocell includes the serialiser/deserialiser and provides the interface to the HS-Link. The Serial Macrocell Interface provides the character coding/encoding along with the link boot protocol and interfaces between the HS-Link macrocells and the router core. The parallel router core (R3P) operates with a character wide data path.

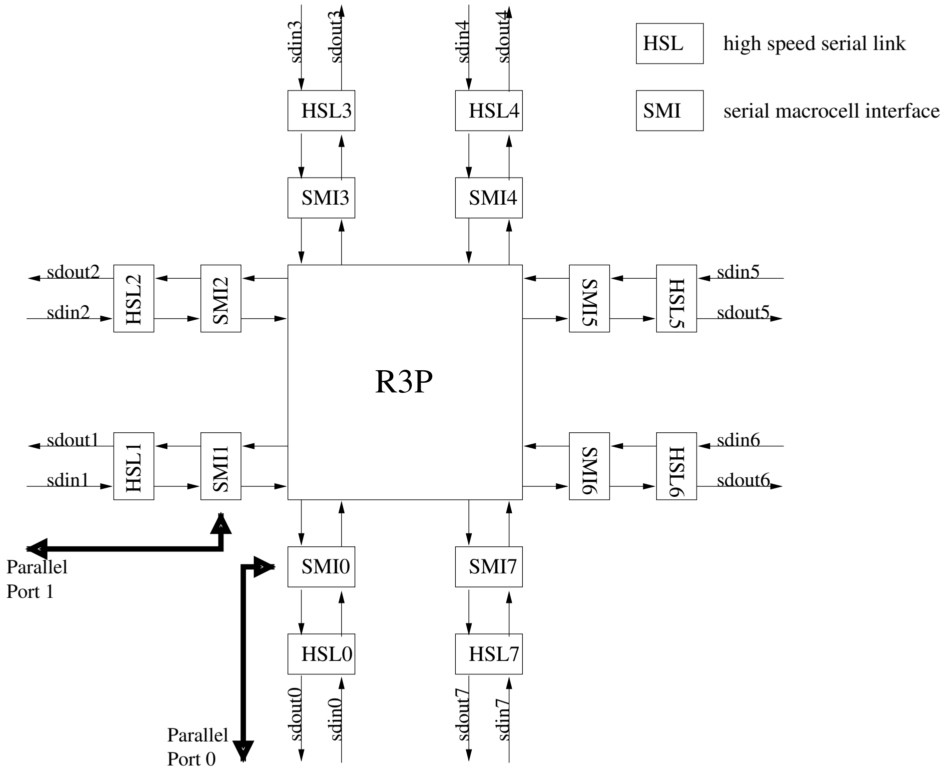

Figure 3: Block diagram of the RCube router

Two of the HS-Link interfaces can be bypassed, which allows for character wide parallel access to the router. This was implemented for device testing, but can also be used as a parallel interface to an HS-Link network. The parallel interface of the RCube does however not have the simple FIFO interface of the Bullit chip and requires careful design of the interface logic if this feature is to be used.

4 An HS-Link Evaluation Board

Initial work has been carried out to investigate the behaviour of the two silicon components described in section 3. An evaluation board containing a single 8-port RCube switch and two Bullit link adapters has been constructed. The reason for designing this board was to identify problems with the available RCube and Bullit chips and their interoperability as well as acquiring experience with the HS-Link technology. This board was designed such that long term tests at the full link data bandwidth could be carried out with different link clock frequencies, power supplies and cable types.

4.1 Functionality

The block diagram of the HS-Link evaluation board is shown in figure 4. A T8 microcontroller is used to initialise and monitor the status of the Bullit chips and of the RCube switch.

There are FIFO buffers connected directly to each Bullit so that data can be passed at the full link rate through the parallel port of these devices.

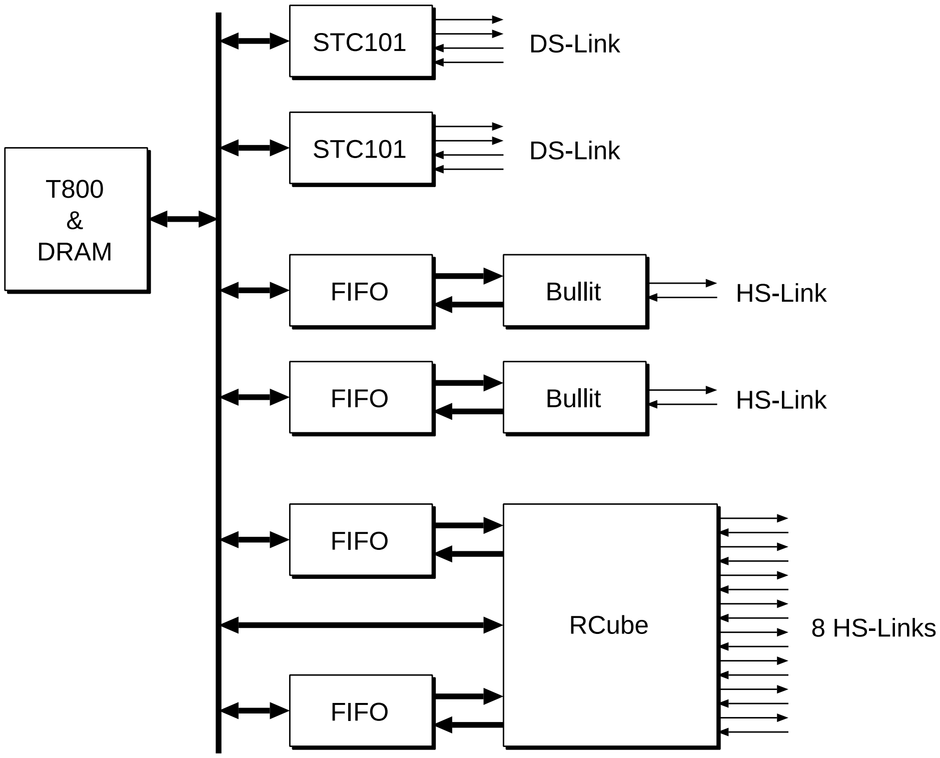

Figure 4: Block Diagram of the HS-Link Evaluation Board

The Bullit chips act as HS-Link data sources and sinks. Before a test begins, the processor loads the data to be transmitted into the associated FIFO buffer. After the transmission has finished the processor can then read the data from the FIFOs to check if any errors have occurred. There are different modes of operation: half-duplex and full-duplex burst data transmission, "ping-pong", i.e. transmission of data from the FIFO on one Bullit into the other one and back again, and continuous wrap-around, where data is circulated bidirectionally between the two Bullit devices.

All the HS-Links connections are made using micro coaxial cables. This allows for direct connections between Bullit chips or via the RCube switch. By looping the links of the RCube switch back onto itself and setting up the routing tables appropriately, data can be sent through all the links of the switch concurrently, even though there are only two data sources.

The STC101 devices have been incorporated to develop support for a DS-Link control network for configuration, control and monitoring of a network of RCube switches, as described in section 5 below.

4.2 Results from the HS-Link Evaluation Board

Long term, error free, full duplex transmission was established, utilising two RCube evaluation boards, with the Bullit chips operating at a clock frequency of 64 MHz and the RCube switches running with a 66 MHz and a 60 MHz clock respectively. All links of each RCube were in operation for this test and AC coupling was used between the boards. The measured data rate was 58 Mbyte/s. The latency for sending the data from one Bullit to the other traversing each RCube four times was $1.5\mu s$. The packet latency across the RCube has been measured to be 180 ns at 66 MHz, which corresponds to 12 clock cycles.

During longer term runs of over 48 hours more than 180 terabytes of data were transmitted without errors for a number of different combinations of cables, clocks and power supply voltages. This results in a bit error rate better than $5 \cdot 10^{-16}$. The main problems to overcome to achieve this result were related to the external control logic which has to handle asynchronous FIFO signals without any metastable conditions.

A few problems with the current HS-Link devices have been found during the testing: the Bullit keeps the EP character in its transmits FIFO until the next character is written into it, the necessity to take the RCube links out of reset synchronously and the HS-Link initialisation in a mixed Bullit and RCube system. Workarounds for these and other problems have been elaborated.

5 Construction of a 64-Node HS-Link and Switch Testbed

Results from large 100MBaud DS-Link switching networks have been previously presented and have demonstrated scalable performance for network of up to 1024 terminal nodes [7]. As part of the ARCHES project a 64 node 1 GBaud HS-Link testbed is being constructed.

5.1 Testbed Architecture

The RCube and Bullit chips are "technology enabling" devices designed to bring the advantages of low power, high speed serial technology to potential applications. To move from these devices to commercially successful implementations depends, among other things, on demonstrating that the technology meets the demands of all of the boundary conditions and how well it performs and scales under full-load conditions. To this end a large test and demonstration switching network of a minimum of 64 ports has been designed that not only will answer these questions but also provide a facility for exploiting the capacity of IEEE 1355 to successfully transport other protocols such as SCI, ATM and Ethernet.

Building on the experiences of the Macrame studies it was decided to base the switch architecture on the multistage Clos network topology while still allowing for the possibility of other topologies. The basic issue is one of packaging, e.g. how many router to router interconnects will be fixed on printed circuits and how many will be user configurable through cabled interconnects?

The choice was made to build a central switch Clos of 32 ports on one printed circuit module. Another module would be used to house four individual RCubes for which all the available links would be brought out to front panel cabling. An HS-Link traffic generator was designed to provide the switch with user controllable traffic patterns capable of operating at the link bandwidth. Therefore the testbed can be constructed from the following three basic modules:

- 4 x 8-way switch modules;
- 32-way switch modules;
- HS-Link network interface cards.

Figure 5 shows the full testbed employing these three modules to build a 64 port 5 stage Clos network. The 64 traffic nodes (T) source and sink data into the first stage of 16 switches which are housed four to a board. These in turn share the traffic between the two 32 way Clos switch boards. Not all of the links are shown for reasons of clarity. Note that for any traffic port to access any other port requires 1, 3 or at most 5 stages of switching.

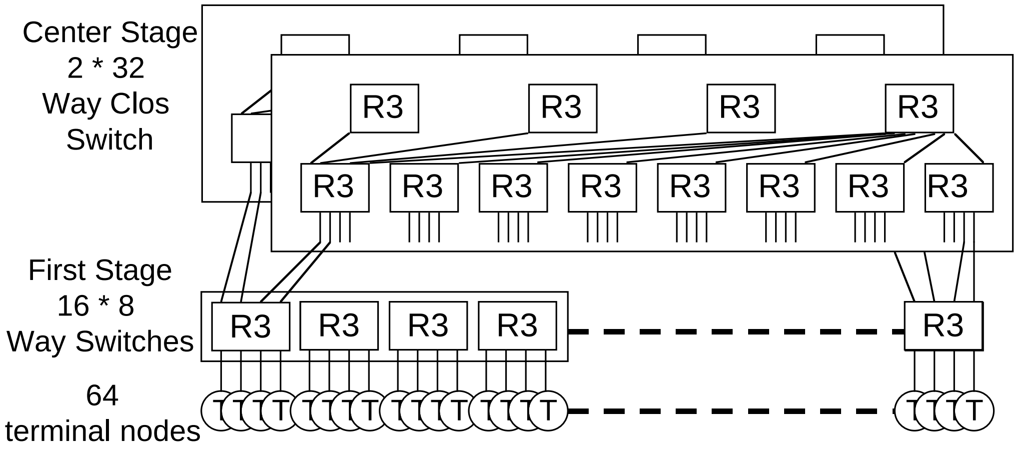

Figure 5: HS-Link Network testbed Architecture

5.2 Testbed Components

The design of the individual modules for the HS-Link network testbed is presented below.

5.2.1 The 4 x 8-way Switch

The 4 x 8-way switch module consists of 4 RCube switches with all their links brought to the front panel for connecting to other switch modules or network interface cards using coaxial cables. The block digram of the switch module is shown in figure 6. A T8 microcontroller is used to configure and monitor the RCube switches. IEEE 1355 DS-Links are used to to control the module. Alternatively RS232 can be used. Two connections are provided, which allows a daisy-chain or a star topology to be used for the control of several such modules.

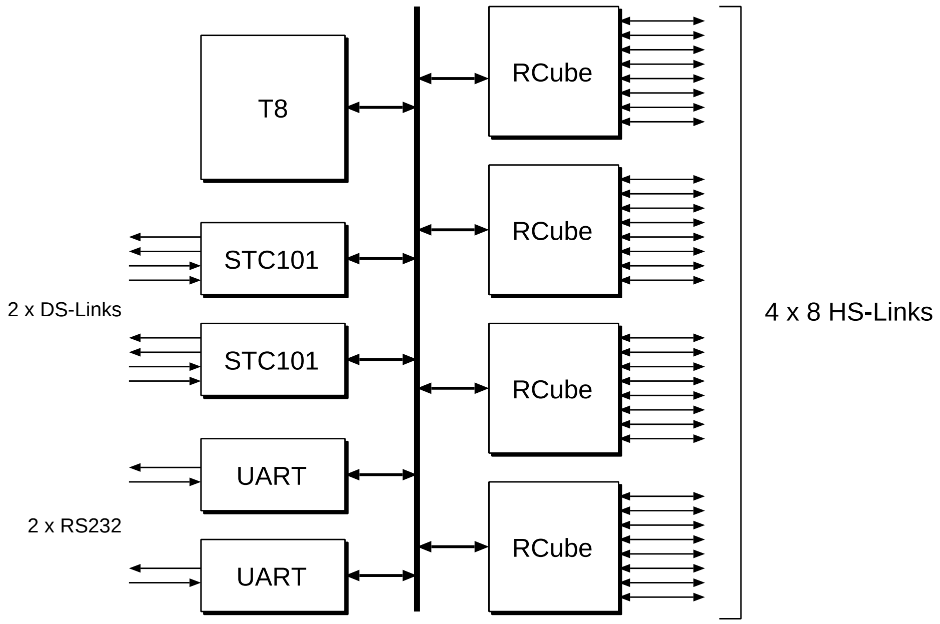

Figure 6: Block diagram of the 4 x 8-way switch

The module is built in 19” mechanics. A prototype of this module has been manufactured and tests have been carried out using a 66 MHz clock for the RCubes. Using the HS-Link evaluation board presented above as a data source, a continuous data stream has been transmitted bidirectionally through the four RCubes. The measured per link bidirectional throughput of 62 Mbyte/s corresponds to the expected value for the Bullit devices on the evaluation board running at 64 MHz.

A total of four of these modules will be built for the full 64-node testbed. They are required for the terminal stage switches in multistage networks and can also be used to construct grid and hypercube topologies.

The Network Description Language (NDL) has been extended to incorporate the RCube device. The low level software package, originally developed for use in configuring and monitoring the Macrame DS-Link network uses these extensions to allow for the control of heterogeneous networks of HS and DS-Links.

5.2.2 The 32-way Switch

The 32-way switch is implemented as three stage Clos network which consists of 8 terminal stage switches and four centre stage switches. The connections between the switches can be seen in figure 5. The module uses the same control structure as the 4x8-way switch module, i.e. a T8 microcontroller and DS-Links or RS232 to initialise and control the switch. The board will also be built in 19" mechanics and the printed circuit board layout is currently in its final stages.

5.2.3 The PCI HS-Link Interface

The HS-Link Network Interface Card can provide two different functions: it can be used as a traffic generator for the network testbed and it can also provide an interface to the PCI bus to allow processor access to the HS-Link. The second feature can be used either for control purposes or using high speed DMA access for processor to processor, high speed, low latency data communications.

The block diagram of the network interface is shown in figure 7. A PCI interface chip provides for user DMA, mailboxes, interrupts and bus interfacing. Glue logic implemented in a FPGA handles the multiplexing of this interface between two HS-Link channels. Each channel handles a HS-Link Bullit interface. State machines to read or write the Bullit input and output FIFOs are implemented in the control FPGA. The HS-Link packets can be transferred to or from the PCI bus directly thus fulfilling the processor to processor communication requirements.

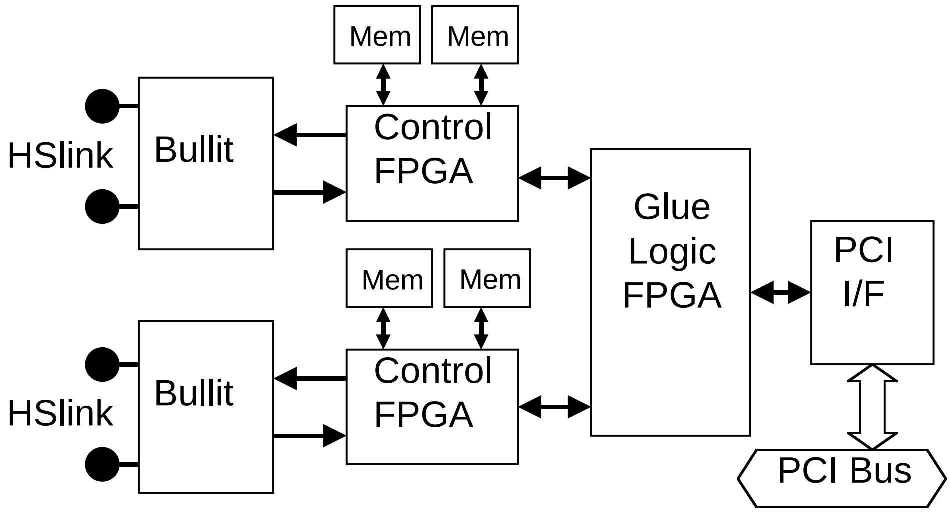

Figure 7: Block Diagram of the PCI HS-Link Interface

Each control FPGA has access to two memory banks. One for the transmitted data and one for the received data. To exercise the links at full bandwidth, data packet descriptors are stored in the transmit memory. These descriptors contains the packets destination address, its length and its required time of despatch. These variables are sufficient to fully define the traffic profile that will traverse the switching network.

State machines in the control FPGA interpret these descriptors and transmit the defined packet as requested. Packets are timestamped on reception, their delays calculated and the results stored in the reception memory which is regularly read out by a control processor across the PCI bus. Control software drivers have been developed for the Linux operating system. The prototype board is currently under test.

6 Current and Future HS-Link Developments

HS-Links are already successfully incorporated in commercial products such as the CCe System of parallel processors[6]. Several research institutions are exploring the communication possibilities of the technologies for use in general purpose multiprocessor networks such as the MPC project at the Universite de Paris Marie Curie (UPMC)[7] or machines specifically designed to run simulation programs such as the RTNN project at the Zaragoza University[8].

The packet level protocol of the HS-Links provides support for any number of higher level message passing protocols which can be implemented in hardware or software. Exploiting this property is done by the PCI-DDC [8] chip developed at UPMC. This chip[9] interfaces an RCube parallel port to the PCI bus and incorporates the DMA engines necessary for running the Direct Deposit Protocol between user space in applications running under FreeBSD and HS packets which can then be routed to any destination processor of choice.

For future technological improvements and extended applications there are a number of ongoing efforts. Research work is underway at UPMC to increase the speed of the HS-Link macrocell and enhance the functionality of the routing technology. Work within the ARCHES project is focussing on the use of HS switching fabrics to carry Gigabit Ethernet frames and demonstrations of feasibility are expected soon.

7 Conclusions

The program of work undertaken by the authors and reported here has shown that the first generation of IEEE 1355 HS-Link and switch devices deliver flow controlled data transmission close to the promised gigabaud line speed together with the required switching functionality. Further work is needed to establish and demonstrate that larger systems scale in both performance and reliability. Commercial interest has already been established and future market opportunities are being explored.

Acknowledgements

We are grateful for the support of the European Union through the Macramé (ESPRIT project 8603) and ARCHES (ESPRIT project 20693) projects. Prof. E. Gabathuler and Dr. M. Houlden from Liverpool University are thanked for their help and encouragement.

[6] http://www.parsytec.de/news/brochures/CC/hardware.html
[7] http://www-asim.lip6.fr/mpc/index.fr.html
[8] http://rtnn.unizar.es/project.html
[9] http://www-asim.lip6.fr/mpc/hard/pciddc/index.fr.html

References

[1] IEEE Std. 1355, *Standard for Heterogeneous Inter-Connect (HIC). Low Cost Low Latency Scalable Serial Interconnect for Parallel System Construction*, IEEE Inc., USA 1995.

[2] R. Marbot et al., "Integration of Multiple Bidirectional Point-to-Point Serial Links in the Gigabits per Second Range", Hot Interconnects Symposium, 1993.

[3] *The Bullit Data Sheet*, Version 2.0, Bull Serial Link Technology, 1995.

[4] *The Rcube Specification*, Version 1.7, Laboratoire MASI, Université de Pierre et Marie Curie, Paris, France, 1997.

[5] W.J. Dally and C.L. Seitz, "Deadlock-free message routing in multiprocessor interconnection networks", *IEEE Transactions on Computers*, vol. 36, no. 5, pp. 547–553, 1987.

[6] *Networks, Routers and Transputers*, edited by M.D. May, P.W. Thompson, P.H. Welch, ISBN 90 5199 129 0.

[7] S. Haas et al., "Results from the Macram 1024 Node IEEE 1355 Switching Network", EMMSEC97, European Multimedia, Microprocessor and Electronics Conference, Florence, Italy, 3-5th November 1997.

[8] *PCIDDC Data Sheet*, Version 1.4, Laboratoire MASI, Université de Pierre et Marie Curie, Paris, France, 1997.