A Gigabit Ethernet Link Source Card

R. E. Blair, J. W. Dawson, G. Drake, W. N. Haberichter, J. L. Schlereth

Argonne National Laboratory, Argonne, IL 60439 USA
reb@hep.anl.gov, jwd@hep.anl.gov, drake@anl.gov, wnh@hep.anl.gov, jls@hep.anl.gov

D. J. Francis,
CERN, 1211 Geneva 23, Switzerland
David.Francis@cern.ch

Abstract

A Link Source Card (LSC) has been developed which employs Gigabit Ethernet as the physical medium. The LSC is implemented as a mezzanine card compliant with the S-Link specifications, and is intended for use in development of the Region of Interest Builder (ROIB) in the Level 2 Trigger of ATLAS. The LSC will be used to bring Region of Interest Fragments from Level 1 Trigger elements to the ROIB, and to transfer compiled Region of Interest Records to Supervisor Processors. The card uses the LSI 8101/8104 Media Access Controller (MAC) [1] and the Agilent HDMP-1636 Transceiver. An Altera 10K50A FPGA [2] is configured to provide several state machines which perform all the tasks on the card, such as formulating the Ethernet header, read/write registers in the MAC, etc. An on-card static RAM provides storage for 512K S-Link words, and a FIFO provides 4K buffering of input S-Link words. The LSC has been tested in a setup where it transfers data to a NIC in the PCI bus of a PC.

I. History of the Gigabit Ethernet LSC

The Second Level Trigger of ATLAS is organized so that it executes the trigger algorithms only on data from Regions of Interest. This scheme has the advantage that the amount of event data which must be transferred on a First Level Accept to the Second Level Processors is held to a minimum, with resulting reduction of traffic on the Switching System. On a Level One Accept each Level 1 Trigger element forwards to the Second Level a Region of Interest (ROI) Fragment which identifies the location and nature of the ROI in that element. Currently there are thought to be 11 ROI Fragments to be transferred to Level 2 on each Level 1 Accept. These Fragments are received in the Second Level trigger by the Region of Interest Builder (ROIB), which accepts the ROI Fragments, builds ROI Records from the received Fragments, selects a target Supervisor Processor, and forwards the compiled ROI Record to the target processor. At an early point the decision was made to use S-Link as the medium to transfer Fragments into the ROIB and Records to the Supervisor Processors. At the time we began building prototype hardware for these tests, Link Source and Destination cards were not readily available, and accordingly various implementations were built by us using copper so that Test Beds and Integration programs could proceed.

The ROIB is composed of a number of 9U ROI Builder cards which receive ROI Fragment information in the form of raw Gigabit Ethernet Frames on fiber optic transceivers from as many as 12 Level 1 Trigger elements, and provide ROI Records in standard S-Link format as raw Gigabit Ethernet Frames through J3 to Transition Cards to as many as 16 Supervisor Processors. In addition to the ability to communicate via Gigabit Ethernet Frames, the ROI Builder cards communicate with each other via the VME back-plane, which facilitates the Allocation Algorithm.

The ROIB also utilizes 9U ROI Input cards, each of which has 6 standard S-Link connectors which may receive standard S-Link Link Destination Cards so that any physical medium may be used to import ROI Fragments from Level 1, and provide fan-out in Gigabit Ethernet Frames to fiber optic transceivers. When a Fragment is received in standard S-Link format on an Input Card, the fragment is first written to a FIFO 4K words deep to provide time buffering. When a FIFO is non-empty the Fragments may be read, translated into Gigabit Ethernet Frames, fanned out to 4 Gigabit Ethernet optic transceivers, and relayed in parallel via fiber to as many as 4 ROI Builder cards.

It was obvious that Gigabit Ethernet was an ideal medium for this application for many reasons. Gigabit Ethernet has excellent commercial support both in hardware and software, it lends itself to a COTS approach to hardware realization, and it has the bandwidth to support the 100 KHz Level 1 Accept rate which was our requirement. When we undertook to develop a Gigabit Ethernet LSC, shown in Fig. 1, there was a thought popular in ATLAS that it was advantageous to store the global event data on a Level 1 Accept at the Readout Drivers rather than to transfer all the data to the Readout Buffers on each Level 1 Accept. Since 99% of the event data is overwritten after Level 2 Rejects, traffic is reduced considerably by leaving the event data at the readout drivers, and forwarding it to the event builder only when Level 2 accepts the event. In this scheme the only data to be routinely transferred on a Level 1 Accept would be the ROI Fragments and the event data relating to the Regions of Interest for processing by the Level 2 Processors. The Gigabit Ethernet

LSC was designed with these functions in mind, and it was thought that these prototypes could be used for testbed studies and integration involving Readout Drivers.

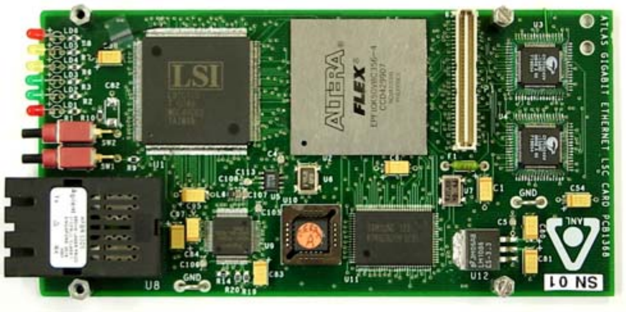

Fig. 1 Gigabit Ethernet LSC.

A. Architecture of the Gigabit Ethernet LSC

Because it was initially foreseen that these cards might store event data, a large synchronous RAM was provided which is 512K words deep. It was thought that event data would be written in RAM, with a table maintained in the FPGA containing event ID, event word count, and starting address in the synchronous RAM. When ROI or event data was requested, a Gigabit Ethernet frame would be received requesting the data with reference to an event number, and a state machine in the FPGA would handle the request. Accordingly, the hardware supports full duplex operation of Gigabit Ethernet. This mode of operation is not currently foreseen for the ATLAS Detector.

As shown in Fig. 2, the Gigabit Ethernet LSC is organized around an Altera 10K50A FPGA. A number of state machines are implemented in the FPGA configuration, and everything that happens on the card is managed by one or another state machine in the FPGA. Because these state machines are implemented in configurable logic, the functioning of the card may be tailored for various different applications. One configuration which has undergone considerable testing is as follows: Records in S-Link format are received via the connector, and written to FIFO as 32 bit words along with the control bit. The card expects that input records will be in the standard ATLAS format with control word header and trailer, and when a record has been received in the FIFO it is read from the FIFO by a state machine and written directly in S-Link format to the synchronous RAM while counting the words. When the complete record has been written to RAM, another state machine generates the Ethernet header with word count, and transfers the header followed by the S-Link record from RAM to the Transmit FIFO of the MAC for transmission via Gigabit Ethernet also providing the Start of Frame and End of Frame Markers.

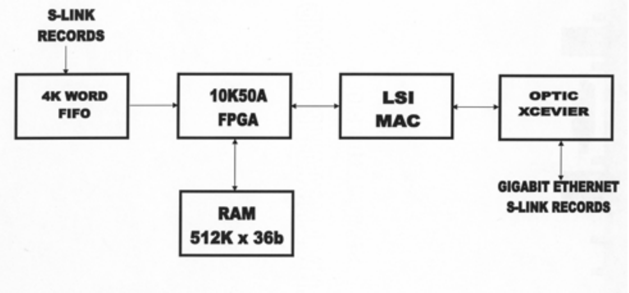

Figure 2: Block diagram of Gigabit Ethernet LSC.

Another configuration which was tested for some time, was where the S-Link words are transferred immediately as soon as the input FIFO is non-empty to a FIFO implemented in the Embedded Array Block (EAB) of the FPGA while the S-Link words are being counted, and then transferred to the Transmit FIFO of the MAC, again controlled by a state machine which formulates the header, embeds the word count, and provides the Start of Frame and End of Frame Markers.

Other state machines in the FPGA initialize registers in the MAC, request Autonegotiation if the link is not up, etc. Since the configuration of the MAC is widely variable depending on register contents, and all registers of the MAC can be written and/or read by the FPGA, the operational configuration of the card is extremely flexible. Flow Control at the S-Link input can be raised by a logical combination of the Almost Full of the input FIFO and/or the Watermark of the Transmit FIFO in the MAC. Flow control in Gigabit Ethernet is accomplished by the use of Pause Control Frames.

B. Operation of the Gigabit Ethernet LSC

Drivers to run under Linux supporting the operation of the Gigabit Ethernet LSC were written at Argonne, and the card was extensively tested in a number of configurations at Argonne. S-Link input records were provided by a CES Rio2, and the S-Link records transferred via Gigabit Ethernet were received by a NIC in the PCI bus of a PC. Records received in the PC were checked for errors against those transferred from the Rio for a large variety of data types, and operation was found to be satisfactory. Throughput for the card in a number of configurations was measured and found to be substantially in agreement with calculations.

Gigabit Ethernet LSC cards were subsequently operated in test beds at CERN, both by the ROS Group within ATLAS, and by the Electronics Group within LHCb. Results of the tests made by these two groups are available[3, 4].

Our current efforts relating to Gigabit Ethernet focus on the design and development of a Gigabit Ethernet LSC which uses copper as the physical medium in accordance with 1000 Base-T. A prototype run has been fabricated and is being assembled, and we expect to begin testing these prototype cards shortly.

C. Summary

A Link Source Card which accepts standard S-Link records, reformats them as Gigabit Ethernet Frames, and transfers them via Gigabit Ethernet has been built and tested both at Argonne and at CERN. A prototype run of these cards exists, and we would be agreeable to furnish one or more cards to groups who would like to use them in test beds or in prototype tests. A version of the Gigabit Ethernet LSC utilizing copper as the physical medium has been designed, and a prototype run is being assembled. After testing we would also be agreeable to furnish one or more of these copper Gigabit Ethernet cards to interested parties.

D. Acknowledgements

The authors with to acknowledge the advice and assistance of Stefan Haas of CERN who shared his extensive experience with Gigabit Ethernet, and Beniamino Di Girolamo of the ROS Group of ATLAS, who performed a number or measurements of the Gigabit Ethernet LSC at CERN to investigate the applicability to the ROS. Beat Jost, Richard Jacobsson, Niko Neufeld, and Joaquim Neves of LHCb deserve special thanks for their suggestions, simulation, and testing of the hardware.

II. References

[1] 8101/8104 Gigabit Ethernet Controller Technical Manual, LSI Corp.

[2] Altera Flex 10K Applications Guide.

[3] Presentation by Beniamino Di Girolamo to ATLAS ROS Group. Private communication 25 March 2002

[4] Presentation by Beat Jost to LHCb Electronics Meeting, 27 May 2002