DS–Link over Fiber: A High Speed Interconnect for Cluster Computing

Yogindra Abhyankar, Anil Degwekar and Abhay Karandikar
Centre for Development of Advanced Computing,
Pune University Campus,
Pune 411 007, INDIA.
Email: {yogindra,aad,karandi}@parcom.ernet.in

Abstract

Recently, there has been a considerable upsurge in cluster based computing. Centre for Development of Advanced Computing, at Pune, India is also offering a cluster computing solution based on its proprietary network[1]*. This network is built around a high speed interconnect which uses DS–link*[2] *protocol, a part of IEEE 1355 standard. One of the disadvantages that prevents DS links to operate in a geographically distributed computing environment is its distance limitation. In this paper, we have sought to remove this limitation by suggesting a scheme for implementing DS links over optical fibers. This scheme has been designed and implemented by us. Here, we discuss the salient features of our scheme. Our preliminary studies indicate that a cluster computing solution using this extended link will be an attractive solution.*

1 Inroduction

In recent years, there has been a considerable interest in high performance computing. Massively Parallel Processing (MPP) machines with various communication architectures provide a necessary high performance scalable environment for scientific and engineering problems. The last few years have witnessed a notable change in the definition of parallel computing. With the availability of very high performance desktop workstations from several vendors, cluster computing provides an alternative to MPPs as high performance computing platform. Cluster computing creates a virtual parallel platform by connecting various workstations through a high speed interconnect. The performance of cluster computing critically depends upon the communication among workstations which in turn depends upon the network links and the architecture employed.

Distributed cluster computing on standard networks like Ethernet, FDDI and even ATM networks have been studied by various authors [2]. We, at C–DAC, are providing a cluster computing solution on a proprietary communication network architecture [1]. This architecture is built around a high speed, low latency, wormhole packet switch router which makes use of DS–link protocol. The DS–link protocol is a part of the IEEE standard for Heterogeneous InterConnect (IEEE 1355). Each DS–link can operate upto 100 Mb/s providing a bidirectional bandwidth of 19 MB/s. The link protocol supports virtual channels and dynamic message routing [3]. However, the DS–links, which were originally designed for transputer connections by INMOS, restricts the distance only upto 30 meters. This severly limits the network to operate in a geographically coupled distributed network computing environment.

In this paper, we address the issue of removing this limitation of DS–link protocol. Our main concern is to provide a cost effective solution to this problem. Specifically, we suggest a scheme for implementing DS–links over optical fibers. This scheme is expected to greatly enhance the performance of the link and our preliminary studies have indicated that a network built around these modified high speed links will provide a competitive solution for network computing.

We begin this paper by first discussing the DS–link and its protocol. We then discuss the limitations of the DS–link. Section 3 outlines the DS link over Fiber protocol. In Section 4, we discuss a scheme for implementing the extended link. Finally, we summarize this paper and indicate some directions for deploying this extended link in a network computing environment.

2 DS–Link Protocol

DS–links were originally introduced by INMOS for providing connectivity among their transputers. These links use a protocol with two wires in each direction. One is called D line and the other is called S line. The D line carries data

[1] More information is available on request from Business Division, C–DAC, Pune 411 007, India. The work reported in this paper is funded by Department of Electronics, Govt. of India.

[2] DS–link is a trademark of SGS-Thomson, UK.

bits, a control bit to distinguish between data and control tokens and a parity bit. The protocol is essentially a bit level protocol. The S line is a strobe signal that changes its level every bit time when the D signal does not change. This effectively provides a Gray code between D and S signals. One advantage of gray level encoding is that the receivers can receive data at whatever speed it is sent. This is made possible because the received data can be decoded from the sequence of D and S [3]. Moreover, because of this particular coding, the signal can be received without a phase locked loop. In the following subsections, we briefly discuss about the link format and low level flow control.

2.1 DS–Link Format

As mentioned above, each DS pair carries tokens and an encoded clock [3]. The format of data and control tokens are shown in Figure 1. Data tokens are 10 bits long. These 10 bits contain 8 bits of data, a parity bit and a flag which is set to 0 to indicate a data token. Control tokens are 4 bits long. These 4 bits contain a parity bit, a flag which is set to 1 to indicate a control token and 2 bits to identify the type of control token. The parity bit in any token is calculated over the parity of the data or control bits in the previous token. Various control tokens are End of Packet (EOP), End of Message (EOM), Flow Control Token (FCT) and Null Token. The coding of the control tokens is given in [3]

2.2 DS–Link Flow Control

The DS–Link protocol employs a token level flow control [3]. In this flow control protocol, each receiving link–input contains a buffer for atleast 8 tokens. Whenever the link input has sufficient buffer space available to receive further 8 tokens, a flow control token (FCT) is transmitted on the associated link output. The sender then transmits 8 tokens and waits for another FCT before sending any additional data.

2.3 Limitations of DS Link

The DS links were primarily meant for connecting chips on the same PCB or on different PCB, but in the same box. In order to exploit the capability of a DS link switch router for parallel computing applications, it may be necessary to bring the links outside the box for longer distance connections. The distance can be extended by using buffers and it has been claimed by INMOS that it is possible to work with DS-links upto 30 meters by using differential buffers.

However, for still larger distances, the skew of the cable for D and S signals in each direction is a major issue of concern. This problem can be successfully alleviated by isolation between the two ends of the connection. In this paper, as suggested in [5], we propose to use optical fibers to provide such necessary isolation together with an interface. This interface provides the buffering and electrical to optical conversion and vice versa.

3 DS–Link over Fiber Protocol

Note that the fiber medium is a single fiber connection and in order to carry DS links over fiber, both D and S needs to be transmitted as a single signal by appropriate encoding [5]. Moreover, the single connection should include sufficient transitions for easy recovery of the clock. The scheme that we have proposed here, adopts a commercially available Parallel–DS link adaptor[3]. This Parallel–DS link adaptor converts the DS link signals into 8 bit parallel data (and vice-versa).

After the Parallel DS link adaptor, 8B:10B line encoding is used (See the next section for details). In this encoding, each byte of data is converted to a 10 bit "field". Thus control tokens EOP and EOM are also 10 bit wide. The DS link protocol as described above, operates between the Parallel–DS link adaptor and the DS link Transmit/Receive engines. However, we have defined our own protocol over the fiber. This protocol does not use signals like FCT and NUL. In the following section, we describe this flow control procedure.

3.1 Flow Control for DS link over Fiber

The token level flow control employed by the DS link protocol works well for smaller distances. However, at the data rate of 10 MB/s, 1 Km of optical fiber can hold nearly 50 bytes of data. The "bit length" of the link in this case is very close to the packet length and hence the original permit scheme described above does not work efficiently.

In the flow control mechanism adopted by us, we have defined two flow control signals, called 'XON' and 'XOFF'. When the destination end is ready to receive data, it sends an XON signal to the intended sending end. The sender can then start sending data. When the destination buffer reaches a certain threshold, it sends an XOFF signal to the sending transmitter. After receiving XOFF signal, the sender stops sending any data till it receives another XON. Note that due to the propogation delay, the transmitter might have actually pumped some data into the link before it receives an XOFF indication. Therefore, the size of the destination buffer and the threshold value is to be chosen carefully so that the data

[3] The adaptor used is ST C101 available from SGS–Thomson. See [4].

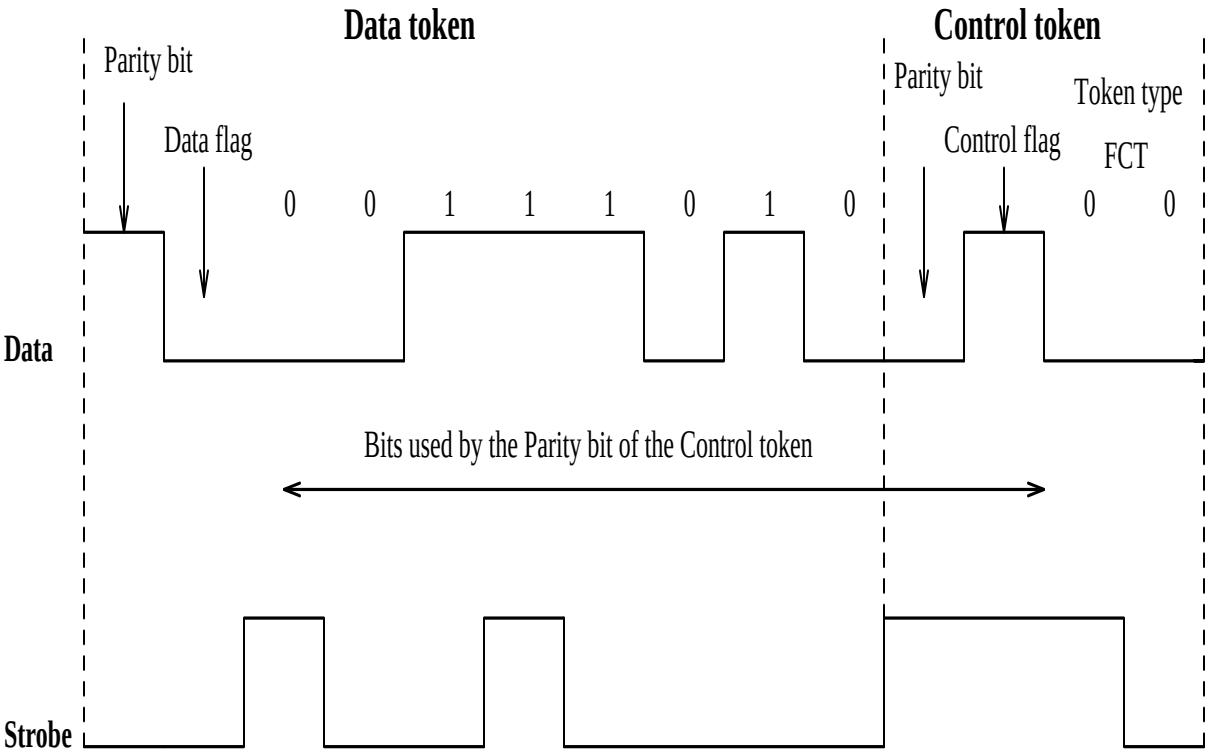

Figure 1: DS–Link Format (See [3])

already in transit may not cause any buffer overflow at the receiver.

4 DS–Fiber Interface Design

In this section, we describe the design of the interface that allows DS link communication between two hosts over the optical fibers. As shown in Figure 2, the interface has DS signals on one end and optical signals over fibers at the other end. The length of the fiber connecting the interface cards could be well extended upto 2 kms. A more detailed block diagram of the interface card is shown in Figure 3. It consists of seven major blocks:– DS–Link Adaptor, Control logic, Microcontroller, Transmitter/Receiver, Fiber–Optic Transceiver and FIFO.

The DS–Link adaptor is a high speed parallel to serial DS link (and vice–versa) convertor. It has two independent transmit and receive channels, each with 64 bytes of FIFOs for optimized packet processing. It converts the serial DS inputs from the host into a 8 bit parallel bus. This multiplexed bus may carry data or control information. The distinction is flagged separately by the adaptor. The parallel transmit bus serves as one of the inputs to the control logic.

The control logic consists of Transmit, Receive and Master control blocks. A dedicated hardware portion of this logic handles EOP and EOM tokens, while the XON, XOFF and other tokens are handled in software. Since EOP and EOM occur more frequently during transmission, their hardware processing is justified over software processing that may add additional latency.

A Transmitter operating at 160 to 330 Mb/s, takes in signals through the 8 bit bus from the transmit control logic and encodes it into 10 bits. The output of this transmitter is a serial data stream, shifted out of the pseudo ECL serial port at 200 Mb/s. When there is no data at the transmitter input, it sends out a special character (SYNC) that allows to maintain the link synchronization between the transmitter and the receiver.

The differential bit stream is fed to a fiber optic 1300 nm LED transceiver. The transceiver used in our design is HFBR 5105 from HP. Other equivalent transceivers may also be used. This transceiver has a duplex SC interface. The transceiver generates shaped optical output in response to the input signal from the transmitter.

The receiver accepts the serial bit stream at its differential line input and using an integrated PLL clock synchronizer recovers the timing information. The serial data is converted

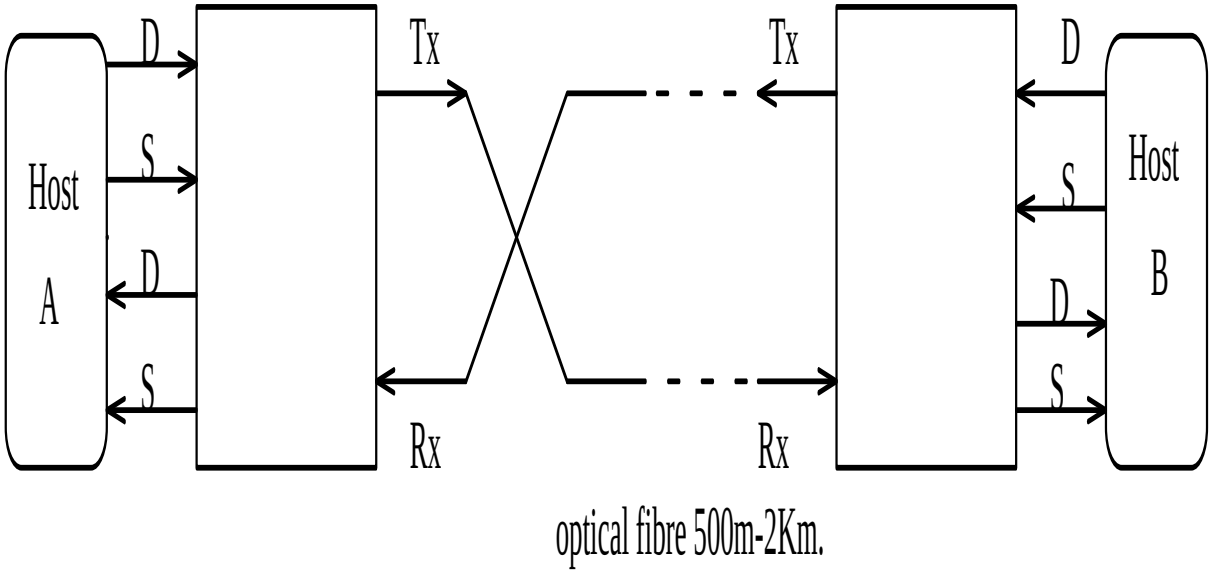

Figure 2: DS Link over Fiber

into parallel 10 bit data and then decoded into a byte.

A FIFO buffering is provided between the receiver and the DS link adaptor. This is especially useful when the receiving host is busy and not able to pickup the received data immediately. As mentioned in Section 3.1, the size of the FIFO is kept proportional to the length of the fiber. We recommend a FIFO of 512 bytes for 2 Km fiber. This FIFO is used in series with the internal 64 byte receive channel FIFO of the DS link adaptor. The FIFO is controlled by the receive and master control logic of the control block. A 8 bit microcontroller is used to initialize the DS link adaptor, handle interrupts and to provide a RS-232 interface. The controller also processes various flow control tokens like XON, XOFF *etc.*

The interface card has provision for local loop back testing at the DS link adaptor level or at the transmitter/receiver level. Necessary filtering for high speed components is also provided on the card. The fiber used in our prototype is a duplex cable with mulimode fibers and SC connectors.

4.1 Interface Card on Start Up

The interface card, after power on, sends XON token to the corresponding card at the other end and waits for an XONACK to be received. This initial handshaking is required to ensure that other card is also powered up. After this procedure, the DS link on the card is initialized and the DS link adaptor on the card starts responding to signals from the host side.

5 Conclusions

In this paper, we have suggested a scheme for implementing DS links over fibers. A prototype of this scheme has been designed by us. This extended link can be deployed in cluster computing applications. One such possible scheme can be implemented using a DS link switch hub. The DS link switch hub can be built around a low latency wormhole router available from INMOS. Commercially available workstations from SUN, Digital *etc.* can be connected through communication adaptor cards (similar to those which have been used in PARAM 9000 [1]) alongwith this interface card. As a future activity, it is planned to put all logic in a field programmable gate array technology. This will make the card more compact.

Acknowledgements

This project is funded by Department of Electronics (DoE), Government of India under its Fiber Optics Systems and Products Project. We gratefully acknowledege Dr. AK Chakravarty, Senior Director, DoE and Dr. VP Bhatkar, Executive Director, C–DAC, Pune for their constant support. The contributions of our colleague Mr. Praveen Shekokar are also acknowledged.

References

[1] C–DAC, Pune. *PARAM9000 Product Brochure*, 1995.

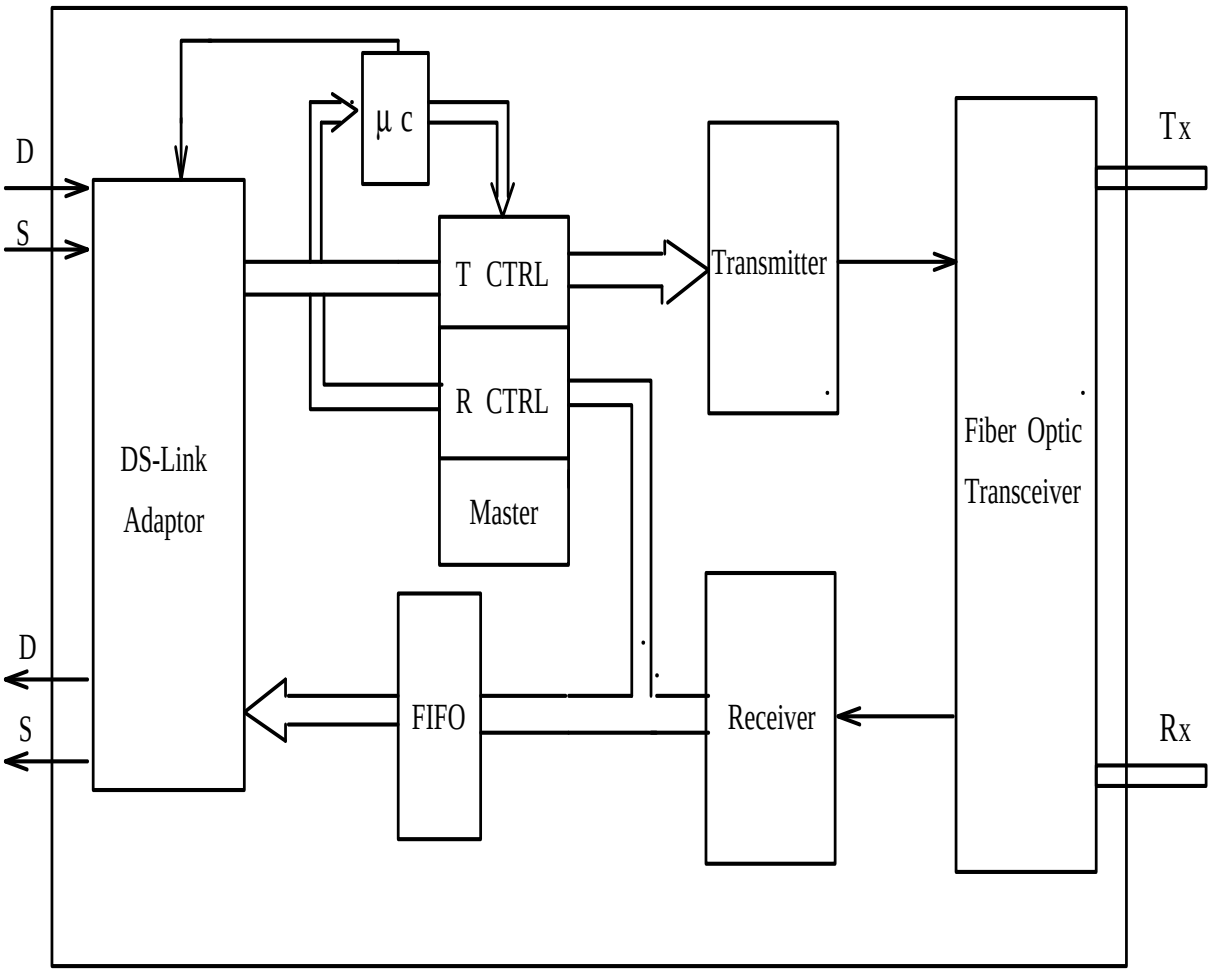

Figure 3: DS–Fiber Interface

[2] R. Fatoohi and S. Weeratunga. Performance evaluation of three distributed computing environments. In *Supercomputing94*, pages 400–409, November 1994.

[3] SGS–Thomson Microelectronics Group. *The T9000 Transputer Hardware Reference Manual*, first edition, 1993.

[4] SGS–Thomson Microelectronics Group. *ST C101 Data Sheet*, June 1994.

[5] P. Walker. Connecting 100 Mbaud T9000 transputer links. Technical Note 70, SGS–Thomson Microelectronics Group, April 1991.