A SERIAL LINK FOR STORAGE SUBSYSTEMS

15 November 1991

I D Judd

IBM (UK) Ltd
Mail Point 200
Hursley Park
Winchester
England SO21 2JN
Tel: 0703-701421
Fax: 0703-705106

A Serial Link for Storage Subsystems

I. D. Judd, IBM (UK) Ltd., Hursley Park, Winchester.

Serial interfaces offer several advantages for the attachment of input/output devices to computers and work-stations. They can improve performance by providing full-duplex communication with frame multiplexing. These features minimise overheads and queuing delays, particularly if an interface is shared by several devices. By comparison with traditional parallel interfaces serial links require smaller cables and connectors and they consume less power. These aspects are increasingly important because of the trend to small form-factor devices, eg. 2.5″ hard disk drives. Finally point-to-point serial links can provide better reliability and serviceability than a multi-drop parallel interface.

IBM 9333 Disk Drive Subsystem

The IBM[1] 9333 High-Performance Disk Drive Subsystem (IBM 9333) for the RISC System/6000 uses serial links for all of its internal interfaces. It is available in both rack-mounted and desk-side enclosures.

As shown in Figure 1, each IBM 9333 enclosure contains four 857 MB hard disk drives, a shared controller card and a power supply. Each disk drive is internally connected to the controller card via a dedicated serial link. In turn the controller card attaches to the using system via a serial interface cable and a MicroChannel[1] adapter card. For improved availability the controller can also be attached to a second adapter in an alternate system. Each MicroChannel adapter card can attach up to 4 IBM 9333 enclosures for a total of 16 disk drives per subsystem. Finally, a RISC System/6000 can attach up to 4 subsystems.

All of the serial links have a common transport layer that provides full-duplex communication with frame multiplexing. The link provides a data rate of 8 Mbytes/s each way using a twisted-pair cable that can be up to 10 Metres long.

For software compatibility the IBM 9333 provides the using system with a SCSI-2 command set and queuing model. However the serial disk drives use a lower-level 'order' set.

Division of function

The attachment function is divided between the subsystem components as follows:

- The adapter card fetches SCSI commands from system memory and it forwards each command to a controller card over the appropriate serial link. When instructed by messages from the controllers the adapter transfers read or write data between the serial links and system memory. At the end of each command the controller returns SCSI status and the adapter raises an interrupt to present it to the system.

- The controller card queues the SCSI commands and executes them by issuing low-level orders to the attached disk drives. The controller has a 1 MB data buffer to prefetch write data and buffer read data. The data buffer also provides a segmented read-ahead cache for each disk drive. The controller is responsible for recovering any disk errors.

[1] Trademark of IBM Corporation

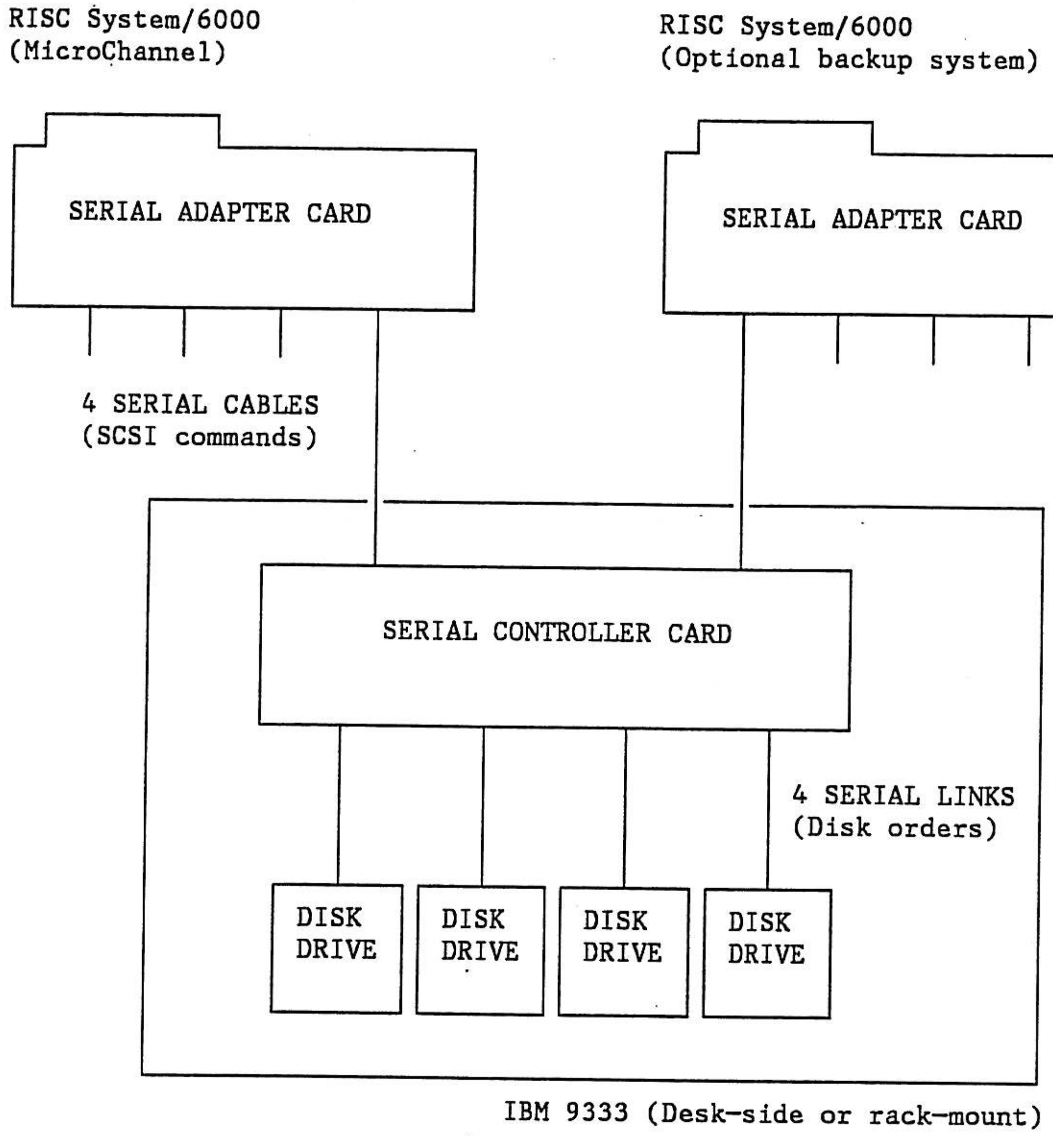


Figure 1. The IBM 9333 Disk Drive Subsystem

- The serial disk drive manages the format of gaps, header fields and data fields on each track. It can read or write multiple sectors with a single order. There is only limited (frame) buffering of the read and write data in the disk drive.

Performance

The IBM 9333 provides excellent performance in disk-intensive applications. Many of the performance benefits are due to the design of the serial interfaces.

The disk orders optimise performance by minimising the overall microcode overhead and providing tight control of the device. Since the controller already provides command queuing, a SCSI command set and data buffering these functions are not duplicated in the disk drive. The orders allow sequential back-to-back writes to be performed on the same revolution. They also support split reads and split writes. This feature reduces the effective rotational latency by allowing data transfer to start anywhere in the requested range of sectors and not just at the first sector. Finally the controller can terminate read-ahead immediately if another command arrives from the RISC System/6000.

The serial link between the adapter and a controller allows commands, data and status to be multiplexed frame-by-frame on behalf of the 4 attached disk drives. Full-duplex communication supports simultaneous transfer of both read and write data. In the case of a split read data can be stored out of order directly into system memory.

Serial interfaces

This section describes the common transport layer that is used by both the adapter-controller link and the controller-disk link. It also explains the different upper-level protocols that are used by the two links.

Physical and electrical characteristics

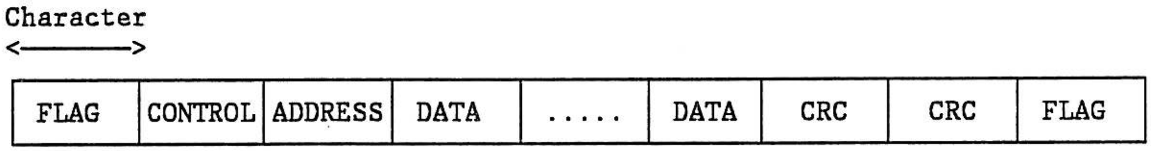
The link uses a dual 100-ohm twisted-pair cable that can be up to 10 Metres in length. One twisted-pair is used for communication in each direction.

The line driver is a 10 mA differential current sink with a common mode range of +/- 2V to accommodate ground shift. The line receiver is a differential comparator with hysteresis. Both the driver and receiver incorporate comparators to detect invalid voltages that would indicate a line fault.

The link operates with synchronous clocking at 80 Mbits/s. Each data byte is encoded into a 10-bit data character using a 4B/5B code and transmitted using non-return-to-zero-inverted (NRZI) modulation. Four additional 10-bit characters are defined for protocol functions. This encoding guarantees sufficient transitions for clock recovery at the receiver and it also provides a measure of DC balance.

Frame format

The unit of communication on the link is a frame, as shown in Figure 2.



FLAG	CONTROL	ADDRESS	DATA	.....	DATA	CRC	CRC	FLAG

Figure 2. Format of a serial link frame

Each frame is divided into 10-bit characters as follows:

- The FLAG protocol character delimits the start and end of a frame. The trailing FLAG of one frame can also be the leading FLAG of the next frame. FLAG characters are sent continuously when the link is idle to maintain byte synchronisation.

- The control field is a single data character that is managed by the transport layer. It contains a 2-bit frame sequence number to detect lost frames. The control field can also specify a Link_reset for error recovery or a Total_reset to initialise the remote node.

- The address field is a single data character that specifies the destination of the frame in the remote node. It may select a hardware DMA channel (for data) or interrupt the microprocessor (for messages).

- The data field contains a variable number of data characters up to some maximum determined by the size of the implemented frame buffers. The IBM 9333 uses a maximum data field length of 128 characters on the adapter-controller link and 32 characters on the controller-disk link.

The data field is managed entirely by the upper-level protocol. It may contain a message (eg. a command or status) or customer data which is stored on a disk drive.

- The Cyclic Redundancy Check (CRC) consists of 2 data characters which cover the control, address and data fields.

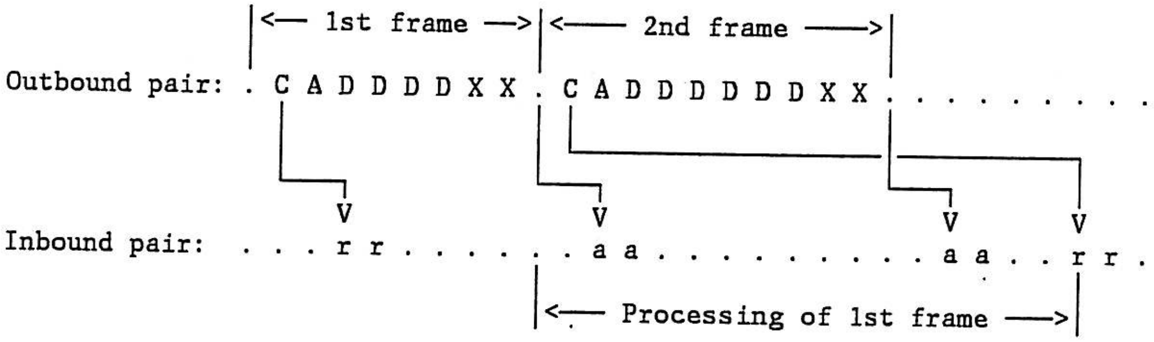
Protocol

When a node transmits a frame it expects to receive 2 responses from the remote node:

- An 'Acknowledgement', which is a pair of consecutive ACK protocol characters. The Acknowledgement indicates that the frame was received without error and it is returned shortly after the trailing FLAG of the corresponding frame.
- A 'Receiver_ready', which is a pair of consecutive RR protocol characters. Receiver_ready indicates that the remote node can accept another frame. Depending on the buffer space available, it may be returned anytime after the control field of the corresponding frame

The responses are special protocol characters which are duplicated for checking. In full-duplex operation a node can insert a response for a received frame within a frame that it is currently transmitting. This minimises latency and it reduces the amount of buffering that is needed in each node to sustain continuous data transfer.

An example of the protocol for a half-duplex transfer is shown in Figure 3. This is depicted from the point of view of the source node and it assumes that the destination node can buffer 2 received frames.


|<— 1st frame —>|<—— 2nd frame ——>|
Outbound pair: . C A D D D D X X . C A D D D D D D X X . . . .
V
V
V
V
Inbound pair: . . . r r . . . . a a . . . . a a . . r r .
|<— Processing of 1st frame —>|

Data characters:

C – Control

A – Address

D – Data

X – CRC

Protocol characters:

. – FLAG

a – ACK

r – RR

Figure 3. Example of a half-duplex transfer

Error handling

The serial link provides excellent reliability and serviceability.

Each node has a wrap mode that routes the serialiser output internally to the deserialiser input for power-on self-test.

The hardware provides comprehensive error detection. This includes cable faults, illegal characters, CRC errors, non-sequential frame sequence numbers, protocol errors and missing acknowledgements.

If an error is detected there is an architected recovery procedure. Both nodes exchange Link_reset frames which contain the receive sequence numbers for the respective nodes. Each node compares its local transmit sequence number with the remote receive sequence number to determine how many outbound frames have been lost. It then retransmits those frames from its outbound frame buffers. If successful, recovery is transparent to the upper-level protocol.

Finally the point-to-point physical connections facilitate fault isolation and concurrent maintenance.

Implementation

The logic and frame buffers for each node require approximately 10,000 equivalent gates. This is fully integrated in 1-micron CMOS standard-cell technology, including the line drivers and receivers. The 80 MHz bit clock is derived from a low-cost 20 MHz crystal by an on-chip phase-locked loop. The deserialiser uses a digital clock-recovery scheme that takes 5 samples of each bit.

Some chips in the IBM 9333 contain 3 link nodes and other functions as well. The total power consumed by both nodes of a single link is typically less than 1 W.

Controller command set

Except for the serial transport layer the controller implements the SCSI-2 architecture as defined by the American National Standards Institute. The controller can queue up to 64 tagged commands on behalf of the 4 disk drives that it attaches. The controller implements 19 different commands including Test Unit Ready, Inquiry, Mode Select, Read, Write, etc. These are specified using standard SCSI Command Descriptor Blocks..

At the end of each command the controller returns a SCSI status byte. If an error occurred (ie. a SCSI 'check condition') the controller also generates 32 bytes of sense data for retrieval by a Request Sense command.

The addressing model also conforms to SCSI architecture. The two adapters appear to be alternate 'initiators', the controllers are 'targets' and the disk drives are 'logical units'.

Naturally the implementation of the SCSI functions on a serial link is somewhat different to a parallel bus. The link provides point-to-point full-duplex communication with frame multiplexing. These features eliminate the need for arbitration, selection, disconnection, reconnection and the SCSI 'Attention' signal.

The serial link uses 'messages' to implement all of the SCSI functions apart from the actual data transfer. In the context of the serial link a message is a single frame that is identified by a unique value in the address field. Messages transfer control information from the microprocessor in one node to the microprocessor in the other. The first byte of the data field identifies the function to be performed and subsequent bytes provide the parameters. Messages are defined to issue a SCSI command, to initiate a data transfer, to return SCSI status, to abort a SCSI command and to perform various selective resets.

Disk drive orders

A single 'Read' or 'Write' order can access up to 64K sequential sectors. The disk drive manages the track format and it performs head switches or cylinder seeks as necessary. The disk drive also manages the header fields and it skips defective sectors automatically. Finally the disk drive checks or generates the ECC bytes for the data field in each sector.

The Read and Write orders support split transfers. If the rotational delay to reach the first requested sector exceeds a threshold then the disk drive terminates the order and indicates the sector that is currently under the head.

An 'Extend' order can be overlapped with a Read or Write order in order to dynamically increase the number of sectors to be accessed. This supports back-to-back writes and it allows read-ahead to be extended after a hit in the controller's read-ahead cache.

There are several orders to allow the controller to recover errors. For example, there are orders to obtain an ECC correction pattern and displacement, to offset the head and to recover from a missing header field.

The 'Format_sectors' order allows the controller to initially format the entire disk drive or to reassign a sector in the event of a grown defect.

Summary

The serial links contribute to the excellent performance of the IBM 9333. The adapter-controller link benefits from full-duplex communication with frame multiplexing. The disk orders allow tight control of the device and avoid duplicated function.

The serial link also improves the subsystem reliability and serviceability. It has a high degree of error detection and recovery. The point-to-point connections facilitate fault isolation and concurrent maintenance.

Finally the serial link is compatible with the trend to smaller device form-factors. It needs little power and it uses compact cables and connectors.