X3T9.2/92-67

SSA-PH Appendix
Daisy Chain

Version 0.9
January 29, 1992

J. S. Best

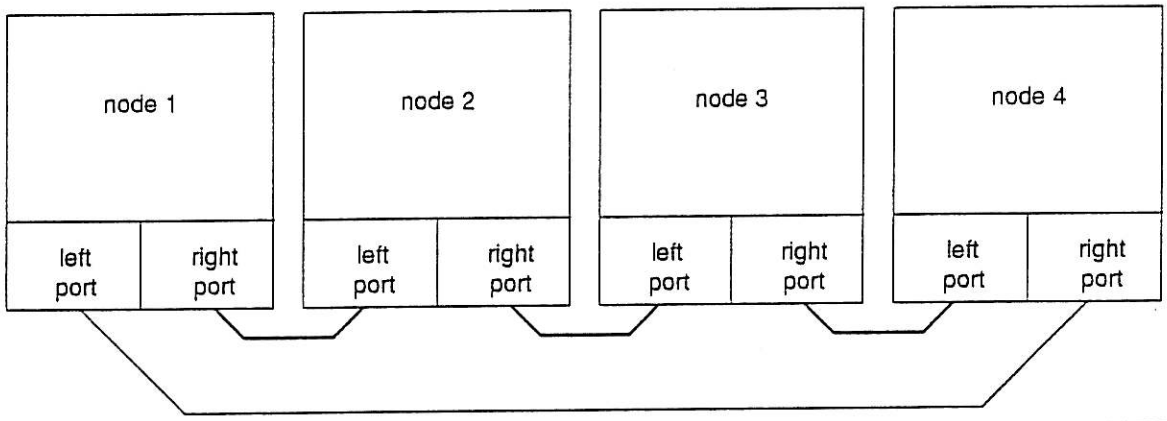
IBM Corporation

Introduction

This appendix defines an extension to the Serial Storage Architecture which implements a daisy chain link with up to 16 nodes on a single link. It is suitable for use as a general inter-connection for computer peripherals. The architecture provides full duplex communication between adjacent nodes, and the capability of handling simultaneous full bandwidth traffic on distinct paths through the daisy chain.

Physical Configuration

Each node includes two bi-directional SSA serial ports. Multiple nodes are connected together by connecting each node to one or two neighbor nodes. Up to 16 nodes may be connected in a single link, with either a straight, simply connected topology, or a single loop topology, as shown in the figure below. The two ports at each node are linked with switch logic which directs frames appropriately to the device containing the node, or through to the next node in the daisy chain.



Definition of terms

Device — A complete peripheral device, such as a disk drive or printer, consisting of the physical device and the serial link node.

Node — The serial link hardware and associated control software, including both of the bi-directional ports and the hardware and software to manage link traffic through both of the ports.

Port — One of the bi-directional serial link interfaces.

Link — The complete set of nodes and connecting cables interconnecting a group of devices.

Link Segment — The two ports and cable between them which connect two adjacent nodes on a link.

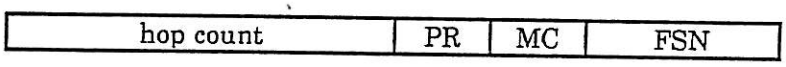
Synchronous Traffic — Time critical traffic which is generated at well defined time intervals, and must arrive within a maximum delay time following its origination. A real time digital video or audio data stream feeding a video display or speaker is typical of this type of data.

Asynchronous Traffic — Typical traffic with an unpredictable origination rate. Ordinary data read and writes to disk drives consist of asynchronous traffic.

Frame Format

The frame format is as previously described in the main SSA-PH document. The CONTROL field and ADDRESS field bit definitions are modified to provide the addressing mechanism for routing frames to the proper destination nodes.

CONTROL field



The control field bit allocation is modified from that point-to-point SSA definition. Usage of the bits is as follows:

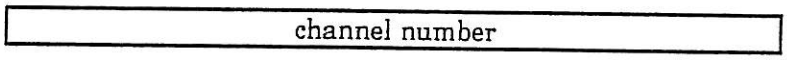
hop count — The hop count provides the primary addressing mechanism, and serves to prevent erroneous frames from circulating indefinitely in a loop configuration. The hop count is initialized by the source node for the frame. It is decremented by each node on receipt of the frame. The node decrementing the hop count to zero is normally a destination node, and will not forward the frame to the next node in the link. Intermediate nodes receiving a frame on one port, forward the frame out the other port. The source sets the hop count to zero for control frames used for local link resets.

PR — This priority bit is set to identify high priority frames, which are used to carry time critical (synchronous) traffic. The pacing mechanism guarantees that priority frames will arrive at all destination nodes within a maximum delay time following initiation of the frame at its source node.

MC — This multicast bit is set to identify frames with multiple destination nodes. Each destination node uses the contents of the address field to recognize multicast frames which it should receive. Non-multicast messages, defined in a higher level protocol are used to prepare each destination node to receive multicast frames with a specific address field value.

FSN — The frame sequence number is defined identically to the manner is which it is defined for dedicated links in SSA-PH. The frame sequence number is independent for each link segment, and is therefore changed by each node through which a frame passes on the way to its destination.

ADDRESS field



The address field is used in essentially the same manner as for point-to-point SSA-PH.

For non-multicast data frames, the interpretation of the address field is higher level protocol implementation dependent. The intended use is that a value of 0xFF is used to address the destination microprocessor. The data field of such frames is used to carry

messages required by the higher level protocol. Other address field values are used to select a particular DMA channel at the destination node for data transfer.

For control frames (hop count = 0), the address field is defined by the reset and link ERP protocol, as in the dedicated link protocol defined in the main SSA-PH document.

For multicast data frames, the address field is chosen by the source node. Each destination node is informed of the address field value to be used for a given multicast transfer using non-multicast commands defined in the higher level protocol. The content of the address field therefore becomes part of the addressing mechanism for directing the frame to each of its destination nodes.

Control Frames and Broadcast Messages

Control frames provide for point-to-point resets across a single link segment. Control field and address field bit definitions are as follows:

	control character								address character
link reset	0	0	0	0	1	0	0	0	link status

Note that the link reset definition is the same as that for the non-daisy chain topology as defined in the main SSA-PH document.

Broadcast messages are used for global resets and configuration messages. They are described in detail in the configuration section of this document. The bit definitions for the currently defined messages are as follows:

	control character					address character							
total reset	hop count	0	1	x	x	0	0	0	0	1	1	1	1
network reset	hop count	0	1	x	x	0	0	0	1	1	1	1	1
"here I am"	hop count	0	1	x	x	0	0	1	0	1	1	1	1

Note that the broadcast messages are not control frames (hop count = 0). They are multicast data frames with an address field of the form 0xnF, where n is any 4 bit value. Only 3 of 16 possible broadcast messages are defined in this appendix. The other broadcast messages are left undefined and may be used by the upper level protocol. If the source node knows the link topology (number of nodes and loop / not-loop status), it can use an appropriate hop count for the message. The message in the table above must all be sent with no assumptions about configuration, hence the source of the message should set the hop count to 0xF and send the message out both of its ports, to insure that the message reaches all other nodes on the link.

Special Character Usage

The main SSA-PH document provides for 7 'user-defined' characters which are available to the upper level protocol. This daisy-chain appendix specifically reserves four of those characters. Two more are reserved by the SSA-SCSI upper level protocol.

SSA-PH Daisy Chain Appendix Reserved

character	usage
CANCEL	"Just kidding" character used in place of the flag character at the end of a frame which is found to contain errors (i. e. CRC error) but has already started transmission forward to the next node.

SAT, SAT′, SAT″	Tokens reserved for use as part of the pacing mechanism to provide fairness among various source nodes, and to enable guaranteed arrival time for priority frames. Handling of these characters by each node will be defined in the pacing section of this appendix.

Higher Level Protocol Reserved

character	usage
SPINDLE SYNC	This character is used to provide a timing reference for spindle synchronization of multiple disk drives in an array configuration. Each node except the node originating this character immediately forwards this character to the next node in the link on its receipt (subject to the rule of never inserting a character between a pair of RR's or ACK's).
TIME SYNC	This character is used for synchronizing other time critical events among multiple nodes. It is treated in the same manner as the SPINDLE SYNC character.

Addressing

Ordinary Transfers

For ordinary transfers (one source and one destination), the multicast (MC) control field bit is set to 0. The source node sets the hop count to the number of link segments that must be traversed to reach the destination. Each node decrements the hop count of an inbound frame. The node which decrements the hop count to zero is the destination. Intermediate nodes pass the frame on to the next node.

The control field is protected against transmission errors by the CRC characters, thereby providing detection for incorrect addressing due to transmission errors.

The source node is not explicitly identified by the SSA-PH level defined fields. For SSA-1 and 9333 SCSI-like upper level protocols, the return address is only needed for messages, not for data transfers (messages will have defined an address field value for the set of data frames associated with a given command). The upper level command protocol (SSA-1 or SSA-2) defines the mechanism for including the source ID, when required, in the frame DATA field.

The address field is used to direct data transfers to the correct destination within a node. The address field for data transfers is determined by the destination node and is communicated to the source node using messages (address = 0xFF) defined in the upper level protocol. A typical hardware implementation would use the address field to select a DMA channel within the destination node. The addressing mechanism, coupled with this use of the address field, provides for up to 255 simultaneous non-multicast data transfers into a given node (with frames for the multiple transfers interleaved in time). The upper level protocol may also provide for tagged command queuing. That mechanism is separate from this use of the address field, and allows for a much larger number of outstanding commands between nodes.

Multicast Transfers

For multicast transfers, the multicast (MC) bit of the CONTROL field is set to 1.

A multicast transfer is initiated using ordinary (point to point) messages defined in the upper level protocol. This message sequence must define a specific ADDRESS to be placed by the source node in each frame for the multicast transfer. Each node to receive the multicast transfer must be set up to receive multicast frames with that specific ADDRESS.

To send the multicast frame, the source sets the hop count to most distant node (provides efficient spatial multiplexing). Each node detecting the MC bit in an incoming frame compares the ADDRESS field contents with addresses of any expected multicast transfers. If there is a match, the node receives the frame. Each node decrements the hop count, as for ordinary transfers, and passes the frame onto the next node if the hop count is not decremented to 0.

Each source is allocated a block of 15 of the potential address field values for multicast transfers. This prevents multiple multicast sources from setting up transfers with conflicting addresses. The configuration section of this appendix describes the mechanism for allocating the block of address field values. This allows for up to 15 independent, simultaneous multicast transfers out of each multicast source.

Messages

Non-multicast messages are used by the upper level protocols for transferring commands and messages between nodes (for example, to send a READ command to a disk drive, or to report completion status). These messages are not required by the SSA-PH level protocol. They are defined in the upper level protocol to be ordinary, non-multicast messages with an address field value of 0xFF.

The resets required to invoke the local link error recovery protocol, as defined in the main SSA-PH document are issued using control frames (hop count = 0). Control frames are local only, and are not forwarded beyond the nodes on either side of a single link segment. Several global resets are required for configuration and catastrophic error recovery. These are issued using a broadcast message.

A multicast frame with an address of the form 0xnF, where n is any 4 bit value, is defined as a broadcast message. Broadcast messages are used for global reset and configuration messages. All nodes, even those not capable of receiving multicast data transfers, must receive the total reset and network reset broadcast messages. The definitions of these messages are given in the configuration section of this appendix.

Inter-link Transfers

This section is not yet fully defined.

The general idea will be that any node supporting multiple links will provide a switch for messages and data from a node on one link to a node on the next link. Extensions to the addressing mechanism which allow peer-to-peer transfers between links are planned.

Configuration

Each node is responsible for detecting the need for a configuration based upon a change in the existence of a node on the other end of either of its two link segments. Any node can request configuration by sending the broadcast network reset message, as defined in the frame format section of this appendix.

All traffic on the link must be halted (configuration may change the addressing for reaching specific nodes, so that frames in process will not properly each their destination). For good system usability, the configuration process should *not* be invoked every time a printer on the link is turned on or off. This would cause problems with multimedia priority traffic (the

video screen would go blank while the configuration was taking place). Mechanisms for handling this problem will be addressed in the power section of this appendix.

Each node on the link sends a "here I am" broadcast message with its device type and a unique serial number in the data field. The exact format of this data is not yet defined. The serial number should include a unique vendor identification field plus a vendor specified field.

Nodes which are not initiators may ignore the "here I am" messages; they receive the return address for all message and data transfers from the contents of messages defined in the higher level protocol. Nodes which are initiators build a table of the direction and hop count to other nodes on the link. Upper level protocol device ID commands may be issued to and from specific nodes to obtain detailed descriptions of the other nodes.

Multicast sources must capture the unique serial number from all other nodes. They then sort the serial numbers in numerical order. The sort index (smallest magnitude serial number is index 0) of the multicast source is then used for the most significant four bits of its address field allocation for multicast data transfers.

Worm-hole Routing

Intermediate nodes in the path of a frame from source to destination should forward the frame to the next node with as short a delay as possible. This is referred to as "worm-hole routing." Each node may determine if a given frame needs to be forwarded by examining only the hop count in the control field. After modifying the hop count and frame sequence number bits, it may send the frame on to the next node.

A delay of at least three characters is required. The CRC characters at the end of the frame must be changed to reflect the correct value following modification of the control field. This requires detecting the trailing FLAG character to recognize the end of the frame. Following detection of the trailing FLAG, the CRC is checked to insure that the frame was received correctly. The CRC characters are replaced with the new CRC characters and forwarded to the next node. This requires that the trailing two CRC characters and FLAG character must be received before the first new CRC character may be sent.. Hence, there must be a delay of at least three characters in the worm-hole forwarding process. Earlier parts of the frame may not be sent with a smaller delay, because the receiving frame has no information about the minimum length of the frame until the trailing FLAG is received.

If a CRC error (or any other link error) is detected, the trailing FLAG character should be replaced with a CANCEL character. Downstream nodes are to ignore the entire frame terminated by the CANCEL character. The local link ERP will be invoked for the first link segment on which the error occurs. A recoverable error will then result in re-transmission of the frame, and proper forwarding to downstream nodes.

Pacing Mechanism and Priority (Synchronous) Traffic

Synchronous traffic is handled in the following manner. Each node contains a separate FIFO buffer in each direction for pass-through priority frames. The FIFO buffer capacity is equal to the maximum frame length. Priority frames are not paced with the normal RR handshake. A node may originate a frame (either synchronous or asynchronous) only if the synchronous traffic throughput FIFO is empty. If a synchronous frame starts to arrive just after a node has started to originate a frame, then the synchronous frame is buffered in the FIFO. The FIFO buffered frame is forwarded as soon as the frame ahead of it is completed. Thus, the FIFO begins to empty as soon as the originated frame has finished (without waiting for RR). A following synchronous frame will therefore never overflow the FIFO buffer, but will fill the buffer immediately behind the frame ahead of it.

Once a node has originated a synchronous frame, the worst case delivery delay is one frame per node traversed. This occurs when each intervening node starts to originate a frame just before the synchronous frame arrives, or when all of the intervening FIFO's already hold one synchronous frame each.

An additional delay occurs at the source if there is pending synchronous traffic in its FIFO from upstream sources. The worst case delay occurs when a high bandwidth upstream synchronous source sends traffic through a number of intermediate nodes, each of which is originating asynchronous traffic destined for the next downstream node. In that case, the delays caused by the asynchronous traffic can cause clumping of the originally evenly spaced synchronous frames. With N such intermediate nodes, and with the upstream synchronous source bandwidth a fraction, B of the full bandwidth of the link, the worst case delay is N×B/(1-B) frame times. This problem is serious only in case of an extremely heavily loaded link, with very unusual asynchronous traffic patterns on the link at the same time.

A more reasonable scenario might involve 2 compressed HDTV 3 MB/s synchronous sources upstream, a single high bandwidth asynchronous source between the two synchronous sources and a third synchronous source downstream, with 3 other intermediate nodes passing the data through. A TV monitor just downstream of the third synchronous source might be sinking the 3 synchronous data streams. Each HDTV source uses only 1/7 of the total bandwidth, so sends 1 frame followed by 6 maximum frame times doing nothing. Frames from the two upstream synchronous sources could clump, resulting in 2 adjacent frames, followed by 5 frames times of nothing. After passing through the asynchronous source, a delay may reduce the 5 frame time gap to 4 frame times. Further clumping would require 4 additional asynchronous sources before the frames pass through the third synchronous source. The worst case delay for the third source is then only 3 frame times. The transmission delay for the first source could be up to 6 frame times, due to the number of intermediate nodes that it must traverse.

The suggested design point is to limit total synchronous traffic bandwidth on any link segment to 90% of the link bandwidth, and to design the buffers for synchronous traffic receivers to handle an arrival time uncertainty of 34 maximum frames times (allowing for 16 hops with worst case 1 frame time delays plus an additional clumping induced delay of 2× 9 frame times. Using 64 byte frames at 20 MB/s, this requires a 117.3μs buffer capability in each receiver.

The high level protocol must supply a mechanism for allocating synchronous traffic bandwidth. The following algorithm, to be invoked by each priority source before starting to originate priority traffic, is suggested:

- The source sends a broadcast message along the same link path as the priority traffic. The message contents should include the required bandwidth for the traffic.
- Only synchronous sources can add to the synchronous traffic bandwidth. Only other synchronous sources along the path of the new traffic needs to recognize the request for synchronous traffic broadcast message. Each such source keeps a total of allocated synchronous bandwidth for its outbound links.
- If the requested additional bandwidth increases that sum to greater than 90% of the total link bandwidth, then a reject message is returned to the requesting source node.
- If no reject message is received, then the synchronous source may proceed. .

- A synchronous source must send a de-allocate broadcast message to the appropriate nodes if the original allocation request is rejected or if it no longer requires use of the allocated synchronous bandwidth.

Because there is no RR pacing mechanism, all nodes receiving synchronous traffic must be capable of accepting each frame immediately. Synchronous traffic should be routed to its destination in the node without delay.

Because there is no RR pacing mechanism, synchronous traffic cannot participate in the normal ERP protocol for retransmitting erroneous frames. Synchronous traffic must be forwarded even while the link is in the middle of an ERP. Synchronous frames do not participate in the normal frame sequence number process. Asynchronous frame sequence numbers function as described earlier in this appendix, and do not include intervening synchronous frames in the sequence. The frame sequence number for synchronous frames is set by the source and is not changed by forwarding nodes. This allows the destination node to detect a missing synchronous frame, deleted due to a framing error or canceled due to a CRC error.

A pacing protocol and fairness algorithm for asynchronous traffic is planned. This would serve two purposes:

1. It would pace asynchronous sources under heavy traffic conditions to prevent the possible of long delays in synchronous traffic.
2. It would prevent one asynchronous source from locking out another asynchronous source.
3. It would prevent buffer full backups under heavy loading conditions which could cause long transmissions delays for asynchronous traffic.

The protocol will follow the SAT/SAT' algorithm proposed by Cidon and Ofek (see References).

Power

This section has not yet been completed. Power is required in the cable to keep a minimal section of the link protocol chip operating so that traffic can be routed through the node even when it is powered off. Power handling hardware and management specifications will be added to be consistent with system requirements to be specified by potential SSA daisy chain system integrators.

References

Isarail Cidon and Yoram Ofec, "MetaRing - A Full-duplex Ring with Fairness and Spatial Reuse"