X3T9, 2/87-25

Los Alamos

Los Alamos National Laboratory
Los Alamos, New Mexico 87545

Computing and Communications Division

TO : ANSI X3T9 Subcommittees and Other Interested Persons

FROM : Don Tolmie Don Tolmie

SUBJECT : High Speed Open Channel

DATE : February 16, 1987

We see the need for a standard computer interface in the 800 Mbit/s range. The current need is in the supercomputer area for data transfers machine to machine, machine to disk, and machine to graphics display. This will be a forward looking standard, the application area is quite new and there are few channels of this speed in use today. A major problem is the lack of a public channel to interconnect different vendor's equipment. The highest speed computer interconnect standardization work currently in progress is FDDI at 100 Mbit/s, a factor of eight slower than our requirements.

Computers have enough number-crunching power today to drive displays at video rates, giving tremendous user productivity increases using movie like output. It allows a closer match between the human eye-brain pattern recognition ability and the number crunching capability of supercomputers. Our current goal for real-time video is 1K x 1K pixels, 24 bits of color, and a refresh rate of 24 frames per second, translating to a sustained data rate of 600 Mbit/s.

Not only do we want to support real-time video rates, but we also want to increase network data rates for machine to machine, and machine to storage system transfers. Todays network speeds of 50 to 80 Mbit/s were really determined by the fastest disks available at the time the networks were designed, and disk and processor I/O rates are now above what even FDDI can handle.

We feel that the basic requirement to satisify these needs is a high speed open (public) channel specification. The community of interested vendors and users is larger than we originally thought, and growing rapidly. We have talked to vendors about products in this range, and they all uniformly agree that we need a channel of this speed, but no one wants to promote their proprietary channel as the proposed standard. We have also looked at current standards efforts, i.e., IPI, SCSI, LDDI, and FDDI, and do not feel that they can be effectively upgraded to give the required bandwidth and functionallity. As such, we have drafted a proposed set of goals and a channel specification for use as strawmen.

We propose that an ad hoc working group meet to discuss this topic, and possibly petition X3T9 to sponser work on standardizing the physical and link layers of a high speed open channel. For further information contact:

Don Tolmie
Los Alamos National Laboratory
C-5, MS-B255
Los Alamos, New Mexico 87545
Phone: (505) 667-5502.

72

*** DRAFT ***

High Speed Open Channel Goals

*** DRAFT***

Don Tolmie
Los Alamos National Laboratory
February 12, 1987

1. Burst speed of at least 800 Mbit/s - This rate is derived from three application areas presently identified. (1) Mainframe to disk transfer rates are currently at 100 Mbit/s and disk speeds are going up. 800 Mbit/s is greater than presently neded, but will provide room for overhead and future growth. (2) Supercomputer to supercomputer communications, expecially when using the supercomputers as cooperating processors of a multiprocessor. (3) Mainframe to video displays where sustained transfer rates of 600 Mbit/s are required to support displays with 1K x 1K pixels, 24 bits of color, and 24 frames per second refresh rate.

2. Point-to-point connections - In the interest of signal quality and reliability, point-to-point was chosen over multi-drop configurations. We plan to use a crossbar switch to route data between multiple nodes. We propose to limit this standards effort to the point-to-point connection, allowing hooks for the crossbar but not specifying it.

3. Symmetrical, full-duplex - The channel shall support simultaneous transfers in both the transmit and receive directions. The symmetrical requirement allows for loop-back testing by simple cable changes.

4. Peer-to-peer - No assumptions are made about master-master or master-slave configurations, this is controlled by higher layer protocols.

5. 50 meter basic length - This assumes parallel conductor copper cables in a machine room environment. Various versions of fiber optic repeaters can be used for longer distances, and of course at a greater cost. We do not propose to work on standardizing a fiber optic implementation at this time, if ever.

6. No bidirectional signal lines - This relates to the requirement for symmetrical paths for loop-back, and for ease in building a crossbar switch.

7. Hardware flow control - Flow control is necessary to keep from overrunning the receiver, and must be at the hardware level to minimize the overhead. The flow control should also allow for double buffering or pre-acknowledgement. In our proposed channel, hardware flow control is used on 1 KByte bursts, allowing distances of up to about 1 km with the sustained data rate equal to the burst rate.

8. Packet size independent of physical level - To keep the overhead down, very large packets must be used (on the order of 50 Kbytes or larger). It is highly desirable to allow upper protocol layers to define the packet size, for example a 3 MByte packet containing one video frame. Small packets must not be precluded, but they will not sustain full bandwidth.

9. Allow the use of existing higher level protocols - We envision a protocol like TCP/IP which allows for sending multiple packets, with packet acknowledgements coming at a later time.

10. Error control - Include sufficient error control to detect single and multiple bit errors, and possibly correct single bit errors. Any retransmissions should be triggered by higher protocol layers so that when transmitting video data, a few bad bits can be ignored, and certainly should not cause retransmission.

11. Implementable with off-the-shelf parts - This is a forward looking standard, but we feel that it can be implemented using today's parts.

High Speed Open Channel Proposal

January 15, 1987

Don Tolmie
Michael McGowen
Gene Dornhoff

Los Alamos National Laboratory

73

1.0 INTRODUCTION

1.1 General

This specification is presented in five sections. The first is an introduction and definition of terms. The second and third are the logical definition of the channel signals and examples illustrating their timing relationships. The fourth is the formal specification of timing and other quantitative data, and section five is a brief discussion of possible future implementations.

1.2 Objectives

The specification defines the physical link layer of a simple high speed channel for the transmission of digital data between pieces of data processing equipment. The channel is optimized for predictable transfers of large blocks of data such as used for raster scan graphics terminals or file transfers between supercomputer class machines, but will also accommodate smaller messages. The channel is kept as simple as practical to minimize implementation cost and to speed throughput. It is expected to be used for point-to-point links, but could be implemented as a multi-drop bus or in a cross-point switch for special requirements.

A block diagram of a simple point-to-point link is shown below.

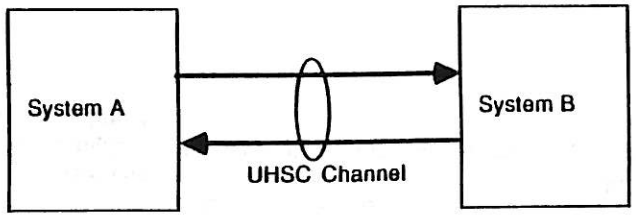


Figure 1.0. UHSC System Block Diagram

1.3 Definition of Terms

Ultra High Speed Channel (UHSC): The specifications and rules for connection and the transfer of data between pieces of equipment. The only physical equipment included is the cable and connectors used to make the physical connection between interfaces. The specification is symmetrical, and accommodates full duplex communications.

Ultra High Speed Parallel Interface (UHSPI): The equipment used to connect a single piece of equipment to an Ultra High Speed Channel. It will normally convert an external, non-compatible channel to conform to the UHSC specification. It may or may not include additional functions as desired for a specific application. For full support of the full duplex nature of the channel, each UHSPI will include both a Source and Destination as defined below.

Source: The equipment at the end of the link from which data flows. May be used interchangeably with the terms Transmitter or TX Interface.

Destination: The equipment at the end of the link to which the data flows. May be used interchangeably with the terms Receiver or RX Interface. When it is necessary to distinguish between a Destination physically connected to a link and an addressed Destination elsewhere in a network, the physically connected Destination will be referred to as the Local Destination. This is often the case when a switching node is between the Source and the ultimate addressed Destination.

Word: 32 bits of data transferred across the channel from the source to the destination by a single strobe edge.

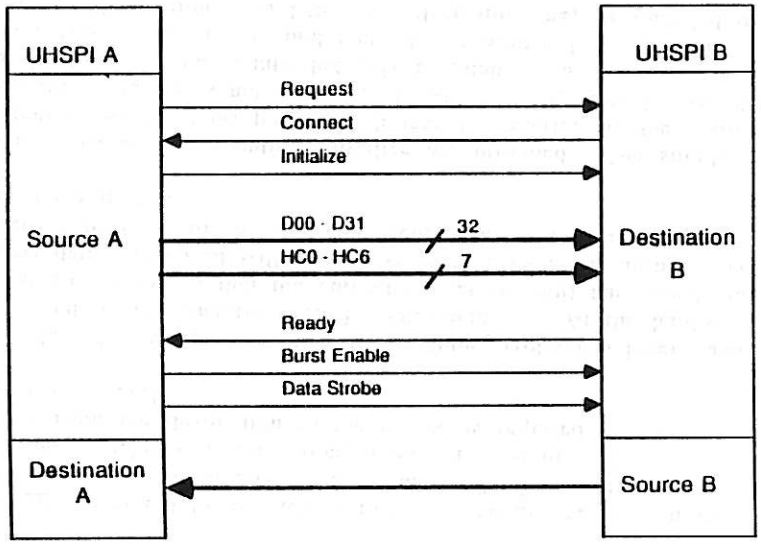
Burst: A group of up to Bmax words sent by a Source to a Destination. No acknowledgement by the destination is required during the burst, but handshakes are required before the first burst and for each subsequent burst.

Packet: A group of data sent during one logical connection of the channel, composed of one or more bursts. No maximum size is specified by the channel, but a maximum may be required by a given interface or by a higher level software protocol.

Request Information Field: A single word of information that may optionally be sent from a source to a destination as part of the sequence of operations establishing a link from a source to a destination. The contents of this field must be defined at a higher level. An example of the type of information that may be useful would be an address.

2.0 SIGNAL DEFINITIONS

A diagram of a simple system including two UHSPI's and one UHSC showing the directions of the signals is shown below. Only one half of the full duplex channel is shown in detail. The second half is the mirror image.



B to A signals are the mirror image of A to B signals.

Figure 2.0. System Functional Block Diagram

2.1 Channel Signal Groups

For convenience, the signals have been ordered into groups of related signals.

2.1.1 Data Bus (D00 - D31)

The data bus consists of 32 parallel lines of data or address information.

2.1.2 Error Control Bus (HC0 - HC6)

Each data or address word is accompanied by a 7 bit Hamming code generated according to the parity matrix shown in Attachment (x). This is a standard code capable of single bit error correction and double bit error detection (SEC/DED).

2.1.3 Packet Control Lines

Request: Asserted by the source to notify the destination that the transmission of a packet is desired. The Request will remain asserted for the duration of the packet. When not asserted, Request indicates that none of the other source channel signals except Initialize are valid, and that the source will ignore any signals asserted by the destination.

The Request may be accompanied by a Request Information Field (I-Field) on the data bus. This is a single word whose meaning is defined at the next higher link layer. If used, the word must be asserted on the data bus RIsetup before Request is asserted, and must remain asserted until the destination returns Connect.

If a Source times out an attempted connection before Connect is returned to the Source, the Source must implement a second time delay before attempting a new connection. The period of these delays will be system dependent.

Connect: Asserted by a destination to notify the source that it is operational and available for data transfers. It may be dropped at any time during a transfer to indicate to the source that the connection is no longer available. When not asserted, the source will

assume that none of the other signals asserted by the destination are valid. Whenever it is asserted, Ready must be in a valid state.

When both Request and Connect are asserted, a Connection is said to be established, or the Link is Connected.

If the optional I-Field is used, the Destination must not assert Connect until it is acceptable for the Source to remove the I-field word from the data bus.

Initialize: Asserted by the source to notify the destination that the source is going to "back up 15 yards and punt". The destination is responsible to take whatever action, if any, it must to start over. No response from the destination to the source is required at the physical link level.

The value of this signal is primarily to higher (software) levels, and will depend on the specific UHSPI implementation. At the hardware level, it simply means that the current packet (if any) has ended, and the next data transferred will be a new block packet. It normally is asserted in response to an error condition, or as part of an initial power up sequence.

Two methods of implementing Initialize are provided. The simplest is a purely asynchronous pulse which may be asserted by the Source at any time Request is not asserted. It will not propagate beyond the Local Destination, and is intended only to establish the lowest level of communication. For the second method, the Source first establishes a link to the Destination. An I-field of information may be included if so defined for the interfaces involved. The Source then asserts the Initialize pulse, and completes the sequence by negating Request.

2.1.4 Burst Control Lines

Ready: Asserted by the destination after a connection is established to indicate that a burst not to exceed Bmax words may be sent. The destination must negate Ready between bursts, and the source must not begin a second burst until Ready has been re-asserted.

A single burst lookahead function may be implemented in the destination such that the destination may drop Ready at any time following the start of the burst and re-assert Ready as soon as the destination can predict that it will be able to accept a second burst

immediately following the current burst (for example, a double-buffered destination could accept two bursts before waiting). A source must implement the look-ahead function to be compatible with all destinations.

A source will recognize only one pending Ready beyond the burst currently in progress.

Burst Enable: Asserted by the source as a burst delimiter to the destination. Setup and hold times TBEsetup and TBEhold are measured with respect to the nearest data strobe edge. The signal is asserted for the first and subsequent words including the final word in the burst. It is negated after the final word.

Data Strobe: Asserted by the source as a timing reference for data word transfers. Both edges of the strobe pulses are active (data is clocked by both rising and falling edges of the signal) as long as Burst Enable is asserted. Timing of the strobe is asynchronous in that only the minimum time between strobe edges is specified. Pauses between strobes are permitted as long as Burst Enable is left asserted. Extraneous strobes between bursts may be ignored, and a source may optionally return the strobe signal to an idle state between bursts if a burst containing an odd number of words has left the signal in the opposite state. (This allows a source to return to a known state between bursts, if desired.)

Note: If the final parallel cable implementation uses multiple cables, as is almost certain, each cable carrying data lines will probably have its own strobe. This will accommodate the inevitable timing skew between cables.

3.0 TRANSACTION EXAMPLES

Logical timing diagrams are shown for several sample channel transactions.

3.1 Channel Selection

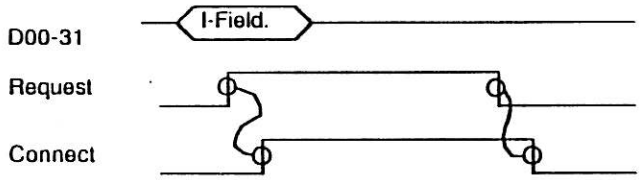


Figure 3.1. Link Connection Sequence

The Source places the I-field (if used) on the data bus, and then asserts Request. The Local Destination responds to Request by asserting Connect when the I-field may be removed. The Local Destination will wait until any remote destination returns Connect, if a remote destination is involved, before it asserts Connect to the source. The Source is then free to remove the I-field from the data bus.

After the end of the data transfer, the Source will drop Request. The Destination will respond by dropping Connect.

3.2 Aborted Channel Selection

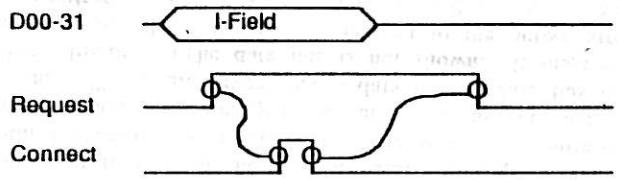


Figure 3.2. Aborted Channel Selection Sequence

The Source places the I-field (if any) on the data bus, and then asserts Request. The Local Destination responds, and finds the Destination is unavailable. (For example, the Local Destination is a port into a switch, and the Destination is a host machine connected to another port on the switch, with that machine being down at this time.) The Local Destination signals to the Source that a link cannot be established by asserting Connect and then negating Connect before asserting Ready.

Note that this transaction is just a degenerate special case of an illegal end where the destination terminates a transfer. In this case, the destination terminates the transfer without accepting any data.

3.3. Initialization Sequences

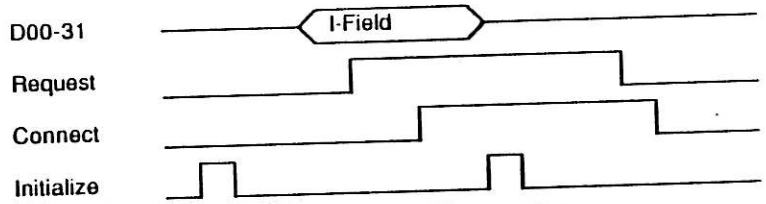


Figure 3.3. Initialization Sequences

The Source asserts and negates Initialize without Request to reset the Local Destination. For the simplest Initialize, that is all that need be done. If the next level of Initialize is required, the Source would begin by placing the I-field (if any) on the data bus, asserting Request, and waiting for Connect. When Connect is received, the Source would then send an Initialize pulse to the Destination. The Source would then negate Request, ending the connection.

3.4 Normal Data Transfer

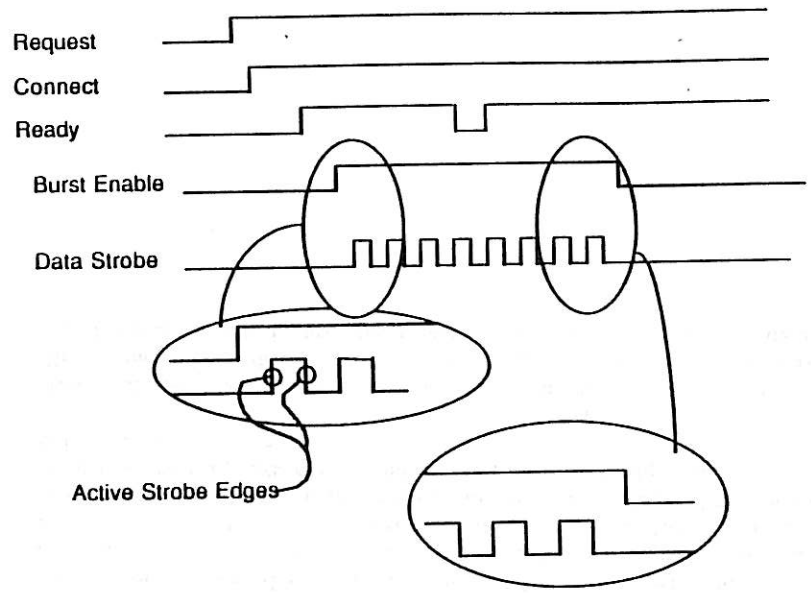


Figure 3.4. Normal Data Transfer

The start of a packet transfer is shown. For clarity, only 16 words are shown in the burst. The sequence begins with a normal link connection operation, with the Source asserting Request and the Destination responding with Connect. The sequence continues with the Destination asserting Ready to signal to the Source that a burst may be sent. The Source then places data on the data bus and asserts Burst Enable. (The data bus is not shown. It must simply meet the setup and hold times with respect to the active strobe edges as specified in section 4.) Data is clocked out of the Source and into the Destination by the strobe as shown. After the final word of the burst, Burst Enable is negated following the hold time specified in section 4.

The Destination may drop Ready at any time following the start of the burst, then re-assert it to indicate the Destination can accept the

next burst. Note that this destination has implemented the lookahead Ready sequence, by dropping and re-asserting Ready during the burst. If the Source had more data available and ready for transfer, it would have permission to begin sending the second burst immediately after the end of the first. Following bursts are sent exactly as the first.

At the end of the packet, the Source will drop Request and the Destination will drop Connect.

It should be noted that the specification does not require the burst to be sent as a continuous stream - - there may be pauses between strobe edges as long as Burst Enable is left asserted. Flow control is exercised by the Destination by not asserting Ready until it is able to accept a full burst of up to Bmax words at the maximum channel speed.

3.5 Illegal End Transfer

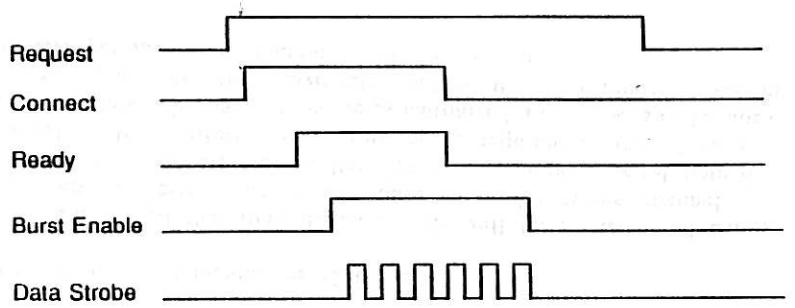


Figure 3.5. Illegal End Transfer

The Source and Destination start a transfer normally. At some time during the transfer, the Destination detects a non-recoverable error. To terminate the transfer, the Destination negates Connect (and Ready, since Ready should not be asserted without Connect). A link propagation delay later, the Source detects the negation of Connect. In response, the Source will drop Burst Enable and stop sending data strobes. Request will also be negated as soon as possible, but this may happen more slowly.

4.0 SIGNAL SPECIFICATIONS

TDmin	Minimum time/word	40ns
BPSmax	Maximum data rate	800 Mbit/s
TCmin+	Minimum time Connect is asserted	100ns
TCmin-	Minimum time Connect is negated	100ns
TImin	Minimum time Init is asserted	100ns
TImax	Maximum time Init is asserted	1000ns
Tsetup	Min. time data is stable before strobe	25ns
Thold	Min. time data is stable after strobe	10ns
TBEsetup	Min. time BE is stable before strobe	100ns
TBEhold	Min. time BE is stable after strobe	100ns
TBEmin-	Minimum time BE is negated	100ns
Bmax	Max. number of words/burst	256
RIsetup	Minimum setup time for I-field	100ns

5.0 Relevant System Implementations

5.1 Switching Nodes

An early anticipated usage of the UHSC is in the hub of a star network of machines interconnected with UHSPI's and UHSC's. The hub would be implemented as a fully connected crossbar switch. It is in anticipation of this application that the optional I-field has been included as part of the Connect sequence. In this application, the I-field will contain an address for a machine elsewhere in the network, not just at the far end of the link physically connected to the source machine.

The 32-bit address field is large enough to allow several logical fields to be included, if desired, to identify such things as message type as well as physical or logical addresses for the actual destination.

5.2 Channel Repeaters

Due to the inherent limitations of the transmission cables, the maximum length of any UHSC is fairly short. The specification will accommodate fairly simple channel repeaters in a manner transparent to the source and to the remainder of the network.

5.3 Serial Implementations (Fiber optics)

It is anticipated that long distance links will be implemented using fiber optics. Serial connections would be used between channel adapters that convert the parallel UHSC to a suitable serial optical format. FIFO buffering will probably be required in the adapters, and the data will have to be re-synchronized to a new synchronous clock. Except for the unavoidable delays, the fiber repeaters should be transparent to the remainder of the system.