The Role of HiPPI Switches in Mass Storage Systems: A Five Year Prospective

T. A. Gilbert

Network Systems Corporation
Vienna, Virginia

Introduction

New standards are evolving which provide the foundation for novel multi-gigabit per second data communication structures. The lowest layer protocols are so generalized that they encourage a wide range of application. Specifically, the ANSI High Performance Parallel Interface (HiPPI) is being applied to computer peripheral attachment as well as general data communication networks.

This paper introduces the HiPPI standards suite and technology products which incorporate the standards. The use of simple HiPPI crosspoint switches to build potentially complex extended "fabrics" is discussed in detail. Several near term applications of the HiPPI technology are briefly described with additional attention to storage systems. Finally, some related standards are mentioned which may further expand the concepts above.

The High Performance Parallel Interface

History

The HiPPI standard evolved from efforts begun and still lead by individuals at The Los Alamos National Laboratory. Originally known as HSC or "High Speed Channel", HiPPI was derived from the Cray Research HSX supercomputer channel.

The original framers of what has become the HiPPI standard had several objectives in mind which in retrospect have been crucial to the rapid acceptance of this standard by many users and vendors:

- An interface capable of data transfer in the gigabit per second range. HiPPI is defined for 800 Mbps and 1.6 Gbps rates.

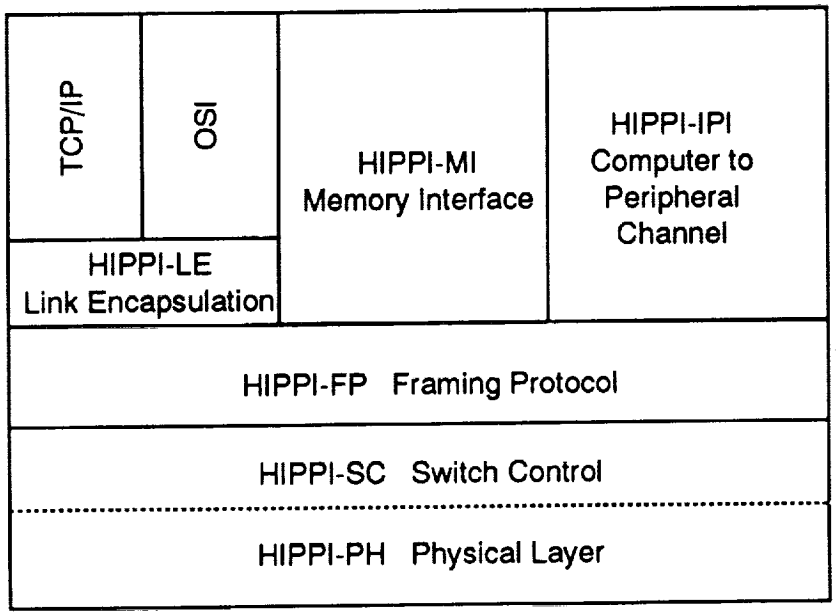
- standard interface which could be implemented by a broad range of vendors without the need for exotic or expensive technology. HiPPI physical layer interfaces can be built from off the shelf components which have been available for two decades.

- standard which is stratified such that the most fundamental common layers impose the least possible restriction on the nature of the digital datastream. HiPPI is being proposed for use in traditional networks, for the attachment of peripherals to host channels, for digital HDTV, and for connecting isochronous streams of imagery and digitized voice.

HiPPI standards efforts are under the auspices of ANSI X3T9.3 which this year will finalize most if not all of the constituent standards relevant to the directions discused in this paper. Related or follow on efforts are discused below.

The ANSI HIPPI Standards Suite

The X3T9.3 committee has defined six HiPPI component standards. Three are common to all others and comprise what may be likened to the media access layer in the ISO Open Systems Interconnection protocol model. However, this analogy implies that HiPPI is but another data link component in the traditional data communications hierarchy. It can serve that role but this understates its generality of application as discussed below.



The six standards are:

HIPPI-PH - The physical layer definition which includes mechanical and electrical interface definitions. It also specifies the signaling rates of 800 and 1600 Mbps. Important HIPPI-PH characteristics are:

- ❑ 800 or 1600 Mbps isochronous interface
- ❑ parallel 32 or 64 bit wide data line interface
- ❑ 25 meter maximum cable length
- ❑ simplex interface
- ❑ parity and LRC data protection
- ❑ ready resume flow control

HIPPI-SC - An optional extension of the physical layer standard which defines a switch control interface. HiPPI connections may be switched to achieve multi-point connectivity. Multiple addressing modes are defined.

HIPPI-FP - Defines a common framing protocol for all other standards.

HIPPI-LE - The link encapsulation definition designed to support traditional data communication protocols such as TCP/IP and OSI. LE essentially creates an IEEE 802.2 LLC compatibility layer on top of HIPPI-FP.

HIPPI-IPI - This is really more of a place holder to designate the use of ANSI IPI2 or IPI3 channel protocols over a HiPPI connection.

HIPPI-MI - Is a memory interface definition which provides for a communication controller to mediate memory to memory data transfers. MI attempts to avoid the overhead in traditional protocols and create mechanisms useful for cooperative processing.

HiPPI Technology

A handful of equipment vendors have actually shipped HiPPI compliant products to date. However, many more have announced intentions to do so over the next year. Products are available as of the first half of 1991 to begin implementing several of the advanced applications mentioned later. Examples of existing products are described in this section.

Computer Channels

IBM was the first computer manufacturer to announce and ship a HiPPI channel for their mainframe products. Subsequently, other vendors in the technical computing market have begun to deliver HiPPI channels. Most notable has been Cray Research who have also aggressively pursued software support in their standard operating system UNICOS.

Peripherals

One of the earliest effects of the HiPPI standards effort was to stimulate peripheral manufacturers efforts. Broad support of a high performance channel by the computer vendors immediately created a "plug compatible" peripheral market. Disk arrays, tape cartridge drives and frame buffers are early examples of announced product which also require the high data transfer rates achievable with HiPPI.

Switches

Network Systems has been an active member of the X3T9.3 committee since its inception and was perhaps the first vendor to ship a HiPPI compliant product in the form of a switch. HiPPI switches provide for the very rapid connection of input channels to output channels. Currently, products support up to eight input and eight output ports per chassis. Switches may be cascaded to form larger fabrics as described below. Thirty-two port switches have been announced for availability later this year.

Extenders

HiPPI's twenty-five meter cable length imposes a severe restriction on most applications. Several companies have delivered fiber extenders for full rate HiPPI channel extension. Using either multi-mode or single-mode fiber pairs, distances of several kilometers can be reached. Extenders may attach switches to one another enabling switched high speed connections over campus distances.

Work has recently begun at Network Systems to couple HiPPI fabrics using SONET (Synchronous Optical NETwork) facilities at the OC-12 signaling rate which is about 622 Mbps. This will initially be targeted to metropolitan area distance requirements. As the technology matures, it is intended that this interface would incorporate an ATM cellifier and data rates up to OC-24 which is 1.244 Gbps. ATM will support variable data rates and the creation of virtual circuits to multiple remote destinations.

Gateways

Traditional internetworks have firmly entrenched the role of bridges and routers in any but the simplest of networks. As the potential applications for HiPPI grow to demand extended "fabrics", perhaps over geographic distances, there will be a need for gateways engineered to operate at HiPPI rates.

Network Systems is currently developing a family of HiPPI gateways as part of its work in the Carnegie Mellon NECTAR project. In NECTAR, the gateways are known as CABs for Communication Accelerator Boards. Indeed, one of the projected uses for HiPPI gateways involves the interfacing of existing bus based systems to the fabric; this was the original intent of the CAB in the NECTAR architecture.

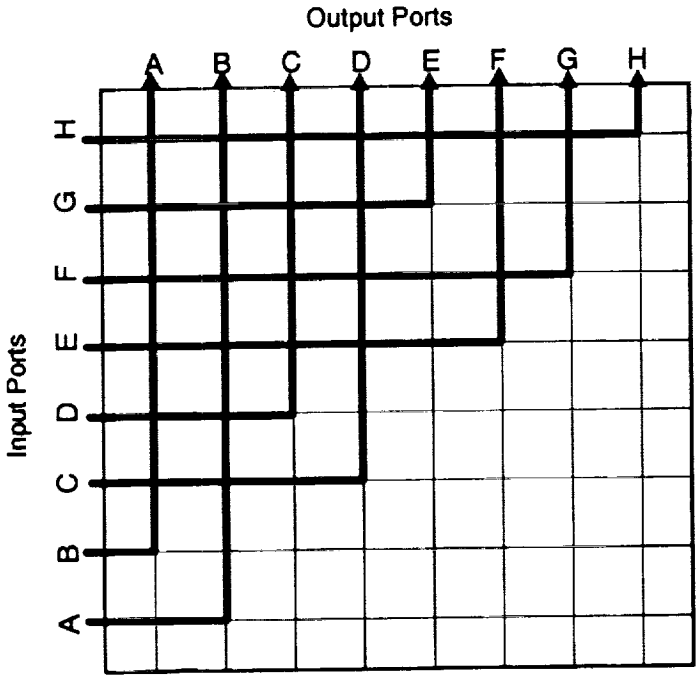
CABs will also exist within HiPPI networks to provides various types of bridging functions. For instance, where long haul extenders are inserted into a network it may be prudent to interface each end via a CAB. The CABs keep a permanent HiPPI connection up between them. Each CAB is prepared to accept HiPPI connections from the user side for forwarding over the extender. This design avoids the latency necessary to establish an end to end HiPPI circuit before the first word can be transmitted. The existence of the CAB will generally be transparent to the user nodes.

Other functions proposed for CABs include security functions to enforce network level access control. Current research is focused on ways CABs may be used to perform outboard protocol assist functions for host computers.

Building Crosspoint Switch "Fabrics"

HiPPI is fundamentally a connection oriented interface standard. One must actually create a HiPPI connection (via control circuits in the physical layer) before data can flow. This is true even for point to point HiPPI cable connections. The basic idea of a crosspoint switch is familiar to anyone with even the barest understanding of telephony. Any input port may be switched in some fashion to any not-busy output port. Once connected, data may flow at the nominal port rate without regard to other connections though the switch.

So far, Network Systems HiPPI switches are true crosspoint switches in that there are no shared data paths. All ports may simultaneously move data at the nominal rate without any contention effects providing for impressive aggregate throughput. Also, the switches are "non-blocking" internally which means that as long as the output port is not busy any input port may connect regardless of other connections in the switch.



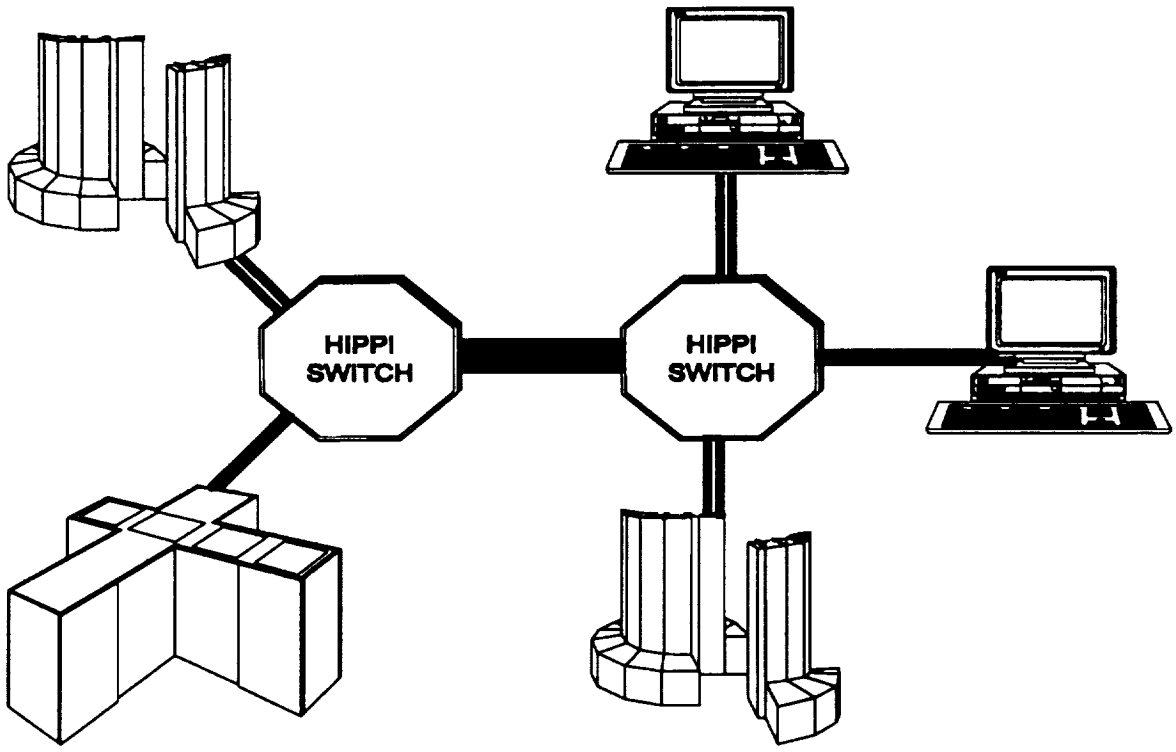
For near term applications in backend networking for supercomputers or attachment of peripherals, single stage switches with four to thirty-two port pairs are probably adequate. However, the limits of board to board connector technology means that we are rapidly approaching the limits of current switch architecture. Therefor, requirements which dictate greater HiPPI connectivity will probably use multi-stage switches constructed by cascading existing switches.

Cascading HiPPI Switches

The output port of a HiPPI switch may be connected to the input port of another (or the same) switch. At each stage, the input port may be switched to any not busy output port of that switch. Switches are designed to propagate the necessary switching signals from input to output such that the existence of the multiple switching stages is essentially transparent to the end points. Once a HiPPI circuit is established through a multi-stage switch fabric, the only noticeable difference from direct cable connections would be a negligible amount of additional data latency.[1] The process of creating a HiPPI switch connection is dependent upon the switch interpreting an in-band address designated by the originator. Note that in a multi-stage switch configuration, each prior stage becomes the originator for each subsequent switch stage until the end-point is reached.

---

[1]Current HiPPI switch products add approximately 160 nsecs of latency to data. This is roughly comparable to the latency due to 25 meters of cable.



Addressing

The basis for HiPPI connection switching is something called the "I-Field" in the HIPPI-PH standard. The I-Field is the contents of the 32 bit wide address circuits of the HiPPI channel at the time the connection request control circuit is raised. The high order octet carries control flags and the low order twenty-four bits are used for the actual addressing.

The HIPPI-SC standard defines two modes of addressing. Either may be used to create multi-stage HiPPI switch connections.

Source Routed Addressing

In source routed addressing, each switch stage examines the several low order bits in the I-Field necessary to address an output port. For instance, an eight port box requires three bits to address ports 0 through 7.

To support multi-stage switching, the switch can optionally rotate the field to bring the next *N* bits into position for the next switch stage. Preservation of the path information is important for the last stage switch. It may be set to automatically create a reverse HiPPI circuit for dual simplex connections.[2]

Source routed address interpretation in switches will typically be performed in hardware providing for very high performance switching.[3] The disadvantage of source routing is that the end point systems must keep a record of the switch fabric topology. The route to each resource will be different for each from-point complicating address table administration. For small networks, this has not been judged to be a problem. But recently, requirements have started to surface for multi-thousand port HiPPI fabrics.

Isomorphic Addressing

Most people are more familiar with isomorphic addressing than the source routing approach. This is the same concept as in Ethernet networks. Each attachment to the network has a unique address which is unrelated to the network topology and need not change when the node is moved to a new point on the network.

Second generation switches support the use of isomorphic addressing which is selected with one of the flag bits in the I-Field. The address portion of the I-Field is split into two twelve bit fields; a to address and a from address. The "to" address is interpreted by each switch stage to determine the next outbound port. Obviously, isomorphic addressing limits HiPPI switch fabrics to a maximum of 4096 addressable nodes.[4]

With isomorphic addressing, boundary nodes are relieved of the need to know about the network topology. Instead they rely upon the collective knowledge contained in the switch forwarding tables. The HiPPI standards do not specify how these tables are created or inserted into the fabric. This is the subject of a current project at Network Systems concerned with switch management.

---

[2]Many of the planned HiPPI applications do not require duplex connections. For instance, frame buffers are essentially simplex, write only devices. Connectionless protocols which use IEEE 802.2 procedures also do not require immediate reverse connections.

[3]The original Network Systems P8 first generation switch is capable of establishing source routed connections in 240 nsecs.

[4]Notice that for multi-stage switch arrays only boundary ports need to consume isomorphic address space. Inter-switch ports may be addressed if necessary using source routed addressing modes.

Switch Management

The switch management project is focused on the practical details of constructing and using arbitrarily large switch fabrics. It is also directing switch features which contribute to the resiliency of the fabric when inevitable failures occur.

Auto-configuration

The foregoing discussion of isomorphic addressing makes clear the necessity of some automatic means for a large multi-switch network to configure itself. By this, we mean the creation of the forwarding tables using connectivity data received from neighboring switches. This is analogous to the techniques used by spanning tree bridge networks to automatically discover the "best" path to a destination.

This process is also intended to support alternate pathing since most practical HiPPI fabrics will contain many possible ways to route a connection from the originating port to a destination. Frequent updating of the tables through the automatic process also provides for routing around failed components. Lastly, the switch management features will provide a means for address resolution similar to that done in internetworks.

None of the switch management features will preclude the use of the HiPPI network for attachment of simple peripherals. Participation in advanced services by boundary nodes is optional.

Additional Services

Closely related to switch address management are the provision of two additional services under consideration. Multi-cast delivery of data is an outgrowth of the address resolution function. It will be possible for boundary nodes to be joined to a multicast group. A sending node may address a HiPPI connection to a multi-cast group address. The switches will provide a best efforts delivery to each node in the multi-cast group.

Network access control services will be provided through forwarding table management. This will allow an administrator to restrict the possible connections from any boundary node.

HiPPI Applications

The HiPPI standards are still being finalized and related products have only been available for a short time. There are many applications for which HiPPI has been proposed. Few of these have been proven for

practical application as of mid 1991. However, the following should be considered representative of the potential breadth of use for this new technology.

Device Connections

Since HiPPI is directly descended from the Cray HSX channel, it seems obvious that it will be used as an open standard computer to peripheral channel. Currently available disk array controllers capable of 500 to 800 Mbps transfer rates clearly demand HiPPI rates. High density tape cartridge systems can read and write in the hundreds of megabits per second range. Some types of telemetry recording devices are being adapted to HiPPI which are capable of Gbps rates.

Another special type of peripheral is the frame buffer used to image animated high resolution displays of complex scientific data. At 24 frames per second, this application requires over 700 Mbps data rates.

The availability of HiPPI switches leverages the advantage of a multi-vendor standard peripheral channel. Any peripheral on a HiPPI switch fabric is potentially shareable by any other nodes on the fabric. Although this sounds like the old Block Mux Channel switch often seen in IBM shops, the rapid switching rates and high transfer rates make this a feasible application even in supercomputer environments.

Backend Networks

The earliest "production" uses of HiPPI are expected to be computer to computer file blasting applications. Standard protocols such as TCP/IP will be supported by most computer vendors who have HiPPI channels on their hosts. This, in turn, will allow higher speed FTP and NFS based data access from host based file servers.

There is a general misconception that TCP/IP is not capable of achieving gigabit per second network speeds. However, multiple researchers have found that there is no intrinsic reason that TCP/IP should not perform in the super gigabit range.[5] In most instances, poor implementation or operating system interference have delivered disappointing network performance.

---

[5]See "How Slow Is One Gigabit Per Second?" by Craig Partridge; BBN Systems and Technology Corporation, Report No. 7080, June 5, 1989.

The availability of high performance networks based upon HiPPI is expected to stimulate vendor efforts in improving protocol performance. Cray Research has, so far, been the leader in this effort.

Backbone Networks

Interestingly, the rapid connect processing of the HiPPI switches makes them suitable for the delivery of short message traffic. It is entirely feasible to "dial-up" a HiPPI connection for each datagram. Each port on current switch products can potentially deliver several million short packets per second.

Todays bridges and routers are not capable of forwarding millions of packets per second. However, HiPPI switches are relatively inexpensive and provide a high performance "media" for the interconnection of high performance bridge routers. Network Systems will deliver HiPPI interfaces for its bridge routers towards the end of this year.

Isochronous Data Routing

A fascinating application of HiPPI involves the transfer of arbitrary digital information. As long as the peak transfer rate requirement does not exceed the HiPPI burst rate of 800 or 1600 Mbps, virtually any type of data can be carried. Continuous or bursty, chunked or non-protocoled, HiPPI imposes minimal constraints on the datastream.

Examples of digital data types considered for HiPPI channels are:

- Digital High Definition TV
- Digitized voice
- Imagery
- Telemetry data

Potential Storage Subsystem Application

The client server model has been applied to files servers from PCs to supercomputers. Despite this success it has serious flaws in the current implementations. One or more computers manage a catalog of files on behalf of one or more client systems so as to facilitate sharing. However, the management computer is also used to retrieve (read) the data from storage peripherals and send a copy (write) to the client.

The server computer is clearly a bottleneck to performance. This design does not scale well and in the supercomputer range literally requires a supercomputer to provide effective file service.

Since the advent of HiPPI switch attachable storage media such as RAIDS and cartridge tape systems, a new file server model has begun to evolve. The obvious but essential idea being that the management computer and the client computers can share direct access to storage peripherals. Access to catalog information by clients need not be across the HiPPI fabric since it is a low bandwidth application.

Most who first consider this concept are aghast that the storage peripheral is left so exposed to unmediated access. The fear of unauthorized access or worse, erasure of valuable data immediately arises.

However, let's consider the following:

- The catalog information will probably exist on private media for optimized access by the server system.
- HiPPI is inherently a simplex media (with flow control). A "read-only" connection can be established from the peripheral to the client system to prevent unauthorized erasure.
- HiPPI switch fabrics will support access control mechanisms such that connection to specific ports may be restricted to specified clients.
- Adequately intelligent peripherals may be instructed by the server computer to stage data, create a simplex connection to the client and then transfer the data as flow controlled via the HiPPI connection.

Many objections can be raised about this concept but equally many solutions have been discussed. No one has yet demonstrated such a system but the author has reason to believe that a commercial implementation will be available in less than a year. The advantages, both technical and economic are so compelling that it must be taken seriously.

Related Emerging Standards

Although this paper has focused on HiPPI because it is here now, there are other standards that will augment or in some cases replace HiPPI for similar needs.

Fiber Channel

Fiber channel is also an emerging computer/peripheral interface spanning a wide performance spectrum up to roughly a Gbps. Like HiPPI, it is also fundamentally a point to point, connection oriented interface.

Network Systems expects to see a demand for fiber channel to HiPPI bridges. Fiber Channel is also well suited for multi-pointing via switches.

SONET

The Synchronous Optical NETwork standards have been adopted by most telecommunications companies on a world wide basis. Signaling rates and multiplex framing standards have been defined from 51.84 Mbps (OC-1) to 2.488 Gbps (OC-48). The large telephony market is expected to create a supply of inexpensive SONET standard components which may be used for data oriented applications.

SONET is also seen as the basis for a national communications infra-structure capable of supporting gigabit per second data applications. As previously stated, a HiPPI over SONET bridge is under development at Network Systems.

ATM

Asynchronous Transfer Mode is associated with SONET and is also promulgated by the telephony industry. Based upon cell relay concepts, ATM will eventually support the economic carriage of bursty data over wide area or metropolitan virtual circuits.

Currently envisioned data applications hide the existence of the cell fabric from the user. The effect, however, will be to allow the cost effective extension of gigabit scale networks over geographic distances.

Conclusion

The HiPPI standards and HiPPI switches are expected to have a significant near term impact on the design and use of mass storage systems. The least optimistic projections recognize the availability of a widely supported standard which offers an order of magnitude improvement over currently available data rates for access to data. Additionally, the creation of an open computer peripheral channel standard is stimulating the development of high performance, cost competitive peripherals accessible from many computer platforms.

More far reaching is the possibility of new client server implementations for mass storage access. The first step implementations are expected this year, with multiple vendor support for direct device access by

1993. Related standards promise geographic access to mass storage libraries at gigabit per second data rates by the mid 1990s.