

ORGANISATION EUROPEAN POUR LA RECHERCHE NUCLEAIRE
CERN EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

CERN/ECP 97-007
15 March 1997
update 28 January 1998

**What is GSN and can it be used for high-energy
physics data acquisition ?**

Arie Van Praag, CERN, ECP, 15 March 1997

Abstract:

In 1989, a completely new technology emerged for fast data transfer and networking using non-blocking full crossbar switches, called HIPPI. For high-energy physics this new technology brought a number of new possibilities not available before, such as fast data distribution and event building. Using HIPPI for data distribution over a number of workstations is very successfully demonstrated by the NA48 experiment.

And now, almost 10 years after its introduction, the Compass experiment plans to do event building using a large HIPPI switch.

Today a new standard, the Gigabyte System Network (GSN), is emerging for computer networking using fast, full-duplex connections with a total bandwidth of 12.8 Gbits/s. This paper describes the Gigabyte System Network, including the switch structure. Some examples will be given to show how this standard can be used for future high-energy physics data acquisition.

What is the Gigabyte System Network and can it be used for high-energy physics data acquisition ?

Arie VanPraag, CERN, ECP, 15 March 1997

Abstract:

In 1989, a completely new technology emerged for fast data transfer and networking using non-blocking full crossbar switches, called HIPPI. For high-energy physics this new technology brought a number of new possibilities not available before, such as fast data distribution and event building. Using HIPPI for data distribution over a number of workstations is very successfully demonstrated by the NA48 experiment.

And now, almost 10 years after its introduction, the Compass experiment plans to do event building using a large HIPPI switch.

Today a new standard, the Gigabyte System Network (GSN), is emerging for computer networking using fast, full-duplex connections with a total bandwidth of 12.8 Gbits/s. This paper describes the Gigabyte System Network, including the switch structure. Some examples will be given to show how this standard can be used for future high-energy physics data acquisition.

Introduction

In 1989, a research group of the Los Alamos National Laboratories (LANL) computer center started work on a standard for fast data transport between mainframe computers, which was accepted by ANSI under the name HIPPI. It was later called HIPPI-800, and registered as ANSI X3.183-1991 [1]. Its speed of 100 MBytes/s for a simplex connection and its relatively easy implementation, added to the possibility of building networks with data switches [2], meant that it was rapidly accepted in a number of computer centers. HIPPI also proved to be an attractive solution for data collection [3,4] and event building [5,6,7,8] in high-energy physics (HEP) data acquisition. It was also in 1989 that the first activities around the HIPPI standard began at CERN. The standardization of the high-speed Peripheral Interface bus (PCI), for workstations, and the introduction of HIPPI to PCI interfaces has brought HIPPI-800 to platforms such as Pentium PCs and Alpha workstations. [9,10,11,12]. The same modules made in the PMC form-factor brought HIPPI networking to the VMEbus world.

HIPPI-800 in the NA48 experiment

The first application of HIPPI-800 in high-energy physics data acquisition was the

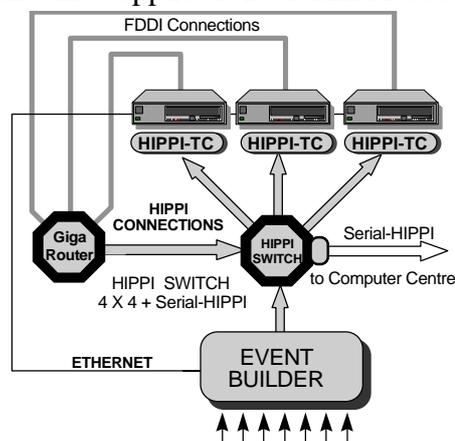


Fig 1: The NA48 data switch

NA48 CP-violation experiment that started to operate in 1964 (Fig 1). Data coming from the level-2 trigger is collected in a large memory in the event builder and distributed to one of the level-3 APX 5000/200 workstations equipped with Turbo-channel-HIPPI interfaces. As the drivers for these interfaces can handle ASCII data only, and some form of protocol is needed to transfer data to the central computer center, an FDDI output is used from these workstations. A Giga router is needed to couple the data into the 10 Km Serial-HIPPI

link to the computer center. The installation is calculated to handle a 250 MByte block of data every 15 seconds. An upgrade of the installation will be necessary if larger data blocks need to be handled, entailing the reconstruction or replacement of the event builder and an extension of the number of switch ports used. At the same time, the Gigarouter, which has a limited bandwidth, should be equipped with a faster interface. If HIPPI interfaces and drivers that include a TCP/IP protocolised transfer are adapted the Gigarouter can be avoided altogether.

Compass

On the Compass experiment, which is currently under construction, it is planned to apply a large HIPPI switch as a third-level trigger and for event building, coupled to a workstation farm. The data are transferred to the computer center using the same Serial-HIPPI link as NA48. Two set-ups are being considered, the first bringing the data from VMEbus units to the switch and delivering complete events to the individual workstations. The architecture is simple but not very flexible. Due to a mechanical problem, HIPPI interfaces for VMEbus modules using the PMC form-factor have the connector staying out in the front. The second set-up will deliver the data directly to the workstations. The workstations are interconnected by a HIPPI switch so that data can be exchanged to build the events. The first solution has the advantage of simplicity for both software and hardware architecture. The second has the advantage of being more flexible, although the software is more complex. In both schemes it is also easy to implement the TCP/IP protocol, the standard for transferring data to the computer center.

Introduction and Application

It can be concluded from the above two examples that it takes about 5 years for the first application of technology with promising characteristics for HEP data acquisition to become operational, and almost 10 years before it is used to its full potential. During this period a number of new standards have emerged in the same speed range. However, new technologies lead to the development of a new networking standard with much higher performances. According to its bandwidth of 800 MBytes/s in each of the full duplex directions or 1.6 GBytes/s total bandwidth it is called "Gigabyte System Network" or GSN. Compatibility with HIPPI-800 and Serial-HIPPI is part of the new standard offering backward compatibility with existing installations.

What is GSN?

The GSN PH (PHysical) specification [13] describes the physical level for a point-to-point full-duplex link interface, using flow-control for reliable transmission of user data. The speed is 6400 Mbits/s in both directions. Distances of 50 m can be bridged with parallel copper cables, while distances of 200 m can be reached with parallel fiber-optic cables.

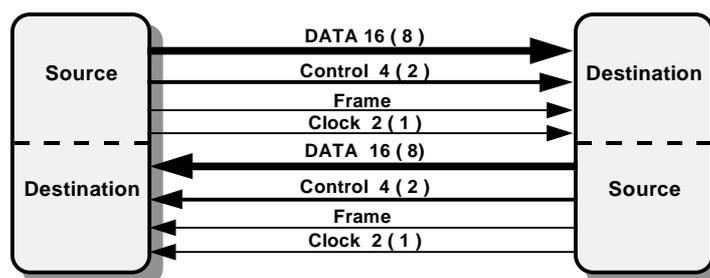
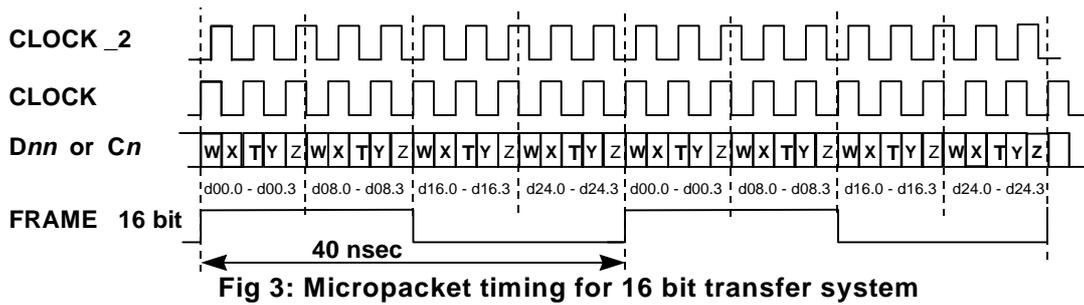


Fig 2: GSN Principle

Small, fixed-size micropackets provide an efficient, low-latency structure for small transfers and can be combined with a component for large data transfers.

The link possesses a



symmetrical structure in both directions. The opposite direction is used to return feedback messages. The connection possesses either 8 or 16 data-lines (Fig 2), one or two control lines, a frame signal and either one clock signal in 8-bit systems or two clock signals with constant phase shift of about 90^0 , in 16-bit systems. Messages and data are sent in micropackets of 32 data-bytes and 64 control bits. They are delimited in the data stream by the frame signal. (Fig 3) A "4-to-5" encoding is used to keep the DC balance constant. The data sent by the source are synchronized to the clock signals of 500 MHz in an 8-bit system and 250 MHz in a 16-bit system. In both cases, each

No Data Bits	No Control Bits	Frame Signal	Clock + Freq.	Use
8	1	1	1 500 MHz	Parallel Fiber
16	2	1	2 250 MHz	Copper Cable

Table 1: Connection Signal Overview

half-phase of the clock carries a set of data bits with the result that the transfer of a micropacket takes 40 nsec. Open spaces are filled with "Null micropackets" to maintain the DC level balanced. Table 1 gives an overview of the different signals. Data are sent over the link in the form of a message which is formed by one or more micropackets. Header packets are added at the start of a message. The MAC Header contains the 48-bit ULA network address, identifying the final destination and the 64 bit SNAP header that defines the protocol type. Flow control is maintained by the destination sending control micropackets in the return way.

Link Structure

The connection is devised in 4 virtual channels in each direction, in order to use efficiently the full bandwidth of the link. (Fig 4:). They are called VC0 to VC3. VC0 is for messages with up to 68 data micropackets (2176 Bytes). VC1 and VC2 are both used for messages with a maximum size of 4100 data micropackets (128 KBytes), and for admin request messages. VC2 also carries the returning admin micropackets

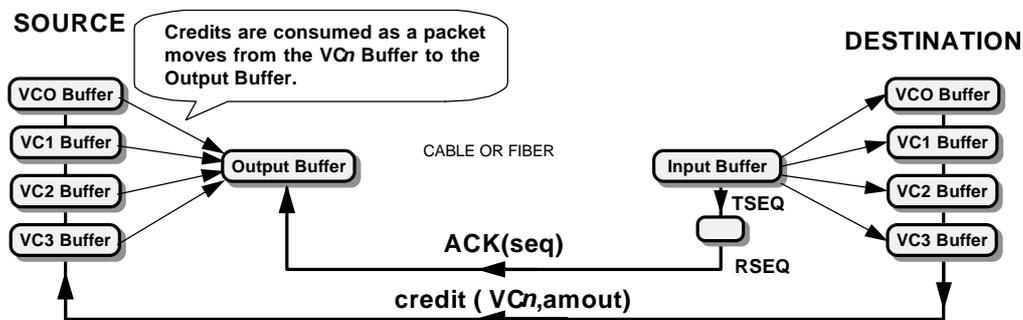


Fig 4: GSN Channel Distribution

for the opposite direction. VC3 is used for messages with a maximum size of 4 GBytes. Common to all VCs is a header micropacket and a tail micropacket which can be the same if the length = 1.

Flow Control

The control word (Fig 5) contained in the micropacket handles all the information needed for flow control. To transfer data, the source sends a request. If the latter is accepted by the destination, a number is returned that represents credits. These credits correspond to the number of micropackets that can be received. Each VC output handles its own credits, as indicated by the pointer in the VCR field. The output buffer of each VC subtracts from the credits the number of micropackets sent and adds new credits received from the corresponding VC destination. Acknowledgment (Ack) is done by comparing the sequence numbering of the micropackets sent by the link source buffer (TSEQ) and the sequence numbering of the destination buffer (RSEQ). Equal

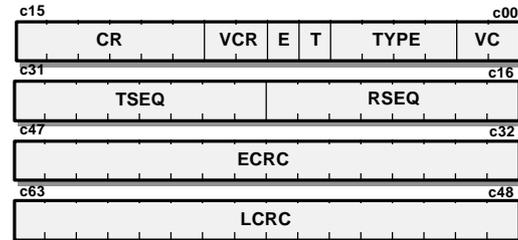


Fig 5: The Micropacket Control Word

NAME	No Bits	FIELD	CONTROL FUNCTION
VC	2	C01 - C00	VC Selector
TYPE	4	C05 - C02	Information Type
T (AIL)	1	C06	Last Micropacket
E (ERROR)	1	C07	ERROR
VCR	2	C08 - C09	Virtual Channel for Credit Addition
CR	6	C10 - C15	Number of Credits
RSEQ number	8	C16 - C23	ACK. Sequence
TSEQ number	8	C24 - C31	Transmission Sequence
ECRC	16	C32 - C47	End to End Checksum
LCRC	16	C48 - C63	Link Level Checksum

Table 2: Control Word Functions

numbers mean that all micropackets sent by the source are received by the destination. Table 2 gives an overview of the control word functions while Table 3 shows the formats of all types of micropackets.

	Reset / Initialize	Null	Credit Only	Header	Data	Admin
Data Byte Contents	0	0	0	32 Bytes header Information	32 Bytes Data	Administrative Information
VC	0	0	0	any	any	Request on VC1 Request on VC2
TYPE(hex)	2,3,4,5	7	A	9	8	F
Tail	1	0	0	= 1 on last micropacket of message	= 1 on last micropacket of message	1
ERROR	0	0	0	= 1 if Error	= 1 if Error	= 1 if Error
TSEQ	xFF	xFF	increments	increments	increments	increments
RSEQ	1	ACK	ACK	ACK	ACK	ACK
VCR	0	0	any	any	any	any
CR	0	0	any	any	any	any
LCRC	single	single	single	single	single	single
ECRC	single	single	single	accumulating	accumulating	single

Table 3: Summary of Micropacket Contents

Scheduled Transfers

A scheduled transfer provides the mechanism for the source and the destination to agree, in advance, on a number of parameters, so that block size and message size are defined in advance for both ends. This mechanism gives some extra overhead during the set-up of a connection. However in the case of equally repeating transfers this has to be done only once. The actual execution of the scheduled transfer will be handled by the adapter. In GSN, the scheduled transfer includes the parameters needed to implement a fully inter-operational memory-to-memory transfer that bypasses the operating system. Software latency is therefore limited to the 1-time set-up, and speed is limited only by the hardware involved, such as network bandwidth and DMA channels.

Error Checking

Error checking is done by two 16-bit Cyclic Redundancy Checks (CRC), the Link-CRC (LCRC) and the End-to-end CRC (ECRC). The LCRC covers all the data bytes and the control bits in a single micropacket, except itself. It acts on the link only and not on the VCs. The CRC formula for the LCRC is:

$$X^{16} + X^{12} + X^5 + 1$$

The ECRC checks all the data bytes of a message, and can thus cover more than one micropacket. It does not check the control bits. As the ECRC checks the data contents of a message it is calculated and maintained independently for each VC. The CRC formula for the LCRC is:

$$X^{16} + X^{12} + X^3 + X + 1$$

GSN Switches

The GSN SC (Switch Control) specification [15] describes the way the non-blocking switch handles the connections. It also specifies how the switch should be addressed to select the data path and handle different protocols. In HIPPI-800 a request - connect [16] handshake makes a physically locked connection between a source and a destination. A GSN switch (Fig 6), on the other hand, constructs a flagged or virtual connection between a VC source and a VC destination with the same number that lasts for the period of a single message. This allows the possibility of interleaving messages from different input ports or VCs via the same physical link. Initialization

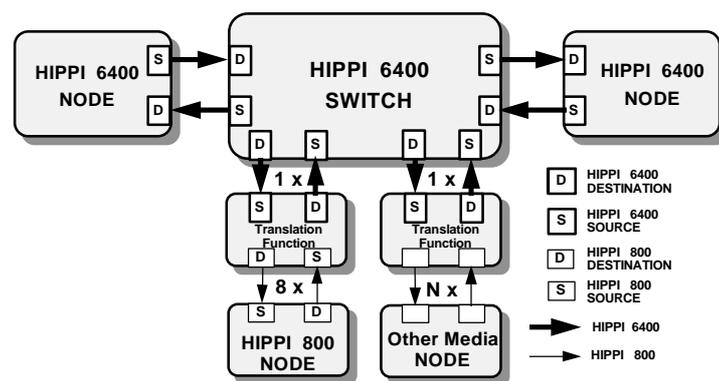


Fig 6: GSNbasic switch

of a GSN switch is done with admin micropackets, as shown in Table 4. However a destination address is contained in the header micropacket in the as a standard 48-bit ULAs [17] conform to the IEEE 802 standard. To be backward compatible with HIPPI-800 a 36-bit ULA prefix that extends with the 12-bit logical address [16] to 48-

bit will be reserved. Source and destination addressing is not supported. The switch specification provides multiple-path addressing and broadcast possibilities. HIPPI-800 physically connects to so-called "Translation Function" adapters that do this conversion transparently. Up to eight HIPPI-800 ports connect in this way to a single GSN port.

Byte	Function
0	Key
1	Hop Count
2-3	Destination Admin Register designates a local register within an element
4-7	Destination Admin Element Address Destination element address in a GSN domain
8	Admin Command
9	Status Flags Return Hop Count
10-12	Source Admin Register designates a local register within an element
12-15	Source Admin Element Address Source element address in a GSN domain
16-31	Data Register

Table 4: Adminmicropacketformat

Switch latency

The first admin micropacket should arrive before a switch starts to decode the routing. Such latency takes at least 40 nsec, to which should be added the decoding and set-up time for the VC. The latter should be short due to the high clock frequency. Altogether the latency is supposed to be shorter than 0.5 μ sec. For HIPPI-800 connections the time to transfer the I-Field into a header micropacket, which is again a simple insertion, should be added. The conclusion is that the latency for a classical GSN connection will be no longer than in a HIPPI-800 switch.

Prospects and Future Products

The GSN specifications are very well advanced and should be sent for official standardization in Spring 1998. At the same time, work on hardware is advancing. The first prototypes of a silicon Integrated Circuit are already available since late 1997. Prototypes of a GSN switch that use the Integrated Circuit should be ready for β testing around the Summer 1998. A tester for HIPPI-800 and GSN is under construction and should also be available early in 1998. Most of the early products will initially use the copper connection. Some of the more complicated features, such as broadcasting and multi-point distribution will be included later. Also some workstation manufacturers are working on interfaces.

Using a GSN for HEP data acquisition systems

Event building with a GSN Switch for LHC Experiments

In HEP data acquisition, event building is one of the most interesting applications for a fast data switch. HIPPI-800 switches have been demonstrated to be able to execute this function limited by the maximum available switch size of 32 X 32 [18]. The combination of several switches results in either a speed penalty or in the doubling of the bypass

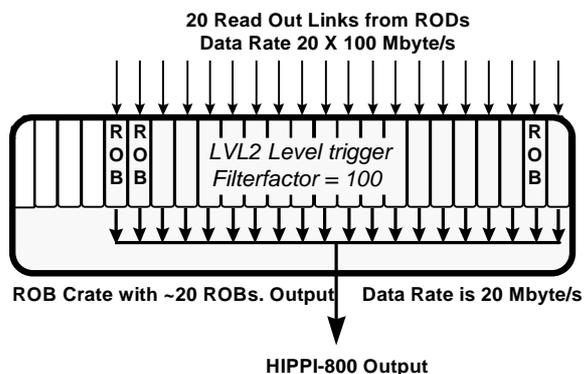


Fig 7: The ROB Crate as used in the example.

connections. By comparison, the announced GSN switch possesses a capacity of 32 X 32 ports. Using the adaption function, each port is able to connect up to eight HIPPI-800 ports (Fig:6).

If a future workstation farm has GSN inputs, the event building speed that can be

GSN Ports	HIPPI-800 Ports	GSN Bandwidth	HIPPI-800 Bandwidth	Average Bandwidth for Event Building
0	256	n.a.	100	100
2	240	1 600	24 000	6.6
4	224	3 200	22 400	14.2
8	192	6 400	19 200	33.3
12	160	9 600	16 000	60
16	128	12 800	12 800	100
20	96	16 000	9 600	60
24	64	19 200	6 400	33.3
28	32	22 400	3 200	14.2
32	0	25 600	n.a.	800

Table 5: Average Event building bandwidth versus port distribution in MBytes/s

expected is a function of the port relation used for GSN and HIPPI-800 connections (see Table 5). Practically average speeds will be about 15 % lower because of the VC distribution. Taking as example an experiment where a LVL2 trigger reduces event rates by a factor 100 (as is done in Atlas [20,21]) and brings the data rate down to 100 Mbytes/s for each of the 2000 channels. If 20 of this channels are combined in a data concentrator, the result is 100 outputs with 20 Mbytes/s bandwidth each. This bandwidth can be handled by without any problems by HIPPI-800 connections. From the many possibilities as given in Table 5, the highlighted range can be used.

The optimum is 104 HIPPI-800 inputs and 19 GSN channels for the output. This combination allows an event building speed of 70 Mbyte / HIPPI-800 port or a maximum 7 GBytes/s. The total data rate at the event building level is however not more than 2 GBytes/s. This can be handled with only three GSN ports. How many of

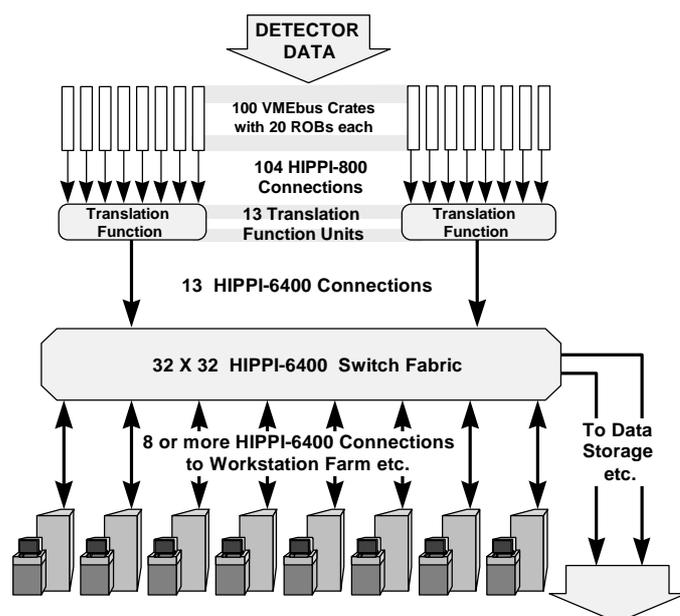


Fig 8: A typical event building application using ATLAS as an example

these fast ports are actually needed depends in reality on the capacity of the processors in the farm. In the example shown in Fig 8: eight processing clusters are connected. The remaining channels can be used for communication and for connection to storage servers. A factor 3 in spare bandwidth is available for detector upgrades.

Protocol insertion can be fast using look-up tables. If scheduled transfers are applied the link set-up has to be done only ones at the start. Event numbers can be handled by the data register of the admin

micropackets. For connections between the VMEbus Read Out Buffers (ROB) and the switch, the use of fiber optic connections such as Serial-HIPPI [19] would be an advantage as it avoids copper cables, which are difficult to handle, and spans longer distances. The same principles as shown here for the ATLAS experiment can be adapted to the other large LHC experiments.

A data concentrator for small experiments

Using the same principles as described before, the data acquisition system for a small experiment can be built around a GSN switch and its translation function adapters. The model uses some of the data concentrator principles of the NA48 experiment [11]. The data coming from the detector is stored in a number of buffer memories using VMEbus modules (Fig 8). These buffer memories are sent sequentially to the translation function of the switch. The data are assembled event by event and sent either to a workstation or to a storage vault. In the same manner, the trigger using either HIPPI-800, Serial-HIPPI, Fiber Channel or 100-Base T Ethernet data can be assembled and sent to a trigger processor. The different control functions find a return path through the switch and the data channels. The flexibility of this architecture is advantageous as it enables partial event data to be sent to the trigger processor. And at the same time these trigger data and trigger decisions can become available to the workstations through the switch. When broadcast functions are used these data can even be sent to both parts at the same time. If connections to a central computer centre are necessary, this can be accomplished by using either a switch port, or a port on a translation function.

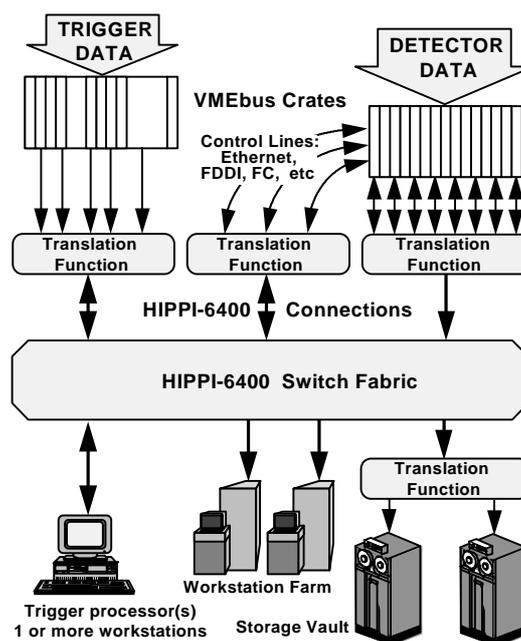


Fig 8: Hypothetical data acquisition for a small experiment

Or a more universal solution using PCI

Both examples use translation function adapters to connect slower network technologies to a GSN switch. In both cases HIPPI-800 and Serial-HIPPI are mentioned for this data collection. In large data acquisition systems, however, a multitude of data transfer technologies are used, depending on data rate, distances and costs. A single translation function should thus be able to accept a whole range

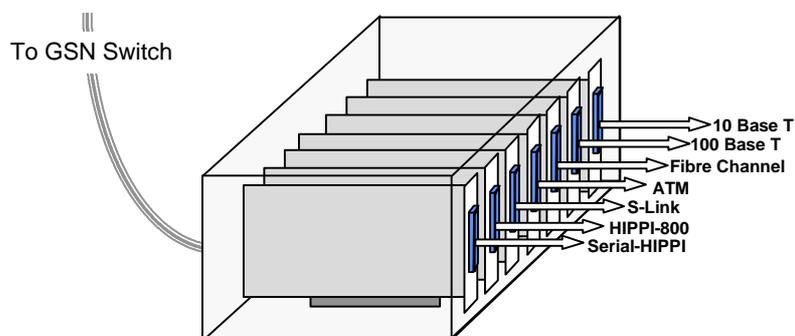


Fig 9 A GSN to PCI Interface Box

of different technologies to be successful in a large number of HEP data acquisition systems. It should not be forgotten that a low price per channel is important as the number of channels becomes very high. If PCI bus circuits available in industry can be used, then the latter condition is fulfilled. To combine this adapters in the best way to the bandwidth the GNS they should use the 64 bit version of the PCI specification and be able to run 66 MHz. In conclusion, a translation function adapter that couples 8 to 10 PCI slots to the GSN channel can be very successful in HEP data acquisition .

Conclusion

GSN certainly brings many advantages in high-energy physics data acquisition, because of the switches and the large fan-in translation function adapters. Furthermore, if these translation function adapters connect the PCI bus to the GSN channel, a very fast and flexible medium would be obtained. The channel interfaces for GSN are under development as CMOS chips which can keep costs of interfaces and switches to a relatively acceptable value.

If Serial-HIPPI becomes available at acceptable prices it might become an attractive way to couple many partial event buffers to GSN. As the partial events are buffered in VMEbus ROB modules, cheap Serial-HIPPI modules with a high sustained throughput will hopefully become available in the PMC form-factor. An even more flexible option is made possible by using the GSN channel to interface to a black box with a multitude of generally accepted interconnect possibilities such as PCI. In this case a multitude of technologies can be coupled into the gigabyte networking devices for large bandwidth data acquisition and event building.

In practice it is shown that the time span from the introduction of a promising new technology or standard to a real data acquisition application is 5 years or more. This time-span is needed to learn the technology and test the usefulness of the standard and its components for the specialized application that is HEP data acquisition. If GSN is to be considered as a serious candidate for third-level triggering and event building in the LHC era, it may be a good idea to start evaluation activities early.

References

- [1] High Performance Parallel Interface Mechanical, Electrical, and Signaling Specification, HIPPIPH, ANSIX3.183-1991 Rev 8.3.
- [2] High Performance Parallel Interface Mechanical, Electrical, and Signaling Specification, HIPPI SC, ANSIX3.210-1992, Rev 4.4.
- [3] HIPPI Developments for CERN Experiments, A. Van Praag, et al, CERN/ECP 91-28, 7 November 1991. Presented at IEEE NSS 199, <http://www.cern.ch/HSIhippi/applic/otherapp/hppidef.htm>
- [4] Data Transfer and Distribution at 70 MBytes/s, J-P.Matheys, et al., CERN/ECP 93-7. Presented at IEEE RT 1993
- [5] Atlas Technical Proposal, CERN/LHCC/94-43, LHCC/P2, 15 December 1994, ISBN 92-9083-067-0, WWW: http://atlasinfo.cern.ch/Atlas/GROUPS/TPTP_ps.html
- [6] Ralf Spiwoks, Evaluation and Simulation of Event Building Techniques at the LHC, Ph.D. Thesis. University of Dortmund, Germany, 1995 (CERN-Thesis- 96-002)..
- [7] Testing HIPPI Switch Configurations for Event Building Applications, Arie Van Praag, Ralf Spiwoks, Robert van der Vlugt, CERN, CERN/ECP 96-15, September 1995, Presented at the SOZOPOL-96 workshop on Relativistic Nuclear Physics, Sozopol, Bulgaria, October 1996
- [8] Switching techniques in data acquisition systems for future experiments. M.F.Letheren, CERN Geneva Switzerland. Presented: CERNsummerschool of computing 1994.
- [9] Low Cost, High Performance LAN's Constructed with Serial HIPPI, G.Mcalpine, McAlpine Research and Development Inc. and Dr. J.R. Wilson, Avaika Network Corporation. January 1995. Presented at Interop March 1995.
- [10] Overview of the use of the PCI Bus in Present And Future high-energy physics Data Acquisition Systems, Arie Van Praag et al, CERN Geneva, CERN/ECP 95-4, 3 January 1995, <http://www.cern.ch/HSIhippi/applic/pcihippi/pcihippi.htm>.
- [11] PCI - HIPPI Interface Modules, G. Antchev, G. Georgiev, S. Piperov, I. Vankov, INRNE-BAS, Sofia, Bulgaria. R. A. McLaren, A. van Praag, CERN, Geneva, Switzerland. O. Orel, IHEP, Prodvino. D. Gillot, A. Guglielmi, Digital Equipment Corporation, Joint project office, CERN, CERN/ECP 96-14, 18 September 1996. Presented at the SOZOPOL-96 workshop on Relativistic Nuclear Physics, Sozopol, Bulgaria,
- [12] HIPPI: It's Not Just for Supercomputers Anymore. DonTolmie, LANL, DonFlanagan, HNF. Data Communications Magazine, may 8, 1995.
- [13] High-Performance Parallel Interface -6400Mbit/s Physical Layer (GSN-PH), DonTolmie LANL et al, X3T11/Project 1213-D/REV 1.0, January 28, 1997, <http://www.noc.lanl.gov/~det/c6400PH.html>
- [14] GSN-PH, Electrical Interface Architecture Specification, Hansel Collins, SGI, January 2 1997, <http://www.noc.lanl.gov/~det/c6400PH.html>
- [15] High-Performance Parallel Interface -6400Mbit/s Physical Switch Control (GSN-SC), Roger Roland e-Systems, et al, X3T11/Project 1231-D/REV 0.8, January 24 1996, <http://www.noc.lanl.gov/~det/c6400SC.html>.
- [16] HIPPI-SC, High-Performance Parallel Interface -Physical Switch Control (HIPPI-SC), ANSI X3.222-1993, April 9, 1996
- [17] The following documents handel the basic assignment of specific logical addressing for network services: RFC 1042, RFC 2067, RFC 1112, RFC 1131, ISO/IEC 9542:1988, ISO/IEC 10589:1992, ANSI/IEEE 802.1D-1990.
- [18] Testing HIPPI Switch Configurations for Event Building Applications, Arie Van Praag, Ralf Spiwoks, Robert van der Vlugt, CERN/ECP 96-15 (18 September 1995), Presented at SOZOPOL-96 workshop, Sozopol, Bulgaria, 30 September - 6 October 1996
- [19] HIPPI 800 and 1600 Serial Specification (HIPPI-Serial Rev 2.6), DonTolmie et al, ANSI X3.300-199x, June 11, 1996.
- [20] Atlas Technical Proposal, CERN/LHCC/94-43
- [21] The Atlas DAQ and Event Filter Prototype -1 Project, The Atlas DAQ group. Presented at CHEP'97