Potential HEP Applications of a New High Performance Networking Technology

Arie Van Praag[11], Ben Segal[1]

[1] CERN, IT/PDP/TE,

Abstract

In 1989, a completely new 100MByte/s technology emerged for fast data networking using non-blocking full crossbar switches, and was called HIgh Performance Parallel Interface (HIPPI). For high-energy physics this new technology brought a number of new possibilities such as fast data distribution and event building. Using HIPPI for data distribution between an experiment's data Acquisition and a number of workstations has been very successful in the NA48 experiment.
Today a new standard, the Gigabyte System Network (GSN), is emerging for computer networking using fast, full-duplex connections with an effective bandwidth of 800 MByte/s in each direction. This paper describes GSN, including the switch structure and its very low latency protocol called Scheduled Transfer (ST). An overview of available components will be given, together with some examples of how this standard can be applied in high end computing and in future high-energy physics data acquisition.

keywords GSN,Network,SAN,high performance,DAQ

1 Introduction

In 1988, a research group of the Los Alamos National Laboratories (LANL) computer center started to 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 MByte/s for a simplex connection and its relatively easy implementation and the possibility of building networks with data switches[2], made it 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 in 1989 that the first activities dealing with HIPPI began at CERN and cumulated in successful use in the NA 48 experiment. Now, 10 years later, a new very high speed networking standard has emerged.

1.1 HIPPI-800 in the NA48 experiment

The NA48 CP-violation experiment started to operate in 1994 with the first application of HIPPI-800 in high-energy physics data acquisition (Fig: 1). The data coming from the level-2 trigger was collected in a large memory in the event builder and distributed to several level-3 APX 5000/200 workstations equipped with Turbo-channel-HIPPI interfaces. The drivers for these interfaces could handle raw data only. To transfer data to the central computer center TCP/IP protocol was needed, so FDDI output was used from the workstations and a Gigarouter was used to convert the multiple FDDI TCP/IP streams into a single HIPPI TCP/IP stream. A 10

[1] Corresponding Author: A. Van Praag, CERN, a.van.praag@cern.ch

Km Serial-HIPPI link[9] between the experiment and the computer center moved the data to the

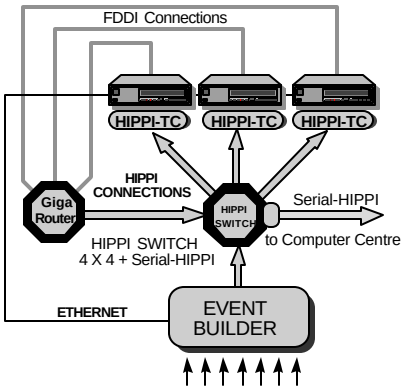

Fig 1: The NA48 data switch

central storage facility. The installation was able to handle a 250 MByte block of data every 15 seconds, and was very successful in the first three years of data taking.

1.2 Time for new technologies

The conclusion from the NA48 example was that it takes several years before a new technology with promising characteristics for HEP data acquisition finds an application, and that it can be usefully applied during a period of 3 to 5 years. During this period, a number of new standards such as Fibre Channel and Gigabit Ethernet have emerged within the same speed range as HIPPI (1 Gbit/s). However, the need for higher transfer speeds, advances in silicon technology and newly developed software methods led to the development of a new networking standard with much higher performance. Using the possibility to transfer data at speeds of 10Gbit/s a new standard was worked out. Without the link overhead it has an effective speed of 800 MByte/s in each of the full duplex directions or a total effective bandwidth of 1.6 GByte/s. Due to its speed it is called "Gigabyte System Network" or GSN. Compatibility with other high performance networking standards such as HIPPI-800, Fibre Channel, Gigabit Ethernet and even with SCSI is part of the new standard and offers backward compatibility with existing installations.

2 A short description of GSN

The Gigabyte System Network (GSN) is the name under which commercial products for this new network standard will be available. The development project was named HIPPI-6400, and

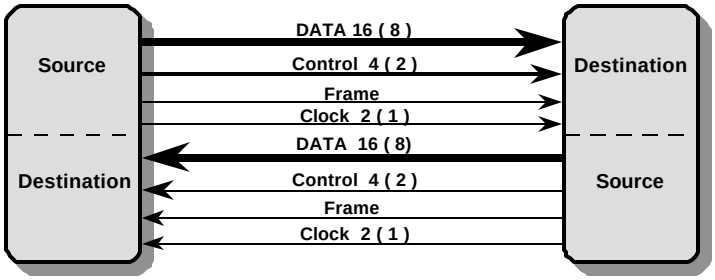

Fig 2: GSN Principle

this is still the name of the different standardization documents. The HIPPI-6400 PH (PHysical) specification [10],[11] 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 800 MByte/s in both directions. Distances of 50 m can be bridged with parallel copper cables, while distances of 200

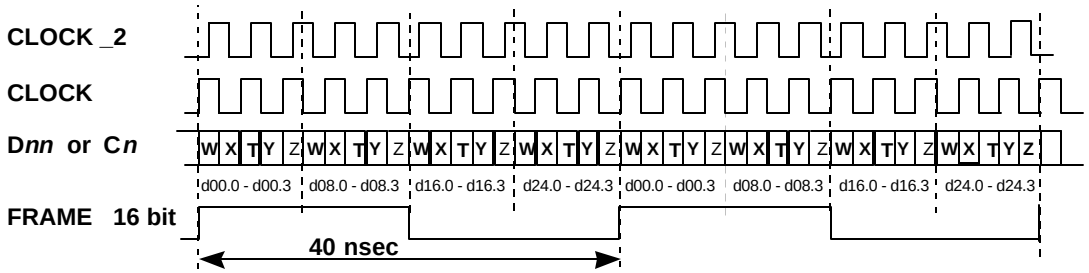

Fig 3: Micropacket timing for 16 bit transfer system

m can be reached with parallel fiber-optic cables. Small, fixed-size micropackets provide an efficient, low-latency structure for small transfers. At the same time the specification foresees that large data transfers will not influence this latency structure.

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 that contain 32 data-bytes and 64 control bits. The frame signal changes polarity for each micropacket. (Fig 3) A "4-to-5" encoding is used to keep the DC balance constant. The

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

data sent by the source is 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 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 is sent over the link in the form of a message that is formed by one or more micropackets. Header packets that contain a length field are added at the start of a message. The MAC Header

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

contains 48-bit ULn network addresses, identifying the destination and source and the protocol type. Depending on the type of traffic, a 64 bit SNAP header can define the protocol type. Table 2 gives an overview of the functions of the Control micropacket.

2.1 Link Structure

In order to use efficiently the full bandwidth of the link, the internal architecture of the link is divided into 4 virtual channels in each direction (Fig 4:). They are called VC0 to VC3. VC0 is

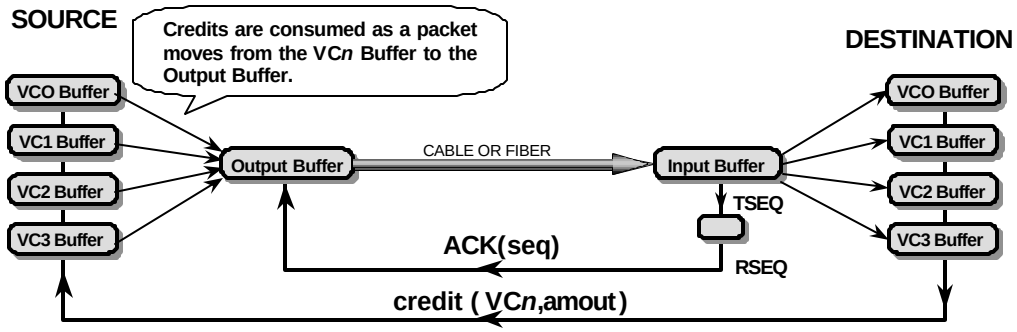

Fig 4: GSN Channel Distribution

only allowed 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 furthermore carries the returning admin micropackets for the opposite direction. VC3 is used for messages up to a maximum size of 4 GBytes. A header micropacket and a tail micropacket (which can be the same if the length = 1) is common to all VCs.

2.2 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

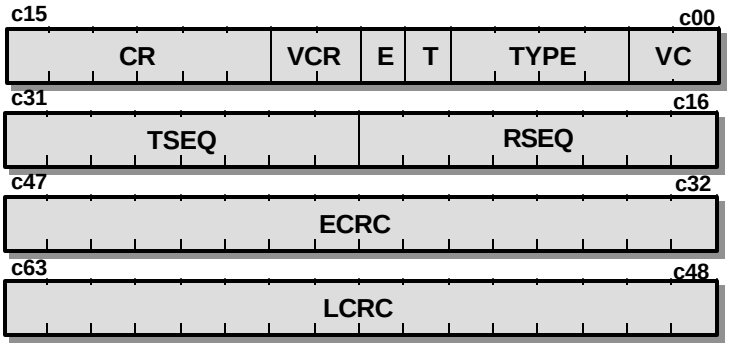

Fig 5: The Micropacket Control Word

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 numbers mean that all micropackets sent by

	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

the source were 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.

2.3 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:

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

2.4 Silicon Implementation

.

Implementation of the HIPPI-6400 PH, that uses very high speed logic, can be problematic. In order to avoid problems a silicon implementation has been developed in parallel with the standard. This silicon chip, called SuMAC, (SuperHIPPI Media Access Controller) has three interface ports; the IC port with separate 64 bit input, and 64 bit output connections; the AC port that is configurable for the two GSN modes for copper or optical connections, the DC port is reserved for manufacturer purposes. The SuMAC is mounted in a 624 pin ceramic column grid area and is commercially available.

3 GSN Switches

The GSN SC (Switch Control) specification [12] 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[13] 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

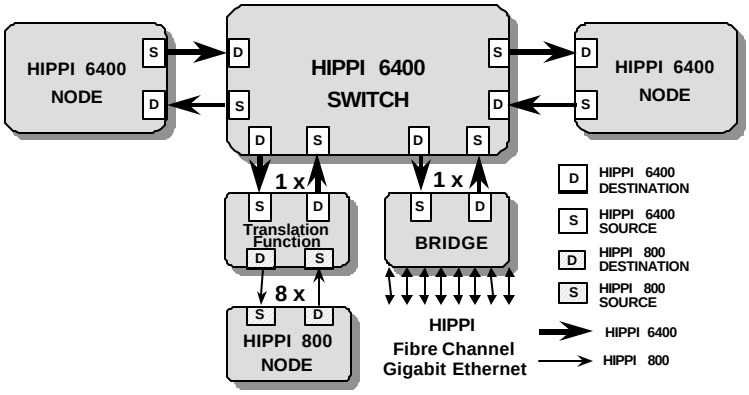

Fig 6: GSN basic switch

with the same number lasting for a single message. This allows the possibility to interleave messages from different input ports or VCs via the same physical link. Initialization of a GSN switch is done with admin micropackets, as shown in Table 4. The addressing is in accordance with the IEEE standards and uses 48 bit address fields. This means that the destination address in

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: Admin micropacket format

the header micropacket is a standard 48-bit ULA[14] conforming to the IEEE 802 standard. The switch specification provides for multiple-path addressing and broadcasting. HIPPI-800 physically connects to so-called "Translation Function" adapters that do the conversion transparently. Up to eight HIPPI-800 ports can be connected in this way to a single GSN port.

3.1 Switch latency

Before a switch action starts at least the first admin micropacket must be received and decoded, with latency 40 nsec. To this time should be added the decoding and set-up time for the VC. The high clock frequency of the Sumac chip and its related logic makes this latter quite short. Altogether the latency is supposed to be less than 0.5 μsec.

3.2 GSN Bridges

Part of the HIPPI-6400 SC[14] specification describes backward compatibility to HIPPI-800. This is done in a so called "Translation Function". In this case the 48 bit ULA uses a reserved 36-bit ULA prefix that extends with the 12-bit logical address[13] to 48-bits. Source and Destination addressing is not supported. Some extra time for transfer of the I-Field into a header micropacket must be added. This is a simple insertion that should not add more than 0.5 μsec to the switch latency. The conclusion is that, including the HIPPI to GSN conversion, the switching time may be about the same time as for a classical HIPPI switch.
Using other translation functions, a GSN bridge can perform conversions to different network technologies. Two types of bridge are known by now: a storage bridge and a network bridge. Going from the GSN bandwidth of 800 MByte/s to other network technologies that have a bandwidth of around 100 MByte/s, a fully loaded bridge will do best with eight output ports.
A Storage Bridge is intended to couple GSN to up to eight Fibre Channel ports, where each port supports an arbitrated loop. Logically the storage bridge connects to devices such as Fibre Channel tapes, disks or disk arrays. The total storage capacity depends on the size of those devices.
A Network Bridge has the function to couple GSN to different network technologies. Plug-in daughter boards can be mixed and are available for HIPPI, Fibre Channel and Gigabit Ethernet. Since the protocol conversion is done in the daughter board hardware, transfer latency can be low. However it will be different for each of the target technologies, and will correspond to the complexity of the protocol translation.

4 Scheduled Transfers

As already experienced with HIPPI, Gigabit Ethernet and Fibre Channel, the total time to transfer a unit of data is very much dependent on three parameters. The first is the necessary use of protocol stacks in the operating system. The second is packet size and the time to build the packets, and the third parameter (which is only valid for HIPPI), is that large transfers can block the connection for smaller ones using the same switching route. GSN takes special care of the last two points by the use of small Micropackets, Virtual Channels and freedom of frame size. To take care of the first point the "Scheduled Transfer" (ST) protocol[15,16] was developed together with GSN; however the use of ST is not limited to GSN and has demonstrated to be successful with other data transfer standards.
One function of the ST protocol is to bypass the operating system, by providing a mechanism for the source and the destination to agree in advance on a number of parameters. The block size, message size and memory parameters, can thus be defined in advance for both ends. To do so, a rich set of instructions is available. This mechanism gives some extra overhead during the set-up of a connection; however, in the case of repeating transfers this has to be done only once. The

resulting ST data transfer is limited to a memory-to-memory transfer that bypasses the operating system. Software latency is therefore limited to the one-time set-up, and speed is limited only by

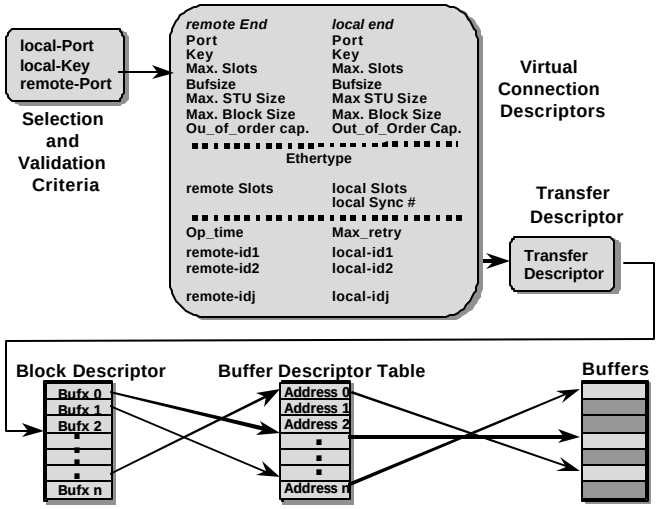

Fig: 7 Scheduled Transfer Principle

the hardware involved, such as memory bandwidth and speed of DMA channels.
The ST protocol is not exclusive to GSN and can also be used on other network technologies. Mapping recommendations are part of the standard and exist or are almost finished for ST to HIPPI, ST to Fibre Channel, ST to Gigabit Ethernet[15] and SCSI over ST[17].

5 Commercial Products Available.

This is a list of commercial products available now or coming on the market very soon.

Silicon Integrated Functions

Silicon Graphics	**Sumac chip**	**Port Interface**	**available**

Interfaces

Silicon Graphics	**Interfaces**	**Origin series**	**available**
Genroco	**Interfaces PCI 64/66**	**Sun, Compaq**	**4 Q 1999**

Switches

ODS Essential	**32 X 32**		**available**
Genroco	**8 X 8**	**Limited Addressing**	**available**
Genroco	**8 X 8**	**Full Addressing**	**4 Q 1999**
PMR	**8 X 8**		**4 Q 1999**

Bridges

ODS Essential	**Translation Function**	**HIPPI**	**available**
Genroco	**Storage bridge**	**Fibre Channel**	**available**
Genroco	**Network bridge**	**HIPPI,**	**available**
		Fibre Channel	**available**
		Gigabit Ethernet	**3 Q 1999**

			S-Link adapter	**3 Q 2000**
Cables and Connectors				
Berg	**Copper cable + Connectors**			**available**
Optical Connections [18]				
Optobahn	**Hybrid Optical**		**Parallel Fibre (12)**	**No date**
Siemens	**Paroli Optical**	**DC**	**Parallel Fibre (12)**	**available**
Siemens	**Paroli Optical**	**AC**	**Synchronous (22)**	**available**
Gore	**Noptical**		**Parallel Fibre (12)**	**3 Q 1999**

Optical fibre connections using a single fibre are still under study.

6 GSN Applications in High Performance Computing

Traditionally High Energy Physics has always been coupled to high performance computing. The simplest application will be a very high-speed backbone to connect different systems. Flexibility can be extended with the use of bridges in combination with ST and its different mappings to other technologies. A good example is the use of the ST protocol with a mapping to SCSI. The result is a GSN interface to Fibre Channel storage, where the bridge hardware will translate the GSN/ST/SCSI to F.C./SCSI. In this way, the GSN hosts can access standard F.C. disk or disk arrays in a Storage Area Network (SAN) (Fig: 8). With the choice of a SAN

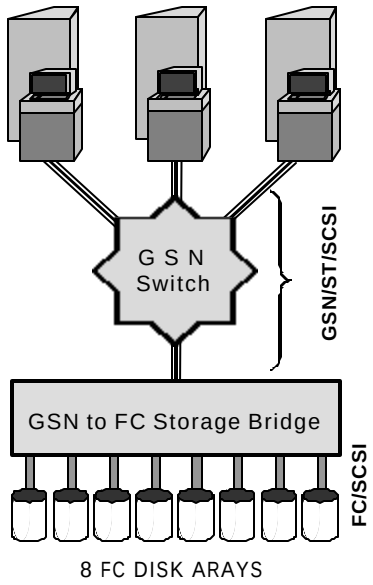

Fig: 8 A Storage Area Network

compliant file system, fast and flexible storage systems can be built. Combining this with a network bridge and different mappings of ST, a multi-protocol network can be built that combines GSN with HIPPI and Gigabit Ethernet. The result is a network where every connected device has access to a SAN (and can equally well handle IP traffic) (Fig: 9).

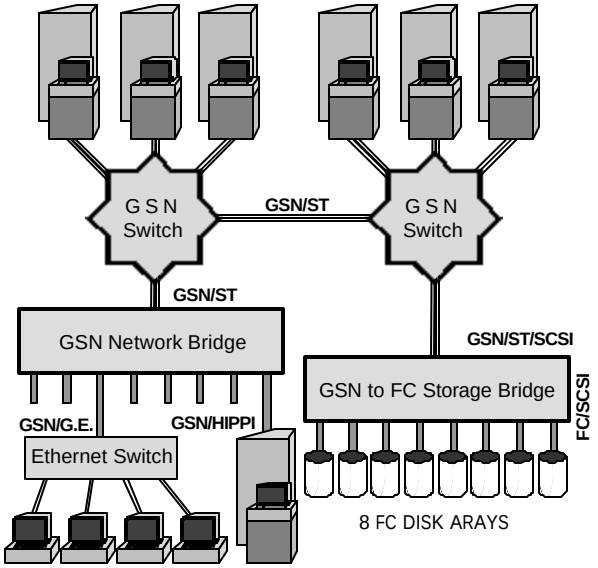

Fig 9: a GSN Network including SAN

7 GSN Applications in High Energy Physics

As GSN is a very high speed networking technology, its role in HEP data acquisition is mostly at the level of second and third level trigger and event building. Depending on data throughput of individual data strings in a detector, two 32 X 32 switches and a number of bridges should be able to handle the data of a large detector. This will be demonstrated in the following example of event building in the imaginary detector of Fig: 10 only to illustrate the principle.

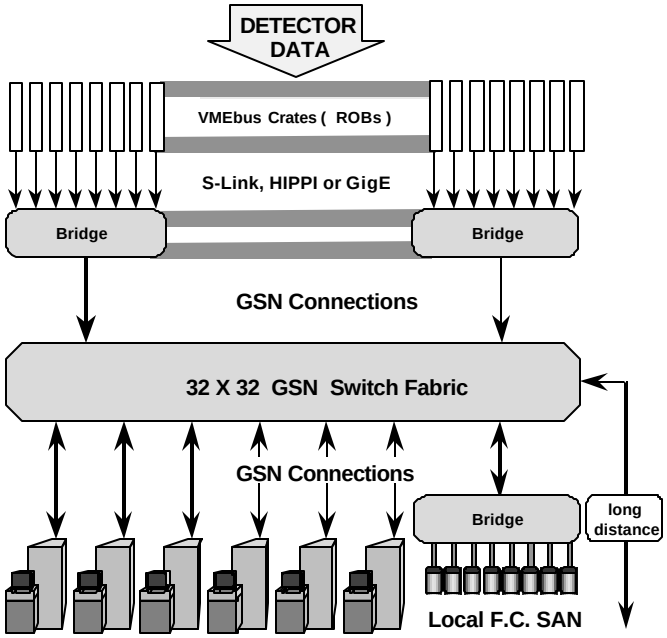

Fig 10: An example of an event building application.

The use of a single GSN switch is assumed. Today this limits the total number of ports to 32 [18]. We assume that a processor farm with 6 GSN ports (for a sustained full duplex bandwidth of 8 X 2 X 800 = 12 800 MByte/s) is sufficient to handle incoming events.
At the same time we reserve one port for local storage and one port for a remote connection to central data handling. The 24 free ports can be used to couple bridges. Given a simultaneous input bandwidth of 100 MByte/s for each port the total input bandwidth is 24 X 8 X 100 = 19 200 MByte/s. The fan-in can consist of a combination of technologies such as HIPPI, Gigabit Ethernet, Fibre Channel or S-Link. This latter can be especially interesting as it permits 4 or eventually 6 inputs for each port. In this case it would mean a total of respectively 768 or

GSN Ports	Bridge Ports	GSN Bandwidth	Bridge Ports Total Bandwidth	Average Bandwidth for Event Building	S-Link Ports
0	256	n.a.	n.a.	n.a.	1024
2	240	1 600	24 000	6.6	960
4	224	3 200	22 400	14.2	896
8	192	6 400	19 200	33.3	768
12	160	9 600	16 000	60	640
16	128	12 800	12 800	100	512
20	96	16 000	9 600	60	384
24	64	19 200	6 400	33.3	256
28	32	22 400	3 200	14.2	128
32	0	25 600	n.a.	n.a.	0

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

1152 S-Link inputs. Bandwidth of the inputs should be balanced as the total bandwidth of the bridge and its ports need to be respected.
(If more inputs are necessary a second switch can be added. However at least two to four ports should be reserved for their interconnections, resulting in a fabric with 60 ports. This would allow around 2000 S-Link channels.)
The S-Link Bridge modules can be made to insert the GSN and ST headers in hardware so that packets of raw detector data can be accepted. After a set-up time to load the ST tables this will substantially reduce the transfer latency. At the same time the ROB's can be simplified as no protocol insertion is necessary.
For the Processor farm, the type of interfaces should be taken into consideration, because native interfaces are limited by the memory bandwidth in DMA mode. (With a PCI interface, a 66 MHz by 64 bit interface can as a maximum do 528 MByte/s.)
The total available bandwidth that can be obtained depends on balancing the ports of the GSN Switch fabric between bridge ports and processor ports. Table 5 gives an overview for a 32 X 32 switch. Bandwidth is given in Mbyte/s and the number of possible S-Link ports is given on a base of 4 per bridge port.
According to the characteristics of the experiment, it can be defined which port relation has to be used and how many S-Link channels are necessary. Indeed, it should be remembered that as soon as the maximum bandwidth is not used on certain inputs to a bridge port or to a switch port, some load balancing takes place. Often the best application will be slightly outside the optimum zone.

8 Conclusion

GSN certainly brings many advantages in two fields that have always been important in high-energy physics, namely computing and storage facilities, and data acquisition, since it can handle a large fan-in and high data throughput. Even more important is the ST protocol that allows very low latency at high speed and brings interoperability between different transfer and networking technologies.
The fact that a CMOS chip is available for the GSN port logic can altogether keep costs of port interfaces low and grant good interoperability between different manufacturers.
Practice in high energy physics has shown that the time span from the introduction of a promising new technology or standard to a real data acquisition application takes at least 5 years. This time-span is needed to learn and master the technology and test the usefulness of the standard and its components for such a specialized application as 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 is a good idea to start evaluation activities early.

9 References

1 High Performance Parallel Interface Mechanical, Electrical, and Signaling Specification, HIPPI PH, ANSI X3.183-1991 Rev 8.3.

2 High Performance Parallel Interface Mechanical, Electrical, and Signaling Specification, HIPPI SC, ANSI X3.210-1992, 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/HSI/hippi/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/TP/TP_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: CERN summerschool of computing 1994.

9 HIPPI 800 and 1600 Serial Specification (HIPPI-Serial Rev 2.6), Don Tolmie et al, ANSI X3.300-199x, June 11, 1996.

10 High-Performance Parallel Interface -6400, Mbit/s Physical Layer (HIPPI-6400 PH), Don Tolmie LANl et al, ANSI NCITS 323-1998. 2.4, June 19, 1999, ISO/IEC 11518-10, http://www.hippi.org/c6400PH.html

11 HIPPI-6400 PH, Electrical Interface Architecture Specification, Hansel Collins, SGI, January 2 1997, http://www.noc.lanl.gov/~det/c6400PH.html

12 High-Performance Parallel Interface -6400, Mbit/s Physical Switch Control (HIPPI-6400 SC), Roger Roland, PMR, et al, ANSI NCITS 324-1999, 2.5 , January 27, 1999, ANSI NCITS 324-1999, d ISO/IEC 11518-xx http://www.hippi.org/c6400SC.html.

13 HIPPI-SC, High-Performance Parallel Interface -Physical Switch Control (HIPPI-SC), ANSI X3.222-1993, April 9, 1996

14 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.

15 Scheduled Transfer Protocol (ST), T11.1/Project 1245-D/Rev 3.2, ANSI NCITS xxx-199x, ISO/IEC 11518-xx. http://www.hippi.org/cST.html

16 Scheduled Transfer - Application Programming Interface Mappings (ST-API), NCITS xxx-199x, ISO/IEC 11518-xx., http://www.hippi.org/cSTAPI.html.

17 SCSI on Scheduled Transfer (SST), ANSI NCITS xxx-199x, ISO/IEC 11518-xx, http://www.hippi.org/cSCSI.html

18 HIPPI-6400-OPT Rev 1.2, ANSI NCITS xxx-199x, ISO/IEC 11518-xx., http://www.hippi.org/c6400OPT.html