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

CERN-IT-2001/003
10 Mai 2001

Coupling GSN to SONET-OC48/SDH16

Arie van Praag, Ben Segal, CERN div. IT/PDP, Geneva, Switzerland
Tvetomir Angulov, INRNE, Sofia, Bulgaria
10 Mai 2001

Abstract:

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 systems.

Coupling GSN to SONET-OC48/SDH16

Arie van Praag, Ben Segal, CERN, IT Division, Geneva, Switzerland
Tvetomir Angulov, INRNE, Sofia, Bulgaria
10 Mai 2001

Abstract

The first 10 Gbit/s network standard to become available is the Gigabyte System Network (GSN) [1]. The specification is for a so called: "Secure Network". One of the consequences is that several feedback channels are necessary, for error checking, flow control and hardware retransmission. The necessary feedback makes latency distance dependent, which means that GSN is not effective over distances longer than several 100 meters. In order to overcome this drawback a coupling to SONET/SDH was developed, including the extension of the GSN and Internet standards. The actual development is done as a joint development project between CERN in Geneva, INRNE in Sofia and Genroco in Slinger, USA. It makes use of the TURBOfibre® TBH-864 GSN bridge. Some emphasis will be given on how to handle the critical points in the very high-speed electronics.

Introduction

The GSN to SONET converter was developed at CERN as part of the Netstore project, where it brings the possibility to access network oriented storage over longer distances. Using an international standard such as SONET gives the possibility to use standard telecommunication material. It was also in mind that this kind of connections would be able to serve as connection between the future LHC experiments and the central computer center. At the same time Genroco saw this kind of connection as a good way to make a long distance connection for GSN, with either 1/4 speed with one module, 1/2 speed with 2 modules or full bandwidth with 4 modules. Using IP protocol on the SONET/SDH connection avoids the distance-related latency.

What is SONET/SDH

The Synchronous Fibre Optics NETwork (SONET) is an ANSI standard from 1985. In 1986 ITU joined the standards body and participated with the extension of the standard under their own coding of SDH. Between SONET and SDH are some subtle differences but communication between them is no problem. It is a pure physical

Optical Level	Europe ITU	Electrical Level	Line Rate (Mbps)	Payload (Mbps)	Overhead (Mbps)
OC - 1		STS - 1	51.840	50.112	1.728
OC - 3	SDH1	STS - 3	155.520	150.336	5.184
OC - 12	SDH4	STS - 12	622.080	601.344	20.736
OC - 48	SDH16	STS - 48	2488.320	2405.376	82.944
OC- 192	SDH48	STS - 192	9953.280	9621.504	331.776
OC- 768	SDH192	STS - 768	39813.120	38486.016	1327.104
OC-3072	SDH768	STS - 3072	159252.48	153944.064	5308.416

Table 1: SONET / SDH Overview

vehicle to move digital data over long distances. Several protocols are possible; the communication industry tends to use ATM (Asynchronous Transfer Mode) for the lower bandwidths up to OC12. For OC48 and higher speeds the tendency is oriented towards larger frame sizes. In POS mode (Packet Over SONET) frame sizes of up to 32 KByte are allowed, with an option for double size or 64 KByte. POS frames use PPP protocol (Point to Point protocol) and have an HDLC structure, as described in RFC2615 [6,7].

There is no CRC as such but the frames are scrambled on the transmitter side and descrambled on reception. Scrambling is done by a $X^{43}+1$ polynomial. Scrambling is easier to implement in hardware running at high clock frequencies and introduces less latency than generating a CRC. A transfer error is detected if the descrambling register is not back to zero at the end of a frame.
A second advantage of the POS mode is that the mapper allows two sideband channels for serial data. The E1-E2 channel is 3 bytes wide and foreseen for two 64kbit/s voice channels. The Data Communication Channel DCC transfers at 576 Kbit/s and is 9 Bytes wide. The DCC channel is part of the frame header and the 9 Bytes are transferred together with each frame. As long as a point to point dedicated connection or dark fibre is used these 9 bytes can be used for message handling such as flow control.
With advances in technology new speed variations with a multiplier factor of 4 are added. For the GSN to SONET transfer OC48 was chosen for two reasons: components became available at that time and network service providers were mostly equipped for OC48, in contrast to OC192.

Protocol Conversion

IP over GSN to SONET

In order to minimize overhead on a point to point connection such as the SONET channel in PPP mode, the MAC header and SNAP header are stripped off and only the IP payload itself is sent. In compliance with the PPP field assignments [5] the value of 021 should be inserted in the 16 bits protocol field for IPv4 (Fig. 2). For the moment it is not foreseen to include the compression keys for IPv6.

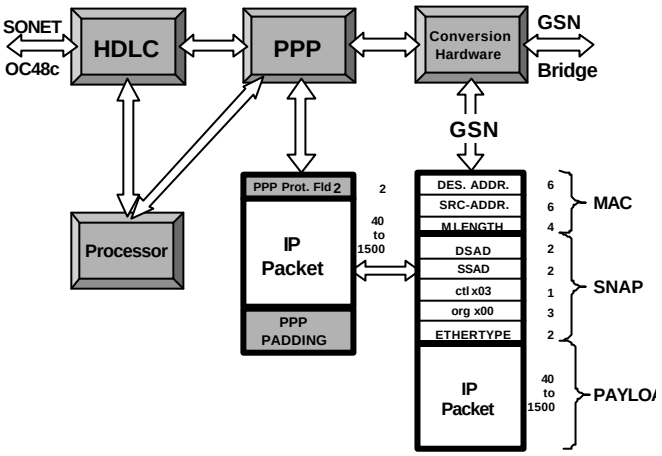

Fig 1: Conversion Etherent Frames to SONET

GSN to SONET

GSN frames are converted to IP frames to be moved over SONET. In both cases, IP over GSN or GSN frames, the GSN "End of Frame" flag is used to end a SONET frame. If frames are concatenated a single 16 bit PPP flag is inserted in the data stream and the next frame is transmitted directly after that. If no data is following the frame will be padded. If no new frame is started the transmission will return to idle patterns.

Fig 2. IP conversion compliant to RFC2615

ST to SONET

Together with the development of GSN, the Scheduled Transfer Protocol was proposed. ST [4] allows to bypass the operating system. The ST protocol is handled as payload in GSN frames. However, ST relies on the properties of a secure network such as GSN that does automatic hardware retransmission of the concerned micropackets in case of errors. Therefore ST does not have to cope with out of order

frames. In order to set up and handle the ST exchange properly it needs the "mlength" value from the MAC header. Recalculating the length value would introduce transfer latency. It was therefore concluded that the most efficient way to move ST over SONET is to move the full GSN frame over SONET. The PPP field assignment and control assignment are defined with the Internet standards body (Table 2). It has to be remarked that SONET itself does not have the flow control ST needs. It should also be noted that OC48c is 1/4 of the bandwidth of GSN. This means that in some way these parameters have to be set in the ST set-up tables during the protocol set-up. It also means that using this media for network storage can introduce congestion problems and as a consequence file corruption.

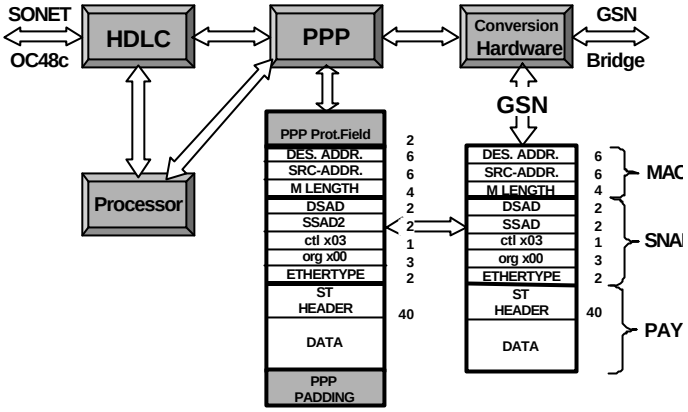

Fig. 3: GSN to SONET conversion

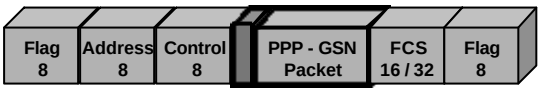

Fig. 4: GSN conversion

Protocol Name	PPP Link Protocol
STP Scheduled Transfer Protocol	020b
STP Control Protocol	820d

Table 2: PPP control words for GSN over SONET

Block Diagram and Realization

Building an electronic circuit with some parts running at 2.5 GHz is only realistic if sufficient parts are available in Silicon. A SONET protocol chip was found, the so-called mapper chip, the AMCC S4801 "Amazon" [8]. This chip includes all functions necessary for interfacing ATM data or raw data to SONET. The Serializer [9] and Deserializer [10] circuits to connect to and from the optical component are external circuits and are parts of a mixed chip set delivered by the same supplier.

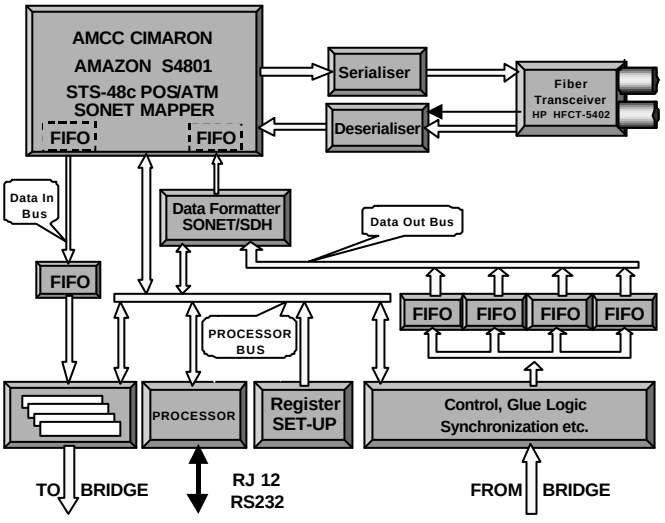

Fig. 5: Block Diagramm

Line Side Interfacing.

The connections between the Amazon and the serializer chip and deserializer chip are both 16 bit busses running at 155 MHz. Differential lines are used with LVPECL levels. They have to be terminated differentially at the destination point with a value twice the characteristic impedance or 100 Ohm. Pull down resisters of 220 ohm must be near the signal source. The connections to the

optical component are 2.5 Gbit serial data. Careful layout of the differential pairs without signal crossings to non-screened layers is mandatory here. It is good to surround these two differential pairs with ground planes and with via's connecting between this different ground layers at regular distances. Differential terminations are again a mandatory. Sometimes these terminations are already builtin in the destination components. Circuits working at these speeds should have all power connections decoupled with a 10 µHenry Ferrox component and decoupling condensers.

Clock Signals

Clock extraction from the incoming Serial Data is done in the optical component. The SONET specification for frequency stability at 2.488.320 GHz is very stringent. At the same time the clock generators for the high frequency parts are generally of a lower frequency and are internally multiplied by a PLL (Phase Locked Loop). The 19.44 MHz reference clock for the optical component and the 155.52 MHz reference clock for Serializer and Deserializer must be compliant with the jitter specifications Bellcore GR-253-core, SONET TI.105.06 and SDH G.958. This means in general a jitter specification better than 3 psec. on the final frequency (Fig. 6.). Only a few Crystal Oscillator manufacturers respect this condition, which makes a very careful selection necessary.

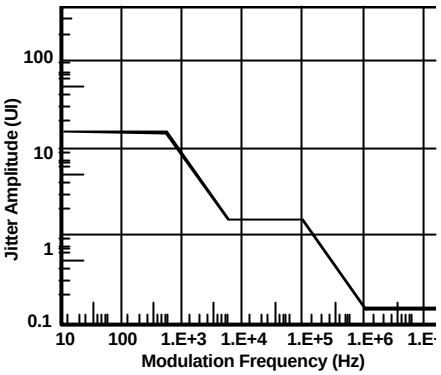

Fig. 6: Jitter Specification.

Some hardware and architectural details

The form factor chosen by Genroco for the plug-in boards of the TBH-864 GSN bridge is that of PCI boards. The interface uses 64 bit wide data and delivers bunches of four words, corresponding to a GSN micropacket. The necessary control signals are given as logic flags. Frames are delimited by a full micropacket having the "Start of

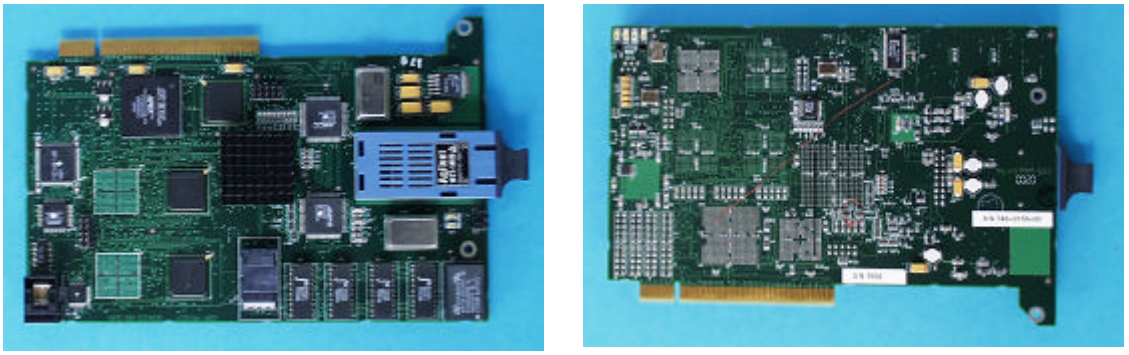

Fig. 7: Component and Wiring side of the GSN to SONET Module.

Frame" (SOF) and the "End of Frame" (EOF) flags set. For data output from GSN to SONET, the frames are first rebuilt and stored in one of the FIFO memories, one for each GSN virtual channel. For input, the frames are stored in a single FIFO memory and are split into micropackets of four 64-bit words as needed by the bridge interface. Flags are set for "Start of Frame" and "End of Frame" with the corresponding micropackets. The conversion from the incoming protocols, GSN, ST and IP to the

formats specified by SONET and vice-versa is done in programmable logic. Also timing and synchronization between the two network technologies is for a large part done in the same way. Fig. 7: gives a view of both sides of the board.

FIFO output logic and frame synchronization

All protocol conversions are done in hardware, thus reducing the latency to a minimum. Data coming from the bridge is micropacket oriented. Each micropacket contains 4 words of 64 bits. The first micropacket of a frame has the Start Of Frame bit set on all four words. The end of a frame is similarly flagged with the End Of Frame flag. The data is stored in large FIFOs, one for each virtual channel. The flags are saved in two extra bits in the same FIFO, SOF with the first word only and EOF with only the last word of the frame. (Fig. 8). Thus the boundaries of the incoming frame are well defined and kept synchronous with the data up to the output of the VC FIFOs. Here the flags are used to indicate to the Amazon chip the start of a message and the end of a message. The Amazon internally defines the PPP frames going to SONET. With large messages it will concatenate frames. It will also do the necessary padding (abending) of incomplete SONET frames or with small messages. A large state machine controls the synchronization of the different logic manipulations. Except for the FIFOs all the logic is included in a FPGA.

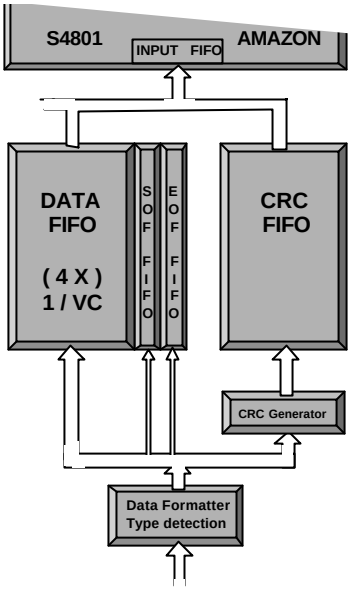

Fig. 8: Data Formating

Converting IP over GSN messages to SONET

The hardware algorithm for the protocol conversion (Fig. 8) is kept as simple as possible and executes as a pipelined function. An incoming IP message is detected from its header. Compliant to the specifications of RFC 2615 and to the ISO 3309 HDLC specification, the SNAP Header and the MAC header are deleted and the PPP header with the corresponding protocol number is formed, ready to be sent to the HDLC processor in the Amazon chip. Up to four IP streams can be handled simultaneously by GSN. Each of them is stored in the corresponding VC FIFO until the message is complete. A complete message or a FIFO almost full starts the transmission cycle. For large messages the FIFO works as a pipeline. Messages already memorized in other VC FIFOs can only start to be sent once the active message is transmitted.

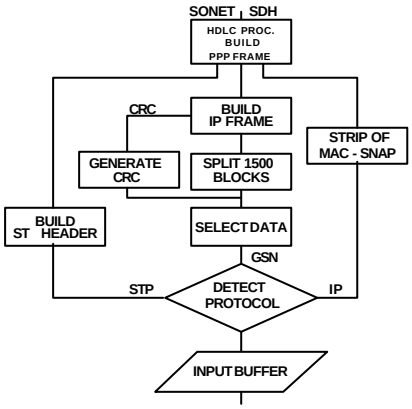

Fig. 8: Principle of Protocol Convertion

Converting GSN-ST messages to SONET

The ST protocol relies strongly on the "MLENGTH" parameter in the header. As this indicates the full length of a message including the headers it was decided to move the full message over a SONET PPP connection. Again the corresponding PPP header has to be built with the corresponding control and protocol values.

Converting GSN messages to SONET.
Messages with raw data in GSN format are transformed into IP packets. To do so the data is split into 1500 byte frames. The IP header is generated from a table. As data streams into the FIFO, the CRC checksum is generated and stored in a separate CRC-FIFO. Transfer to the AMAZON starts only if a full or abended IP frame is stored in the FIFO. This means that data coming from different sources send a sequence that is the IP header from a stored table, followed by the data from a VC FIFO and finished with the CRC Checksum from the corresponding CRC FIFO. As each of the VC FIFOs can store several 15 byte frames, no additional latency is introduced once the first frame of a message is started. If the host interleaves data over several VCs no latency will be seen at all.

Building micro packets from incoming data
Incoming data from SONET is analyzed for the data type and stored in a single FIFO. Dependent on the information available the frames are converted from IP over SONET to IP over GSN. The SOF and EOF frames are stored in the same FIFO. Since the required information is missing the SNAP header and MAC header are omitted. The ST frames are converted to GSN frames but stay essentially the same. Other data frames are converted to GSN frames. At the output of the FIFO the frames are converted to four times 64 bit micropackets and sent with the corresponding flags to the bridge.

Function of the on board Processor
The onboard processor is used for several purposes. It loads the firmware to the FPGAs, loads its own firmware and loads the set-up registers of the Amazon. With a terminal connection a series of test software modules can be activated. As therewas no support available at CERN for the Patriot Scientific PSC1000, all the software and most of the surrounding hardware have been developed by the Industrial partner Genroco.

Results and Performance

The SONET modules have been demonstrated between two Genroco TBH-864 Bridges connected with 10 km of single mode fibre. Both machines were equipped with PCI-GSN interfaces. Results obtained with this set-up were heavily dependent on the bandwidth of the PCI busses in the machines used. In one direction with an IBM

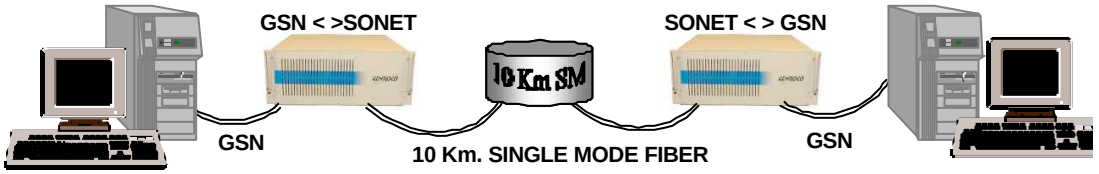

Fig. 9: Demonstration Set-Up

machine as source and the data dropped at the destination, a bandwidth of over 290 MByte/s was seen. This corresponds to wire speed on the OC48c line. In the other direction with a COMPAQ machine as source the bandwidth seen was only 160 MByte/s. The demonstration set-up was used by Genroco at SC2000 in Dallas moving 4 streams of video data simultaneously. Together with NASA this has been the two first demonstrations that show video over OC48, and the very first one with multiple streams of IP video over OC48c.

Future Work

As there is no flow control in the SONET specification it may be problematic to use this protocol converter for network storage. However, it is possible to use the DDC channels to include flowcontrol on a credit base. The best way may be to implement a simple RPC that uses a very limited sub-set of the ST protocol. At the same time this will help to regulate traffic in the host network, as OC48c is 1/4 of the speed of GSN.

Acknowledgements

The collaboration during this joint development project to develop a GSN to OC48c converter, with Genroco in Slinger, USA, has been very good. We have to thank especially Carl Pick for funding the project and Brian Breuer and Joe Nordman for the technical assistance, Alberto Guglielmi for his assistance and for synchronizing the communication with Genroco.

We would also like to thank Prof. Jordan Stamenov and Prof. Vladimir Genschev from INRNE in Sofia, Bulgaria, for their support given to Tzvetomir Anguelov to work on this exciting project.

References

[1] 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

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

[3] Potential HEP Applications of a New High Performance Networking Technology, Arie Van Praag, Ben Segal, CERN, IT/2000-007, 21 August 1999. (presented at CHEP 2000) http://hsi.web.cern.ch/HSI/gsn/reports/GSN-CHEP.PDF

[4] 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

[5] Point-to-point protocol field assignments, ppp dll protocol numbers, 30 March 2001, Information Sciences Institute, University of Southern California. http://www.isi.edu/in-notes/iana/assignments/ppp-numbers

[6] Request for Comments: 2615, Network Working Group, A. Malis, Ascend Communications, Inc. OW. Simpson, Category: Standards Track, DayDreamer, June 1999. http://www.faqs.org/rfcs/rfc2615.html

[7] Packet over SONET, Application Note, Cisco. 2 January 2001. http://www.cisco.com/warp/public/cc/pd/rt/12000/prodlit/gspos_an.htm

[8] AMCC S4801 Amazon data sheet, http://www.amcc.com/pdfs/S4801.pdf

[9] AMCC datasheet S3043, http://www.amcc.com/pdfs/S3043.pdf

[10] AMCC Datasheet S 3044, http://www.amcc.com/pdfs/S3044.pdf

Some useful WWW References

I. Genroco http://www.genroco.com

II. An Index of all RFC bulletins: http://www.faqs.org/rfcs/rfc-titles.html

III. RFC 2615: PPP over SONET http://www.faqs.org/rfcs/rfc2615.html

IV. point-to-point protocol field assignments: http://www.isi.edu/in-notes/iana/assignments/ppp-numbers

V. High Performance Networking Forum http://www.hnf.org

VI. HNF-Europe. http://hsi.web.cern.ch/HSI/HNF-Europe/Welcome.htm

VII. AMCC http://www.amcc.com/

VIII. High Speed Interconnects http://hsi.web.cern.ch/HSI/Welcome.html