High Performance Networking

Did the Pharaoh's have ETHERNET

High Performance Networking

1 High Performance Networking as sign of its time.
- A Historical Overview of Hardware and Protocols

2 Yesterdays High Performance Networks
- Ultranet, HIPPI, SCI, Fibre Channel, Myrinet, Gigabit Ethernet

TO-DAY

3 GSN (the first 10 Gbit/s network and secure)
- Physical Layer, Error Correction, ST Protocol, SCSI-ST

4 Infiniband (the imitating 2.5 – 30 Gbit/s interconnect)
- Physical Layer, Protocols, Network Management

5 SONET and some facts about DWDM,
- 10 Gigabit Ethernet, Physical Layers, Coupling to the WAN

Arie Van Praag CERN IT/ADC
1211 Geneva 23 Switzerland
E-mail a.van.praag@cern.ch

High Performance Networking

High Performance in its Time

This three standards Did not have a large influence on the development of Multi Gbit/s Networks

Year	Type	Bandwidth Mbit/s	Physical	Interf.	Protocol
1984	FDDI	100		fiber	TCP/IP, Dedicated
1988	Ultranet	100	Dedicated	copper	TCP/IP
	ATM	155 – 625	Dedicated	fiber	ATM & encaps.
1989	HIPPI	800	HIPPI-800 HIPPI-Ser.	copper fiber	Dedicated, TCP/IP, IPI3
1991 1999	Fibre Channel	255 - 510, 1020 - 2040	FC-Phys	fiber	Dedicated TCP/IP, IPI3, SCSI
1995 2000	Myrinet	1 Gbit/s 2 Gbit/s	Dedicated fiber		Dedicated, IP TCP/IP
1996	Gigabit Ethernet	1.25 Gbit/s	FC + IEEE 802.ae	copper fiber	TCP/IP

Obsolete or Commodity now to day

High Performance Networking

Ultranet

- A company private solution to allow fast switched point to point connections
- 100 MByte/s channel speed but not full crossbar
- Introduced in 1990 survived up to 1997
- Not compliant with known network standards
- In fact not a network device but a star-point connecting switch
- Compliant with TCP/IP and some other protocols
- Used in the CERN computer center from 1992 up to 1996 to connect the large main frames (Cray, IBM, Siemens/Fujitu

High Performance Networking

ATM = Asynchronous Transfer Mode

- Work on ATM started in 1984 and was released as a standard in 1988.
- ATM is a flexible communication industry standard made for data, voice and video.
- ATM is a packet switching system with small cells of 48 bytes payload + 5 bytes header.
- ATM allows to set-up virtual channels.
- Virtual channels rise problems at congestion, either data cells are lost or in secure mode result in overall high latency.
- ATM had a certain popularity as a network with speeds of 155 Mbit/s and 622 Mbit/s.
- Some people professed ATM as network and back plane replacement, connecting processor, memory and graphics parts aswell as the network. It never happened
- ATM is still used at 2,5 Gbit/s OC48/SDH16 and 10 Gbit/s OC192/SDH48 but tendency is to go at the higher throughputs to the larger frames of POS and IP.

High Performance Networking

The HIPPI History

- With a need to connect supercomputers, Storage systems, and graphics I/O, LANL started to develop a new kind of Interconnect in 1987, guided by Don Tolmie.
- Guidelines for this first High Performance Interface: 100 MByte/s, Simple and Cheap.
- All big computer manufacturers participated in the development, under guidance of ANSI T11.3 and managed by HNF, the High-performance Networking Forum.
- HIPPI-Phy: Physical Standard for point to point connections was accepted in October 1991, first equipment was demonstrated in 1990.
- The standard was extended from point to point connections with the introduction of Data Switches and a switch protocol HIPPI-SW
- Serial-HIPPI was proposed in 1991, and by CERN's influence accepted as a standard in 1995
- Protocols:

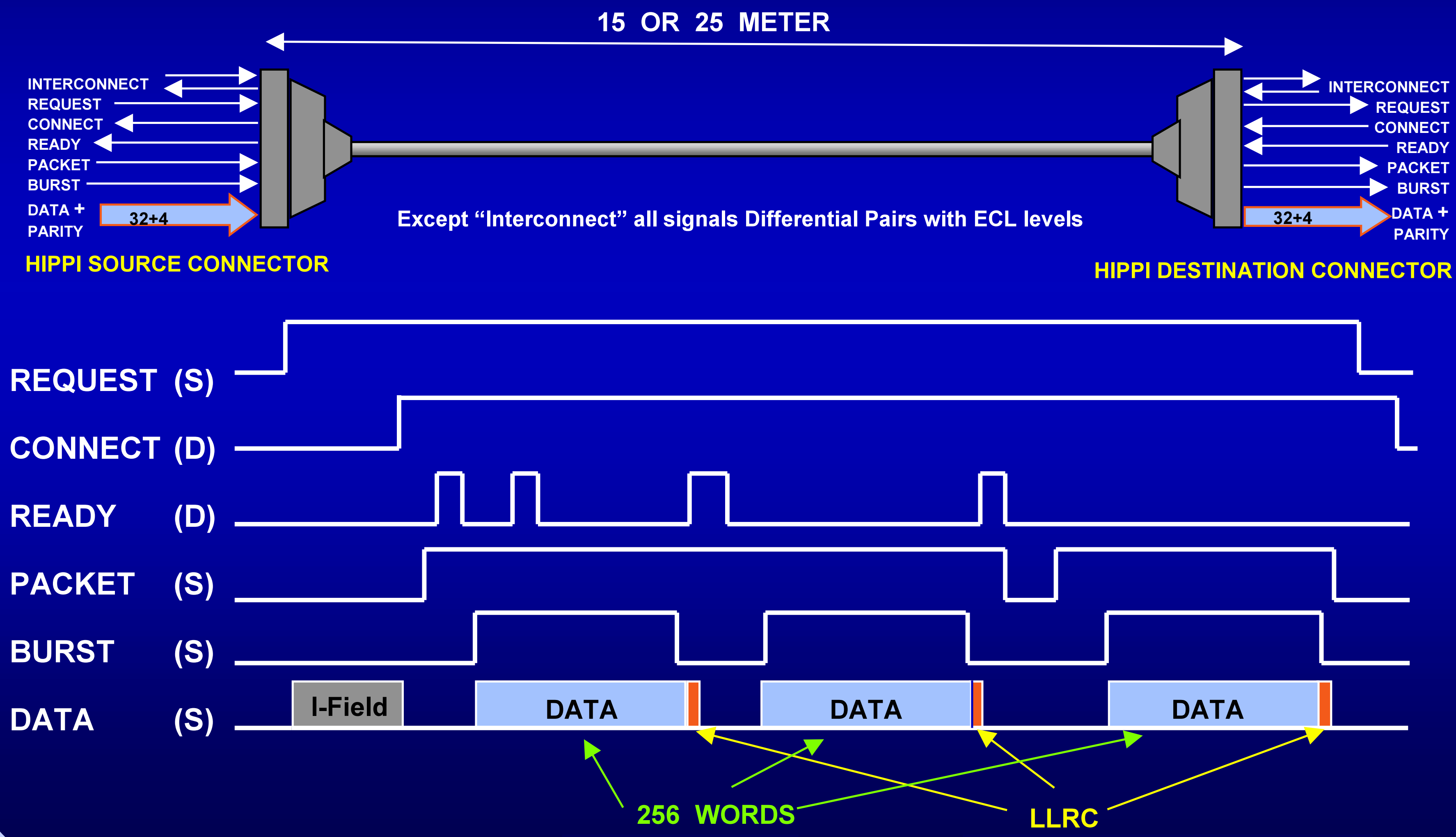
Protocols:	HIPPI framing protocol	(not much used),
	TCP/IP	(for communication)
	IPI 3	(mainly used for storage).

- In 1989 CERN started to work with HIPPI, with a great success in NA48 and in the computer center and was finally dismantled in 2002.

Let's have a quick look at the HIPPI technology

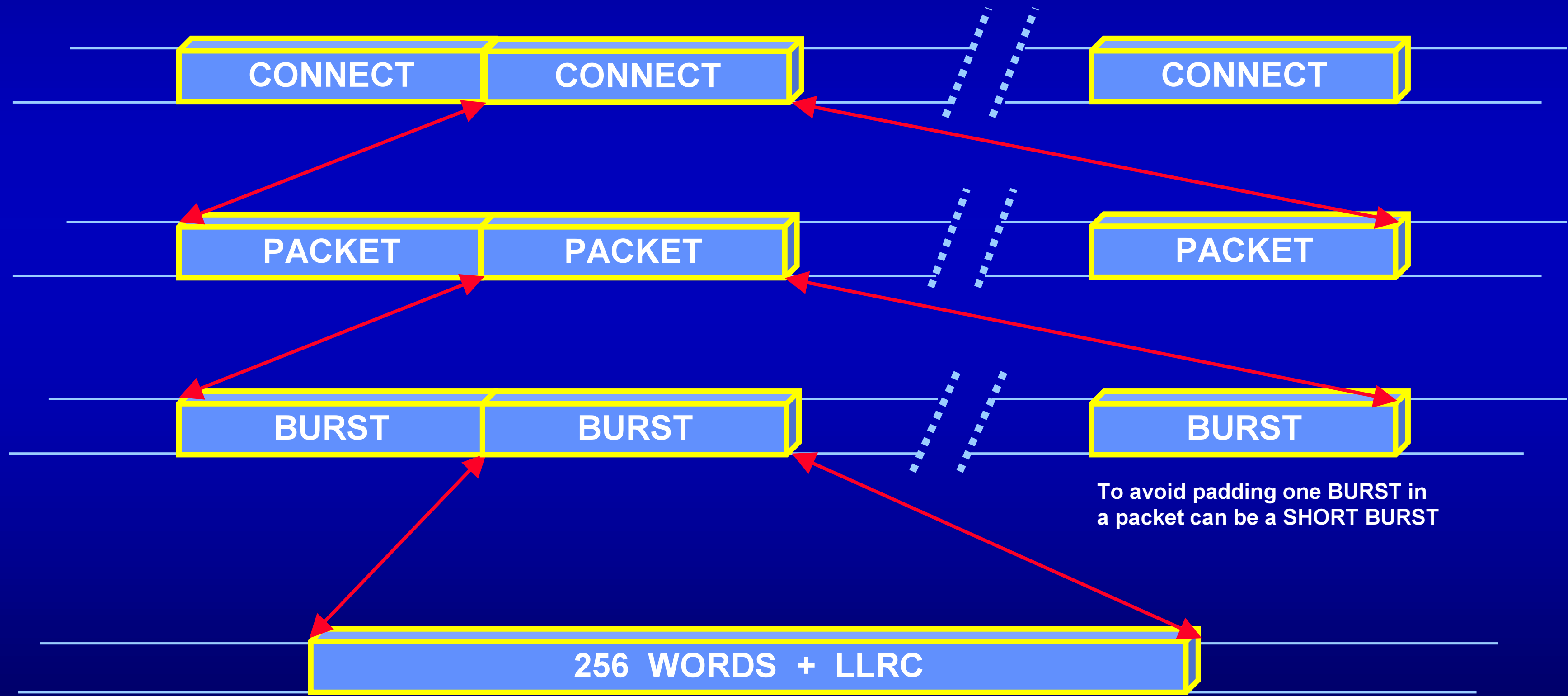
High Performance Networking

HIPPI – 800 Timing Sequence



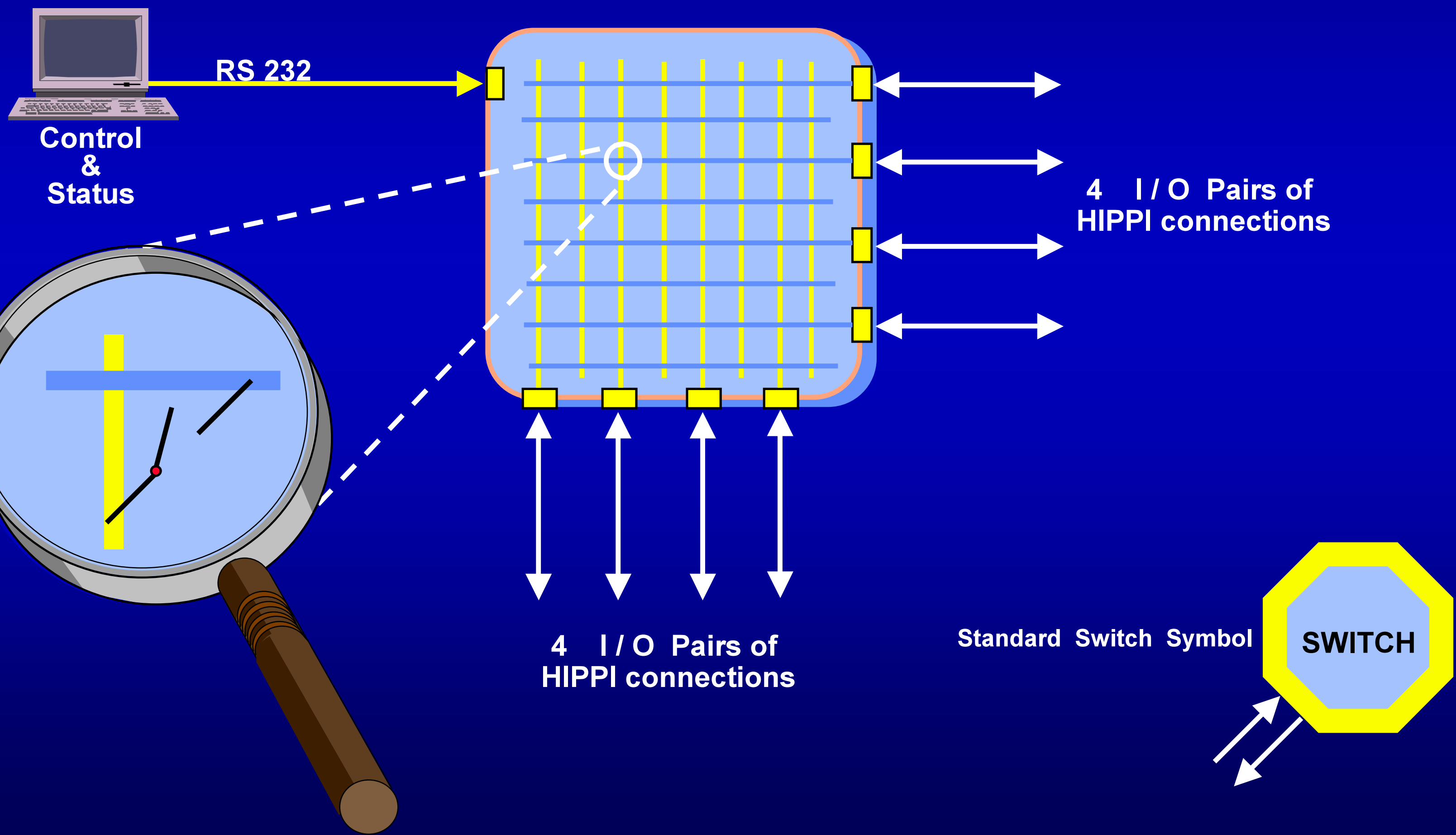
ARIE VAN PRAAG CERN IT PDP E-Mail: a.van.praag@cern.ch

HIPPI 800 Framing Hierarchy



High Performance Networking

HIPPI 800 Crossbar Switch



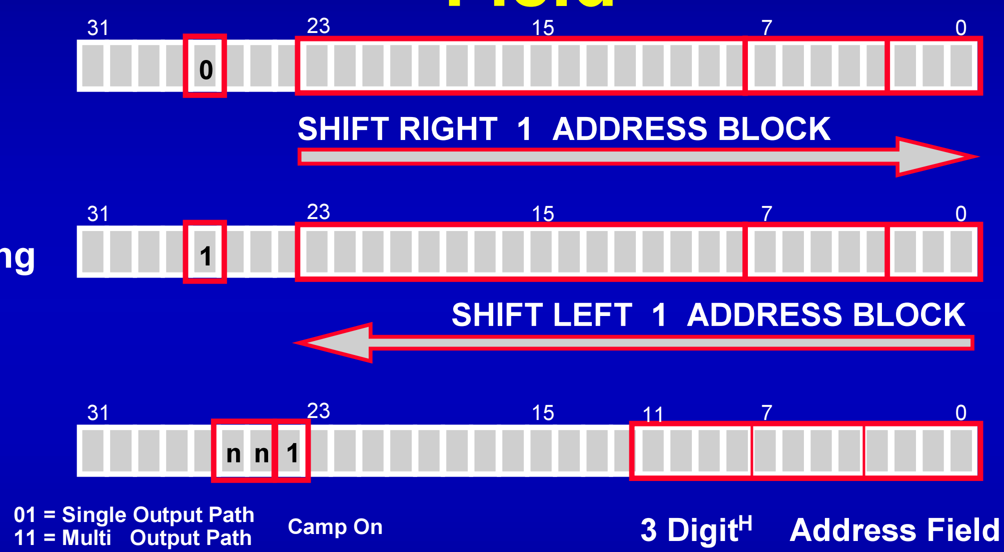
High Performance Networking

HIPPI 800 Addressing Modes using the I-Field

Source Addressing

Destination Addressing

Logical Addressing

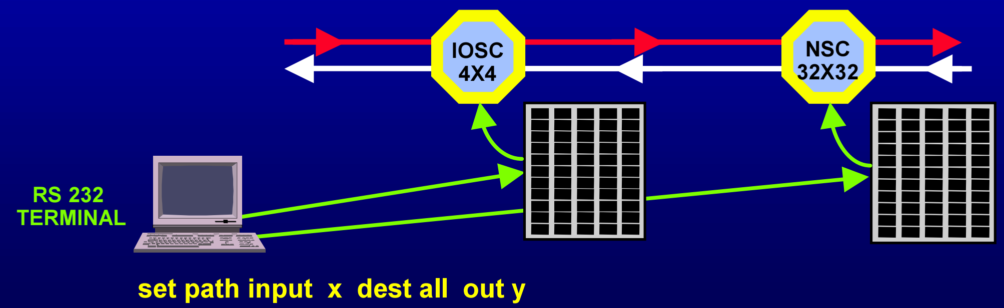


4X4 3 bit Output Field
8X8
32X32 5 Bit Output Field

4X4 3 bit Output Field
8X8
32X32 5 Bit Output Field

I-Field Logical Addresses are stored in the switch. The table is filled in by hand using an RS232 Terminal or a PC.

With some switches the table can be downloaded using RS232, Ethernet or a HIPPI channel



High Performance Networking

Serial-HIPPI

- In 1991 the Serial-HIPPI specification ware handed over to CERN.
- From 1991 to 1993 Michael S. Haben did his PhD to serialize HIPPI signals in a collaboration with Birmingham University, H.P and CERN.
- H.P made the chips and optical components that enabled Serial-HIPPI
- In 1994 work started to connect the NA 48 experiment to the computer center using Serial HIPPI. With the result that ANSI accepted in 1995 as a standard HIPPI – Serial.
- Serial HIPPI uses the same interface and the same protocols as HIPPI.
  Such transparent to the user.
- Serial HIPPI has two physical standards:

Long Wavelength	1350 nm	Lasers	10 Km.
Short Wavelength	850 nm	Diodes	200 m.

- 20b/24b up-coding is used for link security,
  incrementing bandwidth from 800 Mbit/s to 1 Gbit/s

High Performance Networking

Upcoding; what is it, why ?

½ of a 4b/5b (bad) example

0 0 0 0 0	1 0 0 0 0
0 0 0 0 1	1 0 0 0 1
0 0 0 1 0	1 0 0 1 0
0 0 0 1 1	1 0 0 1 1
0 0 1 0 0	1 0 1 0 0
0 0 1 0 1	1 0 1 0 1
0 0 1 1 0	1 0 1 1 0
0 0 1 1 1	1 0 1 1 1
0 1 0 0 0	1 1 0 0 0
0 1 0 0 1	1 1 0 0 1
0 1 0 1 0	1 1 0 1 0
0 1 0 1 1	1 1 0 1 1
0 1 1 0 0	1 1 1 0 0
0 1 1 0 1	1 1 1 0 1
0 1 1 1 0	1 1 1 1 0
0 1 1 1 1	1 1 1 1 1

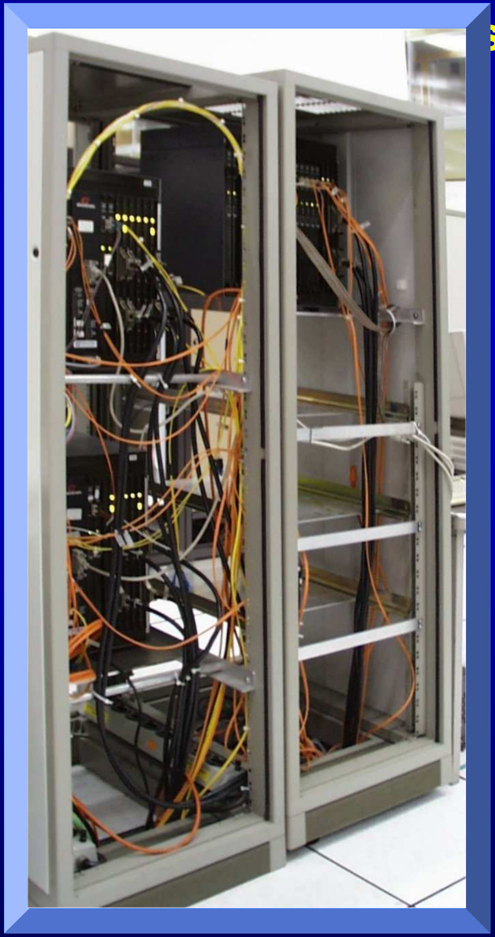
0 0 0 0	0
00 0 1	1
0 0 1 0	2
0 0 1 1	3
0 1 0 0	4
0 1 0 1	5
0 1 1 0	6
0 1 1 1	7
1 0 0 0	8
1 0 0 1	9
1 0 1 0	10
1 0 1 1	11
1 1 0 0	12
1 1 0 1	13
1 1 1 0	14
1 1 11	15

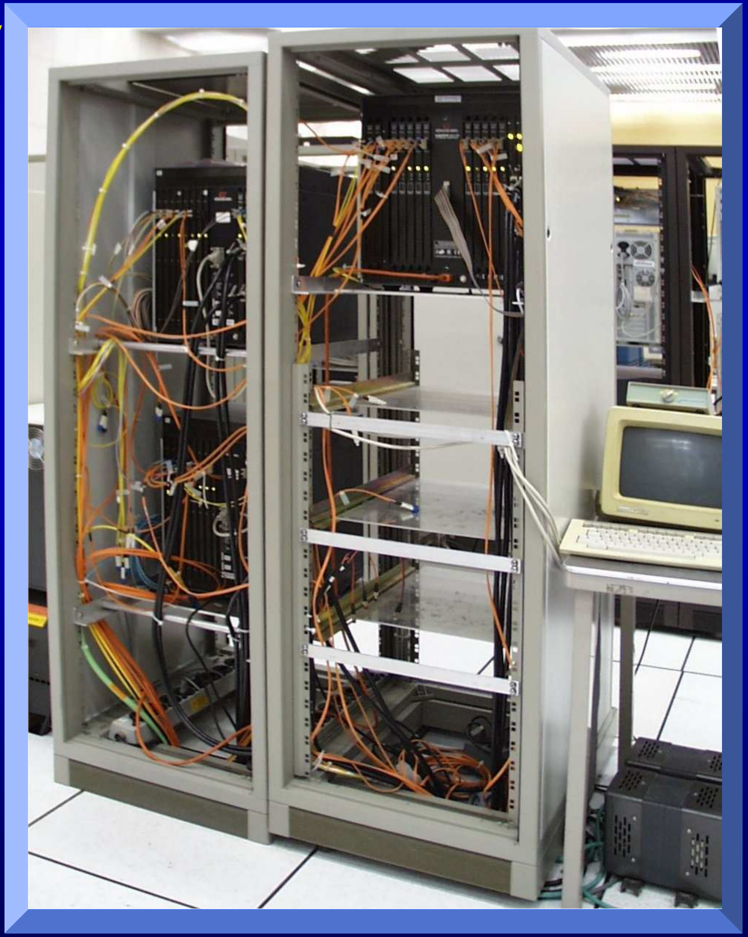
- Several ways of upcoding are possible, a number of them are standardized: 4b/5b, 10b/12b, 20b/24b, but sometimes patented.
- This coding uses a even parity placed in the middle of the four bits + parity.
(a bad example)
- They all have a different influence on the total bandwidth with an average of ± 20 %.
- Data words are spread over a wider pattern, sometimes combined with bit shuffling, such to avoid cross talk.
- Forbidden data words on non critical places in the coding are used for link synchronization.
- The goal is to make link errors as low as possible BERR 10^{-12} to 10^{-15} is normal.

High Performance Networking

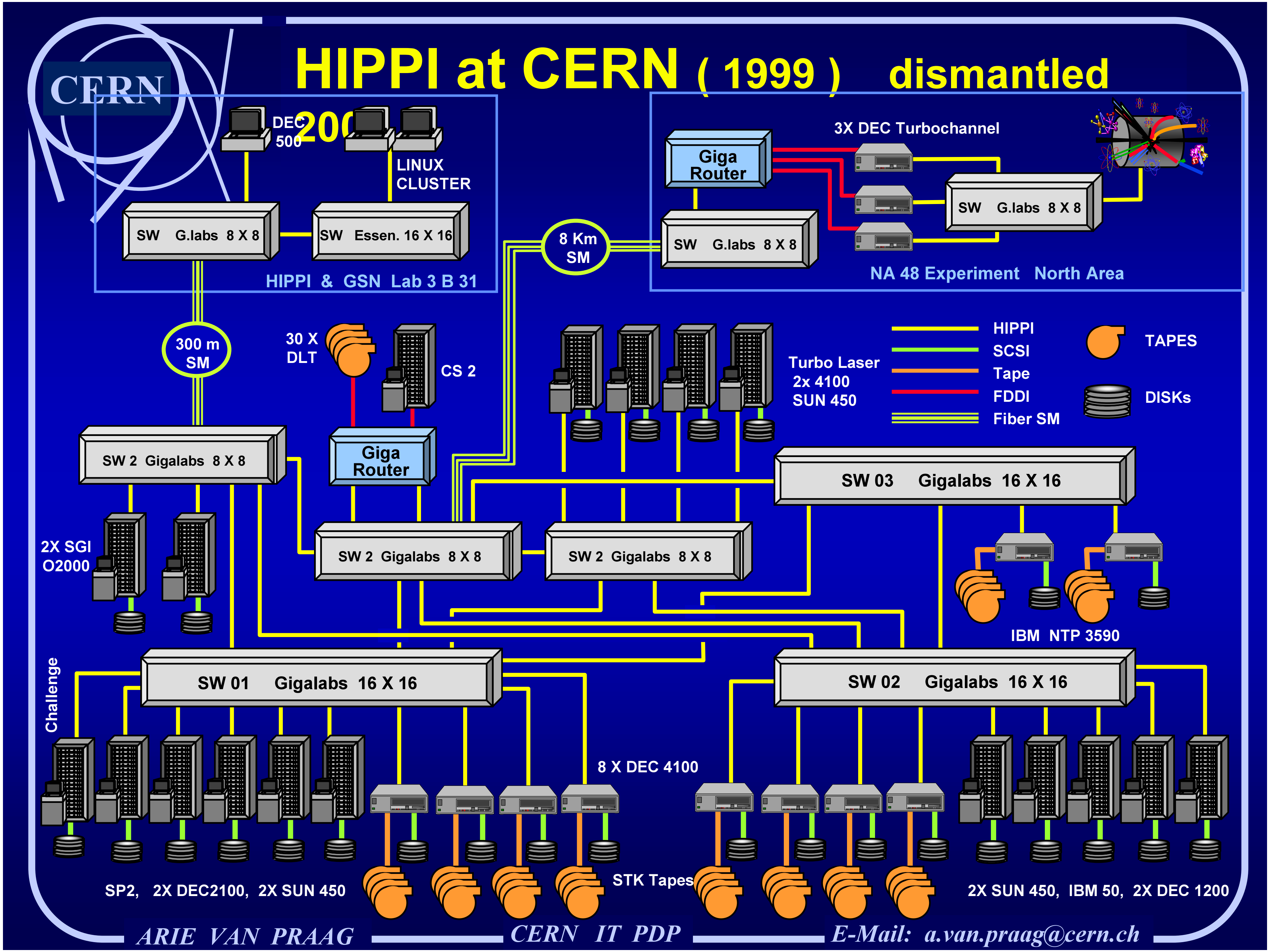
HIPPI 800 in the CERN Computer Center

shortly





 E-Mail: a.van.praag@cern.ch

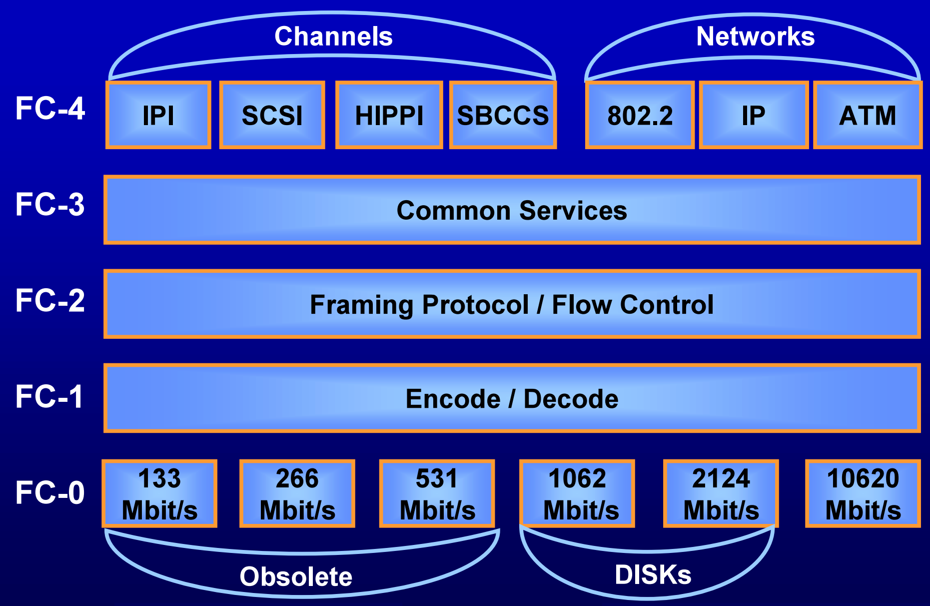


CERN
HIPPI at CERN (1999) dismantled
200
DEC 500
LINUX CLUSTER
SW G.labs 8 X 8
SW Essen. 16 X 16
HIPPI & GSN Lab 3 B 31
8 Km SM
Giga Router
3X DEC Turbochannel
SW G.labs 8 X 8
SW G.labs 8 X 8
NA 48 Experiment North Area
300 m SM
30 X DLT
CS 2
Turbo Laser 2x 4100 SUN 450
HIPPI
SCSI
Tape
FDDI
Fiber SM
TAPES
DISKs
SW 2 Gigalabs 8 X 8
Giga Router
SW 03 Gigalabs 16 X 16
2X SGI O2000
SW 2 Gigalabs 8 X 8
SW 2 Gigalabs 8 X 8
IBM NTP 3590
Challenge
SW 01 Gigalabs 16 X 16
SW 02 Gigalabs 16 X 16
8 X DEC 4100
STK Tapes
SP2, 2X DEC2100, 2X SUN 450
2X SUN 450, IBM 50, 2X DEC 1200
ARIE VAN PRAAG
CERN IT PDP
E-Mail: a.van.praag@cern.ch

High Performance Networking

Fibre-Channel

- Fibre Channel is also a standard handled at ANSI T 3.11
- A point to point connection that can perform network functionality by means of FABRIC's.
- Fibre Channel introduces "OPEN FIBER CONTROL" that shuts down the laser as the connector is open. EYE Security
- Fibre Channel introduces a Layer structure from the Physical Level on.
- FC Protocol.



High Performance Networking

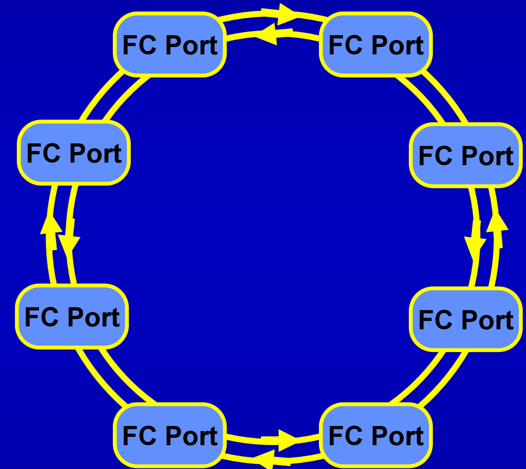
More about Fibre Channel

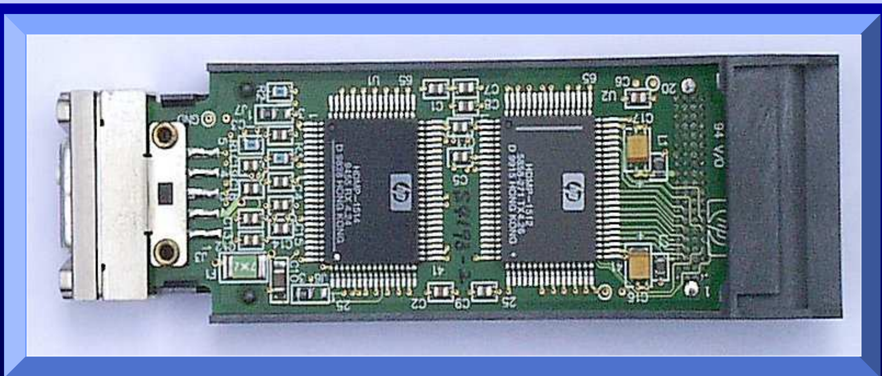
FC Protocol Structure

4 Byte	24 Byte	2112 Byte DATA FIELD		4 Byte	4 Byte
Start of Frame	Frame Header	64 Byte Optional Header	2048 Byte Payload	CRC Error Check	End of Frame

CTL	Source Address	Destination Address	Type	Seq_Cnt	Seq_ID	Exchange ID

FC Arbitrated Loop





Ready to use plug in GLM modules helped to make FC popular

- Fibre Channel made itself a problem with the large number of bandwidth versions and with the numerous protocol options.
- Such Fibre Channel is not successful as a network replacement but crystallizes out to excel in certain niches.

MOST SUCCESFULL:

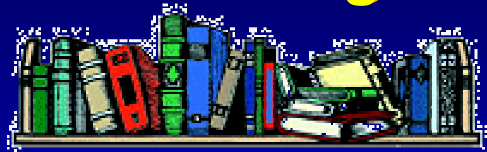
- Storage Area with FC-disks used in Networked storage:

JBOD, NAS, SAN

High Performance Networking

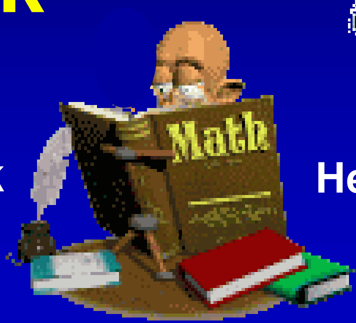
Gigabit Ethernet 1000 Base-T

- The success of HIPPI and the partial success of Fibre Channel forced the Ethernet community in 1995 to react.
- Based on 1 Gigabit payload the total bandwidth with upcoding and overhead should not be more than 1.2 Gigabit.
- The Gigabit Ethernet community struggled a long time with technological problems on the physical link level.
- Finally The Gigabit Ethernet group adapted the 1 Gigabit Fibre Channel Technology and over-clocked it by 20%.
- This over-clocking used the technology at its limits and was in the beginning source of many link errors.
- A certain company used illegally chip technology patented for Serial HIPPI, however the result ware the first reliable and successful PCI interfaces.
- Finally the standards 802.3z describing the modified Fibre-Channel Physical link layer and 802.3ab which introduced new physical link layer with set-up algorithms that proofed successful are accepted in 1999.
- Both standards forgot to solve the problem of the small 1500 byte Ethernet frames.



1000 Base-T Link Initiation

If a clever man goes to read a book



He first looks at the Index

As the link opens up, it looks what the other end contains:
AUTO-NEGOTIATION PAGES

Contents:

RSYNC	= Receive Word Synchronization
SD	= Signal Detect
LINK	= Link Detect
AN_NP	=AutoNegotiation Next page Status
AN_TX_NP	=AutoNegotiation TX Next page Status
AN_RX_NP	=AutoNegotiation RX Next page Status
AN_RX_BP	=AutoNegotiation RX Base page Status
AN_RMTRST	=AutoNegotiation Remote Restart Status

Other Gigabit Ethernet Properties:

8B / 10B Upcoding

One of the non Data Codes is used for a PAUSE_Frame with a value defining Pause time.

If PAUSE_Frame value = 0 transmission restarts.

This Flow control send by the Receiver, at the end of a frame, stops the transmitter at the end of the frame.

High Performance Networking

1000 Base-T and the frame size problem

Given: 1) A 16 MByte file to be transferred over the network.
2) Ethernet maintains the 1500 Byte Maximum Frame Size.

Technology	Frames to be handle	Move Instructions (32 bit words)	Transmitter Interrupts	Receiver Interrupts	Transfer Time without latency or interrupts
10Base-T	± 10 000	4 000 000	1 (end of file)	± 10 000	1.5 seconds
100Base-T	± 10 000	4 000 000	1 (end of file)	± 10 000	0.15 seconds
1000Base-T	± 10 000	4 000 000	1 (end of file)	± 10 000	0.015 seconds
10000Base-T	± 10 000	4 000 000	1 (end of file)	± 10 000	0.0015 seconds

If we only look at the number of Instructions/sec** to be handled (ignoring the Interrupts) with:

Ethernet	or	10Base-T needs	5	Mega	Instructions/second to fill the Pipe	A 5 MHz processor is enough	but in 1990
Fast Ethernet	or	100Base-T needs	50	Mega	Instructions/second to fill the Pipe	A 50 MHz processor is enough	but in 1995
GigE	or	1000Base-T needs	0.5	Giga	Instructions/second to fill the Pipe	A 0.5 GHz processor just now	but in 1998
10GE	or	10000Base-T needs	5	Giga	Instructions/second to fill the Pipe	No 5 GHz processor available	

Solutions ?

The first interfaces did only 30 to 40 Mbyte/s   Why ?   Look at the receiving end

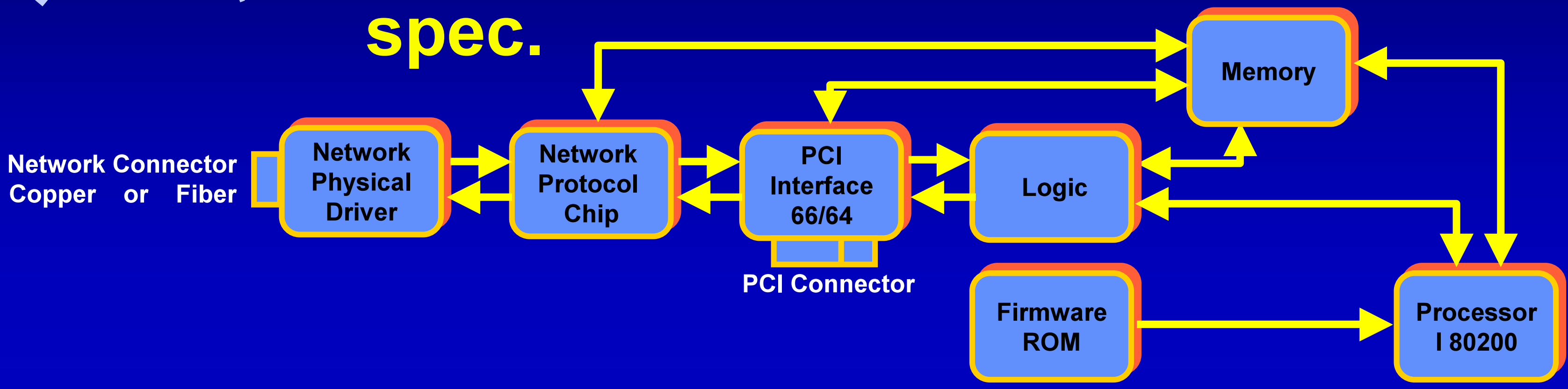
Frame:	A 1500 Byte frame at 10 bit a Byte = 15 µsec
Interrupts:	5 µsec for a 300 MHz processor = 1500 Clock Cycles.
Move and Check frames:	minimum 2 X 1500 cycles = 3000 Clock Cycles = 10 µsec
Transfer Time:	15 + 2.5 +10 = 27.5 µsec/frame
Bandwidth:	1 / 275.10^{-5} (1500) = 37 MByte/s

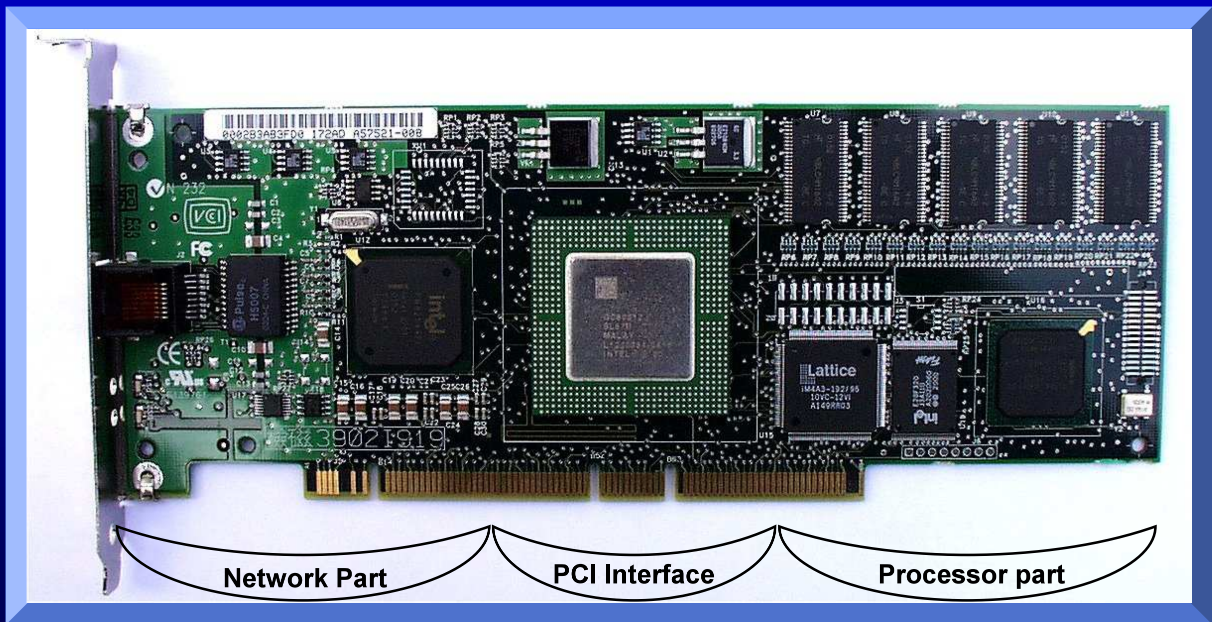
Solutions:

- Use faster processors
- Hardware CRC Generation = The interface generates the CRC code at the fly and keeps it available to be read by the processor
- Interrupt Aggregation = collecting all the frames of a file in local memory in the Interface and generate an interrupt (end_of_file) as the transfer is complete (or if local memory is almost full).
- Traffic Offload Engines (also called TCP/IP offload Engine or TOE) = handling all TCP manipulations in a local processing engine at the interface level and transfer the Data by DMA.

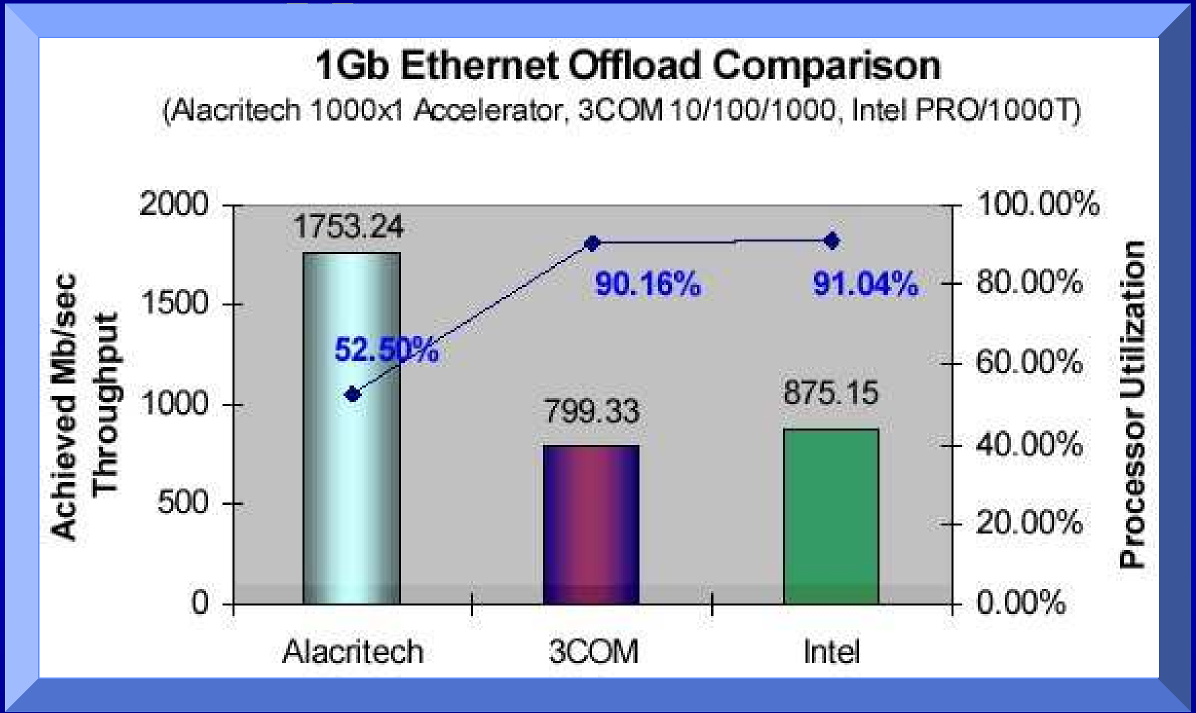
High Performance Networking

TOE if there is nothing in the spec.





High Performance Networking
TOE



This Statistics come from Mellanox: are they reliable ?? are they honest ??

Alacritech can never do 1700 Mbit/s on a 1000Mbit/s link.
It is a 2.5 Gbit network and uses a network DMA in stat of TCP/IP

Intel PRO/1000T is an early pre-production module that can do better.

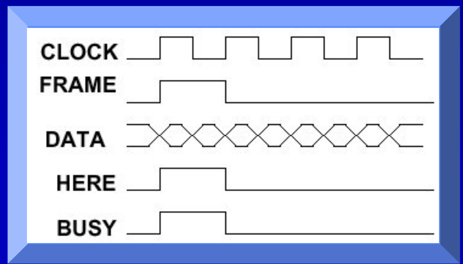
High Performance Networking

SCI = Scalable Coherent interface

- First ideas on SCI started in 1987 resulting in an accepted standard: ANSI/IEEE 1596/1992.
- SCI first of all an OPEN distributed bus, with network capabilities.
- Physical connections are based on parallel serial links with ring oriented interconnects able to connect up to 64 K nodes.
- Including switches, multiple rings can be combined in a network like structure.
- Dedicated frame format and dedicated protocol, but active IP encapsulation possible.
- Algorithms for Cache Coherency and Memory Locks are foreseen.
- Not foreseen: Low latency, Fast priority handling, But transfers use DMA.
- Link speeds: depending IC technology from 125 Mbit/s (1993) to 3 Gbit/s (2000).

High Performance Networking

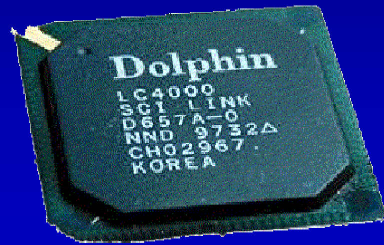
SCI = Scalable Coherent interface

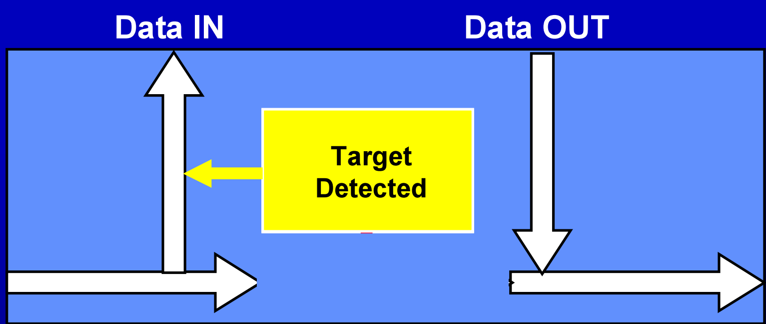


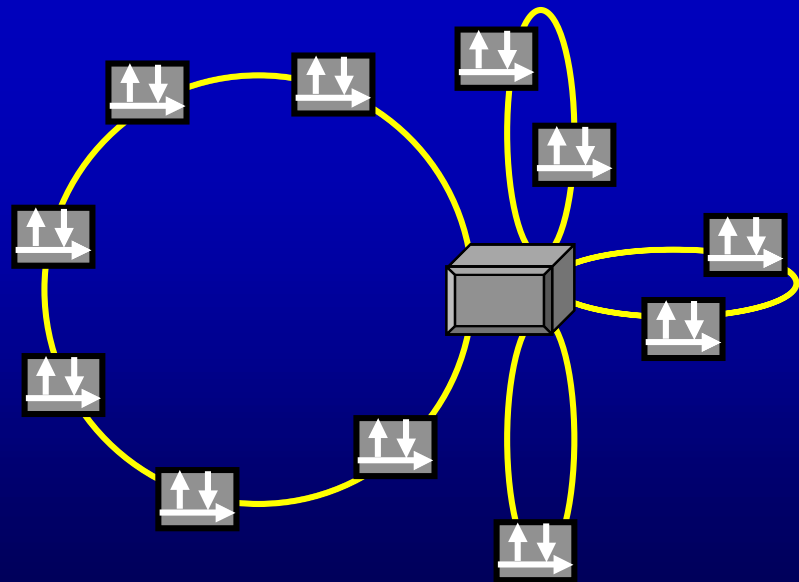
TIME

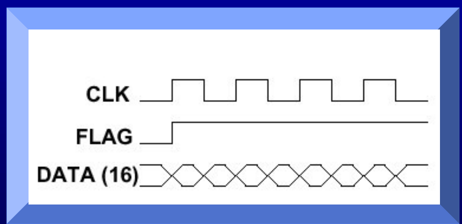
code	tgt ....
Embedded SCI Packet	

crc	post resrv parity







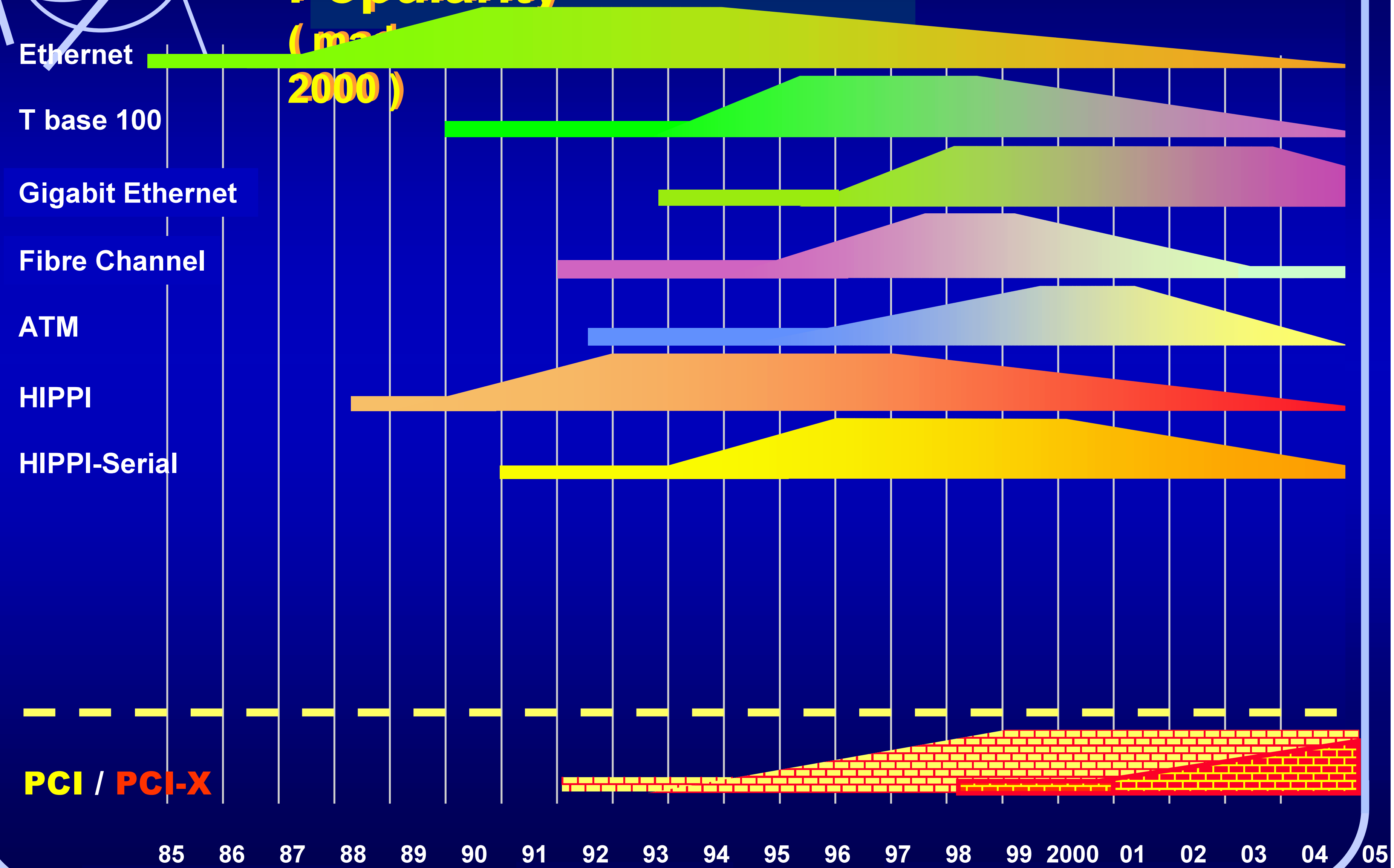


tgt	cmd	src	cnt	addr/data	crc

ARIE VAN PRAAG *CERN IT PDP* *E-Mail: a.van.praag@cern.ch*

Standards & Popularity

2000)



High Performance Networking

References:

- HIPPI PH, High Performance Parallel Interface-Mechanical, Electrical, and Signalling Specification, ANSI X3.183-1991 Rev 8.3. 2.
- HIPPI-SC, High-Performance Parallel Interface -Switch Control, ANSI X3.222-1996, Rev 3.2, ISO/IEC 11518-6, April 9, 1997.
- Applications of Optoelectronics in High Energy Physics, Ph. D. Michael S. Haben, Univ. Birmingham, November 1993.
- HIPPI 800 and 1600 Serial Specification (HIPPI-Serial Rev 2.6), Don Tolmie et al, ANSI X3.300-199x, June 11, 1996.
- Testing a Long Distance Serial-HIPPI Link for NA48, NA48 note, A. Van Praag, CERN div. ECP, 18 October 1994
- High Performance Networking Forum (HNF): http://www.hnf.org/
- HIPPI at the CERN High Speed Interface pages (HIS) http://hsi.web.cern.ch/HSI/hippi/
- For Fibre Channel information http://www.fibrechannel.org/index.html http://data.fibrechannel-europe.com/index.html
- Gigabit Ethernet, 1000Base T, Whitepaper, 1997 http://www.10gea.org/GEA1000BASET1197_rev-wp.pdf
- SEEQ 8101 Gigabit Ethernet Controller (Data Sheet), April 27, 1998
- The Scalable Coherent Interface (SCI) is an approved ISO/ANSI/IEEE Standard, 1596-1992

High Performance Networking

END 2nd Part

Coming Next

Gigabyte System Network **GSN**

Started: **1995** ANSI T3.11 as HIPPI-6400 status: available

★ 3 **GSN** (the first 10 Gbit/s network and the first secure network)

- Physical Layer, Error Correction, ST Protocol, SCSI-ST