

CERN

H igh P erformance N etworking

NA 48

DISKS

IBM SP/2

SHIFT2 SGI CHALLENGE XL.IS

SHD 04 CHALLENGE L (disk server)

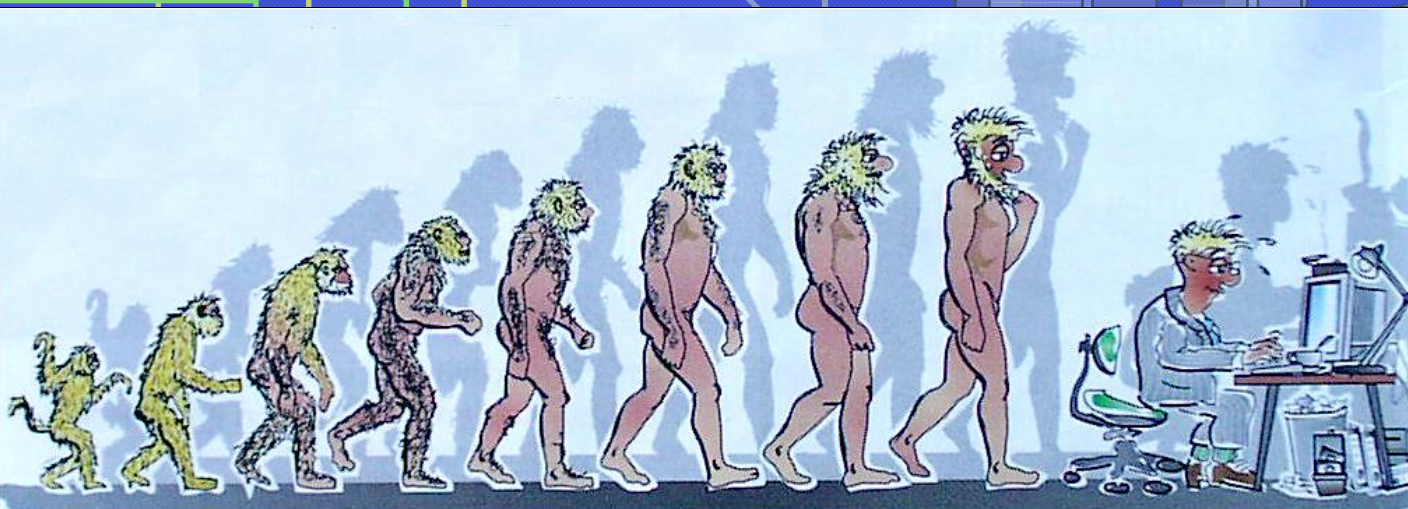
EXPERIMENT

FFDI

GIGA ROUTER

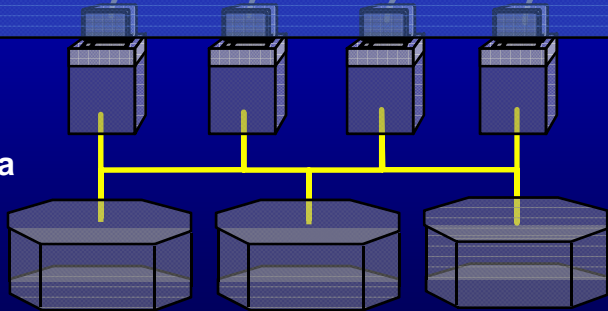
DATA up to 25 2.5 s ev

NA48 BUIL



NA 48 Physics Data from Detector

Flow NA 48 Physics Data



Stagetek Tape Robots

SHIFT7 SGI CHALLENGE XL

digital

4 FFDI Connections

14 DEC DLT Tape drives

CERN CS2 QUADRICS QS 2

GIGA Switch

H igh P erformance N etworking

★ High Performance Networking as sign of its time.

● A Historical Overview

H.P.N Now to day means 10 Gbit/s

● Infiniband IB

● 10 Gigabit Ethernet 10 GigE

● Gigabyte System network GSN

★ The Ideal Application(s)

★ More Virtual Applications for HEP and Others

★ Some thoughts about Network Storage

Arie Van Praag CERN IT/PDP

1211 Geneva 23 Switzerland

E-mail a.van.praag@cern.ch

CERN

High Performance Networking

Wireless Networks

Every New Network has been High Performance in its Time

The Very First Networks have been Wireless !!

With wavelength multiplexing

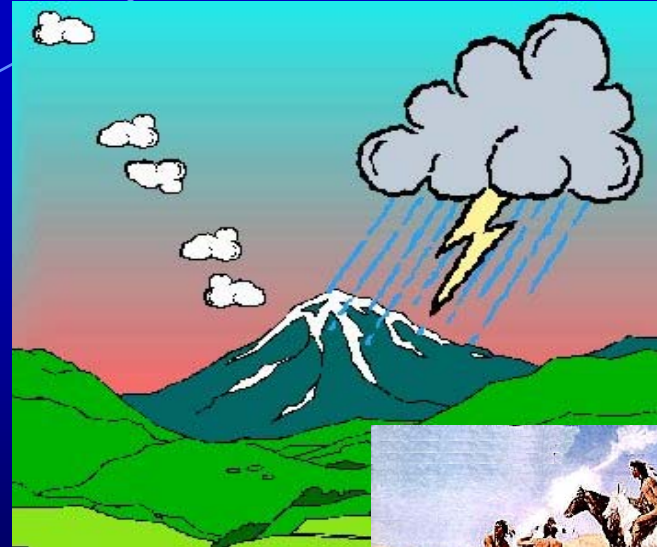


Distance : About 5 Km
Bandwidth: 0.02 Baud
Remark: Faster than a running slave

300 B Chr.



Some Clever People Invented Broadcasting
Distance: 2 - 5 Km



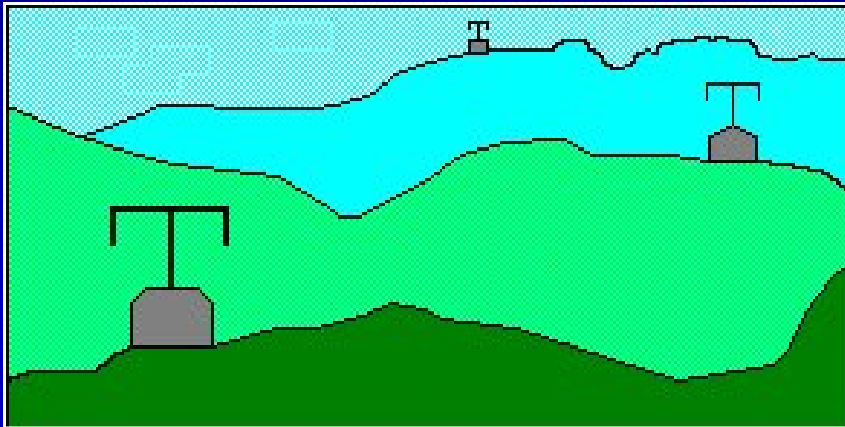
>> 1850



CERN

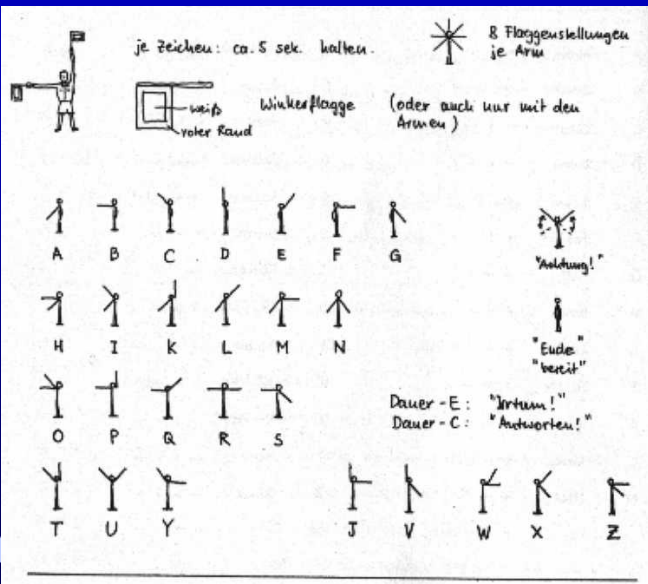
H igh P erformance N etworking Semaphores

etworking



Semaphore type of Networks came in use around **1783**.

And they were in use until the late **50^s** to Indicate Water Level or Wind. Static Message.



It was also the first time a machine language was written.

A living language that is still used by scouts.

1 Byte/s

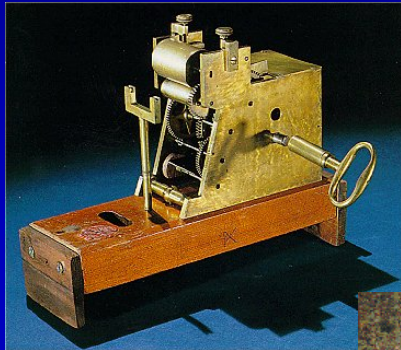
And still exists as monuments



What About: **Data Security**

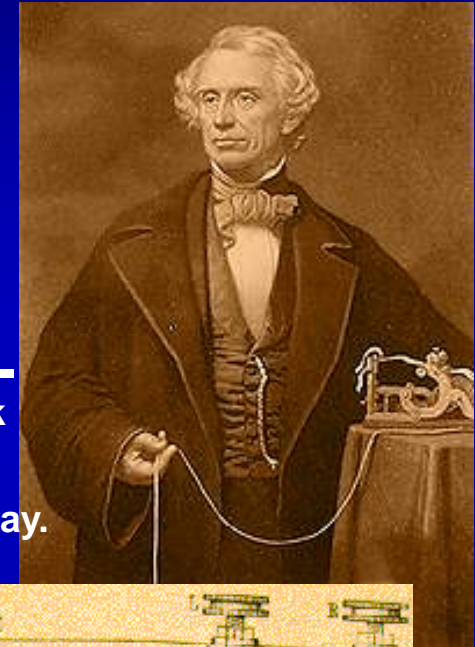
H igh P erformance N etworking

Samuel Morse

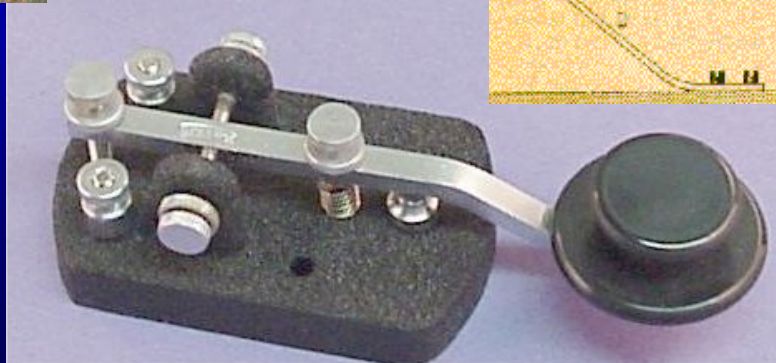
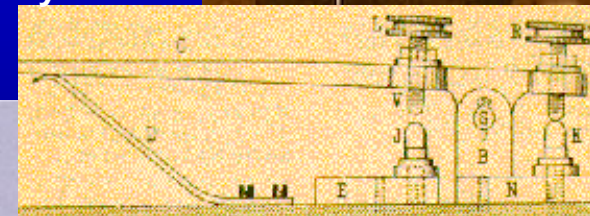


A Printer and a Sounder
1870

Pulling the cables for the first WAN



Invented the first Electric Network in **1845** and a corresponding language: **MORSE**. Still used today.
Bandwidth: ± 30 Bytes/s



CERN

H igh P erformance N etworking

The Telephone: It is a Speech handling Media, not a Data Network. Well is it ?



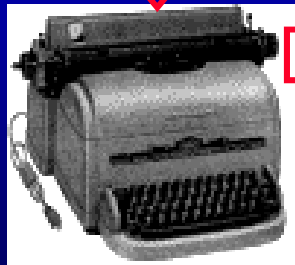
1876



1960

WWW

ASCII + RS 232



Flexowriter 10 Byte/s



Teletype 30 Byte/s



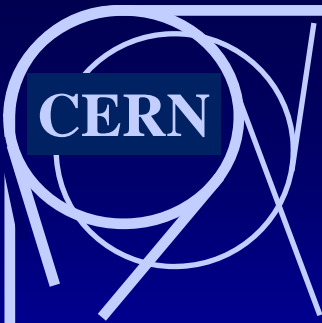
1971 at Stanford

The first Modem 120 Byte/s



The first commercial Modem 120 Byte/s

The Flexowriter interconnect made a standard character-set necessary: **ASCII**



High Performance Networking ARPANET

1966 Start of ARPANET in the USA.



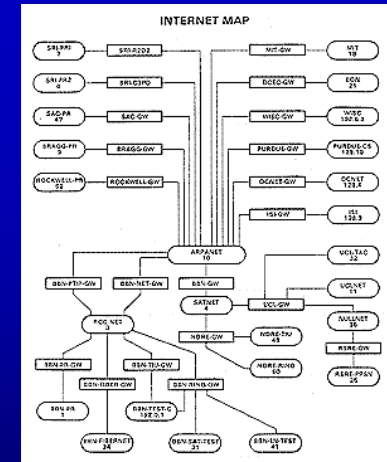
Larry Roberts
Designs and oversees ARPAnet,
which evolves into the Internet



Robert Taylor
starts ARPAnet project; organizes
computer group at Xerox PARC

ARPAnet first connection	1969		
connected	in 1971	13 machines	
connected	in 1977	60 machines	
connected	in 1980	10 000 machines	

Initial speed	2.4 Kbit/s
Incremented later to	50 Kbit/s



Protocols: NCP **IP** →

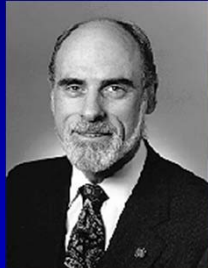
CERN

High Performance Networking

What's New in ARPAnet



Bob Kahn



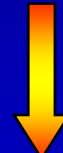
Vinton Cerf



NCP



TCP



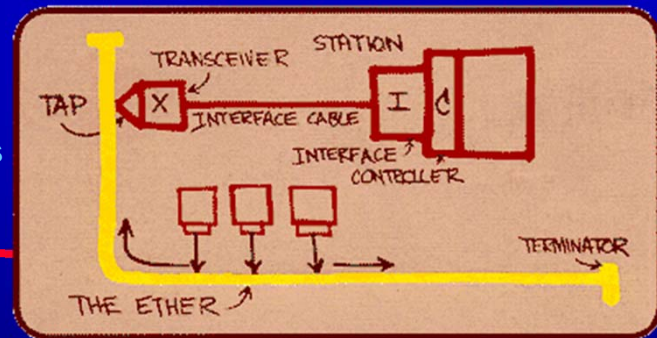
IP



TCP/IP



Bob Metcalfe's Ethernet idea



This developments leads finally to:

By Industry → Digital (DEC) & XEROX → DIX - Ethernet

By IEEE → 802.3 - Ethernet

ARPAnet → Internet



CERN

H igh P erformance N etworking

In Europe (at CERN)


- 1971** A PDP 11 in the Central Library is coupled to the CDC6600 in the Central Computer center using the terminal distributing system. 9600 Bit/s Distance 2 Km.
- 1973** Start of CERNnet with a 1 Mbit/s Link between the computer center and experiments 2 Km away.
Protocols: CERN changed progressively during **1980's** to TCP/IP
- 1985** HEPnet in Europe
Developed to connect CERN computers to a number of Physics Institutes.
- 1987** Inside CERN 100 machines
Outside CERN 6 Institutes (5 in Europe, 1 in USA)
- 1989** CERN connects to the Internet.
- 1990** CERN becomes the Largest Internet site in Europe.



CERN

H igh P erformance N etworking

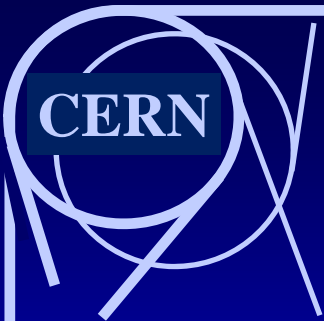
High Performance in its Time



Year	Type	Bandwidth Mbits/s	Physical	Interf.	Protocol
1974	ETHERNET	1	IEEE 802.n	copper	TCP/IP (XNS)
1976	10 Base T	10	IEEE 802.n	copper	TCP/IP (XNS)
1992	100 Base T	100	IEEE 802.n	copper	TCP/IP
1984	FDDI	100			
1989	HIPPI	800	HIPPI-800	copper	Dedicated,
1991			HIPPI-Ser.	fiber	TCP/IP, IPI3
1991	Fibre Channel	255 - 510,	FC-Phys	fiber	Dedicated
1999		1020 - 2040			TCP/IP, IPI3, SCSI
1995	Myrinet	1 Gbit/s	Dedicated		Dedicated,
2000		2 Gbit/s	fiber		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 SONET

Synchronous Optical Network

1985 SONET was born by the  standards body T1 X1 as Synchronous Fibre Optics Network for Digital communications.

1986  CCITT (now ITU ) joined the movement.

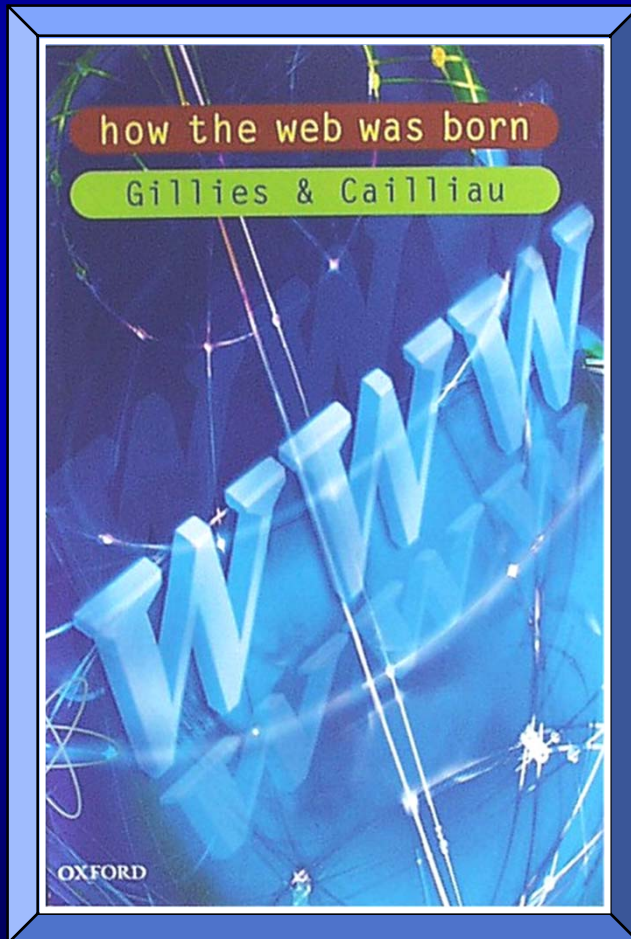
Implemented

	Optical Level	Europe ITU	Electrical Level	Line Rate (Mbps)	Payload (Mbps)	Overhead (Mbps)	H Equivalent
1989	OC - 1	---	STS - 1	51.840	50.112	1.728	---
1992	OC - 3	SDH1	STS - 3	155.520	150.336	5.184	STM- 1
1995	OC - 12	SDH4	STS - 12	622.080	601.344	20.736	STM- 4
1999	OC - 48	SDH16	STS - 48	2488.320	2405.376	82.944	STM-16
2001	OC-192	SDH48	STS-192	9953.280	9621.504	331.776	STM-64

CERN

H igh P erformance N etworking

HOW THE WEB WAS BORN



HOW THE WEB WAS BORN

James Gillies
Robert Cailliau

Oxford University Press
Great Clarendon street
Oxford OX2 6DP

ISBN0-19-286207-3

SFr. 20.- (at CERN)

CERN

H igh P erformance N etworking

About bandwidth



Bandwidth:
Load a Lorry with 10 000 Tapes
100 G Byte each.

Move it over 500 Km

Drive time is 10 Hours

Bandwidth = $10^{15} / 10 \times 3600 = 270$ GByte/s

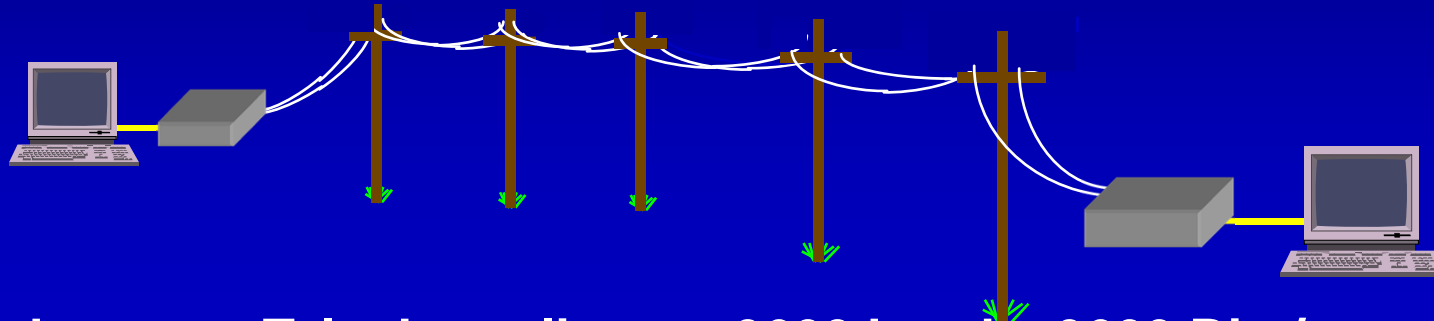
Corresponds to SONET OC 51 152

Latency
Latency

10 Hours
Distance Dependent

High Performance Networking

About Latency



Modem over Telephone lines 9600 baud = 9600 Bits/s

1 Byte = 8 bits >> 8 X 100 usec >> 800 u sec

A 1 MHz Clock Processor does 800 instruction in this time.

1 Peta Byte of data needs $1 \cdot 10^8$ sec or 3 Years to transfer

**Latency is only important as it gets large
in relation to the transfer time**

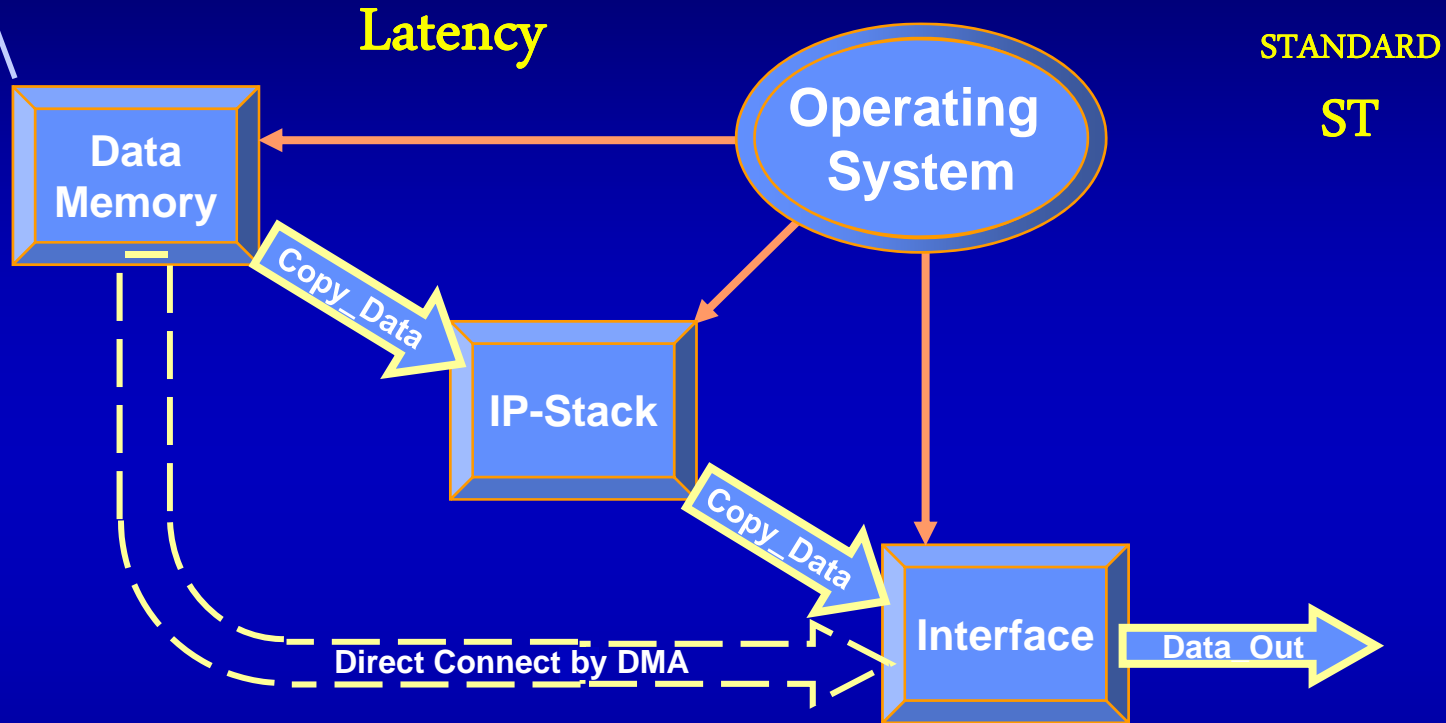
High Performance Networking

Some Statements

- ★ The higher the Bandwidth the more important gets Latency.
- ★ Flow control brings Security but also Latency.
 - Flow control is good for the LAN.
 - No flow control is better for the WAN.
- ★ Latency is always Distance Dependent. (except satellite connections)
 - Without Flow control the Pipe has to be filled.
 - With Flow Control the distance has to be done multiple times.
- ★ Small Frame Sizes kill Processor Efficiency.
- ★ High Performance Networks need Operating System Bypass.
- ★ A Network technology transparent for protocols is the better.
- ★ A Network technology transparent for Frame Size is the better.

CERN

H igh P erformance N etworking



IP Transfers are under control of the Operating System.
Most O.S. copy the Data from Memory to an IP-Stack
and copy from the IP-Stack to the Interface.

In Very High Speed Networks this translates to high losses of transfer capacity

Solution: Go direct From Memory to Interface by a DMA transfer.

QUESTION: How:

ANSWER:

Using Scheduled Transfer (ST) →

CERN

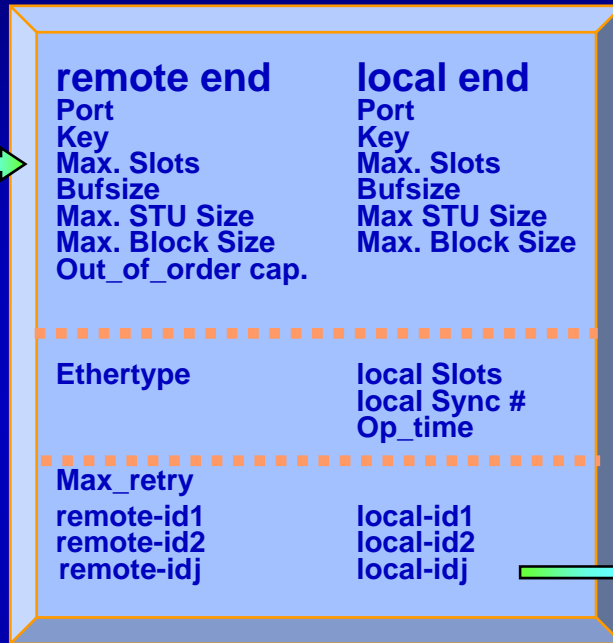
H igh P erformance N etworking

Scheduled Transfer

STANDARD
ST

local-Port
local-Key
remote-Port

Selection
and
Validation
Criteria



Virtual
Connection
Descriptors

Transfer
Descriptor

Transfer
Descriptor

Block Descriptor

Buff 0
Bufx 1
Bufx 2
...
Bufx n

Buffer Descriptor Table

Address 0
Address 1
Address 2
...
Address n

Buffers





CERN

H igh P erformance N etworking

High Performance Network Standards now Today

High Performance Networking Today means

10 Gbit/s



Gigabyte System Network

GSN

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

status: available



Infiniband

IB

Started: **1998** Industry Standard status: standard in progress
standard expected: Dec 2002



10 Gigabit Ethernet

10 GigE

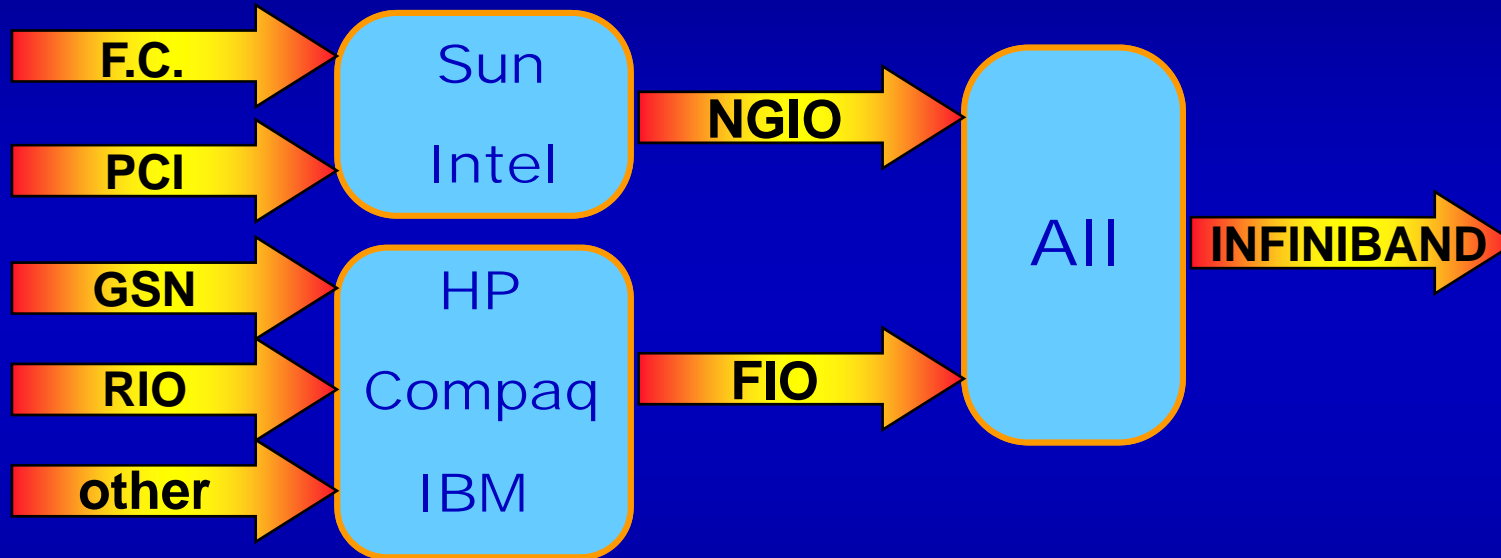
Started: **1999** IEEE 802.3z

status: standard in progress
standard expected: March 2002

CERN

High Performance Networking

INFINIBAND



INFINIBAND Specifications for :
INFINIBAND Specifications for :
INFINIBAND Specifications for :
INFINIBAND Specifications for :
INFINIBAND Specifications for :

ULP
XPORT
PHY
LINK
NET

Link Layer Protocol
Port interface
Physical Layer
Switch Protocol
Network interface



CERN

H igh P erformance N etworking

INFINIBAND

etworking

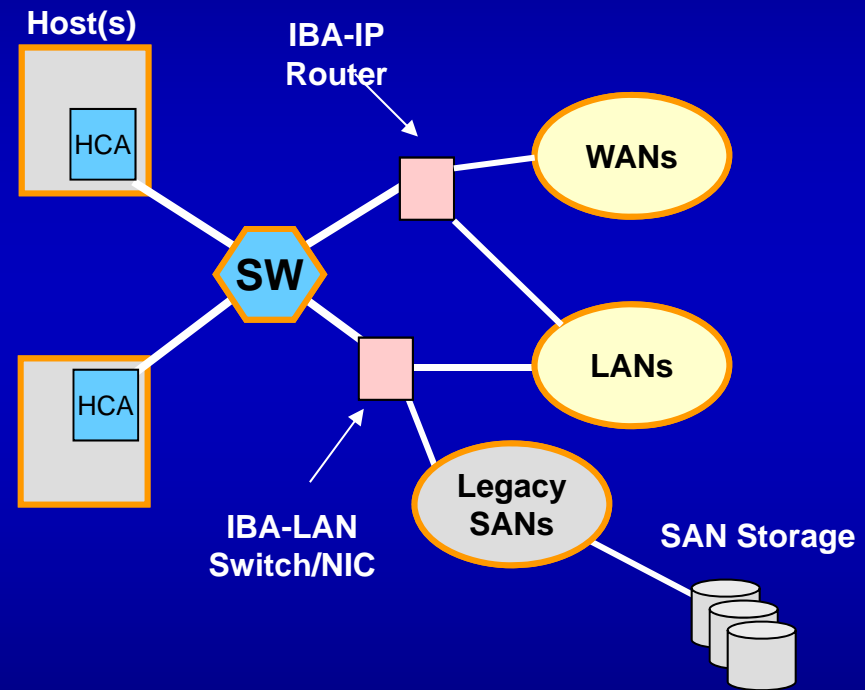
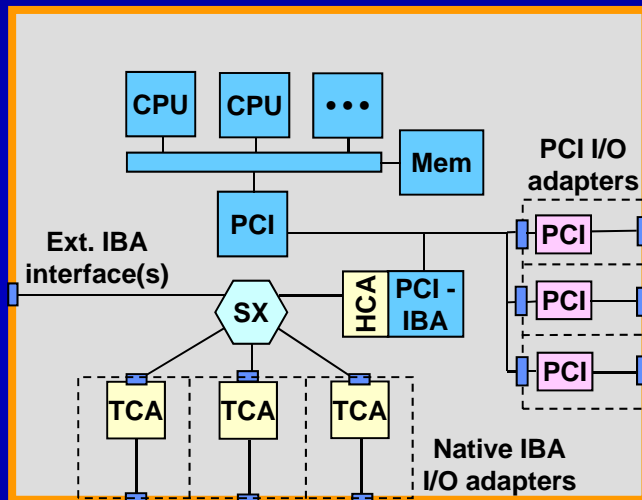


Specifications:

- ★ Bandwidth in Gbits/s Basic 2.5. Payload: ???
- Wire Bandwidth Basic, Striped 2X, 4X, 12X.
- 4 Different Standard Speeds 2.5 Gbit/s, 5 Gbit/s, 10 Gbit/s, 30 Gbit/s.
- 1 or 4 or 12 individual fibers.
- ★ Distance Covered: 25 m. 200 m
- ★ Many Transfer Protocol Options foreseen !!
- ★ Switches and Routers are specified.
- ★ Considered to replace the PCI bus and to be a Crate Interconnect
- ★ Standard Finished in: **2001 / 2002**
- ★ First Commercial hardware: **2002 / 2003**



INFINIBAND Examples



Products: No products seen by now
First proof of concept hardware expected 4 Q 2001

10 Gigabit Ethernet

- ★ Bandwidth: 12.5 Gbit/s Payload: 10 Gbit/s
- ★ Physical: Single Fiber,
 - 4 Fibers at 1/4 speed, ● 4X Coarse Wavelength Multiplexing
- ★ Distance Covered (single fiber):
 - 300 m. Multi mode ● 50 Km Single mode
- ★ Transfer: Full Duplex Fibers
- ★ Frame size: 1500 Bytes Ethernet
- ★ Protocol: TCP/IP *follows IEEE 802.3 full 48 bit addressing*
- ★ Non Blocking Switches and Routers are foreseen.
- ★ **WAN Connections: Direct transfer on OC192**
- ★ Standard IEEE 802.3ae: to be Finished in **2002**
- ★ First Commercial hardware: **2002 / 2003**

CERN

H igh P erformance 10 Gigabit Ethernet

networking



SILICON CHIPS:

EZ-Chips, Broadcom, Infineon, AMCC, PMC-Sierra, (who have a quite good white paper)

Announced



Optical interfaces:

Infineon, Agilent, Mitel

Announced



Interfaces ???

10 Gbit/s = 830 000 frames of 1500 bytes / s, or 1.5 ns / frame.
= 2 X 830 000 Interrupts/s for transmission and for reception.
Without an operating System Bypass it will be extremely difficult



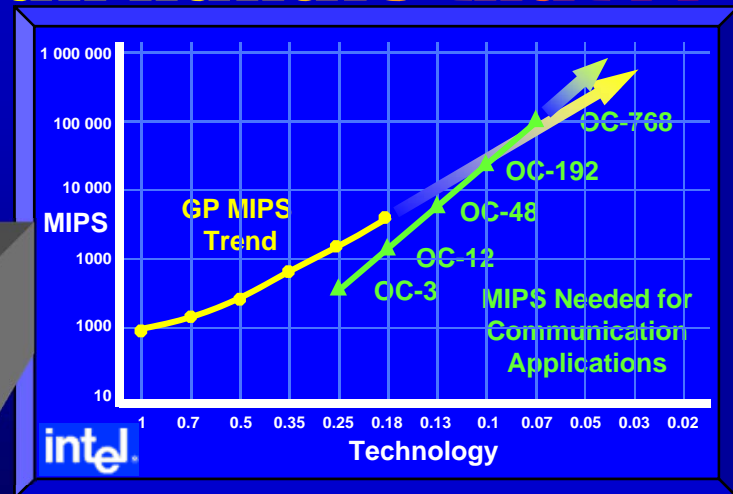
Which processor can handle that !!



Switches and Routers

???

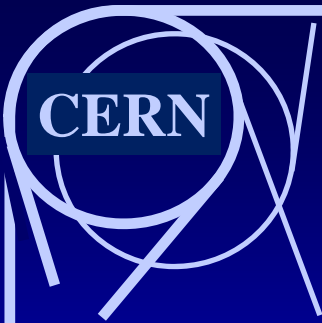
- The first products will be bandwidth concentrators.
- A Kind of proof of concept model is delivered by Cisco to LANL



10 X Gigabit Ethernet



10 Gigabit Ethernet
or / and
OC192 PPP-POS

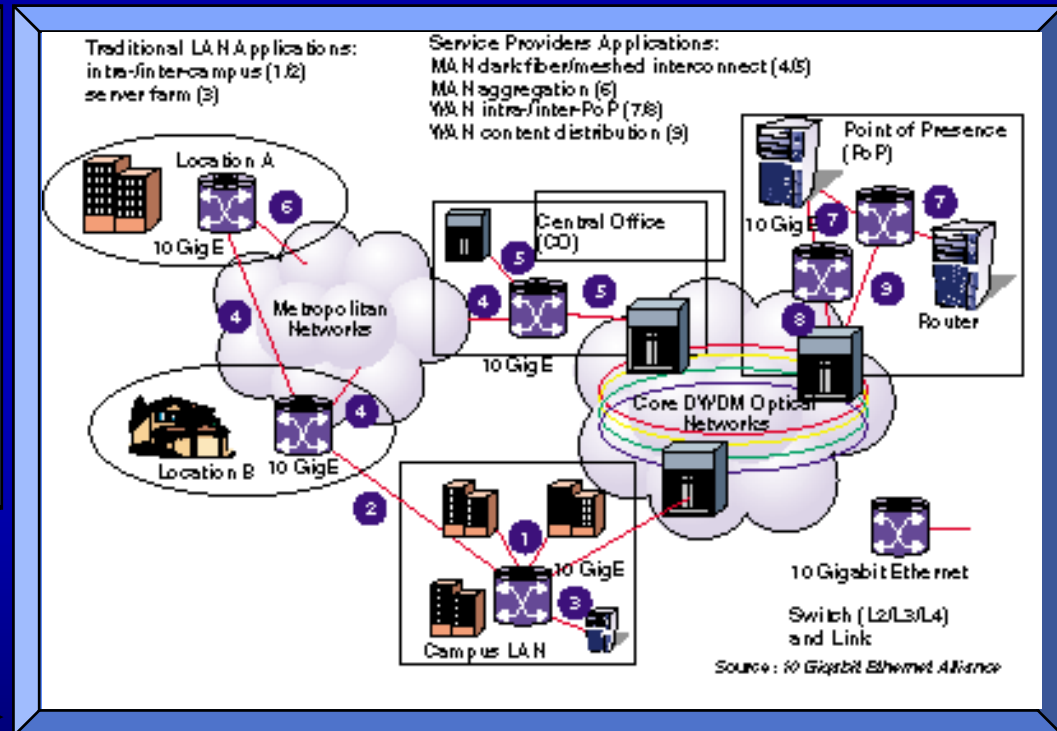
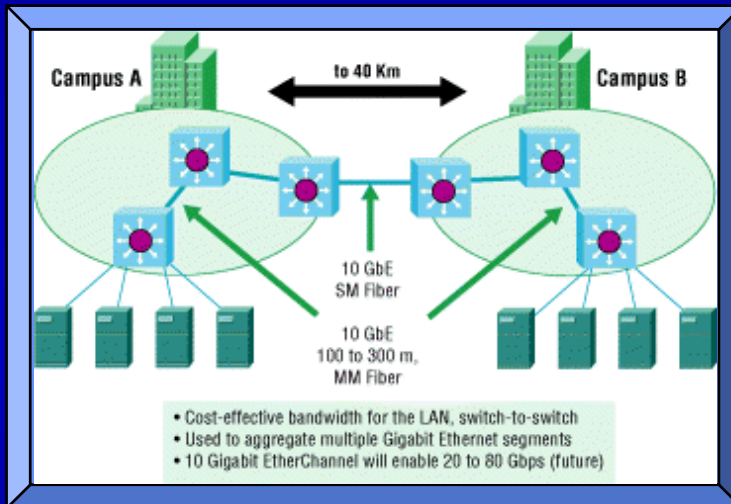


H igh P erformance

etworking



10 GigE Examples



Examples of Future Applications by Cisco and the 10 Gigabit Ethernet Alliance

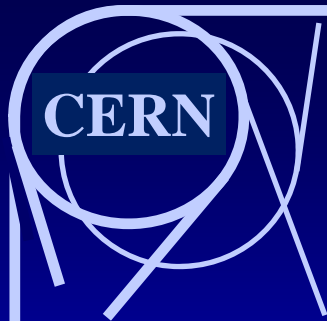


CERN

High Performance Networking

GSN (Gigabyte System Network)

- ★ Bandwidth: 10 Gbit/sec Payload: 800 MByte/s
- ★ Physical: Parallel Copper, Distance 50 m.
● Parallel Fiber, Distance 75 > 200 m.
- ★ Transfer: Full Duplex
- ★ Frame size: Micropackets → Transfer independent of file size
- ★ Protocol: ST, TCP/IP, FC, SONET and SST (SCSI over ST)
- ★ Low latency due to Operating System Bypass
- ★ Non Blocking Switches and Routers available.
- ★ WAN Connections: Bridge Connection to OC48
- ★ First Commercial hardware: **1998**
- ★ Standards →



H igh P erformance N etworking

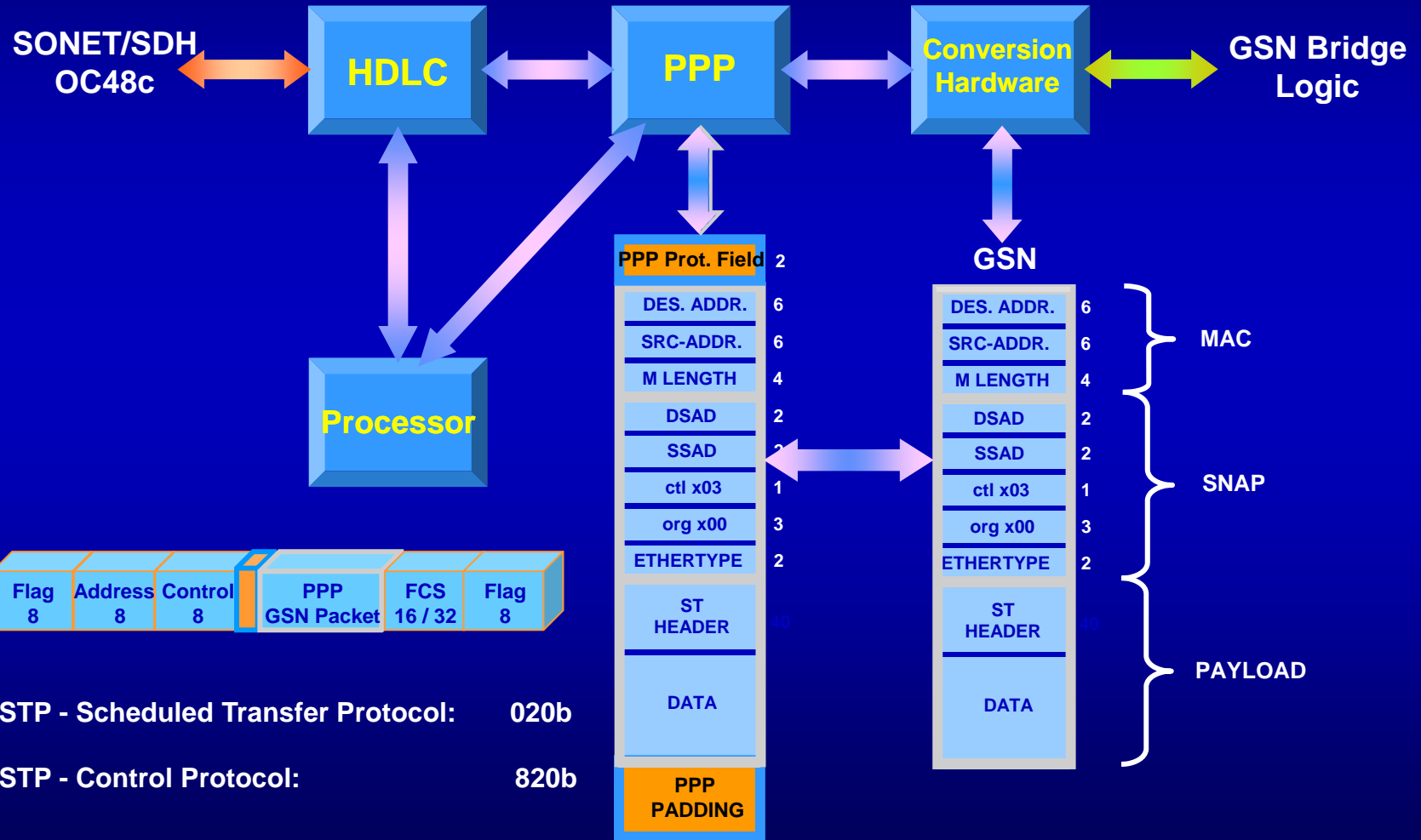


GSN Standards: Project name HIPPI-6400

Document:	Description:	Status:
★ HIPPI-6400 PH	Physical Layer 6400 Mbit/s <i>or 800 MByte/s network</i>	ANSI T11 NCITS 323-1998 ISO ISO/IEC 11518-10
★ HIPPI-6400 SC ●	Switch Standard <i>follows IEEE 802.3 full 48 bit addressing</i>	ANSI T11 NCITS 324-1999
★ HIPPI-6400 OP	Optical Connection	ANSI T11 NCITS Submitted
★ ST	Scheduled Transfer	ANSI T11 NCITS submitted
★ SCSI over ST	SCSI commands over ST	ANSI T11 NCITS Standard ANSI T10 SCSI T10 R-00
★ Sub-standards: ● ● ●	GSN & ST conversions to: Fibre-Channel, ● Gigabit Ethernet, ● ATM.	HIPPI, SONET,

High Performance Networking

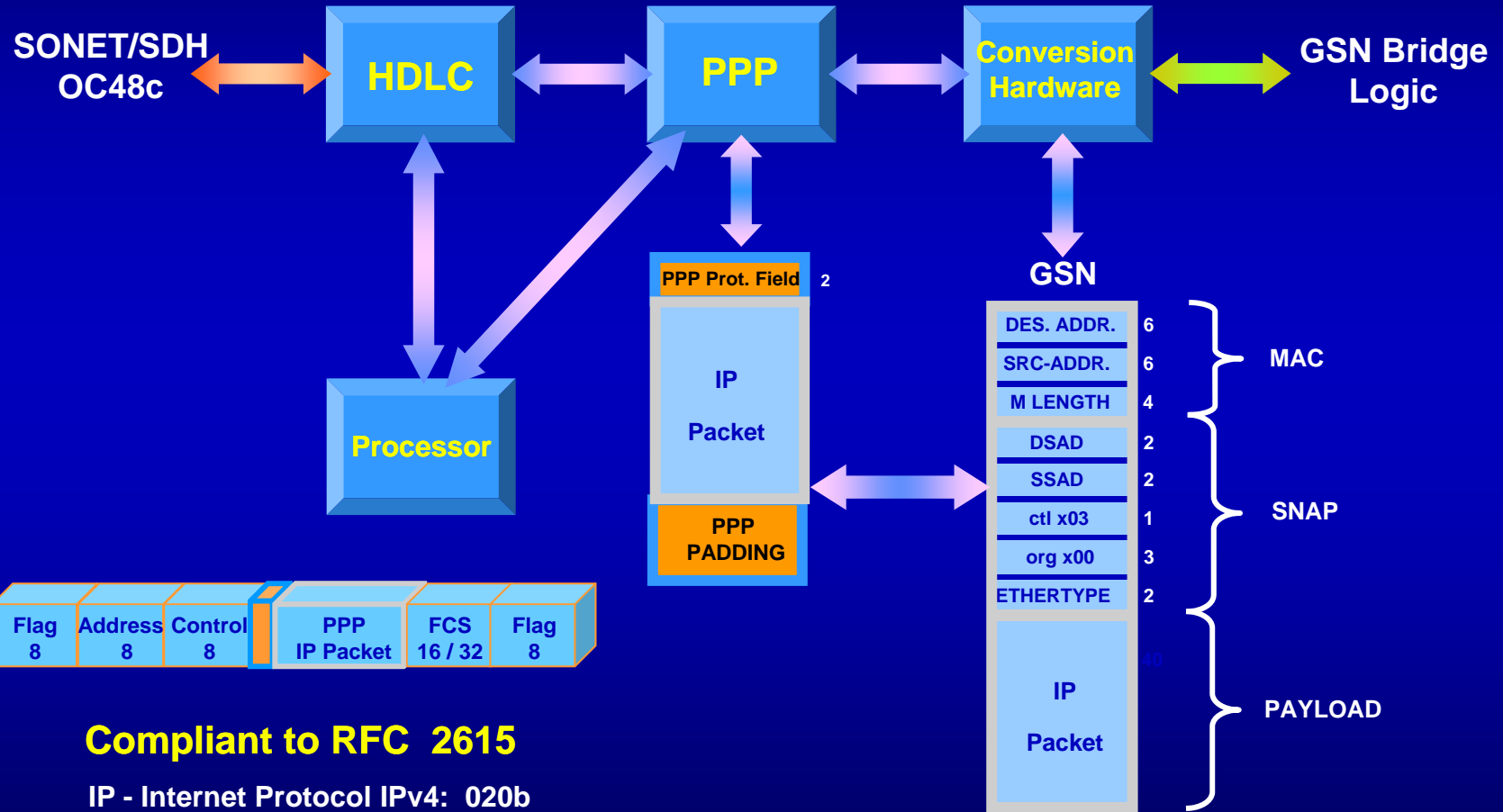
OC48c GSN ST Header Conversion



CERN

High Performance Networking

OC48c GSN IP Header Conversion





CERN

High Performance Networking

GSN Products as of January 2000



sgi™

SILICON CHIPS

Silicon Graphics

Available



GENROCO

INTERFACES:

Silicon Graphics

Origin Series

Available



ESSENTIAL
AN ODS NETWORKS COMPANY

PCI Interface 64/66

Genroco

1 Q 2000

PCI/X Interface

Essential

3 Q 2001



FCI
FRAMATOME GROUP

CABLES:

FCI - Berg Copper cables and Connectors

Available



Infineon
technologies

COMPONENTS for OPTICAL CONNECTIONS:

Infineon

Paroli DC Modules and Fibres

Available

MOLEX

Paroli DC Modules and Fibres

2 Q 2001

Gore

Optical Modules and Fibres

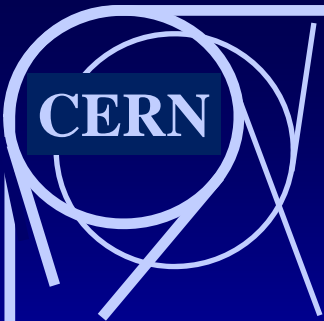
1 Q 2001



GORE
Creative Technologies
Worldwide

GSN Native Optical Connections

2 Q 2000



High Performance Networking

GSN Products as of January 2000

Networking



SWITCHES:



ODS - Essential 32 X 32

Available



ODS - Essential 8 X 8

Available



Genroco 8 X 8

Available

PMR 8 X 8

Available



BRIDGES:



ODS-Essential Translation Function HIPPI-800

Available



Genroco Storage Bridge Fibre Channel

Available

Genroco Network Bridge HIPPI

Available

Fibre Channel

Available

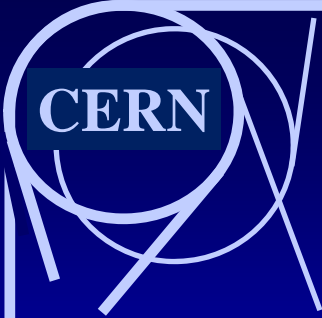
Gigabit Ethernet

Available

OC48c

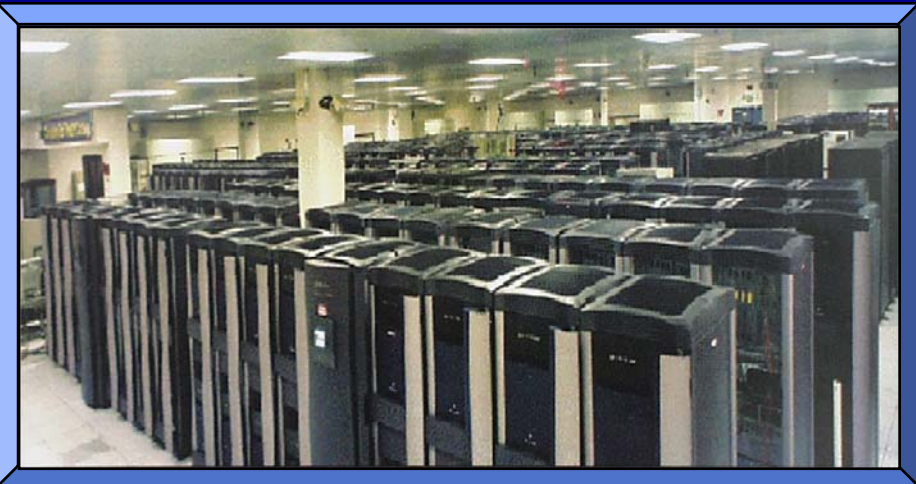
Available



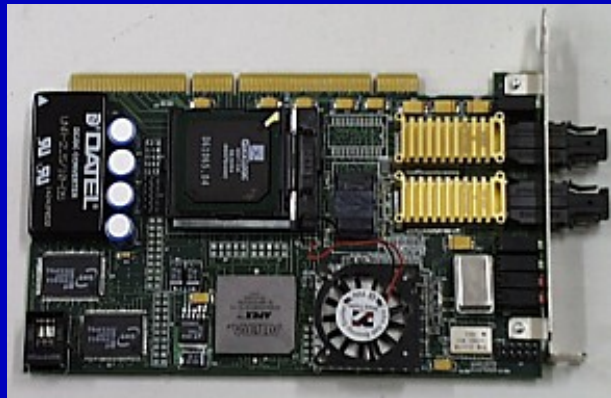


High Performance Networking

GSN Applied



Los Alamos National Laboratory: Blue Mountain Project



PCI > GSN Interface



Switches and a bridge

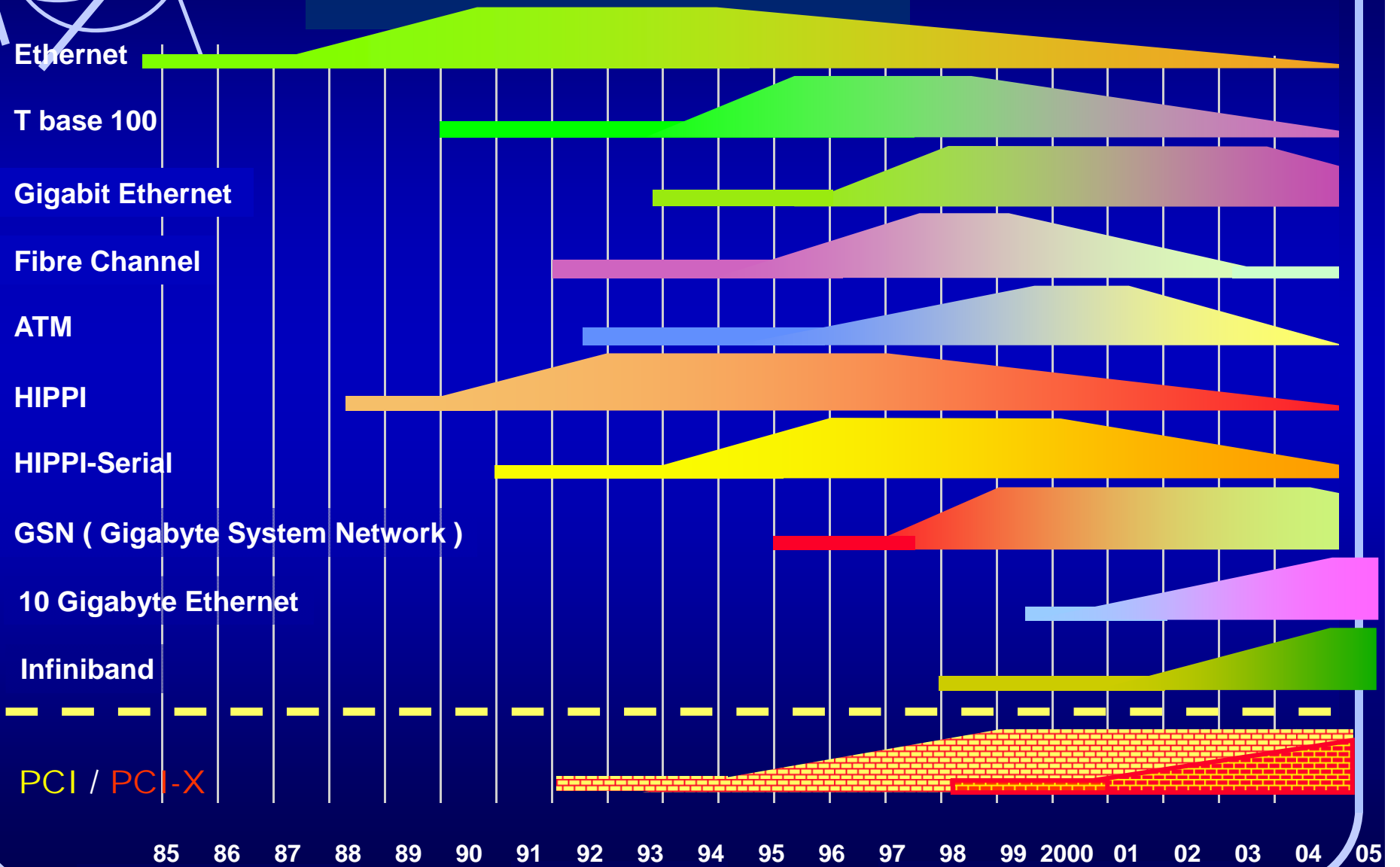


In total there are about 20 active applications worldwide

CERN

Standards & Popularity

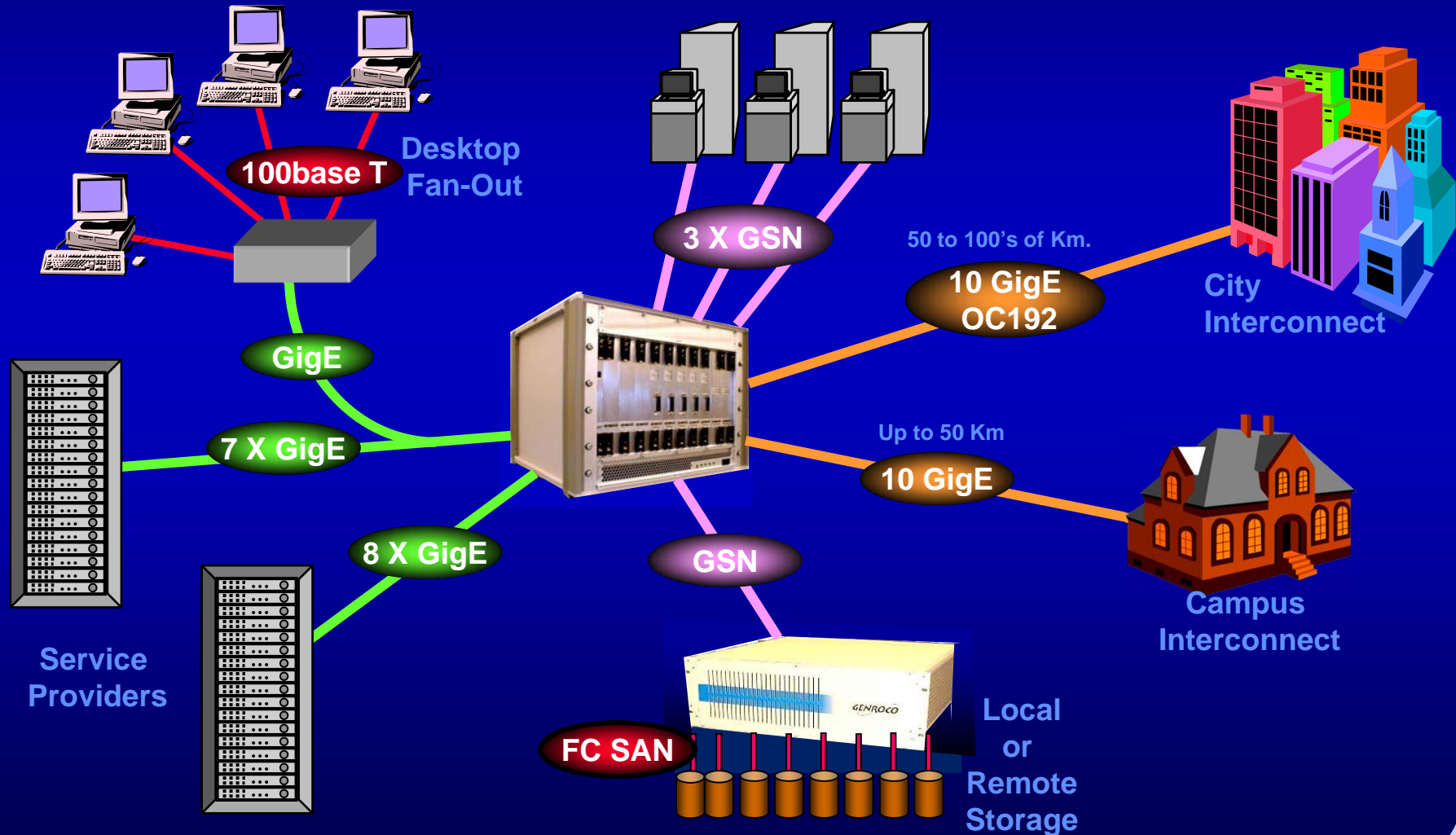
(made in 1995 and extended 2000)



CERN

H igh P erformance N etworking

The Ideal Network with all this Components



Event Building with a Switch



1 0-100 TByte/s.

DETECTOR DATA

100- 1000 Bytes/s.

VMEbus Read Out Buffers (ROB)

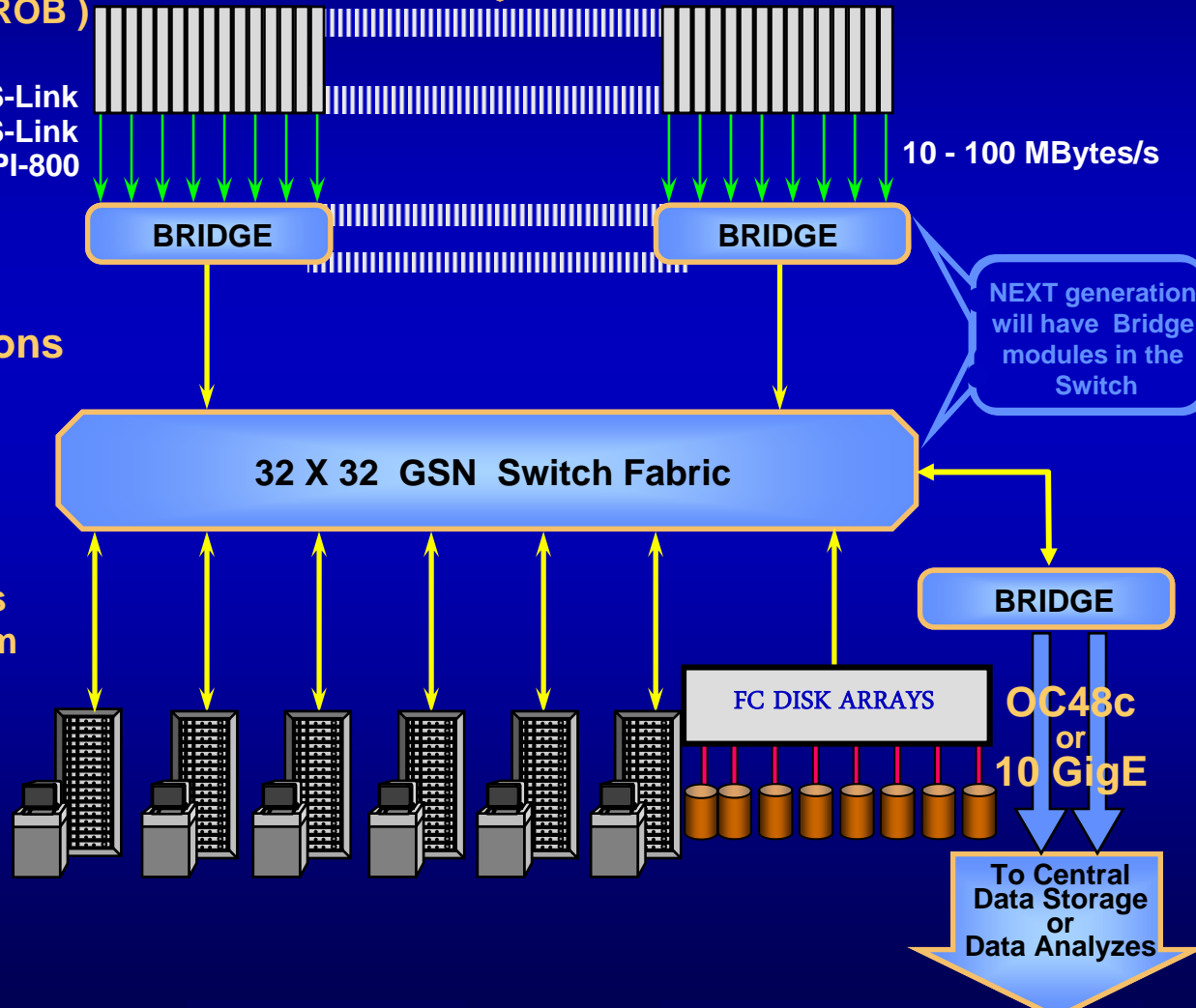
CONNECTIONS

768 (4) S-Link
or 1152 (6) S-Link
or 192 HIPPI-800

24 GSN Bridges

24 GSN Connections

8 GSN Connections to Workstation Farm



10 - 100 MBytes/s

BRIDGE

BRIDGE

32 X 32 GSN Switch Fabric

NEXT generation will have Bridge modules in the Switch

BRIDGE

FC DISK ARRAYS

OC48c or 10 GigE

To Central Data Storage or Data Analyzes

CERN

High Performance Networking

LHC Experiments:

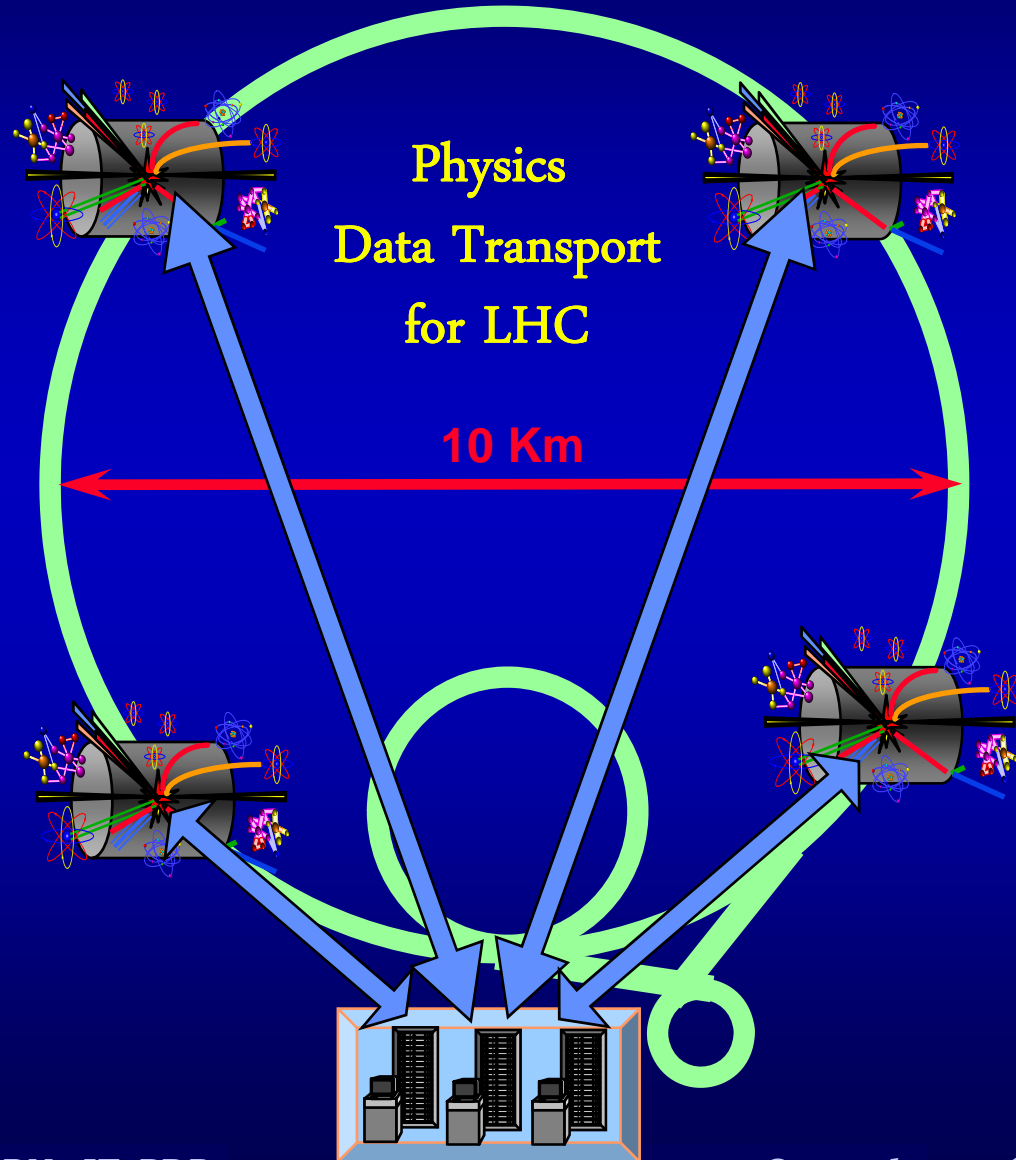
Each experiment Transmits
at least 100 - 250 MBytes/s

How to get this data
to the computer center ?

OC 48c does 310 MByte/s
and
IP over OC192 pos 1GByte/s

Atlas Alice

LHCB CMS



CERN

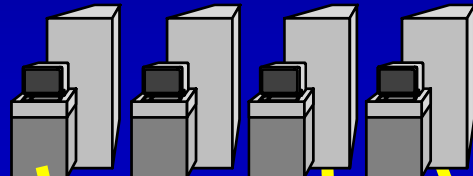
H igh P erformance

IP Video on Demand

Networking

MPEG2 - DVB ASI
Coaxial Copper cable

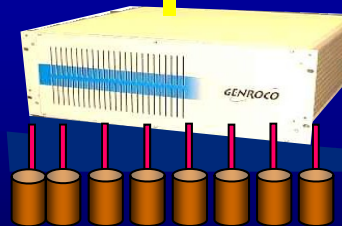
SERVERS



GSN
Connections

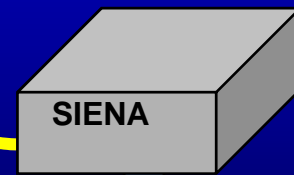


Storage Bridge



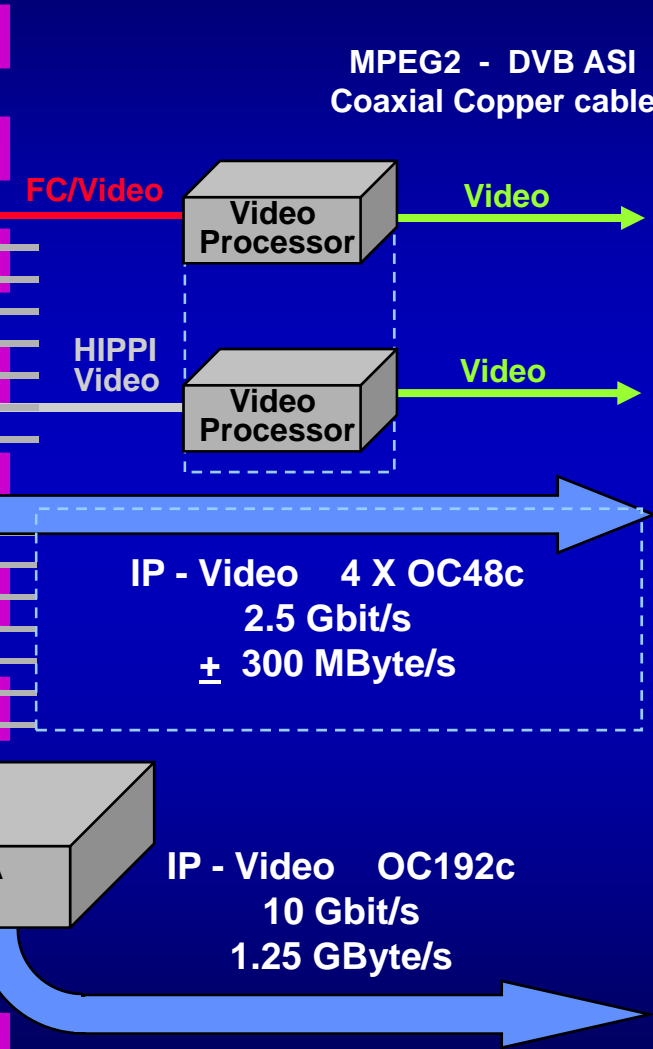
Large Storage array
on Fiber Channel Arbitrated Loop Base

8 X 256 Disks = 25 Terra byte



IP - Video 4 X OC48c
2.5 Gbit/s
± 300 MByte/s

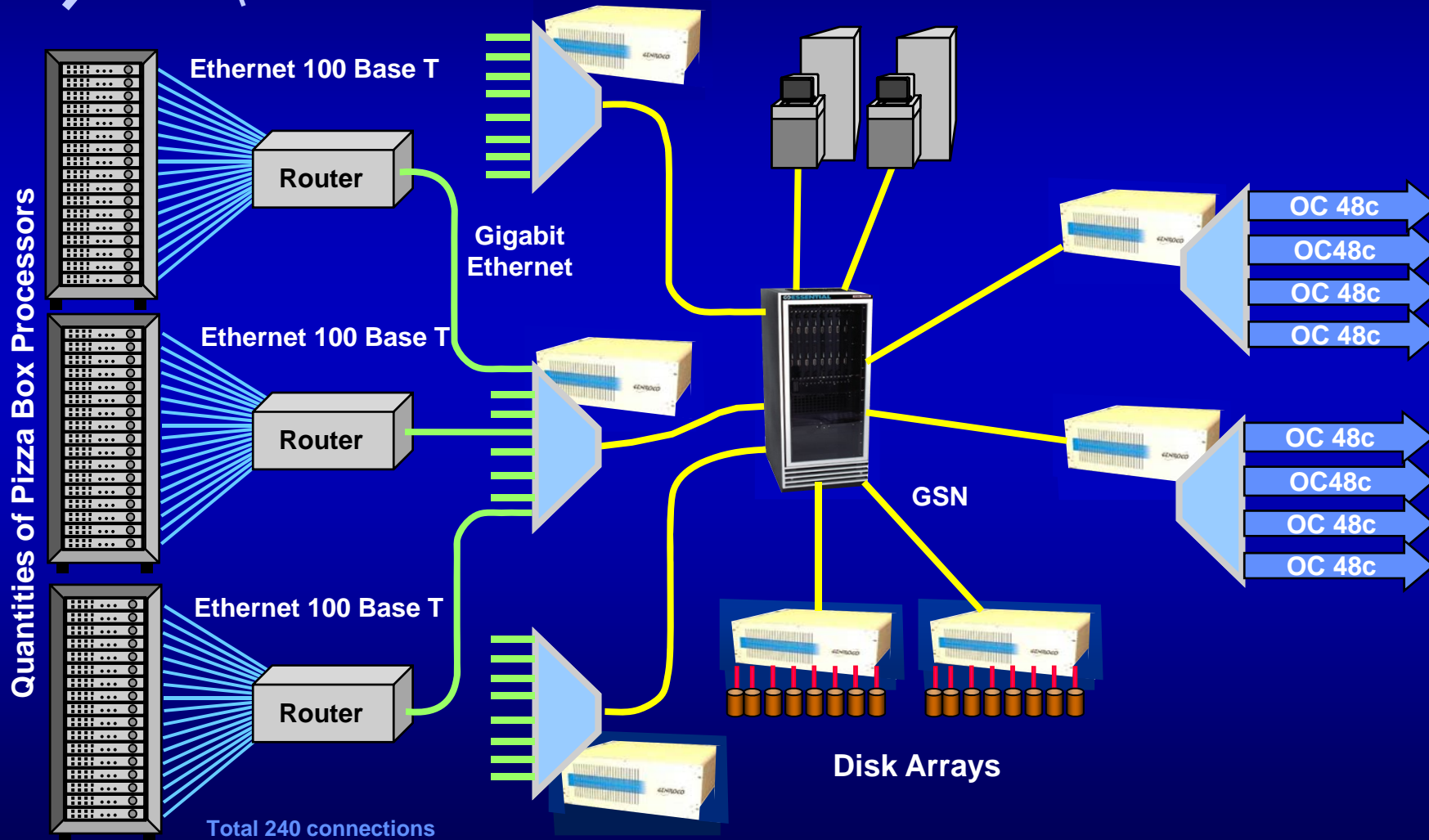
IP - Video OC192c
10 Gbit/s
1.25 GByte/s



CERN

High Performance Networking

Internet Service Provider Computing



CERN

High Performance Networking

Radio Astronomy (Jive)



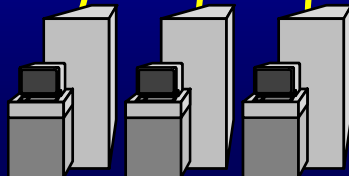
Possible now
to day:
OC48/SDH16
=2.5 Gbit/s



Bridge to
Gigabit Ethernet



Today and
Tomorrow
GSN + ST
10 Gbit/s



Dark Fiber

For Tomorrow:
10 GigE on
OC192/SDH48
=10 Gbit/s

H igh P erformance N etworking

Definitions for Network Storage

Secure Networks:

- ★ Network Integrity is built into the network technology (hardware)
- ★ Data Integrity is built into the network technology (hardware)
- Examples: GSN, Avionics Networks, Automotive Networks

Flow Controlled Networks:

- ★ Flow Control regulates the data stream on a Data Block base.
- ★ Used to avoid Buffer Overflow.
- Examples: SCSI, Fiber Channel, HIPPI

P & P Networks:

- ★ **P**ush the data on the Network & **P**ray it will arrive at the Destination
- ★ Data Integrity is build into the protocol not in the Network (TCP)
- Examples: All IP-only networks, Ethernet, etc.

High Performance Networking

Secure & Flow Controlled Networks




GSN with ST and SST

- Secure network makes the connections safe.
- ST protocol makes the end to end transfer safe.
- The host sees only SCSI commands.
- Read Cycles and heavy traffic conditions are not the problem as flow control and STU control in ST regulate the data streams.



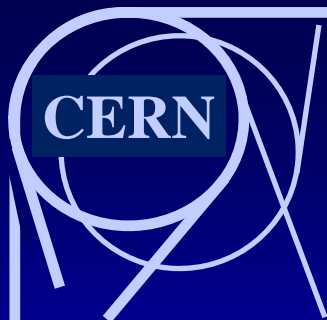
Flow Controlled Networks

- If flow control is endpoint to endpoint the behavior is almost as safe as a Secure Network but Bandwidth is influenced
- SCSI commands have to be encapsulated in TCP.
- If flow control is only point to point it has the dangers of P&P networks. See 

High Performance Networking

Storage on P & P Networks

- ★ **On a TCP/IP network:**
 - Network Congestion can evolve in corrupted or lost frames.
 - Switch Errors of all kind lead to corrupted or lost frames.
 - TCP will correct, but re-transmits and re-ordering frames brings high Latency → Throughput can drop as low as **40 %**
- ★ **iSCSI uses TCP/IP Protocol**
 - An enormous effort goes in the Standards work at 3T10 and someday it will work satisfactory, **efficiency factor ??**
- ★ **On networks Without Protocol or with IP only:**
 - Network Congestion can lead to corrupted or lost frames.
 - There is no Mechanism to Detect and Correct these errors, **Read** can be Reread, **Write** results in a Corrupted File.



H igh P erformance N etworking

Useful Information on the Web

GSN

<http://www.hnf.org>
<http://www.cern.ch/HSI/>
<http://www.cern.ch/HSI/HNF-Europe/>
<http://ext.lanl.gov/lanp/technologies.html>



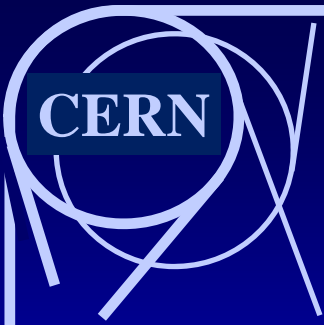
IB

http://developer.intel.com/design/servers/future_server_io/
<http://www.infinibandta.org/home.php3>

10 GigE

<http://www.10gea.org/>
<http://www.10gigabit-ethernet.com/>
<http://grouper.ieee.org/groups/802/3/ae/index.html>
<http://www.10gea.org/10GEA%20White%20Paper%20Final3.pdf>

Arie Van Praag CERN /PDP 1211 Geneva 23 Switzerland
Tel +41 22 7675034 e-mail a.van.praag@cern.ch



High **P**erformance **N**etworking

E N D