Technology Overviews: LAN Networks

LHC Computing Grid Workshop

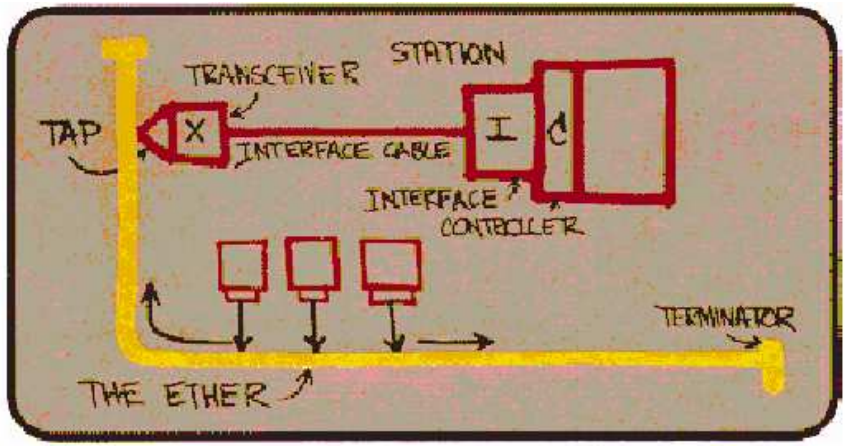
12 March 2002

Gaetano Maron

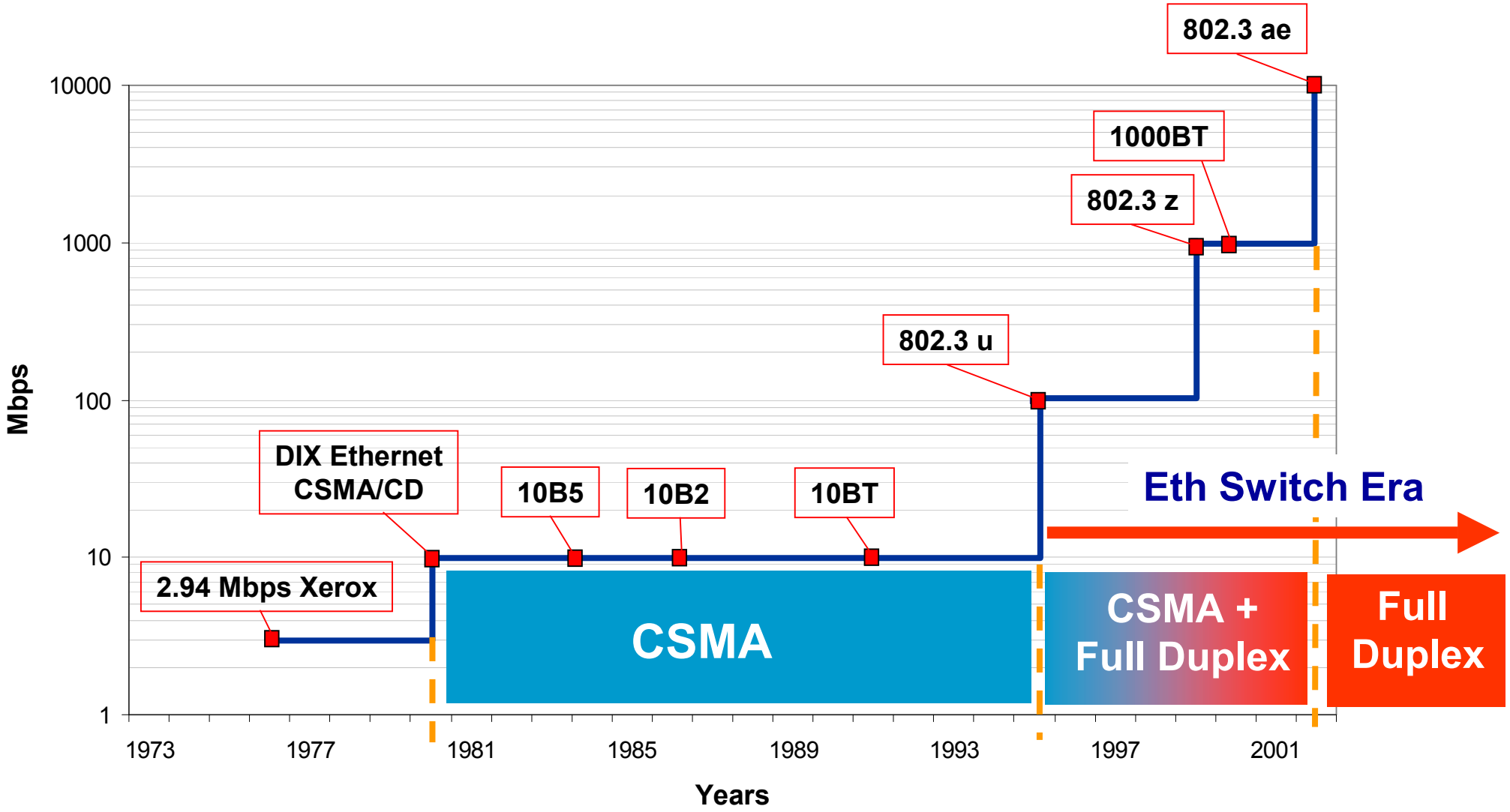
INFN – Laboratori Nazionali di Legnaro

Ethernet

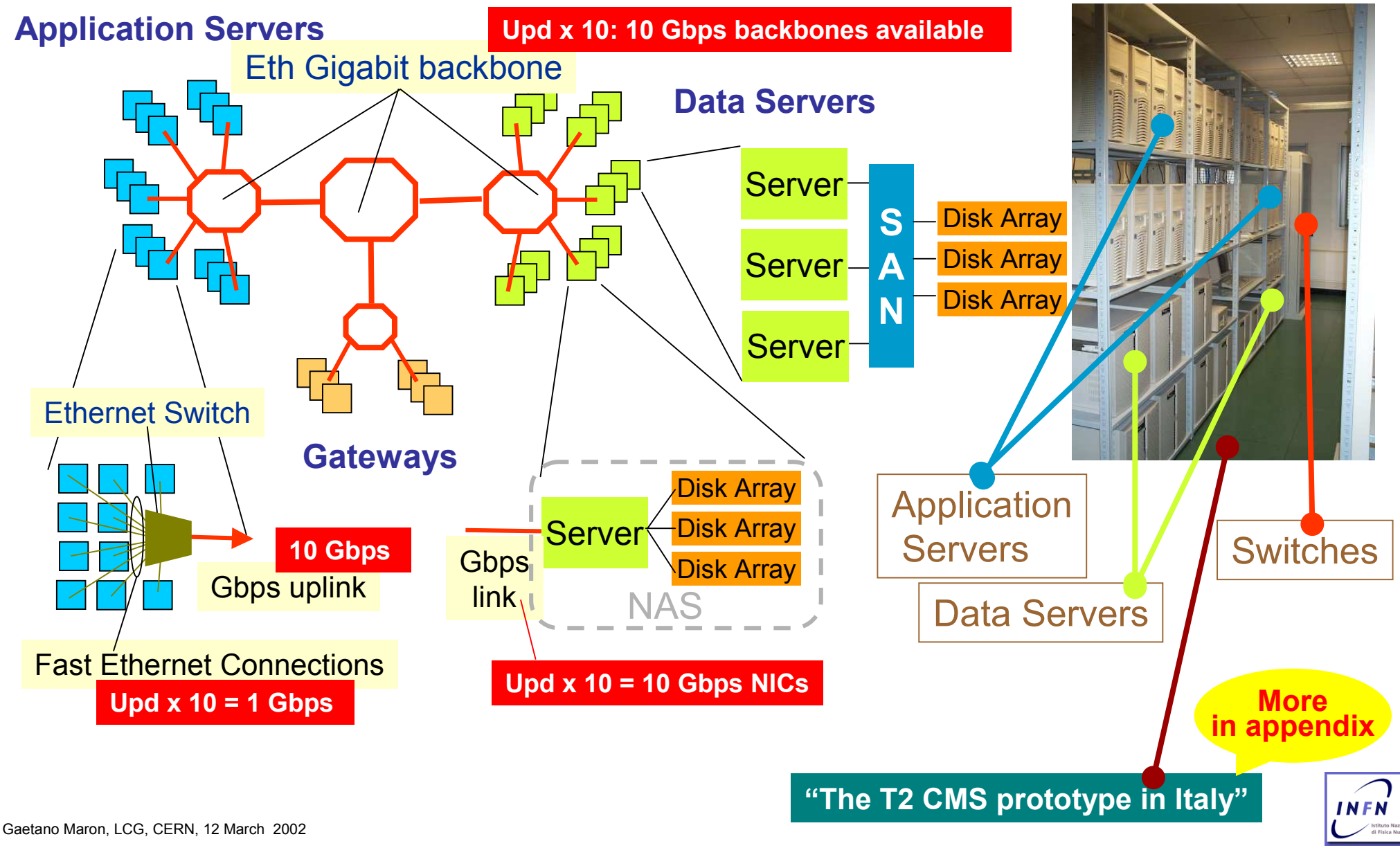
- Ethernet is the dominant technology in the LAN (10, 100, 1000 Mbps). It is everywhere
- 10 Gbps connections over long distances (40 kM) make it attractive also for MAN
- Most of the Internet data traffic is originated from a ethernet LAN and ends to a ethernet LAN
- All this started with this simple draw showing the first ethernet design (R. Metcalfe, Xerox, 1973) featuring an initial transmission rate of 2.49 Mbps.



Ethernet Evolution

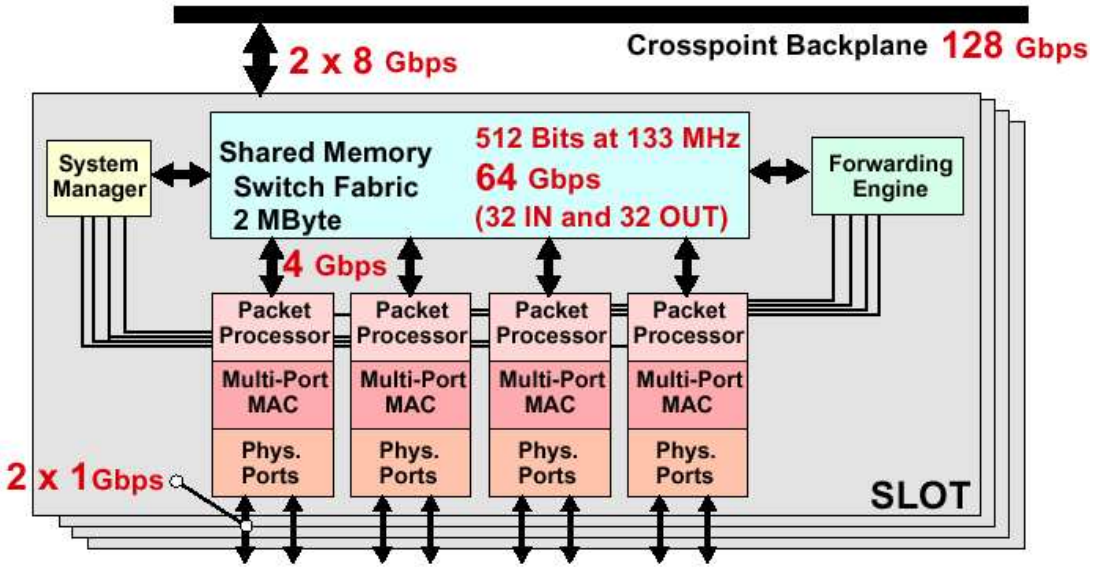


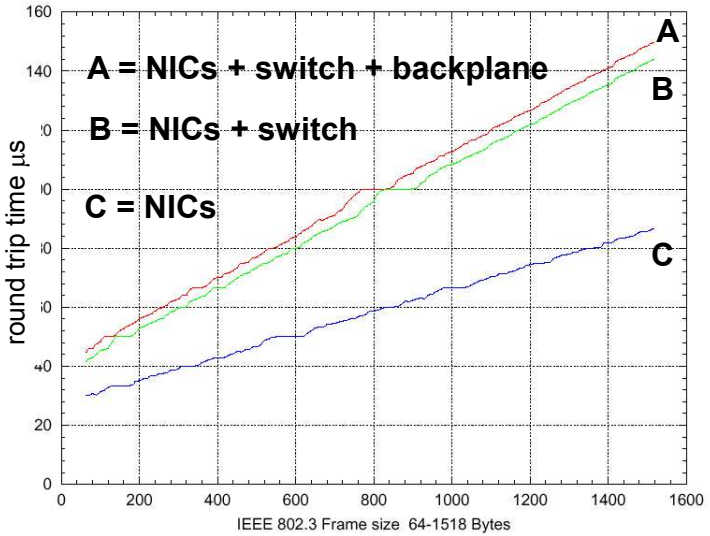
Switched Ethernet, Commodity PCs and Computing Farms

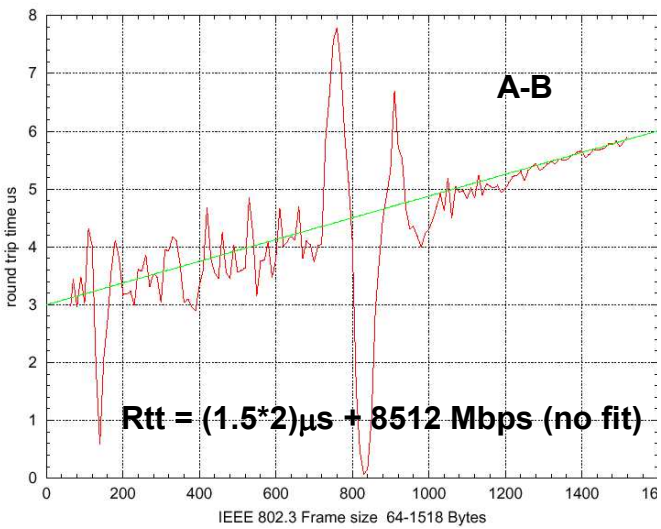


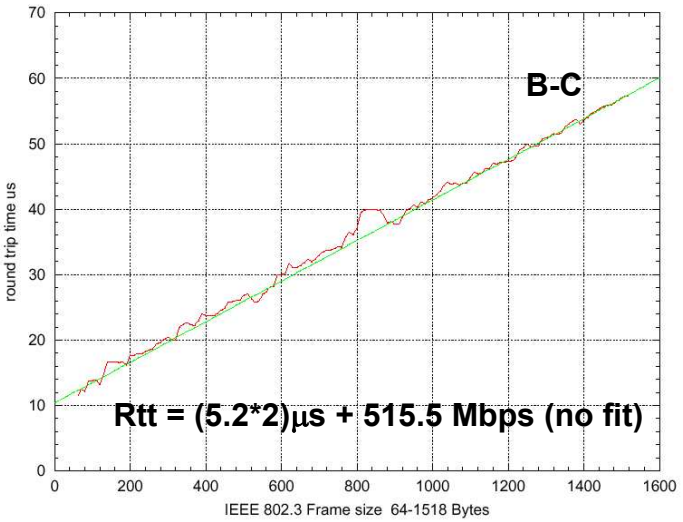
Ethernet Switches

O (10) 10 Gb ports
O (100) 1 Gb ports
O (1000) 0.1 Gb ports



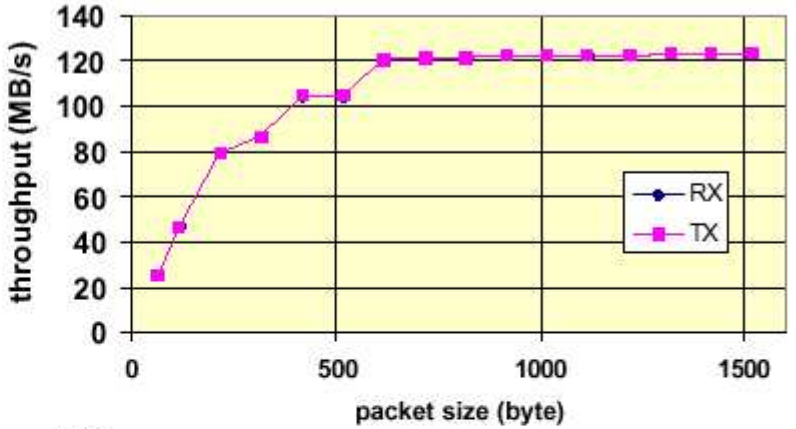


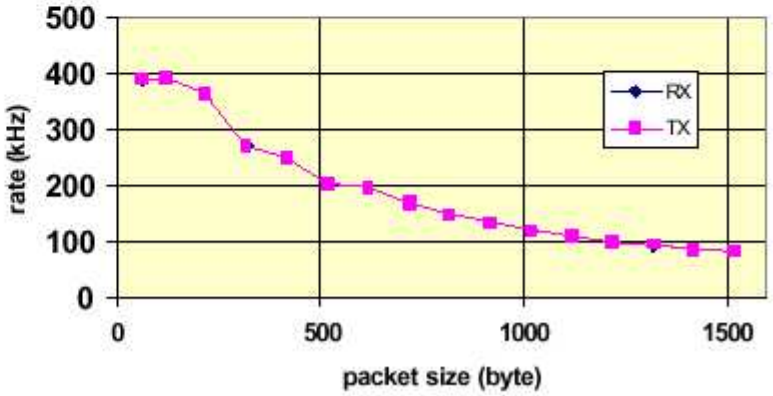




GigaEthernet Point to Point measurements

PC: supermicro PIIIDME
(i840) 64b/66MHz
NIC: SK9821





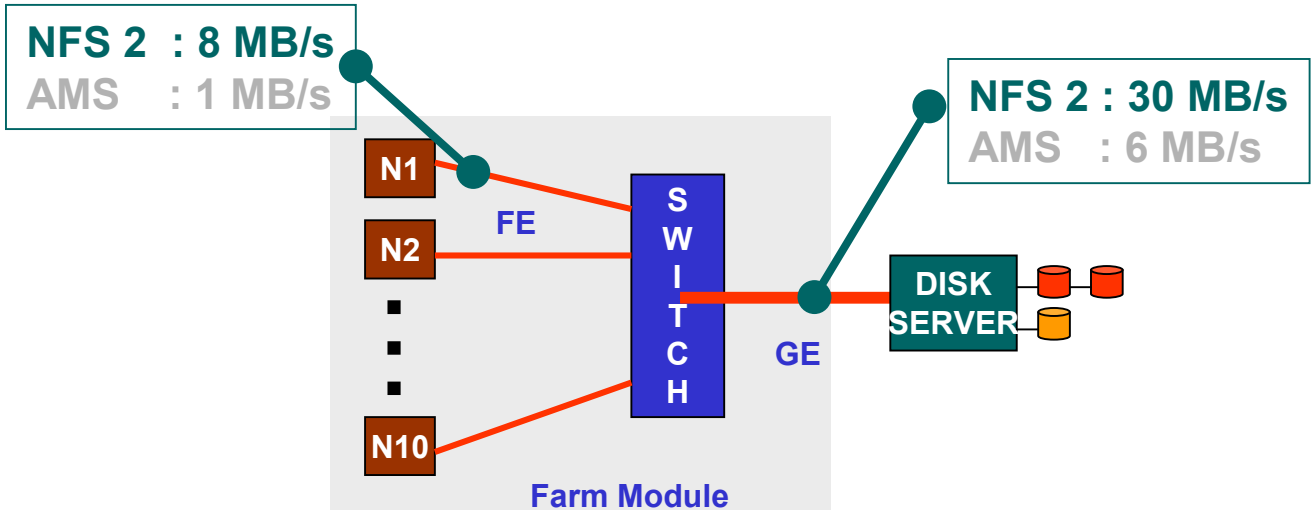
- Throughput up to 123 MB/s
  GE link 125 MB/s
- no packet loss

More in appendix

- **LAN based Event Builder**
- **CMS GE based Event Builder**
- **CMS Myrinet based Event Builder**

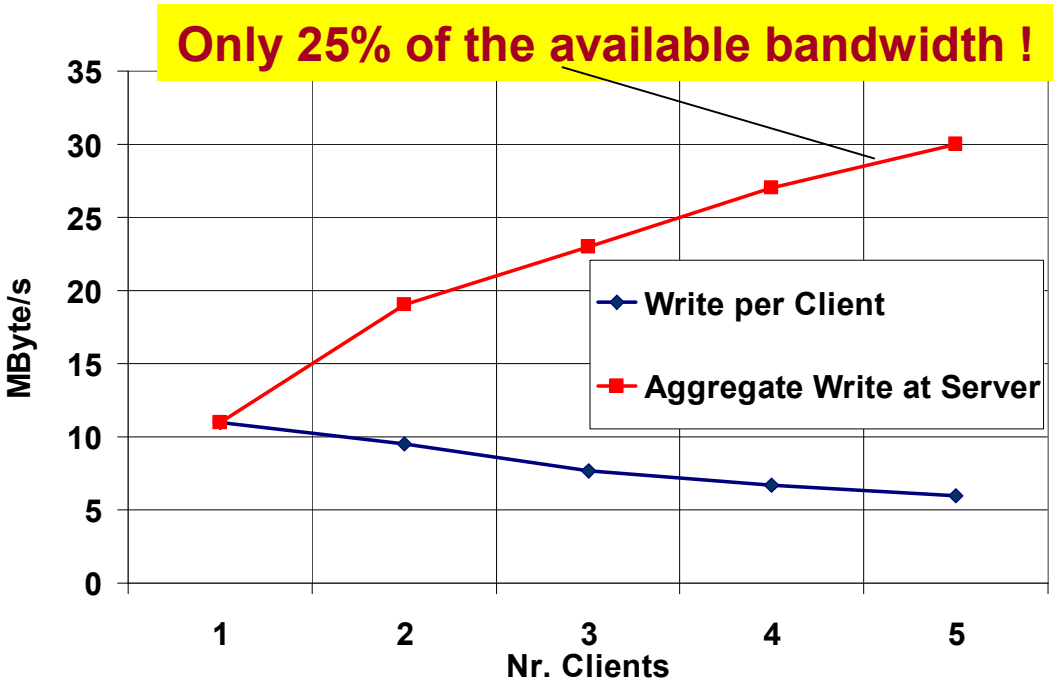
- no high level protocol used
- zero copy mechanisms

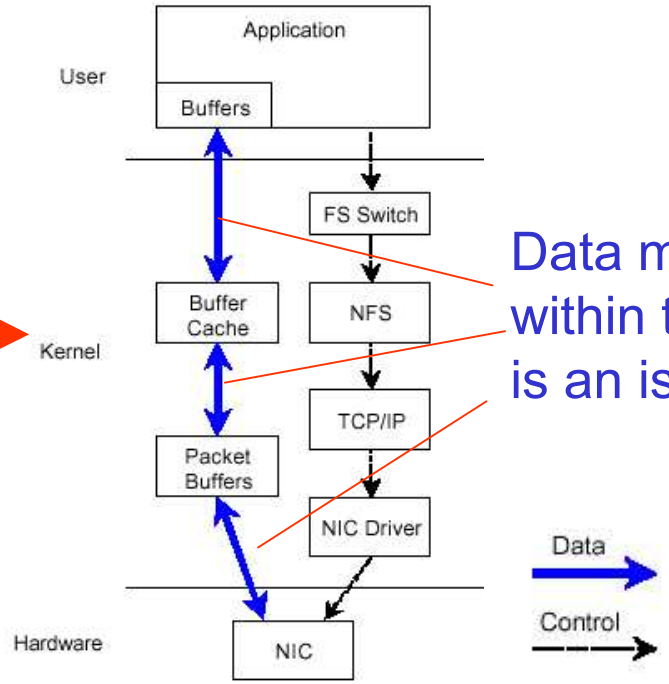
From the test beds to the production centers



NFS test bed (2001)

- 20 PC Dual PIII 800 MHz
- 3 server 1.3 TB disks
- Switch FE/GE: Extreme Summit 48
- Bench: iozone e bonnie with 1 Gbyte file size.





Data moving within the host is an issue

Are the IP protocols suitable for the 1-10 Gbps LANs?

- A key issue is the data moving within the host
- The host internal memory bandwidth is involved; faster CPUs does not help too much
- Interrupt rate and checksum calculation
- Basic structural issues involved:
  - interaction among NIC, OS, APIs, protocols
- Implementation should be reviewed for:
  - protocols (no data copy)
  - NICs (host CPU offloading)

More in appendix

• **Checksum offloading and zero copy**

INFN Istituto Nazionale di Fisica Nucleare

The Virtual Interface (VI): an interconnection architecture

- Originally developed by Compaq, Intel and Microsoft with the aim to define a standard paradigm for interconnecting computers in cluster.

- It is a standard interface for clustering software independent of the underlying network technology. Basically it is distributed messaging technology

- Basically there are two new capabilities:
  - Direct memory to memory transfer. It allows bulk data to bypass the protocol processing and to be transferred directly between the buffers on the communicating machines (**Remote DMA - RDMA**);
  - Direct application access; application processes can queue data transfer operation directly to VI compliant network interfaces without operating system involvement.

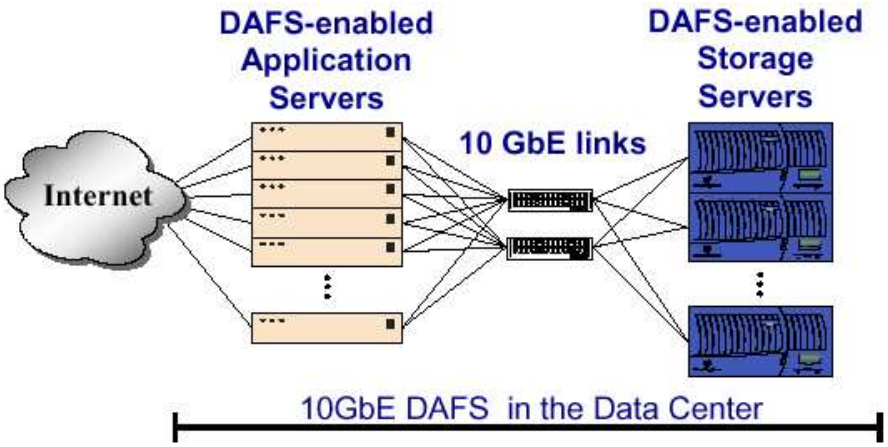
- All this:
  - improves the CPU utilization
  - reduces the latency
  - enables zero copy protocols

Current VI implementations :
- Fibre Channel – FC-VI
- 1 – 10 Gbps Ethernet (over IP) - VIIP
- Infiniband - VIPL
- proprietary interconnection networks
  - GigaNet cLAN
  - Compaq Servernet II

Storage Networking

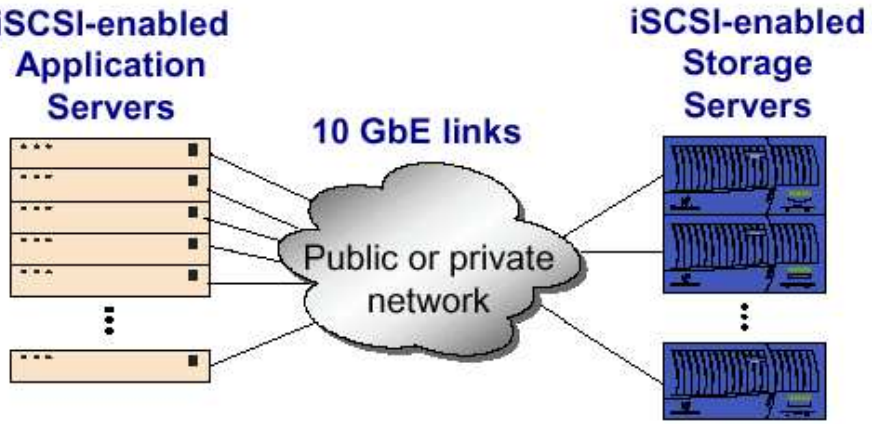
- Farm model uses a network based storage system
- LAN evolution (hardware, protocols, etc.) impacts then heavily on the storage networking evolution and then the design of out future farms

File Access



- TCP-offload Engines (TOEs)
- VI/IP transport
- 10 GbE switches
- File system & data management on storage server

Block Access



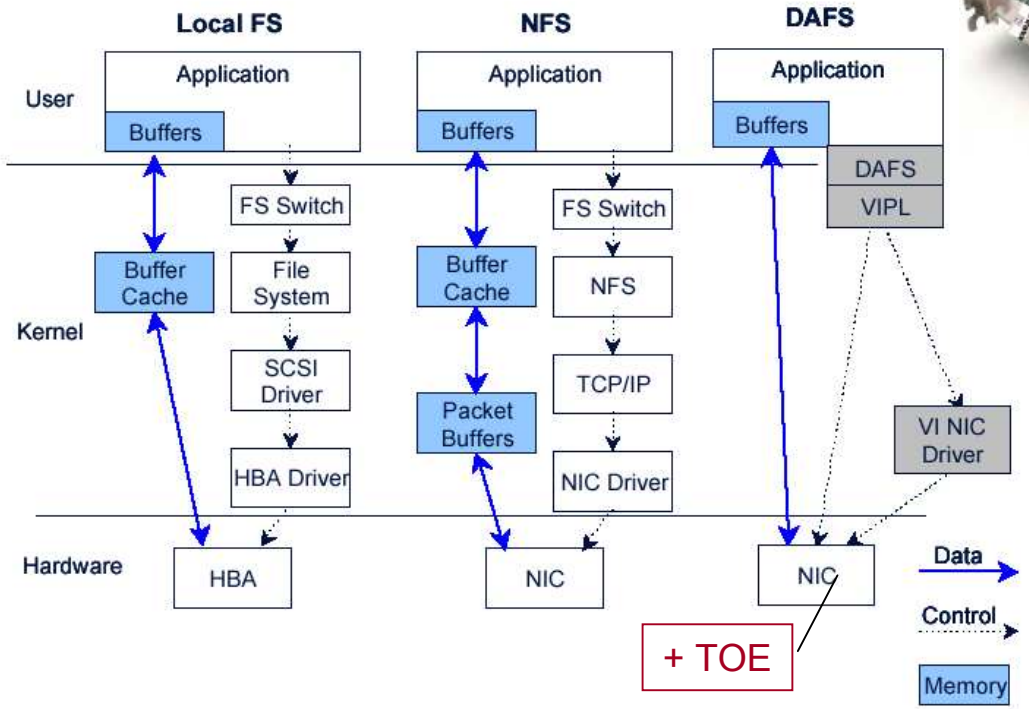
- iSCSI drivers on application servers
- TCP-offload Engines (TOEs)
- 10 GbE switches
- File system on application server
- Data management on storage server

Direct Access File System (DAFS)

- DAFS is a file access protocol based on NFSv4 that has been designed to take advantage of the VI memory to memory (RDMA) interconnect technologies

- NFSv4 used as starting point, but significantly improvements on:
  - fencing
  - predictable reply cache behavior for fail-over
  - enhanced locking

DAFS Collaborative:
- Intel
- IBM
- Compaq
~ 60 Companies



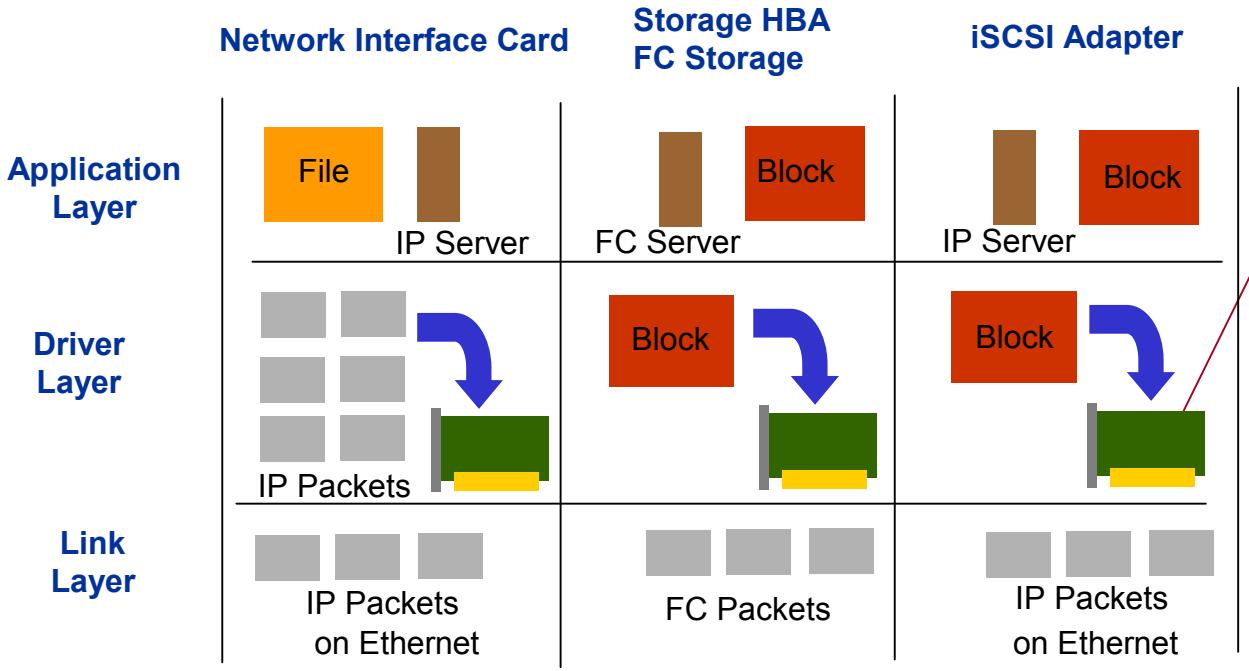
Emulex Ge VI/IP card

Troika FC-VI card

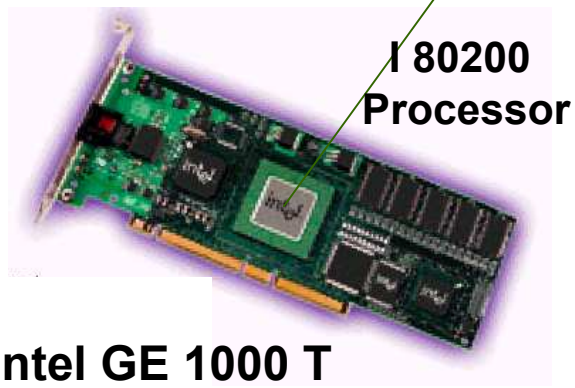
Storage over IP: iSCSI

- Internet SCSI (iSCSI) is a draft standard protocol for encapsulating SCSI command into TCP/IP packets and enabling I/O block data transport over IP networks

- iSCSI adapters combines NIC and HBA functions.

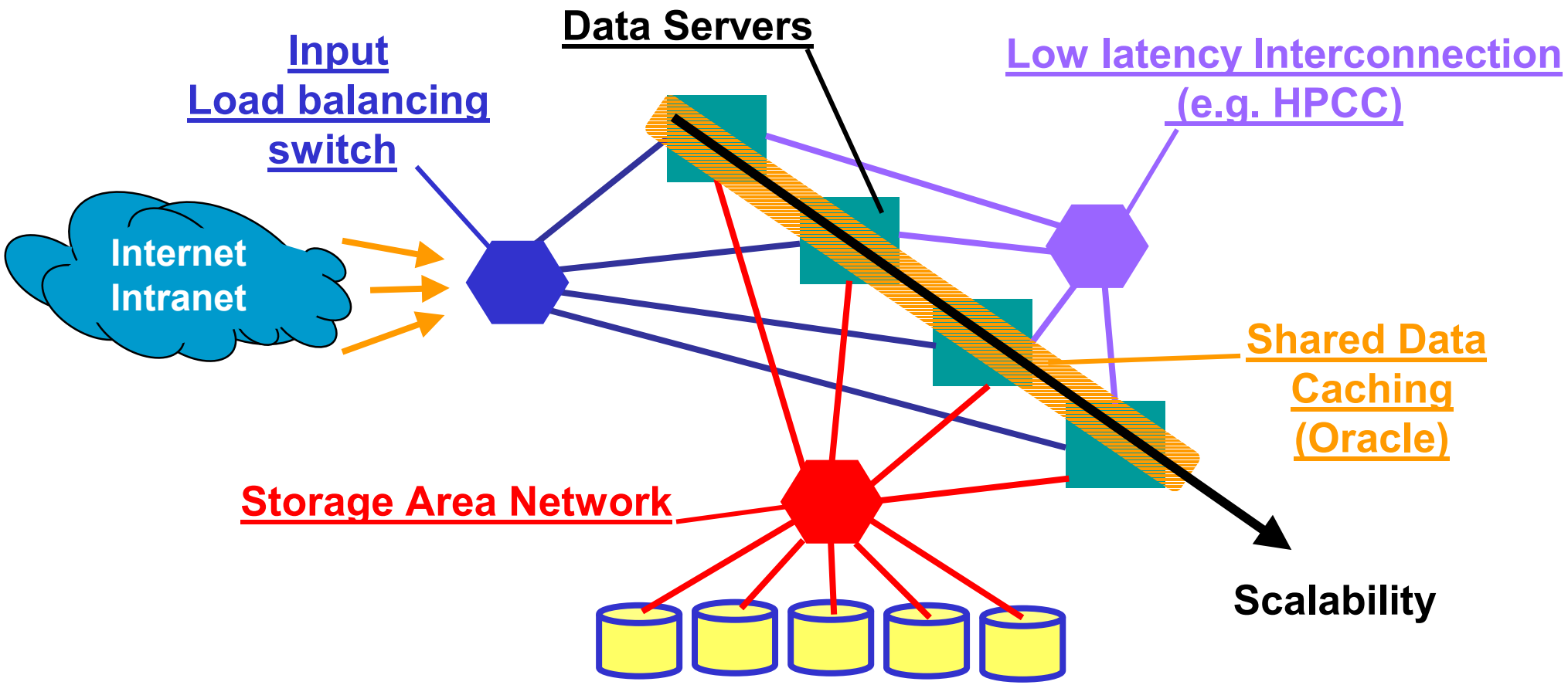


1. take the data in block form
2. handle the segmentation and processing with TCP/IP processing engine
3. send IP packets across the IP network

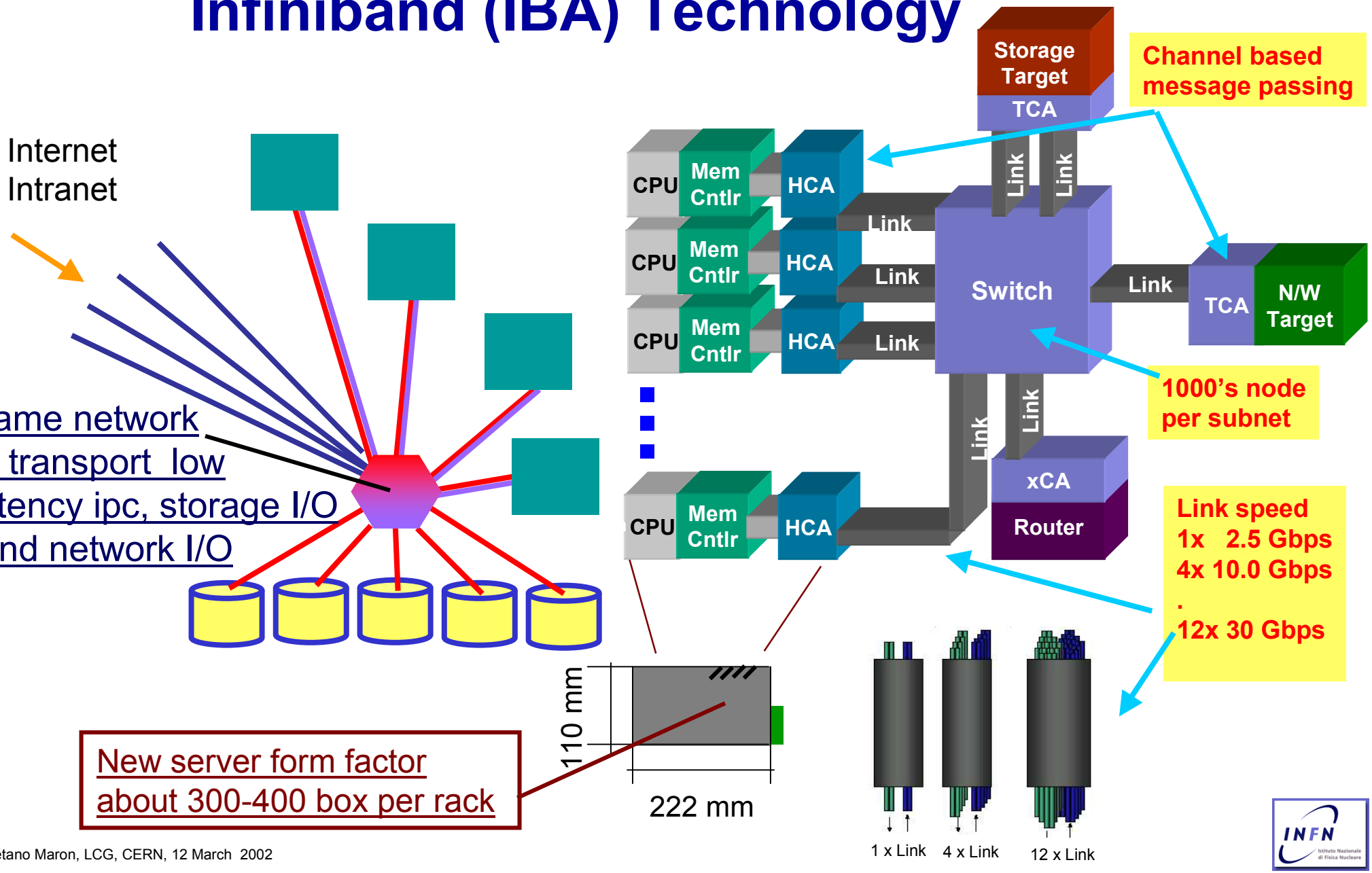


Intel GE 1000 T IP Storage Adapter

Parallel Servers to handle fast DBs



A single "transport" for everything: The Infiniband (IBA) Technology

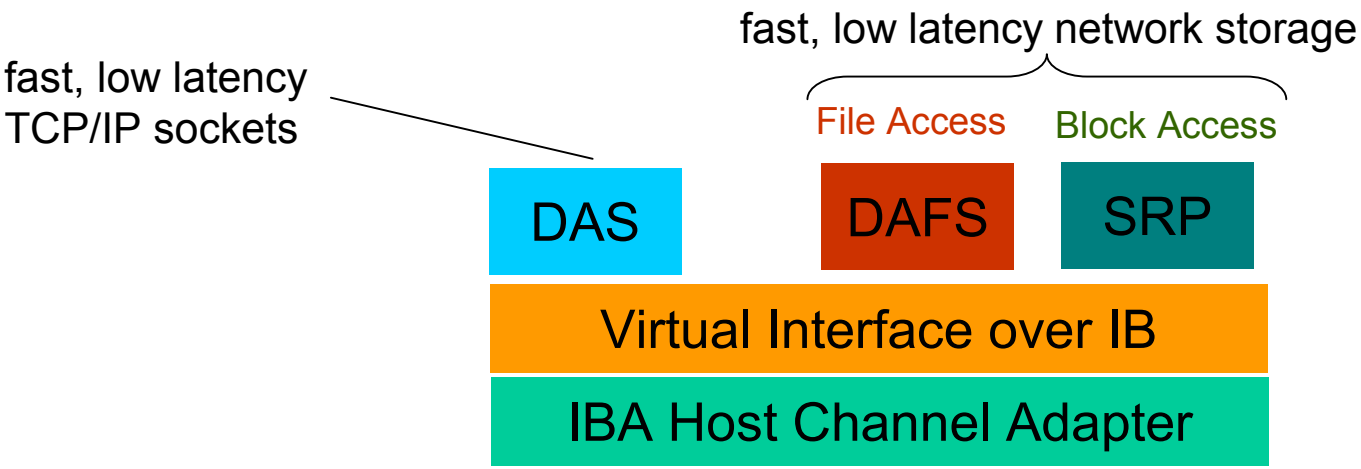


Infiniband transport protocols

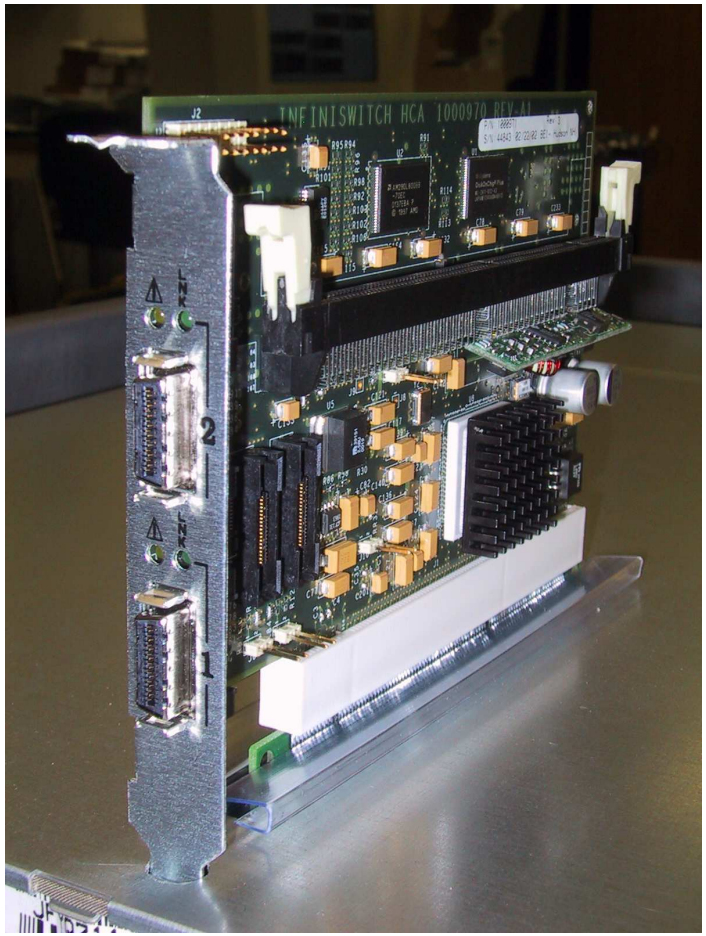
- IBA has been developed with Virtual Interface in mind. VIPL 2.0 includes IBA extensions and RDMA operations.
- SCSI RDMA Protocol (**SRP**). It is a T10 standard.
  - SRP defines mapping to IBA architecture
  - it is the transport protocol over IBA
  - SRP is based on VI
- Direct Access Files System (**DAFS**)
- Direct Access Socket (**DAS**)
  - TCP/IP functionality over VI/IB

More in appendix

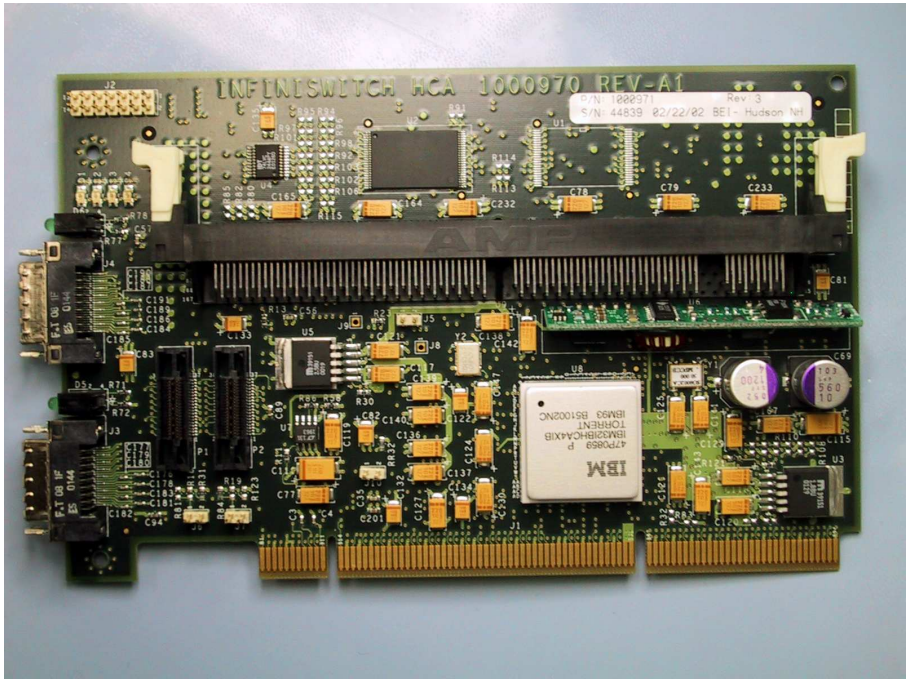
• **Infiniband Main Characteristics**



IBA Host Channel Adapters



Source: InfiniSwitch Corporation



Source: InfiniSwitch Corporation

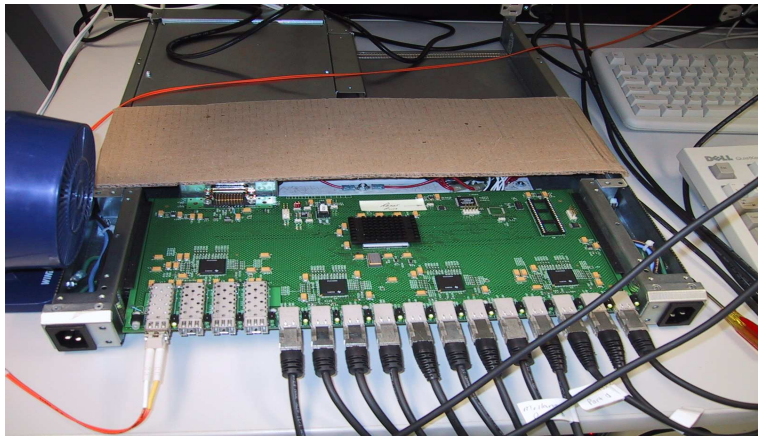
Torrent HCA

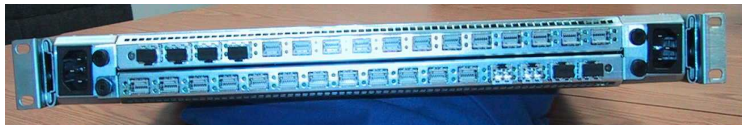
- 64K Queue Pairs
- Expandable Memory (256M)
- Dual 4X (10 Gbps)
- Integrated PPC 405
- Small Form Factor PCI / PCI-X
- Copper and/or Fibre

IBA Switches

1x Leaf Switch

32 1x (2.5 Gbps) ports in 1 U chassis

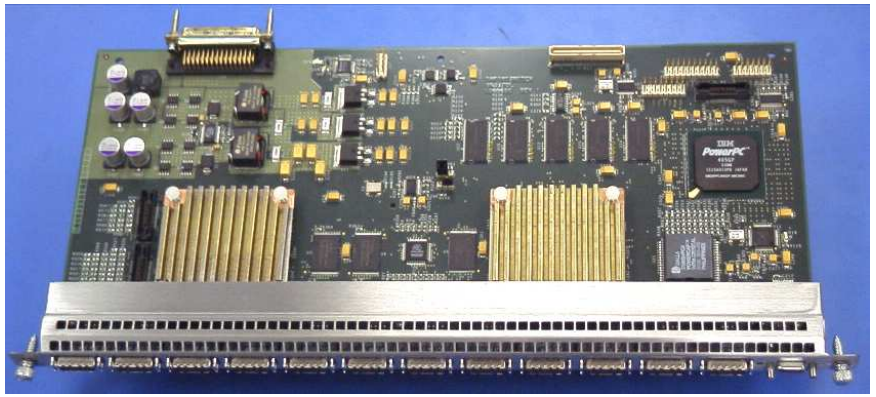




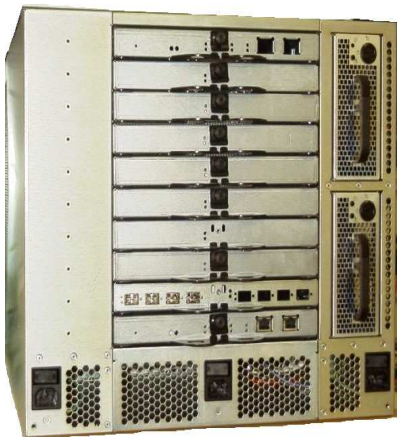
Source: InfiniSwitch Corporation

4x Leaf Switch

24 4x (10 Gbps) ports in 1 U chassis

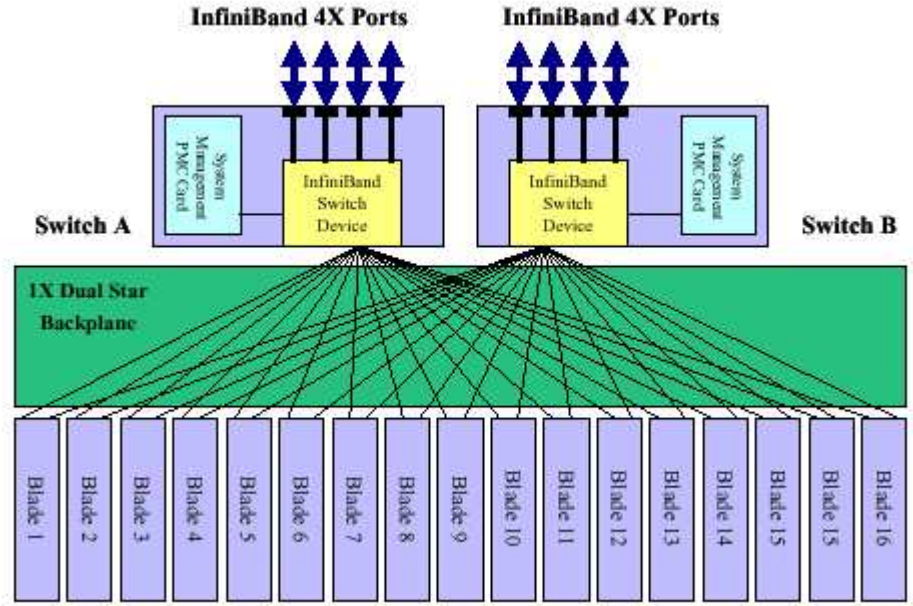


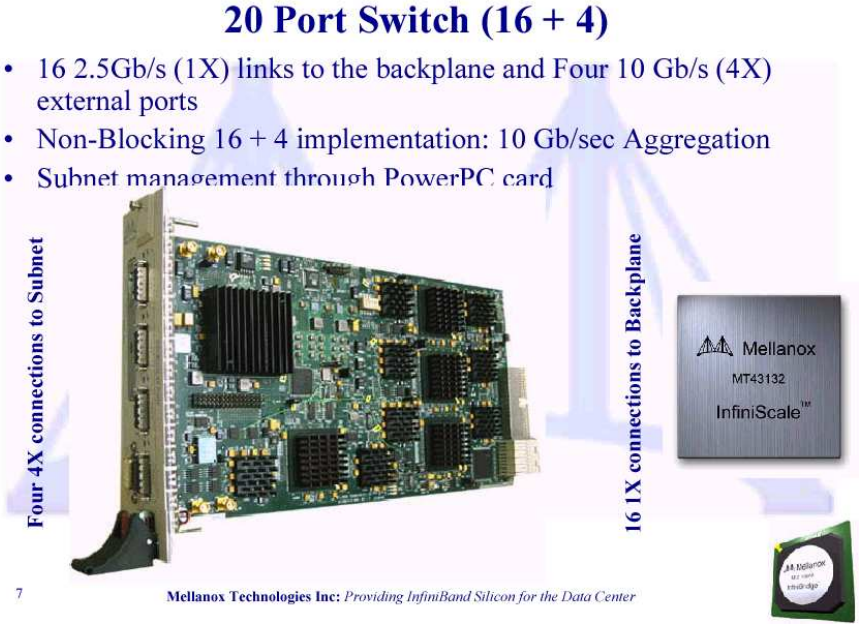
Director Switch



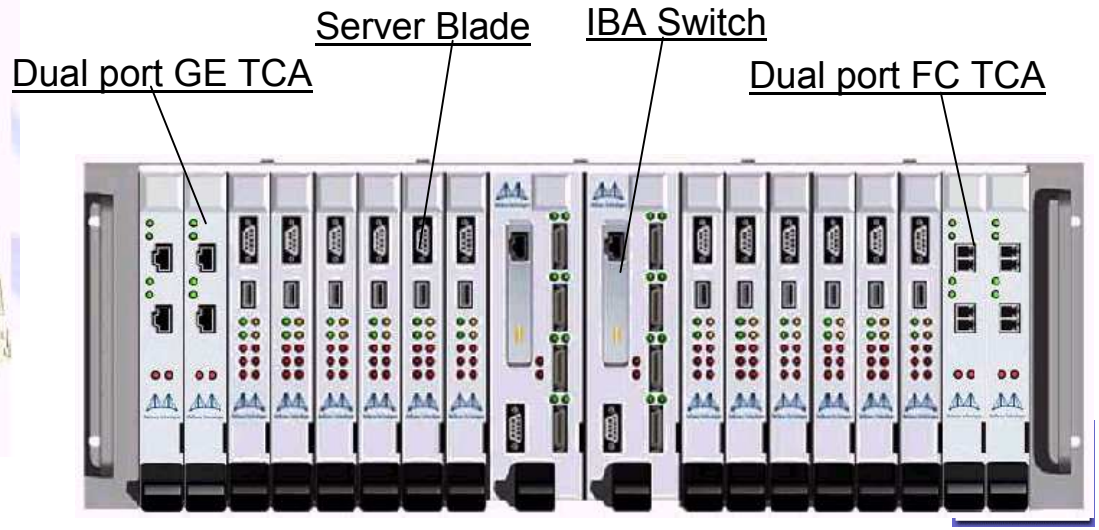
- 64-port capable, increasing to 256
- 1X, 4X, 12X
- Flexible I/O Support
  - Fibre Channel
  - GbE
  - iSCSI
- Flow control

IBA Blades





IBA Server Blade
1.26 GHz PIII Tualatin
35 W/Blade
ServerWorks LE



INFN Infiniband Pilot Project

- formal agreement with:
  - Intel
  - Infiniswitch
  - Mellanox (in discussion)

- aim of the project:
  - IB link characterization
  - IB based farm to experience
    - low latency farming
    - storage over IB (both block and file access)
    - blade server configuration
  - IB based event builder

- Status:
  - first 4 HCA + 1 Leaf switch to be delivered beginning of April

Ethernet and IBA

ETHERNET

16 port Gbps +
1 port 10 Gbps
On chip

Level of Integration
1 10 Gb switch port
fits in a 9U module
(Man Application)

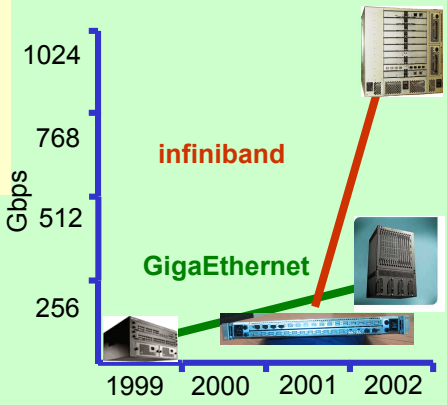
Switch Latency	5 μs
Connection Type	unreliable (reliable TCP)
Speed	1 Gbps 10 Gbps

10 Gbps — DAFS/VI NFS/VI — iSCSI — VI MPI/VI

MAN	LAN	NAS	Storage AN	System AN

Level of Integration
- 8 x 10 Gbps on chip switch
- 160 Gbps aggregate bndw
- Integrated phy layer
- 520 pin package

DAFS/VI — SRP/VI — VI MPI/VI



INFINIBAND

Switch Latency	0.5 μs
Connection Type	reliable hdw
Speed	2.5 Gbps 10 Gbps – 30 Gbps

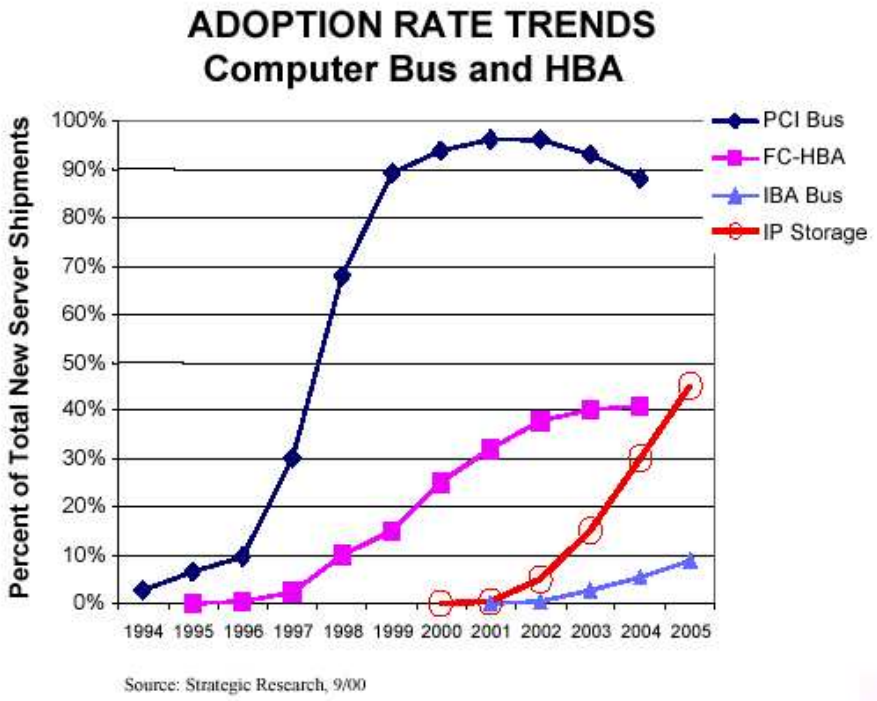
Final Remarks (I)

- Future LANs will continue to be dominated by Ethernet
- Recent Ethernet developments (TOE, VI, etc.) extends its application field to :
  - Storage Area Network (iSCSI)
  - High speed NAS (DAFS)
  - Low Latency IPC
- This can have a significant impact on the design of our future farms
- These fields are new, but they are maturing quickly (rate of announces is impressive)
- Storage over IP at 1 Gbps will lead to a medium performance commodity storage network
- Storage over IP needs 10 Gbps to be competitive with SCSI, FC and IBA.
- 10 Gethernet is at the moment focused on the backbones and on the MAN applications (high cost x port). To extend it to the LAN applications is needed to have:
  - Small, compact and “cheap” switches
  - NICs and HBA to to entry in the host systems
  - PCI-X based host systems (ok)
- Storage over IP is an excellent way to connect different storage technologies, as it speaks the “lingua franca” of the networks: TCP/IP.
- We need to experience all this

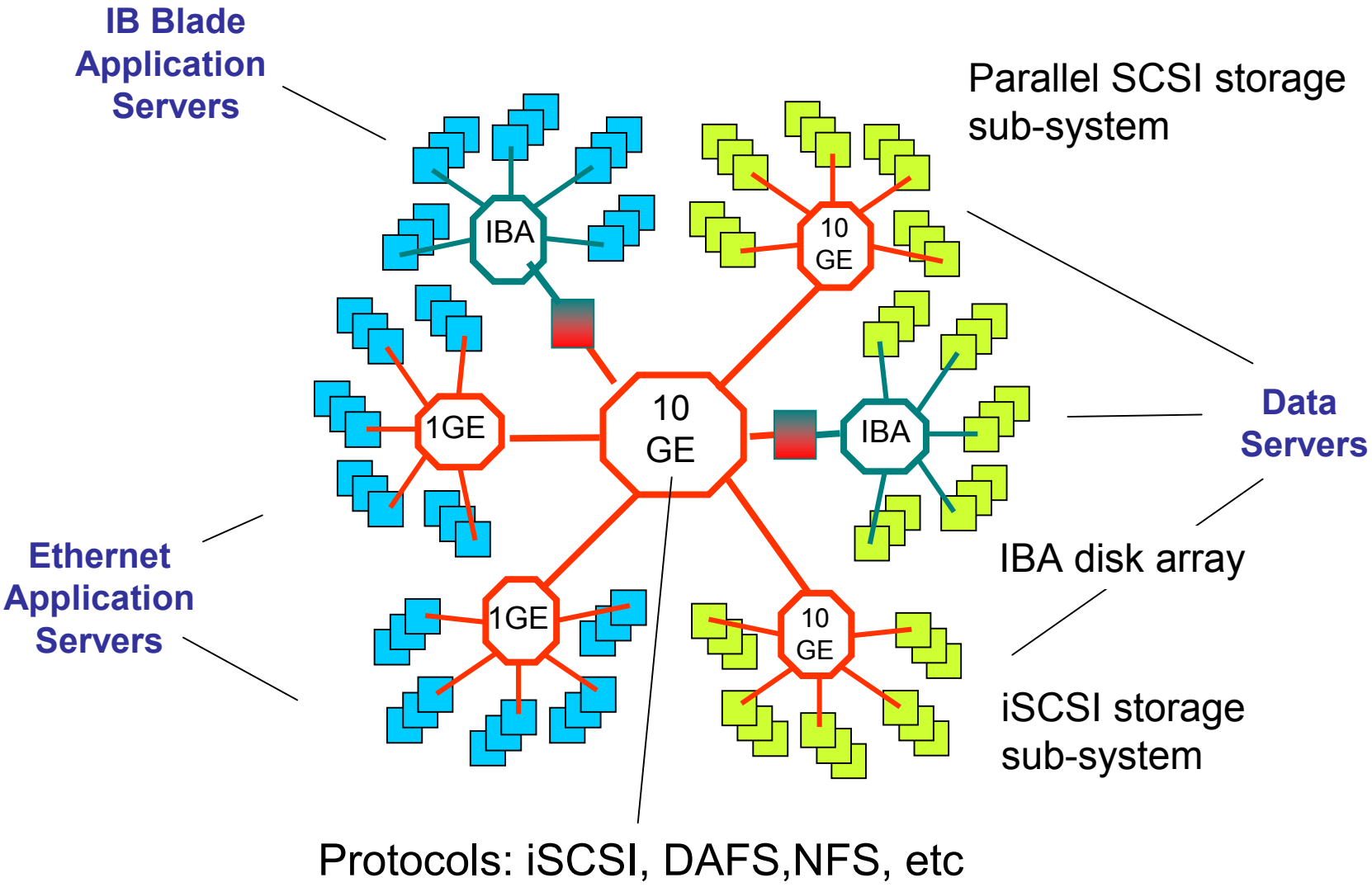
Final Remark (II)

- IBA has been designed for low latency high speed clustering. Max copper cable length is < 20 m.
- IBA has been design with VI in mind and then is optimized by definition for:
  - Storage Area Network (SRP)
  - High speed NAS (DAFS)
  - Low Latency IPC
- It features high level of integration, multiple speed range (2.5, 10, 30 Gbps). This should lead to low cost per port (also if it will be not classified as “commodity component”)
- First (few) products only now. Real take off not before 2003; full deployment of the technology will require long time.
- IB based blade servers are interesting for our farms.
- IB based farm backbone will be prob. cheaper than Ethernet ones.

- Storage over IP can be transported over IBA
- IB native storage (?)
- Needs for test beds



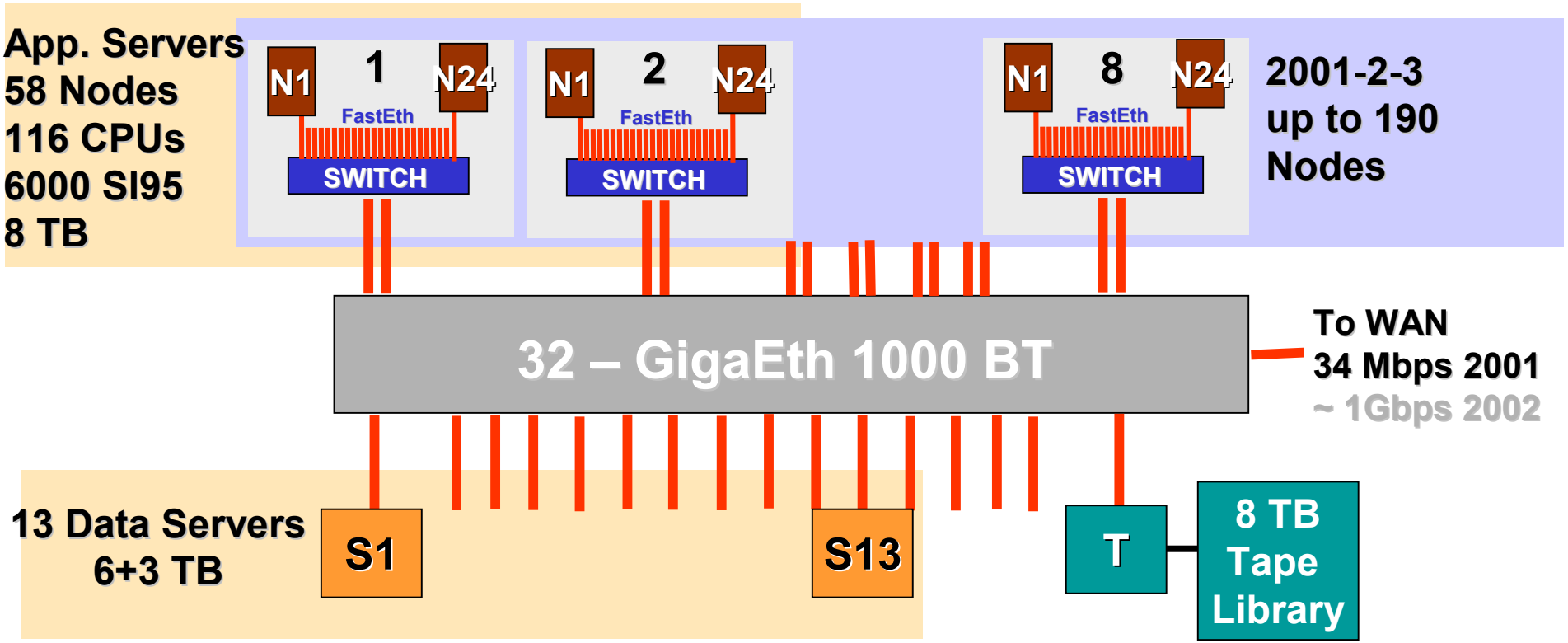
A resulting scenario



Conclusions

Appendix

A1 : The CMS T2 Prototype in Italy

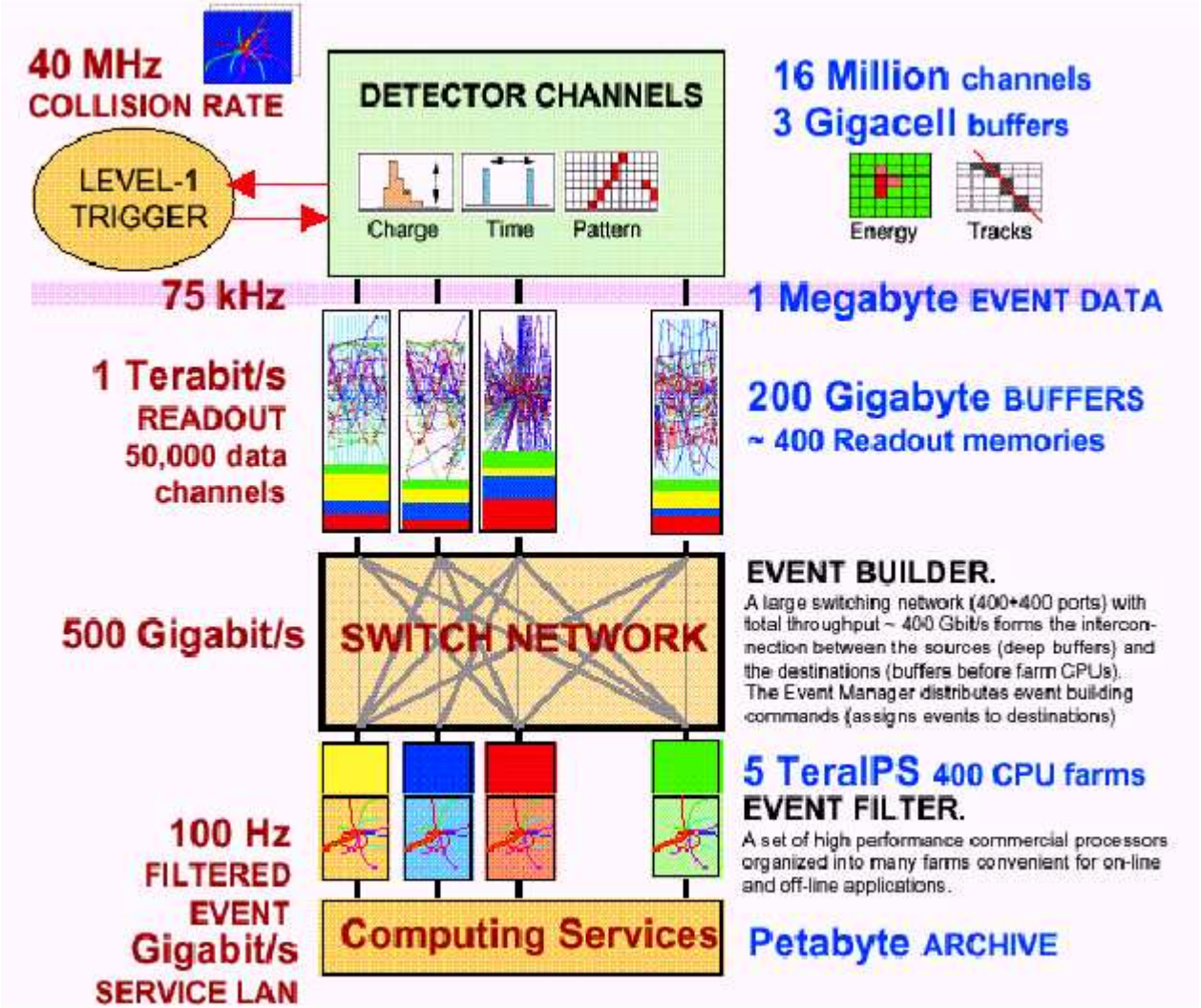


Application Server
Dual PIII – 1 GHz/ 1.26 GHz Tualatin
512 MB
2x80 GB Eide disk + 1x20 GB for O.S.

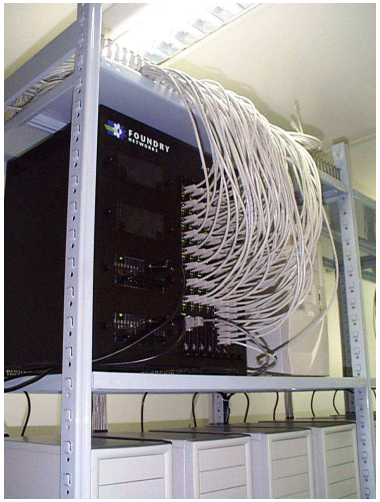
Data Server
Dual PIII – 1 GHz
Dual PCI (33/32 – 66/64 512 MB)
8x80 GB Eide Raid disks or
6x180 GB SCSI Raid disks
1x20 GB disk O.S.

A2: A severe test bed: LAN based Event Builder

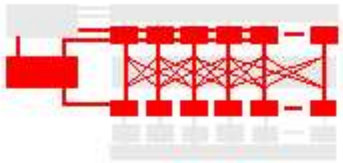
- Event Builder (EVB) in the LHC experiments are performed on switched networks
- CMS EVB needs a Tbps network. Two approaches under investigation:
  - Gigaethernet based EVB
  - Myrinet based EVB
- EVB demonstrators and related simulations are severe test beds to validate new networking technologies

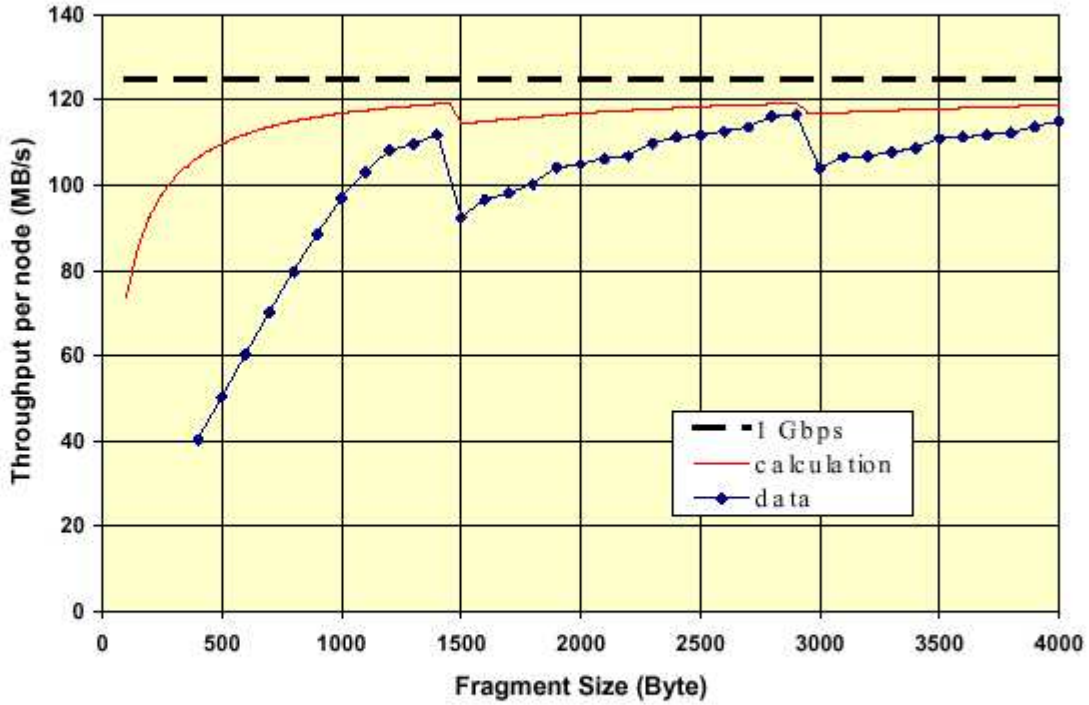


A3: The CMS GE based Event Builder Demonstrator



EVB 15x15 performance - Throughput





- Throughput up to 116 MB/s, ie 93% link speed
- sawtooth due to MTU (no event aggregation)
- no packet loss observed
- scales
- aggregate throughput ~15 Gbps

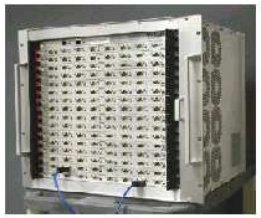
A4: Myrinet based Event Builder

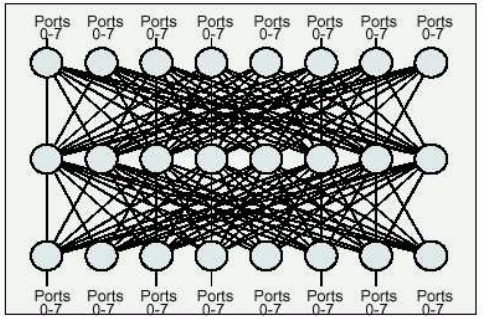
- Myrinet typically used as **cluster interconnect**
- **point to point links**, byte wide, full-duplex, **2 Gbps** per direction, very low error rate

- **packet structure**: routing header, payload and tail
  each crossbar switch strips leading byte from routing header
- **wormhole routing** (versus store-and-forward)
  no buffering, low latency, arbitrary length packets

- byte based **flow control** (STOP/GO)
- **no packet loss** inside switching fabric

- basic unit Xbar16 (8x8)
- CLOS networks, eg CLOS-128 switch





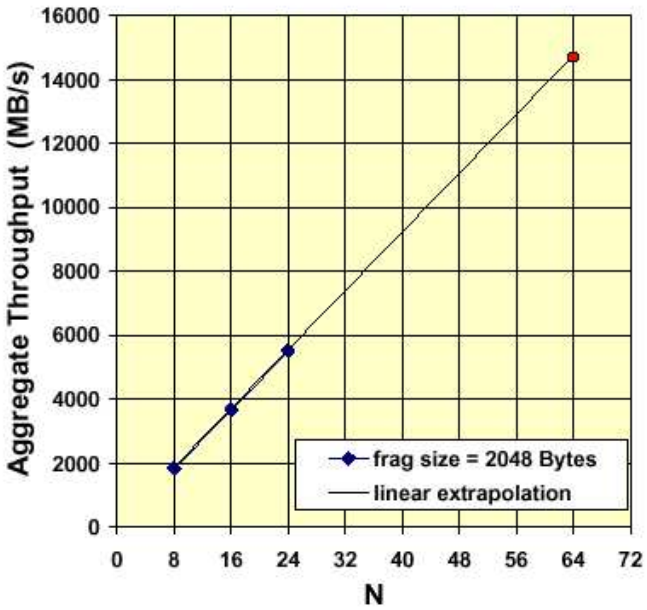
8-port line cards with one Xbar16 each

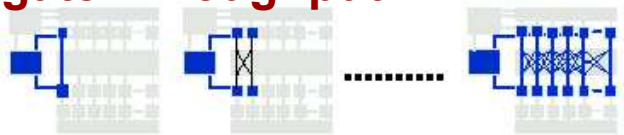
pre-wired network on backplane

8-port line cards with one Xbar16 each

128-Port Clos Switch

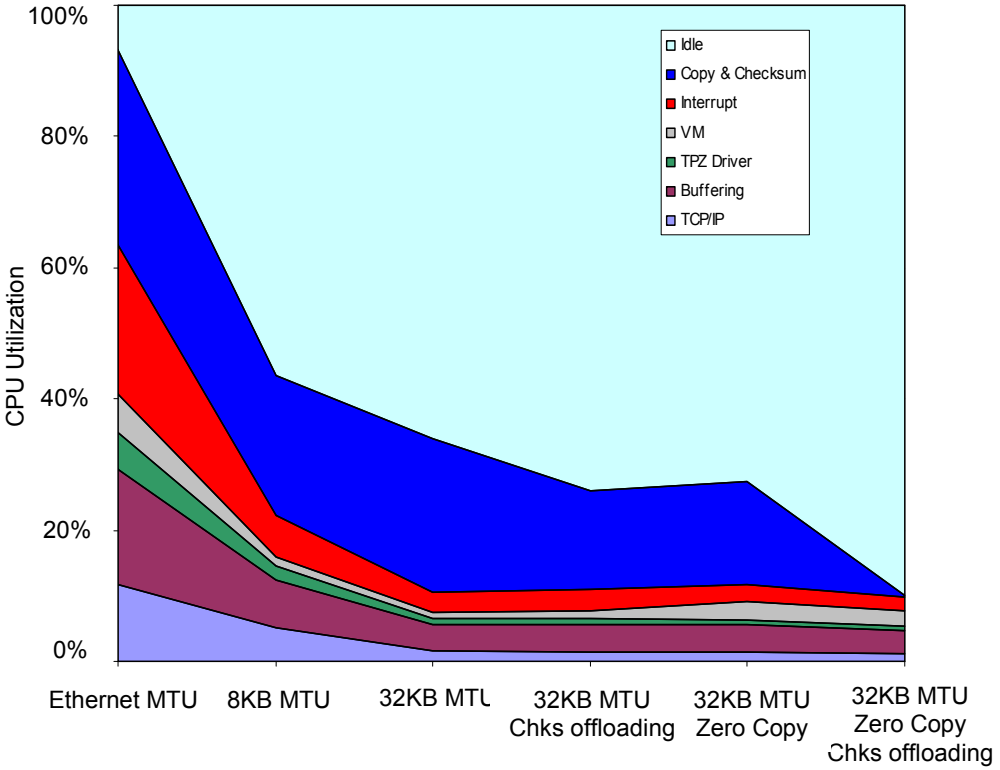
24x24 EVB Aggregate Throughput

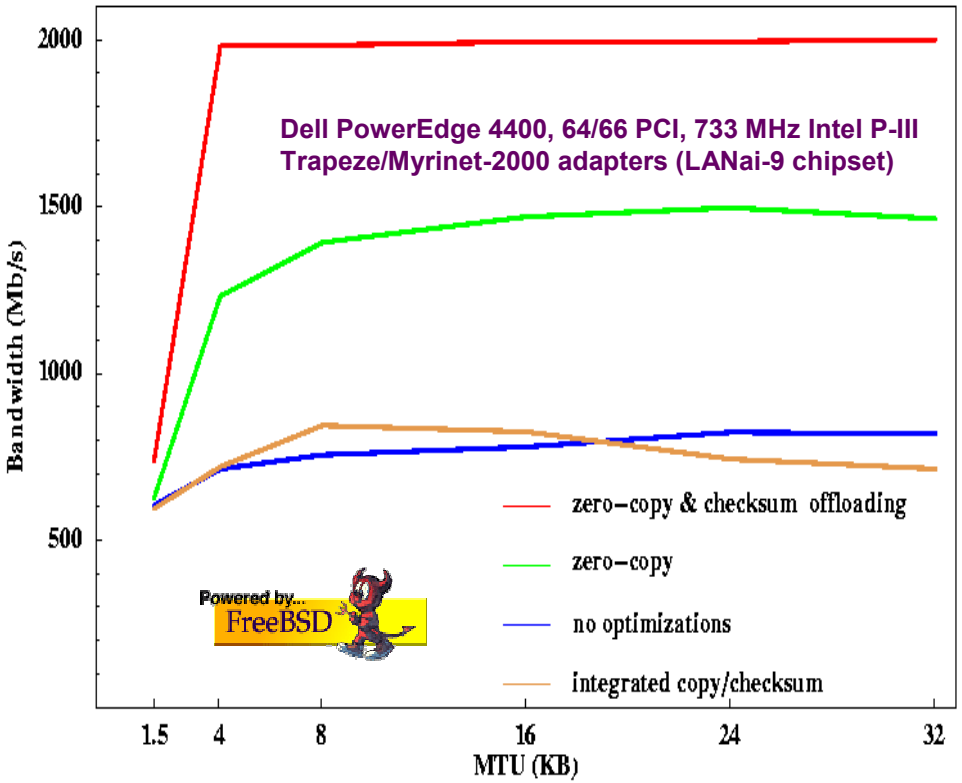




Assuming scaling
with fully populated Clos-128
64x64 EVB with
14 Gbyte/s aggr. throughput

A5: Checksum offloading and zero copy

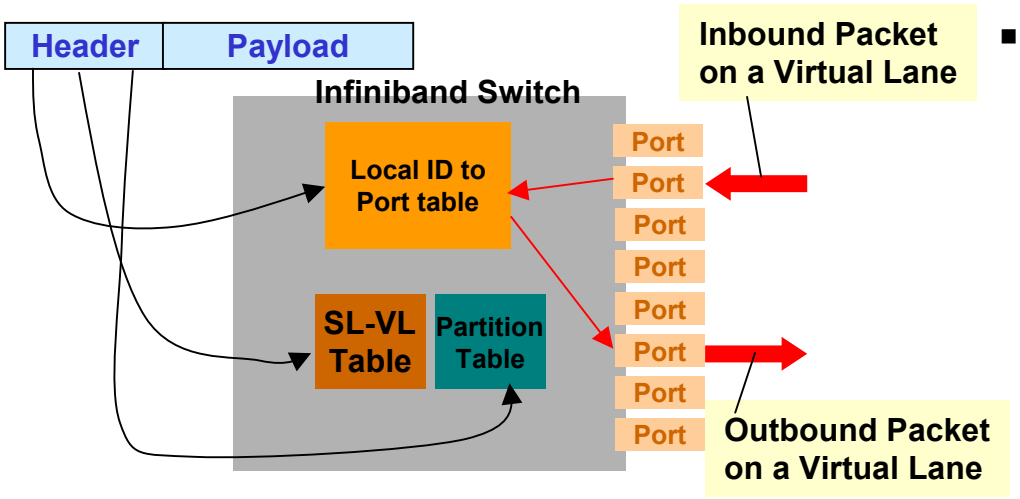




A6: Infiniband main characteristics

- **Topology**
  - Switched Fabric
  - Thousands of nodes per sub-net
  - Multiple subnets bridged w/routers
    - IPv6 addressing x-subnet
- **Fabric Transactions**
  - Unified fabric for IPC,
  - Networking, and Storage
  - Channel based interconnect
  - QoS (Service Levels, Virtual Lanes)

- **Reliability**
  - Automatic fail-over in switch
  - Support for redundant fabrics
- Physical Layer
  - Four wire link (2 pairs)
  - 2.5Gb/sec signaling rate, dual-simplex
  - Copper & Fiber support
    - Copper - 17M
    - Fiber - 1X - 100M - 10KM
  - Multiple link widths
    - X1, X4, X12



- **Protocol**
  - 16 bit local address / multiple MTUs (256-4096)
  - VI-based service types with extension
    - connected, datagram, reliable datagram, raw datagram, atomic operations, multicast
  - Ordering guarantees for connected services
  - HW acknowledge
  - Credit-based link flow control
  - End to end flow control
  - In-band management