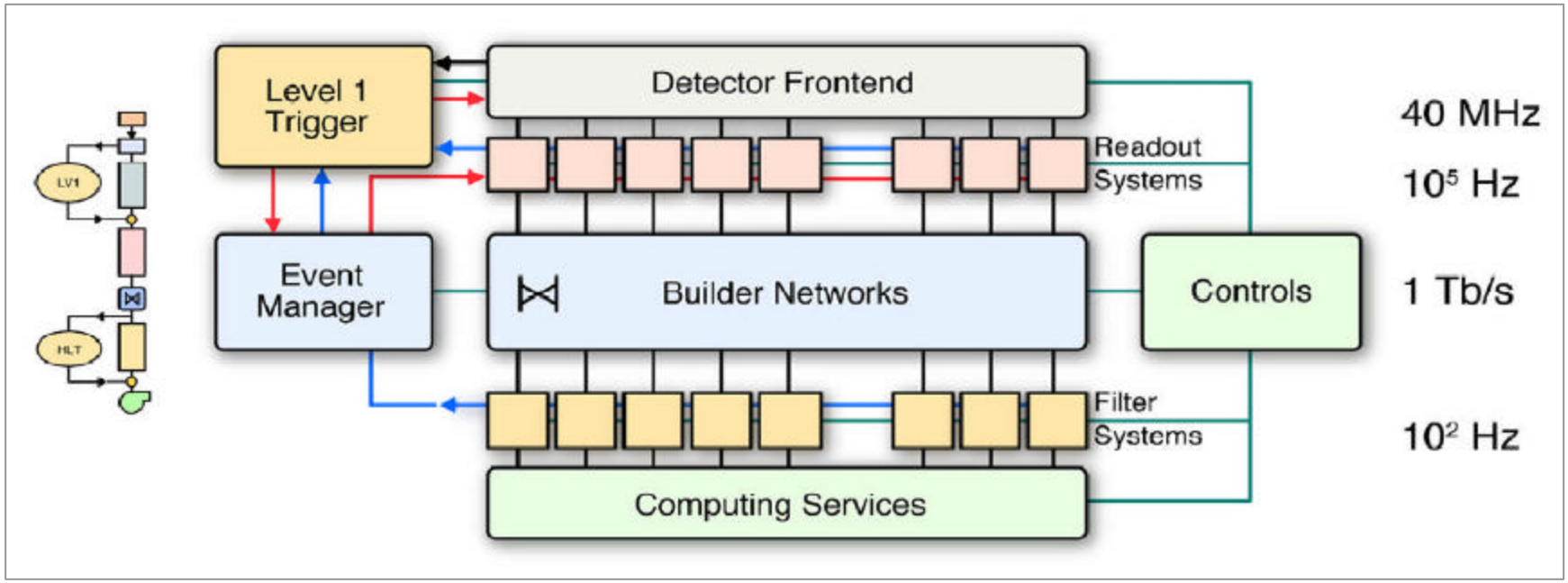
The CMS Builder Unit and Filter Farm

Jim Branson, UC San Diego

BU Requirements

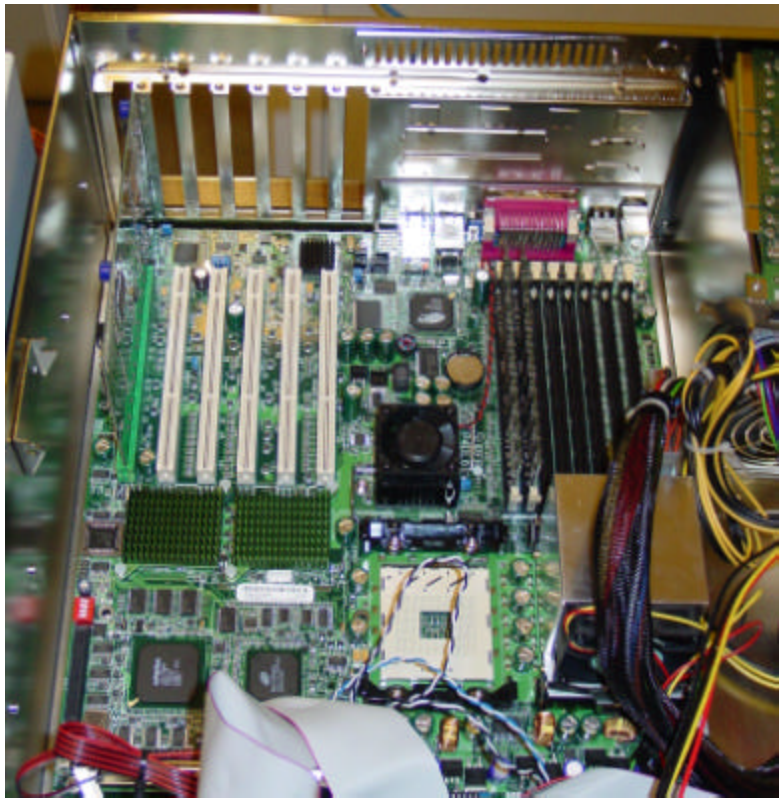


- 200 MByte/sec in and out
- Event buffering and assembly
  - Up to 200 events in queue
- Communication
  - Event Manager
  - Filter CPUs (data on demand)
  - Monitor (full access)

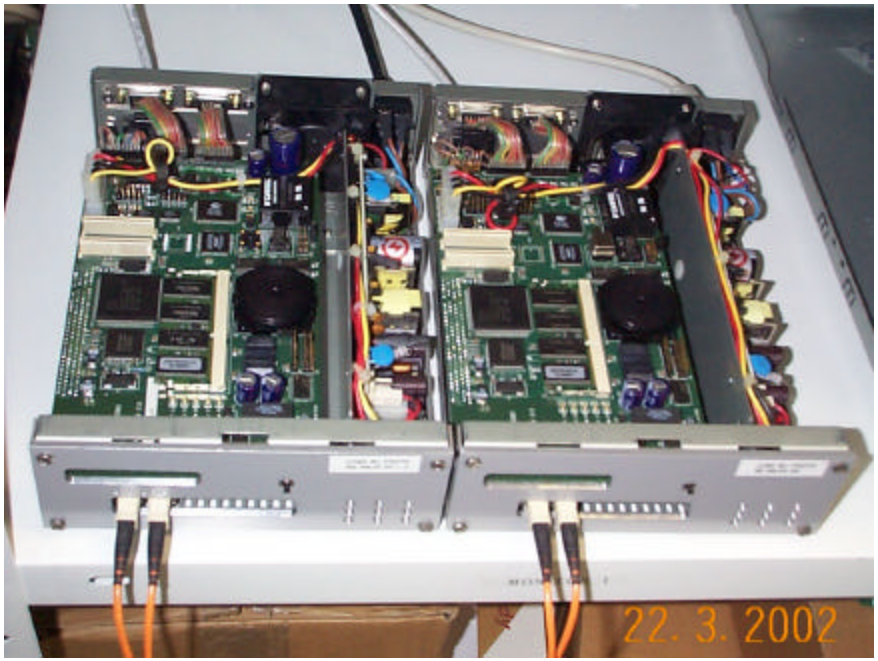
PC and PPC Prototypes

Pursued Two Prototypes

PC

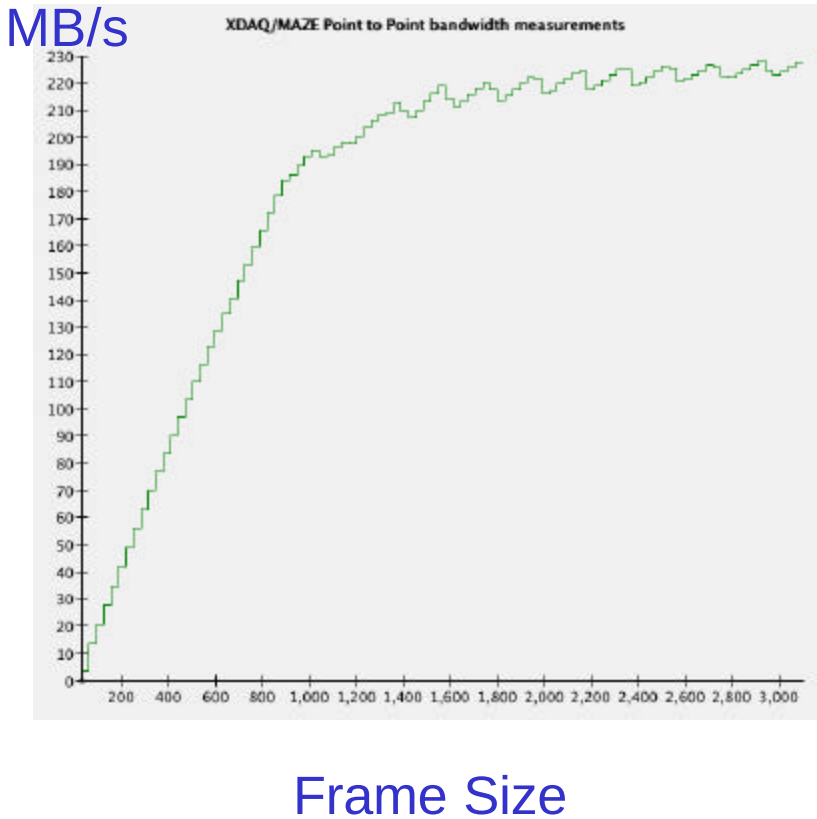


PPC



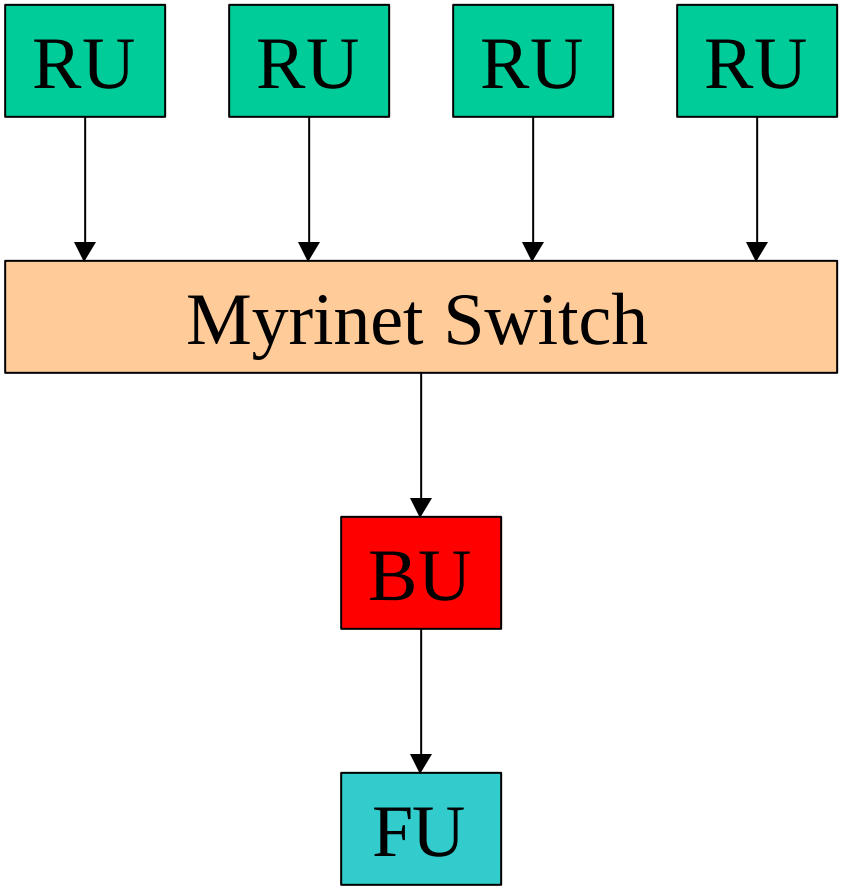
XDAQ/MAZE preliminary tests

- On PC linux, PCI 64/66 Myri2000 1x1, point to point
  - ~220Mb/s,2Kb
  - <11 usecs/32 bytes
- XDAQ software with standard Myrinet drivers for Linux



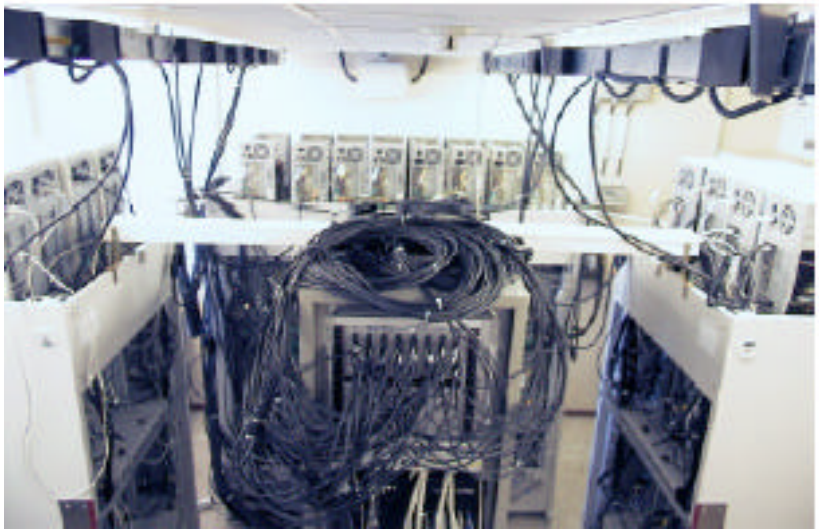
XDAQ/MAZE BU Tests with PCs

- Dell PowerEdge PC as BU
- PC linux, PCI 64/66
- Myri2000 interfaces
- XDAQ software on BU…
- Event building
- FU measures throughput
- 194 MB/s for 8 kB fragments.
- **204 MB/s** for16 kB fragments.



Meet CMS Requirements with last year's hardware

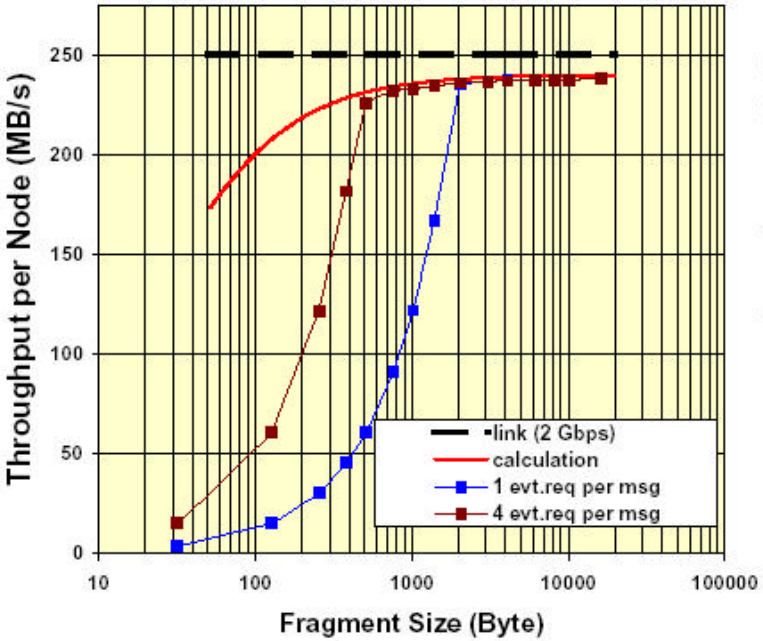
EB with Request Aggregation



Event Builder performance also meets requirements with last years hardware.

Our plans are very realistic now.

32x32 noBM perf. - request aggregation



- fixed size event fragments
- vary # fragm. req. per message (BU to RU)

- for small fragment sizes: limited by req. message handling
- crossover point: corresponds to 105 kHz msg rate for 1 req/msg: about 2kB
  4 1/2 kB

XDAQ

- A lean cross-platform toolkit for building distributed data acquisition systems
  - Goal 1: small additional processing latency
  - Goal 2: high configurability and adaptability to embedded environments
- Based on the Intelligent IO specification
  - Event-driven processing model
  - OS and hardware platform independent
- Full integration of the W3C SOAP/XML standard
  - Remote configuration and control via HTTP
  - Access to remote HTTP servers and services (e.g. logging)

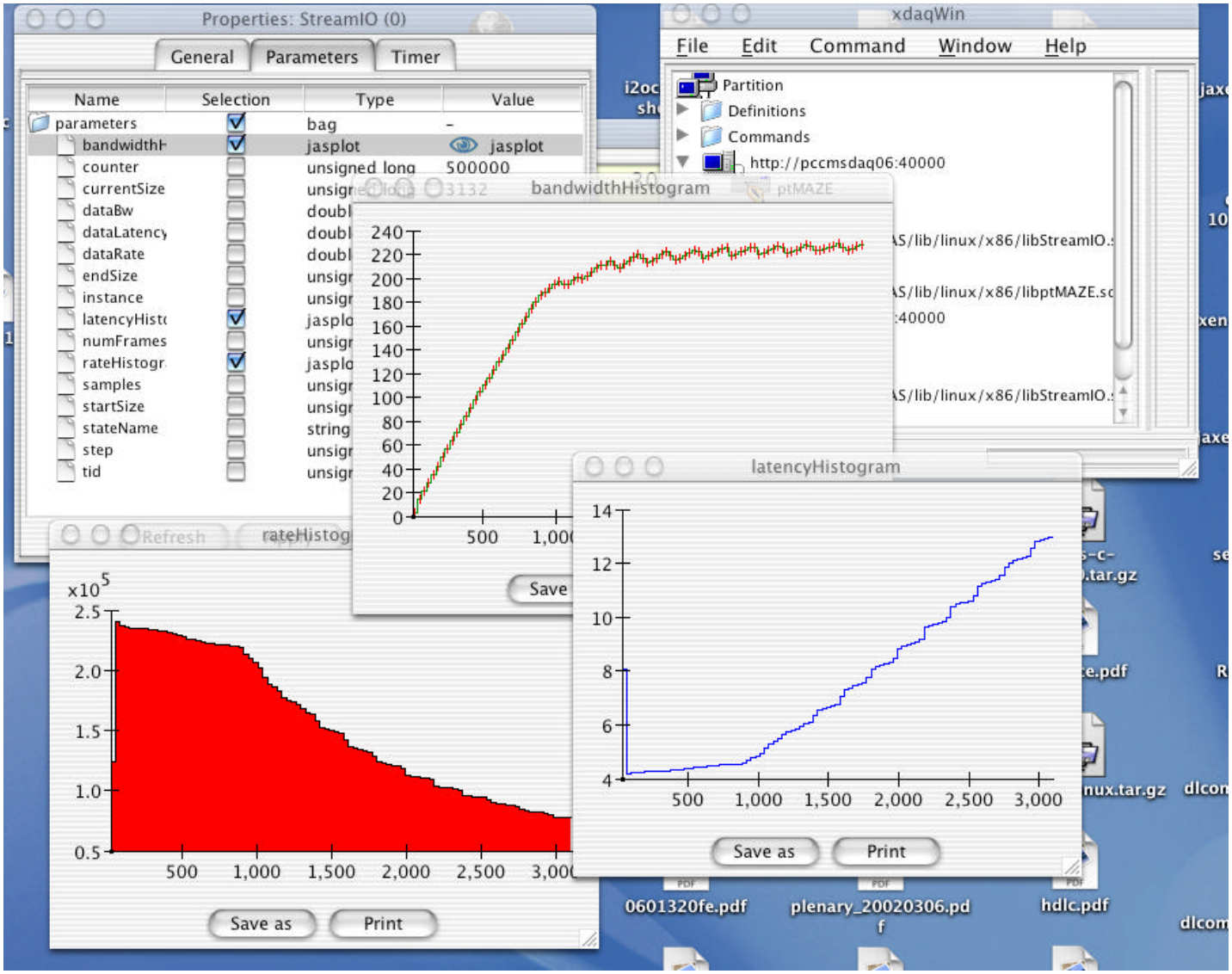
XDAQ Software Process

- Mil498 standard engineering environment
  - Focus on interface requirements and design documentation
  - Systems requirements and design documentation for all subsystems
    - Builder unit, event builder communication, interface to filter farm
    - User support documentation
- Collaborative working environment

Applications and Usage

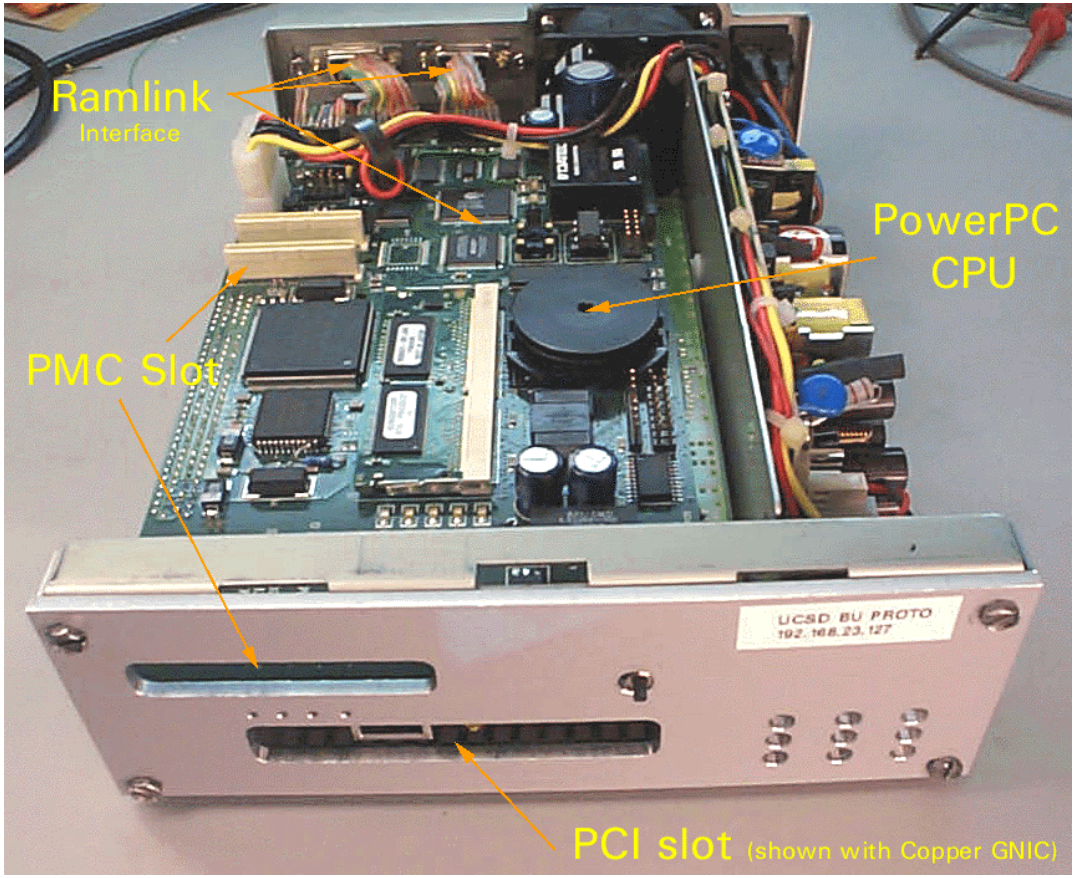
- All DAQ components are available as XDAQ applications
  - RU, EVM, BU
- Usage
  - Muon production in INFN Legnaro, Italy
  - Muon testbeam at CERN
  - Event Manager test stand in Fermilab, US
  - Tests for high-level trigger code integration (filter unit)
  - Integration with Myrinet2000 barrel shifter library from CERN

XDAQ/JAS support



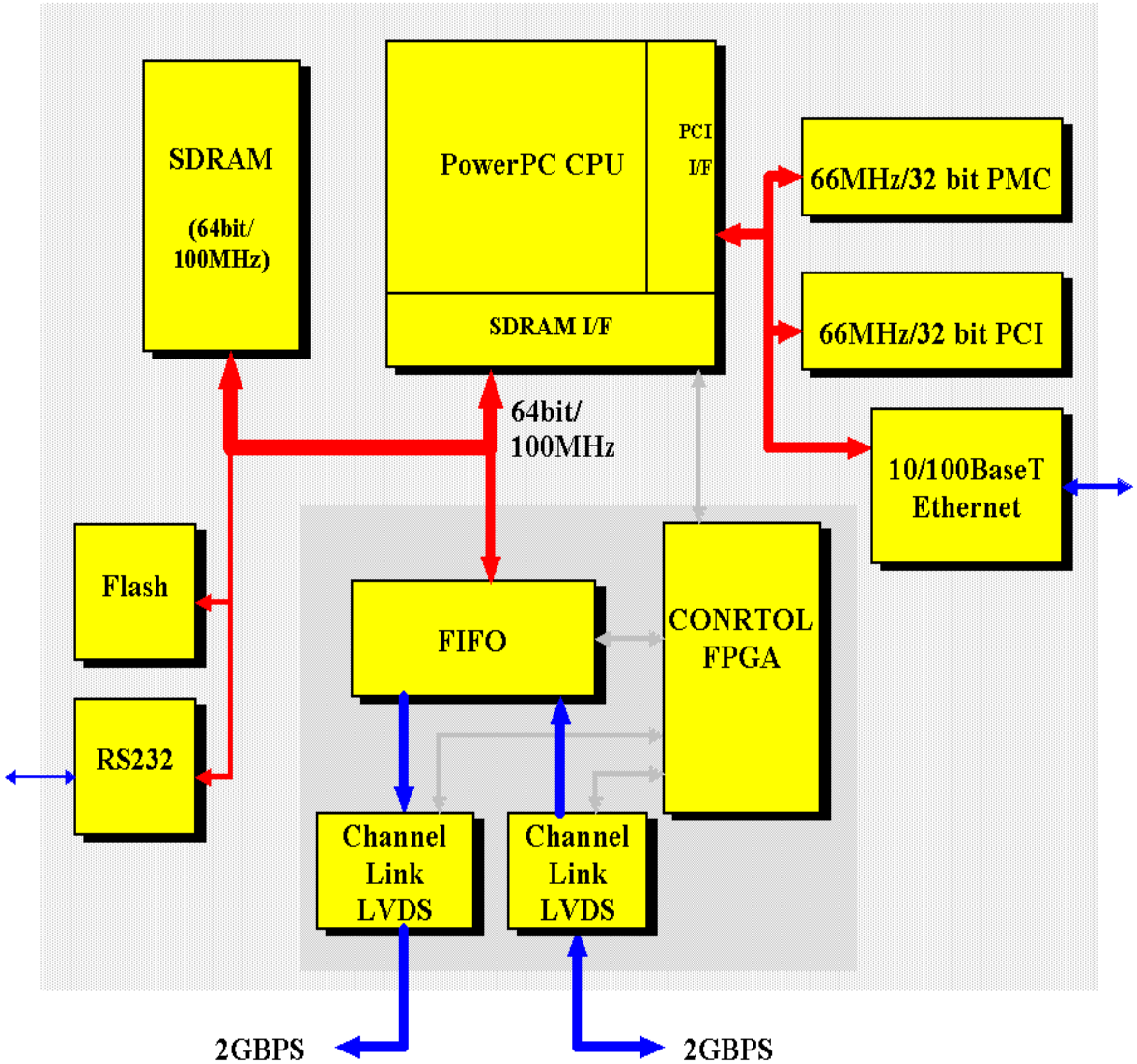
PPC BU Prototype

- Alternative to PC
  - Low Cost
  - VxWorks
- One PCI bus
  - Supports Gigabit Ethernet
  - Or Myrinet
- RamLink (LVDS) to connect systems
- Onboard Ethernet



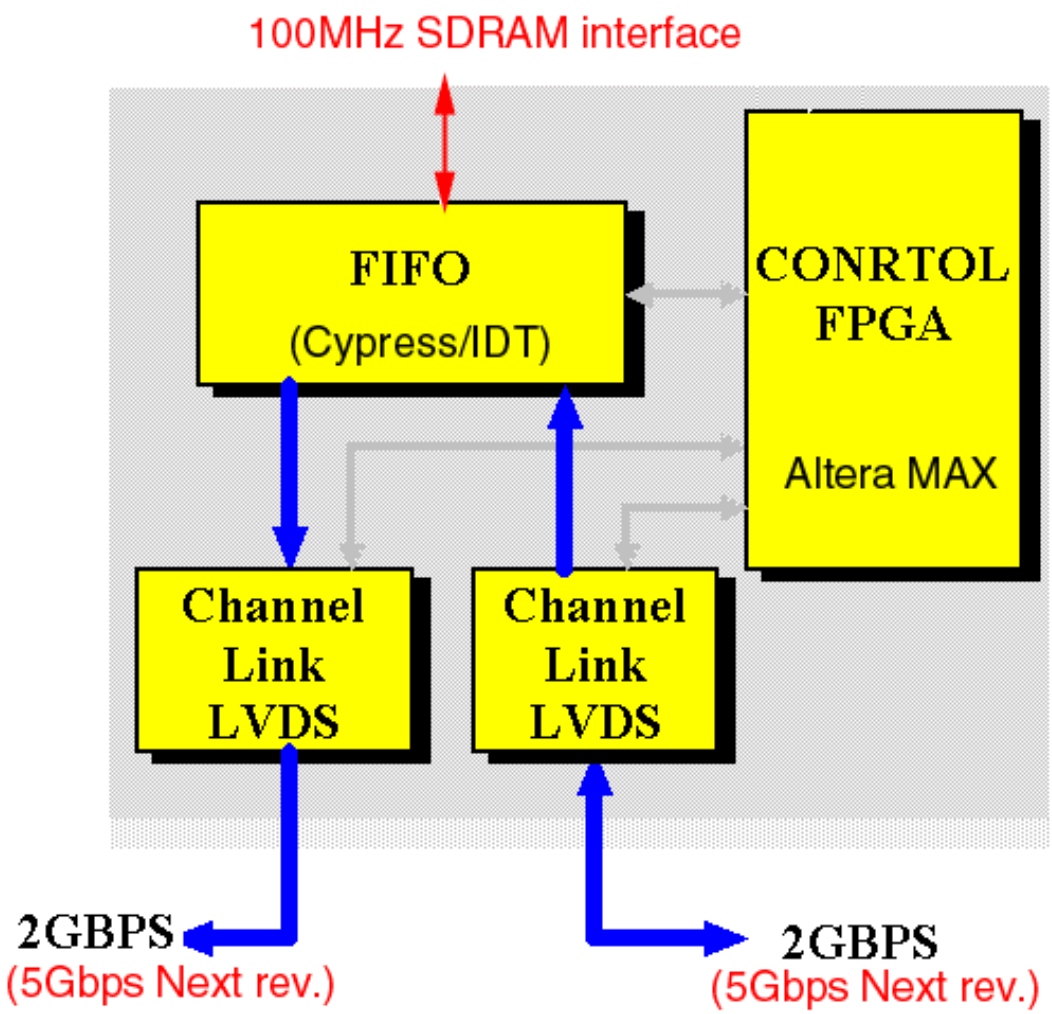
Current Architecture

- MPC8240 PowerPC CPU
- 66MHZ/32bit PCI interface (one PMC + one PCI interface)
- 128 MB, 100-133MHz SDRAM (up to 256Megabytes)
- On board 10/100BaseT Ethernet Interface
- Modular hardware expansion via "Ramlink" 2 Gbps serial interface



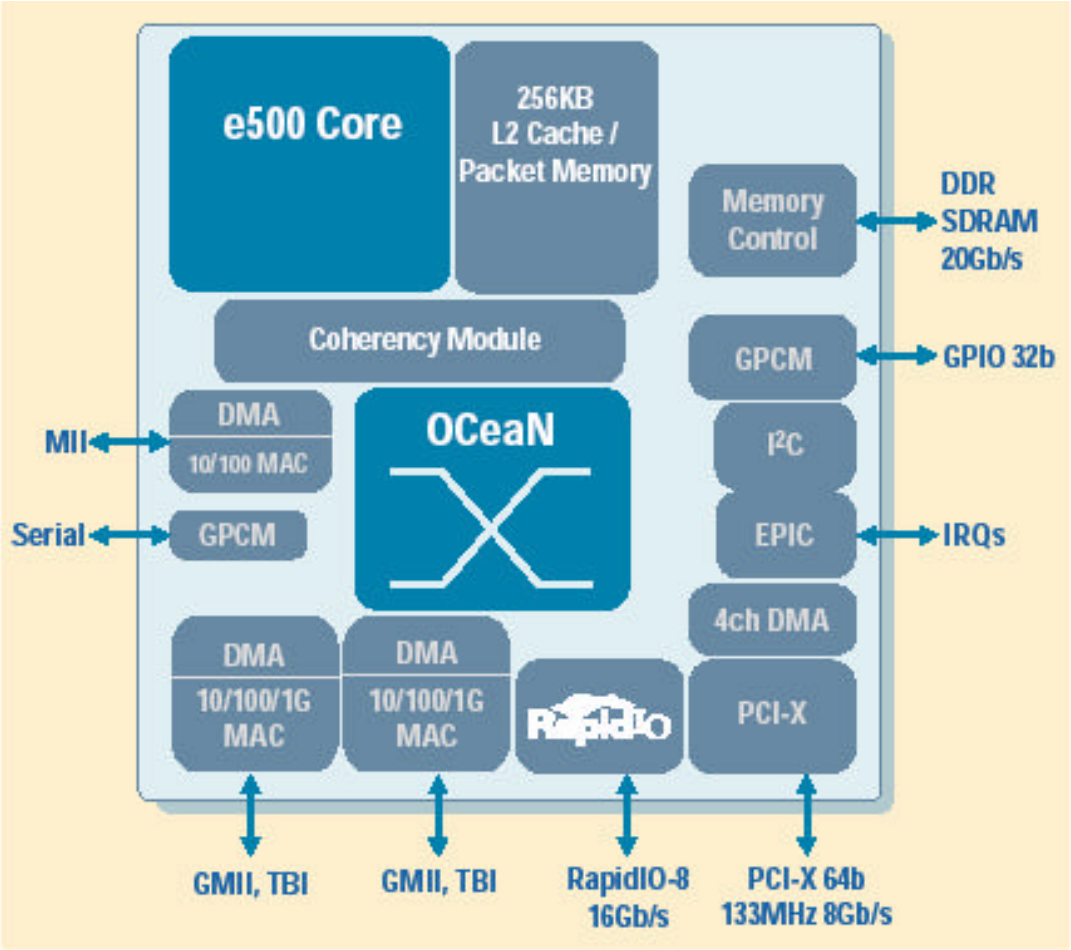
“RamLink” Interface

- SDRAM memory based interface. (currently 32bit,400Mbyte/sec)
- National Semiconductor LVDS Channel links at 600 MHz multiple serial. (up to 800 MHz available)
- Ring architecture allows data flow-through without CPU interference.
- VERY FAST, and SIMPLE
- Implemented in Verilog (models available from michael@tempestinc.com)



Next Generation PPC CPU

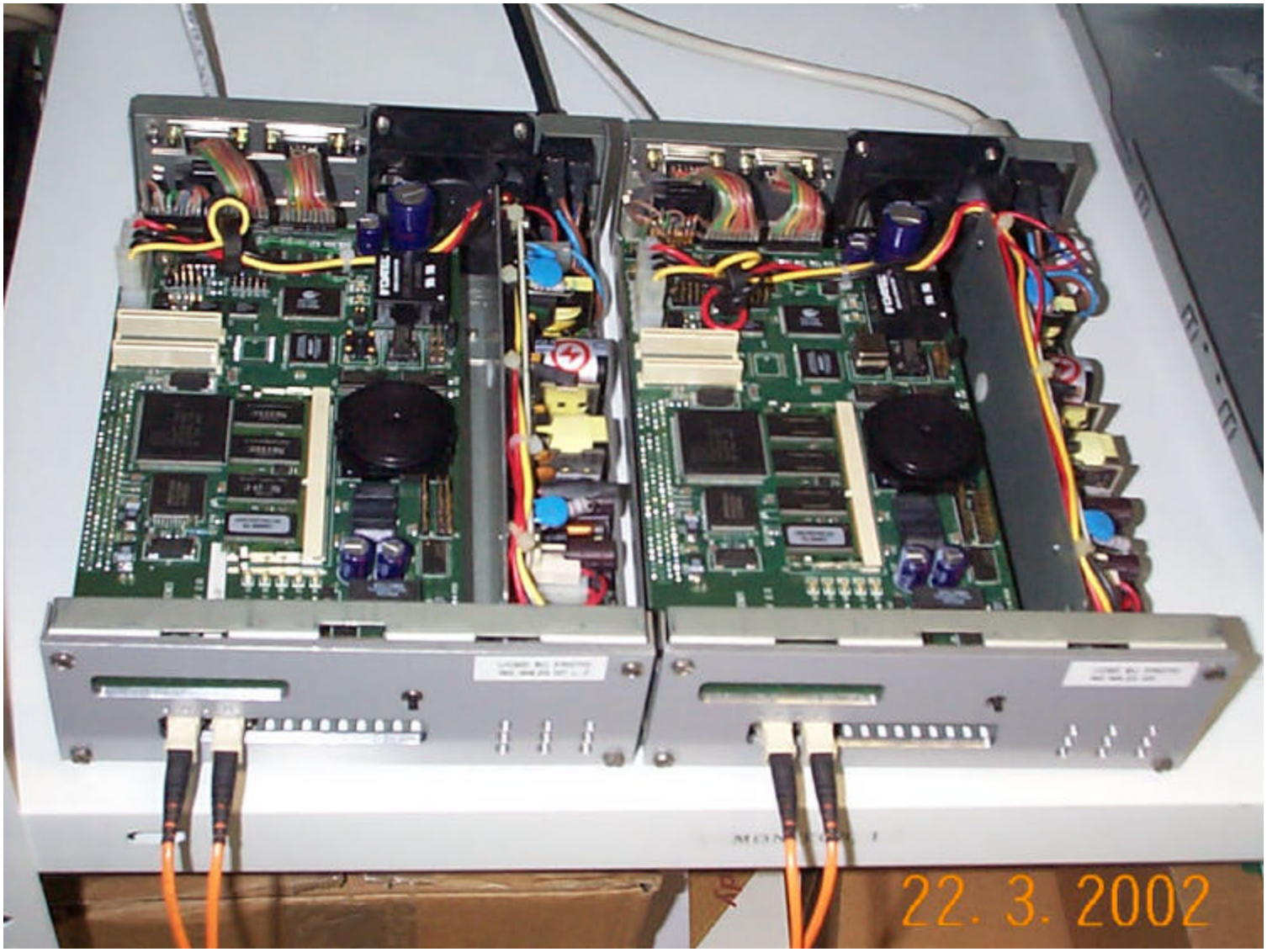
- 1GHz PowerPC
- 64b 133MHz PCI-X (1GB/s)
- 2x Gbit Ethernet + 1 standard Ethernet
- 64bit 333MHz DDR memory (2.4GB/s)
- 16 Gb/s Ramlink type interface on chip
- Integrated on Single Chip
- Due Q3 2002
- We can try this with evaluation board.



MPC8540 Block Diagram

http://e-www.motorola.com/brdata/PDFDB/docs/MPC8540FACT.pdf

Geminis in the test…



XDAQ on PPC Prototype

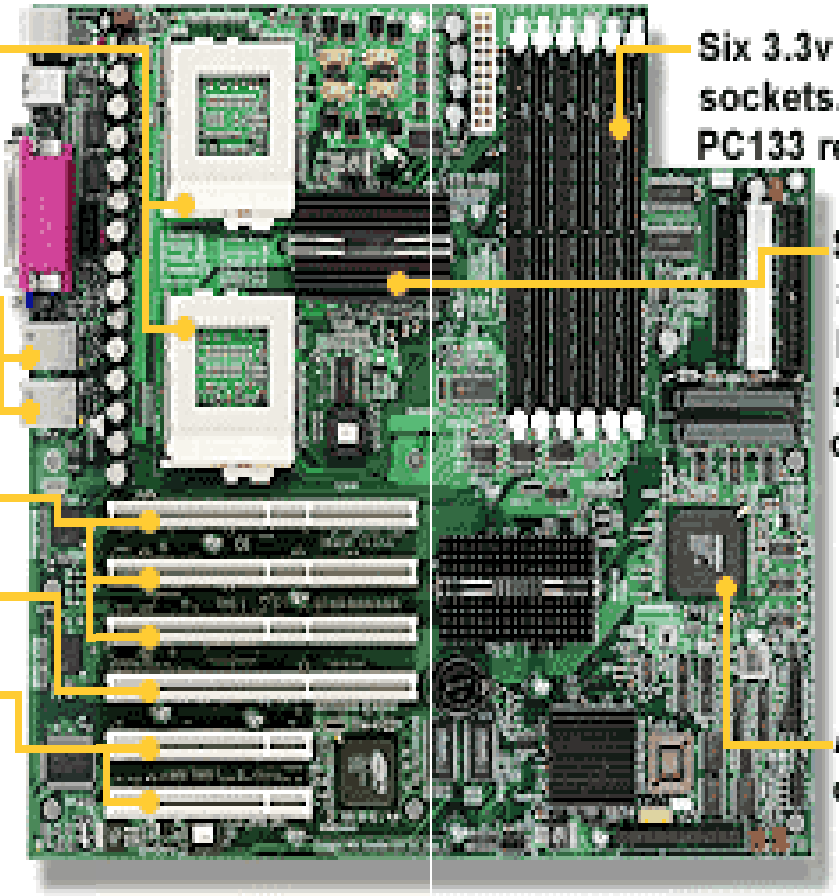
- XDAQ runs on custom Gemini system
  - Raw Ethernet transfers tested
- Creating support for Gemini in the XDAQ framework was not a lot of work
  - Some problem from 24 bit address in local branch instructions: (memory too large)

Industry Improves PCs for Us

- Increased Front Side Bus speeds.
- New chipsets with more, faster PCI busses.
- Onboard Gigabit Ethernet.
- Inexpensive high end Server motherboards.
- Cluster technology improvements.
- The usual exponential increase in
  - CPU power
  - Storage capacity
  - Memory size

How long will this go on without consumer application requiring it?

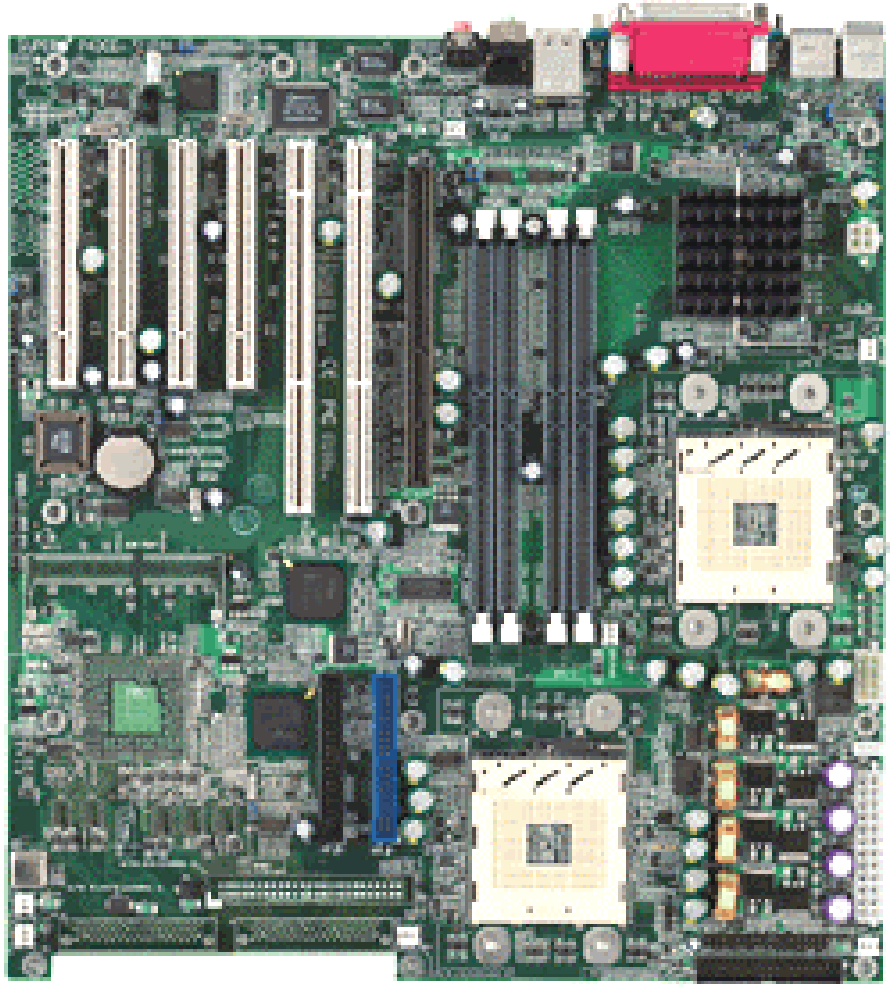
Tyan Thunder HE sl-T



PIII, Serverset III HE-SL3 chipset, 2X 64/66 PCI + 32/33 PCI

SuperMicro P4DCE+

- Dual P4 Xeon processor (1.4 - 2.2 GHz or faster)
- Intel 860 Chipset
- 64/66 PCI bus with 2 slots
- 32/33 PCI bus with 4 slots
- RamBus memory gives good performance



PCI bus performance test

- Tested individual PCI bus performance
- Used Franz Meyer's Myrinet test software to run DMA test.
- **Tyan Thunder (Serverworks HE-SL3 chipset)**
  - Individual 64/66 bus performance:

    330 MBps DMA read

    380 MBps write
- **SuperMicro P4DCE+ (Intel 860 chipset)**
  - 64/66 bus performance:

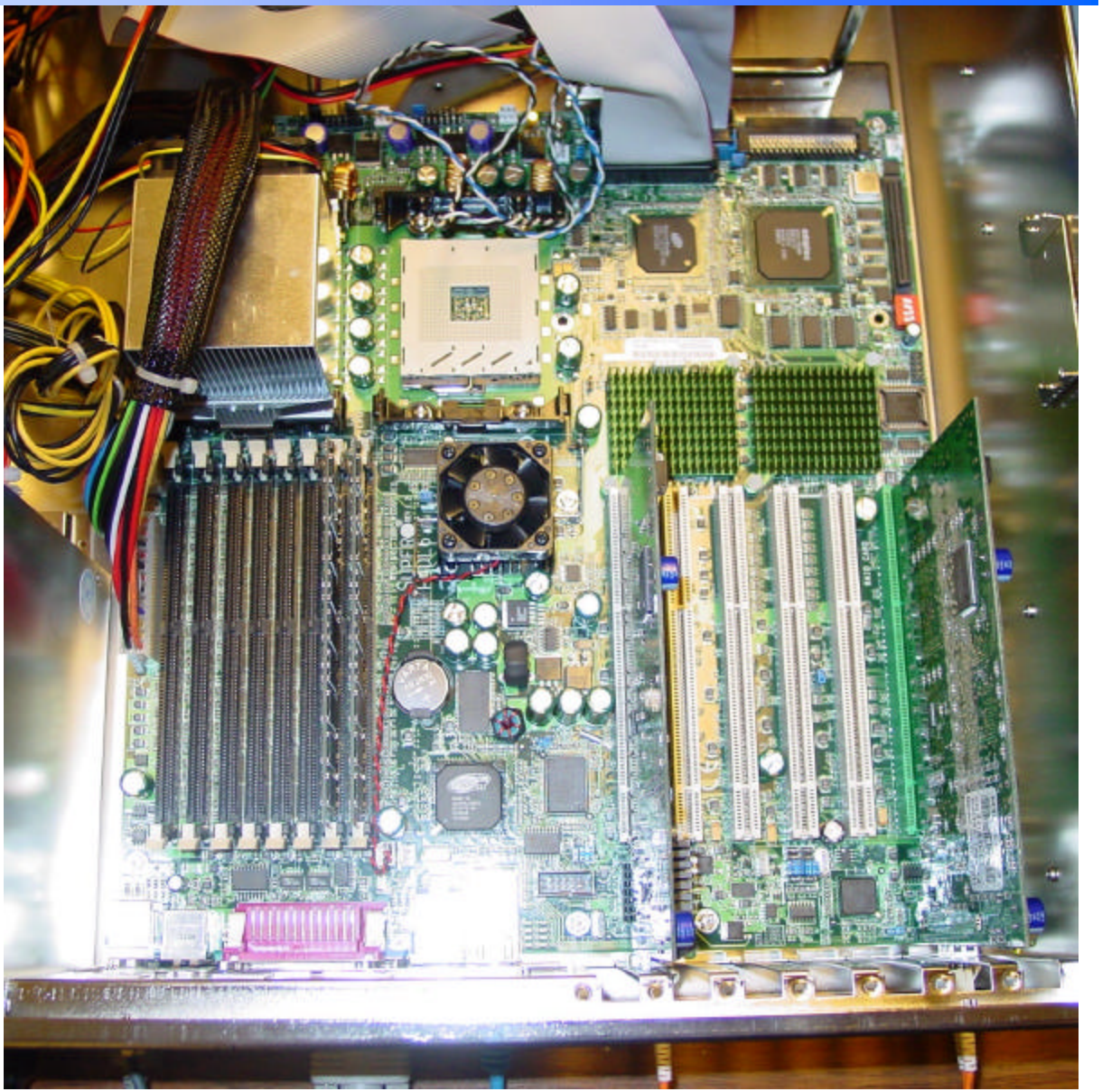
    150 MBps DMA read

    300 MBps write

    significantly worse than Dell PowerEdge
- These were somewhat disappointing
- Really interested in 2-bus performance.

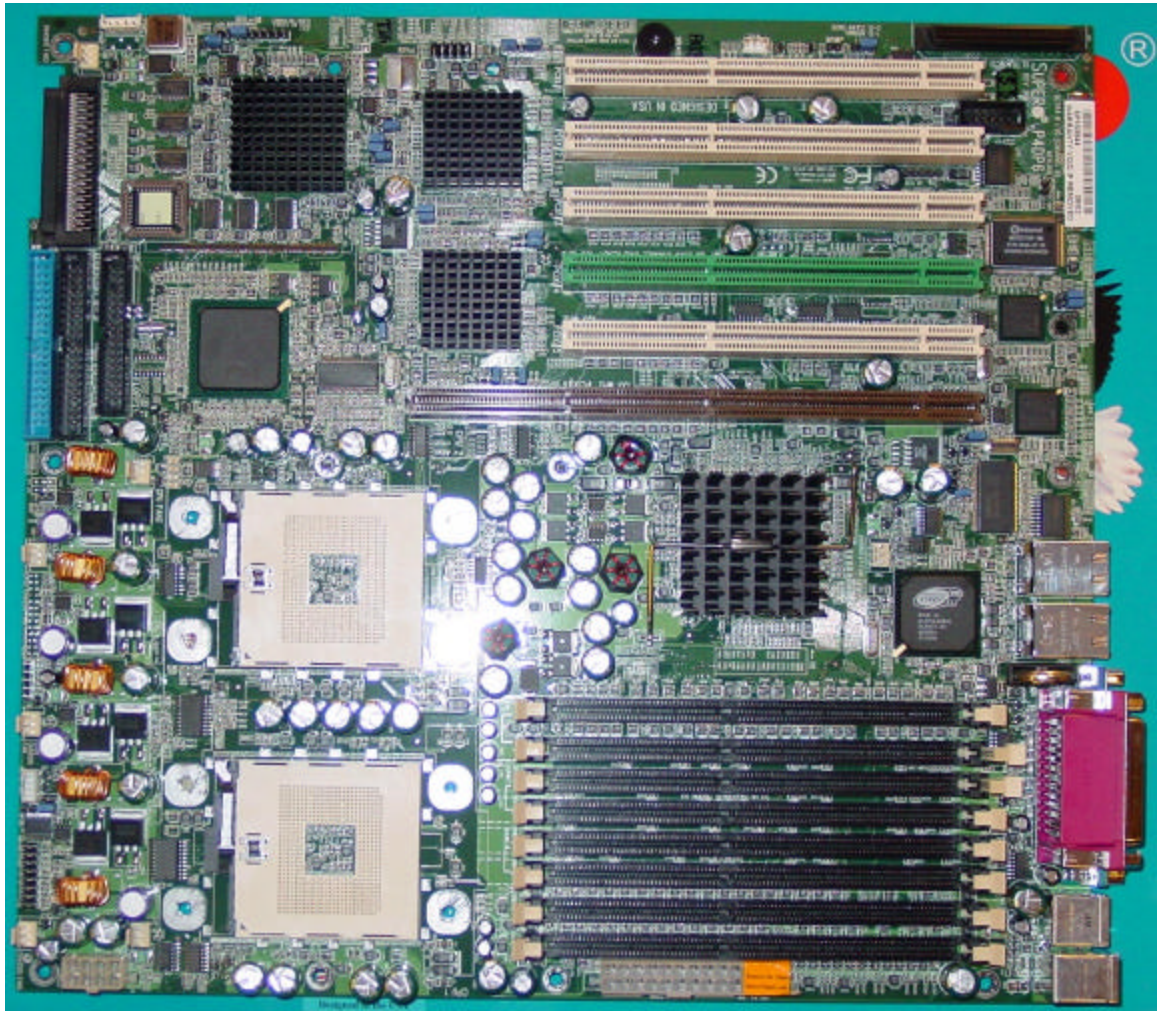
Serverworks Grand Challenge Chipset

- SuperMicro dual P4 motherboard P4DP6
- Intel E7500 chipset
- 4 independent 64 bit PCI buses
  - 33, 66, 100 or 133 MHz select
  - 6 slots
- Onboard GEth. and fast Eth.
- Interleaved DDR memory



New Intel Chipset

- SuperMicro dual P4 motherboard P4DP6
- Intel E7500 chipset
- 4 independent 64 bit PCI buses
  - 33, 66, 100 or 133 MHz select
  - 6 slots
- Onboard GEth. and fast Eth.
- Interleaved DDR memory



New PCI bus performance test

- Tested 2 PCI busses simultaneously
- Used Franz Meyer's Myrinet test software to run DMA test.
- **SuperMicro P4DL6(Serverworks HE-SL3 chipset)**
  - Two 64/66 bus performance:

    330 MBps read + 520 MBps write
- **SuperMicro P4DP6 (**Intel E7500 **chipset)**
  - Two 64/66 bus performance:

    443 MBps read + 512 MBps write
  - **First good Intel chipset in some time**
- (Memory performance worse than RamBus.)
- **What would we get at 133 MHz?**

BU Cost Estimate (PC)

Component	Cost ($)	No.
ServerWorks Dual P4 Motherboard (GC 4 133MHz PCIX + Geth. + Eth.)	680	1
P4 CPU	250	2
Memory (256 MB Reg. ECC DDR)	56	2
Rack-mount case with dual hot swappable power supplies	600	1
80 GB Disk	84	1
Integration and warranty	200	1
Gigabit Ethernet NIC (2nd Geth. NIC)	125	1
Myrinet 2000	1015	1
Total	3366	

The Future

- Factor of 2 speedup of Myrinet (5 Gbit/s) soon
- Gigabit Ethernet has made it to the desktop
  - Expect 2 onboard links on good motherboards
- PCIX is available on server motherboards
  - GEthernet cards available
  - Others will follow
- 10 Gbit Ethernet single chip interface announced
- PCI Express = 3GIO: 10X higher I/O w/o software impact
  - Improvement due to serial link (no clocked bus)
  - Due second half 2003
  - RapidIO (similar product) for PPC… due Q3 2002
- Infiniband: connect large clusters
  - future server market
  - 10-30 Gbit/s
  - Available now or soon (but not all the software)
  - Intel drops out of Silicon for Infiniband

Filter Farm

- New aspect of US project.
- Low cost commodity computers.
  - CPU intensive task: Moore's Law improvement
  - GEthernet now very low cost BU-FU interface
    - On motherboard
    - $60 NIC will do
    - $600 8-port GEthernet switches now available
  - Minimum memory on systems 512 MB is probably more than we need.
- Significant US expertise in building systems and clusters.
- Many studies have be done at CERN.

Filter Farm

- Ultimately about 4000 CPUs
  - 8 DAQ segments
  - Subfarms of about 100 CPU ⇒ 5 subfarms per segment
  - Hopefully many CPUs per “box”: 2-100
- Subfarm manager: monitor, configuration,control, download
- Filter Farm networks move data and messages.
- 10 KHz Level-1 rate ⇒ 25 Hz per CPU
- Level-2 decision can be based on fraction of data reducing network traffic significantly.
- As FF evolves, mixed architectures are likely.

Prototype Tier2 Center at UCSD/SDSC

- Dual CPU PIII compute nodes in 2U rack-mount cases.
  - Front end manages cluster using ROCKS software.
  - RAID arrays.
- Will upgrade to about 100 CPUs, approximately the size of a Filter Subfarm.
- Ian Fisk is US grid testbed facilities (hardware) group leader.

Cluster Management with Rocks

- National Partnership for Academic Computing (NPACI) has developed the Rocks cluster management system at SDSC.
  - Ensures an identical node configuration over large clusters
    - Linux kickstart configuration sections are stored modularly in XML.
  - Allows updating of all software components quickly and automatically
    - A kernel is installed on a partition of each system. This can be used to reinstall the OS for an entire cluster with one command
  - Enables the addition of new hardware or easy reorganization of existing hardware
    - New systems are recognized and installed using a common configuration file
    - Changing the functionality of a system involves changing the pointer to a configuration file and triggering a reinstall.

Configuration

- Currently one “frontend” node serves OS’s to computational nodes. Reinstalling a moderate sized cluster of 40 systems is a 15 minute exercise.
  - The frontend nodes can be replicated and arranged hierarchically so that the system scales.
  - The OS used on the trigger farm could be replaced in a similar length of time.

- Complete reinstallation may not be the desired mode for small trigger software updates.
  - The hierarchical frontend nodes could be used to propagate application software changes to the compute nodes.
  - Maintain ability to upgrade system components when necessary and 20 minutes are available, while being able to update application components in a matter of seconds.

FU Cost Estimate

Component	Cost ($)	No.
Dual P4 Motherboard (Geth. + Eth.) SM P4DPE	470	1
P4 CPU (around optimum price performance)	200	2
Memory (512 MB Reg. ECC DDR)	112	2
1U Rack-mount case with power supply	400	1
80 GB Disk	84	1
Integration and warranty	100	1
Total	1678	

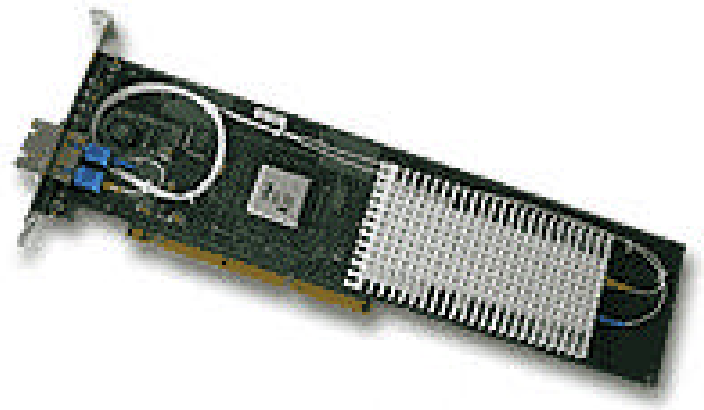
Right now, you can get the same performance a few hundred dollars cheaper with Athlons.

Quad (or larger) systems are coming (with cheap processors?).

Summary

- Builder Unit and Filter Farm procurement is not a problem in new baseline of US DAQ project.
  - Working on BU for years
  - Filter Farm is similar to other computing clusters
  - Continue to work with strong CERN group
- BU requirments can be met with today's hardware
  - XDAQ software is crucial to use of commodity hardware
- FU needs can be met with standard PC's (+ Moore's law) and today's inexpensive networking.
- Developments in technologies to move data will likely change/improve our DAQ systems before procurement.

Single Chip 10Gb Adapter



Intel® PRO/10GbE LR Server Adapter
Is now sampling. It's based on Intel® 82597EX 10 Gigabit Ethernet Controller and the 10GbE IEEE

802.3ae specification.Intel® TXN17401 Optical Transceiver
Intel designed the TXN17401 for 10-Gigabit Ethernet equipment such as enterprise switches and routers as well as core router applications. The transceiver provides an interface between the photonic physical layer and the electrical section layer. It is assembled in an industry Multi-Source Agreement (MSA) and comprises a small cross section for high port density. The MSA supports the proposed IEEE 10 Gigabit Ethernet interoperability standard, which specifies a uniform form factor, size, connector type and electrical pin-out. The transceiver has hot-plug capability for flexibility in system design and production and uses Intel uncooled optics for exceptional performance at temperature.
10GbE LAN Applications

LAN backbone infrastructure
- Connections inside server farms
- Connecting multiple site LANs

PCI Express

Overview

3RD GENERATION I/O

PCI Express

Third Generation PCI

(Also known as NGIO or 3GIO)

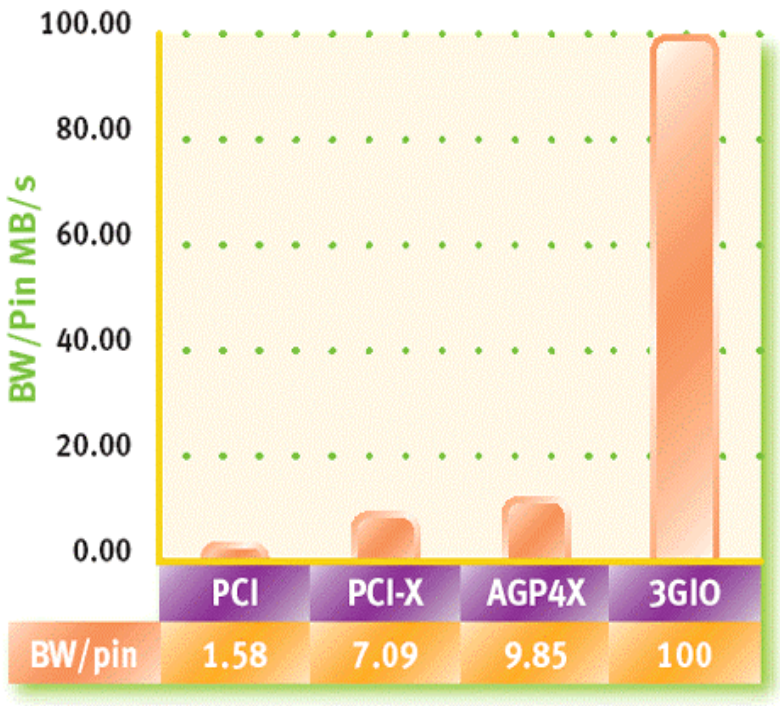
New electrical interface

- Replacement of existing PCI (not enhancement)
- 2.5-Gbit/second serial I/O scheme
  - Serial link gives higher performance.
- 1-40 bits wide (to 100 Gb/sec)
- Software provides backward compatibility
- New form factors
- Tentative timeline is Second half 2003

3RD GENERATION I/O

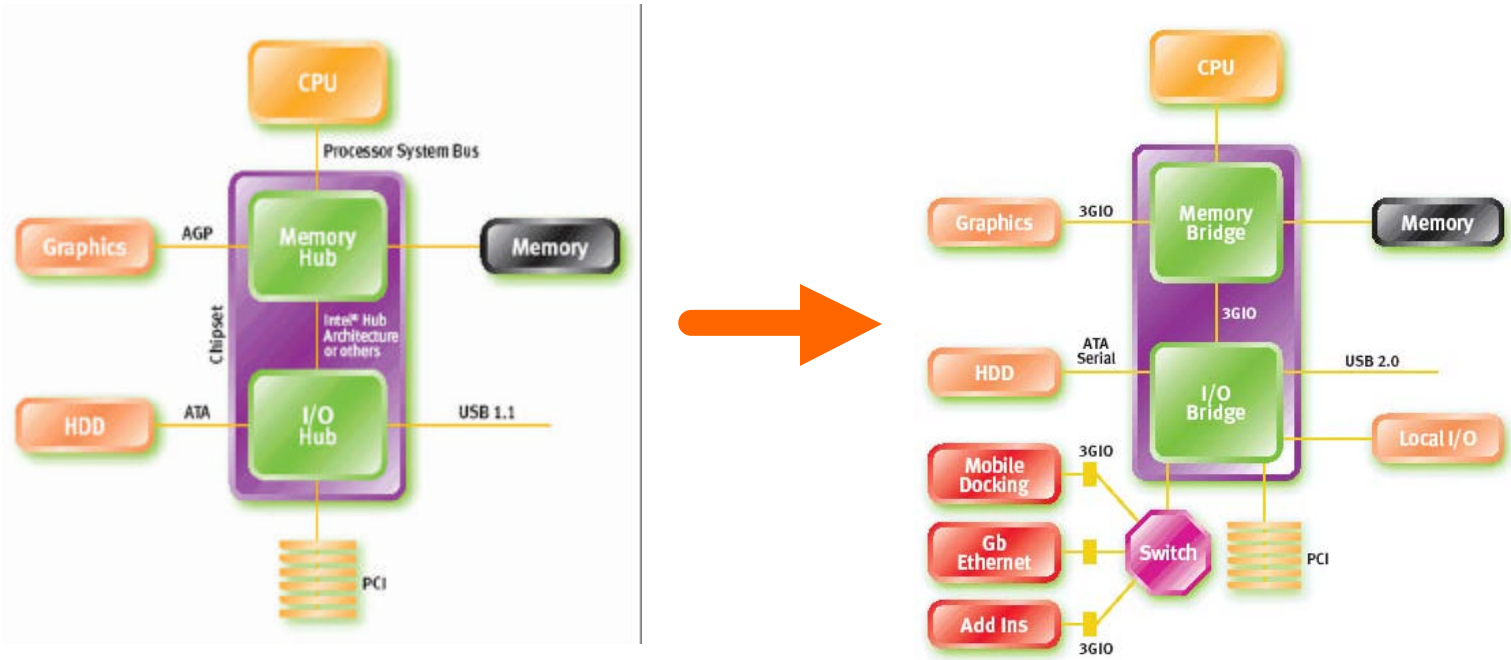
Comparison

- PCI 32bit - 33MHz
- PCI-X 64bit – 133MHz
- AGP4 32bit – 533MHz
- 3GIO 40bit – 2.5GHz



Source: Intel Corporation

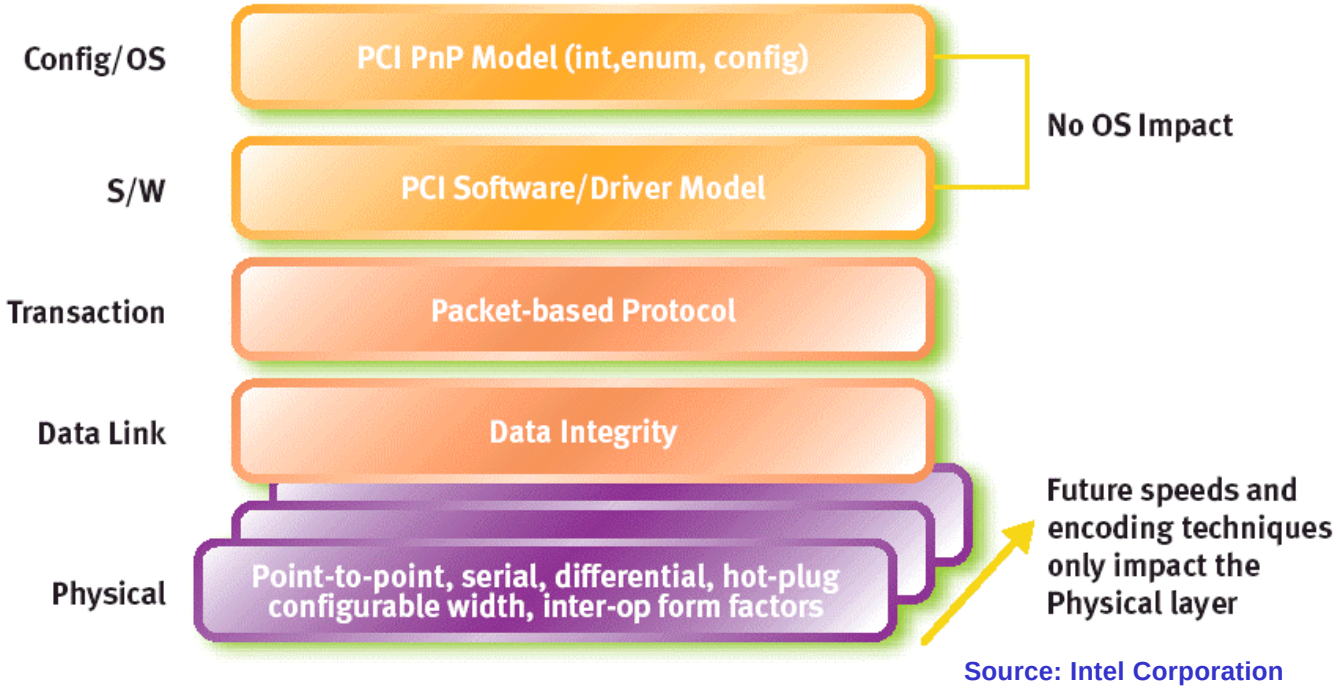
PCI Express Architecture



Initially PCI Express takes over High bandwidth functions in the next generation PC (in particular Graphics and High speed networking) co-existing with the PCI, with the aim to eventually displace PCI

Source: Intel Corporation

Software



Source: Intel Corporation

No new software required at Operation System or Application layers - transparent

PCI Express News

PCI Express may alter PC interconnect scape

By Rick Merritt

SAN MATEO, CALIF. – As engineers start preparations for the arrival of PCI Express (formerly 3GIO), predictions vary on the extent of change that the fast serial interconnect will bring to the architectural landscape of the PC.

While most see PCI Express as a straight PCI replacement, they also voice concerns about a fragmenting market and additional design costs. But a few say it may enable a repartitioning of the PC, possibly creating silicon opportunities–albeit largely for existing players.

PCI Express is expected to hit the PC market as a 2.5-gigabit/second interface that will initially be used on chip sets, graphics processors and gigabit Ethernet controllers. The new link is expected to sample late next year.

The initial expectations of many systems and chip makers are that the link will add throughput speed and development cost to their products, but some are beginning to dig deeper.

"If the industry can standardize on one high-speed interconnect, there are lots of interesting

the plumbing for Nvidia to integrate Gigabit Ethernet, USB 2.0 and 1394 interconnects into a future version of that chip, which the company calls its media and communications controller.

Others envision an even more radical move, in which the traditional PC north and south bridge chips are eliminated to create a system that has the PCI Express switch at its center. "A pure-switched architecture has a lot of advantages. There's a lot of architectural discussions about that now," said a systems designer who asked not to be named.

"I believe this technology allows a variety of partitioning options. Different companies and market segments will adopt different partitionings," said Ajay Bhatt, who directs the research lab at Intel Corp. Bhatt said, however, that he thinks PCI Express switches will see limited use in workstations and servers, and that they are not likely to find a spot in mainstream PCs.

Wherever the PCI Express switches are placed, they should have peer-to-peer communications capabilities, said Roger Tipley, chairman of the PCI

with PCI parts, he said.

PCI Express will be deployed in an evolutionary way in first-generation systems in 2004, linking graphics and Gigabit Ethernet to new chip sets, said Bob Gregory, Intel's top spokesman for the interconnect.

"You will see an evolution from first-generation platforms that add the capability to second-generation [ones] that optimize around it to third-generation systems that fully embrace it. The technology will support different ways to repartition the platform. Everyone in the industry is starting to work through how you might partition a second-generation platform," Gregory said.

Tony Tamasi of Nvidia: Design costs 'will be painful.'

Intel is expected to use PCI Express to link its traditional north and south bridge chips in its next-generation chip sets. Bhatt's group, which designed PCI Express, also designed the proprietary Hub Link Intel uses to connect these bridges today.

would be a pretty fundamental shift in how the PC is put together," Tamasi said.

Dean McCarron, analyst with Mercury Research (Scottsdale, Ariz.), said he doubts PCI Express will spark major architectural changes in the PC. That's because the existing bridge system is already cost-effective and flexible enough for designers. "I don't think you will see anything remarkably different, because what drives the PC architecture is not new technology, but cost," he said.

Indeed, some see the advent of new interconnects like PCI Express and HyperTransport from AMD as necessary but painful transitions that require additional engineering resources without a guarantee of additional revenue.

"The market is fragmenting and that makes it much more difficult," said Fred Leung, associate vice president for sales and marketing at chip set maker

The additional design costs for PCI Express "will be painful," said Nvidia's Tamasi, in part because the link is incompatible with the Accelerated Graphics Port used to link graphics to Intel-based chip sets. Nvidia expects to support both links on products for an uncertain time, doubling the circuit and physical design, validation and qualification times for the I/O portions of its products.

While graphics will still embrace PCI Express to get much-desired bandwidth boost, server designers are more reticent. The server group at Compaq–now part of Hewlett-Packard–has stated it plans to eschew PCI Express for PCI-X on all its servers.

"We don't see PCI Express getting into servers," said Raju Vegesna, founder of ServerWorks Inc., a sever chip set supplier that counts the former Compaq group as one of its biggest customers. "I'm sure some people will have it, but no one has convinced me of the benefits of PCI Express for servers," Vegesna added.

Intel's Bhatt countered that PCI Express brings ice-chrome

InfiniBand

Overview

Motivation for InfiniBand

INTEGRATION:

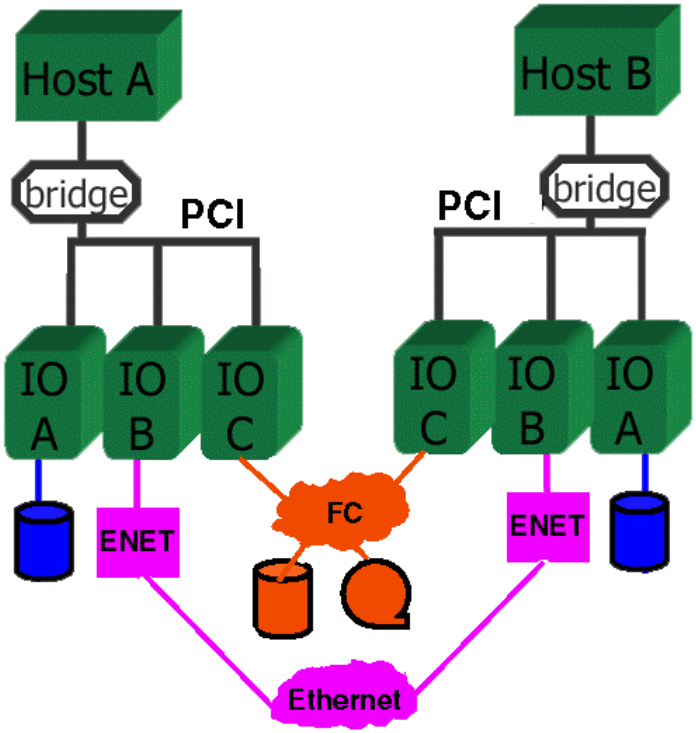
Single Integrated **interconnect for all traffic. A switch based point to point protocol.**

- **Inter-processor communication (e.g. Myrinet)**
- **Storage (e.g. SCSI/fiberchannel)**
- **Network (e.g. Gigabit Ethernet)**

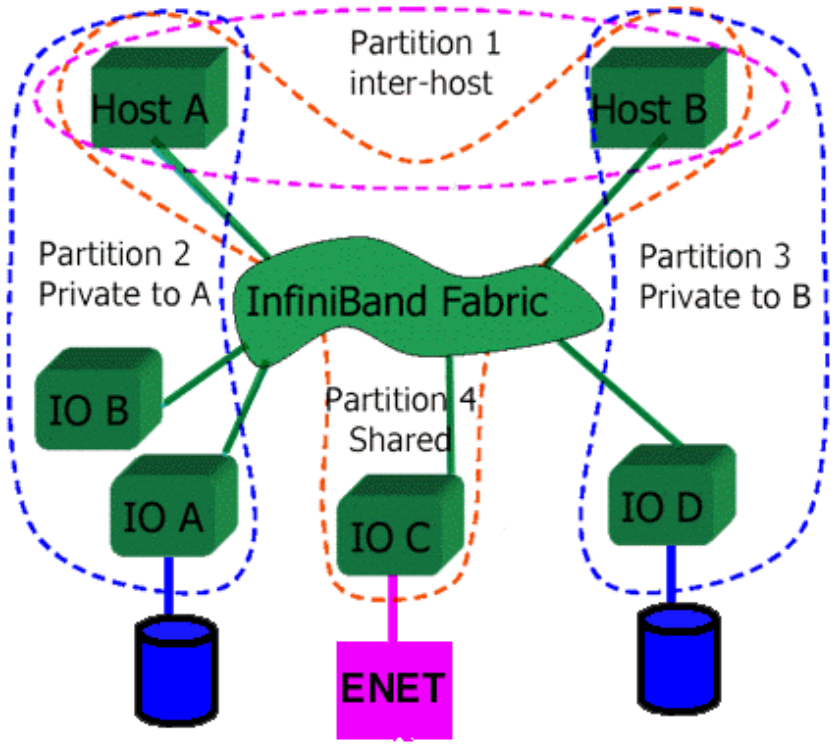
SCALABLE COMPUTING:

- **Scalability to 1000s of nodes copper or optical**
- **Low cost and high Bandwidth 2.5 Gbit/sec x1, x4, x12 (to 30 Gbit/sec per link) interconnects**
- **Reliable, high availability and serviceability**

InfiniBand Topology



Traditional PCI based approach



InfiniBand

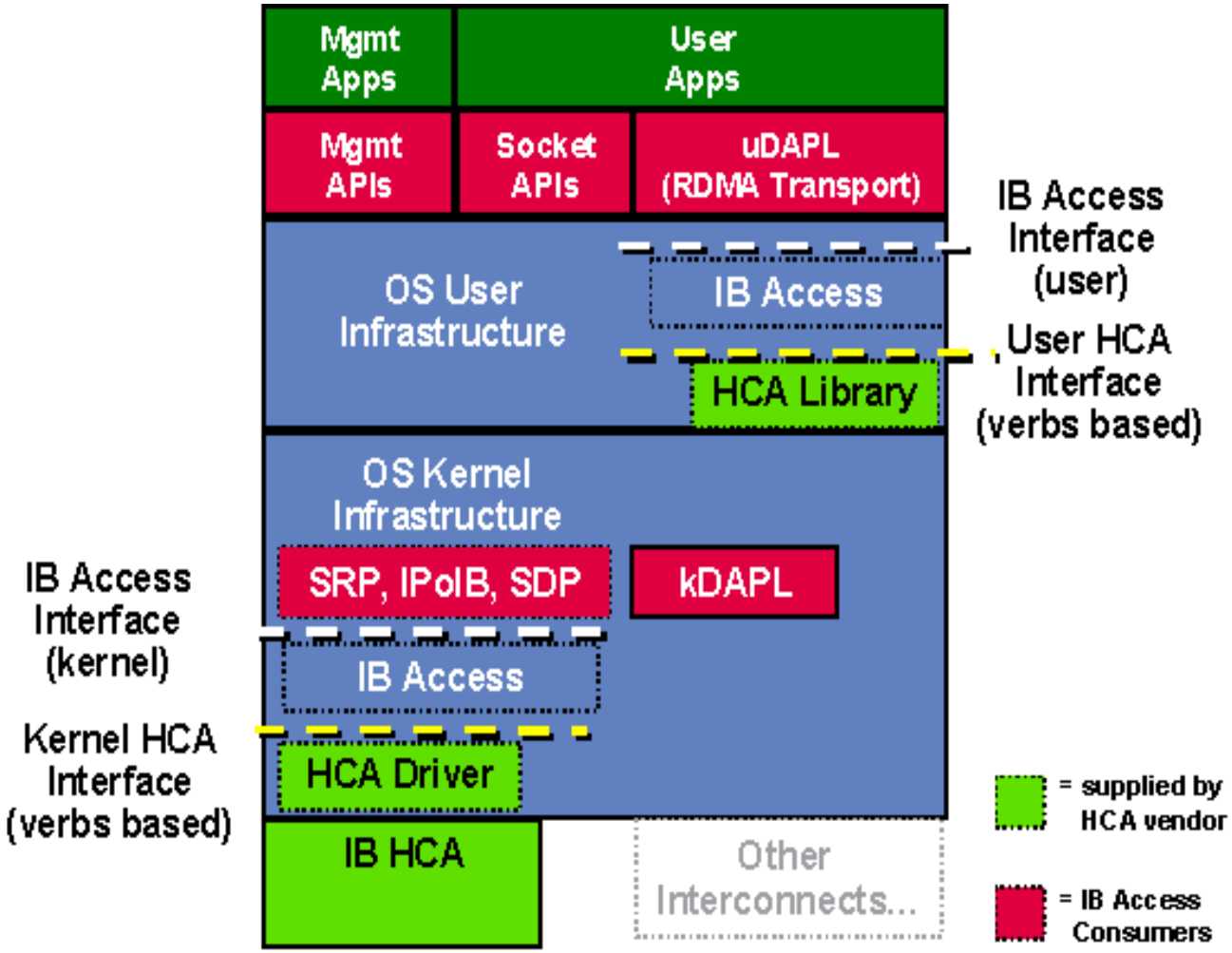
Infiniband SW overview

Linux support for the following Infiniband fabric

- Hardware drivers for fabric access
- IP Networking, Storage, and Shared memory
- Fabric management

http://infiniband.sourceforge.net/

Infiniband on Linux



Fabric Access

- **Hardware Drivers –**
  - **Host Channel Adapter (HCA)** similar to PCI bus master
- **Infrastructure Drivers**:
  - **InfiniBand Access –**exports the HCA implementations to higher-level software
  - **Fabric Boot –**fabric boot support for operating systems
- **Subnet Management –** Provides the basic Subnet Manager functionality. Subnet Management handles several areas related.

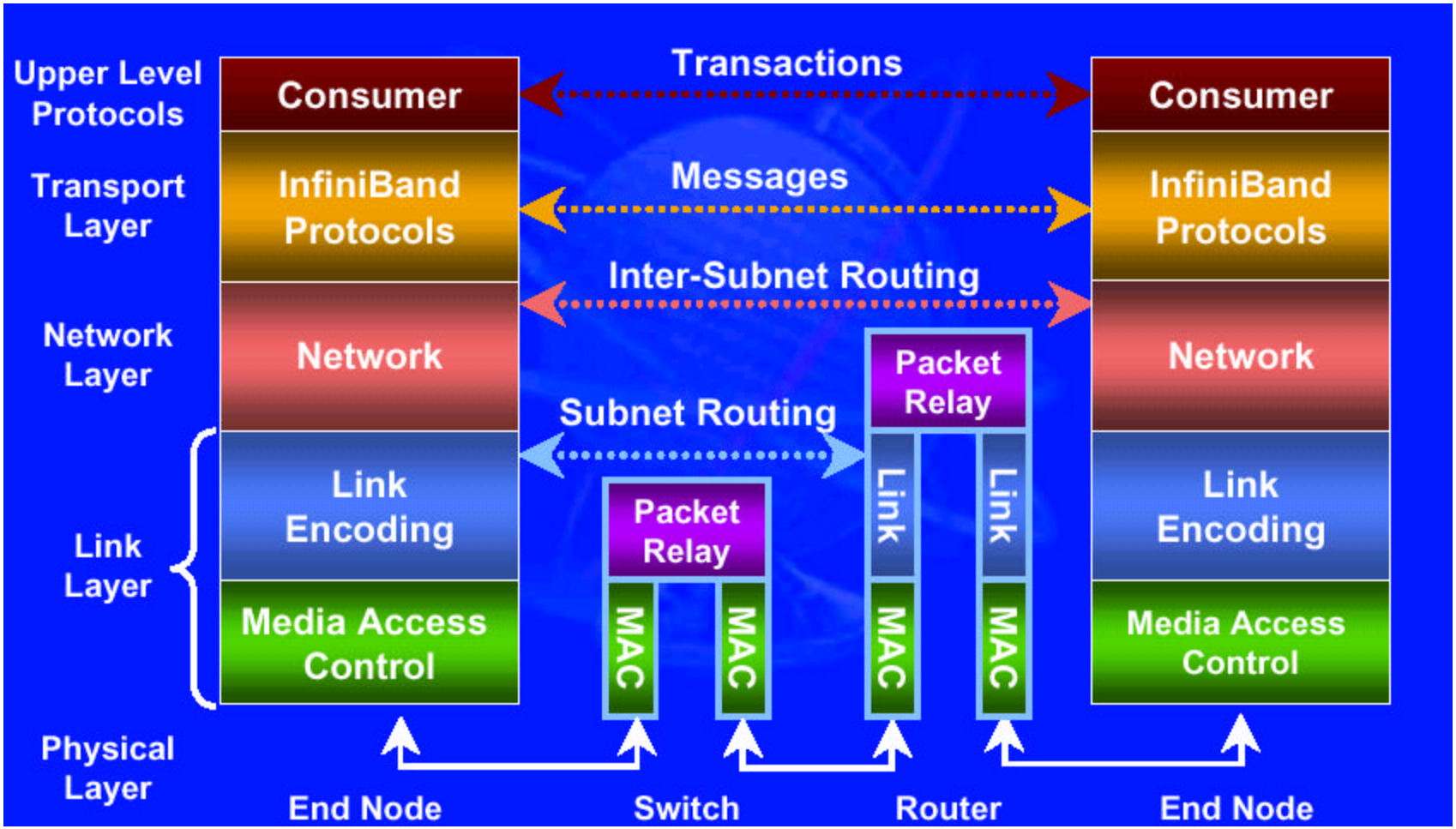
Infiniband Networking

- **Networking Drivers**:
  - **IPoIB –** IPoIB provides standardized IP encapsulation over InfiniBand fabrics
  - **SDP –** The Sockets Direct Protocol (SDP) is an InfiniBand specific networking protocol
- **VNIC –** The Virtual NIC (VNIC) driver enables a host on an InfiniBand fabric to access nodes on an external Ethernet without requiring an Ethernet NIC to be installed in that host.

Storage and Memory

- **Storage Drivers**:
  - **SRP –** The SCSI RDMA Protocol (SRP)
- **IPC Drivers**:
  - **kDAPL –**kernel Direct Access Provider Library. It is a high performance Remote Direct Access Memory (RDMA) API for the kernel..
  - **uDAPL –**It is a high performance Remote Direct Access Memory (RDMA) API for user-mode.

Protocol Stack



Source: IDF spring 2001, Mazin Yousif, Ph. D. Mazin

Hardware Available now

- Intel kit includes switch, 2 HCAs, cables, Driver kits for Win2K, Linux and VxWorks (US $10K)
- 32 port switch available now from InfiniSwitch
- Chips from several vendors available now
- InfiniBand storage devices (Seagate.)

Our Interest

- High performance, commodity Event Builder
  - (Onboard low cost) 4X Infiniband interfaces
  - Inexpensive switches (Myrinet for example)
- Large, scalable computing clusters
  - Probably not needed for Filter Farm application
  - Useful for data access offline

Infiniband News

Intel exits Infiniband silicon arena

By Anthony Cataldo

SAN MATEO, CALIF. – Intel Corp. has abandoned plans to provide Infiniband silicon and instead will rely on third-party host control adapters that will be tied to future Intel server chip sets using the I/O standard.

Intel, which helped to spearhead the box-to-box interconnect standard and was among the first companies to field prototype chips, said it wants to shift its engineering resources to other areas, such as the emerging PCI Express I/O standard (see related story, page 6), formerly known as 3GIO.

"The Infiniband ecosystem is maturing, and several vendors are planning to supply products," an Intel spokeswoman said. "Intel's strategy is to focus on developing interfaces like PCI Express for future chip sets and to ensure high bandwidth for these third-party host controller adapter components, which will ensure Infiniband connectivity on Intel-based platforms."

The spokeswoman said Intel plans to unveil a server chip set next year that will have hooks to Infiniband. She cited IBM Corp. and Mellanox Technologies Inc. as potential Infiniband silicon providers.

Heavyweights hit, too

Jonathan Eunice, an analyst with Illuminata Inc., said the economic downturn has forced Intel and others to reexamine their product lines. "In terms of the economic flattening, even companies like Intel and IBM feel constrained," he said.

Servers used in Web farms were one of Intel's targets for Infiniband, and Intel had fielded a 2.5-Gbit/second (or Infiniband 1x) chip that would serve that market. But after that market soured, the company was left behind by other Infiniband chip vendors that had developed higher-bandwidth 10-Gbit/second (Infiniband 4x) controllers for back-end data centers. "It would have required a new investment if they had gone back to do the 4x," Eunice said.

There's still sufficient demand for 1x speeds, but 4x is coming on strong. "Lots of people will be happy with 1x, but it looks like the market will require 10-Gbit-per-second performance as well," said Kevin Deierling, vice president of product marketing for Mellanox Technologies.

Intel's decision not to produce Infiniband chips does not reflect poorly on the market for the I/O standard, Deierling said. By 2005, he expects up to 4 million servers to be deployed that are Infiniband-enabled, which should mean more than a $1 billion market for Infiniband silicon.

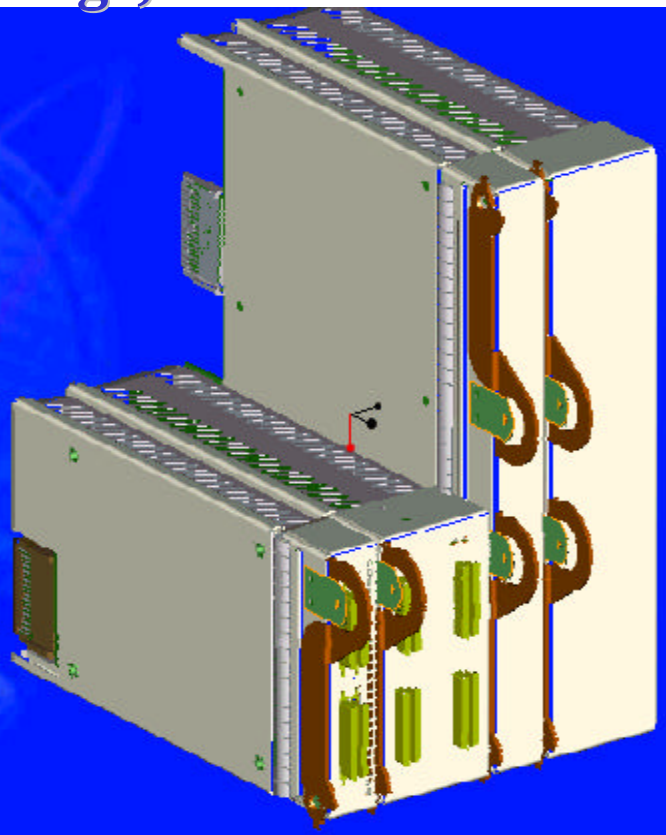
Volume production of Infiniband chips should start late this year and early next year. "Initially, it will be used with the Oracles and IBM DB2s and the data centers at the back end. That's still growing," Deierling said.

With one fewer heavyweight competitor to worry about, Mellanox and others are relishing the opportunity to pick up more customers. "It opens up the market opportunity and makes it more credible that we can get a larger market share," Deierling said.

Eventual Goal

End of server based on PCs. modular format has "Blades" for CPU+memory, Storage, etc.

- Four Adapter form Factors:
  - Single-Wide/Single-Height
    - "*Standard*"
  - Single-Wide/Double-Height
    - "*Standard-Tall*"
  - Double-Wide/Single-Height
    - "*Standard-Wide*"
  - Double-Wide/Double-Height
    - "*Tall-Wide*"
- Variety of possible Adapter solutions



Source: IDF spring 2001, Mazin Yousif, Ph. D. Mazin

RapidIO

Overview

RapidIO

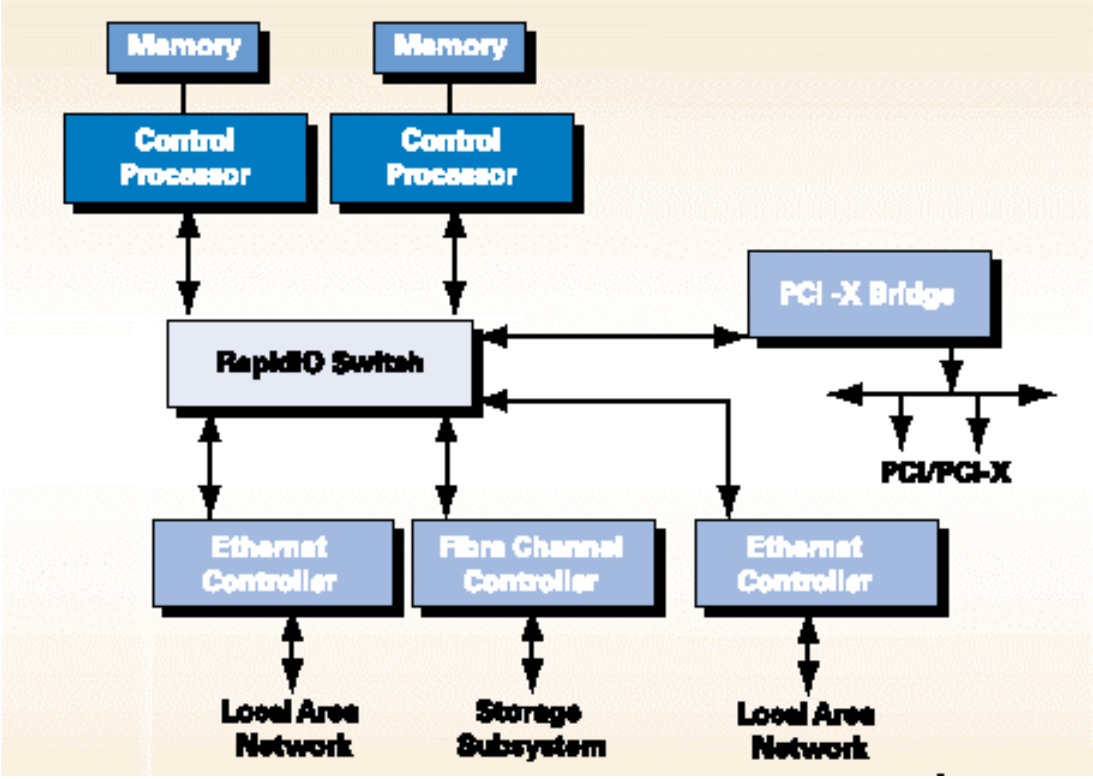
Competitor to PCI

Electronic serial interconnect scheme for chips on a circuit-board and on a backplane

- 10 Gbps bandwidth using 8bit data ports
- Low latency
- Transparent to Software
- Low complexity
- Multiprocessing support
- Flexible switch based topology
- Open standard

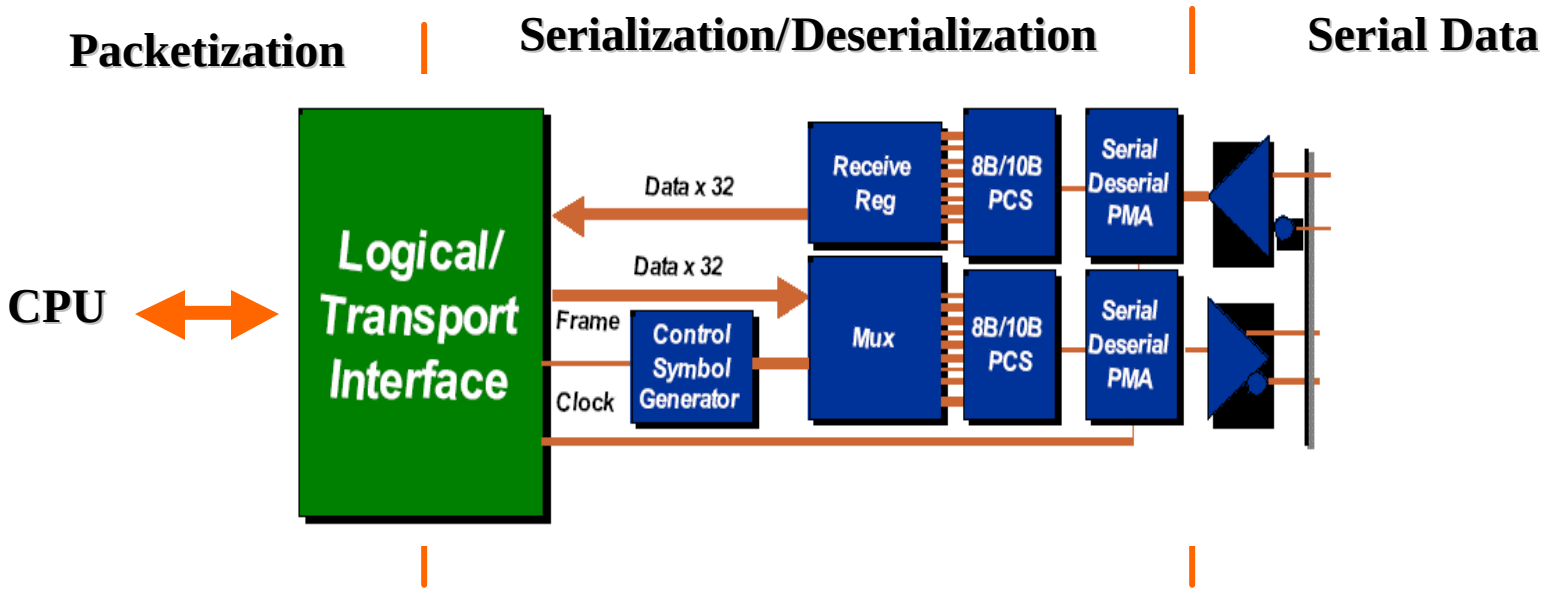
Concept

- Similar to PCI Express/3GIO
- Interconnect CPUs and Peripherals via high speed serial links



Source: RapidIO.org

RapidIO Physical Layer



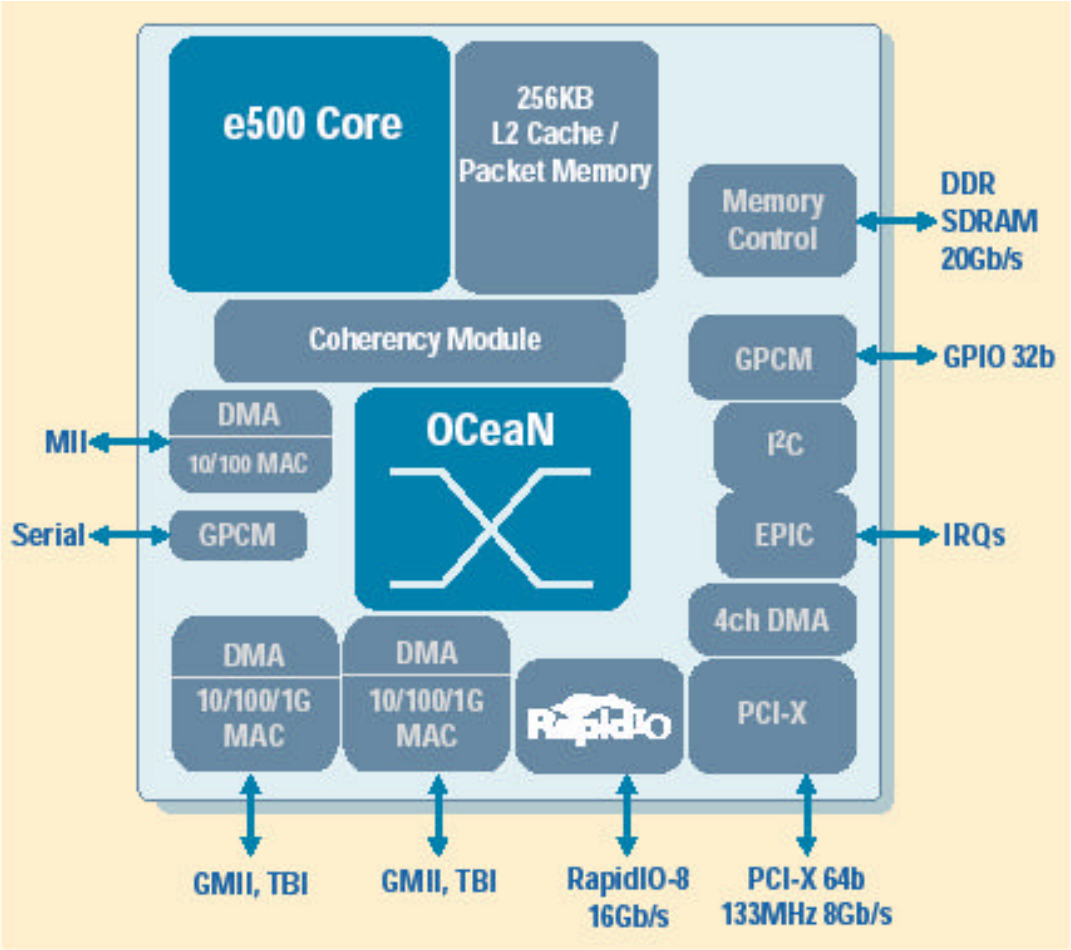
PCI RapidIO uses Physical Coding Sublayer (PCS) and Physical Media Attachment (PMA) to organize packets into a serial bit stream.

Source: Intel Corporation

RapidIO CPU example

Motorola MPC8540

- 1GHz PowerPC
- 256KB On-chip L2
- 128Gb/s On-chip Fabric
- 333MHz DDR
- RapidIO
- 64b PCI-X
- Dual Gbit Ethernet + 1 Standard Ethernet
- Serial ports
- RapidIO interface
- Due Q3 2002



MPC8540 Block Diagram

XDAQ

XDAQ Framework

Tools and Libraries (C,C++)
HAL
XOAP
MAZE
XDAQ SHELL
PVSS

XDAQ core (C++)
Distributed programming environment

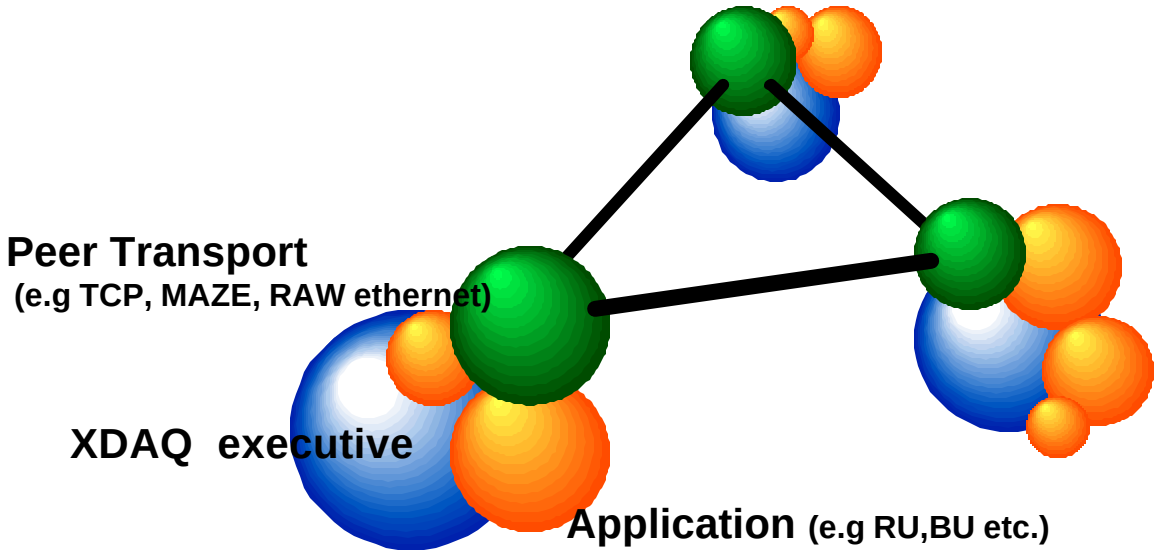
Applications (C++)
Benchmark test suite
Event Builders
Testbeams
Peer Transports (GM, TCP etc.)
FEDKit
Oracle DB access

XDAQ Win (Java)
XDAQ configuration and control
Monitoring
Scripting facilities
basic APIs for RunControl development

External Libraries and Drivers (C, C++, Java)
Xerces
Mathpack
GM
JAXM, JACL etc.

Concerted effort of active users and contributors within CMS collaboration

XDAQ Environment



Ongoing activities

- Linux bigphys memory allocation on XDAQ (use of logical and physical addresses)
- Loadable MAZE peer transport for XDAQ (5 protocol to be supported, BS IN, BS OUT, EVB IN, EVB OUT, IMMEDIATE*)
- FEDKit new allocation scheme (support for buffer loaning scheme in XDAQ and compatibility with MAZE and GM peer transport)
- XDAQ polling scheme (alternative to multiple threads when running performance tests)
- GM on bigphys area (patch to GM 1.5.1 from Myricom for direct use of bigphys memory chunks on GM)
- XDAQ/JAS integration using SOAP (support for histograms preparation and display from XDAQ applications)
- XDAQ documentation
- XDAQ applications (benchmark test suite RoundTrip, StreamIO, event builder models) for reference measurements

Re-usable components

- Benchmark test suite (any transport)
  - Round Trip
  - Stream IO point to point, full duplex
  - Host full I/O
- Event Builders
  - Direct and Indirect models implementations
  - Support for performance measurements
  - Direct re-use in test beam applications
- DAQ Column test suite
  - RU throughput I/O with buffer loaning. Combination of input FED/Myrinet with different output means.

1.1 or later

- 1.1 planned end of May 2002
- Native Oracle database access
  - Read/Write access to oracle from XDAQ (data as DOM)
- SOAP attachments
  - XOAP library to support (MIME) attachments for transfer of binary embedded information
- Re-usable Event Builders and DAQ column test suites
- MAZE peer transport BS in/out protocol
  - EVB/FRL input, requires support for multiple NIC on same host
- Performance measurements on computer FARM
  - COTS and custom protocols
- Integration with RunControl Components
  - Resource and Information Services
- HAL extension

XDAQ on Gemini

- Creating support for Gemini in XDAQ framework
- Preliminary tests at UCSD
- Event builder tests @ CERN
- Raw ethernet test @CERN

S. Bhattacharya, J. Gutleber, L. Orsini

XDAQ on Gemini

- Creating support for Gemini platform within the XDAQ framework was easy.
- XDAQ already supports ppc platform for VxWorks architecture (MV2304).
- Migrating from the mv2304 platform support to Gemini required only a few changes,
- And a few e-mail exchanges with the developers at CERN.
- XDAQ documentation was useful!

XDAQ on Gemini (contd..)

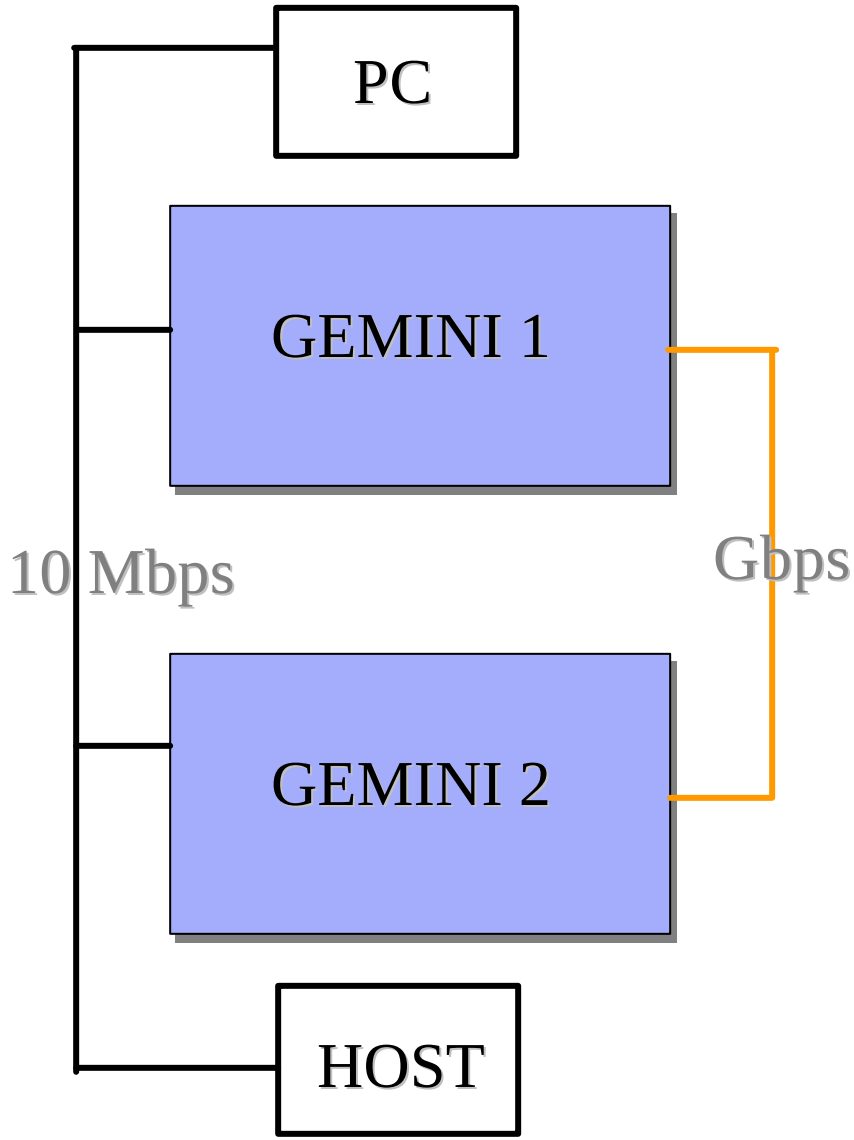
- Migrated from the already supported vxWorks based PPC platform - MV2304.
- Gemini does not have a DMA universe chip
  - build XDAQ without DMA universe support
- Raw ethernet intrface for XDAQ uses the ethrnet interface name.
  - In VxWorks network interface name is vendor dependent.
  - The interface name is hardcoded in XDAQ.
  - Vendor dependent interface name for Gemini was added in the XDAQ framework.

Preliminary test at UCSD

- Included ‘RoundTrip’ – a simple XDAQ based application in the framework.
- In RoundTrip one peer sends an empty I2O message, on successful receipt, the other peer sends back the packet… and so on.
- Performed RoundTrip test between linux pc and one Gemini using 10BaseT ethernet.
- And between two Gemini modules using gigabit interface.

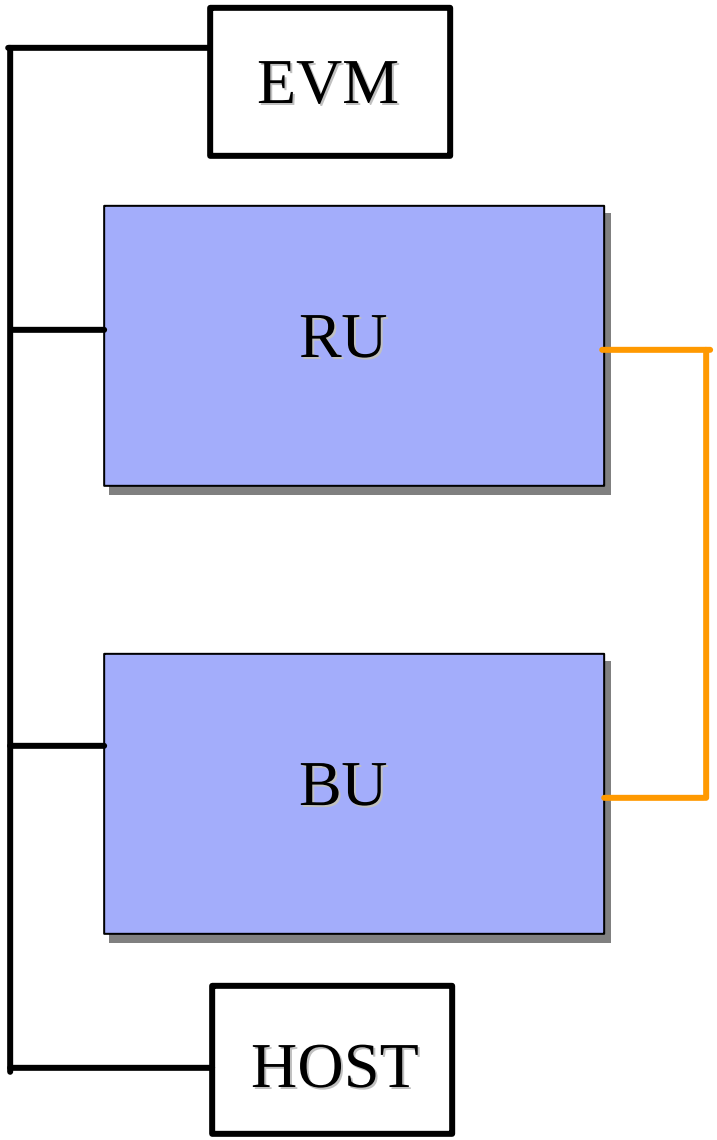
Test setup at CERN

- 2 Gemini modules
- Boots and loads XDAQ applications from a sun host
- Connected to the host and a pcPentium running linux, over regular ethernet private network
- Each Gemini has a GNIC interface, connected by fiber



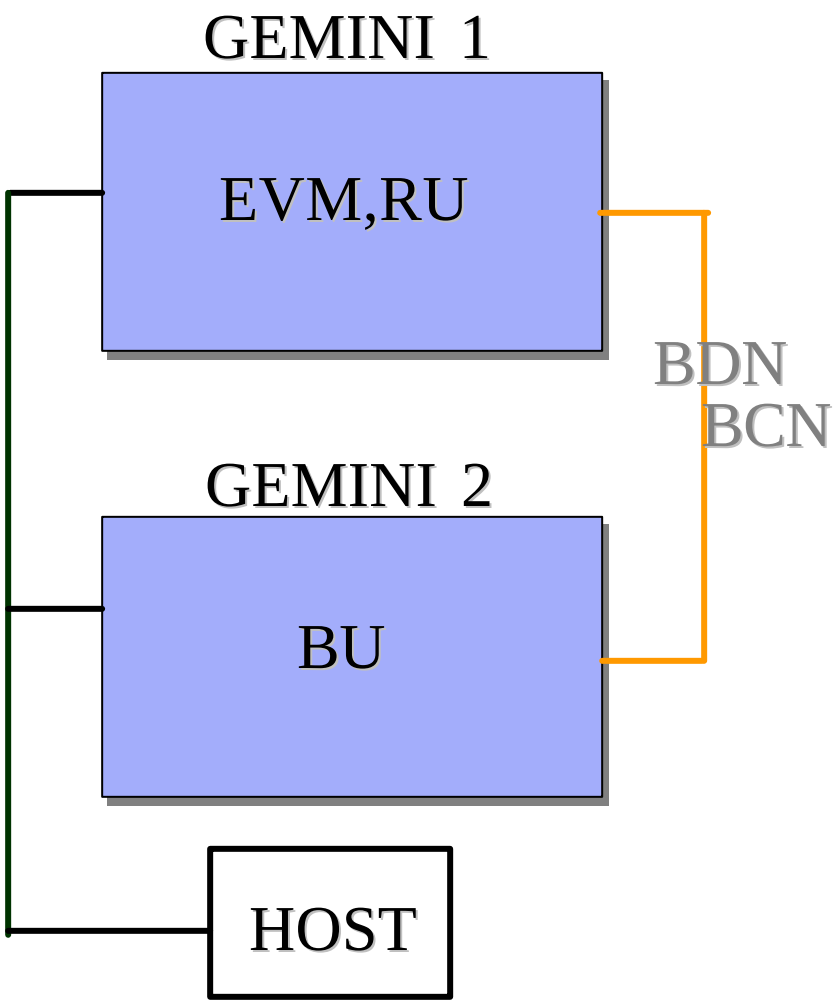
EVB test over Ethernet

- PC runs EVM application
- One Gemini runs RUI/RUO application and the other runs BU
- XdaqWin runs on the host
- Peer to peer communication over ptTCP
- Only one network



EVB Test over Gigabit Link

- EVM and RUI/RUO applications on one Gemini
- BU application on the second Gemini
- All applications communicate over the gigabit link, using ptTCP
- Ran a long test for 4 hours successfully.
- Message size had to be set < ethernet MTU size.



XDAQ/Gemini issues

- Gemini has 128 MB RAM
- However XDAQ does not spawn if the entire memory is made available to vxWorks
- Runs with 32 MB.
- Reason:
  - vxWorks loads applications at the available top, vxWorks kernel near the bottom.
  - PPC relative branch instruction limited to +/- 32 MB (24 bits)
  - ‘Relative branch’ from application to system functions can’t be resolved.
  - Possible solution: compile without ‘relative branch’

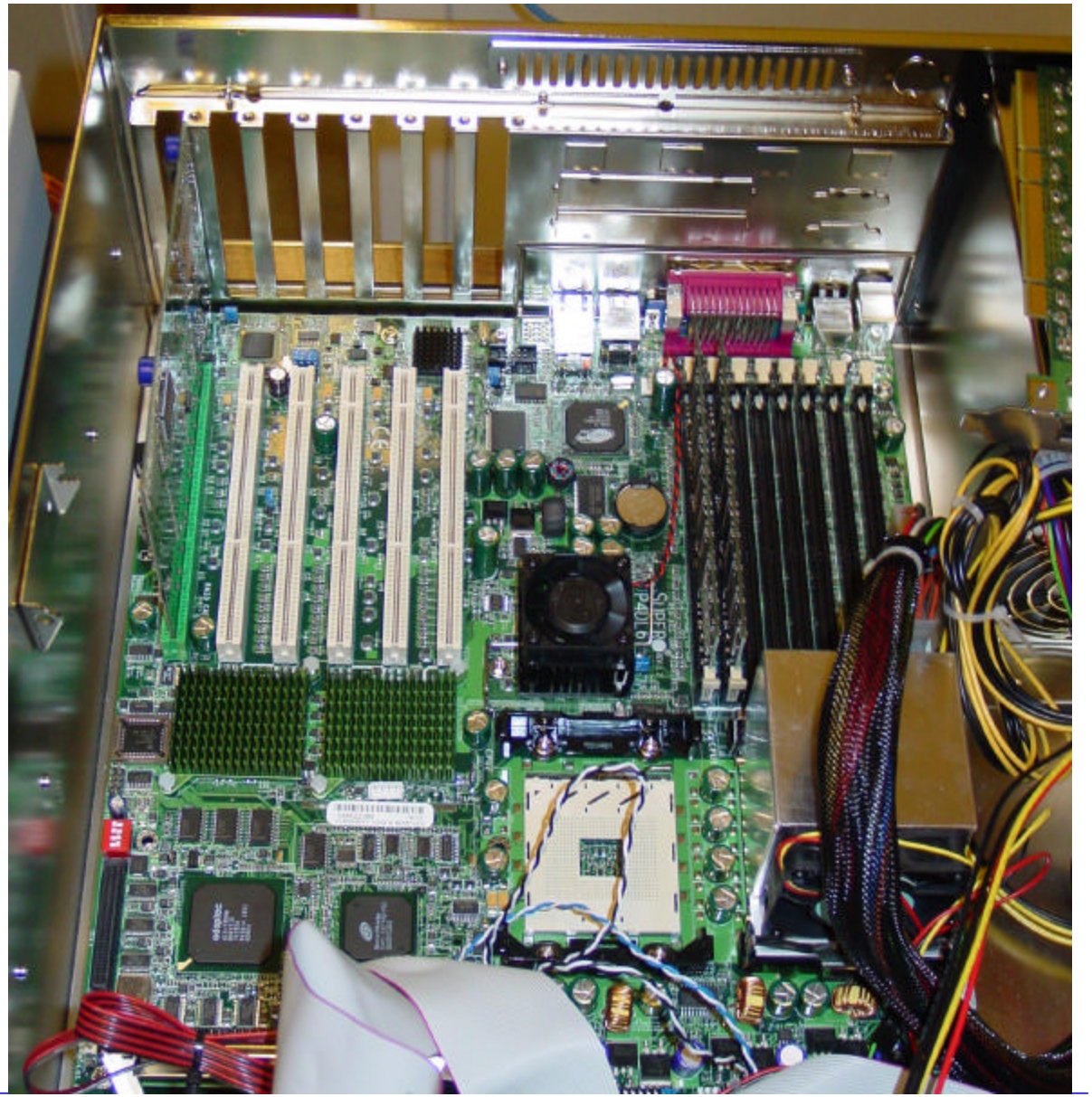
XDAQ/Gemini …

- Access to Gemini from host through host shell (WindShell)
  - Target server cache memory needs to be configured properly for successful loading of XDAQ.
- Available memory 32 MB
  - Object files should be small
  - Strip debug information

XDAQ Builder Unit

- A software component for XDAQ
  - Allows full or multi-level event building
  - Fully XML configurable
  - Code running and tested on
    - Solaris, Linux, VxWorks operating systems
    - Sparc, Intel, PowerPC hardware environments
    - Raw FastEthernet, TCP/IP, Myrinet communication technologies
    - Integration with a prototype trigger card done (used in testbeams)

Raw Ethernet test

- We have successfully tested the raw Ethernet peer transport on Gemini also
  - Used ethernet interface
  - Ran round trip test using ptEther
- XDAQ peer transport using raw ethernet (ptEther) associates a unique network interface name and device number with each platform.
- Gemini supports two different ethernet interfaces, a regular ethernet on board and a copper/fiber gigabit interface with different interface names.
- With current version of ptEther switching from one network interface to other needs recompiling

Another View of GC