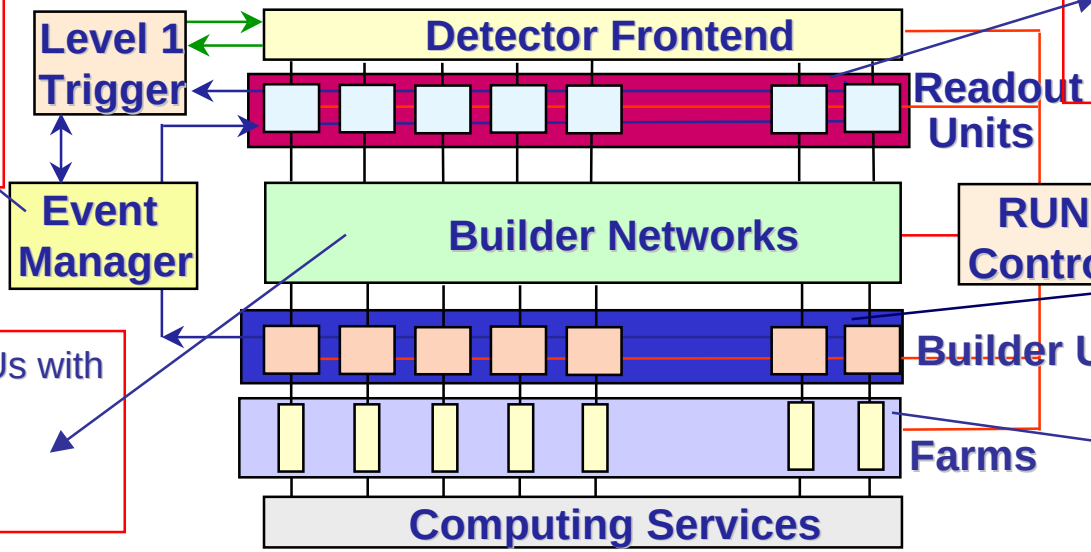
The CMS Event Builder Demonstrators based on GigaEthernet Switched Network

- Marco Bellato	INFN-Padova
- Luciano Berti	INFN-Lab. Naz. Legnaro (LNL)
- Michele Gulmini	INFN-LNL
- Gaetano Maron	INFN-LNL
- Nicola Toniolo	INFN-LNL
- Gabriele Vedovato	INFN-LNL
- Sandro Ventura	INFN-Padova
- XiaoQing Yang	INFN-LNL

CMS DAQ Layout

- Init the RO process from the detector frontend
- Assign Event Identification
- Assign events to destin.
- Throttle Lv-1 trigger



Where the event fragments accepted by Lv-1 Trigger are buffered

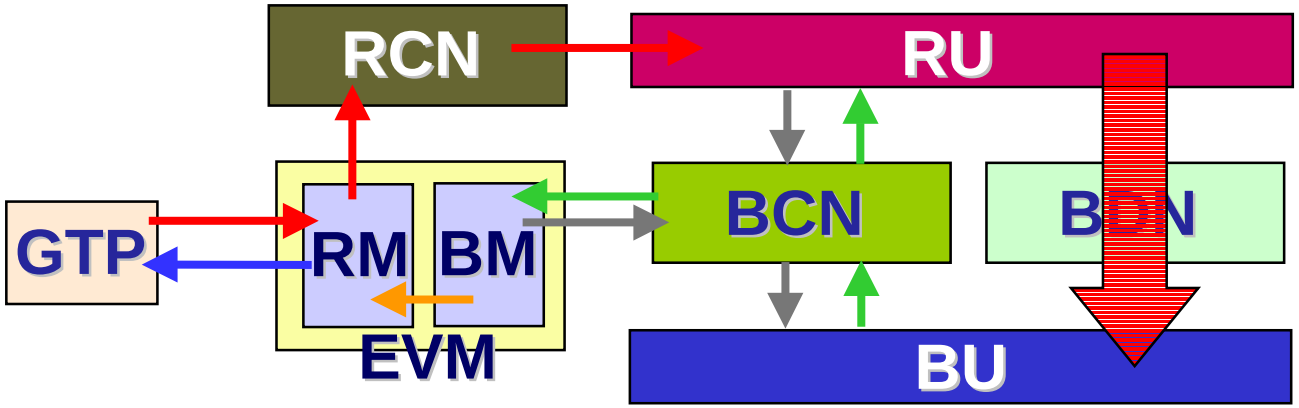
Where the fragments of an event are assembled

Interconnects RUs with BUs with
- low latency
- fast switching capability
- high bandwidth

Where the High Level Triggers are performed

Collision rate	**40 MHz**
Level-1 Maximum trigger rate	**100 kHz**
Average event size	**1 Mbyte**
Builder Network (512-512 port) bandwidth >=	**500 Gbit/s**
Event filter computing power	**5 10^6 MIPS**
Data production	**Tbyte/day**
High Level Trigger acceptance	**1 - 10 %**
Overall dead time	**< 2%**

Data and Control flows

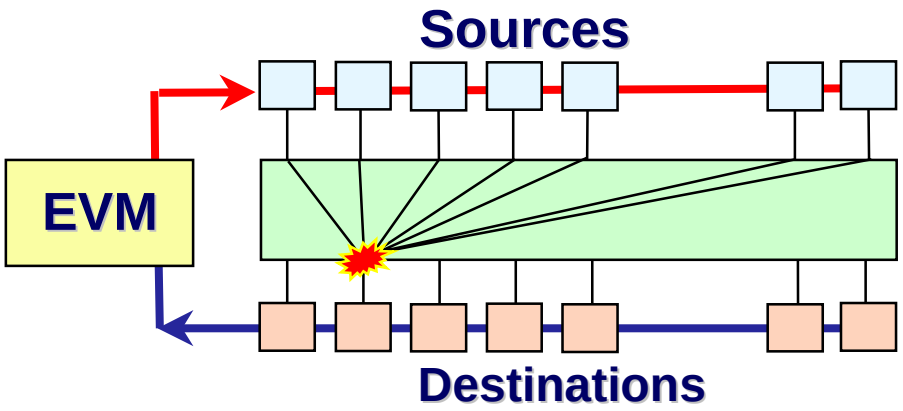


EVM	Event Manager
RM	Readout Manager
BM	Builder Manager
RCN	Readout Control Network
BCN	Builder Control Network

GTP	Global Trigger Processor
RU	Readout Units
BU	Builder Units
BDN	Builder Data Network

Event Builder Requirements

- **Event Size: 1 MByte**
- **Fragment Size (500 Rus) : 2 kByte**
- **Raw Througput at 100 kHz: 1 Tbps**
- **Command messages rate: about 10^6/s**



Output link congestion is intrinsic in the event builder application as multiple sources always compete for the same destination.

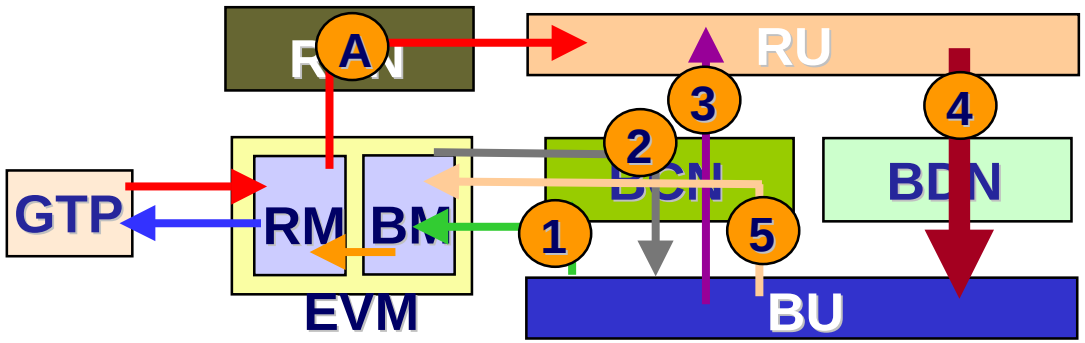
A traffic shape mechanism is then needed to avoid clashes on the output link

- barrel shifter
- rate division
- credit based protocols
- ….

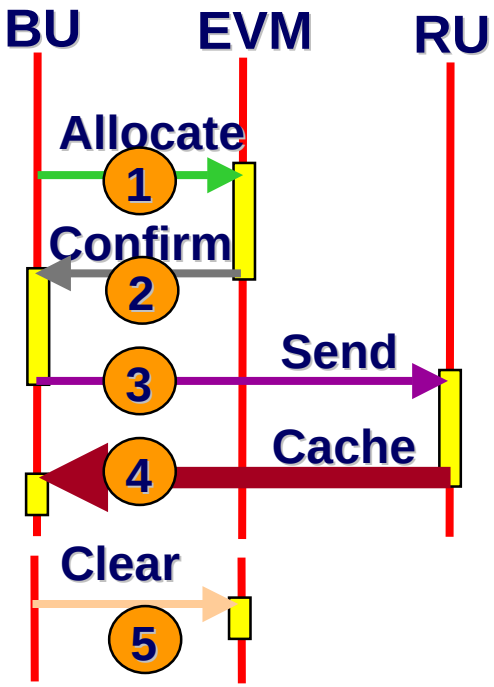
Demonstrators and Simulations

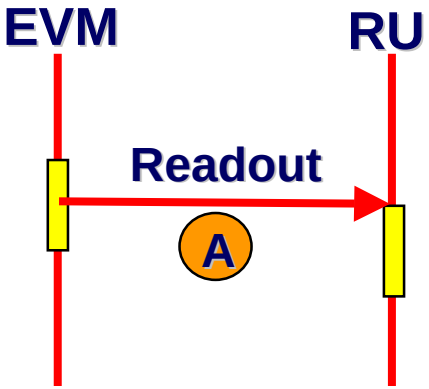
- The needed 1000 ports switch will be probably built connecting several small commercially available switches
- The behaviour of a such system is difficult to predict. Detailed simulations are then necessary
- Technological evaluation of small scale prototypes or demonstrators (16x16 or 32x32) are usefull:
  - different transmission technologies can be compared (at the moment Myrinet (see F. Meijer talk) and Gigaethernet).
  - provide input parameters to the simulations that can extrapolate the results to larger configurations.
  - exercise and compare different type of event builder protocols and traffic shaping mechanism.

Event Building Protocol for the Demonstrator

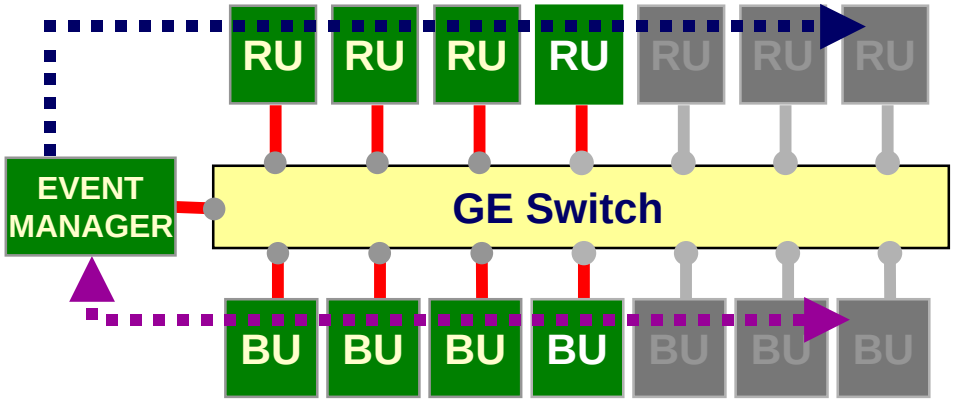


Allocate: assign a new event
Confirm: return evtID
Send: send data from evtID
Cache: returns data evtID
Clear: free evtID





The Gigaethernet Demonstrator



- A GigaEthernet based 7x7 Event Builder demonstrator has been set up. Only 4x4 is in operation. We found problems using the switch backplane.

- BDN, BCN and RCN in the same switch

- Full event builder

- Data moved from RU memory to BU memory

- One network packet per message

Traffic Shaping

- Output buffer overflow is prevented controlling, at BU level, the number of events are concurrently in the building phase. No packet loss.

- Outside switch traffic shaping is not used. According to the EVB protocol adopted , each BU asks, with a round robin scheme, event fragments to each RU; no broadcast is used. This can help to “shape” the traffic. But the main “shape” is done internally by the switch itself.

Hardware and Software Components

HARDWARE

- RU, BU and EVM are based on PIII - 450 MHz commodity PC based on a single 32 bit/33 MHz PCI IO bus.
- Intel 6000 GigaEthernet Switch is used
- Both Intel Pro 1000 PCI/GE and GNIC II Packet Engines interfaces are used

SOFTWARE

- All the nodes run vxWorks 5.4
- Standard vxWorks drivers have been implemented according to the Pro 1000 and GNIC II specifications (under NDA)
- Optimised software for Event Manager, Readout Unit and BU has been written.

The Intel 6000 GigaEthernet switch

Chassis + Fan + Power Supply

4 slots for media Modules

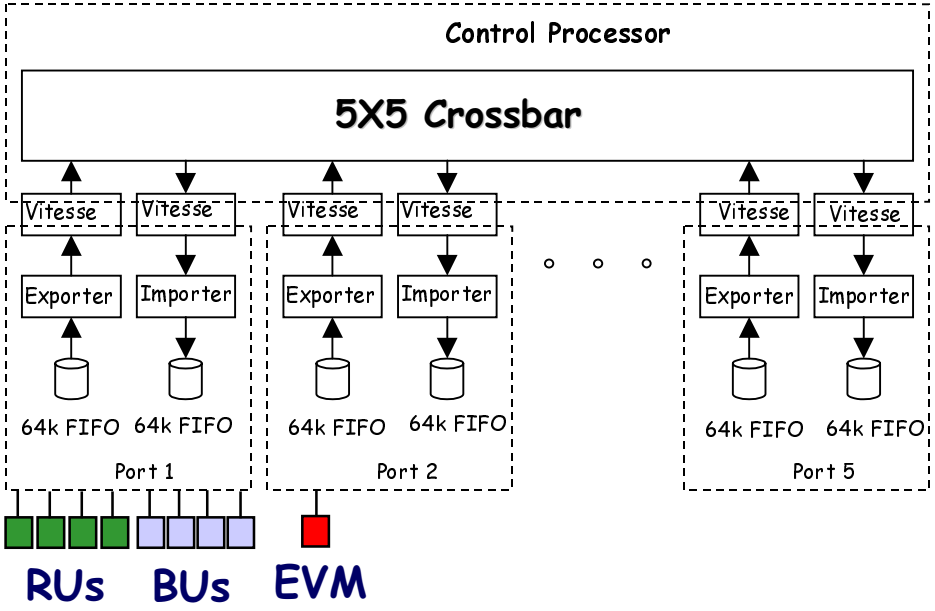
Control Processor

- MIPS RISC processor (NKK NR4700 - 64-bit, 175 Mhz)
- TFTP upgradeable software loads from flash PROM
- 5x5 X-Bar interface
  - 40 Gbps capacity (Full duplex)
- Management ports
  - DB-9 and 10/100 console ports
- Redundant CP optional

Media Modules

- Non-blocking L2/L3+
- 8-port Gigabit Switch
  - 1000BaseSX (at FCS)
  - 4+4 SX/LX (post-FCS)
  - 1000BaseT (post-FCS)
- 24-port 10/100 Ethernet
- 12-port 100BaseFX

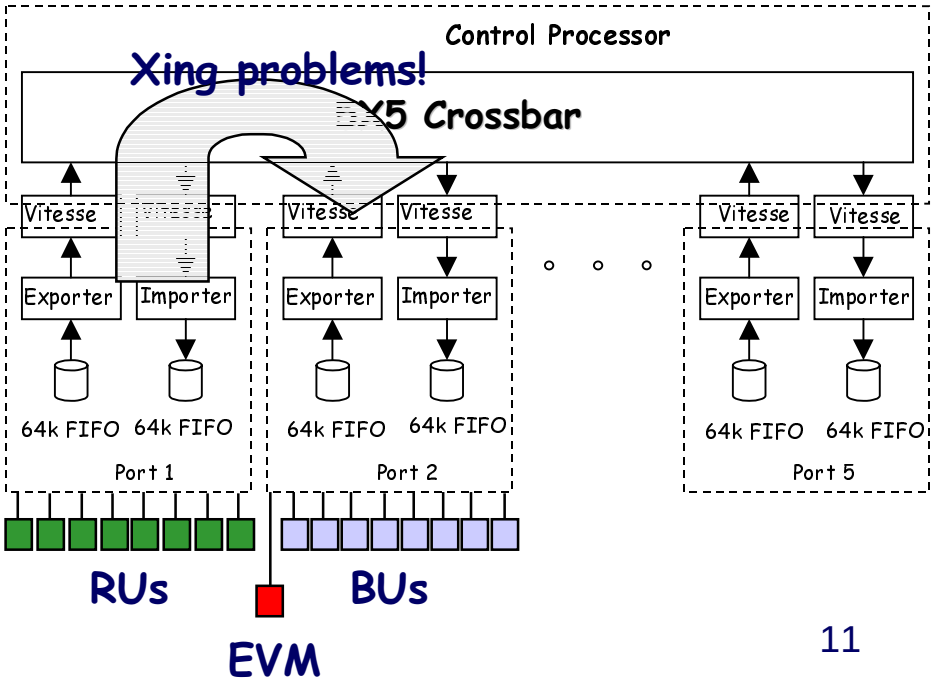
Switch Configuration used



Running Configuration

Faulting Configuration

Any eb configuration from 1x1 to 8x8 fails in this case



The GE Demonstrator

Software Issues

- **Message rate is very high**
- **Low Overhead is crucial**
- **No high level protocol used**
- **Direct mapping between the network interface into the application's address space**

Low Latency GigaEthernet vxWorks drivers

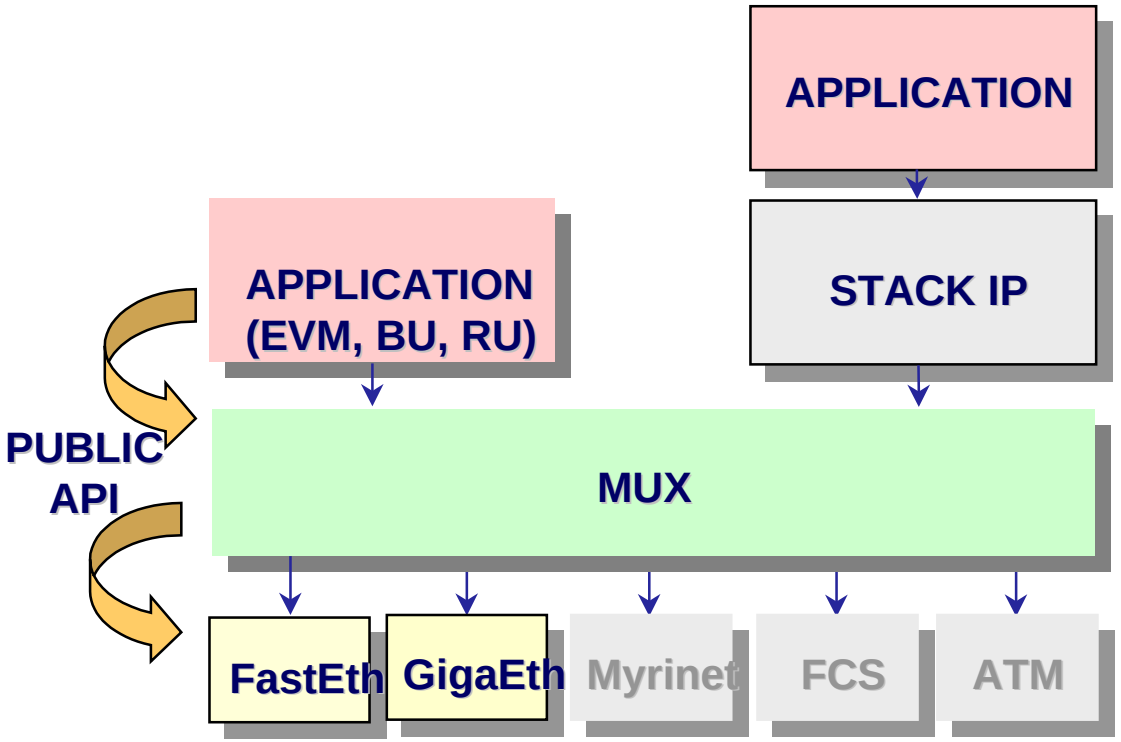
- Enhanced Network Device drivers (END) have been written both for Intel Pro 1000 and GNIC II network interfaces.

- END drivers exploits the low latency communication layer provided by vxWorks SENS (Scalable Enhanced Network Stack).

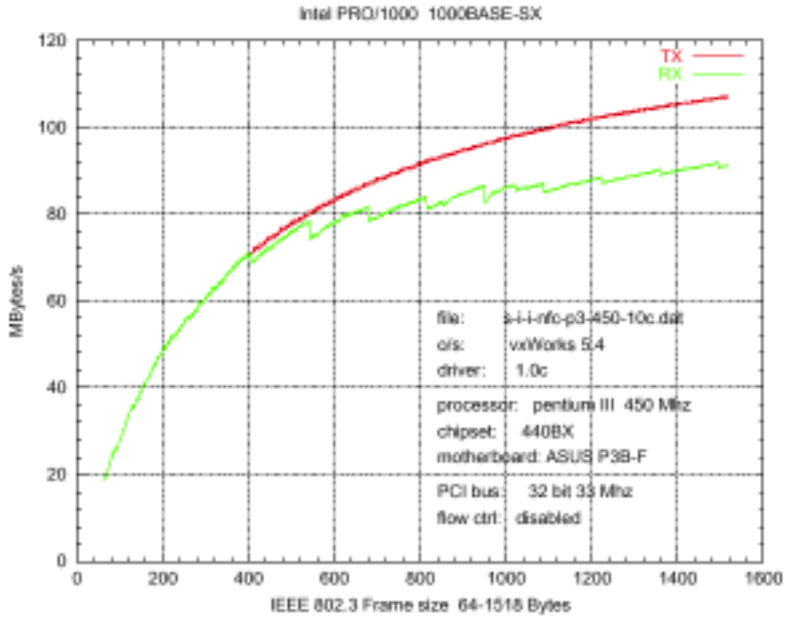
- The EVB protocols used in our demonstrator are implemented using SENS with direct control of the ethernet frames.

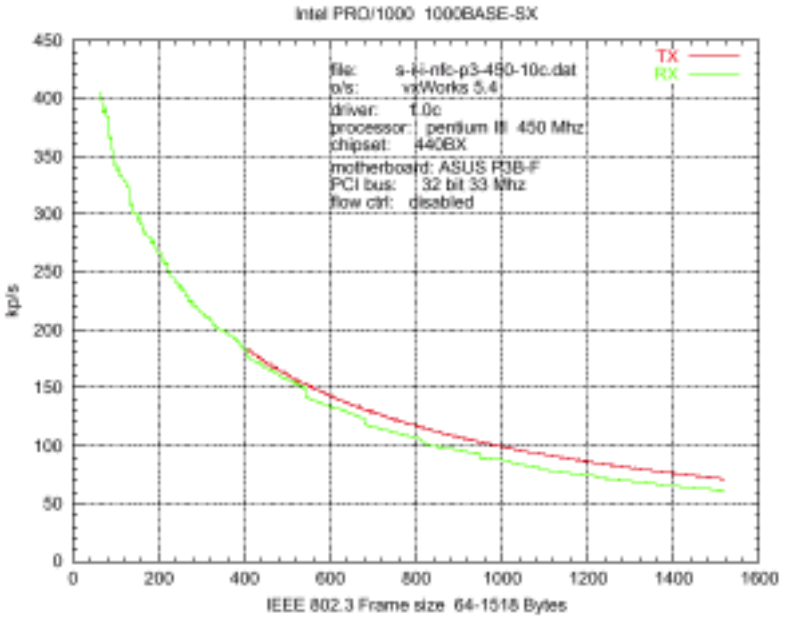
- Other systems (Unix, WNT, etc.) can communicate with SENS systems: 1) having direct access to the ethernet frames; 2) defining a common communication layer (e.g. DLPI, VIA, etc.)

The SENS Stack



GigaEthernet point to point tests (I)

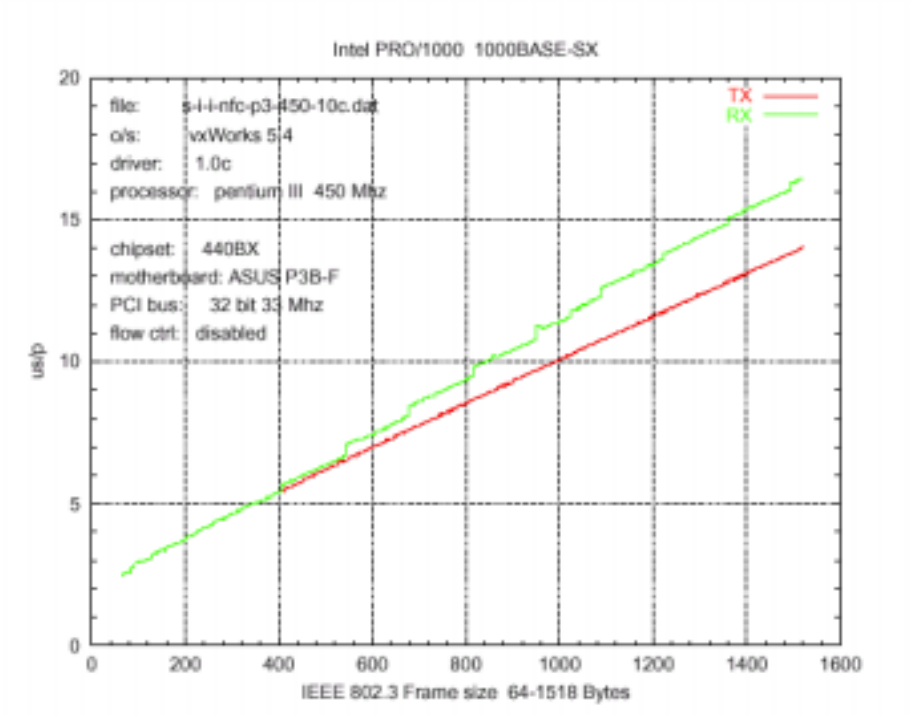


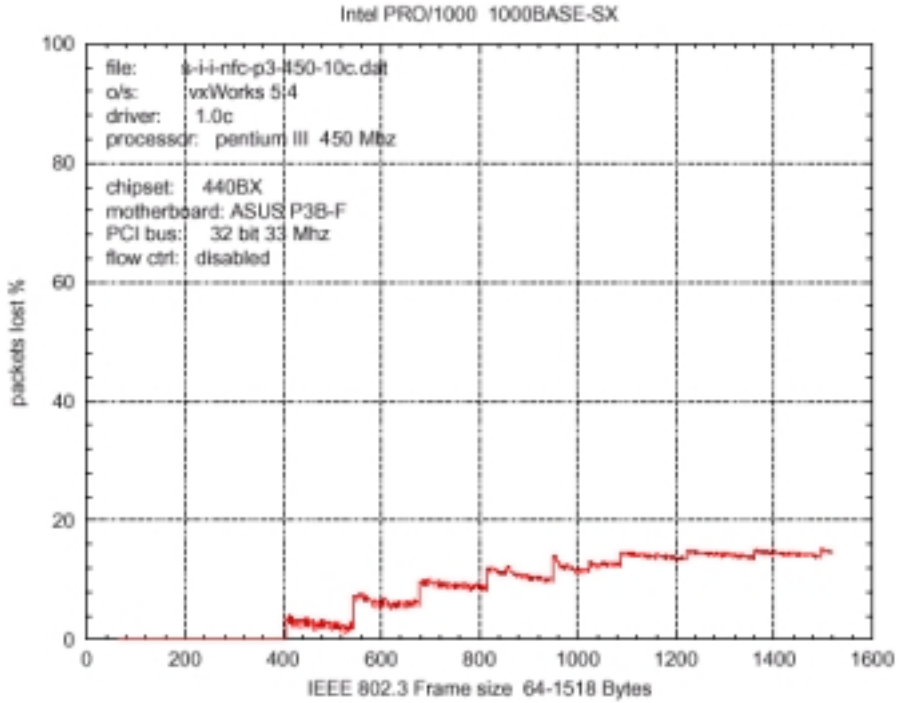


Ptp Test for Intel Pro 1000 NIC

- The sender (110 MByte/s) saturates the receiver (90 MByte/s)

GigaEthernet ptp tests (II)



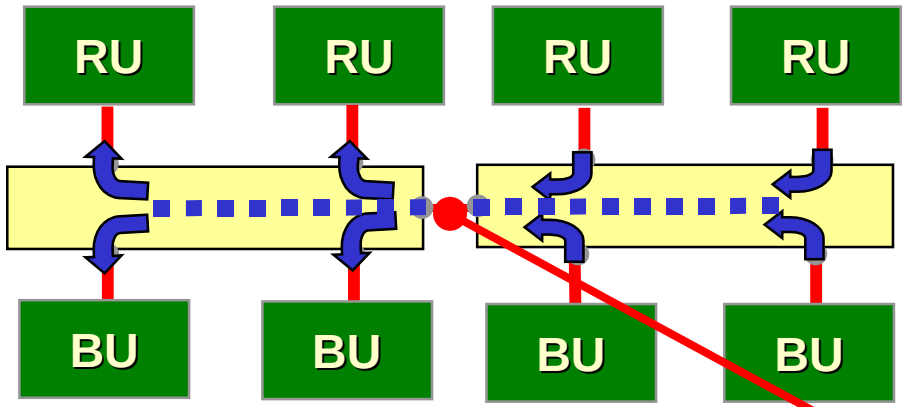


- **Sofware overhead close to 2 μs**

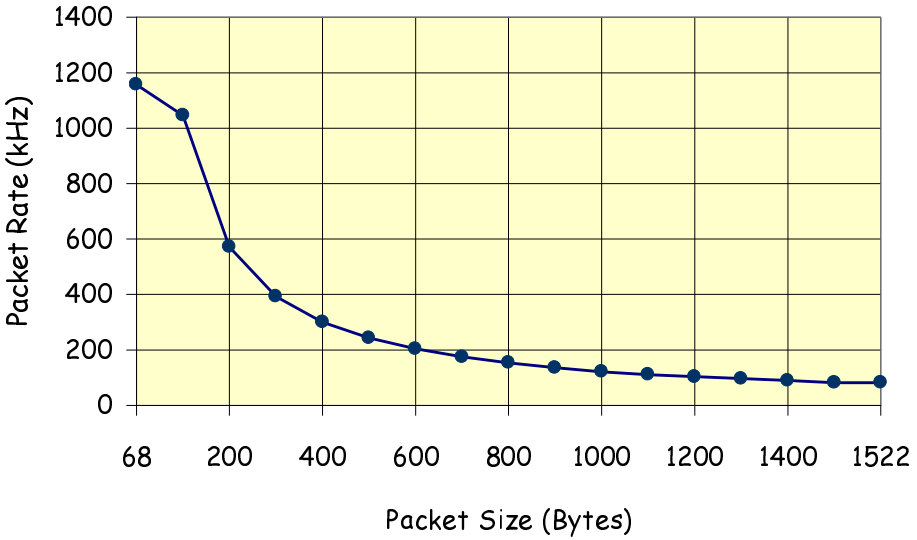
- **Packet loss (no flow control) up to 16%**

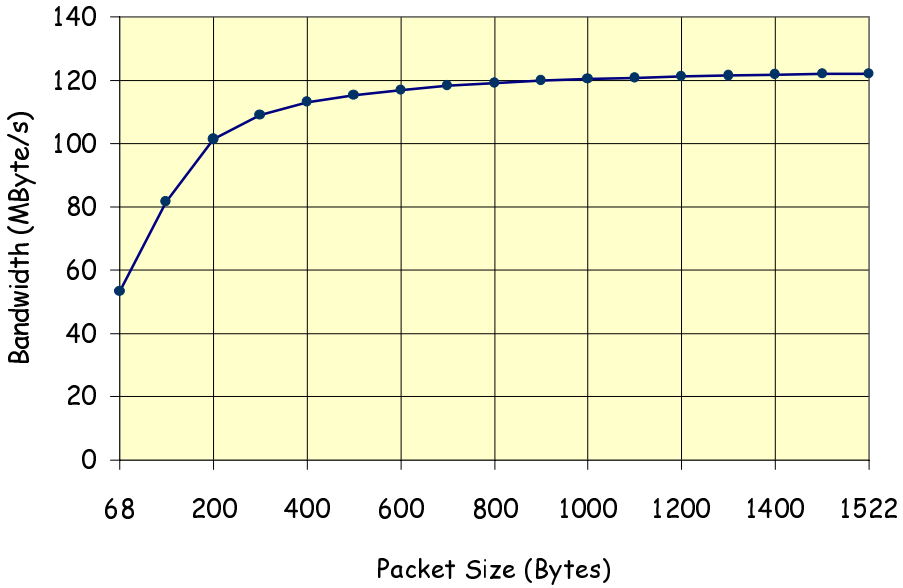
- **With flow control enabled, no packet loss**

GE Intrinsic Limits



2 Intel GigaExpress switches are used in this case





Low Latency Message Layer

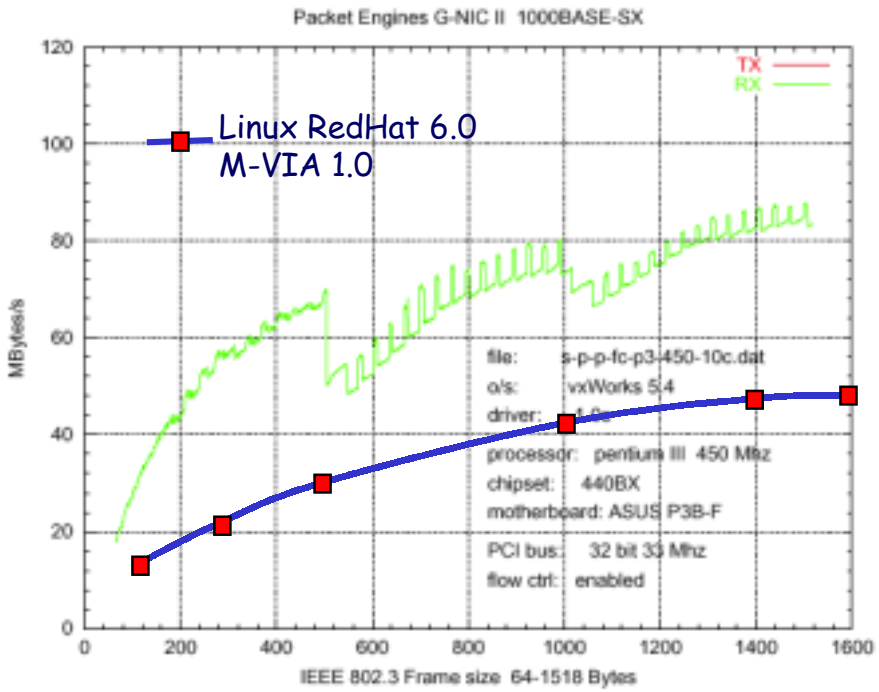
- High speed networks require the direct mapping between the network interface into the application's address space avoiding in this way the overhead of the operating system calls.

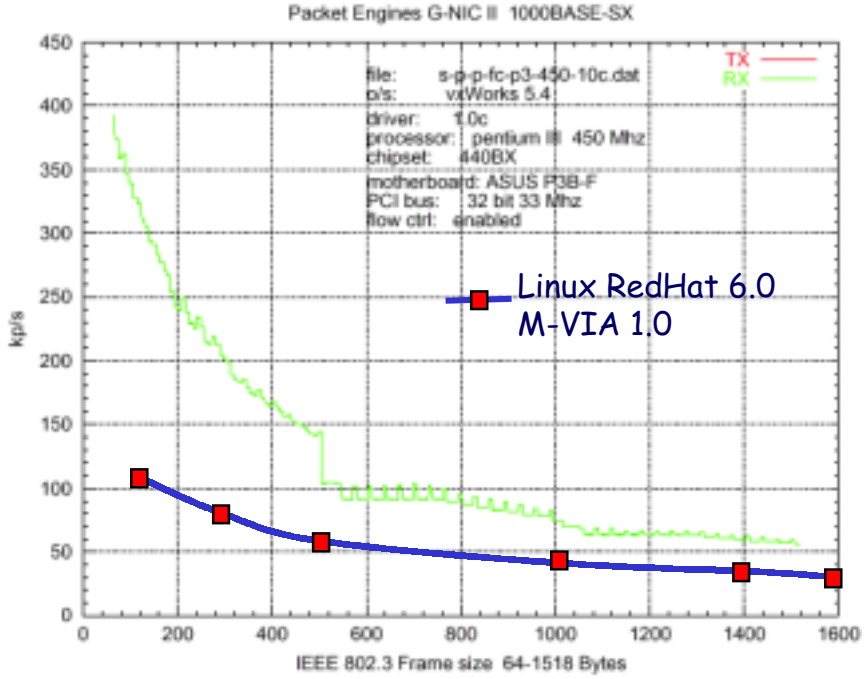
- The research community is producing several low latency messaging layers like
  - Fast Messages (e.g. Illinois Fast Messages)
  - Active Messages (e.g. Berkley's Active Messages)
  - U-NET
  - etc.

- Intel, Compaq and Microsoft proposed a cluster interface standard called VIA (Virtual Interface Architecture) that appears to become a widely used standard interface for high performance communication. VIA features include:
  - user-level protected network access
  - gather-scatter interfaces (zero copy)
  - early demultiplexing of incoming traffic

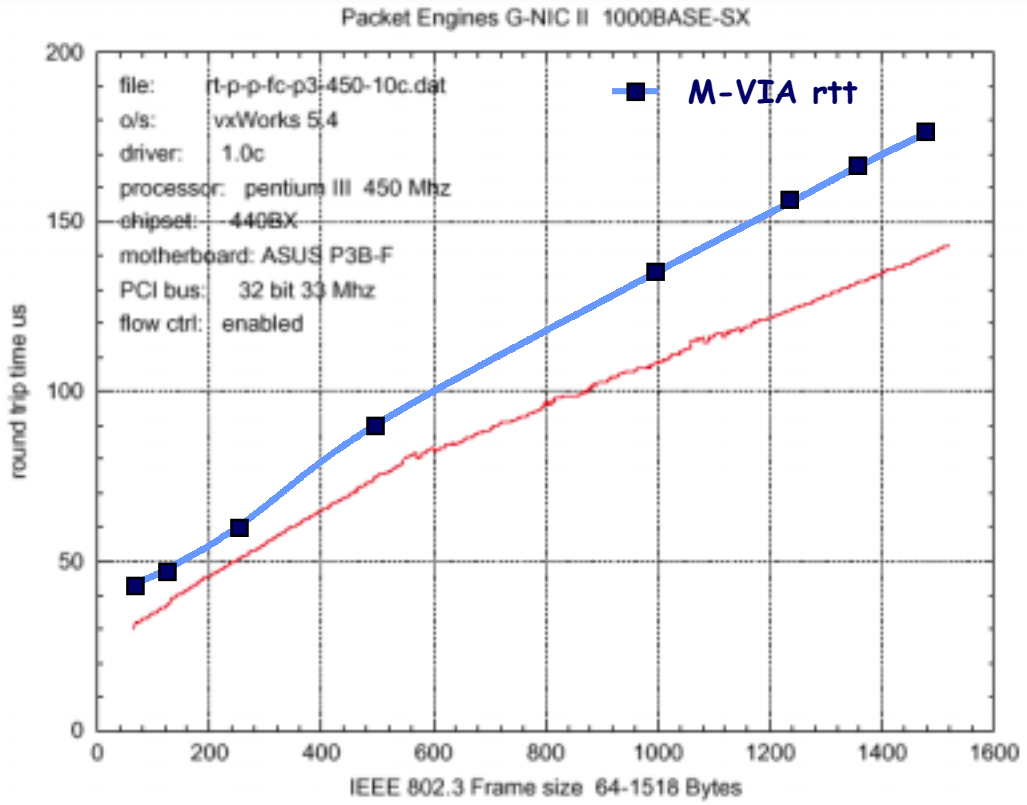
- A simple point to point test has been performed to "taste" VIA with Gethernet. We used the "Modular VIA for Linux" as provided by NERSC http://www.nersc.gov/research/FTG/via/ over PacketEngines GNIC II interfaces

VIA ptp test


Packet Engines G-NIC II 1000BASE-SX
TX
RX
Linux RedHat 6.0
M-VIA 1.0
file: s-p-p-fc-p3-450-10c.dat
o/s: vxWorks 5.4
driver:
processor: pentium III 450 Mhz
chipset: 440BX
motherboard: ASUS P3B-F
PCI bus: 32 bit 33 Mhz
flow ctrl: enabled
MBytes/s
IEEE 802.3 Frame size 64-1518 Bytes

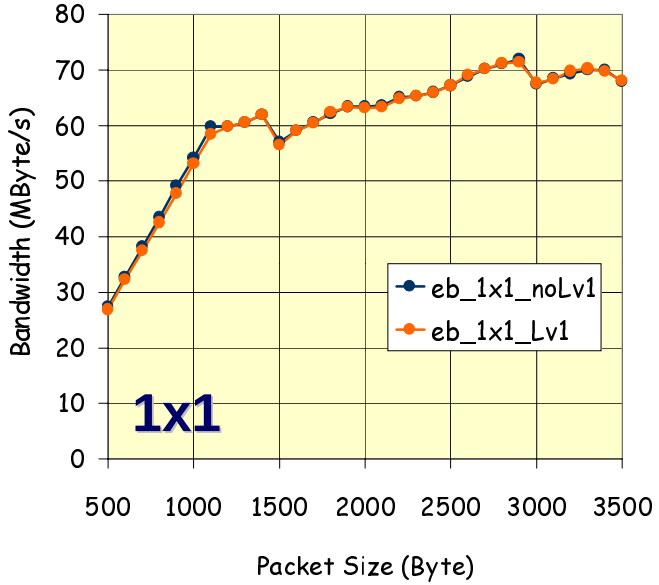

Packet Engines G-NIC II 1000BASE-SX
TX
RX
file: s-p-p-fc-p3-450-10c.dat
o/s: vxWorks 5.4
driver: 1.0c
processor: pentium III 450 Mhz
chipset: 440BX
motherboard: ASUS P3B-F
PCI bus: 32 bit 33 Mhz
flow ctrl: enabled
Linux RedHat 6.0
M-VIA 1.0
kp/s
IEEE 802.3 Frame size 64-1518 Bytes

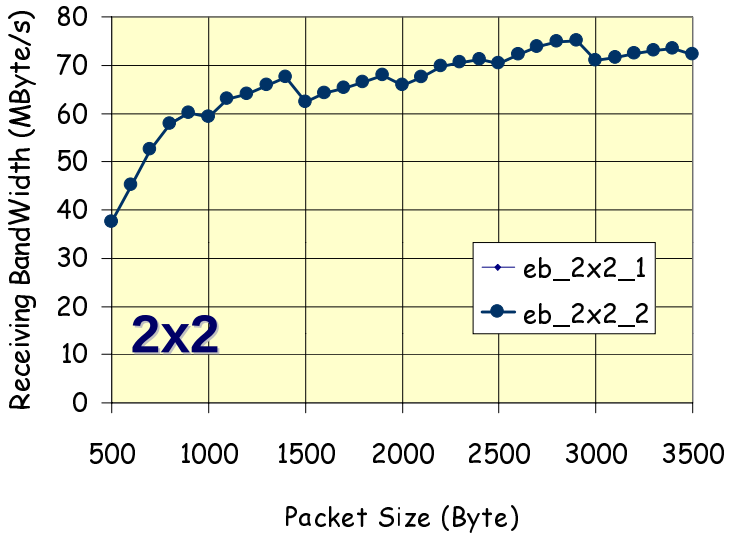
VIA round trip time test

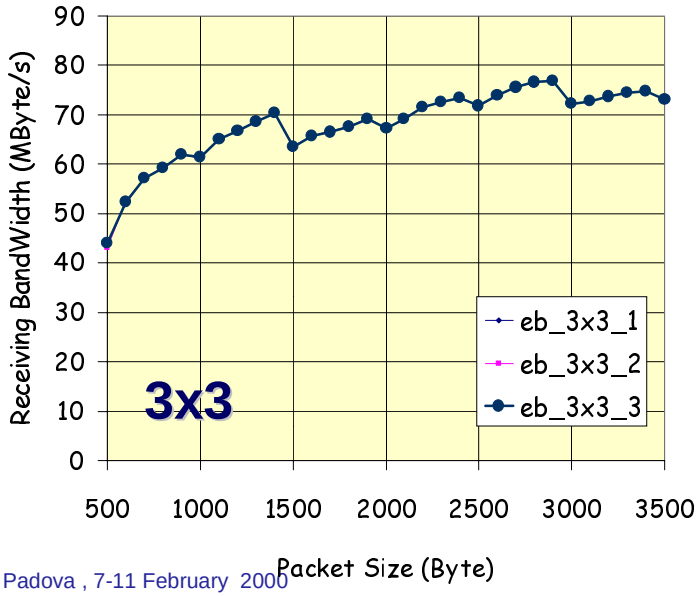


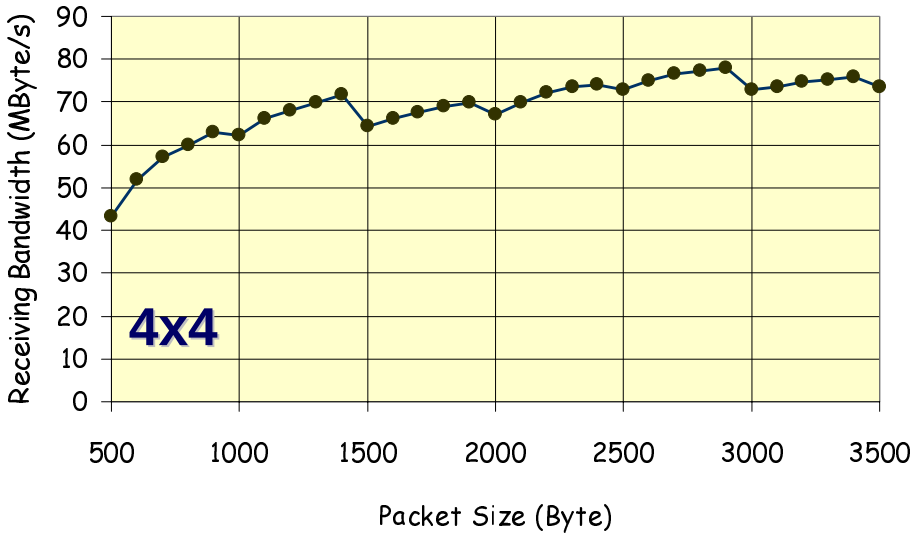
Event Builder Tests with Fixed Event Size

GE Event Building BU Receiving Bandwidth

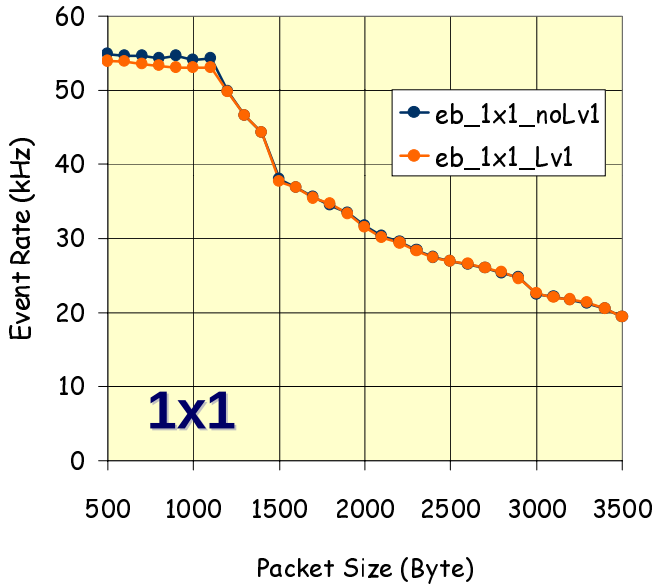


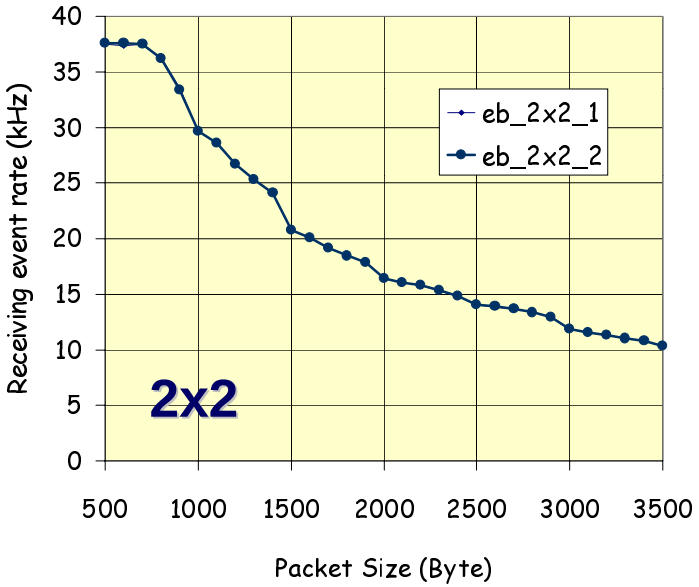


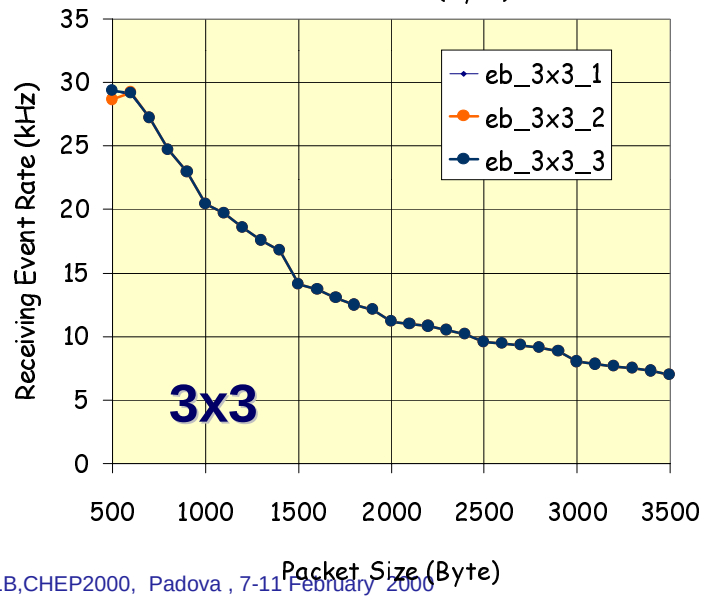


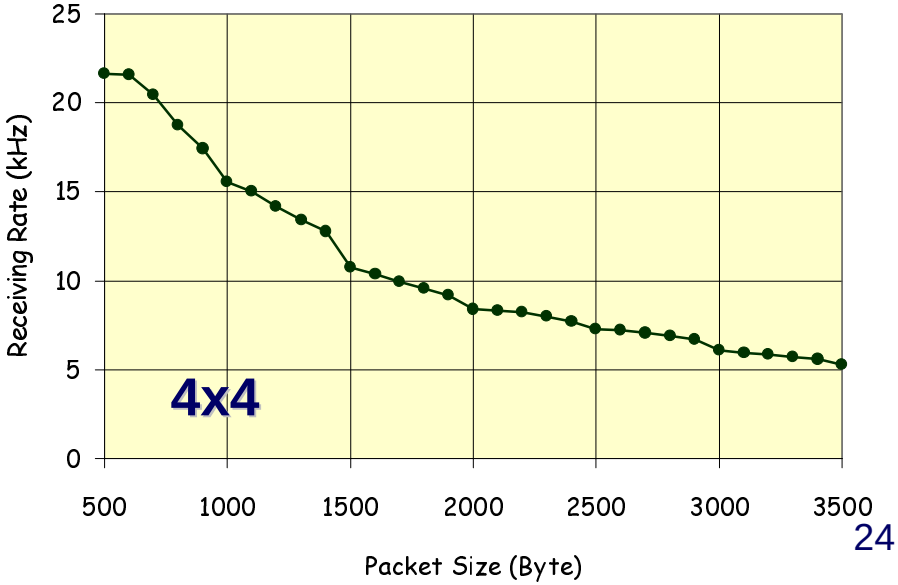


GE Event Building BU Event Rate







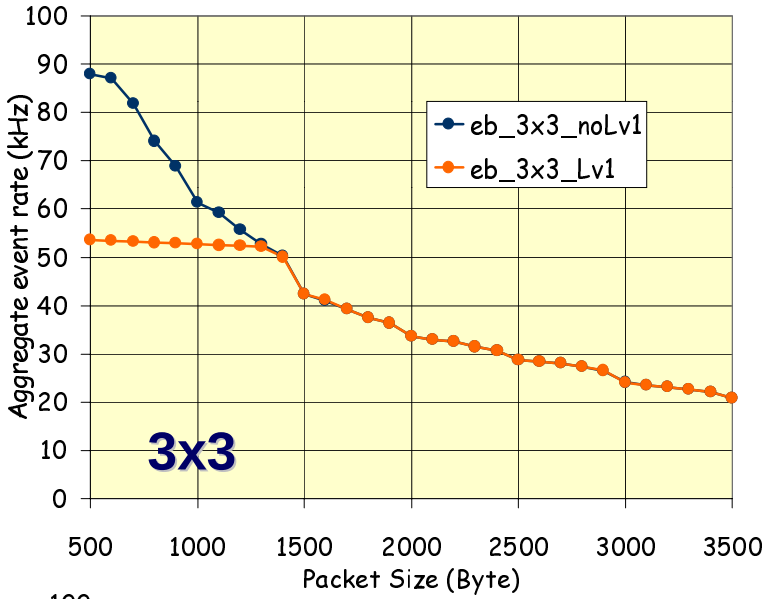


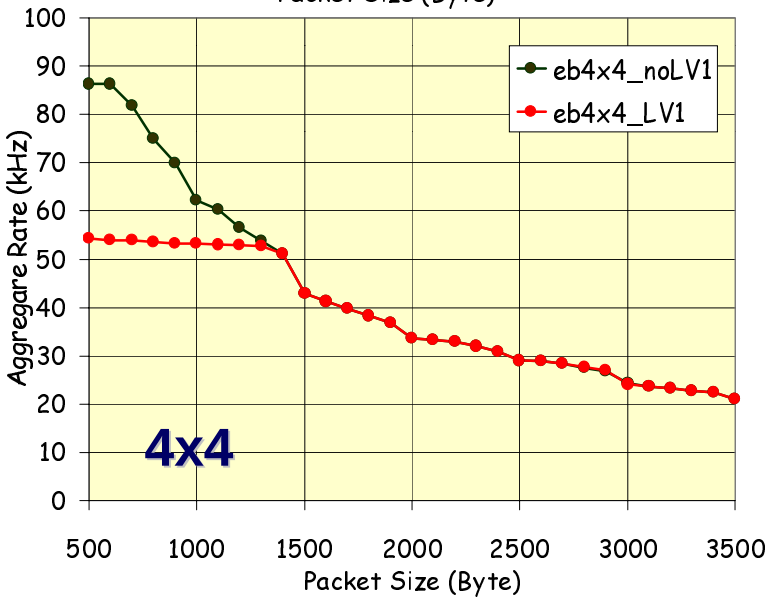
GE Event Building Aggregate Rate

- Output links well balanced
- Througput per node does not change with N (up to 4)

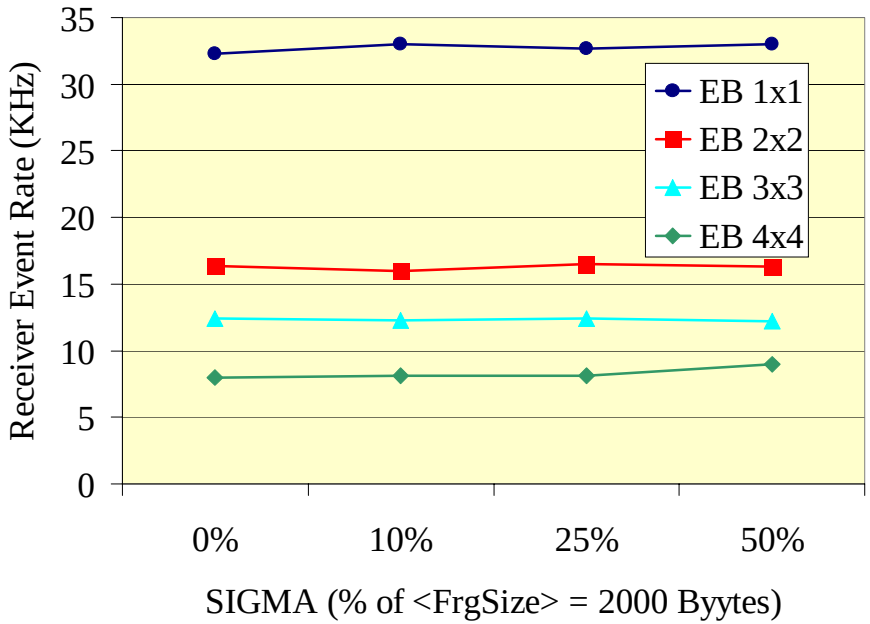
- Aggregate event rate does not change with N (up to 4)

- Lv-1 trigger informations on the same network: affect only small fragment size





Event Builder with Variable Event Size



4x4 GE Event Builder Summary (I)

For 2 kByte Fragment size:
- **event rate 30 kHz**
- **aggregate bandwidth 280 MByte/s**

Main limitation: CPU and PCI IO bandwidth

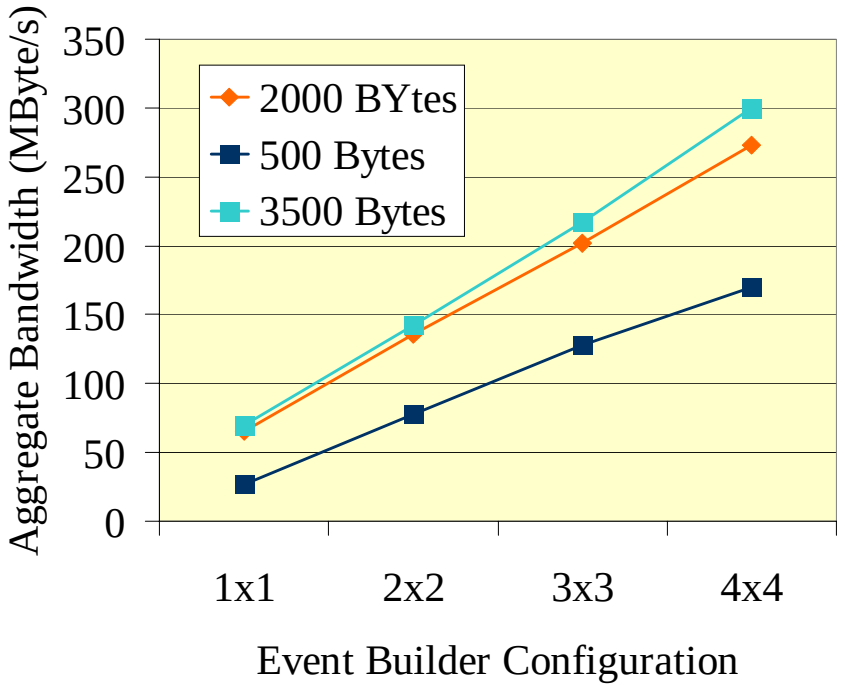
The switch is non-blocking

BDN and BCN in the same network work fine

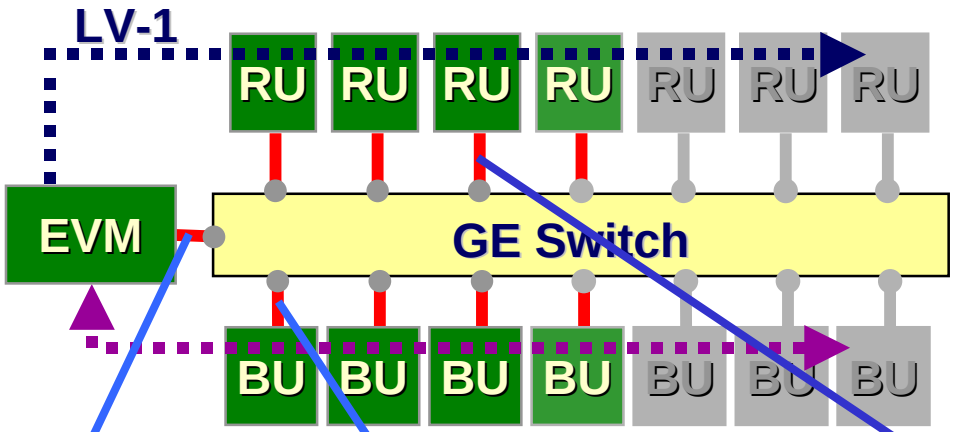
RCN has been also introduced. CPU and IO limits at small fragment size.

No performances impact for event variable size

Extension to multistage switches required detailed simulations



4x4 GE Event Builder Summary (II)



Max Rate per RU (1x4) = 90 kHz
Max Band per RU (1x4) = 85 MB/s

EVM: - 90 kHz (no Lv1)
- 55 kHz (Lv1)

Max Rate per BU (1x1) = 55 kHz
Max Band per BU (4x4) =80 MB/s

Future Work

- Gigaethernet on Copper is coming (1000 Base T). We have a ptp test in progress.

- 16x16 system based on 1000 Base T

- Investigations of asymmetric EVB 500x5000.