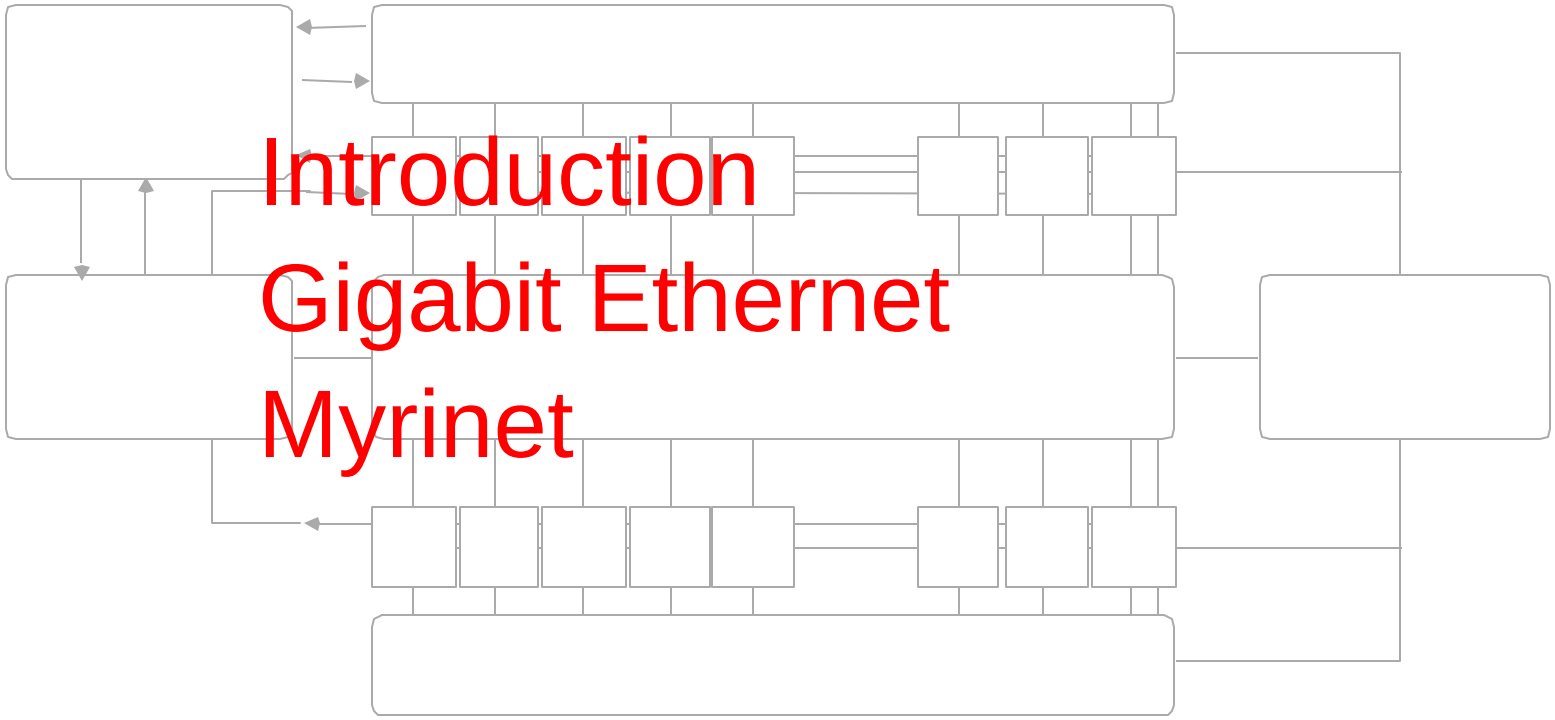
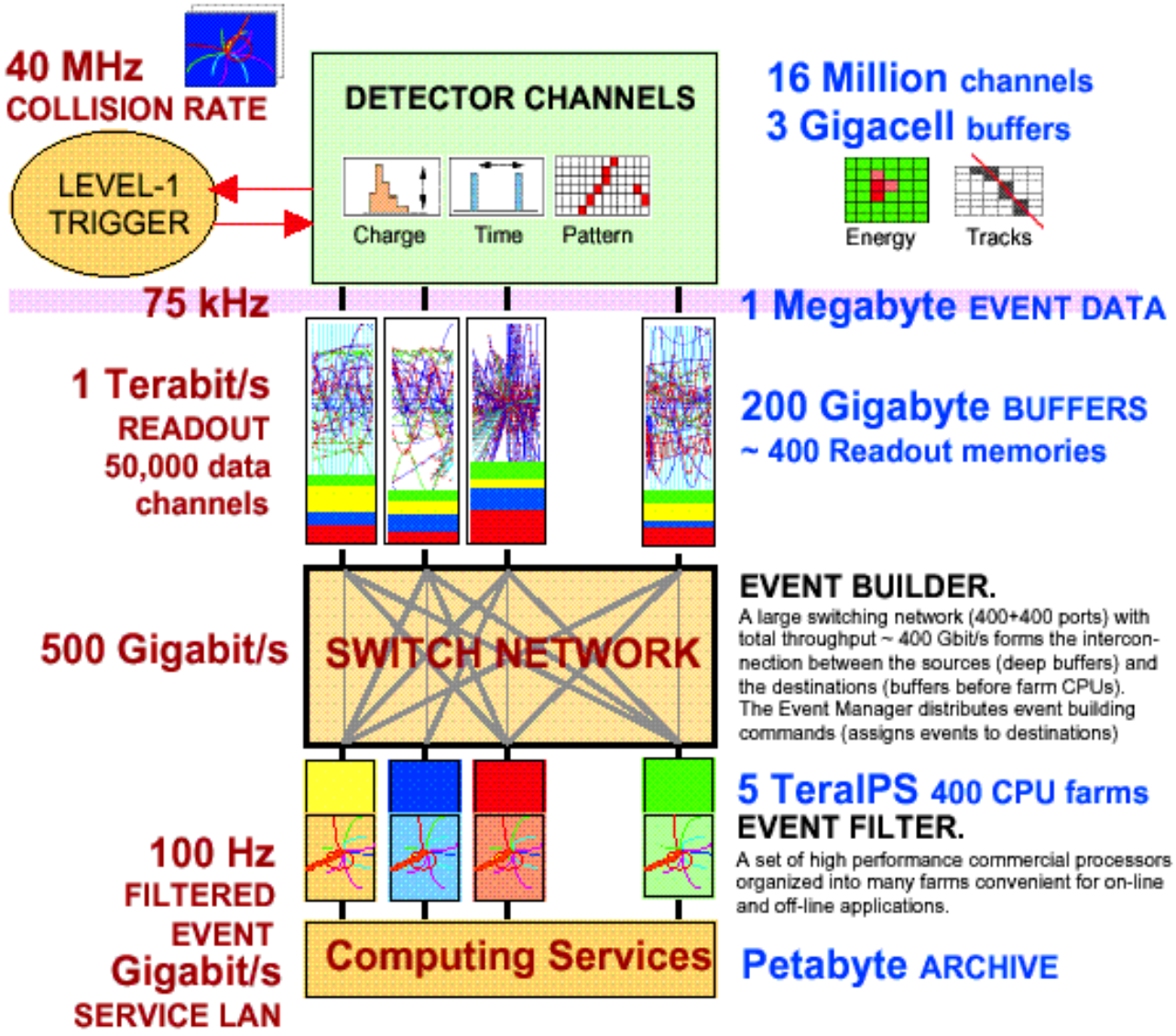
Studies for the CMS Event Builder

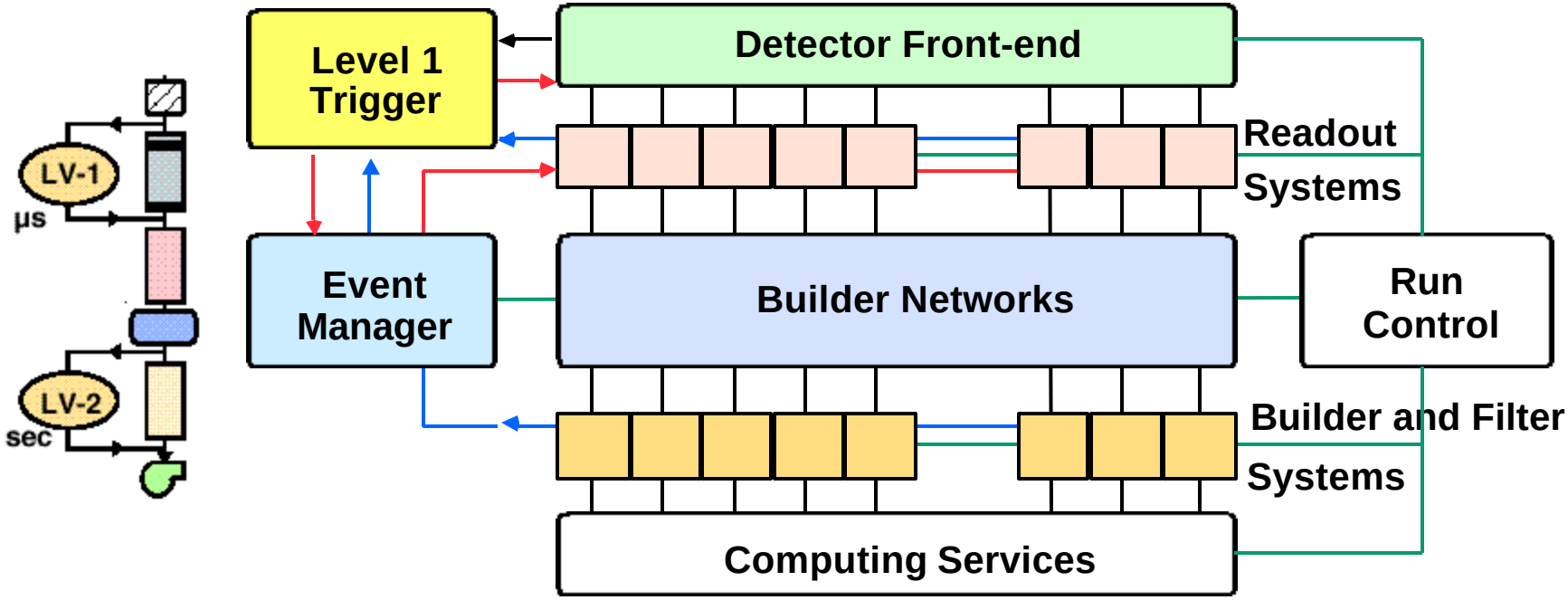
Frans Meijers CERN/EP
on behalf of the CMS DAQ group
DAQ 2000 at NSS-MIC, Lyon, France, 20 Oct 2000



CMS Trigger and Data Acquisition Summary

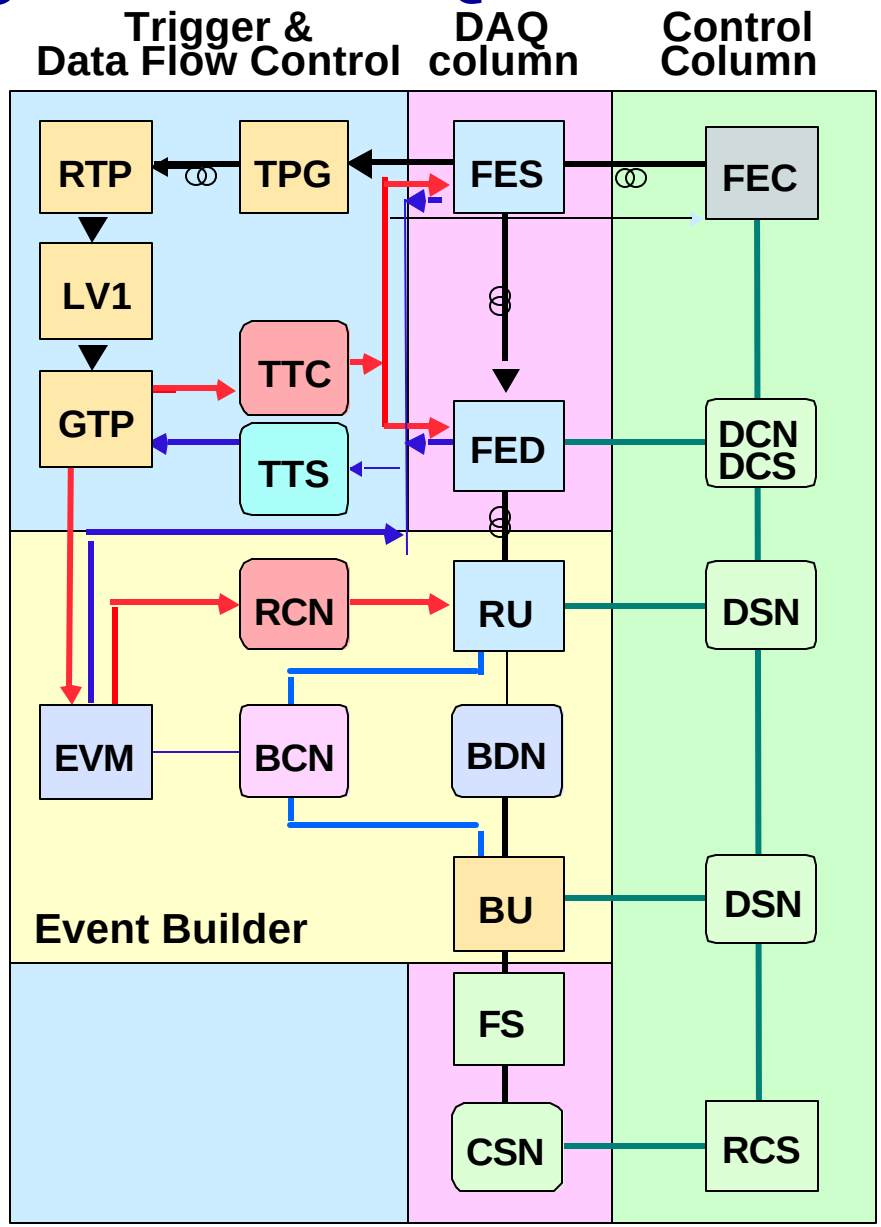


DAQ architecture and EVB parameters

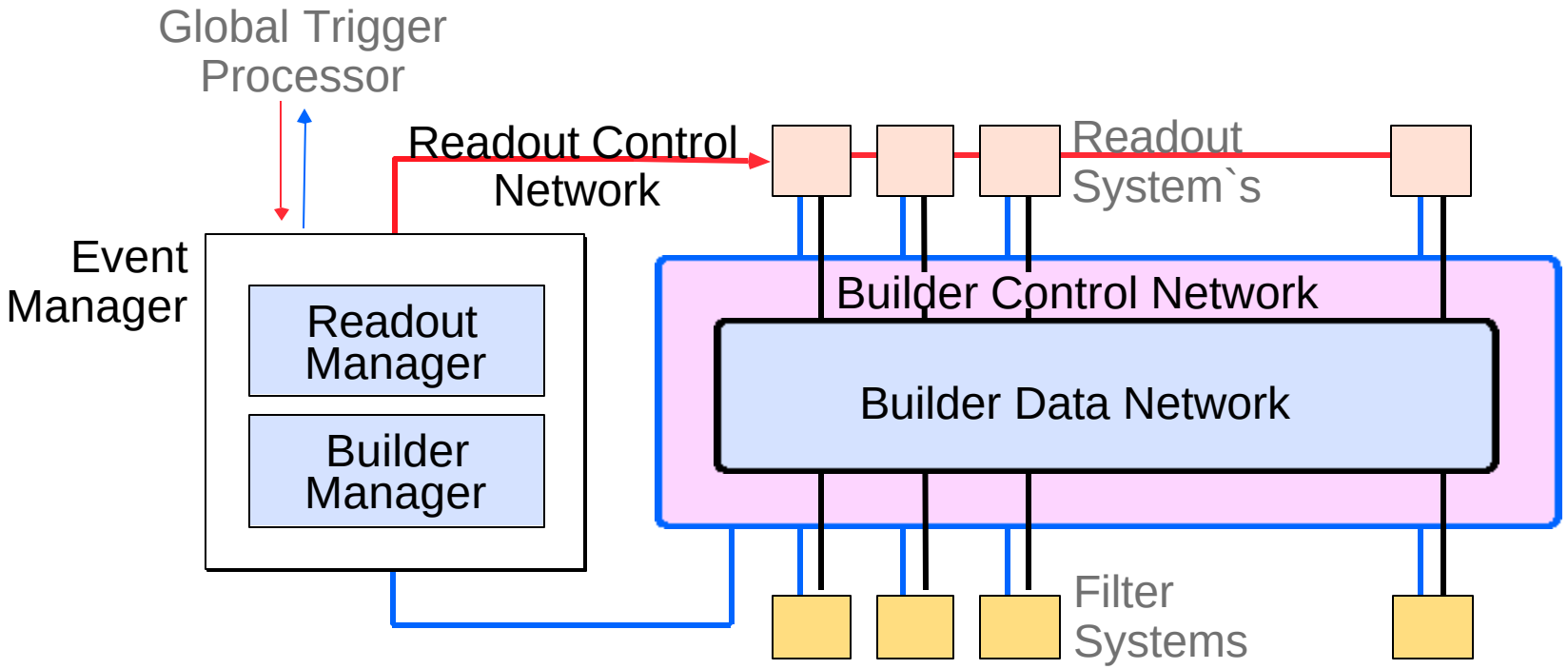


Level-1 Maximum trigger rate	**100 kHz**
High Level Trigger acceptance	**1 - 10 %**
Average event size	**1 Mbyte**
Number of Readout Units	**512**
Average event fragment size	**2 kbyte**
Effective throughput per port	**2 Gbps**
Builder network (512x512 port) aggregate throughput	**1 Tbps**

Trigger and DAQ main subsystems



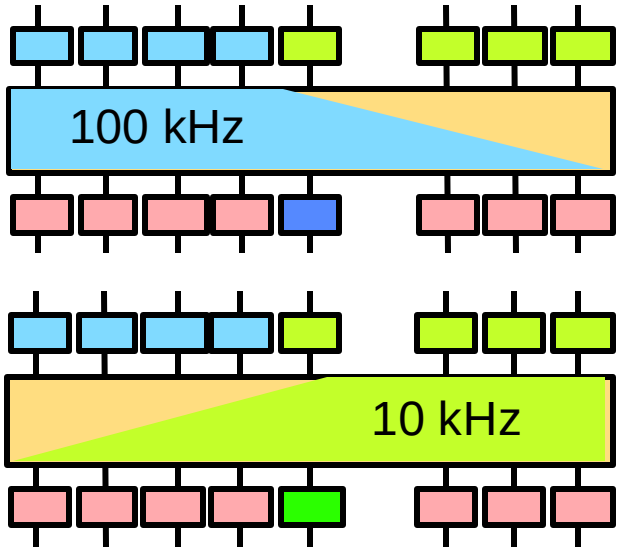
Event Building Networks



Builder Data Network for event data: high throughput required
Builder Control Network for fast control messages

Logical entities: can be implemented (partly) on same physical network

Event Building in Steps



Step 1: at 100 kHz
Rejection factor 10 with 0.25 of the data from High Level Trigger

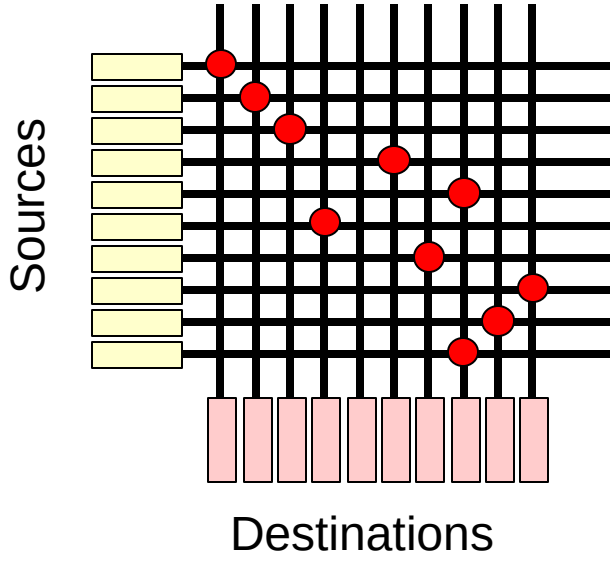
Step 2: at 10 kHz
Remaining 0.75 of the data

Throughput reduced by 0.25+0.1x0.75=0.33, ie factor 3

At the cost of
- control complexity,
- increased latency,
- unbalanced RU inputs

Crossbar Switch

Crossbar: **N**x**N** matrix (space division)



simultaneous transfers between any arbitrary set of sources and destinations

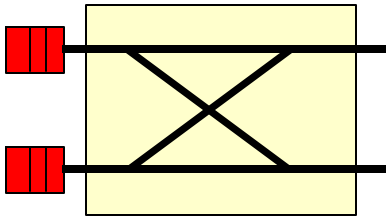
Does not solve **output contention** issue

- The maximum switch load for **random traffic** is about 59%
- Need **traffic shaping** for high efficiency (example: barrel shifter 100%)
- Large crossbar not practical:

→ multistage network

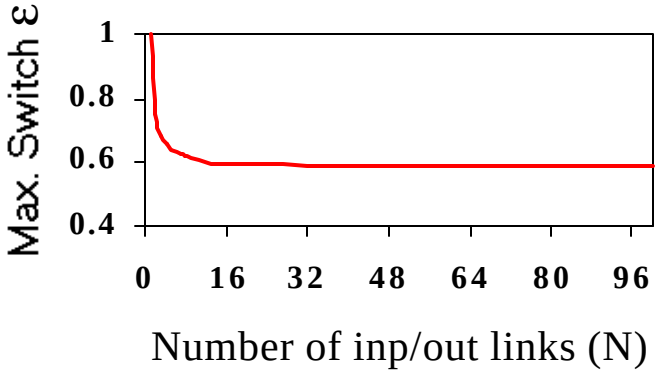
Input versus Output queuing switches

Input Queuing



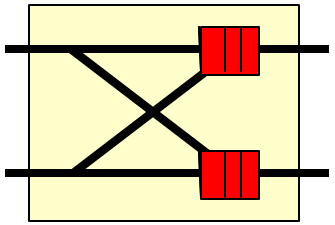
random traffic

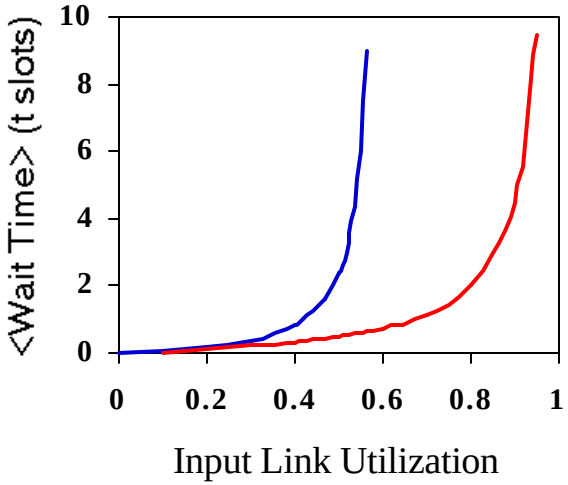
$e \approx 0.59 \quad \text{for } N \rightarrow \infty$



Limited by head-of-line (HOL) blocking

Output Queuing





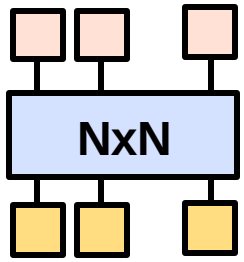
- best performance
- not very scalable

(FIFOs must operate N times faster than input link speeds)

There are other types (CIOQ, shared memory, ..)

EVB DEMONSTRATORS

EVB Demonstrator



- **Small scale prototype**
- **Emulate RU (sources) and BU (sinks)**
- **Evaluate switch technologies for data and fast control**
  - **Gigabit Ethernet**
  - **Myrinet**
- **Study event data acquisition protocols**
- **Extract parameters for simulation of large systems**

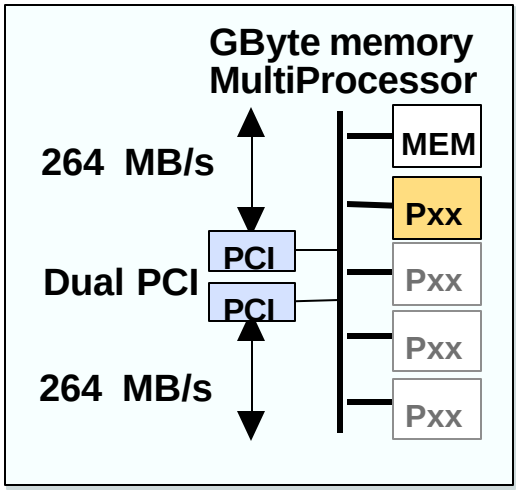
Input for **Technical Design Report** due end 2001

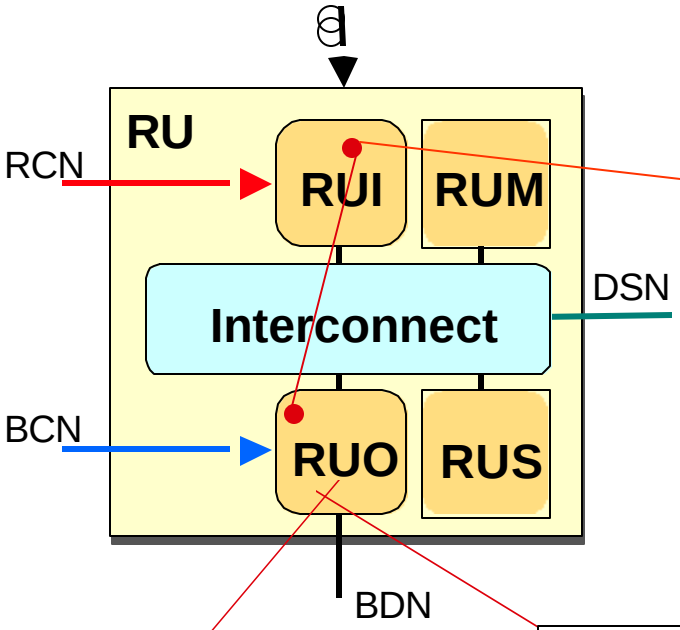
Demonstrator Set-ups

- **CERN**
  - 16x16 + EVM: PCs (Linux)
  - Myrinet, Gigabit Ethernet
- **Legnaro-INFN**
  - 16x16 + EVM: PCs (vxWorks)
  - Gigabit Ethernet
- **UCSD**
  - PCs (Linux), custom BU (vxWorks)
  - Gigabit Ethernet
- **FNAL-MIT**
  - PCs (Linux), EVM and control network

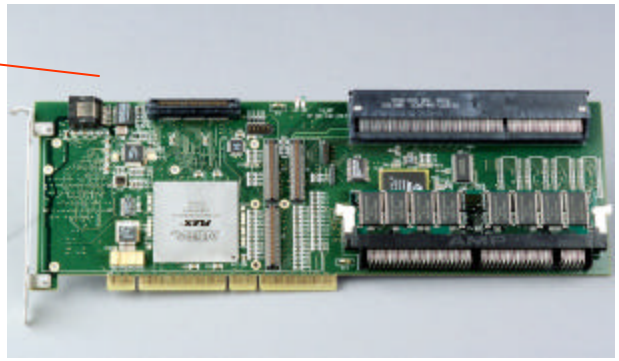
ReadoutUnit H/W S/W implementation

PC based





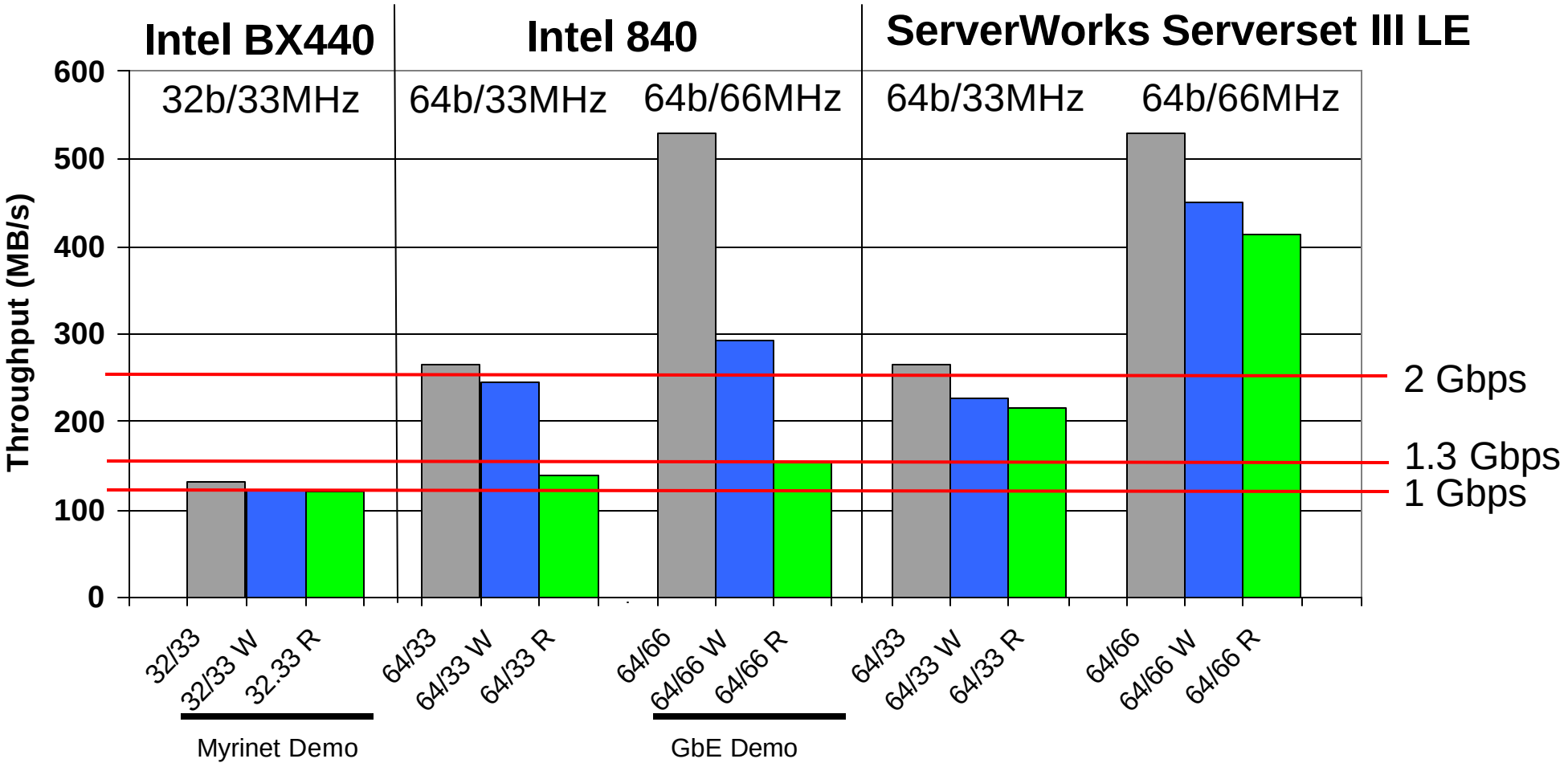
FPGA based



ASIC NIC

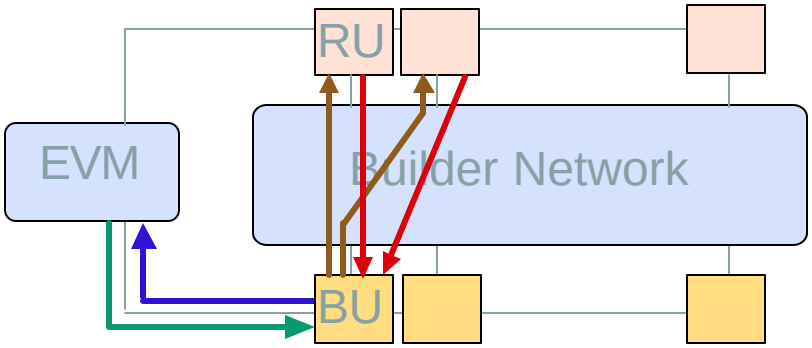
NIC with processor
- programmable firmware
eg Myrinet, Alteon GbE
- custom interface
- for FPGA RU, avoids host intervention per packet

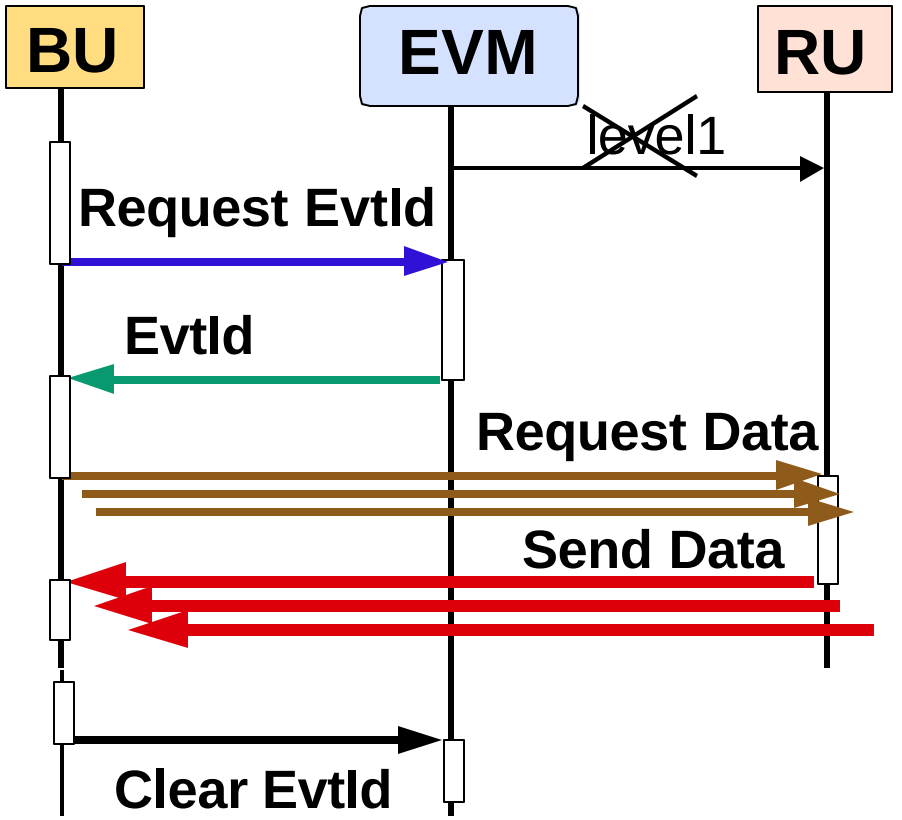
PCI performance on PC



DMA engine: M2M-PCI64A burst: 4kbyte
Motherboards eg: Supermicro PIII DME, Supermicro 370 DLE

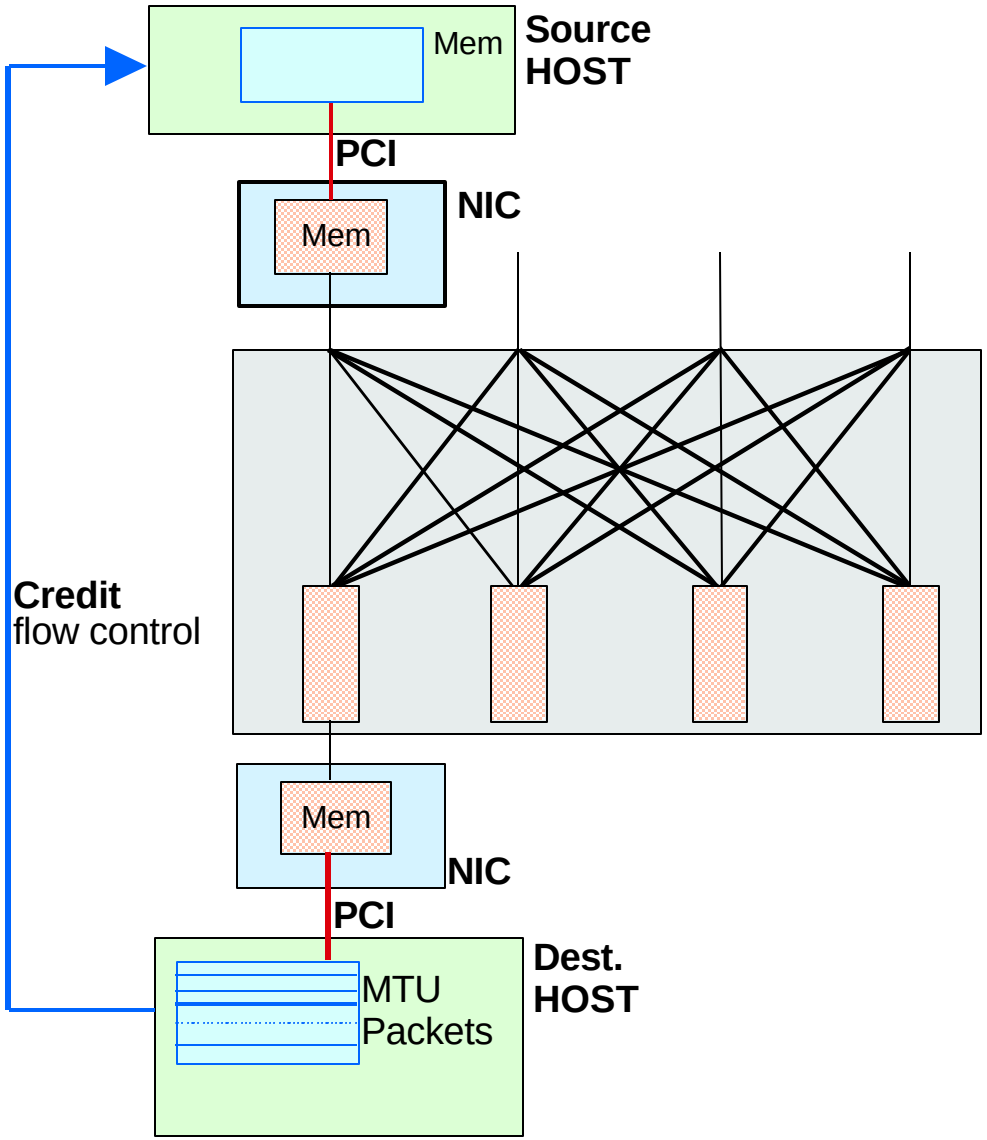
Event building protocol





Several messages can be grouped in a single packet

Flow control



Depending on technology can have packet loss in

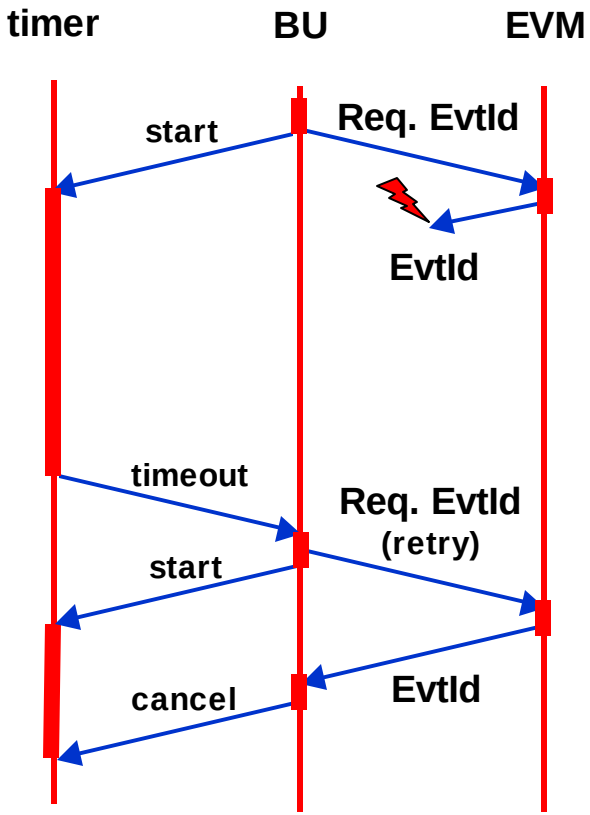
- host
- NIC
- switch

Application flow control to avoid / reduce

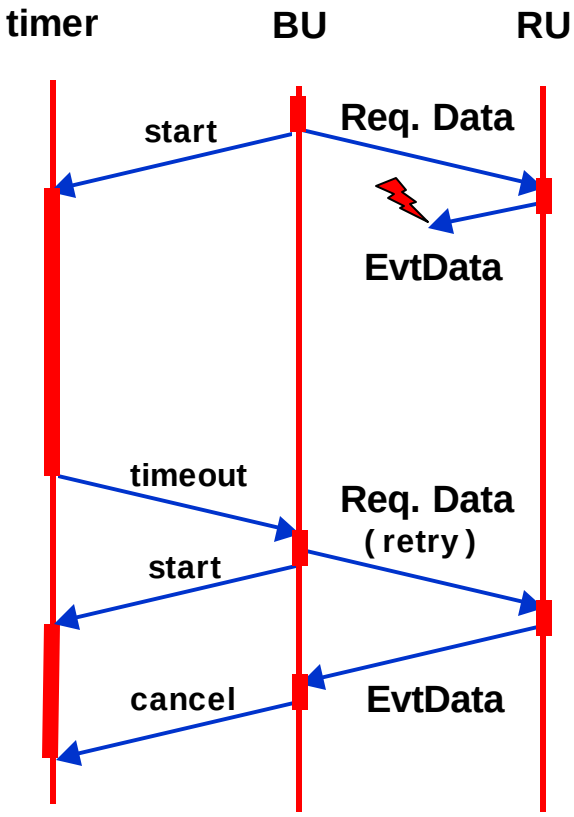
- buffer overflow in receiver
- packet loss in switches

Recovery from lost packets

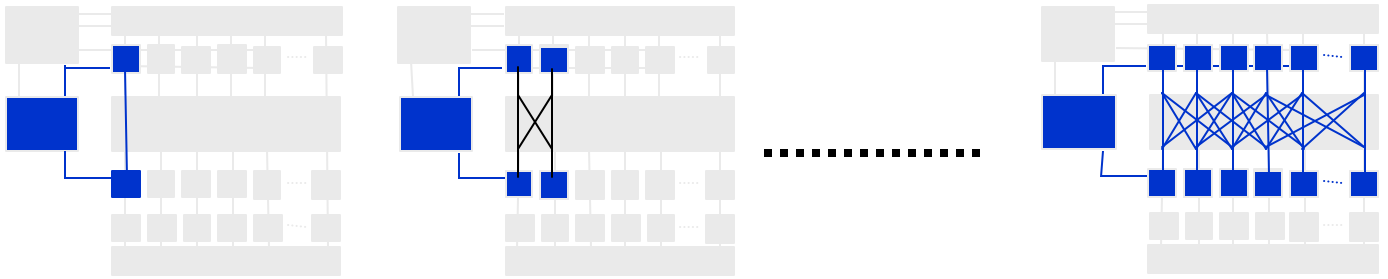
BU – EVM communication



BU – RU communication



EVB observables and scaling



Invariant (if scaling with N):
- Throughput per node (RU or BU) or
- Fragment rate per node = Event rate

- no level-1 implemented
- saturation measurements

GIGABIT ETHERNET

NIC and Switch Evaluation

NIC	fiber	UTP5	custom FW
Packet Engines GNIC II	✓		
Intel Pro 1000	✓		
SysKonnect SK9821		✓	
Alteon ACEnic	✓	✓	✓

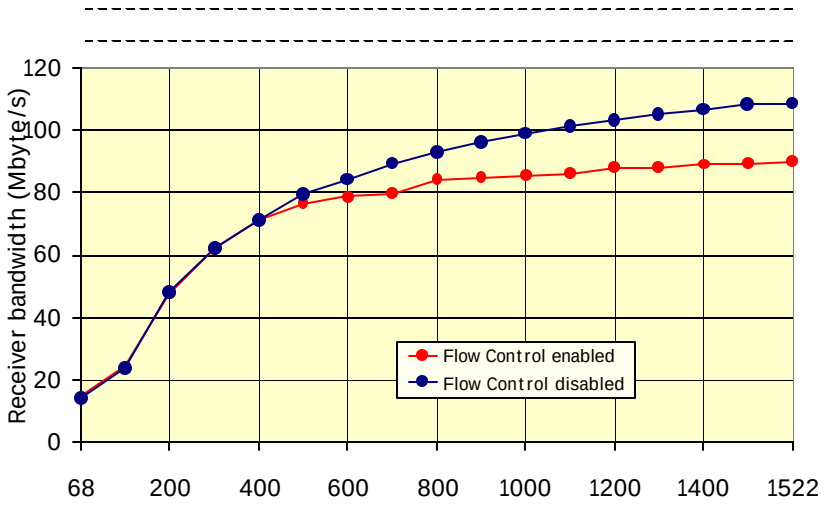
SWITCH	max # ports	fiber	UTP5
Extreme Networks Summit1	8	✓	
Packet Engines PowerRail2200	8	✓	
Intel 6000	32	✓	
Foundry BigIron 8000	64		✓

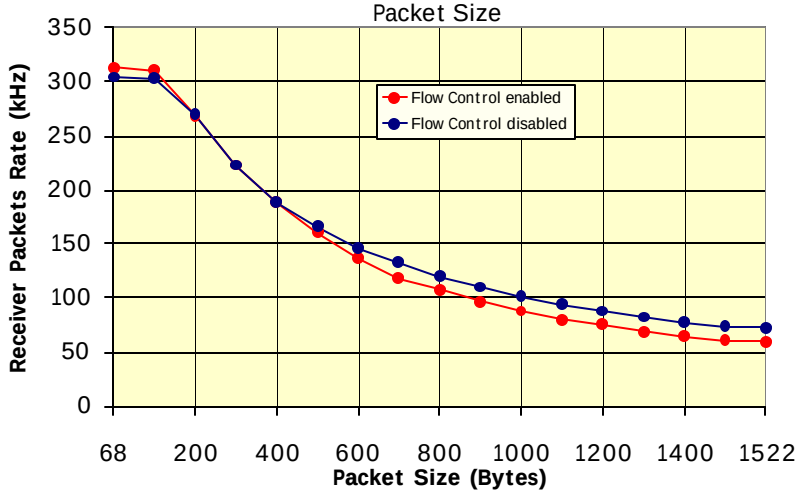
GbE point-to-point (I)

1 sender saturating 1 receiver

PCs with BX PCI 32b/33MHz

PCI → GE →





- throughput up to 90 MB/s

PCI 132 MB/s, GE link 125 MB/s

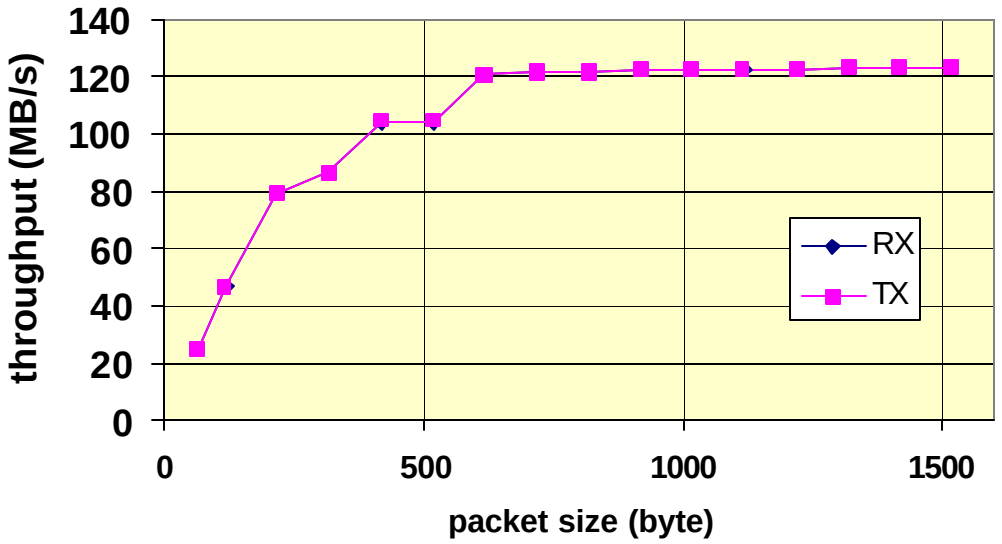
- without flow control: packet loss up to 16%

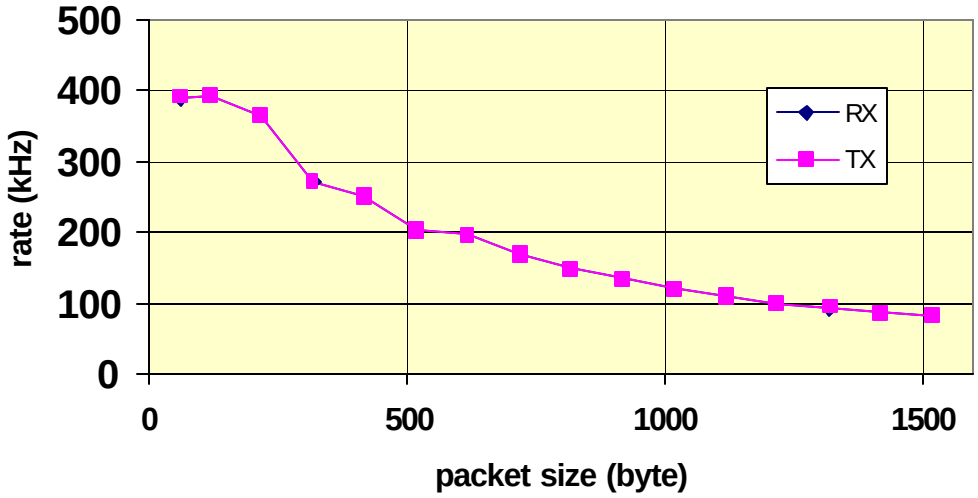
NIC: Intel Pro1000

GbE: point-to-point (II)

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

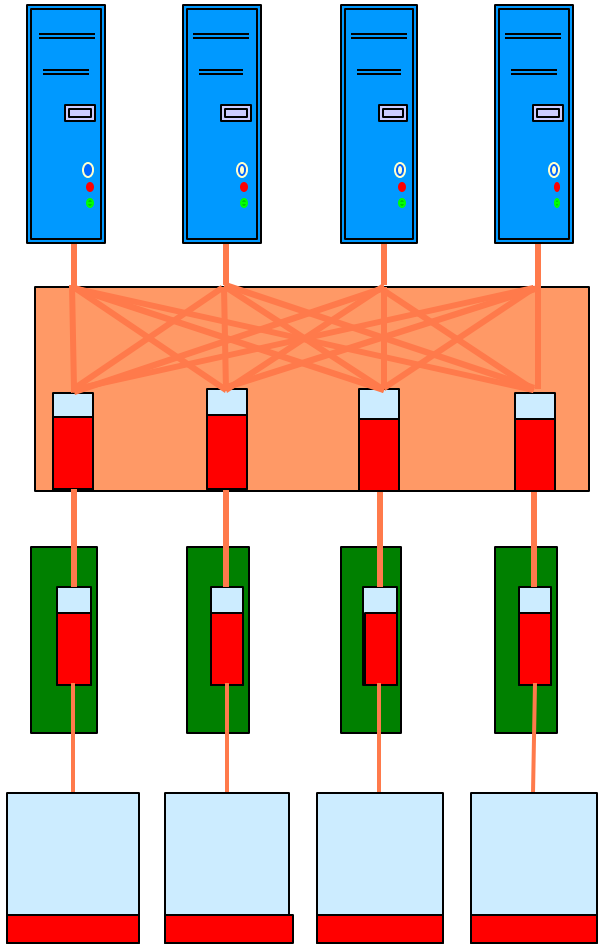
- Throughput up to 123 MB/s
  GE link 125 MB/s
- no packet loss





GbE Switch Tests

- switches have **internal buffer** memory (of finite size)
- **flow control** not always implemented inside switch
- **flow control** reduces throughput for EVB traffic (HOL)
- avoid **packet loss** by credit based protocol (to prevent buffer overflow)
- this method was used to probe buffer size by varying assumed size



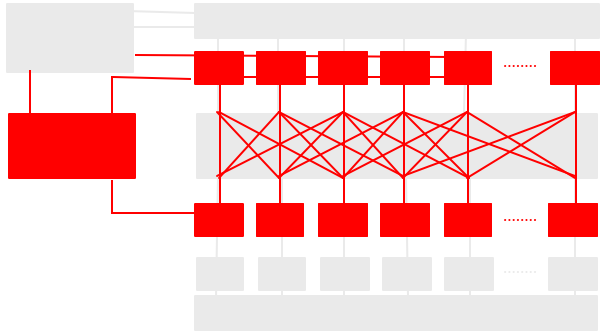
GbE Event Builder

Components:

Hardware:

- switch: FoundryNet BigIron
- NIC: SysKonnect SK9821
- PC: Supermicro PIII (i840)

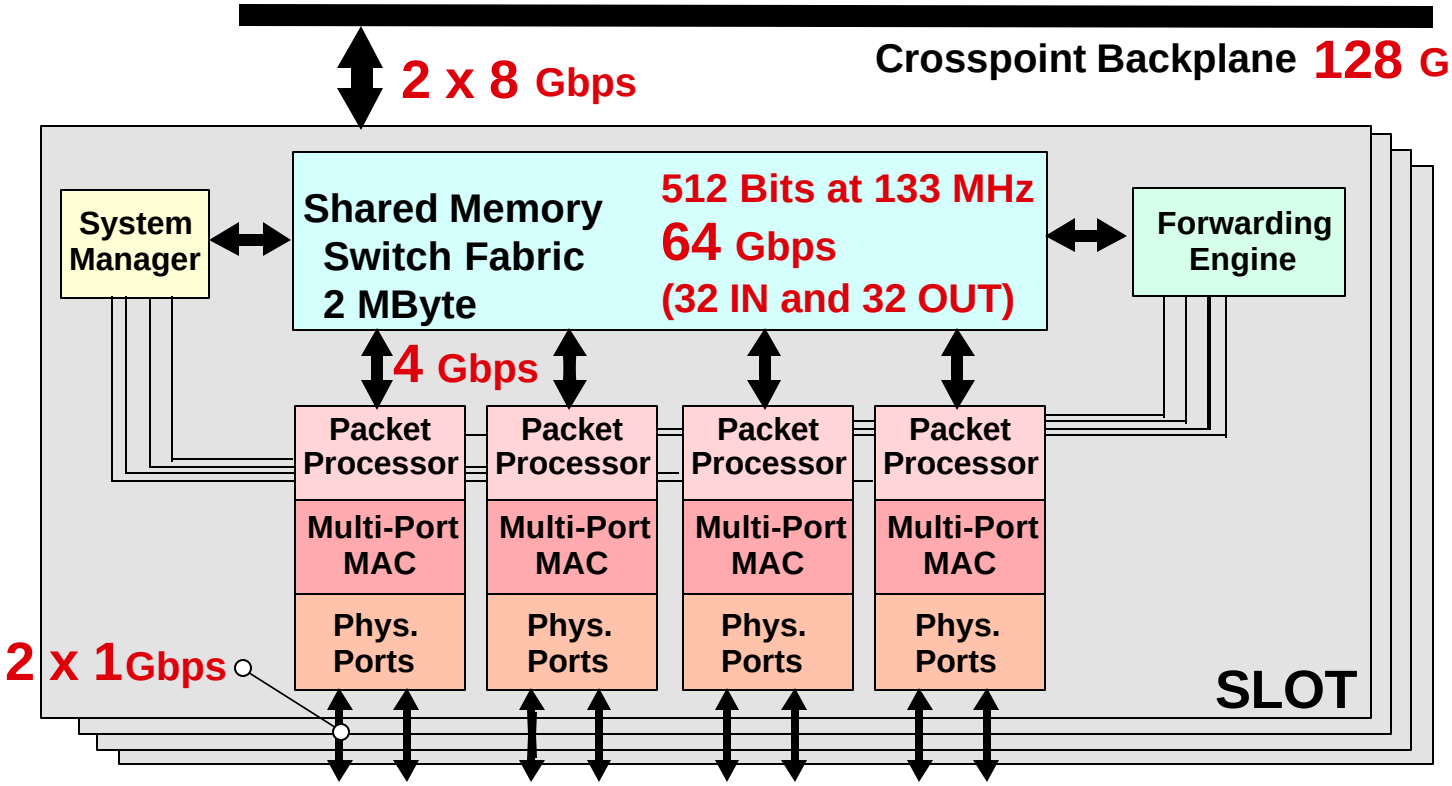
Software: vxWorks based



Test conditions:

- no command or event aggregation (each packet transports a command or data frame relative to a single event)
- full data transfer from/to PC memory
- recovery from packet loss
- fixed fragment sizes are varied 400-4000 bytes

BigIron Architecture



-- **Switch Fabric:**

non-blocking : **32 Gbps** is twice the aggregate total port bandwitdh of the 8 physical ports (16Gbps)

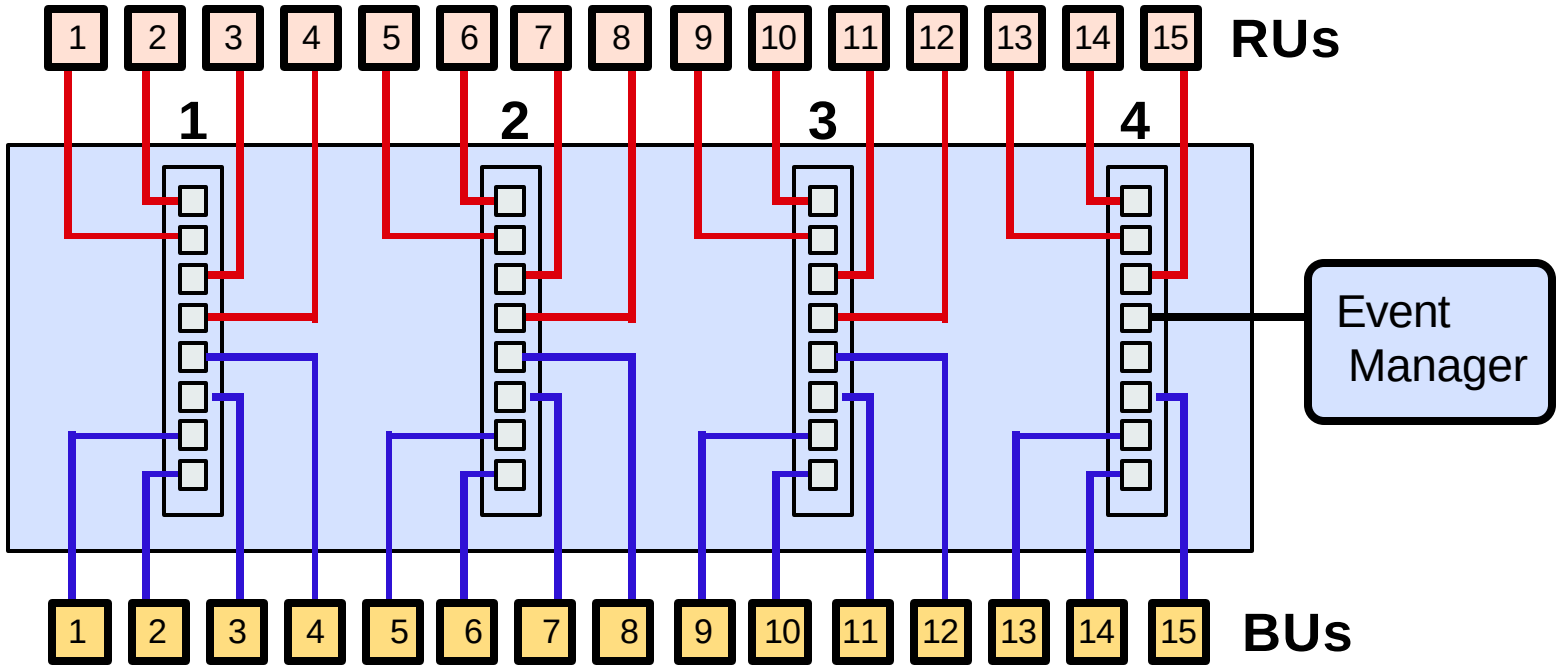
-- **Crosspoint Backplane Connection:**

bottleneck ? **128 Gbps** = 8 slots bandwidth (16Gbps) connection to the backplane

-- **NO Head of Line Blocking:**

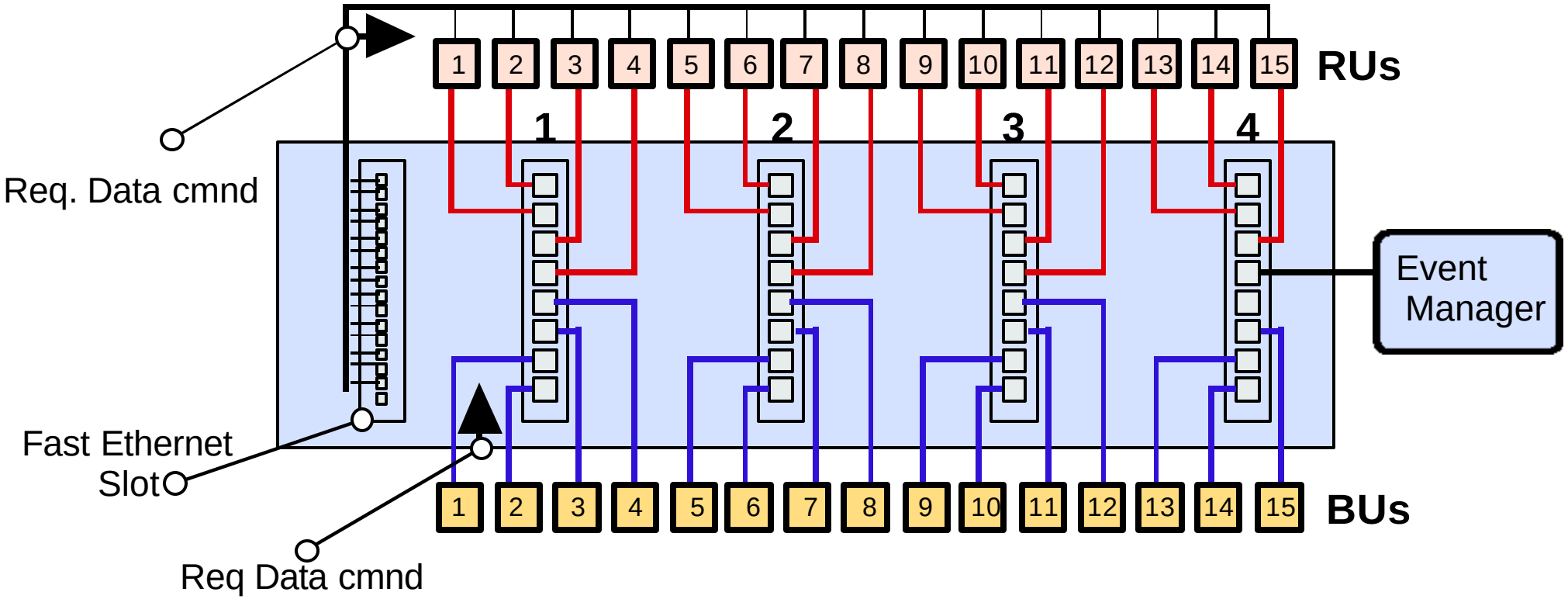
multiple detination output priority queues + multiple input source buffers per output port

EVB layout



RUs and BUs distributed in all switch slots
→ part of traffic localised within slot
→ reduces switch backplane utilisation

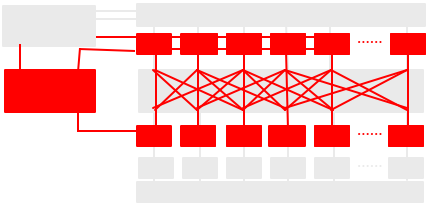
EVB layout - modified

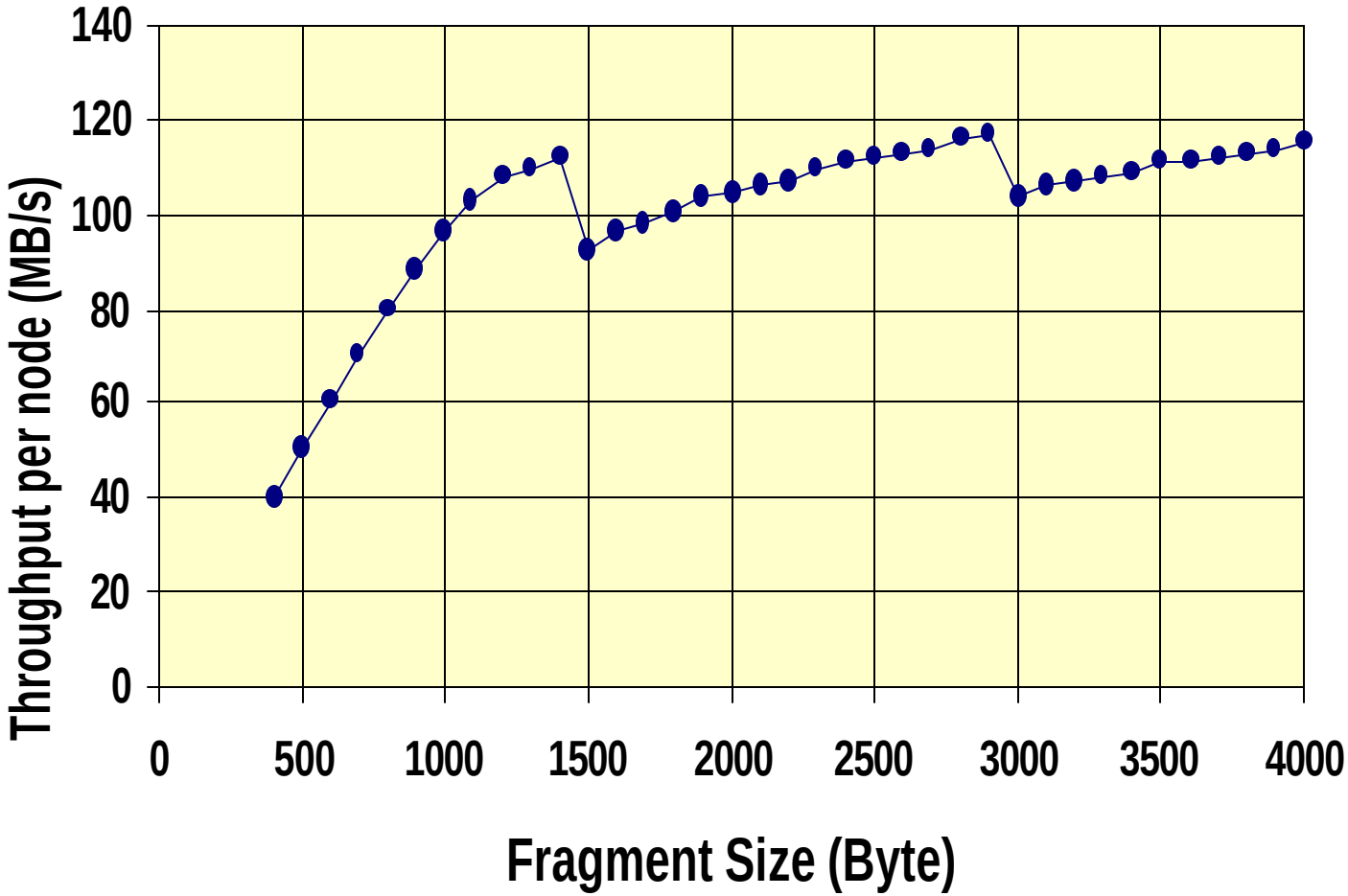


Part of fast control messages over Fast Ethernet
Separates on RUs

- data transfer on 64b/66MHz PCI
- fast control on 32b/33MHz PCI

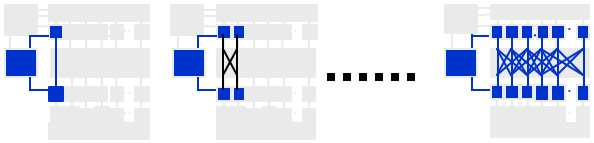
EVB 15x15 performance - Throughput

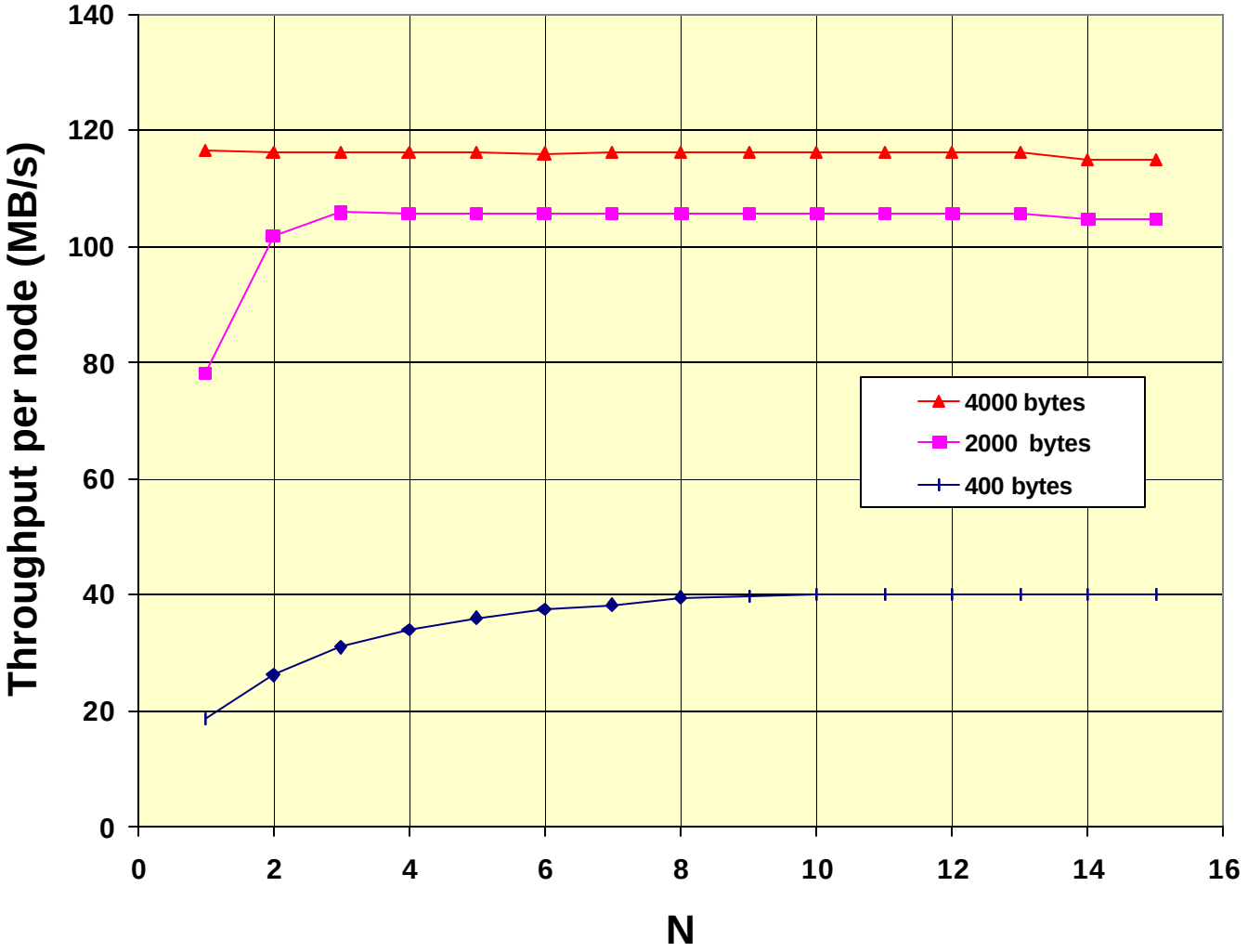




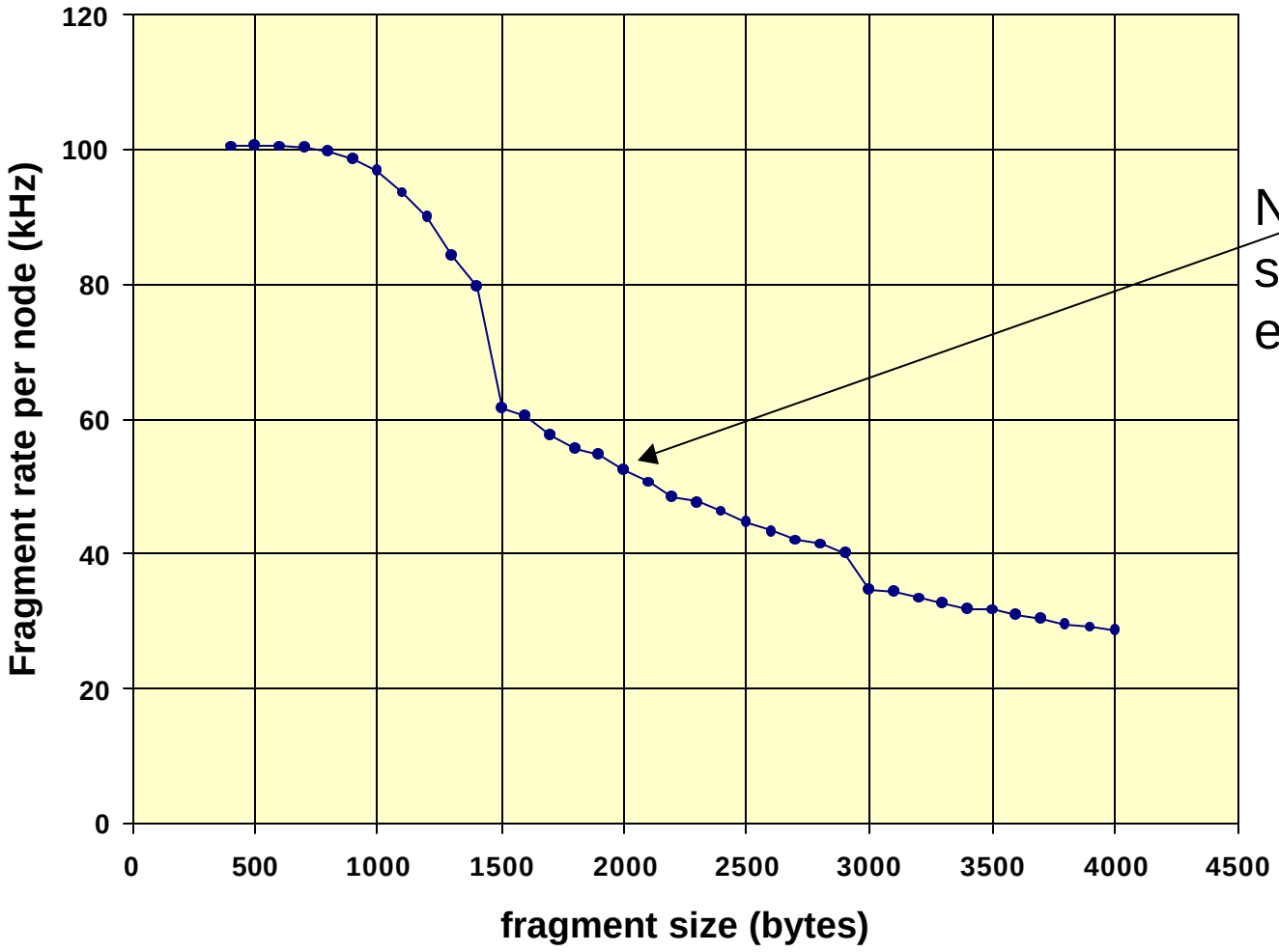
- Throughput up to 116 MB/s, ie 93% link speed
- sawtooth due to MTU
- no packet loss observed (as expected)

EVB scaling

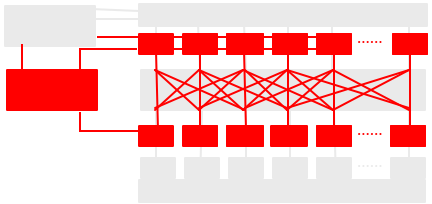



Throughput per node (MB/s)
140
120
100
80
60
40
20
0
0
2
4
6
8
10
12
14
16
N
4000 bytes
2000 bytes
400 bytes

EVB performance - Event Rate



Nominal fragment size 2kbytes:
event rate = 52 kHz

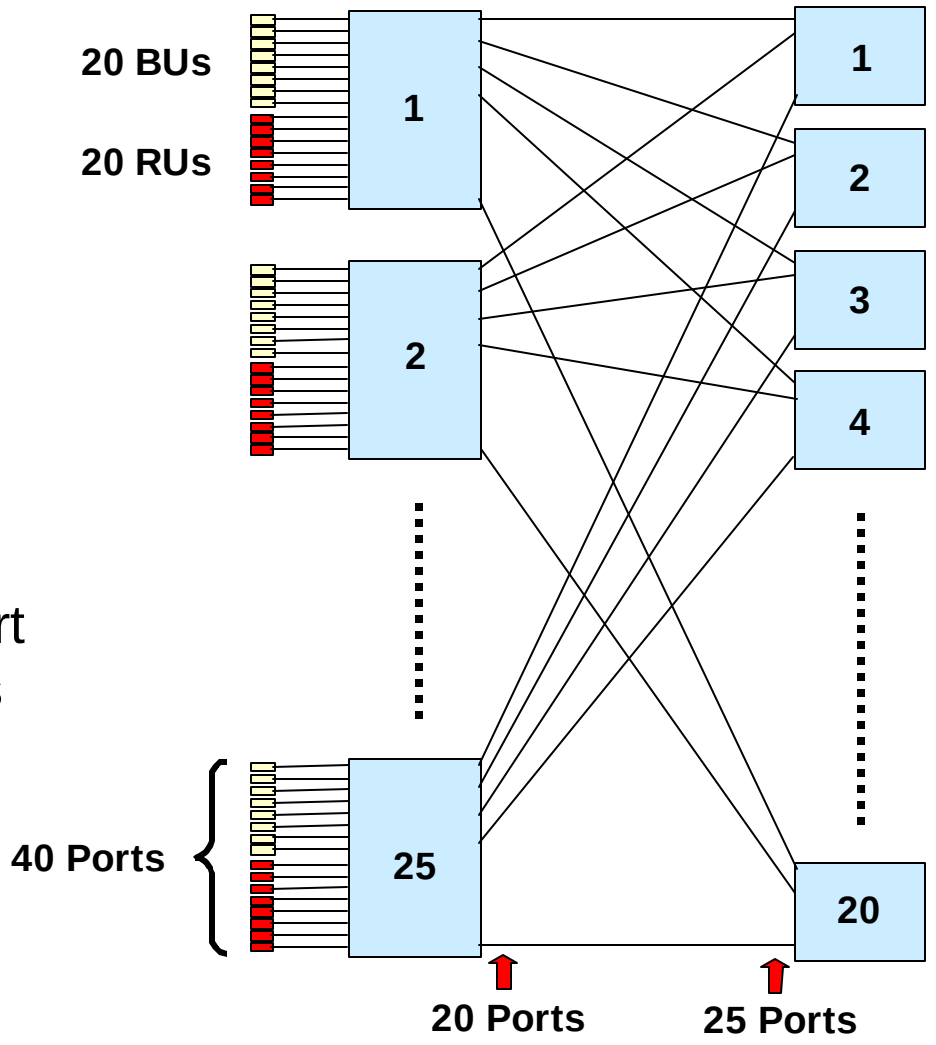


15 x 15

Large (500x500) multistage GbE network (I)

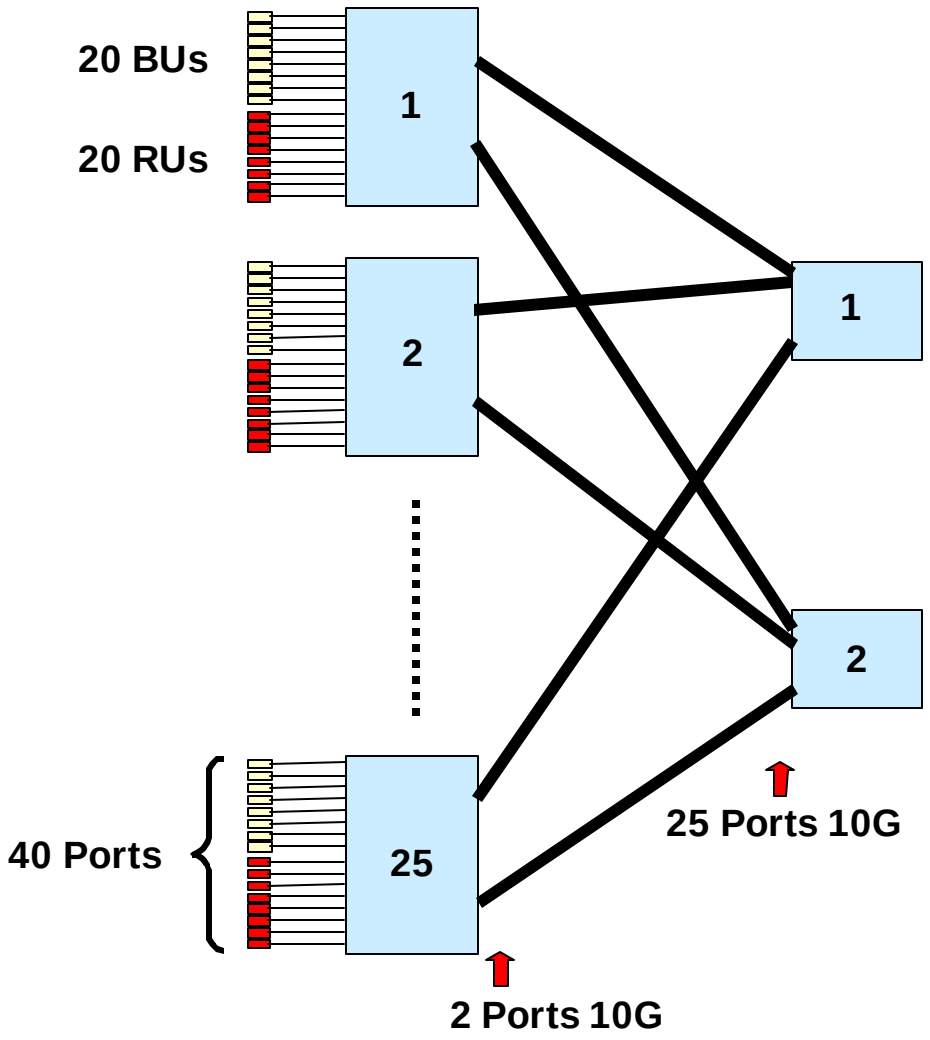
- 25 switches with 60 x 1Gb ports
- 20 switches with 25 x 1 Gb ports

To have 2 Gbps per RU/BU port
→ duplicate network and NICs



Large (500x500) multistage GbE network (II)

- 25 switches with
  40 x 1Gb ports +
  2 x 10 Gb uplinks
- 2 switches with
  25 x 10 Gb ports



Gigabit Ethernet - Summary

- **Switches and NIC evaluation**
  - Switches are fully non-blocking (within memory buffer limits)
  - Switches differ on internals, relevant to us
- **Event Building in Prototype (15x15)**
  - Throughput up to ~90% link speed, scaling, no packet loss
  - For 2 kbyte fragment size, event rate 50 kHz
- **Large (500x500) Multistage Switch**
  - Routing in multistage switch
  - Size of switch internal buffers
  - Can not use flow control inside and in between switches
  - How to prevent / handle packet loss ?
  - Need detailed simulation

Myrinet

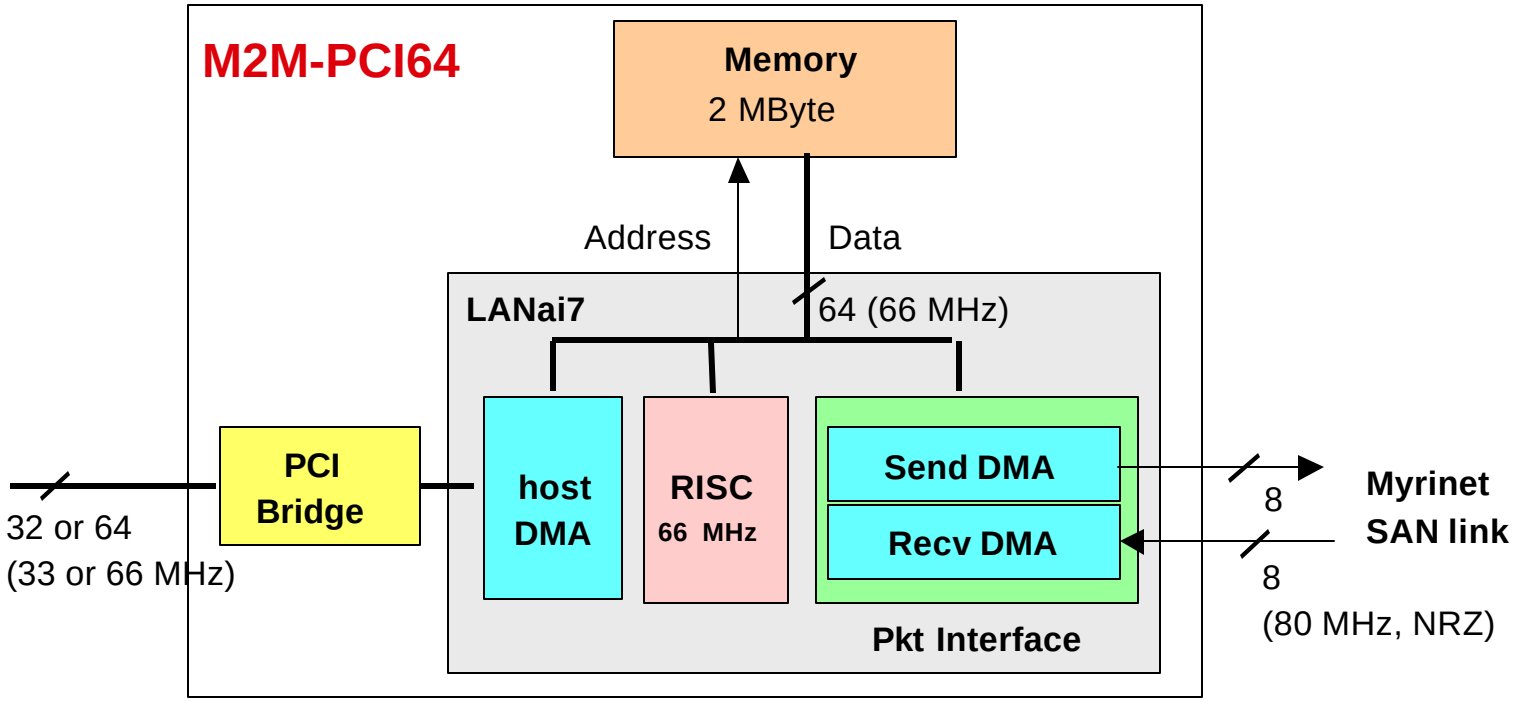
Myrinet features

- Myrinet typically used as **cluster interconnect**
- **point to point links**, byte wide, full-duplex, 1.3 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
- 3Q 2000: link speed from 1.3 Gbps to 2.0 Gbps (Myrinet 2000)

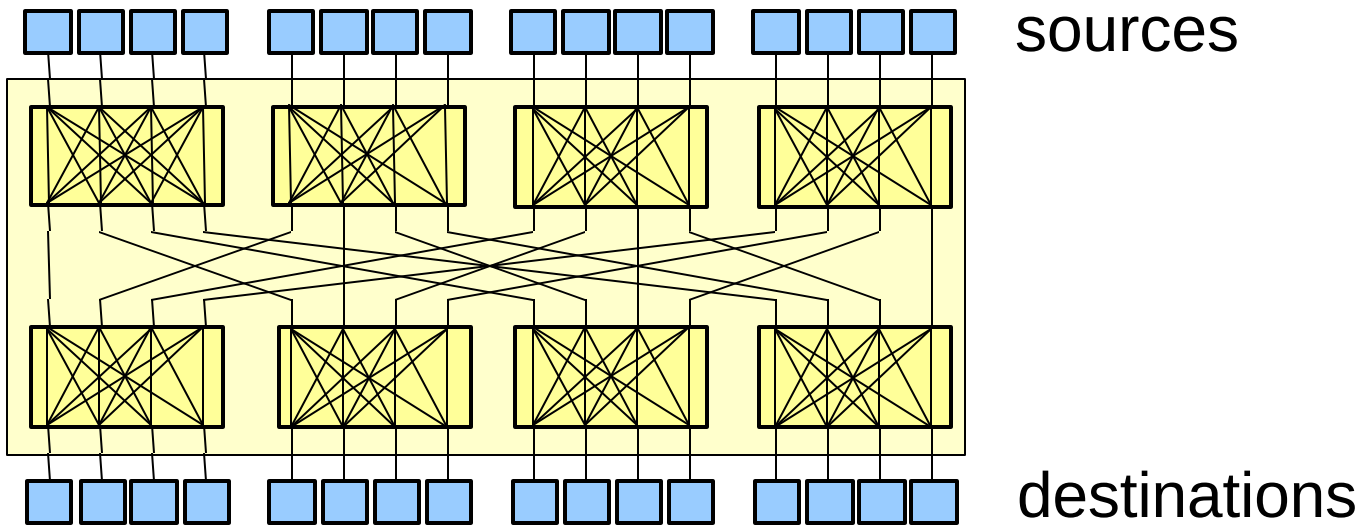
Network interface card



Developed a custom Myrinet Control Program

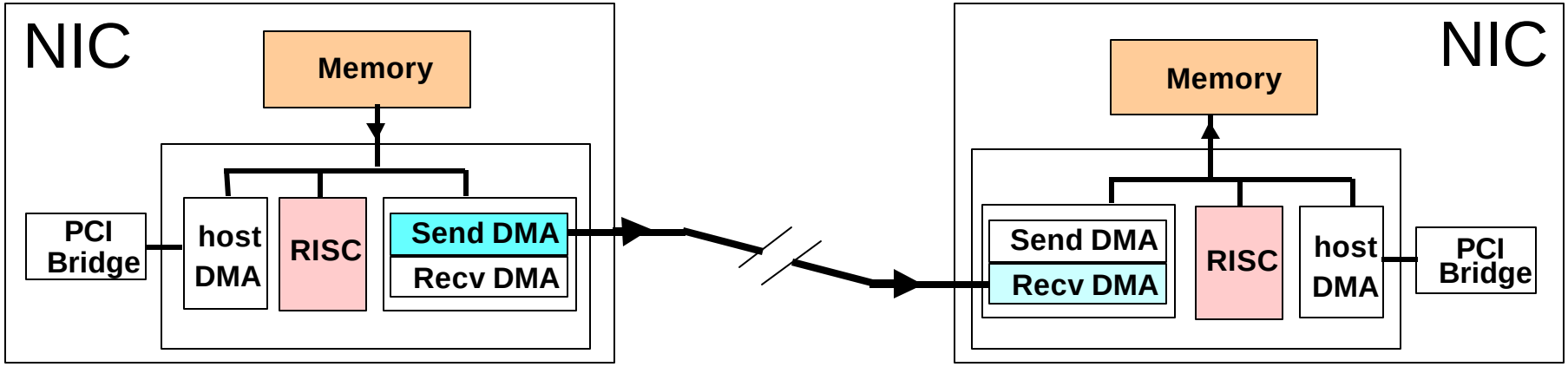
- controls DMA engines
- implements low-level communication protocol

Set-up for NIC and Switch Tests



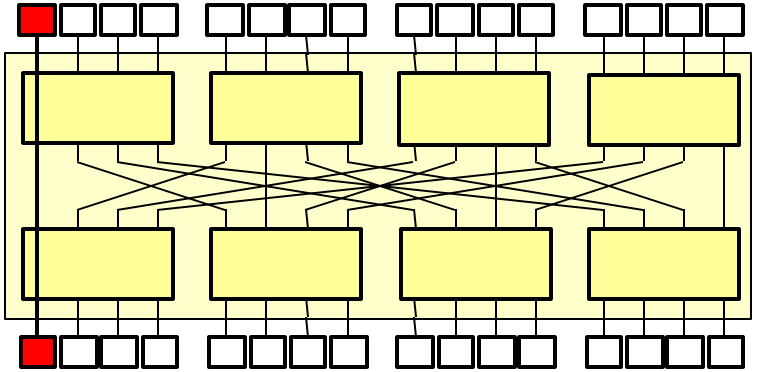
- 32 nodes Linux PCs
- PC: 450/700 MHz PII BX PCI 33 MHz/32bit
- Myrinet switch: M2M-OCT-SW8, NIC: M2M-PCI64[A]
- two-stage Banyan (Delta) network out of 4x4 crossbars (Xbar8)

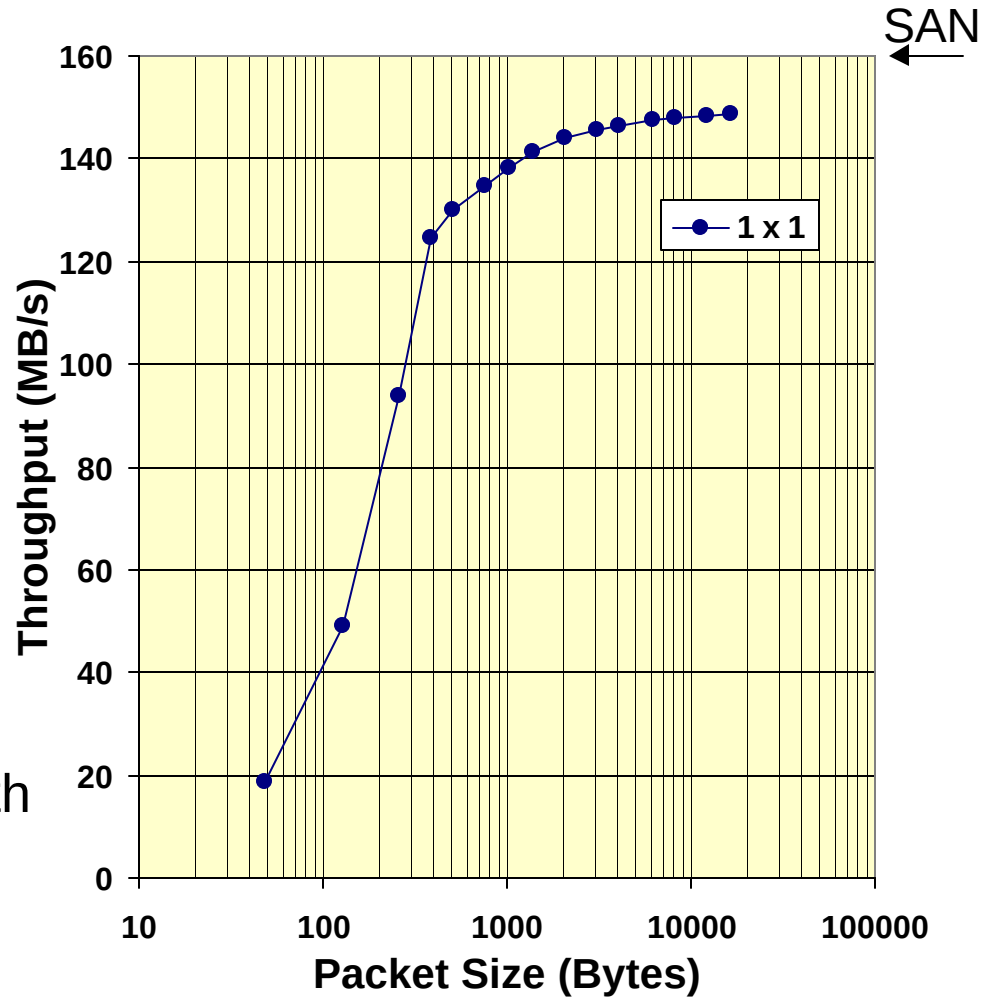
Tests Conditions



- Custom firmware (MCP)
- Packets moved from/to NIC memory
- Host collects statistics
- No packet loss
- Allows to load switch to maximum
- Measure throughput at all NIC sources/destinations
- Saturation measurement
- Packet sizes varied 46 - 16384 bytes

Point-to-point 1x1





1 source saturating 1 destination

Throughput approaches SAN bandwith

Parameters point-to-point 1x1

time per packet = *offset* + **size** / *speed*

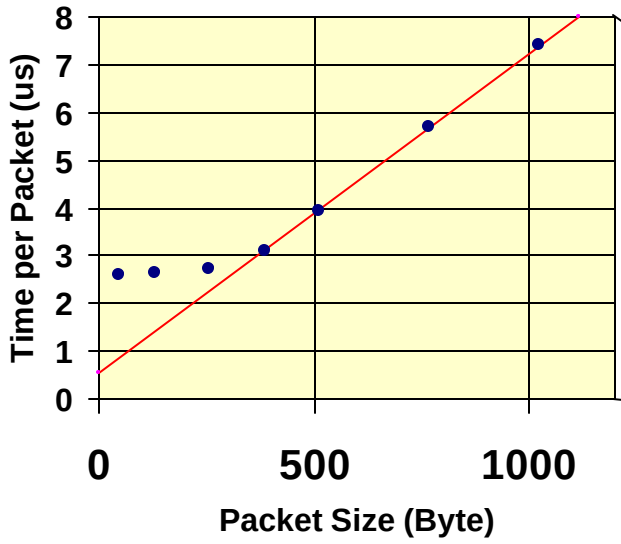
- > 400 bytes: **linear behaviour:**

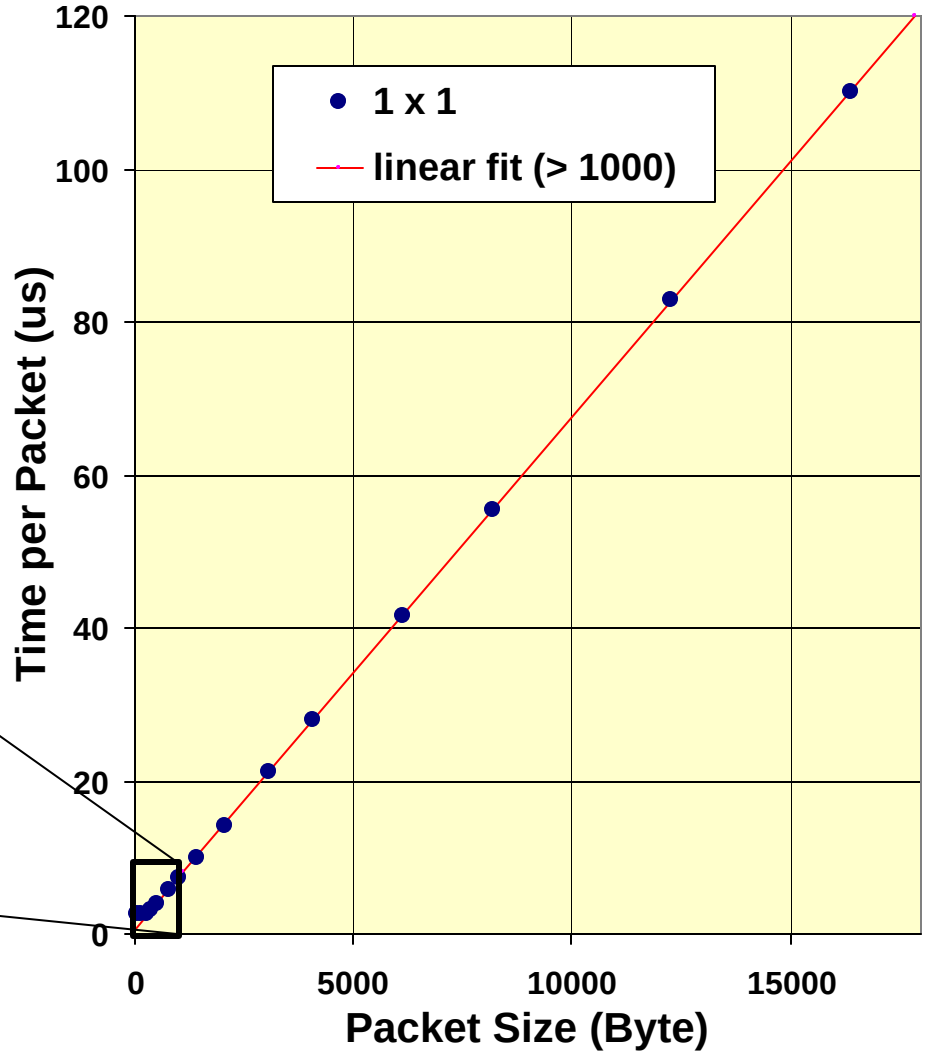
speed: 149 Mbyte/s; 93% SAN

- < 400 bytes:

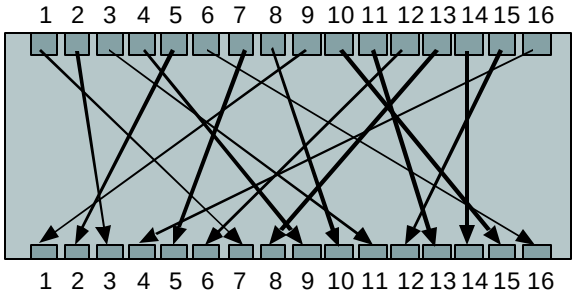
offset : 0.5 µs

plateau**:** 2.7 µs (overhead)





Random Traffic

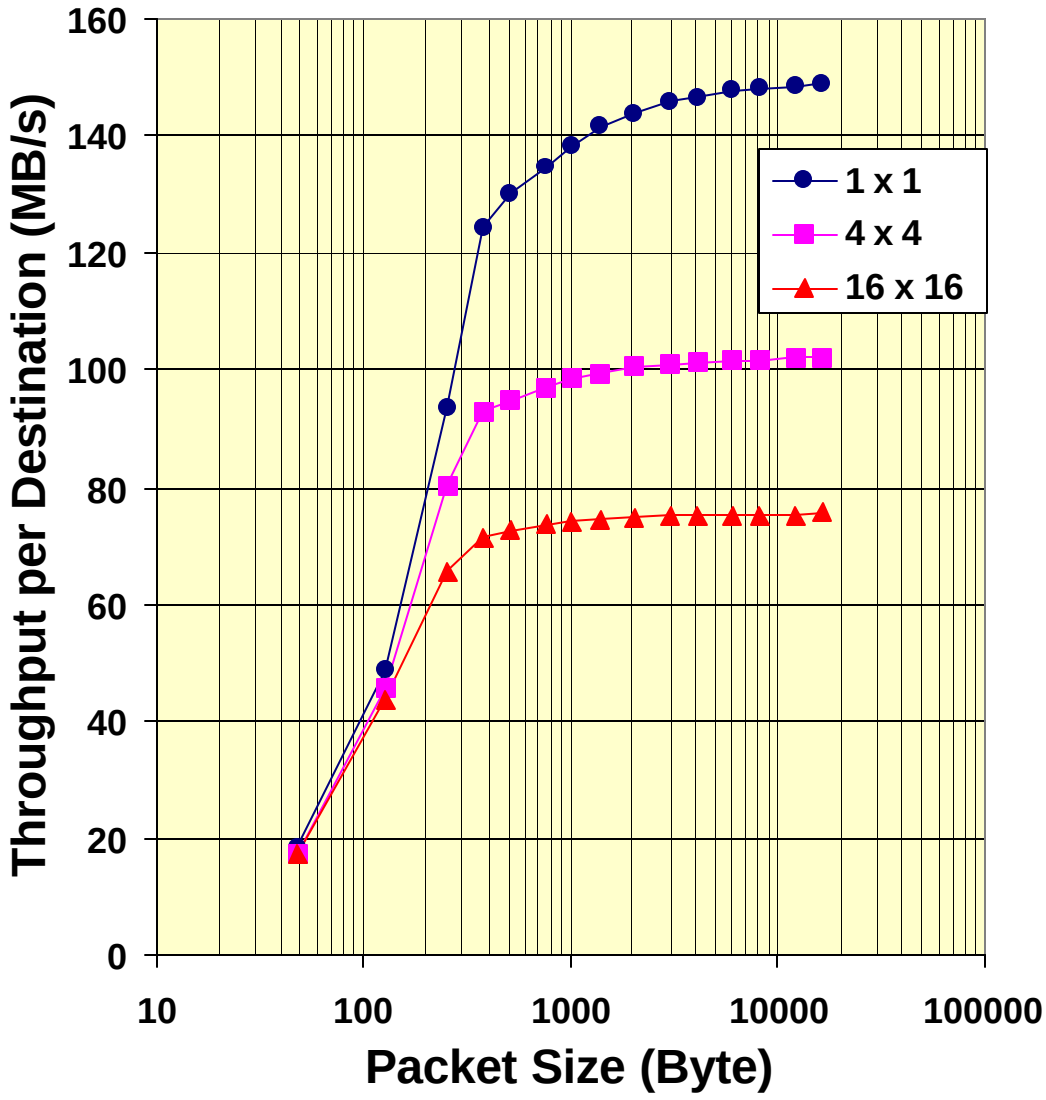


sources send, independently, to a **random destination** according to a uniform distribution

Efficiency:

4x4:	69%	expect	66%
16x16:	51%		[48%]

limited by head-of-line blocking



Event Building with Myrinet

- high link bandwidth
- very low bit error rate
- arbitrary length packets
- no packet loss inside switching fabric

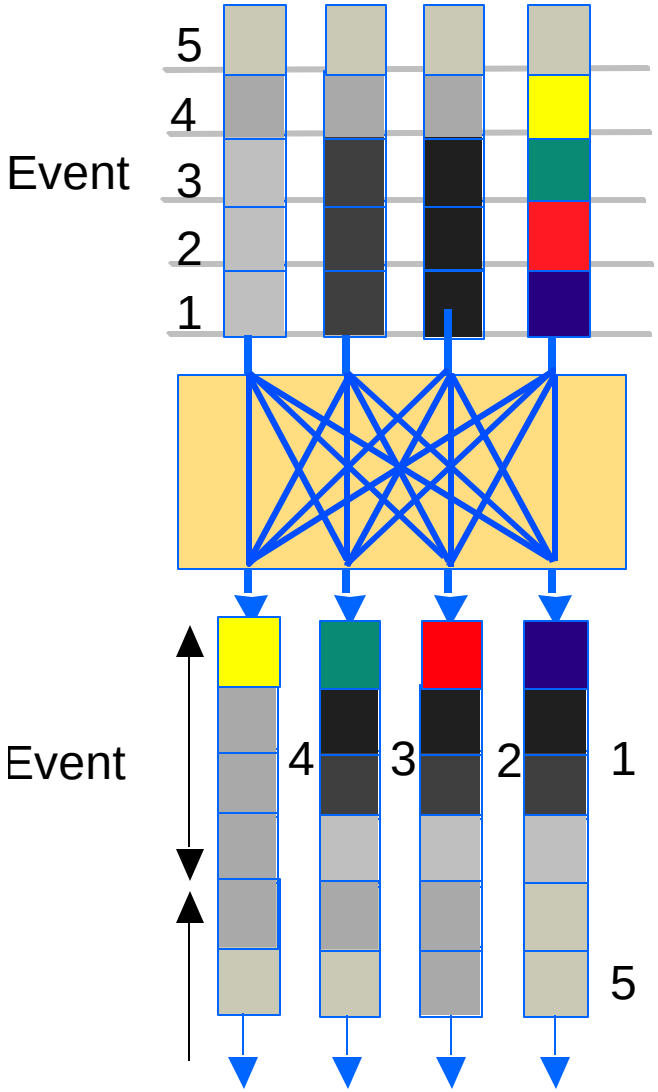
without traffic shaping

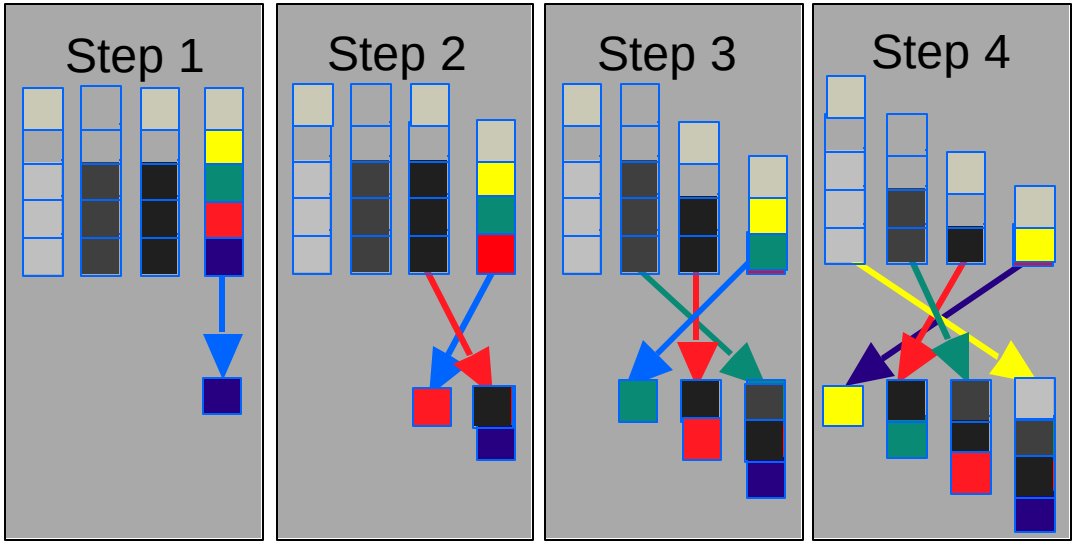
- **pro**: simple, robust
- **con**: reduced efficiency due to HOL blocking
  increases by adding FIFO buffers between stages or by adding stages to provide multiple paths from source to destinations
  ultimate efficiency ~ 60% (fully randomised)

with traffic shaping eg barrel shifter

- **pro**: high efficiency, in principle minimal topology sufficient
- **con**: complex, poor redundancy / fault tolerance

EVB traffic shaping: barrel shifter

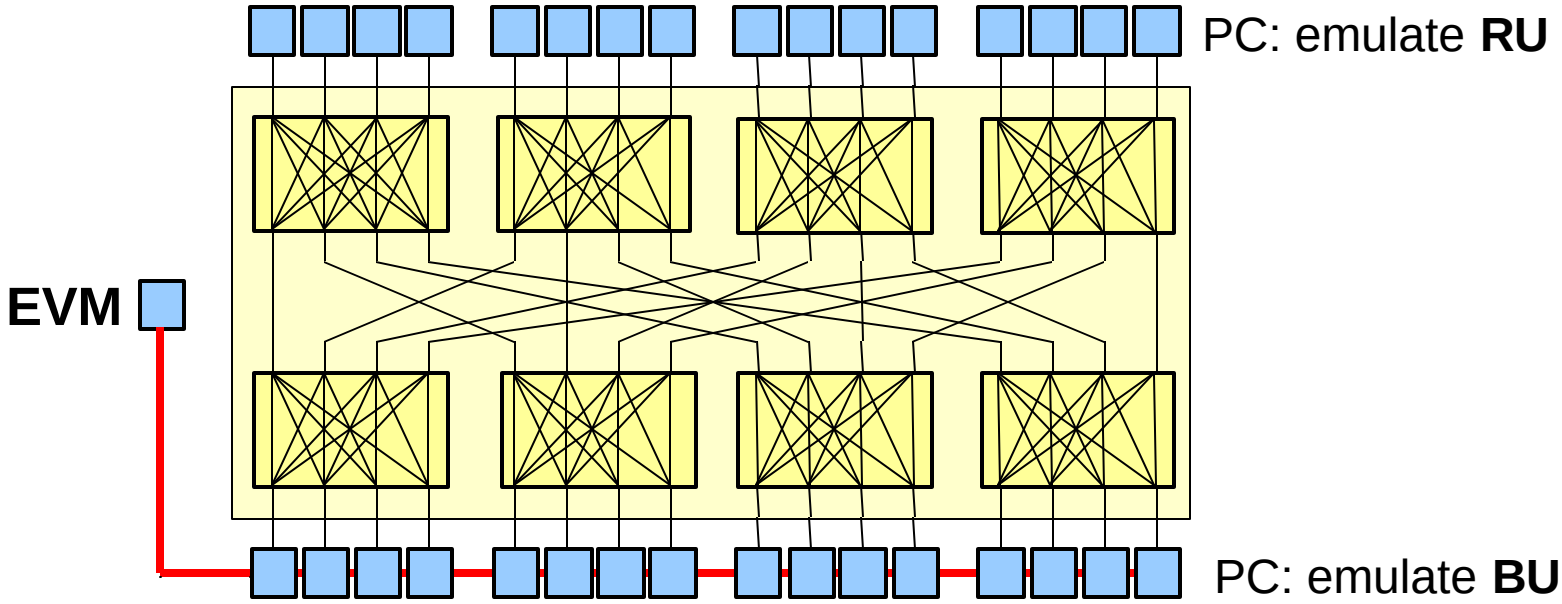




sources emit to mutually exclusive destinations in a cycle

- works only for fixed size chunks
- needs synchronisation

EVB set-up

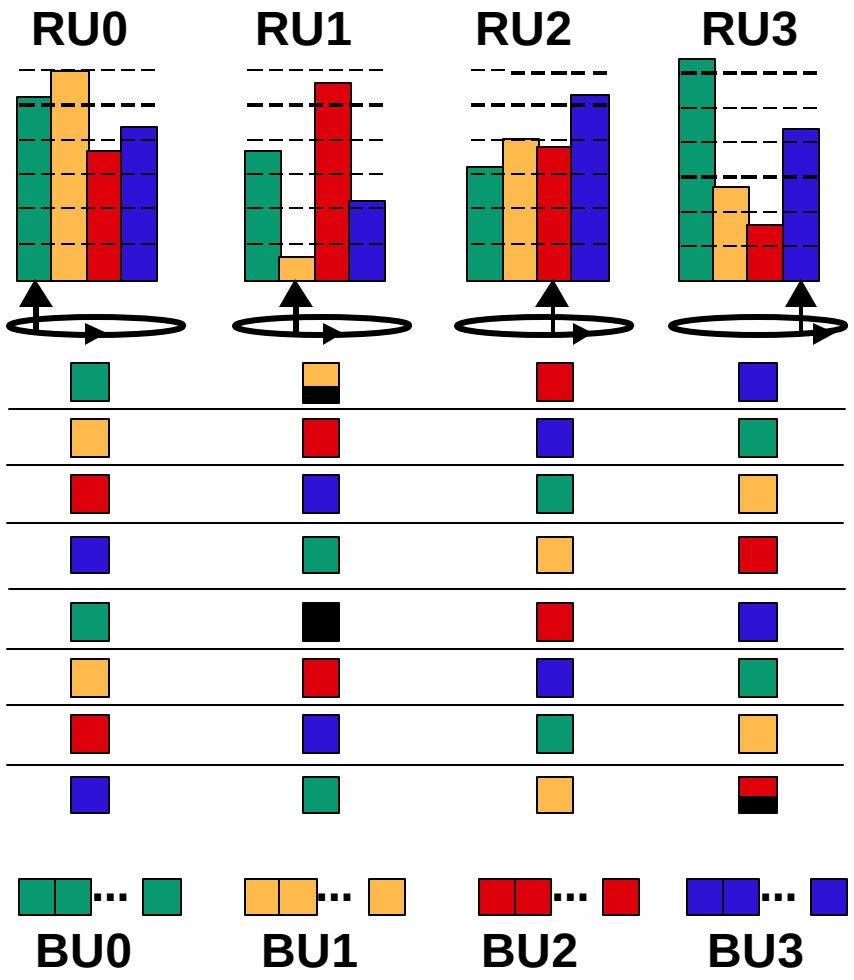


- 32+1 Linux PCs [450/700 MHz PII BX PCI 33 MHz/32b]
- Myrinet switch: M2M-OCT-SW8, NIC: M2M-PCI64[A]
- 16x16 two-stage Banyan (Delta) network out of 4x4 crossbars
- **Myrinet** between **RU**s and **BU**s (full duplex). N-to-N traffic
- **Fast Ethernet** between **BU**s and **EVM**. N-to-1 traffic
- For data messages only partial host-NIC DMA (33 MHz/32b)

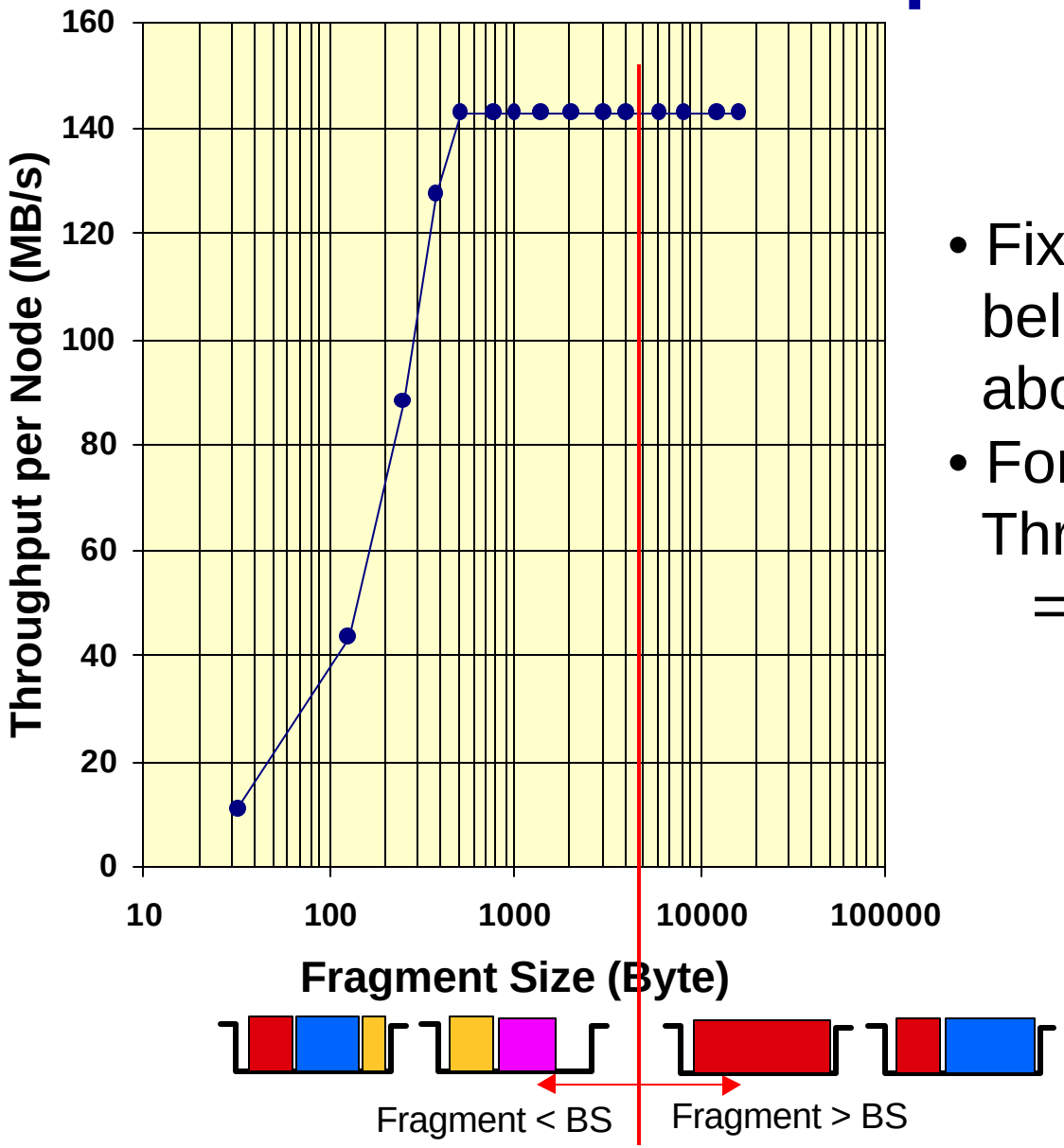
Barrel shifter Traffic shaping

• Sources divide fragments into fixed size packets (blocks) and cycle through all destinations
• Fragments can span more than one packet and a packet can contain data of more than one fragment

Implementation:
• BS performed by **NIC**
• Block size set to 4 kbyte (30 μs cycle)
• **Barrel shifter** without external synchronisation (Myrinet back pressure by HW flow control)
• Packets can be (partially) empty
• In principle works for large multistage switch as well



EVB 16x16 performance



- Fixed size event fragments
  below 4k: Fragment < BS block
  above 4k: Fragment > BS block
- For fragment sizes > 500 bytes:
  Throughput/node = 143 MB/s
  = 90 % of link BW

EVB 16x16 parameterisation

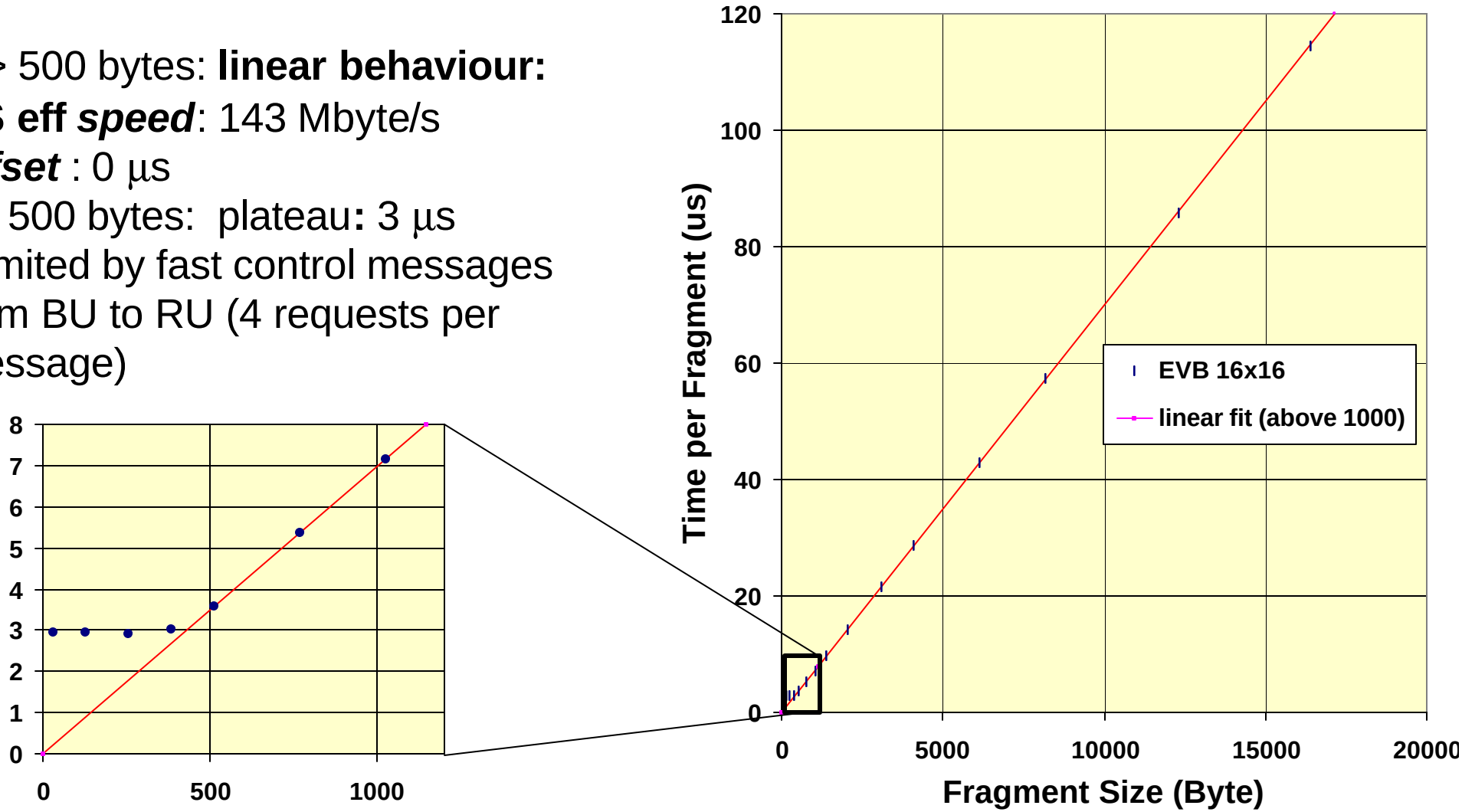
time per fragment = *offset* + **size** / *speed*

- > 500 bytes: **linear behaviour:**

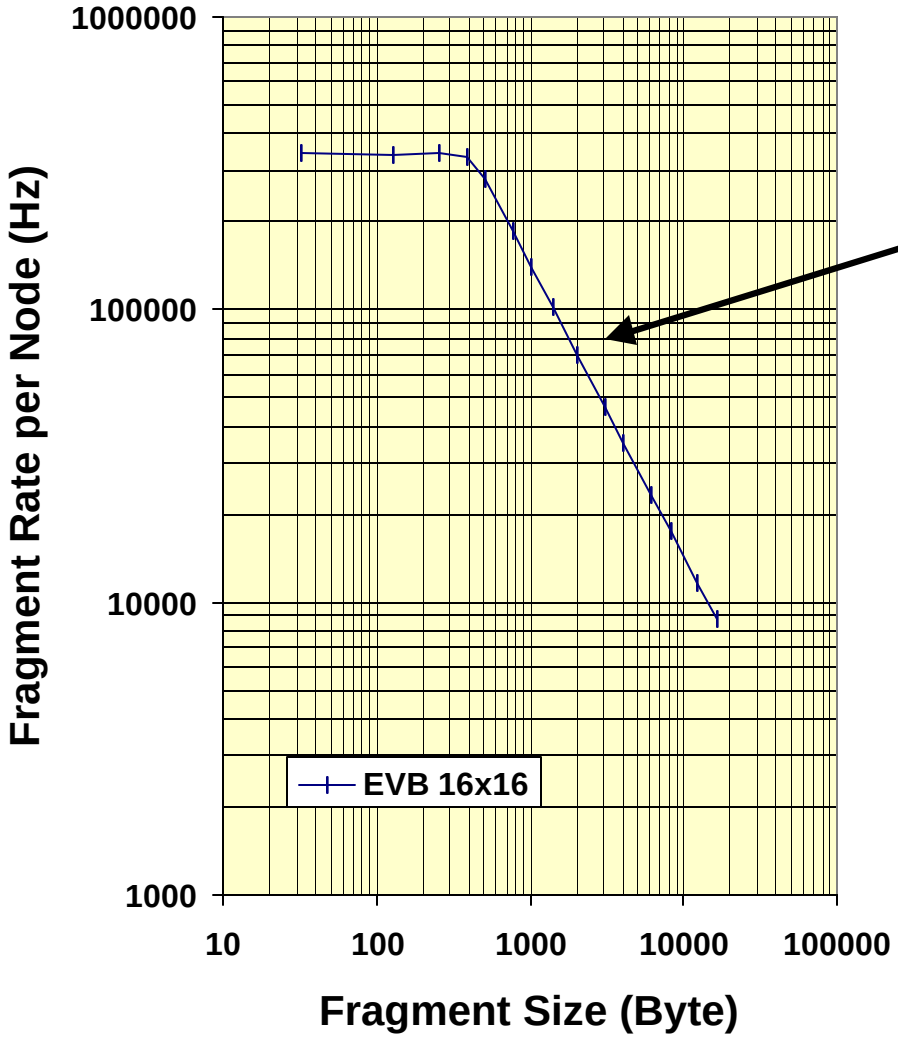
BS eff *speed*: 143 Mbyte/s

offset : 0 µs

- < 500 bytes: plateau**:** 3 µs
- limited by fast control messages from BU to RU (4 requests per message)

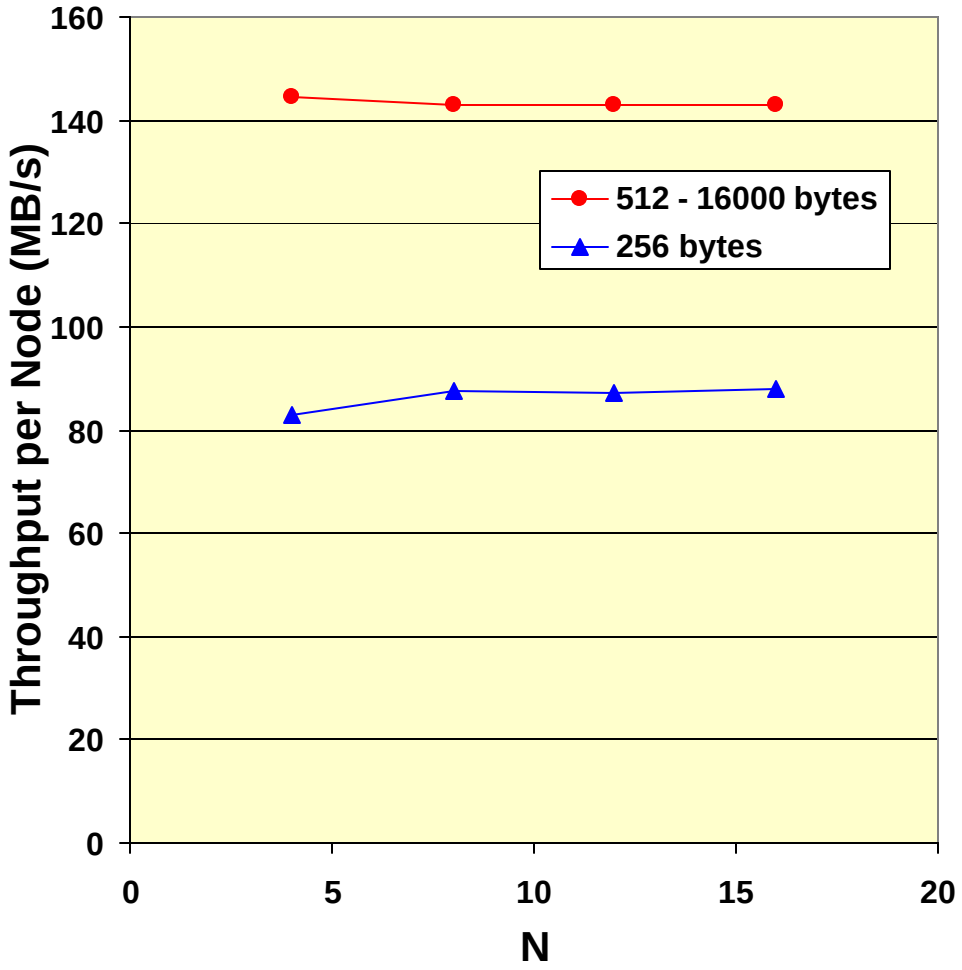


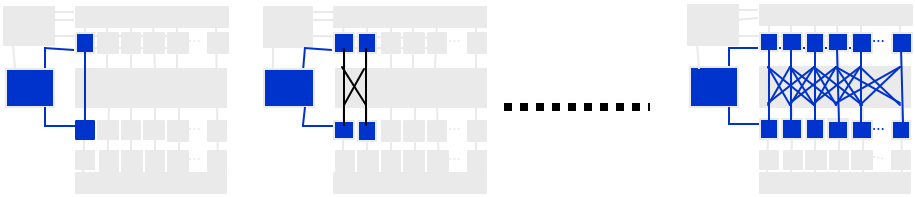
EVB performance - Event Rate



For nominal 2 kbyte fragment size:
Event rate = 70 kHz

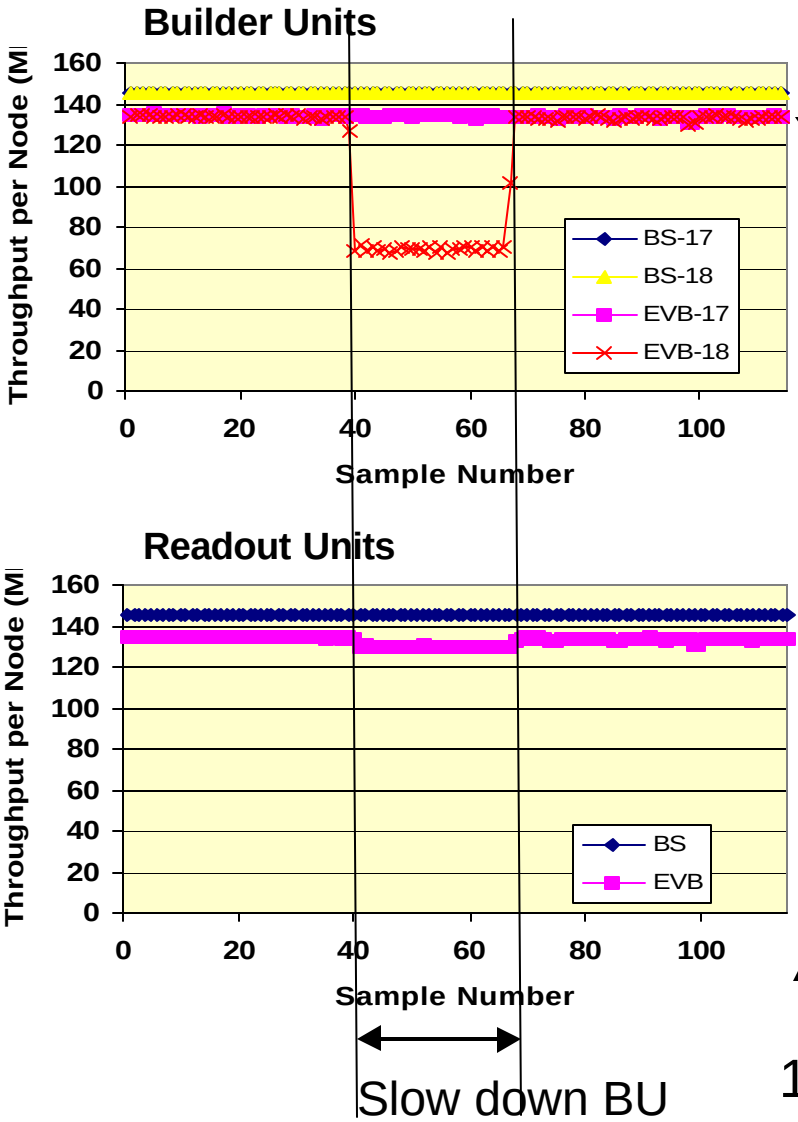
EVB scaling





From 4x4 → 16x16:
Scaling observed
(as expected from barrel shifter)

Traffic shaping - time evolution



BS cycling rate * block size

Event Building Throughput

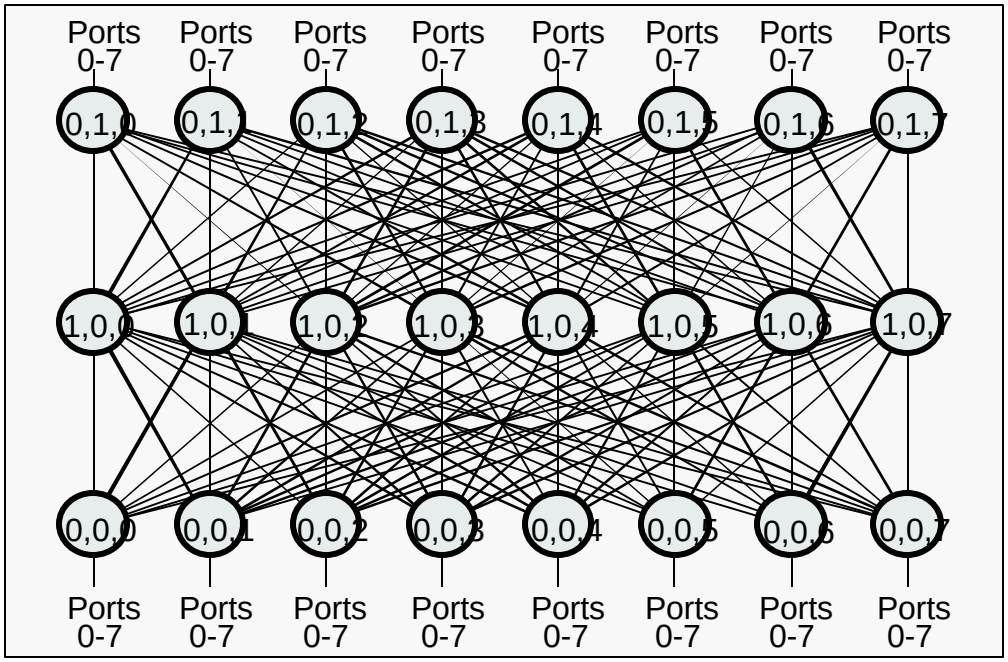
Slow down one BU by 50%
→ TS barrel shifter stays in sync
→ BS decoupled from EVB traffic
→ Event Manager adapts to BU capacity

[variable size event fragments with 2048 bytes average]

10 m (= 2 10^7 cycles, 1Tbyte moved)

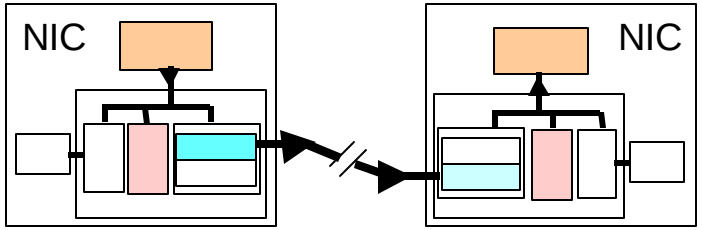
Myrinet 2000

- link speed 1.3 → 2 Gbps
- Lanai7 → Lanai9
- RISC 66 MHz → 132 MHz
- switch based on Xbar16
- also serial PHY medium



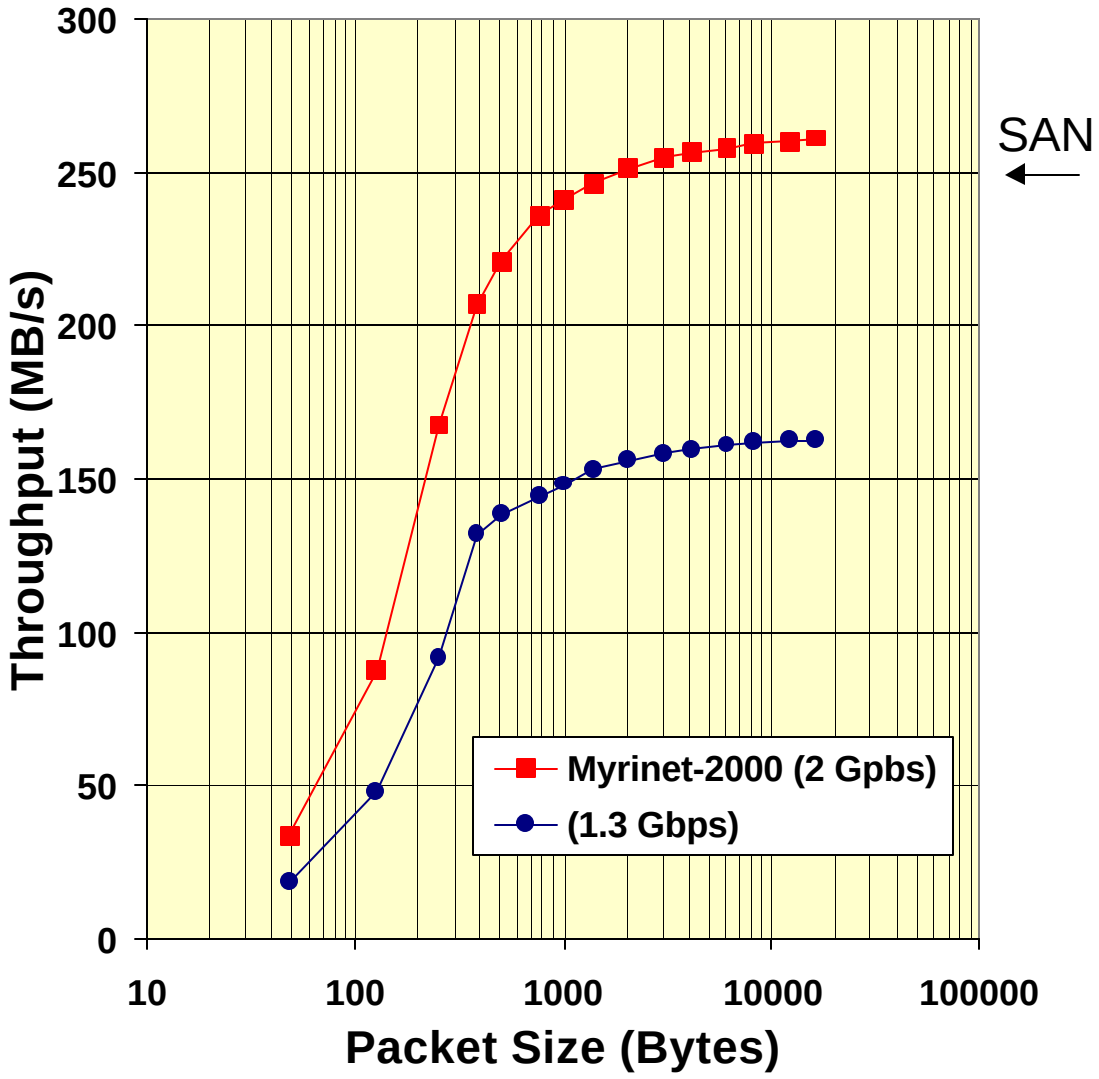
128-Port Clos Switch

Myrinet-2000 Point-to-point



1 source saturating 1 destination
NIC memory to NIC memory

Throughput approaches
2 Gpbs SAN bandwith
> 250 Mbyte/s

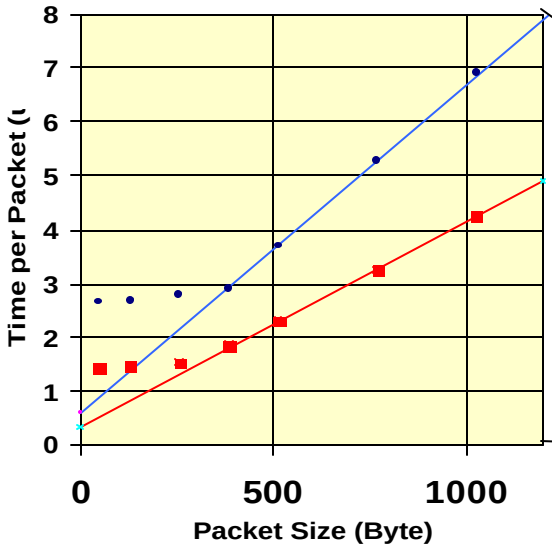


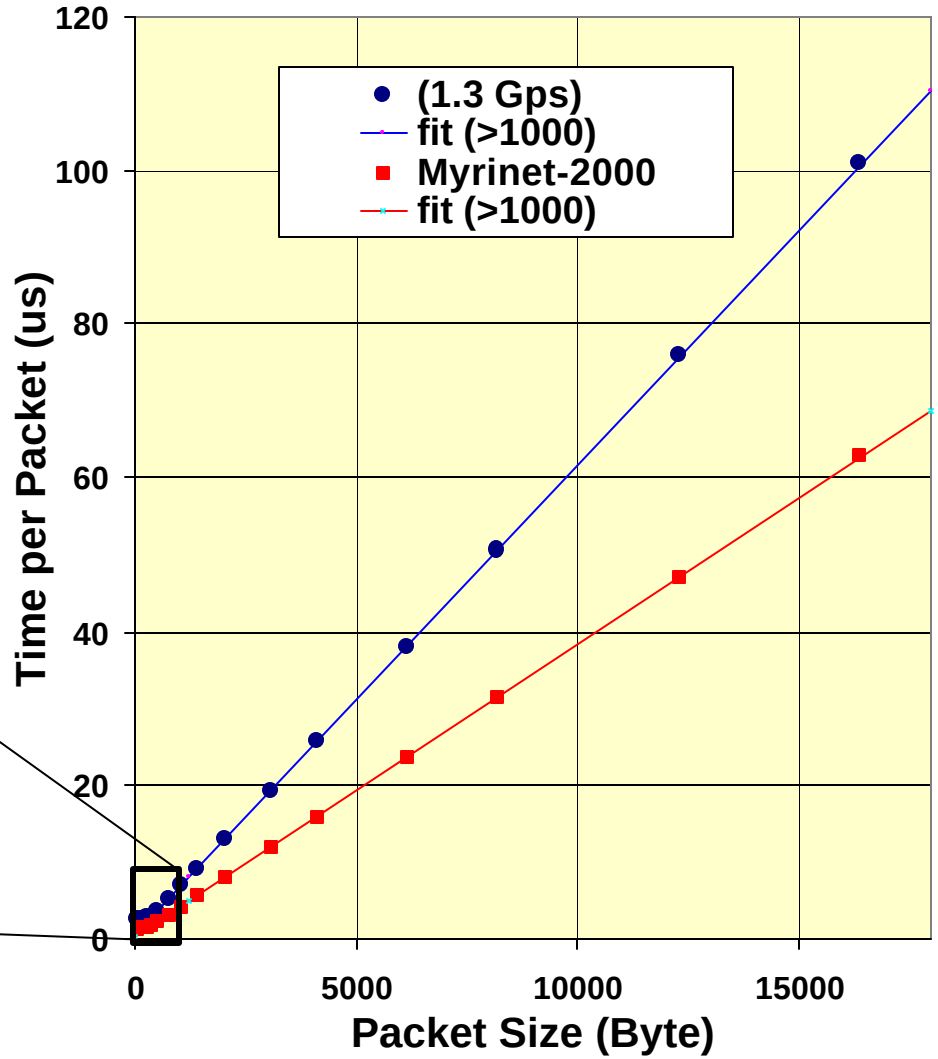
Parameters point-to-point Myrinet 2000

time per packet = *offset* + **size** / *speed*

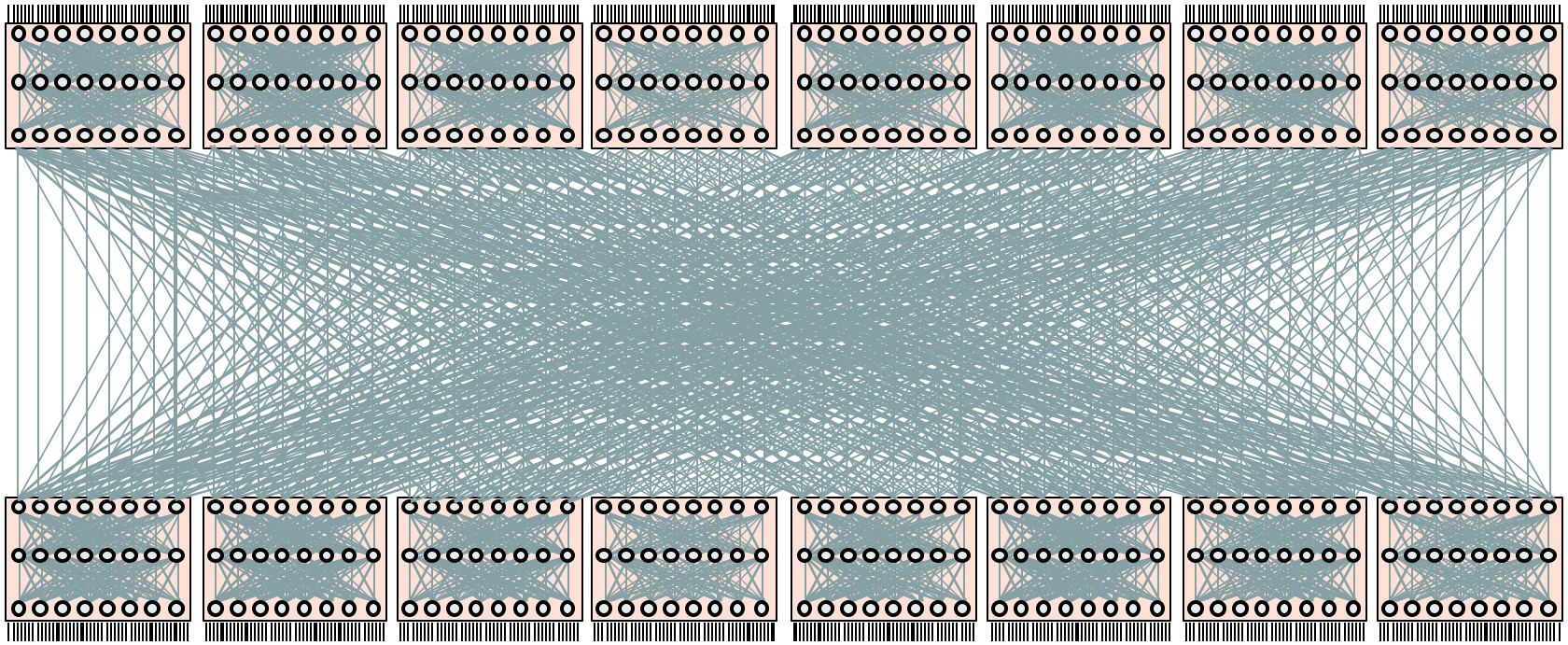
linear behaviour size >400 bytes

		Myrinet 2000
speed	164 MB/s	263 MB/s
offset	0.61 us	0.33 us
plateau	2.7 us	1.4 us





A 512x512 Switching Fabric

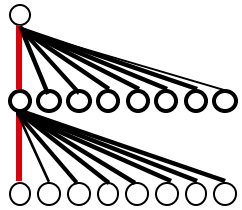


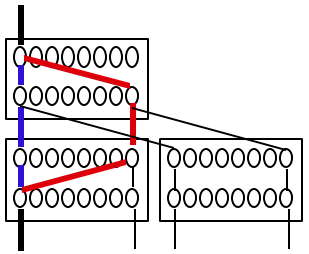
2 rows of 8 Clos-128 switches

- total bisection bandwidth 1 Tbps
- 6 layer - 512 minimal routes for each source - destination pair

Switch Fabric Performance

Various topologies for 512x512 multistage network out of 8x8 crossbars
Performance results from simulation†





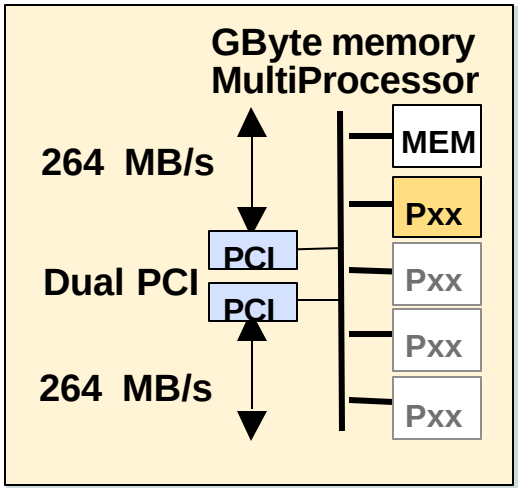
	Delta network	Symmetric	Symmetric
		4 - layer	6 - layer
# hops	3	4	6
# routes S-D pair	1	8	512
Eff. Random traffic	35%		
Eff. EVB traffic	15%	30%	

With barrel shifter traffic shaping efficiency close to 100%, in principle

† B. Rensch

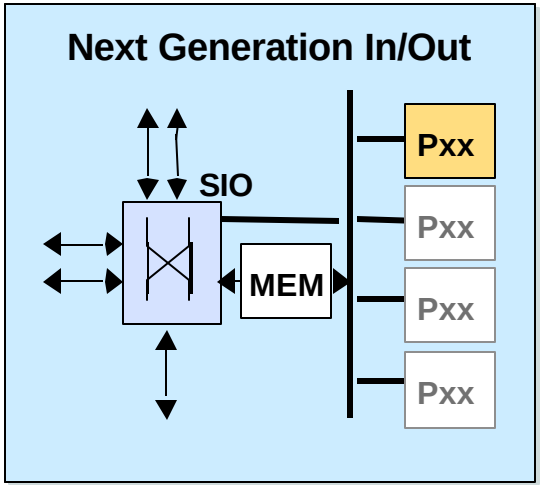
Myrinet - InfiniBand

1990’ PCI



Desktop/Server
current architecture
Peripheral IO**bus PCI:**
33/66 MHz x 32/64 bit
100/200/400 MB/s

2000’ InfiniBand?

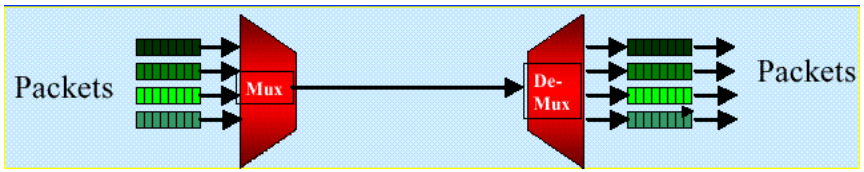


Future :

Devices, Memory, CPU unit communicate via **data channels** (e.g. 1 GB/s).
A **cross bar switch** substitutes the internal bus

InfiniBand: standard for future I/O and cluster interconnect

Myrinet roadmap: convergence to InfiniBand
now: Myrinet 2000 same physical layer as InfiniBand 1x
Lanai10: add features required for IB (eg virtual lanes)



Virtual lanes can increase efficiency in the face of HOL blocking

Myrinet - Summary

- **Switches and NIC evaluation**
  - depending on traffic pattern can have reduced throughput due to HOL blocking (as expected)
- **Event Building in Prototype (16x16)**
  - Throughput up to 90% link speed
  - With BS traffic shaping: scaling
  - For 2 kbyte fragment size, rate 70 kHz
- **Large (500x500) Multistage Switch**
  - no packet loss inside switch fabrics
  - without traffic shaping reduced performance due to HOL blocking
  - will BS traffic shaping work for large system?
  - need detailed simulation

Conclusion and Outlook

- Established EVB demonstrators GbE / Myrinet
  - test-bed of technology
  - event building studies (protocols, traffic shaping, ..)
  - so far 16x16, expand to 32x32 in 2000, ..
- Will also be used for DAQ column (full functionality)
- Need simulation and extrapolation to large systems