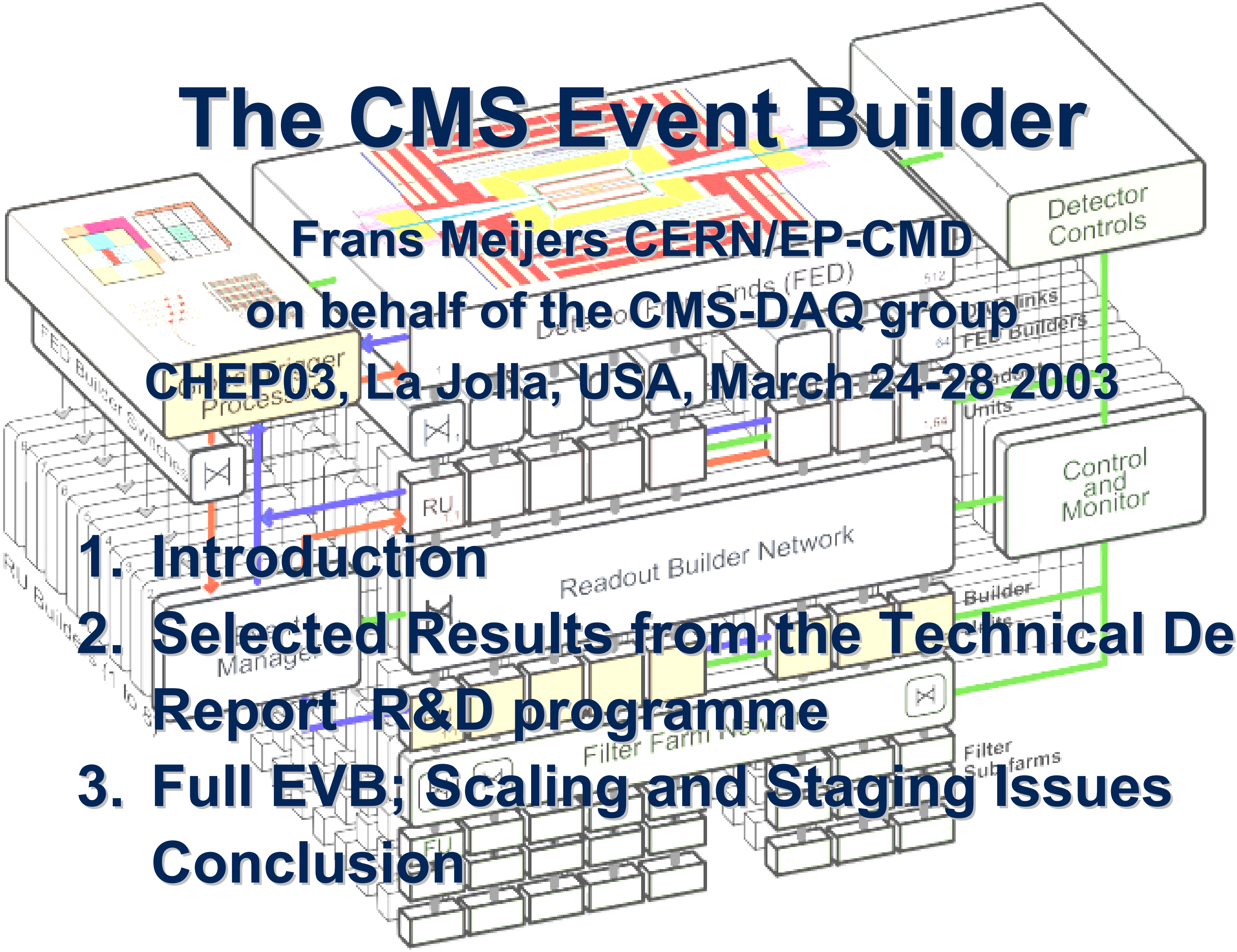
The CMS Event Builder

Frans Meijers CERN/EP-CMD

on behalf of the CMS-DAQ group

CHEP03, La Jolla, USA, March 24-28 2003



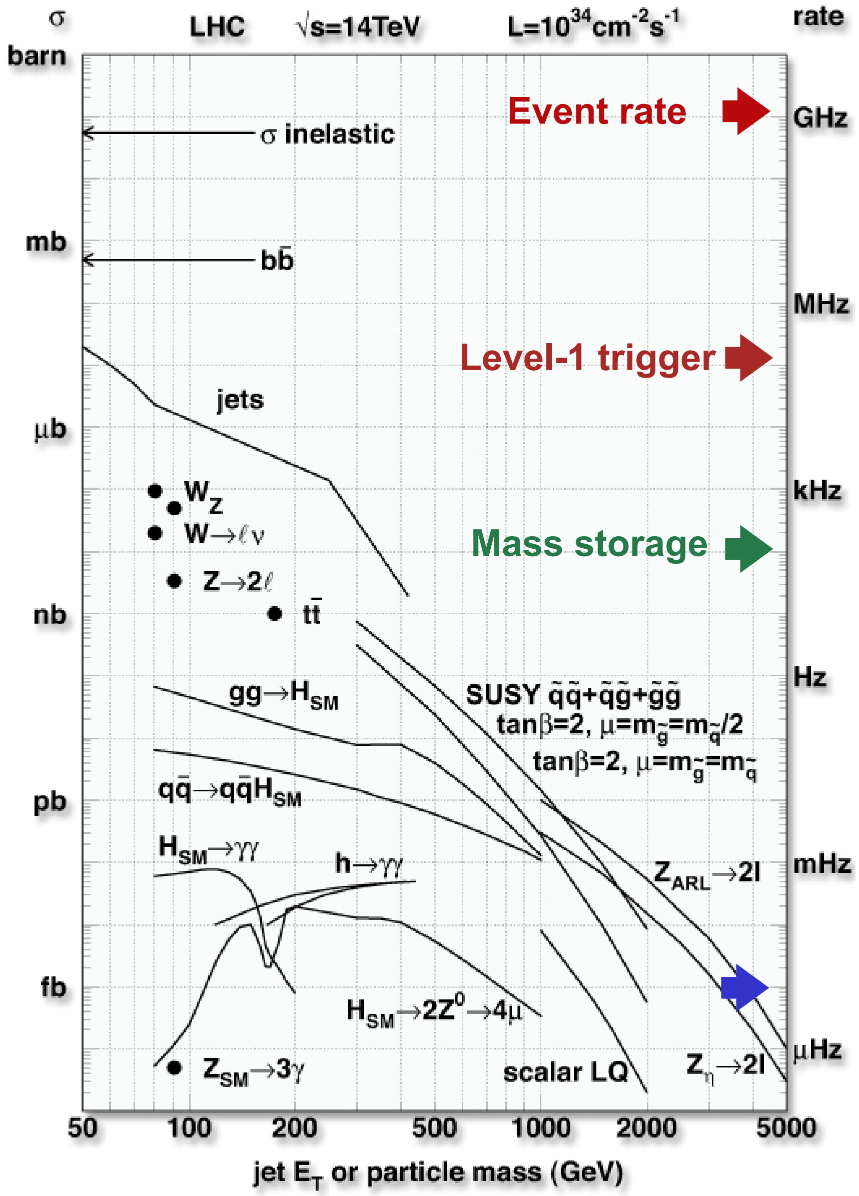
1. Introduction
2. Selected Results from the Technical Design Report R&D programme
3. Full EVB; Scaling and Staging Issues

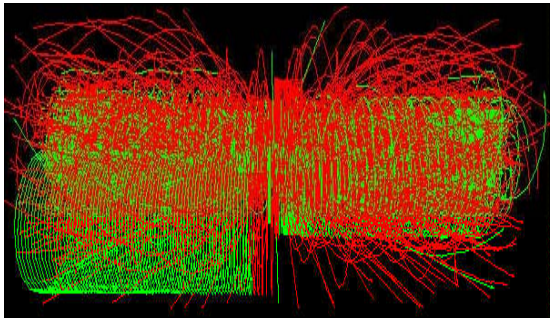
   Conclusion

contact: frans.meijers@cern.ch

Introduction

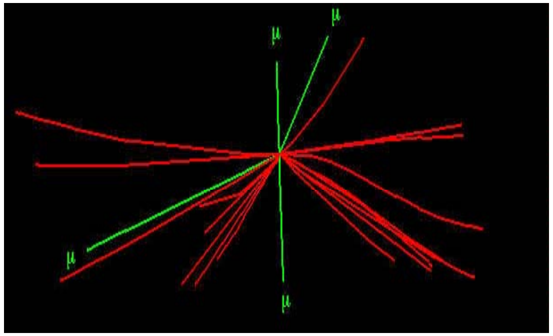
p-p collisions at LHC





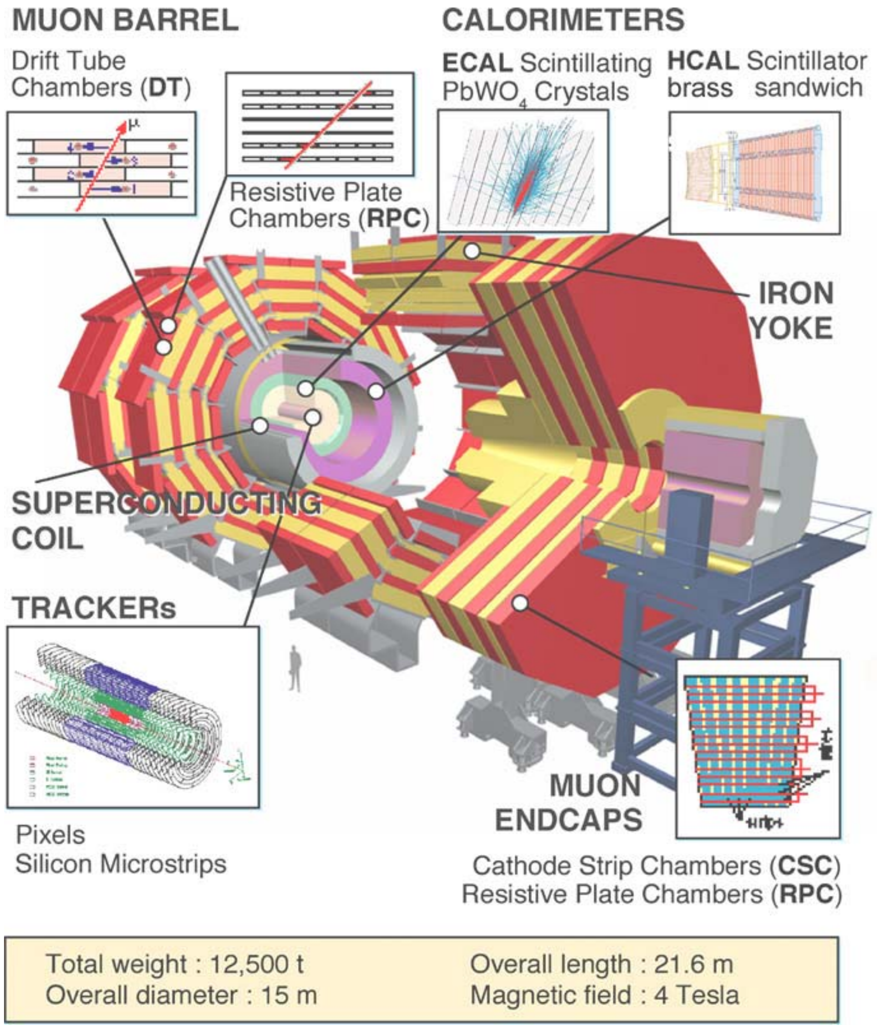
Event Rates:	~10^9 Hz

Level-1 Output	100 kHz
Mass storage	100 Hz
Selection Online:	~1/10^6
Event Selection:	~1/10^{13}



Requirements and design parameters

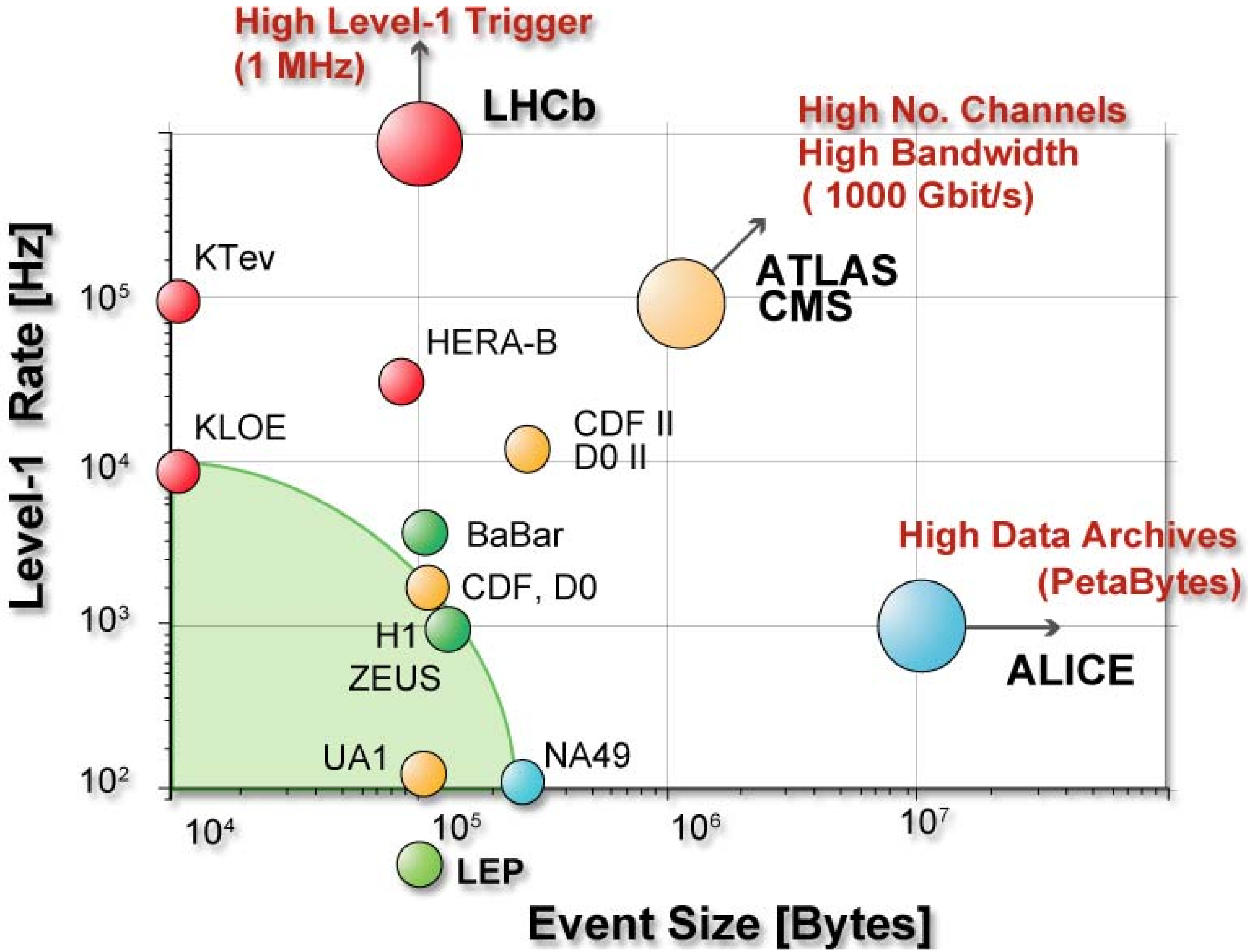
Detectors



Detector	Channels	Ev. Data	EVB-inputs
Pixel	60000000	50 (kB)	36
Tracker	10000000	650	442
Preshower	145000	50	50
ECAL	85000	100	60
HCAL	14000	50	24
Muon DT	200000	10	5
Muon RPC	200000	5	3
Muon CSC	400000	90	8
Trigger		16	8

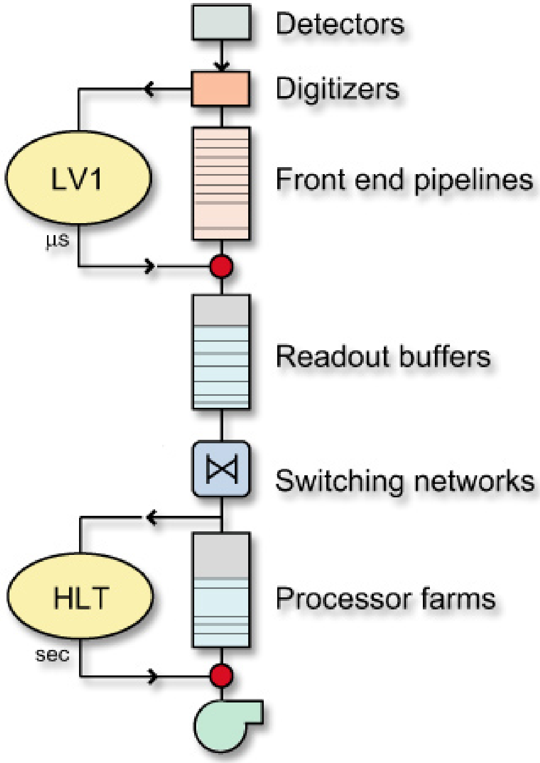
Event size (after reduction)	**1 MByte**
EVB-DAQ inputs	**636**
Max LV1 Trigger	**100 kHz**
Online rejection	**99.999%**
System dead time	**~ %**

Trigger and data acquisition trends



CMS DAQ structure: 2 physical triggers

40 MHz
Clock driven
Custom processors
Decision time O(μs)



Level-1 Trigger
Custom design

100 kHz
Event driven
PC network
Decision time O(s)

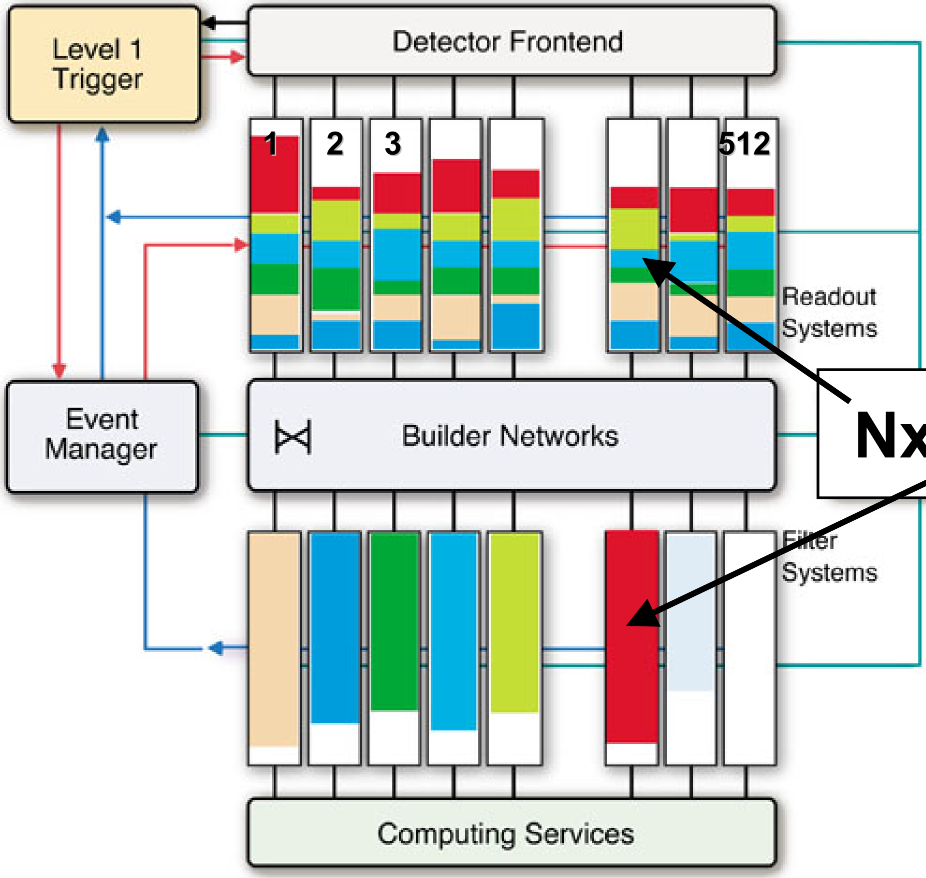
High-Level Trigger
"Off-line" code

Level-1 output / HLT input	**100 kHz**
EVB network throughput	**1 Terabit/s**
HLT output	**10^2 Hz**

Invest in data transportation and CPU

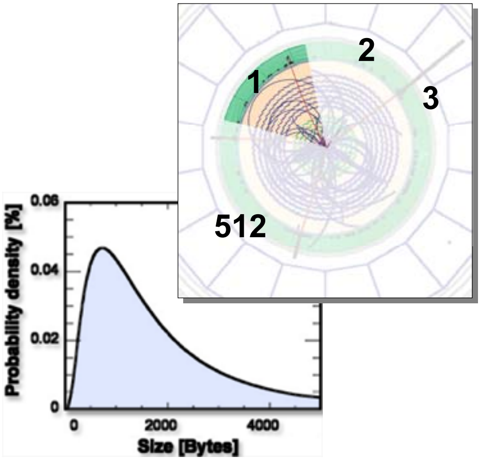
Building the event

Event builder :
Physical system interconnecting data sources with data destinations. It has to move each event data fragments into a same destination

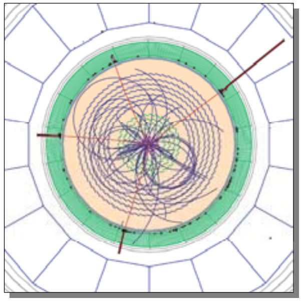


NxM EVB

Event fragments :
Event data fragments are stored in separated physical memory systems

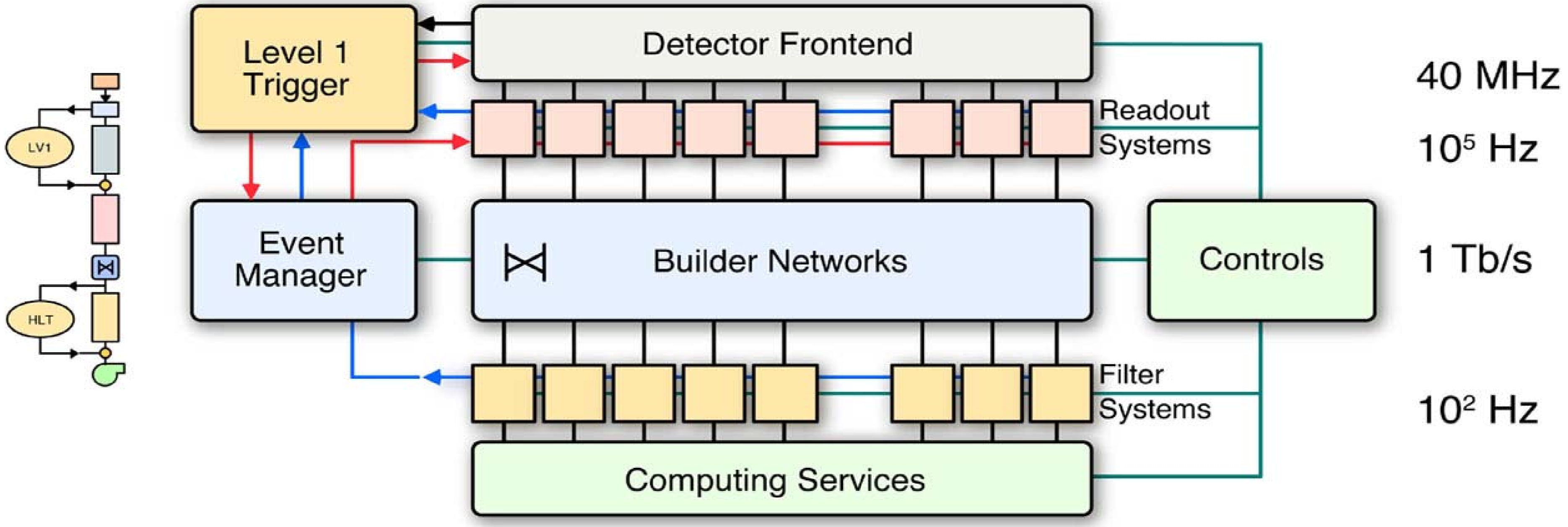


Full events :
Full event data are stored into one physical memory system associated to a processing unit



512 Data sources for 1 MByte events
~1000s HTL processing nodes

DAQ baseline structure

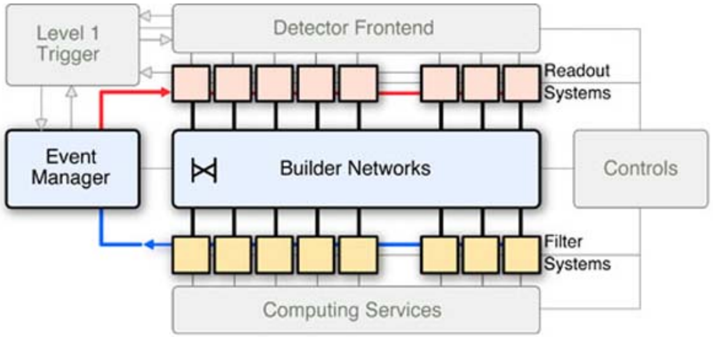


Collision rate	40 MHz	No. of In-Out units	**512**
Level-1 Maximum trigger rate	**100 kHz**	**EVB network throughput**	**≈ 1 Terabit/s**
Average event size	**≈ 1 MByte**	**Event filter computing power**	**≈ 10^6 SI95**
Event Flow Control	≈ 10^6 Mssg/s	Data production	≈ Tbyte/day
		No. of PC motherboards	≈ Thousands

Selected Results from the Technical Design Report R&D programme

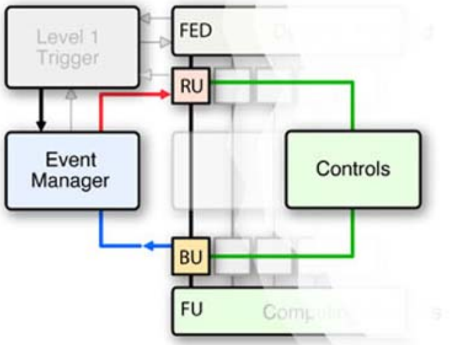
TDR R&D programme

TDR development programme decoupling functionality from performance



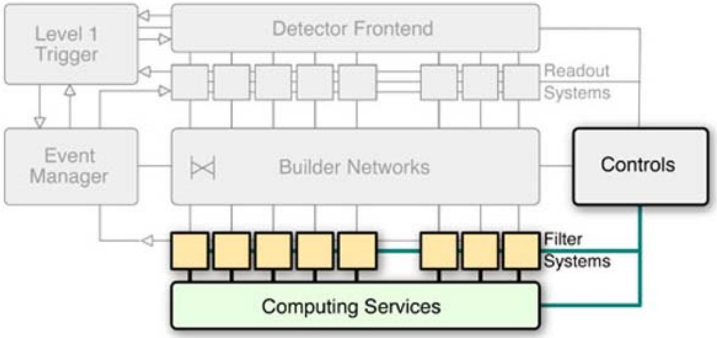
DEMONSTRATORs: Event Builder (EVB)

- Evaluate **network technologies**
- Study **EVB protocols**
- Performance studies by **test benches** and **simulation**



PROTOTYPEs: DAQ Column

- Evaluate **PC platforms** applied to IO systems
- Detector readout and Trigger/**DAQ interfaces**
- Readout hardware/software prototypes integration
- Data flow and Control prototypes
- **Test beams DAQ/DCS systems**

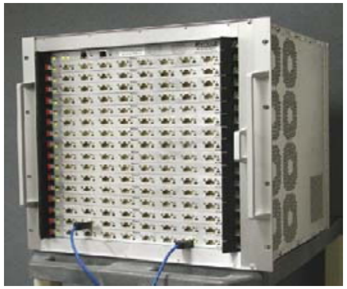


DEVELOPMENTS: Software and Event Filter

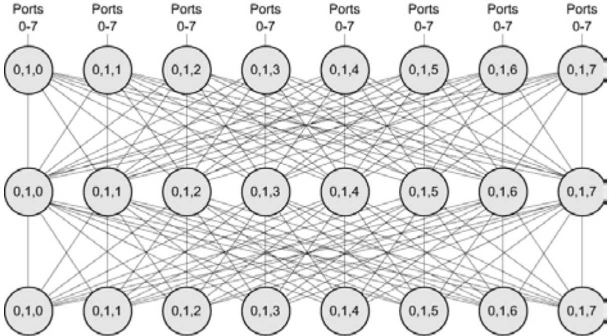
- **Online software** framework
- **Run Control** and Web services
- Farm event distribution and computing services
- HLT application **framework and HLT controls**
- HLT algorithms
- **Farm management** and control
- Data streams and mass storage

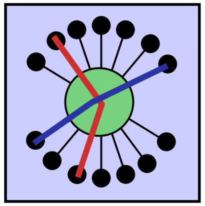
EVB and switch technologies

Myrinet 2000 (from Myricom)

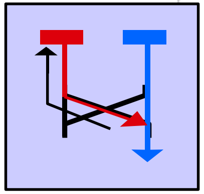


- Market applications: High Performance Parallel Computing
- Switch: **Clos-128 x 2.0** Gbit/s port
- **NIC**: M3S-PCI64B-2 (LANai9)





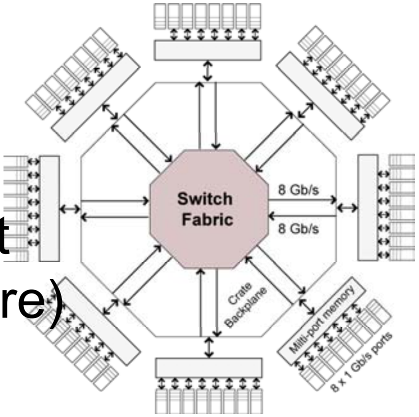
Implementation :
16 port X-bar capable of channeling data between any two ports.

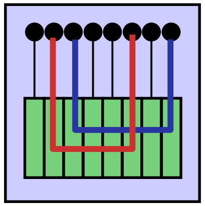


wormhole routing
transport with flow control at all stages

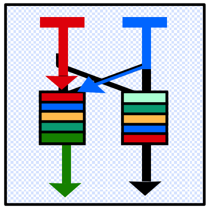
Gigabit Ethernet

- Market application: Cluster net, LAN, WAN
- **Switch**: Foundry **FastIron 64 x 1.0** Gbit/s port
- **NIC**: Alteon AceNIC (running standard firmware)



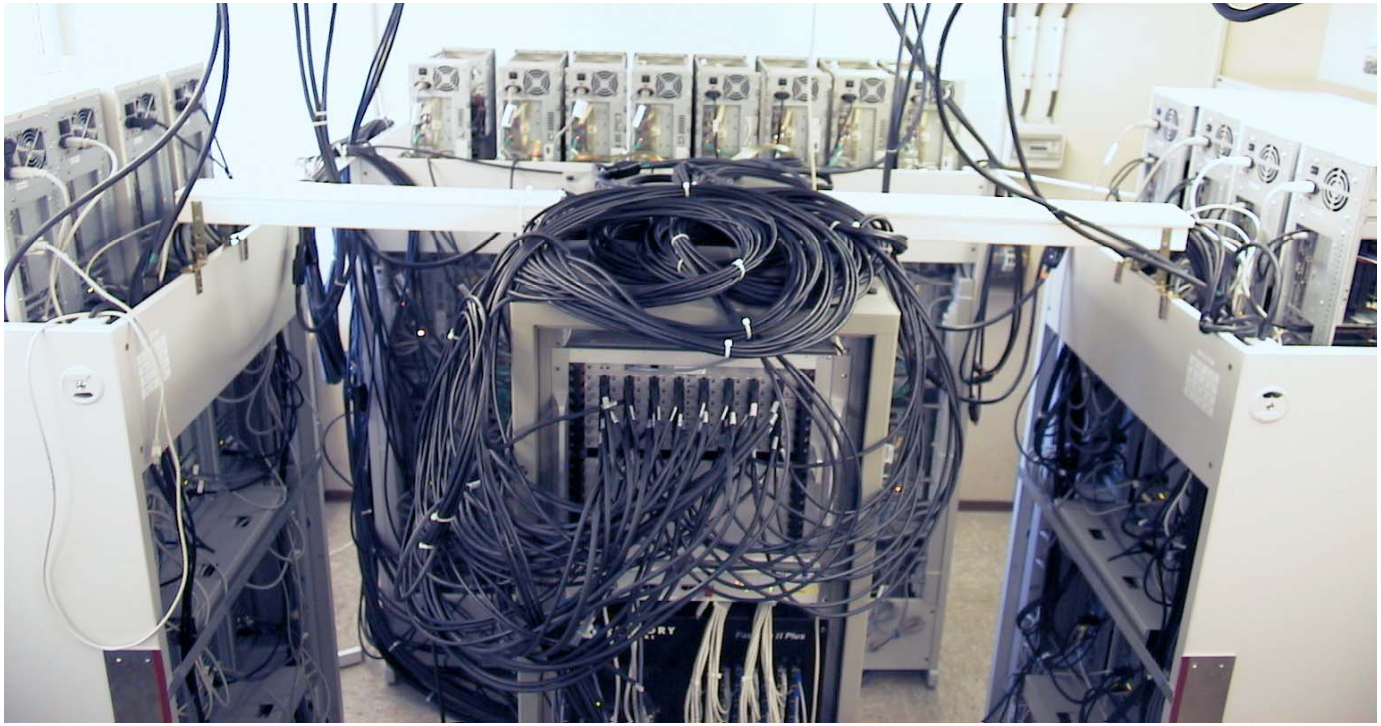


Implementation:
Multi-port memory system of R/W access bandwidth greater than the sum of all port speeds



Packet switching
Contention resolved by Output buffer. Packets can be lost.

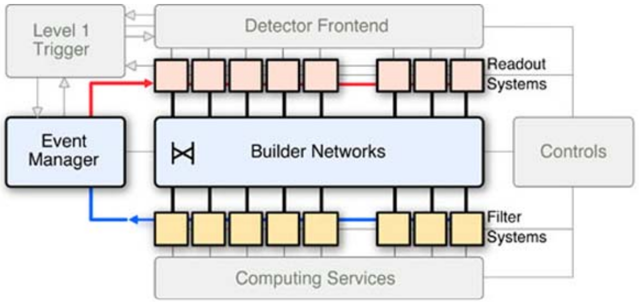
EVB demonstrator test bench 32x32



- **64 PCs**
    - SuperMicro 370DLE (733 MHz and 1 GHz **Pentium3**), 256 MB DRAM
    - ServerWorks LE chipset **PCI**: 32b/33MHz + **64b/66 MHz** (2 slots)
    - Linux 2.4
- **Myrinet2000** Clos-128 switch (64 ports equipped) and M3M-PCI64B NICs
- GB Ethernet 64 port **FastIron8000** and Alteon **AceNIC** NICs

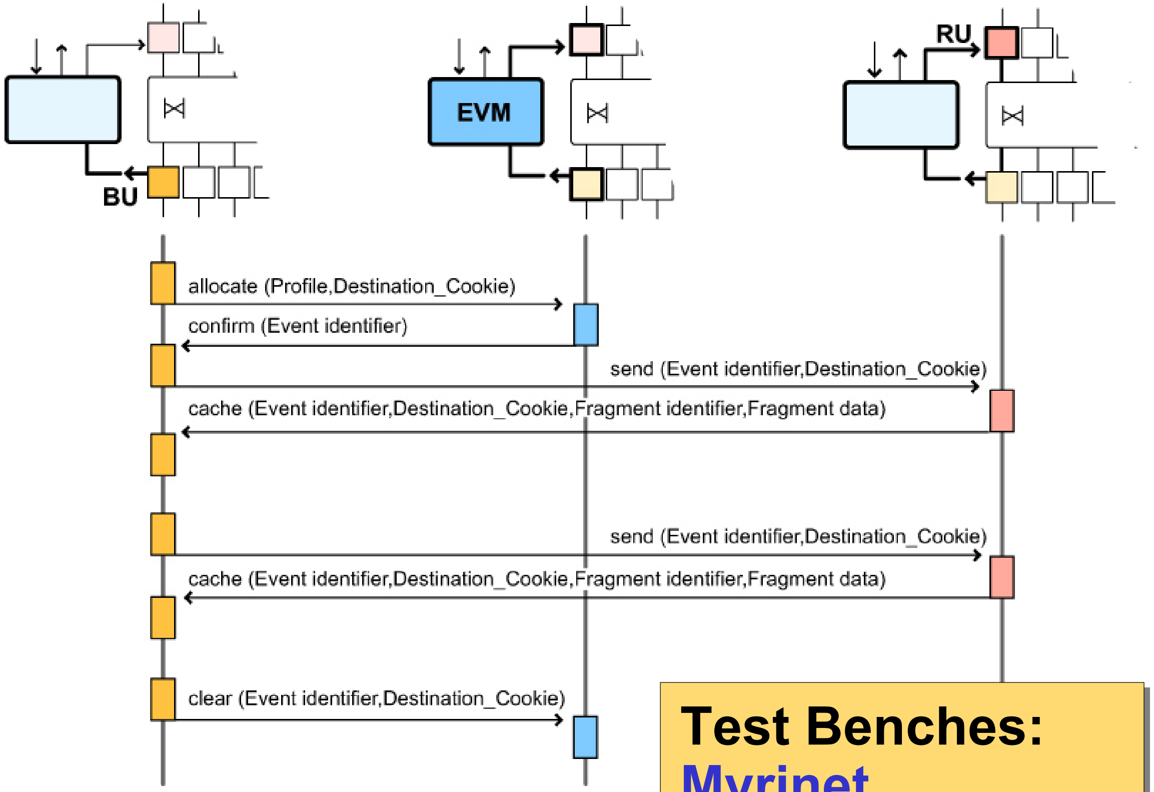
“State-of-the-Art” in 2001

EVB test bench measurements



EVB DAQ Protocol (PULL):

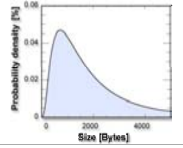
Event allocation and event data fragments are requested by destination. The event manager (EVM) handles the status of event during the EVB operation


BU
EVM
RU
allocate (Profile,Destination_Cookie)
confirm (Event identifier)
send (Event identifier,Destination_Cookie)
cache (Event identifier,Destination_Cookie,Fragment identifier,Fragment data)
send (Event identifier,Destination_Cookie)
cache (Event identifier,Destination_Cookie,Fragment identifier,Fragment data)
clear (Event identifier,Destination_Cookie)

Measurements:

- Throughput (at EVB application) per node (RU or BU)
- No. of ports and performances (scaling), packet loss,

With fixed and variable (log norm distribution) fragment sizes



Test Benches:

Myrinet

Random traffic

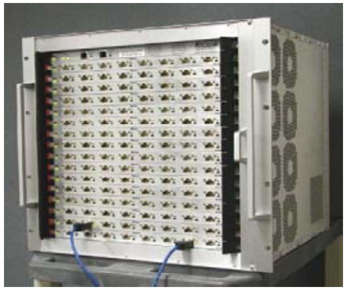
Barrel shifter

GbEthernet

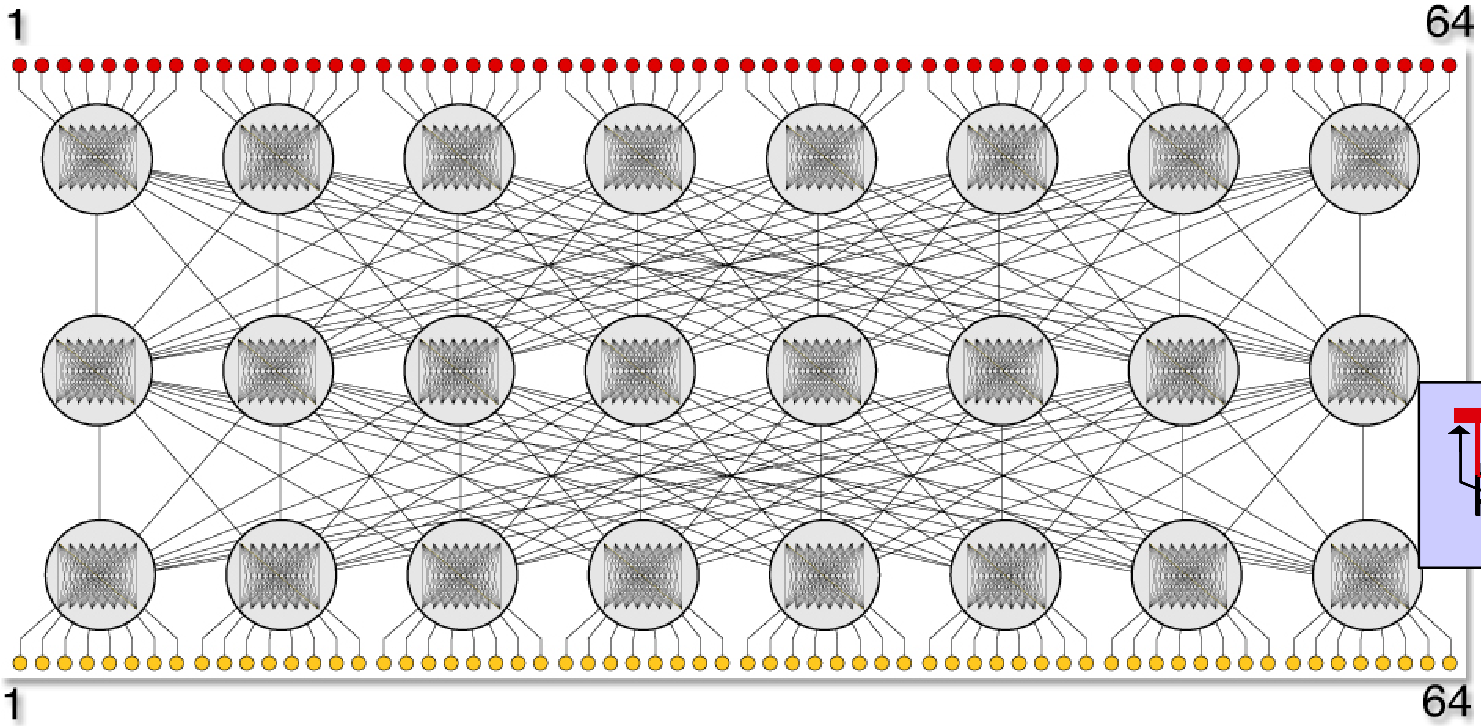
Raw packets

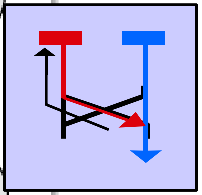
TCP/IP

Myrinet



- network built out of crossbars (Xbar16)
- wormhole routing, built-in back pressure (no packet loss)
- **switch**: 128-Clos switch crate
  - **64x64 x 2.0** Gbit/s port (bisection bandwidth **128 Gbit/s**)
- **NIC**: M3S-PCI64B-2 (LANai9 with RISC), custom Firmware



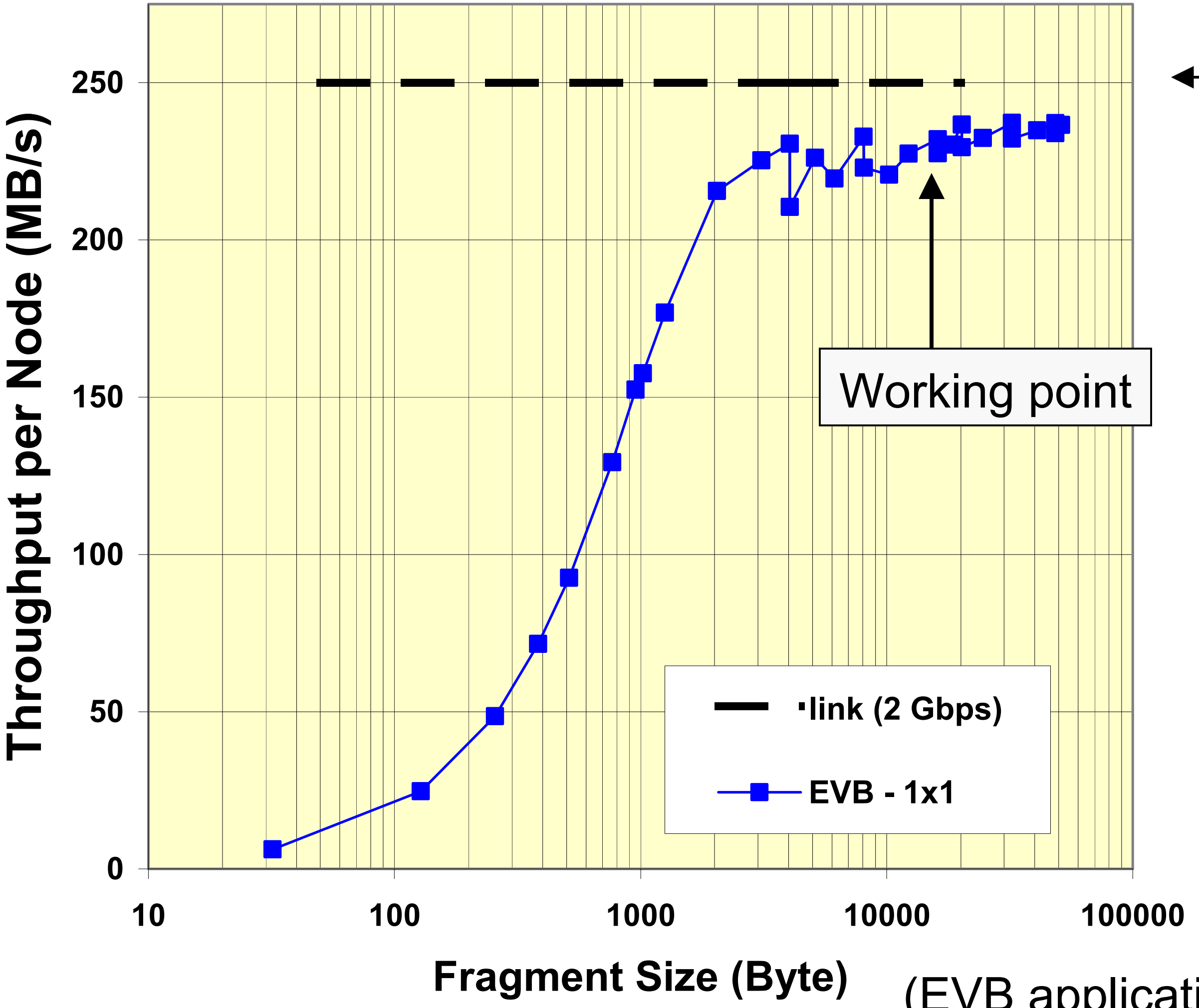


Wormhole routing
transport with flow
control at all stages

Myrinet

intermezzo

Myrinet EVB with random traffic (I)

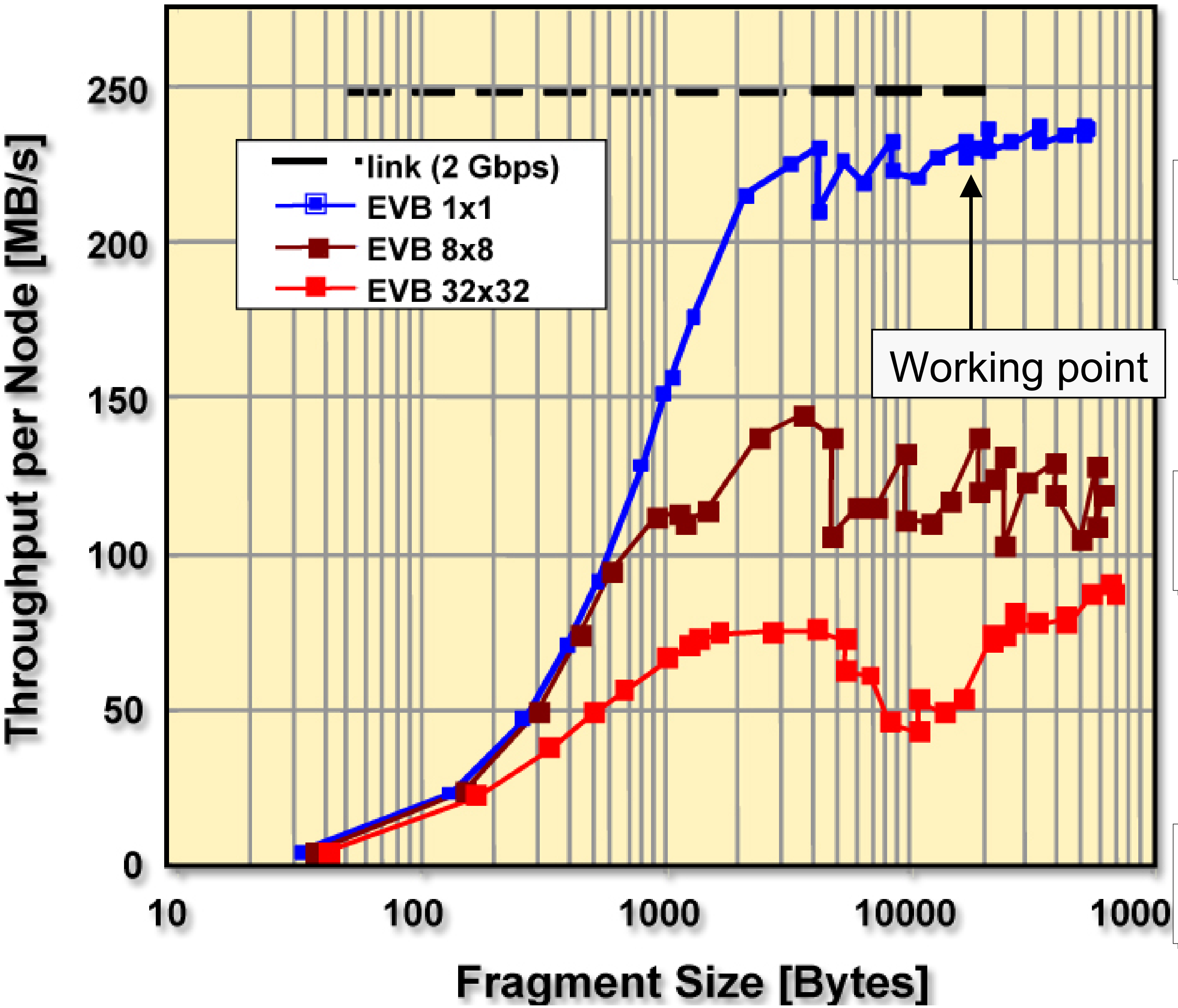


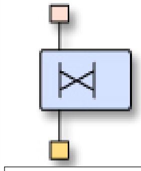
link speed (2 Gbit/s)

(EVB application payload)

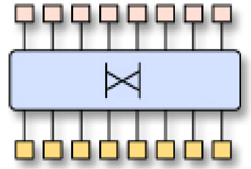
1x1: single stage:
max. **utilisation: > 90%**

Myrinet EVB with random traffic (II)

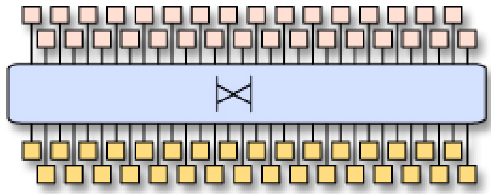




1x1: single stage:
max. **utilisation: > 90%**

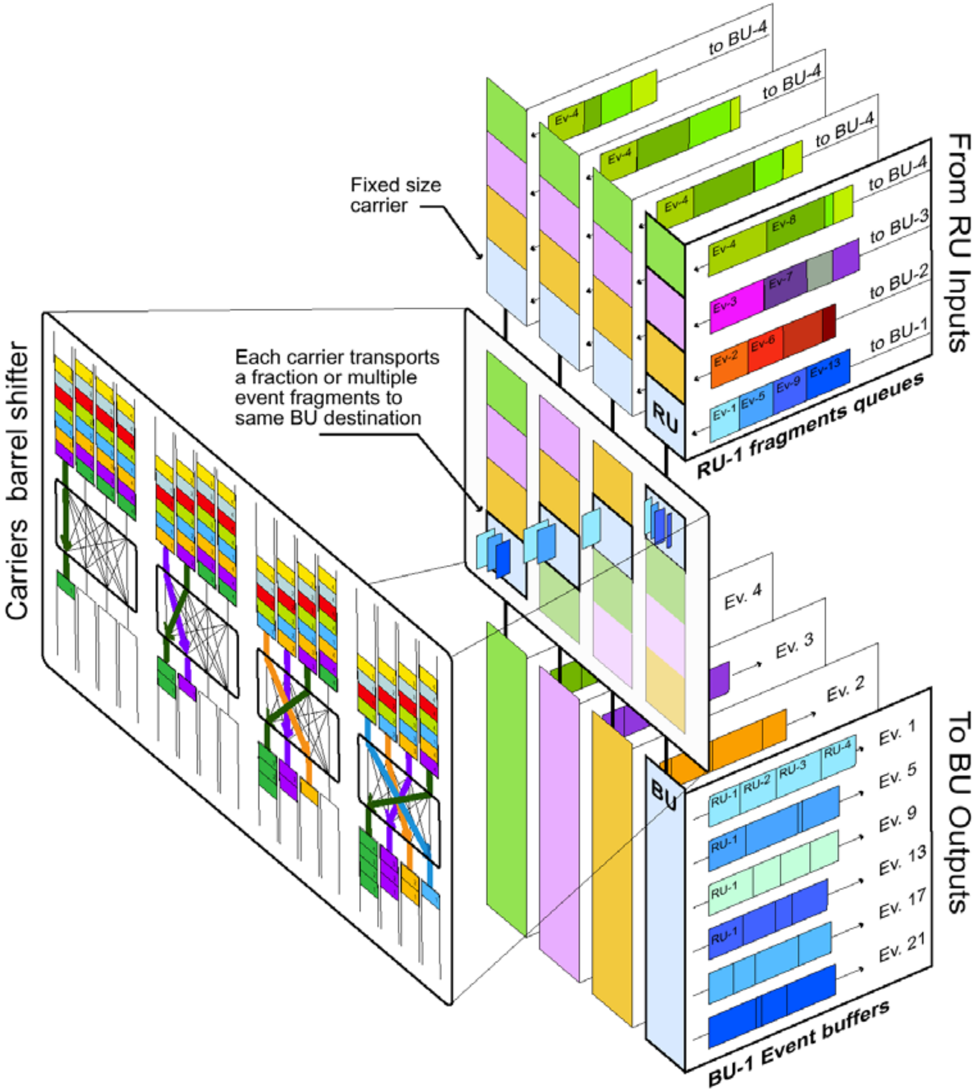


8x8: single stage:
max. **utilisation: ≈ 50%**



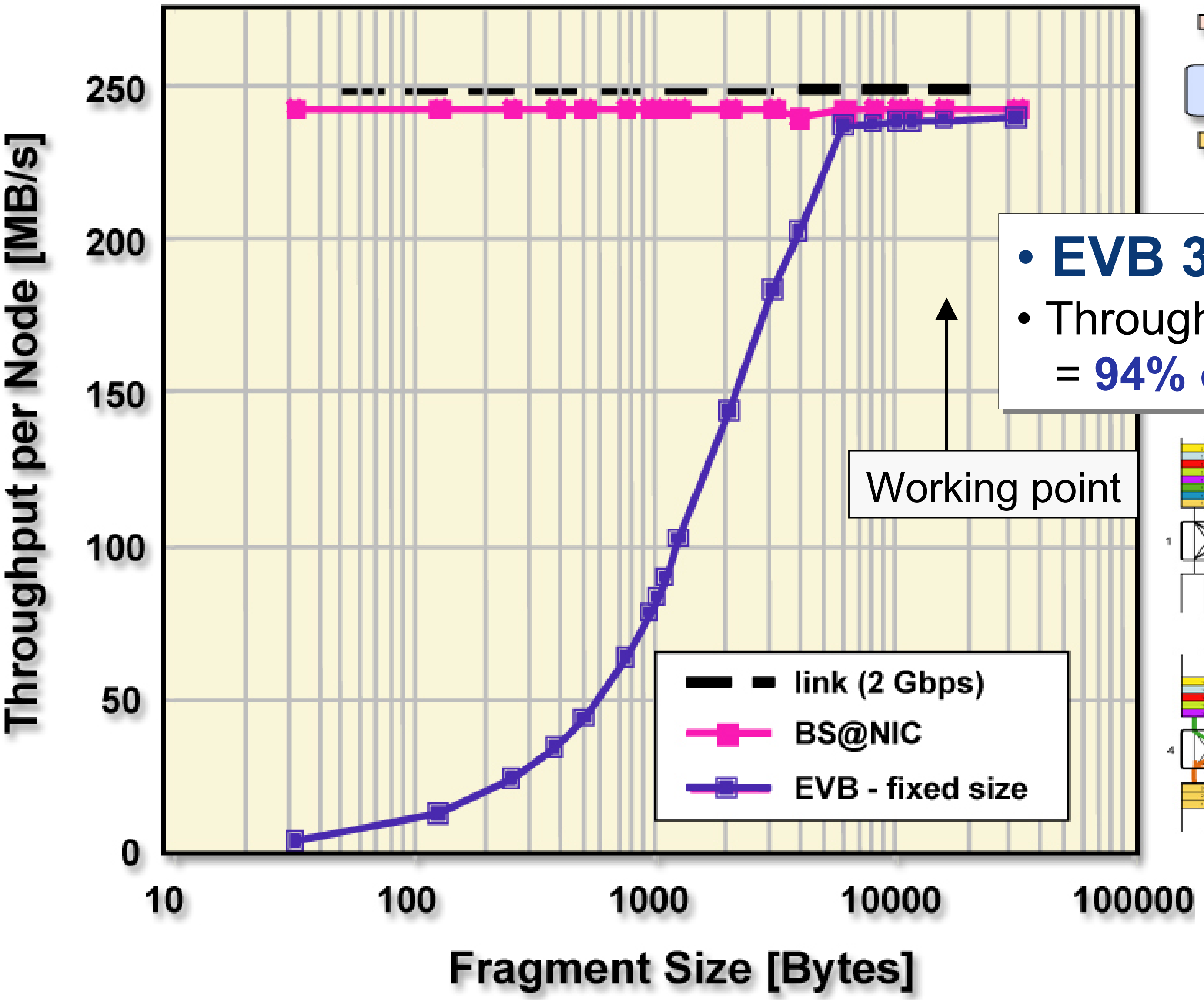
32x32: two stage network
max. **utilisation ≈ 30%**

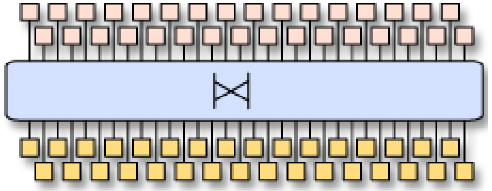
Myrinet EVB: Barrel shifter



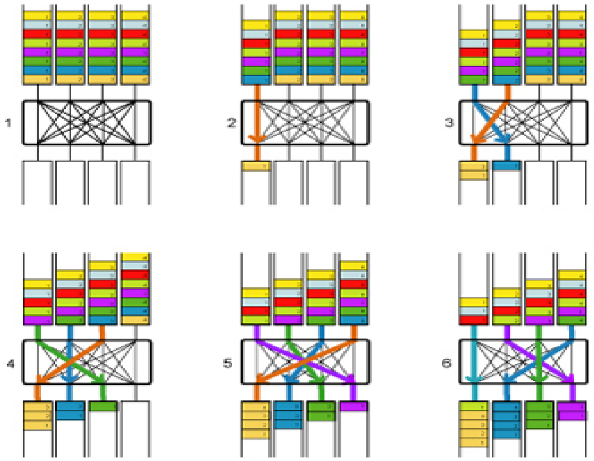
- Barrel shifter **implemented in NIC firmware**
  - Each source has message queue per destination
  - Sources divide messages into fixed size packets (carriers) and cycle through all destinations
  - Messages can span more than one packet and a packet can contain data of more than one message
  - No external synchronization (relies on Myrinet back pressure by HW flow control)
- zero-copy, **OS-bypass**
- **principle works** for multi-stage switches

Myrinet: EVB with barrel shifter protocol

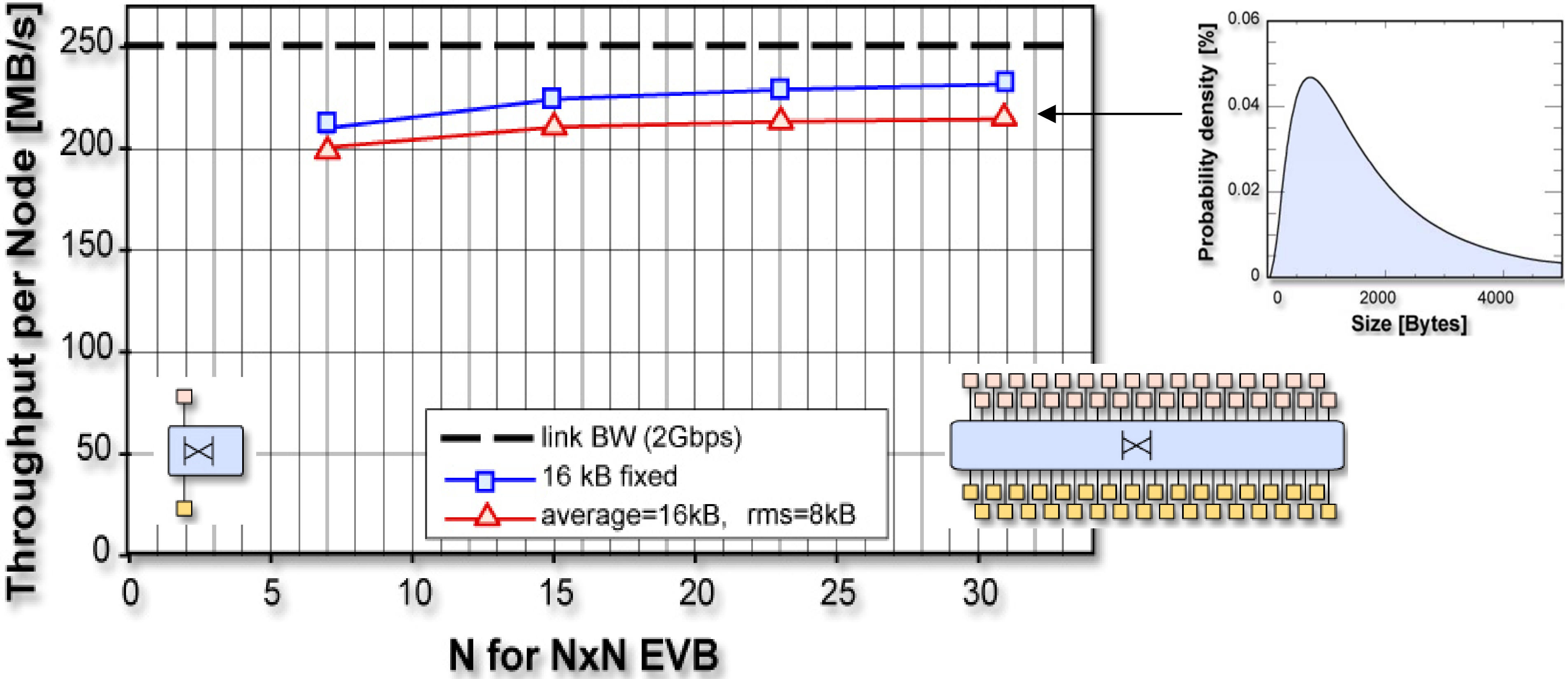




- **EVB 32x32**
- Throughput at **234 MByte/s**
  = **94% of link Bandwidth**



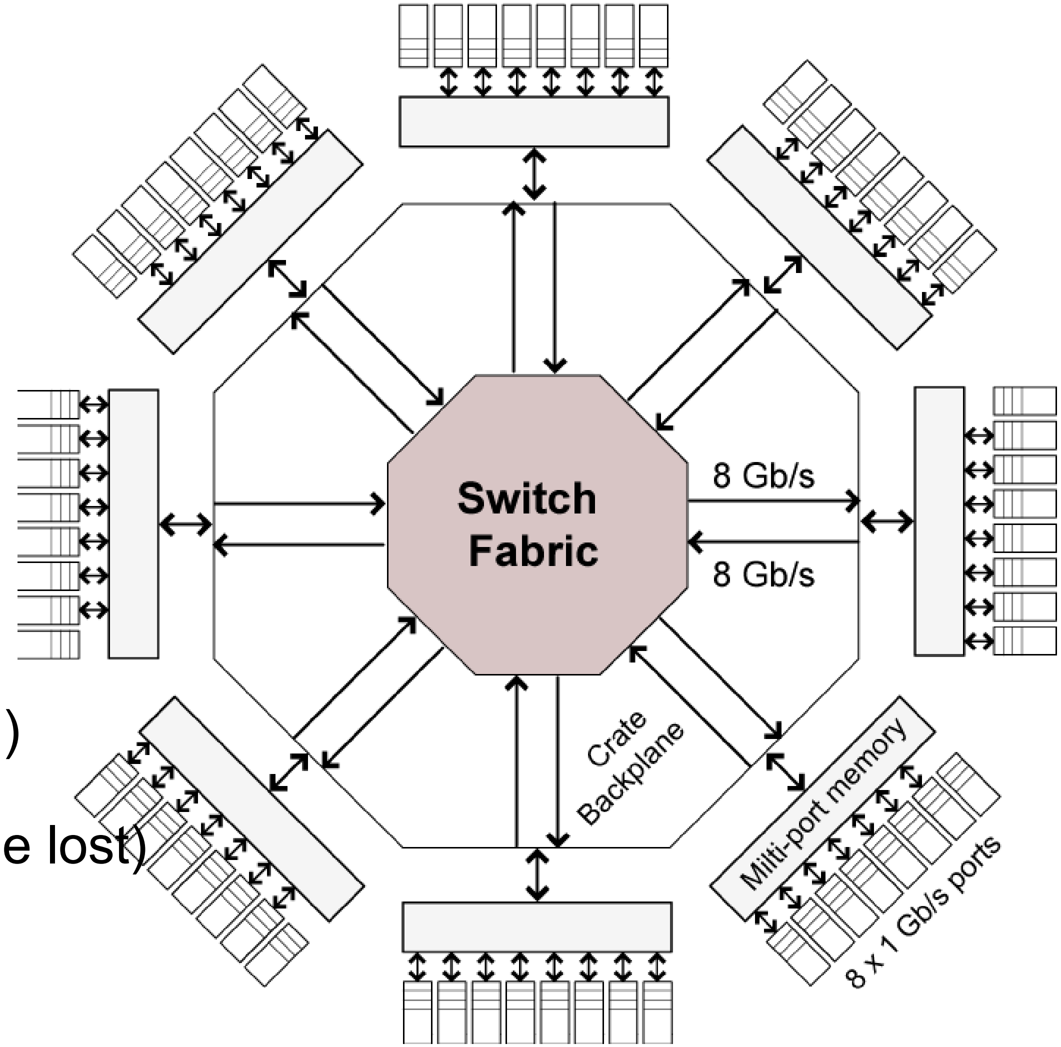
Barrel shifter EVB scaling



From 8x8 to 32x32: Scaling observed (as expected from barrel shifter)

Aggregate EVB Throughput 32 x 200 MB/s = 6 GByte/s
Fully populated Clos-128 (64x64 EVB): 12 GByte/s

GigaBit Ethernet



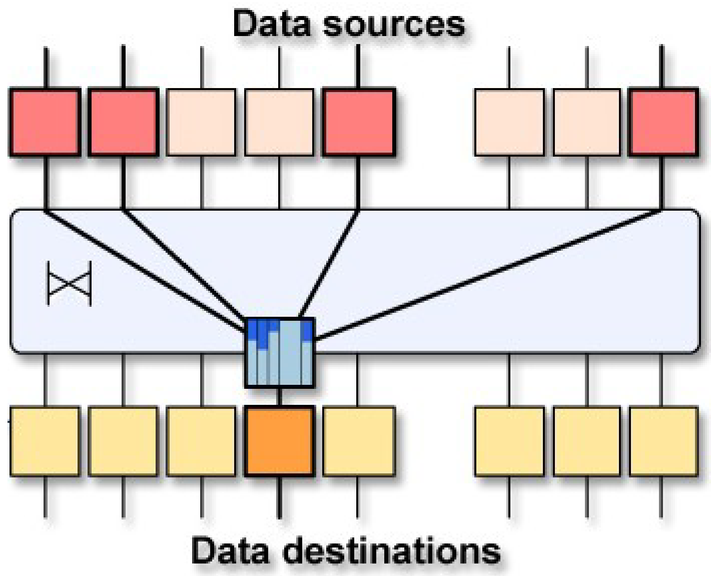
Switch: Foundry **FastIron8000**

- **32x32 x 1.0** Gbit/s port (**32 Gbit/s**)
- **Packet switching.**

  Input/Output buffers (packets can be lost)

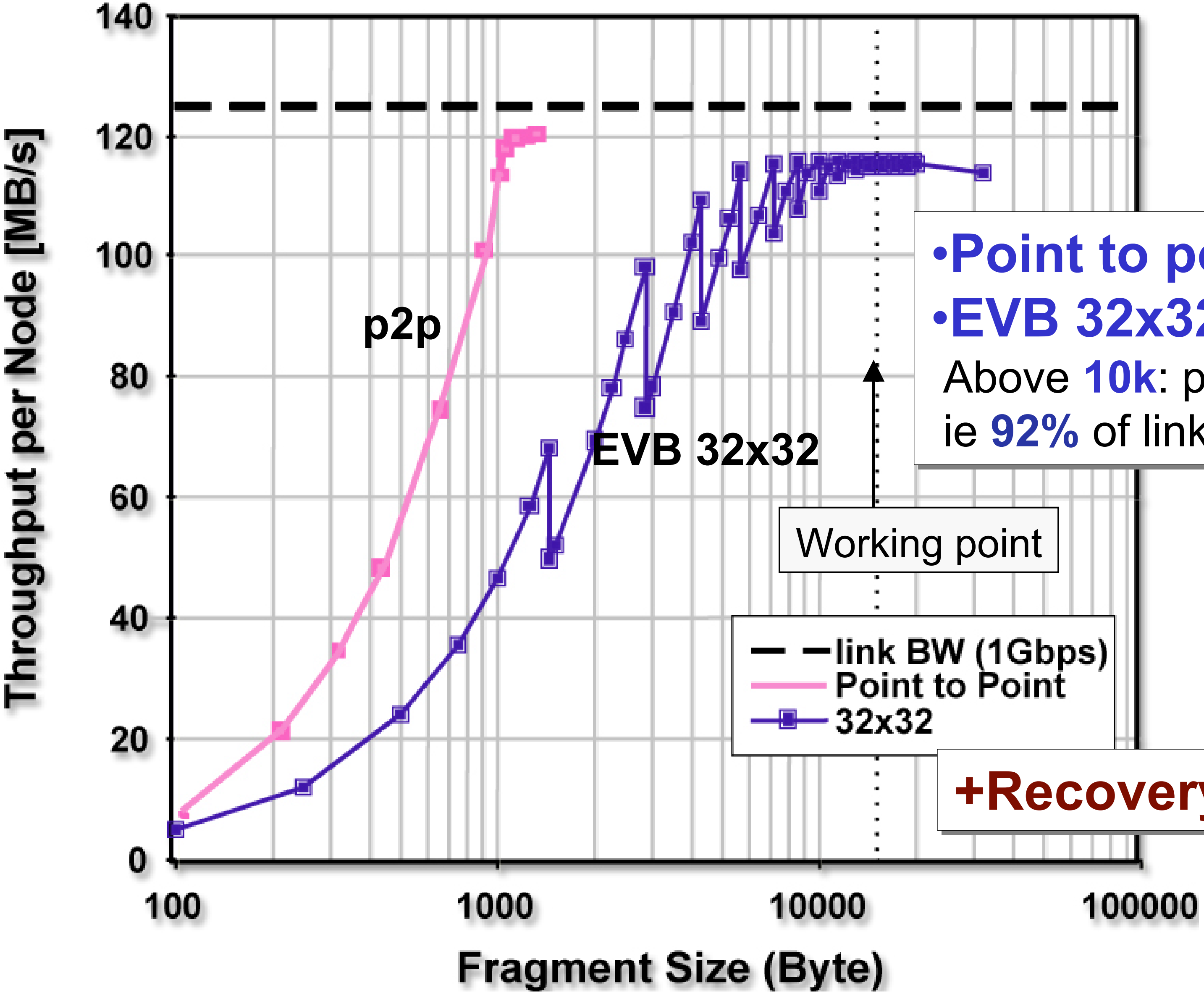
NIC: **Alteon** AceNIC (running standard firmware)

GbE: Destination based Traffic Shaping



- Ethernet switches typically have **no flow-control** through the switch
- **Packet loss** when buffer capacity exceeded during bursty traffic
- Solution:
  - EVB protocol is **destination driven** (pull)
  - Limit or avoid loss by **requesting limited number of packets**
  - done at EVB application level
- Depends on **internal switch architecture** (sizes of memory buffers)

GbE-EVB: raw packet & special driver



- •**Point to point 120 MB/s**
- •**EVB 32x32** sawtooth due to MTU.
  Above **10k**: plateau of **115 MB/s**
  ie **92%** of link speed (1Gbps)

+Recovery from packet loss

- Alteon AceNIC
- FastIron-8000
- standard MTU (1500 B payload)
- PentiumIII Linux2.4

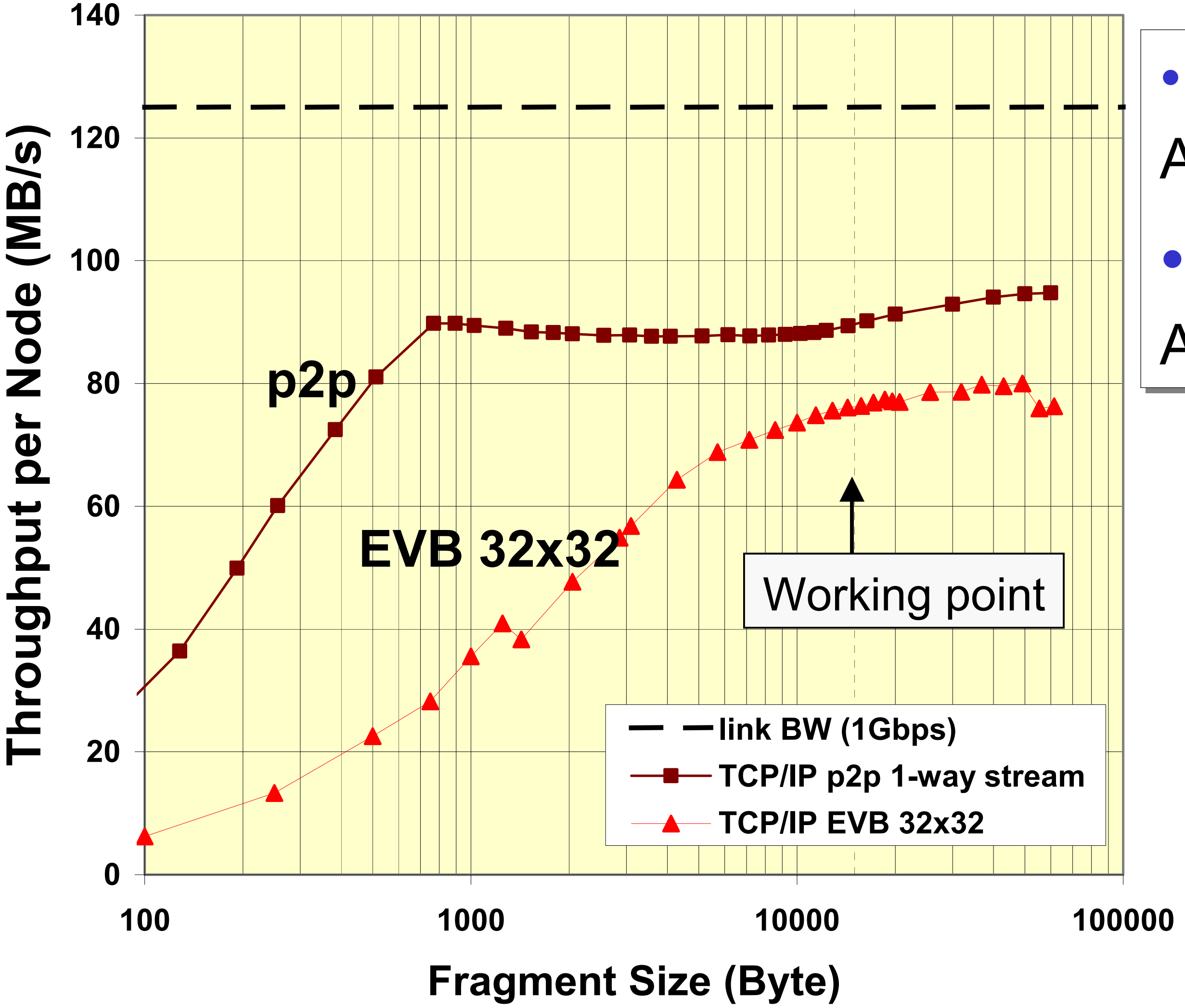
GbE EVB: Raw Packets versus TCP/IP

	Layer-2 Frames	*TCP/IP*
host	computer or more basic (eg FPGA)	computer
reliability	packet loss if congestion	reliable
zero-copy	yes	no
CPU usage	low	high (rule: 1 Hz per bps)
EVB - traffic shaping	required	maybe
EVB - recovery	at application level	built-in
EVB - latency	medium	high

Assumes:

- want EVB throughput close to wire speed
- switch does not propagate flow control end-to-end (typical)

GbE-EVB: TCP/IP full standard



- **Point to point :**

At **1 kB** 88 MB/s (**70%**).

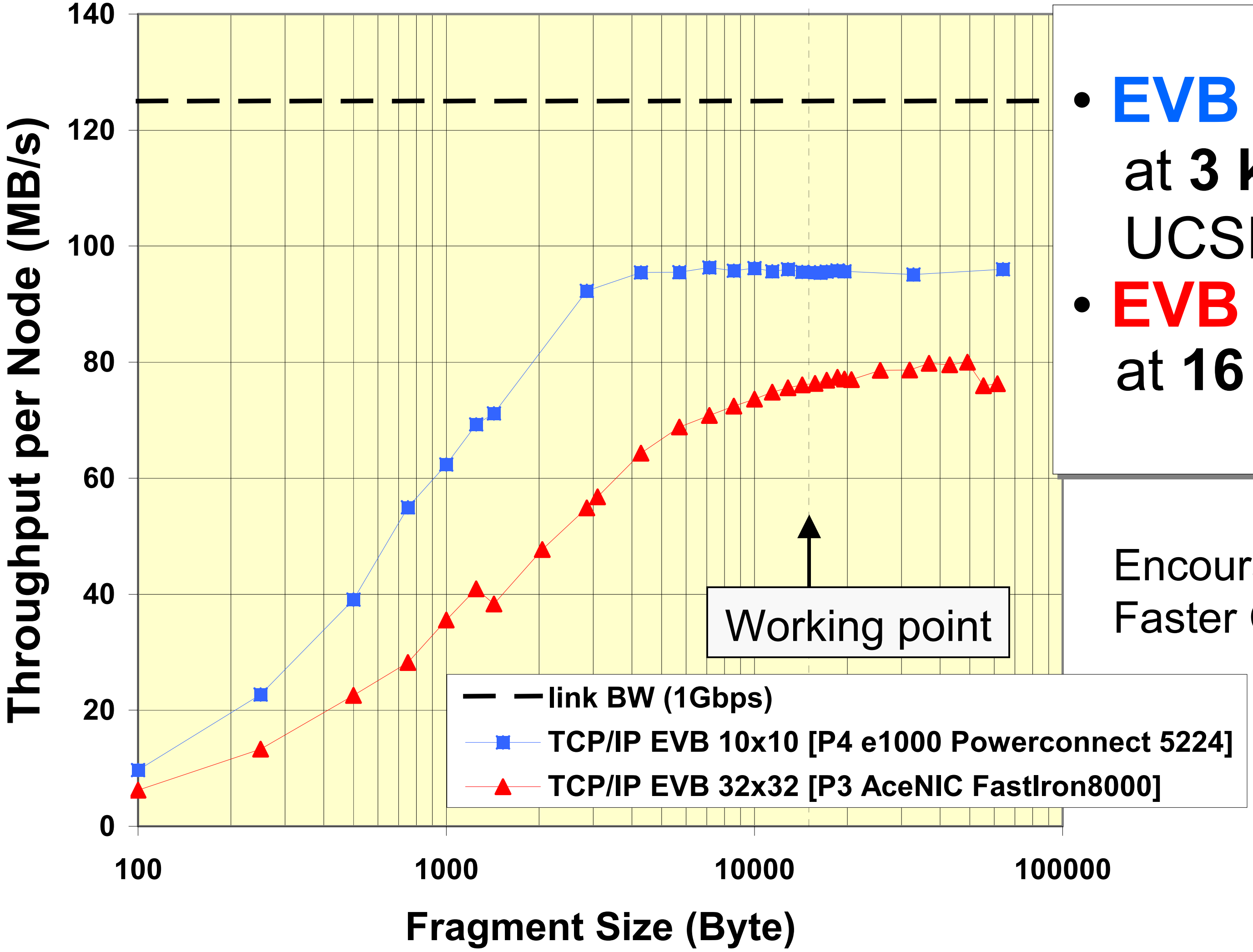
- **EVB 32x32 :**

At **16 kB** 75 MB/s (**60%**)

CPU load ~ 100%

- Alteon AceNIC
- FastIron-8000
- standard MTU (1500 B payload)
- PentiumIII Linux2.4

GbE-EVB: TCP/IP – 2003 equipment



- **EVB 10x10 2003 HW:** at **3 kB** 95 MB/s (**75%**) UCSD T2 (preliminary)
- **EVB 32x32 2001 HW:** at **16 kB** 75 MB/s (**60%**)

Encouraging: Jumbo frames, Faster CPU, better TCP stack impl.

- standard MTU (1500 B payload)
- Linux2.4

EVB software: see Parallel 5 ; “Using XDAQ in application scenarios of the CMS experiment”

EVB demonstrators summary

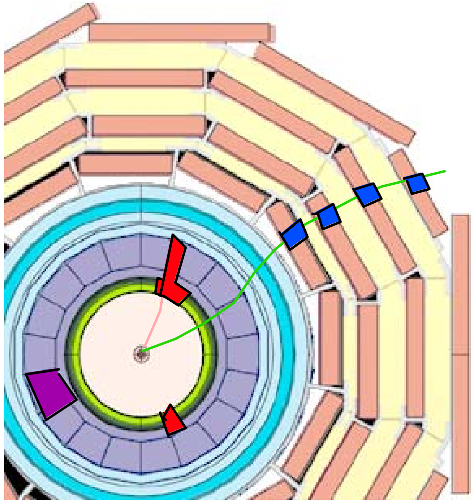
	Myrinet 2000	GbE raw packet	GbE TCP/IP
Test bench	32x32	32x32	32x32
Port speed	2.0 Gbit/s	1.0 Gbit/s	1.0 Gbit/s
Random traffic	30%	50%, 92% (*)	30%, 60% (*)
Barrel shifter	94%	-	-
CPU load	Low	Medium	High
1 Tbit/s EVB	512x512	1024x1024	1536x1536
No. switches	8 128-Clos	16 256-port	24 256-port

Industry standards

Proprietary standards

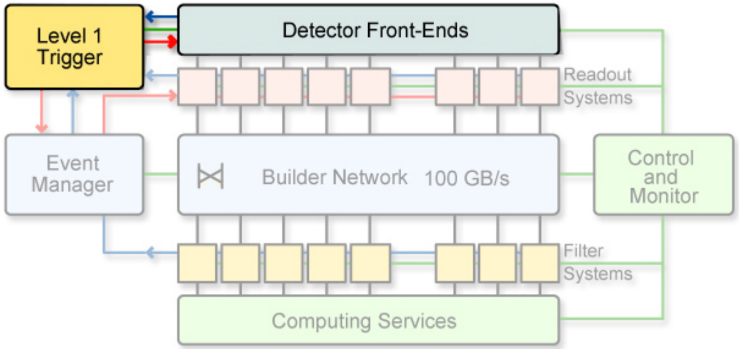
(*) with fragment sizes **larger than 16kB**

Two trigger levels

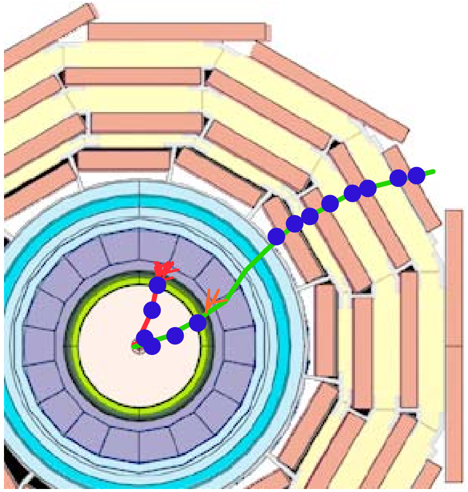


Level-1: Specialized processors
40 MHz synchronous

- Local pattern recognition and energy evaluation on prompt macro-granular information from **calorimeter** and **muon** detectors

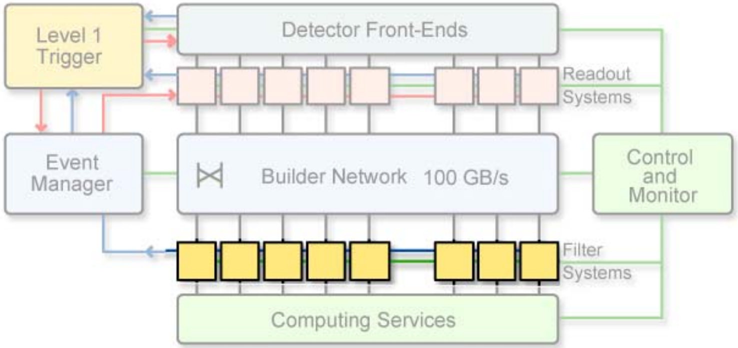


99.99 % rejected 0.01 % Accepted



High trigger levels: CPU farms
100 kHz asynchronous farms

- “off-line” code
- HLT has access to **full event data** (full granularity and resolution)
- Only limitations:
  - **CPU time**
  - **Output selection rate (~10^2 Hz)**
  - **Precision of calibration constants**



100-1000 Hz. Mass storage
Reconstruction and analysis.

99.9 % rejected 0.1 % Accepted

High Level Trigger: CPU usage

- Based on full simulation, full analysis and “offline” HLT code
- All numbers for a 1 GHz, Intel Pentium-III CPU

Trigger	CPU (ms)	Rate (kHz)	Total (s)
1e/γ, 2e/γ	160	4.3	**688**
1μ, 2μ	710	3.6	**2556**
1τ, 2τ	130	3.0	**390**
Jets, Jet * Miss-E_T	50	3.4	**170**
e * jet	165	0.8	**132**
B-jets	300	0.5	**150**

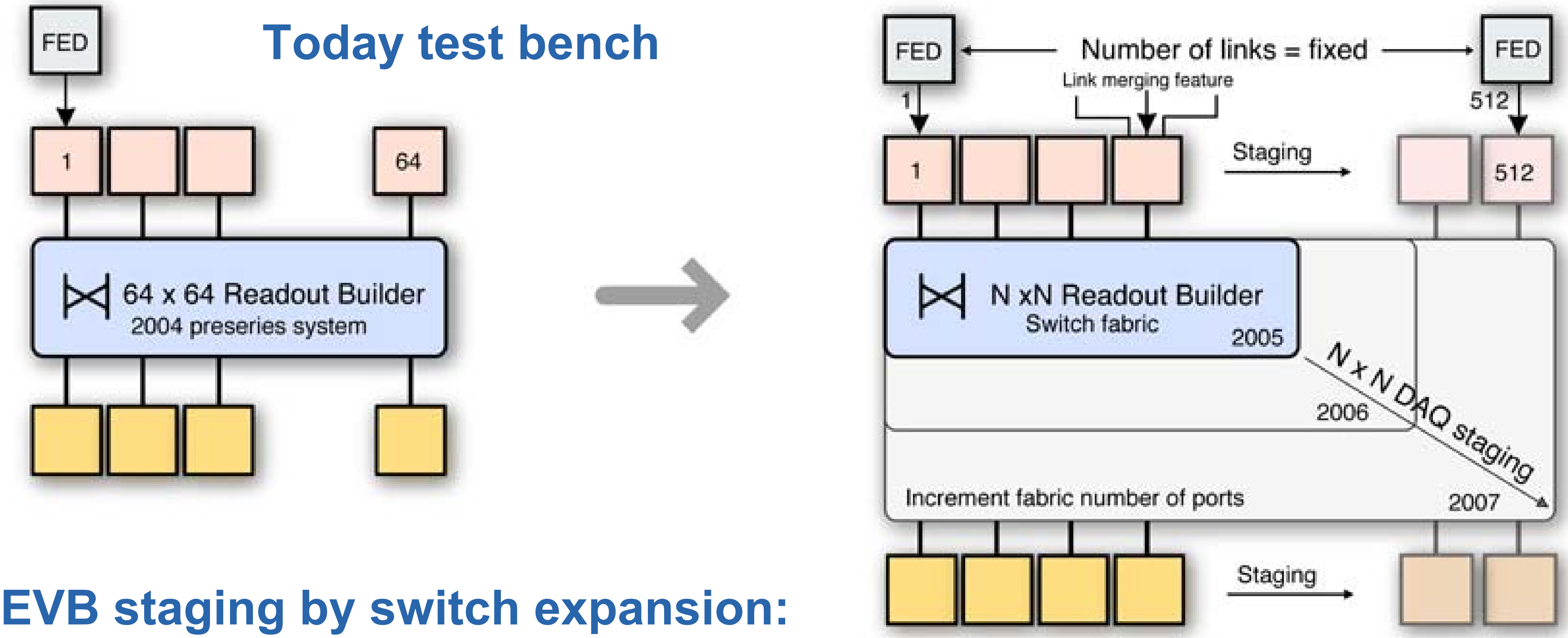
Total: 4092 s for 15.1 kHz → 271 ms/event
Expect improvements, additions.
Therefore, a 100 kHz system requires 1.2x10^6 SI95
Corresponds to 2,000 dual-CPU boxes in 2007
(assuming factor 8 from Moore’s law)

Full EVB;
Scaling and Staging Issues

From demonstrator to final EVB

EVB **staging**: commissioning 2006; low lumi 2007; high lumi 2009?

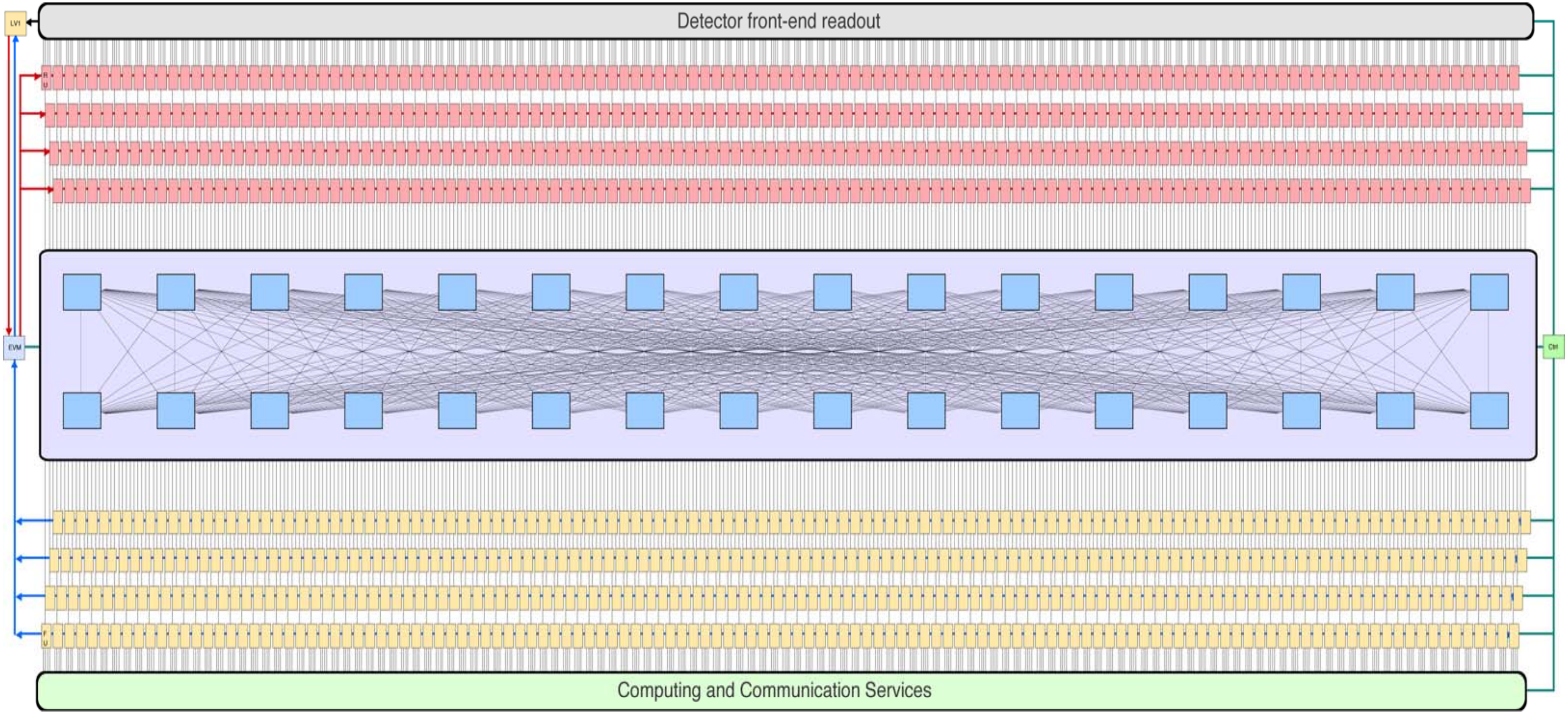
Today test bench



EVB staging by switch expansion:

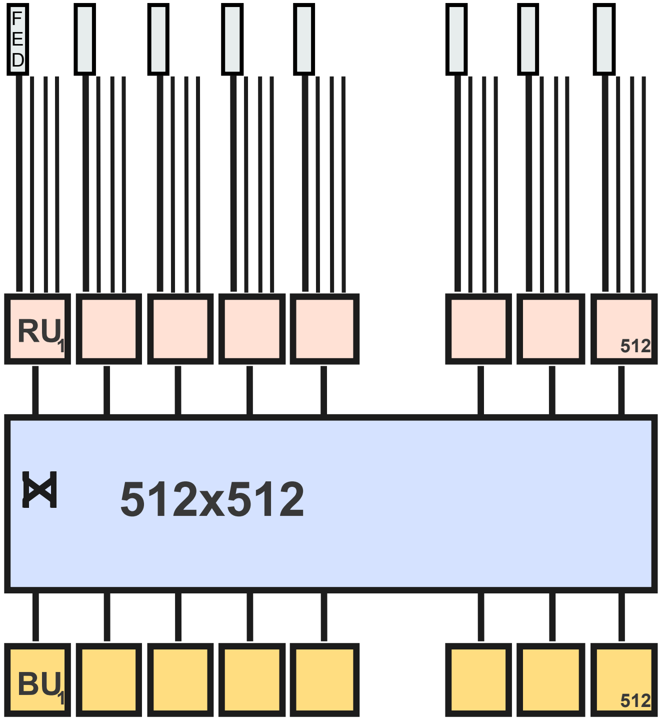
- Readout unit must allow multi-FED link merging
- Expand the switch via a switch fabric structure
- **Early choice of technology (2004)**
- EVB stages are based on the **same technology**
- Performances must scale with size. **To be demonstrated, today only by simulation**
- System failures are highly factorized (failures in one RU or one switching node halt the entire system)

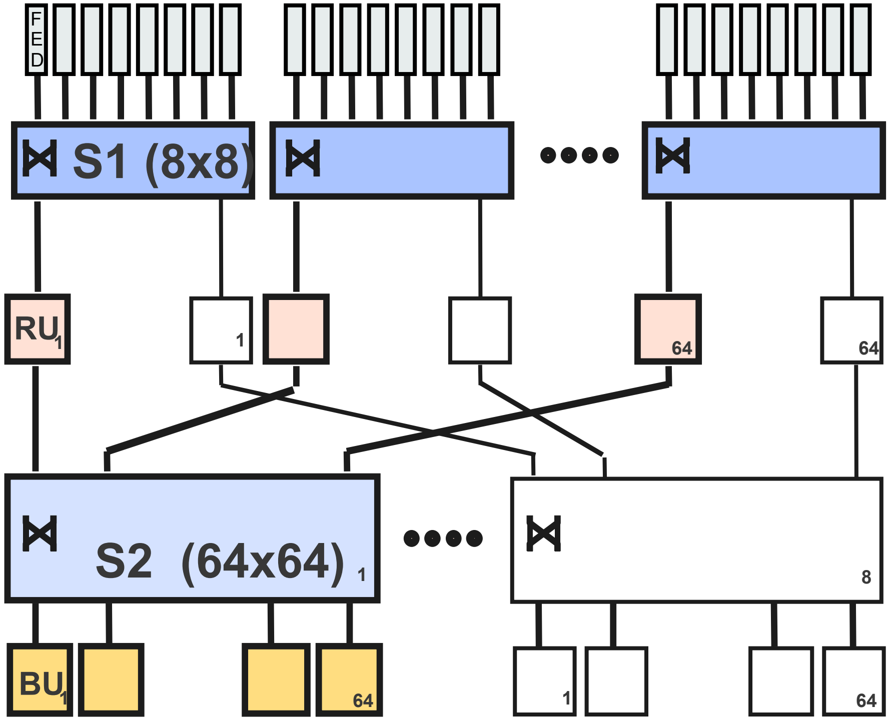
EVB 512x512 (out of 32x32)


LV1
Detector front-end readout
R U
EVM
Ctrl
F U
Computing and Communication Services

- Performance scaling (by factor 10)?
- Fault tolerance?

From Demo to Final EVB: Two-stage EVB





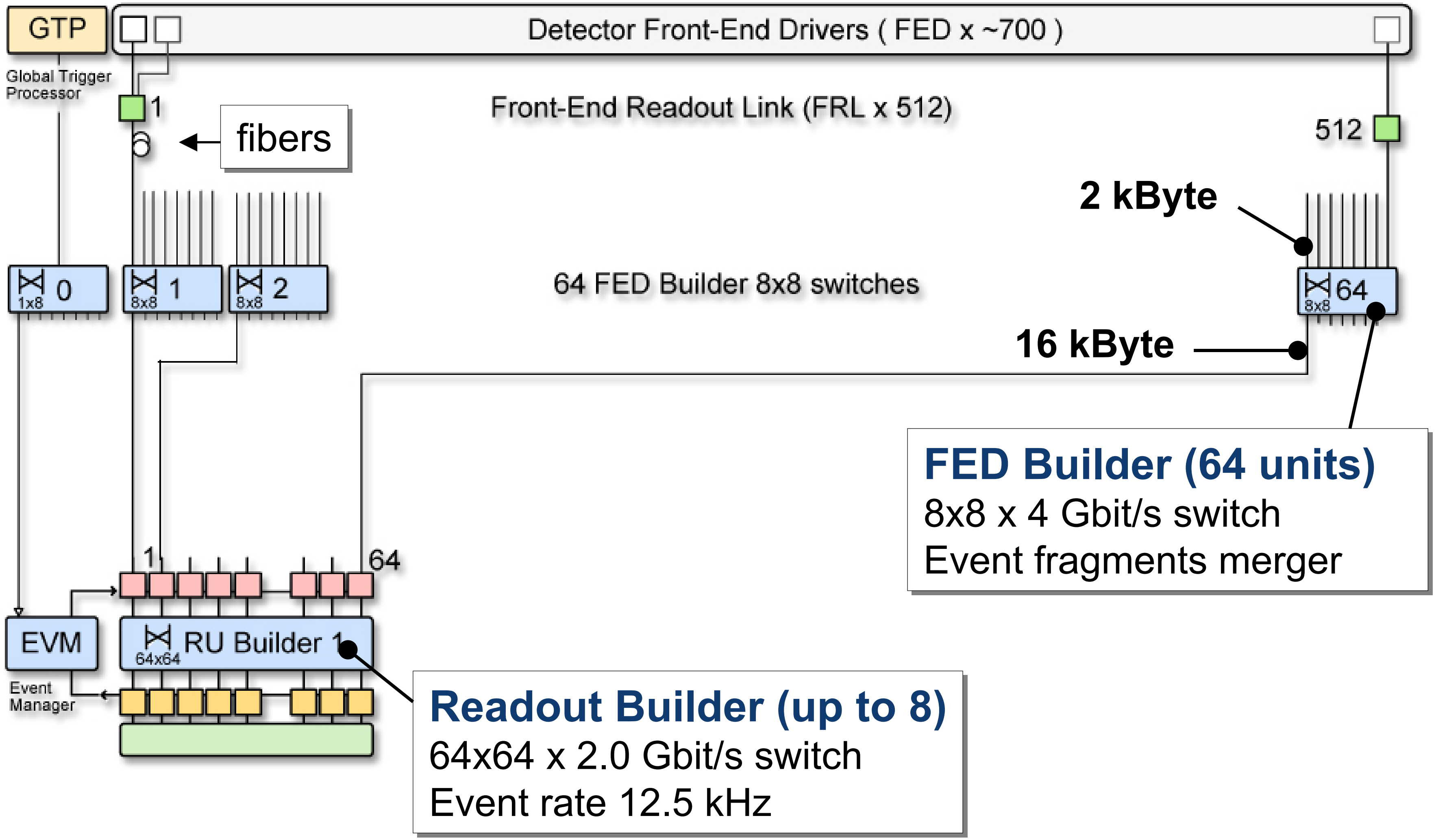
Large monolithic switching fabric

Two stages, separated by large intelligent buffers (PCs)

- Stage One (pre-builder): 8x8 acts as concentrator and multiplexer
- Stage Two (final-builder): 64x64

2 stages: Data to surface & Readout Builder

Data to surface (rate decimation)



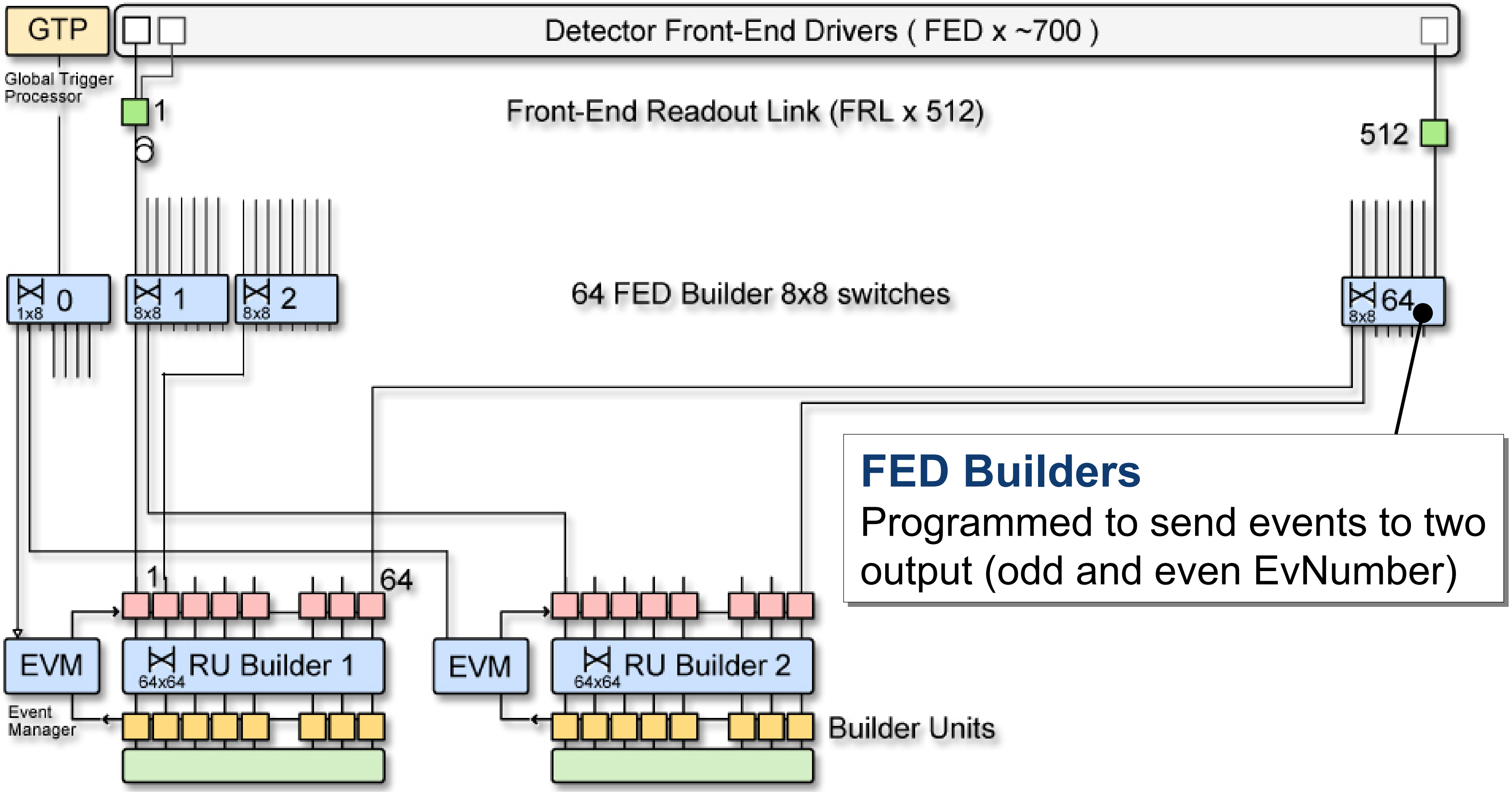
FED Builder (64 units)
8x8 x 4 Gbit/s switch
Event fragments merger

Readout Builder (up to 8)
64x64 x 2.0 Gbit/s switch
Event rate 12.5 kHz

DAQ staging : 2 RBs = 25 kHz

Data to surface (rate decimation)

Readout Builders (modular staging)

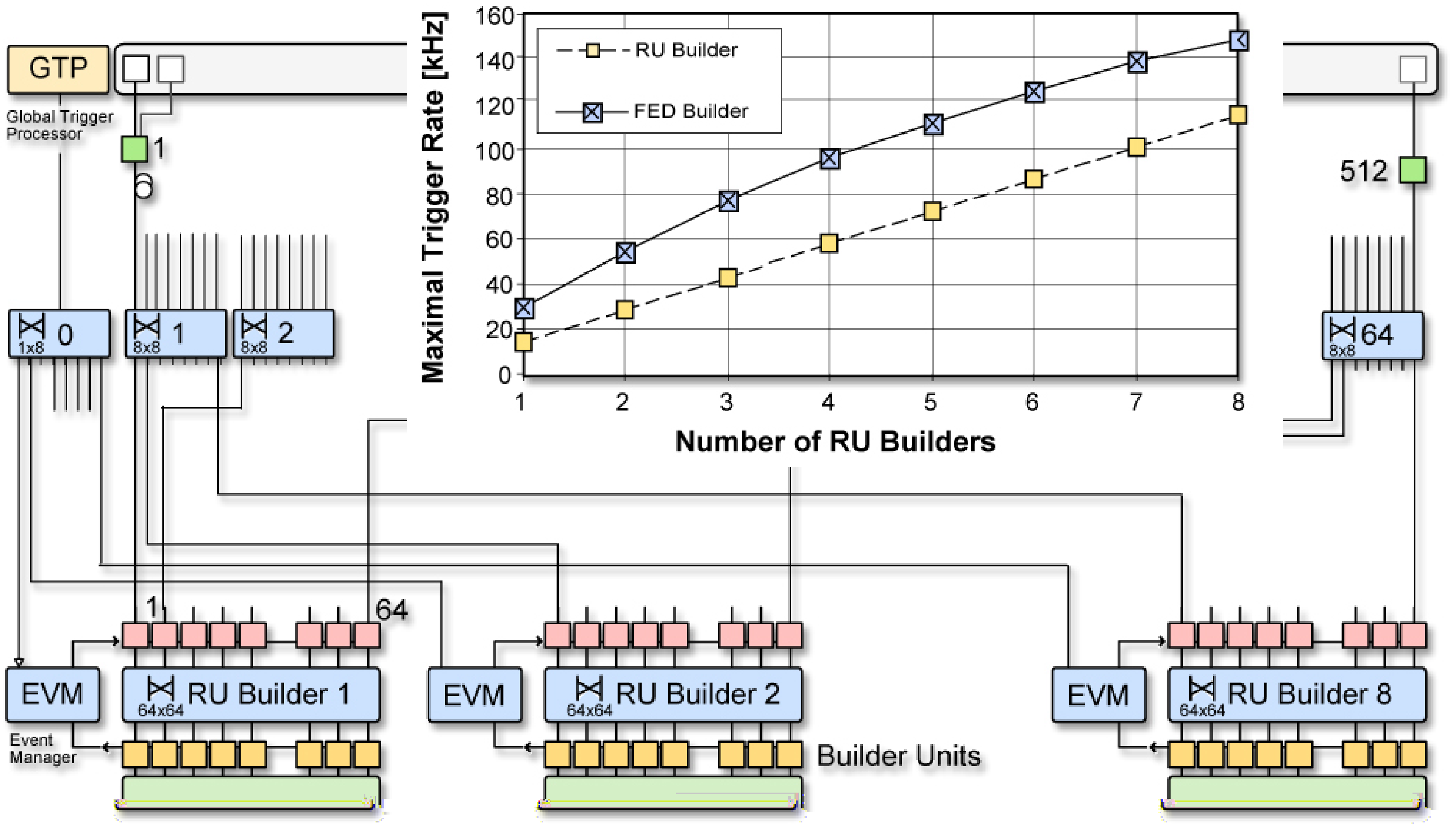


FED Builders
Programmed to send events to two output (odd and even EvNumber)

DAQ staging : 8 RBs = 100 kHz

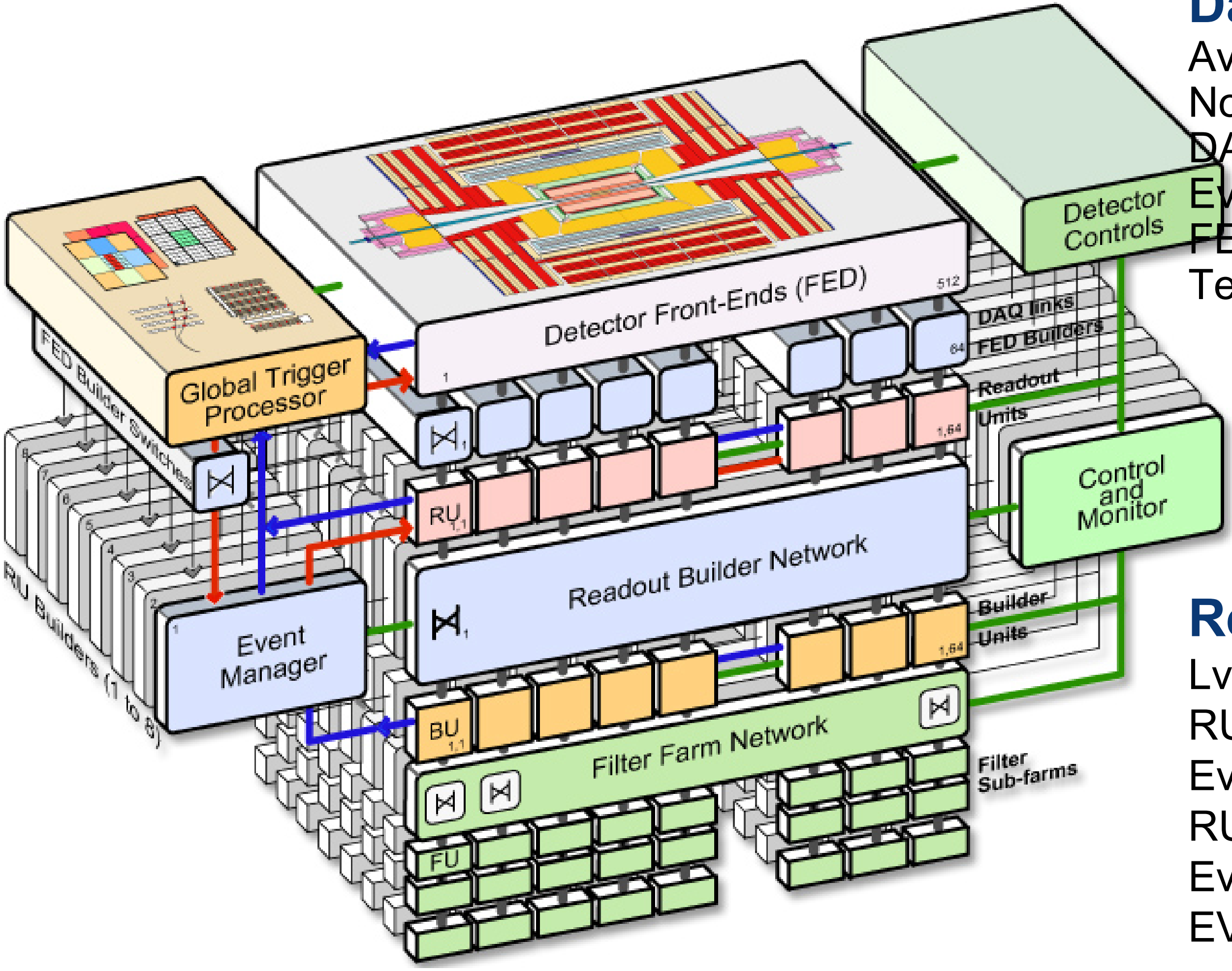
Data to surface (rate decimation)

Readout Builders (modular staging)



Readout Builders are not necessarily based on the same technology

DAQ staging and scaling: 8 x (64x64)



Data to surface (pre-builders):

Average event size	1 Mbyte
No. FED S-link64 ports	700
DAQ links (2.0 Gbit/s)	512+512
Event fragment size	2 kB
FED builders (8x8 dual)	64
Technology(2004)	Myrinet

Readout Builders (x8):

Lv-1 max. trigger rate	12.5 kHz
RU Builder (64x64)	.125 Tbit/s
Event fragment size	16 kB
RU/BU systems	64
Event filter power	10^5 SI95
EVB technology (2006)	Open

8 x (12.5 kHz DAQ units)

Conclusion

The presented DAQ design fulfills the major CMS requirements:

✓**100 kHz level-1 readout**

✓**Event builder:**

- **Built full events**
- **A scalable structure that can go up to 1 Terabit/s**

✓**High-Level Trigger:**

- **By commodity processors having access to full event data**
- **Single-farm design providing maximum flexibility in the physics selection**