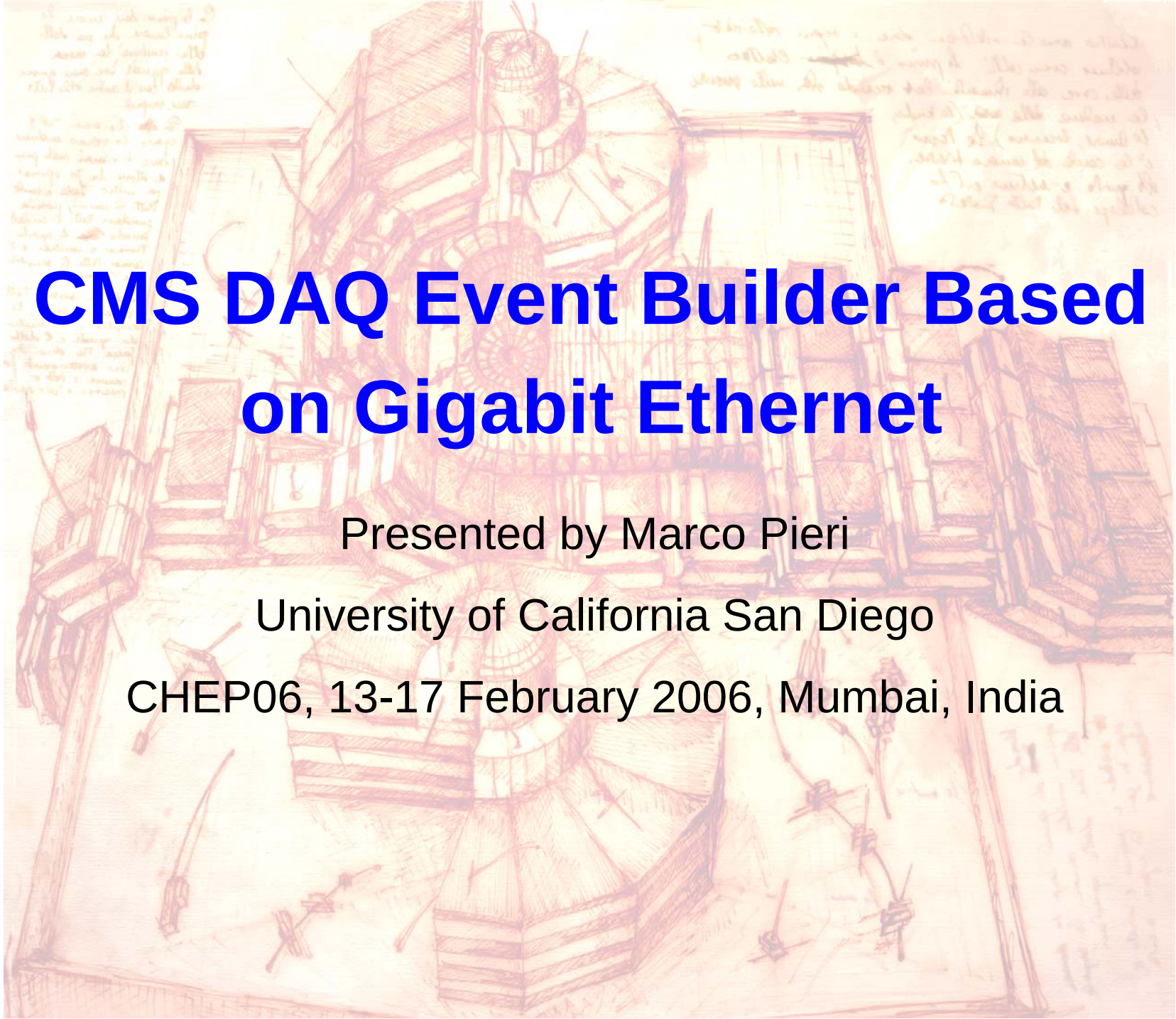


CMS DAQ Event Builder Based on Gigabit Ethernet

Presented by Marco Pieri

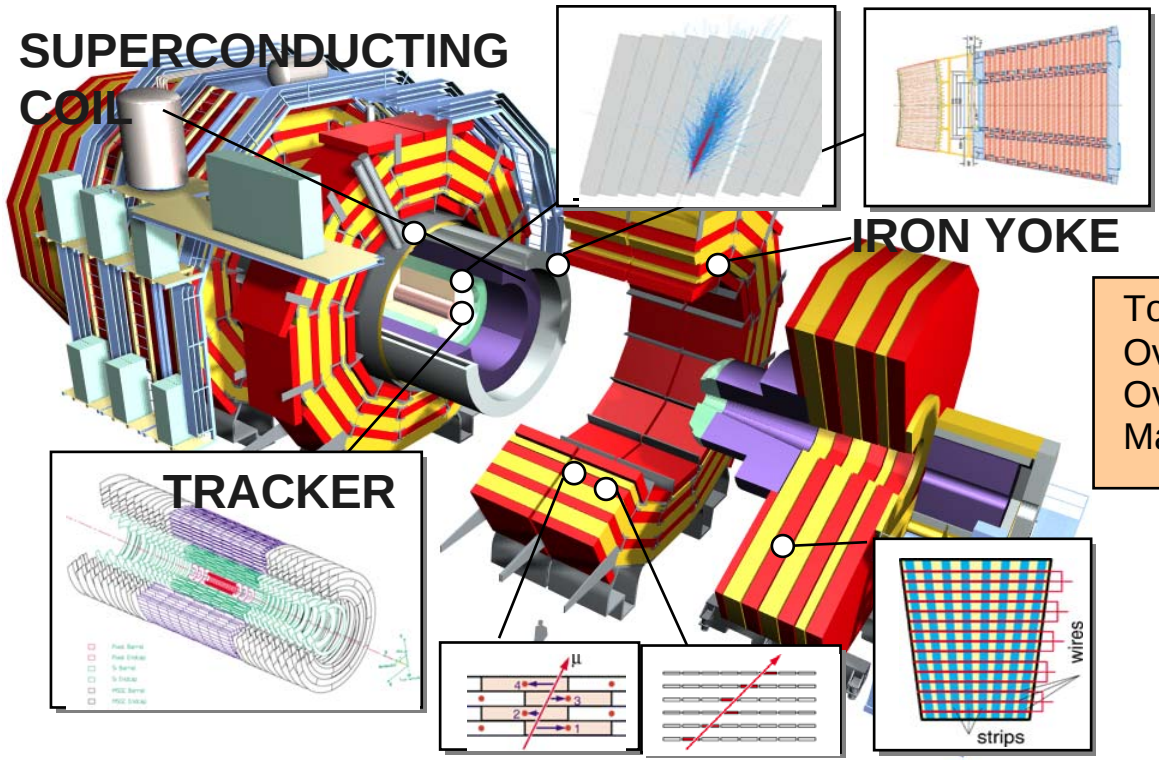
University of California San Diego

CHEP06, 13-17 February 2006, Mumbai, India

Outline

- Introduction
- CMS DAQ architecture
- Readout Builder and Gigabit Ethernet networking
- Results of the tests
- Summary

DAQ Requirements

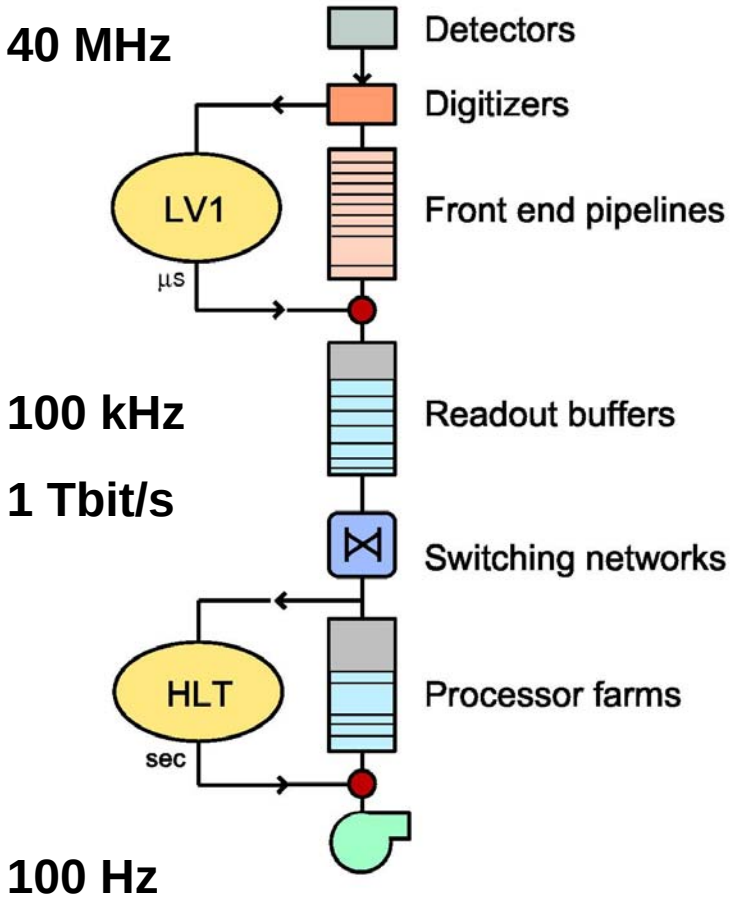


Total weight : 12,500 t
Overall diameter : 15 m
Overall length : 21.6 m
Magnetic field : 4 Tesla

Beam crossing rate	40 MHz
Interaction rate	1 GHz
Max Level 1 trigger rate	100 kHz
Event size	1 MByte
EVB inputs	~700

CMS DAQ Structure

Only two physical trigger stages



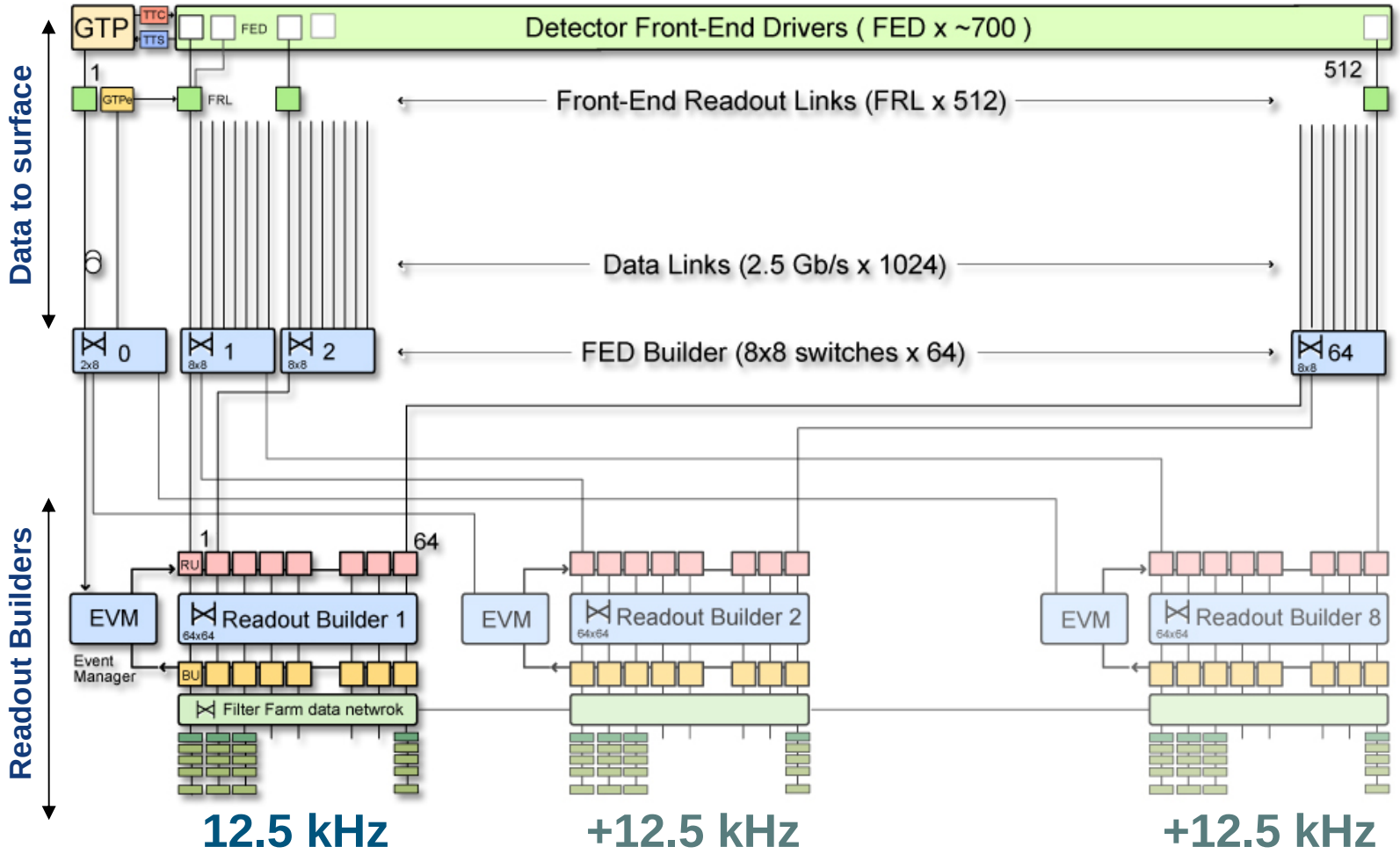
Level 1 trigger

Custom design, implemented in ASIC, FPGA

High Level Trigger (HLT)

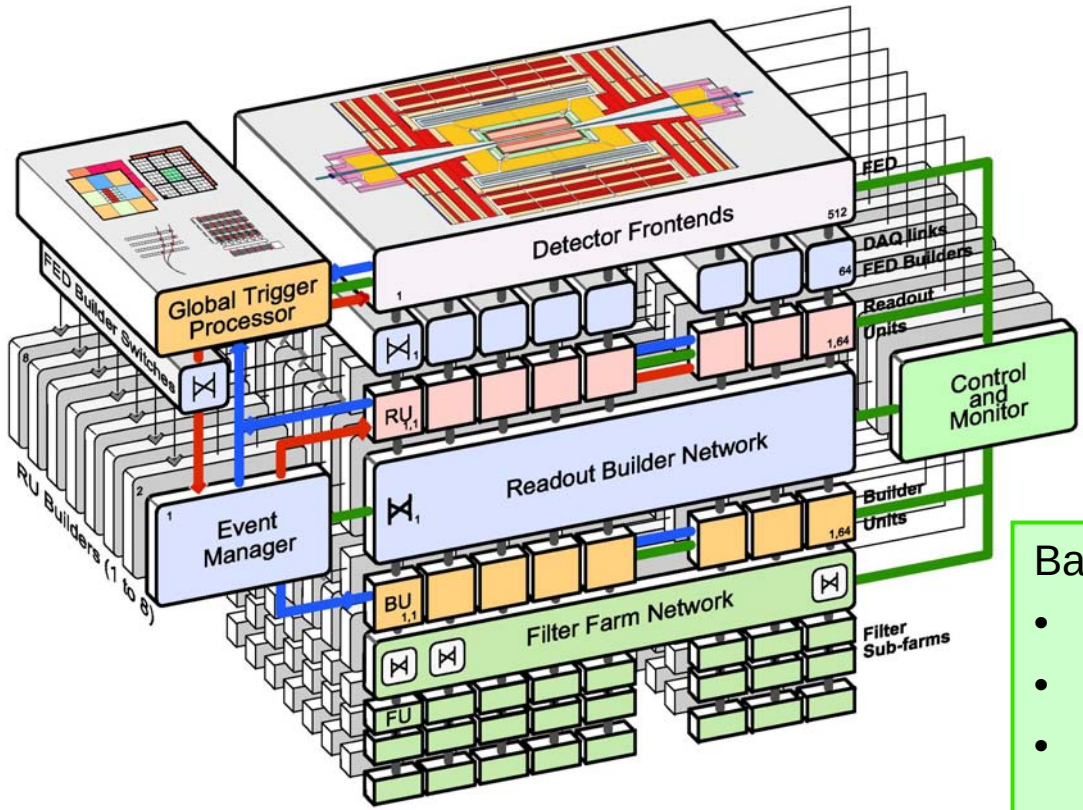
“Offline” code running on PC farm

EVB Architecture



- Events are built in two stages
- The required throughput is 200 MBytes/sec per node

EVB Architecture (II)



Baseline DAQ Configuration

- 512 inputs
- 64 8x8 FED Builders
- 8 64x64 RU Builder slices:
  - 1 EVM
  - 64 RUs
  - 64 BUs
  - 256 FUs
- Probably for the first few years only 4 slices

Event Builder Networking

Myrinet (http://www.myri.com)

- Data rate 2 Gbit/sec per link, low latency
- Flow-control and backpressure in hardware
- No load on the PCs, exploit the NIC CPU
- Relatively easy to implement custom drivers in the NIC

Gigabit Ethernet (GbE)

- Reliable transfer protocols widely used available (TCP/IP)
- Standard technology steadily improving
- CPUs are heavily loaded with TCP/IP

- FED Builder – Use Myrinet, it also provides a rather cost effective optic fiber solution for the Data to Surface data transfer
- Readout Builder – Myrinet or GbE
- Filter data network – GbE –Unpractical to have Myrinet on the FUs

Note: 10Gbit Ethernet not yet considered because still too expensive

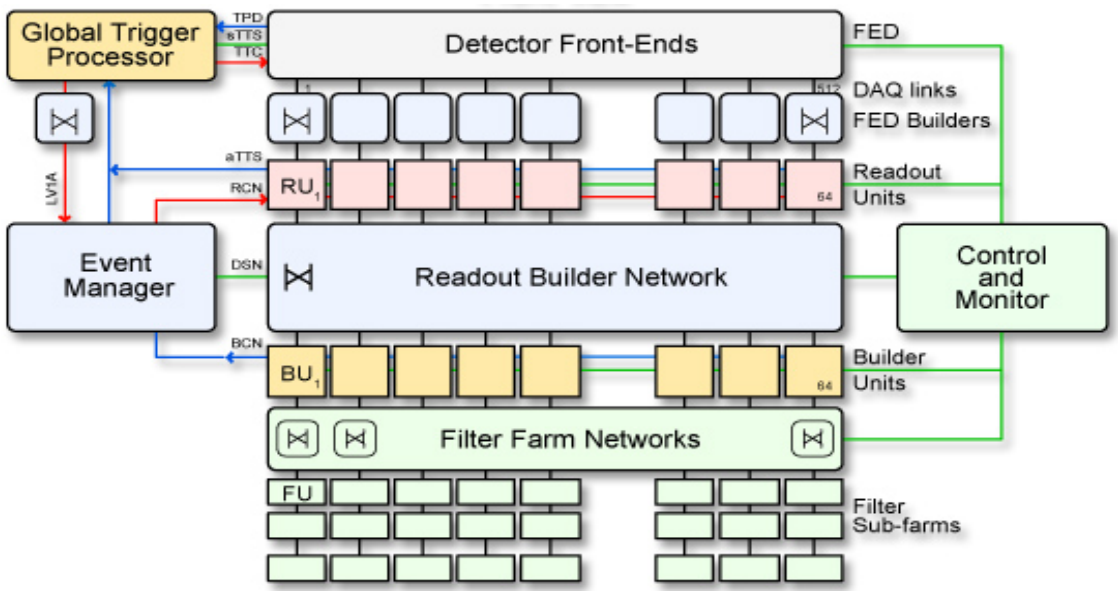
FED Builder

- Use Myrinet technology
- Link speed 2 Gbit/s for data
- Networking efficiency ~50%
- Use two links

- 8x8 FED Builder
- Throughput above 200 MByte/s at the nominal input fragment size of 2 kBytes
- Average output fragment size 16 kBytes
- Also possible to merge up to 16 inputs into 16 outputs

Readout Builder Options

- Baseline: RU, BU, FU in different PCs
  - 64+64 PCs needed for one slice



- Folded: RU-BU in the same PC
  - Only 64 PCs used for one slice
  - Exploit full duplex features of interfaces and switches
  - Throughput on the PCs is double (200 MB/s input Myrinet, 200+400 MB/s in+out GbE)
- Trapezoidal: BU-FU in the same PC

Readout Builder Based on Myrinet

- In 64 RU x 64 BU (or 32x32) configuration head of line blocking further reduces the throughput with no traffic shaping
- 2 links are not enough for the Readout Builder

- With older hardware we have reached 94% link utilization with traffic shaping (barrel-shifter) in the NIC driver
  - Sources divide messages into fixed size packets and cycle through all destinations
  - Sources send to destinations in such a way that in each moment they take different paths inside the switch
- Not yet implemented for the final two port Myrinet NICs

Gigabit Ethernet Readout Builder

- Gigabit Ethernet option
  - Uses standard hardware, firmware and software
  - Easy to upgrade to 10Gbit Ethernet when it will become affordable
- Use TCP/IP
  - No development and maintenance of drivers
  - Reliable protocol – no worry about packet loss that may occur when operating links at wire speed

- We must be able to use the GbE links with almost 100% efficiency, not a trivial task

- We addressed and solved all problems related to TCP/IP usage in large configurations
- We implemented a software component integrated inside our software framework for optimal use of TCP/IP

ATCP – Asynchronous TCP

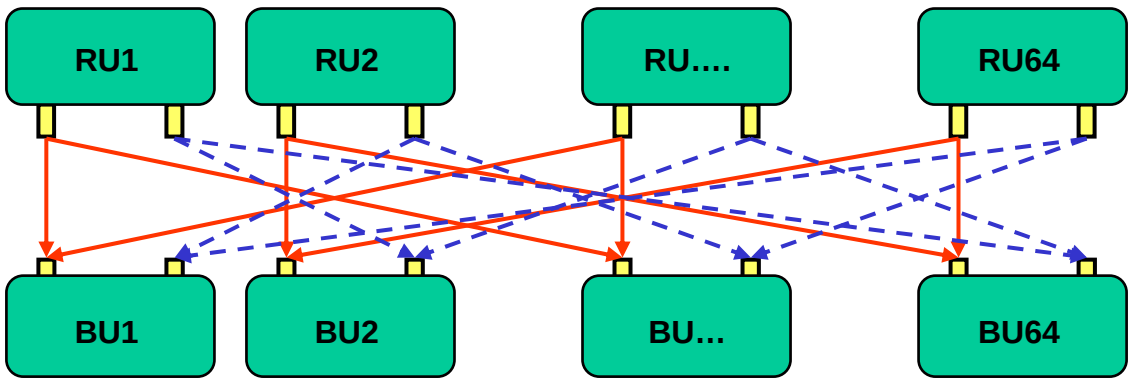
- ATCP is fully integrated in XDAQ, the CMS DAQ software framework
- XDAQ, C++ platform for a distributed Data Acquisition System, implements:
  - configuration (parametrization) and control
  - communication over multiple network technologies concurrently
  - high-level provision of system services (memory management, tasks, ...)

- The ATCP Peer transport
  - non blocking TCP/IP communication
  - Puts all messages to be sent in different queues and asynchronously processes them in another thread
  - for sending and receiving it operates until the current socket is blocked and as soon as it blocks it moves to another open socket and continues
- **No blocking observed for any tested configuration (~100% link usage)**

Design Choices for the Use of TCP/IP

- Use of Jumbo frames
  - By increasing the Maximum Transmission Unit (MTU) to ~7000 we observe an increase in performance of ~50%
  - Need switches that support Jumbo Frames

- Use of multiple links (multi-rail)
  - To achieve the design performance we need at least two GbE links
  - Use different physical networks depending on the source and destination host



Design Choices for the Use of TCP/IP (II)

- Difference between sending data and control messages
  - Control messages that steer the RU Builder operation must be delivered with low latency and have low throughput
  - TCP Nagle algorithm on gives sometimes large latency $O(10^{-2})$ sec when few messages in the pipeline
  - TCP Nagle algorithm off causes a throughput decrease with time when sending data
  - Optimal performance is obtained by using different sockets for data and control messages and setting:
    - Nagle on for data
    - Nagle off for control messages

Pre-series System

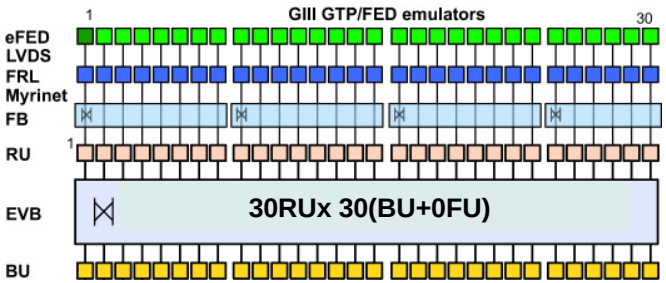
- Our current test-bed is installed at the LHC interaction point P5
- It is roughly equivalent to 1/16 of the full DAQ EVB system with all its functionalities

Readout Builder PCs;

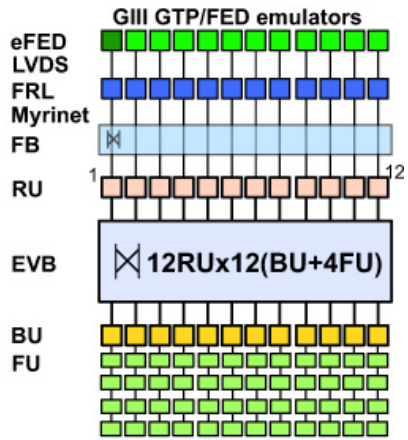
- 64 dual Xeon 2.6 GHz RU-BU PCs
- Myrinet +GbE interfaces
- 16 dual Xeon 2.6 GHz Filter nodes
- OS Linux 2.4

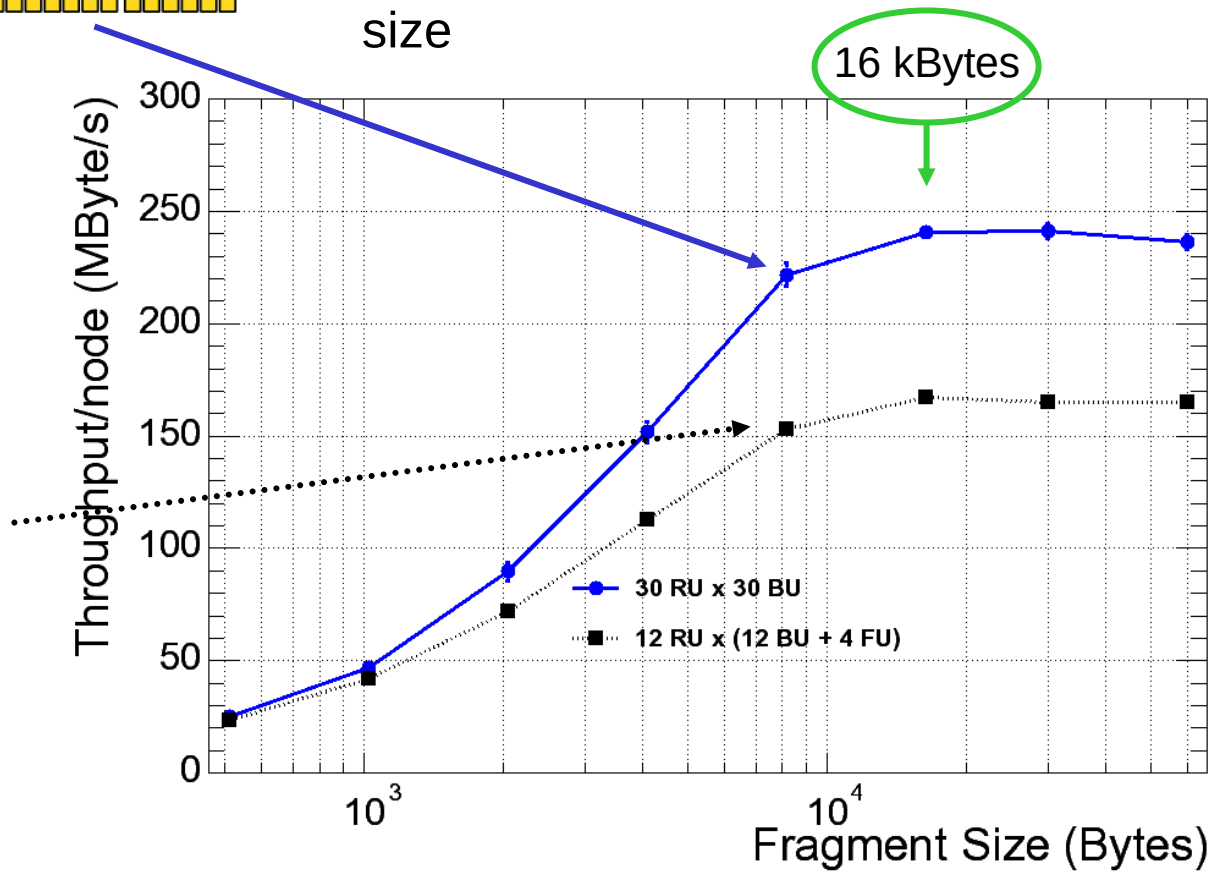
– Interconnected with GbE with a FORCE10 E1200 switch
– The switch can also be used in the first RU builder slices

Baseline Readout Builder Configuration

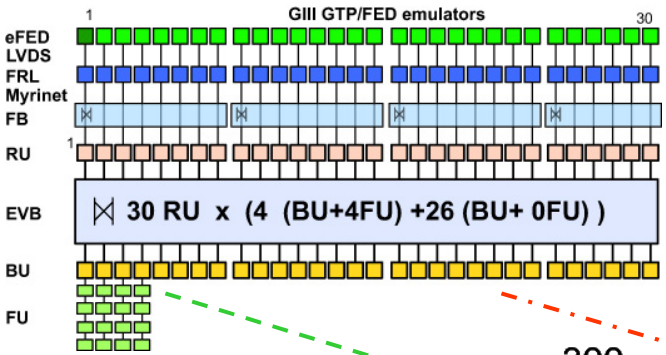


- Use FED Builder Myrinet input
- BU connected with FUs or not
- BU with FUs have a throughput of 165 MByte/s at the nominal fragment size

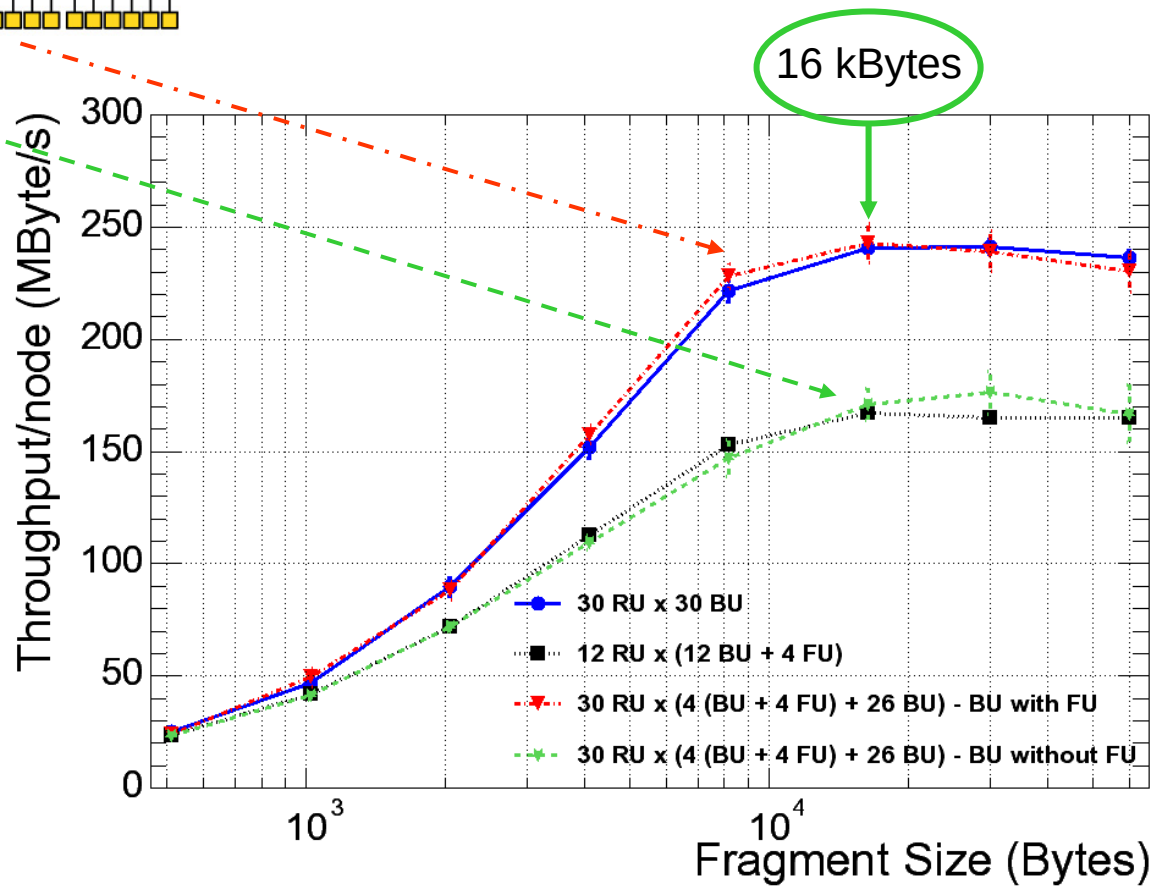




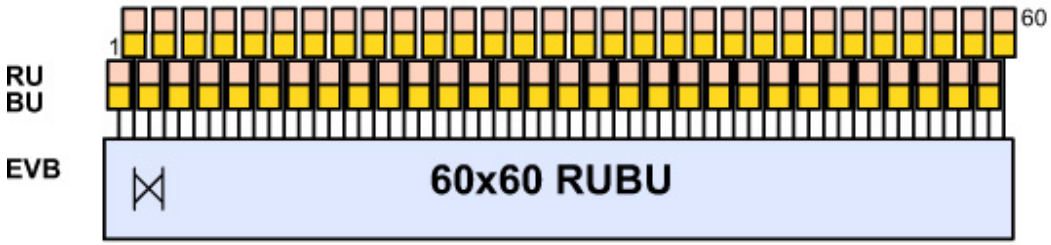
Baseline Readout Builder Configuration (II)

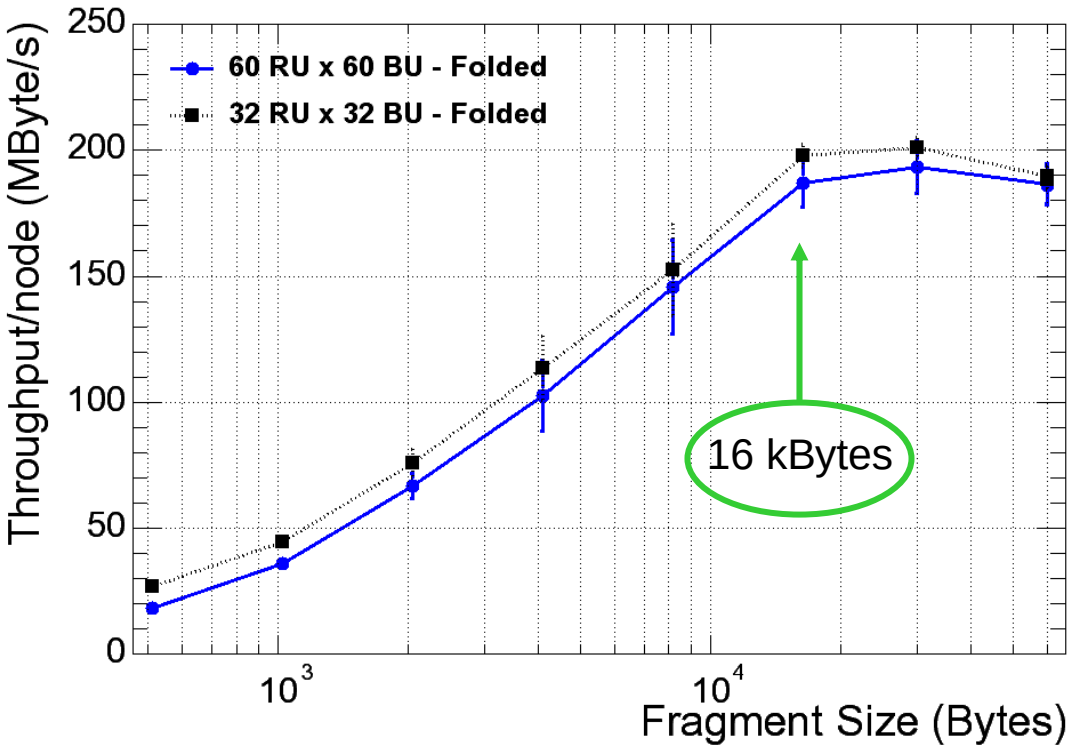


- Configuration 30 RU x 28 BU x 16 FU (4 BUs have 4 FUs each)
- No limitations from the network, only from PC resources



Folded RU Builder Configuration

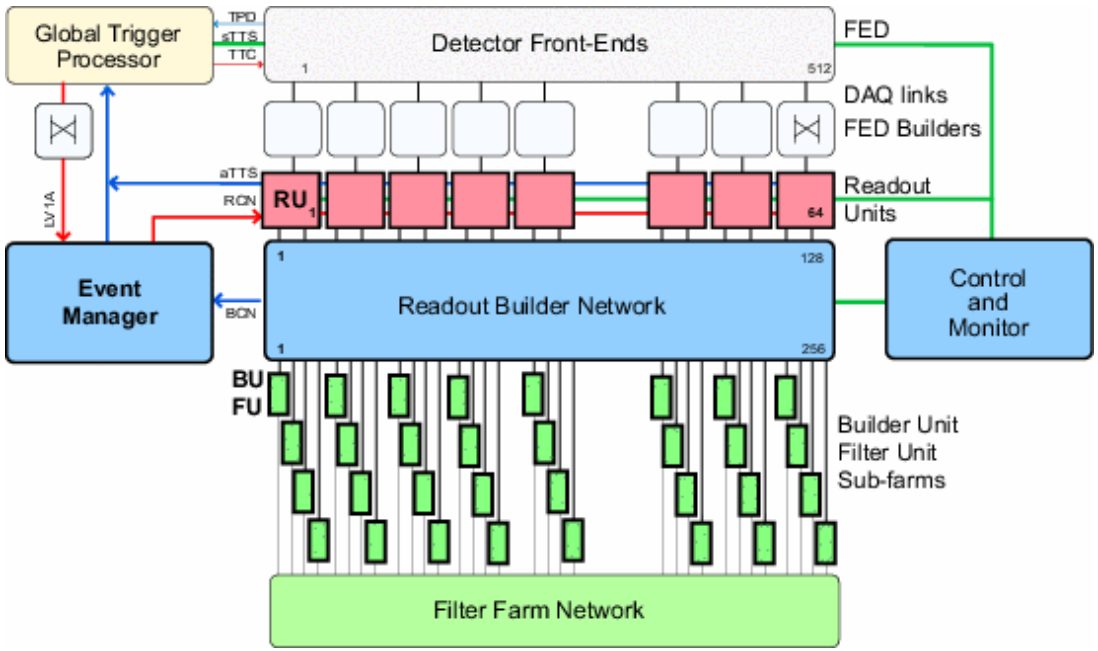




- No Myrinet input to the RUs, no GbE output to the FUs
- Verified scaling and full duplex performances for the RU-BU communication up to 60 RU x 60 BU
- Full duplex does not introduce any penalty

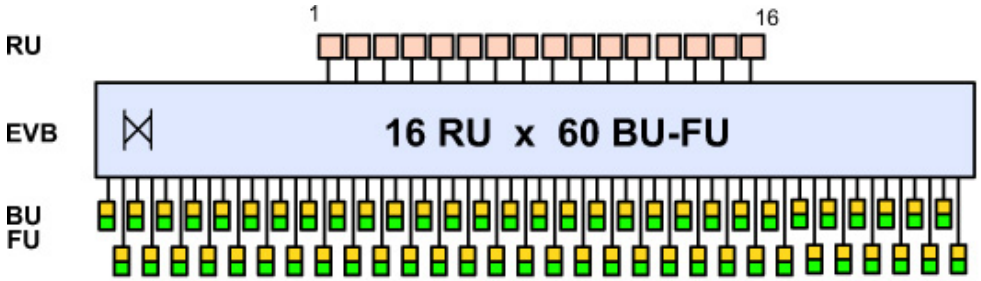
Trapezoidal RU Builder

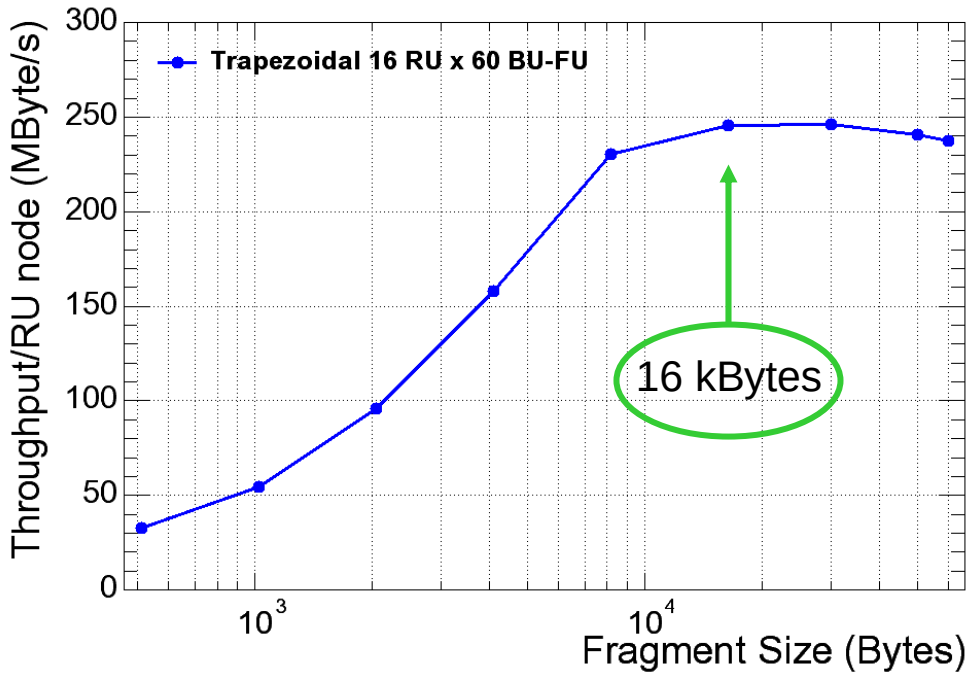
- In the baseline configurations Builder Units are the most heavily loaded part, TCP sending/receiving
- Events are built inside the FU PCs (BU-FU)



- The number of connections from a RU increases from 64 to 256
- The lower throughput on the switch output ports reduces the stress on the switch – possible to use oversubscribed switches
- CPU load on the FUs due to event building is 5-10%

Trapezoidal Readout Builder Test





- Tested almost 1/4 of the final RU Builder 16 RU x 60 BU-FU system
- Throughput ~240 MByte/s at the nominal fragment size
- To approach the full scale test in terms of number of connections we also run multiple BU-FU applications in each of the BU-FU PCs – Similar results up to 16 RU x 240 virtual BU system

GbE RU Builder Performance

- Baseline Readout Builder (nRUs=nBUs) currently at ~165 MByte/sec for 16 kByte fragment size
- Trapezoidal Readout Builder is the most promising configuration (240 MByte/sec), it might allow to reduce the number of RUs
- 30% gain in throughput with current 32 bit CPUs, expect further improvement in performances from the 64 bit and dual core CPUs
- Will try to carry out full scale RU Builder tests for one slice before finalizing the system

Summary

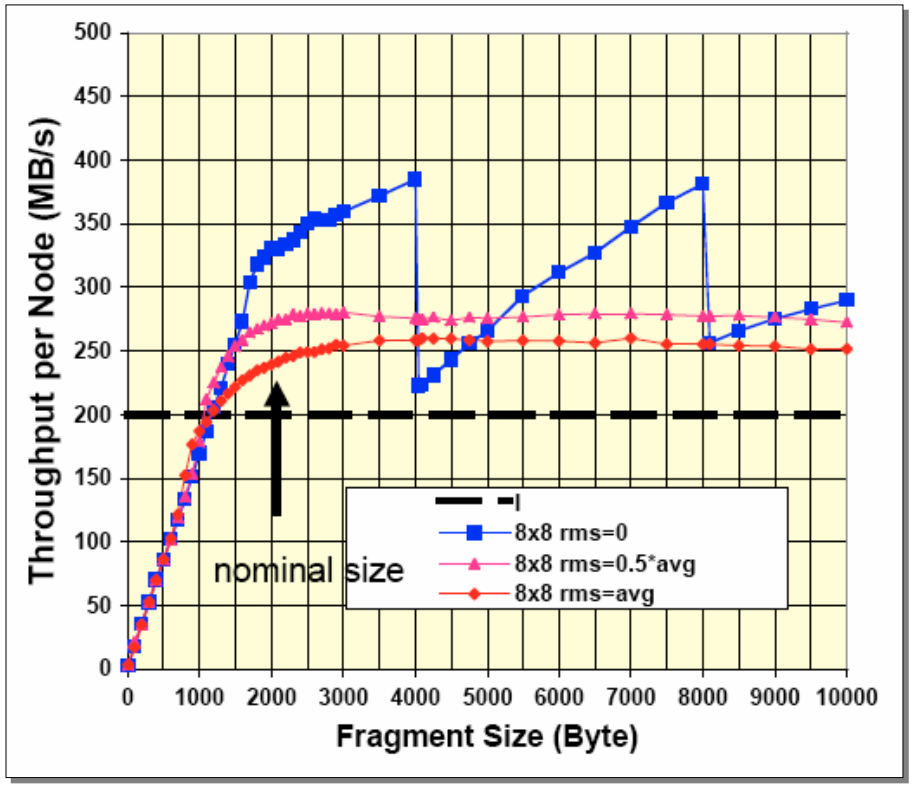
- CMS Event Builder:
  - Myrinet FED Builder at the required level of performance
  - GbE Trapezoidal RU Builder satisfies the requirements for all tested configurations
  - Myrinet RU Builder was already shown to be feasible

- Now finalizing the architecture of the Readout Builders:
  - Test the largest possible Event Builder system
  - Test new dual socket, dual core CPU based PCs
  - Order the RU Builder hardware equipment, install and commission the system to be ready for data taking in summer 2007

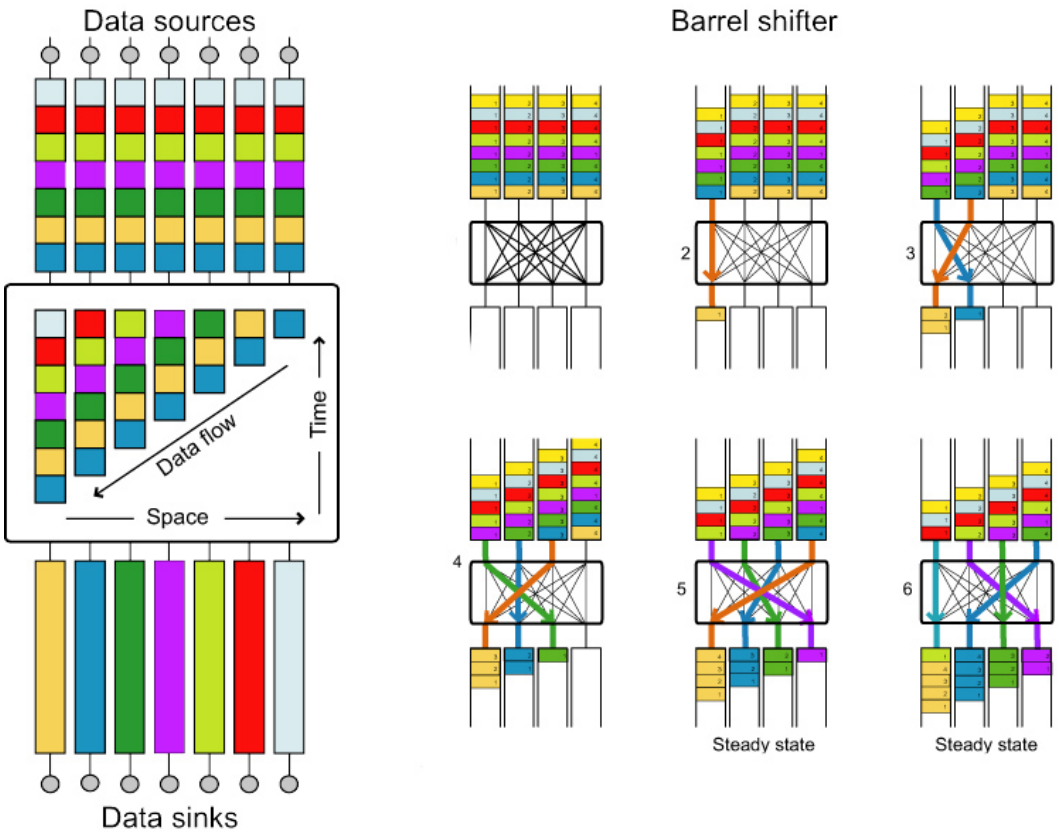
Backup Slides

BACKUP SLIDES

FED Builder Performances



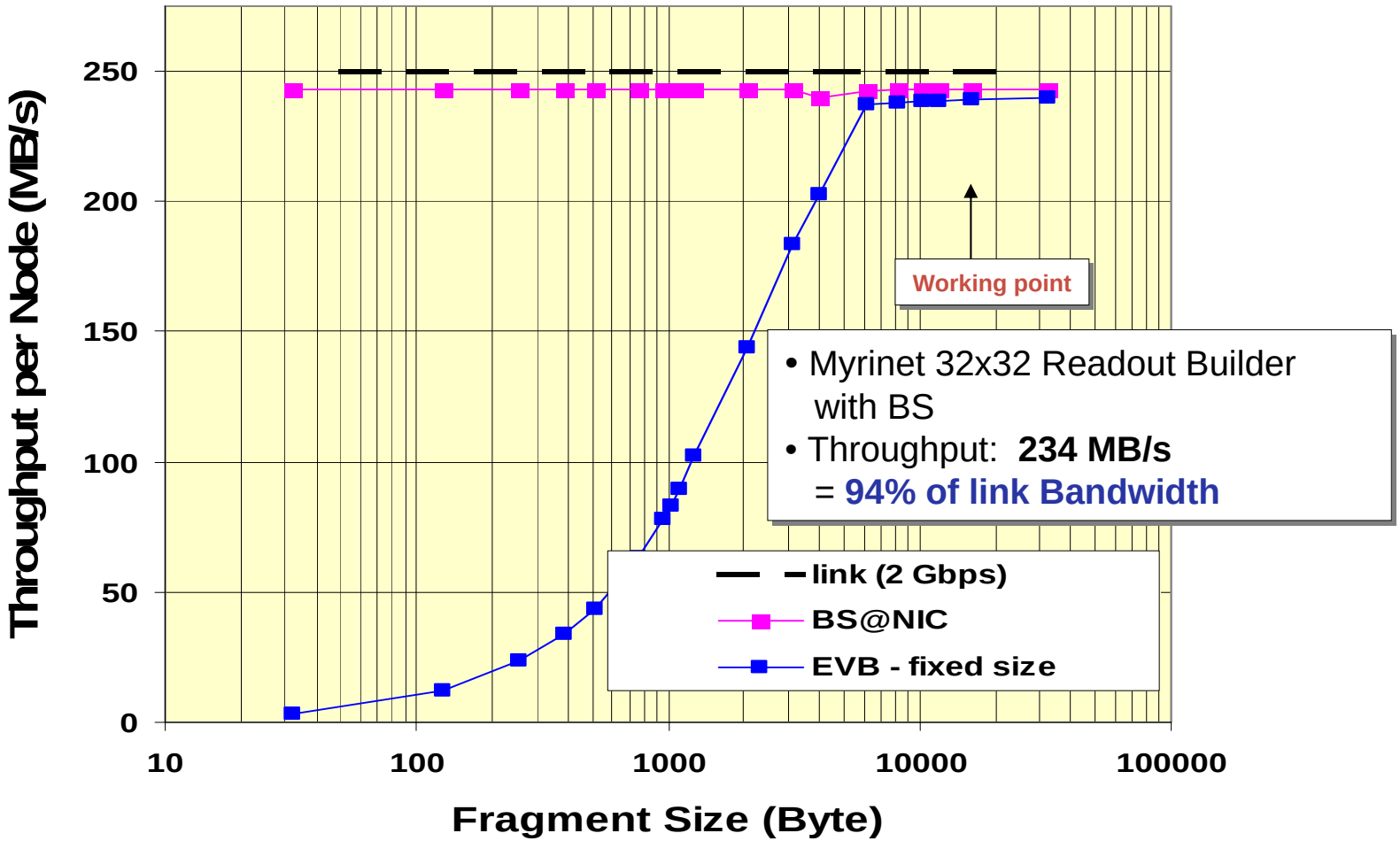
Myrinet Barrel-Shifter



- 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)
- Principle works for multi-stage switches

Myrinet Readout Builder Traffic Shaping

- Possible to use the links with an efficiency approaching 100% with fully synchronized barrel-shifter traffic shaping in the NIC driver



Pre-Series System