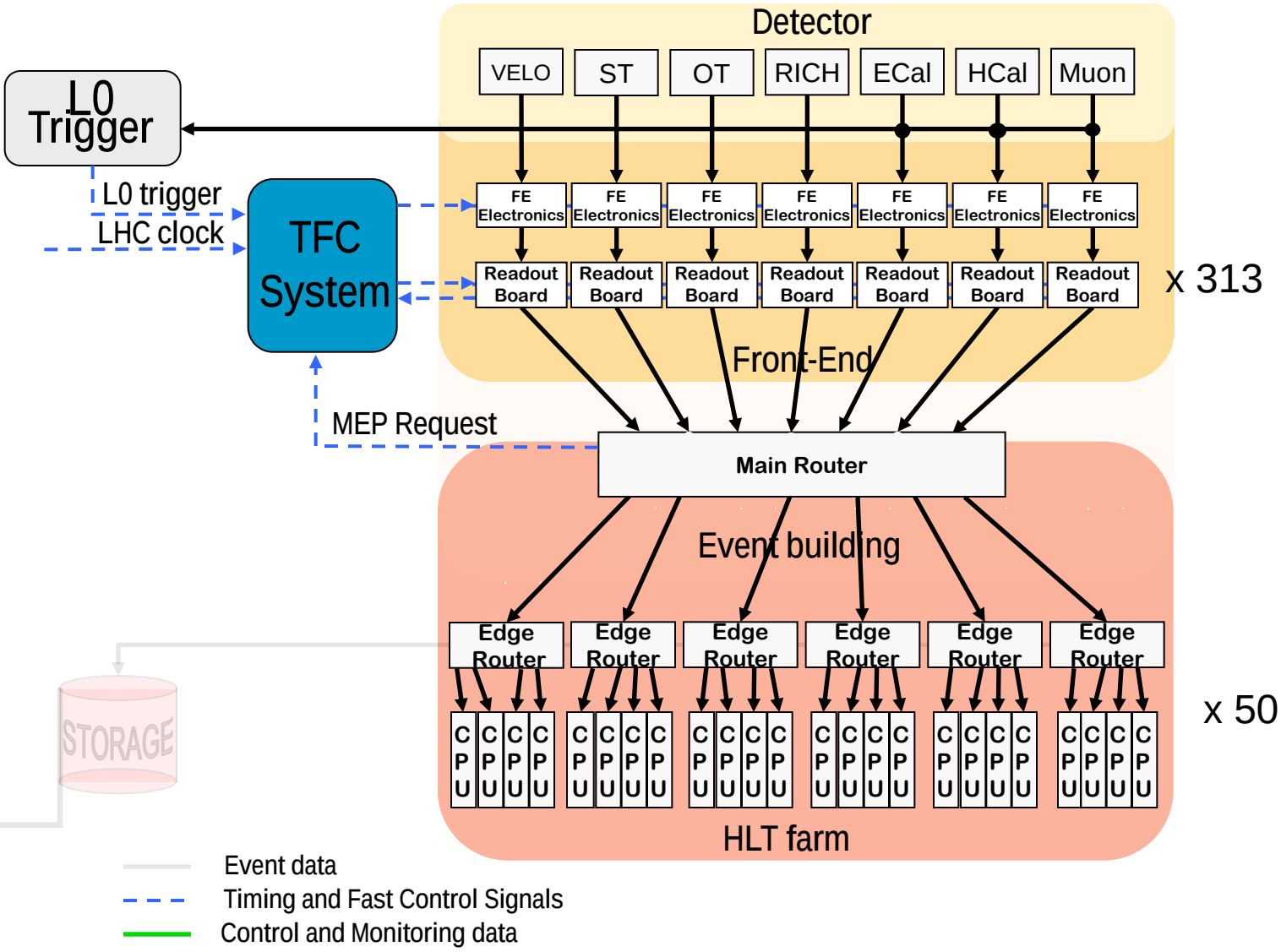
The LHCb Event-Builder

Markus Frank, Jean-Christophe Garnier, Clara Gaspar, Richard Jacobson, Beat Jost, Guoming Liu, Niko Neufeld, CERN/PH

17th Real-Time Conference
Instituto Superior Tecnico
Lisboa 2010

Architecture



Key Parameters

# links (UTP Cat 6)	~ 3000
Event-size (total – zero-suppressed)	35 kB
Read-out rate	1 MHz
# read-out boards	313
output bandwidth / read-out board	up to 4 Gigabit/s (4 Ethernet links)
# farm-nodes	550 (will grow to 1500)
input bandwidth / farm-node	1 Gigabit (1 dedicated Ethernet link)
# core-routers	1 (1260 ports)
# edge routers	50 (48 ports)
Event-rate to storage	2000 Hz (nominal)

Physical Installation

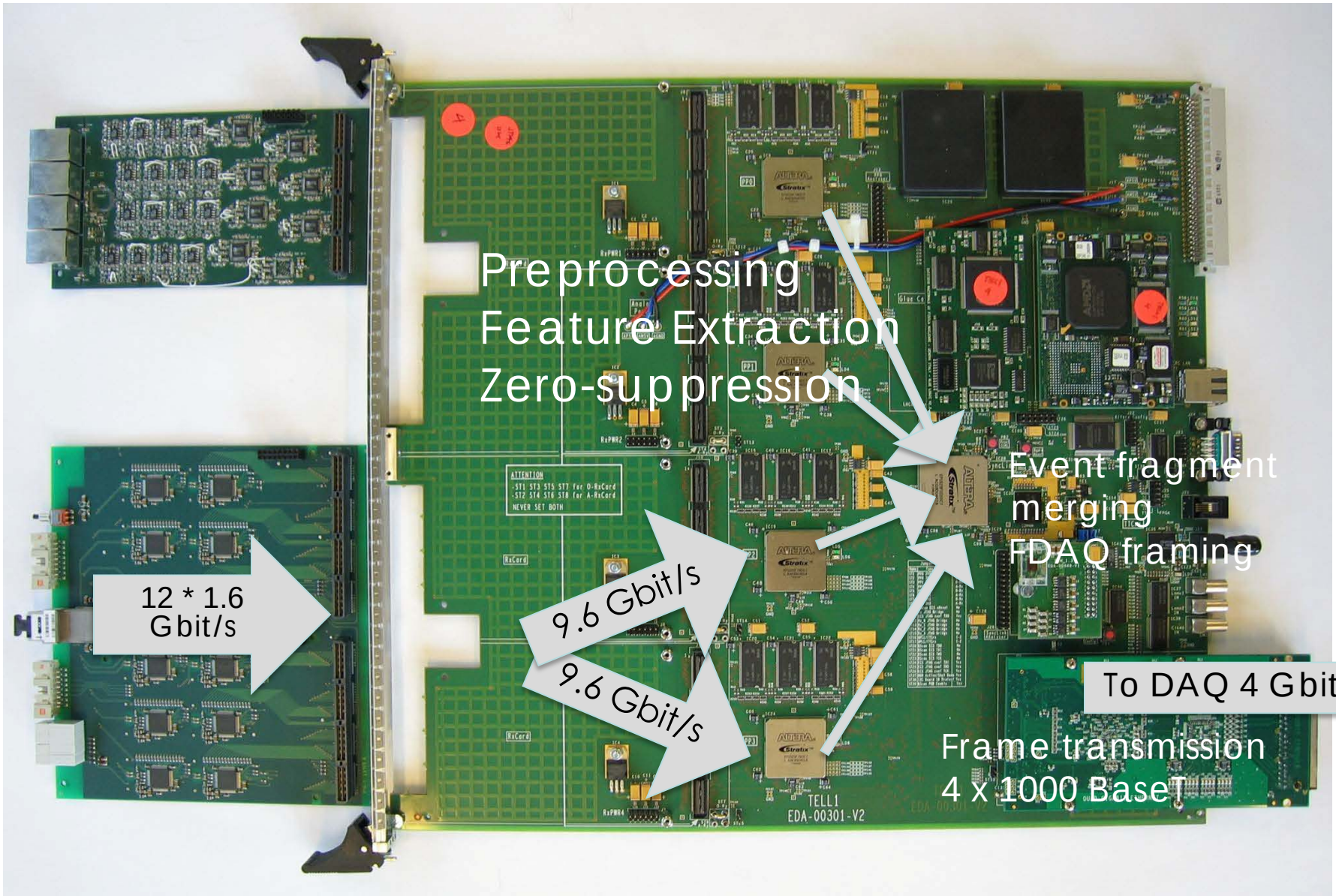
3 readout boards

2 DAQ network

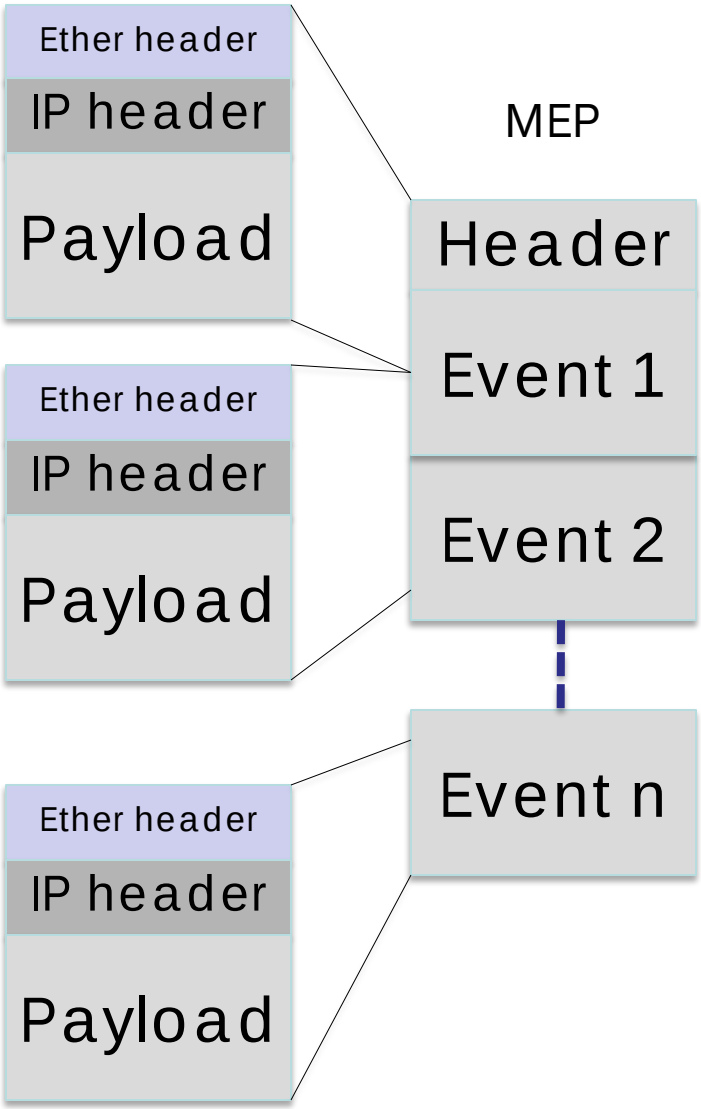
1 event-filter farm

- More than 4000 UTP Cat 6 patch connections
- 3 floors of ~ 80 m2
  - 3rd (top) read-out boards,
  - 2nd (middle): DAQ network,
  - 1st (bottom): event-filter farm
- Cat 6 patch-cords RJ45
- High-density RJ21 6-port connectors for main-router

The Read-out Board ("TELL1")

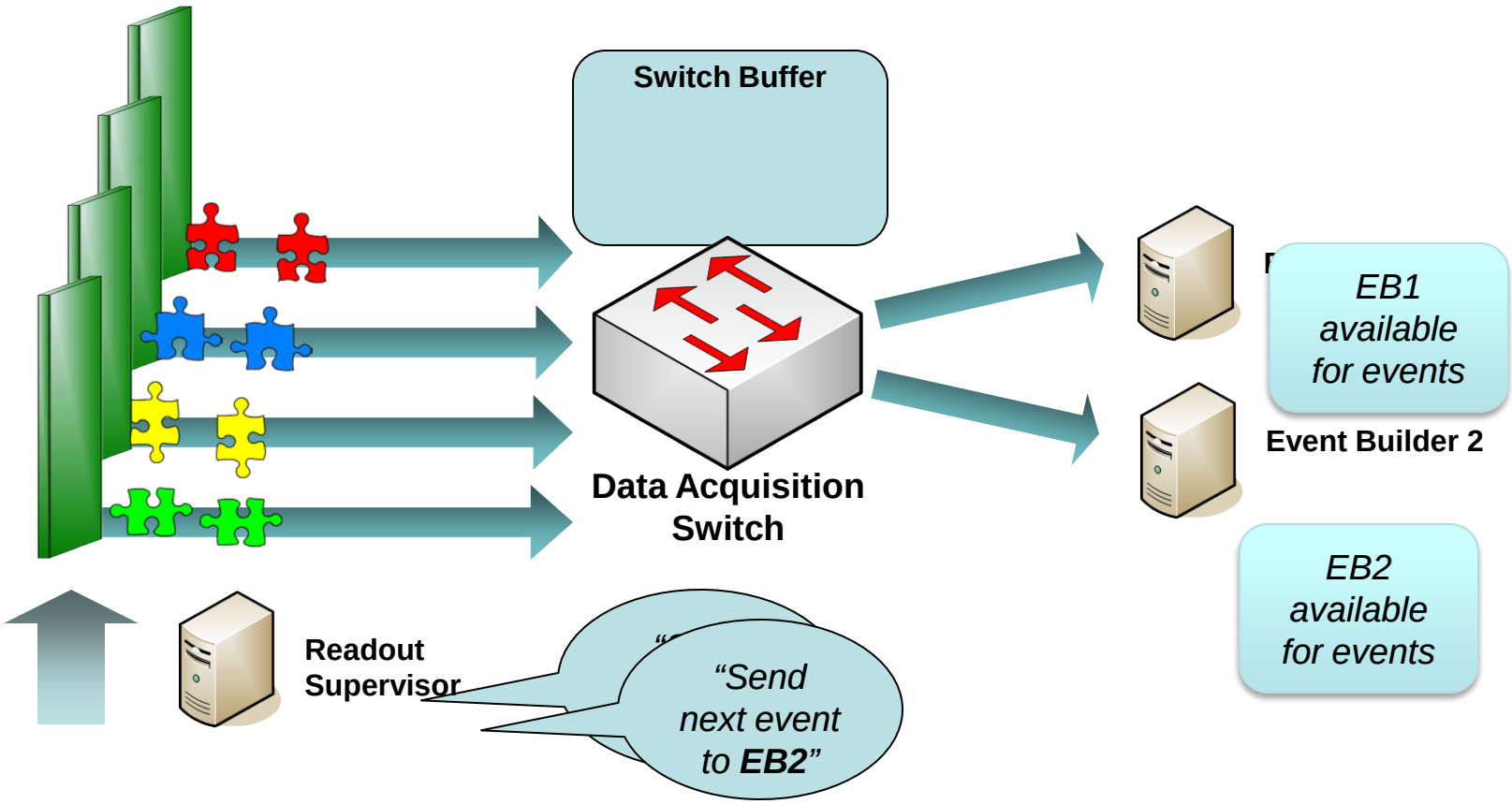


The Protocol



- Event fragments from up to 16 triggers packed into one Multi-Eventfragment Packet (MEP)
- A MEP must fit into one IPv4 packet (64 kB max!)
- IPv4 packets will be fragmented into Ethernet frames of MTU size
- MEP header is 12 bytes only (Event-number + length)
- IP address information is used in event-building
- Unreliable / no duplication / no re-transmission

Dataflow in LHCb



1 Event Builders inform Readout supervisor about availability

2 Readout Supervisor tells readout boards where events must be sent (round-robin)

3 Readout boards do not buffer, so switch must

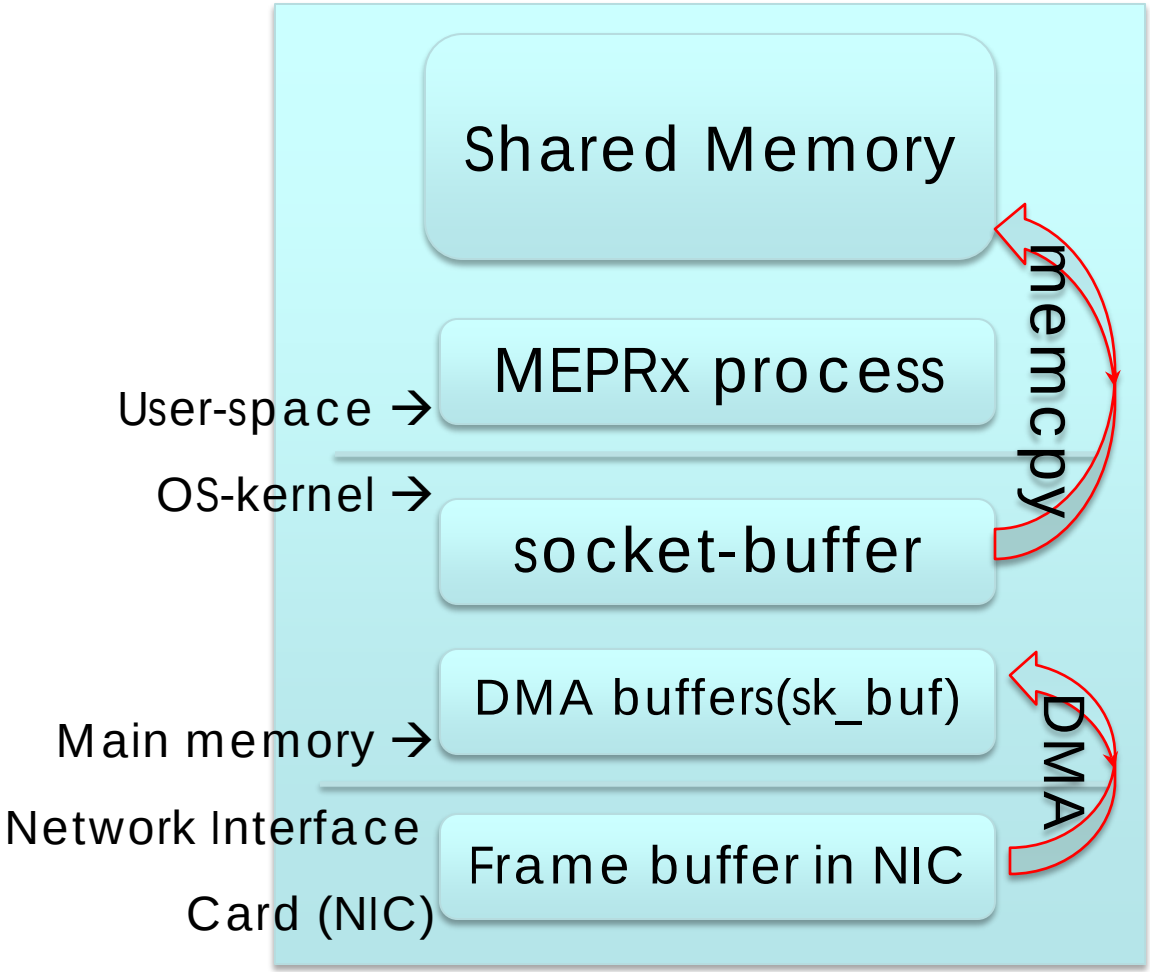
The many sources of packet loss

- Packet corruption (normally < 10^{-12}) – this usually is a badly handled overflow in the FPGAs (very rare now)
  - Depending on the corruption these events may be dropped already in the network
- Packet loss due to buffer overflow (ingress, egress) in core-router, edge-switches or farm-node (between 10^{-9} and 10^{-11})
  - This packet loss often occurs in bursts
- Packet loss in core-router (silent!) < 10^{-13}
  - →could only be fixed in session with company engineers

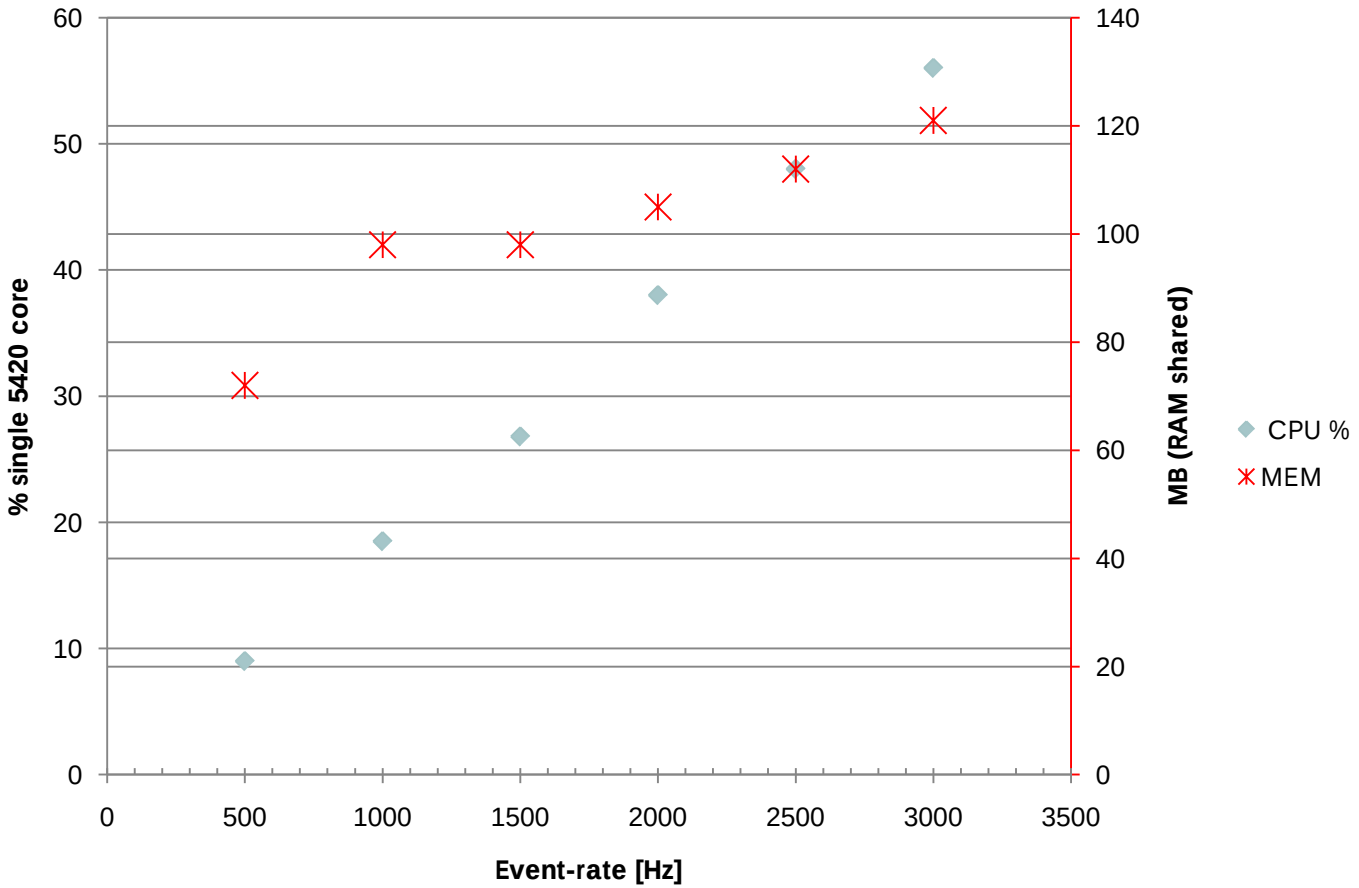
Diagnostic: or "What the shifter saw"

Event assembly / buffering in farm -node

- IP packets re-assembled by OS
- Event-building process (MEPRx) puts data in shared buffer
- Monitoring in kernel difficult: need good (and generous) tuning!



MEPRx resource usage

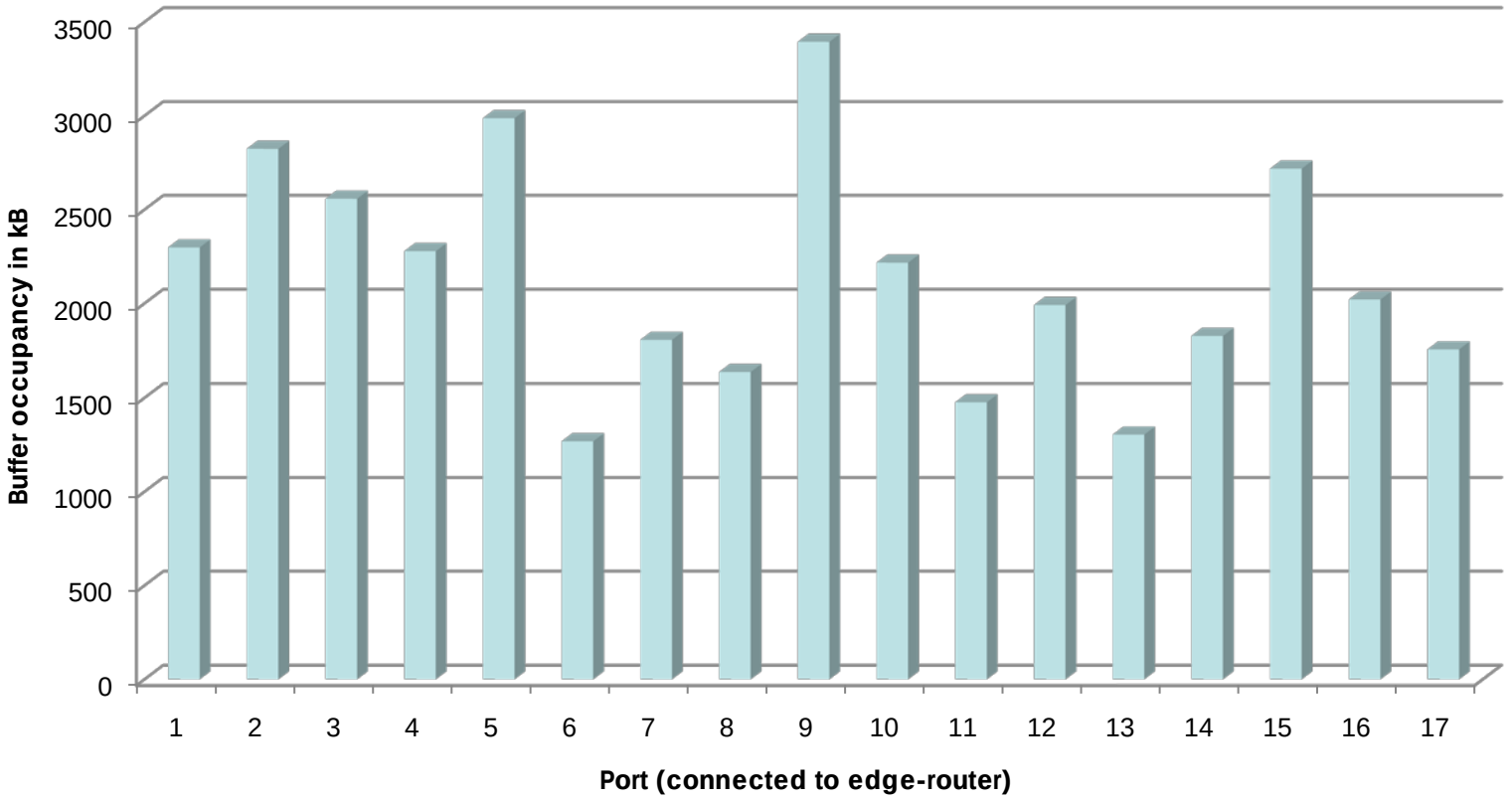


% CPU of one Intel 5420 core (a 4 core processor running at 2.5 GHz
MEM is resident memory (i.e. pages locked in RAM)
Precision of measurements is about 10% for CPU and 1% for RAM

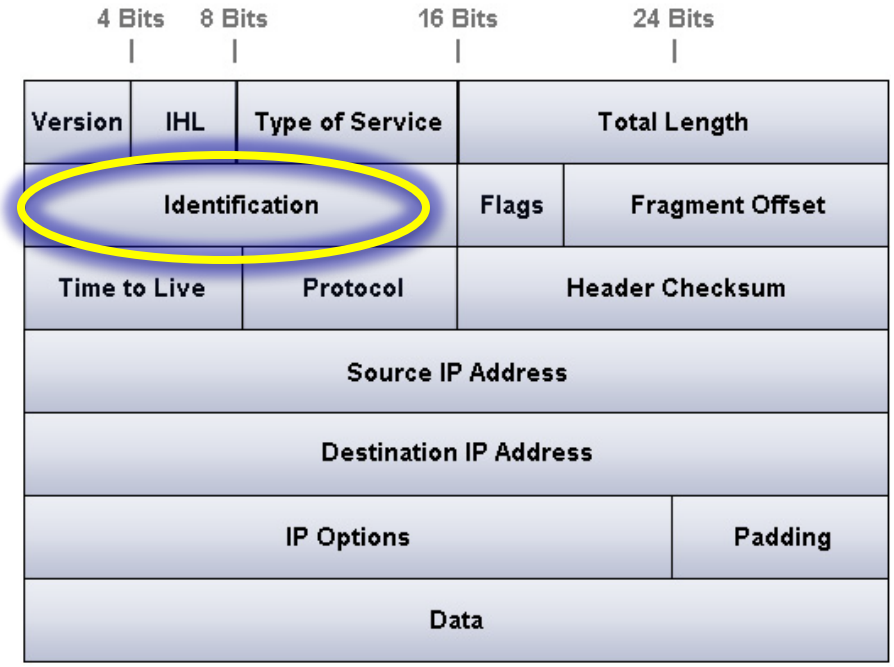
Buffer Usage

Buffer usage in core-router with a test using 270 sources @ 350 kHz event-rate

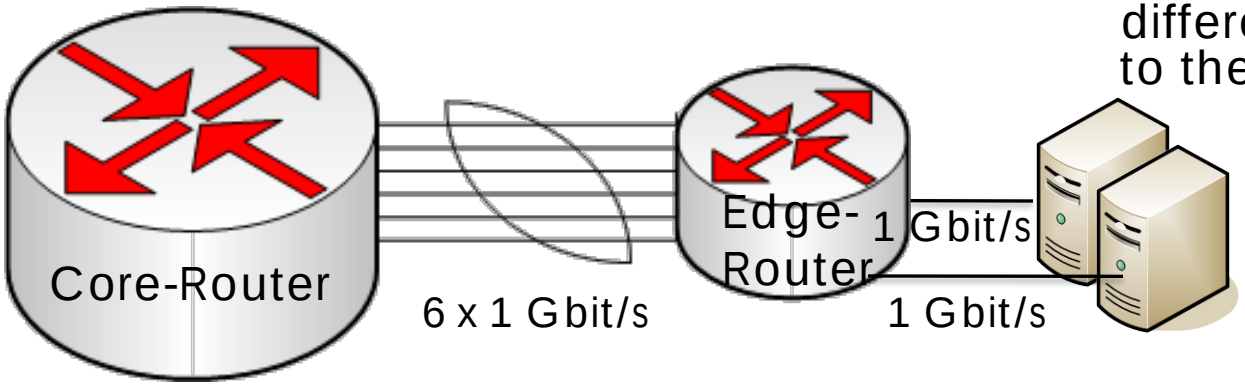
- 256 MB shared between 48 ports
- 17 ports used as "output"
- Non-uniformity under investigation



Link Aggregation



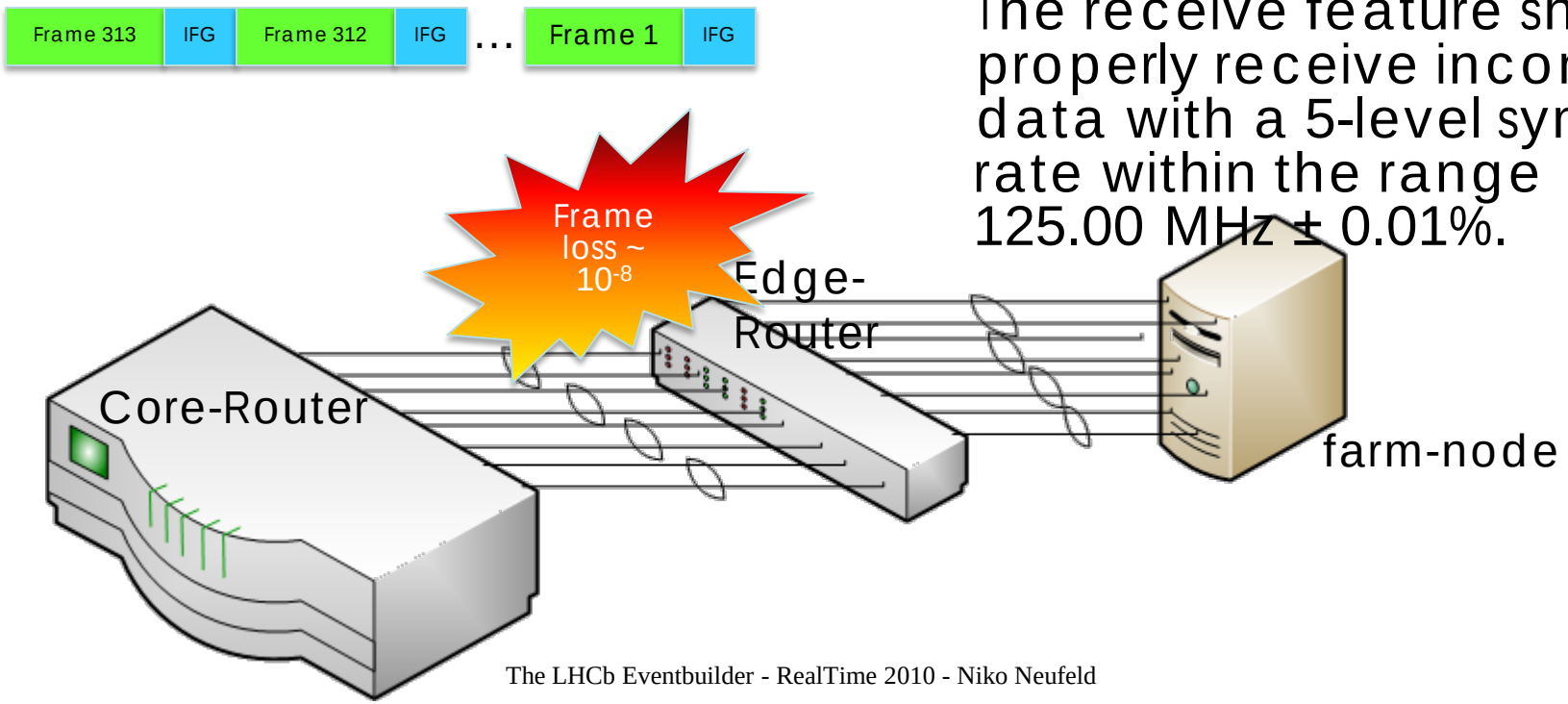
- IEEE 802.3ad (now actually IEEE 802.1AX)
- Does not define how traffic is balanced over multiple links
- Available balancing algorithms are made to preserve frame-order
- Typically use a hash out of destination and source Ethernet and/or IP address
  - → all links will be used when different destinations sent to the same host



Clock Tolerance in IEEE 802.3ab

802.3ab transmits on 5 levels over
4 twisted pairs @ 125 MHz
Inteframe gap (IFG) is 96 bits

- **40.6.1.2.6 Transmit clock frequency**
  The quinary symbol transmission rate on each pair of the master PHY shall be 125.00 MHz ± 0.01%.
- **40.6.1.3.2 Receiver frequency tolerance**
  The receive feature shall properly receive incoming data with a 5-level symbol rate within the range 125.00 MHz ± 0.01%.



Not covered

- Run-control (see talk by Clara Gaspar)
- Monitoring
- Process Management in farm-nodes (see poster of Juan Caicedo)
- Event aggregation & storage (see poster of Jean-Christophe Garnier)

Conclusions

- A (almost) pure push protocol for a 40 Gigabyte/s DAQ works
  - With very good hardware
  - And a lot of hard work
- More than 3000 UTP links are no problem and you do not need TCP
- Commercial will most of time not work off the shelf
- We will now go on to see how it works at 40 MHz (1.2 Terabyte/s)