

The LHCb Event-Builder

Markus Frank, Jean-Christophe Garnier, Clara
Gaspar, Richard Jacobson, Beat Jost, Guoming Liu,

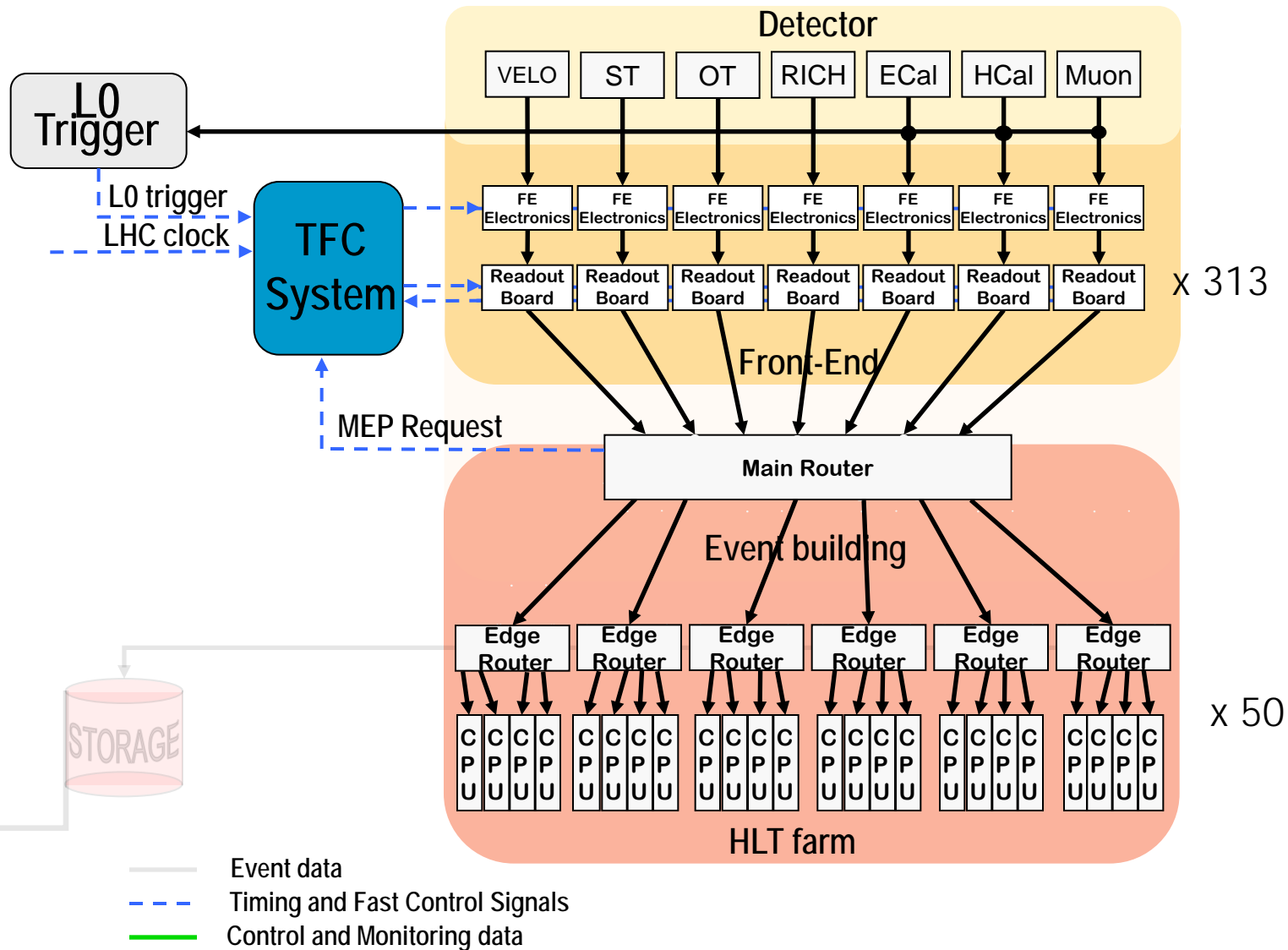
Niko Neufeld, CERN/PH

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

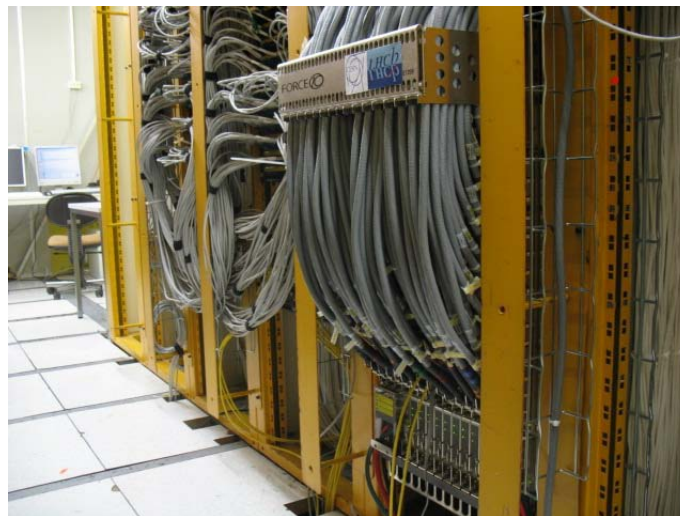


3 readout boards

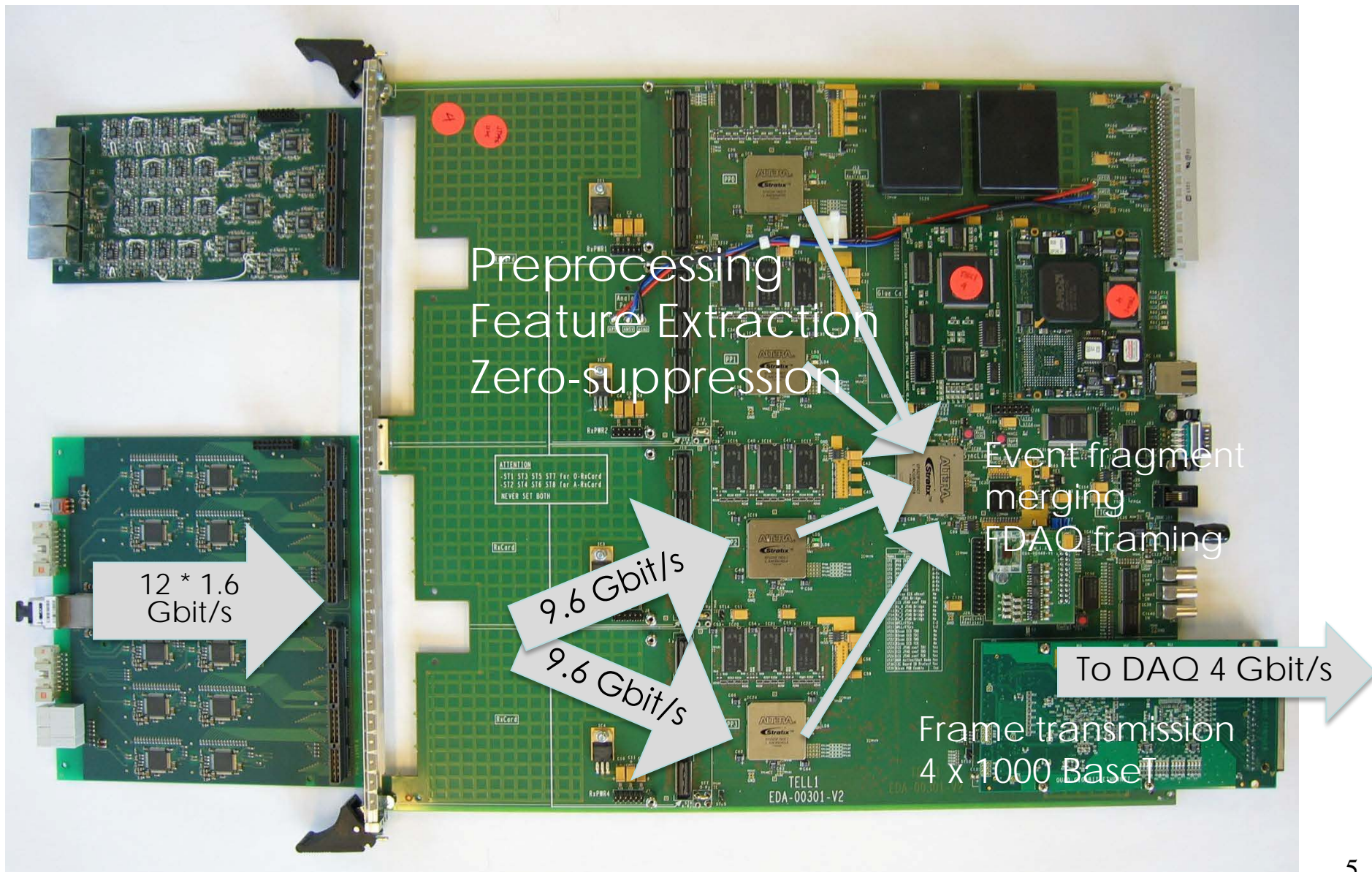
2 DAQ network

1 event-filter farm

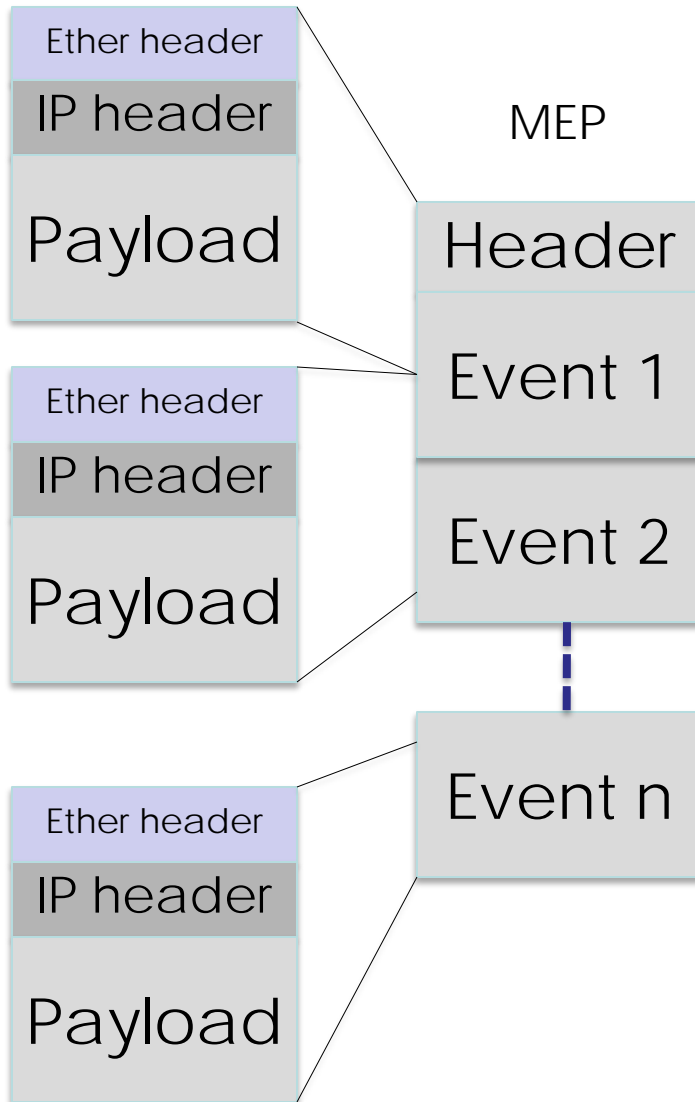
- More than 4000 UTP Cat 6 patch connections
- 3 floors of ~ 80 m²
 - 3rd (top) read-out boards,
 - 2nd (middle): DAQ network,
 - 1st (bottom): event-filter farm
- Cat 6 patch-cords RJ45
- High-density RJ21 6-port connectors for main-router



The Read-out Board ("TELL1")

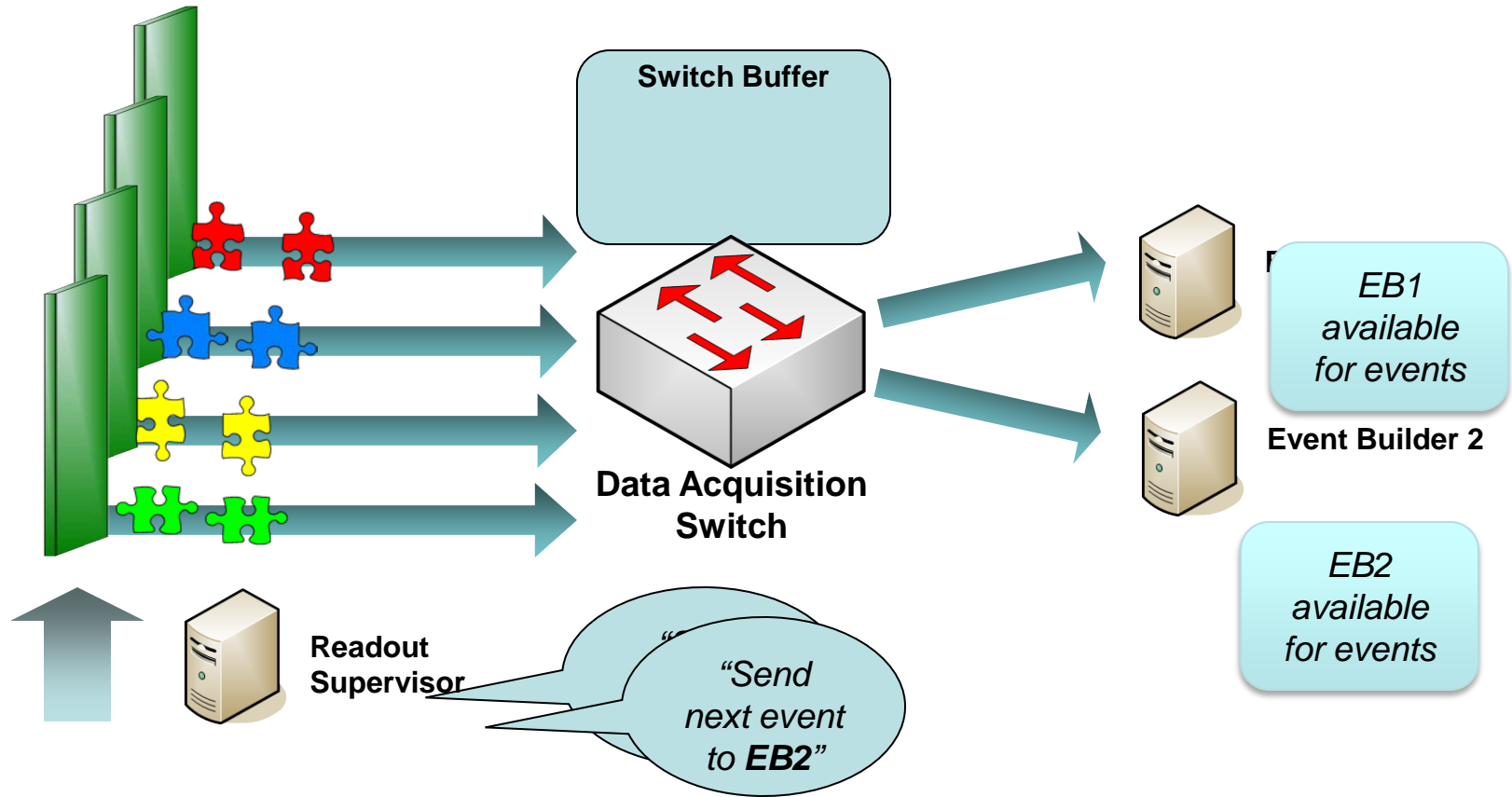


The Protocol



- 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"

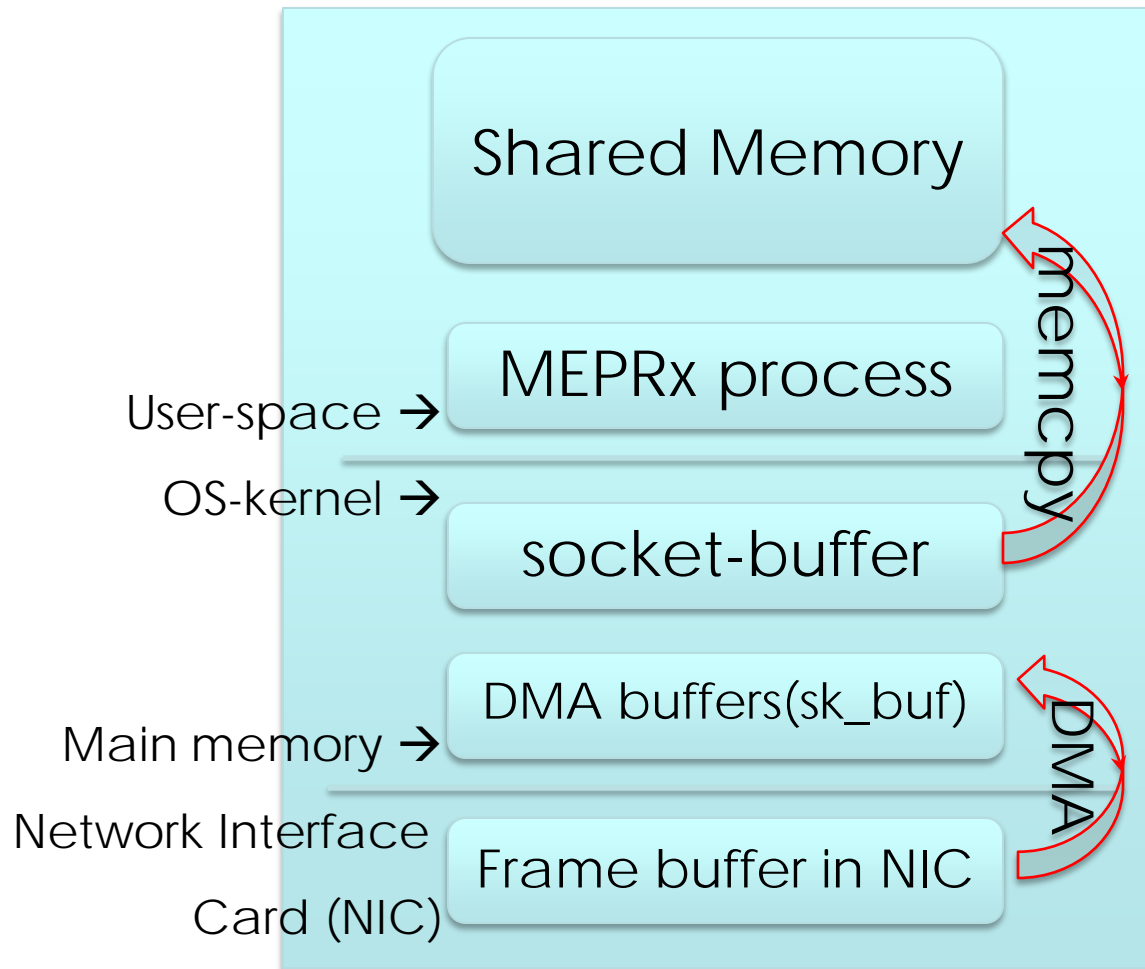
```

LHCb: Message display
May19-164016[WARN] wona0804: Gaudi.exe(LHCb_MONA0804_TTNZShon_00): ToolSvc.ITNoiseCalculationTool: ST::STNoiseCalculationTool: The WARNING message is suppressed : 'Data is empty'
May19-164016[WARN] wona0804: Gaudi.exe(LHCb_MONA0804_TTNZShon_00): ToolSvc.ITChSNoiseCalculationTool: ST::STChSNoiseCalculationTool: The WARNING message is suppressed : 'Data is empty'
May19-164016[WARN] wona0802: Gaudi.exe(LHCb_MONA0802_TTNZShon_00): ToolSvc.ITNoiseCalculationTool: ST::STNoiseCalculationTool: Data is empty
May19-164016[WARN] wona0802: Gaudi.exe(LHCb_MONA0802_TTNZShon_00): ToolSvc.ITChSNoiseCalculationTool: ST::STChSNoiseCalculationTool: Data is empty
May19-164016[WARN] wona0802: Gaudi.exe(LHCb_MONA0802_TTNZShon_00): ToolSvc.ITNoiseCalculationTool: ST::STNoiseCalculationTool: Data is empty
May19-164016[WARN] wona0802: Gaudi.exe(LHCb_MONA0802_TTNZShon_00): ToolSvc.ITChSNoiseCalculationTool: ST::STChSNoiseCalculationTool: Data is empty
May19-164016[WARN] wona0802: Gaudi.exe(LHCb_MONA0802_TTNZShon_00): ToolSvc.ITNoiseCalculationTool: ST::STNoiseCalculationTool: Data is empty
May19-164016[WARN] wona0802: Gaudi.exe(LHCb_MONA0802_TTNZShon_00): ToolSvc.ITChSNoiseCalculationTool: ST::STChSNoiseCalculationTool: Data is empty
May19-164016[WARN] wona0802: Gaudi.exe(LHCb_MONA0802_TTNZShon_00): ToolSvc.ITNoiseCalculationTool: ST::STNoiseCalculationTool: Data is empty
May19-164016[WARN] wona0802: Gaudi.exe(LHCb_MONA0802_TTNZShon_00): ToolSvc.ITChSNoiseCalculationTool: ST::STChSNoiseCalculationTool: Data is empty
May19-164016[WARN] wona0802: Gaudi.exe(LHCb_MONA0802_TTNZShon_00): ToolSvc.ITNoiseCalculationTool: ST::STNoiseCalculationTool: The WARNING message is suppressed : 'Data is empty'
May19-164016[WARN] wona0802: Gaudi.exe(LHCb_MONA0802_TTNZShon_00): ToolSvc.ITChSNoiseCalculationTool: ST::STChSNoiseCalculationTool: The WARNING message is suppressed : 'Data is empty'
May19-164017[WARN] wona0802: Gaudi.exe(LHCb_MONA0802_LODUDAGNon_00): LODURptMonitor: LODURptMonitor:: New configuration tck found : reset all histos
May19-164025[ERROR]lcald0701: Gaudi.exe(CALD0701_MPRx_1): MPRx:002: Run # 72046 - Incomplete Event #324143 No packet from: vetella01 vetella02 vetella03 vetella04 vetella05 vetella06 vetella07 vetella08 vetell
a09 vetella10 vetella11 vetella12 vetella17 vetella18 vetella19 vetella20 vetella21 vetella22 vetella23 vetella24 vetella25 vetella26 vetella27 vetella28 vetella29 vetella30 vetella31 vetella32 vetella33 vetell
a34 vetella35 vetella36 vetella37 vetella38 vetella39 vetella40 vetella41 vetella42 vetellc01 vetellc02 vetellc03 vetellc04 vetellc05 vetellc06 vetellc07 vetellc08 vetellc09 vetellc10 vetellc11 vetellc12 vetell
c13 vetellc14 vetellc15 vetellc16 vetellc17 vetellc19 vetellc20 vetellc21 vetellc22 vetellc23 vetellc24 vetellc25 vetellc26 vetellc27 vetellc28 vetellc29 vetellc30 vetellc31 vetellc32 vetellc33 vetell
c34 vetellc35 vetellc36 vetellc37 vetellc39 vetellc39 vetellc40 vetellc41 vetellc42 ittell08 ittell09 ittell10 ittell11 ittell12 ittell13 ittell14 ittell101 ittell102 ittell103 ittell104 ittell105 ittell106 ittell07
ittell26 ittell27 ittell28 pstell101 pstell102 pstell103 pstell104 pstell105 pstell106 pstell107 pstell108 ectell110 ectell111 ectell112 ectell113 hctell101 hctell102 hctell103 hctell104 hctell105 hctell106 hctell107 hct
ell108 mutella07 mutella06 tmutellq01 tmutellq02 tputell101 tputell102 tputell103 tputell104 tputell105 tfoodin04-d2
May19-164025[ERROR]lcald0701: Gaudi.exe(CALD0701_MPRx_1): MPRx:001: Run # 72046 - Incomplete Event #324155 Only packets from: ittell124 ittell117 ittell118 ittell119 ittell120 ittell121 ittell138 ittell139 ittell140 i
ttell141 ittell142 ittell129 ittell130 ittell132 ittell133 ittell134 ittell135 ottella01 ottella02 ottella03 ottella04 pstell102 pstell107 ectell109 ectell110 ectell111 ectell112 ectell113 ectell101 hctell102 hctell104 hctell105 t
ctell106 hctell107 hctell108 tmutellq01 tmutellq02 tmutellq12 tmutellq34
May19-164025[ERROR]lcald0701: Gaudi.exe(CALD0701_MPRx_1): MPRx:000: Run # 72046 - Incomplete Event #499669 No packet from: rluk1102 rluk1104 ectell101 ectell102 ectell103 ectell104 ectell105 ectell106 ectell107 ectel
l08 ectell114 ectell115 ectell116 ectell117 ectell118 ectell119 ectell120 ectell121 ectell122 ectell123 ectell124 ectell125 ectell126 mutella04 mutella03 mutella02 mutella01 mutellc07 mutellc06 mutellc05 mutellc04 mutellc02
mutellc02 mutellc01 tcatell101 tcatell102 tmutellq03 tmutellq04
May19-164026[ERROR]lcald0701: Gaudi.exe(CALD0701_MPRx_1): MPRx:002: Run # 72046 - Incomplete Event #499679 No packet from: ittell130 ittell131 ittell133 ittell134 ottella01 ottella13 ottella14 ottella03 ottella15
ottella04 ottella16 ottella05 ottella17 ottella16 ottella18 ottella19 ottella08 ottella20 ottella09 ottella21 ottella10 ottella22 ottella11 ottella23 ottella12 ottella24 ottellc01 ottellc13 ottellc02
ottellc14 ottellc03 ottellc15 ottellc16 ottellc16 ottellc05 ottellc17 ottellc06 ottellc18 ottellc07 ottellc08 ottellc09 ottellc08 ottellc02 ottellc21 ottellc10 ottellc22 ottellc11 ottellc23 ottellc12 ottellc24
rluk1101 rluk1102 rluk1103 rluk1104 rluk1105 rluk1106 rluk1107 rluk1101 rluk1102 rluk1103 rluk1104 rluk1105 rluk1106 rluk1107 rluk1108 pstell101 pstell102 pstell103 pstell105 pstell106 pstell107 pstell108 ectell101 ect
ell02 ectell103 ectell104 ectell105 ectell106 ectell107 ectell108 ectell109 ectell110 ectell114 ectell115 ectell116 ectell117 ectell118 ectell119 ectell120 ectell121 ectell122 ectell123 ectell124 ectell125 ectell126 hctell101 hctell
03 hctell105 hctell108 mutella05 mutella04 mutella03 mutella02 mutella01 mutellc07 mutellc06 mutellc05 mutellc04 mutellc03 mutellc02 mutellc01 tcatell101 tcatell102 tmutellq01 tmutellq02 tmutellq03 tmutellq04
May19-164053[ERROR]hltc1009: Gaudi.exe(HLTC1009_GaUCHoJob_7): createVeloLiteClusters: DecodeVeloRawBuffer:: Deleted all lite VEO clusters as more than limit 10000 in the event
May19-164056[ERROR]hltc0110: Gaudi.exe(HLTC0110_MPRx_1): MPRx:002: Run # 72046 - Incomplete Event #1973137 No packet from: mutella01
May19-164353[WARN] hltc0703: Gaudi.exe(HLTC0703_GaUCHoJob_7): HltSelReportsWriter: HltSelReportsWriter:: HltSelReports is huge 121.376 kBytes. Saved in 2 separate RawBanks
May19-164609[ERROR]hlcb0907: Gaudi.exe(HLTC0907_GaUCHoJob_4): createVeloLiteClusters: DecodeVeloRawBuffer:: Deleted all lite VEO clusters as more than limit 10000 in the event
May19-164647[ERROR]hlta0804: Gaudi.exe(HLTA0804_MPRx_1): MPRx:002: Run # 72046 - Incomplete Event #18922012 No packet from: mutella01
May19-164658[ERROR]hlta0809: Gaudi.exe(HLTA0809_GaUCHoJob_8): createVeloLiteClusters: DecodeVeloRawBuffer:: Deleted all lite VEO clusters as more than limit 10000 in the event
May19-164821[ERROR]hltc1001: Gaudi.exe(HLTC1001_MPRx_1): MPRx:002: Run # 72046 - Incomplete Event #23430435 No packet from: mutella01
May19-165019[ERROR]hltd0101: Gaudi.exe(HLTD0101_GaUCHoJob_6): createVeloLiteClusters: DecodeVeloRawBuffer:: Deleted all lite VEO clusters as more than limit 10000 in the event
May19-165034[ERROR]hltc0710: Gaudi.exe(HLTC0710_GaUCHoJob_8): createVeloLiteClusters: DecodeVeloRawBuffer:: Deleted all lite VEO clusters as more than limit 10000 in the event
May19-165329[ERROR]hltc0911: Gaudi.exe(HLTC0911_MPRx_1): MPRx:002: Run # 72046 - Incomplete Event #38301687 No packet from: mutella01

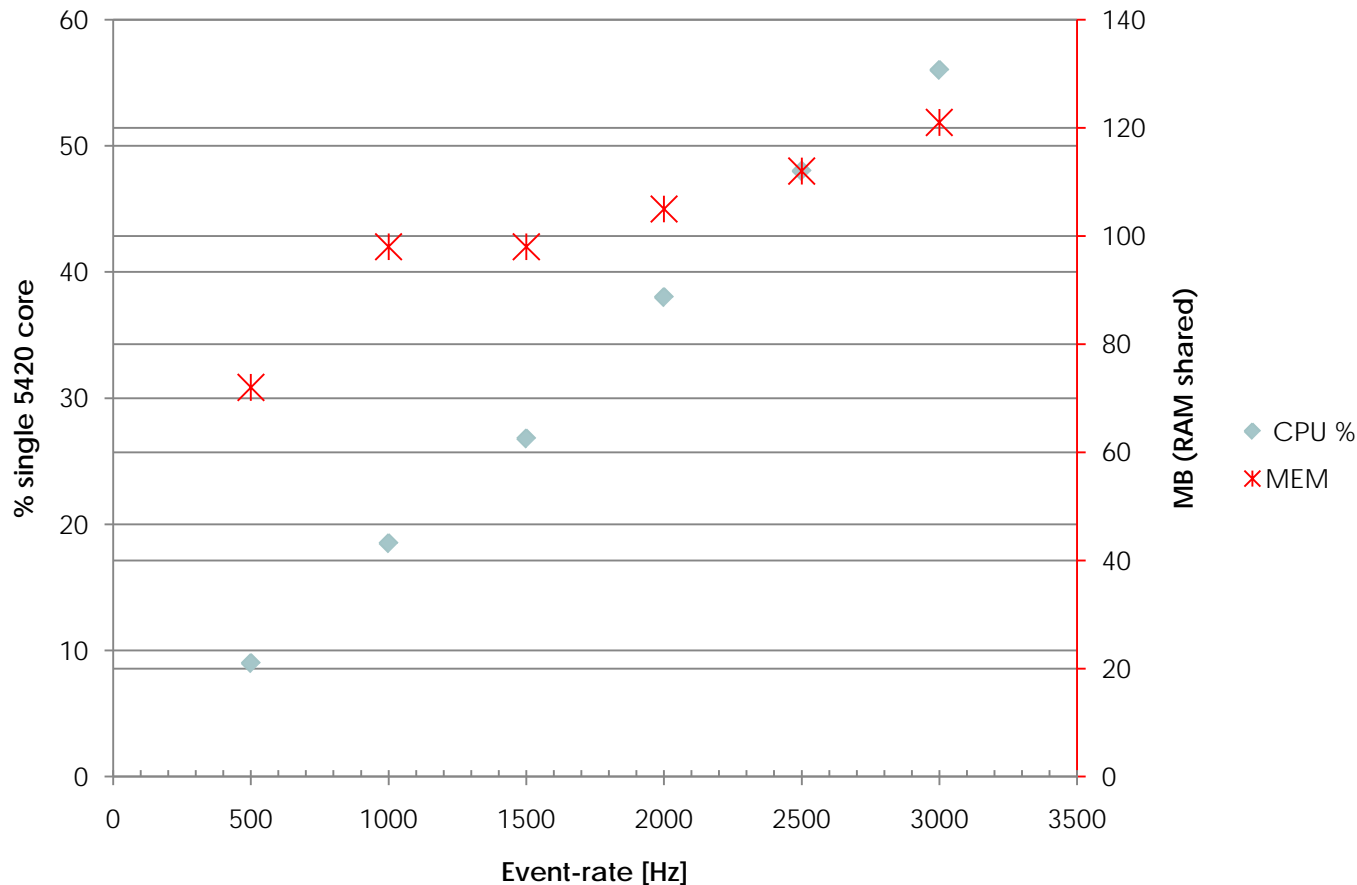
```

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

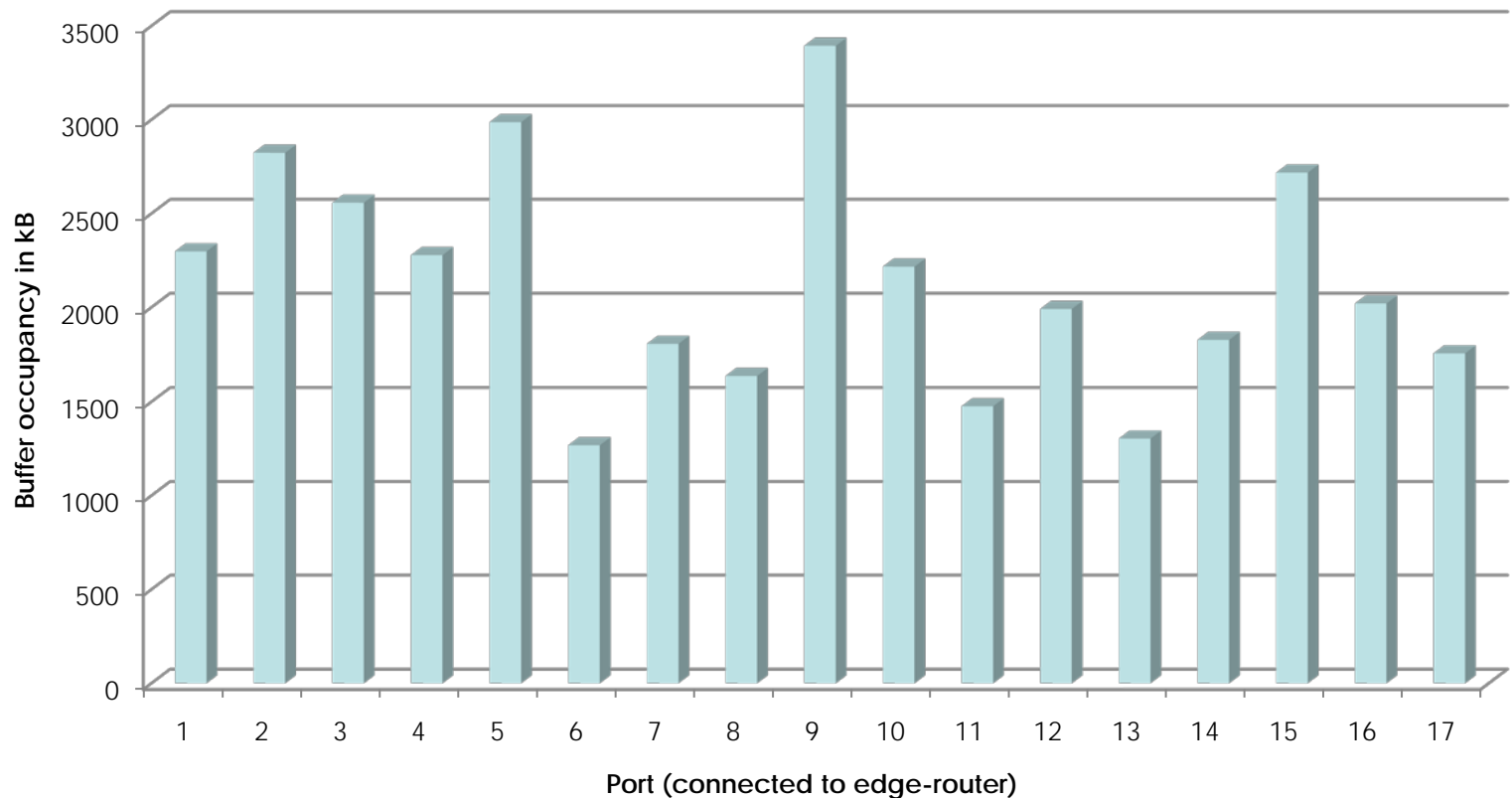


% CPU of one Intel 5420 core (a 4 core processor running at 2.5 GHz
 MEM is resident memory (i.e. pages locked in RAM)
 Precision of measurements is about 10% for CPU and 1% for RAM

Buffer Usage

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

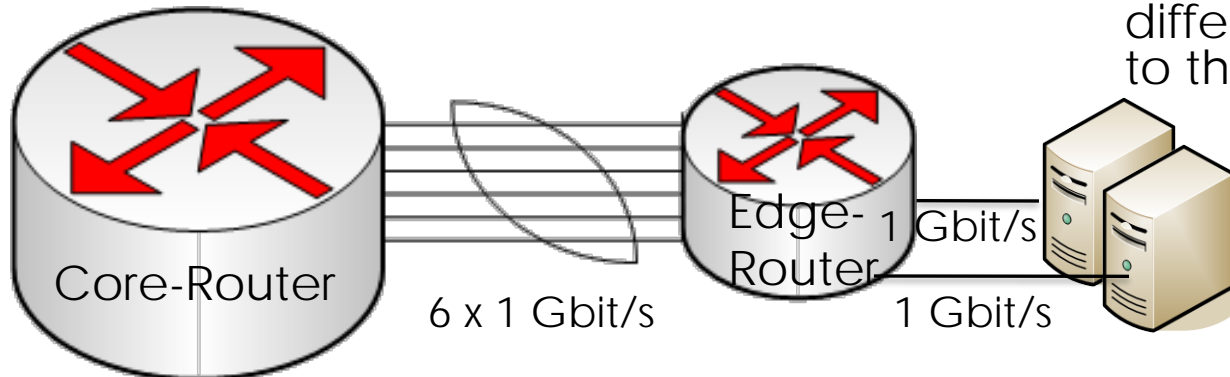
- 256 MB shared between 48 ports
- 17 ports used as "output"
- Non-uniformity under investigation



Link Aggregation

4 Bits	8 Bits	16 Bits	24 Bits
Version	IHL	Type of Service	Total Length
Identification		Flags	Fragment Offset
Time to Live	Protocol	Header Checksum	
Source IP Address			
Destination IP Address			
IP Options			Padding
Data			

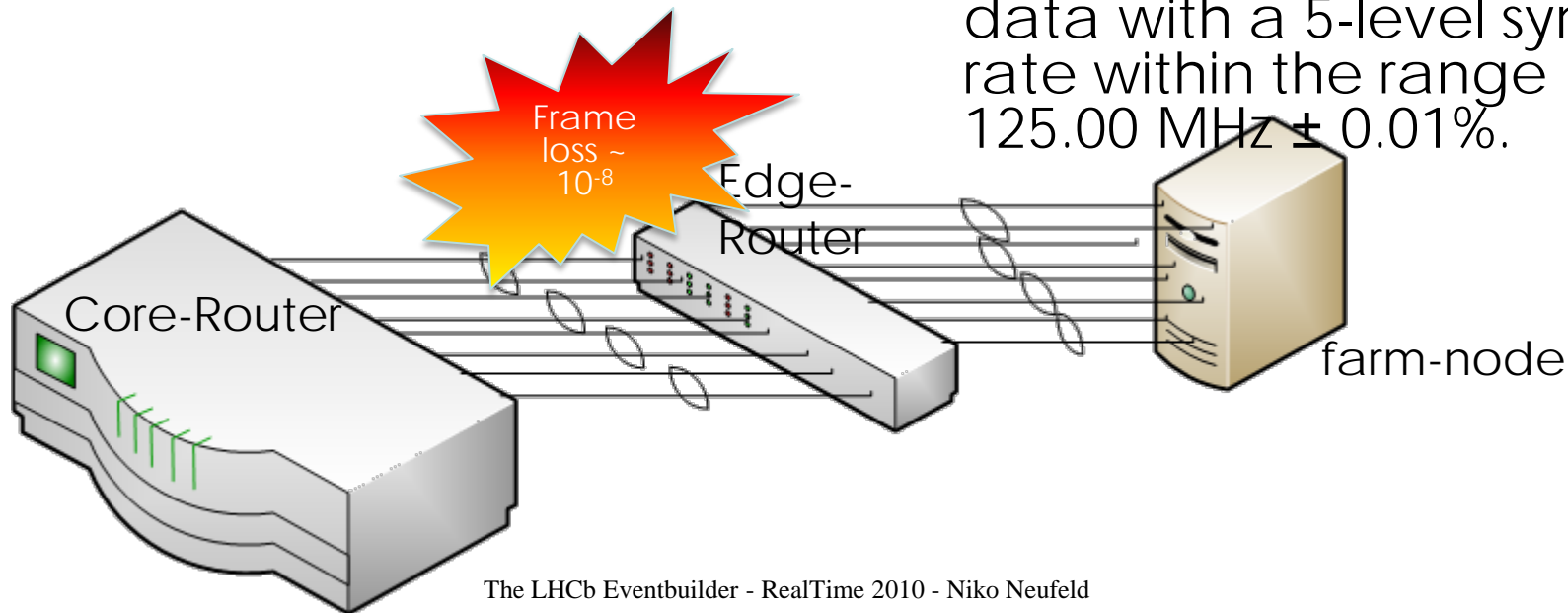
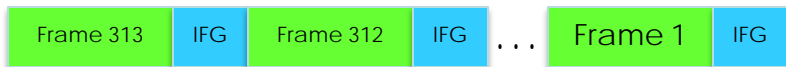
- IEEE 802.3ad (now actually IEEE 802.1AX)
- Does **not** define how traffic is balanced over multiple links
- Available **balancing algorithms** are made to **preserve frame-order**
- Typically use a hash out of destination and source Ethernet and/or IP address
 - → **all** links will be used when different 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 $125.00 \text{ MHz} \pm 0.01\%$.
- **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 $125.00 \text{ MHz} \pm 0.01\%$.



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)