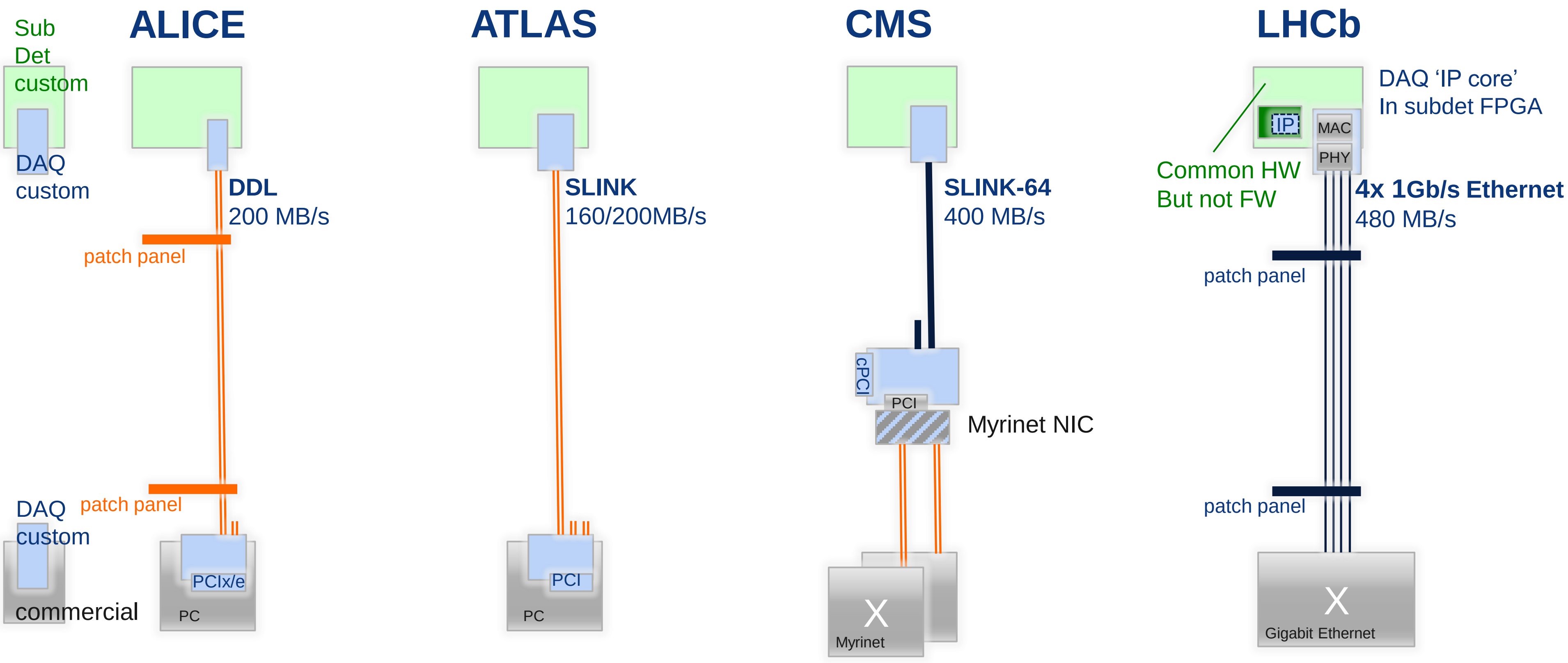
CERN

DAQ Readout Links at the LHC Commissioning & Robustness

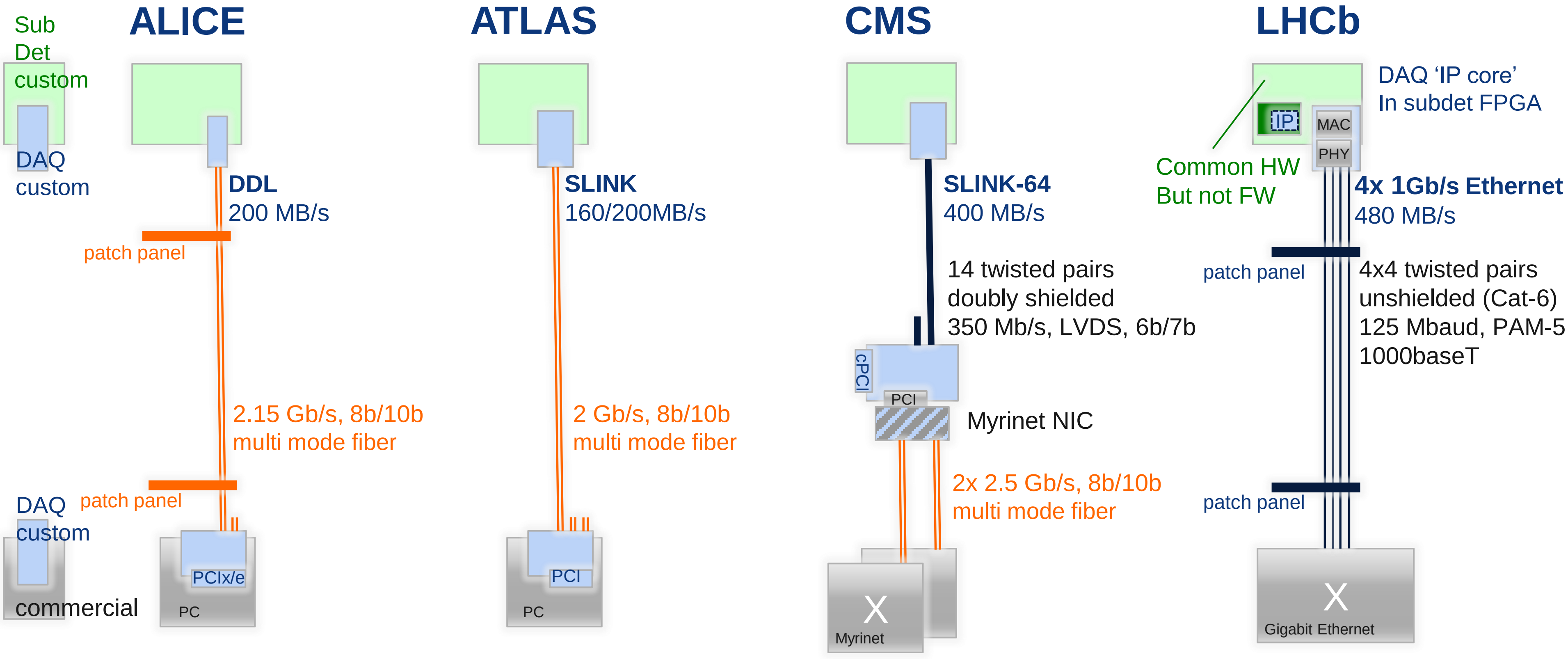
DAQ@LHC workshop, 13 March 2013

Hannes Sakulin / PH-CMD (CMS Data Acquisition and Trigger)

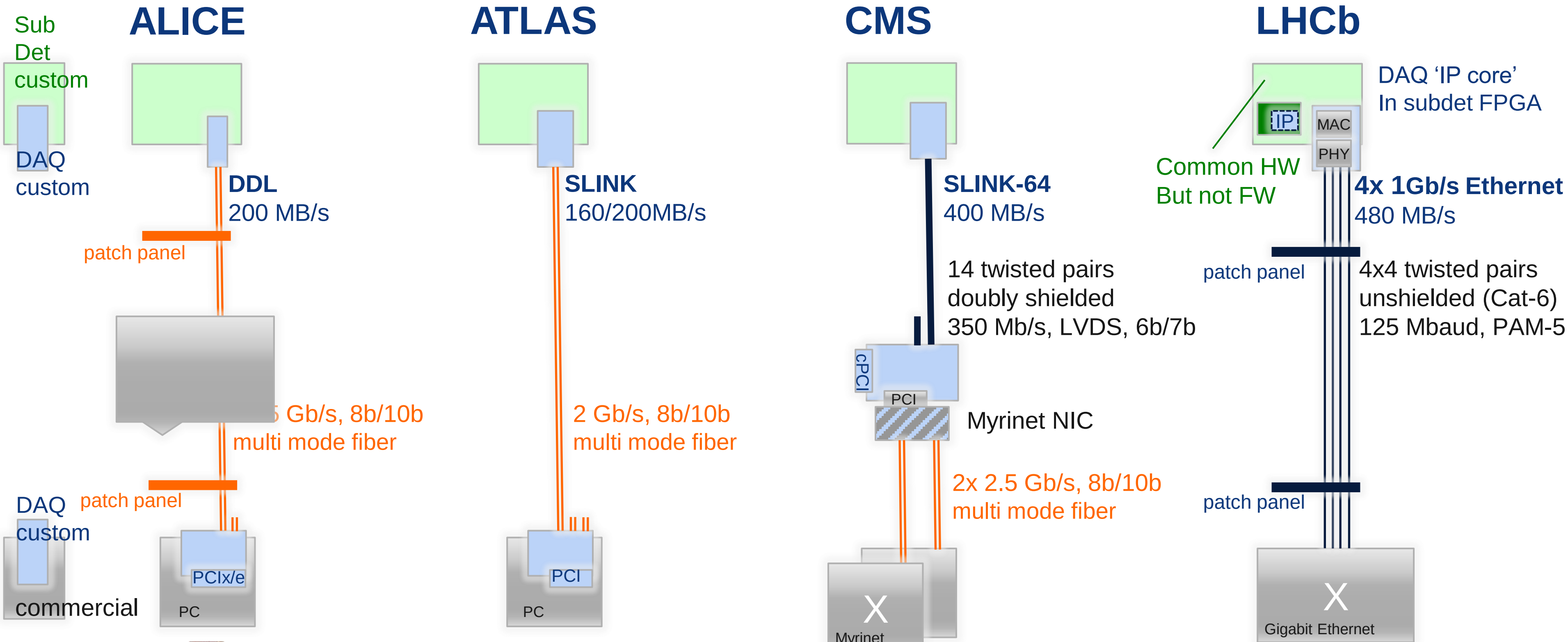
With input from: Filippo Costa & Csaba Soos (ALICE); Markus Joos & Stefan Haas (ATLAS), Dominique Gigi, Attila Racz, Christoph Schwick & Konstanty Sumorok (CMS), Niko Neufeld (LHCb)


Sub
Det
custom
DAQ
custom
DAQ
custom
commercial
ALICE
DDL
200 MB/s
patch panel
patch panel
PCIx/e
PC
ATLAS
SLINK
160/200MB/s
PCI
PC
CMS
SLINK-64
400 MB/s
cPCI
PCI
Myrinet NIC
X
Myrinet
LHCb
DAQ 'IP core'
In subdet FPGA
IP
MAC
PHY
Common HW
But not FW
4x 1Gb/s Ethernet
480 MB/s
patch panel
patch panel
X
Gigabit Ethernet

The Links


Sub
Det
custom
DAQ
custom
ALICE
DDL
200 MB/s
patch panel
2.15 Gb/s, 8b/10b
multi mode fiber
patch panel
DAQ
custom
commercial
PCIx/e
PC
ATLAS
SLINK
160/200MB/s
2 Gb/s, 8b/10b
multi mode fiber
PCI
PC
CMS
SLINK-64
400 MB/s
14 twisted pairs
doubly shielded
350 Mb/s, LVDS, 6b/7b
cPCI
PCI
Myrinet NIC
2x 2.5 Gb/s, 8b/10b
multi mode fiber
X
Myrinet
LHCb
DAQ 'IP core'
In subdet FPGA
IP
MAC
PHY
Common HW
But not FW
4x 1Gb/s Ethernet
480 MB/s
patch panel
4x4 twisted pairs
unshielded (Cat-6)
125 Mbaud, PAM-5
1000baseT
patch panel
X
Gigabit Ethernet

The Links



SIU

HOLA

SLINK CMC

TELL-1

DIU

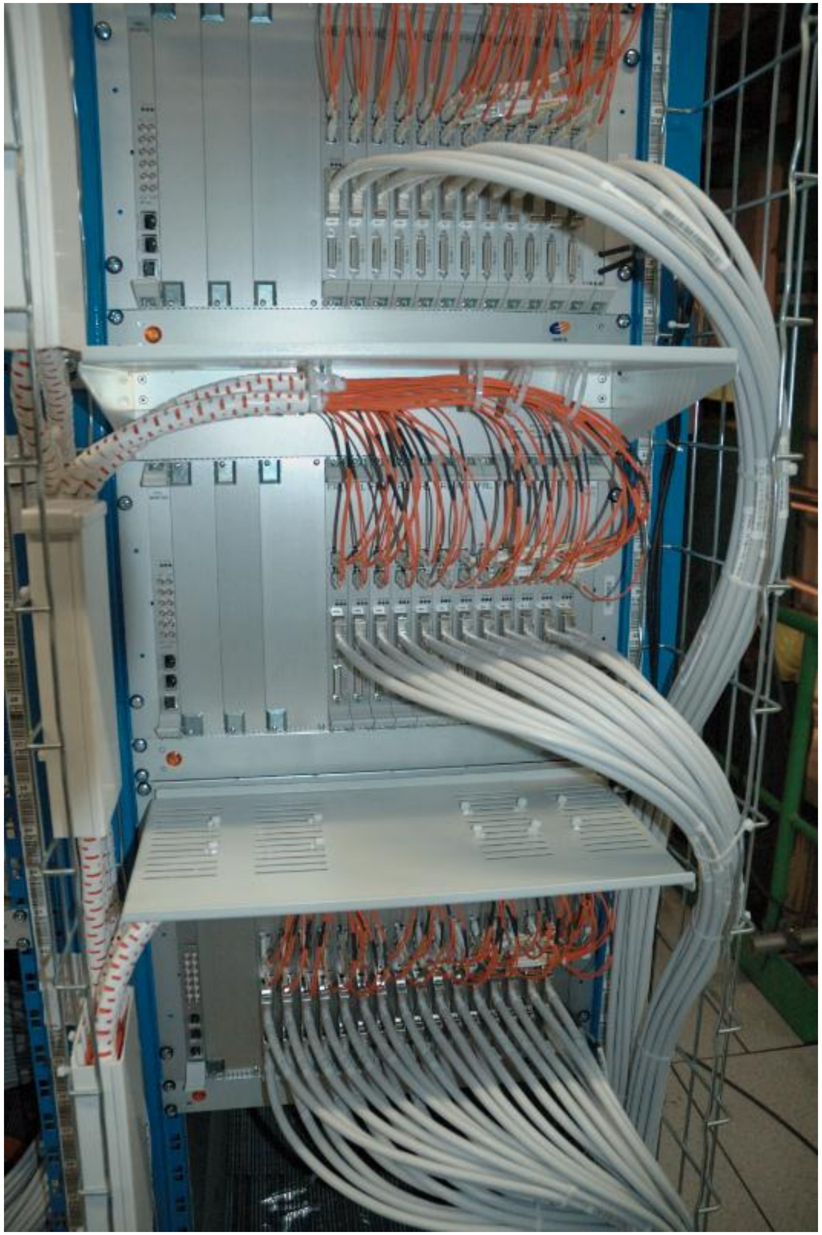
ROBIN

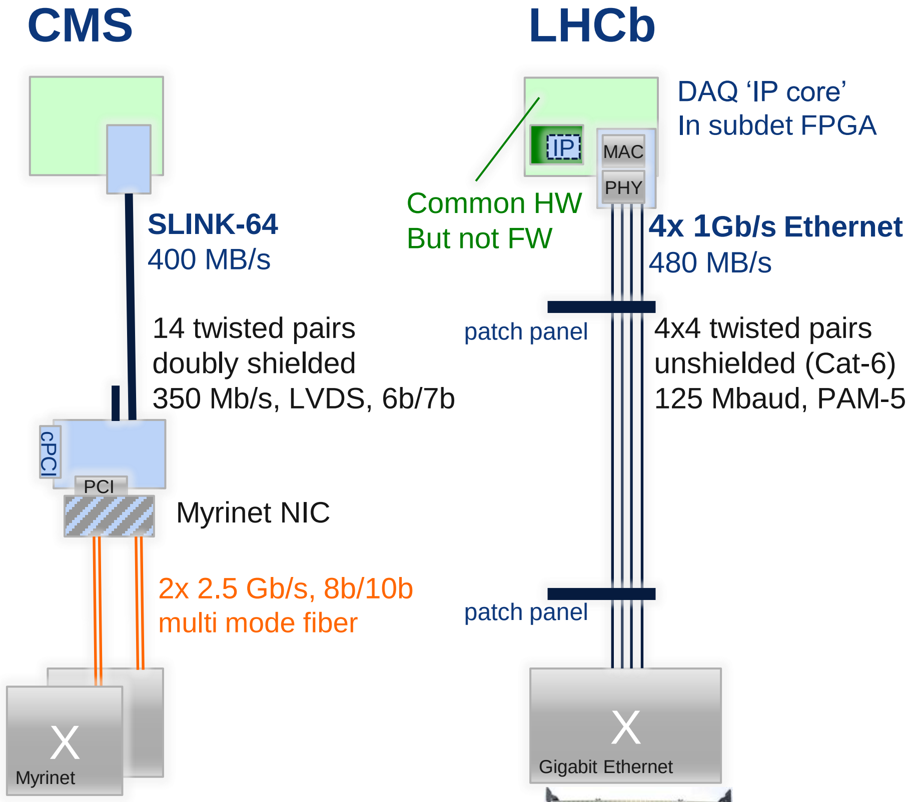
FRL

Switch
Force-10

The Links

The Links





CMS

SLINK-64
400 MB/s

14 twisted pairs
doubly shielded
350 Mb/s, LVDS, 6b/7b

cPCI

PCI

Myrinet NIC

2x 2.5 Gb/s, 8b/10b
multi mode fiber

LHCb

DAQ ‘IP core’
In subdet FPGA

Common HW
But not FW

4x 1Gb/s Ethernet
480 MB/s

patch panel

4x4 twisted pairs
unshielded (Cat-6)
125 Mbaud, PAM-5

patch panel

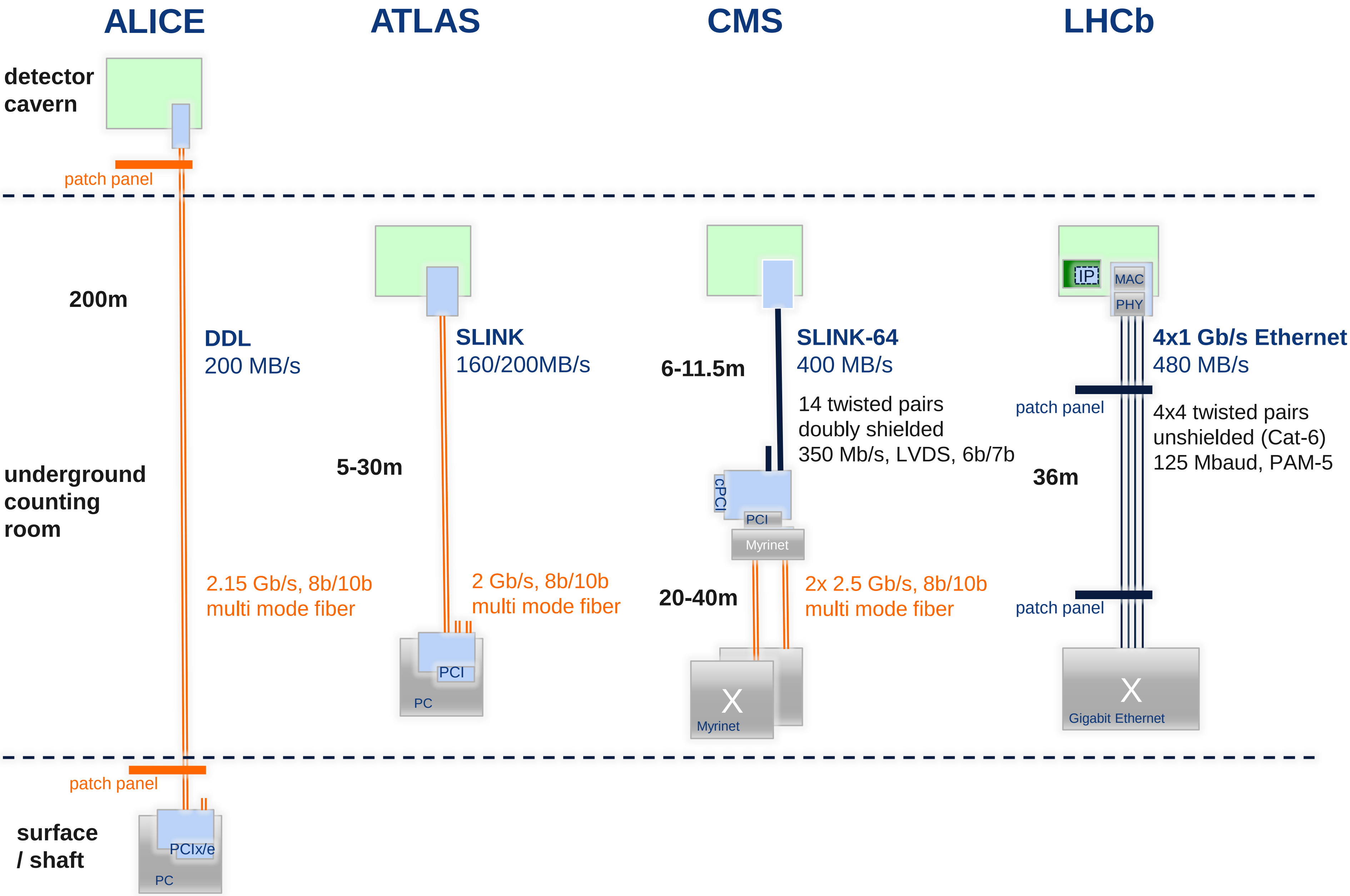
SLINK
CMC

TELL-1

FRL

Switch
Force-10

From Where to Where



ALICE

ATLAS

CMS

LHCb

detector cavern

underground counting room

surface / shaft

ALICE: patch panel; 200m; DDL 200 MB/s; 2.15 Gb/s, 8b/10b multi mode fiber; patch panel; PCIx/e; PC

ATLAS: SLINK 160/200MB/s; 5-30m; 2 Gb/s, 8b/10b multi mode fiber; PCI; PC

CMS: SLINK-64 400 MB/s; 6-11.5m; 14 twisted pairs doubly shielded 350 Mb/s, LVDS, 6b/7b; cPCI; PCI; Myrinet; 20-40m; 2x 2.5 Gb/s, 8b/10b multi mode fiber; X Myrinet

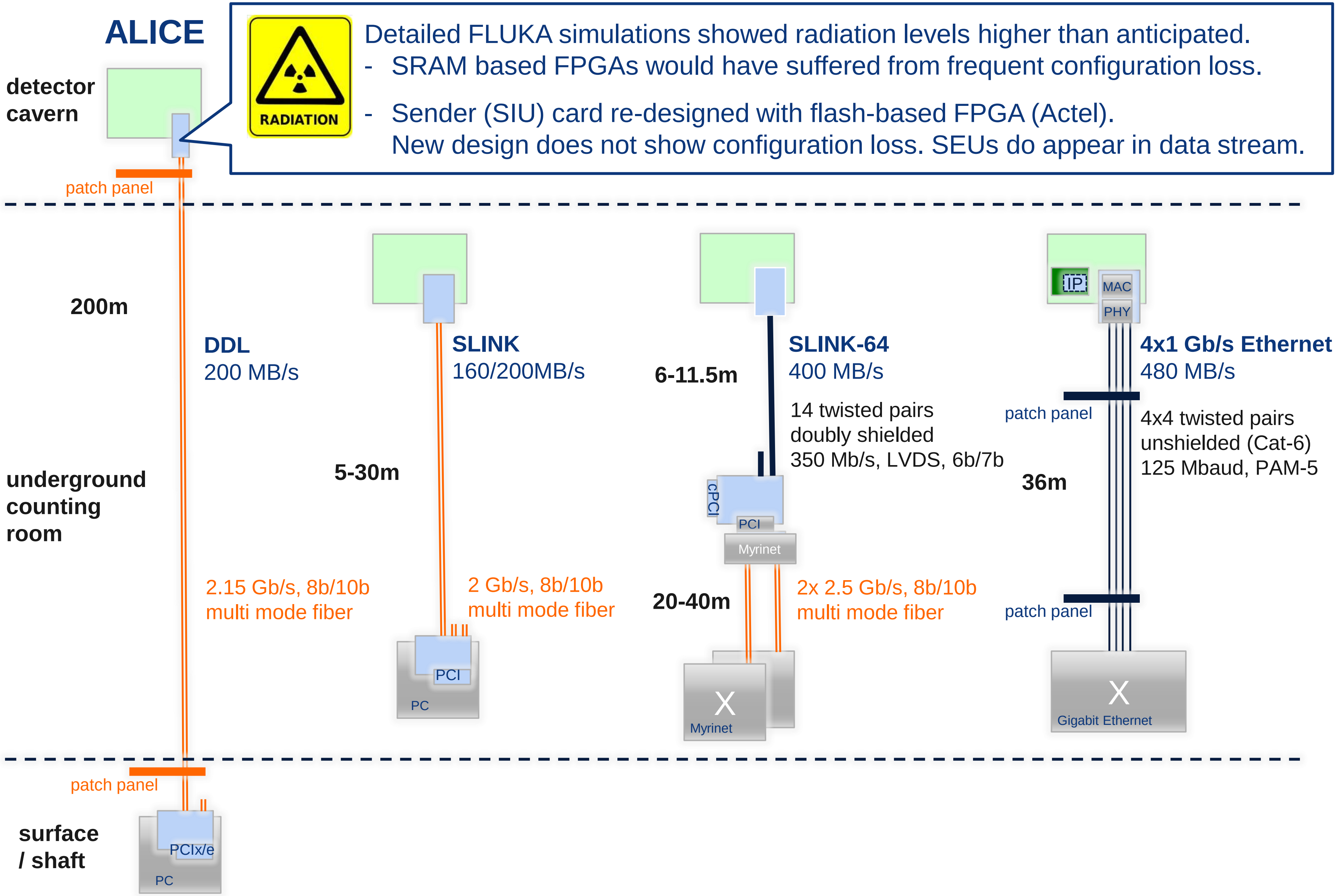
LHCb: IP; MAC; PHY; 4x1 Gb/s Ethernet 480 MB/s; patch panel; 4x4 twisted pairs unshielded (Cat-6) 125 Mbaud, PAM-5; 36m; patch panel; X Gigabit Ethernet

From Where to Where

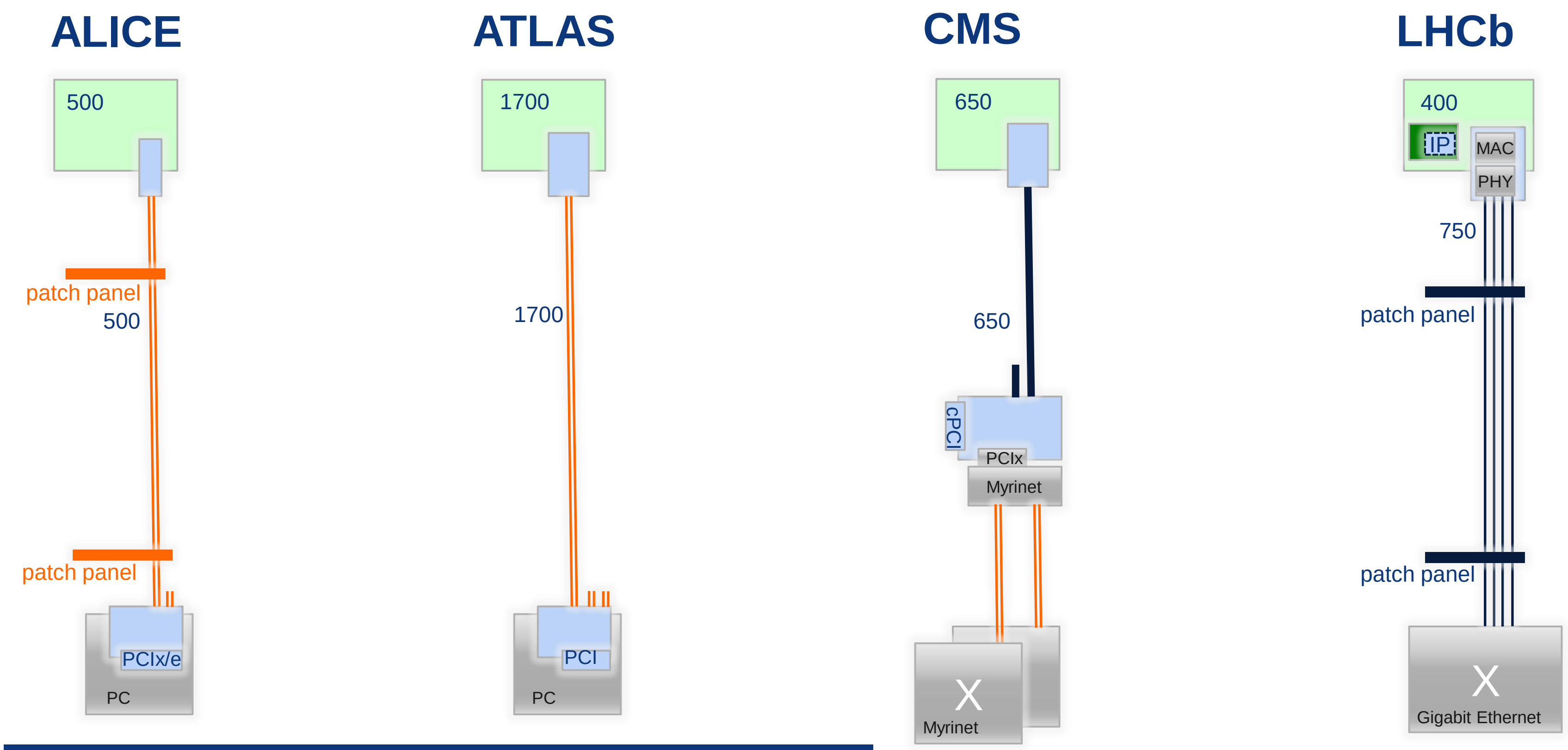
Detailed FLUKA simulations showed radiation levels higher than anticipated.

- SRAM based FPGAs would have suffered from frequent configuration loss.

- Sender (SIU) card re-designed with flash-based FPGA (Actel).
  New design does not show configuration loss. SEUs do appear in data stream.



Try to break it



Pre-production tests: finding the limits of the hardware

Test with proton and neutron irradiation.

Found <1 data error/hour in all 500 links.

Discovered bug in one batch of FPGAs.

Test link with 7 dB attenuator.

Tested maximum speed on ~30 cables (10m).
Observed bit errors
- above 520 MB/s in worst cable
- above 680 MB/s in best cable

Set nominal speed to **400 MB/s (=50 MHz) Keep some margin !**

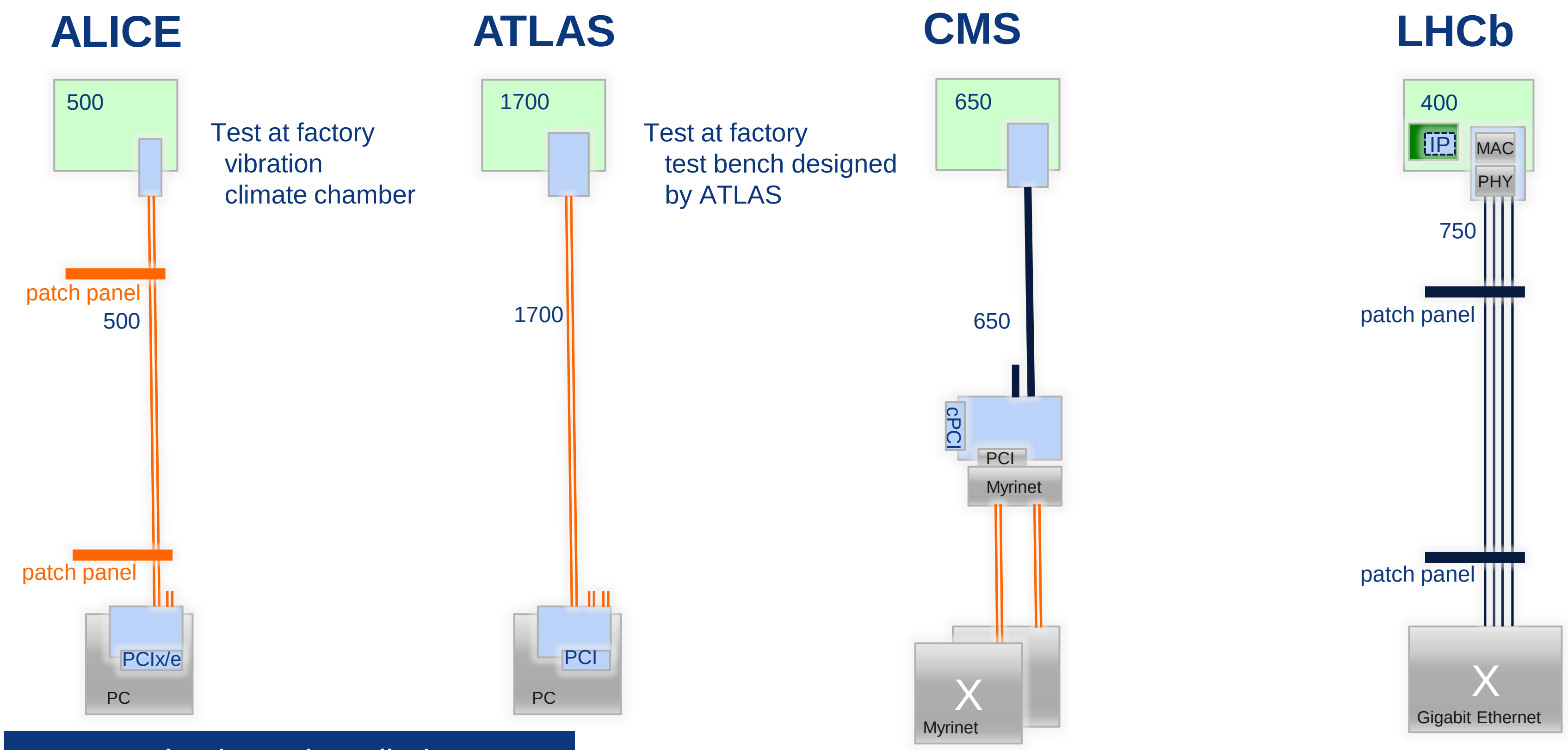
Bug in LVDS receiver found workaround implemented.

Test with long cables (> 100m)

Test in hostile environment
- roll with small radius
- near fluorescent light

System still working fine Under these conditions

Production tests



Test at factory
vibration
climate chamber

Test at factory
test bench designed
by ATLAS

Post-production & installation tests

Systematic test at factory
- vibration
- climate chamber

Test of links with data generator.

Systematic tests at factory with test designed bench by ATLAS - repeated at CERN.

Test at low and high supply voltage margins.

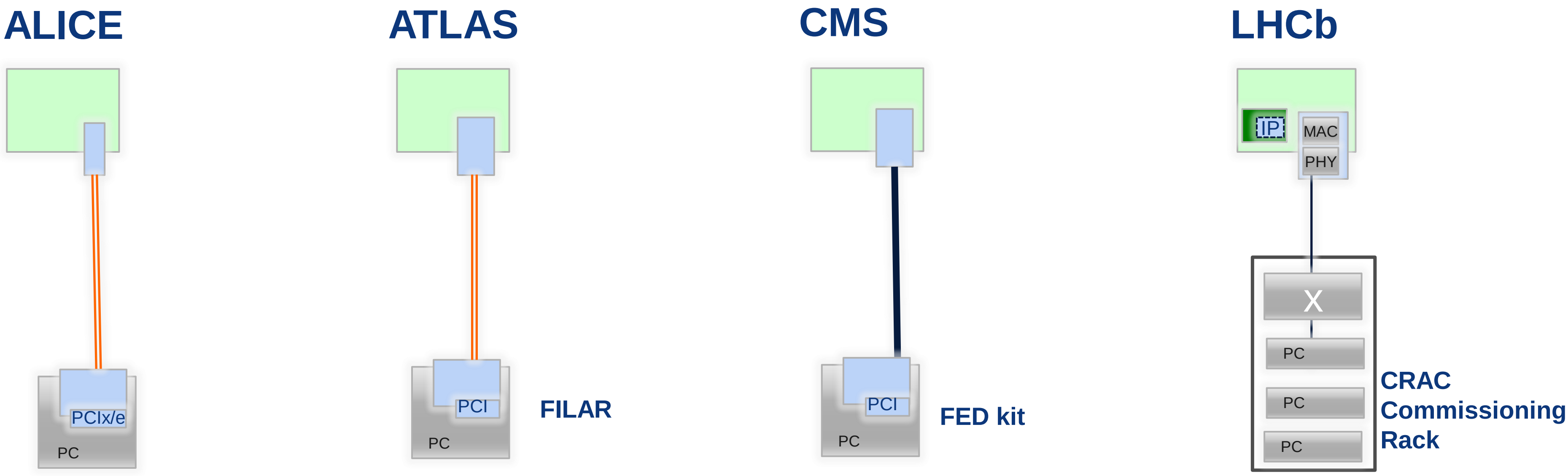
BER tests on few cards.

Fiber attenuation tested in-situ.

1) visual, voltages, currents - 5min
2) Pattern test at full speed
   - All senders - 12h
   - All receivers -12h
     5% of cards with problems (mostly soldering)
   - All senders+receivers+cables
     - 24h, 0 errors
3) Burn-in
4) All cables + receivers post-installation w. mobile tester – 2 min

Systematic production tests of Tell-1 board but no temperature cycles.

Structured cabling (long UTP cables and patch panels) tested by company.



Kits given to sub-detectors to test their readout in the lab.

Sub-detector commissioning

All (central) DAQ groups gave a link and software to the sub-detectors so that they could commission their readout.

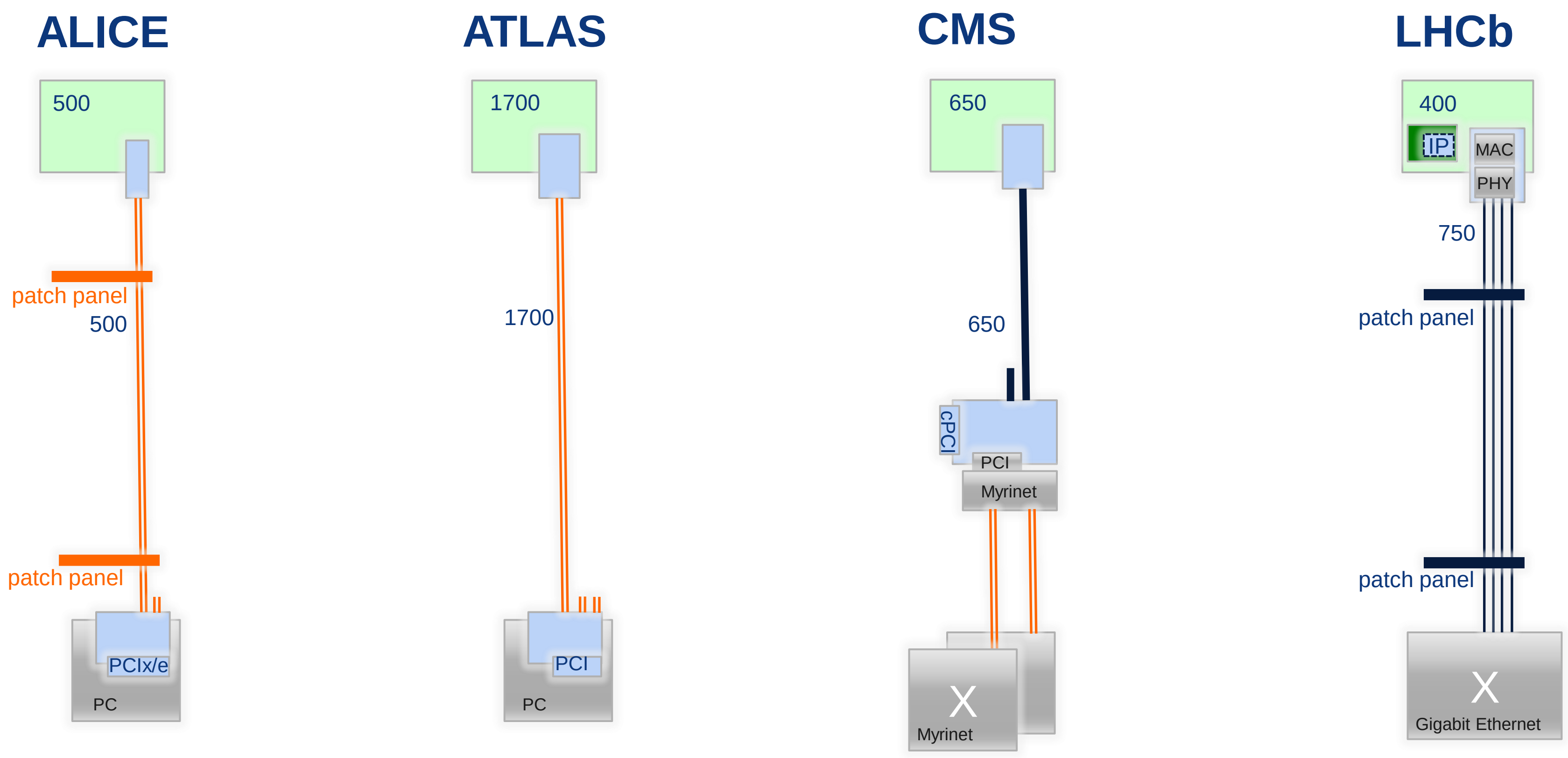
Links were used in test beams / cosmic tests etc.

Then tested with one sub-detector at a time at the pit.

Then global runs.

Sub-detector commissioning

Commissioning at the Pit



Problems found in commissioning

No big problems.

SFPs in receiver side failing. Needed to replace all in 2008.

Plugged it in and it worked.

Some sub-dets occasionally skipping data.
=> Protection added against missing trailers.

Voltage dips in power supply in case of big events in one detector (2008 splashes). Upgraded some senders to v2 which has a voltage regulator.

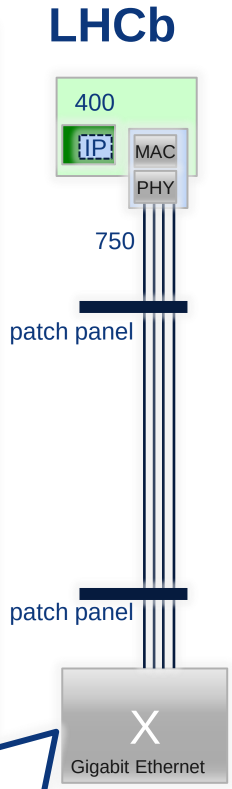
Problems between detector-specific firmware and DAQ "IP core" (both in same FPGA) found. Especially at high rate.

Specification and "management" problem.

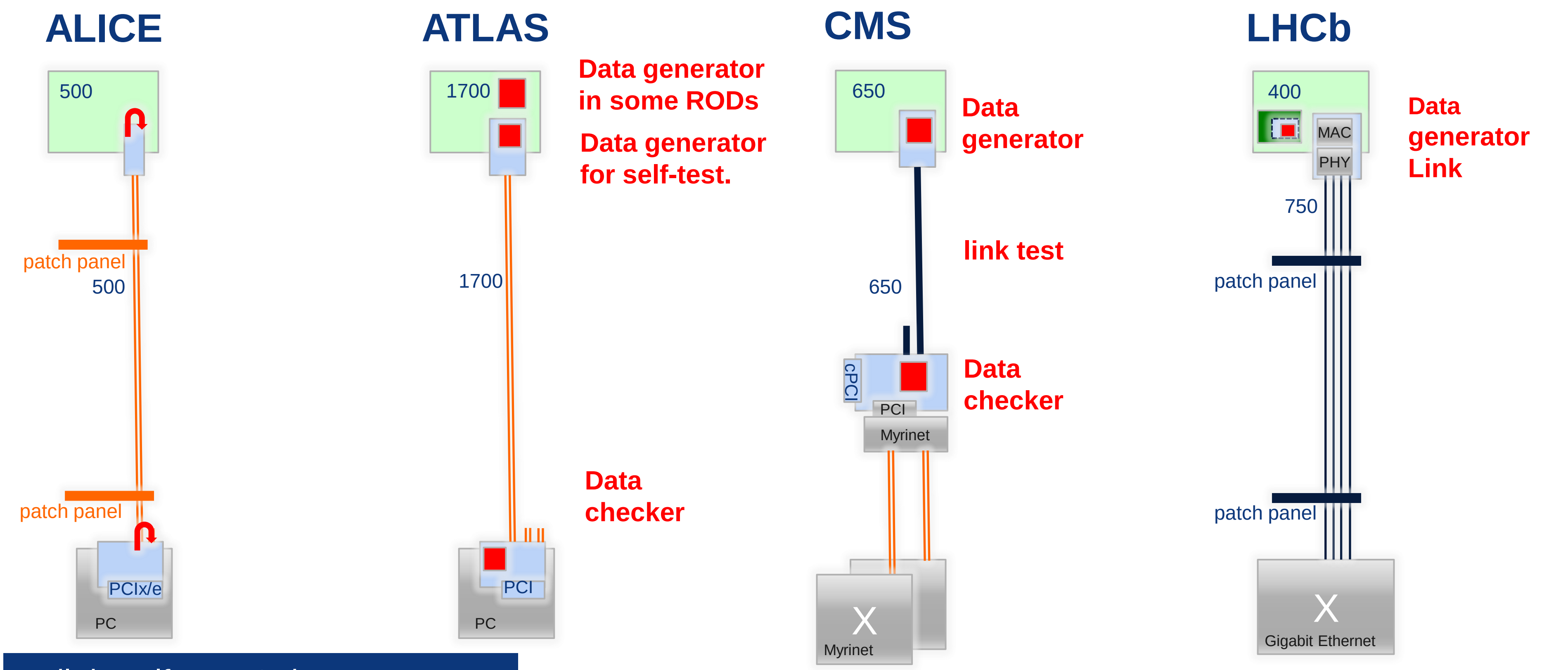
Excursion – commissioning the LHCb event builder

- Link based on 1Gb/s Ethernet is simple and powerful.
  No (direct) flow-control.

- **But the devil is in the detail**
- Found only one affordable switch on the market (at the time) that was able to buffer the event building traffic (synchronized push of all sources to same destination) – Force-10
- Not all switches able to handle jumbo frames (packet drops).
- Burst of losses after configuration (reset) of senders.
  Fixed by switch firmware update.
- Scheduling in switch needed to be accelerated.
- Buffer distribution in switch needed to be fine-tuned.
- Corrective measures against tails in event size distribution. (otherwise drops of large events possible).
- Link aggregation between main and edge router tuned to use one link per multi-event packet to avoid packet drops.
- Small clock difference between main and edge routes causing packet drops. Frame gap introduced.
- IRQ coalescence needed in receiver PC.
- Monitoring had adverse effect on switch performance.



Built-in self tests



Built-in self tests & data generators

Loop-back tests.
At various levels.

Initiated by receiver.

Often used to debug problems.

Self-test feature of link.
Not much used because
It does not test the interface from ROD to sender.

Links are tested by generating data in the RODs or by taking calibration / cosmics data.

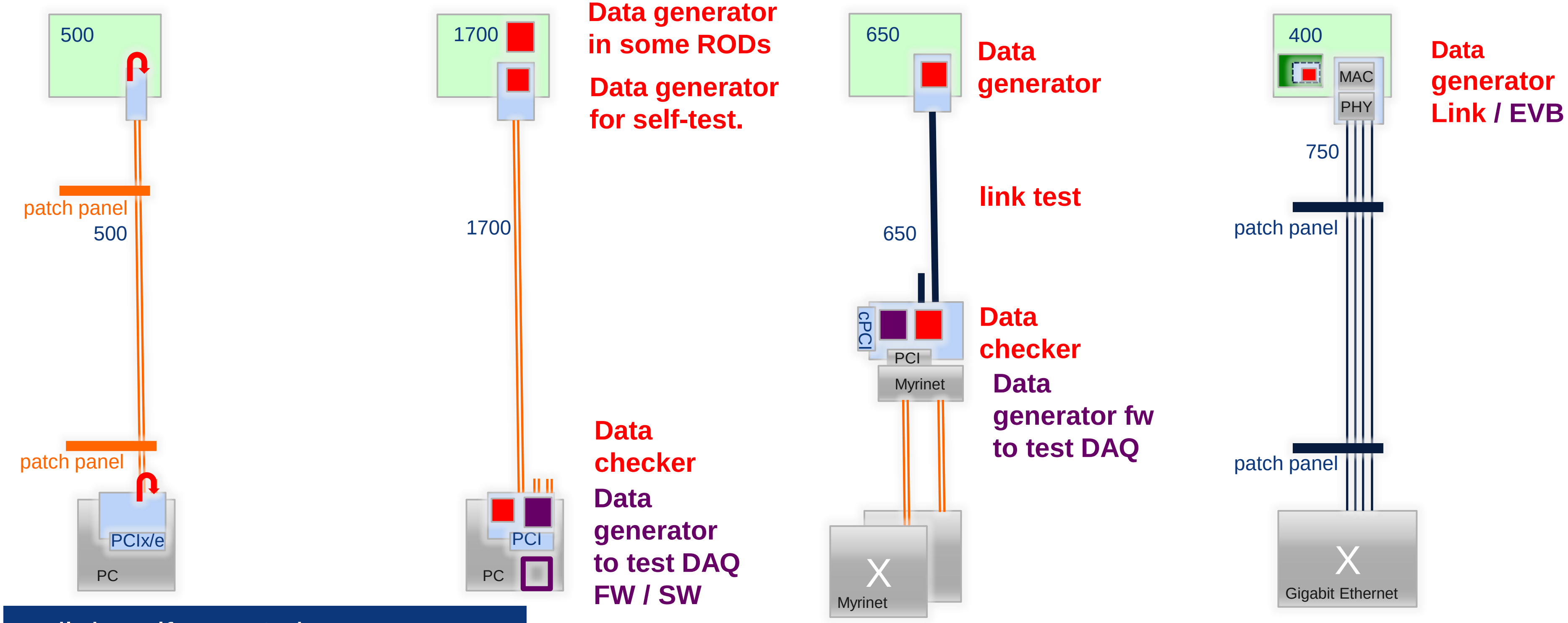
Link test.
Initiated by receiver.
All firmware.

Done every time DAQ is started.

Data generator firmware for tests of DAQ.

Data generator mode of readout board.

Built-in self tests



ALICE

ATLAS

Data generator in some RODs

Data generator for self-test.

Data checker

Data generator to test DAQ FW / SW

CMS

Data generator

link test

Data checker

Data generator fw to test DAQ

LHCb

Data generator Link / EVB

Built-in self tests & data generators

Loop-back tests.
At various levels.

Initiated by receiver.

Often used to debug problems.

Self-test feature of link.
Not much used because
It does not test the interface from ROD to sender.

Links are tested by generating data in the RODs or by taking calibration / cosmics data.

Link test.
Initiated by receiver.
All firmware.

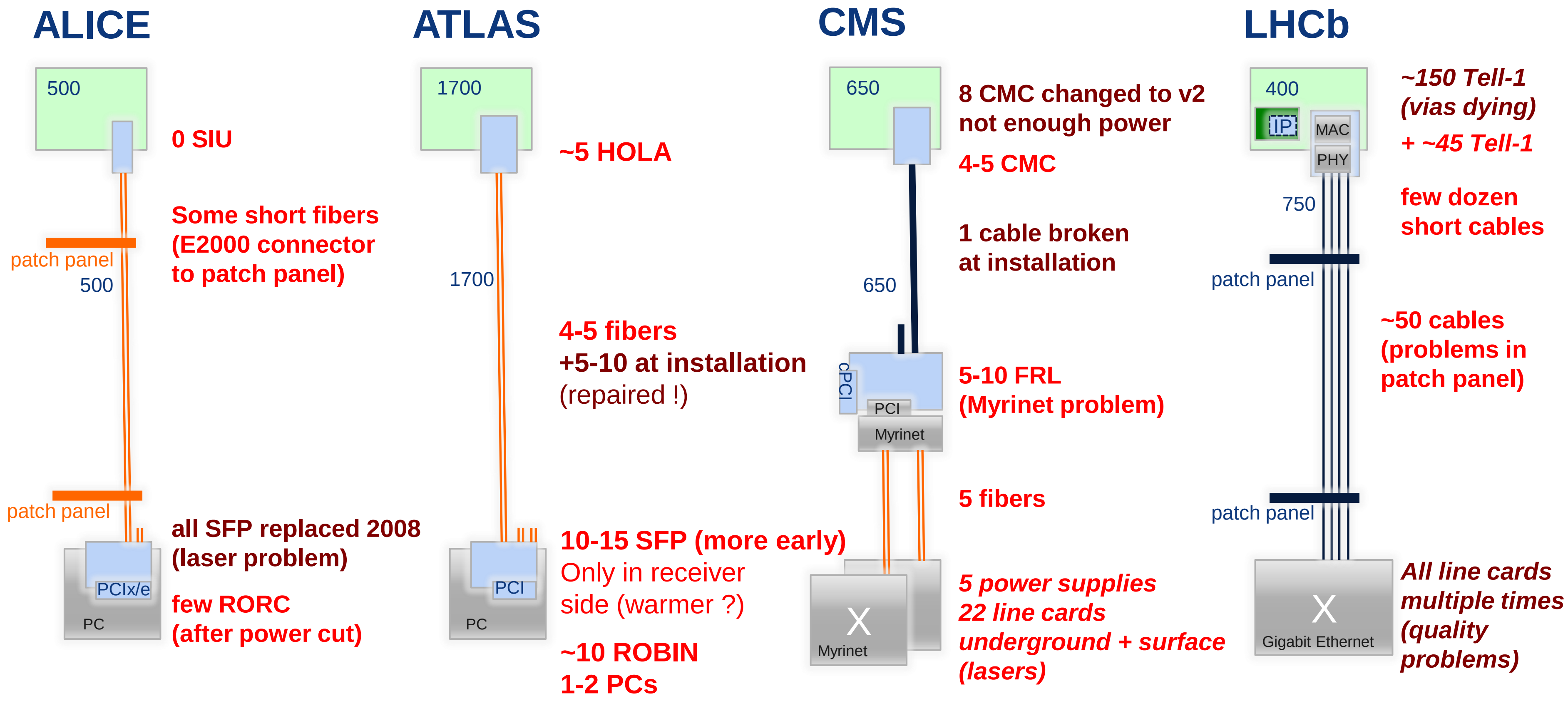
Done every time DAQ is started.

Data generator firmware for tests of DAQ.

Data generator mode of readout board.

Robustness

What parts broke ?



ALICE

0 SIU

Some short fibers (E2000 connector to patch panel)

all SFP replaced 2008 (laser problem)

few RORC (after power cut)

ATLAS

~5 HOLA

4-5 fibers
+5-10 at installation
(repaired !)

10-15 SFP (more early)
Only in receiver side (warmer ?)

~10 ROBIN
1-2 PCs

CMS

8 CMC changed to v2 not enough power

4-5 CMC

1 cable broken at installation

5-10 FRL (Myrinet problem)

5 fibers

5 power supplies
22 line cards underground + surface (lasers)

LHCb

~150 Tell-1 (vias dying)
+ ~45 Tell-1

few dozen short cables

~50 cables (problems in patch panel)

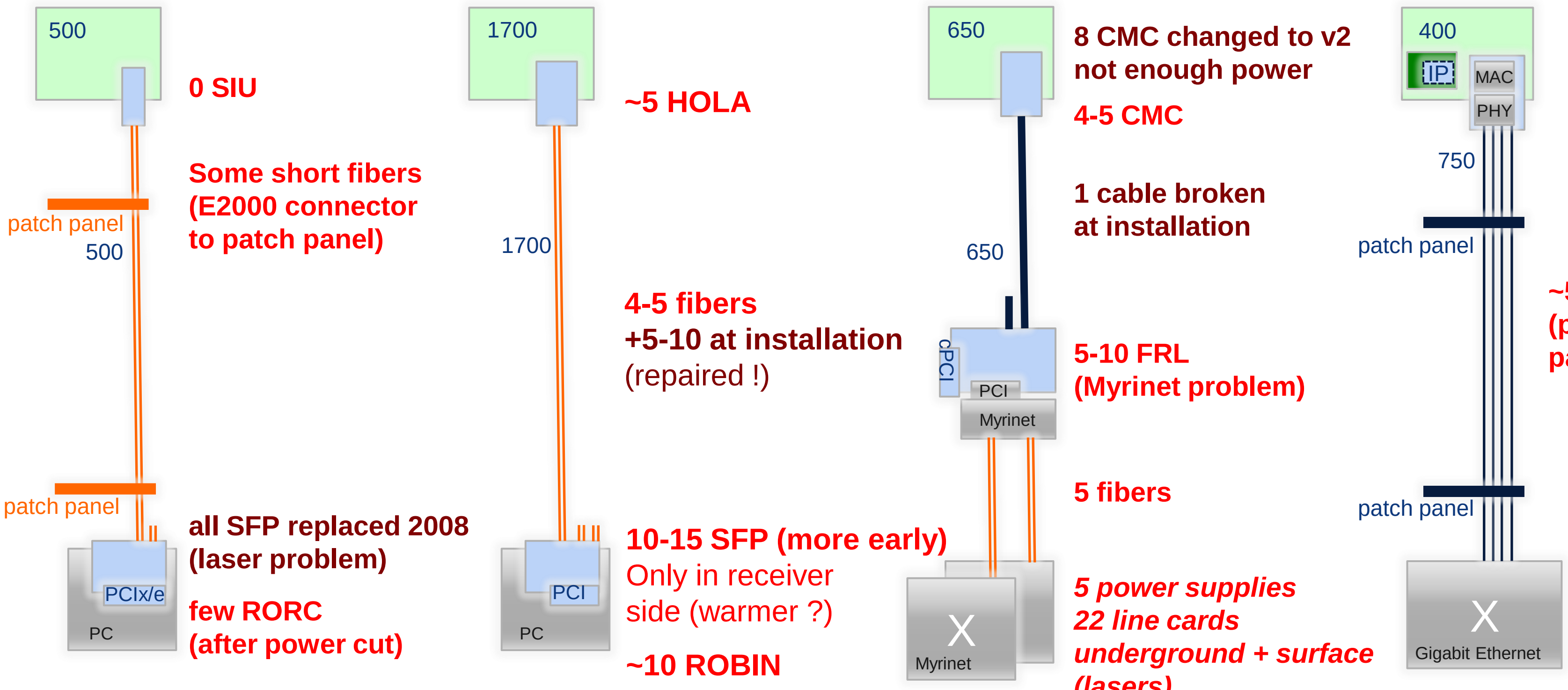
All line cards multiple times (quality problems)

# items replaced during commissioning

# items replaced in 2010-2013 operations

Down times caused by the links

ALICE ## ATLAS ## CMS ## LHCb



ALICE

0 SIU

Some short fibers (E2000 connector to patch panel)

all SFP replaced 2008 (laser problem)

few RORC (after power cut)

ATLAS

~5 HOLA

4-5 fibers
+5-10 at installation
(repaired !)

10-15 SFP (more early)
Only in receiver side (warmer ?)

~10 ROBIN
1-2 PCs

CMS

8 CMC changed to v2 not enough power

4-5 CMC

1 cable broken at installation

5-10 FRL (Myrinet problem)

5 fibers

5 power supplies
22 line cards
underground + surface (lasers)

LHCb

~150 Tell-1 (vias dying)
+ ~45 Tell-1

few dozen short cables

~50 cables (problems in patch panel)

All line cards multiple times (quality problems)

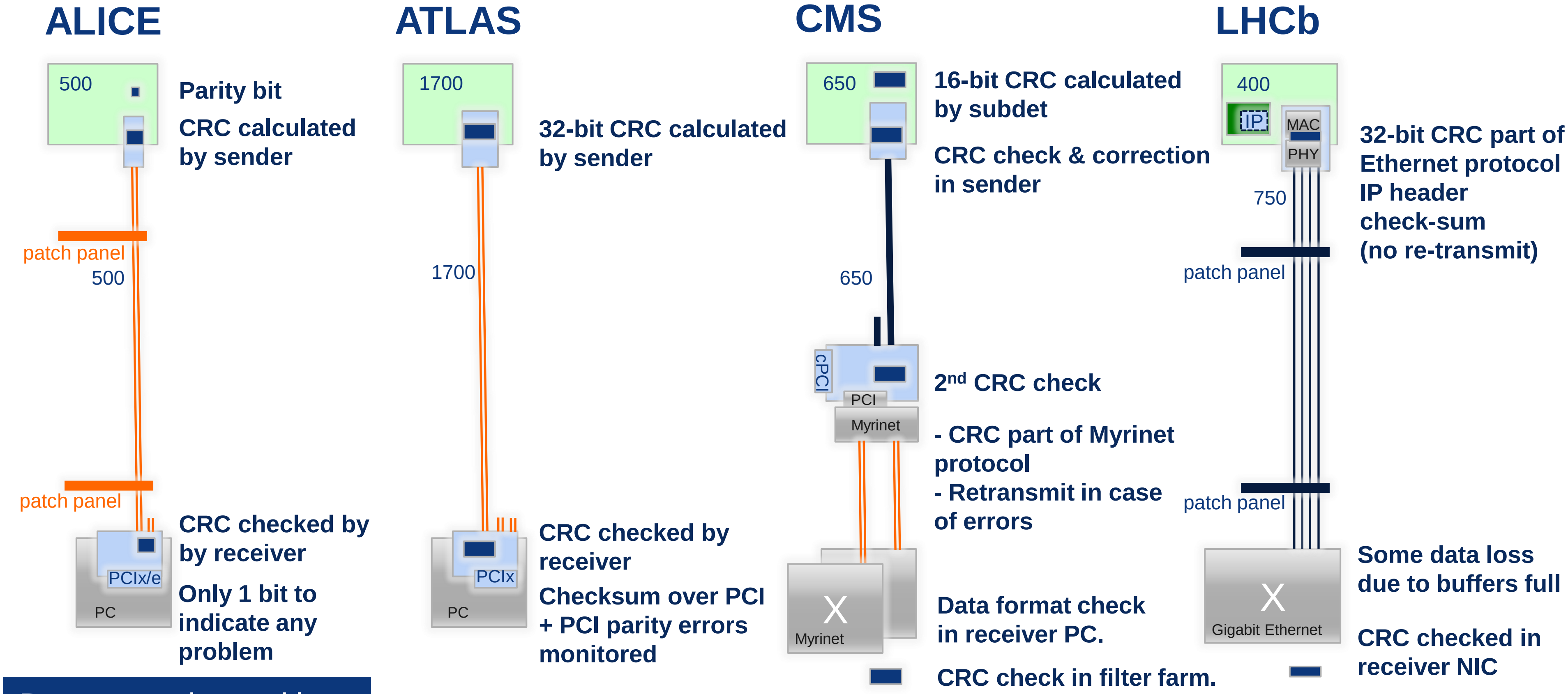
# items replaced during commissioning

# items replaced in 2010-2013 operations

Down time due to readout link

very small	very small (1-2 h per year) (2011: few hours due to incompatibility btw. new ROS PC and old NIC)	very small 1-2h / year	Very rare but down time due to Tell-1 boards and Force-10 line cards

Robustness – Data Corruption / Loss



ALICE

Parity bit

CRC calculated by sender

CRC checked by by receiver

Only 1 bit to indicate any problem

ATLAS

32-bit CRC calculated by sender

CRC checked by receiver

Checksum over PCI + PCI parity errors monitored

CMS

16-bit CRC calculated by subdet

CRC check & correction in sender

2nd CRC check

- CRC part of Myrinet protocol
- Retransmit in case of errors

Data format check in receiver PC.

CRC check in filter farm.

LHCb

32-bit CRC part of Ethernet protocol IP header check-sum (no re-transmit)

Some data loss due to buffers full

CRC checked in receiver NIC

Data corruption and loss

ALICE

Data corruption does happen

Difficult to debug since no information about where error occurred

ATLAS

No CRC errors / CRC errors rate so low that never investigated

initial PCI errors fixed by BIOS update – **no PCI errors since**

CMS

Few SLINK CRC errors per day on 1-2 links; Not increasing over time

Subdet-to-Sender CRC errors - in case of large (splash) events in fwd muon chambers

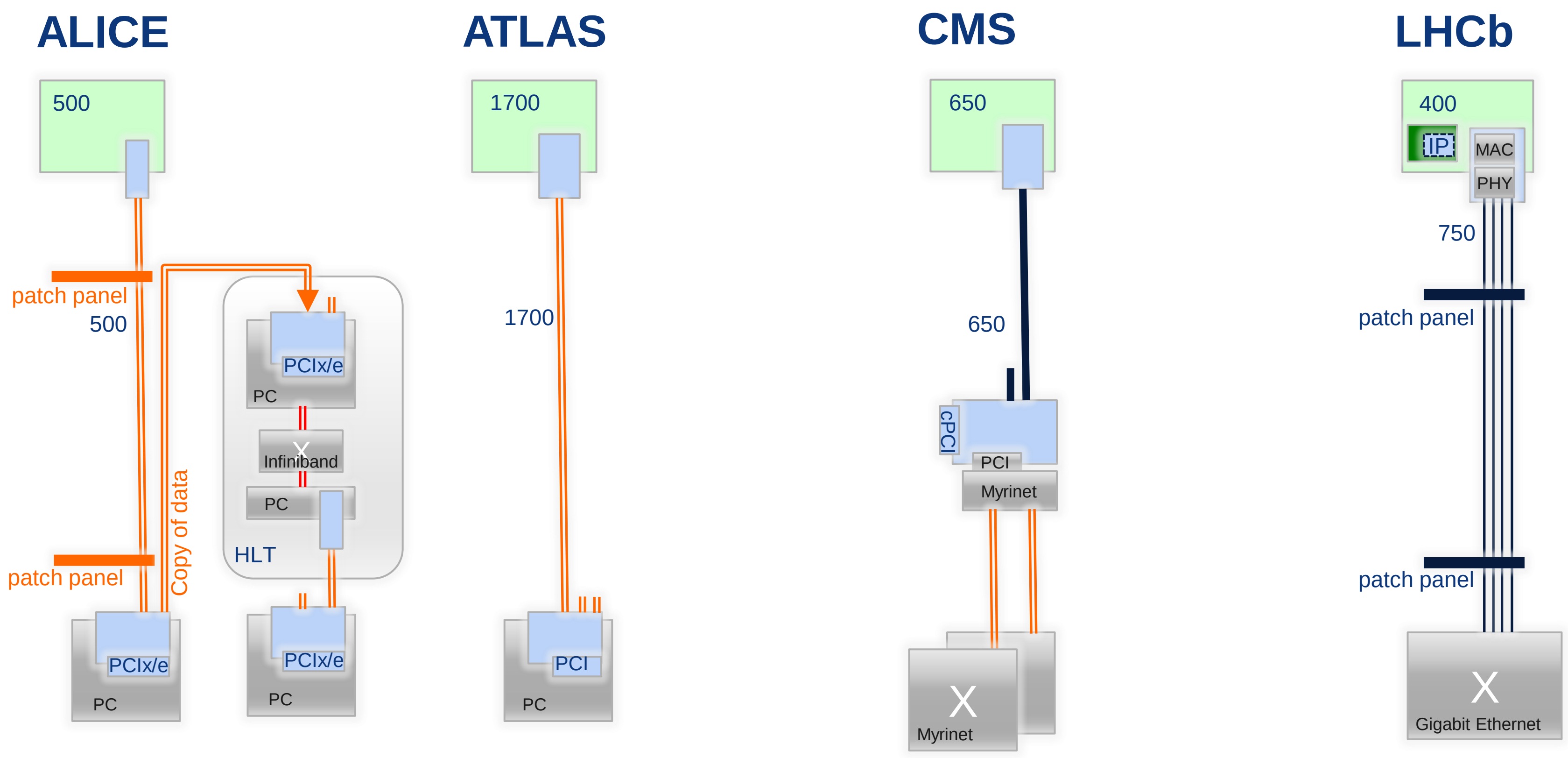
Myrinet CRC errors in bursts, recovered by retransmit. may point to dying laser

LHCb

No CRC errors, or a lot if contact problems

2-3 MEP packets (12 ev) lost per minute in switch

Some de-synchronization But usually a detector problem.



Dealing with corrupted input

Run stopped after 10 CRC errors.

DAQ able to deal with re-synchronization, but HLT not.

Mildly corrupted data served to HLT.

All corrupted data sent to debugging stream in event monitoring.

DAQ able to recover from missing fragments / de-synchronization.

FED-to-SLINK CRC errors flagged in event header. Events with CRC/Data Format errors dumped to disk (first 10).

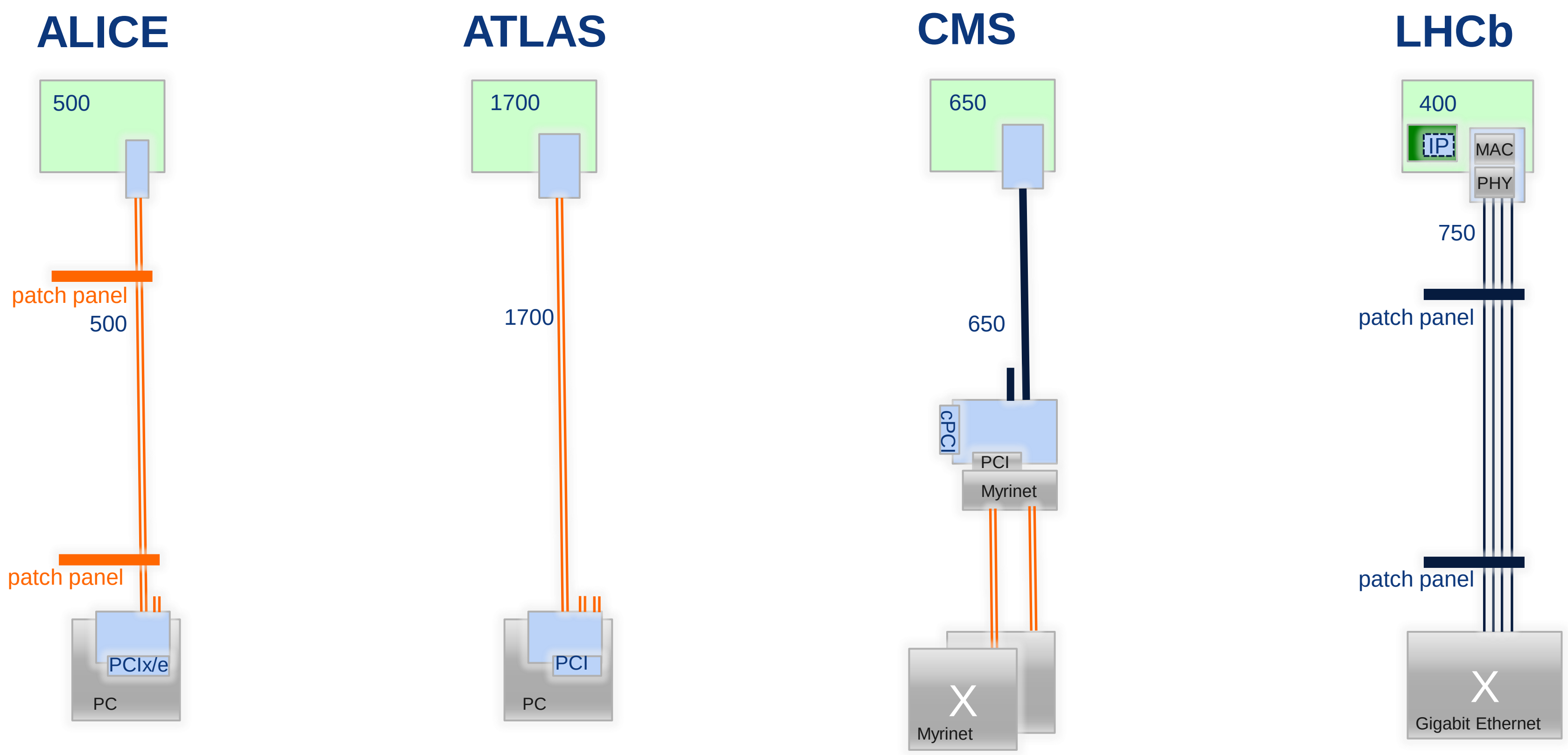
Stop run if fragment out of order detected (no dump).

DAQ gets stuck if re-sync not at same event number in all inputs

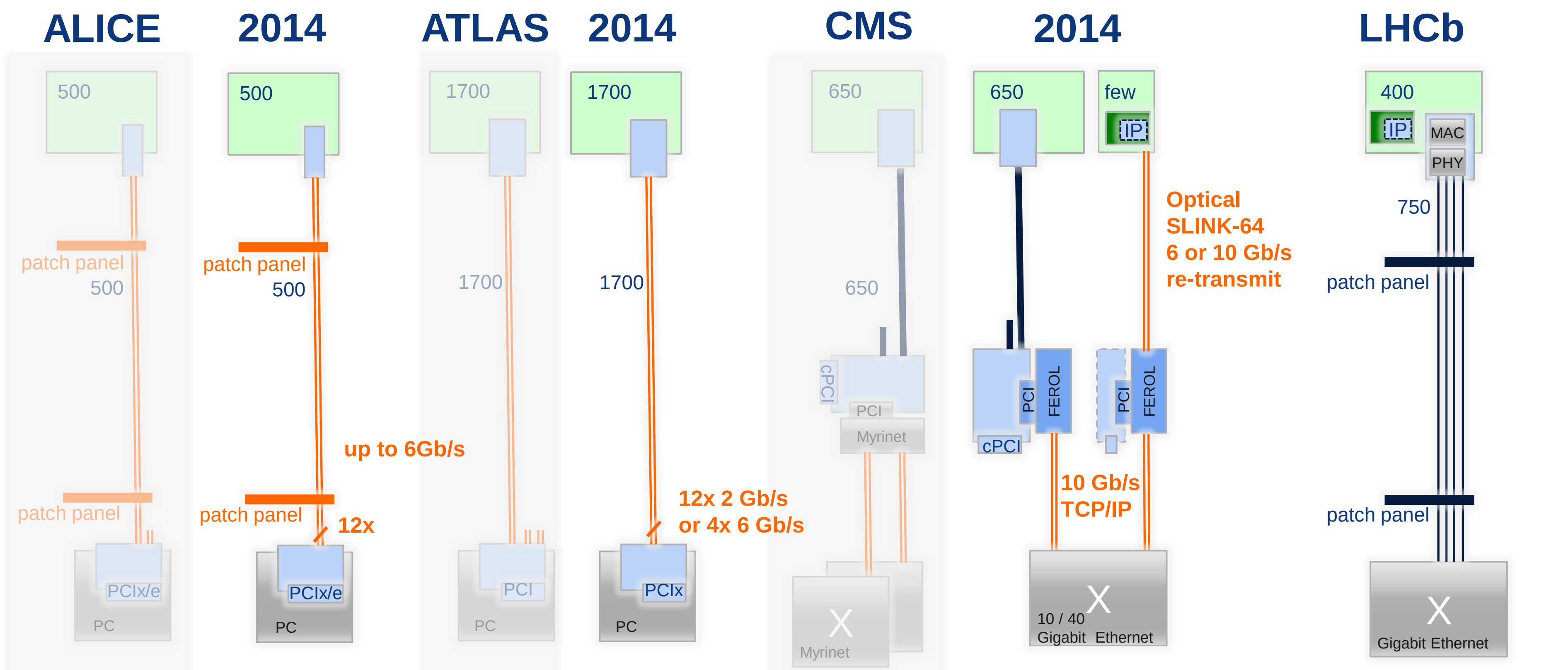
Corrupted data are dropped.

Dealing with Corrupted Input

Other Pros & Cons



	ALICE	ATLAS	CMS	LHCb
The Advantages of the link	Bi-directional. Radiation hard.	Simplicity. Ease of interfacing on Readout-Driver side.	High throughput Double CRC check easy to spot errors Cost effective.	Simple protocol. Copper more cost effective than fiber
The inconveniences	Data corruption difficult to debug. Not enough error bits in protocol	Little monitoring of senders. LC connectors may be connected shifted.	Clumsy cables	Deep buffers needed in switches.



Upgrade plans

Replace some receivers with new cRORC in LS1.

Probably same HW

Keep current protocol
Up to 6 Gb/s
Compatible with current senders.

IP core senders.

Replace all receivers with new ROBIN-NP
In LS1.

Keep current protocol.
Up to 6 Gb/s.
Compatible with current senders.

Add 100-150 links.

Some new FEDs sending over **optical SLINK-64**.
- IP core instead of mezzanine.
- **Re-transmit.**

Replace Myrinet with **10 Gb/s TCP/IP directly from FPGA**.

No changes over LS1.

The Future

Summary

- All four experiment have robust readout links
  - Very little down times due to the links
  - Very little data corruption
- Important to keep some margin in the specification
- Rigorous testing pays off.
  - Much better to find problems early
- Important to foresee error detection in the protocols
- Monitoring of the sender is useful
- Interface needs to be very well defined when giving IP cores to sub-detectors. We should keep this in mind for the upgrades

CERN