LHCb High Level Trigger in remote IT datacentre

CHEP 2018 – Sofia, BG - 12th July 2018
edoardo.martelli@cern.ch

Presentation agenda

- Rationale
- Setup
- Results
- DWDM transmission

Rationale

New DAQ systems

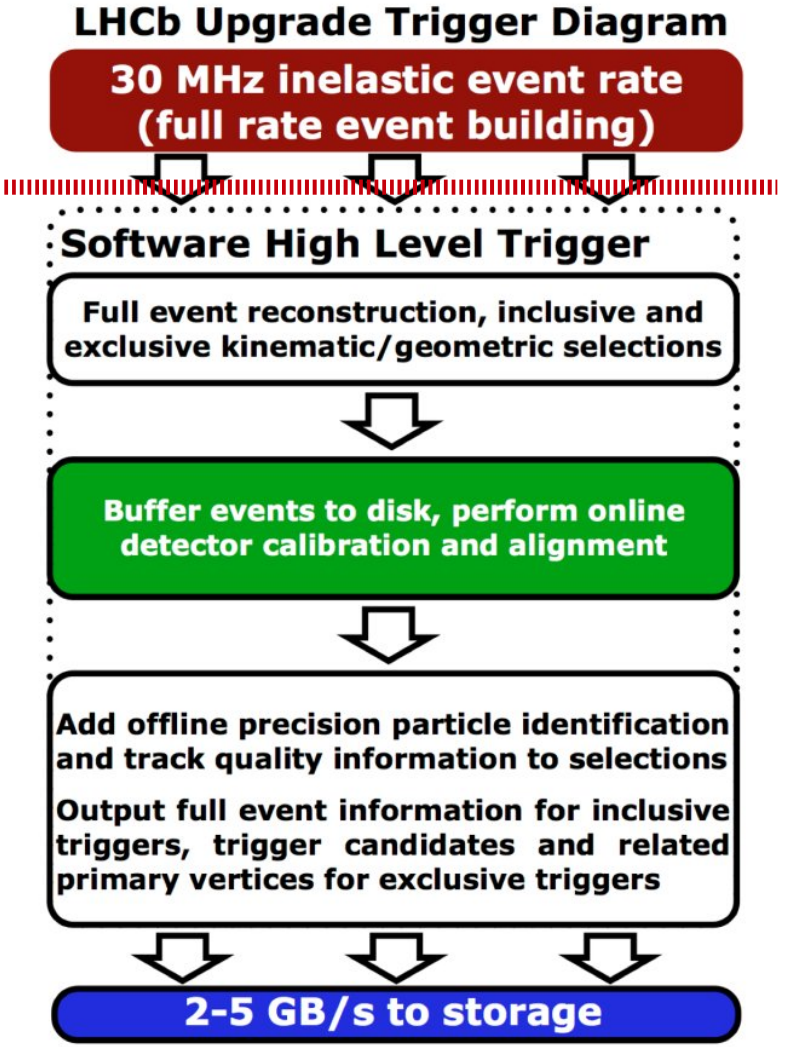
In the recent years the high-speed fabrics have been greatly improved for the needs of High Performance Computing.

Different technologies of 100 Gbit/s network cards are available.

CPU allows the implementation of complex Event Building algorithms (aggregation of the data parts), easier to develop and to maintain with respect to custom FPGAs based solutions.

Using servers allows the choice of the technology to be made at the latest moment: more bandwidth per price unit and a safer, more future proof choice.

An interface card, to read out the detector and to inject events fragments into a server, exploiting PCIe can be effectively plugged to a server: events buffering in the large servers' memory is extremely cheap.



[Credits: Umberto Marconi – INFN Bologna – LHCb]

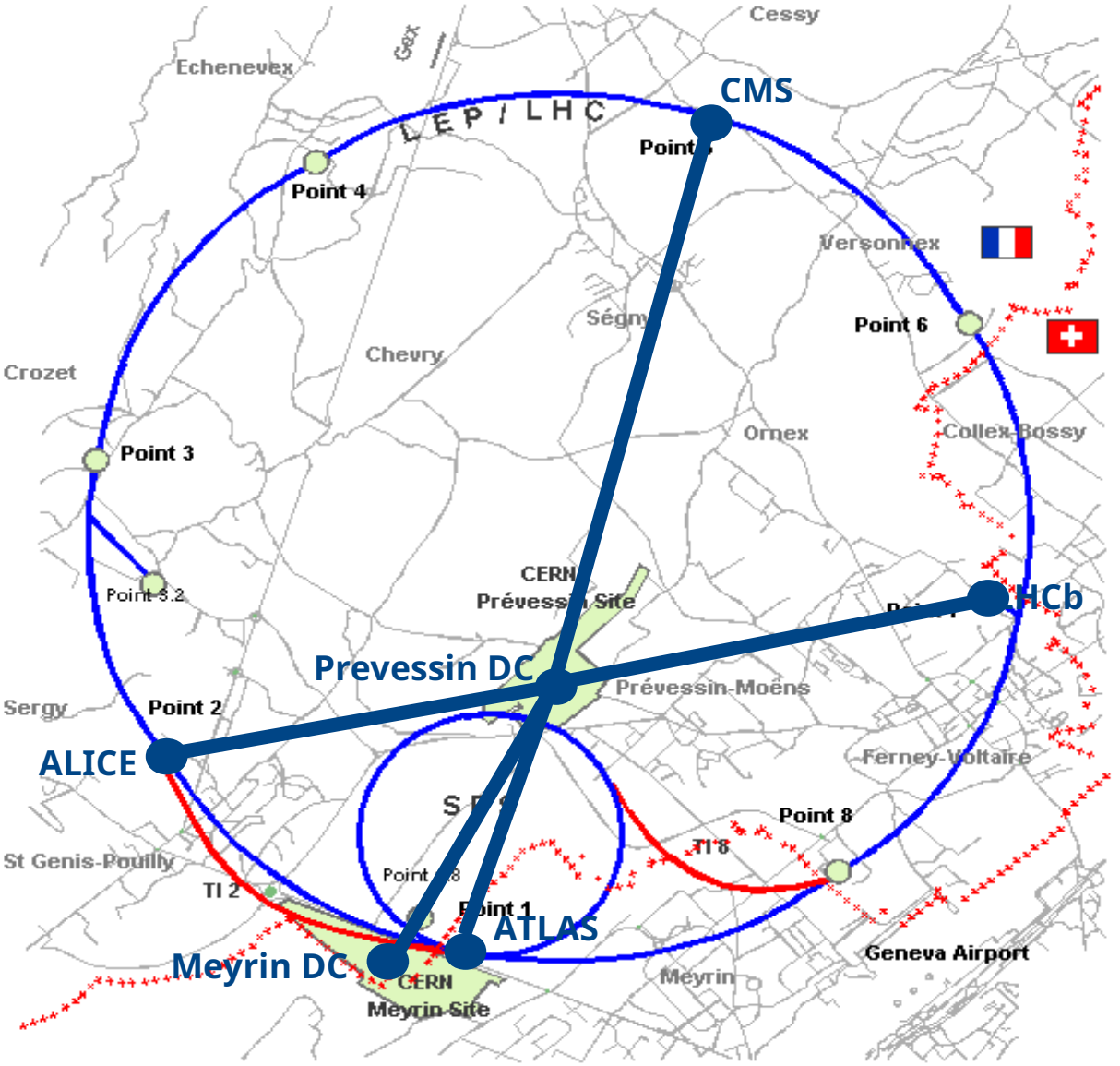
[Image credits: Concezio Bozzi – INFN Ferrara – LHCb]

A shared HLT

Feasibility of a shared facility for the HLT farms of the four LHC was reviewed in 2017

Advantages:

- re-purpose of servers when no data taking
- shared cost of infrastructure and operations



[Disclaimer: hypothetical location and connections of a shared data-centre]

Challenges

- Very high data rates from detectors to HLTs:
  - Run 3:
    - ALICE: 4Tbps
    - LHCb: 40Tbps
  - Run4:
    - ATLAS: 40Tbps
    - CMS: 40Tbps

- Re-purposing of servers among different management domains (IT, Experiments)

Setup

Locations

LHCb Experiment

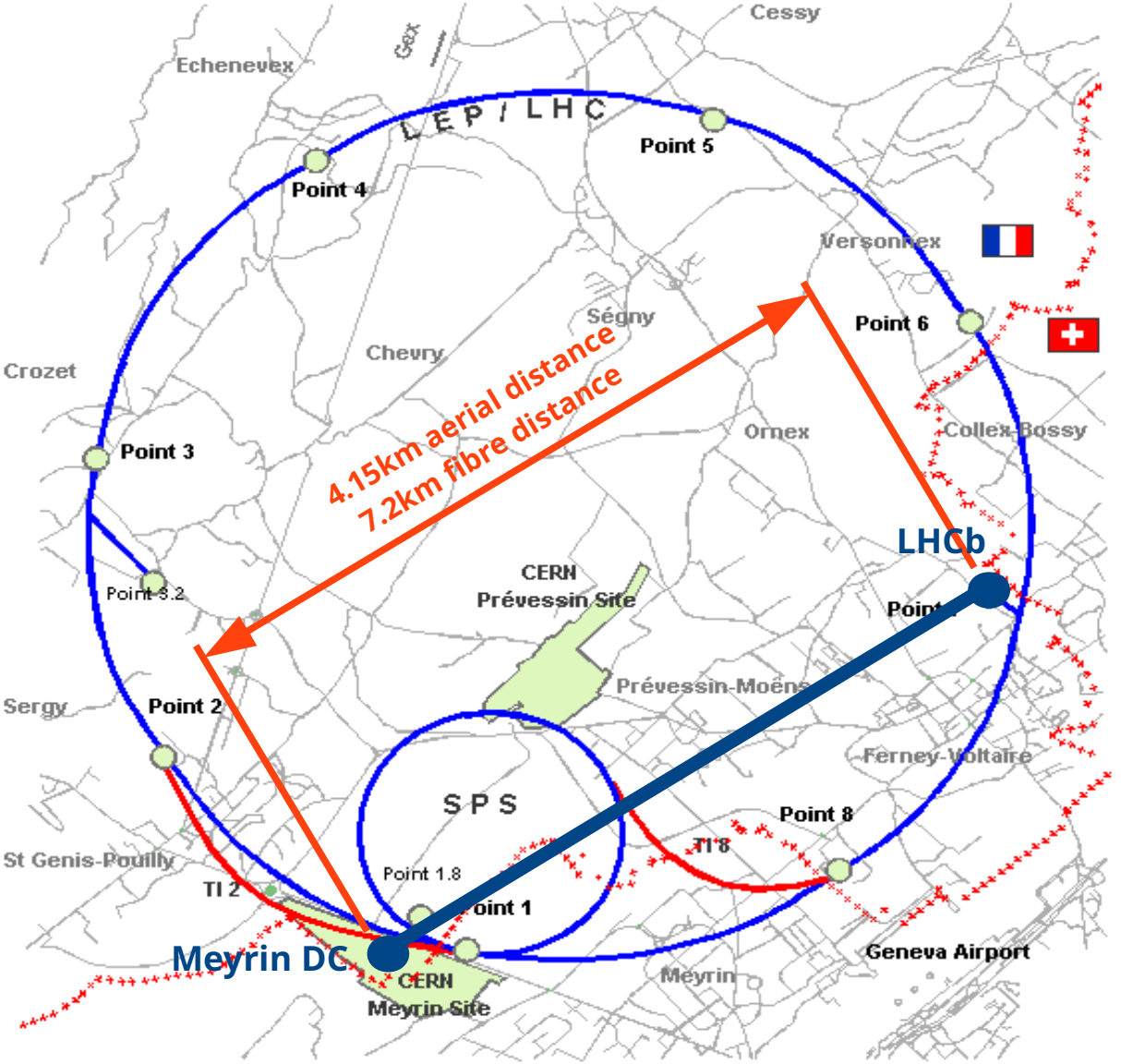
Point 8 in Ferney-Voltaire

- the data source

CERN IT datacentre

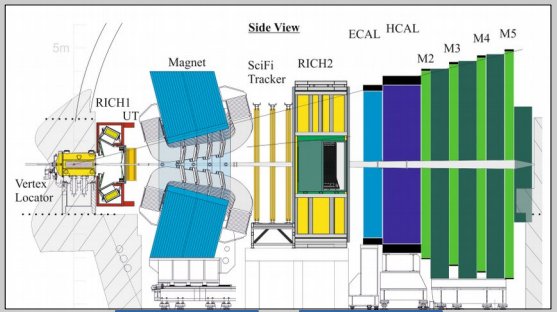
Building 513 in Meyrin

- home of the HLT servers



Setup

LHCb detector - Point8



FORCE10 FORCE10

20x 10G-SR

20x 10G-SR

DATA-VLANs

Juniper QFX5110-48S

4x 100G-LR4

DATA + MGMT VLANs

MGMT VLANs

LHCb online system

IT datacentre - B513

Juniper QFX5200-32C

MGMT 1x 40G

DATA 12x 40G

Rack N040

Ruckus ICX 7750

nb-n040-01

nb-n040-02

nb-n040-12

MGMT 12x 10G

VMGMT 1x 40G

DATA 12x 40G

Rack N044

Ruckus ICX 7750

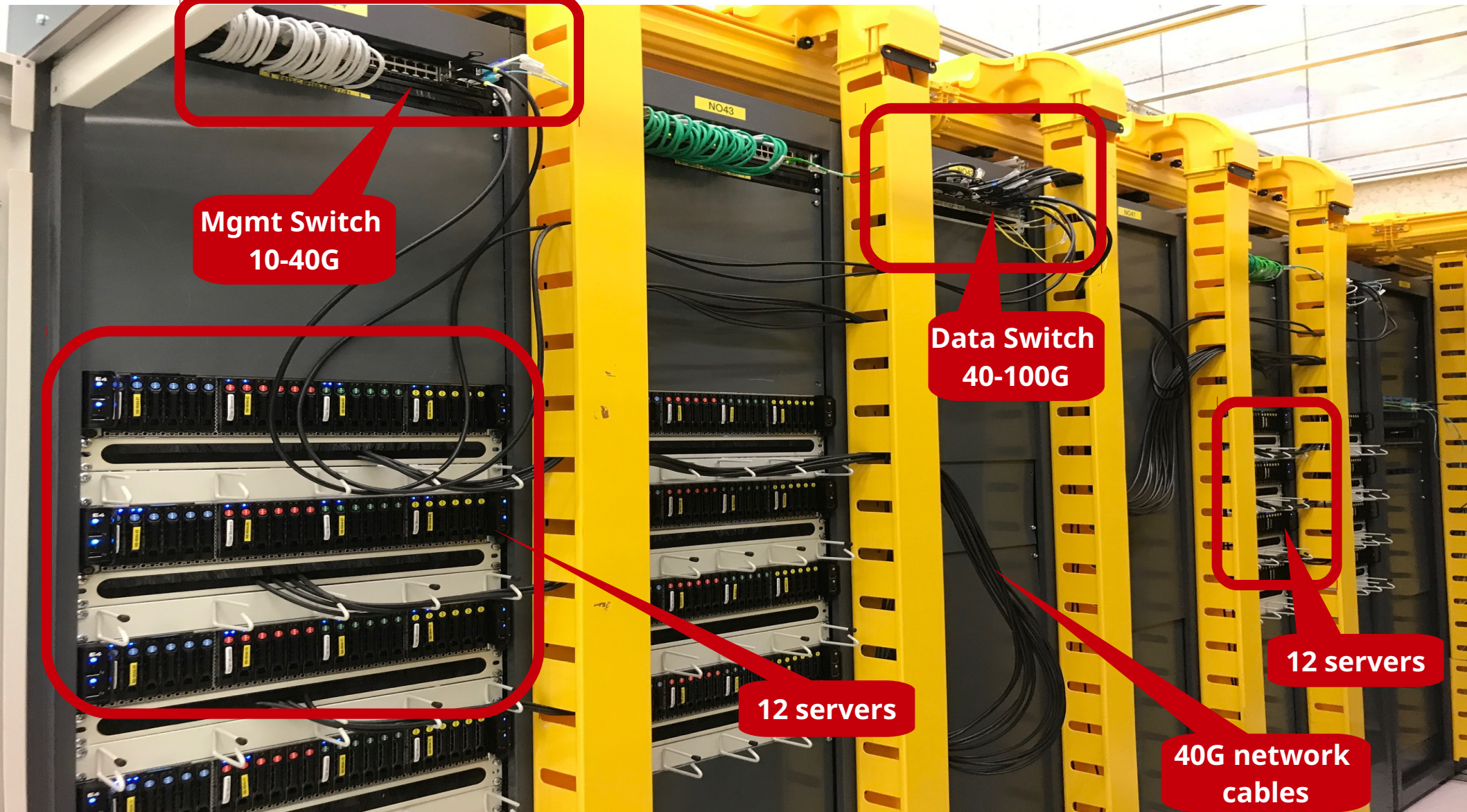
nb-n044-01

nb-n044-02

nb-n044-12

MGMT 12x 10G

Equipment in B513



Results

Server automation

Servers originally part of the CERN Openstack cloud

Handed over to LHCb Online System by redirecting PXE boot requests to it [(1)]

The LHCb Online System could configure the servers as Offline Simulation nodes or HLT nodes, according to the LHC status in order to minimize idle time [(2)]

[(1)] *Redirection of PXE boot packets required some manual cabling and switch configuration*

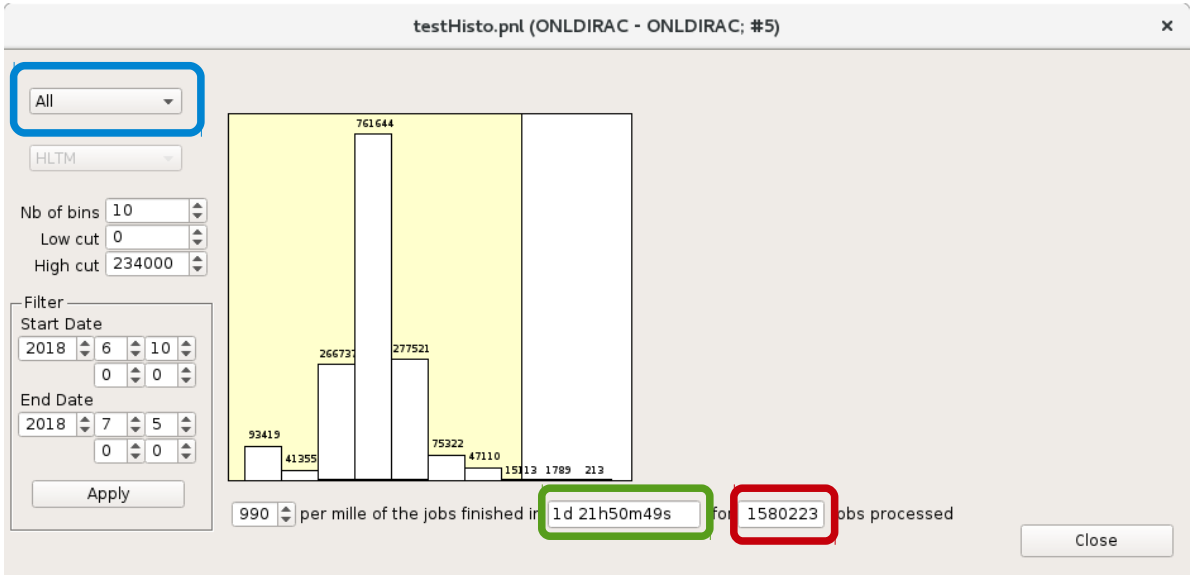
[(2)] *In a production configuration switching the servers between the IT cloud and the Experiment cloud according to need would also be possible*

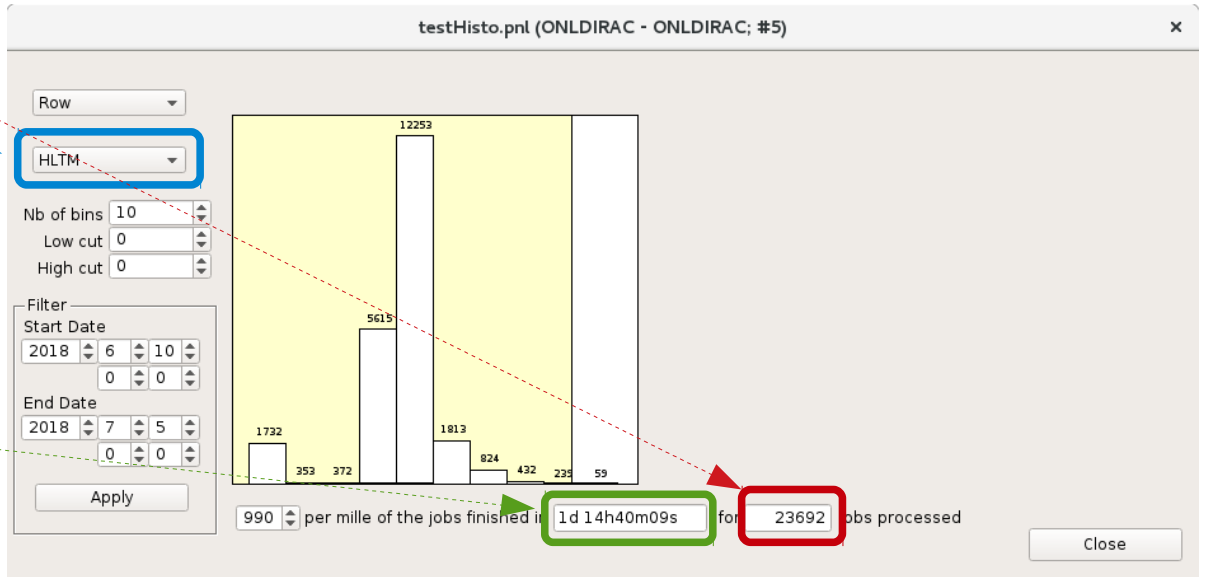
Offline Simulation

In the first period of the test the nodes were used for off-line Montecarlo simulation

In the period, the nodes in B513 [HLTM] ran 23692 jobs, 1.5% of the jobs run in the whole LHCb farm

The jobs in the B513 farm were executed a bit faster than the whole average





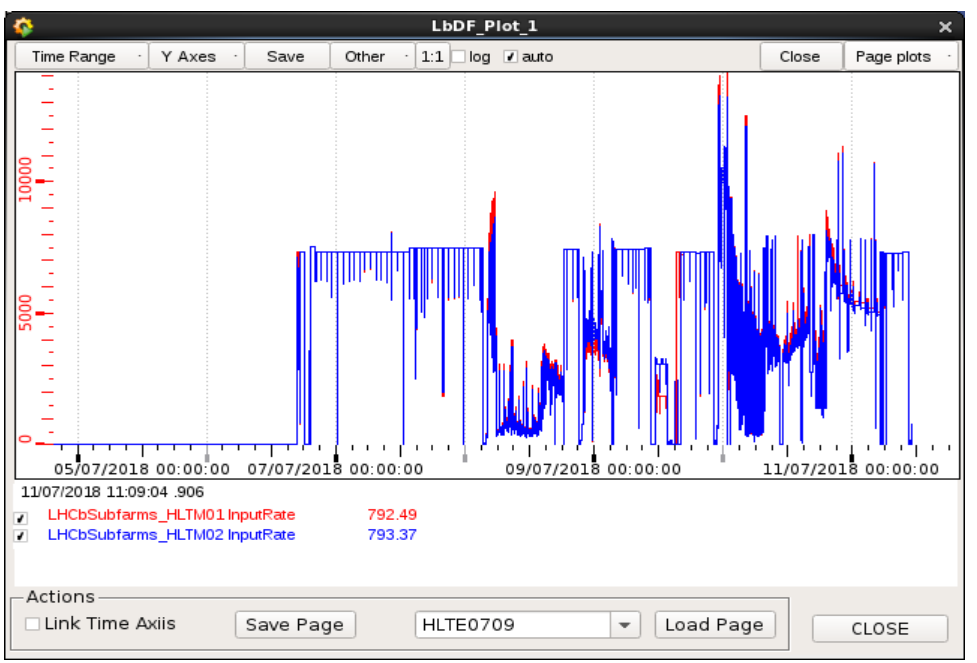
High Level Trigger in Meyrin (HLTM)

LHC has delivered physics only on the 10th of July, just in time for this presentation :-)

LHCb online commented: “Operationally and performance-wise, it seems to be the same as running a sub-farm at Point 8” [LHCb pit]

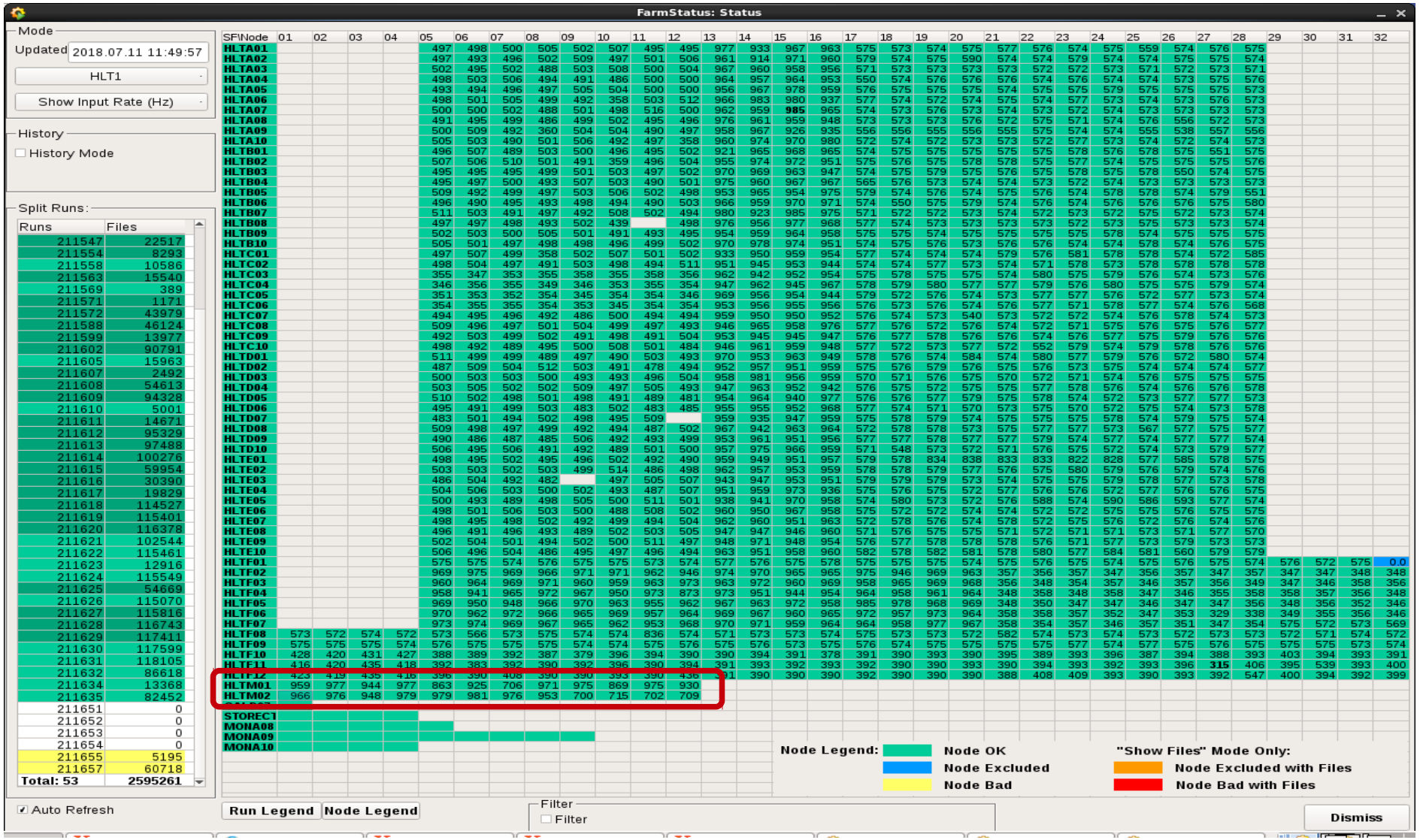
Whole LHCb HLT farm

HLTM01 and HLTM02



events/second

LHCb HLT input rate node by node



DWDM transmission test

DWDM Transmission

LHCb may need to send up to 40Tbps to the remote HLT, but there are few available fibres between LHCb Point8 and the IT data-centre in Building 513

For such high data rates, the cost of traditional DWDM equipment can easily exceed the cost of digging few km of trenches for new fibres.

IT-CS wanted to explore a new cost effective DWDM solution based on the emerging PAM-4 (Pulse Amplitude Modulation technology

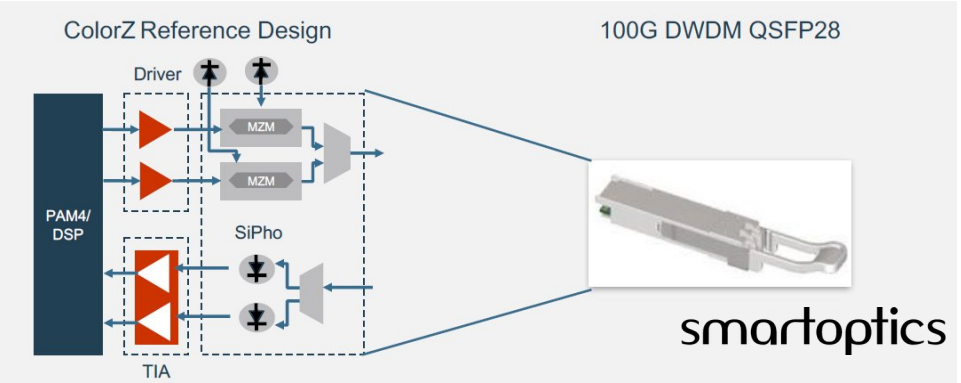
PAM4 DWDM QSF28 transceiver

100G DWDM Transceiver

- 2 wavelengths (lanes) on a 50GHz grid
- Transceiver output power -11 dBm
- Minimum required input power -2 dBm
- Dispersion tolerance +/-6 km on G.652 fiber
- High OSNR required (>31 dB)

Requires an active line system to address these parameters

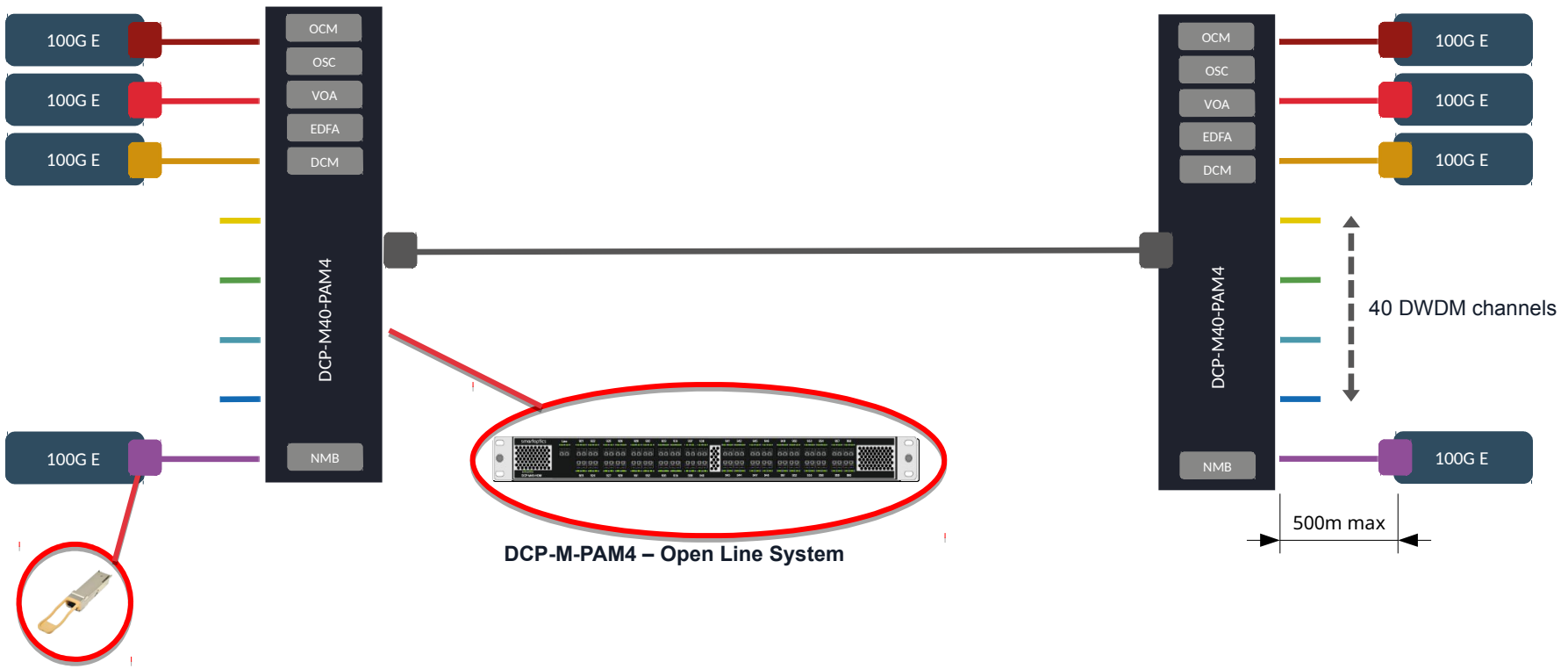
Can be plugged directly into standard switches and NICs



[Credits: Smartoptics]

Smartoptics PAM4 DCI application

100G p-t-p embedded over <80km distance based on a **cost effective solution**



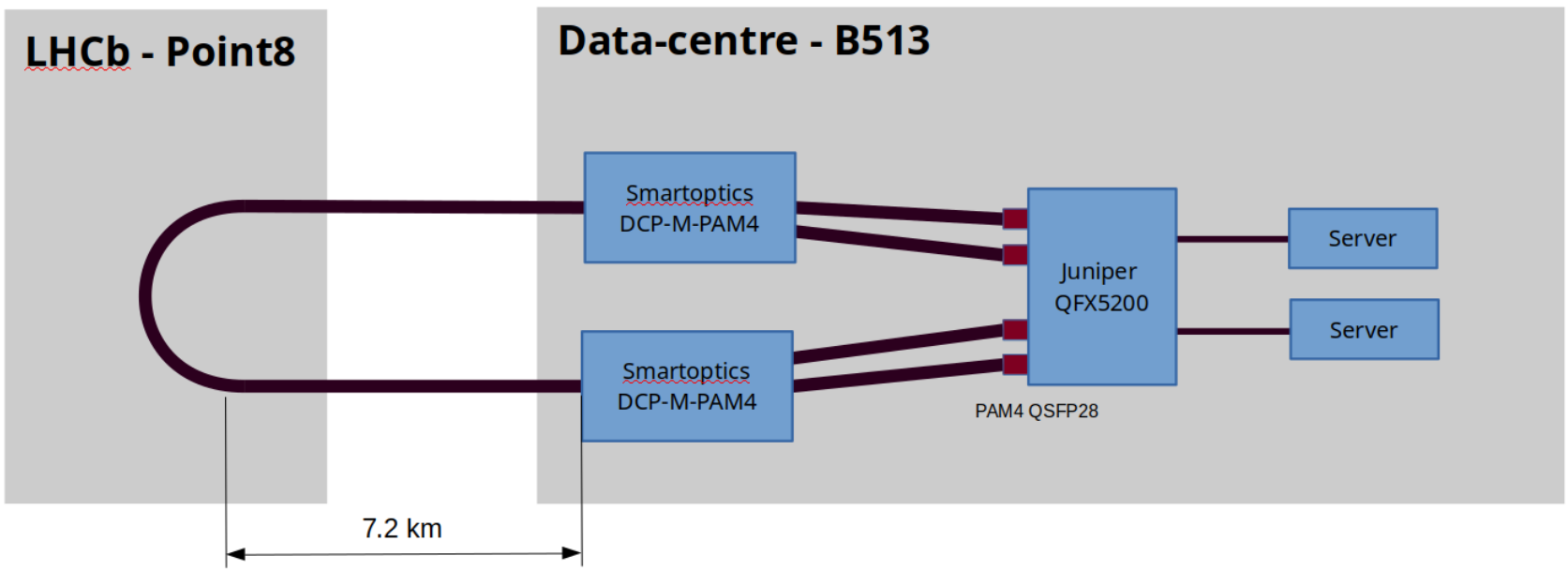
[Credits: Smartoptics]

Setup

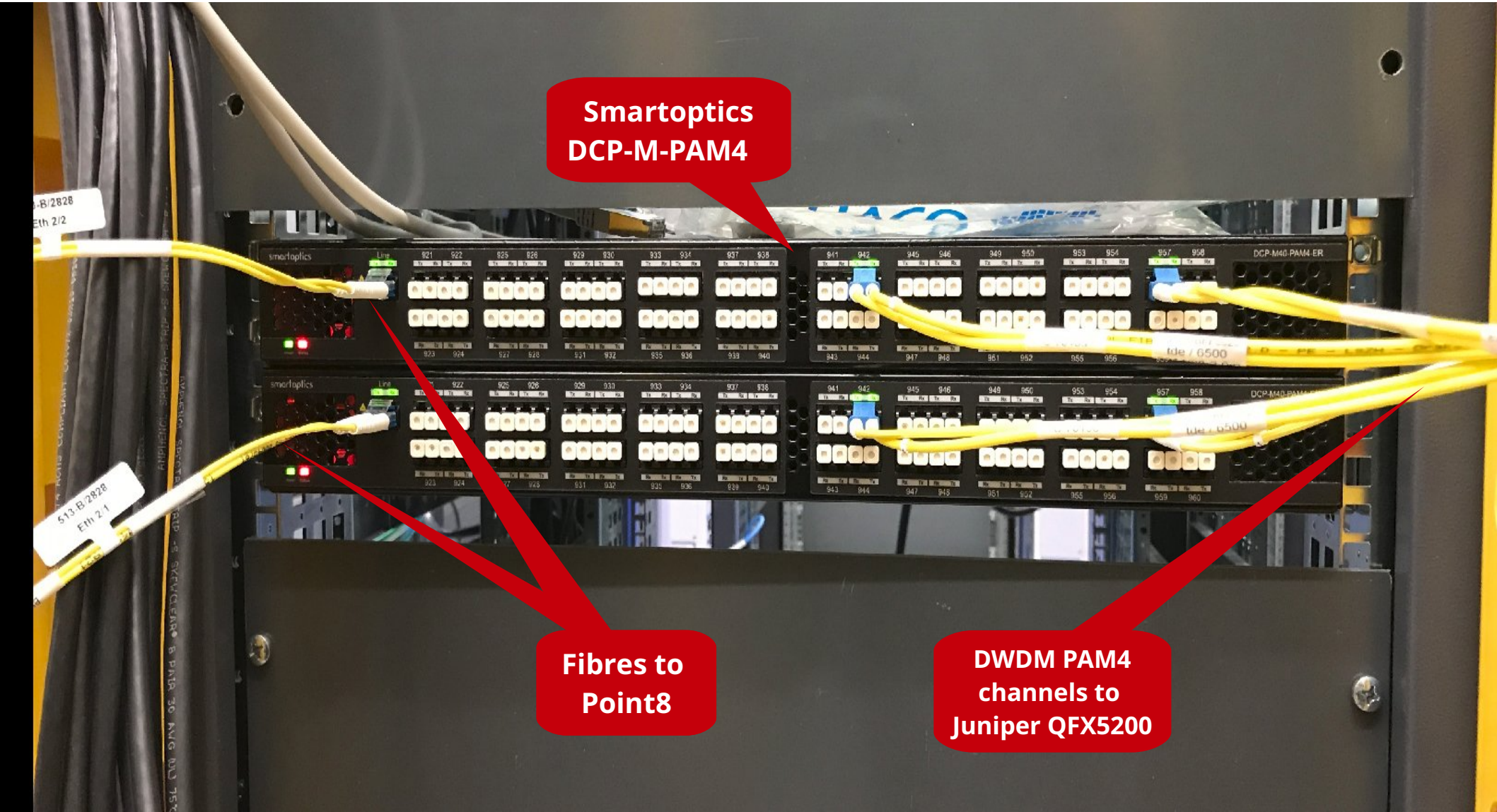
Equipment:

- 2x Smartoptics DCP-M-PAM4 active mux/demux
- 4x Smartoptics PAM4 QSFP28 100Gbps DWDM
- 1x Juniper QFX5200-32C
- 1x Juniper QFX5110-48C

The QFX5110 in Point8 was not able to light up the PAM4 QSFP28 because didn't support high power optics; thus the two pairs of fibres were looped back in Point8



Open-line system



Findings

Smartoptics PAM4 optics worked well on Juniper QFX5200

On the QFX5110 the optics were recognized, but lasers couldn’t be turned on. Support for high power consumption optics is mandatory (similar case to ER4)

The PAM4 optics are not tuneable

PAM4 lines were loaded with sustained 150Gbps traffic for several hours with no transmission errors

Conclusions

Conclusions

A common computing facility for a shared Software High Level Trigger has been proved feasible. “Operationally and performance-wise, it seems to be the same as running a sub-farm at Point 8”

DWDM PAM4 is an affordable and ready to use technology for Tbps transmission lines

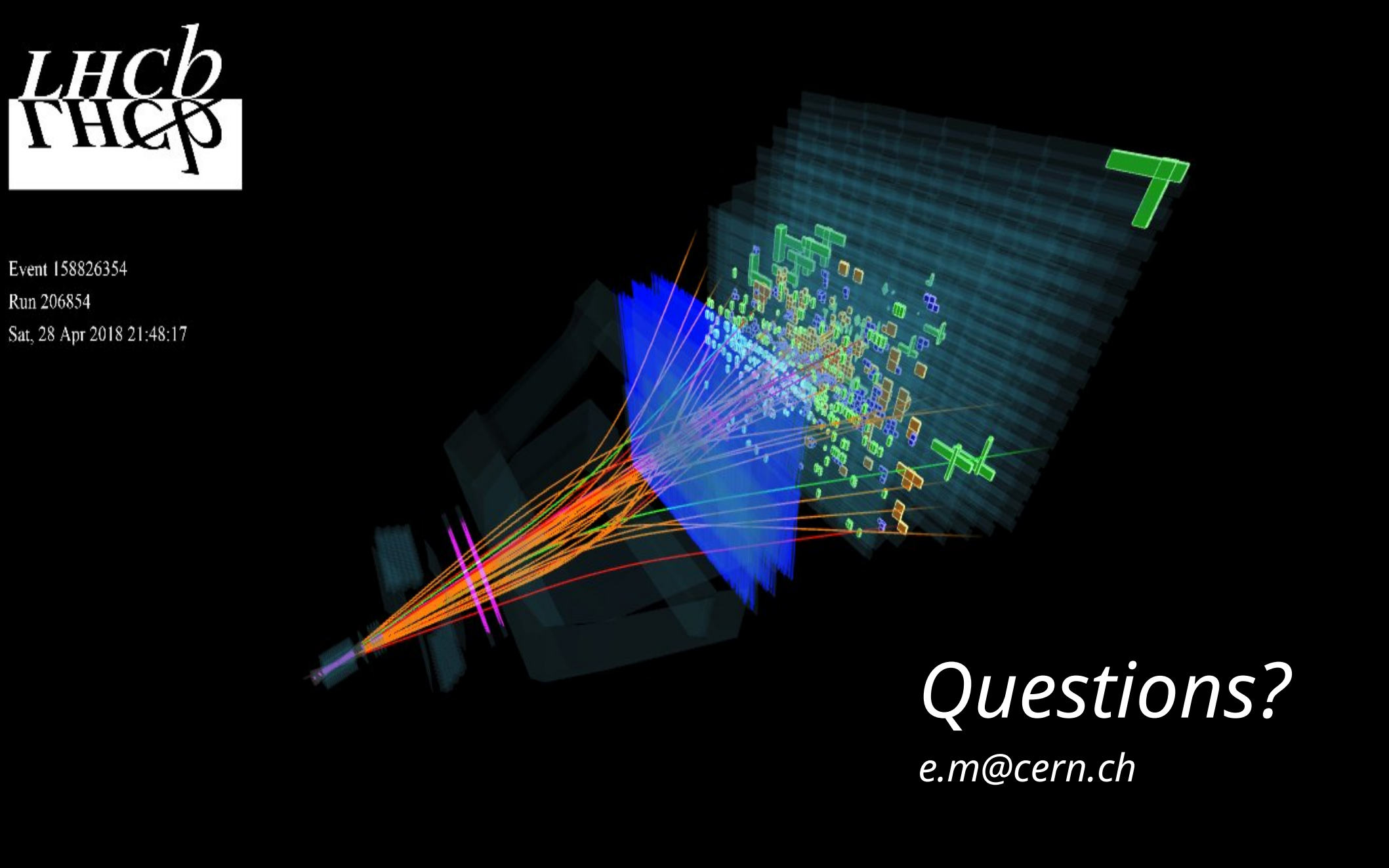
Credits

LHCb

- Loïc Brarda
- Luis Granado Cardoso
- Niko Neufeld
- Tommaso Colombo

CERN IT-CS

- Edoardo Martelli
- Marc Collignon
- Tony Cass

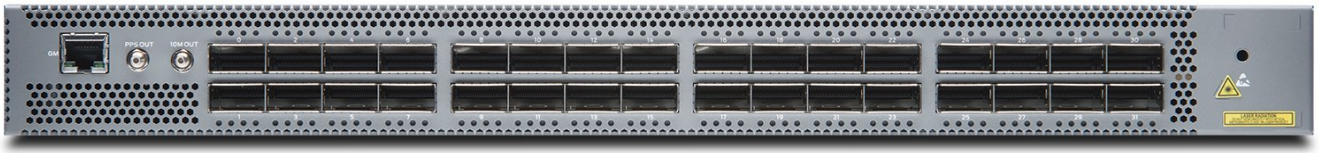

LHCb
Event 158826354
Run 206854
Sat, 28 Apr 2018 21:48:17
Questions?
e.m@cern.ch

Additional information

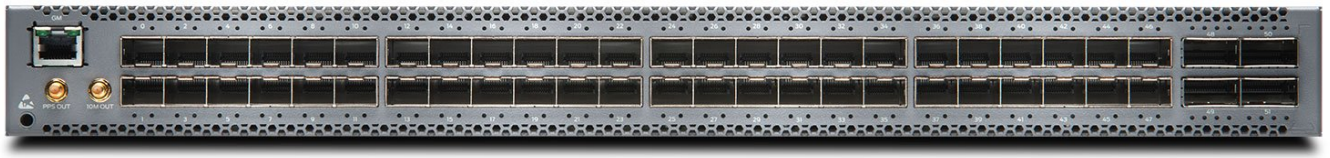
Network Hardware

Connections between two sites:

- Juniper QFX5200-32C used for 4x100G and 24x 40G



- Juniper QFX5110-48S used for 4x100G and 40x 10G



Servers in B513 data-centre:

- 24x servers, 10G NIC for control, 40G NIC for data
- 2x Ruckus ICX7750 with 48x 10GbaseT and 6x 40G

Juniper QFX commands

To run the PAM4 QSF28 it is necessary JUNOS 17.3 or later

These commands were necessary to turn on the lasers of the PAM4 QSF28 optics

```
set interfaces et-0/0/51 gigether-options fec none
set interfaces et-0/0/51 otn-options laser-enable
set interfaces et-0/0/51 otn-options is-ma
set interfaces et-0/0/51 otn-options fec none
```

Show commands:

```
user@n513-c-xjukh-1# run show chassis hardware
    [...]
    Xcvr 29      REV 01   740-061409   1GCQA234062       QSFP-100GBASE-LR4
    Xcvr 30               NON-JNPR     L173800805        QSFP-100GE-DWDM2
    Xcvr 31               NON-JNPR     L173700234        QSFP-100GE-DWDM2

user@n513-c-xjukh-1# run show interfaces diagnostics optics et-0/0/31
Physical interface: et-0/0/31
    [...]
  Lane 0
    Laser bias current                          :  60.673 mA
    Laser output power                          :  0.111 mW / -9.56 dBm
    Laser receiver power                        :  1.739 mW / 2.40 dBm
    [...]
  Lane 1
    Laser bias current                          :  78.699 mA
    Laser output power                          :  0.111 mW / -9.55 dBm
    Laser receiver power                        :  1.595 mW / 2.03 dBm
    [...]
```