

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.

LHCb Upgrade Trigger Diagram

**30 MHz inelastic event rate
(full rate event building)**

Software High Level Trigger

Full event reconstruction, inclusive and exclusive kinematic/geometric selections

Buffer events to disk, perform online detector calibration and alignment

Add offline precision particle identification and track quality information to selections
Output full event information for inclusive triggers, trigger candidates and related primary vertices for exclusive triggers

2-5 GB/s to storage

[Credits: Umberto Marconi – INFN Bologna – LHCb]

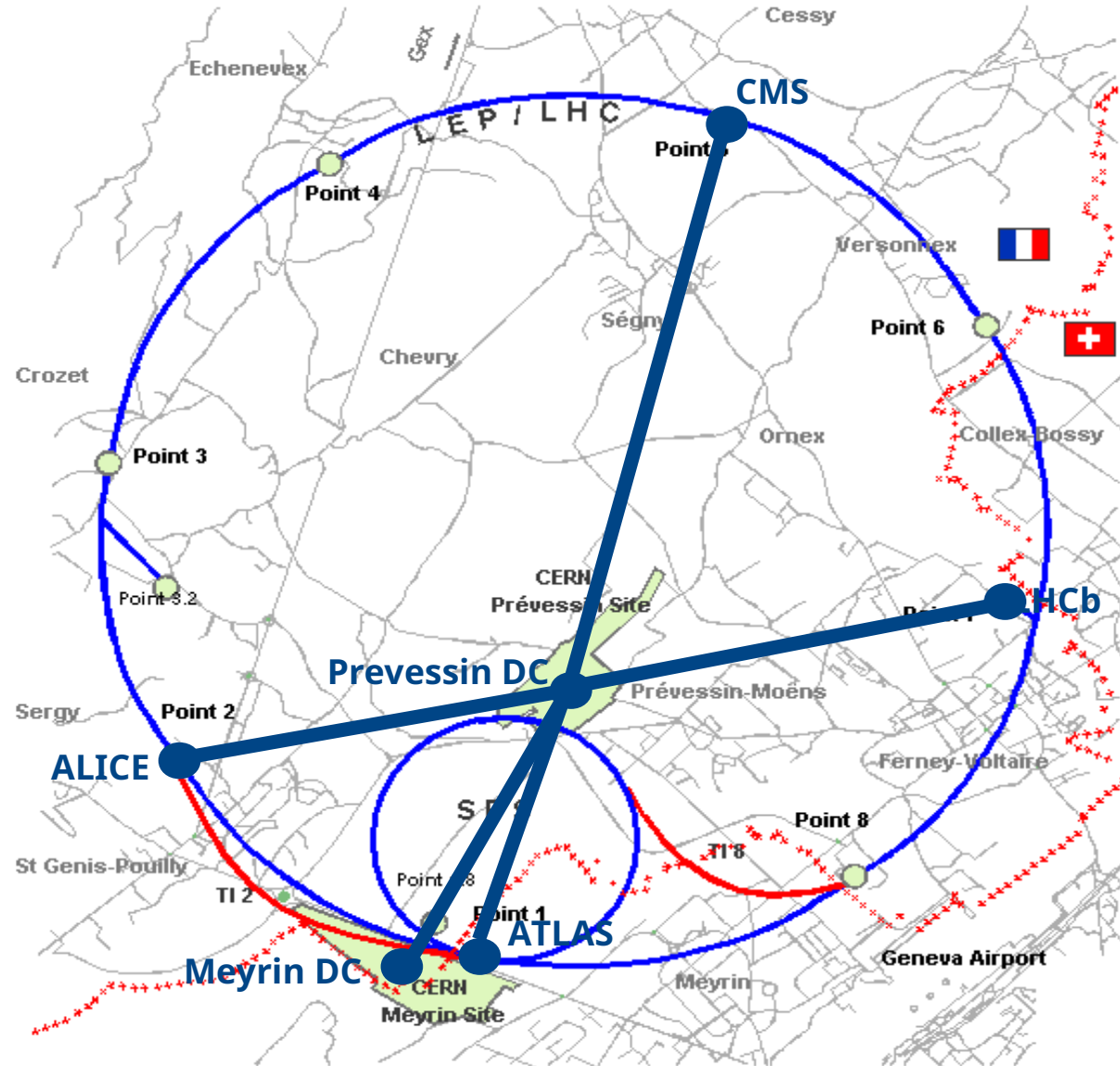
[Image credits: Concezio Bozzi – INFN Ferrara – LHCb]

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

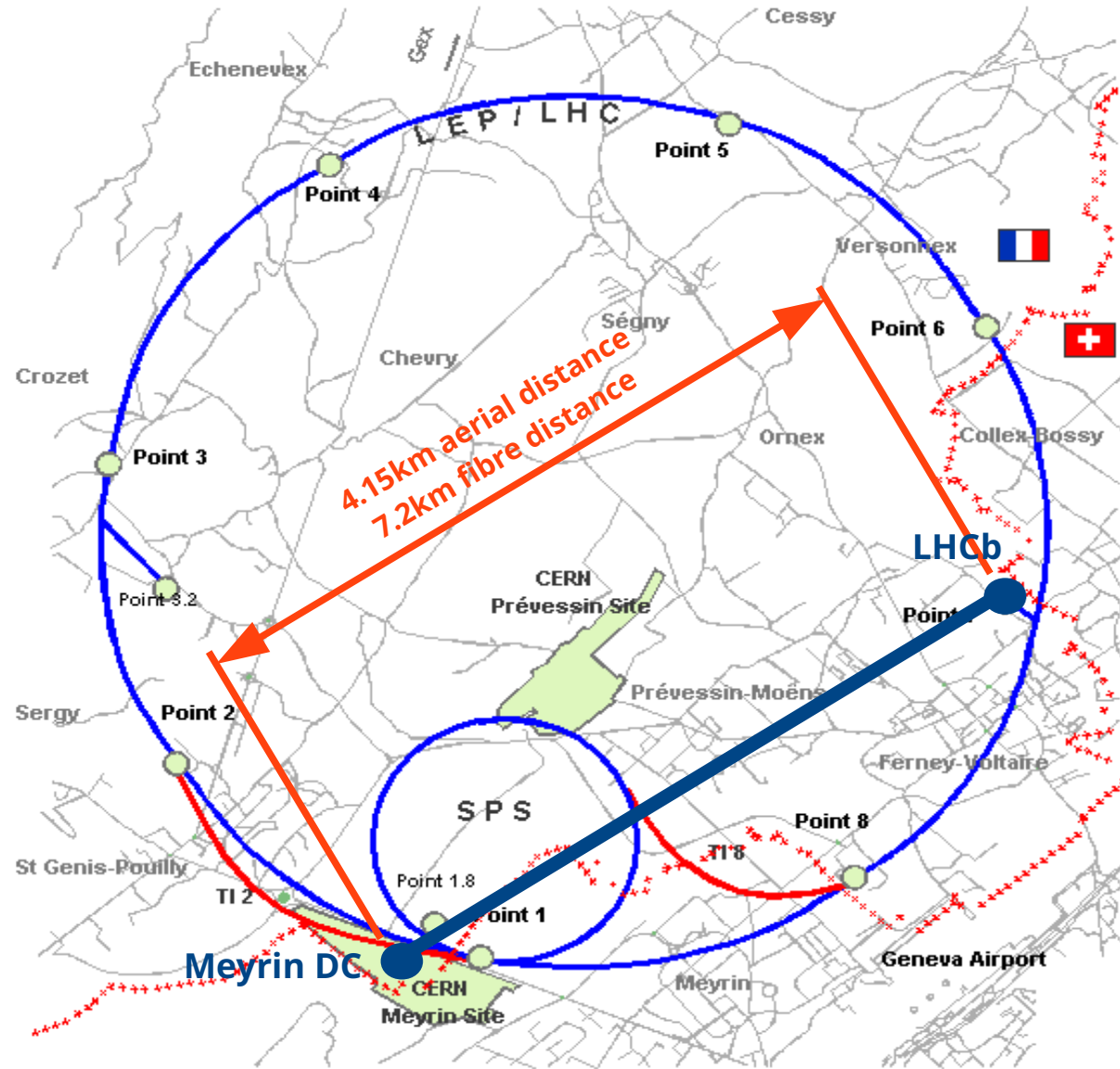
Locations

LHCb Experiment

Point 8 in Ferney-Voltaire
- the data source

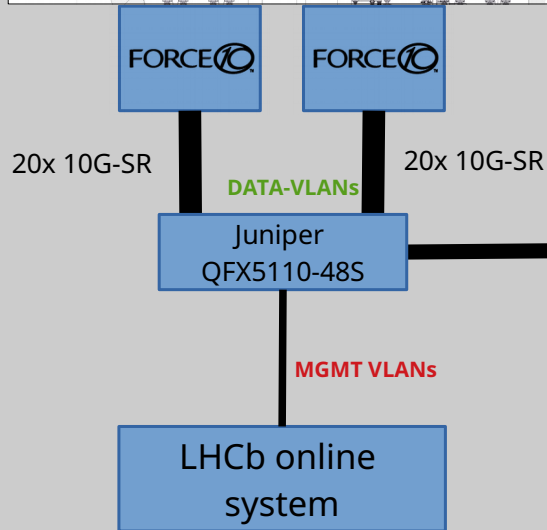
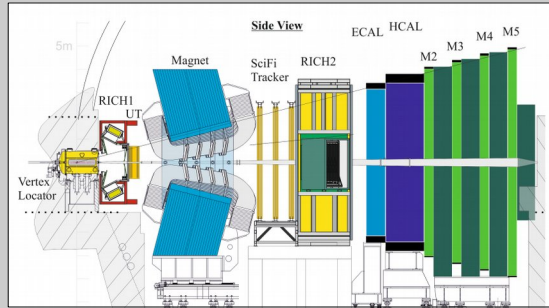
CERN IT datacentre

Building 513 in Meyrin
- home of the HLT servers



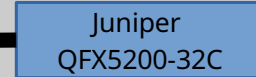
Setup

LHCb detector - Point8

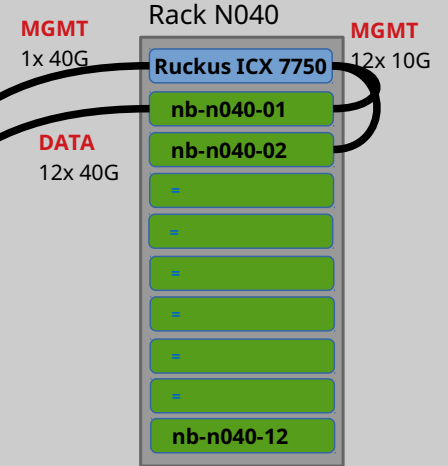
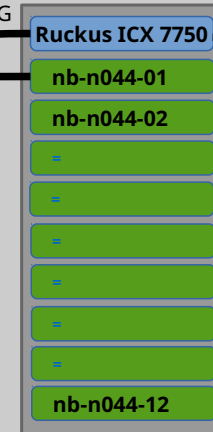


IT datacentre - B513

4x 100G-LR4
DATA + MGMT VLANs



VMGMT 1x 40G
DATA 12x 40G



Equipment in B513

**Mgmt Switch
10-40G**

**Data Switch
40-100G**

12 servers

12 servers

**40G network
cables**

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

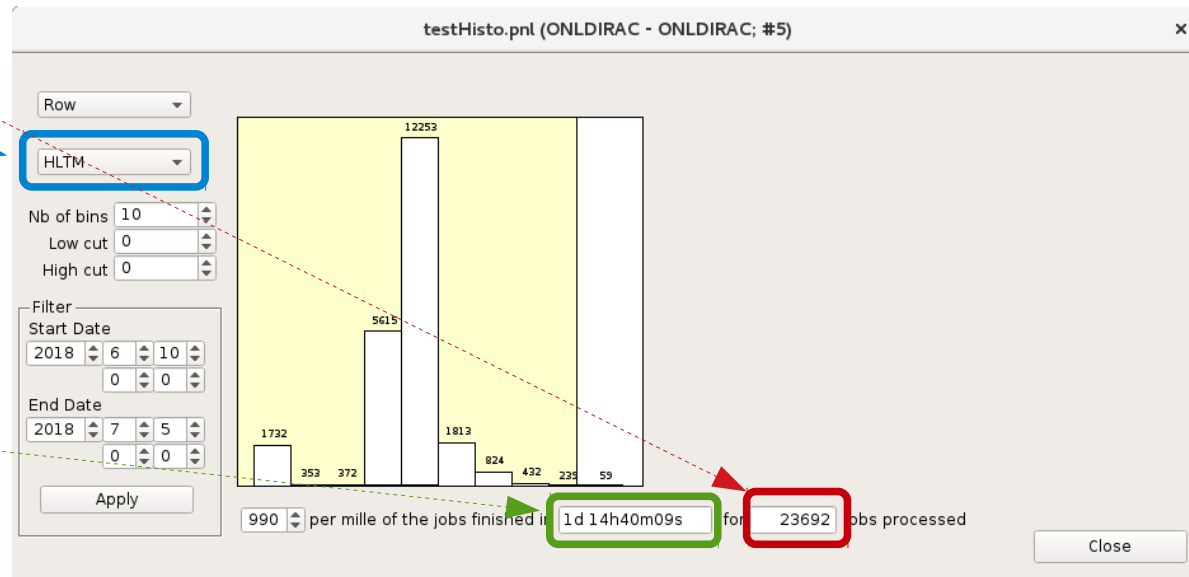
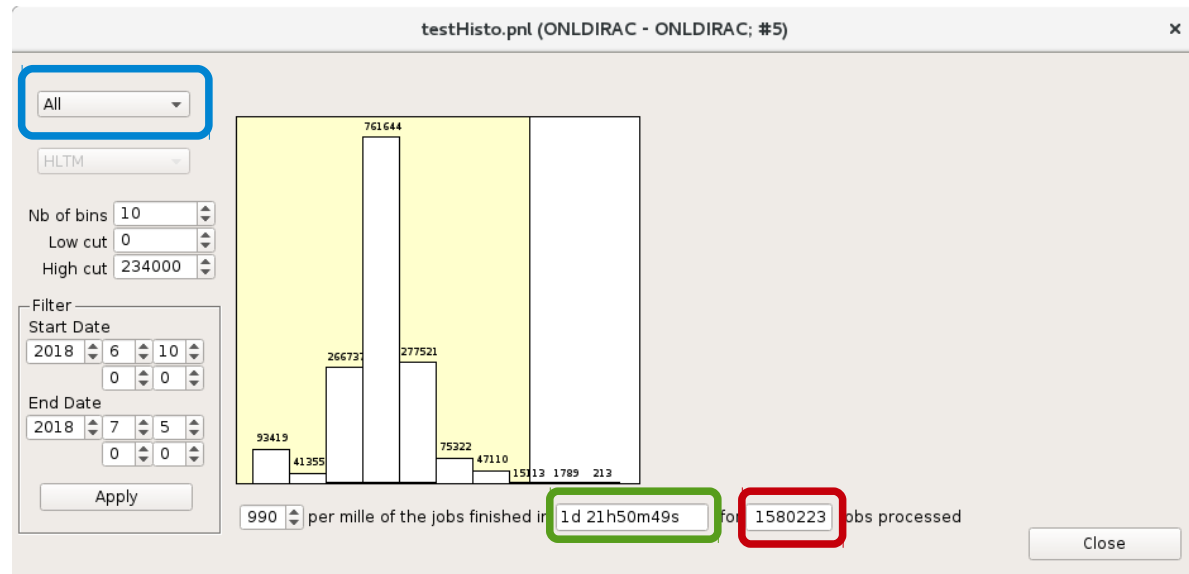
(2) In a production configuration switching the servers between the IT cloud and the Experiment cloud according to need would also be possible

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

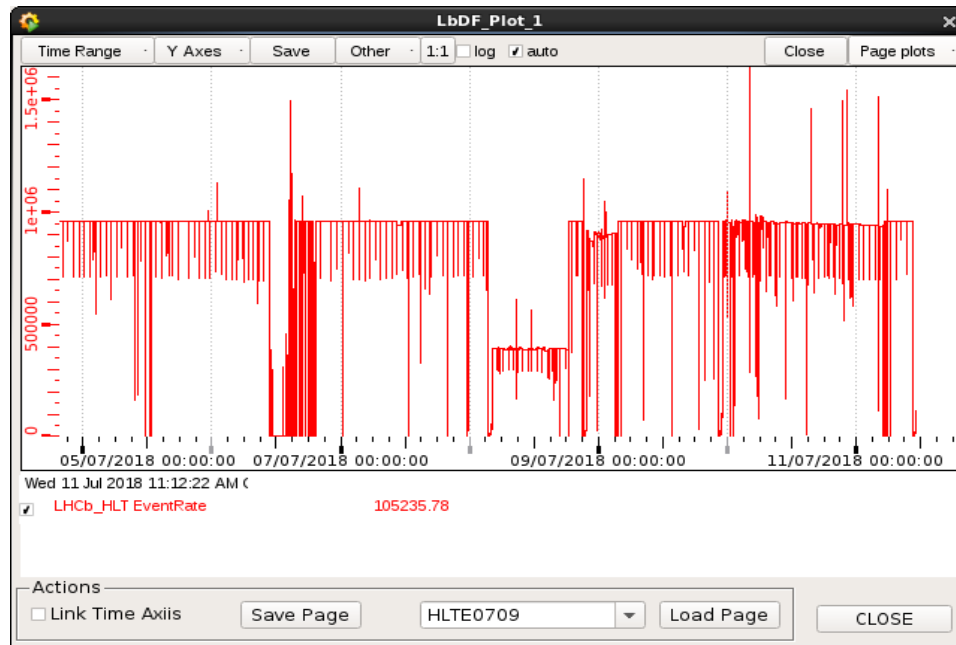


High Level Trigger in Meyrin (HLTM)

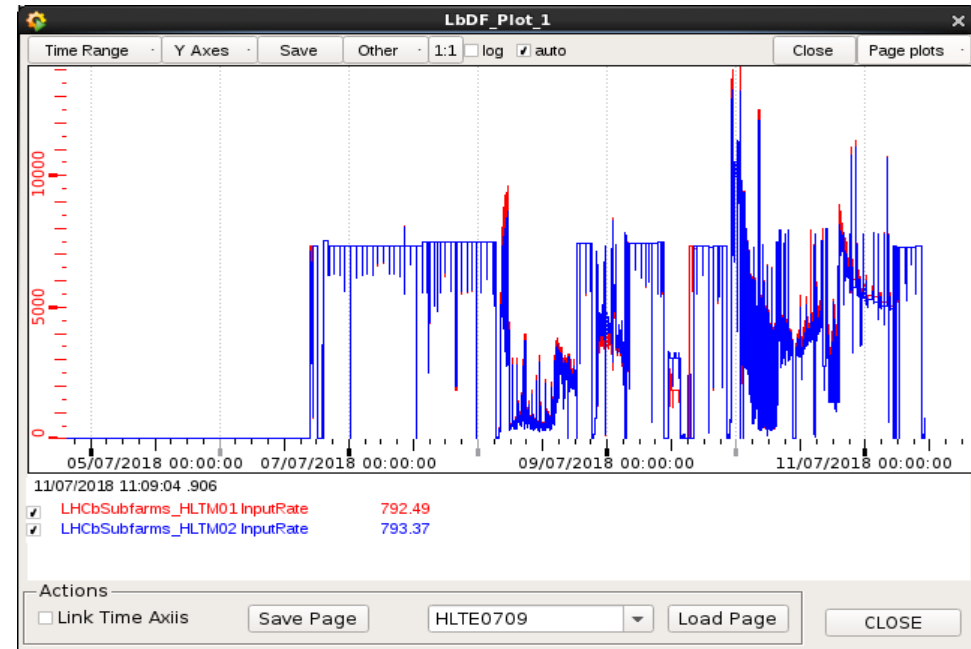
LHC has delivered physics only on the 10th of July, just in time for this presentation :-)

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



LHCb HLT input rate node by node

Mode
Updated: 2018.07.11 11:49:57
HLT1
Show Input Rate (Hz)

History
 History Mode

Split Runs:
Runs: 211547, 211554, 211558, 211563, 211569, 211571, 211572, 211588, 211599, 211602, 211605, 211607, 211608, 211609, 211610, 211611, 211612, 211613, 211614, 211615, 211616, 211617, 211618, 211619, 211620, 211621, 211622, 211623, 211624, 211625, 211626, 211627, 211628, 211629, 211630, 211631, 211632, 211634, 211635
Files: 22517, 8293, 10586, 15540, 389, 1171, 43979, 46124, 13977, 90791, 15963, 2492, 54613, 94328, 5001, 14671, 95329, 97488, 100276, 59954, 30390, 19829, 114527, 115401, 116378, 102544, 115461, 12916, 115549, 54669, 115070, 115816, 116743, 117411, 117599, 118105, 86618, 13368, 82452
Total: 53 2595261

FarmStatus: Status		01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	
HLTA01	497	498	500	505	502	507	495	495	977	933	967	963	575	573	574	575	577	576	574	575	559	574	576	575	576	575	576	575	576	575	576	575		
HLTA02	497	493	496	502	509	497	501	506	961	914	971	960	579	574	575	590	574	574	579	574	574	574	574	574	574	574	574	574	574	574	574	574		
HLTA03	502	495	502	488	503	508	500	504	967	960	958	956	571	573	573	573	573	572	572	573	573	572	572	573	573	571	572	573	575	571	571	571		
HLTA04	498	503	496	494	491	486	500	500	964	957	964	963	550	574	576	576	576	576	574	576	574	576	574	576	574	574	574	574	574	574	574	574		
HLTA05	493	494	496	500	506	505	505	505	967	978	959	978	575	575	575	575	575	575	575	575	575	575	575	575	575	575	575	575	575	575	575	575		
HLTA06	498	501	505	499	492	358	503	512	966	983	980	937	577	574	572	574	575	574	575	574	575	574	572	574	573	574	573	573	573	573	573	573		
HLTA07	500	500	502	488	501	498	516	500	962	959	985	965	574	573	576	573	574	573	574	573	572	574	573	572	574	573	573	573	573	573	573	573		
HLTA08	491	495	499	486	499	502	495	496	976	961	959	948	573	573	573	576	572	575	571	574	575	574	572	574	576	576	576	572	573	573	573	573		
HLTA09	500	509	492	360	504	504	490	497	958	967	926	935	556	556	556	556	556	556	556	556	556	556	556	556	556	556	556	556	556	556	556	556		
HLTA10	506	503	490	501	506	492	487	358	960	974	970	980	572	574	572	573	573	572	573	573	572	573	573	572	573	574	572	574	572	574	573	573		
HLTB01	496	507	489	503	500	496	495	502	921	965	968	965	574	575	575	575	575	575	575	575	575	575	575	575	575	575	575	575	575	575	575	575		
HLTB02	507	506	510	501	491	359	496	504	955	974	972	951	575	576	575	578	578	578	575	577	574	575	575	575	575	575	575	575	575	575	575	575	575	
HLTB03	495	495	495	499	501	503	497	502	970	969	963	947	574	575	575	579	575	576	575	576	575	576	575	576	575	578	575	578	550	574	575	575		
HLTB04	495	497	500	493	507	503	490	501	975	960	967	967	565	576	573	574	574	574	573	572	574	573	572	574	573	572	574	573	573	573	573	573		
HLTB05	509	492	499	497	503	506	502	498	953	965	954	967	575	579	574	576	574	575	576	574	573	574	574	574	574	574	574	574	574	574	574	574	574	
HLTB06	496	490	495	493	498	494	490	503	966	959	970	971	574	550	573	575	579	574	576	574	576	574	576	576	576	576	576	576	576	576	576	576	576	
HLTB07	511	503	491	497	492	508	502	494	980	923	985	975	571	572	572	573	574	572	573	573	574	572	573	572	573	572	575	572	573	574	574	574		
HLTB08	497	497	498	493	502	439	498	497	956	977	968	977	574	573	573	573	573	573	572	573	573	573	573	573	573	573	573	573	573	573	573	573	573	
HLTB09	502	503	500	505	501	491	493	495	954	959	964	958	575	575	574	575	575	575	575	575	575	575	575	575	574	578	574	575	575	575	575	575		
HLTB10	505	501	497	498	496	499	502	978	974	978	974	951	574	575	576	573	576	576	574	578	574	578	574	578	574	578	574	576	576	576	576	576		
HLTC01	497	507	499	358	502	507	501	502	933	950	959	954	577	574	574	574	574	579	576	581	578	578	578	578	578	578	578	578	578	578	578	578	578	
HLTC02	498	504	497	491	503	498	494	511	951	945	953	944	574	574	574	577	573	574	571	578	573	578	578	578	578	578	578	578	578	578	578	578	578	
HLTC03	356	347	353	355	358	355	358	356	962	942	952	954	575	578	575	576	574	580	575	579	576	574	576	576	576	576	576	576	576	576	576	576	576	
HLTC04	346	356	356	354	346	353	355	354	967	964	967	964	578	578	578	578	577	579	576	580	575	577	579	576	580	575	575	575	575	575	575	575	575	
HLTC05	351	353	352	354	354	354	354	354	966	956	954	944	579	572	576	574	573	577	577	576	572	572	572	572	572	572	572	572	572	572	572	572	572	572
HLTC06	354	355	355	354	353	345	354	354	953	956	955	956	576	576	573	576	574	576	577	571	578	577	574	576	577	574	576	576	576	576	576	576	576	
HLTC07	494	495	496	492	486	500	494	494	959	950	950	952	576	574	573	540	573	572	572	574	576	574	576	574	576	574	576	574	576	574	574	574	574	
HLTC08	509	496	497	501	504	499	497	493	946	965	958	976	577	576	572	576	574	572	571	575	576	575	576	575	576	575	576	576	576	576	576	576	576	
HLTC09	492	503	499	502	491	498	491	498	943	953	945	945	947	576	577	578	576	576	576	574	576	577	575	579	576	579	576	579	576	579	576	576	576	
HLTC10	498	492	489	495	500	508	501	484	946	961	959	948	577	572	573	577	572	552	579	574	579	578	576	576	576	576	576	576	576	576	576	576	576	
HLTD01	511	499	499	489	497	490	503	493	970	953	963	949	578	576	574	584	574	580	577	574	580	577	574	580	577	574	580	577	574	574	574	574	574	
HLTD02	487	509	504	512	503	491	478	494	952	957	951	959	575	576	579	576	575	576	575	576	573	575	574	574	574	574	574	574	574	574	574	574	574	
HLTD03	500	503	503	500	493	493	496	504	958	981	956	959	570	571	576	575	570	572	571	574	576	575	575	575	575	575	575	575	575	575	575	575	575	575
HLTD04	503	505	502	502	509	497	505	493	947	963	952	942	576	575	572	575	575	577	578	576	574	576	574	576	574	576	574	576	576	576	576	576	576	
HLTD05	510	502	498	501	498	491	489	491	954	940	949	977	576	576	577	579	575	577	579	575	578	574	572	573	577	577	577	577	577	577	577	577	577	577
HLTD06	495	491	499	503	483	502	483	485	955	955	952	968	577	574	571	570	573	575	570	572	575	574	572	575	574	572	575	574	573	578	573	578	573	
HLTD07	483	501	484	502	496	495	509	959	935	947	959	959	575	578	578	574	575	575	575	578	574	575	575	575	575	575	575	575	575	575	575	575	575	575
HLTD08	509	498	497	499	492	494	487	502	967	942	963	964	572	578	578	573	575	577	577	573	567	577	573	567	577	573	567	577	573	573	573	573	573	
HLTD09	490	486	487	485	506	492	493	499	953	961	951	956	577	577	578	577	577	579	574	577	574	577	574	577	574	577	574	577	574	574	574	574	574	
HLTD10	506	495	506	491	492	489	501	500	957	975	966	959	571	548	573	572	571	576	575	572	574	573	579	573	579	573	579	573	579	573	579	573	579	
HLTE01	498	495	502	495	496	502	492	490	959	949	951	957	579	578	834	838	833	833	822	828	827	585	578	575	575	575	575	575	575	575	575	575	575	575
HLTE02	503	503	502	503	499	514	486	498	962	957	953	959																						

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

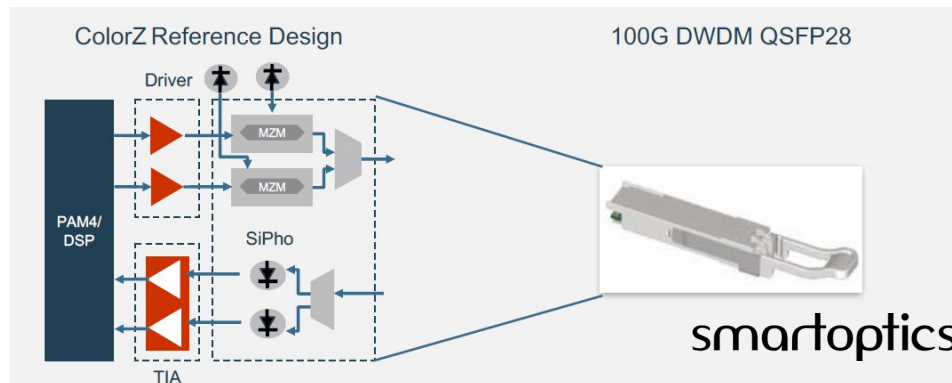
PAM4 DWDM QSF28 transceiver

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

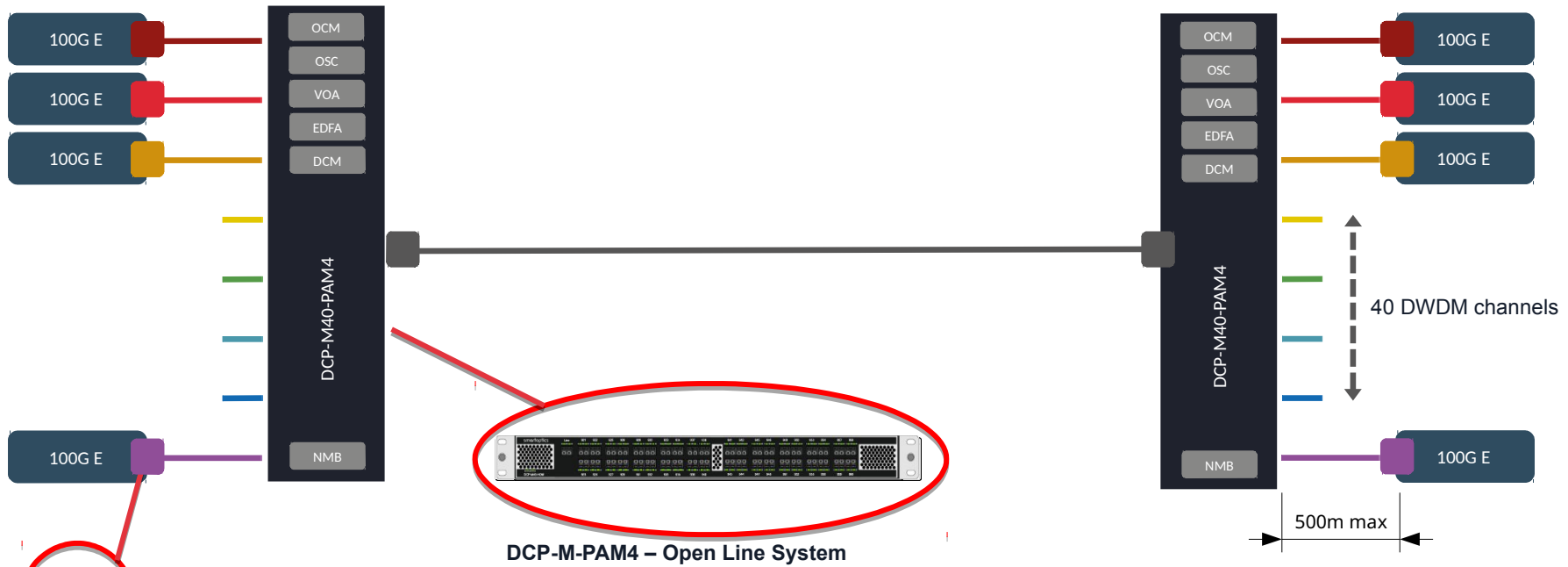
Can be plugged directly into standard switches and NICs



[Credits: Smartoptics]

Smartoptics PAM4 DCI application

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



PAM4 QSFP28 DWDM TRX
Embedded in to 100G switch

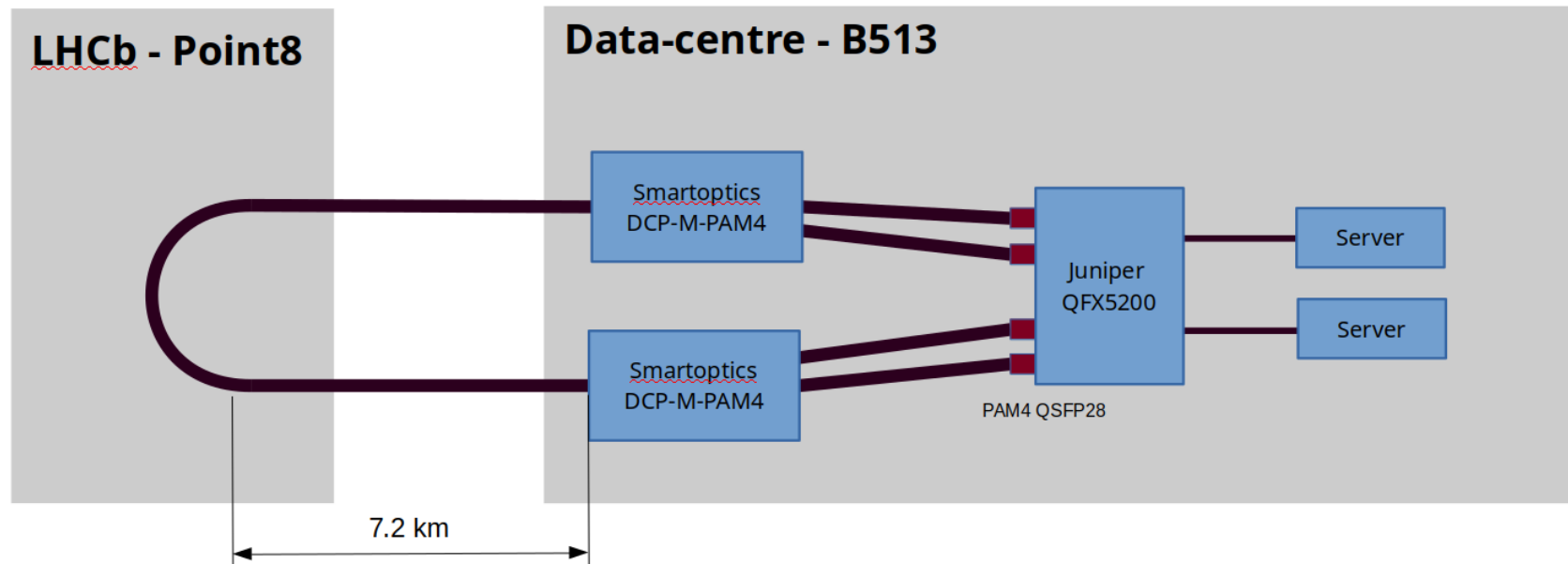
[Credits: Smartoptics]

Setup

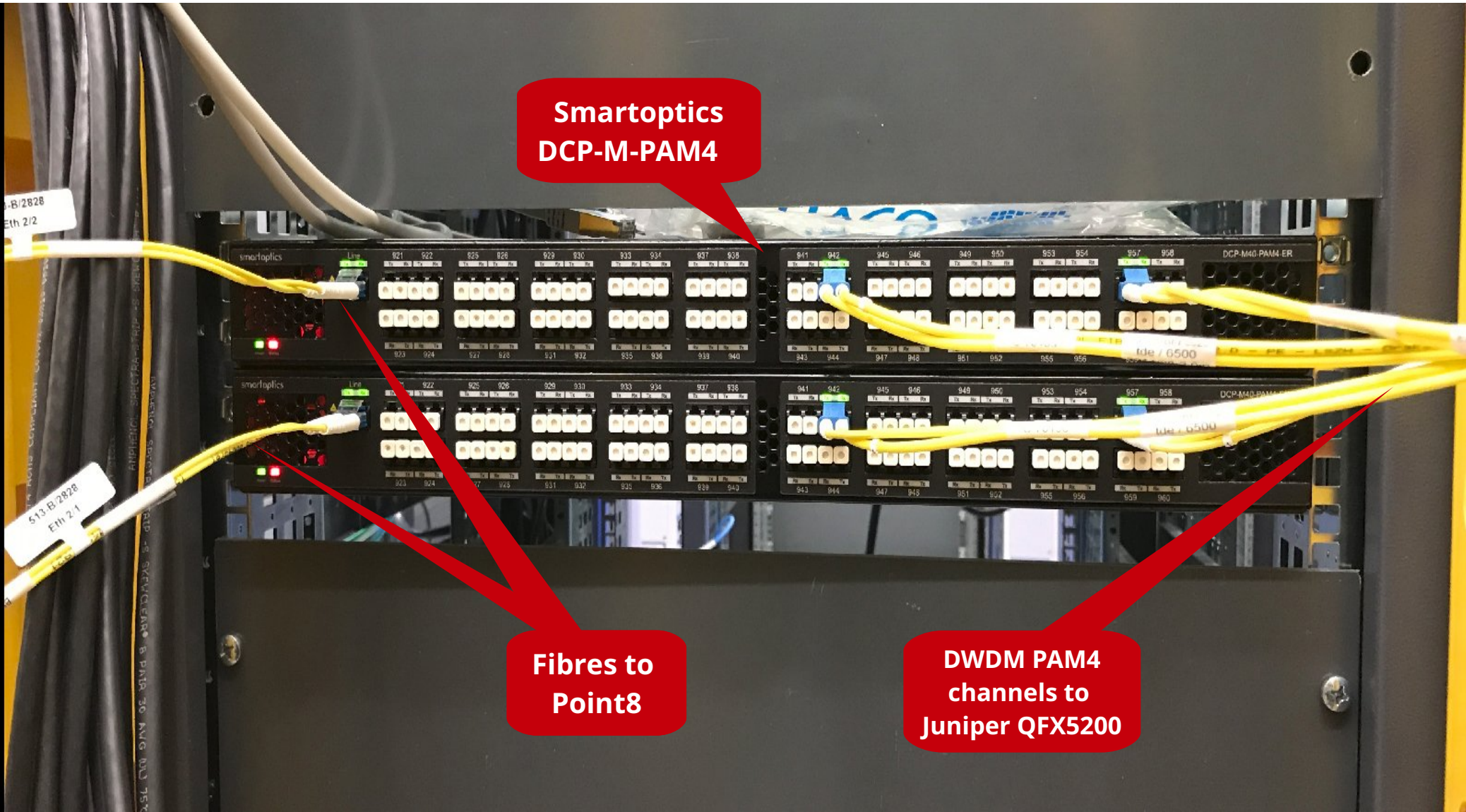
Equipment:

- 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



Smartoptics
DCP-M-PAM4

Fibres to
Point8

DWDM PAM4
channels to
Juniper QFX5200

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

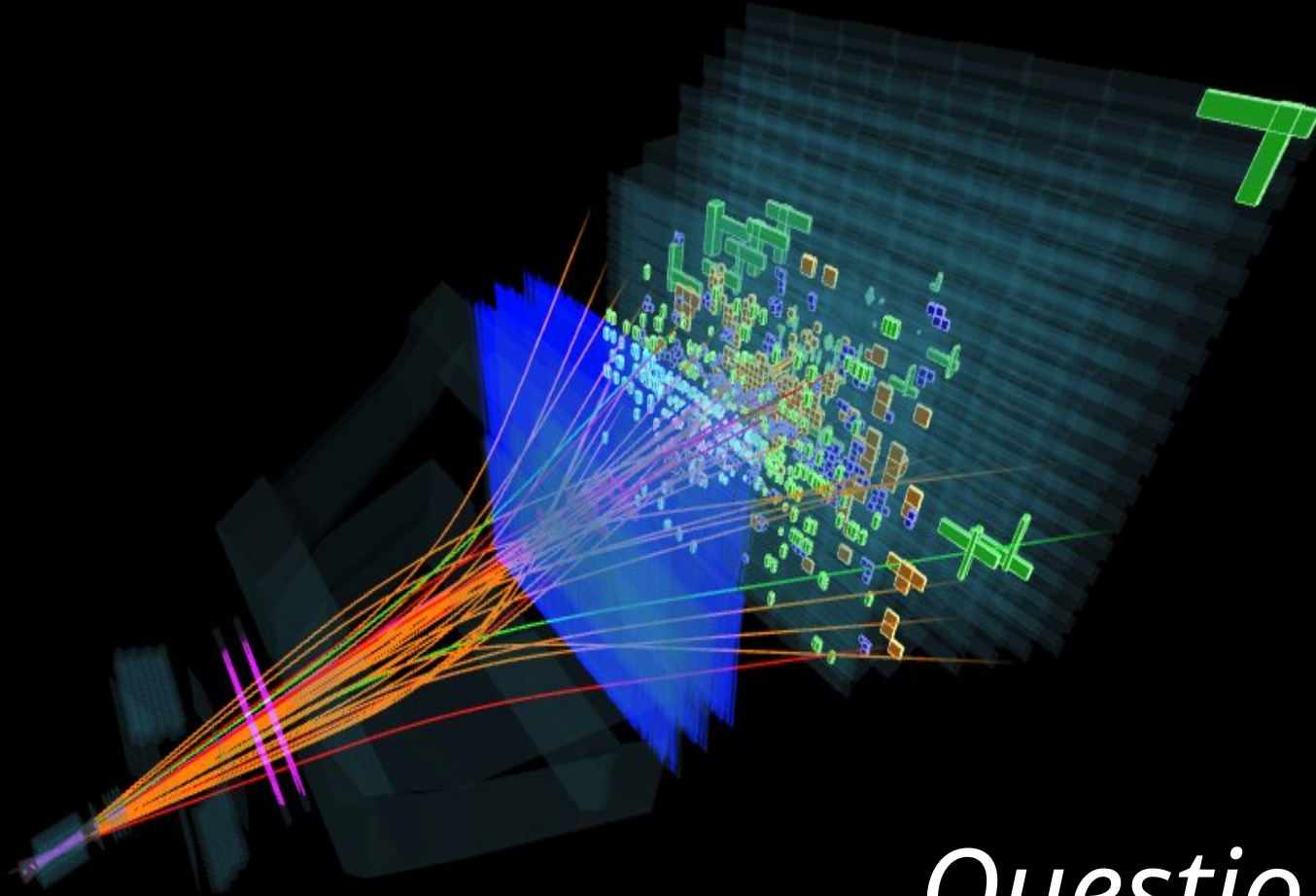
- Edoardo Martelli
- Marc Collignon
- Tony Cass



Event 158826354

Run 206854

Sat, 28 Apr 2018 21:48:17



Questions?

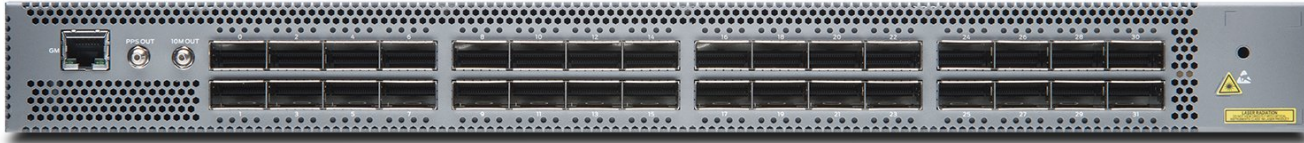
e.m@cern.ch

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

```
[...]
```