LHCb High Level Trigger in remote IT datacentre

Edoardo Martelli[1,*], *Loïc* Brarda[2,**], *Luis* Granado Cardoso[2,***], *Marc* Collignon[1,****], *Niko* Neufeld[2,†], *Tommaso* Colombo[2,‡], and *Tony* Cass[1,§]

[1]CERN (Geneva, Switzerland), IT-CS group
[2]CERN (Geneva, Switzerland), LHCb experiment

Abstract. In 2017 the CERN IT department and the major LHC experiments reviewed the possibility to build a common High Level Trigger facility, to share costs and for a more efficient utilization of computing resources. This paper describes a Proof-of-Concept built by the CERN IT-CS group and the LHCb online and offline teams to demonstrate the feasibility of the proposal: the dedicated, high speed connection between the LHCb detector and the IT data-centre, the setup of the servers and how they were used for offline simulation and online data taking, the different technologies considered and used, the challenges encountered and how they were tackled. The Proof of Concept exceed expectations and the remote servers were used for real data taking for more than two months during the summer 2018.

1 Introduction

The feasibility of a shared facility for the HLT farms of the four LHC Experiments was reviewed in 2017. The IT department proposed a state of the art data-centre building to be built on the French site of Prévessin, connected with very high bandwidth network links to the four major LHC experiments and to the IT data-centre in Meyrin [1].
The connections of the envisaged shared facility are depicted in figure 1.
The main advantages of the proposed solution are:

- the possibility to re-purpose the servers during periods without data taking, and
- a reduction in aggregate costs as infrastructure and operational overheads would be shared between IT and the four experiments.

The main challenges are:

- managing and automating the re-purposing of servers among different management domains (IT, LHC experiments), and
- supporting the very high data rates —up to 40 Tbps —from the detectors to the HLTs.

[*] e-mail: edoardo.martelli@cern.ch - main author
[**] e-mail: loic.brarda@cern.ch
[***] e-mail: luis.granado@cern.ch
[****] e-mail: marc.collignon@cern.ch
[†] e-mail: niko.neufeld@cern.ch
[‡] e-mail: tommaso.colombo@cern.ch
[§] e-mail: tony.cass@cern.ch

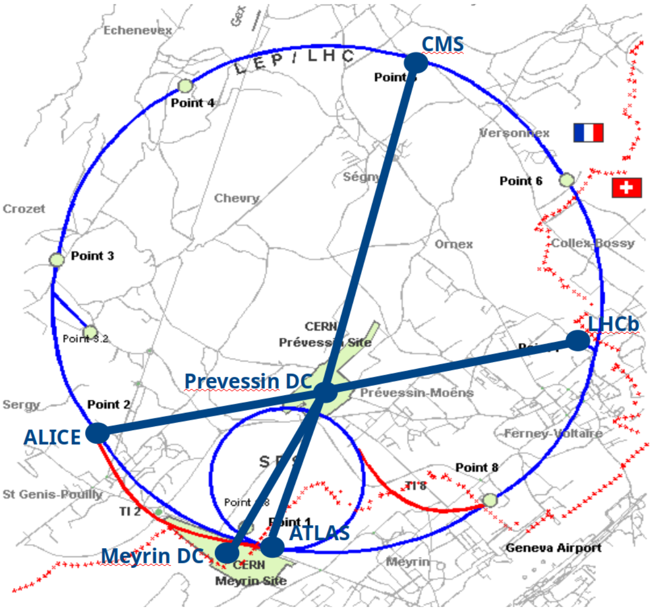


Figure 1. A proposal for a shared HLT facility.

Unfortunately, construction of such a shared data-centre could not be confirmed early enough to ensure its construction on the timescale necessary to meet the needs of ALICE and LHCb Run3 and these experiments are thus currently constructing dedicated facilities to house their HLT farms near their detectors.

Despite the lack of an immediate need, however, it was considered valuable to perform a proof-of-concept study to inform future discussions about a possible common facility. In fact the remote HLT could be implemented also in the existing data-centre in Building 513.

2 Setup of the case study

The IT-CS group and the LHCb online and offline teams agreed to use their facilities to set up this case study.

The locations involved were:

- the LHCb Experiment at Point 8 in Ferney-Voltaire as the data source, and
- the CERN IT data-centre in Building 513 in Meyrin which hosted the test HLT servers.

The locations are depicted in figure 2.

Network connectivity between the two sites was provided by two Juniper QFX switches equipped with four 100 Gbps interfaces each, connected to four fibre pairs.

The switch on the LHCb side was connected to A) the LHCb detector via the production Force10 switch (currently connecting the hardware trigger used for Run1 and Run2); B) to the management network switch. The switch in the IT data-centre was connected to C) two production top-of-rack (ToR) switches for the control/management interfaces of the servers; D) directly to the data interfaces of the test HLT servers.

The network setup is depicted in figure 3.

The list of the equipment used for the case-study is:

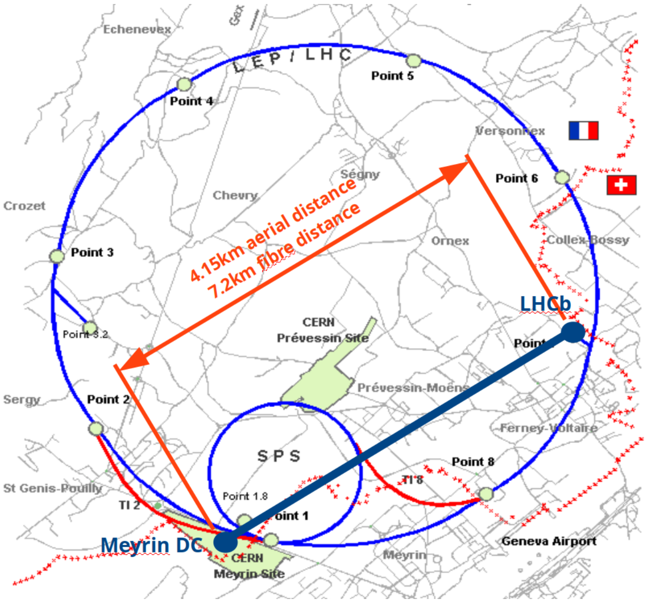


Figure 2. Case study locations.

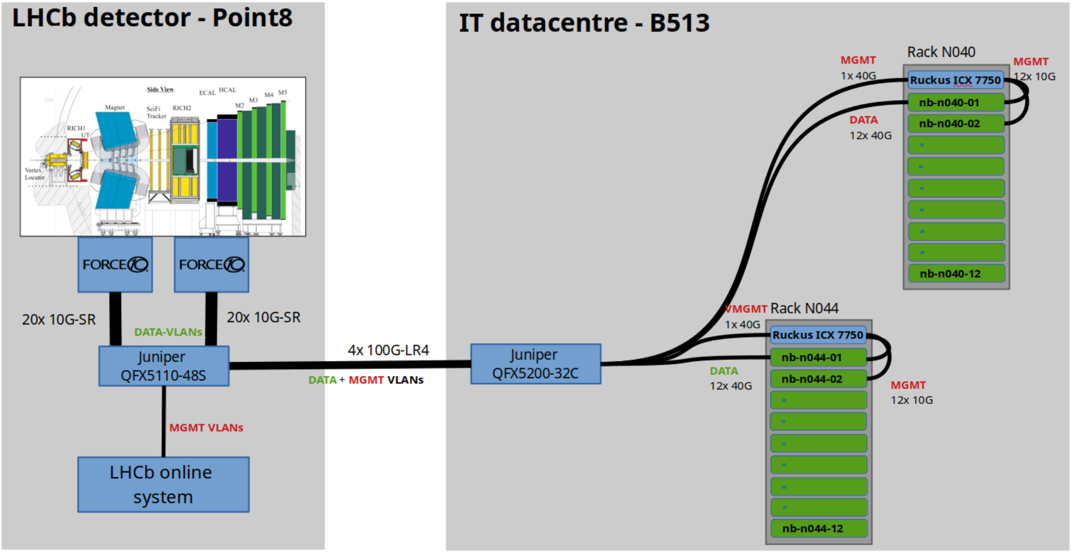


Figure 3. Case study setup

- Network Hardware
  - 1x Juniper QFX5200-32C with 4x100 Gbps and 24x 40 Gbps interfaces,
  - 1x Juniper QFX5110-48S with 4x100 Gbps and 40x 10 Gbps interfaces, and
  - 2x Ruckus ICX7750 with 48x 10GbaseT and 6x 40 Gbps interfaces.

- Servers in B513 data-centre:
    - 24x servers Supermicro Quanta server with Intel Xeon CPU E5-2630 2.20 GHz, 16GB of RAM, 1x 10 Gbps NIC for control and 1x 40 Gbps NIC for data traffic.

3 Observed results

3.1 Server automation

The servers in building 513 were originally part of the CERN IT Openstack cloud.

For the hand over to LHCb, the IT ToR switches were reconfigured to be connected at Layer2 to the LHCb Online System, thus routing PXE boot requests from the test HLT servers to LHCb managed DHCP servers, rather than those managed by the IT Department.

This configuration effectively handed control of the machines to LHCb and the Online System was able to configure the servers as offline Simulation nodes or as HLT nodes according to the LHC accelerator status, just as for the production HLT servers located at the experiment.

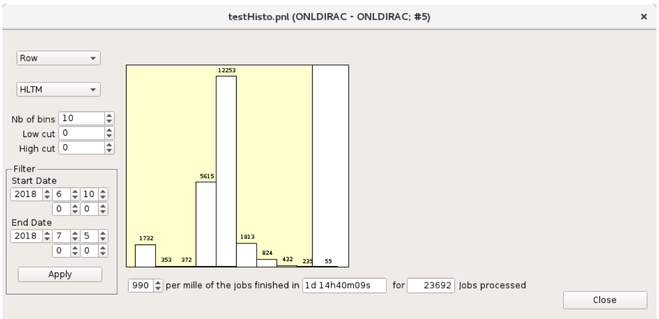
In this proof-of-concept setup, the redirection of the PXE boot packets required some manual cabling and manual reconfiguration of the ToR switches in the Meyrin Data Centre. In a production setup, however, no manual reconfiguration would be required as switching servers between the IT cloud and the Experiment cloud could easily be automated by the IT-CS Network Management System.

3.2 Offline Simulation

In the first phase of the case study, the nodes were used for offline Monte Carlo simulation.

In this period, the 24 nodes in Building 513 (identified as HLTM01 and HLTM02) ran 23692 jobs, 1.5% of the jobs run in the whole LHCb farm in the same period. It was noticed that the jobs in the B513 farm were executed 15% faster than the whole average. Each of the node had 4 CPUs with 10 cores.

These results were confirmed by the values reported in the figures 4 and 5



testHisto.pnl (ONLDIRAC - ONLDIRAC; #5)

Row

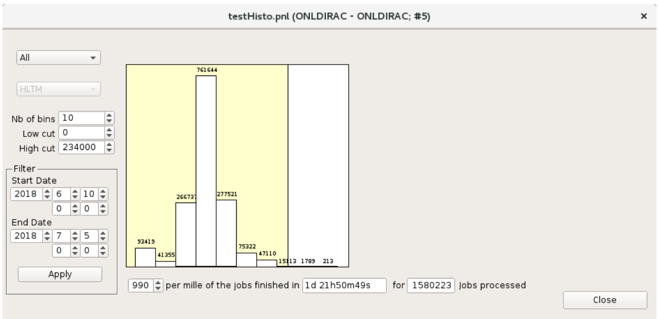
HLTM

Nb of bins 10
Low cut 0
High cut 0

Filter
Start Date 2018 6 10 0 0
End Date 2018 7 5 0 0

990 per mille of the jobs finished in 1d 14h40m09s for 23692 Jobs processed

Figure 4. Number of jobs and their average duration ran on the HLTM servers during offline simulation.



testHisto.pnl (ONLDIRAC - ONLDIRAC; #5)

All

HLTM

Nb of bins 10
Low cut 0
High cut 234000

Filter
Start Date 2018 6 10 0 0
End Date 2018 7 5 0 0

Apply

990 per mille of the jobs finished in 1d 21h50m49s for 1580223 Jobs processed

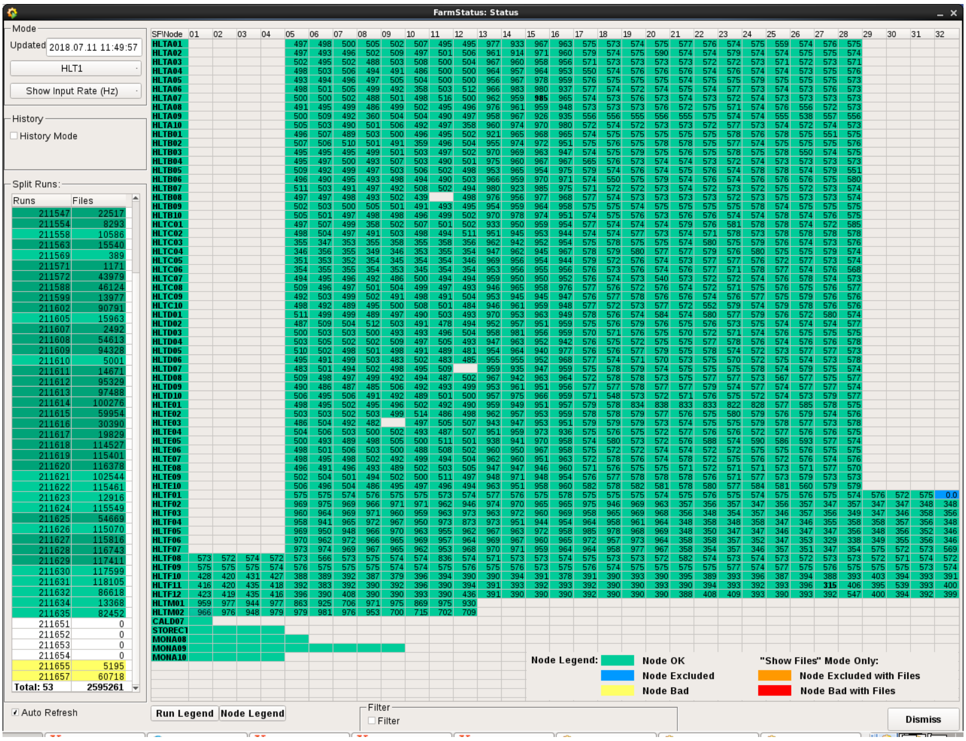
Close

Figure 5. Number of jobs and their average duration ran on the whole LHCb HLT servers during offline simulation.

3.3 Online Remote High Level Trigger

When the LHC finally started delivering collisions in July 2018, the servers were converted to HLT nodes and added to two sub-farms (HLTM01 and HLTM02) of the production LHCb HLT farm.

The remote nodes behaved correctly and performed exactly as the production nodes. The LHCb online team actually commented: “Operationally and performance-wise, it seems to be the same as running a sub-farm at Point 8".



FarmStatus: Status

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

History
□ History Mode

Split Runs:

Runs	Files
211547	22517
211554	8293
211558	10586
211563	15540
211569	389
211571	1171
211572	43979
211588	46124
211599	13977
211602	90791
211605	15963
211607	2492
211608	54613
211609	94328
211610	5001
211611	14671
211612	95329
211613	97488
211614	100276
211615	59954
211616	30390
211617	19829
211618	114527
211619	115401
211620	116378
211621	102544
211622	115461
211623	12916
211624	115549
211625	54669
211626	115070
211627	115816
211628	116743
211629	117411
211630	117599
211631	118105
211632	86618
211634	13368
211635	82452
211651	0
211652	0
211653	0
211654	0
211655	5195
211657	60718
Total: 53	**2595261**

☑ Auto Refresh

SF\Node	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				
HLTA02					497	493	496	502	509	497	501	506	961	914	971	960	579	574	575	590	574	574	579	574	574	575	575	574				
HLTA03					502	495	502	488	503	508	500	504	967	960	958	956	571	573	573	573	573	572	572	573	571	572	573	571				
HLTA04					498	503	506	494	491	486	500	500	964	957	964	953	550	574	576	576	576	574	576	574	574	573	575	576				
HLTA05					493	494	496	497	505	504	500	500	956	967	978	959	576	575	575	575	574	575	575	579	575	575	575	573				
HLTA06					498	501	505	499	492	358	503	512	966	983	980	937	577	574	572	574	575	574	577	573	574	573	576	573				
HLTA07					500	500	502	488	501	498	516	500	962	959	**985**	965	574	573	576	573	574	573	572	574	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	576	556	572	573				
HLTA09					500	509	492	360	504	504	490	497	958	967	926	935	556	556	555	556	555	575	574	574	555	538	557	556				
HLTA10					505	503	490	501	506	492	497	358	960	974	970	980	572	574	572	573	573	572	577	573	574	572	574	573				
HLTB01					496	507	489	503	500	496	495	502	921	965	968	965	574	575	575	575	575	575	578	576	578	575	551	575				
HLTB02					507	506	510	501	491	359	496	504	955	974	972	951	575	576	575	578	578	575	577	574	575	575	575	576				
HLTB03					495	495	495	499	501	503	497	502	970	969	963	947	574	575	579	575	576	575	578	575	578	550	574	575				
HLTB04					495	497	500	493	507	503	490	501	975	960	967	967	565	576	573	574	574	573	572	574	573	573	573	573				
HLTB05					509	492	499	497	503	506	502	498	953	965	954	975	579	574	576	574	575	576	574	578	578	574	579	551				
HLTB06					496	490	495	493	498	494	490	503	966	959	970	971	574	550	575	579	574	576	574	576	576	576	575	580				
HLTB07					511	503	491	497	492	508	502	494	980	923	985	975	571	572	572	573	574	572	573	572	575	572	573	574				
HLTB08					497	497	498	493	502	439		498	976	956	977	968	577	574	573	573	573	573	572	573	575	573	573	574				
HLTB09					502	503	500	505	501	491	493	495	954	959	964	958	575	575	574	575	575	575	575	578	574	575	575	575				
HLTB10					505	501	497	498	498	496	499	502	970	978	974	951	574	575	576	573	576	576	574	574	578	574	576	575				
HLTC01					497	507	499	358	502	507	501	502	933	950	959	954	577	574	574	574	579	576	581	578	578	574	572	585				
HLTC02					498	504	497	491	503	498	494	511	951	945	953	944	574	574	577	573	574	571	578	573	578	578	578	578				
HLTC03					355	347	353	355	358	355	358	356	962	942	952	954	575	578	575	575	574	580	575	579	576	574	573	576				
HLTC04					346	356	355	349	346	353	355	354	947	962	945	967	578	579	580	577	577	579	576	580	575	575	579	574				
HLTC05					351	353	352	354	345	354	354	346	969	956	954	944	579	572	576	574	573	577	577	576	572	577	573	574				
HLTC06					354	355	355	354	353	345	354	354	953	956	955	956	576	573	576	574	576	577	571	578	577	574	576	568				
HLTC07					494	495	496	492	486	500	494	494	959	950	950	952	576	574	573	540	573	572	572	574	576	578	574	573				
HLTC08					509	496	497	501	504	499	497	493	946	965	958	976	577	576	572	576	574	572	571	575	576	575	576	577				
HLTC09					492	503	499	502	491	498	491	504	953	945	945	947	576	577	578	576	576	574	576	577	575	579	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				
HLTD01					511	499	499	489	497	490	503	493	970	953	963	949	578	576	574	584	574	580	577	579	576	572	580	574				
HLTD02					487	509	504	512	503	491	478	494	952	957	951	959	575	576	579	576	575	576	573	575	574	574	574	577				
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				
HLTD04					503	505	502	502	509	497	505	493	947	963	952	942	576	575	572	575	575	577	578	576	574	576	576	578				
HLTD05					510	502	498	501	498	491	489	481	954	964	940	977	576	576	577	579	575	578	574	572	573	577	577	573				
HLTD06					495	491	499	503	483	502	483	485	955	955	952	968	577	574	571	570	573	575	570	572	575	574	573	578				
HLTD07					483	501	494	502	498	495	509		959	935	947	959	575	578	579	574	575	575	575	578	574	579	575	574				
HLTD08					509	498	497	499	492	494	487	502	967	942	963	964	572	578	578	573	575	577	577	573	567	577	575	577				
HLTD09					490	486	487	485	506	492	493	499	953	961	951	956	577	577	578	577	577	579	574	577	574	577	577	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	577				
HLTE01					498	495	502	495	496	502	492	490	959	949	951	957	579	578	834	838	833	833	822	828	577	585	578	575				
HLTE02					503	503	502	503	499	514	486	498	962	957	953	959	578	578	579	577	576	575	580	579	576	579	574	576				
HLTE03					486	504	492	482		497	505	507	943	947	953	951	579	579	579	573	574	575	575	579	578	577	573	578				
HLTE04					504	506	503	500	502	493	487	507	951	959	973	936	575	576	575	572	577	576	576	572	577	576	576	575				
HLTE05					500	493	489	498	505	500	511	501	938	941	970	958	574	580	573	572	576	588	574	590	586	593	577	574				
HLTE06					498	501	506	503	500	488	508	502	960	950	967	958	575	572	572	574	574	572	572	575	575	576	575	575				
HLTE07					498	495	498	502	492	499	494	504	962	960	951	963	572	578	576	574	578	572	575	576	572	576	574	576				
HLTE08					496	491	496	493	489	502	503	505	947	947	946	960	571	576	575	575	571	572	571	571	573	571	577	570				
HLTE09					502	504	501	494	502	500	511	497	948	971	948	954	576	577	578	578	578	576	571	577	573	579	573	573				
HLTE10					506	496	504	486	495	497	496	494	963	951	958	960	582	578	582	581	578	580	577	584	581	560	579	579				
HLTF01					575	575	574	576	575	575	573	574	577	576	575	578	575	575	575	574	575	576	575	574	575	576	575	574	576	572	575	0.0
HLTF02					969	975	969	966	971	971	962	946	974	970	965	965	975	946	969	963	357	356	357	347	356	357	347	357	347	347	348	348
HLTF03					960	964	969	971	960	959	963	973	963	972	960	969	958	965	969	968	356	348	354	357	346	357	356	349	347	346	358	356
HLTF04					958	941	965	972	967	950	973	873	973	951	944	954	964	958	961	964	348	358	348	358	347	346	355	358	358	357	356	348
HLTF05					969	950	948	966	970	963	955	962	967	963	972	958	985	978	968	969	348	350	347	347	346	347	347	356	348	356	352	346
HLTF06					970	962	972	966	965	969	957	964	969	967	960	965	972	957	973	964	358	358	357	352	347	353	329	338	349	355	356	346
HLTF07					973	974	969	967	965	962	953	968	970	971	959	964	964	958	977	967	358	354	357	346	357	351	347	354	575	572	573	569
HLTF08	573	572	574	572	573	566	573	575	574	574	836	574	571	573	573	574	575	573	573	572	582	574	573	574	573	572	573	573	572	571	574	572
HLTF09	575	575	575	574	576	575	575	575	574	574	575	576	575	576	573	575	576	574	575	575	575	575	577	574	577	575	576	575	575	575	573	574
HLTF10	428	420	431	427	388	389	392	387	379	396	394	390	390	394	391	378	391	390	393	390	395	389	393	396	387	394	388	393	403	394	393	391
HLTF11	416	420	435	418	392	383	392	390	392	396	390	394	391	393	392	393	392	390	390	393	390	394	393	392	393	396	**315**	406	395	539	393	400
HLTF12	423	419	435	416	396	390	408	390	390	393	390	436	391	390	390	390	392	390	390	390	388	408	409	393	390	393	392	547	400	394	392	399
HLTM01	959	977	944	977	863	925	706	971	975	869	975	930																				
HLTM02	966	976	948	979	979	981	976	953	700	715	702	709																				
CALD07	🟩																															
STOREC1	🟩	🟩	🟩	🟩																												
MONA08	🟩	🟩	🟩	🟩	🟩																											
MONA09	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩																							
MONA10	🟩	🟩	🟩	🟩																												

Node Legend: 🟩 Node OK, 🟦 Node Excluded, 🟨 Node Bad
"Show Files" Mode Only: 🟧 Node Excluded with Files, 🟥 Node Bad with Files

Run Legend | Node Legend

Filter
□ Filter

Dismiss

Figure 6. LHCb HLT instantaneous input rate (Hz) node by node.

Figure 6 shows input rate of all the nodes used in the LHCb farm. The rates on the remote nodes are similar to those of other nodes.

4 DWDM transmission test

During Run3, LHCb will need to send data to the HLT farm at rates of up to 40 Tbps. While data transmission at such rates is affordable over very short distances, the transmission over distances exceeding a few tens of meters can quickly become very expensive, especially if fibre connections are scarce, as is the case between LHCb Point 8 and the IT data-centre in Building 513.

Given the very few fibre pairs available, two options were costed during the design studies for the proposed shared facility on the Prévessin site: installation of additional fibres and the use of DWDM (Dense Wavelength Division Multiplexing) transmission equipment. At the time of those studies, the only DWDM equipment available was that traditionally used by Telecom operators to cover longer distances and the cost of the DWDM solution exceeded the estimated cost of digging the relatively short trenches (<10km) required for the installation of additional fibres.

Since those studies, however, a more cost effective DWDM solution has appeared based on a new emerging technology, PAM-4 (Pulse Amplitude Modulation with 4 levels of pulse modulation). The Remote HLT proof-of-concept study provided a valuable opportunity to test this technology in a real-world scenario.

The PAM-4 technology could be tested thanks to the collaboration with Smartoptics[2], a Scandinavian manufacturer of transmission equipment, which provided a demo kit.

The significant cost advantages of the PAM-4 technology based DWDM solution is that the QSFP-28 100 Gbps optical transceivers can be plugged directly into standard switches or servers' NICs. In our case, the transceivers were plugged into the Juniper switches already deployed for the case study.

Unfortunately, though, these transceivers have the peculiarity of having a transmission power level (-11 dBm) lower of its input sensitivity (-2 dBm). For this reason, an active line system is always necessary to amplify the signal. In the Smartoptics solution, the active line system was also the multiplexer/demultiplexer that merges the different signals into a single fibre pair.

4.1 Network setup

Initially it was planned to replace the four fibre pairs providing a 400 Gbps with 100GbaseLR4 optics by a single fibre pair using the PAM4/DWDM system to multiplex the four 100 Gbps streams.

However, the demo kit consisted of only two 100 Gbps DWDM channels, so the bandwidth had to be reduced to 200 Gbps. Additionally, the Juniper switch at Point 8, a QFX5110, was not able to supply sufficient power to meet the requirements of the laser of the PAM-4 optics. As a consequence, two fibre pairs were looped together in Point 8 and the test was performed not with real data from the experiment but rather with a simulation of traffic generated by servers in B513. This change to our plans, however, does not affect the utility of the tests for our purposes—and, in any case the two-day period during which the demo kit was available did not correspond with a period of data taking. Figure 7 depicts the configuration of the network setup.

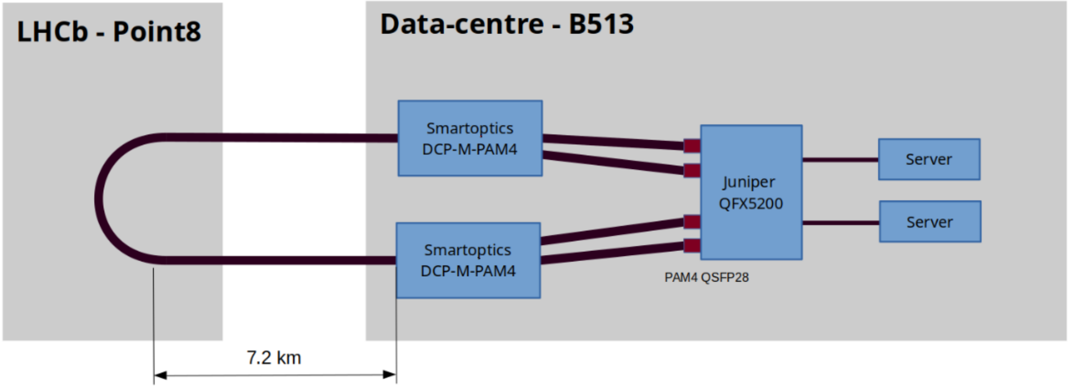


Figure 7. DWDM PAM-4 test setup.

4.2 Findings

The PAM-4 technology proved to be suitable for affordable DWDM transmission over the link distance tested, around 15 km.

In particular, the tests showed that:

- PAM-4 optical transceivers are usable with devices, such as the Juniper QFX5200-32C which support high power optical modules,
- the PAM-4 transceivers place no special requirements on the operating system; the Juniper switch used for the tests had a standard Operating System version installed, and.
- the two channels ran with high load for more than 12 hour without any transmission errors.

Some references to Juniper JUNOS commands used in the test are reported in Appendix A.

5 Conclusions

The tests demonstrated that a common computing facility for a shared Software High Level Trigger has been proved feasible. In this case the servers were shared between IT and LHCb, but the setup could be easily extended to other Experiments.

Additionally, we have shown that PAM-4 DWDM technology is an affordable and ready-to-use option to support Terabit data transmission over medium-range distances using a limited number of fibre pairs.

Appendix A: Juniper JUNOS

Version

The PAM-4 QSF28 optical transceiver are supported on Juniper QFX running JUNOS version 17.3 or later [3].

Configuration

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

This command shows the recognized PAM-4 transceivers

```
# run show chassis hardware
    [...]
    QSFP-100GE-DWDM2
    QSFP-100GE-DWDM2
```

This command shows the two 50 Gbps lines up and their power levels

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

References

[1] I.Bird, H.Meinhard, *Evolution of CERN facilties* (HSF workshop, San Diego (US) - January 2017)
https://indico.cern.ch/event/570249/

[2] Smartoptics, Ryensvingen 7 NO-0680 Oslo, Norway
https://www.smartoptics.com/

[3] Juniper Networks, "Junos" Operating System, version 17.3, not publicly available.
https://www.juniper.net/us/en/products-services/nos/junos/