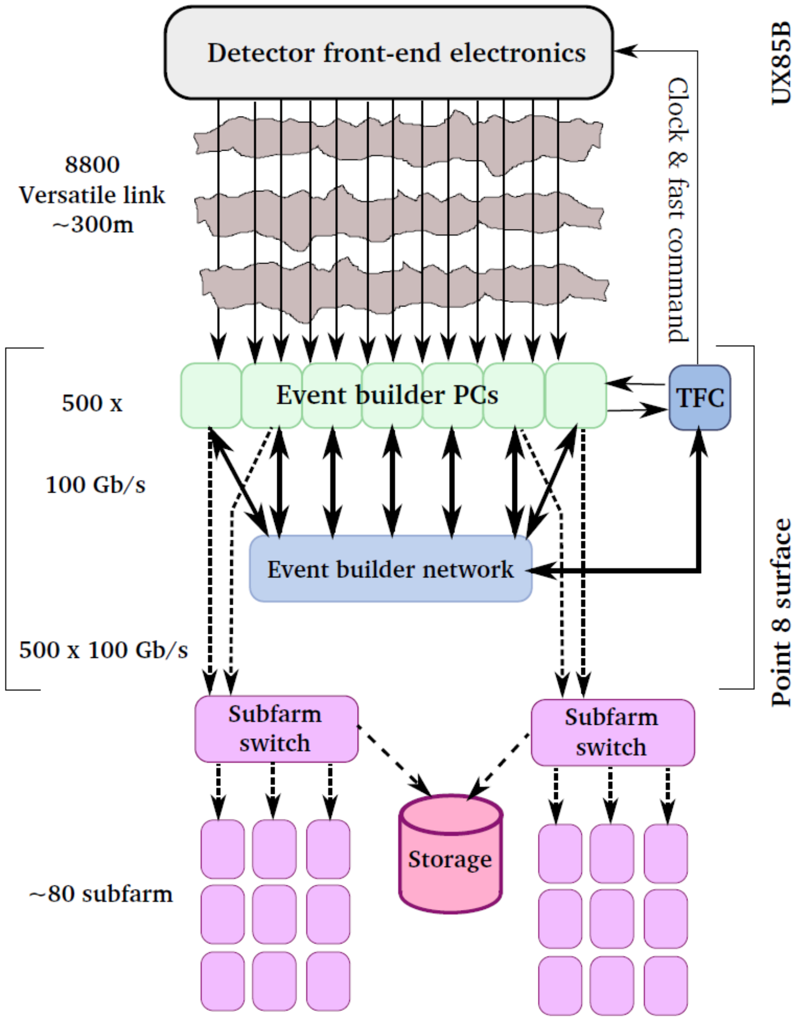
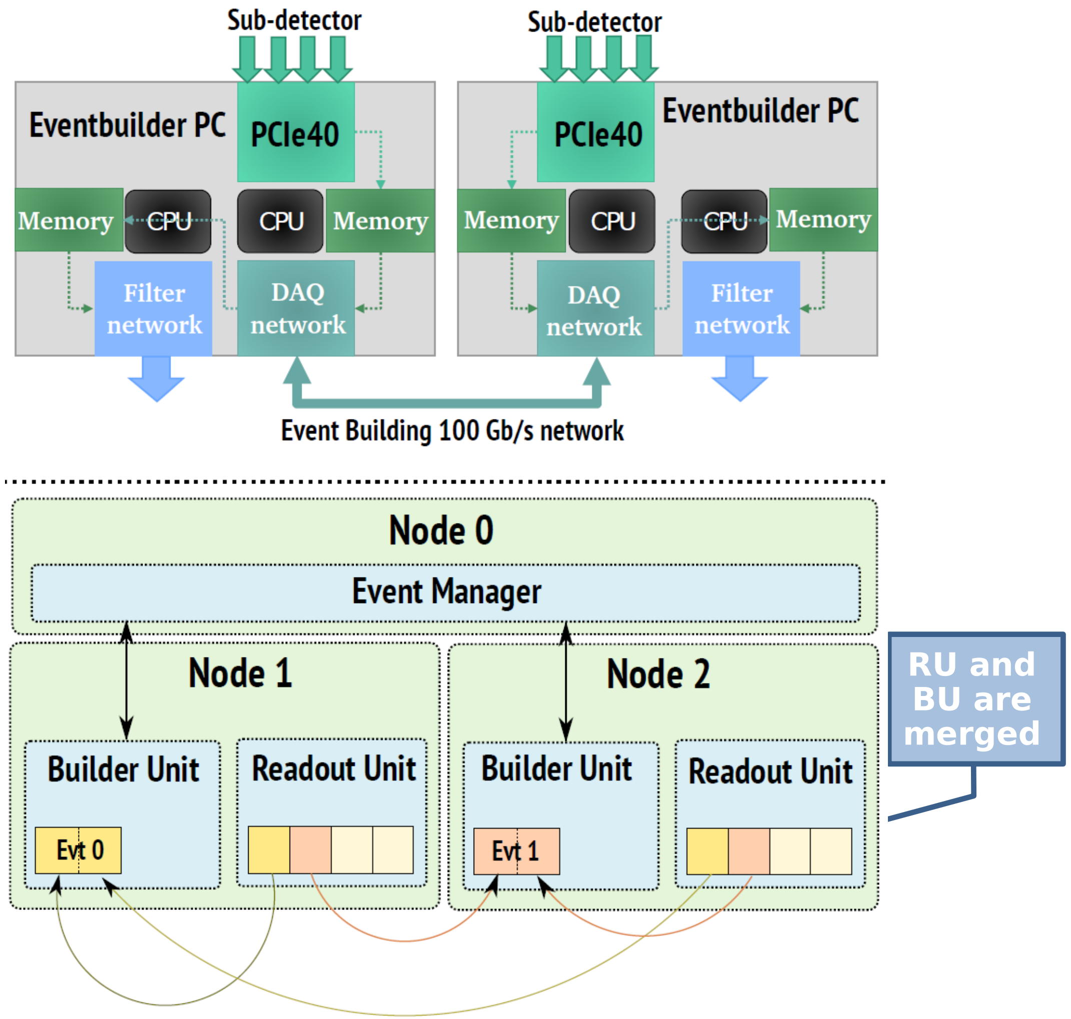
Evaluation of 100 Gb/s LAN networks for the LHCb DAQ upgrade

Balázs Vőneki and Sébastien Valat

The architecture of the upgraded LHCb experiment readout-system:



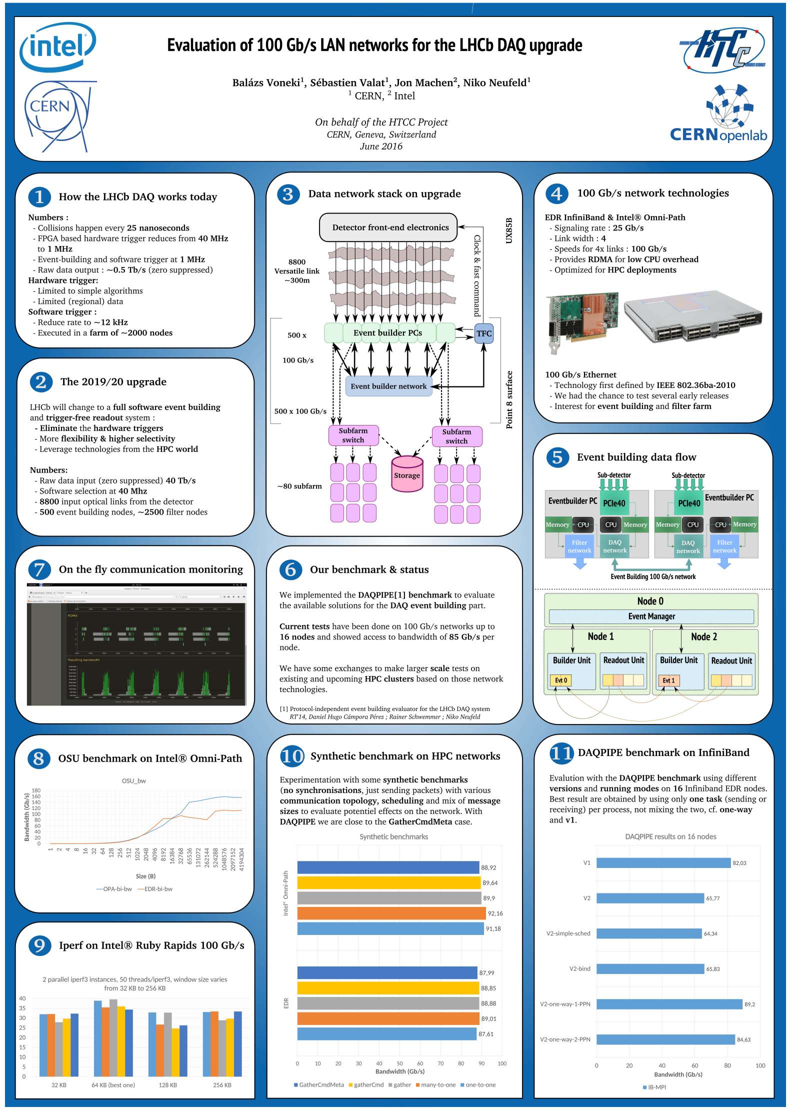
Event building data flow:



Evaluation of 100 Gb/s LAN networks for the LHCb DAQ upgrade

Balázs Vőneki and Sébastien Valat

Poster 229



Evaluation of 100 Gb/s LAN networks for the LHCb DAQ upgrade

Balázs Voneki[1], Sébastien Valat[1], Jon Machen[2], Niko Neufeld[1]
[1] CERN, [2] Intel

On behalf of the HTCC Project
CERN, Geneva, Switzerland
June 2016

1 How the LHCb DAQ works today

Numbers :
- Collisions happen every **25 nanoseconds**
- FPGA based hardware trigger reduces from **40 MHz** to **1 MHz**
- Event-building and software trigger at **1 MHz**
- Raw data output : **~0.5 Tb/s** (zero suppressed)

Hardware trigger:
- Limited to simple algorithms
- Limited (regional) data

Software trigger :
- Reduce rate to **~12 kHz**
- Executed in a **farm of ~2000 nodes**

2 The 2019/20 upgrade

LHCb will change to a **full software event building** and **trigger-free readout** system :
- **Eliminate** the **hardware triggers**
- More **flexibility & higher selectivity**
- Leverage technologies from the **HPC world**

Numbers:
- Raw data input (zero suppressed) **40 Tb/s**
- Software selection at **40 Mhz**
- **8800** input optical links from the detector
- **500** event building nodes, **~2500** filter nodes

3 Data network stack on upgrade

4 100 Gb/s network technologies

EDR InfiniBand & Intel® Omni-Path
- Signaling rate : **25 Gb/s**
- Link width : **4**
- Speeds for 4x links : **100 Gb/s**
- Provides **RDMA** for **low CPU overhead**
- Optimized for **HPC deployments**

100 Gb/s Ethernet
- Technology first defined by **IEEE 802.36ba-2010**
- We had the chance to test several early releases
- Interest for **event building** and **filter farm**

5 Event building data flow

6 Our benchmark & status

We implemented the **DAQPIPE[1] benchmark** to evaluate the available solutions for the **DAQ event building** part.

Current tests have been done on 100 Gb/s networks up to **16 nodes** and showed access to bandwidth of **85 Gb/s** per node.

We have some exchanges to make larger **scale** tests on existing and upcoming **HPC clusters** based on those network technologies.

[1] Protocol-independent event building evaluator for the LHCb DAQ system
RT'14, Daniel Hugo Cámpora Pérez ; Rainer Schwemmer ; Niko Neufeld

7 On the fly communication monitoring

8 OSU benchmark on Intel® Omni-Path

9 Iperf on Intel® Ruby Rapids 100 Gb/s

2 parallel iperf3 instances, 50 threads/iperf3, window size varies from 32 KB to 256 KB
40
35
30
25
20
15
10
5
0
32 KB
64 KB (best one)
128 KB
256 KB

10 Synthetic benchmark on HPC networks

Experimentation with some **synthetic benchmarks** (**no synchronisations**, just sending packets) with various **communication topology, scheduling** and mix of **message sizes** to evaluate potentiel effects on the network. With **DAQPIPE** we are close to the **GatherCmdMeta** case.

11 DAQPIPE benchmark on InfiniBand

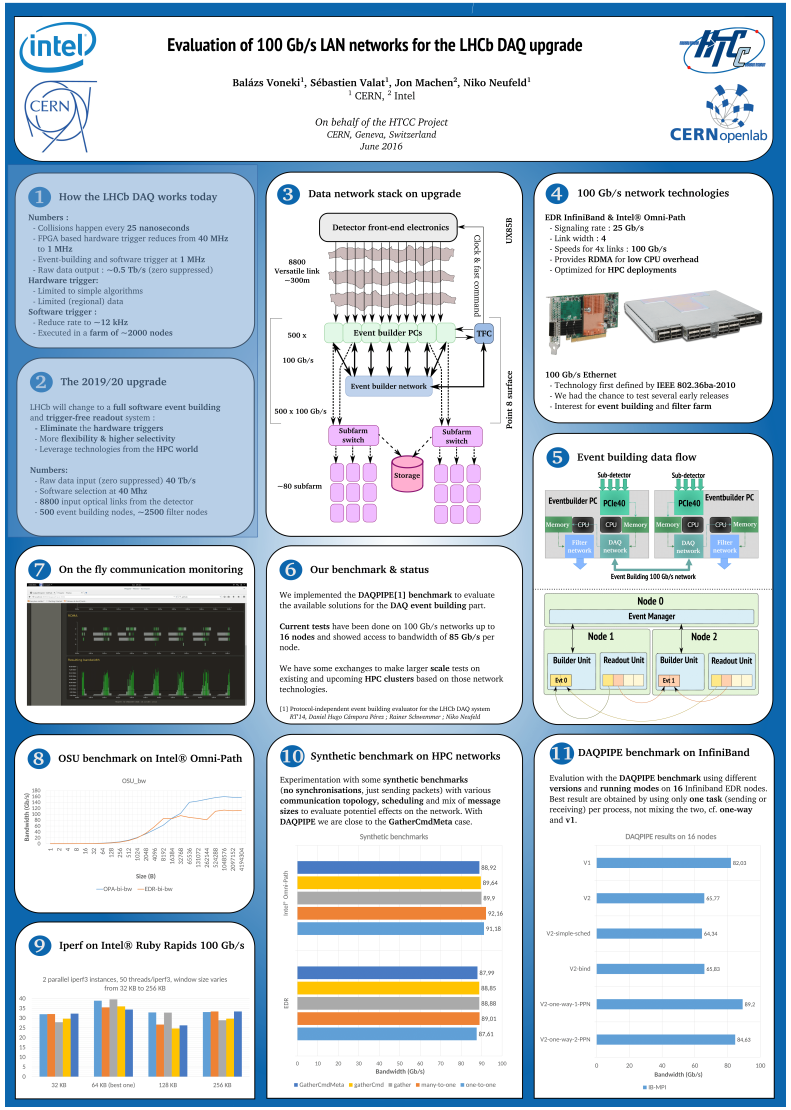
Evalution with the **DAQPIPE benchmark** using different **versions** and **running modes** on **16** Infiniband EDR nodes. Best result are obtained by using only **one task** (sending or receiving) per process, not mixing the two, cf. **one-way** and **v1**.

Evaluation of 100 Gb/s LAN networks for the LHCb DAQ upgrade

Balázs Vőneki and Sébastien Valat

Poster 229

Introduction



Evaluation of 100 Gb/s LAN networks for the LHCb DAQ upgrade

Balázs Voneki[1], Sébastien Valat[1], Jon Machen[2], Niko Neufeld[1]
[1] CERN, [2] Intel

On behalf of the HTCC Project
CERN, Geneva, Switzerland
June 2016

1 How the LHCb DAQ works today

Numbers :
- Collisions happen every **25 nanoseconds**
- FPGA based hardware trigger reduces from **40 MHz** to **1 MHz**
- Event-building and software trigger at **1 MHz**
- Raw data output : **~0.5 Tb/s** (zero suppressed)

Hardware trigger:
- Limited to simple algorithms
- Limited (regional) data

Software trigger :
- Reduce rate to **~12 kHz**
- Executed in a **farm of ~2000 nodes**

2 The 2019/20 upgrade

LHCb will change to a **full software event building** and **trigger-free readout** system :
- **Eliminate** the **hardware triggers**
- More **flexibility & higher selectivity**
- Leverage technologies from the **HPC world**

Numbers:
- Raw data input (zero suppressed) **40 Tb/s**
- Software selection at **40 Mhz**
- **8800** input optical links from the detector
- **500** event building nodes, **~2500** filter nodes

3 Data network stack on upgrade

4 100 Gb/s network technologies

EDR InfiniBand & Intel® Omni-Path
- Signaling rate : **25 Gb/s**
- Link width : **4**
- Speeds for 4x links : **100 Gb/s**
- Provides **RDMA** for **low CPU overhead**
- Optimized for **HPC deployments**

100 Gb/s Ethernet
- Technology first defined by **IEEE 802.36ba-2010**
- We had the chance to test several early releases
- Interest for **event building** and **filter farm**

5 Event building data flow

7 On the fly communication monitoring

6 Our benchmark & status

We implemented the **DAQPIPE[1] benchmark** to evaluate the available solutions for the **DAQ event building** part.

Current tests have been done on 100 Gb/s networks up to **16 nodes** and showed access to bandwidth of **85 Gb/s** per node.

We have some exchanges to make larger **scale** tests on existing and upcoming **HPC clusters** based on those network technologies.

[1] Protocol-independent event building evaluator for the LHCb DAQ system *RT'14, Daniel Hugo Cámpora Pérez ; Rainer Schwemmer ; Niko Neufeld*

8 OSU benchmark on Intel® Omni-Path

9 Iperf on Intel® Ruby Rapids 100 Gb/s

2 parallel iperf3 instances, 50 threads/iperf3, window size varies from 32 KB to 256 KB
40
35
30
25
20
15
10
5
0
32 KB
64 KB (best one)
128 KB
256 KB

10 Synthetic benchmark on HPC networks

Experimentation with some **synthetic benchmarks** (**no synchronisations**, just sending packets) with various **communication topology, scheduling** and mix of **message sizes** to evaluate potentiel effects on the network. With **DAQPIPE** we are close to the **GatherCmdMeta** case.

11 DAQPIPE benchmark on InfiniBand

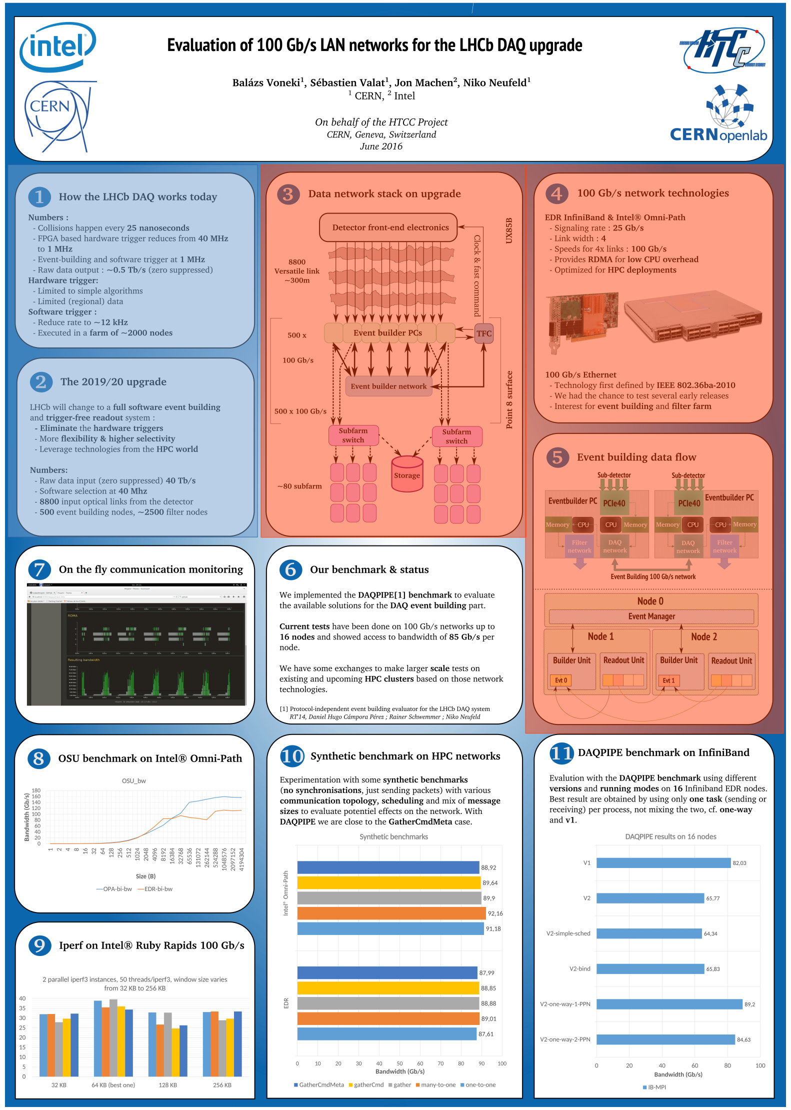
Evalution with the **DAQPIPE benchmark** using different **versions** and **running modes** on **16** Infiniband EDR nodes. Best result are obtained by using only **one task** (sending or receiving) per process, not mixing the two, cf. **one-way** and **v1**.

Evaluation of 100 Gb/s LAN networks for the LHCb DAQ upgrade

Balázs Vőneki and Sébastien Valat

Poster 229

Introduction



Evaluation of 100 Gb/s LAN networks for the LHCb DAQ upgrade

Balázs Voneki[1], Sébastien Valat[1], Jon Machen[2], Niko Neufeld[1]
[1] CERN, [2] Intel

On behalf of the HTCC Project
CERN, Geneva, Switzerland
June 2016

1 How the LHCb DAQ works today

Numbers :
- Collisions happen every **25 nanoseconds**
- FPGA based hardware trigger reduces from **40 MHz** to **1 MHz**
- Event-building and software trigger at **1 MHz**
- Raw data output : **~0.5 Tb/s** (zero suppressed)

Hardware trigger:
- Limited to simple algorithms
- Limited (regional) data

Software trigger :
- Reduce rate to **~12 kHz**
- Executed in a **farm of ~2000 nodes**

2 The 2019/20 upgrade

LHCb will change to a **full software event building** and **trigger-free readout** system :
- **Eliminate** the **hardware triggers**
- More **flexibility & higher selectivity**
- Leverage technologies from the **HPC world**

Numbers:
- Raw data input (zero suppressed) **40 Tb/s**
- Software selection at **40 Mhz**
- **8800** input optical links from the detector
- **500** event building nodes, **~2500** filter nodes

3 Data network stack on upgrade

4 100 Gb/s network technologies

EDR InfiniBand & Intel® Omni-Path
- Signaling rate : **25 Gb/s**
- Link width : **4**
- Speeds for 4x links : **100 Gb/s**
- Provides **RDMA** for **low CPU overhead**
- Optimized for **HPC deployments**

100 Gb/s Ethernet
- Technology first defined by **IEEE 802.36ba-2010**
- We had the chance to test several early releases
- Interest for **event building** and **filter farm**

5 Event building data flow

7 On the fly communication monitoring

6 Our benchmark & status

We implemented the **DAQPIPE[1] benchmark** to evaluate the available solutions for the **DAQ event building** part.

Current tests have been done on 100 Gb/s networks up to **16 nodes** and showed access to bandwidth of **85 Gb/s** per node.

We have some exchanges to make larger **scale** tests on existing and upcoming **HPC clusters** based on those network technologies.

[1] Protocol-independent event building evaluator for the LHCb DAQ system
RT'14, Daniel Hugo Cámpora Pérez ; Rainer Schwemmer ; Niko Neufeld

8 OSU benchmark on Intel® Omni-Path

9 Iperf on Intel® Ruby Rapids 100 Gb/s

2 parallel iperf3 instances, 50 threads/iperf3, window size varies from 32 KB to 256 KB

10 Synthetic benchmark on HPC networks

Experimentation with some **synthetic benchmarks** (**no synchronisations**, just sending packets) with various **communication topology**, **scheduling** and mix of **message sizes** to evaluate potentiel effects on the network. With **DAQPIPE** we are close to the **GatherCmdMeta** case.

Synthetic benchmarks

	GatherCmdMeta	gatherCmd	gather	many-to-one	one-to-one
Intel® Omni-Path	88,92	89,64	89,9	92,16	91,18
EDR	87,99	88,85	88,88	89,01	87,61

11 DAQPIPE benchmark on InfiniBand

Evalution with the **DAQPIPE benchmark** using different **versions** and **running modes** on **16** Infiniband EDR nodes. Best result are obtained by using only **one task** (sending or receiving) per process, not mixing the two, cf. **one-way** and **v1**.

DAQPIPE results on 16 nodes

	IB-MPI
V1	82,03
V2	65,77
V2-simple-sched	64,34
V2-bind	65,83
V2-one-way-1-PPN	89,2
V2-one-way-2-PPN	84,63

Architecture

Evaluation of 100 Gb/s LAN networks for the LHCb DAQ upgrade

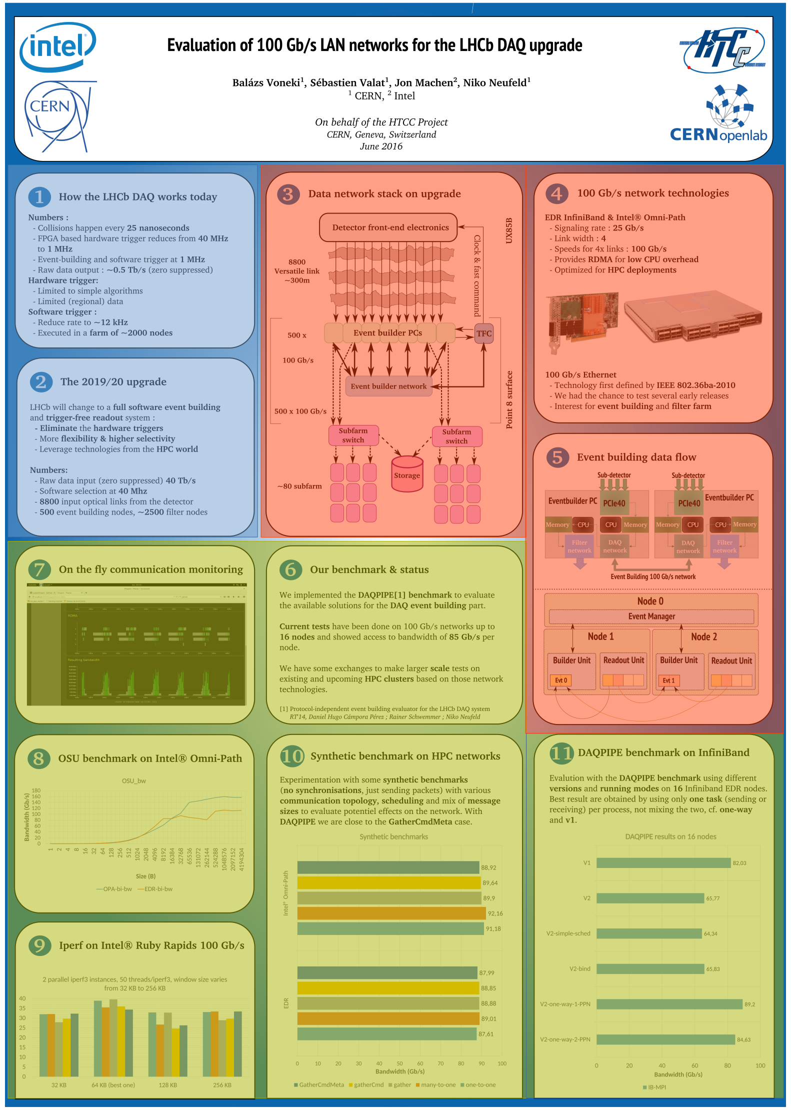
Balázs Vőneki and Sébastien Valat

Poster 229

Introduction

Architecture

Various benchmark results



Evaluation of 100 Gb/s LAN networks for the LHCb DAQ upgrade

Balázs Voneki[1], Sébastien Valat[1], Jon Machen[2], Niko Neufeld[1]
[1] CERN, [2] Intel

On behalf of the HTCC Project
CERN, Geneva, Switzerland
June 2016

1 How the LHCb DAQ works today

Numbers :
- Collisions happen every **25 nanoseconds**
- FPGA based hardware trigger reduces from **40 MHz** to **1 MHz**
- Event-building and software trigger at **1 MHz**
- Raw data output : **~0.5 Tb/s** (zero suppressed)

Hardware trigger:
- Limited to simple algorithms
- Limited (regional) data

Software trigger :
- Reduce rate to **~12 kHz**
- Executed in a **farm of ~2000 nodes**

2 The 2019/20 upgrade

LHCb will change to a **full software event building** and **trigger-free readout** system :
- **Eliminate** the **hardware triggers**
- More **flexibility & higher selectivity**
- Leverage technologies from the **HPC world**

Numbers:
- Raw data input (zero suppressed) **40 Tb/s**
- Software selection at **40 Mhz**
- **8800** input optical links from the detector
- **500** event building nodes, **~2500** filter nodes

3 Data network stack on upgrade

4 100 Gb/s network technologies

EDR InfiniBand & Intel® Omni-Path
- Signaling rate : **25 Gb/s**
- Link width : **4**
- Speeds for 4x links : **100 Gb/s**
- Provides **RDMA** for **low CPU overhead**
- Optimized for **HPC deployments**

100 Gb/s Ethernet
- Technology first defined by **IEEE 802.36ba-2010**
- We had the chance to test several early releases
- Interest for **event building** and **filter farm**

5 Event building data flow

7 On the fly communication monitoring

6 Our benchmark & status

We implemented the **DAQPIPE[1] benchmark** to evaluate the available solutions for the **DAQ event building** part.

Current tests have been done on 100 Gb/s networks up to **16 nodes** and showed access to bandwidth of **85 Gb/s** per node.

We have some exchanges to make larger **scale** tests on existing and upcoming **HPC clusters** based on those network technologies.

[1] Protocol-independent event building evaluator for the LHCb DAQ system *RT'14, Daniel Hugo Cámpora Pérez ; Rainer Schwemmer ; Niko Neufeld*

8 OSU benchmark on Intel® Omni-Path

9 Iperf on Intel® Ruby Rapids 100 Gb/s

2 parallel iperf3 instances, 50 threads/iperf3, window size varies from 32 KB to 256 KB

10 Synthetic benchmark on HPC networks

Experimentation with some **synthetic benchmarks** (**no synchronisations**, just sending packets) with various **communication topology, scheduling** and mix of **message sizes** to evaluate potentiel effects on the network. With **DAQPIPE** we are close to the **GatherCmdMeta** case.

Synthetic benchmarks

11 DAQPIPE benchmark on InfiniBand

Evalution with the **DAQPIPE benchmark** using different **versions** and **running modes** on **16** Infiniband EDR nodes. Best result are obtained by using only **one task** (sending or receiving) per process, not mixing the two, cf. **one-way** and **v1**.

DAQPIPE results on 16 nodes