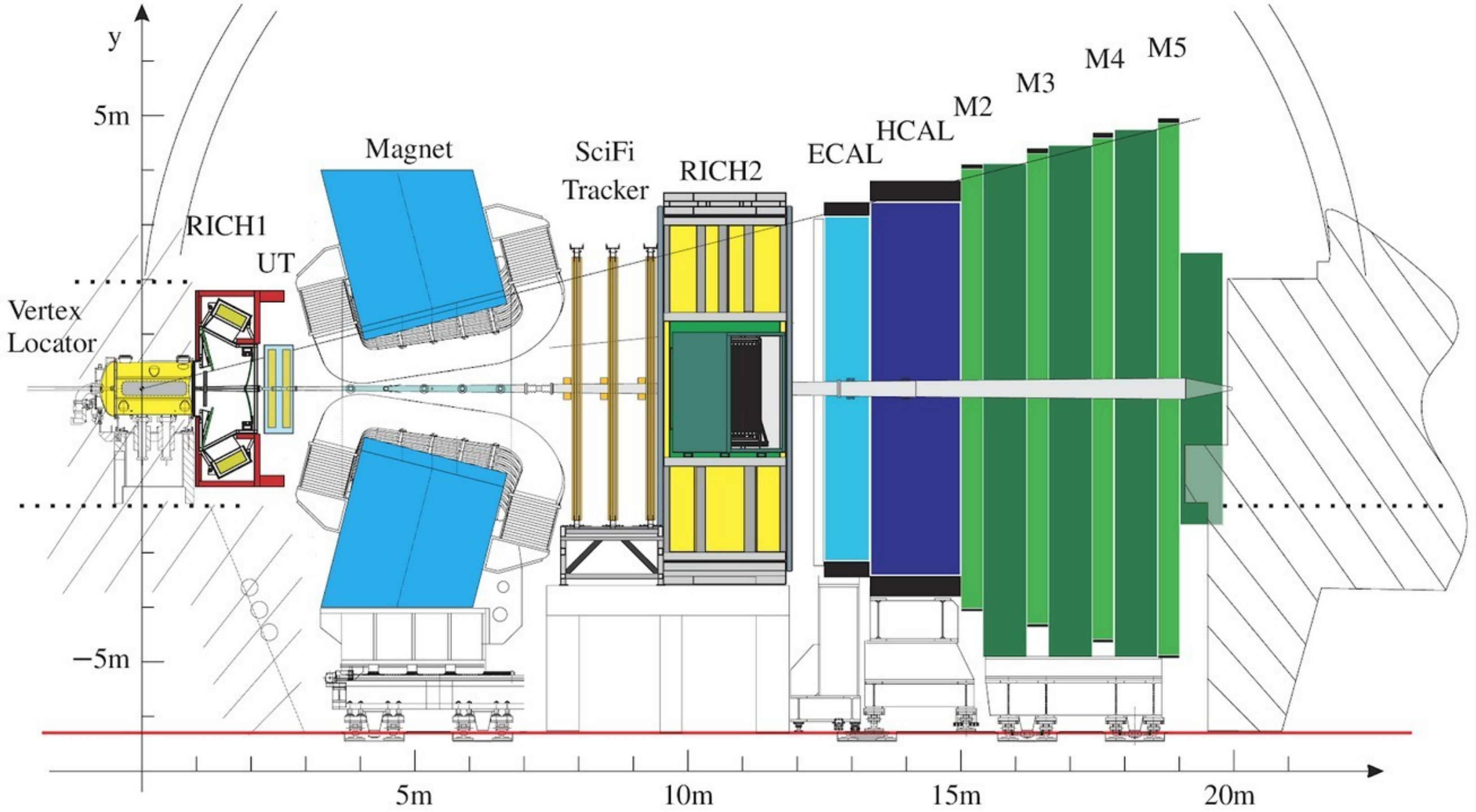
Data distribution over Ethernet for the LHCb filtering farm

Rafał Dominik Krawczyk
rafal.dominik.krawczyk@cern.ch
on behalf of the LHCb Online Team
CERN

vCHEP 2021
18 May 2021

The LHCb experiment

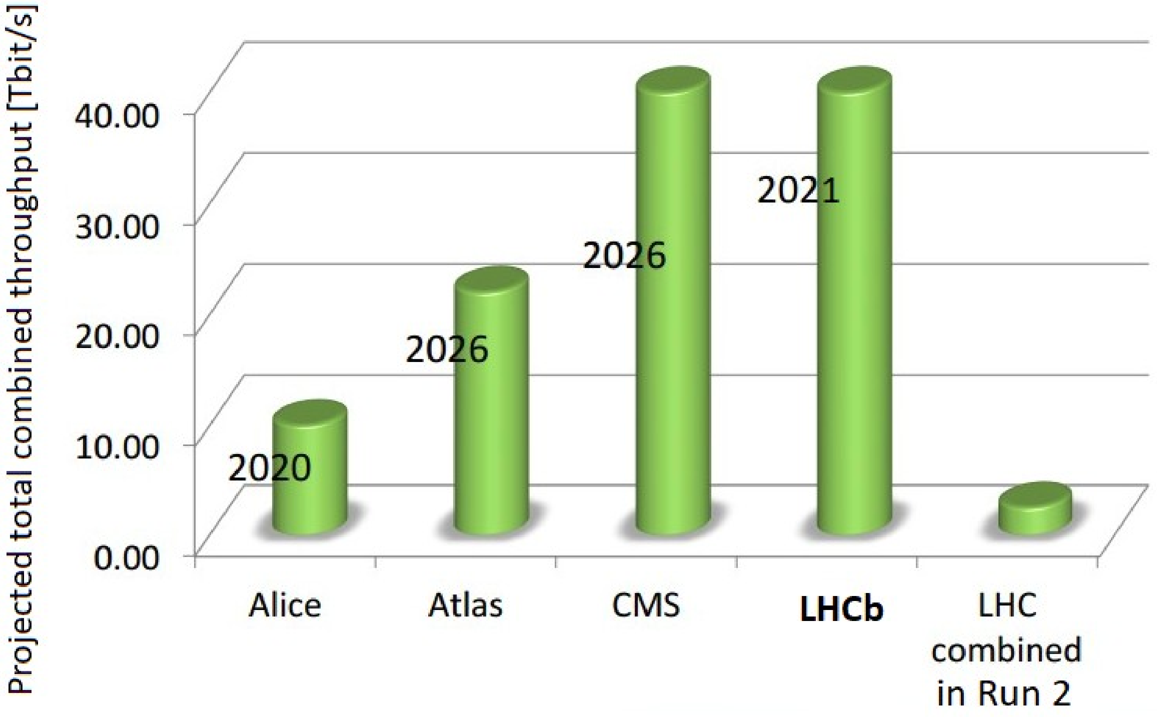
- One of four main LHC detectors
- Purpose: measure CP violations
- p-p bunch crossing rate: 30 MHz
- Luminosity: 2×10^{33} cm^{-2} s^{-1}
- Independent subdetectors
- Have to assemble events first
- Event Building (all-to-all traffic)



Challenge

Abandoned hardware trigger for Run 3:

- LHCb traditional hardware trigger does not profit from higher luminosity



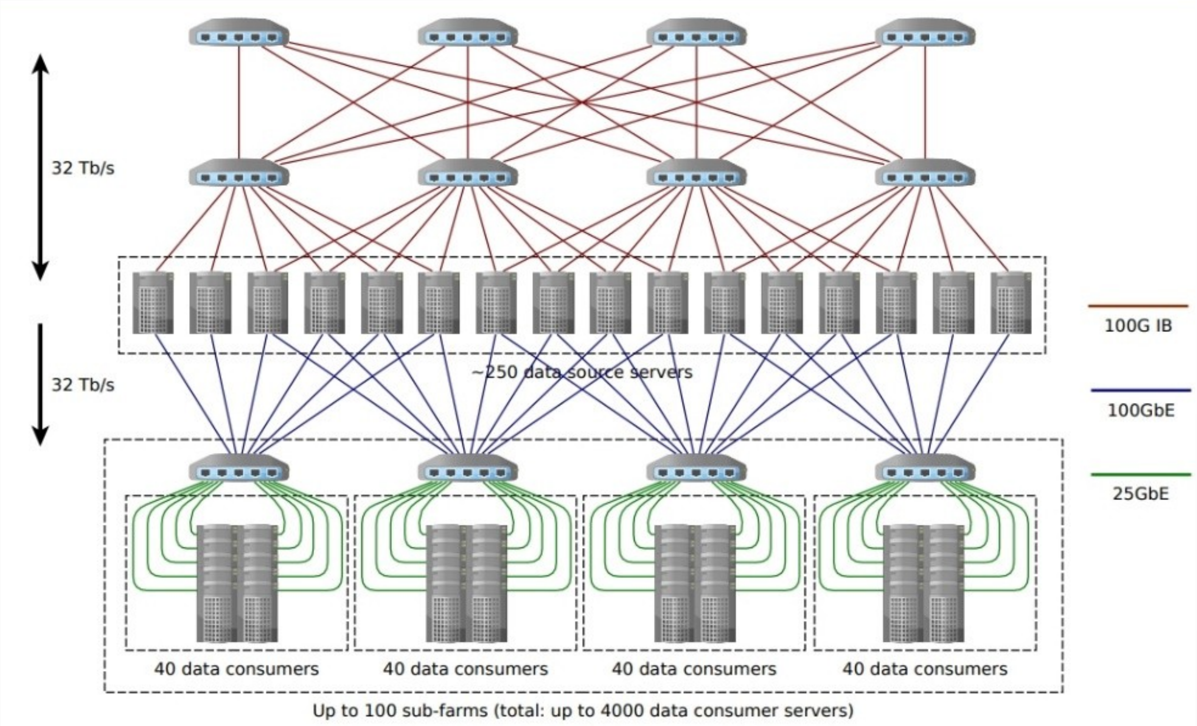
Going trigger-less instead:

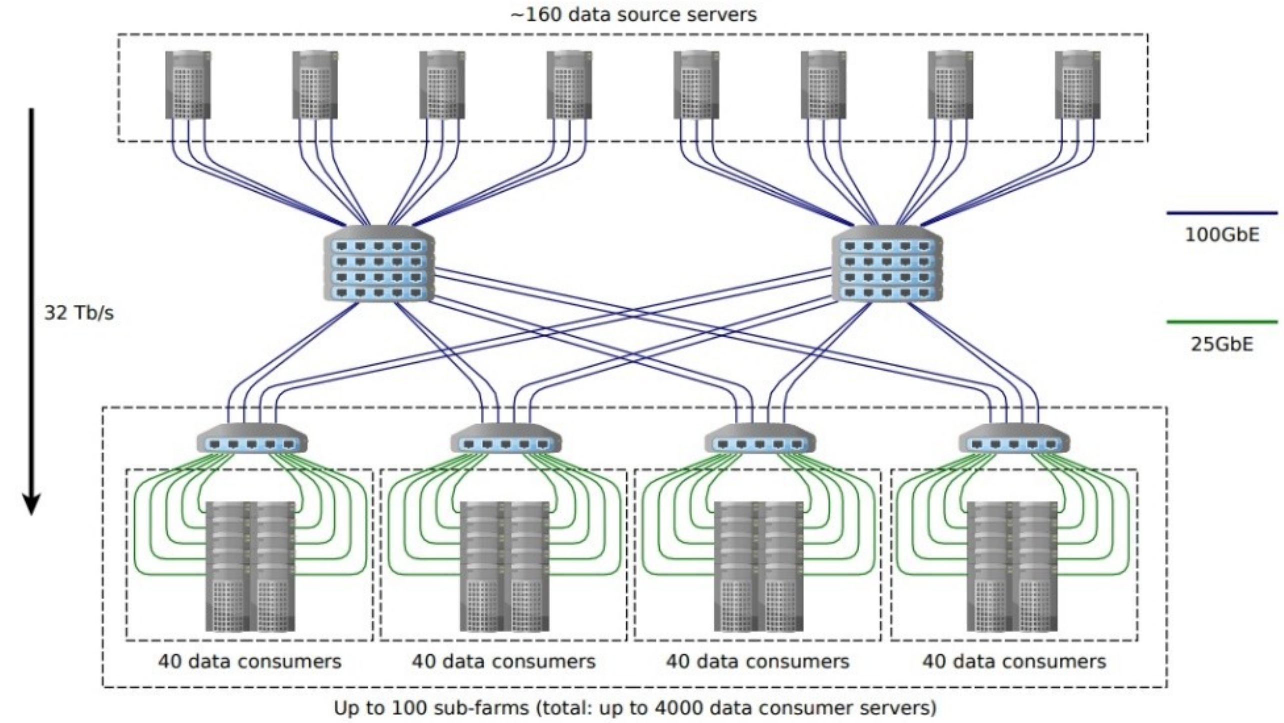
- Detector geometry: fibres/cables not "in the way"
- Relatively low radiation levels → FPGAs in many detector front-ends
- Zero-suppression on the detectors & total event size small (~ 100 kB)
- Software-defined online selection and throughput reduction

- **Commissioning allegedly largest real-time data acquisition (32 Tbit/s) system in the World in 2021**

Evaluation of architectures and fabrics

TWO VARIANTS INITIALLY CONSIDERED





InfiniBand & RoCE Event Building

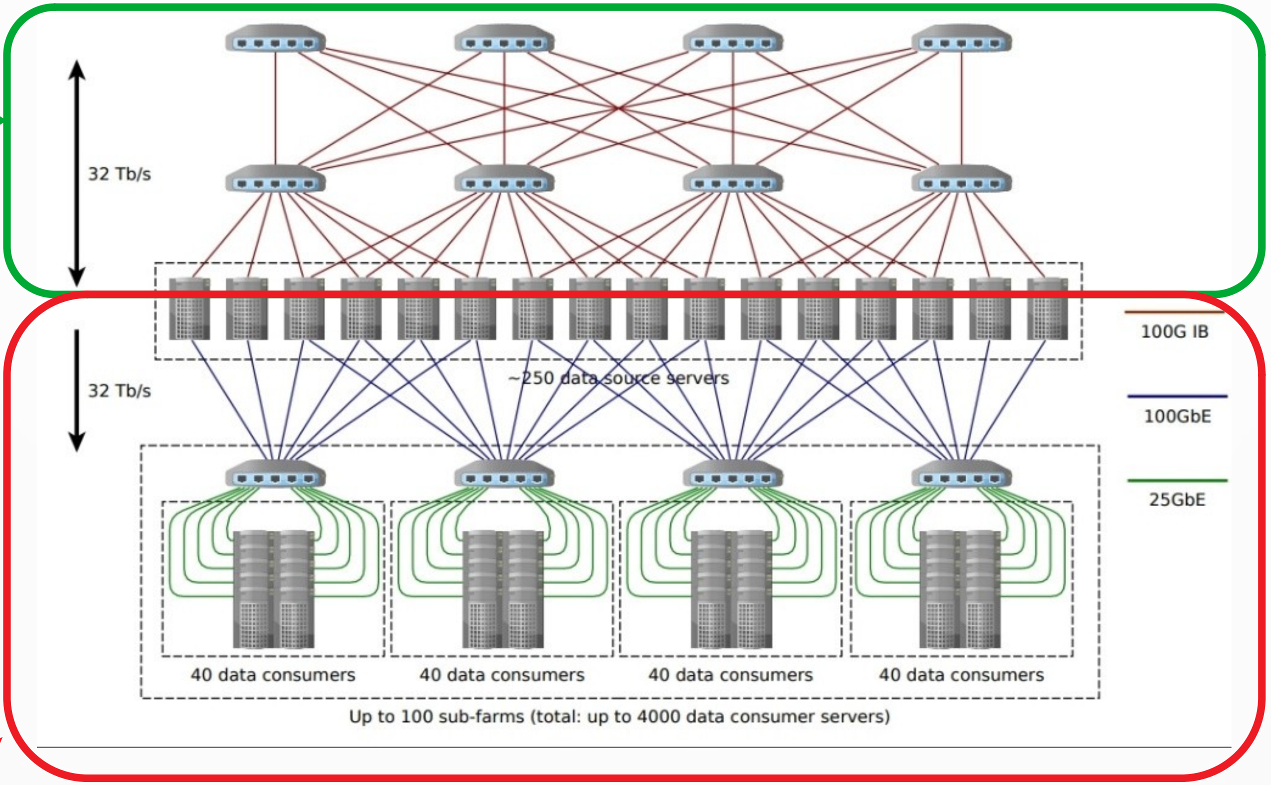
commissioned for Run 3

Ethernet–only Event Building

abandoned for Run 3
(see CHEP19 results)

Ethernet RoCE v2 in commissioned EB

works fine
commissioned

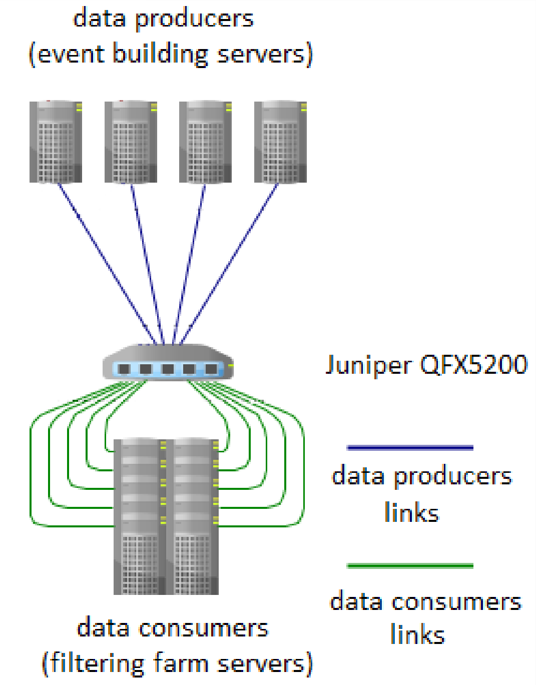


Question: can **this** stage of EB perform well-enough with Ethernet RoCE v2 ?

- Ethernet **is a must** because of combined link speeds and costs
- RoCE v2 → zero-copy protocol with possible flow control
- One-to-many distribution of assembled events, consumers can be temporarily busy

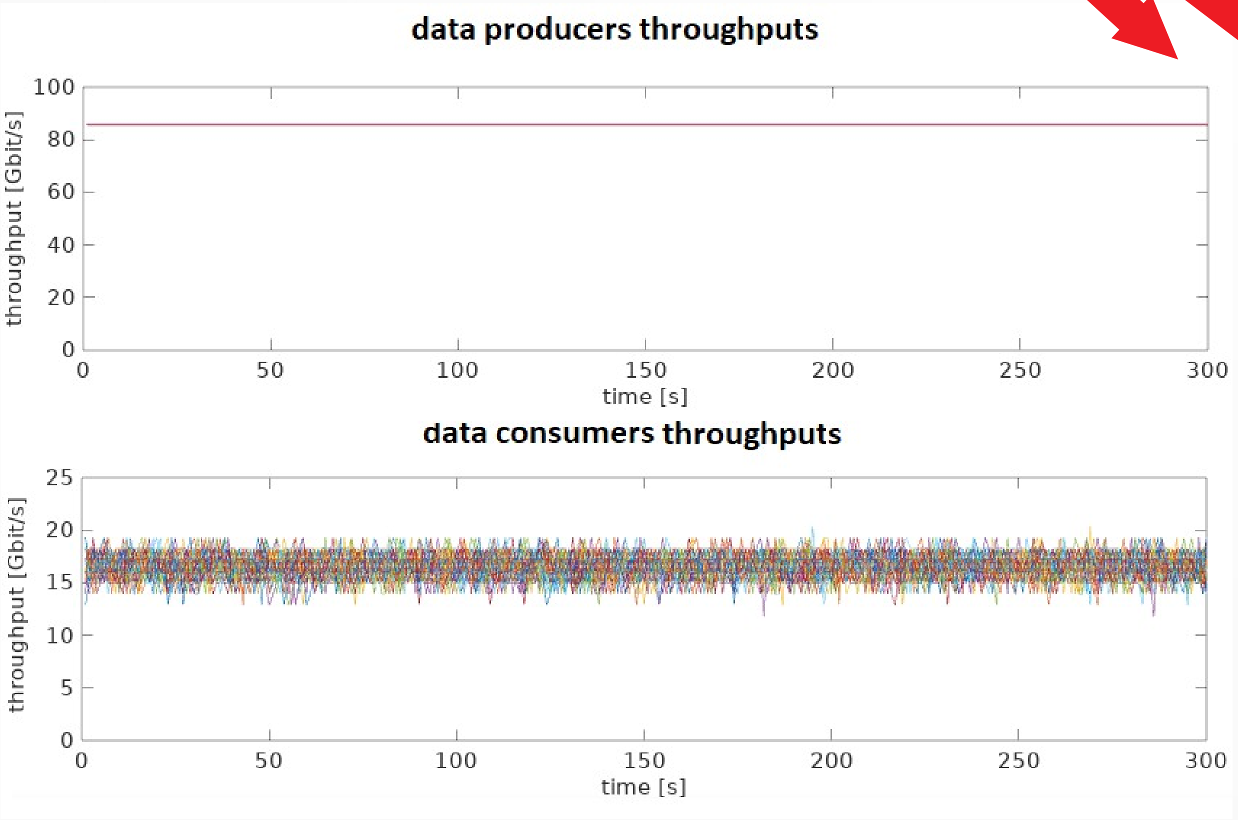
Test bench

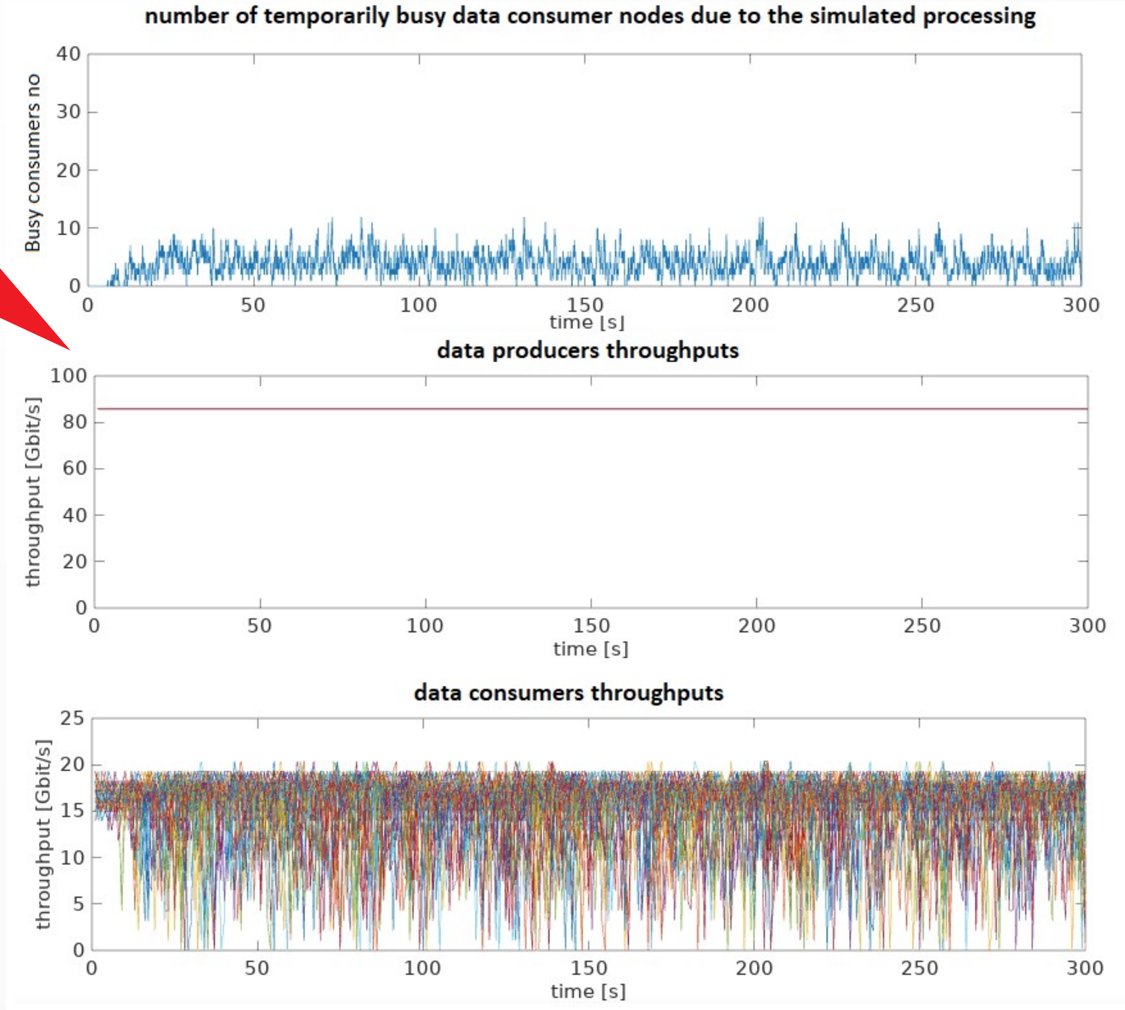
- Small cluster of data producers and data consumers
- Switch → Shallow-buffered Juniper switch
- Producers → 100 Gbit/s links
- Consumers → 2 scenarios tested → 25 or 100 Gbit/s links
- Implemented custom C++ MPI benchmark
  - One-to-many, LHCb-like transmissions
  - Scheduling sends
  - Periodically probing network
  - Simulating temporary data consumers busyness
- **Goal – check if real-time transmissions can be sustained without saturating the buffers → optimally no throughput drops on producers**



Combined 25 and 100 Gbit/s links tests

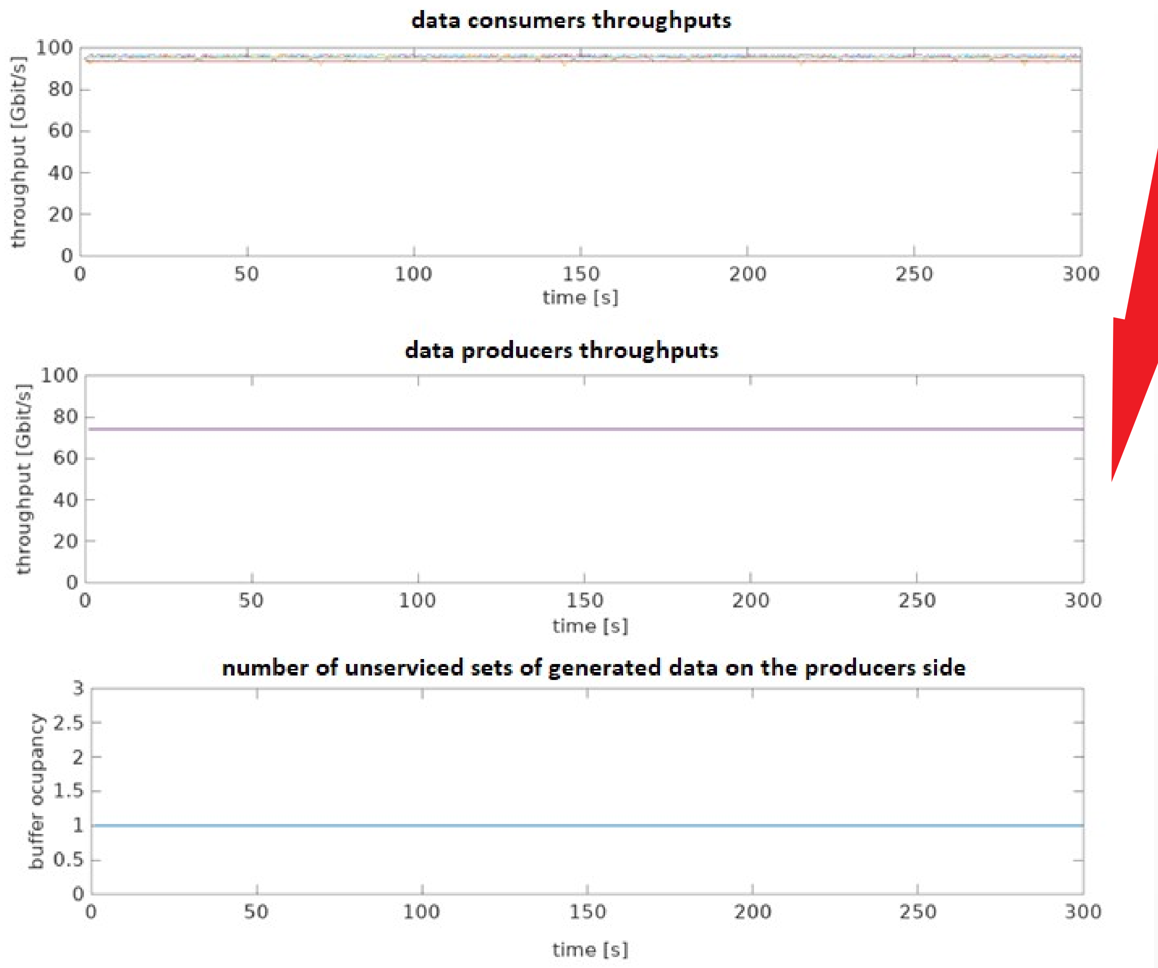
- 14 producers, 72 consumers, 5-minute runs
- Sustained 86 Gbit/s on producers, even with temporary consumers busyness
- **Real-time operation sustained**

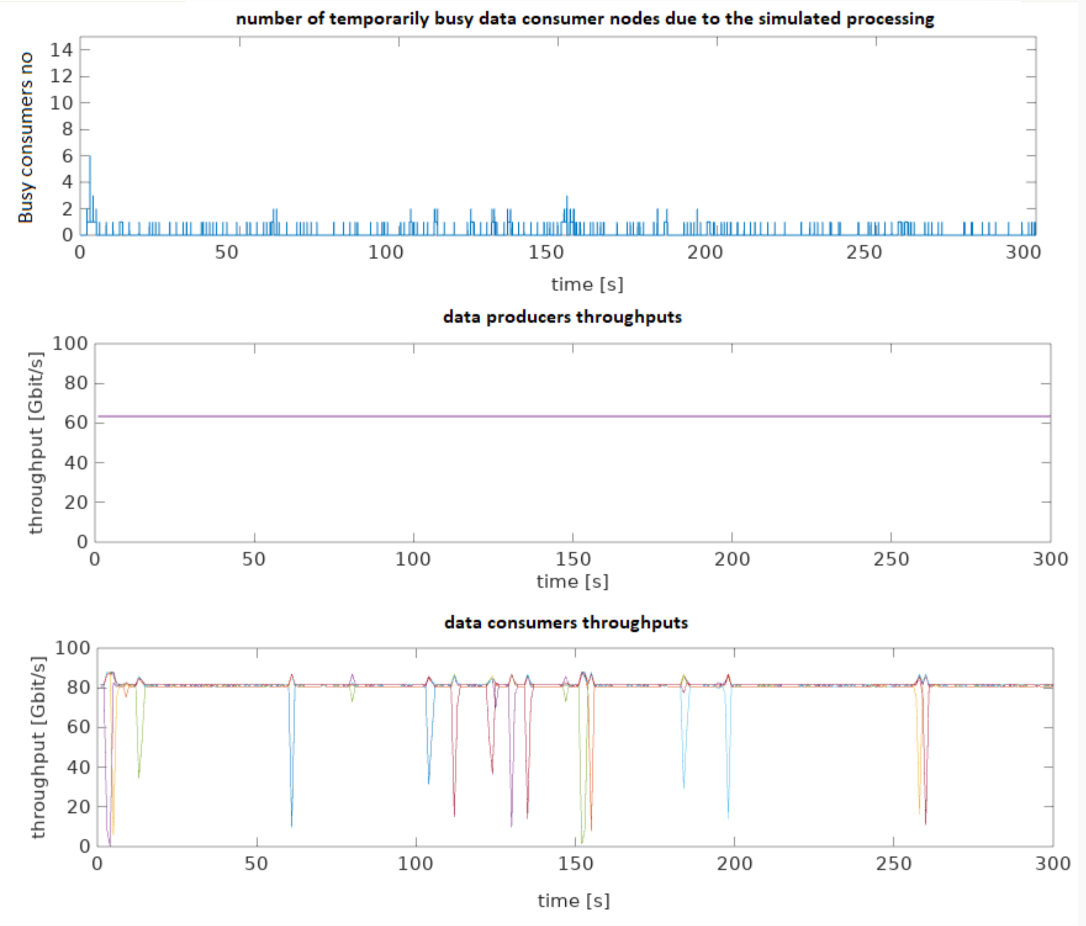




100 Gbit/s links tests

- 18 producers, 14 consumers, 5-minute runs
- Stable stress-test with 74 Gbit/s on producers and 95.2 Gbit/s on consumers
- Real-time operation sustained at 63.3 Gbit/s with temporary consumers busyness





Summary

- Sufficient real-time operation of the LHCb-like traffic over Ethernet
- Hybrid InfiniBand + Ethernet EB applicable
- Proof of concept made → LHCb EB **can** handle 32 Tbit/s readout
- Ethernet RoCE v2 evolution followed → full EB for Run 4 ?

Thank you !