Network simulation of a 40 MHz event building system for the LHCb experiment

05/11/2019

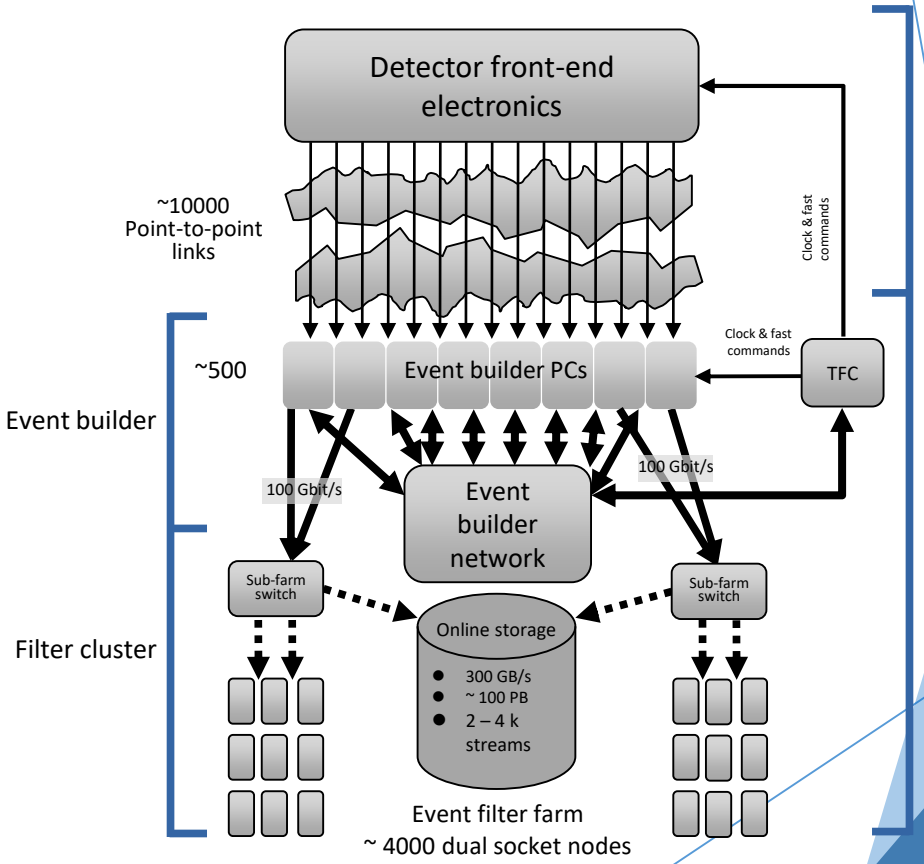
Flavio Pisani

CERN, Università & INFN Bologna

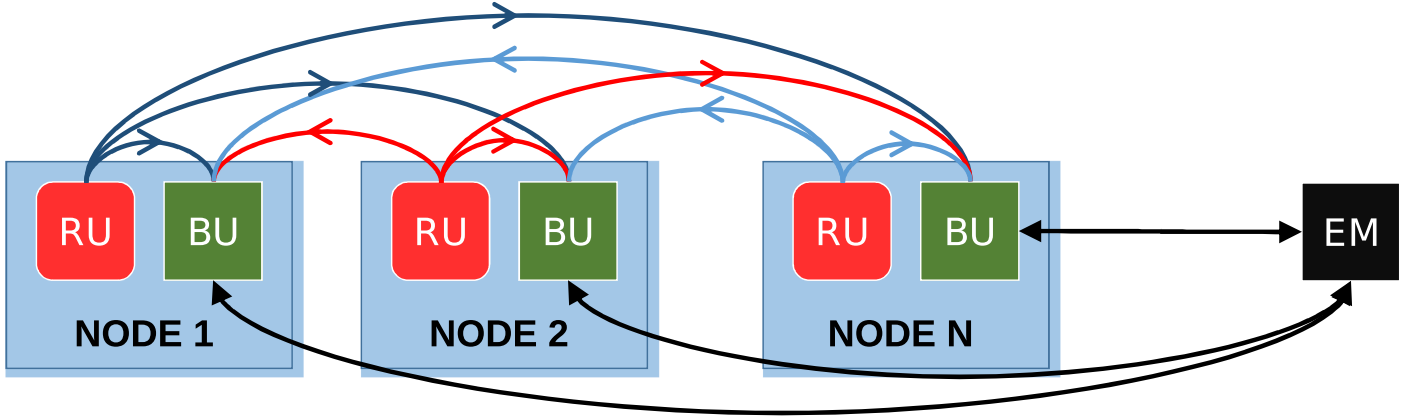
On behalf of the LHCb online group

LHCb DAQ upgrade

- Full software trigger @ 40MHz
- Two separated networks
  - Event building
    - 500 ports @ 80 Gb/s IN/OUT
  - Filter farm
    - 500 ports @ 80 Gb/s OUT
    - 2000 port @ 20 Gb/s IN
- Candidates for event building network:
  - Mellanox InfiniBand (EDR/HDR)
  - 100 Gb/s Ethernet(?)
- Candidate for filter farm network:
  - 25/50/100 Gb/s Ethernet



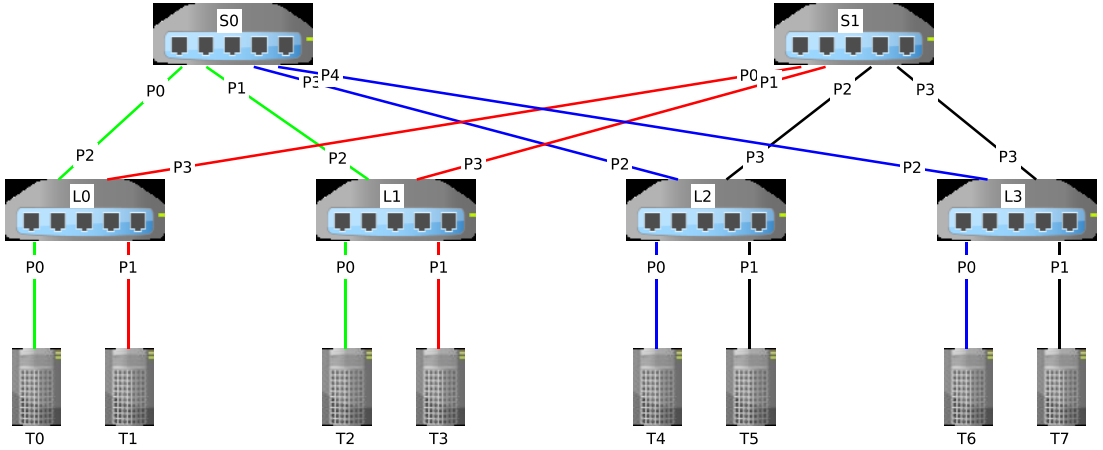
The event building process



- Every event is divided into multiple fragments
- Every **Readout Unit (RU)** receives a fragment of the event
- Every **Builder Unit (BU)** has to gather all the fragments of the event
- The **Event Manager (EM)** arbitrates the transfer

The many-to-one nature of the traffic generates network congestion

The linear shifting scheduling



- The processing of N events is divided into N phases
- In every phase one RU sends data to one BU, and every BU receives data from on RU
- During phase n RU x sends data to BU (x+n)%N
- All the units switch synchronously from phase n to phase n+ 1

Congestion-free traffic on “selected networks” (i.e. non blocking networks)

Event building benchmarks

DAQPIPE

- Linear shifting-like traffic generator
- No strong synchronization is enforced
- Multiple events are processed by every BU (credits)
- Multiple sends in parallel for every credit (parallel sends)

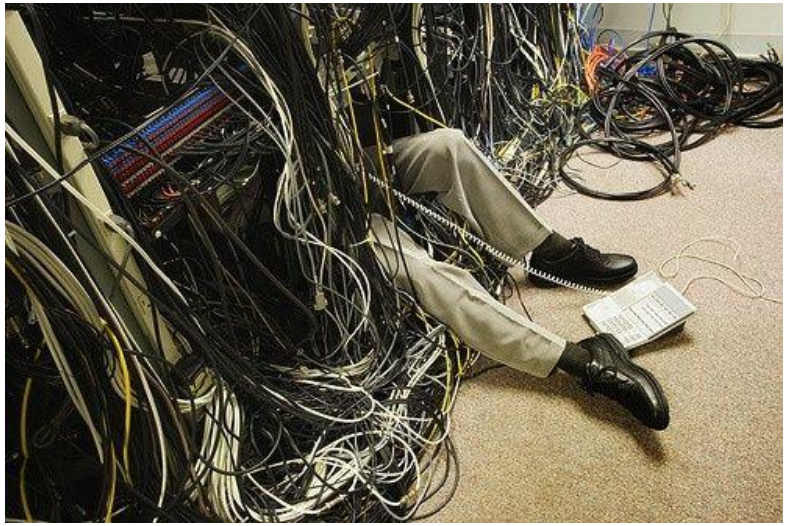
Requires good congestion handling

a2a

- Synchronous linear shifting traffic generator
- Phase shift is triggered by a fixed time window
- Idle period between every phase change to prevent desynchronization

Requires a conflict-free network

Testing on real systems



- The availability of large scale EDR InfiniBand test system is limited
- The network configuration is usually optimised for HPC
- Modification to the network configuration are usually impossible

Low level network simulations are a powerful tool

The OMNeT++ framework

What is OMNeT++?

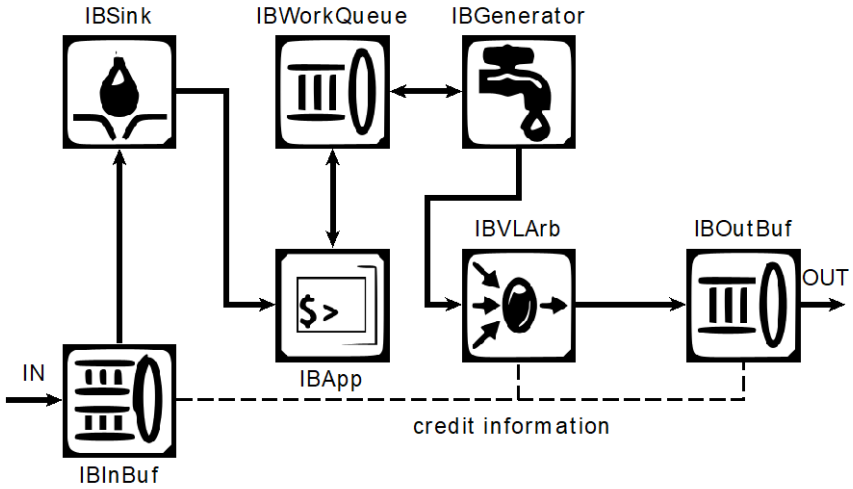
- A discrete event simulation Framework written in C++

What does it offer?

- A modular object-oriented structure flexible and configurable
- A scripting language for easy definition of network topologies
- The possibility of collecting statistics during the simulation
- Native support to parallel processing through OpenMPI

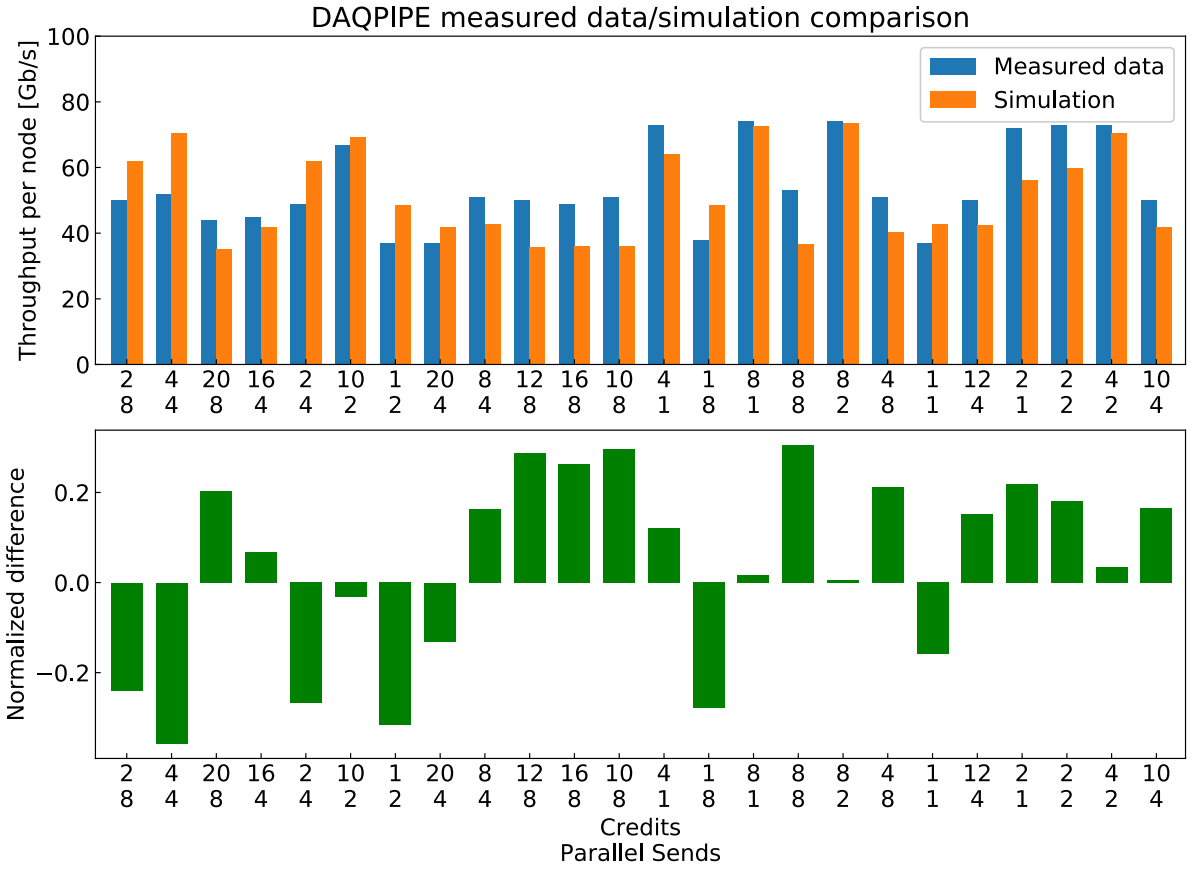
The simulation model

- Initial low level InfiniBand model developed by Mellanox Technologies
- Update of the model to match EDR InfiniBand specifications
- Measure of key features of real InfiniBand EDR hardware (software stack latency, link layer latency, switch buffer size, ecc.)
- Implementation of DAQPIPE and a2a traffic injectors

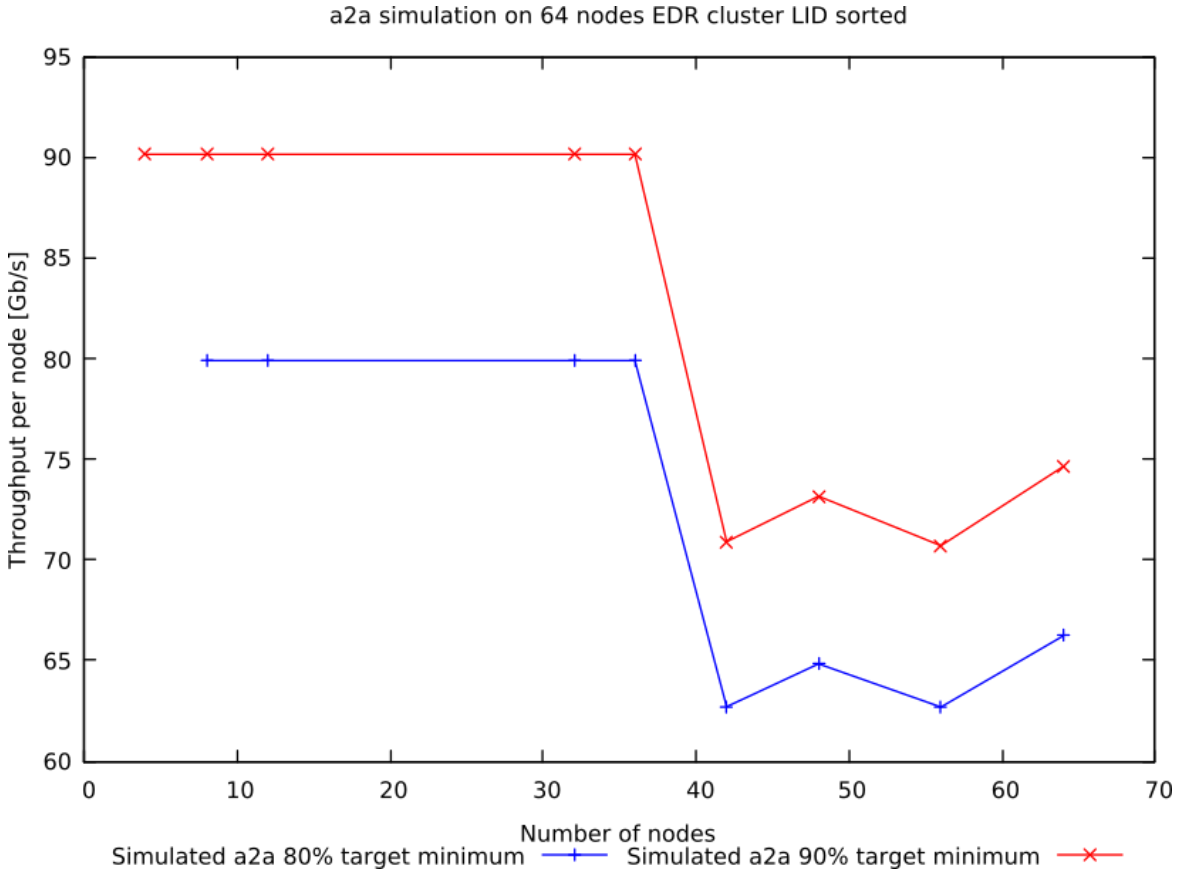


Schematic representation of an end node

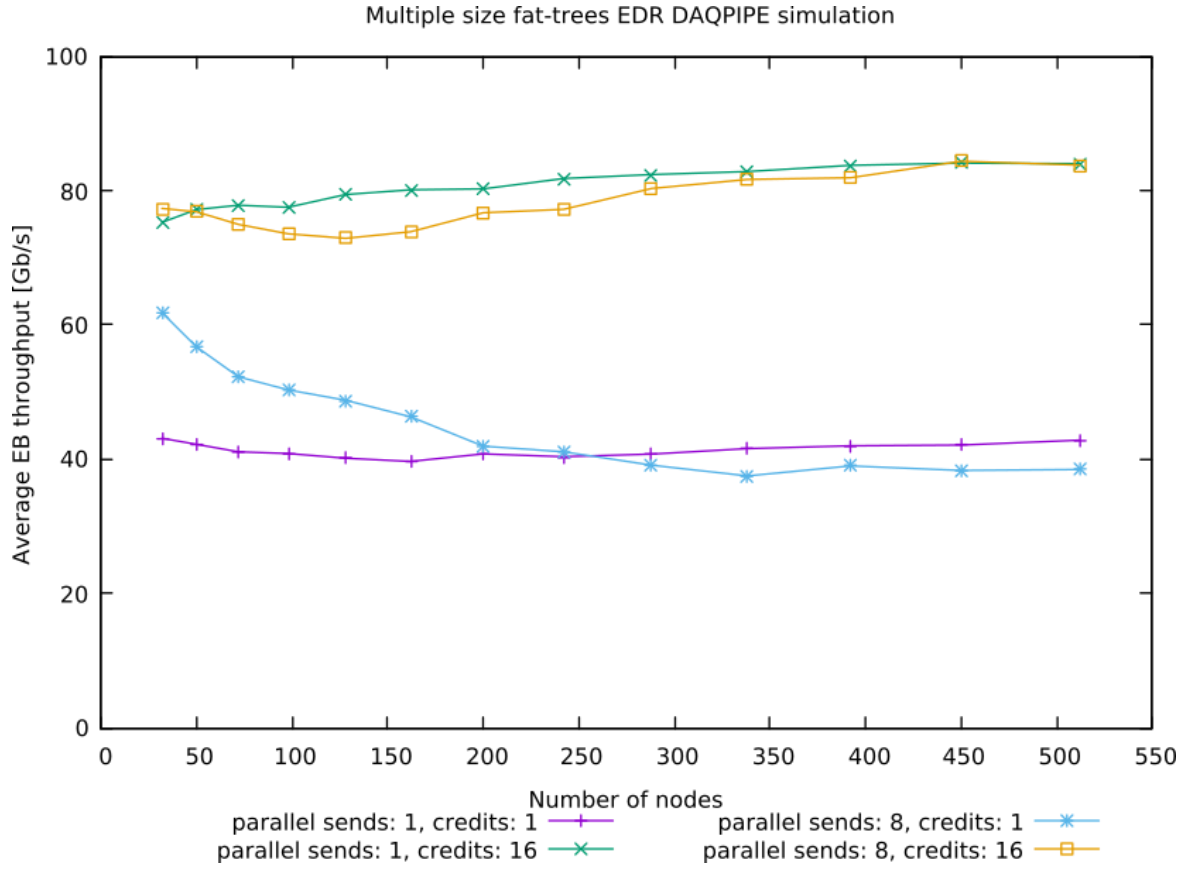
Validation of the simulation model



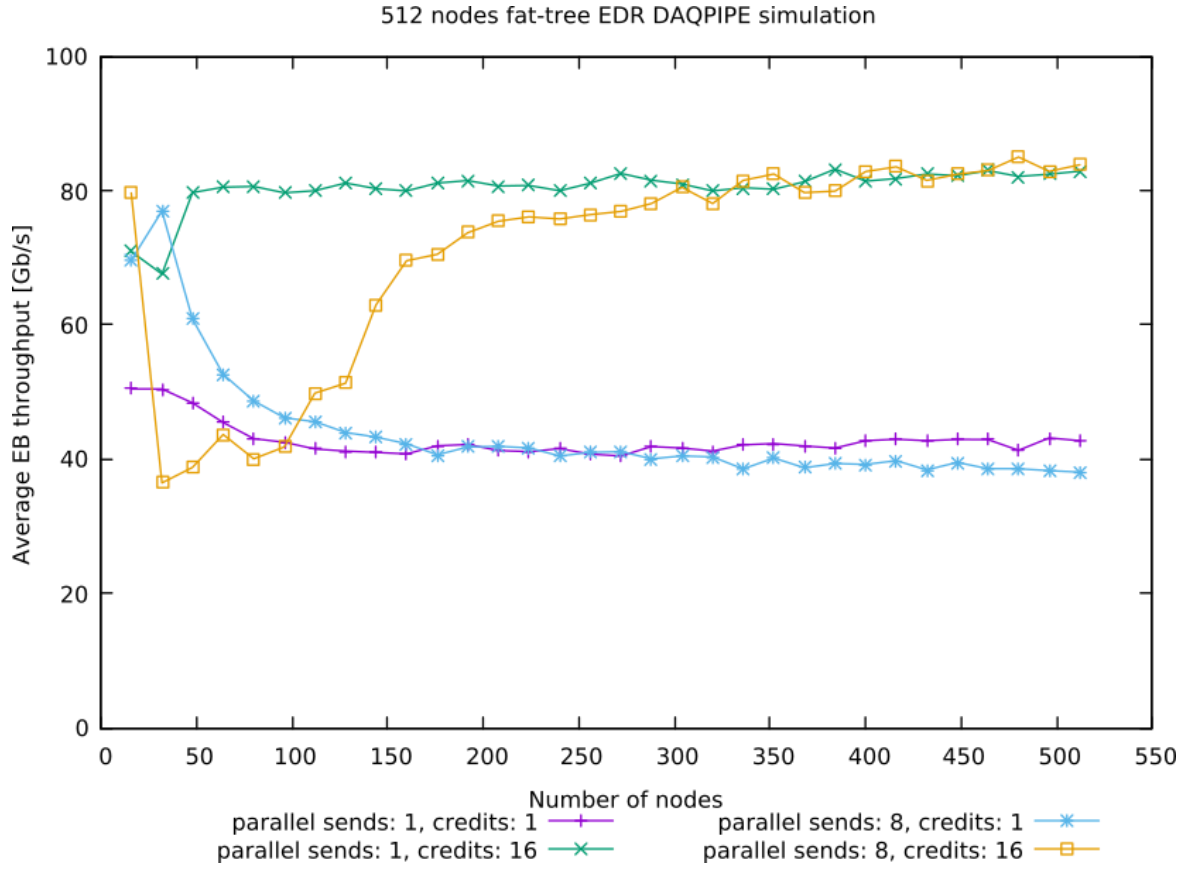
A2a congestion test



Full scale test



Full scale test



Conclusions

- A low-level InfiniBand simulation model has been implemented
- Event builder-like traffic injectors have been added to the simulation framework
- The accuracy of the model has been validated against real data from HPC clusters
- The impact of suboptimal network configuration has been evaluated
- The scalability of the full system has been simulated
- In all the simulated scenarios the network can deliver the target throughput of 80 Gb/s

THANK YOU FOR YOUR ATTENTION

Further reading

- DAQPIPE wiki
  - https://gitlab.cern.ch/lhcb-daqpipe/lhcb-daqpipe-v2/wikis/home
- Simulation model
  - https://doi.org/10.1109/TNS.2019.2905993
  - https://gitlab.cern.ch/flpisani/ib_flit_sim
- Status of the Ethernet testing
  - Poster 475 - 07/11/19 - 15:30 by Rafal Krawczyk