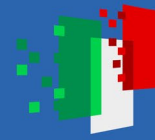




Finanziato
dall'Unione europea
NextGenerationEU



Ministero
dell'Università
e della Ricerca



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA



Future
Artificial
Intelligence
Research

October 19 - 25, 2024

CHEP
2024

Conference on Computing in
High Energy and Nuclear Physics

October 19-25, 2024, Kracow, Poland



**Real-Time implementation of Artificial Intelligence compression
algorithm for High-Speed Streaming Readout signals.**

Fabio Rossi (presenter), **Marco Battaglieri**
Istituto Nazionale di Fisica Nucleare
Genova (Italy)

Edoardo Ragusa, Paolo Gastaldo
SEALab Università di Genova
Genova (Italy)

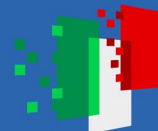
Gagik Gavalian
Jefferson Lab
Newport News (Virginia)



Finanziato dall'Unione europea
NextGenerationEU



Ministero dell'Università e della Ricerca



Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA



Future Artificial Intelligence Research

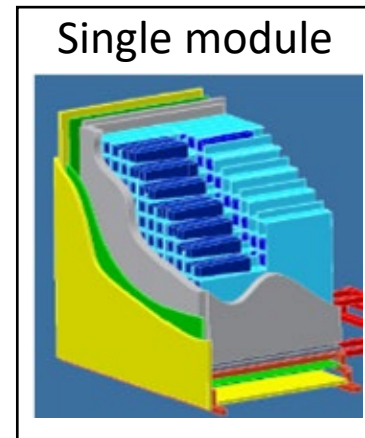
High Energy Physics Experiment: Beam Dump eXperiment (BDX)

Jefferson Lab

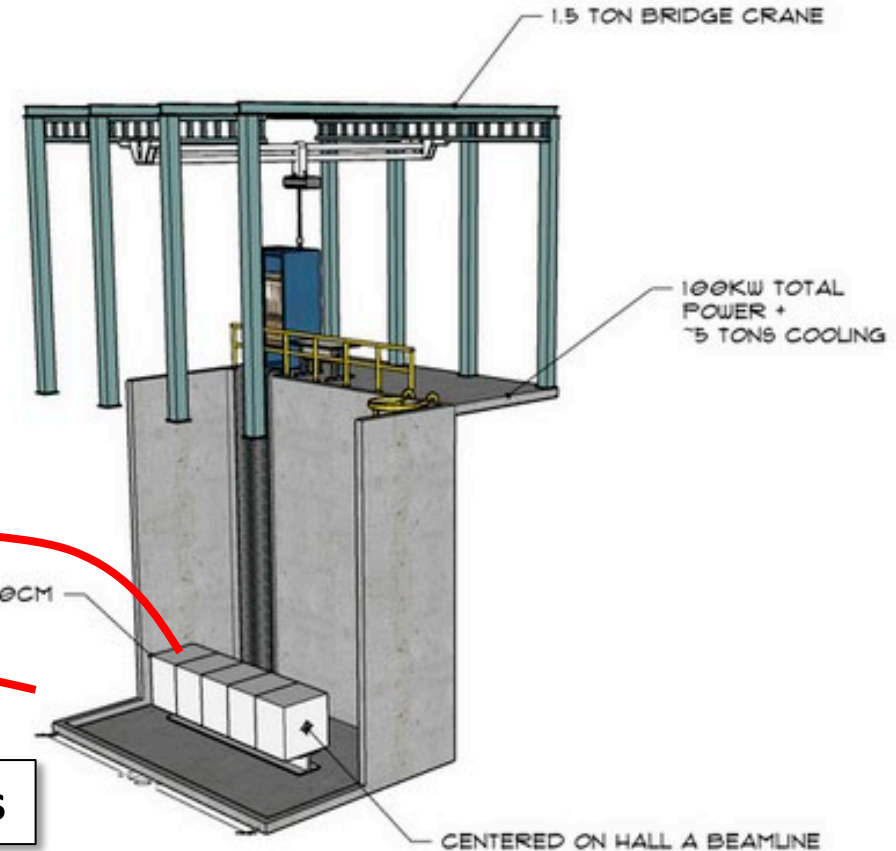


≈ 1000
Calorimeter channels
(30MB/s)

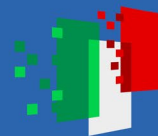
≈ 300
Veto channels
(500MB/s)



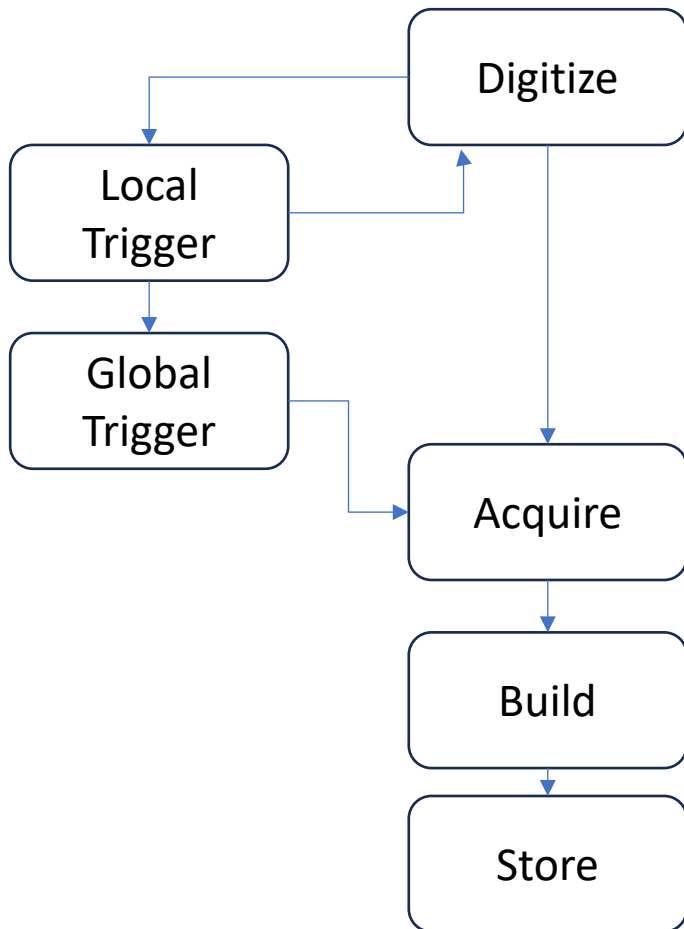
5 MODULES AT 80CM X 100CM X 100CM



Very rare occurrence of Dark Matter events



Traditional triggered DAQ VS Streaming Readout SRO



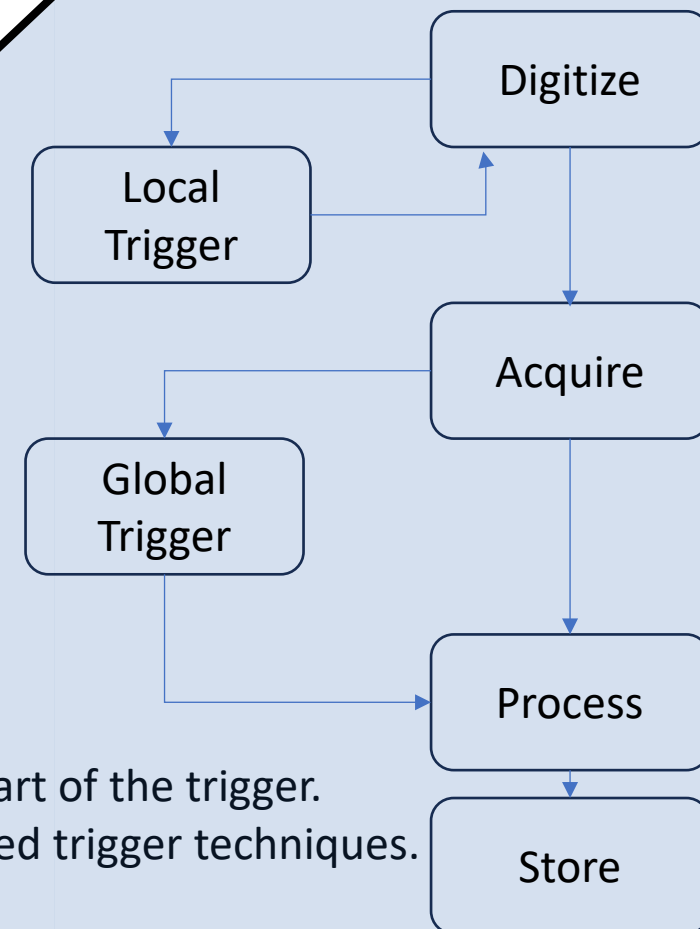
Cons:

Only few information form the trigger.
Trigger logic difficult to implement and debug.
Not easy to adapt to different condition.

Pros:

It works reliably.

Triggered
Streaming



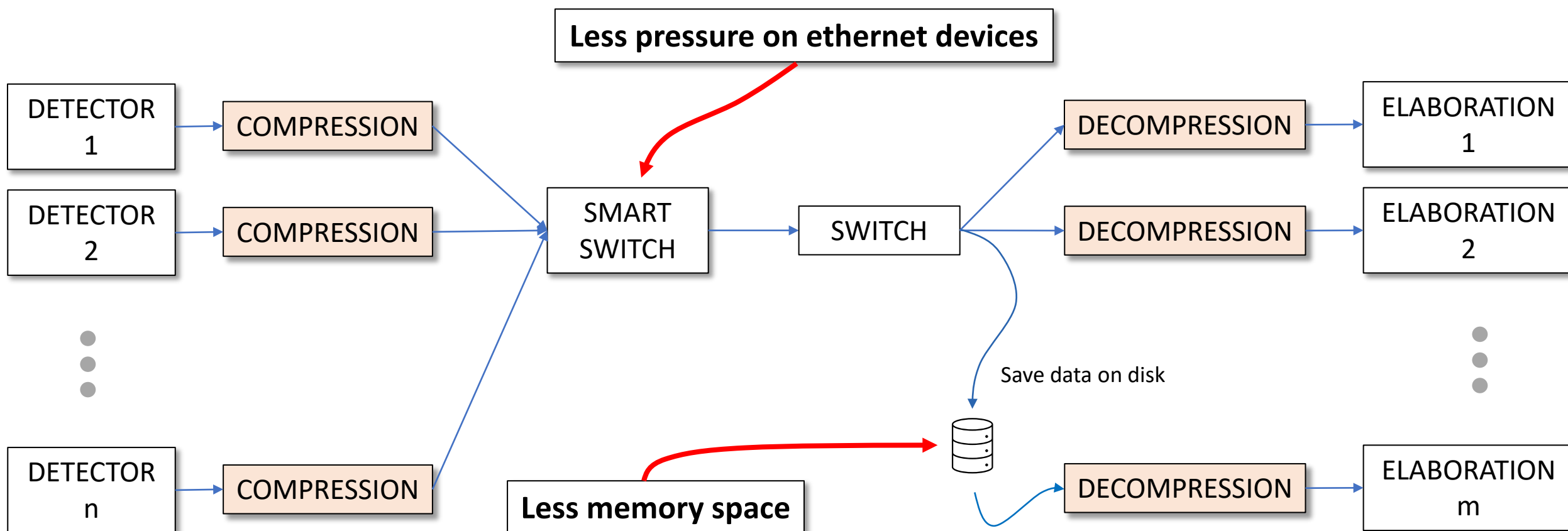
Cons:

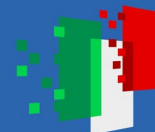
High data rate.
New design.

Pros:

All channels can be part of the trigger.
High level sophisticated trigger techniques.
Software trigger.

Block scheme of data flow





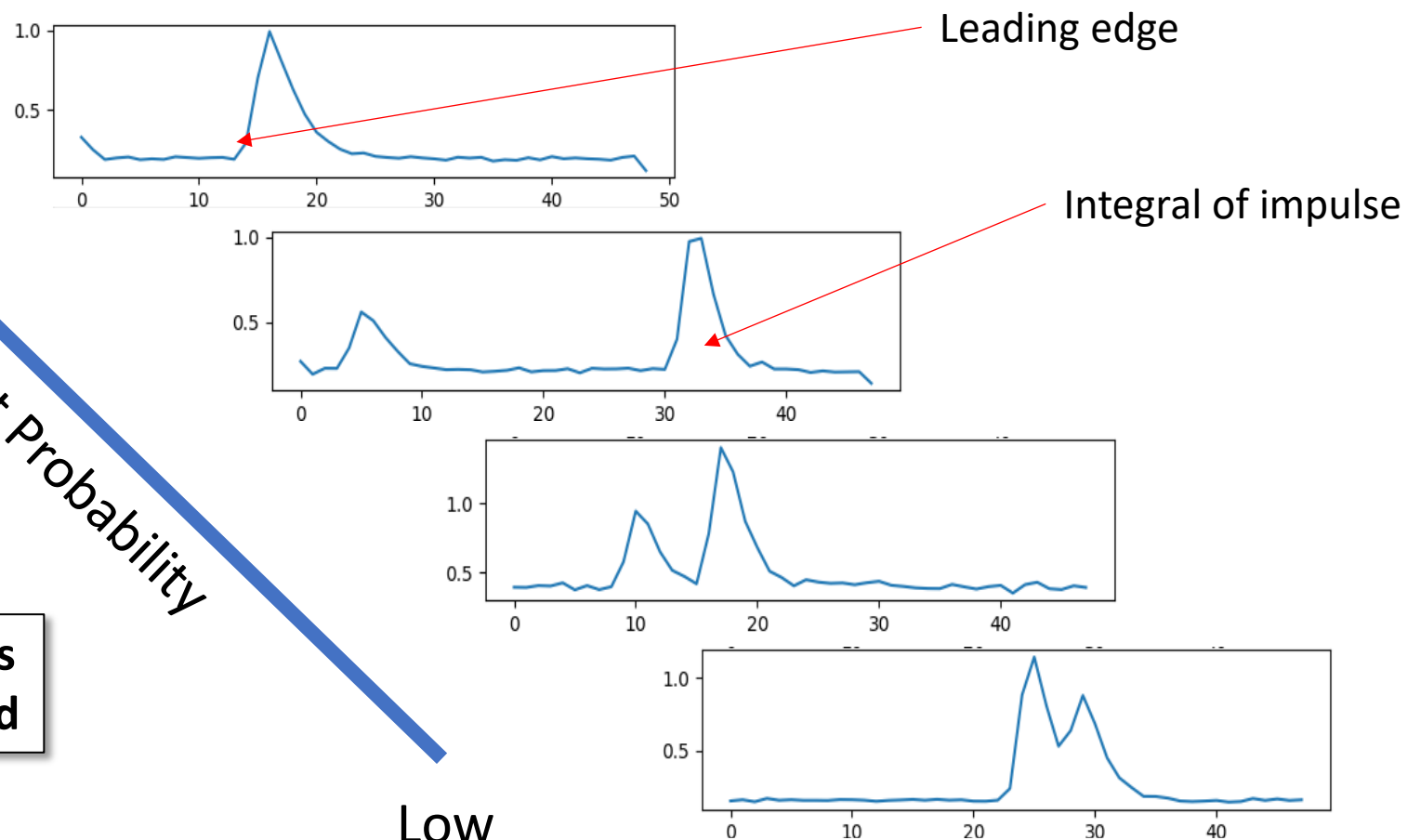
Data from physical Experiment

High

Event Probability

Very-Low probability signals could be sent uncompressed

Low



Autoencoder Definition

Machine Learning Algorithm

Artificial Neural Network

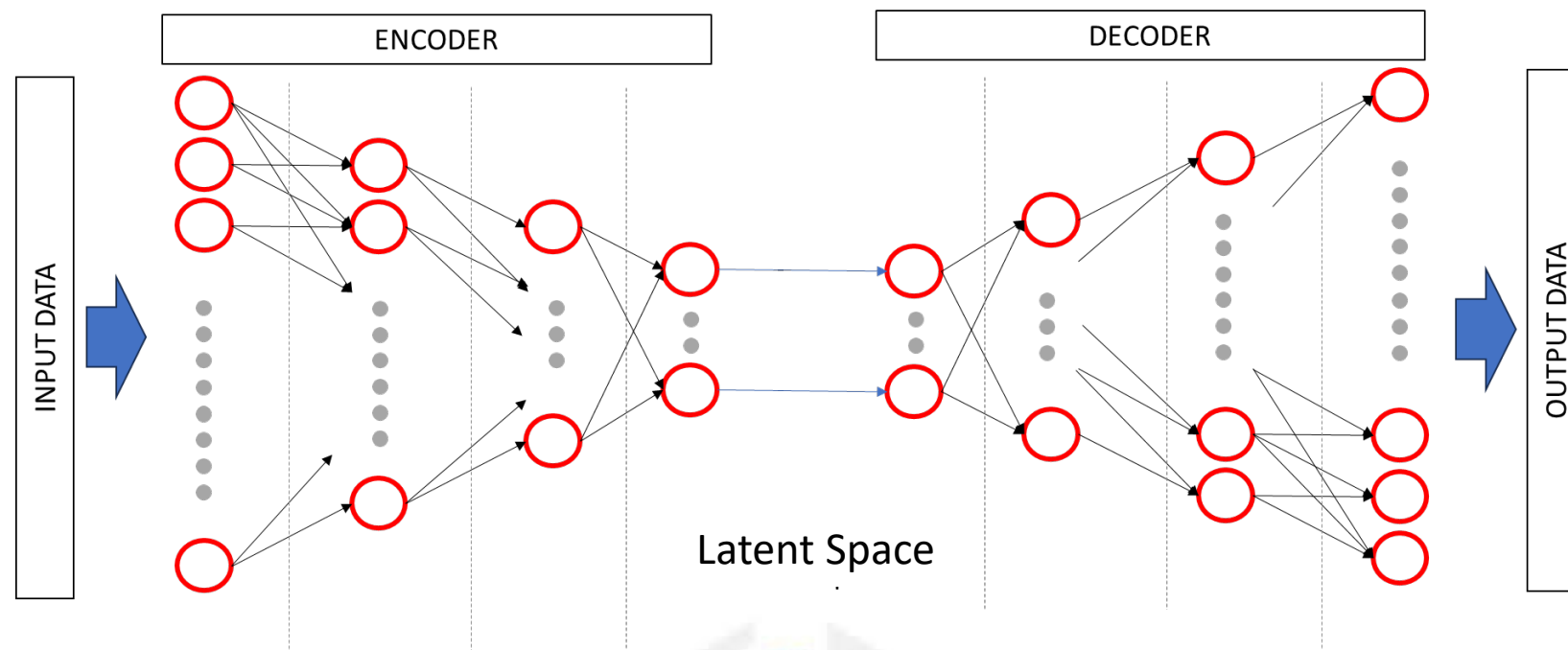
Unsupervised learning

Dimensionality reduction

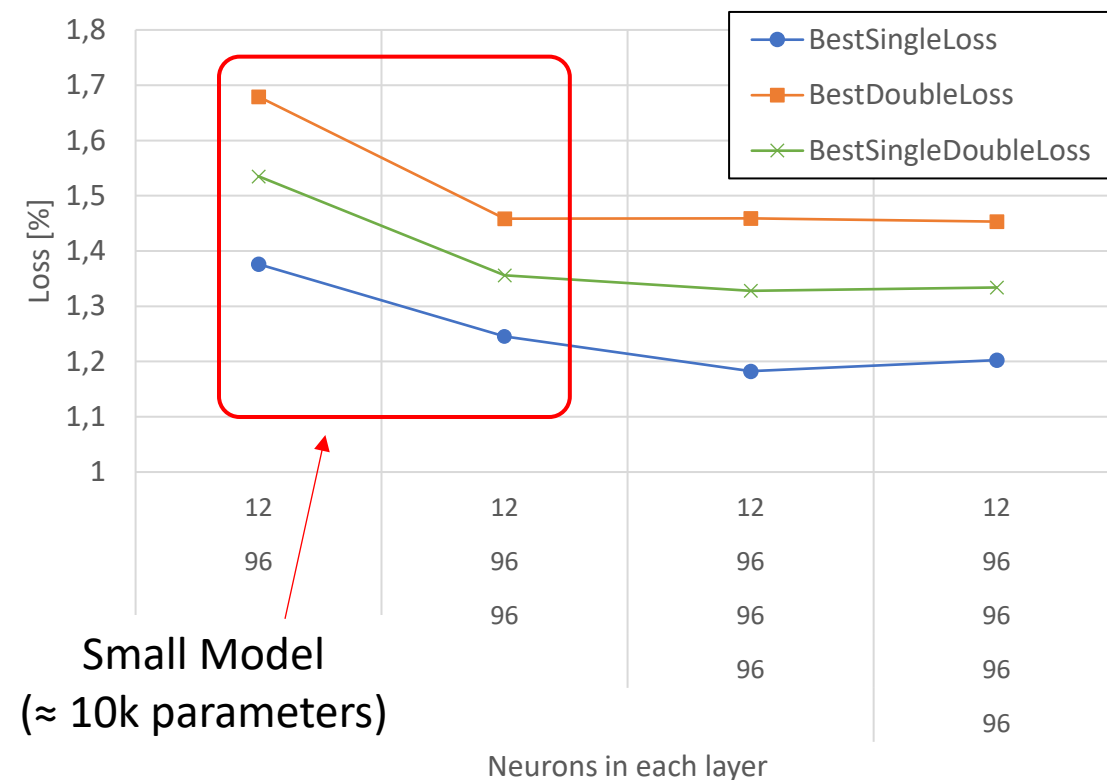
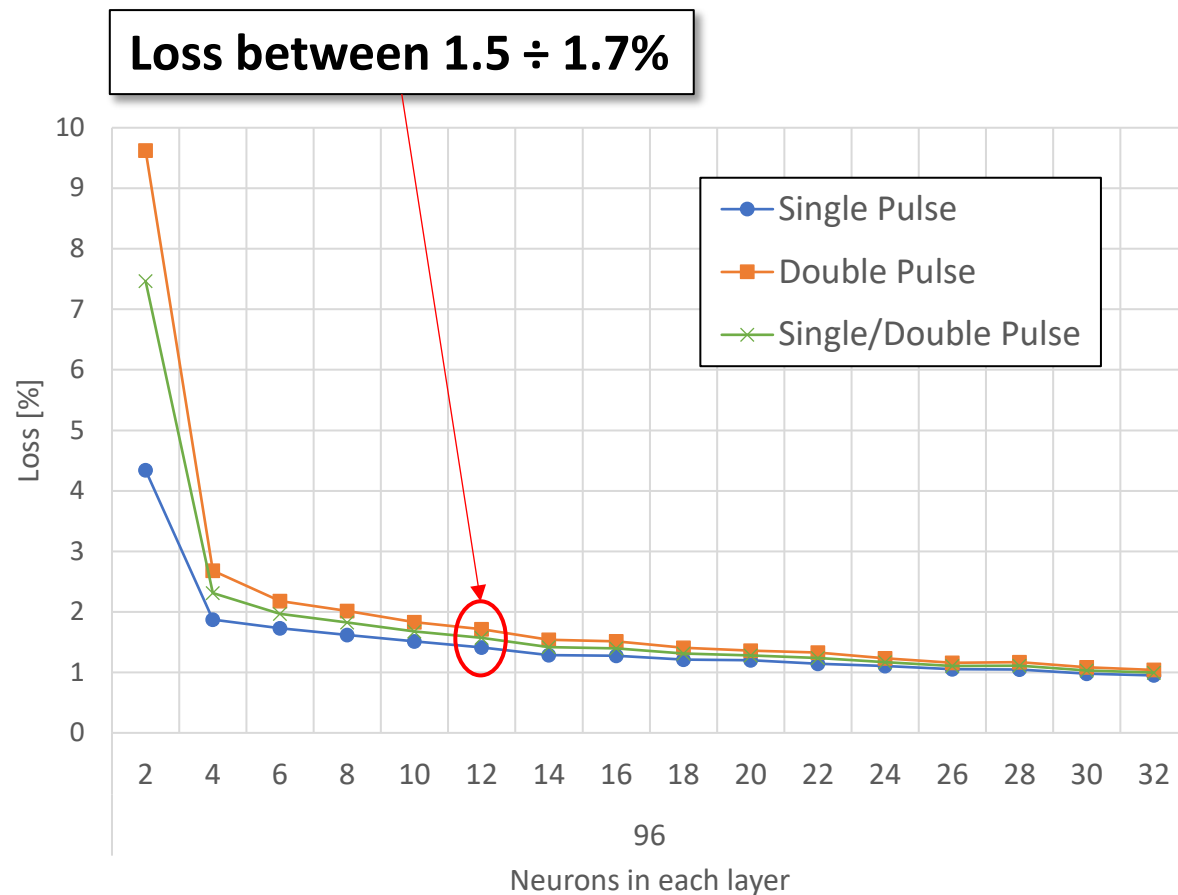
Composed of two function:

- encoding
- decoding

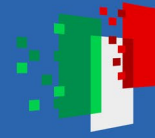
FULLY CONNECTED AUTOENCODER WITH DENSE LAYER



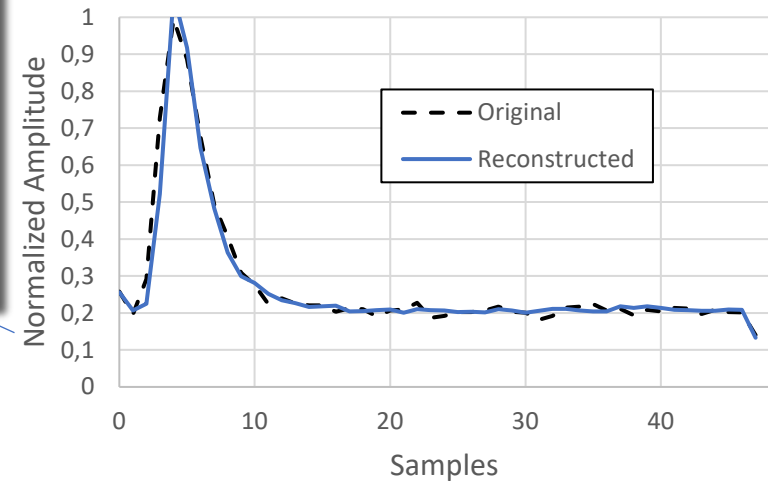
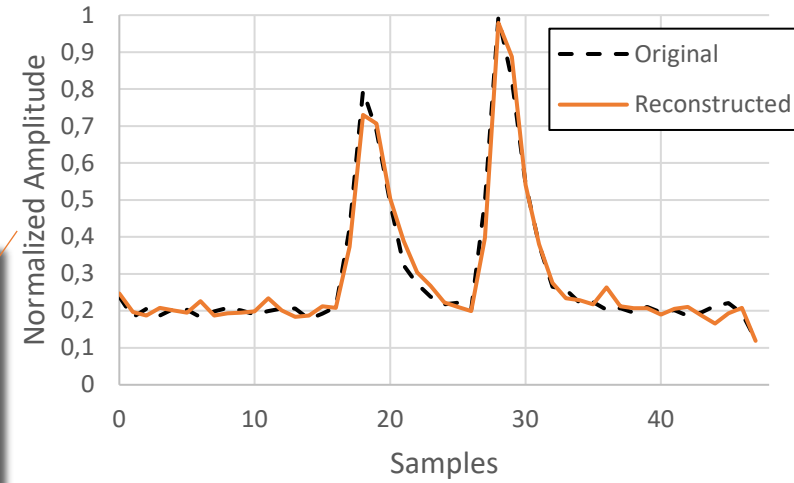
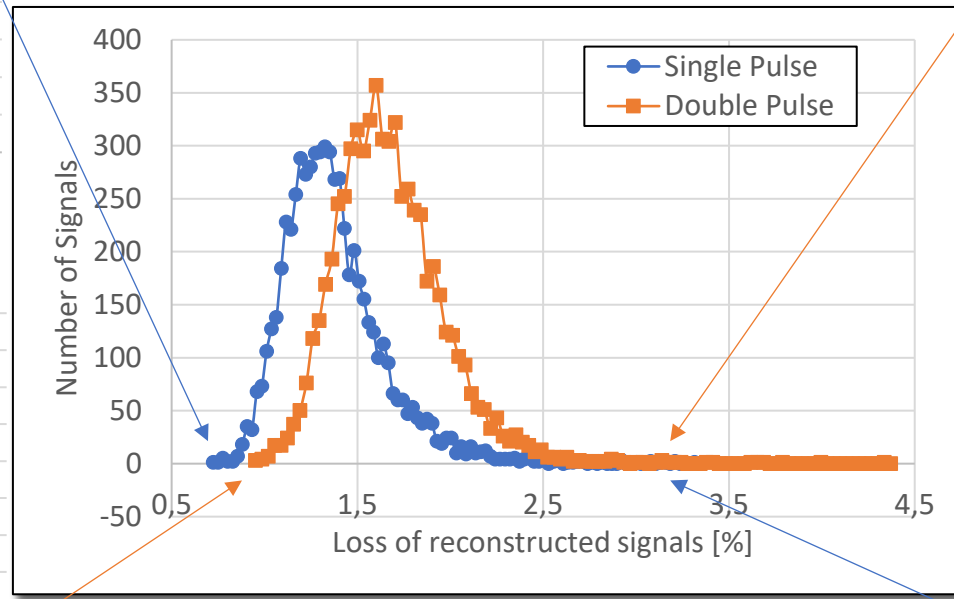
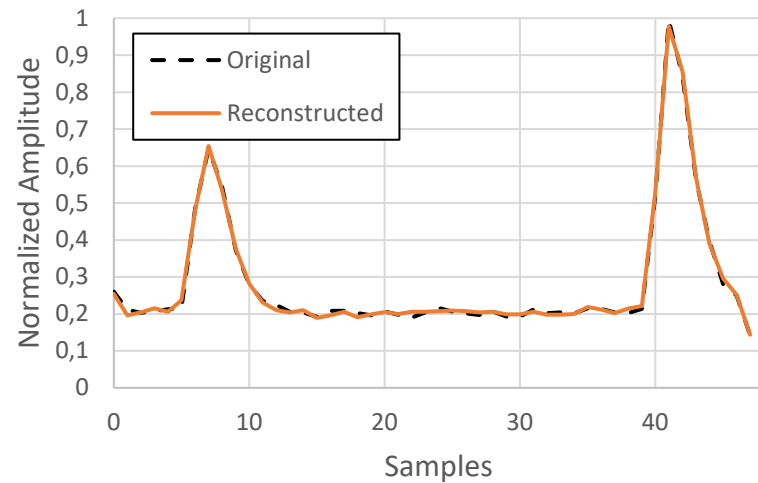
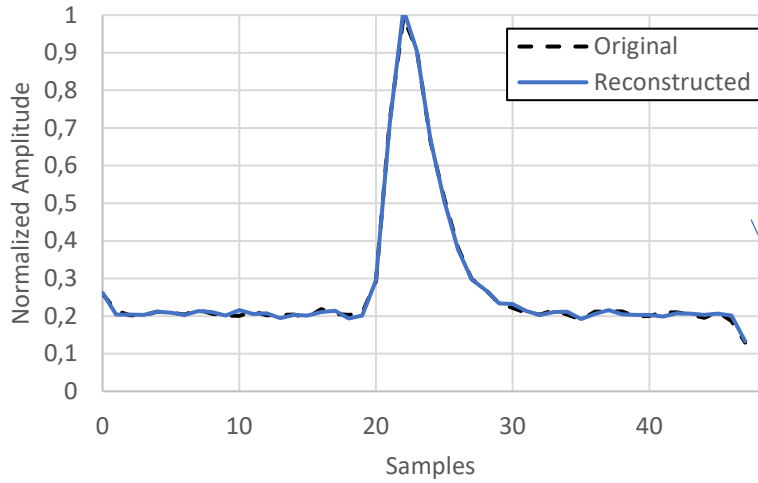
Training of the Autoencoder model



More detail in: Rossi, F., et al. "Artificial intelligence data reduction algorithm for streaming readout in high energy physics experiment." 2024 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)



Signals Compression

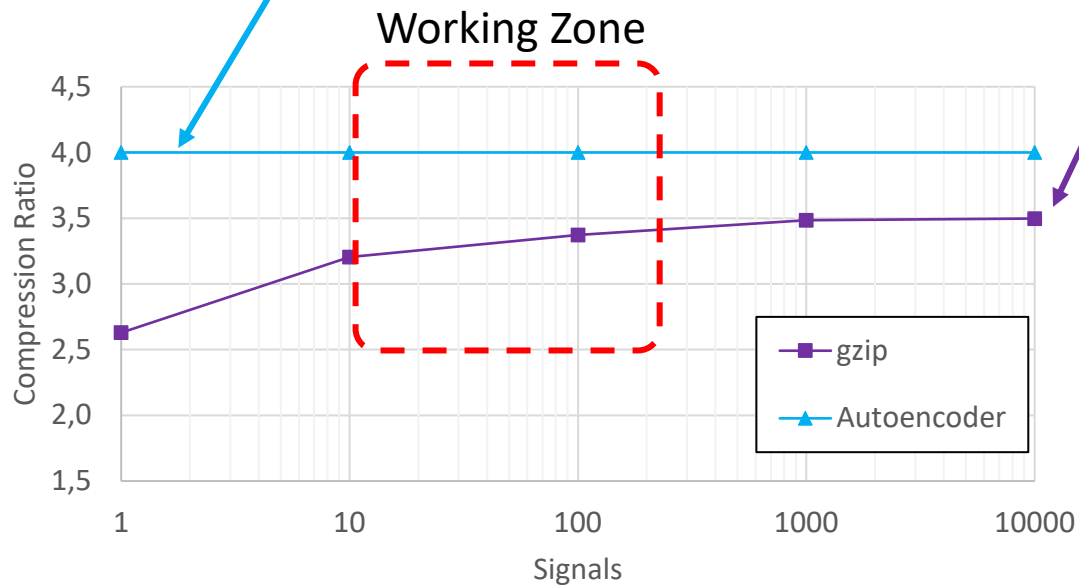




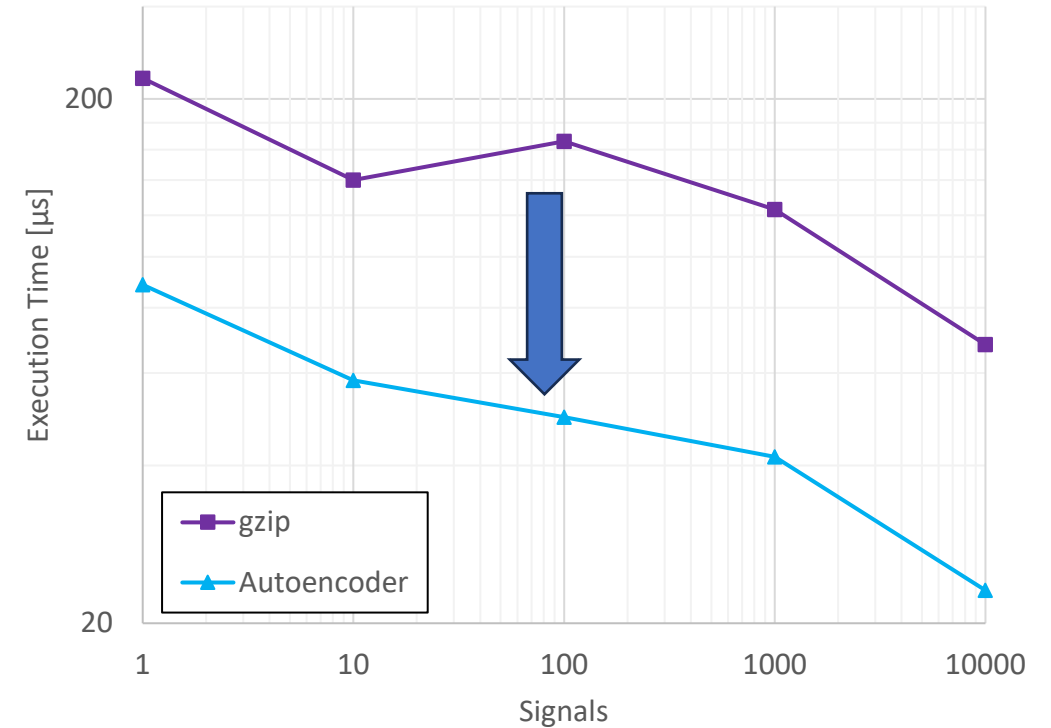
Comparison with standard lossless compression

Autoencoder compression ratio is a Parameter

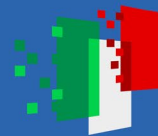
Gzip compression ratio depends on signals number



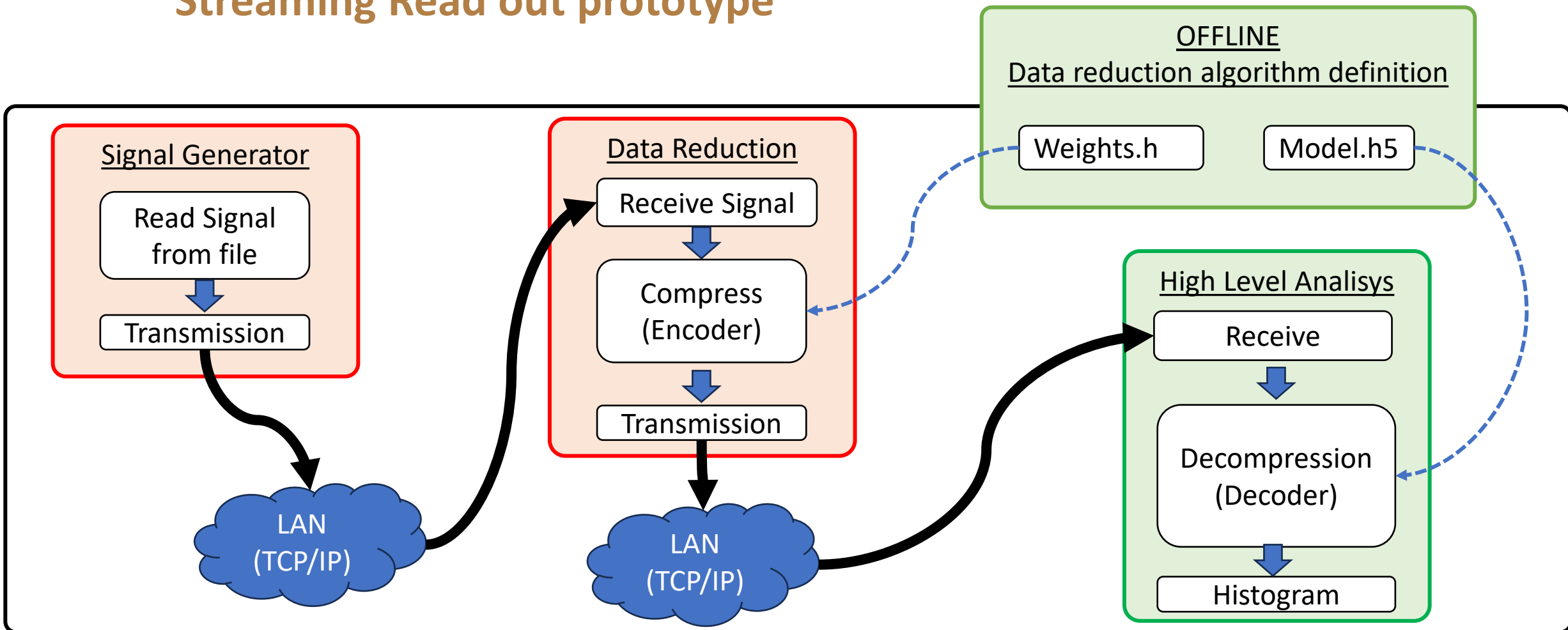
Better compression ratio

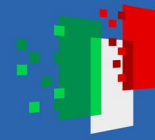


Better also on execution time



Streaming Read out prototype





Implementation of Data Reduction Node

4 x NVIDIA Tesla V100 GPU



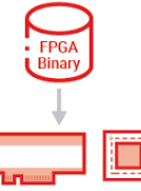
Data Reduction

Receive Signal

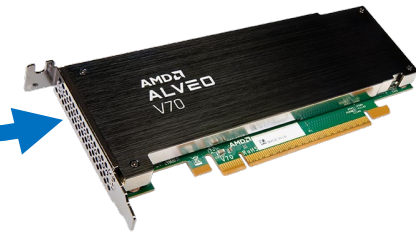
Compress (Encoder)

Transmission

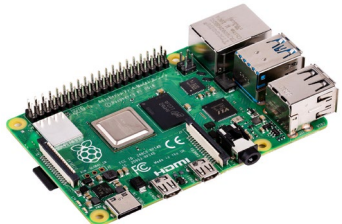
Xilinx XRT



ALVEO V70 FPGA



Raspberry Pi 4 Rev. B

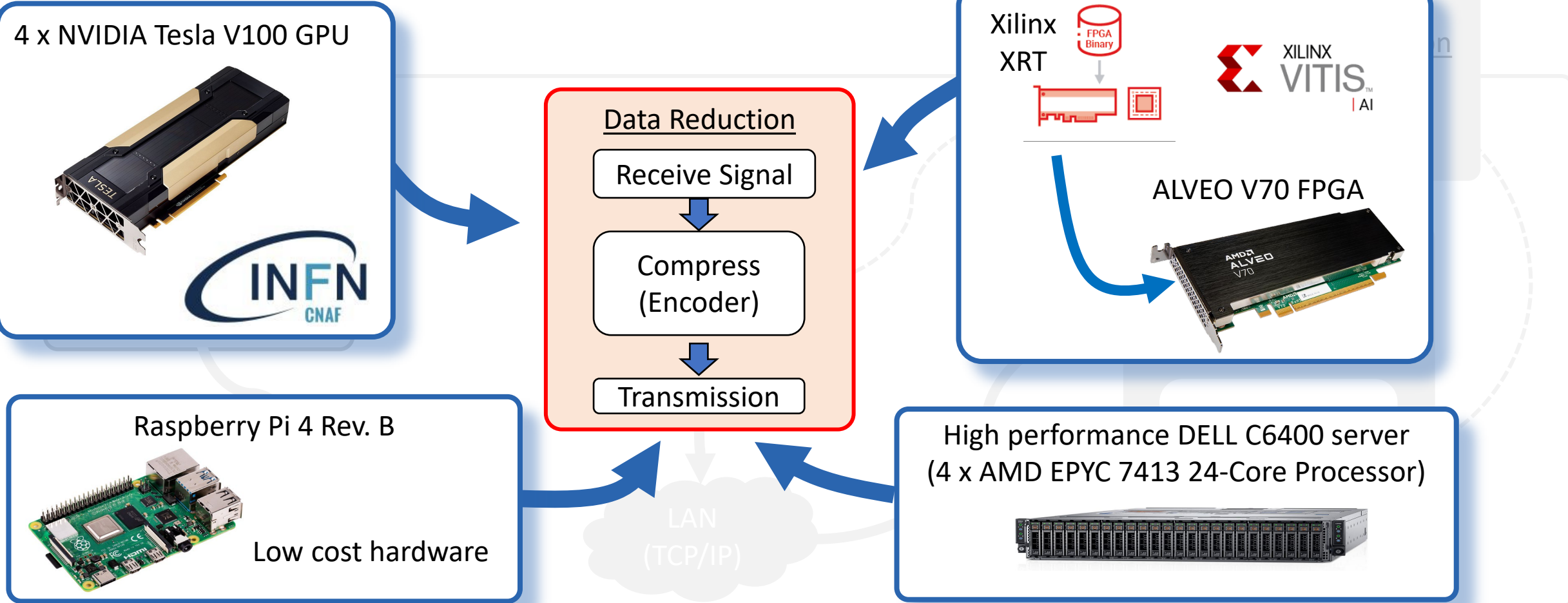


Low cost hardware

High performance DELL C6400 server
(4 x AMD EPYC 7413 24-Core Processor)



LAN (TCP/IP)





Finanziato
dall'Unione europea
NextGenerationEU



Ministero
dell'Università
e della Ricerca



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA



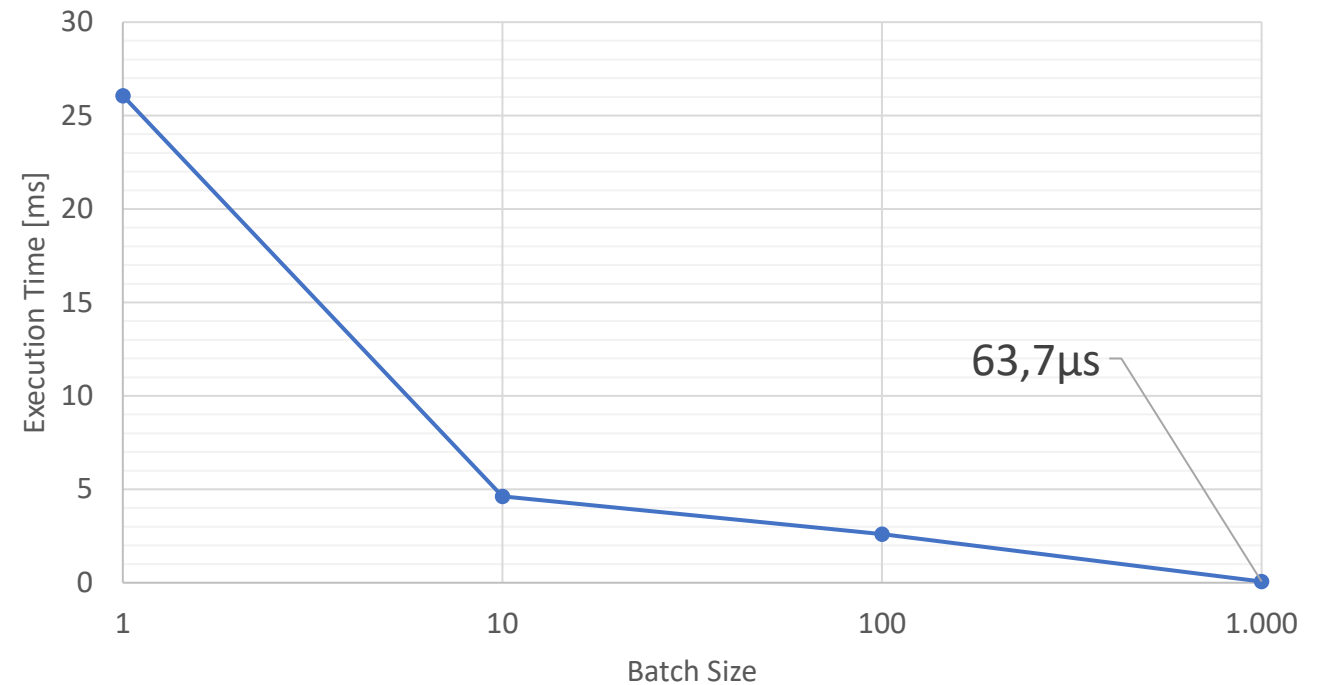
Future
Artificial
Intelligence
Research

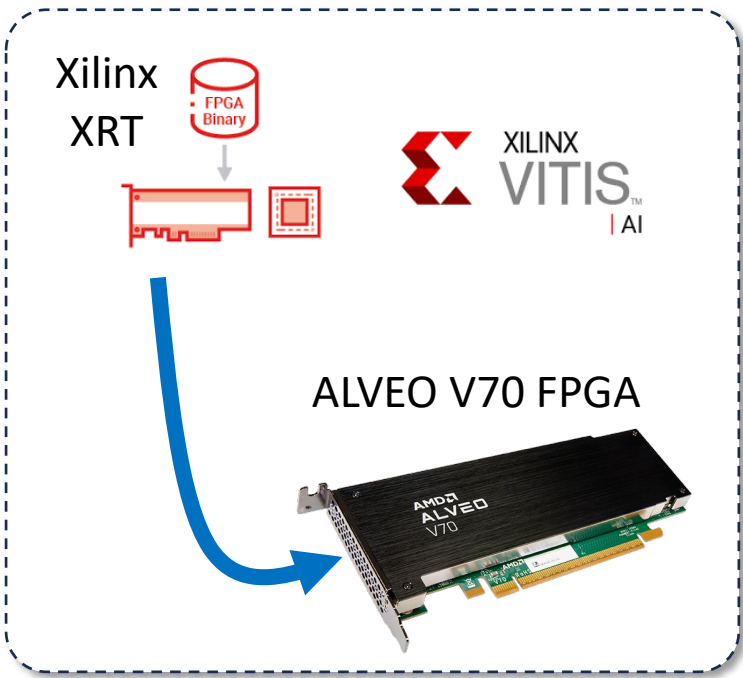
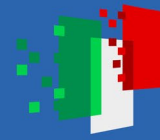
Implementation: GPU

4 x NVIDIA Tesla V100 GPU

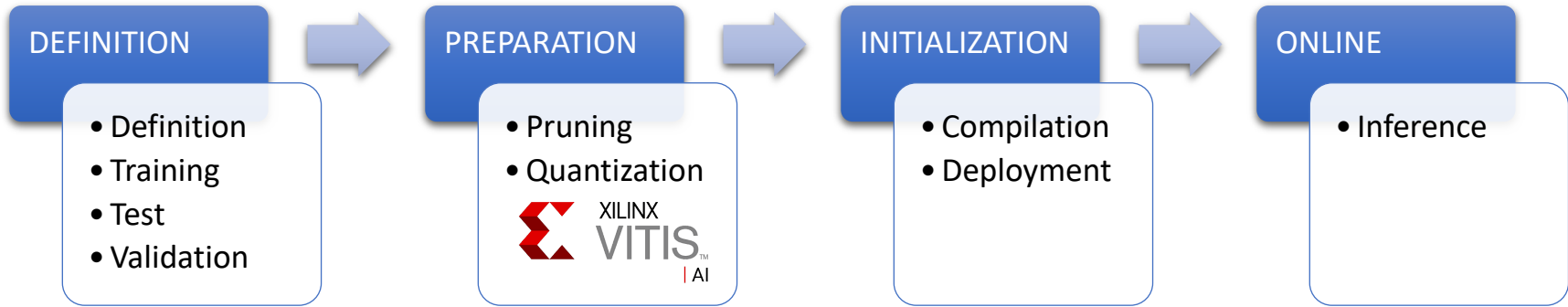


**Execution time not enough
for the application!**



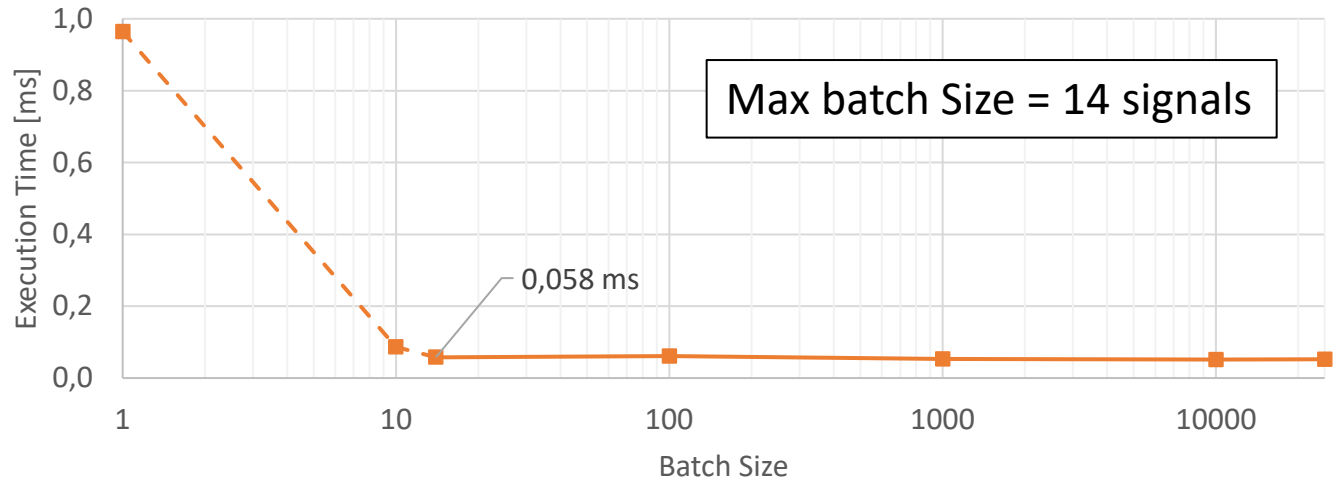


Implementation: FPGA



Execution time still not enough for the application!

Compression time of single signal





Finanziato
dall'Unione europea
NextGenerationEU



Ministero
dell'Università
e della Ricerca



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA



Future
Artificial
Intelligence
Research

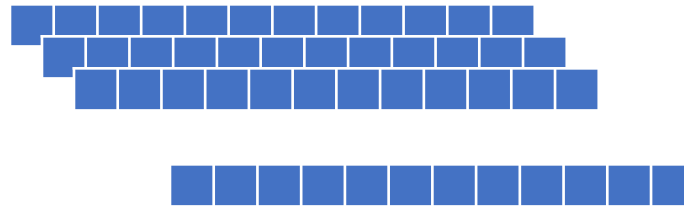
Implementation: High performance server

High performance
DELL C6400 server

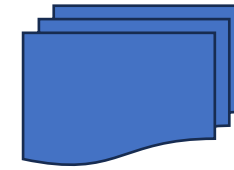


4 x AMD EPYC 7413
24 Core Processor

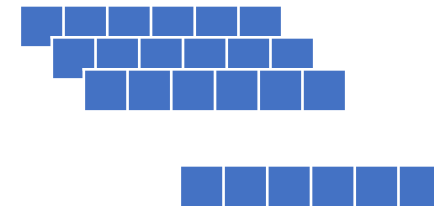
Input Batch



Parallel Execution
(openmp)



Compressed Batch



Single process

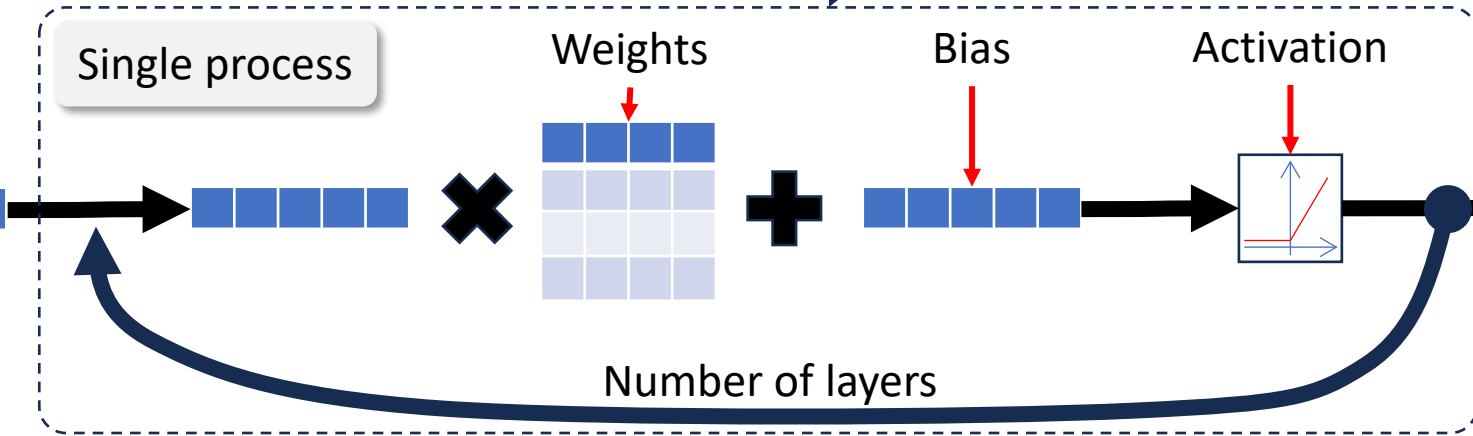
Weights

Bias

Activation



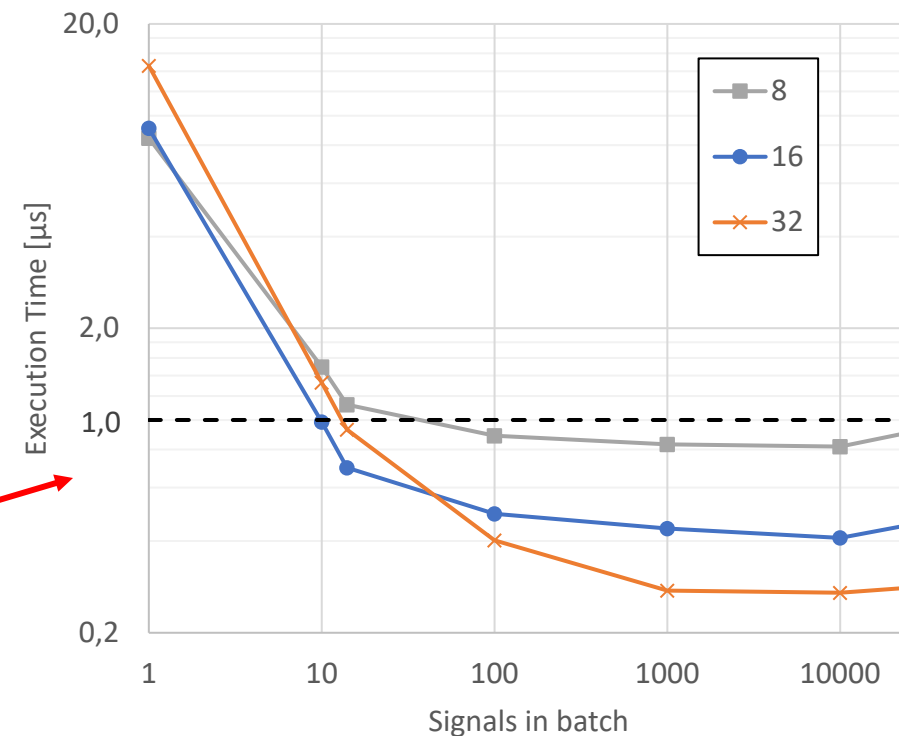
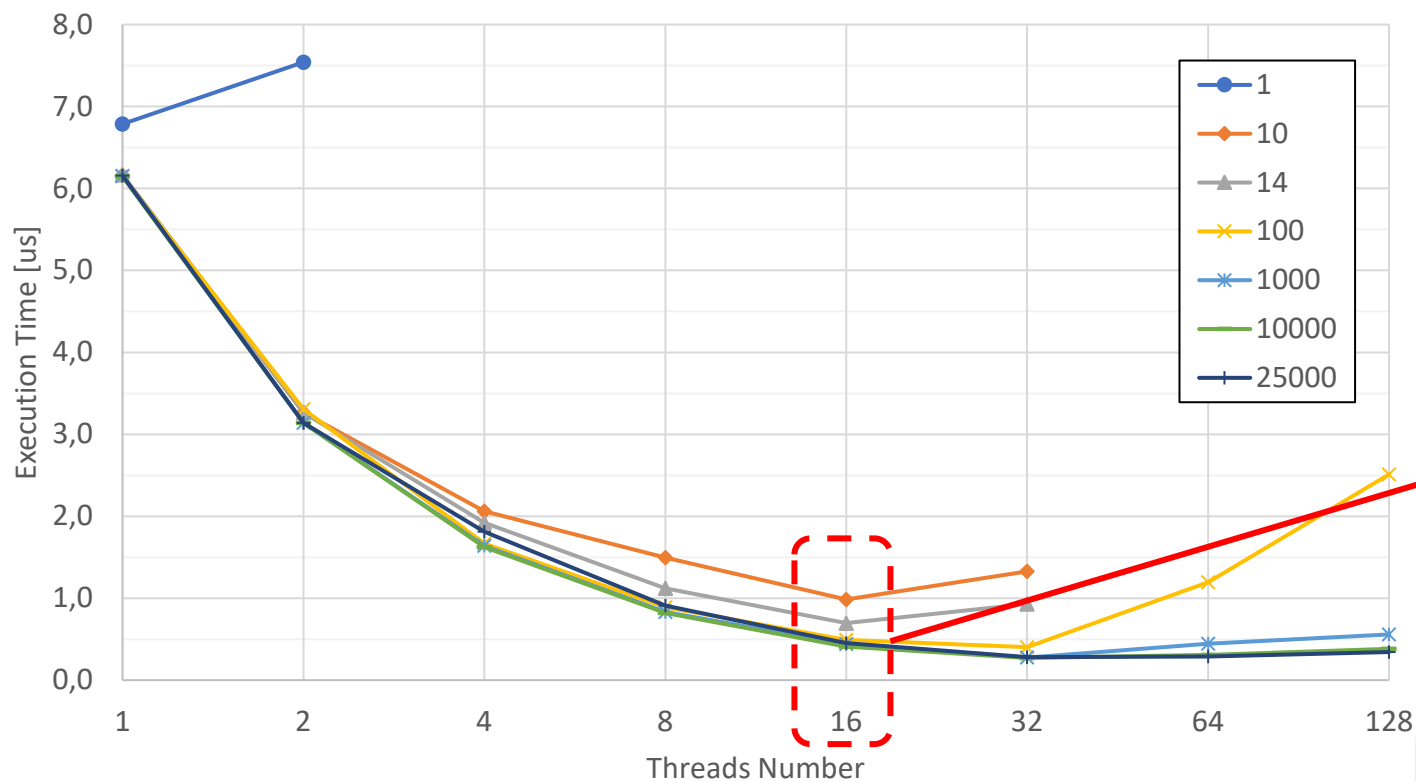
Original Signal



Compressed Signal

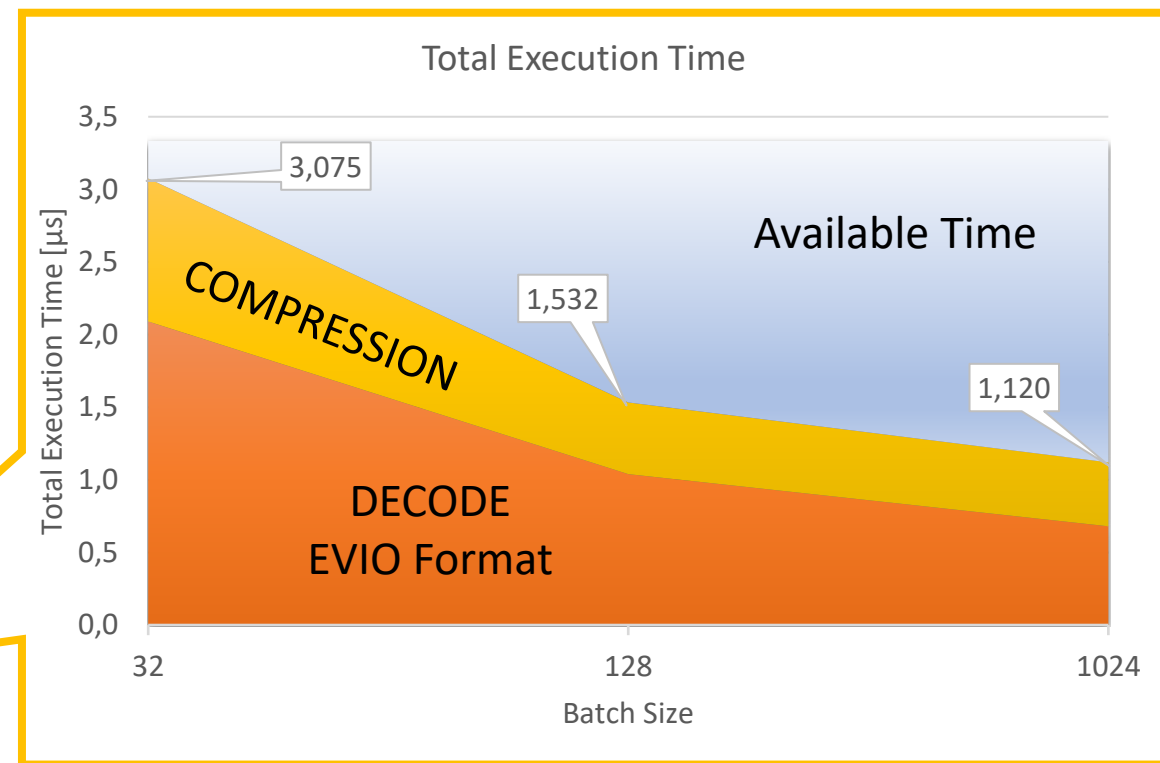
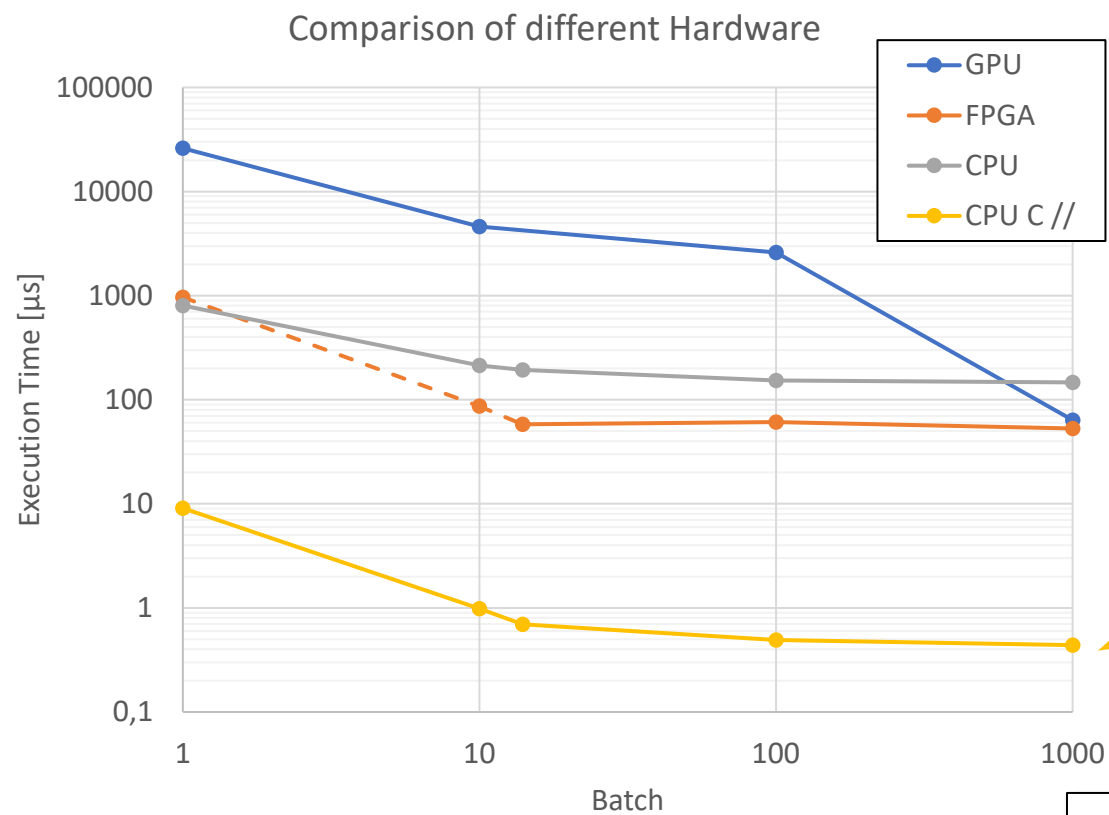
Implementation: High performance server

Execution time of different batches and threads number

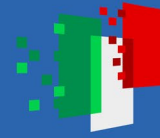


Chosen 16 Threads
Reasonable execution time for the application

Conclusion



Rate can be managed for EVIO packet with at least 32 signals

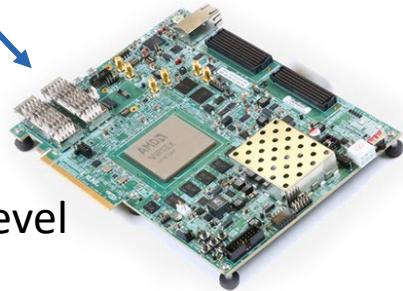
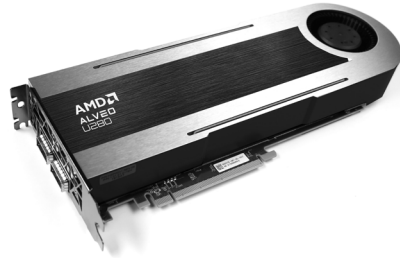


Further Studies

Low level FPGA implementation

Dedicated connectivity
(2xQSFP28 @ 100GbE)

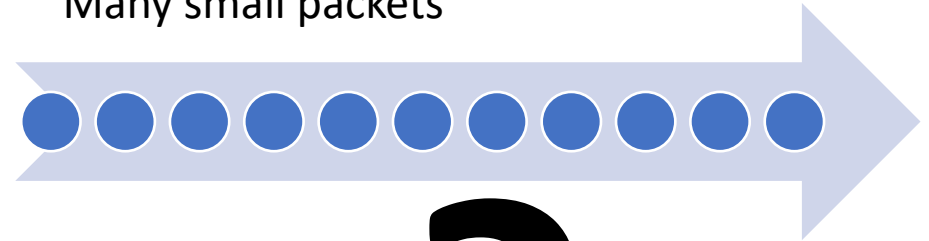
...or very Low level



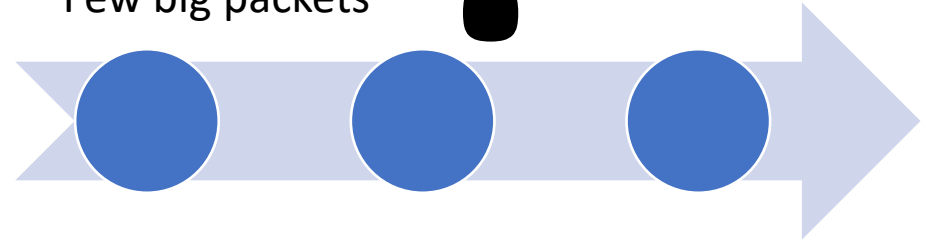
Reduce execution time and maybe save money

Statistical analysis of signals in each EVIO packet

Many small packets



Few big packets



? TBD

Estimate performance on real acquisition



Finanziato
dall'Unione europea
NextGenerationEU



Ministero
dell'Università
e della Ricerca



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA



Future
Artificial
Intelligence
Research

Thank you for your attention



FUTURE AI RESEARCH

<https://fondazione-fair.it>



<https://www.jlab.org>



<https://www.ge.infn.it>



<https://sealab.unige.it>

ACKNOWLEDGMENT

Authors have received support from: **FAIR - Future Artificial Intelligence Research, funded by the European Union Next-Generation EU (Italy)Research) – spoke 6.**