Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Streaming Readout Workshop

SRO-XII

December 2-4, 2024
University of Tokyo, Japan

Real-Time data reduction with Artificial Intelligence for SRO.

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

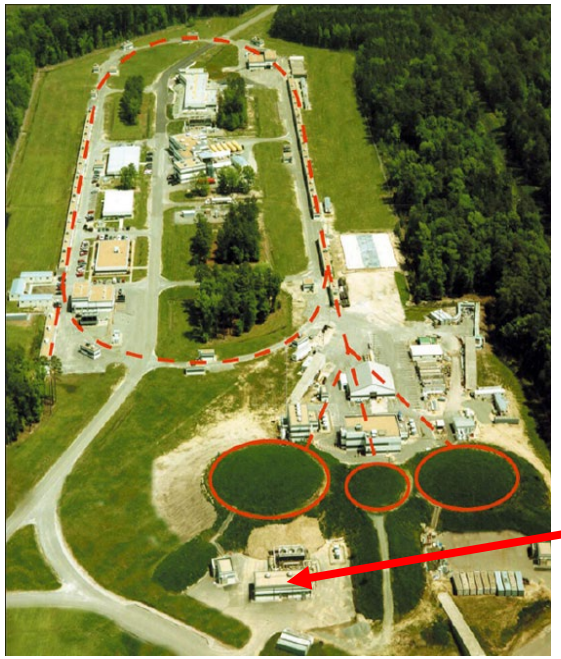
Edoardo Ragusa, **Paolo Gastaldo**
SEALab Università di Genova (DITEN)
Genova (Italy)

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
FAIR Future Artificial Intelligence Research

High Energy Physics Experiment: Beam Dump eXperiment (BDX)

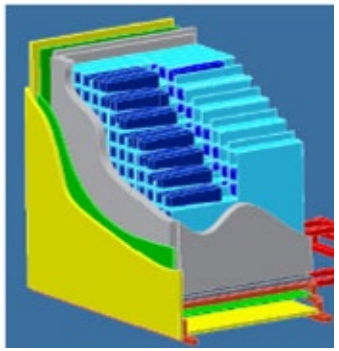
Jefferson Lab

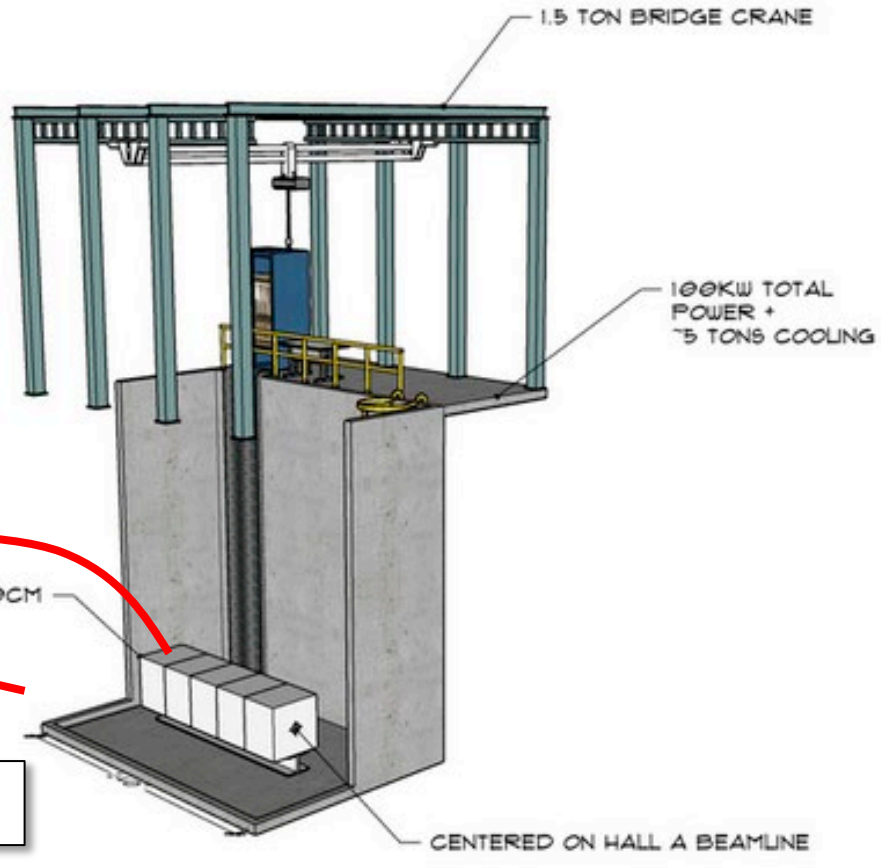


≈1000
Calorimeter channels
(30MB/s)

≈ 300
Veto channels
(500MB/s)

Single module



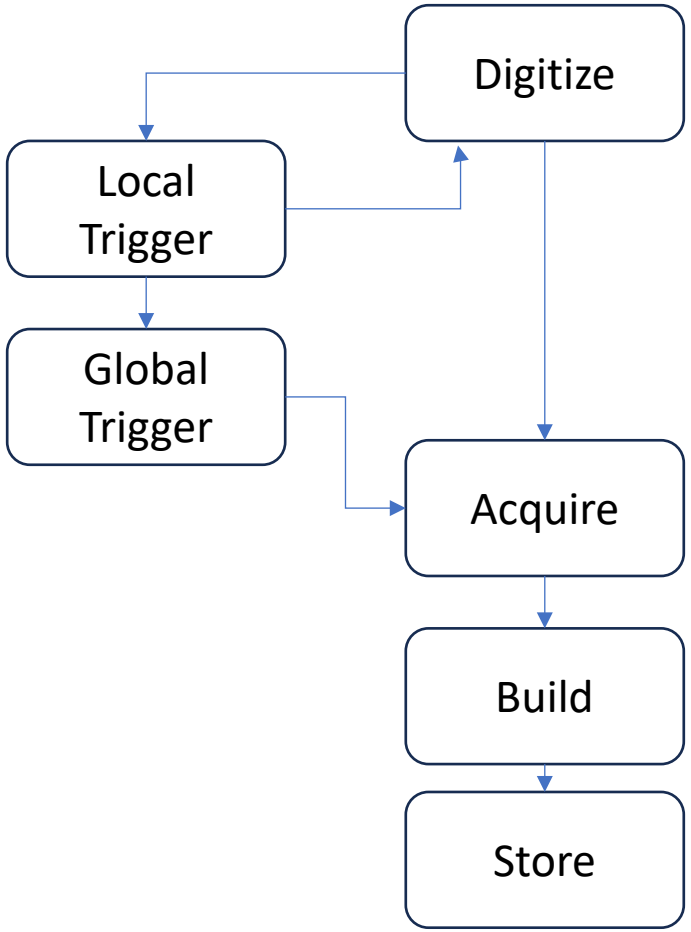


Very rare occurrence of Dark Matter events

Retrieved from: Battaglieri, M., et al. "Dark matter search in a Beam-Dump eXperiment (BDX) at Jefferson Lab." arXiv preprint arXiv:1607.01390 (2016).

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Traditional triggered DAQ VS Streaming Readout



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

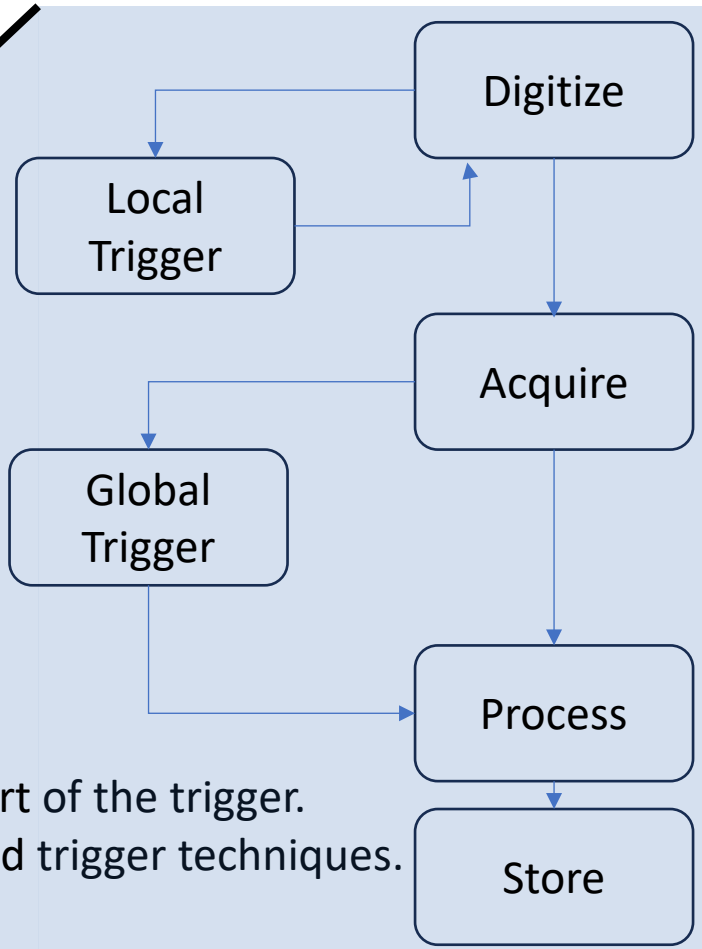
Pros:
It works reliably.

Triggered

Streaming

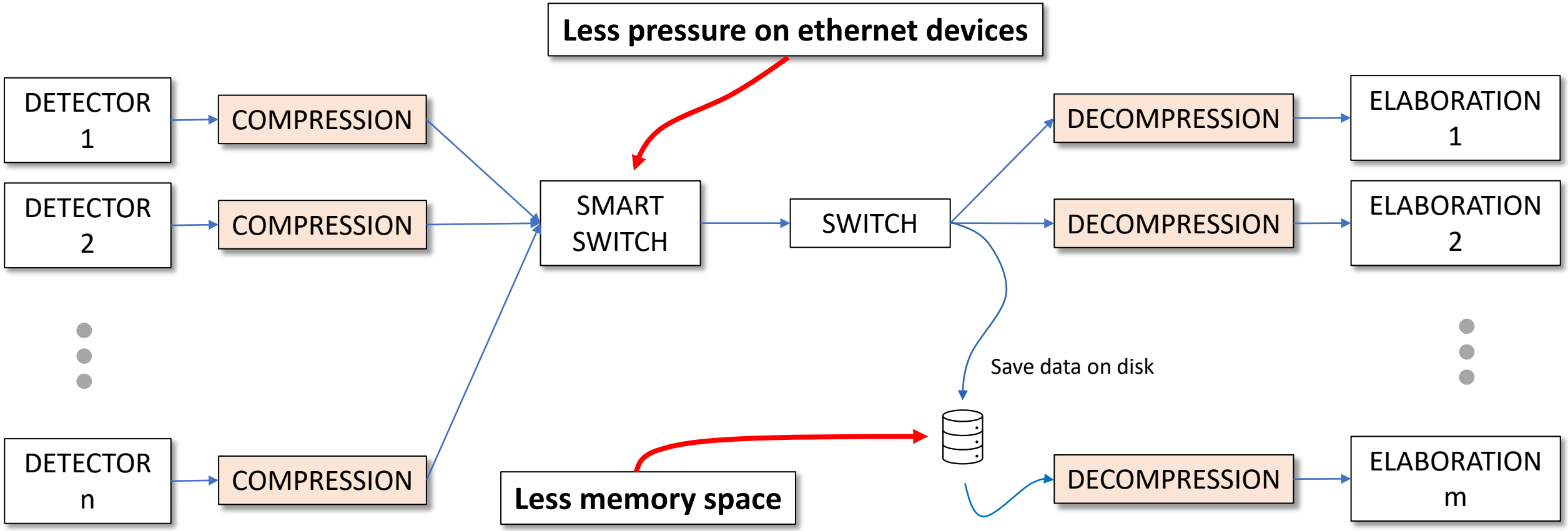
Cons:
<u>High data rate.</u>
New design.

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



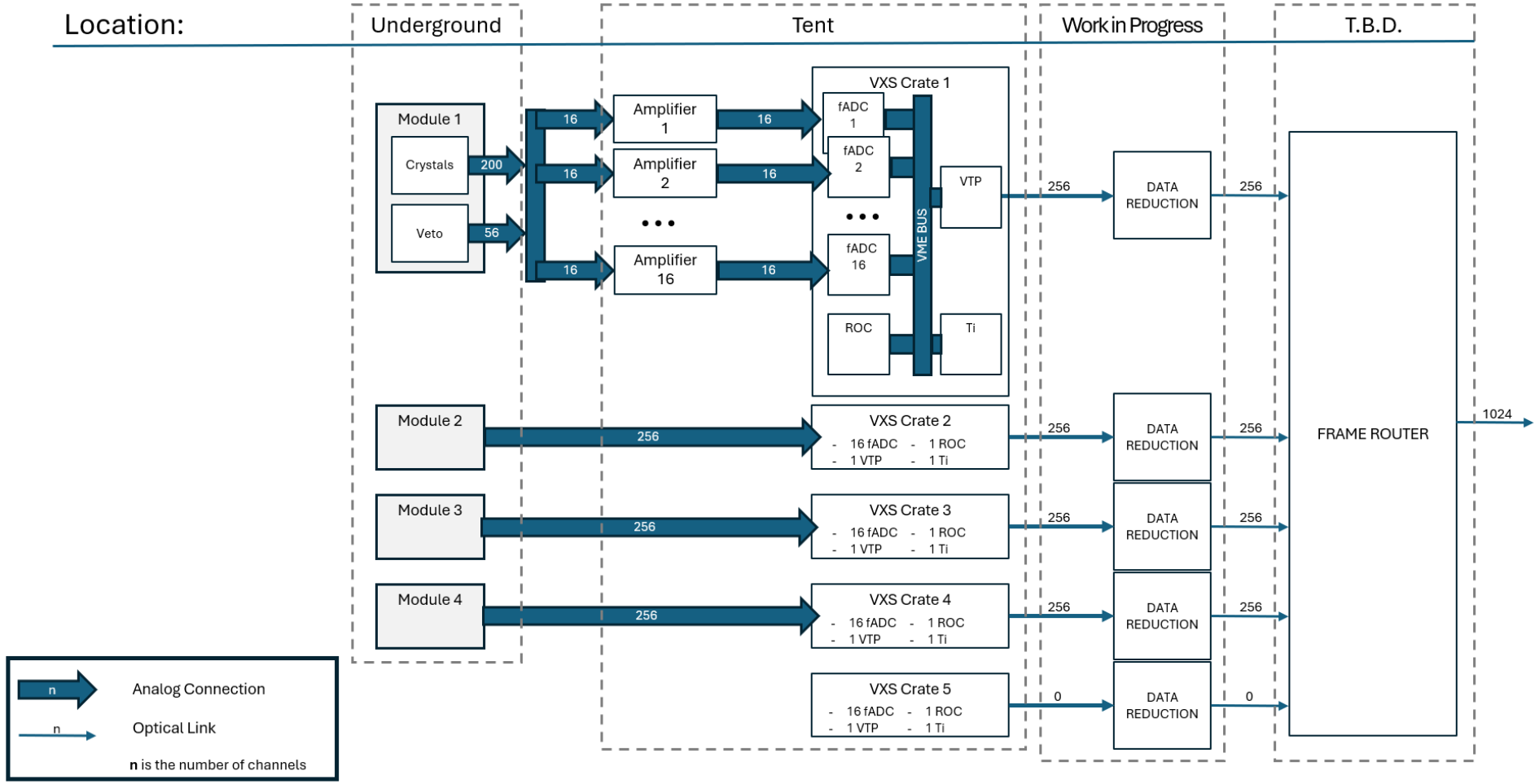
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Block scheme of data flow



Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

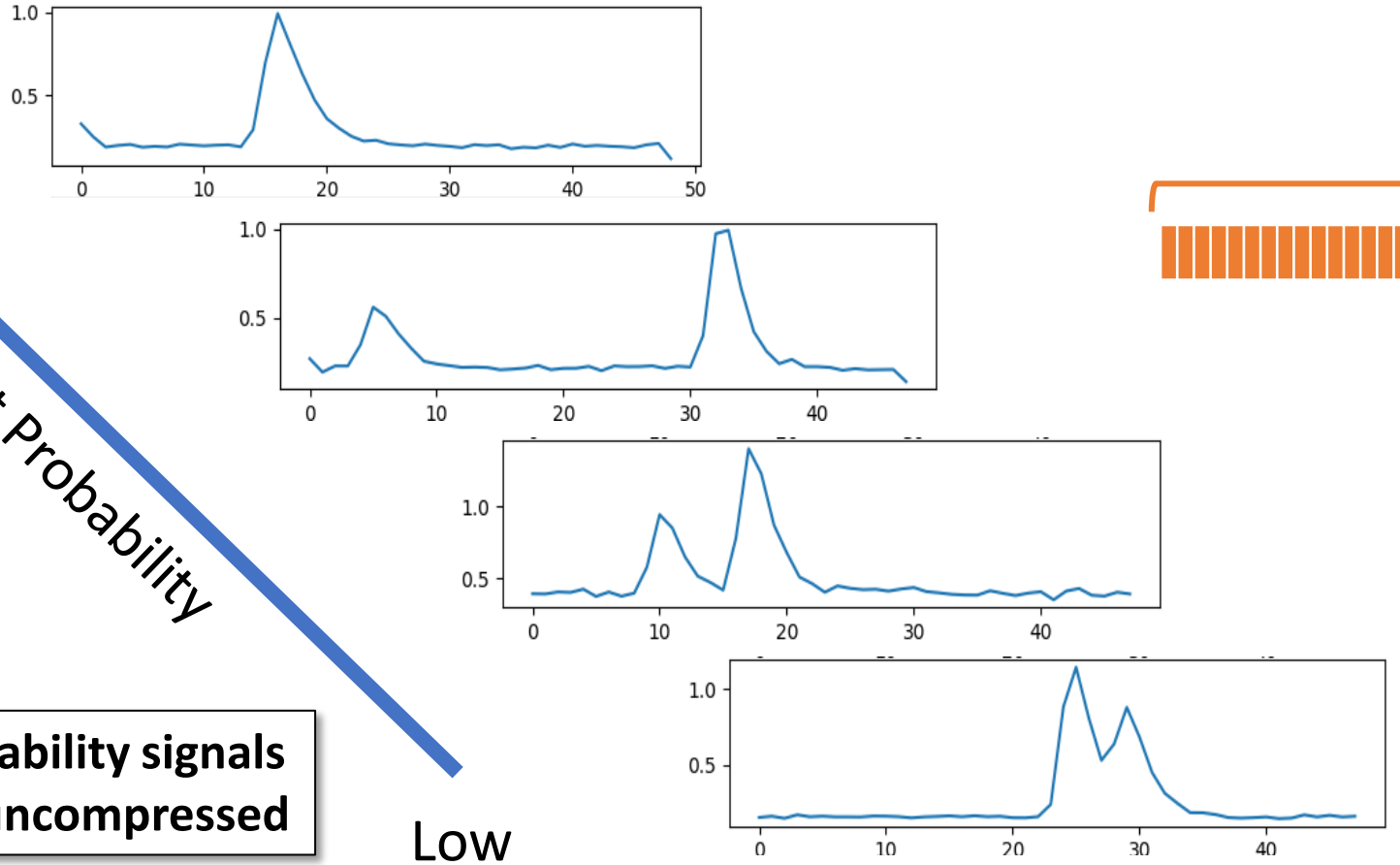
Detailed BDX data flow scheme



Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Data from physical Experiment

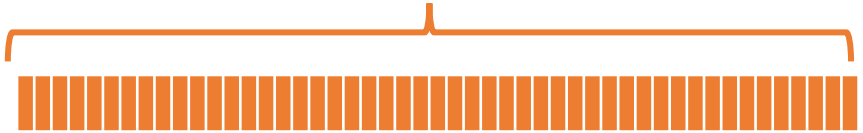
High



Event Probability

Low

48 samples



Very-Low probability signals could be sent uncompressed

Data reduction algorithm: Autoencoder

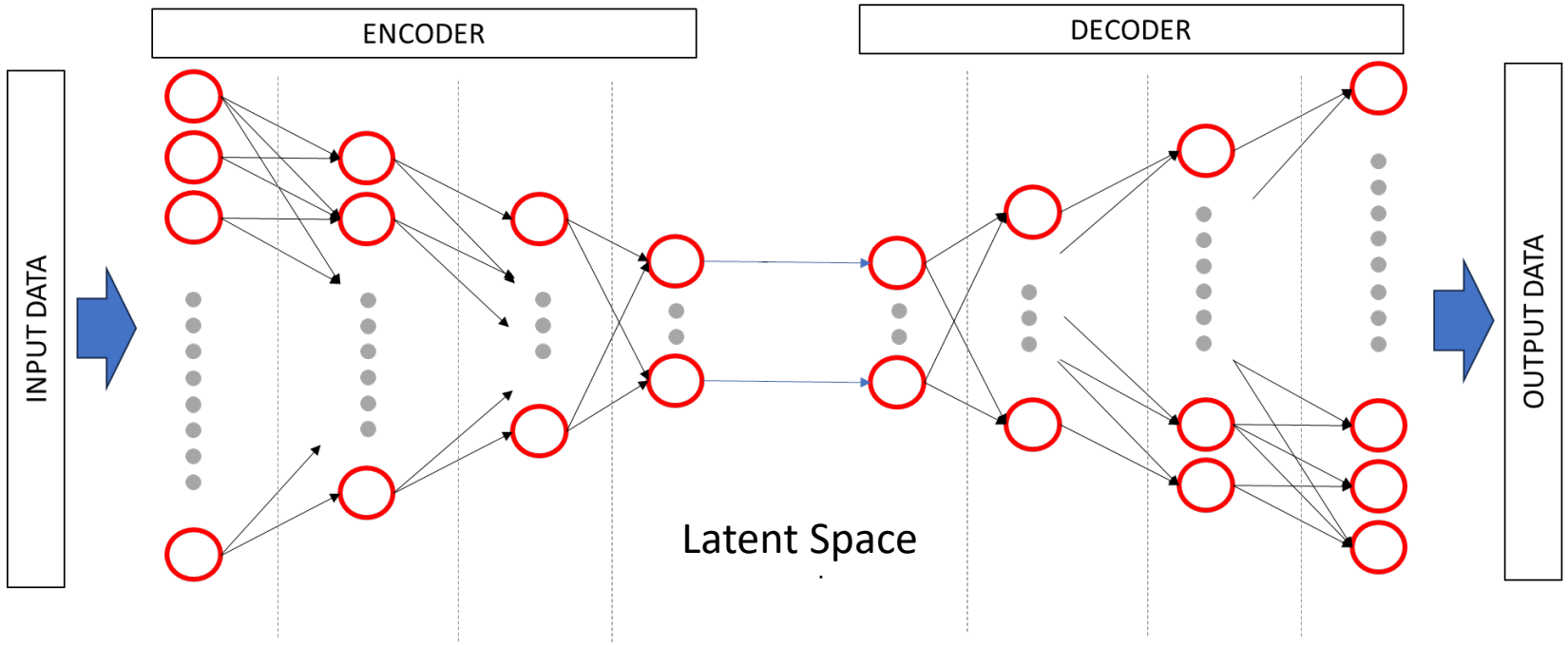
Machine Learning Algorithm

Dimensionality reduction

Unsupervised learning

Artificial Neural Network

Composed of two function:
- encoding
- decoding



FULLY CONNECTED AUTOENCODER WITH DENSE LAYER

Lossy compression algorithm

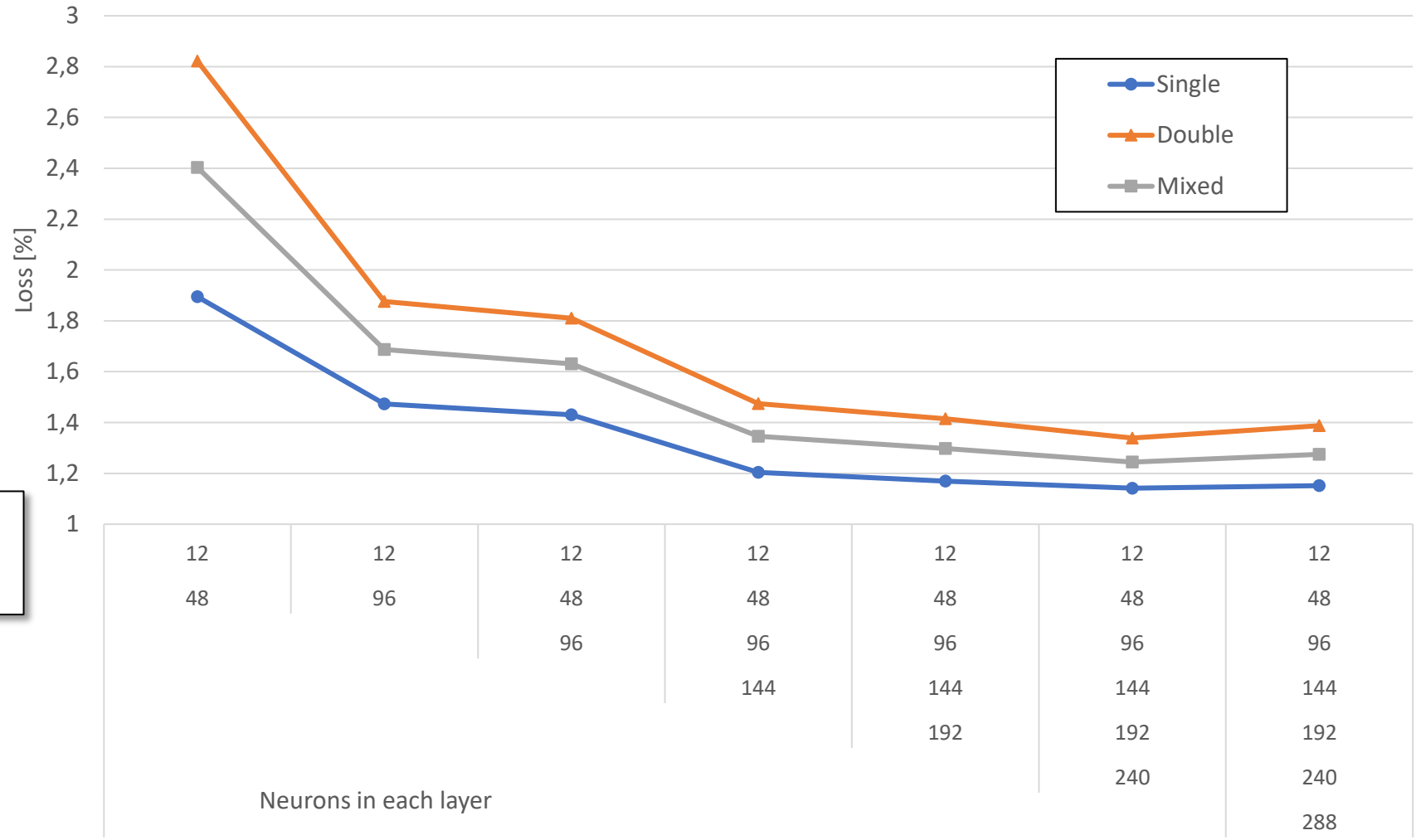
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Autoencoder Training: Different configuration

Test increasing:

- Layer number
- Neurons in each layer

Few improvement adding more and more parameters



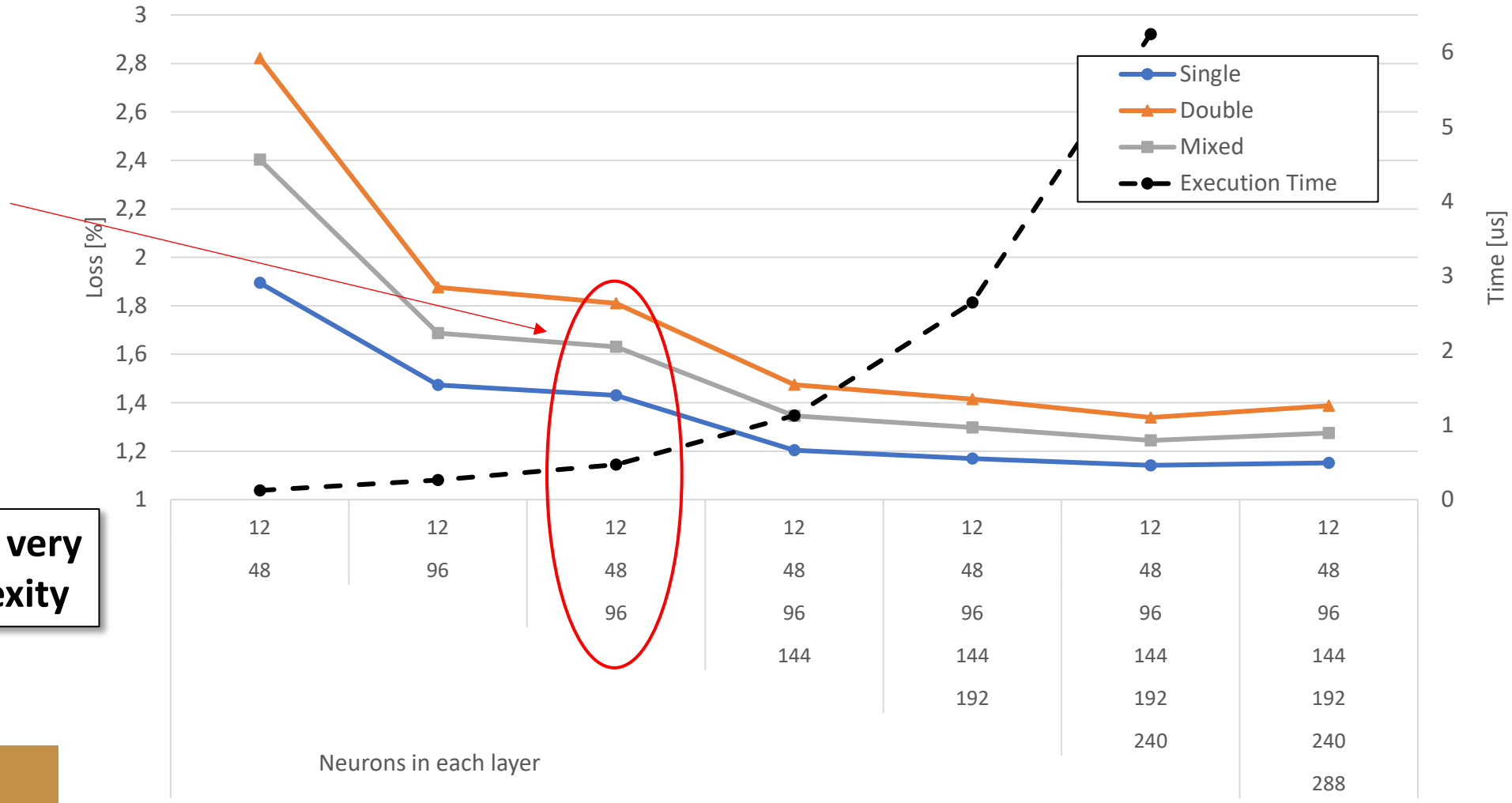
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Autoencoder Training: Different configuration

Chosen configuration

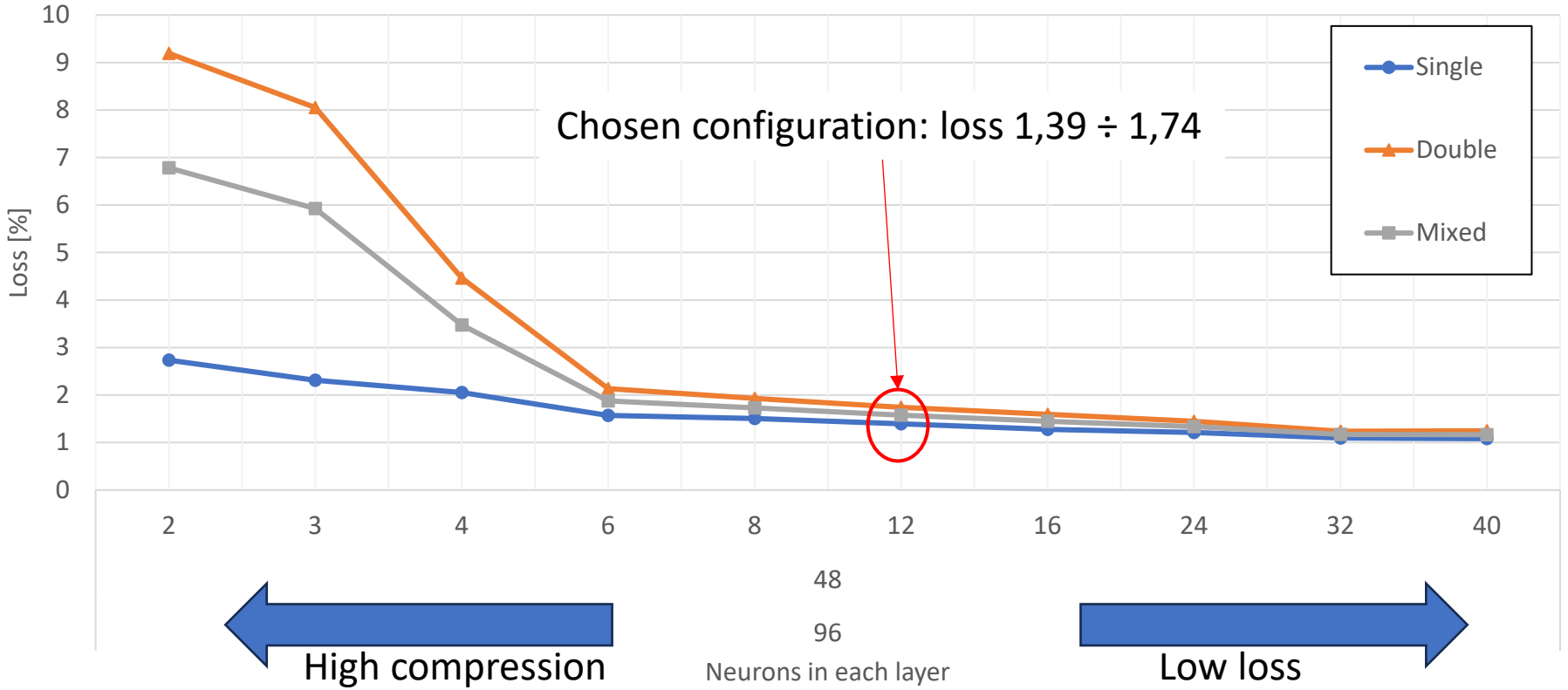
Layers: 3

Neurons:
96
48
12



Execution time increase very fast with model complexity

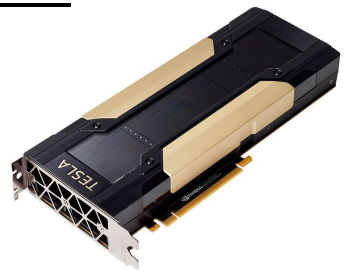
Autoencoder Training: Different latent space



Compression ratio is a parameter and could be chosed as loss tradeoff

Autoencoder: Training time

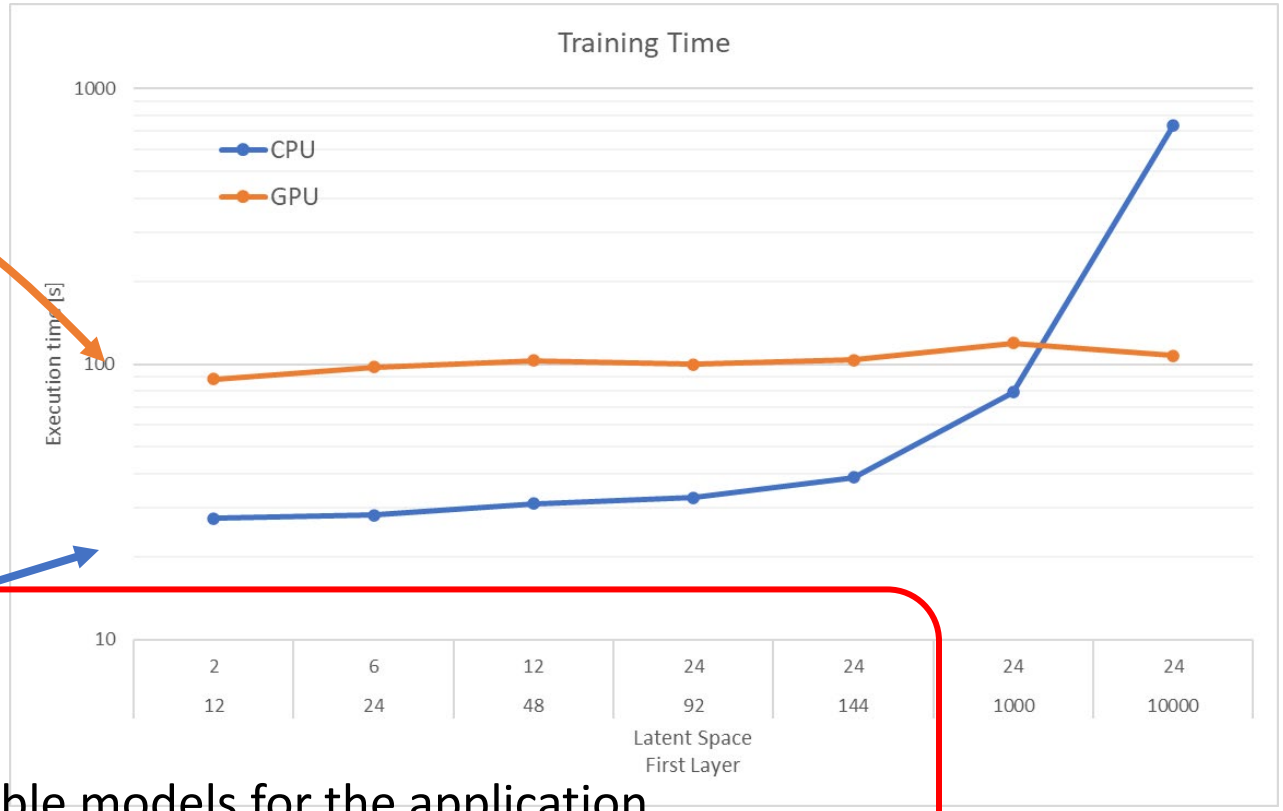
GPU



CPU

AMD Ryzen 5 5600U

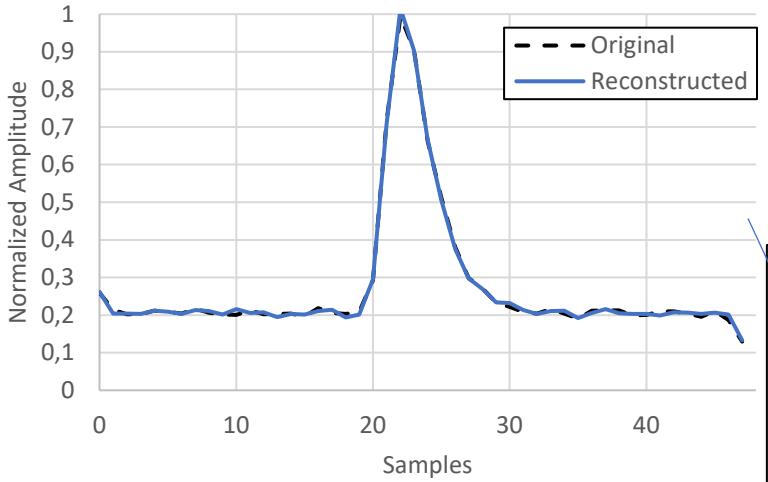
6 core, 12 Logic processor

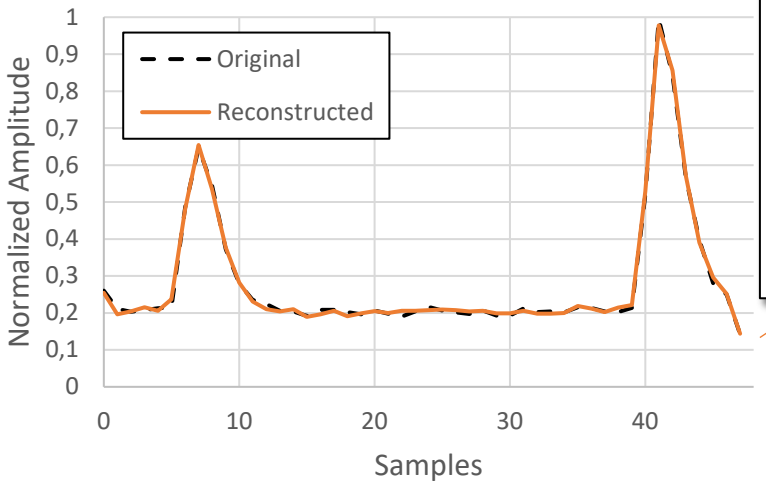


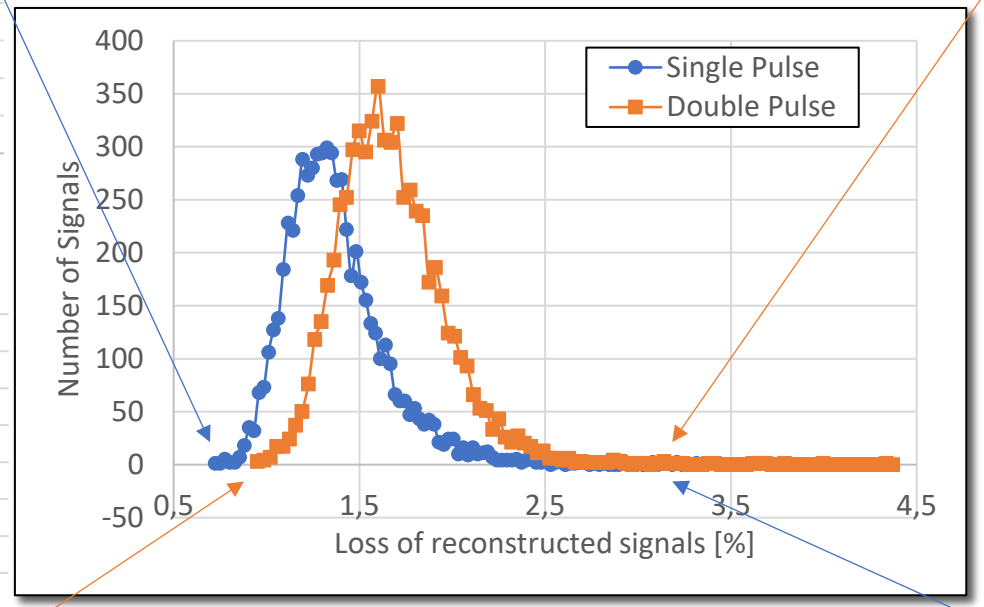
Suitable models for the application

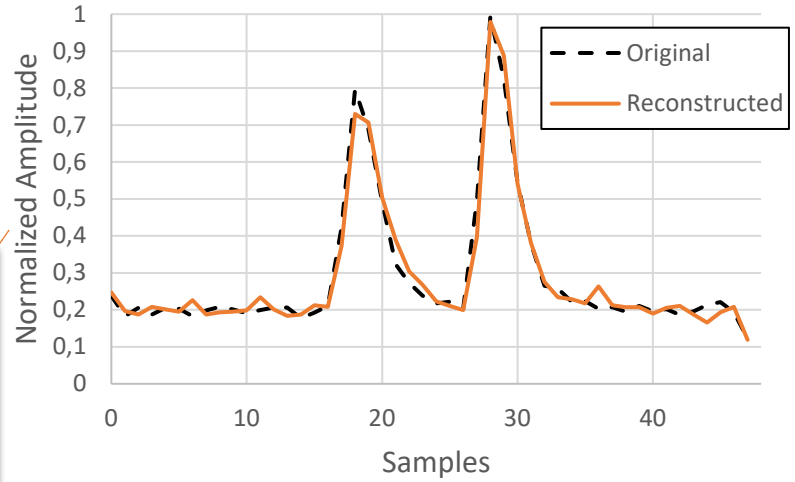
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

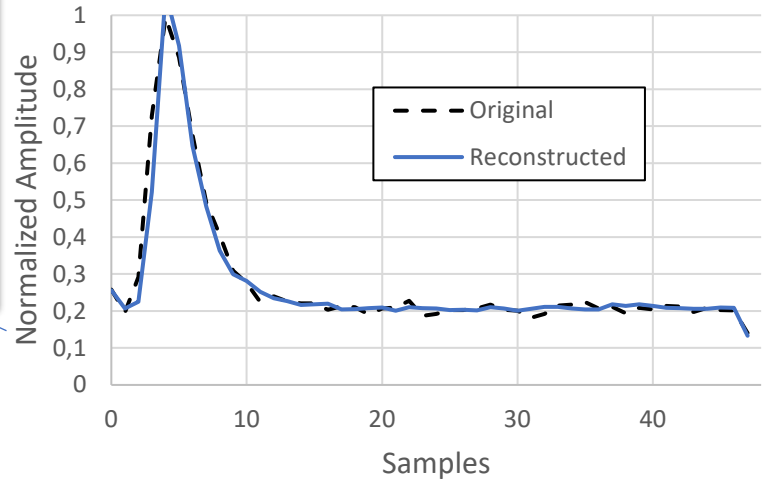
Signals Compression





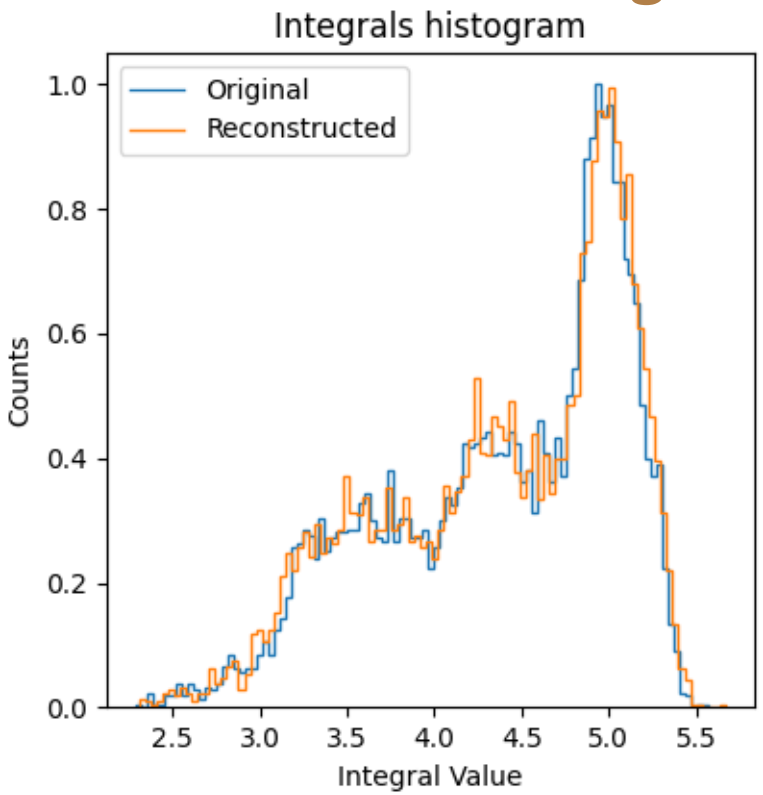


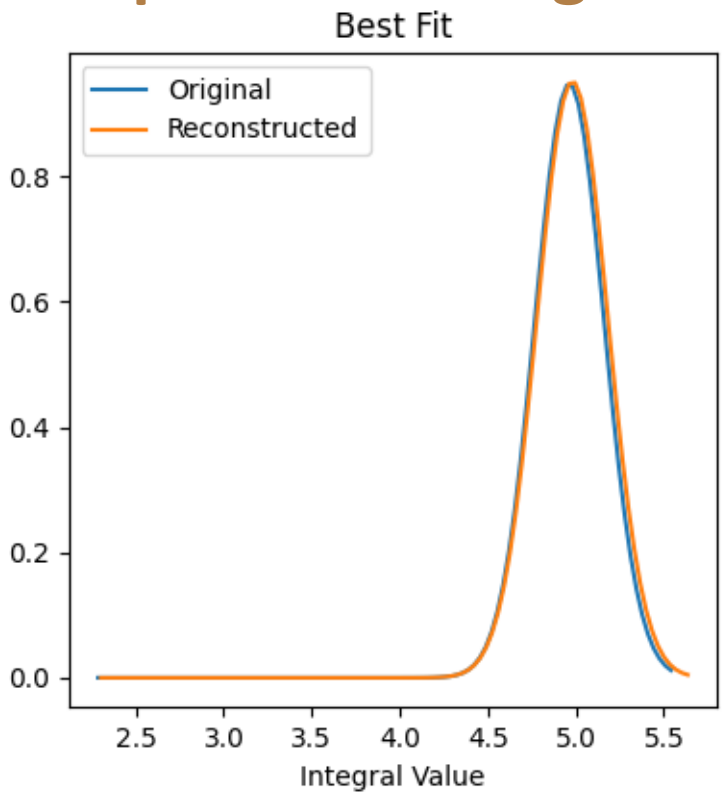




Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Signals Compression: Integral and spectrum



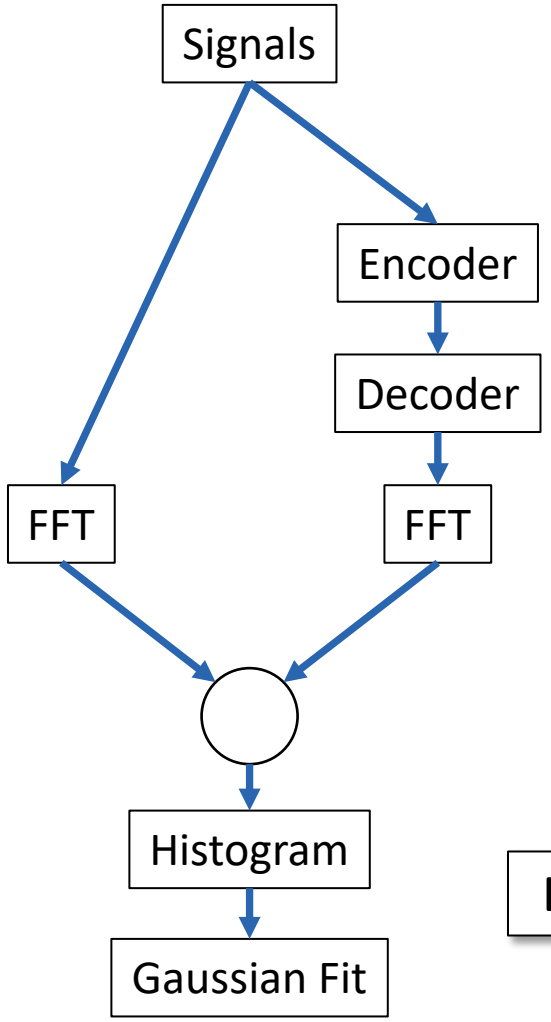


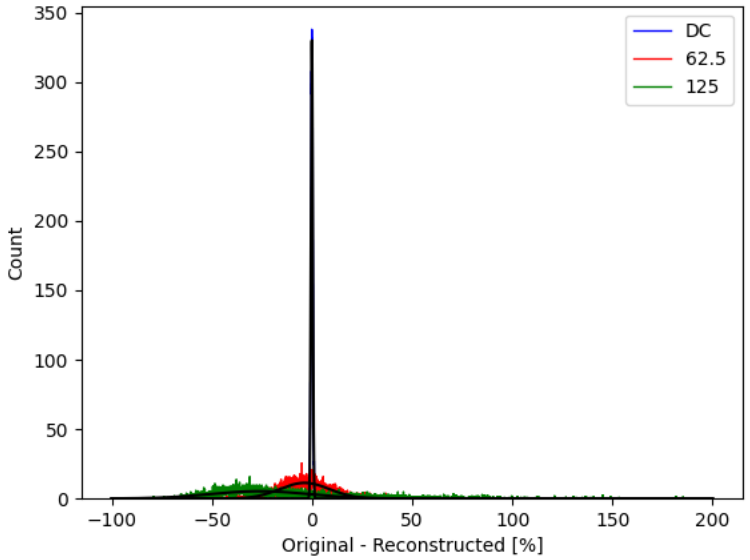
	A	μ	σ
Original	199,124	4,963	0,1952
Recon.	199,992	4,981	0,2006
Diff. [%]	0,44	0,35	2,77

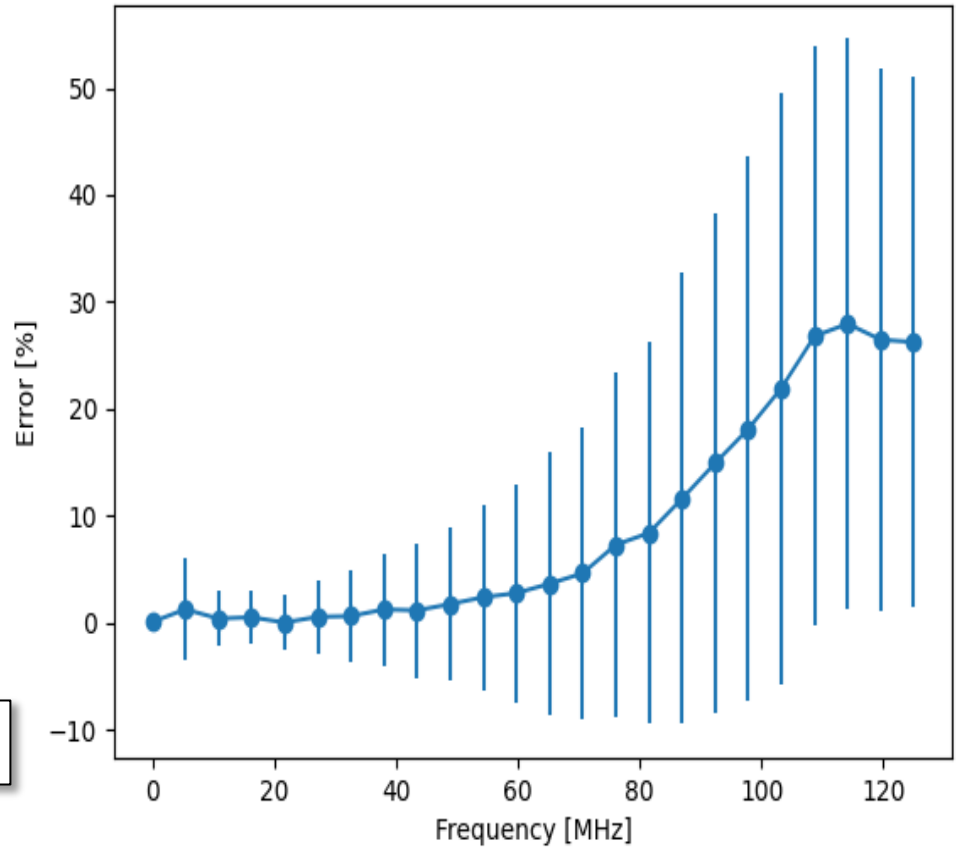
Good performance also on the derived quantities for physical analysis

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Signals Compression: FFT analysis



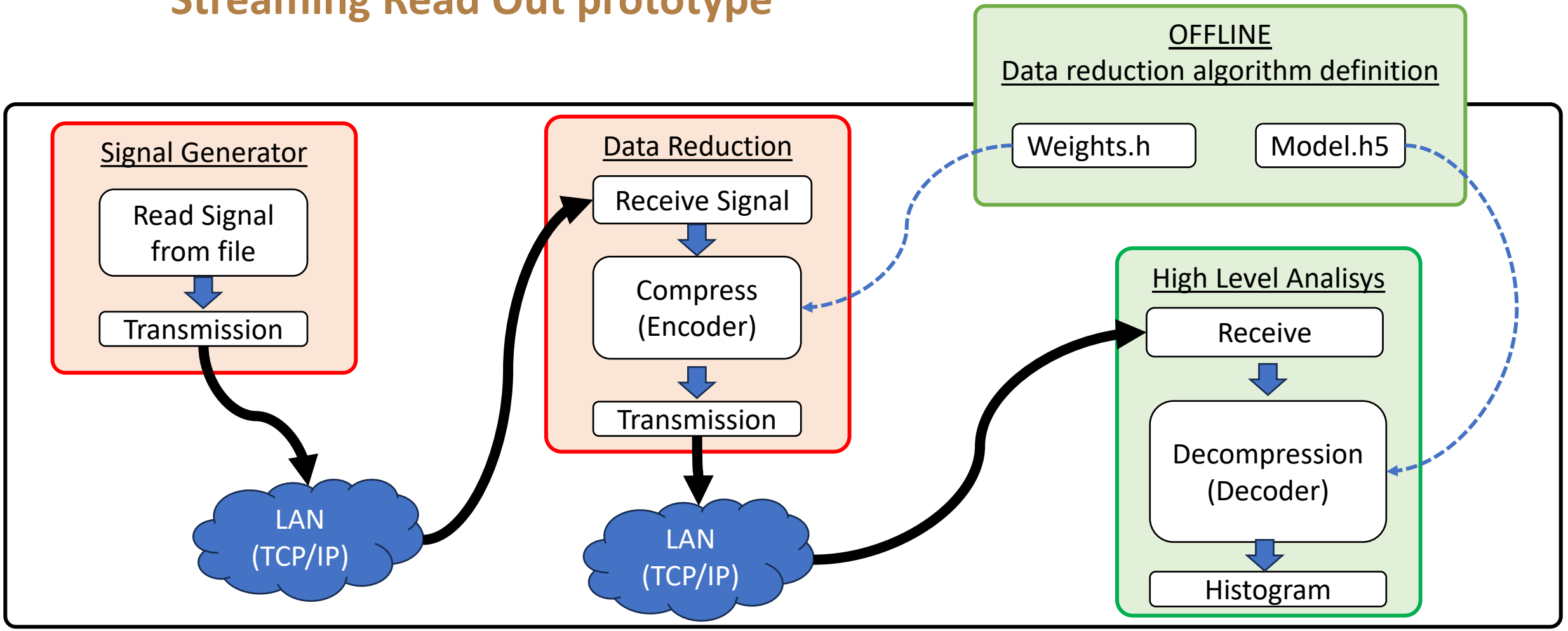




Better reconstruction of low frequency

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Streaming Read Out prototype



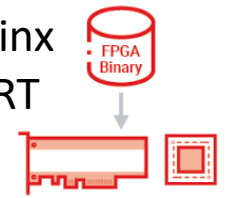
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Implementation of Data Reduction Node

4 x NVIDIA Tesla V100 GPU

INFN CNAF

Xilinx XRT



XILINX VITIS AI

ALVEO V70 FPGA

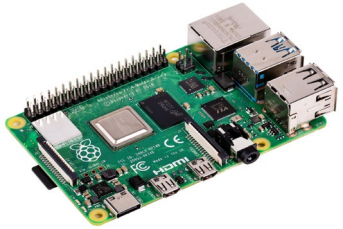
Data Reduction

Receive Signal

Compress (Encoder)

Transmission

Raspberry Pi 4 Rev. B



Low cost hardware

LAN (TCP/IP)

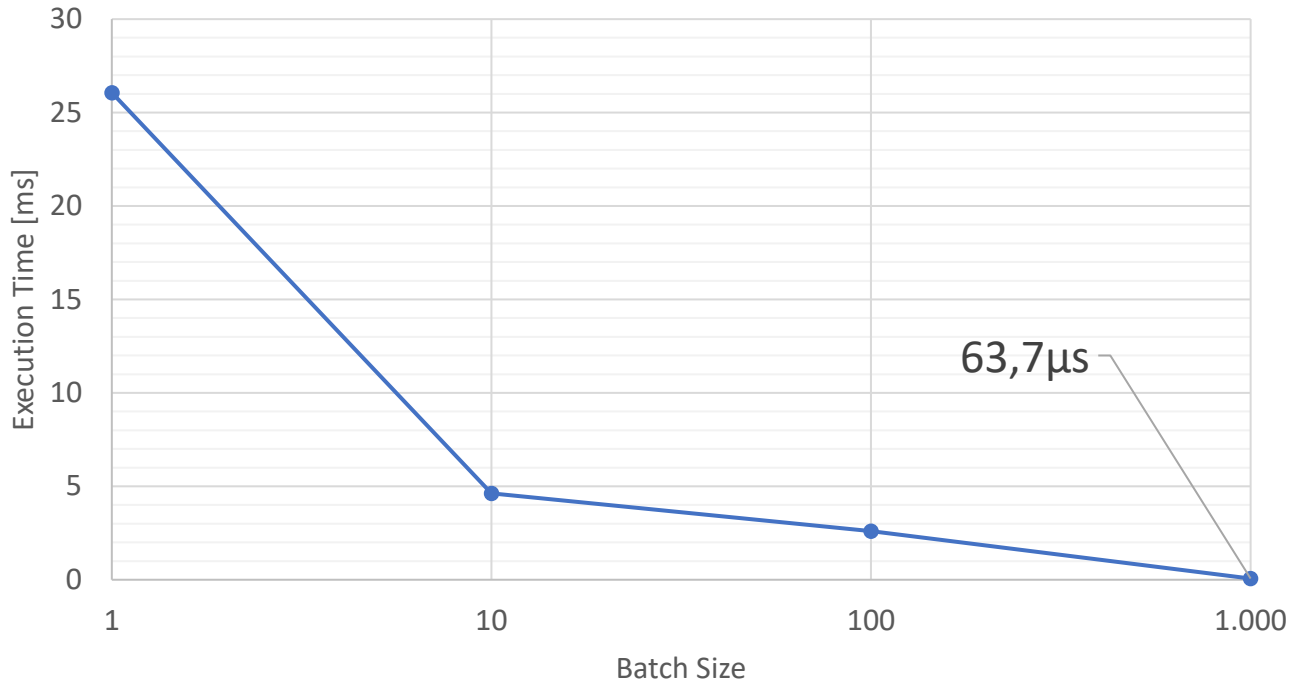
High performance DELL C6400 server (4 x AMD EPYC 7413 24-Core Processor)

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Implementation: GPU

4 x NVIDIA Tesla V100 GPU

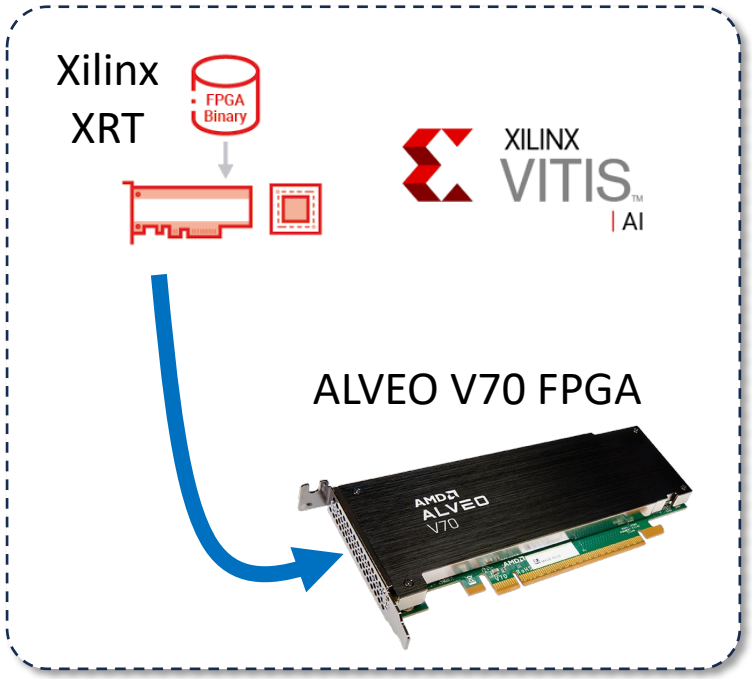
INFN CNAF



Execution time not enough for the application!

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Implementation: FPGA



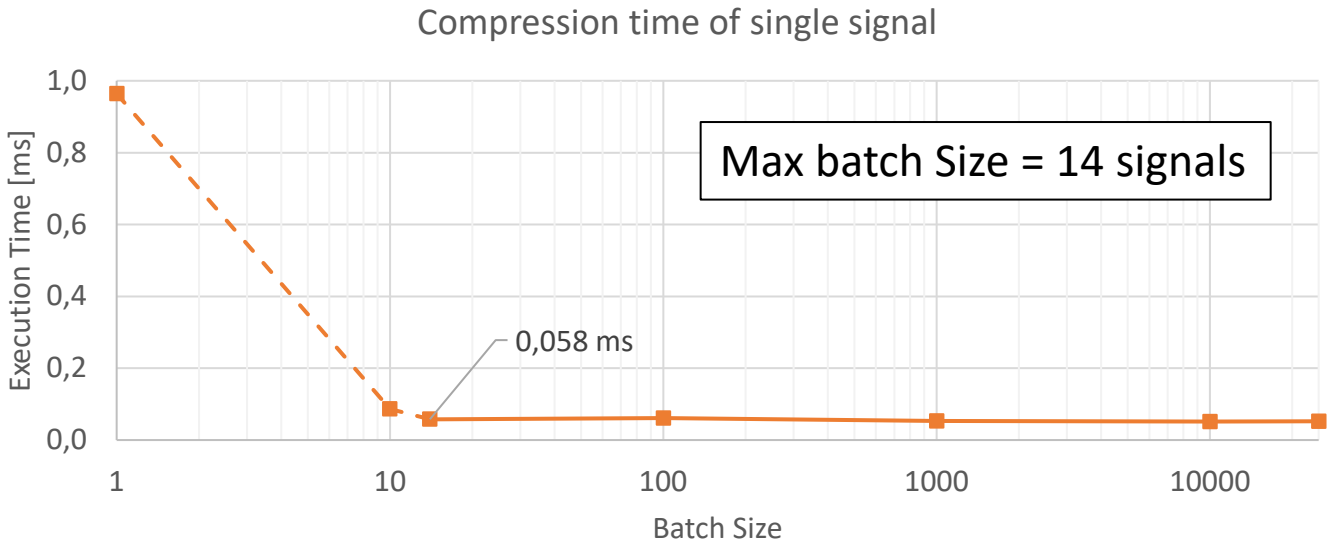
DEFINITION
- Definition
- Training
- Test
- Validation

PREPARATION
- Pruning
- Quantization

XILINX VITIS AI

INITIALIZATION
- Compilation
- Deployment

ONLINE
- Inference



Execution time still not enough for the application!

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR 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

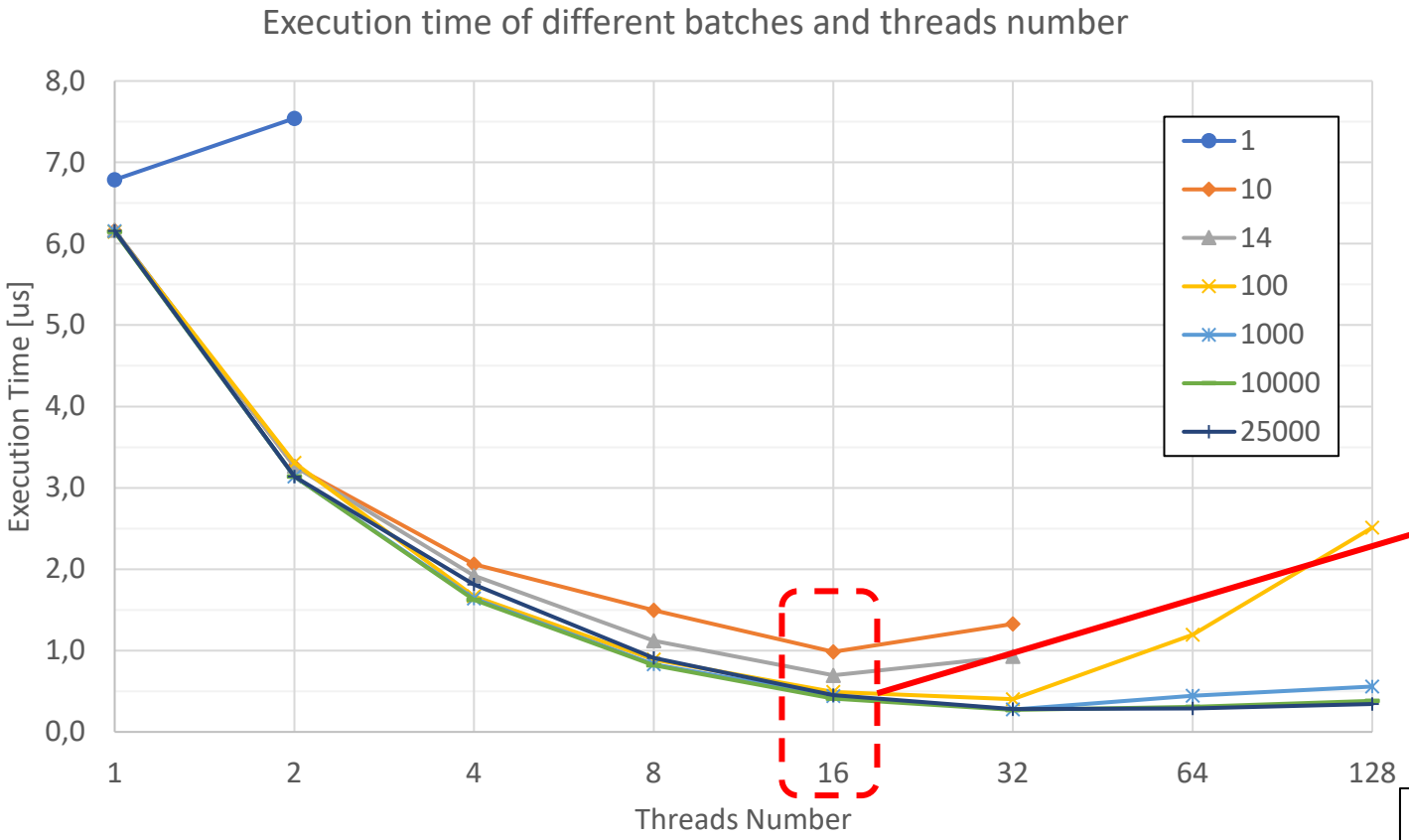
Original Signal

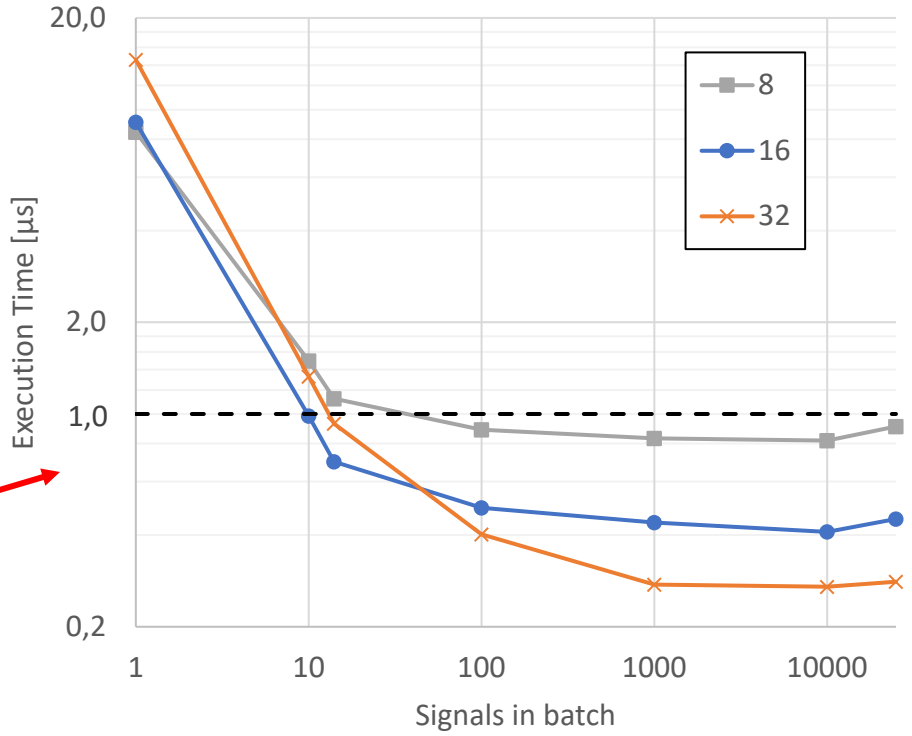
Compressed Signal

Number of layers

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Implementation: High performance server

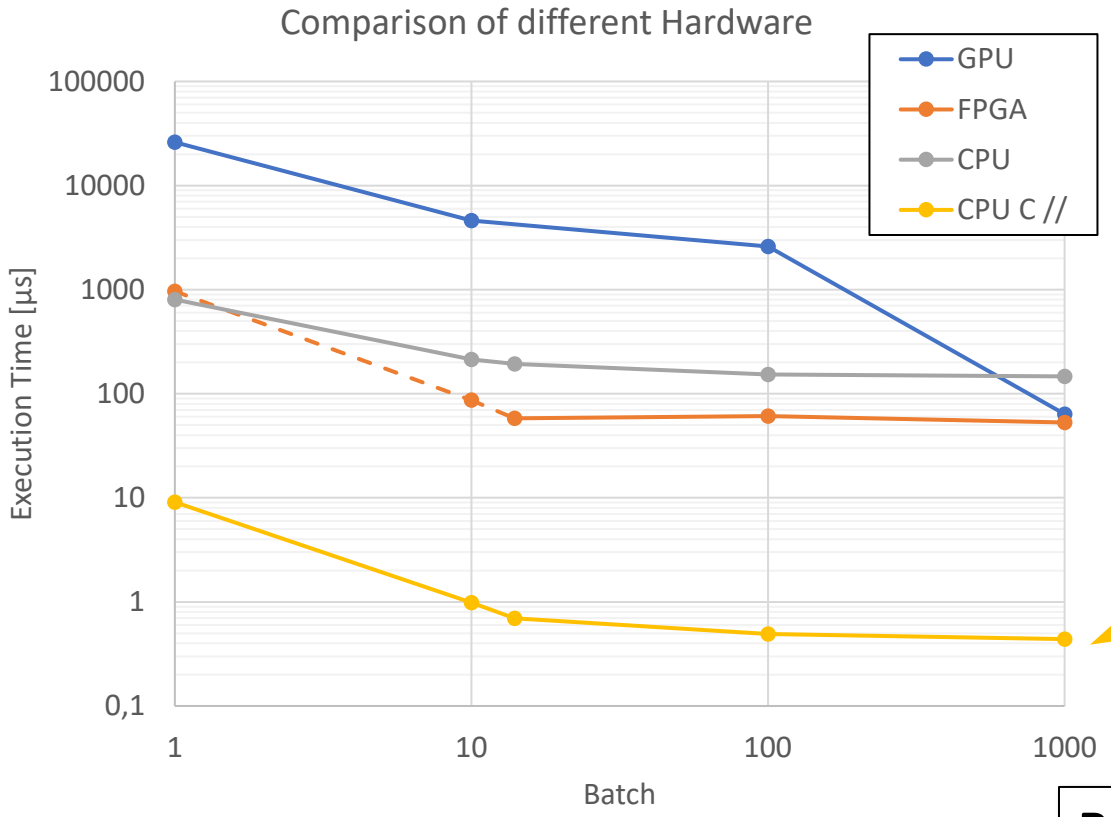


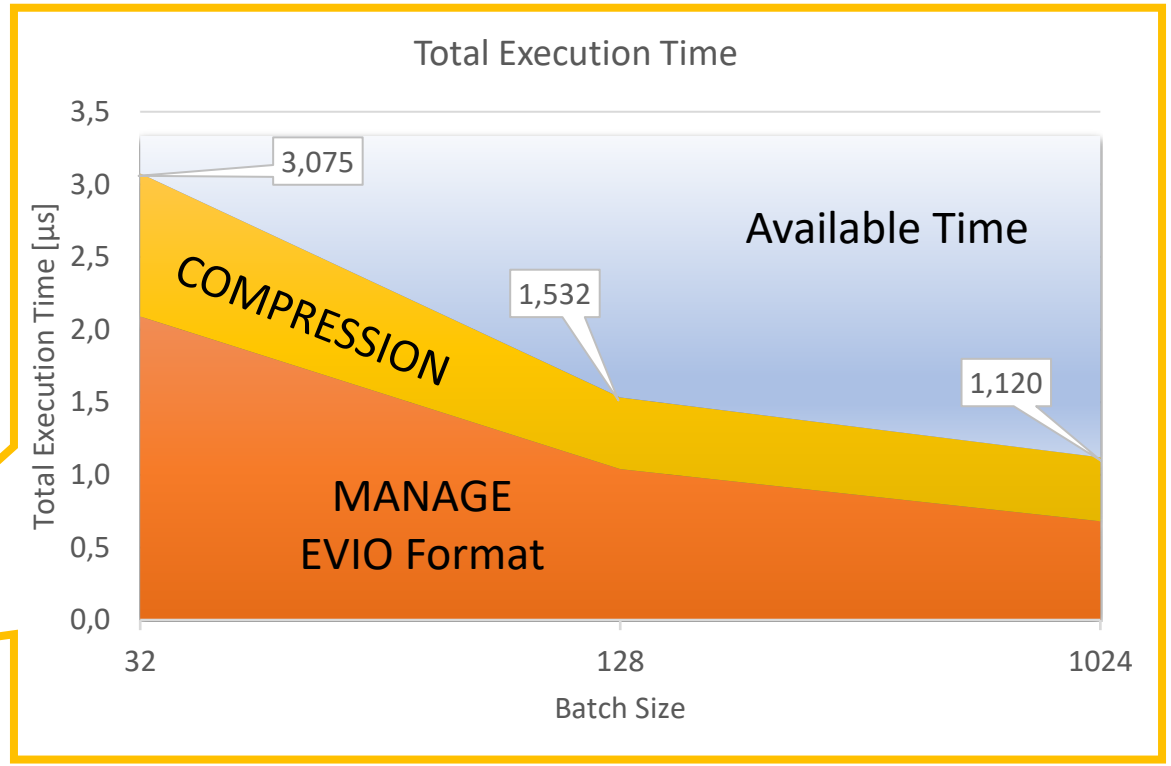


Chosen 16 Threads
Reasonable execution time for the application

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Conclusion





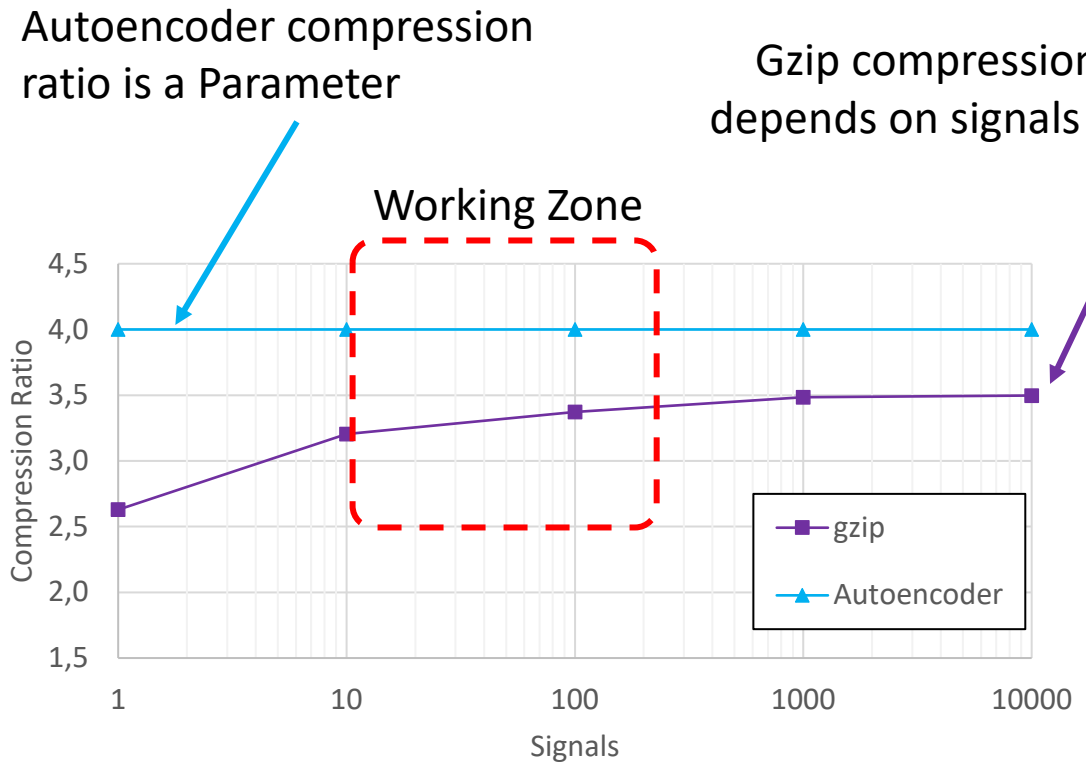
Rate can be managed for EVIO packet with at least 32 signals

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

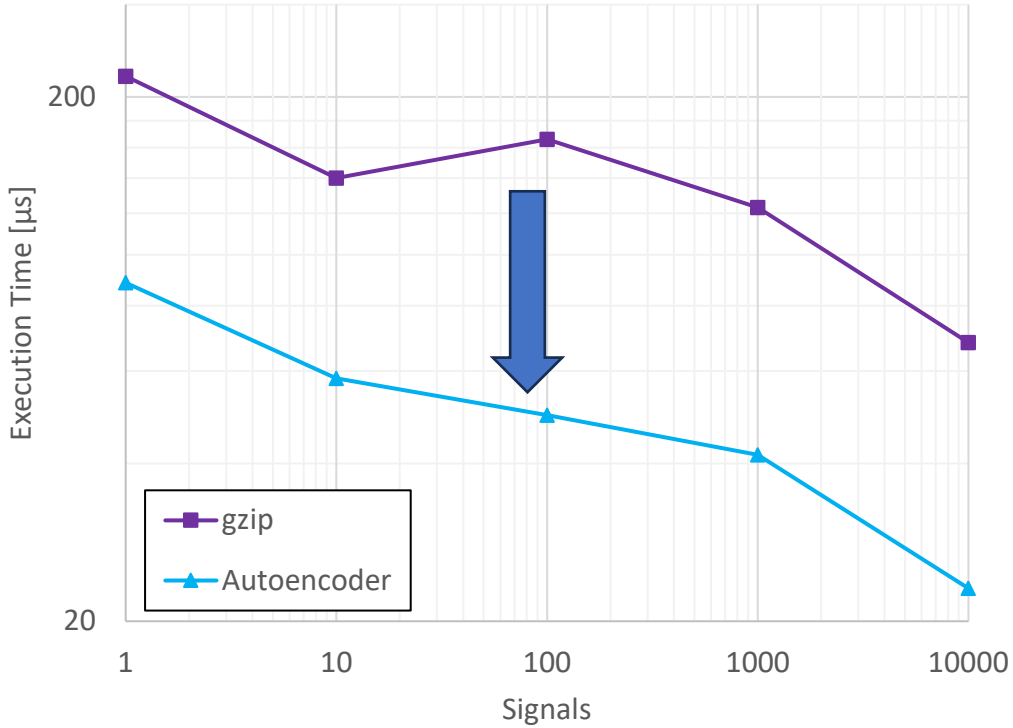
Comparison with standard lossless compression

Autoencoder compression ratio is a Parameter

Gzip compression ratio depends on signals number



gzip

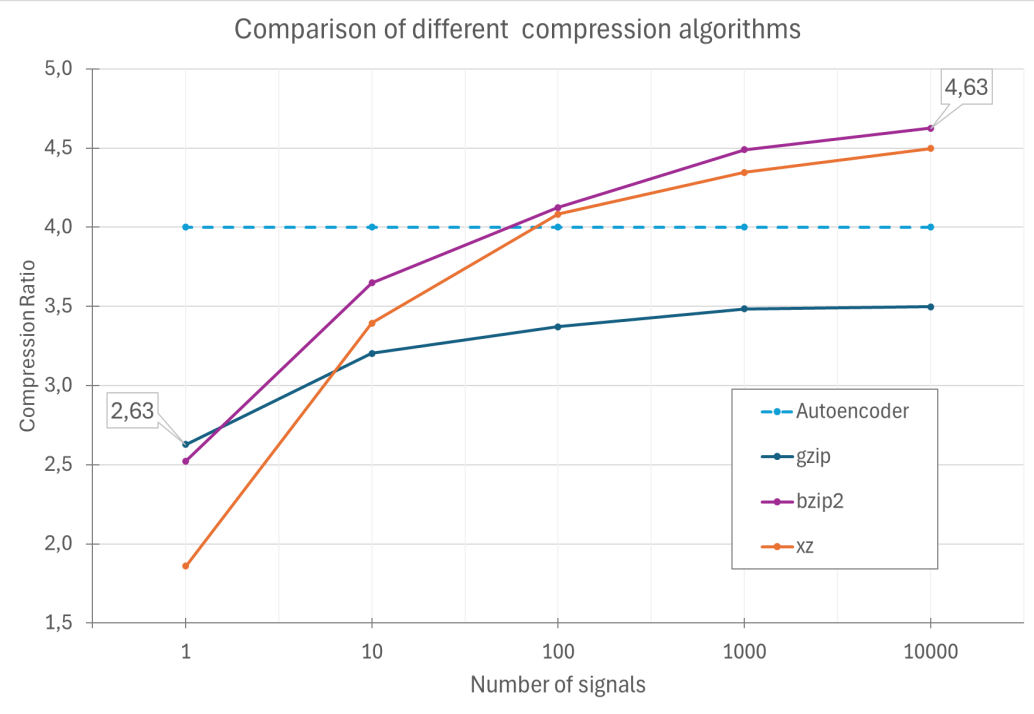


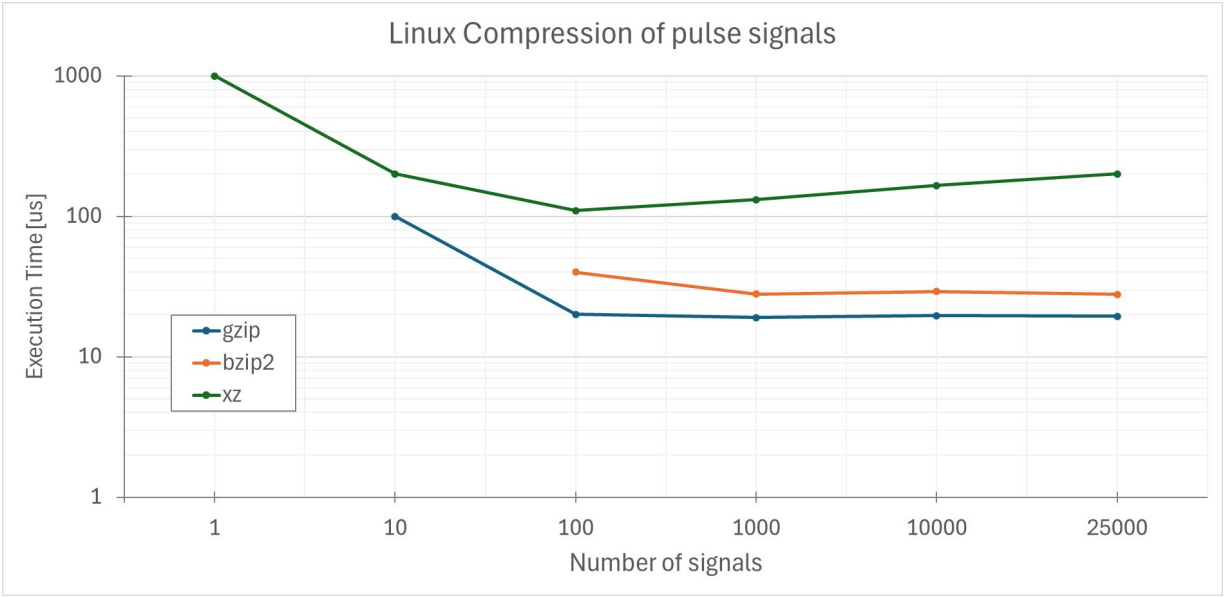
Better compression ratio

Better also on execution time

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Comparison with standard lossless compression





High compression ratio with Bzip2 and X

bzip2

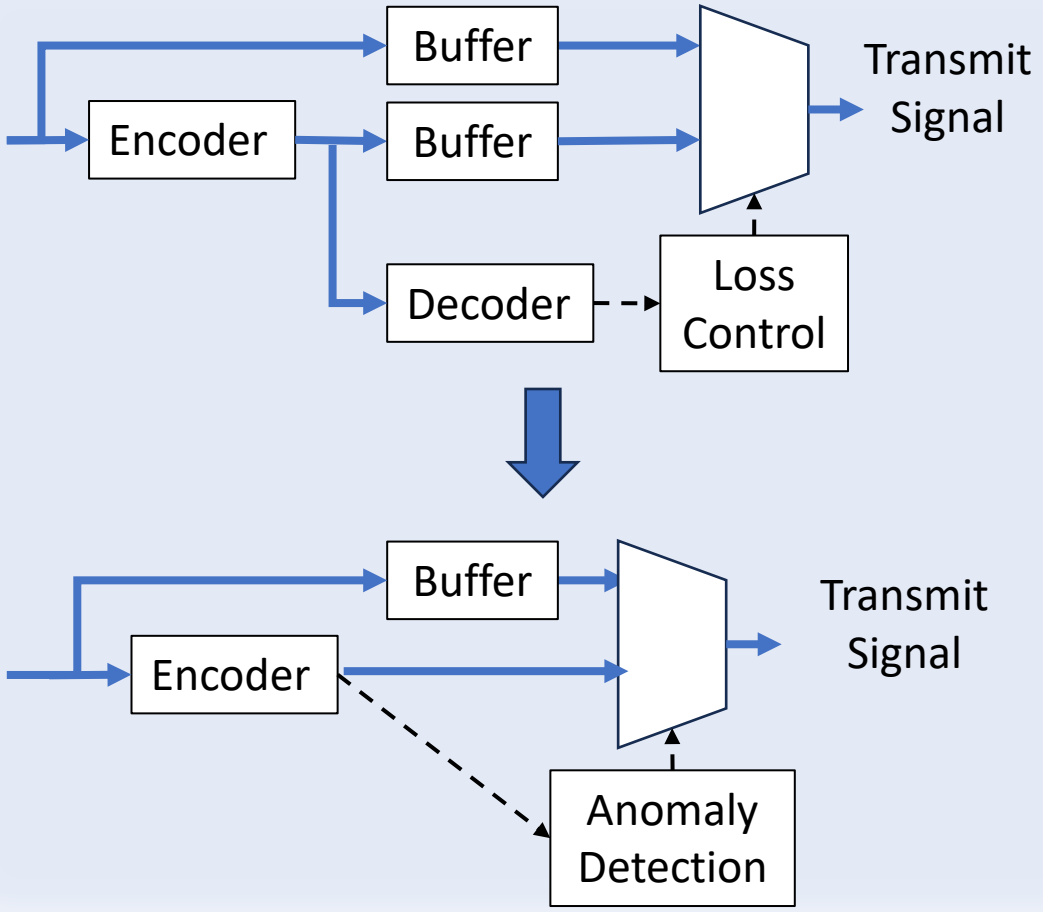
XZ

Very poor execution times:

bzip2	**50% slower**
XZ	**10 times slower**

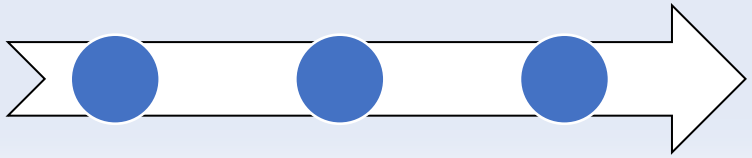
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Further Studies



Reduce execution time removing the decoder

Statistical analysis of signals in each EVIO packets



Estimate performance on real acquisition

Low level FPGA implementation

Dedicated connectivity (2xQSFP28 @ 100GbE)

...or very Low level

Reduce execution time and maybe save money

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Thank you for your attention

FUTURE AI RESEARCH

https://fondazione-fair.it

Jefferson Lab

https://www.jlab.org

INFN GENOVA

https://www.ge.infn.it

SEAlab

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.