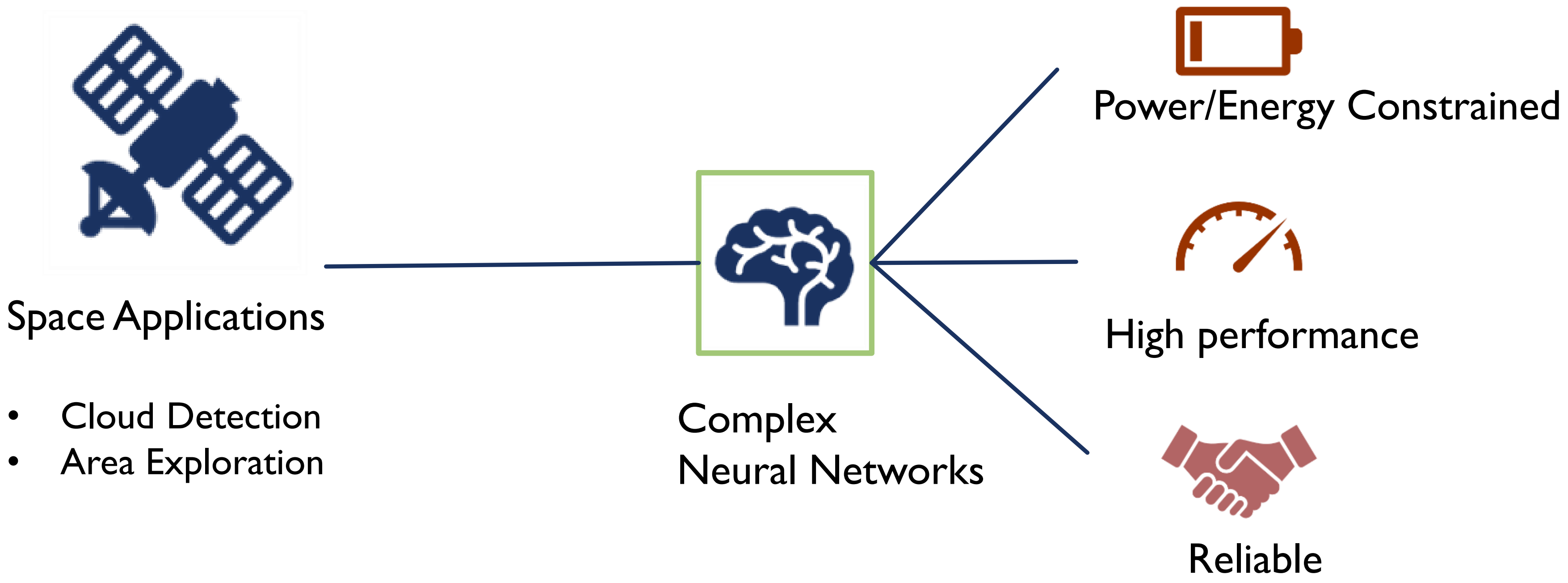
TRADEOFF BETWEEN PERFORMANCE AND RELIABILITY IN FPGA ACCELERATED DNNS FOR SPACE APPLICATIONS

Ioanna SOUVATZOGLOU* | **Dimitris AGIAKATSIKAS** *| Aitzan SARI* | Vasileios VLAGKOULIS* | Antonios TAVOULARIS** | Mihalis PSARAKIS*

*UNIVERSITY OF PIRAEUS, ** EUROPEAN SPACE AGENCY

EUROPEAN DATA HANDLING & DATA PROCESSING CONFERENCE – EDHPC 2023

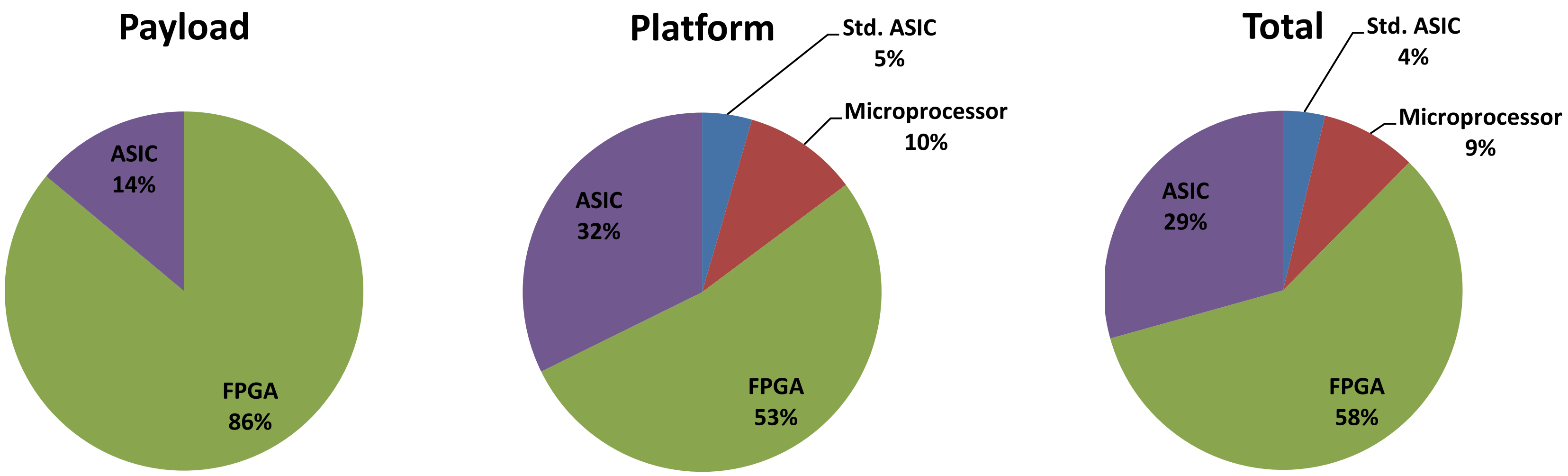
INTRODUCTION



FPGA IS SPACE
EXAMPLE → ESA SENTINEL 2 MISSION

Why:

- High performance
- Low non-recurring engineering costs
- Flexibility

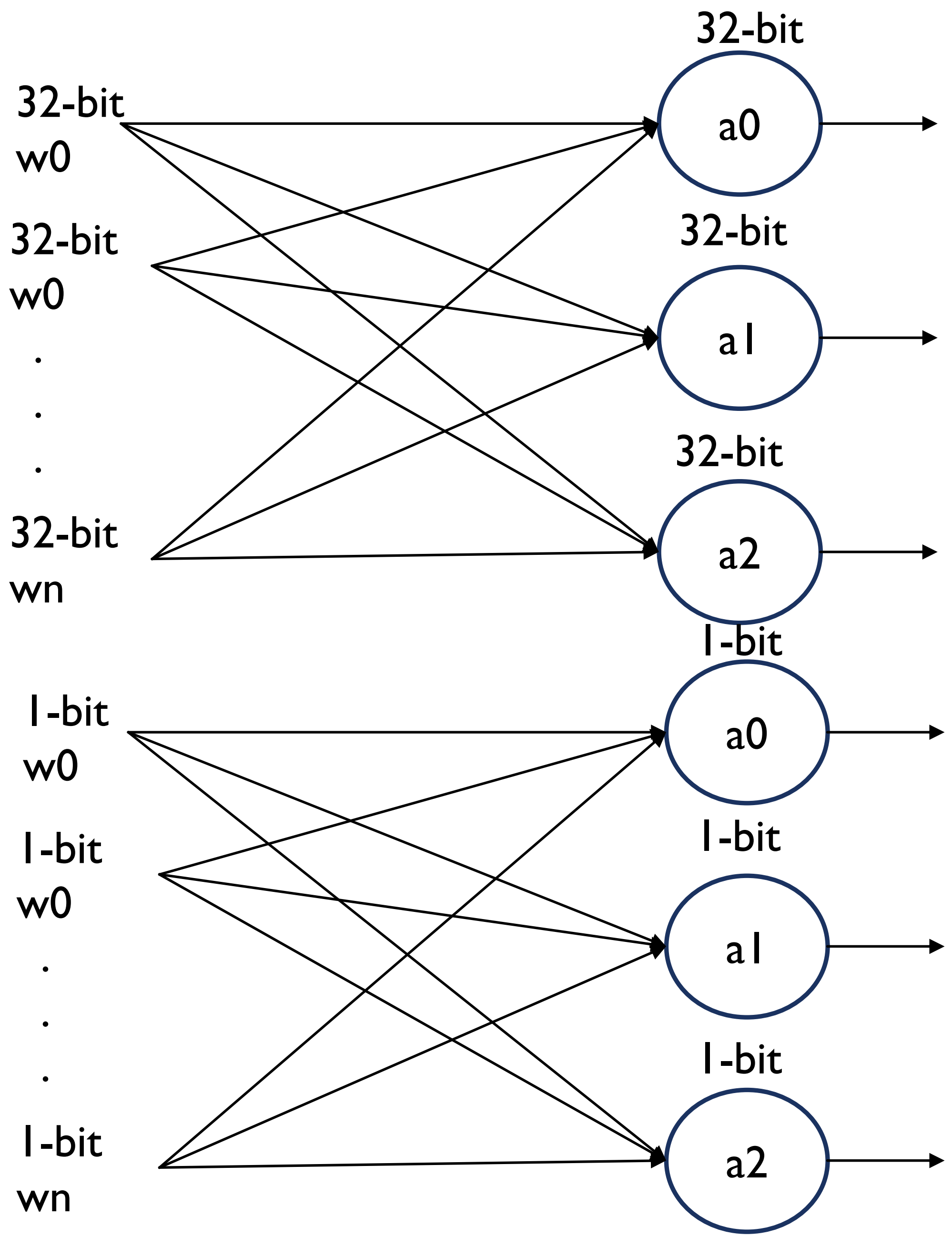


58% of all computing platforms are FPGAs compared to ASICs, Std. ASICs and Microprocessors

Trends and patterns in ASIC and FPGA use in space missions and impact in technology roadmaps of the European Space Agency, Roger Boada Gardenyes, Master Thesis, T. U. Delft and ESA, 15th August 2012

QUANTIZED NEURAL NETWORKS

- Approximate representation of data (on activations/weights)
- Binarized Neural Networks
  - 1-bit for Activation
  - 1-bit for Weights
- (+)Suitable for FPGAs
- (+)Consume less power
- (-)Slight degradation in the network accuracy

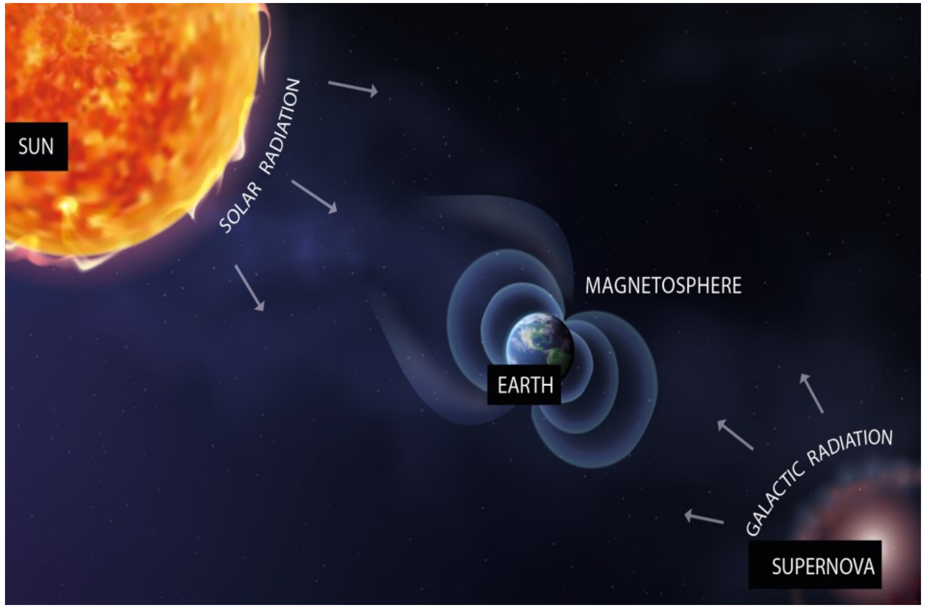


STANDAR NEURAL NETWORK

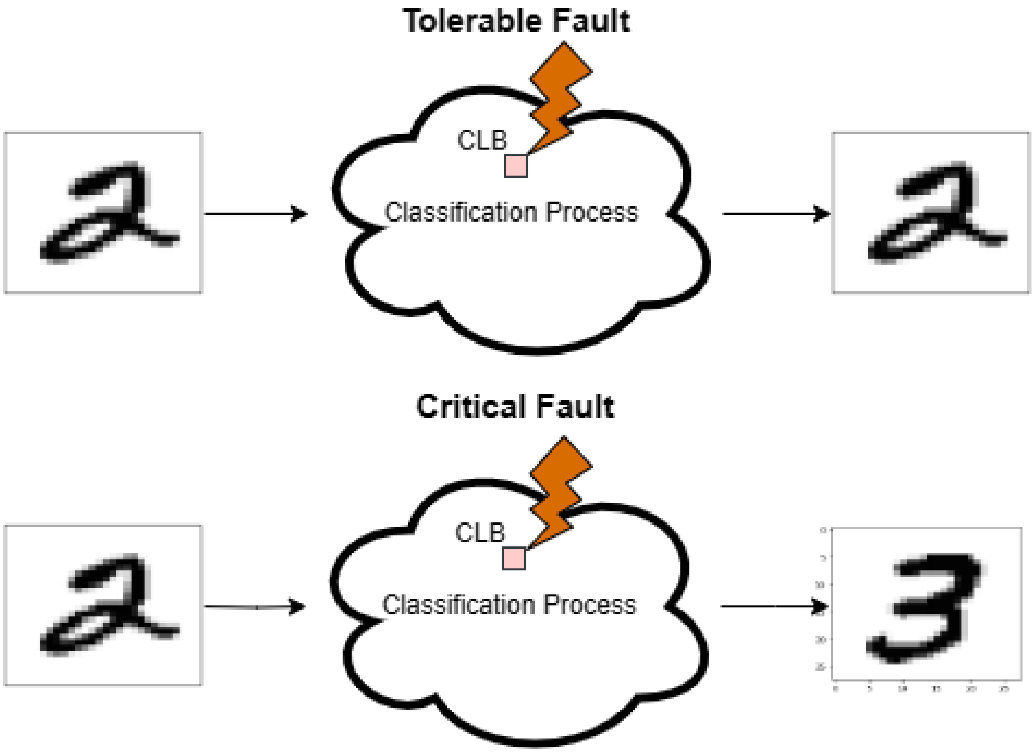
BINARIZED NEURAL NETWORK

THE RELIABILITY PROBLEM

COTS SRAM FPGAs are vulnerable SEEs caused by cosmic radiation



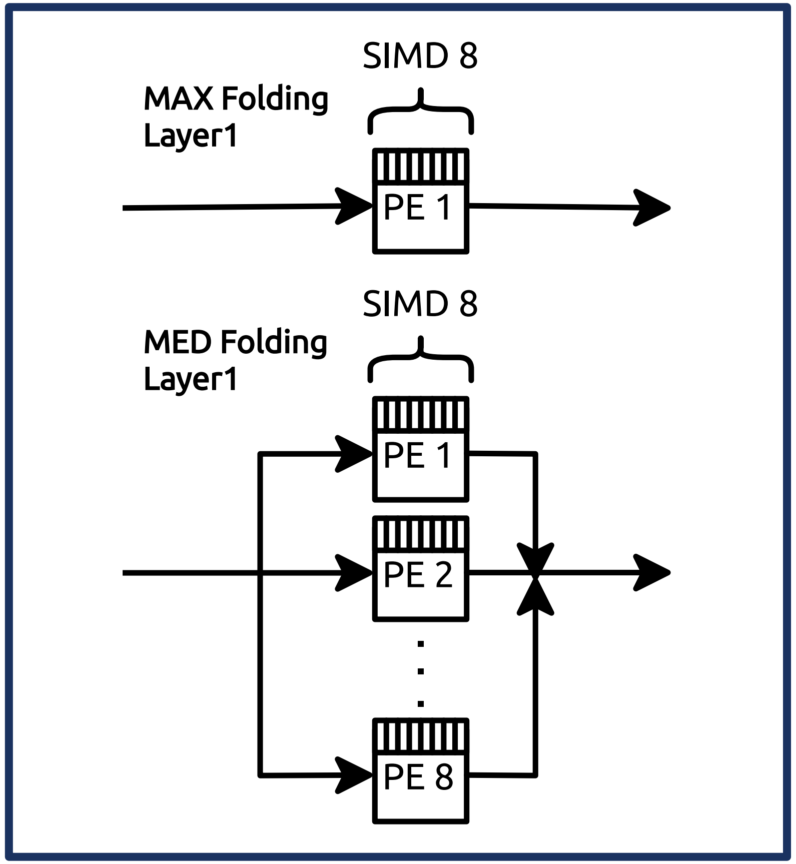
Inherently resilient to computational errors



DESIGN OPTIMIZATIONS FOR SRAM FPGA NNs

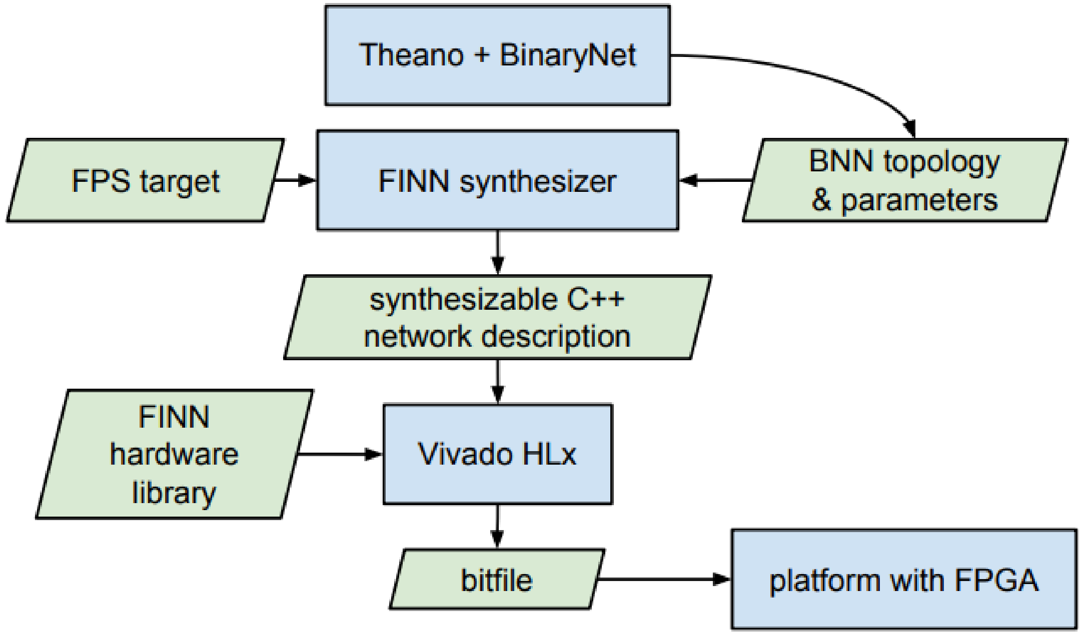
- Multiple optimizations for high demanding application (e.g., space applications):
  - Approximate computing
    - Approximate Adders/Multipliers
  - Quantization/Binarization
    - Int16, Int8, 2-bit, 1-bit Data representation of Weights/activations
  - Architecture optimizations
    - Parallelization/Folding of the Design
- In this work we investigate the effect of the folding parameter on the reliability and the performance of a binarized neural network

RELIABILITY VS DESIGN PARAMETERS (FOLDING)



- The folding design parameter indicates the level of parallelization
- When increasing the no. of PEs:
  - more resources are used
  - more neurons are active per cycle
  - reduces the classification execution time
- Example: 64 neurons:
  - Max. folding: 1 PEs → 64 cycles per neuron operation
  - Med. folding; 8 PEs → 8 cycles per neuron operation
  - Min. folding: 32 PEs → 2 cycles per neuron operation

CASE STUDY: FINN

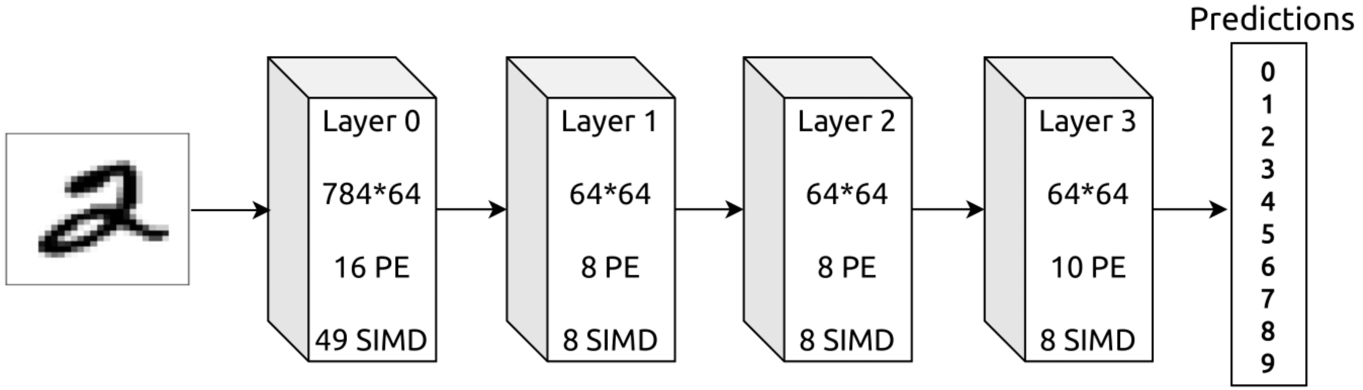


[*] Umuroglu, Yaman, et al. "Finn: A framework for fast, scalable binarized neural network inference." Proc. of the ACM/SIGDA Intl. Symp. on Field-Programmable Gate Arrays. 2017.

- FPGA Neural Network Accelerator
- Automated design flow
- Customized architectures for different network topologies
  - Fully Connected, Convolutional, Pooling
  - Wide range of data precisions
  - Performance parameters:
    - PEs, SIMD, folding, FIFOs, Memory components
- Training NN with :
  - Theano

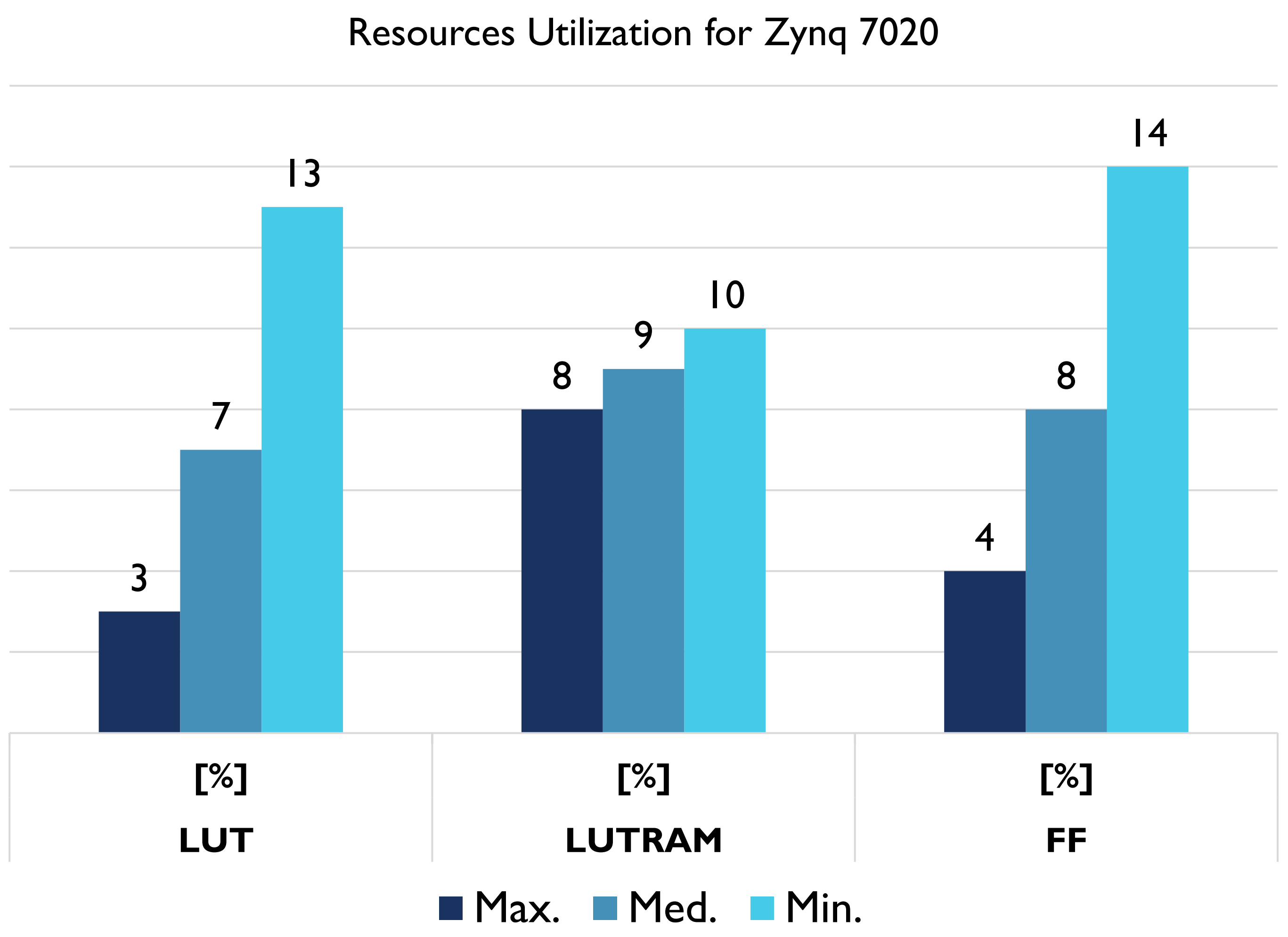
CASE STUDY: BINARIZED NEURAL NETWORK

- Binarized Neural Network – BNN
- Generated by FINN
- Fully Connected
- MNIST dataset
- Customizable Processing Elements and SIMD

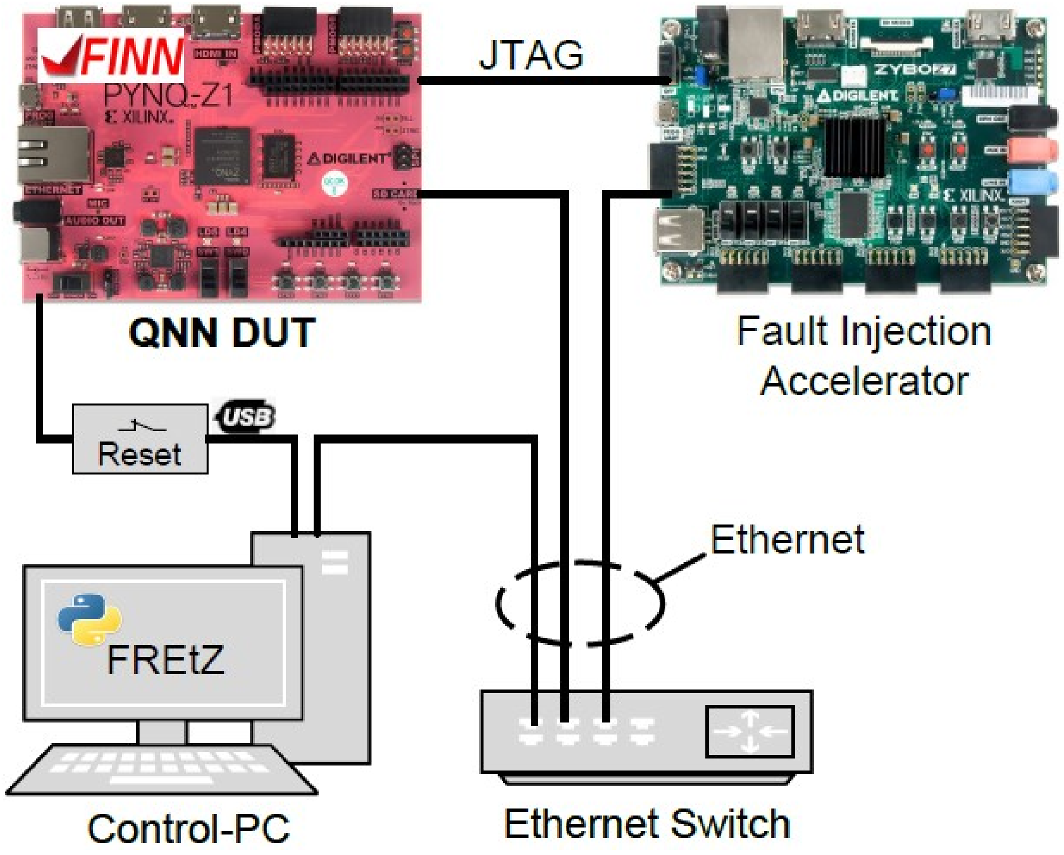


FOLDING: AREA VS PERFORMANCE

- 3 Designs (Max,Med,Min)
- Different no. of Processing Elements
- Execution Time [uS]:
  - Max. folding - 21.4
  - Med. folding - 2.16
  - Min. folding - 0.88



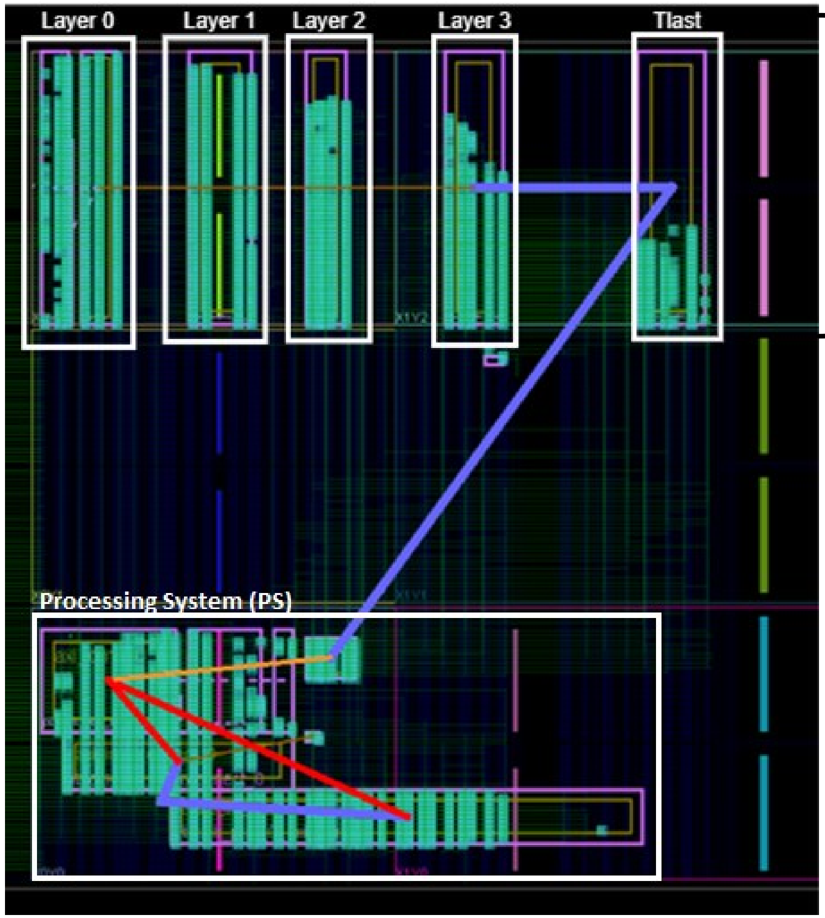
FAULT INJECTION CAMPAIGN: SETUP



- DUT: FINN MNIST CLASSIFICATION on Zynq-7020
- Opensource FREtZ Framework
  - Bitstream manipulation
  - Fault injection
  - Read/Write configuration memory
- Fault Injection Accelerator:
  - JTAG connection with the DUT
  - Accelerates the fault injection procedure
- Fault Model
  - SBU in the essential configuration bits
  - FPGA statistical fault injection campaign
    - Confidence Interval (CI) = 99%
    - Margin of error = 0.3%

FAULT INJECTION CAMPAIGN: FLOW

- Statistical Fault injection through JTAG
  - Extract Essential Bits, through Bitstream Manipulation
  - Divide them per Layer (Constrained placement)
  - Create a Fault List
  - For each Fault in Fault List:
    1. Synchronize with ARM (Boundary Scan)
    2. Read Configuration Frame
    3. Bit Flip essential bit of Frame
    4. Write Configuration Frame



Bit flips injected only in the layers of the NN

RELIABILITY ANALYSIS METRICS

Reliability

$$\text{AVF} = \frac{\#\ \text{Critical faults}}{\#\ \text{Bit Flips}}$$

$$\text{MTBF} = \frac{1}{\lambda} = \frac{1}{\text{Ebits} \times \text{AVF} \times \lambda_{\text{CRAM}}}$$

Reliability & Performance

$$\text{MEBF} = \frac{\text{MTBF}}{\text{Mean classification time per image}}$$

FAULT INJECTION CAMPAIGN RESULTS & AVF

Folding	Bits		Failure Rates [%]							AVF
	Essential Bits	Upsets injected	Layers	Tolerable	**Total Failures**	Crashes	Critical	Zeroes		
Max.	684666	145211	**Overall**	**3.86**	**3.76**	**2.79**	**0.62**	**0.35**		**3.755E-02**
			0	1.76	1.93	1.44	0.26	0.23		1.925E-02
			1	0.6	0.76	0.57	0.14	0.05		7.623E-03
			2	0.75	0.69	0.52	0.13	0.04		6.852E-03
			3	0.75	0.38	0.27	0.09	0.02		3.829E-03
Med.	1289901	161258	**Overall**	**7.29**	**1.72**	**0.84**	**0.72**	**0.16**		**1.720E-02**
			0	3.76	1.05	0.50	0.44	0.11		1.053E-02
			1	0.74	0.23	0.13	0.08	0.02		2.270E-03
			2	1.1	0.22	0.13	0.07	0.02		2.096E-03
			3	1.69	0.23	0.08	0.14	0.01		2.301E-03
Min.	2220495	170174	**Overall**	**6.88**	**1.36**	**0.66**	**0.55**	**0.15**		**1.371E-02**
			0	3.57	0.82	0.38	0.34	0.10		8.262E-03
			1	0.93	0.2	0.11	0.07	0.02		1.992E-03
			2	1.39	0.21	0.11	0.08	0.02		2.104E-03
			3	0.99	0.14	0.05	0.07	0.02		1.352E-03

Folding Max → Min

- Essential bits increase
- AVF decreases

Per layer analysis

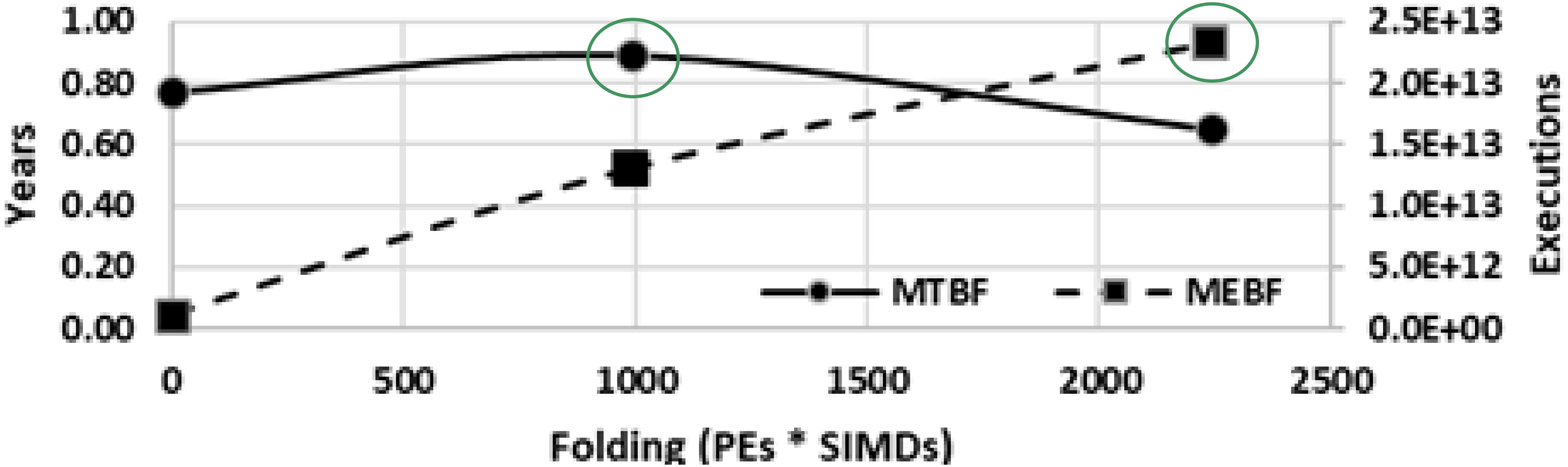
- Layer 0 most vulnerable

RELIABILITY METRICS: MTBF & MEBF

Ebits = 685Kbit,
AVF = 3.755E-02
Exec = 21.4uS
Max.

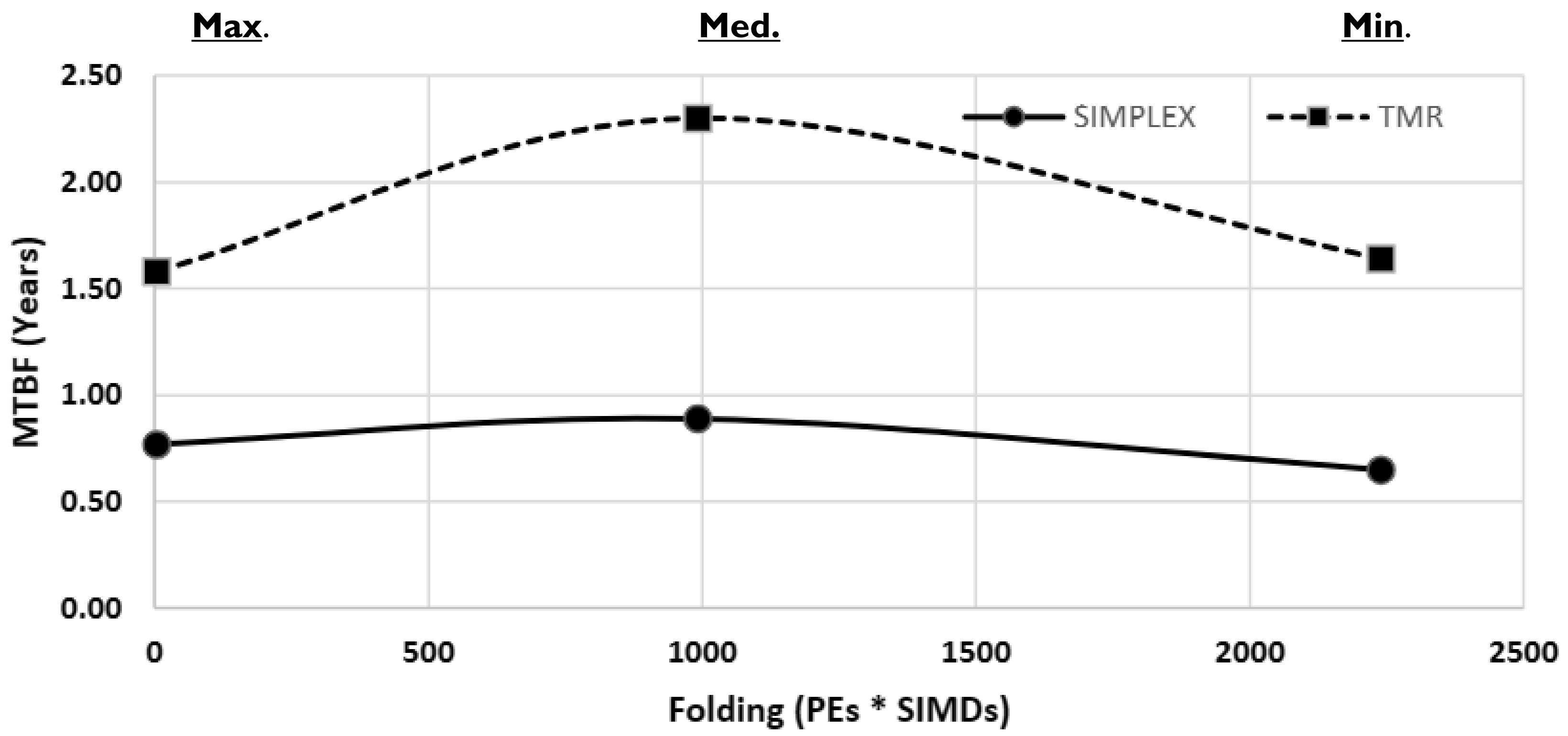
Ebits = 1,290Kbit,
AVF = 1.720E-02
Exec = 2.16uS
Med.

Ebits = 2,220Kbit,
AVF = 1.371E-02
Exec = 0.88uS
Min.



Assuming 4.48 upsets/device/day as for a mission in Low Earth Orbit (404 km perigee, 407 km apogee and 51.64°)

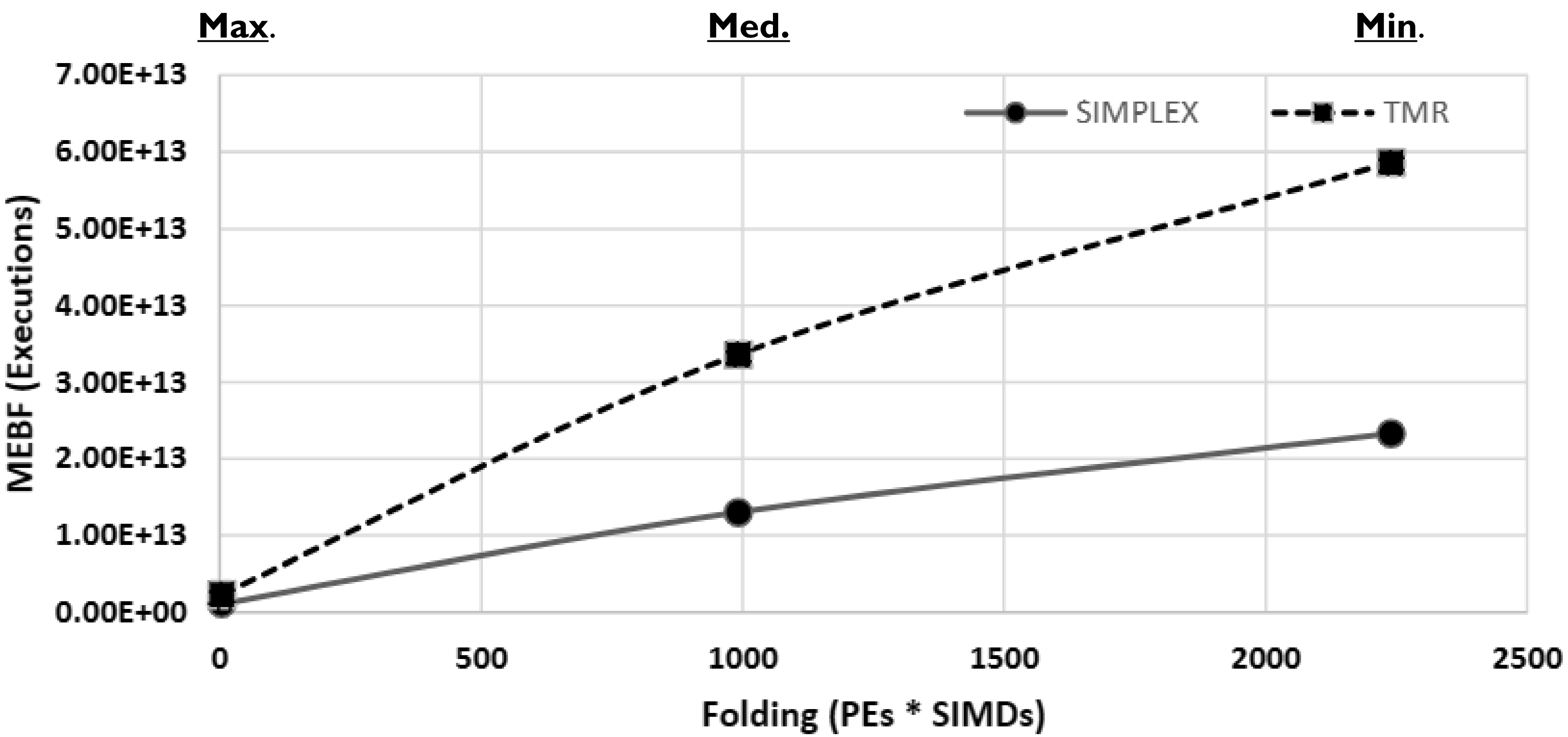
SELECTIVE TMR - LAYER0: MTBF



MTBF with Selective TMR :

- ~x2.5 times better in Med and Min QNN
- ~x2 times better in Max QNN

SELECTIVE TMR - LAYER0 : MEBF



- Selective TMR lead to ~x2 Times Better MEBF in every QNN

CONCLUSION

- For highest MEBF → Highest parallelization
- For highest MTBF→ Need design exploration of the folding factor

UNIVERSITY OF PIRAEUS ES-Labs esa

THANK YOU

CONTACT INFO:

ISOUVATZ@UNIPI.GR