FPG-AI: a Technology Independent Framework for Edge AI Deployment Onboard Satellite, and its Characterisation on NanoXplore FPGAs

ESA OSIP Idea - Early Technology Development

Recent Advances in European Space FPGAs: Technologies and Applications

Pisa, 11 June 2024

Presenter: Pietro Nannipieri

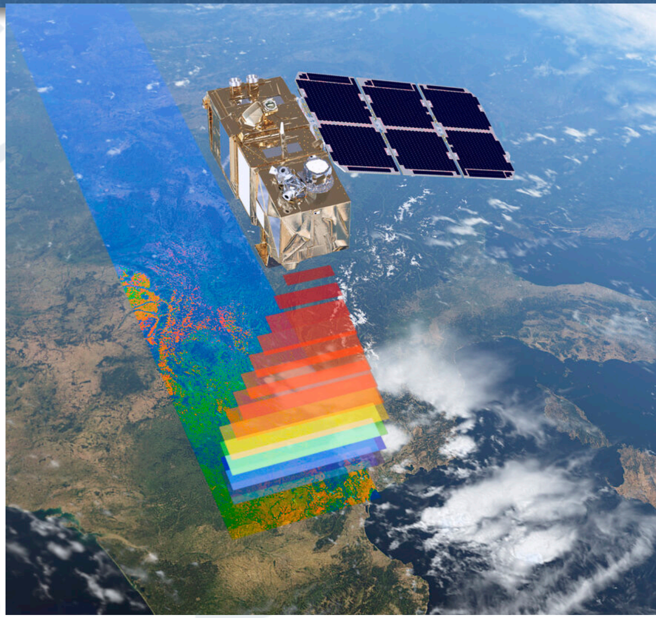
Presentation Outline

- Activity Context and Background
- Proposal Objectives
- Implementation on NX FPGAs
- Hardware Prototyping
- Conclusions

Presentation Outline

- Activity Context and Background
- Proposal Objectives
- Implementation on NX FPGAs
- Hardware Prototyping
- Conclusions

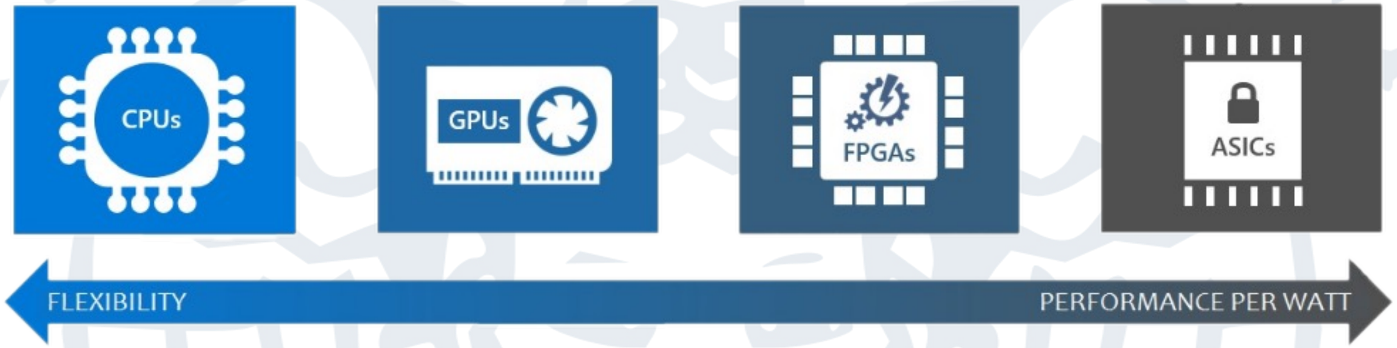
Activity Context



- Growing interest in **AI for space applications**:
  - Weather and Atmospheric Monitoring
  - Object Detection and Tracking
  - Ground Classification
  - Fault Detection, Isolation, and Recovery for Reliability
  - Autonomous Spacecraft Navigation
- AI deployment onboard the satellite constitutes an **open challenge**
- **Satellites** are **resource-** and **power- constrained** devices operating in a **harsh environment**
- The **complexity of AI models** collides with the limitations of satellite platforms

Activity Context

- Multiple hardware technologies are being investigated for AI acceleration onboard the satellite:



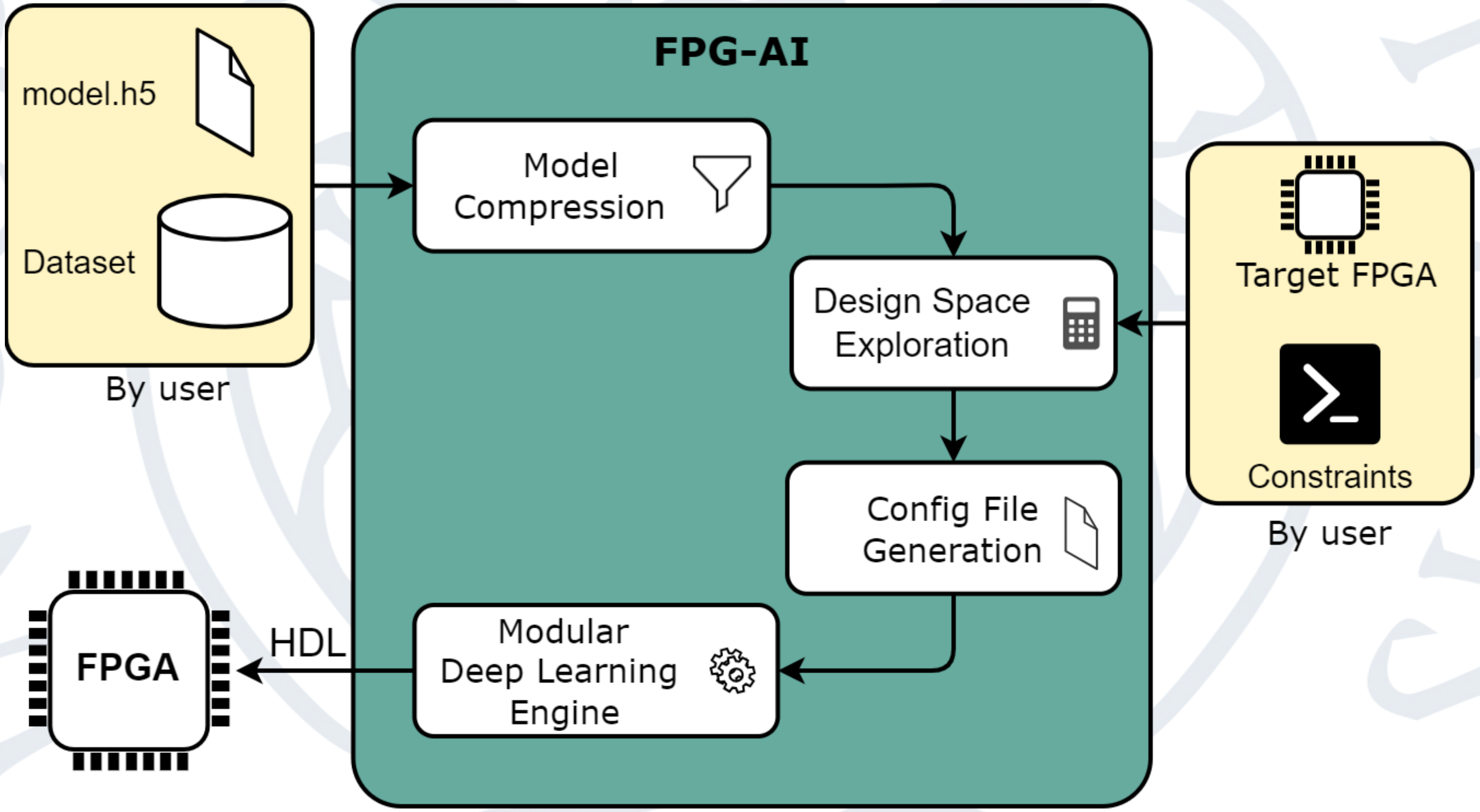
- FPGAs are a promising technology for AI acceleration for their energy efficiency and radiation tolerance

- The design of FPGA-based accelerators for AI typically requires high design expertise and long time-to-market

- Growing interest in **DNN-to-FPGA automation toolflows** for rapid AI deployment onboard the satellite

Background: FPG-AI Toolflow for CNNs

- Automation toolflow for efficient deployment of pre-trained CNN models on FPGA technology [1], [2]



FPG-AI Key Features

- **Easy integration in user-defined SoCs:**
  - Providing as output the accelerator HDL sources and not the final bitstream
  - Possibility to tune the resource consumption according to the requirements of other IPs
  - No workload sharing with the Host-CPU
- **Unmatched device portability of the Modular Deep Learning Engine (MDE) thanks to:**
  - Absence of third-party IPs
  - High scalability in terms of DSP/On-chip memory usage
  - Fine-grain configurable through a .vhd file

➡ Enabling the implementation on FPGAs from different vendors and heterogeneous resource budgets!

XILINX MICROCHIP intel

Presentation Outline

- Activity Context and Background
- **Proposal Objectives**
- Implementation on NX FPGAs
- Hardware Prototyping
- Conclusions

FPG-AI: Extension to NanoXplore FPGAs

→ THE EUROPEAN SPACE AGENCY

ACTIVITY

FPG-AI: A TECHNOLOGY INDEPENDENT FRAMEWORK FOR EDGE AI DEPLOYMENT ONBOARD SATELLITE, AND ITS CHARACTERISATION ON NANOXPLORE FPGAS

Overview Events

RUNNING

Prime contractor
UNIVERSITA DI PISA

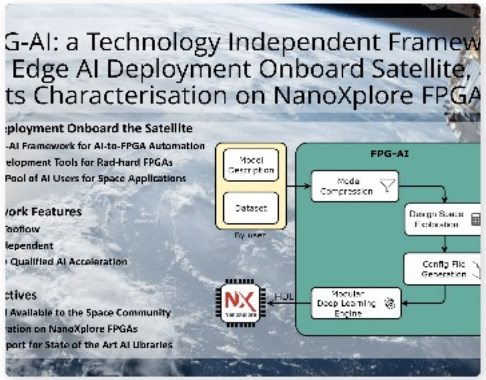
Organisational Unit
TEC-SF

START
31 March 2023

ESTIMATED END
01 October 2024

DURATION: 18 MONTHS

Activity Type
Early technology development



https://activities.esa.int/index.php/4000141108

Proposal Objectives

- "Extending and consolidating the framework to a wider set of supported AI algorithms, e.g. Recurrent Neural Networks (RNNs)"
- "Ensuring that all state-of-the-art devices are supported by the tool, especially focusing NanoXplore (NX) FPGAs, enabling the use of these devices for AI applications and pursuing European sovereignty"
- "Evaluating the tool capability with a prototype hardware demonstrator"

Presentation Outline

- Activity Context and Background
- Proposal Objective and Organization
- Implementation on NX FPGAs
- Hardware Prototype
- Conclusions

Implementation on NX FPGAs

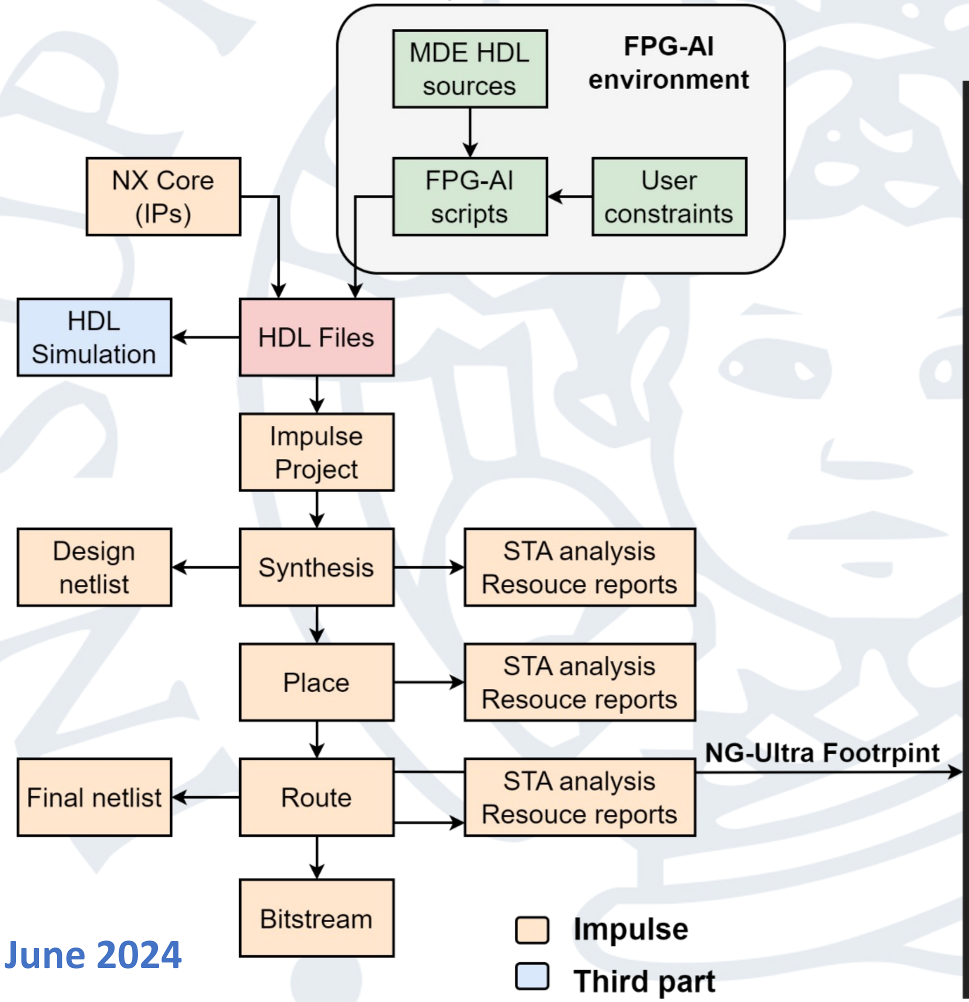
- **Study of NX Technology and Design Suite**:
  - Ramp up on NX Design Suite (Impulse 23.3.0.2)
  - Study of NanoXplore on-chip memory and DSP resources
  - **20/07/2023**: One-day visit to NanoXplore headquarters in Paris for early feedbacks on Impulse flow and for identifying the hardware platform

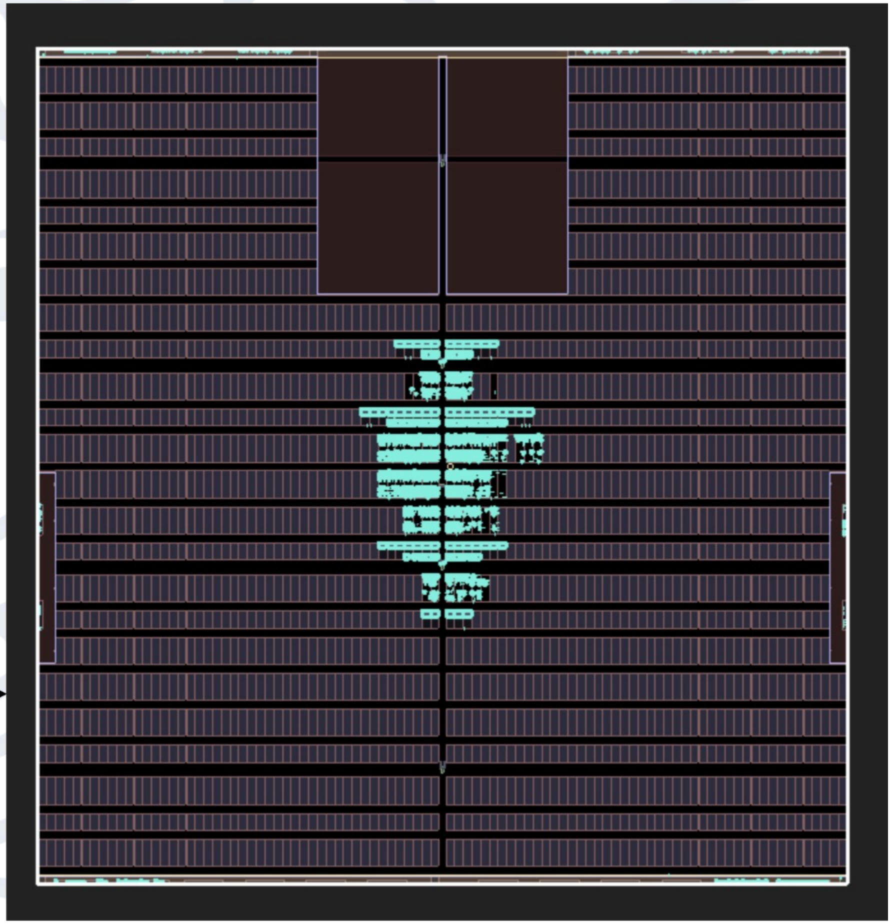
    ➔ **NG-Ultra FPGA** selected as the target device

- **Selected Case Studies**:
  - **LeNet-5**:
    - Digits recognition on MNIST dataset
    - Layers: 2x[Conv + AvgPool] + 3 Dense
    - 44K parameters (~1.36 Mbit)
  - **Network in Network (NiN)**:
    - Image classification on CIFAR10 dataset
    - Layers: 9 Conv + 1 GlobalAvgPool
    - 969K parameters (~29.68 Mbit)

Implementation on NX FPGAs:

Design Flow: NX Impulse





Implementation on NX FPGAs

- **Upgrade of FPG-AI hardware architecture for NX technology**:
  - Main issue: low implementation frequency
  - Pin-point changes to the architectural stage for CNNs to reduce the critical path
  - Frequency increased from **28.6 MHz up to 43.0 MHz for LeNet-5, 15 MHz from up to 25.6 MHz for NiN** (for the MDE only)

- **Exploitation of FPG-AI and NX development tools to obtain implementation results:**
  - Summary of the collected results on **NG-Ultra**:

	LeNet-5	NiN	MobileNet	VGG16
1 PE	✓	✓	On-going (actively working with NX support team)	
16 PE	✓	✓		

Implementation on NX FPGAs

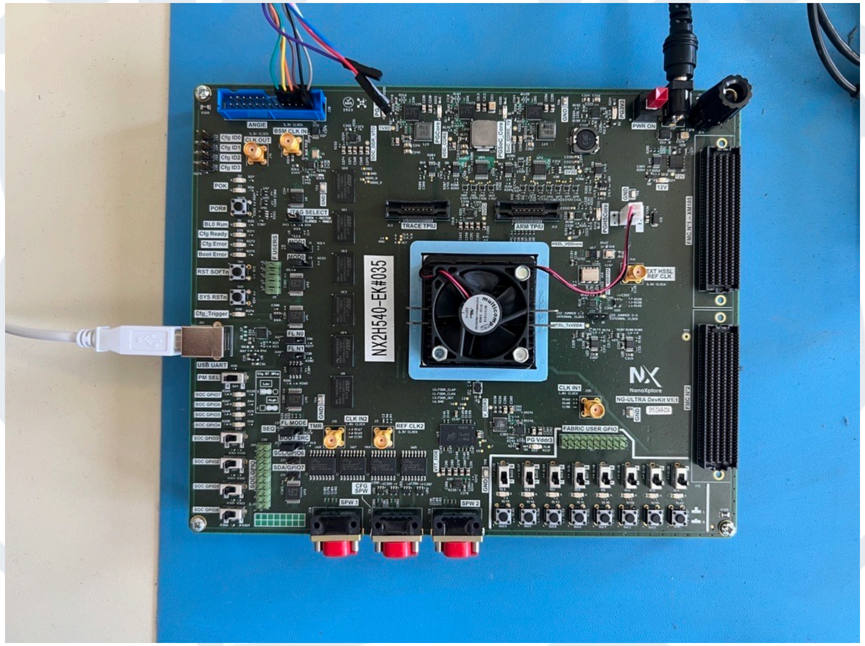
Model	#PE	LUT	FF	Register File Block	DPRAM	DSP	MDE Frequency [MHz]	AXI Frequency [MHz]
LeNet	1	9197 (2%)	4631 (1%)	89 (4%)	29 (5%)	51 (4%)	30.82	33.7
	16	13252 (3%)	5415 (2%)	38 (2%)	149 (23%)	426 (32%)	24.16	34.17
NiN	1	29503 (5.5%)	11433 (2.3%)	0 (0%)	340 (50.6%)	41 (3.1%)	20.8	24.4
	16	38247 (8%)	12450 (3%)	0 (0%)	297 (45%)	415 (31%)	19.5	33.8

Presentation Outline

- Activity Context and Background
- Proposal Objectives
- Implementation on NX FPGAs
- Hardware Prototype
- Conclusions

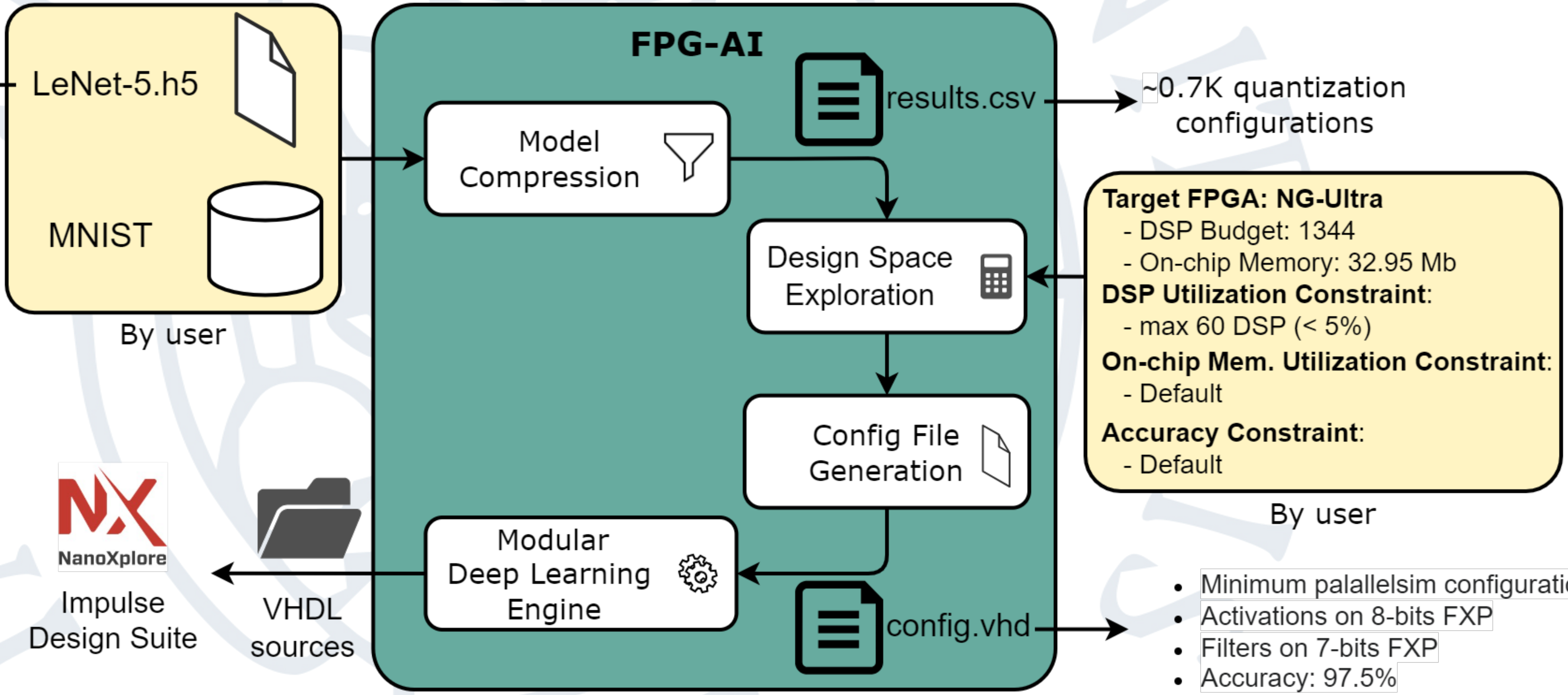
Model and Platform Selection

- **Selection of a development platform hosting a NX FPGA**:
  - NG-Ultra Devkit Board v1.1, suggested by NanoXplore and kindly received on loan from ESA Microelectronic Section
- **Identification of the DNN model to be characterized in hardware**:
  - LeNet-5 selected as the target model (light and commonly used model)



Accelerator Generation with FPG-AI

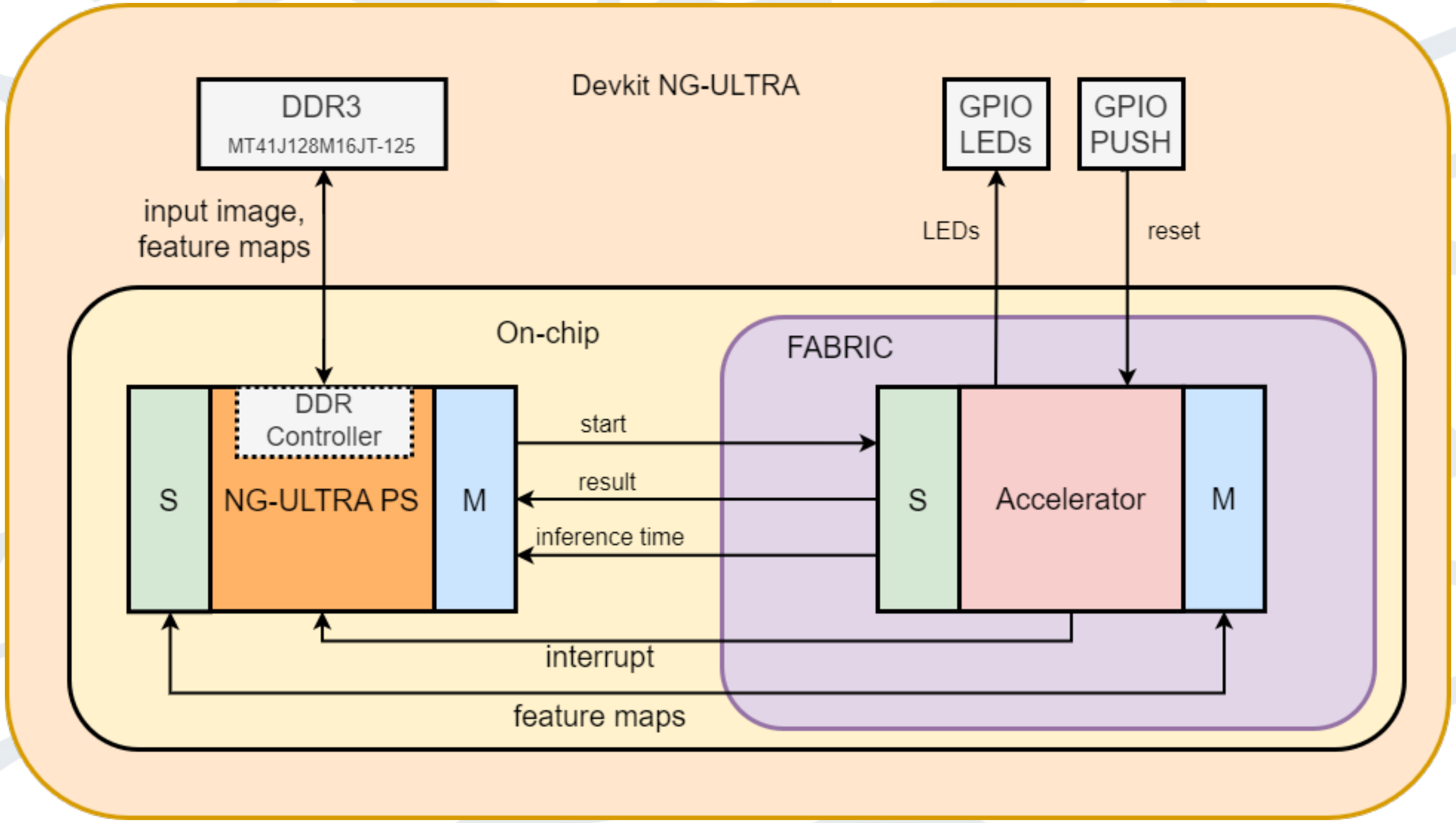
- Activations on 32-bits FP
- Filters on 32-bits FP
- Accuracy: 97.5%



- Minimum palallelsim configuration
- Activations on 8-bits FXP
- Filters on 7-bits FXP
- Accuracy: 97.5%

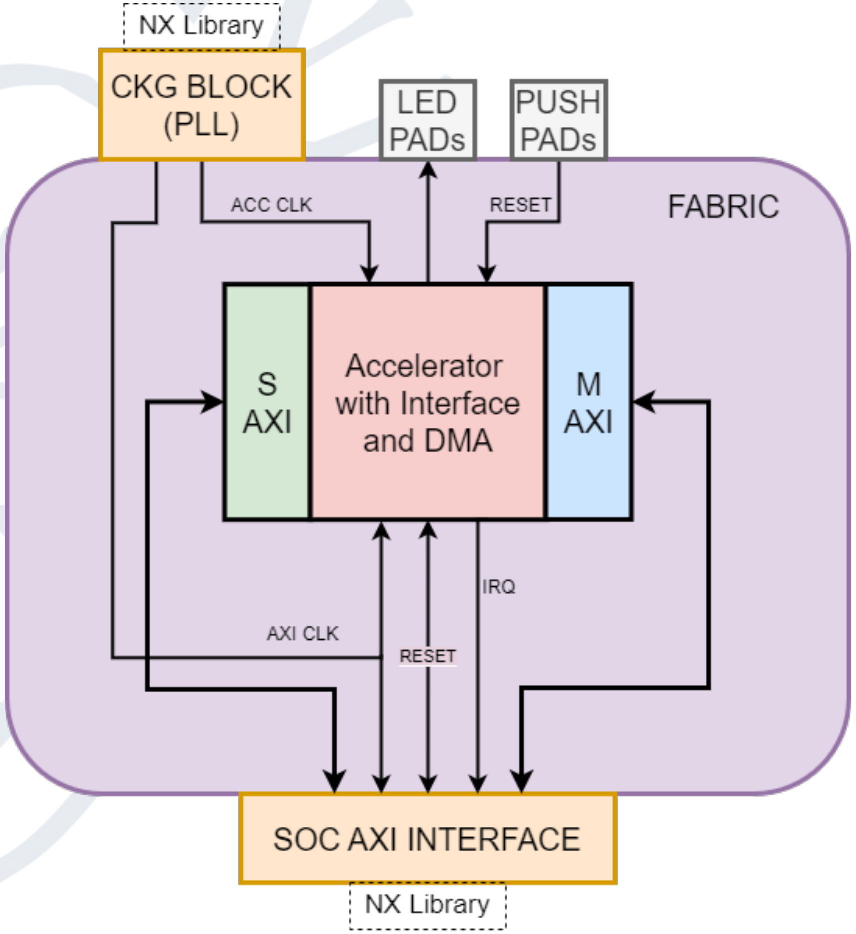
FPG-AI Integration on NG-Ultra Devkit SoC

➢ **Hardware prototype concept:**



PL/PS Interface

- Extension of the accelerator AXI interface to 128-bit to ensure compliance with NG-Ultra SoC
- Instantiation of the AXI SoC interface component (NX Library) in the VHDL top level to connect the PS and the PL
- Development of C code to initialize the DDR memory and to control the accelerator
- Clock generation with PLL
- Performed test to validate communication between PS and accelerator's register file (AXI Slave interface)
  - Currently working on that, thanks to **recently released NX-scope logic analyzer support for NG-Ultra Dev-kit**
  - **FPGA demo to be announced soon**



Detailed Results & Benchmarking

- **Selected FPGAs for the comparison (T5.1):**
  - NanoXplore NG-ULTRA (28 nm)
  - Microchip Polarfire MPF500T (28 nm), similar resources to RTPF500T
  - Microchip RTG4 (65 nm)
  - AMD Space-Grade Kintex Ultrascale XQRKU060 (20 nm)
  - AMD Zynq 7000 XC7Z045 (28 nm)

- **Selected case studies (T5.2)**:
  - **LeNet-5**:
    - Digits recognition on MNIST dataset
    - Layers: 2x[Conv + AvgPool] + 3 Dense
    - 44K parameters (~1.36 Mbit)
  - **Network in Network (NiN)**:
    - Image classification on CIFAR10 dataset
    - Layers: 9 Conv + 1 GlobalAvgPool
    - 969K parameters (~29.68 Mbit)

Detailed Results & Benchmarking

#PE	1				
Device	NG-ULTRA	MPF500T	RTG4	XQRKU060	XC7Z045
LUT	8745 (2%)	8889 (1.8%)	8928 (5.9%)	4824 (1.5%)	6014 (2.8%)
FF	3983 (1%)	4751 (1.0%)	5048 (3.3%)	4031 (0.6%)	3688 (0.8%)
RF/LUTRAM/ µSRAM	16 (1%)	22 (0.5%)	15 (7.1%)	148 (0.1%)	172 (0.2%)
DPRAM/BRAM/LSRAM	44 (7%)	30 (2.0%)	32 (15.3%)	16.5 (1.5%)	16.5 (2.11%)
DSP	51 (4%)	62 (4.2%)	62 (13.4%)	59 (2.1%)	59 (6.56%)
MDE Frequency [MHz]	29.4	67.6	51.5	114.9	100
AXI Frequency [MHz]	41.7	125.0	82.6	161.3	200
Timing Efficiency [GOP/s]	0.862	1.724	1.316	2.92	2.55
DSP Efficiency [GOP/s/#DSP]	0.0169	0.0278	0.0212	0.049	0.043
RAM [Mbit]	2.071	0.602	0.772	0.589	0.591
RAM Efficiency [GOP/s/Mb]	0.416	2.864	1.705	4.958	4.315

Detailed Results & Benchmarking

#PE	1			
Device	NG-ULTRA	MPF500T	XQRKU060	XC7Z045
LUT	29503 (5.5%)	20058 (4.2%)	18794 (5.7%)	24485 (11.2 %)
FF	11433 (2.3%)	8807 (1.8%)	11992 (1.8%)	13813 (3.2%)
RF/LUTRAM/ µSRAM	0 (0%)	0 (0%)	0 (0%)	0 (0%)
DPRAM/BRAM/LSRAM	340 (50.6%)	918 (60.4%)	340 (31.5%)	344.5 (63.2%)
DSP	41 (3.1%)	39 (2.6%)	35 (1.3%)	35 (3.9%)
MDE Frequency [MHz]	20.8	55.6	79.37	83.33
AXI Frequency [MHz]	24.4	126.1	161.3	161.3
Timing Efficiency [GOP/s]	0.423	1.127	1.610	1.690
DSP Efficiency [GOP/s/#DSP]	0.0103	0.0289	0.046	0.048
RAM [Mbit]	15.94	17.93	11.95	12.11
RAM Efficiency [GOP/s/Mb]	0.027	0.063	0.135	0.140

Presentation Outline

- Activity Context and Background
- Proposal Objective and Organization
- FPG-AI Extension to RNNs
- Implementation on NX FPGAs
- Hardware Prototype
- Conclusions

Project Outcome

- **FPG-AI: end-to-end toolflow for the acceleration of DNNs on FPGAs**
  - Technology-independent flow: possibility to target FPGAs from Xilinx, Intel, Microsemi, and NanoXplore
  - Easy integration in user-defined SoCs and high degree of customization

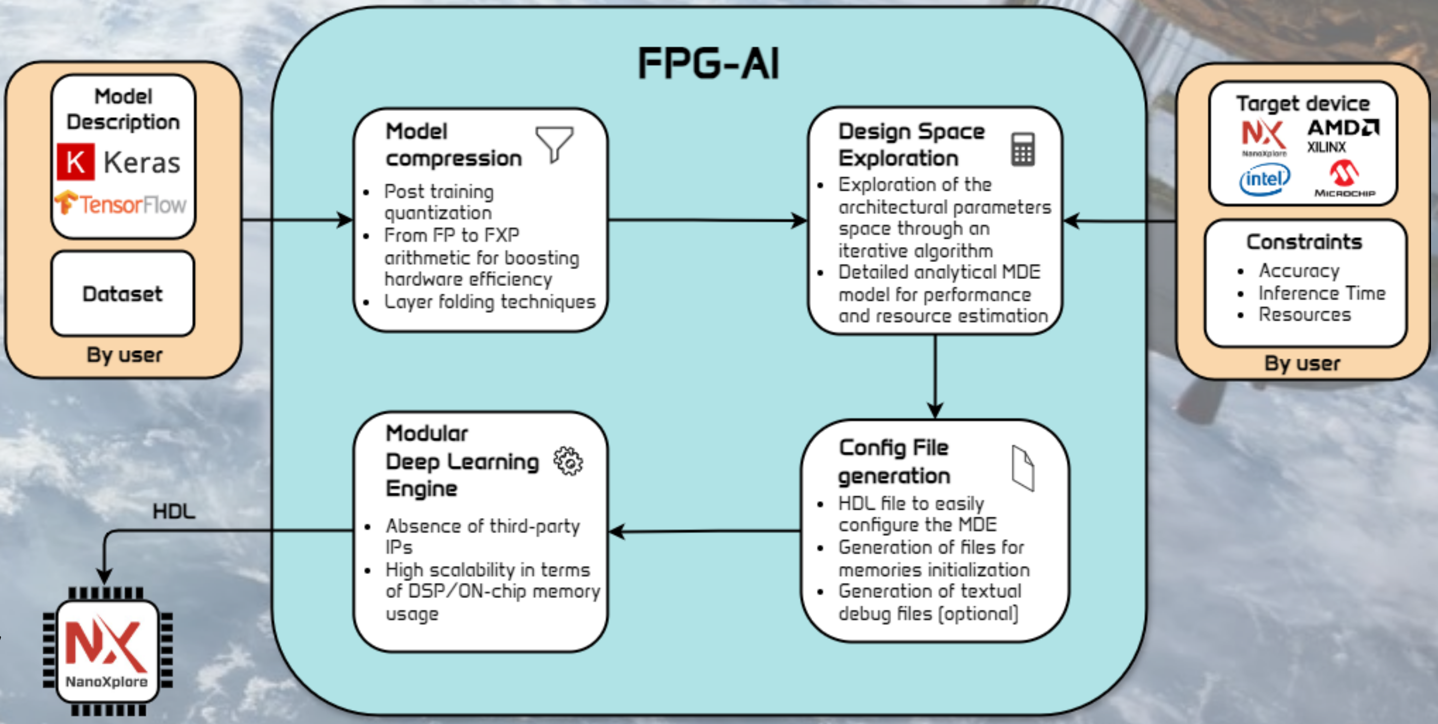
- **Extension to Recurrent Neural Networks (RNNs):**
  - Achieved implementation results on multiple RNN-FPGA pairs
  - Toolflow characterized for Fault Detection and Sequence Classification tasks

- **Extension to NanoXplore technology:**
  - Achieved implementation results for two CNN models targeting the NG-Ultra device
  - Deployed FPG-AI’s accelerator on a Zynq ZCU106 Development Board to evaluate the flow
  - Built a solid expertise on NX flow that will be used to finalize the hardware prototype on NG-ULTRA

Thanks for the attention!

FPG-AI Framework Features

- Ready-to-use Tooflow
- Supporting for both CNN and RNN models
- Technology Independent HDL
- Extremely portable solution
- Enabling Space Qualified AI Acceleration



Project Technical Outcomes

- ✓ Made FPG-AI Available to the Space Community
- ✓ Designed support for LSTM and GRU RNN layers
- ✓ First AI Implementation on NanoXplore NG-ULTRA FPGA

Contacts:

- Prof. Luca Fanucci, luca.fanucci@unipi.it
- Assistant Prof. Pietro Nannipieri, pietro.nannipieri@unipi.it
- Research Fellow Tommaso Pacini, tommaso.pacini@phd.unipi.it