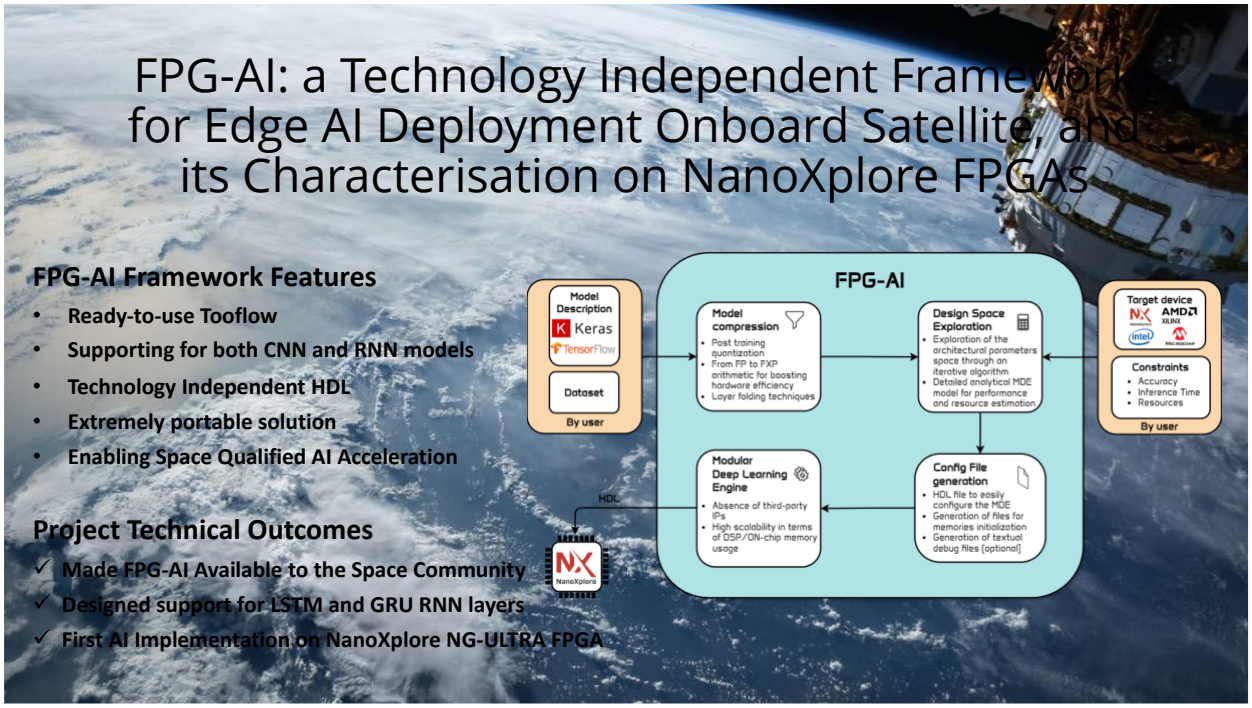

FPG-AI: a Technology Independent Framework for Edge AI Deployment Onboard Satellite, and its Characterisation on NanoXplore FPGAs

Executive summary

Early Technology Development

Open Discovery Idea Channel

Affiliation(s): Department of Information Engineering, University of Pisa, Italy

Activity summary:
The project aims to develop the first AI-to-FPGA toolflow supporting all state-of-the-art FPGAs, including NanoXplore devices. The objective is to facilitate AI deployment onboard satellites and demonstrate the applicability to NanoXplore technology, enhancing European sovereignty. We extended the FPG-AI design for compliance with NanoXplore technology, adding RNNs to the list of supported models and creating a hardware prototype. The results obtained during the benchmarking indicate successful AI model acceleration on rad-hardened FPGAs, reducing development time and cost and increasing the performances concerning state-of-the-art HLS approaches. In conclusion, we made significant steps forward toward enabling efficient, radiation-tolerant AI applications on spacecraft.

Publishing Date: 10/6/2024
Contract Number: 4000141108/23/NL/GLC/ov
Implemented as ESA EXPRESS PROCUREMENT

1 Introduction

1.1 Scope

This document represents the Executive Summary of the activity “FPG-AI: a Technology Independent Framework for Edge AI Deployment Onboard Satellite, and its Characterisation on NanoXplore FPGAs” initiated by the European Space Agency under Contract Number 4000141108.
The Department of Information Engineering of the University of Pisa has performed the work.

2 Executive Project Summary

FPG-AI is a technology-independent toolflow designed to automate the acceleration of deep neural networks (DNNs) on FPGAs. The tool receives as input a pre-trained DNN model together with its application dataset. FPG-AI first prepares the target network for the hardware acceleration, optimizes model topology, and shifts the arithmetic from floating-point to fixed-point. Subsequently, a Design Space Exploration (DSE) process selects the optimal parameters based on user-defined constraints for application metrics, resource consumption, and performance.
The Design Space Exploration (DSE) generates a configuration file to fine-tune the hardware architecture. The core of the accelerator in FPG-AI is the Modular Deep Learning Engine (MDE), a customizable Hardware Description Language (HDL)-based design that does not rely on third-party intellectual property (IP) and ensures compatibility with different Field Programmable Gate Arrays (FPGA) devices and vendors.
FPG-AI produces Hardware Description Language (HDL) files that describe the customized accelerator for a given model and device, allowing users to use the remaining FPGA resources for other tasks. In contrast to other solutions, FPG-AI provides the HDL sources instead of the final bitstream, enabling greater flexibility and customization.
The main objectives of the ongoing project are to reach TRL4 of FPG-AI (starting from TRL2) and make it available to the space community.

To achieve this goal, the project addressed several upgrades to the toolflow, including the extension to Recurrent Neural Networks (RNNs).
The designed post-training quantization algorithm for RNNs converts each floating-point value into its fixed-point version. To quantize a generic RNN model, the tool runs the network on the test set for multiple quantization and truncation settings.
The accelerator for RNN inference includes two key processing units: the Recurrent Processing Unit (RPU) and the Dense Processing Unit (DPU). The RPU handles Recurrent layers, while the DPU manages Dense layers. Each unit contains multiple specialized blocks designed to accelerate specific network layers arranged in a pipeline to simultaneously execute multiple layers during the regime phase. This layer-level customization ensures high resource efficiency, which is suitable for Recurrent networks due to their typically smaller number of layers and parameters than CNNs. Dedicated hardware in the Activation Block accelerates inter-layer activation functions within both units.
The Design Space Exploration (DSE) step in FPG-AI involves identifying an optimal design configuration for the accelerator based on user-provided constraints:

1. Device Selection: The tool can work with various FPGA devices from different vendors, accommodating diverse technological and resource characteristics. This is the only mandatory constraint.
2. Application Metric: Users can specify the best or lower bound application metric. The DSE prioritizes lower bit-width configurations that meet the metric. By default, it selects the lowest bit width that doesn't degrade the metric by more than 5%, or the best available metric if no configurations meet this condition.
3. Inference Time: The tool aims to achieve minimum inference time by adjusting the configuration to meet user-defined time constraints. If it can't meet the time constraint, it reports the issue.
4. Initial Design Space Point: Users can specify the initial point for analysis, allowing evaluation of specific cases.
5. Resource Usage: Users can limit DSP usage on the FPGA to leave room for additional IPs.

The DSE process starts by selecting a quantization setting from the Model Compression phase results, following the Application Metric criteria. It then initializes architectural parameters, either from user input or default settings. The DSE uses an analytical model to estimate DSP utilization and inference time, adjusting parameters iteratively to meet resource and time constraints.

Regarding Recurrent models, most state-of-the-art solutions report implementation results for LSTM-based networks trained on the IMDb dataset. Unfortunately, most existing frameworks have not yet been publicly released. This limitation has forced us to also consider hand-tuned accelerators that offer high performance at the cost of low configurability.
The performed comparison indicates that FPG-AI can be competitive with other state-of-the-art approaches. The timing efficiency is mainly due to the chosen Streamline architecture and the high level of parallelism in the architecture stage of the RNNs. The DSP efficiency aligns with similar work, specially optimized for Intel FPGAs.

For characterizing the tool for the NanoXplore devices (NG-MEDIUM, NG-LARGE, NG-ULTRA, NG-ULTRA300), we have chosen to implement two commonly used models from the literature, LeNet-5 for digit classification and Network in Network (NiN) for image categorization. The board selected for the first AI acceleration is the NG-ULTRA. Indeed, even if the selected networks are not complex (below 4 MB of memory parameter), the required logical resources may not be compatible with the reduced memory and DSP availability of the NG-MEDIUM FPGA. We configured the tool with minimum parallelism to facilitate the synthesizer and minimize hardware complexity, trading off the inference time.
To achieve compatibility with NanoXplore technology, the following updates were made to the MDE architecture of FPG-AI and the DSE script:

1. **Synthesis Attributes:** Added synthesis attributes to map on-chip memories to NanoXplore DPRAM and Register File blocks using the “NX_USE” directive.
2. **Synchronous Reset:** Added support for synchronous reset to use NX primitives that do not support asynchronous reset, such as NX DSPs.
3. **Architectural Changes:**
    a. **Removed Redundant Multiplier:** A redundant hardware multiplier within the activation cache system was removed to reduce the critical path logic depth by precomputing constant multiplications.

   b. **Control Signal Register:** Introduced a register to delay a control signal within the cache system for activations by one clock cycle, ensuring correct layer computation end detection without malfunctions.
4. **DSE Script Upgrade:** Updated DSE script to introduce the NG-ULTRA device, incorporating chip resource information like memory and DSP availability. Designed algorithms to predict the NX memory resource consumption and DSP utilization for each MDE architectural block.

All architectural changes were verified using bit-true simulations, comparing the accelerator's RTL outputs with the quantized model's outputs from FPG-AI Python scripts.
After verifying accelerator operations, implementation results were collected using the NX Impulse design suite.

We then developed the high-level architecture to deploy each accelerator provided by the FPG-AI toolflow on hardware. We tuned and adapted this system according to the model topology to design the hardware demonstrator for LeNet 5.
The system consists of the accelerator, equipped with a dual AXI 64-bit interface, the NG-ULTRA PS, which contains four ARM Cortex-R52, and the DDR3 memory, which stores images and weights to be processed.
Due to the lack of available example designs for NG-ULTRA memory-mapped interfaces, the system was tested and validated on the AMDXilinx ZCU106 board. This involved prototyping on the Xilinx ZU7EV FPGA, which features a quad-core Arm Cortex-A53 applications processor and a dual-core Cortex-R5 real-time processor, enabling communication with DDR4 memory and accelerator IP. The validation aims to prepare for future prototyping on the NG-ULTRA once support for the PS-PL interface and necessary communication procedures are confirmed.
The performance of LeNet-5 and NiN on the NG-ULTRA was compared with results from several FPGAs: Microsemi RTPF500T, Microsemi RTG4G150, Xilinx Kintex XQRKU060, and Xilinx Zynq 7000 XC7Z045. The Microsemi and Kintex FPGAs were chosen for their radiation resistance, while the Zynq 7000 was selected for its similar technology node (28 nm). The analysis found that resource consumption was similar across devices, with differences mainly in on-chip memory use due to various configurations. Frequency showed the greatest variability, impacting efficiency metrics. NanoXplore devices exhibit inferior performance compared to other technologies, indicating the need for enhanced design efforts to improve performance by reducing critical paths and logic depth.
We also compare the performance of FPG-AI on NanoXplore technology with Bambu, another solution for automatic accelerator generation on NanoXplore FPGAs. The results show different intents and functionalities of the two systems: FPG-AI focuses on the automatic acceleration of AI accelerators with customizable HDL code, prioritizing high-efficiency metrics, while Bambu is a general-purpose framework prioritizing broad support for heterogeneous applications, not limited to AI, and not optimizing efficiency metrics.

3 Technological Achievements

- Extension of the framework to Recurrent Neural Networks (RNNs) algorithms.
- Extension of the framework to European Nanoxplore innovative devices.

- Upgraded FPG-AI MDE hardware structure to ensure improved performance with NanoXplore products.
- The first implementation of AI accelerators on NanoXplore NG-ULTRA FPGA with FPG-AI framework.
- Evaluated the tool capability with a prototype hardware demonstrator.

4 Dissemination

The project team actively engaged with the broader scientific and industrial community, presenting FPG-AI at several prominent events:

- **NX Brave Days 2023**: This event provided a platform to showcase our advancements to the NanoXplore FPGA user community.
- **EDHPC 23 Conference**: Our participation here allowed us to share our technical findings and engage with peers working on onboard data handling and processing in satellites.
- **ASI Workshop on Enabling Technologies for Space**: We presented our innovative AI-to-FPGA toolflow, highlighting its potential applications in space technology.
- **ASI Workshop on Bridging Knowledge on Artificial Intelligence**: This workshop facilitated discussions on the intersection of AI and FPGA technologies, promoting knowledge exchange and collaboration.
- **Morpheus 2024**: ESA Workshop on Edge AI and Neuromorphic Hardware Accelerators: At this workshop, we discussed the future of edge AI solutions in space, positioning our project within the context of cutting-edge space technologies.
- **TEC-ED Final Presentation Days**: We presented in detail the development made within the project and the results obtained.

To strengthen our partnership with NanoXplore, we visited their headquarters, which was instrumental in consolidating our ongoing collaboration. This visit facilitated deeper technical exchanges and alignment on future development goals. In collaboration with NanoXplore, we are organizing a workshop in Pisa titled "Recent Advances in European Space FPGAs: Technologies and Applications," scheduled for June 11, 2024. This workshop aims to bring together experts from academia and industry to discuss the latest advancements in FPGA technologies and their applications in space, further cementing our project's impact and fostering future collaborations.

An important part of the dissemination involved writing academic articles. In detail, we presented the outcomes of the FPG-AI project at the EDHPC 2023 conference, with the following contribution, focusing on the extension to RNNs:

> *T. Pacini, E. Rapuano, L. Tuttobene, P. Nannipieri, L. Fanucci and S. Moranti, "Towards the Extension of FPG-AI Toolflow to RNN Deployment on FPGAs for On-board Satellite Applications," 2023 European Data Handling & Data Processing Conference (EDHPC), Juan Les Pins, France, 2023, pp. 1-5, doi: 10.23919/EDHPC59100.2023.10396607.*

We are currently working on an extension of that conference paper to be published on an international journal. At the same time, we are working on a journal article about the extension of FPG-AI on the NanoXplore devices. As we mentioned previously, we are waiting for new releases of the toolchain to enable effective PS-PL communication and, therefore, finalize the test of our prototype. At that point, we will conclude the article submission process.