Efficiency and Performance Tradeoffs in FPGA-based Embedded Computer Vision Applications

UNIVERSITÀ DI TORINO

Qaisar Farooq
Researcher PHD Student

Motivation

The increasing complexity and energy demands of modern AI models, such as Vision Transformers (ViTs), pose challenges for their deployment in resource-constrained and real-time environments. This research investigates field-programmable gate arrays (FPGAs) as an efficient hardware platform for AI acceleration. By minimizing and adapting these models for FPGAs, we aim to:

- **Reduce Energy Footprint**
- **Optimize Model Size**
- **Enhance Efficiency**

Goals

- To minimize and synthesize modern AI models, such as Vision Transformers (ViTs), for small-scale scenarios.
- To analyze the trade-offs, energy consumption, and performance of these models when deployed on FPGAs.
- To evaluate the **feasibility** of using FPGAs as an alternative hardware platform for deploying transformer models in real-time classification tasks.

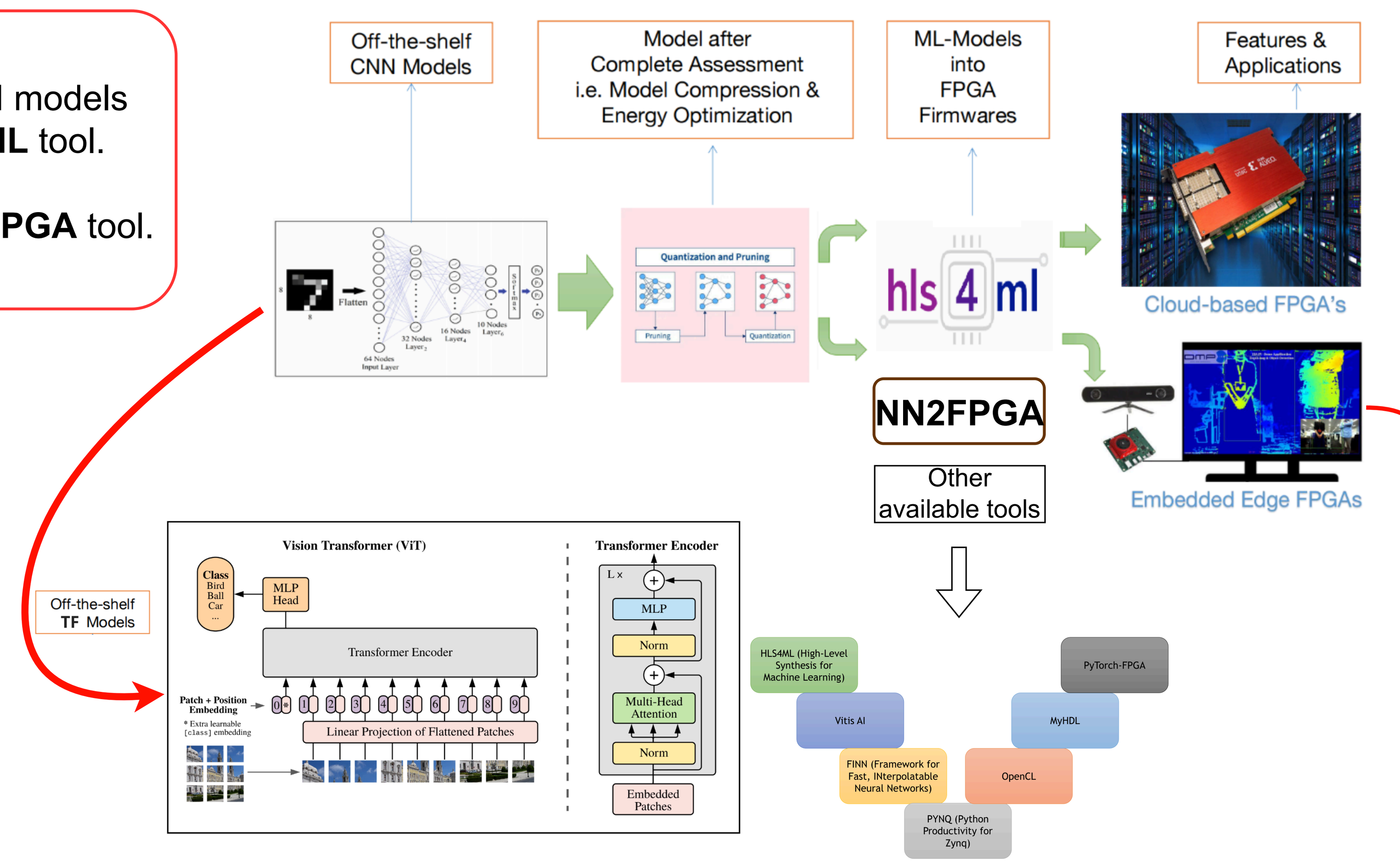
Methodology

First attempt: **CNN** models using the **HLS4ML** tool.

Next: **ViTs** on **NN2FPGA** tool.

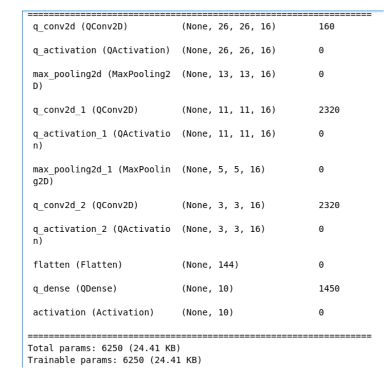
Key Challenges:

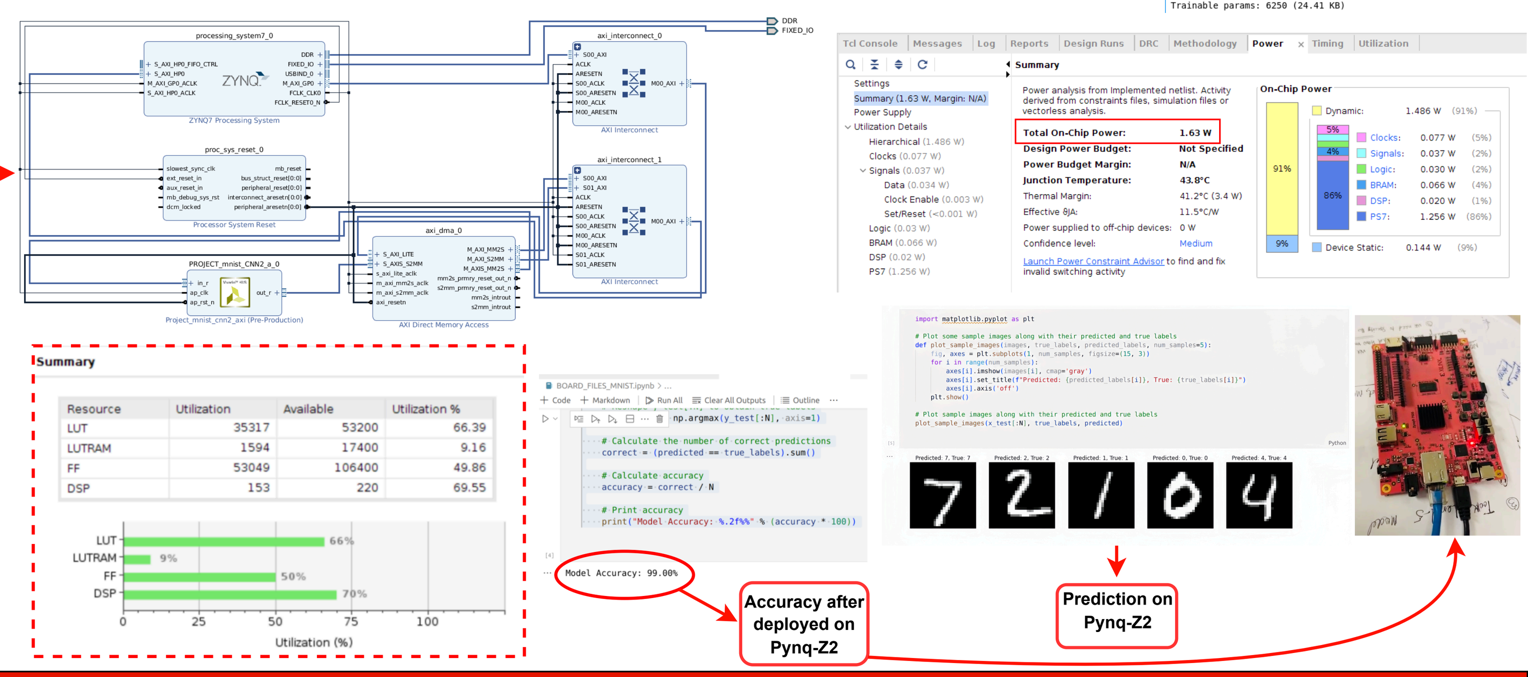
- Replacing **LN** with **BN**.
- Computation strategies for **Softmax** and **GELU**.
- ...



Initial Results

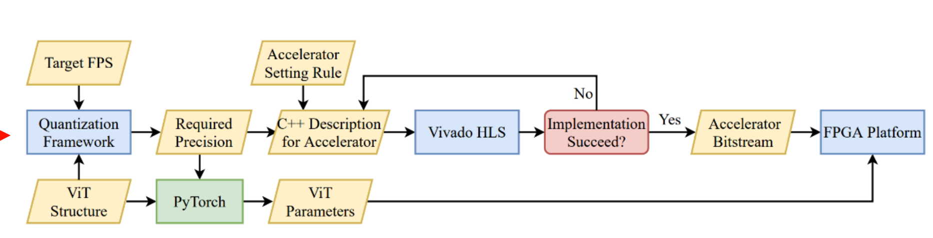
Initial results demonstrate successful deployment of a high-accuracy CNN on a Pynq-Z2 FPGA for MNIST digit classification using HLS4ML. Optimization techniques achieve 99% accuracy with reduced resource usage (LUT 66.39%, LUTRAM 9.16%, FF 49.86%, DSP 69.55%) and power consumption, highlighting the potential for efficient deep learning on edge devices.





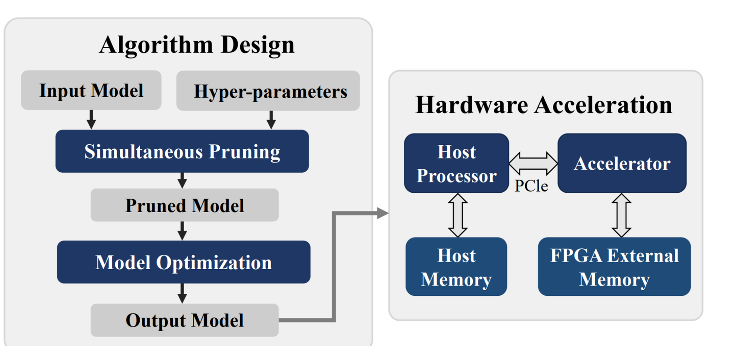
Related Work

VAQF is a framework that automatically builds efficient, real-time Vision Transformer accelerators on FPGAs by optimizing quantization and hardware parameters.



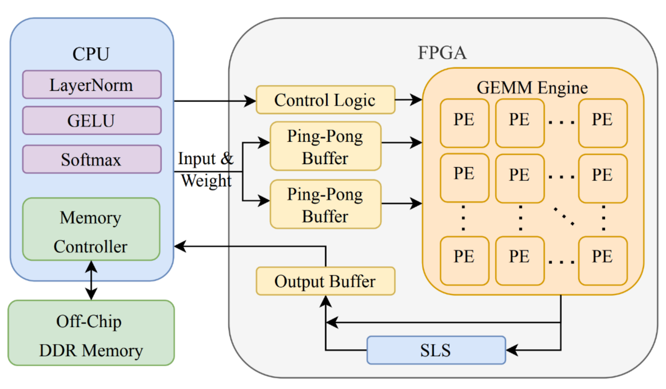
Overall Flow of VAQF

A novel algorithm-hardware codesign combines both static weight and dynamic token pruning for efficient Vision Transformer execution on a new accelerator.

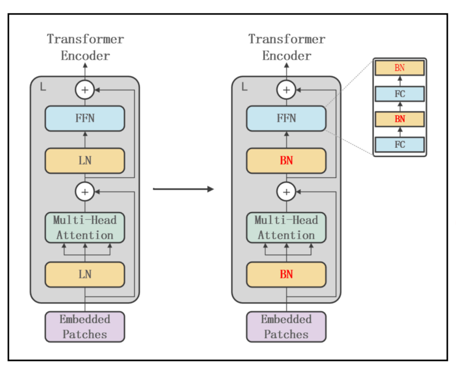


The proposed algorithm-hardware codesign by Dhruv Parikh et.al.

Quasar-ViT is a framework that designs efficient and accurate Vision Transformers for edge devices through hardware-aware quantization and architecture search, achieving high inference speed on FPGAs.

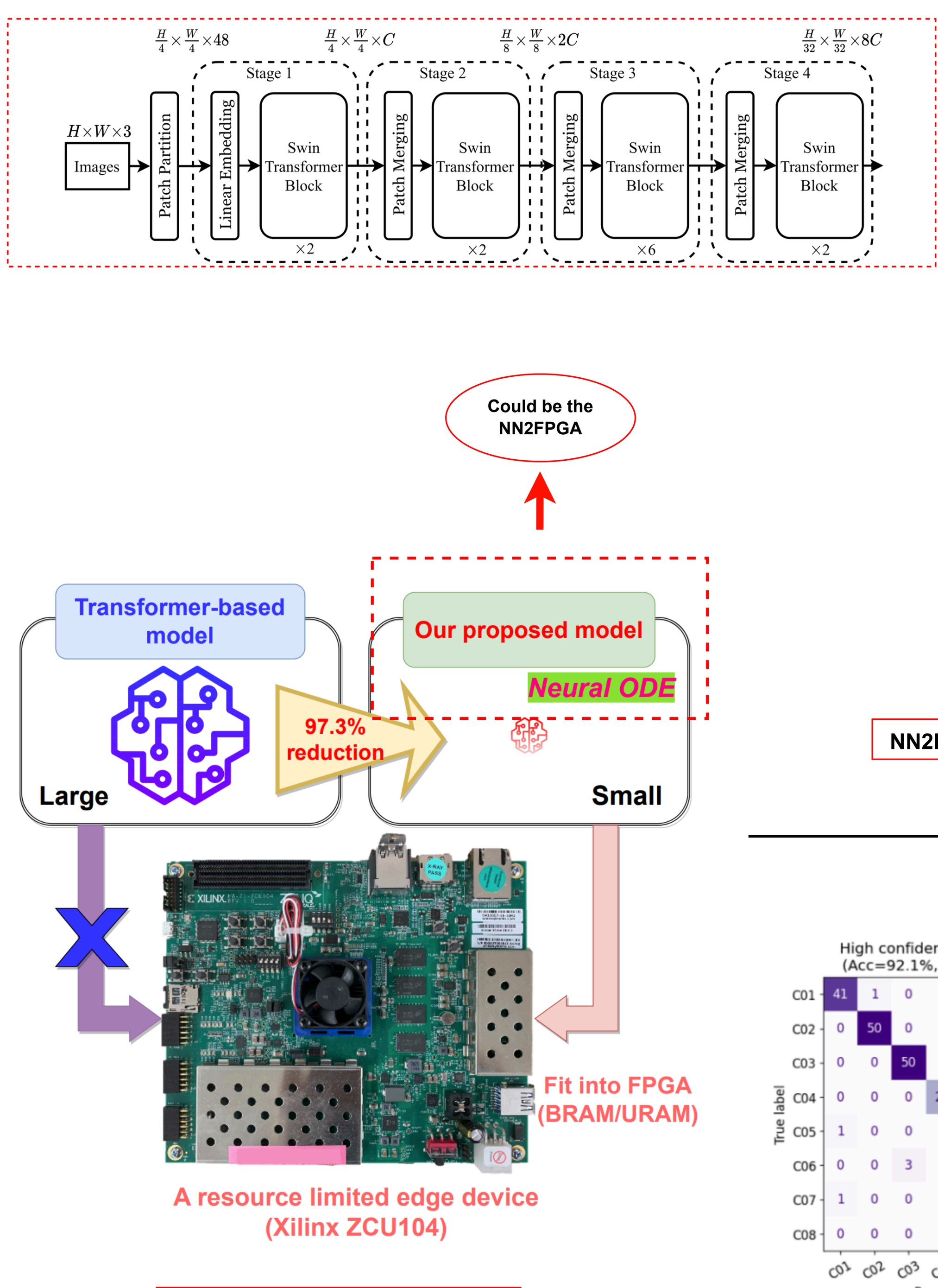


Quasar-ViT hardware architecture

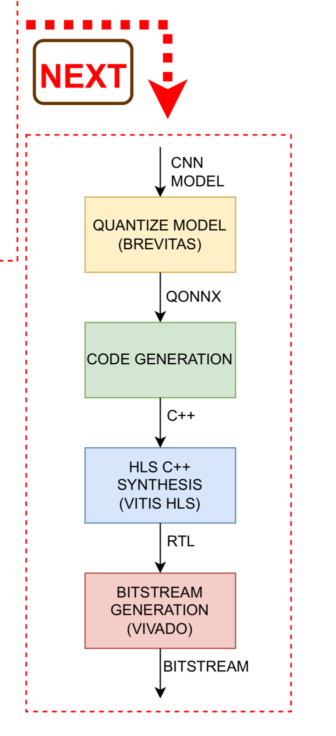


Layer Normalization (LN) was replaced with Batch Normalization (BN) to enable fusion with linear layers, improving inference efficiency on the FPGA with minor accuracy loss.

Current Work



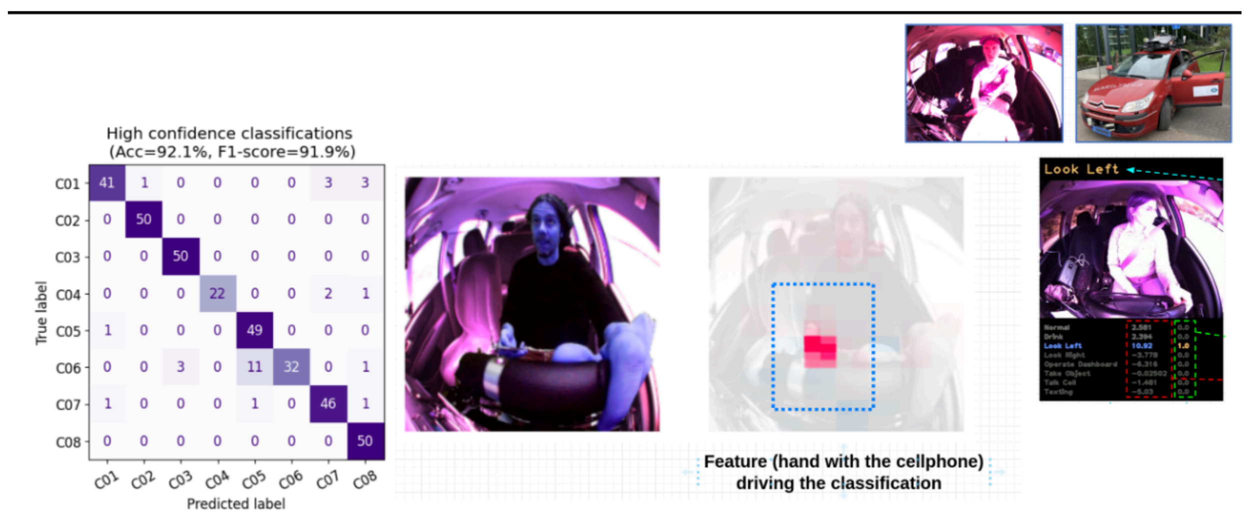
Neural ODE Architecture



NN2FPGA Implementation Flow

- **Dataflow architecture**
- **Fixed-point quantization**
- **Compatible with AMD-Xilinx boards**
- **High throughput/low power tasks**
- **Optimized design for skip connections**

Possible application: comparison with GPU-based alternative



Swin Transformers > 90% (83% best results with CNNs)

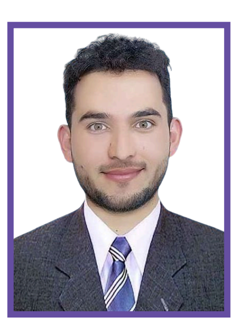
Results are better "explainable" The classifier focuses on "right features"

Conclusions & Future Work

- Our initial experiments using the HLS4ML framework on the Pynq-Z2 board achieved promising results, demonstrating the feasibility of deploying complex neural networks on FPGAs.
- After successfully deploying ResNet models using NN2FPGA on Kria KV-260 and Ultra96-v2 boards, we are now exploring its compatibility with ViT models and identifying any unsupported parameters.
- **Test Transformer Models:** Implement Swin TF model using NN2FPGA, focusing on maintaining accuracy.
- **Select FPGA Platform:** Choose the best FPGA for deployment, comparing cloud and edge options.
- **Compare GPU and FPGA:** Evaluate performance and energy use for models deployed on GPU and FPGA.

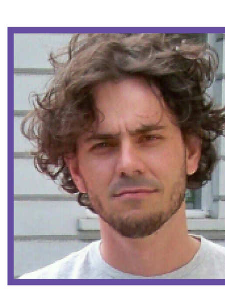
References

1. Fahim et al. (2021). hls4ml: An open-source codesign workflow to empower scientific low-power machine learning devices, https://arxiv.org/abs/2103.05579
2. Dosovitskiy (2020). An image is worth 16x16 words, https://arxiv.org/abs/2010.11929.
3. Casu et al. (202X). Machine Learning Inference Acceleration Using Embedded and Datacenter-Class FPGAs, https://tinyurl.com/DET-CAS-4-ML.
4. Liu et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. https://arxiv.org/abs/2103.14030.
5. Minnella et al. (2023). Design and Optimization of Residual Neural Network Accelerators for Low-Power FPGAs Using High-Level Synthesis, https://arxiv.org/abs/2309.15631.
6. Sun et al. (2022). VAQF: Fully automatic software-hardware co-design framework for low-bit vision transformer, https://arxiv.org/abs/2201.06618
7. Parikh et al. (2024). Accelerating ViT Inference on FPGA through Static and Dynamic Pruning, https://arxiv.org/abs/2403.14047
8. Li et al. (2024). Quasar-ViT: Hardware-Oriented Quantization-Aware Architecture Search for Vision Transformers, Proceedings of the 38th ACM International Conference on Supercomputing, https://dl.acm.org/doi/abs/10.1145/3650200.3656622
9. Liu et al. (2023). An Efficient FPGA-Based Accelerator for Swin Transformer, https://arxiv.org/abs/2308.13922
10. Okubo et al. (2023). A Lightweight Transformer Model using Neural ODE for FPGAs, 2023 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). https://ieeexplore.ieee.org/abstract/document/10196666



Qaisar Farooq
Ph.D. Candidate
- Machine Learning & AI
- Embedded Systems
- FPGA

qaisar.farooq@unito.it



Idilio Drago
Associate Professor
- Italian Consortium Coordinator
- Cyber Security
- Machine Learning & AI

idilio.drago@unito.it

UniTO in DistriMuSe Chips JU

EU Project