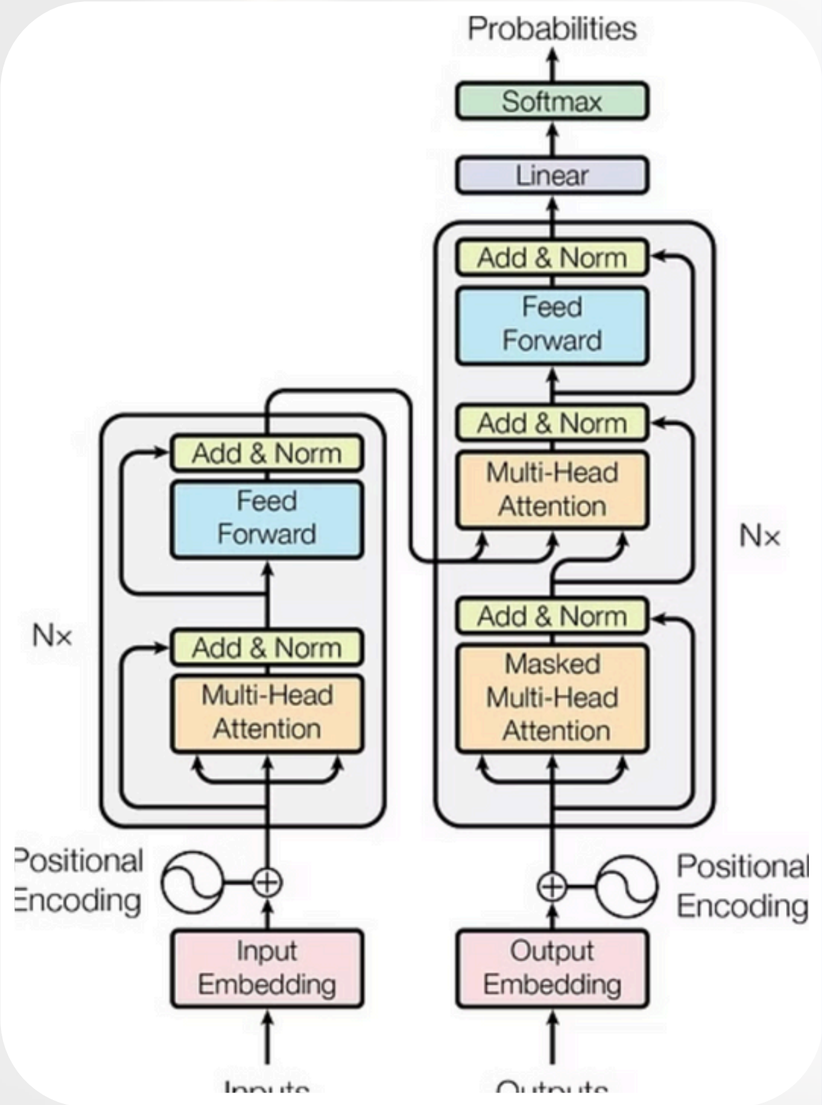
LARGE LANGUAGE MODELS ON

FPGA

This project explores deploying a Large Language Model (LLM) onto an FPGA for reconfigurability and reduced energy consumption.



LARGE LANGUAGE MODEL

- A type of artificial intelligence (AI) model designed to understand and generate human-like text
- its archtecture is Transformer Architecture

LOCAL LARGE LANGUAGE MODEL

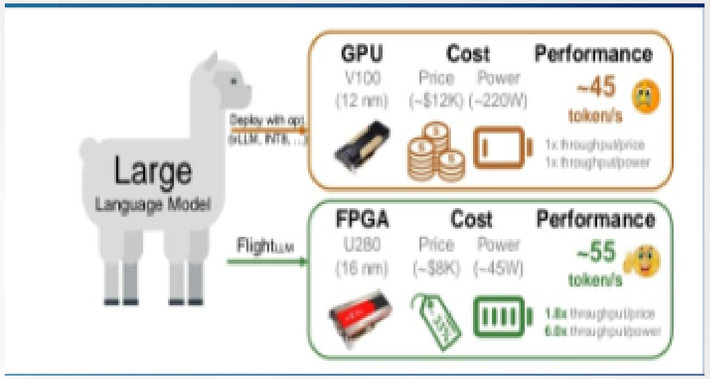
- Local LLMs can be a game changer by eliminating the need dependency on GPUs, you can unlock the full potential of LLMs for your application development needs.

Tiny Bert Model

- TinyBERT is a smaller, efficient version of BERT optimized for NLP tasks in resource-limited settings. It has around 14.5 million prameters , providing a good balance between performance and efficiency.

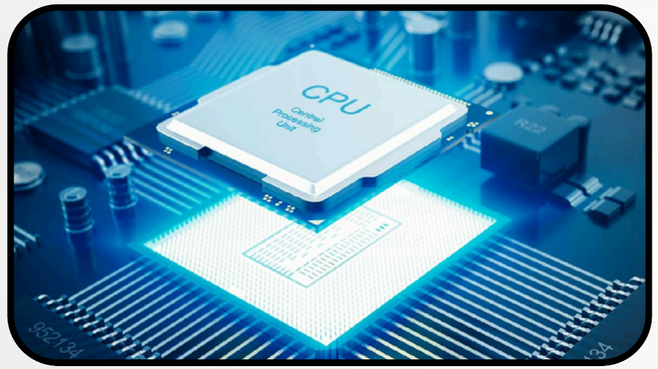
COMPARISON

This analysis compares the performance of different hardware platforms for running the TinyBert model



CPU

Standard processors provide a general-purpose platform for LLM execution, but performance can be limited.

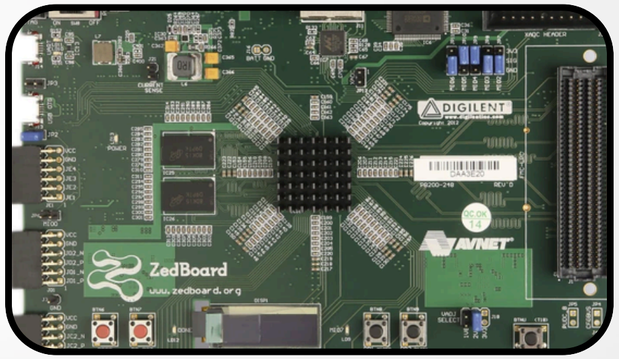


GPU

Graphics processing units are optimized for parallel computation, offering significant speedups for LLM inference.

FPGA

Field-Programmable Gate Arrays offer highly customizable hardware, enabling specialized acceleration for LLMs.



Generating Vivado XSA File

The Vivado Design Suite is used to create a custom hardware design for the TinyBert model.

Hardware Design

Specify the desired FPGA architecture and configure the memory interfaces.

1

Model Implementation

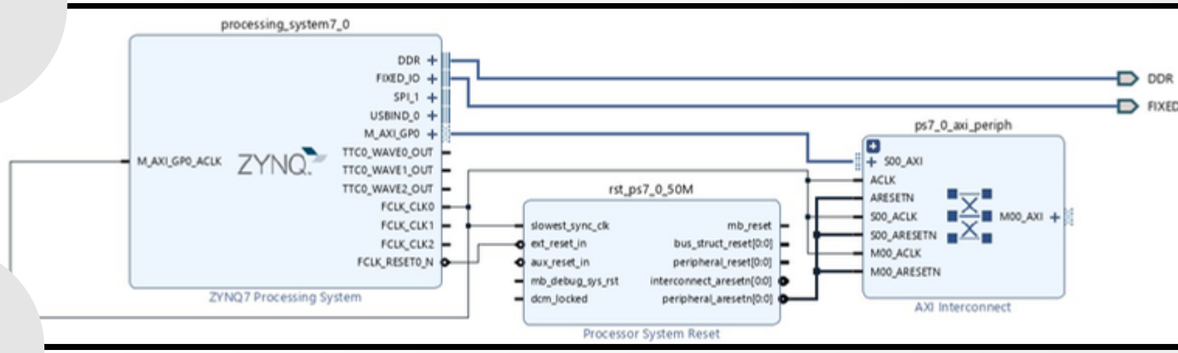
Map the TinyBert model's layers and operations onto the FPGA hardware

2

XSA Generation

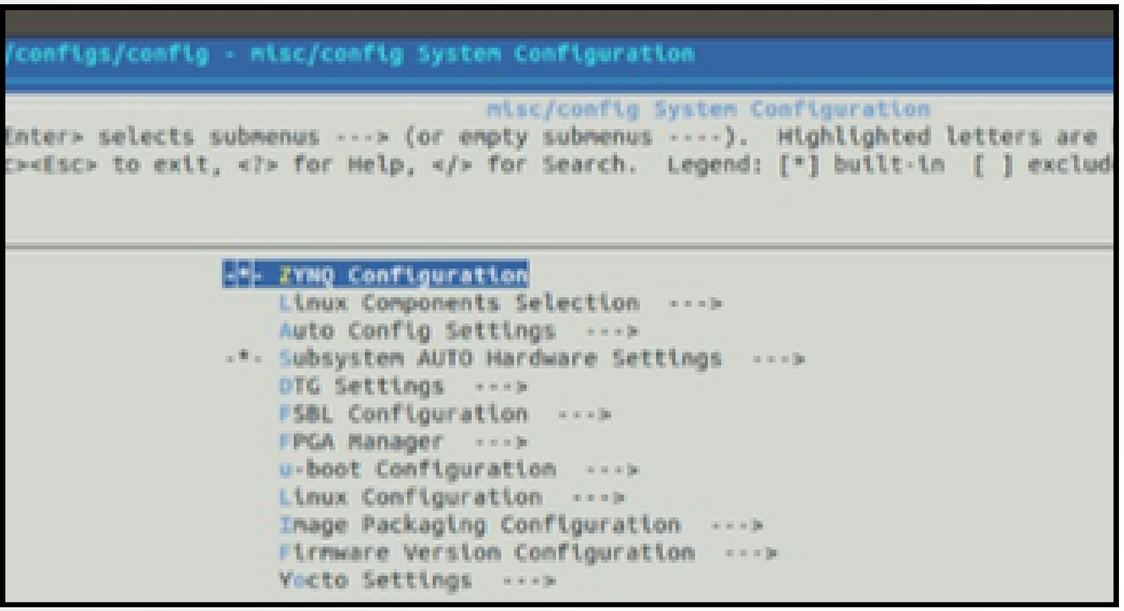
Create an XSA file, which encapsulates the hardware design and model configuration.

3



PETALINUX CONFIGURATION WITH BSP’S

PetaLinux, a Linux distribution optimized for Xilinx devices, provides a software framework for the FPGA deployment.



BOARD SUPPORT PACKAGES

- Select the appropriate BSP for the Zed Board to ensure compatibility and optimized performance.

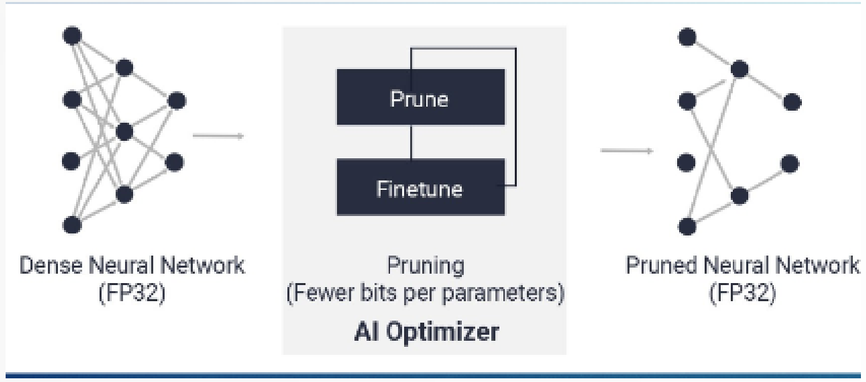
KERNEL CONFIGURATION

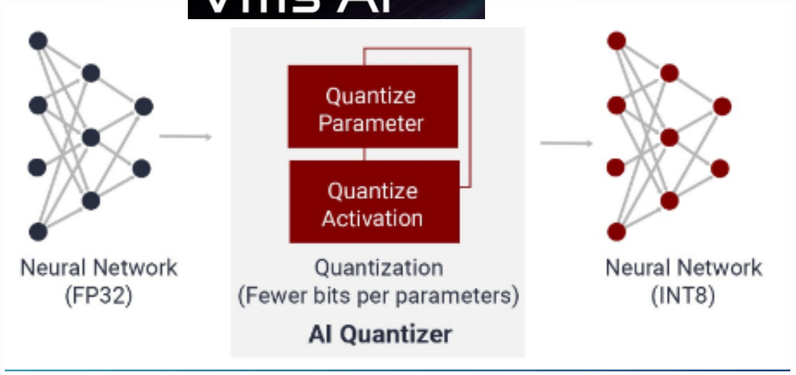
- Configure the Linux kernel to support the hardware acceleration features of the FPGA.

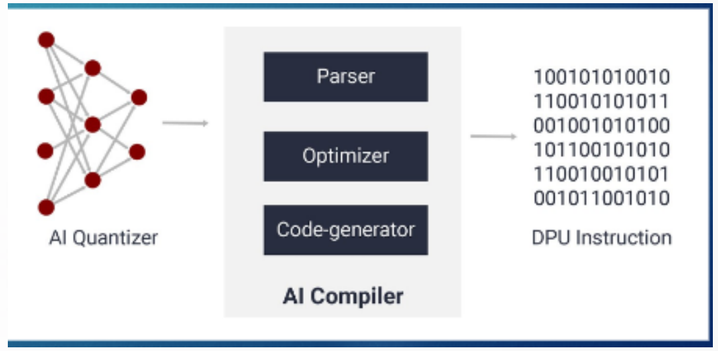
VITIS AI INTEGRATION FOR HARDWARE DEPLOYMENT

Vitis AI provides tools and libraries for accelerating AI inference on Xilinx devices.

AMD Vitis AI







Model Quantization

Convert the TinyBert model to a quantized format compatible with the FPGA hardware.

Hardware Acceleration

Map the quantized model onto the FPGA using Vitis AI tools for optimal performance.

Software Integration

Develop software interfaces to interact with the accelerated model on the FPGA.

BLOCK DIAGRAM:

FUTURE SCOPE:

We want to implement a Large language Models like BARD or BERT or LLaMA on FPGA , not only Text to Text generation it may extend to Text to image generation and animation generation

THANK YOU

Presented by:
Y.Sai Tharun Reddy
P.Supriya
K.Lohitha Reddy
U.Sathwik