[Public]

# Accelerating Large Model Inference on FPGA

Si Qiang,Su Xuyang,Tan Liwei

Nan Jing University

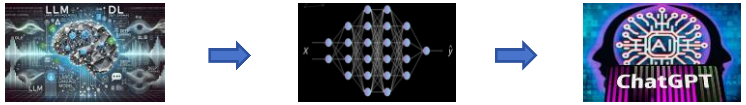
OpenHW2024

AMD

*On board test by AMD XCVU 440*

## INTRODUCTION

Recent advances in large language models with billions of parameters have created a huge demand for efficient deployment of inference workloads. This work uses FPGA **reconfigurable** and **low-power** characteristics to design customized **deep learning operators** on FPGA, and splits multi-operator computing **units** into a large language model in the terminal design script to realize the acceleration of large-model reasoning.

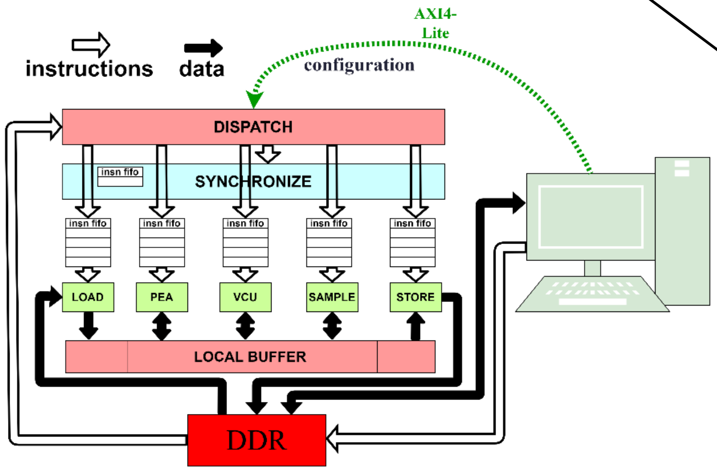


### Typical application case

Compared with GPU and ASIC acceleration, FPGA-based AI accelerators show significant progress in performance optimization, flexibility and energy efficiency improvement, and are widely used in deep neural network (DNN) reasoning. The reconfigurability of FPGAs enables them to optimize the hardware configuration for different tasks and adapt to complex AI inference tasks. **Key technical issues addressed:**

1. we take the one-layer encoder structure in **Bert model** as the model basis, use FPGA to design the convolutional operator, multi-head attention mechanism, addition operator, normalization and other **personalized internal operators**.
2. The **top-level architecture** model is written for the interaction between the host, ddr, and npu.
3. The inference acceleration of Bert is realized by using the **high speed** and low power of FPGA.

## CREATIVE DESING



### System Architecture

Decoupled computing ideas:

1. The calculation process of neural network reasoning is **divided** into data reading and writing, tensor calculation and vector calculation;
2. The hardware can be **modified** quickly to adapt to emerging neural network structures.

Macro instruction set:

1. Based on the decoupled architecture, a macro instruction set supporting general large model acceleration is proposed.;
2. It can realize **flexible loading** of weights and activation data, convolution of different parameter configurations and matrix multiplication calculation;
3.It can realize **a variety of** activation function calculation, intermediate data cache write back and other functions.

## RESULT

This work achieves the FPGA hardware design and accomplishes the software development of key operators. It possesses strong computing power and high on-chip resource utilization. Eventually, the **Bert model** is fully constructed in the form of a macro instruction set, and the large model accelerated reasoning is accomplished.

For FPGA deployment, a **hierarchical** neural network inference compilation process is proposed, which is divided into **network data parsing layer**, o**perator execution flow planning layer**, and **run-time binary file generation layer** for operator level implementation.

Next, we will attempt to **train** the bert model after FPGA acceleration, verify the acceleration effect of FPGA on the training of large models, acquire the **key parameters** that are more suitable for the accelerated model, and ultimately obtain a more satisfactory outcome of the FPGA-accelerated **text question and answer**. We anticipate that in the future, FPGAs will make further progress in the field of AI and replace GPU acceleration in more areas, saving us more energy and time.