GPU Accelerated AI-RAN System: Transformer Based Estimation with FPGA Implemented OFDM Transceiver

Su Hyun Kim, Kae Won Choi*

Sungkyunkwan Univ., * Sungkyunkwan Univ.

rkawns5909@skku.edu, *kaewonchoi@skku.edu

GPU 가속 AI-RAN 시스템: FPGA 구현 OFDM 송수신기 및 트랜스포머 기반 추정

김수현, 최계원*

성균관대학교, *성균관대학교

Abstract

In this paper, we present a GPU accelerated AI-RAN system combined Transformer based channel and phase estimation with an FPGA implemented OFDM transceiver, high-throughput, and low-latency wireless communication.

Ⅰ. Introduction

The growing importance of AI systems in mmWave communications has underscored the need for GPU based high performance processing and practical system implementation [1]. In this paper, we propose a GPU based AI-RAN system that incorporates Transformer based training for channel estimation and phase noise (PN) compensation, implements ultra-high-speed communication on a field programmable gate array (FPGA), and analyzes GPU based channel estimation and simulation results. Finally, we discuss plans for developing a 28 GHz prototype transceiver testbed under real channel conditions.

Ⅱ. Method

A. System Model

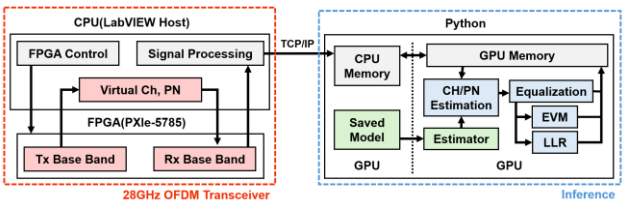


Fig 1. System Structure

The OFDM transceiver comprises a PXIe-5785 FPGA module. OFDM frames are transmitted from the FPGA to a LabVIEW host, propagated through a virtually generated channel and PN environment, and then received by the FPGA. The LabVIEW host program extracts the Rx frames and forwards them to a Python-based processing module over Transmission Control Protocol/Internet Protocol (TCP/IP). In the inference module, the saved Transformer-based estimator model is loaded, and high-speed estimation, equalization, log-likelihood ratio (LLR) computation, and error-vector magnitude (EVM) calculation are performed on the GPU.

B. Transformer Training

The Transformer is trained on NYUSIM channel data from five scenarios (UMi, UMa, RMa, InH, InF), each under both LOS and NLOS conditions. OFDM symbols include Demodulation Reference Signals (DM-RS) in the first symbol of each slot for channel estimation and Phase Tracking Reference Signals (PT-RS) at every 48n + 6th subcarrier for PN estimation. The Transformer uses these signals as inputs and outputs estimated channel and PN components along with attention weights.

C. OFDM Transceiver

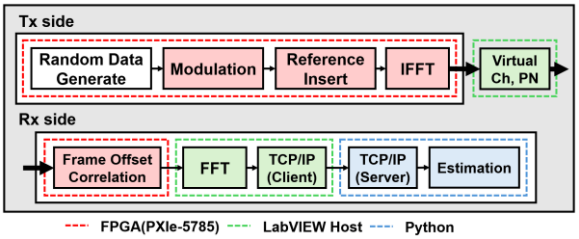


Fig 2. Structure of OFDM Transceiver

In the OFDM transceiver subsystem, the FPGA generates OFDM frames and receives incoming signals. The LabVIEW host performs virtual channel modeling, Fast Fourier Transform (FFT) processing, and TCP/IP communication with a Python server. The Python server then relays the received data and performs signal estimation and compensation using the saved Transformer model.

D. GPU Based Channel and PN Inference

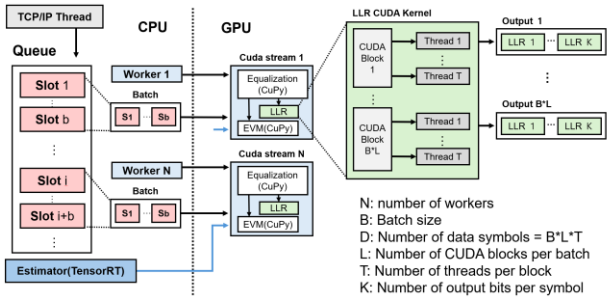
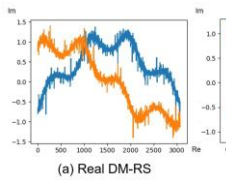


Fig 3. Inference Structure

Prior to inference, the pretrained Transformer model is optimized as a TensorRT engine for GPU execution. Signals received via TCP/IP are queued and batched by workers, then transferred to GPU. Each batch is assigned to a separate CUDA stream, where channel state and PN estimation occur, followed sequentially by CuPy-based equalization and EVM calculation. Finally, LLR computation is performed using a dedicated CUDA kernel invoked once per stream, internally generating multiple CUDA blocks and threads for parallel processing.

Ⅲ. Simulation Result



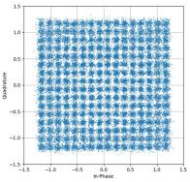
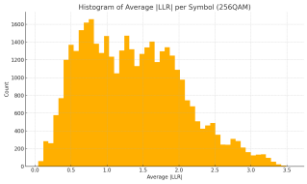


Fig 4. DM-RS VS Estimated CH Fig 5. Constellation

As shown in Fig. 4, the real and imaginary components of DM-RS after propagation through the virtual channel closely trace the original channel response when reconstructed by the estimator, demonstrating high-fidelity recovery. Fig. 5 shows the 256 QAM constellation diagram, where symbol clusters are tightly separated with negligible overlap.



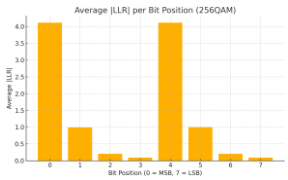


Fig 6. LLR Histogram Fig 7. LLR Value

As shown in Fig. 6 and 7, the 256 QAM LLR histogram and per-bit average |LLR| values demonstrate excellent demodulation performance.

BER	EVM(dB)	SNR(dB)
5.8×10^{-3}	-19.78	22.14

Table 1. System Performance

As shown in Table 1, the average BER is 5.8×10^{-3}, EVM is -19.78dB, and SNR is 22.14 dB, confirming excellent reception performance.

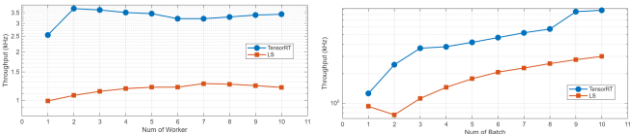


Fig 8. Throughput of GPU worker and batch

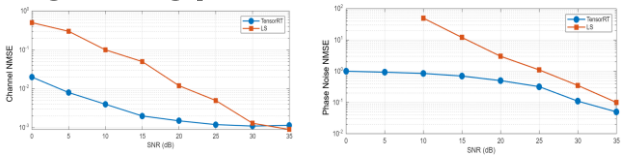


Fig 9. NMSE of Channel and PN

Fig. 8 shows that throughput increases with more GPU workers and larger batch sizes, where one varies while the other is fixed to 3. Further gains are observed with TensorRT acceleration. Fig. 9 shows normalized mean square error (NMSE) versus SNR for channel and PN estimation.

All graphs consistently show that performance improves due to GPU techniques involving increased workers, larger batch sizes, and TensorRT acceleration.

Ⅳ. Conclusion

This paper presents a GPU accelerated AI-RAN mmWave system that combines a Transformer based channel and phase noise estimator with an FPGA OFDM transceiver. The pretrained model, optimized as a TensorRT engine and executed across multiple CUDA streams, enables high throughput channel and phase noise estimation. Simulation results demonstrate a BER of 5.8×10^{-3}, an EVM of - 19.78 dB, and a SNR of 22.14 dB, with robust NMSE performance. Future work includes developing a 28 GHz prototype transceiver, extending the system optimizing hardware and model architectures for 6G RAN and IoT applications.

ACKNOWLEDGMENT

This work was supported by the BK21 FOUR Project.

REFERENCES

[1] N. A. Khan and S. Schmid, "AI-RAN in 6G Networks: State-of-the-Art and Challenges," in IEEE Open Journal of the Communications Society, vol. 5, pp. 294-311, 2024