Published: 27 February 2025

Citation in BibTeX format

FPGA '25: The 2025 ACM/SIGDA International Symposium on Field Programmable Gate Arrays
February 27 - March 1, 2025
CA, Monterey, USA

Conference Sponsors:
SIGDA

POSTER

FMC-LLM: Enabling FPGAs for Efficient Batched Decoding of 70B+ LLMs with a Memory-Centric Streaming Architecture

WENHENG MA, Tsinghua University, Beijing, China

XINHAO YANG, Tsinghua University, Beijing, China

SHULIN ZENG, Tsinghua University, Beijing, China

TENGXUAN LIU, Tsinghua University, Beijing, China

LIBO SHEN, Chinese University of Hong Kong, Hong Kong, Hong Kong

HONGYI WANG, Tsinghua University, Beijing, China

View all

Open Access Support provided by:

Chinese University of Hong Kong

Tsinghua University

Infinigence AI

Shanghai Jiao Tong University

FPGA '25: Proceedings of the 2025 ACM/SIGDA International Symposium on Field Programmable Gate Arrays (February 2025)
https://doi.org/10.1145/3706628.3708863
ISBN: 9798400713965

FMC-LLM: Enabling FPGAs for Efficient Batched Decoding of 70B+ LLMs with a Memory-Centric Streaming Architecture

Wenheng Ma*
Xinhao Yang*
Shulin Zeng
Tengxuan Liu
Tsinghua University
Infinigence-AI
Beijing, China

Libo Shen
The Chinese University
of Hong Kong
Hong Kong, China

Hongyi Wang
Shiyao Li
Tsinghua University
Infinigence-AI
Beijing, China

Jiewen Wang
Yuhan Zhang
Hao Guo
Jintao Li
Infinigence-AI
Beijing, China

Ziming Zhang
Zhenhua Zhu
Xuefei Ning
Tsinghua University
Beijing, China

Tsung-Yi Ho
The Chinese University
of Hong Kong
Hong Kong, China

Guohao Dai†
Shanghai Jiao Tong University
Infinigence-AI
Shanghai, China

Yu Wang†
Tsinghua University
Beijing, China

Abstract

For large language model (LLM) acceleration, FPGAs face two challenges: insufficient peak computing performance and unacceptable accuracy loss of model compression. This paper proposes FMC-LLM to enable FPGAs for efficient batched decoding of 70B+ LLMs.

CCS Concepts

• **Hardware → Hardware accelerators**.

Keywords

LLM, multi-FPGA, memory-centric, streaming architecture

ACM Reference Format:
Wenheng Ma*, Xinhao Yang*, Shulin Zeng, Tengxuan Liu, Libo Shen, Hongyi Wang, Shiyao Li, Jiewen Wang, Yuhan Zhang, Hao Guo, Jintao Li, Ziming Zhang, Zhenhua Zhu, Xuefei Ning, Tsung-Yi Ho, Guohao Dai†, and Yu Wang†. 2025. FMC-LLM: Enabling FPGAs for Efficient Batched Decoding of 70B+ LLMs with a Memory-Centric Streaming Architecture. In *Proceedings of the 2025 ACM/SIGDA International Symposium on Field Programmable Gate Arrays (FPGA '25), February 27-March 1, 2025, Monterey, CA, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3706628.3708863

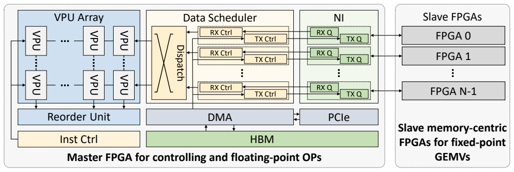

Figure 1: Architecture of FMC-LLM with N+1 FPGAs.

1 Introduction

Existing FPGA-based systems provide low-latency acceleration but lack sufficient performance for efficient LLM batched decoding. This paper introduces **FMC-LLM**, a distributed FPGA-based Memory-Centric streaming architecture. FMC-LLM achieves higher throughput and a lower total cost of ownership (TCO) compared to GPU systems for 70B+ LLM batched decoding.

2 Architecture and Experiments

FMC-LLM employs FPGAs in a master-slave architecture with a star topology, as shown in Figure 1. The Vector Processing Unit (VPU) on the master FPGA connects to N slave FPGAs via the Network Interface (NI). An instruction controller on the master FPGA schedules compute units and dispatches commands to the slave FPGAs. FMC-LLM adopts a non-blocking streaming data flow, eliminating control and backpressure logic to improve operating frequency and resource utilization. Using a V80 FPGA as the master and eight U55C FPGAs as slaves, FMC-LLM achieves a throughput of 563.34 tokens/s under a 50 ms time-per-output-token (TPOT) budget on Llama-3.1-70B-Instruct. Compared to two A100 and eight RTX 4090 GPU systems, FMC-LLM delivers 13.68× and 3.48× higher throughput, as well as 9.13× and 2.42× better cost efficiency, respectively.

[*] Both authors contributed equally to this research.
[†] Corresponding authors: yu-wang@tsinghua.edu.cn, daiguohao@sjtu.edu.cn

FPGA '25, Monterey, CA, USA

ACM ISBN 979-8-4007-1396-5/25/02
https://doi.org/10.1145/3706628.3708863