

LOgIQ: A Novel Multiplier-free Systolic Array Hardware Accelerator for Large Language AI Models

07/28/2026



LOgIQ: A Novel Multiplier-free Systolic Array Hardware Accelerator for Large Language AI Models

Electrical and Computer Engineering

Abstract

Large language models (LLMs) have rapidly advanced state-of-the-art performance across domains, including question answering, summarization, code generation, and dialogue. As these models scale into billions of parameters and longer context windows, deploying them efficiently becomes increasingly difficult, especially in constrained environments like edge accelerators or low-power inference chips. To reduce the memory and compute overhead of LLM inference, quantization has emerged as a widely adopted strategy in both software and hardware optimization pipelines. Most quantization techniques neglect the hardware-level effects, which leads to multiplier-heavy calculations and comparisons that slow down processing time and require more space. This hardware accelerator is multiplier-free and uses significantly less space and energy for LLM inference calculations.

Problem

Edge calculation in small devices is growing in demand with heightened LLM ability. Calculations need to be made in great numbers quickly, cheaply, and use minimal physical space.

Solution

This method simplifies data (quantization) before calculation for faster and lighter results. After simplifying data, this method does not use multiplication (multiplier-free) in data processing for faster and easier calculations.

Value Proposition

Through higher process efficiency, this technology provides powerful calculation acceleration with low power costs, less physical space than existing alternatives, and less memory usage.

Benefit

Problem: Large Language Models (LLM) are growing rapidly. Inference quality, depth, and thinking capacity all become real concerns when hardware comes into play. Each calculation consumes time and energy, and the hardware performing these computations takes up valuable space, especially in edge-computing devices like laptops, satellites, and drones.

As LLM inference expands worldwide and reaches a larger population, this issue becomes even more significant. Hardware accelerators must operate on individual devices so that intensive computations aren't left to centralized servers that can't handle all the data or respond quickly to users. With limited power, cooling, and space, each hardware accelerator must function within tight constraints while still delivering satisfying performance.

Solution: This method simplifies data (quantization) before calculation for faster and lighter results. After simplifying data, this method does not use multiplication (multiplier-free) in data processing for faster and easier calculations.

Value Proposition: Through higher process efficiency, multiplier-free quantization offers powerful calculation

acceleration with low power costs, less physical space than existing alternatives in hardware accompaniment, and less memory usage.

Researchers

Sanghamitra Roy

Koushik Chakraborty

Tanzeel-Ur-Rehman Khan

USU Department: **Electrical and Computer Engineering**

Funding

This invention was made with government support awarded by the NSF. The government has certain rights in the invention.

USU Reference No. C26013

Intellectual Property

Pending U.S. Non-Provisional Patent Application filed December 2025.