



Towards Agentic AI Inference Acceleration on Heterogeneous CGRA Systems

Motivation

The rapid expansion of the LLM domain in the final years generated the need for finer-grained LLM agent management, leading to Agentic AI, where applications repeatedly invoke language models to plan actions, call tools, interpret observations, and revise decisions [1]. These inference requests involve diverse computational kernels whose requirements change with model architecture, context length, batch size, and numerical precision.

Within transformer inference, prompt processing typically exposes matrix–matrix computations, while low-batch token generation involves matrix–vector operations with different data-reuse patterns. Attention kernels vary with sequence length and KV-cache organization, while normalization, softmax, and activation functions require additional arithmetic and reduction operations (non-linear functions). Precision introduces further diversity: quantized integer kernels require scaling and accumulation support, while numerically sensitive operations may retain floating-point arithmetic.

This diversity motivates the need for hardware accelerators to be capable of great parallelism when fitted to the processing needs of each agent’s inference, also providing a need for relatively fast reconfiguration on the targeted devices. Coarse-Grained Reconfigurable Arrays can provide this level of performance/flexibility trade-off by providing customization options along with fast reconfiguration times, when supporting SIMD configurations.

Prior works have demonstrated Transformer-Based inference [2], as well as LLM kernel evaluation and profiling [3] on CGRA accelerators, without exposing proper

heterogeneous configuration exploration both in software (kernel mapping) and on the hardware side (architecture scaling, multi-CGRA interfacing), which is crucial for adaptation of this architectural paradigm to the needs of modern Agentic Inference.

Scope

CGRA Design

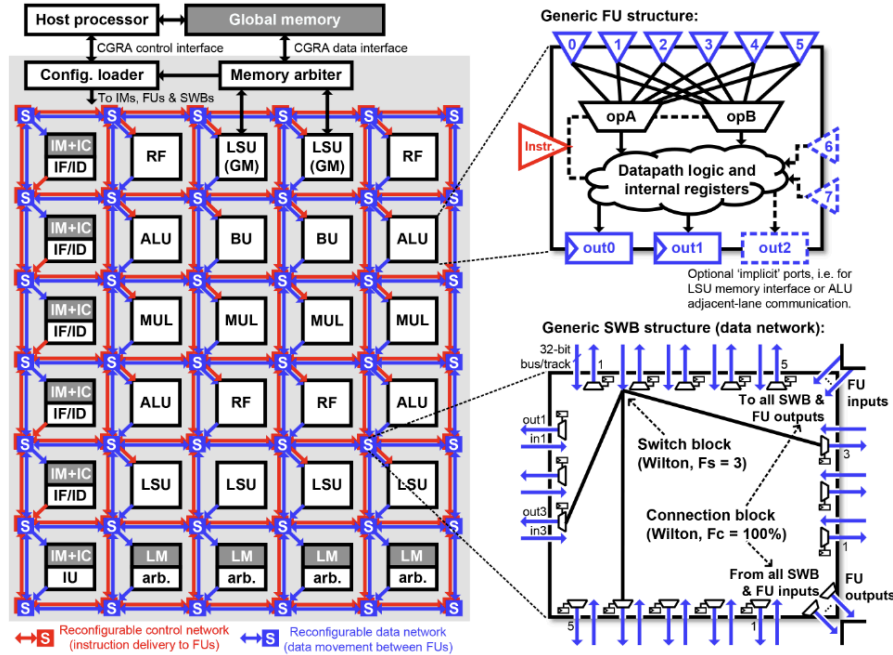


Figure 1: R-Blocks Architecture

As our target CGRA accelerator will use the R-Blocks CGRA [4], a state-of-the-art heterogeneous CGRA approach with each processing element providing a subset of the total ISA provided, transforming the classing CGRA paradigm into a “vectorized CPU” accelerator, with custom Instruction Decode units per computation unit (scalar mode) or N units (SIMD mode with N lines). Two switchbox networks connect all the CGRA elements together, one carrying data and one carrying instruction information. Memory-wise, R-Blocks provide Local Memories for faster access and better locality, and are also configured to be accessed in SIMD mode, as well as external access to external memory with the Load Store Unit (LSU). The Accumulation and Branching Unit (ABU) is responsible for controlling the execution of the program mapped on the CGRA accelerator, sharing a common program counter with all the Instruction Decode units, as

well as necessary stalling signals (e.g. when a load from the external memory is performed).

Thesis Targets

This thesis will investigate distributed CGRA integration on Alveo [5] acceleration platforms and mapping optimization for agentic AI inference. It will exploit R-Blocks heterogeneity, architectural customization, and runtime reconfiguration to accommodate variations in computational kernels, numerical precision, and parallelism across inference workloads.

On the hardware side, we will extend the existing R-Blocks FPGA emulation framework to integrate multiple cooperating CGRA instances into a distributed system on the Alveo FPGA platform. Instances will support combinations of processing element types, vector widths, and arithmetic precisions. The system will provide communication, synchronization, and configurable kernel execution across instances, exploiting FPGA resources to increase parallelism. We will also investigate integration with Alveo’s High-Bandwidth Memory (HBM) interface to efficiently store and move model weights, intermediate tensors, and attention key–value caches.

On the software side, a custom vLLM [6] backend will (ideally) expose the emulated CGRA system through its hardware platform-plugin interface. A worker and model runner will translate scheduled inference batches into accelerator commands, with attention and cache management adapted to FPGA memory. The execution runtime will select compatible kernel mappings and CGRA instances by workload dimensions, precision requirements, and data locality. Extensions will coordinate multiple Alveo FPGA systems with different hardware configurations, enabling distributed inference across heterogeneous CGRAs while accounting for communication overhead.

In conclusion, this thesis will provide both a scaled multi-CGRA system capable of accelerating demanding and heterogeneous Agentic AI workloads, accompanied by a mapping framework that can provide optimized execution, utilizing resource-aware profiling (PE type, precisions supported, memory availability), providing a complete software-hardware co-designed framework.

RELATED MATERIAL:

- [1] Ritik Raj and Souvik Kundu and Ishita Vohra and Hong Wang and Tushar Krishna, "Towards Understanding, Analyzing, and Optimizing Agentic AI Execution: A CPU-Centric Perspective", <https://arxiv.org/abs/2511.00739>
- [2] Y. Luo et al., "ML-CGRA: An Integrated Compilation Framework to Enable Efficient Machine Learning Acceleration on CGRAs," 2023 60th ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 2023, pp. 1-6, access: <https://ieeexplore.ieee.org/document/10247873>
- [3] Dionysios Kefallinos, Georgios Alexandris, Alexis Maras, Panagiotis Chaidos, Manil Dev Gomony, Henk Corporaal, Dimitrios Soudris, and Sotirios Xydis. Performance Modeling & Mapping of LLM Inference on Heterogeneous Vectorized CGRAs. In PARMA-DITAM 2026. <https://doi.org/10.4230/OASlcs.PARMA-DITAM.2026.8>
- [4] Barry de Bruin, Kanishkan Vadivel, Mark Wijtvliet, Pekka Jääskeläinen, and Henk Corporaal. 2024. R-Blocks: an Energy-Efficient, Flexible, and Programmable CGRA. ACM Trans. Reconfigurable Technol. Syst. 17, 2, Article 34 (June 2024), 34 pages. <https://doi.org/10.1145/3656642>
- [5] <https://www.amd.com/en/products/accelerators/alveo/u50/a-u50-p00g-pq-g.html>
- [6] <https://vllm.ai/>

CONTACT INFORMATION:

- Georgios Alexandris, Ph.D Student, NTUA: (galexandris@microlab.ntua.gr)
- Panagiotis Chaidos, Ph.D Student, NTUA: (pchaidos@microlab.ntua.gr)
- Alexis Maras, Ph.D Student, NTUA: (amaras@microlab.ntua.gr)
- Assistant Professor Sotirios Xydis, NTUA: (sxydis@microlab.ntua.gr)