Design and Integration of a Multi-Precision Systolic GEMM Accelerator in RISC-V Vector Clusters via AME Extensions for Edge AI Applications

Recent advances in open instruction-set architectures and modular hardware design are enabling a new generation of energy-efficient computing platforms for artificial intelligence at the edge, where demanding workloads such as inference and on-device adaptation/fine-tuning must be executed under tight power, area, and latency constraints. In this context, RISC-V provides a strategic foundation for innovation, thanks to its openness, extensibility, and growing ecosystem of software and hardware tools. In particular, the emergence of RISC-V vector and matrix-oriented extensions offers a promising path to support dense linear algebra kernels—most notably General Matrix Multiplication (GEMM), which constitutes the computational backbone of modern AI workloads—through architectures that combine programmability, scalability, and domain-specific acceleration.

A key challenge remains the efficient integration of matrix-centric acceleration into heterogeneous RISC-V compute clusters, in a way that preserves software usability and enables high performance across multiple numerical precisions. While vector processing provides flexibility and broad applicability, AI inference and fine-tuning workloads can substantially benefit from dedicated systolic GEMM accelerators capable of exploiting data reuse and regular communication patterns to improve throughput and energy efficiency. In this scenario, the adoption of RISC-V Matrix / AME (Architectural Matrix Extension) mechanisms is particularly relevant, as it enables a principled and standardized interface between the programmable core complex and the matrix accelerator, facilitating co-processor integration, software toolchain support, and future portability across implementations. This approach is especially important for edge SoCs, where tight HW/SW co-design and efficient orchestration between vector, scalar, and matrix compute domains are essential to achieve competitive performance within constrained resources.

This Incarico di Ricerca, aligned with the objectives of the DARE (Digital Autonomy with RISC-V in Europe) initiative and its broader vision of strengthening European digital autonomy in HPC and AI through open, European RISC-V technologies, focuses on the design and evaluation of matrix-acceleration solutions for AI workloads. In particular, the research activity will address: (1) the design of a multi-precision floating-point systolic GEMM accelerator targeting AI inference and fine-tuning on edge SoCs; (2) the integration of the accelerator as a co-processor within a RISC-V vector cluster through RISC-V AME extensions, with attention to architectural interfacing and software accessibility; and (3) the functional validation, as well as performance, power, and area characterization, and benchmarking on representative AI kernels. The overall goal is to advance the state of the art in open, efficient, and portable AI acceleration, contributing to the development of European RISC-V-based computing technologies and design methodologies that are consistent with the long-term roadmap of the DARE project.

Progettazione e integrazione di un acceleratore GEMM sistolico multi-precisione in cluster RISC-V vettoriali tramite estensioni AME per applicazioni di AI all'edge

I recenti progressi nelle architetture di processore aperte e modulari stanno abilitando una nuova generazione di piattaforme di calcolo ad alta efficienza energetica per l'intelligenza artificiale all'edge, in cui carichi di lavoro complessi quali l'inferenza e l'adattamento/fine-tuning on-device devono essere eseguiti nel rispetto di stringenti vincoli di potenza, area e latenza. In questo contesto, RISC-V rappresenta una base strategica per l'innovazione, grazie alla sua natura aperta, alla sua estensibilità e alla crescente maturità del relativo ecosistema hardware e

software. In particolare, l'emergere di estensioni vettoriali e matriciali per RISC-V offre una prospettiva particolarmente promettente per il supporto efficiente dei kernel di algebra lineare densa—e in particolare della General Matrix Multiplication (GEMM), che costituisce il nucleo computazionale di gran parte dei moderni carichi di lavoro AI—mediante architetture che combinano programmabilità, scalabilità e accelerazione domain-specific.

Una sfida centrale riguarda tuttavia l'integrazione efficiente dell'accelerazione matriciale all'interno di cluster di calcolo RISC-V eterogenei, in modo da preservare l'usabilità software e al contempo garantire elevate prestazioni su molteplici precisioni numeriche. Sebbene l'elaborazione vettoriale offra flessibilità e ampia applicabilità, i carichi di lavoro di inferenza e fine-tuning AI possono trarre un beneficio sostanziale dall'impiego di acceleratori GEMM sistolici dedicati, in grado di sfruttare il riuso dei dati e pattern regolari di comunicazione per incrementare throughput ed efficienza energetica. In tale scenario, l'adozione di estensioni RISC-V Matrix / AME (Architectural Matrix Extension) risulta particolarmente rilevante, poiché consente di definire un'interfaccia architetturale standardizzata e coerente tra il complesso di core programmabili e l'acceleratore matriciale, facilitando l'integrazione come co-processore, il supporto da parte della toolchain software e la futura portabilità tra diverse implementazioni. Tale approccio è di particolare interesse per gli edge SoC, nei quali un attento co-design HW/SW e un'orchestrazione efficiente tra domini di calcolo scalari, vettoriali e matriciali sono essenziali per ottenere prestazioni competitive entro risorse limitate.

Il presente Incarico di Ricerca, in coerenza con gli obiettivi dell'iniziativa DARE (Digital Autonomy with RISC-V in Europe) e con la più ampia visione di rafforzamento dell'autonomia digitale europea nei settori HPC e AI attraverso tecnologie RISC-V aperte e sviluppate in Europa, è finalizzato alla progettazione e alla valutazione di soluzioni di accelerazione matriciale per carichi di lavoro di intelligenza artificiale. In particolare, l'attività di ricerca riguarderà: (1) la progettazione di un acceleratore GEMM sistolico floating-point multi-precisione, destinato a applicazioni di inferenza e fine-tuning AI su edge SoC; (2) l'integrazione dell'acceleratore come co-processore all'interno di un cluster RISC-V vettoriale mediante estensioni RISC-V AME, con attenzione agli aspetti di interfacciamento architetturale e accessibilità software; e (3) la validazione funzionale, nonché la caratterizzazione in termini di prestazioni, potenza e area, e il benchmarking su kernel AI rappresentativi. L'obiettivo complessivo è avanzare lo stato dell'arte nell'accelerazione AI aperta, efficiente e portabile, contribuendo allo sviluppo di tecnologie e metodologie di progettazione RISC-V europee coerenti con la roadmap di lungo periodo del progetto DARE.