Design and Evaluation of Reliable Multi-Precision Systolic GEMM Acceleration for AI on RISC-V Edge Systems

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, enables high performance across multiple numerical precisions, and increasingly addresses reliability and fault-tolerance requirements. 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. At the same time, the deployment of AI at the edge in safety-critical and harsh environments, including emerging edge-space scenarios, requires accelerator architectures and compute flows that can sustain execution under transient faults and radiation-induced errors.

This Incarico di Ricerca position, aligned with the activities of the ARCHYTAS project, focuses on the design and evaluation of reliable matrix-acceleration solutions for AI workloads. In particular, the research activity will address:

1) the extension of the architecture and micro-architecture of a systolic GEMM accelerator targeting AI inference and fine-tuning on edge SoCs, with support for emerging AI-oriented data formats and multi-precision computation;
2) the reliability and fault-tolerance analysis of AI algorithms and their execution on the GEMM accelerator, including the study of the intrinsic reliability properties of different numerical/data formats and of alternative micro-architectures for the basic compute units (e.g., FPUs/MAC units) used to implement them;
3) the integration of the multi-precision GEMM accelerator into heterogeneous compute systems, targeting edge platforms and extending the investigation to edge-space application scenarios;
4) the functional validation, performance/power/area characterization, and benchmarking on representative AI kernels, assessing not only performance and energy efficiency, but also error resilience under radiation-induced faults.

The overall goal is to advance the state of the art in reliable and energy-efficient acceleration in multi-tile heterogeneous compute systems, contributing to the development of European sovereignity for computing technologies and design methodologies that are consistent with the long-term roadmap of the ARCHYTAS project.

Progettazione e Valutazione di un'Accelerazione Sistolica GEMM Multi-Precisione Affidabile per l'IA su Sistemi Edge basati su RISC-V

I recenti progressi nelle architetture aperte a set di istruzioni e nel design modulare dell'hardware stanno abilitando una nuova generazione di piattaforme di calcolo efficienti dal punto di vista energetico per l'intelligenza artificiale all'edge, dove carichi di lavoro impegnativi come l'inferenza e l'adattamento/fine-tuning on-device devono essere eseguiti sotto rigidi vincoli di potenza, area e latenza. In questo contesto, RISC-V fornisce una base strategica per l'innovazione grazie alla sua apertura, estensibilità e al crescente ecosistema di strumenti software e hardware. In particolare, l'emergere delle estensioni vettoriali e orientate alle matrici di RISC-V offre un percorso promettente per supportare i kernel di algebra lineare densa — in particolare la Moltiplicazione Generale di Matrici (GEMM), che costituisce la spina dorsale computazionale dei moderni carichi di lavoro AI — attraverso architetture che combinano programmabilità, scalabilità e accelerazione domain-specific.

Una sfida chiave rimane l'integrazione efficiente dell'accelerazione matrix-centrica in cluster di calcolo RISC-V eterogenei, in modo da preservare la fruibilità del software, abilitare alte prestazioni su più precisioni numeriche e rispondere in misura crescente ai requisiti di affidabilità e tolleranza ai guasti. Sebbene il processamento vettoriale offra flessibilità e ampia applicabilità, i carichi di lavoro di inferenza AI e fine-tuning possono trarre notevole beneficio da acceleratori GEMM sistolici dedicati, capaci di sfruttare il riuso dei dati e i pattern di comunicazione regolari per migliorare il throughput e l'efficienza energetica. Allo stesso tempo, il dispiegamento dell'AI all'edge in ambienti safety-critical e ostili — inclusi gli emergenti scenari edge-spaziali — richiede architetture di acceleratori e flussi di calcolo in grado di sostenere l'esecuzione in presenza di guasti transienti ed errori indotti da radiazioni.

Questa posizione di Incarico di Ricerca, allineata con le attività del progetto ARCHYTAS, si concentra sulla progettazione e valutazione di soluzioni affidabili di accelerazione matriciale per carichi di lavoro AI. In particolare, l'attività di ricerca affronterà:

1. l'estensione dell'architettura e della micro-architettura di un acceleratore GEMM sistolico destinato all'inferenza AI e al fine-tuning su SoC edge, con supporto per formati di dati emergenti orientati all'AI e calcolo multi-precisione;
2. l'analisi di affidabilità e tolleranza ai guasti degli algoritmi AI e della loro esecuzione sull'acceleratore GEMM, incluso lo studio delle proprietà intrinseche di affidabilità dei diversi formati numerici/di dati e di micro-architetture alternative per le unità di calcolo di base (ad es., FPU/unità MAC) utilizzate per implementarli;
3. l'integrazione dell'acceleratore GEMM multi-precisione in sistemi di calcolo eterogenei, con target su piattaforme edge ed estensione dell'indagine a scenari applicativi edge-spaziali;
4. la validazione funzionale, la caratterizzazione di prestazioni/potenza/area e il benchmarking su kernel AI rappresentativi, valutando non solo le prestazioni e l'efficienza energetica, ma anche la resilienza agli errori in presenza di guasti indotti da radiazioni.

L'obiettivo complessivo è far avanzare lo stato dell'arte nell'accelerazione affidabile ed efficiente dal punto di vista energetico in sistemi di calcolo eterogenei multi-tile, contribuendo allo sviluppo della sovranità europea nelle tecnologie di calcolo e nelle metodologie di progettazione coerenti con la roadmap a lungo termine del progetto ARCHYTAS.