Reliable AI Deployment on Heterogeneous Host-Accelerator Systems with Programmable RISC-V Accelerators by Exploiting Hardware Reliability Features

Recent advances in open and modular computing architectures are enabling a new generation of heterogeneous platforms for Artificial Intelligence applications, where a host processor cooperates with specialized programmable accelerators. In particular, platforms based on RISC-V components provide a promising foundation to combine programmability, energy efficiency, and acceleration capabilities for AI inference workloads and data-intensive signal-processing applications. However, effective application deployment on such systems remains a major challenge, due to the need to coordinate heterogeneous compute elements, manage data movement across host and accelerators, and preserve execution predictability.
In addition to classical performance and efficiency objectives, deployment reliability is becoming increasingly important, especially in the presence of transient faults, hardware variability, and non-ideal operating conditions. In this scenario, a key research challenge is the development of deployment methodologies and software support capable of systematically exploiting the heterogeneous platform while integrating hardware-supported monitoring, protection, and mitigation mechanisms.
This Incarico di Ricerca position, aligned with the activities of the ARCHYTAS project, focuses on the development of methodologies and software tools for the reliable deployment of AI applications on a heterogeneous system composed of a host and two programmable accelerators: a RISC-V scalar multi-core accelerator and a RISC-V vector accelerator. The work will build on deployment frameworks and toolflows (including Deeploy) to support application mapping, code generation, and execution orchestration across host and accelerators, while integrating reliability-aware strategies that exploit hardware reliability features (e.g., error detection/correction mechanisms, integrity checks, status monitoring, and diagnostic support).
The research activities will include:

1. developing reliability-aware deployment and partitioning strategies for AI applications, including the allocation of operators and computational phases across the host, the RISC-V scalar multi-core accelerator, and the RISC-V vector accelerator;
2. extending the compilation/deployment toolflow and runtime software support for heterogeneous host–accelerator platforms, including protection policies for critical data and operators;
3. designing runtime monitoring and mitigation mechanisms that exploit hardware reliability features to detect anomalies and enable fallback, reconfiguration, or workload reallocation strategies;
4. defining metrics and evaluation methodologies to quantify the impact of the proposed techniques on accuracy, fault coverage, latency, and performance/energy overhead, including experimental campaigns and fault injection;
5. experimentally validating the proposed solutions on representative AI inference case studies.

The overall goal is to advance the state of the art in portable, efficient, and reliable AI deployment on heterogeneous host–accelerator systems based on programmable RISC-V accelerators, through an integrated approach combining software toolflows, runtime techniques, and hardware reliability features.

Deployment affidabile di applicazioni AI su sistemi eterogenei con host e acceleratori programmabili RISC-V mediante sfruttamento di funzionalità hardware di affidabilità

I recenti progressi nelle architetture di calcolo aperte e modulari stanno abilitando una nuova generazione di piattaforme eterogenee per l'esecuzione di applicazioni di Intelligenza Artificiale,

in cui un processore host coopera con acceleratori programmabili specializzati. In particolare, le piattaforme basate su componenti RISC-V rappresentano una soluzione promettente per coniugare flessibilità programmabile, efficienza energetica e capacità di accelerazione per workload di inferenza AI e di elaborazione di segnali ad alta intensità di dati. Tuttavia, il deployment efficace delle applicazioni su tali sistemi rimane una sfida aperta, a causa della necessità di coordinare elementi di calcolo eterogenei, gestire i movimenti di dati tra host e acceleratori e garantire prevedibilità del comportamento in esecuzione.
Accanto agli obiettivi classici di prestazioni ed efficienza, assume inoltre crescente rilievo il tema dell'affidabilità (reliability) del deployment, soprattutto in presenza di fault transitori, variabilità hardware e condizioni operative non ideali. In questo scenario, una problematica di ricerca centrale consiste nello sviluppo di metodologie di deployment e supporti software capaci di sfruttare in modo sistematico le caratteristiche della piattaforma eterogenea e, al tempo stesso, integrare meccanismi di monitoraggio, protezione e mitigazione supportati dall'hardware.
Questo Incarico di Ricerca, allineato con le attività del progetto ARCHYTAS, è focalizzato sullo sviluppo di metodologie e strumenti software per il deployment affidabile di applicazioni AI su un sistema eterogeneo composto da un host e due acceleratori programmabili: un acceleratore RISC-V multi-core scalar e un acceleratore RISC-V vettoriale. L'attività farà leva su framework e toolflow di deployment (tra cui Deeploy) per supportare la mappatura delle applicazioni, la generazione del codice e l'orchestrazione dell'esecuzione tra host e acceleratori, integrando strategie reliability-aware che sfruttino funzionalità hardware di affidabilità (ad es. meccanismi di error detection/correction, controlli di integrità, monitoraggio dello stato e supporto alla diagnosi).
Le attività di ricerca includeranno:

1. lo sviluppo di strategie di deployment e partizionamento reliability-aware per applicazioni AI, con allocazione di operatori e fasi computazionali tra host, acceleratore RISC-V multi-core scalar e acceleratore RISC-V vettoriale;
2. l'estensione del toolflow di compilazione/deployment e del supporto runtime per piattaforme eterogenee host–accelerator, includendo politiche di protezione per dati e operatori critici;
3. la progettazione di meccanismi di monitoraggio e mitigazione in esecuzione, capaci di sfruttare funzionalità hardware di reliability per rilevare anomalie e abilitare strategie di fallback, riconfigurazione o riallocazione del carico;
4. la definizione di metriche e metodologie di valutazione per quantificare l'impatto delle tecniche proposte su accuratezza, fault coverage, latenza e overhead prestazionale/energetico, anche mediante campagne sperimentali e fault injection;
5. la validazione sperimentale delle soluzioni proposte su casi di studio rappresentativi di inferenza AI.

L'obiettivo complessivo è avanzare lo stato dell'arte nel deployment portabile, efficiente e affidabile di applicazioni AI su sistemi eterogenei host–accelerator basati su acceleratori programmabili RISC-V, attraverso un approccio integrato che coniughi toolflow software, runtime e funzionalità hardware di affidabilità.