Software for AI Accelerators with Advanced Memory Integration

Recent advances in open and modular processor architectures are enabling a new generation of heterogeneous many-core systems that tightly integrate general-purpose cores with domain-specific hardware accelerators for machine learning. The performance and energy efficiency promised by these accelerators, however, can only be realized if the software stack is able to exploit them - and this becomes especially challenging when the architecture incorporates advanced memory integration techniques such as analog and digital in-memory computing, non-volatile CMOS-compatible PCM devices, multi-chiplet scaling, and 3D-stacked memories. These emerging memory technologies break the assumptions of conventional programming and compilation flows: data placement, tiling, and movement across a deep and heterogeneous memory hierarchy become first-class concerns that must be modeled, explored, and optimized before silicon is available. Virtual platforms that faithfully capture the behaviour and cost of advanced and 3D memory integration, together with retargetable compiler infrastructures such as MLIR, are therefore essential to close the gap between novel AI accelerator hardware and the deployment of real applications onto it.

This Incarico di Ricerca position, which is aligned with the activity of the **ARCHYTAS** project, focuses on the development of the software stack required to program and deploy real applications onto AI accelerators that adopt advanced memory integration. The incaricato di ricerca will focus on: 1) development and extension of virtual platforms that model advanced and 3D memory integration (e.g. in-memory computing, PCM-based and 3D-stacked memories) for early, pre-silicon evaluation of AI accelerators; 2) design of compiler infrastructure based on MLIR for mapping, tiling and scheduling of AI workloads onto the heterogeneous and deep memory hierarchy; 3) end-to-end application deployment and design-space exploration, using the virtual platform and the compiler flow to optimize data placement and movement across the integrated memory subsystem.

Software for AI Accelerators with Advanced Memory Integration

I recenti progressi nelle architetture di processori aperti e modulari stanno abilitando una nuova generazione di sistemi many-core eterogenei che integrano strettamente core general-purpose con acceleratori hardware domain-specific per il machine learning. Le prestazioni e l'efficienza energetica promesse da questi acceleratori, tuttavia, possono essere realizzate solo se lo stack software è in grado di sfruttarli — e ciò diventa particolarmente difficile quando l'architettura incorpora tecniche avanzate di integrazione della memoria quali il computing in-memory analogico e digitale, i dispositivi PCM compatibili con CMOS e non volatili, lo scaling multi-chiplet e le memorie 3D-stacked. Queste tecnologie di memoria emergenti infrangono le assunzioni dei flussi convenzionali di programmazione e compilazione: il posizionamento dei dati, il tiling e il loro movimento attraverso una gerarchia di memoria profonda ed eterogenea diventano aspetti di primo piano, che devono essere modellati, esplorati e ottimizzati prima che il silicio sia disponibile. Piattaforme virtuali che catturino fedelmente il comportamento e il costo dell'integrazione di memoria avanzata e 3D, insieme a infrastrutture di compilazione retargetabili come MLIR, sono pertanto essenziali per colmare il divario tra il nuovo hardware degli acceleratori AI e il dispiegamento di applicazioni reali su di esso.

Questa posizione di Incarico di Ricerca, allineata con le attività del progetto **ARCHYTAS**, si focalizza sullo sviluppo dello stack software necessario per programmare e dispiegare applicazioni reali su acceleratori AI che adottano un'integrazione di memoria avanzata. L'incaricato di ricerca si concentrerà su: 1) sviluppo ed estensione di piattaforme virtuali che modellino l'integrazione di memoria avanzata e 3D (ad es. computing in-memory, memorie basate su PCM e 3D-stacked) per la valutazione precoce e pre-silicio degli acceleratori AI; 2) progettazione di un'infrastruttura di compilazione basata su MLIR per il mapping, il tiling e lo scheduling di carichi di lavoro AI sulla gerarchia di memoria eterogenea e profonda; 3) dispiegamento end-to-end di applicazioni ed esplorazione dello spazio di progetto, utilizzando la

piattaforma virtuale e il flusso di compilazione per ottimizzare il posizionamento e il movimento dei dati attraverso il sottosistema di memoria integrato.