ASSEGNO DI RICERCA - FONDI EPI (Prof. Luca Benini) e WIPLASH (Prof. Davide Rossi)

Tutor: Prof. Davide Rossi

Heterogeneous Computing Platforms for AI acceleration

The context of this work is the increasing irrelevance of Moore's Law, which observed that the number of transistors that could be put on a chip at the same price doubled every 18 to 24 months. For more to fit they had to get smaller, which let them run faster, albeit hotter, so performance rose over the years — but so did expectations. Today, those expectations remain, but processor performance has plateaued. This leads to the search of new architectures with higher energy efficient that just faster single processors.

The goal is to develop micro-architectural and macro-architectural and software concepts related to low-power many core heterogeneous systems exploiting in-package wireless communication and in-memory computing accelerators in both high-performance and deeply embedded domain. This is in line with the goal of the EPI project i.e. develop heterogeneous many-core systems exploiting accelerators for artificial intelligence workloads with improved power efficiency and scalability exploiting specialized multi-core architectures, also exploiting new concepts such as wireless plasticity and in-memory computing envisioned by Wiplash project.

The main topics that will be studied in this work are:

1. Scalable many-core heterogeneous architectures
2. Hardware-software optimization
3. Modelling of accelerators for CNN workloads
4. Simulation od wireless channels for chip-to-chip communication

Architetture di Calcolo Eterogenee per Accelerazione di algoritmi IA

Il contesto di questo lavoro è la crescente irrilevanza della Legge di Moore, che era basata sull'osservazione empirica che il numero di transistor che potrebbero essere messi su un chip allo stesso prezzo veniva raddoppiato ogni 18-24 mesi. Cio' implica che i transistori CMOS devono essere sempre più piccoli, il che permette loro di funzionare più velocemente, anche se a temperature operative piu' alte. In questo modo così le prestazioni aumentarono nel corso degli anni, ma anche le aspettative. Oggi, queste aspettative rimangono, ma le prestazioni dei processori si sono stabilizzate in termini di frequenza operativa, a causa dell'intollerabile aumento della densita' di potenza dissipata. Questo porta alla ricerca di architetture con una maggiore efficienza energetica piuttosto che focalizzarsi esclusivamente su processori singoli più veloci.

L'obiettivo è lo sviluppo di concetti micro-architetturali e macro-architetturali e software relativi a sistemi scalabili a bassa potenza sia nel dominio ad alte prestazioni che in quello profondamente integrato. Questo è in linea con l'obiettivo del progetto EPI, ovvero sviluppare acceleratori paralleli con efficienza energetica e scalabilità migliorate sfruttando architetture multi-core eterogenee, sfruttando anche concetti emergenti di plasticità wireless per la comunicazione on-chip ed off-chip e accelerazione in-memory computing, obiettivi del progetto Wiplash.

Gli argomenti principali che saranno studiati in questo lavoro sono:

1. Architetture multi-core eterogenee
2. Ottimizzazione hardware-software
3. Modellazione di acceleratori per carichi computazionali convoluzionali
3. Simulazione di canali wireless per la comunicazione off-chip