Contribution ID: **112** Type: **Presentazione orale**

Sviluppo di acceleratori per il Machine Learning e sistemi di Inference as a Service su FPGA

Wednesday, 24 May 2023 11:30 (30 minutes)

I Field Programmable Gate Arrays (FPGAs) sono una tecnologia rivoluzionaria per l'inferenza di Machine Learning (ML) grazie alla loro architettura altamente parallela, basso consumo energetico e capacità di eseguire algoritmi personalizzati. Lo sviluppo di progetti di sintesi di alto livello [1, 2] che semplificano la programmazione HDL ha portato ad un aumento significativo dell'uso di FPGA nel settore ML, abilitando l'uso di questi dispositivi in sistemi di ML as a Service per il calcolo scientifico [3]. In questa presentazione, descriveremo la nostra esperienza nella creazione di acceleratori per implementare algoritmi ML su FPGA, a partire dalla generazione del firmware utilizzando un nuovo tipo di architettura [4] adatta a modelli computazionali, fino all'utilizzo ad alto livello dell'acceleratore stesso. La flessibilità del modello proposto consente molte tipologie di ottimizzazione, sia per la precisione numerica che per l'architettura, e verranno mostrati i risultati ottenuti in termini di: utilizzo delle risorse, velocità di inferenza ed efficienza energetica.
Infine, mostreremo un prototipo di ecosistema OpenSource che facilita l'uso di FPGA per il calcolo scientifico, rendendolo più accessibile e indipendente dal fornitore. Il progetto proposto è costruito attorno a KServe, uno dei software Inference as a Service più flessibili nell'ecosistema cloud-native, estendendo le sue capacità con un framework di Generation as a Service del firmware FPGA.

Primary authors: SPIGA, Daniele (Istituto Nazionale di Fisica Nucleare); CIANGOTTINI, Diego (INFN Perugia); SURACE, Giacomo (Istituto Nazionale di Fisica Nucleare); BIANCHINI, Giulio (Istituto Nazionale di Fisica Nucleare); STORCHI, Loriano (Istituto Nazionale di Fisica Nucleare); MARIOTTI, Mirko (Istituto Nazionale di Fisica Nucleare)

Presenter: BIANCHINI, Giulio (Istituto Nazionale di Fisica Nucleare)

Session Classification: Tecnologie ICT Hardware e Software

Track Classification: Tecnologie ICT (Harware e Software)