پژوهشگاه دانش‌های بنیادی
پژوهشکده علوم کامپیوتر

IPM

سخنرانی‌های علمی

HPIML: Heterogeneous Process In Memory for Machine Learning Algorithms

By: Dr. Hajar Falahati
IPM School of Computer Science

Abstract

Advent of big data applications makes Machine Learning (ML) algorithms such as those involving convolutional and deep neural networks (CNNs and DNNs) popular in a wide range of applications. ML algorithms are categorized as compute-intensive ones which carry out considerable computations on huge amount of data in both learning and inference phases. Research trend also shows that ML algorithms are growing in size so rapidly. Their number of layers, weights, and input vectors are increasing over the years. Owing to the inherent parallel nature of ML algorithms, a lot of platforms such as FPGAs, GPUs, and ASICs have been proposed to accelerate ML algorithms. However, due to their huge memory footprints, the current processing platforms fail to provide their required computational and memory demands. To keep computational resources busy, these accelerators need to transfer huge amount of data which makes memory subsystem a serious bottleneck in accelerating ML algorithms.

To tackle the memory gap, process-in-memory (PIM) has been suggested as a solution to bridge the gap; however, power and area optimizations are serious concerns with PIM. Moreover, units placed in the memory side need to be designed in general and reconfigurable manner to be capable to execute various kinds of ML algorithms. To address these challenges, we propose a novel heterogeneous accelerator for ML algorithms, which coupled Near Data Processing with Pattern-Aware execution in order to solve the gradient decent of a wide range of ML algorithms. First, through analysis of different ML algorithms, we show that different ML algorithms whose objective functions are totally different can be presented by known patterns. Then, we map these pattern-based descriptions into a suite of logical parts implemented by simple hardware blocks. These specific hardware blocks satisfy both the ML requirements and 3D-DRAM limitations. To support new ML algorithms whose size is growing rapidly, our proposed idea follows heterogeneous execution style by distributing hardware blocks over a various kind of processing platforms such as 3D-stacked memories and state-of-the art FPGAs. Considering the inherent parallelism of ML algorithms, we devise a smart partitioning scheme that minimize the communication overhead and fully utilize the resources of both platforms.

Biography

Hajar Falahti is a Postdoc researcher within the School of Computer Science at \Institute for Research in Fundamental Sciences (IPM). She received her BSc in Hardware Engineering in 2009, in Isfahan University of Technology, Master of Computer Architecture in 2011, and PhD in Computer Architecture in late 2016, in Sharif University of Technology, Iran. Her research interests include computer architecture, high-performance computing, low-power design, hardware accelerators (e.g., GPU), and bioinformatics.

زمان : پنج‌شنبه ۹۶/۸/۱۱ – ساعت ۱۵

مکان : فرمانیه – خیابان شهید لواسانی – جنب برج کوه نور – نبش خیابان فربین – پژوهشگاه دانش‌های بنیادی – طبقه همکف

***** شرکت برای عموم علاقه‌مندان آزاد است *****