Ultra-Efficient Processing In-Memory for Data Intensive Applications

Title	Ultra-Efficient Processing In-Memory for Data Intensive Applications
Publication Type	Conference Paper
Year of Publication	2017
Authors	Imani, Mohsen, Gupta, Saransh, Rosing, Tajana
Conference Name	Proceedings of the 54th Annual Design Automation Conference 2017
Publisher	ACM
Conference Location	New York, NY, USA
ISBN Number	978-1-4503-4927-7
Keywords	analogical transfer, analogies, Emerging computing, Human Behavior, human factors, non-volatile memory, Processing in-memory, pubcrawl, Transference

Abstract

Recent years have witnessed a rapid growth in the domain of Internet of Things (IoT). This network of billions of devices generates and exchanges huge amount of data. The limited cache capacity and memory bandwidth make transferring and processing such data on traditional CPUs and GPUs highly inefficient, both in terms of energy consumption and delay. However, many IoT applications are statistical at heart and can accept a part of inaccuracy in their computation. This enables the designers to reduce complexity of processing by approximating the results for a desired accuracy. In this paper, we propose an ultra-efficient approximate processing in-memory architecture, called APIM, which exploits the analog characteristics of non-volatile memories to support addition and multiplication inside the crossbar memory, while storing the data. The proposed design eliminates the overhead involved in transferring data to processor by virtually bringing the processor inside memory. APIM dynamically configures the precision of computation for each application in order to tune the level of accuracy during runtime. Our experimental evaluation running six general OpenCL applications shows that the proposed design achieves up to 20x performance improvement and provides 480x improvement in energy-delay product, ensuring acceptable quality of service. In exact mode, it achieves 28x energy savings and 4.8x speed up compared to the state-of-the-art GPU cores.

URL https://dl.acm.org/citation.cfm?doid=3061639.3062337
DOI 10.1145/3061639.3062337
Citation Key imani_ultra-efficient_2017

Analogical Transfer analogies Emerging computing Human behavior Human Factors non-volatile memory Processing in-memory pubcrawl Transference