IPM

پژوهشگاه دانشهای بنیادی
پژوهشکده علوم
کامپیوتر

سخنرانی علمی

SPAGHETTI: Streaming Accelerators for Highly Sparse GEMM on FPGAs

By: Reza Hojabr
University of Tehran

Abstract

Generalized Sparse Matrix-Matrix Multiplication (Sparse GEMM) is widely used across multiple domains, but the computation's regularity is dependent on the input sparsity pattern. The majority of sparse GEMM accelerators are based on the inner product method and propose new storage formats to regularize computation. We find that these storage formats are more suited for denser matrices. Accelerators adopting the outer product algorithm are more suitable for highly sparse inputs, since they support CSC/CSR storage formats. In this talk, we introduce *Spaghetti*, an open-source Chisel generator for creating FPGA-optimized outer product accelerators. The key novelty in *Spaghetti* is a new pattern-aware software scheduler that analyzes the sparsity pattern and schedules row-col pairs of the inputs onto the fixed microarchitecture. *Spaghetti* takes advantage of our observation that the rows in the input matrix lead to mutually independent rows in the final output. Thus the scheduler can partition the input into tiles that maximize reuse and eliminate re-fetching the partial matrices from the DRAM. The microarchitecture template we create has the following key benefits: i) we can statically schedule the inputs in a streaming fashion and maximize DRAM utilization, ii) we can parallelize the merge phase and generate multiple rows of the output in parallel maximally using the output DRAM bandwidth, iii) we can adapt to the varying logic resources and bandwidth across various FPGA devices and attain maximal roofline performance (only limited by memory bandwidth). We auto-generate sparse GEMM accelerators on Amazon AWS FPGAs and demonstrate that we can achieve performance improvement over CPUs and GPUs between 1.1--34.5×. Compared to the state-of-the-art outer product accelerator, our design improves performance by an average of 2.6×, and reduces DRAM accesses by an average of 4×.

Biography

Reza Hojabr is a last year PhD candidate in Computer Engineering at the University of Tehran, Iran under supervision of Dr. Ahmad Khonsari. Currently, he is a researcher in the Computer Systems Lab at Simon Fraser University, Canada under supervision of Dr. Arrvindh Shriraman. He received his BSc and MSc from the University of Tehran and K.N.Toosi University of Technology in 2014 and 2012, respectively. His research is centered on application-specific accelerators, approximate computing and agile hardware design, published in top conferences and journals such as HPCA, MICRO, DAC, ISCAS and IEEE TC.

زمان : چهارشنبه ۱۳۹۹/۱۰/۱۰ - ساعت ۱۳:۰۰
ارائه به صورت مجازی انجام خواهد شد.

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