NJIT | INSTITUTE FOR DATA SCIENCE

Data Science Seminar Series

Exploiting Sparsity in Deep Neural Network Accelerator Hardware

Joel Emer, Ph.D.
Professor
Massachusetts Institute of Technology

Date:	Wednesday, April 20th, 2022
Time:	4:00 PM – 5:00 PM EDT
Web Link:	https://njit-institute-for-datascience.eventbrite.com/

Recently it has increasingly been observed that exploiting sparsity in hardware for linear algebra computations can result in significant performance improvements. This is because for data that has many zeros compression can reduce reduce both storage space and data movement. In addition, it is possible to take advantage of the simple mathematical equality that anything times zero equals zero because it results in what is commonly referred to as an ineffectual operation. Eliminating spending time do ineffectual operations and the data accesses associated with them can result in a considerable performance and energy improvements over hardware that performs all computations both effectual and ineffectual. One especially popular domain for exploiting sparsity is in deep neural network (DNN) computations, where the operands are often sparse because the input activations have zeros in them introduced by the non-linear RELU operation and the weights may have been explictly pruned such that many of them are zero. Previously proposed deep neural network accelerators have employed a variety of computational dataflows and techniques to compress data to optimize performance and energy efficiency.

In an analogous fashion to our prior work that categorized DNN dataflows into patterns like weight stationary and output stationary, this talk will try to characterize the range of sparse DNN accelerators. Thus, rather than presenting a single specific combination of a dataflow and concrete data representation, I will present a generalized framework for describing dataflows and their manipulation of sparse tensor operands. In this framework, the dataflow and the representation of the operands are expressed independently in order to better facilitate the exploration of the wide design space of sparse DNN accelerators. Therefore, I will begin by presenting a format-agnostic abstraction for sparse tensors, called fibertrees. Using the fibertree abstraction, one can express a wide variety of concrete data representations, each with its own advantages and disadvantages. Furthermore by adding a set of operators for activities, like traversal and merging of tensors, the fibertree notation can be used to express dataflows independent of the concrete data representation used for the tensor operands. Thus, using this common language, I will describe a variety of previously proposed sparse neural network accelerator designs, highlighting the choices they made. Finally, I will present the some work on how this framework can be used as the basis of an analytic framework for evaluating the effectiveness of various sparse optimizations in accelerator designs.

For over 40 years, Joel Emer held various research and advanced development positions investigating processor microarchitecture and developing performance modeling and evaluation techniques. He has made architectural contributions to a number of VAX, Alpha and X86 processors and is recognized as one of the developers of the widely employed quantitative approach to processor performance evaluation. He is also well known for his contributions to the advancement of deep learning accelerator design, spatial and parallel architectures, processor reliability analysis, cache organization and simultaneous multithreading. Currently he is a professor at the Massachusetts Institute of Technology and spends part time as a Senior Distinguished Research Scientist in Nvidia's Architecture Research group. Previously, he worked at Intel where he was an Intel Fellow and Director of Microarchitecture Research. Even earlier, he worked at Compaq and Digital Equipment Corporation. He earned a doctorate in electrical engineering from the University of Illinois in 1979. He received a bachelor's degree with highest honors in electrical engineering in 1974, and his master's degree in 1975 -- both from Purdue University. Recognitions of his contributions include an ACM/SIGARCH-IEEE-CS/TCCA Most Influential Paper Award for his work on simultaneous multithreading, and six other papers that were selected as IEEE Micro's Top Picks in Computer Architecture. Among his professional honors, he is a Fellow of both the ACM and IEEE, and a member of the NAE. In 2009 he was recipient of the Eckert-Mauchly award for lifetime contributions in computer architecture.