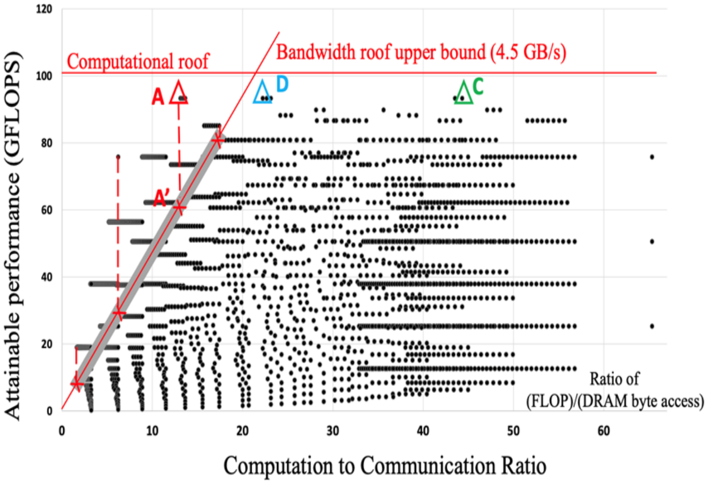
Optimizing FPGA-based accelerator design for deep convolutional neural networks

Chen Zhang, Peng Li, Guangyu Sun, Yijin Guan, Bingjun Xiao, and Jason Cong

Published: *Proceedings of the 2015 ACM/SIGDA 23rd international symposium on Field programmable gate arrays (FPGA'15), February 2015, pp. 161–170.*

This paper is an essential read for researchers who are interested in AI hardware and architecture. For the first time, this work presents a comprehensive, holistic methodology that optimizes both computation and memory access simultaneously for FPGA-based CNN Accelerators. By leveraging the roofline model for an in-depth analysis of the design space, it presents a framework that can serve as a guiding principle for future AI hardware accelerator research.

The paper introduces innovative loop-based analysis techniques that provide a complete and accurate abstraction for CNN design variants. It covers key aspects such as computation engines, memory accesses, buffer sizes, and hardware constraints, offering a well-rounded view of the design space. By integrating computational optimization and memory access strategies, the paper uses the roofline model to systematically explore all the legal solutions within the design space. This thorough examination ensures that the selected solutions meet the constraints of FPGA platform bandwidth and computational resources. The final design is chosen based on the compute-to-communication (CTC) ratio, which strikes an optimal balance between performance and hardware resource efficiency. This method reduces the need for excessive I/O ports, look-up tables (LUTs), and hard-wired connections in the data transfer engine, showcasing exceptional hardware resource optimization.



Overall, this paper introduces a groundbreaking approach to quantitative hardware modeling and optimization for AI workloads, making a significant impact to the field of AI accelerator design. Its contributions have not only advanced the understanding of FPGA-based accelerator design but also influenced the broader landscape of AI hardware architecture. With over 2,500 citations to date, this work has been widely recognized by both academia and industry. It has been referenced by leading companies in the field of deep learning acceleration, including Google, NVIDIA, Intel, Microsoft, and Xilinx (now part of AMD), underscoring its vital role in shaping the future of AI hardware design.

Endorsement by: Yun (Eric) Liang, Professor, Peking University

DOI: https://dl.acm.org/doi/10.1145/2684746.2689060