T5-a physical-design-driven, learning-based prediction framework is proposed . the framework combines power, performance, and area (PPA) analysis with frontend performance simulation . the framework includes a fully automated DSE technique for optimizing backend and system metrics through an automated search of architectural and backend parameters . it predicts backend PPA and system metrics with an average prediction error of 7% or less for ASIC implementations of two deep learning accelerator platforms .

Bart- An Open-Source ML-Based Full-Stack Optimization Framework for Machine Learning Accelerators. Parameterizable machine learning (ML) accelerators are the product of recent breakthroughs in ML. To fully enable their design space exploration (DSE), a learning-based prediction framework is proposed. This framework combines power, performance, and area (PPA) analysis with frontend performance simulation to estimate backend PPA and system metrics such as runtime and energy realistically. Additionally, the framework includes a fully automated DSE technique for optimizing backend andsystem metrics through an automated search of architectural and backend parameters. Experimental studies demonstrate that this approach consistently predicts back-end metrics with an average prediction error of 7% or less for ASIC implementations of two deep learning accelerator platforms. The framework is open-source and free to download from the GitHub repository.

Gemini- This framework facilitates the design space exploration of parameterizable ML accelerators by combining PPA analysis with frontend performance simulation, providing realistic estimates of backend PPA and system metrics. It also offers an automated DSE technique that optimizes backend and system metrics by searching architectural and backend parameters. Experiments show accurate predictions of backend PPA and system metrics, with an average prediction error of 7% or less for ASIC implementations of two deep learning accelerator platforms.

Gpt-40 The text introduces an innovative open-source machine learning (ML)-based optimization framework specifically designed for ML accelerators, which are emerging from recent advancements in ML technology. This framework is crucial for exploring the design space of these accelerators, which are essential for both deep neural network (DNN) and non-DNN ML algorithms. The proposed framework integrates a comprehensive analysis of power, performance, and area (PPA) with frontend performance simulation. This integration enables the realistic estimation of backend PPA and critical system metrics such as runtime and energy consumption.

A key feature of this framework is its fully automated design space exploration (DSE) technique. This technique enhances the optimization of backend and system metrics by automating the search for the most effective architectural and backend parameters. This automation is pivotal in refining the design and efficiency of ML accelerators, ensuring they meet the specific needs of various applications while maintaining high performance and energy efficiency.

The effectiveness of this framework has been validated through experimental studies, which have shown that it can predict backend PPA and system metrics with a high degree of accuracy. Specifically, the framework achieves an average prediction error of 7% or less when applied to ASIC implementations of two different deep learning accelerator platforms. This level of precision in prediction underscores the framework's potential in significantly advancing the development and optimization of ML accelerators. By providing a reliable and efficient tool for DSE, the framework supports the progression towards more sophisticated and tailored ML hardware solutions, catering to the growing demand for faster and more efficient ML computations.

Gpt-3 The research paper introduces an innovative open-source machine learning (ML)-based full-stack optimization framework tailored for ML accelerators. These accelerators, a result of recent ML advancements, require a comprehensive design space exploration (DSE) approach. The proposed framework leverages physical-design-driven, learning-based predictions to optimize hardware-accelerated deep neural network (DNN) and non-DNN ML algorithms. By integrating power, performance, and area (PPA) analysis with frontend performance simulation, the framework accurately estimates backend PPA and system metrics like runtime and energy consumption. Moreover, it incorporates an automated DSE technique that systematically optimizes backend and system metrics by exploring architectural and backend parameters. Experimental results showcase the framework's effectiveness, demonstrating consistent predictions of backend PPA and system metrics with an average error rate of 7% or less for application-specific integrated circuit (ASIC) implementations on two deep learning accelerator platforms. This framework not only streamlines the optimization process but also enhances the efficiency and accuracy of ML accelerator design, offering a valuable tool for researchers and developers in the field.