COMMENTS AND OPINIONS

Journal of Semiconductors
(2020) 41, 020301
doi: 10.1088/1674-4926/41/2/020301

Reconfigurable computing: a promising microchip architecture for artificial intelligence

Shaojun Wei

Department of Microelectronics and Nanoelectronics, Tsinghua University, Beijing 100084, China

Citation: S J Wei, Reconfigurable computing: a promising microchip architecture for artificial intelligence[J]. *J. Semicond.*, 2020, 41(2), 020301. http://doi.org/10.1088/1674-4926/41/2/020301

Shaojun Wei is the Professor of the Department of Micro- and Nano-electronics, Tsinghua University, Beijing, China. His research activities range from integrated circuit design methodologies, electronic design automation (EDA), embedded system design to reconfigurable computing technologies. Prof. Wei is the fellow of IEEE and the fellow of Chinese Institute of Electronics (CIE).

Today, integrated circuit technology is approaching the physical limit. From performance and energy consumption perspective, reconfigurable computing is regarded as the most promising technology for future computing systems with excellent feature in computing and energy efficiency. From the perspective of computing performance, compared with single thread performance stagnation of general purpose processors (GPPS), reconfigurable computing may customize hardware according to application requirements, so as to achieve higher performance and lower energy consumption. From the perspective of economics, a microchip based on reconfigurable computing technology has post-silicon reconfigurability, which can be applied in different fields, so as to better share the cost of non-recurring engineering (NRE). High computing and energy efficiency together with unique reconfigurability make reconfigurable computing one of the most important technologies of artificial intelligent microchips.

What is reconfigurable computing?

Different from the traditional time domain programming computing mode, reconfigurable computing performs computing on both temporal and spatial programmable architecture. Its connotation and implementation have been evolving with the progress of semiconductor technology and target applications. Field-programmable gate arrays (FPGA), which was born in 1980s, is a typical reconfigurable microchip. It was developed for logic emulation, but soon became widely used devices because its reconfigurability provides the possibility to implement various algorithms. By eliminating the instruction fetch and decode of GPPs, FPGAs are much more energy-efficient than GPPs.

However, because of large amount of the configuration context caused by FPGA's fine-grained architecture and its static reconfiguration mechanism, the computing efficiency and energy efficiency of FPGA are not ideal. For example, its look-up table (LUT) structure results in 95% of the logic used for definition rather than for computation, so that most energy consumption is not directly related to computing. Furthermore, the static programming architecture determines that only when a whole circuit design is loaded into FPGA, can its function be realized. Therefore, a 10 million gates FPGA can only achieve several hundred thousand gates circuit design.

Recently, with the emerging of artificial intelligence (AI), FPGA is used to implement different AI algorithms. However, its low programming efficiency and static reconfiguration characteristics also show that in order to truly realize AI application, especially those terminal side applications that need high energy efficiency and high flexibility, it is necessary to find a new microchip architecture.

Coarse-grained reconfigurable architectures (CGRA) is another way of implementation of the reconfigurable computing concept. Through redundant deployment of computing resources, the arithmetic logic, memory subsystem and interconnection of CGRA can be flexibly customized according to the application requirements, so as to improve the computing and energy efficiency. The emerging dynamic reconfigurable computing technology can realize the real-time dynamic configuration of CGRA according to software (application), which may be considered as an ideal AI microchip architecture.

Why is reconfigurable computing suitable for AI applications?

Modern AI applications, such as computer vision and voice recognition, are based on the computation of artificial neural networks (NNs), which are characterized by complex computation involving massive data, parameters and frequent layer to layer communication. Although AI technology has made great progress, AI algorithms are still evolving, and one artificial NN (algorithm) only adapts to one application, so an ideal AI microchip must be able to adapt to the continuous evolution of algorithms, to support different artificial NNs according to requirements, and to switch between different artificial NNs flexibly. Obviously, by enabling customization in computation pattern, computing architecture and memory hierarchy, microchips based on reconfigurable computing technology might be able to efficiently support different NNs with high-throughput computations and communications. Many researches achieve astonish performance on diverse

NNs by reconfiguring data paths to minimize energy consumption in data movement.

What is the recent progress?

Recently, reconfigurable computing has achieved many remarkable progresses on AI applications accelerations. At first, an optimal NN is always formed of several kinds of layers, such as convolutional and fully-connected layers. In order to achieve end-to-end AI applications, efficient computing must be supported on these layers. Most AI processors designed reconfigurable computing units, instead of independent hardware resources, to support various layers to improve overall performance of the entire networks. Secondly, memory access, especially DRAM access, is the bottleneck of AI acceleration. For example, in AlexNet, to support its 724M MACs, nearly 3000M DRAM accesses will be required, which requires up to 200x energy than one MAC. Four dataflows, including weight stationary, output stationary, no local reuse and row stationary, are proposed to improve data reuse and reduce memory access. Every time a piece of data is moved from an expensive level to a lower cost level in terms of energy, this piece of data should be reused as much as possible to minimize subsequent accesses to the expensive levels, which is the target of the optimized dataflow. The challenge, however, is that the capacity of these low cost memories is limited. Thus different dataflows should be explored to maximize reuse under these constraints. Different from application specific integrated circuits (ASICs) that support specialized processing dataflows, more and more processors proposed to design reconfigurable architectures to dispatch one of four dataflow for different AI applications, which can maximize data reuse and significantly improve overall flexibility. Thirdly, the AI applications are implemented by processors always based on quantization, namely from floating point to fixed point. The ultimate goal is to minimize the error between the reconstructed data from the quantization levels and the original data, and sometimes to reduce the number of operations. The quantization methods can be classified into linear quantization and non-linear quantization. Linear quantization is simpler but can lose more accuracy while non-linear quantization can maintain higher accuracy but is more complex. Meanwhile, as the importance of weights and activations in different layers are various, different methods of quantization can be used for weights and activations, and different layers, filters, and channels in the network. Therefore, reconfigurable computing is more and more attractive to recent researchers to support different quantization methods. Based on the reconfigurable computing, high accuracy and less operations and operands can be achieved. Fourth, ReLU is a popular form of non-linearity activation function used in AI applications that sets all negative values to zero. As a result, the output activations of the feature maps after the ReLU are sparse; for instance, the feature maps in AlexNet have sparsity between 19% to 63%. This sparsity can be exploited for energy, cycle and area savings using compression, prediction and network pruning particularly for off-chip DRAM access which is expensive. Compression methods can skip reading the weights and performing the MAC for zero-valued activations without accuracy loss, but complex control logic is required; prediction methods sacrifices accuracy to reduce operations corresponding to zero-valued activations; The pruning methods is to eliminate the low-valued activations to make network even more sparse, but accuracy can be effected significantly. As these methods performs variously in different networks, some reconfigurable computing architectures are proposed to combine these methods to reduce operations as much as possible with marginal loss. Finally, some methods proposed compact network architectures to reduce the number of weights and computations in AI applications. The main trend is to replace a large filter with a series of smaller filters, which can be applied during network architecture design. As each compact network architectures are designed for specific AI applications, some reconfigurable computing architectures try to support all kinds of compact networks, which can maximally reduce the number of operations and model size for different compact networks in their specific situations with marginal accuracy loss.

Remained challenges and prospects

Up to now, the main research of AI microchips is focused on multilayer perceptual neural networks. The latest development of AI researches require that AI microchips can also accelerate the newly emerging neural networks, such as graphical neural networks and memory networks. Another promising direction is to use artificial intelligence technology to guide the design of reconfigurable computing system. Traditionally, they are designed and programmed using empirical methods. With the increasing complexity of microchips, designers can use AI to better build and manage complex reconfigurable systems.