ERNEST JAMRO

Architektury zaawansowanych układów arytmetycznych w sprzętowych systemach rekonfigurowalnych

Streszczenie

W pracy przedstawiono wyniki koncepcyjne oraz aplikacyjne realizacji wybranych zaawansowanych układów arytmetycznych w rekonfigurowalnych strukturach FPGA. Praca opisuje użycie układów FPGA w wybranych zastosowaniach w systemach wbudowanych oraz komputerach dużej mocy obliczeniowej, pokazuje metody projektowania układów FPGA oraz ich współprojektowania w przypadku użycia procesorów CPU.

Głównym celem niniejszej pracy jest przedstawienie nowatorskich architektur układów arytmetycznych, takich jak sieci neuronowe, mnożenie o skróconej szerokości, akumulacji zmiennoprzecinkowej czy też mnożenia macierzy rzadkich. Przedstawiona architektura równoległo-szeregowa (PSAN) sieci neuronowej umożliwia silną parametryzację struktury sieci, dzięki czemu możliwe jest dobranie poziomu zrównoleglenia i szeregowości obliczeń (optymalizacja zajmowanych zasobów sprzętowych) na podstawie takich wymagań jak, szybkość obliczeń, liczba neuronów wejściowych i wyjściowych w każdej warstwie sieci czy też sposobu wprowadzania i wyprowadzania danych.

Następnym zagadnieniem poruszanym w niniejszym opracowaniu jest mnożenie o skróconej szerokości, które zdecydowanie zmniejsza zajmowane przez układ mnożący zasoby sprzętowe kosztem niewielkiego dodatkowego błędu obliczeń, którego wartość często jest na poziomie błędu zaokrąglenia. Przestudiowano dostępne rozwiązania oraz pokazano, że niektóre z nich są oparte na błędnych założeniach. Podano własne ulepszone rozwiązania.

Układy FPGA stają się alternatywną platformą do prowadzenia obliczeń dużej mocy w stosunku do procesorów CPU. Mnożenie macierzy (czyli głównie operacja mnożenia i akumulacji) jest jedną z podstawowych operacji wykonywanych w tego typu obliczeniach. Zaprezentowano nowe rozwiązania dotyczące akumulatora zmiennoprzecinkowego. Wielki potencjał redukcji zajmowanych zasobów sprzętowych może być związany z prowadzeniem obliczeń z mniejszą szerokością bitową i dodatkową kontrolą błędu obliczeń. Warto podkreślić, że podczas obliczeń zmiennoprzecinkowych to błąd znoszenia może generować największe błędy obliczeń, dlatego jego wykrywanie w procesie akumulacji daje duże perspektywy prowadzenia obliczeń ze zmienną szerokością bitową, domyślnie małą i ewentualnie zwiększaną w przypadku wystąpienia nieakceptowalnego błędu. Podobnie możliwa jest zdecydowana redukcja zajmowanych zasobów sprzętowych w przypadku mnożenia macierzy rzadkich, dla których, jak udowodniono w niniejszej pracy, dominującym zadaniem obliczeniowym jest wyszukiwanie odpowiadających sobie indeksów mnożonych macierzy, czyli operacji wydajnie realizowanych w układach FPGA. Podsumowując, zastosowanie zaproponowanych nowatorskich układów arytmetycznych może zdecydowanie zmniejszyć wymagane zasoby sprzętowe i dzięki temu umożliwić większy stopień zrównoleglenia (szybkości) prowadzonych obliczeń.

ERNEST JAMRO

Advanced arithmetic architectures in reconfigurable hardware systems

Summary

This monograph deals with architectures and design methods of selected advanced arithmetic modules implemented in Field Programmable Gate Arrays (FPGAs). This work approaches areas of High Performance Computing (HPC) and embedded systems, describes FPGA design and hardware-software codesign.

This work describes novel arithmetic architectures including neural networks, reduced width multipliers, floating-point accumulations and sparse matrix multiplications. A novel Parallel Serial Architecture for Neural networks (PSAN) was introduced. This architecture is strongly parameterised thus the parallel and serial levels can be adjusted according to the required computation speed, the number of input and output neurons in each layer or the input / output interfaces.

In this work reduced width multiplier was also approached. This architecture significantly reduces hardware resources by the cost of an extra computation error. This error value may be as low as the rounding error. Available solutions were studied, and it was proved that some of them are based on incorrect assumptions. Consequently, improved solutions were presented.

FPGAs becomes an alternative platform to CPU for High Performance Computing. Matrix multiplication (mainly multiply accumulate operations) is a very fundamental operation for HPC. Therefore a novel architecture for floating-point accumulation was presented. Hardware resources can be significantly saved by reducing computation precision. This can be achieved by controlling the computation error, which main source is often a catastrophic cancellation. Thus detection of cancellation error during floating-point accumulation might allow to carry out main computations in a lower precision and occasionally increase the precision in the case when unacceptable computation error level in detected. Similar reduction of hardware resources can be obtained during sparse matrix multiplication, which, as it was proved in this work, is dominated by search for the matching indices in multiplied matrices. This search operation ideally fits in FPGA structures. Summing up, implementation of the proposed arithmetic architectures may significantly reduced hardware resources. Consequently, the level of parallelism and thus computation speed might be significantly increased.