IP-КОМПОНЕНТ КОНВЕЙЕРНОГО УСКОРИТЕЛЯ СВЁРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ СНК СЕРИИ ZYNQ-7000

Осипов А. С., Петровский Н. А.
Кафедра электронных вычислительных средств[1],
Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: axelosip3345@gmail.com, nick.petrovsky@bsuir.by

В работе представлена аппаратная реализация IP-ядра ускорителя вычисления нейронных сетей на базе ПЛИС типа FPGA серии Zynq-7000. Вычислитель реализован в виде конвейрного процессорного ядра с микропрограммным управлением, поддерживаются операции перемножения матриц и свертки с заданными параметрами. Проведён синтез для кристалла макетной платы Zybo-Z7 сверточных архитектура на основе вещественно значимых и кватернионовых нейронных сетей.

Введение

Свёрточные нейронные сети (CNN) являются фундаментальным компонентом глубокого обучения и лежит в основе многих мультимедийных приложений. Однако высокая вычислительная сложность требует энергоэффективной для носимых устройств. Проектирование ускорителей для FPGA [1] позволяет обеспечить низкую латентность системы, так и оценить энергоэффективность в другом технологическом базисе.

I. Архитектура сопроцессора

Для реализации были выбраны основные типы слоёв: полносвязный, сверточный, max-pooling и flatten для упрощения адресации. Помимо этого предусмотрены отдельные стадии для приема и передачи данных по интерфейсам AXI-Stream.

Для хранения весов используется внешняя DDR-память FPGA из-за большого объёма данных. Ввиду синхронным операциям чтения и записи, а также с целью сокращения критического пути, применена двухстадийная конвейерная архитектура. Первая стадия посвящена формированию управляющих сигналов и запроса на чтение памяти, вторая – вычисление операции MAC (Multiply and Accumulate).

Структурная схема конвейера с внутренними модулями приведена на рисунке 1. Устройство управления отвечает за координацию внешних интерфейсов: формирование адресов и установку управляющих сигналов вычислительному модулю. Для предотвращения конфликтов сигналы задерживаются на один такт при передаче между стадиями.

Вычислительный модуль помимо операции MAC выполняет функцию $max(x_{max},x)$ для слоя *pool-max* и применяет функцию активации *leaky_relu* с фиксированным наклоном 2^{-N} перед сохранением данных в RAM. Память RAM, организованная двухстранично на основе внутрикристального BRAM, хранит входные и промежуточные данные слоя. По завершении вычислений происходит логическая перестановка страниц, а выходные данные с предыдущего слоя становятся входными для следующего. Из-за ограничений BRAM чтение и запись в один такт невозможны, что требует завершить транзакцию записи перед перестановкой страниц. Для этого вводится дополнительная задержка конвейера на такт работы.

Дополнительно модуль включает вспомогательные блоки (*preprocessor* и *conditioner*) для подготовки данных и синхронизации интерфейсов.

II. Устройство управления

Для настройки и контроля сопроцессора используются Control and Status регистры (CSRs). Регистр CS обеспечивает программный сброс, указывает причину исключения и номер текущего этапа.

Группа регистров `config0-15` задаёт структуру нейронной сети: этап (fullycon, conv2d, poolmax, flatten, приём или выдача данных) и его параметры, размерность и объём входных данных. Согласованность с предыдущим слоем определяются автоматически, что позволяет сократить размерность регистров. Некорректная конфигурация сети вызывает исключение и приостанавливает работу сопроцессора до сброса.

III. Интерфейсы IP-компонента

IP-компонент включает один интерфейс AXI-Lite для регистров конфигурации, два интерфейса AXI-Stream для потоковой передачи входных и выходных данных, а также интерфейс доступа к внешней DDR-памяти на основе механизма request–acknowledge.

Стандартный протокол request–acknowledge требует установки запроса и ожидания ответа на последующем такте, что требует два такта работы на чтение данных. Для снижения латентности ме-

[1]Работа выполнена в совместной учебной лаборатории БГУИР-YADRO
`https://www.bsuir.by/ru/kaf-informatiki/yadro`

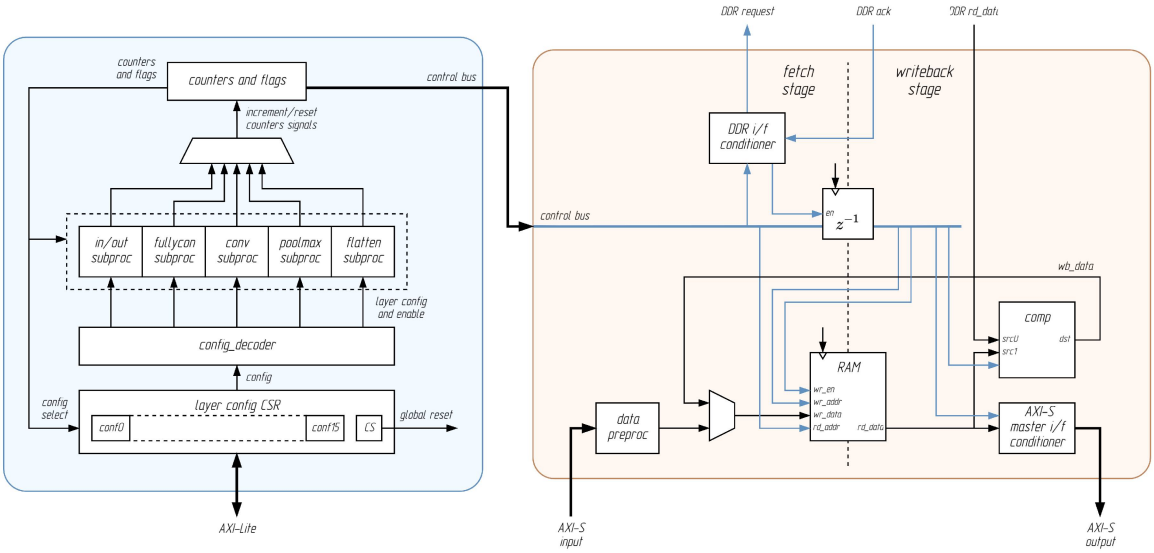

Рис. 1 – Структура сопроцессора

ханизм был модифицирован: сопроцессор может выдать следующий адрес без ожидания ответа от DDR. При своевременном отклике конвейер работает непрерывно, при задержке со стороны DDR – приостанавливается, повторно используя предыдущий адрес. Данный подход обеспечивает непрерывность работы конвейера, однако может возникнуть критический путь при прохождении сигнала acknowledge к IP и обратно.

IV. Результаты синтеза аппаратуры и заключение

В таблице IV приведены результаты синтеза двух конфигураций нейронных сетей: вещественнозначимых (real-valued) и с применением алгебры кватерниов (quaternionic) [2]. Макетирование выполнено для макетной платы Zynq Z7-10. Полученный критический путь проходит через BRAM→умножитель→аккумулятор→BRAM. Для нейронной сети на рисунке 2 расчёт латентности представлен в таблице IV; обработка одного изображения 32 × 32 занимает 77мс при $f_{CLK} = 128$ МГц.

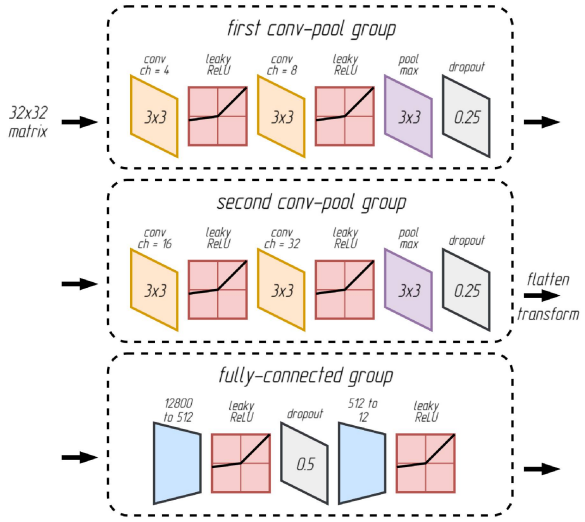

Рис. 2 – Пример структуры нейронной сети

Таблица 1 – Результаты синтеза сопроцессора для кристалла xc7z010clg400-1

Результаты синтеза	
Модель CNN	Quaternionic (4×Q16.13)
Кол-во config CSR	16
RAM size	8 × 32 × 32 words, 64-bit word wide
LUT	2222 / 17600 (12.6%)
FF	1269 / 35200 (3.6%)
BRAM	32 / 60 (53.3%)
DSP	48 / 80 (60%)
f_{CLK}	106 МГц.
Модель CNN	Real-valued (Q16.13)
Кол-во config CSR	16
RAM size	32 × 32 × 32 words, 16-bit word wide
LUT	1289 / 17600 (6.9%)
FF	1236 / 35200 (3.5%)
BRAM	32 / 60 (53.3%)
DSP	1 / 80 (1.3%)
f_{CLK}	128 МГц.

Таблица 2 – Латентность модели Real-valued

Этап	**Latency**	**Этап**	**Latency**
input 32x32	1 024	conv:3x3 ch:4	36 000
conv:3x3 ch:8	232 064	pool:3x3	48 672
conv:3x3 ch:16	672 768	conv:3x3 ch:32	2 245 760
pool:3x3	115 200	flatten	12 800
fc out:512	6 554 112	fc out:12	6 156
output	12	interm.	11
Итого			9 892 179

1. Kastner, R. Parallel Programming for FPGAs [Electronic resource] / R. Kastner, J. Matai, S. Neuendorffer // arXiv e-prints. – 2018. –Mode of access: https://arxiv.org/abs/1805.03648. – Date of access: 24.10.2025.
2. Osipov, A., Petrovsky, N. FPGA Implementation of Quaternionic Fully Connected Neural Network for Image Classification // Pattern Recognition and Information Processing (PRIP'2025): Proc. of the 17th Int. Conf., 16–18 Sept. 2025, Minsk, Belarus. – Minsk : UIIP NASB, 2025. – P. 230-234.