Module
НАУЧНО-ТЕХНИЧЕСКИЙ ЦЕНТР

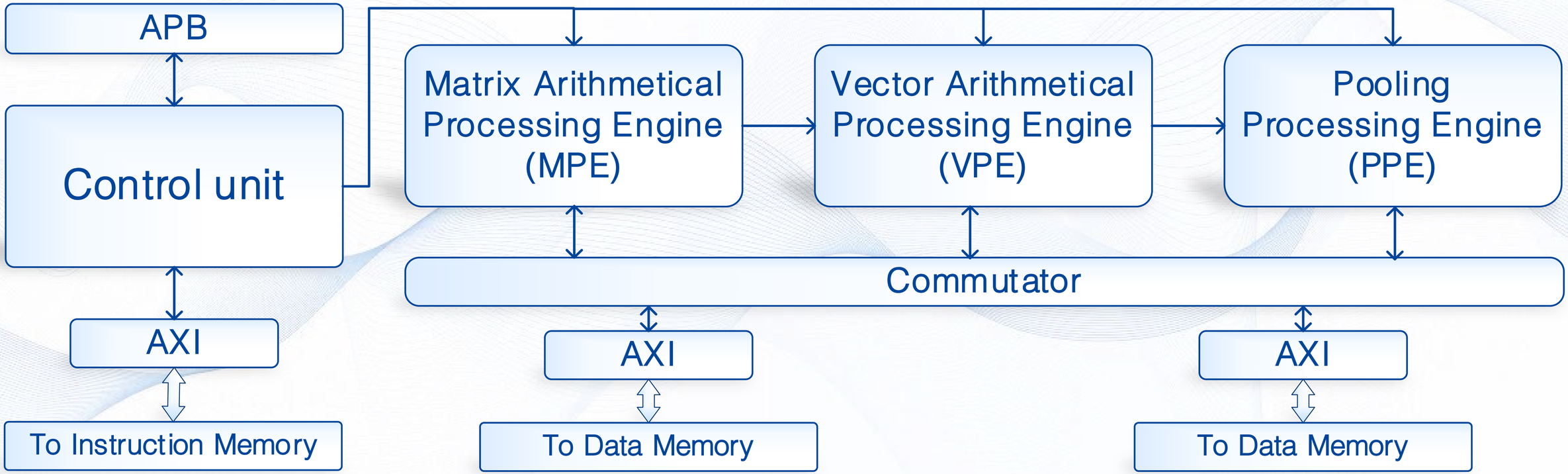
РОССИЙСКИЙ ФОРУМ
МИКРОЭЛЕКТРОНИКА 2023

Программное обеспечение для реализации нейросетевых алгоритмов на сложнофункциональном блоке нейросетевого ускорителя

Черникова Анна Дмитриевна
Главный специалист

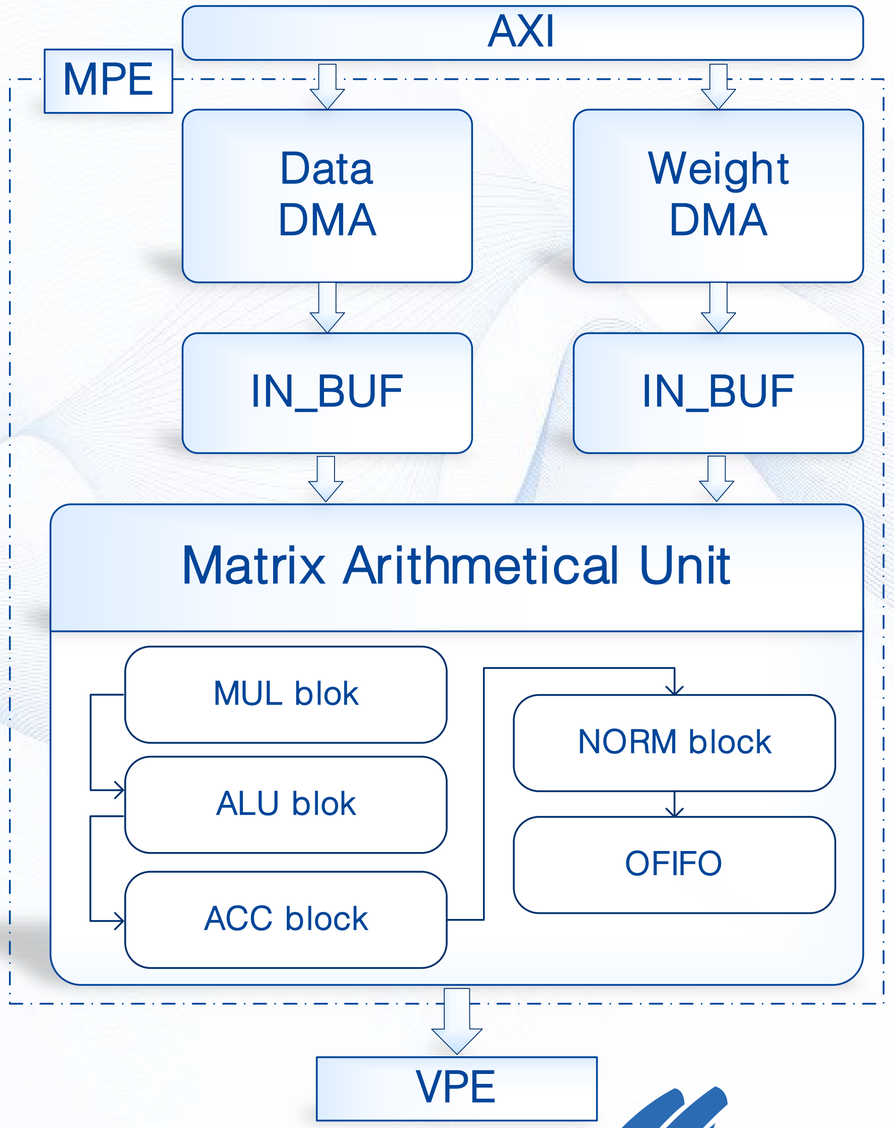
2023

СЛОЖНО-ФУНКЦИОНАЛЬНЫЙ БЛОК НЕЙРОСЕТЕВОГО УСКОРИТЕЛЯ



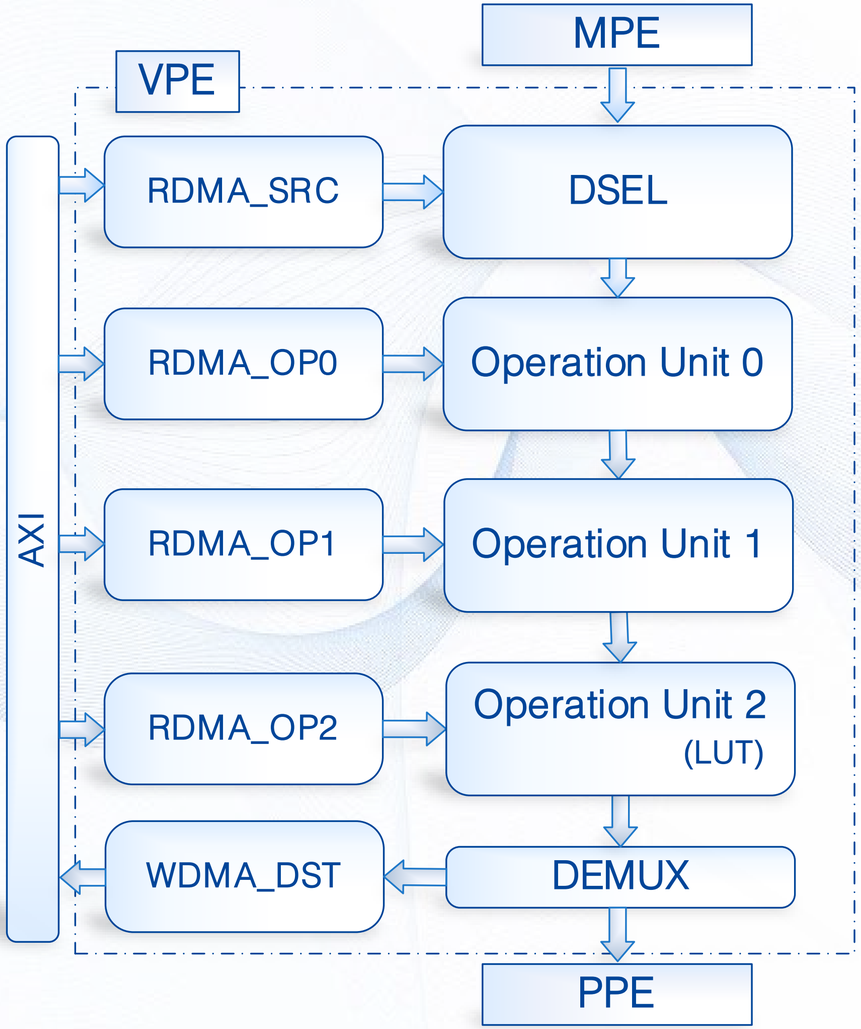
МАТРИЧНЫЙ АРИФМЕТИЧЕСКИЙ БЛОК

- Блоки ПДП обеспечивают подкачку данных в соответствии со всеми требуемыми параметрами свертки
- Блок аккумуляторов обеспечивает хранение частичных сумм до 1024 циклов работы без потери точности на int16/int8
- Есть возможность нормализации данных в типах int16/int8
- Производительность:
  - 1024 MAC/такт для форматов int16/fp16
  - 4096 MAC/такт для формата int8



ВЕКТОРНЫЙ АРИФМЕТИЧЕСКИЙ БЛОК

- Получение данных непосредственно от матричного арифметического блока;
- Возможность поэлементной и поканальной обработки вектора данных;
- Возможность конвертации типов данных;
- Состав операционных узлов;
  - OP0/OP1: ALU/SUB/MAX/MIN; MUL/Prelu; Relu;
  - OP2: MUL/Prelu; ALU/SUB/MAX/MIN; LUT;



БЛОК СУБДИСКРЕТИЗАЦИИ

Ограничения по размерам ядра:

Для типа Fp16:

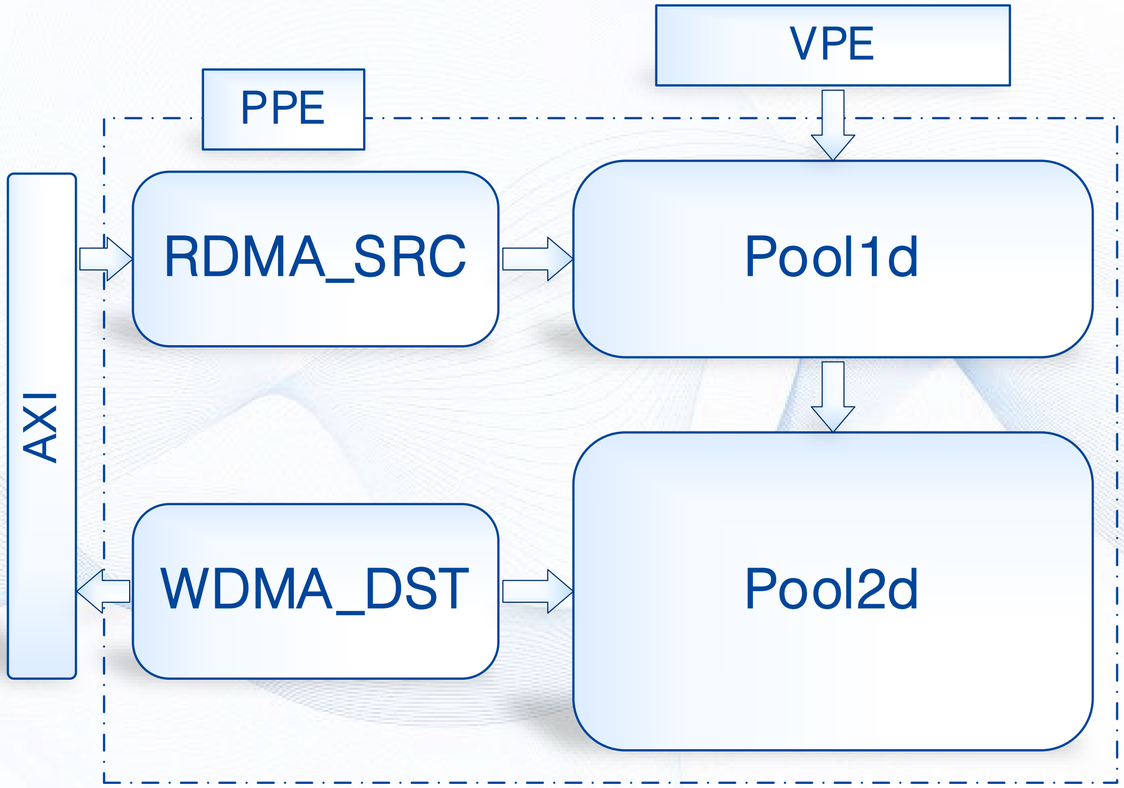
- 4096 по любой из сторон куба

Для типа Int16:

- 64 по горизонтали
- W*H <= 4096

Для типа int8:

- 4096 по любой из сторон куба



ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ

Основные цели:

- Поддержка различных фреймворков
- Поддержка максимального количества нейросетевых операций

- Библиотека TVM является универсальным средством для обработки нейросетевых моделей из различных фреймворков (Tensorflow, PyTorch, Onnx)
- Оптимизации выполняются на уровне графа модели в соответствии с особенностями СФ-блока нейроускорителя
- Библиотека прикладных функций является инструментом для настройки СФ-блока нейросетевого ускорителя
- Компиляция и запуск происходят с учетом архитектуры управляющего ядра

ПОДДЕРЖИВАЕМЫЕ ОПЕРАЦИИ

Сверточные слои матричного блока:

- Convolution 2D
- Convolution 3D
- DepthWise Convolution 2D
- Transpose Convolution 2D

Операции блока субдискретизации:

- Max Pooling 1D, Max Pooling 2D
- Min Pooling 1D, Min Pooling 2D
- Average Pooling 1D, Average Pooling 2D

Операции векторного блока:

- Mul
- Add
- Subtract
- Power
- Transpose
- Strided Slice
- MemCopy

Операции нормализации и активации:

- Batch Normalization
- ReLU/PReLU/LeakyReLU
- Clip
- LUT Activations:
  - Mish,
  - Erf,
  - Sigmoid, Tanh,
  - Exponent,
  - Elu, Selu,
  - Gelu, Silu,
  - SoftPlus,
  - Arctan, Arcctan

Поддержанные нейронные сети :

- SqueezeNet;
- ResNet50;
- Inception v4;
- MobileNet v3;
- Yolo v2/v3;
- VGG16;
- RetinaNet;
- GAN;
- SOD;
- DBNet;
- CRNN;
- DeapSpeach;

ПРОГРАММНЫЕ И АППАРАТНЫЕ ОСОБЕННОСТИ

- Объединение операций в композиты (исключение лишних обращений в память);
- Многократное использование данных из входного буфера матричного блока;
- Пакетный и одиночный режимы обработки данных;
- Псевдо-пакетный режим;
- Склейка коротких каналов (C < 16);

ПРОГРАММНЫЕ И АППАРАТНЫЕ ОСОБЕННОСТИ

- Аппаратная поддержка последовательного и параллельного запусков блоков;
- Организация возможности самостоятельной подкачки команд СФ-блоком нейросетевого ускорителя;
- Предварительная подготовка весовых коэффициентов нейросетевых слоев в соответствии с архитектурными особенностями;
- Модификация графа нейросети для приведения различных табличных функций активации к одной табличной функции;

ЭФФЕКТИВНОСТЬ

Название сети	Кол-во операций	Время, такты	Эффективность*
YOLOv2	$14.7 * 10^9$	$21.8*10^6$	65.9%
YOLOv3	$32.9 * 10^9$	$52 * 10^6$	61.8%
Resnet50	$4.1 * 10^9$	$6.2 * 10^6$	64.5%
SqueeseNet	$0.8 * 10^9$	$1.4 * 10^6$	58.0%

** Эффективность это отношение фактически-полученной производительности к пиковой.*

Module
НАУЧНО-ТЕХНИЧЕСКИЙ ЦЕНТР

РОССИЙСКИЙ ФОРУМ
МИКРОЭЛЕКТРОНИКА 2023

Спасибо за внимание!

www.module.ru

2023