XILINX®

Макро Групп – авторизованный дистрибьютор Xilinx

МАКРО ГРУПП

Портирование свёрточных нейронных сетей на платформу Xilinx Zynq Ultrascale Plus и ускорение их работы

Владимир Викулин, инженер по применению Xilinx
Николай Щебнев, инженер по нейросетям департамента RnD

29.09.2021 г.

План вебинара

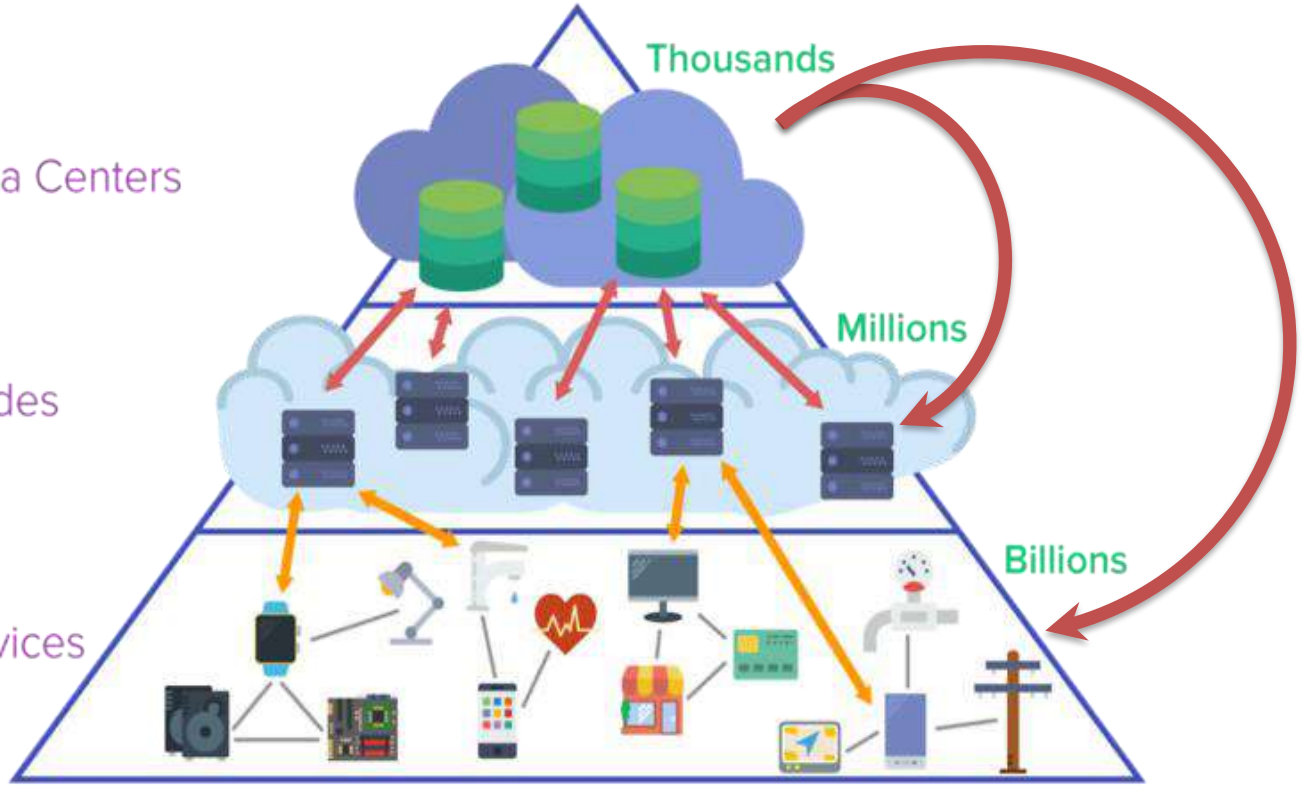
- Миграция ИИ на конечные устройства (EDGE AI) – современная тенденция развития глобальной IT инфраструктуры
- Сверточные нейросети и платформы Xilinx для их портирования
- Инструментальные средства для портирования сверточных сетей на платформы Xilinx
- Маршрут портирования, ключевые этапы маршрута
- Сквозной пример – от описания сети до демонстрации ее работы на плате ZCU104
- Сравнение скорости работы сверточной нейросети на разных платформах при разной степени ее оптимизации.
- Рекомендации по портированию нейросетей на платформы Xilinx
- Ответы на вопросы

1. Миграция ИИ на конечные устройства

1) **CLOUD** I Data Centers

2) **FOG** I Nodes

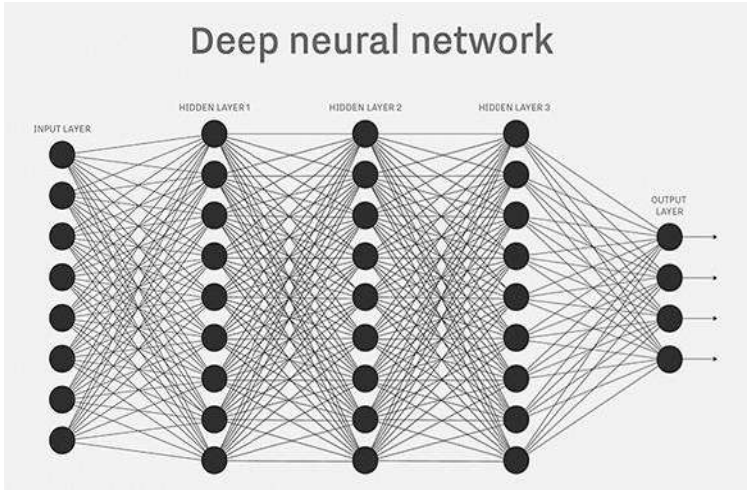
3) **EDGE** I Devices



2. Сверточные нейросети и платформы Xilinx для их имплементации

Общая структура сверточной Н.С.

Входной слой (напр. цветное изображение)



Выходной слой – вектор, в котором, например, для задачи классификации, каждое значение соответствует вероятности принадлежности изображения к определенному классу

Внутренние слои и связи
(Благодаря отсутствию обратной связи можно обеспечить детерминированное время отклика)

2. Сверточные нейросети и платформы Xilinx для их имплементации (2)

Аппаратные платформы Xilinx

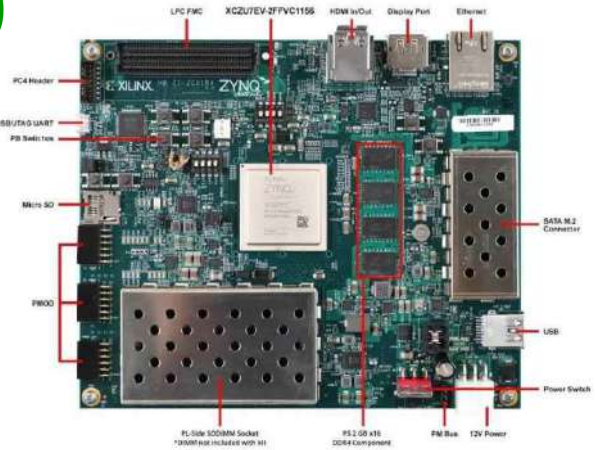
- Реализуются на СнК либо на ускорителях Xilinx
- Содержат хотя бы одно IP-ядро “искусственного интеллекта” (DPU)
- Могут быть “стандартными” либо “кастомными”

Edge (СнК)	Ускорители	Тип DPU
	Alveo	Soft
Zynq7000		Soft
Zynq US+		Soft
Versal AI		Hard

На этом вебинаре мы будем работать с платформой на базе платы ZCU104

2. Сверточные нейросети и платформы Xilinx для их имплементации (3)

Пример “стандартной” платформы: ZCU104



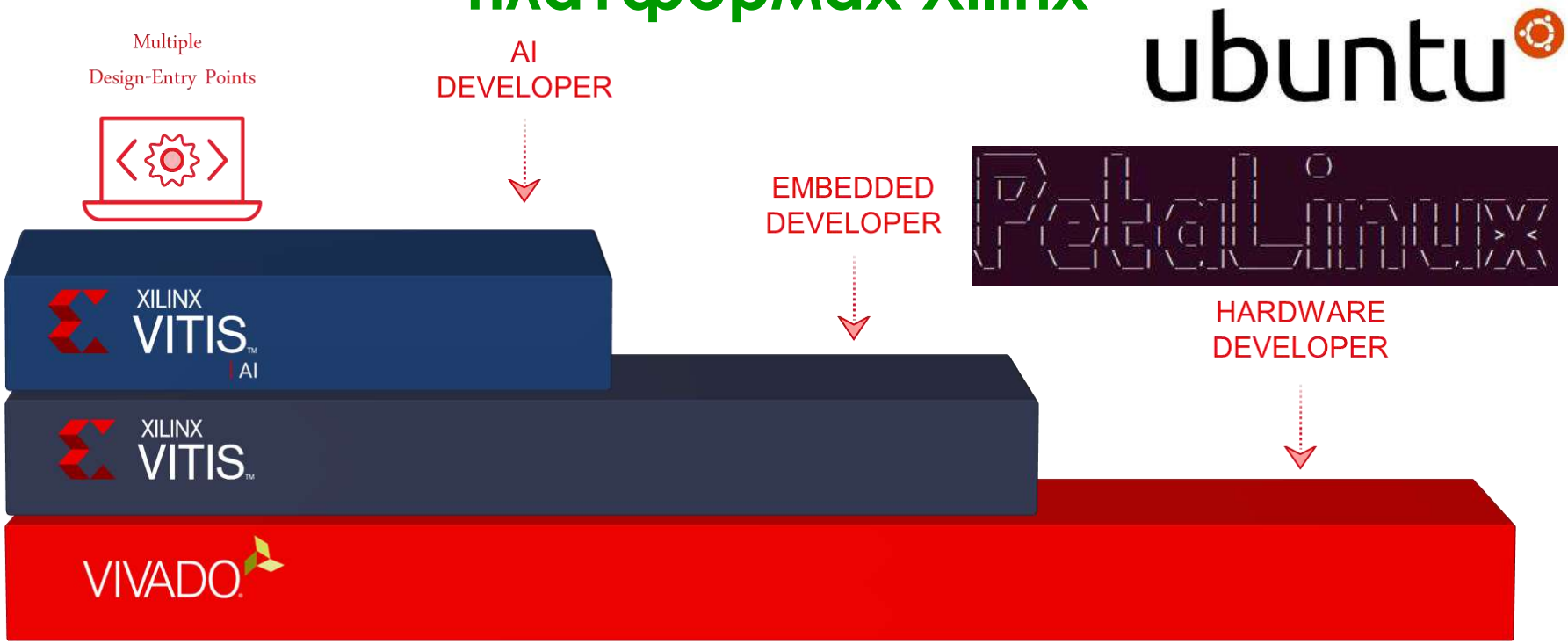
Платформа – это:

- Аппаратура, например отладочная плата ZCU104
- Конфигурация СнК или ПЛИС, должна включать имплементацию хотя бы одного ядра DPU
- Программное обеспечение, включая OS Linux

Для EDGE-платформ вся конфигурационная информация и ПО обычно записывается на SD-карточку

SD-Образ платформы для ZCU104 скачивается отсюда (требуется регистрация):
https://www.xilinx.com/member/forms/download/design-license-xef.html?filename=xilinx-zcu104-dpu-v2021.1-v1.4.0.img.gz

3. Инструментальные средства для имплементации сверточных сетей на платформах Xilinx



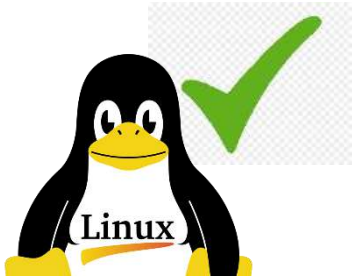
3. Инструментальные средства для имплементации сверточных сетей на платформах Xilinx (2)

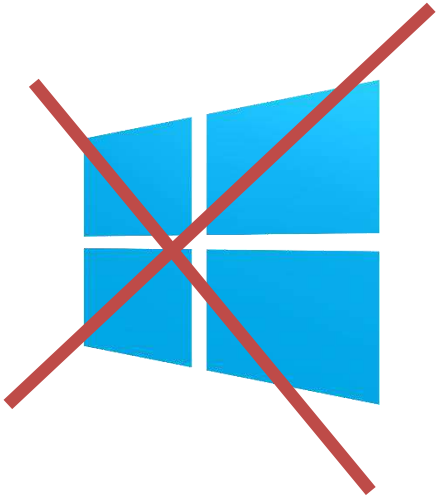
Что необходимо для работы

1. Компьютер под Linux (напр. Ubuntu 20.04 LTS)
   - Графические акселераторы Nvidia (для обучения)
2. Стандартный набор средств Xilinx: Vitis, Vivado, Petalinux
3. Vitis AI
4. Docker (часто входит в Ubuntu)
5. Дополнительно: среда для работы с сорцами программ

i) Версии ПО должны быть согласованы

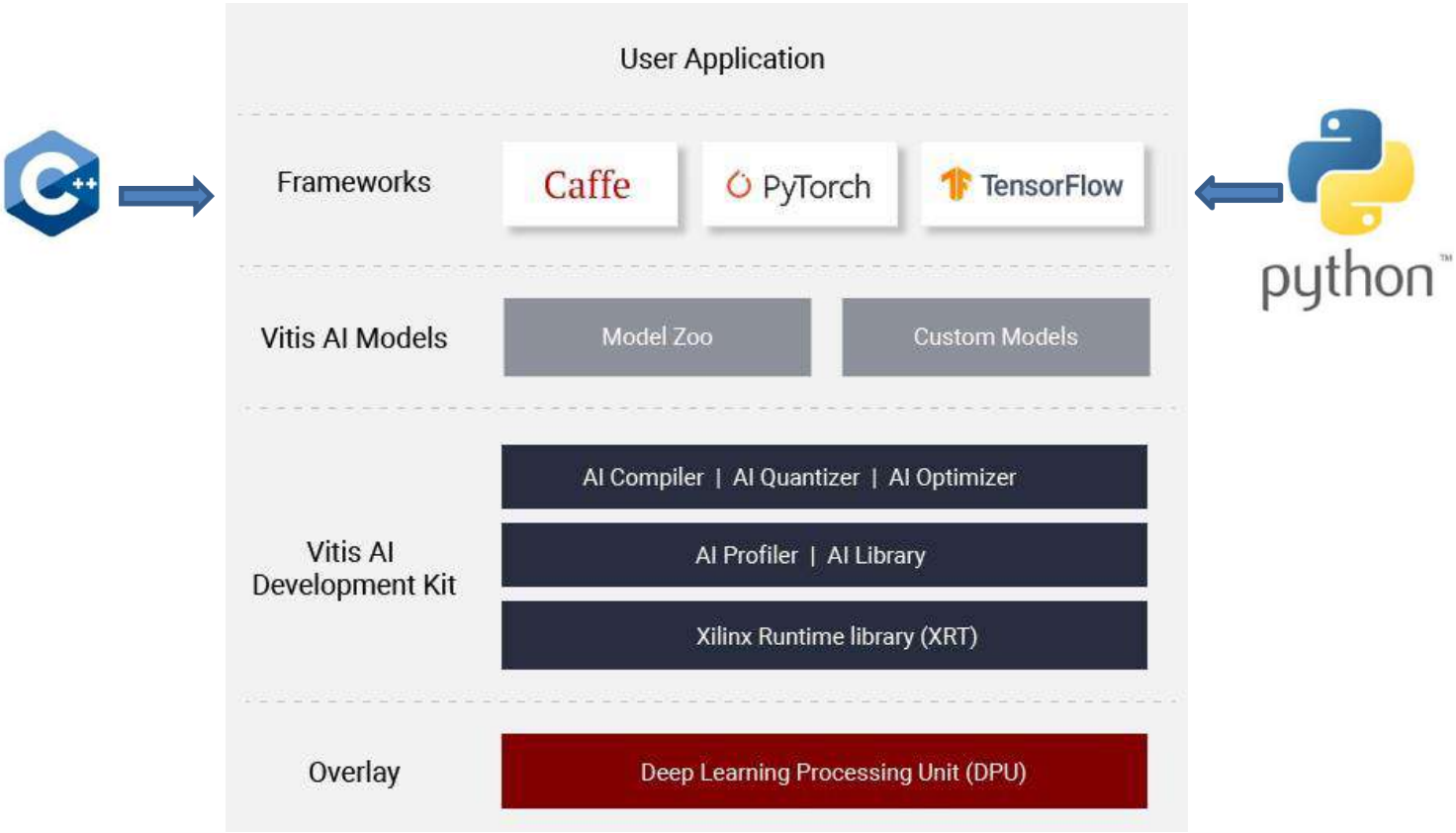
ii) Пакет Vitis AI должен быть не только загружен, но и правильно настроен (set up)





3. Инструментальные средства для имплементации сверточных сетей на платформах Xilinx (3)

Структура Vitis AI



4. Инструментальные средства для имплементации сверточных сетей на платформах Xilinx (4)

Конфигурирование Vitis AI

- Vitis AI сконфигурирован в Docker-контейнере
  - Предварительно сконфигурированная версия для CPU здесь: https://hub.docker.com/r/xilinx/vitis-ai/tags?page=1&ordering=last_updated
  - Версию для GPU требуется собирать с помощью скриптов: https://github.com/Xilinx/Vitis-AI/tree/master/setup/docker

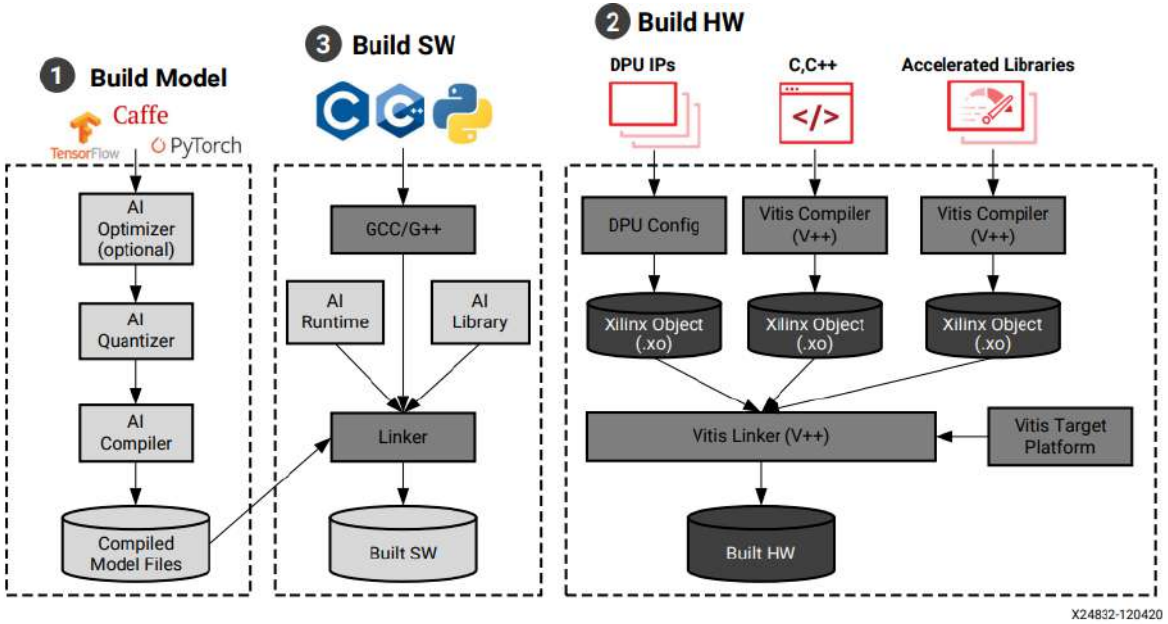
Дополнительная информация:

- Стартовая страница: https://github.com/Xilinx/Vitis-AI
- Руководство: https://www.xilinx.com/html_docs/vitis_ai/1_4/index.html
- Форум: https://forums.xilinx.com/t5/AI-and-Vitis-AI/bd-p/AI

4. Маршрут проектирования, ключевые этапы маршрута

Маршрут проектирования

1. Создание НС
2. Обучение НС
3. Оптимизация НС
4. Кросс-компиляция НС под платформу
5. Написание программы для работы с НС



Работа выполняется на Python (Tensorflow, Pytorch) или на C++ (Caffe).
Представление НС:

- Protobuf (.pb, .h5 –Tensorflow, .pth – Pytorch, .prototxt/.caffemodel – Caffe)
- Скомпилированная под платформу Xilinx НС: .xmodel

4. Маршрут проектирования, ключевые этапы маршрута (2)

Создание НС

- Послойное описание сети на языке Python с помощью средств фреймворков Caffe, PyTorch и TensorFlow

- Использование сторонних источников, где НС уже имплементирована в одном из фреймворков. Важно поддерживать совместимость версий, для Vitis AI 1.4:
  Pytorch 1.5-1.7.1,
  Tensorflow 1.15,
  Tensorflow 2.3

4. Маршрут проектирования, ключевые этапы маршрута (3)

Обучение НС

1. Подготовка наборов данных: обучающего, контрольного и валидационный
2. Написание обучающего скрипта с помощью средств фреймворков Caffe, PyTorch и TensorFlow
3. Обучение сети
4. Оценка качества работы обученной НС
5. Сохранение графа и весов НС

i) При обучении структура сети не меняется, а меняются только значения коэффициентов
ii) Коэффициенты как правило представляются в формате fp32
iii) Методы обучения весьма разнообразны. Наиболее распространённый – обучение с учителем
iv) Окончание обучения производится по достижению некоторого значения целевой метрики, например, для задачи классификации: точность (accuracy), полнота (recall), точность (precision)

4. Маршрут проектирования, ключевые этапы маршрута (4)

Оптимизация НС

1. Прореживание (Vitis AI Optimizer)
   - Цель: сократить количество вычислений путем удаления структур (фильтров свертки) либо параметров, слабо влияющих на результат
   - Бывает структурное и неструктурное (в Vitis представлен структурный)
   - Требуется дообучение
   - Утилита от Xilinx платная. Цена лицензии - $50K. Эффективность: уменьшение размера графа до 90%
   - Можно произвести частичную оптимизацию с помощью сторонних средств

2. Квантизация (Vitis AI Quantizer)
   - Уменьшение разрядности весов модели (fp32 -> int8)
   - Может выполняться с дообучением либо без
   - Бесплатный

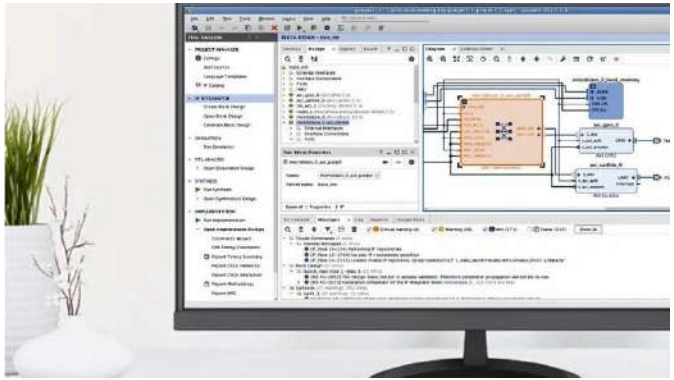
4. Маршрут проектирования, ключевые этапы маршрута (5)

Кросс-комплиция и написание программы для работы с НС под платформу

1. Компиляция сети
   - Обученная (и оптимизированная) НС конвертируется в формат Vitis AI (.xmodel), производится оптимизация графа, создаются инструкции для выполнения на DPU
   - Создание приложения
   - Выполняется на хосте с использованием вызовов из Xilinx Runtime Libraries
   - Собирается: a) в Vitis b) В командной строке c) При сборке Linux
2. Инициализация платформы
   - Загрузка с SD-карты
3. Интеграция с платформой
   - Пересылка каталога и приложения на целевую платформу (либо добавление их сразу в сборку Linux)
4. Выполнение приложения

4. Маршрут проектирования, ключевые этапы маршрута (6)

Кастомные платформы



Когда нужны кастомные платформы?

- Когда вы разворачиваете сети на собственном оборудовании
- Когда на одной платформе развернуты несколько сетей
- Когда используется нестандартная периферия или источники данных

Как создать кастомную платформу?

- Рекомендуем начинать не с нуля, а взять за основу одну из “стандартных” платформ
- Модифицировать проект в Vivado и имплементировать его
- Собрать кастомную версию Linux с помощью Petalinux

5. Сравнение эффективности работы различных вариантов реализации сверточной нейросети с приведением численных характеристик производительности

Сеть AlexNet (2012 г.) для задач классификации, содержит ~62 млн параметров
Характеристики сети:

	Оригинальная сеть, FPS		Pruned (80%), FPS	
	Без пред-обработки	Без пред-обработки	С пред-обработкой	Без пред-обработки
CPU (AMD Ryzen 9 3900XT)	72	72	192	183
GPU (RTX 2070 SUPER)	526	514	701	677
DPU (ZCU104), квантованная		213.4		564

6. Сквозной пример – от описания сети до демонстрации ее работы на плате ZCU104

7. Рекомендации по работе с нейросетями и платформами Xilinx от RnD Макро

Рекомендации

1. Начать с самого простого – имплементация готовой сети из ModelZoo на стандартной платформе Xilinx
2. Использовать стандартные конфигурации для разработки
3. Работать по инструкции и не торопиться
4. Привлечь специалистов требуемой квалификации и в необходимом количестве, провести их обучение
5. Вовремя консультироваться
6. Тщательно выбирать целевые платформы (Макро поможет)

Спасибо за внимание !

Ваши вопросы

Наши ответы

Официальный дилер Xilinx

Контакты

Тел.: 8 (800) 333-06-05
email: SALES@MACROGROUP.RU
Продукция Xilinx и техподдержка: fpga@macrogroup.ru
Силовая электроника: power@macrogroup.ru

Олег Болихов – руководитель направления “Цифровая электроника”
Владимир Викулин, Дмитрий Шадрин – техподдержка Xilinx
Сергей Салмин – руководитель департамента RnD

Ссылки

Страница TrenZ Electronic: https://www.trenz-electronic.de/en/

TrenZ Electronic EDDP: https://wiki.trenz-electronic.de/display/PD/EDDP+Resources

Совместная программа обучения

- Технологии Xilinx сложны и разнообразны, поэтому квалификация разработчиков имеет решающее значение для успешного выполнения проектов
- Обучение по программам и стандартам лидера европейского обучения PLC2 – авторизованного партнера тренинг-партнера Xilinx (XTP)
- На русском языке
- Сертификат PLC2 + Макро

Xilinx – полезные ссылки

- Сайт Xilinx: https://www.xilinx.com/
- Сайт Developer: https://developer.xilinx.com/
- Форум: https://forums.xilinx.com/
- Обучение: https://xilinxprod-catalog.netexam.com/
- Репозиторий: https://github.com/Xilinx
- Отладочные платы и платформы: https://www.xilinx.com/products/boards-and-kits/see-all-evaluation-boards.html

Xilinx – материалы с сайта

Справочные и методологические материалы на сайте Xilinx

- DocNav – все в одном месте на вашем компьютере
- Selection guides – руководства по выбору
- User guides – руководства по применению
- UFDM – методология проектирования
- AR – ответы на вопросы

Xilinx – как получить техподдержку

Техподдержка

- Сначала посмотреть на форуме Xilinx
- Поискать на ресурсах Xilinx: https://www.xilinx.com/support.html
- Обратиться в Macro: fpga@macrogroup.ru
- Открыть service request: https://service.xilinx.com/sservice_prod/start.swe

Xilinx – на чем разрабатывать и отлаживать свои проекты

Отладочные платы и платформы

Подберем отладку под ваши задачи

- https://www.macrogroup.ru/catalog/partgroup/378
- https://www.xilinx.com/products/boards-and-kits.html