XILINX®

Макро Групп – авторизованный дистрибьютор Xilinx

МАКРО ГРУПП

Многократное ускорение: новые продукты и средства проектирования от Xilinx

- ПЛИС, СнК, платформы и средства разработки
- Новые подходы и технологии проектирования

Владимир Викулин,
Инженер по применению Xilinx

02.2020

План семинара

Часть 2

- Обзор линейки ускорителей Alveo и области их применения.
- Облачные сервисы
- Дата-центры: вычисление, хранение, передача данных.
- Нейронные сети и машинное обучение, методы повышения производительности нейронных сетей.
- Vitis AI для работы с нейронными сетями на платформе Xilinx
- Заключительная часть
  - Демонстрация: Простейшая программа для Alveo в Vitis
  - Вопросы-ответы

Ускорители ALVEO
Для серверных/облачных вычислений

Обзор платформы Alveo

Ускорители ALVEO
Для серверных/облачных вычислений

- На базе чипов Virtex UltraScale+
- Высокоуровневое проектирование в vitis
- Низкоуровневое проектирование в Vivado
- Можно комбинировать высокоуровневое и низкоуровневое проектирование
- Разработка в облаке (AWS, Nimbix)

Ускорители ALVEO
Для серверных/облачных вычислений

- Alveo U50
- Alveo U200
- Alveo U250
- Alveo U280

Решаемые задачи:

- Ускорение вычислений
- ИИ, нейросети
- Обработка big data
- Транскодирование видео на лету
- …

https://www.xilinx.com/support/documentation/selection-guides/alveo-product-selection-guide.pdf

Ускорители ALVEO

	Product Name	Alveo U200	Alveo U250	Alveo U280	Alveo U50
Dimensions	Width	Dual Slot	Dual Slot	Dual Slot	Single Slot
	Form Factor, Passive Form Factor, Active	Full Height, ¾ Length Full Height, Full Length	Full Height, ¾ Length Full Height, Full Length	Full Height, ¾ Length Full Height, Full Length	Half Height, ½ Length
Logic Resources [1]	Look-Up Tables	1,182K	1,728K	1,304K	872K
	Registers	2,364K	3,456K	2,607K	1,743K
	DSP Slices	6,840	12,288	9,024	5,952
DRAM Memory	DDR Format	4x 16GB 72b DIMM DDR4	4x 16GB 72b DIMM DDR4	2x 16GB 72b DIMM DDR4	–
	DDR Total Capacity	64GB	64GB	32GB	–
	DDR Max Data Rate	2400MT/s	2400MT/s	2400MT/s	–
	DDR Total Bandwidth	77GB/s	77GB/s	38GB/s	–
	HBM2 Total Capacity	–	–	8GB	8GB
	HBM2 Total Bandwidth	–	–	460GB/s	316GB/s[4]
Internal SRAM	Total Capacity	43MB	57MB	43MB	28MB
	Total Bandwidth	37TB/s	47TB/s	35TB/s	24TB/s
Interfaces	PCI Express®	Gen3 x16	Gen3 x16	Gen3 x16, 2xGen4 x8, CCIX	Gen3 x16, 2xGen4 x8, CCIX
	Network Interface	2x QSFP28	2x QSFP28	2x QSFP28	U50[2] - 1x QSFP28 U50DD[3] - 2x SFP-DD
Power and Thermal	Thermal Cooling	Passive, Active	Passive, Active	Passive, Active	Passive
	Typical Power	100W	110W	100W	50W
	Maximum Power	225W	225W	225W	75W
Time Stamp	Clock Precision	–	–	–	IEEE Std 1588
Compute Performance	INT8 TOPs	18.6	33.3	24.5	16.2
	Machine Learning	Machine Learning Solution Brief			
	Acceleration Applications	Acceleration Application Solutions			

Alveo™ Data Center Accelerator Cards

Notes
1. Logic resources shown without shell usage; refer to card user guides for shell resource usage
2. U50 is a production qualified card for volume deployment
3. U50DD is an engineering sample card not for volume production
4. For A-U50DD-P00G-ES3-G and A-U50-P00G-PQ-G measured 316 GB/s peak HBM2 bandwidth, 201 GB/s nominal

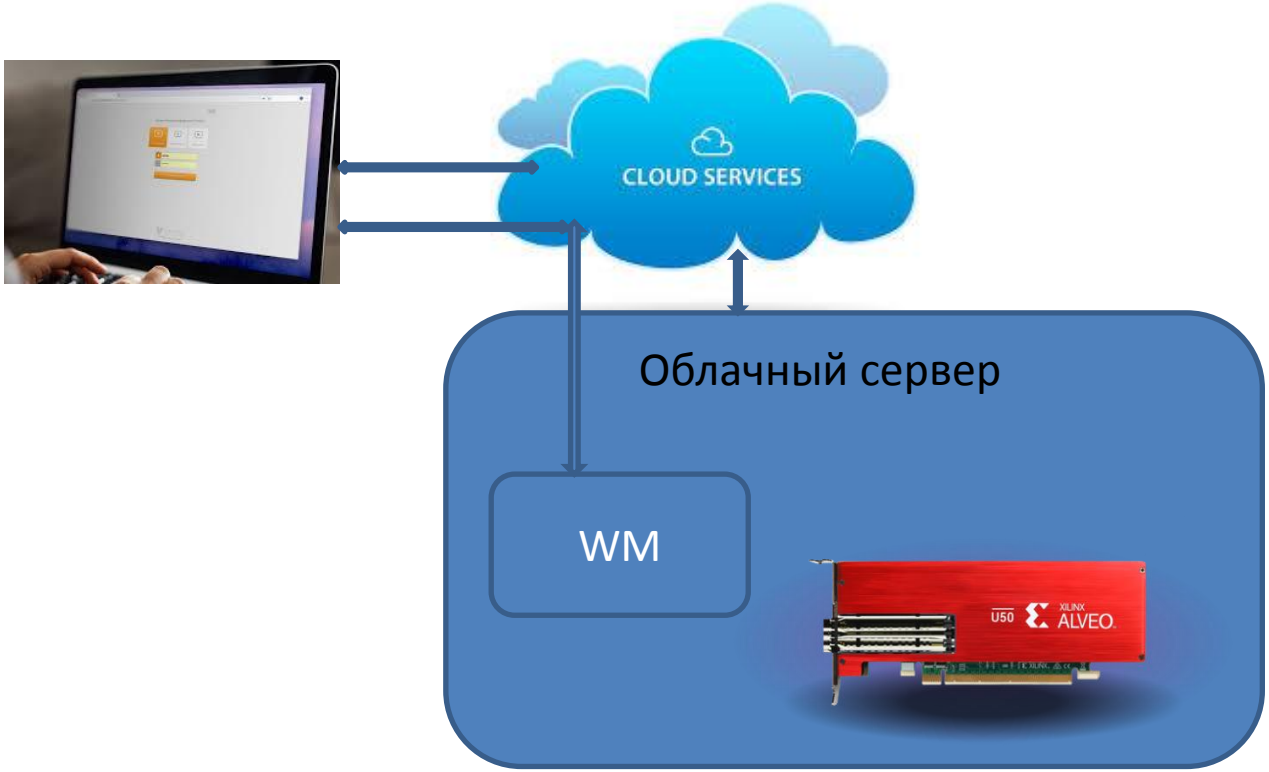
Page 2

XILINX.

Alveo
Удаленная разработка в облаке

Как это устроено

1. Управление через Web-интерфейс
2. Доступ к виртуальной машине
3. Удаленная отладка на реальном оборудовании



Alveo
Удаленная разработка в облаке

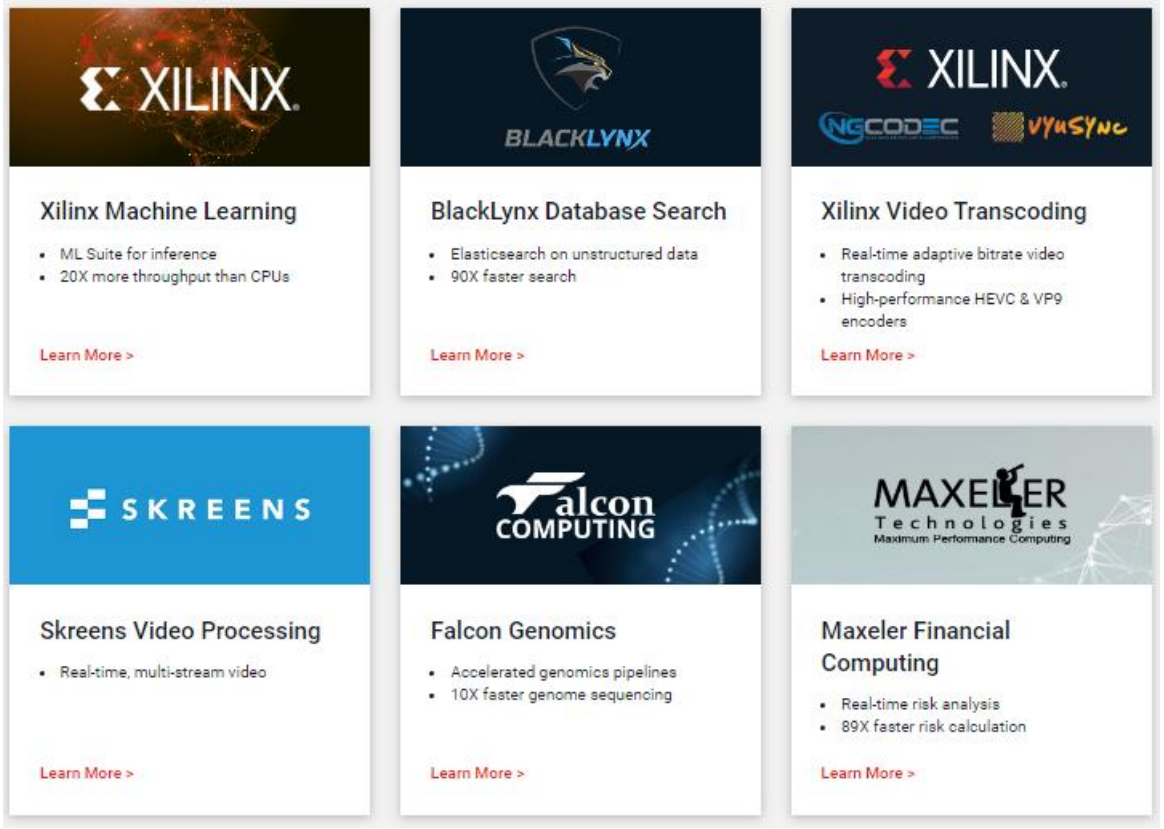
Преимущества облачных технологий за небольшую плату

- Все средства разработки уже установлены и правильно сконфигурированы
- Собственные виртуальные машины на мощных серверах
- Доступ к реальным отладочным платам
- Обмен данными с системой клиента
- Доступ к предустановленным приложениям

Облачные сервисы Amazon и Nimbix
AWS - https://www.xilinx.com/products/design-tools/acceleration-zone/aws.html
Nimbix - https://www.xilinx.com/xilinxtraining/assessments/portal/alveo/intro_nimbix_cloud/story.html

ALVEO в датацентрах

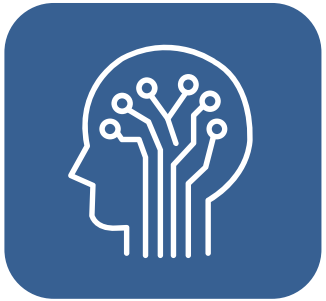
Экосистема приложений для облачных вычислений и датацентров

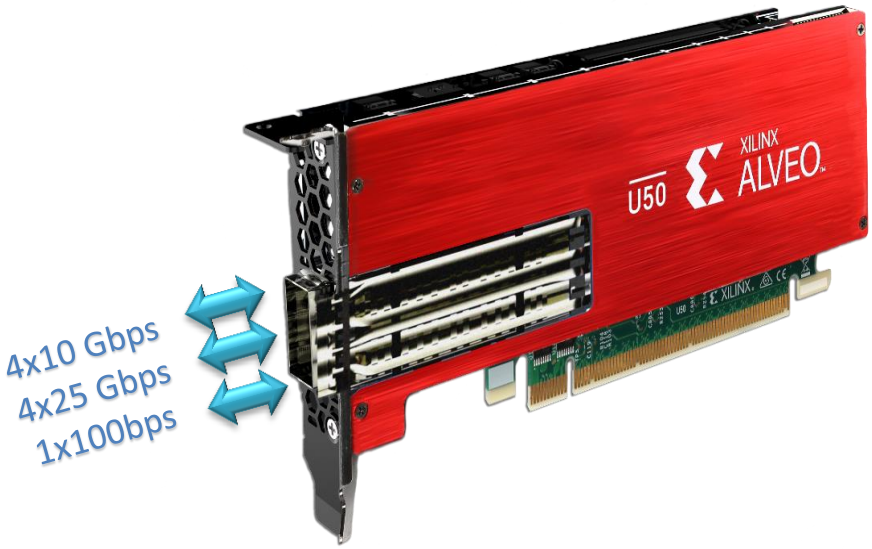


ALVEO в датацентрах

Fintech

Machine Learning





Storage

Database

Датацентры - пример

Транскодирование видео

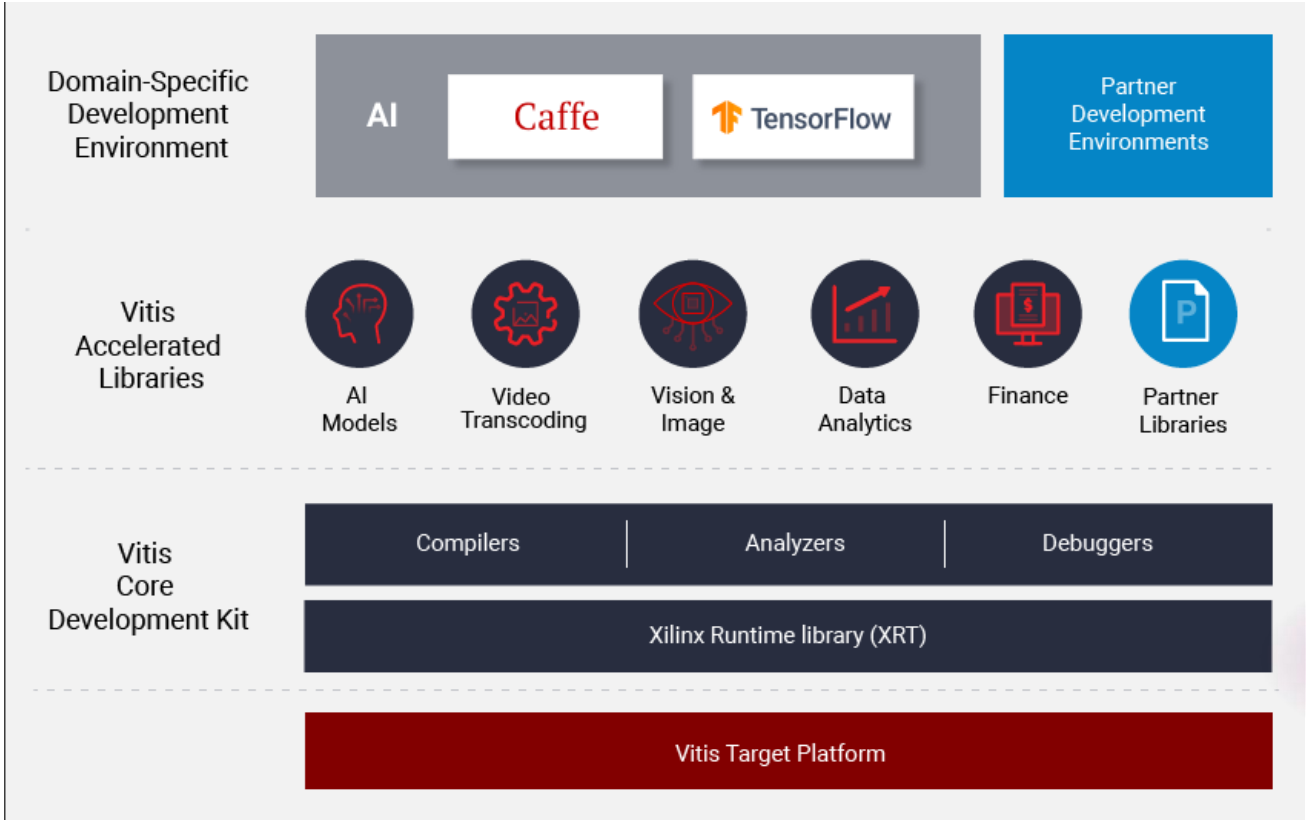
40x Xeon Gold
H.265 very-high quality
20x 1080p30

Alveo U50 HEVC Video Compression

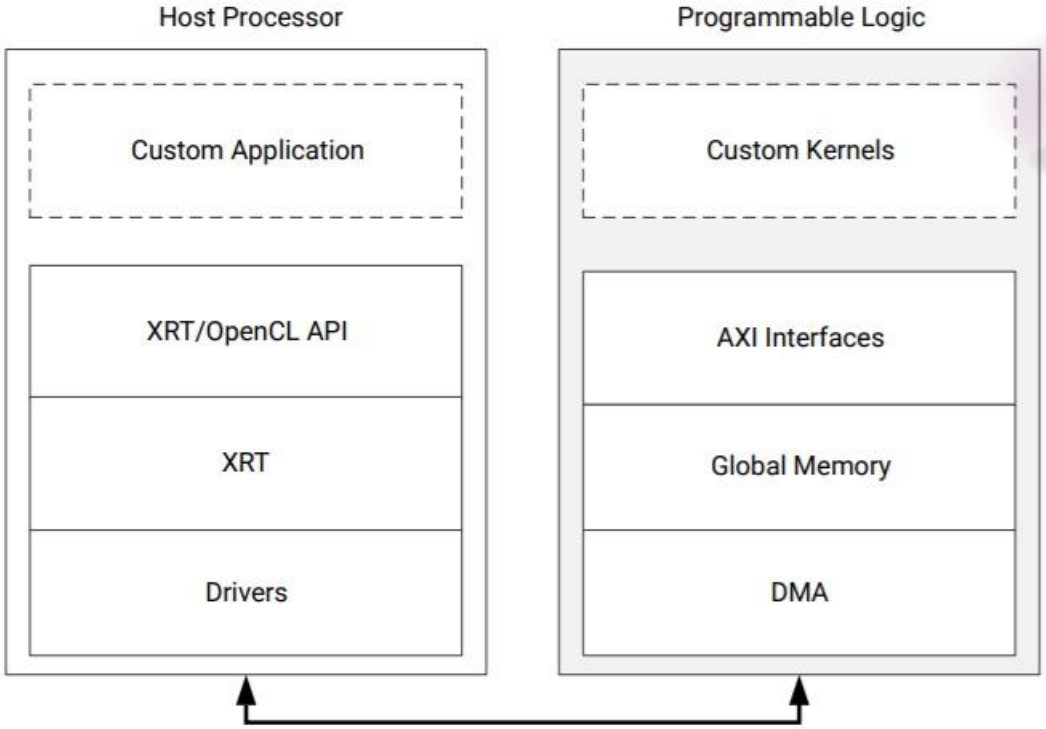
20x Throughput Per Node
8x Lower HW Cost
23x Lower Power Cost

5x Alveo U50
NGCodec HEVC Very-High Quality
20x 1080p30

Vitis для работы с Alveo

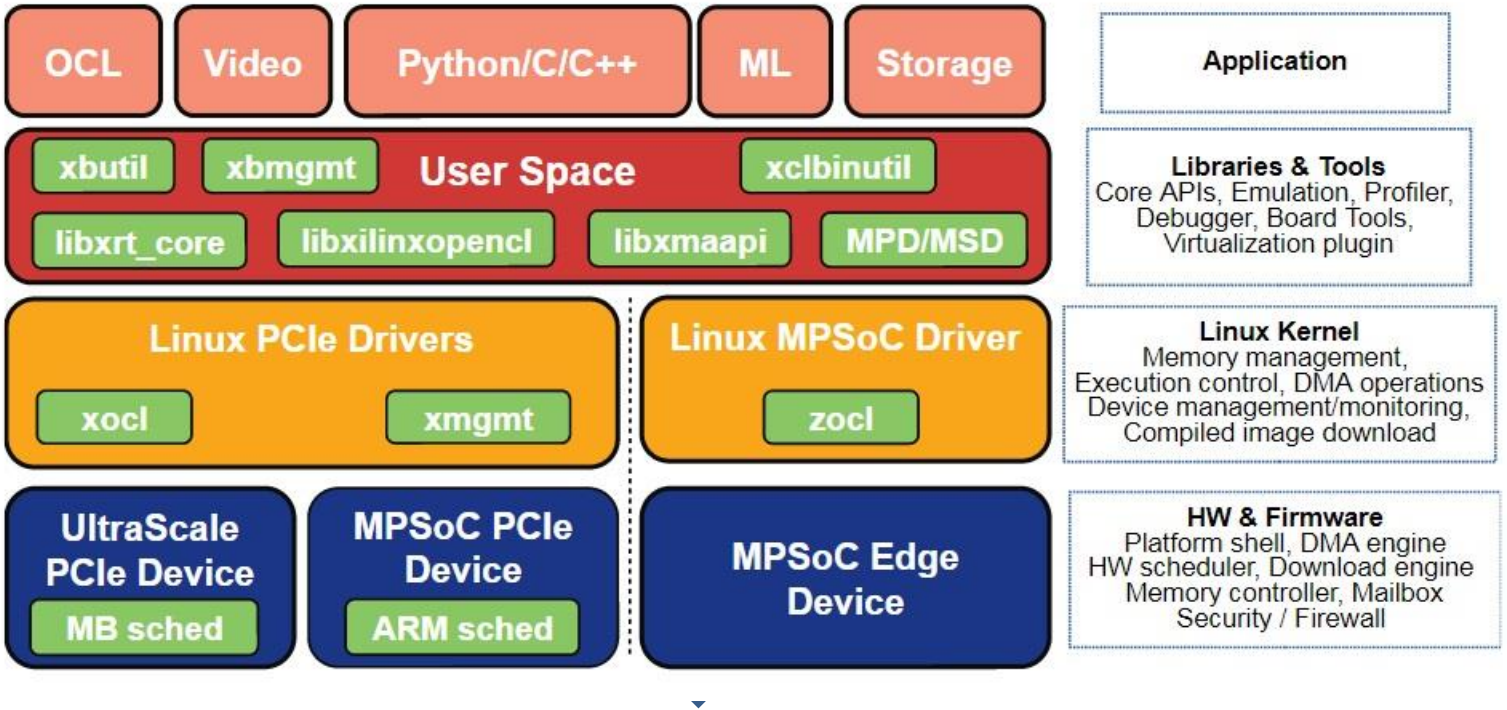


Vitis – клиент-серверная модель



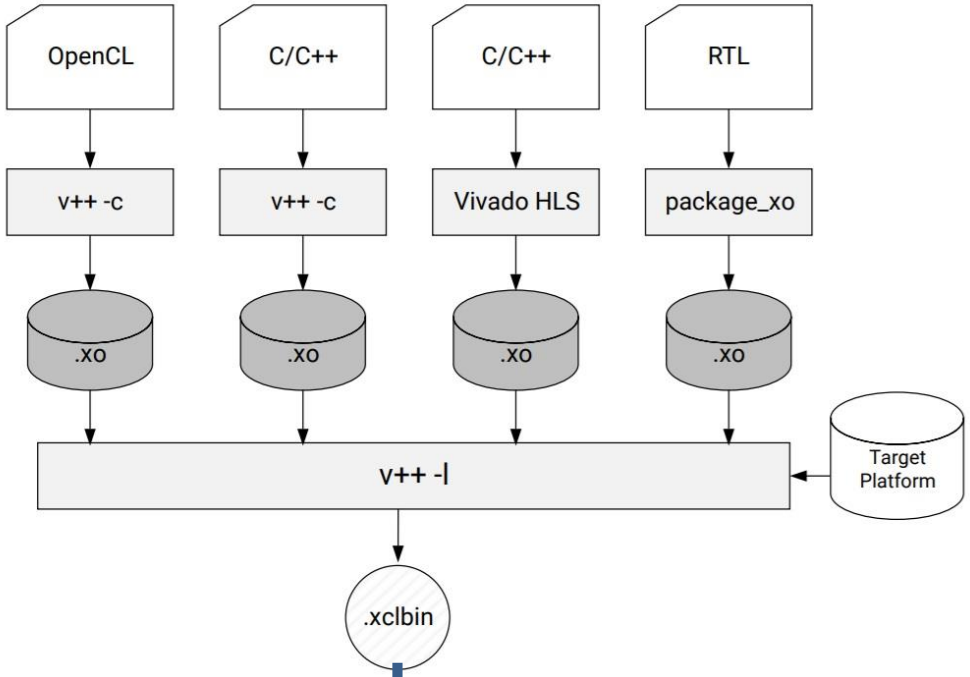
UG1393 - Документация

Vitis – XRT



https://github.com/Xilinx/XRT

Vitis – программирование ПЛИС

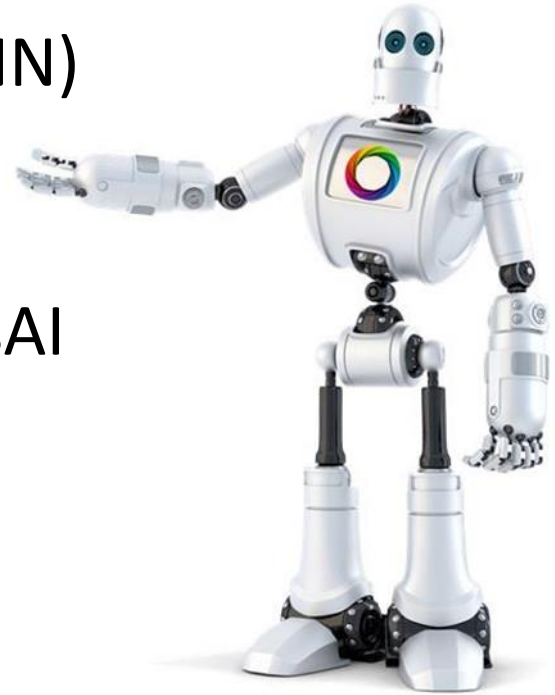


(Alveo U50 Platform)

UG1393 - Документация

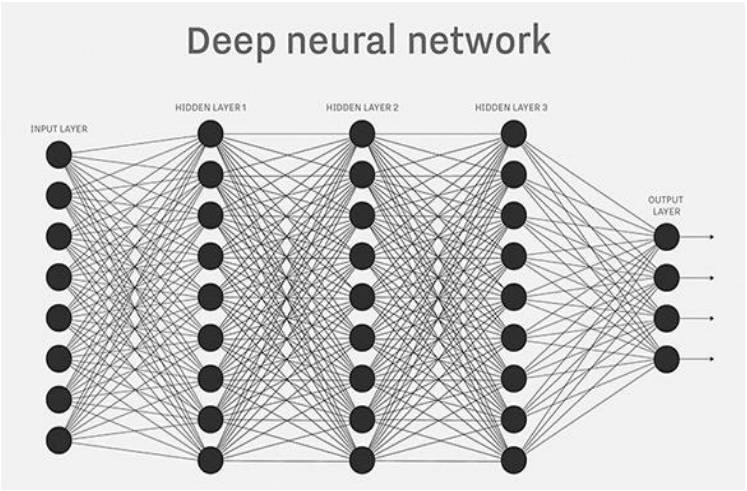
Нейронные сети

- Сверточные нейронные сети (CNN)
  - Обучение (Deep Learning)
  - Вычисление (Inference)
- Разработка CNN с помощью VitisAI



Структура CNN

Входной слой (напр. двумерное изображение)



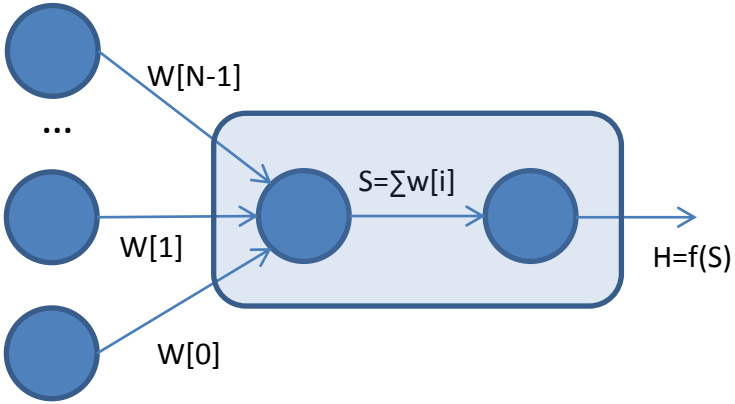
Выходной слой (каждый кружок – возможный вариант)
Напр.:

- Кошечка
- Собачка
- Человек
- Неведома зверюшка
- Все остальное

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

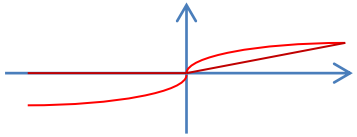
Перцептрон

Основной структурный элемент Н.С. - перцептрон (аналог нейрона)



Выход H

1. f(S) – нелинейная функция
2. Выход **H** нормирован в диапазоне [-1..+1]



Вычисление CNN

Вычисление CNN производится на специализированных Soft либо Hard ядрах:

- DPU (разные для Alveo и Zynq)
- AI engine (VersalAI)

Соответственно, CNN реализуется как программный код для этих ядер. В аппаратуре имплементируются соответсвующие ядра – по одному ядру на сеть. В одной ПЛИС или СнК можно реализовать несколько сетей. DPU различаются по производительности

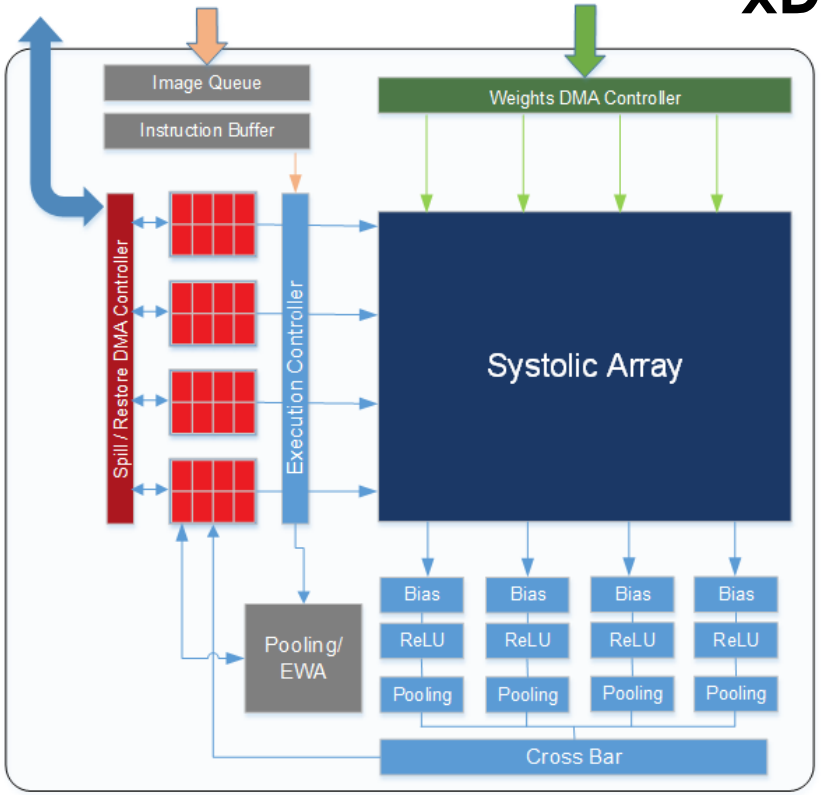
DPU для различных платформ Xilinx

Платформы	DPU
Alveo	Soft: xDNN
Zynq-7000, ZynqUS+,Versal	Soft: Bxxxx
VersalAI	Hard: AI Engines

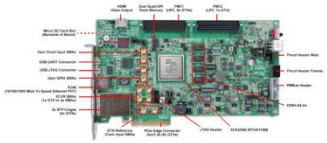
Alveo: https://github.com/Xilinx/ml-suite/blob/master/docs/ml-suite-overview.md
Zynq: https://www.xilinx.com/support/documentation/ip_documentation/dpu/v3_1/pg338-dpu.pdf

Структура DPU Alveo

xDNN IP Core

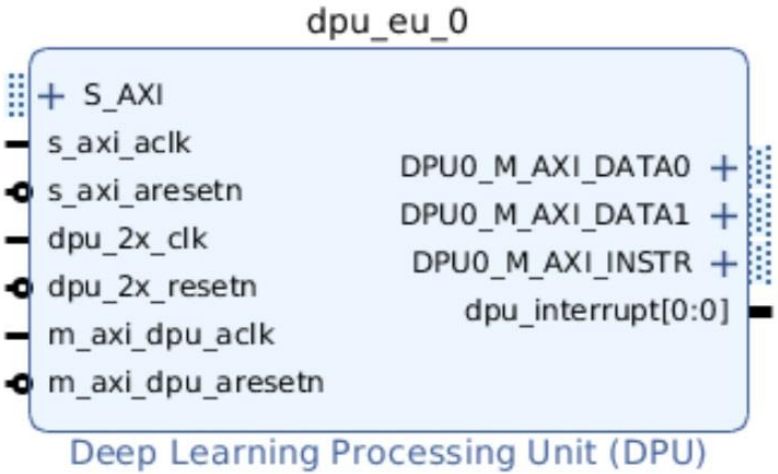


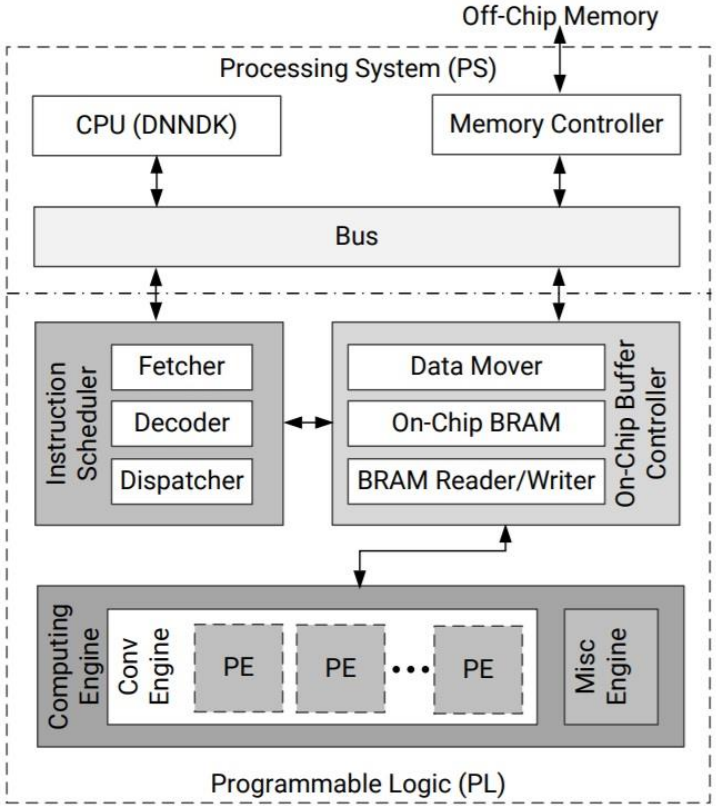
DSP Array Configuration	Total Image Memory per PE	Total DSPs in Array	16-bit GOP/s	8-bit GOP/s
96x16	9 MB	1536	2150 @700MHz	4300 @700MHz
28x32	4 MB	896	896 @500MHz	1792 @500MHz
56x32	6 MB	1792	1792 @500MHz	3584 @500MHz

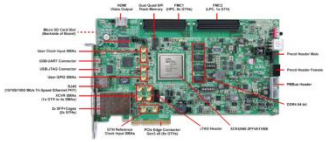


Структура DPU

Системы на кристалле: Zynq, ZynqUS+, Versal





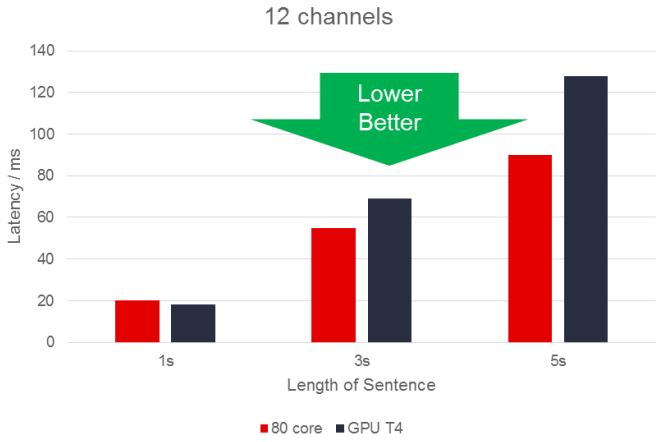


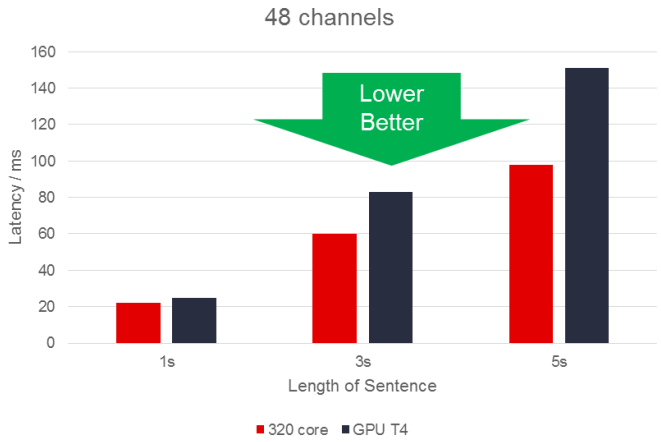
Характеристики ядер DPU

High DSP Usage					Low DSP Usage				
Arch	LUT	Register	BRAM	DSP	Arch	LUT	Register	BRAM	DSP
B512	20055	28849	69.5	98	B512	21171	33572	69.5	66
B800	21490	34561	87	142	B800	22900	33752	87	102
B1024	24349	46241	101.5	194	B1024	26341	49823	101.5	130
B1152	23527	46906	117.5	194	B1152	25250	49588	117.5	146
B1600	26728	56267	123	282	B1600	29270	60739	123	202
B2304	39562	67481	161.5	386	B2304	32684	72850	161.5	290
B3136	32190	79867	203.5	506	B3136	35797	86132	203.5	394
B4096	37266	92630	249.5	642	B4096	41412	99791	249.5	514

Сравнение производительности

Comparison between Versal ACAP and nVidia T4



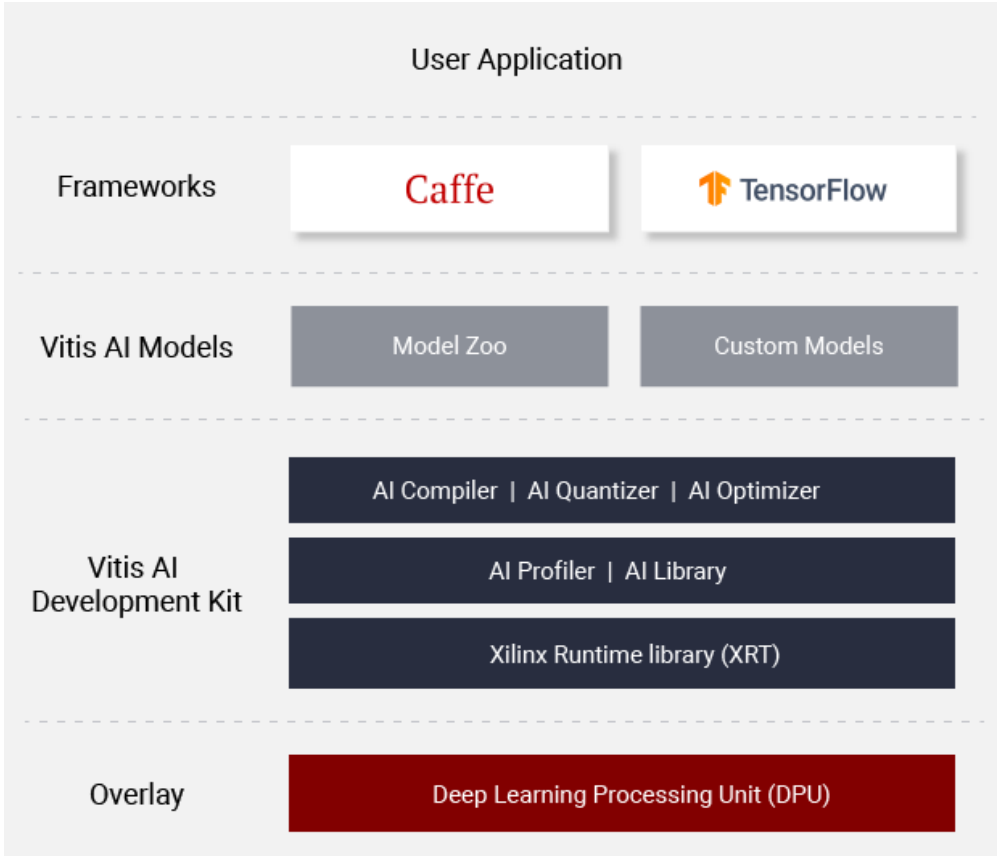


	1s	3s	5s	Power
Versal 80-core AIE	20	55	90	21.9w
GPU T4 fp16	18	69	128	75w

	1s	3s	5s	Power
Versal 320-core AIE	22	60	98	50.8w
GPU T4 fp16	25	83	151	75w

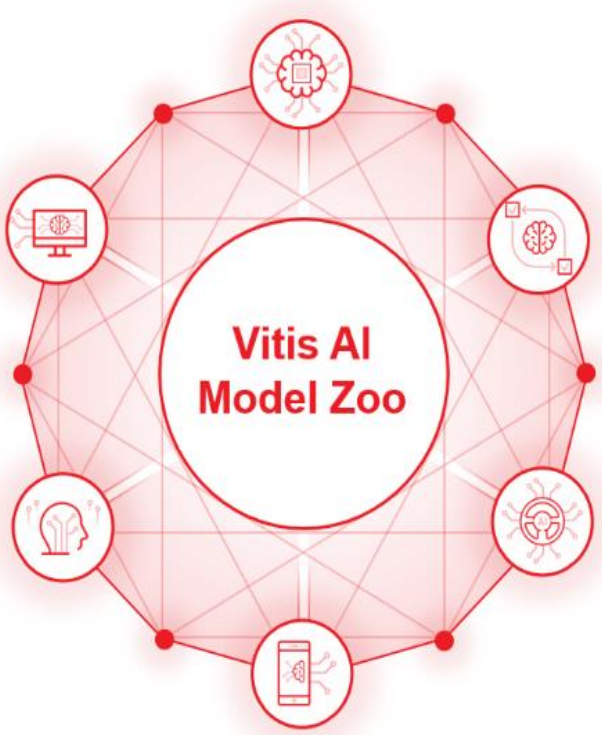
Units in time (ms)

Структура Vitis AI



Единый пакет как для Alveo, так и для Edge платформ

Vitis AI: библтотека моделей



Rich Models from Tensorflow and Caffe

Модели из Tensorflow и Caffe

Open and Free on Github for All Developers

Доступны на гитхабе для всех

Advanced Optimization, Including Pruning, Applied

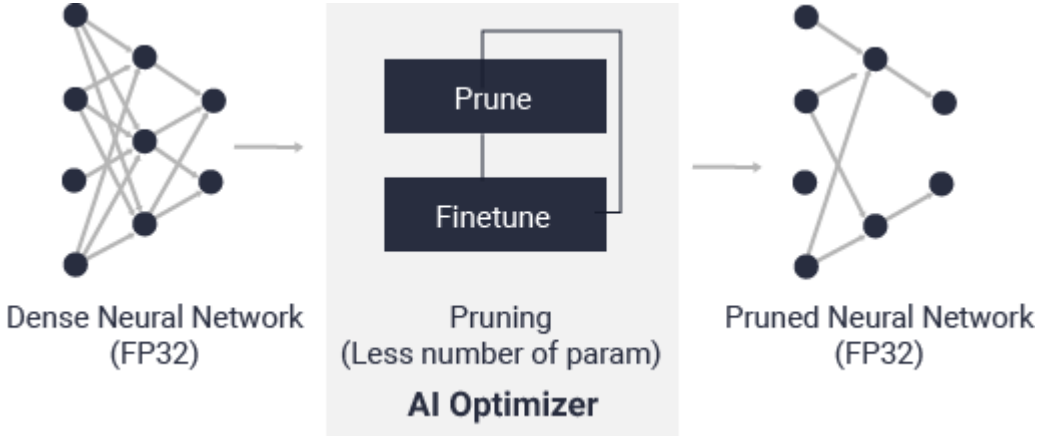
Оптимизация, включая прореживание

Retrainable with Custom Dataset

Переобучение на пользовательских датасетах

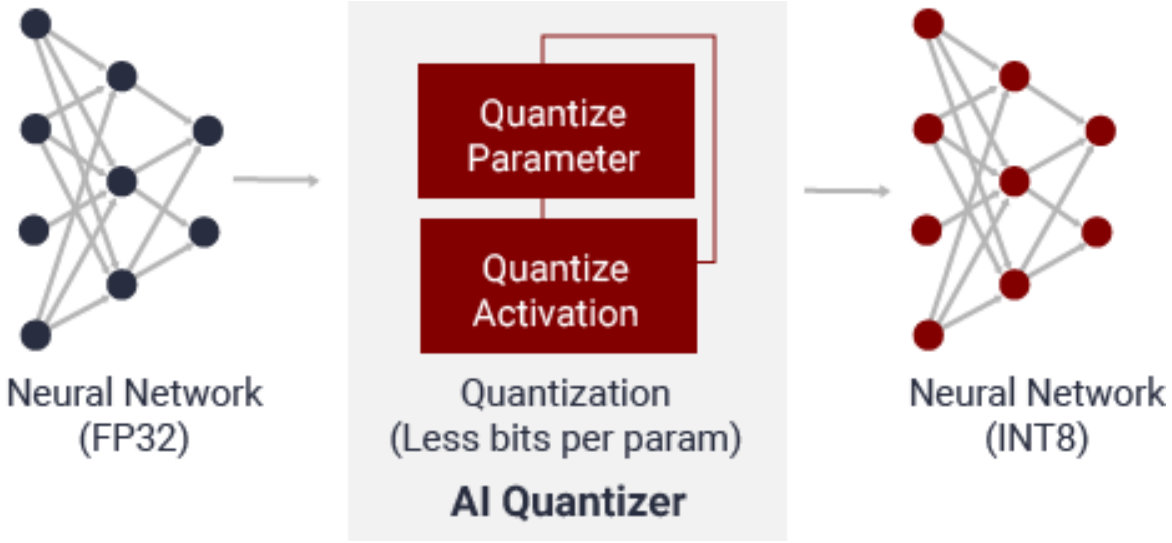
Vitis AI: прореживание

Прореживание(Pruning) – ликвидация избыточных связей между узлами.
Используются Проприетарные разработки Xilinx
Этап опциональный, лицензия платная и дорогая



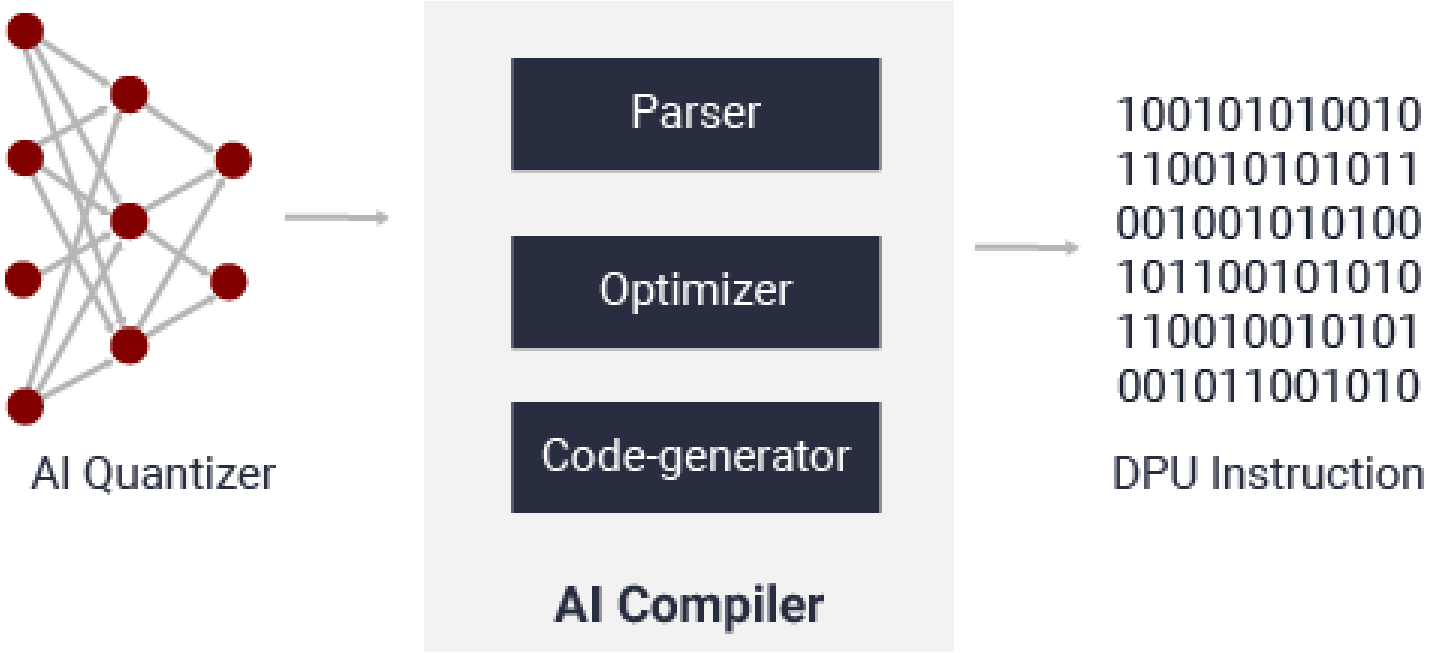
Vitis AI: Квантизатор

Конвертация из FP32 в FP16 существенно повышает быстродействие практически без потери точности



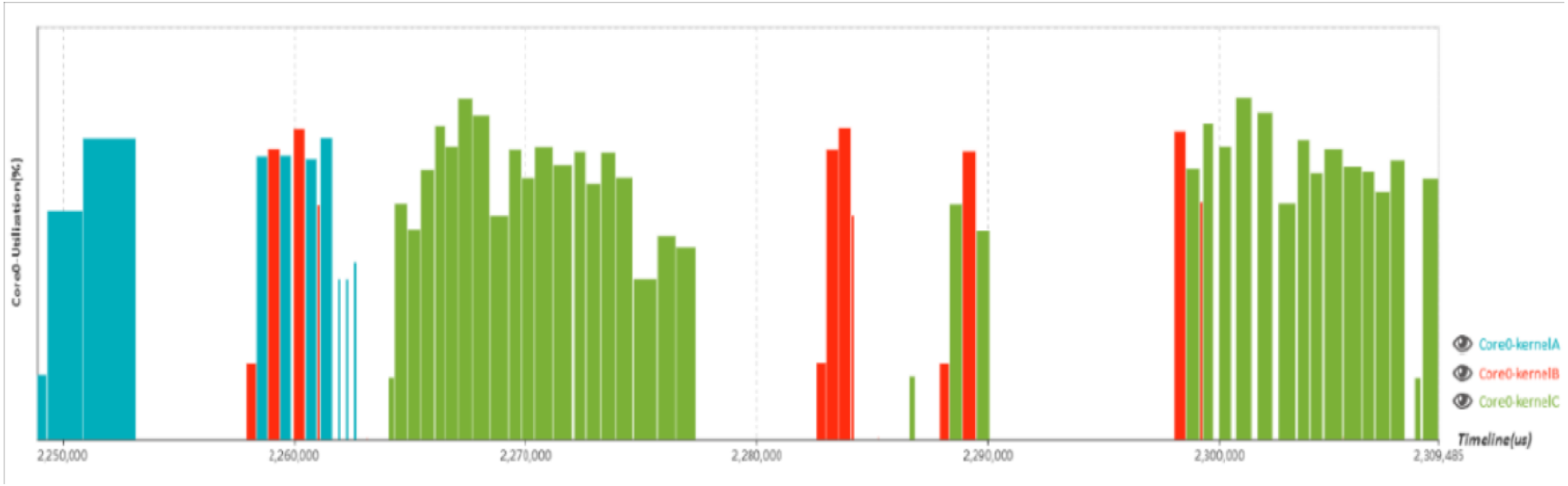
Vitis AI: Компилятор

Компилирует сеть в программу, выполняющуюся на DPU

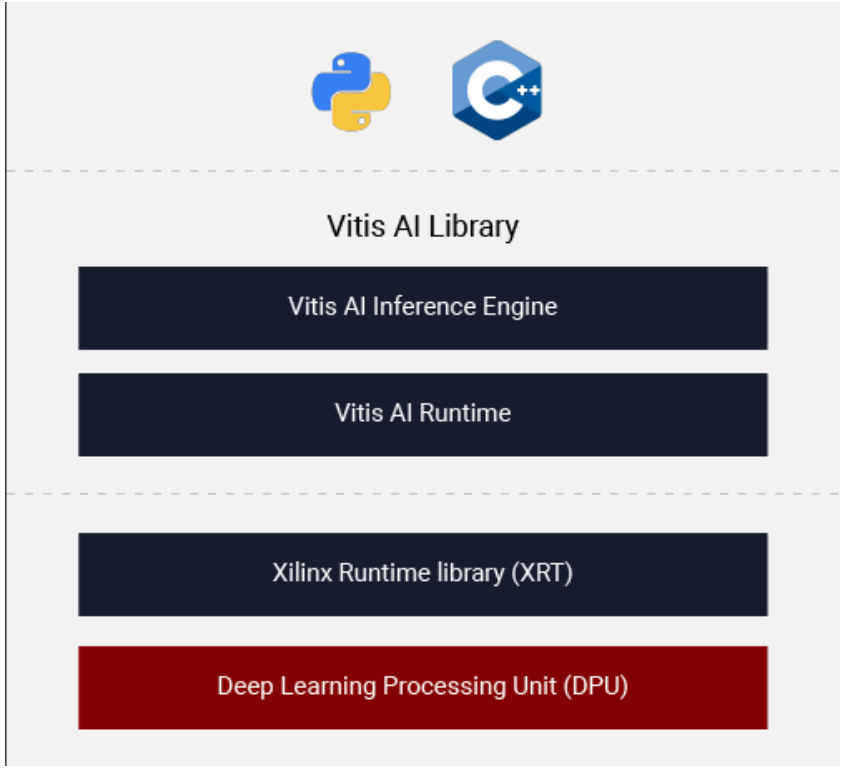


Vitis AI: Профайлер

Анализ производительности и требуемых ресурсов



Vitis AI: библиотеки



Библиотеки времени выполнения: набор API на языках C++ и Python для встраивания в собственные приложения.

Спасибо за внимание

Ваши вопросы – наши ответы

МАКРО ГРУПП

Официальный дилер Xilinx

Контакты

Тел.: 8 (800) 333-06-05
email: SALES@MACROGROUP.RU

Продукция xilinx и техподдержка: fpga@macrogroup.ru
Олег Болихов – руководитель направления
“Цифровая электроника”
Дмитрий Хорьков – руководитель направления Xilinx
Владимир Викулин – техподдержка Xilinx