



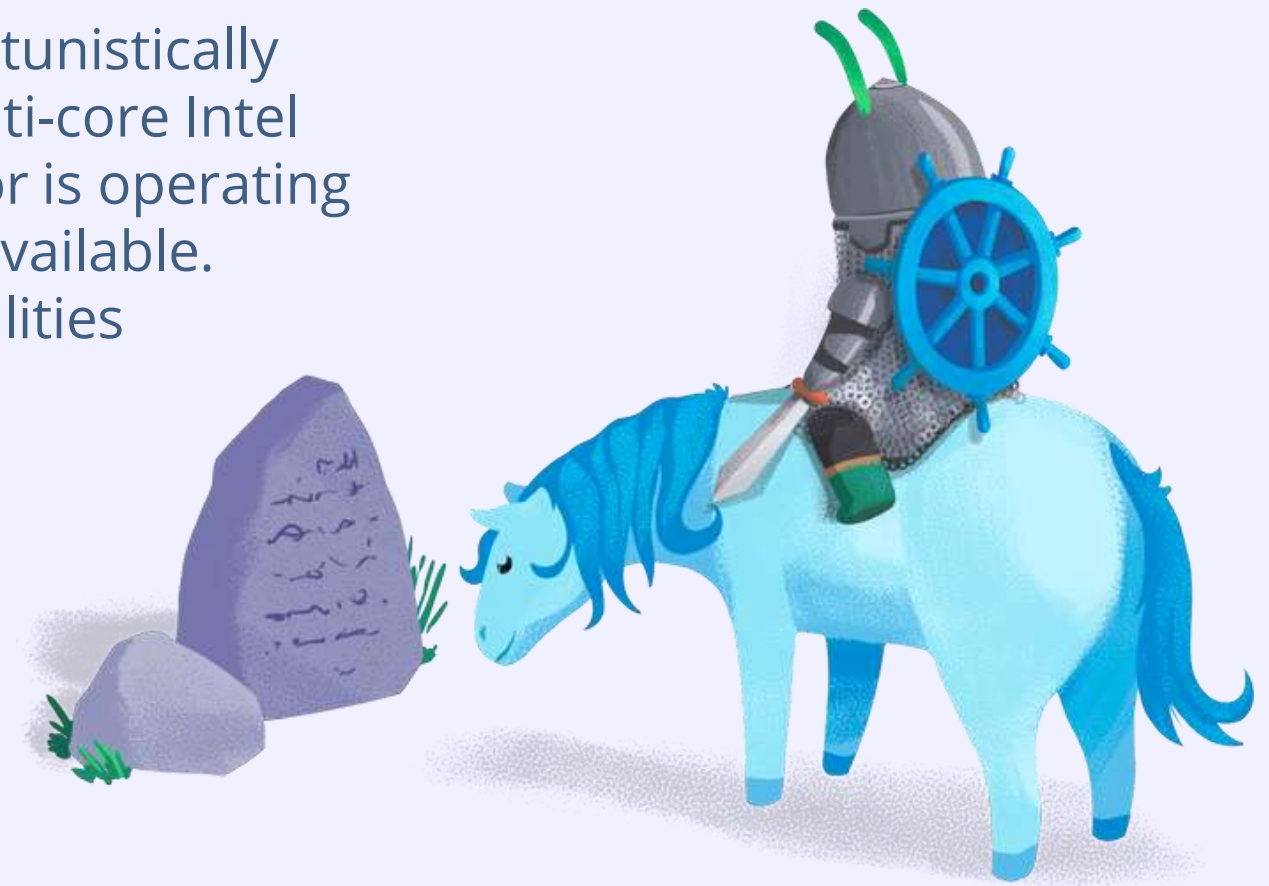
Урок 16. Решения для Deep & Machine Learning

Оптимизированные под процессоры Intel решения

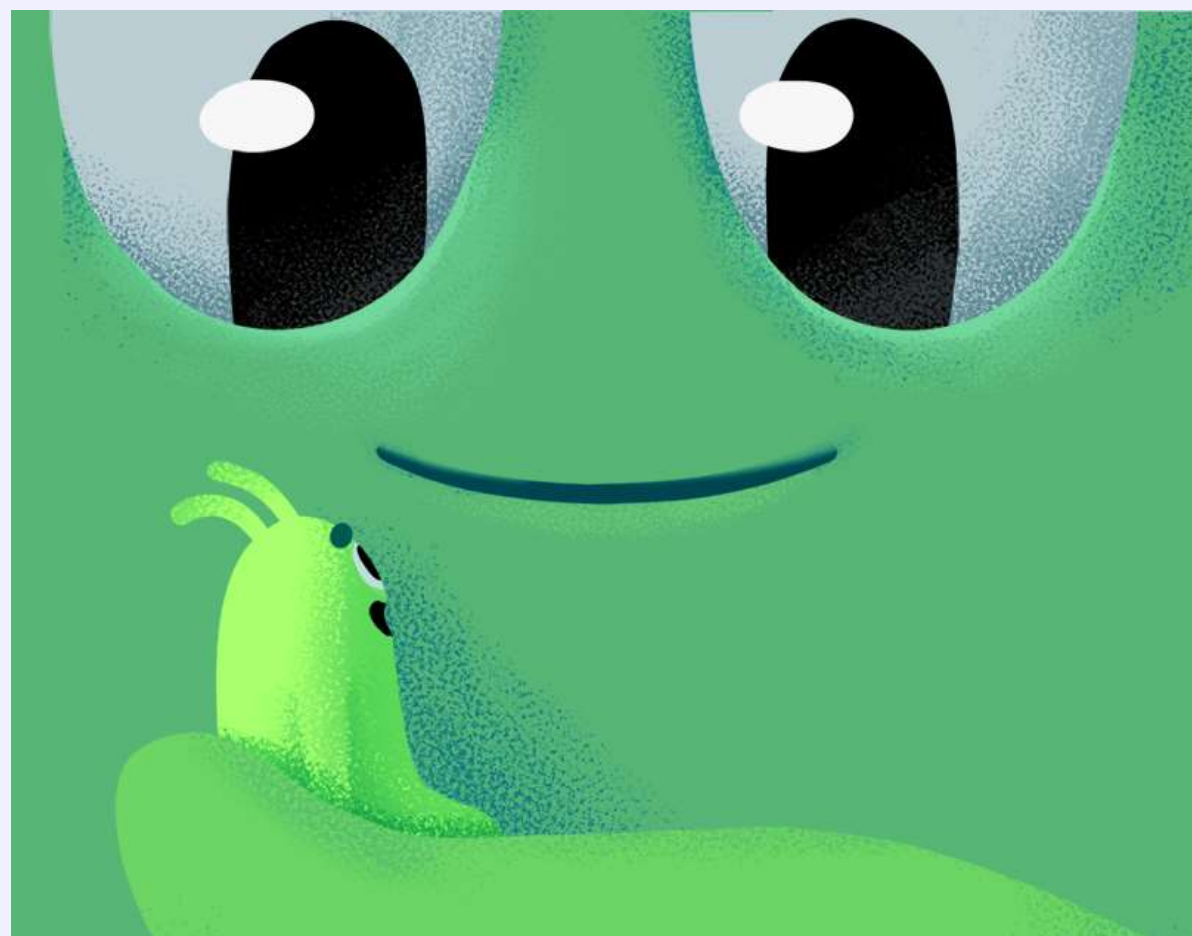
Дмитрий Сивков
Инженер-консультант
Intel Russia

Notices and Disclaimers

- Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex
- Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details.
- No product or component can be absolutely secure.
- Includes the effect of Intel Thermal Velocity Boost, a feature that opportunistically and automatically increases clock frequency above single-core and multi-core Intel Turbo Boost Technology frequencies based on how much the processor is operating below its maximum temperature and whether turbo power budget is available. The frequency gain and duration is dependent on the workload, capabilities of the processor and the processor cooling solution.
- Code names are used by Intel to identify products, technologies, or services that are in development and not publicly available. These are not "commercial" names and not intended to function as trademarks.
- © Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

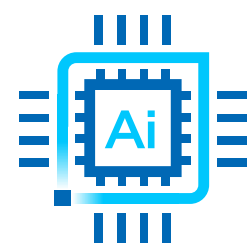


Представим мир, наполненный ИИ

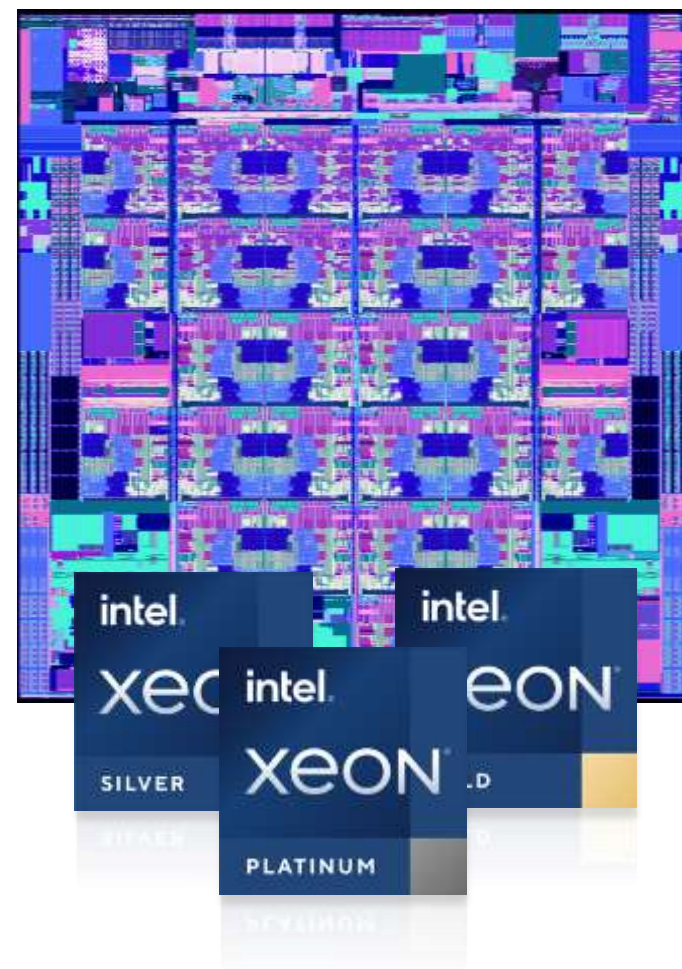


Хеон повсюду

Встроенное ускорение
ИИ в x86 CPU для ЦОД
(Intel® Deep Learning
Boost - Intel® DL Boost)



Инструменты data science
и экосистема решений



Безопасное
федеративное обучение
Intel® Software Guard
Extensions (Intel® SGX)



Возможность
обрабатывать терабайты
данных с
Intel® Optane™ Persistent
Memory

3-е поколение Intel® Xeon® Scalable

Гибкое ускорение ИИ

CPU

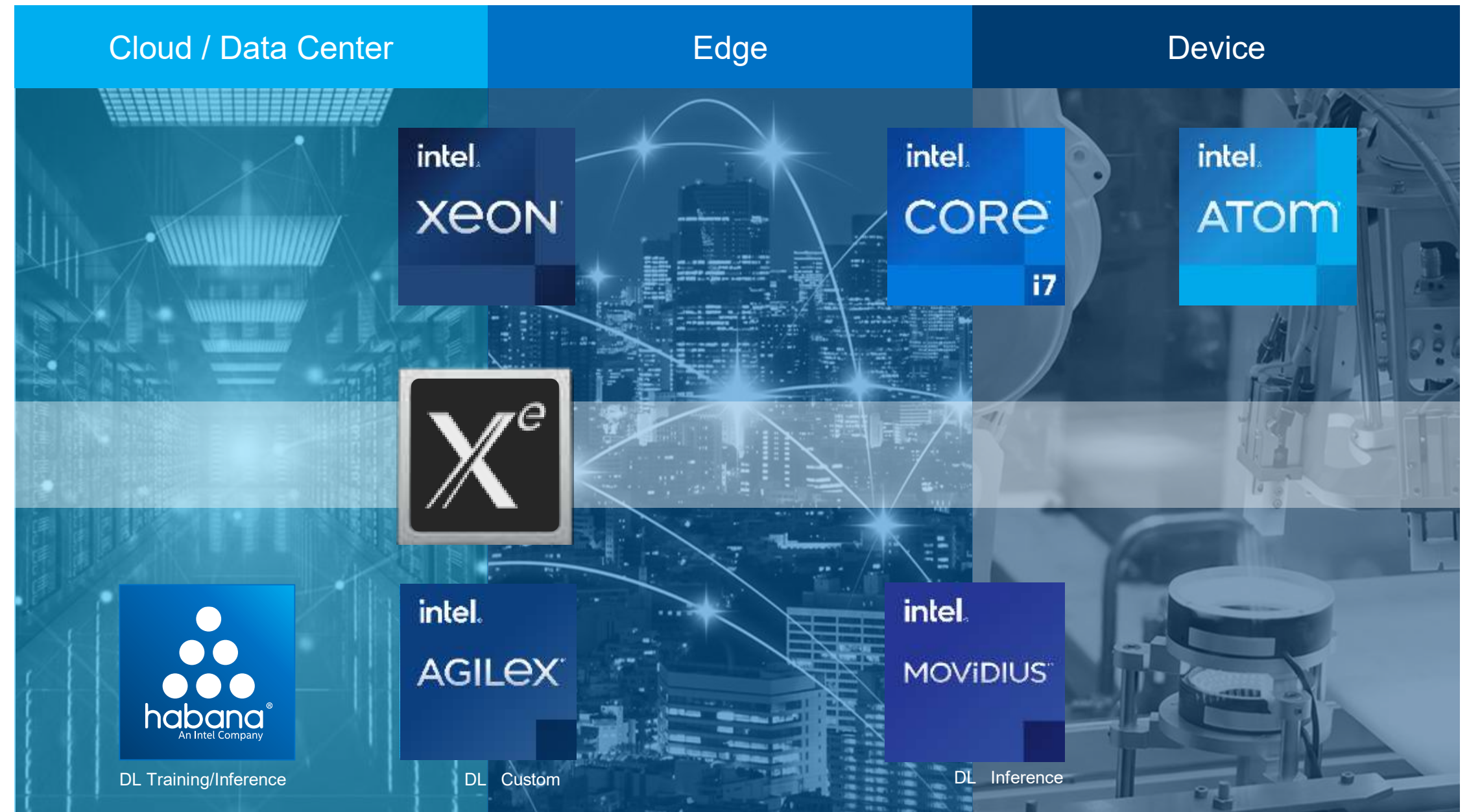
Ускорение для основных сценариев ИИ

CPU + GPU

Доминирование вычислительных сценариев AI, HPC, фотореалистичной графики и риаптайм обработки медиа

CPU + спецускорители

Доминирование сценариев с глубоким обучением



Intel® Stratix® 10 NX FPGA

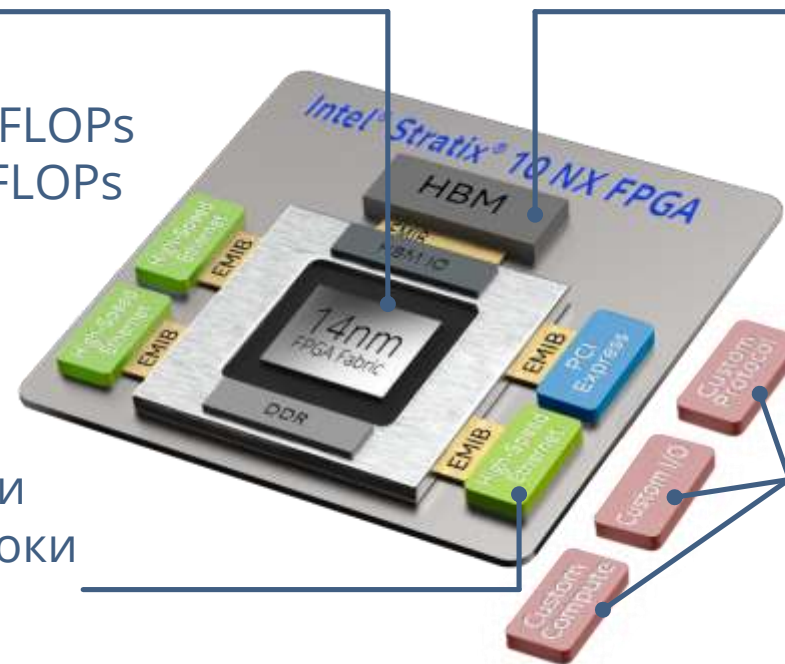
Первый оптимизированный Intel FPGA

AI Tensor Blocks

- 3,960 AI Tensor Blocks
- 286 INT4 TOPs и Block FP12 TFLOPs
- 143 INT8 TOPs и Block FP16 TFLOPs
- 1-2 TOPs/W или TFLOPs/W

Быстрая сеть

- До 57.8G PAM4 трансиверов и аппаратные Intel Ethernet блоки
- Гибкий и настраиваемый интерконнект
- Для масштабирования



Много памяти Near-Compute

- Настраиваемая иерархия памяти
- Встроенные 8 GB или 16 GB HBM памяти

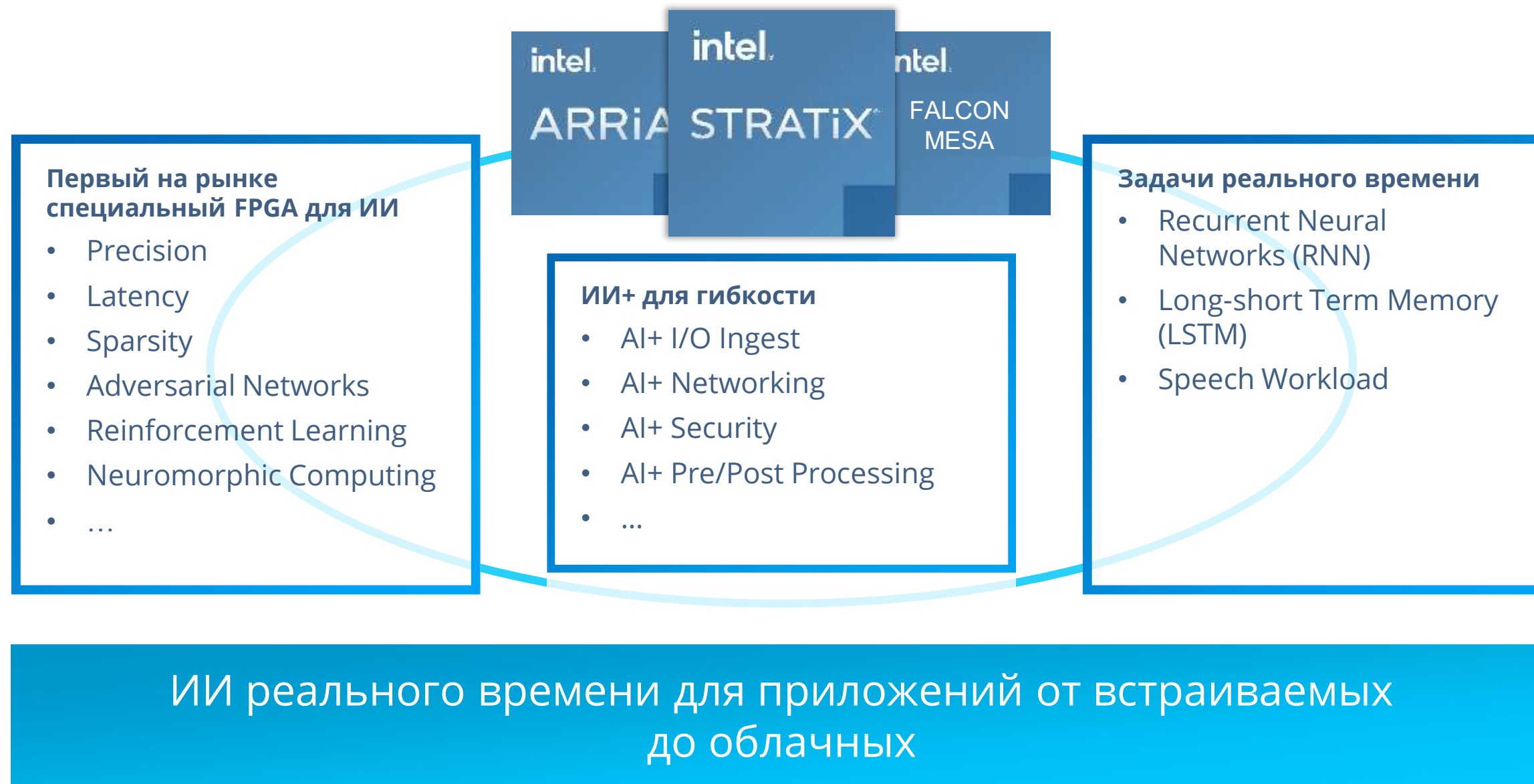
Расширяемая

- Чиплеты дают интерфейс к легкой адаптации и кастомизации и расширения функций ASIC

**Tensor Compute, Near Memory и Networking дают
Высокую производительность для решений AI**

Peak performance for INT8 precision is calculated as follows: (3,960 AI Tensor Blocks) * (30 multiplications per AI Tensor Block) * (2 operations per multiplication) = 237,600 operations.
Assuming 600 MHz maximum frequency: (237,600 operations) * (600 MHz) = 142.56 TOPS ~143 TOPS. The peak performance for INT4 precision is double that of INT8, or approximately 286 TOPS
See the following white paper for more details: <https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/a1040767-wp-01301-pushing-ai-boundaries-with-scalable-compute-focused-fpgas.pdf>

Intel FPGA для ИИ

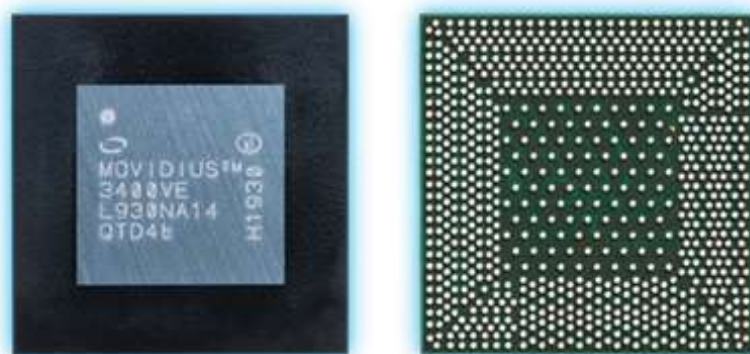


All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

Intel® Movidius™ VPU

Edge AI

- ✓ Deep learning inference + computer vision + media
- ✓ Faster memory bandwidth
- ✓ Groundbreaking high-efficiency architecture
- ✓ OpenVINO toolkit enabled



Выбор форм-факторов



Приложения



X^e HPC (Ponte Vecchio)

Производительность для требовательных ИИ задач



<https://www.youtube.com/watch?v=JzbN1IOAcwY>

>40 активных тайлов, более 100 миллиардов транзисторов в одном ускорителе

Компоновка нескольких технологий производства
EMIB (2D) и Foveros (3D)

POWERING AURORA

Mobileye – an Intel Company



Visit www.mobileye.com

АВТОПИЛОТ

All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

СЛЕРМ при поддержке

 VK Cloud Solutions

+

 intel.

Habana – an Intel Company



Visit www.habana.ai

ИИ-ускорители

All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

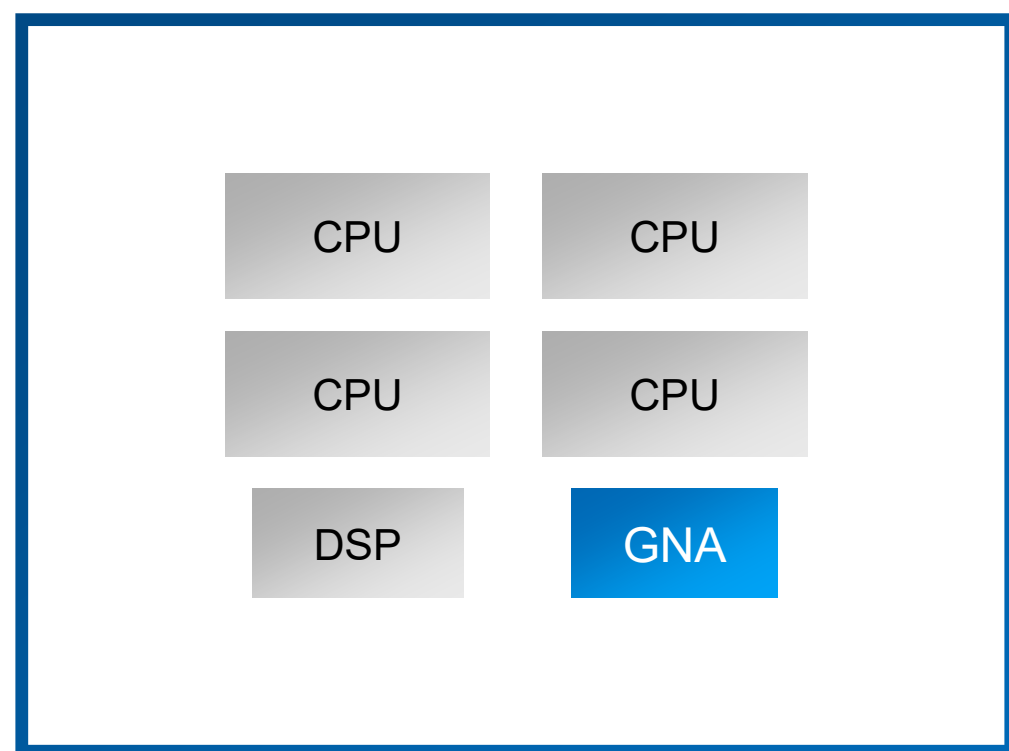
СЛЕРМ при поддержке

 VK Cloud Solutions

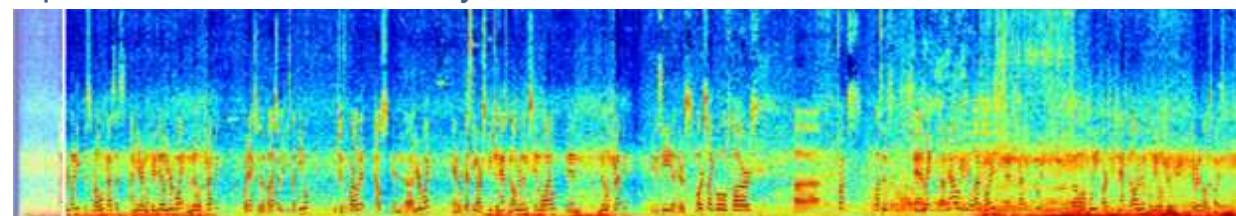
+

 intel.

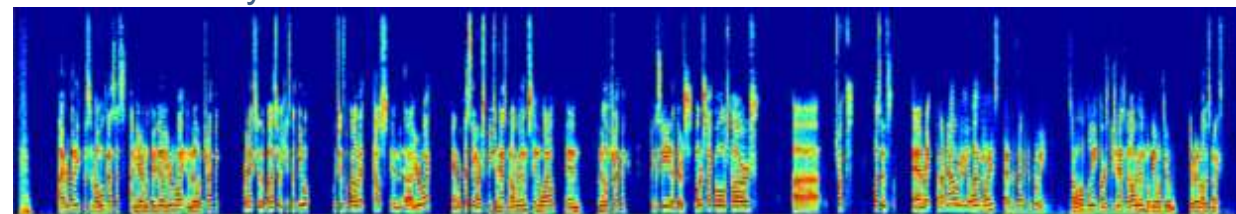
Intel® GNA – Gaussian & Neural Accelerator



Оригинальная запись с шумом



Запись без шума



Intel® GNA = Intelligent Noise Cancellation

- Отдельный IP-блок: самое низкое энергопотребление из ускорителей ИИ Intel
- Может работать, пока основная SOC в режиме энергосбережения
- Помогает CPU: снижает энергопотребление и высвобождает ресурсы CPU

ИИ повсюду: От технологий к решениям

Нажми на кнопку – получишь ИИ
200+ решений «под ключ» &
их поставщики

Solutions

Intel Solutions
Marketplace



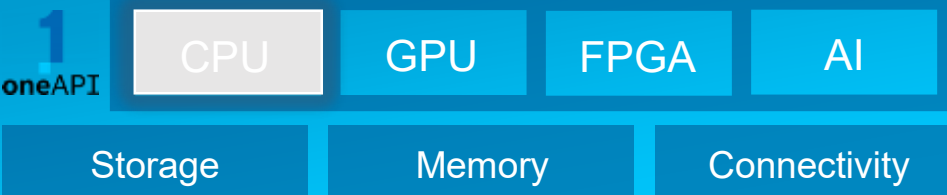
Умные удобные инструменты
150+ контейнеров для data science

Tools



Ускорь ИИ сегодня
1.5x vs. AMD и 1.3x vs. Nvidia на 20 ИИ
задачах

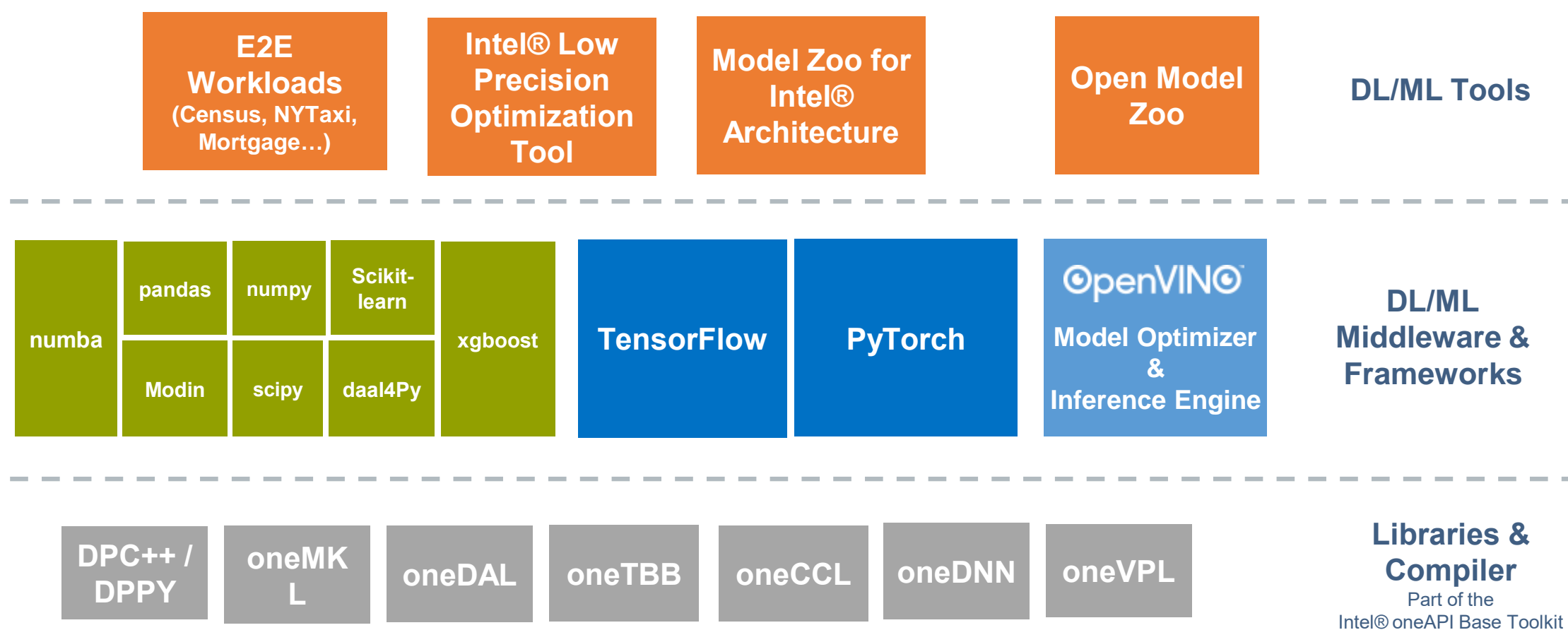
Technology



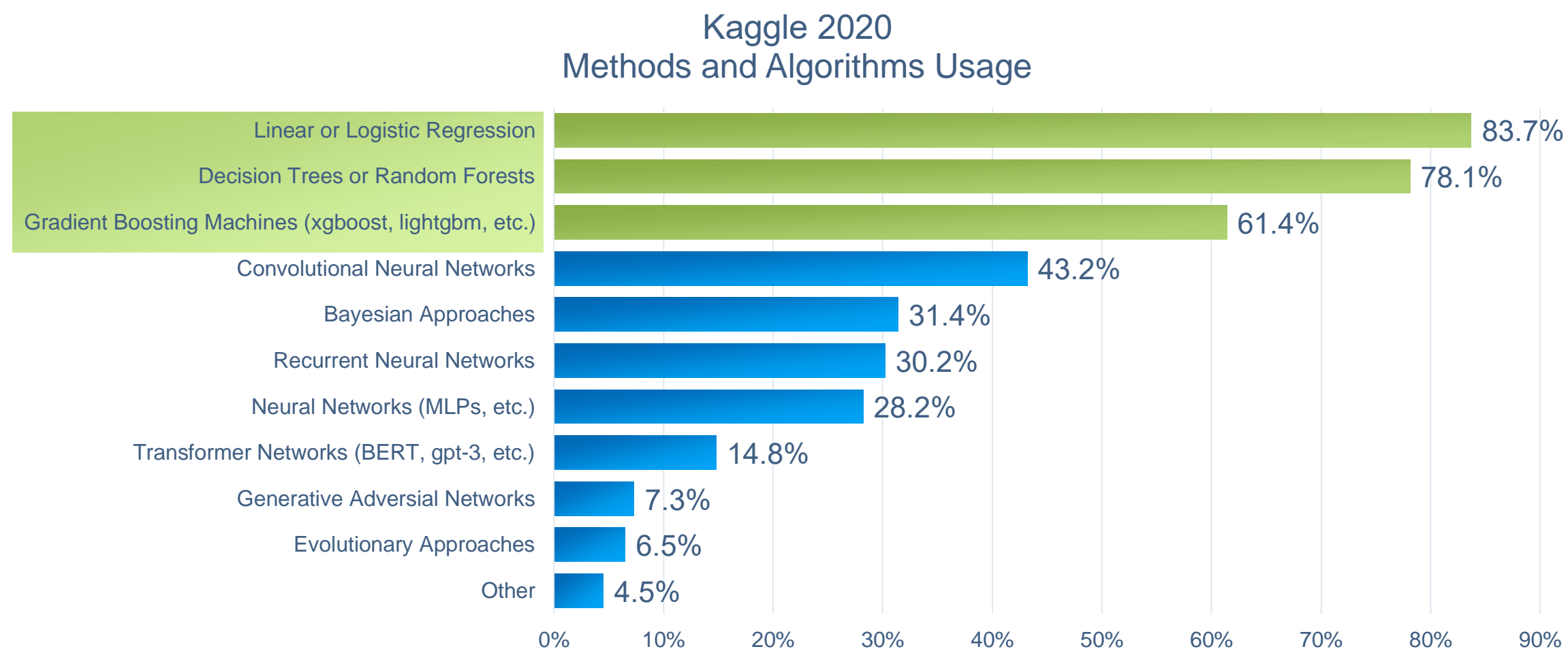
See claims [43, 44] at www.intel.com/3gen-xeon-config for workloads and configurations. Results may vary.

Программный стек ИИ для Intel XPU

Intel дает полный стек ПО для оптимизации производительности



Исследователи и инженеры чаще используют машинное обучение, чем глубокое обучение



Source: <https://www.kaggle.com/kaggle-survey-2020>

Intel Distribution for Python

Преимущества для разработчика



Installing Intel® Distribution for Python*

Установщик

Download full installer from
<https://software.intel.com/en-us/intel-distribution-for-python>

Anaconda.org

Anaconda.org/intel channel

```
> conda config --add channels intel  
> conda install intelpython3_full  
> conda install intelpython3_core
```

PyPI

```
> pip install intel-numpy  
> pip install intel-scipy  
> pip install mkl_fft  
> pip install mkl_random
```

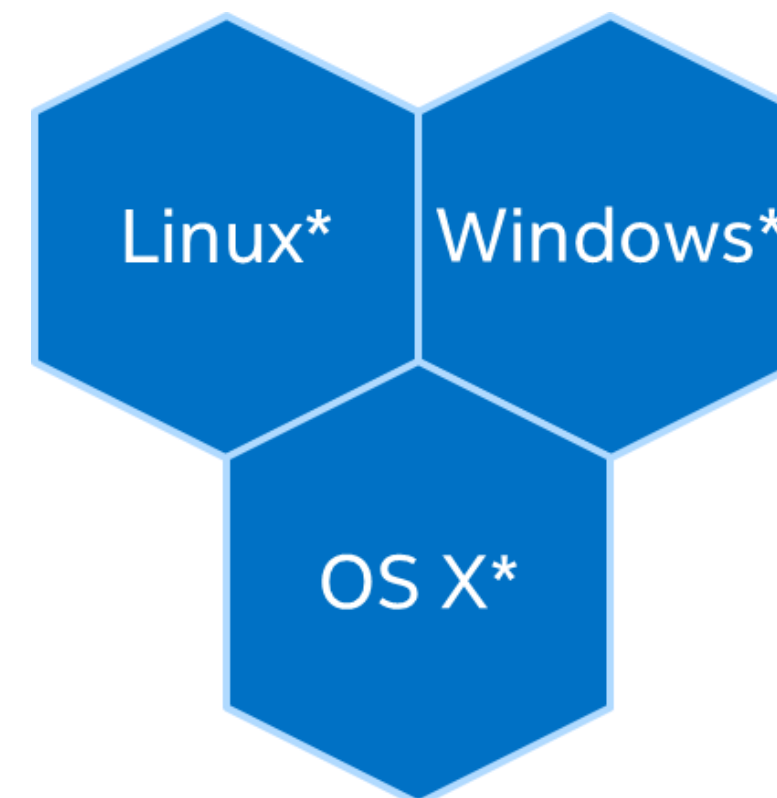
+ Intel library Runtime packages
+ Intel development packages

Docker Hub

```
docker pull intelpython/intelpython3_full
```

YUM/APT

Access for yum/apt:
<https://software.intel.com/en-us/articles/installing-intel-free-libs-and-python>



Intel oneAPI Data Analytics Library (oneDAL)

API независимое от аппаратных решений и вендора

Основан на Data Parallel C++ (DPC++) and C++17

Управление данными

Алгоритмы

Независимое от устройств представление
Поддержка разнородных и разреженных данных
Совместимость с форматом Apache Arrow
Расширяемый пользовательские форматы
Поддержка сериализации/десериализации
Эффективная реализация алгоритмов
машинного обучения
Поддержка распределенных и параллельных
вычислений

Семейства алгоритмов

GLM	SVM	k-NN	k-Means
Statistics	Random Forest	Decision Trees	PCA
DBSCAN	ALS	Naïve Bayes	...

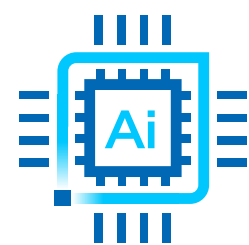
*Зависит от оптимизаций

Available as part of Intel® oneAPI Base Toolkit : intel.com/oneAPI-BaseKit

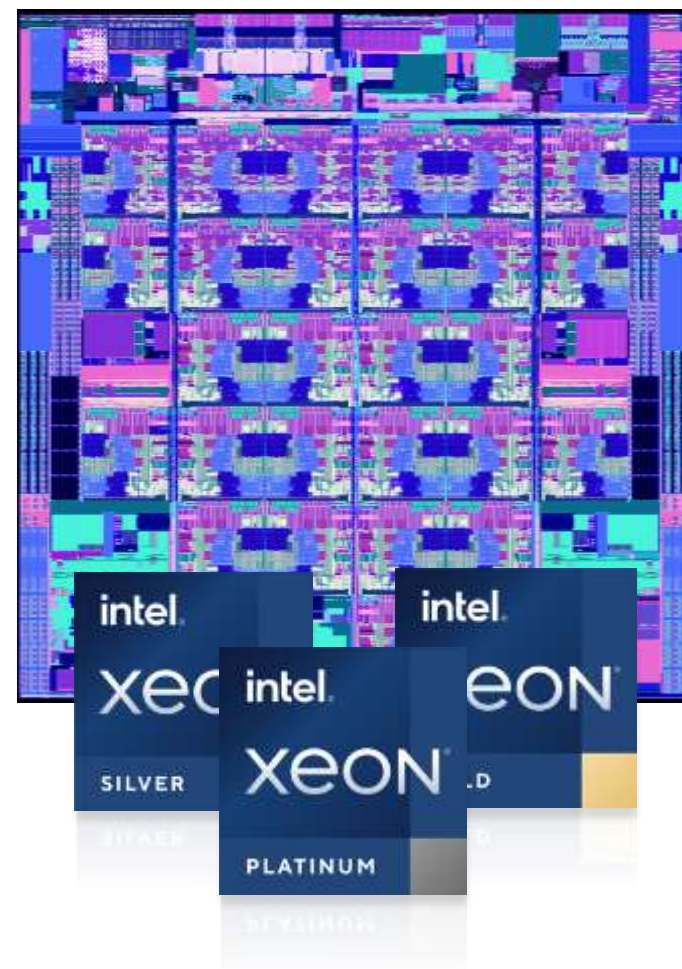
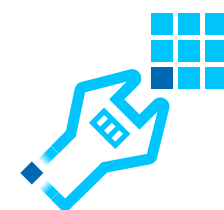
Disclaimer: API shown in the presentation is subject to change

Хеон повсюду

Встроенное ускорение
ИИ в x86 CPU для ЦОД
(Intel® Deep Learning
Boost - Intel® DL Boost)



Инструменты data science
и экосистема решений



Безопасное
федеративное обучение
Intel® Software Guard
Extensions (Intel® SGX)



Возможность
обрабатывать терабайты
данных с
Intel® Optane™ Persistent
Memory

3-е поколение Intel® Xeon® Scalable

Intel Deep Learning Boost

A Vector Neural Network Instruction (VNNI)

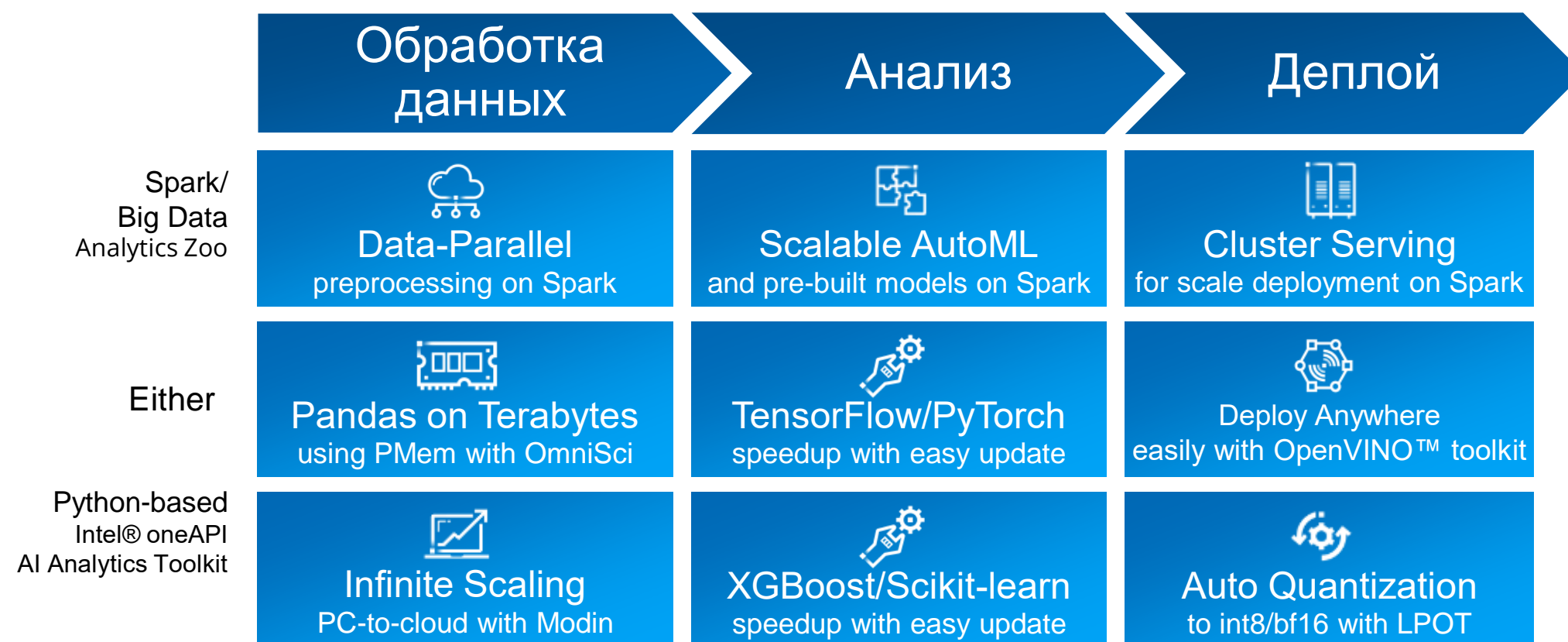
Расширяет Intel AVX-512 для ускорения AI/DL



1. See [123] at www.intel.com/3gen-xeon-config. Results may vary.

Оптимизации на всех этапах

Powered by 



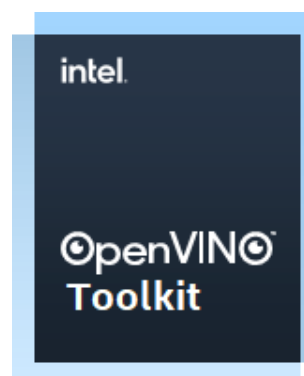
Popular tools optimized, 3 Intel toolkits, 150+ containers

Наборы инструментов в экосистеме oneAPI



Intel® oneAPI AI Analytics Toolkit

- Accelerate machine learning & data science pipelines with optimized DL frameworks & high-performing Python libraries
- Data Scientists, AI Researchers, DL/ML Developers



Intel® Distribution of OpenVINO™ Toolkit

- Deploy high performance inference & applications from edge to cloud
- AI Application, Media, & Vision Developers



Intel® oneAPI Base Toolkit

Includes oneDNN, oneCCL & oneDAL

- Optimize primitives for algorithms and framework development
- DL Framework Developers - Optimize Algorithms for Machine Learning & Analytics

Intel® Distribution of OpenVINO™ Основан на oneAPI

Ускорение инференса глубокого обучения

Инструменты для ускорения разработки решений с инференсом глубокого обучения, машинного зрения. Оптимизирует использование ускорителей Intel CPUs, GPUs, FPGAs, VPUs.

Кому нужен этот продукт

Разработчики машинного зрения, искусственного интеллекта

Датасайнтисты

ОЕМы, ISV, системные интеграторы

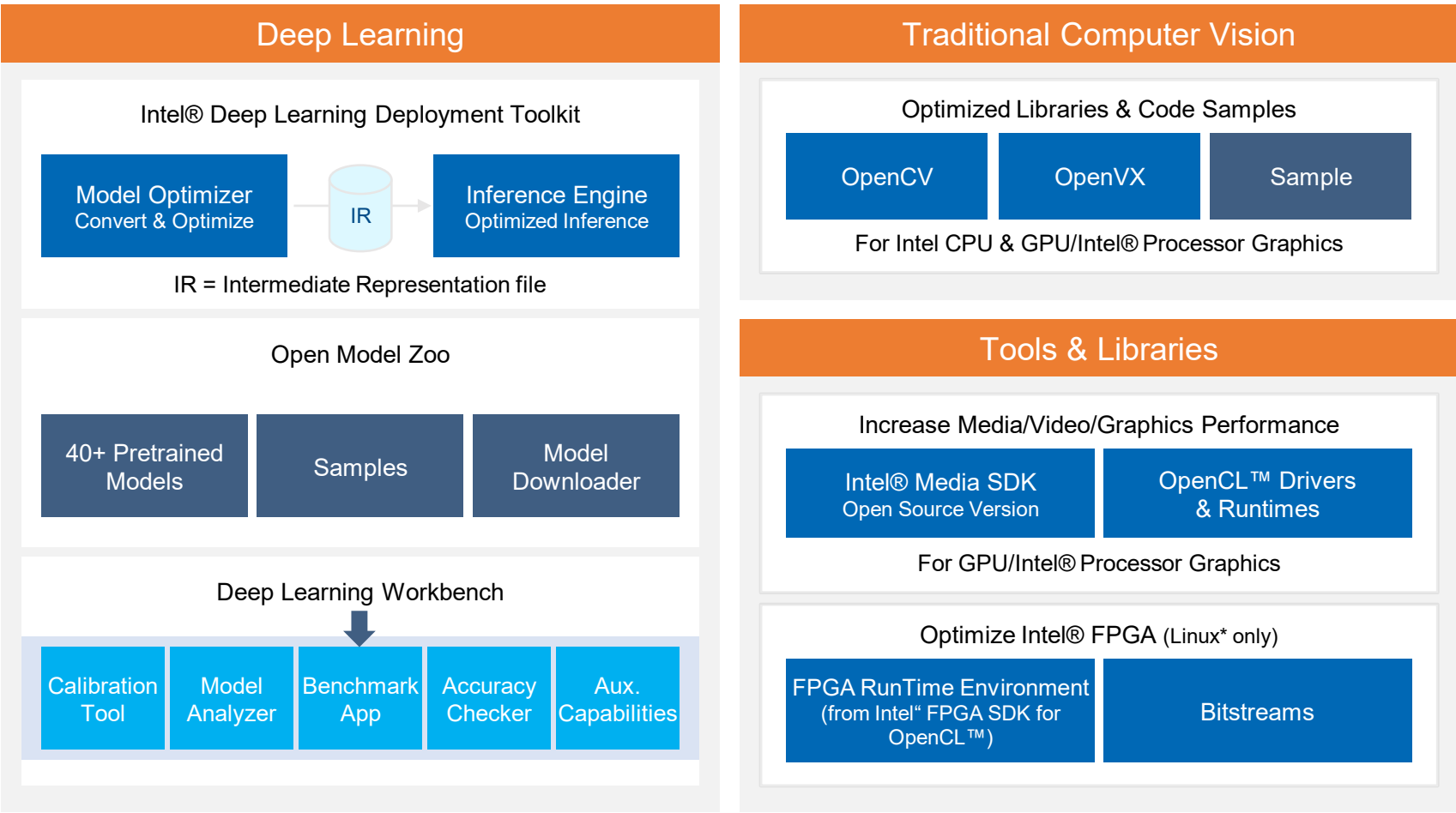
Использование

Системы безопасности, робототехника, продажи, медицина, ИИ, автоматизация офисов, транспорт, голос, язык и многое другое

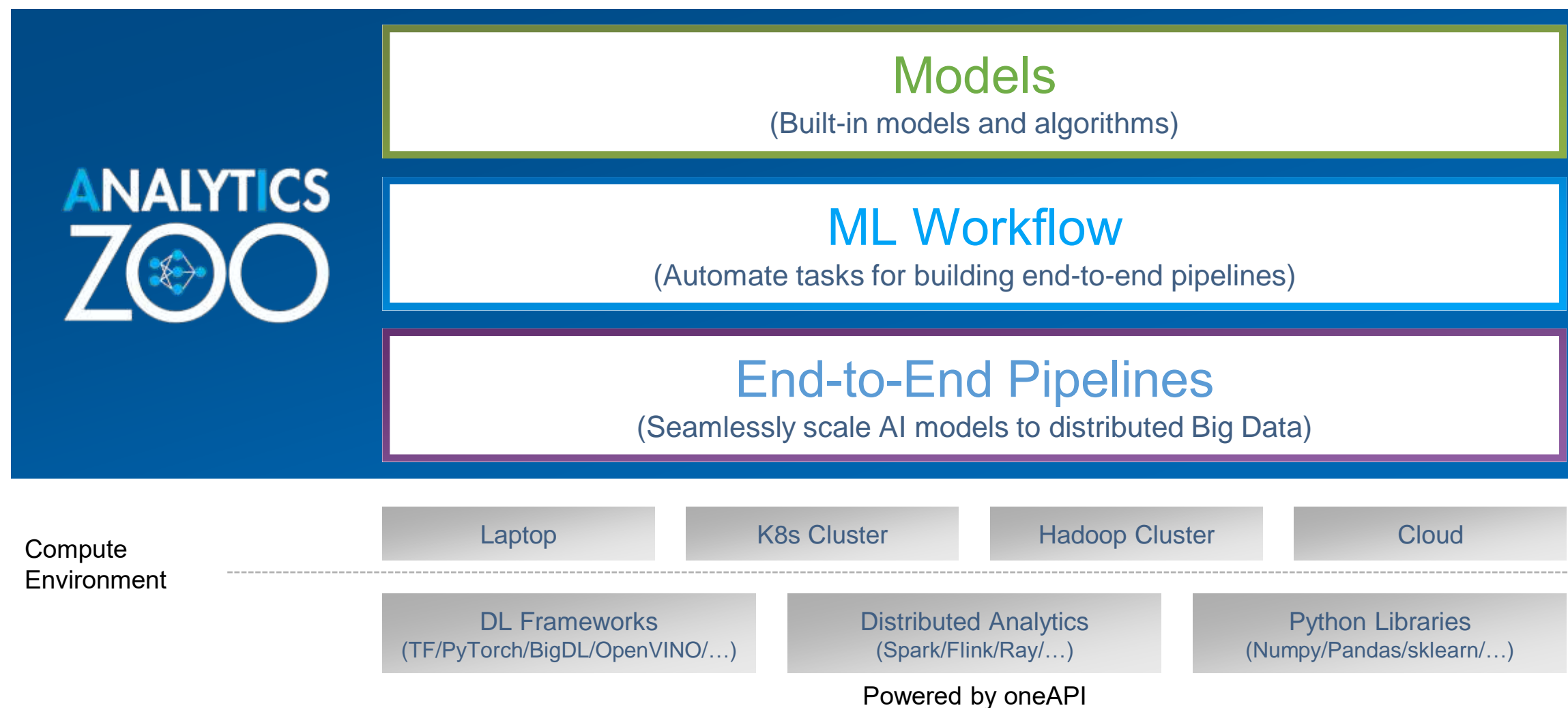
Edge AI &
Vision Alliance



Intel® Distribution of OpenVINO™ toolkit

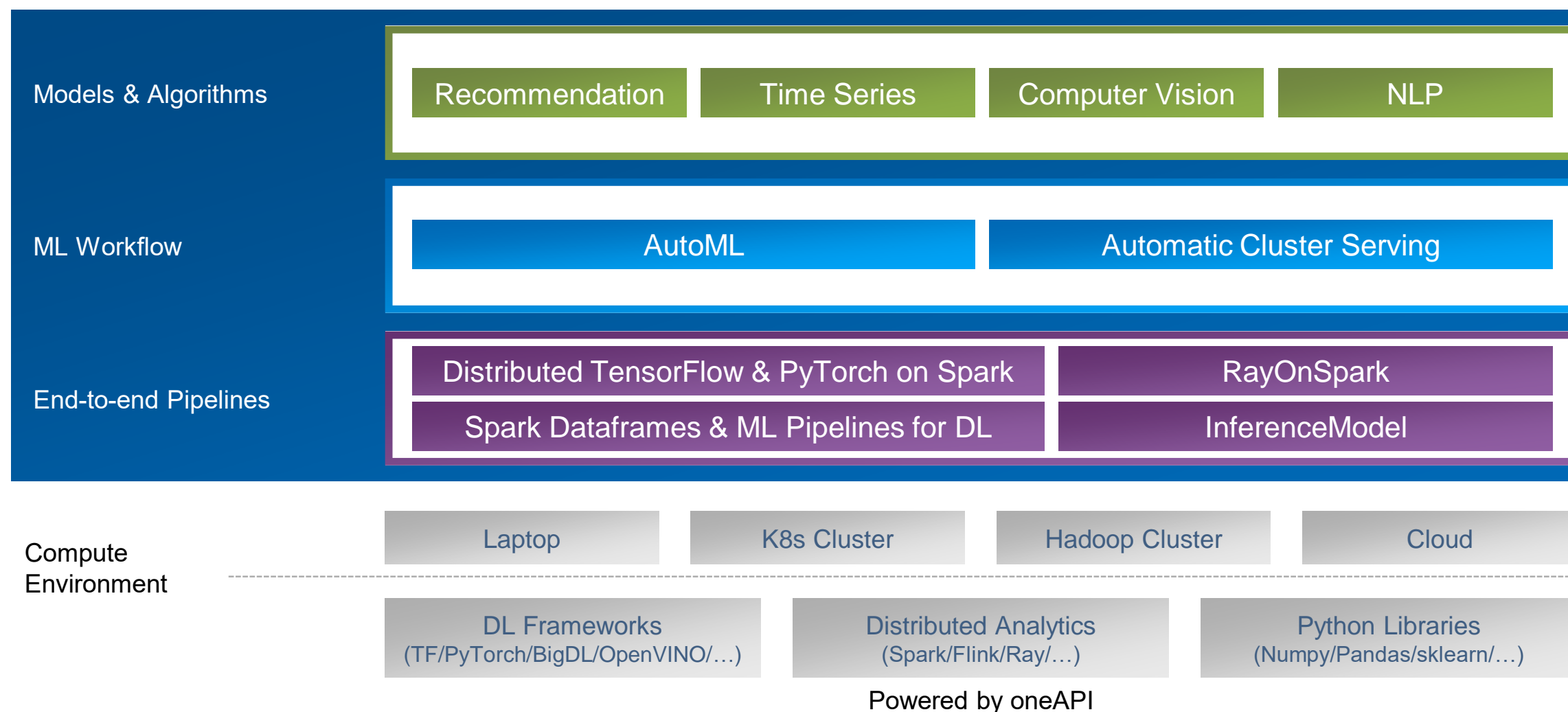


Analytics Zoo: Платформа ИИ Больших Данных



<https://github.com/intel-analytics/analytics-zoo>

Analytics Zoo: Платформа ИИ Больших Данных



<https://github.com/intel-analytics/analytics-zoo>

СЛЕПМ при поддержке



VK Cloud Solutions



intel.

Спасибо!





Мы ищем разработчиков:



Все вакансии



Experienced:

- Build/DevOps Engineer
- DevOps Engineer (Integration, Computer Vision)
- IT Infrastructure Engineer
- Software Validation Engineer (DevOps, Computer Vision)
- Infrastructure and DevOps engineer
- Python Backend Developer
- Cloud Orchestration SW Engineer (Computer Vision)



Interns (для студентов очной формы):

- DevOps intern
(Parallel Runtimes Engineering team)
- Infrastructure and DevOps Intern