Intel® oneAPI для xPU:
ни один транзистор не останется без дела

Intel
Дмитрий Сивков

oneAPI

Инструменты Intel® oneAPI
на платформе ML Space

Попробуйте сейчас

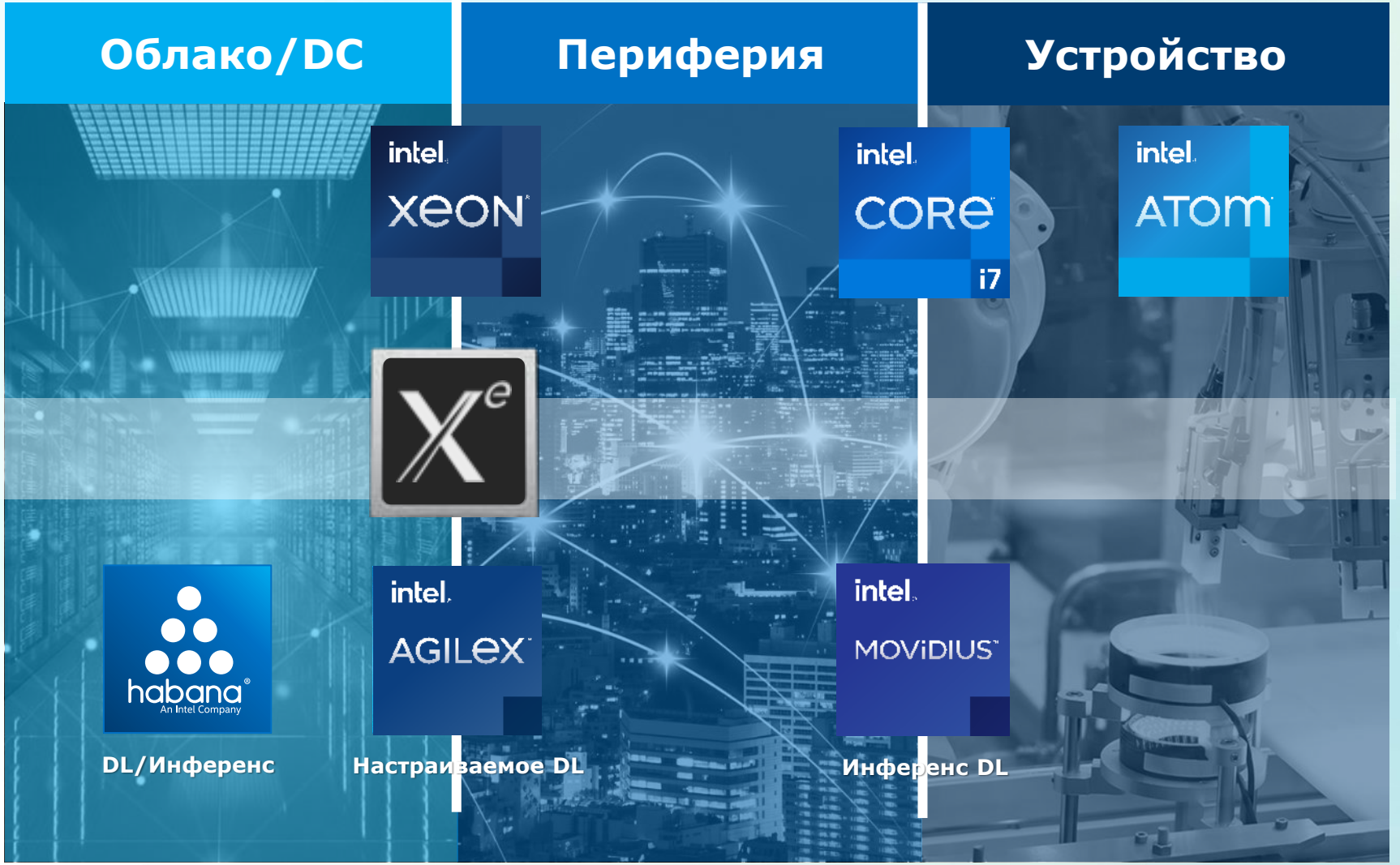
Обеспечение ИИ вычислений от облака до устройства

FLEXIBLE ACCELERATION

Только CPU
Для типичных задач ИИ

CPU + GPU
Когда в вычислениях преобладает ИИ, HPC, графика и/или мультимедиа в реальном времени

CPU + CUSTOM
Когда в вычислениях преобладает глубокое обучение (DL)



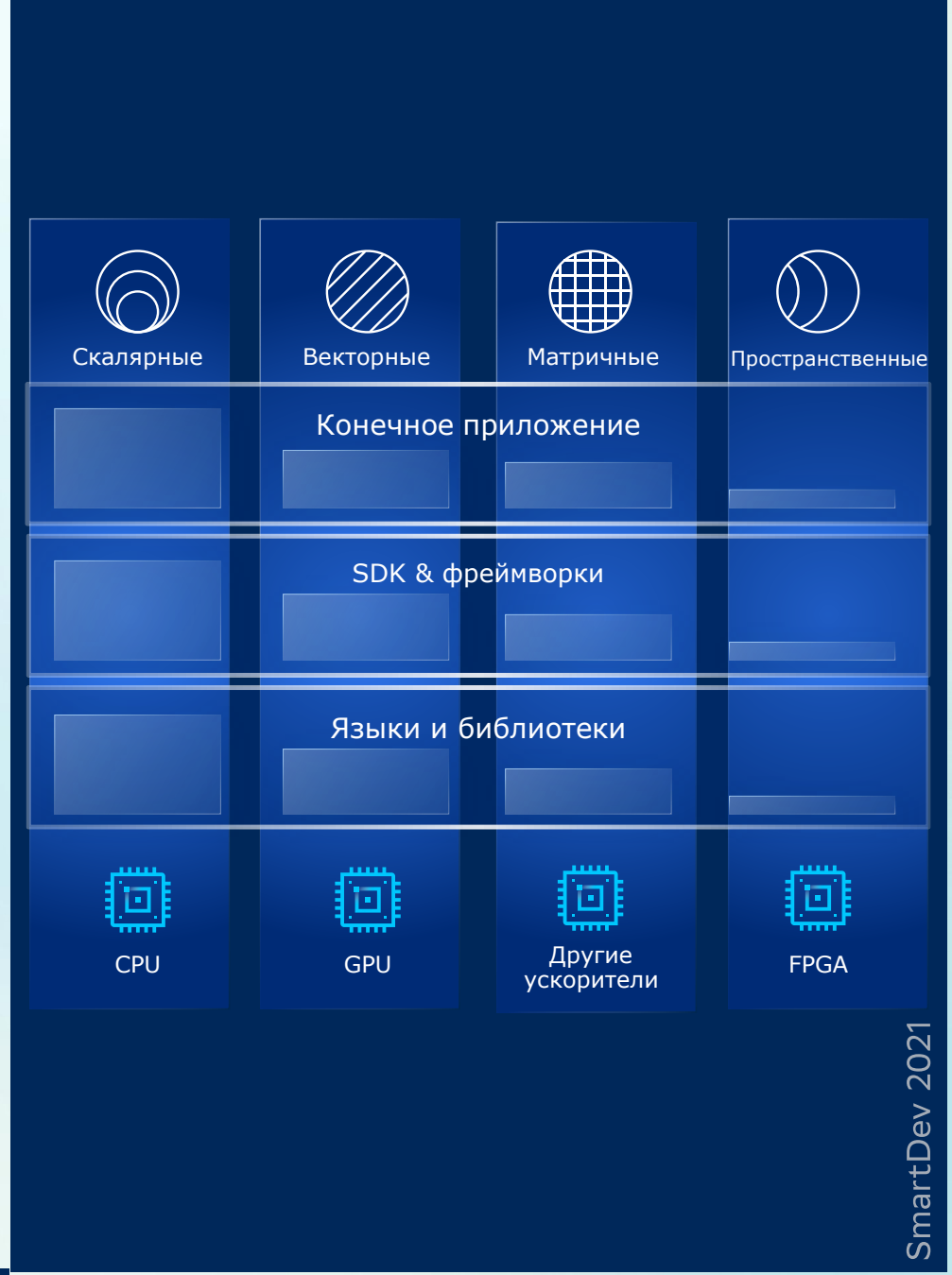
DC = Дата-центр
DL = Глубокое обучение

Обеспечение ИИ вычислений от облака до устройства

Сложности программирования

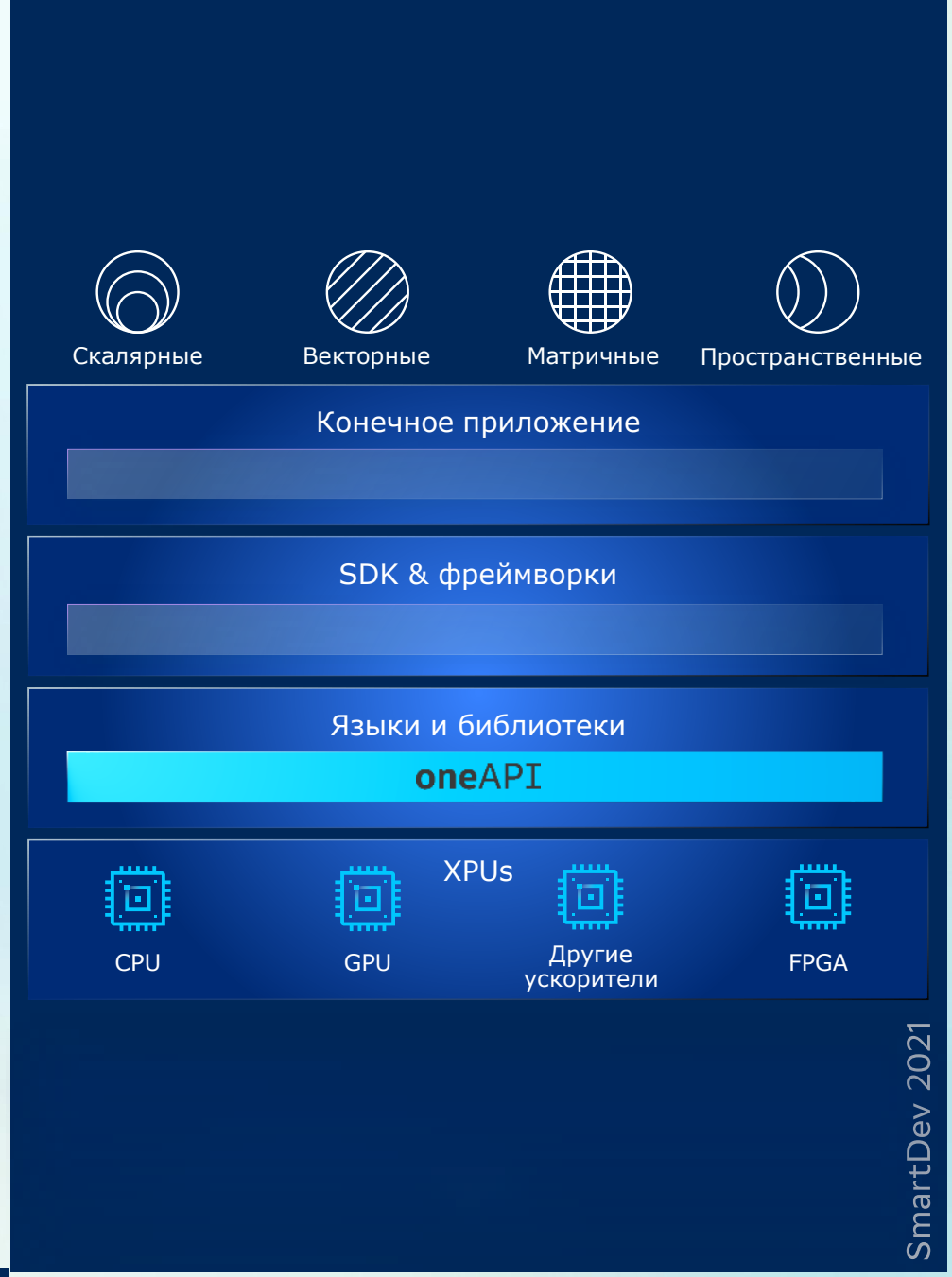
для различных архитектур

- Увеличение объемов компонентов разных типов
- Требуются специализированные ускорители
- Для каждой архитектуры необходимы специализированные модели программирования и инструменты
- Сложность разработки программного обеспечения ограничивает свободу выбора архитектуры



Представляем Intel® oneAPI

- Кросс-архитектурное решение позволяет использовать оптимальное аппаратное обеспечение
- Основано на отраслевых стандартах и открытых спецификациях
- Использует передовые функциональные возможности новейшего оборудования
- Совместимо с существующими высокопроизводительными языками и моделями программирования, включая C++, OpenMP, Fortran и MPI



Посетите сайт oneapi.com, чтобы узнать больше деталей

Отраслевая инициатива oneAPI

Разорвите привязку к вендору

- Кросс-архитектурный язык, основанный на стандартах C++ и SYCL
- Эффективные библиотеки, предназначенные для ускорения предметно-ориентированных функций
- Уровень абстракции низкоуровневого аппаратного обеспечения
- **Открытый стандарт для использования сообществом и промышленностью**
- **Обеспечивает возможность адаптации кода для различных архитектур и вендоров**

SmartDev 2021

Продуктивный и умный путь к освобождению ускоренных вычислений от экономического и технического бремени проприетарных моделей программирования

Посетите сайт oneapi.com, чтобы узнать больше деталей

Data Parallel C++

Стандартизированный, кросс-архитектурный язык DPC++ = ISO C++ и Khronos SYCL

Параллелизм, производительность и быстродействие для процессоров и ускорителей

- Обеспечивает ускоренное вычисление за счет демонстрации аппаратных возможностей
- Позволяет повторно использовать код на аппаратных объектах, а также адаптировать его для конкретных ускорителей
- Обеспечивает открытое межотраслевое применение для проприетарных закрытых решений

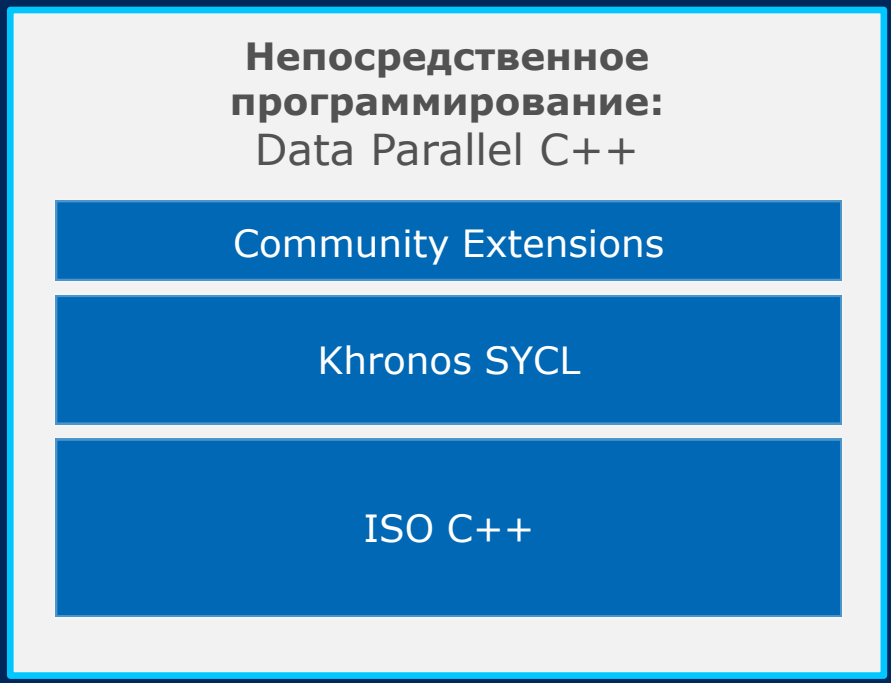
На основе C++ и SYCL

- Предоставляет преимущества в производительности C++, используя общие знакомые конструкции C и C++
- Включает SYCL от Khronos Group для поддержки параллелизма данных и гетерогенного программирования

Community Project для улучшения языка

- Предоставляет расширения для упрощения программирования с распараллеливанием данных
- Постоянное развитие на основе открытого сотрудничества

Непосредственное программирование:
Data Parallel C++



SmartDev 2021

Вместо того, чтобы в очередной раз переписывать код для новой платформы, вы сможете применить свои навыки для создания инноваций.

Открытый исходный код и поддержка компиляторов Intel DPC++/C++ для CPU, GPU и FPGA от Intel.
Codeplay объявили о выпуске компилятора DPC++ для GPU от Nvidia.

Мощные библиотеки oneAPI

- Предназначены для ускорения предметно-ориентированных функций
- Уже оптимизированы для платформ Intel для максимальной производительности

SmartDev 2021

Посетите сайт oneapi.com, чтобы узнать больше деталей

SBER | intel

oneAPI: поддержка экосистемы

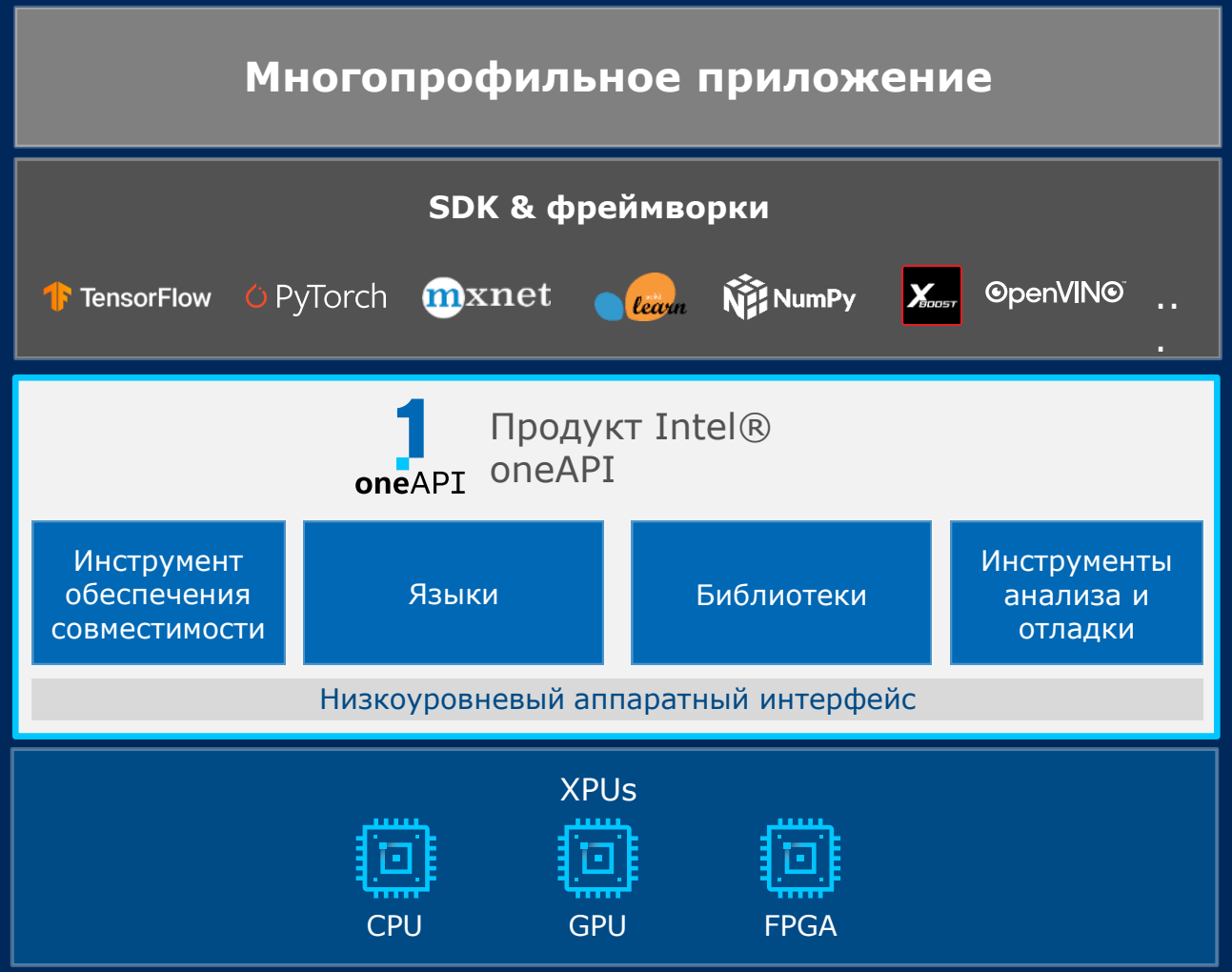
Перечисленные организации поддерживают концепцию инициативы OneAPI для единой, унифицированной кросс-архитектурной модели программирования. Это не подразумевает соглашения о покупке или использовании продукции Intel. *Прочие названия и бренды могут являться собственностью других компаний.

Продукт Intel® oneAPI

Создан на основе богатого опыта Intel в разработке CPU Адаптирован для xPU

Полный набор продвинутых компиляторов, библиотек, а также инструментов для портирования, анализа и отладки.

- Ускоряет вычисления за счет использования передовых аппаратных возможностей
- Работая с существующими моделями программирования и кодовыми базами (C++, Fortran, Python, OpenMP и т.д.), разработчики могут быть уверены в полной совместимости с oneAPI
- Облегчает переход на новые системы и ускорители единообразным способом, давая разработчикам возможность уделять больше времени инновациям



Доступно уже сейчас

SmartDev 2021

Посетите сайт oneapi.com, чтобы узнать больше деталей

Некоторые возможности могут отличаться в зависимости от архитектуры, требуется индивидуальная настройка.
В дальнейшем планируется поддержка других ускорителей.

Инструменты для анализа и отладки

Эффективное использование аппаратных мощностей

Проектирование

Intel® Advisor

- Эффективный оффлоад кода на GPU
- Оптимизация вашего кода CPU/GPU
- Обеспечение большей степени векторизации и параллелизма и повышение эффективности
- Анализ возможностей применения многопоточности в однопоточных приложениях

Отладка

Intel® Distribution for GDB

- Поддержка нескольких ускорителей с эмуляцией CPU, GPU, FPGA
- Глубокая общесистемная отладка кода Data Parallel C++ (DPC++), C, C++ и Fortran.

Настройка

Intel® VTune™ Profiler

- Анализ DPC++
- Настройка для GPU, CPU и FPGA
- Оптимизация производительности выгрузки
- Поддержка DPC++, C, C++, Fortran, Python, Go, Java и комбинации языков

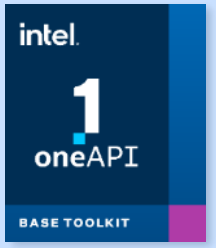
Используйте передовые инструменты для эффективной отладки и профилирования кода на всех уровнях абстракции.

Intel® oneAPI Toolkits

Полный набор проверенных инструментов для разработчиков с поддержкой CPU и xPU

Intel® oneAPI **Base** Toolkit

Разработчики платформенно-ориентированного кода



Основной набор высокопроизводительных инструментов для создания приложений на C++, Data Parallel C++ и приложений на базе библиотек oneAPI

Добавление объектно-специфического инструментария

Специализированные рабочие нагрузки

Intel® oneAPI Tools for HPC

Разрабатывайте масштабируемые приложеняй на Fortran, OpenMP и MPI

Intel® oneAPI Tools for IoT

Создавайте эффективные и надежные решения, работающие на локальных устройствах

Intel® oneAPI Rendering Toolkit

Создавайте эффективные визуальные приложения фотореалистичного качества

oneAPI Toolkits

Специалисты по обработке данных и разработчики ИИ

Intel® AI Analytics Toolkit

Ускорение машинного обучения и передачи данных с помощью оптимизированных фреймворков DL и высокопроизводительных библиотек Python

Intel® Distribution of OpenVINO™ Toolkit

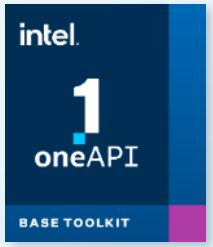
Создание высокопроизводительных приложений на всех платформах, от облака до периферии

Intel® oneAPI Toolkits

Полный набор проверенных инструментов для разработчиков с поддержкой CPU и xPU

Intel® oneAPI **Base** Toolkit

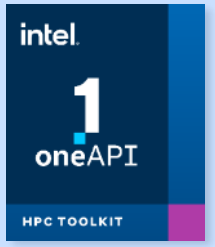
Разработчики платформенно-ориентированного кода



Основной набор высокопроизводительных инструментов для создания приложений на C++, Data Parallel C++ и приложений на базе библиотек oneAPI

Добавление объектно-специфического инструментария

Специализированные рабочие нагрузки



Intel® oneAPI Tools for HPC

Разрабатывайте масштабируемые приложеняй на Fortran, OpenMP и MPI

Intel® oneAPI Tools for IoT

Создавайте эффективные и надежные решения, работающие на локальных устройствах

Intel® oneAPI Rendering Toolkit

Создавайте эффективные визуальные приложения фотореалистичного качества

oneAPI Toolkits

Специалисты по обработке данных и разработчики ИИ

Intel® AI Analytics Toolkit

Ускорение машинного обучения и передачи данных с помощью оптимизированных фреймворков DL и высокопроизводительных библиотек Python

Intel® Distribution of OpenVINO™ Toolkit

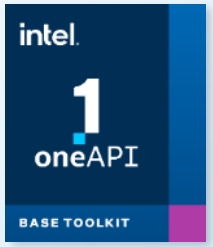
Создание высокопроизводительных приложений на всех платформах, от облака до периферии

Intel® oneAPI Toolkits

Полный набор проверенных инструментов для разработчиков с поддержкой CPU и xPU

Intel® oneAPI **Base** Toolkit

Разработчики платформенно-ориентированного кода



Основной набор высокопроизводительных инструментов для создания приложений на C++, Data Parallel C++ и приложений на базе библиотек oneAPI

Добавление объектно-специфического инструментария

Специализированные рабочие нагрузки

Intel® oneAPI Tools for HPC

Разрабатывайте масштабируемые приложеняй на Fortran, OpenMP и MPI

Intel® oneAPI Tools for IoT

Создавайте эффективные и надежные решения, работающие на локальных устройствах

Intel® oneAPI Rendering Toolkit

Создавайте эффективные визуальные приложения фотореалистичного качества

oneAPI Toolkits

Специалисты по обработке данных и разработчики ИИ

Intel® AI Analytics Toolkit

Ускорение машинного обучения и передачи данных с помощью оптимизированных фреймворков DL и высокопроизводительных библиотек Python

Intel® Distribution of OpenVINO™ Toolkit

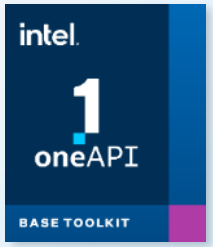
Создание высокопроизводительных приложений на всех платформах, от облака до периферии

Intel® oneAPI Toolkits

Полный набор проверенных инструментов для разработчиков с поддержкой CPU и xPU

Intel® oneAPI **Base** Toolkit

Разработчики платформенно-ориентированного кода



Основной набор высокопроизводительных инструментов для создания приложений на C++, Data Parallel C++ и приложений на базе библиотек oneAPI

Добавление объектно-специфического инструментария

Специализированные рабочие нагрузки

Intel® oneAPI Tools for HPC

Разрабатывайте масштабируемые приложеняй на Fortran, OpenMP и MPI

Intel® oneAPI Tools for IoT

Создавайте эффективные и надежные решения, работающие на локальных устройствах

Intel® oneAPI Rendering Toolkit

Создавайте эффективные визуальные приложения фотореалистичного качества

oneAPI Toolkits

Специалисты по обработке данных и разработчики ИИ

Intel® AI Analytics Toolkit

Ускорение машинного обучения и передачи данных с помощью оптимизированных фреймворков DL и высокопроизводительных библиотек Python

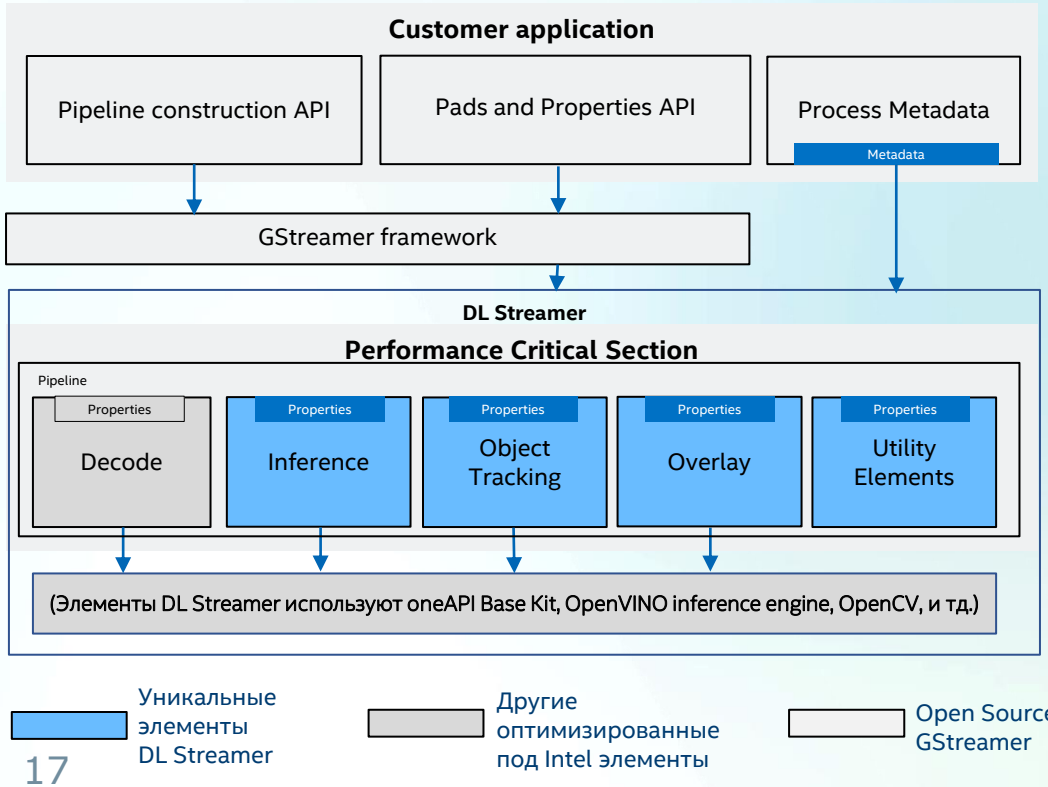
Intel® Distribution of OpenVINO™ Toolkit

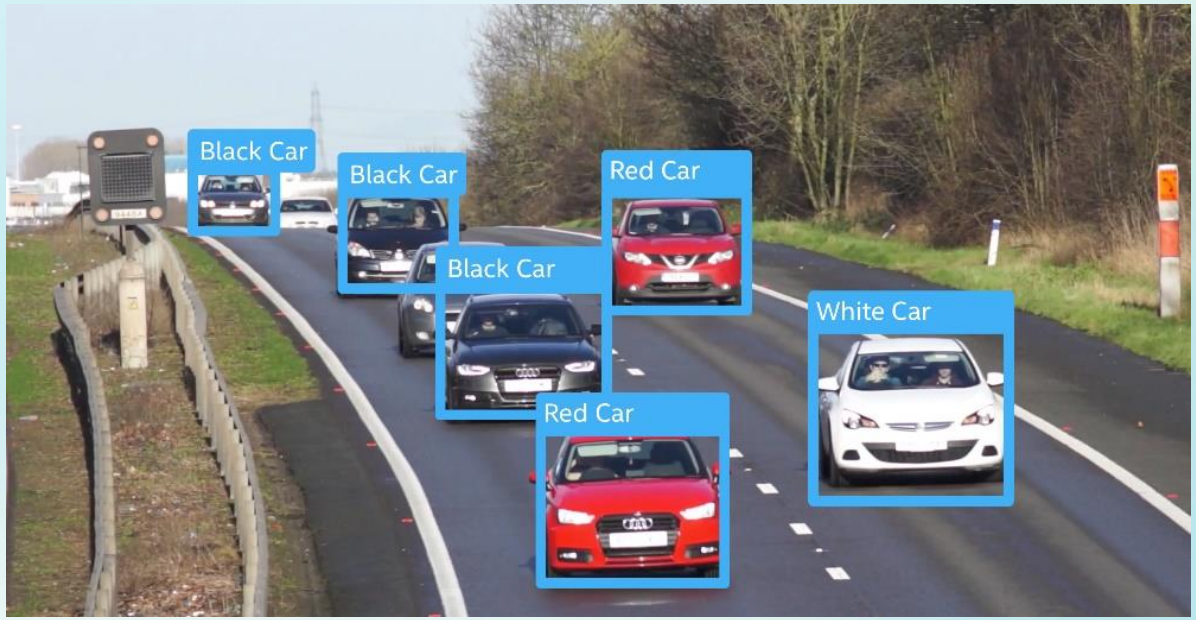
Создание высокопроизводительных приложений на всех платформах, от облака до периферии

OpenVINO™ DL Streamer

DL Streamer – это оптимизированный фреймворк для медиа аналитики, который дополняет опен-сорсный GStreamer* новыми возможностями для ИИ. Он оптимизирован под архитектуры Intel xPU.

С DL Streamer вы пишете меньше кода, и получаете лучшую производительность

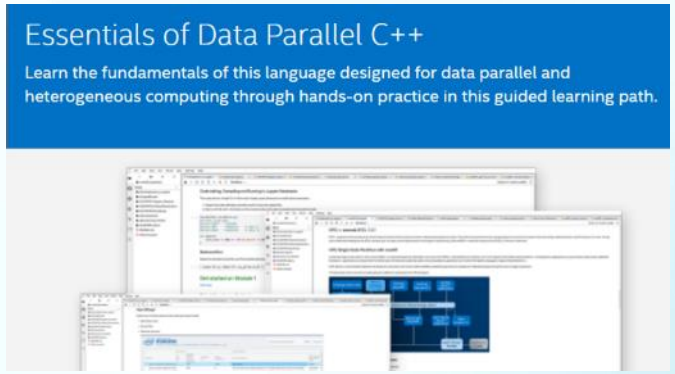




Пример: детекция и классификация автомобилей с визуализацией

Поддержка экосистемы

Обучение

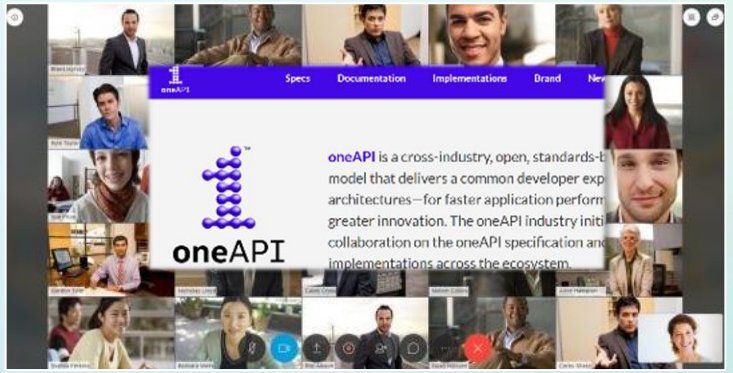


Вебинары и онлайн курсы, руководства для разработчиков, примеры кода

Наука

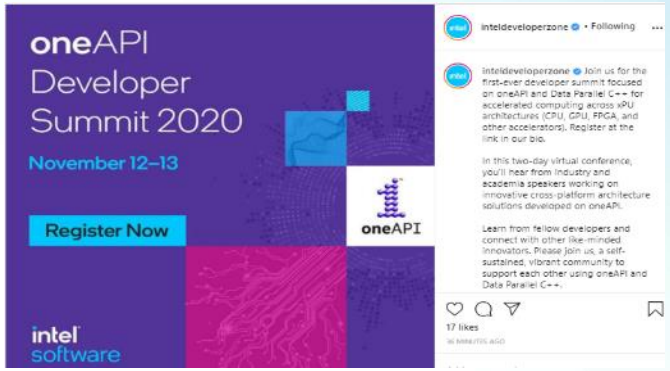
oneAPI Centers of Excellence: исследования, обучение программированию, учебный план, преподавание

Сообщество



Открытая спецификация oneAPI, инновации DevMesh, форумы сообщества

Конференции и семинары



Прямой эфир и записи занятий
Сессии сообществ

Отраслевые эксперты

Intel® DevCloud

Ресурсы oneAPI

software.intel.com/oneapi

Учитесь и работайте

- software.intel.com/oneapi
- Обучение
- Документация
- Примеры кода

oneAPI

Отраслевые инициативы

- oneAPI.com
- Промышленная спецификация
- oneAPI
- Открытые исходные коды

oneAPI

Экосистема

- Форумы сообщества
- Академическая программа
- Инновационные проекты Intel® DevMesh

Footnotes and System Configuration

1. Up to 1.93x higher AI training performance with 3rd Gen Intel Xeon Scalable processor supporting Intel DL Boost with BF16 vs. prior generation on ResNet50 throughput for image classification – New: 1-node, 4x 3rd Gen Intel Xeon Platinum 8380H processor (pre-production 28C, 250W) on Intel Reference Platform (Cooper City) with 384 GB (24 slots / 16GB / 3200) total memory, ucode 0x700001b, HT on, Turbo on, with Ubuntu 20.04 LTS, Linux 5.4.0-26,28,29-generic, Intel 800GB SSD OS Drive, ResNet-50 v 1.5 Throughput, https://github.com/Intel-tensorflow/tensorflow -b bf16/base, commit#828738642760358b388d8f615ded0c213f10c99a, Modelzoo: https://github.com/IntelAI/models/ -b v1.6.1, Imagenet dataset, oneDNN 1.4, BF16, BS=512, test by Intel on 5/18/2020. Baseline: 1-node, 4x Intel Xeon Platinum 8280 processor on Intel Reference Platform (Lightning Ridge) with 768 GB (24 slots / 32GB / 2933) total memory, ucode 0x4002f00, HT on, Turbo on, with Ubuntu 20.04 LTS, Linux 5.4.0-26,28,29-generic, Intel 800GB SSD OS Drive, ResNet-50 v 1.5 Throughput, https://github.com/Inteltensorflow/tensorflow -b bf16/base, commit#828738642760358b388d8f615ded0c213f10c99a, Modelzoo: https://github.com/IntelAI/models/ -b v1.6.1, Imagenet dataset, oneDNN 1.4, FP32, BS=512, test by Intel on 5/18/2020.
2. Up to 1.92x higher performance on cloud data analytics usage models with the new 3rd Gen Intel Xeon Scalable processor vs. 5-year old 4-socket platform – New: 1-node, 4x 3rd Gen Intel Xeon Platinum 8380H processor (pre-production 28C, 250W) on Intel Reference Platform (Cooper City) with 1536GB (48 slots / 32 GB / 3200 (@2933) total memory, microcode 0x700001b, HT on, Turbo on, with Ubuntu 18.04.4 LTS, 5.3.0-53-generic, 1x Intel 240GB SSD OS Drive, 4x P4610 3.2TB PCIe NVME, 4 x 40 GbE x710 dual port, CloudXPRT vCP - Data Analytics, Kubernetes, Docker, Kafka, MinIO, Prometheus, XGBoost workload, Higgs dataset, test by Intel on 5/27/2020. Baseline: 1-node, 4x Intel Xeon processor E7-8890 v3 on Intel Reference Platform (Brickland) with 1024 GB (64 slots / 16GB / 1600) total memory, microcode 0x0000016, HT on, Turbo on, with Ubuntu 18.04.4 LTS, 5.3.0-53-generic, 1x Intel 400GB SSD OS Drive, 4x P3700 2TB PCIe NVME, 4 x 40 GbE x710 dual port, CloudXPRT vCP - Data Analytics, Kubernetes, Docker, Kafka, MinIO, Prometheus, XGBoost workload, Higgs dataset, test by Intel on 5/27/2020.
3. Up to 1.7x more AI training performance with 3rd Gen Intel Xeon Scalable processor supporting Intel DL Boost with BF16 vs. prior generation on BERT throughput for natural language processing – New: 1-node, 4x 3rd Gen Intel Xeon Platinum 8380H processor (pre-production 28C, 250W) on Intel Reference Platform (Cooper City) with 384 GB (24 slots / 16GB / 3200) total memory, ucode 0x700001b, HT on, Turbo on, with Ubuntu 20.04 LTS, Linux 5.4.0-26,28,29-generic, Intel 800GB SSD OS Drive, BERT-Large (QA) Throughput, https://github.com/Intel-tensorflow/tensorflow -b bf16/base, commit#828738642760358b388d8f615ded0c213f10c99a, Modelzoo: https://github.com/IntelAI/models/ -b v1.6.1, Squad 1.1 dataset, oneDNN 1.4, BF16, BS=12, test by Intel on 5/18/2020. Baseline: 1-node, 4x Intel Xeon Platinum 8280 processor on Intel Reference Platform (Lightning Ridge) with 768 GB (24 slots / 32GB / 2933) total memory, ucode 0x4002f00, HT on, Turbo on, with Ubuntu 20.04 LTS, Linux 5.4.0-26,28,29- generic, Intel 800GB SSD OS Drive, BERT-Large (QA) Throughput, https://github.com/Inteltensorflow/tensorflow -b bf16/base, commit#828738642760358b388d8f615ded0c213f10c99a, Modelzoo: https://github.com/IntelAI/models/ -b v1.6.1, Squad 1.1 dataset, oneDNN 1.4, FP32, BS=12, test by Intel on 5/18/2020.
4. AliCloud PAI Customized TextCNN on TF1.14 Run Time Performance on 3rd Gen Intel Xeon Scalable Processor: New: Tested by Intel as of 4/23/2020. 4 socket 3rd Generation Intel Xeon Processor Scalable Family (Ali Customized SKU) Processor using Intel Reference Platform, 24 cores HT On Turbo ON Total Memory 384 GB (24 slots/ 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (microcode:0x7000017), Storage: Intel SSDPE2KX010T7, NIC: 2x Intel Ethernet Controller 10G X550T, OS: CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: TF1.14 https://pypi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb6be6148ec4f971/intel_tensorflow-1.14.0-cp36-cp36m-manylinux1_x86_64.whl, Compiler: gcc 8.3.1, oneDNN version: DNNLv1.3, Customized TextCNN(Confidential), BS=32, Dummy data, 4 instances/4 socket, Datatype: BF16
   Baseline: Tested by Intel as of 4/23/2020. 4 socket 3rd Generation Intel Xeon Processor Scalable Family (Ali Customized SKU) Processor, using Intel Reference Platform 24 cores HT On Turbo ON Total Memory 384 GB (24 slots / 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (microcode:0x7000017), Storage: Intel SSDPE2KX010T7, NIC: 2x Intel Ethernet Controller 10G X550T, OS: CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: TF1.14 https://pypi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb6be6148ec4f971/intel_tensorflow-1.14.0-cp36-cp36m-manylinux1_x86_64.whl, Compiler: gcc 8.3.1, MKL version: 2020.1.217, Customized TextCNN(Confidential), BS=32, Dummy data, 4 instances/4 socket, Datatype: FP32
5. AliCloud PAI Customized BERT on TF1.14 Latency Performance on 3rd Gen Intel Xeon Scalable Processor:
   New: Tested by Intel as of 4/23/2020. 4 socket Intel Xeon Platinum 83xxH (Ali Customized SKU) Processor using Intel Reference Platform, 24 cores HT On Turbo ON Total Memory 384 GB (24 slots/ 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (microcode:0x7000017), Storage: Intel SSDPE2KX010T7, NIC: 2x Intel ethernet Controller 10G x550T, OS: CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: TF1.14 https://pypi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb6be6148ec4f971/intel_tensorflow-1.14.0-cp36-cp36m-manylinux1_x86_64.whl, Compiler: gcc 8.3.1, oneDNN version: DNNLv1.3, Customized BERT(Confidential), BS=1, MRPC data, 12 instance/4 socket, Datatype: BF16
   Baseline: Tested by Intel as of 4/23/2020. 4 socket Intel Xeon Platinum 83xxH (Ali Customized SKU) Processor using Intel Reference Platform, 24 cores HT On Turbo ON Total Memory 384 GB (24 slots / 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (microcode:0x7000017), Storage: Intel SSDPE2KX010T7, NIC: 2x Intel ethernet Controller 10G x550T, OS:CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: TF1.14 https://pypi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb6be6148ec4f971/intel_tensorflow-1.14.0-cp36-cp36m-manylinux1_x86_64.whl, Compiler: gcc 8.3.1, MKL version: 2020.1.217, Customized BERT(Confidential), BS=1, MRPC data, 12 instance/4 socket, Datatype: FP32
6. Alibaba Ant Financial Inference and Training on 3rd Gen Intel Xeon Scalable Processor: Tested by Intel as of 4/20/2020. 4 socket 3rd Gen Intel Xeon Scalable processor (18-core, 170W, pre-production) Processor using Intel Reference Platform, 18 cores HT OFF, Turbo ON Total Memory 768 GB (24 slots / 32GB / 2666), BIOS Version: 166.08 (6BC51780-BFDE-1000-03E6-000000000000) Microcode: 0x8600000b, CentOS 7.7.1908, 3.10.0-957.el7.x86_64, Deep Learning Framework: Pytorch Intel optimized Pytorch-1.0.0a0+3ca7205 https://gitlab.devtools.intel.com/cce-ai/pytorch, dnnl (mkldnn) commit id:7b53785 https://github.com/oneapi-src/oneDNN, Model: 3d CNN I3D, Compiler: gcc 7.3.1, Libraries: dnnl (mk-dnn), Dataset: UCF101 (size: 13320 shape: 3x64x224x224, Baseline Training: BS=24*4, FP32, New Training: BS=24*4, BF16; Baseline Inference: BS=32, 4 instances/4sockets, FP32, New Inference: BS=32, 4 instances/ 4 sockets, BF16.
7. Hisign Facial Recognition Throughput Performance on 3rd Gen Intel Xeon Scalable Processor:
   NEW: Tested by Intel as of 5/15/2020. 1-node, 4x Intel Xeon Platinum 8380H (pre-production) Processor on Intel Reference Platform, 28 cores HT On Turbo ON Total Memory 768 GB (24 slots/ 32GB/ 3200 MHz), BIOS: WLYDCRB1.SYS.0015.D19.2002140555 (microcode: 0x87000016), NIC: Intel X550T; Storage: 1x Intel 800GB SSD, OS: RedHat 8.1, 4.18.0-147.8.1.el8_1.x86_64, Framework: Internal Tensorflow 2.1 Branch: UTB, Commit id: 4c711446a4d42fa1ef8759602345fb75f50154ee, ssh://git@gitlab.devtools.intel.com:29418/TensorFlow/Direct-Optimization/private-tensorflow.git, Topology/ML Algorithm: customized FaceResNet, Compiler: GCC 8.3.1, MKL DNN, Dataset: Customer provided 4906 images, 128x128x3, Precision: BF16
   BASELINE: Tested : 0x400002C), NIC: Intel X550T; Storage: 1x Intel 800GB SSD, OS: RedHatby Intel as of 5/15/2020. 1-node, 4x Intel Xeon Platinum 8280L Processor on Inspur NF8260M5, 28 cores HT On Turbo ON Total Memory 768 GB (24 slots/ 32GB/ 2933 MHz), BIOS: Inspur 4.1.10 (microcode 8.1, 4.18.0-147.8.1.el8_1.x86_64, Framework: Internal Tensorflow 2.1 Branch: UTB, Commit id: 4c711446a4d42fa1ef8759602345fb75f50154ee, ssh://git@gitlab.devtools.intel.com:29418/TensorFlow/Direct-Optimization/private-tensorflow.git, Topology/ML Algorithm: customized FaceResNet, Compiler: GCC 8.3.1, MKL DNN, Dataset: Customer provided 4906 images, 128x128x3, Precision: FP32

Footnotes and System Configuration

8. 1.86x ResNet-50 Training Throughput Performance Improvement on Catalina platform with BF16: 1-node, 8x 3rd Gen Intel Xeon Platinum 8380H processor(28C) on Catalina with 768 GB (48 slots / 16GB / 3200) total memory, microcode 0x86000017, HT on, Turbo on, Ubuntu 20.04 LTS(Host | Ubuntu 18.04 (Docker) Kernel 5.4.0-28-generic (Host), 1x INTEL_SSDSC2BX01, 8x Intel E810-C, ResNet-50 v 1.5 Throughput, Intel optimized TensorFlow 2.2, https://github.com/Intel-tensorflow/tensorflow/commits/bf16/base, https://github.com/IntelAI/models/blob/v1.6.1/models/image_recognition/tensorflow/ResNet50v1_5/training/mlperf_resnet/resnet_model.py, gcc version 7.5.0 (docker) , ImageNet Challenge 2012 Dataset, oneDNN v1.4, FP32 and BF16, test by Intel on 05/24/2020, *16-node projected performance
9. TensorFlow on Neusoft Pathology Inference Throughput Performance on 3rd Gen Intel Xeon Scalable Processor:
NEW: Tested by Intel as of 5/15/2020. 1-node, 4x Intel Xeon Platinum 8380H (pre-production) Processor on Intel Reference Platform, 28 cores HT On Turbo ON Total Memory 768 GB (24 slots/ 32GB/ 3200 MHz), BIOS: WLYDCRB1.SYS.0015.D19.2002140555 (microcode: 0x87000016), NIC: Intel X550T; Storage: 1x Intel 800GB SSD, OS: RedHat 8.1, 4.18.0-147.8.1.el8_1.x86_64, Framework: Internal Tensorflow 2.1 Branch: UTB, Commit id: 4c711446a4d42fa1ef8759602345fb75f50154ee, ssh://git@gitlab.devtools.intel.com:29418/TensorFlow/Direct-Optimization/private-tensorflow.git, Topology/ML Algorithm: customized DNN topology, Compiler: GCC 8.3.1, MKL DNN, Dataset: Customer provided 1728 images, 32x32x3, Precision: BF16
BASELINE: Tested by Intel as of 5/15/2020. 1-node, 4x Intel Xeon Platinum 8280L Processor on Inspur NF8260M5, 28 cores HT On Turbo ON Total Memory 768 GB (24 slots/ 32GB/ 2933 MHz), BIOS: Inspur 4.1.10 (microcode: 0x400002C), NIC: Intel X550T; Storage: 1x Intel 800GB SSD, OS: RedHat 8.1, 4.18.0-147.8.1.el8_1.x86_64, Framework: Internal Tensorflow 2.1 Branch: UTB, Commit id: 4c711446a4d42fa1ef8759602345fb75f50154ee, ssh://git@gitlab.devtools.intel.com:29418/TensorFlow/Direct-Optimization/private-tensorflow.git, Topology/ML Algorithm: customized DNN topology, Compiler: GCC 8.3.1, MKL DNN, Dataset: Customer provided 1728 images, 32x32x3, Precision: FP32
10. Tencent Search Engine Customized NLP model on TF1.14 Throughput Performance on 3rd Generation Intel Xeon Scalable Processor:
New: Tested by Intel as of 4/28/2020. 4 socket 3rd Generation Intel Xeon Processor Scalable Family(CPX pre-production SKU) Processor, 26 cores HT On Turbo ON Total Memory 384 GB (24 slots/ 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (microcode:0x86000017), CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: TF1.14 https://pypi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb6be6148ec4f971/intel_tensorflow-1.14.0-cp36-cp36m-manylinux1_x86_64.whl, Compiler: gcc 8.3.1, OneDNN version: DNNLv1.3, Customized NLP model(Confidential), BS=1, MRPC data, 8 instances/4 socket, Datatype: BF16
Baseline: Tested by Intel as of 4/28/2020. 4 socket 3rd Generation Intel Xeon Processor Scalable Family (CPX pre-production SKU) Processor, 26 cores HT On Turbo ON Total Memory 384 GB (24 slots / 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (microcode:0x86000017),CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: TF1.14 https://pypi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb6be6148ec4f971/intel_tensorflow-1.14.0-cp36-cp36m-manylinux1_x86_64.whl, Compiler: gcc 8.3.1, OneDNN version: DNNLv1.3, Customized NLP model(Confidential), BS=1, MRPC data, 8 instances/4 socket, Datatype: FP32
11. Tencent Cloud Xiaowei Customized WaveRNN on MXNetv1.7 Throughput Performance on 3rd Generation Intel Xeon Scalable Processor:
Opt. BF16 Solution: Tested by Intel as of 4/28/2020. 4 socket 3rd Generation Intel Xeon Processor Scalable Family(CPX pre-production SKU) Processor, 26 cores HT On Turbo ON Total Memory 384 GB (24 slots/ 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (microcode:0x86000017), CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: MXNet1.7 https://github.com/apache/incubator-mxnet/tree/v1.7.x, Compiler: gcc 8.3.1, oneDNN version: DNNLv1.3, Customized WaveRNN(Confidential), BS=1 , Customer Provided data, 104 Instances/4 socket, Datatype: BF16
BASELINE(Opt. FP32 Solution): Tested by Intel as of 4/28/2020. 4 socket 3rd Generation Intel Xeon Processor Scalable Family(CPX pre-production SKU) Processor, 26 cores HT On Turbo ON Total Memory 384 GB (24 slots / 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (microcode:0x86000017),CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: MXNet1.7 https://github.com/apache/incubator-mxnet/tree/v1.7.x, Compiler: gcc 8.3.1, oneDNN version: DNNLv1.3, Customized WaveRNN(Confidential), BS=1, Customer Provided data, 104 Instances/4 socket, Datatype: FP32
12. Tencent Cloud Xiaowei TTS P_Wavenet on TF1.14 Run Time Performance on 3rd Generation Intel Xeon Scalable Processor:
New: Tested by Intel as of 5/11/2020. 4 socket 3rd Generation Intel Xeon Processor Scalable Family(CPX pre-production SKU) Processor, 26 cores HT On Turbo ON Total Memory 384 GB (24 slots/ 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (microcode:0x86000017), CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: TF1.14 https://pypi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb6be6148ec4f971/intel_tensorflow-1.14.0-cp36-cp36m-manylinux1_x86_64.whl, Compiler: gcc 8.3.1, oneDNN version: DNNLv1.3, Customized TTS Pwavenet(Confidential), BS=1, Customer Provided data, 4 instances/4 Socket, Datatype: BF16
Baseline: Tested by Intel as of 5/11/2020. 4 socket 3rd Generation Intel Xeon Processor Scalable Family (CPX pre-production SKU) Processor, 26 cores HT On Turbo ON Total Memory 384 GB (24 slots / 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (microcode:0x86000017),CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: TF1.14 https://pypi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb6be6148ec4f971/intel_tensorflow-1.14.0-cp36-cp36m-manylinux1_x86_64.whl, Compiler: gcc 8.3.1, oneDNN version: DNNLv1.3, Customized TTS Pwavenet(Confidential), BS=1, Customer Provided data, 4 instances/4 Socket, Datatype: Datatype: FP32
13. Intel® Agilex™ FPGA + Quartus Prime 20.4 Software FPGA performance made flexible ~2x Better Fabric Performance per Watt vs. Versal: The Agilex and Versal devices (part number/speed grade) used in the perf/watt comparison are as follows: Agilex: AGF014-2, Versal: Equivalent density to AGF014-2 in 2M speed grade,ted March 2021 by Intel. Design profile used for the comparison: Base Stratix 10 frequency: 450MHz, Agilex Fmax = 450 * 1.59 = 716Mhz, Versal Fmax = 450 * 1.19 = 536Mhz, Resource usage: 60% of AGF014 resource (logic, M20K memory, DSP), power at the respective Fmax; Version of all the tools used for this data :Agilex: Quartus 20.4/PTC 21.1 b149, Versal :Vivado 2020.2/XPE: 2020.2
14. 50% faster Video IP performance: Derived from a set of five video IP designs comparing Fmax of each design achieved in Xilinx Versal ACAP devices with the Fmax achieved in Intel® Agilex™ devices, using Intel® Quartus® Prime Software (version 20.4) and Xilinx Vivado Software (version 2020.2). On geomean, designs running in the mid speed grade of Intel® Agilex™ FPGAs achieve a 50% higher in Fmax compared to the same designs running in the mid speed grade of Xilinx Versal devices (-2M speed grade), and 42% higher in Fmax compared to the same designs running in the fast speed grade of Xilinx Versal devices (-2H speed grade) and 24% higher in Fmax compared to the same designs running in the mid speed grade of Xilinx 16nm VUP devices (-2 speed grade) , tested January 2021. Up to 49% faster fabric performance compared to prior generation FPGA for high-speed 5G fronthaul gateway applications –Derived from comparing the Fmax result of Agilex FPGA and Stratix 10 FPGA in a fronthaul gateway reference example using Quartus Prime 20.4 software, tested in February, 2021. Software Configurations: Tests were done by running internal builds of Intel® Quartus® Prime Pro Design Software on a wide variety of internal benchmarks. The computer systems used for the evaluations were Intel® Skylake CPU @ 3.3GHz 256G Memory class machines running SUSE Linux Enterprise Server 12 operating system. The performance results represent average improvements across a wide variety of internal benchmarks, and results may vary for each testcase. Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates.

Краткое изложение

- Разнообразие рабочих нагрузок приводит к потребности в гетерогенных вычислительных архитектурах, но каждая архитектура требовала отдельных моделей программирования.
- Модель кросс-архитектурного программирования oneAPI обеспечивает свободу выбора. Вместо того, чтобы в очередной раз переписывать код для новой платформы, вы сможете применить свои навыки для создания инноваций.
- Продукты Intel® oneAPI в полной мере используют преимущества ускоренных вычислений, максимизируя производительность CPU, GPU и FPGA от Intel.
- Быстрая и эффективная разработка благодаря полному набору кросс-архитектурных библиотек и продвинутых инструментов, которые взаимодействуют с существующими моделями разработки.

Юридическая информация

Использование технологий Intel может потребовать соответствующего оборудования, программного обеспечения или активации обслуживания. Никакая продукция или компоненты не являются абсолютно безопасными. Ваши расходы и результаты могут варьироваться. Оптимизация Intel для компиляторов и другой продукции может не осуществляться в той же мере для продукции других производителей.

Intel не контролирует содержание и не проводит аудит информации, предоставленной партнерами. Для оценки достоверности такой информации вам следует проверять другие источники.

intel®