oneAPI

2023 SIAM Conference on Computational Science and Engineering

Intel® Developer Tools for Serious SYCL

Nisha Patel (Director of Developer Tools Customer Engineering EMEA, WW customer engagements)
February 2023

intel.

SYCL 2020: *Khronos standard ready to support accelerators*

Based on modern C++ / Decoupled from OpenCL

Unified address space

```
sycl::malloc_host
sycl::malloc_device
sycl::malloc_shared
```

Atomic ops on non-atomic objects

```
sycl::atomic_ref
```

Subgroups (similar to CUDA warp)
Subgroup and work group algorithms

```
Any_of, all_of, non_of, reduce, exclusive_scan,
inclusive_scan, shift_left, shift_right, select, permute
```

Reductions

```
sycl::reduction
```

SYCL 1.2.1

```
float *sum = (float *)malloc(sizeof(float));
float *data = (float *)malloc(N*sizeof(float));
const size_t tc = N/ITEMS_PER_THREAD;
const size_t itb = ITEMS_PER_THREAD*BS;
float *output = (float *)malloc(tc*sizeof(float));
buffer<float, 1> buf(data, N);
buffer<float, 1> ans(output, tc);
q.submit([&](handler &h) {
  auto buf_acc = buf.get_access<access::mode::read>(h);
  auto ans_acc = ans.get_access<access::mode::discard_write>(h);
  h.parallel_for<class SumK>(nd_range<1>(thread_count, BS),
    [=](nd_item<1> it) {
      const size_t st = it.get_group(0)*itb+it.get_local_id(0);
      float lsum = 0;
      for (size_t i = st; i <st+itb; i += it.get_local_range(0)) {
        lsum += buf_acc[i];
      }
      ans_acc[it.get_global_id(0)] = lsum;
    });
}).wait();
auto ans_acc = ans.get_access<access::mode::read>();
*sum = 0;
for (size_t i = 0; i < tc; ++i) { *sum += ans_acc[i]; }
```

SYCL 2020

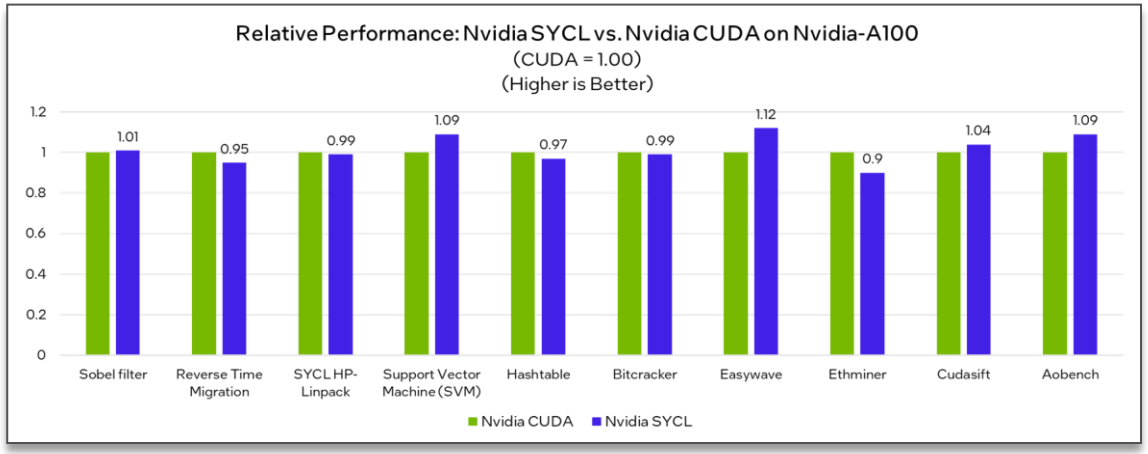
```
float* sum = malloc_shared<float>(1, q);
float* data = malloc_shared<float>(N, q);


q.parallel_for(N, reduction(sum, std::plus<>()),
  [=](size_t i, auto& sum) {
    sum += data[i];
});
```

Slide credit : Paul Peterson @ IISWWC'22

Accelerating Choice with SYCL

Khronos Group Standard

- Open, standards-based
- Multiarchitecture performance
- Freedom from vendor lock-in
- Comparable performance to native CUDA on Nvidia GPUs
- Extension of widely used C++ language
- Speed code migration via open source SYCLomatic or Intel® DPC++ Compatibility Tool



Testing Date: Performance results are **based on testing by Intel as of August 15, 2022** and may not reflect all publicly available updates.

Configuration Details and Workload Setup: Intel® Xeon® Platinum 8360Y CPU @ 2.4GHz, 2 socket, Hyper Thread On, Turbo On, 256GB Hynix DDR4-3200, ucode 0x000363. GPU: Nvidia A100 PCIe 80GB GPU memory. Software: SYCL open source/CLANG 15.0.0, CUDA SDK 11.7 with NVIDIA-NVCC 11.7.64, cuMath 11.7, cuDNN 11.7, Ubuntu 22.04.1. SYCL open source/CLANG compiler switches: -fsycl-targets=nvptx64-nvidia-cuda, NVIDIA NVCC compiler switches: -O3 --gencode arch=compute_80, code=sm_80. Represented workloads with Intel optimizations.

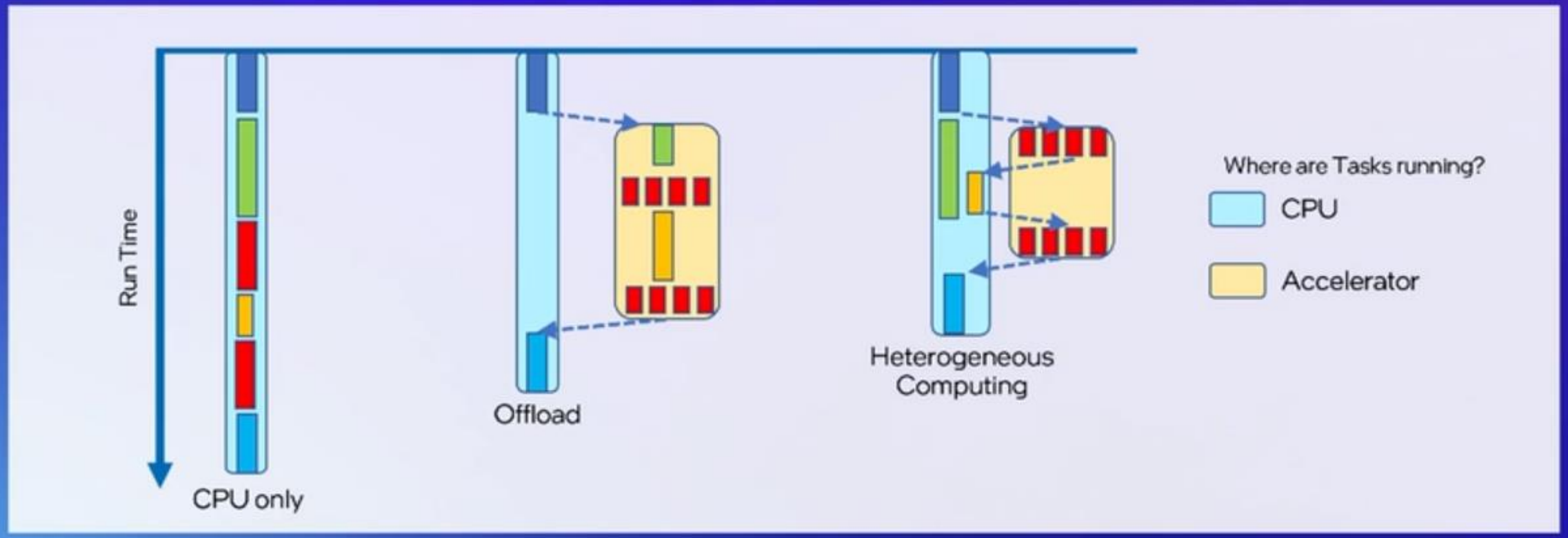
Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See configuration disclosure for details. No product or component can be absolutely secure.

Performance varies by use, configuration, and other factors. Learn more at www.Intel.com/PerformanceIndex. Your costs and results may vary.

Architectures	Intel | Nvidia | AMD CPU/GPU | RISC-V | ARM Mali | PowerVR | Xilinx

Offload vs. Heterogeneous Computing

Heterogeneous computing – not just offload, but leveraging CPUs & accelerators in the same application



Offload: The CPU moves work to an accelerator and waits for the answer.

Heterogeneous Computing: Run sub-problems in parallel on the hardware best suited to them.

Slide credit : Tim Mattson @Intel innovation conference '2022

What is oneAPI?

Open-source software stack

Built with industry standard components
(CLANG, LLVM, SPIR-V)

Intel LLVM	oneAPI Open-Source Projects

An open community

github/oneAPI-TAB

Technical Advisory Boards	Global Support	Implementations	
SYCL	US Dept of Energy	Intel	
oneMKL	Silicon Pearl	NVidia	AMD
oneDNN	Fugaku	Xilinx	ARM

An open specification

oneAPI
oneAPI.io

Building on other open standards
(SYCL 2020, OpenMP, BLAS, ...)

creative commons

Intel Products

oneAPI
software.intel.com/oneapi

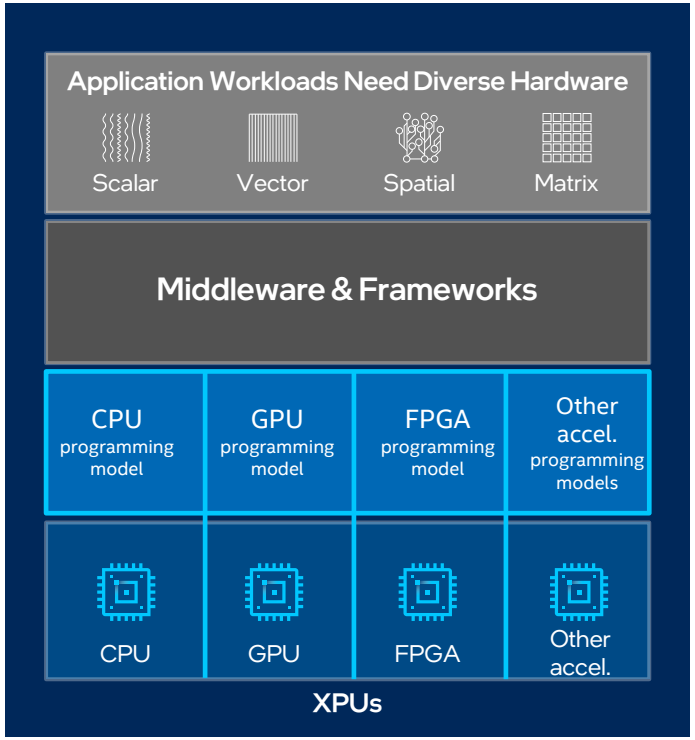
Intel® oneAPI Toolkits
Release 2023

Available on

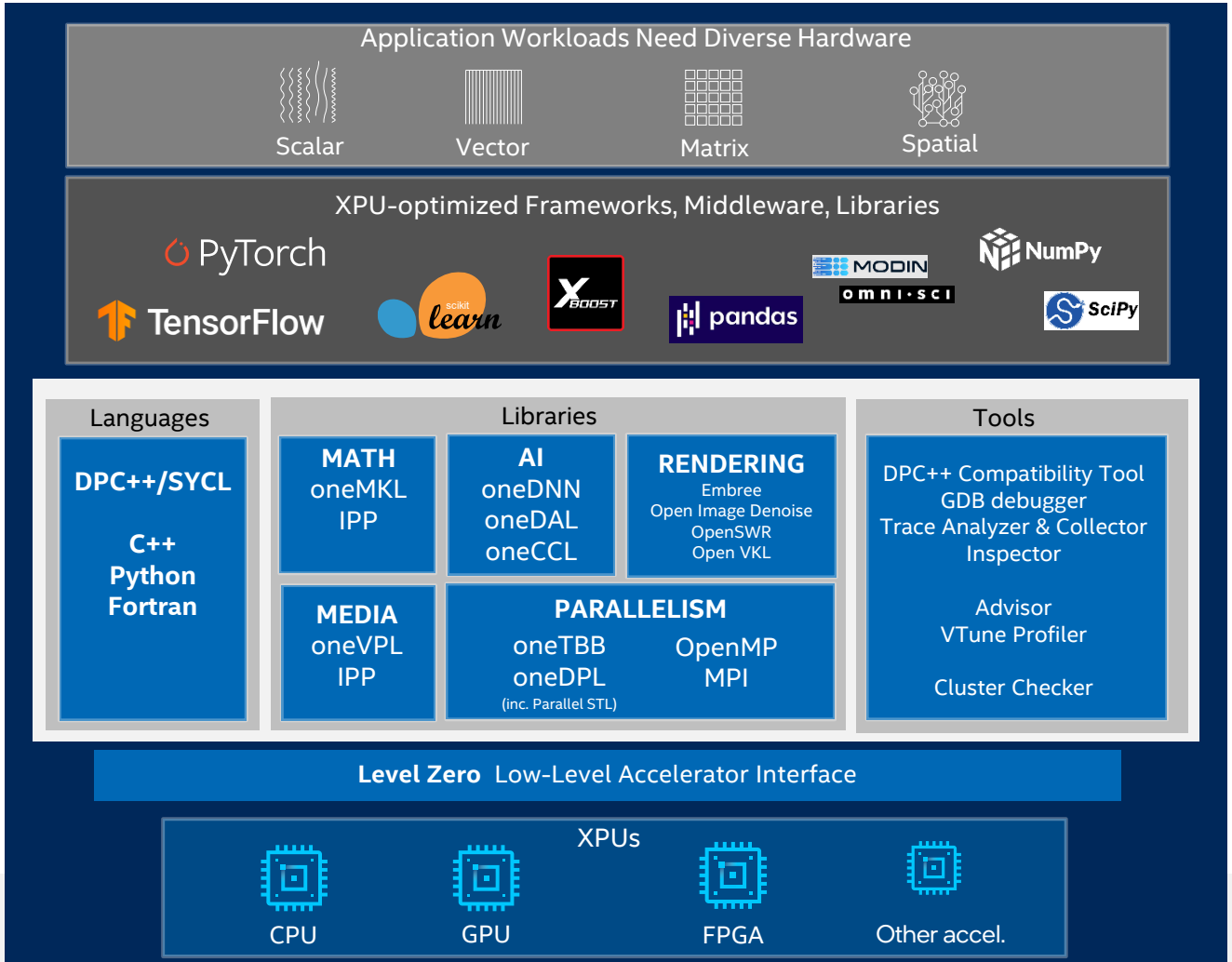
apt · GitHub · CentOS yum/dnf · intel · openSUSE zypper · nuget · CONDA · Maven · Spack

oneAPI enables cross-platform XPU programming

Programming challenges for multiple architectures



oneAPI open specification provides a standard programming language, libraries, and hardware abstraction layer



oneAPI Libraries

Domain	Name	Description	Open Spec	Open Source
Parallel Programming	oneDPL	Data Parallel C++ Library including Parallel STL	Yes	Yes
	oneTBB	Threading Building Blocks	Yes	Yes
AI & ML	oneDNN	Deep Neural Networks	Yes	Yes
	oneCCL	Collective Communications	Yes	Yes
	oneDAL	Data Analytics and Machine learning	Yes	Yes
Math	oneMKL	Math Kernels: linear algebra, FFT, random numger generation	Yes	Partial
Video	oneVPL	Video Processing: encode, decode, transcode	Yes	Yes
Ray Tracing	Embree, VKL, OID, OSPRay	Geometric & Volumetric Ray Tracing, Image Denoise, Scalable Rendering	Yes	Yes

Slide credit : Paul Peterson @ IISWWC'22

Architectures supported

Developer Tools for Intel® Data Center GPU Flex Series

Multiarchitecture Acceleration for Media, Visual Inferencing & Cloud Workloads

Intel® oneAPI Video Processing Library for GPU, and GStreamer + FFmpeg for CPU speed media processing, delivery and cloud gaming streaming.

Intel® oneAPI DPC++/C++ Compiler & oneVPL, both part of the Intel® oneAPI Base Toolkit, scale AI visual inference workloads up to 150 TOPS INT8 between CPU and GPU with the same code.

Intel® VTune™ Profiler analyzes application performance and accelerates compute-intensive tasks by identifying the most time-consuming parts of GPU code and optimizing GPU offload schema and data transfers for SYCL, OpenCL code, Microsoft DirectX*, or OpenMP* offload code.

Intel® Graphics Performance Analyzers optimize visual compute and graphics workloads on CPUs and GPUs.

Intel® OpenVINO™ toolkit, powered by oneAPI, supports popular deep learning frameworks such as TensorFlow and PyTorch to streamline AI visual inferencing and deploy quickly.

Intel® AI Analytics Toolkit, powered by oneAPI, accelerates machine learning & data science pipelines end-to-end with optimized DL & ML frameworks & high-performing Python libraries

intel. 14

Developer Tools for Intel® Data Center GPU Max Series

Build Multiarchitecture Applications with Breakthrough Performance for HPC & AI

Accelerate HPC

oneAPI DPC++/C++ and **Fortran compilers** and **performance libraries (oneMKL, oneDNN, Intel® MPI Library)** activate Xe Matrix Extensions (Intel® XMX) for acceleration and Intel® Xe Link for direct GPU-to-GPU communications to speed up applications running on multiple GPUs.

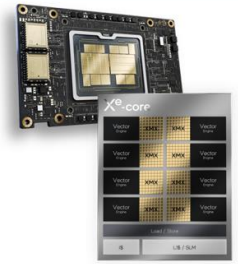
Analysis and debug tools help developers analyze and accelerate compute-intensive tasks by identifying the most time-consuming parts of GPU code, ensure correctness and efficiently debug applications on discrete GPUs. Get actionable advice to design code that runs optimally on Intel GPUs.

Boost AI & Deep Learning Inference

oneDNN in the **Intel® oneAPI Base Toolkit** utilizes Intel XMX to accelerate AI training/inference in deep learning frameworks.

Intel® OpenVINO™ toolkit powered by oneAPI, streamlines AI visual inferencing and accelerates deployment.

Both tools support popular deep learning frameworks such as TensorFlow and PyTorch. GPU support for tools in the Intel® AI Analytics Toolkit is coming soon.



Create Multiarchitecture Code Efficiently with Code Migration Tools

Migrate CUDA code to SYCL to create a single source code base for easy portability across multiple vendors' architectures - including Intel Max Series GPU.

The Intel® DPC++ Compatibility Tool, based on open source SYCLomatic, automates most of the process. Save significant time on ongoing code maintenance.

Coming Soon – High-Performance, High-Fidelity Ray Tracing & Rendering

Intel® Embree, fully ported to C++ with SYCL, will take advantage of hardware ray tracing units and expanded memory capacity to deliver increased rendering performance across larger datasets.

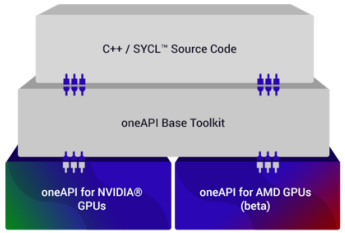
Intel® Open Image Denoise will leverage Intel XMX delivering high fidelity, high-performance denoising capabilities and AI optimization.

intel. 16

oneAPI for NVIDIA and AMD GPUs

Adds support for NVIDIA and AMD GPUs to the Intel oneAPI Base Toolkit.

Develop code using SYCL and run on NVIDIA and AMD GPUs.



Download free plugins from developer.codeplay.com

intel. 17

oneAPI for NVIDIA GPUs

Free and Open Source

- Download free plugin from Codeplay website
- Code is open source

Paid Priority Support

- Raise support issues direct to Codeplay engineers
- Escalate defect reports

Compatibility

CUDA SDK >= 11.7

GPUs with at least sm_50

intel. 18

Intel® oneAPI Toolkits

Intel® oneAPI Base Toolkit

A core set of high-performance libraries and tools for building C++, SYCL, C/OpenMP, and Python applications

Add-on Domain-specific Toolkits

For HPC developers

Intel® oneAPI Tools for HPC
Deliver fast Fortran, OpenMP & MPI applications that scale

For Edge & IoT developers

Intel® oneAPI Tools for IoT
Build efficient, reliable solutions that run at network's edge

For visual creators, scientists, and engineers

Intel® oneAPI Rendering Toolkit
Create performant, high-fidelity visualization applications

Toolkits powered by oneAPI

For AI developers and data scientists

Intel® AI Analytics Toolkit
Accelerate machine learning & data science pipelines end-to-end with optimized DL & ML frameworks & high-performing Python libraries

For deep learning inference developers

Intel® OpenVINO™ toolkit
Deploy high performance inference & applications from edge to cloud

Download at **intel.com/oneAPI**
Or run tools in the Intel® Developer Cloud

Latest version available 2023

Intel® oneAPI and AI Tools 2023 – Highlights

Optimized, Standards-based Support for Powerful New Architectures

Optimized support for Intel's upcoming portfolio of CPU and GPU architectures

4th Gen Intel® Xeon® Scalable Processors with Intel® Advanced Matrix Extensions, Quick assist Technology, Intel® AVX-512, bfloat16, and more built-in accelerators

Intel® Xeon® Max Series CPUs with high-bandwidth memory

Intel® Data Center GPUs, including Flex Series with hardware AV1 encode and Max Series with datatype flexibility, Intel® Xe Matrix Extensions, vector engine, XE-Link, and other features

Compilers & SYCL Support

Intel® oneAPI DPC++/C++ Compiler improves CPU and GPU offload performance and broadens SYCL language support for improved code portability and productivity

Intel® oneAPI DPC++ Library (oneDPL) expands support of the C++ standard library in SYCL kernels with additional heap and sorting algorithms and adds the ability to use OpenMP for thread-level parallelism.

Intel® DPC++ Compatibility Tool (based on the open source SYCLomatic project) improves migration of CUDA library APIs, including those for runtime and drivers, cuBLAS, and cuDNN.

Intel® Fortran Compiler implements coarrays, eliminating the need for external APIs such as MPI or OpenMP, expands OpenMP 5.x offloading features, adds DO CONCURRENT GPU offload, and improves optimizations for source-level debugging.

Performance Libraries

Intel® oneAPI Math Kernel Library increases CUDA library function API compatibility coverage for BLAS and FFT; for 4th Gen Xeon, leverages Intel® XMX to optimize matrix multiply computations for TF32, FP16, BF16, and INT8 data types; and provides interfaces for SYCL and C/Fortran OpenMP offload programming.

Intel® oneAPI Threading Building Blocks improves support and use of the latest C++ standard for parallel_sort, offers an improved synchronization mechanism to reduce contention when multiple task_arena calls are used concurrently, and adds support for Microsoft Visual Studio 2022 and Windows Server 2022.

Intel® oneAPI Video Processing Library supports the industry's only hardware AV1 codec in the Intel Data Center GPU Flex Series and Intel® Arc™ processors; expands OS support for RHEL9, CentOS, Stream 9, SLES15Sp4, and Rocky 9 Linux; and adds parallel encoding feature to sample_multi_transcode.

New & enhanced features announced December 2022. For full details: Check out What's New

Intel® oneAPI and AI Tools 2023 – Highlights

Optimized, Standards-based Support for Powerful New Architectures

Analysis & Debug

Intel® VTune™ Profiler enables ability to identify MPI imbalance issues via its Application Performance Snapshot feature; delivers visibility into Xe Link cross-card traffic for utilization, bandwidth consumption, and other issues; and adds support for 4th Gen Intel® Xeon® Scalable Processors, Intel® Data Center GPU Max Series, and 13th Gen Intel® Core™ processors.

Intel® Advisor adds automated roofline analysis for Data Center GPU MAX Series to identify and prioritize memory, cache, or compute bottlenecks and understand their causes, and delivers actionable recommendations for optimizing data-transfer reuse costs of CPU-to-GPU offloading.

AI & Analytics

Intel® AI Analytics Toolkit can now be run natively on Windows with full parity to Linux except for distributed training (GPU support is coming in Q1 2023).

Intel® oneAPI Deep Neural Network Library further supports delivery of superior CNN performance by enabling advanced features in 4th Gen Xeon CPUs including Intel AMX, AVX-512, VNNI, and bfloat16.

Intel® Distribution of Modin integrates with new heterogeneous data kernels (HDK) solution in the back end, enabling AI solution scale from low-compute resources to large- or distributed-computed resources.

Beta additions for Intel® Distribution for Python include compute-follows-data model extension to GPU, data exchange between libraries and frameworks, and data-parallel extensions for NumPy and Numba packages.

Rendering & Visual Computing

Intel® oneAPI Rendering Toolkit includes the Intel® Implicit SPMD Program Compiler runtime library for fast SIMD performance on CPUs.

Intel® Open Volume Kernel Library increases memory-layout efficiency for VDB volumes and adds an AVX-512 8-wide CPU device mode for increased workload performance.

Intel® OSPRay and Intel® OSPRay Studio add features for multi-segment deformation motion blur for mesh geometry, primitive, and objects; face-varying attributes for mesh and subdivision geometry; new light capabilities such as photometric light types; and instance ID buffers to create segmentation images for AI training.

New & enhanced features announced December 2022. For full details: Check out What's New

oneAPI: A Bridge to Our Heterogeneous/Distributed Future

My vision for how we bring oneAPI into a future dominated by power-optimized heterogenous chips organized into distributed systems:

Research

Machine Programming
Radical portability across distributed systems

Distributed Data Structures
Collection of distributed data containers for common core structures

Partitioned Global Address Space
OpenSHMEM or MPI 3 one-sided communication

oneAPI Languages
SYCL, OpenMP, TBB + common high-level APIs

Foundation of solid oneAPI engineering

The key to making this work? The programmer is in control and chooses the level of abstraction based on the programming task.

Slide credit : Joe Curley & Tim Mattson

Conclusion

- The present and future of computing is heterogeneous
- oneAPI proposes an open specification for programming heterogenous systems
- Results demonstrate the promise of multi-architecture, multi-vendor oneAPI
- Future opportunities : continued ease of development, multi-architecture, distributed, disaggregated

- Call to action: Interested? Join the community at oneapi.io

Slide credit : Joe Curley & Tim Mattson

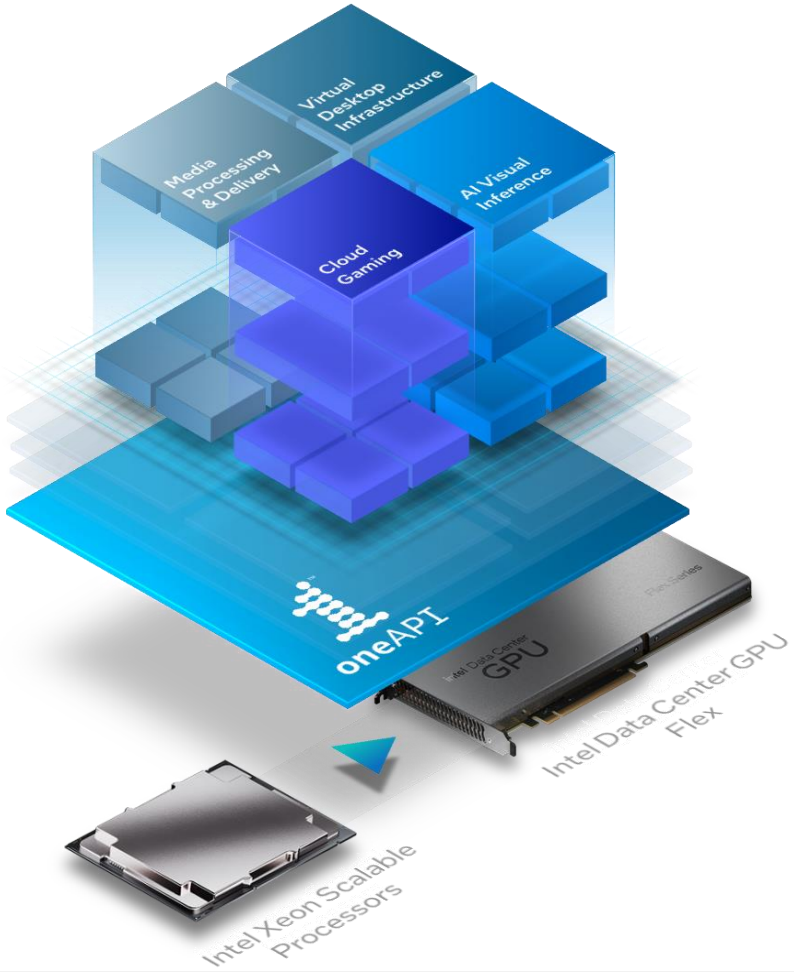
Developer Tools for Intel® Data Center GPU Flex Series

Multiarchitecture Acceleration for Media, Visual Inferencing & Cloud Workloads

Intel® oneAPI Video Processing Library for GPU, and GStreamer + FFmpeg for CPU speed media processing, delivery and cloud gaming streaming.

Intel® oneAPI DPC++/C++ Compiler & oneVPL, both part of the Intel® oneAPI Base Toolkit, scale AI visual inference workloads up to 150 TOPS INT8 between CPU and GPU with the same code.

Intel® VTune™ Profiler analyzes application performance and accelerates compute-intensive tasks by identifying the most time-consuming parts of GPU code and optimizing GPU offload schema and data transfers for SYCL, OpenCL code, Microsoft DirectX*, or OpenMP* offload code.



Intel® Graphics Performance Analyzers optimize visual compute and graphics workloads on CPUs and GPUs .

Intel® OpenVINO™ toolkit, powered by oneAPI, supports popular deep learning frameworks such as TensorFlow and PyTorch to streamline AI visual inferencing and deploy quickly.

Intel® AI Analytics Toolkit, powered by oneAPI, accelerates machine learning & data science pipelines end-to-end with optimized DL & ML frameworks & high-performing Python libraries

Developer Tools for Intel® Data Center GPU Max Series

Build Multiarchitecture Applications with Breakthrough Performance for HPC & AI

Accelerate HPC

oneAPI DPC++/C++ and Fortran compilers and performance libraries (oneMKL, oneDNN, Intel® MPI Library) activate Xe Matrix Extensions (Intel® XMX) for acceleration and Intel® Xe Link for direct GPU-to-GPU communications to speed up applications running on multiple GPUs.

Analysis and debug tools help developers analyze and accelerate compute-intensive tasks by identifying the most time-consuming parts of GPU code, ensure correctness and efficiently debug applications on discrete GPUs. Get actionable advice to design code that runs optimally on Intel GPUs.

Boost AI & Deep Learning Inference

oneDNN in the Intel® oneAPI Base Toolkit utilizes Intel XMX to accelerate AI training/inference in deep learning frameworks.

Intel® OpenVINO™ toolkit powered by oneAPI. streamlines AI visual inferencing and accelerates deployment.

Both tools support popular deep learning frameworks such as TensorFlow and PyTorch. GPU support for tools in the Intel® AI Analytics Toolkit is coming soon.

Create Multiarchitecture Code Efficiently with Code Migration Tools

Migrate CUDA code to SYCL to create a single source code base for easy portability across multiple vendors' architectures - including Intel Max Series GPU.

The Intel® DPC++ Compatibility Tool, based on open source SYCLomatic, automates most of the process. Save significant time on ongoing code maintenance.

Coming Soon – High-Performance, High-Fidelity Ray Tracing & Rendering

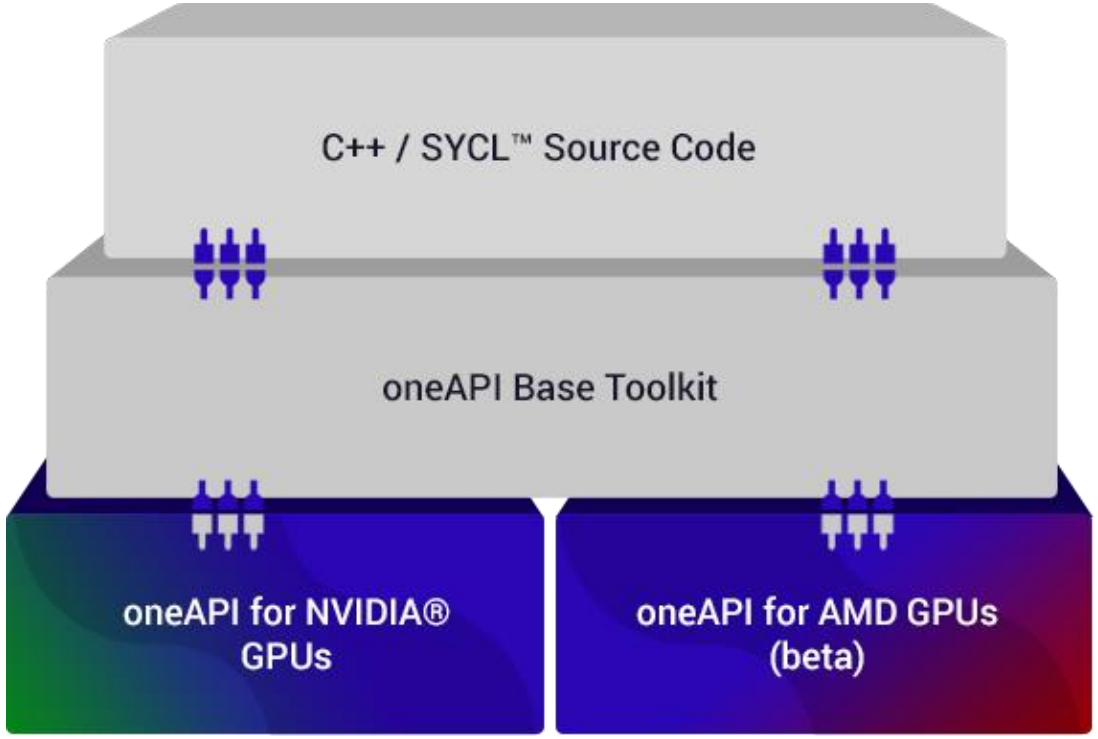
Intel® Embree, fully ported to C++ with SYCL will take advantage of hardware ray tracing units and expanded memory capacity to deliver increased rendering performance across larger datasets.

Intel® Open Image Denoise will leverage Intel XMX delivering high fidelity, high-performance denoising capabilities and AI optimization.

oneAPI for NVIDIA and AMD GPUs

Adds support for NVIDIA and AMD GPUs to the Intel oneAPI Base Toolkit.

Develop code using SYCL and run on NVIDIA and AMD GPUs.



Download free plugins from developer.codeplay.com

oneAPI for NVIDIA GPUs

Free and Open Source

- Download free plugin from Codeplay website
- Code is open source

Paid Priority Support

- Raise support issues direct to Codeplay engineers
- Escalate defect reports

Compatibility

CUDA SDK >= 11.7

GPUs with at least sm_50

Backup

oneAPI Industry Momentum

End Users

Accrad Accelerated Radiology
archanan
FOUNDRY MODO
BioDataAnalysis GmbH GeoEast Next Generation Bio-Medical Image Analysis
BENTLEY
CINESITE
LAIKA
GE
kt
PHILIPS
AUTODESK ARNOLD
中国石油集团东方地球物理勘探有限责任公司 BGP INC.,CHINA NATIONAL PETROLEUM CORPORATION
iOmniscient Autonomous
EURECOM Sophia Antipolis
Brightskies
大势智慧 DASPATIAL
WeBank
SAMSUNG MEDISON
TANGENT ANIMATION
allegro.ai
ILLUMINATION MACGUFF
Verizon
MediaKind

National Labs

Argonne NATIONAL LABORATORY
SiPearl
CERN openlab
UT-BATTELLE Oak Ridge National Laboratory
CINECA
Peraton Labs
SANKHYA SUTRA LABS
lrz Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities
Laboratório Nacional de Computação Científica

ISVs & OSVs

codeplay
CANONICAL
CERTIFACE
ANACONDA
Ansys
MAX PLANCK COMPUTING & DATA FACILITY
MPCDF
AI SINGAPORE
SAS
Red Hat OpenShift Data Science
KATANA GRAPH
CGG
AIBLE
CHAOSGROUP
MAXON
Ospirent
FOUNDRY
mercenaries engineering Guerilla
Hisense Medical 海信医疗
SENAI CIMATEC FIEB
SAP
E4 COMPUTER ENGINEERING
YUAN
GIGASPACES innovate with confidence
MEGH COMPUTING
KFBIO
Tech Mahindra
AsiaInfo 亚信科技
SUSE We adapt. You succeed.
UNITED IMAGING

OEMs & SIs

Bittware a molex company
Hewlett Packard Enterprise
DELL Technologies
Atos
Lenovo
HCL
rENIAC
MEGWARE SUPERCOMPUTING TECHNOLOGY

Universities & Research Institutes

LOBACHEVSKY UNIVERSITY
TECHNION Israel Institute of Technology
UNIVERSITY OF CAMBRIDGE
中国科学院计算技术研究所 INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES
Ben-Gurion University of the Negev
Berkeley UNIVERSITY OF CALIFORNIA
UC DAVIS UNIVERSITY OF CALIFORNIA
CtG
北京大学软件与微电子学院 School of Software & Microelectronics
THE UNIVERSITY OF TENNESSEE KNOXVILLE
UNIVERSIDAD COMPLUTENSE MADRID
OLD DOMINION UNIVERSITY
ZIB
ILLINOIS
THE UNIVERSITY OF UTAH
UNIVERSITY OF OREGON
TÉCNICO LISBOA
inesc id lisboa
University of Stuttgart Germany
FACULTY OF MATHEMATICS AND PHYSICS Charles University
Durham University
University College London
Indian Institutes of Technology Delhi / Kharagpur / Roorkee
सी डैक CDAC
Science and Technology Facilities Council Scientific Computing
TACC
SDSC SAN DIEGO SUPERCOMPUTER CENTER
Indian Institute of Science Bangalore
IISER PUNE Indian Institute of Science Education & Research Pune
UNIVERSIDAD DE MÁLAGA
URZ HEIDELBERG UNIVERSITY COMPUTING CENTRE
Northern Illinois University
University of BRISTOL
Stockholm University
KTH

CSPs & Frameworks

Microsoft Azure
Alibaba Cloud
Google Cloud
TensorFlow
Taboola
Tencent 腾讯
DataRobot
Baidu 百度
飞桨 PaddlePaddle
NAVER CLOVA
PyTorch

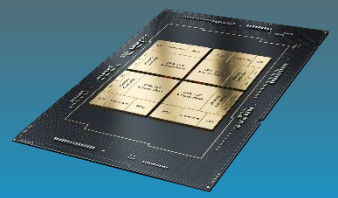
These organizations support the oneAPI initiative for a single, unified programming model for cross-architecture development.
It does not indicate any agreement to purchase or use of Intel's products. *Other names and brands may be claimed as the property of others.

Modern Applications Demand Increased Processing

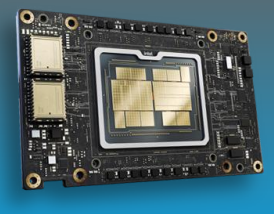
Diverse accelerators needed to meet today's performance requirements:

48% of developers target heterogeneous systems that use more than one kind of processor or core[1]

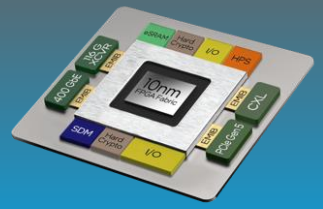
CPU



GPU



FPGA



Other Accelerators

Developer Challenges: Multiple Architectures, Vendors, and Programming Models

oneAPI

Open, Standards-based, Multiarchitecture Programming

1. Evans Data Global Development Survey Report 22.1, June 2022

oneAPI Industry Initiative

Break the Chains of Proprietary Lock-in

Freedom to Make Your Best Choice

- One programming model for multiple architectures and vendors
- Cross-architecture code reuse for freedom from vendor lock-in

Realize all the Hardware Value

- Performance across CPU, GPUs, FPGAs, and other accelerators
- Expose and exploit cutting-edge features of the latest hardware

Develop & Deploy Software with Peace of Mind

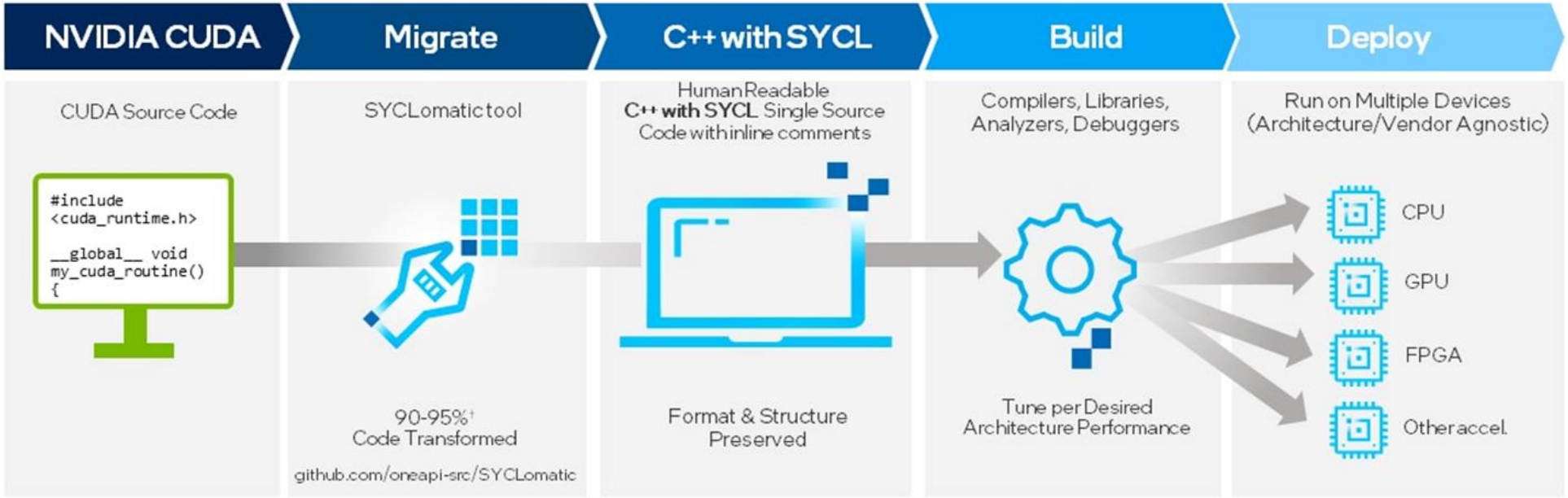
- Open industry standards provide a safe, clear path to the future
- Compatible with existing languages and programming models including C, C++ with SYCL, Python, OpenMP, Fortran, and MPI
- Powerful libraries for acceleration of domain-specific functions

oneAPI

The productive, smart path to freedom for accelerated computing from the economic and technical burdens of proprietary programming models

CUDA to SYCL Migration Made Easy

Open Source SYCLomatic Tool Reduces Code Migration Time



Assists developers migrating code written in CUDA to C++ with SYCL, generating **human readable** code wherever possible

~90-95% of code typically migrates automatically[1]

Inline comments are provided to help developers finish porting the application

Intel® DPC++/C++ Compatibility Tool is Intel's implementation, available in the Base Toolkit

[1]Intel estimates as of September 2021. Based on measurements on a set of 70 HPC benchmarks and samples, with examples like Rodinia, SHOC, PENNANT. Results may vary.

Codeplay Compiler Plug-ins for Nvidia and AMD GPUs

Adding support for NVIDIA and AMD GPUs to the Intel® oneAPI Base Toolkit

oneAPI for NVIDIA & AMD GPUs

- Free Codeplay download of latest binary plugins to the Intel DPC++/C++ compiler:
  - Nvidia GPU
  - AMD Beta GPU
- Availability at the same time as the Intel oneAPI Base Toolkit
- Plug-ins updated quarterly in-sync with oneAPI

Priority Support

- Sold by Intel and Codeplay and our channel
- Requires Intel Priority support for Intel DPC++ /C++ compiler
- Intel takes first call and Codeplay delivers backend support
- Codeplay access to older versions of plugins

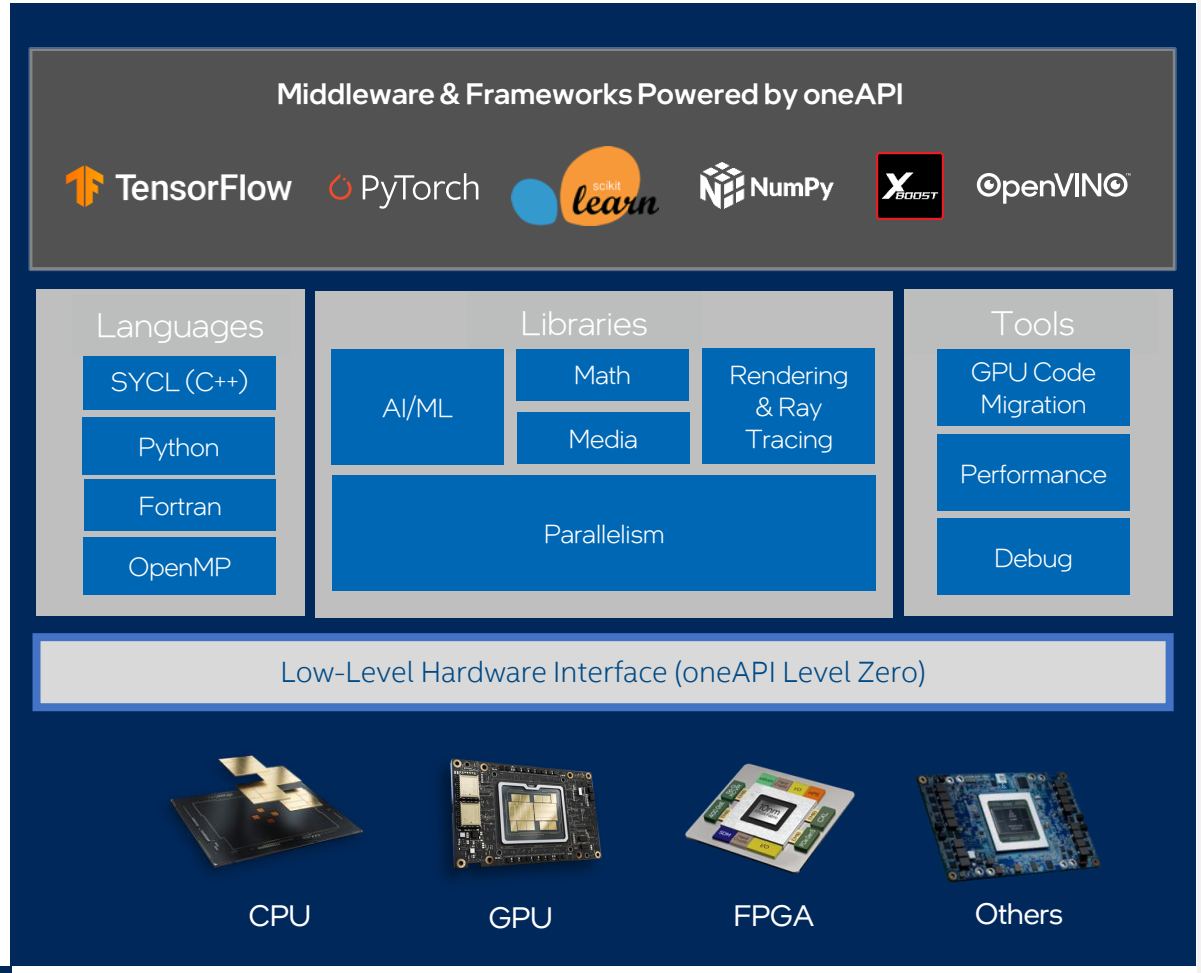
Nvidia GPU plug-in | AMD GPU plug-in | Codeplay blog | Codeplay press release

Intel® Developer Tools Supporting oneAPI

A Complete Set of Proven Tools Expanded from CPU to Accelerators

oneAPI

- Advanced compilers, libraries, and analysis, debug, and porting tools
- Full support for C, C++ with SYCL, Python, Fortran, MPI, OpenMP
- Intel® Advisor determines device target mix before you write your code
- Intel's compilers optimize code to take full advantage of multiarchitecture workload distribution.
- Intel® VTune™ Profiler analyzes hotspots to optimize code performance
- Intel AI tools support acceleration of major deep learning and machine learning frameworks



oneAPI Commercial & Community Support Available

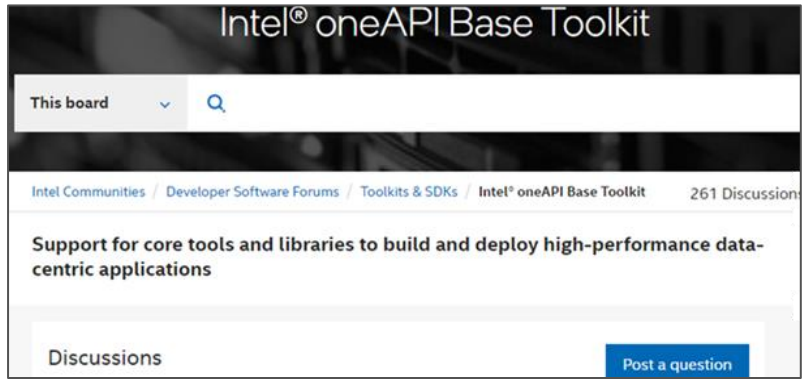
Priority Support for Intel® oneAPI Toolkits

Every paid version of Intel® oneAPI Developer Toolkits includes Priority Support for that toolkit (Intel oneAPI Base, HPC, IOT, & Rendering Toolkits)

- Direct, private interaction with Intel software support engineers
- Accelerated response time
- Access to—and support for—previous Intel products such as Fortran compiler versions, previous toolkit versions, and more
- Intel Technical Consulting Engineers for on-site or online training and consultation at a reduced cost

Free Community Support

- Support via the Intel public Community Forum
- Access to only the latest versions of oneAPI Toolkits
- Access to online tutorials and self-help forums



Maximize Your Performance

With Intel Developer Tools & Hardware Platforms

HPC & Data Center

AI & Visualization

Embedded Systems & IoT

Performance

Optimize compute performance on the latest Intel CPUs, GPUs and FPGAs

Maximize built-in accelerators

Accelerate across AI frameworks

Productivity

Familiar languages and standards

Easily integrate w/ legacy code

Easily migrate CUDA to SYCL

Minimize code re-writes

Freedom

Open alternative to proprietary lock-in

Enables easy architecture retargeting

Code longevity for future hardware

Details about Intel® oneAPI Toolkits

Intel® oneAPI Base Toolkit

Intel® oneAPI Base Toolkit

Accelerate Data-centric Workloads

A core set of core tools and libraries for developing high-performance applications on Intel® CPUs, GPUs, and FPGAs.

Who Uses It?

- A broad range of developers across industries
- Add-on toolkit users since this is the base for all toolkits

Top Features/Benefits

- Data Parallel C++ compiler, library and analysis tools
- SYCLomatic / DPC++ Compatibility tool helps migrate CUDA code to C++ with SYCL
- Python distribution includes accelerated scikit-learn, NumPy, SciPy libraries
- Optimized performance libraries for threading, math, data analytics, deep learning, and video/image/signal processing

Learn More & Download

Intel® oneAPI Base Toolkit

Direct Programming	API-Based Programming	Analysis & debug Tools
Intel® oneAPI DPC++/C++ Compiler	Intel® oneAPI DPC++ Library oneDPL	Intel® VTune™ Profiler
Intel® DPC++ Compatibility Tool	Intel® oneAPI Math Kernel Library - oneMKL	Intel® Advisor
Intel® Distribution for Python	Intel® oneAPI Data Analytics Library - oneDAL	Intel® Distribution for GDB
Intel® FPGA Add-on for oneAPI Base Toolkit	Intel® oneAPI Threading Building Blocks - oneTBB	
	Intel® oneAPI Video Processing Library - oneVPL	
	Intel® oneAPI Collective Communications Library oneCCL	
	Intel® oneAPI Deep Neural Network Library - oneDNN	
	Intel® Integrated Performance Primitives - Intel® IPP	

intel. 1 oneAPI BASE TOOLKIT

Productive and Performant SYCL Compiler

Intel® oneAPI DPC++/C++ Compiler

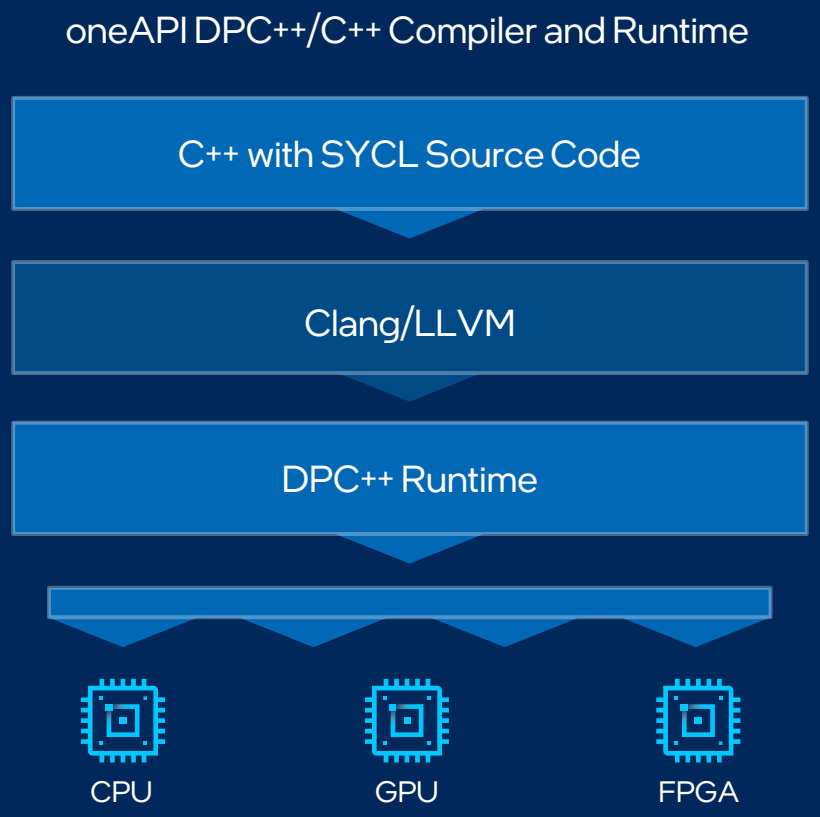
Uncompromised parallel programming productivity and performance across CPUs and accelerators

- Allows code reuse across hardware targets, while permitting custom tuning for a specific accelerator
- Open, cross-industry alternative to single architecture proprietary language

Khronos SYCL Standard

- Delivers C++ productivity benefits, using common and familiar C and C++ constructs
- Created by Khronos Group to support data parallelism and heterogeneous programming

Builds upon Intel's decades of experience in architecture and high-performance compilers



Learn More & Download

There will still be a need to tune for each architecture.

Intel® DPC++ Compatibility Tool

Minimizes Code Migration Time

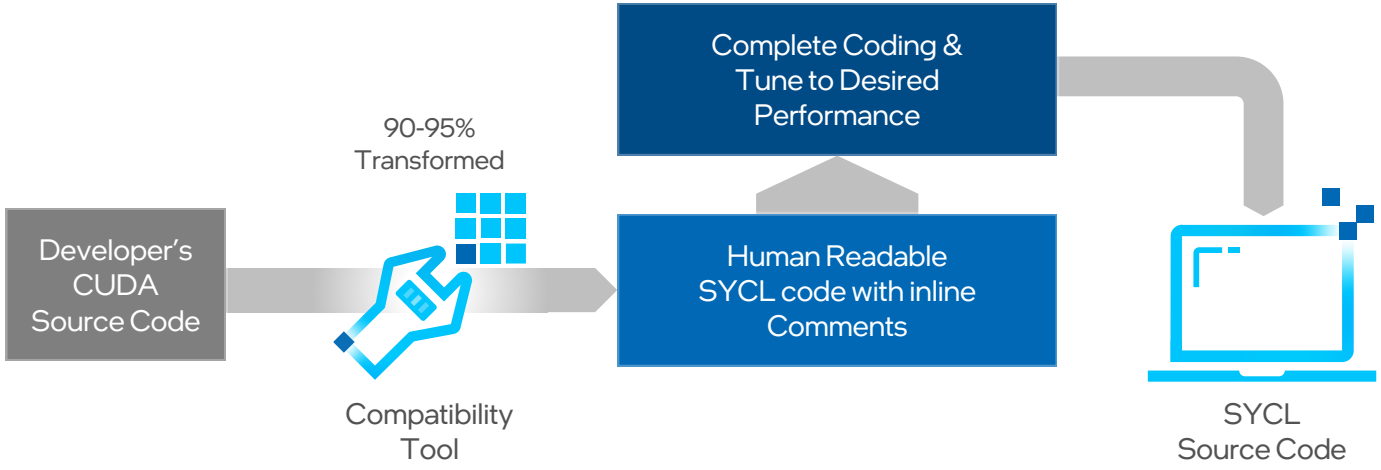
Assists developers migrating code written in CUDA to C++ with SYCL once, generating human readable code wherever possible

~90-95% of code typically migrates automatically[1]

Inline comments are provided to help developers finish porting the application

SYCLomatic tool is the open source version

Intel DPC ++ Compatibility Tool Usage Flow



Free Yourself from Vendor Lock-in – SYCL* Adoption Made Easy

Learn More & Download

[1]Intel estimates as of September 2021. Based on measurements on a set of 70 HPC benchmarks and samples, with examples like Rodinia, SHOC, PENNANT. Results may vary.

Analysis & Debug Tools

Get More from Diverse Hardware

Design

Intel® Advisor

- Efficiently offload code to GPUs
- Optimize your CPU/GPU code for memory and compute
- Enable more vector parallelism and improve efficiency
- Add effective threading to unthreaded applications

Debug

Intel® Distribution for GDB

- Multiple accelerator support with CPU, GPU and FPGA
- Enables deep, system-wide debug of SYCL, C, C++, OpenMP and Fortran cross-architecture applications
- IDE Integration into Microsoft Visual Studio, VS Code and Eclipse

Tune

Intel® VTune™ Profiler

- Tune for GPU, CPU, and FPGA
- Optimize offload performance
- Supports SYCL, C, C++, Fortran, Python, Go, Java or a mix of languages

Powerful Performance Libraries

Rich Functionality

Optimized performance libraries for every use case:

Threading, offload, math, data analytics, data processing, rendering, ray tracing, DNN, comms, crypto, and more.

Realize all the Hardware Value

Designed for acceleration of key domain-specific functions

Freedom of Choice

Pre-optimized for each target platform for maximum performance

Explore the Libraries

Intel® oneAPI Math Kernel Library **oneMKL**	Intel® oneAPI Deep Neural Network Library **oneDNN**
Intel® oneAPI Video Processing Library **oneVPL**	Intel® oneAPI Data Analytics Library **oneDAL**
Intel® oneAPI Threading Building Blocks **oneTBB**	Intel® oneAPI Collective Communications Library **oneCCL**
Intel® oneAPI DPC++ Library **oneDPL**	Intel® oneAPI Rendering Toolkit Rendering and Ray-Tracing Libraries
Intel® Integrated Performance Primitives **Intel® IPP**	Intel® MPI Library

The One-Stop Shop for All your Developer Library Needs

Intel® oneAPI DPC++ Library (oneDPL)

Accelerate SYCL C++ Kernels on Intel CPUs, GPUs & FPGAs

Optimized C++ Standard Algorithms

Contains 75 parallelized C++17 algorithms and utilities for efficient application development and deployment on a variety of hardware.

Based on parallel libraries that C++ developers are already familiar with

Incorporates popular libraries Parallel STL and Boost. Compute for easier developer adoption.

Integrated with Intel® DPC++ Compatibility Tool

Complements all oneAPI DPC++ components to simplify migration of developers' CUDA* code to DPC++ code.

Bringing Multi-Architecture Compute to C++

Learn More & Download

Intel® oneAPI Deep Neural Network Library (oneDNN)

Deliver High Performance Deep Learning

Helps developers create high performance deep learning frameworks

Abstracts out instruction set & other complexities of performance optimizations

Same API for both Intel CPUs and GPUs, use the best technology for the job

Supports Linux, Windows

Open sourced for community contributions

Learn More & Download

Intel® oneAPI Video Processing Library (oneVPL)

Accelerated Video Processing with a Unified Programming API

Deliver fast, high-quality video transcoding from camera to cloud

Boost media and video application performance with hardware-accelerated codecs and programmable graphics on Intel CPUs and GPUs

Simple API that works the same on CPU and GPU

Using the API, developers have full control over codec visual quality and performance

Learn More & Download

Intel® oneAPI Collective Communications Library (oneCCL)

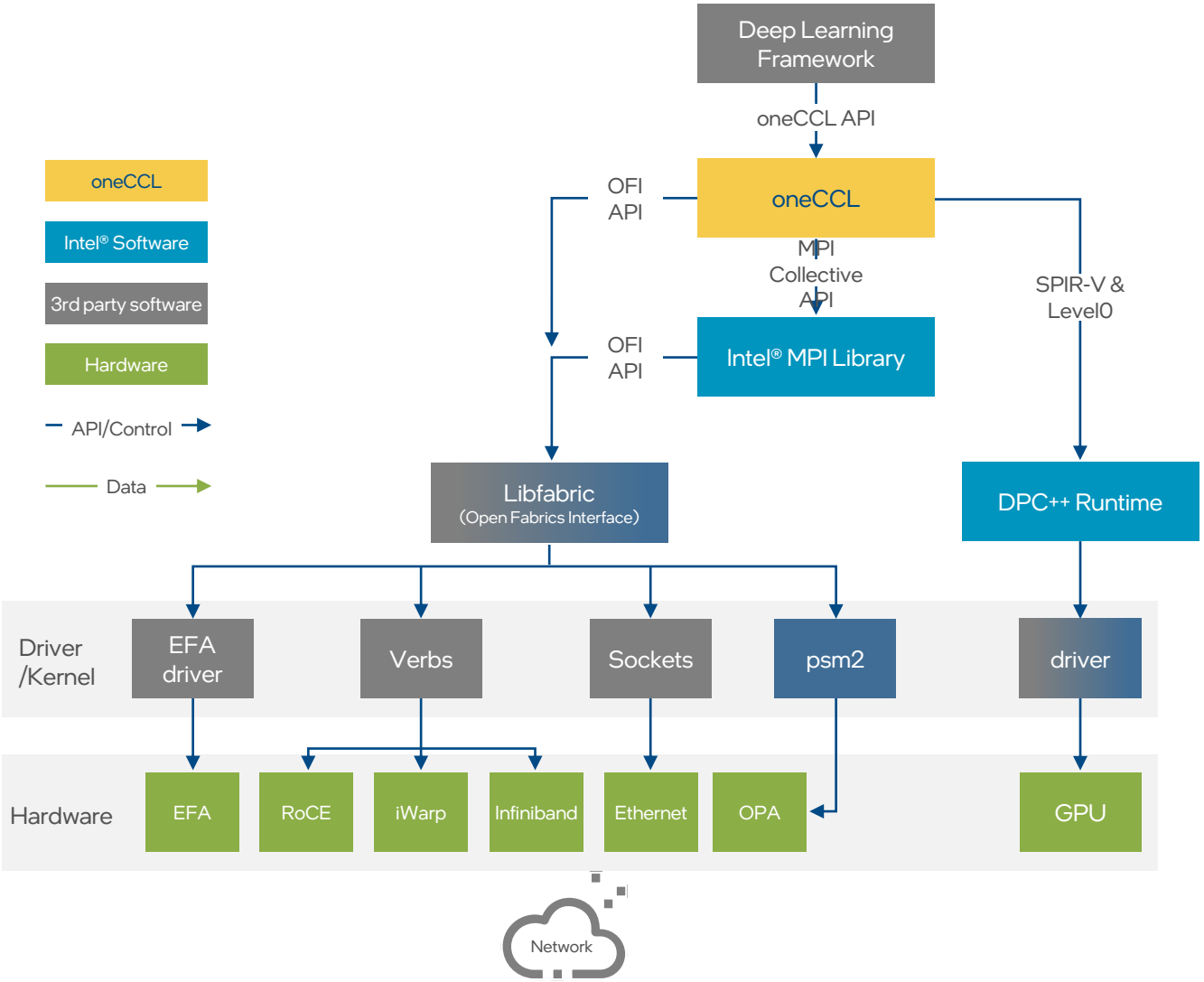
Optimize Communication Patterns

Provides optimized communication patterns for high performance on Intel CPUs & GPUs to distribute model training across multiple nodes

Transparently supports many interconnects, such as Intel® Omni-Path Architecture, InfiniBand, & Ethernet

Built on top of lower-level communication middleware-MPI & libfabrics

Enables efficient implementations of collectives used for deep learning training-all-gather, all-reduce, & reduce-scatter



Learn More & Download

Intel® VTune™ Profiler

Tune for CPU, GPU & FPGA

Analyze SYCL code

See the lines of SYCL that consume the most time

Tune for Intel CPUs, GPUs & FPGAs

Optimize for any supported hardware accelerator

Optimize Offload

Tune OpenMP offload performance

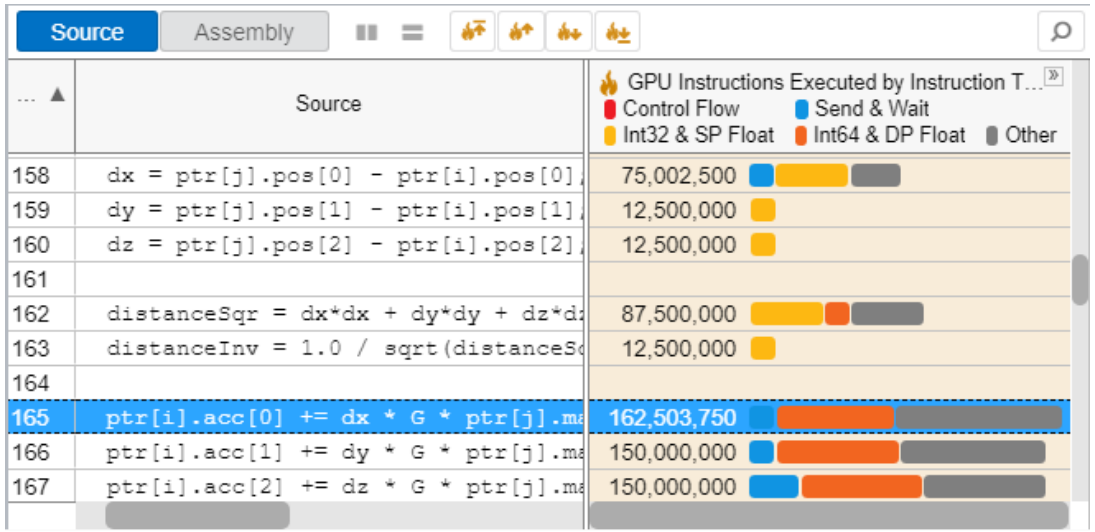
Wide Range of Performance Profiles

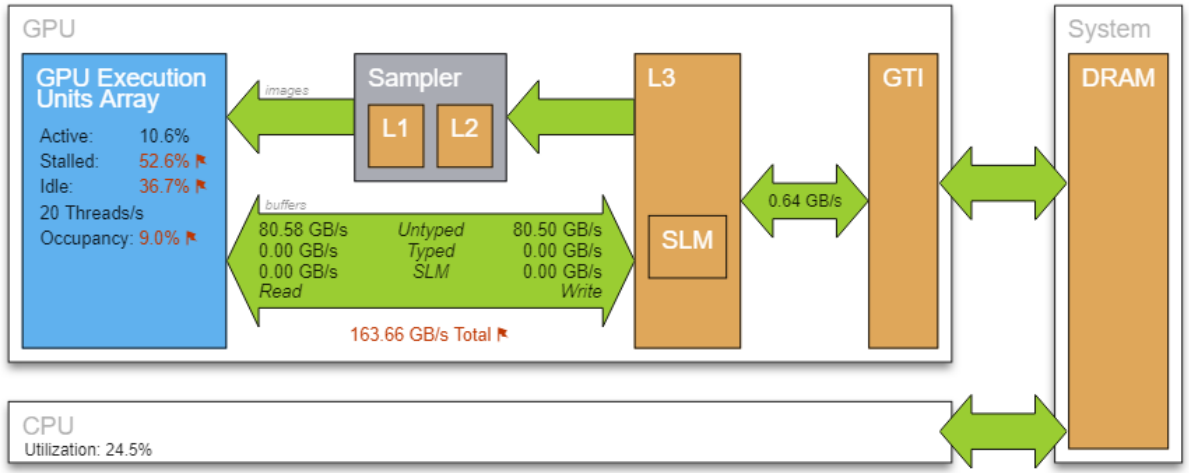
CPU, GPU, FPGA, threading, memory, cache, storage...
Flame graph display improves visualization of hot spots

Supports Popular Languages

SYCL, C, C++, Fortran, Python, Go, Java, or a mix

Learn More & Download





Images above show analysis of SYCL code and GPU Offload profiling.

There will still be a need to tune for each architecture.

Intel® Advisor

Configure Your Accelerated Computing Solution

Offload Advisor

Estimate performance of offloading to an accelerator

Roofline Analysis

Optimize CPU/GPU code for memory and compute

Vectorization Advisor

Add and optimize vectorization

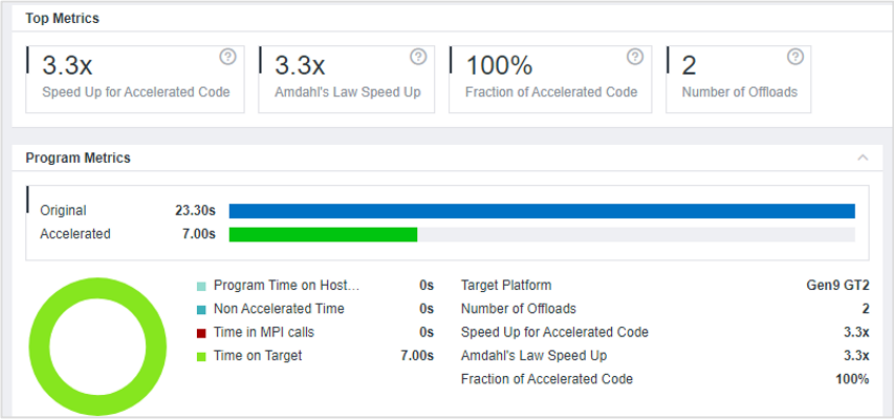
Threading Advisor

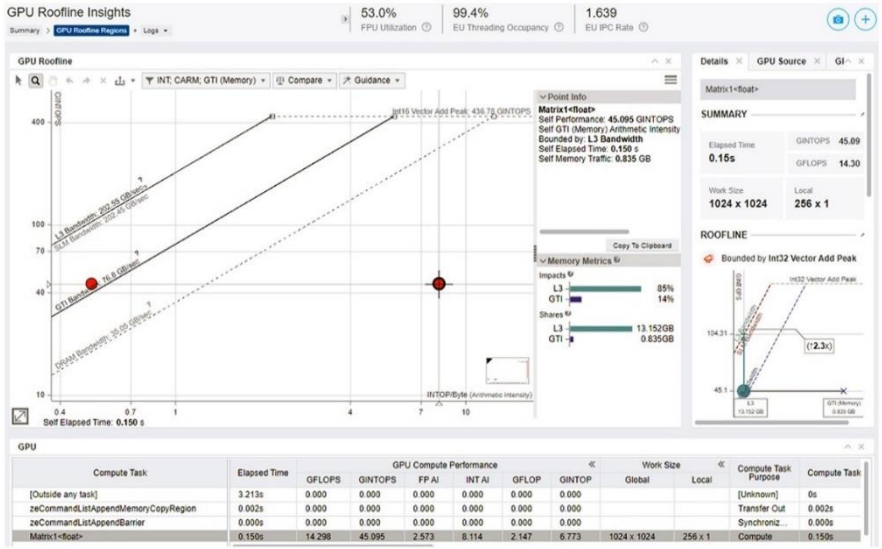
Add effective threading to unthreaded applications

Flow Graph Analyzer

Create and analyze efficient flow graphs

Learn More & Download





Know Before You Offload

There will still be a need to tune for each architecture.

Intel® Distribution for GDB*

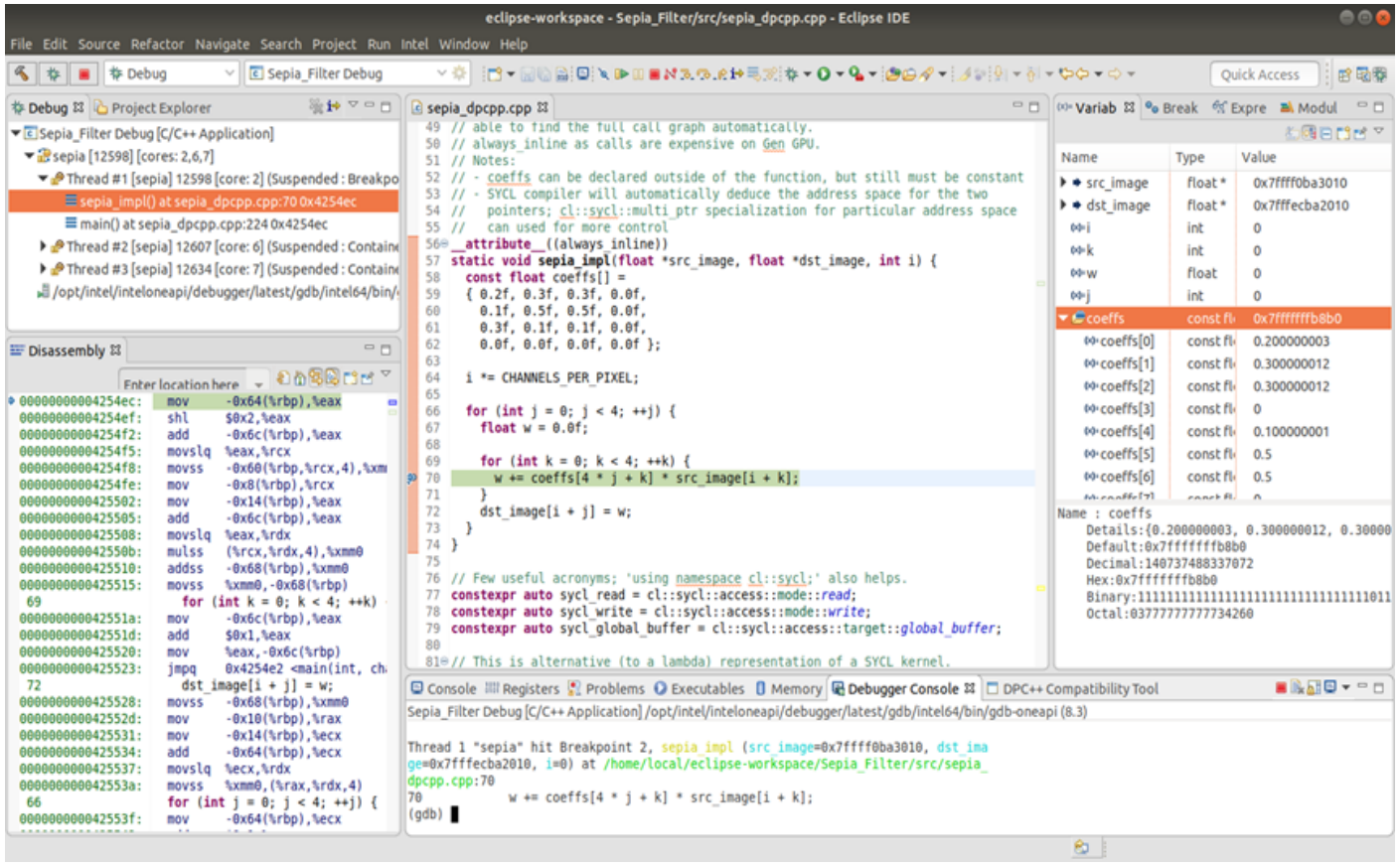
Heterogeneous Application Debug

High-level language debug support

Multiple accelerator support:
Intel CPU, GPU, FPGA emulation

Auto-detect accelerator architecture
during application runtime

Non-proprietary open-source solution
based on GDB



Learn More & Download

oneAPI FPGA Add-On

SYCL Coding for Spatial Architecture

For Experienced FPGA Developers

Ease of Use

The same heterogeneous SYCL code. Experienced FPGA users can take advantage of a streamlined programming model using Data Parallel C++

Runtime Analysis Support

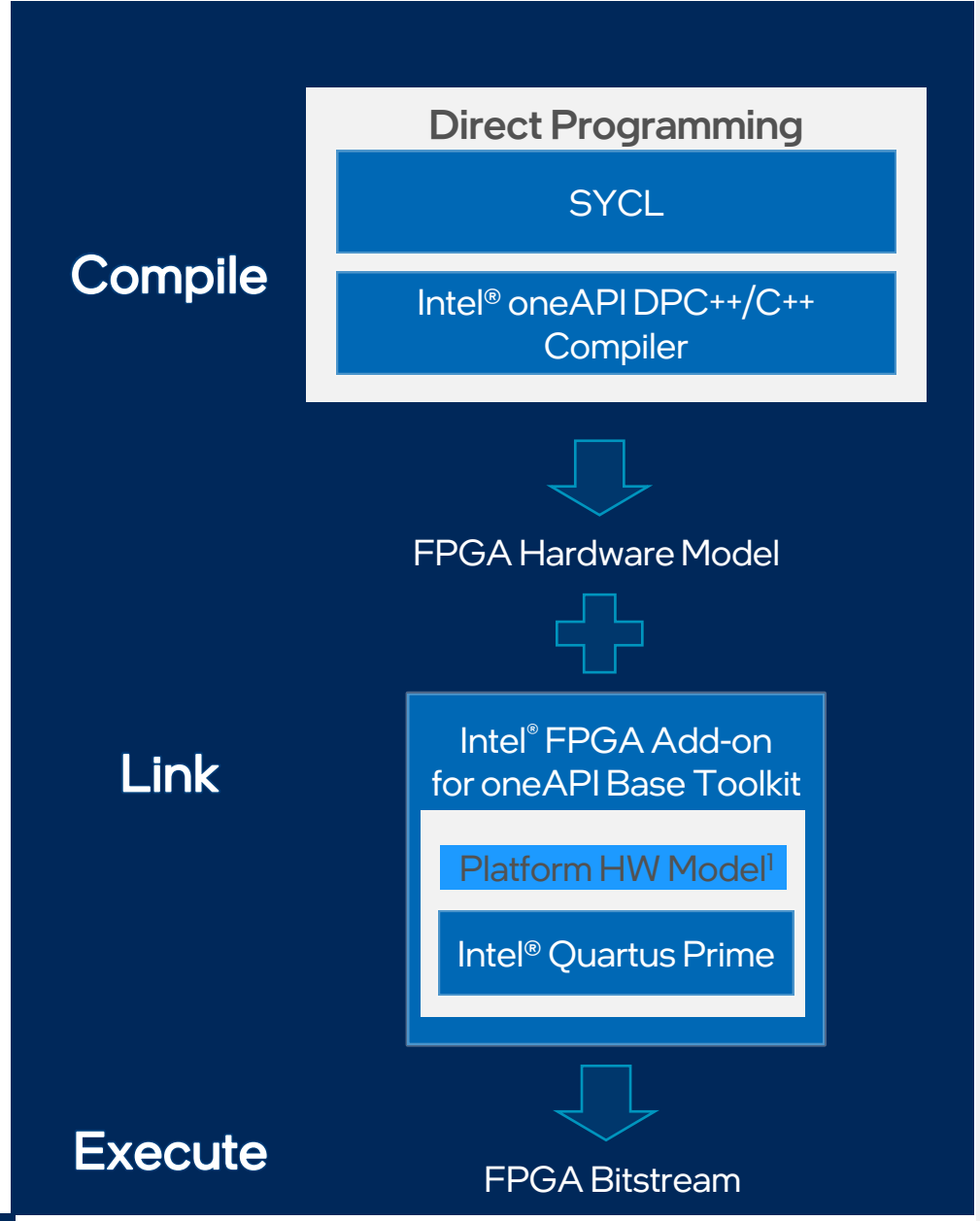
Collect profiling data at runtime to analyze CPU and FPGA interaction with Intel® VTune™ Profiler

Real-time Processing

Process data faster with deterministic low latency, lower power and high throughput

Device Specific Optimizations

One-day class provides experienced FPGA developers training to begin optimizing oneAPI code for FPGA



1. Available from 3rd Party Platform Vendors

Where to get started?

	1 Optimizing your source code?	2 Writing a new program or porting existing CUDA code?	3 Adding accelerators to your hardware config?
Compilers & Code Migration	Intel® DPC++ Compatibility Tool Intel® oneAPI DPC++/C++ Compiler Intel® Fortran Compiler Intel® Distribution for Python	Intel® DPC++ Compatibility Tool Intel® oneAPI DPC++/C++ Compiler Intel® Fortran Compiler Intel® Distribution for Python	Intel® DPC++ Compatibility Tool Intel® oneAPI DPC++/C++ Compiler Intel® Fortran Compiler Intel® Distribution for Python
Libraries	Intel® oneAPI Libraries	Intel® oneAPI Libraries	Intel® oneAPI Libraries
Analyzers	Intel® VTune Profiler Intel® Advisor	Intel® VTune Profiler Intel® Advisor	Intel® Advisor Intel® VTune Profiler
	Start with VTune to analyze code performance in your application. **Using an older compiler?** Re-compiling with Intel's newest compiler can offer real speed-up.	**New Program? Start with our industry leading compliers & libraries** found in our toolkits for your best performance. **Porting? Use DPC++ Compatibility Tool** to migrate your CUDA code to SYCL for porting across architectures	**Start with Intel Advisor** to define your optimal device mix before optimizing and target the right accelerators. **Use VTune** to further analyze and tune code performance in your application

	4 Working on new hardware?		
Bring-Up		Intel® System Bring Up Toolkit	**Start with Intel System Bring Up Kit** to speed up system bring-up and validation of new hardware

oneAPI Resources

software.intel.com/oneapi

Get Started

- software.intel.com/oneapi
- Documentation + dev guides
- Code Samples
- Intel® Developer Cloud

Industry Initiative

- oneAPI.io
- oneAPI open Industry Specification
- Open-source Implementations

Learn

- Training: Webinars & courses
- Intel® DevMesh Innovator Projects
- Summits & Workshops: Live & on-demand virtual workshops, community-led sessions
- Training by certified oneAPI experts worldwide for HPC & AI

Ecosystem

- Community Forums
- Intel® DevMesh Innovator Projects
- Academic Programs: oneAPI Centers of Excellence: research, enabling code, curriculum, teaching

Domain-specific Toolkits for Specialized Workloads

- Intel® oneAPI HPC Toolkit
- Intel® AI Analytics Toolkit
- Intel® oneAPI Rendering Toolkit
- Intel® oneAPI IoT Toolkit
- Intel® Distribution of OpenVINO™ toolkit, powered by oneAPI

Intel® oneAPI Tools for HPC

Intel® oneAPI HPC Toolkit

Deliver Fast Applications that Scale

What is it?

A toolkit that adds to the Intel® oneAPI Base Toolkit for building high-performance, scalable parallel code on C++, Fortran, SYCL, OpenMP & MPI from enterprise to cloud, and HPC to AI applications.

Who needs this product?

- OEMs/ISVs
- C++, Fortran, OpenMP, MPI Developers

Why is this important?

- Accelerate performance on Intel® Xeon® & Core™ processors & Intel accelerators
- Deliver fast, scalable, reliable parallel code with less effort built on industry standards

Learn More & Download

Intel® oneAPI Base & HPC Toolkits

Direct Programming	API-Based Programming	Analysis & debug Tools
Intel® C++ Compiler Classic	Intel® MPI Library	Intel® Inspector
Intel® Fortran Compiler Classic	Intel® oneAPI DPC++ Library oneDPL	Intel® Trace Analyzer & Collector
Intel® Fortran Compiler	Intel® oneAPI Math Kernel Library - oneMKL	Intel® Cluster Checker
Intel® oneAPI DPC++/C++ Compiler?	Intel® oneAPI Data Analytics Library - oneDAL	Intel® VTune™ Profiler
Intel® DPC++ Compatibility Tool	Intel® oneAPI Threading Building Blocks - oneTBB	Intel® Advisor
Intel® Distribution for Python	Intel® oneAPI Video Processing Library - oneVPL	Intel® Distribution for GDB
Intel® FPGA Add-on for oneAPI Base Toolkit	Intel® oneAPI Collective Communications Library oneCCL	
	Intel® oneAPI Deep Neural Network Library - oneDNN	
	Intel® Integrated Performance Primitives – Intel® IPP	

Intel® oneAPI **HPC** Toolkit +
Intel® oneAPI **Base** Toolkit

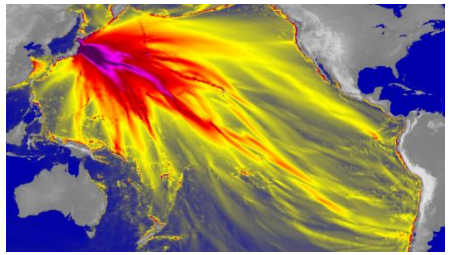
intel. 1 oneAPI HPC TOOLKIT

Deliver Fast HPC Applications that Scale

Customer Use Cases – Intel® oneAPI Base & HPC Toolkits

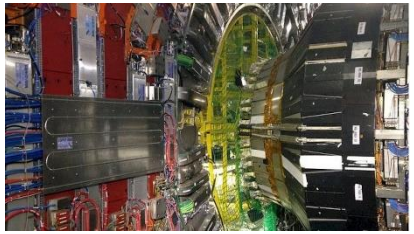
SAMPLE USE CASES & PROOF POINTS

Intel oneAPI tools help prepare code for Aurora.
Aurora, Argonne Leadership Computing Facility's Intel-HPE/Cray supercomputer, will be one of the U.S.'s 1st exascale systems



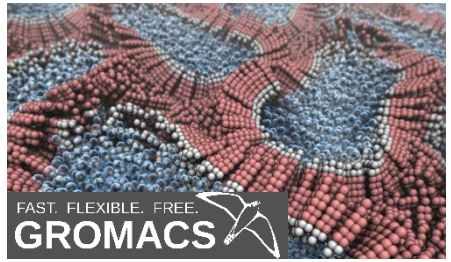
Zuse Institute Berlin (ZIB) ported the *easyWave* tsunami simulation application from CUDA to Data Parallel C++ (DPC++) **delivering performance on Intel CPUs, GPUs, FPGAs**, & Nvidia P100

Accelerating Google Cloud for HPC
C2 provides great performance for HPC workloads: **40% higher performance/core.** Runs on Intel® Xeon® processors + AMD, optimized by Intel® oneAPI **Base** & **HPC** Toolkits. Video
Video | Podcast



Acceleration for HPC & AI Inferencing
CERN, SURFsara, and Intel are investigating approaches driving **breakthrough performance on simulations** used in scientific, engineering, and financial applications*.

Texas Advanced Computing Center (TACC)
Frontera SuperComputer Visualization & Filesystem Use Cases Show Value of Large Memory Fat Nodes on Intel® Xeon® processors & Intel® Optane Persistent Memory*



University of Stockholm/KTH
GROMACS, a simulation application used to design new drugs, was optimized by oneAPI. CUDA code was migrated to oneAPI to create new cross-architecture code targeting Intel CPUs and multiple accelerators.

Learn more: *oneAPI Discussions with HPC Thought Leaders* Video [2.20]

*Uses Intel® oneAPI Rendering Toolkit

Intel® AI Analytics Toolkit

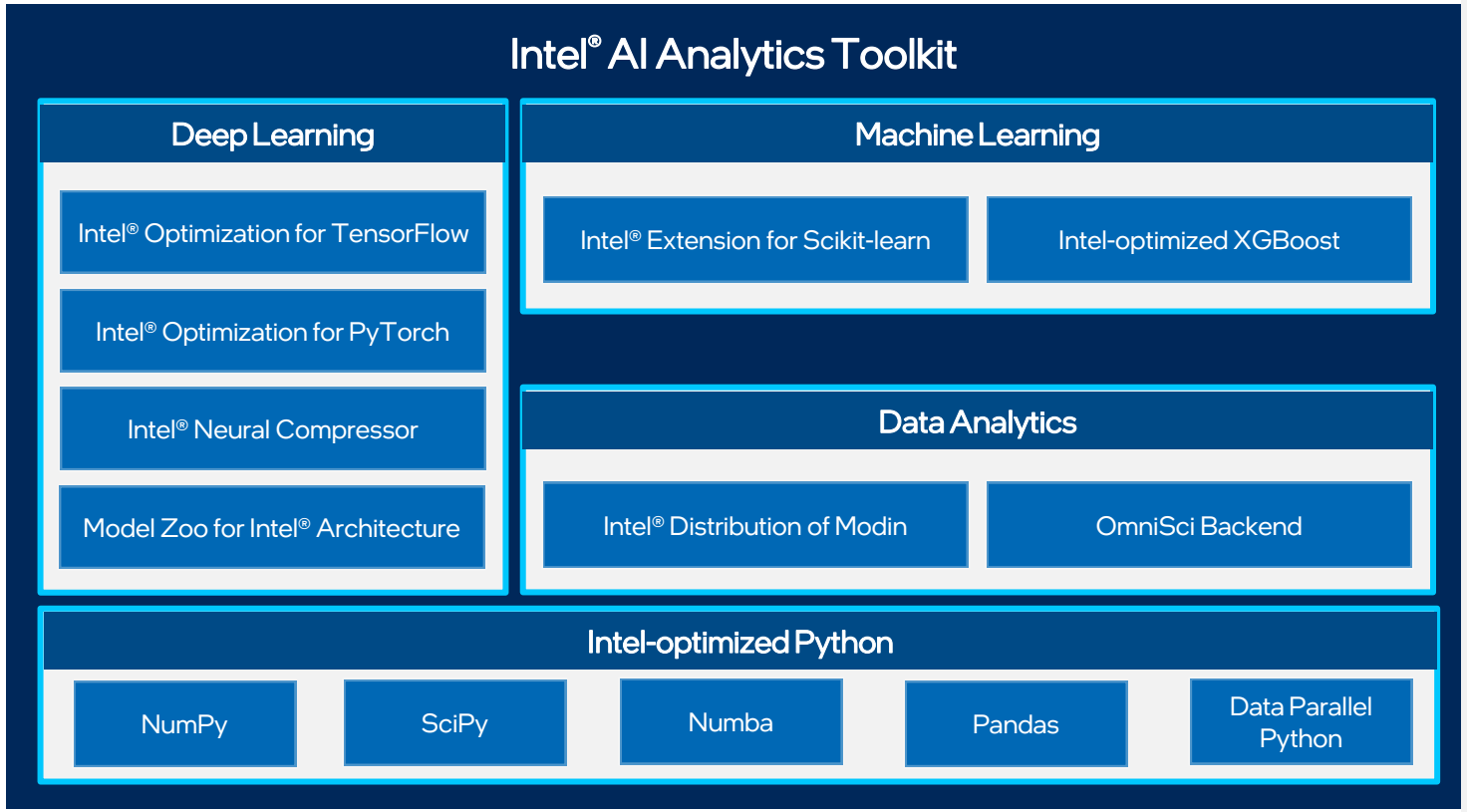
Accelerate end-to-end AI and data analytics pipelines with libraries optimized for Intel® architectures

Who needs this product?

Data scientists, AI researchers, ML and DL developers, AI application developers

Top Features/Benefits

- Deep learning performance for training and inference with Intel optimized DL frameworks and tools
- Drop-in acceleration for data analytics and machine learning workflows with compute-intensive Python packages



CPU GPU

Hardware support varies by individual tool. Architecture support will be expanded over time.

Get the Toolkit HERE or via these locations

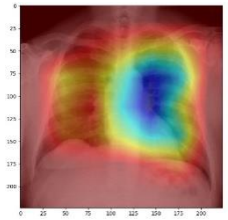
Intel Installer | Docker | Apt, Yum | Conda | Intel® DevCloud

Achieve End-to-End Performance for AI Workloads–Customer Use Cases

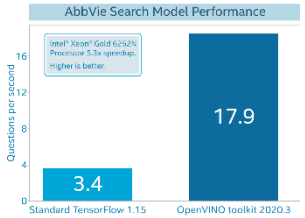
Accelerate Training + Inference - Most are optimized by Intel® AI Analytics Toolkit

May also use Intel® Distribution of OpenVINO™ toolkit, Intel® oneAPI Base Toolkit, or Intel® oneAPI Rendering Toolkit

SAMPLE USE CASES & PROOF POINTS > View More Case Studies

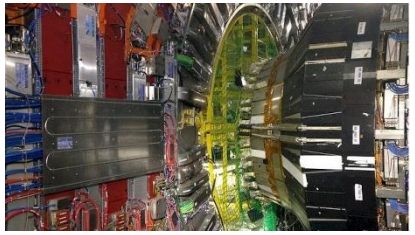


Accrad AI-based Solution Helps Accelerate Lung Disease Diagnosis – Acceleration for training + inference



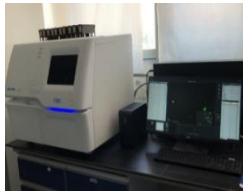
AbbVie Machine Translation Solution accelerates natural language processing inference models using processor optimized capabilities

Optimized by Intel® AI Analytics Toolkit & Intel® Distribution of OpenVINO™ toolkit



Acceleration for HPC & AI Inferencing - CERN, SURFsara, & Intel are driving breakthrough performance on scientific, engineering, and financial simulations. Includes strong inference benchmark.

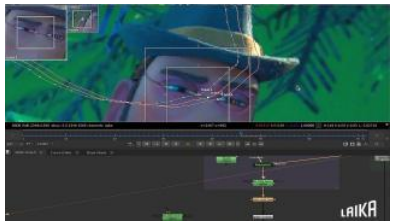
30 Days to AI value – Aible has optimized the machine learning & AI tasks, Its AI portfolio of products targets the end-to-end AutoML lifecycle help Enterprise achieve significant performance improvements and speed in implementation.



KFBIO AI-based Solution Helps Accelerate Tuberculosis Diagnosis demonstrating Intel inference performance leadership

AsiaInfo AI-based Solution optimized the AI end-to-end workflow for performance, and helps Accelerate 5G Network Intelligence

Bentley Motors Limited's Car Configurator uses real-time, accurate, visualization & AI for 1.7M+ rendered images in providing options to customers.
PR | Video | Demo | Session



LAIKA Studios & Intel Join Forces to Expand What's Possible in Stop-Motion Filmmaking – decreases production time by transitioning manual work to AI-based processing
Article | Video

Intel & Facebook Accelerate PyTorch Performance on training workloads using Intel® Deep Learning Boost on 3rd gen Intel® Xeon® Scalable processors

Red Hat Optimizing Data Science Workflows
Red Hat Optimizing Data Science Workflows

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Intel® oneAPI Rendering Toolkit

Powerful Libraries for High-Fidelity Visualization Applications

- Deliver high-performance, high-fidelity visualization applications on Intel® architecture
- Create amazing visual, hyper-realistic renderings via ray tracing with global illumination
- Access all system memory space to create renderings using the largest data sets
- Flexible, cost-efficient development using open-source libraries

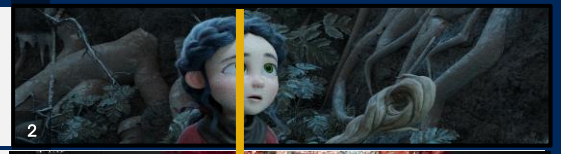
YOUR WORK. MORE VIVID.

intel. 1 oneAPI RENDERING TOOLKIT

Intel oneAPI Rendering & Ray Tracing Libraries

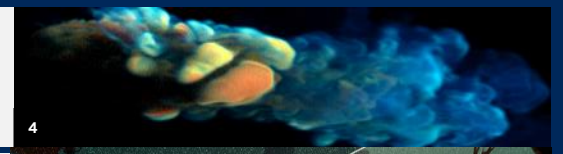
Intel® Embree
Award winning, High-Performance, Feature-Rich Ray Tracing & Photorealistic Rendering

Intel® Open Image Denoise
AI-Accelerated Denoiser for Superior Visual Quality



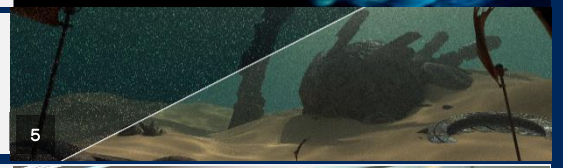
Intel® OpenSWR
High-Performance, Scalable, OpenGL-Compatible Rasterizer

Intel® Open Volume Kernel Library
Render & Simulate 3D Spatial Data Processing



Intel® Open Path Guiding Library
Path Tracing Importance Sampling, higher fidelity at same or lower compute cost

Beta Released Q1'22!



Intel® OSPRay
Scalable, Portable, Distributed Rendering API
HPC, SciVis and Pro Render Path Tracer

Intel® OSPRay Studio
Real-time rendering through a graphical user interface with this new scene graph application

Intel® OSPRay for Hydra
Connect the Rendering Toolkit libraries to Universal Scene Description Hydra Rendering subsystem via plugin

Learn More: intel.com/oneAPI-RenderKit

Image Courtesy
[1] The Addams Family 2 - Digital Domain, Marvel Studios, Chaos Group V-Ray
[2] Scene © Blender Foundation | cloud.blender.org/spring
[3] Model from Leigh Orf at University of Wisconsin. For more tornado visualization, visit Leigh Orf's site
[4] Smoke volume, data courtesy OpenVDB example repository
[5] Blender Institute: Nishita Demo (CC-BY-ND)
[6] Copyright Animal Logic Pty Limited 2021. All rights reserved.

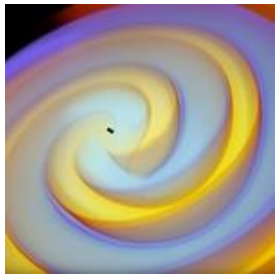
Intel® oneAPI Rendering Toolkit

Render Your Vision in Highest Fidelity: Your Open Path to Advanced Ray Tracing

intel. 1 oneAPI RENDERING TOOLKIT

SCIENTIFIC & TECHNICAL AWARDS Intel® Embree

SAMPLE USE CASES & PROOF POINTS > View more case studies



Stephen Hawking Centre for Cosmology Visualizes Cosmos Physics

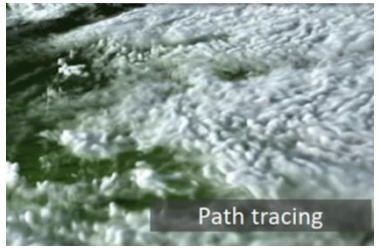
Chaos Group
Case Study | Video [2:22] | Session [17:42]

Maxon's Cinema 4D Enhances CG Workflows with AI-Trained Denoising
Case Study | Session CG Workflow [7:03] | Tutorial Scene Creation/3D Scans [16:33]

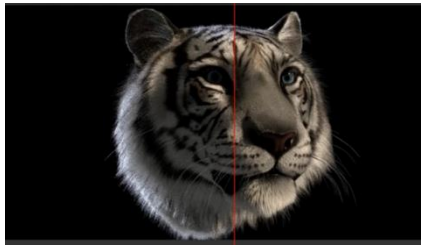
Bentley Motors Limited's Car Configurator uses visualization & AI for 1.7M+ rendered images
PR | Video [2:31] | Demo [6:58] | Session [10:37]

For The Addams Family 2, Cinesite gained 10% to up to 25% efficiency in rendering.

Tangent Studios gained 5X-6X reduction in renders using Intel® Embree & AI-based Intel® Open Image Denoise
Video [2:35]

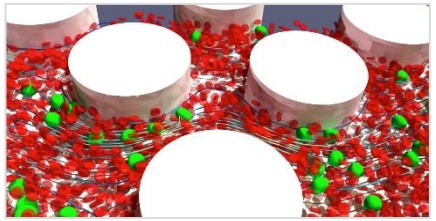


Texas Advanced Computing Center (TACC) Frontera SuperComputer Visualization & Filesystem Use Cases Show Value of Large Memory Fat Nodes



Mercenaries Engineering is delivering AI-based Animation & VFX through its production-ready tools and improved rending times by up to 138% in renders.*

View a Customer Showcase Reel [Video 2:34]



Intel oneAPI tools help prepare code for Aurora.
Aurora, Argonne Leadership Computing Facility's Intel-HPE/Cray supercomputer, will be one of the U.S.'s 1st exascale systems

[1]Courtesy Baozou Production in association with Tangent Animation using Blender with Intel® Embree. Media courtesy of Netflix, Inc. Now streaming on Netflix. Netflix subscription required.
*See slide notes for configuration details.
Refer to software.intel.com/articles/optimization-notice for more information regarding performance & optimization choices in Intel software products.
For workloads and configurations visit www.Intel.com/PerformanceIndex. Results may vary. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy. .

Intel® oneAPI Tools for IoT

Intel® IoT Toolkit

Accelerate development of IoT applications for smart, connected devices that run at the network's edge for healthcare, smart homes, industrial, retail, aerospace, and more.

Who needs this product?

- OEMs, ODMs, SIs, ISVs
- C, C++, SYCL, OpenMP, Python* Developers

Top Features/Benefits

- Leverage more cores & built-in technologies in IA-based platforms through optimized compilers & libraries
- Easily connect sensors to devices, and devices to cloud with IoT Connection Tools
- Speed development & maintenance of Yocto Project platform projects
- Develop with confidence with powerful analysis tools to identify threading, memory & offloading optimization opportunities
- DPC++ compatibility tool helps migrate existing code written in CUDA

Learn More & Download

Intel oneAPI Tools for IoT Development

Direct Programming	API-Based Programming	Analysis & debug Tools
Intel® C++ Compiler Classic	Intel® oneAPI DPC++ Library oneDPL	Intel® Inspector
Eclipse IDE	Intel® oneAPI Math Kernel Library oneMKL	Intel® VTune™ Profiler
Linux Kernel Build Tools	Intel® oneAPI Data Analytics Library - oneDAL	Intel® Advisor
Intel® oneAPI DPC++ /C++ Compiler	Intel® oneAPI Threading Building Blocks - oneTBB	Intel® Distribution for GDB
Intel® DPC++ Compatibility Tool	Intel® oneAPI Video Processing Library - oneVPL	
Intel® Distribution for Python	Intel® oneAPI Collective Communications Library oneCCL	
Intel® FPGA Add-on for oneAPI Base Toolkit	Intel® oneAPI Deep Neural Network Library - oneDNN	
	Intel® Integrated Performance Primitives – Intel® IPP	

Intel® oneAPI **IoT** Toolkit +

Intel® oneAPI **Base** Toolkit +

intel. 1 oneAPI IoT TOOLKIT

Intel® Distribution of OpenVINO™ toolkit

Powered by oneAPI

Deliver High-Performance Deep Learning Inference

A toolkit to accelerate development of high-performance deep learning inference & computer vision in vision/AI applications used from edge to cloud. It enables deep learning on hardware accelerators & easy deployment across Intel® CPUs, GPUs, FPGAs, VPUs.

Who needs this product?

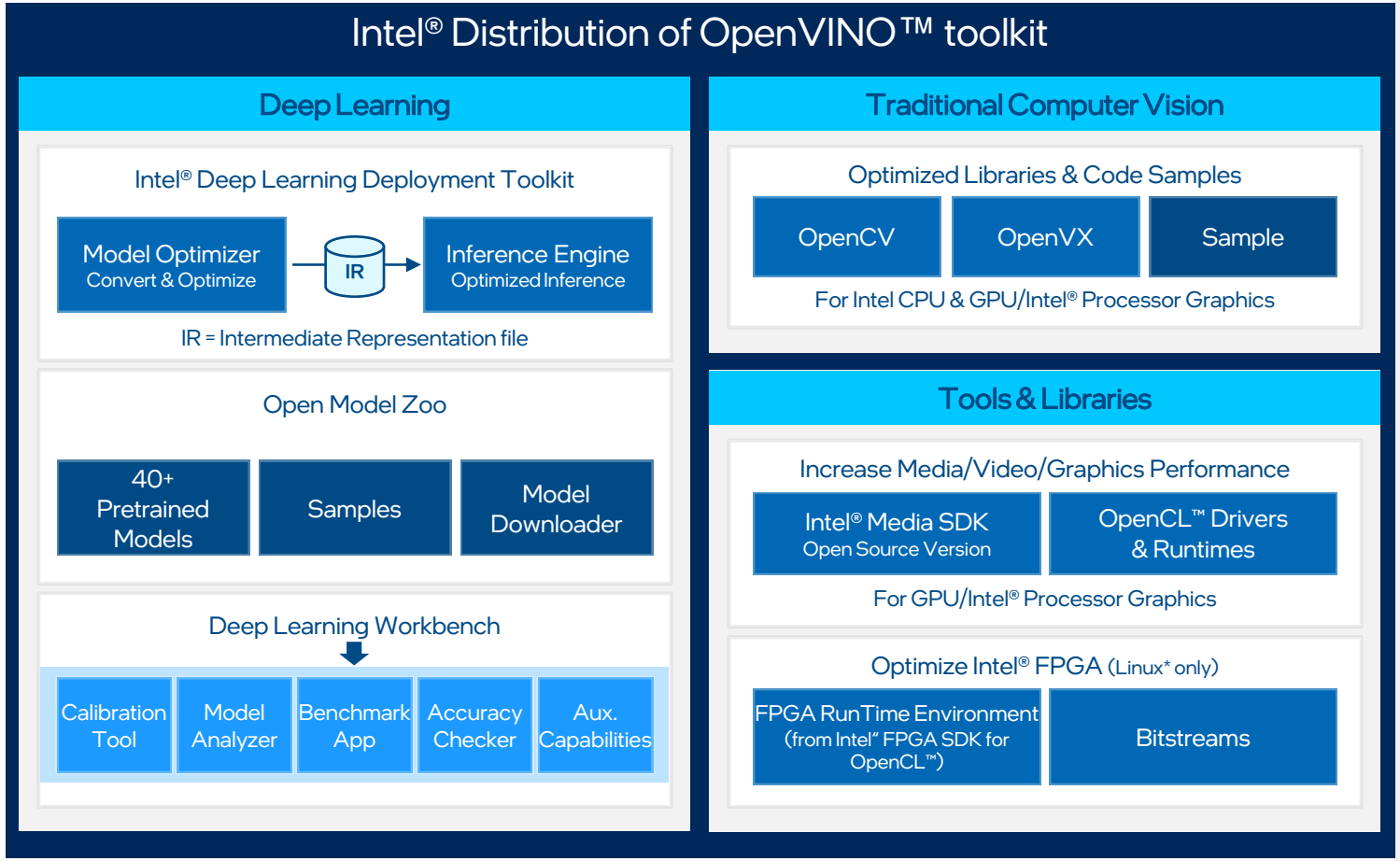
- Computer vision, deep learning software developers
- Data scientists
- OEMs, ISVs, System Integrators

Usages

Security surveillance, robotics, retail, healthcare, AI, office automation, transportation, non-vision use cases (speech, NLP, Audio, text) & more

Learn More & Download

Edge AI & Vision Alliance 2019 VISION PRODUCT OF THE YEAR



Accelerate Development of Smart, Connected Devices

Customer Use Cases

May be optimized one or a combination of the Intel® oneAPI Base & IoT Toolkits, Intel® AI Analytics Toolkit, & Intel® Distribution of OpenVINO™ toolkit

SAMPLE USE CASES & PROOF POINTS

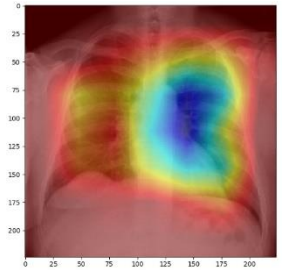
Samsung Medison Uses oneAPI to Power Obstetric Ultrasound Systems

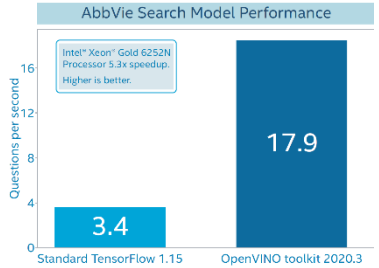
Intel® oneAPI Base Toolkit & Intel® Distribution of OpenVINO™ toolkit help accelerate image processing at the edge for consistent measurement accuracy & improved workflows.[1]

Intel PR News Byte 09/10/20 | Video

United Imaging Successfully Ported CUDA* Code to oneAPI

United Imaging develops advanced medical imaging diagnosis and treatment products and innovative medical IT solutions. It used Intel oneAPI Base Toolkit for code migration and optimizations.





Optimized by Intel® oneAPI Analytics Toolkit & Intel® Distribution of OpenVINO™ toolkit

Accrad AI-based Solution Helps Accelerate Lung Disease Diagnosis – Acceleration for training + inference.

Learn more in the solution brief

AbbVie Machine Translation Solution accelerates natural language processing inference models using processor optimized capabilities.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Notices & Disclaimers

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex. Results may vary.

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.

Texas Advanced Computing Center (TACC) Frontera references
Article: *HPCWire: Visualization & Filesystem Use Cases Show Value of Large Memory Fat Notes on Frontera*.
www.intel.com/content/dam/support/us/en/documents/memory-and-storage/data-center-persistent-mem/Intel-Optane-DC-Persistent-Memory-Quick-Start-Guide.pdf
software.intel.com/content/www/us/en/develop/articles/introduction-to-programming-with-persistent-memory-from-intel.html
wreda.github.io/papers/assise-osdi20.pdf

KFBIO
KFBIO m. tuberculosis screening detectron2 model throughput performance on 2nd Intel® Xeon® Gold 6252 processor: NEW: Test 1 (single instance with PyTorch 1.6: Tested by Intel as of 5/22/2020. 2-socket 2nd Gen Intel® Xeon® Gold 6252 Processor, 24 cores, HT On, Turbo ON, Total Memory 192 GB (12 slots/16 GB/2666 MHz), BIOS: SSE5C620.86B.02.01.0008.031920191559 (ucode: 0x500002c), Ubuntu 18.04.4 LTS, kernel 5.3.0-51-generic, mitigated Test 2 (24 instances with PyTorch 1.6: Tested by Intel as of 5/22/2020. 2-socket 2nd Gen Intel Xeon Gold 6252 Processor, 24 cores, HT On, Turbo ON, Total Memory 192 GB (12 slots/16 GB/2666 MHz), BIOS: SSE5C620.86B.02.01.0008.031920191559 (ucode: 0x500002c), Ubuntu 18.04.4 LTS, kernel 5.3.0-51-generic, mitigated BASELINE: (single instance with PyTorch 1.4): Tested by Intel as of 5/22/2020. 2-socket 2nd Gen Intel Xeon Gold 6252 Processor, 24 cores, HT On, Turbo ON, Total Memory 192 GB (12 slots/16 GB/2666 MHz), BIOS: SSE5C620.86B.02.01.0008.031920191559 (ucode: 0x500002c), Ubuntu 18.04.4 LTS, kernel 5.3.0-51-generic, mitigated.

Tangent Studios
Configurations for Render Times with Intel® Embree, testing conducted by Tangent Animation Labs. Render farm: 8x Intel® Core™ processors +hyperthread*2 + 128gig. In-office workstations: Intel® Xeon® processors HP blade c7000 chassis, with HP460 gen8 blades - 2x Intel Xeon E5-2650 V2, Eight Core 2.6GHz-128GB. Software: Blender 2.78 with custom build using Intel® Embree. For more information on Tangent's work with Embree, watch this video:
www.youtube.com/watch?time_continue=251&v=_2Ia4h8q3xs&feature=emb_logo
Recreation of the performance numbers can be recreated using Agent327, Blender and Embree.

Chaos Group - Up to 90% Memory Reduction for Displacement
Testing conducted by Chaos Group with Intel® Embree 2020. Software Corona Renderer 5 with Intel Embree. Up to 90% memory reduction calculated using Corona Renderer 5 with regular displacement grids per triangle of 154 bytes versus Corona Renderer 5 with Intel Embree, which has a displacement capability grid of 12 bytes per grid triangle. (12/154 = 7.8% usage or >90% memory reduction.) Recreation of the performance numbers can be accomplished using Corona Renderer 5 and Embree. For more information, visit the Corona Renderer Blog: blog.corona-renderer.com/corona-renderer-5-for-3ds-max-released/

The Addams Family 2 - **Gained a 10% to 20%—and sometimes 25%—efficiency in rendering, saving thousands of hours in rendering production time**.
Testing Date: Results are based on data conducted by Cinesite 2020-21. 10% to up to 25% rendering efficiency/thousands of hours saved in rendering production time/15 hrs per frame per shot to 12-13 hrs.
Cinesite Configuration: 18-core Intel® Xeon® Scalable processors (W-2295) used in render farm, 2nd gen Intel Xeon processor-based workstations (W-2135 and -2195) used. Rendering tools: Gaffer, Arnold, along with optimizations by Intel® Open Image Denoise.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

intel®

oneAPI Ecosystem Support

These organizations support the oneAPI initiative for a single, unified programming model for cross-architecture development.
It does not indicate any agreement to purchase or use of Intel's products. *Other names and brands may be claimed as the property of others.

Intel® oneAPI Toolkits – Proven Performance

Top Customer Use Cases

- HPC Cross-architecture
  - Argonne National Labs is running exascale-class applications efficiently on current and future Intel CPUs and GPUs.
  - Zuse Institute Berlin (ZIB) ported the tsunami simulation *easyWave* application from CUDA to DPC++ delivering performance across multiple architectures and vendors.
  - GROMACS, accelerated by oneAPI open programming and multiarchitecture tools, runs on Intel Xe architecture-based GPUs with strong performance.
- HPC & AI – CERN uses Intel® DL Boost and oneAPI to speed simulations with inference acceleration by nearly 2x without accuracy loss*
- High-Fidelity Visualization Using Advanced Ray Tracing – Bentley Motors Limited's AI-based car configurator processes 1.7M+ images with up to 10B possible configurations per model. The Addams Family 2, Cinesite used Intel's denoiser achieving 25% rendering efficiency.
- IoT – Samsung Medison accelerates ultrasound image processing at the edge on multiple Intel® architectures for improved accuracy and fast diagnosis
- Major CSPs & Framework endorse oneAPI – Microsoft Azure, Google Cloud, TensorFlow
- FPGA – Using oneAPI, Bittware had its application running **in days** vs. what typically would take several weeks using Verilog or VHDL*
- And more… 250+ applications developed using oneAPI tools > view catalog

Video [3:45]

*Detailed slides per customer are noted in the oneAPI Customer Use Cases deck. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy. See Notices & Disclaimers for more details.

Intel® oneAPI Toolkits Free Availability

Get Started Quickly

Code Samples, Quick-start Guides, Webinars, Training

software.intel.com/oneapi

Run the tools locally

- Downloads
- Repositories
- Containers

Or run the tools in Intel® Developer Cloud

- No hardware acquisition
- No download, install or configuration
- Sample code & documentation
- Ready-to-use deployment & development environments
- Beta trial open now

Get Up & Running In Seconds!

Intel® oneAPI Priority Support

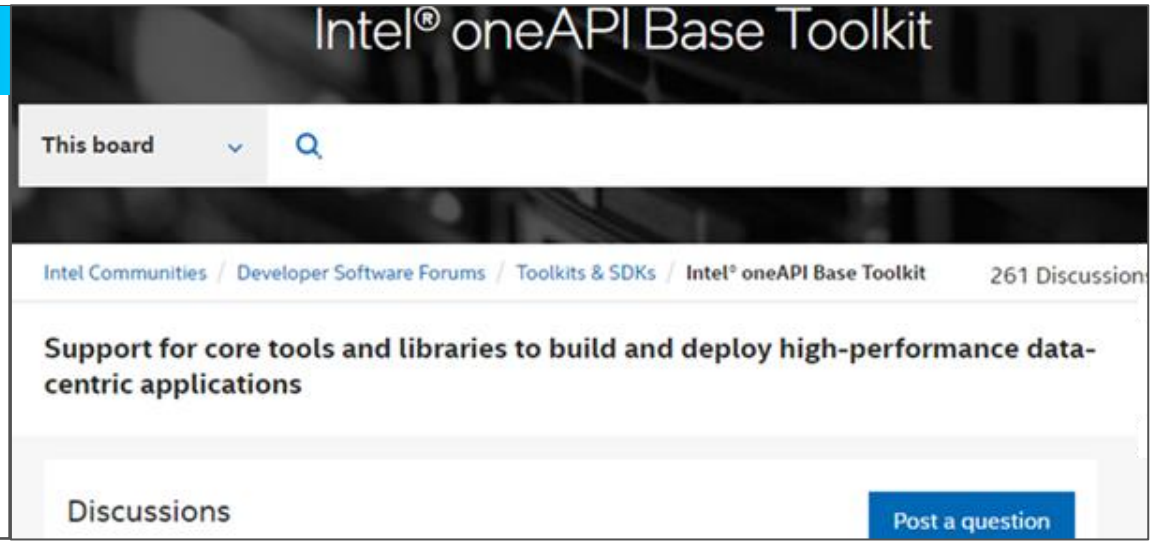
Priority Support for Intel® oneAPI Toolkits

- Every paid version of Intel® oneAPI Developer Toolkits includes Priority Support for that toolkit (Intel oneAPI Base, HPC, IOT, & Rendering Toolkits)
  - Direct, private interaction with Intel software support engineers
  - Accelerated response time
  - Access to—and support for—previous Intel products such as Fortran compiler versions, previous toolkit versions, and more
  - Intel Technical Consulting Engineers for on-site or online training and consultation at a reduced cost

Intel® oneAPI Community

Free Intel® Community Support

- Support via the Intel public Community Forum
- Access to only the latest version of oneAPI Toolkits
- Access to online tutorials and self-help forums



Why SYCL?

```
sycl::queue q(cpu_selector{});

auto A = sycl::malloc_shared<float>(n, q);
auto B = sycl::malloc_shared<float>(n, q);

q.parallel_for( sycl::range<1>{n},
  [=] (sycl::id<1> i) {
      B[i] += A[i] * A[i];
  }
).wait();
```

```
sycl::queue q(gpu_selector{});

auto A = sycl::malloc_shared<float>(n, q);
auto B = sycl::malloc_shared<float>(n, q);

q.parallel_for( sycl::range<1>{n},
  [=] (sycl::id<1> i) {
      B[i] += A[i] * A[i];
  }
).wait();
```

- Standard C++17 aids time to developer productivity
- Syntax for accelerators (device selection, offload, memory transfer)
- Unified shared memory
- Single source (host & device code)
- Multi-Architecture (CPU, GPU, FPGA & other targets)
- Stack based on standards & open specifications (CLANG, LLVM, SPIR-V, Level Zero)

*Example uses static **targets in bold**, but can be programmed to be dynamic as well

Data Parallel C++:
oneAPI's implementation of SYCL

DPC++ = ISO C++ and Khronos SYCL and community extensions

Freedom of Choice: Future-Ready Programming Model

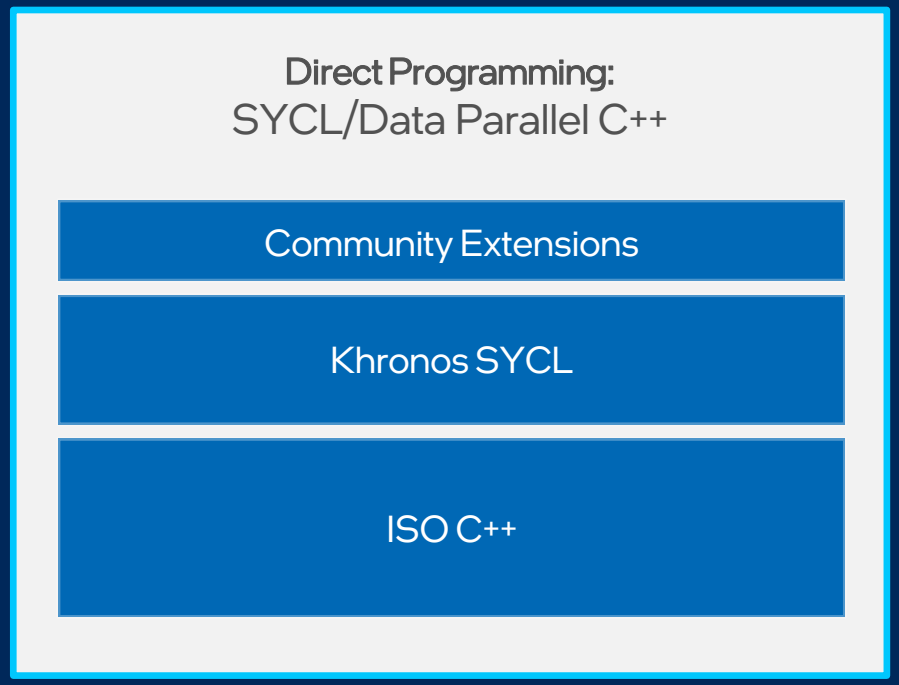
- Allows code reuse across hardware targets
- Permits custom tuning for a specific accelerator
- Open, cross-industry alternative to proprietary language

DPC++ = ISO C++ and Khronos SYCL and community extensions

- Designed for data parallel programming productivity
- Provides full native high-level language performance on par with standard C++ and broad compatibility
- Adds SYCL from the Khronos Group for data parallelism and heterogeneous programming

Community Project Drives Language Enhancements

- Provides extensions to simplify data parallel programming
- Continues evolution through open and cooperative development



The open source and Intel DPC++/C++ compiler supports Intel CPUs, GPUs, and FPGAs.
Codeplay announced a DPC++ compiler that targets Nvidia GPUs.