oneAPI software stack: ONNX support for xPU hardware

Intel® oneAPI Base Toolkit

DIRECT PROGRAMMING

Intel® oneAPI DPC++ Compiler

Intel® DPC++ Compatibility Tool

Intel® Distribution for Python*

Intel® FPGA Add-on for oneAPI Base Toolkit

Intel® C++ Compiler with OpenMP*

Intel® Fortran Compiler with OpenMP*

Intel® C++ Compiler

Eclipse* IDE

Linux* Kernel Build Tools

API-BASED PROGRAMMING

Intel® oneAPI DPC++ Library

Intel® oneAPI Math Kernel Library

> Intel® oneAPI Data Analytics Library

Intel® oneAPI Threading Building Blocks

Intel® oneAPI Video Processing Library

Intel® oneAPI Collective Comms. Library

Intel® oneAPI Deep Neural Network Library

Intel® Integrated Performance Primitives

Intel® MPI Library

IoT Connection Tools

ANALYSIS & DEBUG TOOLS

Intel® VTune™ Profiler

Intel® Advisor

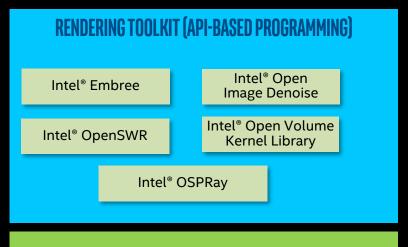
GDB*

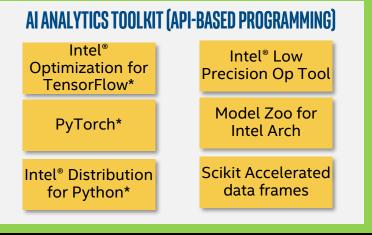
Intel® Inspector

Intel® Trace Analyzer & Collector

Intel® Cluster Checker

Intel® System Debugger







Rendering Toolkit +



AI Analytics Toolkit +

oneDNN

- Intel® oneAPI Deep Neural Network Library
- improves productivity and enhance performance of deep learning frameworks
- supports key data type formats, e.g. fp16, fp32, bfloat16, and int8
- implements ops, e.g. convolution, matrix multiplication, pooling, batch normalization, activation functions
- supports DL instructions and accelerators in Intel hardware, e.g. DL Boost (VNNI), AMX/TMUL, Intel® GPUs

Why onnxruntime with oneDNN

- Intel cpu and gpu will have accelerators for deep learning software
- oneDNN provides a unified interface to utilize these accelerators
- oneDNN library abstracts away complexity of programming to the accelerators

onnxruntime one DNN execution provider features

- **2**020
 - 32 bit floating point, inference, CNN, CPU
 - didn't have GPU support
- **2**021
 - GPU, NLP/transformer, training, int8