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 (API-BASED PROGRAMMING)

- Intel® Embree
- Intel® Open Image Denoise
- Intel® OpenSWR
- Intel® Open Volume Kernel Library
- Intel® OSPRay

AI ANALYTICS TOOLKIT (API-BASED PROGRAMMING)

- Intel® Optimization for TensorFlow*
- Intel® Low Precision Op Tool
- PyTorch*
- Model Zoo for Intel Arch
- Intel® Distribution for Python*
- Scikit Accelerated data frames

OneAPI base

HPC Toolkit +

IOT Toolkit +

Rendering Toolkit +

AI Analytics Toolkit +

Github: oneAPI Open Source | oneAPI

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 oneDNN execution provider features

- 2020
  - 32 bit floating point, inference, CNN, CPU
  - didn't have GPU support
- 2021
  - GPU, NLP/transformer, training, int8