oneAPI software stack: ONNX support for xPU hardware
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