oneAPI

AI Workshop - oneAPI DevSummit 2023

Intel® Optimizations for TensorFlow

Anish Kumar
Vishnu Madhu

Agenda

- Workshop pre-requisites
- Intel® AI Optimizations
- AI Handson Workshop

Workshop Pre-requisites

- Register for accessing Intel® Developer Cloud (5 mins)
  visit → cloud.intel.com
  Sign up --> Create Account
- Setup SSH access to Intel® Developer Cloud (5 mins)
  Login to Intel® Developer Cloud
- Clone the workshop repository on Developer cloud
  https://tinyurl.com/oneapi-ai-workshop
- Laptop with open internet access (preferred)

Intel® Developer Cloud

Intel® Developer Cloud

a service platform for developing and running workloads in Intel®-optimized deployment environments with the latest Intel® processors

oneAPI

- Landing page :
  - https://cloud.intel.com
- Instructions to get started:
  - **http://tinyurl.com/ReadmeIDC**

oneAPI

Intel® AI Optimizations

Diverse Compute Requirements

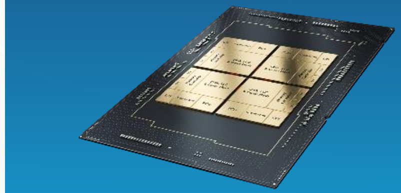
Diverse accelerators needed to meet today's performance requirements:
48% of developers target heterogeneous systems
that use more than one kind of processor or core[1]

Developer Challenges: Multiple Architectures, Vendors, and Programming Models

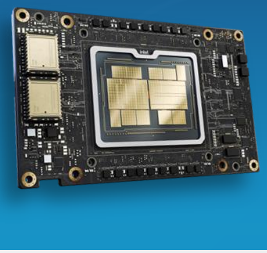
oneAPI

Open, Standards-based, Multiarchitecture Programming

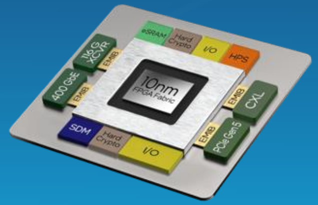
CPU



GPU



FPGA



Other Accelerators

[1] - https://evansdata.com/reports/viewRelease.php?reportID=40

oneAPI Industry Initiative

Break the Chains of Proprietary Lock-in

Freedom to Make Your Best Choice

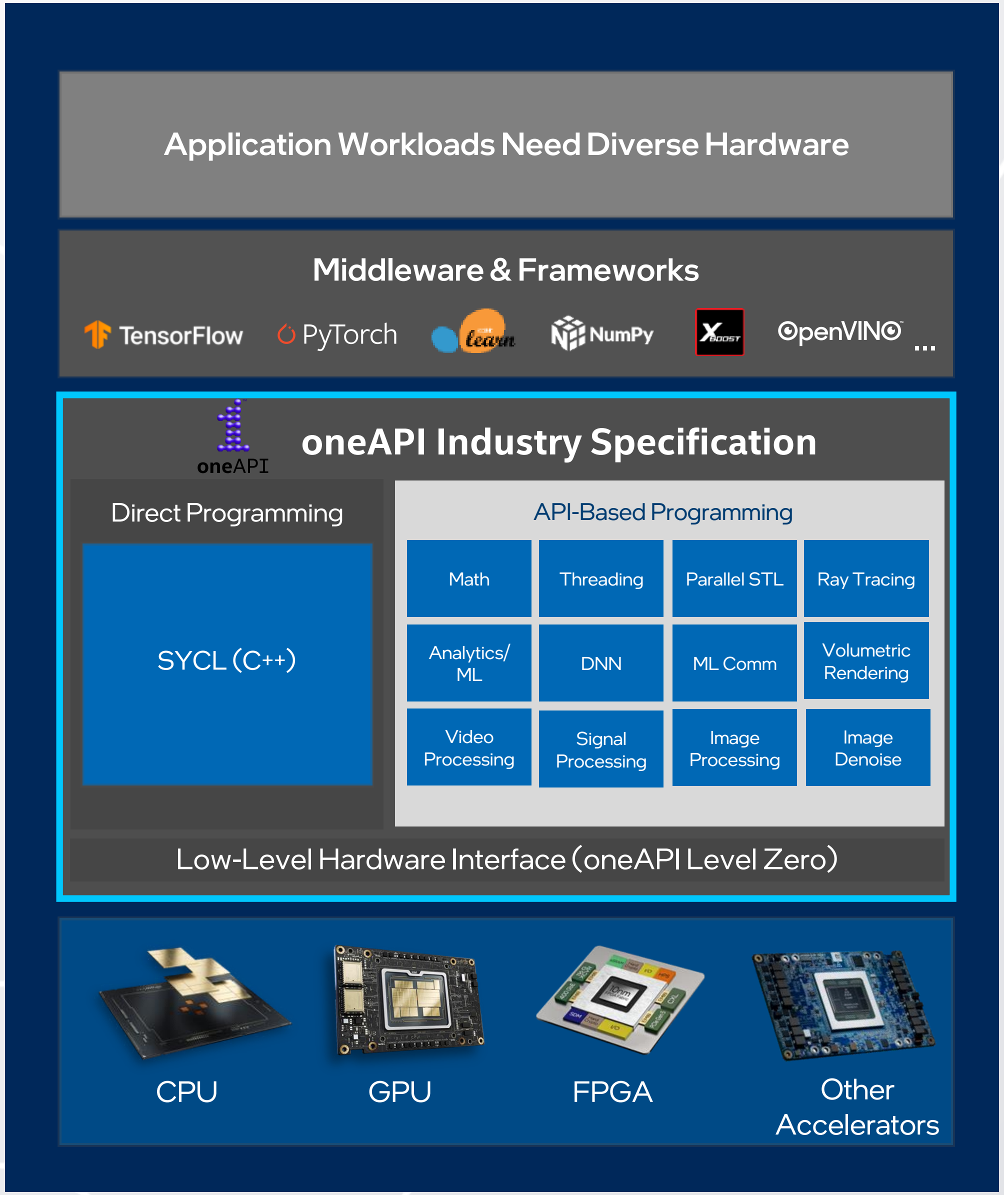
- C++ programming model for multiple architectures and vendors
- Cross-architecture code reuse for freedom from vendor lock-in

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
- Interoperable with familiar languages and programming models including Fortran, Python, OpenMP, and MPI
- Powerful libraries for acceleration of domain-specific functions



Application Workloads Need Diverse Hardware

Middleware & Frameworks

TensorFlow, PyTorch, learn, NumPy, XBoost, OpenVINO ...

oneAPI Industry Specification

Direct Programming

SYCL (C++)

API-Based Programming

Math, Threading, Parallel STL, Ray Tracing

Analytics/ML, DNN, ML Comm, Volumetric Rendering

Video Processing, Signal Processing, Image Processing, Image Denoise

Low-Level Hardware Interface (oneAPI Level Zero)

CPU, GPU, FPGA, Other Accelerators

oneAPI

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

oneAPI Industry Momentum

End Users

Accrad Accelerated Radiology

@rchanan

FOUNDRY. MODO

BioDataAnalysis GmbH GeoEast Next Generation Bio-Medical Image Analysis

BENTLEY

CINESITE

LAIKA

kt

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 AI

EURECOM Sophia Antipolis

PHILIPS

Brightskies technologies

大势智慧 DASPATIAL

WeBank

SAMSUNG MEDISON

TANGENTANIMATION

allegro.ai

ILLUMINATION MACGUFF

Verizon

MediaKind

National Labs

Argonne NATIONAL LABORATORY

UT-BATTELLE Oak Ridge National Laboratory

oneAPI

CERN openlab

CINECA

Peraton Labs

Laboratório Nacional de Computação Científica

lrz Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities

SANKHYA SUTRA LABS

ISVs & OSVs

codeplay®

CANONICAL

CERTIFACE

ANACONDA.

Ansys

MPCDF MAX PLANCK COMPUTING & DATA FACILITY

AI SINGAPORE

sas

Red Hat OpenShift Data Science

KATANA GRAPH

CGG

AIBLE

CHAOSGROUP

MAXON

Spirent

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 technology for health

Tech Mahindra

AsiaInfo 亚信科技

SUSE We adapt. You succeed.

vmware

UNITED IMAGING

OEMs & SIs

BittWare a molex company

Hewlett Packard Enterprise

DELL Technologies

Atos

Lenovo

HCL

ENIAC

MEGWARE SUPERCOMPUTING · TECHNOLOGY

Universities & Research Institutes

LOBACHEVSKY UNIVERSITY

TECHNION Israel Institute of Technology

UNIVERSITY OF CAMBRIDGE

ICT 中国科学院计算技术研究所 INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

UNIVERSIDAD COMPLUTENSE MADRID

ctg

Ben-Gurion University of the Negev

Berkeley UNIVERSITY OF CALIFORNIA

UCDAVIS UNIVERSITY OF CALIFORNIA

THE UNIVERSITY OF TENNESSEE KNOXVILLE

OLD DOMINION UNIVERSITY

北京大学软件与微电子学院 School of Software & Microelectronics

ILLINOIS

THE UNIVERSITY OF UTAH

UNIVERSITY OF OREGON

TÉCNICO LISBOA

inesc id lisboa

ZIB

FACULTY OF MATHEMATICS AND PHYSICS Charles University

PURDUE UNIVERSITY Elmore Family School of Electrical and Computer Engineering

Indian Institutes of Technology Delhi / Kharagpur / Roorkee

University College London

Durham University

University of Stuttgart Germany

SDSC SAN DIEGO SUPERCOMPUTER CENTER

सीडैक CDAC

UKRI Science and Technology Facilities Council Scientific Computing

TACC

Indian Institute of Science Bangalore

IISER PUNE Indian Institute of Science Education & Research Pune

UNIVERSIDAD DE MÁLAGA

URZ HEIDELBERG UNIVERSITY COMPUTING CENTRE

NIU 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

oneAPI DevSummit South-East Asia 2023

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

A complete set of proven developer tools expanded from CPU to Accelerators

Intel® oneAPI **Base** Toolkit

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

Add-on **Domain-specific** Toolkits

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

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

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

Toolkits **powered by oneAPI**

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

OpenVINO™

Intel® Distribution of OpenVINO™ Toolkit
Deploy high performance inference & applications from edge to cloud

Latest version available 2023.1

Intel® AI Analytics Toolkit

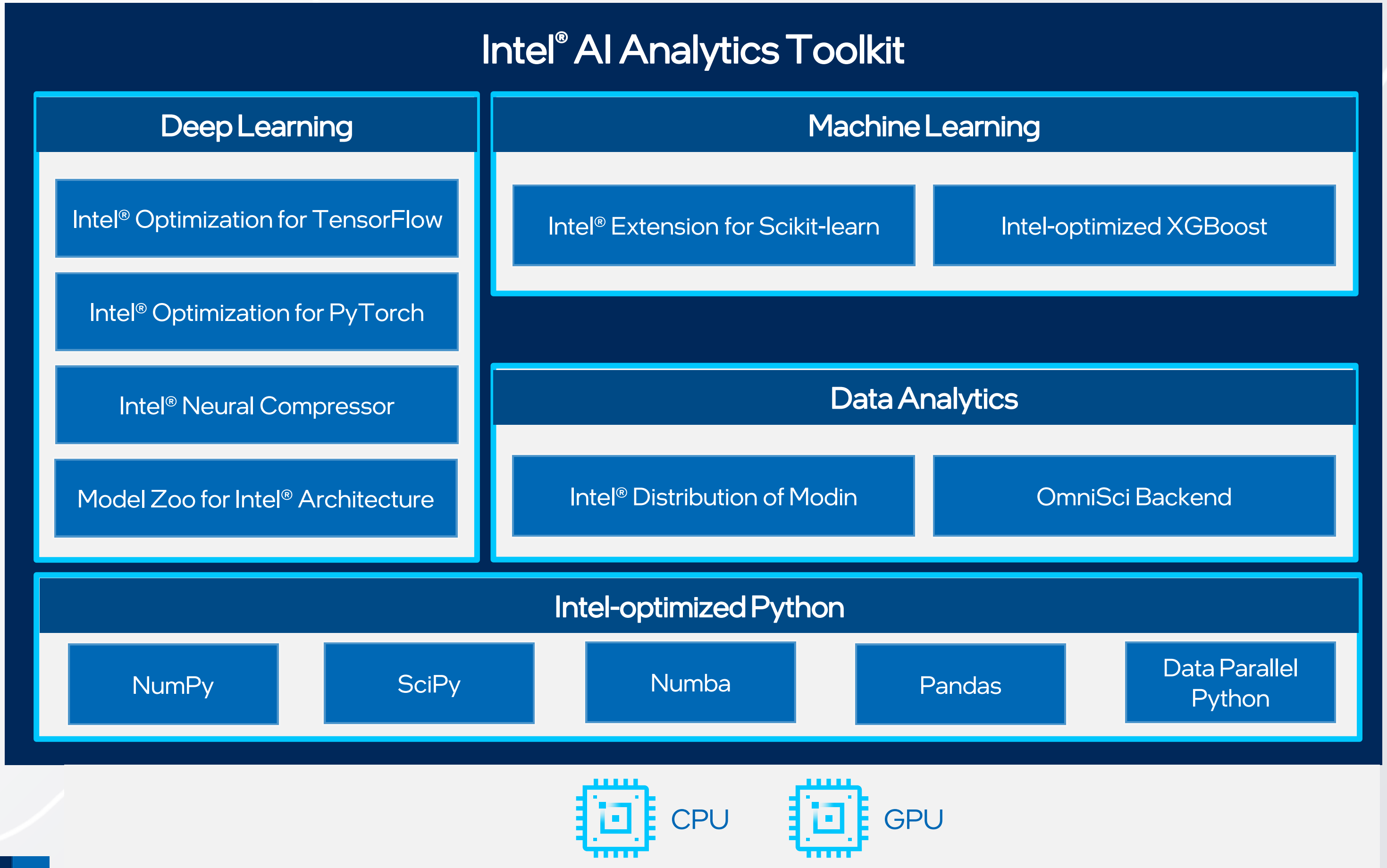
Accelerate end-to-end AI and data analytics pipelines with libraries optimized for Intel® architectures

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



Intel® AI Analytics Toolkit

Deep Learning
- Intel® Optimization for TensorFlow
- Intel® Optimization for PyTorch
- Intel® Neural Compressor
- Model Zoo for Intel® Architecture

Machine Learning
- Intel® Extension for Scikit-learn
- Intel-optimized XGBoost

Data Analytics
- Intel® Distribution of Modin
- OmniSci Backend

Intel-optimized Python
- NumPy
- SciPy
- Numba
- Pandas
- Data Parallel Python

CPU GPU

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

Get the Toolkit HERE or via these locations

https://software.intel.com/content/www/us/en/develop/tools/oneapi/download.html#aikit

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

oneAPI

Intel® Tensorflow Optimizations

Intel® Optimization for Tensorflow

What's New for TensorFlow Optimization?

- oneDNN is in official TensorFlow release!
- The platforms use the Intel® oneAPI Deep Neural Network Library (oneDNN), an open-source, cross-platform performance library for Deep-Learning applications
- Enable those Intel ® oneDNN CPU optimizations by setting the environment variable TF_ENABLE_ONEDNN_OPTS=1 for the official x86-64 TensorFlow after v2.5.
- Since TensorFlow **v2.9** and above, the oneAPI Deep Neural Network Library (oneDNN) optimizations are enabled by default

Features

- Operator optimizations: Replace default (Eigen) kernels by highly-optimized kernels (using Intel® oneDNN)
- Graph optimizations: Fusion, Layout Propagation
- System optimizations: Threading model

oneAPI

oneAPI Deep Neural Network Library (oneDNN)

Intel® Optimization for Tensorflow

Features

- Supports FP32, FP16, Bfloat16, and int8.
- Leverages Intel® DL Boost, AVX512 instructions and processor capabilities
- Fused operations for optimized performance

Support Matrix

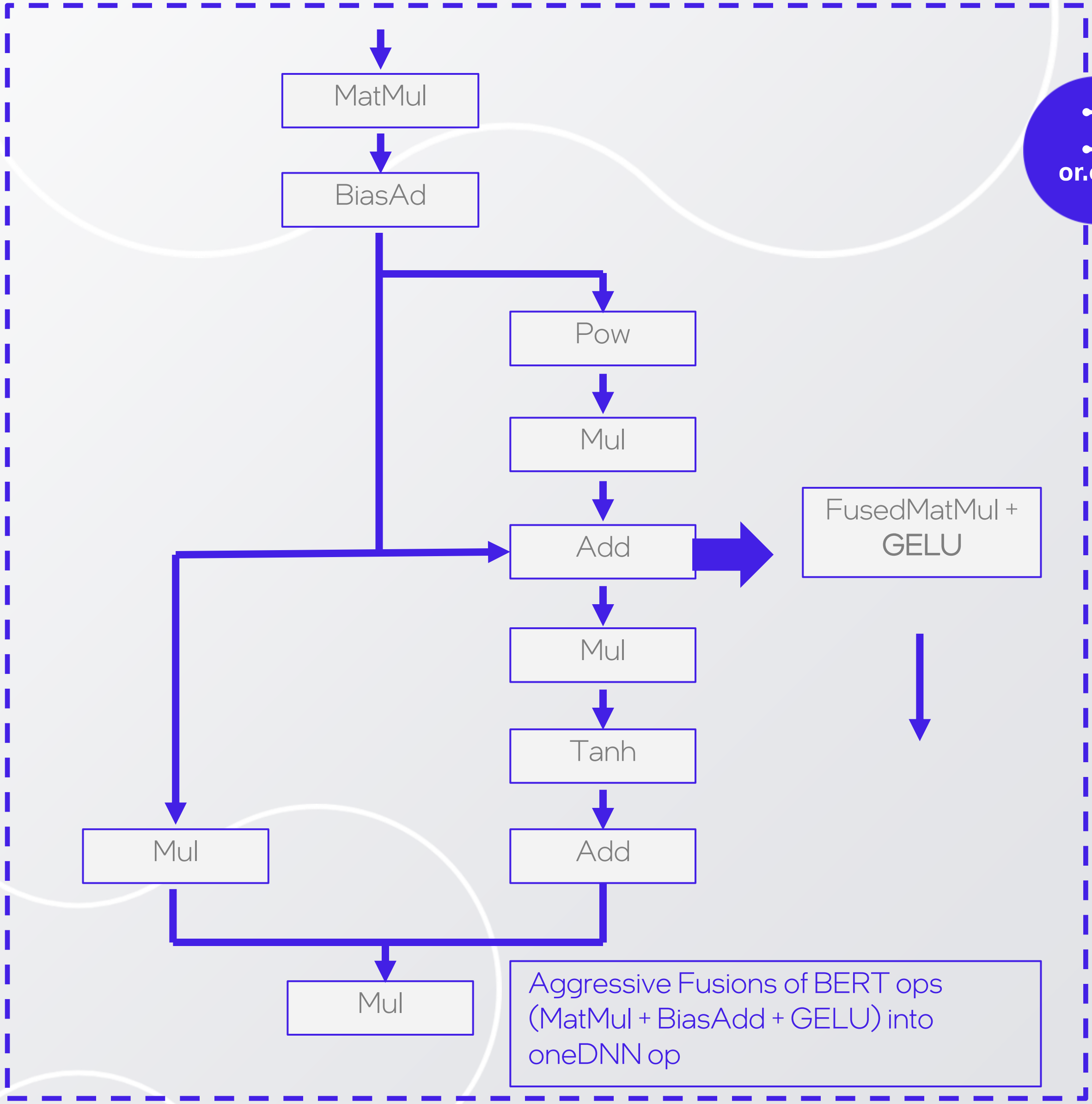
- Compilers: Intel® oneAPI DPC++ / C++ Compilers
- OS: Linux, Windows, macOS
- CPU: Intel ® Atom, Intel® Core™, Intel® Xeon®, Intel® Xeon® Scalable processors
- GPU: Intel® Processor Graphics Gen9, Intel® Processor Graphics Gen 12

Category	Functions
Compute intensive operations	• (De-)Convolution • Inner Product • RNN (Vanilla, LSTM, GRU) • GEMM
Memory bandwidth limited operations	• Pooling • Batch Normalization • Local Response Normalization • Layer Normalization • Elementwise • Binary elementwise • Softmax • Sum • Concat • Shuffle
Data manipulation	• Reorder

oneDNN Integration with TensorFlow

Features

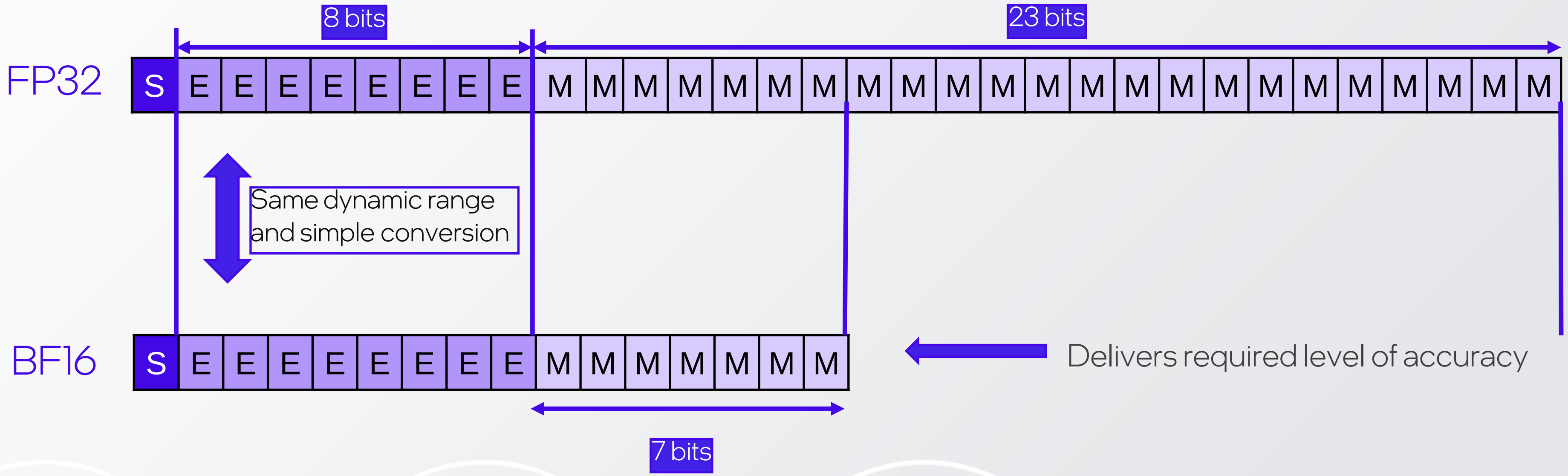
- Replaces compute-intensive standard TF ops with highly optimized custom oneDNN ops
- Aggressive op fusions to improve performance of Convolutions and Matrix Multiplications
- Primitive caching to reduce overhead of calling oneDNN Graphics Gen 12



Aggressive Fusions of BERT ops (MatMul + BiasAdd + GELU) into oneDNN op

https://www.intel.com/content/www/us/en/docs/onednn/developer-guide-reference/2023-1/gelu.html

BFloat16 Data Type

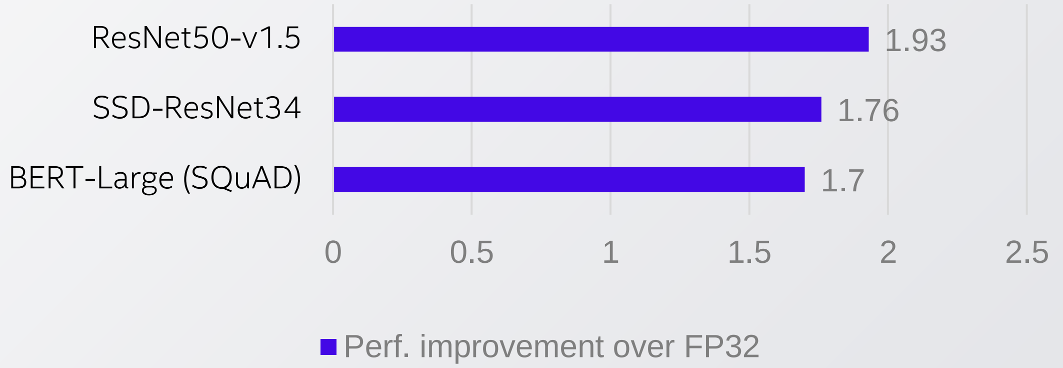

8 bits
23 bits
FP32
S E E E E E E E E M M M M M M M M M M M M M M M M M M M M M M M
Same dynamic range and simple conversion
BF16
S E E E E E E E E M M M M M M M
7 bits

Delivers required level of accuracy

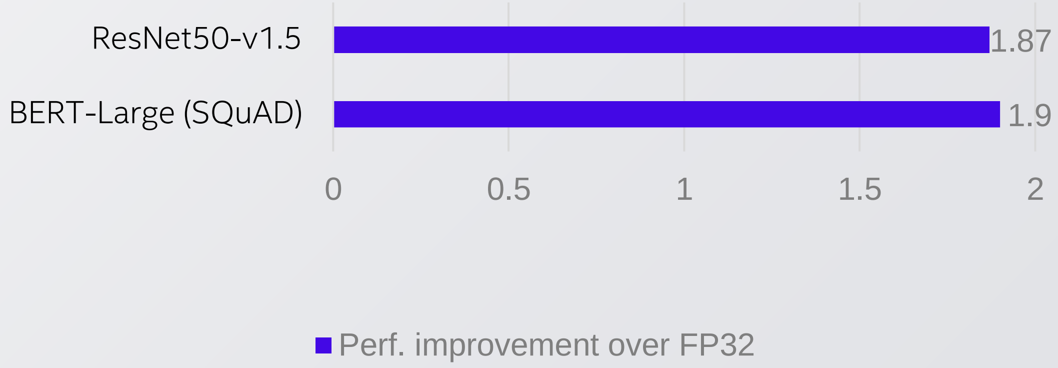
Bfloat16 Optimization

- Bfloat16 - 16-bit data type with the same dynamic range as FP32
- Benefits
  - Reduced bandwidth
  - Improved performance with hardware support
- Easy to use
  - No special handling for loss scaling
  - No hyperparameter tuning for training, can reuse FP32 hyperparameters
- Up to 2x improvement on training and inference with negligible accuracy loss (< 0.20%)
- AMP (Automatic Mixed Precision) in tensorflow automatically converts model to use bfloat16 data type.
- Supports both Keras and arbitrary graph based models.

Mixed precision training with Bfloat16



Inference with Bfloat16



https://blog.tensorflow.org/2020/06/accelerating-ai-performance-on-3rd-gen-processors-with-tensorflow-bfloat16.html

How to Install Intel ® optimization for Tensorflow

- Intel®optimization for Tensorflow is included in AI kit. If you have AI Analytics toolkit Intel-Tensorflow conda environment can be activated.
- Install via Pip: pip install intel-tensorflow==2.11.0
- For Stock-tensorflow: pip install tensorflow==2.11.0
  Since TensorFlow v2.9, the oneAPI Deep Neural Network Library (oneDNN) optimizations are enabled by default.
- With Conda: conda install tensorflow -c intel

https://www.intel.com/content/www/us/en/developer/tools/oneapi/ai-analytics-toolkit-download.html?operatingsystem=linux&distributions=yoctoproject
https://www.intel.com/content/www/us/en/developer/articles/guide/optimization-for-tensorflow-installation-guide.html

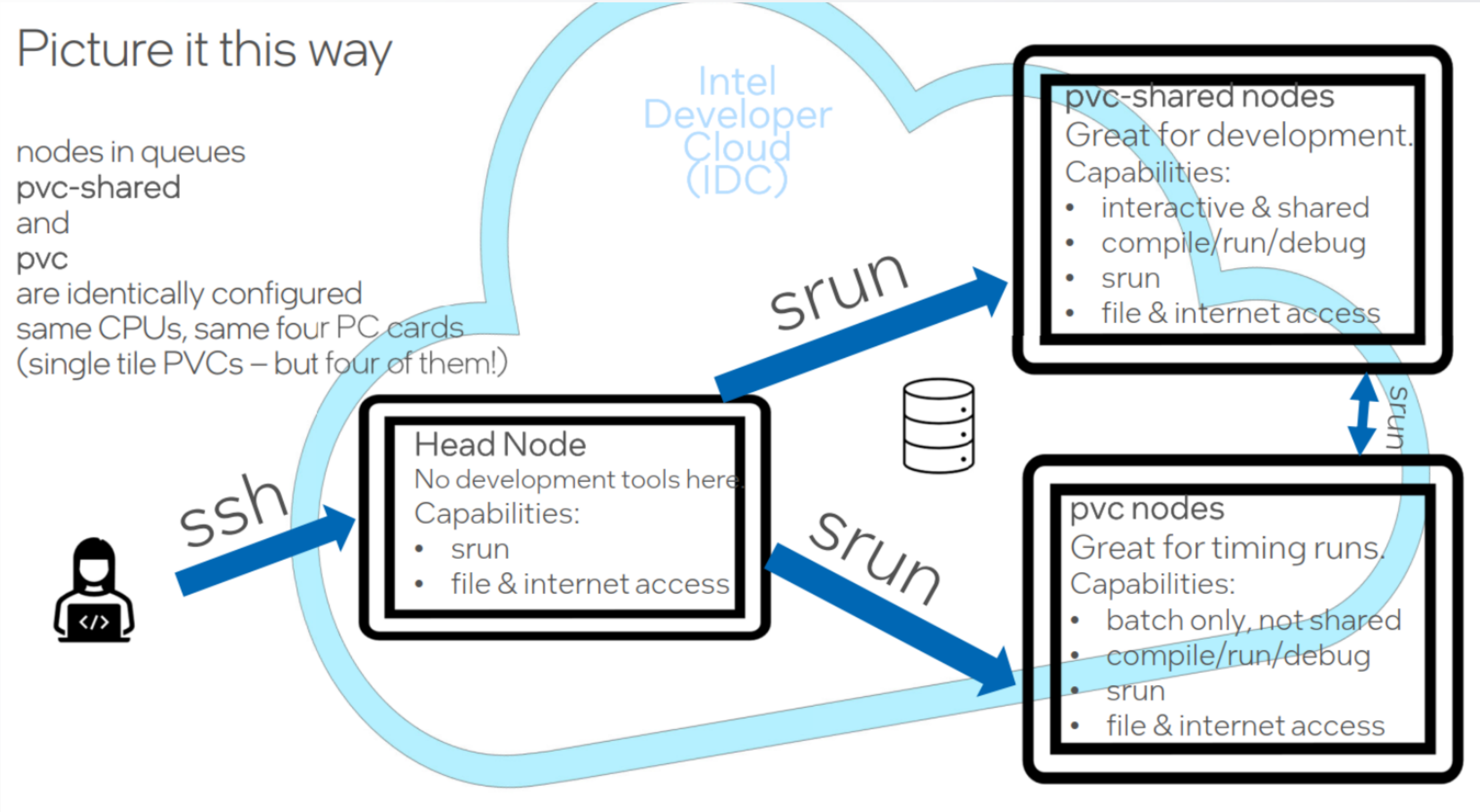
Handson Workshop

Github Repo

https://tinyurl.com/oneapi-ai-workshop

IDC Access Architecture

- SLURM based
- Access valid for 20 days
- Expires if unused for last 7 days
- 20 GB NFS storage



Picture it this way

nodes in queues
pvc-shared
and
pvc
are identically configured
same CPUs, same four PC cards
(single tile PVCs – but four of them!)

Head Node
No development tools here.
Capabilities:
- srun
- file & internet access

pvc-shared nodes
Great for development.
Capabilities:
- interactive & shared
- compile/run/debug
- srun
- file & internet access

pvc nodes
Great for timing runs.
Capabilities:
- batch only, not shared
- compile/run/debug
- srun
- file & internet access

https://github.com/bjodom/idc

Intel Data Center GPU Max Series Products & Form Factor

	Max 1550 GPU (600W OAM)	Max 1350 GPU (450W OAM)	Max 1100 GPU (300W PCIe)
Architecture	X^e HPC		
X^e Cores	128	112	56
Memory	HBM2E 128 GB	HBM2E 96 GB	HBM2E 48 GB
Cache	L1 64 MB L2 408 MB	L1 48 MB L2 216 MB	L1 28 MB L2 108 MB
Max TDP	600W	450W	300W
Form Factor	OAM		PCIe AIC
Host Interconnect	PCIe Gen5		
Physical Ports	X^e Link 53 GB/s 16 ports		X^e Link 53 GB/s 6 ports

One Generation → 30x AI Performance Gain

oneAPI

TensorFlow

Baseline

13.06 images/s

(FP32)
Official TensorFlow on 3rd Gen Intel® Xeon® Scalable Processor

oneAPI one DNN

1.5x

20.54 images/s

(FP32)
TensorFlow with oneDNN enabled

Intel® Neural Compressor

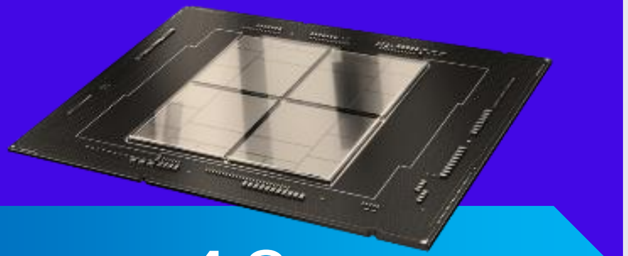
3.9x

81.66 images/s

(INT8)
Model quantization with Intel® Neural compressor

3rd Gen Intel® Xeon™ Scalable Processors

4th Gen Intel® Xeon® Scalable Processor



4.8x

394 images/s

(INT8)
Intel® AMX optimization on Sapphire Rapids

4th Gen Intel® Xeon® Scalable processors

Results may vary. See www.intel.com/InnovationEventClaims for workloads and configurations.

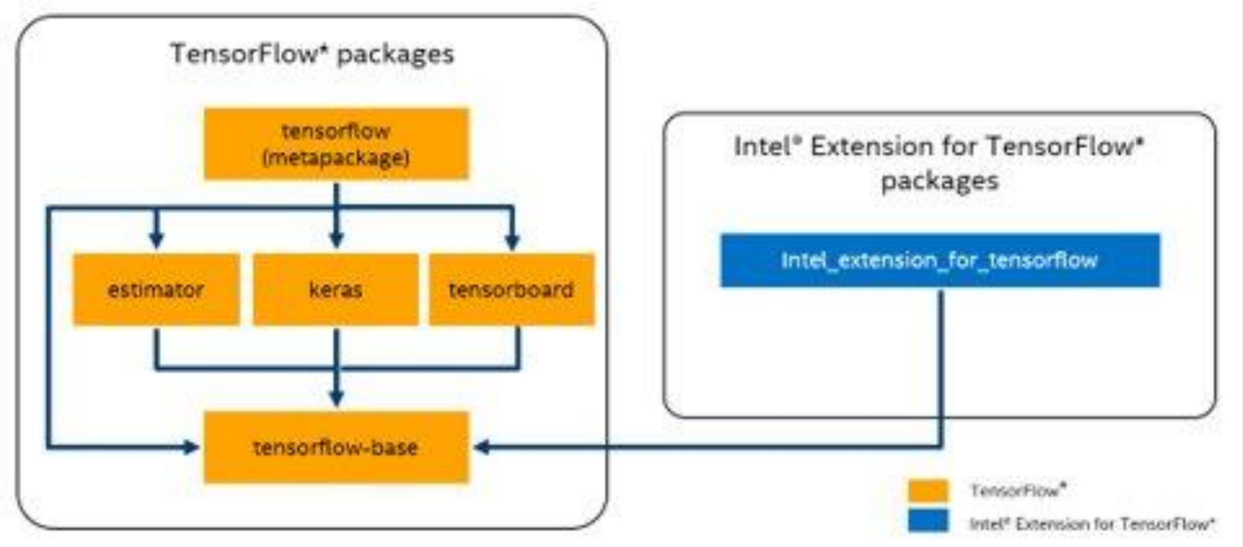
SSD-ResNet-34 Inference Throughput (Batch Size =1)
For workloads and configurations visit www.intel.com/InnovationEventClaims. Results may vary.

Slurm Commands

- sinfo – Lists available partitions and node allocations
- squeue – lists the queued jobs
- srun – Sends a job to the queue for execution
- scancel – Deletes a queued job

Intel® Extension for TensorFlow*

- Intel® Extension for TensorFlow* is a heterogeneous, high performance deep learning extension plugin based on TensorFlow PluggableDevice interface to bring Intel XPU(GPU, CPU, etc) devices into TensorFlow .
- Good performance using default ITEX setting with no code change
- More performance optimizations with minor code change using simple frontend Python API
- GitHub: https://github.com/intel/intel-extension-for-tensorflow


TensorFlow* packages
tensorflow (metapackage)
estimator
keras
tensorboard
tensorflow-base
Intel® Extension for TensorFlow* packages
Intel_extension_for_tensorflow
🟧 TensorFlow*
🟦 Intel® Extension for TensorFlow*

Intel® Extension for TensorFlow* PyPI packages and dependencies

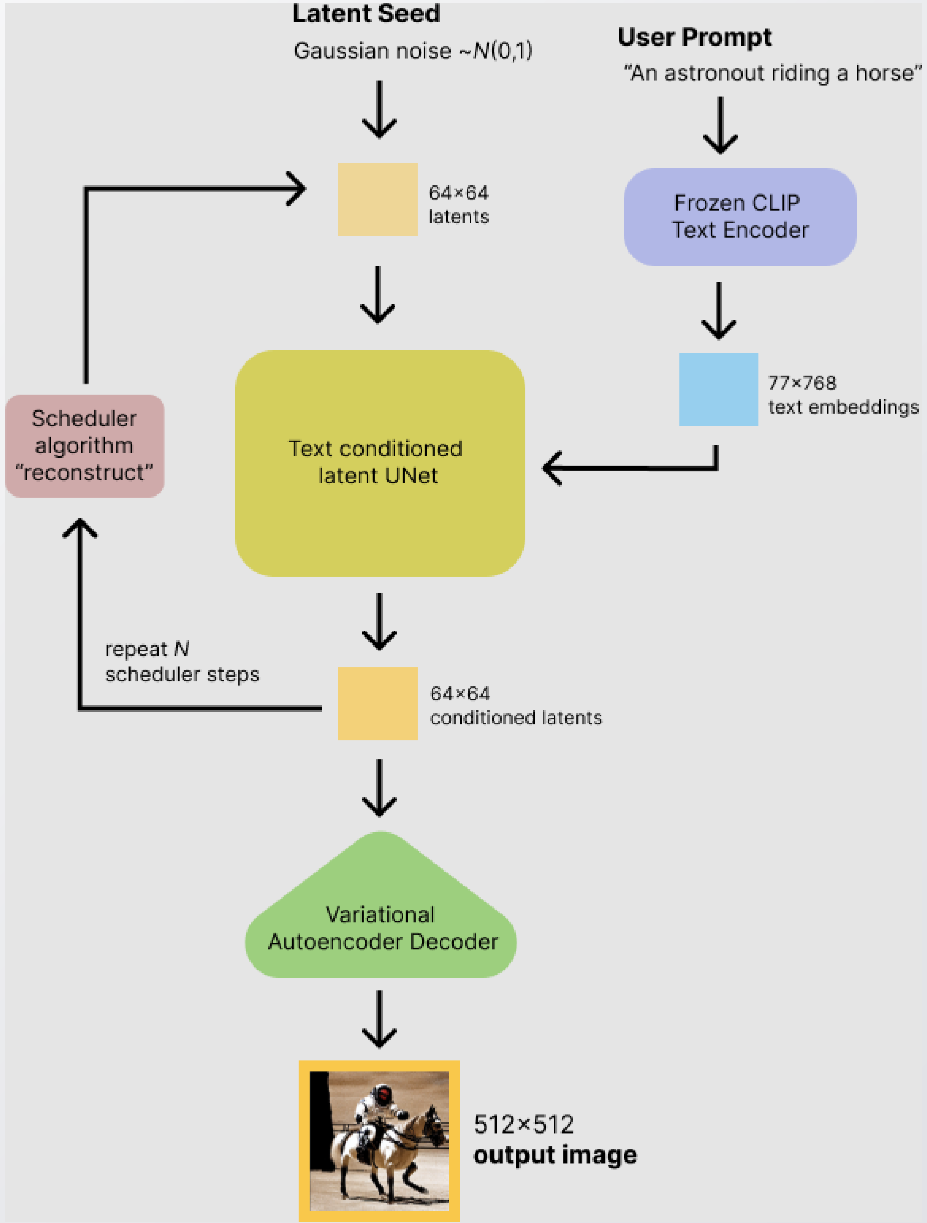
Intel® Extension for TensorFlow* - Features

Features:

- Auto Mixed Precision (AMP)
  - support of AMP with BFloat16 and Float16 operations
- Channels Last
  - support of channels_last (NHWC) memory format
- DPC++ Extension
  - mechanism to create operators with custom DPC++ kernels running on the XPU device
- Optimized Fusion
  - support of SGD/AdamW fusion for both FP32 and BF16 precision
  - a set of fusion patterns for inference

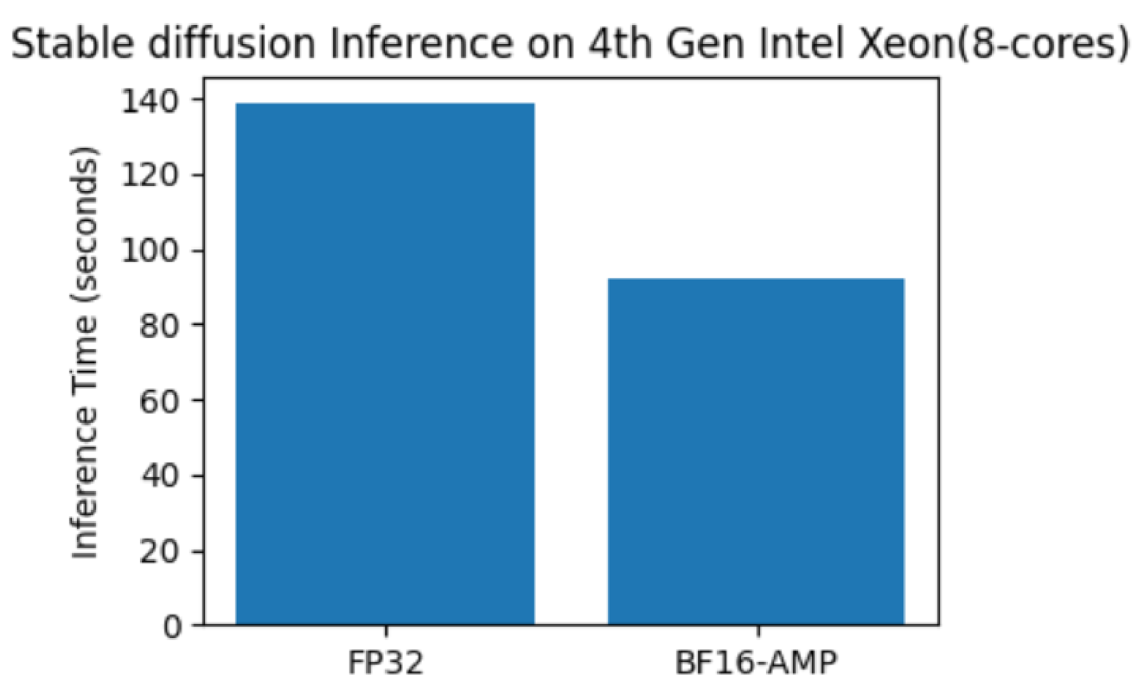
Stable Diffusion

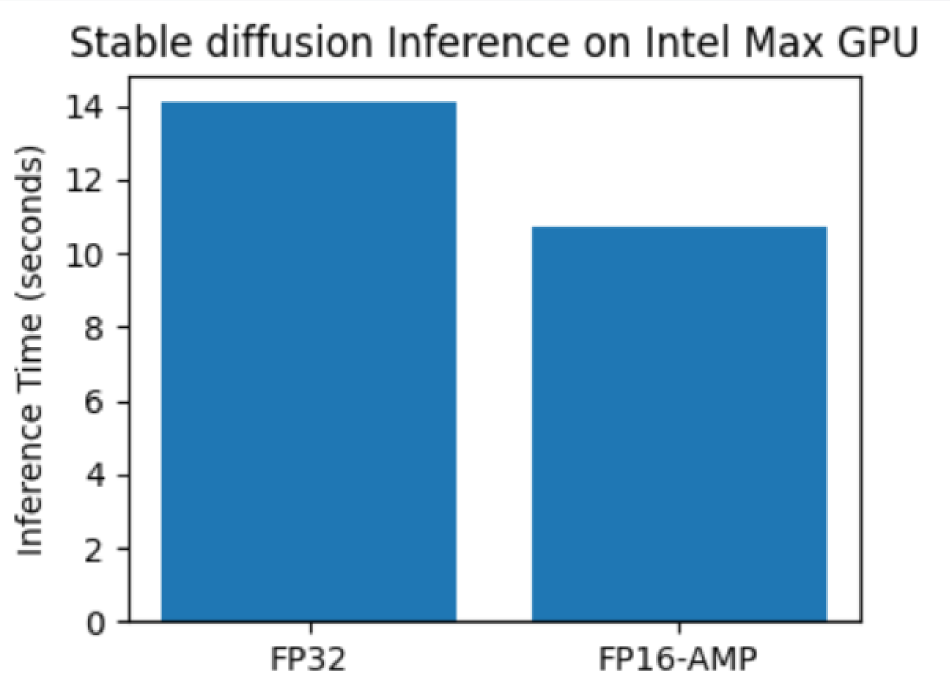
- Latent(space) Diffusion Models
  - like DALL-E, Midjourney etc.
- Various tasks text2image, inpainiting, image2image etc.
- 3 main components:
  - Text Encoder – CLIPText
  - Diffusion Model – Unet
  - Image Decoder – VAE
- ~ 1B parameters


Latent Seed
Gaussian noise ~N(0,1)
User Prompt
"An astronout riding a horse"
64×64 latents
Frozen CLIP Text Encoder
77×768 text embeddings
Scheduler algorithm "reconstruct"
Text conditioned latent UNet
repeat N scheduler steps
64×64 conditioned latents
Variational Autoencoder Decoder
512×512
output image

https://huggingface.co/blog/stable_diffusion

Performance Estimates SPR vs PVC





Thank you!!

oneAPI

FAQ!!