November 2022

AI at Intel

Adel Chaibi – AI TSS

intel®

If you've heard of it, Intel has likely optimized for it

ECOSYSTEM & PARTNERS

Data | Model | Deploy

MODIN | NumPy | SciPy | XGBoost | scikit learn | TensorFlow | PyTorch | OpenVINO

oneAPI

Open, Standards-Based Programming Model & AI Libraries

intel ATOM | intel CORE | intel XEON | Xe | intel habana

Intel® Xeon® Scalable Processors

The Only Data Center CPU with Built-in AI Acceleration

Intel Advanced Vector Extensions 512

Intel Deep Learning Boost (Intel DL Boost)

Intel Optane Persistent Memory

Shipping

Cascade Lake

New Intel DL Boost (VNNI)
New memory storage hierarchy

Cooper Lake

Intel DL Boost (BFLOAT16)

April 2021

Ice Lake

Intel DL Boost (VNNI) and new Intel Software Guard Extensions (Intel® SGX) that enable new AI use cases like federated learning

2022

Sapphire Rapids

Intel Advanced Matrix Extensions (AMX) extends built-in AI acceleration capabilities on Xeon Scalable

Leadership performance

The AI pipeline runs on Xeon

Data	Model	Deploy
Majority of data science and data prep already happens on Intel CPUs	**1.3x** higher ML performance than NV A100 across 20 key customer workloads (geomean)[1]	**70%** of data center inferencing runs on Xeon[2]

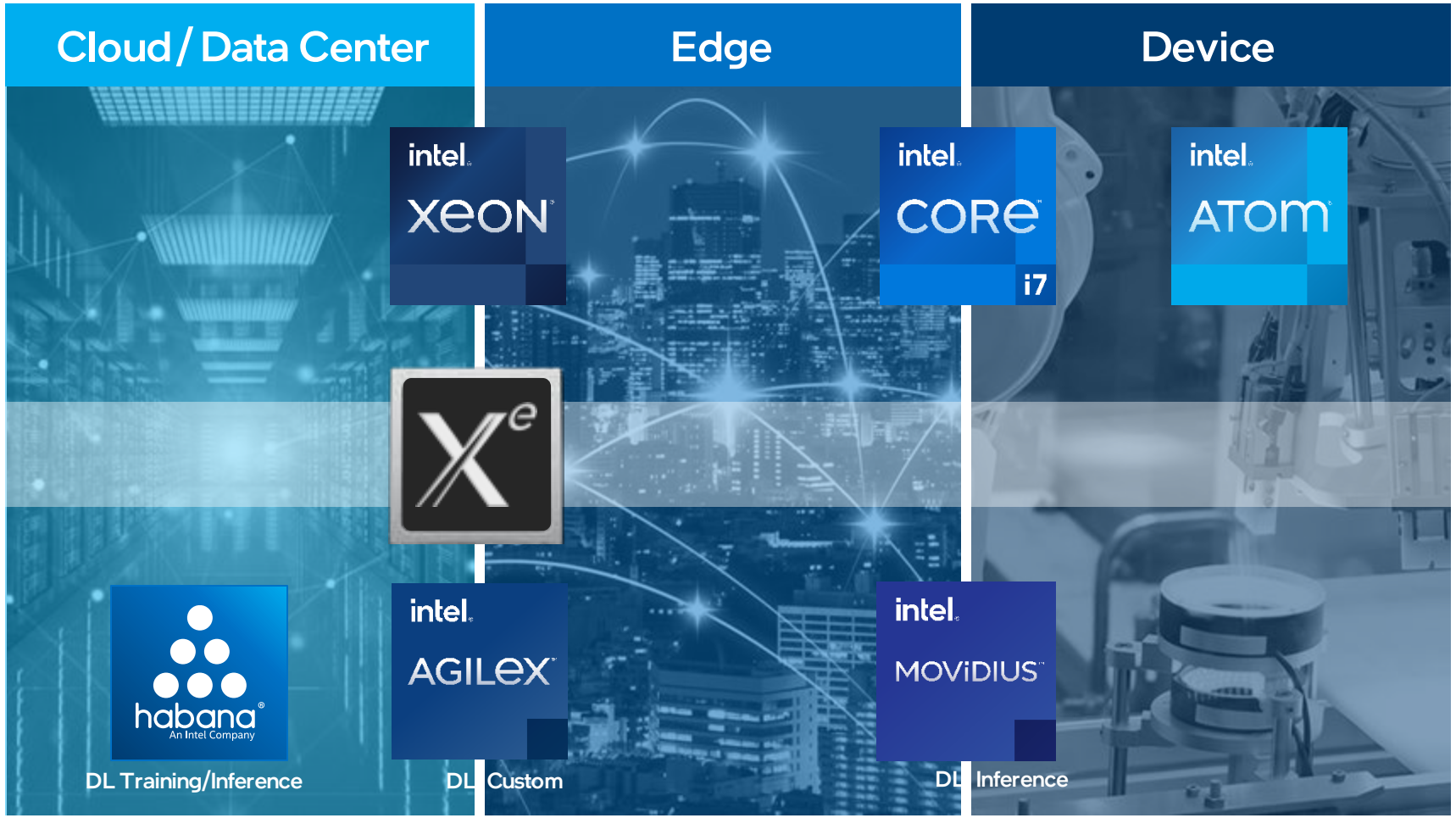
1 See claim 44 at www.intel.com/3gen-xeon-config for workloads and configurations. Results may vary.
2 Based on Intel market modeling of the worldwide installed base of data center servers running AI Inference workloads as of December 2021.

Flexible AI Acceleration

CPU *only*
Built-in AI acceleration for mainstream AI use cases

CPU + GPU
When compute is dominated by AI, HPC, graphics, and/or real-time media

CPU + custom
When compute is dominated by deep learning (DL)

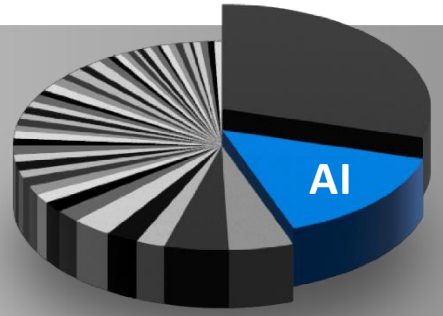


Alternate Version

Intel AI XPU Portfolio

Broad Market

AI is a part of a broader app and/or workload mix

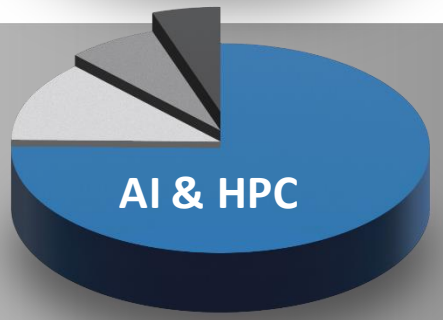


CPU only

intel. XEON | intel. CORE i7

AI/HPC/Graphics

High percentage of AI, HPC, real-time media and/or graphics

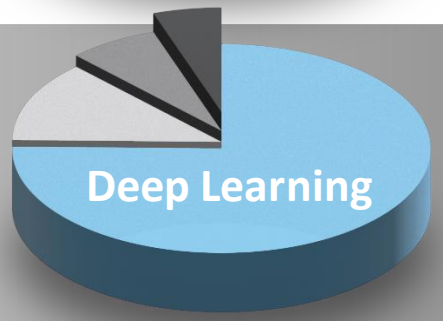


CPU + GPU

Xe

Dedicated Deep Learning

High percentage of DL training and/or inference



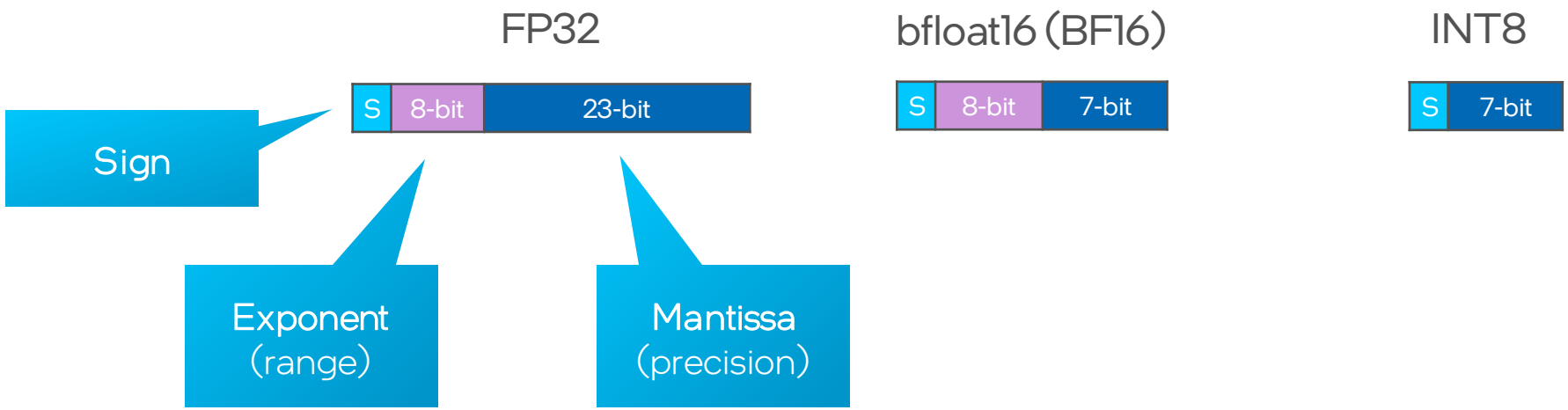
CPU + Dedicated AI ASIC

intel. habana

Intel® DL Boost on Xeon

Additional content for base presentation

Refresher on Data Types



Refresher on Data Types

Supported in 3rd Gen (Cooper Lake) and 4th Gen Xeon (Sapphire Rapids)

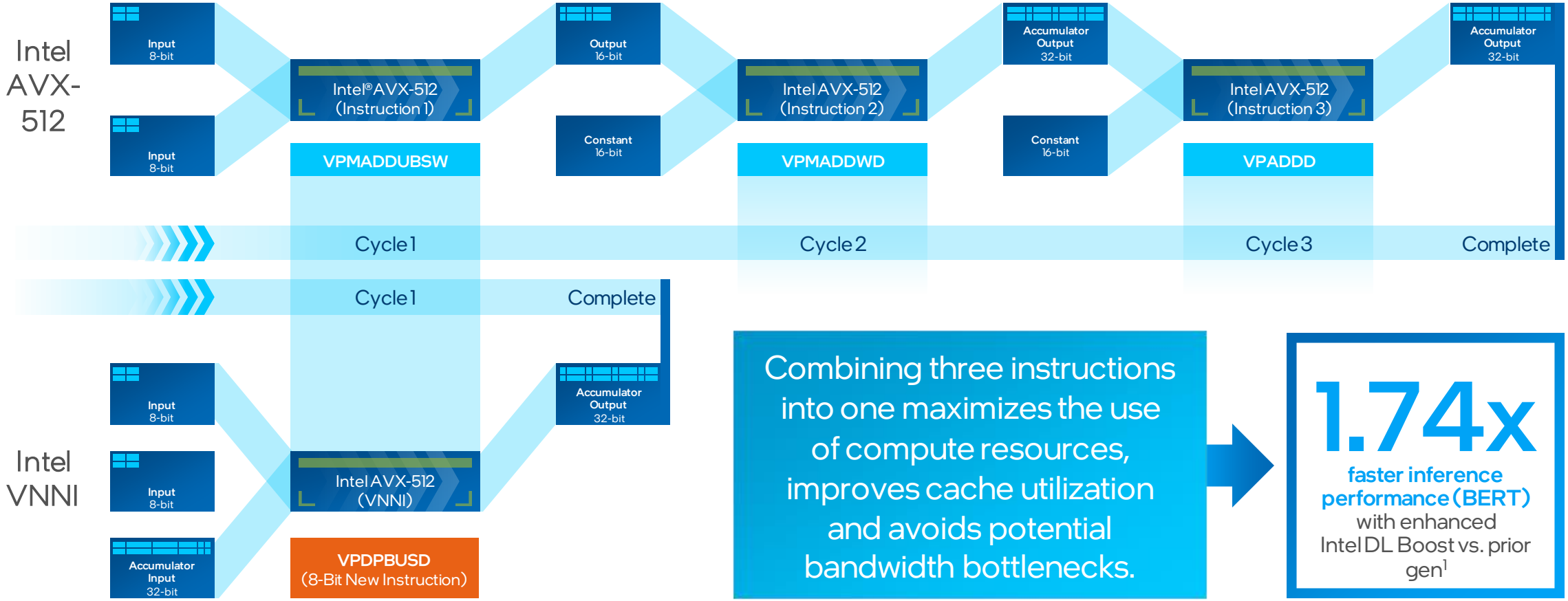
	FP32	bfloat16 (BF16)	INT8
	S 8-bit 23-bit	S 8-bit 7-bit	S 7-bit
Precision/Accuracy	Higher		Lower
AI Usage	Training		Inference

Good enough accuracy for most TRAINING

Higher accuracy for INFERENCING if needed

Intel Deep Learning Boost

A Vector Neural Network Instruction (VNNI) Extends Intel AVX-512 to Accelerate AI/DL Inference



Combining three instructions into one maximizes the use of compute resources, improves cache utilization and avoids potential bandwidth bottlenecks.

1.74x faster inference performance (BERT) with enhanced Intel DL Boost vs. prior gen[1]

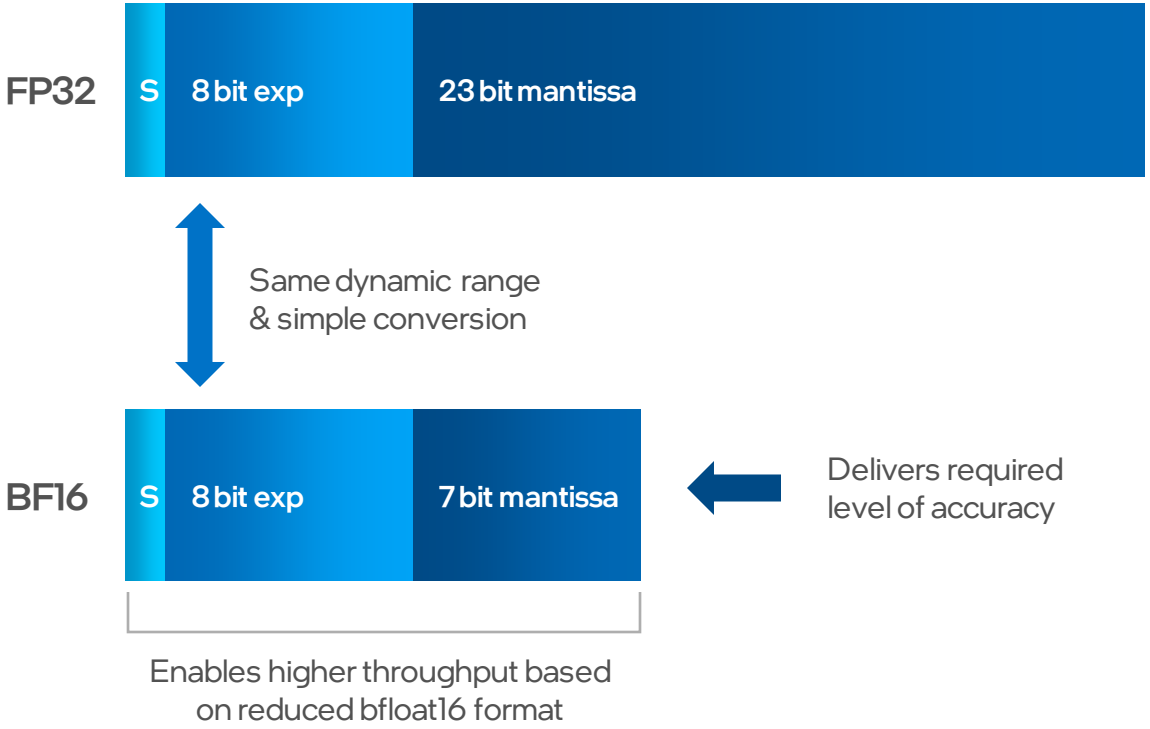
1. See [123] at www.intel.com/3gen-xeon-config. Results may vary.

Intel Deep Learning Boost

Increase Training and Inference Throughput Using Bfloat16

Available on 3rd Gen Intel Xeon Scalable Processors on 4S Cedar Island

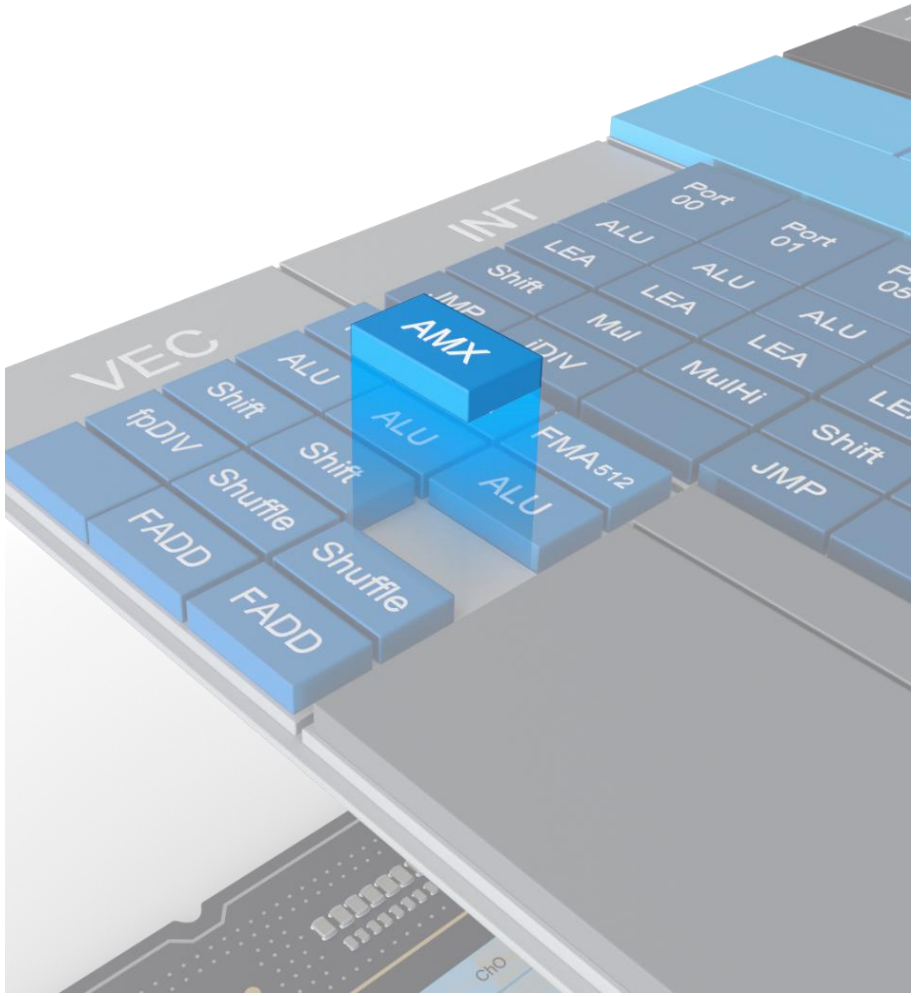
- Training & inference acceleration
- Native support for bfloat16 datatype
- 2x bfloat16 peak throughput/cycle vs. fp32
- Improved throughput and efficiencies
- Seamless integration with popular AI frameworks



New built-in AI-acceleration capabilities in select 3rd Generation Intel Xeon Scalable processors target higher training and inference performance with the required level of accuracy

What is Intel® Advanced Matrix Extensions (AMX)?

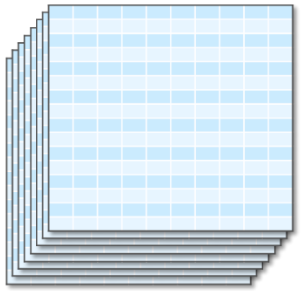
Dedicated silicon on each Sapphire Rapids core that helps accelerate **deep learning** workloads



What is Advanced Matrix Extensions (AMX)?

"Tiles"

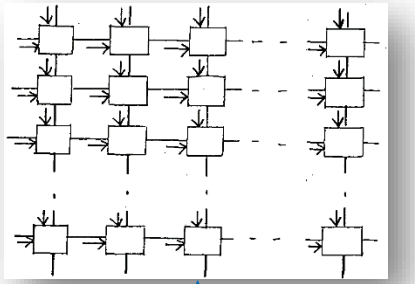
2D Register Files



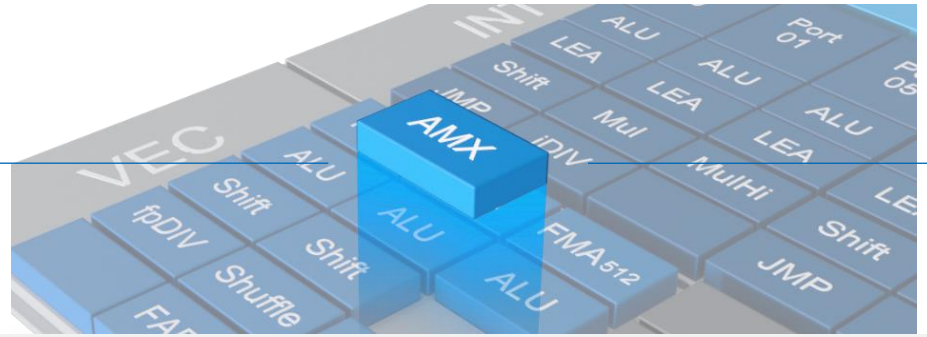
+

"TMUL"

Tile Matrix Multiply



Store bigger chunks of **DATA** in each core



INSTRUCTIONS that compute larger matrices in a single operation

Intel® AMX supports BF16 and INT8 data types

	3rd Gen Intel® Xeon® Scalable processor		4th Gen Intel® Xeon® Scalable processor
	Cooper Lake	Ice Lake	Sapphire Rapids
AVX512	FP64, FP32, bfloat16	FP64, FP32	FP64, FP32
VNNI	INT8	INT8	INT8
AMX			**bfloat16, INT8**

FP32 supported through AVX512 instructions

Intel® AMX supports BF16 and INT8

Intel® AMX accelerates DEEP LEARNING use cases

Intel® Advanced Matrix Extensions (AMX)

BF16 and INT8 data types

Recommender Systems

Natural Language Processing

Image Recognition
Object Detection

Intel® Advanced Vector Extensions (AVX-512)

FP32 and FP64 data types

Data Analytics

Classical Machine Learning

FUN FACT:

Many DL workloads are "mixed precision"
and
4th Gen Xeon can seamlessly transition between AMX and AVX-512 as needed

AI Software

oneAPI

Additional content for base presentation

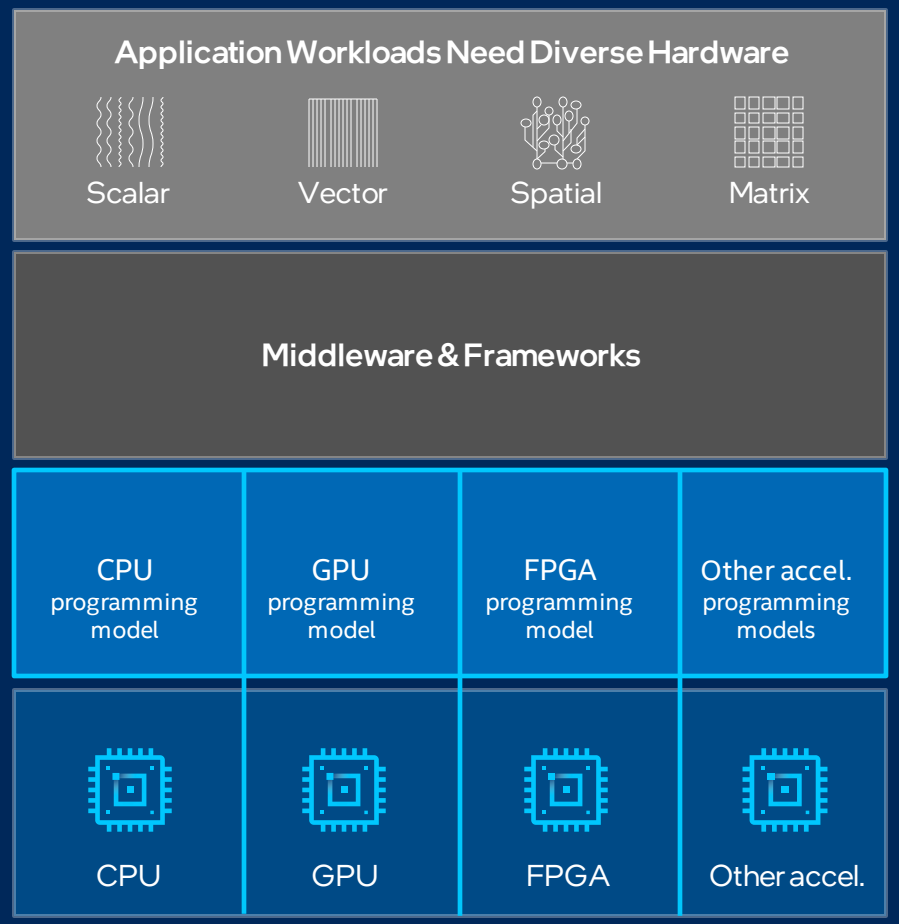
Programming Challenges
for Multiple Architectures

Growth in specialized workloads

Variety of data-centric hardware required

Separate programming models and toolchains for each architecture are required today

Software development complexity limits freedom of architectural choice



oneAPI

One Programming Model for Multiple Architectures and Vendors

Freedom to Make Your Best Choice

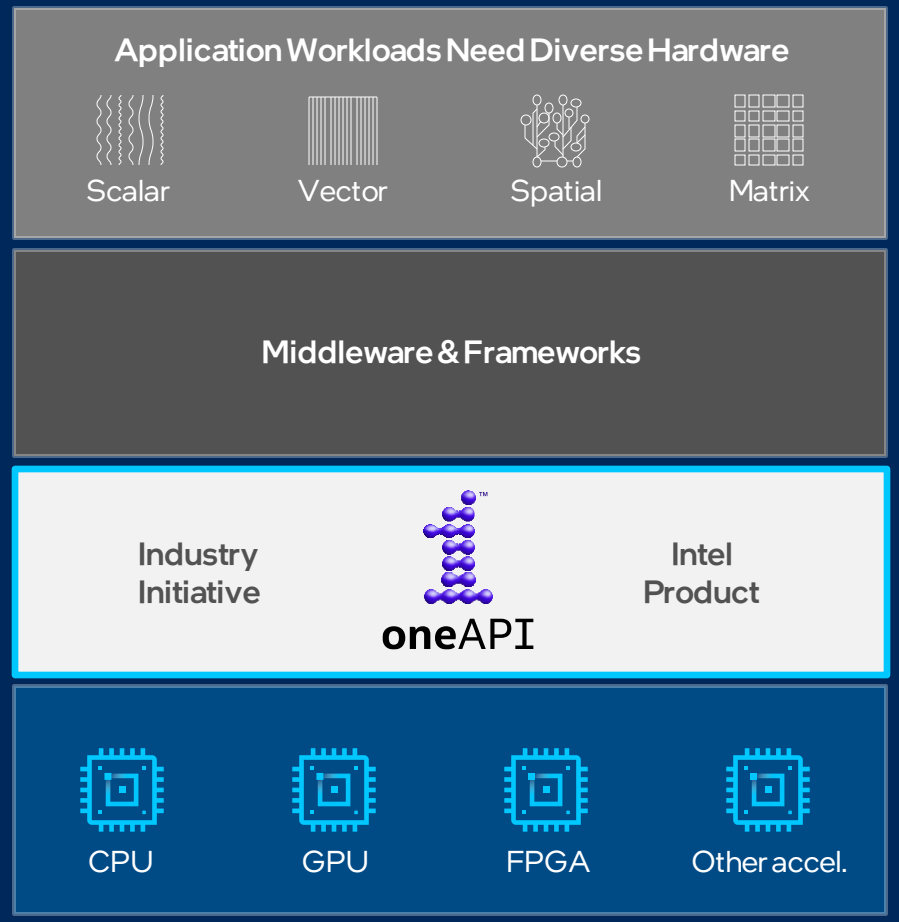
- Choose the best accelerated technology the software doesn't decide for you

Realize all the Hardware Value

- Performance across CPU, GPUs, FPGAs, and other accelerators

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++, Python, SYCL, OpenMP, Fortran, and MPI



oneAPI Industry Initiative

Break the Chains of Proprietary Lock-in

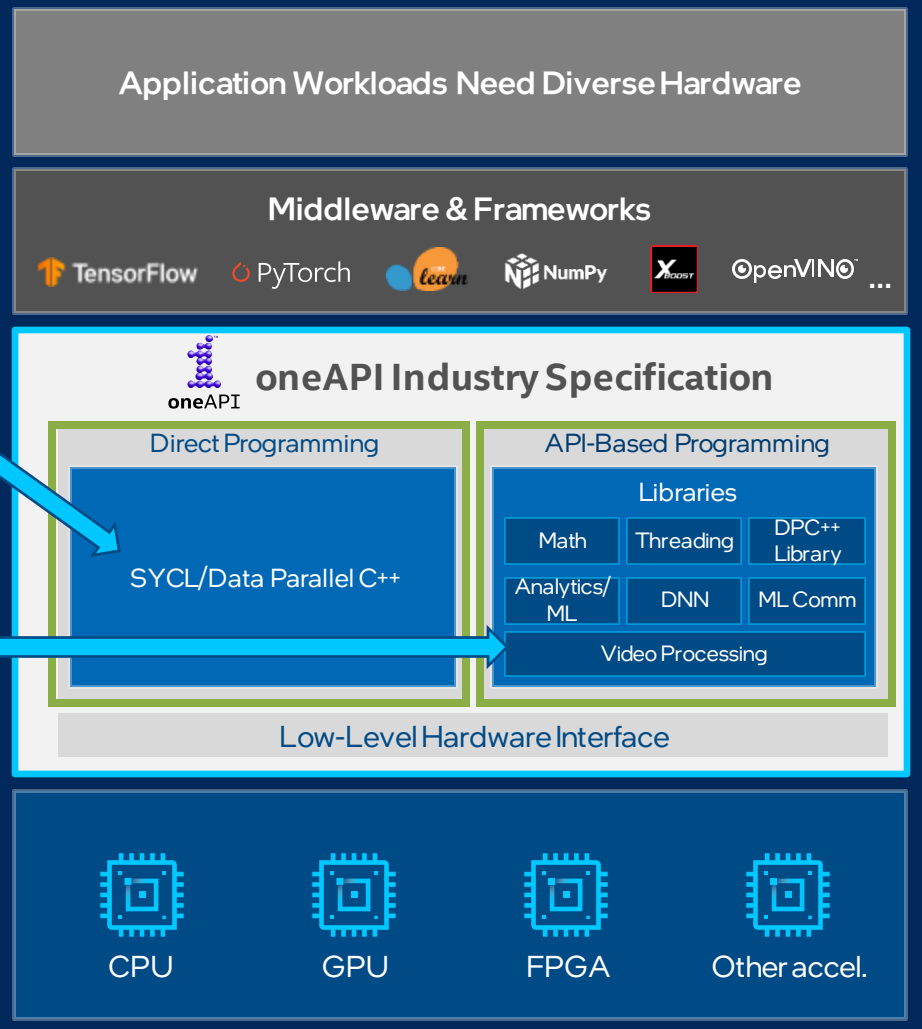
Open to promote community and industry collaboration

Enables code reuse across architectures and vendors

A cross-architecture language based on C++ and SYCL standards

Powerful libraries designed for acceleration of domain-specific functions

Low-level hardware abstraction layer



oneAPI

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

Intel® oneAPI Toolkits

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

oneAPI

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

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

Latest version available 2022.2

oneAPI AI Analytics Toolkit

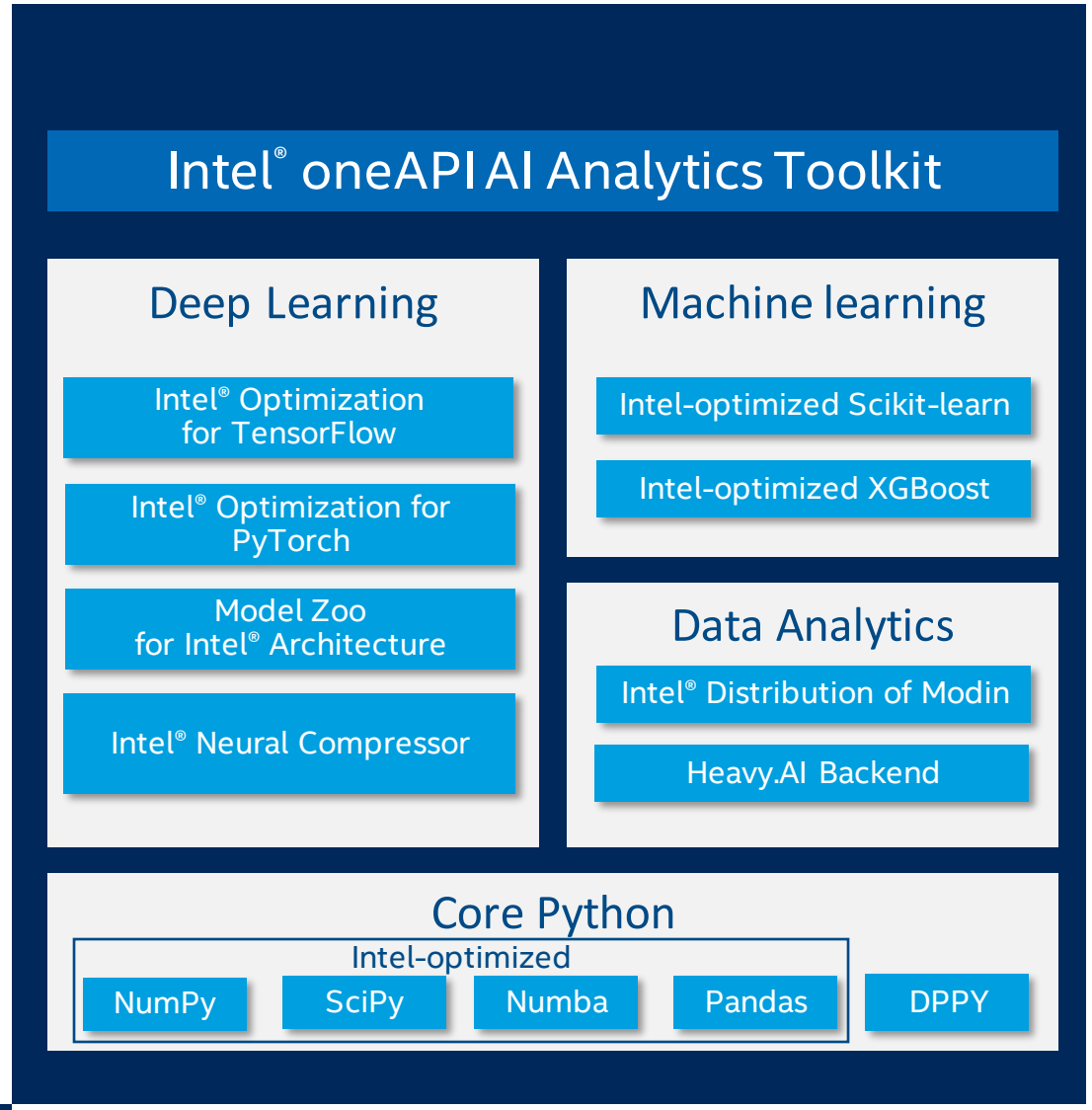
Intel® oneAPI AI Analytics Toolkit

Accelerates end-to-end Machine Learning and Data Analytics pipelines with frameworks and libraries optimized for Intel® architectures

Who Uses It?

Data scientists, AI Researchers, Machine and Deep Learning developers, AI application developers

Learn More: intel.com/oneAPI-AIKit



Engineer Data → Create Machine Learning & Deep Learning Models → Deploy

AI Platforms & Kits

Most Popular Tools and Frameworks

Performance Libraries

GP Compute	Vector Accl	Matrix Accl	Memory
#Cores, #Frequency	Intel AVX2, AVX-512, VNNI	Intel AMX	Cache, DDR5, HBM, Intel® Optane™, Frequency

AVX – Advanced Vector Extensions, VNNI – Vector Neural Network Instructions, AMX – Advanced Matrix Extensions

Engineer Data → Create Machine Learning & Deep Learning Models → Deploy

AI Platforms & Kits

Most Popular Tools and Frameworks

oneGNN oneAPI | SYCLomatic | oneDAL | oneDNN | oneCCL | oneMKL

GP Compute	Vector Accl	Matrix Accl	Memory
#Cores, #Frequency	Intel AVX2, AVX-512, VNNI	Intel AMX	Cache, DDR5, HBM, Intel Optane, Frequency

oneDAL – Intel oneAPI Data Analytics Library, oneDNN – Intel oneAPI Deep Neural Networks Library, oneCCL – Intel oneAPI Collective Communications Library, oneMKL- Intel oneAPI Math Kernel Library
AVX – Advanced Vector Extensions, VNNI – Vector Neural Network Instructions, AMX – Advanced Matrix Extensions

Engineer Data → Create Machine Learning & Deep Learning Models → Deploy

AI Platforms & Kits

Data Analytics Scale

MODIN, SciPy, pandas, NumPy

Optimized Frameworks and Middleware

TensorFlow, PyTorch, mxnet, PaddlePaddle, scikit learn, ONNX, LightGBM, XGBoost, CatBoost

Optimize Models

Automate Model Tuning AutoML

Automate Low-Precision Optimization

SigOpt

Intel Neural Compressor

Intel® Neural Coder

w/ Intel Optimizations

oneAPI: oneDAL, oneDNN, oneCCL, oneMKL

GP Compute
#Cores, #Frequency

Vector Accl
Intel AVX2, AVX-512, VNNI

Matrix Accl
Intel AMX

Memory
Cache, DDR5, HBM, Intel Optane, Frequency

oneDAL – Intel oneAPI Data Analytics Library, oneDNN – Intel oneAPI Deep Neural Networks Library, oneCCL – Intel oneAPI Collective Communications Library, oneMKL- Intel oneAPI Math Kernel Library
AVX – Advanced Vector Extensions, VNNI – Vector Neural Network Instructions, AMX – Advanced Matrix Extensions

Engineer Data → Create Machine Learning & Deep Learning Models → Deploy

Accelerate End-to-End Data Science and AI — AI Analytics Toolkit

Data Analytics Scale

MODIN, SciPy, pandas, NumPy

Optimized Frameworks and Middleware

TensorFlow, PyTorch, mxnet, PaddlePaddle, scikit learn, ONNX, LightGBM, XGBoost, CatBoost

Optimize Models

Automate Model Tuning AutoML — SigOpt

Automate Low-Precision Optimization — Intel Neural Compressor

w/ Intel Optimizations

oneAPI — oneDAL, oneDNN, oneCCL, oneMKL

GP Compute
#Cores, #Frequency

Vector Accl
Intel AVX2, AVX-512, VNNI

Matrix Accl
Intel AMX

Memory
Cache, DDR5, HBM, Intel Optane, Frequency

oneDAL – Intel oneAPI Data Analytics Library, oneDNN – Intel oneAPI Deep Neural Networks Library, oneCCL – Intel oneAPI Collective Communications Library, oneMKL- Intel oneAPI Math Kernel Library
AVX – Advanced Vector Extensions, VNNI – Vector Neural Network Instructions, AMX – Advanced Matrix Extensions

Engineer Data → Create Machine Learning & Deep Learning Models → Deploy

Connect AI to Big Data — Spark — BigDL (previously "Analytics Zoo")

Accelerate End-to-End Data Science and AI — AI Analytics Toolkit

Data Analytics Scale

MODIN, SciPy, pandas, NumPy

Optimized Frameworks and Middleware

TensorFlow, PyTorch, mxnet, PaddlePaddle, scikit learn, ONNX, LightGBM, XGBoost, CatBoost

Optimize Models

Automate Model Tuning AutoML — SigOpt

Automate Low-Precision Optimization — Intel Neural Compressor

w/ Intel Optimizations

oneAPI — oneDAL — oneDNN — oneCCL — oneMKL

GP Compute
#Cores, #Frequency

Vector Accl
Intel AVX2, AVX-512, VNNI

Matrix Accl
Intel AMX

Memory
Cache, DDR5, HBM, Intel Optane, Frequency

oneDAL – Intel oneAPI Data Analytics Library, oneDNN – Intel oneAPI Deep Neural Networks Library, oneCCL – Intel oneAPI Collective Communications Library, oneMKL- Intel oneAPI Math Kernel Library
AVX – Advanced Vector Extensions, VNNI – Vector Neural Network Instructions, AMX – Advanced Matrix Extensions

Engineer Data → Create Machine Learning & Deep Learning Models → Deploy

Connect AI to Big Data — Spark — BigDL (previously "Analytics Zoo")

Accelerate End-to-End Data Science and AI — AI Analytics Toolkit

Data Analytics Scale

MODIN, SciPy, pandas, NumPy

Optimized Frameworks and Middleware

TensorFlow, PyTorch, mxnet, PaddlePaddle, scikit learn, ONNX, LightGBM, XGBoost, CatBoost

Optimize Models

Automate Model Tuning AutoML

Automate Low-Precision Optimization

SigOpt

Intel Neural Compressor

OpenVINO™ Toolkit

Write Once

Deploy Auto-Optimized

Anywhere

w/ Intel Optimizations

oneAPI — oneDAL — oneDNN — oneCCL — oneMKL

GP Compute
#Cores, #Frequency

Vector Accl
Intel AVX2, AVX-512, VNNI

Matrix Accl
Intel AMX

Memory
Cache, DDR5, HBM, Intel Optane, Frequency

oneDAL – Intel oneAPI Data Analytics Library, oneDNN – Intel oneAPI Deep Neural Networks Library, oneCCL – Intel oneAPI Collective Communications Library, oneMKL- Intel oneAPI Math Kernel Library
AVX – Advanced Vector Extensions, VNNI – Vector Neural Network Instructions, AMX – Advanced Matrix Extensions

Engineer Data → Create Machine Learning & Deep Learning Models → Deploy

Container Repository	MLOps	Developer Sandbox	Annotation/Training/Optimization
oneContainer	**Cnvrg.io**	**Intel® Developer Cloud**	**Intel® GETi**

Connect AI to Big Data — Spark — BigDL (previously "Analytics Zoo")

Accelerate End-to-End Data Science and AI — AI Analytics Toolkit

Data Analytics Scale

MODIN, SciPy, pandas, NumPy

Optimized Frameworks and Middleware

TensorFlow, PyTorch, mxnet, PaddlePaddle, scikit learn, ONNX, LightGBM, XGBoost, CatBoost

Optimize Models

Automate Model Tuning AutoML — Automate Low-Precision Optimization

SigOpt — **Intel Neural Compressor**

OpenVINO Toolkit

Write Once

Deploy Auto-Optimized

Anywhere

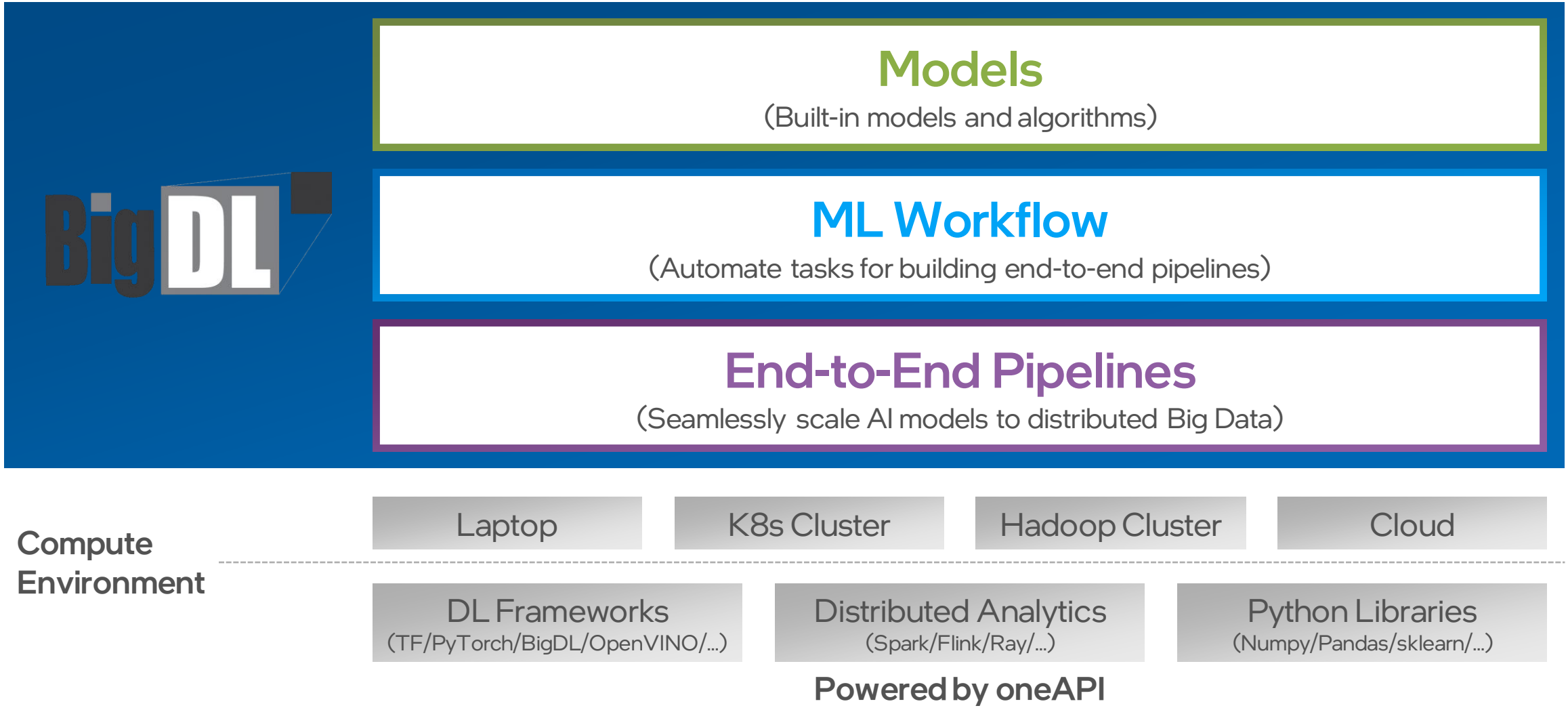
w/ Intel Optimizations

oneAPI — oneDAL — oneDNN — oneCCL — oneMKL

GP Compute	**Vector Accl**	**Matrix Accl**	**Memory**
#Cores, #Frequency	Intel AVX2, AVX-512, VNNI	Intel AMX	Cache, DDR5, HBM, Intel Optane, Frequency

oneDAL – Intel oneAPI Data Analytics Library, oneDNN – Intel oneAPI Deep Neural Networks Library, oneCCL – Intel oneAPI Collective Communications Library, oneMKL- Intel oneAPI Math Kernel Library
AVX – Advanced Vector Extensions, VNNI – Vector Neural Network Instructions, AMX – Advanced Matrix Extensions

BigDL: Software Platform for Big Data AI



https://github.com/intel-analytics/analytics-zoo

BigDL: Software Platform for Big Data AI

Models & Algorithms

- Recommendation
- Time Series
- Computer Vision
- NLP

ML Workflow

- AutoML
- Automatic Cluster Serving

End-to-end Pipelines

- Distributed TensorFlow & PyTorch on Spark
- RayOnSpark
- Spark Dataframes & ML Pipelines for DL
- InferenceModel

Compute Environment

- Laptop
- K8s Cluster
- Hadoop Cluster
- Cloud

- DL Frameworks (TF/PyTorch/BigDL/OpenVINO/...)
- Distributed Analytics (Spark/Flink/Ray/...)
- Python Libraries (Numpy/Pandas/sklearn/...)

Powered by oneAPI

https://github.com/intel-analytics/analytics-zoo

Intel Has the Developer Tools Companies Use to Scale AI Everywhere

Upstream

Integrated acceleration to popular open source software

Modin, XGBoost, TF, PT, PDPD, MxNet, more …

Intel Extension

Easily pluggable extensions to open source software

Scikit-Learn Extension, Optimized Analytics Package, IPEX, more …

Intel Distro

Intel Optimized Distributions of open source software

Modin, Intel TF, IDP

Intel Tools

Tools / Kits which improve productivity and perf on Intel HW

AIKIT, OpenVINO™, BigDL, oneContainer Portal, Cnvrg.io, Intel Neural Compressor, SigOpt,

Across major software channels (PyPI, Anaconda, Intel, Apt, Yum, Docker)

Intel® Xeon® Scalable Processor AI Software Approach Map

Categories	SW Product	Open Source	Optimizations Upstreamed	Intel Extension	Intel Distro	Intel Tool / Kit
Orchestration	Cnvrg.io	No				■
Packaged Software Toolkits	AI Kit	Yes				■
	BigDL	Yes				■
	OpenVINO	Yes			■	■
Optimize	Neural Compressor	Yes				■
	SigOpt	No				■
DL Frameworks	TensorFlow	Yes	■		■	
	PyTorch	Yes	■	■		
	ONNX	Yes	■			
	PDPD	Yes	■			
	MxNet	Yes	■			
ML Frameworks	XGBoost	Yes	■			
	Scikit-Learn	Yes		■		
	CatBoost	Yes	■			
	LightGBM	Yes	■			
Data Prep	Modin	Yes	■		■	
	Spark	Yes	■	■		

Animated
[additional information under blue boxes]

Intel® Xeon® Scalable Processor Software Value Approach Map

Categories	SW Product	Optimizations Upstreamed	Intel Extension	Intel Distro
DL Frameworks	TensorFlow	Differentiation upstreamed to Open Source	+ Differentiation in Intel package	
	PyTorch			
ML Frameworks	Scikit-Learn			
Data Prep	Spark / OAP			
	Modin			

See link below for workloads and configurations. Results may vary
https://techdecoded.intel.io/resources/one-line-code-changes-to-boost-pandas-scikit-learn-and-tensorflow-performance/#gs.bzkn2n

oneAPI Available on
Intel® DevCloud

A development sandbox to develop, test and run workloads across a range of Intel CPUs, GPUs, and FPGAs using Intel's oneAPI software.

Get Up & Running In Seconds!

software.intel.com/devcloud/oneapi

intel. DevCloud

1 Minute to Code

No Hardware Acquisition

No Download, Install or Configuration

Easy Access to Samples & Tutorials

Support for Jupyter Notebooks, Visual Studio Code

intel®