INTRODUCTION TO ARTIFICIAL INTELLIGENCE USING INTEL® HARDWARE PLATFORM

Dr. Fabio Baruffa

Sr. Technical Consulting Engineer, Intel IAGS

NAVIGATING THE AI PERFORMANCE PACKAGE

INTRODUCTION TO AI
Overview of Deep Learning Software

INTEL® XEON® SCALABLE PROCESSORS
First and second generation: Skylake / Cascade Lake

INTEL® DEEP LEARNING BOOST
Intel® AVX-512 Vector Neural Network Instructions (VNNI)

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

NAVIGATING THE AI PERFORMANCE PACKAGE

INTRODUCTION TO AI

Overview of Deep Learning Software

- What are AI, Machine Learning, and Deep Learning?
- Deep Learning Software breakdown
- Popular AI Neural Networks and their uses
- Intel's AI software tools

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

WHAT IS AI?

Regression

Classification

Clustering

Decision Trees

Data Generation

Image Processing

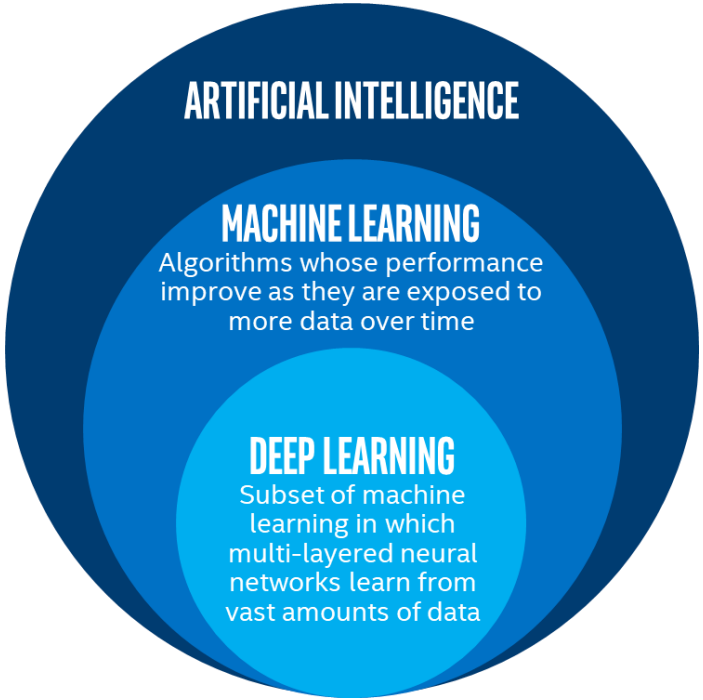
Speech Processing

Natural Language Processing

Recommender Systems

Adversarial Networks

Reinforcement Learning



ARTIFICIAL INTELLIGENCE

is the ability of machines to learn from experience, without explicit programming, in order to perform cognitive functions associated with the human mind

No one size fits all approach to AI

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

UNSUPERVISED LEARNING EXAMPLE

MACHINE LEARNING

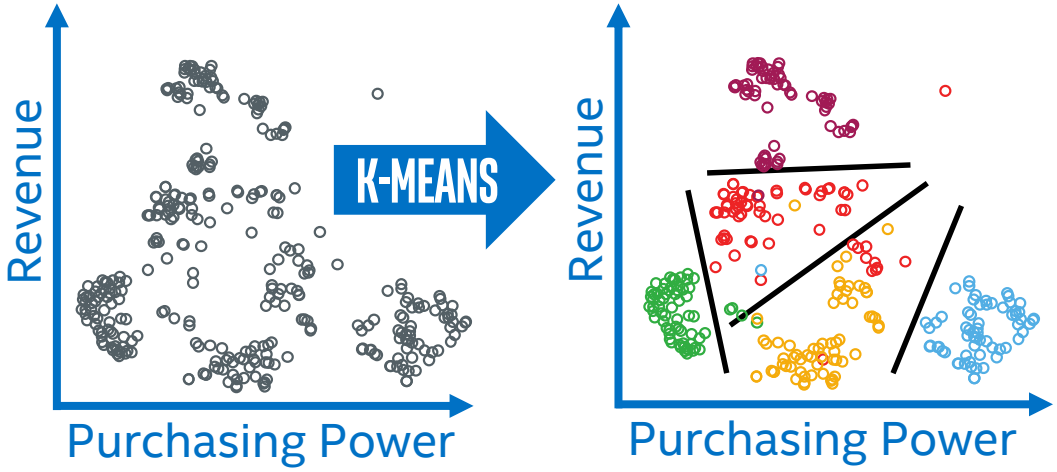
- Regression
- Classification
- **Clustering**
- Decision Trees
- Data Generation

DEEP LEARNING

- Image Processing
- Speech Processing
- Natural Language Processing
- Recommender Systems
- Adversarial Networks
- Reinforcement Learning

MACHINE LEARNING (CLUSTERING)

An 'Unsupervised Learning' Example



Choose the right AI approach for your challenge

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

SUPERVISED LEARNING EXAMPLE

MACHINE LEARNING

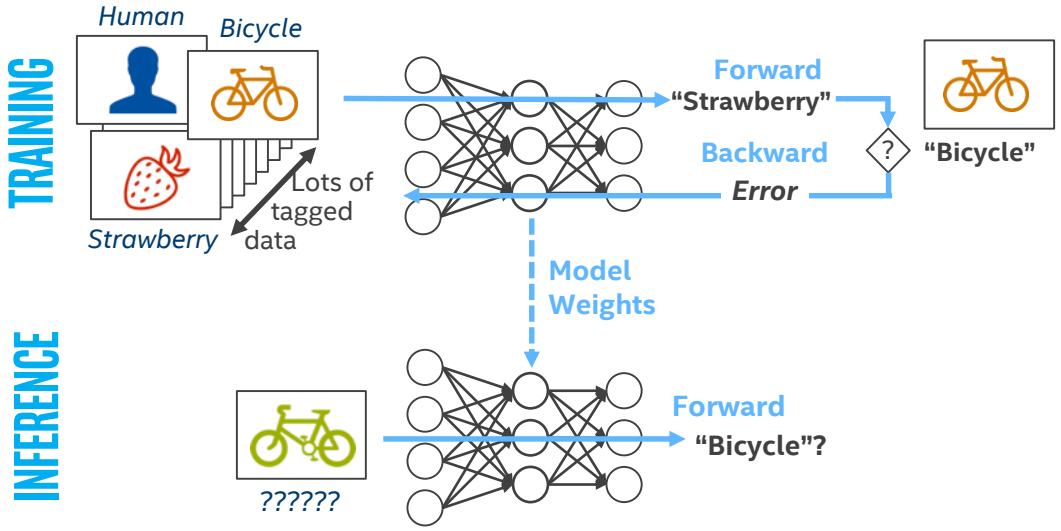
- Regression
- Classification
- Clustering
- Decision Trees
- Data Generation
- Image Processing

DEEP LEARNING

- Speech Processing
- Natural Language Processing
- Recommender Systems
- Adversarial Networks
- Reinforcement Learning

DEEP LEARNING (IMAGE RECOGNITION)

A 'Supervised Learning' Example



Choose the right AI approach for your challenge

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

DEEP LEARNING GLOSSARY

LIBRARY

MKL-DNN
DAAL
Spark MlLib
Scikit-Learn
Intel® Distribution for Python
Mahout
NumPy
Pandas

Hardware-optimized mathematical and other primitive functions that are commonly used in machine & deep learning algorithms, topologies & frameworks

FRAMEWORK

TensorFlow
BigDL for Spark
mxnet
PaddlePaddle
PYTORCH
Caffe

Open-source software environments that facilitate deep learning model development & deployment through built-in components and the ability to customize code

TOPOLOGY

Wide variety of algorithms modeled loosely after the human brain that use neural networks to recognize complex patterns in data that are otherwise difficult to reverse engineer

Translating common deep learning terminology

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

DEEP LEARNING USAGES & KEY TOPOLOGIES

Image Recognition

Resnet-50
Inception V3
MobileNet
SqueezeNet

Object Detection

R-FCN
Faster-RCNN
Yolo V2
SSD-VGG16, SSD-MobileNet

Image Segmentation

Mask R-CNN

Language Translation

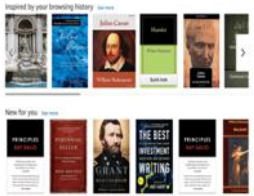
GNMT

Text to Speech

Wavenet

Recommendation System

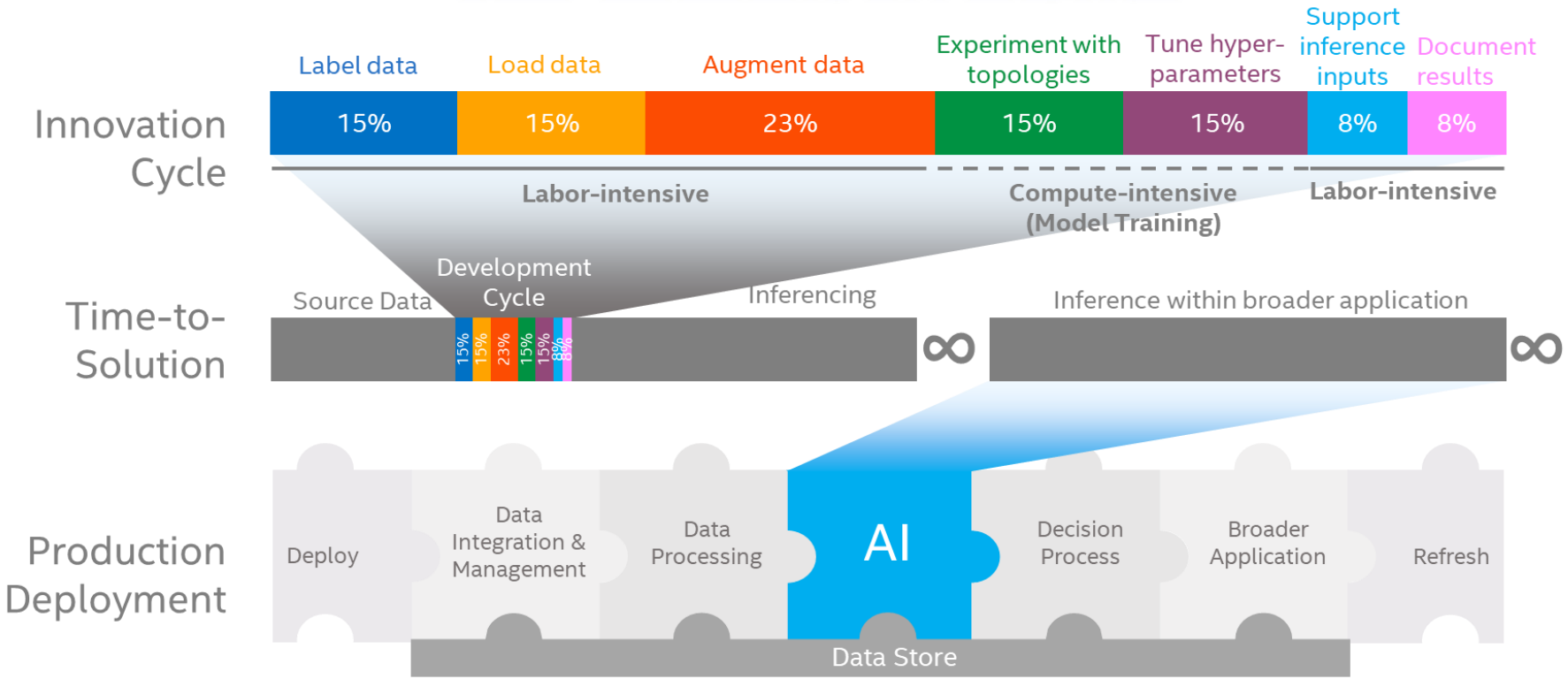
Wide & Deep, NCF



There are many deep learning usages and topologies for each

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

DEEP LEARNING IN PRACTICE



Time-to-solution is more significant than time-to-train

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

Visit:
www.intel.ai/technology

SPEED UP DEVELOPMENT

using open AI software

MACHINE LEARNING ⟷ DEEP LEARNING

TOOLKITS
App developers

ANALYTICS ZOO
Open source platform for building E2E Analytics & AI applications on Apache Spark* with distributed TensorFlow*, Keras*, BigDL

OpenVINO™
Deep learning inference deployment on CPU/GPU/FPGA/VPU for Caffe*, TensorFlow*, MXNet*, ONNX*, Kaldi*

NAUTA
Open source, scalable, and extensible distributed deep learning platform built on Kubernetes (BETA)

LIBRARIES
Data scientists

Python
- Scikit-learn
- Pandas
- NumPy

R
- Cart
- Random Forest
- e1071

Distributed
- MlLib (on Spark)
- Mahout

Intel-optimized Frameworks
TensorFlow*, Caffe2*, ONNX*, mxnet*, PyTorch*, BigDL

And more framework optimizations underway including PaddlePaddle*, Chainer*, CNTK* & others

KERNELS
Library developers

Intel® Distribution for Python*
Intel distribution optimized for machine learning

Intel® Data Analytics Acceleration Library (Intel® DAAL)
High performance machine learning & data analytics library

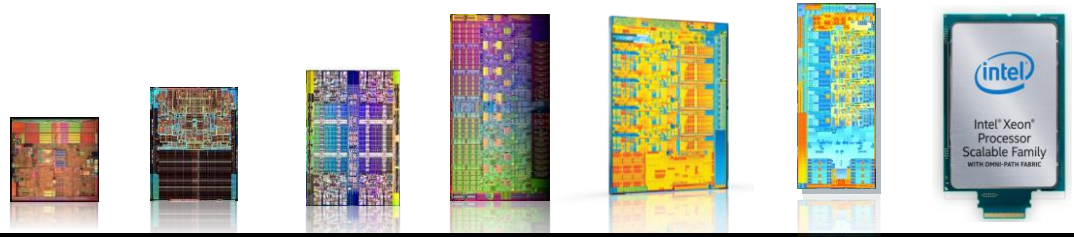
Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN)
Open source DNN functions for CPU / integrated graphics

nGraph
Open source compiler for deep learning model computations optimized for multiple devices (CPU, GPU, NNP) from multiple frameworks (TF, MXNet, ONNX)

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

THE EVOLUTION OF MICROPROCESSOR PARALLELISM

More cores → More Threads → Wider vectors

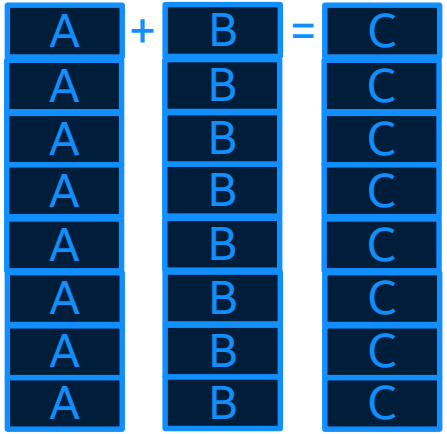


	Intel® Xeon® Processor 64-bit	Intel® Xeon® Processor 5100 series	Intel® Xeon® Processor 5500 series	Intel® Xeon® Processor 5600 series	Intel® Xeon® Processor E5-2600 v2 series	Intel® Xeon® Processor E5-2600 v3 series v4 series	Intel® Xeon® Scalable Processor[1]
Up to Core(s)	1	2	4	6	12	18-22	28
Up to Threads	2	2	8	12	24	36-44	56
SIMD Width	128	128	128	128	256	256	512
Vector ISA	Intel® SSE3	Intel® SSE3	Intel® SSE4- 4.1	Intel® SSE 4.2	Intel® AVX	Intel® AVX2	Intel® AVX-512

1. Product specification for launched and shipped products available on ark.intel.com.

Scalar

SIMD



Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

INTEL® OPTIMIZED SOFTWARE AND LIBRARIES

Optimization techniques for getting performance

Scaling

- Improve load balancing
- Reduce synchronization events, all-to-all comms

Utilize all the cores

- OpenMP, MPI
- Reduce synchronization events, serial code
- Improve load balancing

Vectorize/SIMD

- Unit strided access per SIMD lane
- High vector efficiency
- Data alignment

Efficient memory/cache use

- Blocking
- Data reuse
- Prefetching
- Memory allocation

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

PRODUCTIVITY WITH PERFORMANCE VIA INTEL® PYTHON*

Intel® Distribution for Python*

scikit learn | NumPy | SciPy | Numba | pandas $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$ | mpi4py | SMP | •••

Easy, out-of-the-box access to high performance Python

- Prebuilt accelerated solutions for data analytics, numerical computing, etc.
- Drop in replacement for your existing Python. No code changes required.

Learn More: software.intel.com/distribution-for-python

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

INSTALLING INTEL® DISTRIBUTION FOR PYTHON* 2019

Standalone Installer

Download full installer from
https://software.intel.com/en-us/intel-distribution-for-python

Anaconda.org
Anaconda.org/intel channel

```
> conda config --add channels intel
> conda install intelpython3_full
> conda install intelpython3_core
```

PyPI

```
> pip install intel-numpy
> pip install intel-scipy
> pip install mkl_fft
> pip install mkl_random
```

+ Intel library Runtime packages
+ Intel development packages

Docker Hub

```
docker pull intelpython/intelpython3_full
```

YUM/APT

Access for yum/apt:
https://software.intel.com/en-us/articles/installing-intel-free-libs-and-python

2.7 & 3.6
(3.7 coming soon)

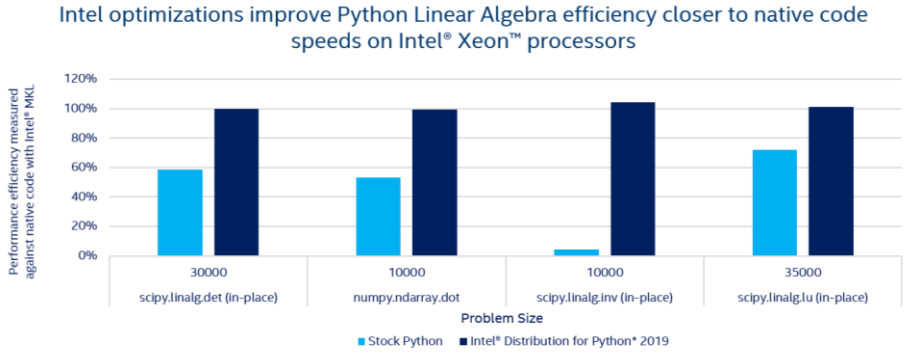
Linux*

Windows*

OS X*

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

FASTER PYTHON* WITH INTEL® DISTRIBUTION FOR PYTHON 2019



Performance results are based on testing as of **July 9, 2018** and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information, see Performance Benchmark Test Disclosure.

Testing by Intel as of July 9, 2018. Configuration: Stock Python: python 3.6.6 hc3d631a_0 installed from conda, numpy 1.15, numba 0.39.0, llvmlite 0.24.0, scipy 1.1.0, scikit-learn 0.19.2 installed from pip; Intel Python: Intel® Distribution for Python* 2019 Gold: python 3.6.5 intel_11, numpy 1.14.3 intel_py36_5, mkl 2019.0 intel_101, mkl_fft 1.0.2 intel_np114py36_6,mkl_random 1.0.1 intel_np114py36_6, numba 0.39.0 intel_np114py36_0, llvmlite 0.24.0 intel_py36_0, scipy 1.1.0 intel_np114py36_6, scikit-learn 0.19.1 intel_np114py36_35; OS: CentOS Linux 7.3.1611, kernel 3.10.0-514.el7.x86_64; Hardware: Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz (2 sockets, 18 cores/socket, HT:off), 256 GB of DDR4 RAM, 16 DIMMs of 16 GB@2666MHz

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice revision #20110804.
For more complete information about compiler optimizations, see our Optimization Notice.

Linear Algebra functions in Intel Distribution for Python are faster than equivalent stock Python functions

High Performance Python Distribution

- Accelerated NumPy, SciPy, scikit-learn well suited for scientific computing, machine learning & data analytics
- Drop-in replacement for existing Python. No code changes required
- Highly optimized for latest Intel processors
- Take advantage of Priority Support – connect direct to Intel engineers for technical questions[2]

What's New in 2019 version

- Faster Machine learning with Scikit-learn: Support Vector Machine (SVM) and K-means prediction, accelerated with Intel® Data Analytics Acceleration Library
- Integrated into Intel® Parallel Studio XE 2019 installer. Also available as easy command line standalone install.
- Includes XGBoost package (Linux* only)

Software & workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark & MobileMark, are measured using specific computer systems, components, software, operations & functions. Any change to any of those factors may cause the results to vary. You should consult other information & performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to http://www.intel.com/performance.

Optimization Notice

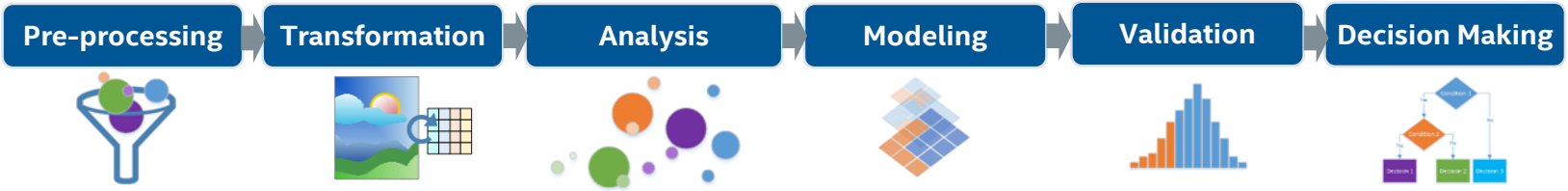
[2]Paid versions only.

Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

INTEL® DATA ANALYTICS ACCELERATION LIBRARY (INTEL® DAAL)

Building blocks for all data analytics stages, including data preparation, data mining & machine learning



Common Python*, Java*, C++ APIs across all Intel hardware

Optimizes data ingestion & algorithmic compute together for highest performance

Supports offline, streaming & distributed usage models for a range of application needs

Flexible interfaces to leading big data platforms including Spark*

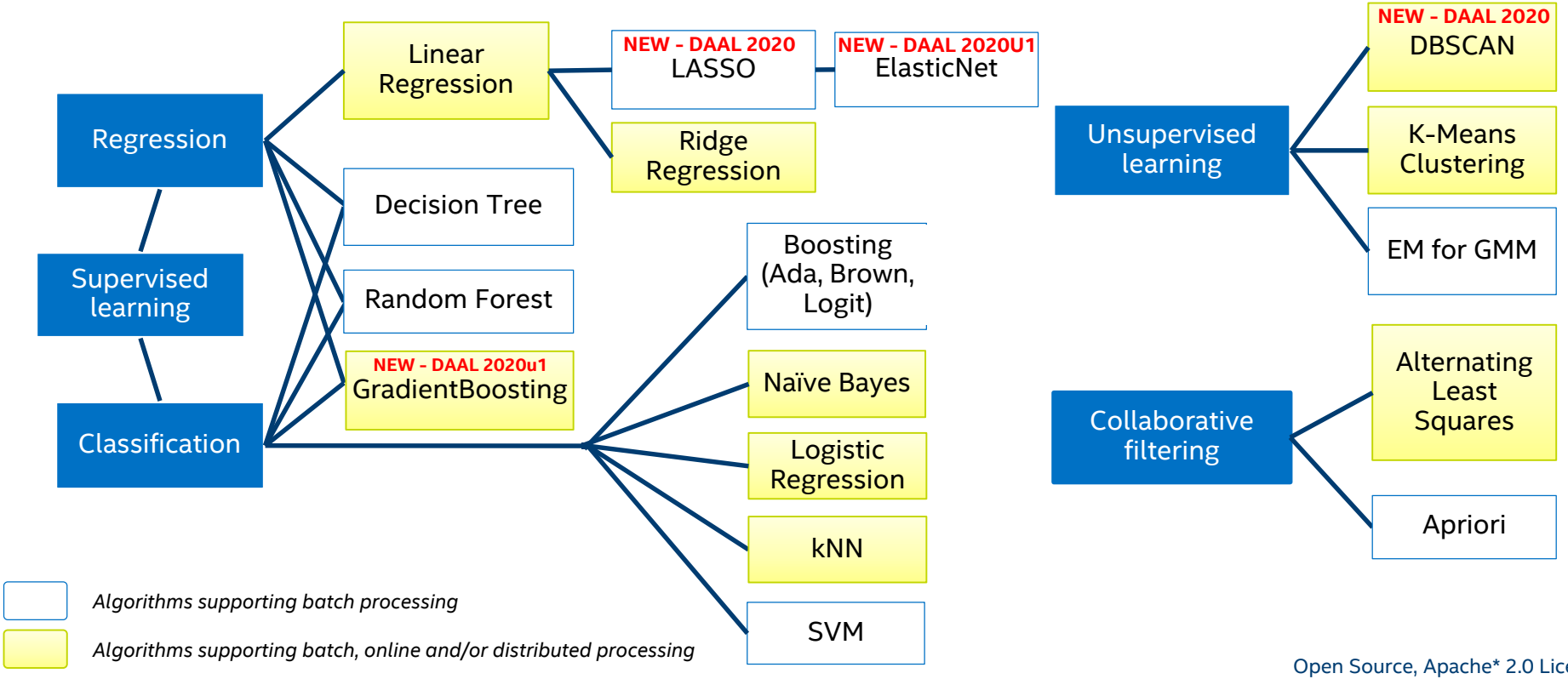
Split analytics workloads between edge devices & cloud to optimize overall application throughput

High Performance Machine Learning & Data Analytics Library

Open Source, Apache* 2.0 License

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

MACHINE LEARNING ALGORITHMS



Open Source, Apache* 2.0 License

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

DATA TRANSFORMATION & ANALYSIS ALGORITHMS

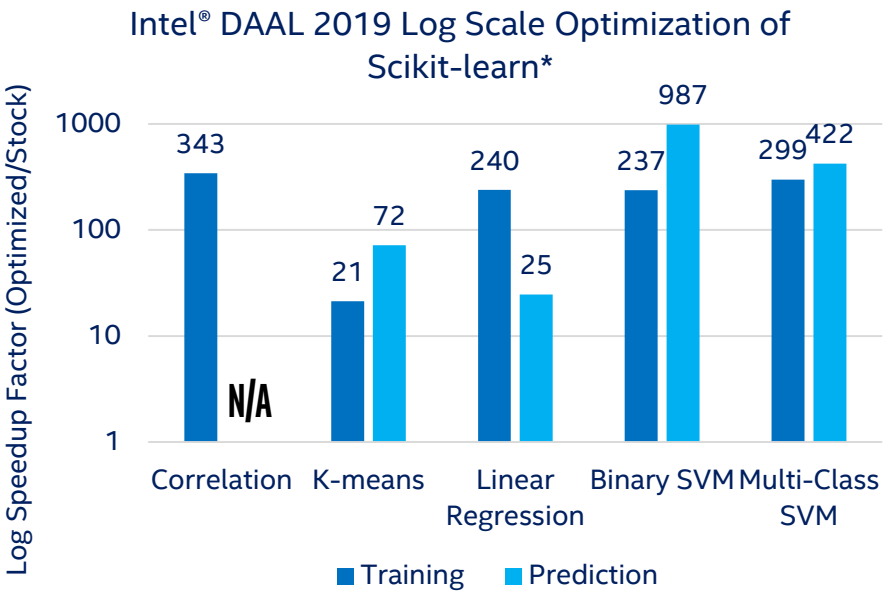
- **Basic statistics for datasets**
  - Low order moments
  - Quantiles
  - Order statistics
- **Correlation and dependence**
  - Cosine distance
  - Correlation distance
  - Variance-Covariance matrix
- **Matrix factorizations**
  - SVD
  - QR
  - Cholesky
  - NEW - DAAL 2020 tSVD
- **Dimensionality reduction**
  - PCA
  - Association rule mining (Apriori)
  - Optimization solvers (SGD, AdaGrad, lBFGS)
- **Outlier detection**
  - Univariate
  - Multivariate
  - Math functions (exp, log,...)

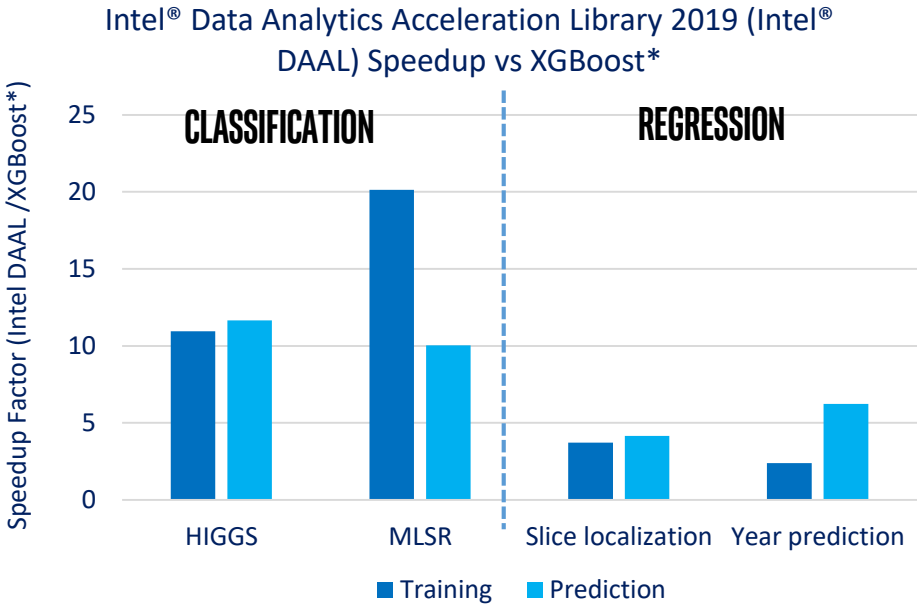
Algorithms supporting batch processing

Algorithms supporting batch, online and/or distributed processing

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

INTEL® DAAL - PERFORMANCE





Performance results are based on testing as of **July 9, 2018** and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that proc when combined with other products. For more complete information, see Performance Benchmark Test Disclosure.

Testing by Intel as of July 9, 2018. Configuration: Intel® Xeon® Platinum 8180 H0 205W, 2x28@2.50GHz, 192GB, 12x16gb DDR4-2666, Intel® Data Analytics Acceleration Library (Intel® DAAL 2019), RHEL 7.2

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Inte not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizatior specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice revision #20110804.
For more complete information about compiler optimizations, see our Optimization Notice.

Performance results are based on testing as of **July 9, 2018** and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information, see Performance Benchmark Test Disclosure.

Testing by Intel as of July 9, 2018. Configuration: Intel® Xeon® Gold 6140 CPU, 2x18@2.30GHz, 256GB, 16x16gb DDR4-2666, Intel® Data Analytics Acceleration Library (Intel® DAAL 2019), Optimized: Scikit-learn*_intel 0.19.1, Numpy*_intel 1.14.3 Stock: Scikit-learn* 0.19.2, Numpy* 1.15.0, CentOS Linux 7.3.1611

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice revision #20110804.
For more complete information about compiler optimizations, see our Optimization Notice.

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

FASTER, SCALABLE CODE WITH INTEL® MATH KERNEL LIBRARY

LINEAR ALGEBRA

- BLAS
- LAPACK
- ScaLAPACK
- Sparse BLAS
- Iterative sparse solvers
- PARDISO*
- Cluster Sparse Solver

FFTS

- Multidimensional
- FFTW interfaces
- Cluster FFT

VECTOR RNGS

- Congruential
- Wichmann-Hill
- Mersenne Twister
- Sobol
- Neirderreiter
- Non-deterministic

SUMMARY STATISTICS

- Kurtosis
- Variation coefficient
- Order statistics
- Min/max
- Variance-covariance

VECTOR MATH

- Trigonometric
- Hyperbolic
- Exponential
- Log
- Power
- Root

AND MORE

- Splines
- Interpolation
- Trust Region
- Fast Poisson Solver

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

INTEL® MATH KERNEL LIBRARY FOR DEEP NEURAL NETWORKS (INTEL® MKL-DNN)

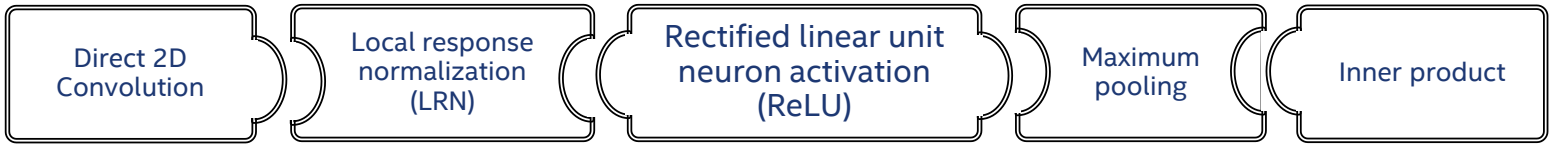
For developers of deep learning frameworks featuring optimized performance on Intel hardware

Distribution Details

- Open Source
- Apache* 2.0 License
- Common DNN APIs across all Intel hardware.
- Rapid release cycles, iterated with the DL community, to best support industry framework integration.
- Highly vectorized & threaded for maximal performance, based on the popular Intel® Math Kernel Library.

github.com/01org/mkl-dnn

Examples:



Accelerate Performance of Deep Learning Models

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

INTEL® AI OPTIMIZED FRAMEWORKS

Popular DL Frameworks are now optimized for CPU!

CHOOSE YOUR FAVORITE FRAMEWORK

TensorFlow™* | Caffe* | mxnet* | BigDL FOR Apache Spark* | PYTORCH

See installation guides at ai.intel.com/framework-optimizations/

More under optimization: Caffe2* PYTORCH* PaddlePaddle*

SEE ALSO: Machine Learning Libraries for Python (Scikit-learn, Pandas, NumPy), R (Cart, randomForest, e1071), Distributed (MLlib on Spark, Mahout)
**Limited availability today*
Other names and brands may be claimed as the property of others.

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

INTEL® DISTRIBUTION OF OPENVINO™ TOOLKIT

OpenVINO™

DEEP LEARNING

Caffe | TensorFlow | ONNX | mxnet | KALDI

- Model Optimizer
- Inference Engine
- Supports 100+ public models, incl. 30+ pretrained models

COMPUTER VISION

OpenCV | OpenCL™ | OpenVX™

- Computer vision library (kernel & graphic APIs)
- Optimized media encode/decode functions

SUPPORTS MAJOR AI FRAMEWORKS

Rapid adoption by developers

CROSS-PLATFORM FLEXIBILITY

Multiple products launched based on this toolkit

HIGH PERFORMANCE, HIGH EFFICIENCY

Breadth of product portfolio

Strong Adoption + Rapidly Expanding Capability

software.intel.com/openvino-toolkit

Obtain open source version at 01.org/openvinotoolkit

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

INTEL® DEEP LEARNING DEPLOYMENT TOOLKIT

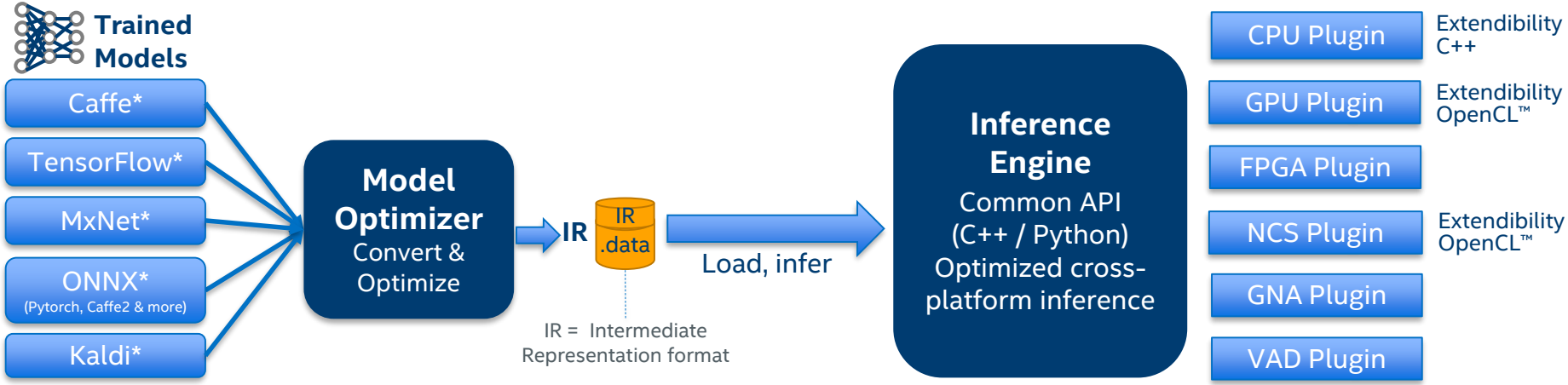
For Deep Learning Inference – Part of Intel® Distribution of OpenVINO toolkit

Model Optimizer

- **What it is**: A Python*-based tool to import trained models and convert them to Intermediate representation.
- **Why important**: Optimizes for performance/space with conservative topology transformations; biggest boost is from conversion to data types matching hardware.

Inference Engine

- **What it is**: High-level inference API
- **Why important**: Interface is implemented as dynamically loaded plugins for each hardware type. Delivers best performance for each type without requiring users to implement and maintain multiple code pathways.



GPU = Intel CPU with integrated graphics processing unit/Intel® Processor Graphics

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

NAVIGATING THE AI PERFORMANCE PACKAGE

INTRODUCTION TO AI

Overview of Deep Learning Software

INTEL® XEON® SCALABLE PROCESSORS

First and second generation: Skylake / Cascade Lake

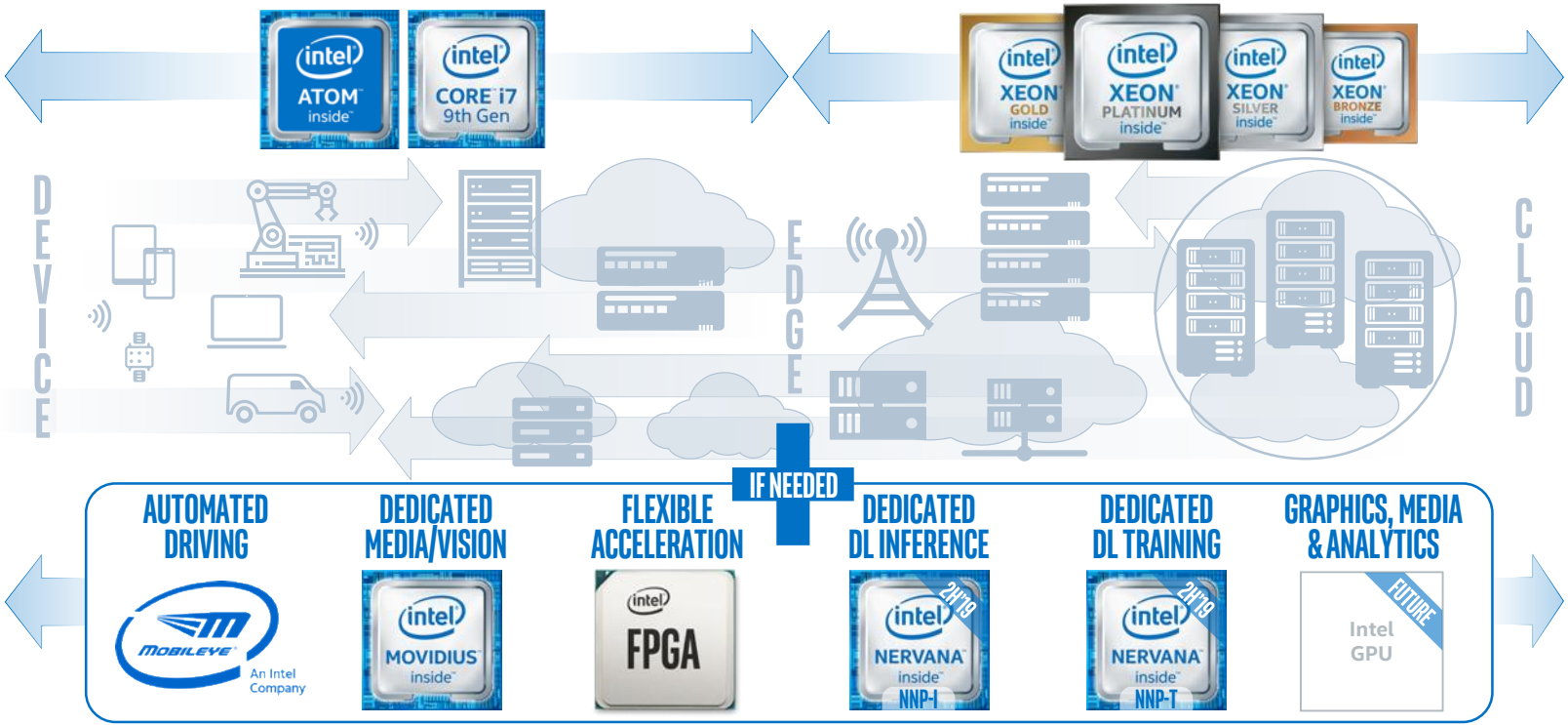
- Deploy AI Everywhere on Intel® Architecture
- Intel® AVX-512 (Advanced Vector Instructions)

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

Visit:
www.intel.ai/technology

DEPLOY AI ANYWHERE

with unprecedented hardware choice



All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

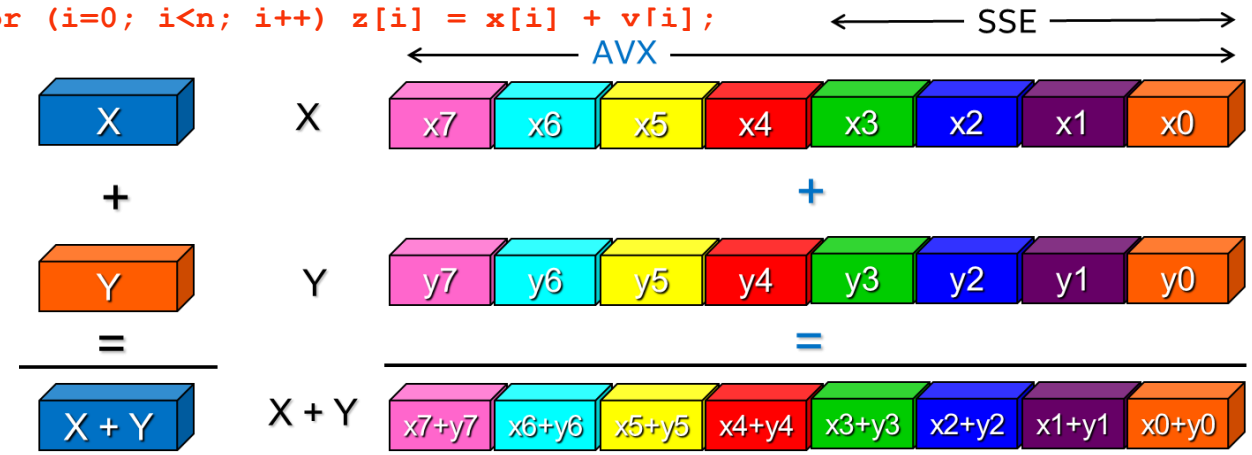
Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

SIMD PROCESSING

Single instruction multiple data (**SIMD**) allows to execute the same operation on multiple data elements using larger registers.

- Scalar mode
  - one instruction produces one result
  - E.g. vadd**s**s, (vadd**s**d)
- Vector (SIMD) mode
  - one instruction can produce multiple results
  - E.g. vadd**p**s, (vadd**p**d)

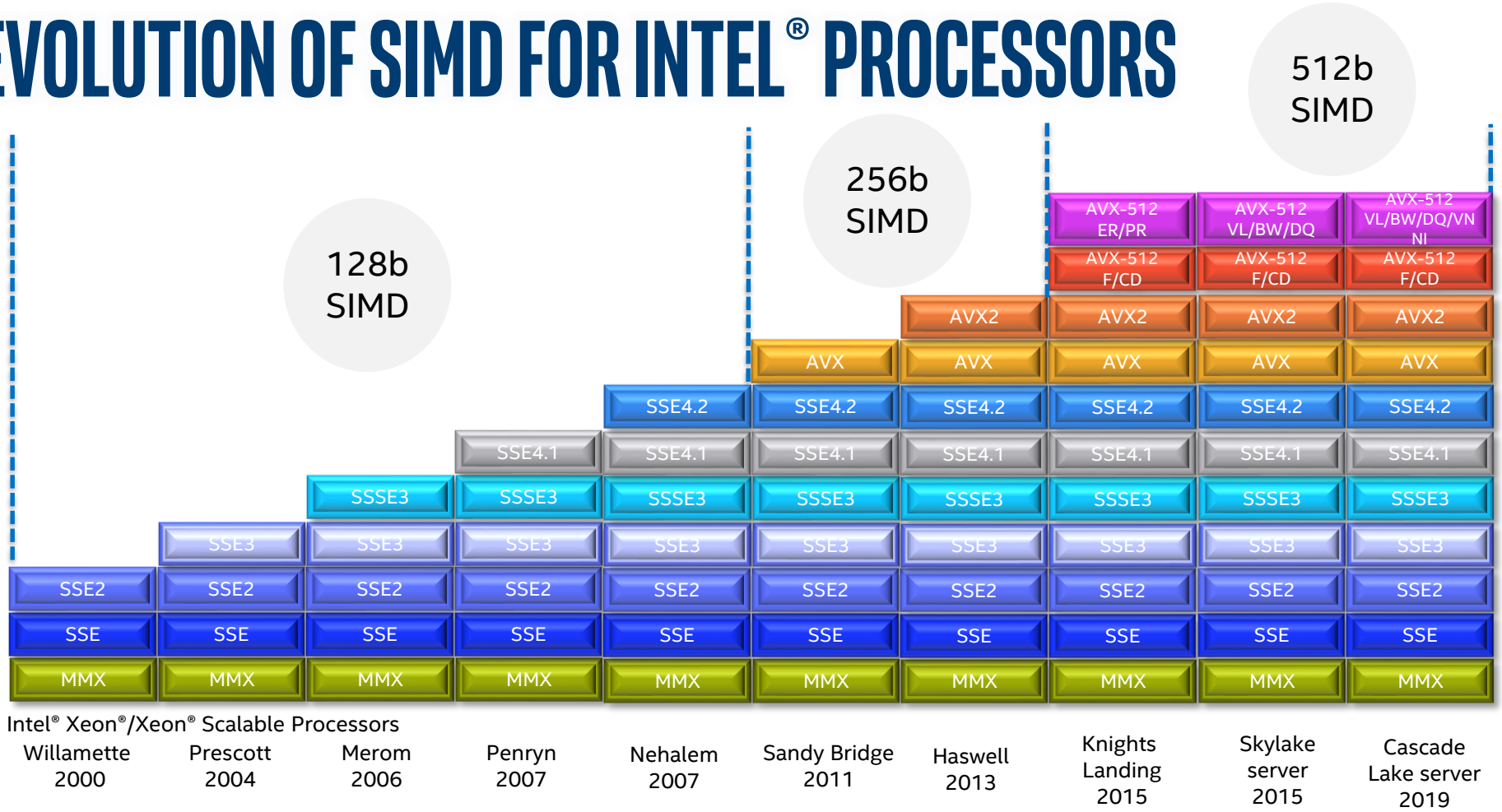
```
for (i=0; i<n; i++) z[i] = x[i] + v[i];
```



- SSE (128 Bits reg.): -> 4 floats
- AVX (256 Bits reg.): -> 8 floats
- AVX512 (512 Bits reg.): -> 16 floats

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

EVOLUTION OF SIMD FOR INTEL® PROCESSORS



Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

AVX-512 COMPRESS AND EXPAND

VCOMPRESSPD|PS|D|Q	Store sparse packed floating-point values into dense memory
VEXPANDPD|PS|D|Q	Load sparse packed floating-point values from dense memory

double/single-precision/doubleword/quadword

```
vcompresspd YMMWORD PTR [rsi+rax*8]{k1}, ymm1
```

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

COMPRESS LOOP PATTERN

AUTO-VECTORIZATION

https://godbolt.org/z/x7gNfb

```
int compress(double *a, double * __restrict b, int na)
{
    int nb = 0;
    for (int ia=0; ia <na; ia++)
    {
        if (a[ia] > 0.)
            b[nb++] = a[ia];

    }

    return nb;
}
```

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

COMPRESS LOOP PATTERN

AUTO-VECTORIZATION

https://godbolt.org/z/x7gNfb

```
int compress(double *a, double *
__restrict b, int na)
{
    int nb = 0;
    for (int ia=0; ia <na; ia++)
    {
        if (a[ia] > 0.)
            b[nb++] = a[ia];

    }

    return nb;
}
```

Targeting Intel® AVX2

-xcore-avx2 -qopt-report-file=stderr -qopt-report-phase=vec

LOOP BEGIN

remark #15344: loop was not vectorized: vector dependence prevents vectorization.

remark #15346: vector dependence: assumed FLOW dependence between b[nb] (7:4) and a[ia] (7:4)

LOOP END

Targeting Intel® AVX-512

-xcore-avx512 -qopt-report-file=stderr -qopt-report-phase=vec

LOOP BEGIN

remark #15300: LOOP WAS VECTORIZED

LOOP END

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

COMPRESS LOOP PATTERN

AUTO-VECTORIZATION

https://godbolt.org/z/x7gNfb

```
int compress(double *a, double * __restrict b, int na)
{
    int nb = 0;
    for (int ia=0; ia <na; ia++)
    {
        if (a[ia] > 0.)
            b[nb++] = a[ia];

    }

}
```

```
movsxd      rax, eax
xor         r11d, r11d
kmovw       r8d, k1
popcnt      r11d, r8d
vcompresspd YMMWORD PTR [rsi+rax*8]{k1}, ymm1
add         eax, r11d
```

Key Take Aways

Compress/Expand loop pattern doesn't vectorize on architectures like Intel® AVX2 and the previous ones and does with Intel® AVX512

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

COMPRESS LOOP PERFORMANCE

Compiler Options	Speedup (in C)	Speedup (in FORTRAN)
Simple Loops (–O2 –xCORE-AVX2)	1.0x	1.0x
(-O2 –xCORE-AVX512)	12.8x	12.2x

Performance tests are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary.
The results above were obtained on an Intel® Xeon® Platinum 8168 system, frequency 2.7 GHz, running Red Hat* Enterprise Linux* Server 7.2 and using the Intel® Fortran Compiler version 18.0 update 1.

Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.
Notice Revision #20110804.

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

NAVIGATING THE AI PERFORMANCE PACKAGE

INTRODUCTION TO AI

Overview of Deep Learning Software

INTEL® XEON® SCALABLE PROCESSORS

First and second generation: Skylake / Cascade Lake

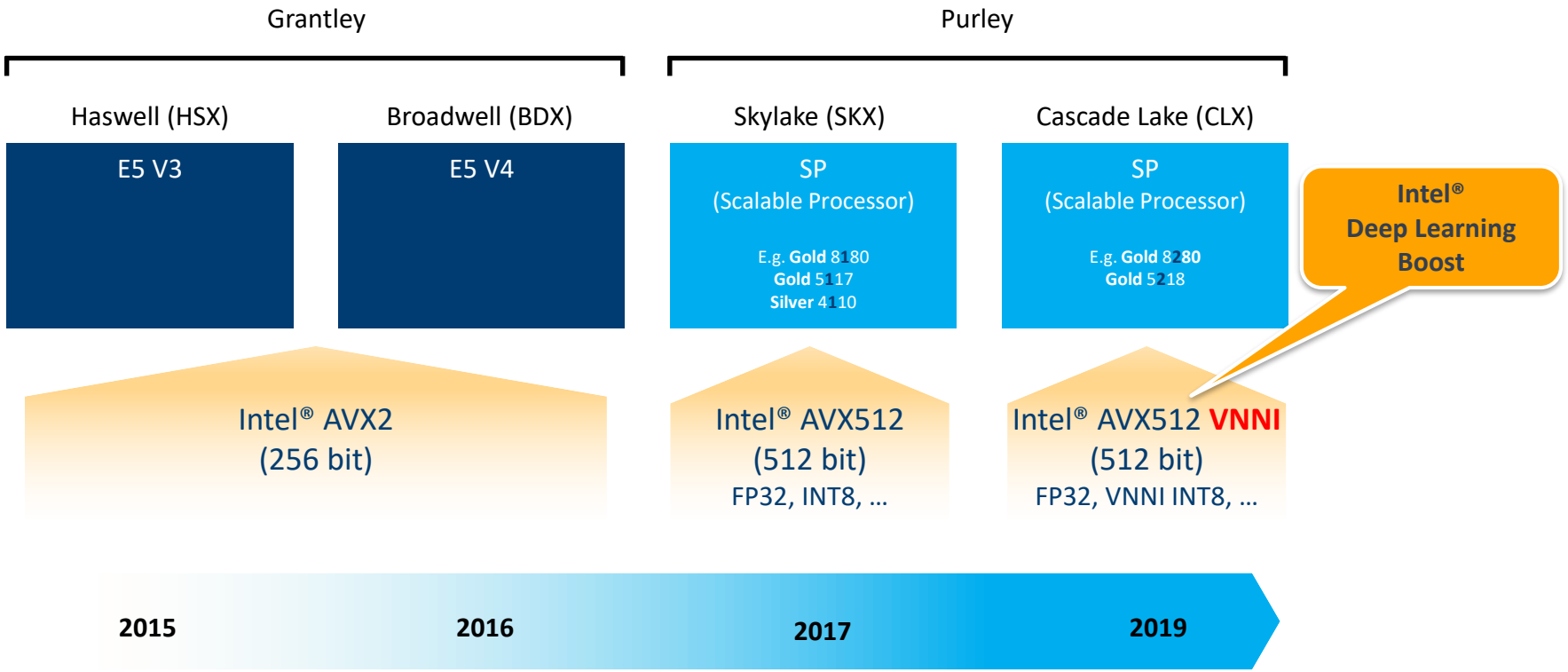
INTEL® DEEP LEARNING BOOST

Intel® AVX-512 Vector Neural Network Instructions (VNNI)

- Boost Deep Learning Inference with VNNI

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

FAST EVOLUTION OF AI CAPABILITY ON INTEL® XEON® PLATFORM



Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

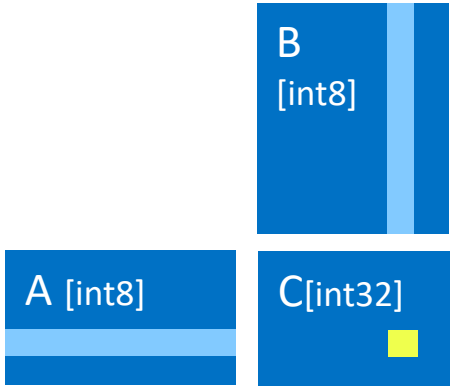
- ***Intel® Deep Learning Boost*** **is a new set of AVX-512 instructions *designed to deliver significant,***

more efficient Deep Learning (Inference) acceleration on second generation Intel® Xeon® Scalable processor (codename "Cascade Lake")

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

DEEP LEARNING FOUNDATIONS

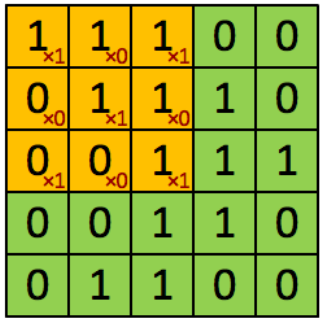
- Matrix Multiplies are the foundation of many DL applications
  - **Multiply** a row*column values, **accumulate** into a single value
- Traditional HPC and many AI training workloads use floating point
  - Massive dynamic range of values (FP32 goes up to ~2^128)
- Why INT8 for Inference?
  - More power efficient per operation due to smaller multiplies
  - Reduces pressure on cache and memory subsystem
  - Precision and dynamic range sufficient for many models
- What's different about INT8?
  - Much smaller dynamic range than FP32: 256 values
    - Requires *accumulation into INT32* to avoid overflow (FP handles this "for free" w/ large dynamic range)



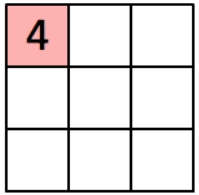
Matrix Multiply
A x B = C

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

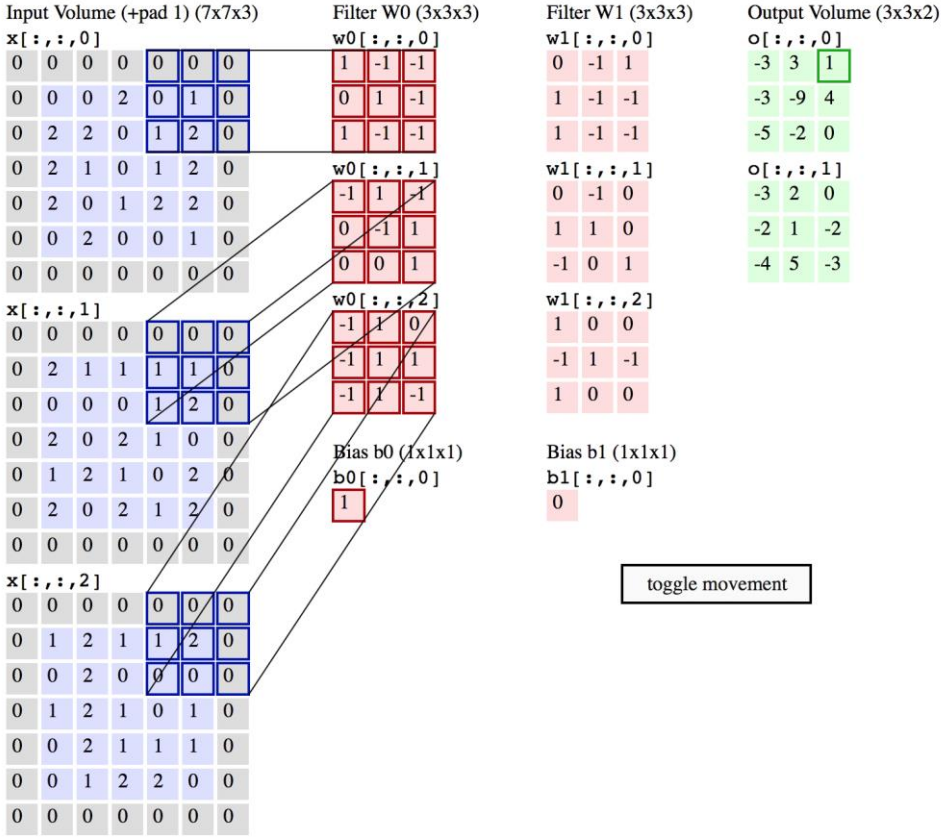
FAST EVOLUTION OF AI CAPABILITY ON XEON PLATFORM



Image



Convolved Feature



Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

PRECISION FOR DEEP LEARNING

- Precision is a measure of the detail used for numerical values primarily measured in bits

FP32 TRAINING → BF16 BIT TRAINING

Better cache usage
Avoids bandwidth bottlenecks
Maximizes compute resources
Lesser silicon area & power

FP32 INFERENCE → 8 BIT INFERENCE

NEED - FP32 WEIGHTS → INT 8 WEIGHTS FOR INFERENCE

BENEFIT - REDUCE MODEL SIZE AND ENERGY CONSUMPTION

CHALLENGE - POSSIBLE DEGRADATION IN PREDICTIVE PERFORMANCE

SOLUTION - QUANTIZATION WITH MINIMAL LOSS OF INFORMATION

DEPLOYMENT - OPENVINO TOOLKIT OR DIRECT FRAMEWORK (TF)

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

INSTRUCTIONS FUSION

VPMADDUBSW

Multiply and Add Packed Signed and Unsigned Bytes

SRC 1 8-bit	A_0	A_1	A_2	A_3	. . . .	A_{63}
SRC 2 8-bit	B_0	B_1	B_2	B_3	. . . .	B_{63}
DEST 16-bit	$A_0*B_0 + A_1*B_1$	$A_2*B_2 + A_3*B_3$	. . . .			$A_{62}*B_{62} + A_{63}*B_{63}$

VPMADDWD

effectively upconvert to 32-bit and horizontal add of neighbors

SRC 1 16-bit	$A_0*B_0 + A_1*B_1$	$A_2*B_2 + A_3*B_3$	. . . .	$A_{62}*B_{62} + A_{63}*B_{63}$
SRC 2 16-bit	1	1	. . . .	1
DEST 32-bit	$A_0*B_0 + A_1*B_1 + A_2*B_2 + A_3*B_3$	. . . .	$A_{60}*B_{60} + A_{61}*B_{61} + A_{62}*B_{62} + A_{63}*B_{63}$	

VPADDD

Add Packed Double-Precision Floating-Point Values

SRC 1 32-bit	$A_0*B_0 + A_1*B_1 + A_2*B_2 + A_3*B_3$	. . . .	$A_{60}*B_{60} + A_{61}*B_{61} + A_{62}*B_{62} + A_{63}*B_{63}$
SRC 2 32-bit	C_0	. . . .	C_{15}
Dest 32-bit	$A_0*B_0 + A_1*B_1 + A_2*B_2 + A_3*B_3 + C_0$	. . . .	$A_{60}*B_{60} + A_{61}*B_{61} + A_{62}*B_{62} + A_{63}*B_{63} + C_{15}$

ISA Fusion

VPDPBUSD

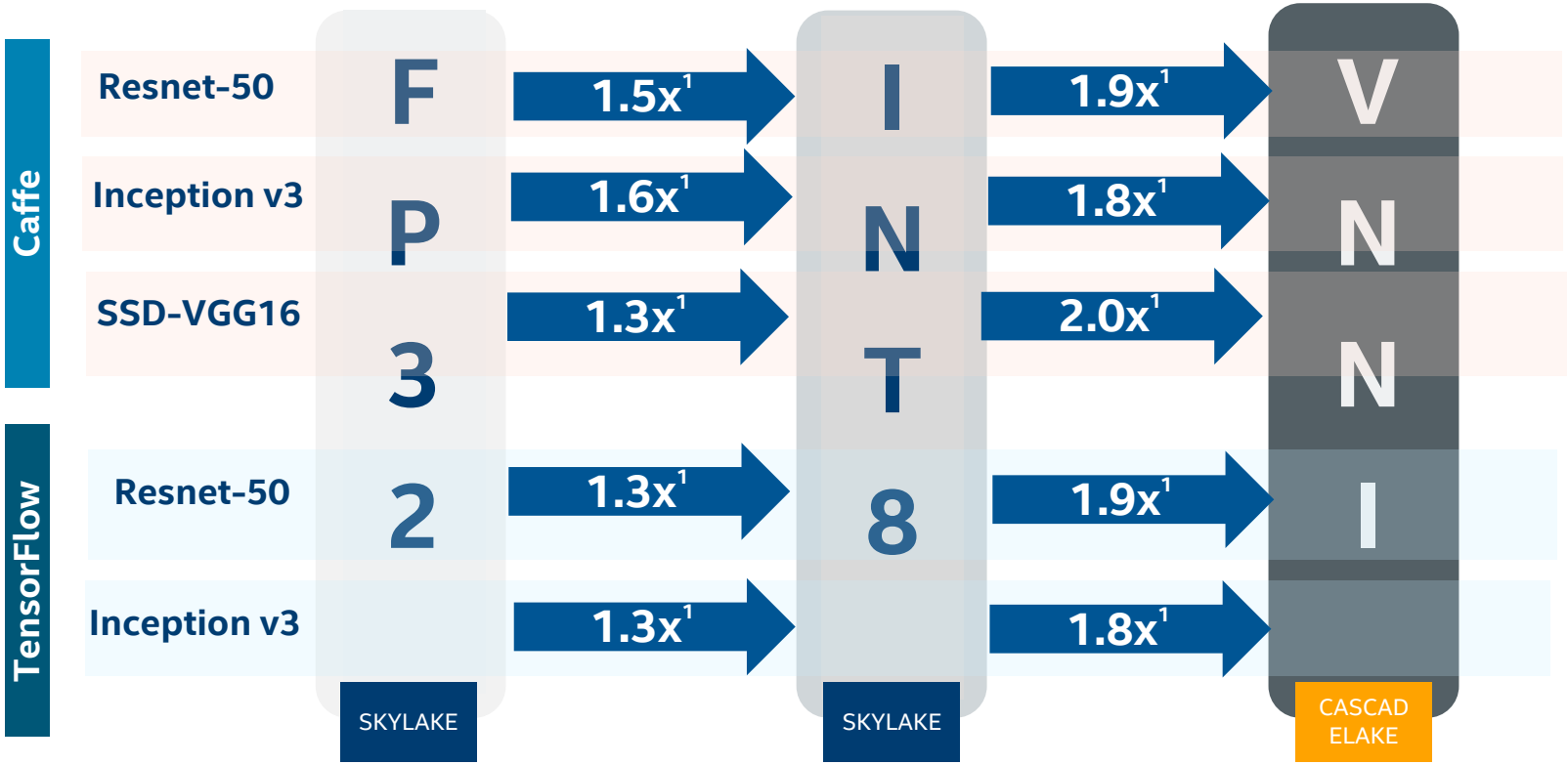
Multiply and Add Unsigned and Signed Bytes

SRC 1 8-bit	A_0	A_1	A_2	A_3	. . . .	A_{63}
SRC 2 8-bit	B_0	B_1	B_2	B_3	. . . .	B_{63}
SRC3 / DEST 32-bit	C_0				. . . .	C_{15}
	$A_0*B_0 + A_1*B_1 + A_2*B_2 + A_3*B_3 + C_0$				. . .	$A_{60}*B_{60} + A_{61}*B_{61} + A_{62}*B_{62} + A_{63}*B_{63} + C_{15}$

VPMADDUBSW + VPMADDWD + VPADDD fused into VPDPBUSD (3x peak OPs)

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

GENERATIONAL PERFORMANCE PROJECTIONS ON INTEL® SCALABLE PROCESSOR FOR DEEP LEARNING INFERENCE FOR POPULAR CNNS



[1]8/24/2018) Results have been estimated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. .Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit: http://www.intel.com/performance

Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice
*Other names and brands may be claimed as the property of others.

SO WHAT?

AVX512_VNNI is a new set of AVX-512 instructions to boost Deep Learning performance

- VNNI includes FMA instructions for:
  - 8-bit multiplies with 32-bit accumulates (u8 x s8 ⇒ s32)
  - 16-bit multiplies with 32-bit accumulates (s16 x s16 ⇒ s32)
- Theoretical peak compute gains are:
  - 4x int8 OPS over fp32 OPS and ¼ memory requirements
  - 2x int16 OPS over fp32 OPS and ½ memory requirements
- Ice Lake and future microarchitectures will have AVX512_VNNI

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

ENABLING INTEL® DL BOOST ON CASCADE LAKE

THEORETICAL IMPROVEMENTS: FP32 VS. INT8 & DL BOOST

UP TO 4X BOOST IN MAC/CYCLE

UP TO 4X IMPROVED PERFORMANCE / WATT

DECREASED MEMORY BANDWIDTH

IMPROVED CACHE PERFORMANCE

UP NEXT: MICROBENCHMARKING WITH INTEL® MKL-DNN'S

Workloads

Topologies

TensorFlow, mxnet, Caffe2, Caffe, PaddlePaddle

Frameworks

Intel® MKL-DNN Libraries

Intel® Processors

Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.

intel®

Software