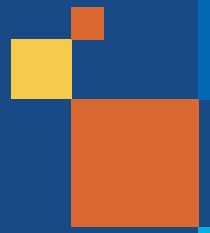
9/29/2020

Intel AI Overview

Intel Corporation

APJ Datacenter Group Sales

Hiroshi Ouchiyama

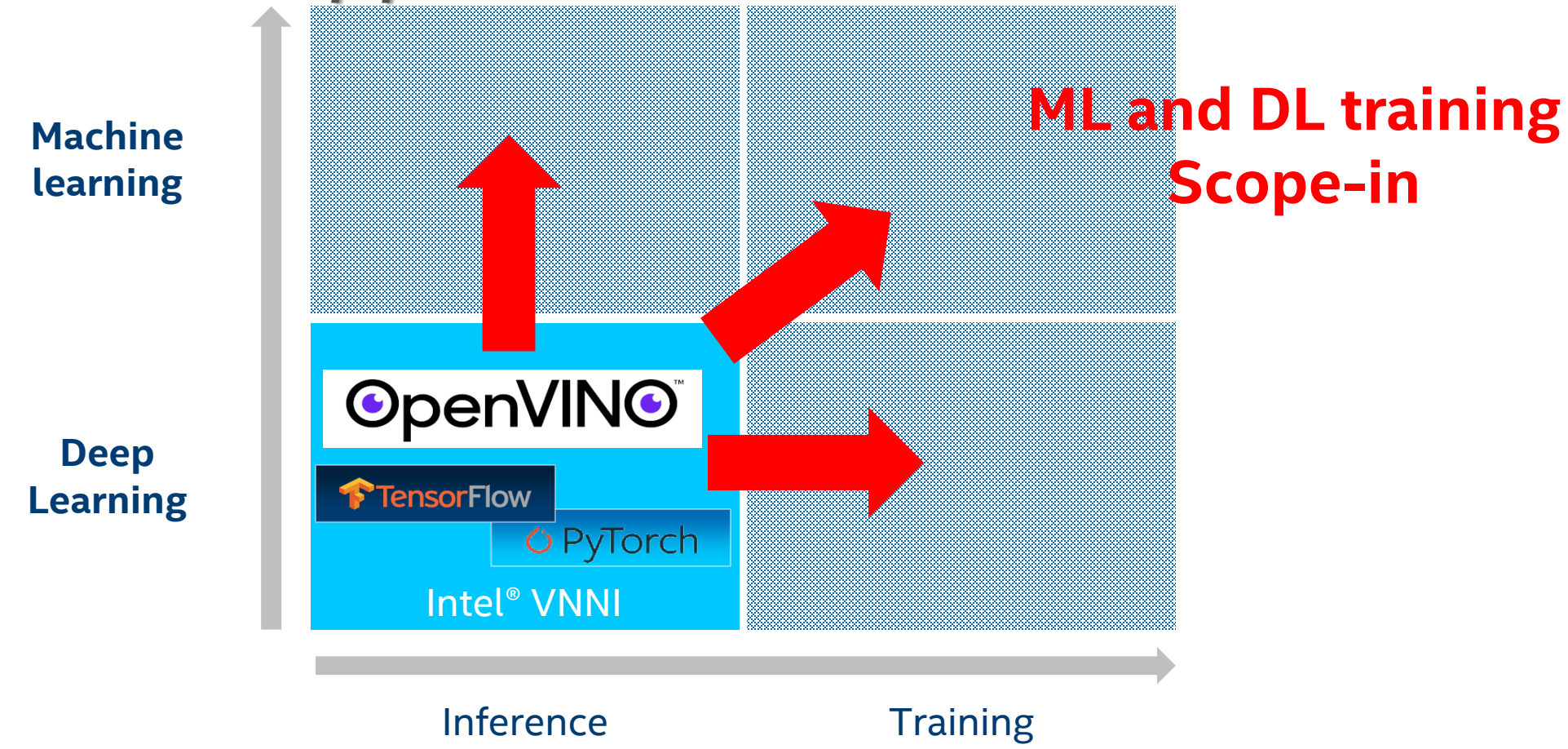


intel®

AI also can be run on CPU

- Versatility and flexibility for all workloads are the hallmarks of the CP

Recent Intel AI
~Supports a wider AI workload~



AVX-512 & DL Boost on Intel CPU

- AVX-512 (SIMD) is installed, contributing to the improvement of parallel processing performance. Furthermore, further acceleration can be expected with the dedicated Deep Learning Boost instruction.

Intel® AVX-512

(Intel® Advanced Vector Extensions 512)

+

Intel® DL Boost

(Intel® Deep Learning Boost)

Inside

From 10th gen Ice Lake

Inside

From Skylake, AVX-512
From Cascade Lake, DL Boost

Intel® AI Software: ML and DL

DEVELOPER TOOLS

App Developers

SW Platform Developer

TOPOLOGIES & MODELS

Data Scientist

FRAMEWORKS

Data Scientist

GRAPH

ML Performance Engineer

KERNEL

ML Performance Engineer

MACHINE LEARNING

ANALYTICS ZOO

- **Intel Distribution for Python** (SKlearn, Pandas)

python

- Intel Data Analytics Acceleration Library (Intel DAAL)
- Intel Math Kernel Library (Intel MKL)

DEEP LEARNING

OpenVINO

TensorFlow

PyTorch

- Intel Machine Learning Scaling Library (Intel MLSL)
- Intel® Deep Neural Network Library (DNNL)

MANAGEMENT TOOLS

Containers

kubernetes

Kubeflow

Deep Learning Reference Stack

Data Analytics Reference Stack

Architect & DevOps

CPU

CPU ▪ GPU ▪ FPGA ▪ 専用

Red font products are the most broadly applicable SW products for AI users

Deep Learning Framework (Optimizations by Intel)

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

TensorFlow

mxnet

Caffe2 PyTorch

BigDL for Spark

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

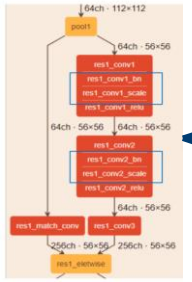
More framework optimizations underway (e.g., PaddlePaddle*, CNTK* and more)

SEE ALSO: Machine Learning Libraries for Python (Scikit-learn, Pandas, NumPy), R (Cart, randomForest, e1071), Distributed (MlLib on Spark, Mahout)
*Limited availability today
Optimization Notice

DL推論

Optimization and Quantization of Deep Learning Models for Further Performance Improvement of Inference Processing

- Optimization: Make models smarter by removing unnecessary Ops, integrating multiple Ops, etc.
- Quantization*: Make models slim by converting internal numerical representation of the model from FP32 to INT8.

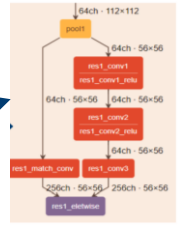


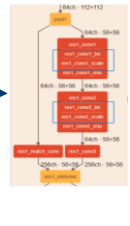
元のモデル
（TensorFlow*、PyTorch* などで作成）

Optimization

Quantization

Optimization & Quantization





by OpenVINO™

by TensorFlow PyTorch

A quantization tool is available for each framework.

by OpenVINO™

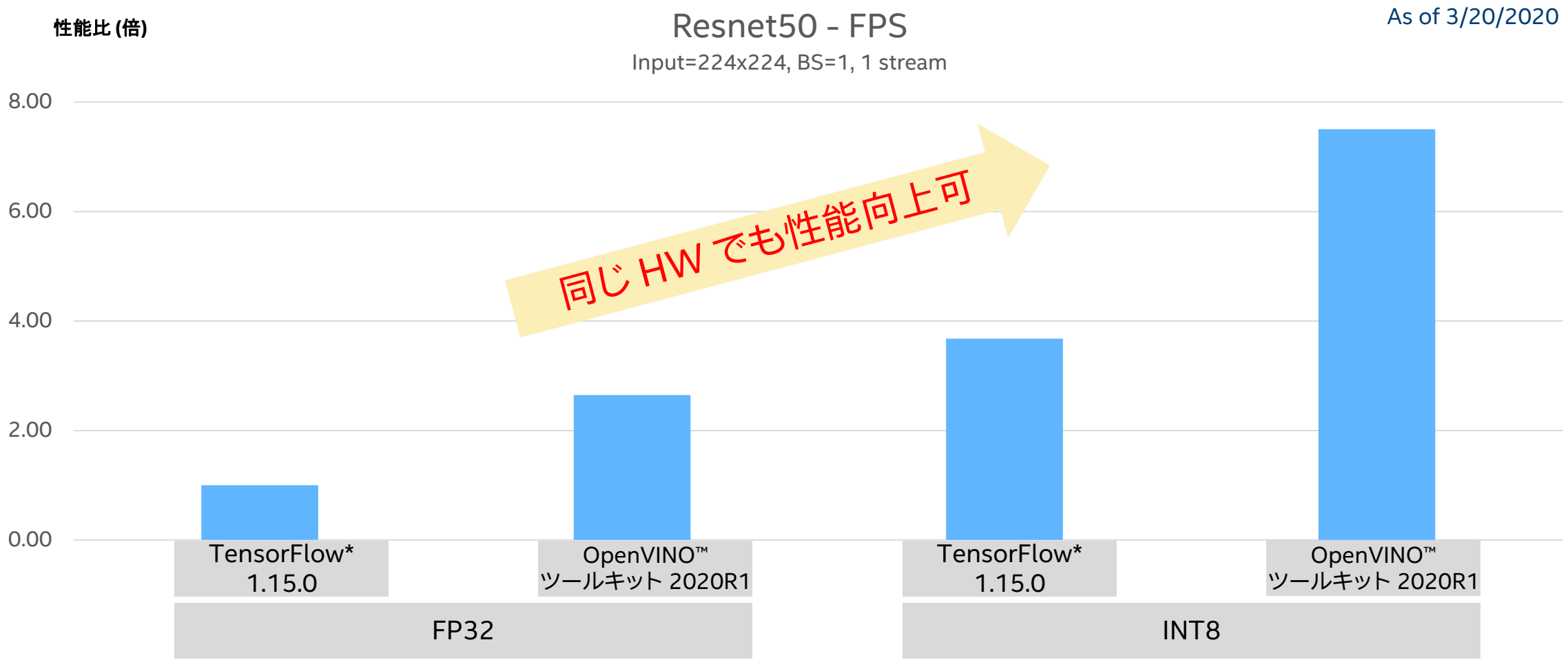
* 2nd generation Intel® Xeon® scalable processors and later with Intel® Deep Learning Boost (VNNI) as of May 2020, effective on 10th generation Intel® Core™ processor family (Ice Lake† only) and later

参考値

DL推論

Deep Learning Inference Processing Benchmark

Intel® Xeon® Gold 6254 processor @ 2.10GHz (18 cores x 1 sockets)



注)インテル社員による性能確認のための個人的なベンチマーク結果であり、インテルの公式結果ではありません。

CheXNet Performance Optimization by OpenVINO™

Before Optimization←

→ After Optimization

Measured the batch inference performance with 22K images

Optimization

- Export the model to ONNX.
- Convert the ONNX to IR by OpenVINO's Model Optimizer.
- Run the IR on OpenVINO's inference engine.

Quantization

- Quantize the IR to INT8 format by OpenVINO's quantization tool
- Run the IR on OpenVINO's inference engine. (on VNNI)

Parallelization

- Change the source code to use asynchronized processing and multi threading on OpenVINO's inference engine (8 parallel).

11,177 sec (Baseline) on Xeon 6252 → x10.0 → 1,116 sec → x3.1 → 359 sec → x1.4 → 251 sec

x 44.5 against Baseline

Please refer the link below to find the specific source code.
https://github.com/taneishi/CheXNet

Training with Huge Memory ~U-Net Training by NUS~

DL学習

NUS National University of Singapore | Saw Swee Hock School of Public Hea

GPU-based env

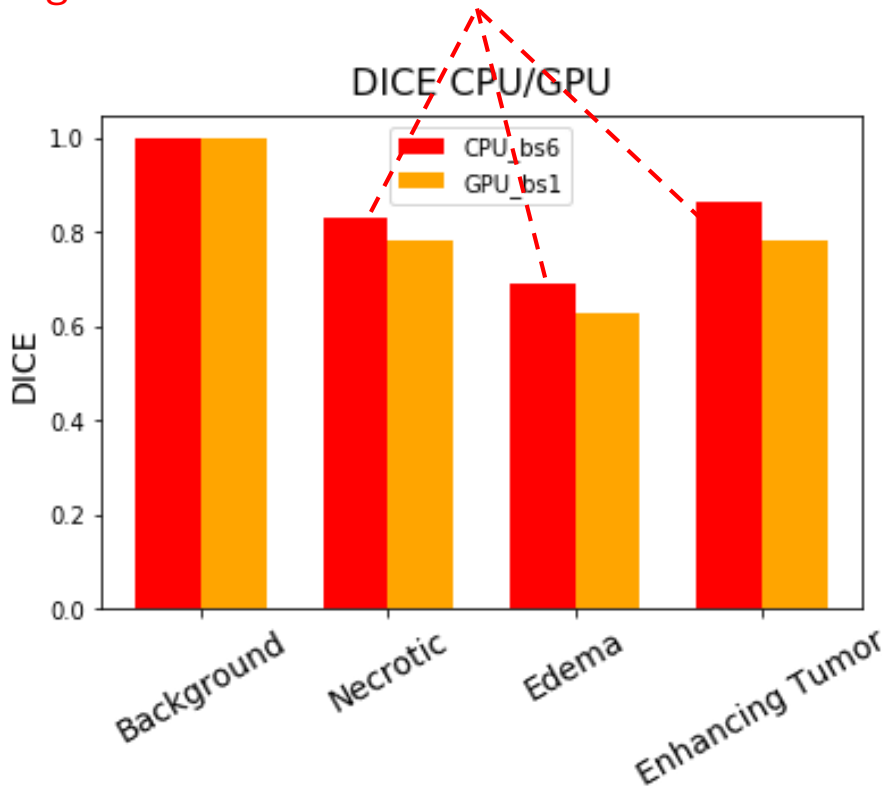
- V100 GPU (32GB memory)
- 10 CPU cores
- 126GB RAM
- **Batch size of 1**

CPU-based env

- 2 x Intel Platinum CPUs.
- 2 x 24 CPU cores
- 384GB RAM
- **Batch size of 6**

Result

The DICE (model accuracy) is on average 5% higher for models trained on an Intel® CPU.



What if I want performance?

↓

Use multiple CPUs in a bundle
In other words, Distributed Training ☝

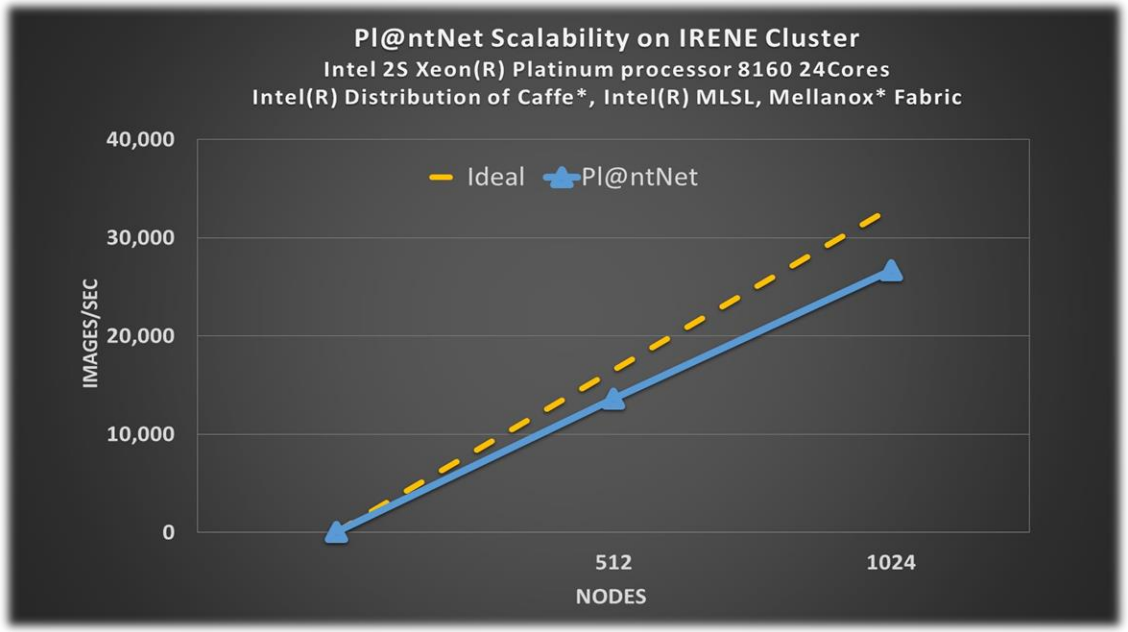
DL学習

Scaling Efficient Deep Learning on Existing Infrastructure: The Case of GENCI and CERN

GENCI

French research institute focused on numerical simulation and HPC across all scientific and industrial fields

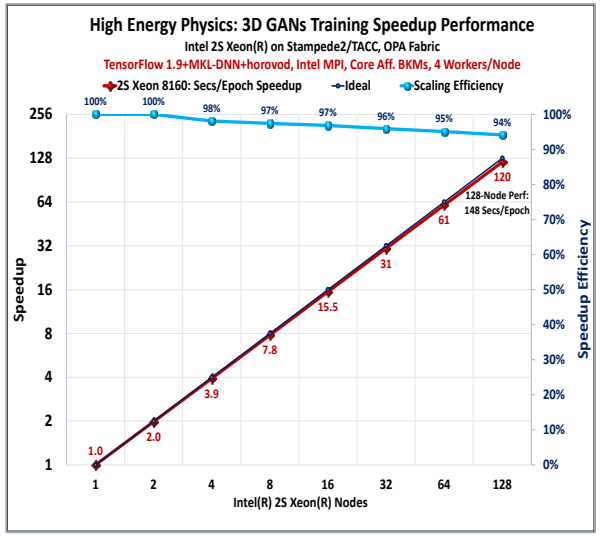
Succeeded in training a plant classification model for 300K species, 1.5TByte dataset of 12 million images on 1024 2S Intel® Xeon® Nodes with Resnet50.

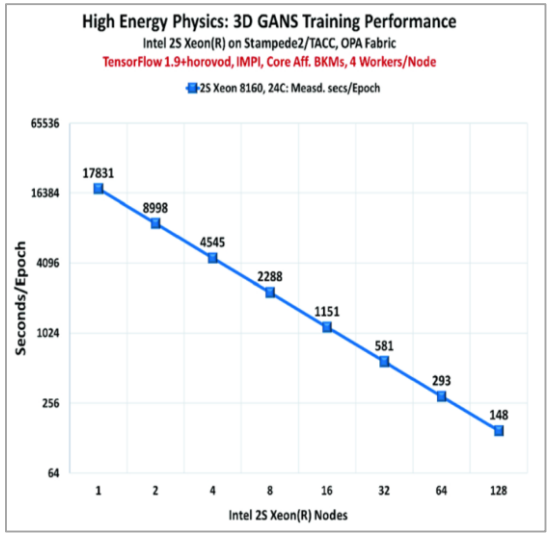


CERN

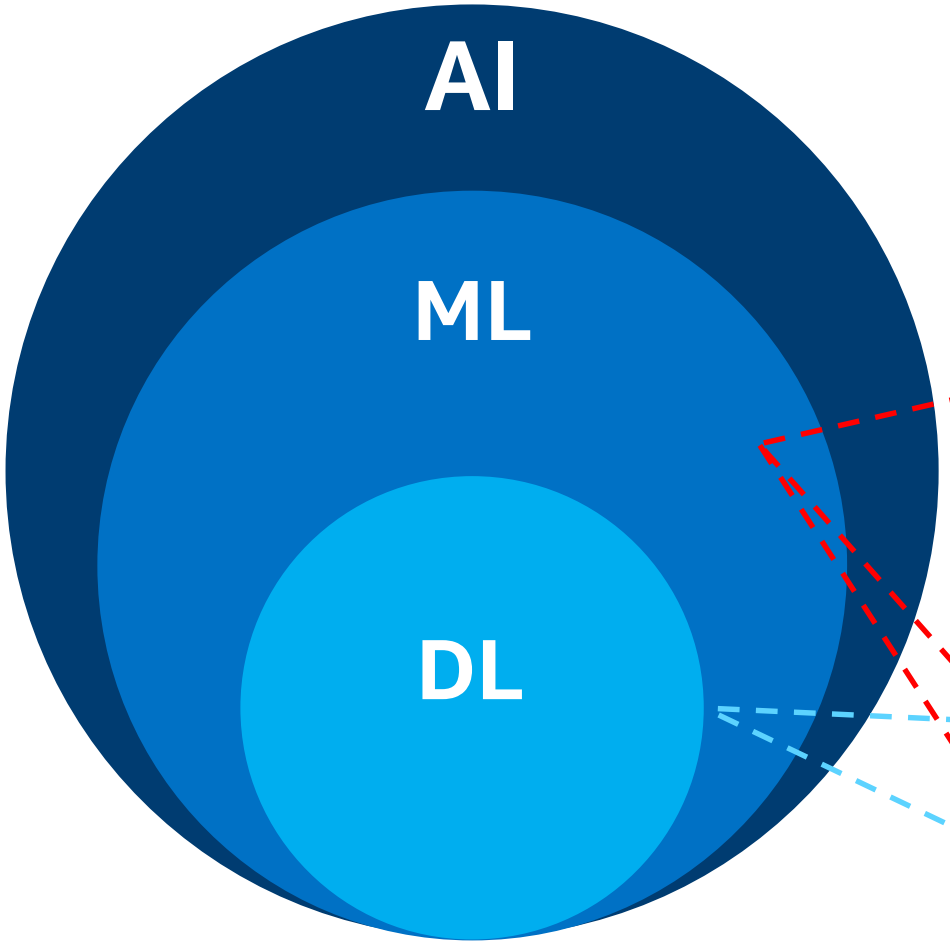
the European Organization for Nuclear Research, which operates the Large Hadron Collider (LHC), the world's largest particle accelerator

94% scaling efficiency up to 128 nodes, with a significant reduction in training time per epoch for 3D-GANs

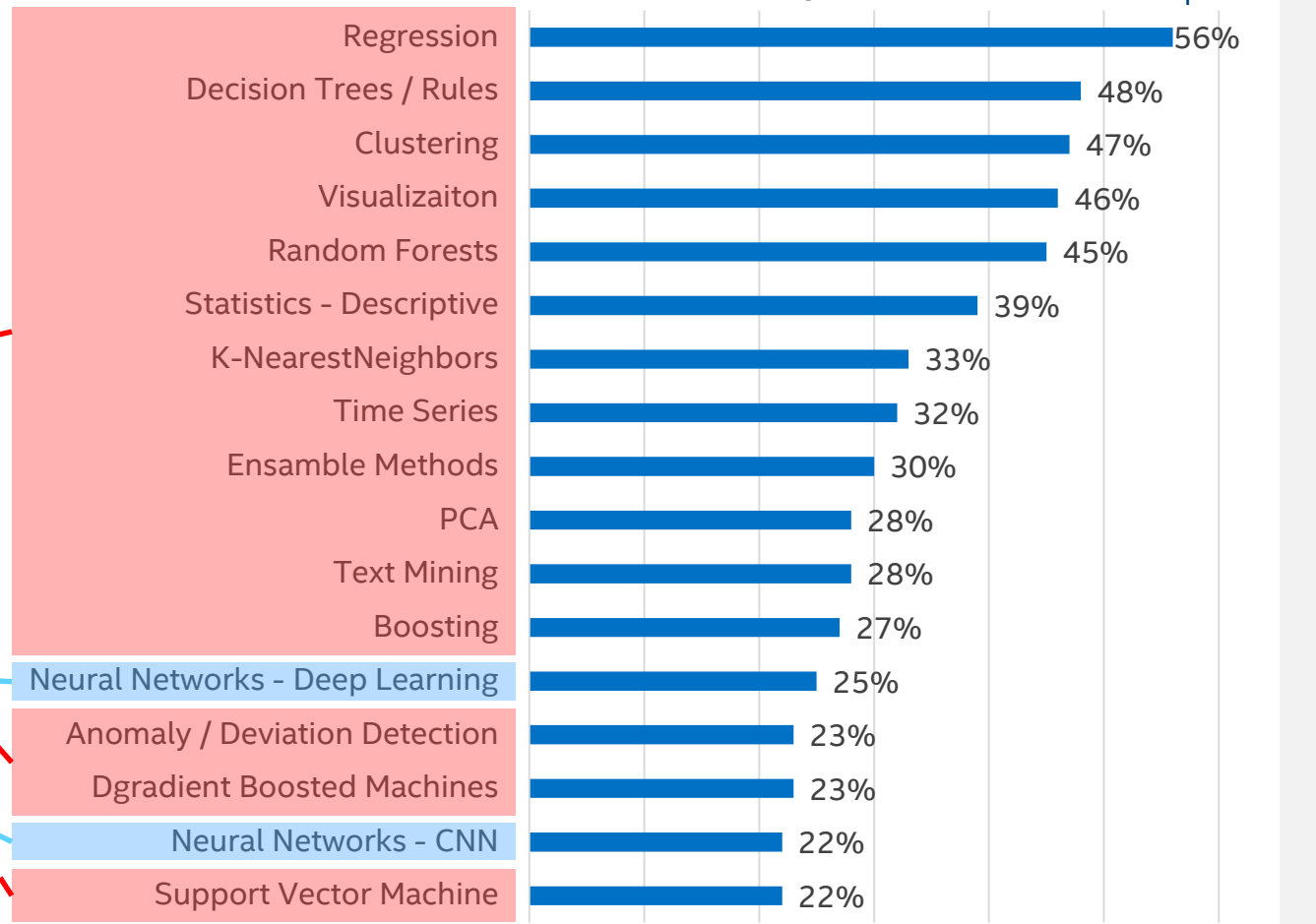




ML is still important



Top Data Science, Machine Learning Methods used in 2018/2019



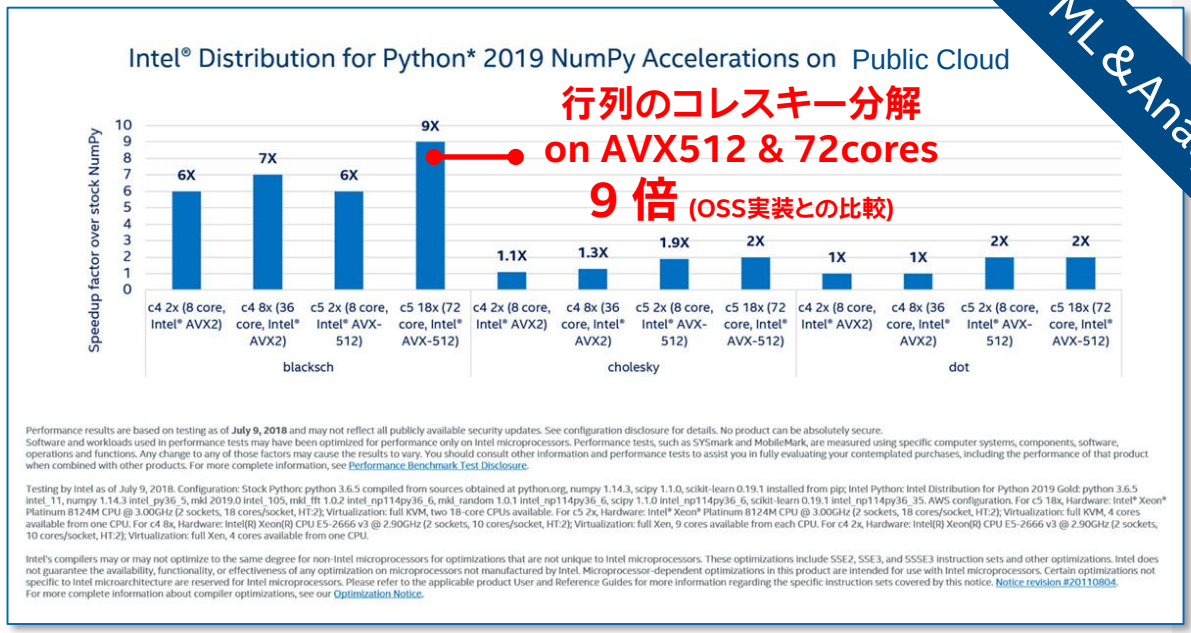
引用元：https://www.kdnuggets.com/2019/04/top-data-science-machine-learning-methods-2018-2019.html

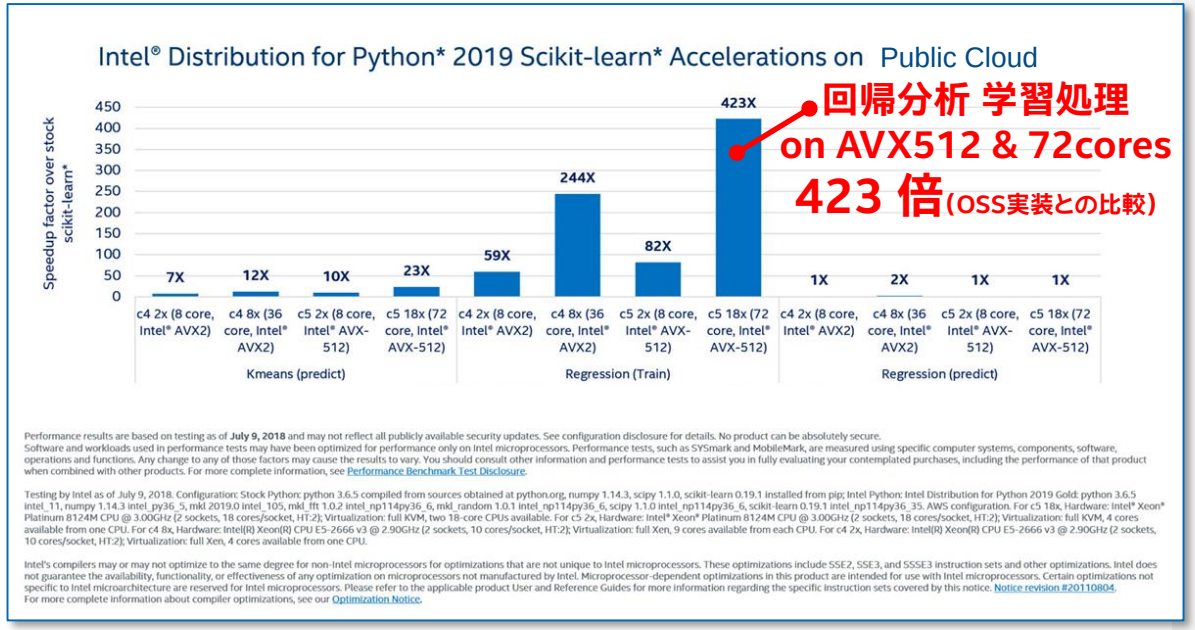
Intel® Distribution for Python*

Intel's implementation and optimization of Python and related libraries

- Numpy
- Pandas
- Scipy
- Scikit-learn
- XGBoost
- TensorFlow
- etc..

ML & Analytics





https://software.intel.com/en-us/distribution-for-python/benchmarks

Intel® AI Library & oneDAL

ML & Analytics

Math	ML & Analytics	DL	Collective Communicaiton
Intel® oneAPI Math Kernel Library (oneMKL)	Intel® oneAPI Data Analytics Library (oneDAL)	Intel® oneAPI Deep Neural Network Library (oneDNN)	Intel® oneAPI Collective Communication Library (oneCCL)
daal4py	scikit learn	APACHE Spark, databricks	Partner Solution
pip install daal4py	*pip install intel-scikit-learn*	*Able to install into Spark*	*http://www.intel.com/analytics*

https://www.oneapi.com/

New demand, New Technology

先進技術

Security

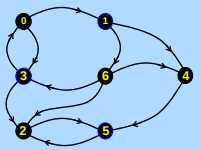
PPML

(**P**rivacy **P**reserving **M**achine **L**earning)

Machine learning technology with an emphasis on privacy protection

Data

Graph



Analysis of graph data or pattern detection using machine learning

Algorithm

SLIDE

(**S**ub-**L**Inear **D**eep **l**earning **E**ngine)

Collaboration with Rice University. Deep learning's training algorithms have been fundamentally redesigned to achieve higher learning performance on the CPU than on the GPU.

Intel Technology Blog on Graph Analysis

先進技術

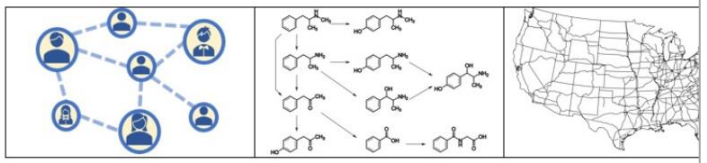
Measuring Graph Analytics Performance

The Diverse Landscape of Graph Analytics Requires a Comprehens
Benchmark

Henry Gabb Follow
Feb 8 · 6 min read

What Is Graph Analytics And Why Does It Matter?

A graph is a good way to represent a set of objects and the relations be
them (**Figure 1**). Graph analytics is the set of techniques to extract
information from connections between entities.

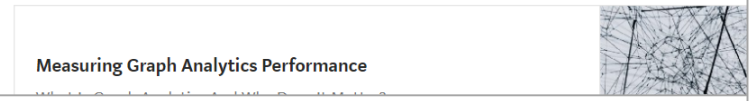


Adventures in Graph Analytics Benchmarking

It's Important to Use a Benchmark for Its Intended Purpose

Henry Gabb Follow
Feb 18 · 5 min read

With all the attention graph analytics is getting lately, it's increasingly important to measure its performance in a comprehensive, objective, and reproducible way. I covered this in a previous article, in which I recommended using an off-the-shelf benchmark like the GAP Benchmark Suite from the University of California, Berkeley. There are other graph benchmarks, of course, like LDBC Graphalytics, but they can't beat GAP f ease of use. There's significant overlap between GAP and Graphalytics, bu the latter is an industrial-strength benchmark that requires a special software configuration.

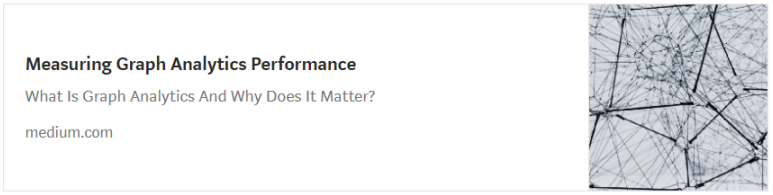


You Don't Have to Spend $800,000 to Compute PageRank

There's a Better Way to Do Large-Scale Graph Analytics

Henry Gabb Follow
Jul 8 · 3 min read

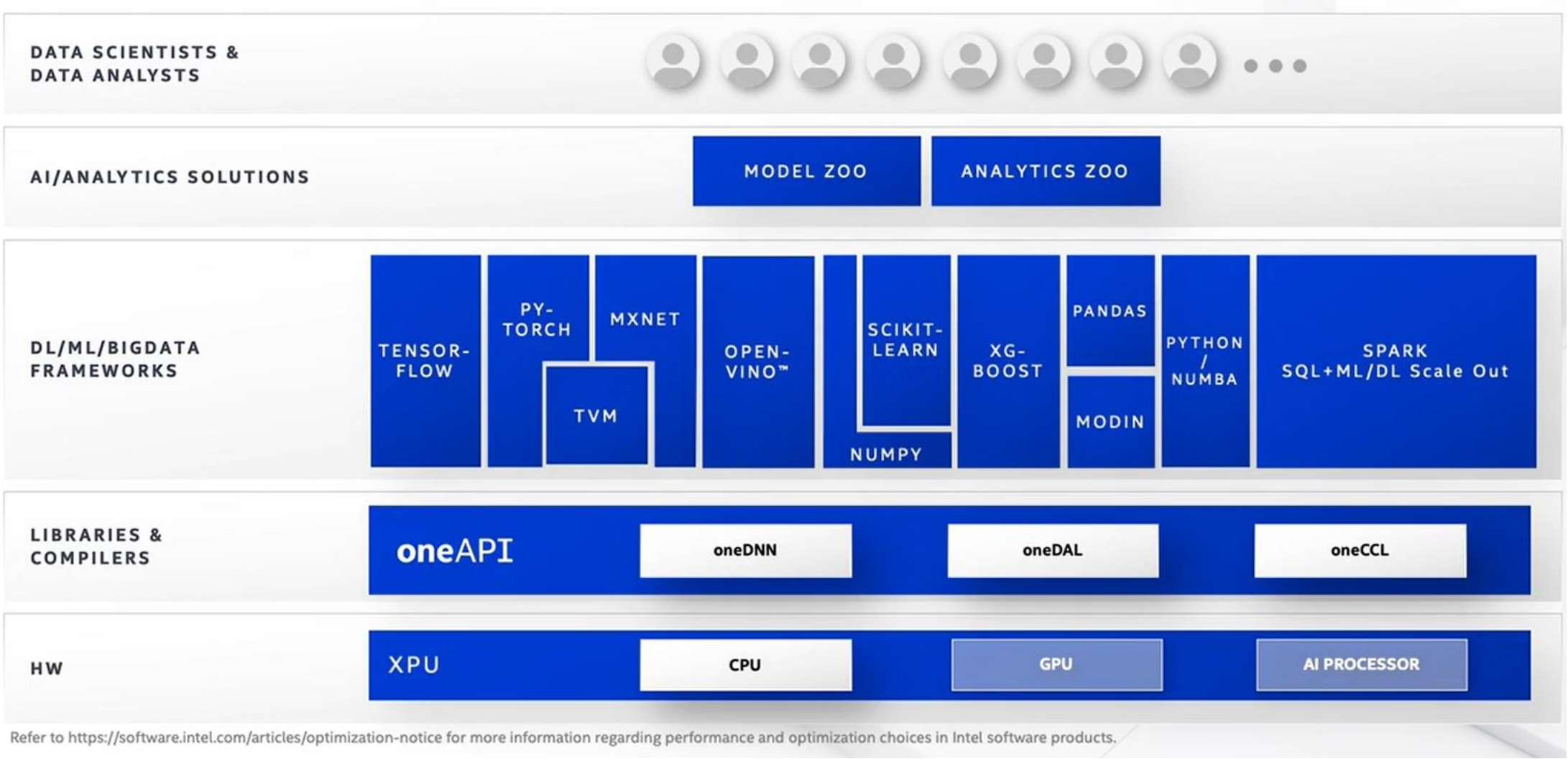
Benchmarking isn't my favorite topic, but I have a passing interest in graph analytics benchmarking:



I'll occasionally dissect benchmarks that I think are inaccurate or misleading:

https://medium.com/intel-analytics-software

AI Software Ecosystem on Intel



Refer to https://software.intel.com/articles/optimization-notice for more information regarding performance and optimization choices in Intel software products.

Accelerate Your AI Journey with Intel

Intel Xeon Scalable Processor: The only data center CPU with built-in AI acceleration

DISCOVERY *of possibilities & next steps*

DATA *setup, ingestion & cleaning*

DEVELOP *models using analytics/AI*

DEPLOY *into production & iterate*

ECOSYSTEM

INTEL® AI BUILDERS Over 100 vertical & horizontal ecosystem solutions

OPTIMIZED CLOUD Amazon Web Services, Baidu Cloud, Google Cloud Platform, Microsoft Azure & More

AI-OPTIMIZED CONFIGURATIONS intel select solution

SOFTWARE

DATA ANALYTICS Over 50 optimized software platforms

MACHINE LEARNING Intel Distribution for Python; Analytics Zoo

DEEP LEARNING TensorFlow, PyTorch, OpenVINO

HARDWARE

MOVE Ethernet, Barefoot, Silicon Photonics

STORE Intel Optane DC Persistent Memory, Intel Optane DC Solid State Drive, Intel 3D NAND SSD

PROCESS Intel Xeon, Intel Core i7, Intel Movidius; In development: Intel Xe, habana

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


intel