ARTIFICIAL INTELLIGENCE

June 2019
Public

NOTICES AND DISCLAIMERS

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration.

No product or component can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. For more complete information about performance and benchmark results, visit http://www.intel.com/benchmarks .

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/benchmarks .

Intel® Advanced Vector Extensions (Intel® AVX)* provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at http://www.intel.com/go/turbo.

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.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

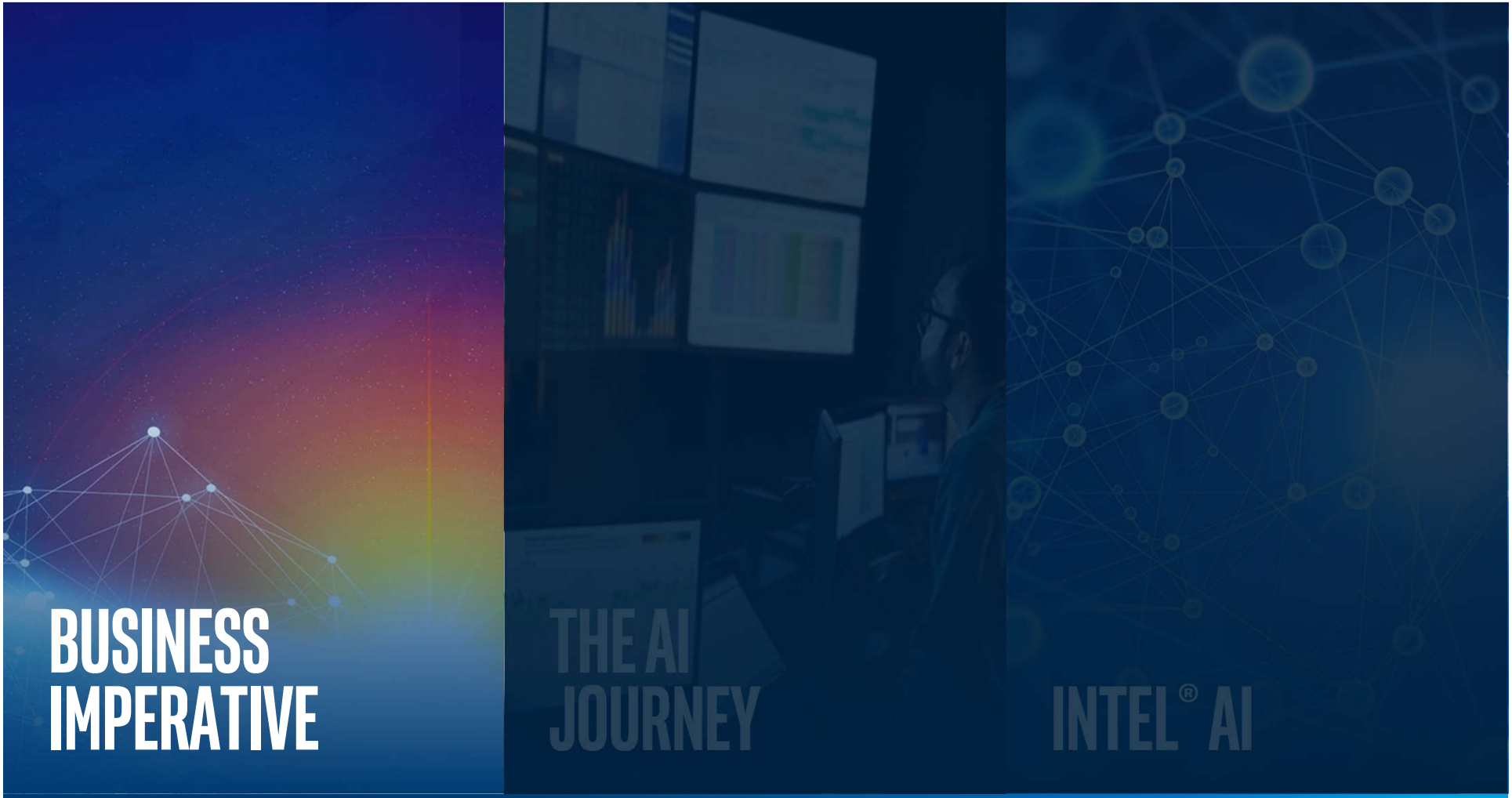
Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Arria, Celeron, Intel, the Intel logo, Intel Atom, Intel Core, Intel Nervana, Intel Optane, Intel Xeon, Iris, Movidius, OpenVINO, Pentium, Stratix and the Stratix logo and are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as property of others.

ABOUT THE SPEAKER

- Joined Intel since 2015 via acquisition. Currently Based in Taiwan Sales and Marketing Group as Solution Architect
- Responsible for solutions for acceleration technology
- FPGA support lead at Altera
- ASIC designer at NXP, Trex Enterprises
- Support AI activities since 2012



THE AI MANDATE

"AI technologies are evolving fast and growing increasingly **critical** to firms' ability to win, serve, and retain customers."

"...strategic technologies for 2019 with the potential to drive significant **disruption** and deliver **opportunity** over the next five years"

"...**70%** of CIOs will aggressively apply data and AI to IT operations, tools, and processes by 2021."

THE TIME TO BEGIN AI ADOPTION IS NOW

*Other names and brands may be claimed as the property of others
Source: https://www.forrester.com/report/The+Forrester+Tech+Tide+Artificial+Intelligence+For+Business+Insights+Q3+2018/-/E-RES143252
Source: https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2019
Source: https://www.idc.com/getdoc.jsp?containerId=prUS44420918

WHAT IS AI?

Regression

Classification

Clustering

Decision Trees

Data Generation

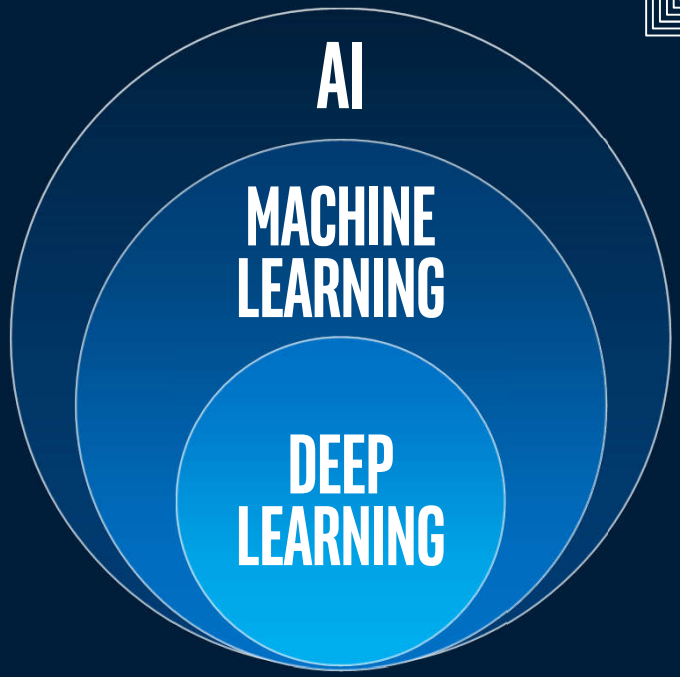
Image Processing

Speech Processing

Natural Language Processing

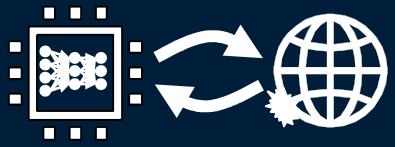
Recommender Systems

Adversarial Networks



SUPERVISED LEARNING

UNSUPERVISED LEARNING



REINFORCEMENT LEARNING

NO ONE SIZE FITS ALL APPROACH TO AI

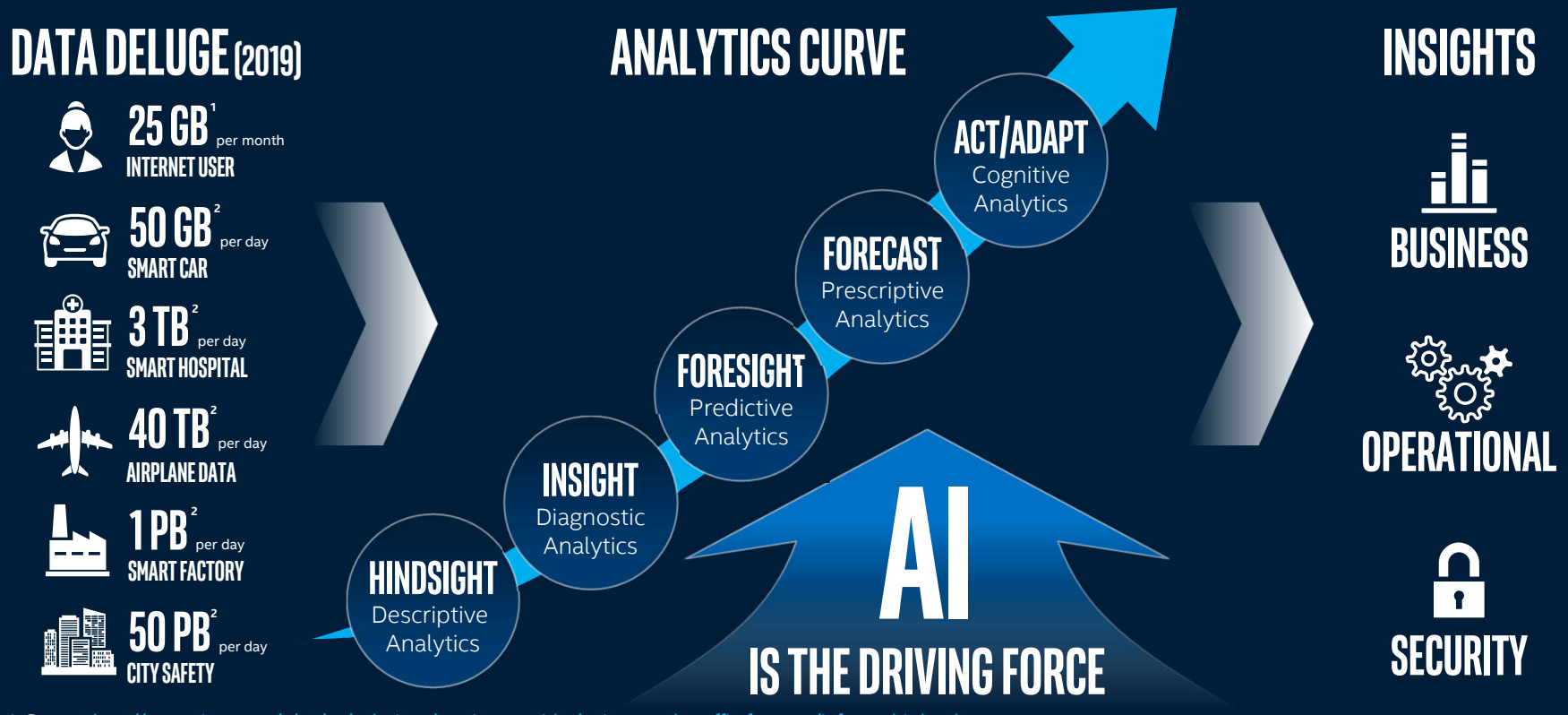
WHAT IS AI?

MANUFACTURING EXAMPLE

	QUESTION	METHOD	APPROACH
1	**How many parts should we manufacture?**	Historical supply & demand analysis	**Statistical Methods**
2	**What will our production yield be?**	Algorithm learns which variables correlate to yield	**Machine Learning** ***(Unsupervised)***
3	**Which parts have visual defects?**	Algorithm learns to identify defects in images	**Deep Learning** ***(Supervised)***
4	**Can my robotic arm adapt to conditions in real-time?**	Algorithm that acts and adapts based on feedback	**Deep Learning** ***(Reinforcement)***

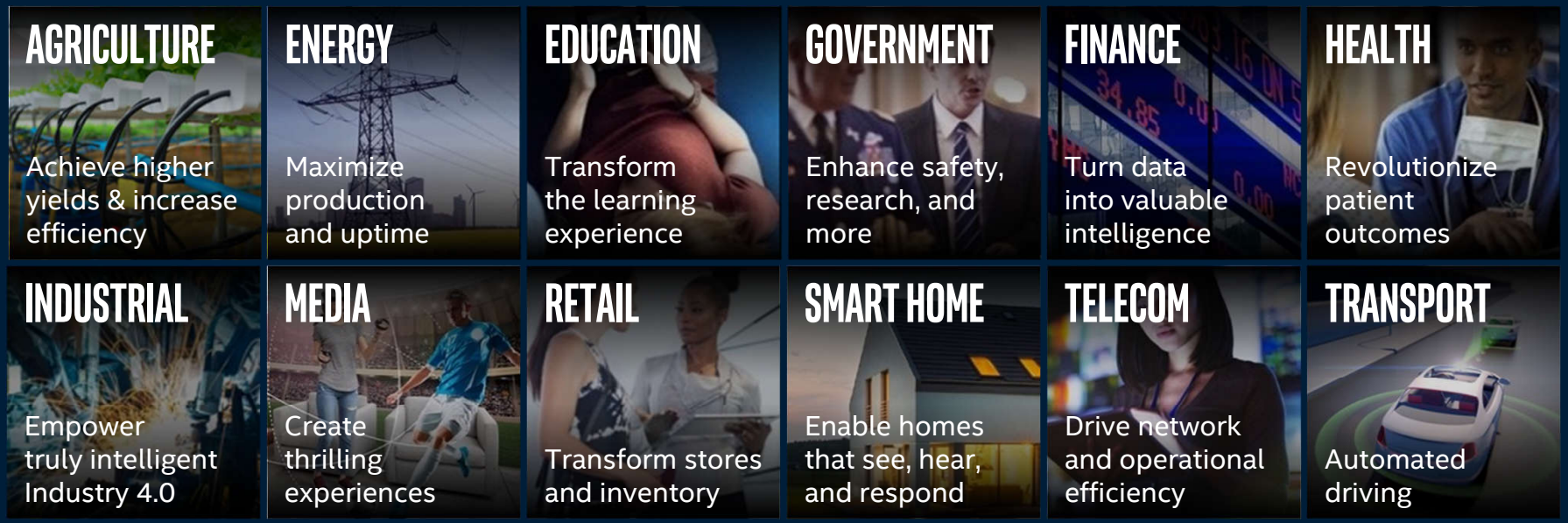
CHOOSE THE RIGHT AI APPROACH FOR YOUR CHALLENGE

WHY AI NOW?



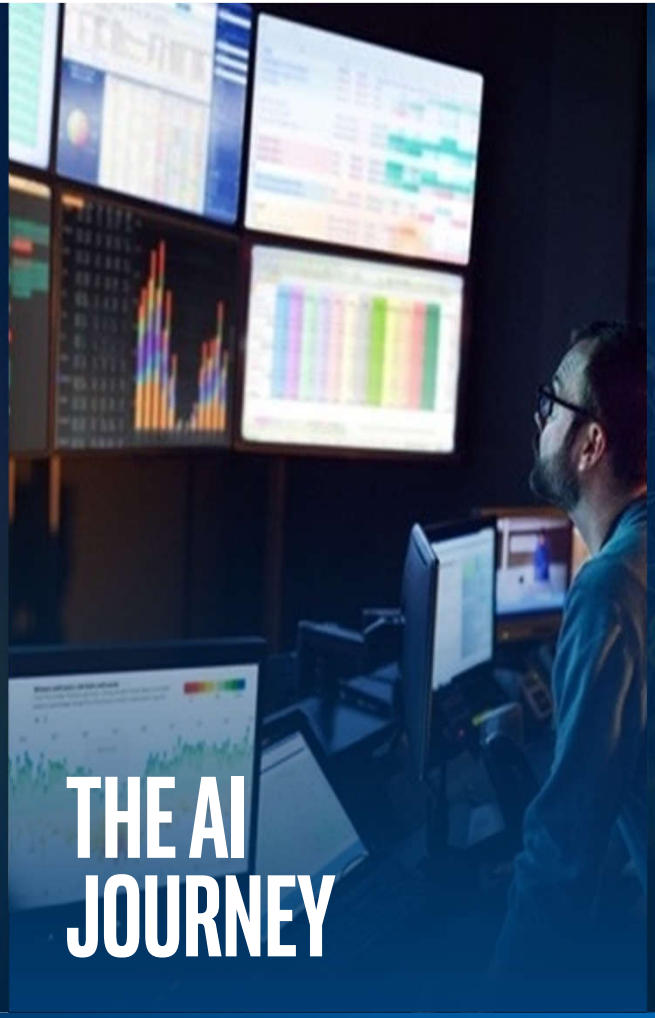
1. Source: http://www.cisco.com/c/en/us/solutions/service-provider/vni-network-traffic-forecast/infographic.html
2. Source: https://www.cisco.com/c/dam/m/en_us/service-provider/ciscoknowledgenetwork/files/547_11_10-15-DocumentsCisco_GCI_Deck_2014-2019_for_CKN__10NOV2015_.pdf

AI SOLUTIONS IN EVERY MARKET



OUR PARTNERS ARE DRIVING REAL-WORLD VALUE WITH INTEL AI

BUSINESS IMPERATIVE

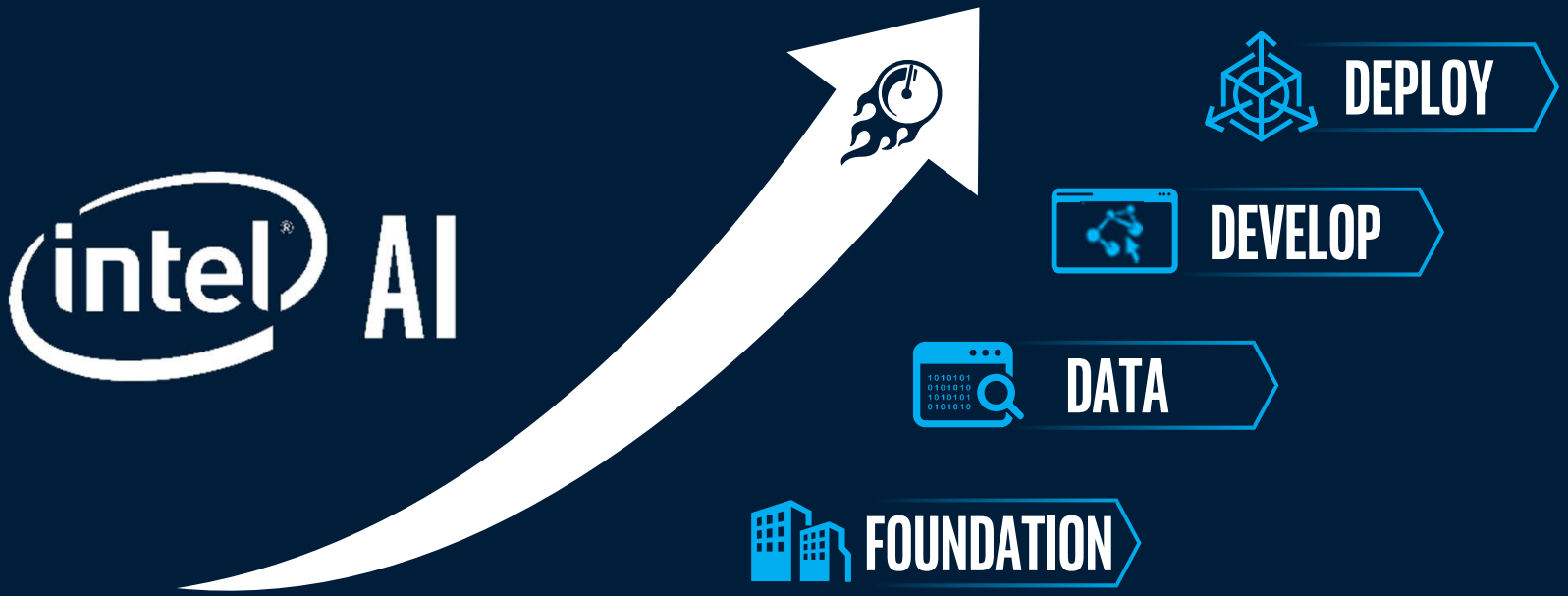


THE AI JOURNEY

INTEL® AI

THE AI JOURNEY

ACCELERATE WITH INTEL AI



INTEL® AI CASE STUDY

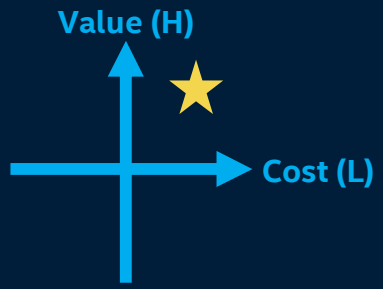
FOUNDATION

IDENTIFY

Identify prospects internally and using the 70+ AI solutions in Intel's portfolio then assess business value of each one

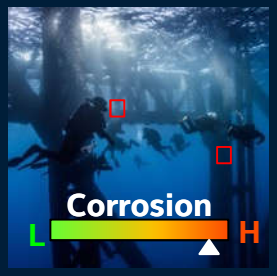
PRIORITIZE

Prioritize projects based on business value & cost to solve with Intel guidance; choose industrial defect detection via DL[1]



CONSIDER

Consider ethical, social, legal, security & other risks and mitigation plans with Intel advisors prior to kickoff



ORGANIZE

Organize internally to get buy-in, support new development philosophy & grow developer talent via Intel AI

intel AI DEVELOPER PROGRAM

INTEL® AI CASE STUDY

DATA

011010110110
110101101011
001011010100
011010110110
110101101011
001011010100
011010110110
110101101011
001011010100
011010110110
110101101011
001011010100

INGEST

Ingest streaming data from drones using a popular software tool among the many that run on the CPU

STORE

Store data in block storage (for high-performance) in a data lake with guidance from an Intel storage partner

PROCESS

Process data by performing cleanup & integration using popular software tools that run on the CPU

MANAGE

Manage data via a popular framework for distributed computation on CPU infrastructure

12 weeks

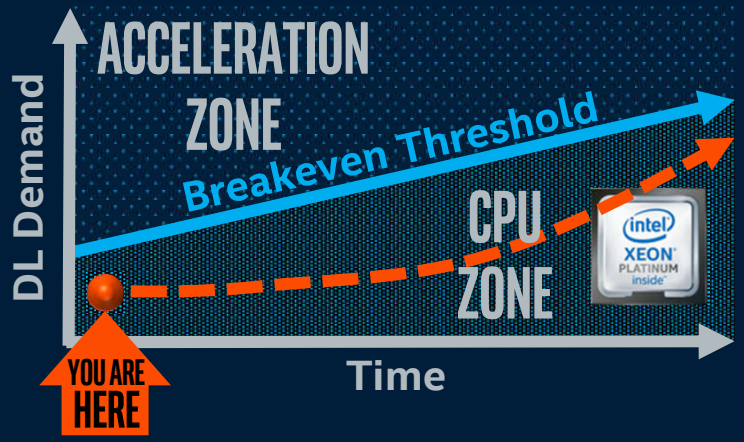
Source Data	Transmit Data	Ingest Data	Cleanup Data	Integrate Data	Stage Data
10%	*5%*	*5%*	*60%*	*10%*	*10%*

INTEL® AI CASE STUDY

DEVELOP

SETUP

Setup compute environment; DL training (~7% of journey) acceleration NOT worthwhile due to high setup time & cost



MODEL

Model development through training a deep neural network using an Intel-optimized DL framework

TEST

Test the deep learning model using a control data set to determine if accuracy meets requirements

DOCUMENT

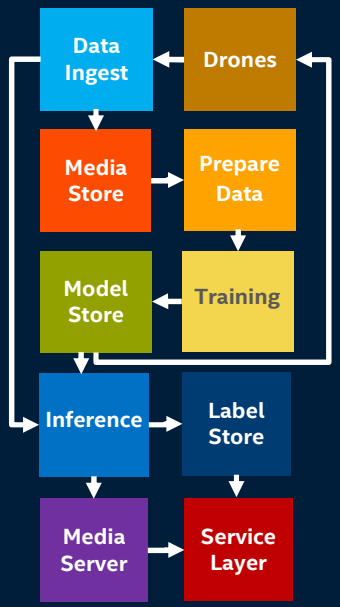
Document the code, process, and key learnings for future reference

12 weeks

Train *(Topology Experiments)*	Train *(Tune Hyper-parameters)*	Test Inference	Document Results
30%	*30%*	*20%*	*10%*

INTEL® AI CASE STUDY

DEPLOY



ARCHITECT

Architect AI deployment with Intel AI Builders

IMPLEMENT

Implement AI in production environment

SCALE

Scale to more sites & users as demand grows

ITERATE

Iterate on the models with new data over time

REMOTE DEVICES

Drone
Drone
Drone

10 Drones

Real-time object detection and data collection

Drone
Drone
Drone

Per Drone

1x Intel® Core™ processor
1x Intel® Movidius™ VPU

MEDIA SERVER

Media Store
Media Store
Media Store

110 Nodes

8 TB/day per cameraf
10 cameras
3x replication
1-year retention
4 mgmt nodes

Media Store
Media Store
Media Store

Per Node

1x 2S 61xx 20x 4TB SSD

MULTI-USE CLUSTER

Data Ingestion
Data Ingestion
Data Ingestion
Data Ingestion
Inference
Inference
Inference
Inference
Prepare Data
Prepare Data
Service Layer
Service Layer
Service Layer
Media Server
Media Server
Media Server

4 Nodes
One ingestion per day, one-day retention
4 Nodes
20M frames per day
2 Nodes
Infrequent op
3 Nodes
Simultaneous users
3 Nodes
10k clips stored

DATA STORE

Model Store
Model Store
Model Store
Model Store
Label Store
Label Store
Label Store
Label Store

Per Node

1x 2S 81xx
5x 4TB SSD

4 Nodes
1-year of history
4 Nodes
Labels for 20M frames /day

ADV. ANALYTICS

Training

16 Nodes
Intermittent use
1 training/month
for <10 hours

Training

Per Node

1x 2S 81xx
1x 4TB SSD

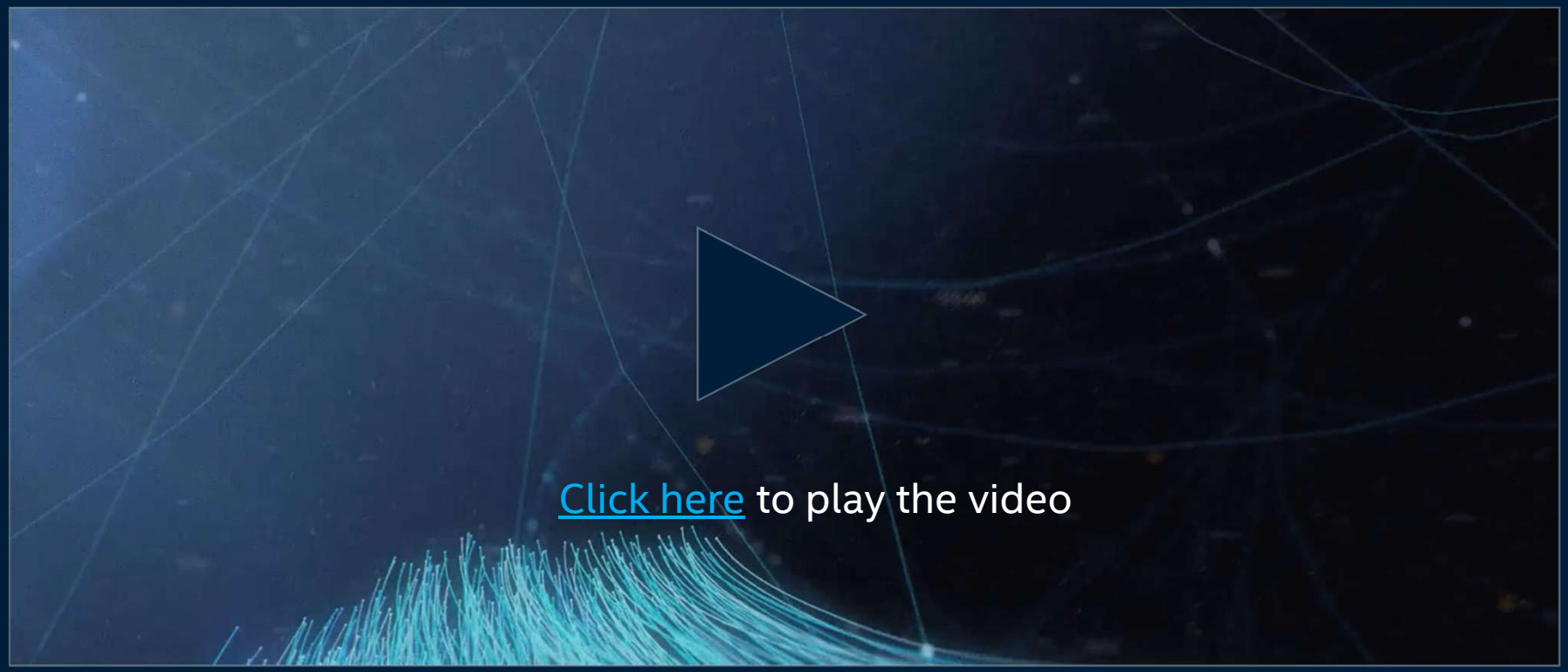
SOFTWARE

- *OpenVino™ Toolkit*
- *Intel® Movidius™ SDK*
- *TensorFlow**
- *Intel® MKL-DNN*

A NEW DAWN OF COMPUTING

intel AI

INTEL® AI TRAILER



INTEL AI STRATEGY

SEED & DRIVE THE ECOSYSTEM

- Seed emerging use cases
- Attract & develop top talent
- Pioneer leading-edge AI

SHAPE & WIN INDUSTRY OPEN SOFTWARE STACKS

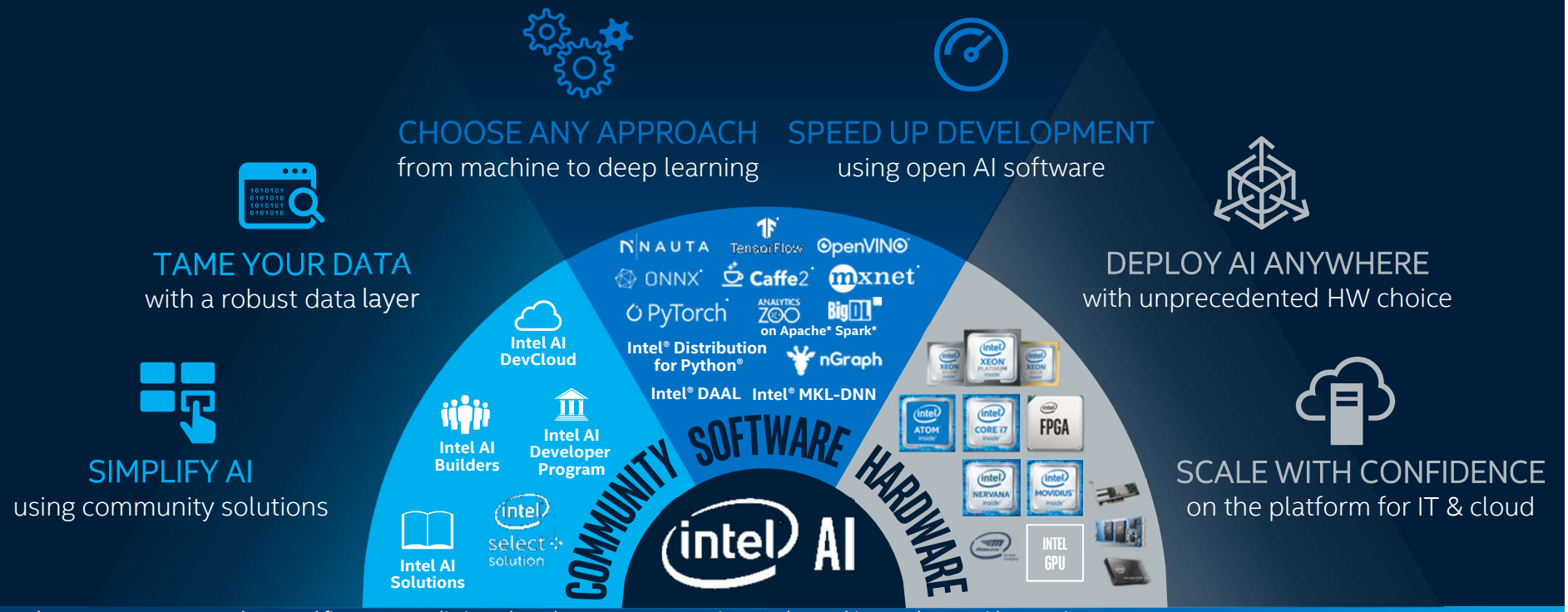
- Optimize customer software
- Build a unified API
- Evangelize to developers

DELIVER THE BEST AI PLATFORMS

- Extend the CPU
- Most complete portfolio
- Best integrated platforms

BREAKING BARRIERS BETWEEN AI THEORY AND REALITY

PARTNER WITH INTEL® TO ACCELERATE YOUR AI JOURNEY



SIMPLIFY AI

USING COMMUNITY SOLUTIONS

PARTNER

***Solve** your challenge using one of 100+ AI solutions in the Intel AI Builders program*

Visit: builders.intel.com/ai

OR

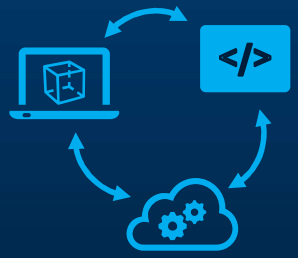
DEPLOY

***Deploy** AI-optimized systems including Intel® Select Solutions, Dell*, HPE*, Lenovo*, Inspur*, more*

Visit: builders.intel.com/ai

OR

DEVELOP



***Develop** your own AI solutions using Intel's FREE¥ developer courses, tools and cloud access*

Visit: software.intel.com/ai

¥Free = available to download/access at no cost to qualified developers who are enrolled in the program
*Other names and brands may be claimed as the property of others.

TAME YOUR DATA
WITH A ROBUST DATA LAYER

SOURCE(S)?
STRUCTURED?
VOLUME?
DURABILITY?
STREAMING?
LOCALITY?
GOVERNANCE?
OTHER?

INGEST

Tool for live streaming data ingestion from Internet of Things (IoT) sensors in endpoint devices

STORE

File, block, or object-based storage solution given cost, access, volume and perf requirements

PROCESS

Integration, cleaning, normalization and more transformations on batch and/or streaming data

ANALYZE

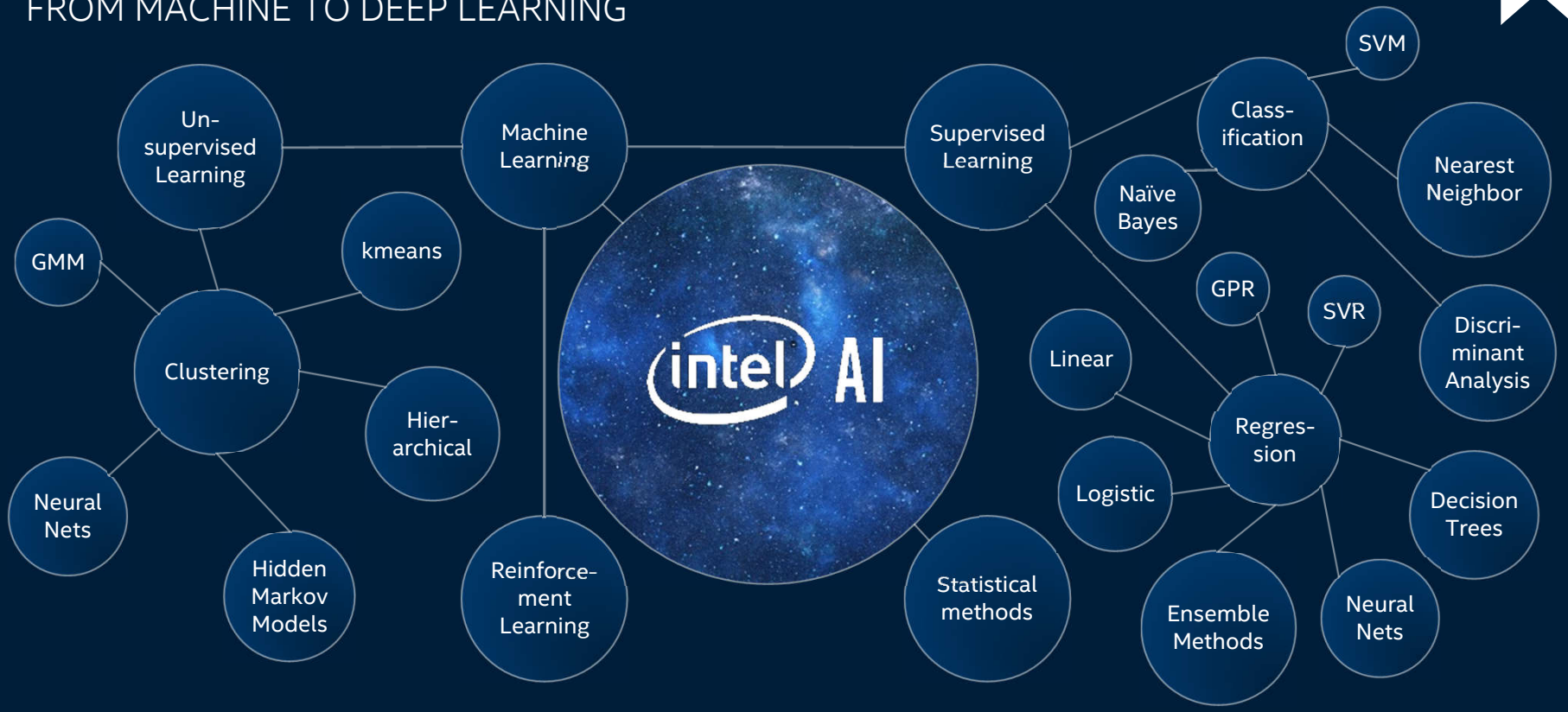
Applications in HPC, Big Data, HPDA, AI & more that have access to a common compute and data pool

Visit: *intel.com/content/www/us/en/analytics/tame-the-data-deluge.html*

CHOOSE ANY APPROACH

FROM MACHINE TO DEEP LEARNING

Visit: software.intel.com/ai/courses



SPEED UP DEVELOPMENT
USING OPEN AI SOFTWARE

Visit: www.intel.ai/technology

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 (DAAL)
High performance machine learning & data analytics library

Intel® Math Kernel Library for Deep Neural Networks (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)

[1] An open source version is available at: 01.org/openvinotoolkit *Other names and brands may be claimed as the property of others.
Developer personas show above represent the primary user base for each row, but are not mutually-exclusive
All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

SPEED UP DEVELOPMENT
USING OPEN AI SOFTWARE

Visit: www.intel.ai/technology

MACHINE LEARNING | DEEP LEARNING

TOOLKITS
App developers

ANALYTICS ZOO | OpenVINO | NAUTA

LIBRARIES
Data scientists

Python
- Scikit-learn
- Pandas
- NumPy

R
- Cart
- Random Forest
- e1071

Distributed
- MLLib (on Spark)
- Mahout

Intel-optimized Frameworks
TensorFlow* | Caffe2* | ONNX* | And more... | mxnet* | PyTorch* | BigDL

KERNEL/GRAPH
Library developers

Intel® Distribution for Python*

Intel® Data Analytics Acceleration Library (DAAL)

Intel® Math Kernel Library for Deep Neural Networks (MKL-DNN)

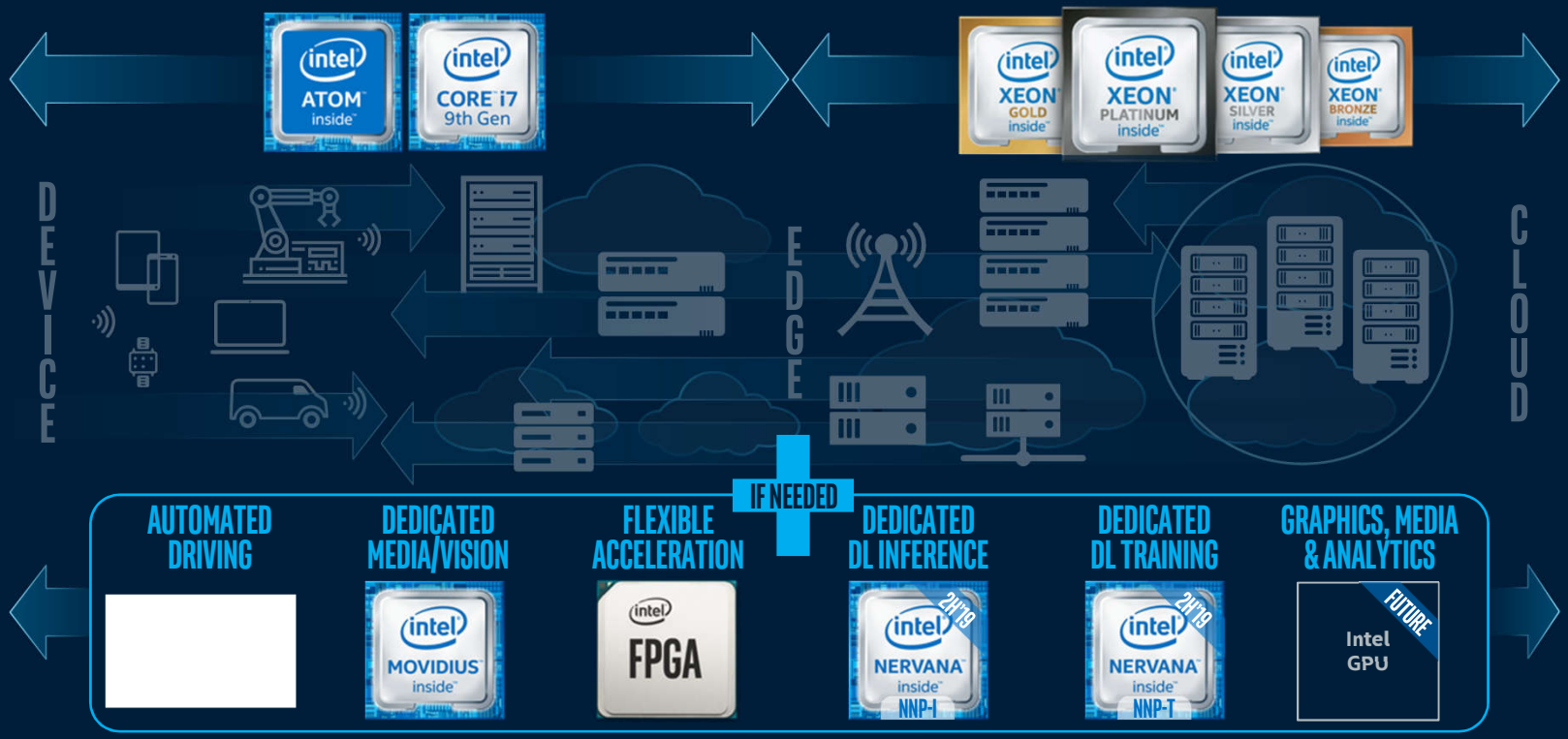
nGraph | plaidML

[1] An open source version is available at: 01.org/openvinotoolkit *Other names and brands may be claimed as the property of others.
Developer personas show above represent the primary user base for each row, but are not mutually-exclusive
All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

DEPLOY AI ANYWHERE

WITH UNPRECEDENTED HARDWARE CHOICE

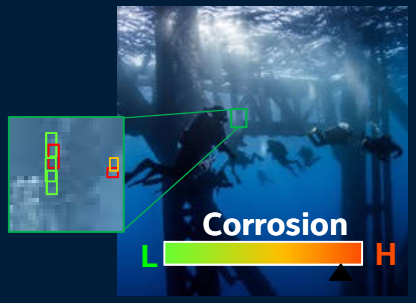
Visit: www.intel.ai/technology



SCALE WITH CONFIDENCE
ON THE PLATFORM FOR IT & CLOUD

Visit: builders.intel.com/ai

Model development is a challenge, but scaling to production is also daunting



Drones → Data Ingest → Media Store → Prepare Data → Training → Model Store → Inference → Label Store → Solution Layer

Inference → Media Server → Service Layer

REMOTE STORE

Drone
Drone
Drone

10 Drones
Real-time object detection and data collection

Drone
Drone
Drone

Per Drone

1x Intel® Core™ processor
1x Intel® Movidius™ VPU

MEDIA STORE

Media Store
Media Store
Media Store

110 Nodes
8 TB/day per camera
10 cameras
3x replication
1-year video retention
4 mgmt nodes

Media Store
Media Store
Media Store

Per Node 1x 2S 61xx 20x 4TB SSD

MULTI-USE CLUSTER

Data Ingestion
Data Ingestion
Data Ingestion
Data Ingestion

4 Nodes One ingestion per day, one-day retention

Inference
Inference
Inference
Inference

4 Nodes 20M frames per day

Prepare Data
Prepare Data

2 Nodes Infrequent op

Service Layer
Service Layer
Service Layer

3 Nodes Simultaneous users

Media Server
Media Server
Media Server

3 Nodes 10k clips stored

DATA STORE

Model Store
Model Store
Model Store
Model Store

4 Nodes 1-year of history

Label Store
Label Store
Label Store
Label Store

4 Nodes Labels for 20M frames /day

1x 2S 81xx
5x 4TB SSD

SOFTWARE

- OpenVino™ Toolkit
- Intel® Movidius™ SDK
- TensorFlow*
- Intel® MKL-DNN

ADV. ANALYTICS

Training

16 Nodes
Intermittent use
1 training/month
for <10 hours

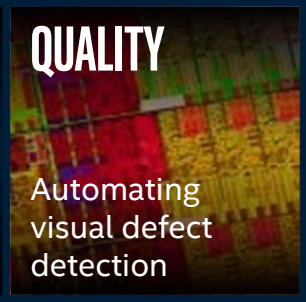
Training

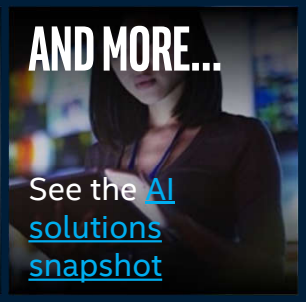
Per Node

1x 2S 81xx
1x 4TB SSD

*Other names and brands may be claimed as the property of others

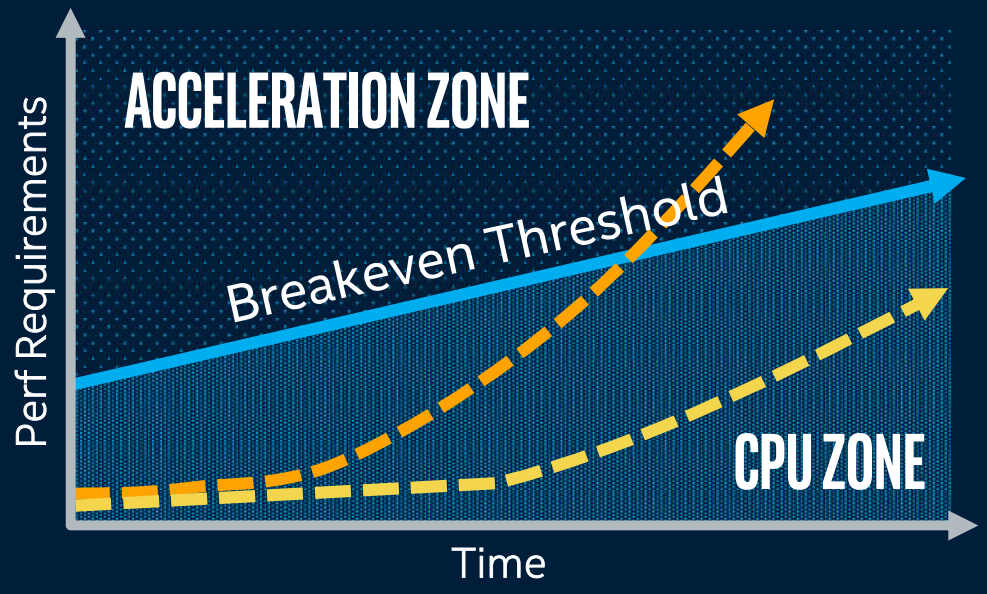
AI INSIDE INTEL





INTEL® IS INFUSING AI INTO EVERYTHING WE DO

BUSTING THE DEEP LEARNING MYTH



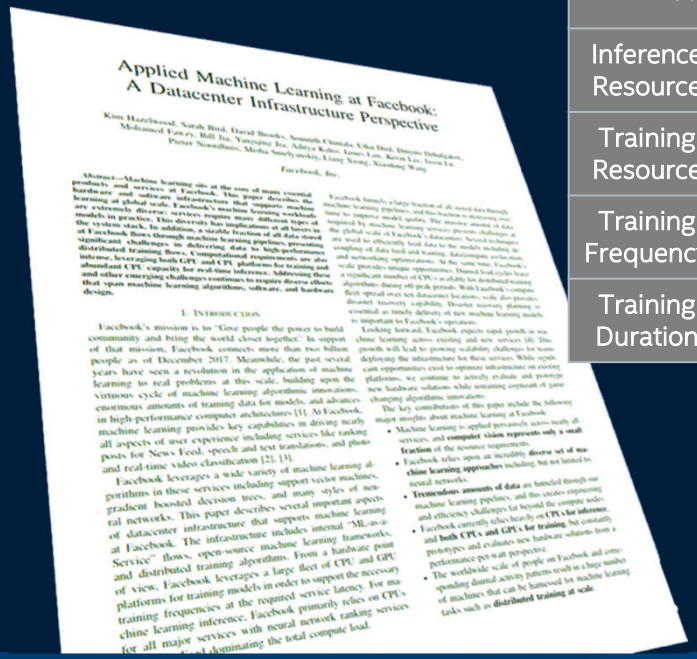
"A GPU is required for deep learning..." **FALSE**

- Most enterprises (---) use CPU for machine & deep learning needs
- Some early adopters (---) may reach a deep learning tipping point when acceleration is needed[1]

1"Most" of enterprise customers based on survey of Intel direct engagements and internal market segment analysis

DEEP LEARNING IN PRACTICE

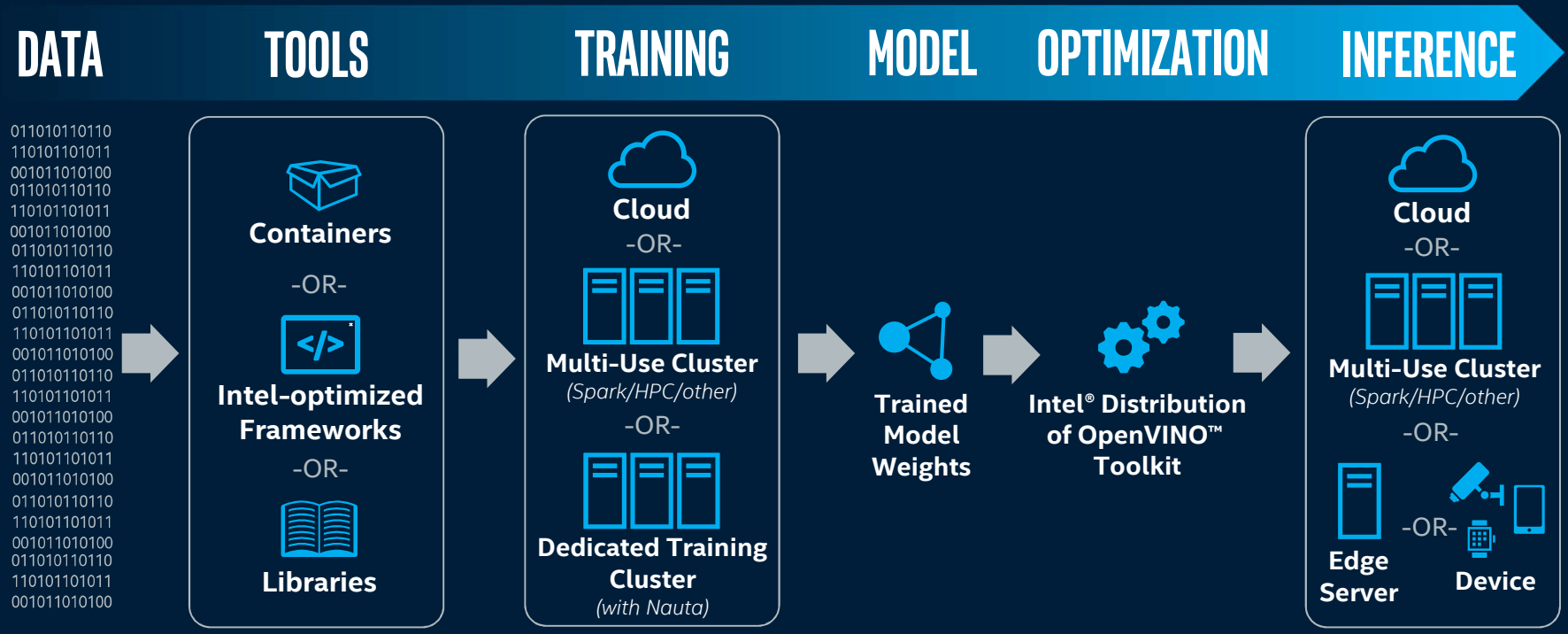
Source Paper:
research.fb.com/wpcontent/uploads/2017/12/hpca-2018-facebook.pdf


Applied Machine Learning at Facebook:
A Datacenter Infrastructure Perspective

Services	Ranking Algorithm	Photo Tagging	Photo Text Generation	Search	Language Translation	Spam Flagging	Speech
Model(s)	MLP	SVM,CNN	CNN	MLP	RNN	GBDT	RNN
Inference Resource	CPU	CPU	CPU	CPU	CPU	CPU	CPU
Training Resource	CPU	GPU & CPU	GPU	Depends	GPU	CPU	GPU
Training Frequency	Daily	Every N photos	Multi-Monthly	Hourly	Weekly	Sub-Daily	Weekly
Training Duration	Many Hours	Few Seconds	Many Hours	Few Hours	Days	Few Hours	Many Hours

LARGE CLOUD USERS EMPLOY CPU EXTENSIVELY FOR DEEP LEARNING

DEEP LEARNING DEPLOYED



END-TO-END DEEP LEARNING ON INTEL

2ND GENERATION INTEL® XEON® SCALABLE PROCESSOR

FORMERLY KNOWN AS CASCADE LAKE

Drop-in compatible CPU on Intel® Xeon® Scalable platform

TCO/FLEXIBILITY

Begin your AI journey efficiently, now with even more agility…

- ✓ IMT – Intel® Infrastructure Management Technologies
- ✓ ADQ – Application Device Queues
- ✓ SST – Intel® Speed Select Technology

PERFORMANCE

Built-in Acceleration with Intel® Deep Learning Boost…

Deep Learning throughput![1]

Throughput (img/s)

SECURITY

Hardware-Enhanced Security…

- ✓ Intel® Security Essentials
- ✓ Intel® SecL: Intel® Security Libraries for Data Center
- ✓ TDT – Intel® Threat Detection Technology

1 Based on Intel internal testing: 1X,5.7x,14x and 30x performance improvement based on Intel® Optimization for Café ResNet-50 inference throughput performance on Intel® Xeon® Scalable Processor. See Configuration Details 3
Performance results are based on testing as of 7/11/2017(1x) ,11/8/2018 (5.7x), 2/20/2019 (14x) and 2/26/2019 (30x) and may not reflect all publically available security updates. No product can be absolutely secure. See configuration disclosure for details. ,
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. 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

HARDWARE

MOVE FASTER

INTEL® SILICON PHOTONICS

INTEL® ETHERNET

INTEL® OMNI-PATH FABRIC

STORE MORE

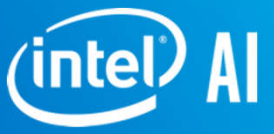
PROCESS EVERYTHING

CPU

AI ACCELERATORS

FPGA, GPU

POWERING THE FUTURE OF COMPUTE AND COMMUNICATIONS

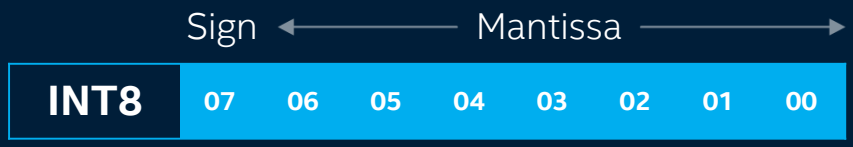


APPENDIX A – PRODUCT BRIEFS

ANIMATED

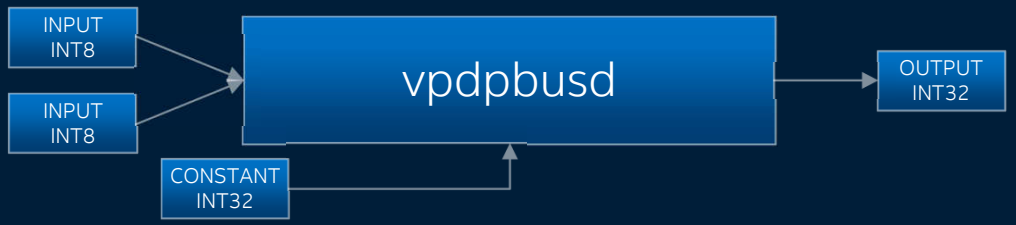
INTEL® DEEP LEARNING BOOST (DL BOOST)

FEATURING VECTOR NEURAL NETWORK INSTRUCTIONS (VNNI)



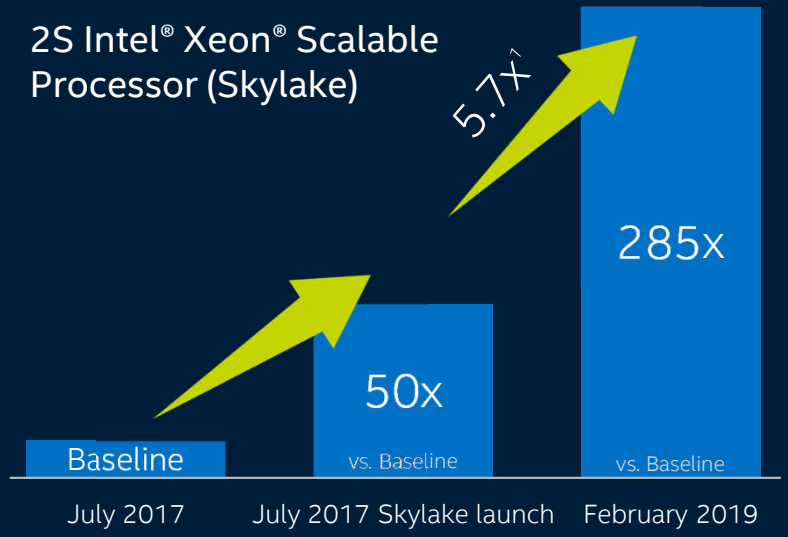
Current AVX-512 instructions to perform INT8 convolutions: **vpmaddubsw, vpmaddwd, vpaddd**

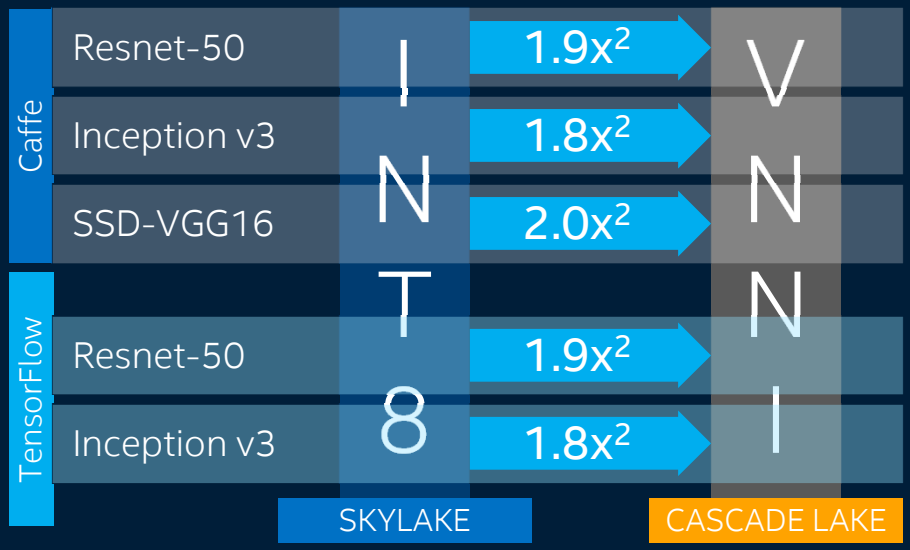
NEW *Future* AVX-512 (VNNI) instruction to accelerate INT8 convolutions: **vpdpbusd**



DEEP LEARNING PERFORMANCE ON CPU

HARDWARE + SOFTWARE IMPROVEMENTS FOR INTEL® XEON® PROCESSORS





1 5.7x inference throughput improvement with Intel® Optimizations for Caffe ResNet-50 on Intel® Xeon® Platinum 8180 Processor in Feb 2019 compared to performance at launch in July 2017. See configuration details on Config 1

Performance results are based on testing as of dates shown in configuration and may not reflect all publicly available security updates.

28/24/2018) 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. No product can be absolutely secure. See configuration disclosure for details. 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. 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

INTEL® NERVANA™ NEURAL NETWORK PROCESSORS (NNP)¥

NNP-T

DEDICATED
DL TRAINING

Fastest time-to-train with high bandwidth AI server connections for the most persistent, intense usage

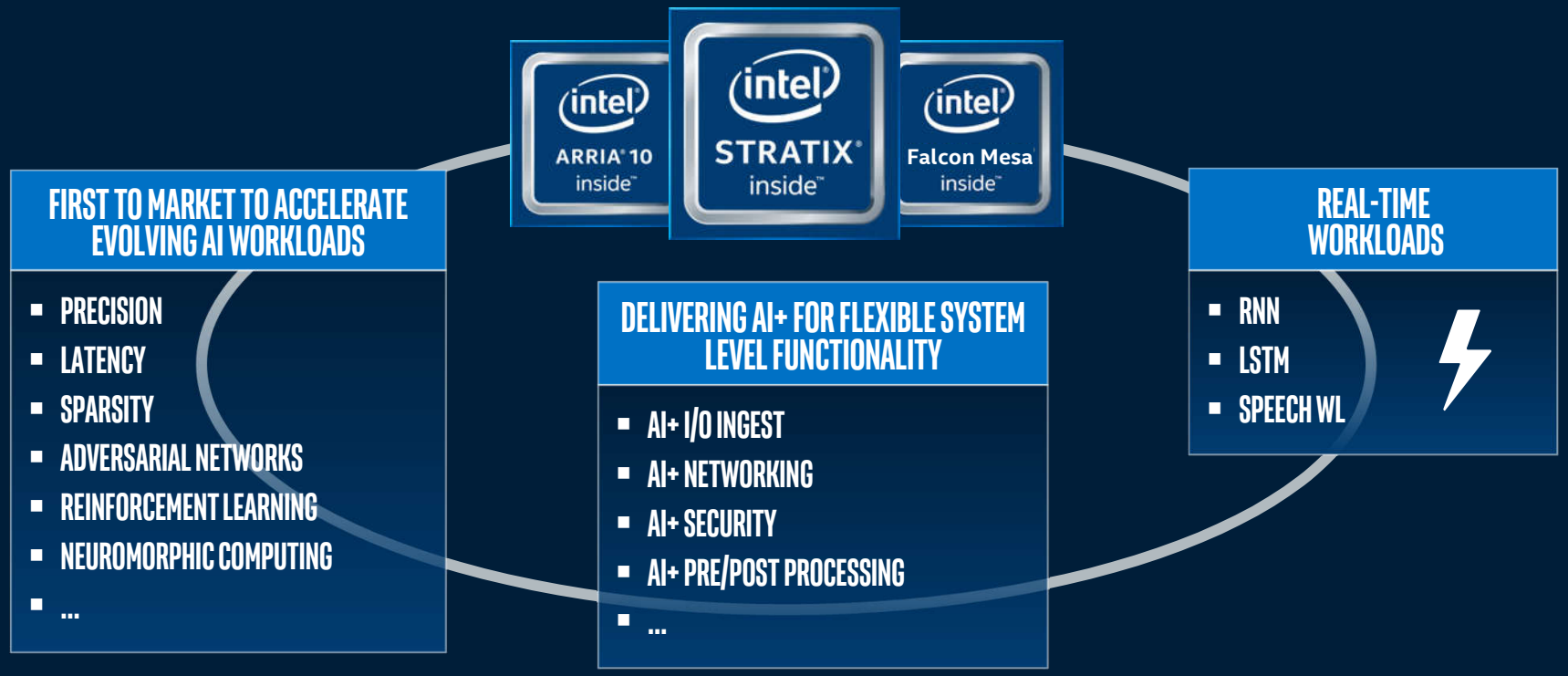
NNP-I

DEDICATED
DL INFERENCE

Highly-efficient multi-model inferencing for cloud, data center and intense appliances

¥ The Intel® Nervana™ Neural Network Processor is a future product that is not broadly available today
All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

INTEL® FPGA FOR AI



ENABLING REAL-TIME AI IN A WIDE RANGE OF EMBEDDED, EDGE AND CLOUD APPS

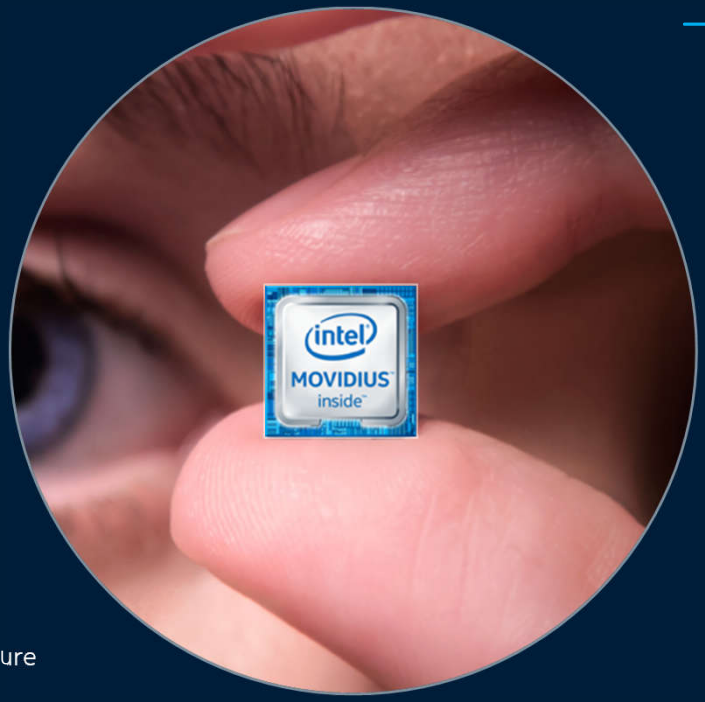
All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

INTEL® MOVIDIUS™ VISION PROCESSING UNIT (VPU)

SERVICE ROBOTS
- Navigation
- 3D Vol. mapping
- Multimodal sensing

SURVEILLANCE
- Detection/classification
- Identification
- Multi-nodal systems
- Multimodal sensing
- Video, image capture

WEARABLES
- Detection, tracking
- Recognition
- Video, image, session capture



DRONES
- Sense and avoid
- GPS denied hovering
- Pixel labeling
- Video, image capture

AR-VR HMD
- 6DOF pose, position, mapping
- Gaze, eye tracking
- Gesture tracking, recognition
- See-through camera

SMART HOME
- Detection, tracking
- Perimeter, presence monitoring
- Recognition, classification
- Multi-nodal systems
- Multimodal sensing
- Video, image capture

POWER-EFFICIENT IMAGE PROCESSING, COMPUTER VISION & DEEP LEARNING FOR DEVICES

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

INTEL® NEURAL COMPUTE STICK 2

USB STICK FORM FACTOR
for neural network acceleration

REAL-TIME ON-DEVICE INFERENCE
no cloud connectivity required

NO ADDITIONAL PERIPHERALS
needed to start deploying solutions

ACCELERATE DEVELOPMENT
with Intel® Distribution of OpenVINO™ toolkit

INDUSTRY LEADING PERFORMANCE
with Intel® Movidius™ Myriad™ X VPU

HIGHER PERFORMANCE
on deep neural networks compared to Myriad™ 2 VPU

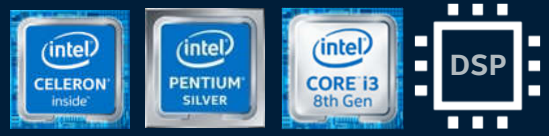
[1]Testing by Intel as of October 12th, 2018
Deep Learning Workload Configuration. Comparing Intel® Movidius™ Neural Compute Stick based on Intel® Movidius™ Myriad™ 2 VPU vs. Intel® Neural Compute Stick 2 based on Intel® Movidius™ Myriad™ X VPU with Asynchronous Plug-in enabled for (2xNCE engines). As measured by images per second across GoogleNetV1. Base System Configuration: Intel® Core™ I7-8700K 95W TDP (6C12T at 3.7GHz base freq and 4.7GHz max turbo freq), Graphics: Intel® UHD Graphics 630 Total Memory 65830088 kB Storage: INTEL SSDSC2BB24 (240GB), Ubuntu 16.04.5 Linux-4.15.0-36-generic-x86_64-with-Ubuntu-16.04-xenial, deeplearning_deploymenttoolkit_2018.0.14348.0, API version 1.2, Build 14348, myriadPlugin, FP16, Batch Size = 1

INTEL® GAUSSIAN NEURAL ACCELERATOR (GNA)

AMPLE THROUGHPUT
For speech, language, and other sensing inference

LOW POWER
<100 mW power consumption for always-on applications

FLEXIBILITY
Gaussian mixture model (GMM) and neural network inference support



TRY IT TODAY!

Intel® Speech Enabling Developer Kit

https://software.intel.com/en-us/iot/speech-enabling-dev-kit

Learn more: https://sigport.org/sites/default/files/docs/PosterFinal.pdf

STREAMING CO-PROCESSOR FOR LOW-POWER AUDIO INFERENCE & MORE

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

INTEL® INTEGRATED PROCESSOR GRAPHICS

UBIQUITY/SCALABILITY

- Shipped in 1 billion+ Intel® SoCs
- Broad choice of performance/power across Intel Atom®, Intel® Core™, and Intel® Xeon® processors

MEDIA LEADERSHIP

- Intel® Quick Sync Video – fixed-function media blocks to improve power and performance
- Intel® Media SDK – API that provides access to hardware-accelerated codecs

HARDWARE INTEGRATION

POWERFUL, FLEXIBLE ARCHITECTURE

- Rich data type support for 32bitFP, 16bitFP, 32bitInteger, 16bitInteger with SIMD multiply-accumulate instructions

MEMORY ARCHITECTURE

- Shared memory architecture on die between CPU and GPU to enable lower latency and power

SOFTWARE SUPPORT

MacOS (CoreML and MPS[1])
Windows O/S (WinML)
OpenVINO™ Toolkit (Win, Linux)
clDNN

BUILT-IN DEEP LEARNING INFERENCE ACCELERATION

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

CONNECTIVITY

INTEL® SILICON PHOTONICS
Connects memory and compute, integrating connectivity technologies onto a single die for affordable, scalable solutions

COMING SOON
SMARTNIC (CASCADE GLACIER)
Enables optimized performance for Intel® Xeon® processor-based systems

INTEL® OMNI-PATH ARCHITECTURE
Provides low-latency interconnect to scale to hundreds of thousands of nodes without losing performance or reliability

HIGH-SPEED CONNECTIVITY FOR MASSIVELY PARALLEL AND DISTRIBUTED AI

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

MEMORY AND STORAGE

INTEL® RACK SCALE DESIGN (INTEL® RSD)
Pools huge resources of on-demand compute, storage, and connectivity

INTEL® OPTANE™ TECHNOLOGY
Helps you affordably expand a large pool of memory closer to the CPU, so you can train on a much larger data set

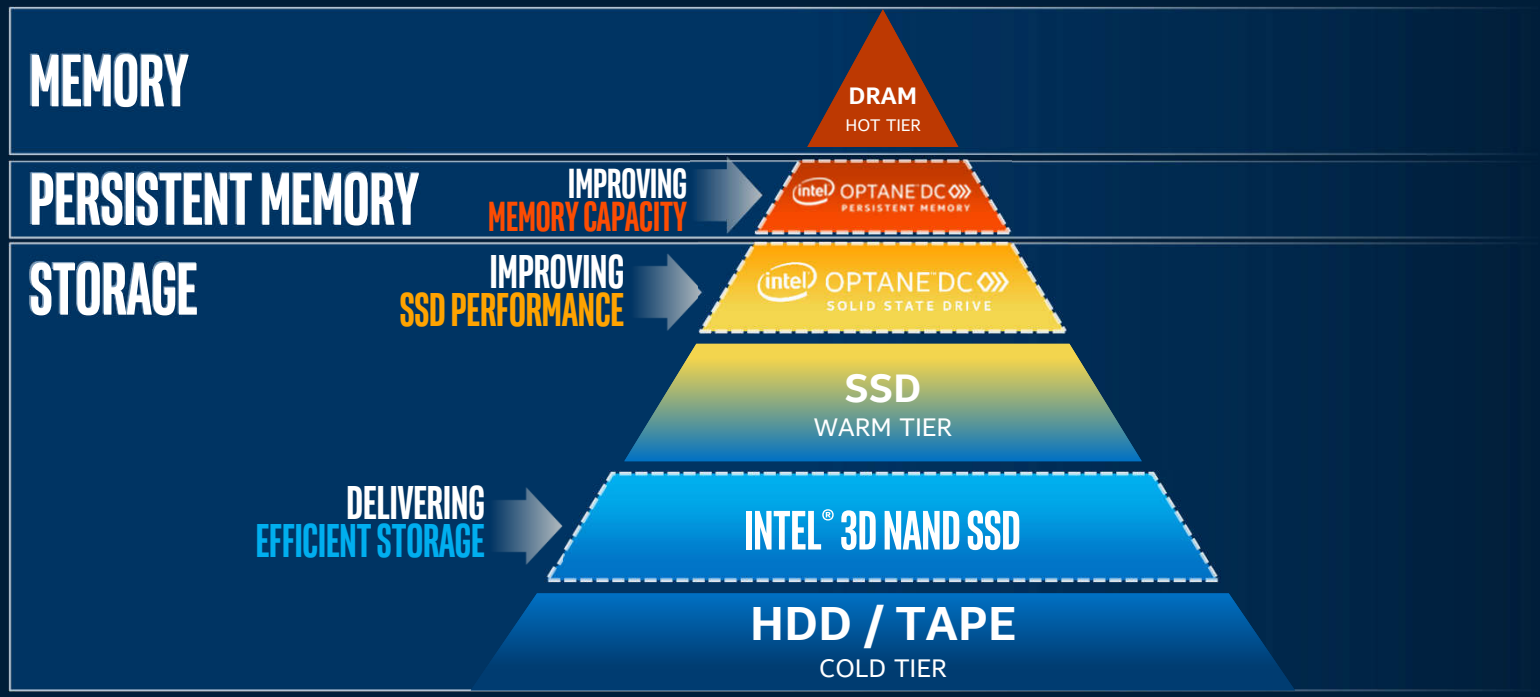
HIGH-BANDWIDTH MEMORY (HBM)
Specialized form of stacked DRAM integrated with processing units to increase speed while reducing latency, power, and size

OVERCOMING BOTTLENECKS IN DATA FLOW

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

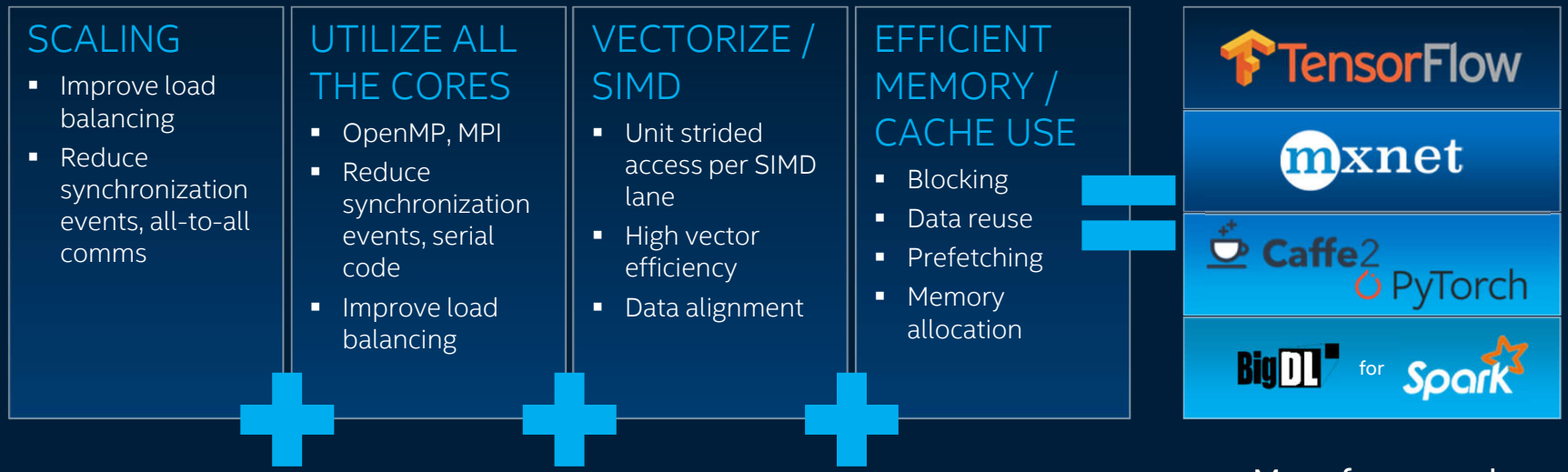
ANIMATED

RE-ARCHITECTING THE MEMORY / STORAGE HIERARCHY



A NEW CLASS OF MEMORY & STORAGE IS BORN

DEEP LEARNING FRAMEWORK (OPTIMIZATIONS BY INTEL®)



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

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

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.

AI ON APACHE Spark™

BigDL

HIGH-PERFORMANCE
DEEP LEARNING FRAMEWORK
FOR APACHE SPARK

software.intel.com/bigdl

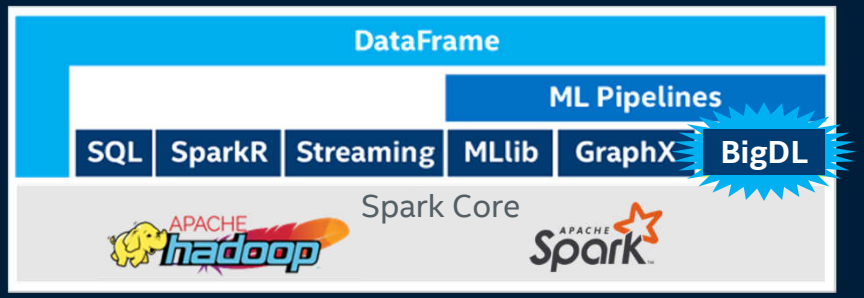
ANALYTICS ZOO

UNIFIED ANALYTICS + AI PLATFORM
DISTRIBUTED TENSORFLOW, KERAS AND BIGDL ON APACHE SPARK
Reference Use Cases, AI Models,
High-level APIs, Feature Engineering, etc.

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

UNIFYING ANALYTICS + AI ON APACHE SPARK

HIGH PERFORMANCE DEEP LEARNING FOR APACHE SPARK* ON CPU INFRASTRUCTURE



BigDL is an **open-source** distributed deep learning library for Apache Spark* that can run directly on top of existing Apache Spark or Apache Hadoop* clusters with direct access to stored data and tool/workflow consistency!

Designed and Optimized for Intel® Xeon®

No need to deploy costly accelerators, duplicate data, or suffer through scaling headaches!

Feature Parity with TensorFlow*, Caffe* and Torch*

Lower TCO and improved ease of use with existing infrastructure

Deep Learning on Big Data Platform, Enabling **Efficient Scale-Out**

Powered by Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN) and multi-threaded programming

software.intel.com/bigdl

ANALYTICS ZOO

BUILD E2E ANALYTICS & AI APPLICATIONS FOR BIG DATA AT SCALE

UNIFIED BIG DATA ANALYTICS + AI OPEN SOURCE PLATFORM


Play the Video

Reference Use Cases	▪ Anomaly detection, sentiment analysis, fraud detection, image generation, chatbot, sequence prediction, etc.
Built-In Deep Learning Models	▪ Image classification, object detection, text classification, recommendations, GANs, Sequence to Sequence, etc.
Feature Engineering	Feature transformations for ▪ Image, text, 3D imaging, time series, speech, etc.
High-Level Pipeline APIs	▪ Distributed Tensorflow* & Keras* on Apache Spark/BigDL ▪ Support for Autograd*, transfer learning, Spark DataFrame and ML Pipeline ▪ Model serving API for model serving/inference pipelines
Backends	Apache Spark, TensorFlow*, BigDL, Python, etc.

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

INTEL® DISTRIBUTION FOR PYTHON*

software.intel.com/intel-distribution-for-python

FOR DEVELOPERS USING THE MOST POPULAR AND FASTEST GROWING PROGRAMMING LANGUAGE FOR AI

EASY, OUT-OF-THE-BOX ACCESS TO HIGH PERFORMANCE PYTHON

- Prebuilt, optimized for numerical computing, data analytics, HPC
- Drop in replacement for your existing Python (no code changes required)

DRIVE PERFORMANCE WITH MULTIPLE OPTIMIZATION TECHNIQUES

- Accelerated NumPy/SciPy/Scikit-Learn with Intel® MKL
- Data analytics with pyDAAL, enhanced thread scheduling with TBB, Jupyter* Notebook interface, Numba, Cython
- Scale easily with optimized MPI4Py and Jupyter notebooks

FASTER ACCESS TO LATEST OPTIMIZATIONS FOR INTEL® ARCHITECTURE

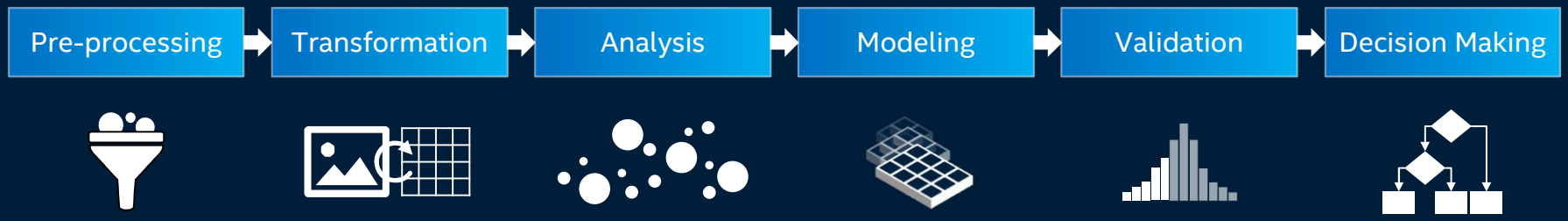
- Distribution and individual optimized packages available through conda and Anaconda Cloud
- Optimizations upstreamed back to main Python trunk

ADVANCING PYTHON* PERFORMANCE CLOSER TO NATIVE SPEEDS

All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.
Other names and brands may be claimed as the property of others.

INTEL® DATA ANALYTICS ACCELERATION LIBRARY (INTEL® DAAL)

BUILDING BLOCKS FOR ALL DATA ANALYTICS STAGES, INCLUDING DATA PREPARATION, DATA MINING & MACHINE LEARNING



Open Source | Apache* 2.0 License

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

Optimized for large data sets including streaming and distributed processing

Flexible interfaces to leading big data platforms including Spark* and range of data formats (CSV, SQL, etc.)

HIGH PERFORMANCE MACHINE LEARNING AND DATA ANALYTICS LIBRARY

INTEL® MATH KERNEL FOR DEEP LEARNING 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® MKL library.

github.com/01org/mkl-dnn

EXAMPLES:

- Direct 2D Convolution
- Local response normalization (LRN)
- Rectified linear unit neuron activation (ReLU)
- Maximum pooling
- Inner product

Accelerate Performance of Deep Learning Models

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

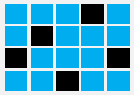
OpenCV · OpenCL™ · OpenVX™

- 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

INTEL® DEEP LEARNING DEPLOYMENT TOOLKIT

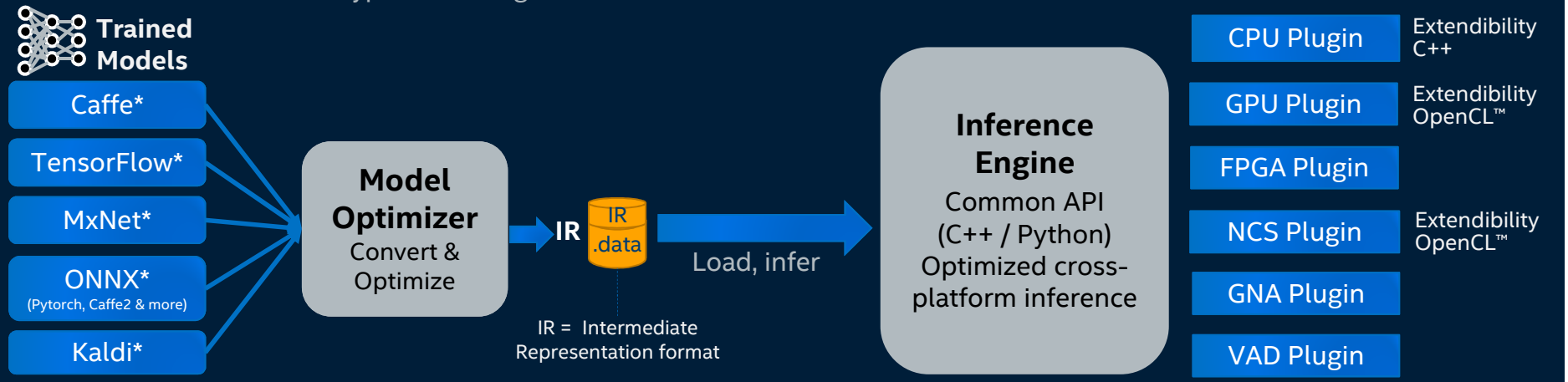
FOR DEEP LEARNING INFERENCE – PART OF INTEL® DISTRIBUTION OF OPENVINO TOOLKIT

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

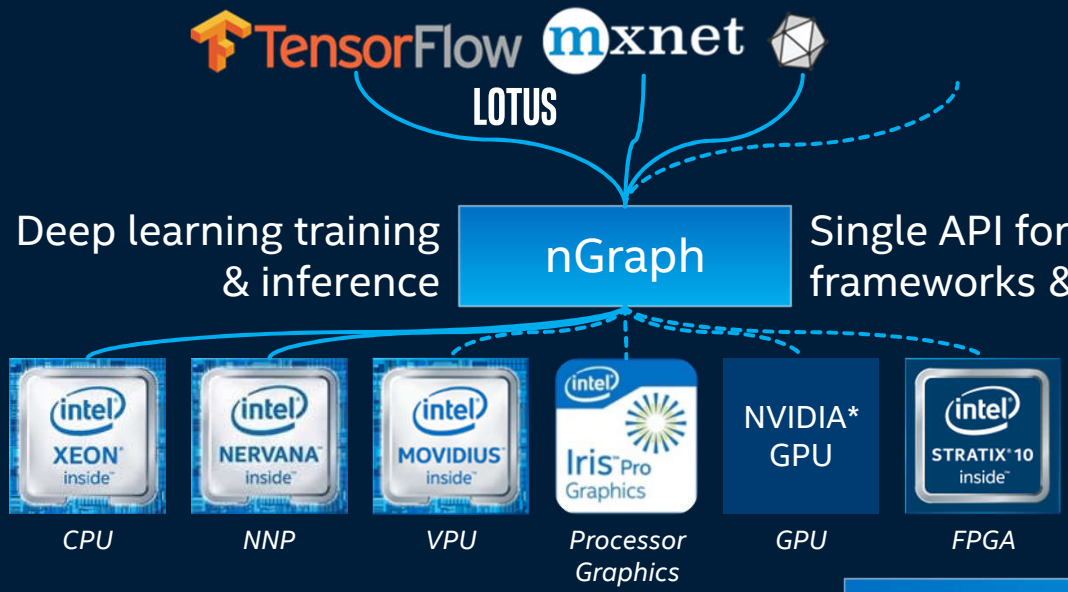
VAD = Vision Accelerator Design Products; includes FPGA and 8 MyriadX versions

Also available in the open source version (01.org/openvinotoolkit)

INTEL® NGRAPH™ COMPILER

---------- Work in progress

NOW IN BETA!



github.com/NervanaSystems/ngraph

OPEN-SOURCE C++ LIBRARY, COMPILER & RUNTIME FOR DEEP LEARNING ENABLING FLEXIBILITY TO RUN MODELS ACROSS A VARIETY OF FRAMEWORKS AND HARDWARE

All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.
*Other names and brands may be claimed as the property of others.

NAUTA

NOW IN BETA!

BUILD

Multi-user collaboration

Interactive sessions

Template functionality

TRAIN

Fast training

Batch training

Experiment tracking

Multi-node distribution

Analytics & visualization using TensorBoard*

EVALUATE

Batch inference

Inference deployment

Export to edge devices

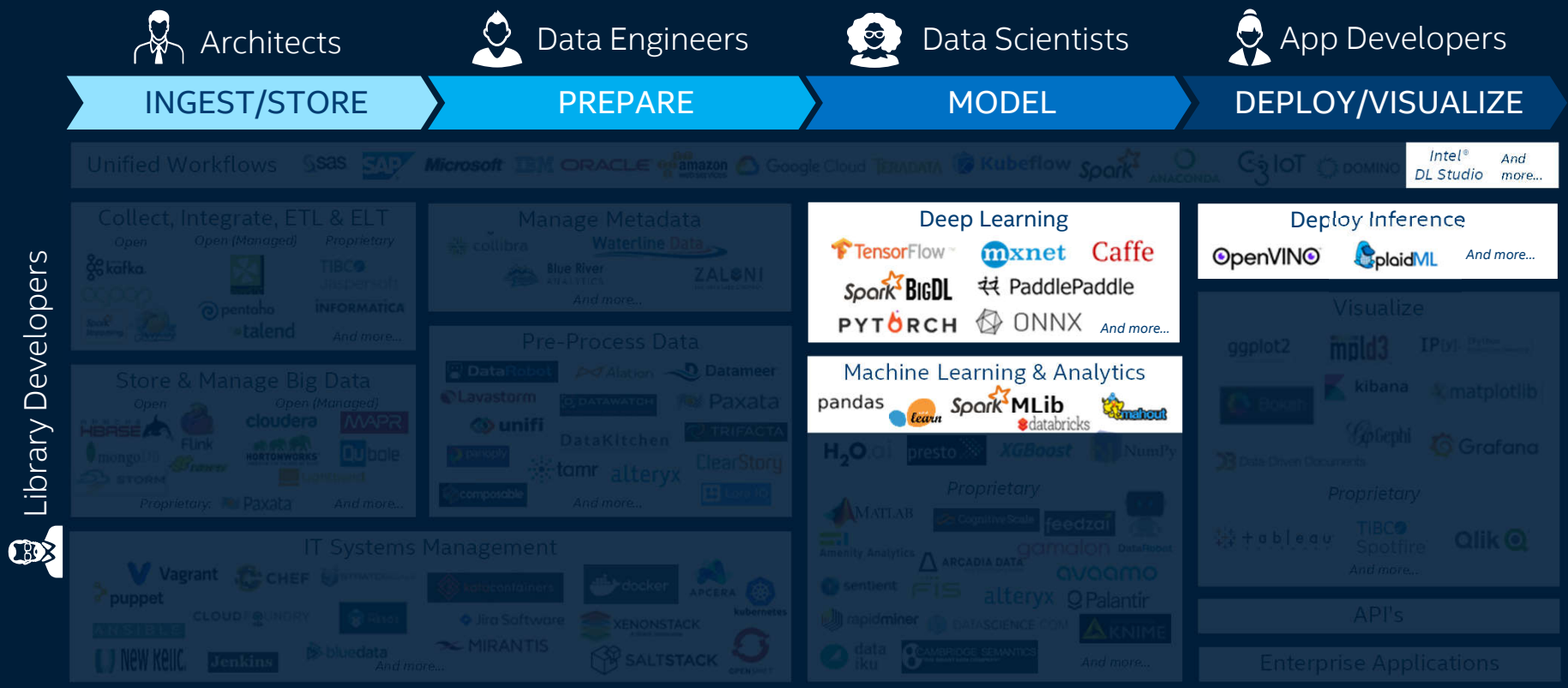
github.com/intelAI/Nauta

OPEN-SOURCE DISTRIBUTED DEEP LEARNING PLATFORM FOR KUBERNETES*

All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.
* Other names and brands may be claimed as the property of others.

ANIMATED

AI SOFTWARE IN THE BIG PICTURE



Other names and brands may be claimed as the property of others.
Note: displayed logos are not a complete list of solutions/providers in each category, personas are not mutually-exclusive by workflow step, and categorization is generalized

INTEL® AI BUILDERS BENEFITS FOR PARTNERS

builders.intel.com/ai

Partner Activation

Available to all partners

TECH ENABLEMENT

- Account mgt
- Tech enablement
- Intel® AI Dev Cloud for Builders
- Learning resources

+

Partners demonstrating solutions on Intel AI

CO-MARKETING

- Event demos & speakerships
- Builders Solutions Library
- Digital content/ social channels

+

Partners optimized on Intel AI

MATCH-MAKING

- Match-making of optimized partners with Intel enterprise customers

+

Select start ups

INVESTMENT

- Considered for investment by our dedicated Intel Capital AI investment team

INTEL® AI BUILDERS BENEFITS FOR END USERS

builders.intel.com/ai

Road Map Alignment → Engagement and Support from Planning to Deployment → **Faster and Safer Deployments**

CONNECT

with preferred solution, providers, get access to state of the art solutions and benefit from best practices

+ TEST

your roadmap of interoperable, scalable, and flexile solutions by utilizing the resources available only to Intel® AI Builders members

+ DEPLOY

optimized, tested, and reliable solutions, and continue to drive the transformation of your system

INTEL® SELECT SOLUTIONS FOR AI

All Intel® Select Solution configurations and benchmark results are **VERIFIED BY INTEL**

intel select solution

intel.com/selectsolutions

SIMPLIFIED EVALUATION — Tightly-specified HW and SW components, eliminating guesswork

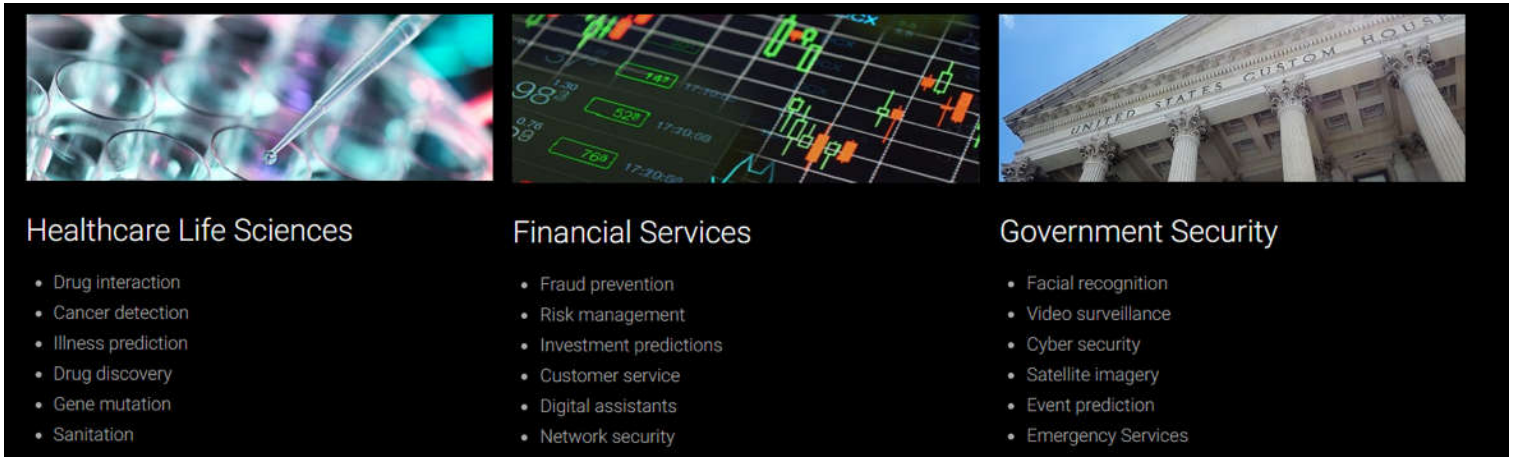
FAST AND EASY TO DEPLOY — Pre-defined settings and system-wide tuning, enabling smooth deployment

WORKLOAD OPTIMIZED — Designed and benchmarked to perform optimally for specific workloads

AI SOLUTIONS FOR (1) DL INFERENCE AND (2) ANALYTICS+DL ON BIGDL (SPARK*)

*Other names and brands may be claimed as the property of others

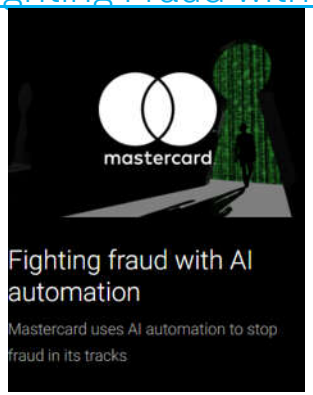
DELL OVERVIEW AND OFFERINGS



- Dell Ready Solutions for AI: Dell Ready Solutions
  - Ready Solution for ML (BigDL, Spark, Hadoop) Now
  - Ready Solution for DL (Nauta + Xeon Cluster) May
- System Alignment Focus:
- Dell R740 and converged C6420 Platforms

DELL AI USE CASES

Fighting Fraud with AI



"The solution can index, match and sort results using several search algorithms and new scoring capabilities that were previously impractical to implement on a legacy platform.

The end result of these efforts is a more trustworthy transaction experience for legitimate cardholders and merchants and more digital barriers to stop the criminals who try to exploit vulnerabilities in the payment systems."

Genomic Data Analysis

"By identifying this gene mutation, with help from our Dell EMC HPC cluster, TGen will potentially be able to identify other children who have the disorder,"

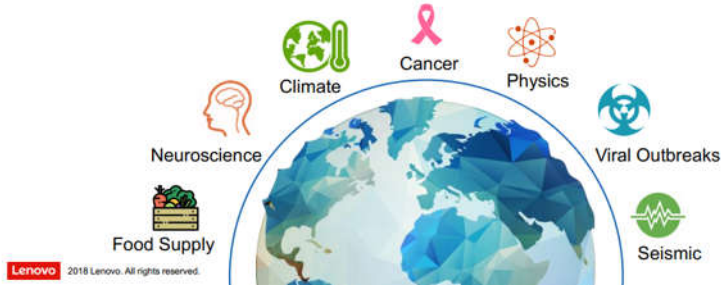
> *"By identifying this gene mutation, with help from our Dell EMC HPC cluster, TGen will potentially be able to identify other children who have the disorder"*

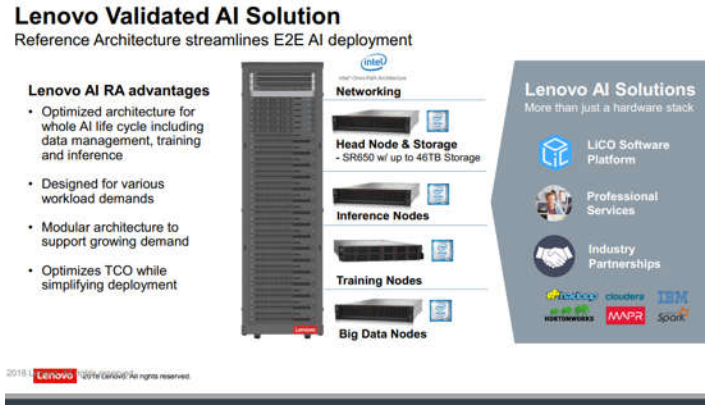
Note: Links to PDF full descriptions

LENOVO OVERVIEW AND OFFERINGS

IPC Sponsor

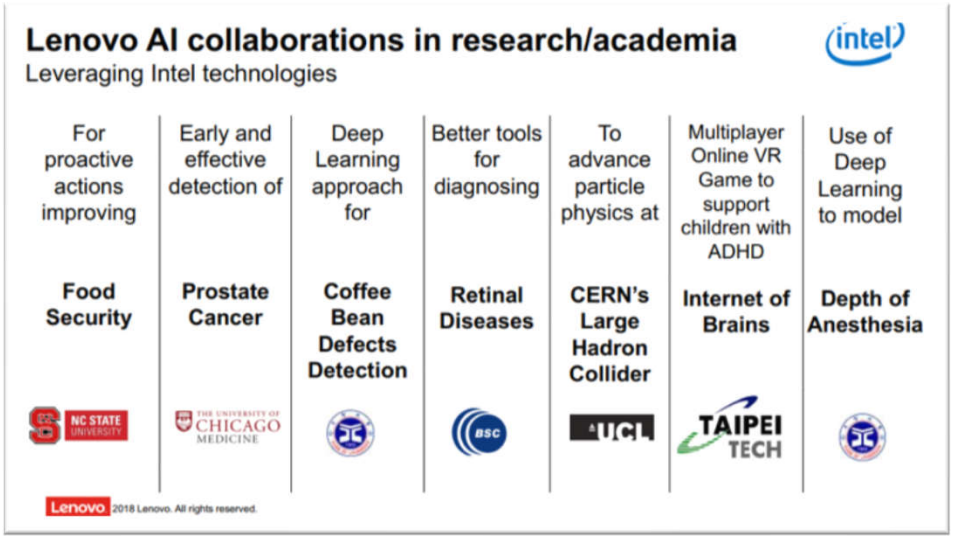
Lenovo HPC & AI: Helping Solve Humanity's Greatest Challenges





- Vertical Markets served: **Mfg, Retail, Financial, HLS**
- LICO - Lenovo Intelligent Computing Orchestration
  - Working w/ Lenovo to optimize and integrate Intel SI and SW tools elements to expedite solutio
  - Lenovo FPGA acceleration for LICO
- Systems Alignment: Focus on ThinkSystem SR670 Rack Server

LENOVO AI USE CASES



Significant Research in AI with Academia/Research

Computer Vision in a manufacturing environment
Check out demo in the Lenovo booth

INSPUR AI SOLUTIONS BASED ON INTEL

Inference accelerators based on Intel Xeon + FPGA

F10A | FPGA Accelerator

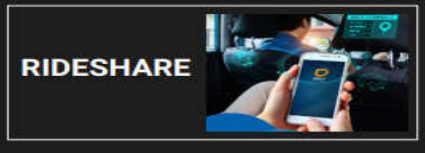
• Altera based • 1.366 TFlops

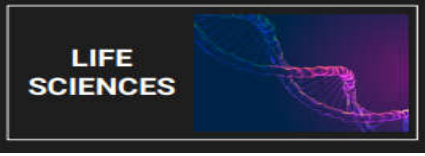
S10 | FPGA Accelerator

Training and Inference based on Xeon

NF5468M5-V

INDUSTRIES EMPOWERED BY INSPUR AI





Hyperscale Customer

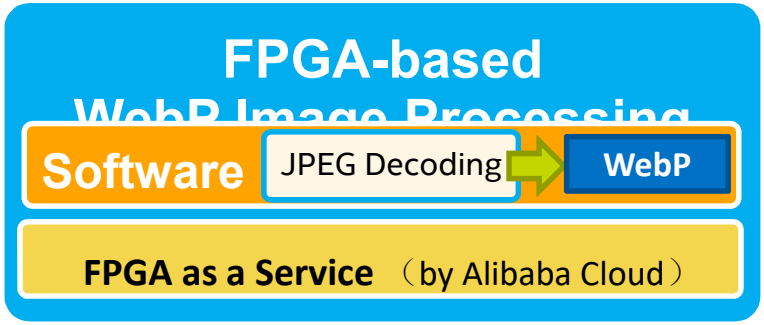
XEON + FPGA BASED WEBP IMAGE PROCESS SOLUTION

Critical Challenge

Affected by the slow decoding speed of JPG & PNG images, few companies in China are able to use WebP, an image compression format reducing the size of JPEG images.

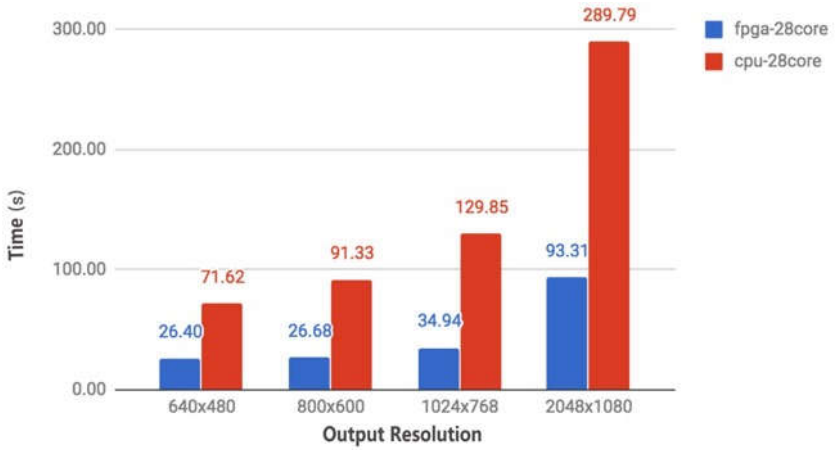
FPGA Solution

This enables users to utilize JPEG decoding system and WebP coding system to convert JPEG format into WebP in batches.



User Benefit

This F10A-based image processing solution achieves several times the performance of image processing based on CPU.



HPE AI SOLUTIONS BASED ON INTEL

ProLiant Servers

Apollo Servers

Autonomous Vehicles

Fraud Detection

Video Surveillance

Prescriptive Maintenance


intel AI
APPENDIX B – BONUS GOLD DECK SLIDES

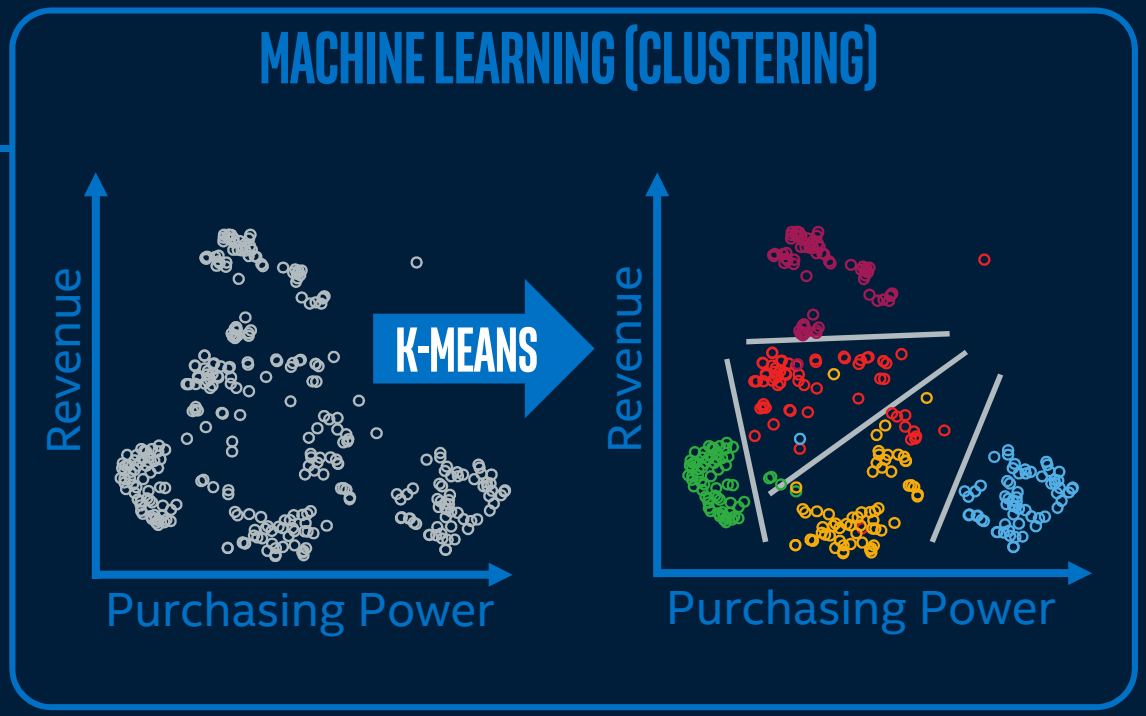
HOW DOES MACHINE LEARNING WORK?

MACHINE LEARNING

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

DEEP LEARNING

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



CHOOSE THE BEST AI APPROACH FOR YOUR CHALLENGE

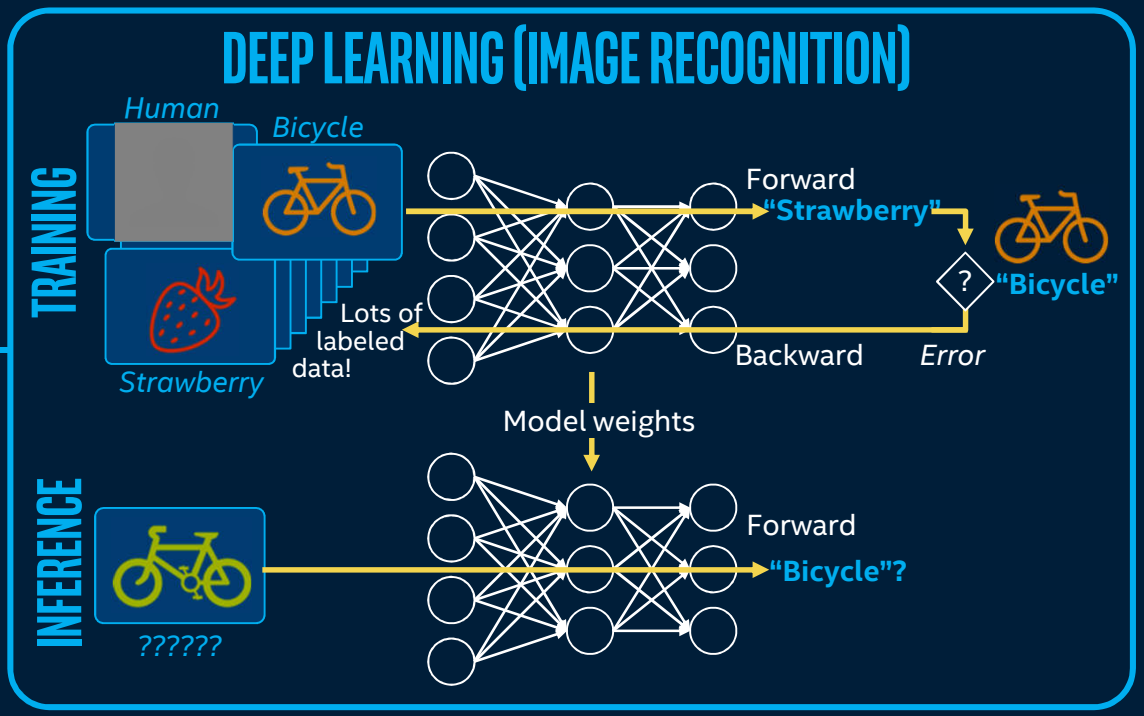
HOW DOES DEEP LEARNING WORK?

MACHINE LEARNING

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

DEEP LEARNING

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



CHOOSE THE BEST AI APPROACH FOR YOUR CHALLENGE

HOW DOES DEEP LEARNING WORK?

MACHINE LEARNING

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

DEEP LEARNING

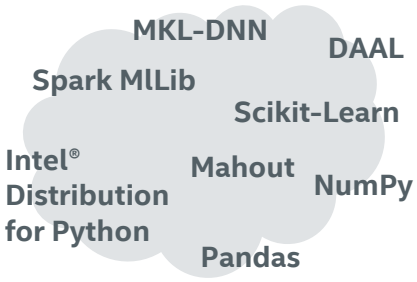
- Speech Processing
- Natural Language Processing
- Recommender Systems
- Adversarial Networks
- Reinforcement Learning

DEEP LEARNING (REINFORCEMENT LEARNING)

CHOOSE THE BEST AI APPROACH FOR YOUR CHALLENGE

DEEP LEARNING GLOSSARY

LIBRARY



Optimized primitive functions for AI

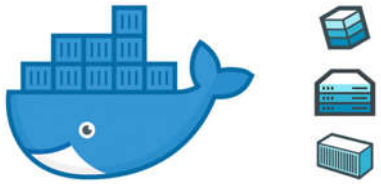
FRAMEWORK

Open-source development environments

TOPOLOGY

Specific neural network implementations

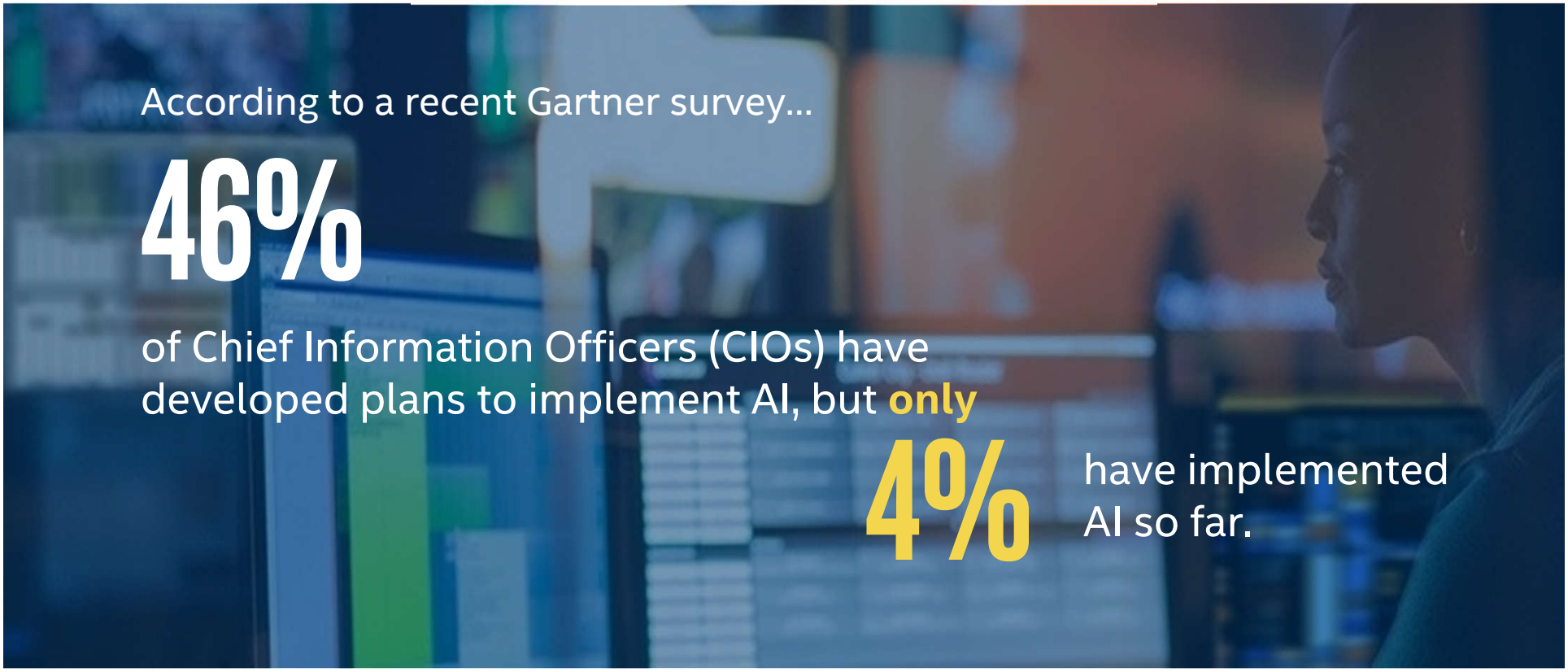
CONTAINER



Pre-configured AI environments ready to depoy

Learn more about “what is AI” at software.intel.com/ai/course

AI ADOPTION IS NASCENT



Source: Gartner Says Nearly Half of CIOs Are Planning to Deploy Artificial Intelligence. February 2018 (https://www.gartner.com/newsroom/id/3856163)

AI OPPORTUNITY ASSESSMENT

Brainstorm and Prioritize Business Challenges

What business challenges am I facing today?

What business value is tied to each challenge?

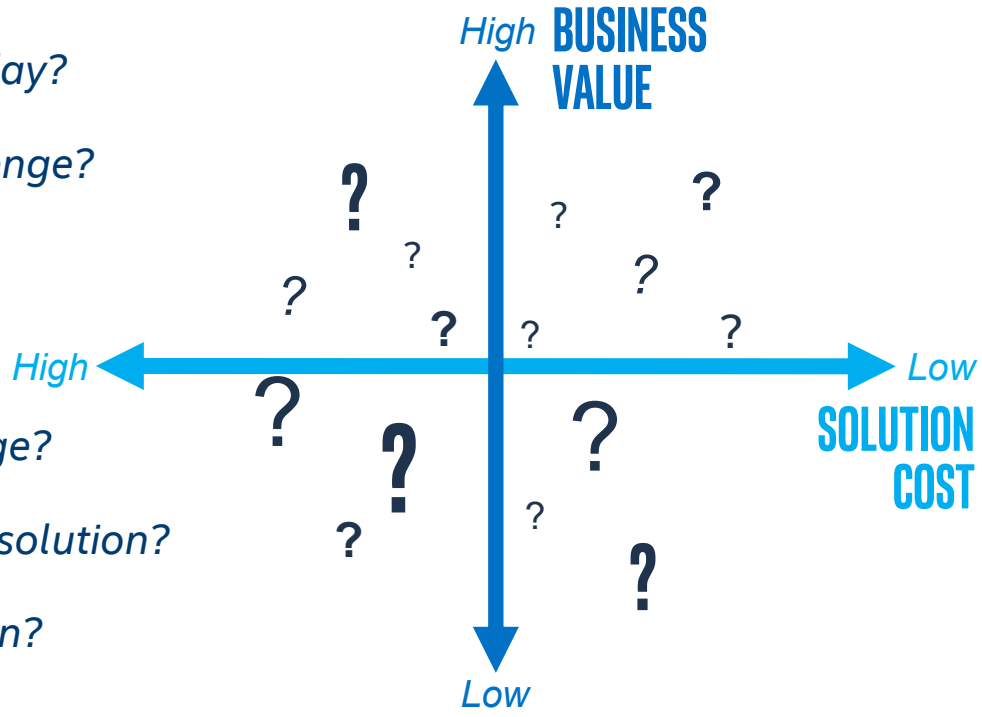
What are my solution requirements?

What data do I have at my disposal?

Do I know how to approach each challenge?

Do I have what I need to implement each solution?

How costly is it to implement each solution?



MACHINE VS. DEEP LEARNING

MACHINE LEARNING

How do you engineer the best features?

$N \times N$

$(f_1, f_2, \cdots, f_K)$

Roundness of face
Dist between eyes
Nose width
Eye socket depth
Cheek bone structure
Jaw line length
...etc.

CLASSIFIER ALGORITHM

SVM
Random Forest
Naïve Bayes
Decision Trees
Logistic Regression
Ensemble methods

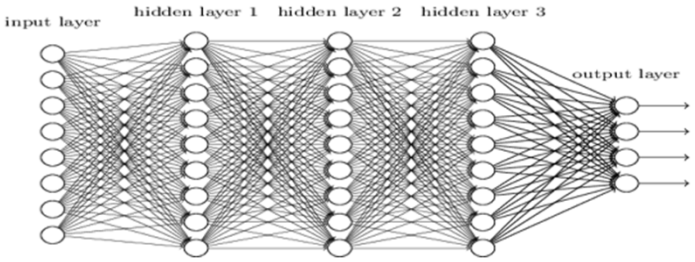
Arjun

DEEP LEARNING

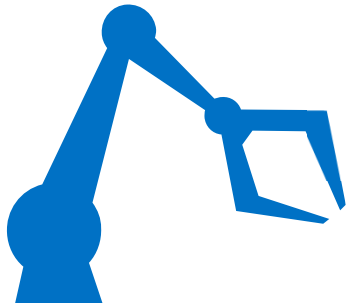
How do you guide the model to find the best features?

$N \times N$

NEURAL NETWORK



Arjun



WHICH APPROACH IS RIGHT?

A large **manufacturer** uses data to improve their operations, with each challenge using a different approach to deliver maximum business value at the lowest possible cost

CHALLENGE	BEST APPROACH	APPROACH	ANSWER
How many widgets should we manufacture?	Analyze historical supply/demand	Analytics/ Business Intelligence	10,000
What will our yield be?	Algorithm that correlates many variables to yield	Statistical/ Machine Learning	At current conditions, yield will be at 90% with 10% loss expected
Which widgets have visual defects?	Algorithm that learns to identify defects in images	Deep Learning	Widget 1003, Widget 1094 . . .

LEARN MORE IN THE NEXT SLIDES

AI CLOSER LOOK

MACHINE LEARNING

Algorithms designed to deliver better insight with more data

Regression (Linear/Logistic)

Classification (Support Vector Machines/SVM, Naïve Bayes)

Clustering (Hierarchical, Bayesian, K-Means, DBSCAN)

Decision Trees (RandomForest)

Extrapolation (Hidden Markov Models/HMM)

More...

DEEP LEARNING

Neural networks used to infer meaning from large dense datasets

Image Recognition (Convolutional Neural Networks/CNN, Single-Shot Detector/SSD)

Speech Recognition (Recurrent Neural Network/RNN)

Natural Language Processing (Long-Short Term Memory/LSTM)

Data Generation (Generative Adversarial Networks/GAN)

Recommender System (Multi-Layer Perceptron/MLP)

Time-Series Analysis (LSTM, RNN)

Reinforcement Learning (CNN, RNN)

More...

REASONING

Hybrid of analytics & AI techniques designed to find meaning in diverse datasets

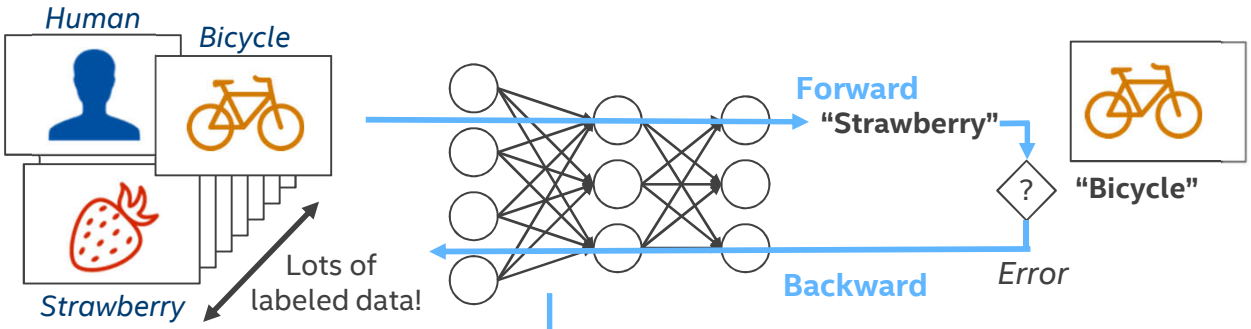
Associative Memory (Intel® Saffron AI memory base)

← **See also:** machine & deep learning techniques

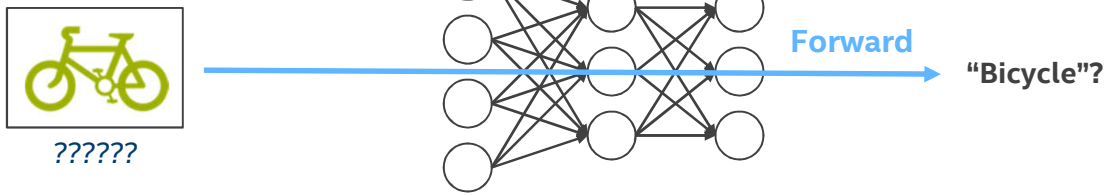
More...

DEEP LEARNING BASICS

TRAINING

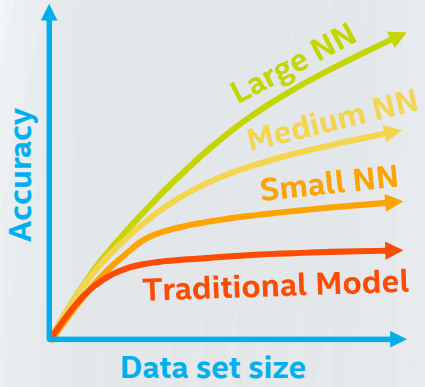


INFERENCE



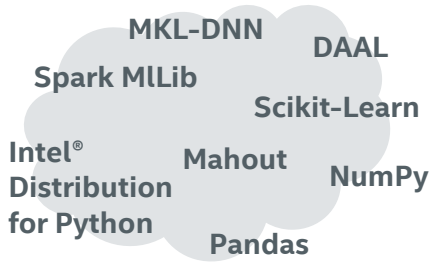
DID YOU KNOW?

Training with a large data set AND deep (many layered) neural network often leads to the highest accuracy inference



DEEP LEARNING GLOSSARY

LIBRARY



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

FRAMEWORK

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

DEEP LEARNING USAGES & KEY TOPOLOGIES

Image Recognition

Resnet-50
Inception V3
MobileNet
SqueezeNet

Language Translation

GNMT

Object Detection

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

Text to Speech

Wavenet

Image Segmentation

Mask R-CNN

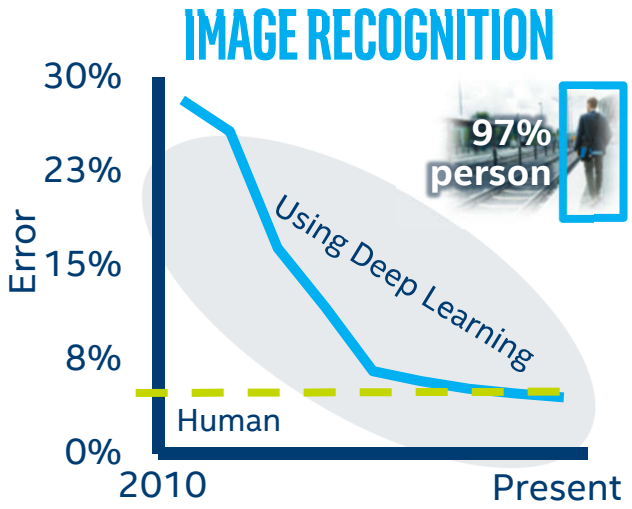
Recommendation System

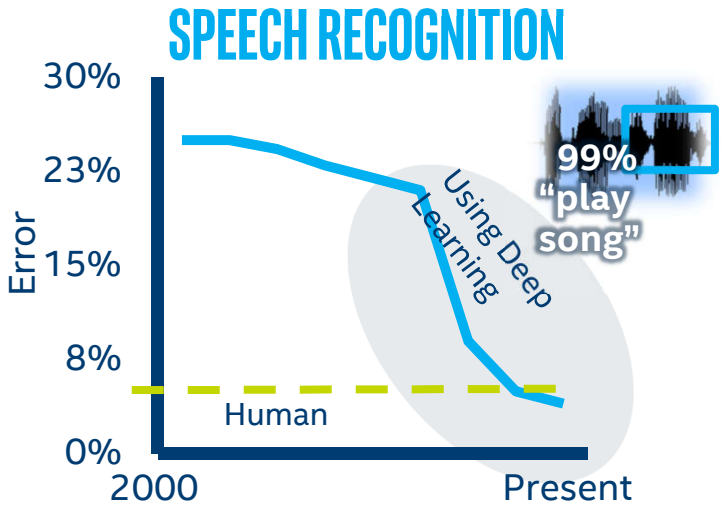
Wide & Deep, NCF

There are many deep learning usages and topologies for each

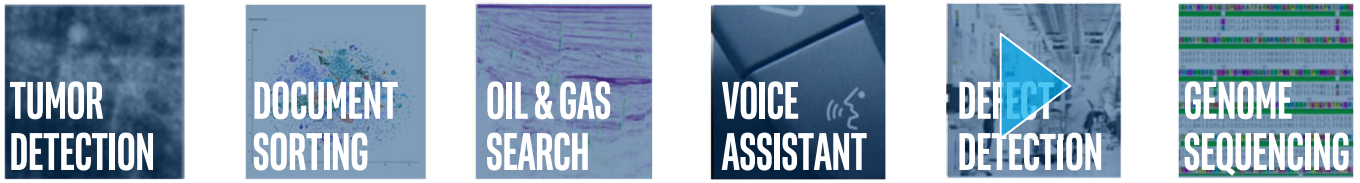
DEEP LEARNING BREAKTHROUGHS

Machines able to meet or exceed human image & speech recognition





e.g.



Source: ILSVRC ImageNet winning entry classification error rate each year 2010-2016 (Left), https://www.microsoft.com/en-us/research/blog/microsoft-researchers-achieve-new-conversational-speech-recognition-milestone/ (Right)

HARDWARE

Multi-purpose to purpose-built
AI compute from cloud to device

intel AI

	MAINSTREAM	INTENSIVE
DEEP LEARNING → TRAINING	intel XEON PLATINUM inside	+ intel NERVANA inside (COMING 2019)
DEEP LEARNING → INFERENCE	intel ATOM / CORE 8th Ge / XEON PLATINUM inside	+ intel STRATIX 10 / CORE i7 8th Gen / IRIS / MOVIDIUS / GNA
MOST OTHER AI	intel ATOM / CORE 8th Ge / XEON PLATINUM inside	

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

ONE SIZE DOES NOT FIT ALL

intel

HARDWARE

Multi-purpose to purpose-built
AI compute from device to cloud

intel AI

END POINT

User-touch end point devices with lower power requirements such as laptops, tablets, smart home devices, drones

EDGE

Small scale data centers, small business IT infrastructure, to few on-premise server racks and workstations

DATA CENTER

Large scale data centers such as public cloud or comms service providers, gov't and academia, large enterprise IT

Varies to <1ms — <5ms — <10-40ms — ~100ms

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

AI IS EXPANDING

intel

HARDWARE

Multi-purpose to purpose-built
AI compute from device to cloud

intel AI

END POINT

IOT SENSORS
(Security, home, retail, industrial...)

Vision & Inference

Speech

SELF-DRIVING VEHICLES

Autonomous Driving

DESKTOP & MOBILITY

SOC

M.2 Card

Display, Video, AR/VR, Gestures, Speech

EDGE

SERVERS, APPLIANCES & GATEWAYS

Most Use Cases

+ Special Purpose

Dedicated Media & Vision Inference

Latency-Bound Inference

Basic Inference, Media & Vision

DATA CENTER

SERVERS & APPLIANCES

Most Use Cases

NNP-T

Most Intensive Use Cases

+ Special Purpose

Flexible & Memory Bandwidth-Bound Use Cases

Varies to <1ms — <5ms — <10-40ms — ~100ms

[1]GNA=Gaussian Neural Accelerator

All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice. Images are examples of intended applications and are not an exhaustive list.

ONE SIZE DOES NOT FIT ALL

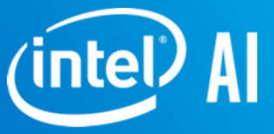
intel

intel AI

Bring Your AI Vision to Life Using Our Complete Portfolio

DATA
Intel analytics ecosystem to get your data ready

FUTURE
Driving AI forward through R&D, investments and policy



APPENDIX C – CONFIGURATION DETAILS

Config 3

CONFIGURATIONS FOR CONTINUED INNOVATION DRIVING DL GAINS ON XEON® (MARCH' 2019)

1x inference throughput improvement on Intel® Xeon® Platinum 8180 processor (July 2017) baseline : Tested by Intel as of July 11th 2017: Platform: 2S Intel® Xeon® Platinum 8180 CPU @ 2.50GHz (28 cores), HT disabled, turbo disabled, scaling governor set to "performance" via intel_pstate driver, 384GB DDR4-2666 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD DC S3700 Series (800GB, 2.5in SATA 6Gb/s, 25nm, MLC).Performance measured with: Environment variables: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, CPU Freq set with cpupower frequency-set -d 2.5G -u 3.8G -g performance. Caffe: (http://github.com/intel/caffe/), revision f96b759f71b2281835f690af267158b82b150b5c. Inference measured with "caffe time --forward_only" command, training measured with "caffe time" command. For "ConvNet" topologies, synthetic dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50),and https://github.com/soumith/convnet-benchmarks/tree/master/caffe/imagenet_winners (ConvNet benchmarks; files were updated to use newer Caffe prototxt format but are functionally equivalent). Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.20170425. Caffe run with "numactl -l".

5.7x inference throughput improvement on Intel® Xeon® Platinum 8180 processor (December 2018) with continued optimizations : Tested by Intel as of November 11th 2018 :2 socket Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores HT ON , Turbo ON Total Memory 376.46GB (12slots / 32 GB / 2666 MHz). CentOS Linux-7.3.1611-Core, kernel: 3.10.0-862.3.3.el7.x86_64, SSD sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB , Deep Learning Framework Intel® Optimization for caffe version: 551a53d63a6183c233abaa1a19458a25b672ad41 Topology::ResNet_50_v1 BIOS:SE5C620.86B.00.01.0014.070920180847 MKLDNN: 4e333787e0d66a1dca1218e99a891d493dbc8ef1 instances: 2 instances socket:2 (Results on Intel® Xeon® Scalable Processor were measured running multiple instances of the framework. Methodology described here: https://software.intel.com/en-us/articles/boosting-deep-learning-training-inference-performance-on-xeon-and-xeon-phi) Synthetic data. Datatype: INT8 Batchsize=64 vs Tested by Intel as of July 11th 2017:2S Intel® Xeon® Platinum 8180 CPU @ 2.50GHz (28 cores), HT disabled, turbo disabled, scaling governor set to "performance" via intel_pstate driver, 384GB DDR4-2666 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD DC S3700 Series (800GB, 2.5in SATA 6Gb/s, 25nm, MLC).Performance measured with: Environment variables: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, CPU Freq set with cpupower frequency-set -d 2.5G -u 3.8G -g performance. Caffe: (http://github.com/intel/caffe/), revision f96b759f71b2281835f690af267158b82b150b5c. Inference measured with "caffe time --forward_only" command, training measured with "caffe time" command. For "ConvNet" topologies, synthetic dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50). Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.20170425. Caffe run with "numactl -l".

Config 3

CONFIGURATIONS FOR CONTINUED INNOVATION DRIVING DL GAINS ON XEON® (MARCH' 2019) – CONTINUED

14x inference throughput improvement on Intel® Xeon® Platinum 8280 processor with Intel® DL Boost: Tested by Intel as of 2/20/2019. 2 socket Intel® Xeon® Platinum 8280 Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2933 MHz), BIOS: SE5C620.86B.0D.01.0271.120720180605 (ucode: 0x200004d), Ubuntu 18.04.1 LTS, kernel 4.15.0-45-generic, SSD 1x sda INTEL SSDSC2BA80 SSD 745.2GB, nvme1n1 INTEL SSDPE2KX040T7 SSD 3.7TB, Deep Learning Framework: Intel® Optimization for Caffe version: 1.1.3 (commit hash: 7010334f159da247db3fe3a9d96a3116ca06b09a) , ICC version 18.0.1, MKL DNN version: v0.17 (commit hash: 830a10059a018cd2634d94195140cf2d8790a75a, model: https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS=64, syntheticData, 4 instance/2 socket, Datatype: INT8 vs Tested by Intel as of July 11th 2017: 2S Intel® Xeon® Platinum 8180 CPU @ 2.50GHz (28 cores), HT disabled, turbo disabled, scaling governor set to "performance" via intel_pstate driver, 384GB DDR4-2666 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD DC S3700 Series (800GB, 2.5in SATA 6Gb/s, 25nm, MLC).Performance measured with: Environment variables: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, CPU Freq set with cpupower frequency-set -d 2.5G -u 3.8G -g performance. Caffe: (http://github.com/intel/caffe/), revision f96b759f71b2281835f690af267158b82b150b5c. Inference measured with "caffe time --forward_only" command, training measured with "caffe time" command. For "ConvNet" topologies, synthetic dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50),. Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.20170425. Caffe run with "numactl -l".

30x inference throughput improvement on Intel® Xeon® Platinum 9282 processor with Intel® DL Boost : Tested by Intel as of 2/26/2019. Platform: Dragon rock 2 socket Intel® Xeon® Platinum 9282(56 cores per socket), HT ON, turbo ON, Total Memory 768 GB (24 slots/ 32 GB/ 2933 MHz), BIOS:SE5C620.86B.0D.01.0241.112020180249, Centos 7 Kernel 3.10.0-957.5.1.el7.x86_64, Deep Learning Framework: Intel® Optimization for Caffe version: https://github.com/intel/caffe d554cbf1, ICC 2019.2.187, MKL DNN version: v0.17 (commit hash: 830a10059a018cd2634d94195140cf2d8790a75a), model: https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS=64, No datalayer syntheticData:3x224x224, 56 instance/2 socket, Datatype: INT8 vs Tested by Intel as of July 11th 2017: 2S Intel® Xeon® Platinum 8180 CPU @ 2.50GHz (28 cores), HT disabled, turbo disabled, scaling governor set to "performance" via intel_pstate driver, 384GB DDR4-2666 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD DC S3700 Series (800GB, 2.5in SATA 6Gb/s, 25nm, MLC).Performance measured with: Environment variables: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, CPU Freq set with cpupower frequency-set -d 2.5G -u 3.8G -g performance. Caffe: (http://github.com/intel/caffe/), revision f96b759f71b2281835f690af267158b82b150b5c. Inference measured with "caffe time --forward_only" command, training measured with "caffe time" command. For "ConvNet" topologies, synthetic dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50),. Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.20170425. Caffe run with "numactl -l".

CONFIGURATION DETAILS

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer. No computer system can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Intel, the Intel logo, Intel inside, the Intel inside logo, Xeon, the Xeon logo, Xeon Phi, the Xeon Phi logo, Core, the Core logo, Atom, the Atom logo, Movidius, the Movidius logo, Stratix, the Stratix logo, Arria, the Arria logo, Myriad, Nervana and others are trademarks of Intel Corporation in the U.S. and/or other countries.
*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. Notice Revision #20110804.

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.

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

Config 9

CONFIG FOR – ACCELERATOR LIKE PERFORMANCE ON INTEL® XEON® PROCESSORS WITH INTEL® DL BOOST

Nvidia data source: https://developer.nvidia.com/deep-learning-performance-training-inference

Max Inference throughput at <7ms

Intel® Xeon® Platinum 8180 processor: Tested by Intel as of 2/26/2019. 2S Intel® Xeon® Platinum 8280(28 cores per socket), HT ON, turbo ON, Total Memory 384 GB (12 slots/ 32 GB/ 2933 MHz), BIOS: SE5C620.86B.0D.01.0348.011820191451, Centos 7 Kernel 3.10.0-957.5.1.el7.x86_64, Deep Learning Framework: Intel® Optimization for Caffe version: https://github.com/intel/caffe Commit id: 362a3b3, ICC 2019.2.187 for build, MKL DNN version: v0.17 (commit hash: 830a10059a018cd2634d94195140cf2d8790a75a), model: https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS=10, synthetic Data:3x224x224, 2 instance/2 socket, Datatype: INT8; latency: 6.16 ms

Intel® Xeon® Platinum 9242 Processor: Tested by Intel as of 2/26/2019 2S Intel® Xeon® Platinum 9242(48 cores per socket), HT ON, turbo ON, Total Memory 768 GB (24 slots/ 32 GB/ 2933 MHz), BIOS: SE5C620.86B.0D.01.0403.022020190327, Centos 7 Kernel 3.10.0-957.5.1.el7.x86_64, Deep Learning Framework: Intel® Optimization for Caffe version: https://github.com/intel/caffe Commit id: 362a3b3, ICC 2019.2.187 for build, MKL DNN version: v0.17 (commit hash: 830a10059a018cd2634d94195140cf2d8790a75a), model: https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS= 2, synthetic Data:3x224x224, 16 instance/2 socket, Datatype: INT8; latency: 6.90 ms

Intel® Xeon® Platinum 9282 Processor: Tested by Intel as of 2/26/2019. DL Inference: Platform: Dragon rock 2S Intel® Xeon® Platinum 9282(56 cores per socket), HT ON, turbo ON, Total Memory 768 GB (24 slots/ 32 GB/ 2933 MHz), BIOS: SE5C620.86B.0D.01.0241.112020180249, Centos 7 Kernel 3.10.0-957.5.1.el7.x86_64, Deep Learning Framework: Intel® Optimization for Caffe version: https://github.com/intel/caffe Commit id: 362a3b3, ICC 2019.2.187 for build, MKL DNN version: v0.17 (commit hash: 830a10059a018cd2634d94195140cf2d8790a75a), model: https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS=10, synthetic Data:3x224x224, 4 instance/2 socket, Datatype: INT8; latency: 6.91 ms

Max Inference throughput

Intel® Xeon® Platinum 8180 processor: Tested by Intel as of 2/26/2019. 2S Intel® Xeon® Platinum 8280(28 cores per socket), HT ON, turbo ON, Total Memory 384 GB (12 slots/ 32 GB/ 2933 MHz), BIOS: SE5C620.86B.0D.01.0348.011820191451, Centos 7 Kernel 3.10.0-957.5.1.el7.x86_64, Deep Learning Framework: Intel® Optimization for Caffe version: https://github.com/intel/caffe Commit id: 362a3b3, ICC 2019.2.187 for build, MKL DNN version: v0.17 (commit hash: 830a10059a018cd2634d94195140cf2d8790a75a), model: https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS=8, syntheticData:3x224x224, 14 instance/2 socket, Datatype: INT8

Intel® Xeon® Platinum 9242 Processor: Tested by Intel as of 2/26/2019 2S Intel® Xeon® Platinum 9242(48 cores per socket), HT ON, turbo ON, Total Memory 768 GB (24 slots/ 32 GB/ 2933 MHz), BIOS: SE5C620.86B.0D.01.0403.022020190327, Centos 7 Kernel 3.10.0-957.5.1.el7.x86_64, Deep Learning Framework: Intel® Optimization for Caffe version: https://github.com/intel/caffe Commit id: 362a3b3, ICC 2019.2.187 for build, MKL DNN version: v0.17 (commit hash: 830a10059a018cd2634d94195140cf2d8790a75a), model: https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS=128, synthetic Data:3x224x224, 4 instance/2 socket, Datatype: INT8

Intel® Xeon® Platinum 9282 Processor: Tested by Intel as of 2/26/2019. DL Inference: Platform: Dragon rock 2S Intel® Xeon® Platinum 9282(56 cores per socket), HT ON, turbo ON, Total Memory 768 GB (24 slots/ 32 GB/ 2933 MHz), BIOS: SE5C620.86B.0D.01.0241.112020180249, Centos 7 Kernel 3.10.0-957.5.1.el7.x86_64, Deep Learning Framework: Intel® Optimization for Caffe version: https://github.com/intel/caffe Commit id: 362a3b3, ICC 2019.2.187 for build, MKL DNN version: v0.17 (commit hash: 830a10059a018cd2634d94195140cf2d8790a75a), model: https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS=8, synthetic Data:3x224x224, 14 instance/2 socket, Datatype: INT8

BKMs for running multi-stream configurations on Xeon: https://www.intel.ai/wp-content/uploads/sites/69/TensorFlow_Best_Practices_Intel_Xeon_AI-HPC_v1.1_Q119.pdf

CONFIGURATION DETAILS (CONT'D)

Configuration: AI Performance – Software + Hardware

INFERENCE using FP32 Batch Size Caffe GoogleNet v1 128 AlexNet 256.

The benchmark results may need to be revised as additional testing is conducted. The results depend on the specific platform configurations and workloads utilized in the testing, and may not be applicable to any particular user's components, computer system or workloads. The results are not necessarily representative of other benchmarks and other benchmark results may show greater or lesser impact from mitigations. 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 Source: Intel measured as of June 2017 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.

Configurations for Inference throughput

Platform :2 socket Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores HT ON , Turbo ON Total Memory 376.28GB (12slots / 32 GB / 2666 MHz),4 instances of the framework, CentOS Linux-7.3.1611-Core , SSD sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB , Deep Learning Framework caffe version: a3d5b022fe026e9092fc7abc7654b1162ab9940d Topology:GoogleNet v1 BIOS:SE5C620.86B.00.01.0004.071220170215 MKLDNN: version: 464c268e544bae26f9b85a2acb9122c766a4c396 NoDataLayer. Measured: 1449.9 imgs/sec vs Platform: 2S Intel® Xeon® CPU E5-2699 v3 @ 2.30GHz (18 cores), HT enabled, turbo disabled, scaling governor set to "performance" via intel_pstate driver, 64GB DDR4-2133 ECC RAM. BIOS: SE5C610.86B.01.01.0024.021320191901, CentOS Linux-7.5.1804(Core) kernel 3.10.0-862.3.2.el7.x86_64, SSD sdb INTEL SSDSC2BW24 SSD 223.6GB. Framework BVLC-Caffe: https://github.com/BVLC/caffe, Inference & Training measured with "caffe time" command. For "ConvNet" topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. BVLC Caffe (http://github.com/BVLC/caffe), revision 2a1c552b66f026c7508d390b526f2495ed3be594

Configuration for training throughput:

Platform :2 socket Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores HT ON , Turbo ON Total Memory 376.28GB (12slots / 32 GB / 2666 MHz),4 instances of the framework, CentOS Linux-7.3.1611-Core , SSD sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB , Deep Learning Framework caffe version: a3d5b022fe026e9092fc7abc765b1162ab9940d Topology:alexnet BIOS:SE5C620.86B.00.01.0004.071220170215 MKLDNN: version: 464c268e544bae26f9b85a2acb9122c766a4c396 NoDataLayer. Measured: 1257 imgs/sec vs Platform: 2S Intel® Xeon® CPU E5-2699 v3 @ 2.30GHz (18 cores), HT enabled, turbo disabled, scaling governor set to "performance" via intel_pstate driver, 64GB DDR4-2133 ECC RAM. BIOS: SE5C610.86B.01.01.0024.021320191901, CentOS Linux-7.5.1804(Core) kernel 3.10.0-862.3.2.el7.x86_64, SSD sdb INTEL SSDSC2BW24 SSD 223.6GB. Framework BVLC-Caffe: https://github.com/BVLC/caffe, Inference & Training measured with "caffe time" command. For "ConvNet" topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. BVLC Caffe (http://github.com/BVLC/caffe), revision 2a1c552b66f026c7508d390b526f2495ed3be594

CONFIGURATION DETAILS (CONT'D)

Configuration: AI Performance – Software + Hardware

1.4x training throughput improvement in August 2019:

Tested by Intel as of measured August 2nd 2019. Processor: 2 socket Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores HT ON , Turbo ON Total Memory 376.46GB (12slots / 32 GB / 2666 MHz). CentOS Linux-7.3.1611-Core kernel 3.10.0-693.11.6.el7.x86_64, SSD sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB , Deep Learning Framework Intel® Optimizations for caffe version:a3d5b022fe026e9092fc7abc7654b1162ab9940d Topology::resnet_50 BIOS:SE5C620.86B.00.01.0013.030920190427 MKLDNN: version: 464c268e544bae26f9b85a2acb9122c766a4c396 NoDataLayer. Measured: 123 imgs/sec vs Intel tested July 11th 2017 Platform: Platform: 2S Intel® Xeon® Platinum 8180 CPU @ 2.50GHz (28 cores), HT disabled, turbo disabled, scaling governor set to "performance" via intel_pstate driver, 384GB DDR4-2666 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD DC S3700 Series (800GB, 2.5in SATA 6Gb/s, 25nm, MLC).Performance measured with: Environment variables: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, CPU Freq set with cpupower frequency-set -d 2.5G -u 3.8G -g performance. Caffe: (http://github.com/intel/caffe/), revision f96b759f71b2281835f690af267158b82b150b5c. Inference measured with "caffe time --forward_only" command, training measured with "caffe time" command. For "ConvNet" topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from https://github.com/intel/caffe/tree/master/models/intel_optimized_models (GoogLeNet, AlexNet, and ResNet-50), https://github.com/intel/caffe/tree/master/models/default_vgg_19 (VGG-19), and https://github.com/soumith/convnet-benchmarks/tree/master/caffe/imagenet_winners (ConvNet benchmarks; files were updated to use newer Caffe prototxt format but are functionally equivalent). Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2019.0.20170425. Caffe run with "numactl -l".

5.4x inference throughput improvement in August 2019:

Tested by Intel as of measured July 26th 2019 :2 socket Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores HT ON , Turbo ON Total Memory 376.46GB (12slots / 32 GB / 2666 MHz). CentOS Linux-7.3.1611-Core, kernel: 3.10.0-862.3.3.el7.x86_64, SSD sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB , Deep Learning Framework Intel® Optimized caffe version:a3d5b022fe026e9092fc7abc7654b1162ab9940d Topology::resnet_50_v1 BIOS:SE5C620.86B.00.01.0013.030920190427 MKLDNN: version:464c268e544bae26f9b85a2acb9122c766a4c396 instances: 2 instances socket:2 (Results on Intel® Xeon® Scalable Processor were measured running multiple instances of the framework. Methodology described here: https://software.intel.com/en-us/articles/boosting-deep-learning-training-inference-performance-on-xeon-and-xeon-phi) NoDataLayer. Datatype: INT8 Batchsize=64 Measured: 1233.39 imgs/sec vs Tested by Intel as of July 11th 2017:2S Intel® Xeon® Platinum 8180 CPU @ 2.50GHz (28 cores), HT disabled, turbo disabled, scaling governor set to "performance" via intel_pstate driver, 384GB DDR4-2666 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD DC S3700 Series (800GB, 2.5in SATA 6Gb/s, 25nm, MLC). **Performance measured with:** Environment variables: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, CPU Freq set with cpupower frequency-set -d 2.5G -u 3.8G -g performance. Caffe: (http://github.com/intel/caffe/), revision f96b759f71b2281835f690af267158b82b150b5c. Inference measured with "caffe time --forward_only" command, training measured with "caffe time" command. For "ConvNet" topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50). Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2019.0.20170425. Caffe run with "numactl -l".

11X inference thoughput improvement with CascadeLake:

Future Intel Xeon Scalable processor (codename Cascade Lake) 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 vs Tested by Intel as of July 11th 2017: 2S Intel® Xeon® Platinum 8180 CPU @ 2.50GHz (28 cores), HT disabled, turbo disabled, scaling governor set to "performance" via intel_pstate driver, 384GB DDR4-2666 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD DC S3700 Series (800GB, 2.5in SATA 6Gb/s, 25nm, MLC). **Performance measured with:** Environment variables: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, CPU Freq set with cpupower frequency-set -d 2.5G -u 3.8G -g performance. Caffe: (http://github.com/intel/caffe/), revision f96b759f71b2281835f690af267158b82b150b5c. Inference measured with "caffe time --forward_only" command, training measured with "caffe time" command. For "ConvNet" topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50),. Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2019.0.20170425. Caffe run with "numactl -l".

CONFIGURATION DETAILS (CONT'D)

Intel® Arria 10 – 1150 FPGA energy efficiency on Caffe/AlexNet up to 25 img/s/w with FP16 at 297MHz

Vanilla AlexNet Classification Implementation as specified by http://www.cs.toronto.edu/~fritz/absps/imagenet.pdf, Training Parameters taken from Caffe open-source Framework are 224x224x3 Input, 1000x1 Output, FP16 with Shared Block-Exponents, All compute layers (incl. Fully Connected) done on the FPGA except for Softmax, Arria 10-1150 FPGA, -1 Speed Grade on Altera PCIe DevKit with x72 DDR4 @ 1333 MHz, Power measured through on-board power monitor (FPGA POWER ONLY), ACDS 16.1 Internal Builds + OpenCL SDK 16.1 Internal Build, Compute machine is an HP Z620 Workstation, Xeon E5-1660 at 3.3 GHz with 32GB RAM. The Xeon is not used for compute.