
AIDC
INTEL AI DEVCON 2018

AIDC
INTEL AI DEVCON 2018

COST EFFECTIVE MODEL DEPLOYMENT USING INTEL DEEP LEARNING DEPLOYMENT TOOLKIT

HOW TO DEPLOY AI ON YOUR EXISTING CPU

Prashant Shah, Intel
Dave Chevalier, GE Healthcare - CT imaging

23-May-2018

intel AI

DEEP LEARNING @ GE HEALTHCARE CT

Tomo in the CAT scan machine
Credit: Grahm S. Jones/Columbus Zoo and Aquarium

COMPUTERIZED AXIAL TOMOGRAPHY

COMPUTED TOMOGRAPHY (CT)

80+ million CT exams performed in United States each year
source: IMV 2016 CT Market Outlook

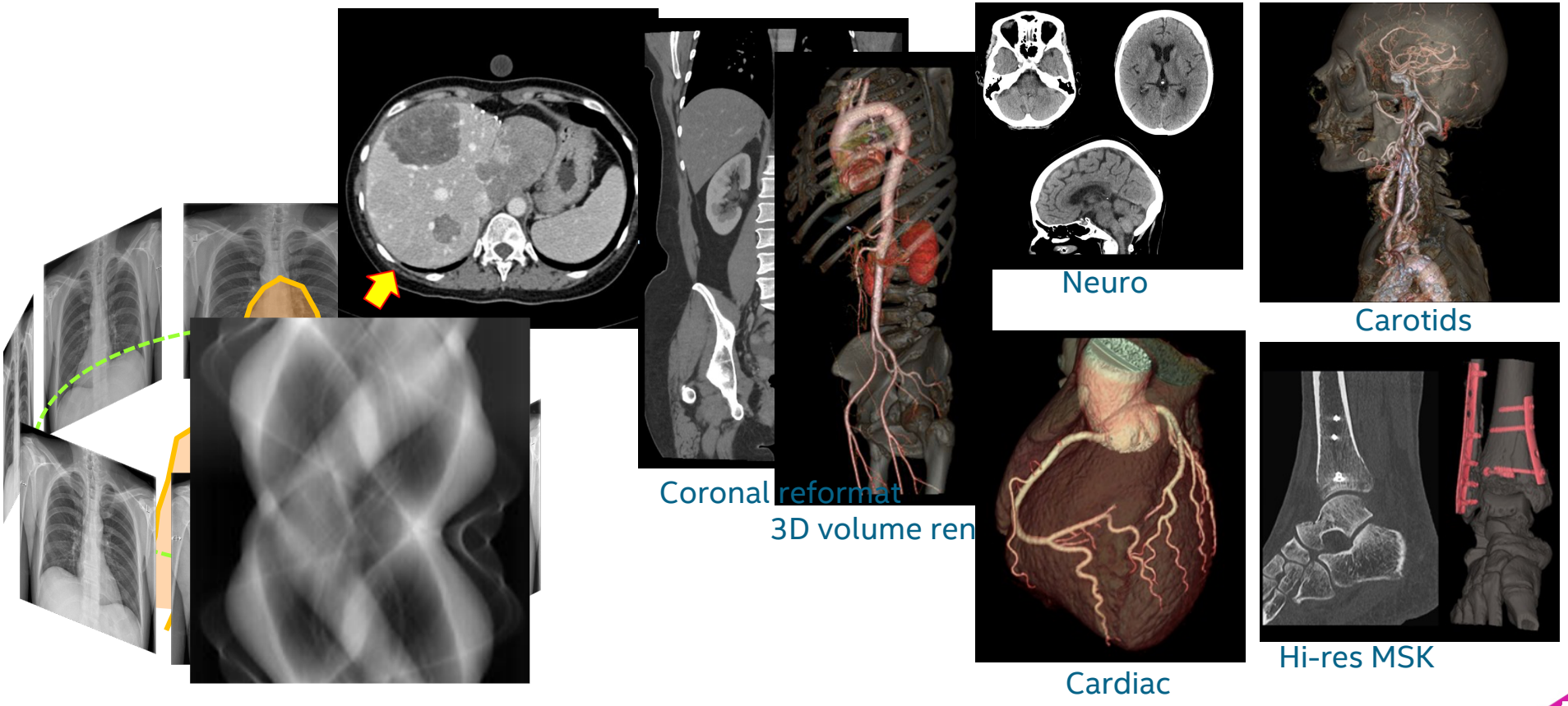
14% of Emergency Department (ED) visits result in a CT scan
source: "National Trends in Use of Computed Tomography in the Emergency Department." Annals of Emergency Medicine. Nov 2011.

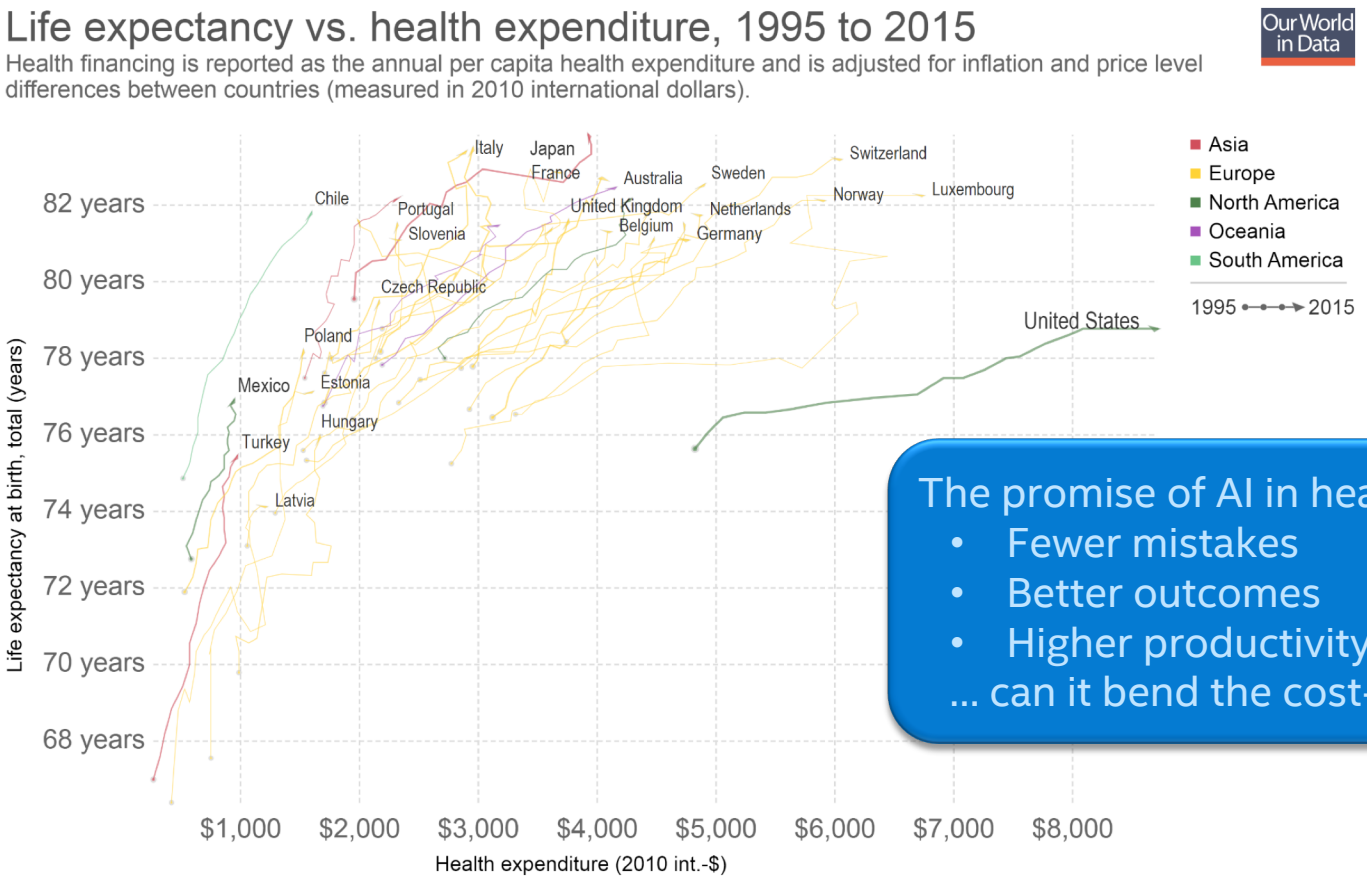
Common CT procedures

1. Abdomen/Pelvis
2. Brain
3. Neck or Spine
4. Chest
5. CT Angiography

source: IMV 2016 CT Market Outlook

PRINCIPLES OF COMPUTED TOMOGRAPHY





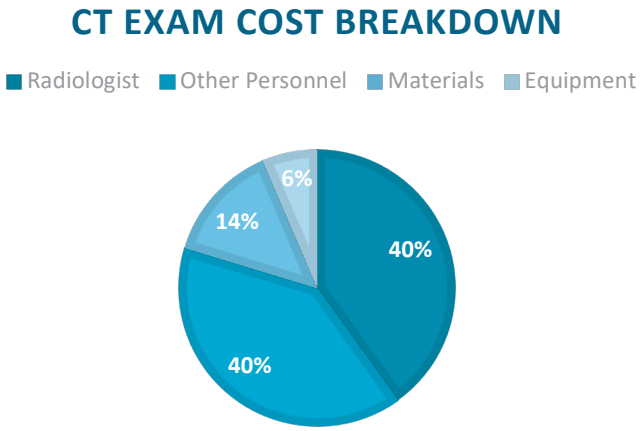
The promise of AI in healthcare –

- Fewer mistakes
- Better outcomes
- Higher productivity

... can it bend the cost-benefit slope?

Source: World Bank – WDI, Health Expenditure and Financing - OECDstat (2017)
OurWorldInData.org/the-link-between-life-expectancy-and-health-spending-us-focus • CC BY-SA

AI AND HEALTHCARE ECONOMICS



80% of the cost of a CT exam is human labor

Source: Dissecting Costs of CT Study: Application of TDABC (Time-driven Activity-based Costing) in a Tertiary Academic Center. Anzai Y, Heilbrun ME2, Haas D, Boi L, Moshre K, Minoshima S, Kaplan R, Lee VS. 2017 Feb; 24(2):200-208. doi: 10.1016/j.acra.2016.11.001. Epub 2016 Dec 14.

Productivity = cost savings

Automation of routine tasks – AI is the enabler

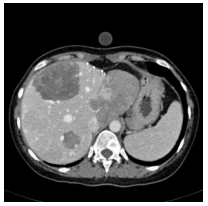
Inference deployment – keep the costs down

- Deploy AI on existing CPU
- CPU-based inference allows embedded AI across the full range of our applications and products

Reduce OpEx without increasing CapEx

AI: PUTTING THE "SEE" IN "CT"

AXIAL SLICE CLASSIFICATION

 = ABDOMEN

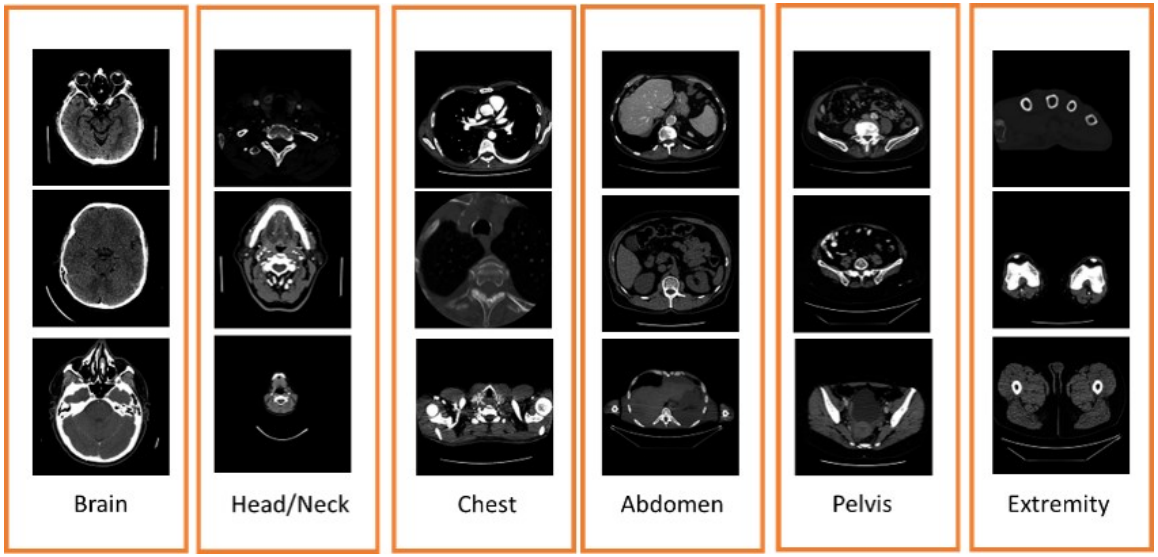
POTENTIAL USE CASES

- IMAGE QUALITY OPTIMIZATION – ANATOMY SPECIFIC TECHNIQUES
- QUALITY ASSURANCE – DID I SCAN WHAT I WANTED TO SCAN?
- POST-PROCESSING APPLICATION SELECTION
- DATA TAGGING – FOR LATER RETRIEVAL, RESEARCH, AI TRAINING
- CONTENT-BASED RETRIEVAL - COMPARE TO NORMALS

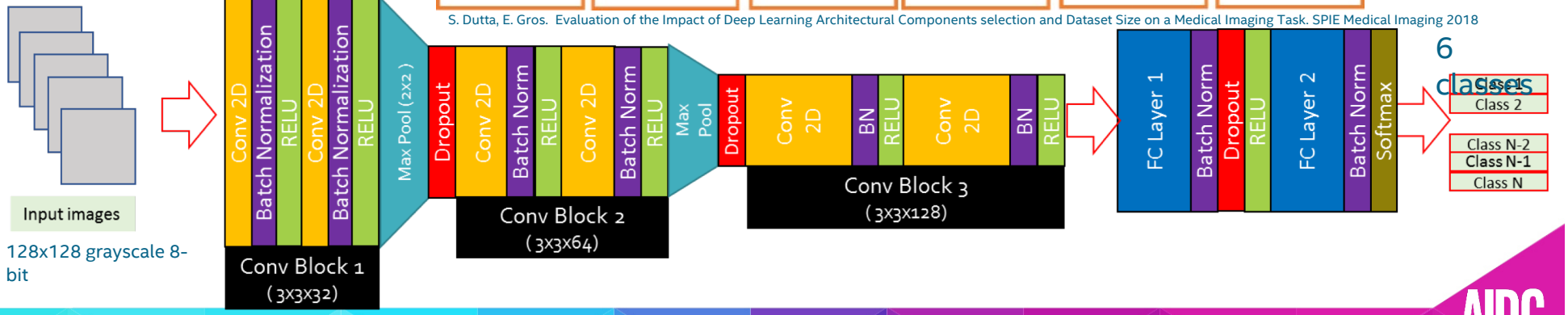
CT AXIAL IMAGE CLASSIFIER

ANATOMY DATASET:

- 223 EXAMS, ~30,000 IMAGES
- EACH AXIAL CT IMAGE LABELED INTO 6 ANATOMIC REGIONS

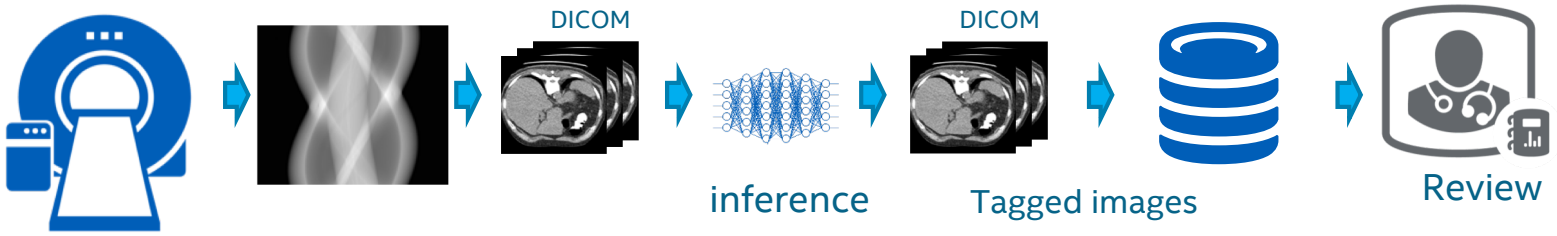


S. Dutta, E. Gros. Evaluation of the Impact of Deep Learning Architectural Components selection and Dataset Size on a Medical Imaging Task. SPIE Medical Imaging 2018



INFERENCE PERFORMANCE GOAL

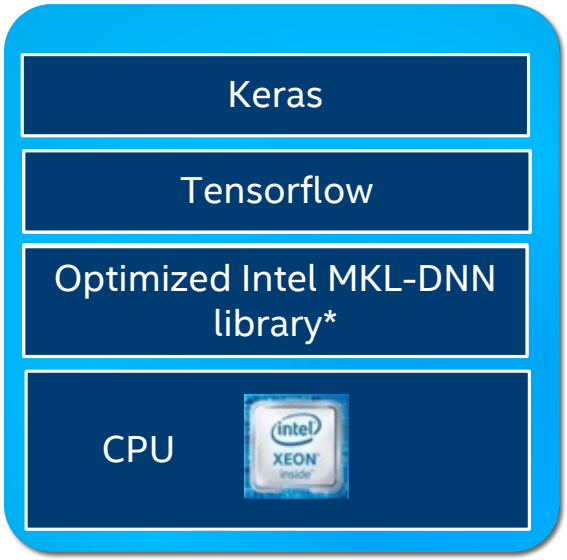
Clinical Goal: Keep pace with the imaging pipeline

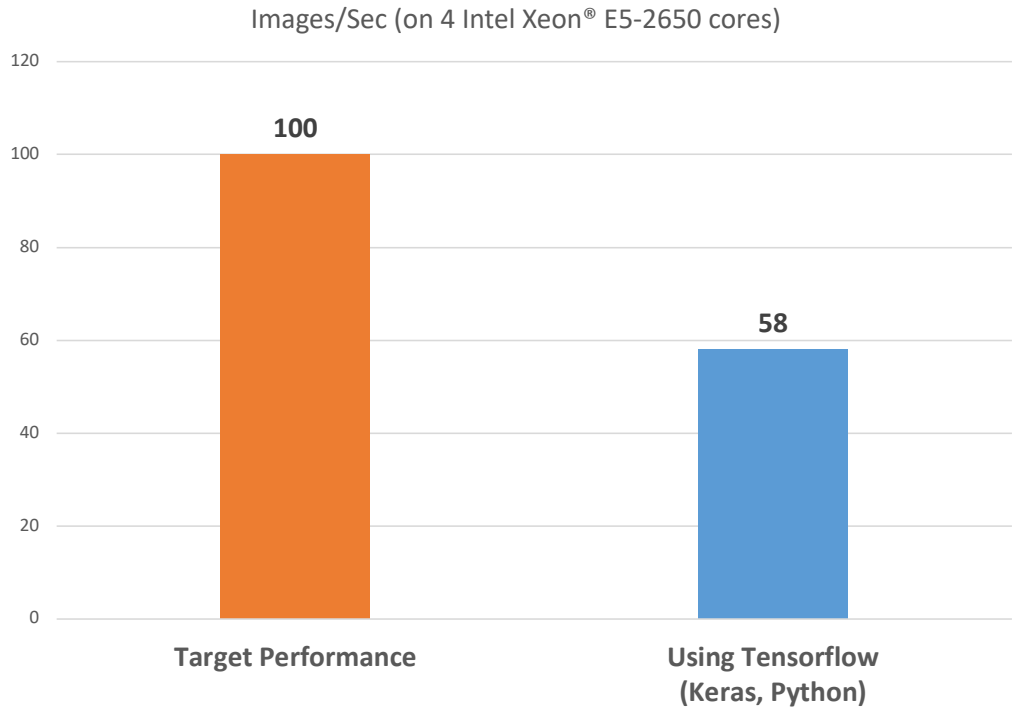


Performance target: 100 images/sec or ~10msec latency

– up to 4 cores available without impacting the imaging pipeline

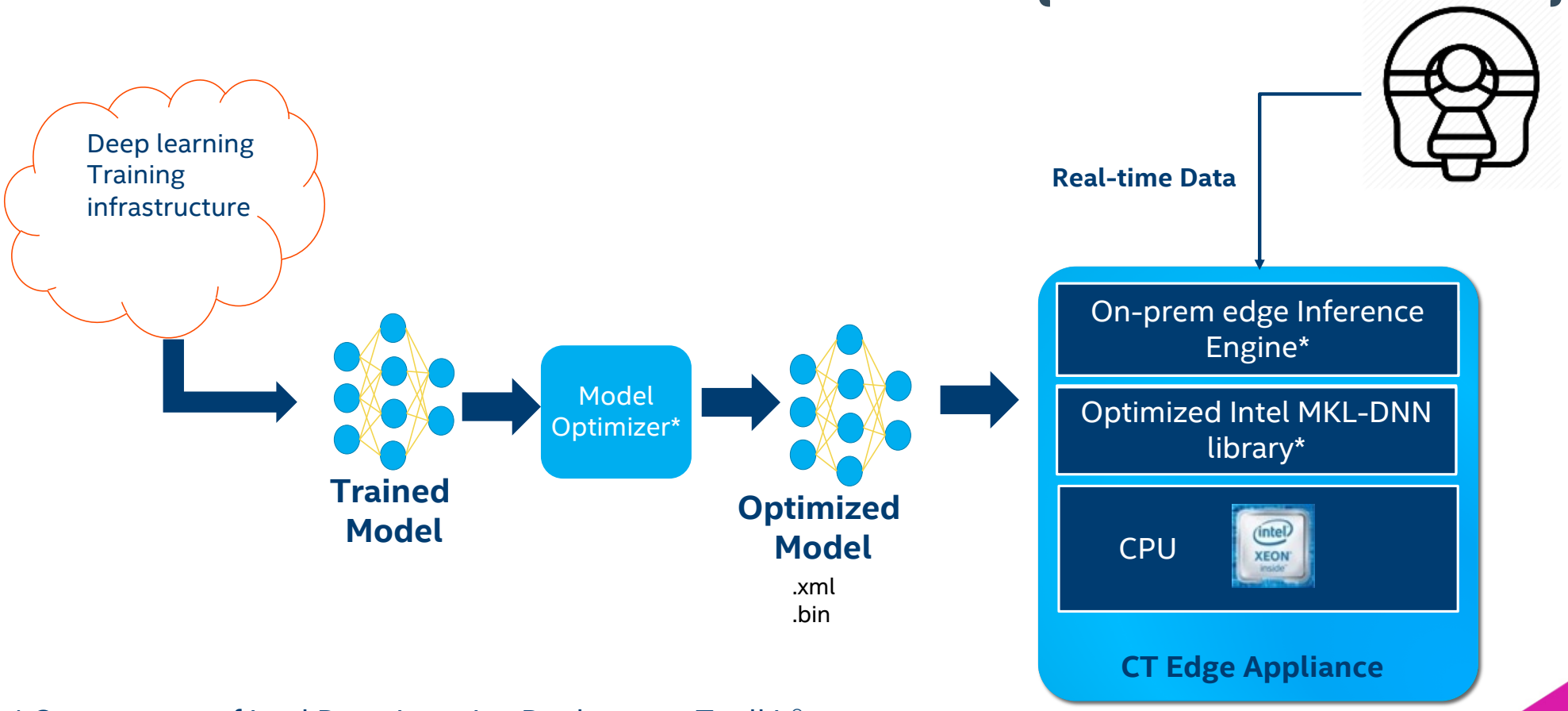
BASELINE BENCHMARKING WITH TENSORFLOW





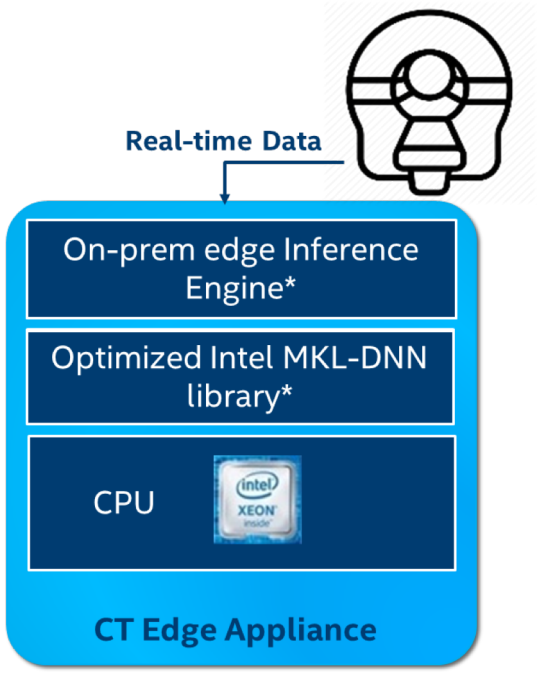
Configuration Details:
® Xeon® processor E5-2650 v4 at 2.20GHz, and configured with 264 GB of memory, Intel® Solid State Drive Data Center 480 GB, and CentOS Linux* 7.4.1708; Tensorflow version 1.4 Compiled with MKL-DNN. Model and Dataset: GE proprietary

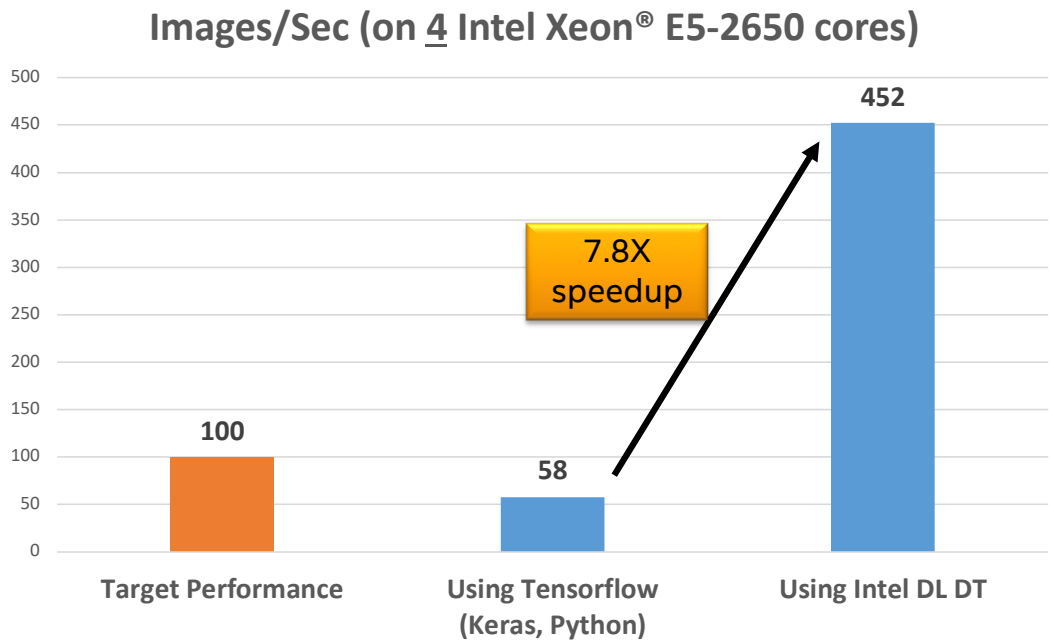
SOLUTION – INTEL DEEP LEARNING DEPLOYMENT TOOLKIT® (PART OF INTEL® OPEN VINO™)



* Components of Intel Deep Learning Deployment Toolkit®

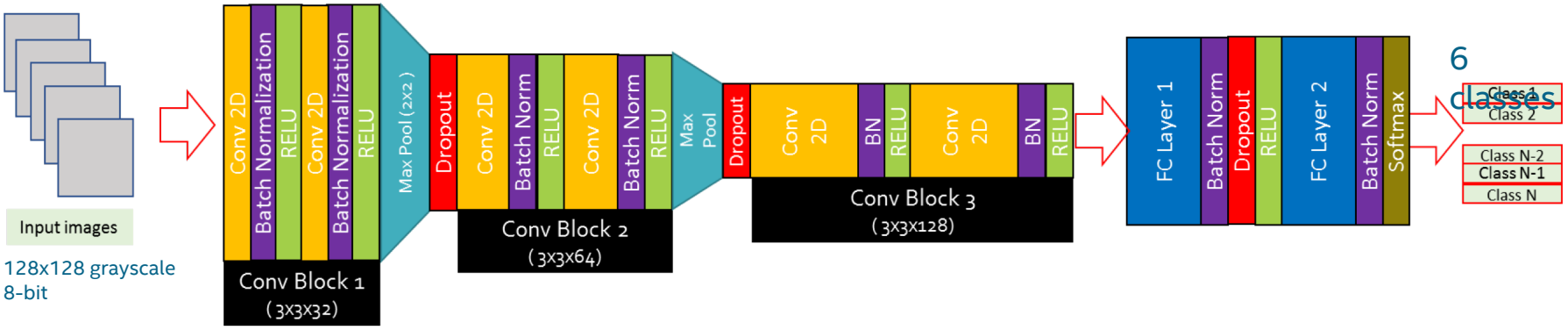
BASELINE BENCHMARKING WITH INTEL® DEEP LEARNING DEPLOYMENT TOOLKIT® (PART OF INTEL OPEN VINO™)

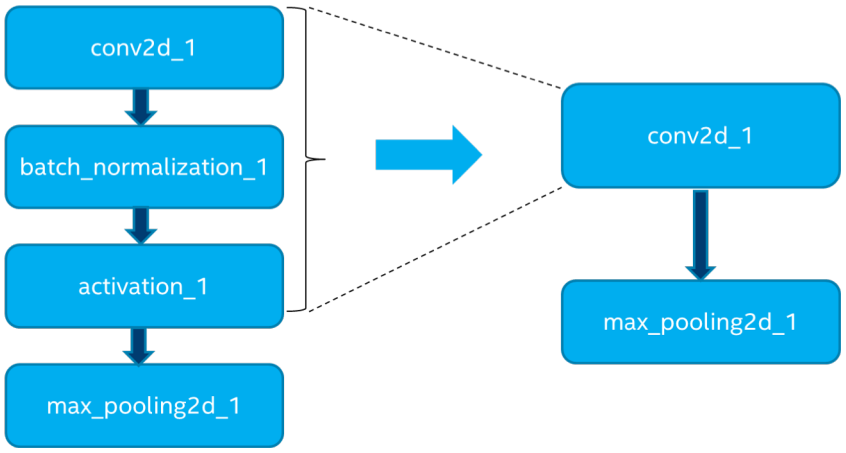




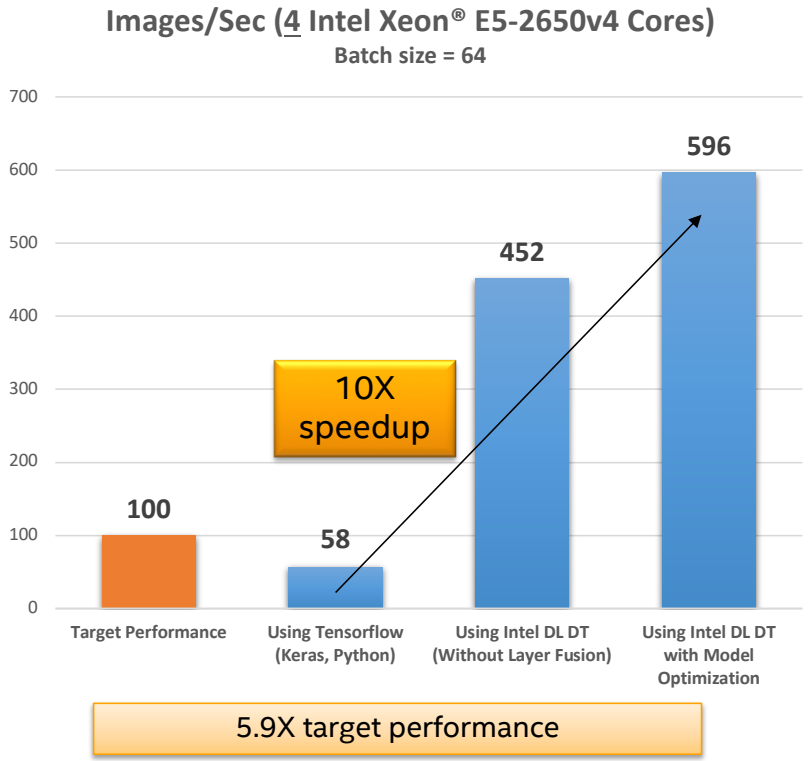
Configuration Details:
® Xeon® processor E5-2650 v4 at 2.20GHz, and configured with 264 GB of memory, Intel® Solid State Drive Data Center 480 GB, and CentOS Linux* 7.4.1708; Tensorflow version 1.4 Compiled with MKL-DNN. Model and Dataset: GE proprietary

OPTIMIZATIONS DONE ON THE GE AXIAL CT MODEL:

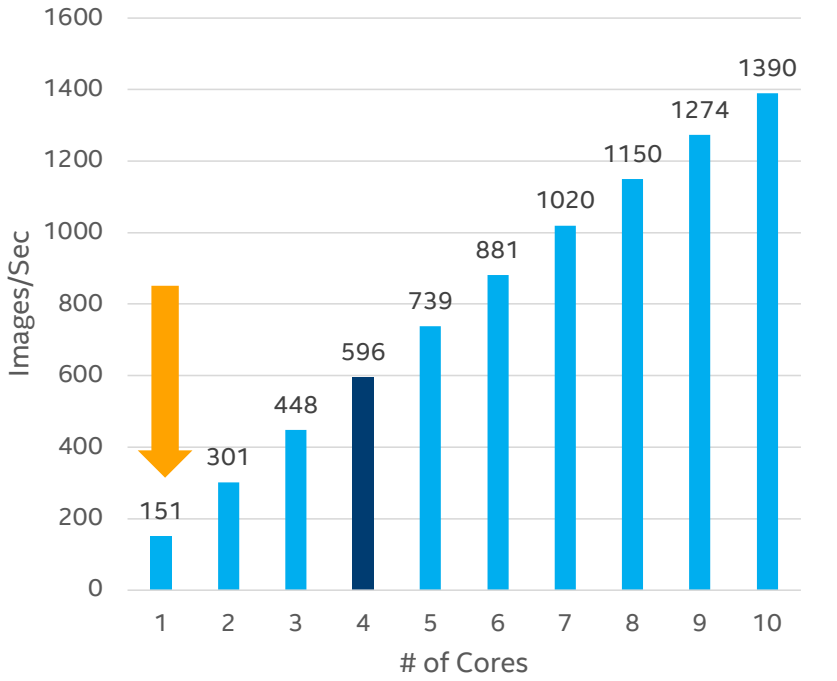




RESULTS



5.9X target performance



Exceeds GE's requirement with just 1 Intel® Xeon® core

Configuration Details:
® Xeon® processor E5-2650 v4 at 2.20GHz, and configured with 264 GB of memory, Intel® Solid State Drive Data Center 480 GB, and CentOS Linux* 7.4.1708; Tensorflow version 1.4 Compiled with MKL-DNN. Model and Dataset: GE proprietary

DEEP LEARNING DEPLOYMENT TOOLKIT (PART OF INTEL® OPEN VINO™)

Train

Train a DL model

TensorFlow

Caffe

mxnet

Prepare Optimize

Model Optimizer
-Convert
-Optimize
- node merging
- BN elimination
- Constant folding
- Horizontal fusion

-Prepare to inference

(device agnostic, Generic optimization)

.prototxt
.caffemodel

Run Model Optimizer

IR
.xml
.bin

Inference

Inference-Engine
a lightweight API
To use in your application for inference

User Application

Inference Engine

Optimize/ Hetero

Inference-Engine
Support multiple devices
For heterogeneous flows

Device level optimization

MKL-DNN → CPU: Xeon/Core/Atom

clDNN → GPU

DLA → FPGA

Myriad → Movidius Myriad 2/X

Extend

Inference-Engine
Support extensibility
And allow custom
Kernels for various
devices

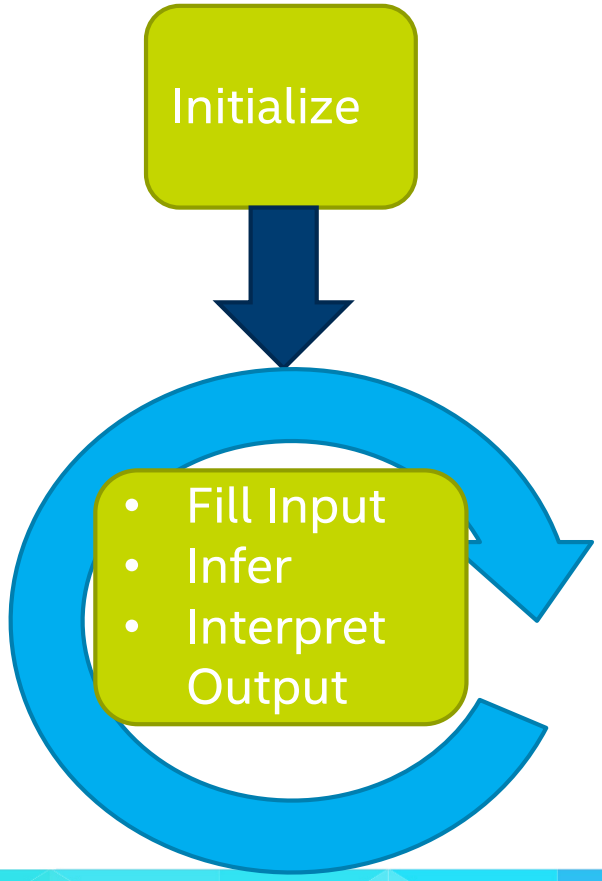
Extensibility **C++**

Extensibility **OpenCL**

Extensibility **OpenCL/TBD**

Extensibility **TBD**

DEEP LEARNING DEPLOYMENT TOOLKIT INFERENCE WORKFLOW



Initialization

Load model and weights
Set batch size (if needed)
Load Inference Plugin (CPU, GPU, FPGA)
Load network to plugin
Allocate input, output buffers

Main loop

Fill input buffer with data
Run inference
Interpret output results

AI PORTFOLIO

SOLUTIONS

Data Scientists | Technical Services | Reference Solutions

PLATFORMS

Intel® AI DevCloud

Intel® Deep Learning System‡

Intel Saffron™ REASONING

TOOLS

Intel® Deep Learning Studio‡

Intel® Deep Learning Deployment Toolkit†

Intel® Open Vino ™ (Formerly Computer Vision SDK)

Intel® Movidius™ Software Development Kit (SDK)

FRAMEWORKS

TensorFlow* Caffe* mxnet* BigDL on Spark* neon Caffe2*‡ PYTORCH*‡ Microsoft CNTK*‡ PaddlePaddle*‡

LIBRARIES

Intel® MKL/MKL-DNN, clDNN, DAAL, Intel Python Distribution, etc.

Intel® nGraph™ Compiler^α

CPU Transformer† | NNP Transformer‡ | Other

TECHNOLOGY

Xeon | Nervana‡ | Arria 10 | Atom | Core i7 | Iris Graphics | Movidius | Altera

END-TO-END COMPUTE

SYSTEMS & COMPONENTS

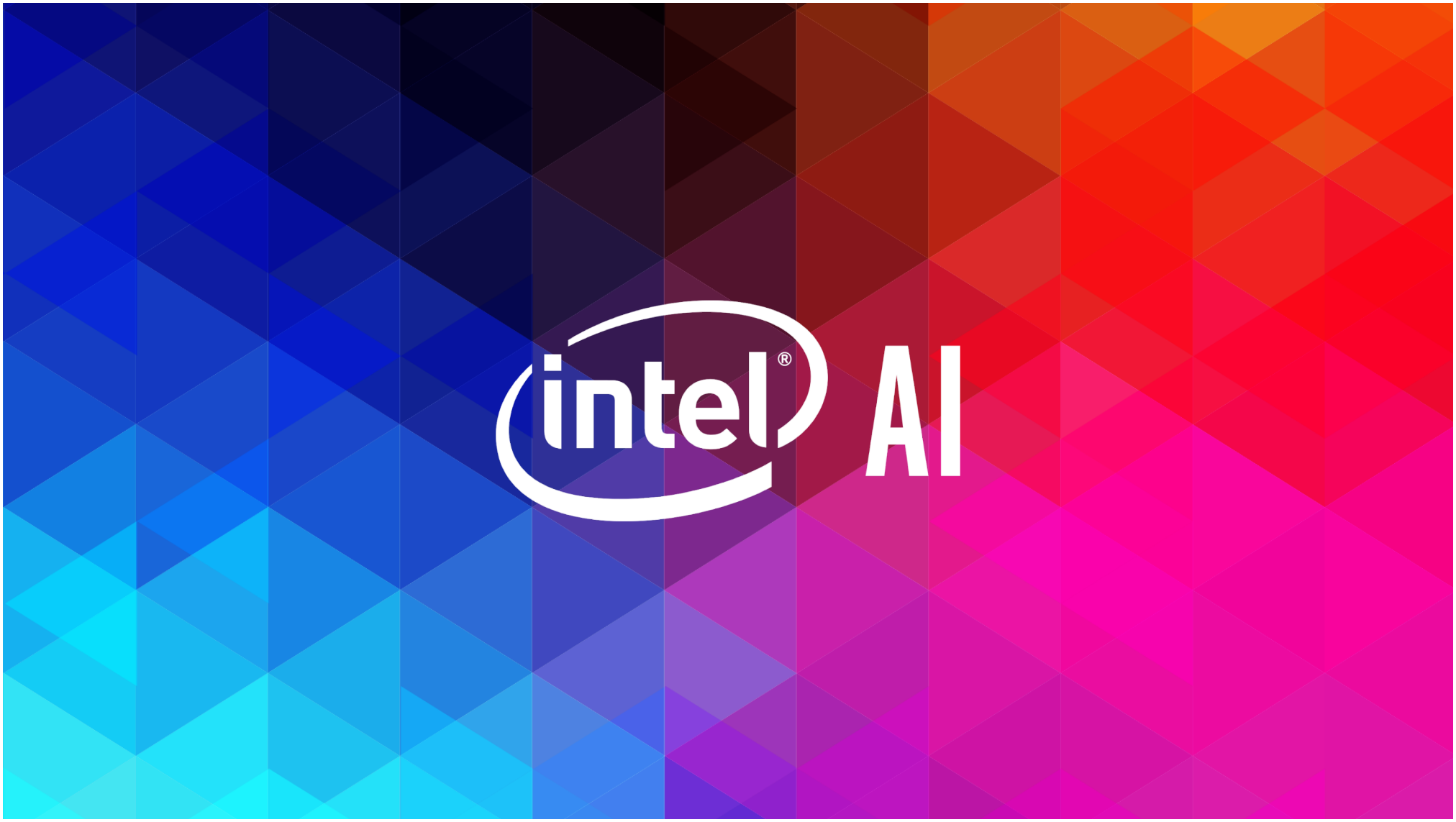
^α Alpha available
† Beta available
‡ Future
*Other names and brands may be claimed as the property of others.
All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

AIDC INTEL AI DEVCON 2018

SUMMARY

- Using Intel Deep Learning Deployment Toolkit®, we were able to achieve **6X the target performance**, and **10X the performance over native Tensorflow** without the need of add-in cards.
- Inference on CPUs provides GE the flexibility of deployment architecture. CPUs are ubiquitous in embedded devices, workstations and datacenter/cloud.
- DL DT can be used to efficiently deploy models on CPUs, Intel Integrated GPUs, FPGAs and Movidius inference accelerators.


AIDC
INTEL AI DEVCON 2018
Q&A
intel AI


intel
AI