
AIDC
INTEL AI DEVCON 2018

AIDC
INTEL AI DEVCON 2018

ARTIFICIAL INTELLIGENCE
CASE STUDY

Hongwei Yi

2018/08/08

intel AI

LEGAL NOTICES & DISCLAIMERS

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.

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.

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.

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†

OpenVINO™ Toolkit

Intel® Movidius™ Software Development Kit (SDK)

FRAMEWORKS

TensorFlow* Caffe* mxnet* BigDL ON Spark* Caffe2*‡ PYTORCH*‡ CNTK*‡ PaddlePaddle*‡

LIBRARIES

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

DIRECT OPTIMIZATION

Intel® nGraph™ Compiler[α]

CPU Transformer† | NNP Transformer‡ | Other

TECHNOLOGY

XEON | NERVANA‡ | ARRIA 10 | ATOM | CORE i7 | Iris Graphics | Movidius

END-TO-END COMPUTE

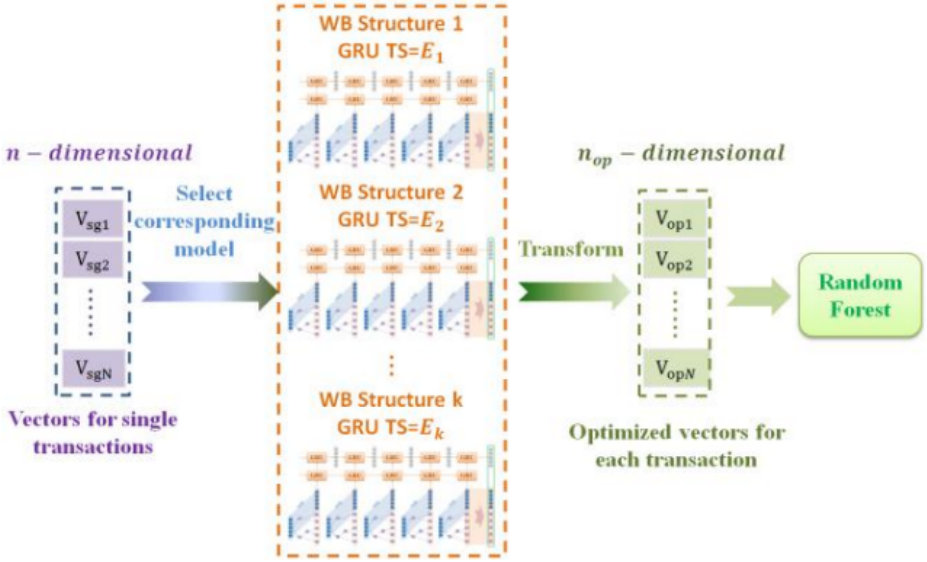
SYSTEMS & COMPONENTS

[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

CASE 1: TRANSACTION FRAUD DETECTION WITH DEEP LEARNING & MACHINE LEARNING

TRANSACTION FRAUD DETECTION WITH DEEP LEARNING & MACHINE LEARNING

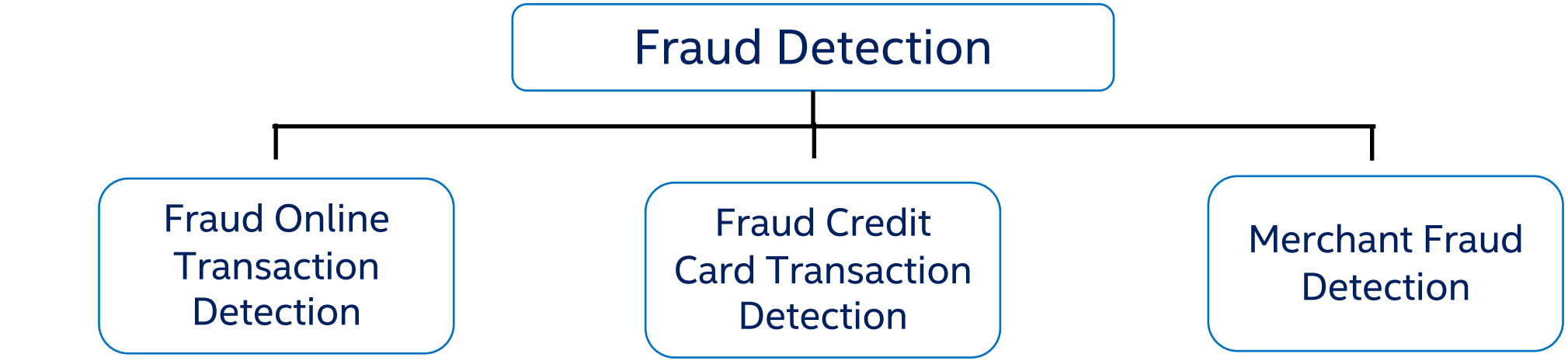


Problem: no single algorithm delivers high-enough accuracy

Solution: ensemble models of classical machine learning with deep learning achieved industry recorded high detection accuracy

The COMPLETE End-to-End detection runs on Intel® Xeon® Scalable Processor

THE FRAUD CREDIT CARD DETECTION USAGE OF POC



- Protect online payment from identity stolen
- ~60M transactions per day for UnionPay. Detection is offline.

BigDL adopted in 2016

- Detect transactions using fake/clone cards through POS, ATM, online payment
- ~300M transactions per day for UnionPay. Real time detection is required

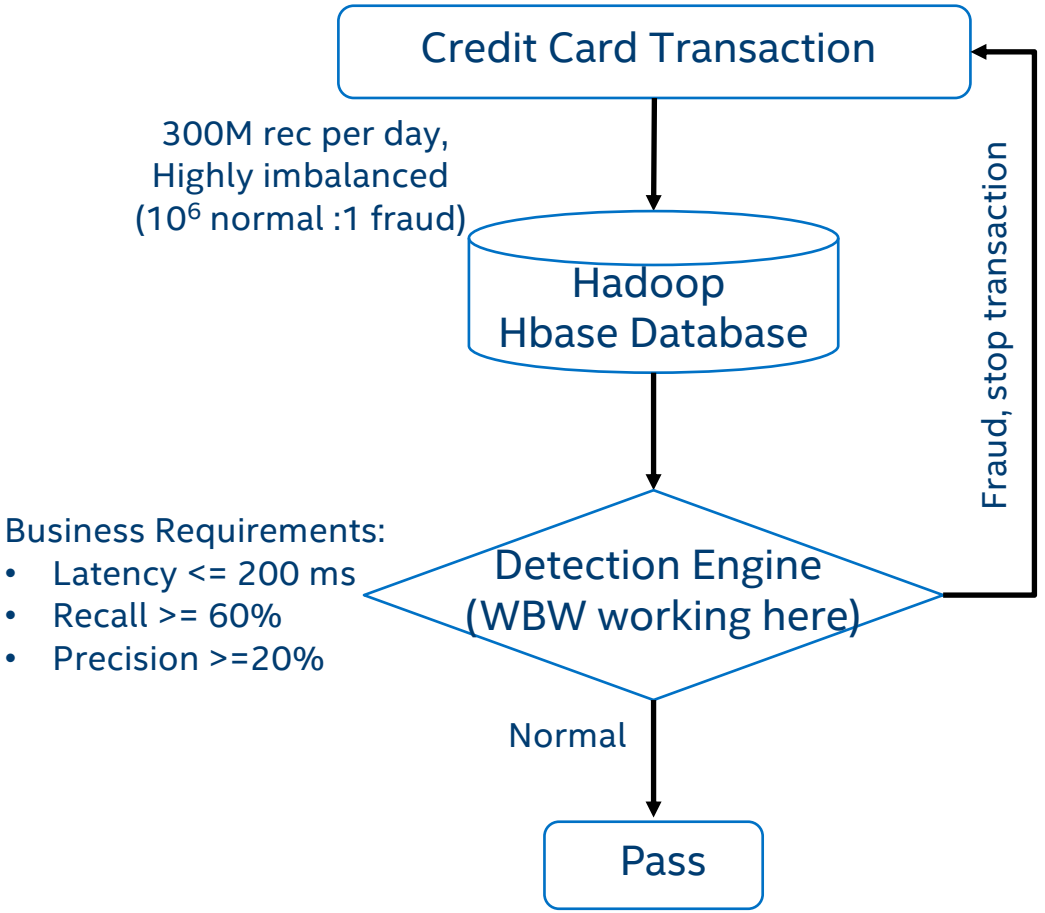
2017 focus

- Detection for fake merchants, money laundering, illegal cash out, fraud compensation, malicious collapse, ...

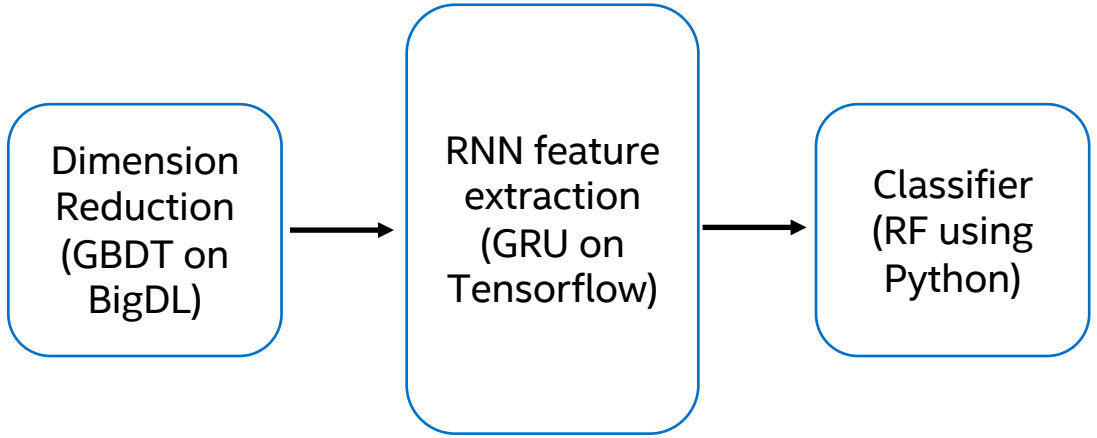
Future Opportunity

THE WITHIN-BETWEEN-WITHIN(WBW) SANDWICH-STRUCTURED SEQUENCE LEARNING ARCHITECTURE

Fraud Credit Card Transaction Detection Flow Diagram



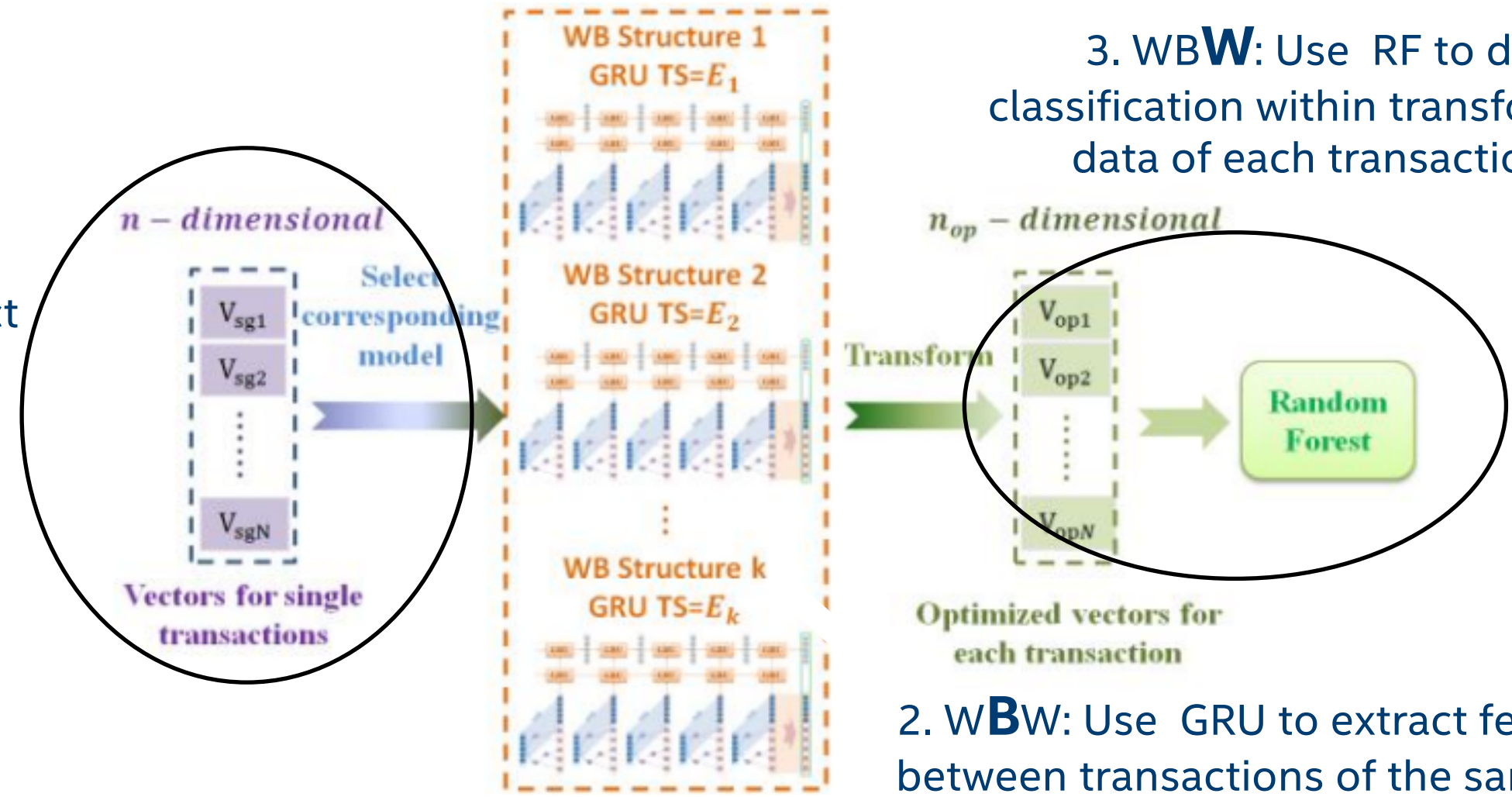
The Within-Between-Within(WBW) Architecture



- WBW model achieved the record high >25% precision on >60% recall rate, with latency of <200 ms
- GBDT & RF don't run well on GPU due to irregular computing in the algorithms

"WBW" (WITHIN-BETWEEN-WITHIN) SANDWICH-STRUCTURED SEQUENCE LEARNING ARCHITECTURE

1. **W**BW: Use GBDT to extract data within each single transaction



3. WB**W**: Use RF to do classification within transformed data of each transaction

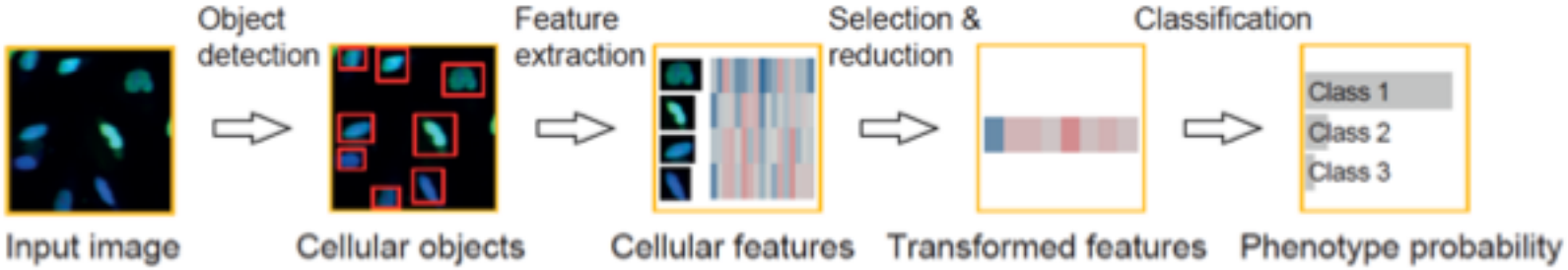
2. W**B**W: Use GRU to extract features between transactions of the same a/c

CASE 2 - HIGH-CONTENT IMAGING IN DRUG DISCOVERY

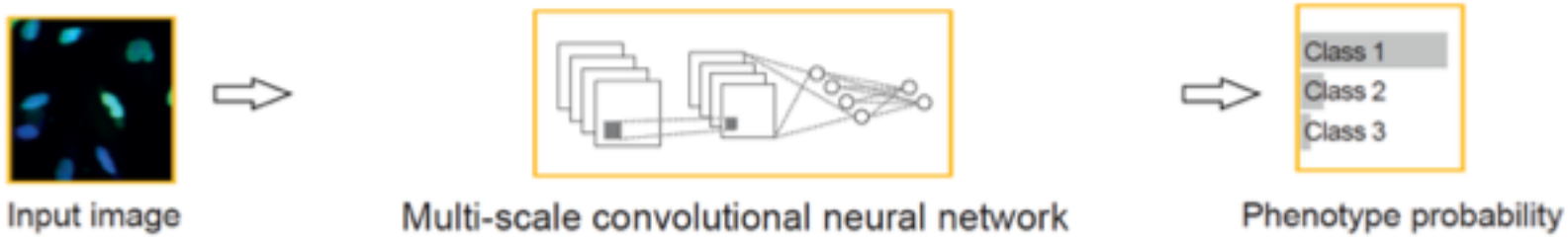
HIGH-CONTENT IMAGING (HCI) IN DRUG DISCOVERY

High-content imaging (HCI) or screening (HCS) has seen increased application in systems biology & drug discovery. Images acquired through microscopy-based assays provide visual information to investigate cellular phenotypes induced by genetic or chemical treatments.

Conventional HCS analysis pipeline



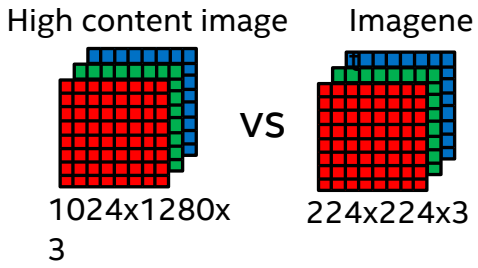
HCS analysis using Multi-Scale Convolutional Network[1]



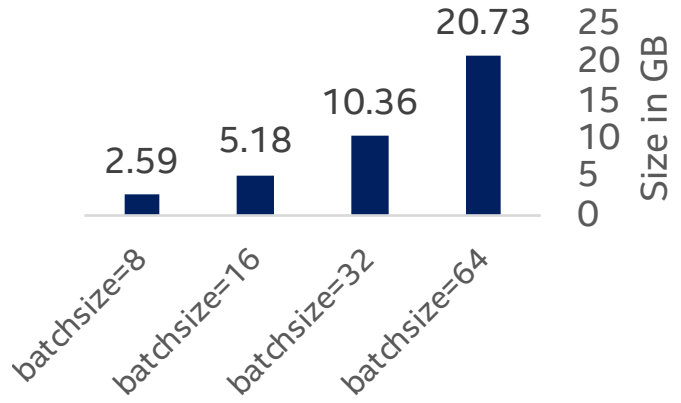
[1] Godinez et al, A multi-scale convolutional neural network for phenotyping high-content cellular images. Bioinformatics, 2017

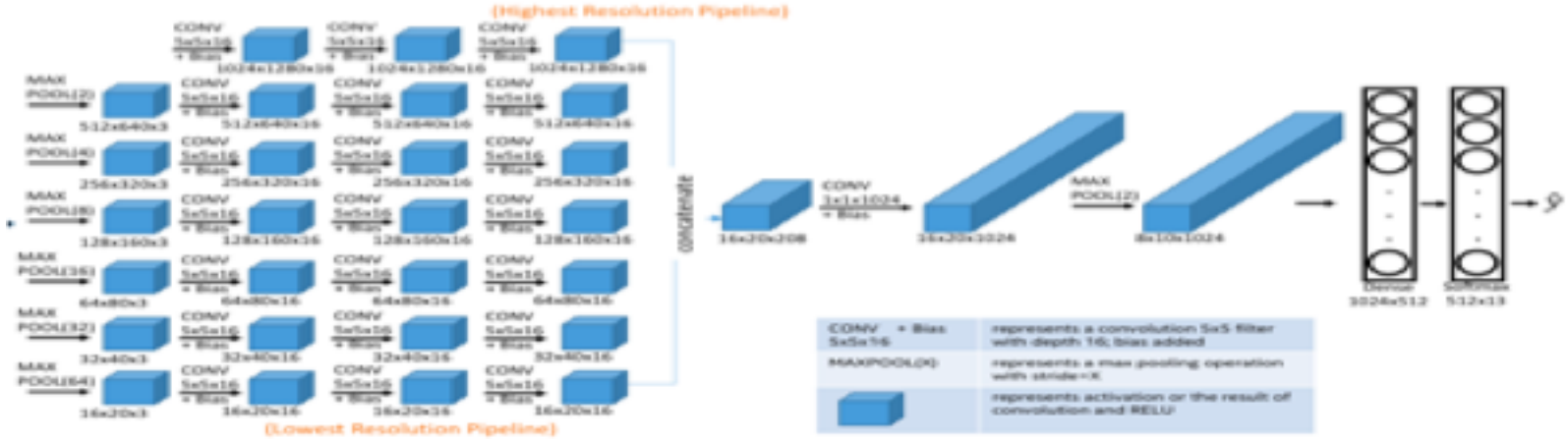
MCNN HAS A LARGE MEMORY FOOTPRINT

High content images are 26x larger than images from ImageNet dataset



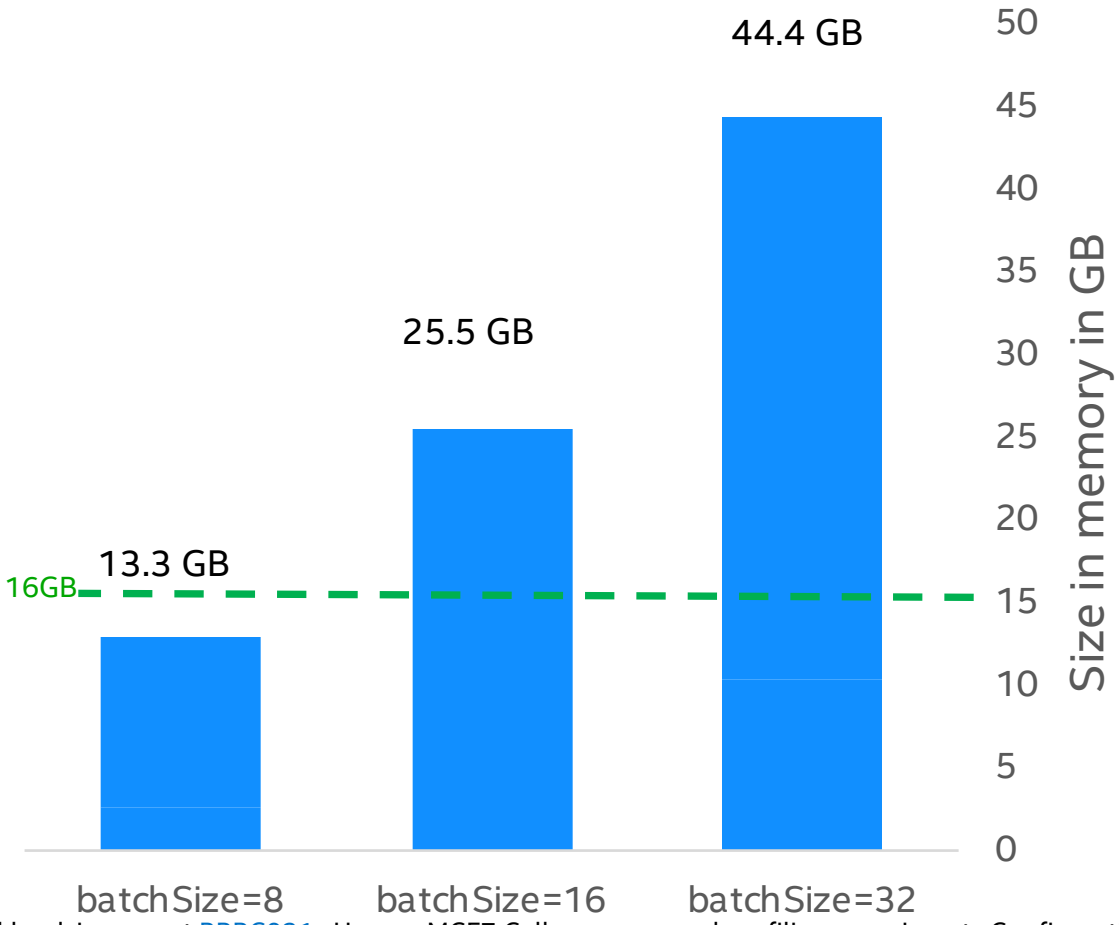
MCNN size grows linearly with batch size used in each step of training



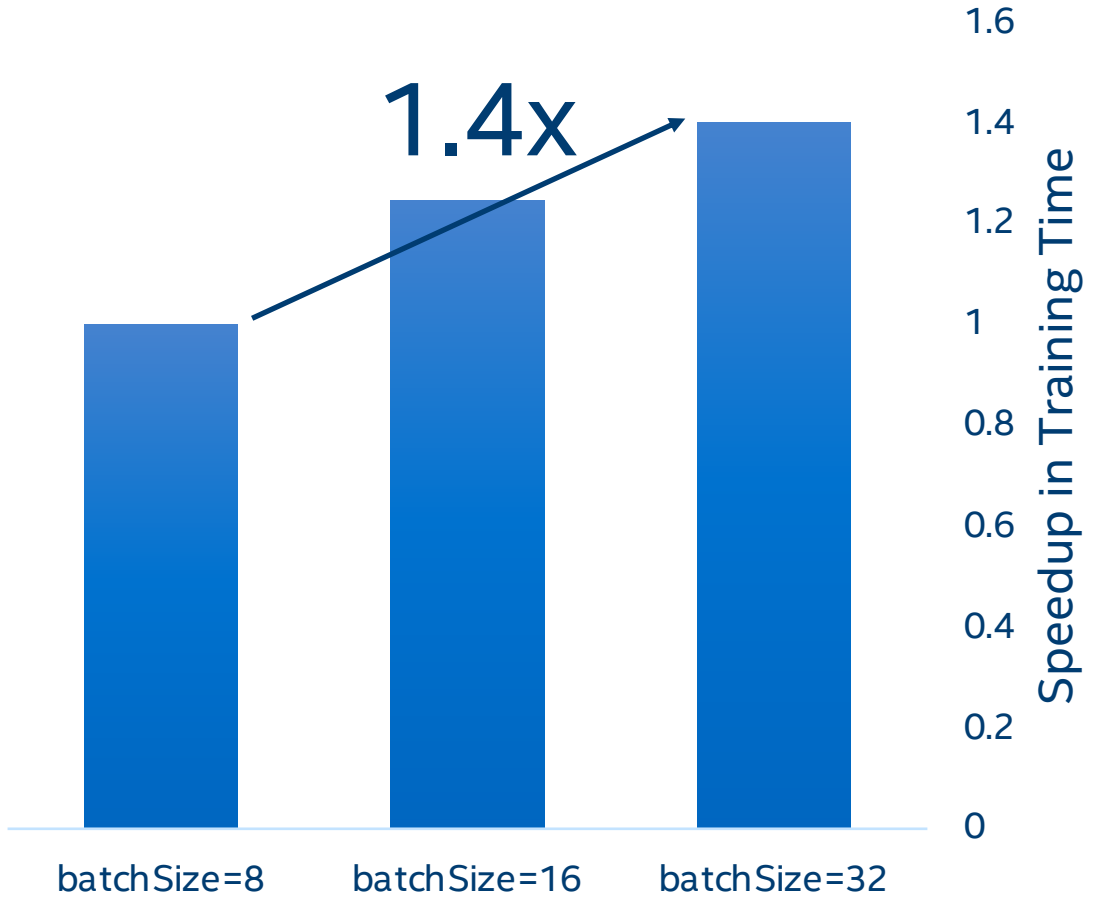


INTEL XEON® ENABLES LARGE BATCH TRAINING WITH LARGE MEMORY

Peak memory utilization in MCNN training can scale well beyond 16GB



Training Time Improves with large batches

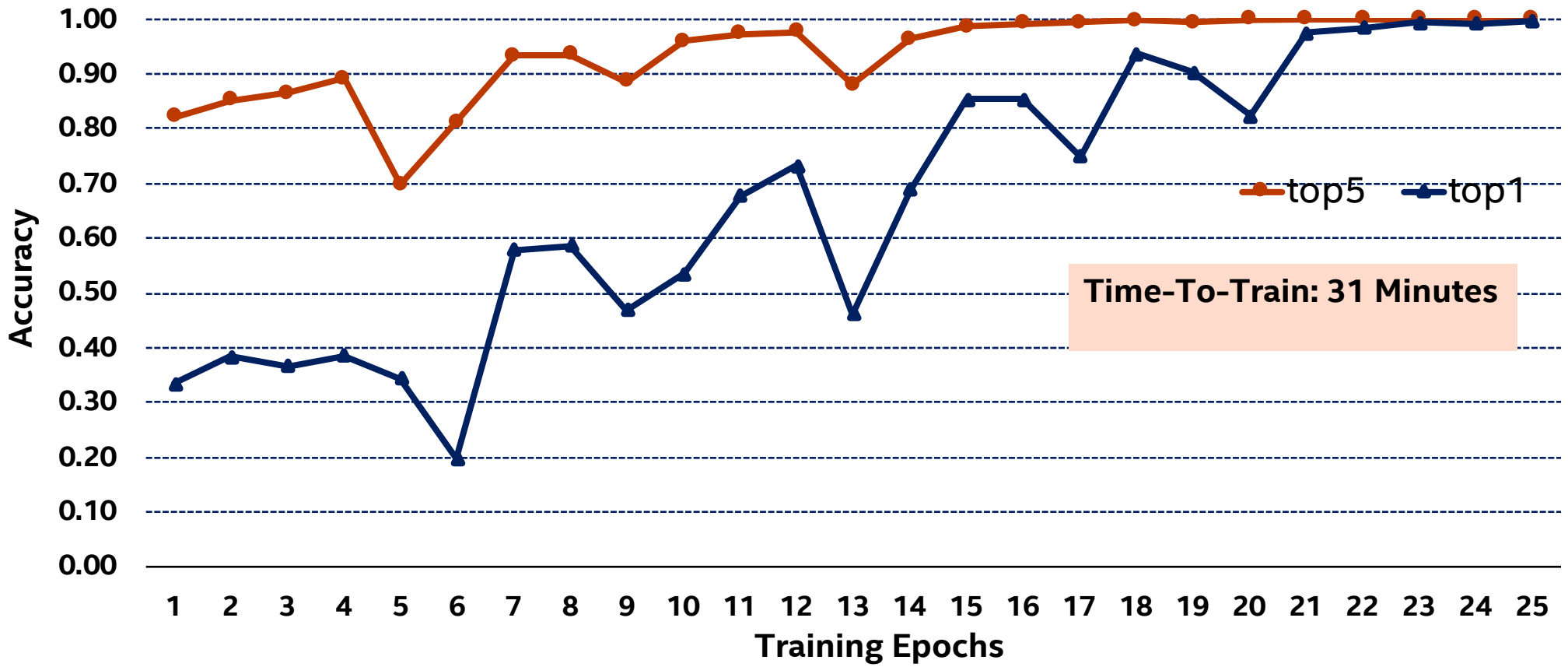


[1] Workload: Image set BBBC021 : Human MCF7 Cells – compound profiling experiment. Configuration details in backup

[2] Godinez et al, A multi-scale convolutional neural network for phenotyping high-content cellular images. Bioinformatics, 2017

LARGE BATCH TRAINING M-CNN REACHES SOTA ON 8 XEON® SERVERS

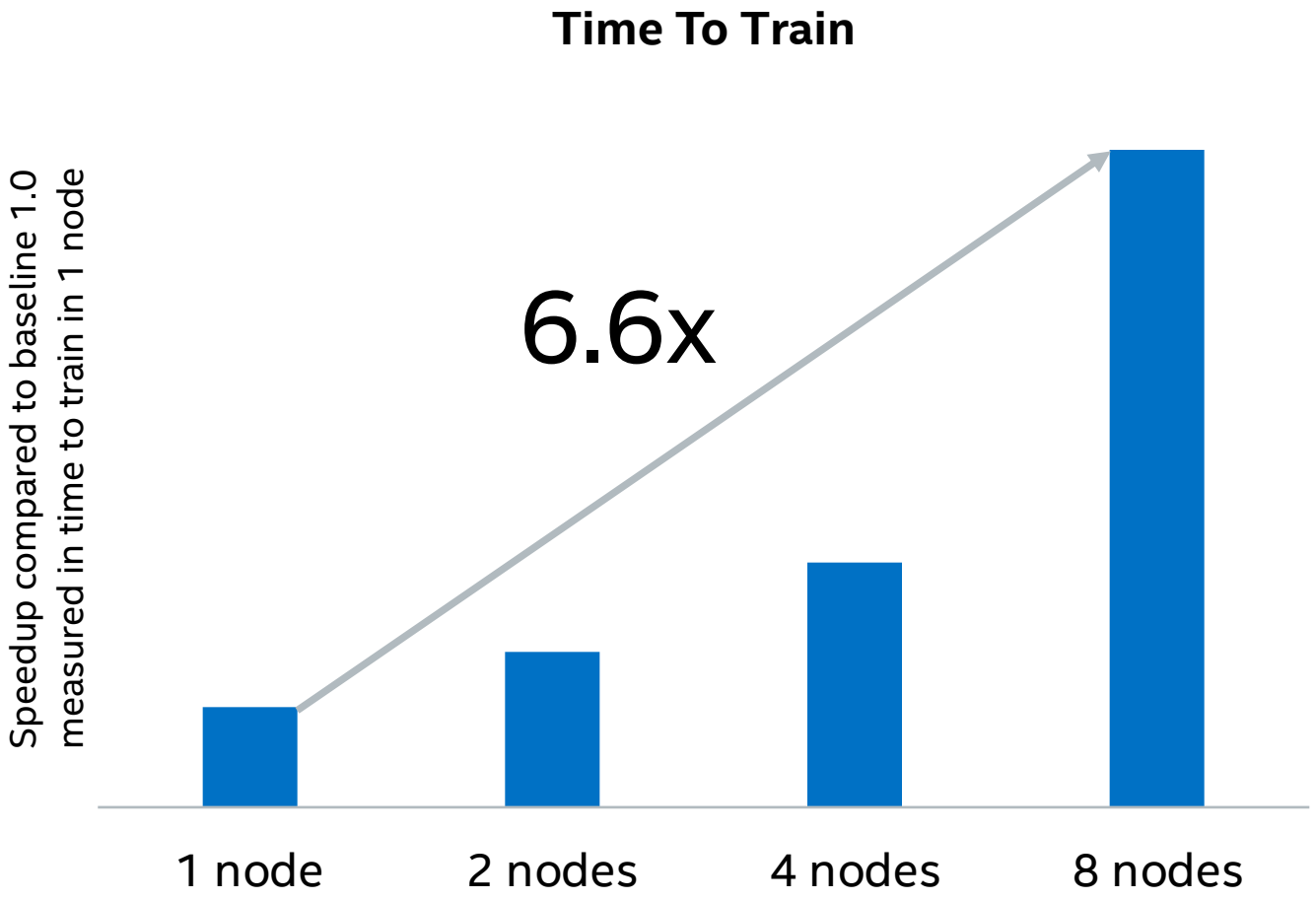
High Content Screening/M-CNN Training on 8 Node Intel® 2S Xeon® 6148 processor cluster
TensorFlow 1.7, Horovod, OpenMPI, BS=32/Node, GBS=256, OPA Fabric

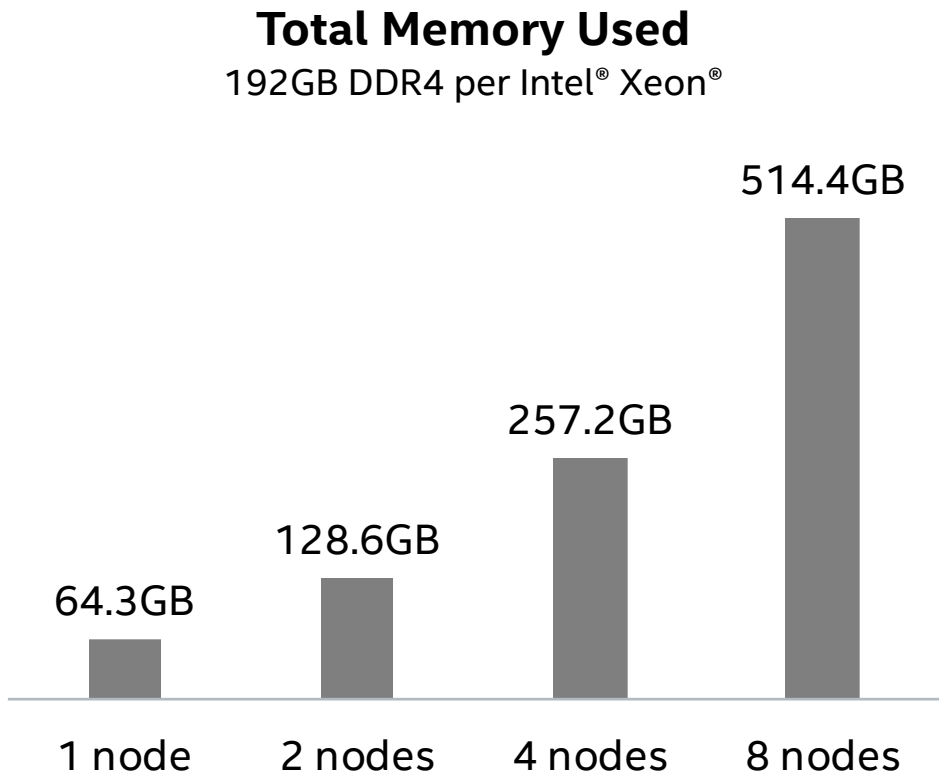


Workload: Image set BBBC021 : Human MCF7 Cells – compound profiling experiment.

TensorFlow: 1.7.0, Python: 2.7.5, Horovod: 0.12.1: OMP_NUM_THREADS=10

HIGH PERFORMANCE AT SCALE WITH INTEL® XEON® SCALABLE PROCESSOR





Workload: Image set BBBC021 : Human MCF7 Cells – compound profiling experiment.

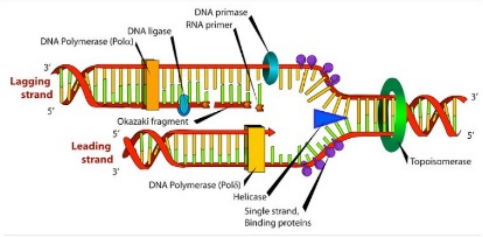
TensorFlow: 1.7.0, Python: 2.7.5, Horovod: 0.12.1, OpenMPI 3.0.0

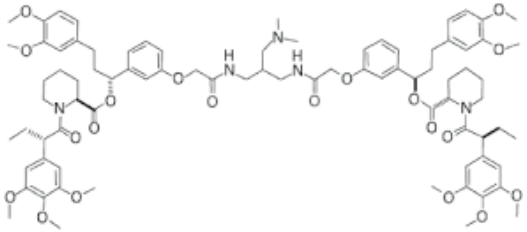
CASE 3: DEEP LEARNING FOR DRUG DISCOVERY

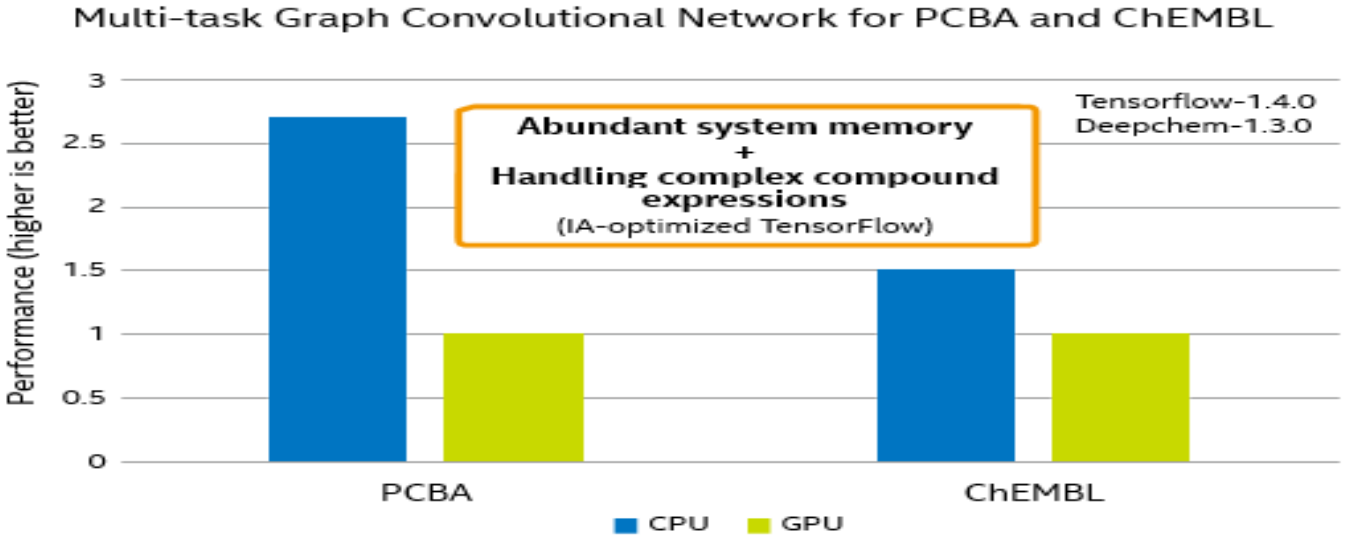
DEEP LEARNING FOR DRUG DISCOVERY

Problem: Need high efficiency platform for clinical genome analysis on DeepChem Topologies: GCN

Solution: Optimized data feeding and computing architecture on Xeon Processor Scalable Family faster than GPGPU







Drug discovery (predicting compound-protein interactions)

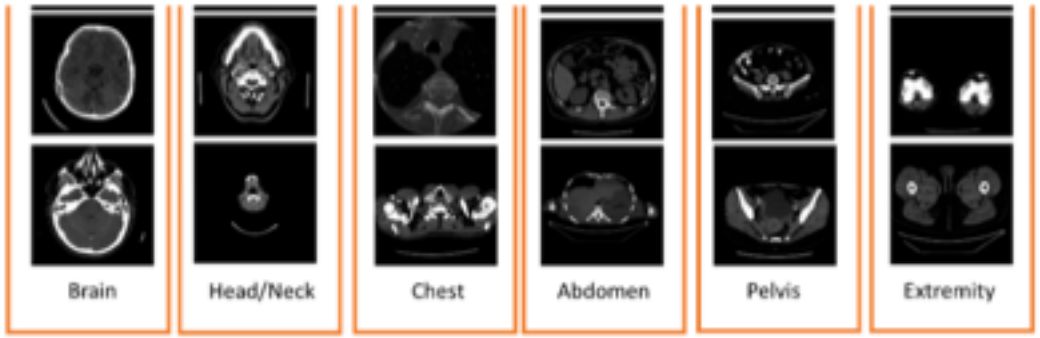
CASE 4: DIAGNOSTIC IMAGING – BODY PART CLASSIFICATION IN CT SCANS

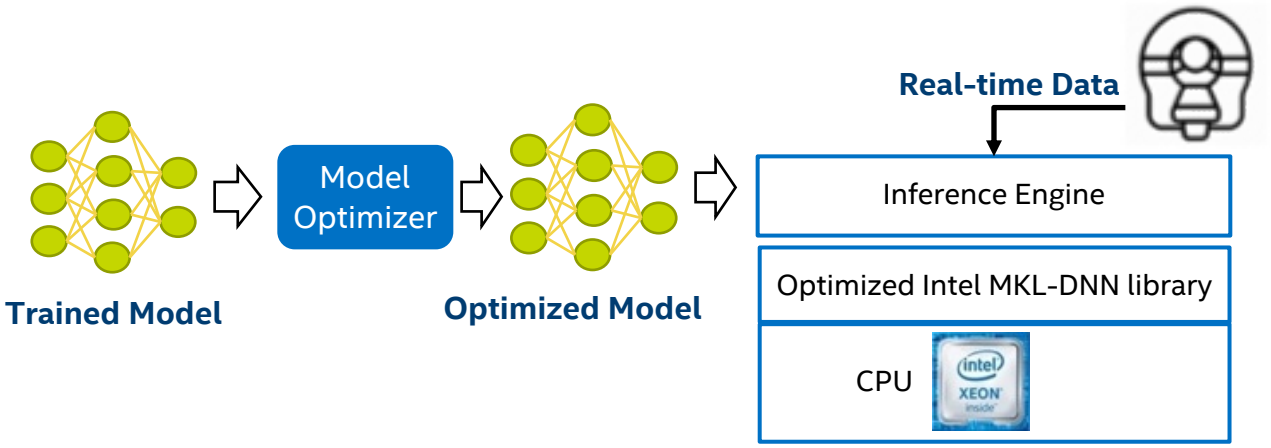
DIAGNOSTIC IMAGING – BODY PART CLASSIFICATION IN CT SCANS

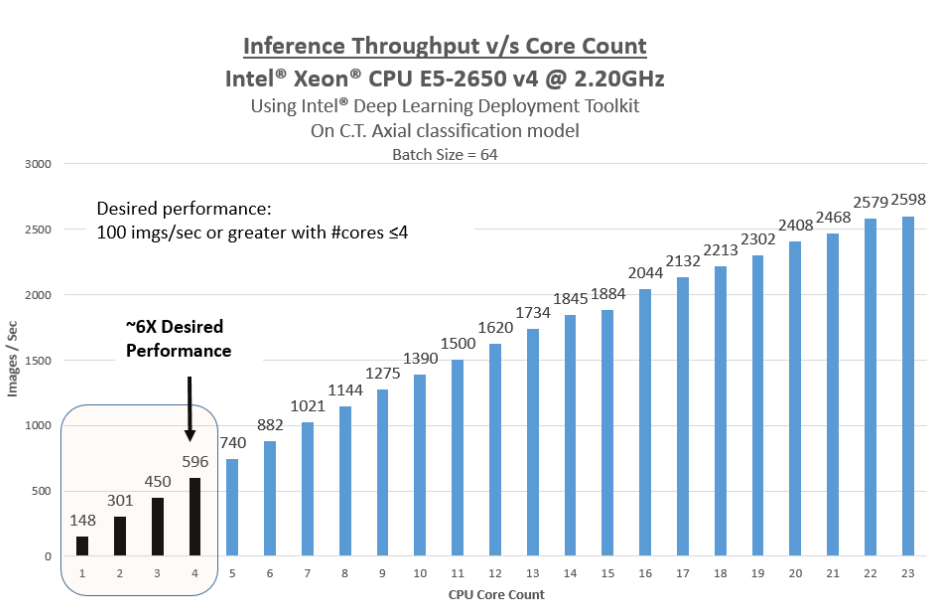
Problem: Need fast inferencing to classify images coming off of a CT scanner

Solution:
High performance inferencing at ~600 images/sec. Exceeded customer goals by 6x and provided 10x performance boost over un-optimized version

CT scan images



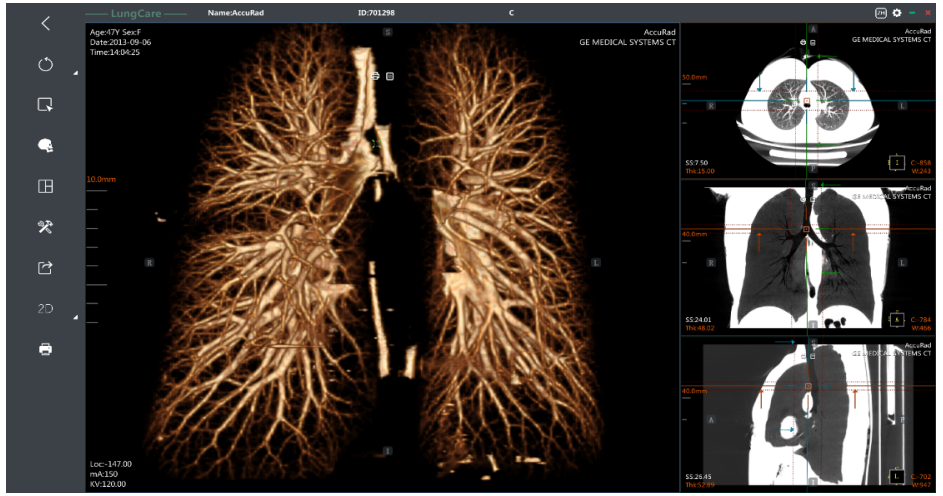


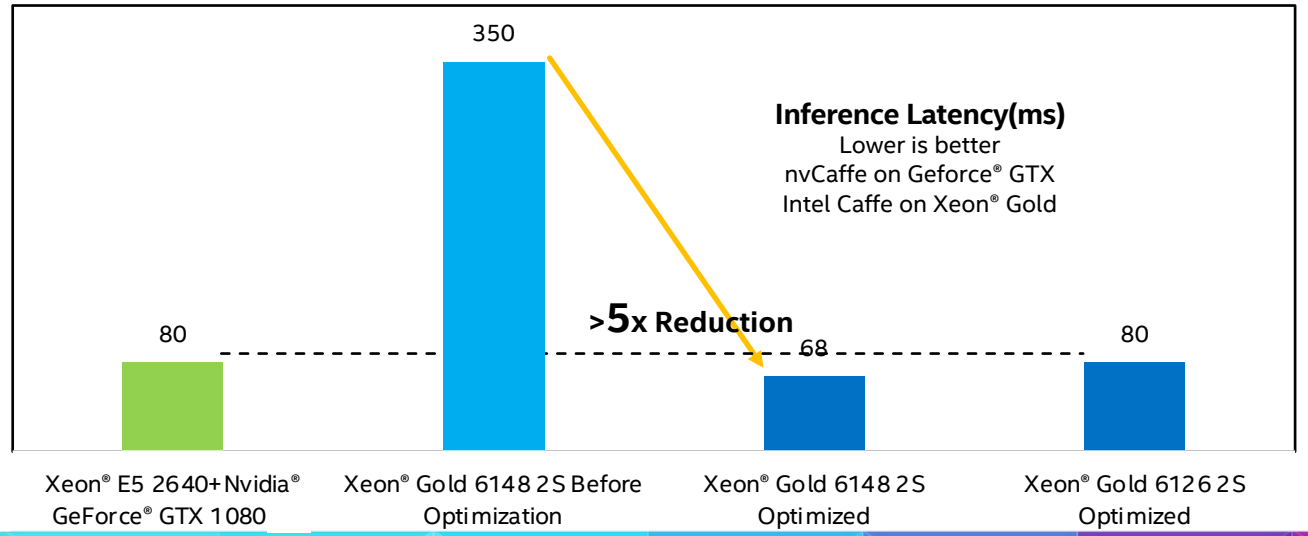


CASE 5: MEDICAL IMAGE ANALYSIS APPLICATION

AI INFERENCE OPTIMIZATION FOR USER EXPERIENCE & TCO

- **Customer**: A leading Medical Image solution company in China wants to port an inference application from NVidia GPGPU to Intel® Xeon® platform for production deployment of the solution on cloud. The motivation is to reduce the deployment cost and simplify the cloud infrastructure requirements.
- **Challenge**: The Medical Image Analysis application requires fast responsiveness for end user experience. The original latency of inference on Xeon® is too high.
- **Support Requirement**: Inference latency on Xeon® can be equal to NVidia® GeForce® GTX 1080
- **Result**: Inference latency on Xeon® Gold 2S system is equal or lower than Nvidia® GPU





WHAT CUSTOMER SHOWED AT INTEL AI WORKSHOP

Transfer to IA Platform

AMD GPU - Rendering

Nvidia GPU - AI computation

CPU - Business processing

Xeon

Low cost、high density、large clustering、Seamless transfer

Performance comparison

Framework : Caffe & Tensorflow

Model : RFCN

Data : Single image in Dicom format (Chest radiograph)

GTX 1080	Xeon 6148
80ms	68ms

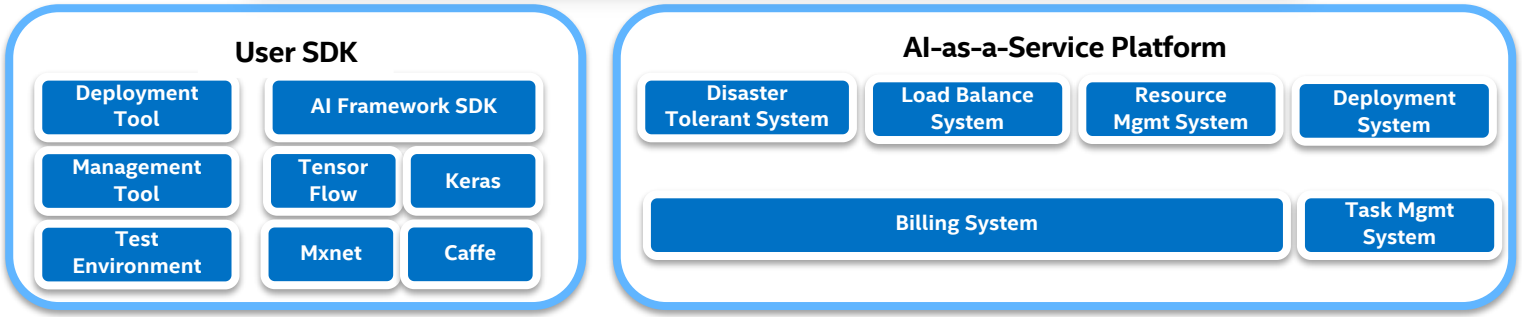
Good performance and Seamless Migration

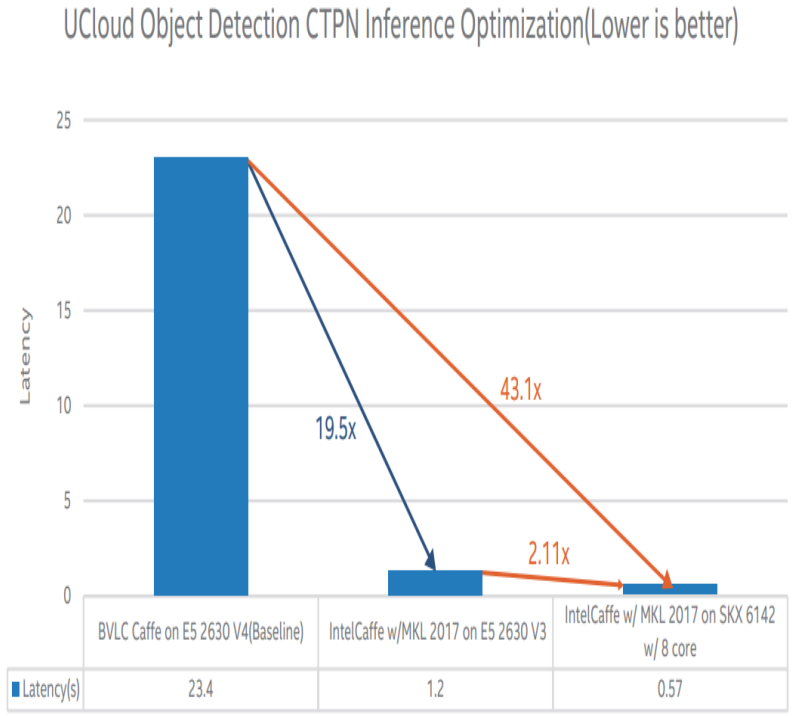
CASE 6: AI AS A SERVICE – AI CLOUD

UCLOUD UAI-INFERENCE

- UCloud - Top5 public cloud service providers in PRC, served more than 50,000 enterprises
- UCloud launched AI online service – UAI-inference using Intel® Xeon® SKX based servers to build up best TCO.
- UCloud has introduced a better performing AI framework: Intel Caffe with 43x higher performance

UCLOUD UAI-Service: The first commercial deployment of Intel optimized Caffe solution as AI differentiated service on SKX in public cloud provider





FIND OUT MORE

LEARN

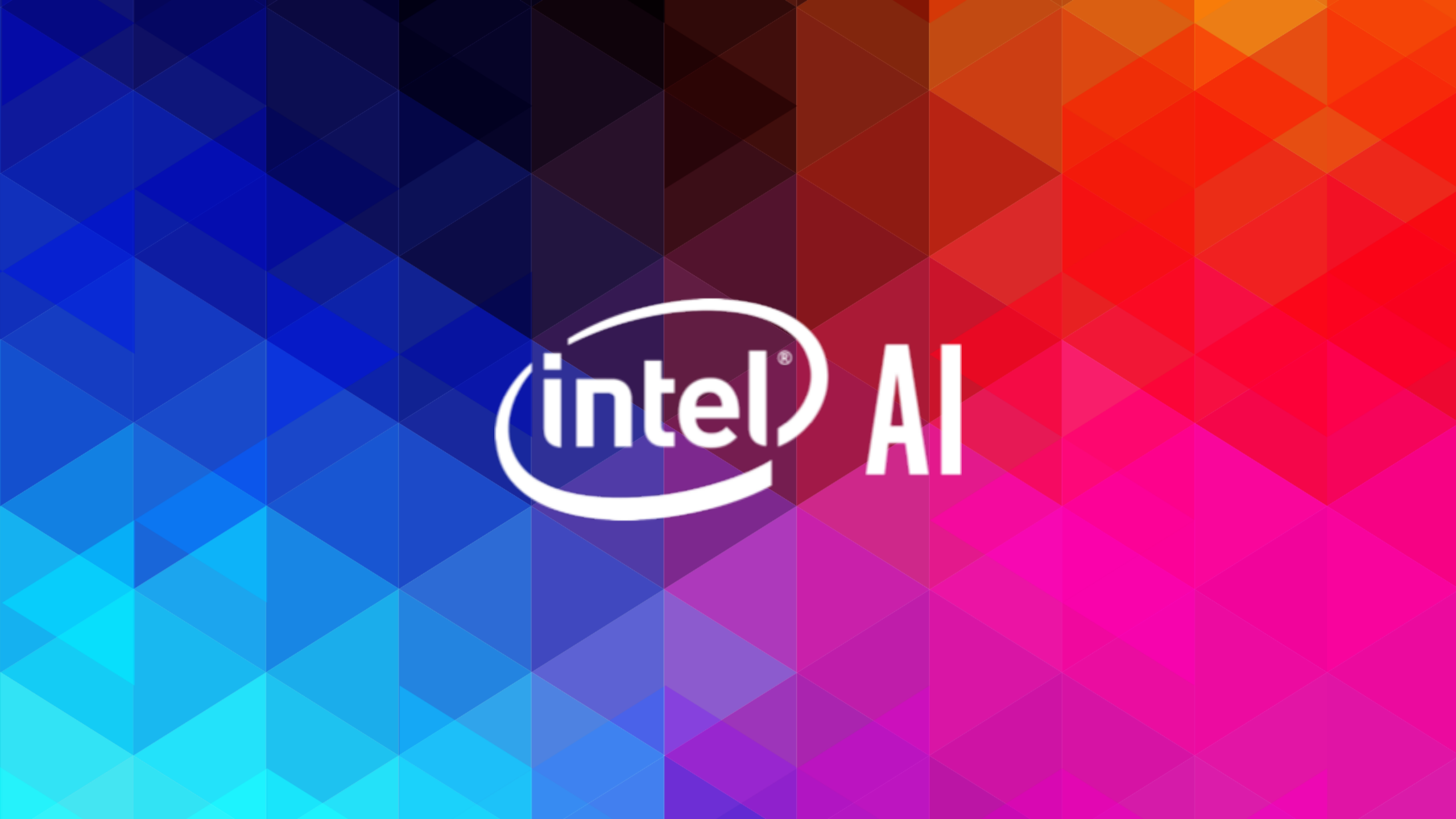
More information at ai.intel.com

EXPLORE

Use Intel's performance-optimized libraries & frameworks

ENGAGE

Contact your Intel representative for help and POC opportunities


intel
AI