GADI SINGER

VICE PRESIDENT,
ARCHITECTURE GENERAL MANAGER,
AI PRODUCTS GROUP - INTEL

THE AI REVOLUTION IS REALLY A COMPUTING EVOLUTION

AIDC
INTEL AI DEVCON 2018

ARTIFICIAL INTELLIGENCE

TYPES OF ANALYTICS/ML
(PARTIAL LIST)

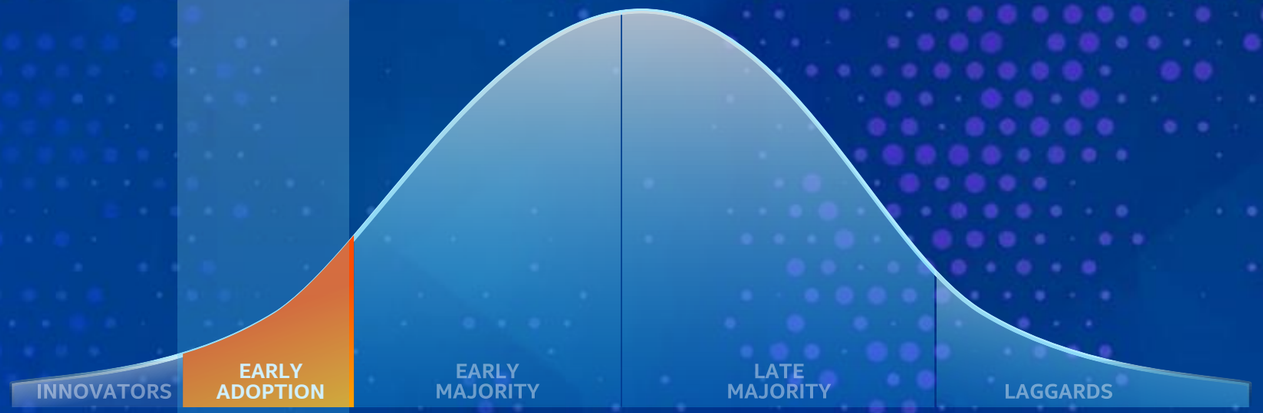
AIDC
INTEL AI DEVCON 2018

AI IS TRANSFORMING ['14/15]

SAMPLE-SCALE BENCHMARKS

COMPUTE RATIO

Inference ~= Training



PROGRAMMING MODEL

C++, Proprietary APIs

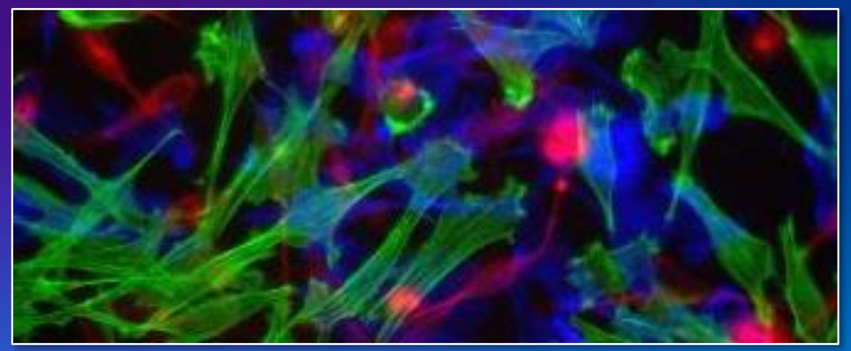
HARDWARE ARCHITECTURE

CPU, GPU

AIDC
INTEL AI DEVCON 2018

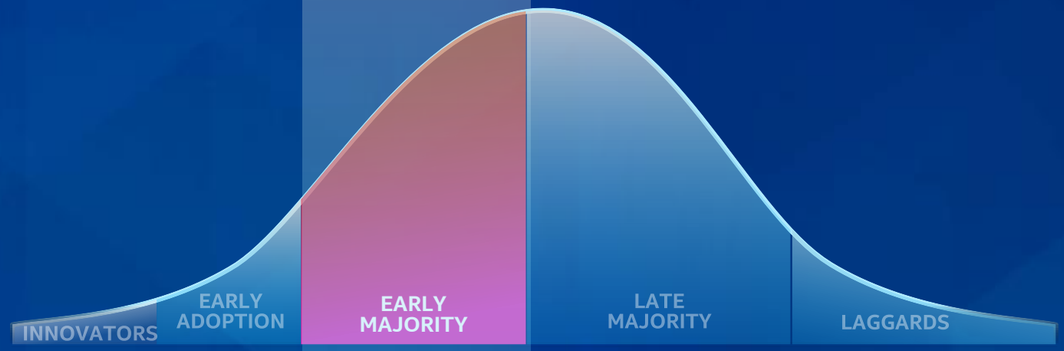
DL COMING OF AGE [‘19/20]

AT-SCALE DEPLOYMENT OF SCIENCE AND REAL-LIFE SOLUTIONS



COMPUTE RATIO

Inference >> Training



ML FRAMEWORKS

‘Democratization of Data Science’

mxnet PaddlePaddle PYTORCH Caffe2 APACHE Spark Microsoft CNTK BigDL

HARDWARE ARCHITECTURE

CPU, DL Acceleration/NNPs, GPUs

AIDC INTEL AI DEVCON 2018

AI IN ACTION – LEADERSHIP AT SCALE

INTEGRATION WITH EXISTING WORKLOADS

BLENDED REAL-LIFE AI WORKLOADS

TCO-TOTAL COST OF OWNERSHIP

END DEVICES TO DATACENTER

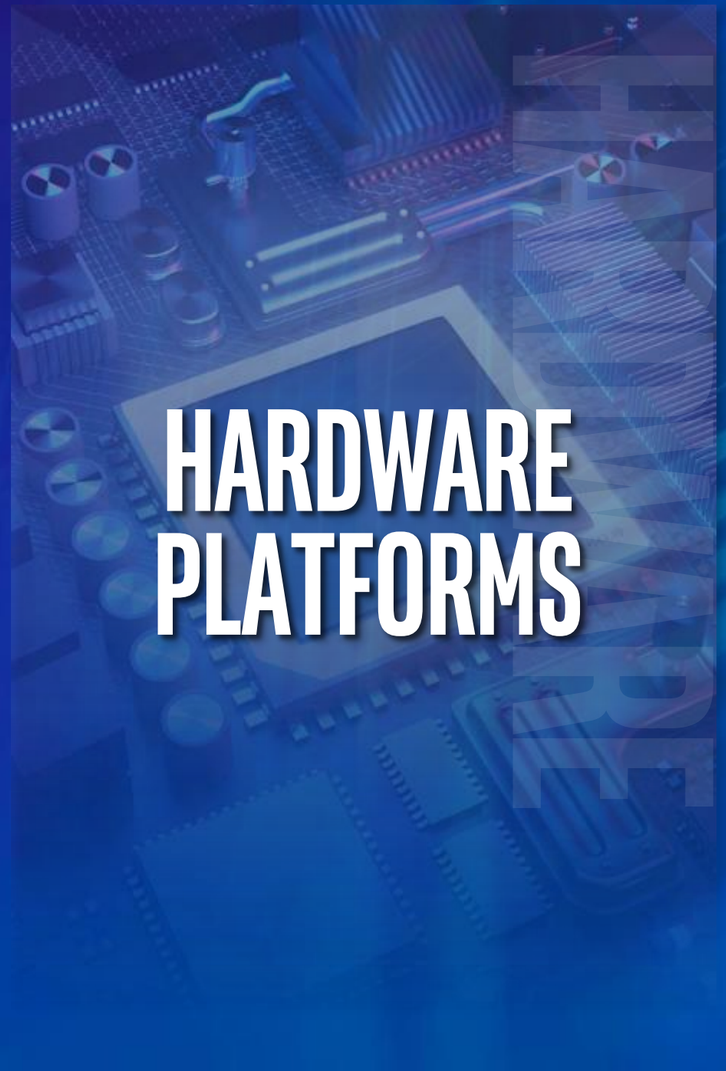
PERFORMANCE VS PERF/W VS LATENCY

FULL PLATFORM OPTIMIZATION

BUILT FOR AI. DESIGNED TO SCALE.

AIDC
INTEL AI DEVCON 2018

ENABLE DEVELOPERS TO ACHIEVE THEIR AI VISION

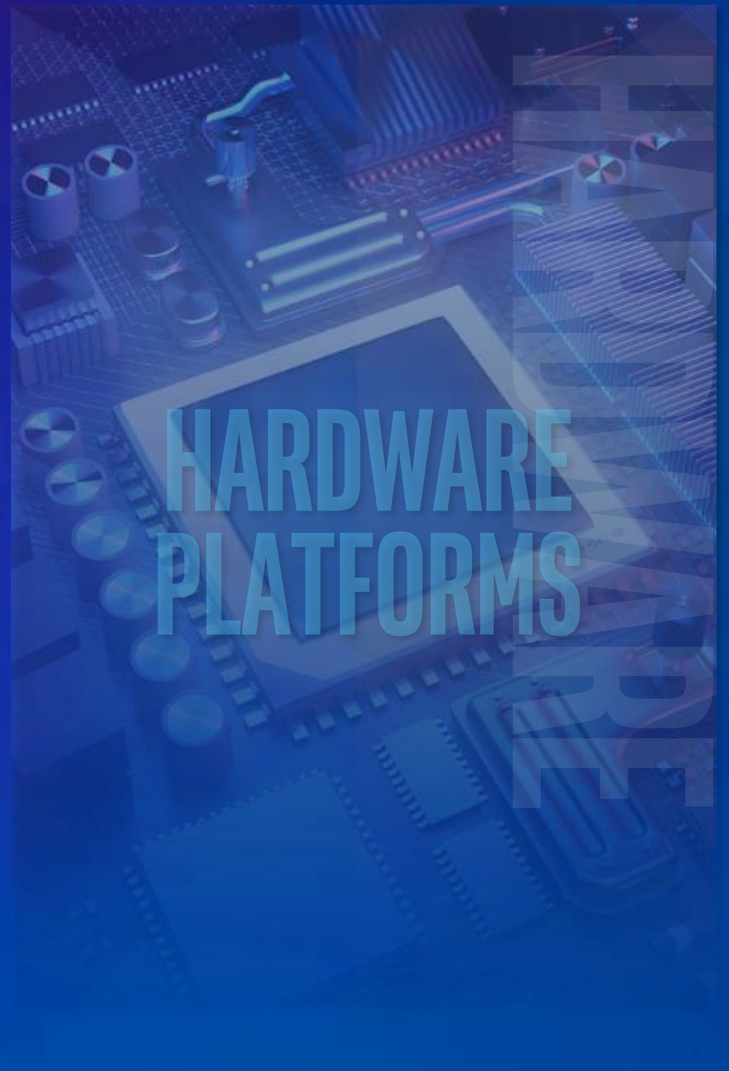


AIDC
INTEL AI DEVCON 2018

AMIR KHOSROWSHAHI

VICE PRESIDENT, AI PRODUCTS GROUP
CHIEF TECHNOLOGY OFFICER - INTEL

ENABLE DEVELOPERS TO ACHIEVE THEIR AI VISION



AIDC
INTEL AI DEVCON 2018

OPEN SOURCE

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

HELPS REALIZE THE INCREDIBLE BENEFITS OF DIRECT OPTIMIZATION

MATRIX MULTIPLICATION

BATCH NORM

POOLING

NORMALIZATION

ACTIVATION

CONVOLUTION

PRIMITIVES

Intel/mkl-dnn

AIDC
INTEL AI DEVCON 2018

TensorFlow

OPTIMIZING TENSORFLOW

TO SUPERCHARGE AI WORKLOADS

AIDC
INTEL AI DEVCON 2018

Arya.ai

ENTERPRISE DEEP LEARNING PLATFORM

VINAY KUMAR SANKARAPU

FOUNDER & CEO, ARYA.AI

AIDC
INTEL AI DEVCON 2018

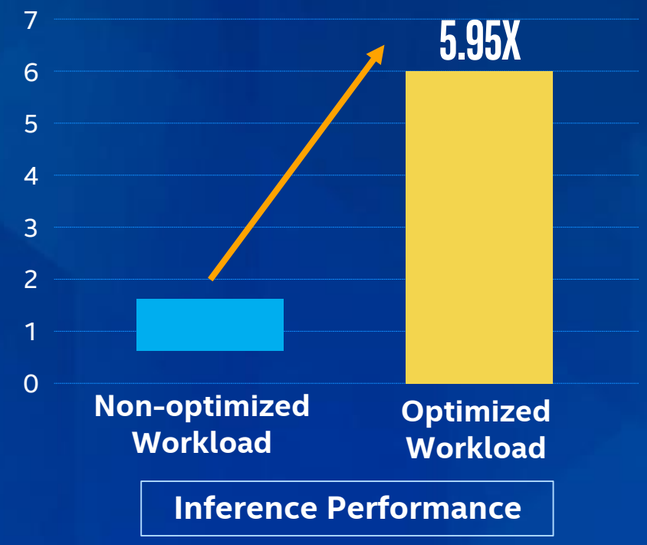
HIGH PERFORMANCE DELIVERED ON INTEL® XEON® SCALABLE PROCESSOR

USE CASE: AUTOMATED CLAIMS PROCESSING for health insurance

CUSTOMER IMPACT

Inferencing Performance on Intel® Xeon® Scalable using Intel distribution for Python* & Intel-optimized TensorFlow* enabled the arya platform to reduce claim processing time from 48hrs to 0.2 seconds

PERFORMANCE WITH THE INTEL® XEON® SCALABLE PROCESSOR



AIDC
INTEL AI DEVCON 2018

nGraph™

OPEN SOURCE COMPILER ENABLING FLEXIBILITY
TO RUN MODELS ACROSS A VARIETY OF FRAMEWORKS
AND HARDWARE

NervanaSystems/ngraph

AIDC
INTEL AI DEVCON 2018

BIGDL: DISTRIBUTED DEEP LEARNING

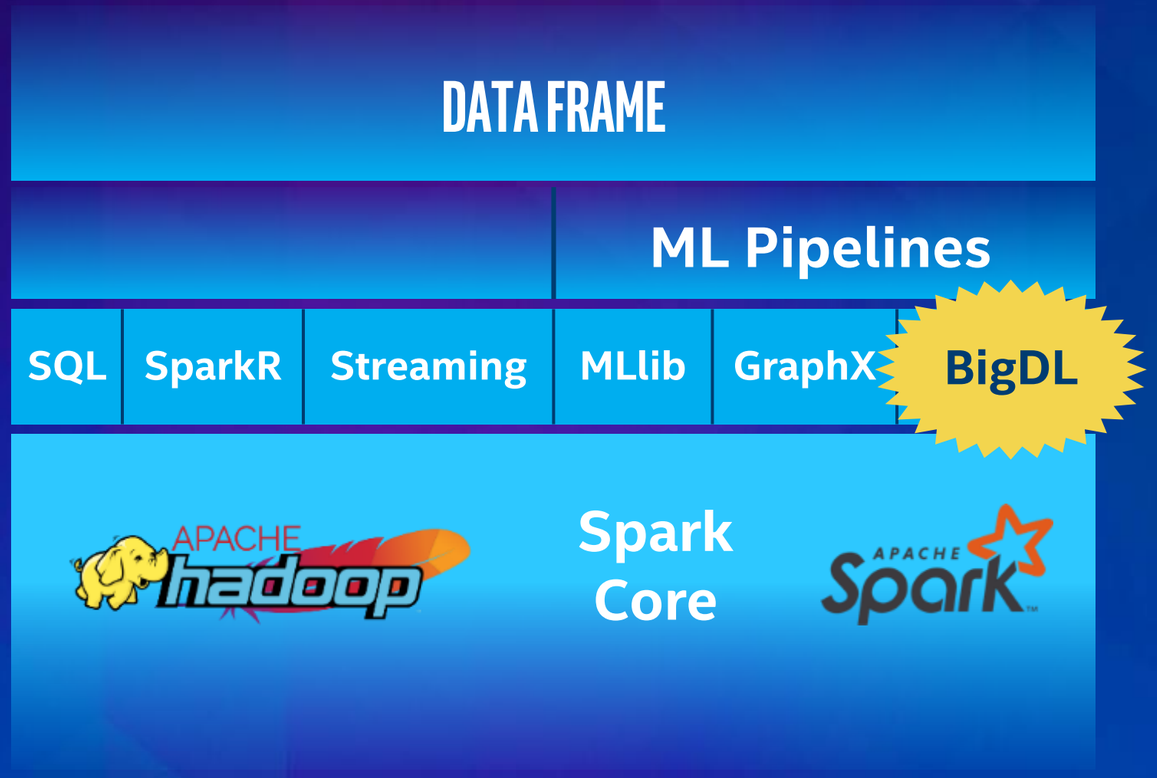
AKANKSHA BALANI

Country Lead - Intel Software Developer Products

AIDC
INTEL AI DEVCON 2018

BigDL

HIGH PERFORMANCE DEEP LEARNING FOR APACHE SPARK* ON CPU INFRASTRUCTURE



DESIGNED AND OPTIMIZED FOR INTEL® XEON® PROCESSOR

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

Feature Parity	Lower TCO, improved ease of use	Efficient Scale-Out

Powered by Intel® MKL-DNN

AIDC INTEL AI DEVCON 2018

CASE STUDY: IMAGE RECOGNITION

JD.京东 JD.COM

RESULT

4X GAIN

Processing ~380M images

BigDL Apache Spark

Intel® Xeon® CPU

CLIENT:

JD.Com, 2nd largest online retailer in China, ~ 250 M users.

CHALLENGE:

Building deep learning applications such as image similarity search without moving data.

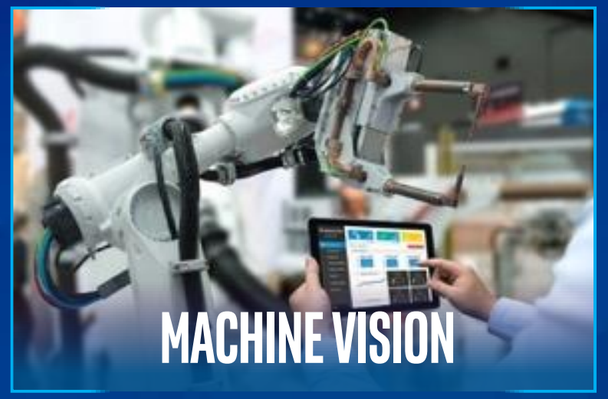
SOLUTION:

Switched from GPU to CPU cluster. Using Apache Spark* with BigDL, running on Intel® Xeon® processors

https://software.intel.com/en-us/articles/building-large-scale-image-feature-extraction-with-bigdl-at-jdcom

AIDC INTEL AI DEVCON 2018

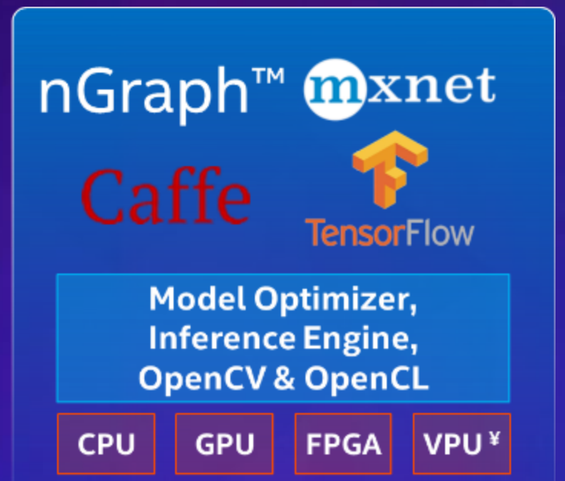
VIDEO IS THE ULTIMATE IOT SENSOR



AIDC
INTEL AI DEVCON 2018

ANNOUNCING: OPENVINO™ SOFTWARE TOOLKIT

VISUAL INFERENCING AND NEURAL NETWORK OPTIMIZATION



DEPLOY COMPUTER VISION AND DEEP LEARNING CAPABILITIES TO THE EDGE

- WRITE ONCE + SCALE TO DIVERSE ACCELERATORS
- BROAD FRAMEWORK SUPPORT
- HIGH PERFORMANCE, HIGH EFFICIENCY FOR THE EDGE

Other names and brands may be claimed as the property of others
VPU = Vision Processing Unit (Movidius)

AIDC
INTEL AI DEVCON 2018

ENABLE DEVELOPERS TO ACHIEVE THEIR AI VISION

AIDC
INTEL AI DEVCON 2018

intel AI

BUILT FOR AI. DESIGNED TO SCALE.

intel XEON PLATINUM inside
intel Future NERVANA inside
intel CORE i7 8th Gen
intel Iris Graphics
intel ATOM inside
MOBILEYE An Intel Company
intel STRATIX 10 inside
intel MOVIDIUS inside
intel GNA inside

AIDC
INTEL AI DEVCON 2018

FOUNDATIONAL FOR ARTIFICIAL INTELLIGENCE

AIDC
INTEL AI DEVCON 2018

ARTIFICIAL INTELLIGENCE AND INTEL® XEON® PROCESSORS

AIDC
INTEL AI DEVCON 2018

Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective

Kim Hazelwood, Sarah Bird, David Bro
Mohamed Fawzy, Bill Jia, Yangqing
Pieter Noordhuis, Mis
Soumith Chintala, Utku D

Abstract—Machine learning sits at the core of many products and services at Facebook. This paper describes the hardware and software infrastructure that supports machine learning at global scale. Facebook's machine learning workloads are extremely diverse: services require many different types of models in practice. This diversity has implications at all layers in the system stack. In addition, a sizable fraction of all data stored at Facebook flows through machine learning pipelines, presenting significant challenges in delivering data to high-performance distributed training flows. Computational requirements are also intense, leveraging both GPU and CPU platforms for training and abundant CPU capacity for real-time inference. Addressing these and other emerging challenges continues to require diverse efforts that span machine learning algorithms, software, and hardware design.

I. Introduction

Facebook's mission is to "Give people the power to build community and bring the world closer together." In support of that mission, Facebook connects more than two billion people as of December 2017. Meanwhile, the past several years have seen a revolution in the application of machine learning to real problems at this scale, building upon the virtuous cycle of machine learning algorithmic innovations, enormous amounts of training data for models, and advances in high-performance computer architectures [1]. At Facebook, machine learning provides key capabilities in driving nearly all aspects of user experience including services like ranking posts for News Feed, speech and text translations, and photo and real-time video classification [2], [3].

Facebook leverages a wide variety of machine learning algorithms in these services including support vector machines, gradient boosted decision trees, and many styles of neural networks. This paper describes several important aspects of datacenter infrastructure that supports machine learning at Facebook. The infrastructure includes internal "ML-as-a-Service" flows, open-source machine learning frameworks, and distributed training algorithms. From a hardware point of view, Facebook leverages a large fleet of CPU and GPU platforms for training models in order to support the necessary training frequencies at the required service latency. For many [illegible] inference. Facebook primarily relies on CPUs for all major [illegible] neural network ranking services [illegible] Feed dominating the [illegible] compute load.

[illegible] learning pipelines, and this fraction is increasing over time as improve model quality. The massive amount of data required by machine learning services presents challenges at the global scale of Facebook's datacenters. Several techniques are used to efficiently feed data to the models including data coupling of data feed and training, data/compute co-location, and networking optimizations. At the same time, Facebook's scale provides unique opportunities. Diurnal load cycles leave a significant number of CPUs available for distributed training algorithms during off-peak periods. With Facebook's compute fleet spread over ten datacenter locations, scale also provides disaster recovery capability. Disaster recovery planning is essential as timely delivery of new machine learning models is important to Facebook's operations.

Looking forward, Facebook expects rapid growth in machine learning across existing and new services [4]. This growth will lead to growing scalability challenges for teams deploying the infrastructure for these services. While significant opportunities exist to optimize infrastructure on existing platforms, we continue to actively evaluate and prototype new hardware solutions while remaining cognizant of game-changing algorithmic innovations.

The key contributions of this paper include the following major insights about machine learning at Facebook.

- Machine learning is applied pervasively across nearly all services, and **computer vision represents only a small fraction** of the resource requirements.
- Facebook relies upon an incredibly **diverse set of machine learning approaches** including, but not limited to, neural networks.
- **Tremendous amounts of data** are funneled through our machine learning pipelines, and this creates engineering and efficiency challenges far beyond the compute nodes.
- Facebook currently relies heavily on **CPUs for inference**, and **both CPUs and GPUs for training**, but constantly prototypes and evaluates new hardware solutions from a performance-per-watt perspective.
- The worldwide scale of people on Facebook and corresponding diurnal activity patterns result in a huge number of machines that can be harnessed for machine learning tasks such as **distributed training at scale**.

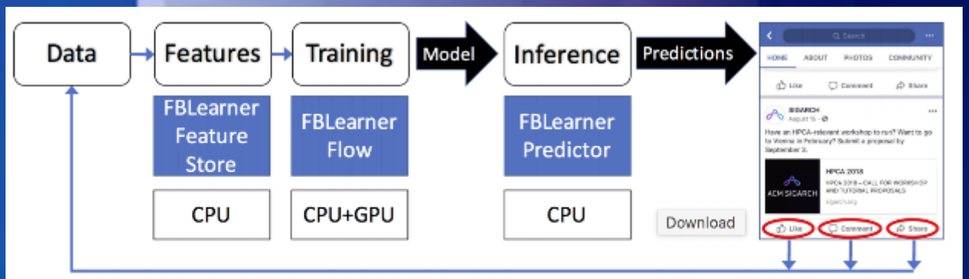


Fig. 1. Example of Facebook's Machine Learning Flow and Infrastructure.

Services	Relative Capacity	Compute	Memory
News Feed	100X	Dual-Socket CPU	High
Facer	10X	Single-Socket CPU	Low
Lumos	10X	Single-Socket CPU	Low
Search	10X	Dual-Socket CPU	High
Language Translation	1X	Dual-Socket CPU	High
Sigma	1X	Dual-Socket CPU	High
Speech Recognition	1X	Dual-Socket CPU	High

TABLE III
RESOURCE REQUIREMENTS OF ONLINE INFERENCE WORKLOADS.

wipro holmes

FROM CHIP TO APPLIED AI

TAPATI BANDOPADHYAY
GENERAL MANAGER AND PRACTICE HEAD
WIPRO HOLMES

AIDC
INTEL AI DEVCON 2018

AI-OPTIMIZED CHIP TO BUSINESS SERVICES

VISION

Driving AI-empowered Business Services

Ensuring Adoption, Scale & ROI from AI Investments

SOLUTION DEPLOYED

wipro holmes

intel XEON PLATINUM inside

Performance Benchmarking in Image Detection & Recognition

Leveraging CNNs on Intel® Xeon® Platinum

OUTCOME

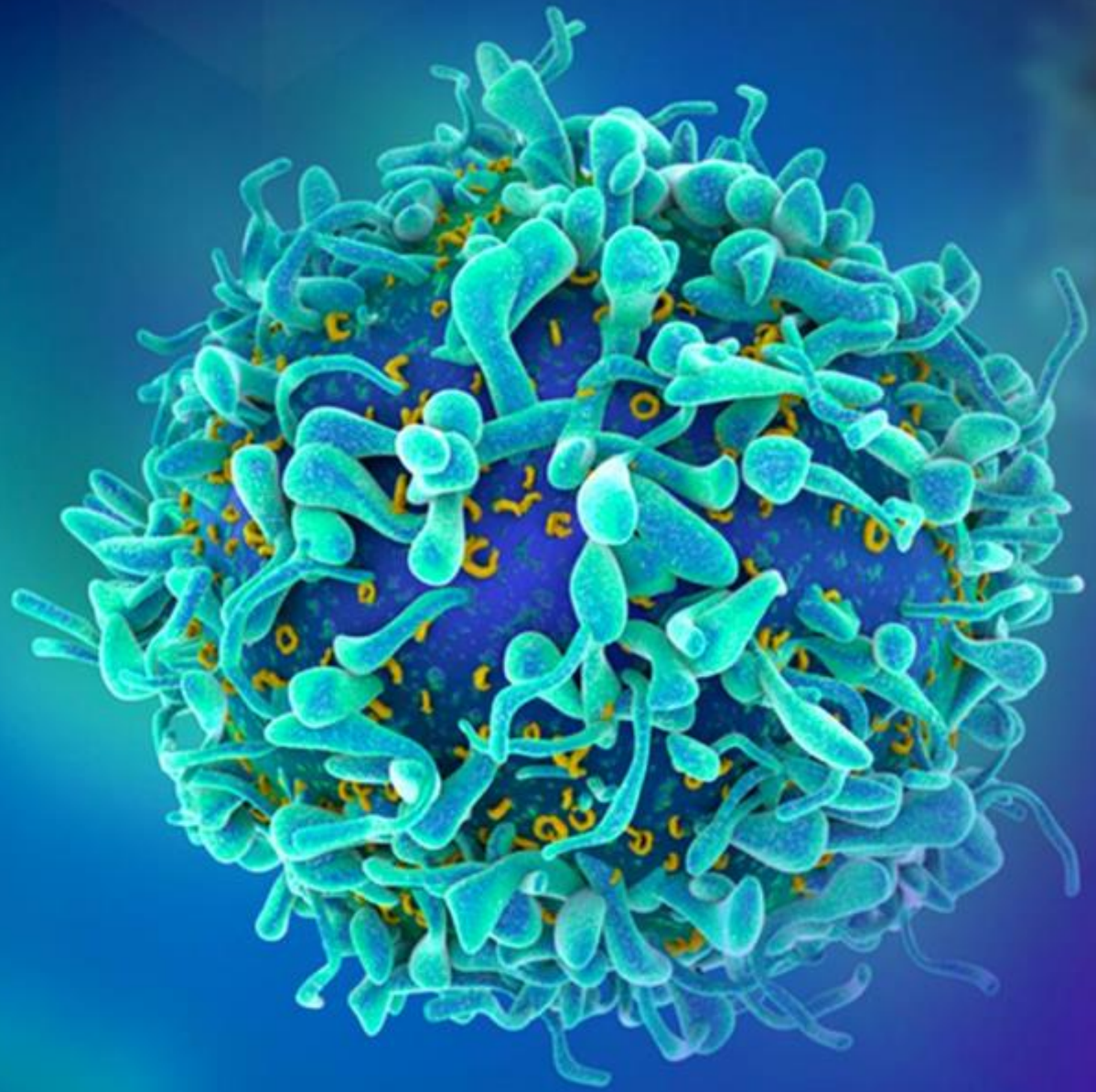
320+
Accounts

Analyst Reports Leadership **20 FY18**

Automation Arbitrage **19,000** people equivalent productivity

3,200+
Bots Deployed FY18

AIDC
INTEL AI DEVCON 2018



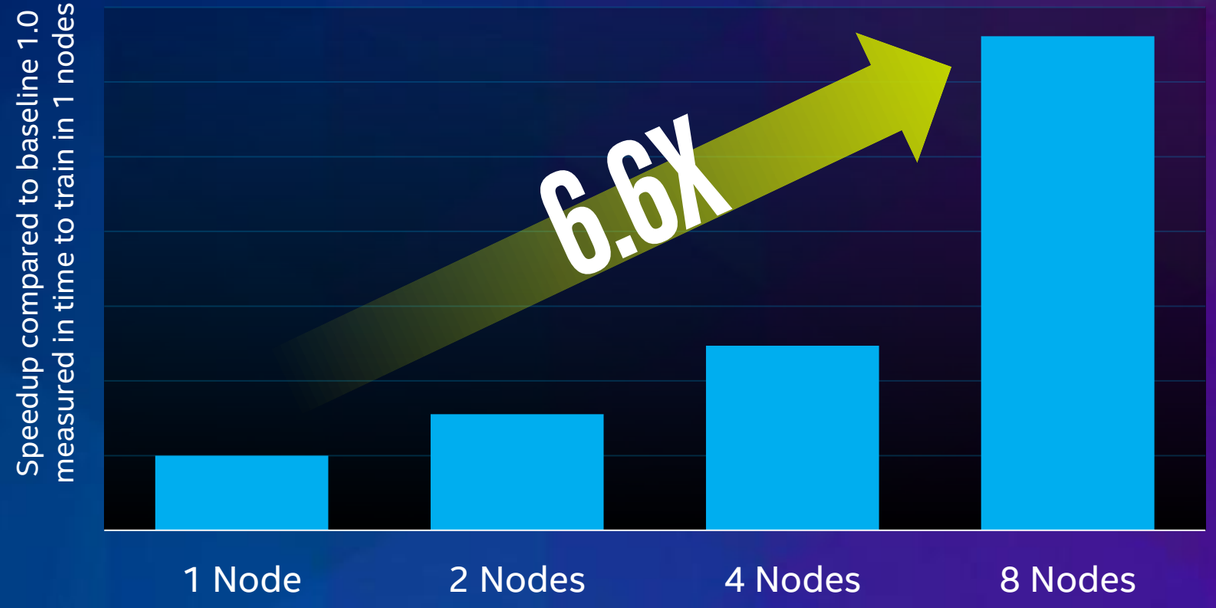
A SCIENTIFIC COLLABORATION BETWEEN INTEL AND NOVARTIS

AIDC
INTEL AI DEVCON 2018

HIGH PERFORMANCE AT SCALE

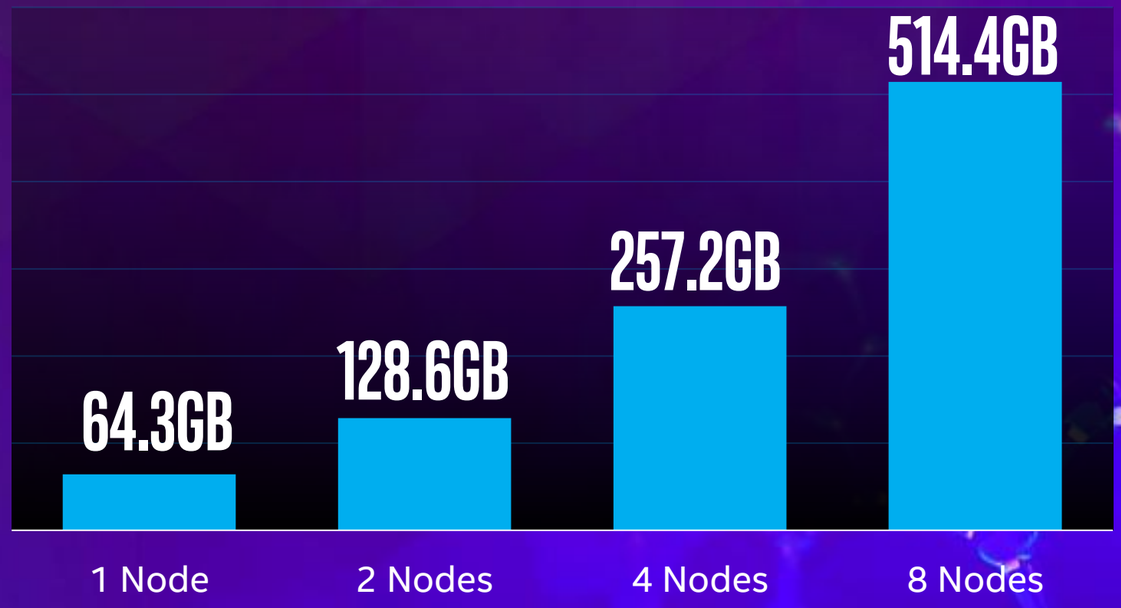
SCALING OF TIME TO TRAIN

Intel® Omni-Path Architecture, Horovod and TensorFlow®



TOTAL MEMORY USED

192GB DDR4 PER INTEL® 2S XEON® 6148 PROCESSOR



MULTISCALE CONVOLUTION NEURAL NETWORK

TensorFlow

OPTIMIZED LIBRARIES

Intel® MKL/MKL-DNN, clDNN, DAAL

INTEL® OMNI-PATH ARCHITECTURE

Other names and brands may be claimed as the property of others
§ Configuration: CPU: Xeon 6148 @ 2.4GHz, Hyper-threading: Enabled. NIC: Intel® Omni-Path Host Fabric Interface, TensorFlow: v1.7.0, Horovod: 0.12.1, OpenMPI: 3.0.0. OS: CentOS 7.3, OpenMPU 23.0.0, Python 2.7.5
Time to Train to converge to 99% accuracy in model
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.

AIDC
INTEL AI DEVCON 2018

FLEXIBLE REAL-TIME INFERENCING
FPGA PRODUCTS

// AI For Earth

GPU

STOP

1 1 10

45,546 images processed

FPGA

STOP

1 1

308,500 images processed

AIDC
INTEL AI DEVCON 2018

DEEP NEURAL NETWORKS – PROJECT BRAINWAVE

DR. VIVEK SESHADRI

RESEARCHER – MICROSOFT LABS

AIDC INTEL AI DEVCON 2018

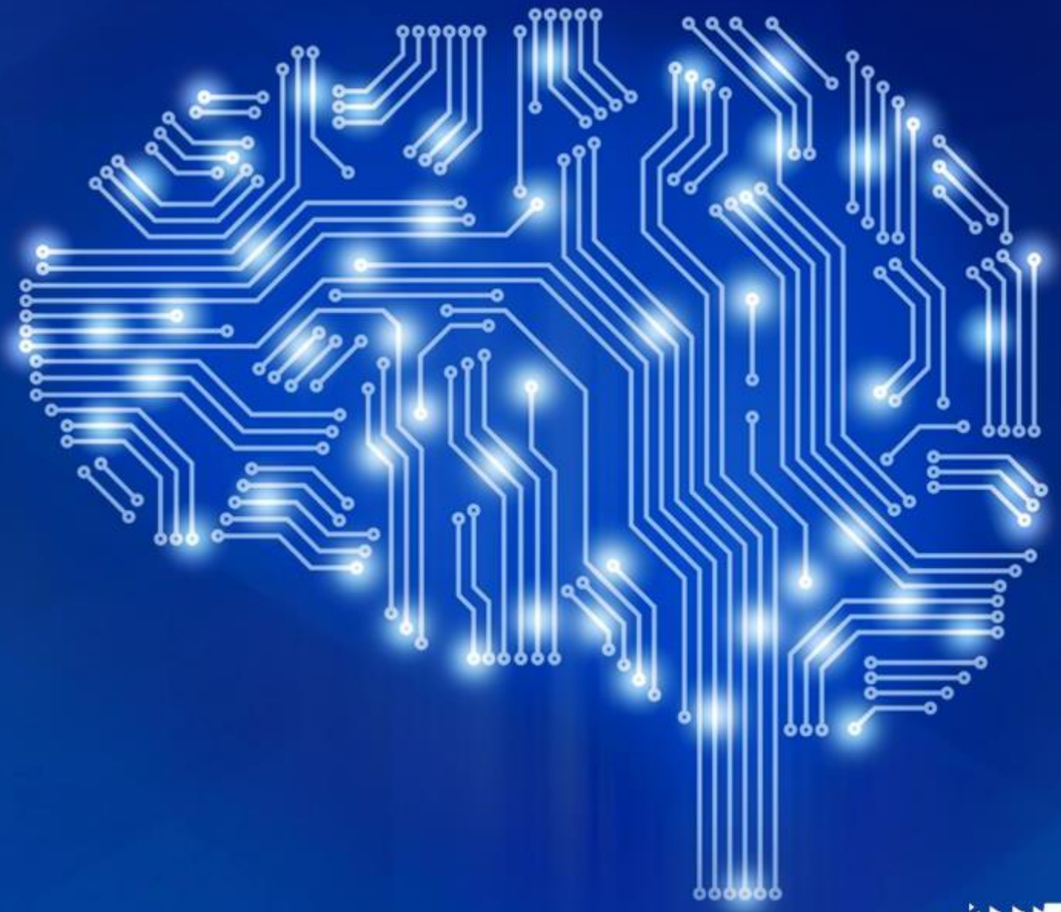
PROJECT BRAINWAVE

Microsoft

A SCALABLE FPGA-POWERED SERVING PLATFORM FOR DEEP NEURAL NETWORKS

INDUSTRY-LEADING LATENCY (<2MS) AT ULTRA LOW COST

DEPLOYABLE TO INTELLIGENT CLOUD OR INTELLIGENT EDGE THROUGH AZURE IOT



Azure

http://aka.ms/aml-real-time-ai
brainwave-edge@microsoft.com

AIDC
INTEL AI DEVCON 2018

intel® Movidius™

AIDC
INTEL AI DEVCON 2018

intel® Movidius™

Deploy DNN and Computer Vision at the Edge

Native FP16 and Fixed Point 8 bit support

4 TOPS with 1 TOPS of DNN Compute at 1W

AIDC
INTEL AI DEVCON 2018

intel Movidius™

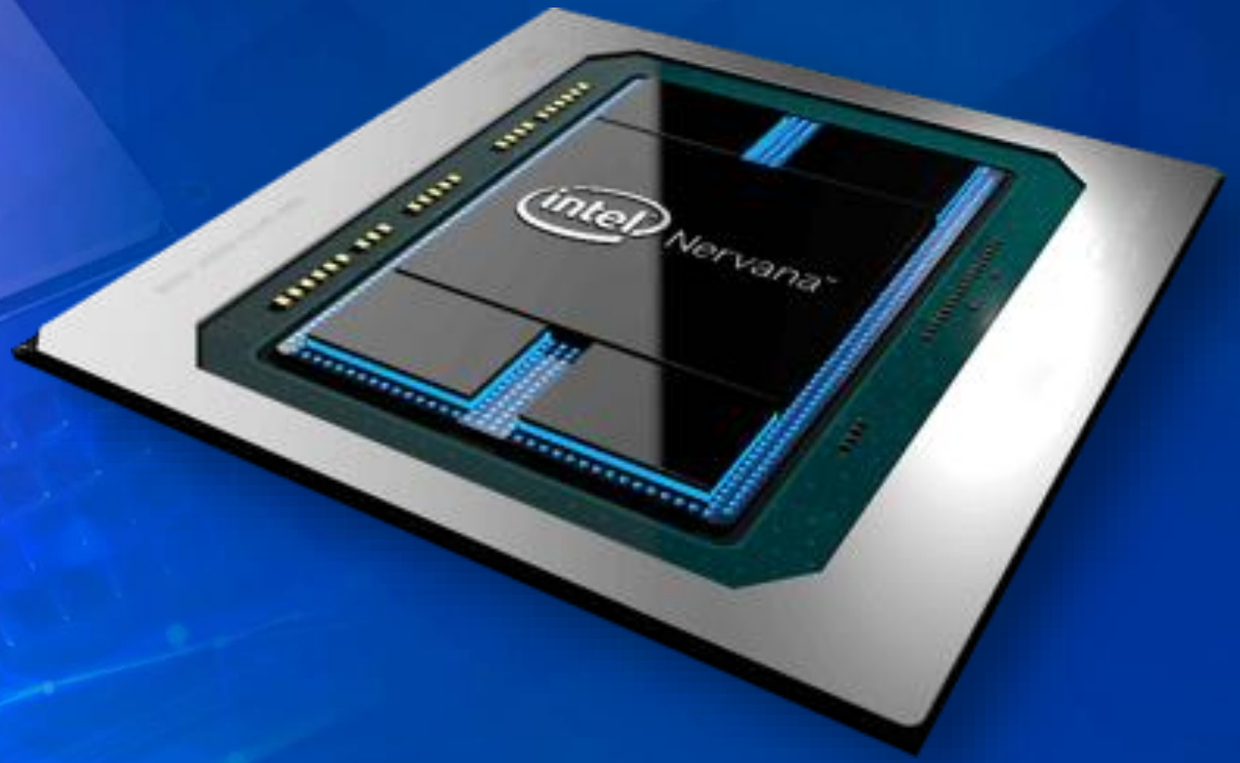
AI GENERATED MUSIC

KEVIN DOUCETTE
KATE KALLOT

AIDC
INTEL AI DEVCON 2018

intel® Nervana™

NEURAL NETWORK PROCESSORS



AIDC
INTEL AI DEVCON 2018

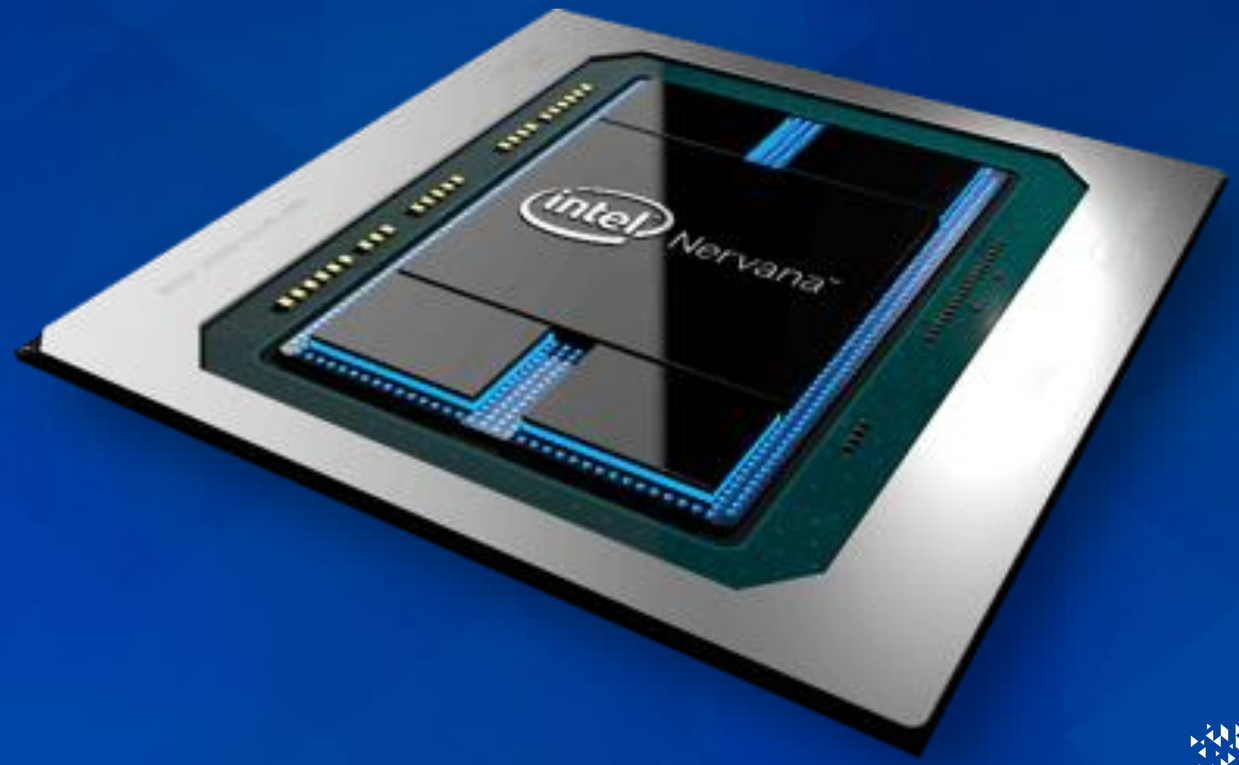

intel Nervana™
GOALS – HIGH UTILIZATION
& MODEL PARALLELISM
AIDC
INTEL AI DEVCON 2018

INTEL® NERVANA™ NNP L-1000 in 2019

3-4x training performance of first generation Lake Crest product

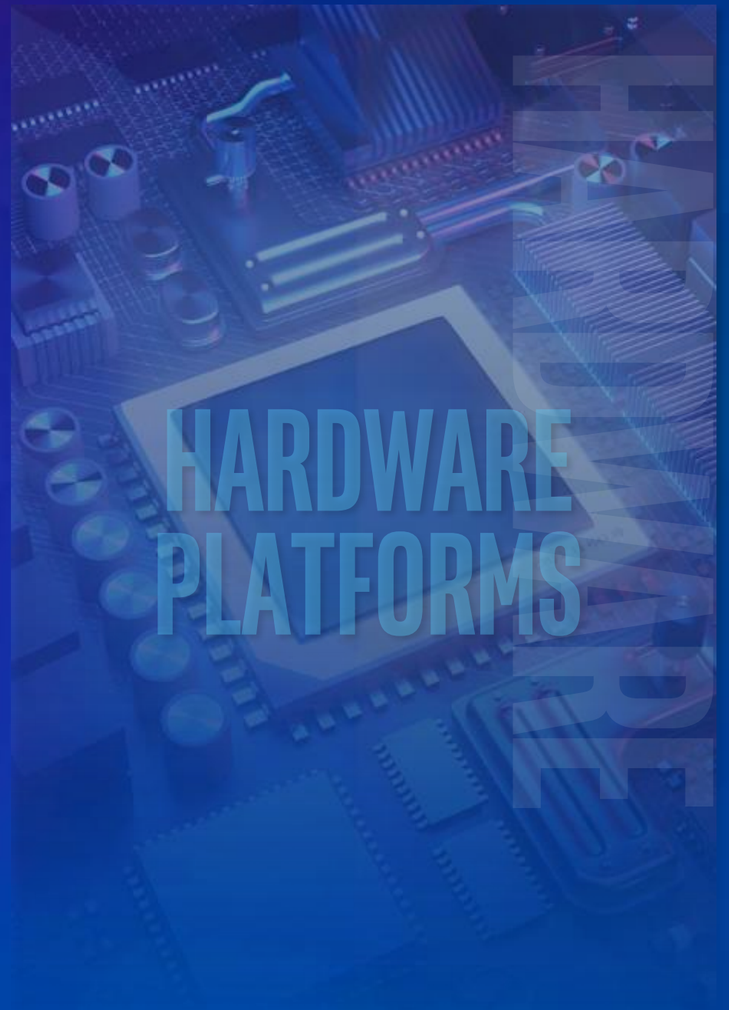
SPRING CREST

PURPOSE BUILT DESIGN OPTIMIZED ACROSS MEMORY BANDWIDTH, UTILIZATION, AND POWER



AIDC
INTEL AI DEVCON 2018

ENABLE DEVELOPERS TO ACHIEVE THEIR AI VISION



AIDC
INTEL AI DEVCON 2018

AIDC
INTEL AI DEVCON 2018

PRAKASH MALLYA

MANAGING DIRECTOR, SALES & MARKETING GROUP
INTEL INDIA

THE PROMISE OF DATA DRIVEN INDIA

DIGITIZATION

- 1B+ citizen data digitized; payment digitization
- 4G infra momentum, content explosion

CAPABILITIES

- 3rd largest footprint of AI startups*
- 2nd largest developer ecosystem in the world

OPPORTUNITIES

- Our demographics
- High friction verticals: Healthcare, Agriculture and Education

COMMUNITY BUILDING CRITICAL TO UNLEASH THE TRUE POTENTIAL OF AI

AIDC
INTEL AI DEVCON 2018

** Tracxn Startup Tracker ** McKinsey Qtrly Data # Intel MS&F and Economic Times*

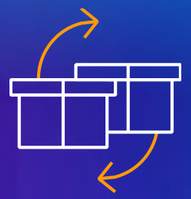
SUPERCHARGE YOUR ML MODEL

ATANU ROY

AI SPECIALIST SOLUTIONS ARCHITECT -AMAZON

MACHINE LEARNING AT AMAZON: A LONG HERITAGE

Personalized recommendations



Fulfillment automation / inventory management

amazon Prime Air

Cargo

amazon alexa

Voice driven interactions

amazon go

Inventing entirely new customer experiences

AIDC
INTEL AI DEVCON 2018

AMAZON SAGEMAKER

BUILD

Pre-built notebooks for common problems

Built-in, high performance algorithms

TRAIN

One-click training

Hyperparameter optimization

DEPLOY

One-click deployment

Fully managed hosting with auto-scaling

AIDC
INTEL AI DEVCON 2018

PHILIPS
Healthcare

PHILIPS BONE-AGE DL INFERENCE

RAVI RAMASWAMY

PHILIPS HEALTHCARE - SENIOR DIRECTOR

AIDC
INTEL AI DEVCON 2018

PHILIPS BONE-AGE DL INFERENCE MODEL ON INTEL® XEON® PROCESSOR PLATFORM

1

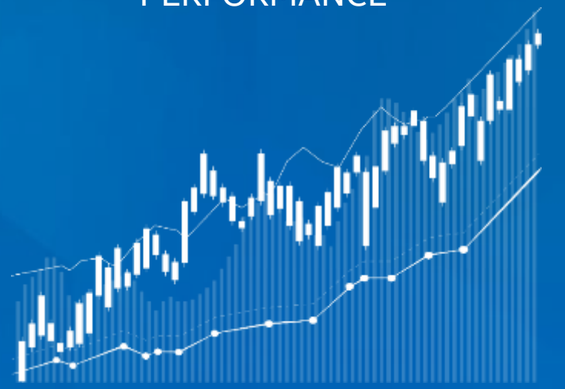
BIOMEDICAL IMAGE BASED ANALYTICS FOR DISEASE SCREENING/DIAGNOSIS

2

HEALTHCARE ANALYTICS WLs **OPTIMIZATION ON INTEL® XEON® PROCESSOR**

3

40X IMPROVEMENT IN INFERENCE PERFORMANCE



AIDC
INTEL AI DEVCON 2018

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

AI FOR SAFE MOBILITY

DR. C. V. JAWAHAR
AMAZON CHAIR PROFESSOR, CVIT,
IIIT-HYDERABAD

AIDC
INTEL AI DEVCON 2018

INDIA DRIVING DATA SET

AN INTEL & IIIT-HYDERABAD INITIATIVE

DATA MAKES AI POSSIBLE

Volume

+

Variety

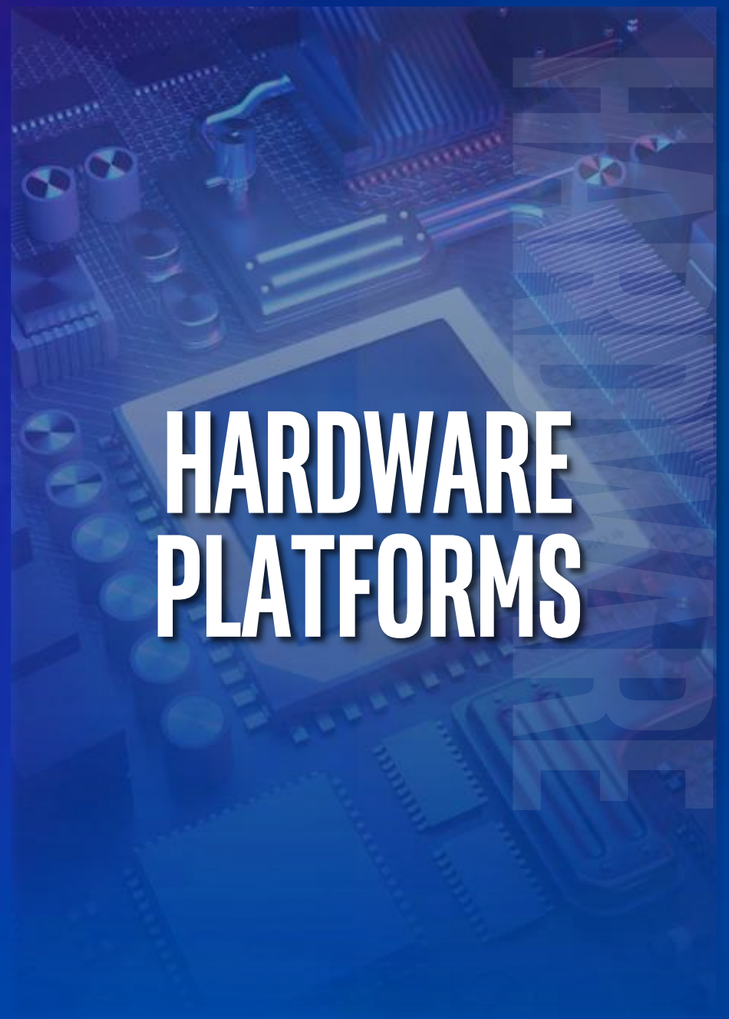
Better Reasoning and Actions

Enabling **AI Research in Unstructured Driving Conditions & Innovations** around Safe Mobility

WORLD'S FIRST PUBLIC DATA SET OF INDIAN DRIVING CONDITIONS

AIDC
INTEL AI DEVCON 2018

ENABLE DEVELOPERS TO ACHIEVE THEIR AI VISION



AIDC
INTEL AI DEVCON 2018

HELPING MAKE YOUR AI VISION A REALITY

TOOLS

HARDWARE

COMMUNITY

INTEL AI DEVCON 2018

RISK FACTORS

Today's presentation contains forward-looking statements. All statements made that are not historical facts are subject to a number of risks and uncertainties, and actual results may differ materially. Please refer to our most recent earnings release, Form 10-Q and 10-K filing available on our website for more information on the risk factors that could cause actual results to differ.