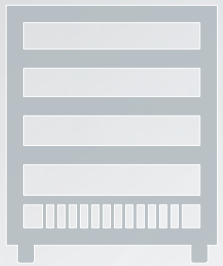
POWERING AI

Werner Schueler

Enterprise Account Manager, Intel

THE NEXT BIG WAVE



MAINFRAMES

STANDARDS-BASED SERVERS

CLOUD COMPUTING

- DATA DELUGE
- COMPUTE BREAKTHROUGH
- INNOVATION SURGE

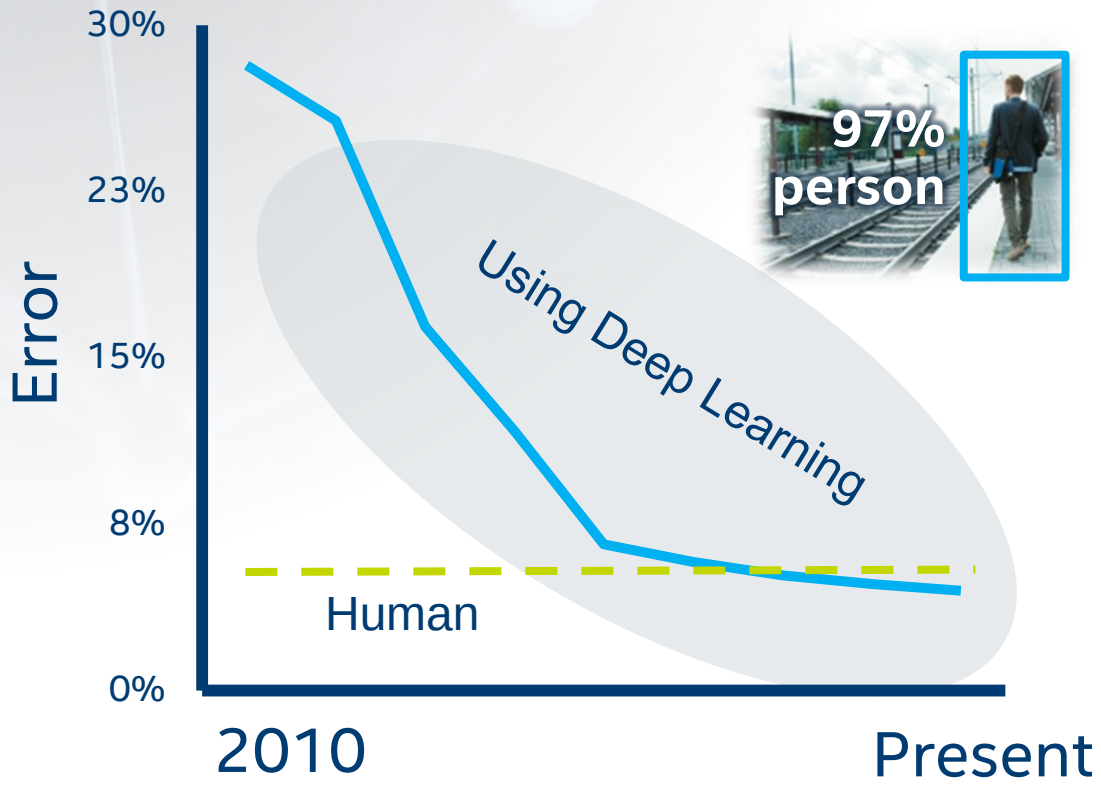
ARTIFICIAL INTELLIGENCE

AI COMPUTE CYCLES WILL GROW 12X BY 2020

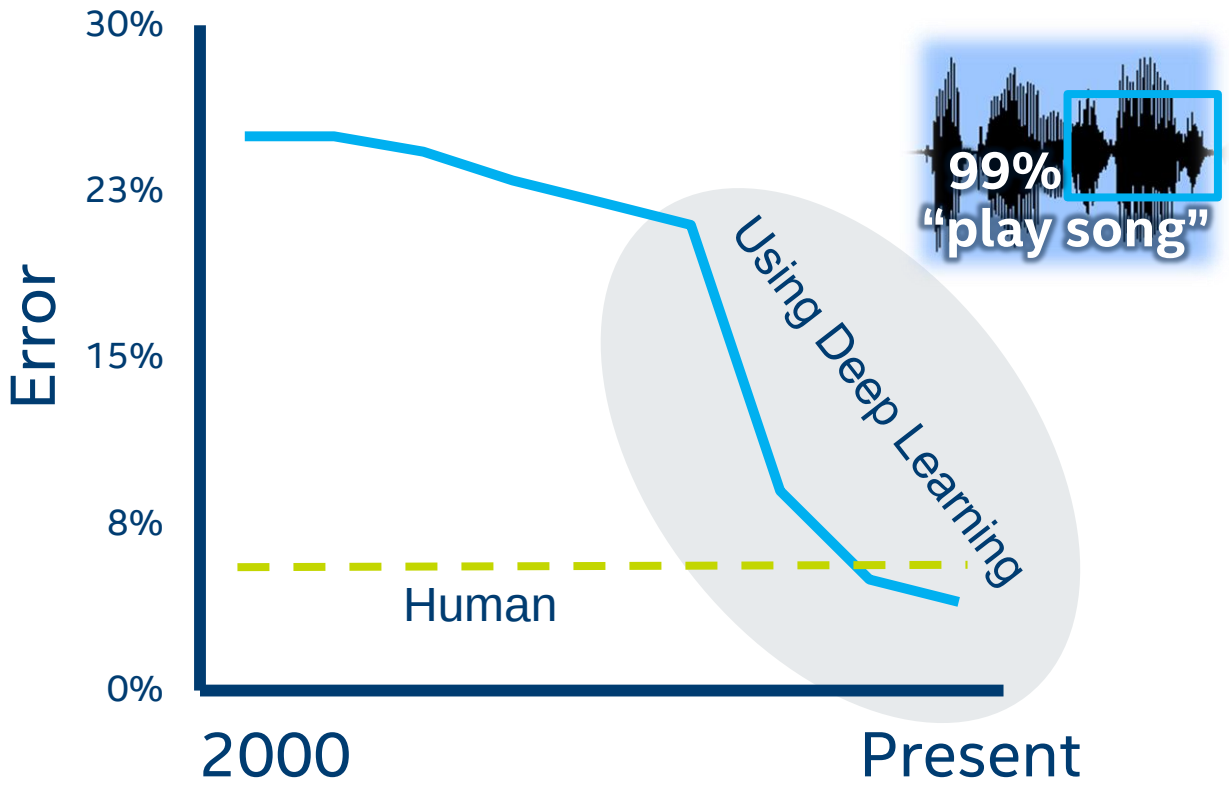
Source: Intel forecast

DEEP LEARNING BREAKTHROUGHS

IMAGE RECOGNITION



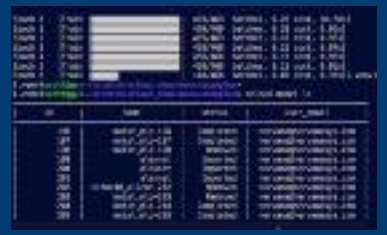
SPEECH RECOGNITION



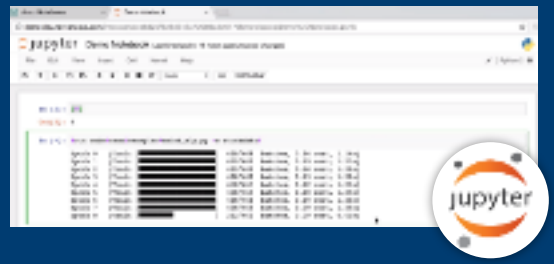
High-Level Workflow

Multiple Interface options

ncloud
Command Line Interface



Interactive Notebooks



User Interface

Label

Dataset

Import Dataset

Build Model

Train

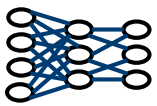
Deploy

Cloud/Server

Edge

Model Library

Trained Model



INTEL® NERVANA™ PORTFOLIO

EXPERIENCES

PLATFORMS

Intel® Nervana™ Cloud & System

Intel® Nervana™ Deep Learning Studio

Intel® Computer Vision SDK

Movidius™ Technology

intel Saffron™

FRAMEWORKS

Spark MLlib BigDL

neon

TensorFlow

Microsoft CNTK

mxnet

torch

Caffe

Caffe2

Chainer

theano

LIBRARIES

python Intel® Python Distribution

Intel® Data Analytics Acceleration Library (DAAL)

Intel® Nervana™ Graph*

Intel® Math Kernel Library (MKL, MKL-DNN)

HARDWARE

Compute

Memory & Storage

Networking

INSIDE AI

*Future

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

END-TO-END AI COMPUTE

DATACENTER

Many-to-many hyperscale for stream and massive batch data processing

Ethernet & Wireless

GATEWAY

1-to-many with majority streaming data from devices

Wireless and non-IP wired protocols

- ✓ *Secure*
- ✓ *High throughput*
- ✓ *Real-time*

EDGE

1-to-1 devices with lower power and often UX requirements

- Intel® Xeon® Processors
- Intel® Xeon Phi™ Processors*
- Intel® Core™ & Atom™ Processors
- CPU+ Intel® Processor Graphics
- Intel® FPGA
- Crest Family (Nervana ASIC)*
- Movidius VPU
- Intel® GNA (IP)*

*Future

AI DATACENTER

ALL PURPOSE

Intel® Xeon® Processor Family

MOST AGILE AI PLATFORM

Scalable performance for widest variety of AI & other datacenter workloads – including deep learning training & inference

HIGHLY-PARALLEL

Intel® Xeon Phi™ Processor (Knights Mill†)

FASTER DL TRAINING

Scalable performance optimized for even faster deep learning training and select highly-parallel datacenter workloads*

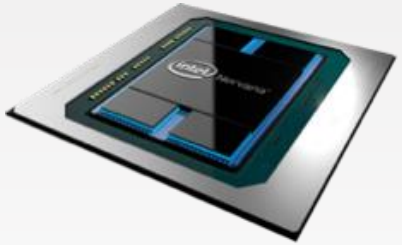
FLEXIBLE ACCELERATION

Intel® FPGA

ENHANCED DL INFERENCE

Scalable acceleration for deep learning inference in real-time with higher efficiency, and wide range of workloads & configurations

DEEP LEARNING



Crest Family†

DEEP LEARNING BY DESIGN

Scalable acceleration with best performance for intensive deep learning training & inference

†Codename for product that is coming soon
All performance positioning claims are relative to other processor technologies in Intel's AI datacenter portfolio
*Knights Mill (KNM); select = single-precision highly-parallel workloads generally scale to >100 threads and benefit from more vectorization, and may also benefit from greater memory bandwidth e.g. energy (reverse time migration), deep learning training, etc.
All products, computer systems, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.

AI GATEWAY/EDGE

ALL PURPOSE

Intel® Processors

AGILE AI PLATFORMS

Range of performance and power for widest variety of AI, gateway & edge workloads – including deep learning inference

GRAPHICS ACCEL

CPU+ Intel® Processor Graphics*

HIGHER DL INFERENCE

Higher deep learning inference throughput and performance for graphics-intensive gateway & edge workloads**

FLEXIBLE ACCEL

Intel® Arria® 10 FPGA

ENHANCED DL INFERENCE

Acceleration for deep learning inference in real-time with higher efficiency, and wide range of workloads & configurations

VISION

Movidius Myriad

COMPUTER VISION

Ultra-low power computer vision engine using deep learning inference in gateway and devices

SPEECH

GNA

Intel® GNA† (Future)

SPEECH RECOGNITION

Ultra-low power speech recognition engine using deep learning inference in devices

†GNA = Gaussian Mixture Model (GMM) and Neural Network Accelerator (IP block)
*Intel® HD Graphics, Intel® Iris™ Graphics, Intel® Iris™ Pro Graphics
**Graphics-intensive workloads include, for example, transcode, edge detection, image enhancement, 3D remote workstation, rendering, encoding, OpenCL/GL...
All products, computer systems, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.

INTEL® XEON® PROCESSOR SCALABLE FAMILY

Scalable performance for widest variety of AI & other datacenter workloads – including deep learning

MOST AGILE AI PLATFORM

BUILT-IN ROI

Begin your AI journey today using existing, familiar infrastructure

POTENT PERFORMANCE

Train in ~~days~~ HOURS with up to 113X[2] perf vs. prior gen (2.2x excluding optimized SW[1])

PRODUCTION-READY

Robust support for full range of AI deployments

[1,2]Configuration details on slide: 4, 5, 6
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 Source: Intel measured as of November 2016
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

INTEL® XEON PHI™ PROCESSOR (KNIGHTS MILL)

COMING 2017

FASTER DL TRAINING

*Scalable performance optimized for even faster deep learning training and select highly-parallel datacenter workloads**

FASTER TIME-TO-TRAIN

- Delivers up to **4X** deep learning performance over Knights Landing†
- New instructions sets deliver enhanced lower precision performance
- Time-to-train reduction is the primary benchmark to judge deep learning training performance

EFFICIENT SCALING

- Direct access of up to 400 GB of memory with no PCIe performance lag (vs. GPU:16GB)
- Efficient scaling further reduces time-to-train when utilizing scaled Knights Mill systems

FUTURE READY

- Up to **400X** deep learning performance on existing HW via Intel SW optimization
- Share deep learning software investments across Intel Platforms via Intel deep learning software tools
- Binary-compatible with Intel® Xeon® processor

†Knights Landing is the former codename for the Intel® Xeon Phi™ processor family that was released in 2016
Configuration details on final slides
**Knights Mill (KNM); select = single-precision highly-parallel workloads generally scale to >100 threads and benefit from more vectorization, and may also benefit from greater memory bandwidth e.g. energy (reverse time migration), deep learning training, etc. 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 Source: Intel measured as of November 2016*
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

INTEL® ARRIA 10 FPGA

ENHANCED DL INFERENCE

Scalable acceleration for deep learning inference in real-time with higher efficiency, and wide range of workloads & configurations

POWER EFFICIENCY

- Up to **80%** reduction in power consumption (vs. Intel® Xeon® processor), with inference up to **25 images/sec/watt** on Caffe running Alexnet

HIGH-THROUGHPUT, LOW LATENCY

- Deterministic, real-time, inline processing of streaming data without buffering.

  (as low as **<3ms**[1] latency)

FLEXIBLE INFRASTRUCTURE

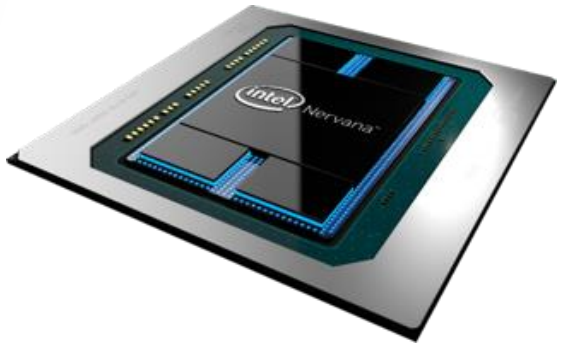
- Future proof for new neural network topologies, arbitrary precision data types (FloatP32 => FixedP2, sparsity, weight sharing), inline & offload processing
- Reconfigure accelerator for a variety of workloads with fast switching

Note: available as discrete or Xeon with Integrated FPGA (Broadwell Proof of Concept)
Configuration details on final slides
[1]Includes system and PCIe latency (GoogleNet); NVIDIA only publishes latency for more compute intensive topologies, doesn't specify what's included, and is in the ~5-25ms range
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 Source: Intel measured as of November 2016
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

CREST FAMILY

1ST SILICON 2017

DEEP LEARNING BY DESIGN



Scalable acceleration with best performance for intensive deep learning training & inference, period

CUSTOM HARDWARE

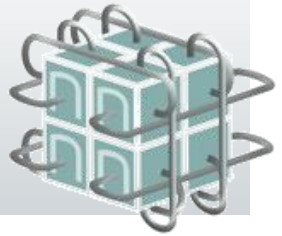
- Unprecedented compute density
- Large reduction in time-to-train

BLAZING DATA ACCESS

- 32 GB of in package memory via HBM2 technology
- 8 Tera-bits/s of memory access speed

HIGH-SPEED SCALABILITY

- 12 bi-directional high-bandwidth links
- Seamless data transfer via interconnects



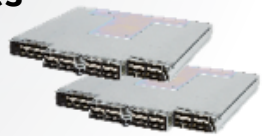
[1]Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance Software and workloads used in performance 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: http://www.intel.com/performance Source: Intel measured as of November 2016

INTEL® OMNI-PATH ARCHITECTURE

World-Class Interconnect Solution for Shorter Time to Train

HFI Adapters
Single port
x8 and x16

Edge Switches
1U Form Factor
24 and 48 port



Director Switches
QSFP-based
192 and 768 port

Software
Open Source
Host Software and Fabric Manager

Cables
Third Party Vendors
Passive Copper
Active Optical

Fabric interconnect for breakthrough performance on scale-out workloads like deep learning training

STRONG FOUNDATION

- Highly leverage existing Aries & Intel True Scale fabrics
- Excellent price/performance ⍰⍰ price/port, 48 radix
- Re-use of existing OpenFabrics Alliance Software
- Over 80+ Fabric Builder Members

BREAKTHROUGH PERFORMANCE

- Increases price performance, reduces communication latency compared to InfiniBand EDR[1]:
  - Up to **21%** Higher Performance, lower latency at scale
  - Up to **17%** higher messaging rate
  - Up to **9%** higher application performance

INNOVATIVE FEATURES

- Improve performance, reliability and QoS through:
  - Traffic Flow Optimization to maximize QoS in mixed traffic
  - Packet Integrity Protection for rapid and transparent recovery of transmission errors
  - Dynamic lane scaling to maintain link continuity

[1]Intel® Xeon® Processor E5-2697A v4 dual-socket servers with 2133 MHz DDR4 memory. Intel® Turbo Boost Technology and Intel® Hyper Threading Technology enabled. BIOS: Early snoop disabled, Cluster on Die disabled, IOU non-posted prefetch disabled, Snoop hold-off timer=9. Red Hat Enterprise Linux Server release 7.2 (Maipo). Intel® OPA testing performed with Intel Corporation Device 24f0 – Series 100 HFI ASIC (B0 silicon). OPA Switch: Series 100 Edge Switch – 48 port (B0 silicon). Intel® OPA host software 10.1 or newer using Open MPI 1.10.x contained within host software package. EDR IB* testing performed with Mellanox EDR ConnectX-4 Single Port Rev 3 MCX455A HCA. Mellanox SB7700 - 36 Port EDR Infiniband switch. EDR tested with MLNX_OFED_Linux-3.2.x. OpenMPI 1.10.x contained within MLNX HPC-X. Message rate claim: Ohio State Micro Benchmarks v. 5.0. osu_mbw_mr, 8 B message (uni-directional), 32 MPI rank pairs. Maximum rank pair communication time used instead of average time, average timing introduced into Ohio State Micro Benchmarks as of v3.9 (2/28/13). Best of default, MXM_TLS=self,rc, and -mca pml yalla tunings. All measurements include one switch hop. Latency claim: HPCC 1.4.3 Random order ring latency using 16 nodes, 32 MPI ranks per node, 512 total MPI ranks. Application claim: GROMACS version 5.0.4 ion_channel benchmark. 16 nodes, 32 MPI ranks per node, 512 total MPI ranks. Intel® MPI Library 2017.0.064. Additional configuration details available upon request.

AI FRAMEWORKS

SELECT YOUR FAVORITE AI FRAMEWORK

and more frameworks enabled via Intel® Nervana™ Graph (future)

Intel®'s reference deep learning framework committed to best performance on all hardware
intelnervana.com/neon

✓OPTIMIZED FOR INTEL ARCHITECTURE

See **Roadmap** for availability

INTEL DISTRIBUTION FOR PYTHON

Advancing Python Performance Closer to Native Speeds

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

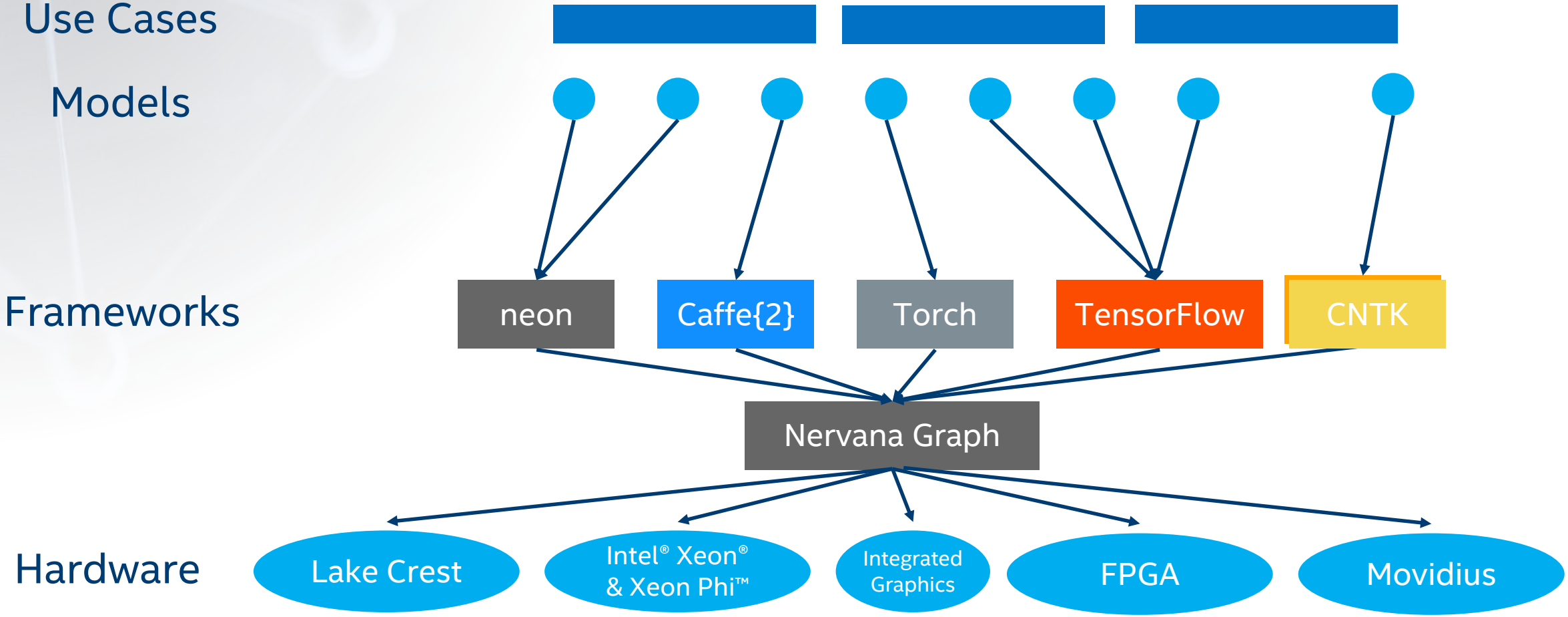
- Distribution and individual optimized packages available through conda and Anaconda Cloud
- Optimizations upstreamed back to main Python trunk

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

INTEL® NERVANA™ GRAPH

COMING SOON

High-Performance Execution Graph for Neural Networks



INTEL® NERVANA™ PLATFORM

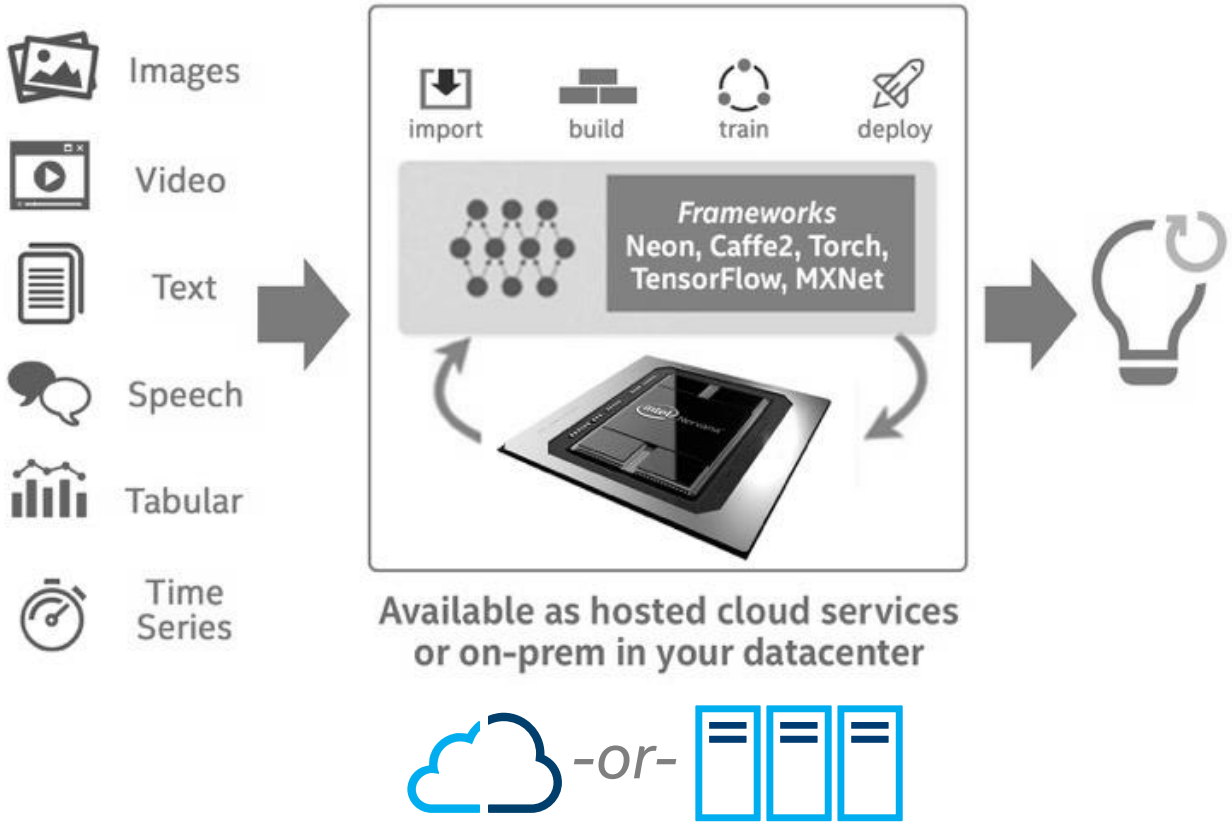
World's Most Advanced Deep Learning Platform

A full stack, user-friendly & turnkey system that enables businesses to develop and deploy high-accuracy AI solutions in record time:

Compress the Development Cycle

Pure Acceleration

Benefits Beyond the Box



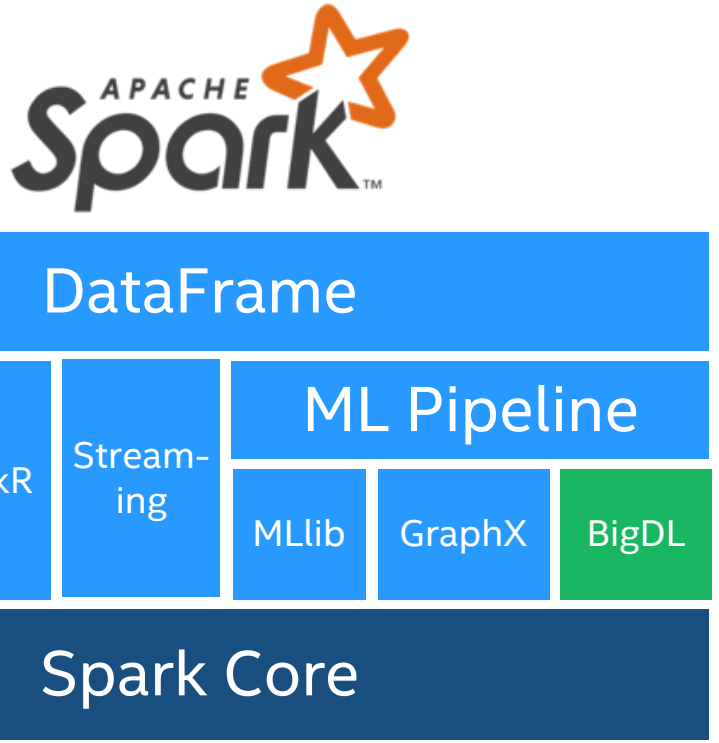
*Other names and brands may be claimed as the property of others.

BIGDL

Bringing Deep Learning to Big Data

For developers looking to run deep learning on Hadoop/Spark due to familiarity or analytics use

- **Open Sourced** Deep Learning Library for Apache Spark*
- **Make Deep learning more Accessible** to Big data users and data scientists.
- **Feature Parity** with popular DL frameworks like Caffe, Torch, Tensorflow etc.
- **Easy Customer and Developer Experience**
  - Run Deep learning Applications as Standard Spark programs;
  - Run on top of existing Spark/Hadoop clusters (No Cluster change)
- **High Performance** powered by Intel MKL and Multi-threaded programming.
- **Efficient Scale out** leveraging Spark architecture.



github.com/intel-analytics/BigDL

INTEL® MKL-DNN

Math Kernel Library for Deep Neural Networks

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:

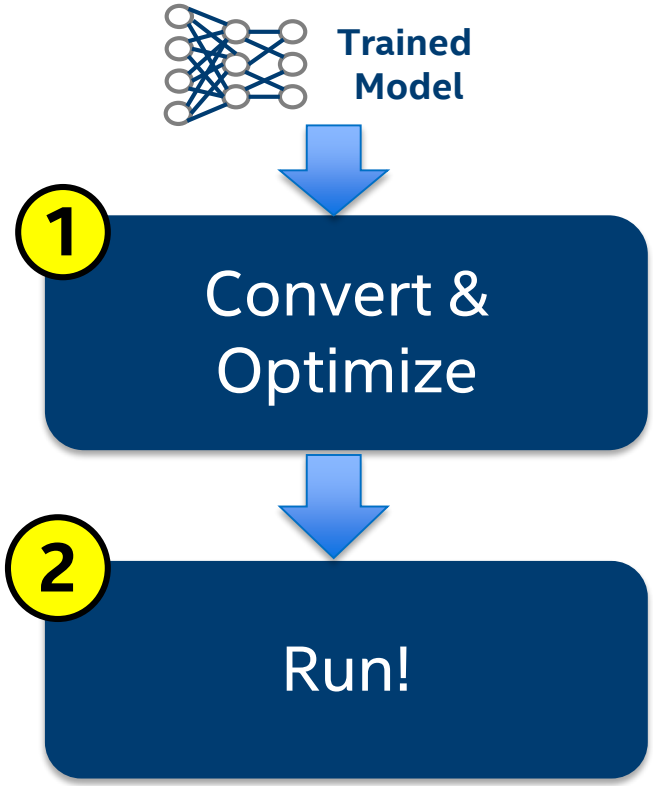
INTEL® DEEP LEARNING DEPLOYMENT TOOLKIT

BETA Now Available!

For developers looking to run deep learning models on the edge

1
- Imports trained models from popular DL framework regardless of training HW
- Enhances model for improved execution, storage & transmission

2
- Optimizes Inference execution for target hardware (computational graph analysis, scheduling, model compression, quantization)
- Enables seamless integration with application logic
- Delivers embedded friendly Inference solution

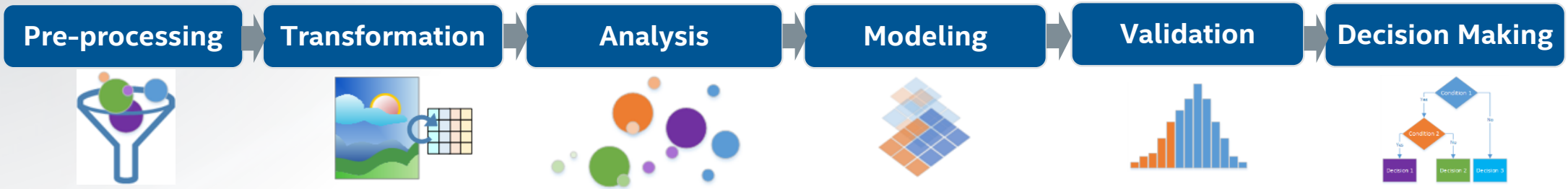


Ease of use + Embedded friendly + Extra performance boost

INTEL® DATA ANALYTICS ACCELERATION LIBRARY (INTEL® DAAL)

High Performance ML and Data Analytics library

Building blocks for all data analytics stages, including data preparation, data mining & machine learning



Open Source • Apache 2.0 License

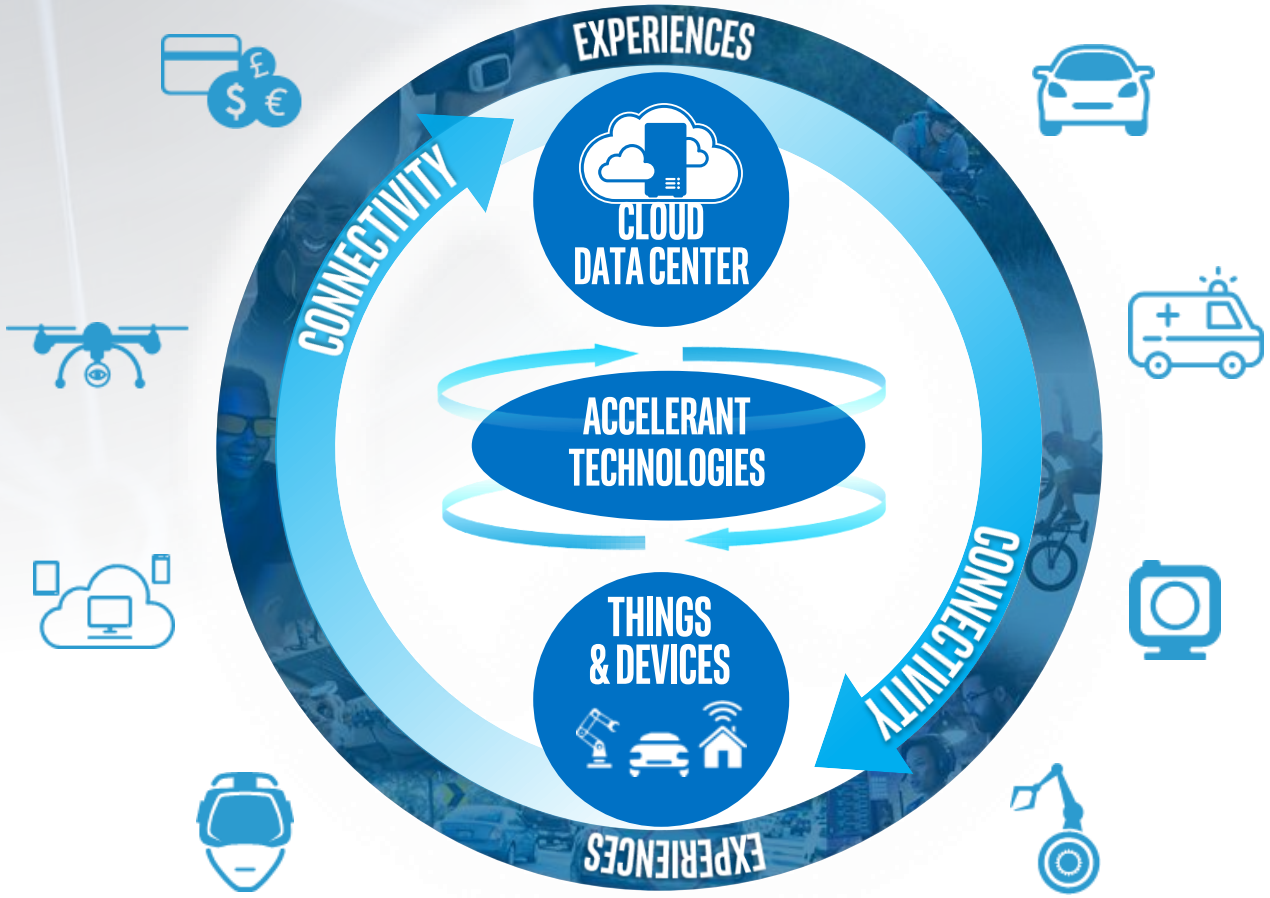
Common Python, Java and C++ APIs across all Intel hardware

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.)

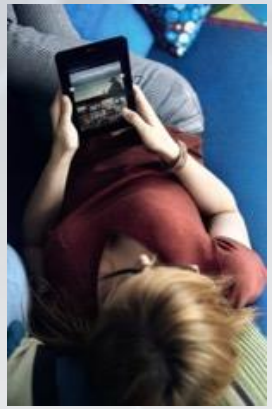
END-TO-END AI

Intel Has a Complete End-to-End Portfolio



- MACHINE/DEEP LEARNING
- REASONING SYSTEMS
- PROGRAMMABLE SOLUTIONS
- COMPUTER VISION
- TOOLS & STANDARDS
- MEMORY/STORAGE
- NETWORKING
- COMMUNICATIONS

AI IS TRANSFORMATIVE









CONSUMER

Smart Assistants

Chatbots

Search

Personalization

Augmented Reality

Robots

HEALTH

Enhanced Diagnostics

Drug Discovery

Patient Care

Research

Sensory Aids

FINANCE

Algorithmic Trading

Fraud Detection

Research

Personal Finance

Risk Mitigation

RETAIL

Support

Experience

Marketing

Merchandising

Loyalty

Supply Chain

Security

GOVERNMENT

Defense

Data Insights

Safety & Security

Resident Engagement

Smarter Cities

ENERGY

Oil & Gas Exploration

Smart Grid

Operational Improvement

Conservation

TRANSPORT

Autonomous Cars

Automated Trucking

Aerospace

Shipping

Search & Rescue

INDUSTRIAL

Factory Automation

Predictive Maintenance

Precision Agriculture

Field Automation

OTHER

Advertising

Education

Gaming

Professional & IT Services

Telco/Media

Sports

Source: Intel forecast

THE COMING FLOOD OF DATA

BY 2020...

The average internet user will generate
~1.5 GB OF TRAFFIC PER DAY

Smart hospitals will generate over
3,000 GB PER DAY

Self driving cars will generate over
4,000 GB PER DAY... EACH

A connected plane will generate over
40,000 GB PER DAY

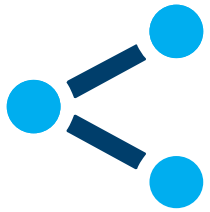
A connected factory will generate over
1,000,000 GB PER DAY

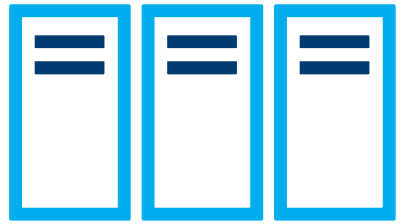
RADAR	~10-100 KB	PER SECOND
SONAR	~10-100 KB	PER SECOND
GPS	~50 KB	PER SECOND
LIDAR	~10-70 MB	PER SECOND
CAMERAS	~20-40 MB	PER SECOND

All numbers are approximated
http://www.cisco.com/c/en/us/solutions/service-provider/vni-network-traffic-forecast/infographic.html
http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.html
https://datafloq.com/read/self-driving-cars-create-2-petabytes-data-annually/172
http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.html
http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.html

END-TO-END EXAMPLE

Automated Driving





VEHICLE

- Driving Functions
  Autonomous Driving Functions
  Trajectory Enumeration, Path Planning, Selection & Maneuvering
  Driving Policy, Path Selection
- Real Time Environment Modeling
  Localization
- Sensor Processing & Fusion
  Object ID & Classification

Anomaly Detection

NETWORK

Data Storage

Dataset Management & Traceability

Compressed Models

OTA SW / FW Updates

Data Formatting

DATACENTER

- Neural Network Design for Target Hardware, & usage
  (Vision, Data Driven, etc.)
- Model Training
  Single & Multi-node optimized Frameworks
- Model Inference >than Real Time
  Model Simulation & Verification

LEADERSHIP AI

COMMITMENT from CEO to lead in AI, including major acquisitions and AIPG formation

RESOURCES including end-to-end portfolio required to unleash the full potential of AI

ECOSYSTEM installed base running majority of AI compute on Intel architecture today

TECHNOLOGY built on the most dense and power efficient transistors on the planet

EXPERIENCE from previously driving several successful computing transitions

Thank You

werner.schueler@intel.com

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.

CONFIGURATION DETAILS

2S Intel® Xeon® processor E5-2697A v4 on Apache Spark™ with MKL2017 up to 18x performance increase compared to 2S E5-2697 v2 + F2JBLAS machine learning training
BASELINE: Intel® Xeon® Processor E5-2697 v2 (12 Cores, 2.7 GHz), 256GB memory, CentOS 6.6*, F2JBLAS: https://github.com/fommil/netlib-java, Relative performance 1.0

Intel® Xeon® processor E5-2697 v2 Apache* Spark* Cluster: 1-Master + 8-Workers, 10Gbit/sec Ethernet fabric, Each system with 2 Processors, Intel® Xeon® processor E5-2697 v2 (12 Cores, 2.7 GHz), Hyper-Threading Enabled, 256GB RAM per System, 1-240GB SSD OS Drive, 12-3TB HDDs Data Drives Per System, CentOS* 6.6, Linux 2.6.32-642.1.1.el6.x86_64, Intel® Intel® MKL 2017 build U1_20160808 , Cloudera Distribution for Hadoop (CDH) 5.7, Apache* Spark* 1.6.1 standalone, OMP_NUM_THREADS=1 set in CDH*, Total Java Heap Size of 200GB for Spark* Master and Workers, Relative performance up to 3.4x

Intel® Xeon® processor E5-2699 v3 Apache* Spark* Cluster: 1-Master + 8-Workers, 10Gbit/sec Ethernet fabric, Each system with 2 Processors, Intel® Xeon® processor E5-2699 v3 (18 Cores, 2.3 GHz), Hyper-Threading Enabled, 256GB RAM per System, 1-480GB SSD OS Drive, 12-4TB HDDs Data Drives Per System, CentOS* 7.0, Linux 3.10.0-229.el7.x86_64, Intel® Intel® MKL 2017 build U1_20160808 , Cloudera Distribution for Hadoop (CDH) 5.7, Apache* Spark* 1.6.1 standalone, OMP_NUM_THREADS=1 set in CDH*, Total Java Heap Size of 200GB for Spark* Master and Workers, Relative performance up to 8.8x

Intel® Xeon® processor E5-2697A v4 Apache* Spark* Cluster: 1-Master + 8-Workers, 10Gbit Ethernet/sec fabric, Each system with 2 Processors, Intel® Xeon® processor E5-2697A v4 (16 Cores, 2.6 GHz), Hyper-Threading Enabled, 256GB RAM per System, 1-800GB SSD OS Drive, 10-240GB SSDs Data Drives Per System, CentOS* 6.7, Linux 2.6.32-573.12.1.el6.x86_64, Intel® MKL 2017 build U1_20160808 , Cloudera Distribution for Hadoop (CDH) 5.7, Apache* Spark* 1.6.1 standalone, OMP_NUM_THREADS=1 set in CDH*, Total Java Heap Size of 200GB for Spark* Master and Workers, Relative performance up to 18x

Machine learning algorithm used for all configurations : Alternating Least Squares ALS Machine Learning Algorithm https://github.com/databricks/spark-perf

Intel® Xeon Phi™ Processor 7250 GoogleNet V1 Time-To-Train Scaling Efficiency up to 97% on 32 nodes
32 nodes of Intel® Xeon Phi™ processor 7250 (68 Cores, 1.4 GHz, 16GB MCDRAM: flat mode), 96GB DDR4 memory, Red Hat* Enterprise Linux 6.7, export OMP_NUM_THREADS=64 (the remaining 4 cores are used for driving communication) MKL 2017 Update 1, MPI: 2017.1.132, Endeavor KNL bin1 nodes, export I_MPI_FABRICS=tmi, export I_MPI_TMI_PROVIDER=psm2, Throughput is measured using "train" command. Data pre-partitioned across all nodes in the cluster before training. There is no data transferred over the fabric while training. Scaling efficiency computed as: (Single node performance / (N * Performance measured with N nodes))*100, where N = Number of nodes

Intel® Caffe: Intel internal version of Caffe

GoogLeNetV1: http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/43022.pdf, batch size 1536

Intel® Xeon Phi ™ processor 7250 up to 400x performance increase with Intel Optimized Frameworks compared to baseline out of box performance
BASELINE: Caffe Out Of the Box, Intel® Xeon Phi™ processor 7250 (68 Cores, 1.4 GHz, 16GB MCDRAM: cache mode), 96GB memory, Centos 7.2 based on Red Hat* Enterprise Linux 7.2, BVLC-Caffe: https://github.com/BVLC/caffe, with OpenBLAS, Relative performance 1.0

NEW: Caffe: Intel® Xeon Phi™ processor 7250 (68 Cores, 1.4 GHz, 16GB MCDRAM: cache mode), 96GB memory, Centos 7.2 based on Red Hat* Enterprise Linux 7.2, Intel® Caffe: : https://github.com/intel/caffe based on BVLC Caffe as of Jul 16, 2016, MKL GOLD UPDATE1, Relative performance up to 400x

AlexNet used for both configuration as per https://papers.nips.cc/paper/4824-Large image database-classification-with-deep-convolutional-neural-networks.pdf, Batch Size: 256

Intel® Xeon Phi™ Processor 7250, 32 node cluster with Intel® Omni Path Fabric up to 97% GoogleNetV1 Time-To-Train Scaling Efficiency
Caffe: Intel® Xeon Phi™ processor 7250 (68 Cores, 1.4 GHz, 16GB MCDRAM: flat mode), 96GB DDR4 memory, Red Hat* Enterprise Linux 6.7, Intel® Caffe: https://github.com/intel/caffe, not publically available yet

export OMP_NUM_THREADS=64 (the remaining 4 cores are used for driving communication)
MKL 2017 Update 1, MPI: 2017.1.132, Endeavor KNL bin1 nodes, export I_MPI_FABRICS=tmi, export I_MPI_TMI_PROVIDER=psm2, Throughput is measured using "train" command. Split the images across nodes and copied locally on each node at the beginning of training. No IO happens over fabric while training.

GoogLeNetV1: http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/43022.pdf, batch size 1536

Intel® Xeon Phi ™ processor Knights Mill up to 4x estimated performance improvement over Intel® Xeon Phi™ processor 7290
BASELINE: Intel® Xeon Phi™ Processor 7290 (16GB, 1.50 GHz, 72 core) with 192 GB Total Memory on Red Hat Enterprise Linux* 6.7 kernel 2.6.32-573 using MKL 11.3 Update 4, Relative performance 1.0

NEW: Intel® Xeon phi™ processor family – Knights Mill, Relative performance up to 4x

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.

CONFIGURATION DETAILS (CONT'D)

- 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
- http://www.intel.com/performance/datacenter. Tested by Intel as of 14 June 2016. Configurations:

1. Faster and more scalable than GPU claim based on Intel analysis and testing
   - Up to 2.3x faster training per system claim based on AlexNet* topology workload (batch size = 1024) using a large image database running 4-nodes Intel Xeon Phi processor 7250 (16 GB MCDRAM, 1.4 GHz, 68 Cores) in Intel® Server System LADMP2312KXXX41, 96GB DDR4-2400 MHz, quad cluster mode, MCDRAM flat memory mode, Red Hat Enterprise Linux* 6.7 (Santiago), 1.0 TB SATA drive WD1003FZEX-00MK2A0 System Disk, running Intel® Optimized DNN Framework (internal development version) training 1.33 million images in 10.5 hours compared to 1-node host with four NVIDIA "Maxwell" GPUs training 1.33 million images in 25 hours (source: http://www.slideshare.net/NVIDIA/gtc-2016-opening-keynote slide 32).
   - Up to 38% better scaling efficiency at 32-nodes claim based on GoogLeNet deep learning image classification training topology using a large image database comparing one node Intel Xeon Phi processor 7250 (16 GB MCDRAM, 1.4 GHz, 68 Cores) in Intel® Server System LADMP2312KXXX41, DDR4 96GB DDR4-2400 MHz, quad cluster mode, MCDRAM flat memory mode, Red Hat* Enterprise Linux 6.7, Intel® Optimized DNN Framework with 87% efficiency to unknown hosts running 32 each NVIDIA Tesla* K20 GPUs with a 62% efficiency (Source: http://arxiv.org/pdf/1511.00175v2.pdf showing FireCaffe* with 32 NVIDIA Tesla* K20s (Titan Supercomputer*) running GoogLeNet* at 20x speedup over Caffe* with 1 K20).
3. Up to 6 SP TFLOPS based on the Intel Xeon Phi processor peak theoretical single-precision performance is preliminary and based on current expectations of cores, clock frequency and floating point operations per cycle. FLOPS = cores x clock frequency x floating-point operations per second per cycle
4. Up to 3x faster single-threaded performance claim based on Intel estimates of Intel Xeon Phi processor 7290 vs. coprocessor 7120 running XYZ workload.
5. Up to 2.3x faster training per system claim based on AlexNet* topology workload (batch size = 1024) using a large image database running 4-nodes Intel Xeon Phi processor 7250 (16 GB MCDRAM, 1.4 GHz, 68 Cores) in Intel® Server System LADMP2312KXXX41, 96GB DDR4-2400 MHz, quad cluster mode, MCDRAM flat memory mode, Red Hat Enterprise Linux* 6.7 (Santiago), 1.0 TB SATA drive WD1003FZEX-00MK2A0 System Disk, running Intel® Optimized DNN Framework, Intel® Optimized Caffe (internal development version) training 1.33 billion images in 10.5 hours compared to 1-node host with four NVIDIA "Maxwell" GPUs training 1.33 billion images in 25 hours (source: http://www.slideshare.net/NVIDIA/gtc-2016-opening-keynote slide 32).
   Up to 38% better scaling efficiency at 32-nodes claim based on GoogLeNet deep learning image classification training topology using a large image database comparing one node Intel Xeon Phi processor 7250 (16 GB MCDRAM, 1.4 GHz, 68 Cores) in Intel® Server System LADMP2312KXXX41, DDR4 96GB DDR4-2400 MHz, quad cluster mode, MCDRAM flat memory mode, Red Hat* Enterprise Linux 6.7, Intel® Optimized DNN Framework with 87% efficiency to unknown hosts running 32 each NVIDIA Tesla* K20 GPUs with a 62% efficiency (Source: http://arxiv.org/pdf/1511.00175v2.pdf showing FireCaffe* with 32 NVIDIA Tesla* K20s (Titan Supercomputer*) running GoogLeNet* at 20x speedup over Caffe* with 1 K20).
   Up to 50x faster training on 128-node as compared to single-node based on AlexNet* topology workload (batch size = 1024) training time using a large image database running one node Intel Xeon Phi processor 7250 (16 GB MCDRAM, 1.4 GHz, 68 Cores) in Intel® Server System LADMP2312KXXX41, 96GB DDR4-2400 MHz, quad cluster mode, MCDRAM flat memory mode, Red Hat Enterprise Linux* 6.7 (Santiago), 1.0 TB SATA drive WD1003FZEX-00MK2A0 System Disk, running Intel® Optimized DNN Framework, training in 39.17 hours compared to 128-node identically configured with Intel® Omni-Path Host Fabric Interface Adapter 100 Series 1 Port PCIe x16 connectors training in 0.75 hours. Contact your Intel representative for more information on how to obtain the binary. For information on workload, see https://papers.nips.cc/paper/4824-Large image database-classification-with-deep-convolutional-neural-networks.pdf.
   Up to 30x software optimization improvement claim based on customer CNN training workload running 2S Intel® Xeon® processor E5-2680 v3 running Berkeley Vision and Learning Center* (BVLC) Caffe + OpenBlas* library and then run tuned on the Intel® Optimized Caffe (internal development version) + Intel® Math Kernel Library (Intel® MKL).