INSIDE AI:

INTEL® NERVANA™ TECHNOLOGY

RALPH DE WARGNY
INTEL SOFTWARE

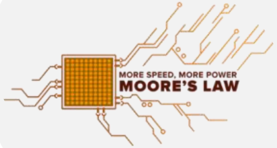
SEPTEMBER 2017

ARTIFICIAL INTELLIGENCE TODAY

Bigger Data

Numbers: 5 KB / record
Text: 500 KB / record
Image: 1000 KB / picture
Audio: 5000 KB / song
Video: 5,000,000 KB / movie
High-Res: 50,000,000 KB / object

Faster Hardware



Transistor density doubles 18m
Computation / kwh doubles 18m

CPUs at over 3 TFlops

Cost / Gigabyte in 1995: $1000.00
Cost / Gigabyte in 2015: $0.03

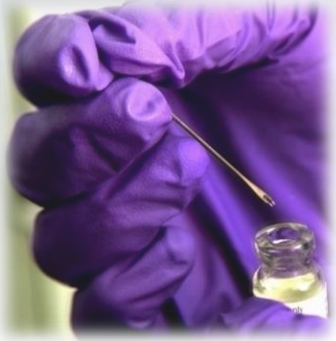
Smarter Algorithms

Theoretical advances in training multi-layer feedforward neural networks led to better accuracy

New mathematical techniques for optimization over non-convex curves led to better learning algorithms

ARTIFICIAL INTELLIGENCE IS CHANGING THE WORLD

On the Scale of the Agricultural, Industrial and Digital Revolutions



ACCELERATE

Large scale solutions

Cure Diseases
Prevent Crime
Unlock Dark Data

UNLEASH

Scientific Discovery

Explore New Worlds
Decode the Brain
Uncover New Theories

EXTEND

Human Capabilities

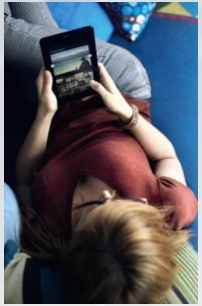
Personalize Learning
Enhance Decisions
Optimize Time

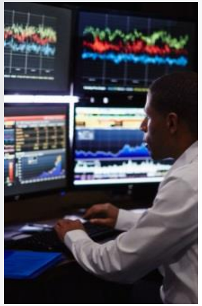
AUTOMATE

Undesirable Tasks

Automate Driving
Save Lives in Danger
Perform Chores

AI IS TRANSFORMATIVE





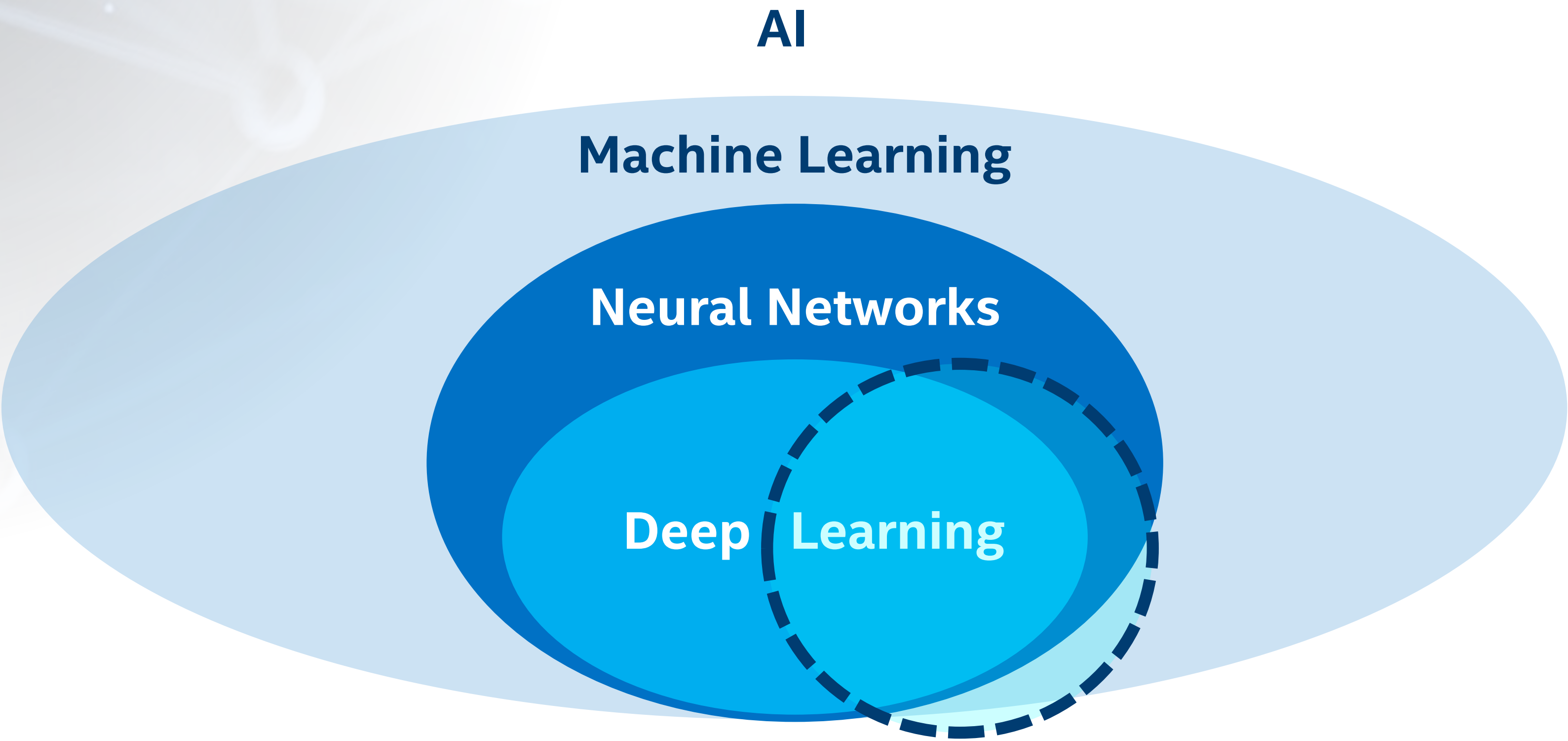




CONSUMER	HEALTH	FINANCE	RETAIL	GOVERNMENT	ENERGY	TRANSPORT	INDUSTRIAL	OTHER
Smart Assistants	Enhanced Diagnostics	Algorithmic Trading	Support	Defense	Oil & Gas Exploration	Autonomous Cars	Factory Automation	Advertising
Chatbots	Drug Discovery	Fraud Detection	Experience	Data Insights	Smart Grid	Automated Trucking	Predictive Maintenance	Education
Search	Patient Care	Research	Marketing	Safety & Security	Operational Improvement	Aerospace	Precision Agriculture	Gaming
Personalization	Research	Personal Finance	Merchandising	Resident Engagement	Conservation	Shipping	Field Automation	Professional & IT Services
Augmented Reality	Sensory Aids	Risk Mitigation	Loyalty	Smarter Cities		Search & Rescue		Telco/Media
Robots			Supply Chain					Sports
			Security					

Source: Intel forecast

THE NEXUS OF AI TODAY

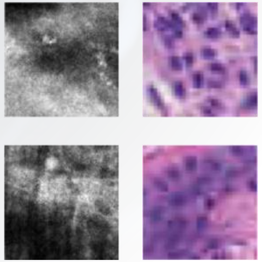


DEEP LEARNING IN PRACTICE

Healthcare: Tumor detection

Positive:

Negative:

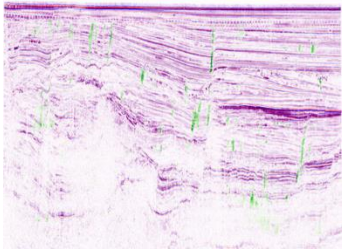


Industry: Agricultural Robotics

Positive:

Negative:

Energy: Oil & Gas

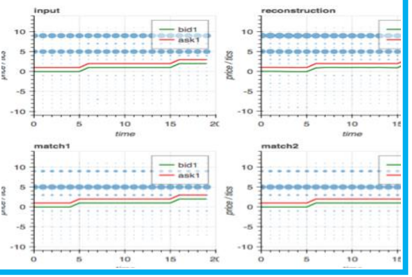


Automotive: Speech interfaces

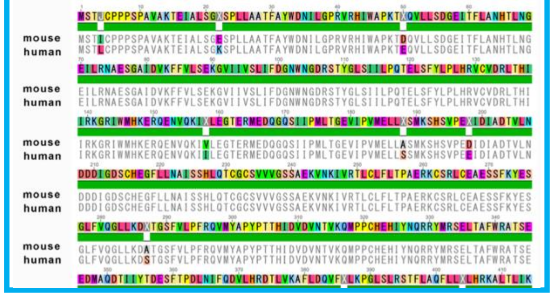
Finance: Time-series search engine

Query:

Results:



Genomics: Sequence analysis



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

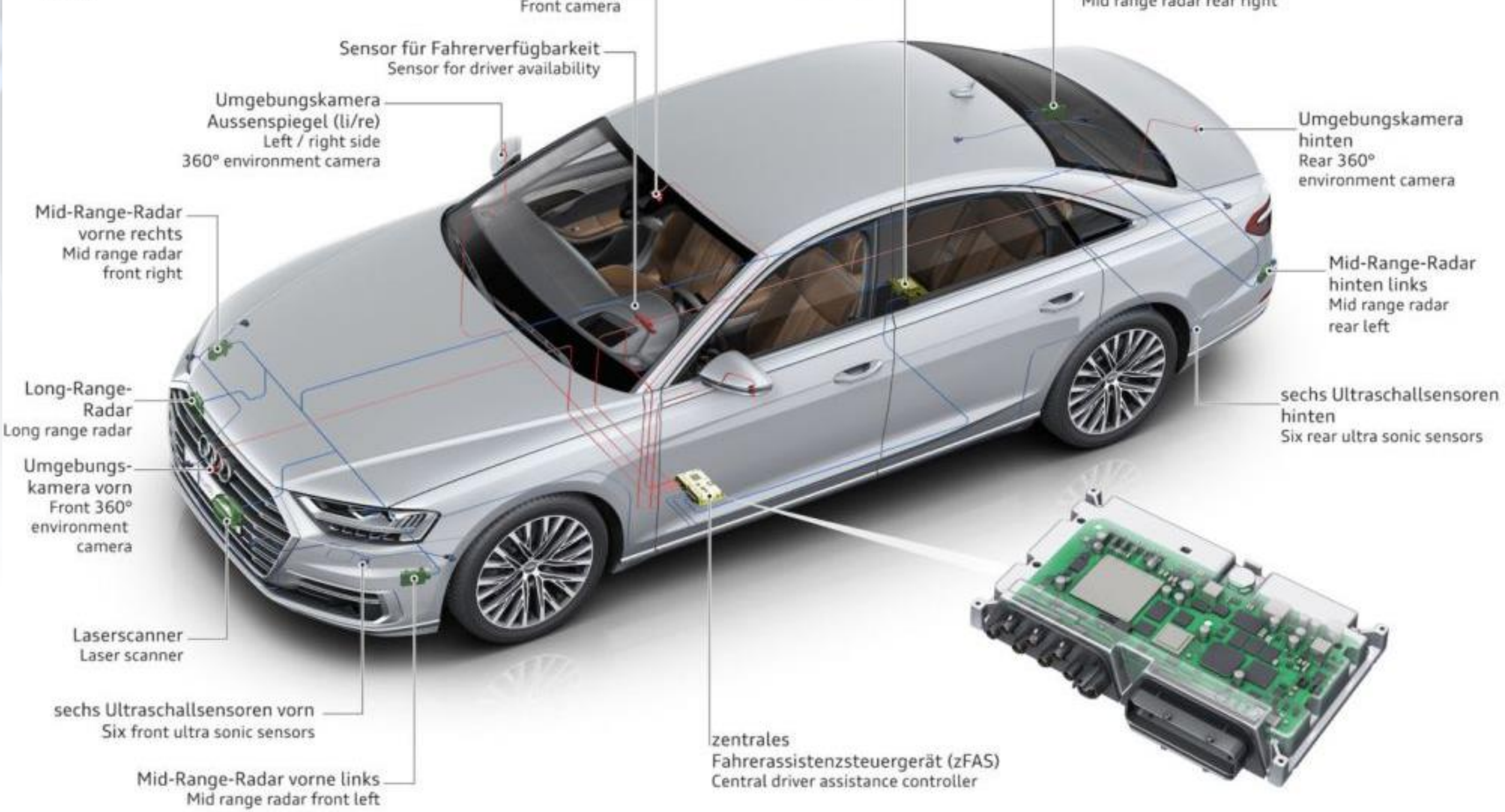
All numbers are approximated
http://www.cisco.com/c/en/us/solutions/service-provider/vni-network-traffic-forecast/infographic.html
http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.html
https://datafloq.com/read/self-driving-cars-create-2-petabytes-data-annually/172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

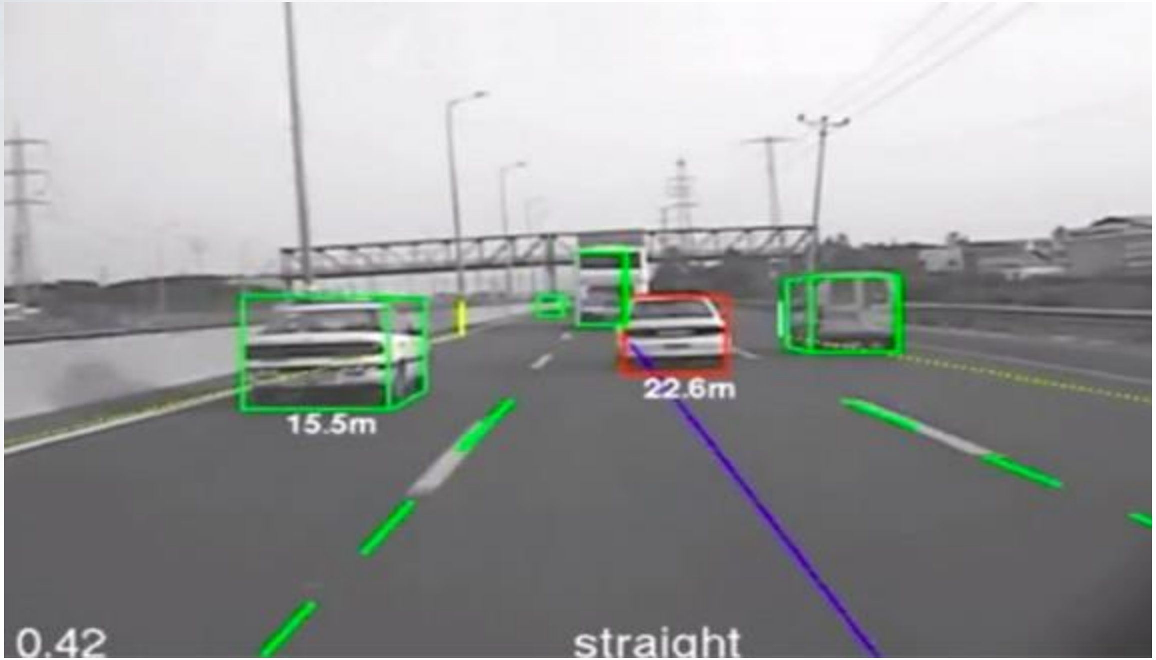
Audi A8

Vernetzung der Sensoren und Kameras

Topology of the sensors and cameras

07/17





An Intel Company

END-TO-END EXAMPLE: AUTONOMOUS DRIVING

IN-VEHICLE

Autonomous Driving Functions
Trajectory Enumeration, Path Planning, Path Selection, Driving Policy, Maneuvering

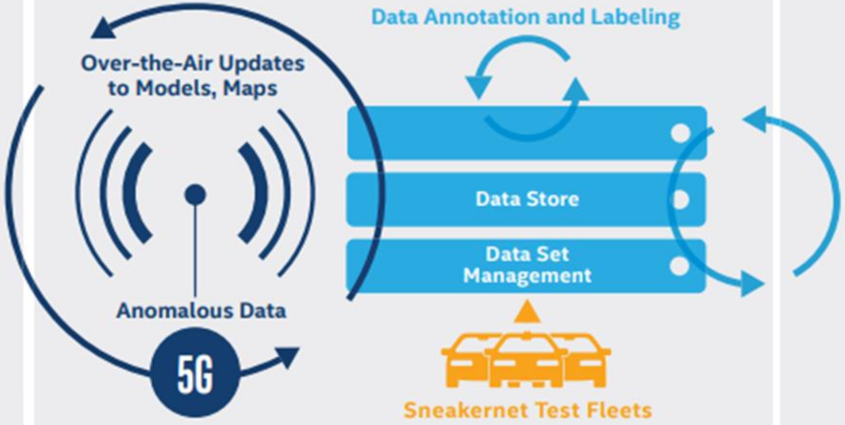
Real-Time Environment Modeling
Localization

Sensor Processing and Fusion
Object ID and Classification, Multimodal, Time-Synchronized

ANOMALY DETECTION

Deep Learning Scoring

COMMUNICATION



DATA CENTER

End Point Management
Geographical Tracking, OTA Updates

Deep Learning

Big Data and Statistical Analytics

- **Model Training**
  Multinode/Intel® Architecture-Optimized Frameworks


intel
HYBRID
GARAGE

INTEL® GO™ AUTOMATED DRIVING SOLUTIONS

CAR

Intel® GO™ development platforms for automated driving deliver the incredibly high compute performance needed for cars to react to changes on the road with split-second agility, support for advanced human-machine interface (HMI) experiences that build trust, and the broadest compute portfolio to let developers code how they need.

CONNECTIVITY

By working closely with device manufacturers and network operators, Intel is paving a path to 5G. We're enabling a more streamlined design process and accelerating prototype development with the launch of our Intel® GO™ automotive 5G platform in 2017.

CLOUD

Intel® technologies for the data center support Intel® GO™ automated driving solutions by scaling to meet the demands of new workloads, including artificial intelligence (AI).

intel.com/automotive

INTEL® NERVANA™ PORTFOLIO

EXPERIENCES

PLATFORMS

Intel® Nervana™ Cloud & Appliance

Intel® Nervana™ DL Studio

Intel® Computer Vision SDK

Movidius Fathom

intel Saffron™

FRAMEWORKS

Apache Spark: MLlib, BigDL

neon

TensorFlow

mxnet

Microsoft CNTK

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: intel XEON inside, intel XEON PHI inside, intel ARRIA 10 inside, *, intel CORE i7 inside, intel ATOM inside, More

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

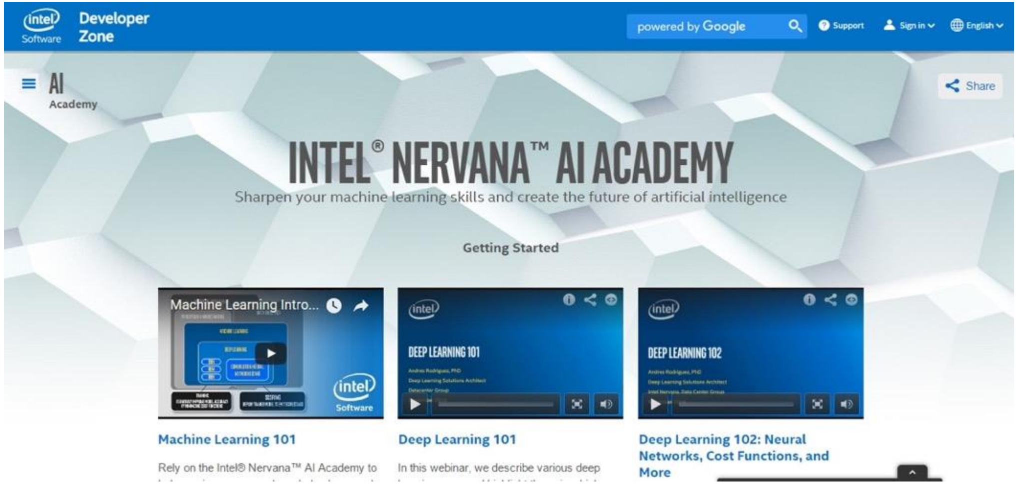
1-to-1 devices with lower power and often UX requirements

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

*Future

INTEL® NERVANA™ AI ACADEMY

- ✓ Intel Developer Zone for Artificial Intelligence
- ✓ Deep Learning Frameworks, libraries and additional tools
- ✓ Workshops, Webinars, Meet Ups & Remote Access



SOFTWARE.INTEL.COM/AI/ACADEMY

intel® Nervana™

CONFIGURATION DETAILS

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).