AIDC

INTEL AI DEVCON 2018

INTEL AI DEVCON 2018

DEEP-LEARNING BASED PHENOTYPE CLASSIFICATION IN HIGH CONTENT CELLULAR IMAGING ON INTEL® ARCHITECTURE

Kushal Datta, PhD
Research Scientist

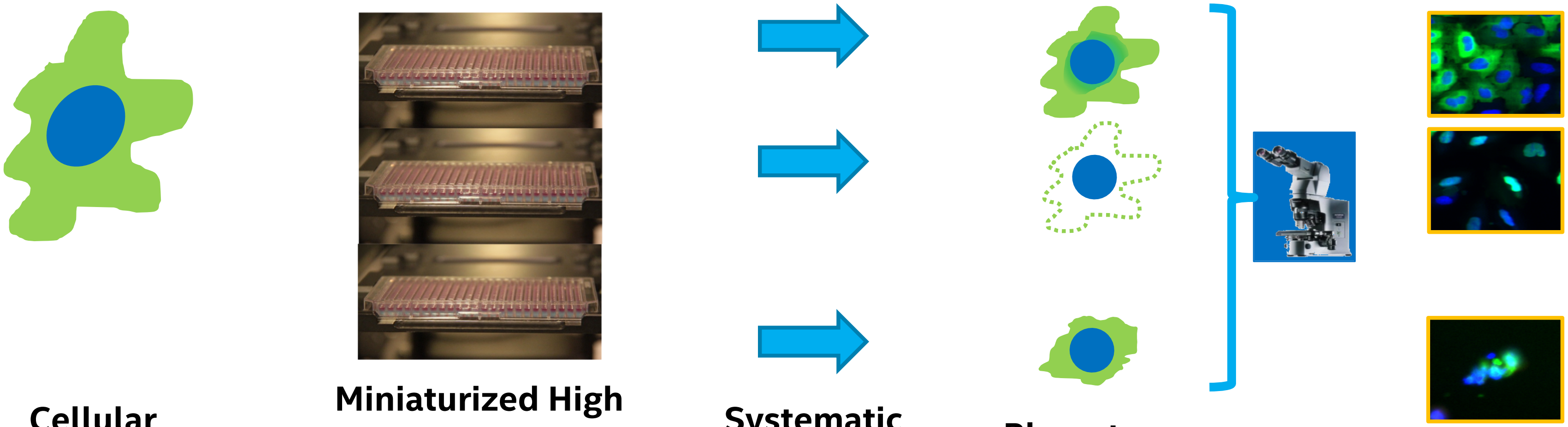
Sun Choi, Vikram Saletore PhD (PE), Mike Demshki, Kyle Ambert PhD

05/24/2018

OPTIMIZATION OF MULTI-SCALE CNNS FOR HIGH-CONTENT SCREENING IN DRUG DISCOVERY

A scientific collaboration between Intel and Novartis

HIGH-CONTENT SCREENING (HCS) IN DRUG DISCOVERY



Cellular Experiment

Miniaturized High Throughput Screening setup

Systematic perturbations

Phenotypes

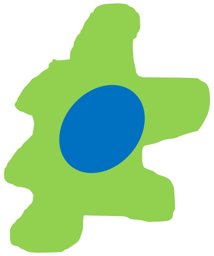
Microscopy images

Multiple experiments and lead discovery projects

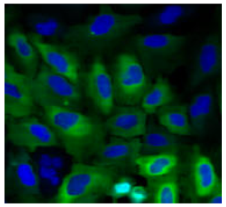
1000s to > 1 Mi compound treatments
per experiment

Multiple channels per image, for different subcellular components
Typically 2560x2160 images, 2-4 channels
Additional multipliers: Dose variations, Time-dependent and 3D measurements.

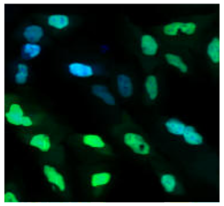
HIGH-CONTENT SCREENING (HCS) IN DRUG DISCOVERY



Cellular Experiments



Positive control



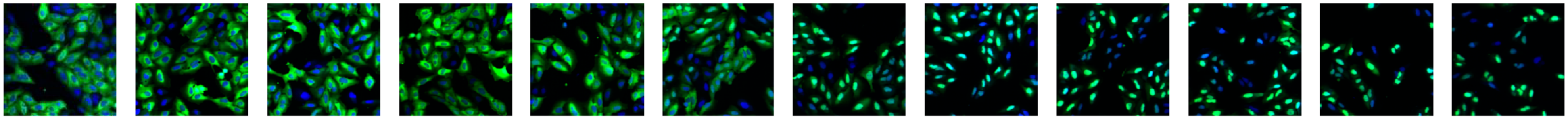
Negative control

BBBC013: Cytoplasm to nucleus translocation of the Forkhead (FKHR-EGFP) fusion protein in stably transfected human osteosarcoma cells, U2OS

Negative control: Wild type

Positive control: Fat

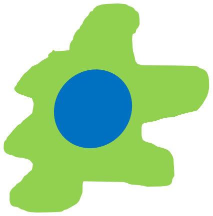
BBBC011: RNA interference screen for genes affecting fat mass in the *C.elegans*. The animals are stained with oil red O, a non-flourescent dye

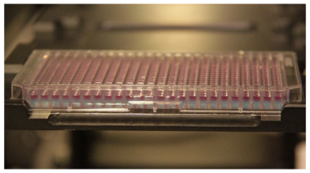


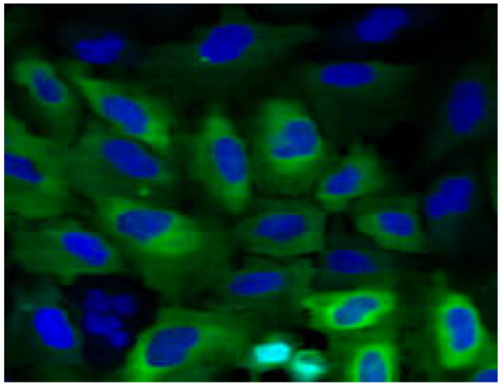
Example from BBBC013 dataset: compound concentration dependent phenotype change

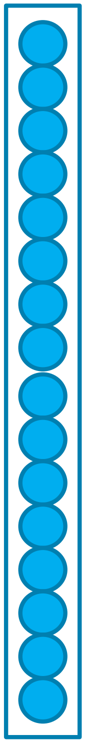
Ljosa V, Sokolnicki KL, Carpenter AE (2012). Annotated high-throughput microscopy image sets for validation. Nature Methods

HIGH-CONTENT SCREENING (HCS) IN DRUG DISCOVERY









High Content Feature Set

Morphometric features

- Nucleus/Cell shape and size
- Position and orientation
- Geometric properties
- Contour – Curvature and Fourier descriptors
- Invariant moment

Densitometric features expressing total intensity

- Intensity features from different regions
- Invariant moments from the extinction image

Textural features

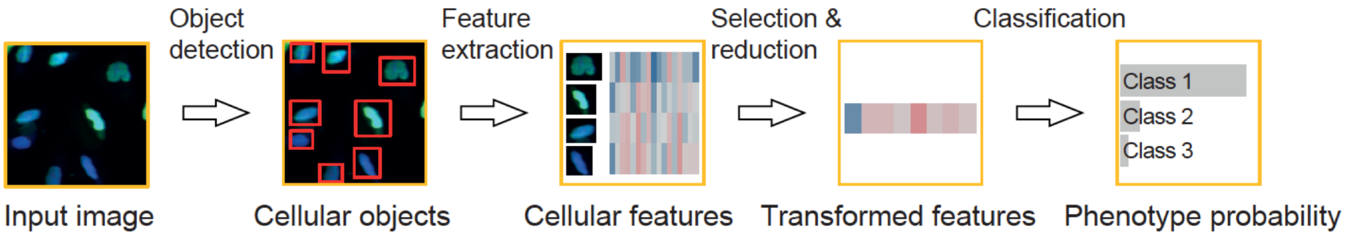
- Gradient/Laplace/Flat texture image features
- Topological gradients and rice fields

Others

- Orlov, Nikita et al. “WND-CHARM: Multi-Purpose Image Classification Using Compound Image Transforms.”, *Pattern recognition letters*, 2008
- Rodenacker K, Bengtsson E. A Feature Set for Cytometry on Digitized Microscopic Images. *Analytical Cellular Pathology : the Journal of the European Society for Analytical Cellular Pathology*, 2003
- Carpenter, Anne E et al. “CellProfiler: Image Analysis Software for Identifying and Quantifying Cell Phenotypes.” *Genome Biology*, 2006

HIGH-CONTENT SCREENING (HCS) IN DRUG DISCOVERY

Conventional HCS analysis pipeline



Object detection

Feature extraction

Selection & reduction

Classification

Class 1
Class 2
Class 3

Input image

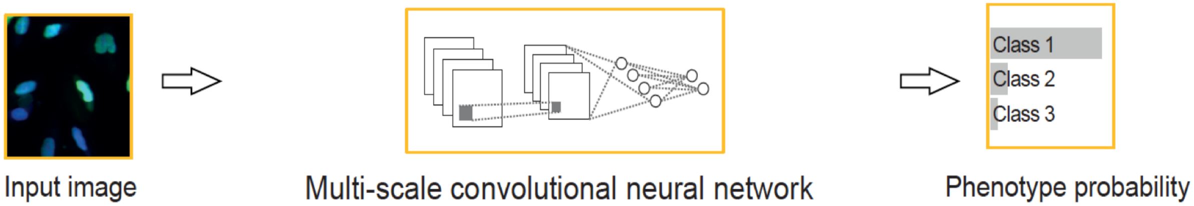
Cellular objects

Cellular features

Transformed features

Phenotype probability

HCS analysis using Multi-Scale Convolutional Network[1]



Class 1
Class 2
Class 3

Input image

Multi-scale convolutional neural network

Phenotype probability

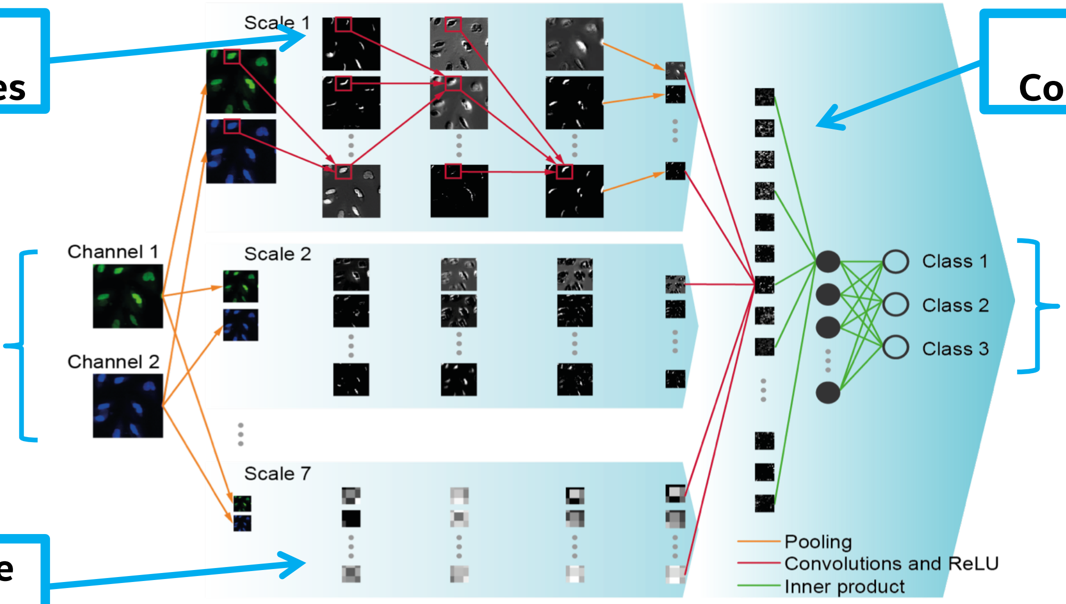
[1] Godinez et al, A multi-scale convolutional neural network for phenotyping high-content cellular images. Bioinformatics, 2017

MULTI-SCALE CONVOLUTIONAL NEURAL NETWORK FOR PHENOTYPE PREDICTION

Analyze **in parallel** different scales and features

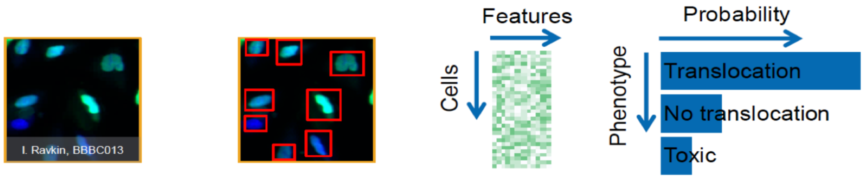
Fine features

Feature Co-localization



Coarse features

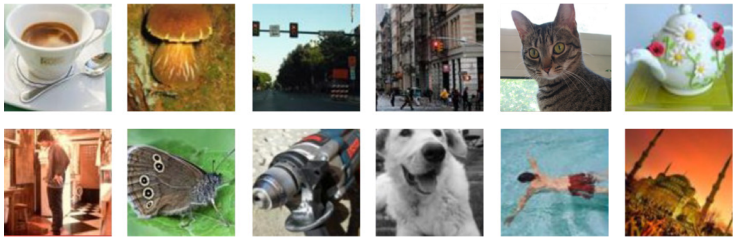
COMPARING DEEP LEARNING APPROACHES IN HCS



	Multi-cellular image	Individual cellular objects	Cellular features	Phenotype prediction
Kandaswamy *et al.*, J. Biomol. Screen 2016			→	→
Dürr & Sick, J. Biomol. Screen 2016		→	→	→
Kraus *et al.*, Mol. Syst. Biol. 2017		→	→	→
Godinez *et al.*, Bioinformatics 2017	→	→	→	→
Pawlowski *et al.*, Proc. NIPS MLCP 2016		→	→	
Ando *et al.*, BioArxiv 2017		→	→	
Sommer *et al.*, BioArxiv 2017		→	→	

IMAGE SIZES

ImageNet

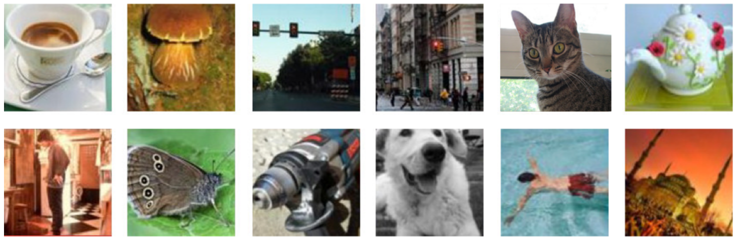


224x224x3

Olga Russakovsky et al, **ImageNet Large Scale Visual Recognition Challenge**. *IJCV*, 2015

IMAGE SIZES

ImageNet



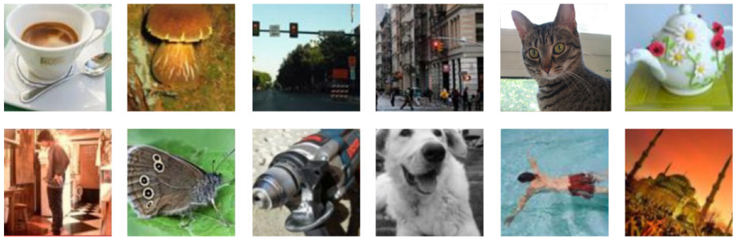
224x224x3

26x

Olga Russakovsky et al, **ImageNet Large Scale Visual Recognition Challenge**. *IJCV*, 2015

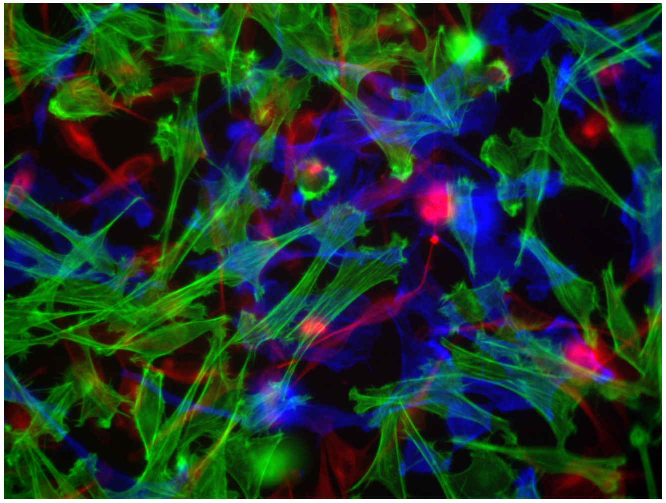
HIGH CONTENT CELLULAR IMAGES ARE LARGE

ImageNet



224x224x3

26x



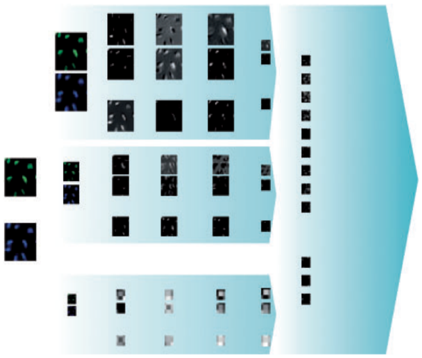
1024x1280x3

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

Olga Russakovsky et al, **ImageNet Large Scale Visual Recognition Challenge**. *IJCV*, 2015

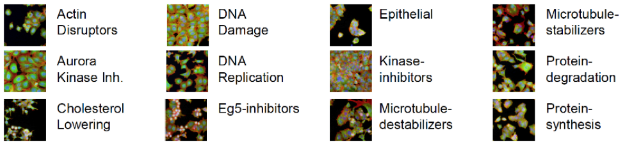
2017: M-CNN ACHIEVES PROMISING RESULTS IN HIGH CONTENT SCREENING

THE COLLABORATION STARTED FROM



M-CNN

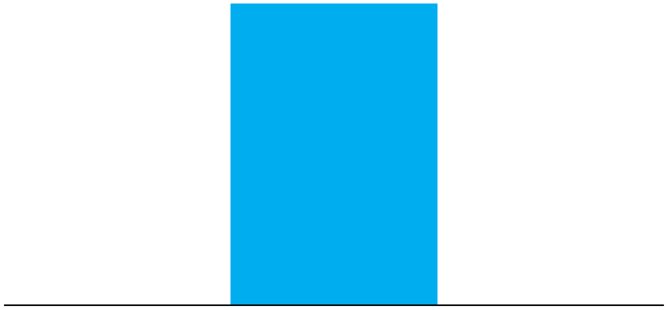
Cell line	Human breast cancer cells (MCF7)
Modality	Fluorescence
Channels	Actin (red) Tubulin (green), DNA (blue)
Treatment	113 compounds at 8 concentrations in triplicate
Task	<u>Predict mechanism-of-action (MoA) of compounds</u>
Number of images	13200
Validation images	1292 - 103 compound-concentration pairs with annotated MoA



Broad Bioimage Benchmark Collection 021

Intel® Xeon Phi™ 7290F processor

Training Time: 11hrs
(Batch Size: 8)



1 Node

Intel® Distribution of Caffe* 2017

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

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

STRATEGY

USE LATEST INTEL® XEON® PROCESSOR BASED PLATFORM

IMPROVE PERFORMANCE WITH TENSORFLOW*

MAXIMIZE PERFORMANCE ON A SINGLE NODE

SCALEOUT PERFORMANCE WITH MULTIPLE NODES

STRATEGY

USE LATEST INTEL® XEON® PROCESSOR BASED PLATFORM

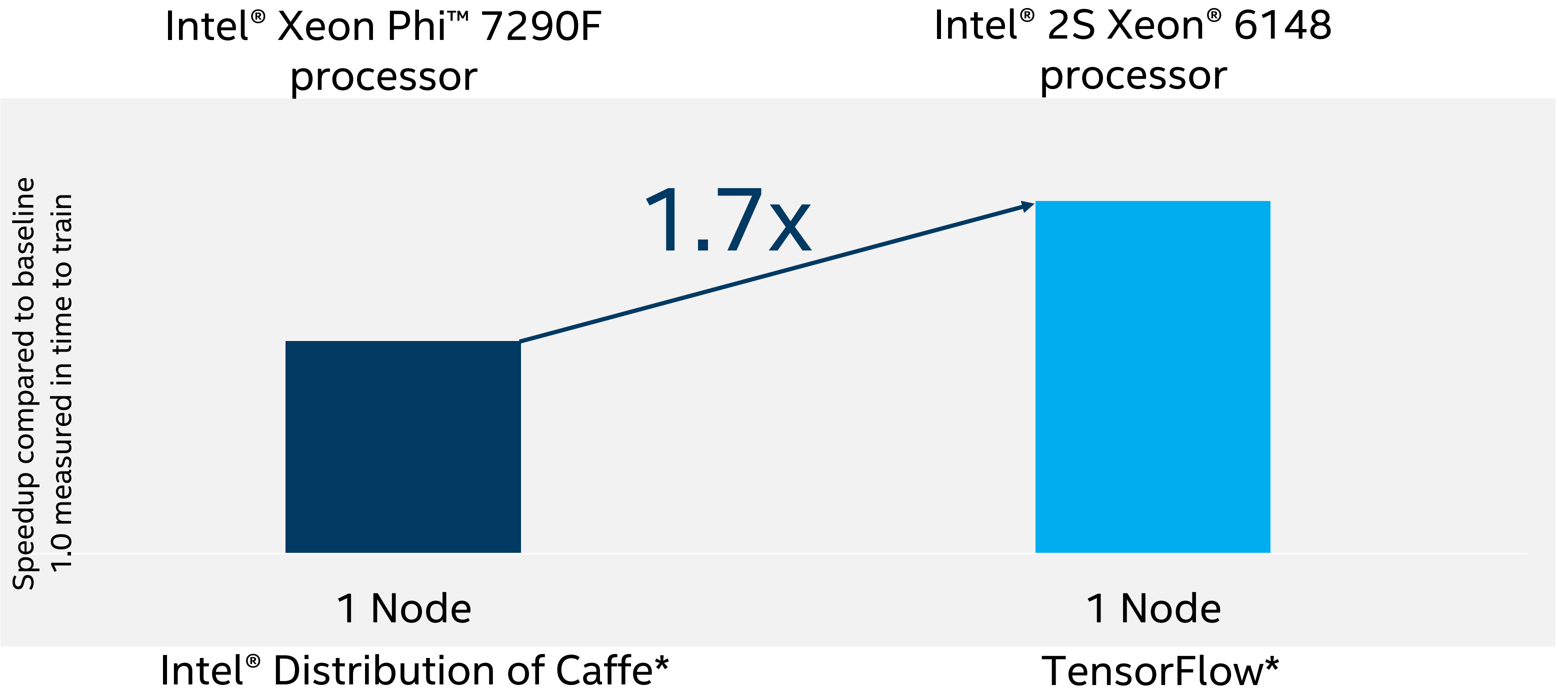
IMPROVE PERFORMANCE WITH TENSORFLOW*

MAXIMIZE PERFORMANCE ON A SINGLE NODE

SCALEOUT PERFORMANCE WITH MULTIPLE NODES

SPEEDUP IN TIME TO TRAIN WITH INTEL® XEON® GOLD SCALABLE PROCESSOR

BATCH SIZE 8



TensorFlow* - 1.3, Caffe* 2017

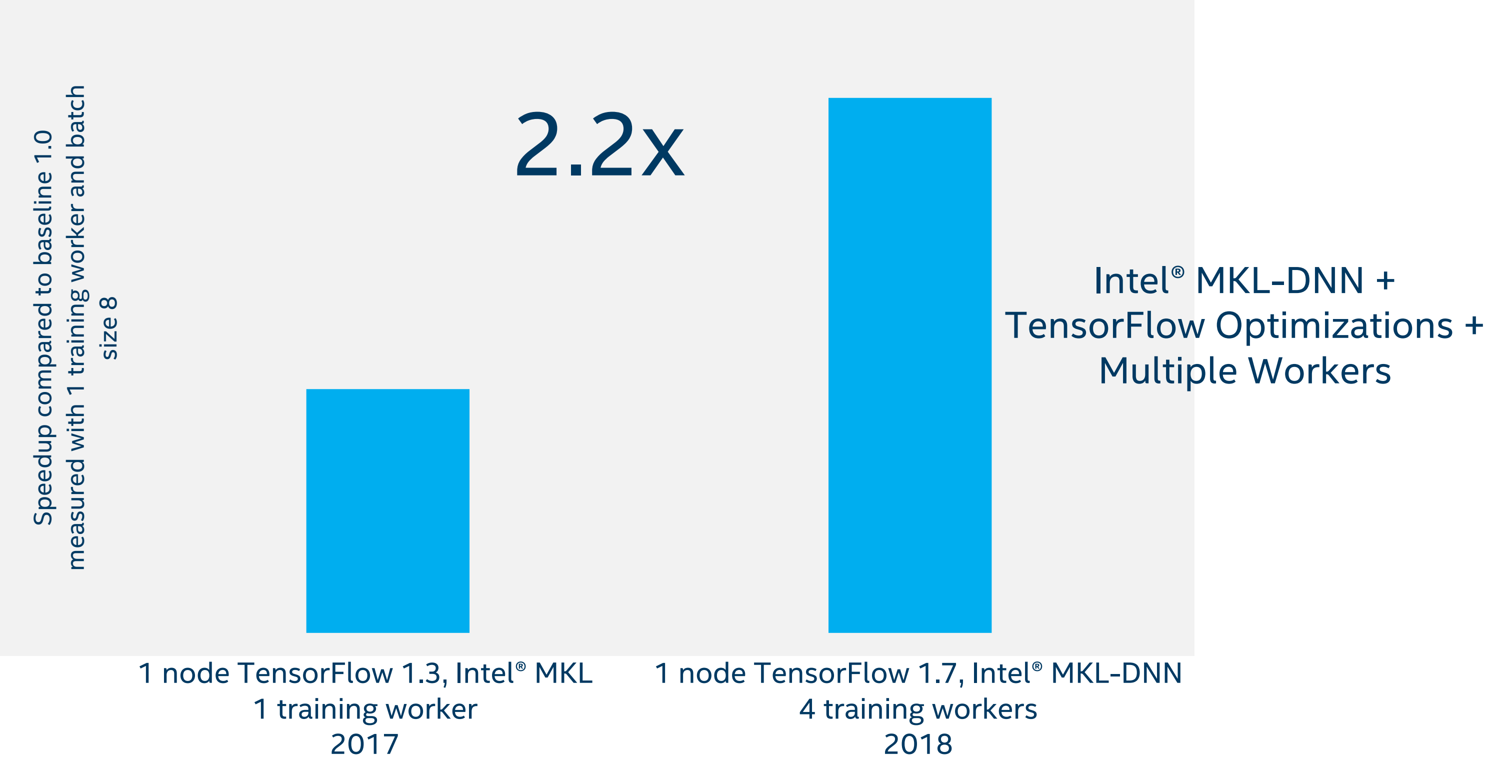
Godinez et al, A multi-scale convolutional neural network for phenotyping high-content cellular images. Bioinformatics, 2017

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

STRATEGY

- USE LATEST INTEL® XEON® PROCESSOR BASED PLATFORM ✓
- IMPROVE PERFORMANCE WITH TENSORFLOW* ✓
- **MAXIMIZE PERFORMANCE ON A SINGLE NODE**
- SCALEOUT PERFORMANCE WITH MULTIPLE NODES

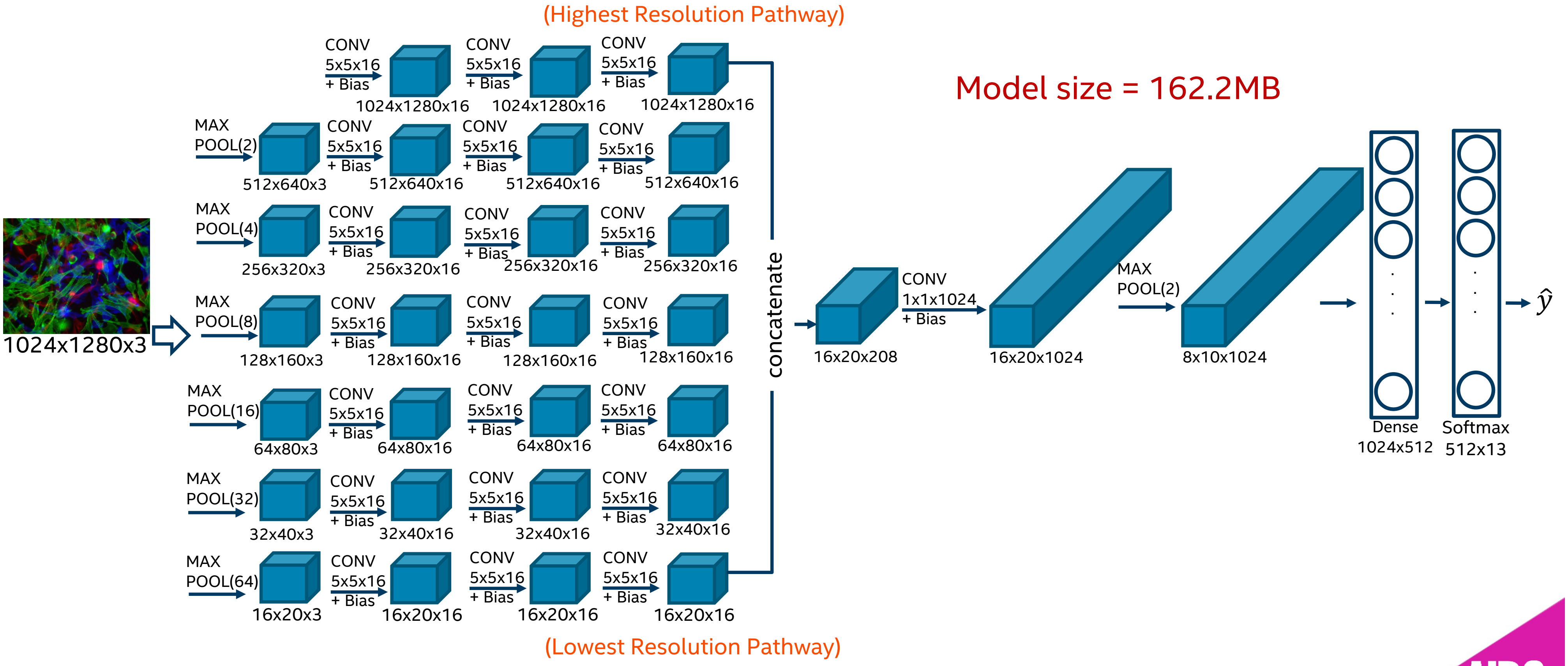
>2X PERFORMANCE IMPROVEMENT IN 6 MONTHS



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

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

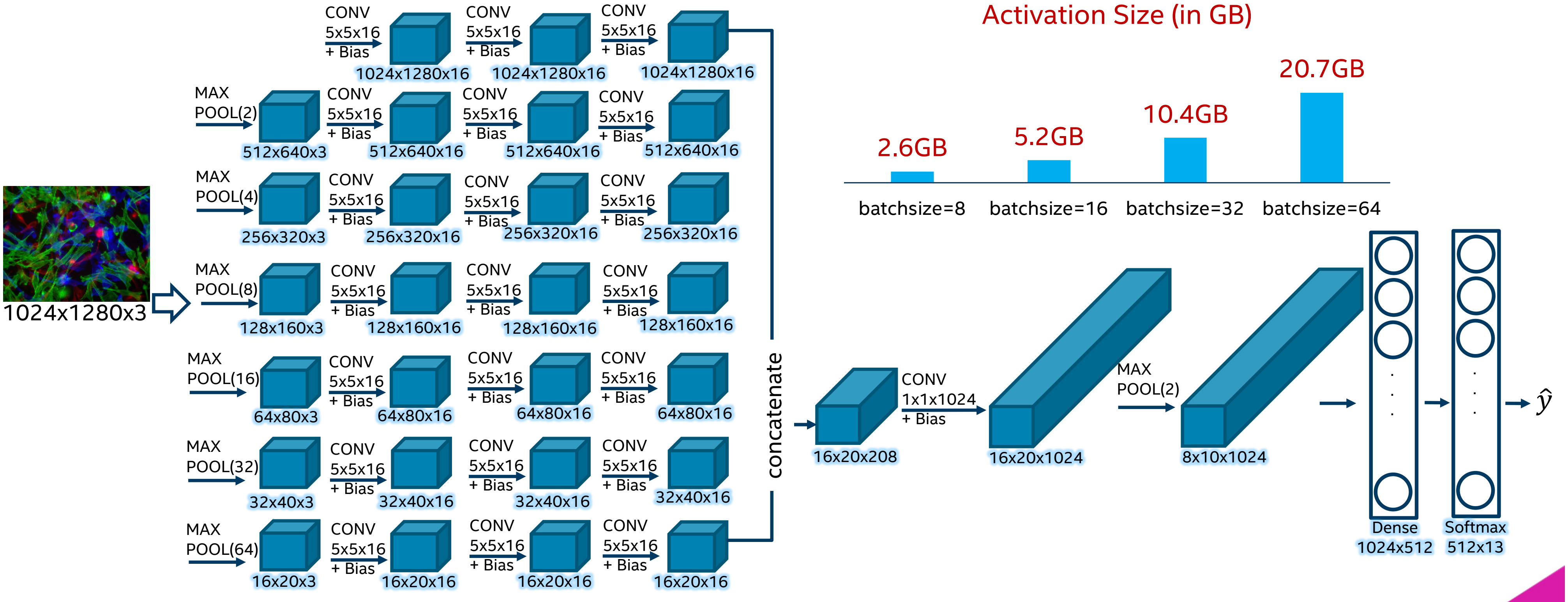
OPERATIONS AND KERNELS IN M-CNN



Model size = 162.2MB

SIZE OF ACTIVATION INCREASES LINEARLY WITH BATCH SIZE

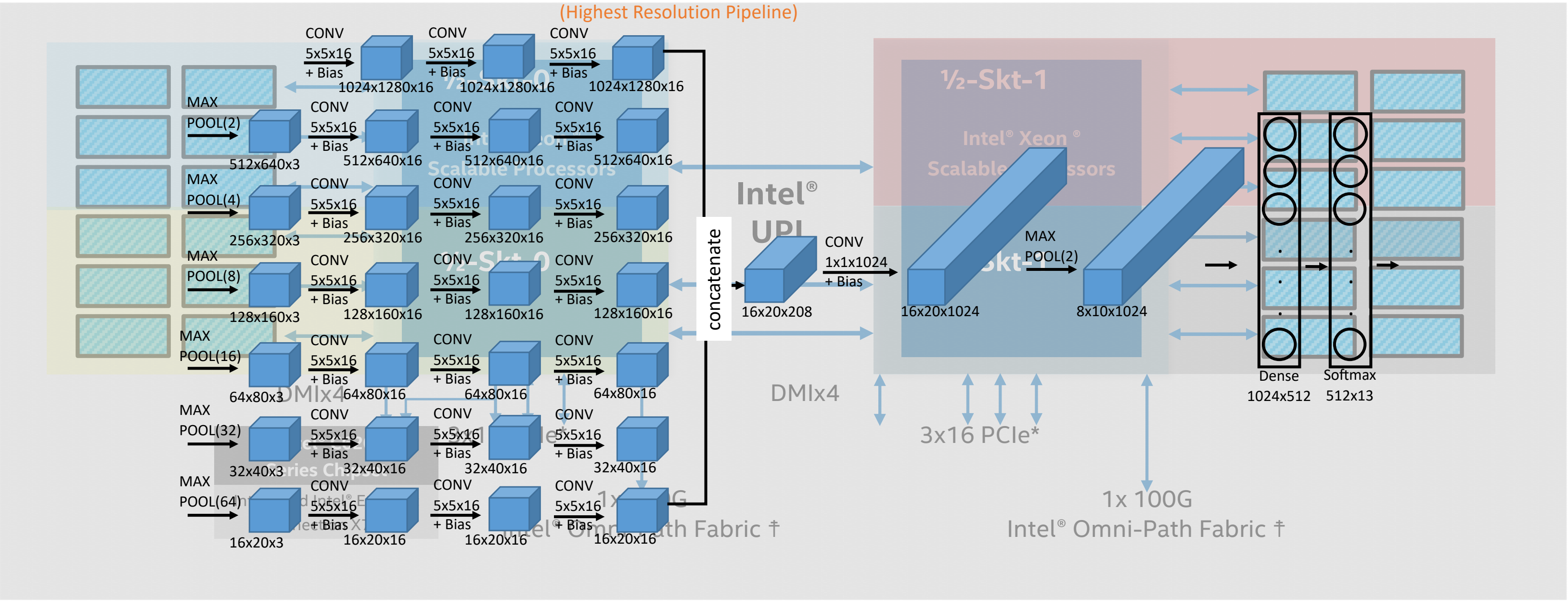
(Highest Resolution Pathway)



(Lowest Resolution Pathway)

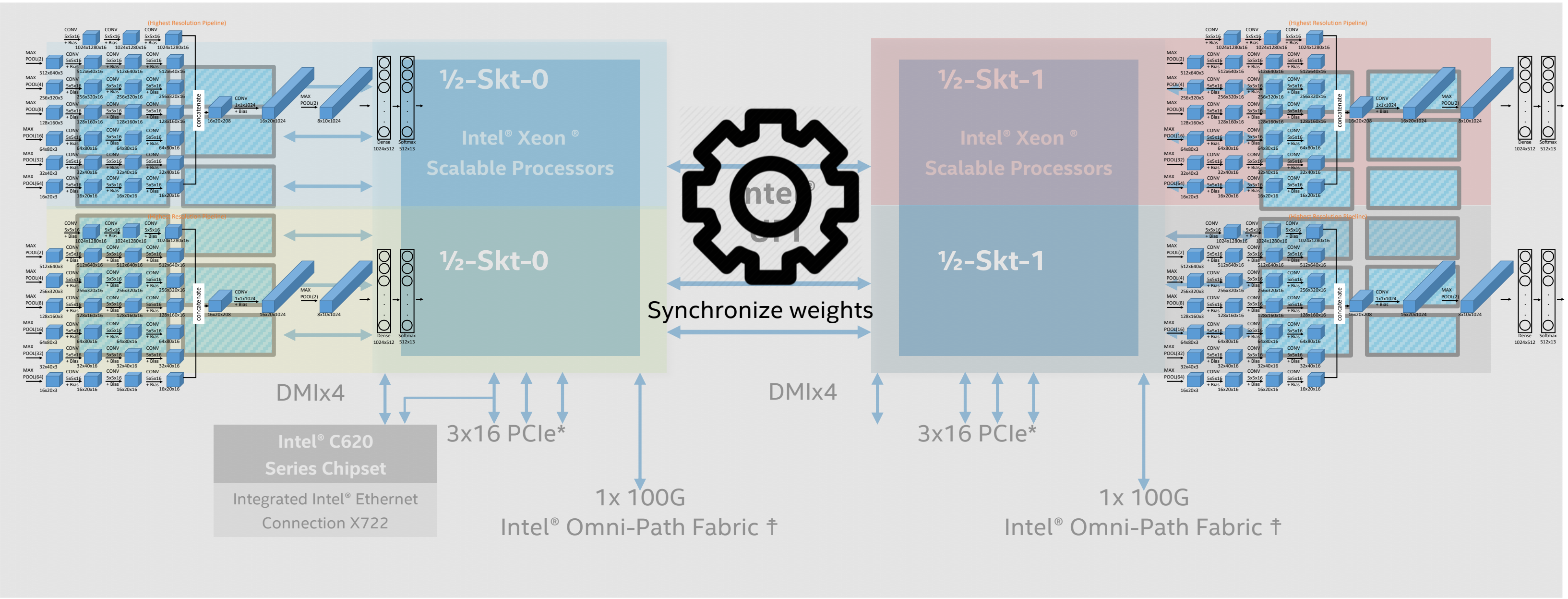
SINGLE WORKER ON INTEL® 2S XEON® 6148 PROCESSOR

Xeon® 2Socket configuration

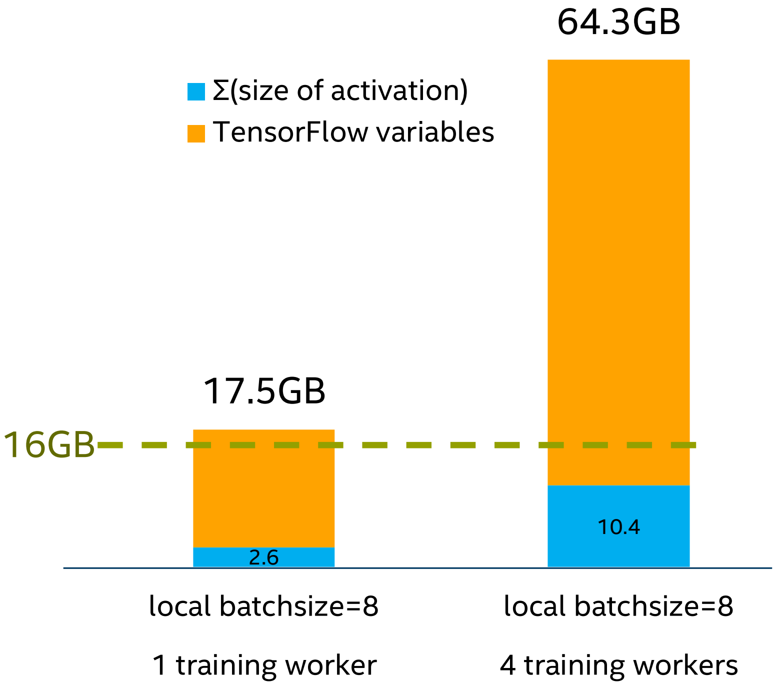


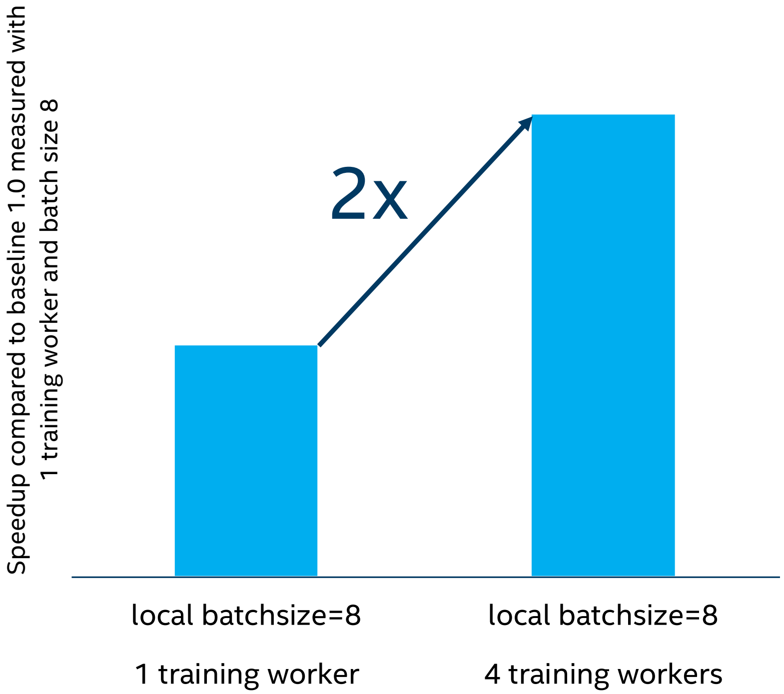
NUMA CONFIGURATION IN TWO SOCKETS

Xeon® 2Socket configuration



PERFORMANCE IMPROVED WITH 4 WORKERS IN A SINGLE INTEL® 2S XEON® 6148 PROCESSOR





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

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=40/10

STRATEGY

- USE LATEST INTEL® XEON® PROCESSOR BASED PLATFORM ✓
- IMPROVE PERFORMANCE WITH TENSORFLOW* ✓
- MAXIMIZE PERFORMANCE ON A SINGLE NODE ✓
- **SCALEOUT PERFORMANCE WITH MULTIPLE NODES**

METHOD TO SCALE TO MULTIPLE NODES

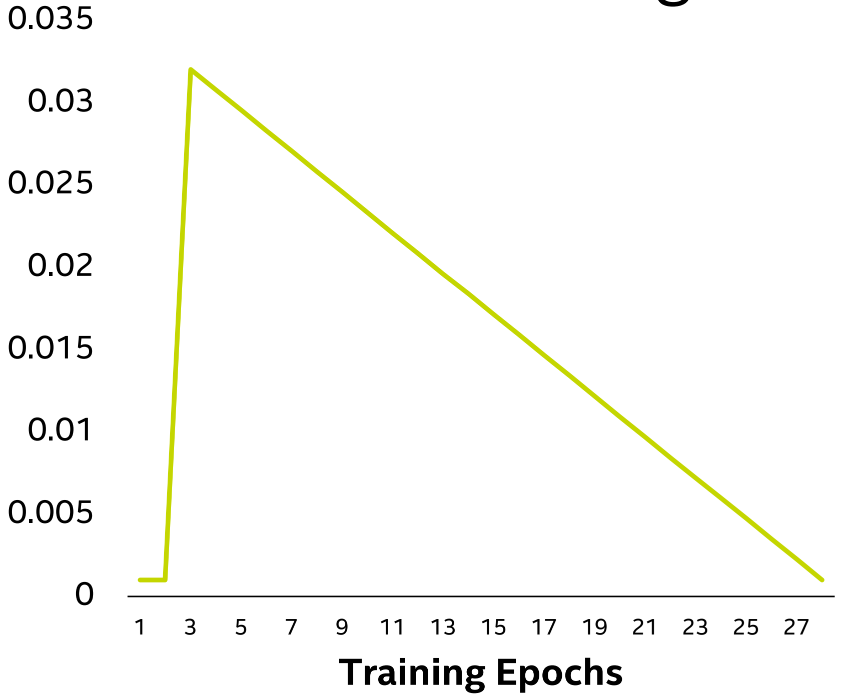
4 TRAINING WORKERS ON 8 NODE INTEL® XEON® CLUSTER

USE HOROVOD AND MPI TO SYNCHRONIZE GRADIENTS

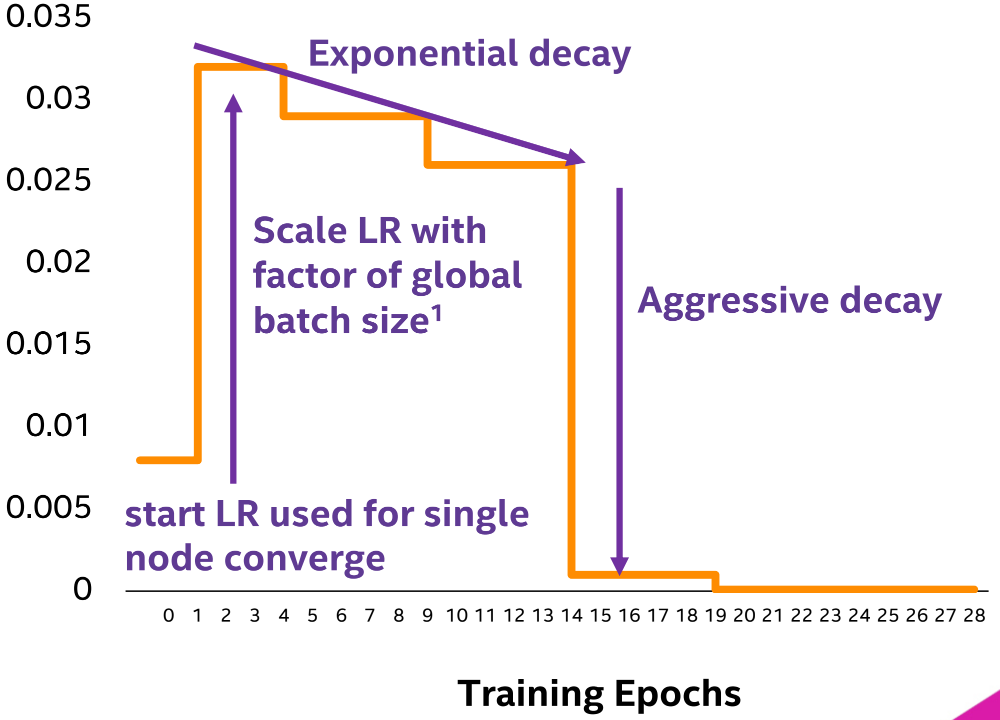
ADJUST LEARNING RATE TO ACHIEVE CONVERGENCE

LEARNING RATE SCHEDULE TO SCALE TO LARGE BATCHES

Established scaling rule[1]



What worked for us



Exponential decay

Scale LR with factor of global batch size[1]

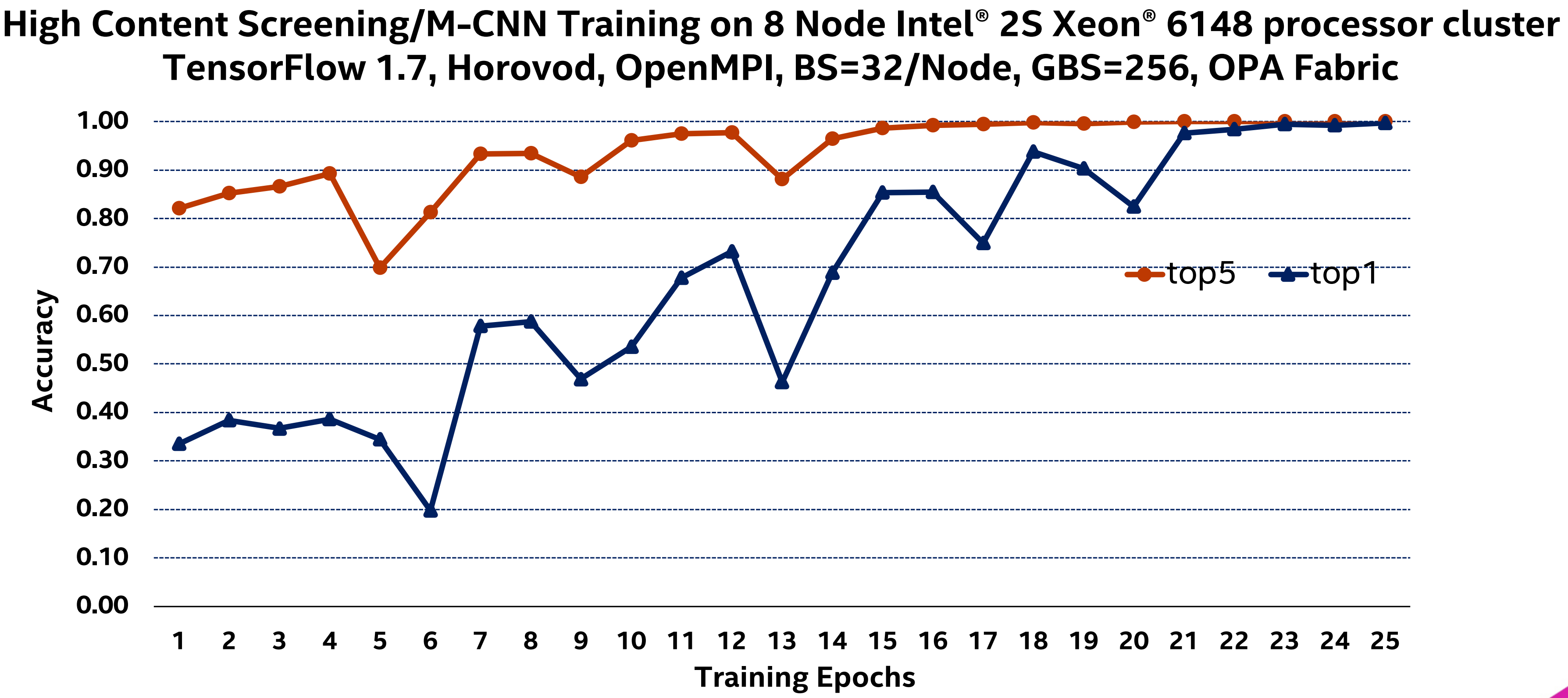
Aggressive decay

start LR used for single node converge

[1] Goyal et, 2017, "Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour"

SUMMING IT ALL UP

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

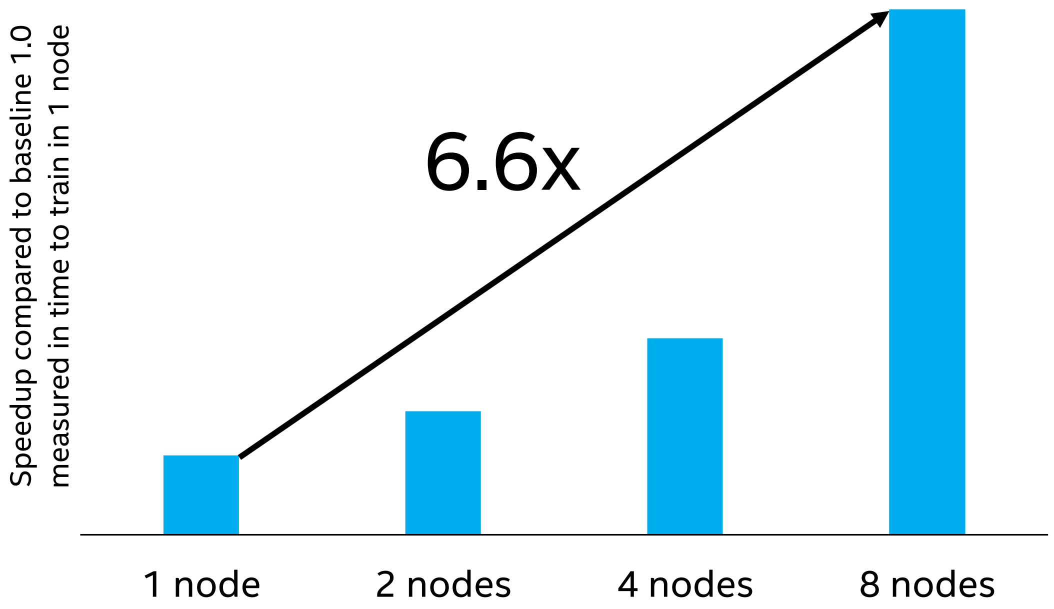


Workload: Image set BBBC021 : Human MCF7 Cells – compound profiling experiment.

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

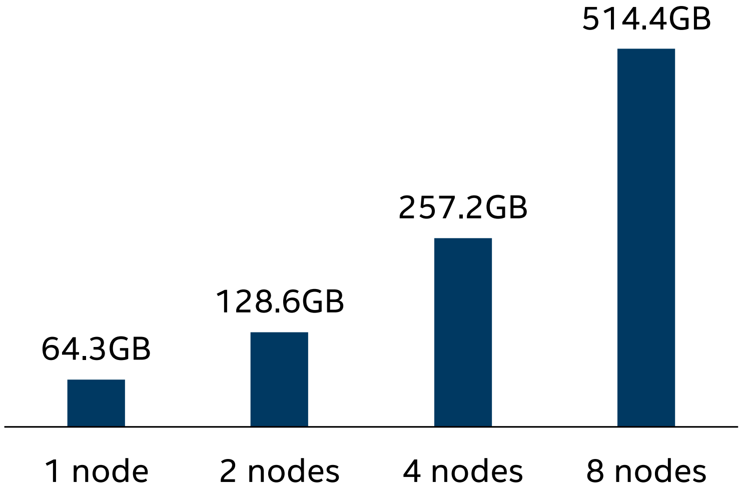
HIGH PERFORMANCE AT SCALE WITH INTEL® XEON® SCALABLE PROCESSOR

Time To Train



Total Memory Used

192GB DDR4 per Intel® Xeon®



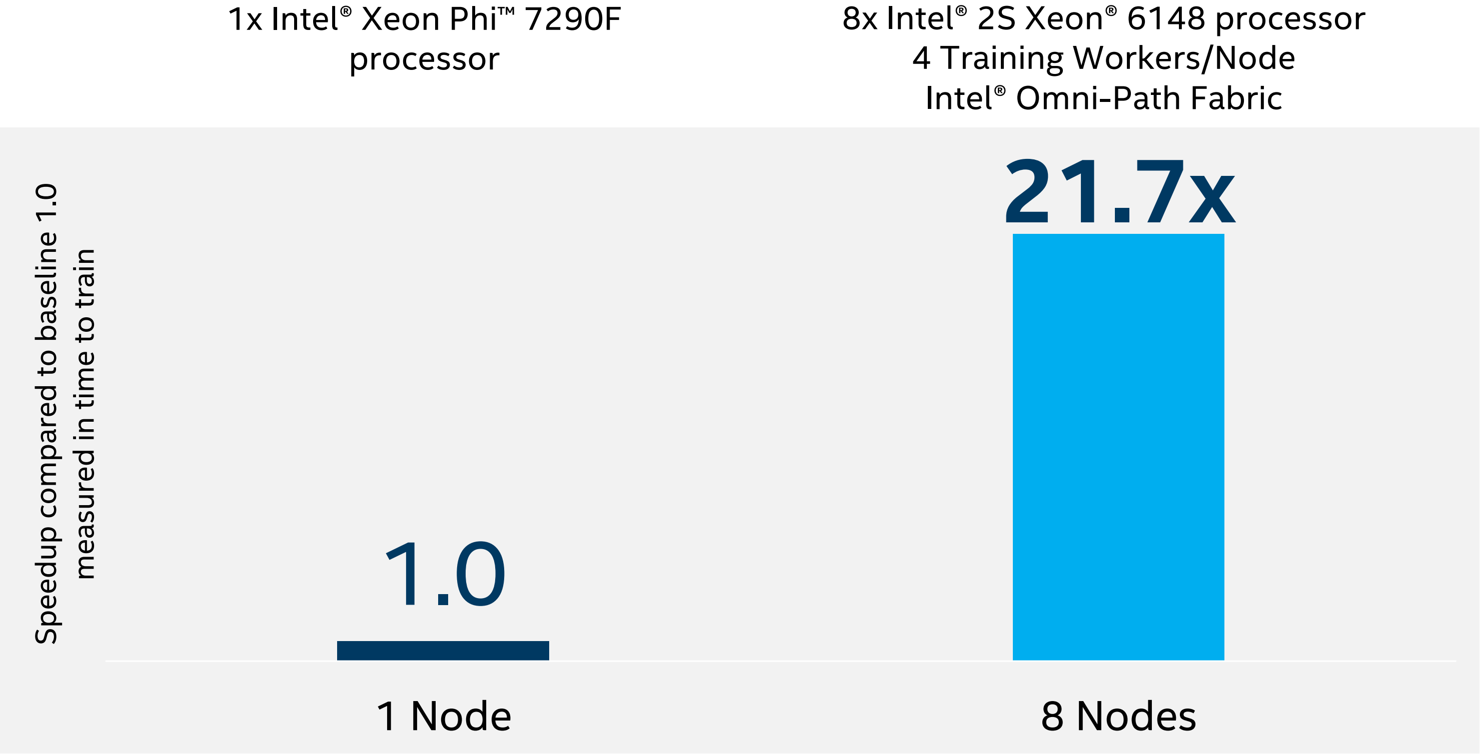
Workload: Image set BBBC021 : Human MCF7 Cells – compound profiling experiment.

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

STRATEGY

- USE LATEST INTEL® XEON® PROCESSOR BASED PLATFORM ✓
- IMPROVE PERFORMANCE WITH TENSORFLOW* ✓
- MAXIMIZE PERFORMANCE ON A SINGLE NODE ✓
- SCALEOUT PERFORMANCE WITH MULTIPLE NODES ✓

OVERALL SPEEDUP ACHIEVED IN TIME TO TRAIN



Workload: Image set BBBC021 : Human MCF7 Cells – compound profiling experiment.
Godinez et al, A multi-scale convolutional neural network for phenotyping high-content cellular images. Bioinformatics, 2017

FUTURE WORK

- Reduce Time to Train of M-CNN by
  - Scaling training to 256 nodes
  - Explore larger datasets with millions of images

- Reduce expensive annotation of HCS images
  - Clustering with Deep Autoencoders

LEGAL NOTICES AND DISCLAIMERS

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.

Normalized performance is calculated by assigning a baseline value of 1.0 to one benchmark result, and then dividing the actual benchmark result for the baseline platform into each of the specific benchmark results of each of the other platforms, and assigning them a relative performance number that correlates with the performance improvements reported.

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.

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.

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.

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.

Intel, Xeon, AVX-512, Omni-Path Architecture, Quick Assist, Optane, Trusted Execution Technology, Key Protection Technology, Platform Trust Technology, Resource Director Technology, the Intel logo and others are trademarks of Intel Corporation and its subsidiaries in the U.S. and/or other countries.

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

© 2018 Intel Corporation.

8 NODE CLUSTER CONFIGURATION DETAILS

Compute Nodes: 2 sockets Intel® Xeon® Gold 6148 CPU with 20 cores each @ 2.4GHz for a total of 40 cores per node. 2 Threads per core. L3 Cache: 27.3MB, 192GB of DDR4, Intel® Omni-Path Host Fabric Interface, dual-rail, Software: OpenMPI library 3.0.0, 100Gbps Intel® Intel® Omni-Path Host Fabric Interface, 480GB Intel® SSD OS drive, 1.6TB Intel® SSD data drive, CentOS* Linux 7.3, Horovod 0.12.1, Python 2.7.5

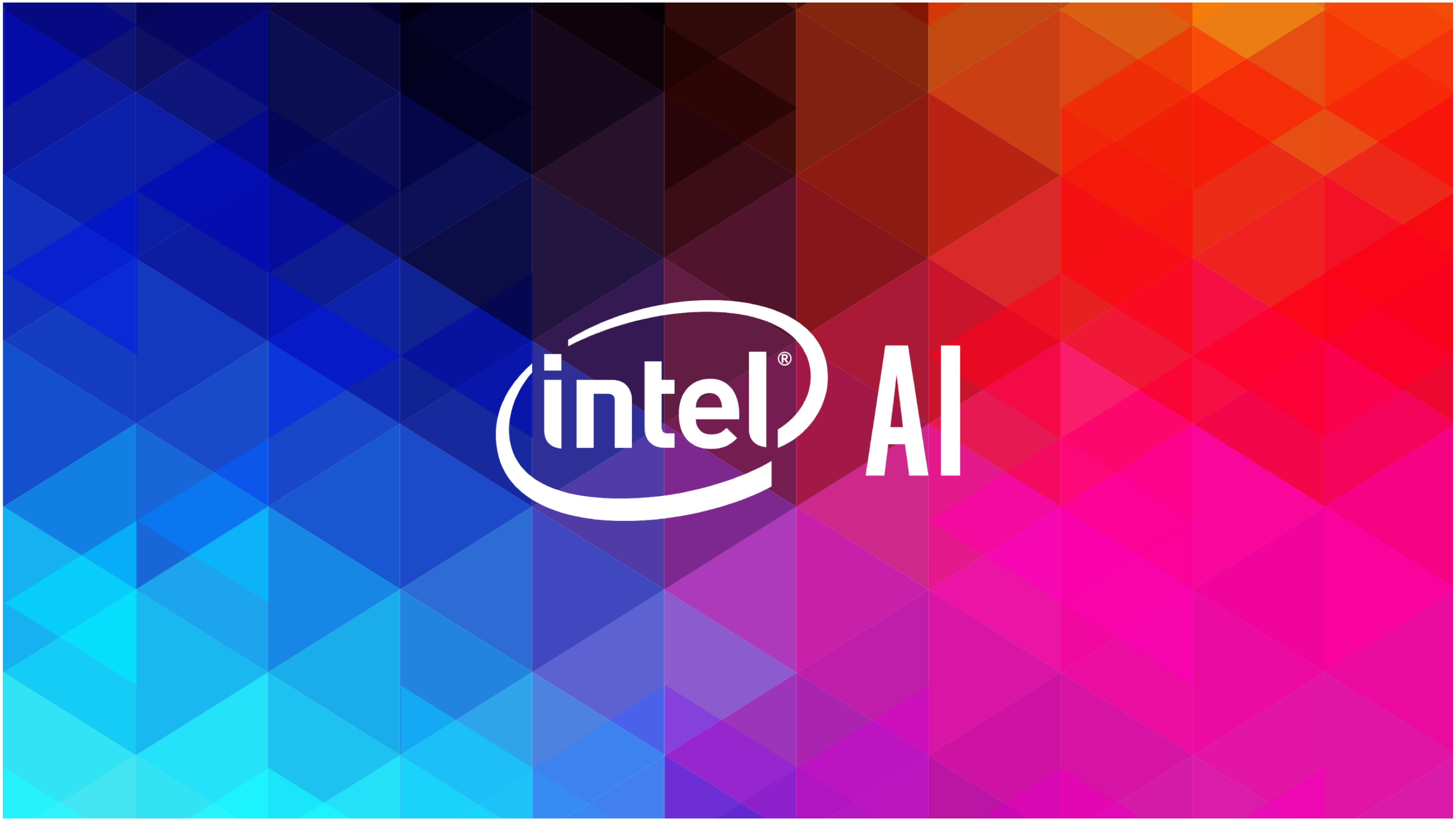
Top of the rack Switch: 48-port Intel® Omni-Path Edge Switch 100 Series

TensorFlow*: Intel® optimized TensorFlow version 1.7.0 https://github.com/tensorflow/tensorflow/tree/v1.7.0

Model: As defined by Godinez et al, A multi-scale convolutional neural network for phenotyping high-content cellular images. Bioinformatics, 2017

Performance Measured With:

```
OMP_NUM_THREADS=10 mpirun -np 32 -cpus-per-proc 10 --map-by socket -hostfile HOSTFILE --report-bindings --oversubscribe -x LD_LIBRARY_PATH -x
PATH -x OMP_NUM_THREADS -x HOROVOD_FUSION_THRESHOLD  numactl -l python tf_cnn_benchmarks.py --model=mcnn --batch_size=8 --
data_format=NCHW --data_dir=INPUT_DATA_DIR --data_name=mcnn --num_intra_threads=10 --num_inter_threads=2 --num_batches=2000 --
num_warmup_batches=70 --display_every=5 --momentum=0.9 --weight_decay=0.00005 --optimizer=momentum --resize_method=bilinear --distortions=False
--sync_on_finish=True --device=cpu --mkl=True --kmp_affinity=="granularity=fine,compact,1,0" --variable_update=horovod --local_parameter_device=cpu --
kmp_blocktime=1 --horovod_device=cpu --piecewise_learning_rate_schedule='0.008;2;0.032;5;0.029;10;0.026;15;0.001;20;0.0001' --
train_dir=TRAIN_DATAWRITE_DIR --save_summaries_steps=1 --summary_verbosity=1
```



intel® AI