Open. Together.

OCP SUMMIT

HPC &
CPU/GPU/FPGA
Technology

Next Generation Intel® Xeon® Scalable Processors for Machine Learning

Andres Rodriguez, Sr. Principal Engineer, Intel

Niveditha Sundaram, Director of Engineering, Intel

Jianhui Li, Principal Engineer, Intel

Shivani Sud, Architect, Intel

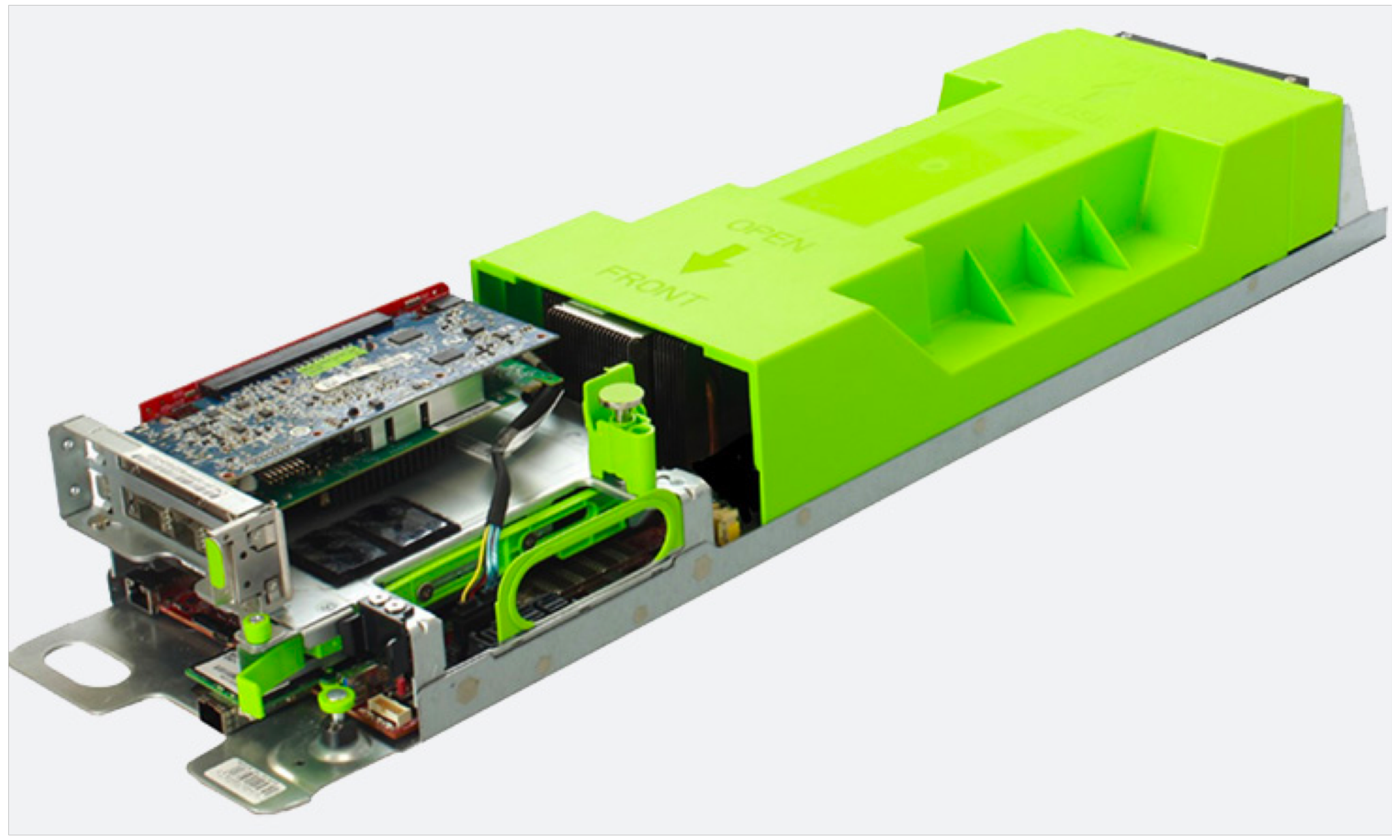
OPEN PLATINUM™

Overview

- SW optimizations
- Lower numerical precision benefits
- Modular system architecture for high density cloud usages
- What's next?

Tioga Pass Intel® Xeon® OCP Platform

ResNet-50 inference images/second per socket					
Batch size	No MKLDNN @FP32	MKLDNN @F32	Gains	MKLDNN @INT8	Gains
1	18.90	101.36	**5.4x**	175.16	**9.3x**
32	21.18	169.49	**8.0x**	331.12	**15.6x**



https://code.fb.com/data-center-engineering/the-end-to-end-refresh-of-our-server-hardware-fleet/

Tested by Intel as of 3/01/2019. 2S Intel® Xeon® Gold 6139 (18 cores), HT ON, turbo ON, Total Memory 128 GB (4 slots/ 32 GB/ 2.30 GHz), BIOS: F08_3A13, Centos 7 Kernel 3.10.0-957.e17.x86_64, Deep Learning Framework: PyTorch w/C2 backend. “No MKLDNN”: https://github.com/pytorch/pytorch.git checkout 4ac91b2d64eeea5ca21083831db5950dc08441d6, wget https://patch-diff.githubusercontent.com/raw/pytorch/pytorch/pull/17464.diff, git apply 17464.diff. “MKLDNN”: PR link: https://github.com/pytorch/pytorch/pull/17464, gcc (Red Hat 5.3.1-6) 5.3.1 20160406, MKLDNN version: v0.17.3 (commit hash: 0c3cb94999919d33e4875177fdef662bd9413dd4), ResNet50: https://github.com/intel/optimized-models/tree/master/pytorch, No datalayer, 1 instance/1 socket, Datatype: INT8 & FP32. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. Additional disclaimers in Slide 21.

Overview of ML requirements

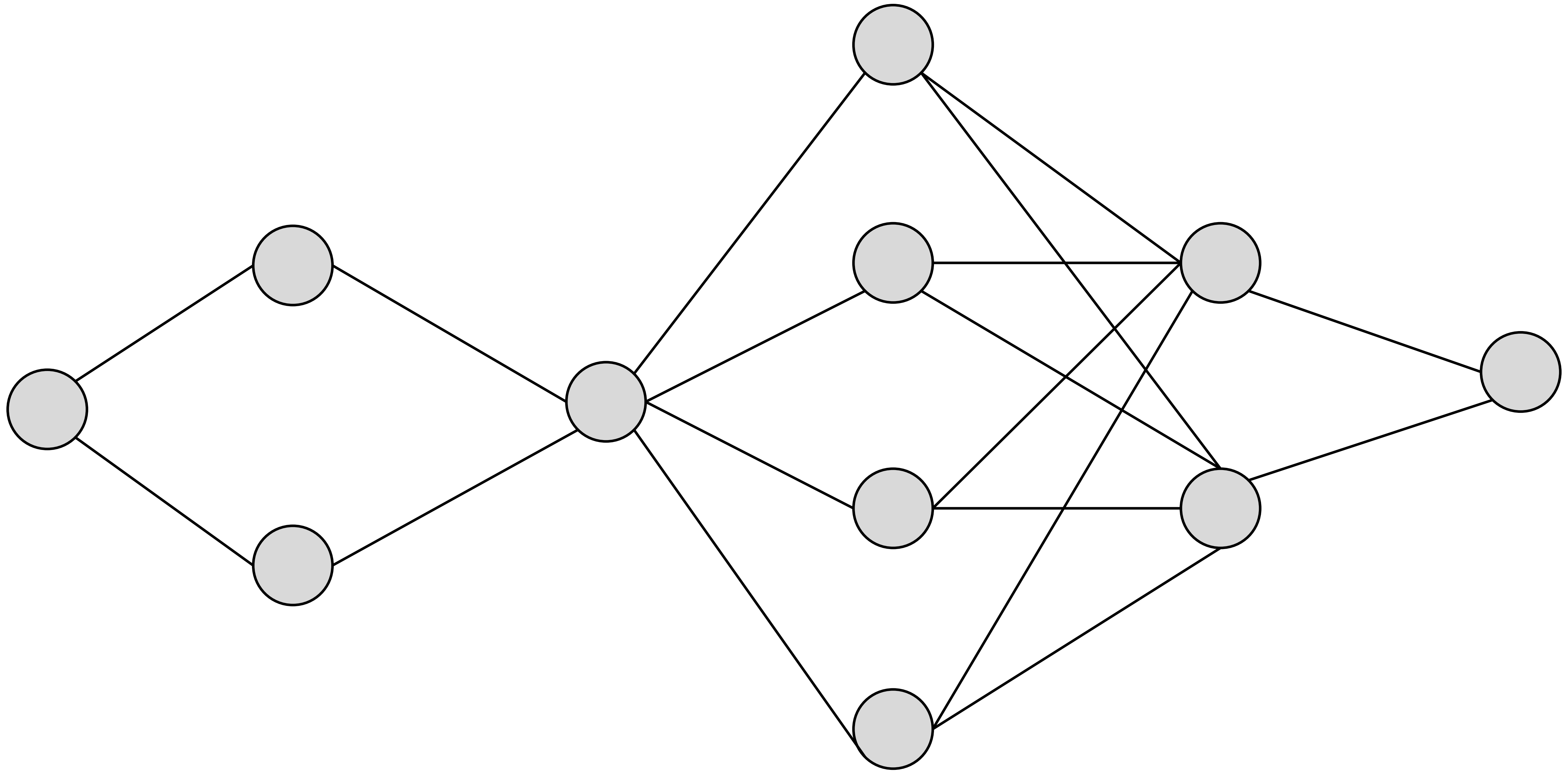
Input

Tensors

2D

3D

Computational graph (i.e. deep learning model)



Output

Tensors

Prob of pedestrian

Operation
(conv, matrix multiply, pooling, ReLU, LSTM, data reorder, …)

Data flow
Dictates how Tensor flows

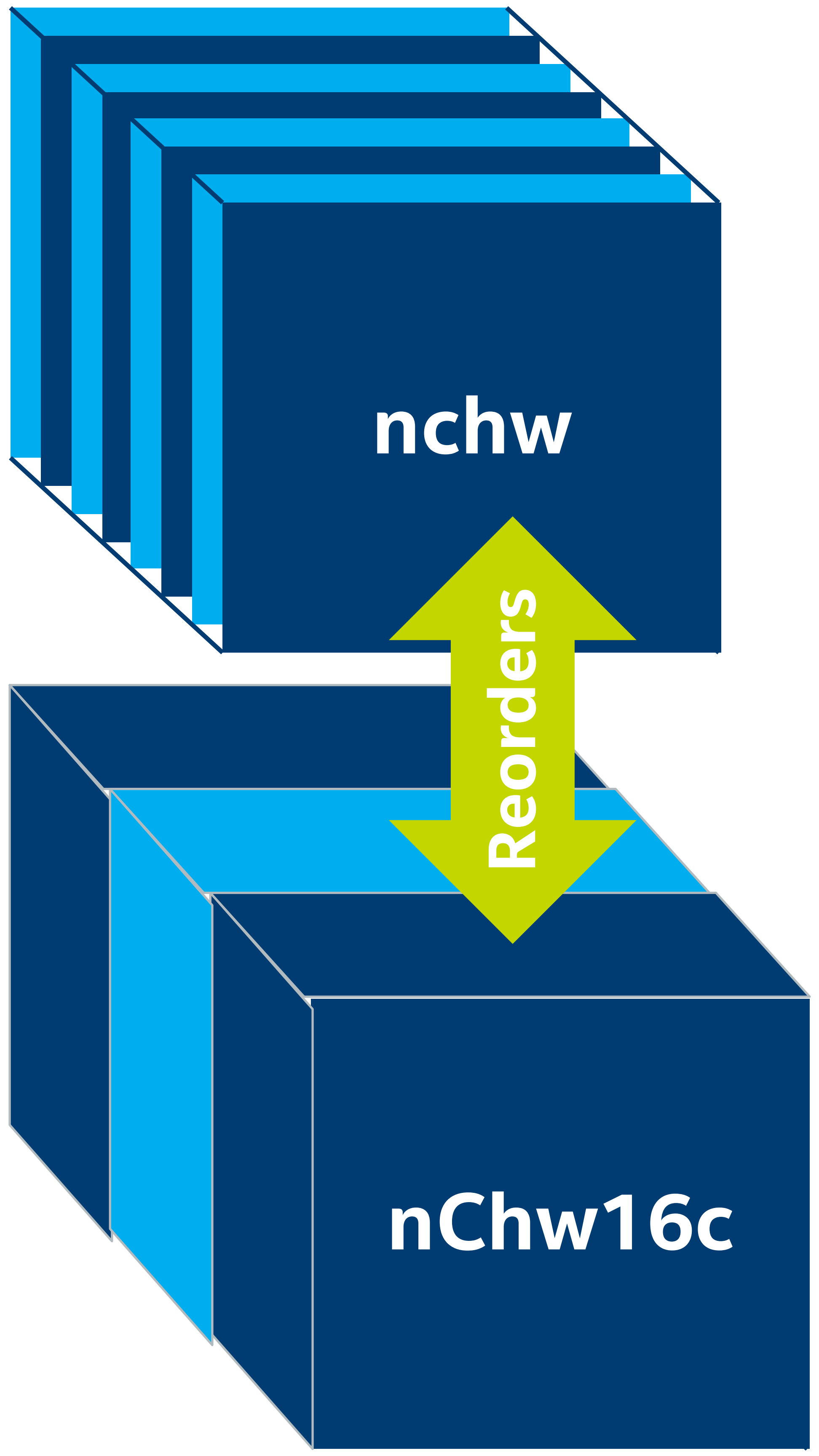
Inference (aka serving): Forward once

Training: Forward and backward many times

Adapted from Harry Kim

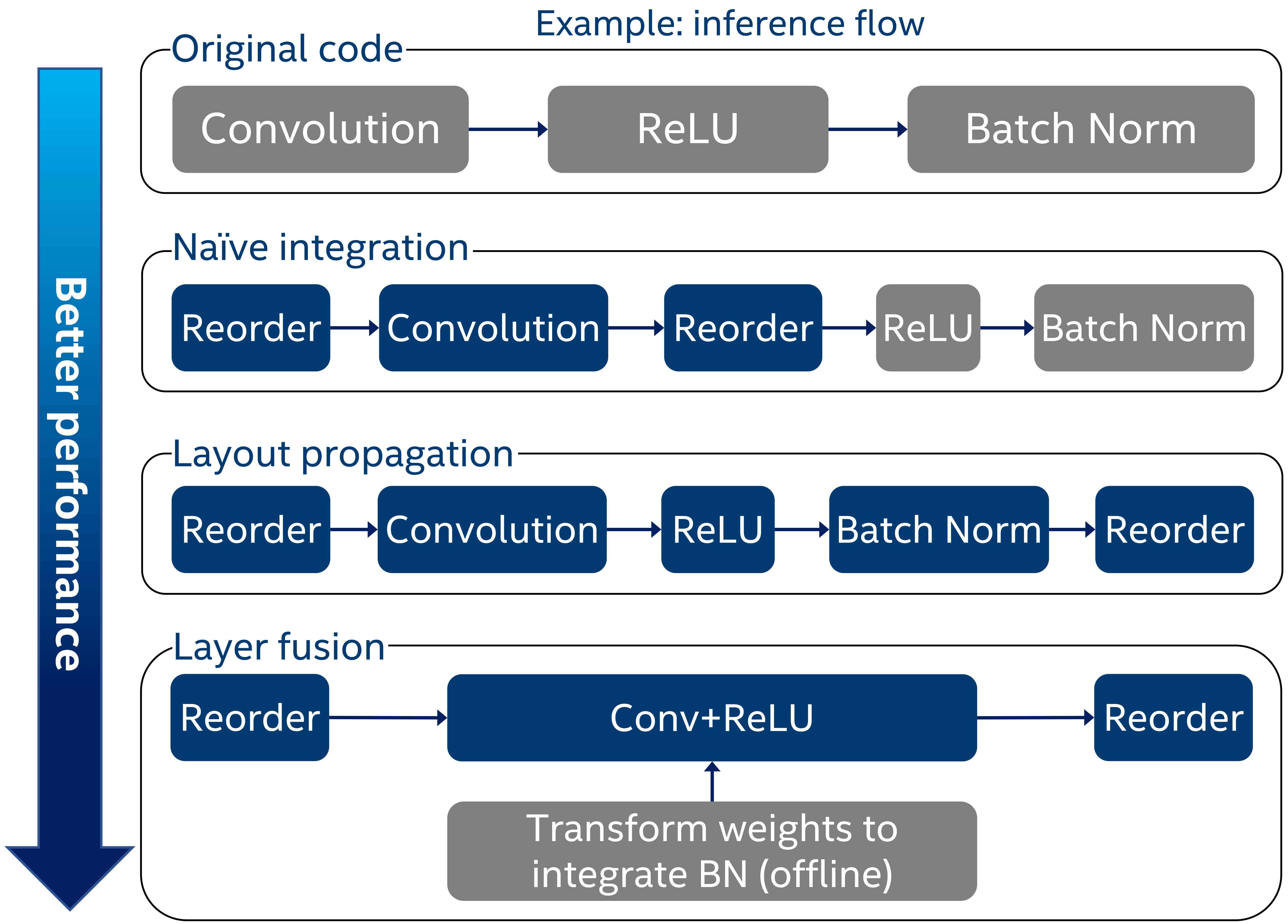
Reorder: Blocking & Mem layouts

- Data reorder (only when necessary) for effective use of CPU caches & registers
- E.g., popular memory layouts for image recognition are **nhwc** and **nchw**
- Vectorization: SIMD
  - **nChw16c:** block feature maps by 16 (bc 16 *fp32* values per AVX512 register)
  - Output feature maps can be computed independently in parallel
- Register blocking: reuse register data & hide FMA latencies
  - Blocking in the spatial domain of the output tensor
- Cache blocking: reuse cache data
- Intel MKLDNN



Adapted from Vadim Pirogov

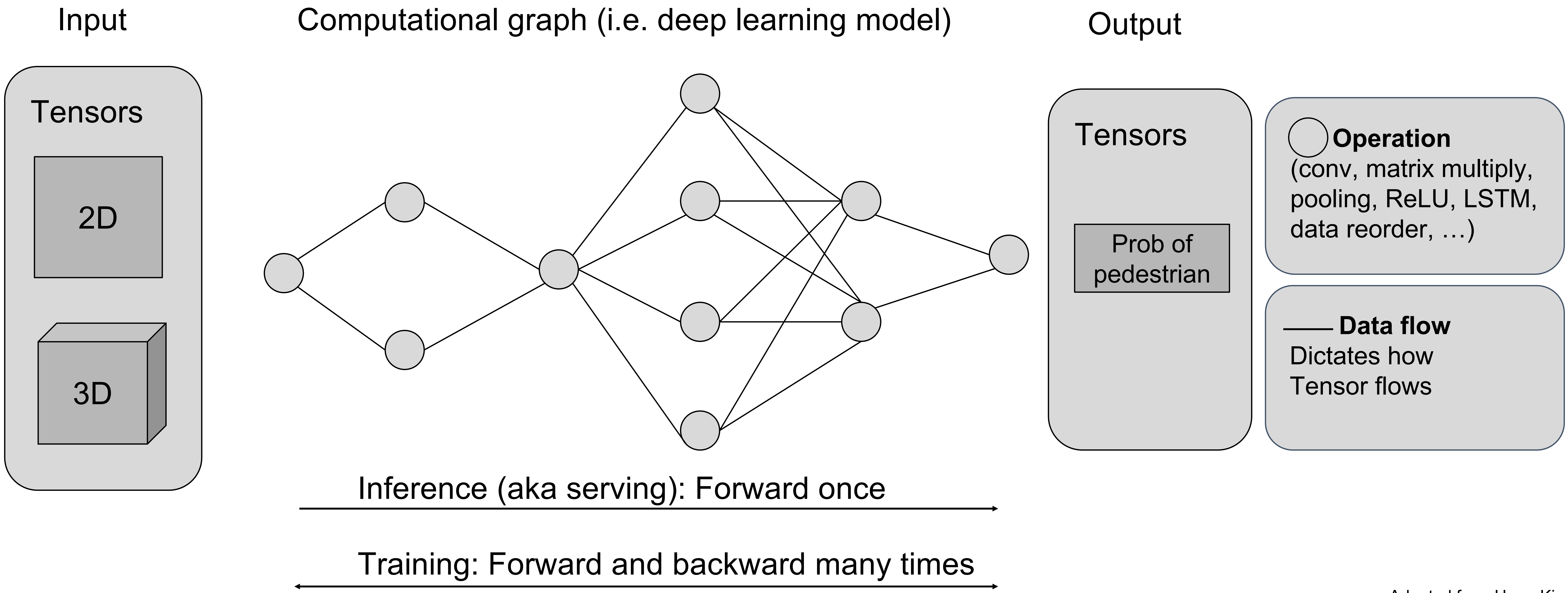
Performance optimizations: node + graph



Graph Nodes / Primitives / Kernels	Class
• (De-)Convolution • Inner Product • RNN, LSTM, GRU	Compute intensive operations
• Pooling AVG/MAX • Batch Normalization • ReLU, Tanh, Softmax • …	Memory bandwidth limited operations
• Reorder • Concatenation	Data movement

Adapted from Vadim Pirogov

Lower precision motivation



Adapted from Harry Kim

Popular numerical precisions

FP32	s	8 bit exp	23 bit mantissa
BF16	s	8 bit exp	7 bit mantissa
FP16	s	5 bit exp	10 bit mantissa
INT16	s	15 bit mantissa	
INT8	s	7 bit mantissa	

- FP32 is usually the default training and inference numerical precision
- BF16 shown to provide virtually the same accuracy for *training* and *inference* as FP32
    - Simulated on various workloads and achieving virtually the same accuracy
    - No hyper-parameters changes compared to FP32 on simulated workloads
- INT8 shown to provide similar accuracy for *inference* as FP32 for various models

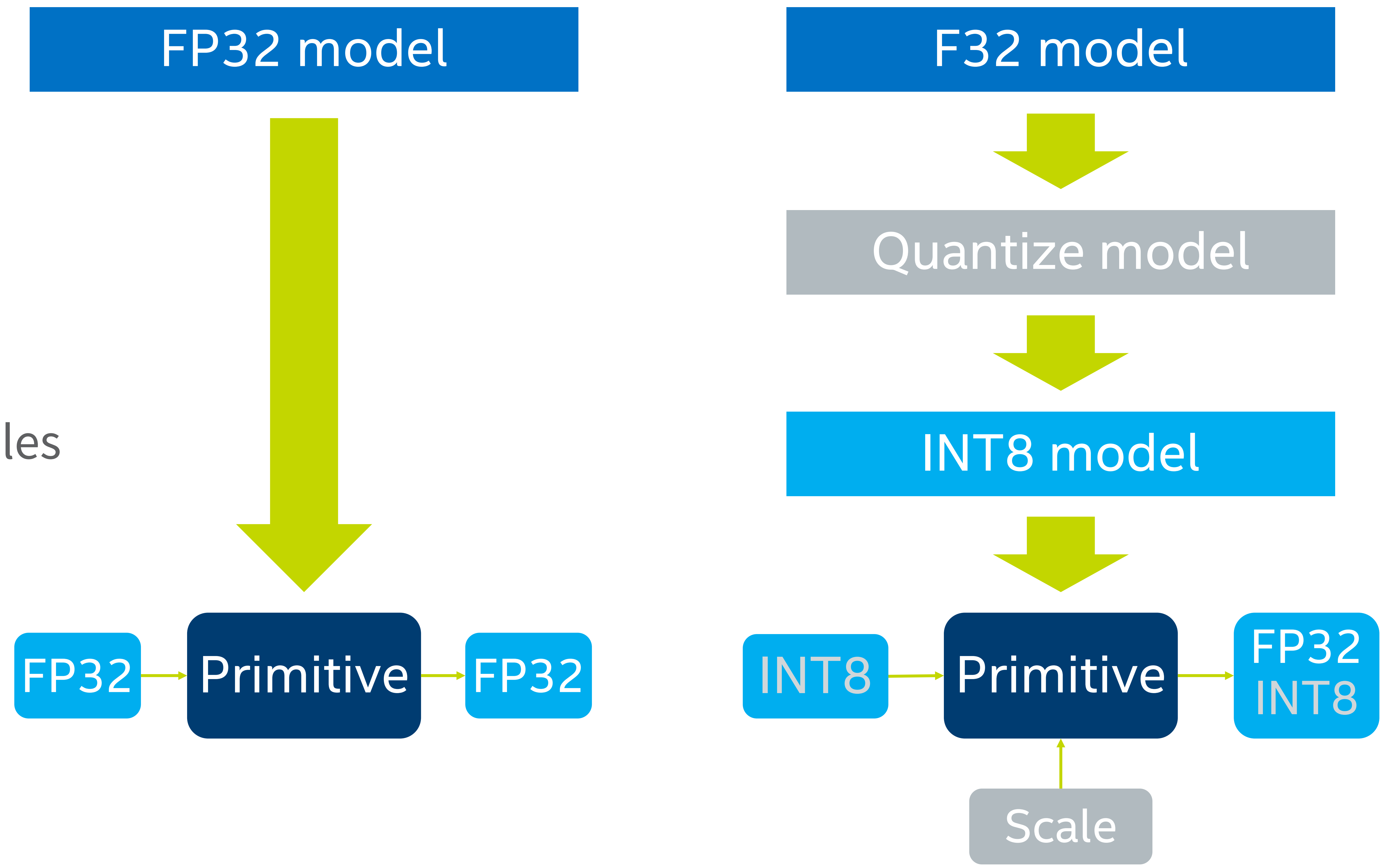
Lower-precision (INT8) inference

One common approach:

- Symmetric quantization (zero shift)
- KL divergence to find a threshold
- Quantize conv & inner product w/channel-wise scales

Offline calibration required to compute scales

Some layers run in higher precision



Adapted from Vadim Pirogov

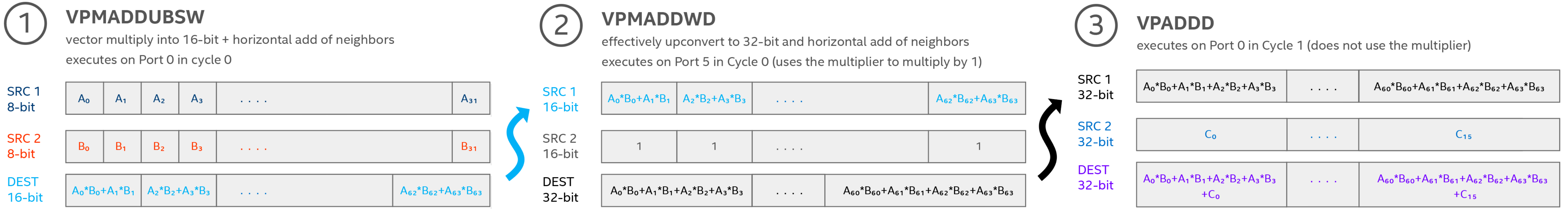
To quantize or not to quantize…

Best known method to determine quantizable layers:

- Quantize entire model
- Compute a metric[(1)] of difference between FP32 and INT8
- While accuracy is not meet:
  - Unquantized layer with worst metric of difference

(1) Examples of metrics: Normalized Root-Mean-Square Deviation (NRMSD); KL Divergence

HW advancements



Intel Xeon Scalable (Skylake)

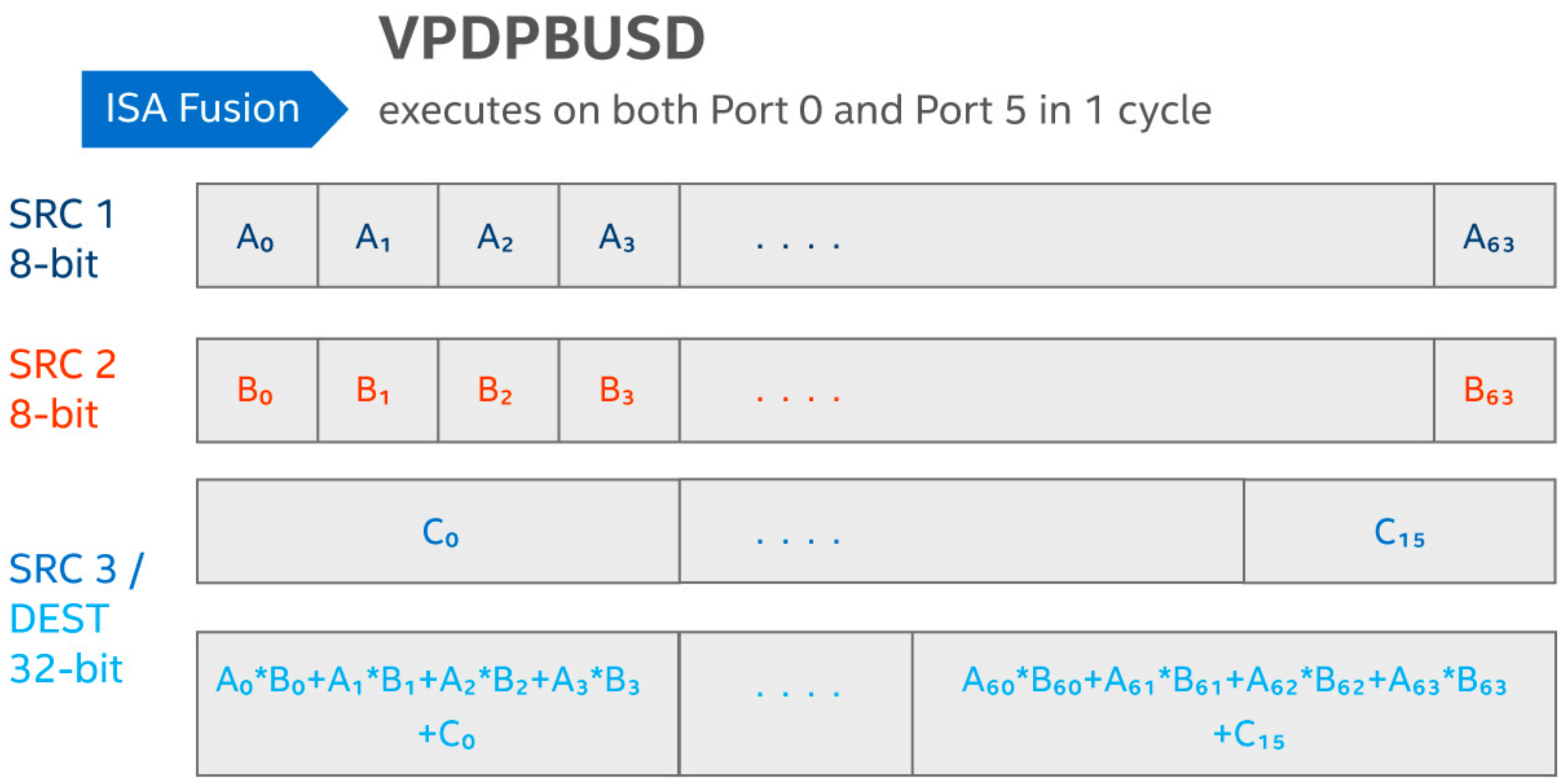
- AVX512
- 2 FMAs/core
- 2x registers
- 2x size of registers
- 1 MB per core L2 cache
- Higher memory bandwidth
- 1.33x INT8 peak throughput vs. fp32

next gen Intel Xeon Scalable (Cascade Lake)

- +AVX512 VNNI instruction
- 4x INT8 peak throughput vs. fp32

future Intel Xeon Scalable (Cooper Lake)

- +New ISA supporting BFLOAT16 [(1)]
- Full ISA specification soon to be published
- 2x bfloat16 peak throughput vs. fp32



(1) https://software.intel.com/en-us/download/bfloat16-hardware-numerics-definition

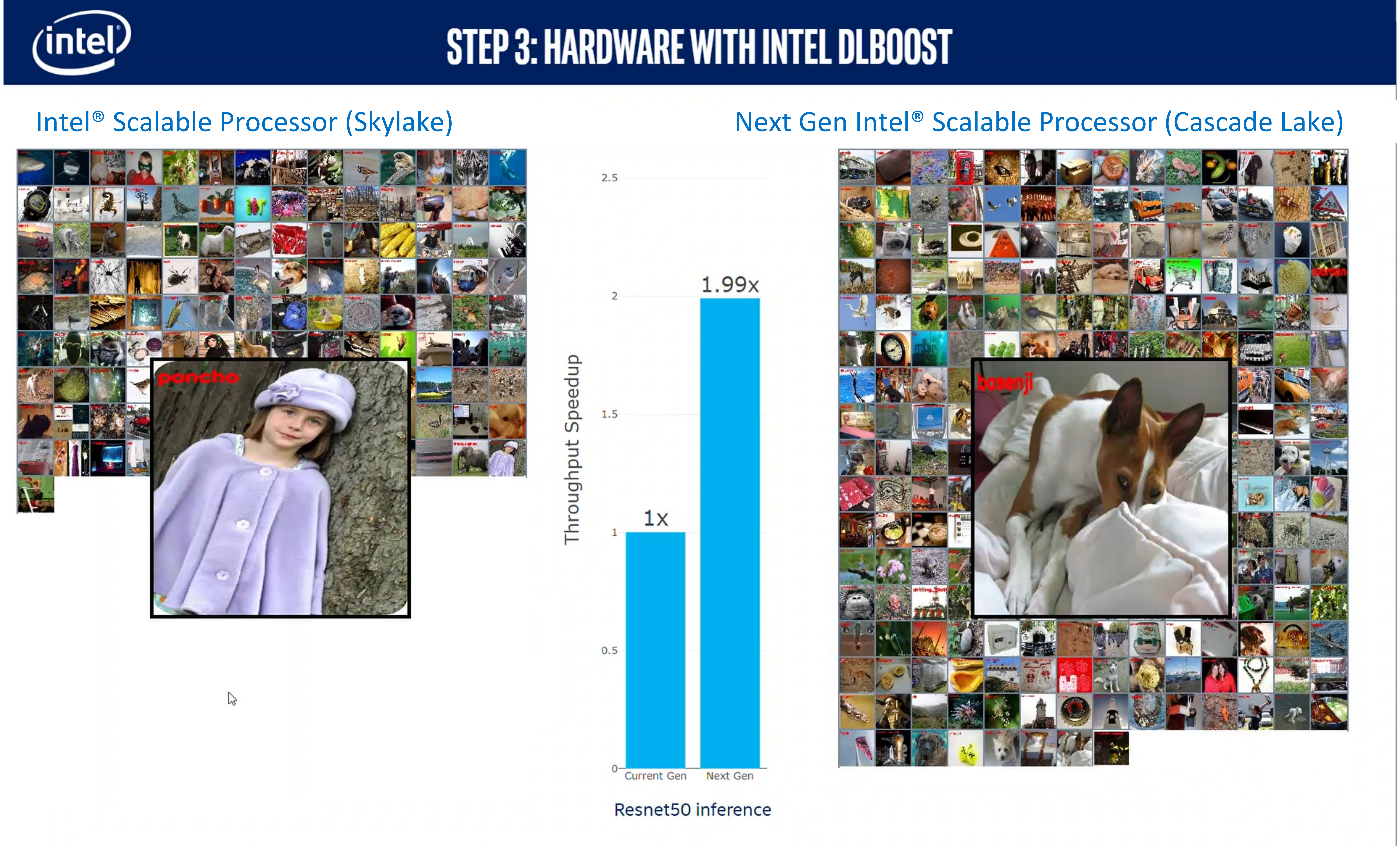
Figures adapted from Israel Hirsh

Performance results with Intel® DL Boost for a range of customer workloads

SIGNIFICANT PERFORMANCE GAINS USING INTEL® DL BOOST ACROSS POPULAR FRAMEWORKS AND DIFFERENT CUSTOMER USECASES

Configuration details in Slides 21-23. Additional disclaimers in Slide 19. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

Demo

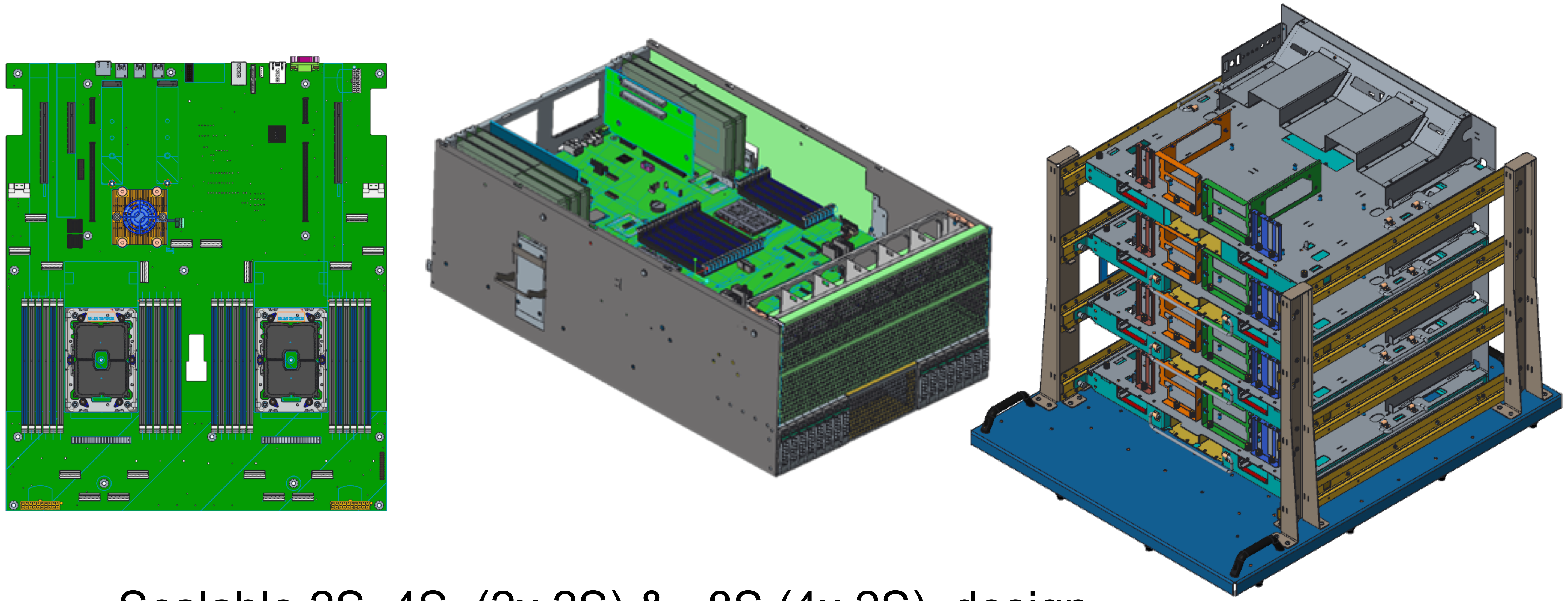


PLAY

2 socket Intel® Xeon® Platinum 8280 Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2933 MHz), BIOS: SE5C620.86B.0D.01.0271.120720180605 (ucode: 0x4000013), Ubuntu 18.04.1 LTS, kernel 4.15.0-45-generic, SSD 1x sda INTEL SSDSC2BA80 SSD 745.2GB, 3X INTEL SSDPE2KX040T7 SSD 3.7TB; 2 socket Intel® Xeon® Platinum 8180 Processor, 28 cores HT On Turbo ON Total Memory 192 GB (12 slots/ 16GB/ 2666 MHz), BIOS: SE5C620.86B.00.01.0015.110720180833 (ucode: 0x200004d), CentOS 7.5, 3.10.0-693.el7.x86_64, Intel® SSD DC S4500 SERIES SSDSC2KB480G7 2.5'' 6Gb/s SATA SSD 480G. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. Additional disclaimers in Slide 19.

Scalable modular system architecture - Enabling high density cloud usages

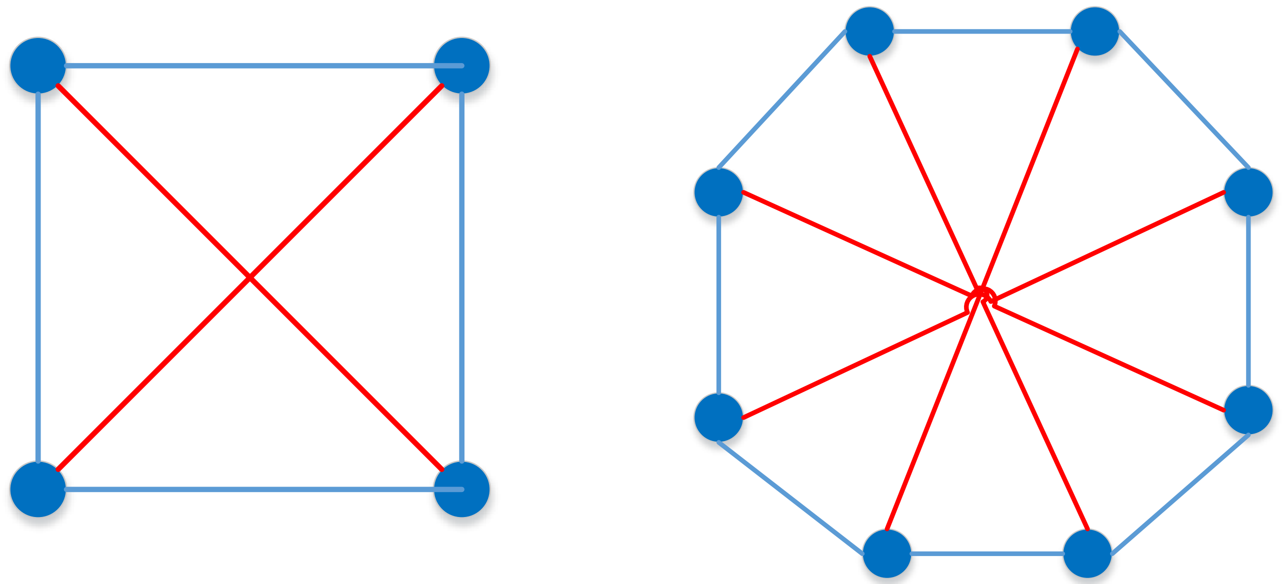
- Glue-less Scalable 2S modular architecture
- Balanced core/memory architecture – UMA/NUMA scaling
- Distributed & unified server management – OpenBMC support
- Advanced RAS capabilities for fault tolerant design
- Use cases across multiple segments (IMDB, cloud IaaS, etc.)



Scalable 2S, 4S (2x 2S) & 8S (4x 2S) design

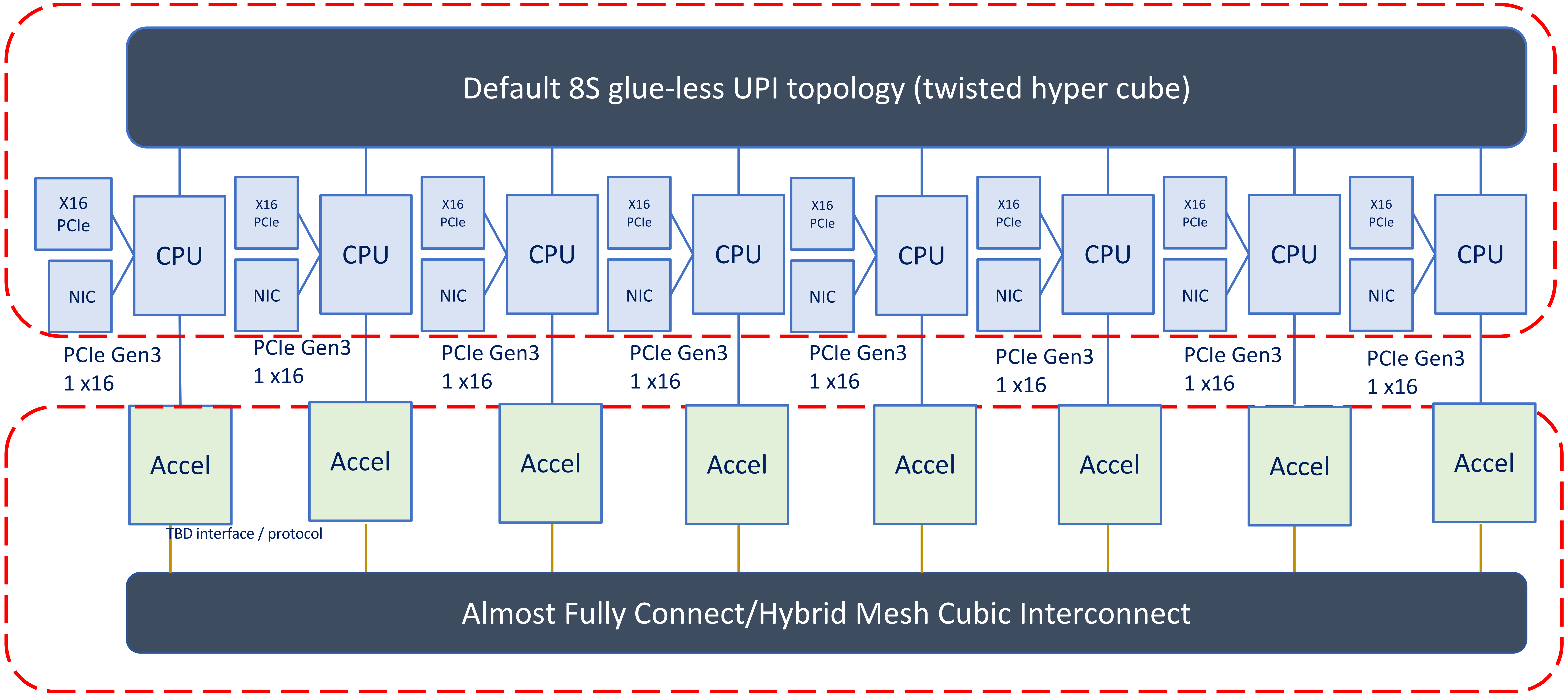
Suitable for large ML workloads

- Large memory footprint applications
- Large compute-bound applications
- Lower TCO



4S : Fully connected Mesh topology; 8S: Pin Wheel topology

Large Memory Unified Training OCP platform: 8S Intel Xeon architecture – glueless pinwheel topology



facebook

intel

What's Next?

- Intel® Nervana® NNP-L (Spring Crest) mezz card will be compliant with the OCP Accelerator Module (in production in 2019)
  - Large HBM memory and local SRAM closer to compute
  - High-speed on- and off-chip interconnects
- Intel® Nervana® NNP-i (Spring Hill) (in production in 2019)
  - Built on Intel 10nm process technology and includes Ice Lake cores
  - Facebook has been a close collaborator on the Intel Nervana NNP-i 1000
- Intel is a proud partner of the Glow community facebook GLOW
- bfloat16 native support in Cooper Lake, NNP-L, FPGA and other future products
  - simulations show virtually the same accuracy with *bloat16* as *fp32* across various training workloads including: ResNet-50, Deep Speech 2, Google GNMT, DC-GAN, etc.

Call to Action

Contribute to OCP 8-Socket reference platform

Use popular frameworks with Intel MKL-DNN

More information:

https://software.intel.com/en-us/articles/lower-numerical-precision-deep-learning-inference-and-training

Notices and Disclaimers

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.

Performance results are based on testing as of 7/11/2017(1x) and may not reflect all publically available security updates. No product can be absolutely. See configuration disclosure for details. 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. 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: www.intel.com/performance. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

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.

The cost reduction scenarios described are intended to enable you to get a better understanding of how the purchase of a given Intel based product, combined with a number of situation-specific variables, might affect future costs and savings. Circumstances will vary and there may be unaccounted-for costs related to the use and deployment of a given product. Nothing in this document should be interpreted as either a promise of or contract for a given level of costs or cost reduction.

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.

No product, component, or computer system can be absolutely secure.

Intel® Advanced Vector Extensions (Intel® AVX)* provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at http://www.intel.com/go/turbo.

Intel processors of the same SKU may vary in frequency or power as a result of natural variability in the production process.

INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS". NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO THIS INFORMATION INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Open. Together.

OCP Global Summit | March 14–15, 2019

OCP SUMMIT

4.0x performance boost with TensorFlow ResNet101: Tested by Intel as of 3/26/2019. 2 socket Intel® Xeon® Platinum 8280L Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2933 MHz), BIOS: SE5C620.86B.0D.01.0348.011820191451 (ucode:0x5000017), CentOS 7.6, Kernel 4.19.5-1.el7.elrepo.x86_64, SSD 1x INTEL SSDSC2KG96 960GB, Deep Learning Framework: TensorFlow: https://hub.docker.com/r/intelaipg/intel-optimized-tensorflow:PR25765-devel-mkl (https://github.com/tensorflow/tensorflow.git commit: 6f2eaa3b99c241a9c09c345e1029513bc4cd470a + Pull Request PR 25765, PR submitted for upstreaming), Compiler: gcc 6.3.0,MKL DNN version: v0.17, ResNet101 : https://github.com/IntelAI/models/tree/master/models/image_recognition/tensorflow/resnet101 commit: 87261e70a902513f934413f009364c4f2eed6642,Synthetic data, Batch Size=128, 2 instance/2 socket, Datatype: INT8 vs Tested by Intel as of 3/26/2019. 2 socket Intel® Xeon® Platinum 8180 Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2666 MHz), BIOS: SE5C620.86B.0D.01.0286.121520181757 (ucode:0x2000057), CentOS 7.6, Kernel 4.19.5-1.el7.elrepo.x86_64, SSD 1x INTEL SSDSC2KG96 960GB, Deep Learning Framework: TensorFlow: https://hub.docker.com/r/intelaipg/intel-optimized-tensorflow:PR25765-devel-mkl (https://github.com/tensorflow/tensorflow.git commit: 6f2eaa3b99c241a9c09c345e1029513bc4cd470a + Pull Request PR 25765, PR submitted for upstreaming), Compiler: gcc 6.3.0,MKL DNN version: v0.17, ResNet101 : https://github.com/IntelAI/models/tree/master/models/image_recognition/tensorflow/resnet101 commit: 87261e70a902513f934413f009364c4f2eed6642,Synthetic data, Batch Size=128, 2 instance/2 socket, Datatype: FP32

3.9x performance boost with TensorFlow ResNet50: Tested by Intel as of 3/26/2019. 2 socket Intel® Xeon® Platinum 8280L Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2933 MHz), BIOS: SE5C620.86B.0D.01.0348.011820191451 (ucode:0x5000017), CentOS 7.6, Kernel 4.19.5-1.el7.elrepo.x86_64, SSD 1x INTEL SSDSC2KG96 960GB, Deep Learning Framework: TensorFlow: https://hub.docker.com/r/intelaipg/intel-optimized-tensorflow:PR25765-devel-mkl (https://github.com/tensorflow/tensorflow.git commit: 6f2eaa3b99c241a9c09c345e1029513bc4cd470a + Pull Request PR 25765, PR submitted for upstreaming), Compiler: gcc 6.3.0,MKL DNN version: v0.17, ResNet50 : https://github.com/IntelAI/models/tree/master/models/image_recognition/tensorflow/resnet50 commit: 87261e70a902513f934413f009364c4f2eed6642, Synthetic data, Batch Size=128, 2 instance/2 socket, Datatype: INT8 vs Tested by Intel as of 3/26/2019. 2 socket Intel® Xeon® Platinum 8180 Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2666 MHz), BIOS: SE5C620.86B.0D.01.0286.121520181757 (ucode:0x2000057), CentOS 7.6, Kernel 4.19.5-1.el7.elrepo.x86_64, SSD 1x INTEL SSDSC2KG96 960GB, Deep Learning Framework: TensorFlow: https://hub.docker.com/r/intelaipg/intel-optimized-tensorflow:PR25765-devel-mkl (https://github.com/tensorflow/tensorflow.git commit: 6f2eaa3b99c241a9c09c345e1029513bc4cd470a + Pull Request PR 25765, PR submitted for upstreaming), Compiler: gcc 6.3.0,MKL DNN version: v0.17, ResNet50 : https://github.com/IntelAI/models/tree/master/models/image_recognition/tensorflow/resnet50 commit: 87261e70a902513f934413f009364c4f2eed6642, Synthetic data, Batch Size=128, 2 instance/2 socket, Datatype: FP32

3.9x performance boost with OpenVino™ ResNet-50: Tested by Intel as of 1/30/2019. 2 socket Intel® Xeon® Platinum 8280 Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2933 MHz), BIOS: SE5C620.86B.0D.01.0271.120720180605 (ucode:0x4000013), Linux-4.15.0-43-generic-x86_64-with-debian-buster-sid, Compiler: gcc (Ubuntu 7.3.0-27ubuntu1~18.04) 7.3.0, Deep Learning ToolKit: OpenVINO R5 (DLDTK Version:1.0.19154 , AIXPRT CP (Community Preview) benchmark (https://www.principledtechnologies.com/benchmarkxprt/aixprt/) BS=64, Imagenet images, 1 instance/2 socket, Datatype: INT8 vs Tested by Intel as of 1/30/2019. 2 socket Intel® Xeon® Platinum 8180 Processor, 28 cores HT On Turbo ON Total Memory 192 GB (12 slots/ 16GB/ 2633 MHz), BIOS: SE5C620.86B.0D.01.0271.120720180605, Linux-4.15.0-29-generic-x86_64-with-Ubuntu-18.04-bionic, Compiler: gcc (Ubuntu 7.3.0-27ubuntu1~18.04) 7.3.0, Deep Learning ToolKit: OpenVINO R5 (DLDTK Version:1.0.19154), AIXPRT CP (Community Preview) benchmark (https://www.principledtechnologies.com/benchmarkxprt/aixprt/) BS=64, Imagenet images, 1 instance/2 socket, Datatype: FP32

3.8x performance boost with MXNet ResNet101: Tested by Intel as of 3/26/2019. 2 socket Intel® Xeon® Platinum 8280L Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2933 MHz), BIOS: SE5C620.86B.0D.01.0348.011820191451 (ucode:0x5000017), CentOS 7.6, Kernel 4.19.5-1.el7.elrepo.x86_64, SSD 1x INTEL SSDSC2KG96 960GB, Deep Learning Framework: MXNet https://github.com/apache/incubator-mxnet.git -b master da5242b732de39ad47d8ecee582f261ba5935fa9, Compiler: gcc 6.3.1, MKL DNN version: v0.17, ResNet101: https://github.com/apache/incubator-MXNet/blob/master/python/MXNet/gluon/model_zoo/vision/resnet.py, Synthetic Data, Batch Size=64, 2 instance/2 socket, Datatype: INT8 vs Tested by Intel as of 3/26/2019. 2 socket Intel® Xeon® Platinum 8180 Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2666 MHz), BIOS: SE5C620.86B.0D.01.0286.121520181757 (ucode:0x2000057), CentOS 7.6, Kernel 4.19.5-1.el7.elrepo.x86_64, SSD 1x INTEL SSDSC2KG96 960GB, Deep Learning Framework: MXNet https://github.com/apache/incubator-mxnet.git -b master da5242b732de39ad47d8ecee582f261ba5935fa9, Compiler: gcc 6.3.1, MKL DNN version: v0.17, ResNet101: https://github.com/apache/incubator-MXNet/blob/master/python/MXNet/gluon/model_zoo/vision/resnet.py, Synthetic Data, Batch Size=64, 2 instance/2 socket, Datatype: FP32

3.7x performance boost with PyTorch ResNet50: Tested by Intel as of 2/25/2019. 2 socket Intel® Xeon® Platinum 8280 Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2933 MHz), BIOS: SE5C620.86B.0D.01.0271.120720180605 (ucode: 0x4000013), Ubuntu 18.04.1 LTS, kernel 4.15.0-45-generic, SSD 1x sda INTEL SSDSC2BA80 SSD 745.2GB, 3X INTEL SSDPE2KX040T7 SSD 3.7TB, Deep Learning Framework: Pytorch with ONNX/Caffe2 backend: https://github.com/pytorch/pytorch.git (commit: 4ac91b2d64eeea5ca21083831db5950dc08441d6) and Pull Request link: https://github.com/pytorch/pytorch/pull/17464 (submitted for upstreaming), gcc (Ubuntu 7.3.0-27ubuntu1~18.04) 7.3.0, MKL DNN version: v0.17.3 (commit hash: 0c3cb94999919d33e4875177fdef662bd9413dd4), ResNet-50: https://github.com/intel/optimized-models/tree/master/pytorch, Synthetic Data, Batch Size=512, 2 instance/2 socket, Datatype: INT8 vs Tested by Intel as of 2/25/2019. 2 socket Intel® Xeon® Platinum 8180 Processor, 28 cores HT On Turbo ON Total Memory 192 GB (12 slots/ 16GB/ 2666 MHz), BIOS: SE5C620.86B.00.01.0015.110720180833 (ucode: 0x200004d), CentOS 7.5, 3.10.0-693.el7.x86_64, Intel® SSD DC S4500 SERIES SSDSC2KB480G7 2.5'' 6Gb/s SATA SSD 480GB, Deep Learning Framework: Pytorch with ONNX/Caffe2 backend: https://github.com/pytorch/pytorch.git (commit: 4ac91b2d64eeea5ca21083831db5950dc08441d6) and Pull Request link: https://github.com/pytorch/pytorch/pull/17464 (submitted for upstreaming), gcc (Ubuntu 7.3.0-27ubuntu1~18.04) 7.3.0, MKL DNN version: v0.17.3 (commit hash: 0c3cb94999919d33e4875177fdef662bd9413dd4), ResNet-50: https://github.com/intel/optimized-models/tree/master/pytorch, Synthetic Data, Batch Size=512, 2 instance/2 socket, Datatype: FP32

3.0x performance boost with MXNet ResNet50: Tested by Intel as of 3/26/2019. 2 socket Intel® Xeon® Platinum 8280L Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2933 MHz), BIOS: SE5C620.86B.0D.01.0348.011820191451 (ucode:0x5000017), CentOS 7.6, Kernel 4.19.5-1.el7.elrepo.x86_64, SSD 1x INTEL SSDSC2KG96 960GB, Deep Learning Framework: MXNet https://github.com/apache/incubator-mxnet.git -b master da5242b732de39ad47d8ecee582f261ba5935fa9, Compiler: gcc 6.3.1, MKL DNN version: v0.17, ResNet50: https://github.com/apache/incubator-MXNet/blob/master/python/MXNet/gluon/model_zoo/vision/resnet.py, Synthetic Data, Batch Size=64, 2 instance/2 socket, Datatype: INT8 vs Tested by Intel as of 3/26/2019. 2 socket Intel® Xeon® Platinum 8180 Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2666 MHz), BIOS: SE5C620.86B.0D.01.0286.121520181757 (ucode:0x2000057), CentOS 7.6, Kernel 4.19.5-1.el7.elrepo.x86_64, SSD 1x INTEL SSDSC2KG96 960GB, Deep Learning Framework: MXNet https://github.com/apache/incubator-mxnet.git -b master da5242b732de39ad47d8ecee582f261ba5935fa9, Compiler: gcc 6.3.1, MKL DNN version: v0.17, ResNet50: https://github.com/apache/incubator-MXNet/blob/master/python/MXNet/gluon/model_zoo/vision/resnet.py, Synthetic Data, Batch Size=64, 2 instance/2 socket, Datatype: FP32

2.5x Performance boost with MXNet SSD-VGG16 Inference: Tested by Intel as of 3/26/2019. 2 socket Intel® Xeon® Platinum 8280L Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2933 MHz), BIOS: SE5C620.86B.0D.01.0348.011820191451 (ucode:0x5000017), CentOS 7.6, Kernel 4.19.5-1.el7.elrepo.x86_64, SSD 1x INTEL SSDSC2KG96 960GB, Deep Learning Framework: MXNet https://github.com/apache/incubator-mxnet.git -b master da5242b732de39ad47d8ecee582f261ba5935fa9, Compiler: gcc 6.3.1, MKL DNN version: v0.17, SSD-VGG16: https://github.com/apache/incubator-MXNet/blob/master/example/ssd/symbol/vgg16_reduced.py, Synthetic Data, Batch Size=224, 2 instance/2 socket, Datatype: INT8 vs Tested by Intel as of 3/26/2019. 2 socket Intel® Xeon® Platinum 8180 Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2666 MHz), BIOS: SE5C620.86B.0D.01.0286.121520181757 (ucode:0x2000057), CentOS 7.6, Kernel 4.19.5-1.el7.elrepo.x86_64, SSD 1x INTEL SSDSC2KG96 960GB, Deep Learning Framework: MXNet https://github.com/apache/incubator-mxnet.git -b master da5242b732de39ad47d8ecee582f261ba5935fa9, Compiler: gcc 6.3.1, MKL DNN version: v0.17, SSD-VGG16: https://github.com/apache/incubator-MXNet/blob/master/example/ssd/symbol/vgg16_reduced.py, Synthetic Data, Batch Size=224, 2 instance/2 socket, Datatype: FP32

2.2x performance boost with Intel® Optimized Caffe SSD-Mobilenet v1: Tested by Intel as of 2/20/2019. 2 socket Intel® Xeon® Platinum 8280 Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2933 MHz), BIOS: SE5C620.86B.0D.01.0271.120720180605 (ucode: 0x4000013), Ubuntu 18.04.1 LTS, kernel 4.15.0-45-generic, SSD 1x sda INTEL SSDSC2BA80 SSD 745.2GB, Deep Learning Framework: Intel® Optimization for Caffe version: 1.1.3 (commit hash: 7010334f159da247db3fe3a9d96a3116ca06b09a) , ICC version 18.0.1, MKL DNN version: v0.17 (commit hash: 830a10059a018cd2634d94195140cf2d8790a75a), model: https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/ssd_mobilenet_int8.prototxt, Synthetic Data, Batch Size=64, 2 instance/2 socket, Datatype: INT8 vs Tested by Intel as of 2/21/2019. 2 socket Intel® Xeon® Platinum 8180 Processor, 28 cores HT On Turbo ON Total Memory 192 GB (12 slots/ 16GB/ 2666 MHz), BIOS: SE5C620.86B.00.01.0015.110720180833 (ucode: 0x200004d), CentOS 7.5, 3.10.0-693.el7.x86_64, Intel® SSD DC S4500 SERIES SSDSC2KB480G7 2.5'' 6Gb/s SATA SSD 480GB, Deep Learning Framework: Intel® Optimization for Caffe version: 1.1.3 (commit hash: 7010334f159da247db3fe3a9d96a3116ca06b09a) , ICC version 18.0.1, MKL DNN version: v0.17 (commit hash: 830a10059a018cd2634d94195140cf2d8790a75a), model: https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/ssd_mobilenet_int8.prototxt, Synthetic Data, Batch Size=64, 2 instance/2 socket, Datatype: FP32

2.6x performance boost with PyTorch RetinaNet: Tested by Intel as of 3/26/2019. 2 socket Intel® Xeon® Platinum 8280 Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2933 MHz), BIOS: SE5C620.86B.0D.01.0271.120720180605 (ucode: 0x4000013), Ubuntu 18.04.1 LTS, kernel 4.15.0-45-generic, SSD 1x sda INTEL SSDSC2BA80 SSD 745.2GB, 3X INTEL SSDPE2KX040T7 SSD 3.7TB , Deep Learning Framework: Pytorch with ONNX/Caffe2 backend: https://github.com/pytorch/pytorch.git (commit: 4ac91b2d64eeea5ca21083831db5950dc08441d6)and Pull Request link: https://github.com/pytorch/pytorch/pull/17464 (submitted for upstreaming), gcc (Ubuntu 7.3.0-27ubuntu1~18.04) 7.3.0, MKL DNN version: v0.17.3 (commit hash: 0c3cb94999919d33e4875177fdef662bd9413dd4), RetinaNet: https://github.com/intel/Detectron/blob/master/configs/12_2017_baselines/retinanet_R-101-FPN_1x.yaml BS=1, synthetic data, 2 instance/2 socket, Datatype: INT8 vs Tested by Intel as of 3/26/2019. 2 socket Intel® Xeon® Platinum 8180 Processor, 28 cores HT On Turbo ON Total Memory 192 GB (12 slots/ 16GB/ 2666 MHz), BIOS: SE5C620.86B.00.01.0015.110720180833 (ucode: 0x200004d), CentOS 7.5, 3.10.0-693.el7.x86_64, Intel® SSD DC S4500 SERIES SSDSC2KB480G7 2.5'' 6Gb/s SATA SSD 480G, ,Deep Learning Framework: Pytorch with ONNX/Caffe2 backend: https://github.com/pytorch/pytorch.git (commit: 4ac91b2d64eeea5ca21083831db5950dc08441d6)and Pull Request link: https://github.com/pytorch/pytorch/pull/17464 (submitted for upstreaming), gcc (Ubuntu 7.3.0-27ubuntu1~18.04) 7.3.0, MKL DNN version: v0.17.3 (commit hash: 0c3cb94999919d33e4875177fdef662bd9413dd4), RetinaNet: https://github.com/intel/Detectron/blob/master/configs/12_2017_baselines/retinanet_R-101-FPN_1x.yaml, BS=1, synthetic data, 2 instance/2 socket, Datatype: INT8 vs FP32

2.1x performance boost with TensorFlow Wide & Deep: Tested by Intel as of 3/26/2019. 2 socket Intel® Xeon® Platinum 8280L Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2933 MHz), BIOS: SE5C620.86B.0D.01.0348.011820191451 (ucode:0x5000017), CentOS 7.6, Kernel 4.19.5-1.el7.elrepo.x86_64, SSD 1x INTEL SSDSC2KG96 960GB, Deep Learning Framework: TensorFlow https://github.com/tensorflow/tensorflow.git A3262818d9d8f9f630f04df23033032d39a7a413 + Pull Request PR26169 + Pull Request PR26261 + Pull Request PR26271, PR submitted for upstreaming, Compiler: gcc 6.3.1, MKL DNN version: v0.18, Wide & Deep: https://github.com/IntelAI/models/tree/master/benchmarks/recommendation/tensorflow/wide_deep_large_ds commit: a044cb3e7d2b082aebae2edbe6435e57a2cc1f8f, Model: https://storage.googleapis.com/intel-optimized-tensorflow/models/wide_deep_int8_pretrained_model.pb, https://storage.googleapis.com/intel-optimized-tensorflow/models/wide_deep_fp32_pretrained_model.pb, Dataset: Criteo Display Advertisement Challenge, Batch Size=512, 1 instance/1 socket, Datatype: INT8 vs Tested by Intel as of 3/26/2019. 2 socket Intel® Xeon® Platinum 8180 Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2666 MHz), BIOS: SE5C620.86B.0D.01.0286.121520181757 (ucode:0x2000057), CentOS 7.6, Kernel 4.19.5-1.el7.elrepo.x86_64, SSD 1x INTEL SSDSC2KG96 960GB, Deep Learning Framework: TensorFlow https://github.com/tensorflow/tensorflow.git A3262818d9d8f9f630f04df23033032d39a7a413 + Pull Request PR26169 + Pull Request PR26261 + Pull Request PR26271 , PR submitted for upstreaming, Compiler: gcc 6.3.1, MKL DNN version: v0.18, Wide & Deep: https://github.com/IntelAI/models/tree/master/benchmarks/recommendation/tensorflow/wide_deep_large_ds commit: a044cb3e7d2b082aebae2edbe6435e57a2cc1f8f, Model: https://storage.googleapis.com/intel-optimized-tensorflow/models/wide_deep_int8_pretrained_model.pb, https://storage.googleapis.com/intel-optimized-tensorflow/models/wide_deep_fp32_pretrained_model.pb, Dataset: Criteo Display Advertisement Challenge, Batch Size=512, 1 instance/1 socket, Datatype: FP32

2.1x performance boost with MXNet Wide & Deep: Tested by Intel as of 3/26/2019. 2 socket Intel® Xeon® Platinum 8280L Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2933 MHz), BIOS: SE5C620.86B.0D.01.0348.011820191451 (ucode:0x5000017), CentOS 7.6, Kernel 4.19.5-1.el7.elrepo.x86_64, SSD 1x INTEL SSDSC2KG96 960GB, Deep Learning Framework: MXNet https://github.com/apache/incubator-mxnet.git commit f1de8e51999ce3acaa95538d21a91fe43a0286ec applying https://github.com/intel/optimized-models/blob/v1.0.2/mxnet/wide_deep_criteo/patch.diff, Compiler: gcc 6.3.1, MKL DNN version: commit: 08bd90cca77683dd5d1c98068cea8b92ed05784, Wide & Deep: https://github.com/intel/optimized-models/tree/v1.0.2/mxnet/wide_deep_criteo commit: c3e7cbde4209c3657ecb6c9a142f71c3672654a5, Dataset: Criteo Display Advertisement Challenge, Batch Size=1024, 2 instance/2 socket, Datatype: INT8 vs Tested by Intel as of 3/26/2019. 2 socket Intel® Xeon® Platinum 8180 Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2666 MHz), BIOS: SE5C620.86B.0D.01.0286.121520181757 (ucode:0x2000057), CentOS 7.6, Kernel 4.19.5-1.el7.elrepo.x86_64, SSD 1x INTEL SSDSC2KG96 960GB, Deep Learning Framework: MXNet https://github.com/apache/incubator-mxnet.git commit f1de8e51999ce3acaa95538d21a91fe43a0286ec applying https://github.com/intel/optimized-models/blob/v1.0.2/mxnet/wide_deep_criteo/patch.diff, Compiler: gcc 6.3.1, MKL DNN version: commit: 08bd90cca77683dd5d1c98068cea8b92ed05784, Wide & Deep: https://github.com/intel/optimized-models/tree/v1.0.2/mxnet/wide_deep_criteo commit: c3e7cbde4209c3657ecb6c9a142f71c3672654a5, Dataset: Criteo Display Advertisement Challenge, Batch Size=1024, 2 instance/2 socket, Datatype: FP32