VNNI DEEP DIVE

WALTER RIVIERA

EMEA AI TECHNICAL SOLUTION SPECIALIST | DATA CENTRE GROUP

AGENDA

VNNI Vector Neural Network Instructions

- What?
- Why?
- How?
- When?
- Summary

WHAT

Intel® Deep Learning Boost (VNNI) **is a new set of AVX-512 instructions** ***designed to deliver significant, more efficient Deep Learning (Inference) acceleration***

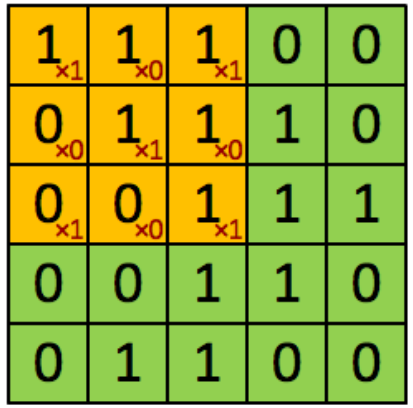
FIELD REQUIREMENT

- Matrix algebra is extensively used in deep learning. The most frequent operation is multiplying a matrix by a matrix (or vector).
- This boils down to computing an inner product

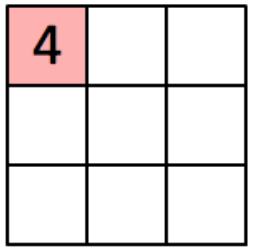
$$x_1.y_1 + x_2.y_2 + x_3.y_3 + \ldots x_n.y_n$$

- Computing this requires a series of multiply-add combinations.
- Fused Multiply-Add (FMA) provides twofold benefits:
  - Performance by only having to perform a single rounding (eliminating an addition operation)
  - Correct rounded division and square root calculations.

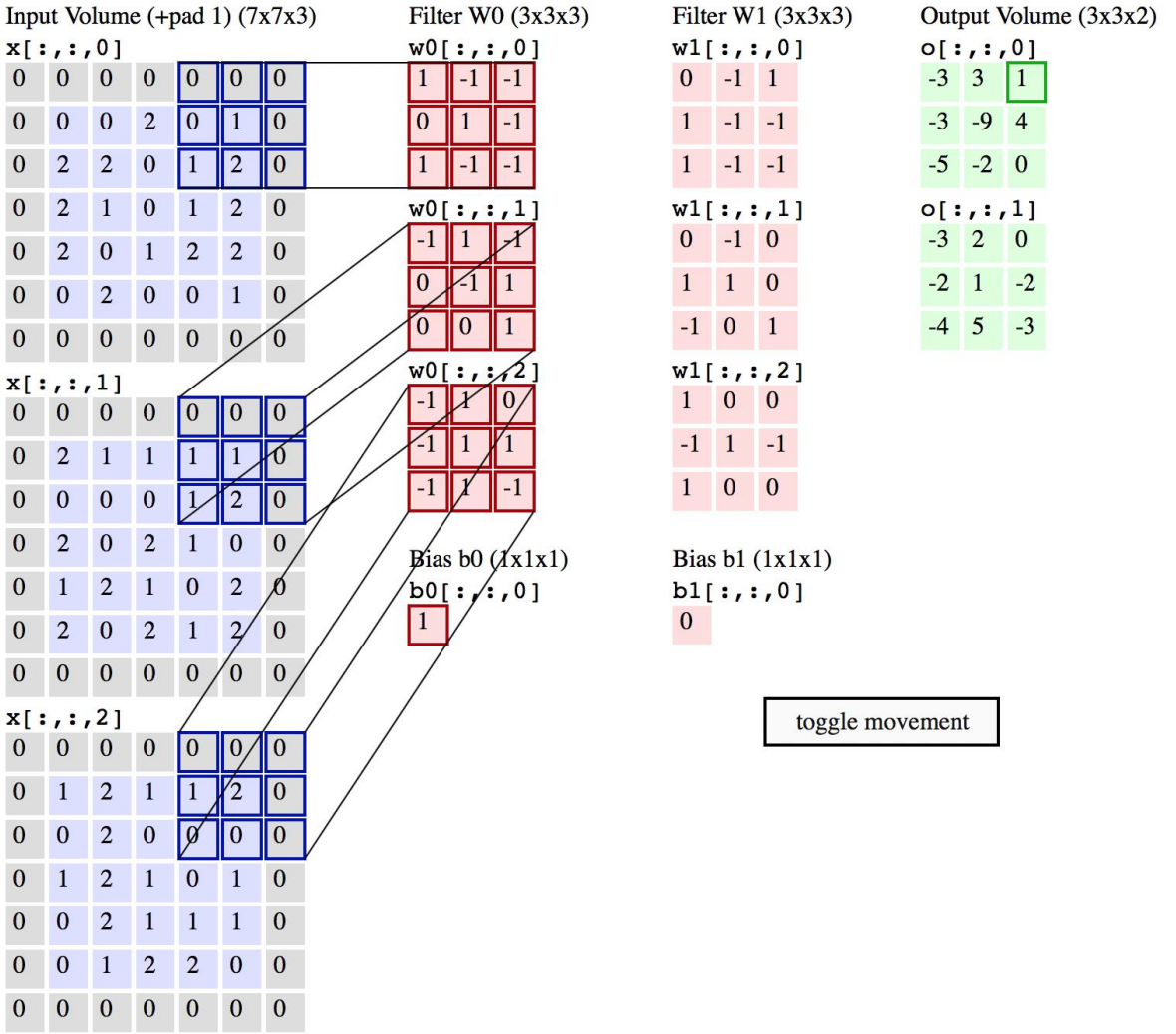
CONVOLUTION = MULTIPLY - ADD OP.



Image



Convolved Feature



SOFTWARE MATTERS!

INFERENCE THROUGHPUT

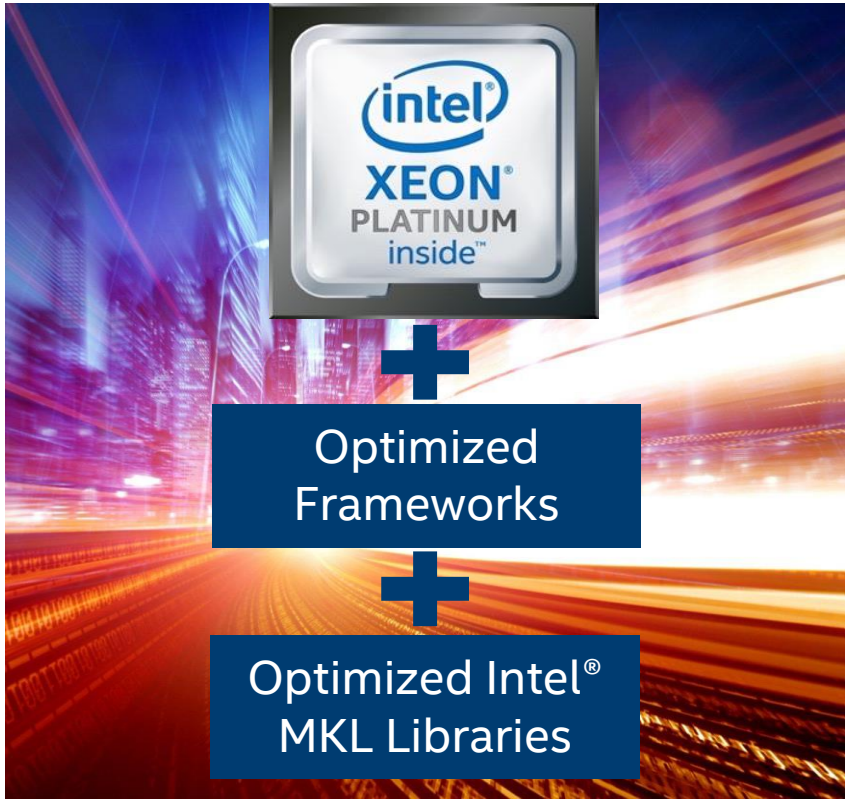
Up to
277x[1]

Intel® Xeon® Platinum 8180 Processor
higher Intel optimized Caffe GoogleNet v1 with Intel® MKL
inference throughput compared to
Intel® Xeon® Processor E5-2699 v3 with BVLC-Caffe
Inference and training throughput uses FP32 instructions

TRAINING THROUGHPUT

Up to
241x[1]

Intel® Xeon® Platinum 8180 Processor
higher Intel Optimized Caffe AlexNet with Intel® MKL
training throughput compared to
Intel® Xeon® Processor E5-2699 v3 with BVLC-Caffe

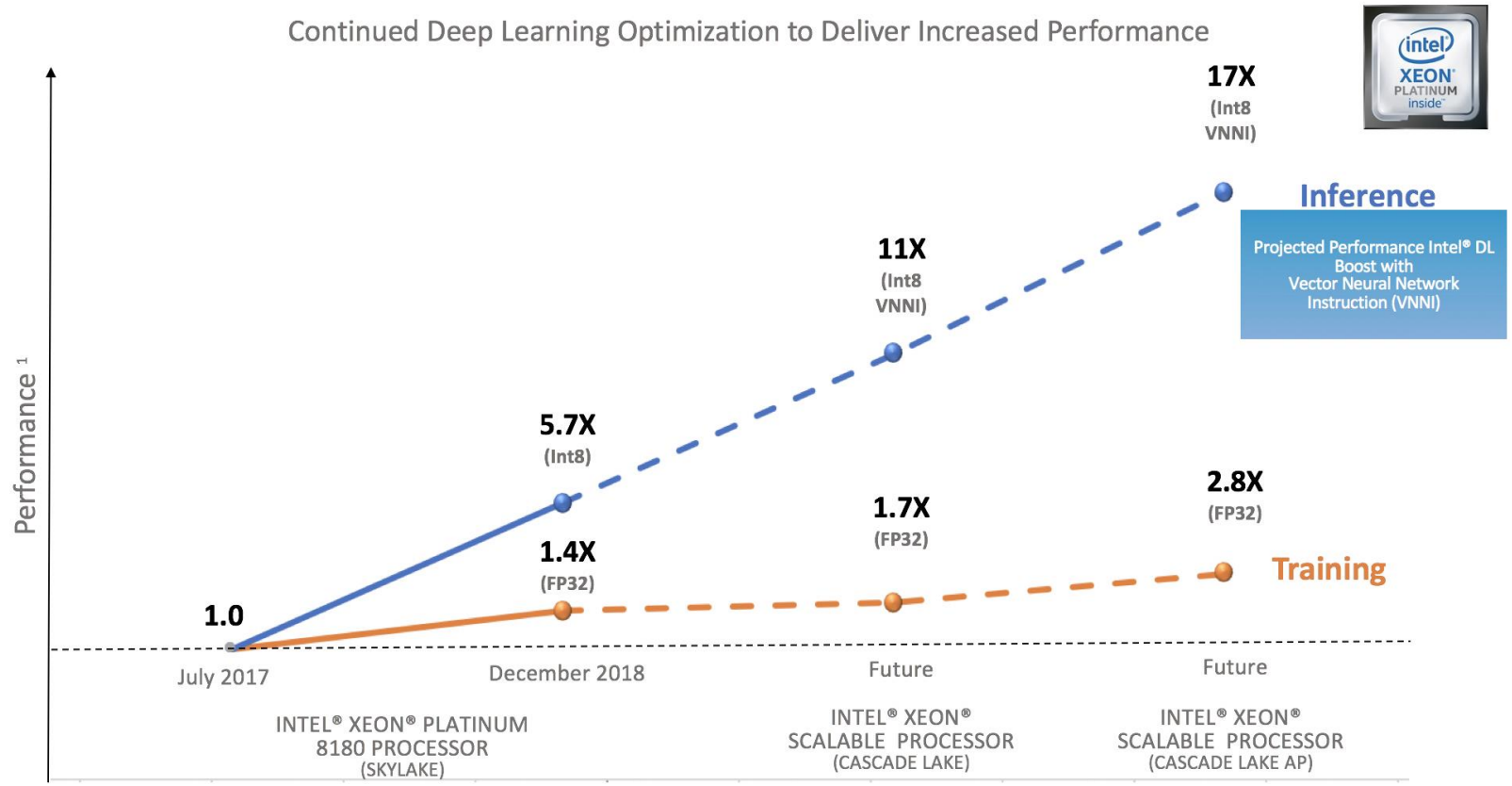


Deliver significant AI performance with hardware and software optimizations on Intel® Xeon® Scalable Family

[1] See configuration disclosure for details (config 49).

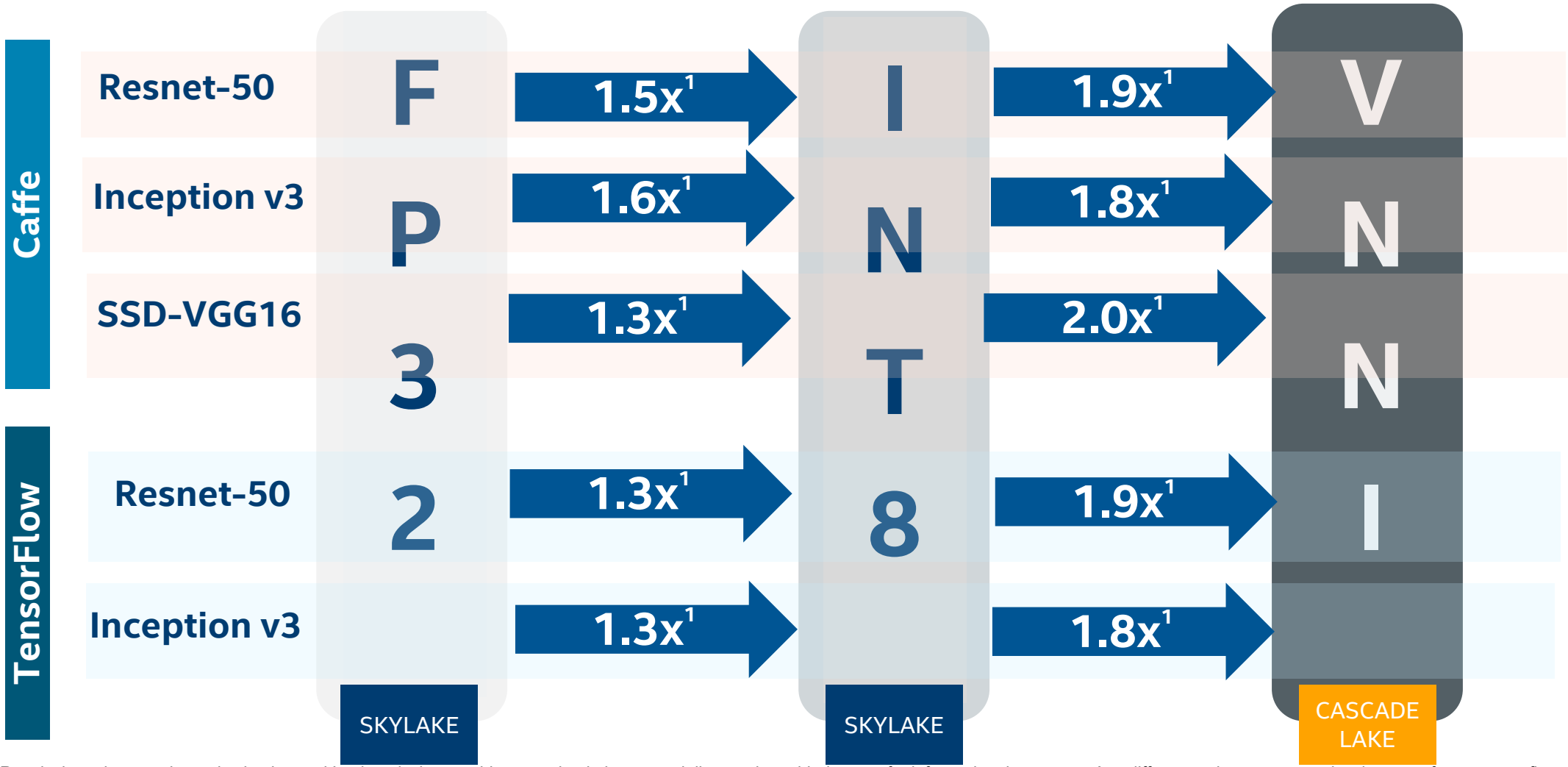
INTEL® XEON® SCALABLE PROCESSORS

Continued Deep Learning Optimization to Deliver Increased Performance



1 Intel® Optimization for Caffe Resnet-50 performan imization for Café Resnet-50 inference throughput performance on Intel® Xeon® Scalable Processor. See Configuration Details 11X (7/25/2018), 17X (10/31/2018), 1.7X (11/5/2018), 2.8X (11/5/2018) Results have been estimated using internal Intel analysis, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance. See configuration disclosure for details. No product can be absolutely secure. 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: http://www.intel.com/performance.

GENERATIONAL PERFORMANCE PROJECTIONS ON INTEL® SCALABLE PROCESSOR FOR DEEP LEARNING INFERENCE FOR POPULAR CNNS



[1]8/24/2018) Results have been estimated 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. 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: http://www.intel.com/performance

SO WHAT? ☺

Theoretical peak compute gains are:

- 4x *int8* OPS over *fp32* OPS and ¼ memory requirements
- 2x *int16* OPS over *fp32* OPS and ½ memory requirements

Reference [1]

INSTRUCTIONS FUSION

VPMADDUBSW

Multiply and Add Packed Signed and Unsigned Bytes

SRC 1 8-bit	A_0	A_1	A_2	A_3	. . . .	A_{63}
SRC 2 8-bit	B_0	B_1	B_2	B_3	. . . .	B_{63}
DEST 16-bit	$A_0*B_0 + A_1*B_1$		$A_2*B_2 + A_3*B_3$		. . . .	$A_{62}*B_{62} + A_{63}*B_{63}$

VPMADDWD

effectively upconvert to 32-bit and horizontal add of neighbors

SRC 1 16-bit	$A_0*B_0 + A_1*B_1$	$A_2*B_2 + A_3*B_3$	. . . .	$A_{62}*B_{62} + A_{63}*B_{63}$
SRC 2 16-bit	1	1	. . . .	1
DEST 32-bit	$A_0*B_0 + A_1*B_1 + A_2*B_2 + A_3*B_3$		. . . .	$A_{60}*B_{60} + A_{61}*B_{61} + A_{62}*B_{62} + A_{63}*B_{63}$

VPADDD

Add Packed Double-Precision Floating-Point Values

SRC 1 32-bit	$A_0*B_0 + A_1*B_1 + A_2*B_2 + A_3*B_3$	. . . .	$A_{60}*B_{60} + A_{61}*B_{61} + A_{62}*B_{62} + A_{63}*B_{63}$
SRC 2 32-bit	C_0	. . . .	C_{15}
Dest 32-bit	$A_0*B_0 + A_1*B_1 + A_2*B_2 + A_3*B_3 + C_0$	. . . .	$A_{60}*B_{60} + A_{61}*B_{61} + A_{62}*B_{62} + A_{63}*B_{63} + C_{15}$

ISA Fusion

VPDPBUSD

Multiply and Add Unsigned and Signed Bytes

SRC 1 8-bit	A_0	A_1	A_2	A_3	. . . .	A_{63}
SRC 2 8-bit	B_0	B_1	B_2	B_3	. . . .	B_{63}
SRC3 / DEST 32-bit	C_0				. . . .	C_{15}
	$A_0*B_0 + A_1*B_1 + A_2*B_2 + A_3*B_3 + C_0$			. . .	$A_{60}*B_{60} + A_{61}*B_{61} + A_{62}*B_{62} + A_{63}*B_{63} + C_{15}$	

VPMADDUBSW + VPMADDWD + VPADDD fused into VPDPBUSD (3x peak OPs)

INSTRUCTIONS SET REFERENCES

- Intel® 64 and IA-32 Architectures Software Developer Manuals
  This boils down to computing an inner product

https://software.intel.com/en-us/articles/intel-sdm

- Intel® Architecture Instruction Set Extensions and Future Features Programming Reference

https://software.intel.com/sites/default/files/managed/c5/15/architecture-instruction-set-extensions-programming-reference.pdf

INTEL® DEEP LEARNING BOOST

SOFTWARE OPTIMIZATION EXAMPLES

FRAMEWORKS
LIBRARIES

Caffe TensorFlow mxnet

Intel® MKL-DNN

VALUE PILLAR

PERFORMANCE

END CUSTOMER VALUE

Designed to accelerate AI/Deep Learning workloads (image classification, speech recognition, language translation, object detection and more)

PROBLEM SOLVED

Intel® AVX-512	VNNI
VPMADDUBSW VPMADDWD VPADDD	VPDPBUSD (8-bit new instruction)

Low Precision Integer Operations

CUSTOMER SEGMENTS

Cloud Service Providers

Enterprise

Comms Service Providers

SUMMARY

AVX512_VNNI is a new set of of AVX-512 instructions to boost Deep Learning performance

- VNNI includes FMA instructions for:
  - 8-bit multiplies with 32-bit accumulates (u8 x s8 ⇒ s32)
  - 16-bit multiplies with 32-bit accumulates (s16 x s16 ⇒ s32)
- Theoretical peak compute gains are:
  - 4x int8 OPS over fp32 OPS and ¼ memory requirements
  - 2x int16 OPS over fp32 OPS and ½ memory requirements
- Ice Lake and future microarchitectures will have AVX512_VNNI

REFERENCES

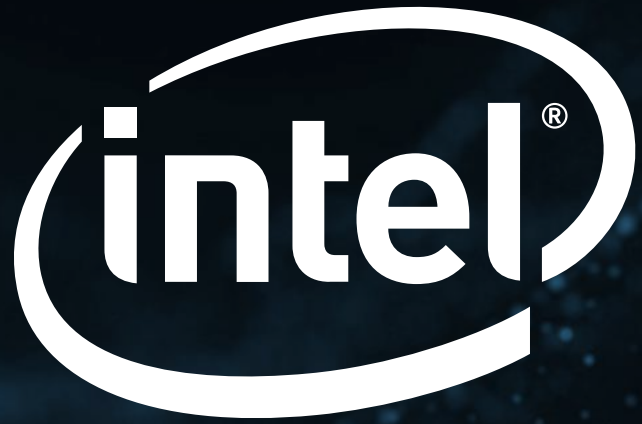
- To replicate the demo:

https://github.com/IntelAI/models/tree/master/benchmarks/image_recognition/tensorflow/inceptionv3#int8-inference-instructions

- VNNI White paper:

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

THANK YOU


intel

CONFIGURATION: SOFTWARE MATTERS

INFERENCE using FP32 Batch Size Caffe GoogleNet v1 128 AlexNet 256.

Configurations for Inference throughput

Tested by Intel as of 6/7/2018:Platform :2 socket Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores HT ON , Turbo ON Total Memory 376.28GB (12slots / 32 GB / 2666 MHz),4 instances of the framework, CentOS Linux-7.3.1611-Core , SSD sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB , Deep Learning Framework caffe version: a3d5b022fe026e9092fc7abc7654b1162ab9940d Topology:GoogleNet v1 BIOS:SE5C620.86B.00.01.0004.071220170215 MKLDNN: version: 464c268e544bae26f9b85a2acb9122c766a4c396 NoDataLayer. Measured: 1449 imgs/sec vs Tested by Intel as of 06/15/2018 Platform: 2S Intel® Xeon® CPU E5-2699 v3 @ 2.30GHz (18 cores), HT enabled, turbo disabled, scaling governor set to “performance” via intel_pstate driver, 64GB DDR4-2133 ECC RAM. BIOS: SE5C610.86B.01.01.0024.021320181901, CentOS Linux-7.5.1804(Core) kernel 3.10.0-862.3.2.el7.x86_64, SSD sdb INTEL SSDSC2BW24 SSD 223.6GB. Framework BVLC-Caffe: https://github.com/BVLC/caffe, Inference & Training measured with “caffe time” command. For “ConvNet” topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. BVLC Caffe (http://github.com/BVLC/caffe), revision 2a1c552b66f026c7508d390b526f2495ed3be594

Configuration for training throughput:

Tested by Intel as of 05/29/2018 Platform :2 socket Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores HT ON , Turbo ON Total Memory 376.28GB (12slots / 32 GB / 2666 MHz),4 instances of the framework, CentOS Linux-7.3.1611-Core , SSD sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB , Deep Learning Framework caffe version: a3d5b022fe026e9092fc7abc765b1162ab9940d Topology:alexnet BIOS:SE5C620.86B.00.01.0004.071220170215 MKLDNN: version: 464c268e544bae26f9b85a2acb9122c766a4c396 NoDataLayer. Measured: 1257 imgs/sec vs Tested by Intel as of 06/15/2018 Platform: 2S Intel® Xeon® CPU E5-2699 v3 @ 2.30GHz (18 cores), HT enabled, turbo disabled, scaling governor set to “performance” via intel_pstate driver, 64GB DDR4-2133 ECC RAM. BIOS: SE5C610.86B.01.01.0024.021320181901, CentOS Linux-7.5.1804(Core) kernel 3.10.0-862.3.2.el7.x86_64, SSD sdb INTEL SSDSC2BW24 SSD 223.6GB. Framework BVLC-Caffe: https://github.com/BVLC/caffe, Inference & Training measured with “caffe time” command. For “ConvNet” topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. BVLC Caffe (http://github.com/BVLC/caffe), revision 2a1c552b66f026c7508d390b526f2495ed3be594

CONFIGURATION DETAILS (CONT'D)

Configuration: AI Performance – Software + Hardware

1.4x training throughput improvement in August 2018:

Tested by Intel as of measured August 2nd 2018. Processor: 2 socket Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores HT ON , Turbo ON Total Memory 376.46GB (12slots / 32 GB / 2666 MHz). CentOS Linux-7.3.1611-Core kernel 3.10.0-693.11.6.el7.x86_64, SSD sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB , Deep Learning Framework Intel® Optimizations for caffe version:a3d5b022fe026e9092fc7abc7654b1162ab9940d Topology::resnet_50 BIOS:SE5C620.86B.00.01.0013.030920180427 MKLDNN: version: 464c268e544bae26f9b85a2acb9122c766a4c396 NoDataLayer. Measured: 123 imgs/sec vs Intel tested July 11th 2017 Platform: Platform: 2S Intel® Xeon® Platinum 8180 CPU @ 2.50GHz (28 cores), HT disabled, turbo disabled, scaling governor set to "performance" via intel_pstate driver, 384GB DDR4-2666 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD DC S3700 Series (800GB, 2.5in SATA 6Gb/s, 25nm, MLC).Performance measured with: Environment variables: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, CPU Freq set with cpupower frequency-set -d 2.5G -u 3.8G -g performance. Caffe: (http://github.com/intel/caffe/), revision f96b759f71b2281835f690af267158b82b150b5c. Inference measured with "caffe time --forward_only" command, training measured with "caffe time" command. For "ConvNet" topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from https://github.com/intel/caffe/tree/master/models/intel_optimized_models (GoogLeNet, AlexNet, and ResNet-50), https://github.com/intel/caffe/tree/master/models/default_vgg_19 (VGG-19), and https://github.com/soumith/convnet-benchmarks/tree/master/caffe/imagenet_winners (ConvNet benchmarks; files were updated to use newer Caffe prototxt format but are functionally equivalent). Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.20170425. Caffe run with "numactl -l".

5.4x inference throughput improvement in August 2018:

Tested by Intel as of measured July 26th 2018 :2 socket Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores HT ON , Turbo ON Total Memory 376.46GB (12slots / 32 GB / 2666 MHz). CentOS Linux-7.3.1611-Core, kernel: 3.10.0-862.3.3.el7.x86_64, SSD sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB , Deep Learning Framework Intel® Optimized caffe version:a3d5b022fe026e9092fc7abc7654b1162ab9940d Topology::resnet_50_v1 BIOS:SE5C620.86B.00.01.0013.030920180427 MKLDNN: version:464c268e544bae26f9b85a2acb9122c766a4c396 instances: 2 instances socket:2 (Results on Intel® Xeon® Scalable Processor were measured running multiple instances of the framework. Methodology described here: https://software.intel.com/en-us/articles/boosting-deep-learning-training-inference-performance-on-xeon-and-xeon-phi) NoDataLayer. Datatype: INT8 Batchsize=64 Measured: 1233.39 imgs/sec vs Tested by Intel as of July 11th 2017:2S Intel® Xeon® Platinum 8180 CPU @ 2.50GHz (28 cores), HT disabled, turbo disabled, scaling governor set to "performance" via intel_pstate driver, 384GB DDR4-2666 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD DC S3700 Series (800GB, 2.5in SATA 6Gb/s, 25nm, MLC).**Performance measured with**: Environment variables: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, CPU Freq set with cpupower frequency-set -d 2.5G -u 3.8G -g performance. Caffe: (http://github.com/intel/caffe/), revision f96b759f71b2281835f690af267158b82b150b5c. Inference measured with "caffe time --forward_only" command, training measured with "caffe time" command. For "ConvNet" topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50). Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.20170425. Caffe run with "numactl -l".

11X inference thoughput improvement with CascadeLake:

Future Intel Xeon Scalable processor (codename Cascade Lake) 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 vs Tested by Intel as of July 11th 2017: 2S Intel® Xeon® Platinum 8180 CPU @ 2.50GHz (28 cores), HT disabled, turbo disabled, scaling governor set to "performance" via intel_pstate driver, 384GB DDR4-2666 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD DC S3700 Series (800GB, 2.5in SATA 6Gb/s, 25nm, MLC).**Performance measured with**: Environment variables: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, CPU Freq set with cpupower frequency-set -d 2.5G -u 3.8G -g performance. Caffe: (http://github.com/intel/caffe/), revision f96b759f71b2281835f690af267158b82b150b5c. Inference measured with "caffe time --forward_only" command, training measured with "caffe time" command. For "ConvNet" topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50),. Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.20170425. Caffe run with "numactl -l".

LEGAL NOTICES & DISCLAIMERS

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.

Intel, the Intel logo, Pentium, Celeron, Atom, Core, Xeon, Movidius and others are trademarks of Intel Corporation in the U.S. and/or other countries.
*Other names and brands may be claimed as the property of others.