KNIGHTS MILL: NEW INTEL PROCESSOR FOR MACHINE LEARNING

Dennis Bradford, Sundaram Chinthamani, Jesus Corbal, Adhiraj Hassan, Ken Janik, Nawab Ali

Legal Disclaimers

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at [intel.com].

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.

§ Configurations: see each performance slide notes for configurations.

§ For more information go to http://www.intel.com/performance.

Relative 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.

SPEC, SPECint, SPECfp, SPECrate, SPECpower, SPECjbb, SPECompG, SPEC MPI, and SPECjEnterprise* are trademarks of the Standard Performance Evaluation Corporation. See http://www.spec.org for more information.

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

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

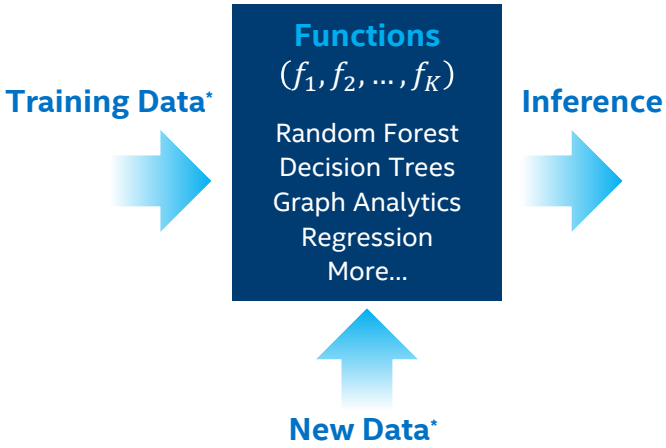
Intel, the Intel logo, Intel Xeon, Xeon logo, Intel Xeon Phi logo, and the Look Inside. Logo are trademarks of Intel Corporation in the U.S. and/or other countries.

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

What is Machine Learning?

CLASSIC ML

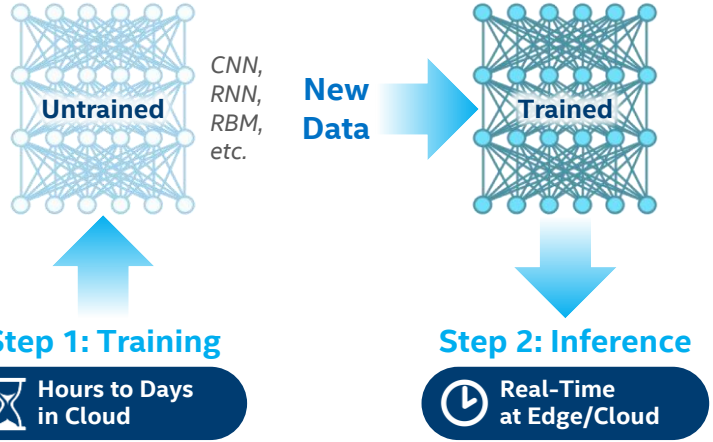
Using functions or algorithms
to extract insights from new data



High amount of Human Interaction

DEEP LEARNING

Using massive data sets to train deep (neural)
graphs that can extract insights from new data



Models Trained Autonomously

*Not all classical machine learning algorithms require separate training and new data sets

AI DATACENTER

ALL PURPOSE

Intel® Xeon® Processor Family

MOST AGILE AI PLATFORM

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

HIGHLY-PARALLEL

Intel® Xeon Phi™ Processor (Knights Mill†)

FASTER DL TRAINING

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

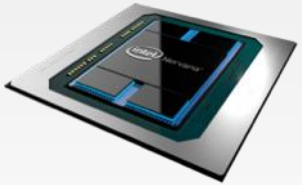
FLEXIBLE ACCELERATION

Intel® FPGA

ENHANCED DL INFERENCE

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

DEEP LEARNING



Crest Family†

DEEP LEARNING BY DESIGN

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

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

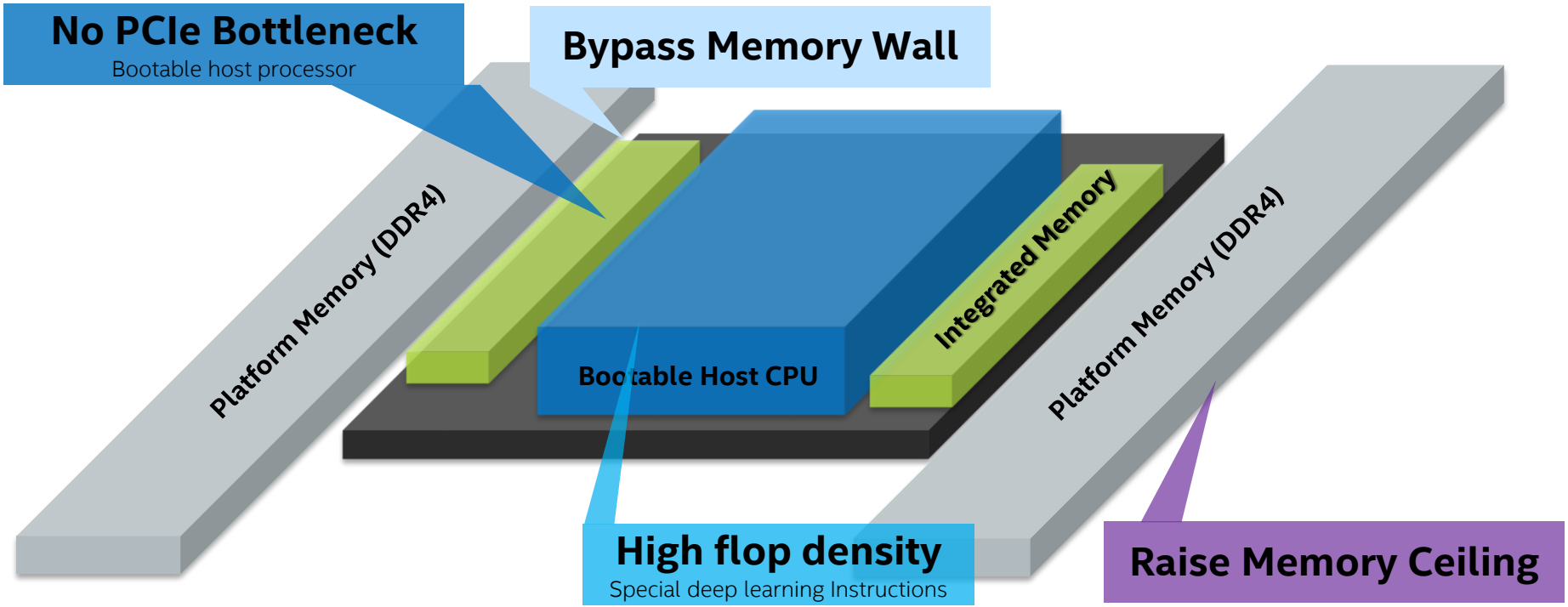
What is Knights Mill?

- First Intel product targeted specifically at Deep Learning training workloads
  - Up to 4x DL Peak performance over Xeon Phi™ 7200 Series[1]
- *Built on top of 2nd generation Intel® Xeon Phi™ processor*
  - Improved efficiency
  - Optimized for scale-out
  - Enhanced variable precision
  - Flexible, high capacity memory

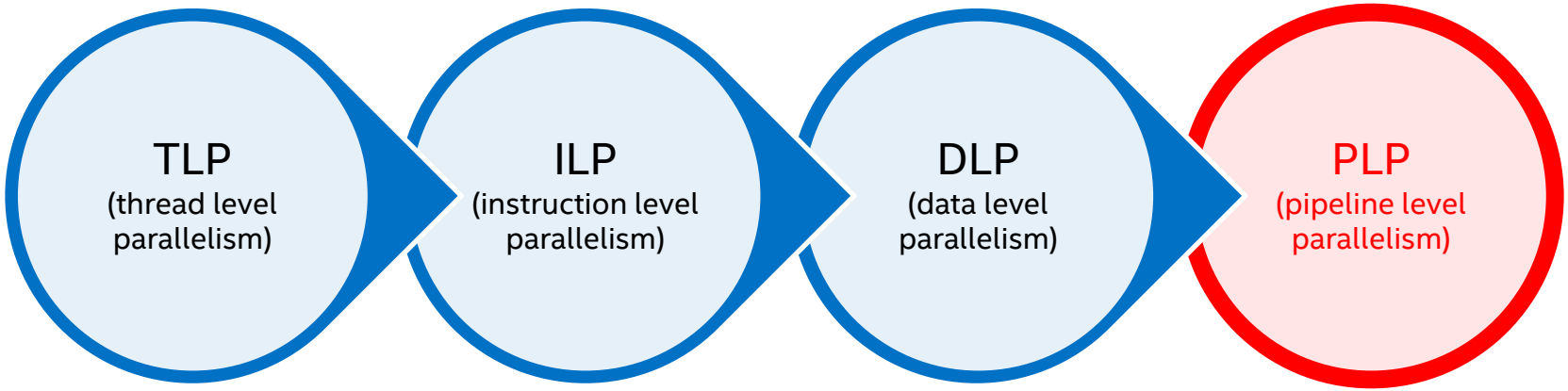
[1]Intel internal estimate: Performance estimate wrt Xeon Phi™ 7290 SKU SGEMM. Performance Calculation= AVX freq X Cores X Flops per Core X Efficiency

Knights Mill – New Intel Processor for Deep Learning

Designed for Deep Learning – "AI on IA"



Knights Mill exploits all ~~3~~ 4 levels of parallelism



- Many core architecture on high performing mesh interconnect
- 4-way SMT

- 2-way superscalar
- OOO execution

- 512-bit SIMD (AVX-512)
- *New* Variable Precision support (VNNI)

- *New* Quad FMA instructions

New Deep Learning ISA: Quad FMA FP32

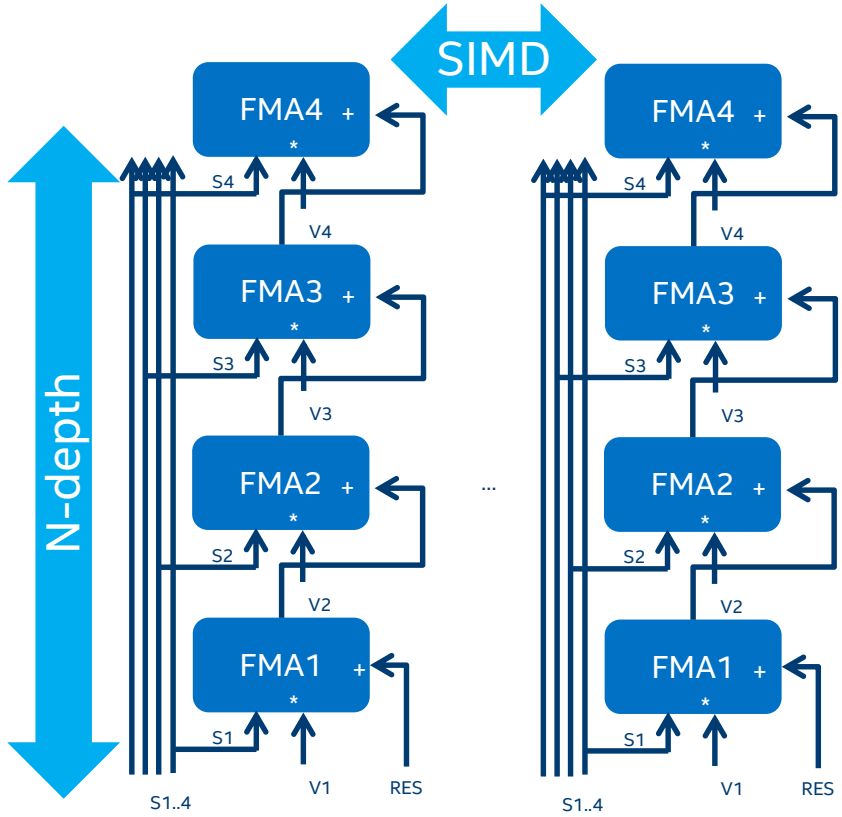
Mnemonic	Format	Description
V4FMADDPS	zmm1 {k1}, zmm2+3, m128	Quadruple packed single-precision multiply and add
…	…	…

"source block" of 4 zmm sources

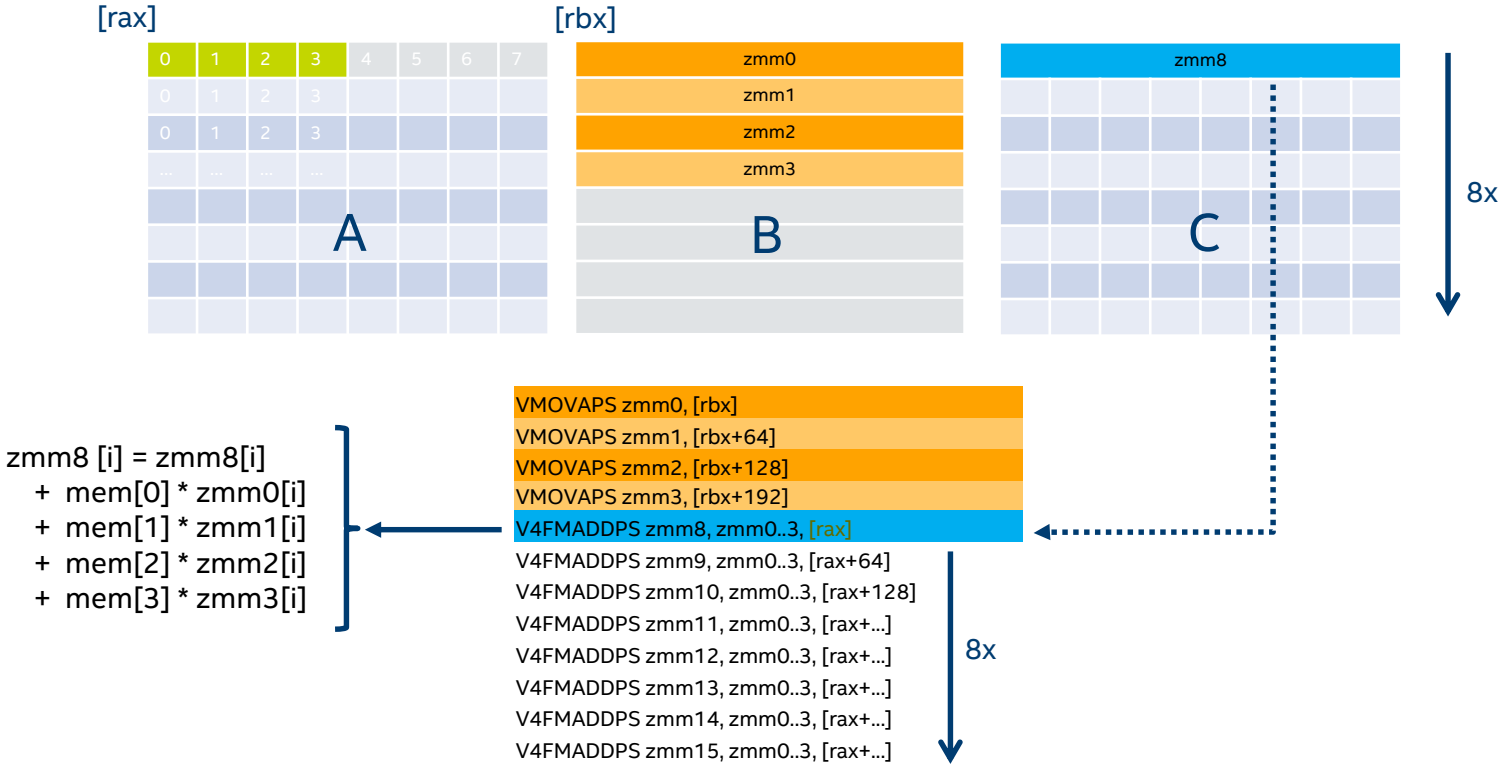
Memory operand packing 4 scalars (4x32)

V4FMADDPS zmm4 {k1}, zmm0+3, m128

```
for i=0..15
   zmm4.fp32[i] = zmm4.fp32[i]
        + zmm0.fp32[i]*m128.fp32[0]
        + zmm1.fp32[i]*m128.fp32[1]
        + zmm2.fp32[i]*m128.fp32[2]
        + zmm3.fp32[i]*m128.fp32[3]
```

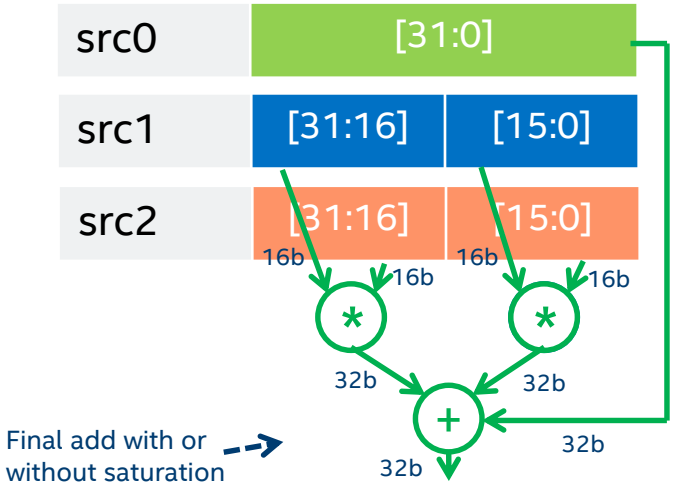


An Example: Using Quad FMA on Matrix Multiply

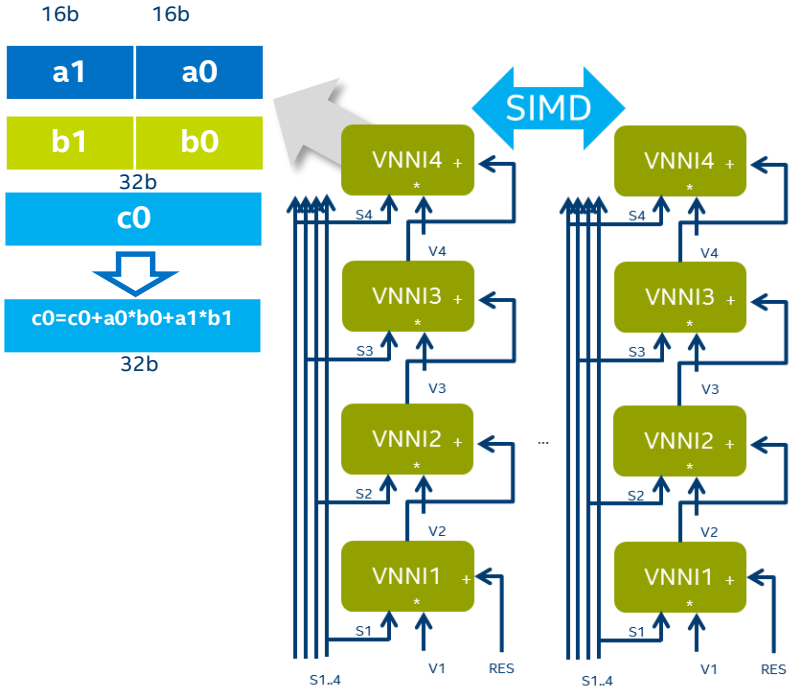


Variable Precision: What is VNNI-16?

- Vector Neural Network Instructions
- Variable precision
  - ***Inputs***: 16-bit INT
  - ***Outputs***: 32-bit INT
- Variable precision is best of both worlds
  - **Same operations/instruction as 'half precision'**
    - 2x OPS vs Single Precision
  - **Similar output precision for optimal training convergence**
    - 31 bits of INT32 vs 24 bits of mantissa in FP32
  - The obvious trade-off is the associated overhead on handling dynamic range in software (fixed precision)



QVNNI = QFMA + VNNI



Instruction	Format	Description
VP4DPWSSD	zmm1 {k1}, zmm2+3, mem128	Quadruple INT16 to INT32 horizontal MAC
VP4DPWSSDS	zmm1 {k1}, zmm2+3, mem128	Quadruple INT16 to INT32 horizontal MAC with signed saturation

- Example
  - VP4DPWSSD zmm4 {k1}, zmm0+3, m128
    - for i=0..15
      - zmm4.int32[i] = zmm4.int32[i]
        + (zmm0.int16[2*i]*m128.int16[0] + zmm0.int16[2*i+1]*m128.int16[1])
        + (zmm1.int16[2*i]*m128.int16[2] + zmm1.int16[2*i+1]*m128.int16[3])
        + (zmm2.int16[2*i]*m128.int16[4] + zmm2.int16[2*i+1]*m128.int16[5])
        + (zmm3.int16[2*i]*m128.int16[6] + zmm3.int16[2*i+1]*m128.int16[7])

Knights Mill Core

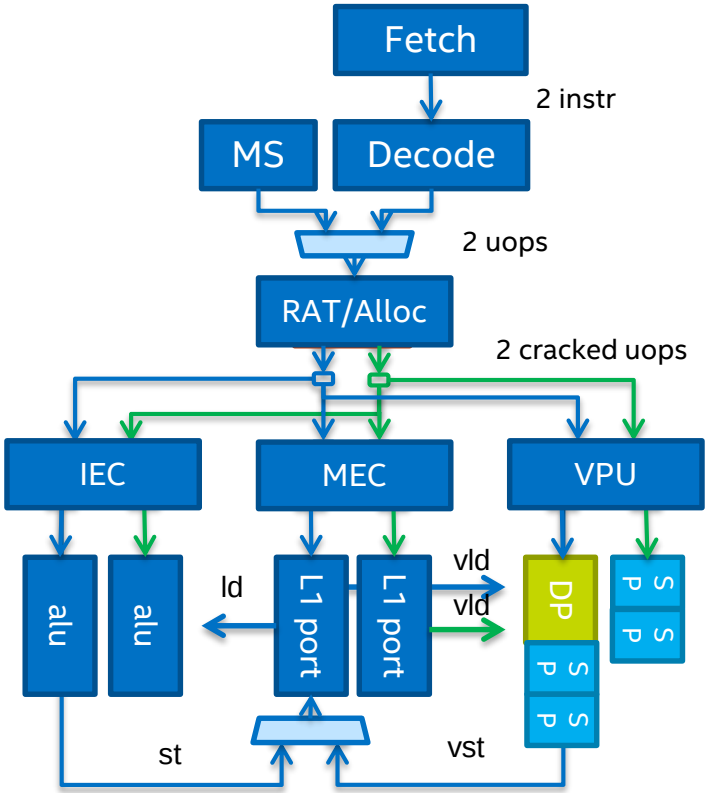
ISA: SSE, AVX, AVX512-F

Double Precision stack

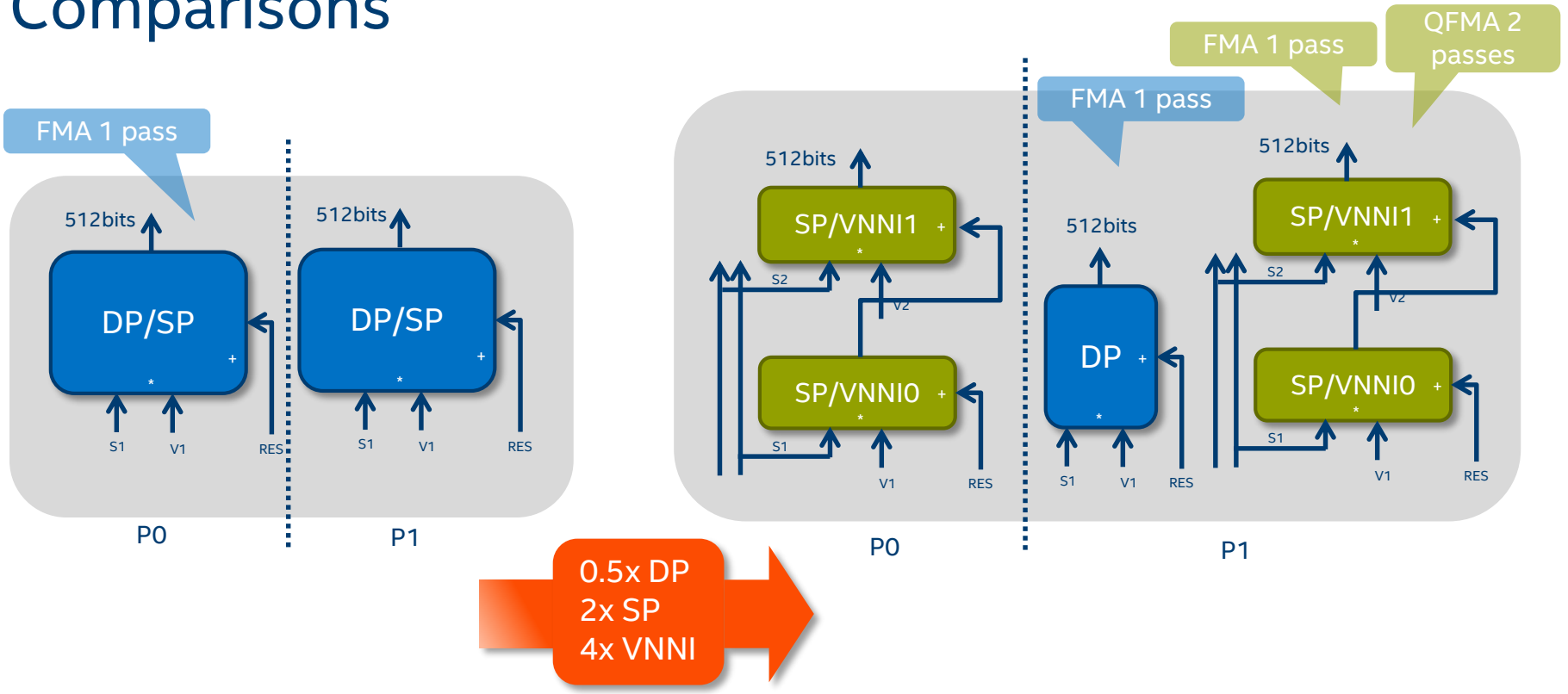
- 1 VPU port/core (512b)

Single Precision/VNNI stack

- 2 stacked FMAs per port



Intel® Xeon Phi™ 7200 Series vs. Knights Mill: Port Comparisons



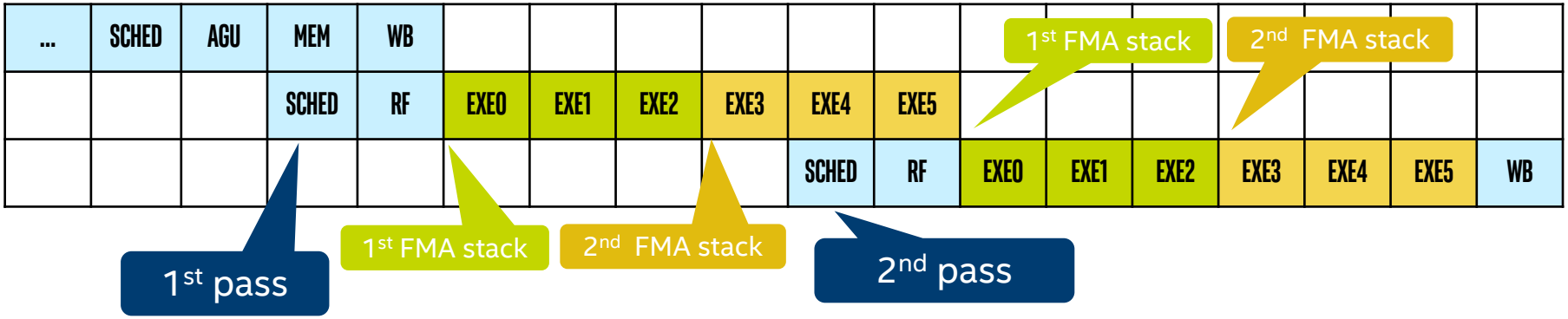
Quad FMA Double-pumped Execution (*)

(*) Included for illustration purposes, not intended as an exact recreation of KNM pipeline stages

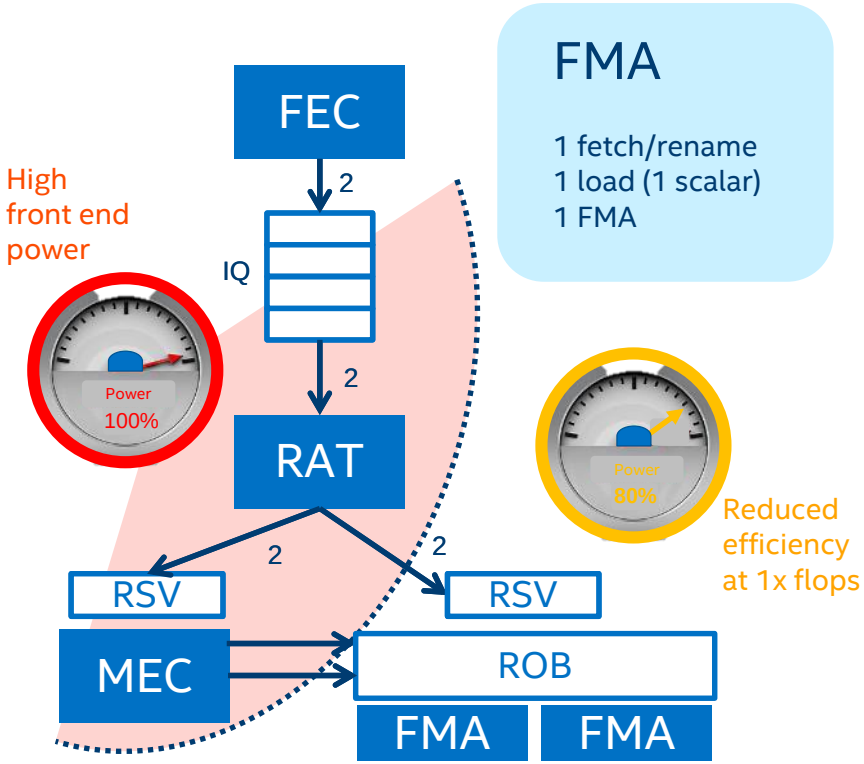
The life of a Single-precision FMA instruction in Knights Mill

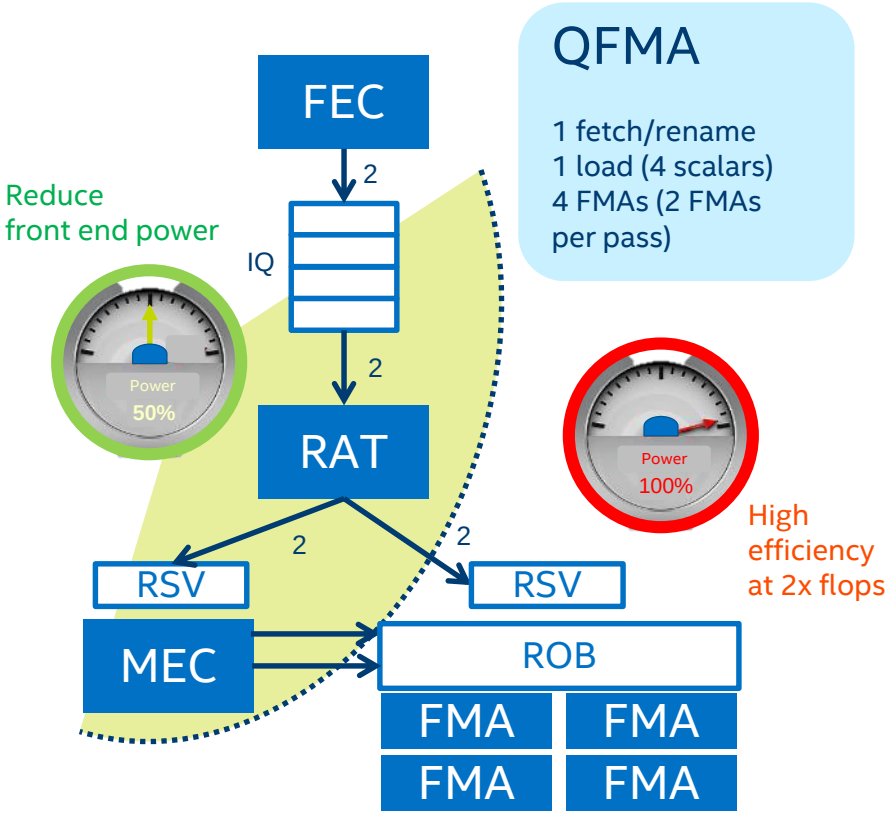
FETCH	DEC	RAT	SCHED	AGU	MEM	WB				
					SCHED	RF	EXE0	EXE1	EXE2	WB

The life of a Single-precision QFMA instruction in Knights Mill



Efficiency of double-pumped execution





Knights Mill: Putting All Together

Knights Mill is Xeon Phi™ 7200 series Derivative

- Xeon Phi™ 7200 series & Knights Mill share the same compute architecture
- Built for different markets
- Xeon Phi™ 7200 series → HPC workloads
- Knights Mill → deep learning training workloads

Xeon Phi™ 7200 series & Knights Mill are the **same** generation of Intel® Xeon Phi™ products

Adding New Instruction Sets

- Knights Mill uses new instructions to adjust performance
- Compared to Xeon Phi™ 7200 series:
  - 2x single precision
  - ½ double precision
  - 4x using new QVNNI

Up to 4x* performance over Knights Landing for Deep Learning workloads via QVNNI

Deep Learning Software Optimizations

- Intel is optimizing library & frameworks used for deep learning training
- Investments apply to Intel® Xeon® and Xeon Phi™ processors, & FPGAs

S/W optimizations give up to 400x performance over non-optimized Intel products**

*Performance estimate wrt Xeon Phi™ 7290 SKU SGEMM. Performance Calculation= AVX freq X Cores X Flops per Core X Efficiency
**Config details in backup

INTEL AI PORTFOLIO

AI FRAMEWORKS

SELECT YOUR FAVORITE AI FRAMEWORK

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

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

✓OPTIMIZED FOR INTEL ARCHITECTURE

INTEL LIBRARIES, FRAMEWORKS & TOOLS

	Intel® MKL	MKL-DNN	Intel® MLSL	Intel® Data Analytics Acceleration Library (DAAL)	Intel® Distribution for python	Open Source Frameworks (neon, TensorFlow, mxnet)	Intel Deep Learning SDK	Intel® Computer Vision SDK
High Level Overview	High performance math primitives granting low level of control	Free open source DNN functions for high-velocity integration with deep learning frameworks	Primitive communication building blocks to scale deep learning framework performance over a cluster	Broad data analytics acceleration object oriented library supporting distributed ML at the algorithm level	Most popular and fastest growing language for machine learning	Toolkits driven by academia and industry for training machine learning algorithms	Accelerate deep learning model design, training and deployment	Toolkit to develop & deploying vision-oriented solutions that harness the full performance of Intel CPUs and SOC accelerators
Primary Audience	Consumed by developers of higher level libraries and Applications	Consumed by developers of the next generation of deep learning frameworks	Deep learning framework developers and optimizers	Wider Data Analytics and ML audience, Algorithm level development for all stages of data analytics	Application Developers and Data Scientists	Machine Learning App Developers, Researchers and Data Scientists.	Application Developers and Data Scientists	Developers who create vision-oriented solutions
Example Usage	Framework developers call matrix multiplication, convolution functions	New framework with functions developers call for max CPU performance	Framework developer calls functions to distribute Caffe training compute across an Intel® Xeon Phi™ cluster	Call distributed alternating least squares algorithm for a recommendation system	Call scikit-learn k-means function for credit card fraud detection	Script and train a convolution neural network for image recognition	Deep Learning training and model creation, with optimization for deployment on constrained end device	Use deep learning to do pedestrian detection

Find out more at software.intel.com/ai

Call to Action: Visit Intel Websites for more info

Intel.com/AI

- Stories and Use Cases
- AI Academy Access
- AI Product Overviews

Intelnervana.com

- Nervana Platform Info
- Optimization resources
- Nervana Graph
- Events & Partners


intel®
experience
what's inside™