ACCELERATE YOUR INFERENCING WITH INTEL® DEEP LEARNING BOOST

Shailen Sobhee

AI Software Technical Consultant

shailen.sobhee@intel.com

Audience pre-requisites

- Familiar with deep learning stages
  - Training and inferencing
- Have a basic knowledge about hardware
  - know what are vector registers, like AVX-512

Outline

- What is Intel® Deep Learning Boost (Intel® DL Boost)
- Why is Intel® DL Boost useful?
- What are **V**ector **N**eural **N**etwork **I**nstructions (**VNNI**)
- Sample results

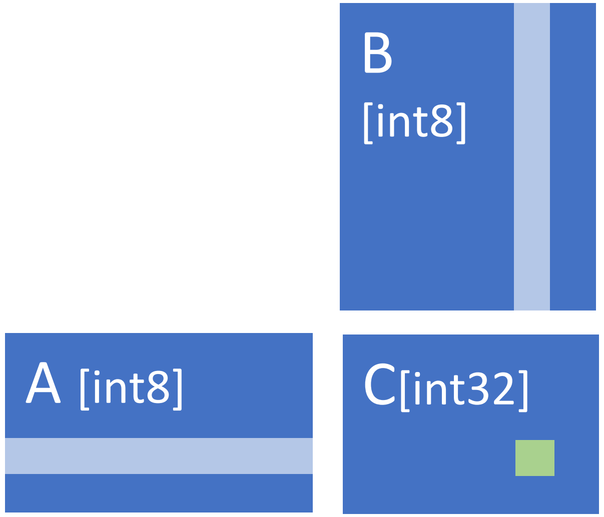
What is Intel® Deep Learning Boost ?

Intel® DL Boost:

- extends the AVX-512 instructions
- designed to deliver **significant** and **more efficient** Deep Learning (Inference) acceleration
- for deep learning workloads optimized to use the Vector Neural Network Instruction (VNNI)
- on Intel® Xeon® Scalable processor
- as from the 2nd generation (codename “Cascade Lake”)

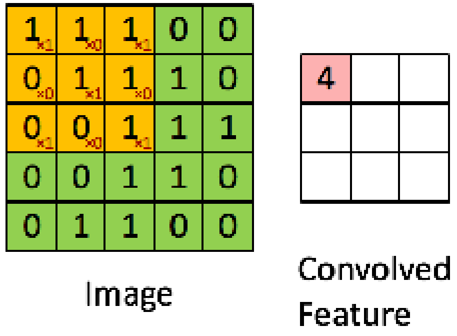
Deep Learning Foundations

- Heavy compute (Matrix Multiplications) are the foundation of many DL applications
  - **Multiply** a row*column values, **accumulate** into a single value
- Traditional HPC and many AI training workloads use floating point
  - Massive dynamic range of values (FP32 goes up to ~2^128)



Matrix Multiply
A x B = C

Convolution operation for inference

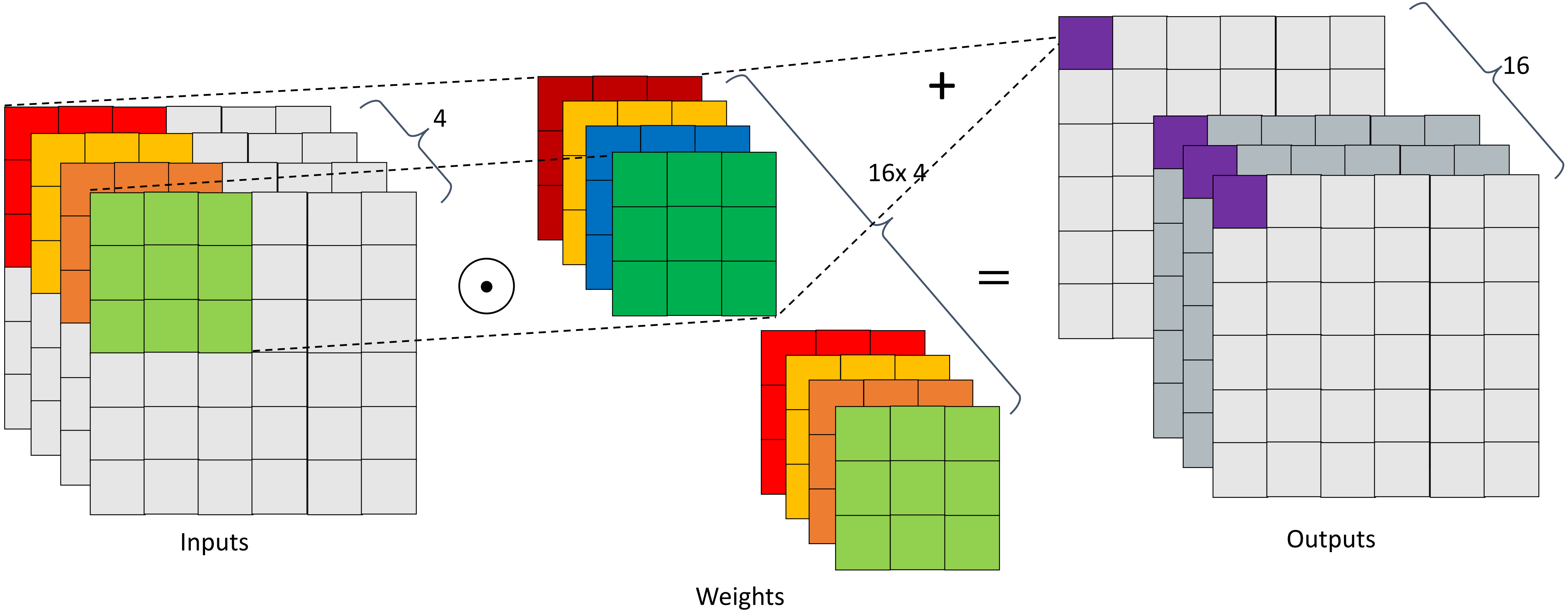


$1_{\times1}$	$1_{\times0}$	$1_{\times1}$	0	0
$0_{\times0}$	$1_{\times1}$	$1_{\times0}$	1	0
$0_{\times1}$	$0_{\times0}$	$1_{\times1}$	1	1
0	0	1	1	0
0	1	1	0	0

Image

4		

Convolved Feature

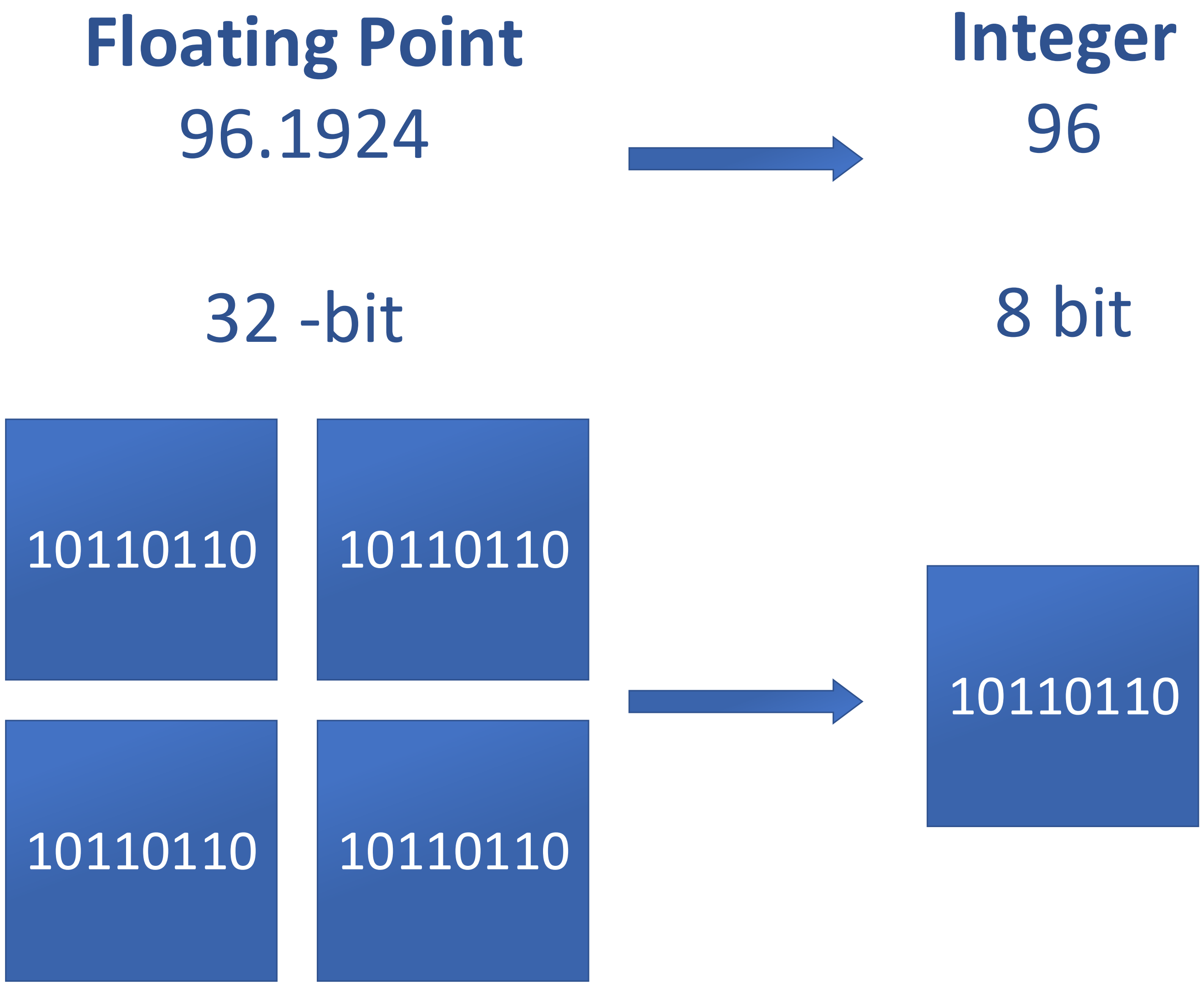


Why do we need Intel® Deep Learning Boost?

The key term:

Quantization

Here’s why Quantization matters



Floating Point

96.1924

32 -bit

10110110

10110110

10110110

10110110

Integer

96

8 bit

10110110

Here’s why Quantization matters

Lower Power

Lower memory bandwidth

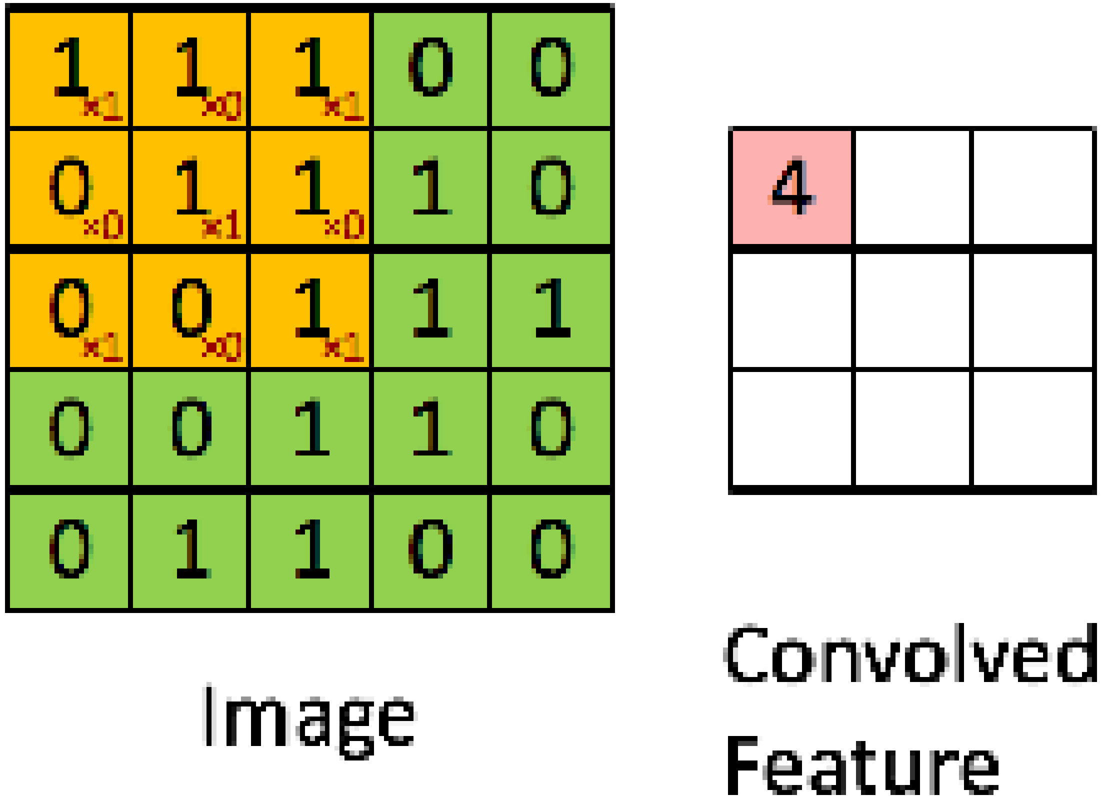
Lower storage

Higher performance

Important: Acceptable accuracy loss

Image credits: see backup

VNNI INSTRUCTION SET



$1_{\times 1}$	$1_{\times 0}$	$1_{\times 1}$	0	0
$0_{\times 0}$	$1_{\times 1}$	$1_{\times 0}$	1	0
$0_{\times 1}$	$0_{\times 0}$	$1_{\times 1}$	1	1
0	0	1	1	0
0	1	1	0	0

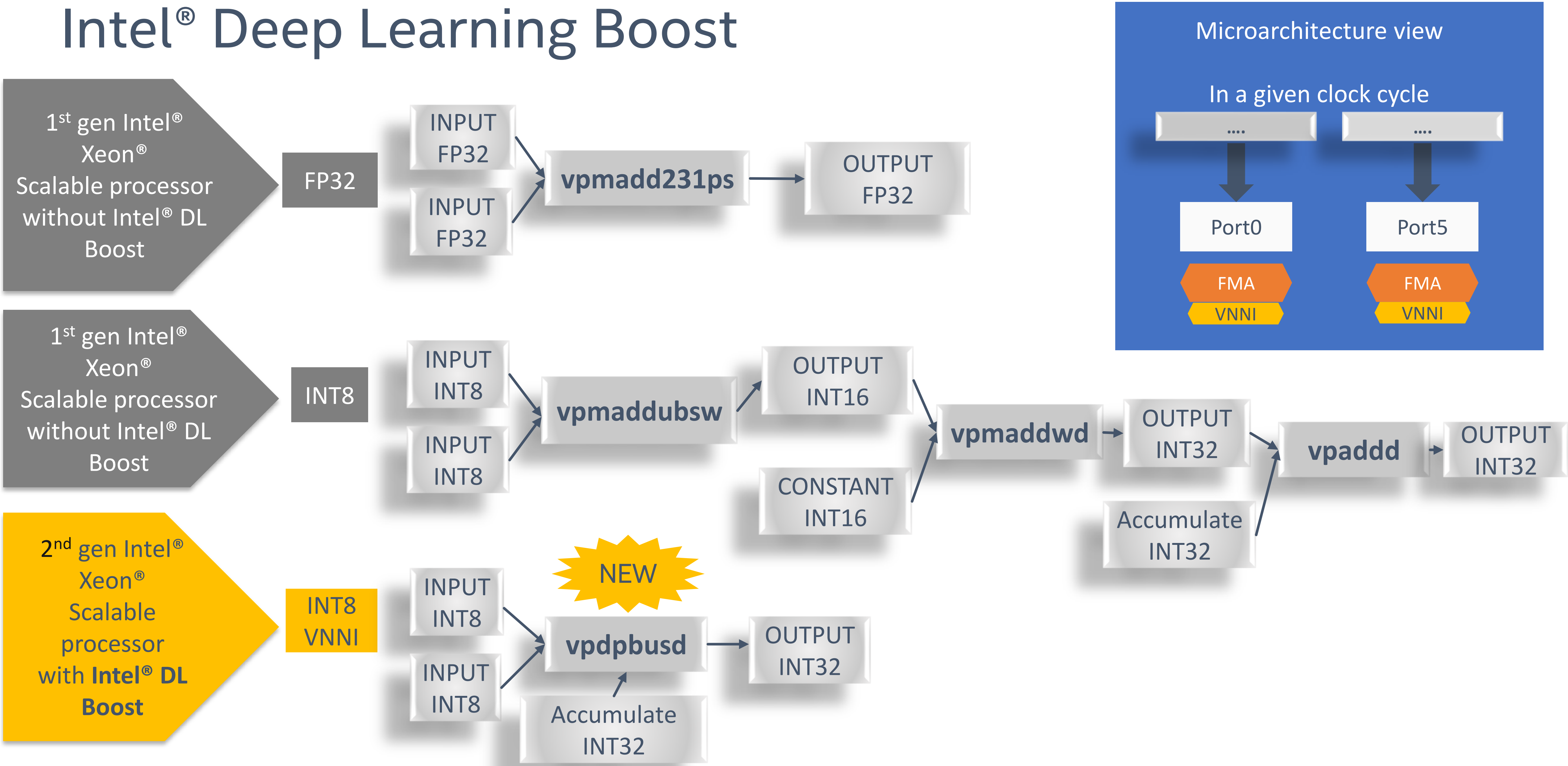
Image

4		

Convolved Feature

Optimizing AI inference

Intel® Deep Learning Boost



1st gen Intel® Xeon® Scalable processor without Intel® DL Boost

FP32

INPUT FP32

INPUT FP32

vpmadd231ps

OUTPUT FP32

1st gen Intel® Xeon® Scalable processor without Intel® DL Boost

INT8

INPUT INT8

INPUT INT8

vpmaddubsw

OUTPUT INT16

CONSTANT INT16

vpmaddwd

OUTPUT INT32

Accumulate INT32

vpaddd

OUTPUT INT32

2nd gen Intel® Xeon® Scalable processor **with Intel® DL Boost**

INT8 VNNI

INPUT INT8

INPUT INT8

NEW

vpdpbusd

Accumulate INT32

OUTPUT INT32

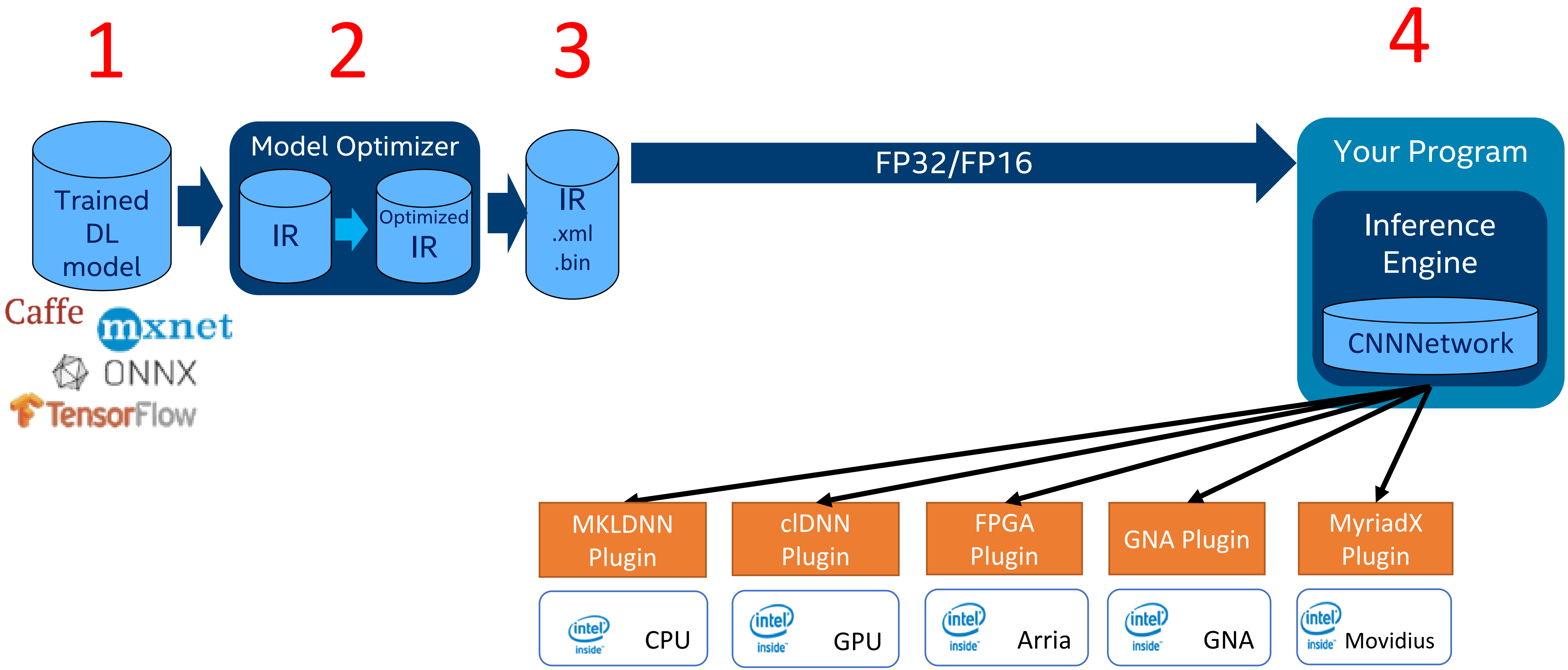
Microarchitecture view

In a given clock cycle

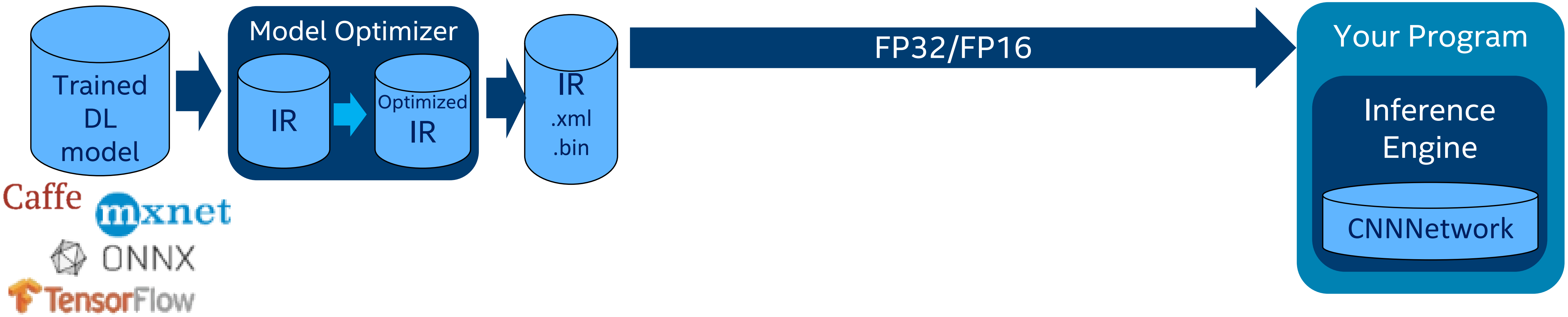
Here’s one tool in your arsenal to do it ☺

OpenVINO™

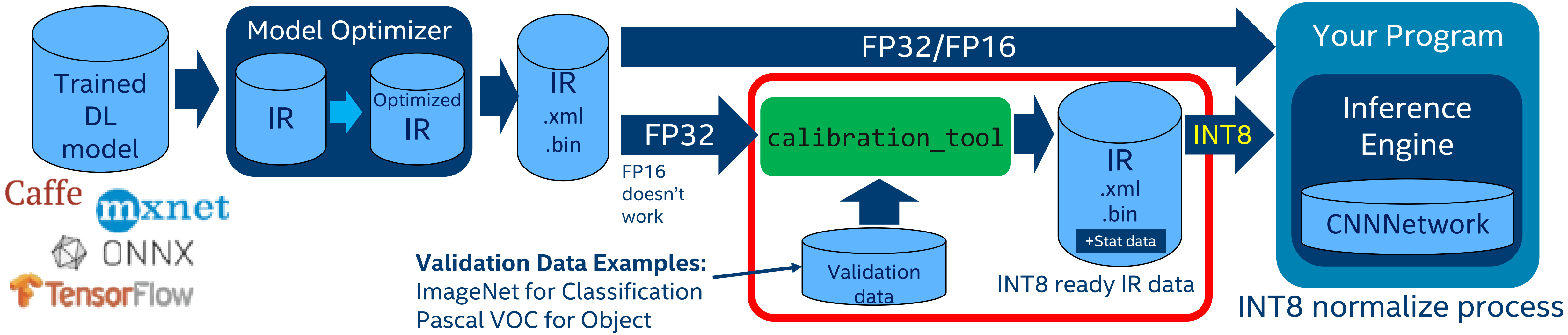
Intel® Distribution of OpenVINO™ in a nutshell



Intel® Distribution of OpenVINO™ in a nutshell

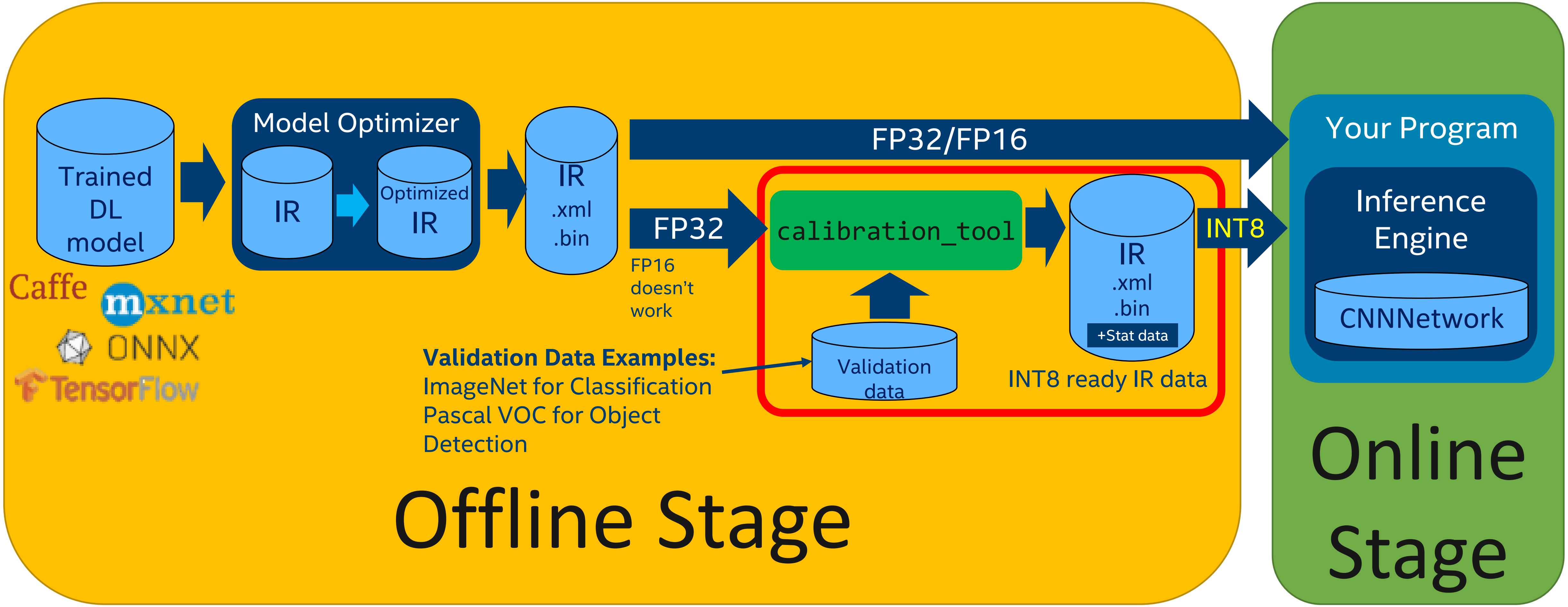


Intel® Distribution of OpenVINO™ in a nutshell



Validation Data Examples:
ImageNet for Classification
Pascal VOC for Object Detection

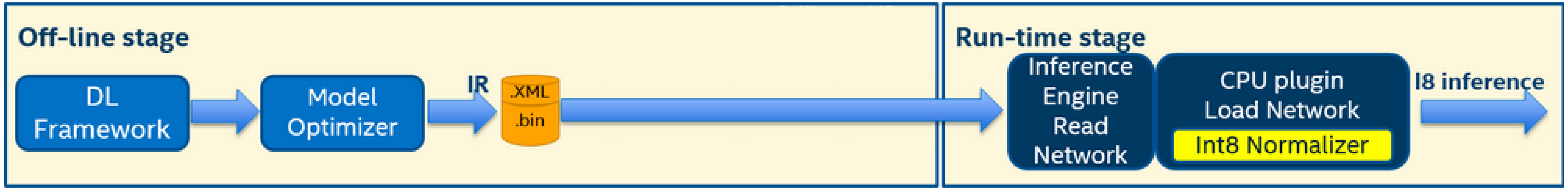
Intel® Distribution of OpenVINO™ in a nutshell



https://docs.openvinotoolkit.org/latest/_docs_IE_DG_Int8Inference.html

Sample results

Demo 1



Off-line stage

DL Framework

Model Optimizer

IR

.XML

.bin

Run-time stage

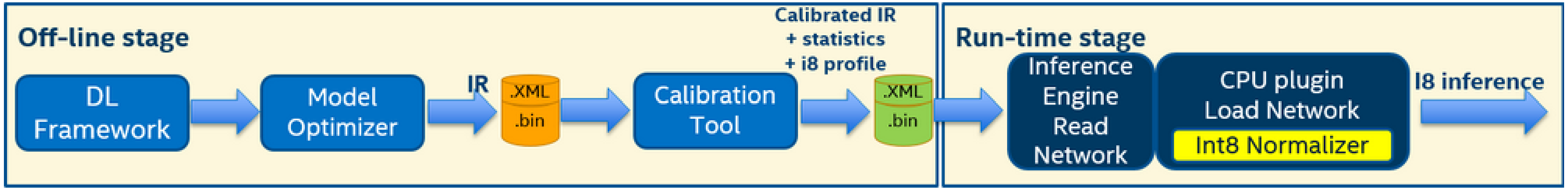
Inference Engine Read Network

CPU plugin Load Network

Int8 Normalizer

I8 inference

Demo 2



Off-line stage

DL Framework

Model Optimizer

IR

.XML

.bin

Calibration Tool

Calibrated IR
+ statistics
+ i8 profile

.XML

.bin

Run-time stage

Inference Engine Read Network

CPU plugin Load Network

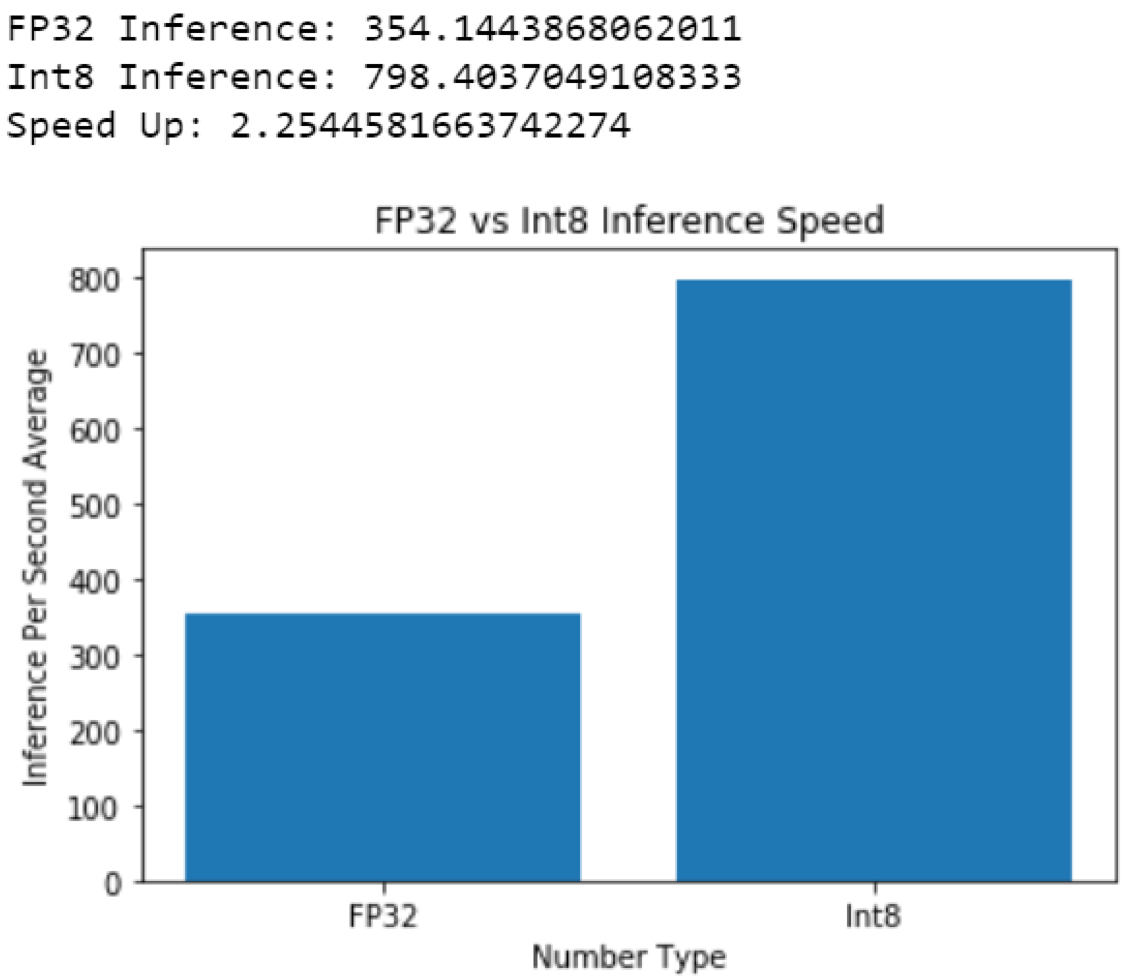
Int8 Normalizer

I8 inference

Both executions on Intel® Cascade Lake CPU

QUICK DEMO

Sample results



```
FP32 Inference: 354.1443868062011
Int8 Inference: 798.4037049108333
Speed Up: 2.2544581663742274
```

Key take away

Try the **Intel® Distribution of OpenVINO™**: https://software.intel.com/en-us/openvino-toolkit

Benefit from **faster** inference speeds with **INT8** leveraging **VNNI** instructions on **Intel® Cascade Lake** CPUs.

Summary

- What is Intel® Deep Learning Boost (Intel® DL Boost)
- What are **V**ector **N**eural **N**etwork **I**nstructions **(VNNI)**
- Why is Intel® DL Boost useful?
- Intel® Distribution of OpenVINO™

THANK YOU!