



ACCELERATE YOUR INFERENCE WITH INTEL® DEEP LEARNING BOOST

Shailen Sobhee

AI Software Technical Consultant

shailen.sobhee@intel.com

010N ▶ TR/01 ▶ 03
010N ▶ TR/01 ▶ 03

▶ RS/0211 SEARCH... A01
▶ RS/0211 SEARCH... A01

▶ SEARCH ▶ TR/01 ▶ 03
▶ SEARCH ▶ TR/01 ▶ 03



Outline

- What is Intel® Deep Learning Boost (Intel® DL Boost)
- Why is Intel® DL Boost useful?
- What are **Vector Neural Network Instructions (VNNI)**
- Sample results

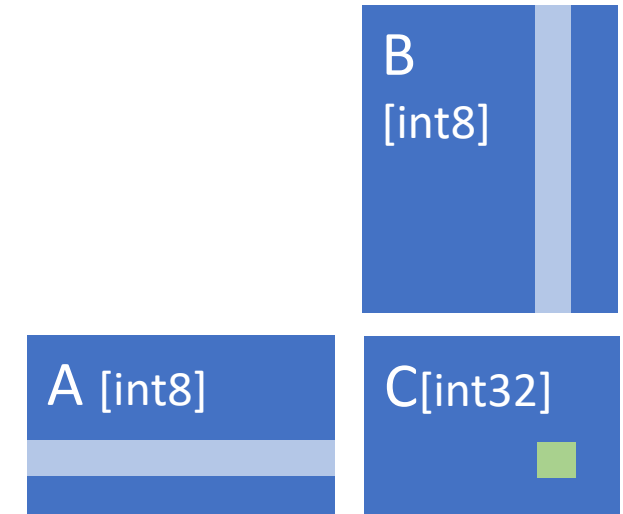
What is Intel® Deep Learning Boost ?

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

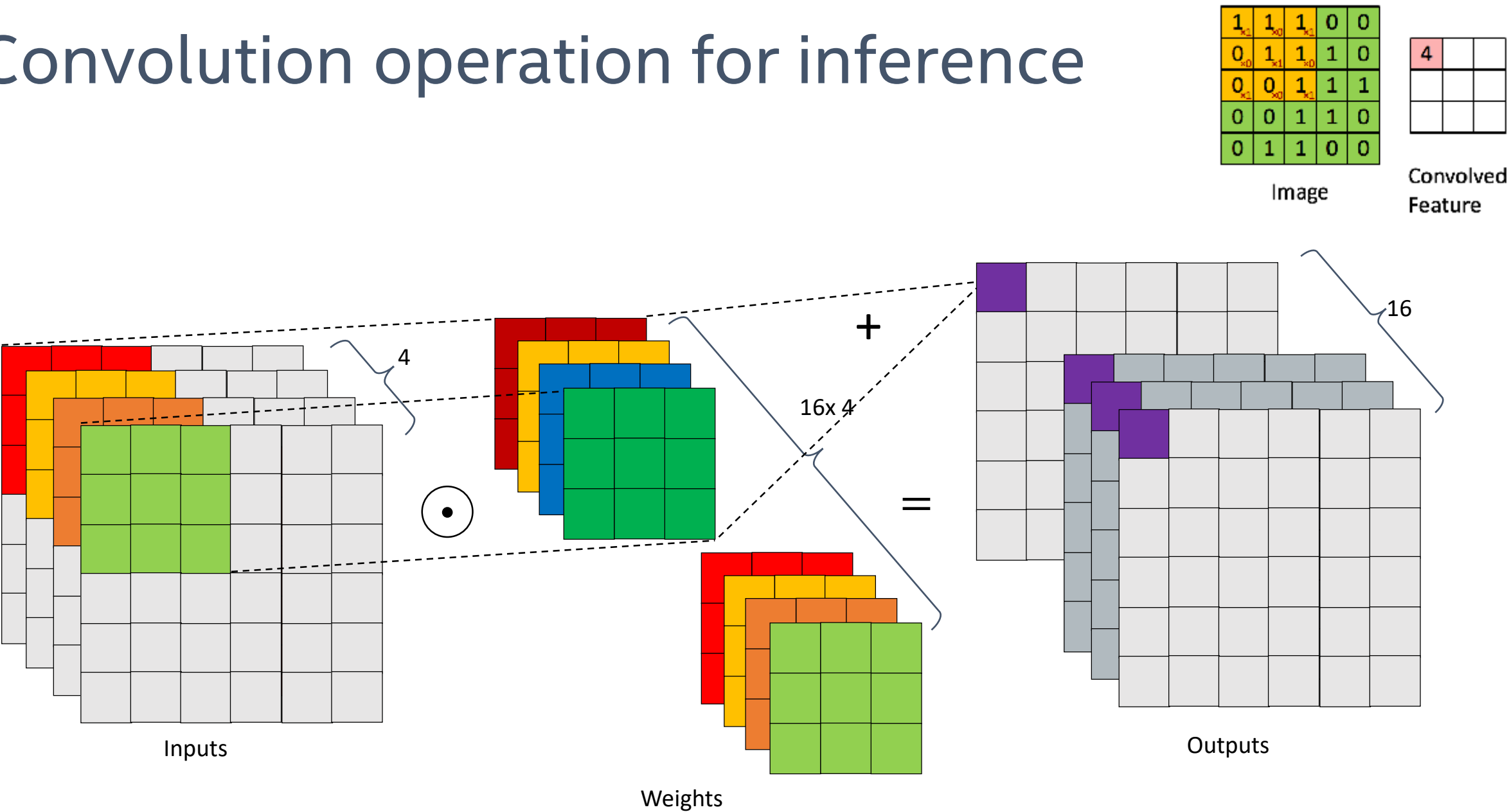
- Heavy compute (Matrix Multiplications) are the foundation of many DL applications
 - **Multiply** a row*column values, **accumulate** into a single value
- Traditional HPC and many AI training workloads use floating point
 - Massive dynamic range of values (FP32 goes up to $\sim 2^{128}$)



Matrix Multiply

$$A \times B = C$$

Convolution operation for inference



Why do we need Intel® Deep Learning Boost?



The key term:

Quantization

Here's why Quantization matters

Floating Point

96.1924

Integer

96



32 -bit

8 bit

10110110

10110110

10110110

10110110

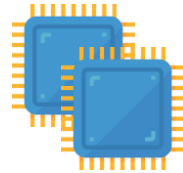


10110110

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

►RS:0211 SEARCH...A01
►RS:0211 SEARCH...A01

►SEARCH►TR/01►03
►SEARCH►TR/01►03

010N ►TR/01►03
010N ►TR/01►03



1 <small>x1</small>	1 <small>x0</small>	1 <small>x1</small>	0	0
0 <small>x0</small>	1 <small>x1</small>	1 <small>x0</small>	1	0
0 <small>x1</small>	0 <small>x0</small>	1 <small>x1</small>	1	1
0	0	1	1	0
0	1	1	0	0

Image

4		

Convolved
Feature

Intel® Deep Learning Boost

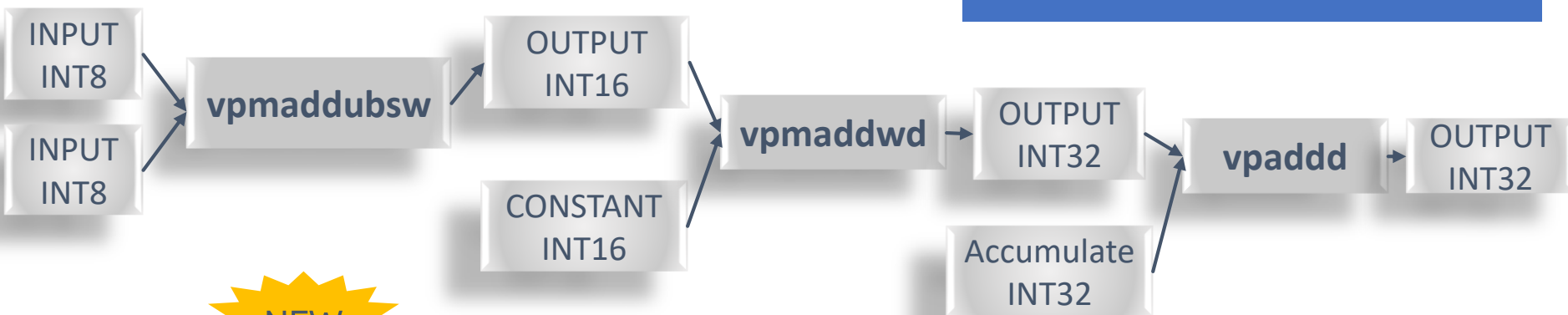
1st gen Intel® Xeon® Scalable processor without Intel® DL Boost

FP32



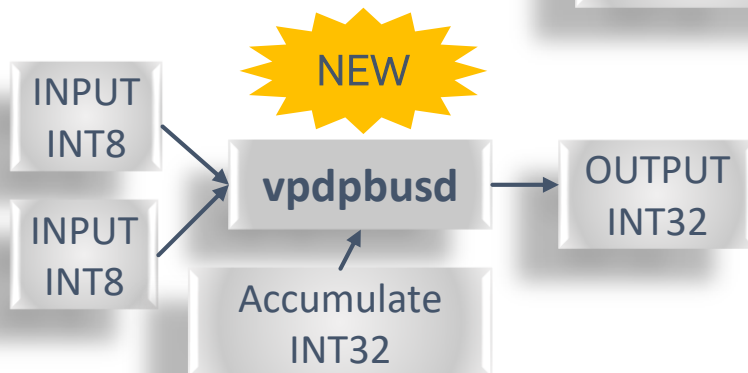
1st gen Intel® Xeon® Scalable processor without Intel® DL Boost

INT8



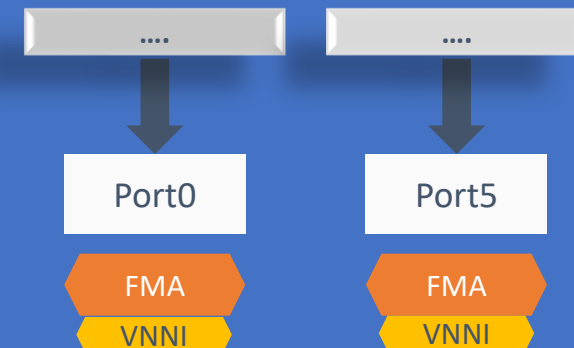
2nd gen Intel® Xeon® Scalable processor with Intel® DL Boost

INT8 VNNI



Microarchitecture view

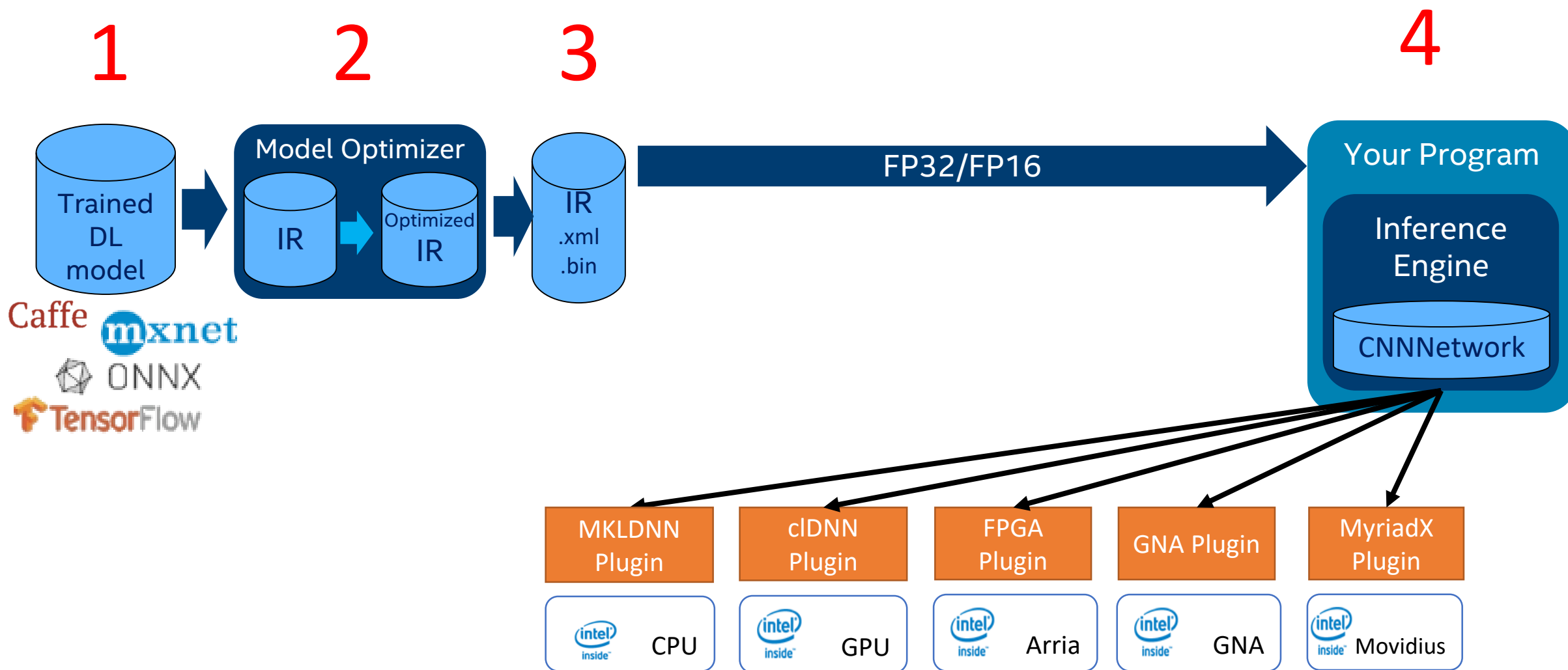
In a given clock cycle



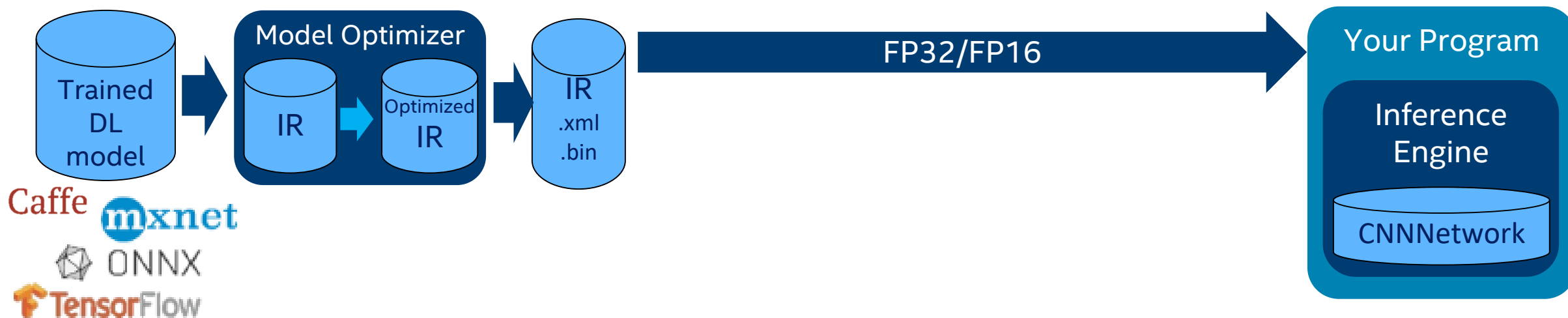
Here's one tool in your arsenal
to do it 😊

OpenVINO™

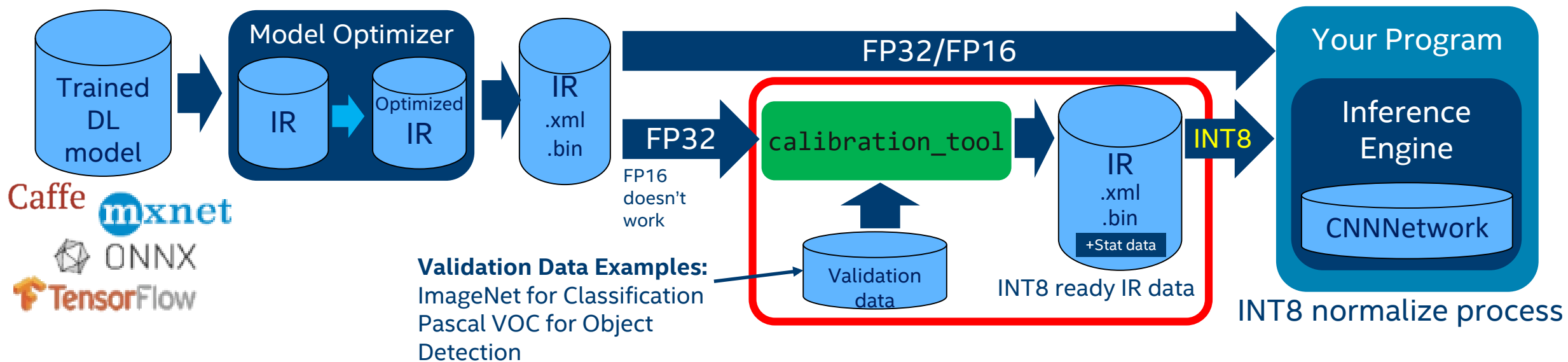
Intel® Distribution of OpenVINO™ in a nutshell



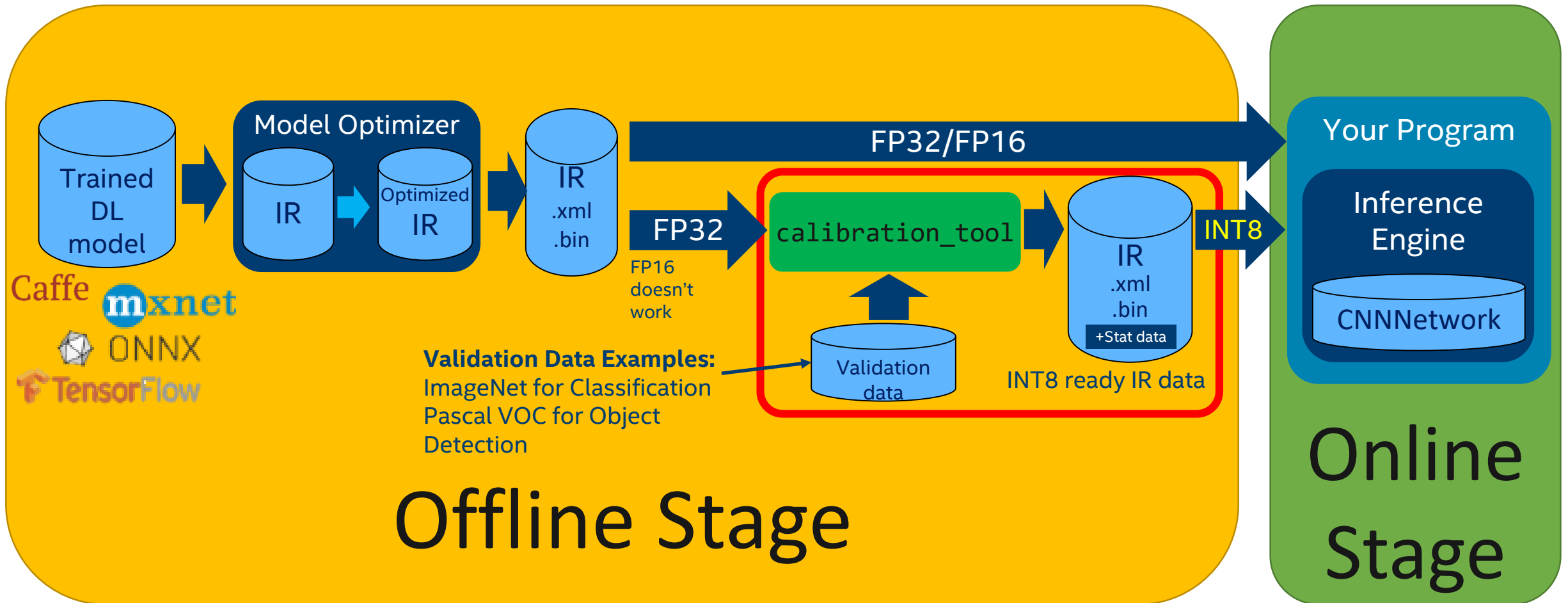
Intel® Distribution of OpenVINO™ in a nutshell



Intel® Distribution of OpenVINO™ in a nutshell



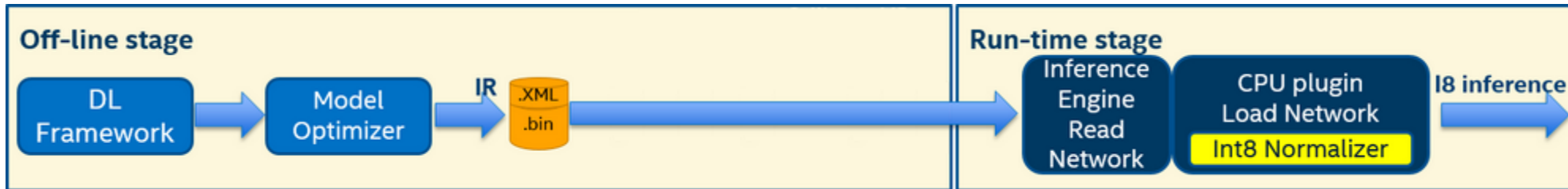
Intel® Distribution of OpenVINO™ in a nutshell



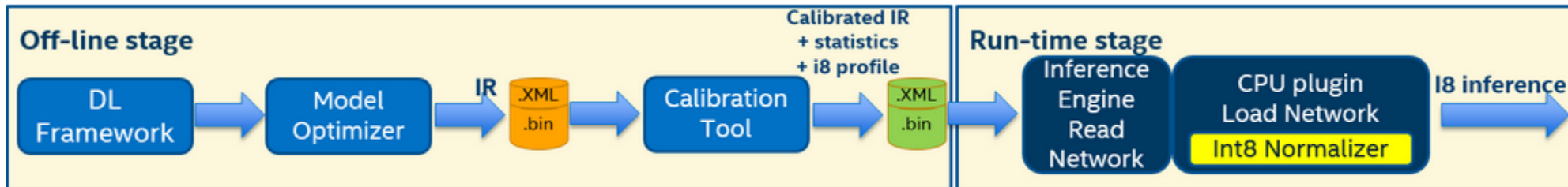
https://docs.openvinotoolkit.org/latest/docs_IE_DG_Int8Inference.html

Sample results

Demo 1



Demo 2



Both executions on Intel® Cascade Lake CPU



QUICK DEMO

►RS/0211 SEARCH...A01
►RS/0211 SEARCH...A01

►SEARCH►TR/01►03
►SEARCH►TR/01►03

010N ►TR/01►03
010N ►TR/01►03

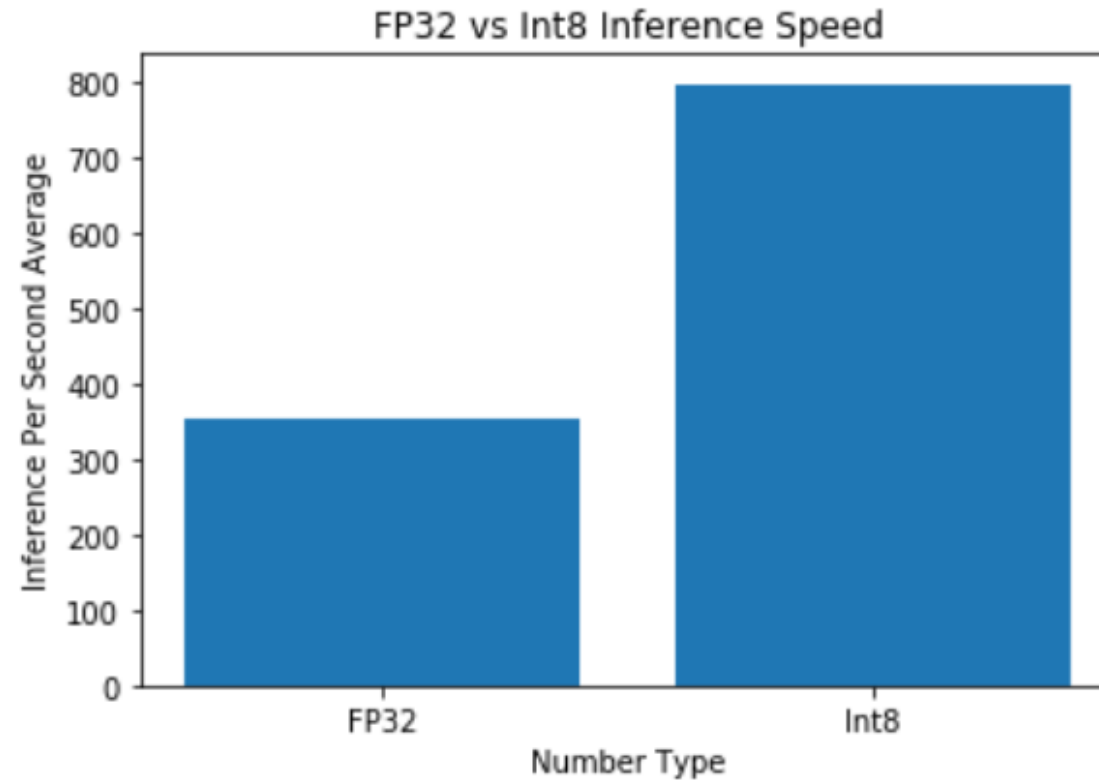


Sample results

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!

►RS/0211 SEARCH...A01
►RS/0211 SEARCH...A01

►SEARCH►TR/01►03
►SEARCH►TR/01►03

010N ►TR/01►03
010N ►TR/01►03

