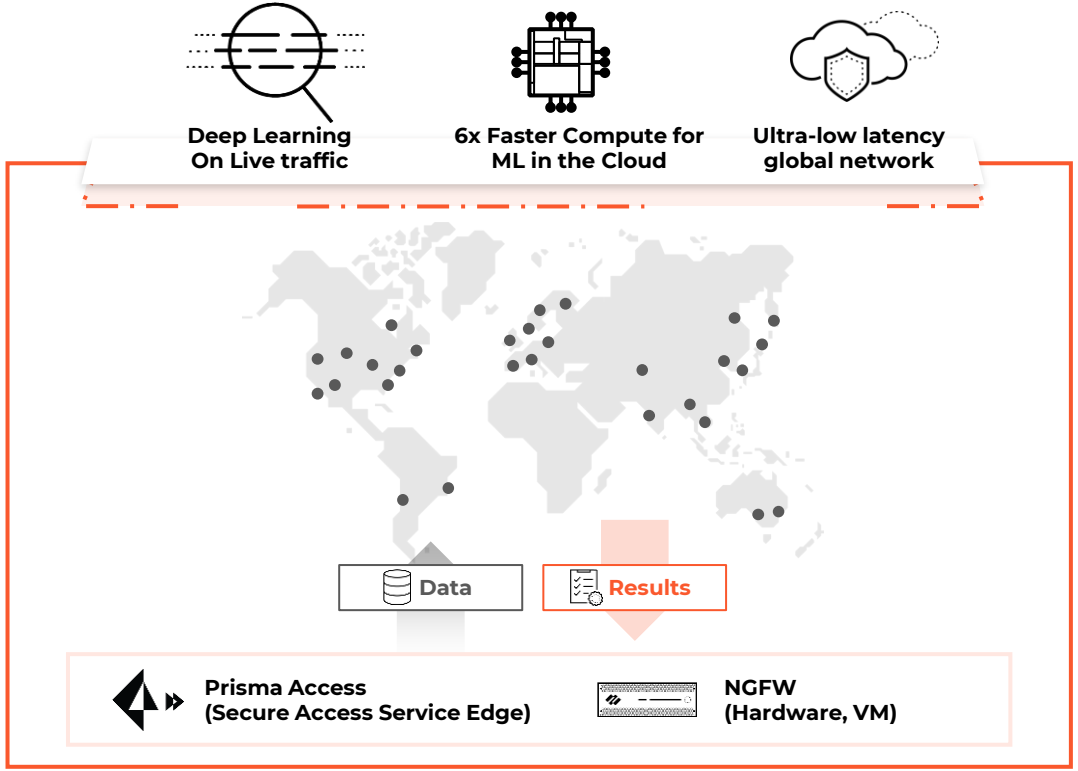
# Next-Gen Security Services with Inline AI Inference Powered by Intel® Xeon® Processor in the Cloud

Suiqiang Deng, Distinguished Engineer & Architect, Palo Alto Networks

David Lu, Platform Solution Architect, Intel

# Palo Alto Networks recently introduced ‘Inline Deep Learning’

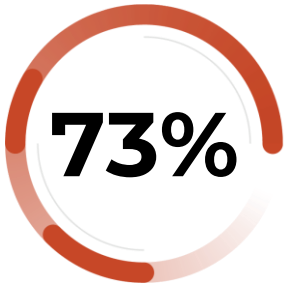
## Stops Evasive Threats. Inline.




- All-new cloud-based **deep learning engine** detects today's most challenging **unknown attacks**
- Brings **next-level prevention inline** to **defend the initial target**
- Ultra-low latency global network streams real data from **cloud-based** and **on-prem** network security deployments

# Stopping Today's C2 Attacks Harder than Ever

**Attackers employing C2 tools to evade signature defenses...**




YoY increase in attacks using Cobalt Strike for post exploitation

**....and encrypting their network communications**


Of malware campaigns use encryption to conceal C2 activity

**Use Public Cloud Infrastructure**

## Recent Examples

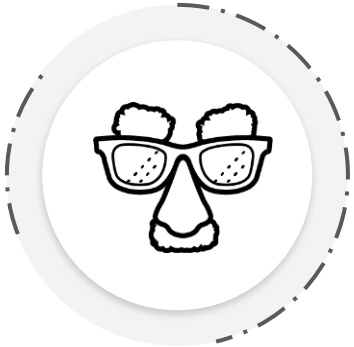


# **IPS is foundational for Network Security…** But only to prevent known threats

## Signatures prevent known threats

Traditional approaches struggle to keep pace with the changing threat landscape

## Cannot prevent the unknown



Malware communications must be observed first in the wild before protections can be released

## Evasive Command and Control on the rise

Threat actors are leveraging highly customizable tooling to evade traditional technologies

# New Advanced Threat Prevention - Industry's first IPS to stop unknown C2

## Best-In-Class prevention of known threats

Block known exploits, C2, and malware across all web and non-web traffic

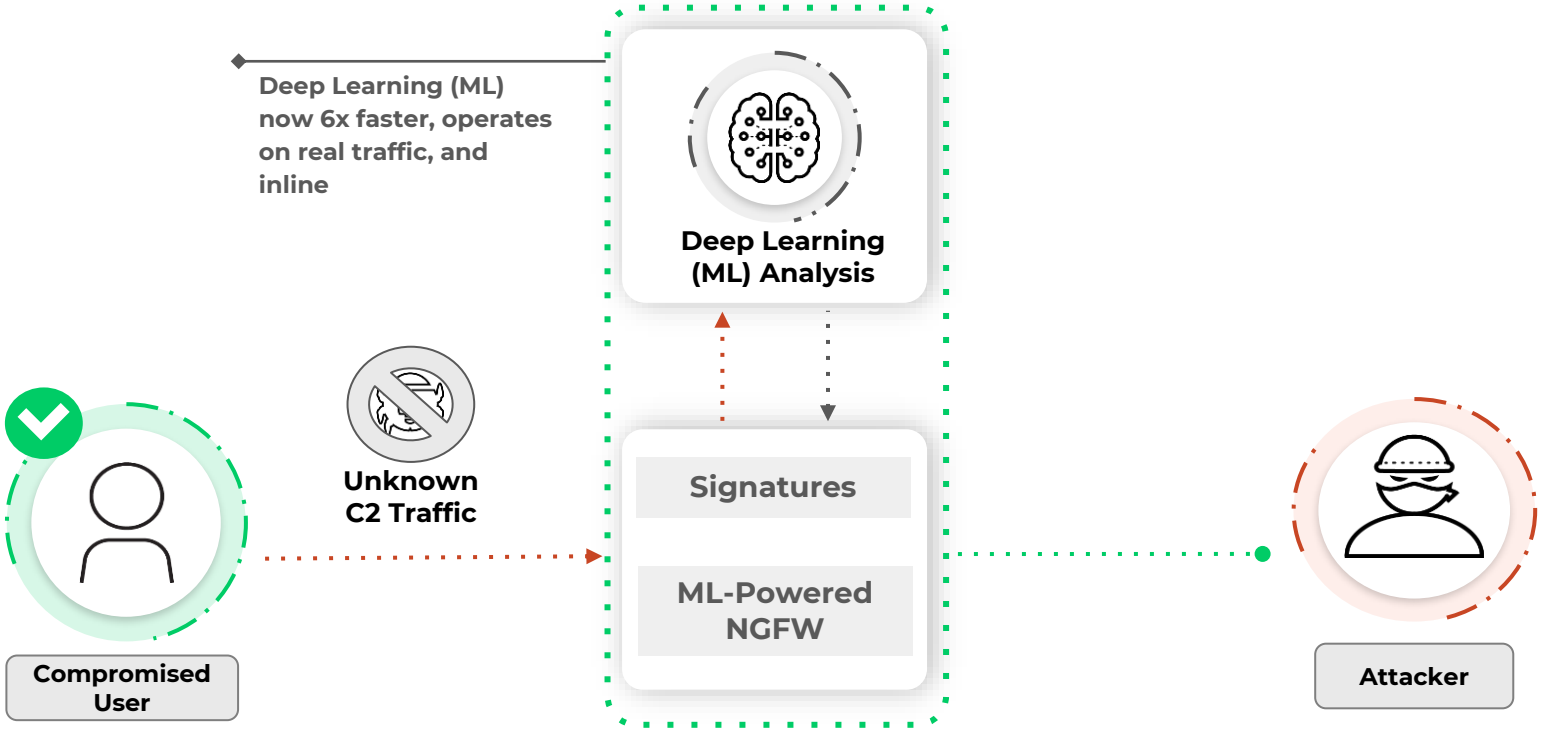
## Prevent Unknown Command-and-Control

Stop communications of highly evasive hack tools (Eg: Cobalt Strike)

## Cloud-Delivered Inline Deep Learning

Stay always up to date with industry first inline deep learning, and easily adopt new capabilities over time

# Advanced Threat Prevention: Utilize Inline Deep Learning for Prevention



# Stop 48% More Unknown C2

**Unknown C2 Detection Engine**

- Patent pending Deep Learning models
- Encrypted Traffic Classifier
- Anomaly Detector
- Malware Family Classifier
- Continuous Model Validation/Training Pipeline
- Prevention against C2 derived from hack tools (Eg: Cobalt Strike)

+ all the leading capabilities from Threat Prevention


**48%**

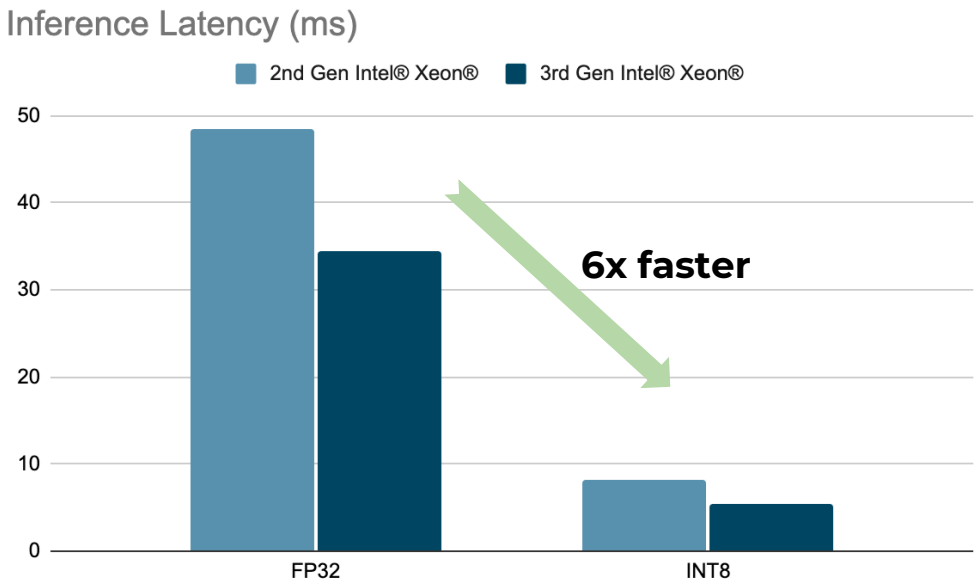
**Increase in C2 threat detection** compared to industry's leading Threat Prevention solution

**96%**

Web based Cobalt Strike C2 communications **blocked inline**

**Supported on PAN-OS 10.2 (Nebula)**

# Results for AI inference latency improvement:



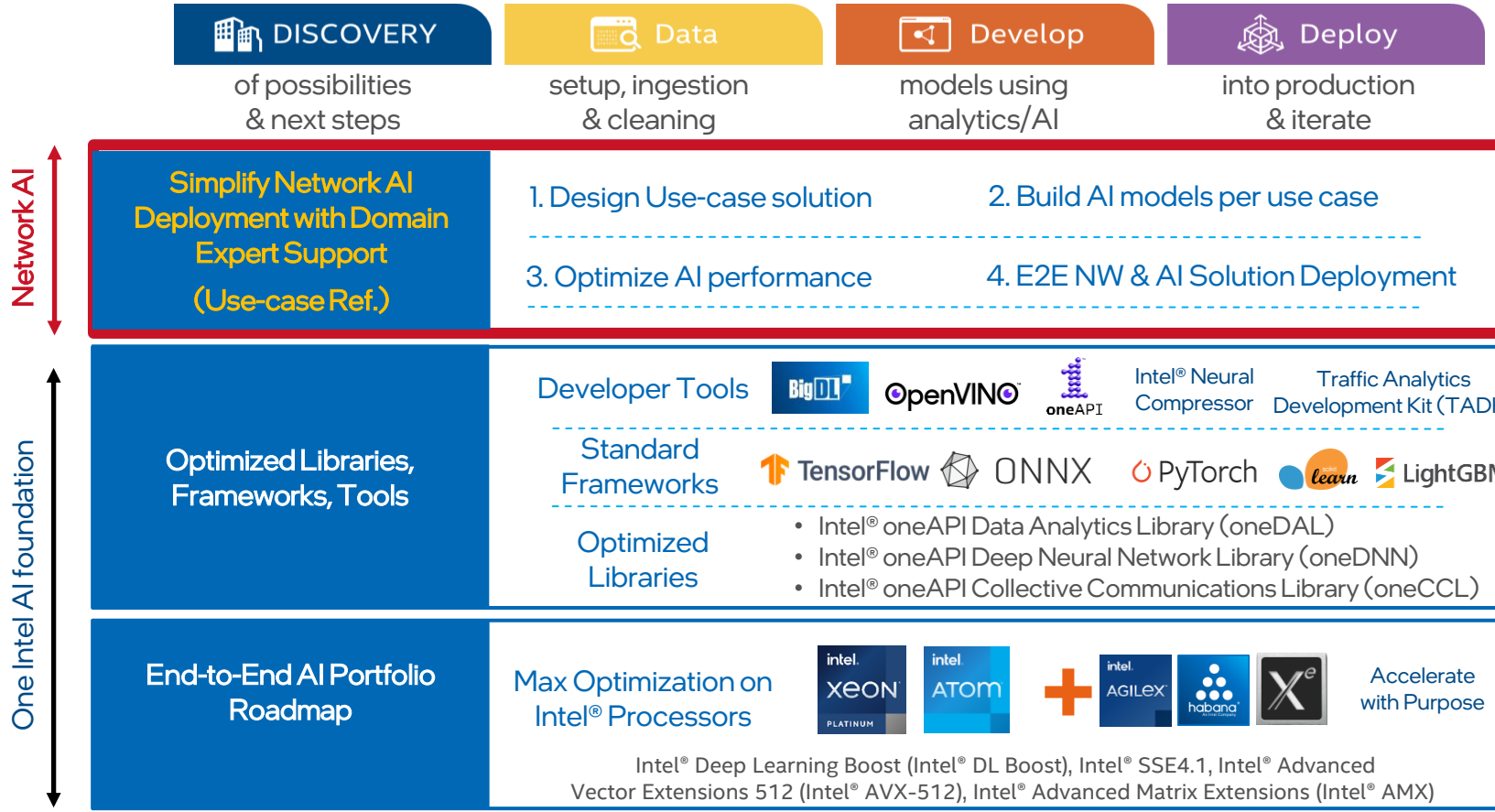

- Tests were done on GCP N2 instances with 8-VCPU 2nd Gen Intel Xeon or 3rd Gen Intel Xeon CPU
- INT8 model is ~6x faster than original SavedModel
- 3rd Gen Intel Xeon is ~30% faster than 2nd Gen Intel Xeon

# Intel Network AI Optimization

# Notices and Disclaimers

- Intel technologies may require enabled hardware, software or service activation.
- No product or component can be absolutely secure.
- Your costs and results may vary.
- © Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

# Intel® Network AI Offering



# Intel® Xeon® Scalable Processors

## DATA CENTER CPU OPTIMIZED FOR AI

INTEL® ADVANCED VECTOR EXTENSIONS 512 (INTEL® AVX-512)
INTEL® DEEP LEARNING BOOST (INTEL® DL BOOST)
INTEL® OPTANE™ DC PERSISTENT MEMORY

| Intel® DL Boost Technologies | | | |
|---|---|---|---|
| Microarchitecture | AVX512_VNNI | AVX512_BF16 | AMX |
| Client | | | |
| Core 10th Gen | ✓ | X | X |
| Server | | | |
| Xeon SP Gen 2 | ✓ | X | X |
| Xeon SP Gen 3H | ✓ | ✓ | X |
| Xeon SP Gen 3 | ✓ | X | X |
| Next Gen Xeon SP | ✓ | ✓ | ✓ |

**2019**

### Gen 2 Xeon SP
Intel® Deep Learning Boost (Intro)
Intel® AVX-512 (VNNI/INT8)

**2020**

### Gen 3H Xeon SP
Intel® Deep Learning Boost
Intel® AVX-512 (VNNI/INT8 & BFloat16)

**2021**

### Gen 3 Xeon SP
Intel® Deep Learning Boost
Intel® AVX-512 (VNNI/INT8)

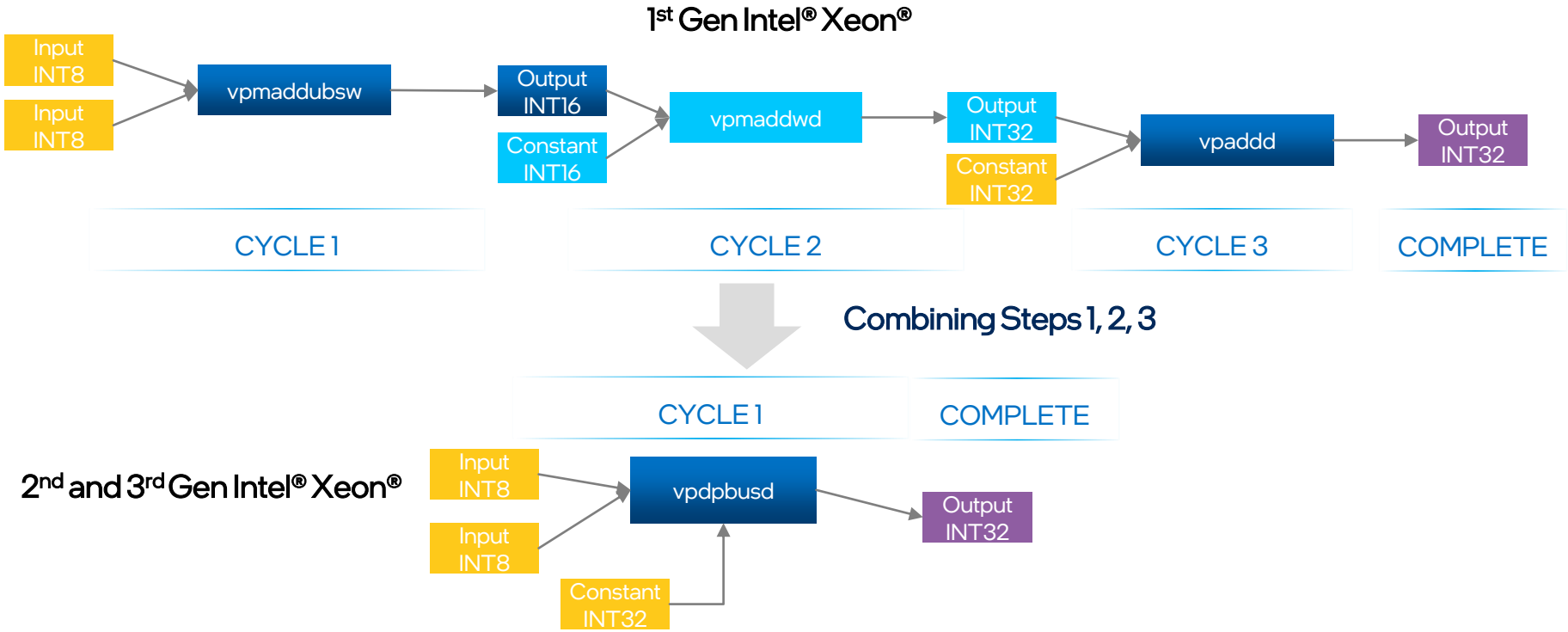
**2022**

### Next Gen Xeon SP
Intel® AMX – INT8 and BFloat16 support
Intel® AVX-512 (VNNI/INT8)

## LEADERSHIP PERFORMANCE

# Intel® Deep Learning Boost

## A Vector Neural Network Instruction (VNNI) Extends Intel® AVX-512 to accelerate AI/DL Inference



# Intel® oneAPI Deep Neural Network Library (oneDNN)

## Features

- Supports FP32, FP16, Bfloat16, and int8.
- Leverages Intel® DL Boost, Intel® AVX-512 instructions, and processor capabilities
- Fused operations for optimized performance

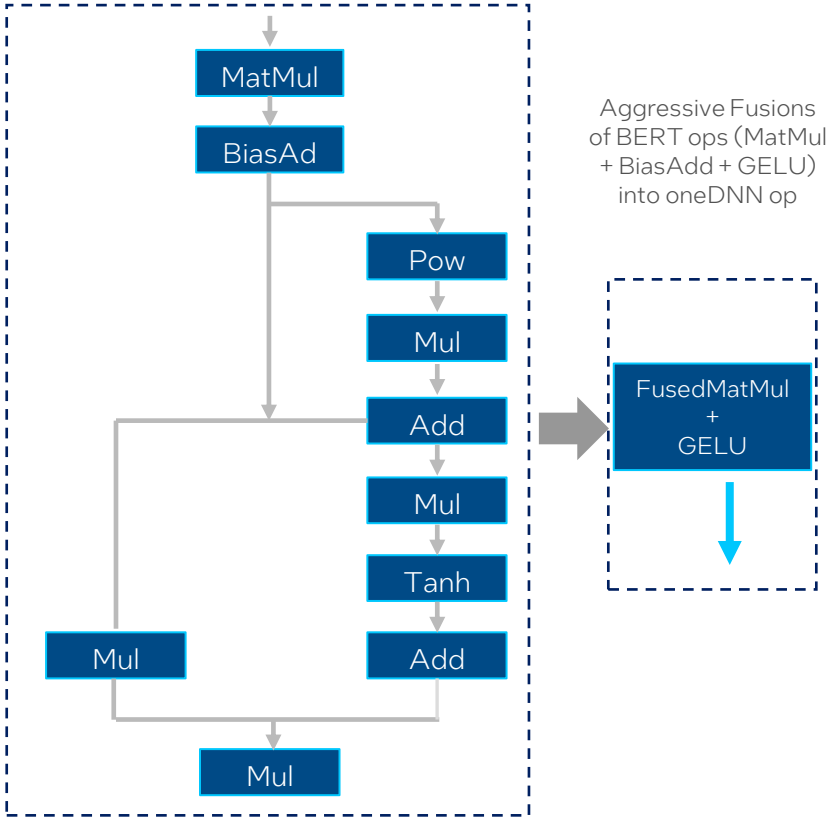
## Support Matrix

- Compilers: Intel® oneAPI DPC++ / C++ Compilers
- OS: Linux, Windows, macOS
- CPU: Intel Atom®, Intel® Core™, Intel® Xeon®, Intel® Xeon® Scalable processors
- GPU: Intel® Processor Graphics Gen9, Intel® Processor Graphics Gen 12

| Category | Functions |
|---|---|
| Compute intensive operations | • (De-)Convolution<br>• Inner Product<br>• RNN (Vanilla, LSTM, GRU)<br>• GEMM |
| Memory bandwidth limited operations | • Pooling<br>• Batch Normalization<br>• Local Response Normalization<br>• Layer Normalization<br>• Elementwise<br>• Binary elementwise<br>• Softmax<br>• Sum<br>• Concat<br>• Shuffle |
| Data manipulation | • Reorder |

# Intel® oneDNN Integration with TensorFlow

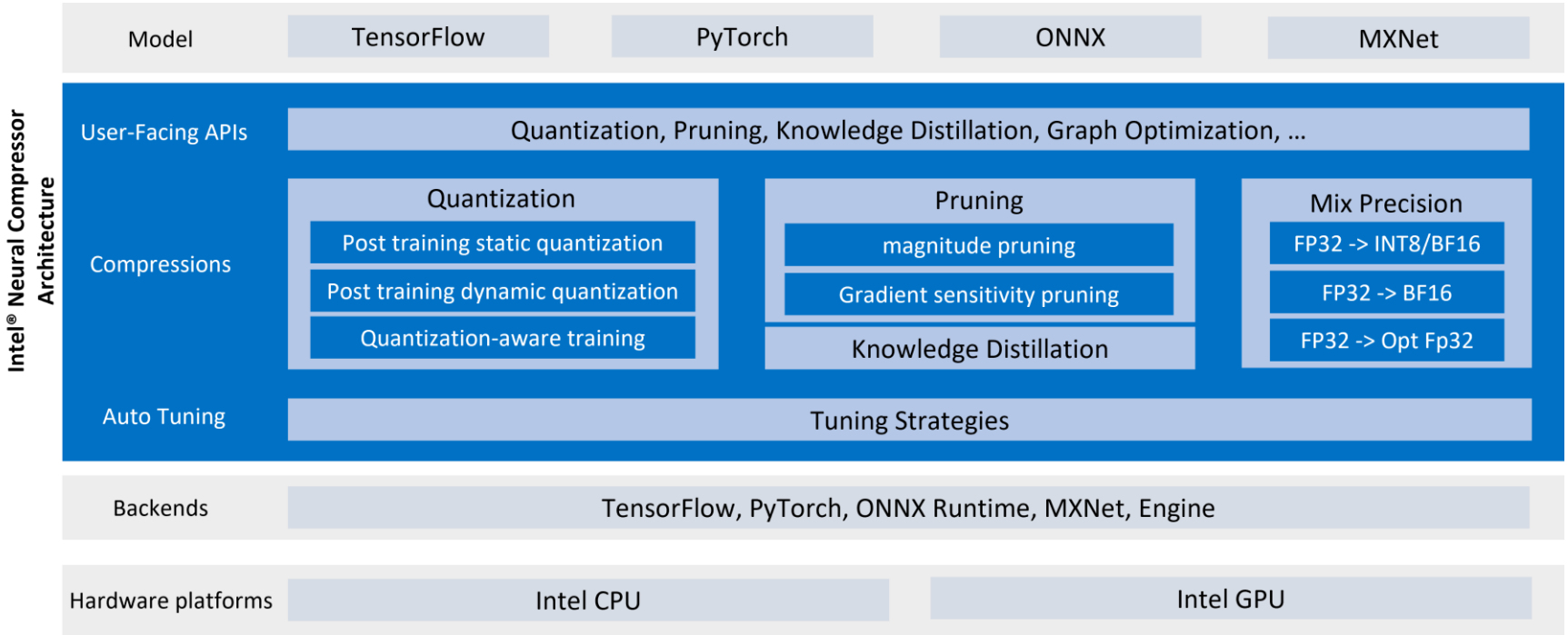
- Replaces compute-intensive standard TF ops with highly optimized custom oneDNN ops
- Aggressive op fusions to improve performance of Convolutions and Matrix Multiplications
- Primitive caching to reduce the overhead of calling oneDNN
- Turn on oneDNN optimizations at runtime in official TensorFlow distributions by setting an environment variable TF_ENABLE_ONEDNN_OPTS=1




https://github.com/tensorflow/community/pull/400
https://medium.com/intel-analytics-software/leverage-intel-deep-learning-optimizations-in-tensorflow-129faa80ee07

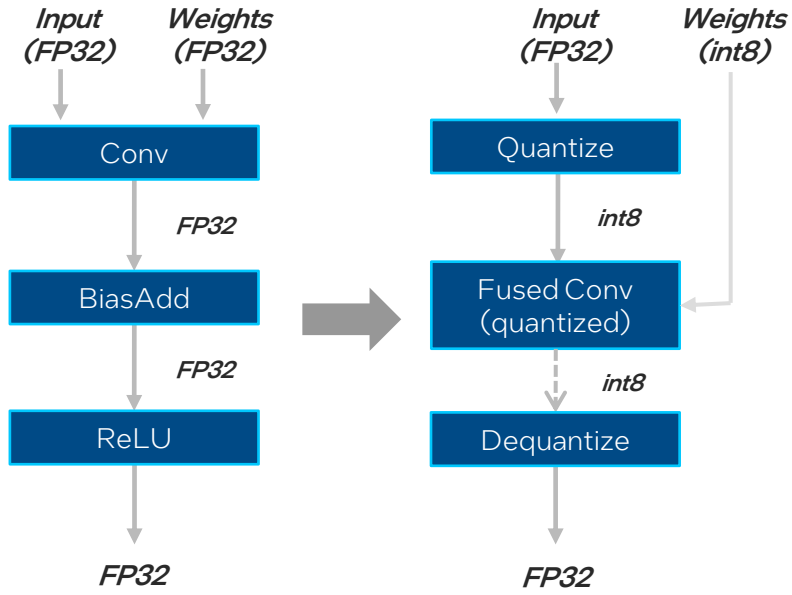
# Intel® Neural Compressor Infrastructure

Opensource Tool for Quantization (https://github.com/intel/neural-compressor)



# Low Precision (8-bit Integer) Inference Optimization

- Quantized models using 8-bit integers gaining adoption
  - Improved performance
  - Trade off accuracy for performance
- Intel® Neural Compressor*
  - Automatically quantizes pre-trained model
  - Additional post-training quantization steps needed
  - Picks quantization scheme to meet specific performance and accuracy needs
- Accelerated by Intel® DL Boost instructions or Intel® AMX




Quantization Process

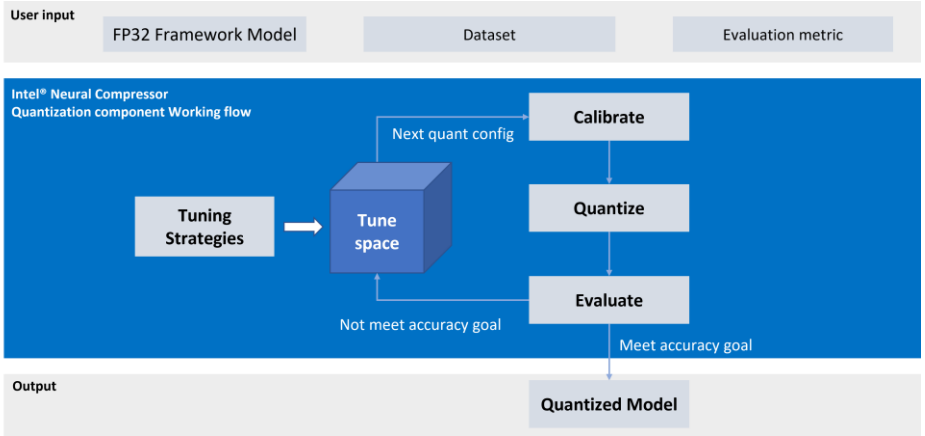
*Formerly Low Precision Optimization Tool (LPOT)

# Examples: Quantize TensorFlow RN50

```
# RN50.yaml
model:
 name: resnet50
 framework: tensorflow

quantization:
 calibration:
  dataloader:
    dataset:
      ImageRecord:
        root: /path/to/calibration/dataset
    transform:
      ResizeCropImagenet:
        height: 224
        width: 224
        mean_value: [123.68, 116.78, 103.94]

evaluation:
 accuracy:
   metric:
      topk: 1
```




```
# tune.py
from neural_compressor.experimental import Quantization, common

quantize = Quantization('./RN50.yaml')
quantize.model = common.Model(self.args.input_graph)
q_model = quantize()
q_model.save(output_model_path)
```

# References

Blog on using ML to detect C2 traffic: https://unit42.paloaltonetworks.com/c2-traffic/

How to improve the performance with Intel® oneDNN and Intel® Neural Compressor under TensorFlow: https://networkbuilders.intel.com/solutionslibrary/intel-deep-learning-boost-boost-network-security-ai-inference-performance-in-google-cloud-platform-gcp-technology-guide

AI Technologies – Unleash AI Innovation in Network Applications: https://networkbuilders.intel.com/solutionslibrary/ai-technologies-unleash-ai-innovation-in-network-applications-solution-brief

# Thank you