Optimizing Deep Learning models

Theory, tools & best-practices

Deltatre Innovation Lab

Clemente Giorio
R&D Senior SW Engineer
deltatre

Gianni Rosa Gallina
R&D Technical Lead
deltatre

AI Day

Platinum Sponsor

Microsoft

avanade

ellycode
WHAT DO YOU NEED?

PORINI
A DGS COMPANY

Gold Sponsor

LOBRA

UNIKEY
Bringing IT knowledge to the people

Technical Sponsor

NET CODE

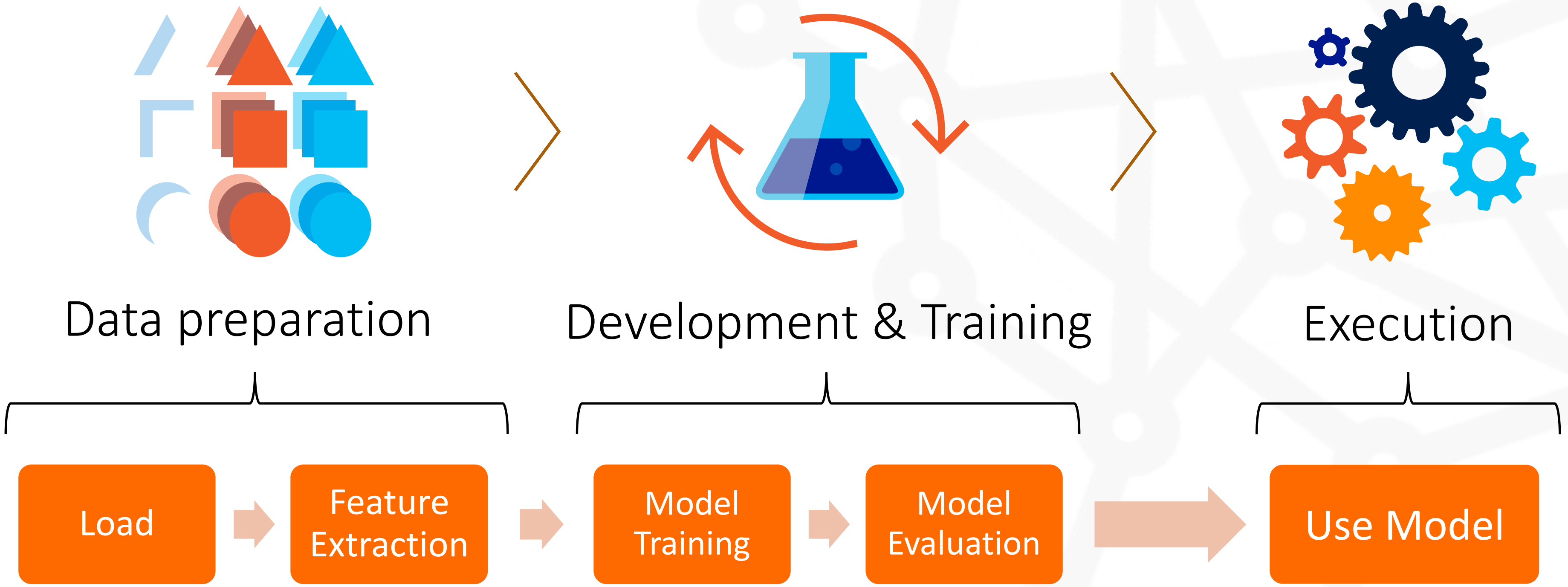
JET BRAINS

Packt

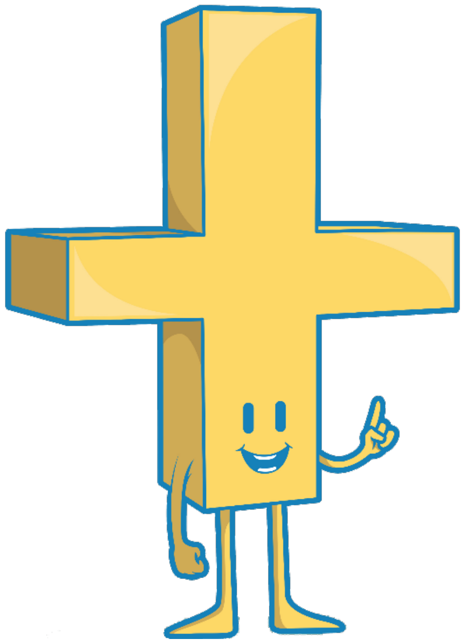
stickermule

Disclaimer

Model lifecycle



Advantages & Disadvantages



Size reduction

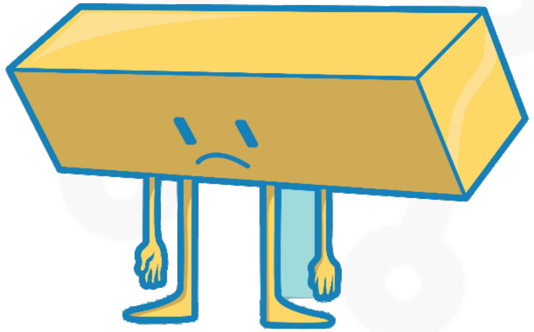
Speed improvements

Energy/Costs savings

Additional steps

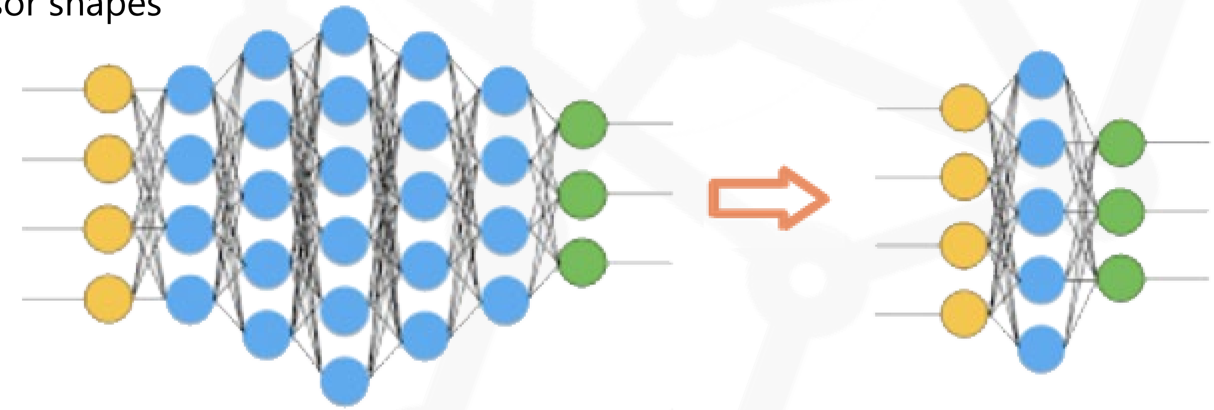
Evaluation metrics drops

Not easy



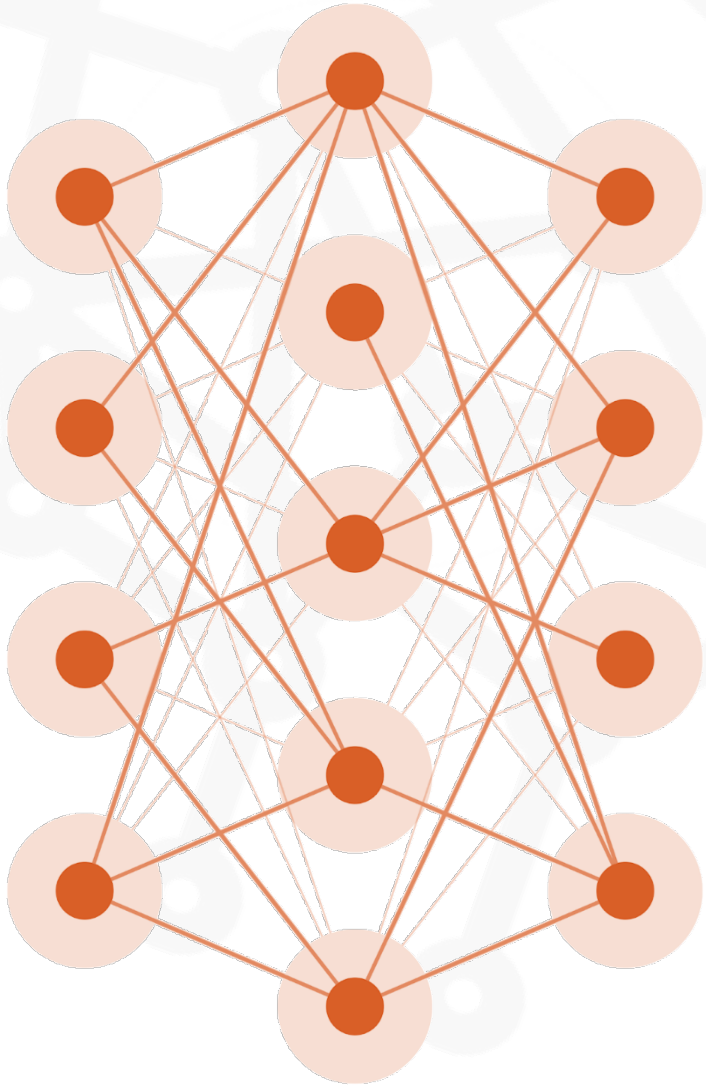
Self Organizing Tree Algorithm & Techniques

- **Unstructured/semi-structured pruning**
  - Remove individual (or groups) weights by masking to 0
  - Algorithms: Magnitude (0^{th} order), Movement (1^{st} order), WoodFisher (2^{nd} order)
- **Structured pruning**
  - Remove large sections of weights by changing tensor shapes
  - Channel, Filter, Layer, Attention head
- **Quantization**
  - Reduce the precision of activations and weights
  - INT8 vs. < INT8; dynamic vs static
- **Distillation**
  - Distil information from a lager, teacher into a student model
  - Response, feature, relation



Applying Optimization to a Model

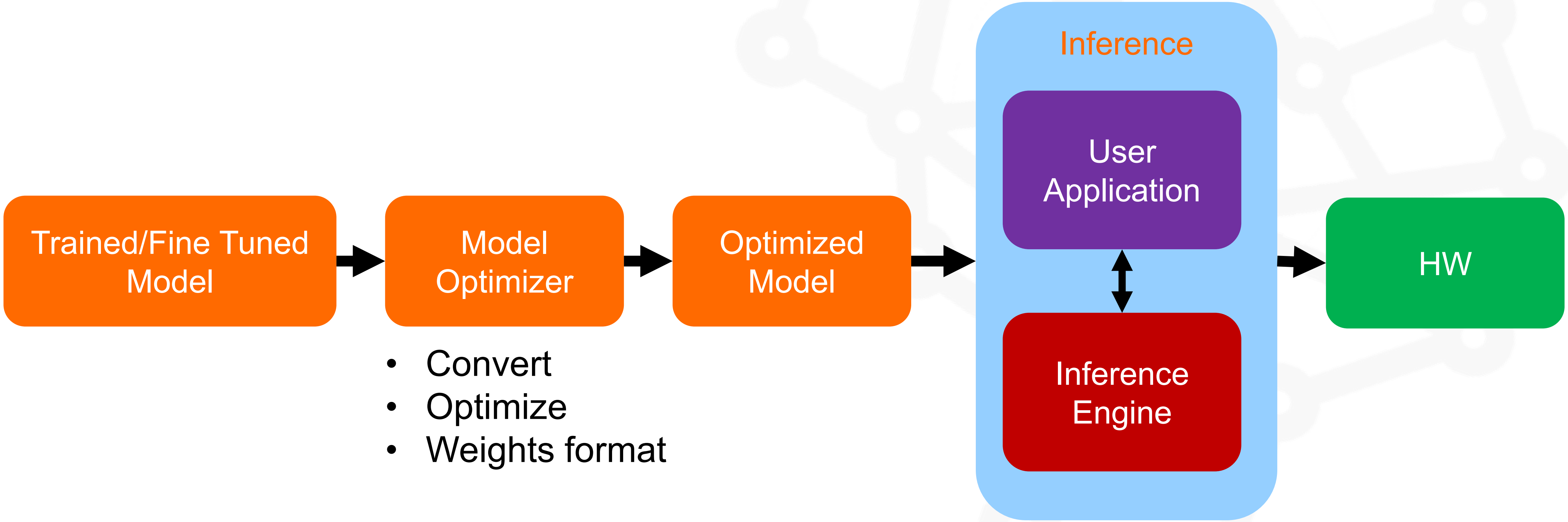
- Post training (One Shot)
  - Applied after training using the model and sample data
  - Works well for dynamic quantization
  - Currently does not work well for pruning
- Training-Aware
  - Optimization are applied during training
  - Works well for pruning and quantization
- Sparse Transfer
  - Fine-tune a pre-sparsified model onto a new dataset



Tools

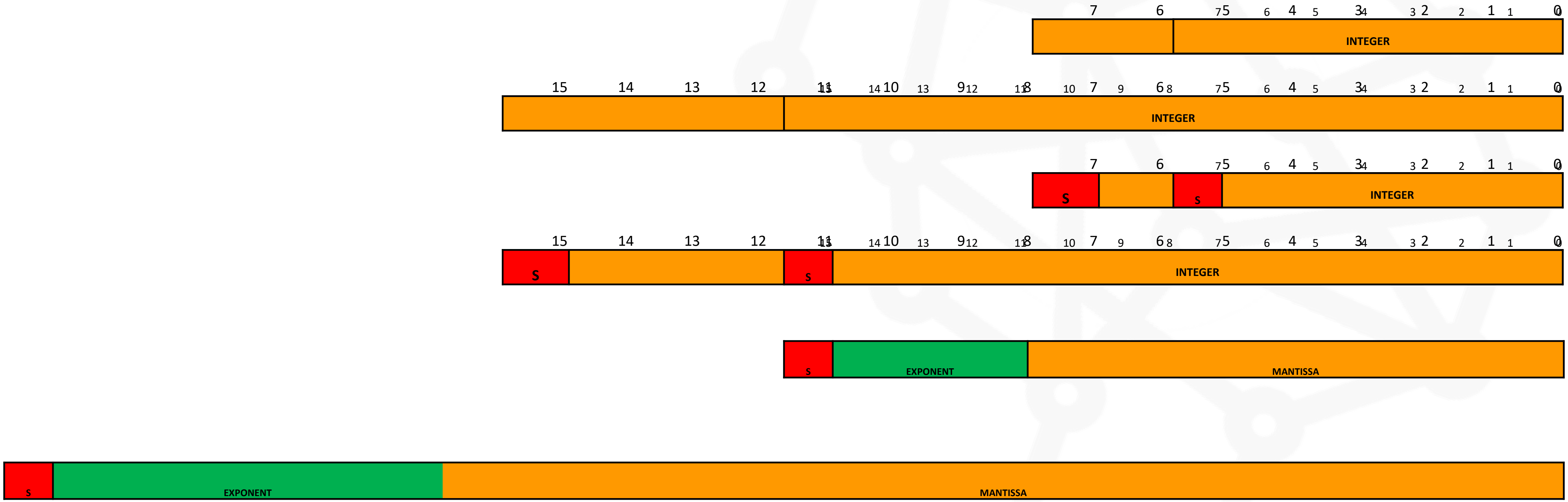
- PyTorch and TensorFlow built-in
  - Limited algorithms support
  - Optimizations defined in code
- NVIDIA TLT/TensorRT and Intel NNCF
  - Good one-shot support
  - Limited integration and training capabilities
- Research Libraries
  - Good single algorithm support
  - Limited integration and multi-model support
- Nebuly, OctoML, Deci
  - Wrap the above tools with some additional algorithm support

Optimizer Flow

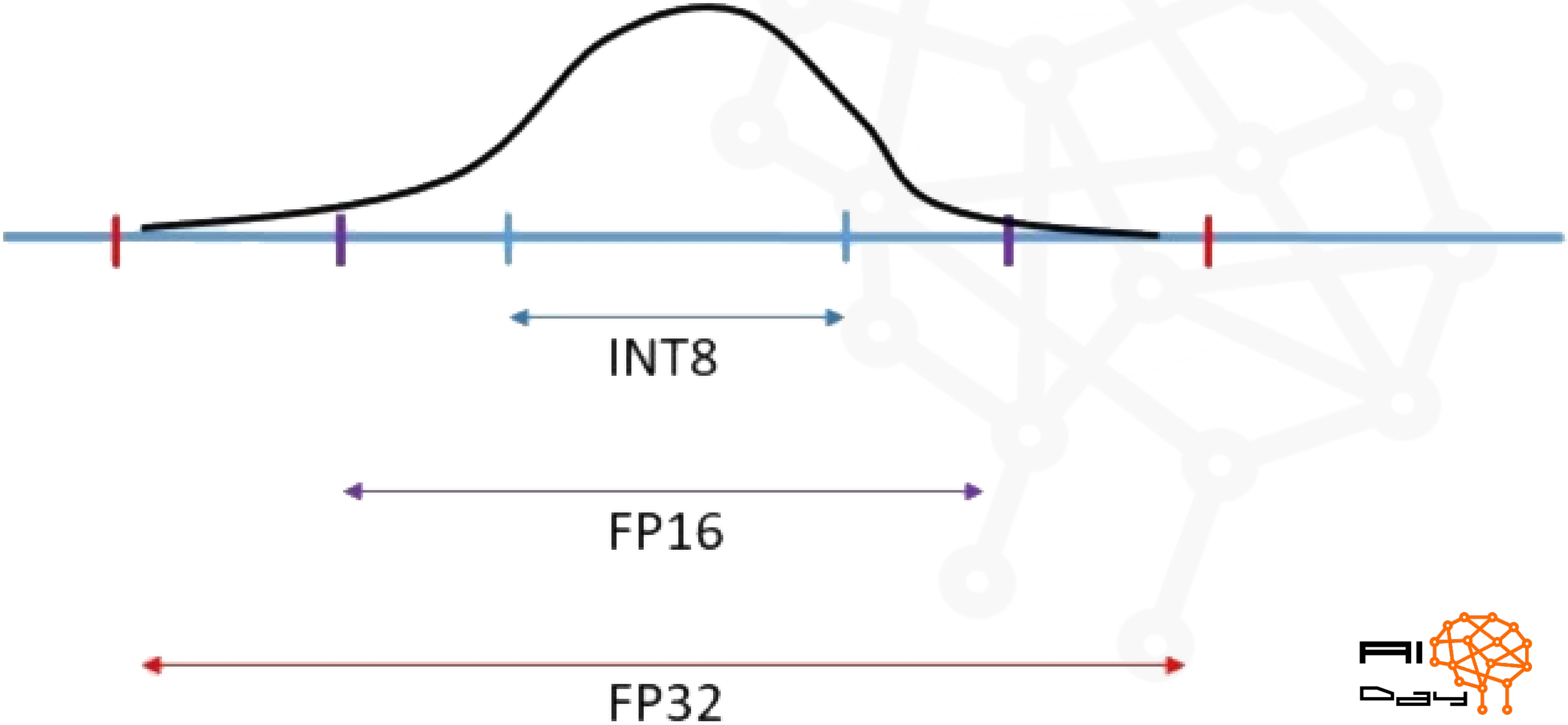


- Convert
- Optimize
- Weights format

Choosing the Right Precision



Choosing the Right Precision



Quantization

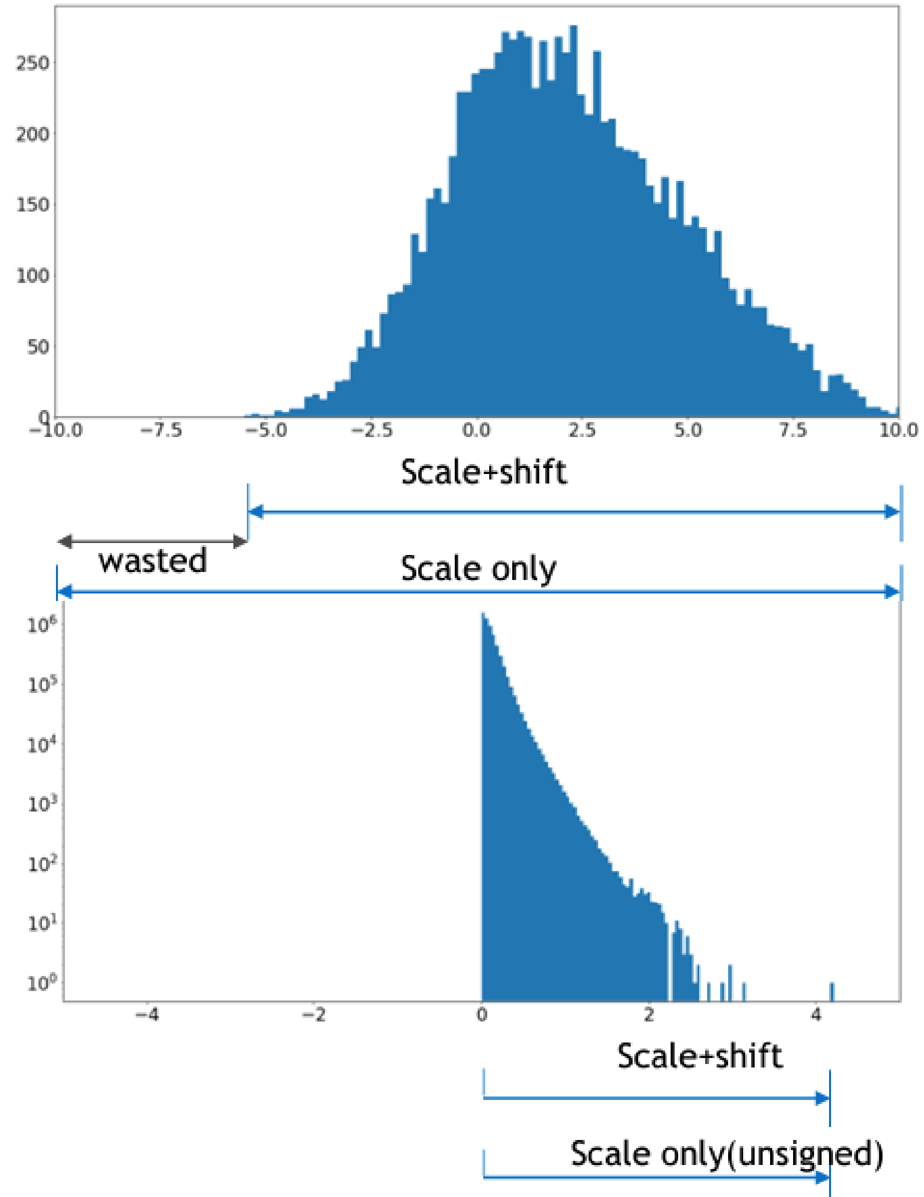


Image Classification, top-1 accuracy

	FP32	Int8 Scale	Int8 Scale+Shift
Mobilenet-v1_1_224	70.90	70.70	70.00
Mobilenet-v2_1_224	71.90	71.10	70.90
Nasnet-Mobile	74.00	73.00	73.00
Mobilenet-v2_1.4_224	74.90	74.50	73.50
Inception-v3	78.00	78.00	78.00
Resnet-v1_50	75.20	75.00	75.00
Resnet-v2_50	75.60	75.00	75.00
Resnet-v1_152	76.80	76.20	76.50

Object Detection, mAP

	FP32	Int8 Scale	Int8 Scale+Shift
faster_rcnn_resnet101_coco*	0.38	0.37	0.38
faster_rcnn_nas_coco*	0.56	0.55	0.55
faster_rcnn_inception_v2_coco	0.28	0.28	0.279

Source: https://arxiv.org/abs/1806.08342

Quantization: Math Throughput

Relative to fp32 math

Input Type	Accumulation Type	Relative math throughput	Bandwidth savings
FP16	FP16	8x	2x
INT8	INT32	16x	4x
INT4	INT32	32x	8x
INT1	INT32	128x	32x

Source: Nvidia.com

Inference Speedup over FP32

Input size 224x224 for all, except 299x299 for Inception networks

	Batch size 1			Batch size 8			Batch size 128		
	FP32	FP16	Int8	FP32	FP16	Int8	FP32	FP16	Int8
MobileNet v1	1	1.91	2.49	1	3.03	5.50	1	3.03	6.21
MobileNet v2	1	1.50	1.90	1	2.34	3.98	1	2.33	4.58
ResNet50 (v1.5)	1	2.07	3.52	1	4.09	7.25	1	4.27	7.95
VGG-16	1	2.63	2.71	1	4.14	6.44	1	3.88	8.00
VGG-19	1	2.88	3.09	1	4.25	6.95	1	4.01	8.30
Inception v3	1	2.38	3.95	1	3.76	6.36	1	3.91	6.65
Inception v4	1	2.99	4.42	1	4.44	7.05	1	4.59	7.20
ResNext101	1	2.49	3.55	1	3.58	6.26	1	3.85	7.39

Tested with TensorRT on Tesla T4 GPU

Source: Nvidia.com

Quantized Inference

- Quantization:
  - Using lower precision to represent weights and activations
  - Using lower precision math

- Benefits:
  - Speed up inference:
    - Math limited layers due to high throughput math
    - Memory limited layers due to bandwidth saving
  - Reduce resource requirements: memory footprint, etc.

- Challenge:
  - Maintaining model accuracy

Supported Model Formats

Plugin	FP32	FP16
CPU	Supported and prefered	Not supported
GPU	Supported	Supported and preferred
FPGA	Supported	Supported
VPU	Not supported	Supported
GNA	Supported	Not supported

Source: https://docs.openvino.ai

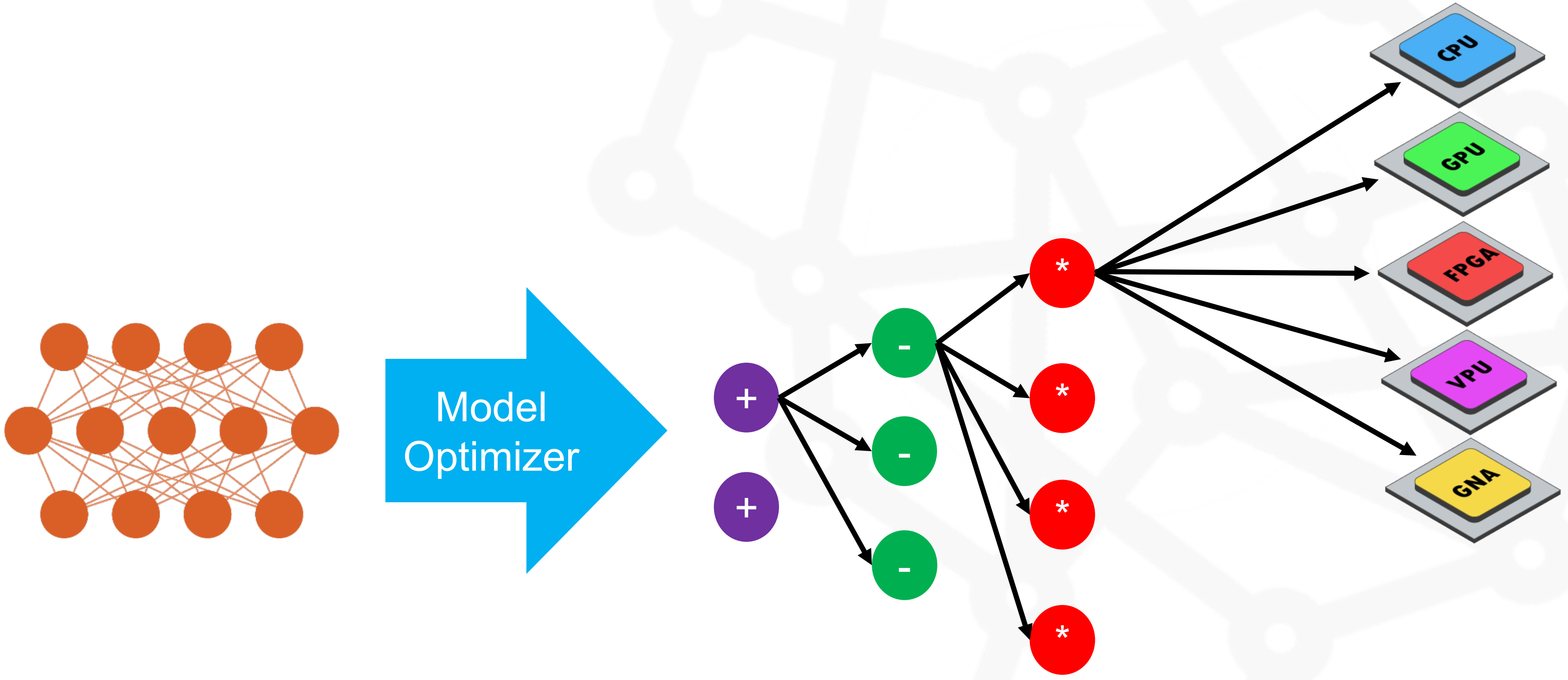
Supported Input Precision

Plugin	FP32	FP16	U8	U16	I8	I16
CPU	Supported	Not Supported	Supported	Supported	Not Supported	Supported
GPU	Supported	Supported	Supported	Supported	Not Supported	Supported
FPGA	Supported	Supported	Supported	Supported	Not Supported	Supported
VPU	Supported	Supported	Supported	Not Supported	Not Supported	Not Supported
GNA	Supported	Not Supported	Not Supported	Not Supported	Supported	Supported

Source: https://docs.openvino.ai

Inference Optimization

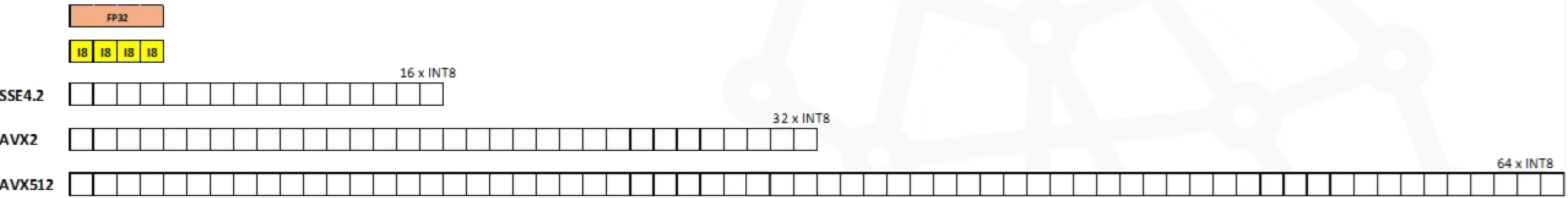
Original model



- Network level optimizations
- Memory level optimizations
- Kernel level optimizations

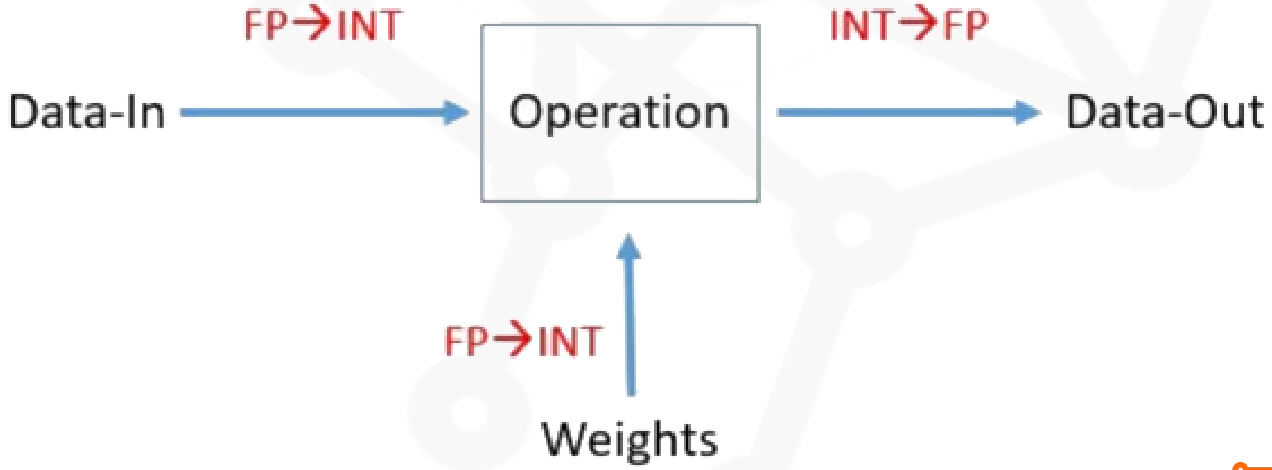
AI Day

INT8 and VNNI



Intel CPUs supporting
- SSE4.2
- AVX2
- AVX512

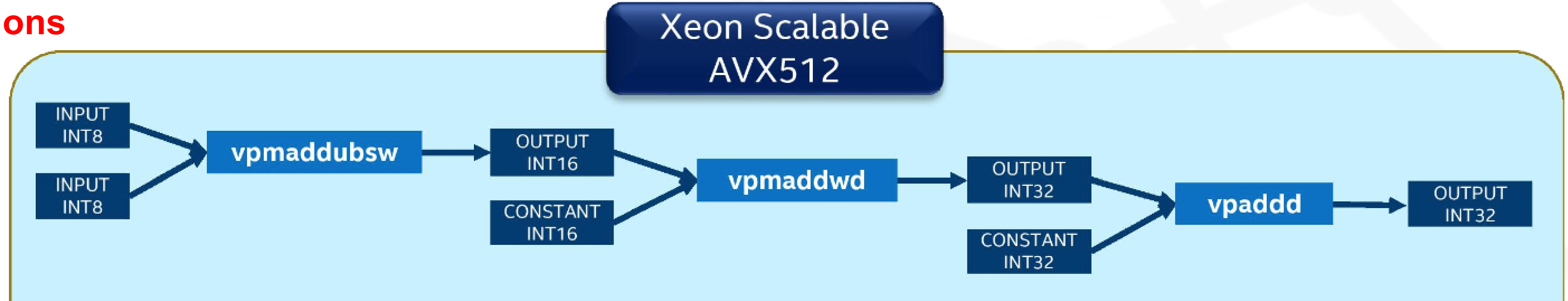
Supported INT8 operations (R1 2019):
- Convolution
- ReLU
- Pooling
- Eltwise
- Concat



https://www.intel.it/content/www/it/it/architecture-and-technology/avx-512-overview.html

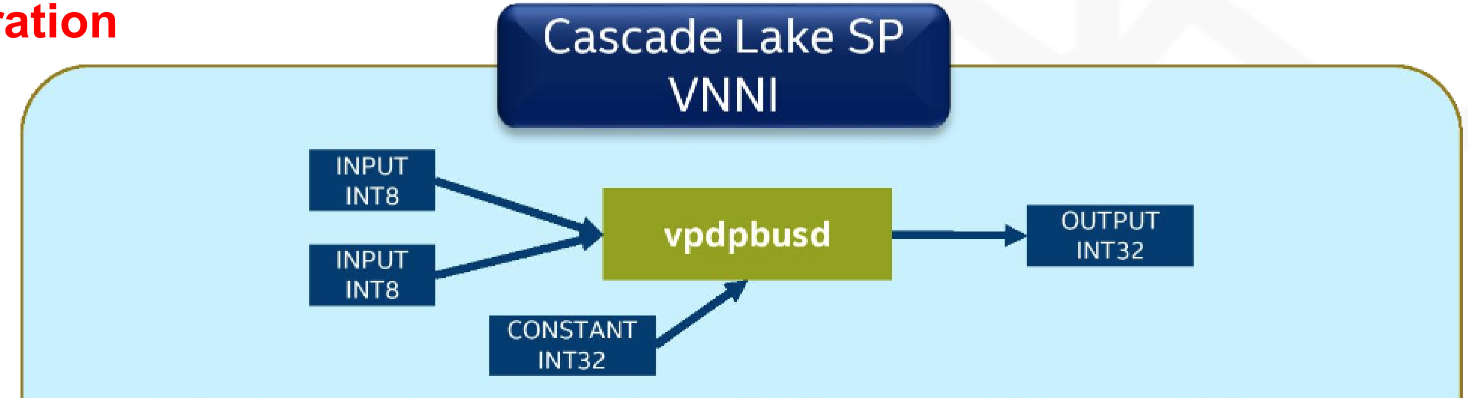
INT8 and VNNI

3 operations



Current AVX-512 instructions to perform INT8 convolutions: **vpmaddubsw, vpmaddwd, vpaddd**

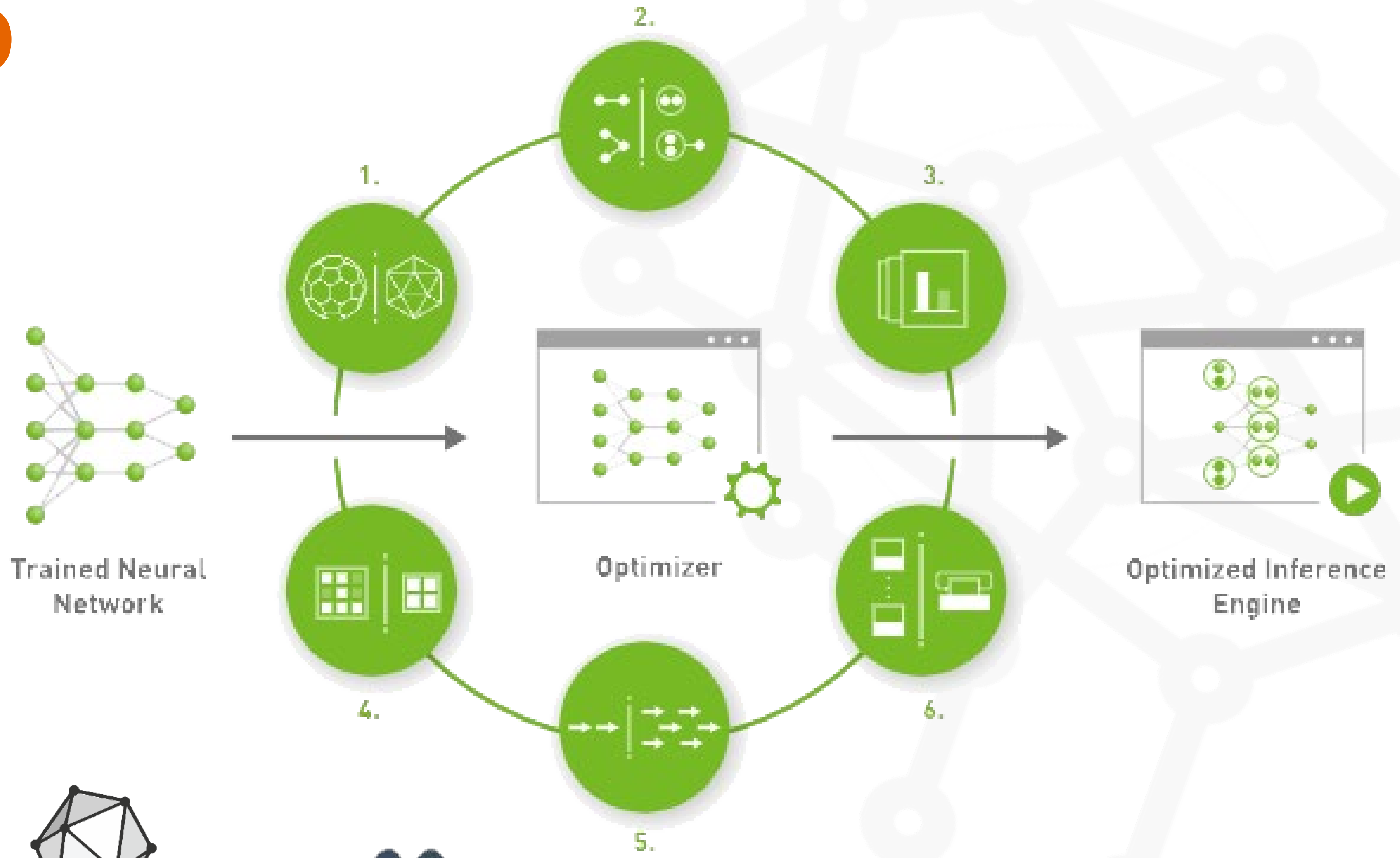
1 operation



VNNI instruction to accelerate INT8 convolutions: **vpdpbusd**

https://www.intel.com/content/www/us/en/artificial-intelligence/deep-learning-boost.html

DEMO



Diaram Image source from Nvidia.com

Take aways

- We can improve the performance of our AI models
- We can reduce costs and optimize resources

!! IT'S NOT EASY !!

- Dedicate proper time and team to optimization
- Requires specific knowledge & skills

Thank You!

εuχαριστώ Salamat Po متشكرم شكراً Grazie

благодаря ありがとうございます Kiitos Teşekkürler 谢谢

ขอบคุณครับ Obrigado شكريه Terima Kasih Dziękuję

Hvala Köszönöm Tak Dank u wel дякую Tack

Mulţumesc спасибо Danke Cảm ơn Gracias

多謝晒 Ďakujem תודה நன்றி Děkuji 감사합니다

Q&A

AI Day

Microsoft

avanade

ellycode
WHAT DO YOU NEED?

PORINI
A DGS COMPANY

LOOBRA

UNIKEY
Bringing IT knowledge to the people

NET CODE

JET BRAINS

Packt>

stickermule

Slides/Demo repository

Deltatre Innovation Lab

DOWNLOAD ME

GitHub

https://github.com/deltatrelabs/deltatre-aiday-2022-demo

About us

Clemente GIORIO @tinux80

R&D Senior Software Engineer @ deltatre

- Augmented/Mixed/Virtual Reality
- Artificial Intelligence, Machine Learning, Deep Learning
- Internet of Things
- Hybrid Clusters
- Multimodal Tracking

MVP Microsoft Most Valuable Professional

dotNET{podcast}

intel Software INNOVATOR

PACKT PUBLISHING Author

FAB LAB NAPOLI

AI Day

About us

Ing. Gianni ROSA GALLINA @giannirg

R&D Technical Lead @ deltatre

- AI, Machine Learning, Deep Learning on multimedia content
- Virtual/Augmented/Mixed Reality
- Immersive video streaming & 3D graphics for sport events
- Cloud solutions, web backends, serverless, video workflows
- Mobile apps dev (Windows / Android / Xamarin)
- End-to-end solutions with Microsoft Azure

MVP Microsoft Most Valuable Professional

Microsoft Specialist
Programming in C#
Programming in HTML5 with JavaScript & CSS3

Microsoft CERTIFIED Solutions Developer
Windows Store Apps Using C#
Web Applications

Microsoft Certified Azure AI Engineer Associate

Microsoft Certified Azure Data Scientist Associate

unity Certified Programmer

PLURALSIGHT

AI Day

Platinum Sponsor

Microsoft

avanade

ellycode
WHAT DO YOU NEED?

PORINI
A DGS COMPANY

Gold Sponsor

LOBRA

UNIKEY
Bringing IT knowledge to the people

Technical Sponsor

NET CODE

JET BRAINS

Packt

stickermule