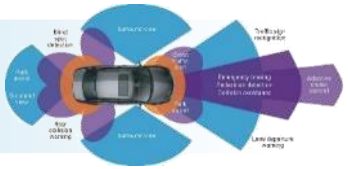
NeuPro-M: Highly Scalable, Heterogeneous and Secure Processor for High-Performance AI/ML in Smart Edge Devices

Yair Siegel
Senior Director, BD
CEVA inc.

AI Technology Challenges

Performance – ML prevails alternatives:
More use-cases ➔ More sensors ➔ More pixels ➔ More data



Scalability – Different use-cases, different needs:
ADAS, Autonomous L1-L5, Powertrain, Infotainment, DMS/OMS, …

Low power – Constrained energy envelope

Flexibility – Technology evolution faster than product deployment

SW Vs. HW

Riskless & Fast Solution – Software abstraction + hardware agnostic
comprehensive and compatible software is key

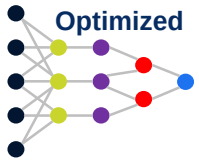
CEVA's Approach

Performance Optimized	▶ Heterogenous, parallel multi-element compute engine ▶ Wide range, mixed precision support
Aggressive Bandwidth Reduction	▶ Latency optimized via hierarchical memory structure ▶ Lossless compression, on-the-fly sparsity
Multiple Use-Cases	▶ Scalable and secure multi-engine, multi-core solution ▶ Automotive ISO26262 safety compliant
Long Term Solution	▶ Innovative **next-gen NN** heterogeneous compute architecture ▶ Vector Processing Unit in every engine enables **flexibility**
Fast, Simple & Riskless Implementation	▶ Comprehensive, system level toolchain, utilizes HW capabilities ▶ Supports open source languages and frameworks

NeuPro-M At-A-Glance

Computational Power
8-1200 TOPS

>250 Neural networks
>450 AI Kernels
>50 Algorithms



Simple integration

SW + HW + Tools

2b
4b
8b
12b
16b
32b

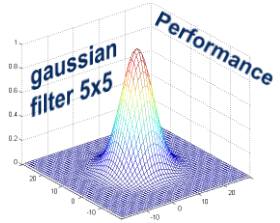
All data types

Maximum Utilization

>93% Utilization
SW **utilize** optimized HW

Power Efficient

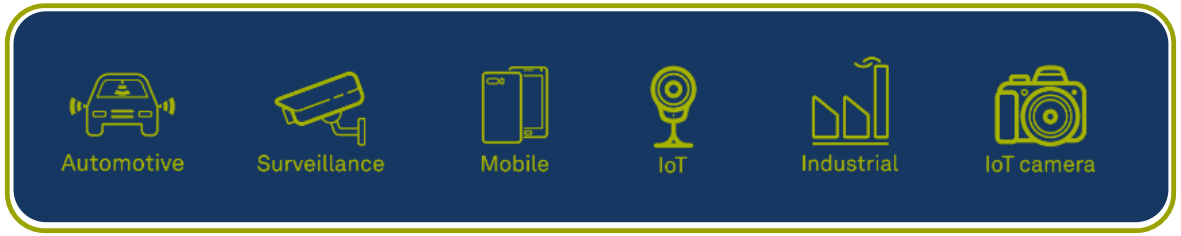
24 $\frac{TOPS}{watt}$



20k [fps]

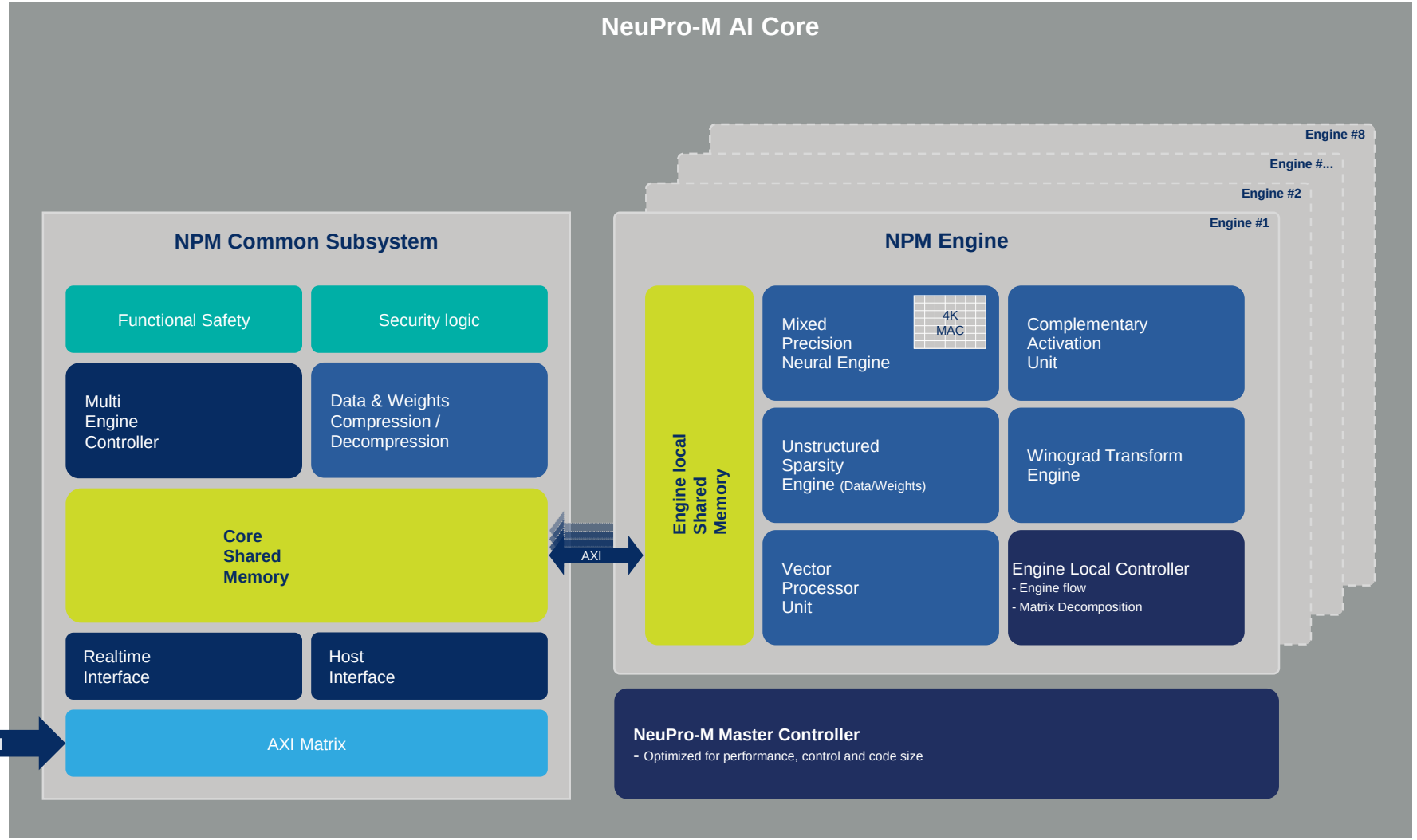
Performance
AI neural network
ResNet-50
5x-15x $\frac{inference}{second}$
vs. previous generation

Proven Experience
Markets



More than **6x**
vs. previous generation

NeuPro-M Block Diagram

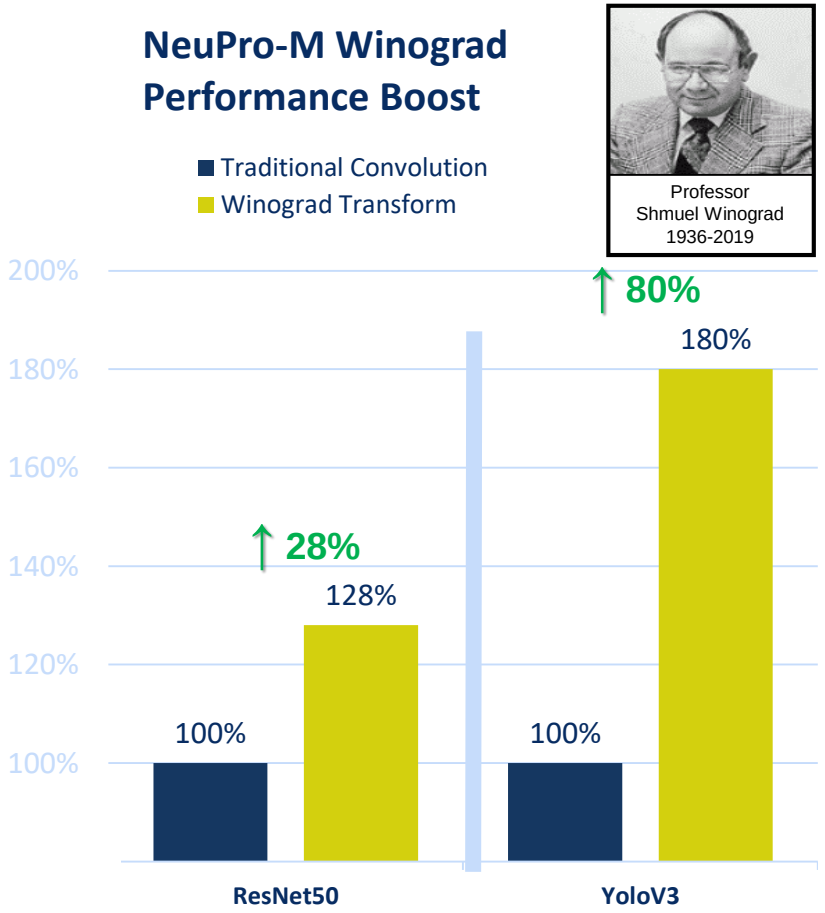


10 Top Distinctive NeuPro-M Features

NeuPro-M Feature List Highlights	Low Power	High Utilization	High Performance	Low Bandwidth	Low Latency	Agility / Scalability
Ultimate Control Scheme	✓	✓	✓	✓	✓	✓
Out-of-the-box Winograd transform	✓	✓	✓			
True (unstructured) Sparsity	✓	✓	✓	✓		
Unique GRID micro architecture *[Activation x Weights]* *(data type diversity with minimal power consumption)*	✓	✓	✓	✓		✓
Programmability		✓	✓	✓		✓
Optimized (reduced) Data Traffic	✓	✓	✓	✓	✓	
Data Compression	✓			✓	✓	
Next Generation AI Features *(e.g. transformers, 3D convolution)*		✓	✓			✓
Matrix Decomposition	✓	✓	✓	✓	✓	
Safety / Security						✓

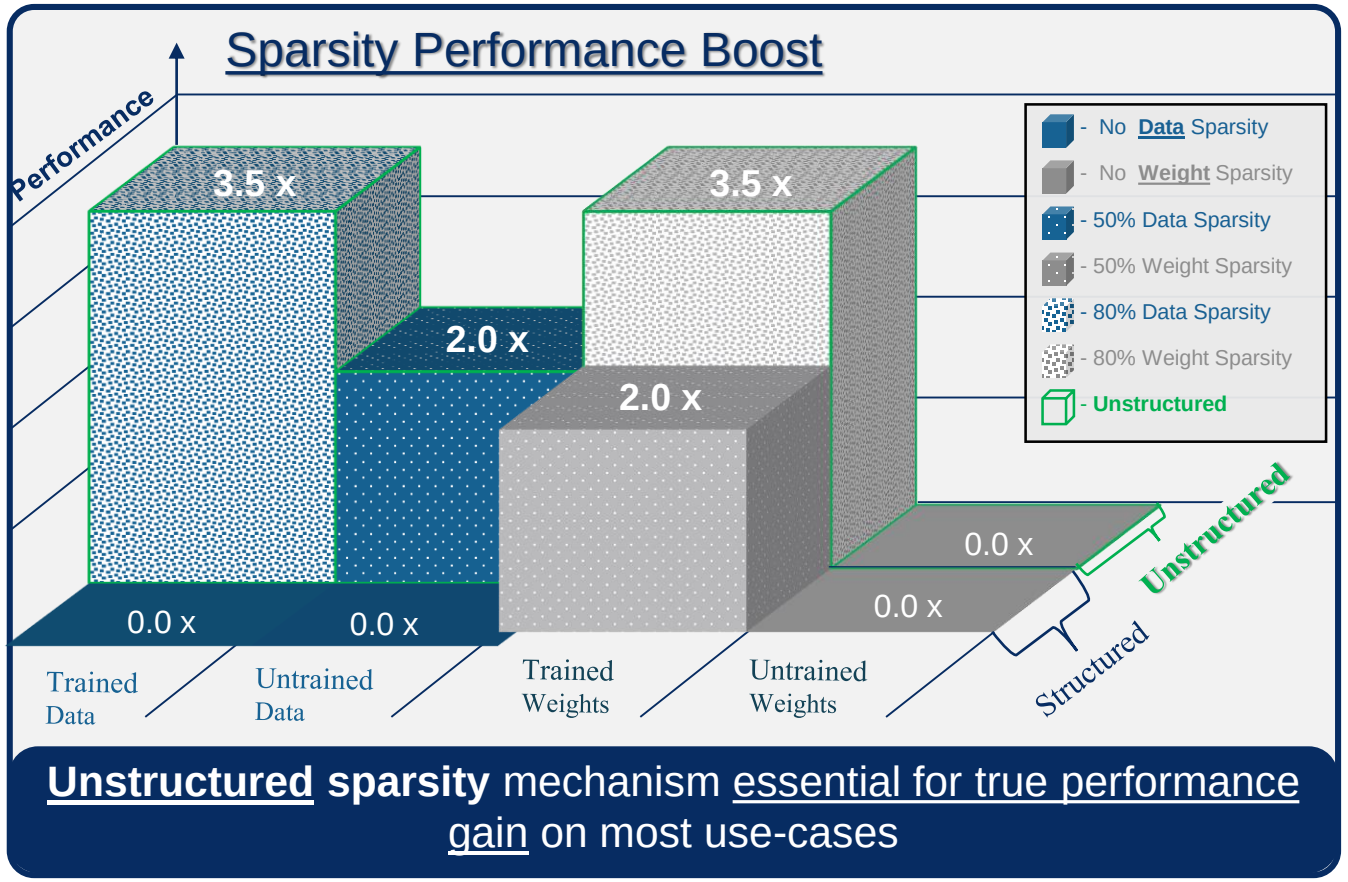
Out-of-the-box (untrained) Winograd Transform

- Alternative efficient way performing convolution
  - Using half the MACs (Multiply And Accumulate) operations
  - **Reduced power consumption**
  - **Negligible precision degradation**
- **2x performance gain** for 3x3 convolution layers
- **Out-of-the-box** Winograd "convolution"
  - Untrained = **No dedicated retraining needed**
    - 8-bit with <0.5% precision degradation
  - Wide range of data types supported
    - 4-bit, 8-bit, 12-bit, 16-bit



True Sparsity: Unstructured Pruning

- Ability to "**skip" zeros in data/activations or weights** along inference process
  - Performance gain by avoiding multiplication by zeros
- Up to **4x performance gain**
  - Bandwidth reduction
  - Power reduction
- **Preserves accuracy**
- Full HW & SW support
  - Training for data optional
- **Unstructured sparsity**
  - Essential for data
  - Higher weights sparsity level



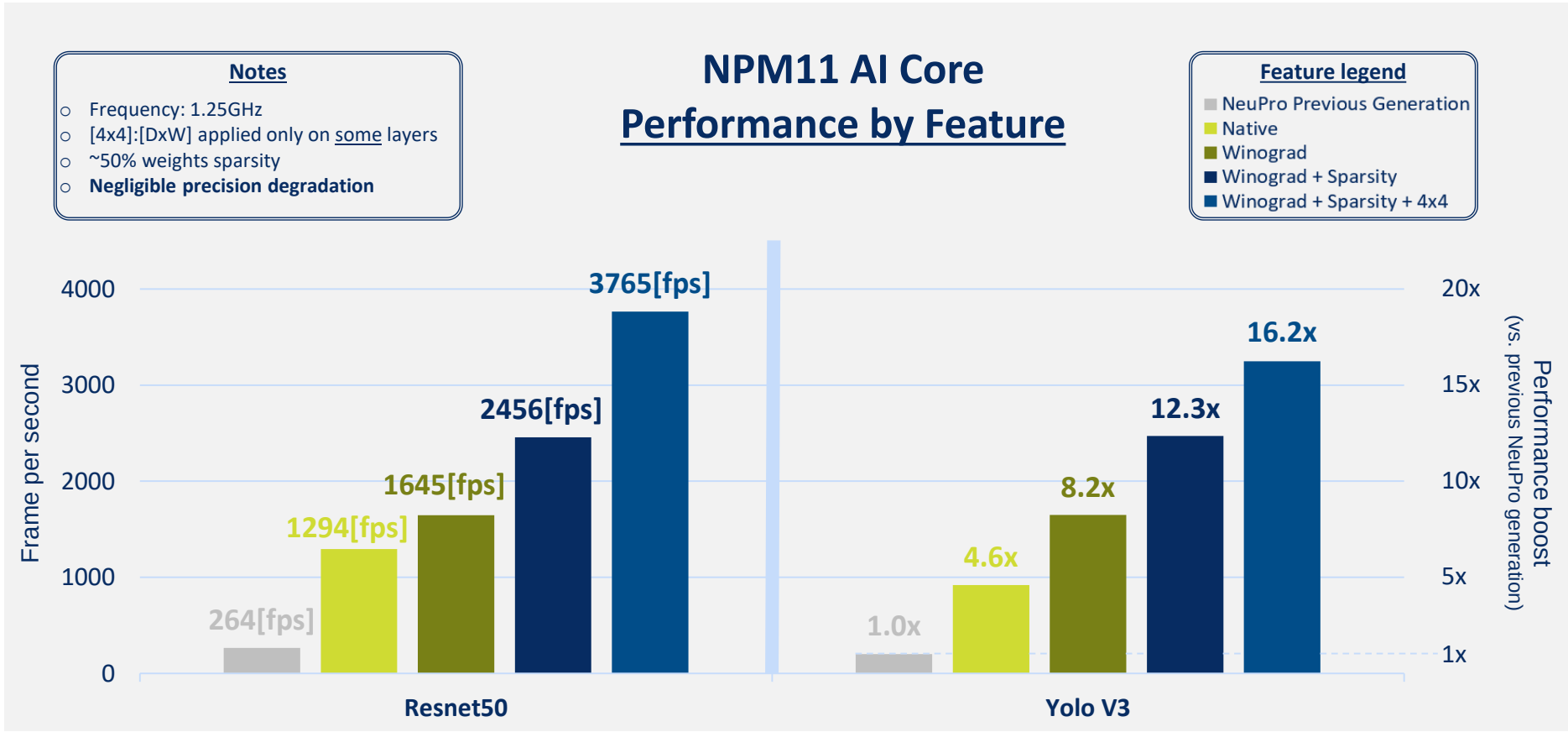
Data Type Diversity / Flexibility

- Unique design of **highly efficient** 4K MACs (Multiply And Accumulate) **mixed precision neural engine**
  - **Various data types** *support with minimal power consumption*
  - Low bit data and weights will generate
    - ➔ ***Lower bandwidth***
    - ➔ ***Reduced power consumption***
    - ➔ ***Performance acceleration***
  - Different layers using different data/weights types, while overall precision kept
- **Flexibility** - tackles different use cases

Type			[bit] x [bit]	# of MACs
2-bit	**Trinary**		2 x 2	6.0 K
			8 x 2	4.0 K
4-bit	**Fixed point**		4 x 4	16.0 K
			4 x 8	8.0 K
			4 x 12	5.3 K
			4 x 16	4.0 K
8-bit	**Fixed point**		8 x 8	4.0 K / 8.0 K*
			8 x 12	2.7 K / 4.0 K*
			8 x 16	2.0 K / 4.0 K*
12-bit	**Fixed point**		12 x 12	1.8 K / 2.3 K*
			12 x 16	1.3 K / 2.0 K*
16-bit	**Fixed point**		16 x 16	1.0 K / 2.0 K*
16-bit	**Half precision**	**Floating-point**	16 x 16	64
32-bit	**Single precision**	**Floating-point**	32 x 32	32

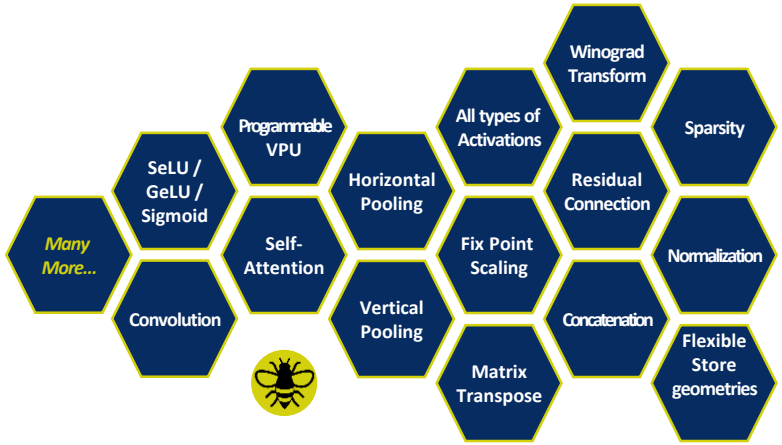
* With 50% sparsity

Single Engine NeuPro-M Core Performance



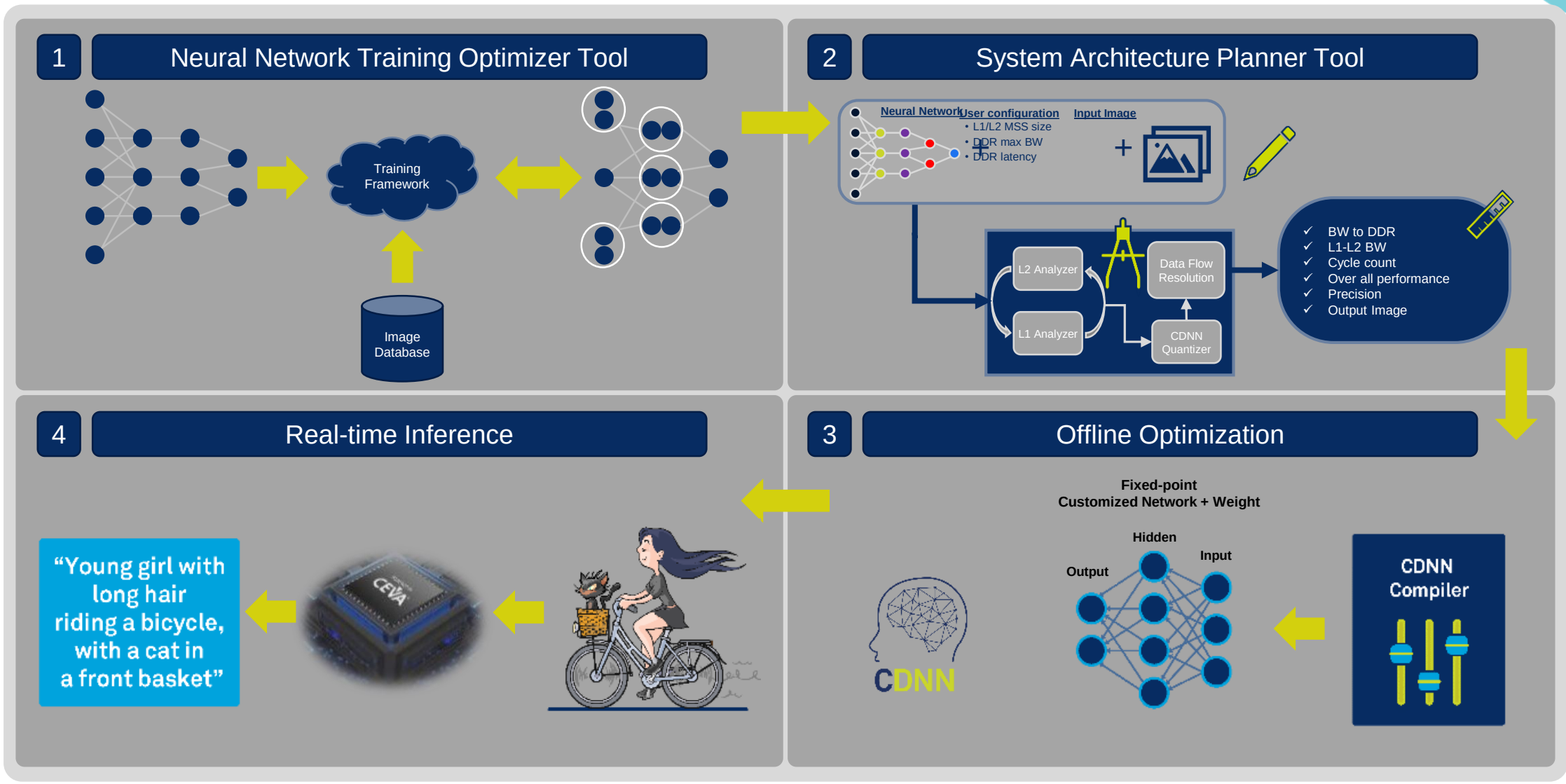
Ultimate Control Scheme

- Unique, optimized and interleaved control scheme
  - **Mixture of software & hardware operations**, orchestrates heterogenous hardware
    - Engines, co-processors, local controllers, DMAs & queue managers
  - Ensures efficient and deterministic performance
    - Optimal parallel interleaved operation
- No need to redispatch program upon layers
  - Reduces system data traffic and software complexity
- **On the fly, 'head to tail' fused operation pipeline**
  - Hardware based task flow control (not software interrupt based)
  - Reduces internal & external memory access e.g. L1, L2, DSP, DDR

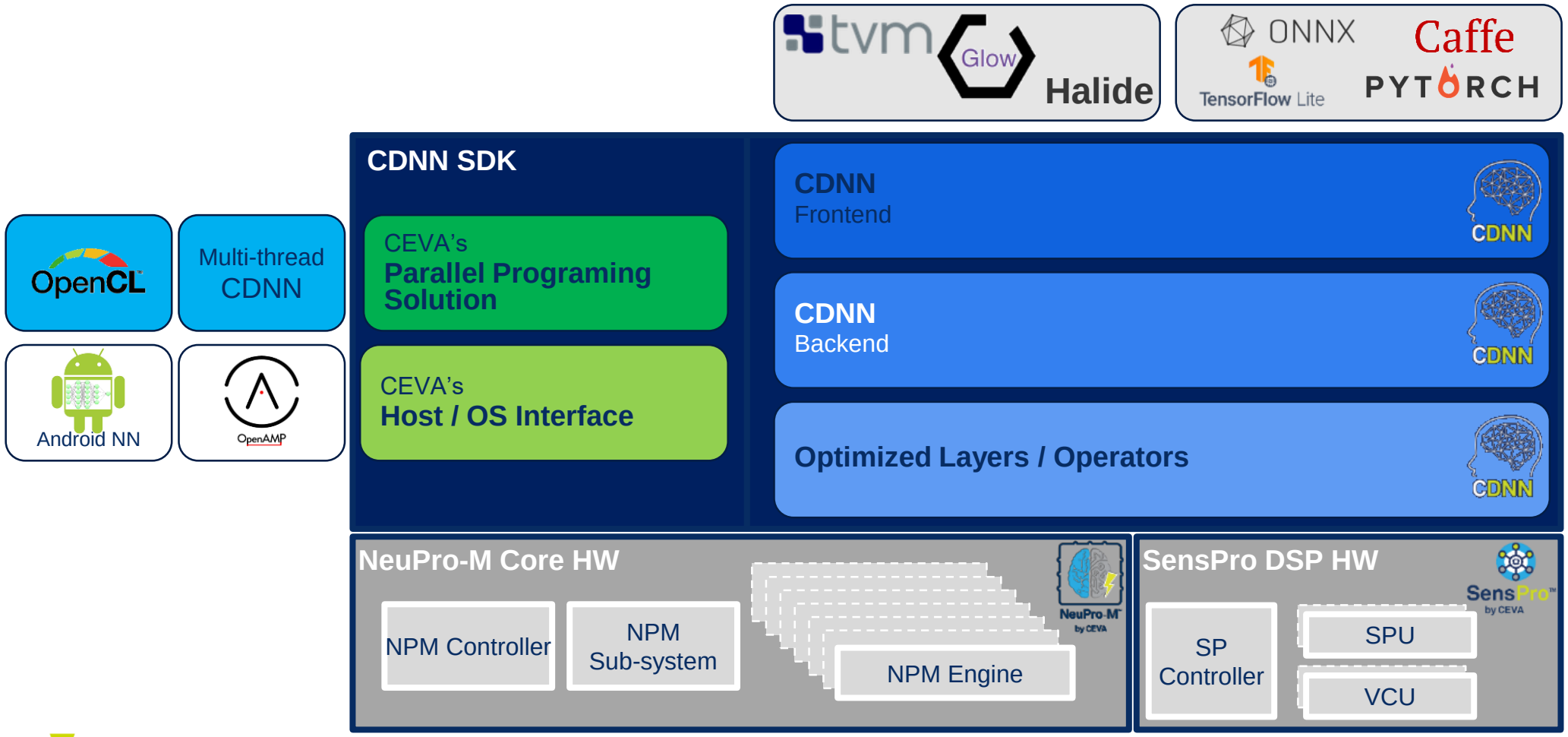


Example of "on-the-fly" fused operations

NeuPro-M Comprehensive Toolchain



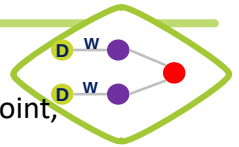
CDNN Supports Open-Source Ecosystem



NeuPro-M: Optimized AI Solution in Every Layer

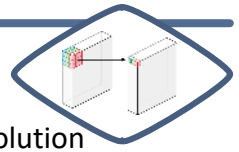
▶ Data & Weights

Mixed Precision: 32/16/12/8/4/2 bits, Fixed-point, Single/Half precision floating-point, Sparsity



▶ Elementary Operation

Winograd, Transformers, 3D-Convolution, Programmable operation, Depthwise-Convolution

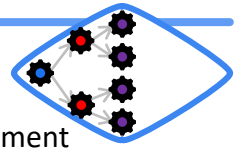


▶ Reduced Data Traffic

Bandwidth, Latency, Memory hierarchy, Data/Weights compression, Realtime clients

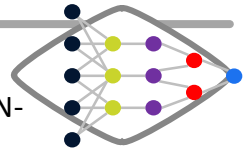
▶ System & Flow Control

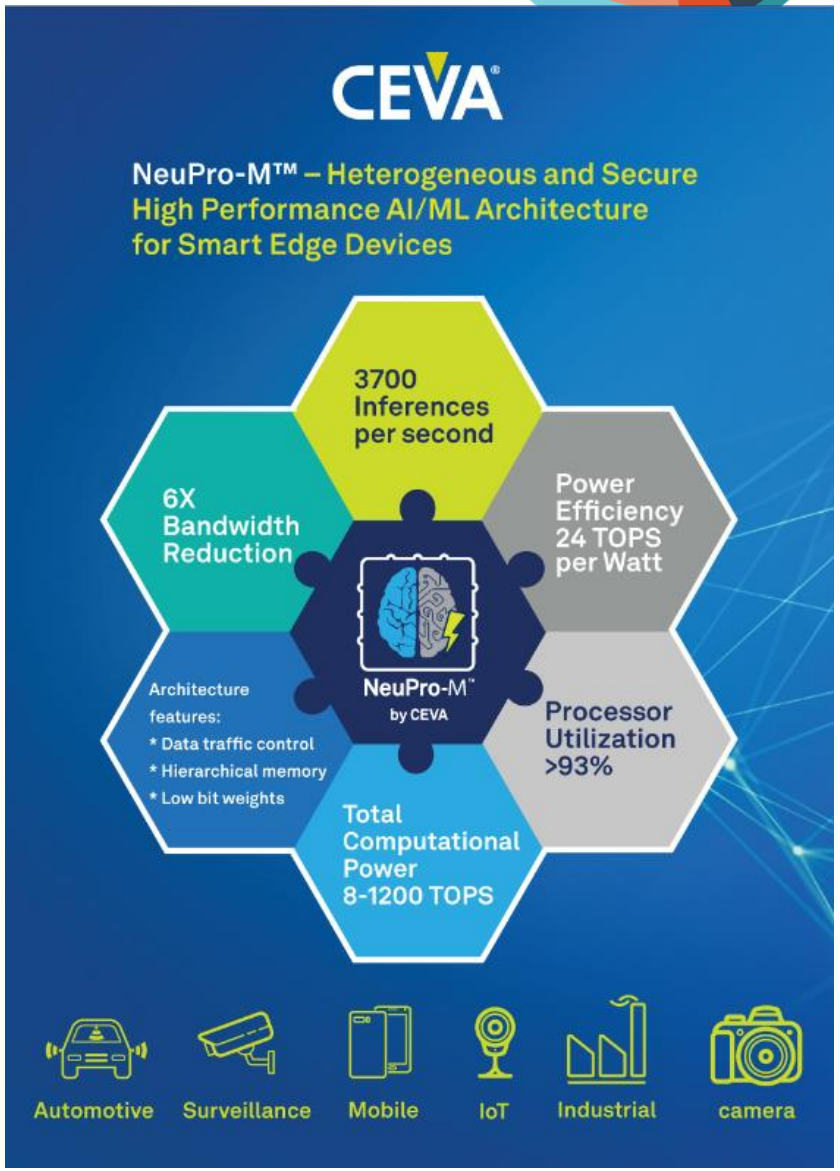
Compound parallel processing, Metrix decomposition, Decentralized data management



▶ Neural Network Models

HW-aware networks, CDNN Compiler, CDNN-Invite, Asymmetric quantization, Pruning





CEVA

Resources

- **NeuPro-M** Overview: https://www.ceva-dsp.com/product/ceva-neupro-m/
- **CDNN** Compiler: https://www.ceva-dsp.com/product/ceva-deep-neural-network-cdnn/
- **Fortrix,** Root-of-Trust and Security IP: https://www.ceva-dsp.com/product/fortrix-secured2d/
- **SensPro**, Sensor-hub and Computer Vision Processor: https://www.ceva-dsp.com/product/ceva-senspro/

THANK YOU