SYNOPSYS®
Silicon to Software™

AI Enabled DSPs to Accelerators – Dialing in the Right Performance

Gordon Cooper, Product Marketing Manager
Markus Willems, Sr. Product Marketing Manager

Synopsys ARC® Processor Summit 2022

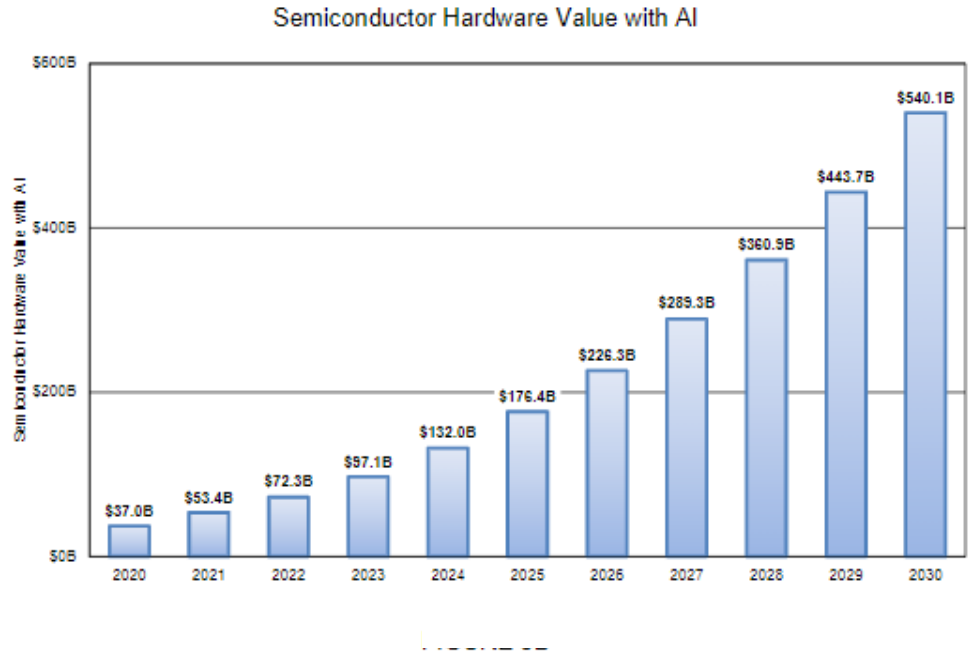
Agenda

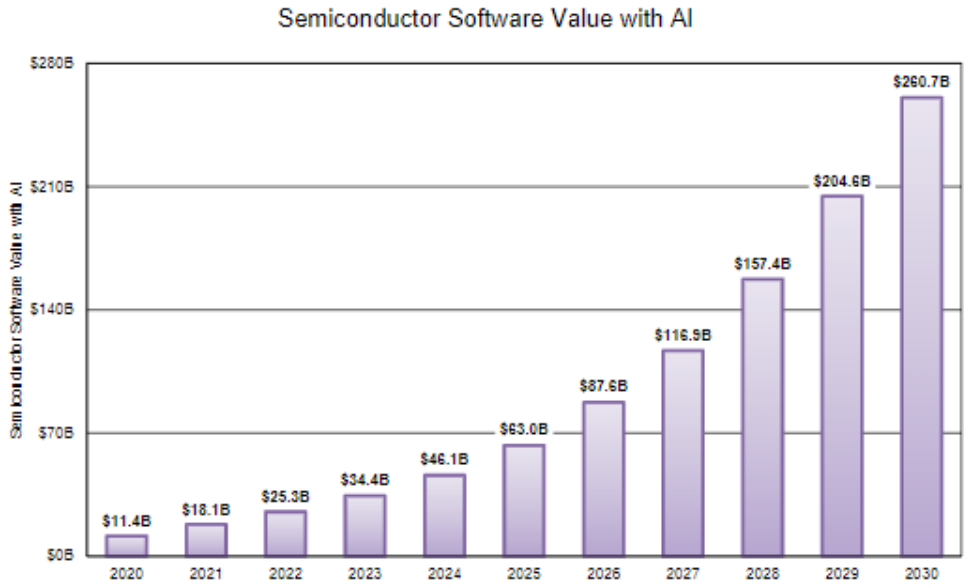
- AI is Proliferating
- AI System-Level Challenges
- DSP vs NPU Implementation Options
- Synopsys Portfolio of DSPs and NPUs
- Summary

AI is Proliferating

Semiconductor market related to AI will be $800.8 billion in 2030 compared to $48.4 billion in 2020, indicating a CAGR of 32.40%

(Global Semiconductor Industry Service report, May '22, IBS)





Neural Network Market (Edge Devices)

Performance Requirements per Application are Increasing

- AIoT
- Human activity recognition

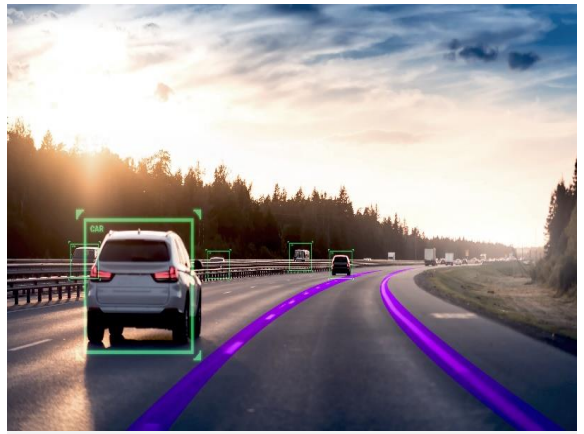
<100 GOPS

- Robotics / Drones
- Automotive Powertrain
- Games/toys
- Audio / Voice control
- Facial detection

100 GOPS to 1 TOPS

- Driver Monitoring Sys
- Surveillance
- Facial recognition
- Digital still cameras
- High End Gaming
- Augmented reality
- Mid-end smartphones
- Facial recognition

1 to 10 TOPS



- ADAS Front Cameras
- ADAS LiDAR/Radar
- High end surveillance
- High-end smartphones
- DTV
- HPC
- Microservers (inference)
- Data center (inference)

10 to 1000+ TOPS

Real-Time AI Examples In the Market

AI for Drones

- Object detection, classification, and tracking to avoid collisions and locate and track targets (ie. to monitor elephant herds and identify poachers)

AI for Automotive

- Object (i.e. pedestrian, street sign) detection
- Semantic segmentation for path finding
- Sensor Fusion to combine vision, radar, lidar data
- Driver monitoring

AI for Mobile / Digital Still Cameras

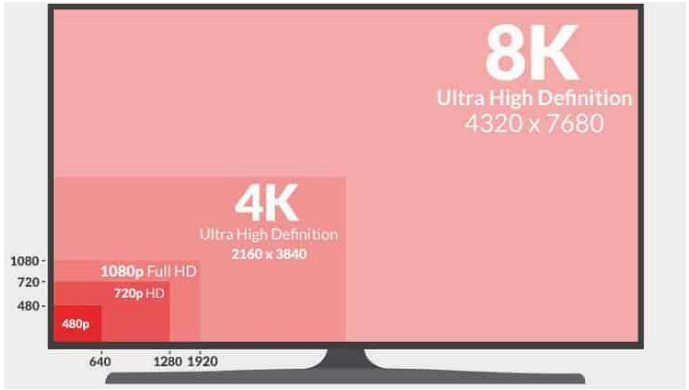
- Facial recognition & other biometrics for security
- Smile recognition
- Voice activation
- Computational Photography
- Super resolution for image enhancement / noise reduction

AI System-Level Challenges

AI Complexity Grows – It's Not Just the Algorithms

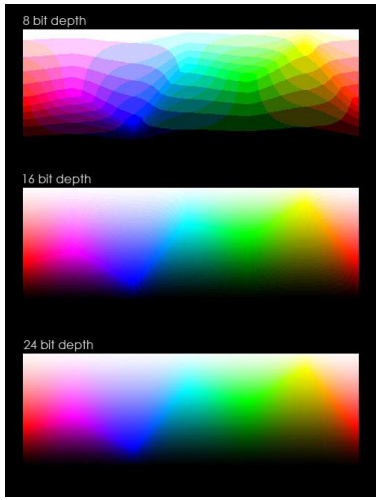
More Bandwidth & Compute Required

High Definition
Requires More Compute & Memory



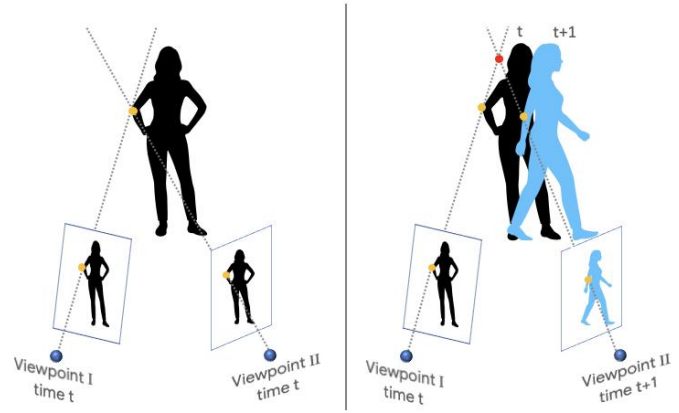
- ImageNET – common AI benchmark dataset: 224x224 pixels
- HD 40x more pixels than ImageNET
- 4K 160x more pixels than ImageNET

Higher Resolution & Dynamic Range
Enable Machine Vision, Real-Time Perception



- RAW16 color depth 2x more data than RAW8
- RAW24 color depth 2x more data than RAW16

Multiple Camera Arrays & 4D Sensing
Enable Depth & Motion Inference



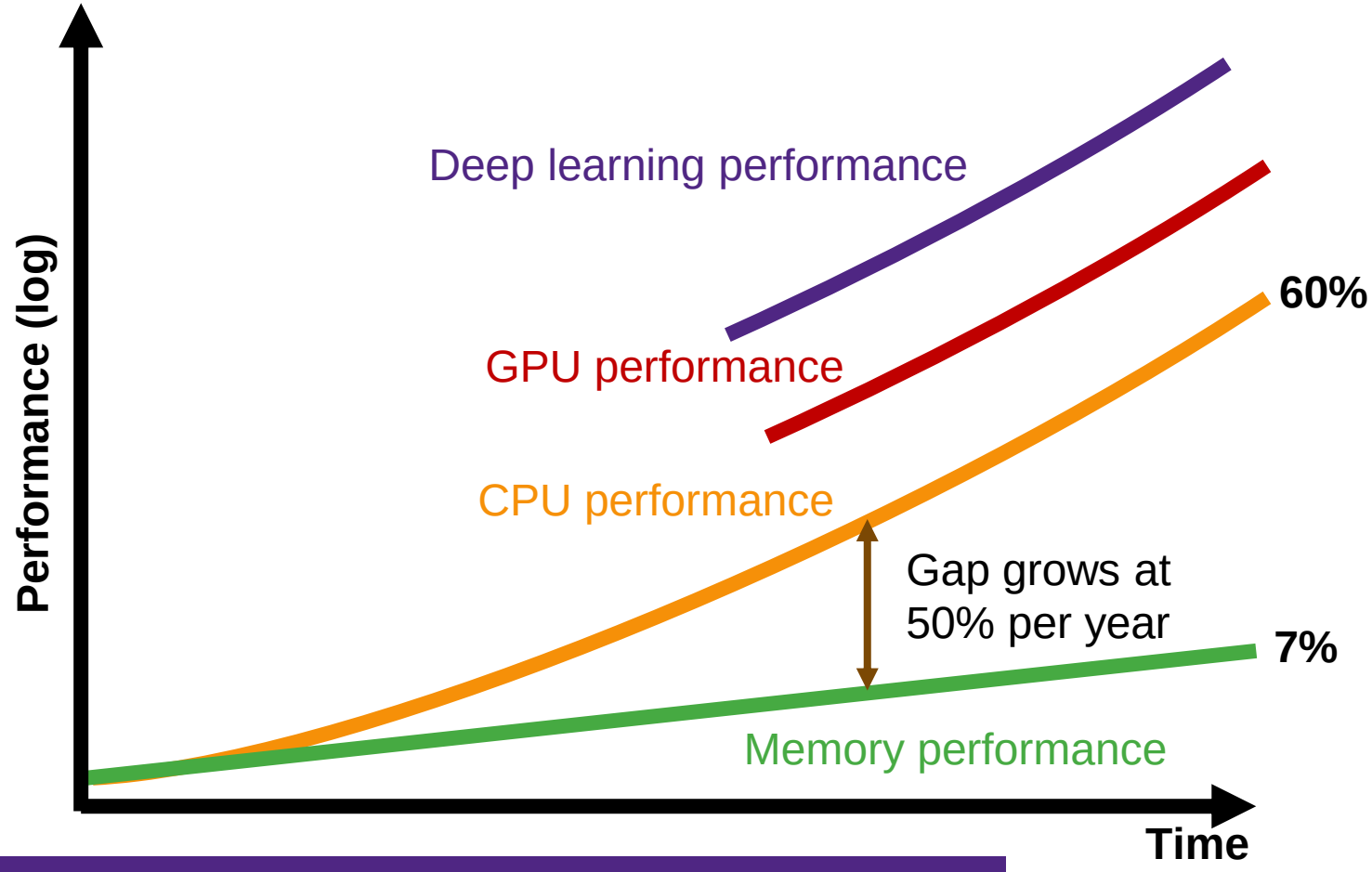
- Additional cameras multiply the required memory and compute

AI performance requirements growing rapidly by application

AI Growth Constrained by Memory Bandwidth

Deep Learning Performance Outpacing Memory

- Moore's Law: CPU performance outpacing memory access speed
- GPUs initiated Deep Learning in 2012, widening the gap
- Deep Learning accelerators outpace GPUs
- Goal - reduce data movement
  - Innovative heterogeneous memory architectures required
  - From on-chip memory compilers to high bandwidth HBM2



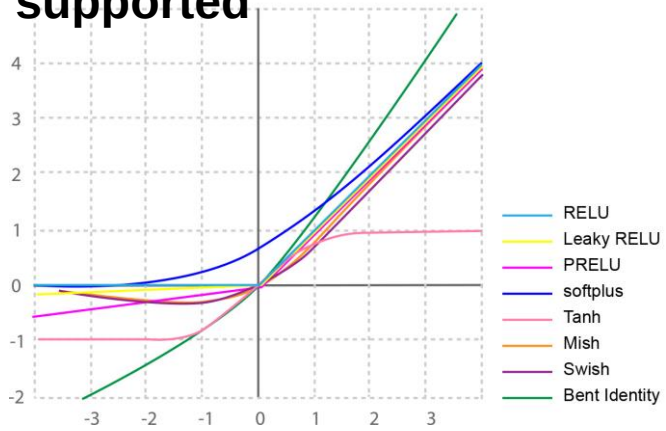
Processing performance is outgrowing memory bandwidth

AI Research Requires Evolving Architectural Improvements

AI Hardware needs to be optimized and flexible to keep up with emerging trends

Evolving AI Research
Emerging Neural Network models (i.e. EfficientNet, Transformers) require more advanced hardware and software techniques

Example: Growing list of activation functions to be supported



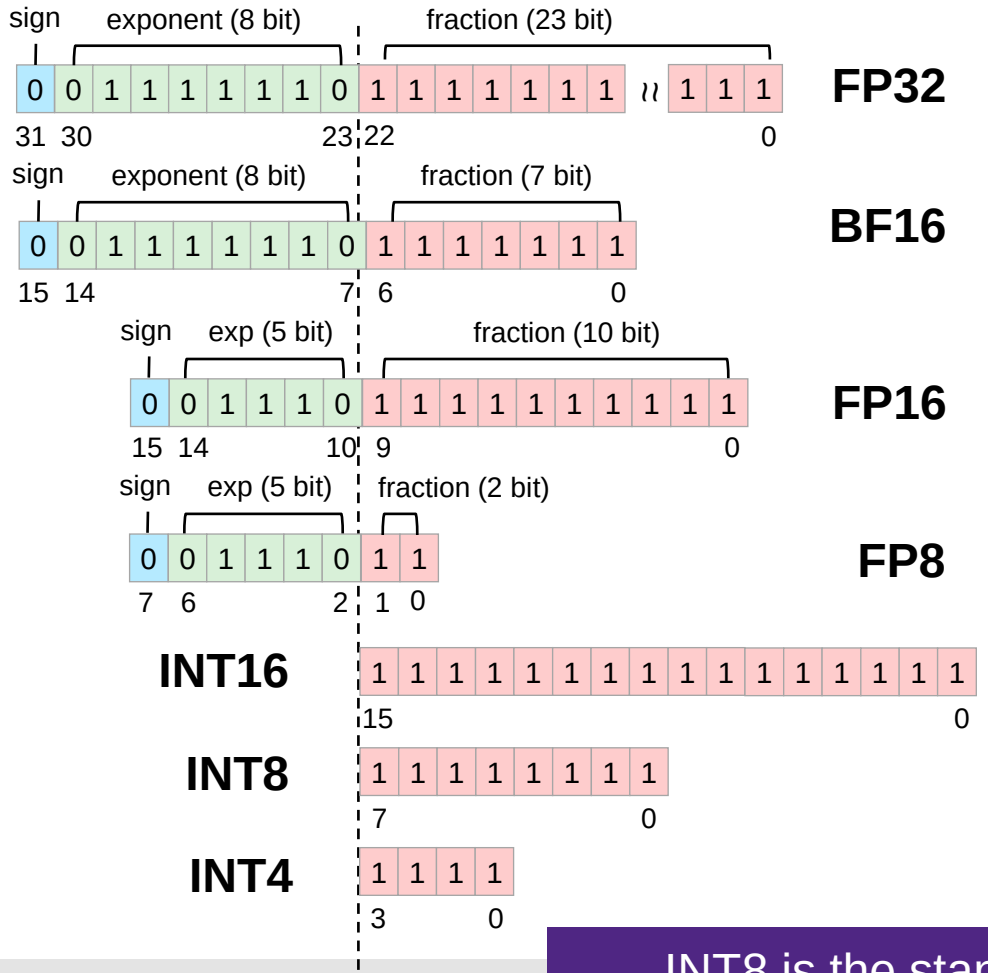
ReLU, Leaky-ReLU, ReLU6, ReLU1, PReLU, Sigmoid, Tanh, Swish, H-swish, Mish, GELU, GLU, etc.

Architectural Enhancements Required

- Look-up table for activation functions
- Improved support for depthwise separable convolutions

Flexibility is crucial for future proofing design – for chips that come out in 1-2 years

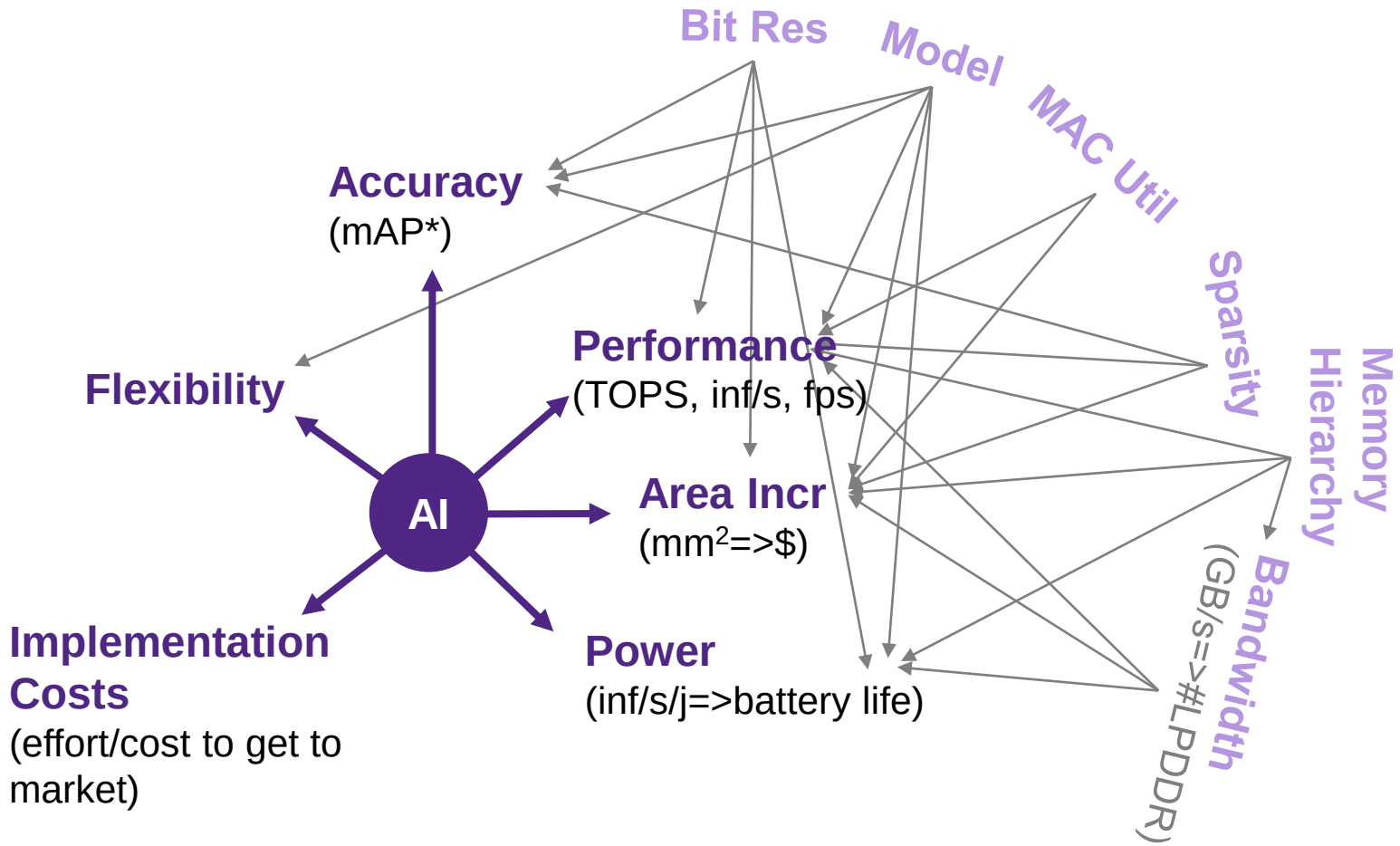
NN Applications Use Wide Range of Data Representations



- **FP32** typical format used in GPUs for NN model training
- **FP16 & BF16** are NOT needed for accuracy over INT8/16 – they make the transition from GPU easier, avoids having to retrain models
- **FP8** has more traction for training than inference
- **INT16** provides accuracy 'insurance' for radar and super resolution (at reduced performance)
- **INT8** standard for neural network object detection
- **INT4** can save bandwidth; not very popular yet

INT8 is the standard resolution for object detection

Customer Decision Factors for Adding AI to Real Time SoC



Key Performance Indicators (KPIs) aka Customer/System requirements drive AI processing selections

*Mean Average Precision (number of correct predictions)

DSP vs NPU Implementation Options

Choosing Hardware for Real Time AI

CPU, GPU, DSPs, NPUs (AI accelerators), etc.

	Performance	Area Eff	Power Eff	Flexibility	Comments
CPU	★	★	★	★★★★★	CPUs don't have math horsepower for fast NN processing
GPU	★★★★★	★	★	★★★★★	GPUs have high performance but large area and higher power
FPGA	★★	★★★	★	★★★	Good for prototyping but are expensive, high-power and limited in frequency
DSP	★★★	★★★	★★★	★★★★★	Vector DSPs have good parallel math but not dedicated for NN processing
NPU	★★★★★	★★★★★	★★★★★	★★★	Optimized for NNs
Hard-wired accelerator	★★★★★	★★★★★	★★★★★	★	Optimized for specific tasks, but not flexible for evolving 'next thing'

Trade-offs for Product Selection: DSP vs NPU

Summary

Vector DSP

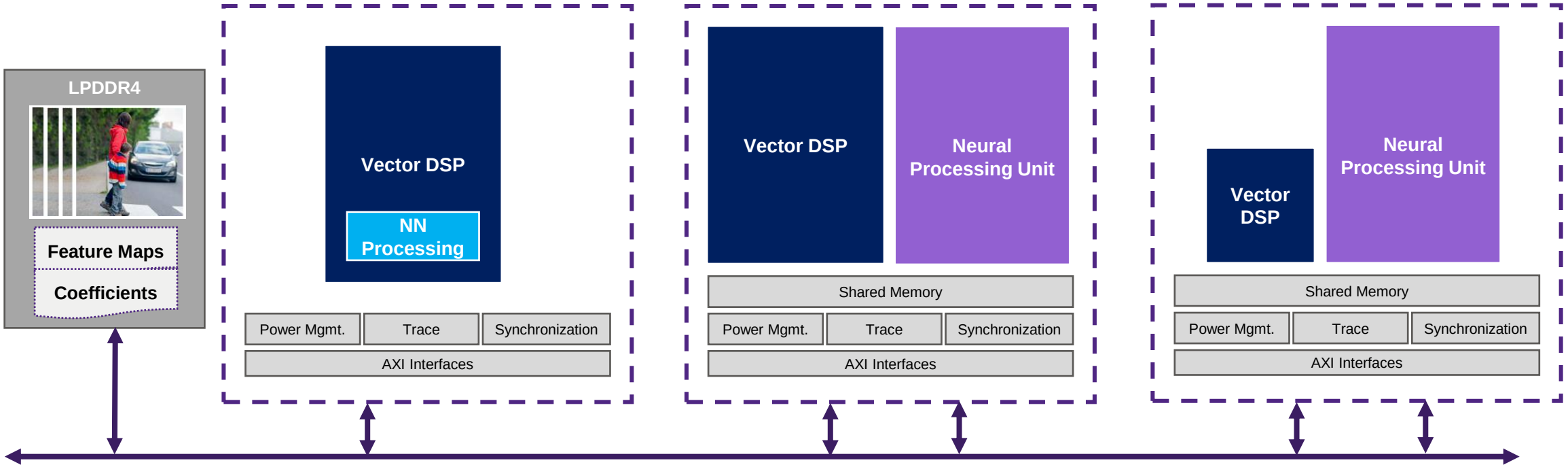
- DSPs provide more **flexibility** than NPUs
- DSPs can perform DSP and lower performance AI **with no additional area**
  - Specifically relevant if AI tasks are small part of application workload and/or are not always active
- Vector DSP used to support functions that can't be processed on NPU

Neural Processing Unit

- NPU accelerates all common AI layers – CNN, RNN, Transformers, Recommenders, etc.
- NPUs more **power** and **area** efficient for mid- to **high-performance** AI needs
- For relevant AI workloads (MAC dominated)
  - NPUs have better power efficiency (3-10x)
  - NPUS have better area efficiency (2-8x)
- NPUs need DSPs for future-proofing

AI Processing Requires both DSP and NN capabilities

Mix of DSP and NN processing depends on use cases

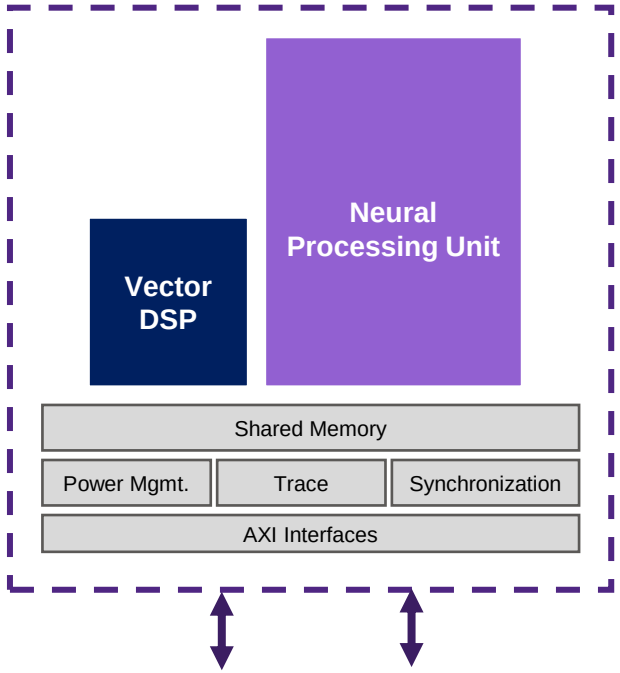


Configuration 1:
Vector DSP only

Configuration 2:
Vector DSP and NPU prioritizing both DSP functions and NN support

Configuration 3:
NPU priority with vector DSP as NN support only

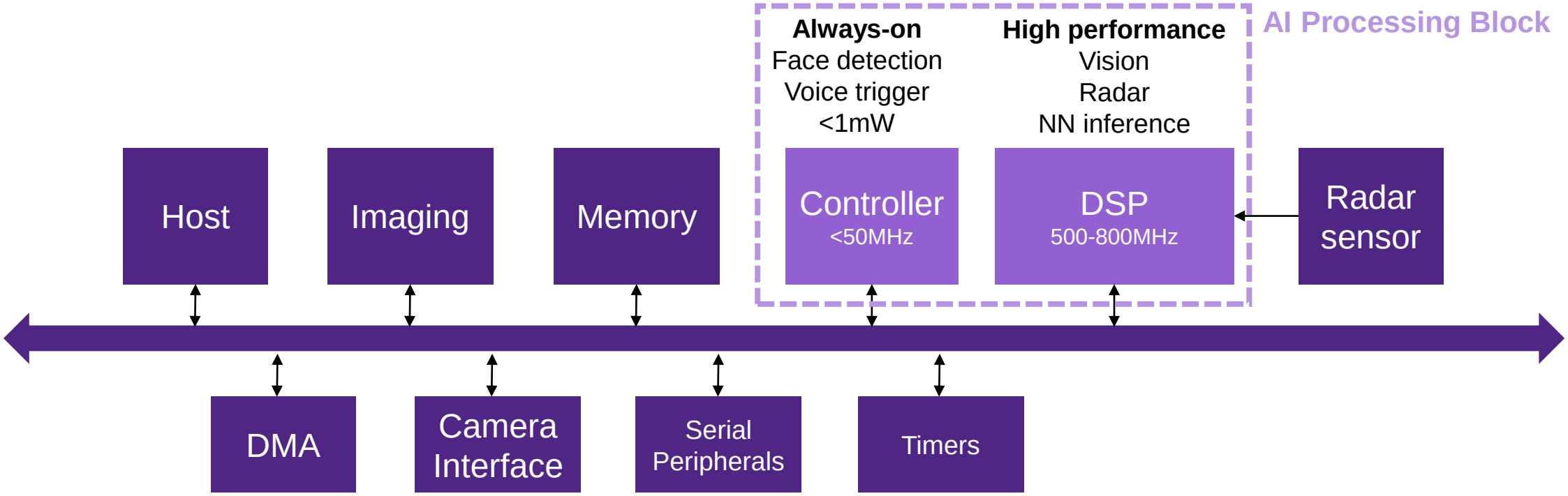
Technical Criteria for Product Selection DSP vs NPU



Configuration 3:
NPU priority with vector DSP as NN support only

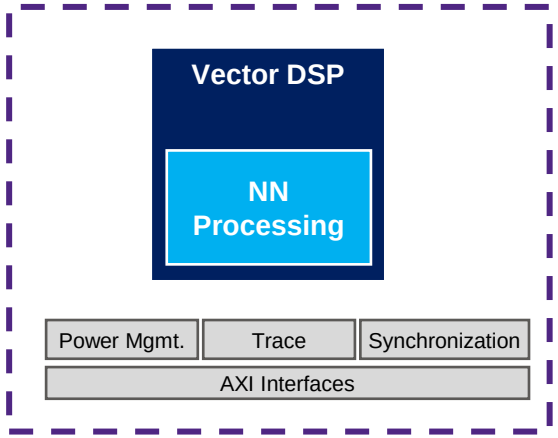
- **Configuration 3: NPU Priority**
  - NPU accelerates all common AI layers – CNN, RNN, Transformers, Recommenders, etc.
  - Vector DSP used to support functions that can't be processed on NPU, e.g.
    - Non-maximum suppression in Object Detection graphs (Used in Yolov4, Yolov5, Resnet34-SSD)
    - Glue code of graphs that consist out of multiple sub-graphs
    - MT-CNN, Faster-RCNN
    - ROI pooling and ROI alignment
    - Faster-RCNN, PVANet
    - Preprocessing for Sparse Lidar graphs
    - SECOND, SMConv
    - Preprocessing of Speech Recognition graphs
    - Deepspeech2, RNN-T, Conformer, Wave2Letter
    - Etc.

Case Study: Smart Home Application



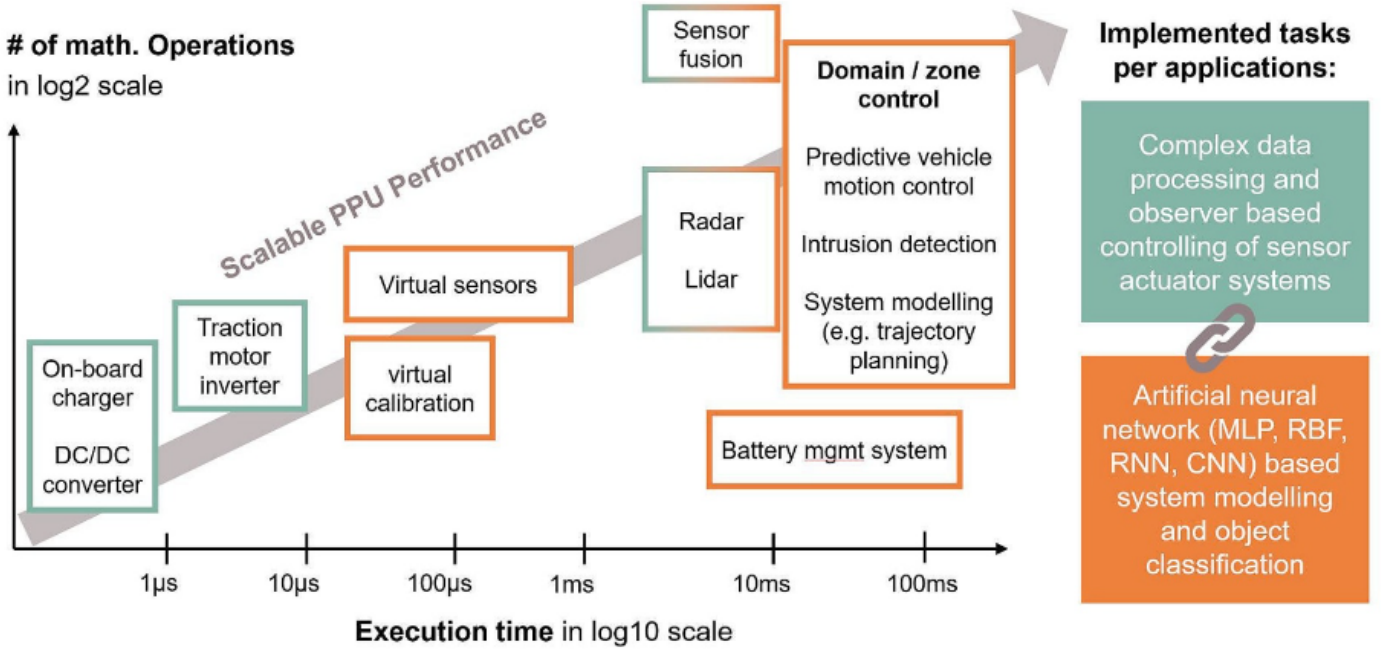
- Meeting 1mW for Always-On demands very low-power processor running at low frequency – **Controller**
  - For example: <20 uW/MHz x <50MHz
- High performance demands few orders of magnitude more performance – **DSP**
  - Vector DSP that meets diverse requirements of Vision, Radar and NN inference
  - Powered down when system is in Always-On mode

Case Study: Powertrain



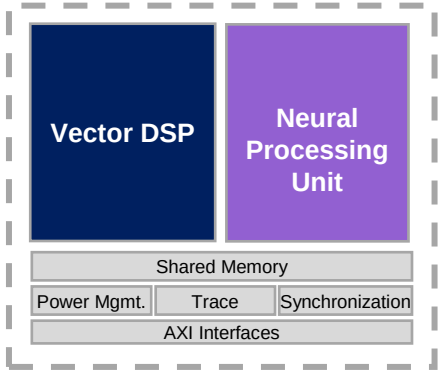
PPU is a flexible architecture to address automotive applications with fast execution times and large data processing requirement

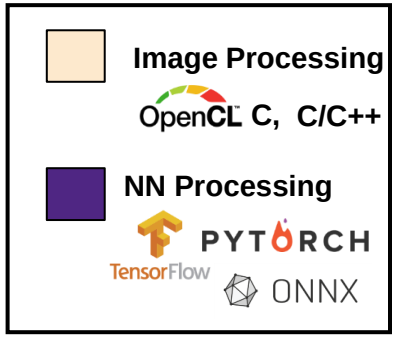
The implemented tasks and use cases will differ per application but two main clusters can be identified. On the one hand, PPU allows complex data processing and observer-based controlling of sensor actuator systems (e.g. for traction motor inverter or DC/DC converter control). On the other hand, it enables the implementation of artificial neuronal networks (MLP, RBF, RNN, CNN) based system modeling (e.g. for virtual sensors, state of health/state of charge optimization in battery management systems, and predictive vehicle motion control for in future domain or zone controllers) and object classification (e.g. Sensor Fusion) solutions.

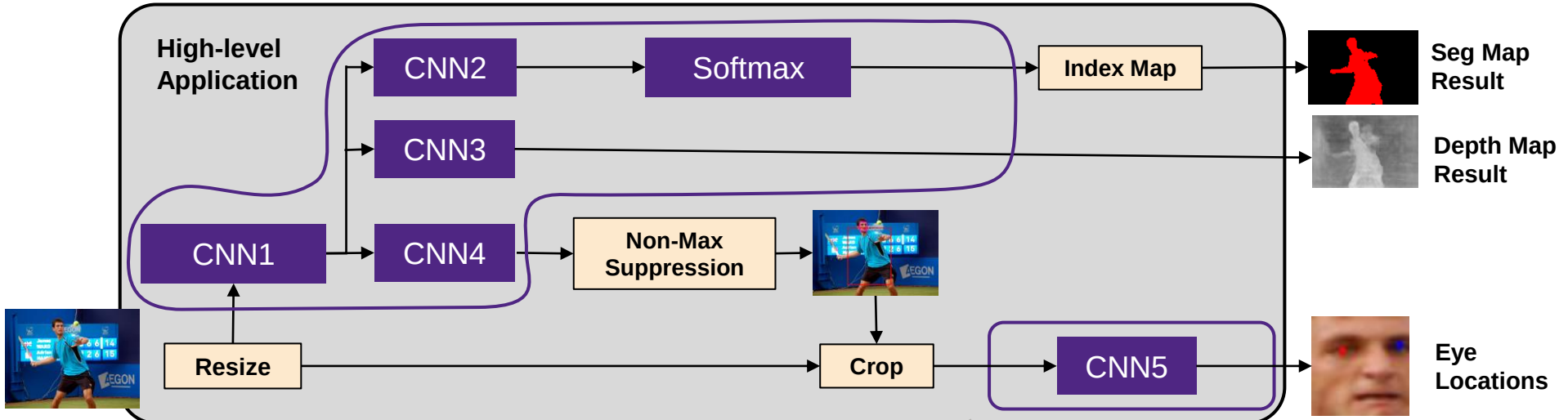


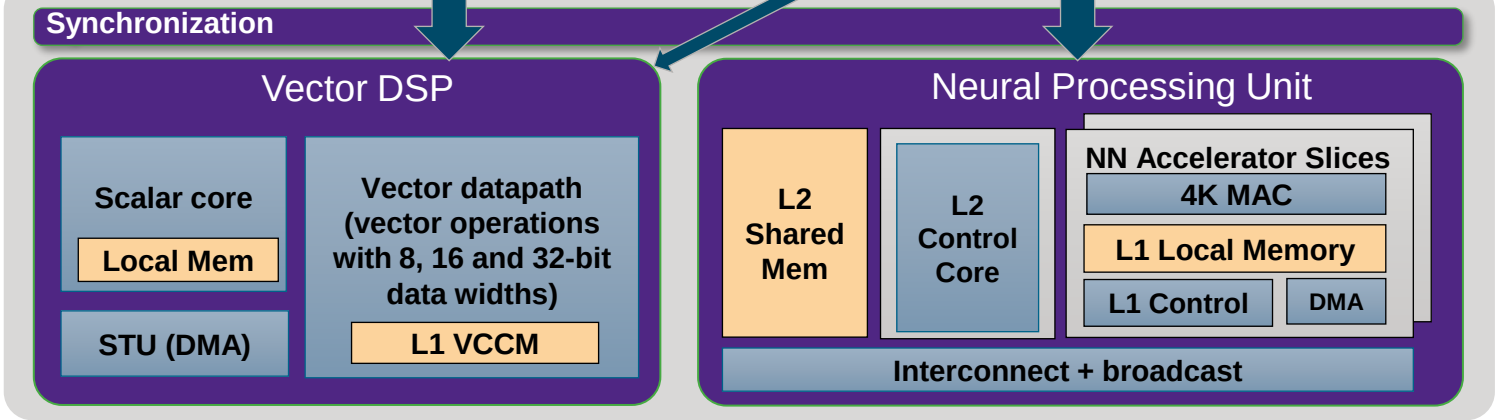
https://www.infineon.com/cms/en/product/promopages/new-ppu-simd-vector-dsp/

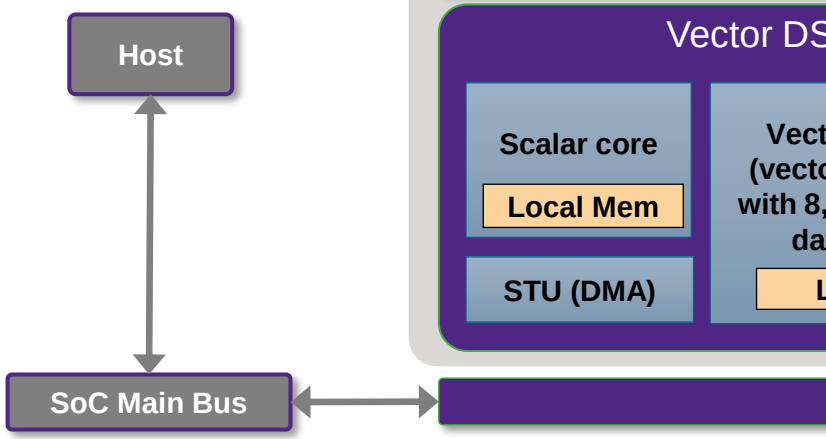
Case Study: Digital Still Camera











Synopsys Portfolio of DSPs and NPUs

Synopsys Processor Portfolio

Unrivaled Efficiency for Embedded Applications

EM Family

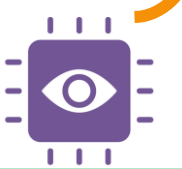
- Optimized for ultra low power IoT
- 3-stage pipeline w/ high efficiency DSP
- Power as low as 3uW/ MHz
- Area as small as 0.01mm^2 in 28HPM

SEM Family

- Security processors for IoT and mobile, including DSP
- Protection against HW, SW, and side channel attacks
- SecureShield for TEEs

HS Family

- Highest performing CPUs, CPU + DSP
- 32- & 64-bit ISAs
- High-speed 10- stage pipeline
- SMP Linux support



EV Family

- Heterogeneous multicore for vision and AI processing
- DNN (Deep Neural Network) Engine
- High productivity, standards-based tool suite

VPX Family

- High performance vector DSP
- SIMD/VLIW design for massive parallel processing
- Multiple vector FP engines for high precision results

NPX Family

- Scalable neural processor units
- From 1 to 250 TOPS
- Supports latest AI applications
- High productivity, standards-based tool suite

ASIP Designer

- Tool automating the creation of application-specific instruction-set processors (ASIPs)
- When processor IP cannot meet PPA requirements and fixed hardware is not flexible enough

Functional Safety (FS) Processors

- Integrated hardware safety features for ARC EM, SEM, HS, VPX, EV and NPX processor families
- Accelerates ISO 26262 certification for safety-critical automotive SoCs

Neural Network Market (Edge Devices)

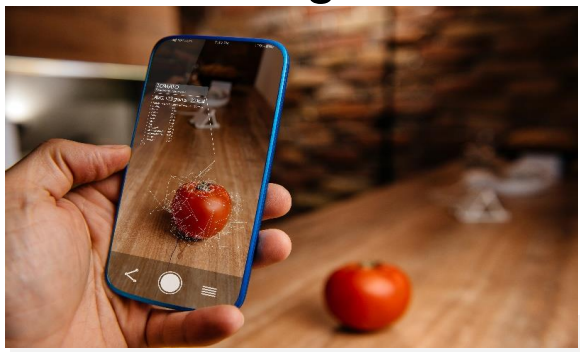
Performance Requirements per Application are Increasing

AIoT; Human activity recognition

<200 GOPS

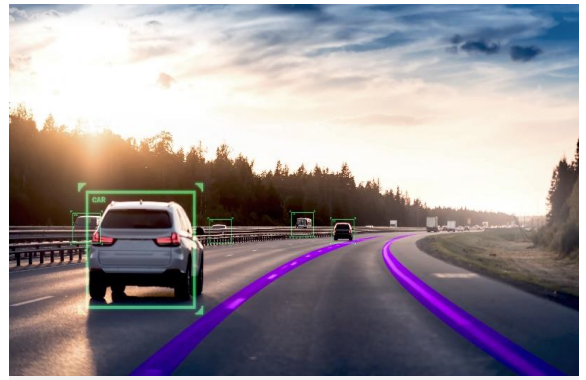
Robotics / Drones; Automotive Powertrain; Games/toys; Audio / Voice control; Facial detection

100 GOPS to 1 TOPS



Driver Monitoring Sys; Surveillance; Facial recognition; Digital still cameras; High End Gaming; Augmented reality; Mid-end smartphones

1 to 10 TOPS



ADAS Front Cameras; ADAS LiDAR/Radar High end surveillance; High-end smartphones; DTV; HPC; Microservers (inference); Data center (inference)

10 to 1000+ TOPS

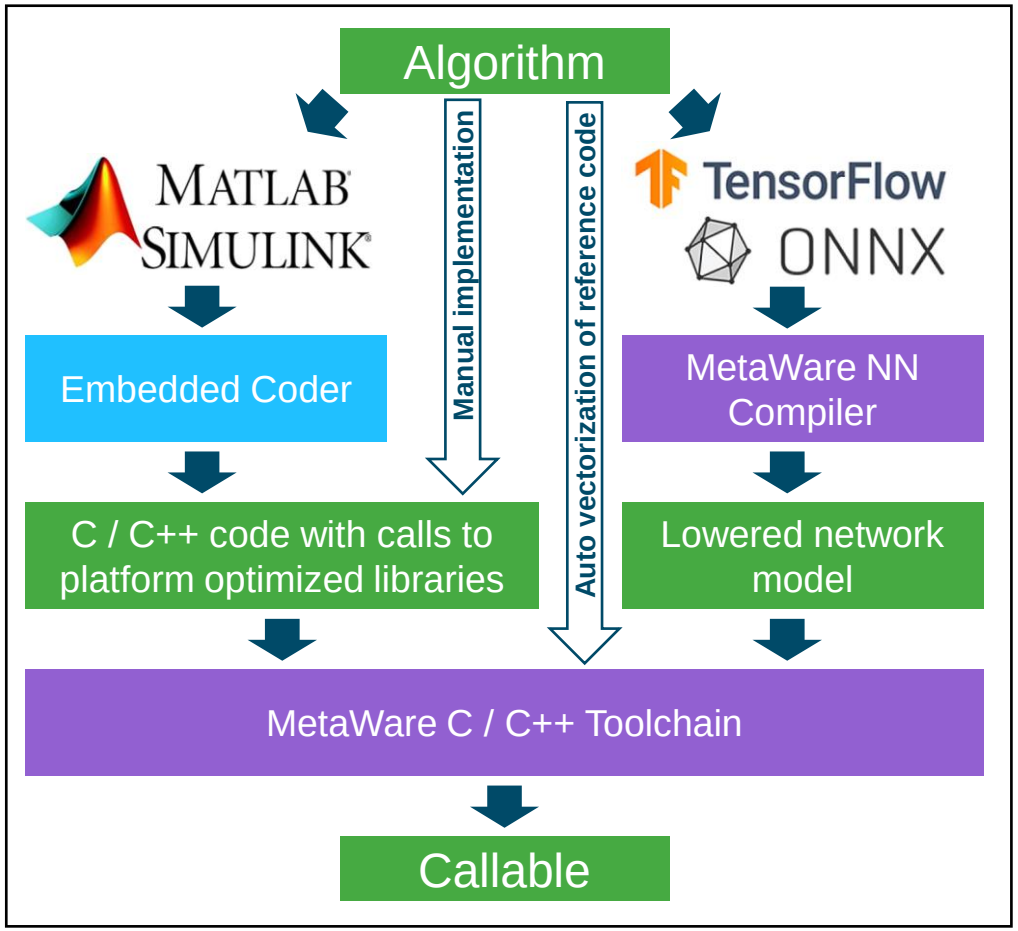
EMxD

VPX2 – VPX5

NPX6-4K – NPX6-96K

2x to 8x NPXs

VPX: Software Tool Support for both DSP and AI

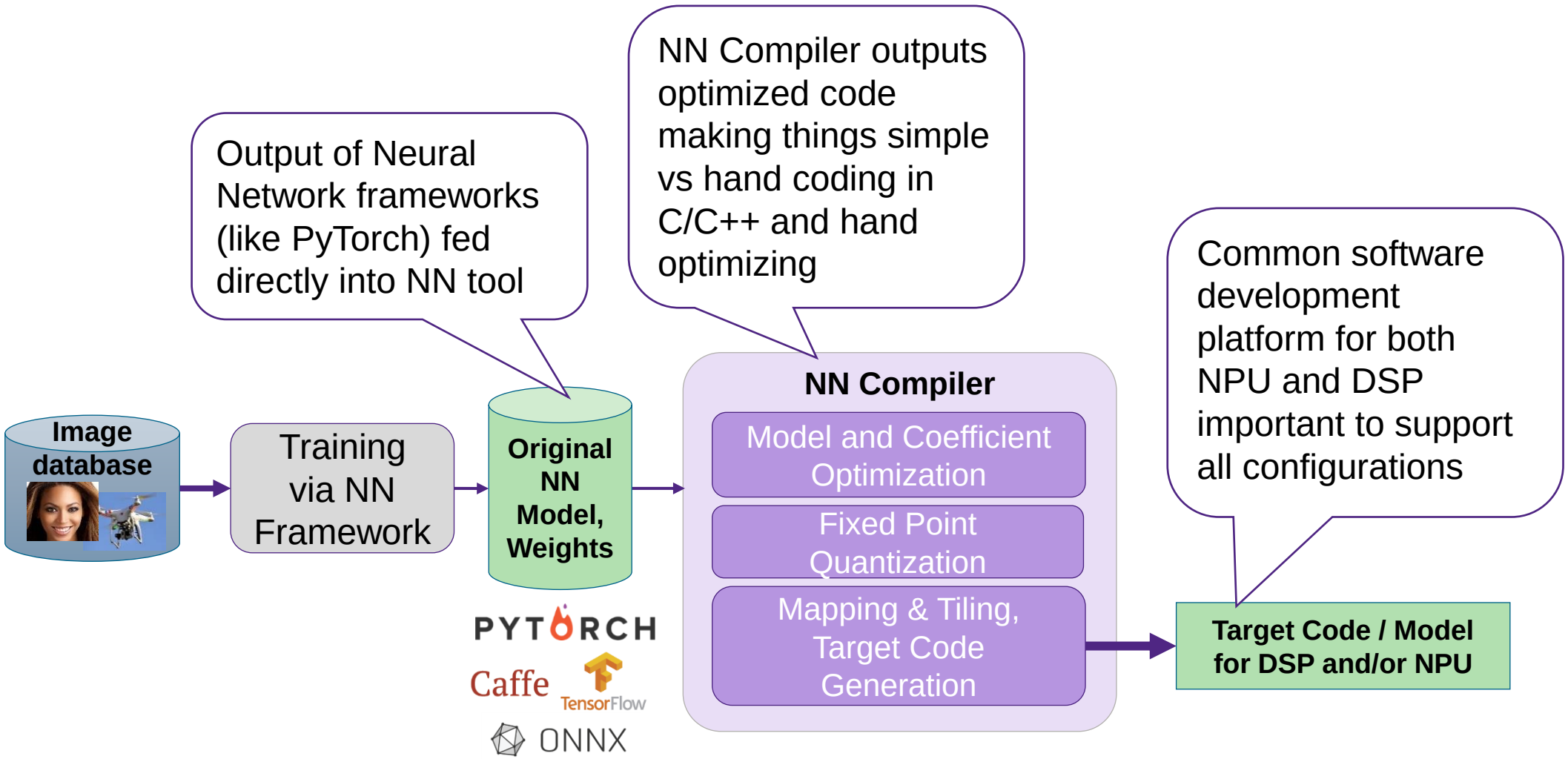


AI adds new requirements to the programming flow

- DSP programming options
  - C/C++ and OpenCL C
  - Model-based design
  - Leveraging processor-optimized libraries, e.g. for DSP, linear algebra, vision processing
- AI programming options
  - C/C++
  - Trained NN Models and Weights
  - Leveraging processor-optimized library for machine-learning

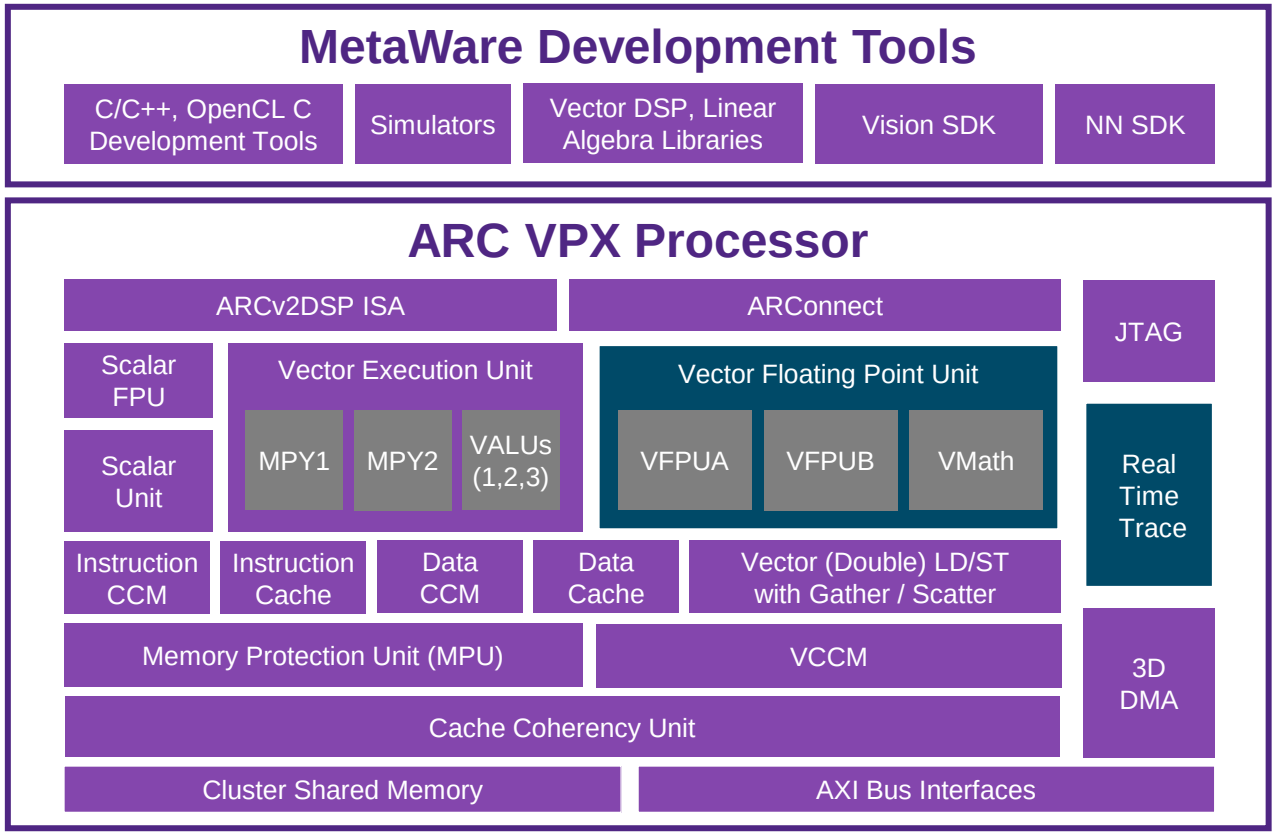
MetaWare MX NN SDK

Common Software Tools Chain for DSP and NPU



ARC VPX DSP Processor IP

Next-Generation DSP Architecture for a Data Centric World



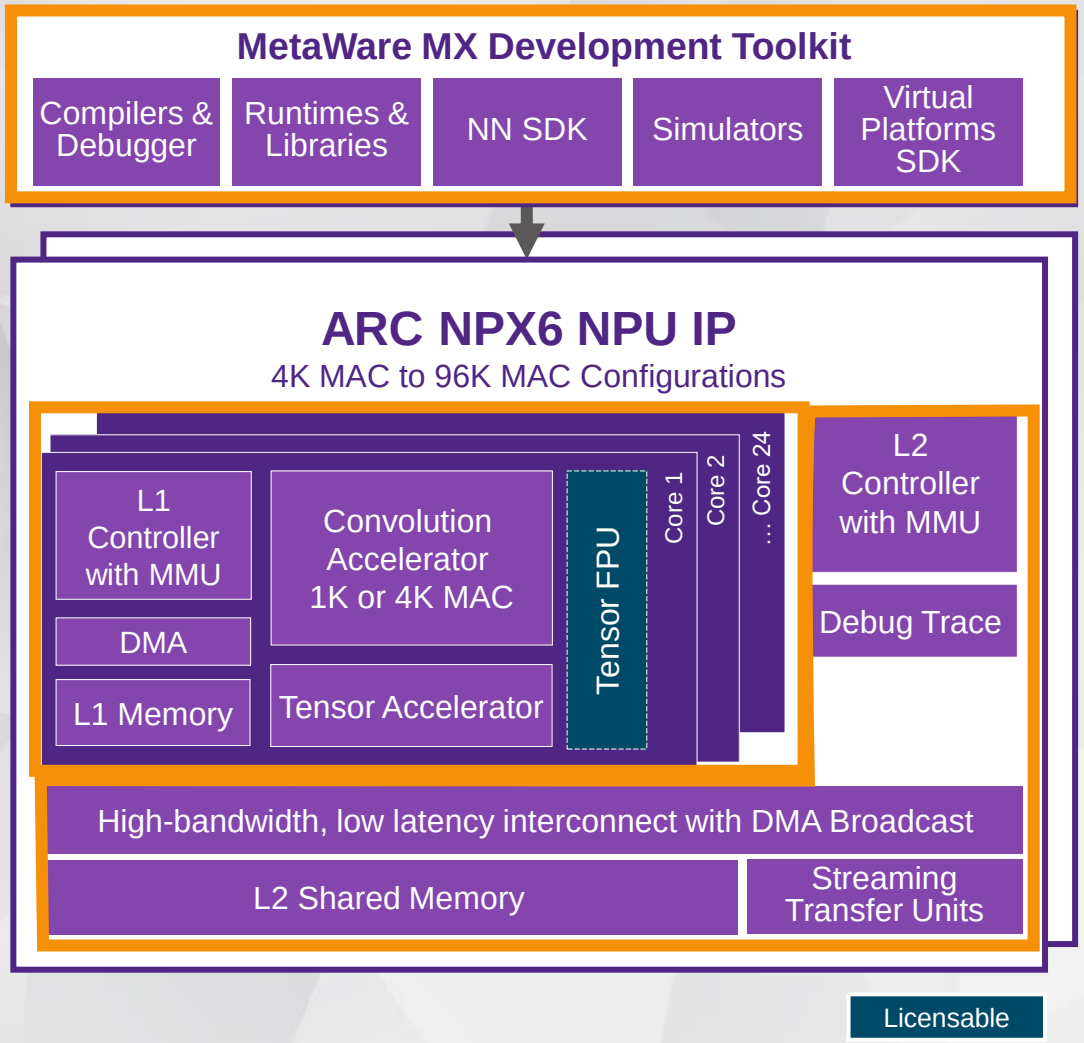
- Advanced **SIMD / VLIW DSP IP** addresses broad range of DSP workloads including RADAR/LiDAR, vision and sensor fusion
  - Vector lengths of **128-bit, 256-bit and 512-bit** enable users to select optimum PPA for required workload
  - **Scalable** and **configurable** to tune performance, area and power for specific SoC requirements
- Special features for precision results on latest algorithms
  - **Ultra high performance floating-point** processing
  - **Hardware acceleration** for linear /non-linear algebra
- Architecture, data types and software libraries optimized for **efficient machine learning**
- ISO 26262 **ASIL B and D** compliant versions for safety

Licensable Option

VPX5	512-bit vector SIMD/VLIW DSP
VPX5x2	Dual Core 512-bit vector SIMD/VLIW DSP
VPX5x4	Quad Core 512-bit vector SIMD/VLIW DSP

NEW	VPX3	256-bit vector SIMD/VLIW DSP
NEW	VPX3x2	Dual Core 256-bit vector SIMD/VLIW DSP
NEW	VPX2	128-bit vector SIMD/VLIW DSP
NEW	VPX2x2	Dual Core 128-bit vector SIMD/VLIW DSP

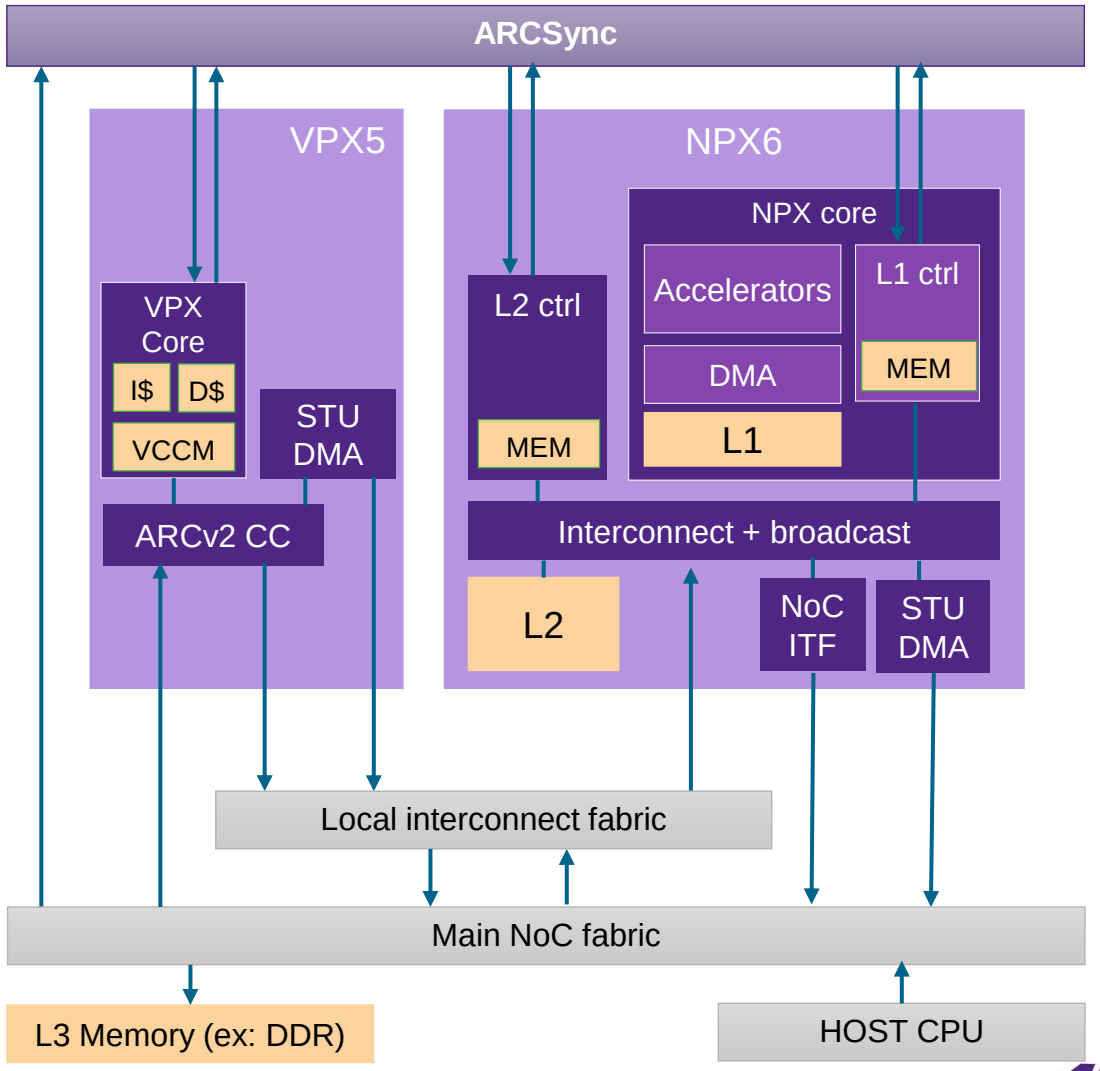
ARC NPX6 w/ 440 TOPS* Performance



- **Scalable NPX6 architecture**
  - 1 to 24 core NPU up to 96K MACS (440 TOPS*)
  - Multi-NPU support (up to eight for 3500 TOPS*)
- Trusted **software tools** scale with the architecture
- **Convolution accelerator –** MAC utilization improvements with emphasis on modern network structures
- **Generic Tensor accelerator –** Flexible Activation & support of Tensor Operator Set Architecture (TOSA)
- **Memory Hierarchy** – high bandwidth L1 and L2 memories
- **DMA broadcast lowers external memory bandwidth** requirements and improves latency

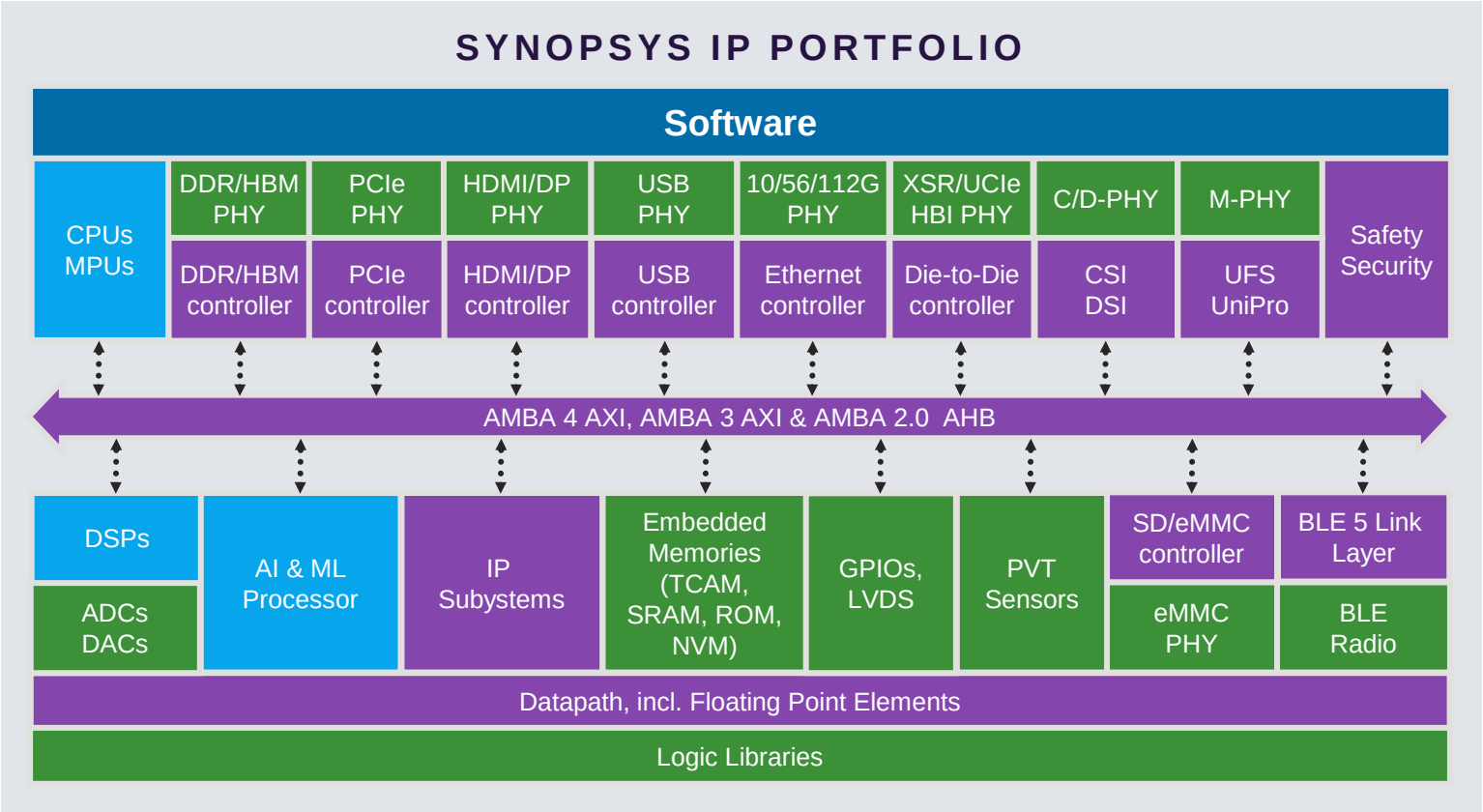
** 1.3 GHz,5nm FFC worst case conditions using sparse EDSR model*

NPX/VPX Closely Coupled for Control and Data



- NPX and VPX were designed to be used in a closely couple format
- ARCSync handles control communication between multiple NPX and VPX
- Data over Local Switch (NoC)
  - VPX can access NPX L2 memory over NPX Interconnect
  - NPX STU reads and writes from external L2 (DDR) memory
  - Hierarchical (local/global NoC) is an example. Customer can choose flat interconnect

Synopsys Broad IP Portfolio



Broadest Portfolio
Foundation IP, analog, interfaces, security, processors, subsystems

Highly Differentiated
Delivering IP at the cutting edge of technology & functionality across key markets

Committed to Your Success
5,600+ IP engineers worldwide dedicated to quality

Summary

DSP and NPU for AI Processing Summary

- AI applications require efficient Neural Network performance – especially accuracy, performance, power, area and flexibility
- NPUs and AI enabled DSPs are critical components in real-time AI SoCs – the amount of each required depends on the use case
  - Choose AI enabled DSP when … flexibility is required, and the application requires a combination of signal-processing and AI
  - Choose an NPU for AI because… an NPU provides dedicated and scalable neural network performance for mid-to-high TOPS AI tasks
  - And you might find that the combination is the right fit ….
- Select IP provider with deep experience with AI chips across broad range of applications
  - High quality IP and comprehensive, easy to use tool chains are needed for fast time to market

SYNOPSYS®
Silicon to Software™

Thank You