Solving Tomorrow's AI Problems Today with Cadence's Newest Processor

Amol Borkar

Director of Product Management & Marketing

Cadence Design Systems

cadence

AI Market Penetration

AR/VR

Automotive

IoT

Mobile

Robotics

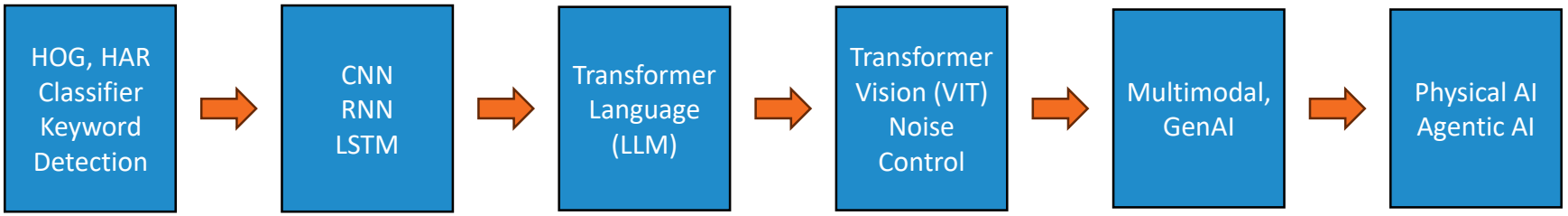
IP Camera

Data Center

Consumer Electronics

AI used for majority of workloads

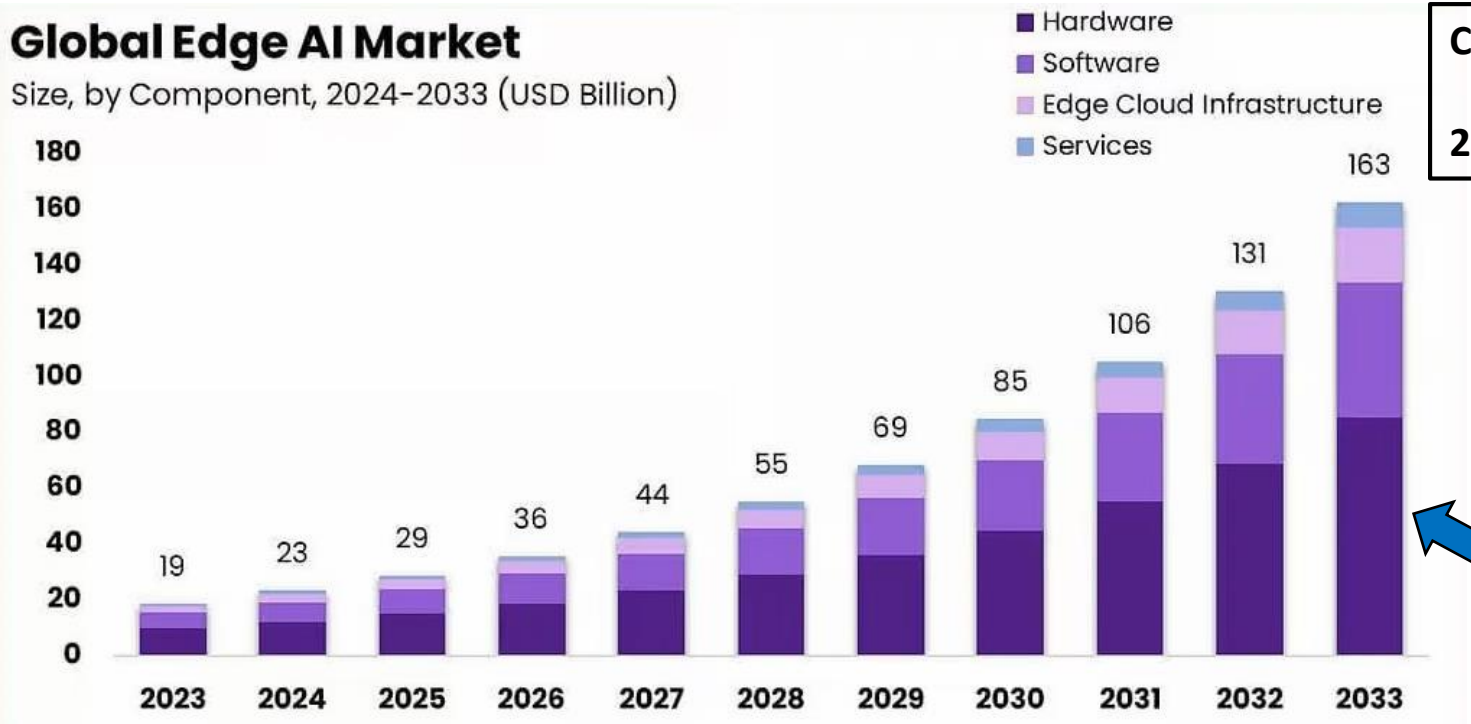
AI Inference Trend:
SoCs Need Flexibility and Programmability



Until 2014	2012 to 2018	2017 to 2021	2020 to 2023	2023 to now	2024 and beyond
	Heavy Convolution Deep NN Depthwise Convolution	Attention Matrix Multiplier Softmax LayerNorm	Multiple VIT Changing attention SWIN Transformers	Multiple data types Prompt based	

- Takes approx. 12 months to get silicon from fab
- AI landscape is constantly evolving
- Architecture and SoC design needs future proofing for the yet-to-be defined AI workload

Business and AI Market



CAGR: 24.1%

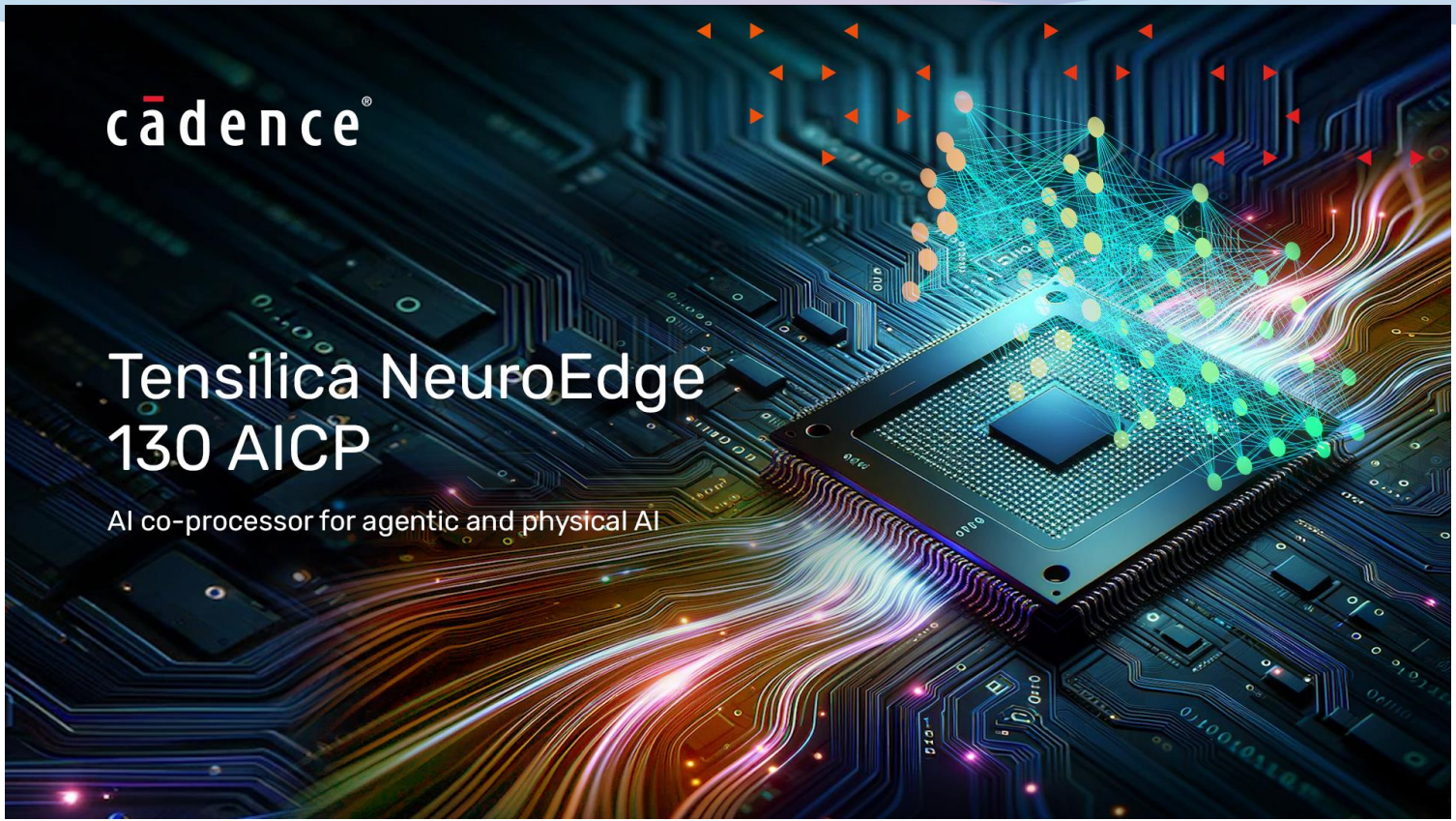
2033 Forecast: $163B

High % on AI hardware

Source: https://market.us/report/edge-ai-market/

AI Hardware != NPU
Not everything can run on NPU

- AI is proliferating & growing
  - Replacing classical methods with transformers, GenAI, agentic, multimodal processing etc.
- In SoC ➔ NPU captures majority of the AI spotlight for inferencing
- BUT! Not everything can run on the NPU! (becoming evident)
  - Large number of non-MAC layers, relu, sigmoid, tanh need to be offloaded
- Typical offload mechanisms for NPU (“unsupported layers”)
  - Offload to CPU, GPU, DSP not ideal
- Is there an ideal solution?
  - There is!



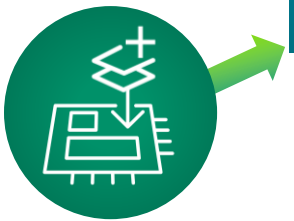
Introducing the Tensilica® NeuroEdge 130 AI Co-Processor

EMBEDDED VISION SUMMIT

New Class of Processor
Purpose-built to assist NPU for AI workloads

NeuroEdge 130 AI Co-Processor (AICP)
NPU Companion for Agentic and Physical AI

Connect to any NPU
In-House Developed
3rd party NPU IP
Cadence Neo NPU



Scalable, Configurable, and Extensible Architecture
Inherited from features, instructions
& ISA Vision DSP family

Fast TTM
Leverages existing AI libraries, tools, frameworks, SDKs for fast development

Efficient Design
512-bit VLIW SIMD
Over 30% area saving*
Over 20% power & energy savings*
1:1 Performance*

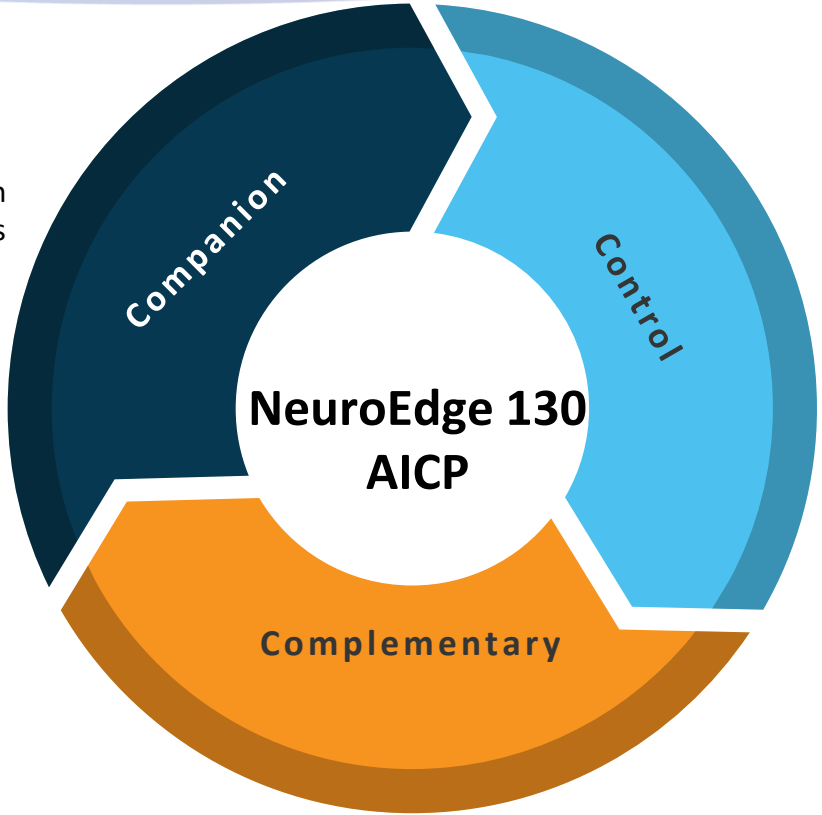
cadence

* Compared to similarly configured Tensilica DSP

Why Co-Processor?

Companion

No AI subsystem is complete with just an NPU (a vector processor is needed)



Control

Instruct NPU and send commands to perform operations

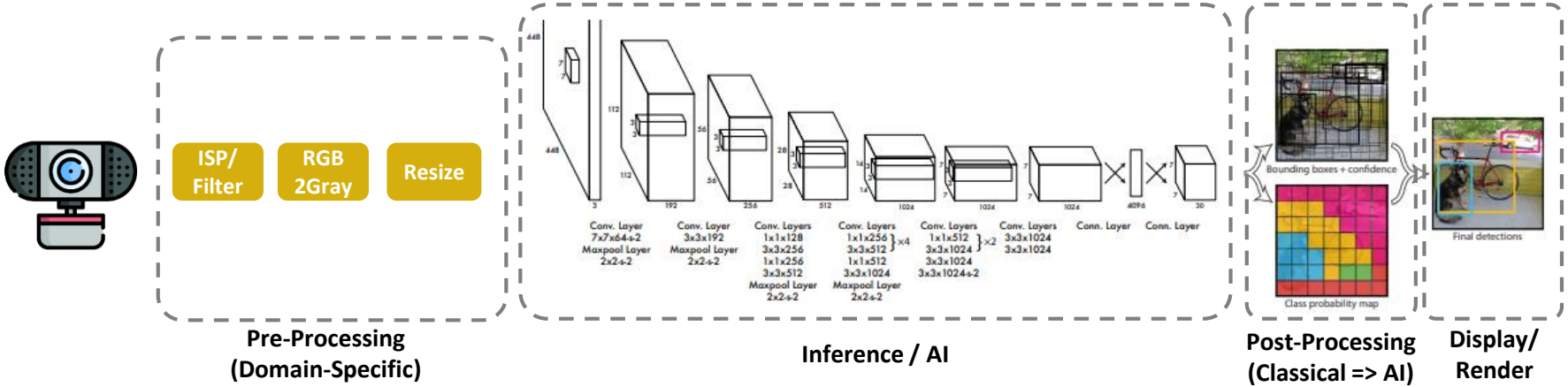
Complementary

can be offloaded to NeuroEdgeWhat does not run on NPU

Why Do We Need It?

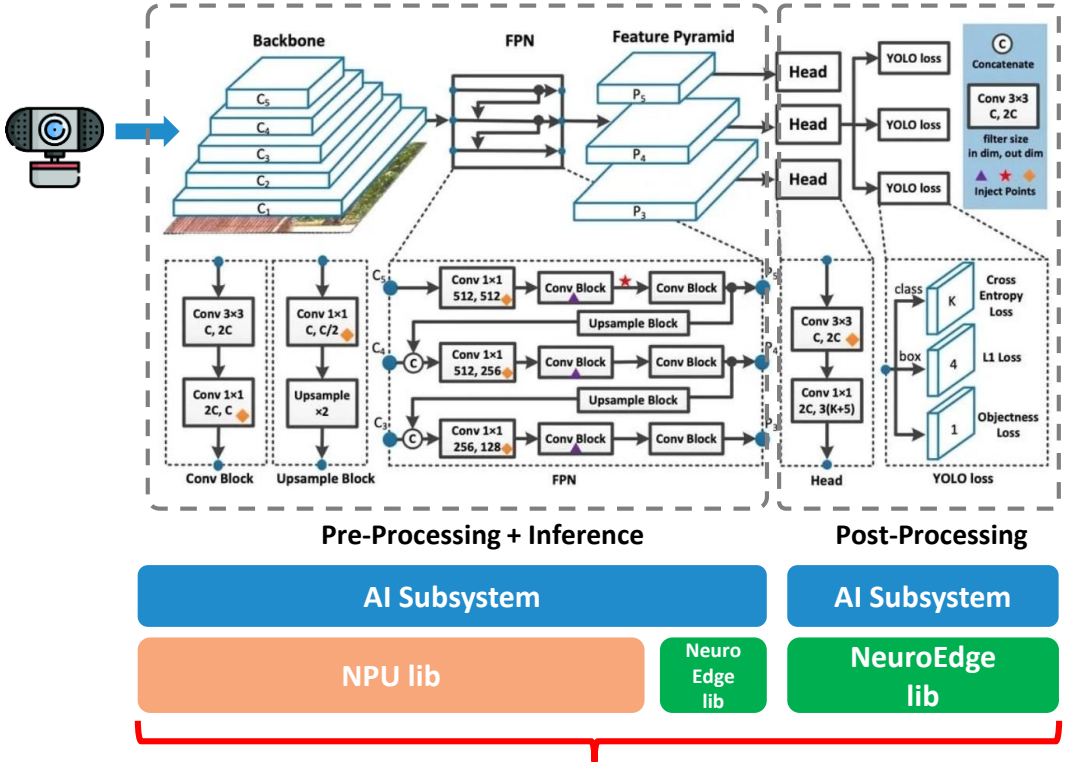
- Workloads are transforming and becoming AI end-to-end
  - Need for classical and domain specific processing is diminishing
- Not everything can run on NPU
  - Pre/post processing, non-MAC, batch norm, non-linear etc. is needed; NPU offload still needed
- AI Co-Processors use case is proven
  - Vector processor placed next to NPU in many designs (will show later)
- Customers **asking for** small AI-focused processor to complement NPU
  - Domain specific DSPs contain extra ISA and options (not for AI); customers asking for **less PPA**
- NeuroEdge AI Co-Processor is purpose-built for AI
  - ➔ Improved PPA, power, energy

NeuroEdge 130 AI Co-Processor Use Case #1 (Yolo)

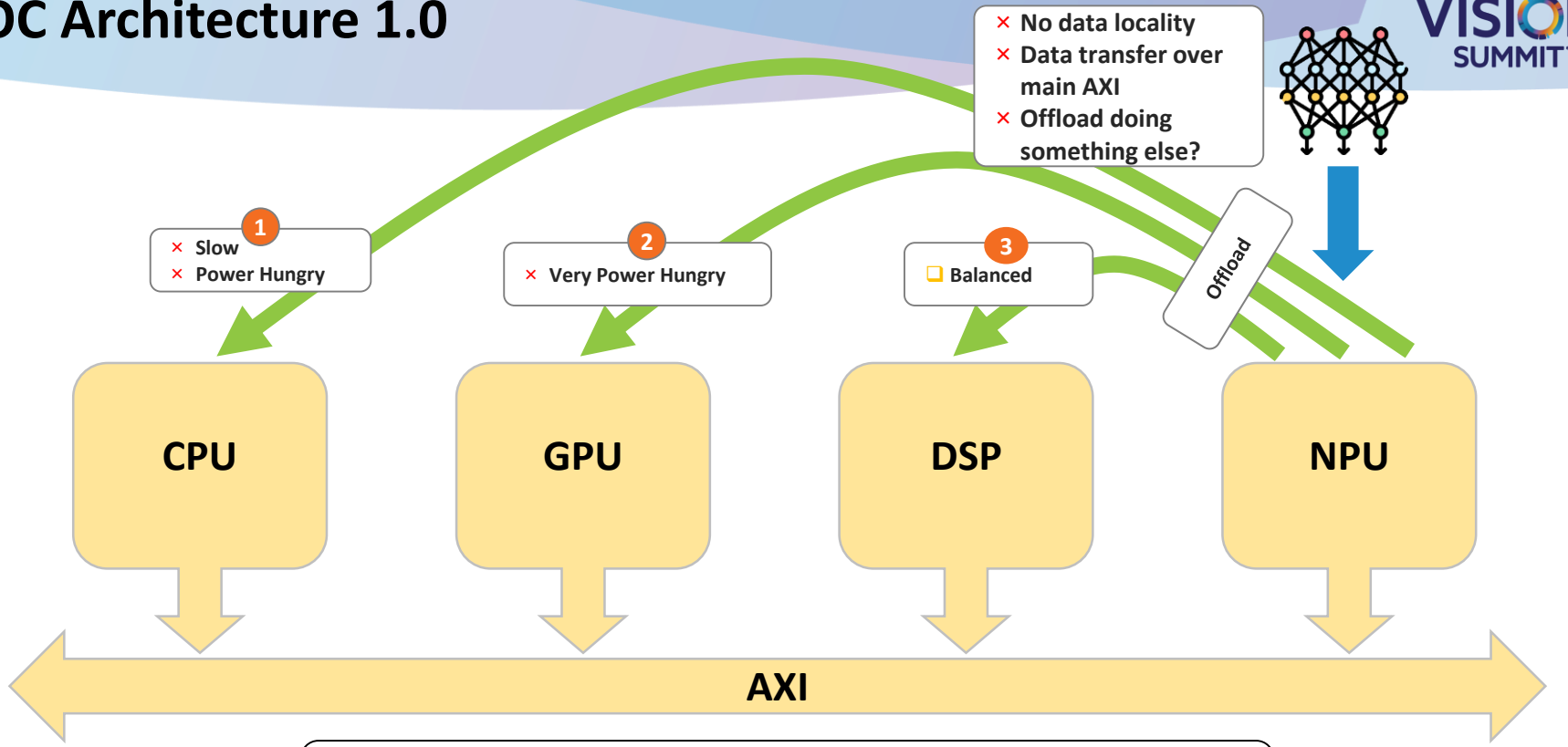


Use Case #2 (YoloS and Yolo8)

- Transformer-based Yolo
- Little to no pre-processing
  - Larger input image / aperture size (also handles resize)
  - RGB/Gray image input
- Any pre-processing is AI-based and part of inference portion
- Camera image is directly fed to network

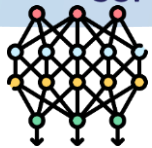


SOC Architecture 1.0



There is room for improvement

SOC Architecture 2.0 – Adding Complexes / Clusters



- ✓Offload within Subsystem
- ✓Local data movement
- ✓Internal fabric for connection
- ✓Control multiple NPUs

CPU Complex
CPU CPU
CPU CPU
CPU CPU

GPU Complex
GPU
2D Engine

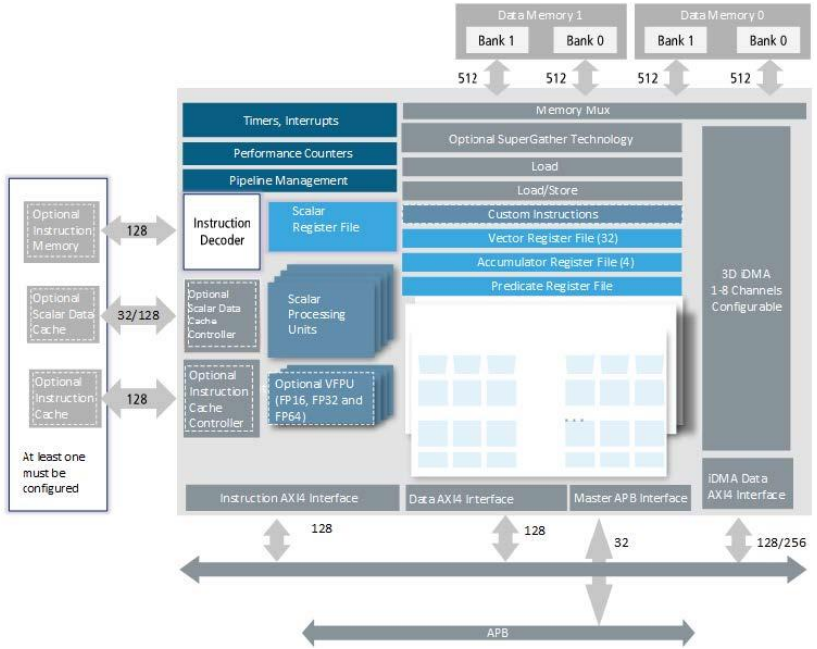
Imaging/CV Cluster "Sensor Hub"
Vision Vision
Vision HiFi
DSP DSP

AI Subsystem
NPU
Neuro Edge
NPU
Interconnect/NOC

AXI8

Architecture Overview

- VLIW SIMD Machine
  - Built upon mature Vision DSP architecture
- Configurable and scalable
  - Up to 512 8x8 MAC, FP16, FP32, BF16 support
  - Quantization Engine, Non-Linear operators
- AXI-based design with enhanced iDMA
- NPU connection
  - AXI based
  - HBDI (High Bandwidth Direct Interface)
- Operate in Manager or Subordinate mode
- ISO 26262 Fusa-ready for automotive markets



NeuroWeave™ : Unified Cadence® AI Compiler Toolchain

EMBEDDED VISION SUMMIT

TensorFlow
PyTorch
Caffe2
ONNX
...

Build AI models

One Single Software Toolchain flow for various deployment scenarios

Floating-Point or Quantized Model Graph

NeuroWeave® SDK

Model Analyzer

Model Optimizer/Mapper

Deploy AI models

NEW
NeuroEdge

Vision / RLC DSPs

Audio DSPs

AI Accelerators

Wake Word

Always ON

Object Localization

Object Classification

Sound Analytics

Emotion Detection

ADAS Systems

GenAI

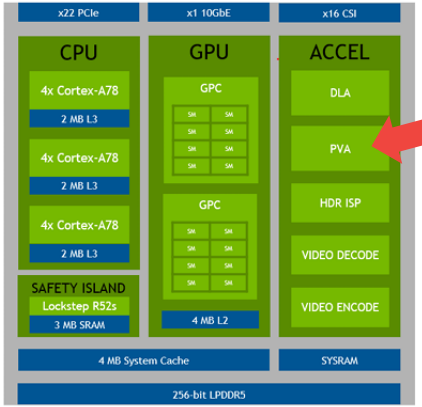
- *No hand-writing code*
- *Mature AI framework to consume networks (200+ networks)*
- *TVM based front-end to support latest NN compiler trends*
- *Connect to your NPU using NeuroConnect API*

cadence®

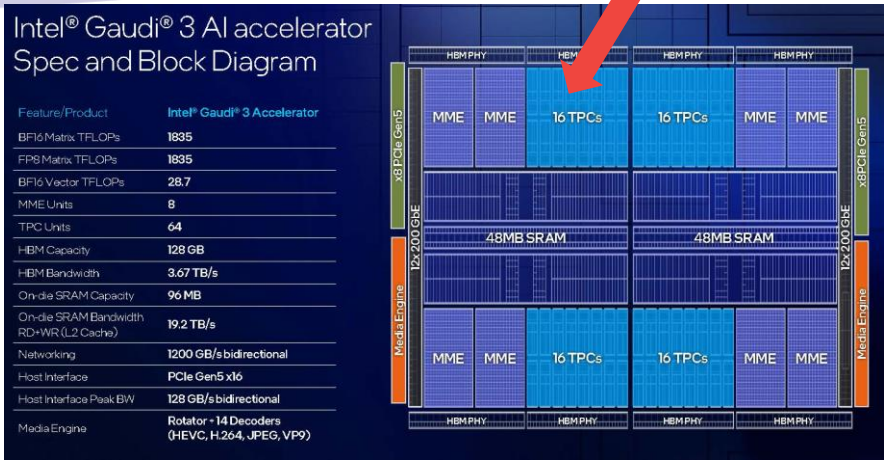
Key Benefits of NeuroEdge 130 AI Co-Processor

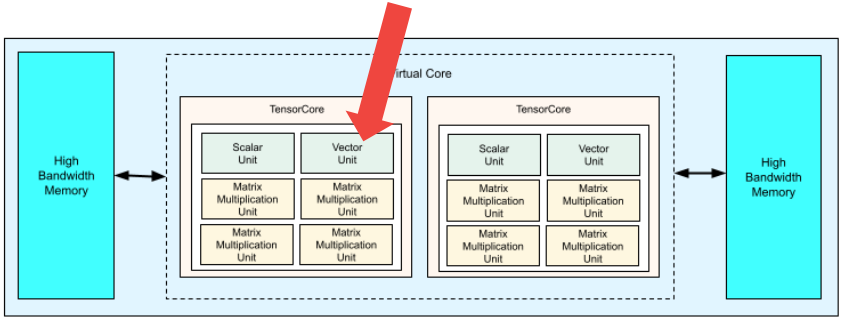
- Purpose Built AI Co-Processor
  - DSPs are vertical products that “can also do AI”
  - An AI Co-Processor for AI workloads of today and tomorrow
- Improved PPA compared to Vision DSPs
  - Over **30%** smaller area for similar config
  - Over **20%** lower power consumption for equivalent workloads
- All of the above with near identical AI performance
  - At network and kernel level
- Same software libraries, compilers, toolchain to enable AI integration
  - Leverage 3rd generation tools and frameworks for rapid development and TTM!

AI Co-Processors – Do they exist?



nVidia Orin Jetson AGX





Google TPU v4

Summary

- Confirmed the trend of growing AI usages
- Need for new products to meet such demands
- Introduced new NeuroEdge 130 AI Co-Processor (AICP) product from Cadence Tensilica®
- Flexibility to connect to any NPU
- Use case is proven and in the wild
- Mature software toolchain and libraries will enable fast prototyping and deployment

Resources

- Cadence Booth: 402
- Global Edge AI Market
- nVidia PVA Technical Brief
- Intel Gaudi 3 AI Accelerator
- Google TPU v4
- Qualcomm Hexagon 780
- Available Now! Logo
- Innovation Logo
- Presentation Template

Thank you!